"Komparative Analyse der humanen Chromosomenregion 11p15.3 um die Gene LMO1/TUB und der orthologen Region in der Maus"

Dissertation

zur Erlangung des Grades

„Doktor der Naturwissenschaften“

am Fachbereich Biologie
der Johannes Gutenberg-Universität
in Mainz

Thomas Brückmann
geboren am 07.12.1967 in Wiesbaden

Mainz 2005

Dekan des Fachbereichs:

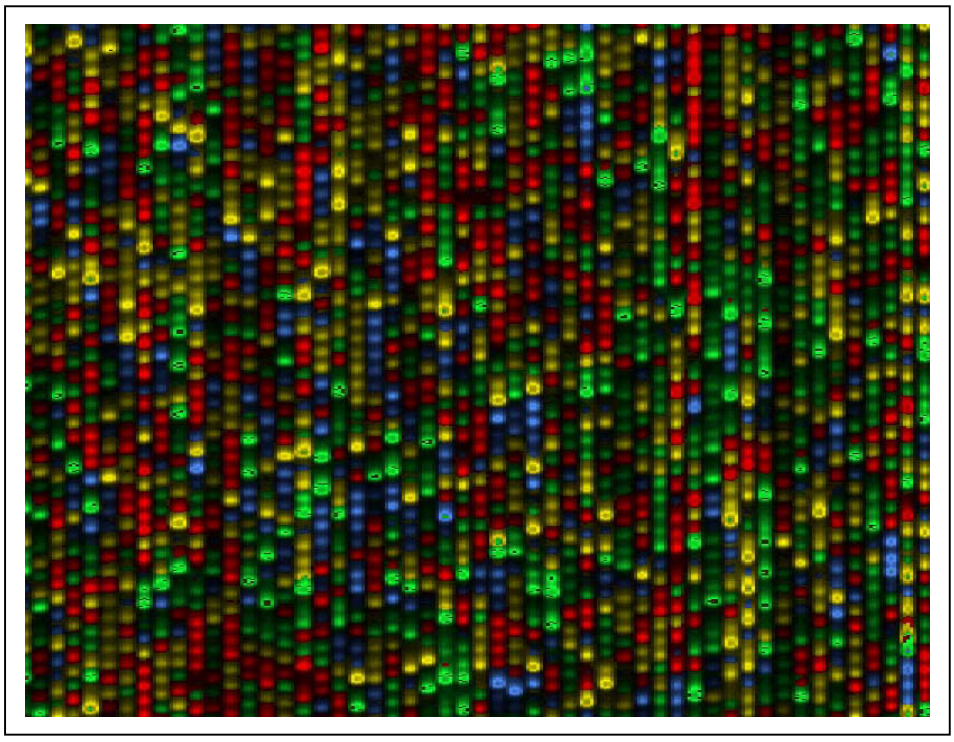
1. Berichterstatter:

2. Berichterstatter:

Tag der mündlichen Prüfung:

Der unermesslich reichen, stets sich erneuernden Natur gegenüber wird der Mensch, soweit er auch in der wissenschaftlichen Erkenntnis fortgeschritten sein mag, immer das sich wundernde Kind bleiben und muss sich stets auf neue Überraschungen gefasst machen.

(MAX PLANCK)



Ausschnitt eines Sequenziergelbildes mit fluoreszenzmarkierten Didesoxynukleotiden

INHALTSVERZEICHNIS

ABBILDUNGSVERZEICHNIS

TABELLENVERZEICHNIS

ABKÜRZUNGEN

Formelgrößen wurden mit den international gebräuchlichen SI-Einheiten, Aminosäuren wurden entsprechend ihres Einbuchstabencodes und Elemente in chemischen Verbindungen gemäß ihrer Bezeichnung im Periodensystem abgekürzt. Für die Abkürzungen gebräuchlicher Wörter der deutschen Sprache wurde gemäß Duden verfahren

Abb.	Abbildung
Acc.-Nr.	Accession Nummer
Amp	Ampicillin
APS	Ammoniumpersulfat
AS	Aminosäuren
ATP	Adenosintriphosphat
BAC	Bacterial Artificial Chromosome
bp	Basenpaare
BSA	Bovines Serumalbumin
BWSCR	Beckwith-Wiedemann-Syndrom
cDNA	"complementary DNA"
CNS	Zentrales Nervensystem
Cen	Centromer
dATP	Desoxyadenosintriphosphat
dCTP	Desoxycytosintriphosphat
dGTP	Desoxyguanosintriphospat
DNA	Desoxyribonukleinsäure
dNTP	Desoxynuleotidtriphosphat
DTT	Dithiothreitol
dTTP	Didesoxythymidintriphospat
dUTP	Desoxyuraciltriphosphat
EDTA	Ethylendiamintetraessigsäure
EMBL	"European molecular biology
EST	"expressed sequence tag"
EtOH	Ethanol
Fa.	Firma
FISH	Fluoreszenz-in situ-Hybridisierung
Gb	Gigabase
GDB	"genome database"
IPTG	Isopropyl-1-thio-β-D-galactosid
Kap.	Kapitel
kb	Kilobasen
kDa	Kilodalton
LB-Medium	Luria-Bertani Medium
LINE	"long intersperced element"
LOH	"loss of heterozygosity"
LOI	"loss of imprinting"
LTR	"long terminal repeat"
MAR	"matrix attachment region"
Mb	Megabase
MGI	TIGR Mouse Gene Index
Mio.	Million
MIT	"Massachusetts Institute of
mM	Millimol
NIH	"National Institute of Health"
nt	Nukleotid
nt	Nukleotid
OLB	"oligo labelling buffer"
OMIM	"online mendelian inheritance in men"
ORF	"open reading frame" = offener Leserahmen
PAA	Polyacrylamid
PAC	"P1-derived artificial chromosome"
PCR	"polymerase chain reaction" = Polymerase-Kettenreaktion
PDB	Brookhaven Protein Datenbank
PEG	Polyethylenglykol
PFGE	Pulsfeldgelelektrophorese
PIP	Percentage identity plot
PIR	National Biomedical Research Foundation (NBRF)-Protein Information Resource
Ψ-*Genname*	Pseudo-Gen
PRF	"Protein Research Foundation" protein sequence database
PWS	Prader-Willy-Syndrom
QTL	"quantitative trait loci"
RNA	Ribonukleinsäure
RT-PCR	Reverse Transcription - Polymerase-Kettenreaktion
SAR	"scaffold attachment region"
SDS	Natriumdodecylsulfat
SINE	"short interspersed element"
SNP	"single nucleotide polymorphism
SSCP	"single stranded conformation polymorphism"
STS	"sequence tagged site"
SV40	Simian Virus 40
Tel	Telomer
TEMED	Tetramethylethylenediamin
TIGR	The Institute for Genomic Research
TRITC	Tetramethylrhodamin isothiocyanat
TZR	T-Zell Rezeptor
Upm	Umdrehungen pro Minute
URL	"uniform resource locator"
UTR	untranslatierte Region
Vers.	Version
Vol.	Volumen
X-Gal	5-Brom-4-Chlor-3-indolyl-β-D-galactosid
YAC	"yeast artificial chromosome"

1 EINLEITUNG

Es bedurfte keine 50 Jahre, um von der erstmaligen Veröffentlichung der Struktur des Moleküls der Desoxyribonukleinsäure (DNA) durch Watson & Crick (1953) bis zur Veröffentlichung der ersten „vollständigen" Sequenz des menschlichen Genoms (IHGSC, 2001; Venter *et al*. 2001) zu gelangen. Die Erkenntnis, dass die DNA-Doppelhelix die Baupläne aller lebenden Strukturen in sich trägt, führte in dieser Zeit zu einer gänzlich anderen Betrachtungsweise von biologischen Zusammenhängen und zu einem neuen Verständnis der Spezies und der Artenvielfalt.

1.1 Methoden zur Genidentifizierung

Ziel der genetischen Forschung war – und ist es noch heute – die informativen Bereiche in der Genomsequenz zu identifizieren und sie zu analysieren. Es gilt insbesondere die exprimierten Genabschnitte zu charakterisieren, die die Informationen für die Aminosäuresequenzen der Polypeptide oder für funktionelle mRNAs kodieren. Gerade im medizinischen Kontext ist die biologische Funktion eines Gens von besonderem Interesse, da genetische Veränderungen meist Auswirkungen auf die Morphologie und die Physiologie des Organismus haben. Viele Erkrankungen basieren auf Prädispositionen, die auf genetischer Ebene manifestiert und die für die Ausprägung von besonderen individuellen Reaktionsweisen des Organismus verantwortlich sind. Im Laufe der genetischen Forschung wurden daher verschiedene Techniken etabliert, die wenigen kodierenden Bereiche der Gene im vergleichsweise riesigen Genom zu finden. Man nimmt an, dass nur ca. 3% der gesamtgenomischen DNA kodierende Informationen für die Proteinwelt besitzt. So waren es auch vor allem vererbbare, individuelle Auffälligkeiten, die erste Hinweise auf genetische Veränderungen in einer kodierenden Gensequenz gaben. Je nach Kenntnisstand über die jeweilige Erkrankung, werden unterschiedliche Vorgehensweisen angewendet, alle mit dem Ziel das entsprechende Gen in seiner gesamten Länge zu bestimmen und für weitere Untersuchung als klonierbare Sequenz isoliert zur Verfügung zu stellen.

Die Vorgehensweise des **Funktionellen Klonierens** setzt z.B. die Kenntnis der biochemischen Ursache einer Erkrankung oder das Vorliegen eines bekannten Proteins oder Enzyms voraus; die chromosomale Lokalisation ist dabei nicht von vordergründigem Interesse. Das mutmaßliche Protein wird biochemisch untersucht und auf seine krankheitsauslösenden Eigenschaften getestet. Mittels Sequenzierung vom amino-terminalen Ende her, kann die mögliche DNA-Gensequenz unter Berücksichtigung der Redundanz des genetischen Codes ermittelt werden, mit dem Ziel letztlich die Klonierung des gesuchten Gens vornehmen zu können. Mit der vorliegenden Information der Gensequenz kann dann zusätzlich auch eine Kartierung im Genom vorgenommen werden. So wurde z.B. das Gen für den Faktor VIII, das bei Hämophilie-Patienten eine Störung der Blutgerinnung hervorruft, über die Methode der Funktionellen Klonierung entdeckt und dem X-Chromosom zugeordnet (Antonarakis *et al*., 1995).

Der Ansatz der **Positionellen Klonierung** geht den reversen Weg. Es beginnt mit einer möglichst genauen Lokalisierung des Krankheitsgens in seiner chromosomalen Subregion. Die Lokalisierung im humanen Genom kann mit zytogenetischen Analysen, wie dem Auffinden von Chromosomen-Abnormalitäten in einem betroffenen Patienten-Kollektiv starten, oder über Kopplungs- und PCR-Analysen mittels bekannter informativer Mikrosatelliten- oder STS-Markern („sequence-tagged-sites") durchgeführt werden. Es folgt die Kartierung der gekoppelten Region mittels eines Klon-Contigs und das sukzessive Einengen dieses genomischen Bereiches („Chromosomen-Walking") bis hin zu einer minimalen Region, die bestenfalls nur noch aus einem Einzelklon besteht, der das gesuchte Gen, bzw. den mutierten Bereich beinhaltet. Mögliche Kandidatengene und offene Leserahmen (ORFs = „open reading frames") werden nach möglichen Mutationen hin untersucht, die Auslöser für den veränderten Phänotyp sein könnten. Sequenzunterschiede der Patientenprobe zur Wildtyp-DNA markieren dann das gesuchte Kandidatengen. Anschließende detaillierte Funktionsanalysen zeigen dann die eigentliche Funktion des Gens auf. Viele genetische Erkrankungen werden durch Mutationen in Genen hervorgerufen, deren Proteinprodukte oder biochemische Funktionen unbekannt sind. In diesen Fällen ist die Positionelle Klonierung die einzige Möglichkeit, Kandidatengene zu identifizieren. Beispiele für Gene, die mit Hilfe der positionellen Klonierung gefunden wurden, sind die Gene für das Fragile-X-Syndrom (Warren *et al.*, 1988), die Muskeldystrophie Typ Duchenne (König *et al.*, 1987) und die Chorea-Huntington-Erkrankung (Huntington's Disease Collaborative Research Group, 1993).

Eine weitere Möglichkeit Gene zu identifizieren, bietet der **Kandidatengenansatz**. Er setzt molekulare Informationen über die Pathogenese der Erkrankung oder physiologische Daten aus Tiermodell-Versuchen mit Anhaltspunkten für die genetische Ursache des veränderten Phänotyps voraus. Im Vordergrund steht hier zuerst die Suche nach potentiellen Genen oder Genbereichen, die mit dem Krankheits-Phänotyp assoziiert sind. Mittels Kopplungsanalysen über Haplotyp-Positionierung oder SNP-Analysen („single nucleotide polymorphism") können die möglichen Kandidatengene dann dem Krankheitsbild zugeordnet werden. Über PCR-SSCP-Analysen („single strand conformation polymorphism") (Hayashi, 1991) und direkter Sequenzierung können Veränderungen von putativen Kandidatengene identifiziert, bzw. Mutationen mit der Ausprägung der Krankheit gekoppelt werden. Die so durch die Mutationsanalyse hergestellte Zuordnung führt schließlich zur Identifizierung der gesuchten neuen Gensequenzen.

Eine Kombination aus beiden Strategien beschreibt der **Positionelle Kandidatengenansatz**. Mit Hilfe von QTL-Analysen („quantitative trait loci"), die komplexe, meist polygene Merkmale einem bestimmten genomischen Bereiche zuweisen, können begrenzte DNA-Abschnitte untersucht und einer gesonderten Genanalyse unterzogen werden. Insbesondere bietet die komparative Genom-untersuchung die Möglichkeit, assoziierte QTLs verschiedener Spezies aufzufinden, um so im Tiermodell Expressionsdaten zu generieren und anschließend vergleichend zu untersuchen. Mit Hilfe einer zusätzlichen Cosegregationsanalyse können gleichzeitig weitere genetische Faktoren aufgedeckt und charakterisiert werden (Bluthochdruck: Hubner & Ganten, 1995; Neurofibromatose: Schotland *et al.*, 1992). Ein weiteres Beispiele für den Erfolg dieser Vorgehensweise ist die Entdeckung des Fibrillin1-Gens auf Chromosom 15q21.1 (Magenis *et al.*, 1991), das als genetische Ursache für das

Marfan-Syndrom (OMIM #154700) angesehen wird. Der zunehmende Datenbestand an genetischen und proteinchemischen Informationen ermöglicht es mit dieser Methode in immer kürzerer Zeit und mit immer geringerem experimentellen Aufwand, neue krankheitsassoziierte Gene zu identifizieren.

Eine andere experimentelle Methode nahezu alle kodierenden Bereiche in einer klar definierten chromosomalen Region zu identifizieren, ist die Methode der **Exon-Amplifikation** (Buckler *et al.*, 1991). Genomische DNA-Fragmente eines zuvor kartierten BAC-, PAC- oder Cosmid-Klons werden in ein spezielles Expressionsvektorsystem mit artifiziellen Spleißstellen subkloniert und anschließend in COS-7-Zellen transfiziert. Diese eukaryontischen Zellen fungieren als ein *in vitro* Expressionssystem, in dem die einklonierten genomischen DNA-Fragmente unter der Kontrolle eines SV40-Promotors effizient transkribiert und bei vorhandener Spleißakzeptor- und Spleißdonorstelle auch prozessiert, d.h. unter Verlust der flankierenden Intronbereiche an die Vektorsequenz gespleißt werden. Befindet sich keine Exonsequenz mit entsprechenden Spleißstellen im klonierten DNA-Abschnitt, so wird das gesamte genomische Integrat aus dem Vektor entfernt. Eine sich anschließende reverse Transkription der zuvor präparierten RNA-Produkte wird gefolgt von zwei Restriktions- und Amplifizierungsschritten, die die DNA auf die gerichtete Klonierung in einen Sequenzierungsvektor (z.B. pBluescript II) vorbereiten. Die Sequenzierung der so generierten cDNA-Fragmente gibt Auskunft über die Basenpaarabfolge der isolierten putativen Exonsequenzen. Mit Hilfe von Datenbankvergleichen können diese dann abschließend mit bereits annotierten Genen oder exprimierten Genombereichen verglichen, und gegebenenfalls in Übereinstimmung gebracht werden.

Für die Identifizierung und Charakterisierung von proteinkodierenden Sequenzabschnitten spielt vor allem in neuerer Zeit die Vielzahl der Datenbanken und die immense Menge der unterschiedlichsten Sequenzdaten eine immer größere Rolle. Mit Hilfe verschiedener Computeralgorithmen können „***in silico***" eine ganze Reihe der unterschiedlichsten Informationen verknüpft und so schnell zu einem relativ komplexen Bild zusammengetragen werden, das abschließend nur noch mit wenigen experimentellen Versuchen verifiziert werden muss. Diese computerbasierenden Analysemethoden beschränken sich nicht nur auf den bloßen Vergleich verschiedenster Sequenzdaten, es lassen sich auch mehrere experimentelle Untersuchungstechniken rechnergestützt simulieren (z.B. „electronic northern"; „electronic PCR"). Sehr umfassende Aussagen können vor allem mit Hilfe der verschiedenen Referenzgenome anderer Spezies getroffen werden. Kodierende oder funktionell wichtige Sequenzabschnitte sind während der Evolutiv konservativer erhalten geblieben als die nichtkodierenden Bereiche. Basierend auf dieser Erkenntnis geben Sequenzhomologien im Interspeziesvergleich einen ersten Hinweis darauf, dass diese Abschnitte mit einer funktionellen Aufgabe oder Information versehen sein müssen. Gleichzeitig ist aufgrund der Interspezies-Homologie eines neu entdeckten Gens, auch das mögliche orthologe Pendant im Vergleichsorganismus bekannt, so dass diese Informationen für experimentelle funktionelle Studien in diesem Tiermodell zur Verfügung steht.

Grundlage für dieses „Datenschürfen" („data mining") sind die über das Internet zugänglichen öffentlichen Datenbanken mit vornehmlich genomischen und proteomischen Sequenzinformationen.

Insbesondere die Sequenzdatenbank GDB („Genome Database") (Fasman *et al.*, 1997), die sämtliche Sequenzinformationen der Primärdatenbanken GenBank („NIH genetic sequence database"), „DNA DataBank of Japan" (DDBJ) und „European Molecular Biology Laboratory" (EMBL) zusammenfasst, dient als großer Datenpool für eigene bioinformatische Auswertungen. Aufgrund der unterschiedlichen Nomenklatur werden Nukleotidsequenz- und Proteinsequenz-Informationen in zwei unterschiedliche Datenbank-Klassen getrennt. Die Nukleotidsequenzen werden gemäß der Art ihrer Herkunft nochmals in verschiedene Untergruppen differenziert (nr = "Non-Redundant"; dbEST = "Expressed Sequence Tags"; dbSTS = „sequence-tagged sites"; dbGSS = „genome survey sequence"; HTGS = „high-throughput genomic sequences"). Diese Eingruppierung fasst Sequenzdaten nach bestimmten Aspekten zusammen, wie z.B. nach dem Kriterium der gleichen methodischen Präparation. So beinhaltet die Datenbank dbEST nur Sequenzinformationen, die aus der cDNA-Sequenzierung stammen. Die Proteinsequenzdaten setzen sich aus Einträgen der Datenbanken Swiss-Prot, PIR, PRF, PDB und aus den translatierten Sequenzen, der kodierenden Bereiche der „Genome Database" zusammen. Für die Identifizierung von neuen Genen ist die UniGene-Datenbank (Schuler, 1997) von besonderem Interesse, da hier die Einträge der dbEST nach überlappenden oder identischen Sequenzbereichen geclustert vorliegen und oftmals zu vollständigen Transkripten, d.h. Gensequenz mit allen Exons, zusammengeführt werden konnten. Auch unvollständige Transkripte können durch die Sequenzierung der entsprechenden cDNA-Klone in einem Contig zu vollständigen Gensequenzen kompletiert werden.

Um die sich derzeit schätzungsweise alle 20 Monate verdoppelnde Informationsmenge an genetischen Daten auszuwerten und zu strukturieren, werden die teils sehr heterogenen bioinformatischen Daten aus den verschiedenen eukaryontischen Genome mit Hilfe des integrierten Genomsequenz-Management-Systems „Ensembl" zusammengefasst und untereinander verknüpft (Hubbard *et al.*, 2002) (http://www.ensembl.org). Damit lassen sich über eine interaktive graphische Oberfläche alle bekannten und vorhergesagten Gene gemäß ihrer genomischen Anordnung mit den Informationen aus den oben beschriebenen Datenbanken abrufen. Der Zugriff auf die Primärdatenbanken ist über Verknüpfungen direkt möglich und erlaubt die Darstellung sämtlicher Informationen der lokalisierten Gentranskripte. Mit dieser Datenbanktechnologie wird sowohl eine vertikale Integration von verschiedenen Daten (DNA, Protein, OMIM, PubMed, etc.), wie auch eine horizontale Vernetzung unterschiedlicher Datensätze (dbEST, dbSTS, HTGS, etc.) verschiedener Spezies erreicht und für die Forschung zugänglich.

1.2 Das Humangenomprojekt

Am ersten Oktober 1990 startete nach mehrjähriger Vorbereitungszeit das internationale „Human Genome Project". Koordiniert von einem multinationalen Forschungsverbund, bestehend aus 20 Institutionen der Nationen USA, Großbritannien, Japan und Frankreich und später auch von Deutschland und China, wurde sich das Ziel gesetzt, die 3,2 Milliarden Basen des menschlichen Erbguts zu entschlüsseln und die darin vorhandenen, damals noch auf bis zu 100.000 geschätzten Gene zu identifizieren. In 15 Jahren sollte vor allem durch die Arbeit der fünf großen Sequenzierzentren in USA und England – *Department of Energy* (DOE) *Joint Genome Institut*, *Baylor College of Medicine*, *Sanger-Centre*, *Washington University Genome Sequencing Center* und *Whitehead Institute/MIT Center for Genome Research* – und einer Reihe weiterer kleinerer Institutionen, die Sequenzierung und Kartierung des menschlichen Genoms realisiert werden. Meilensteine dieses internationalen Großvorhabens waren auch gleichzeitig die genetische Charakterisierung von anderen Modellorganismen für vergleichende Analysen, z.B. das der Maus. Ebenso sollten die benötigten technologischen und bioinformatischen Ressourcen weiterentwickelt werden und zu einer Beschleunigung der Sequenzierkapazität führen. Zielsetzung ist es bis heute, mit den erhaltenen Informationen ein besseres Verständnis für die über 6.000 krankheitsassoziierten Gene des Menschen zu bekommen (Moore, 2001) und die vielen multifaktoriellen Erkrankungen mit genetischer Prädisposition (z.B. Diabetes, Arthrose, Asthma, Alzheimer) besser verstehen zu lernen. Die pharmazeutische Industrie erhoffte sich durch die Interpretation der genetischen Daten neue Wege für die Diagnose, die Behandlung und die Therapie, die in Zukunft auch die Möglichkeit zur Prävention von humanen Erkrankungen beinhalten soll. Man nimmt an, dass den meisten Erkrankungen eine genetische Komponente zugrunde liegt, die entweder von den Eltern vererbt oder *de novo* durch erbgutschädigende Umwelteinflüsse (Strahlung, Toxine, Virusinfektionen) erworben wird. Gleichzeitig stellen die generierten genomischen Daten auch die Grundlage für die Erforschung der natürlichen Variabilität des menschlichen Genoms dar und dienen der Identifizierung und Charakterisierung der vielen Polymorphismen, die als SNPs („single nucleotide polymorphisms") durch ein eigenes Konsortium sequenziert und in entsprechende Datenbanken archiviert werden (z.B. dbSNP: www-genome.wi.mit.edu/snp/human/). Individuelle SNPs werden oft als Ursache für die teils sehr unterschiedlichen Therapieerfolge bei medikamentöser Krankheitsbehandlung angesehen (Adam, 2001). Letztlich sollen die neuen pharmakogenomischen Erkenntnisse einmal dem behandelten Arzt die Möglichkeit geben, erkrankten Personen eine individualisierte Medikamentation und Therapie anzubieten, jeweils abgestimmt auf den individuellen Genotyp.

Mit der nationalen Initiative dem Deutschen Humangenomprojekt begann 1995 die Bundesrepublik Deutschland verstärkt in die Genomforschung zu investieren und sich den internationalen Bemühungen anzuschließen. Gefördert vom Bundesministerium für Wissenschaft, Forschung und Technologie (BMBF) und der Deutschen Forschungsgemeinschaft (DFG) wurden in zwei aufeinander folgenden Runden insgesamt 135 Forschungsprojekte unterstützt (www.dhgp.de). Erklärtes Ziel war

auch für Deutschland, die systematische Identifizierung und Charakterisierung der menschlichen Gene voranzutreiben. Insbesondere sollten Gene mit medizinischer Relevanz, in Bezug auf Struktur, Funktion und Regulation charakterisieren werden. Forschergruppen, vor allem der drei großen deutschen Sequenzierzentren (*Gesellschaft für biotechnologische Forschung (GBF)*, Braunschweig, *Institut für molekulare Biotechnologie*, Jena und *Max-Planck-Institut für molekulare Genetik*, Berlin) konnten durch Ihre Arbeiten zu den internationalen Bemühungen bis Juni 2000 Sequenzinformationen von über 57,8 MB der Chromosomen 2, 3, 7, 8, 9, 11, 17, 21 und X beisteuern. Ein Sequenzierbeitrag wurde auch im Rahmen der vorliegenden Doktorarbeit erzielt, die ein Bestandteil des Kooperationsprojektes zwischen der *Kinderklinik Mainz* und dem *Institut für Molekulargenetik, gentechnologische Sicherheitsforschung und Beratung* darstellt und mit Mitteln des Deutschen Humangenomprojekts gefördert wurde.

Im Rahmen des internationalen Humangenomprojektes waren erste Erfolge bereits 1992 zu verzeichnen, als für die beiden Chromosomen 21 (Chumakov *et al.*, 1992) und Y (Foote *et al.*, 1992) erstmals eine komplette genetische Contig-Karte veröffentlicht werden konnte. Es dauerte allerdings noch fünf Jahre, bis die erste vollständige DNA-Sequenz eines menschlichen Chromosoms vollständig entschlüsselt war. Erst 1997 wurde die 33,4 Mb lange Sequenz von Chromosom 22 (Dunham *et al.*, 1999), dem kleinsten menschlichen Chromosom, mit den assoziierten Erkrankungen Neurofibromatose Typ 2 (Seizinger *et al.*, 1986) und dem DiGeorge-Syndrom (Burn, 1999) veröffentlicht. Als nächstes folgte mit 33,5 MB ein Jahr später die Sequenz von Chromosom 21, welches Gene trägt, die z.B. für das Down-Syndrom oder eine Autoimmunerkrankung („APECED-disease"; Aaltonen *et al.*, 1994) verantwortlich sind und hauptsächlich von deutschen und japanischen Forschern sequenziert wurde (Hattori *et al.*, 2000).

Im Februar 2001 wurde die Veröffentlichung der ersten kompletten Arbeitssequenz („draft sequence") aller 23 humanen Chromosomen bekannt gegeben. In elf Jahren waren 94% des gesamten menschlichen Genoms entziffert und 96% des humanen Euchromatins kartiert worden. Beschleunigt durch die zuvor 1998 entstandene Konkurrenz durch die private Initiative des Unternehmers Craig Venter mit seiner Firma CELERA GENOMICS das humane Genom allein ohne öffentliche Hilfe in nur drei Jahren zu sequenzieren, wurde in der letzten Phase auch für das öffentlich geförderte Genomprojekt die Sequenzierleistung mit zusätzlichen finanziellen Mitteln gesteigert.

Diesen beiden zeitlich parallel verlaufenden Sequenzierprojekten ist es zu verdanken, dass die humane Genomsequenz nun in zwei Versionen vorliegt, die durch zwei alternative methodische Vorgehensweisen generiert wurden. Das Ergebnis des öffentlichen Genomprojektes basierte auf den zuvor durchgeführten umfangreichen physikalischen Kartierungsarbeiten der zu sequenzierenden und in Contigs angeordneten genomischen Klone (YACs, BACs, PACs). Diese distinkten genomischen Abschnitte verpackt in den Klonsequenzen wurden dann jeweils einzeln im sog. „Shotgun"-Sequenzierverfahren (Green, 1997) sequenziert – eine Strategie, die auch in dieser Doktorarbeit als Methode Verwendung fand. Die Genomsequenz wurde später durch Assemblierung vieler überlappender Einzelsequenzen generiert. Die Verknüpfung dieser verschiedenen assemblierten

Klonsequenzen untereinander führte schließlich zu einer durchgehenden genomischen Consensussequenz eines ganzen chromosomalen Abschnitts, bzw. eines gesamten Chromosoms. Diesem hierarchischen „Shotgun"-Verfahren stand der Gesamtgenom-„Shotgun"-Ansatz von CELERA GENOMICS gegenüber. CELERA ersparte sich die stufenweise Subklonierung von Genomfragmenten in Genbanken und die Kartierung dieser Klone und sequenzierte das Genom als Ganzes in Form von Millionen zufällig generierter Einzelsequenzen. Ein massiv paralleler Rechnereinsatz führte dann diese Vielzahl von Fragmente zu einer Gesamtsequenz zusammen (Weber & Meyers, 1997). Da beide Verfahren insbesondere in stark repetitiven Bereichen nicht ganz fehlerfrei arbeiten, und zudem noch Lücken zu schließen waren, wurde in den letzten zwei Jahren an der Verifizierung der in vielen Bereichen nicht ganz eindeutigen „working draft sequence" gearbeitet.

Ein vorläufiger Höhepunkt wurde zu Beginn des Jahres 2003 erreicht, als im April das Sanger Centre die humane Genomsequenz als „finished", als beendet proklamierte und die humangenomische Sequenzierung gemeinsam mit allen anderen Beteiligten bei einer Sequenzgenauigkeit von 99,999% weitestgehend einstellte (http://www.sanger.ac.uk/Info/Press/2003/ 030414.shtml#2). Auf sämtliche Daten kann nun mit dem bereits in Kap. 1.1 erwähnten Genom-Browser „Ensembl" (http://www.ensembl.org/) zugegriffen und mit den Ergebnissen der eigenen experimentellen Arbeit oder mit bereits publizierten und Zusammenhängen verglichen und bioinformatisch untersucht werden.

1.3 Weitere Genomprojekte

In den letzten 20 Jahren wurde in jeweils eigenen Sequenzierprojekten auch an der Entschlüsselung von Genomen anderer Organismen gearbeitet. Meist wurden in den Anfangsjahren kleinere Genome wie die von Viren oder pathogenen Bakterien untersucht. Die erste vollständige Genomsequenz konnte 1995 mit der Basenpaarabfolge des Gram-negativen, fakultativ anaeroben Bakteriums ***Haemophilus influenza*** mit eine Länge von 1,83 MB veröffentlicht werden (Fleischmann *et al.*, 1995). Es folgten die Genome der Bakterien *Mycoblasma genitalium* (Fraser *et al.*, 1995), *Methanococcus jannaschii* (Bult *et al.*, 1996), *Mycoplasma pneumonia* (Himmelreich *et al.*, 1996) und *Synechocystis* Strain PCC6803 (Kaneko *et al.*, 1996). Zur gleichen Zeit konnte außerdem als komplette Genomsequenz das 4,6 MB umfassende Genom des den Wissenschaftlern gut bekannte fakultativ anaeroben Darmbakteriums ***Escherichia coli*** (Stamm K12) publiziert werden (Blattner *et al.*, 1997). *E. coli* stellt eines der bestuntersuchtesten Modellorganismen dar, mit dessen Hilfe nicht nur in der Frühzeit der Genetik grundlegende Zusammenhänge identifiziert werden konnten, wie die DNA als genetisches Material (Hershey & Chase, 1952) oder das Funktionieren von Replikation (Meselson & Stahl, 1958) und Transkription. Auch in der aktuellen molekulargenetischen Forschung ist *E. coli* mit mehr als 4.200 identifizierten proteinkodierenden Genen ein unverzichtbarer Modellorganismus und wird in Form von unterschiedlichsten Stämmen für vielerlei Klonierungsexperimente eingesetzt. Ohne dessen Hilfe wäre z.B. die Sequenzierung der unzähligen rekombinanten Klone mit genomischen DNA-Fragmenten nicht möglich gewesen. *E. coli* besitzt zudem eine medizinische Relevanz, da es von

diesem Bakterium pathogene Stämme gibt, die verantwortlich für verschiedene Entzündungen in Magen/Darmtrakt, Harnwegen, Lunge und Nervensystem sein können (Perna *et al.*, 2001). Insgesamt konnten bisher 1.328 virale, 35 viroide, 168 Phagen- und 112 mikrobielle Genome entschlüsselt werden (http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/micr.htm).

Durch die internationale Zusammenarbeit eines großen Konsortiums gelang es Mitte der 90iger Jahre nach rund sechs Jahren Forschungsarbeit die gesamte genetische Sequenz des ersten eukaryontischen Organismus zu publizieren. Die Bäckerhefe ***Saccharomyces cerevisiea*** (Stamm S288C) (Mewes *et al.*, 1997) konnte mit 16 Chromosomen, über 12 Megabasen an DNA und nahezu 6.000 identifizierten Genen (Goffeau *et al.*, 1996) in ihrer Basenpaarabfolge vorgestellt werden. Als noch einzelliges System, aber mit echtem Zellkern, stellt dieser Organismus die ideale Grundlage für Funktionsanalysen und Charakterisierungen von noch unbekannten Genen dar. Viele Stoffwechsel- und Zellzyklus-aktive Gene, deren humane Homologe mit wenig erforschten Erkrankungen in Assoziation stehen könnten, können an diesem Modellsystem untersucht werden. So erlaubt z.B. das „Yeast Two-Hybrid"-System eine Identifizierung von unbekannten Proteinbindungspartnern und Protein-Protein-Interaktionen, für Gene, deren Funktion oftmals selbst noch völlig unbekannt ist (Fields & Song, 1989). Durch „Mating", d.h. durch die Paarung zweier mit unterschiedlichen Genen transfizierter, haploider Hefe-Klone, werden die zu untersuchenden Gene bzw. ihre Proteine experimentell mit einem etwaigen Bindungspartner so zusammengeführt, dass dabei im Falle der Bindung ein Reportergen-Expressionssystem aktiviert wird. Das Signal des Reportergens gibt daraufhin einen Hinweis auf die Interaktion der Proteinbindungspartner. So konnten Erkrankungen wie die Hyperhomocyteinämie – ein Mangel an Enzym Cystathionin-β-Synthetase mit hohem Risiko zu vaskulären Erkrankungen und zum Schlaganfall – mithilfe von cys4-Deletionsmutanten der Hefe aufgeklärt werden (Kruger & Cox, 1995).

Die vollständig identifizierte Genomsequenz des ersten mehrzelligen Organismus erfolgte Ende der 90iger Jahre mit dem Nematoden ***Caenorhabditis elegans*** (C.elegans Sequencing Consortium, 1998). Als diploider Organismus besteht sein Genom aus fünf Paaren autosomaler und einem Paar Geschlechtschromosomen. Die insgesamt 97 MB DNA kodieren für über 19.000 möglicher Gene, deren putative Genprodukte zu vielen bereits bekannten Proteinen Homologie zeigen. So finden sich im direkten Vergleich zu 74% aller veröffentlichten humanen Proteine homologe Expressionsprodukte im Genom des Fadenwurms (Sonnhammer *et al.*, 1998) und nahezu 30% aller Krankheitsgene des Menschen haben Homologe in *C. elegans* (Mushegian *et al.*, 1998). Eine Besonderheit für die Forschung mit diesem Nematoden ist die Kenntnis über das genaue Schicksal einer jeden einzelnen Körperzelle. Man ist in der Lage die Entstehung der genau 558 Zellkerne der Wurmlarve zu beobachten und weiter bis zum adulten Tier mit insgesamt 959 somatischen Zellkernen zu verfolgen. Man zählt bewusst die Zellkerne und nicht die Zellen, da manche Zellen Syncytien sind und deshalb mehrere Kerne besitzen können. Ebenso ist die komplette Struktur des Nervensystems bekannt. Es lassen sich auch viele komplexe Entwicklungsprozesse, wie z.B. Apoptose (Hengartner, 1994) beobachten. Die kurze, nur 65 Stunden dauernde Generationszeit und die einfache Anatomie in einem fast transparenten Körper ließen diesen Organismus zu einem der wichtigsten Modellorganismen für

Entwicklungsbiologen und Genetiker werden. So können z.B. *in vivo* Genexpressionsstudien mit GFP-markierten Proteinen („Green Fluorescent Protein") durchgeführt und mikroskopisch beobachtet werden. Mit Hilfe von RNAi-Techniken („RNA-mediated interference") lassen sich exprimierte Gene post-transkriptionell ausschalten und die auftretenden physiologisch und morphologischen Veränderungen studieren (Kuwabara & Coulson, 2000).

Die zweite Genomentschlüsselung eines vielzelligen Organismus wurde mit der Publikation des vollständigen Genoms der Fruchtfliege ***Drosophila melanogaster*** abgeschlossen (Adams *et al.*, 2000). *Drosophila* gehört zu den genetisch bestuntersuchtesten Organismen, da mit ihr schon seit T. H. Morgan vor über neun Jahrzehnten als Versuchstier gearbeitet wird und es mittlerweile eine enorme Zahl an charakterisierten Mutanten gibt, die meist durch phänotypische Veränderungen auffallen, wie z.B. durch andere Augenfarbe und -form. Mit 165 MB, verteilt auf vier Chromosomen, umfasst das Genom ca. 13.600 vorhergesagte Gene (Adams *et al.*, 2000) und ist somit interessanterweise genärmer als das Genom von *C. elegans.* Für die vergleichende humane Sequenzanalyse stellt die Fruchtfliege einen sehr wichtigen Organismus dar, da ihre Gene homologer und näher verwandt mit den orthologen Genen der Säugetiere erscheinen als die von *C. elegans* (Mushegian *et al.*, 1998). Einen Beitrag zur komparativen Analyse leistet auch die Analyse der sog. DRES-Gene („Drosophila-related expressed sequences"), eine Sammlung von gemeinsam konservierten Gensequenzen aus Mensch und Maus, für die es ein Homolog als Fliegenmutante gibt (Banfi *et al*, 1997). Diese Zuordnung erlaubt die Charakterisierung von krankheitsrelevanten Genen des Menschen am Tiermodell *Drosophila*. So konnten mithilfe der Daten des vollständig sequenzierten *Drosophila*-Genoms für 177 von 289 untersuchten humanen Krankheitsgenen eine Zuordnung zu orthologen *Drosophila*-Genen vorgenommen werden; dies entspricht einem Verhältnis von 61%. Für humane Onco- und Tumorsuppressorgene ist die Anzahl mit 68% sogar noch etwas höher (Rubin *et al.*, 2000).

Im selben Jahr wie Drosophila wurde ein weiteres Genomprojekt abgeschlossen. Das erste vollständige Genom einer Pflanze wurde mit der DNA-Sequenz des Kreuzblütlers ***Arabidopsis thaliana*** beschrieben (Arabidopsis Genome Initiative, 2000). Das ebenfalls recht kleine Genom, bestehtend aus fünf Chromosomen, umfasst nur 125 MB DNA mit nahezu 25.500 Protein-kodierenden Genen, die zu 11.000 Familien mit ähnlicher funktioneller Diversität wie bei *Drosophila* und *C. elegans* zusammengefasst werden können. Die kurzen Generationszeiten und die hohe Vermehrungszahl machen *Arabidopsis* zu einer interessanten Modellpflanze, an der viele pflanzenspezifische Fragestellungen erforscht werden können, wie die Schädlingsabwehr und die Photomorphogenese, d.h. die lichtinduzierte Steuerung von Wachstum und Differenzierung der Pflanzen.

Der in der genetischen Forschung dem Menschen am nächsten verwandte Modellorganismus ist die Maus ***Mus musculus***, deren vollständige Genomsequenz Ende 2002 veröffentlicht wurde (Waterston *et al.*, 2002). Aufgrund der evolutiven Trennung vor etwa 60 bis 85 Millionen Jahren (Li *et al.*, 1992), besitzt das murine Genom – abgesehen von dem der Primaten – in vielen Bereichen eine außerordentlich hohe Homologie zum humanen Genom. Der diploide Chromosomensatz umfasst

insgesamt 40 Chromosomen (38 Autosomen + 2 Gonosomen) und ist mit 2,7 Gigabasen etwas kleiner als der des Menschen mit 3,2 GB. Da über 90% der Mausgenomsequenz korrespondierende konservierte Syntäniebereiche mit gleicher Anordnung der orthologen Gene aufweisen, stellt die Maus insbesondere für biochemische Studien der Geninteraktivität und für Untersuchung der Modifikationen auf Gen- und Genomebene das ideale Tier dar. Eine sehr große Zahl gut charakterisierter Mausmutanten dienen als Modelle für die verschiedensten menschlichen Erkrankungen. Etablierte gentechnische Methoden für definierte genetische Veränderung, wie etwa die gezielte Mutagenese einzelner Gene („Gene-targeting": Capecchi *et al.*, 1989) oder das Ausschalten ganzer Gene bei transgenen Knock-out Mäusen z.B. mit Hilfe des Cre/loxP Rekombinationssystem (Kuhn *et al.*, 1995, Rajewsky, 1996) geben die Möglichkeit zur Untersuchung der unterschiedlichsten Fragestellungen. Im Rahmen eines großangelegten Screeningverfahrens als Teil des Deutschen Humangenomprojektes wird versucht mit Hilfe der chemisch-induzierten Mutagenese durch das alkylierende Mutagen ENU (Ethyl-Nitroso-Harnstoff) neue Mausmutanten zu erzeugen und diese möglichst umfassend mit über 150 verschiedenen Parameter zu phänotypisieren (Hrabe de Angelis *et al.*, 2000; Nolan *et al.*, 2000). So gelang es z.B. mehrere neue Mausmutanten für komplexe multigene Erkrankungen wie Diabetes, Adipositas und Hypercholesterinämie zu identifizieren (http://www.mgu.har.mrc.ac.uk/mutabase). Erst die spezifische Kombination bestimmter, sog. „Modifier"-Gene führt hier zur Ausprägung des krankhaften Phänotyps (Montagutelli, 2000, Nadeau, 2001). Auch für andere „Volkskrankheiten", wie etwa der rheumatoiden Arthritis, steht ein Mausmodelle zur Verfügung (Nabozny *et al.*, 1996).

Zu der mittlerweile großen Zahl an weiteren Genomprojekten wie *Danio rerio* (Zebrafisch), *Rattus norvegicus* (Ratte), *Culex pipiens* (Stechmücke), *Gallus gallus* (Huhn), *Pan troglodytes* (Schimpanse) kommt dem Genomprojekt des Kugelfischs ***Takifugu rubripes*** eine besondere Bedeutung zu, da sein Genom mit 22 Chromosomen ein ähnlich großes Genrepertoire wie das des Menschen aufweist, es aber insgesamt nur 365 MB umfasst (vgl. Mensch: 3,2 GB)(Hedges & Kumar, 2002). Die sich hieraus ergebende hohe Gendichte wird durch einen nur 10%igen Gehalt an repetitiver DNA (Mensch: 44,8%) und durch sehr kurze Intronsequenzen erreicht, die im Durchschnitt weniger als 300 bp umfassen (Venkatesh *et al.*, 2000). Dies hat für die Sequenzierung des *Fugu*-Genoms die praktische Folge, dass pro Cosmid (40-45 kb) im Durchschnitt 6 bis 8 Gene, bzw. pro BAC-Klon (80-120 kb) 10 bis 15 Gene bestimmt werden können. Da sich während der Evolution die Urahnen von *Fugu* und Mensch bereits vor mehr als 450 Mio. Jahren getrennt haben (Aparicio *et al.*, 2002), gibt die Konservierung genomischer Sequenzbereiche im Interspeziesvergleich einen noch eindeutigeren funktionellen Hinweis, als Sequenzhomologien zwischen den beiden Mammaliern Maus und Mensch, die sich, wie bereits erwähnt, erst vor 60 bis 85 Millionen Jahren (Li *et al.*, 1992) eigenständig entwickelt haben. Ebenso blieb auch die Syntänie in vielen Genbereichen des Teleostiers im Vergleich zum Menschen auffallend konserviert und auch die Exon-Intron-Strukturen der kodierenden Abschnitte ist oftmals erhalten geblieben. In manchen Fällen wurde auch das gleiche alternative Spleißen bestimmter Gene beibehalten, wie am Beispiel des L1-Gens, kodierend für ein neuronales zelluläres Adhäsionsmolekül gezeigt werden konnte (Coutelle *et al.*, 1998).

Tab. 1 Genomgrößen verschiedener Vertebratenspezies im Vergleich. Die effektive, bzw. geschätzte Genomgröße, gemessen in der Anzahl der Basenpaare in Gigabasen, wird verglichen mit dem C-Wert, der die Menge an DNA eines haploiden eukaryontischen Genoms in pg pro Zelle wiederzugeben. Gleichzeitig wird der Genomgröße die Chromosomenanzahl des haploiden Genoms gegenübergestellt. Die Tabelle wurde in Anlehnung nach Venkatesh *et al.*, 2000 erstellt. Änderungen sind mit entsprechendem Index gekennzeichnet: 1: Vinogradov, 1998; 2: Ojima & Yamamoto, 1990; 3: Aparicio *et al.*, 2002; 4: Ohtsuka *et al.*, 1999, 5: Waterston *et al.*, 2002.

Spezies	Genomgröße	C-Wert [pg]	Zahl der Chromosomen (haploid)
Homo sapiens (Mensch)	2,9 GB[5]	3,5	23
Mus musculus (Maus)	2,5 GB[5]	2,45 – 3,25[1]	20
Rattus norvegicus (Ratte)	3,0 GB	3,05[1]	21
Gallus gallus (Huhn)	1,2 GB	1,25	39
Xenopus laevis (Krallenfrosch)	3,1 GB	3,2	18
Xenopus tropicalis	1,7 GB	1,78	10
Brachydanio rerio (Zebrafisch)	1,7 GB	1,8	25
Oryzias latipes Medaka (Reiskärpfling)	0,65 – 0,8[4]	1,1	24
Takifugu rubripes (Pufferfisch)	0,365 – 0,4 GB[3]	0,42[2]	22

1.4 Die Chromosomenregion 11p15

Das Chromosom 11 ist eines der sieben mittelgroßen submetazentrischen Chromosomen (Chr. 6 bis 12) des menschlichen Karyotyps und wird mit einer geschätzten Länge von 144 MB (NCBI; IHGSC, 2001) angeben. Im März 2001 waren insgesamt 128,6 MB sequenziert und im April 2003 lagen davon 106.702 kb in 11 „Contigs", d.h. als durchgängig vorliegende Genomsequenz ohne Lücken als verifizierte („finished") Sequenz vor, was 74,1% der Chromosomen-Gesamtsequenz entspricht (http://www.ncbi.nlm.nih.gov/genome/guide/HsChr11.shtml). Insgesamt konnten 1937 bekannte Gene dem Chromosom 11 zugeordnet werden, wovon erst 510 näher charakterisiert sind und 132 einer Erkrankung zugeordnet werden konnten („OMIM disease-Datenbank": http://www3.ncbi.nlm.nih.gov/Omim/) Insbesondere der kurze Arm des Chromosoms und dort die terminale Region 11p15 zeichnet sich durch eine sehr hohe Gendichte aus. Zahlreiche krankheitsassoziierte Gene konnten mit dieser etwa 20 MB umfassenden chromosomalen Region bisher in Verbindung gebracht werden.

Das Spektrum der gekoppelten Krankheiten reicht von Entwicklungsstörungen mit Fehlbildungen und Anomalien, wie beschrieben beim Beckwith-Wiedemann-Syndrom, dem Freeman-Sheldon Syndrom oder über kardiovasculäre Erkrankungen wie dem Romano-Ward-Syndrom und dem Jervell-Lange Nielsen Syndrom, bis hin zu Stoffwechselkrankheiten, wie Formen der Diabetes, Hyperinsulinämie, Nesidioblastose, Hypoparathyroidismus und Fettsucht, und neurodegenerativer Erkrankungen wie die

Niemann-Pick- und die Jansky-Bielschowsky-Krankheit. Nicht zuletzt Autoimmunkrankheiten (Sjogren-Syndrom), Hämoglobinopathien und krankhafte Schädigungen der Sinnesorgane Augen und Ohr (Usher-Syndrom) zeigen eine Kopplung zu dieser terminal liegenden Region auf dem kurzen Arm von Chromosom 11. Eine Zusammenstellung mit Verweis auf den jeweiligen OMIM-Datenbankeintrag, auf Alternativbezeichnungen und Phänotyp ist in nachfolgender Tabelle 2 aufgeführt.

Tab. 2 Auflistung aller komplexen Erkrankungen, die eine bekannte Kopplung zur chromosomalen Region 11p15 aufweisen. Als Referenz wurde der Verweis auf die Eintragung in der OMIM-Datenbank (www.ncbi.nlm.nih.gov/Omim/) gegeben. Als synonyme Bezeichnung existiert meist ein zweiter deskriptiver Name für die Erkrankung. Die Angabe des chromosomalen Kandidatengen-Locus wurde der OMIM-Annotierung entnommen

Erkrankung	Synonym	Phänotyp	OMIM	Locus
Beckwith-Wiedemann-Syndrom	Exomphalus-Macroglossie-Gigantismus-Syndrom	Frühkindliches Überwuchssyndrom mit erhöhtem Tumorrisiko (z.B. Wilmstumore)	130650	11p15.5
Romano-Ward-Syndrom	Long QT Syndrom I	Herzanomalie mit polymorpher ventrikulärer Arrhythmie	192500	11p15.5
Jervell-Lange Nielsen Syndrom	. / .	Angeborene Taubheit mit funktionellen Herzstörungen	220400	11p15.5
Hyperinsulinämie	. / .	Insulinüberproduktion	176730	11p15.5
Jansky-Bielschowsky-Krankheit	Bernheimer-Seitelberger-Syndrom/ Familiäre amaurotische Idiotie	Lipoidspeichererkrankung mit fortschreitender geistiger Demenz und Erblindung	204500	11p15.5
Sjogren-Syndrom	Sicca-Syndrom	Rheumatoide Arthritis mit autoimmunen Reaktionen	270150	11p15.5
Hämoglobinopathie	Beta-Thalassämie	Störung der Hämoglobinbildung, sichelzellenförmige Erythrozyten, Anämie	141900	11p15.5
Niemann-Pick-Syndrom	Sphingomyelin Lipidose	Lipidansammlungen in Ganglienzellen und dem zentralen Nervensystem	257200	11p15.4 - 11p15.1
Hypoparathyroidismus	. / .	Unterversorgung mit Parathyroidhormon	168450	11p15.3 - 11p15.1
Nesidioblastose	Frühkindliche Hypoglykämie	Erkrankung der Langerhansschen Inselzellen in der Bauchspeicheldrüse	256450	11p15.1
Usher-Syndrom	USH1C	Netzhautdegeneration und Innenohrschwerhörigkeit	276904	11p15.1
Fettsucht	Obesity	Angeborene Fettleibigkeit	601665	11p15
Freeman-Sheldon Syndrom	Craniocarpotarsale Dystrophie	Windmühlenflügeldeformität der Finger	193700	?

Besonders auffallend ist die Assoziation dieser Region zu vielen Krebserkrankungen. Es lässt sich eine Liste an Tumoren zusammenstellen, die Neoplasien der unterschiedlichsten Gewebe umfassen. So können sowohl Leukämien, d.h. maligne Bluterkrankungen, wie auch verschiedene embryonale Tumore, u.a. primitive neuroektodermale Tumore (PNET), Nephroblastome, Rhabdomyosarkome, Rhabdoidtumore, adrenokortikale Karzinome, Hepatoblastome und Neuroblastome mit der Chromosomenregion 11p15 in Verbindung gebracht werden. Dies betrifft außerdem Krebserkrankungen der Harnblase, der testikulären Stammzellen, der Ovarien, der Brust, der Lunge,

des Magens und der neuroepithelialen Gewebe des Gehirns (Prawitt, 1999). Die Untersuchung einiger potentieller Kandidatengene hat zudem gezeigt, dass für die Genregulation in dieser Region auch epigenetische Mechanismen eine Rolle spielen. Diese Veränderungen, die die Transkriptionsregulation eines Gens betreffen, zeigen sich nicht durch eine modifizierte Nukleotidabfolge, sondern sind geprägt durch „*über*-genetische" (= epi-) Strukturmerkmale der DNA, wie die der parentalspezifischen CpG-Methylierung. Allelspezifische Strukturveränderungen der durch das Imprinting regulierten DNA, wie die Extra-Methylierung von bestimmten Cytosin-Nukleotiden der mütterlichen, bzw. des väterlichen Genkopie, führen zu einer Inaktivierung der Expression dieser Genkopie. Störungen dieses allelspezifischen Methylierungsmusters können sich in einem LOI- („loss of imprinting"), also einem Verlust des Imprintings, oder in einem LOH-Ereignis („loss of heterozygosity") zeigen, wenn z.B. die eine elterliche Genkopie verloren gegangen ist. Diese Alterationen sind meist direkt assoziiert mit dem Auftreten bestimmter Erkrankungen, wie sie z. B. für die Entstehung des Wilmstumors (Nephroblastom) gezeigt werden konnte (Mannens *et al.*, 1988).

Ebenso befinden sich in dieser terminalen Region des Chromosoms 11 drei Bruchpunktregionen, die im Zusammenhang mit dem Auftreten des bereits erwähnten Beckwith-Wiedemann-Syndroms (BWS) stehen. Die erste BWSCR-Bruchpunkt-Region („Beckwith-Wiedemann syndrome critical region") befindet sich ca. 200 kb proximal zum Gen *IGF2* (Insulin-like growth factor II) und beinhaltet das Gen *KvLQT1*, welches für einen Kaliumkanal kodiert und krankhafte kardiologische Veränderungen hervorruft (Wang *et al.*, 1996), wie sie beim Jervell-Lange-Nielsen Syndrom beobachtet werden konnten (Neyroud *et al.*, 1997). Ungefähr fünf Megabasen centromerwärts vom BWSCR1 wird die zweite Bruchpunktregion BWSCR2 kartiert, und in geschätzten weiteren 7 MB Entfernung wird die Bruchpunktregion BWSCR3 lokalisiert, was ungefähr dem Bereich der chromosomalen Bande 11p15.3 entspricht (Redeker *et al.*, 1994) und somit mit dem in dieser Arbeit sequenzierten genomischen Abschnitt überlappt.

Aufgrund der Vielzahl an Erkrankungen mit Kopplung zu dieser Region wird angenommen, dass in der Chromosomenregion 11p15 Gene und regulatorische Elemente existieren, deren genetische Veränderungen, z.B. durch Mutationen oder deren gestörte epigenetisch gesteuerte Regulation ursächlich zur Ausbildung der oben beschriebenen Krankheitsbilder führt. Es besteht daher ein großes Interesse diese Gene und die flankierenden Genbereiche zu identifizieren und durch Charakterisierung und Studium ihrer Funktion, die Ursache der Fehlentwicklung zu finden, um in Zukunft neue Ansatzpunkte für pharmakologische und medizinisch-therapeutische Maßnahmen zu entwickeln.

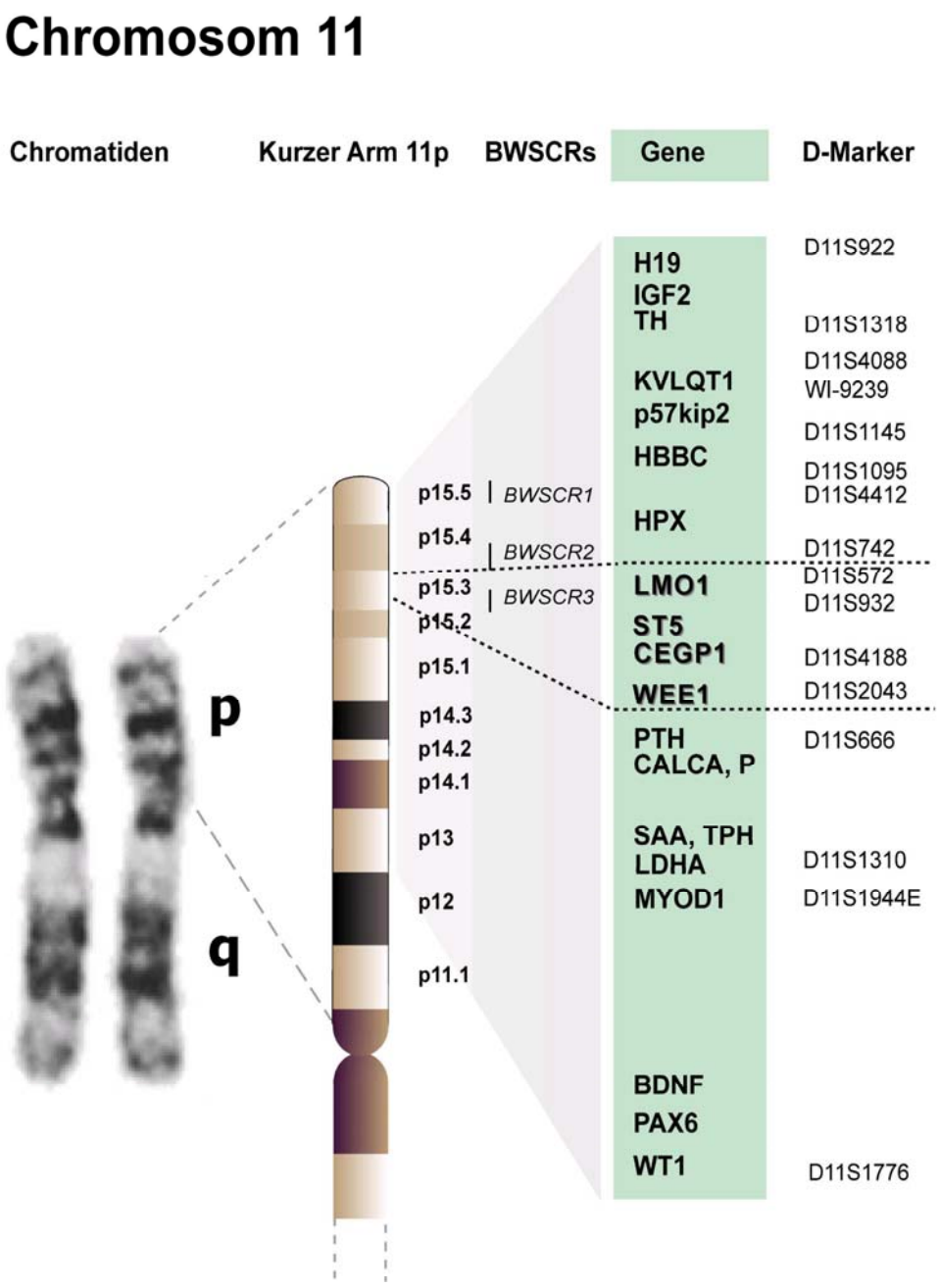


Abb. 1 Karte des humanen Chromosoms 11 mit Vergrößerung der in dieser Arbeit relevanten Bereiche. Darstellung zweier Chromosom 11 Chromatiden und deren G-Bandierung. Der kurze Arm 11p wurde zur Verdeutlichung schematisch vergrößert und gemäß seines Bandenmusters beschriftet, gleichzeitig wurden die „Beckwith-Wiedemann-Syndrom kritischen Regionen" (BWSCR, siehe Kap. 1.4) eingezeichnet. In der Ausschnittsvergrößerung (grünes Feld) sind repräsentative Gene in ihrer physikalischen Reihenfolge aus dem dargestellten Bereich von 11p13 bis 11p15.5 aufgelistet. Ebenso wurden verschiedene D-Sonden zur chromosomalen Orientierung der Genomumgebung zugeordnet.

1.5 Markergene in der chromosomalen Region 11p15.3

Zytogenetische Untersuchungen in der Arbeitsgruppe von Prof. B. Zabel mit fluoreszenzmarkierten DNA-Sonden (FISH-Analysen) zeigten, dass die Subregion 11p15.3 der chromosomale Ort für die vier „Markergene" *LMO1*, *ST5*, *CEGP1* und *WEE1* ist. Gleichzeitig konnten diese vier Gene grob in dieser Reihenfolge von Telomer nach Centromer angeordnet werden (Seipel, 1996).

Das Gen ***LMO1*** („LIM-domaine only protein") oder auch *TTG1* („T-Zell Translokationsgen 1) genannt ist ein Tumorsuppressorgen, welches mit dem Auftreten von akuter lymphoblastischer Leukämie assoziiert ist und durch die Charakterisierung des T-Zell Translokationsbruchpunktes t(11;14)(p15;q11) identifiziert werden konnte (Boehm *et al.*, 1988; Greenberg *et al.*, 1989). Als Mitglied der LIM-Domäne-Proteine, die ihren Namen aus dem Akronym der drei Gene *lin-11* (*Freyd* et al.,1990), *isl-1* (Karlsson *et al.*, 1990) und *mec-3* (Way & Chalfie, 1988) mit diesem Motiv bekommen haben, ist es in der Lage über diese Homöodomäne eine Doppel-Zinkfinger-Struktur auszubilden. Diese scheint in Analogie zu den DNA-bindenden Motiven eines Leuzin-Zippers (Landschultz *et al.*, 1988) oder eines Helix-Loop-Helix-Motivs (Murre *et al.*, 1989) eine Rolle bei der Protein-Protein-Interaktion zu spielen und bei der Transkriptionskontrolle beteiligt zu sein (Sanchez-Garcia & Rabbits, 1994). Das Gen ***ST5*** („suppression of tumorigenicity 5") oder *HTS1*-Gen („HeLa tumor suppression") ist ebenfalls ein Tumorsuppressorgen, dessen Expression in Milz, Gehirn, Muskel, verschiedenen sekretorischen Drüsen und in Niere und Lunge nachgewiesen werden konnte (Lichy *et al.*, 1992). Als dritter chromosomaler Marker diente das Gen ***CEGP1*** („CUB domain and EGF-like repeat containing protein 1") auch beschrieben unter dem Namen *Scube2* (signal peptide-CUB domain-EGF-related, gene 2) bei der Maus (Grimmond *et al.*, 2001). Dieses Gen wird aufgrund seiner Strukturhomologie zur Superfamilie der epidermalen Wachstumsfaktoren gezählt und kodiert für ein Protein bestehend aus einer N-terminalen Transmembrandomäne, sechs EGF-ähnlichen Sequenzwiederholungen und einer CUB-Domäne, eine extrazelluläre Domäne vieler Proteine, die an der Entwicklungsregulation beteiligt sind (Bork & Beckmann, 1993). Die genaue Funktion des Gens ist bislang noch unbekannt. Das Gen ***WEE1*** wird zur Familie der Cyclin-abhängigen Kinasen gezählt und ist das humane Homolog des Gens der Spalthefe *Schizosaccharomyces pompe* (Igarashi *et al.*, 1991). Das Genprodukt wirkt in der Zelle in seiner unphosphorylierten Form als negativer Regulator beim Übertritt der Zelle von der S/G2- in die Mitose-Phase (Watanabe *et al.*, 1995). Als Zellzyklus-Gen schützt *WEE1* den Zellkern durch die Phosphorylierung des cytoplasmatischen Proteins CDC2 vor seinem Eintritt in die Mitose-Phase und lässt die Zelle in der G2-Phase verharren. Während der Mitose ist die *WEE1*-Expression dagegen gesenkt, was durch zusätzliche Degradation des Proteins noch unterstützt wird.

Ein besonderes Interesse galt der Entdeckung neuer noch unbekannter Gene, die in ursächlichem Zusammenhang mit BWS oder mit anderen kartierten Erkrankungen dieser Chromosomenregion stehen könnten.

Mit Hilfe der FISH-Analyse (Fluoreszenz-in situ-Hybridisierung) auf Chromosomenpräparaten der Maus konnte gezeigt werden, dass sich die orthologen murinen Gene in der gleichen Syntäniegruppe auf

Chromosom 7 der Maus wiederfinden. Diese Ähnlichkeit in der Anordnung und der übergeordneten genomischen Architektur der homologen Gene der Region 11p15.3 deutete auf einen gemeinsamen evolutiven Ursprung hin und wurde zum Anlass genommen, den gesamten genomischen Bereich, des auf ca. eine Megabase geschätzten Abschnitts, in beiden Spezies (Mensch und Maus) vollständig zu sequenzieren und mit einem anschließenden Interspeziesvergleich der beiden genomischen Sequenzen zu charakterisieren. Dabei sollte die Homologie und die Konservierung zwischen den beiden Spezies ein Hinweis für die biologische Relevanz bestimmter Sequenzabschnitte geben. Insbesondere von Protein-kodierenden Abschnitten wurde eine höhere Homologie erwartet als für nichtkodierende Intergen- und Intronbereiche. Aber auch regulativ wichtige Abschnitte sollten sich durch eine größere Sequenzähnlichkeit auszeichnen (Oeltjen *et al.*, 1997; Ansari-Lari *et al.*, 1998; Onyango *et al.*, 2000). Der Interspeziesvergleich sollte eine Methode sein, um neue Gene und nicht kodierende regulative Elemente zu identifizieren und einer weiteren Charakterisierung zuzuführen.

Ebenso sollte die lange Zeit unklare relative genomische Lage der vier o.g. Gene eindeutig geklärt werden. Der Wissensstand zu Beginn der Sequenzierarbeit über die chromosomale Abfolge der Gene wird in nachfolgender Grafik (Abb. 2) mit drei verschiedenen Varianten zusammengefasst. Führte das Ergebnis der physikalische Kartierung bestehend aus einer Kombination aus FISH- und Restriktionsfragmentanalyse mittels Pulsfeld-Gelelektrophorese (PFGE) von Redeker & Mitarbeiter (1995) zur Platzierung des Gens *LMO1* an das proximale und *ST5* ans distale Ende, mit *WEE1* zwischen den beiden Genloci, so postulierte Higgins & Mitarbeiter (1994) diese Reihenfolge spiegelverkehrt in invertierter Ausrichtung. Auch Higgins ordnete die Gene nach den Ergebnissen einer Not I-Restriktionsfragmentanalyse an, die insgesamt die chromosomale Bande 11p15 abdeckte. Van Heyningen & Little (1995) setzten dagegen das Gen *WEE1* an das proximale und *LMO1* an das distale Ende; das Gen *ST5* wurde somit von diesen beiden flankiert. Grundlage für diese physikalische Anordnung war eine Kartierung mit YAC-Klonen („yeast artificial chromosoms") aus diesem Bereich, die die drei Gene beinhalteten und die sich an die Ergebnisse von James & Mitarbeiter (1994) anlehnte.

Ergebnisse aus Vorarbeiten der Arbeitsgruppe, insbesondere durch FISH-Kartierungsexperimente von Seipel (1996) konnten diese zuletzt dargestellte Anordnung bestätigen und waren somit Ausgangspunkt für die vorliegende Arbeit.

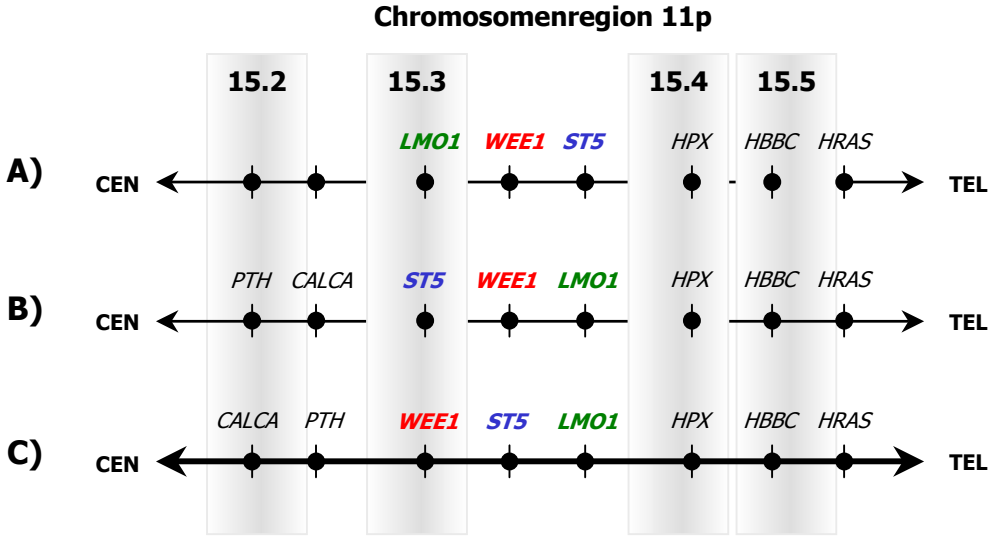


Abb. 2 Vergleich der postulierten Genanordnung nach verschiedenen Autoren. Für die Chromosomenregion 11p15.3 gab es zu Beginn der vorliegenden Arbeit unterschiedliche Vorstellungen über die relative Anordnung der Gene *LMO1*, *ST5* und *WEE1*. **A)** beschreibt die Reihenfolge nach Redeker *et al.* (1995); **B)** nach Higgins *et al.* (1994) und **C)** nach van Heyningen & Little (1995). Darstellung C: gibt hier die reale Abfolge im menschlichen Genom wieder, wie sie durch die Kartierung von Seipel (1996) und später durch die ermittelte Genomsequenz bestätigt werden konnte. Gen-Abkürzungen: CALCA = Calcitonin A; PTH = Parathyroid Hormon (OMIM 168450); HPX = Hämopexin; HBBC = Beta Hämoglobin Cluster; HRAS = „Harvey Rat Sarcoma Viral" Onkogen.

1.6 Zielsetzung der Arbeit

Die vorliegende Dissertation ist Teil des Sequenzierprojekts: „*Vergleichende Sequenzierung und Analyse einer ein Megabasen Region des Menschen (Chromosom 11p15) und der Maus (Chromosom 7)*", das in Kooperation zwischen der Universitäts-Kinderklinik und dem Institut für Molekulargenetik, gentechnologische Sicherheitsforschung und Beratung an der Johannes Gutenberg-Universität Mainz im Rahmen des Deutschen Humangenomprojektes (Förderungs-Nummer 01KW9624) durchgeführt wurde.

Es sollte der distale Abschnitt der im humanen Genom auf etwa eine Megabase geschätzten Region mit den Markergenen *LMO1*, *ST5*, *CEGP1* und *WEE1* über eine Länge von ca. 300 Kilobasen in beiden Spezies parallel sequenziert werden. Als genomischer Anker sollte das in die Chromosomenregion 11p15.3 kartierte Gen *LMO1* bzw. das orthologe murine Gen *Lmo1* auf Maus-Chromosom 7 dienen. Aufgrund von zytogenetischen Vorarbeiten durch *Seipel* (1996) war bereits ein humaner PAC-Klon bekannt. Ziel sollte sein, (i) für die zu untersuchende Region in beiden Spezies einen lückenlosen Contig aus BAC, PAC oder Cosmid-Klonen zu generieren und anhand dieser Klone zu sequenzieren. (ii) Die Sequenzermethodik und das Proben-Management sollte optimiert und auf einen internen Hochdurchsatz-Standart gebracht werden. Die Sequenziergenauigkeit sollte dem internationalen

„Bermuda-Standard" von einer Fehlerrate kleiner als 10^{-4} Genüge tragen. (iii) Anhand der verifizierten genomischen Consensus-Sequenzen von Mensch und Maus sollte eine komparative Sequenzanalyse stattfinden, die zur genomischen Charakterisierung von bekannten und neuen kodierenden DNA-Abschnitten und deren Exon-Intron-Struktur führt. Besonderes Interesse galt auch der Identifizierung von konservierten nicht-kodierenden Abschnitten in beiden Spezies. Mit Hilfe der computergestützten Auswertung und mit verschiedenen bioinformatischen Programmen sollten diese Informationen mit den Ergebnissen von der Vorhersageprogramme verknüpft und mit statischen und funktionellen Daten der Nukleotidsequenz an sich, mit der genomischen Architektur der Basenzusammensetzung und mit dem Anteil an repetitiven Sequenzbereichen korreliert werden. (iv) Ebenso sollte mit Hilfe des Homologievergleichs zu Referenzeinträgen der verschiedenen Datenbanken ein Bezug sowohl zu exprimierten DNA-Sequenzen (cDNA-Klone, ESTs), wie auch zu weiteren konservierten chromosomalen Bereichen des Menschen (NR, HTGS), bzw. zu anderen Modellorganismen hergestellt werden. Diese Zusatzinformationen sollten den prädiktiven Charakter der Programm-Algorithmen mit gesicherten experimentellen Erkenntnissen komplettieren und eine Vorgabe für die weitere gezielte Erforschung dieses chromosomalen Bereiches sein. (v) Schließlich sollte durch die Veröffentlichung der genomischen Sequenzinformation in die öffentlichen Datenbanken ein Beitrag zur Sequenzierung des Chromosoms 11 im Rahmen des Internationalen Humangenomprojektes geliefert werden.

2 MATERIAL UND METHODEN

2.1 Versuchsmaterialien

2.1.1 DNA-Klone

In der vorliegenden Arbeit wurden DNA-Klone aus insgesamt fünf verschiedenen Klonbibliotheken charakterisiert (Details unter Kap. 2.18.4). Die verifizierten Einzelklone wurden beim Ressourcenzentrum des Deutschen Humangenomprojekts (RZPD) in Berlin bezogen. Eine Übersicht aller humanen und murinen DNA-Klone ist in nachfolgender Tabelle zusammengestellt.

human Klone:

Klonbezeichnung	Arbeits-bezeichnung	Bibliothek
LANLc114B11153	cSRL 153B11	# 114
LANLc114D0263	cSRL 63D2	# 114
LANLc114E0244	cSRL 44E2	# 114
LANLc114E06155	cSRL 155E6	# 114
LANLc114G09102	cSRL 102G9	# 114
LANLc114G05109	**cSRL 119G5**	**# 114**
LANLc114H1015	cSRL 15H10	# 114
RPCIP704A061061	PAC 1061A6	# 704
RPCIP704A151184	PAC 1184A15	# 704
RPCIP704B14892	PAC 892B14	# 704
RPCIP704C1375	PAC 75C13	# 704
RPCIP704F16985	PAC 985F16	# 704
RPCIP704G1312	**PAC 12G13**	**# 704**
RPCIP704G21247	PAC 247G21	# 704
RPCIP704I07159	PAC 159I7	# 704
RPCIP704K02830	PAC 830K2	# 704
RPCIP704K03781	**PAC 781K3**	**# 704**
RPCIP704K151174	PAC 1174K15	# 704
RPCIP704K21474	PAC 474K21	# 704
RPCIP704L04217	PAC 217L4	# 704
RPCIP704M01910	PAC 910M1	# 704
RPCIP704M0940	PAC 40M9	# 704
RPCIP704M192	PAC 2M19	# 704
RPCIP704N01597	PAC 597N1	# 704
RPCIP704N14908	PAC 908N14	# 704
RPCIP704N192	PAC 2N19	# 704
RPCIP709K1436	PAC 36K14	# 709
RPCIP709N18227	PAC 227N18	# 709

Tab. 3a/b Auflistung aller charakterisierten DNA-Klone in Mensch und Maus mit Angabe der offiziellen und der Arbeitsklonbezeichnungen inkl. der Referenzbibliothek Nummerierung des RZPD. Fett hervorgehoben sind jene Klone, die in dieser Arbeit sequenziert wurden. Die Bibliothek Nr. 114 umfasste Cosmidklone mit Chromosom 11 spezifischer genomischer DNA; Bibliothek Nr. 704 war eine humane PAC-Klon-Bank mit 16-facher Redundanz. Die Bibliothek Nr. 709 umfasste ebenfalls PAC-Klone, aller nur mit einer 4-fachen Redundanz. Bibliothek Nr. 731 bestand aus einer murinen BAC- und Bibliothek Nr. 711 aus einer murinen PAC-Bank mit jeweils 13, bzw. 11-facher Redundanz. Die Redundanz einer Klonbibliothek stellt ein ungefähres Maß für die Komplexität und für die zu erwartenden positiven Klone für eine bestimmte Region im Genom dar.

murine Klone:

Klonbezeichnung	Arbeits-bezeichnung	Bibliothek
RPCIB731L01282	**BAC 282L1**	**# 731**
RPCIB731P04287	**BAC 287P4**	**# 731**
RPCIP711A08504	PAC 504A8	# 711
RPCIP711C02368	**PAC 368C2**	**# 711**
RPCIP711K08420	PAC 420K8	# 711
RPCIP711L17194	PAC 194L17	# 711

2.1.2 IMAGE-cDNA-Klone

Die Charakterisierung der neuen genkodierenden Sequenzabschnitten wurde zusätzlich durch die Analyse und Sequenzierung von cDNA-Klonen aus dem IMAGE-Konsortium verifiziert. Auch diese Klone, aufgelistet in nachfolgender Tabelle, konnten über das RZPD bezogen werden. Die RIKEN-Klone fanden nur über die annotierten Datenbank-Sequenzen Verwendung.

Tab. 4 Auflistung der IMAGE-cDNA- und Riken-Klonen, die zur Verifizierung der neuen Exonsequenzen verwendet wurden unter Angabe der IMAGE-Klon-ID, bzw. MIG-ID, der Accession-Nummer (Acc.-Nr.), des Vektorstypus und der Bibliothek des IMAGE-/Riken-Klons. Für einige Klone lagen mehr als eine Sequenz in den EST-Datenbanken vor; meist eine für den 3´und für den 5´-Bereich.

Klonbezeichnung	IMAGE-ID/ MIG-ID	Acc.-Nr.	Vektor	Bibliothek
IMAGp956F14133	2028954	AI261674 AI793003 AI793181	pT7T3D-PacI	NCI_CGAP_Kid11
IMAGp998A051337	555532	AI28482 AA118408	pCMV-Sport2	*Mus musculus* Embryo
IMAGp998A12274	46698	H10243	Lafmid BA	Soares infant brain 1NIB
IMAGp998A142098	847765	AA433755	pT7T3D-PacI	*Mus musculus* Herz
IMAGp998A194993	2028954	AI261674 AI793003 AI793181	pT7T3D-PacI	NCI_CGAP_Kid11
IMAGp998B134436	1745580	AI173044	pT7T3D-PacI	*Mus musculus* Testis
IMAGp998C21685	305228	N94985 W19458	pT7T3D-PacI	Soares_parathyroid_tumor NbHPA
IMAGp998D04240	153555	R48327 R48435	pT7T3D-PacI	Soares_2NbHBst
IMAGp998E023854	1522153	AA908701	pT7T3D-PacI	NCI_CGAP_Lu5
IMAGp998E085868	2363887	AI799590	pT7T3D-PacI	Soares_NFS_F8_9W_OT_ PA_P_S1 Foetus
IMAGp998P214518	1846916	AI239685	pT7T3D-Pac	Soares_NFL_T_GBC_S1
IMAGp998O164517	1846503	AI239997	pT7T3D-Pac	Soares_NFL_T_GBC_S1
IMAGp998D224063	1602405	AA987338	pT7T3D-Pac	NCI_CGAP_Lu5
Riken 2410002E12	MGI:1893288	AK010333	Phagemid	Riken Maus
Riken 1110003L04	MGI:1892386	AK003372	Phagemid	Riken Maus – whole body Embryo 18d
Riken 0610037M02	MGI:1892107	AK002778	Phagemid	Riken Maus – adulte Niere
Riken 4921505G21	MGI:1907093	AK014819	Phagemid	Riken Maus – adult Testis

2.2 Isolierung von DNA

Die angewandten Protokolle zur Isolierung von DNA aus Bakterienzellen basierten auf dem Prinzip der alkalischen Lyse nach Birnboim & Doly (1979). In drei aufeinander folgenden Schritten wurden die frei

im Zytoplasma der Bakterienzellen liegende rekombinante Plasmid-DNA aufgeschlossen und in isolierter Form für weitere Schritte bereitgestellt. Im ersten Schritt wurden die Bakterienzellen aus dem Medium einer meist über Nacht inkubierten Kultur abzentrifugiert, gefolgt von einem Resuspendieren der Zellen in einem Lösungspuffer. In einem zweiten Schritt fand nach Zugabe von NaOH/SDS unter alkalischen Bedingungen die Lyse der Bakterienzellen und die Denaturierung der DNA und Proteine statt. Der abschließende dritte Schritt erfolgte nach Zugabe einer hochmolaren Salzkonzentration (z.B. Kaliumacetat). Die Neutralisation führte zur Präzipitation des SDS mitsamt den denaturierten Proteinen, der chromosomalen DNA und den übrigen Zell-Debris, so dass nur noch die frei gelöste Plasmid-DNA nach einem weiteren Zentrifugationsschritt aus dem Probenüberstand isoliert werden musste. Je nach gewünschter Reinheit der DNA wurde die DNA direkt gefällt oder verschiedenen Aufreinigungsschritten unterzogen.

2.2.1 Die Präparation von PAC- und BAC-DNA

Für die Gewinnung hochmolekularer DNA aus den „low-copy-number" PAC- bzw. BAC-Klone wurde das Protokoll der alkalischen Lyse modifiziert.

Kleinere DNA-Mengen, ausreichend für die Klongrößenbestimmung mittels Pulsfeldgelelektrophorese (PFGE) (siehe Kap. 2.4.2) oder die BAC/PAC-Randsequenzierung, wurden nach einem BAC-Mini-Prep-Protokoll von Sheng & Mitarbeitern (1995) isoliert. Dabei wurden 3 bis 5 ml Übernachtkultur in mehreren Zentrifugationsrunden in einem 2 ml Reaktionsgefäß pelletiert und mit 100 µl CBIL-Sol. 1 resuspendiert. Durch Hinzugabe von 200 µl CBIL-Sol. 2 und vorsichtigem Invertieren wurde die Lyse eingeleitet. Nach 2 bis 3 Minuten wurde die Reaktion durch Zugabe von 150 µl eisgekühlter CBIL-Sol. 3 abgestoppt und für weitere 10 Minuten auf Eis inkubiert. Die sich anschließende Zentrifugation trennte die gelöste DNA von den übrigen ausgefällten Bestandteilen, so dass der Überstand abpipettiert und in ein neues Gefäß für eine Ethanolfällung (1 ml) überführt werden konnte. Nach 6 Minuten Zentrifugation (>20.000 g) wurde der Überstand verworfen, die gefällte DNA mit 70% EtOH gewaschen und anschließend getrocknet. Für die Insertgrößenbestimmung folgte eine Not I-Restriktion des gesamten Ansatzes, oder die in H_2O gelöste DNA wurde in einem 50 µl Restriktionsansatz mit Eco RV (20 Units) für 3 Stunden bei 37°C restringiert. Die benötigte DNA-Reinheit für die Sequenzierung wurde nach Abschluss der Restriktion durch eine Aufreinigung über Glassmilch (Geneclean-Kit, BIO101) erzielt.

Größere DNA-Ausbeuten wurden durch Präparation nach einem Protokoll von Pieter de Jong („MIT-Protokoll") erreicht. 300 bis 500 ml Übernacht-Bakterienkulturen wurden bei 4°C und 5.000 g für 15 Minuten abzentrifugiert und mit 20 ml MIT-Sol. 1 resuspendiert. Die alkalische Lyse erfolgte durch Zugabe von 40 ml frisch angesetzter MIT-Sol. 2 für maximal 5 Minuten und wurde durch weitere 30 ml MIT-Sol. 3 neutralisiert. Nach 5 bis 10 Minuten Abkühlung auf Eis wurde bei 10.000 g für 15 Minuten zentrifugiert. Der in ein neues Gefäß überführte DNA-haltige Überstand wurde nach Zugabe von 45 ml Isopropanol und nochmaligem 15-minütigen Zentrifugieren zur Volumenreduktion gefällt.

Das weißliche DNA-Pellet wurde erneut in 9 ml TE (10 mM Tris, pH 8,0 + 50 mM EDTA) gelöst, mit 4,5 ml KAc (7,5 M) aufgefüllt, in ein 50 ml FALCON-Röhrchen überführt und für 30 Minuten bei –70°C eingefroren. Nach 10 Minuten Zentrifugation (5.000 g) des gefrorenen Ansatzes wurde der Überstand in ein mit 27 ml EtOH-gefülltes neues Gefäß abgegossen. Nochmaliges Zentrifugieren (10 Min, 5.000 g) pelletierte die gefällte DNA, so dass diese nach Lösen mit 700 µl TE (50 mM Tris, 50 mM EDTA) für 30 bis 60 Minuten bei 37°C einem RNase-A-Verdau (70 Units, QIAGEN) unterzogen werden konnte. Danach schloss sich zur Aufreinigung der DNA eine Phenol-Chloroform-Extraktion an. Eine abschließende Fällung mit 0,7 Vol. Isopropanol und waschen mit 70% EtOH stellte die isolierte PAC-/BAC-DNA für weitere Anwendungen bereit.

Ein für die BAC-/PAC-Präparation optimiertes Protokoll der Firma QIAGEN für die QIAGEN-Tip 100 Säulen fand ebenso Anwendung. Nach alkalischer Lyse mit den QIAGEN-Puffern P1, P2 und P3 erfolgte eine Isopropanolfällung, um das Ansatzvolumen für das weitere Vorgehen zu reduzieren. Nach Auftragen des DNA-haltigen Ansatzes auf die Filtersäule und dem vorgeschriebenen Waschschritt folgte die Elution der DNA von der Säule mit 5 ml auf 65°C erwärmten QBT-Puffer, der in 5 Aliquots à 1 ml aufgetragen wurde, um ein zu schnelles Abkühlen des Puffer über der Säule zu verhindern. Abschließend erfolgte eine 30-minütige Isopropanolfällung bei Raumtemperatur. Aufgenommen wurde das gewaschene und luftgetrocknete DNA-Pellet in TE oder sterilem Wasser.

Für ein paralleles „Screening" vieler PAC-/BAC-Klone im 96er-Format wurde eine modifizierte Vorschrift des „R.E.A.L Prep 96 Plasmid Kits" (R.E.A.L-Kit QIAGEN) angewandt. Dabei wurde das Kulturvolumen verdoppeln und gleichzeitig zwei 96er Blocks im selben Anordnungsschema, mit je 1,5 ml Kulturmedium angeimpft. Um das Wachstum der Übernachtkulturen zu optimieren, wurde mit angereicherten Medien („CircleGrow"-Medium, BIO101) gearbeitet, und der Gefäßblock mit einer luftdurchlässigen Folie („Air-sheets", QIAGEN) verschlossen. Für die alkalische Lyse wurden die Überstände des zweiten Blocks auf die bereits pelletierten Zellen des ersten Blocks übertragen. Alle weiteren Schritte folgten laut Angaben des Herstellers.

2.2.2 Isolierung von Plasmid-DNA

Die Präparation der subklonierten DNA erfolgte je nach Anzahl der zu bearbeitenden Klone entweder einzeln oder angeordnet im 96er-Format mit verschiedenen kommerziellen Kits. Zur DNA-Gewinnung aus Einzelklonen wurde vor allem mit Kits der Firmen BIO101 (RPM-Kit) und BOEHRINGER („High-Pure-Plasmid-Isolation-Kit") nach Herstellerangaben gearbeitet. Die DNA der in 96er-Platten kultivierten Subklone der „Shot-gun"-Bibliotheken wurde anfänglich mit Hilfe des „QIAprep 96 Turbo Miniprep Kits" der Firma QIAGEN präpariert und durch einen Säulenaufreinigungsschritt ließ sich damit sehr saubere DNA erzielen. Während des gesamten Arbeitsablaufs blieb dabei die 96er Anordnung für nachfolgende Anwendungen insbesondere der Hoch-Durchsatz-Sequenzierung gewahrt. Im Laufe der Arbeit wurde auf den „R.E.A.L Prep 96 Plasmid Kits" (R.E.A.L-Kit, QIAGEN) umgestellt, da zum einen am späteren Sequenzierergebnis zwischen beiden Aufreinigungsmethoden kein Unterschied

ausgemacht werde konnte, und die zum anderen die zweite Variante ohne eine Säulenaufreinigung vor der Ethanolfällung kostengünstiger war.

2.3 Standardmethoden

2.3.1 Fällung

Die Fällung der DNA erfolgte je nach zu isolierender DNA-Fragmentgröße mit Ethanol oder mit Isopropanol. Sollte die Gesamt-DNA inklusive auch sehr kurzer Oligonukleotide wiedergewonnen werden, so geschah dies durch Zugabe von 2 bis 2,5 Vol. Ethanol unter Hinzugabe von monovalenten Kationen in Form von 1/10 Vol. 3 M NaAc-Lösung (pH 4,6). Die Wiedergewinnung von überwiegend längeren Fragmenten unter gleichzeitiger Abtrennung von kleineren Nukleotiden, z.B. beim Fällen von PCR-Reaktionsansätzen, wurde mit 0,7 Vol. Isopropanol und 0,5 Vol. 4 M Ammoniumacetat erreicht. Die 20 bis 30-minütige Zentrifugation fand in einer auf 16°C bis 4° gekühlten Zentrifuge statt, um einer zu starken Probenerwärmung vorzubeugen. Nach einem Waschschritt mit 70%igem EtOH wurde das DNA-Pellet im Vakuum oder bei hochmolekularer DNA an der Raumluft getrocknet.

2.3.2 DNA-Aufreinigung

DNA aus PCR-Ansätzen wurde entweder nach Volumenerhöhung auf 200 µl mit 1 Vol. Isopropanol und Zusatz von einem Volumen 4 M Ammoniumacetat bei Raumtemperatur gefällt oder durch Verwendung des „Geneclean-Kits" (BIO101) bzw. des „QIAquick PCR Purification Kits" (QIAGEN) nach Herstellerangaben aufgereinigt.

Eine weitere Methode der Aufreinigung war die Phenolextraktion der DNA. Durch Zugabe des gleichen Volumens eines Phenol/Chloroform/Isoamylalkohol-Gemisches (im Verhältnis: 25/24/1) konnten Protein-Kontaminationen aus vorhergehenden enzymatischen Reaktionen durch Denaturierung der organischen Verbindungen in die Inter- und organischen Phase abgetrennt werden. Die Beseitigung von Phenolresten vollzog sich durch Mischen und Zentrifugation mit Chloroform/Isoamylalkohol (24/1). Eine abschließende Ethanolfällung der DNA-haltigen wässrigen Phase stellte die DNA für weitere Anwendungen zur Verfügung.

2.3.3 DNA-Restriktion

Der Verdau von DNA mit Restriktionsendonukleasen diente der Charakterisierung von DNA-Fragmenten und der Isolierung von Plasmid- und genomischer DNA. Je nach DNA-Konzentration wurde mit einem Probenvolumen von 20 bis 100 µl gearbeitet. Entsprechend der verwendeten Enzyme wurde der vom Hersteller (meist NEB) angegebene 10x Reaktionspuffer verwendet. Ein 50 µl Standard-Restriktionsansatz beinhaltete z.B. ca. 5 µg DNA und etwa 20 Units des verwendeten

Restriktionsenzyms. Die Inkubation fand im Wasserbad bei 37°C statt. Bei großen Restriktionsansätzen mit mehreren Mikrogramm zu restringierender DNA wurde der Verdau mit einem einstündigen erneuten „Spike" nach Hinzugabe von nochmals 10 bis 20 Units Enzym abgeschlossen.

2.4 Gelelektrophoresen

2.4.1 Agarose-Gelelektrophorese

Die Auftrennung von PCR-Produkten oder restringierter DNA entsprechend ihrer Fragmentgröße fand je nach DNA-Molekülgröße in 0,7 bis 4%igen Agarosegelen in 0,5x TBE-Puffer statt. Als Gelelektrophoreseeinheit dienten Horizontalgelkammern unterschiedlicher Größe mit einer Probenkapazität von 12 bis 192 Spuren. Die Auftrennung fand in Abhängigkeit der gewählten Gelkammergröße und der Agarosekonzentration bei 80 bis 200 Volt statt. Als Molekulargewichtsmarker dienten standardmäßig eine 100 bp- bzw. 123 bp-Leiter (GIBCO) oder Hind III-geschnittene Lambda-DNA (BOEHRINGER). Zur Erhöhung des spezifischen Gewichts der Probe wurde diese vor dem Auftragen mit 6x Orange-Dye (FERMENTAS) versehen.

Zur Darstellung der DNA wurden die Gele in einer Ethidiumbromid-Lösung 10 bis 15 Minuten gefärbt, 15 Minuten in Aqua dest. gewässert und unter UV-Licht (λ = 312 nm) mit einem Gel-Imaging-System (HEROLAB) betrachtet und dokumentiert.

2.4.2 Pulsfeldgelelektrophorese (PFGE)

Um die genomischen Integratgrößen der zu charakterisierenden PAC-/BAC-Klone zu bestimmen, wurde die präparierte und mit Not I oder einem anderen selten schneidenden Enzym („rare-cutter"), wie etwa Xho I, geschnittene BAC-/PAC-DNA mit einer „Contour-clamped-homogeneous-electric-field"-(CHEF) Elektrophorese aufgetrennt. In einem jeweils um 120° alternierenden elektrischen Feld, welches durch hexagonal angeordneten Elektroden induziert wird, konnten so DNA-Fragmente im Megabasenbereich ihrer Molekülgröße nach linear aufgetrennt werden. Vor dem Auftragen wurden DNA-Proben wurden in Agaroseblöckchen eingegossen und dadurch immobilisiert. Die Auftrennung fand in einem 1%igen GTG-„low-melting" Agarosegel statt. Die PFGE-Steuereinheit (BIORAD) wurde zuvor auf einen Fragmentgrößenbereich von 10 bis 150 kb eingestellt (weitere Parameter siehe Tab. 5). Um die Linearität in der Auftrennung des eingestellten Bereiches zu erzielen, wurde mithilfe eines speziellen Algorithmus automatisch eine Zeitverlängerung für den Wechsel der Stromflussrichtung bei einer Gesamtlaufzeit von 20:18 h errechnet und eingestellt. Als Laufpuffer diente 0,5x TBE, der auf 14°C abgekühlt war. Als Molekulargewichtsmarker wurden λ-DNA-Concatemere (PHARMACIA) mit Längen von jeweils 49,5 kb, bzw. einem Vielfachen davon und Hind III-geschnittene Lambda-DNA verwendet.

Tab. 5 Konfigurationsparameter für die Steuereinheit (ROM Version 2.1, Biorad) zur Pulsfeld-gelelektrophorese

Parameter	Wert
Fragmentgrößenbereich	10 kb bis 150 kb
Kalibrierungsfaktor	1,0
Gradient:	6,0 V/cm
Laufzeit:	20 Stunden 18 Minuten
eingeschlossener Winkel	120°
Zeitintervall für Spannungswechsel: zu Beginn / am Ende	0,47 Sekunden / 12,91 Sekunden
„Ramping"-Faktor a	linear

2.4.3 Polyacrylamid-Sequenziergelelektrophorese (PAA-Gele)

Das standardmäßige Polyacrylamidgel zur Sequenzprobenauftrennung wurde nach Angaben des Herstellers (ABI APPLIED BIOSYSTEMS) gegossen und als 5%iges Polyacrylamidgel mit 7M Harnstoff in einem Volumenansatz von 50 ml angesetzt. Die Gellösung bestand aus 8,4 ml 30% PAA (29:1)(ROTH), 21 g Harnstoff (ROTH), 6 ml TBE-Puffer, 20 ml autoklaviertes deionisiertes Wasser. Zur Polymerisation wurden nach vorherigem Sterilfiltrieren und Entgasen 20 µl TEMED und 300 µl 10%iges APS hinzupipettiert. Nach mindestens 90-minütigem Aushärten (siehe auch Kap. 2.12) wurde das Gel mit Aliquots der zu sequenzierenden Proben beladen. Der Gellauf fand für eine durchschnittliche Leseweite von 800 bp über 48 cm Gelplatten unter den Parametern des „SeqRun48A-1200" Laufmoduls statt. Hierzu war eine Laufzeit von 10 Stunden bei 2.400 V und 200 W Leistung notwendig.

2.5 DNA-Isolierung aus Gelen

Die Isolierung der gelelektrophoretisch aufgetrennten und mit einem Skalpell an distinkten Banden aus dem Agarosegel ausgeschnittenen DNA-Fragmente fand unter Zuhilfenahme der Kits der Firmen BIO101 („Geneclean-Kit") und QIAGEN („QIAquick Gel Extraction Kit") statt. Hierbei wurde das Gelstück unter Erwärmung auf 55°C und spezifischen Pufferbedingungen verflüssigt und die freiwerdende DNA in Gegenwart eines chaotropen Salzes an eine Glassmilch gebunden. Nach einem Waschschritt wurde die DNA mit Aqua dest. von der Glassmilch wieder eluiert.

2.6 Isolierung von RNA

Die Isolierung der Gesamt-RNA und die sich anschließende reverse Transkription in cDNA diente dazu, die zuvor bestimmten putativen Exonsequenzen zu verifizieren und ihre exakten Exon-Intron-Grenzen zu bestimmen. Die RNA-Präparationen aus verschiedenen Gewebeproben wurde mit Hilfe des „RNeasy Mini-Kits" der Firma QIAGEN nach Herstellerempfehlung durchgeführt. Dieser Kit arbeitet nach der „Single-Step"-Methode (Chomczynski & Sacchi, 1987), die mit einer chaotropen Salzlösung (Guanidinisothiocyanat) sehr effektiv Proteine (auch RNasen) denaturiert und inaktiviert. Das meist tiefgefrorene Gewebematerial wurde entweder mit einem Potter Homogenisator nach vorherigem Zerkleinern der Probe mittels Mörser und Pistill in flüssigem Stickstoff oder mit Hilfe eines „QIAshredders" (QIAGEN) homogenisiert und 2 Minuten in 350 µl RLT-Puffer zentrifugiert. Nach Zugabe von einem Volumen 70%igem EtOH wurde die Probe auf die RNeasy-Säulen, zum Binden der RNA an dieselben, gegeben und für 15 Sekunden kurz anzentrifugiert. Es folgten zwei Waschschritte mit RW1- und RPE-Puffer und die Elution der Gesamt-RNA mit 30 bis 50 µl RNase-freies DEPC-Wasssers (= Diethylpyrocarbonat). Die Menge an isolierter RNA wurde anschließend photometrisch bestimmt und die Proben bis zur weiteren Verwendung bei –70°C gelagert.

2.7 Polymerasekettenreaktion (PCR)

2.7.1 Standard-PCR

Zur Amplifikation spezifischer DNA-Abschnitte wurde die Polymerasekettenreaktion nach Saiki & Mitarbeitern (1988) eingesetzt. In einem Ansatzvolumen von 35 oder 50 µl bestehend aus 10 mM Tris/HCl (pH 8,3), 50 mM KCl, 1,5 mM $MgCl_2$, 0,1 mg/ml Gelatine, 100 mM dNTPs (dATP, dCTP, dGTP, dTTP) und 10 bis 20 pMol der den gewünschten Bereich flankierenden Primer, wurde die Reaktion zusammen mit 2,5 Units Taq-DNA-Polymerase und 10-100 pg DNA-„Template" auf Eis angesetzt. Die DNA-Vervielfältigung fand während 30 Zyklen in drei Temperaturschritten statt, begrenzt von einer verlängerten ersten Denaturierungsphase von 2 bis 5 Minuten bei 94°C und einer abschließenden Elongationsphase („final extension") von 7 Minuten bei 72°C. Hierbei dauerte der Denaturierungsschritt bei 94°C für 10 sec bis 1 min, das „Annealing" je nach Schmelzpunkt der Primer bei 48°C bis 60°C für 30 sec und die Elongation, je nach Größe des zu amplifizierenden Fragment, bei 72°C zwischen 30 sec und 2 min. Da ausschließlich mit PCR-Geräten mit beheizbarer Abdeckung gearbeitet wurde, entfiel eine Überschichtung der Ansätze mit Mineralöl.

Weitere Modifikationen des Standard-PCR-Protokolls stellten die „Hot-Start"-PCR, mit Zugabe der Taq-Polymerase nach dem initialen Denaturierungsschritt und die „Nested"-PCR, mit einem dritten im PCR-Amplifikat liegenden Primer dar, welche insbesondere bei Reamplifikationen Anwendung fanden.

2.7.2 „Touchdown"-PCR

Bei einem Schmelztemperaturunterschied der eingesetzten Primer von mehr als 5°C wurde eine sog. „Touchdown"-PCR benutzt. Die „Annealing"-Temperaturen wurde hierbei pro Zyklus sukzessive vom oberen auf den unteren Schmelztemperaturwert der beiden Primer abgesenkt, um sich so von stringenten Reaktionsbedingungen während der ersten Runden relaxierteren anzunähern.

2.7.3 Expand-PCR

Zur PCR-basierten Generierung von DNA-Fragmenten mit mehr als 2.000 Basenpaaren Länge wurde eine Expand-PCR benutzt (Cheng *et al.*, 1994), die sich durch ein sukzessives zeitliches Verlängern der Elongation ab dem zweiten Drittel der insgesamt 30 Gesamtzyklen auszeichnete. Nach 2 Minuten Denaturierung bei 94°C und 10 Zyklen mit 15 Sekunden bei 94°C, 30 Sekunden bei 50-55°C (je nach Primer) und 150 Sekunden bei 72°C wurde die Elongationszeit für die folgenden 20 Zyklen um jeweils 30 Sekunden je Zyklus verlängert, so dass das PCR-Programm mit einer abschließenden Extensionszeit von 12 Minuten und 30 Sekunden beendet wurde.

2.7.4 Reverse Transkriptase-Polymerasekettenreaktion (RT-PCR)

Das Transkribieren der präparierten mRNA in cDNA geschah durch reverse Transkription mit einem universellen Thymin-Oligomer (T16-mer) als Primer. Hierzu wurden 2 bis 4 µg RNA in einem Ansatzvolumen von 16 µl mit H_2O aufgenommen und für 10 min bei 70°C inkubiert, um etwaige Sekundärstrukturen der RNA aufzuschmelzen. Die reverse Transkription fand während einer Inkubation bei 37°C für 90 min statt, unter Hinzugabe von 100 pMol T16-mers als Primer, 35 nMol dNTPs, 2 µl (40 U) RNase-Inhibitor, 8 µl 5x 1st Strand Buffer (GIBCO), 4 µl Dithiothreitol (DTT; 0,1 M) und 1,5 µl MMLV-Reverse-Transkriptase (200 U/µl; GIBCO) in ein Gesamtvolumen von 40 µl. Danach folgte ein 10-minütiger Inaktivierungsschritt bei 94°C und ein Abkühlen auf 4°C. Für spätere PCR-Experimente wurde von dieser synthetisierten cDNA 1 bis 2 µl als Matrize eingesetzt.

2.8 Isolierung von Anschlussklonen

Die Suche nach potentiellen Anschlussklonen für die Darstellung des zu untersuchenden genomischen Bereiches in einem lückenlosen Kloncontig geschah mit einer Chromosomen-Walking-Strategie. Die Randbereiche des genomischen Integrats der verifizierten PAC-, BAC- oder Cosmid-Klone wurden über Sequenzierung aus der bekannten Sequenz des Vektors heraus „ansequenziert". Diese Sequenzinformation diente dann als Matrize zur Generierung von Primern für die Amplifikation der Sonden-DNA. Mit Hilfe dieser PCR-Sonden – zuvor radioaktiv markiert – wurde dann eine Koloniefilter-Hybridisierung auf gespotteten „high-density" Filtersets durchgeführt. Die Signal-positiven Klone der

Hybridisierung wurden über PCR, bzw. wieder über das Ansequenzieren der Randbereiche des neuen genomischen Integrats verifiziert und entsprechend der jeweiligen Überlappung der Sequenzflanken im so entstehenden Contig angeordnet.

2.9 Herstellung einer „shot-gun"-Klonbibliothek

Die Sequenzierung der ausgewählten PAC-, BAC- und Cosmid-Klone wurde mit Hilfe einer sog. „shot-gun"-Klonierungsstrategie verwirklicht. Um die hochmolekulare DNA der Ausgangsklone in eine für die Sequenzierung handhabbare Fragmentgröße zu bekommen, musste die DNA in Fragmentgrößen von 500 bis 3.000 bp überführt und in einen Sequenzierungsvektor (pUC18) subkloniert werden. Die rekombinanten Klone wurden im 96er Format angeordnet und über Nacht zum Anwachsen bei 37°C inkubiert und danach bis zur Weiterverarbeitung für die DNA-Präparation bei –70°C tiefgefroren.

2.9.1 „Plasmid-Safe"-Behandlung

Für die Herstellung einer Klonbibliothek dienten 20 bis 30 µg DNA des zu sequenzierenden PACs, BACs oder Cosmids als Ausgangsmaterial. Um bei dieser großen Menge die Kontamination an bakterieller genomischer DNA für die Subklonierung so gering wie möglich zu halten, wurde die isolierte Ausgangs-DNA einer „Plasmid-Safe"-Behandlung (EPICENTRE TECHNOLOGIES) nach Herstellerangaben unterzogen. Mit Hilfe einer ATP-abhängigen DNase wurde dabei selektiv nicht zirkuläre lineare Einzel- und Doppelstrang-DNA in ihre Einzelbausteine hydrolisiert. Zirkulär-geschlossene und „super coiled" DNA des Ausgangsklons blieb bei dieser Prozedur unberührt und fand im weiteren Vorgehen Verwendung.

2.9.2 Nebulisierung und Fragmentgrößen-Selektion

Das Fragmentieren der „Plasmid-Safe"-behandelten DNA geschah unter Verwendung eines Nebulizers, der die zu klonierende DNA in gewünschte Molekülgrößen brachte. Das Auftreten hoher Scherkräfte während des Einsatzes im Nebulizers durch Anlegen von 1 Bar Druckluft aus einer Stickstoffflasche führte dabei zur Fragmentierung der hochmolekularen DNA. Die „Plasmid-Safe"-behandelte DNA wurde dazu mit 1x TE-Puffer auf 2 ml aufgefüllt und für genau eine Minute unter obigen Bedingungen im Nebulizer zirkulieren lassen. Nach Abschluss wurde die nebulisierte DNA unter Hinzugabe von Dialysepuffer (1/10 Vol.) und Ethanol (2,5 Vol.) gefällt. Eine Fragmentgrößenbestimmung und Isolierung fand nach elektrophoretischer Auftrennung in einem 1%igem Gel mit niedrig schmelzender Agarose statt. Die sich als verwischte Bande darstellende fragmentierte DNA wurde aus zwei bzw. drei Fragmentgrößenbereichen aus dem Gel ausgeschnitten und wieder gewonnen, so dass für die Klonierung zwei Fraktionen aus den Fragmentgrößen von 700 bis 1.000 bp und von 1.000 bis 2.000 bp, bzw. drei Fraktionen aus den Größenbereichen 600 bis 1.000 bp, 1 kb bis 1,5 kb und 1,5 kb bis 2,5 kb zur Verfügung standen. Es wurde hierbei vermieden, die zu isolierende DNA mit

Ethidiumbromid in Kontakt zu bringen und dem UV-Licht auszusetzen, um eine artifizielle Modifikation der DNA auszuschließen. Als Referenz für die Bestimmung der auszuschneidenden Bereiche diente ein Aliquot gefärbte Proben DNA derselben Probe in einer extra Spur.

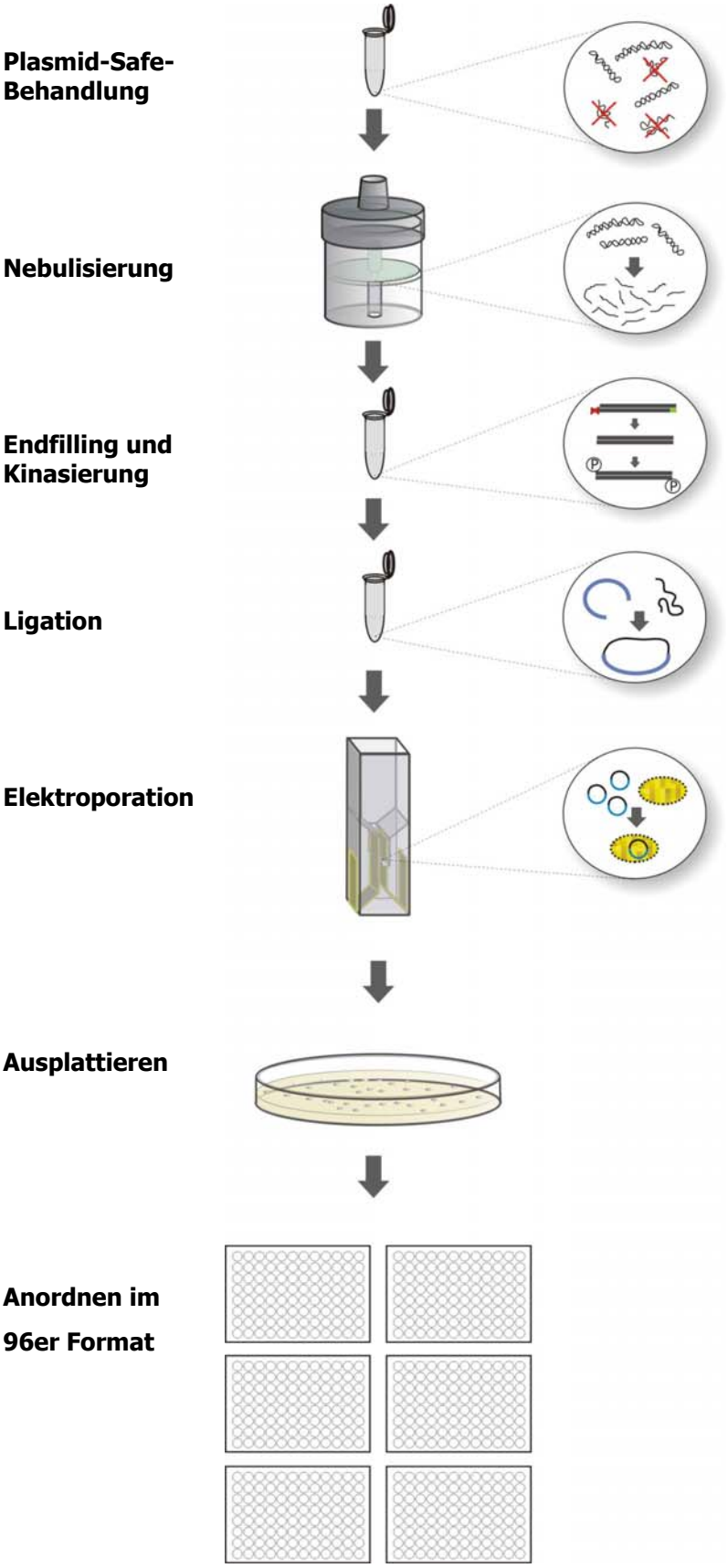


Abb. 3 Verlaufsdiagramm der Herstellung einer „Shotgun"-Klonbibliothek.

Die präparierte DNA, des zu sequenzierenden Klons wurde DNase-behandelt („Plasmid-Safe") und durch Nebulisieren auf subklonierbare Fragmentgrößen gebracht. Die nebulisierte DNA wurde daraufhin gelelektrophoretisch aufgetrennt und nach Fragmentgrößen gepoolt wiedergewonnen. In Vorbereitung auf die Ligation in den Sequenzierungsvektor (pUC18) fand ein „Endfilling" und eine Kinasierung der zu klonierenden DNA-Fragmente statt. Es folgte die Transformation in elektrokompetente *E. coli*-Zellen mit der in pUC18-verpackten DNA. Die elektroporierten Bakterienzellen wurden auf Agar-Medium zur Kultivierung ausplattiert. Nach Übernacht-Inkubation der Platten fand ein Anordnen der rekombinanten Subklone in Flüssigmedium in 96-Loch Mikrotiterplatten statt. Die Subklone wurden eine weitere Nacht bei 37°C inkubiert und danach bis zur Sequenzierung bei minus 80°C gelagert.

2.9.3 „Endfilling" und Kinasierung

Die größenfraktionierte DNA wurde im nächsten Schritt einem „Endfilling" unterzogen, d.h. einem Auffüllen überhängender einzelsträngiger DNA-Enden, die durch die mechanische Beanspruchung des Scherens während des Nebulisierens entstandenen waren. Anschließend folgte die Phosphorylierung der Fragmentenden, um die Effizienz der nachfolgenden Ligation in den „blunt-end"-geschnittenen Klonierungsvektor pUC18 zu erhöhen. Das Auffüllen der Enden vollzog sich in einem 100 µl Reaktionsansatz, bestehend aus der nebulisierten DNA, 0,5 mM dNTPs, 10 µl 10x T4-Polymerase-Puffer, 9 U T4-DNA-Polymerase (NEB) und 10 U Klenow-DNA-Polymerase (NEB), bei Raumtemperatur für 30 Minuten. Für das Endfilling wurde nicht nur die Polymerase-Aktivität des Klenow-Enzyms genutzt, sondern auch gleichzeitig die 3´-5´Exonuklease-Aktivität der T4-DNA-Polymerase (Sambrook *et al.*, 1989). Nach Aufreinigung der DNA mit Hilfe von „QIAquick-Spin-Columns" (QIAGEN) folgte die Kinase-Behandlung in einem 50 µl-Ansatz mit 10 mM rATP, 5 µl 10x Kinase-Puffer und 30 U T4-Polynukleotid-Kinase für 30 Minuten bei 37°C. Abschließend wurde die DNA nochmals über „QIAquick-Spin-Columns" (QIAGEN) aufgereinigt.

2.9.4 Ligation

Als Klonierungsvektor für die Herstellung der Bibliothek diente der Vektor pUC18, der Sma I-restringiert und mit Hilfe der bakteriellen alkalischen Phosphatase dephosphoryliert worden war, erworben von der Firma PHARMACIA. Der Ligationsansatz bestand pro Größenfraktion aus 250 bis 500 ng der vorbehandelten DNA, 25 ng Vektor-DNA, 1/10 Vol. 10x Ligationspuffer inkl. 1 mM ATP (NEB), 5% PEG und 400 U T4-DNA-Ligase (NEB). Die Ligationsreaktion fand zur Vermeidung von chimären Subklonen über Nacht gekühlt bei 4°C statt. Um die Salzkonzentration für die Elektroporation zu reduzieren, fand abschließend eine Säulenaufreinigung („QIAquick PCR Purification Kits", QIAGEN) statt.

2.9.5 Transformation elektrokompetenter E.coli-Zellen

Mit der in pUC18 ligierten Klon-DNA wurden elektrokompetente *E.coli*-Zellen des Stammes XL1-Blue der Firma STRATAGENE transformiert. 40 µl kompetenter Zellen wurden mit 1/10 des Ligationsansatzes in einer eisgekühlten Küvette (0,1 cm Spalt, BIORAD) bei 1,7 kV, 200 Ω und 25 µF (Zeitkonstante: 4,5 bis 5 ms) mit Hilfe einer BIORAD-Elektroporationseinheit gepulst. Nach sofortiger Aufnahme des Ansatzes in 960 µl SOC-Medium folgte eine Inkubation von einer Stunde bei 37°C und 250 Upm im Schüttelgerät. Je 50 µl der Suspension wurden auf eine X-Gal-LB-Agar-Ampicillin (100 mg/l) +Tetracyclin (30 mg/l)-Platte ausgestrichen und über Nacht bei 37°C inkubiert. Pro Fraktion wurden so insgesamt 5 bis 10 Platten hergestellt. Der restliche Transformationsansatz wurde bei 4°C für die Ausplattieren weiterer Platten aufbewahrt.

2.9.6 Selektion und Anordnung der rekombinanten Klone

Verteilt auf zwei, bzw. drei Größenfraktionen wurden pro PAC-/BAC-Bibliothek 2.500 bis 3.000 Klone bzw. 600 bis 800 Klone pro Cosmidbank gepickt und in 96er-Platten (gefüllt mit 200 µl LB-Amp + 7% Glyzerin) einzeln angeordnet. Mit Hilfe des lacZ-Gens im Klonierungsvektor pUC18 konnten für das Picken der Klone über die X-Gal-Umsetzung rekombinante von nicht rekombinanten Subklonen durch die andersartige Färbung unterschieden werden. Weiße rekombinante Klone wurden aufgenommen und in 96er Platten, die mit 250 µl LB-Medium und mit 7% Glycerin versetzt worden waren, angeordnet; bläuliche Klone ohne Integrat wurden verworfen. Nach Inkubation der angeordneten Klone für 20 Stunden bei 37°C wurden diese bis zur weiteren Verwendung bei –80°C eingefroren.

Zur Überprüfung der Klonierungseffizienz wurden pro Fraktion alle Klone einer 96er-Platte einer Standard-PCR mit Vektor-spezifischen M13-Primern zur genomischen Integrat-Amplifikation unterzogen und 1/5 der Ansätze gelelektrophoretisch aufgetrennt. Als PCR-Matrize diente 1 µl Kulturüberstand aus der Mikrotiterplatte.

2.10 Radioaktive Markierung von DNA-Sonden

Die Herstellung von radioaktiven DNA-Sonden erfolgte nach der von Feinberg & Vogelstein (1983) beschriebenen Methode des „Random-primed-oligolabelling". Die Markierung der Sonden fand mit [α^{32}P]dCTP (3.000 Ci/mmol) statt. Der Reaktionsansatz bestand bei der Markierung genomischer DNA aus 6 µl Oligolabelling-Mix mit Hexanukleotid-Primern, 60 µg BSA, 50-100 ng hitzedenaturierter Sonden-DNA, 30 µCi [α^{32}P]dCTP und 3 Units Klenow-DNA-Polymerase (NEB). Bei zu markierenden PCR-Produkten wurde ein Oligolabelling-Mix ohne Hexanukleotide, aber mit Zusatz von je 50 µM der fragmentspezifischen Primer verwendet. Die Markierungsreaktion, bei der die radioaktiven Nukleotide in die DNA-Sequenz inkorporiert wurden, dauerte bei 37°C zwei bis drei Stunden.

Die Aufreinigung von nicht-eingebauten Nukleotiden geschah über Sephadex G-50 Säulen („NICK™ Columns" oder „MicroSpin™ G-50 Columns", PHARMACIA) nach den Vorgaben des Herstellers. Ein Aliquot von einem µl der eluierten Sonde wurde zur Quantifizierung der Radioaktivität im Szintillationszähler (WALLACE 1410) gemessen.

2.11 Hybridisierungstechniken mit radioaktiven Sonden

2.11.1 DNA-Transfer auf eine Filtermembran (Southern-blotting)

Die Übertragung von DNA-Fragmenten aus Agarosegelen auf eine Nylonmembran (Hybond™-N^+, AMERSHAM) fand nach der Methode von Southern (1975) und in Anlehnung an das Protokoll des Membranherstellers AMERSHAM statt. Zur Steigerung der Transfereffizienz wurde die DNA im Gel mit

0,25 M HCl für eine halbe Stunde depuriniert. Beim Blotten der DNA mit 20x SSC als Laufpuffer folgte eine Denaturierung mit 1,5 M NaCl und 0,5 M NaOH für 30 Minuten und eine ebenso lange Neutralisierung (mit 1,5 M NaCl; 0,5 M Tris/HCl, pH 7,2; 1 mM EDTA). Der DNA-Transfer erfolgte in der Regel über Nacht. Die Fixierung der DNA an die mit 2x SSC gewaschenen und luftgetrockneten Filter geschah durch UV-„Cross linking" mittels eines Crosslinking-Ofens (STRATAGENE).

2.11.2 DNA-DNA-Hybridisierung nach Southern

Die Hybridisierung der geblotteten Filter erfolgte über Nacht bei 65°C mit Hybridisierungspuffer nach Church *et al.* (1983). Nach 2 bis 3 Stunden Prähybridisierung wurde die denaturierte, radioaktive Sonde (pro ml Hybridisierungslösung ca. eine Mio. „counts") zu den Filtern hinzugegeben. Die Hybridisierungsreaktion bei kleinen Filtern fand in speziellen Hybridisierungsröhren (TECHNE) im Wärmeofen (Hybridiser HB-2, TECHNE) statt. Nach der Inkubation wurden die Filter mit Lösungen abnehmender $NaHPO_4$-Konzentration (200 mM → 100 mM → 50 mM) bei 65°C bzw. bei Raumtemperatur für jeweils 15 bis 30 Minuten gewaschen. Mit Hilfe eines Müller-Geiger-Zählers wurde die Abnahme der radioaktiven Strahlung überwacht und bei einer Aktivität unter 50 counts/sec wurden die Filter zusammen mit einem Röntgenfilm (Hyperfilm-MP RPN8, AMERSHAM) in Kassetten mit Verstärkerfolie (DU PONT) bei –80°C exponiert. Je nach Strahlungsintensität der Markierungen wurde nach einer Expositionsdauer von 3 Stunden bis mehreren Tagen der Röntgenfilm entwickelt und das Ergebnis ausgewertet.

2.11.3 Koloniefilter-Hybridisierung

Das „Screening" nach Anschlussklonen fand mit der Methode der Koloniefilter-Hybridisierung nach Grundstein & Wallis (1979) statt. Grundlage für die Hybridisierung waren entweder selbsthergestellte Koloniefilter oder automatisch gespottete „high-density" Filterset verschiedener Bibliotheken bezogen vom Ressourcenzentrum des Deutschen Humangenomprojekts in Berlin. Die Hybridisierung geschah mit einem Natriumphosphatpuffer nach Empfehlung des Ressourcenzentrums auf der Grundlage nach Church & Gilbert (1984). Aufgrund des großen Filterformats wurden die Filter für die Hybridisierungsreaktion mit der markierten Sonde in Kunststofffolien geschweißt und geschützt durch eine Plastikkassette in einem Schüttelwasserbad inkubiert.

2.12 Herstellung einer Subtraktionsgenbank

Um die Zahl der zu sequenzierenden Subklone für Überlappungsbereiche zweier benachbarter PAC-Klone zu reduzieren, wurde versucht aus einer der beiden „Shot-gun"-Bibliotheken jene Subklone zu sortiert, die aus dem gemeinsamen genomischen Überlappungsbereich stammen. Die in Mikrotiterplatten angeordneten Subklone des Klons *Eins* wurden auf Koloniefilter übertragen und mit Sonden-DNA hybridisiert, die nicht aus dem Überlappungsbereich mit Klon *Zwei* stammte. Alle durch

die Hybridisierung markierten und somit nicht-redundanten Subklone wurden in einer zweiten Bibliothek der Subtraktionsbank neu angeordnet.

In der vorliegenden Arbeit wurde auf diese Weise die Zahl der zu sequenzierenden Subklone des PAC368C2 verringert, der auf beiden Seiten von bereits sequenzierten Anschlussklonen flankiert war. Die DNA dieses PAC-Klons wurde mit Eco RI restringiert und gelelektrophoretisch aufgetrennt. Das Gel wurde geplottet und die DNA auf Nylonmembran übertragen. Die anschließende Hybridisierung dieses Filters mit radioaktiv markierter DNA aus den beiden Nachbarklonen identifizierte alle Restriktionsbanden, die aus den überlappenden Randbereichen resultierten. Jene Banden des PAC368C2, die durch diese Hybridisierung nicht markiert werden konnten, wurden ausgeschnitten. Die darin enthaltene DNA wurde isoliert, ebenfalls radioaktiv mit [α^{32}P]dCTP markiert und mit den Koloniefiltern bestehend aus den angeordneten Subklonen des PAC368C2 hybridisiert. Um eine Kreuzhybridisierung mit repetitiven Sequenzabschnitten zu verhindern, wurden diese mit Cot I-DNA in der Hybridisierungslösung abgesättigt. Alle markierten Subklone wurden in einer zweiten Subtraktions-Genbibliothek angeordnet und der nachfolgenden Sequenzierung zugeführt.

2.13 Fluoreszenz-in-situ-Hybridisierung

Zur eindeutigen chromosomalen Lokalisierung wurden die ausgewählten Klone mit Hilfe der Fluoreszenz-in-situ-Hybridisierung (FISH) auf gespreiteten und fixierten Metaphase-Chromosomen, gewonnen aus synchronisierten Lymphozytenkulturen (Viegas-Péquignot, 1987), dargestellt.

2.13.1 Herstellung von Metaphasechromosomen

Humane Chromosomenpräparate wurden aus Bromdesoxyuridin (BrdU) synchronisierten Lymphozytenkulturen hergestellt (Lemieux *et al.*, 1992). Fünf Tropfen heparinisiertes Vollblut wurden mit 5 ml Chromosomenmedium IA (GIBCO) für 48 Stunden bei 37°C inkubiert und nach Zugabe von 200 µg BrdU pro ml Chromosomenmedium für weitere 16 bis 17 Stunden bei gleicher Temperatur inkubiert. Nach zweimaligem Waschen mit vorgewärmten Medium IA wurde die Kultur mit frischem 5 ml Medium IA und mit 25 µg Thymidin/ml Chromosomenmedium versetzt. Es folgte eine Inkubation der Kultur bei 37°C, der nach 7 Stunden 50 µl Colcemid (10 µg/ml) zugeführt wurden. Nach nochmals einer Stunde Inkubation wurde die Probe abzentrifugiert, der Überstand verworfen und die Zellen in 5 ml hypotonischer Lösung (0,075 mol/l KCl) resuspendiert. Es schloss sich eine erneute Inkubation der Zellen für 10 Minuten an, die durch einen Zentrifugationsschritt beendet wurde. Der Überstand wurde vorsichtig abgesaugt und die Zellen mit 5 ml kaltem Methanol/Eisessig (3:1) versetzt. Nach zweimaliger Wiederholung dieses Schritts wurden die Zellen nach der Zentrifugation mit wenig Methanol/Eisessig zu einer genügend dichten Zellsuspension aufgenommen. Diese Suspension wurde auf eisgekühlte und entfettete Objektträger aufgetropft und sofort zum Fixieren kurz abgeflammt.

Die Herstellung von Maus-Chromosomenpräparaten geschah aus einer kultivierten Fibroblasten-Zelllinie. Nach Anwachsen der Fibroblasten am Boden der Kulturschalen im Brutschrank bei 37°C und ca. 5% CO_2 wurde das Erstmedium durch 5 ml RPMI-Medium mit 10% fetalem Kälberserum ersetzt. Bei Bildung eines dichten Zellrasens wurde das Medium entfernt, die Zellen mit Trypsin-EDTA abgelöst und bei 1.000 Upm bei Raumtemperatur für 10 Minuten abzentrifugiert. Nach zweimaligem Waschen mit RPMI-Medium folgte die Behandlung der Zellen mit hypotonischer KCL-Lösung und die Fixierung mit Methanol/Eisessig wie oben beschrieben.

2.13.2 Markierung der Sonden-DNA

Für die Markierung mit fluoreszierenden Markerfarbstoffen (Biotin bzw. Digoxygenin) wurde die Sonden-DNA durch Restriktion mit Pst I auf Fragmentgrößen zwischen 300 bp und 700 bp geschnitten. Die sich nach einer Phenolextraktion des Restriktionsansatzes anschließende Nick-Translation wurde unter Verwendung des „Nick-Translations-Kit" der Firma GIBCO BRL durchgeführt. 500 ng bis 800 ng Sonden-DNA wurden zusammen mit 1 µl Biotin-16-dUTP, bzw. Digoxygenin-11-dUTP markierten Nukleotiden, dNTPs und Enzymgemisch für eine Stunde bei 15°C inkubiert. Der Markierungsansatz wurde zusammen mit humaner, bzw. muriner Cot I-DNA, zur Absättigung repetitiver Bereiche, und mit „Salmon sperm"-DNA gefällt.

2.13.3 Hybridisierung

Ein Lösen des Pellets geschah mit Formamid und einem Master-Mix (4x SSC, 20% Dextransulfat) zu gleichen Anteilen. Die Sonde wurde denaturiert und anschließend für 15 bis 30 Minuten bei 37°C inkubiert. Parallel dazu wurden die fixierten Chromosomen auf den Objektträger für drei Minuten bei 75°C mit 70% Formamid und 2x SSC denaturiert und in einer ansteigenden Ethanolreihe (70% → 80% → 95%) dehydriert. Auf die derart vorbehandelten Objektträger wurde je ein Aliquot (40–50 ng) der beiden Sonden aufgetragen und unter einem Abdeckglas verschlossen. Die Hybridisierung fand über Nacht bei 37°C in einer feuchten Kammer statt.

2.13.4 Detektion der markierten Sonden und Chromosomengegenfärbung

Die Detektion der Sonden erfolgte nach 16 Stunden Inkubation am nächsten Morgen nach der Methode von Lichter & Mitarbeiter (1988). Die Hybridisierungen wurden bei 42°C dreimal fünf Minuten in 50% Formamid und 1x SSC, und dreimal fünf Minuten in 0,3x SSC gewaschen und zusätzlich für 15 Minuten bei Raumtemperatur in 4x SSC, 0,1% Tween 20 und BSA abgesättigt. Der Nachweis der Digoxigenin-markierten Sonde geschah mit den monoklonalen Antikörpern: Maus-anti-DIG IgG, TRITC-markierten Hase-anti-Maus-IgG und TRITC-markierten Ziege-anti-Hase-IgG (SIGMA). Die Biotin-markierte Sonde wurde mit den Antikörpern Fluorescein-Avidin, Biotin-anti-Avidin und mit Fluorescein-Avidin (SIGMA) detektiert. Jede Inkubation dauerte 30 Minuten bei 37°C und wurde durch drei Waschschritte mit 4x SSC und 0,1% Tween 20 für jeweils fünf Minuten abgeschlossen. Die

Gegenfärbung der Chromosomen erfolgte lichtgeschützt mit DAPI (BOEHRINGER) in 4x SSC, 0,1% Tween 20 für fünf Minuten bei Raumtemperatur. Um ein Austrocknen der Präparate zu verhindern, wurden diese in „Vectashield Mounting Medium" (VECTOR LABORATORIES) eingebettet. Die Auswertung der Präparate erfolgte mit einem Fluoreszenz-Mikroskop (LEICA). Die Optik bestand aus einem 10x Okular und einem 100x/1,3 „Fluotar" Öl-Immerisions-Objektiv. Die Bildaufzeichnung war CCD-Kamera gestützt und wurde über das Softwareprogramm CYTOVISION 2.21 (APPLIED IMAGING) gesteuert.

2.14 DNA-Sequenzierung

Die automatisierte DNA-Sequenzierung basierte auf dem Prinzip der Kettenabbruchsynthese-Methode nach Sanger & Mitarbeiter (1977), und der Modifikation von Lee & Mitarbeiter (1992), die eine basenspezifische DNA-Markierung am 3´-Ende mit vier verschiedenen fluoreszenzmarkierten Didesoxynukleotiden ermöglicht. Diese Markierung geschah mit Hilfe von fluoreszierenden Chromophoren, die Licht in einem gemeinsamen Frequenzbereich absorbieren, aber in unterschiedlichen Spektralbereichen farbig wieder emittieren.

2.14.1 Markierung der zu sequenzierende DNA

In einem „Cycle-Sequencing"-Prozess wurde die zu analysierende DNA mit Hilfe von proben-spezifischen Primern in einer linearen Amplifikation unter Einsatz von (a) einer hitzestabilen Taq-Polymerase („AmpliTaq™DNA Polymerase", PERKIN ELMER) ohne 3'-5' und 5'-3' Exonuklease-Aktivität und von (b) fluoreszenzmarkierten Terminatoren („Dye Terminators", PERKIN ELMER) farbig markiert. In jedem Amplifikationszyklus endete die DNA-Synthese nach Einbau eines Matrizenstrang-komplementären und fluoreszenzmarkierten Didesoxynukleotid. Da dieser Syntheseabbruch statistisch an jeder Base des zu sequenzierenden DNA-Moleküls stattfinden kann, ließ sich während des 24-30 Zyklen umfassenden „Cycle-Sequencing" jedes Nukleotid farbig markieren. Die Analyse der markierten DNA-Fragmente geschah nach gelelektrophoretischer Auftrennung in einem Polyacrylamidgel (siehe Kap. 2.4.3) und mit Hilfe eines laserbasierten Detektionssystems in einem Sequenzierautomaten. In der vorliegenden Arbeit konnte auf drei Geräte der Firma ABI APPLIED BIOSYSTEMS (ABI PRISM 377-96, ABI PRISM 377XL, ABI 373A) mit Ladekapazitäten zwischen 36 und 96 Einzelproben zurückgegriffen werden. Die Markierungsreaktion fand unter Verwendung der kommerziellen Kits (1) „ABI PRISM™ Dye Terminator Cycle sequencing Ready Reaction Kit", PE APPLIED BIOSYSTEMS, (2) „ABI PRISM™ Big-Dye™ Terminator Cycle Sequencing Ready Reaction Kit" PE APPLIED BIOSYSTEMS, (3) „Thermo Sequenase™ Dye Terminator Cycle Sequencing Pre-Mix Kit", AMERSHAM LIFE SCIENCE während des „Cycle-Sequencings" in einem PCR-Gerät statt. Die Zusammensetzung des „Cycle-Sequencing"-Ansatzes und Modifikationen des PCR-Programms richteten sich nach Art der zu sequenzieren DNA. Die Kits (1) und (3) wurden dabei gleichwertig behandelt. Bei Verwendung der Big-Dye™ Terminatoren (2) konnte auch mit der Hälfte der Menge des angegebenen Enzym-Reaktionsgemisches

(Premix) eine ausreichend starke Markierung erzielt werden. Eine Auflistung der verwendeten Mengen ist in nachfolgender Tabelle 6 zusammengestellt:

Tab. 6 Konzentrationsangaben der verschiedenen „Cycle-Sequencing"-Ansätze. Eingesetzte DNA-, Premix- und Primer-Mengen, je nach Art der zu sequenzierenden DNA und die verwendete Zyklenzahl. Die Premix-Menge konnte bei Verwendung des Big-Dye™-Premix auch bis auf die Hälfte der angegebenen Menge (PCR- und Plasmid-DNA) reduziert werden, ohne die Lesbarkeit der sequenzierten DNA nennenswert zu verschlechtern.

Template	DNA-Menge	Premix-Menge	Primer	Ansatz-Vol.	Zyklenzahl
PCR-Produkte	20 – 100 ng	4 – 6 µl	10 pMol	15 µl	24x
subklonierte Plasmid-DNA	100 – 400 ng	4 µl	5 pMol	10 µl	26x
PAC-, BAC-DNA	800 – 1000 ng	8 – 16 µl	10 pMol	20 – 40 µl	35x
Cosmid-DNA	500 – 1000 ng	8 µl	10 pMol	20 µl	30x

Für das „Cycle-Sequencing" wurden die verwendeten PCR-Geräte (PTC 200™, MJ RESEARCH,INC. (USA) und PE 9700 Cycler, PERKIN-ELMER) mit folgenden Zyklen programmiert: Das Standard-„Cycle-Sequencing"-Programm für M13-Primer umfasste insgesamt 26 Zyklen mit 15 Sekunden Denaturierung bei 96°C, 10 Sekunden „Annealing" bei 50°C und vier Minuten Elongation bei 60°C; bis zur weiteren Verarbeitung wurden die Proben abschließend auf 4°C gekühlt.

Bei schwierigen Matrizen, wie etwa GC-reiche Sequenzen, wurden eine Vorinkubation über fünf Minuten bei 98°C vor Zugabe des Premix durchgeführt, um ein vollständiges Denaturieren der Probe zu gewährleisten. Matrizen, bei denen für ein schlechtes Sequenzierergebnis Sekundärstrukturen verantwortlich gemacht wurden, konnten durch Erhöhung der Premix-Konzentration und durch Anhebung der Denaturierungstemperatur auf 98°C für 20 Sekunden und der „Annealing"-Temperatur auf 55°C für 15 Sekunden bei 26 Zyklen besser sequenziert werden. Hochmolekulare DNA-Templates, wie bei der PAC- bzw. BAC-Klon Direktsequenzierung, wurden generell mit erhöhten Premix-Konzentrationen, einem „Hotstart", d.h. Zugabe des Premix nach vorherigem fünfminütigem Denaturieren bei 95°C, und verlängerter Denaturierung auf 30 Sekunden, bzw. angehobener „Annealing"-Temperatur auf 55°C für 20 Sekunden sequenziert. Wurden Randsequenzen von PAC- oder BAC-Klonen, d.h. das Sequenzieren aus dem flankierenden vektoriellen Anteil des Klons in die genomische DNA des Integrats hinein, sequenziert, führte eine vor dem „Cycle-Sequencing" durchgeführte Eco RV-Restriktion mit anschließender DNA-Aufreinigung („Geneclean-Kit" BIO101) zu einem besser lesbaren und somit besser auswertbaren Fluoreszenzsignal. Um die Signalintensität bei geringen DNA-Ausgangskonzentrationen zu erhöhen, wurde zusätzlich mit einer leicht erhöhten Zyklenzahl gearbeitet. Eine Anhebung auf insgesamt bis zu 30 Zyklen erbrachte in den meisten Fällen signifikant bessere Leseergebnisse. Die Reaktionen fanden entweder einzeln in 0,2 ml PCR-

Reaktionsgefäßen (ADVANCED BIOTECHNOLOGIES) oder in 96er PCR-Platten (ADVANCED BIOTECHNOLOGIES) statt, um die Mikrotiter-Platten-Anordnung der gepickten Subklone zu wahren.

Nicht eingebaute Nukleotide wurden durch eine Ethanolfällung unter Hinzugabe von 1/10 Vol. 3 M NaAc (pH 4,6) abgetrennt oder durch Gelfiltration („QIAquick Spin Columns", QIAGEN) aufgereinigt. Insbesondere beim Sequenzieren der in 96-Loch-Platten angeordneten Subklone während der „Shotgun"-Sequenzierungsphase wurde die Aufreinigung über Sephadex G-50 (PHARMACIA) gewählt, da unter Verwendung des „Multiscreen Systems" (MILLIPORE) die Aufreinigung in einem einzigen Zentrifugationsschritt durchgeführt werden konnte und die Proben abschließend nur noch in der Vakuumzentrifuge eingedampft und getrocknet werden mussten.

2.14.2 Sequenziergel-Herstellung

Die Sequenziergel-Herstellung und die gelelektrophoretische Auftrennung im Sequenziergerät wurde nach Parametern des Herstellers (ABI APPLIED BIOSYSTEMS) durchgeführt. Das Standard-Polyacryamid-(PAA) Gel bestand aus 8,4 ml 30% PAA (29:1)(ROTH), 21 g Harnstoff (ROTH), 6 ml TBE-Puffer, 20 ml autoklaviertes, deionisiertes Wasser; zur Polymerisation wurden nach vorherigem Sterilfiltrieren und Entgasen 20 µl TEMED und 300 µl 10% APS hinzupipettiert. Nach mindestens 90-minütigem Aushärten wurde das Gel in die entsprechende Aufhängevorrichtung des Sequenziergerätes platziert. Der Probenauftrag fand einzeln oder mittels einer 8-Kanal-Pipette (KLOEHN) statt. Je nach Anzahl der Probenplätze wurden 0,5 µl (96er Kamm) bis 3 µl (36er Kamm) der aufgereinigten und fluoreszenzmarkierten Probe pro Sequenzlauf aufgetragen (siehe auch Kap. 2.4.3).

2.14.3 Detektion und Generierung der DNA-Sequenz

Nach gelelektrophoretischen Auftrennung der Proben-DNA fand am unteren Ende des Sequenziergels eine lasergestützte Detektion der emittierten, Nukleotid-spezifischen Fluoreszenzsignale und die rechnergestützte Aufzeichnung der Primärdaten statt. Nach Abschluss der Gelelektrophorese, d.h. nach Durchlaufen aller DNA-Fragmente, wurde unter Verwendung der *DATA COLLECTING SOFTWARE 2.6* (PE BIOSYSTEMS) die Zuordnung der Fluoreszenzsignale zu den aufgetragenen Proben („tracking") durchgeführt. Das Übersetzen des vierfarbigen Probensignals in die entsprechenden Nukleotidbasen Adenin, Cytosin, Guanin und Thymin („basecalling"), d.h. das Auslesen der Primärdaten wurde mit der *SEQUENCING ANALYSIS SOFTWARE 3.4* (PE BIOSYSTEMS) vorgenommen. Bei Sequenzläufen mit 96 Proben wurde zusätzlich ein fünfter Farbstoff („lane-guide") in jede achte Spur eingesetzt, um bei der Gelauswertung eine eindeutige Diskriminierung der vertikalen Spuren („lane-tracking") vornehmen zu können. War diese automatisch generierte Zuordnung nicht eindeutig, musste im Anschluss noch manuell editiert werden.

2.15 Sequenzierungsstrategie und Darstellung der genomischen Sequenz

Die Generierung der genomischen Consensussequenz für die sequenzierten PAC-, BAC- und Cosmid-Klone erfolgte strategisch wie in Abb. 4 skizziert in drei Schritten.

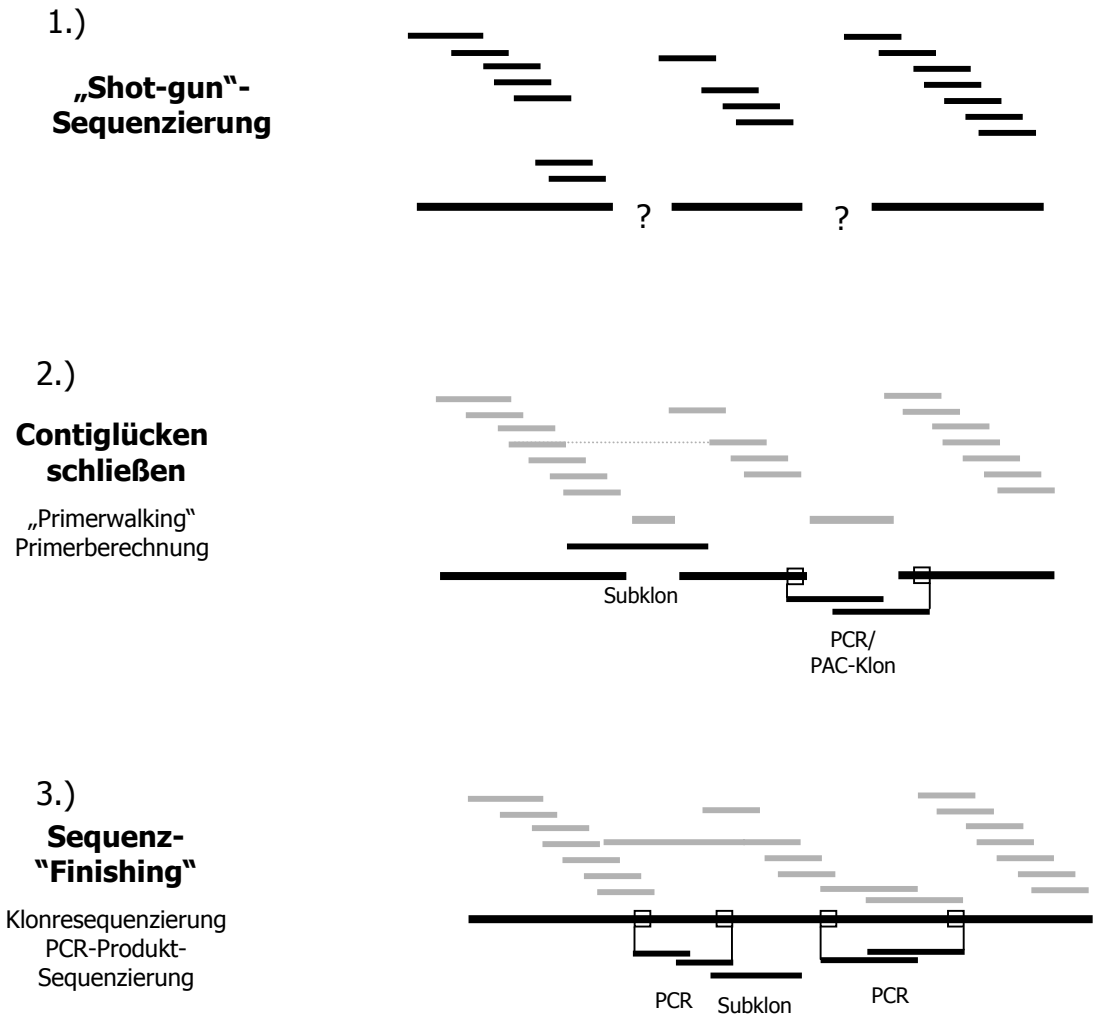


Abb. 4 **1.)** In der „Shot-gun"-Sequenzierungsphase werden die in 96er Platten angeordneten Subklone mit M13-Primern sequenziert. Die Sequenzen werden am Computer aufgrund ihrer überlappenden Bereiche zu „Contigs" angeordnet und zusammengefasst. **2.)** Fehlende Sequenzbereiche zwischen den Contigs werden durch das Sequenzieren von lückenüberspannenden Subklonen, deren Randsequenzen in zwei unterschiedlichen Contigs eingebaut wurden, ermittelt oder durch das direkte Sequenzieren der PAC-Klon-DNA mittels Contig-flankierender Primer erzeugt. Auch das Sequenzieren von lückenüberspannenden PCR-Produkten, die mit Hilfe von Contig-flankierenden Primer synthetisiert wurden, trugen zur Ermittlung der Consensussequenz bei. **3.)** In der „Finishing"-Phase wurden Bereiche, deren Basenpaarabfolge nicht eindeutig bestimmt werden konnte, bzw. die nur durch eine einzige Sequenz repräsentiert waren, resequenziert. Hierzu wurde die DNA der entsprechenden Subklone oder die über PCR amplifizierte DNA der zu verifizierenden Bereiche verwendet.

2.15.1 „Shot-gun"-Sequenzierung

In der ersten Phase, der sog. „Shot-gun"-Sequenzierungsphase wurde versucht, möglichst große, zusammenhängende Sequenzabschnitte aus überlappenden Einzelsequenzen zu generieren. Dazu wurden alle im 96er Format angeordneten Subklone von einer Seite mit Hilfe der Vektorsequenz-spezifischen Primer *M13-forward* sequenziert. Die anfänglich durchgeführte beidseitige M13-Sequenzierung wurde im Laufe der Arbeit eingeschränkt und nur noch bei Subklonen mit großem Integrat (ab 1,5 kb) durchgeführt. In dieser Phase wurden etwa 85% aller genomischen Sequenzdaten produziert. Die Verarbeitung dieser großen Masse an Sequenzinformationen geschah durch das Anordnen der Einzelsequenzen, das sog. „Assembling", mit Hilfe der Programme *SEQUENCHER* (Vers. 3.0 bis 4.1; GENCODES), bzw. dem Software-Packet *PHREPPHRAP* (Vers. 0.96731). Da die zweite Software einen höheren Automatisierungsgrad zuließ und auf einer leistungsstarken Workstation als Plattform (UltraSparc 1, 200 MHz) betrieben werden konnte, wurde dieses Programm im Verlauf der Arbeit ausschließlich verwendet.

2.15.2 Contiglücken schließen

Die zweite Phase beinhaltete das Verbinden der Contigs untereinander durch Schließen der noch verbliebenen Sequenzlücken. Dieser Schritt bestand zum einen durch selektives Sequenzieren des Gegenstrangs bestimmter nur einseitig sequenzierter Subklone, deren Sequenzinformation die Lücke ausfüllte, bzw. durch zusätzliches „Primerwalking" auf dem genomischen Sequenzanteil großer Subklone (1,5 bis 2 kb). Zum anderen wurden Contiglücken, die nicht durch Subklone repräsentiert waren, durch PCR-Amplifikation mittels lückenflankierender Primer und PAC/BAC-DNA, und durch Sequenzierung der PCR-Produkte geschlossen. Am Ende dieses zweiten Schritts stand die lückenlose Consensussequenz des genomischen Integrats des Ausgangsklon zur Verfügung.

2.15.3 Sequenz-„Finishing" und Sequenzannotierung

Der dritte Schritt umfasste das Sequenz-„Finishing" mit dem Ziel, dass jeder Sequenzabschnitt mindestens dreimal einzelsträngig oder einmal doppelsträngig sequenziert vorlag. Diese Vorgabe wurde durch Resequenzierung einzelner Subklone, bzw. durch PCR-Fragment-Sequenzierung mittels spezieller Primer erreicht, die den zu verifizierenden Bereich umschlossen.

Die so erstellte und verifizierte Consensussequenz fand daraufhin über das WEBIN-Serviceportal des Servers des „European Bioinformatics Institute" (EBI) in Hinxton unter Zuweisung einer „Accession number" (Acc.-Nr.) ihren Eintrag in die Nukleotidsequenz-Datenbank („Nucleotide Sequence Database") (EMBL: http://www.ebi.ac.uk/embl/Submission/index.html).

2.16 Übersicht der verwendeten Software zur Sequenzgenerierung und –auswertung

In allen drei Sequenzierungsphase wurde mit den Assemblierungsprogrammen *SEQUENCHER* oder *PHREDPHRAP* gearbeitet. Mit dem Software-Programmpaket *PHREDPHRAP* wurde zudem auch eine direkte Berechnung von Oligonukleotid-Primern auf der Basis der vorhandenen Consensussequenz für das Sequenz-„Finishing" vorgenommen, so dass das sehr zeitintensive Kopieren der entsprechenden Consensussequenzabschnitte in internetbasierte Anwendung entfiel. Insgesamt wurde mit sechs verschiedenen Software-Programmen gearbeitet, die in nachfolgender Grafik (Abb. 5) entsprechend der Chronologie im Prozess der Consensussequenz-Generierung angeordnet sind. Die Vor- und Nachteile der einzelnen Programme werden im Kap. 4.1.3 und 4.1.4 diskutiert.

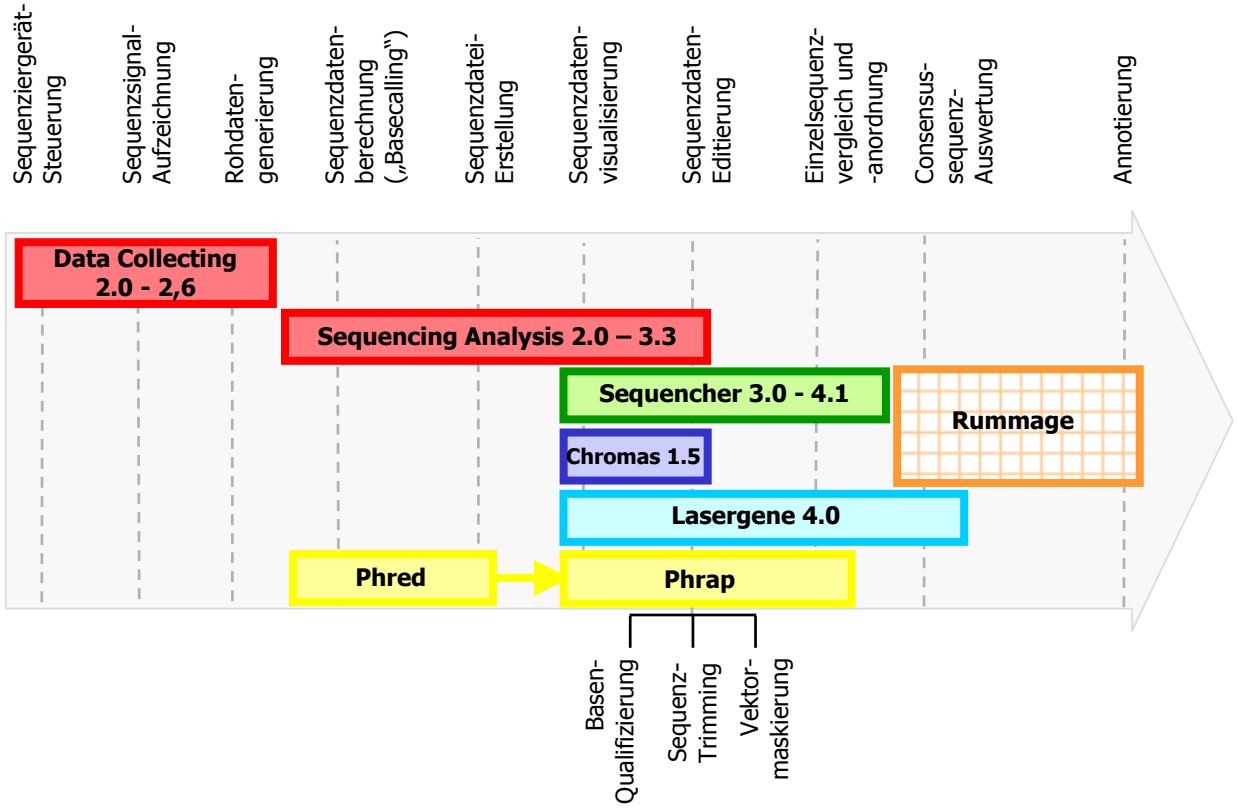


Abb. 5 Übersicht aller in den Prozess der DNA-Entschlüsselung involvierten Software-Programme. Die rot umrandeten Programme wurden vom Sequenzierautomaten-Hersteller (ABI Applied Bioscience) angeboten. Der sehr arbeitsintensive Bereich des Sequenzdaten-Editierens wird von einer Palette von Programmen abgedeckt (Sequencher, Chromas, Lasergene, Phrap). Dieser Arbeitschritt beinhaltete im Detail die drei Handlungsschritte: Basen-Qualifizierung („quality-valuation"), Sequenz-„Trimming" und Vektor-Maskierung, die von den verschiedene Programmen teils unterschiedlich ausgeführt wurden. Das orange umrandete Programm Rummage stellte eine Plattform dar, welche ein ganze Reihe an unterschiedlichen Software-Tools für die Sequenzanalyse miteinander verknüpfte und in Kap. 2.14.1 erläutert wird. Die Software-Hersteller sind unter 2.15.10 aufgeführt.

2.17 Computergestützte DNA-Auswertung

Im Gegensatz zur Generierung der Consensussequenz, die ausschließlich mit Programmen auf lokalen Rechnern vorgenommen wurden, fand die Analyse der Sequenzinformationen der genomischen DNA überwiegend mit Internet-basierenden Programmen auf den Servern der jeweiligen Anbieter statt. Eine Übersicht der eingesetzten Programme ist in Tab. 7 unter Angabe der jeweiligen Server-Internetadresse zusammengetragen.

Tab. 7 **Zusammenstellung aller webbasierenden Programme,** die für die Analyse der genomischen Consensussequenz eingesetzt wurden. Die eingetragene URL gibt die Adresse des Servers mit der Programmapplikation an. Eine detaillierte Beschreibung der einzelnen Programme findet sich im Anschluss unter Kap. 2.17.2 bis 2.17.6.

Analyse	Programm	Internet-Adresse (URL)
CpG-Inseln + GC-Profil	LCP CpG	http://gen100.imb-jena.de/rummage/ http://bioweb.pasteur.fr/seqanal/interfaces/cpgplot.html
Repetitive Elemente	RepeatMasker SST (sensitive search tool) Censor XNUN	http://ftp.genome.washington.edu/cgi-bin/RepeatMasker http://gen100.imb-jena.de/rummage/ http://www.girinst.org/Censor_Server-References.html http://gen100.imb-jena.de/rummage/
Promotor-Analyse	ProScan II GENSCAN	http://menu.hgmp.mrc.ac.uk/menu-bin/run?option=proscan http://genes.mit.edu/GENSCAN.html
Exon-Vorhersage	GRAIL FEXHB MZEF Xpound GENSCAN	http://compbio.ornl.gov/Grail-1.3/intro.html http://searchlauncher.bcm.tmc.edu:9331/gene-finder/gf.html http://argon.cshl.org/genefinder/ http://bioweb.pasteur.fr/seqanal/interfaces/xpound-simple.html http://genes.mit.edu/GENSCAN.html
Genstruktur	GENSCAN DPS ExonSampler	http://genes.mit.edu/GENSCAN.html http://genome.cs.mtu.edu/sas.html http://gen100.imb-jena.de/rummage/docu/paper/paper/node2.html
tRNA-Codierung	TRNA	http://www.genetics.wustl.edu/eddy/tRNAscan-SE/
Homologie- und Motivsuche	BLAST (*homo, mus*) (*fugu*) (*drosophila*) (*C.elegans*) FASTA	http://www.ncbi.nlm.nih.gov:80/BLAST/ http://fugu.hgmp.mrc.ac.uk/Analysis/ http://fly.ebi.ac.uk:7081/ + http://www.fruitfly.org/blast/ http://www.wormbase.org/db/searches/blast http://www.ebi.ac.uk/fasta33/
MAR-Analyse	MarFinder	http://www.futuresoft.org/MAR-Wiz/
PIP-Analyse	PipMaker	http://bio.cse.psu.edu/pipmaker/
Primer-Berechnung	Primer3	http://www-genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi
Protein-Molekulargewichts-berechnung	ExPASy: PeptideMass	http://www.expasy.org/tools/peptide-mass.html

2.17.1 Das RUMMAGE-Analyseprogramm

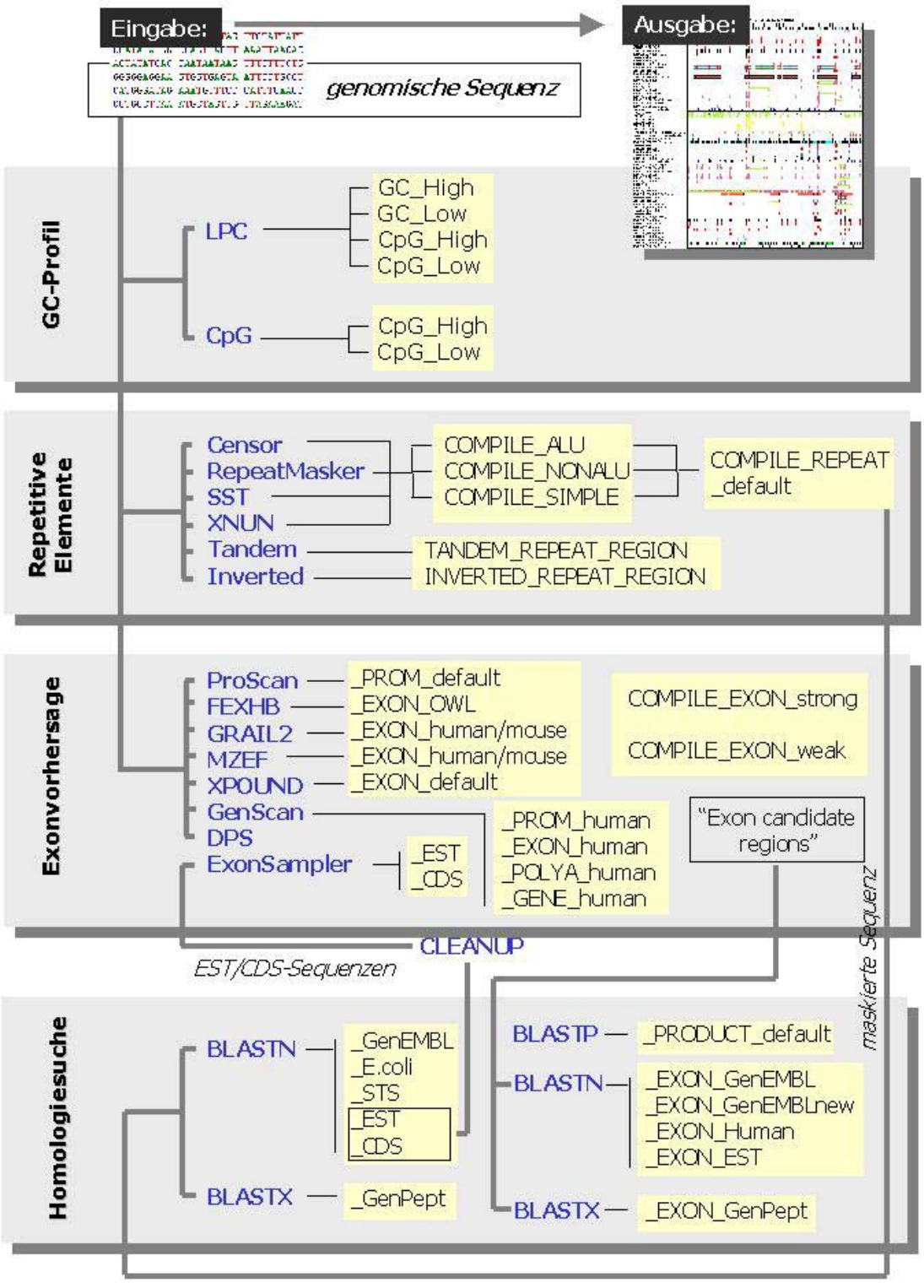


Abb. 6 **Verlaufsdiagramm des Rummage-Annotierungsprogramms**. Die Grafik zeigt die Komplexität der Rummage-Analyse mit allen implementierten Algorithmen und Software-Programmen. Die Analyse-Ergebnisse der in blauer Schrift aufgeführten Unterprogramme werden in die gelb-unterlegten Datenbanken hinterlegt und können von dort entweder direkt oder über die generierte interaktive graphische

Oberfläche abgerufen werden. Die genomische Consensussequenz wurde dabei vor der Analyse in kleinere Abschnitte von ca. 150 kb zerlegt, für die dann parallel der GC-Gehalt, die repetitiven Elemente und die putative Exonbereiche mit Hilfe der in blauer Schrift aufgeführten Programm-Module ermittelt wurde. Maskiert um die repetitiven Sequenzabschnitte mit Hilfe der Programme Censor, RepeatMasker und SST wurde die genomische DNA-Sequenz einem Homologievergleich in verschiedenen Referenz-Datenbanken unterzogen. Ebenso wurden mit den als kodierend hervorgehobenen Bereichsabschnitten der „GenScan"-Analyse eine gesonderte Homologiesuche, bestehend aus einer BlastP-, BlastN- und BlastX-Analyse, vollzogen.

Unter Zuhilfenahme des Annotierungsprogramms *RUMMAGE* (Taudien *et al*, 2000) konnte die genomische Sequenz nahezu umfassend nach funktionellen und organisatorischen Elementen hin untersucht werden. Die gleichzeitige Verknüpfung von insgesamt mehr als 20 Analyseprogrammen ließ eine sehr umfangreiche Auswertung der genomischen DNA-Sequenz zu. Der Einstieg zu den Einzeldaten erfolgte dabei über eine graphische Karte (siehe Abb. 16), die alle Ergebnisse gemäß ihrer räumlichen Lage auf der Consensussequenz zusammenfasste. Über Verknüpfungen konnte von der graphischen Übersicht auf die tabellarisch aufgelisteten Einzelergebnisse zugegriffen werden. Einzelergebnisse, beispielsweise die Homologien nach einer BlastN-Analyse zu bestimmten Datenbankeinträgen, waren ihrerseits verknüpft mit den Annotierungen aus den öffentlichen Datenbanken. So konnte bei der Auswertung der Ergebnisse, jedes ermittelte Ergebnis auf seinen Informationsgehalt und seine Aussagekraft hin überprüft werden. Ebenso wurden vereinzelt Bereiche nochmals einer manuellen Analyse unterzogen. Insbesondere der Homologievergleich mit Einträgen speziesspezifischer Datenbanken war ein wichtiges Instrument, die über das Rummage-Programm generierten Daten zu ergänzen.

Im Einzelnen wurde die Analyse der genomischen Sequenz mit folgenden Programmen vorgenommen:

2.17.2 Bestimmung des GC-Gehaltes

Eine rein statistische Auswertung der Basenpaarabfolge geschah durch die Tools ***LPC*** (Xiaoqiu Huang, 1994) und ***CPG*** (Larsen *et al.*, 1992), die GC-reiche Sequenzabschnitte und CpG-Inseln identifizieren. Das Programm *LPC* wurde mit vier verschiedenen Parametereinstellungen durchlaufen, das Programm *CPG* mit zwei Einstellungsvarianten. Die graphische Darstellung des GC-Gehaltes über den gesamt-sequenzierten Bereich geschah mithilfe des Programms ***GENEQUEST*** aus dem LASERGENE-Paket (DNA-STAR). Die Fenstergröße für die prozentuale Berechnung des GC-Gehaltes betrug 500 bp (siehe Abb. 29, Kap. 3.10.1).

Tab. 8 Parameter der GC-Gehalt Bestimmungsprogramme LPC und CPG. Für die zu untersuchenden Intervalle wurden zwei Werte 100 bp und 250 bp bei jeweils zwei unterschiedlichen Signifikanzgrenze eingegeben. Die Signifikanzgrenze gibt dabei den Schwellenwert (Prozentsatz) an, ab dem ein DNA-Abschnitt mit der Eigenschaft als GC-reich, CpG-reich, CpG-arm, bzw. „CpGlang" versehen wird.

Programm	Intervall	Signifikanzgrenze
LPC_GC_High	100 bp	60%
LPC GC Long	250 bp	55%
LPC_CpG_High	100 bp	10%
LPC CpG Low	250 bp	5%
CPG CpG High	100 bp	50%
CPG_CpG_Long	250 bp	40%

2.17.3 Repetitive Sequenzbereiche

Eine Analyse nach repetitiven und wenig komplexen DNA-Abschnitten wurde mit insgesamt sechs verschiedenen Programmen vorgenommen, deren Resultate zusammengefasst wurden, um die hohe Redundanz der Ergebnisse zu minimieren. In tabellarischer Form untergliedert, wurden drei Bereiche ausgewertet: ALU, Nicht-ALU und einfache „Repeats". Diese wurden nochmals in zu maskierende DNA-Bereiche zusammengefasst und in der Datenbank „COMPILE_REPEAT_default" abgelegt.

Die Programme *CENSOR* 1.1 (Jurka *et al.*, 1996), *REPEATMASKER* (Smit *et al.*, 1996*)* und *SST* (States, unveröffentlicht) identifizierten sowohl Genom-überspannende, wie auch einfache Sequenz-wiederholungen und Klassifizieren diese unter Zuhilfenahme verschiedener Referenzdatenbanken. Das Tool *XNUN* filtert sogenannte „short-period repeats" heraus, die typisch für Mikrosatelliten sind und maskiert diese Sequenzbereiche. Die Programme *TANDEM* und *INVERTED* suchen, wie es die deskriptiven Namen nahe legen, nach tandem-repetitiven, bzw. nach invertiert-repetitiven Sequenzabschnitten und stellen sowohl die sich wiederholenden Sequenzen, wie auch den umfassenden Gesamtbereich dar.

Tab. 9 Konfigurationsparameter der Analyse-Programme nach repetitiven Sequenzabschnitten. Angegeben sind die verwendeten Referenzsequenz-Datenbanken und die Stringenz (Sensitivität) mit der die Homologie überprüft wurde. Für das Programm INVERTED wurde der Intervallbereich angegeben, in dem das invertierte Sequenzmotiv gesucht wurde.

Programm	Datenbank	Sensitivität	Intervall
CENSOR	HumRep/MamRep	moderate	./.
RepeatMasker	Primate/Rodent	high	./.
SST	HumRep	./.	./.
XNUN	micro-satellites	0,1%	./.
TANDEM	micro-satellites	./.	./.
INVERTED	./.	./.	2.000 bp

2.17.4 Promotorvorhersage und Polyadenylierungsstellen-Suche

Putative Polymerase II Promotorbereiche wurden mithilfe des Programms *PROSCAN* (Dan Prestridge, 1995) bestimmt. Ebenso kam die Suchoption nach etwaigen Promotorstellen aus der *GENSCAN*-Analyse (Burge & Karlin, 1997) zum Einsatz. Putative Polyadenylierungsstellen wurden ebenfalls mit Hilfe der *GENSCAN*-Analyse charakterisiert.

2.17.5 Exonvorhersage

Die Vorhersage von putativen Exonsequenzen wurde mit fünf verschiedenen Programmen vorgenommen. Interne Exons, d.h. kodierende Sequenzabschnitte mit offenem Leserahmen und konservierter Spleißdonor- und Spleißakzeptorstelle, versuchen die Anwendungen *FEXHB* (Solovyev *et al.*, 1994), *MZEF* (Zhang *et al.*, 1997) und *XPOUND* (Thomas & Skolnick, 1994) zu bestimmen, jeweils in der optimierten Einstellung der entsprechenden Referenzspezies Mensch oder Maus. Mit Hilfe der Programme *GRAIL2* (Xu *et al.*, 1994) und *GENSCAN* (Burge & Karlin, 1997) war darüber hinaus durch das gleichzeitige Bestimmen von Promotor- und Polyadenylierungssequenzen auch eine Vorhersage für die flankierenden Exons am Anfang und am Ende eines putativen Gens möglich.

Für die effiziente Auswertung der verschiedenen Programmergebnisse wurde mithilfe der COMPILE_Exon-Option der *RUMMAGE*-Analyse gearbeitet. Eine summarische Auflistung aller potentiellen Exonbereiche aus allen fünf verwendeten Programmen erfolgte als Datenpool in der generierten Datenbank „COMPILE-Exon-Weak". Bereiche die mehrfach, von verschiedenen Vorhersage-Programmen ermittelt wurden, wurden in einer zweiten Datenbank namens „COMPILE-Exon-Strong" zusammengestellt.

Das Programm *EXONSAMPLER* (Weber *et al.*, unveröffentlicht) sucht in der genomischen Sequenz nach Übereinstimmungen mit EST-Datenbank-Einträgen und versucht diese Treffer unter Berücksichtigung der Spleißdonor/Spleißakzeptor-Consensussequenz anzuordnen. Die Resultate dieser Analyse wurden mit der Methode *est2genome* (Mott, 1997) vorgenommen und zusammengestellt.

2.17.6 Homologie- und Datenbanksuche

Die Suche nach bekannten Sequenzhomologien fand mit dem bekannten „Basic local alignment search tool": *BLAST* (Altschul *et al.*, 1997) statt. Unter Verwendung der verschiedenen Algorithmen *BLASTN* (Vergleich Nukleotide mit Nukleotide) und *BLASTX* (Vergleich Nukleotide mit Proteinen) wurde sowohl die gesamte genomische Sequenz mit den Datenbanken GenEMBL, EMBLnew, E.coli, dbSTS, dbEST, CDS und GenPept verglichen, wie auch die selektiven Bereiche der Exonvorhersageprogramme aus der Datenbank „COMPILE_EXON". Dies geschah nach der Maskierung der Sequenzbereiche, die in der Datenbank „COMPILE_REPEAT" als repetitive Abschnitte identifiziert worden waren. Eine *BLASTP*-Analyse (Vergleich Aminosäuren mit Proteinen) fand mit den abgeleiteten Aminosäure-Sequenzen der Einträge aus der „COMPILE_EXON"-Datenbank statt, die in einer dritten Datenbank „COMPILE_PROD"

als putativen mRNA-Sequenzen vorlagen. Als Referenz dienten die Einträge der OWL-Datenbank (Bleasby & Wootton, 1996) unter der Adresse <ftp://ncbi.nlm.nih.gov/repository/OWL>.

Tab. 10 **Konfigurationsparameter der benutzten BLAST-Algorithmen**. Ausschlusskriterium für eine signifikante Homologie war die Sequenzübereinstimmung von mindestens <Identität> über einen Sequenzbereich von mindestens <Sequenzlänge> mit einer Trefferquote von weniger als <Wahrscheinlichkeit>. Für den Proteinvergleich durfte zudem der Prozentsatz für positive Austausche nicht niedriger liegen als <Pos.>.

Algorithmus	Sequenzlänge	Identität	Pos.	Wahrscheinlichkeit
BLASTN_HOMO_GenEMBL	100 bp	97 %	./.	$\leq 10^{-20}$
BLASTN_HOMO_Ecoli	100 bp	97 %	./.	./.
BLASTN_HOMO_STS	60 bp	90 %	./.	./.
BLASTN_HOMO_EST	60 bp	65 %	./.	$\leq 10^{-20}$
BLASTN_HOMO_CDS	60 bp	75 %	./.	$\leq 10^{-20}$
BLASTX_HOMO_GenPept	30 AS	50 %	65 %	$\leq 10^{-20}$
BLASTN_EXON_Human	60 bp	80 %	./.	./.
BLASTN_EXON_EST	60 bp	60 %	./.	./.
BLASTN_EXON_EMBLnew	60 bp	80 %	./.	./.
BLASTP_PROD_OWL	65 AS	./.	65 %	./.

2.17.7 MAR-Analyse

Zur Bestimmung von DNA-Sequenzmotiven, die Einfluss auf die DNA-Konformation nehmen und eine Krümmung bzw. Drehung der DNA-Helix verursachen, bzw. für die Anheftung von Strukturelementen wichtig sind, wurden die Consensussequenzen einer sogenannten MAR-Analyse unterzogen. Diese „Matrix-attachment-regions" (MAR) wurden mit Hilfe des Programms *MARFINDER* (http://www.futuresoft.org/MAR-Wiz/) detektiert und graphisch in Form einer durchgehenden Wahrscheinlichkeitskurve (von 0 bis 100%) zur untersuchten Sequenz dargestellt. Für die Berechnung wurden insgesamt sechs verschiedene Regeln berücksichtigt. Nachfolgend sind diese Kriterien mit den entsprechenden Sequenzmotiven tabellarisch zusammengefasst:

Tab. 11 **Sequenzmotive der MarFinder-Analyse** unterteilt nach den zugewiesenen Funktions-eigenschaften der jeweiligen Motive.

MAR-Sequenzeigenschaft	MAR-Sequenzmotive
Replikationsursprung	ATTA, ATTTA, ATTTTA
TG-reiche Region	TGTTTTG, TGTTTTTTG, TTTTGGGG
gekrümmte „curved" DNA	$AAAAn_7AAAAn_7AAAA$, $TTTTn_7$ $TTTTn_7TTTT$, TTTAAA
verdrehte „kinked" DNA	TAn_3TGn_3CA, TAn_3CAn_3TG, TGn_3TAn_3CA, TGn_3CAn_3TA CAn_3TGn_3TG, CAn_3TGn_3TA
Topoisomerase II Erkennungsstellen	RnYnnCnnGYnGKTnYnY, GTnWAYATTnATnnR
AT-reiche Region	WWWWWW

Der Signifikanzwert für eine MAR-Region wurde in der Analyse auf 0,6 eingestellt. Dieser Wert kann als normalisiertes MAR-Potential zwischen 0 und 1 interpretiert und als relative Wahrscheinlichkeit für die reelle Existenz einer MAR-Region verstanden werden. Der Wert 0,6 würde demnach einen Bereich mit 60%igem Potential für eine MAR-Region signalisieren.

2.18 Interspeziesvergleich der genomischen Sequenzen

Eine Darstellung für den Grad der genomischen Konservierung zwischen zwei unterschiedlichen Spezies fand durch Dotplot-Analysen mit dem Programm *MEGALIGN* der Lasergene-Software-Kollektion (DNA-STAR, USA), bzw. in Form einer PIP-Analyse („percent identity plot") mit dem Programm *PIPMAKER* (Schwartz *et al.,* 2000) statt.

Für die Dotplot-Analyse wurden beide verifizierten Consensussequenzen direkt miteinander verglichen und in einem zweidimensionalen Koordinatensystem einander gegenübergestellt. Die Programm-Parameter wurden so gewählt, dass zum einen alle homologen Bereiche dargestellt, zum anderen aber möglichst wenig „Hintergrund" durch repetitive Elemente erzeugt wird. Das beste Ergebnis wurde bei einer Stringenz von mindestens 65% Homologie über einen Sequenzbereich von 50 bp erzielt. Während der Analyse wurde das Sequenzfenster jeweils um 10 bp iterativ verschoben.

Zur differenzierten Darstellung von Homologien zweier unterschiedlicher Sequenzen wurde die PIP-Analyse im „Advanced"-Modus angewendet. Sie stellt Homologien der Ausgangssequenz mit der Vergleichssequenz ab einem Wert von 50% Homologie ebenfalls graphisch in einem zweidimensionalen Koordinatensystem dar. Allerdings wird auf der Koordinatenachse nicht die Vergleichssequenz selbst, sondern der Grad der prozentualen Übereinstimmung aufgeführt. Es entsteht somit über der Abszisse unter gleichzeitiger Angabe der prozentualen Ähnlichkeit immer an den Stellen ein Punkt oder ein Balken, die die Homologie zur Vergleichssequenz besitzen.

Diese hochauflösende lineare Darstellungsform für den Grad der Sequenzhomologie zweier unterschiedlicher Sequenzen erlaubt es außerdem, weitere Aspekte der Sequenzanalyse mit in die Graphik einzubeziehen, um so verschiedene Informationen in einen gemeinsamen Kontext zu bringen. So wurden in die PIP-Darstellung (siehe Kap. 3.8.1.2) sämtliche repetitiven Bereiche aus der *REPEATMASKER*-Analyse eingetragen. Maßstabsgerecht wurden diese Abschnitte über den Bereich ihrer Erstreckung dargestellt und durch verschiedene Symbole je nach Art des repetitiven Motivs wiedergegeben. Ebenso wurden GC-reiche Bereiche in zwei unterschiedlichen Signifikanzwerten ($\geq$0,60 und $\geq$0,75) hervorgehoben. Bezüglich ihrer proteinkodierenden Funktion wurden darüber hinaus alle genkodierenden Abschnitte markiert. Des weiteren fanden alle putativen Genbereiche aus der Datenbank „COMPILE-EXON-strong", berechnet durch die verschiedenen Exonvorhersage-programme, und alle Bereiche mit signifikanten Homologie zu EST-Datenbankeinträgen aus der *RUMMAGE*-Analyse ihren Eintrag.

2.19 Reagenzien und Materialien

2.19.1 Puffer und Lösungen

Agar-Platten	15g Agar-Agar ad 1000 ml LB-Medium
APS-Lösung	10% Ammoniumpersulfat
Antibiotika	Ampicillin: Stammlösung: 50 mg/ml Mediumskonzentration: 50-75 µg/ml Chloramphenikol: Stammlösung: 34 mg/ml Mediumskonzentration: 170 µg/ml Kanamycin: Stammlösung: 25 mg/ml Mediumskonzentration: 25 µg/ml
CBIL-Sol.1	50 mM Glucose 25 mM Tris/HCl, pH 8,0 10 mM EDTA, pH 8,0
CBIL-Sol.2	0,2 N NaOH 1% SDS
CBIL-Sol.3	3 M KOAc 11,5 ml Eisessig ad 100 ml H_2O
Church-Puffer	0,5 M Na-Phosphat, pH 7,2 7% SDS 1 mM EDTA
Dialysepuffer	25 mM Tris 300 mM NaCl 10 mM Na_2EDTA
DNA-Auftragungspuffer (für die Agarose-Gelelektrophorese)	0,25% Bromphenolblau 0,25% Xylencyanol 40% Sucrose
DTM-Puffer	je 100 µl dATP, dGTP, dTTP in 250 mM Tris/HCl, pH 7,0 25 mM $MgCl_2$ 50 mM β-Mercaptoethanol
Ethidiumbromid-Färbelösung	5 µg/ml H_2O
FM-Medium (2x)	65% Glycerin 100 mM $MgSO_4$ 25 mM Tris/HCl, pH 8,0
LB-Medium	10 g NaCl 10 g Tryptone 5 g Hefeextrakt ad 1 l Aqua bidest. pH 7,5 (mit NaOH)

LB-Agar	1 l LB-Medium + 20 g Agar
MOPS (10x)	200 mM 3(N-morpholin)Propan-Sulfonsäure 50 mM Na-Acetat 10 mM EDTA, pH 7,5
MIT-Sol.1	10 mM EDTA, pH 8,0
MIT-Sol.2	0,2 M NaOH 1% SDS
MIT-Sol.3	1,87 M KAc 23 ml Eisessig ad 200 ml H_2O
OL	90 OD U 5´-pd(N6)/ml TE
OLB-Puffer	250 mM Tris/HCl 25 mM $MgCl_2$, pH 8,0 50 mM β-Mercaptoethanol je 100 µM dNTPs 1 mM Tris 1 mM EDTA, pH 7,5 1 M HEPES/DTM/OL (25:25:7)
"Orange-Dye"-Auftragungspuffer	0,175 g Orange G 15 g Sucrose ad 50 ml H_2O
P1-Puffer (QIAGEN)	50 mM Tris/HCl, pH 8,0 10 mM EDTA 100 µg/ml RNase A
P2-Puffer (QIAGEN)	200 mM NaOH 1% SDS
P3-Puffer (QIAGEN)	3 M Kaliumacetat, pH 5,5
QBT-Puffer (QIAGEN)	750 mM NaCl 50 mM MOPS, pH 7,0 15% Isopropanol 0,15% Triton-X 100
QC-Puffer (QIAGEN)	1,0 M NaCl 50 mM Tris/HCl, pH 7,0 15% Isopropanol
QF-Puffer (QIAGEN)	1,25 M NaCl 50 mM Tris/HCl, pH 8,5 15% Isopropanol
RLT-Puffer (QIAGEN)	50 mM Tris/HCl, pH 8,0 140 mM NaCl 1,5 mM $MgCl_2$ 0,5% Nonidet P-40 0,1 M β-Mercaptoethanol 1.000 U/ml RNasin 1 mM DTT
SOB-Medium	20 g Trypton

	5 g Hefeextrakt 0,5 g NaCl ad 1 l H_2O, pH 7,0 10 mM $MgCl_2$ 10 mM $MgSO_4$
SOC-Medium	SOB-Medium 20 mM Glucose
SSC-Puffer (10x)	1,5 M NaCl 150 mM Natriumcitrat pH 7,0
TBE-Puffer (1x)	90 mM Tris 90 mM Borsäure 1,25 mM Na_2EDTA
TE-Puffer (1x)	10 mM Tris/HCl, pH 8,0 1 mM Na_2EDTA
Waschpuffer für FISH	4x SSC 0,1% Triton-X 100
Waschpuffer nach *Church*	0,4 M Na_3PO_4, pH 7,2 0,1% (w/v) SDS
X-Gal-LB-Agar	1 l LB-Agar + 80 mg X-Gal (gelöst in 1 ml Dimethylformamid) + 50 mg IPTG

2.19.2 Radioisotope, Enzyme, Markierungssysteme

ABI PRISM™ Big-Dye™ Terminator Cycle Sequencing Ready Reaction Kit	Fa. APPLIED BIOSYSTEMS (USA)
ABI PRISM™ Ready Reaction Dye Deoxy Terminator Cycle Sequencing Kit	Fa. APPLIED BIOSYSTEMS (USA)
Biotin-16-dUTP	Fa. BOEHRINGER (Mannheim)
Cot 1-DNA	Fa. GIBCO BRL (USA)
Digoxigenin-11-dUTP	Fa. BOEHRINGER (Mannheim)
DNase I	Fa. ROCHE DIAGNOSTICS (Mannheim)
dNTPs	Fa. BOEHRINGER (Mannheim)
Klenow-DNA-Polymerase	Fa. NEW ENGLAND BIOLABS (Frankfurt)
Nick-Translations-System	Fa. GIBCO BRL (USA)
Plasmid-Safe™ ATP-dependent DNase	Fa. EPICENTRE TECHNOLOGIES CORPORATION (USA)
pUC18-DNA x Sma I/BAP	Fa. PHARMACIA (USA)
Restriktionsenzyme	Fa. NEW ENGLAND BIOLABS (Frankfurt)
Reverse Transkriptase	Fa. PERKIN ELMER (USA)
RNase A	Fa. BOEHRINGER (Mannheim)
RNase Inhibitor	Fa. MBI FERMENTAS (St. Leon-Rot)
RNeasy™ Total RNA Kit	Fa. QIAGEN (Hilden)

Salmon sperm DNA	Fa. PHARMACIA (Freiburg)
T4-DNA-Ligase	Fa. NEW ENGLAND BIOLABS (Frankfurt)
Taq-DNA-Polymerase	Fa. GIBCO BRL (USA)
[α-^{32}P] dCTP	Fa. AMERSHAM LIFE SCIENCE (Braunschweig)
α-Satellit Chromosom 11 (Mensch)	Fa. ONCOR (USA)
α-Satellit Chromosom 7 (Maus)	Fa. ONCOR (USA)

2.19.3 Bakterienstämme

XL1-Blue MRF´ (Epicurian Coli®, STRATGENE)	Δ(mcrA)183 Δ(mrcCB-hsdSMR-mrr)173 endA1 supE44 thi-1 recA1 gyrA96 relA1 lac[F´proAB lacl Z ΔM15 Tn10 (Tet^r)]

2.19.4 Klonbibliotheken

Für die Untersuchungen in dieser Arbeit wurden Klone aus folgenden Bibliotheken verwendet, die über Filtersets des Deutschen Ressourcenzentrums (RZPD) in Berlin identifiziert und bezogen wurden:

Bibliothek **Nr. 114** (=LANLc11 Human chromosome 11 cosmid (SRL) Library)

18.048 Klone in 47 x 384-well Platten

- DNA-Quelle: *Homo sapiens*, somatic cell hybrid J1 E
- Chromosom 11 ("flow-sorted")
- hergestellt von: *Larry Deaven*, *J. Longmire*

Bibliothek **Nr. 704** (=RPCI 1,3-5 Human PAC Library)

582.528 Klone auf 15 Filtern

- entspricht einer Redundanz von 16
- DNA-Quelle: *Homo sapiens*, Blut
- hergestellt von: *P. de Jong, P. Ioannou*

Bibliothek **Nr. 709** (=RPCI 6 Human PAC Library)

92.160 Klone in 240 x 384-well Platten

- entspricht einer Redundanz von 4
- DNA-Quelle: *Homo sapiens*, weiblich
- hergestellt von: *P. de Jong*

Bibliothek **Nr. 711** (=RPCI-21 Mouse PAC Library)

258.048 Klone auf 10 Filtern

- entspricht einer Redundanz von 13
- DNA-Quelle: *Mus musculus*, Stamm: 129/SvevTACfBr
- Gewebe: Milz
- hergestellt von: *K. Osoegawa, P. de Jong*

Bibliothek **Nr. 731** (=RPCI-23 Mouse BAC II Library)

184.320 Klone

- entspricht einer Redundanz von 11
- DNA-Quelle: *Mus musculus*, Stamm: C57BL/6J
- Gewebe: Niere+Hirn
- hergestellt von: *P. de Jong, K. Osoegawa*

2.19.5 Primer

Oligonukleotide wurden bei den Firmen ROTH (KARLSRUHE), GIBCO LIFE TECHNOLOGIES (Eggenstein) und GENSET (Paris) in einer entkoppelten und entsalzten Qualität bezogen.

2.19.5.1 Primer für STS-Sonden:

D11S1020-F	5´-CAG GGA TGG CAG TCT CTC GTC TCC-3´
D11S1020-R	5´-TTA CCA CTA TCC TAT GAT CAC TTC-3´
D11S1134-F	5´-CTG CCA AGA TTT CAG AGG ATG-3´
D11S1134-R	5´-GCT CCA CTA GGC AGT GCC-3´
D11S1152-F	5´-ATG AAG GGC GTA GTC CCC-3´
D11S1152-R	5´-TGG CCT GGT CCC TTT AGA G-3´
D11S932-F	5´-TCG TAT AGC ACA CCT TGG C-3´
D11S932-R	5´-CTT ATC ATC TCT GGG TAG TGA AGT C-3´
SHGC-148637-F	5´-TTC CAT TTG TTG CTT TCT TCC AT-3´
SHGC-148637-R	5´-TGC CAT AAC TCT ATA GGG GCA GA-3´
D11S2704-F	5´-AAG TCC CTC TTA AAA CAT GG-3´
D11S2704-R	5´-AAC CTA GGC TCT CTT GTC AG-3´
D11S3261-F	5´-TAA GTT TCC TGA GGT GTC TG-3´
D11S3261-R	5´-TGG GTG ACT GAC TCT ATT TC-3´
D11S3436-F	5´-TGC AGC TTC TCC TAT TGT AC-3´
D11S3436-R	5´-TCT CAA TTC CTT TCT CAA TG-3´
WI-14382-F	5´-CAC CTG TTA TTT TTC ACC TTT CG-3´
Wi-14382-R	5´-CTG AGT TTA TAG AAA GCT ATT GCA-3´
D11S2050-F	5´-ACT TGC CCT CTT GGC CTC-3´
D11S2050-R	5´-CAC AGT GGG GTC CAG AAT G-3´
D11S572-F	5´-GGT AAG ACT GCA GTG AGC CAT GAT TAC-3´
D11S572-R	5´-GTA CAA TAG AAG AAG CTG CAG GTG TCC-3´
D7Mit127-f	5´-AGC CAC CAA CCA ACC CTT C-3´
D7Mit127-r	5´-CCT ATG CTA TAA AAA AAT TTG GGC-3´
D7Mit219-f	5´-TCA GAT GTG GAG CCA CAC AT-3´
D7Mit219-r	5´-CAT ATT CAA GGA GGA CAG AGG C-3´
D7Mit280-f	5´-TGA GTA CCC CTT ACA GAG CAT G-3´
D7Mit280-r	5´-TGG TTA GTG TTC TTT AGG ACA CTC C-3´
D7Mit325-f	5´-TTA TTG TAC CTA CCA AGG TCT CAG C-3´
D7Mit325-r	5´-AGC AAA ACA CAC ATG CCT GA-3´
D7Mill1-f (B13T7f)	5´-GCA TGA GCA GCA CTG GTT TA-3´
D7Mill1-r (B13T7r)	5´-TCC TCT CCT TCT CAG AGC CA-3´

D7Mill2-f (M65Lf)	5´-CGG AAT TCA GCA GTC GTA GA-3´
D7Mill2-r (M65Lr)	5´-TCC TTC TAG GAG CAT GCT GA-3´
D7Mill3-f (M72Rf)	5´-TGC GCA GAA ACA ATC ACC TA-3´
D7Mill3-r (M72Rr)	5´-CAA GAC GTG AAC AAC CTG GA-3´
D7Mill4-f (M65Rf)	5´-GAA CAG GAC TTG GAC GTG GT-3´
D7Mill4-r (M65Rr)	5´-AAC TCT GCA TGA CCC AGG TC-3´
D7Mill5-f (CA37f)	5´-GTA CTT AGT ATT AAC TTC AGG TC-3´
D7Mill5-r (CA37r)	5´-TGA CCA GTG GCA TAC AGA AC-3´
D7Mill6-f (B4Sp6f)	5´-CAT GGG CTC CTC TGC CAC CC-3´
D7Mill6-r (B4Sp6r)	5´-TAA AGT GTG GAC CAA GTA CAC-3´
D7Mill7-f (CA39f)	5´-GAG AAC AGA CTC CAC AAA TTG-3´
D7Mill7-r (CA39r)	5´-ACC TGA GAG ACT ACA TCA GG-3´
D7Mill8-f (M74Rf)	5´-AGA CAG GAA GGG GAA GGA AA-3´
D7Mill8-r (M74Rr)	5´-GGC CCT TCT GGA TGA GCT TA-3´
D7Mill9-f (M70Lf)	5´-TGC CAA TTG TCA AAG CTC TG-3´
D7Mill9-r (M70Lr)	5´-ATC AAA TGC CCC ATT GGA TA-3´
D7Mill10-f (P1Sp6f)	5´-ATG GGT CTC ATC TGG AAG GT-3´
D7Mill10-r (P1Sp6r)	5´-CTG AAA GAC TGG AAG AGG CA-3´
D7Mill11-f (M31Lf)	5´-AGC TCA GAT CAA CCA TAA CTG C-3´
D7Mill11-r (M31Lr)	5´-CCT TCC AAG GAC TGT GCG T-3´
D7Mill13-f (B2Sp6f)	5´-GAT CCC TGA AGG TAG GAC TG-3´
D7Mill13-r (B2Sp6r)	5´-CCT ATA TAC TGT CTG AGA ATG-3´
D7Mill14-f (M36CA1f)	5´-GAT CTG CCC AGC CAC GTC TCC G-3´
D7Mill14-r (M36CA1r)	5´-CCC GGG CGC TGG GGT GTA T-3´

2.19.5.2 Primer für Hybridisierungssonden:

1174K15-28for	5´-TGT CTC AAC GGA GAC ATT GC-3´
1174K15-28rev	5´-ACC CAG GGA AGA GAG GAC AT-3´
1184A15-28for	5´-GAA CAC GAG GTG GTC CAA AT-3´
1184A15-28rev	5´-CAT AGT CCA AGT CGC CAT GA-3´
119G5_T7f	5´-GCA CTT CGA TGC TCA GTC AC-3´
119G5_T7r	5´-TCC CAG AAA TCA ACC CAT TC-3´
12G13uni-F	5´-CAC CAT GCT CAC AGC TGG-3´
12G13uni-R	5´-GCT ACT GGA GGC GTT CAC TC-3´
12G13-28-F	5´-GAC ACA TCA CTG TCC AAC CG-3´
12G13-28-R	5´-AGC CTC AGG GAA GAC ACA GA-3´
12G13-08F	5´-CGG ATT TTT CCG TCA GAT GT-3´
12G13-08R	5´-GCA GAT CCC TAA ACA GCA GC-3´
RBTN-EX2-F	5´-CCG ATG CTC TCC GTC CAG-3´
RBTN-EX2-R	5´-AGG TAG TCG CGT CGG CAC-3´
BAC221D7-T7-f	5´-ACC CCA AAT TCC ACA AAC AA-3´
BAC221D7-T7-r	5´-CGA TTG GTC TCC TCA TCT CTG-3´
287P4-T7-f	5´-TGT CCA AAT CCC CTT CAG AG-3´
287P4-T7-r	5´-AAG GTG TGA AAC GTG GGA AG-3´
287P4-Sp6-f	5´-CAA GGT CTC CCT GGA CAT TAG T-3´

287P4-Sp6-r	5´-CCC ATT TTC TTT TCC TTC TGG-3´

2.19.5.3 Sequenzierprimer:

M13 univers	5´-GTA AAA CGA CGG CCA GT-3´
M13 revers	5´-CAG GAA ACA GCT ATG ACC-3´
Sp6	5´-GAT TTA GGT GAC ACT ATA G-3´
T7	5´-TAA TAC GAC TCA CTA TAG GG-3´
T3	5´-ATT AAC CCT CAC TAA AGG GA-3´
PAC18628	5´-TCT GCC GTT TCG ATC CTC-3´
PAC2808	5´-CGA CGA TAG TCA TGC CCC-3´

2.19.5.4 RT-PCR-Primer:

putTub_EX1-f	5´-GGC ATT CAA AGC AGA ACA GG-3´
putTub_Ex2-f	5´-GAA GGA AGG AAG GGA AAT CG-3´
putTub_EX2-r	5´-CGA TTT CCC TTC CTT CCT TC-3´
Tub_Ex3-r	5´-GCT GCC CTC ATC ATC TAG GA-3´
altSP_Tub_Ex4-f	5´-CCG ACT CGA TTG CCA GTG TA-3´
altSP_Tub_Ex6-r	5´-GCT GGA GCT GTT TTC ATC CTC A-3´
Tub_Ex6-r	5´-TTG CTG TTT AGC TGG GAG GAG-3´
Tub_Ex9-f	5´-GCA CCA AGT TCA CCG TTT ATG A-3´
Tub_Ex10-f	5´-CTC GGA AGA TGA GTG TGA TCG T-3´
Tub_Ex11-f	5´-GGC AGA ACA AGA ACA CGG AGA-3´
Tub_Ex12-r	5´-CTG GAC AGA GCA ATG GCA AAG-3´
Tub_Intr11-r	5´-GCA CAA GCA AGA ATG GTG GAG-3´
hTUB5pr-RT1-f	5´-GGC GAA TAC AGG GAA TTT CA-3´
hTUB5pr-RT1-r	5´-CGA TCT CCC TTC CTT CCT TC-3´
hTUB5pr-RT2-f	5´-GGT CCC GGG GAG GAT AC-3´
hTUB5pr-RT2-r	5´-GCG GCT TGG AAG TCA TGT-3´
hTUBc_Ex1-f	5´-CTC CTG CCT CTG GCA TAA CTG T-3´
hTUBc_Ex2-r	5´-ATG GTG CTG ACG CTC ACA TCT A-3´
hTUB_Ex1-f	5´-GTC CCG GGG AGG ATA CGT C-3´
hTUB_Ex3-f	5´-GCT GGA GCA GAA GCA GAA GA-3´
hTUB_Ex3-r	5´-GCT GCT GAG GTA GGA CTC CA-3´
hTUB_Ex4-r	5´-TGT GCT TTC CCT TCT TCT CC-3´
mLmo1_ex1-f	5´-CGA GAT TCC CCC ATC TCT TT-3´

mLmo1_ex1-f2	5´-CTC TCT TCC CCC TTC TCT CTC C-3´
mLmo1_ex1-f3	5´-AAA TCC GAG GGG AAA ACA CTT T-3´
mLmo1_ex2-f	5´-GAT GCT CTC CGT CCA ACC TA-3´
mLmo1_ex2-r	5´-GAG GTT GGC CTT TGG TGT AGA-3´
mLmo1_ex3-f	5´-TTT TGG CAC CAC AGG AAA CT-3´
mLmo1_ex3-r	5´-GCG AAG CAG TCA AGG TGA TA-3´
mLmo1_ex4-f	5´-TGC GTG GGA GAC AAA TTC TT-3´
mLmo1_ex4-r	5´-GTG GTG GAG AAC AGC CAC CTT C-3´

2.19.6 Molekulargewichtsstandards

3	100 bp-Leiter	4	Fa. GIBCO BRL (USA)
5	123 bp-Leiter	6	Fa. GIBCO BRL (USA)
7	1 kb-Marker	8	Fa. GIBCO BRL (USA)
9	λ x Hind III	10	Fa. BOEHRINGER (Mannheim)
11	λ x Hind III–Concatemere (für PFGE)	12	Fa. PHARMACIA (Freiburg)

2.19.7 Chemikalien

13	Acrylamid	Fa. ROTH (Karlsruhe)
14	Agar-Agar	Fa. DIFCO (USA)
15	Agarose	Fa. EUROGENTEC (Belgien)
16	Ampicillin	Fa. RATIOPHARM (Ulm)
17	Bacto-Hefeextrakt	Fa. DIFCO (USA)
18	Bacto-Trypton	Fa. DIFCO (USA)
19	Chromosomen-Medium IA	Fa. GIBCO/INVITROGEN (Karlsruhe)
20	EDTA	Fa. BOEHRINGER (Mannheim)
21	Ethanol	Fa. RIEDEL-DE-HAEN (Seelze)
22	Ethidiumbromid	Fa. ONCOR (USA)
23	Harnstoff	Fa. ROTH (Karlsruhe)
24	IPTG	Fa. BOEHRINGER (Mannheim)
25	LMP-Agarose	Fa. BMA (USA)
26	Kanamycin	Fa. RATIOPHARM (Ulm)
27	Phenol	Fa. ROTH (Karlsruhe)
28	Phenol-Chloroform	Fa. ROTH (Karlsruhe)
29	SDS	Fa. INC (USA)
30	Tris	Fa. ROTH (Karlsruhe)
31	Tris-HCl	Fa. GERBU (Gaiberg)

32	X-Gal	Fa. EUROGENTEC (Belgien)

33 Alle weiteren Feinchemikalien wurden bei den Firmen MERCK (Darmstadt), ROTH (Karlsruhe), SERVA (Heidelberg) oder SIGMA (Deisenhofen) bezogen.

2.19.8 Materialien

34	DNA-Nebulizer	Fa. GATC (Konstanz)
35	Elektroporationsküvetten	Fa. BIORAD (München)
36	MultiScreen®Assay-System	Fa. MILLIPORE (Eschborn)
37	Nick-Columns	Fa. PHARMACIA (Freiburg)
38	Nylonmenbran (Hybond^{TM}-N+)	Fa. AMERSHAM LIFE SCIENCE (Braunschweig)
39	Röntgenfilme (Hyperfilm^{TM}-MP)	Fa. AMERSHAM LIFE SCIENCE (Braunschweig)
40	Sephadex-G50 (fine, ultrafine)	Fa. PHARMACIA (Freiburg)

2.19.9 Geräte

41	Computer	versch. Apple Macintosh PowerPCs: 7600/132 bis G3
42	mit den Betriebssystemen MacOS 7.x und 8.x	
43	versch. PCs: PI-100MHz, PII-850 MHz	
44	mit den Betriebssystemen Win95, Win2000	
45	SUN Workstation: UltraSparc 1 – 200 MHz	
46	mit dem Betriebssystem Solaris 2.6 auf SunOS 5.x	
47	DNA-Sequenziergeräte	ABI PRISM^{TM} 377-96
48	ABI PRISM™ 377-XL	
49	ABI 373A	
50	Elektrophorese-Kammern	EASY-CAST – System
51	Fa. OWL SCIENTIFIC INC. (USA)	
52	Gel-Imaging-System	E.A.S.Y. System 429K
53	Software: E.A.S.Y. plus Rev. 4.24	
54	Fa. HEROLAB (Wiesloch)	
55	Elektroporationsanlage	GenePulser + Controller + Extender
56	Fa. BIORAD LABORATORIES (München)	
57	Hybridisierungsofen	Hybridiser HB-2
58	Fa. TECHNE (USA)	
59	PCR-Geräte	DNA Thermal Cycler
60	Fa. PERKIN ELMER CETUS (USA)	

PE 9700 Cycler Fa. PERKIN-ELMER		
PTC 200™ und PTC 100™ Fa. MJ RESEARCH,INC. (USA)		
pH-Meter Fa. SCHOTT-GERÄTE (Mainz)	CG-840	
Photometer Fa. PHARMACIA (Freiburg)	Pharmacia LUC. Ultrospec III	
Pulsfeldgelelektrophorese-Anlage Fa. BIORAD (München)	CHEF Mapper-Einheit	
Schüttelinkubator Fa. HERAEUS (Hanau)	Typ B5042	
Spannungsgeräte Fa. PHARMACIA (Freiburg)	LKB-GPS 200/400	
Sterilbank Fa. CLEANAIR NSF49	Antair BSK	
Szintillationszähler Fa. WALLAC (Freiburg)	Liquid Scintillations Counter WALLACE-1410	
UV-„Crosslinker" Fa. STRATAGENE (Heidelberg)	UV-Stratalinker Modell 2400	
Vakuumtrockner Fa. UNI-EQUIP (Martinsried)	Uni-vapo 100H + Uni-jet II	
Waagen Fa. SARTORIUS (Göttingen)	L610 – D2 + R180D – D1	
Wasserbäder Fa. BRAUN (Melsungen) F10 Fa. JULABO (USA)	Thermomix 1420	
2219 Multitemp II Fa. LKB (Schweden)		
Certomat WR (Schüttelwasserbad) Fa. B. BRAUN (Melsungen)		
Zentrifugen Fa. HERAEUS SEPATECH (Hanau)	Biofuge A	
Eppendorf Zentrifuge 5804R + 5415C Fa. EPPENDORF (Hamburg)		
Hettich Zentrifuge EBA12R Fa. HETTICH (Tuttlingen)		

99 Sigma 3K12 Zentrifuge
100 Fa. SIGMA (Deisendorf)

101 Sorvall RT 6000D und Sorvall RC5C
102 Fa. DU PONT (USA)

2.19.10 BioSoftware-Programme

103	*BIOEDIT 5.0.9*	DEPT. OF MICROBIOLOGY, NORTH CAROLINA STATE
104	UNIVERSITY (USA)	
105	*CHROMAS 1.5*	Fa. TECHNELYSIUM PTY LTD (Australia)
106	*CONSED* (Vers. 0.981015)	DEPT. OF MOLECULAR BIOTECHNOLOGY, UNIV. OF
107	WASHINGTON (USA)	
108	*DATA COLLECTING 2.0 – 2.6*	Fa. PE BIOSYSTEMS (USA)
109	*LASERGENE 4.0*	mit den Programmen: EditSeq, MegAlign, GeneQuest
110	Fa. DNASTAR (USA)	
111	*PHREDPHRAP* (Vers. 0.96731)	DEPT. OF MOLECULAR BIOTECHNOLOGY, UNIV. OF
112	WASHINGTON (USA)	
113	*REFERENCEMANAGERPROFESSIONAL 10.0*	RESEARCH INFORMATIONS SYSTEMS (USA)
114	*SEQUENCHER 3.1 – 4.1*	Fa. GENECODES (USA)
115	*SEQUENCING ANALYSIS 2.0 – 3.3*	Fa. PE BIOSYSTEMS (USA)

3 ERGEBNISSE

Die Ergebnisse der vorliegenden Arbeit sind Bestandteil des komparativen Sequenzierprojekts der Johannes Gutenberg-Universität Mainz im Rahmen des Deutschen Humangenomprojektes über eine Megabase genomischer DNA der humanen Chromosomregion 11p15.3 und der syntänen Region auf Chromosom 7 der Maus. Die vorliegenden Resultate charakterisieren den distalen Bereich der Megabasenregion mit den Markergenen *LMO1* und *TUB*.

3.1 Die zu sequenzierende Gesamtregion auf Chromosom 11p15.3 (Mensch), bzw. Chromosom 7 (Maus)

Vorarbeiten von Barbara Seipel (1996) mit Hilfe der Zwei-Farben-Fluoreszenz-in-situ-Hybridisierung (FISH) gaben Hinweise darauf, dass der genomische Abschnitt zwischen den humanen Genen *LMO1* und *WEE1* in der Chromosomenbande 11p15.3 lokalisiert ist und dass sich die Abfolge der dortigen Gene in einem auf etwa eine Megabase geschätzten Bereich über die Evolution hinweg in Mensch und Maus konserviert erhalten hat. Es konnte gezeigt werden, dass die Gene *LMO1*, *ST5*, *CEGP1* und *WEE1* in dieser Reihenfolge von Telomer zu Centromer auf dem humanen Chromosom 11 zu finden sind. Außerdem deutete sich an, dass die gleiche Anordnung dieser vier Gene in invertierter Abfolge (proximal → distal) auf Chromosom 7 der Maus beibehalten wurde.

Die Syntänie dieses chromosomalen Abschnitts konnte in den nachfolgenden Analysen, die Gegenstand dieser Arbeit sein werden, bestätigt werden und beschränkte sich nicht nur auf die Anordnung der einzelnen Gene, sondern konnte ebenso durch viele weitere Sequenzmerkmale wie CpG-Inseln oder repetitive Cluster verifiziert werden.

Da in den genannten Vorarbeiten lediglich die Lage der humanen Gene *LMO1* (telomerwärts) und *WEE1* (centromerwärts) zueinander durch FISH gezeigt wurde (Abb. 7A), stand die Überprüfung für das Genom der Maus noch aus. Erst mit Hilfe der in dieser Arbeit beschriebenen murinen Klone um die Gene *Lmo1* und *Tub*, konnte die anfangs postulierte invertierte chromosomale Anordnung des gesamten Bereichs in der Maus bis zum Gen *Wee1* mittels Fluoreszenz-in-situ-Hybridisierung bestätigt werden. Hierfür wurde der BAC-Klon 287P4, der positiv für das am weitesten distal gelegene Gen *Tub* ist, mit Digoxigenin grün fluoreszierend markiert. Die relative Lage zum Gen *Wee1* wurde durch den Biotin-markierten, rot fluoreszierenden PAC-Klon 256N10 ermittelt. Dieser Klon wurde im Rahmen der Dissertation von Andrea Cichutek (2001) näher charakterisiert. Die nachfolgende Abb. 7B zeigt in der Vergrößerung trotz der kompakten Form der telozentrischen Mauschromosomen die Reihenfolge TEL – *Wee1* – *Lmo1* – CEN, die somit invertiert zum humanen Vergleichschromosom vorliegt.

A) **B)**

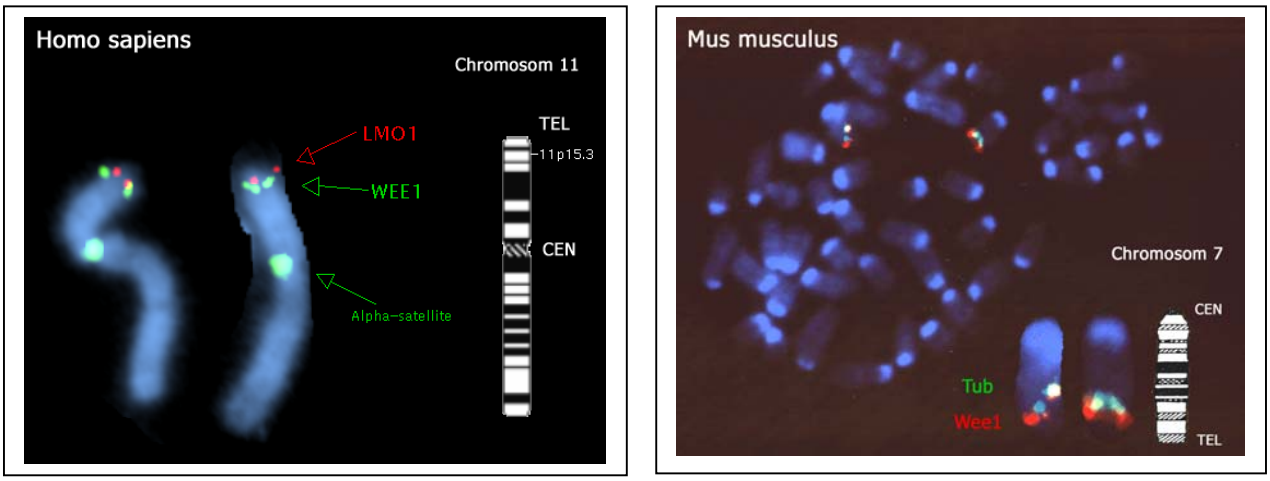


Abb. 7 **Zwei-Farben Fluoreszenz-in-situ-Hybridisierung** mit Gensonden für das distale und proximale Ende der auf eine Megabase geschätzten Region in Mensch und Maus. **A:** Als Sonde für das humane Chromosom 11 dienten die Klone PAC 12G13 (LMO1: → rot markiert) und PAC 142M6 (WEE1: → grün markiert). Die Identifizierung des Chromosoms wurde durch einen Chromosom-11-spezifischen Alpha-Satellitenmarker vorgenommen (Seipel, 1996). **B**: Der Nachweis für das telozentrische Maus-Chromosom 7 wurde mit den Klonen BAC 287P4 (=Tub: → grün markiert) und PAC 256N10 (=Wee1: → rot markiert) erbracht. Die umgedrehte chromosomale Abfolge dieser beiden Gene lässt vermuten, dass der gesamte Bereich in der Maus im Vergleich zum Menschen invertiert vorliegt.

3.2 Kartierung der zu sequenzierenden Region

3.2.1 Die humane Chromosomenregion 11p15.3

3.2.1.1 Verifizierung des humanen PAC-Klons 12G13

Zu Beginn der Arbeit wurde der humane PAC-Klon 12G13, der bereits in Vorarbeiten für die Fluoreszenz-in-situ-Hybridisierungs-Experimente auf Chromosom 11 von Barbara Seipel (1996) verwendet worden war, durch verschiedene Untersuchungen in seiner chromosomalen Lage am distalen Ende der Gesamtregion verifiziert. Mit Hilfe der Southern-Hybridisierung konnte eine eindeutige Zuordnung zum Markergen *LMO1* vorgenommen werden. Als Sonde diente das Exon 2 des *LMO1*-Gens, welches durch PCR mit den Primern *RBTN-EX2-F* und *RBTN-EX2-R* aus gesamtgenomischer DNA generiert und zur Amplifikatüberprüfung anschließend sequenziert worden war. Das 206 bp lange Fragment wurde radioaktiv markiert und zusammen mit einem Blot aus gelelektrophoretisch aufgetrennter, verschieden restringierter PAC 12G13-DNA hybridisiert. Pro Verdau zeigte sich nur eine Bande, jeweils auf dem Restriktionsfragment, das im Besitz der DNA des *LMO1*-Exons war (siehe Abb. 8). Da sich in der Exonsequenz keine Schnittstellen der verwendeten Restriktionsenzyme befanden, durfte durch die Hybridisierung auch nur eine Bande pro Verdau markiert werden.

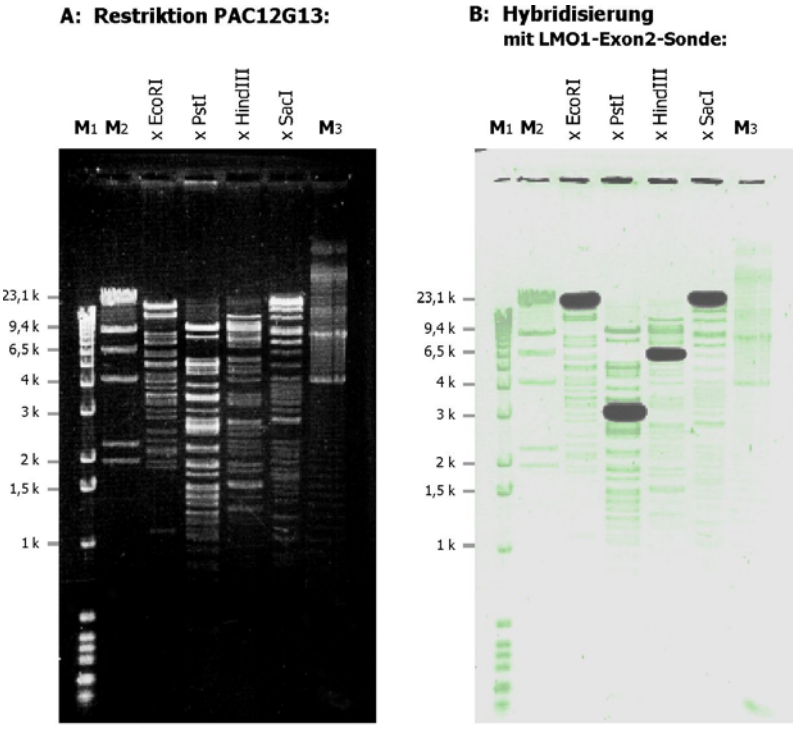


Abb. 8 Verifizierung des PAC-Klons 12G13 mittels Southern-Hybridisierung.

A: PAC12G13-DNA wurde mit den Restriktionsenzymen Eco RI, Pst I, Hind III und Sac I verdaut und gelelektrophoretisch aufgetrennt.
B: Anschließend erfolgte die Hybridisierung der geblotteten PAC12G13-DNA mit einer radioaktiv-markierten DNA-Sonde, bestehend aus dem Exon 2 des humanen *LMO1*-Gens. Je Restriktion ließ sich nur eine Bande spezifisch markieren. Um die Höhe der markierten Banden zu bestimmen, wurde dem Röntgenfilm das Restriktionsbandenmuster (grün eingefärbt) von Gel A bildtechnisch unterlegt!

M_1=100bp-Leiter;
M_2=λxHind III;
M_3=123bp-Leiter

3.2.1.2 Isolierung und Charakterisierung von humanen Anschlussklonen proximal und distal zum Gen LMO1

Als Ausgangspunkt für die weitere Kartierung der zu sequenzierenden humanen Region, diente das genomische Integrat des PAC-Klons 12G13, das nach Not I-Restriktion auf ca. 175 kb geschätzt worden war. Die Suche nach Contig-erweiternden Anschlussklonen geschah über PCR-generierte Hybridisierungssonden aus den Randbereichen der genomischen Sequenz des PAC 12G13. Da zu Beginn der Arbeit keine Aussage über die Centromer/Telomer-Orientierung des Klons gemacht werden konnte, wurden Anschlussklone für beide Flanken ermittelt.

Mit Hilfe der radioaktiv-markierten Sonde *12G13-08* aus dem mit dem Primer *PAC2808* (siehe Kap. 2.18.5) sequenzierten PAC-Randbereich wurden die Koloniefilter der genomischen PAC-Bibliothek #704 (siehe Kap. 2.18.4) hybridisiert und die PAC-Klone 1174K15 und 1184A15 detektiert. Bei näherer Charakterisierung dieser Klone durch den Vergleich der Restriktionsbandenmuster stellten sich aber große Überlappungsbereiche mit dem Ausgangs-PAC 12G13 heraus (75% bzw. 90%, siehe auch Abb. 10), so dass sie für eine Sequenzierung nicht in Frage kamen. Stattdessen wurden die neuen Randsequenzen dieser Klone, die nicht mit dem PAC 12G13-Enden überlappten, zur nochmaligen Sondengenerierung herangezogen und für eine weitere Hybridisierung eingesetzt. In paralleler Hybridisierung mit den Sonden *1174K15-28* + *1184A15-28* konnten fünf neue Klone ermittelt und verifiziert werden (PAC 781K3, PAC 597N1, PAC 892B14, PAC 908N14 und PAC 1061A8). In der sich anschließenden Not I-Restriktionsfragmentanalyse und Endsequenzierung der PAC-Klone zeigte sich

der Klon PAC 781K3 im Vergleich zu den anderen Klonen mit einer Integratgröße von 130 kb und einer Überlappung von 21,3 kb als am günstigsten gelegen. Er wurde daher im Rahmen dieser Arbeit subkloniert und komplett sequenziert.

Für das gegenüberliegende Ende des PAC 12G13 konnten mit Hilfe der Sonde *12G13-28* (siehe Kap. 2.18.5) durch Koloniefilter-Hybridisierung die PAC-Klone 2M19, 2N19, 36K14, 40M9 und 75C13 detektiert werden. Der Vergleich der Überlappungsbereiche zeigte für diese Klone allerdings keine günstige genomische Lage, da sie mit mindestens 50% zum Ausgangsklon PAC 12G13 identisch waren. Daher wurde die Anschlußklonsuche mit einer Chromosom 11-spezifischen Cosmidbibliothek fortgesetzt. Diese Cosmidbank stand zweidimensional angeordnet in Form von Spalten- und Reihen-Pools zur Verfügung und erlaubte ein schnelles Durchsuchen der Bibliothek mittels PCR. Mit den Primern für die Sonde *12G13-28* wurden die Cosmide cSRL14c7, cSRL53f6, cSRL99b8, cSRL119g5 und cSRL174d1 isoliert. Nach Größen- und Lagebestimmung wurde der Klon cSRL119g5 ausgewählt, da er mit nur 6,7 kb Sequenzüberlappung und einem genomischen Integrat von 41 kb den größten Hinzugewinn an neuer genomischer Baseninformation bot. Dieser Klon wurde daraufhin von Silke Schlaubitz (2000) subkloniert und sequenziert.

Eine weitere Suche nach Anschlussklonen wurde mit der Sonde *119g5-T7* aus dem distalen Ende von Cosmid cSRL119g5 durchgeführt. Die Hybridisierung ergab Signale für die PAC-Klone 985F16, 474K21, 830K2 und 910M1. Die Randsequenzierung und die Not I-Fragment-Größenbestimmung ließ die Anordnung im Gesamtcontig zu und erweiterte diesen auf insgesamt 383 Kilobasen.

Um die Orientierung zweier Contigsequenzen innerhalb der PAC 12G13 Sequenz zu bestimmen, wurde Hilfe des STS-Markers *D11S572* fünf weitere Cosmidklone cSRL15H10, cSRL63d2, cSRL102g5, cSRL153B11 und cSRL155e6 identifiziert (siehe auch Abb. 10).

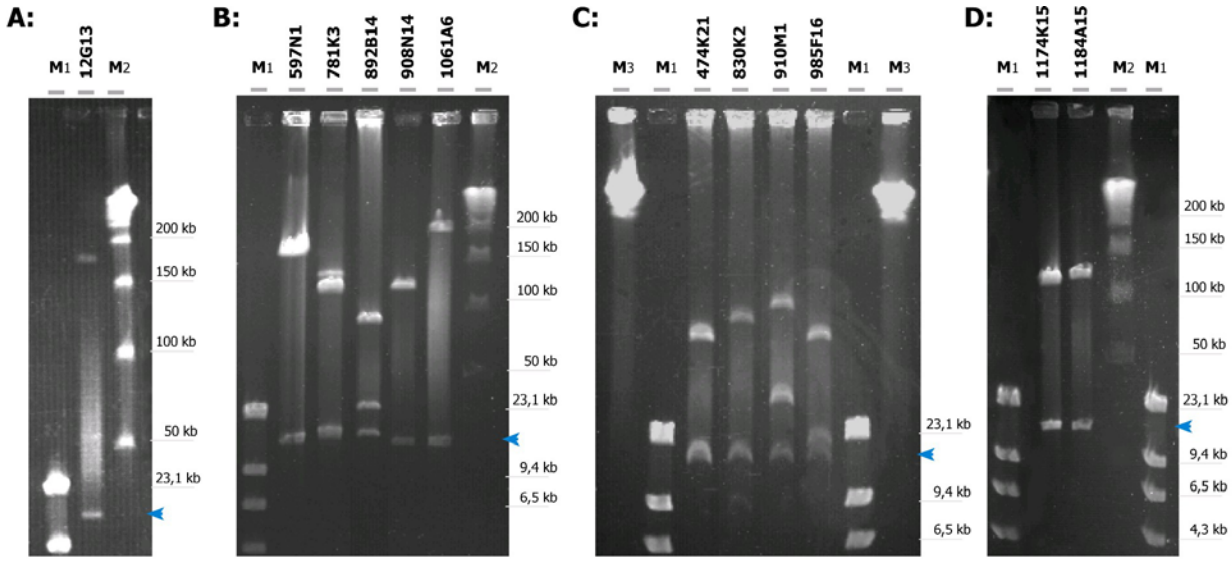


Abb. 9 Pulsfeld-gelelektrophoretisch aufgetrennte, Not I-restringierte PAC-Klone. Zur Größenabschätzung der genomischen Integrate wurden die verifizierten PAC-Klone mit Not I restringiert, um das genomische Integrat vom Vektor zu isolieren und die genomische Fragmentgröße zu bestimmen. Als Längenstandard wurde Hind III restringierte λ-DNA (M1) und λ-DNA-Concatemere (M2), bzw. ein Hefe-Chromosomen-PFGE-Marker (M3) verwendet. Der blaue Pfeil gibt die Höhe der Vektorbande (pCYPAC2) bei 18,7 kb an (Ioannou et al., 1994). A: Ausgangsklon 12G13 mit geschätzter Größe von 175 kb. B: Mit den Sonden 1174K15-28 und 1184A15-28 wurden die PAC-Klone 597N1 = 150 kb; 781K3 = 131 kb (Doppelbande! obere durch unvollständigen Verdau); 892B14 = 110 kb; 908N14 = 115 kb und 1061A6 = 190 kb. isoliert. C: Die PAC-Klone 474K21 = 75 kb; 830K2 = 104 kb; 910M1 = 139 kb und 985F16 = 92 kb wurden durch die Sonde 119g5-T7 identifiziert. D: Durch die Sonde 12G13-08 konnten die PAC-Klone 1174K15 = 120 kb und 1184A15 = 125 kb bestimmt werden.

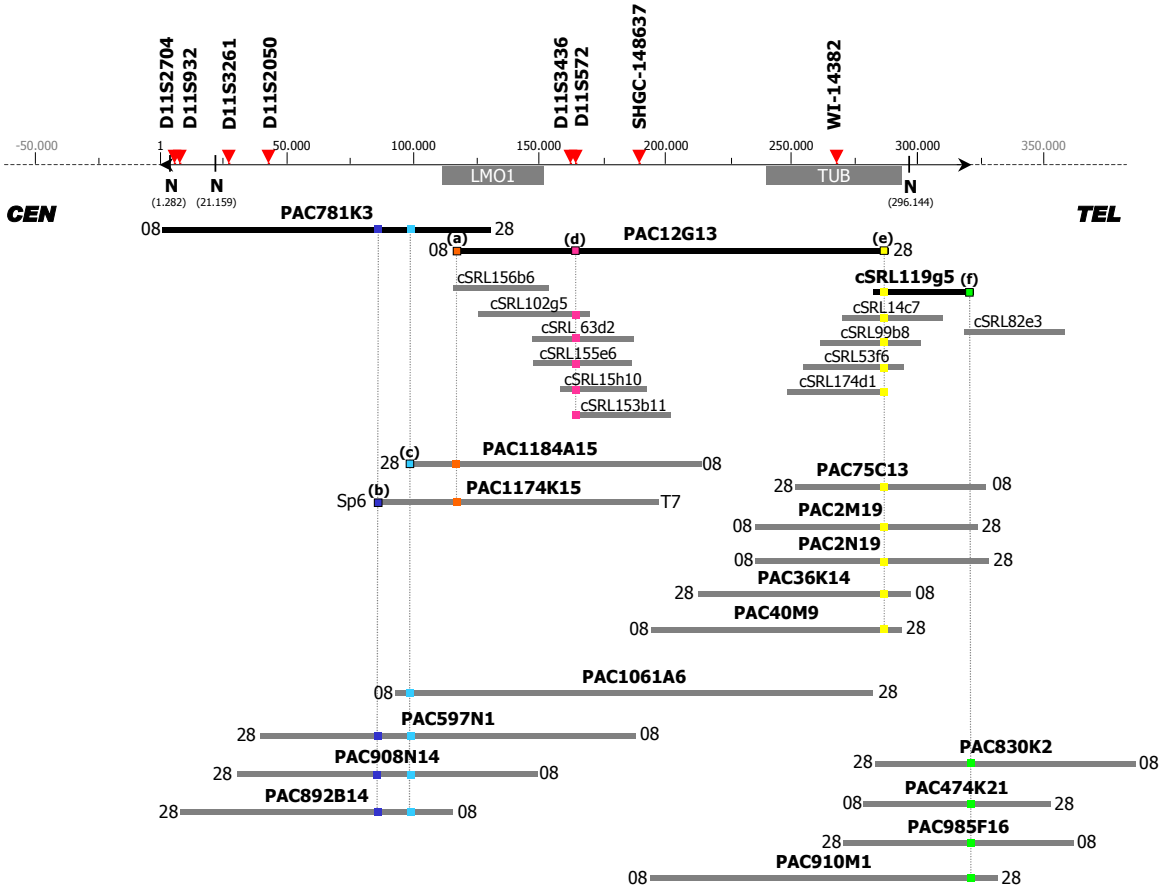


Abb. 10 Klon-Contig mit 29 verifizierten Klonen der zu charakterisierenden genomischen Region des Menschen. Die in dieser Arbeit vollständig sequenzierten Klone PAC 781K3, PAC 12G13 und das Cosmid cSRL 119g5 sind mit einem schwarzen Balken hervorgehoben und repräsentieren einen genomischen DNA-Bereich von ca. 320 kb. Die flankierenden Enden dieser Klone überlappen dabei um 21 kb (PAC 781K3 zu PAC 12G13), bzw. 6 kb (PAC 12G13 zu cSRL 119g5) zueinander. Ausgehend von PAC 12G13 wurden centromerwärts Anschlussklone mit Hilfe der Sonden 12G13-08 (a) und 1174K15-28 (b) + 1184A15-28 (c) ermittelt. Richtung Telomer dienten die Sonden 12G13-28 (e) und 119g5-T7 (f) als Identifizierungsmarken für neue Klone. Für den STS-Marker D11S572 (d) konnten fünf Cosmidklone zur Verifizierung eines internen Bereichs charakterisiert werden. Insgesamt konnten in der untersuchten Region acht verschiedene STS-Marker kartiert werden. „Ns" geben Not I-Restriktionsschnittstellen in der genomischen Sequenz wieder. Zur weiteren Orientierung wurden ebenfalls schematisch die unter Kap. 3.6 näher beschriebenen Gene *LMO1* und *TUB* gemäß ihrer Lage im Contig eingezeichnet.

3.2.2 Die orthologe murine Chromosomenregion proximal und distal zum Mausgen Lmo1

Für den orthologen genomischen Bereich auf Chromosom 7 der Maus existierten aus den eingangs beschriebenen Vorarbeiten keinerlei Ausgangsklone. Mit Hilfe der aus der humanen Sequenz bekannten murinen Markergene *Lmo1* und *Tub* wurde nach PAC-Klonen für diese Region gesucht. In einer gleichzeitig durchgeführten Literaturrecherche fiel eine Veröffentlichung von Kleyn & Mitarbeitern (1996) auf. Kleyn versuchte in einem positionellen Klonierungsansatz den Maus-Locus des *Tub*-Gens physikalisch zu kartieren und stellte in seiner Arbeit einen Klon-Contig vor, der den zu charakterisierenden Bereich in Richtung Centromer bis zum Gen *ST5* abdeckte. Die angeordneten

BAC-Klone wurden bei RESEARCH GENETICS (USA) bestellt und mit verschiedenen STS-Sonden aus dieser Region, veröffentlicht vom WHITEHEAD INSTITUTE/MIT (Database Release, 1998) D7Mit127, D7Mit219, D7Mit280 und D7Mit325, bzw. von MILLENNIUM PHARMACEUTICALS (Kleyn *et al.*, 1996) D7Mill1 und D7Mill2, über PCR auf die Richtigkeit ihrer Lage hin geprüft. Die Verifizierung und Endsequenzierung der bestellten Klone zeigte, dass der BAC-Klon 287P4 große Teile der Sequenz des *Tub*-Gens beinhaltet und positiv ist für die STS-Marker D7Mit127 und D7Mill1. Mit einer Größe von 163 kb an genomischer Sequenz wurde er im Rahmen dieser Arbeit subkloniert und komplett sequenziert. Ein zweiter BAC-Klon mit der Bezeichnung 282L1 beinhaltete das komplette Gen *Lmo1*, wie auch die STS-Marker D7Mit280 und D7Mit325. Auch dieser Klon wurde im Rahmen dieser Arbeit subkloniert und komplett sequenziert. Die zueinanderweisenden Enden der Maus-BAC-Klone 287P4 und 282L1 zeigten im Vergleich zur humangenomischen Sequenz einen geschätzten Abstand von ca. 83 kb zueinander. Über Koloniefilter-Hybridisierung der murinen PAC-Bibliothek #711 wurden mit Hilfe des *Lmo1*-Exons 4 als Sonde, generiert mit den Primern *mLmo1_ex4-f* und *mLmo1_ex4-r*, vier neue PAC-Klone (194L17, 368C2, 420K8 und 504A8) identifiziert, die diese Lücke überspannen. Nach Bestimmung der genauen Lage der vier Klone im Contig, wurde der PAC 368C2 ausgewählt und zum Sequenzieren subkloniert, da er im Vergleich zu den anderen Klonen mit 58 kb den geringsten Überlappungsbereich zu den flankierenden BAC-Klonen aufwies. Die Überschneidung zur BAC 282L1-Sequenz betrug dabei 48.924 bp und zur BAC 287P4-Sequenz 9.144 bp.

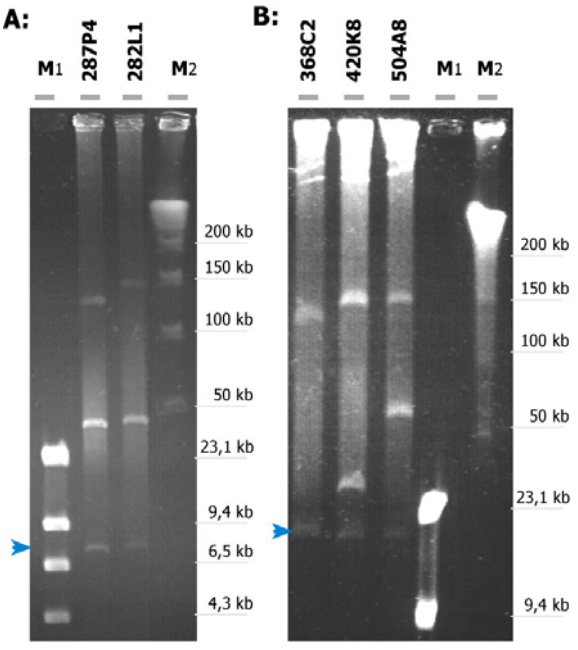


Abb. 11 Pulsfeld-Gelelektrophorese der murinen Not I- BAC bzw. PAC-Klonen nach Not I-Restriktion. Als Längenstandard wurde Hind III restringierte λ-DNA und ein λ-DNA-Concatemer verwendet. Der blaue Pfeil gibt die Höhe der Vektorbande bei 7,4 kb (pBeloBAC11), bzw. bei 18,7 kb (pCYPAC2) an. A: Die Klone BAC 287P4 (124k+40k=163kb) und BAC 282L1 (142k+43k= 185kb) besitzen beide eine interne Not I-Schnittstelle, so dass die genomische Sequenz des Integrats in zwei Fragmente zerfällt. B: Die Klone PAC 368C2 (130k+5,7k=136kb: die kleine 5,7 k Bande ist nicht mehr im Gel zu sehen, konnte aber getrennt nachgewiesen werden**),** PAC 420K8 (150k+30k= 180kb) und PAC 504A8 (150k+60k=210kb) überspannen die Lücke zwischen den beiden BAC-Klonen.

Marker: M_1=λxHind III, M_2=λ-DNA-Concatemere.

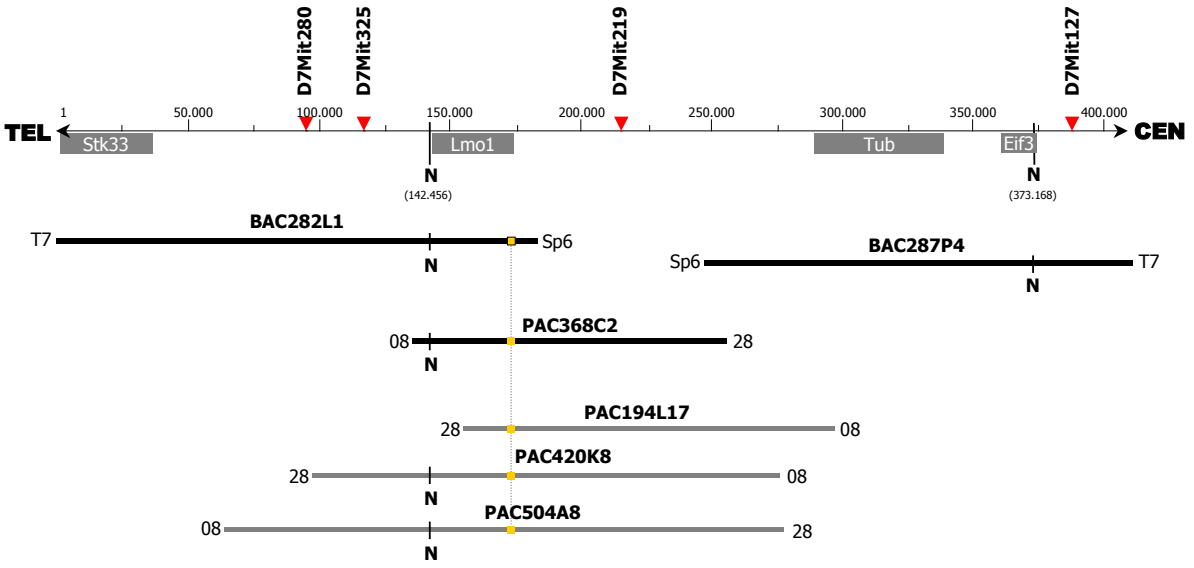


Abb. 12 **Klon-Contig der zu charakterisierenden genomischen Region der Maus**. Die schwarz unterlegten Klone BAC 282L1, BAC 287P4 und PAC 368C2 wurden in dieser Arbeit sequenziert. Mit Hilfe der Sonde Lmo1-Exons 4 (gelb) konnten vier Klone identifiziert werden, die die auf ca. 83 kb geschätzte Lücke zwischen den BAC-Klonen 282L1 und 287P4 überspannen. Da der Klon PAC 368C2 mit insgesamt 58 kb den geringsten Überlappungsbereich zu den benachbarten BAC-Klonen auswies, wurde dieser für die Sequenzierung weiter genutzt. Für diesem Contigbereich konnten insgesamt vier STS-Marker (D7Mit280, D7Mit325, D7Mit219 und D7Mit127; Whitehead Institute/MIT) lokalisiert werden. Der Großbuchstabe „N" gibt Not I-Restriktionsschnittstellen in der genomischen Sequenz wieder.

3.3 Subklonierung der ausgewählten Klone

Für die Sequenzierung der ausgewählten BAC-, PAC- und Cosmid-Klonen mussten diese in einen speziellen Sequenzierungsvektor (pUC18) subkloniert und in Klon-Bibliotheken angeordnet werden. Um den Anteil an linearer Einzel- und Doppelstrang-DNA, der eine Kontamination an bakterieller genomischer DNA in sich bringt, für die Klonierung weitestgehend zu reduzieren, wurde die DNA nach ihrer Präparation einer DNase-Behandlung unterzogen, in der nicht zirkuläre DNA hydrolysiert wurde. (siehe auch Kap. 2.8.1). In Abbildung 13A ist dieser Effekt exemplarisch an isolierter hochmolekularer PAC-DNA und an linearisierter pUC18-Vektor-DNA dargestellt. Die in der Abbildung zu sehende untere breite DNA-Bande der PAC-DNA konnte nach Abschluss der DNase-Behandlung entfernt werden. Noch eindeutiger ließ sich dieser Effekt mit restringierter und somit linearisierter Vektor-DNA demonstrieren. Hier verschwand die Bande der vektoriellen DNA vollständig.

Nach Nebulisierung der zu klonierenden DNA (siehe auch Kap. 2.8.2) wurde diese gefällt und gelelektrophoretisch aufgetrennt. Mit diesem Verfahren konnte die DNA in Fragmentgrößen von durchschnittlich 500 bis 5.000 bp gebracht werden (siehe Abb. 13B). Um die nebulisierte DNA in unterschiedlichen Fraktionen zu klonieren, wurden verschiedene Fragmentgrößenbereichen aus dem

Agarosegel ausgeschnitten und die DNA hieraus getrennt isoliert. Mittels eines Aliquots wurden die wiedergewonnenen DNA nach nochmaliger gelelektrophoretischer Auftrennung auf Ihre tatsächliche Fragmentgröße hin kontrolliert (siehe Abb. 13C).

Nach dem „Endfilling" und der Kinasierung folgte die Ligation in den dephosphorylierten Sequenzierungsvektor pUC18. Die Phosphorylierung der zu ligierenden genomischen Fragmente erhöhte dabei die Ligationseffizient, was sich später in der hohen Zahl an Subklonen mit gewünschter Fragmentgröße zeigte. Zur Kontrolle des Transformationsergebnisses und zur Überprüfung von chimären Subklonen mit concatemeren Integraten wurden die in 96-Loch-Mikrotiterplatten angeordneten Subklone einer PCR mit M13-Primer unterzogen, die zur Amplifikation des genomische Integrate im Vektor führte (Abb. 14). Etwaige Klonierungsartefakte wären bei dieser Überprüfung durch unerwartet große Amplifikatprodukte aufgefallen.

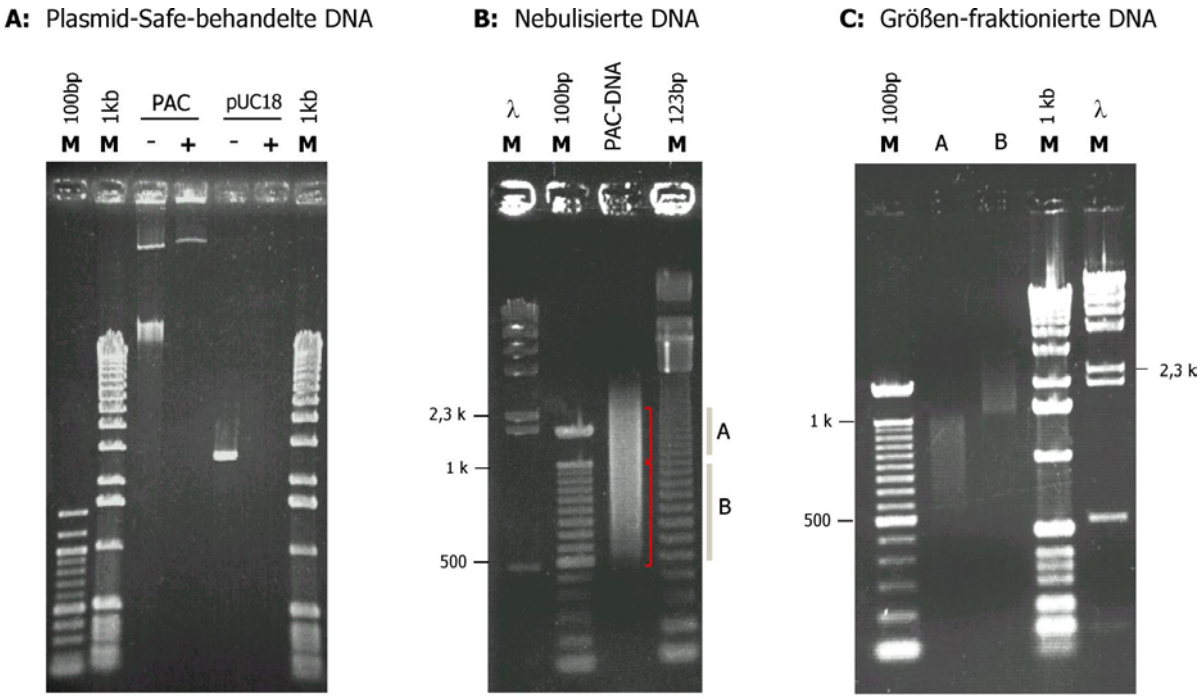


Abb. 13 **Vorbereitung der zu klonierenden DNA. A:** Die „Plasmid-Safe"-Behandlung führte zu einem Entfernen der linearisierter DNA. Beispielhaft konnte dieser Effekt mit PAC-DNA vor (-) und nach (+) der Behandlung gezeigt werden. Nach DNase-Verdau blieb nur die hochmolekulare, „supercoiled" PAC-DNA übrig. Die Kontrolle mit Eco RI-restringierter pUC18-DNA zeigte ebenso ein vollständiges Verschwinden der linearisierten Vektor-DNA. **B:** Die durch das Nebulisieren gescherte PAC-DNA wurde auf einem präparativen Agarosegel aufgetrennt. Der sich zeigende „Schmier" stellte DNA-Fragmente in der Größe von 500 bp bis ca. 5 kb dar. Zwei Fragmentgrößenbereiche (A: 500 bp-1.000 bp + B: 1,0 kb – 2,5 kb) wurden auf dem Gel ausgeschnitten und die darin enthaltene DNA wiedergewonnen. C: Die Kontrolle zeigt die DNA der beiden Fraktionen A und B, die zur Ligation und Transformation weiterverwendet wurde.

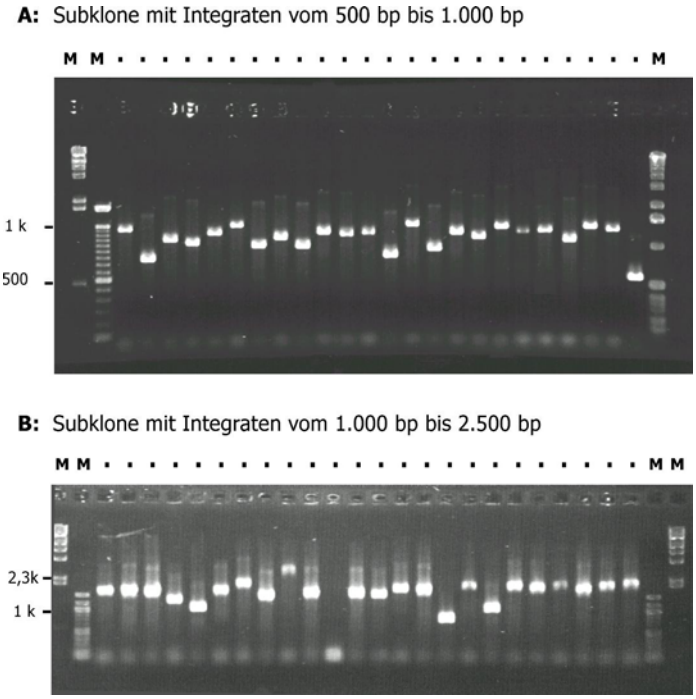


Abb. 14 Transformationskontrolle nach PCR-Amplifikation mit M13-Primern der in 96er Format angeordneten pUC18-Subklone A: Gelelektrophoretisch aufgetrennte PCR-Amplifikate der Proben, die mit genomischen Fragmenten der Größe von 500 bp bis 1.000 bp (Fraktion A) transformiert wurden. B: PCR-Produkte der zweiten Fraktion B mit Fragmentgrößen von 1,0 kb bis 2,5 kb. Alle PCR-Amplifikate liegen in dem gewünschten Fragmentgrößenbereich. Die präparierte DNA aus diesen Klonen wurde anschließend der Hoch-Durchsatz-Sequenzierung zugeführt.

3.4 Sequenzierung der ausgewählten Klone

Die Sequenzierungsstrategie war bei allen Klonen eine kombinierte Vorgehensweise aus „Shotgun"-Sequenzierung der in 96-Loch-Mikrotiterplatten angeordneten Subklone mittels pUC18-Vektor-spezifischer M13-forward/revers-Primern und einer „Primer walking"-Strategie in die verbliebenen Contiglücken hinein mit Hilfe spezifischer Primer aus den flankierenden Randbereichen der Contigs.

In der ersten „Shotgun"-Sequenzierungsphase wurden überwiegend Subklone mit großen Integraten (1 – 2 kb, bzw. 1,5 – 2,5 kb) von beiden Enden her sequenziert. So wurde bei einer Leseweite von 500 bp bis 600 bp möglichst schnell, viel genomische Sequenzinformation ermittelt, ohne dabei zu viel Redundanz durch sich überlappende Sequenzbereiche zu erzeugen. Die Subklone mit kleineren Integratgrößen (600 – 1.000 bp) wurden in einer ersten Runde nur einseitig sequenziert. Die Sequenzierung des Gegenstrang fand nur noch vereinzelt gezielt statt, je nach Lage des Subklons im Gesamtcontig.

3.4.1 Die humanen Klone

Für die Generierung der humangenomischen Sequenz wurden die beiden PAC-Klone 12G13 und 781K3 und das Cosmid cSRL119g5 sequenziert.

Tab. 12 Sequenzierstatistik der sequenzierten humanen PAC-Klone 12G13 und 781K3, und des Cosmids cSRL 119g5. Es werden miteinander die verwendete Assemblierungs-Software, die Zahl der angeordneten Klone, die Größe des genomischen Integrats, die Zahl der Einzelsequenzierungen, die Zahl der verwendeten Oligonukleotid-Sequenzen und die durchschnittliche Redundanz verglichen. Als Wert für die Redundanz der sequenzierten DNA der Klone PAC781K3 und cSRL119g5 konnte nur ein grober Wert, berechnet aus der Zahl der im Contig assemblierten Sequenzen mit einer durchschnittlichen Leseweite von 500 bp, angegeben werden, da das Programm predPhrap keine Statistik über die Zahl der in die Consensussequenz einfließenden Basenpaaren besitzt. Der Cosmidklon cSRL119g5 wurde während der „Shotgun"-Phase mit dem Programm phredPhrap assembliert und während der abschließenden „Finishing"-Phase mit dem Programm Sequencher weiterbearbeitet.

Sequenzierter Klon	**PAC 12G13**	**PAC 781K3**	**cSRL 119g5**
offizielle Klonbezeichnung	**RPCIP704G1312**	***RPCIP704K03781***	***LANLc114G05119***
verwendete Assemblierungs-Software	Sequencher	phredPhrap	phredPhrap + Sequencher
gepickte Klone	20 x 96 Klone = 1.920	20 x 96 Klone = 1.920	6 x 96 Klone = 576
Größe des genom. Integrats	**174.986 bp**	**131.145 bp**	**40.996 bp**
Zahl der Sequenzierungen	1.776 „reads"	2.205 „reads"	725 „reads"
Zahl der Primer	205	143	121
Durchschnittliche Redundanz bei einer Leseweite von 500 bp	5,1	8,4	8,8

Die genomische Sequenz des PAC 12G13 wurde ausschließlich mit Hilfe des Programms *SEQUENCHER* ermittelt. Das Editieren der Sequenzen geschah weitestgehend durch visuelle Betrachtung des Elektropherogramms mit Hilfe der *SEQUENCING ANALYSIS SOFTWARE*. „Base calling"-Fehler konnten durch diese Bearbeitung manuell korrigiert werden. Diesem Umstand verdankt das Projekt (im Vergleich zu den anderen Sequenzierprojekten) seine geringere Zahl an 1.776 importierten Einzelfragmenten, da aufgrund der visuellen Kontrolle nur informative Sequenzen aufgenommen wurden. Die vergleichsweise große Zahl an Primern (205 Oligonukleotide) für des Sequenz-„Finishing" macht indes deutlich, dass die ermittelte 174.986 bp lange Consensus-Sequenz nur in einer geringen Redundanz (5,1) vorlag und mehr Bereiche über Primeramplifikation verifiziert werden mussten.

Das genomische Integrat des PAC 781K3 mit einer Länge von 131.145 bp konnte durch Sequenzierung von 2.205 Einzelsequenzen bestimmt und mittels 143 Primern verifiziert werden. Als Assemblierungssoftware wurde mit dem Programm *PHREDPHRAP* gearbeitet, welches ein eigenes „Basecalling" (siehe auch Kapitel 4.1.3) vornahm und das Assemblieren über zuvor generierte Qualitätswerte für jede Base der Sequenz vollzog („quality value"-Zuweisung). Auf diese Weise konnten alle informativen Rohdaten mit in die Assemblierung eingebracht werden, so dass sich die durchschnittliche Redundanz der generierten DNA-Sequenz auf 8,4 erhöhte.

Die genomische Sequenz des Cosmids cSRL119g5 wurde durch Anordnung von 725 Einzelsequenzen zu einer Länge von 40.996 bp erreicht. Bis zu einem Stadium von fünf Teilcontigs wurde mit dem Programm *PHREDPHRAP* gearbeitet. Auch das „Finishing" der DNA-Sequenz dieser fünf Contigs wurde

mit der *PHREDPHRAP*-Software durchgeführt. Lediglich der Lückenschluss geschah mit dem Programm *SEQUENCHER* über „Primer walking". Insgesamt wurden für die Fertigstellung der annotierten Sequenz 121 Primern eingesetzt. Auch hier zeigte sich mit dem Wert 8,8 eine ähnliche hohe Redundanz wie bei dem PAC-Klon 781K3.

3.4.2 Die sequenzierten murinen Klone

Für die Darstellung des Maus-genomischen Bereiches wurden ebenfalls drei Klone – die BACs 287P4 und 282L1 und der PAC-Klon 368C2 – sequenziert.

Tab. 13 Sequenzierstatistik der sequenzierten murinen BAC-Klone 287P4 und 282L1 und des PAC-Klons 368C2. Es werden miteinander die verwendete Assemblierungs-Software, die Zahl der angeordneten Klone, die Größe des genomischen Integrats, die Zahl der Einzelsequenzierungen, die Zahl der verwendeten Oligonukleotid-Sequenzen und die durchschnittliche Redundanz verglichen. Als Wert für die Redundanz der sequenzierten DNA der Klone PAC 781K3 und cSRL 119g5 konnte nur ein grober Wert, berechnet aus der Zahl der im Contig assemblierten Sequenzen mit einer durchschnittlichen Leseweite von 500 bp, angegeben werden, da das Programm phredPhrap keine Statistik über die Zahl der in die Consensussequenz einfließenden Basenpaaren besitzt.

Sequenzierter Klon	**BAC 287P4**	**BAC 282L1**	**PAC 368C2**
offizielle Klonbezeichnung	***RPCIB731P04287***	***RPCIB731L01282***	***RPCIP711C02368***
verwendete Assemblierungs-Software	phredPhrap	phredPhrap	phredPhrap
gepickte Klone	24 x 96 Klone = 2.304	20 x 96 Klone = 1.920	8 x 96 Klone = 768
Größe des genom. Integrats	**163.456 bp**	**185.688 bp**	**~ 135 kb**
Zahl der Sequenzierungen	2.955 „reads"	2.324 „reads"	827 „reads"
Zahl der Primer	123	121	69
Durchschnittliche Redundanz bei einer Leseweite von 500 bp	9,0	6,2	3,1

Die DNA-Sequenz aller genomischen Integrate, der drei o.g. Klone wurden mit Hilfe des Programms *PHREDPHRAP* generiert; lediglich das Aneinanderfügen der drei Klone zur fertigen Consensussequenzen geschah abschließend mit Hilfe des *SEQUENCHER*-Programms.

Die genomische Sequenz des Maus-BAC-Klons 287P4 wurde durch insgesamt 2.955 Einzelsequenzen und 123 Primern mit einer Länge von 163.456 bp angegeben. Da der repetitive Anteil an LINE-Sequenzen mit 15% vergleichsweise hoch lag – der Anteil im BAC 282L1 betrug nur 8,5% – mussten mehr Sequenzierungen durchgeführt werden, als beim BAC 282L1, so dass die berechnete Redundanz einem Wert von 9 ergab. Für die Sequenzierung des Maus-BAC-Klons 282L1 wurde 2.324 Einzelsequenzen und 121 Primer gebraucht, um den genomischen Anteil von 185.688 Basenpaaren zu beschreiben.

Der dritte auf 135 kb geschätzte Maus-Klon PAC 368C2 zeigte zu den flankierenden BAC-Klonen eine Überlappung von 58.069 bp, was einer Überschneidung von über 42% entsprach. Es wurde daher versucht, die in Mikrotiterplatten angeordneten Subklone mit Hilfe einer Subtraktionshybridisierung in ihrer Zahl auf diejenigen zu beschränken, die den unbekannten Abschnitt zwischen den beiden BAC-Klonen repräsentieren, um so die Zahl der nötigen Sequenzierungen zu reduzieren (siehe Kap. 2.11). Diese Selektion an Subklonen und die Anordnung in einer zweiten Subtraktionsbank hatte zum Erfolg, dass von den ausgewählten sequenzierten Klonen 75% mit ihrer genomischen Sequenzen für den unbekannten, knapp 77 kb großen Bereich zwischen den beiden BAC-Klonen kodierten. Im Verhältnis zum übrigen Sequenzanteil des PAC 368C2, der sich aus 58 kb überlappenden Anteil mit den flankierenden BAC-Klonen und aus 18,7 kb PAC-Vektoranteil zusammensetzt, erbrachte die Subtraktionshybridisierung eine Effizienzsteigerung von 25%. Da ohne die Subtraktion statistisch nur die Hälfte der Klone (klonierte DNA = 135 kb genom. Integrat + 18,7 kb = 153,7 kb : 77 kb DNA der Sequenzlücke = 50,09%) mit Sequenzinformation für die Sequenzlücke geliefert hätten. Somit konnte die Zahl der informativen Subklone für den unbekannten genomischen Bereich um ein Viertel erhöht werden. Dass dieser Wert nicht größer ausgefallen ist, kann auf den relativ hohen Anteil an repetitiven Bereichen in der genomischen Sequenz zurückzuführen sein, der bei der Subtraktionshybridisierung zum einen durch Kreuzhybridisierung zu falsch Positiven, bzw. durch die Absättigung mit Cot I-DNA zum anderen zu einem Verlust von informativen Klonen durch nicht Nichtmarkierung führte.

Mit Hilfe von 615 Einzelsequenzen konnte für einen genomischen Bereich von 62.684 bp die genaue Basenabfolge bestimmt und verifiziert werden. Da dieser Bereich nicht mit den BAC-Klon-Sequenzen 282L1 und 287P4 überlappte, musste er für den unbekannten genomischen Abschnitt von 77 kb kodieren und deckte somit ca. 81% der zu bestimmenden Gesamtsequenz ab. Ein verbleibender Sequenzbereich von zusammengenommen ca. 15 kb (= 19%) konnte nicht eindeutig bestimmt werden, da für diesen Bereich weder Subklone existierten, noch der Versuch über eine „Primer-Walking"-Strategie, aufgrund von repetitiven Flanken der Lücken, zu keinen auswertbaren Ergebnissen führte. Da aber durch den Interspeziesvergleich mit der humanen Genomsequenz die einzelnen Contigsequenzen aufgrund ihrer Homologie zur humanen DNA eindeutig physikalisch angeordnet werden konnten und die fehlenden Bereiche keine kodierenden Abschnitte erwarten ließen, wurden die insgesamt zehn unbestimmten Bereiche von durchschnittlich ca. 1,5 kb in der Referenzsequenz für die spätere komparative Auswertung mit jeweils 100 „Ns" als Platzhalter versehen.

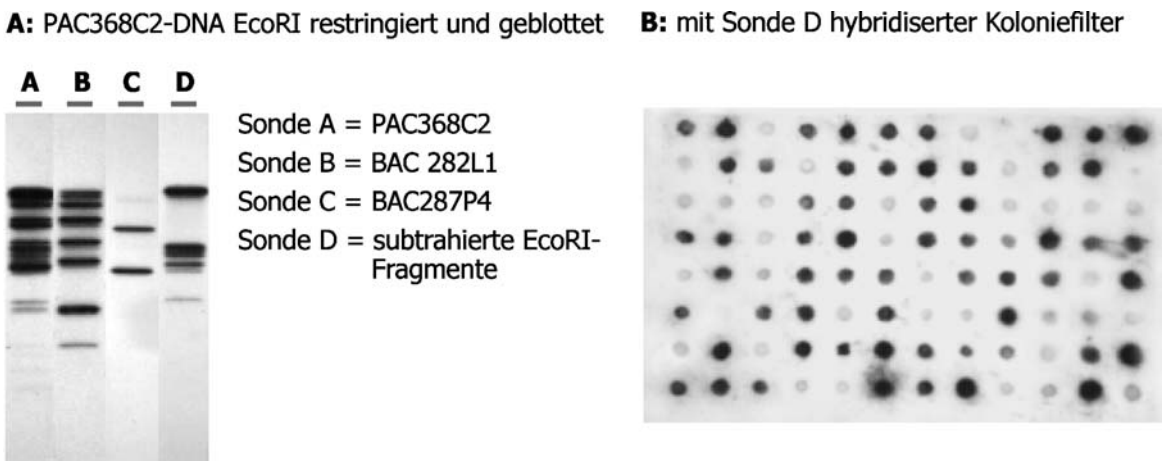


Abb. 15 **Erstellung der PAC 368C2 Subtraktionsgenbank** **A:** Darstellung der für die Hybridisierung eingesetzten Eco RI-Fragmenten des PAC 368C2. Aus dem Eco RI-Restriktionsmuster des PAC-Klons 368C2 (Spur A) wurden die Eco RI-Restriktionsbanden der BAC-Klone 282L1 (B) und 287P4 (C) subtrahiert, da sie den überlappenden Bereich mit dem PAC 368C2 darstellen. Die Spur D zeigt die Auswahl an DNA-Banden, die für die Hybridisierung radioaktiv markierten wurden und zum Detektieren als jener Subklone verwendet wurde, die nicht aus dem genomischen Bereich der BAC-Klone stammen. **B:** Exemplarische Abbildung eines mit der Sonde D hybridisierten Koloniefilters. Die Hybridisierung identifizierte alle Klone mit DNA aus der Lücke der beiden BACs. Diese Klone wurden gepickt und in 96er Platten für die Sequenzierung neu angeordnet.

3.4.3 Zusammenfassung aller sequenzierten Bereiche

Die Anordnung aller verifizierten Klone ergab für den humanen Genomabschnitt einen Gesamtcontig von ca. 383 Kilobasen, wovon 319.119 bp als durchgehende Sequenzinformation vorliegen - repräsentiert durch die Klone: *CEN* – PAC781K3 – PAC12G13 – cSRL119g5 – *TEL*. In diesen Bereich konnten zudem acht verschiedene STS-Marker kartiert und mit den genauen Abständen zueinander bestimmt werden. Aus Richtung Centromer beginnend, lassen sich diese wie folgt mit den jeweiligen Abständen anordnen: CENTROMER → D11S2704 → 1.411 bp → D11S932 → 20.879 bp → D11S3261 → 15.647 bp → D11S2050 → 118.833 bp → D11S572/D11S3436 → 14.420 bp → SHGC-148637 → 90.629 bp → WI-14382 → TELOMER. Zwischen den Markern D11S2050 und D11S572/D11S3436 ist das Gen *LMO1* lokalisiert. Die STS-Sonden D11S572 und D11S3436 liegen so dicht beieinander, dass sie sich einen Sequenzabschnitt von 20 bp gemeinsam teilen, der exakt der Primersequenz entspricht. Der Marker WI-14382 konnte in einem kodierenden Bereich des *TUB*-Gens (siehe Kap. 3.6.2.1) zum Liegen gebracht werden. Jede Consensussequenz dieser drei Klone wurde einzeln unter den Acc.-Nr. **AJ277661** (PAC 781K3), **AJ277661** (PAC 12G13) und **AJ296302** (cSRL 119g5) in der EMBL Nukleotidsequenz-Datenbank hinterlegt.

Im Zuge der vollständigen Sequenzierung konnten indirekt auch die Größe der genomischen Integrate der sechs Cosmidklone bestimmt werden. Demnach haben die genomischen Integrate dieser Cosmidklone folgende Größen: cSRL15H10 = 33.134 bp; cSRL63d2 = 35.563 bp; cSRL102g5 **=** 39.290 bp; cSRL153B11 = 39.218 bp; cSRL155e6 = 34.466 bp.

Für das Mauschromosom 7 konnten durch die drei ausgewählten Klonen in der Anordnung *TEL* – BAC282L1 – PAC368C2 – BAC287P4 – *CEN* ein Bereich von ca. 425 kb an muriner Genomsequenz sequenziert werden. Dieser Bereich umfasste vollständig den oben beschriebenen orthologen Genombereich auf Chromosom 11p15.3 des Menschen. Es konnten insgesamt vier STS-Marker kartiert werden, deren Lage zueinander sich wie folgt darstellte: TELOMER → D7Mit280 → 29.941bp → D7Mit325 → ~101kb → D7Mit219 → ~169kb → D7Mit127 → CENTROMER. Hier geben die Basenpaarangaben zwischen den STS-Marker-Bezeichnungen die Abstände zwischen den Markersonden wieder. Die Entfernung zwischen der Sonde D7Mit325 und D7Mit219, bzw. D7Mit219 und D7Mit127 konnte aufgrund der nicht durchgehend vorliegenden Genomsequenz nur näherungsweise angeben werden. Für diesen Bereich konnten insgesamt vier Gene – *Stk33*, *Lmo1*, *Tub* und *eIf3* – identifiziert und charakterisiert werden. In den Kapiteln 3.6 und 3.7 wird im Detail auf diese Gene eingegangen werden. Die generierten genomische Sequenzinformation der BAC-Klone 282L1 und 287P4 wurde ebenfalls unter den Acc.-Nr. **AJ296304** und **AJ296303** in die EMBL Nukleotidsequenz-Datenbank eingegeben. Der PAC 368C2 wurde aufgrund der stellenweise nicht exakt zu bestimmenden Sequenzinformation nicht in der EMBL-Datenbank annotiert.

In dieser Arbeit wurden durch die Sequenzierung in zwei Spezies zusammengenommen eine genomische Sequenz von 731.946 Basenpaaren ermittelt und nach den „Bermuda-Kriterien" mit einer Genauigkeit von 99,99% verifiziert. Dieser genomische Abschnitt, geteilt auf die beiden orthologen Chromosomenabschnitte in Mensch und Maus, wurde durch die Analyse von 10.812 informativen Einzelsequenzen unter Zuhilfenahme von 782 Primern erreicht. Bei einer durchschnittlichen Leseweise pro Sequenz von ca. 500 bp multipliziert mit der Zahl der Sequenzierungen ergibt sich für alle analysierten Klone eine Gesamtsequenzierleistung an generierten Rohdaten von ca. 5,4 Megabasen.

3.5 Analyse der genomischen Sequenz

Nach Generierung und Verifizierung der genomischen Consensussequenz in beiden Spezies fand die Analyse der insgesamt 732 Kilobasen nach kodierenden, funktionellen oder strukturgebenden Sequenzabschnitten statt. Um diese Auswertung umfassend und effizient durchführen zu können, wurde mit Hilfe des Programms *RUMMAGE* gearbeitet, welches sämtliche Ergebnisse summarisch in einer interaktiven Datenbank mit graphischer Oberfläche zusammenstellt und präsentiert. Das Programm erzeugt darüber hinaus direkte Verknüpfungen zu allen relevanten Datenbankeinträgen, um sie so einer sich anschließenden Ergebnisauswertung zuzuführen. Die Komplexität dieser Zusammenstellung, extrahiert aus mehreren hundert Seiten tabellarisch gelisteter Homologien verdeutlicht exemplarisch die Grafik des Annotierungsergebnisses der humanen Sequenz in der nachfolgenden Abbildung 16. Die Sequenzbereiche der bekannten Gene *LMO1* und *TUB* sind zur Verdeutlichung grau unterlegt. Die Graphik ist insgesamt in drei Felder unterteilt. Während das obere Drittel alle Zuordnungen zur genomischen Consensussequenz in Leserichtung darstellt, finden sich im

unteren Drittel die Ergebnisse für den revers-komplementären Gegenstrang. Das mittlere Feld schematisiert auffällige Basenpaar-Zusammensetzungen (CpG-Inseln, invertiert- und tandem-repetitive Bereiche) unabhängig ihrer Leserichtung. Die Vorgehensweise im Sequenzdaten-management, welche Grundlage für die Generierung des graphischen Ergebnisses ist, wurde bereits im Flussdiagramm in Abbildung 6 schematisch dargestellt.

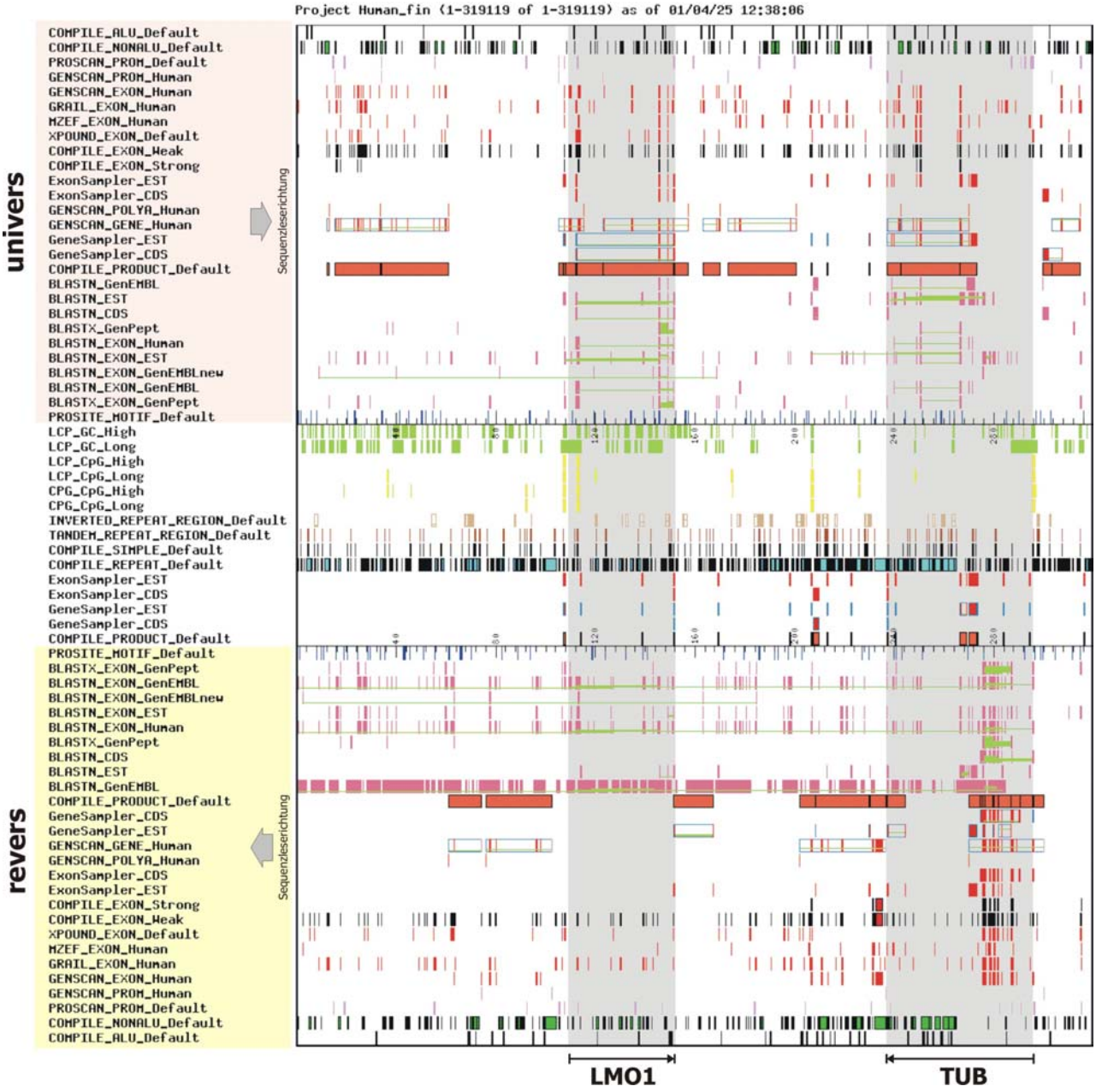


Abb. 16 **Grafische Zusammenfassung der humanen Rummage-Analyse.** Alle Ergebnisse der Rummage-Analyse wurden grafisch in einer Gesamtübersicht als farbige Balken entsprechend ihrer Lage entlang der genomischen Referenz-Sequenz angeordnet (umrahmtes Feld). Die linke Spalte listet untereinander die eingesetzten Bioinformatik-Programme auf, beginnend mit der Analyse der repetitiven Bereiche (schwarz) über die Vorhersage möglicher Promotorbereiche (lila) und möglicher Exonbereichen (rot) bis zu verschiedenen BlastN-Analysen (rosa) mit der genomischen Sequenzen und der prädiktiven Exonsequenzen (BLASTN_EXON) (Erläuterungen im Text unter Kap. 2.16). Die Gesamtgraphik ist in drei Felder unterteilt: Der obere Bereich („univers" - rötlich unterlegt) zeigt alle Programme und die jeweils ermittelten Sequenzbereiche der in Sequenzierrichtung gelesenen Sequenz und der untere Bereich („revers" - gelblich unterlegt) zeigt alle Ergebnisse der revers-komplementär gelesenen Genomsequenz. Das mittlere Feld zwischen den beiden Sequenzskalen hebt Strukturmerkmale hervor (z.B. GC-Gehalt, einfache Sequenzwiederholungen), die nicht Strang-spezifisch sind. Die Sequenzbereiche der beiden Gene LMO1 und TUB wurden grau unterlegt.

3.5.1 Genomische Consensussequenz vs. diverse Datenbanken

In der ersten Charakterisierung der genomischen Consensussequenzen fand eine BlastN- und BlastX-Analyse statt. Um unspezifische Homologien zu repetitiven Sequenzen auszuschließen, wurden vorher alle repetitiven Bereiche maskiert, die mit Hilfe der Programme *CENSOR*, *REPEATMASKER*, *SST* und *XNUN* identifiziert wurden. Für die Homologiesuche dienten in getrennten Analysen die Nukleotidsequenzdatenbanken „EMBL" (alle genomischen und genkodierenden Sequenzen), „dbEST" (sequenzierte Abschnitte aus cDNA-Bibliotheken) und „dbCDS" (vollständig sequenzierte Gene mit Start- und Stoppcodon) als Referenz. Die Zusammenfassungen dieses Homologievergleiche werden in nachfolgender Tabelle 12 mit den jeweiligen Konfigurationseinstellungen aufgezeigt. Aufgrund der teils hohen Redundanz in den EST-Datenbanken durch stark exprimierte und somit überrepräsentierte Gene ist die Summe aller Homologien zu Einträgen in den Datenbanken nicht sehr aussagekräftig. So bezogen sich allein 552 Einträge auf das Mausgen *Eif3*. Daher wurden diese Homologien aus der aus der Summe in der Statistik herausgenommen. Als zweiter Wert wurde die Zahl der genomischen Bereiche angegeben, mit Homologie Datenbankeinträge beziehen. Zudem wurden im ersten Homologievergleich mit der EMBL-Datenbank all jene Übereinstimmungen herausgerechnet, die sich auf die eigenen Annotierungen der selbstsequenzierten Klone bezogen. Ein weiterer Faktor, der die absolute Zahl an Homologien erhöhte, war die Konservierung der bekannten Gene zwischen den verschiedenen Organismen, so dass gerade bei der BlastX-Analyse auf Proteinebene sämtliche orthologen Gene aller bisher genetisch charakterisierten Spezies auffielen.

Tab. 14 Zusammenfassung der Ergebnisse des Homologievergleiches. Mit Hilfe der BlastN-Analyse konnten zahlreiche Sequenzabschnitte der genomischen Consensus-Sequenz von Mensch und Maus verschiedenen Einträgen der Datenbanken GenEMBL, EST und CDS (Erläuterung siehe Text) zugeordnet werden. Die Tabelle zeigt sowohl die Summe aller identifizierten Datenbankeinträge mit einer Homologie zur Consensussequenz, wie auch die Summe der Homologie-zeigenden Sequenzabschnitte auf der Consensus-Sequenz von Mensch und Maus. Als Filter dienten die Einstellwerte wie sie in der Spalte „Programm-Parameter" angegebenen sind. Die gezählten Datenbank-Sequenzen mussten eine Ähnlichkeit (hom.) von mindestens der angegebenen Prozentzahl auf einer Länge (len.) von mindestens der angegebenen Basenpaaren haben und eine Trefferwahrscheinlichkeit (prop.) von unter 10-20 aufweisen. Für die BlastX-Analyse mussten mind. 30 Aminosäuren (AS) eine Übereinstimmung von 50% der Aminosäuren aufweisen (ident.), bzw. 65% aus positiven Austauschen (pos.) bestehen. So konnten über die BlastX-Analyse für 22 Sequenzbereichen der humanen Genomsequenz 233 homologe Proteindatenbankeinträge zugeordnet werden. *) von den 791 Übereinstimmungen zu EST-Datenbankeinträgen fielen allein 552 Sequenzhomologien auf Transkripte des Gens eIF3, so dass diese in der ersten Summe von 189 nicht berücksichtigt wurden.

Algorithmus	Referenz-Datenbank	positive Einträge zur humanen Sequenz	positive Einträge zur murinen Sequenz	Programm-Parameter
BlastN	GenEMBL	256 / 46	287 / 55	hom. 97%, len.100 bp, prop. 10^{-20}
	EST	170 / 36	189 (791*) / 36	hom. 80%, len.60 bp, prop. 10^{-20}
	CDS	67 / 22	124 / 32	hom. 80%, len.60 bp, prop. 10^{-20}
BlastX	GenPept	233 / 22	245 / 32	ident. 50%, len. 30 AS, pos. 65%, prop. 10^{-20}

In dieser ersten Analyse konnten insgesamt 36 genomische Bereiche in beiden Spezies aufgrund ihrer Homologie zu cDNA-Sequenzen eine kodierende Funktion zugesprochen werden. Trotz der gleichen Anzahl sind die angesprochenen Bereiche zwischen Mensch und Maus nicht identisch. Der analysierte murine Genomanschnitt umfasste knapp 30% mehr genomische DNA.

3.5.2 Genomische Consensussequenz vs. Exonvorhersage-Programmergebnis

Im zweiten Analyseschritt fand die Charakterisierung der putativ genkodierenden Sequenzabschnitte statt. Hierzu wurden die Ergebnisse von vier Exonvorhersage-Programmen (*GENSCAN, GRAIL2, MZEF, XPOUND*) ausgewertet. Insgesamt berechneten diese vier verschiedenen Algorithmen für die 319.119 bp menschlicher Genomsequenz 229 unterschiedliche Exonsequenzen. Für die 412.827 bp Genomsequenz der Maus wurden 527 Exonbereiche vorhergesagt. Nutzt man die Unabhängigkeit der vier Vorhersage-Algorithmen aus und addiert nur die vorhergesagten Exonabschnitte miteinander, die gleichzeitig von drei und mehr Programmen berechnet wurden, so lässt sich die Zahl auf 24 vorhergesagte Exonbereiche im Menschen, bzw. auf 49 putativ kodierende Sequenzen bei der Maus reduzieren. Die mehrfach genannten Genomabschnitte wurden in der Datenbank „Compile_EXON_strong" zusammengefasst.

Die Einzelresultate der angewandten Vorhersageprogramme sind in nachfolgender Tabelle 11 addiert zusammengestellt. Des weiteren sind diese als kodierend angesehenen Sequenzabschnitte des Menschen in der Grafik (Abb. 25) im Kapitel 3.8.1.2 der PIP-Analyse eingezeichnet, um sie den konservierten genomischen Bereichen zwischen Mensch und Maus gegenüberstellen zu können.

Tab. 15 Ergebniszusammenfassung der Exonvorhersageprogramme GenScan, Grail2, MZEF und XPOUND. Während GenScan versucht über eine gleichzeitige Bestimmung von Promotorbereich und Polyadenylierungssignal die berechneten Exons zu putativen Genen zusammenzusetzen, nimmt die Grail2-Analyse eine qualitative Bewertung der vorhergesagten Exons in „excellent" (=100% Genauigkeit), „gut" (=69% Genauigkeit) und „grenzwertig" (="marginal") (=16% Genauigkeit) vor. Die „Compile"-Funktion summierte unter „EXON_weak" alle Ergebnisse der vier Vorhersageprogramme zusammen; „EXON_strong" zählt nur die putativen Exons, die unabhängig von mindestens drei der vier Vorhersageprogramme ermittelt wurden. Die letzte Zeile der Tabelle gibt die Zahl der Gen-Exons ohne und mit den in dieser Arbeit neu charakterisierten Exonsequenzen wieder.

Programm	Σ putativer Exons Mensch	Σ putativer Exons Maus
GENSCAN	67 = 14 Gene	74 = 12 Gene
GRAIL2	122 excellent 36 gut 66 grenzwertig 20	203 excellent 79 gut 98 grenzwertig 26
MZEF	55	335
XPOUND	74	112
COMPILE_EXON_WEAK	229	527
COMPILE_EXON_STRONG	24	49
bekannte/verifizierte Exons/Gene	16/21/3	16/31/4

3.5.3 Vorhergesagte Exonsequenzen vs. diverse Datenbanken

Um die Existenz der durch die vier Exonvorhersageprogramme berechneten Exonsequenzen *in silico* verifizieren zu können, wurden diese DNA-Abschnitte im dritten Analyseschritt selektiv einer Homologie-Analyse unterzogen. Wie schon zuvor mit der gesamtgenomischen DNA wurde ein Nukleotidsequenz-Homologievergleich mit den Datenbanken „EMBLnew", „dbEST" und „dbCDS" durchgeführt. Ebenso wurde die Homologie der translatierten DNA in alle drei möglichen Leserahmen auf Proteinebene durch eine BlastX-Analyse untersucht. Das Ergebnis ist summarisch in Tabelle 12 zusammengefasst. Mit Hilfe dieser Analysen konnten zu den 16 bekannten humanen Exonsequenzen der Gene *LMO1* und *TUB* über die Homologie zu cDNA-Klon-Sequenzen (ESTs) weitere 20 neue Bereiche herausgehoben werden, von denen sich 5 Bereiche experimentell über RT-PCR als neue Exonsequenzen verifizieren ließen. Für die Maus-genomische Consensussequenz waren es insgesamt 15 neue Exonsequenzen, die verifiziert werden konnten.

Tab. 16 Ergebniszusammenfassung der Blast-Analysen mit den vorhergesagten humanen und murinen Exonsequenzen. Die mit den Exonvorhersageprogrammen GenScan, Grail2, MZEF und XPOUND berechneten Exonsequenzen wurden einer BlastN- und BlastX-Analyse unterzogen und mit dem Datenbestand der Referenzdatenbanken EMBL, EMBLnew, EST und Rodent, bzw. GenPept verglichen. Die Tabelle zeigt sowohl die Summe aller berechneten Exonsequenzen mit Homologie zu Einträgen in den Datenbanken, wie auch die Summe aller Sequenzeinträge, die eine Homologien zu diesen putativen Exonbereichen aufwiesen. So zeigten z. B. 36 berechnete Exonbereiche aus der untersuchten Humansequenz mit insgesamt 340 EST-Einträgen Homologie. Als Filterkriterien dienten die Werte in der Spalte „Programm-Parameter". hom: Homologie in Prozent; len: kleinste Länge der homologen Sequenz; ident: Prozentsatz der identischen Aminosäuren; AS: Aminosäuren; pos: Prozentsatz der positiven Aminosäurenaustausche. Die große Zahl an Einzelhomologien für EST-, bzw. zu Proteinsequenzen resultierte zum einen aus einer Überrepräsentation stark exprimierter Gene in den Datenbanken und zum anderen aus der Konserviertheit der Aminosäuresequenzen zu allen weiteren Genfamilienmitgliedern.

Algorithmus	Referenz-Datenbank	Putative humane Exonsequenzen mit Homologie zu Datenbankeinträgen [Zahl der Exons / Zahl der Datenbankeinträge]	Putative murine Exonsequenzen mit Homologie zu Datenbankeinträgen [Zahl der Exons / Zahl der Datenbankeinträge]	Programm-Parameter
BlastN	EMBL	36 / 212	./.	hom. 80%, len. 60 bp
	EMBLnew	21 / 44	35 / 99	hom. 80%, len. 60 bp
	EST	36 / 340	136 / 574	hom. 60%, len. 60 bp
	Rodent	./.	12 / 70	hom. 80%, len. 60 bp
BlastX	GenPept	49 / 205	55 / 228	ident. 50%, len. 30 AS, pos. 65%

3.5.4 Vorhergesagte Promotoren und Polyadenylierungsstellen

Ein weiteres wichtiges Kriterium für die Identifizierung neuer Genbereiche ist das Vorhandensein eines Transkriptionsstartpunktes. So konnten im Rahmen der *PROSCAN*-Analyse für den untersuchten humangenomischen Bereich insgesamt 38 putative Promotor-Bereiche angesprochen werden, von denen 15 ein TATA- und ein CAP-Motiv aufwiesen. Die Promotor-Vorhersage der *GENSCAN*-Analyse zeigte insgesamt 12 Bereiche, von denen bis auf zwei Ausnahmen bei Nukleotid 33.945 bis 34.207 und bei Nukleotid 105.362 bis 105.401 im 5´-Bereichs des *LMO1*-Exons 1a kein Bereich mit der *PROSCAN*-Analyse übereinstimmte. Somit konnte bis auf diese beiden Ausnahmen die Vorhersageergebnisse der beiden Programme für insgesamt 48 Promotoren nicht gegenseitig bestätigt werden. Sämtliche Ergebnisse wurde in der nachfolgenden PIP-Analyse (Abb. 25, Kap. 3.8.1.2) grafisch eingezeichnet, um sie im so Gesamtkontext beurteilen zu können.

Bei der Untersuchung der Maus-genomischen DNA konnten durch die *PROSCAN*-Analyse insgesamt 33 putative Promotoren bestimmt werden, von denen 13 ein TATA- und ein CAP-Motiv aufwiesen. Die zum Vergleich durchgeführte Promotor-Vorhersage der *GENSCAN*-Analyse ergab neun vorhergesagte Abschnitte für einen Promotorbereich. Auch hier ergab von insgesamt 41 putativen Promotorbereichen der Ergebnisvergleich nur einen Abschnitt von Nukleotid 377.468 bis 377.725, der in beiden Programmen als möglicher Promotor erkannt wurde.

Die Auswertung der möglichen Polyadenylierungsstellen, die im Rahmen der *GENSCAN*-Analyse bestimmt wurden, ergab für die humangenomische DNA-Sequenz 14 und für die Maus-genomische DNA 12 wahrscheinliche Stellen. Auch hier wurden die Ergebnisse zur humanen Sequenz in die PIP-Grafik in Kap. 3.8.1.2 eingetragen.

3.6 Strukturaufklärung bekannter Gene

3.6.1 Das LIM domain only 1-Gen –- LMO1/Lmo1 (Mensch/Maus)

Das dieser Arbeit als Startpunkt dienende Gen *LMO1* , welches über Fluoreszenz-in-situ-Hybridisierung an das distale Ende der zu sequenzierenden humanen Megabasen-Region kartiert wurde, konnte in seiner genomischen Struktur in beiden Spezies mit drei verschiedenen Spleißvarianten beschrieben werden. Als Referenz diente die von McGuire & Mitarbeitern (1993) annotierte 1.274 bp-lange humane Gensequenz mit der Acc.-Nr. M26682 (oder NM_002315). Die murine Gensequenz war trotz mehrerer Publikationen (Boehm *et al.*, 1990a, 1990b; Foroni *et al.*, 1992) den Datenbanken nicht zu entnehmen und wurde erst im Rahmen dieser Arbeit unter der Acc.-Nr. NM_057173 veröffentlicht.

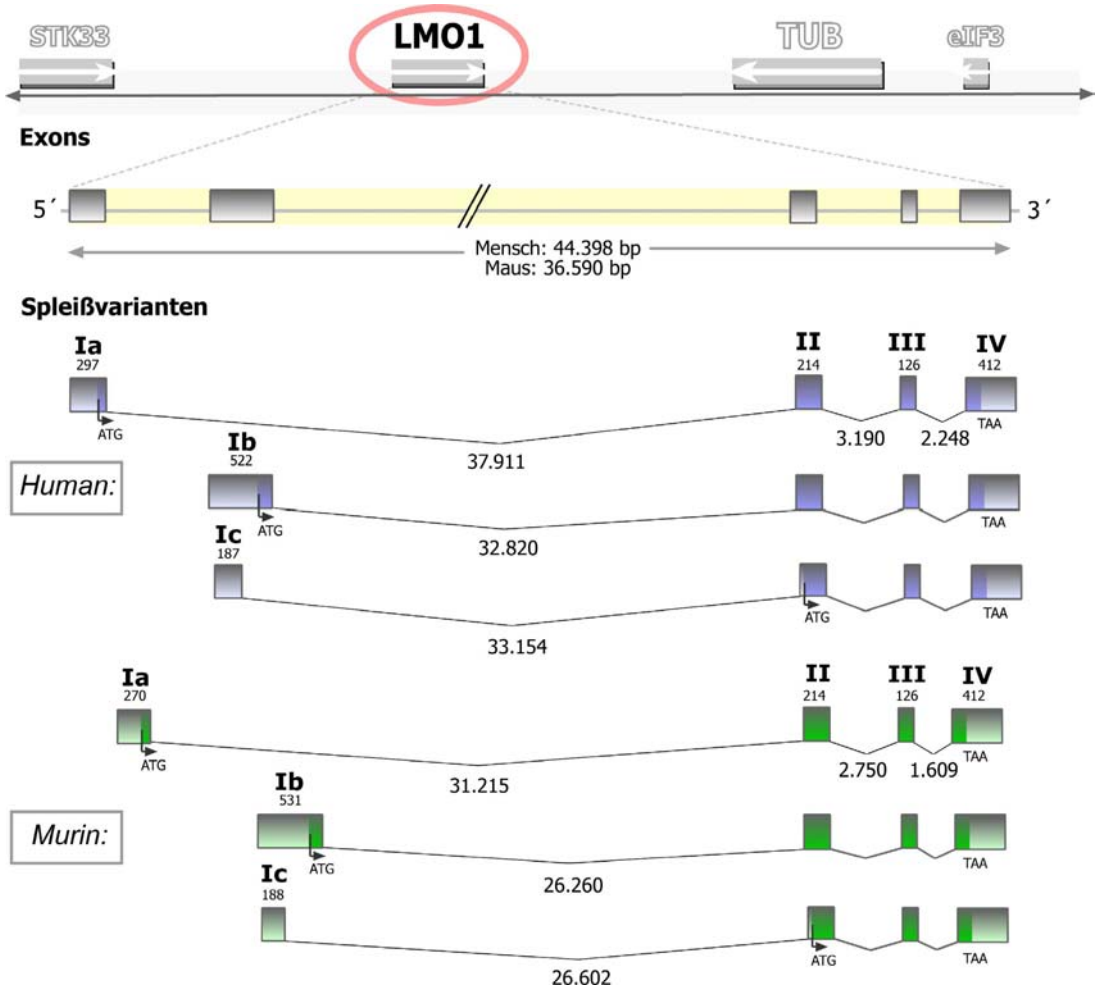


Abb. 17 **Exon-Intron-Struktur des Gens LMO1 in Mensch (blau) und Maus (grün)**. Die mRNA-Sequenz des LMO1-Gens setzt sich aus 4 Exons zusammen, wobei das erste Exon in beiden Spezies in drei alternativen Spleißformen vorliegt. Die Spleißvarianten A und B besitzen im ersten Exon das Startcodon ATG (durch einen abgewinkelten Pfeil gekennzeichnet). Die Spleißvariante C nutzt eine interne Spleißakzeptorstelle des Exons 2, da das erste Exons (1c) kommt, was eine Verschiebung des Translationsstart zum nächsten Startcodon in Exon 2 zur Folge hat. Das LMO1-Gen erstreckt sich im Humangenom über 44.398 bp und im Mausgenom über 36.590 bp. Dieser Sequenzbereich wird vor allem durch das erste Intron dominiert, welches für die Spleißform A allein 85% der gesamten transkribierten Sequenz einnimmt. Die Bereiche des 5´- und 3´-UTRs sind durch eine nur schwache Einfärbung gekennzeichnet.

3.6.1.1 Das humane LMO1-Gen

Das humane *LMO1*-Gen erstreckt sich über 44.398 bp genomischer Sequenz und setzt sich aus insgesamt fünf kodierenden Exonbereichen zusammen, wobei die ersten beiden Exons durch drei unterschiedliche Spleißvariante alternativ genutzt werden. Die Exons zwei bis vier sind in allen Spleißformen identisch. Die Spleißvariante A weist mit 37.911 bp zwischen Exon 1 und 2 die längste Intronsequenz auf. Der Translationsstart befindet sich bei Base 276 im ersten Exon, welches somit die ersten sieben Codons des 155 Aminosäuren langen Proteins kodiert. Die Spleißvariante B, die durch die annotierte Sequenz Acc.-Nr. NM_002315 repräsentiert wird, besitzt das erste Exon (1b) in 5 kb Entfernung distal, so dass sich die Länge des Introns 1b auf 32.820 bp verkürzt. Der Translationsstart beginnt auch hier im ersten Exon 24 bp vor Ende der 522 bp langen Exon 1-Sequenz und wird durch

zwei konsekutive ATG-Codons bestimmt. Die nachfolgenden Codons sind bis auf einen einzigen Nukleotidunterschied zur ersten Spleißvariante identisch. Das sechste Basentriplett weist an erster Stelle die Base Adenin anstelle von Cytosin auf, was ein Austausch der dortigen Aminosäure - von Glutamin nach Lysin - nach sich zieht. Somit wird in dieser Spleißvariante eine saure Aminosäure durch eine Aminosäure mit basischer Seitenkette ersetzt. Das erste Exon der Spleißvariante C, die im Rahmen dieser Arbeit erstmals beschrieben wurde, endet bereits nach 188 bp an einer internen Spleißakzeptorstelle. Dass es sich hierbei um eine reelle und nicht um eine kryptische Spleißstelle handelt, konnte durch Sequenzieren von RT-PCR-Amplifikaten verifiziert werden. Die Nutzung des Exons 1c hat zur Folge, dass sich der Translationsstart in das zweite Exon hinein verschiebt und die Proteinsequenz um 11 Aminosäuren von 156 auf 145 verkürzt wird. Im Gegensatz zur murinen Sequenz konnten für die Spleißformen A und C in der EST-Datenbank keine cDNA-Sequenzen als Referenz bestimmt werden. Alle Exons und ihre genomische Lokalisation sind in nachfolgender Tabelle 15 zusammengestellt.

Im Rahmen der *in silico*-Analyse konnten außerdem für beide alternative Exons 1a und 1b jeweils ein Promotorbereich charakterisiert werden. So weist die *PROSCAN*-Analyse die Bereiche von nt 107.298 bis nt 107.547 (Score: 9,46) und von nt 112.615 bis nt 112.864 (Score 13,93) der Referenz-Consensussequenz als Promotorbereiche aus.

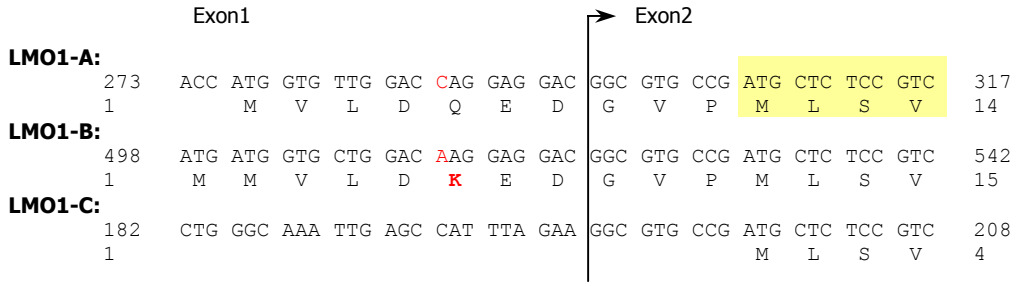


Abb. 18: Sequenzvergleich der alternativ gespleißten 5´-Regionen der drei humanen LMO1-Varianten. Das Exon 1a der Genvariante LMO1-A kodiert für die ersten 7 Aminosäuren der LMO1-Proteinsequenz. Die Spleißvariante LMO1-B weist im kodierenden Bereich des Exons 1b fast die gleichen Codons auf. Außer einem zusätzlichen Methionin-Triplett am Translationsstart zeigt die Exonsequenz 1b im sechsten Codon ein alternatives Nukleotid (rot hervorgehoben). Anstelle der Base Cytosin an gleicher Position der Variante A ist bei Spleißvariante B ein Adenin zu finden. Dies hat zur Folge, dass anstatt der sauren Aminosäure Glutamin ein Lysin mit basischer Seitenkette in die Proteinkette miteingebaut wird. Der Translationsstart der Spleißvariante C beginnt erst im zweiten Exon, so dass sich die Proteinsequenz dieser Spleißvariante um die ersten 11 Aminosäuren verkürzt.

Tab. 17 Genomische Lokalisierung des humanen LMO1-Gens mit Exon-Intron-Grenzen der mRNAs. Im oberen Abschnitt der Tabelle ist die exakte Lage des Gens zur humanen genomischen Referenzsequenz und die Größen der Exons und Introns eingetragen. Gleichzeitig wurden die unmittelbar benachbarten Basen als Sequenzausschnitt der Spleißakzeptor- und der Spleißdonor-Stelle dargestellt. Die

Gensequenz mit Exon 1b entspricht der Datenbank-Annotierung Acc.-Nr. NM_002315, die alternativen Exons 1a und 1c sind erst im Rahmen dieser Arbeit bestimmt worden. Der untere Abschnitt der Tabelle fasst die exakten Positionen des offenen Leserahmens mit Start- und Stoppcodon zur genomischen Referenzsequenz zusammen. Die drei alternativen Startcodons enden alle mit einem gemeinsamen Stoppcodon.

INTRON	genom. Position	Spleiß-akzeptor	EXON (Länge in bp)	Spleiß-donor	genom. Position	INTRON
	107.336	CGGCCC..	**Exon1a** 296	..ACG**gt**agg..	107.632	***Intron1a*** 37.911
	112.202	CAGCGGG..	**Exon1b** 522	..ACG**gt**agg..	112.723	***Intron1b*** 32.820
	112.203	AGCGGGA..	**Exon1c** 188	..AAG**gt**gag..	112.389	***Intron1c*** 33.154
Intron1	145.544	..tgc**ag**GCG..	**Exon2** 214	..GAG**gt**ggg..	145.757	***Intron2*** 3.190
Intron2	148.948	..ggc**ag**GCT..	**Exon3** 126	..GAG**gt**cag..	149.073	***Intron3*** 2.248
Intron3	151.322	..tct**ag**ATT..	**Exon4** 412	..TTCCGGG	151.733	

UTR	genom. Position	Start-Codon	ORF	Stop-Codon	genom. Position	UTR
5´- 275	107.611	..**ATG**GTG..	**ORF-A** 465 bp 155 AS	..CAG**taa**..	151.424	**-3´**
5´-	112.699	..**ATG**ATG..	**ORF-B** 468 bp 156 AS	..CAG**taa**..	151.424	**-3´**
5´-	145.552	..**ATG**CTC..	**ORF-C** 435 bp 145 AS	..CAG**taa**..	151.424	**-3´**

3.6.1.2 Das murine LMO1-Gen

Die genomische Struktur des Maus-*Lmo1*-Gens ist dem humanen Homolog sehr ähnlich und weist ebenfalls drei alternative erste Exons auf. Auch bei der Maus befindet sich der Translationsstart jeder alternativen Spleißvariante auf einem anderen Exon. Die Konservierung ist so hoch, dass selbst die alternative Base an der ersten Position im fünften Basentriplett der murinen Spleißvariante A zur Variante B zu finden ist. Auch bei der Maus ist diese Position alternativ mit einen Cytosin, bzw. Adenin besetzt, mit der gleichen Folge eines Aminosäuresaustausches in der Proteinsequenz. Der Vergleich der mRNA-Sequenzen beider Spezies zeigte für die Spleißvarianten A eine Homologie von 90%, für die Varianten B von 93% und für die Varianten C von 92%. Die genomische Erstreckung der längsten der drei Spleißvarianten ist mit 36.590 bp in der Maus um 7,8 kb kürzer als im Menschen. Die Verkürzung des Exonabstände spiegelt sich in allen Intronsequenzen wieder. Die Unterschiede zwischen Mensch und Maus betragen dabei mit Introngrößen von 2.750 bp (Intron 2) und 1.609 bp (Intron 3) zwischen -16% und -40%. Das sehr große erste Intron (Intron 1a) weist mit 31.154 bp bei der Maus einen DNA-Verlust von 21% bzw. mit 26.260 bp (Intron 1b) von 25% gegenüber dem Humangenom auf. Auch hier zeigte die *ProScan*-Analyse, dass die Bereiche von nt 142.219 bis nt 142.468 und von nt 147.688 bis nt 147.937 als Promotorbereiche der beiden *Lmo1*-Genvarianten anzusehen sind.

Tab. 18 Genomische Lokalisierung des murinen Lmo1-Gens mit Exon-Intron-Grenzen der mRNAs. Im oberen Abschnitt der Tabelle ist die exakte Lage des Gens zur genomischen Referenzsequenz eingetragen. Die Gensequenz mit Exon 1b entspricht dem Datenbankeintrag Acc.-Nr. NM_057173, die im Rahmen dieser Arbeit annotiert wurde. Der untere Abschnitt der Tabelle fasst die exakten Positionen des offenen Leserahmens zur genomischen Referenzsequenz zusammen. Die drei alternativen Startcodons enden wie beim humanen Homolog mit einem gemeinsamen Stoppcodon.

INTRON	genom. Pos.	Spleiß-akzeptor	EXON (Länge in bp)	Spleiß - donor	genom. Pos.	INTRON
	142.456	CGGCCG..	**Exon1a** 270	..ACG**gt**agg..	142.725	**Intron1a** 31.215
	147.150	CAGCGGG..	**Exon1b** 531	..ACG**gt**agg..	147.680	**Intron1b** 26.260
	147.151	AGCGGGA..	**Exon1c** 188	..GAA**gt**gag..	147.338	**Intron1c** 26.602
Intron1	173.939	..cac**ag**GTG..	**Exon2** 214	..GAG**gt**ggg..	174.152	**Intron2** 2.750
Intron2	176.901	..ggc**ag**GCT..	**Exon3** 126	..GAG**gt**gag..	177.026	**Intron3** 1.609
Intron3	178.634	..tct**ag**ATT..	**Exon4** 412	..GGACATG	179.045	

UTR	genom. Pos.	Start-Codon	ORF	Stop-Codon	genom. Pos.	UTR
5´-	142.704	..ATGGTT..	**ORF-A** 465 bp 155 AS	..CAGtaa..	178.736	**-3´**
5´-	147.656	..ATGATG..	**ORF-B** 468 bp 156 AS	..CAGtaa..	178.736	**-3´**
5´-	173.947	..ATGCTC..	**ORF-C** 435 bp 145 AS	..CAGtaa..	178.736	**-3´**

3.6.2 Das Tubby-Gen – TUB/Tub (Mensch/Maus)

In ca. 125 Kilobasen Entfernung telomerwärts vom *LMO1*-Gen befindet sich im humanen Genom das Gen *Tubby*. Erstmals von Noben-Trauth & Mitarbeiter (1996) in der Maus und von Kleyn & Mitarbeiter (1996) im Menschen beschrieben, konnte dieses Gen durch Positionelle Klonierung der Maus-Tubby-Mutation zugeordnet werden, deren Träger durch eine ausgeprägte Fettleibigkeit und Körperfülle im Vergleich zu den gesunden Artgenossen auffielen. Auch dieses Gen zeichnet sich durch mehrere alternative Spleißvarianten aus. Für die kodierende Sequenz konnten insgesamt 18 verschiedene Exonbereiche charakterisiert werden (siehe Abb. 19). Es konnte sowohl im Menschen wie in der Maus jeweils eine unbekannte Spleißvariante beschrieben werden. Alle Exon-Intron-Übergänge weisen die konservierten Spleißdonor- ([Exon]GT..Intron) und Spleißakzeptorstellen (Intron..AG[Exon]) auf.

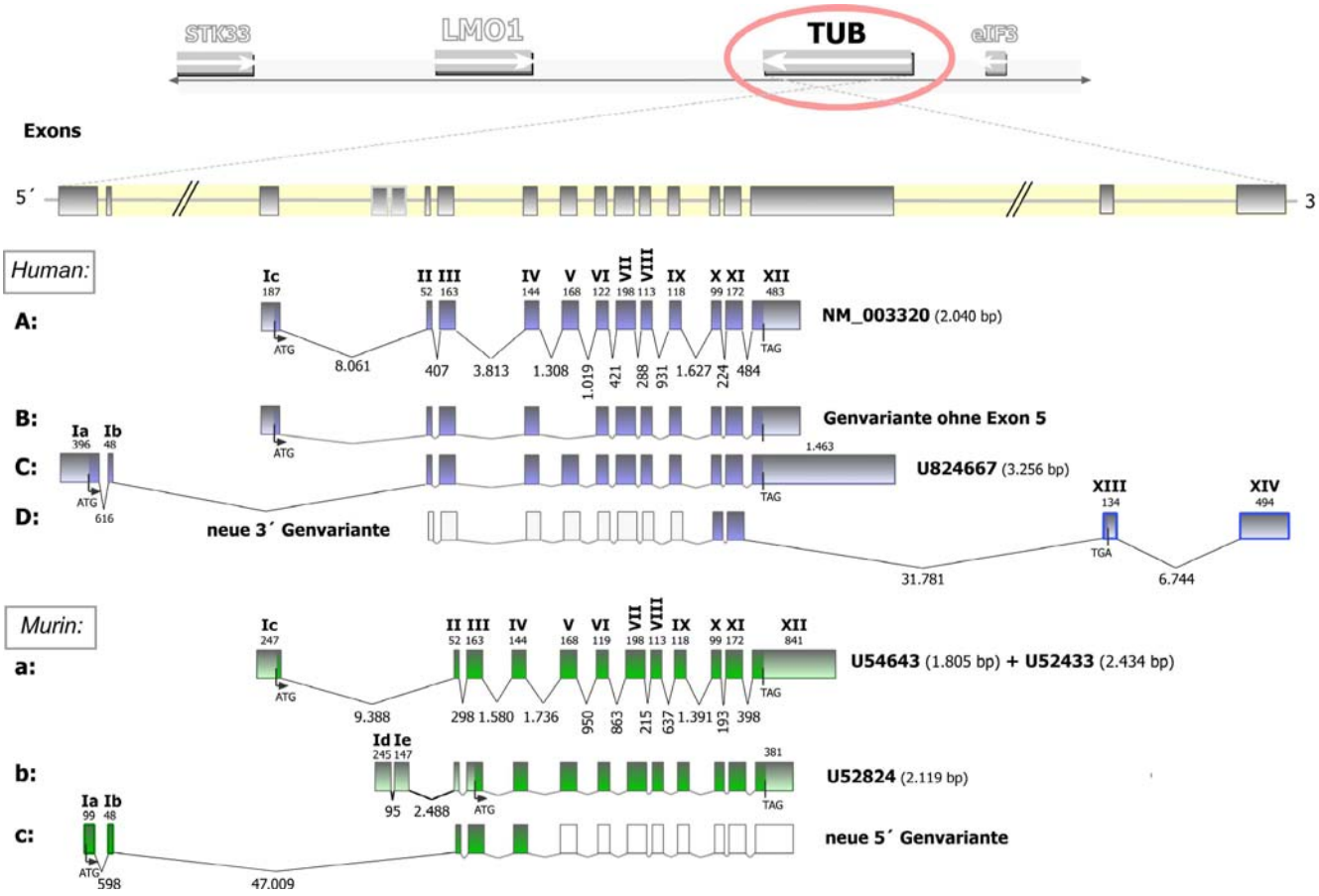


Abb. 19 **Genomische Organisation des Gens „Tubby" in Mensch (blau) und Maus (grün).** Zwischen den Genen LMO1 (proximal) und eIF3 (distal) befindet sich auf dem DNA-Gegenstrang der sequenzierten genomischen Sequenz der kodierende Abschnitt des TUB-Gens. Zur besseren Lesbarkeit wurde die Exon-Intron-Reihenfolge im Vergleich zur genomischen Orientierung invertiert, so dass das Gen in gewohnter Leserichtung vom 5´zu 3´ dargestellt werden konnte. Insgesamt umfasst die „Tubby"-mRNA 18 verschiedene Exonsequenzen, die im humanen Genom mindestens in vier und im murinen Genom mindestens in drei unterschiedlichen Spleißvarianten vorliegen. Die alternativen Exons liegen vor allem am Genanfang und am Genende. Die Exons 2 bis 10 sind in allen Genvarianten beider Spezies vorhanden. Eine Ausnahme bildet lediglich das Exon 5, welches alternativ herausgespleißt werden kann. Für das humane TUB konnten folgende Genvarianten beschrieben werden: Die Genvariante A, mit der annotierten Sequenz Acc.-Nr. NM_003320 als Referenz, umfasst 12 Exons. Der Genvariante B zeichnet sich durch das Fehlen des Exons 5 aus. Die insgesamt 13 Exons umfassende Genvariante C wird durch die Annotierung Acc.-Nr. U82467 repräsentiert und besitzt einen wesentlich weiter distal gelegenen Genstart. Die neu entdeckte Genvariante D besitzt am 3´-Ende die zwei zuvor unbekannten alternativen Exons 13 und 14, die durch ein über 31 kb großes Intron mit der restlichen Gensequenz verbunden sind. Für das Mausgenom konnte die Genvariante a charakterisiert werden, die das Homolog zur humanen Sequenz A darstellt und ebenfalls von 12 Exons gebildet wird. Als Referenz für Variante a dienten die Annotierungen Acc.-Nr. U54643 und U52433, die sich lediglich in einer unterschiedlich langen 3´-UTR-Sequenz unterscheiden. Die Genvariante b, bestehend aus 13 Exons, besitzt einen alternativen 5´-Bereich, der durch die Exons 1d und 1e gebildet wird. Aufgrund eines fehlenden offenen Leserahmens beginnt der Translationsstart bei dieser Genvariante erst in Exon 3. Als Referenz hierzu diente die Datenbanksequenz Acc.-Nr. U52824. Als direktes Homolog zur humanen Variante C kann die neue murine Genvariante c mit den neuen Exons 1a und 1b betrachtet werden, die in einem Abstand von 47 kb zum Exon 2 lokalisiert werden konnten.

3.6.2.1 Das humane TUB-Gen

Die Sequenzierung der humangenomischen Sequenz erlaubte eine genaue Anordnung aller Exons der annotierten cDNA-Sequenzen Acc.-Nr. NM_00320 (2.040 bp) und U82467 (3.256 bp) (siehe dazu

Tab. 17). Während die Sequenz NM_003320 eine insgesamt 12 Exons umfassende Variante über einen genomischen Bereich von 20.601 bp darstellt, beschreibt die annotierte Sequenz U82467 eine cDNA mit insgesamt 13 Exons. Im Unterschied zur ersten cDNA hat diese Variante einen etwa 1.000 bp längeren 3´untranslatierten Bereich, und einen über 49 Kilobasen weiter distal gelegenen Translationsstart, der außerhalb der sequenzierten 319.119 bp Consensus-Sequenz lokalisiert ist und im Anschlusscosmid cSRL82e3 mit zwei Exons (1a und 1b) zum Liegen kommt. Der transkribierte Bereich vergrößert sich dadurch auf insgesamt über 60 Kilobasen. Die wirkliche Distanz konnte aufgrund der fehlenden genomischen Sequenz nur grob angegeben auf ca. 65 kb geschätzt werden und basierte in Bezug auf die Entfernung im murinen Genom. Der Abstand der beiden ersten Exons 1a und 1b zueinander betrug 616 bp.

Als unbekannte Variante wurde die mRNA-Variante E charakterisiert, die im 3´-Bereich des TUB-Gens zwei neue Exons aufweist. Die cDNA-Klone IMAGp998P214518 (Acc.-Nr: AI239685), IMAGp998O164517 (Acc.-Nr: AI239997) und IMAGp998D224063 (Acc.-Nr: AA987338) verlängern die Gensequenz von Exon 11 über ein 31.781 bp großes Intron um 134 bp (Exon 13) und 494 bp (Exon 14) nach weiteren 6.744 bp Intronsequenz. Die Sequenz des Exons 12 wird durch das Spleißen entfernt. Da sich der offene Leserahmen in diesen neuen Bereich erstreckt, führt dies auch zu einer Änderung des carboxyterminalen Endes der *TUB*-Proteinsequenz. Es ist um 34 Aminosäuren kürzer als die bekannten Sequenzen und unterscheidet sich in den letzten 10 Proteinbausteinen. Das Stoppcodon ist in Exon 13 lokalisiert; Exon 14 besteht vollständig aus 3´-untranslatierter Sequenz. Welche Konsequenzen diese Verkürzung für die räumliche Struktur des Proteins hat, wird in Kapitel 4.2.2.1 diskutiert werden. Der Transkriptionsstart dieser mRNA-Variante D konnte aufgrund fehlender, nicht ausreichend langer cDNA-Klone nicht ermittelt werden. Auch eine Sequenzverlängerung durch RT-PCR-Reaktionen brachte keine auswertbaren Ergebnisse, so dass keine Zuordnung der alternativen 5´-Sequenzvarianten ab Exon 10 vorgenommen werden konnte. Die drei identifizierten cDNA-Klone lassen als Expressionsort vor allem die Lunge in Betracht kommen, da Gewebe dieses Organs für alle drei cDNA-Sequenzen als Ursprungsquelle dienten. Als weitere Expressionsorte könnten zusätzlich die Gewebe Hoden und B-Zellen angesehen werden, da die Klone IMAGp998P214518 und IMAGp998O164517 aus einem Pool von drei normalisierten Bibliotheken der Gewebe Lunge/Testis/B-Zellen stammen.

Tab. 19 Genomische Lokalisierung der humanen TUB-Gen mit den Exon-Intron-Grenzen der verschiedenen mRNAs. Im oberen Abschnitt der Tabelle ist die exakte Lage und Länge aller Exons zur human-genomischen Referenzsequenz eingetragen. Die Gensequenz mit Exon 1a und 1b entspricht der Datenbank-Annotierung Acc.-Nr. U824667. Die Gensequenz mit Exon 1c entspricht der Datenbank-Annotierung Acc.-Nr. NM_003320. Alle Exon-Intron-Grenzen wiesen eine konservierte Spleißdonor und Spleißakzeptorstelle mit den Basen „GT" und „AG" auf. Der untere Abschnitt der Tabelle fasst die exakten Positionen des offenen Leserahmens von Start- zum Stoppcodon zur genomischen Referenzsequenz zusammen. Die Genvarianten A und C enden am gemeinsamen Amber-Stoppcodon „UAG", dagegen findet die Genvariante D mit dem Opal-Stoppcodon „UGA" ihren Abschluss.

INTRON	genom. Pos.	Spleiß-akzeptor	EXON (Länge in bp)	Spleiß-donor	genom. Pos.	INTRON
	./.	CTT..	**Exon1a** 396	..TCG**gt**..	./.	**Intron1a** 616
	./.	..	**Exon1b** 48	..GAG**gt**..	./.	**Intron1b** 49.917
	295.980	TGG..	**Exon1c** 187	..CAG**gt**..	295.794	**Intron1c** 8.061
Intron1	287.732	..**ag**TGT..	**Exon2** 52	..CAG**gt**..	287.681	**Intron2** 407
Intron2	287.273	..**ag**CGG..	**Exon3** 163	..AAG**gt**..	287.111	**Intron3** 3.813
Intron3	283.297	..**ag**TCA..	**Exon4** 144	..AAG**gt**..	283.154	**Intron4** 1.308
Intron4	281.845	..**ag**GCA..	**Exon5** 168	..AGG**gt**..	281.678	**Intron5** 1.019
Intron5	280.658	..**ag**GCA..	**Exon6** 122	..AGG**gt**..	280.537	**Intron6** 421
Intron6	280.115	..**ag**GAG..	**Exon7** 198	..AAG**gt**..	279.918	**Intron7** 288
Intron7	279.629	..**ag**GTG..	**Exon8** 113	..GCG**gt**..	279.517	**Intron8** 931
Intron8	278.585	..**ag**GTC..	**Exon9** 118	..TAC**gt**..	278.468	**Intron9** 1.627
Intron9	276.840	..**ag**GAG..	**Exon10** 99	..AAC**gt**..	276.742	**Intron10** 224
Intron10	276.517	..**ag**GAG..	**Exon11** 172	..ACC**gt**..	276.346	**Intron11** 484
Intron11	275.861	..**ag**CGG..	**Exon12** 483	..GGGGAG	275.379	
Intron11b 31.781	244.564	..ccc**ag**AAA..	**Exon13** 134	..ACT**gt**aag..	244.431	**Intron12** 6.744
	237.688	..ggc**ag**GGG..	**Exon14** 494	..TGGTGAT	237.195	

UTR	genom. Pos.	Start-Codon	ORF	Stop-Codon	genom. Pos.	UTR
5´	295.831	.gac**ATG**AC.	**ORFb** NM_003320	.GAG**TAG**.	275.730	**-3´**
5´	?	.gac**ATG**AC.	**ORFa** U82467	.GAG**TAG**.	275.730	**-3´**
			ORFe neue 3´-Variante	.ATC**TGA**.	244.535	**-3´**

3.6.2.2 Das murine Tub-Gen

In der sequenzierten genomischen Maussequenz konnten drei unterschiedliche *Tub*-Genvarianten beschrieben werden. Aus dem Datenbestand der Entrez-nr-Datenbank wurden insgesamt drei vollständige („full-length") cDNA-Sequenzen identifiziert und mit der genomischen Maussequenz verglichen. Die cDNA-Sequenzen Acc.-Nr. U54643 und U52433, die sich lediglich in einem verschieden langen 3´-UTR unterschieden, ließen sich zu der Gensequenz a (siehe Abb. 19) bestehend aus 12 Exons zusammenfassen. Die zweite murine Genvariante b wurde durch die Annotierung Acc.-Nr. U52824 beschrieben und umfasste insgesamt 13 Exons. Aufgrund interner Stoppcodons in den ersten beiden Exons 1d und 1e, wird ein alternativer Translationsstart in Exon 3 genutzt. Die Folge ist eine Verkürzung der Proteinsequenz von Genvariante b um die ersten 46 Aminosäuren.

In einer Entfernung von 47.009 bp auf der genomischen Referenzsequenz konnten über die Homologie zur humanen cDNA-Sequenz Acc.-Nr. U824667 zwei neue murine Exonsequenzen (Exon 1a und Exon 1b) identifiziert werden. RT-PCR-Experimente mit cDNA aus den Geweben Gonaden (Tag 17,5) und Hoden (adult) und den Primern *putTub_EX1-f* und *putTub_EX2-r* (siehe Kap. 2.18.5.4) konnten diese Exons verifizieren. Aus dem Exon 1a konnten 99 bp und für das Exon 1b alle 48 bp generiert werden. Beide Exons sind durch eine kurze Intronsequenz über 598 bp voneinander getrennt. Im Vergleich zur *Tub*-Gensequenzvariante a mit Exon 1c besitzt diese Variante c 78 bp mehr an kodierender Nukleotidsequenz, und verlängert so das *Tub*-Protein c um 26 Aminosäuren am Amino-terminalen Ende. Die mRNA-Sequenz dieser neuen Variante c konnte bis zum Exon 4 experimentell bestätigt und in Hirngewebe der Maus über RT-PCR nachgewiesen werden. Eine weiterreichende cDNA-Sequenz-Komplettierung konnte auch über eine EST-Datenbanksuche nicht erreicht werden. Es gab keine Homologie zu bereits ansequenzierten cDNA-Fragmenten, die diesen neuen 5´-Genbereich hätten erweitern können.

Ebenso wie die neu beschriebene murine 5´-*Tub*-Genvariante c (Acc.-Nr. U52433) mit den Exons 1a und 1b über die Homologie zur humanen Genvariante C entdeckt wurde, konnte die Genvariante a der Maus der humanen mRNA-Sequenz Acc.-Nr. NM_003320 zugeordnet werden; beide Gensequenzen weisen die gleiche Exon-Intron-Organisation auf. Für die murine Genvariante b (Acc.-Nr. U52824) mit den Exon 1d und 1e konnte ein humanes Homolog nicht eindeutig charakterisiert werden, obwohl im direkten Interspezies-Sequenzvergleich in der erwarteten Genomregion des Menschen in ca. 2.250 bp Entfernung zum Exon 2 sich zwei DNA-Bereiche mit erhöhter Konservierung von mehr als 60% über 165 bp für den Abschnitt des Exons 1d und über 93 bp für das Exon 1e identifizieren ließen (siehe Abb. 20). Auch die Länge der humanen Intronsequenz mit 113 bp entspricht ungefähr dem Abstand von 95 bp der beiden Exons im murinen Genom. Der Versuch diesen Bereich über RT-PCR aus Nieren- und Gehirn-Gewebe zu verifizieren, führte allerdings zu keinem Ergebnis.

Für die neu beschriebenen humanen Exons 13 und 14 der *TUB*-Variante C konnte über einen Interspezies-Sequenzvergleich im Mausgenom kein homologer konservierter Bereich charakterisiert

werden, so dass die Existenz dieser Genabschnitte in der Maus nicht abschließend geklärt werden konnte.

Tab. 20 Genomischen Lokalisierung des murinen Tub-Gens mit den Exon-Intron-Grenzen der verschiedenen mRNAs. Im oberen Abschnitt der Tabelle ist die Lage aller Exons mit den konservierten Spleißdonor- und Spleißakzeptorstellen in der genomischen Referenzsequenz eingetragen. Die Gensequenz mit den Exons 1a und 1b (entspricht der Genvariante c in Abb. 19) beschreibt eine neue bisher unbekannte Tub-Genvariante der Maus. Die Gensequenz mit Exon 1c entspricht der Datenbank-Annotierung Acc.-Nr. U52433. Die Gensequenz mit den Exons 1d und 1e entsprechen der Referenzsequenz Acc.-Nr. U52824. Alle Exon-Intron-Grenzen weisen eine konservierte Spleißdonor und Spleißakzeptorstelle mit den Basen „GT" und „AG" auf. Der untere Abschnitt der Tabelle fasst die exakten Positionen des offenen Leserahmens zur genomischen Referenzsequenz zusammen. Enden Genvariante a und b an dem gemeinsamen Amber-Stoppcodon „UAG", so konnte das Transkriptionsende für Genvariante c aufgrund fehlender cDNA-Sequenzen nicht eindeutig bestimmt werden.

INTRON	genom. Pos.	Spleiß-akzeptor	EXON (Länge in bp)	Spleiß-donor	genom. Pos.	INTRON
	335.127	ACT..	**Exon1a** 99	..TAG**gt**..	335.029	Intron1a 598
Intron1a	334.430	..**ag**GAG..	**Exon1b** 48	..TGG**gt**..	334.383	Intron1b 47.009
	297.008	CCT..	**Exon1c** 247	..CAG**gt**..	296.762	Intron1c 9.389
	290.348	ATC..	**Exon1d** 245	..GAG**gt**..	290.104	Intron1d 95
Intron1d	290.008	..**ag**ATG..	**Exon1e** 147	..AAG**gt**..	289.862	Intron1e 2.488
Intron1	287.373	..**ag**GTG..	**Exon2** 52	..CAG**gt**..	287.322	**Intron2** 298
Intron2	287.023	..**ag**CGG..	**Exon3** 163	..AAG**gt**..	286.861	**Intron3** 1.580
Intron3	285.280	..**ag**TTG..	**Exon4** 144	..AAG**gt**..	285.137	**Intron4** 1.736
Intron4	283.400	..**ag**GCA..	**Exon5** 168	..AGG**gt**..	283.233	**Intron5** 950
Intron5	282.282	..**ag**GCA..	**Exon6** 119	..CGG**gt**..	282.164	**Intron6** 863
Intron6	281.300	..**ag**GAG..	**Exon7** 198	..AAG**gt**..	281.103	**Intron7** 215
Intron7	280.887	..**ag**GTG..	**Exon8** 113	..GCG**gt**..	280.775	**Intron8** 637
Intron8	280.137	..**ag**GTC..	**Exon9** 118	..TAT**gt**..	280.020	**Intron9** 1.391
Intron9	278.628	..**ag**GAG..	**Exon10** 99	..AAT**gt**..	278.530	**Intron10** 193
Intron10	278.336	..**ag**Gaa..	**Exon11** 172	..ACC**gt**..	278.165	**Intron11** 398
	277.766	..**ag**CGG..	**Exon12** 841	..CCCA	276.925	

UTR	genom. Pos.	Start-Codon	ORF	Stop-Codon	genom. Pos.	UTR
5´-	296.799	.gac**ATG**AC.	**ORFa** U52433	.GAG**tag**.	277.635	**-3´**
5´-	335.096	.ctc**ATG**GG.	**ORFb** neue 5´-Variante	?	?	**-3´**
5´-	286.975	.ttg**ATG**GT.	**ORFc** U52824	.GAG**tag**.	277.635	**-3´**

put. TUB Exon 1d

```
hum_genom            U52824
(290266>290416)      (189<41)
66.4% identity in 152 residues overlap; Score: 42.0; Gap frequency: 3.3%

    v290270   v290280     v290290   v290300   v290310   v290320
GCCATGGACTATGCA--CAGTTATGCCAGAGGCAGGAGGCTGCTTTCCATGAAAACATTC
||| |||||| | ||  ||||| || | | ||   ||||||| |||||| ||  ||  ||
GCCCTGGACTTTCCATACAGTTGTGTCTGGGGTCAGAGGCTGTTTTCCACGAGCACGGTC
         ^180      ^170      ^160      ^150      ^140      ^130

      v290330   v290340    v290350   v290360   v290370   v290380
CTTTCTCCTCCTCTGCCT-CCACCTCCATTTCCTGATGGGGGAGAATGATAAAGAGCCACA
||| ||   || ||| || ||||    |||||||| | ||| |||  || |  |||||| |
CTTCCTTTCCCACTGTCTTCCAC----ATTTCCTGGTTGGGAAGAGGGACACTGAGCCATA
         ^120      ^110          ^100      ^90       ^80

      v290390   v290400   v290410
AGGCTTCGGTGACCCTGATTACTAGGTAAGATA
 || ||  ||  |  ||  ||  || |||||||
-GGTTTTAGTACCATTGGCTAGCAGATAAGATA
   ^70       ^60       ^50

hum_genom            U52824
(290190>290205)      (245<230)
87.5% identity in 16 residues overlap

v290190   v290200
CTCACTGCCCAGTGCC
|||||||||  |||||
CTCACTGCCTGGTGCC
     ^240      ^230
```

put. TUB Exon 1e

```
hum_genom            U52824
(289983>290077)      (338<244)
67.7% identity in 93 residues overlap

       v289990   v290000   v290010   v290020   v290030   v290040
GGCAGTCACCTGGGC-CATCCCCAGCCCTGGGCTTGGCTGAGCAAGACAAAGAAGATGACT
|||||  ||||||   ||| ||  |   |||||| |||| ||||   ||||||| ||| ||
GGCAGCCACCTGGTATCATTCCAGGTGTTGGTCTCGGCTAAGCATACCAAAGAACATGCCT
        ^330      ^320      ^310      ^300      ^290      ^280

      v290050   v290060   v290070
TCCATGGTGCTGACGCTCACATCTAGAGTCATTTAG
|||  |||||  | |   |  ||||| | ||| |
TCCTCGGTGCCAATGAAAATGTCTAGGGACATCTCA
       ^270      ^260      ^250
```

Abb. 20 Interspezies-Homologievergleich der murinen Tub-Exonsequenzen 1d und 1e (Genvariante c) mit der humanen Genomsequenz. In etwa 2.250 bp Entfernung zum humanen TUB-Exon 2 befinden sich zwei Abschnitte mit erhöhter Homologie zu den ersten Maus-Exonsequenzen der Genvariante b. Tub-Exon 1e zeigt eine Ähnlichkeit von 67% über 93 Basenpaare und Tub-Exon 1d mit 66% Homologie über 149 Basenpaare zum Humangenom. Beide Abschnitte liegen in einem Abstand 113 bp zueinander entfernt. Vergleicht man die humangenomischen Abstände und Längen mit denen der Maus, so lässt die Anordnung der putativen kodierenden Bereiche eine ähnliche Architektur in beiden Spezies erkennen.

3.7 Identifizierung weiterer unbekannter Gene

3.7.1 Murines Homolog zum eukaryontischen Translationsinitiationsfaktor 3, 47 kDa Untereinheit - Eif3s5 (Maus)

Im proximalen Bereich der sequenzierten murinen Genomsequenz konnte in einer Entfernung von 31.066 bp zum *Tub*-Gen ein neuer kodierender Genbereich identifiziert und beschrieben werden. Aufgrund seiner hohen Homologie von 91% auf Proteinebene und 83% auf Nukleotidebene zur 47 kDa Untereinheit des humanen eukaryontischen Translationsinitiationsfaktors 3 (*eIF3S5*, Acc.-Nr. U94855) schien es sich hierbei um das putative murine Homolog zu handeln. Es konnten drei cDNA-Klone (RIKEN-Klon 2410002E12, RIKEN-Klon 1110003L04, RIKEN-Klon 0610037M02) ermittelt werden, die mit ihrer Sequenz den gesamten kodierenden Bereich des Genes umschrieben. Mit 1.251 Basenpaaren überspannte die längste cDNA-Sequenz einen genomischen Bereich von 7.272 bp des acht Exons umfassenden Gens. Die Analyse des Promotorbereiches mit Hilfe des *PROSCAN*-Programms ergab einen putativen Transkriptionsstart bei Nukleotid 373.497 der genomischen murinen Referenzsequenz 50 bp vor dem Startcodon „ATG". Der Translationsstartpunkt entsprach mit der Sequenzfolge „GGC**A**AG$_{\mathbf{AUG}}$**G**" der von Kozak (1999) postulierten Initiationssequenz und wies, fett hervorgehoben, die wichtigen Basen an den Positionen –3 und +4 um das Startcodon „AUG" auf (siehe Abb. 21). In 29 bp Entfernung stromaufwärts vom Transkriptionsstartpunkt konnte zudem im Promotorbereich eine TATA-Box als RNA-Polymerase II-Bindungsstelle identifiziert werden. Des weiteren war dieser genomische Abschnitt von nt 373.093 bis nt 373.455 durch einen erhöhten GC-Gehalt von 73,8% charakterisiert (siehe auch Abb. 29), der sich durch 46 CpG-Dinukleotid-wiederholungen auszeichnete und als Promotorbereich-typische CpG-Insel beschrieben werden kann. Die Analyse des 3´UTRs zeigte nach 1.948 bp stromabwärts zum Stoppcodon „TGA" die Poly-A-Consensussequenz „AATAAA". Für diesen 3´-UTR-Bereich konnten die ersten 152 bp durch die EST-Sequenzen verifiziert werden. Der translatierte Bereich des Gens erstreckte sich über 1.083 bp der mRNA und kodierte für ein 361 Aminosäuren langes Protein mit einem errechneten Molekulargewicht von 37,96 kDa. Die Homologie zu mehr als 550 EST-Sequenzen aus den unterschiedlichsten Geweben zu allen Bereichen der Gensequenz deuteten auf eine ubiquitäre Expression hin. Die Auswertung dieser EST-Sequenzen ergab keinen Hinweis auf ein etwaiges alternatives Spleißen der acht Exons.

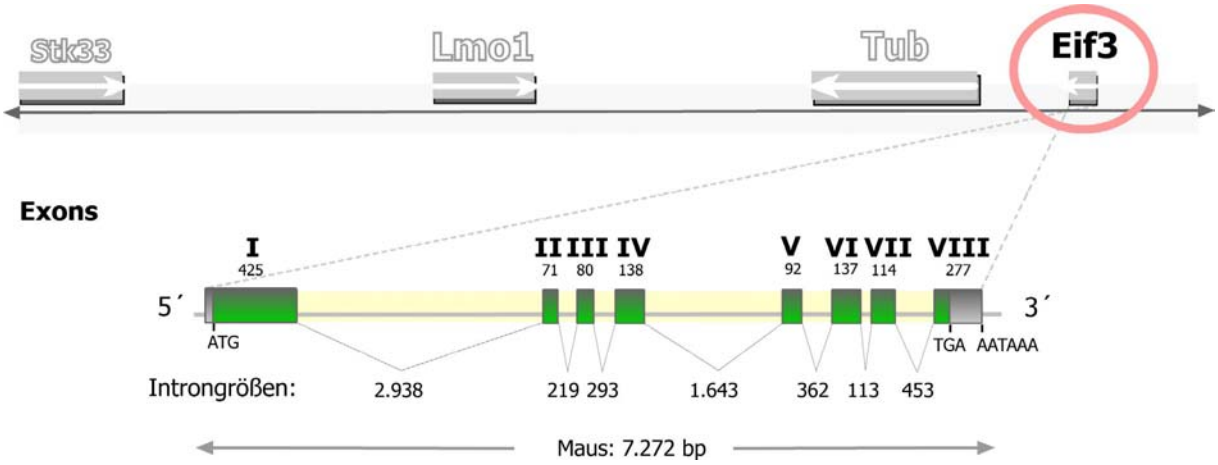


Abb. 21 **Genomische Organisation aller Exons des neuen murinen Gens Eif3s5** (Eukaryontischer Translationsinitiationsfaktor 3 Untereinheit 5). Am proximalen Ende der sequenzierten, genomischen Maussequenz in einem Abstand von 31.066 bp zum Tub-Gen befindet sich das murine Homolog zum humanen eukaryontischen Translationsinitiationsfaktor 3. Der offene Leserahmen (grün) erstreckt sich von nt 50 (in Exon 1) bis nt 1.132 (in Exon 8) der mRNA und umfasst 361 Aminosäuren. Insgesamt überspannt das Gen mit acht Exons einen genomischen Sequenzbereich von 7.272 bp. In 123 bp Entfernung zum Stoppcodon befindet sich ein Polyadenylierungssignal. Sowohl die Exongrößen – über den schematisch dargestellten Exonsequenzen - wie auch die Introngrößen sind angegeben.

Tab. 21 **Genomische Lokalisierung der Exon-Intron-Grenzen der murinen mRNA des Eif3-Gens.** Als Referenz wurde die annotierte Sequenz des cDNA-Klons RIKEN-Klon 0610037M02 herangezogen. Im oberen Abschnitt der Tabelle ist die Lage aller Exons mit den konservierten Spleißdonor- und Spleißakzeptorstellen in der genomischen Referenzsequenz eingetragen. Alle Exon-Intron-Grenzen wiesen eine konservierte Spleißdonor und Spleißakzeptorstelle mit den Basen „GT" und „AG" auf. Die Erstreckung des offenen Leserahmens über 1.083 bp wurde unter Angabe der genomischen Position des Start- und Stoppcodons in der unteren Hälfte der Tabelle vermerkt.

INTRON	genom. Pos.	Spleiß-akzeptor	cDNA	EXON	cDNA	Spleiß-donor	genom. Pos.	INTRON
	373.464	..TCTCTTT..	1	**Exon1** 391	391	..TGG**gt**gag..	373.073	**Intron1** 2.938
Intron1	370.134	..acc**ag**GAA..	392	**Exon2** 71	462	..GAA**gt**gag..	370.064	**Intron2** 219
Intron2	369.844	..ctt**ag**GTG..	464	**Exon3** 80	542	..CTG**gt**aag..	369.765	**Intron3** 293
Intron3	369.471	..tgc**ag**GTA..	543	**Exon4** 138	680	..CAG**gt**gag..	369.334	**Intron4** 1.643
Intron4	367.690	..ccc**ag**CAC..	681	**Exon5** 92	772	..GAG**gt**gag..	367.599	**Intron5** 362
Intron5	367.236	..cac**ag**TTG..	773	**Exon6** 137	909	..CTG**gt**gag..	367.100	**Intron6** 113
Intron6	366.986	..ctc**ag**TCT..	910	**Exon7** 114	1.023	..AAT**gt**gag..	366.873	**Intron7** 453
Intron7	366.419	..tcc**ag**GAC..	1.024	**Exon8** 228	1.251	..TGTAAA	366.194	

UTR	genom. Pos.	Start-Codon	cDNA	ORF	cDNA	Stop-Codon	genom. Pos..	UTR
5´-	373.448	..cggcaag**ATG**..	17 1	**ORF** AS	1099 361	..TGTGAatg..	366.344	**-3´**

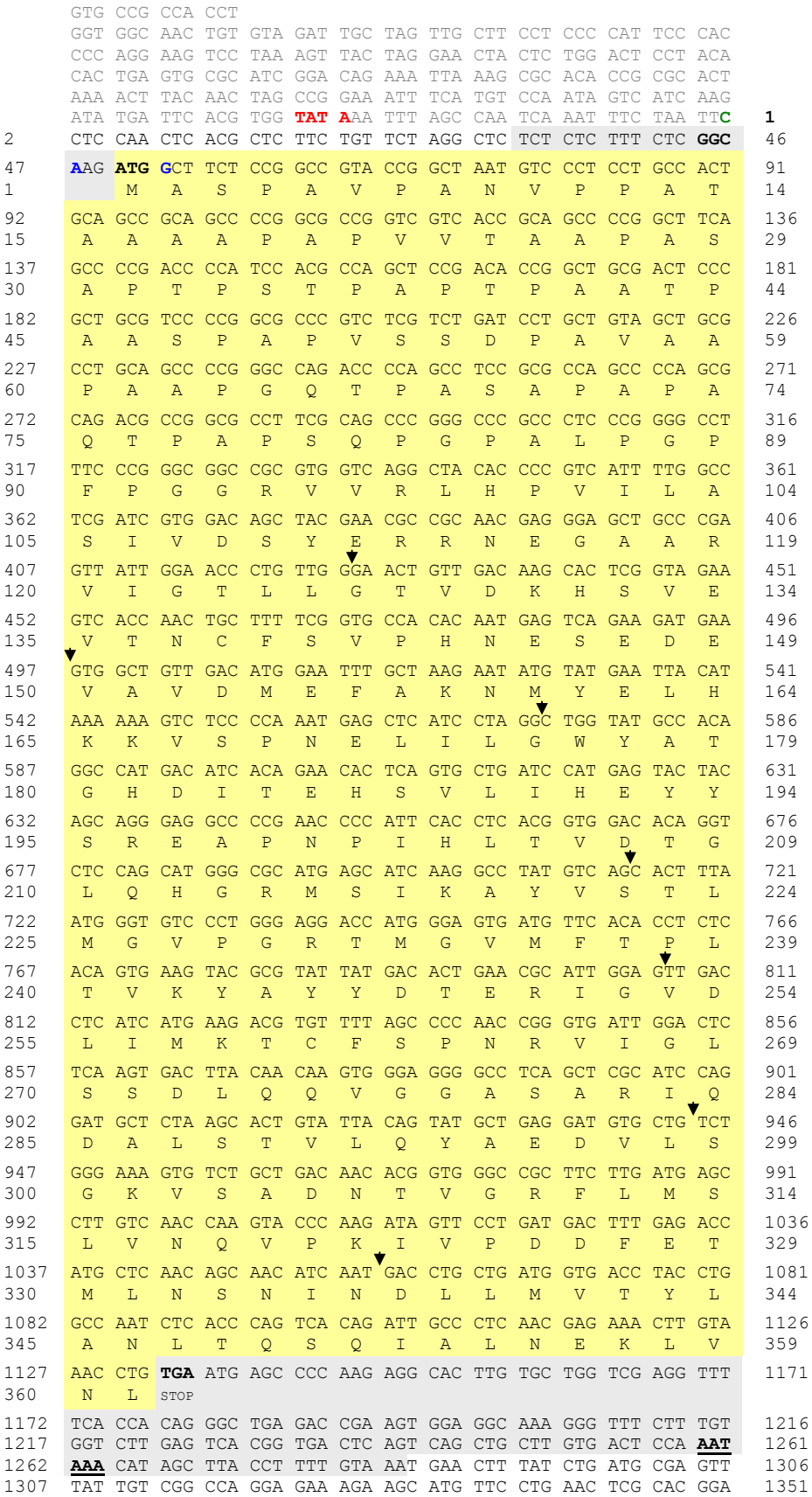

GTG CCG CCA CCT
GGT GGC AAC TGT GTA GAT TGC TAG TTG CTT CCT CCC CAT TCC CAC
CCC AGG AAG TCC TAA AGT TAC TAG GAA CTA CTC TGG ACT CCT ACA
CAC TGA GTG CGC ATC GGA CAG AAA TTA AAG CGC ACA CCG CGC ACT
AAA ACT TAC AAC TAG CCG GAA ATT TCA TGT CCA ATA GTC ATC AAG
ATA TGA TTC ACG TGG TAT AAA TTT AGC CAA TCA AAT TTC TAA TTC 1
2 CTC CAA CTC ACG CTC TTC TGT TCT AGG CTC TCT CTC TTT CTC GGC 46
47 AAG ATG GCT TCT CCG GCC GTA CCG GCT AAT GTC CCT CCT GCC ACT 91
1 M A S P A V P A N V P P A T 14
92 GCA GCC GCA GCC CCG GCG CCG GTC GTC ACC GCA GCC CCG GCT TCA 136
15 A A A A P A P V V T A A P A S 29
137 GCC CCG ACC CCA TCC ACG CCA GCT CCG ACA CCG GCT GCG ACT CCC 181
30 A P T P S T P A P T P A A T P 44
182 GCT GCG TCC CCG GCG CCC GTC TCG TCT GAT CCT GCT GTA GCT GCG 226
45 A A S P A P V S S D P A V A A 59
227 CCT GCA GCC CCG GGC CAG ACC CCA GCC TCC GCG CCA GCC CCA GCG 271
60 P A A P G Q T P A S A P A P A 74
272 CAG ACG CCG GCG CCT TCG CAG CCC GGG CCC GCC CTC CCG GGG CCT 316
75 Q T P A P S Q P G P A L P G P 89
317 TTC CCG GGC GGC CGC GTG GTC AGG CTA CAC CCC GTC ATT TTG GCC 361
90 F P G G R V V R L H P V I L A 104
362 TCG ATC GTG GAC AGC TAC GAA CGC CGC AAC GAG GGA GCT GCC CGA 406
105 S I V D S Y E R R N E G A A R 119
407 GTT ATT GGA ACC CTG TTG GGA ACT GTT GAC AAG CAC TCG GTA GAA 451
120 V I G T L L G T V D K H S V E 134
452 GTC ACC AAC TGC TTT TCG GTG CCA CAC AAT GAG TCA GAA GAT GAA 496
135 V T N C F S V P H N E S E D E 149
497 GTG GCT GTT GAC ATG GAA TTT GCT AAG AAT ATG TAT GAA TTA CAT 541
150 V A V D M E F A K N M Y E L H 164
542 AAA AAA GTC TCC CCA AAT GAG CTC ATC CTA GGC TGG TAT GCC ACA 586
165 K K V S P N E L I L G W Y A T 179
587 GGC CAT GAC ATC ACA GAA CAC TCA GTG CTG ATC CAT GAG TAC TAC 631
180 G H D I T E H S V L I H E Y Y 194
632 AGC AGG GAG GCC CCG AAC CCC ATT CAC CTC ACG GTG GAC ACA GGT 676
195 S R E A P N P I H L T V D T G 209
677 CTC CAG CAT GGG CGC ATG AGC ATC AAG GCC TAT GTC AGC ACT TTA 721
210 L Q H G R M S I K A Y V S T L 224
722 ATG GGT GTC CCT GGG AGG ACC ATG GGA GTG ATG TTC ACA CCT CTC 766
225 M G V P G R T M G V M F T P L 239
767 ACA GTG AAG TAC GCG TAT TAT GAC ACT GAA CGC ATT GGA GTT GAC 811
240 T V K Y A Y Y D T E R I G V D 254
812 CTC ATC ATG AAG ACG TGT TTT AGC CCC AAC CGG GTG ATT GGA CTC 856
255 L I M K T C F S P N R V I G L 269
857 TCA AGT GAC TTA CAA CAA GTG GGA GGG GCC TCA GCT CGC ATC CAG 901
270 S S D L Q Q V G G A S A R I Q 284
902 GAT GCT CTA AGC ACT GTA TTA CAG TAT GCT GAG GAT GTG CTG TCT 946
285 D A L S T V L Q Y A E D V L S 299
947 GGG AAA GTG TCT GCT GAC AAC ACG GTG GGC CGC TTC TTG ATG AGC 991
300 G K V S A D N T V G R F L M S 314
992 CTT GTC AAC CAA GTA CCC AAG ATA GTT CCT GAT GAC TTT GAG ACC 1036
315 L V N Q V P K I V P D D F E T 329
1037 ATG CTC AAC AGC AAC ATC AAT GAC CTG CTG ATG GTG ACC TAC CTG 1081
330 M L N S N I N D L L M V T Y L 344
1082 GCC AAT CTC ACC CAG TCA CAG ATT GCC CTC AAC GAG AAA CTT GTA 1126
345 A N L T Q S Q I A L N E K L V 359
1127 AAC CTG TGA ATG AGC CCC AAG AGG CAC TTG TGC TGG TCG AGG TTT 1171
360 N L STOP
1172 TCA CCA CAG GGC TGA GAC CGA AGT GGA GGC AAA GGG TTT CTT TGT 1216
1217 GGT CTT GAG TCA CGG TGA CTC AGT CAG CTG CTT GTG ACT CCA AAT 1261
1262 AAA CAT AGC TTA CCT TTT GTA AAT GAA CTT TAT CTG ATG CGA GTT 1306
1307 TAT TGT CGG CCA GGA GAA AGA AGC ATG TTC CTG AAC TCG CAC GGA 1351

Abb. 22 mRNA-Sequenz des murinen eukaryontischen Translationsinitiationsfaktors 3, Untereinheit 5 (Eif3S5). Der kodierende Genbereich (gelb) umfasst 361 Aminosäuren. Er wird flankiert von einem 50 bp langem 5´untranslatierten Bereich (Transkriptionsstart: grün). 18 bp dieses Bereiches konnten durch cDNA-Sequenzen (grau) verifiziert werden. In 29 bp Entfernung stromaufwärts zum Transkriptionsstart konnte eine Promotor-relevante Consensussequenz (TATA-Box) identifiziert werden (rot). Die Translationsinitiation bei Nukleotid 50 konnte mit der „Kozak"-Sequenz (GGC**A**AG$_{AUG}$**G**)(blau) in Übereinstimmung gebracht werden (Näheres siehe Text). Der 3´untranslatierte Genbereich besteht bis zum Polyadenylierungssignal „AATAAA" (schwarz, unterstrichen) aus 123 bp. Dieser Bereich konnte ebenfalls mit cDNA-Sequenzen bestätigt werden. Die Grenzen der Exons sind mit Pfeilspitzen in die Nukleotidsequenz eingezeichnet. Die Aminosäuresequenz befindet sich in eines zweiten Zeile direkt unter der korrespondierenden Nukleotidsequenz.

Da der genomische Bereich des humanen *EIF3S5*-Gens nicht mehr durch die eigens generierte humane DNA-Sequenz abgedeckt war, wurde versucht diesen Abschnitt über BlastN-Analysen und Datenbank-Recherchen zu rekonstruieren und für den Interspeziesvergleich zugänglich zu machen. Bei dieser Suche konnten insgesamt vier genomische Sequenzannotationen identifiziert werden. Zwei Sequenzeinträge (Acc.-Nr. AC013694 und AC087648) bezogen sich auf Klone, die unvollständig in ihrer Sequenz vorlagen und für den genkodierenden Gesamtbereich nur lückenhafte Sequenzinformationen boten. Zwei weitere Genomsequenzen mit hoher Homologie zum murinen Gen lagen vollständig sequenziert vor und konnten für eine detailliertere Analyse herangezogen werden. BAC-Klon RP11-334G22 (Acc.-Nr. AC007250) umschrieb eine 180.557 bp lange Genomsequenz, die auf Chromosom 2p16.1 kartiert wurde, und BAC RPCI11-429A20 (Acc.-Nr. AC005906) umspannte auf Chromosom 12p13.3 insgesamt 185.952 bp.

Der Vergleich der genomischen Sequenz von Chromosom 2 (Acc.-Nr. AC007250) zeigte in der PIP-Analyse (siehe Abb. 28) eine Konservierung für alle acht murinen Exons mit einer Übereinstimmung der Basenpaare von 76% (Exon 1) bis 93% (Exon 5). Die Betrachtung der genomischen Chromosom 2-DNA verdeutlichte außerdem, dass die dortige zum *eIF3S5*-Gen homologe Sequenz aus einer einzigen Sequenz besteht, die nicht von Intronsequenzen unterbrochen ist. Direkte Sequenzwiederholungen zum Genanfang und am Genende im Anschluss an den Poly-A-Schwanz wurden als Hinweise auf ein Pseudogen gedeutet, das durch Retrotransposition an diese chromosomale Stelle gelangt ist (siehe auch Kap. 4.2.4).

3.7.2 Serin/Threonin-Kinase-Gen – Stk33 (Maus)

Am distalen Ende der in dieser Arbeit bestimmten murinen Genomsequenz konnten zwei weitere kodierende Sequenzen identifiziert werden, die über 786 bp eine 100%ige Homologie zum murinen RIKEN-cDNA-Klon 4921505G21 (Acc.-Nr. AK014819) aufwiesen. Diese zwei exprimierten Abschnitte stellten das 3´-Ende eines neuen Gens in der Maus dar, das von Mujica & Mitarbeitern (2001) als Serin/Threonin-Kinase-Gens *Stk33* beschrieben wurde. Der Sequenzvergleich zeigte, das diese Abschnitte den Exons 11 und 12 der insgesamt 2.389 bp langen mRNA-Sequenz des Gens zugeordnet werden konnten. Die Exonbereiche von 150 und 634 Basenpaaren wurden dabei durch eine 32.304 bp große Intronsequenz getrennt. Das Stoppcodon des 406 Aminosäuren langen Proteins befand sich

nach BlastN-Analyse an Position 38.002 der murinen Referenz-Genomsequenz. Da eine ausführliche Charakterisierung von Mujica & Mitarbeitern (2001) vorgenommen wurde und die orthologe human Genomsequenz nicht Gegenstand dieser Arbeit war, wurde von weiteren Untersuchungen des Gen abgesehen.

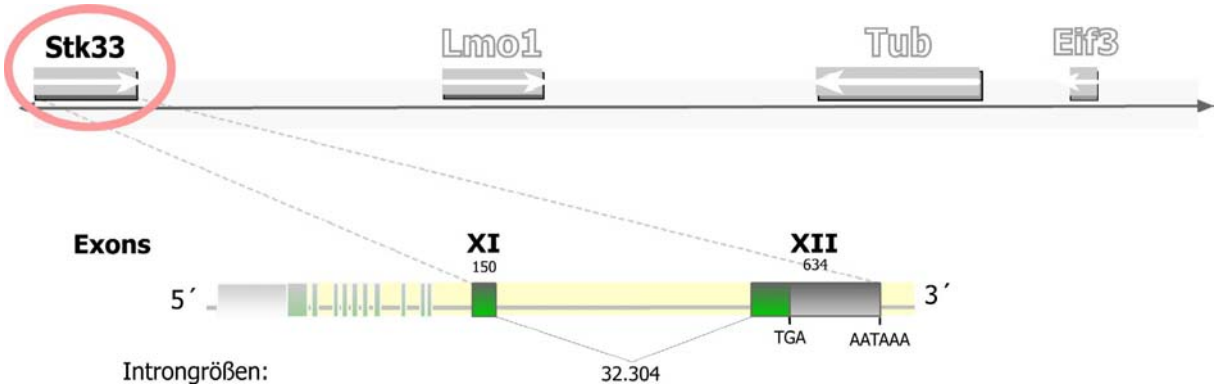


Abb. 23: Genomische Organisation der Exons 11 und 12 des neuen murinen Gens *Stk33*-Gens. Für das 3´-Ende des murinen *Stk33*-Gen konnten in der sequenzierten Mausgenomsequenz die beiden letzten Exon der cDNA-RIKEN-Klon-Sequenz 4921505G21 identifiziert werden. Beide Exons sind durch eine 32.304 bp Intronsequenz voneinander getrennt. Das Stoppcodon des insgesamt 406 Aminosäuren langen Proteins befindet sich an Position 38.002 der murinen Referenz-Genomsequenz.

Tab. 22 Genomische Lokalisierung der Exons 11 und 12 der murinen mRNAs des Stk33-Gens. (oberen Hälfte der Tabelle) Die Erstreckung der beiden letzten Exons als Teil des offenen Leserahmens auf der cDNA des RIKEN-Klons 4921505G21 (Acc.-Nr. AK014819). Die untere Hälfte der Tabelle gibt die Lokalisation des Genendes mit der Position des Stoppcodons auf genomischer und cDNA-Sequenz wieder. Es wurde auch das außerhalb der analysieren Referenz-Genomsequenz liegende Startcodon der 406 Aminosäuren (AS) des offen Leserahmens (ORF) bei Nukleotid 725 der RIKEN-Klon-cDNA-Sequenz aufgeführt.

INTRON	genom. Pos.	Spleiß-akzeptor	cDNA	EXON	cDNA	Spleiß-donor	genom. Pos.	INTRON
...	5.357	..ctt**ag**GGC..	1.604	**Exon11** 150	1.753	..CAG**gt**agg..	5.506	**Intron11** 32.304
Intron11	37.813	..cat**ag**CCC..	1.754	**Exon12** 634	2.387	..CAGAAG	38.446	

UTR	genom. Pos.	Start-Codon	cDNA	ORF	cDNA	Stop-Codon	genom. Pos.	UTR
5´-	n.n.	gcaca**ATG**..	725 1	**ORF** AS	1.942 406	..CTCTAAggt	38.002	**-3´**

3.8 Interspezies-Homologievergleiche der sequenzierten Genomsequenzen

Die verschiedenen in dieser Arbeit benutzten Vorhersageprogramme konnten für die ausgewählten Sequenzbereiche nur eine Wahrscheinlichkeit bezüglich ihrer kodierenden Funktion angeben. Mit Hilfe des Interspeziesvergleichs ist es darüber hinaus möglich gewesen, über die evolutive Konservierung funktionell wichtiger Sequenzbereichen, diese Abschnitte zu bestätigen und mögliche neue Gen zu identifizieren. Konnte die hohe Aussagekraft der Homologie stark konservierter Basenpaarabschnitt bereits anhand der bekannten Gene und ihrer neuen zusätzlichen Exons im vorigen Kap 3.7 demonstriert werden, so wird im diesem Kapitel ein genereller Vergleich aller Nukleotide der in dieser Arbeit generierten Genomsequenzen von Mensch und Maus analysiert.

3.8.1 Vergleich Mensch - Maus

3.8.1.1 Dotplot-Analyse:

Einen Überblick für den Grad der Ähnlichkeit zweier Sequenzen gibt die Dotplot-Analyse, bei der in einem zweidimensionalen Koordinatensystem zwei Sequenzen miteinander verglichen werden. Die hier untersuchten beiden genomischen Bereiche zeigen über die gesamte gemeinsame Distanz Homologiepunkte in einer fast durchgehenden Diagonale. Diese Diagonale belegt den hohen Grad der Syntänie dieses chromosomalen Abschnitts zwischen den Spezies Mensch und Maus. Als Homologie-Parameter für den Dotplot wurde eine minimale Übereinstimmung von 65% über einen Bereich von 50 Basenpaaren eingestellt, so dass jeder 50 bp-Sequenzabschnitt mit gleichgroßer oder höherer Ähnlichkeit sich als Punkt in der nachfolgenden Grafik (Abb. 24) wiederfindet.

Das Ergebnis dieser Analyse zeigte, dass sämtliche Gene in ihrer Orientierung und Zahl der Exons zwischen Mensch und Maus konserviert geblieben sind. Darüber hinaus fand sich nicht nur in den kodierenden Bereichen eine hohe Homologie, auch viele Intergen-Bereiche zeigten eine hohe Sequenzübereinstimmung. Bis auf einen 8,6 kb-Bereich zwischen den Basen nt. 185.445 und nt. 194.086 des Menschen, der im Mausgenom zu fehlen scheint, konnten keine größeren Deletionen oder Insertionen detektiert werden. Schaut man sich die Neigung der Diagonalen an, die sich aus insgesamt 29.793 Einzelpunkten zusammensetzt, so fällt auf, dass sie nicht exakt im 45° Winkel verläuft, wie bei gleichmäßiger Verteilung der Sequenzhomologie zu erwarten wäre. Vielmehr verkleinert sich der Winkel hin zur humangenomischen Sequenz, die Diagonale verläuft flacher als erwartet. Dies deutete auf eine größere Zahl an DNA-Basenpaaren im Menschen hin. Die ca. 319 kb des Menschen ließen sich einem Bereich von ca. 274 kb bei der Maus zuordnen, so dass die Maus ca. 45 kb weniger genomische DNA aufweist, was relativ betrachtet einem Unterschied von etwa 14% entspricht.

Dotplot

Percentage: 65; Window: 50; Min Quality: 10

Human_fin	Maus_fin	Total Diagonals
(1>319119)	(1>412827)	29793

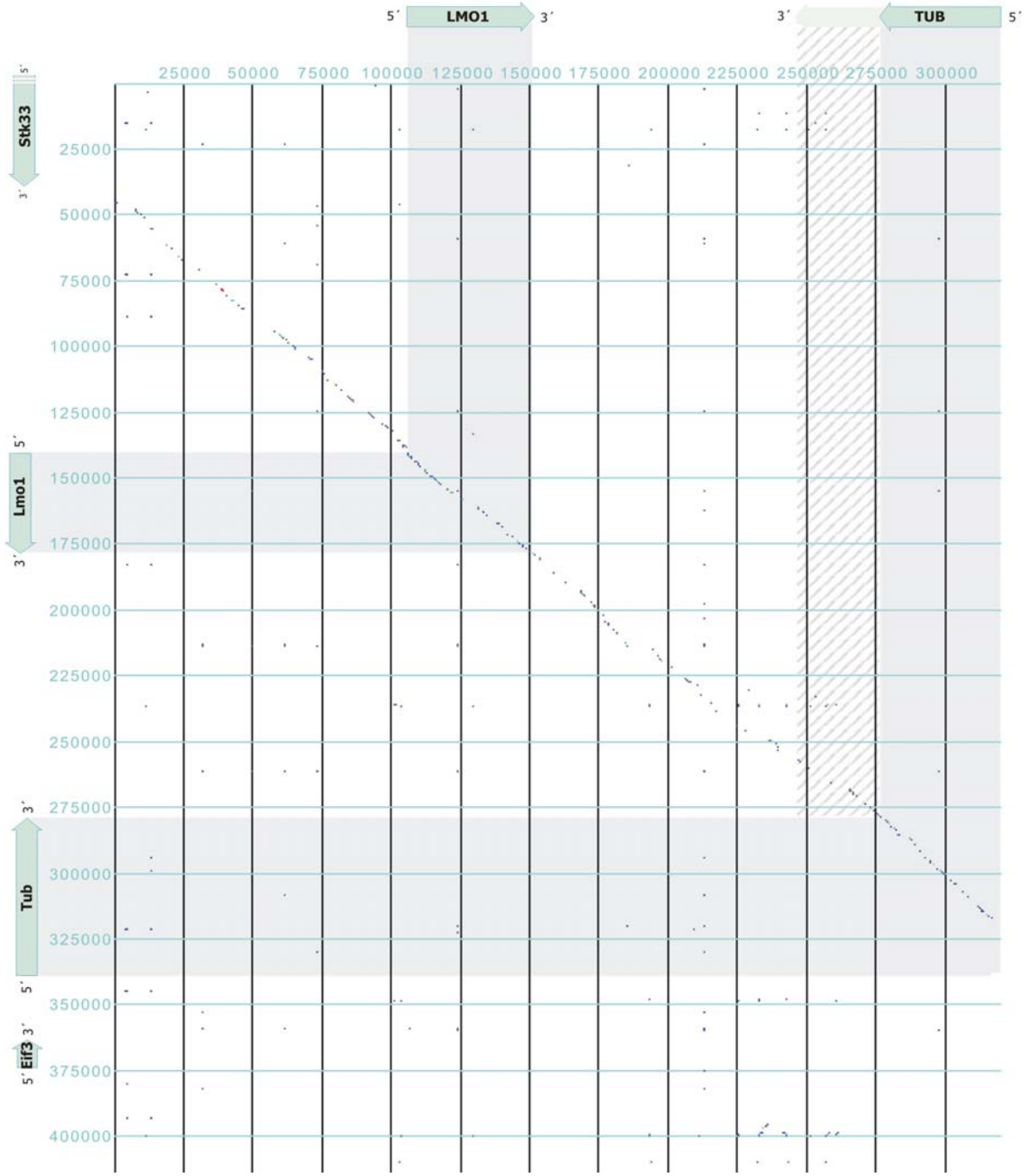


Abb. 24 **Dotplot-Analyse** mit den beiden genomischen Sequenzbereichen aus Mensch und Maus der vorliegenden Arbeit. Die fast durchgehende Diagonale setzt sich aus 29.793 Einzelbereichen zusammen, deren Homologie über eine Distanz von 50 Basenpaaren mindestens 65% beträgt. Die humangenomische Sequenz von über 319 kb (x-Achse) wird nur von insgesamt 274 kb im Mausgenom (y-Achse) repräsentiert. Das entspricht einer Reduktion auf ca. 86 %. Da der sequenzierte murine Genombereich mit insgesamt 412 kb um 138 kb länger ist als der humane, stand für ca. 45 kb am Flankenanfang der Maussequenz und ca. 93 kb am Flankenende derselben keine humane Vergleichssequenz zur Verfügung. Aus diesem Grund sind die murinen Gene Eif3 und

Stk33 nicht durch die Diagonale der Dotplot-Analyse repräsentiert. Der neue 3´-Bereich des humanen TUB-Gens konnte in der Maus nicht durch konservierte Sequenzen bestätigt werden, so dass dieser Bereich nur schräg schraffiert hervorgehoben wurde. Der übrige Genbereich des *TUB*|*Tub*-Gens ist wie die Gensequenz des *LMO1*|*Lmo1*-Gens grau unterlegt.

3.8.1.2 PIP-Analyse:

Bietet die Dotplot-Analyse anhand der Ausrichtung der Diagonalen eine Aussage über die genaue Lage der homologen Bereiche in den Sequenzen beider Spezies und über die Orientierung der Sequenzabschnitte, so diskriminiert diese Darstellungsform nicht den Grad der Homologie eines bestimmten Bereiches. Diese Information kann mit Hilfe einer PIP-Analyse dargestellt werden. Ab einer Homologie von 50% zur Vergleichssequenz wird jeder Bereich als eine horizontale Line über den Basenstrang der Referenzsequenz wiedergegeben. In der hier durchgeführten PIP-Analyse zwischen den genomischen Sequenzen von Mensch und Maus wurden sowohl die repetitiven Bereiche (siehe Kapitel 3.10.3) wie auch die bekannten kodierenden Exonabschnitte graphisch hervorgehoben. Außerdem wurden GC-reiche Sequenzabschnitte (≥60%) (siehe Kapitel 3.10.1) als solche gekennzeichnet. Als zusätzliches Ergebnis wurden sämtliche putativ kodierenden Bereiche der verschiedenen Exonvorhersage-Programme *FEXHB, MZEF, XPOUND, GRAIL2* und *GENSCAN* (siehe Kapitel 3.5.2) in die Graphik integriert, so dass ein direkter Bezug mit dem Grad der Konservierung dieser Bereiche getroffen werden konnte. Des weiteren wurden alle Bereiche der genomischen Sequenz farbig hervorgehoben, die Homologien zu EST-Sequenzen aufwiesen. So konnten in einer einzigen Grafik verschiedenste Ergebnisse der Sequenzanalyse mit dem Vorteil zusammengefasst werden, dass bestimmte Sequenzbereiche deutlich auffallen, die gleichzeitig durch mehrere Analyseprogramme identifiziert werden konnten.
Das Gesamtergebnis dieser PIP-Analyse zeigt ein hohes Maß der Sequenzkonservierung in beiden Genomen (siehe Abb. 25 a-c). Wie schon in der Dotplot-Analyse dargestellt, beschränkt sich diese Homologie nicht nur auf die kodierenden DNA-Abschnitte der charakterisierten Gene, sondern erstreckte sich auch über weite Bereiche der Intergen-Regionen. Von insgesamt 208 putativen Exonsequenzen – ohne die der beschriebenen Gene *LMO1*|*Lmo1* und *TUB*|*Tub* (229-21=208; siehe Tab. 11, bzw. Tab. 12) – konnten 58 durch EST-Sequenzhomologie verifiziert werden. Interessanterweise werden nur 33 dieser vermeintlich kodierenden Bereiche durch eine signifikante Interspezieshomologie (> 50%) bestätigt. Demgegenüber existieren insgesamt 36 Sequenzbereiche mit hoher Homologie von über 80% zwischen Mensch und Maus, die weder durch die Exonvorhersage noch durch die Homologie zu ESTs aufgefallen waren.

Betrachtet man die translatierten Exonbereiche der Gene *LMO1* und *TUB*, so zeigte sich, dass diese in beiden Spezies mit einer Homologie von über 75% streng sequenzkonserviert geblieben sind. Lediglich die untranslatierten Bereiche, wie das Exon 1a des Gens *LMO1* und Exon 12 des Gens *TUB*, wiesen einen Verlust dieser starken Konservierung auf. Auffällig für das *LMO1*-Gen war die hohe Konservierung des angrenzenden proximalen 5´-Bereiches von Exon 1a mit 98% über 171 bp (von Referenzsequenz nt. 107.018 bis 107.188) und 95% über 93 bp (Referenzsequenz nt. 107.190 und

107.282) außerhalb der cDNA-verifizierten Exonsequenzen. Ebenso zeigten sich allein in der ersten Intronsequenz des *LMO1*-Gens sieben Abschnitte mit einer Homologien zwischen 84% und 99% über mehrere hundert Basenpaare. Keines der verwendeten Exonvorhersageprogramme konnte für diese Bereiche eine genkodierende Funktion nahe legen; auch zeigten BlastN-Analysen keine Homologie zu Datenbankeinträgen bekannter EST-Sequenzen. Exemplarisch wurden diese Einzelbereiche des ersten *LMO1*-Introns detailliert in der nachfolgenden Tabelle 21 aufgeführt (siehe auch Abb. 25a).

Tab. 23 Hochkonservierte Bereiche in der genomischen Sequenz des LMO1-Intron 1. Dargestellt wurden die Genomabschnitte anhand ihrer Position in der jeweiligen Genomsequenz von Mensch und Maus unter Angabe der relativen Sequenzähnlichkeit in Prozent und der Bereichsgröße in Basenpaare.

Nr.	Position in humaner Genomsequenz	Position in muriner Genomsequenz	Grad der Homologie	Größe des konservierten Bereiches
1	115.304 – 115.682	149.927 – 150.316	95%	379 bp
2	116.830 – 117.099	151.359 – 151.628	99%	270 bp
3	121.579 – 122.268	155.398 – 156.092	93%	690 bp
4	125.686 – 125.909	158.118 – 158.341	84%	224 bp
5	131.354 – 131.688	161.322 – 161.776	91%	451 bp
6	133.165 – 133.402	162.972 – 163.212	87%	238 bp
7	142.102 – 142.281	171.684 – 171.870	92%	180 bp

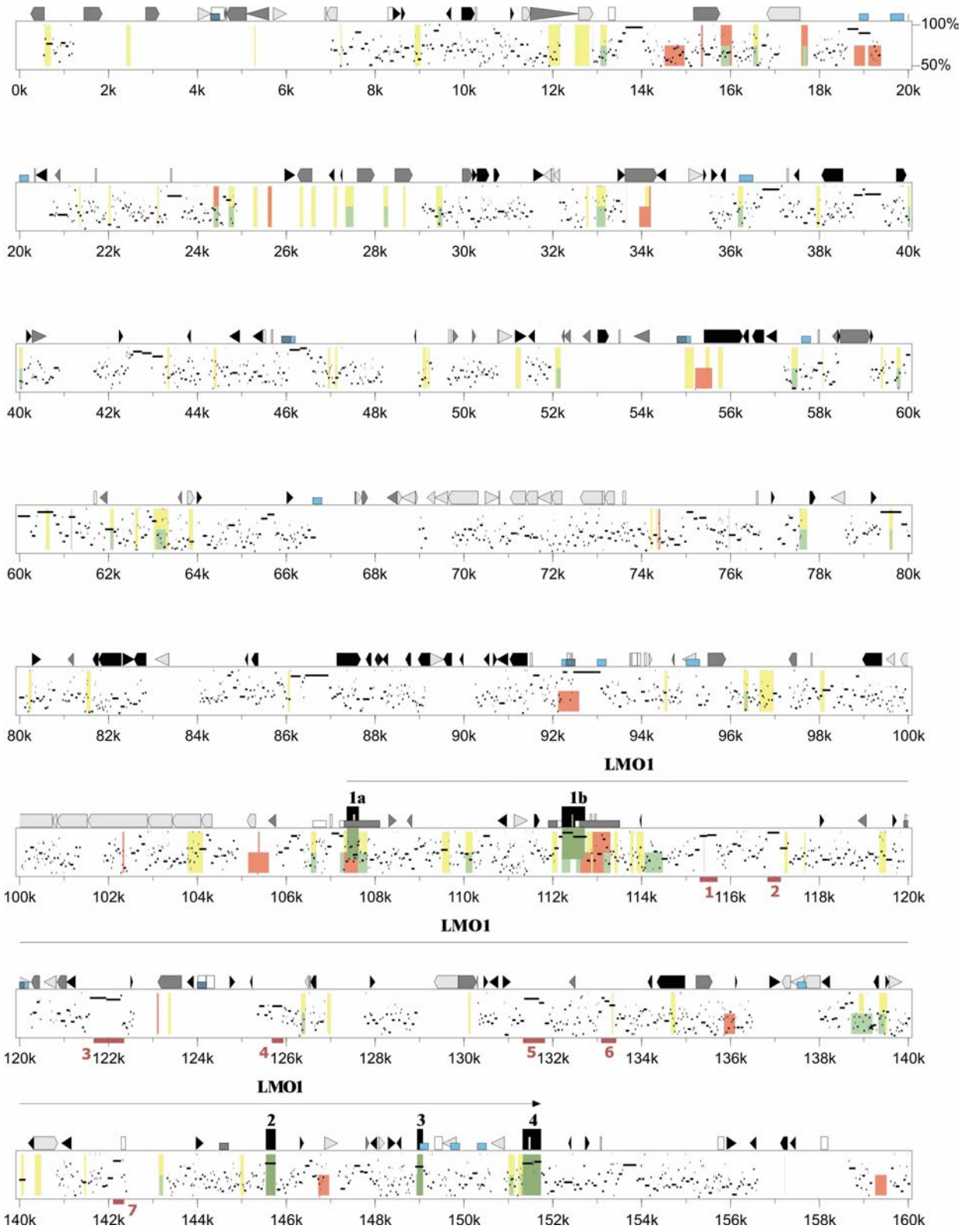


Abb. 25a PIP-Analyse (1. Teil) Base 1 bis 160.000 der humangenomischen Referenzsequenz verglichen mit der Maus-genomischen Referenzsequenz. Alle Sequenzbereiche mit einer Konservierung von 50 bis 100% wurden durch einen kleine Punkt, bzw. schwarzen Balken hervorgehoben. Exonbereiche, repetitive Elemente und GC-reiche Abschnitte sind durch Symbole in der Kopfzeile des PIP-Plots markiert. Farbige Bereiche heben cDNA-Homologien, putative Exonsequenzen, putative Promotor- und Poly-A-Bereiche hervor. Farbkodierung siehe übernächste Seite Abb. 25c. Die in Tab. 21 aufgeführten hochkonservierten Bereiche in LMO1-Intron 1 sind gemäß der Nummerierung in der Tabelle in rot mit der entsprechenden Ziffern versehen.

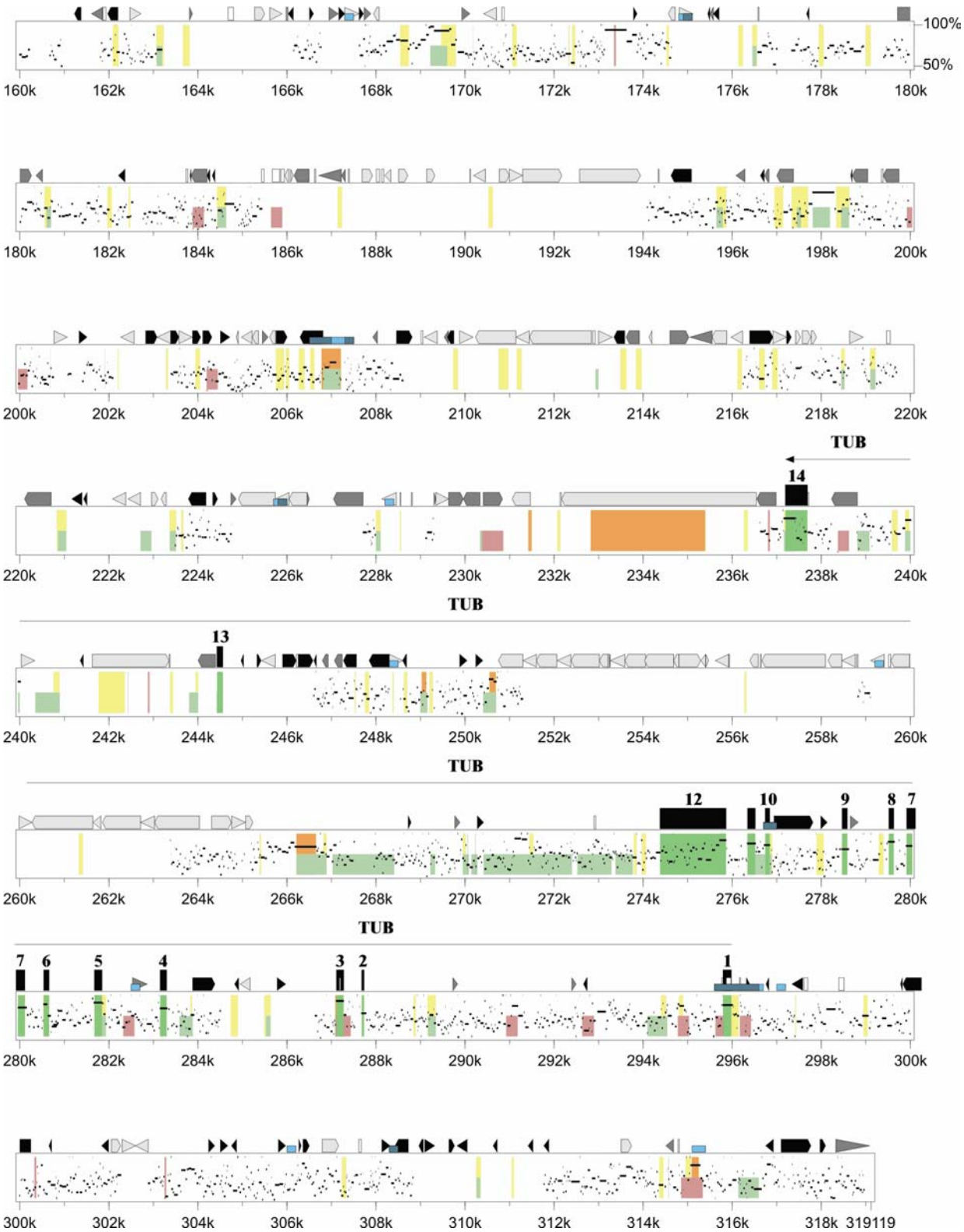


Abb. 25b **PIP-Analyse (2. Teil)** Base 160.001 bis 319.199 der humangenomischen Referenzsequenz verglichen mit der Maus-genomischen Referenzsequenz. Alle Sequenzbereiche mit einer Konservierung von 50 bis 100% wurden durch einen kleine Punkt bzw. schwarzen Balken hervorgehoben. Exonbereiche (schwarze Balken), repetitive Elemente (Dreiecke, bzw. Pfeile) und GC-reiche Abschnitte sind durch Symbole in der Kopfzeile des PIP-Plots markiert. Farbige Bereiche heben cDNA-Homologien, putative Exonsequenzen, putative Promotor- und Poly-A-Bereiche hervor. Farbkodierung siehe nächste Seite, Abb. 25c.

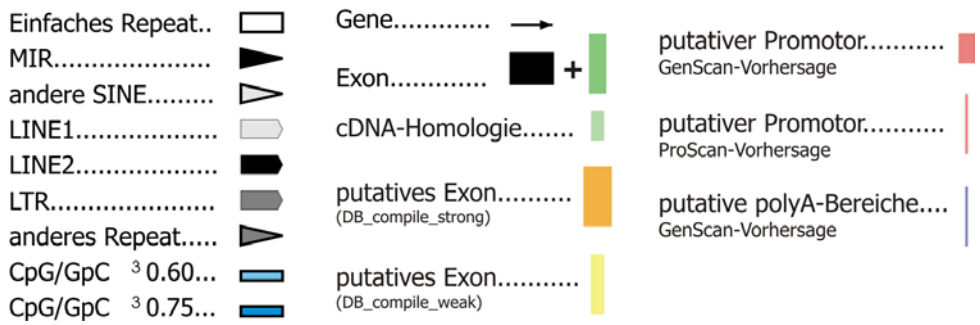


Abb. 25c **Legende zu den vorigen PIP-Plot-Grafiken.** Die repetitiven Elemente wurden mit dem Programm RepeatMasker ermittelt und entsprechend ihrer Klassifizierung nach einfache „Repeats", SINES, LINES, LTR („long terminal repeats") und anderen Sequenzwiederholungen eingezeichnet. Die cDNA-Homologien wurden über BlastN-Analyse gegen die Datenbank dbEST ermittelt. Die putativen Exons wurden mit den Programmen FEXHB, MZEF, XPOUND, GRAIL2 und GENSCAN ermittelt, in der Datenbank „compile_weak" zusammengefasst und in den Plot gelb eingezeichnet. Sequenzbereiche, die von drei oder mehr Programmen vorhergesagt wurden (= Datenbank „compile_strong"), sind orangefarben dargestellt. Putative Promotorbereiche wurden mit den Programmen GENSCAN und ProScan berechnet und sind in roter Farbe in der Graphik vermerkt. Putative Poly-A-Bereiche, ebenfalls mit dem Programm GENSCAN identifiziert, sind als graublaue Linien eingezeichnet.

3.8.2 Vergleich Mensch – Fugu

3.8.2.1 Dotplot- und PIP-Analyse

Um die biologische Relevanz der konservierten Genomabschnitte für den Organismus zu bekräftigen, wurde in einem zweiten Interspeziesvergleich die humane Genomsequenz mit genomischen Fragmenten einer weiteren Tierspezies verglichen, die sich während der Evolution schon wesentlich früher vom Menschen abgespaltet hat als die Maus. Das Genom des Kofferfischs *Fugu rubripes* erschien hierfür sehr geeignet, da diese Spezies zum einen in einem eigenen, weit fortgeschrittenen Genomprojekt sequenziert wurde und zum anderen die Genomgröße im Vergleich zum Menschen bei ähnlicher Genausstattung um das 7,5-fache kleiner war (Elgar *et al.*, 1996).

Für eine zweite Dotplot- und PIP-Analyse wurden zunächst zur humanen Sequenz orthologe Bereiche aus dem Fugu-Genomprojekt ermittelt. Die hierzu identifizierten Fugu-Sequenzen wurden über eine BlastN-Analyse in der "scaffolds"-Datenbank mit der humangenomischen Consensus-Sequenz unter der URL <http://fugu.hgmp.mrc.ac.uk> bestimmt. Für die Suche wurde die Menschsequenz in 50.000 bp Abschnitte unterteilt und abschnittsweise untersucht. Die insgesamt 11 ermittelten Vergleichssequenzen aus *Fugu* (siehe Abb. 26) wurden zu einer 239.131 bp langen Consensussequenz aneinandergereiht und für die Dotplot- und PIP-Analyse eingesetzt. Die Reihenfolge der genomischen *Fugu*-Sequenzen entsprach dabei, wie die Dotplot-Grafik (Abb. 26) zeigt, nicht genau der chromosomalen Anordnung. Für die PIP-Analyse selbst war diese Ungenauigkeit aber nicht von Relevanz, da hier nicht die Positionen der Vergleichssequenz berücksichtigt und dargestellt wurden Ebenso wurde über die Ähnlichkeit zu den Genfamilienmitglieder von *LMO1* und *TUB* auch genomische

Fugu-Sequenzen identifiziert, die aus den orthologen Chromosomenabschnitten der Genfamilien-mitglieder stammen.

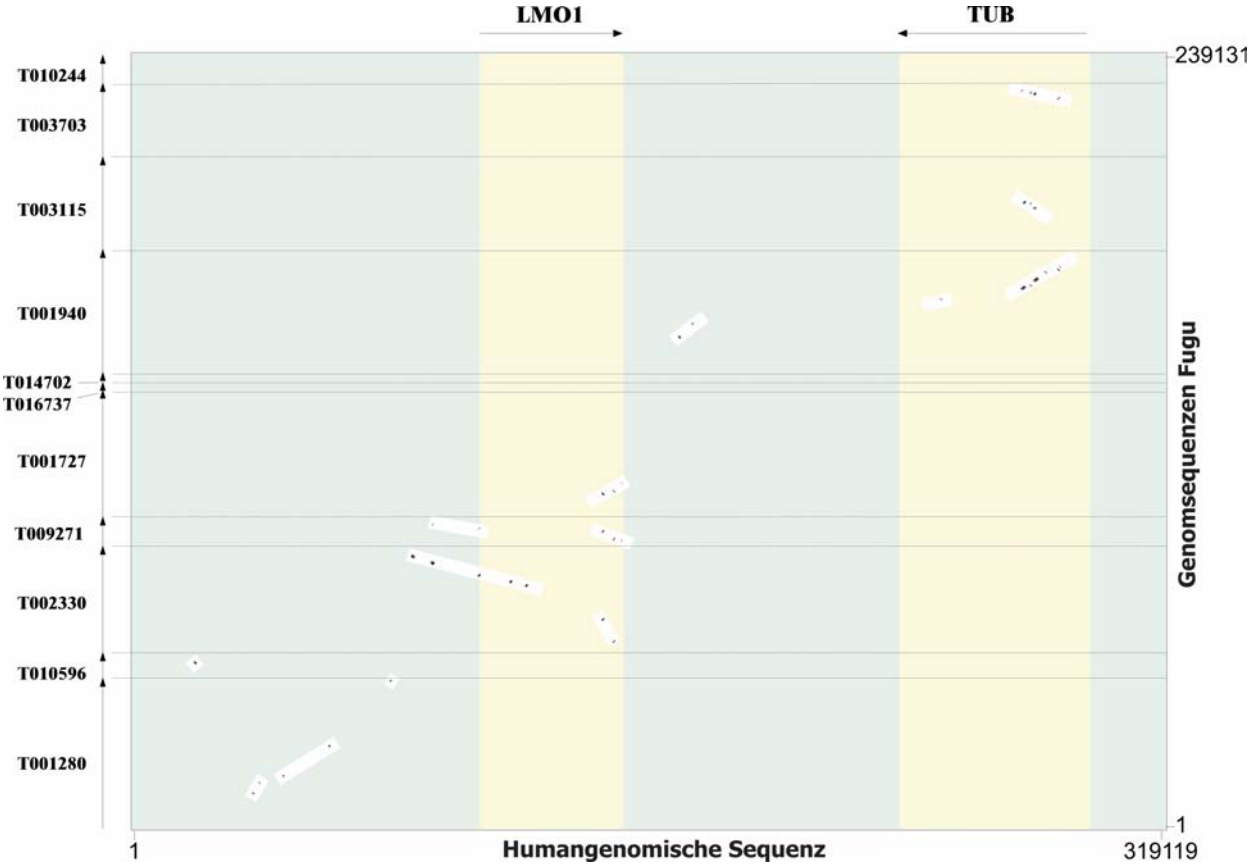


Abb. 26 **Dotplot-Analyse der genomischen Fugu-Sequenzen** mit der humangenomischen Sequenz**.** Die Fugu-Sequenz (Abszisse) ist aus verschiedenen Genomabschnitten zusammengesetzt, die Acc. Nr.: „T0xxx" sind am linken Rand der Grafik in Höhe des jeweiligen Abschnitts angegeben. Die humane Referenzgenomsequenz (Ordinate) ist in Gesamtlänge von 319.119 bp aufgeführt. Gelb hervorgehoben sind die transkribierten Gen-Bereiche. Zur besseren Sichtbarkeit wurden die Homologiebereiche mit den „Dots" durch einen weißen Hintergrund hervorgehoben.

Die Auswertung des Vergleichs zeigte, dass insgesamt 27 Bereiche angesprochen werden konnten, deren Homologie zwischen den beiden Spezies über 50% lag (Abb. 26). 14 Bereiche davon bezogen sich auf die Exons der beiden Gene *LMO1* und *TUB*.

Betrachtete man das Gen *LMO1* genauer, so fiel auf, dass das Exon 1b keine Sequenzkonservierung aufwies. Das Exon 1a der alternativen Spleißform zeigte interessanterweise nur außerhalb des 5´untranslatierten Bereichs der transkribierten Sequenz eine signifikante Homologie; ein Bereich der auch proximal zum *in silico* bestimmten Promotorbereich lag. Der kodierende Bereich des Exon 1 fiel selbst nicht auf. Ebenso wenig war der 3´untranslatierte Bereich konserviert erhalten geblieben. Die Konservierung brach bereits 20 bp vor dem humanen Stoppcodon noch im offenen Leserahmen ab. Es fiel weiter auf, dass in der ersten Intronsequenz sich zwei Abschnitte (#10 und #11 in Tab. 21)

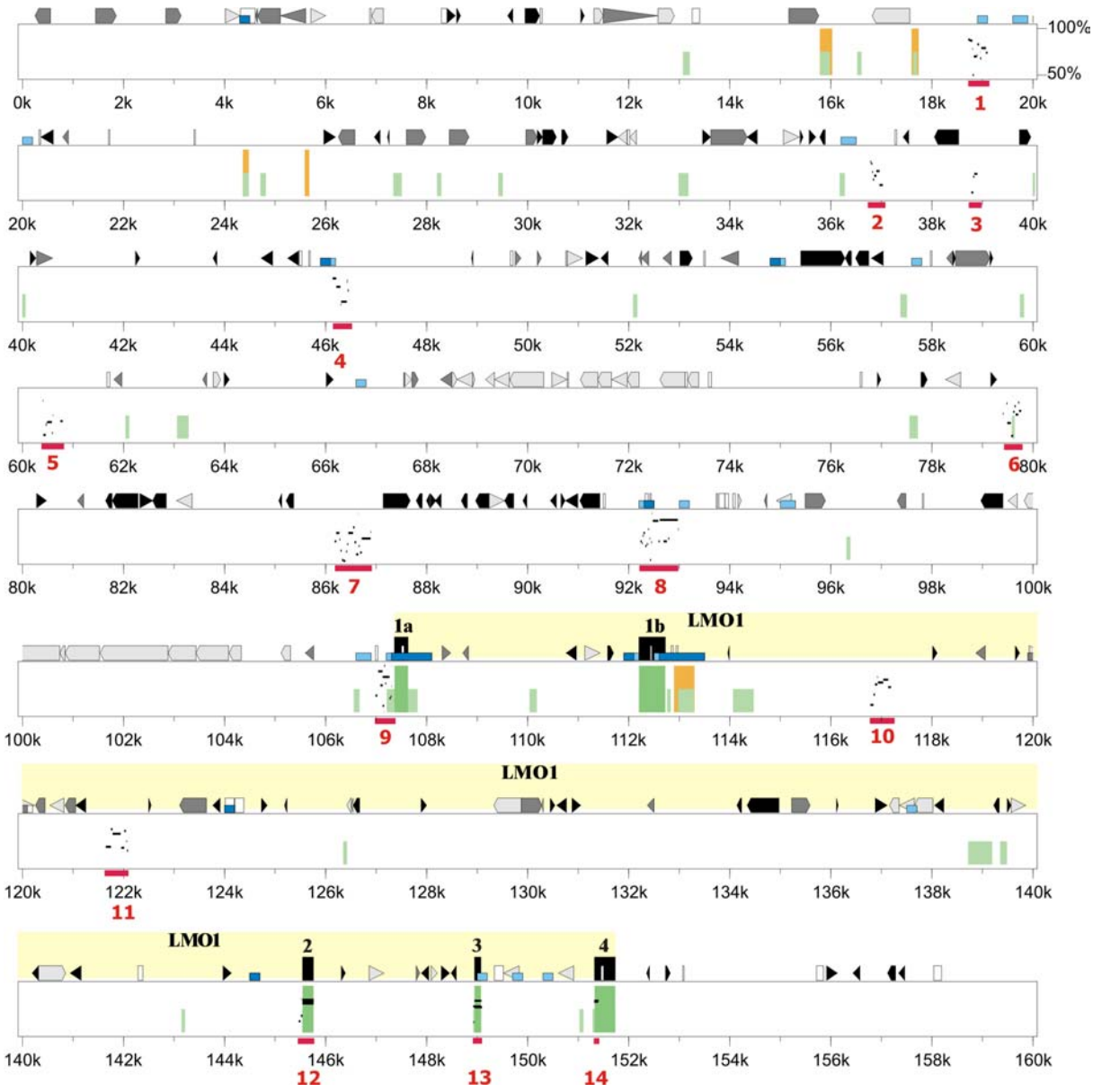
befanden, deren Sequenzähnlichkeit bis zu 88% beträgt und die keinerlei Homologie zu bekannten cDNA-Sequenzen aufwiesen und bisher noch nicht mit dem *LMO1*-Gen in Zusammenhang gebracht wurden. Die Exons 2 und 3 fanden sich in voller Länge mit einer Homologie von nahezu 86% bzw. 84% im Kofferfisch wieder (siehe Tab. 21).

Für das Gen *TUB* zeigten alle Exons bis auf das erste (Exon 1c) und das fünfte eine über 80%ige Homologie. Auch hier ist die fehlende Konservierung für das Exon 5 interessant, da dieses Exon durch alternatives Spleißen entfernt werden kann und somit scheinbar nicht essentiell für die Genfunktion ist. Exon 12 (#18) ist innerhalb des kodierenden Bereichs bis hin zum Stoppcodon mit 86% konserviert geblieben, der 3´UTR weist keine Ähnlichkeiten mehr auf. Dagegen zeigen die Introns 6, 7 und 11 in Fragmentabschnitten von 20 bis 30 bp eine Homologie zur genomischen *Fugu*-Sequenz von 65% bis 85%. Die kürzeren Intronsequenzen von *Fugu* sind in diesen drei Bereichen in voller Länge mit der angegebenen Übereinstimmung zur Humansequenz erhalten geblieben. Bedeutsam ist diese Stabilität der Sequenzinformation insbesondere für das Intron 11 (#19: Abb. 27b), da dessen Spleißakzeptorstelle der Ort für die Tubby-Mutation der Maus ist.

Allen weiteren konservierten Bereiche wurde versucht mit Hilfe der *in silico*-Analyse cDNA-Sequenzen zuzuordnen. Lediglich die Homologiebereiche #6 und #15 zeigten eine jeweils über 60% liegende Übereinstimmung mit einer cDNA-Sequenz der Maus (Acc.-Nr. AV162648), bzw. einer cDNA-Sequenz der Ratte (Acc.-Nr. AA859374). Der Homologiebereich #17 wies eine Übereinstimmung mit der humanen Sequenz Acc.-Nr. AU118609 auf; gleichzeitig wurde dieser Abschnitt auch von mehreren Exonvorhersageprogrammen als kodierend eingestuft. Die verbleibenden 10 Abschnitte, die ebenfalls im Interspeziesvergleich mit der Maus durch eine über 60%ige Konservierung aufgefallen waren, konnten über den Homologievergleich nicht näher charakterisiert werden (siehe Tab. 20).

Tab. 24 **Zusammenfassung aller konservierten Sequenzabschnitte** der humangenomischen Sequenz zur zusammengefügten Referenzsequenzen aus dem Fugu-Genom. Die Homologie einiger Exonbereiche zu mehreren genomischen Fugu-Sequenzen ist auf die Ähnlichkeit zu weiteren Mitglieder der LMO-Genfamilie, bzw. zu den TULP-Genen zurückzuführen. Als orthologe Fugu-Genomsequenzen dürfen aufgrund der höchsten Übereinstimmung die Datenbankeinträge T001280, T002330 und T001940 angesehen werden. Die nummerierten Bereiche finden sich auch in nachfolgender Abb. 27a/b wieder.

Region #	Fugu-Scaffold-Sequenz <http://fugu.hgmp.mrc.ac.uk>	Konservierter Bereich in der humanen Genomsequenz	Homologie zw. Fugu und Mensch	Funktion/Gen bzw. andere Interspezieshomologie
1	T001280	nt 18.721 – nt 19.097 = 378 bp	47% - 88%	?
2	T001280	nt 36.783 – nt 37.029 = 248 bp	62% - 87%	?
3	T001280	nt 38.780 – nt 38.902 = 124 bp	54% - 74%	?
4	T001280	nt 46.139 – nt 46.456 = 319 bp	64% - 92%	?
5	T001280	nt 60.402 – nt 60.795 = 396 bp	54% - 91%	?
6	T001280	nt 79.405 – nt 79.765 = 361 bp	47% - 89%	AV162648 (Mus) (66% auf 66bp)
7	T002330	nt 86.177 – nt 86.899 = 723 bp	49% - 100%	?
8	T002330 T009271	nt 92.230 – nt 92.985 = 756 bp nt 92.688 – nt 92.801 = 114 bp	50% - 96% 47% - 73%	?
9	T002330 T009271	nt 106.987 – nt 107.321 = 335 bp nt 107.083 – nt 107.197 = 115 bp	65% - 100% 57% - 100%	5´UTR - LMO1 Ex1a
10	T002330	nt 116.782 – nt 117.192 = 411 bp	58% - 88%	?
11	T002330	nt 121.648 – nt 122.089 = 442 bp	65% - 88%	?
12	T001727 T002330 T009271	nt 145.467 – nt 145.762 = 296 bp nt 145.511 – nt 145.762 = 252 bp nt 145.541 – nt 145.759 = 219 bp	62% - 86% 68% - 85% 81%	LMO1 Ex2
13	T001727 T009271 T002330	nt 148.944 – nt 149.083 = 140 bp nt 148.919 – nt 149.078 = 140 bp nt 148.925 – nt 148.942 = 18 bp	84% 78% 61% - 77%	LMO1 Ex3
14	T009271 T001727	nt 151.320 – nt 151.404 = 85 bp nt 151.317 – nt 151.391 = 75 bp	84% 83%	LMO1 Ex4
15	T001940	nt 169.179 – nt 169.638 = 460 bp	50% - 96%	AA859374 (Rattus) (65% auf 126 bp)
16	T001940	nt 173.297 – nt 173.568 = 272 bp	67% - 80%	?
17	T001940	nt 250.567 – nt 250.695 = 129 bp	58% - 82%	AU118609 (Human) 100% auf 146 bp
18	T001940 T003703	nt 275.727 – nt 275.864 = 138bp nt 275.702 – nt 275.799 = 98 bp	86% 79% - 85%	TUB Ex12
19	T001940 T003115	nt 276.324 – nt 276.520 = 197 bp nt 276.356 – nt 276.520 = 357 bp	80% 70%	TUB Ex11
20	T001940	nt 276.726 – nt 276.853 = 128 bp	70%	TUB Ex10
21	T001940 T003115 T003703	nt 278.460 – nt 278.613 = 154 bp nt 278.462 – nt 278.591 = 130 bp nt 278.462 – nt 278.600 = 139 bp	67% - 73% 67% - 100% 62%	TUB Ex9
22	T001940 T003703 T003115	nt 279.514 – nt 279.631 = 118 bp nt 279.509 – nt 279.641 = 133 bp nt 279.513 – nt 279.638 = 126bp	79% 75% 67%	TUB Ex8
23	T001940 T003703 T003115	nt 279.914 – nt 280.054 = 141 bp nt 279.913 – nt 280.067 = 155 bp nt 279.916 – nt 280.055 = 140 bp	77% 72% 71%	TUB Ex7
24	T001940	nt 280.529 – nt 280.659 = 131 bp	79%	TUB Ex6
25	T001940	nt 283.148 – nt 283.330 = 183 bp	42% - 94%	TUB Ex4
26	T001940 T003703	nt 287.132 – nt 287.282 = 151 bp nt 287.109 – nt 287.282 = 174 bp	76% 73%	TUB Ex3
27	T001940	nt 287.675 – nt 287.748 = 74 bp	74%	TUB Ex2



Fugu PIP-Plot Legende

Einfaches Repeat..
MIR.....
andere SINE.....
LINE1.....
LINE2.....
LTR.....
anderes Repeat.....
CpG/GpC ³0.60...
CpG/GpC ³0.75...

Gene.....
Exon..... +
cDNA-Homologie.....
putatives Exon.....
(DB_compile_strong)
konservierter Bereich
in Fugu.....

Abb. 27a PIP-Analyse (1. Teil) der humangenomischen Sequenz von Nukleotid 1 bis 160.000 mit genomischen Sequenzen aus dem Fugu-Genomprojekt. Mit rotem Balken versehen sind alle Bereiche aus Tab. 22, die Homologien zur Genomsequenz von Fugu rubripes aufweisen. Zur Orientierung und Charakterisierung der Humansequenz sind durch Symbole repetitive, GC-reiche Bereiche und Gen-kodierenden Abschnitte hervorhoben. Zudem sind Sequenzen mit Homologie zu annotierten mRNAs (hellgrün) und putative Exonbereichen (orange) farbig unterlegt. Die detaillierte Legende zur Graphik ist links zu finden.

Abb. 27b **PIP-Analyse (Teil 2)** der humangenomischen Sequenz von Nukleotid 160.001 bis 300.000 mit genomischen Sequenzen aus dem Fugu-Genomprojekt. Mit rotem Balken versehen sind alle Bereiche aus nachfolgender Tab. 22, die Homologien zur Genomsequenz von Fugu rubripes aufwiesen. Die Legende zur Graphik befindet sich auf der vorhergehenden Seite.

3.9 Homologievergleich ausgewählter paraloger Chromosomenregionen

3.9.1 PIP-Analyse

Aufgrund der hohen Konservierung der Genomsequenzen im Interspeziesvergleich wurde mit Hilfe der PIP-Analyse versucht, auch das genomische Umfeld der paralogen Gene auf Chromosom 12 zu analysieren, um etwaige funktionell wichtige Konservierungen in Intergen- bzw. Intronbereichen charakterisieren zu können. Schon bei der Untersuchung der genkodierenden mRNA-Sequenzen zeigten sich in den BlastN-Analysen hohe Homologien zum humanen Chromosom 12. So konnten für die humanen Gene *TUB*, *LMO1* und auch für das murine Gen *Eif3* genomische Bereiche auf dem kurzen Arm von Chromosom 12 angesprochen werden, die auf weitere Familienmitglieder der drei Gene schließen ließen.

Für die Chromosomenregion 12p12.3 konnten über die Homologie des *LMO1*-Gens zum *LMO3*-Gen, das in den Datenbanken als „Neuronen-spezifischer Transkriptionsfaktor DAT1" (Acc.-Nr. AF258348) geführt wird, die genomischen Sequenzen Acc.-Nr. AC007529 (BAC69C13: 147.239 bp) und Acc.-Nr. AC007552 (BAC424M22: 113.951 bp) identifiziert werden. Interessanterweise zeigte sich eine Sequenzkonservierung nicht nur für die kodierenden Abschnitte der Exons 2 und 3, und den 5´-Bereich des Exons 4, sondern auch für die Intronsequenz 3 zwischen den beiden Exons 3 und 4. Die Sequenzanalyse zeigte allerdings eine 92%ige Ähnlichkeit zu einem Vertreter der AluSx- Subfamilie, so dass es sich hierbei um ein konserviertes Alu-„Repeat" handeln könnte (siehe Abb. 28 - A).

In der Chromosomenregion 12p13.33 erstreckt sich die 1.482 bp lange mRNA-Sequenz des *TULP3*-Gens (Acc.-Nr. NM_003324) und überspannt mit 11 Exons einen genomischen Bereich von 49,46 kb. Als konserviert zur Chromosomenregion 11p15.3 konnten mit Hilfe der genomischen Sequenz Acc.-Nr. AC005911 (BAC372B4: 100.701 bp) die *TUB*-Exons 3, 7 und 9 bis 12 angesprochen werden. Mit einer Homologie von 68% bis 80% hoben sie sich deutlich von den nicht-konservierten Intronbereichen ab (siehe Abb. 28 - B). Auch das genomische Umfeld im 5´- und 3´-Genbereich fiel mit keinerlei Homologien auf.

In der Chromosomenregion 12p13.32, in etwa 2,12 Megabasen Entfernung zum *TULP3*-Gen, zeigte sich für die exprimierten Bereiche des neubeschriebenen murinen *Eif3s5*-Gens eine Konservierung zur Genomsequenz AC005906 (BAC429A20: 185.952 bp). Die Exons 5 bis 8 und der 5´-Bereich des Exons 1 wiesen zu dieser Chromosom 12-genomischen Sequenz eine Homologie zwischen 77% und 87% auf. Der Bereich der internen Exons 2, 3 und 4 stellte sich bei näherer Betrachtung der humanen Genomsequenz als deletiert heraus und konnte somit im PIP-Blot auch nicht in Erscheinung treten (siehe Abb. 28 – C1).

Mit Hilfe der murinen Genomsequenz um das *Eif3s5*-Gens konnte darüber hinaus auch für die Chromosomenregion 2p16.1 eine Konservierung zwischen 73% und 90% zu allen acht Exons nachgewiesen werden. Die Analyse dieses Chromosom 2-Bereiches, repräsentiert durch die humane

Genomsequenz Acc.-Nr. AC007250 (BAC 334G22: 180.557 bp), zeigte, dass die gesamte Gensequenz von einem einzigen Exon, einer 1.234 bp langen durchgehenden mRNA-Sequenz (Acc.-Nr. XM_010886), repräsentiert wurde. Auch in diesem Bereich erstreckte sich die Konservierung nur auf die kodierenden Sequenzabschnitte, die Intronbereiche zeigten keinerlei Homologien (siehe Abb. 28 – C2).

Als dritten chromosomalen Locus konnte über die Homologie zur murinen *Eif3*-Sequenz ein genomischer Bereich auf dem humanen Chromosom 17 angesprochen werden. Hier zeigten die Genomabschnitte der annotierten Sequenz Acc.-Nr: AC087648 (BAC 21M3: 74.105 bp) Homologie zu den Exons 2, 3 und 6, 7 und 8. Auffällig war in diesem Fall ebenfalls eine Konservierung in den Intronsequenzen 2, 3, 5, 6 und 7. Die fehlende Konservierung darüber hinaus dürfte mit der noch unvollständigen Sequenzierung des BAC-Klons zusammenhängen und durch eine unvollständige Genomgenomsequenz erklärt werden (siehe Abb. 28 – C3).

Nach Abschluss der Arbeiten konnte auch die orthologe Genomsequenz des humanen Chromosoms 11p15.3, repräsentiert durch die Sequenz Acc.-Nr: AC124259 (BAC 21N2, 99.973 bp) dem murinen Gen-Locus gegenübergestellt und dessen Homologie mit den der paralogen Regionen im menschlichen Genom verglichen werden. Es zeigte sich, dass die Konservierung der kodierenden orthologen Sequenzanteile nicht höher lag als zu den paralogen Abschnitten. Auffällig ist aber, dass die Intron- und Intergen-Bereiche eine deutlich größere Homologie besitzen (siehe Abb. 28 – C4). Warum der chromosomale Bereich auf Chromosom 17 dieser Abstufung nicht folgt, konnte nicht geklärt werden.

Die Ergebnisse dieser PIP-Analysen mit verschiedenen paralogen Genomabschnitten unterschieden sich aber deutlich von Interspeziesvergleichen der orthologen Genombereiche andere Modellorganismen, wie beispielsweise *Fugu*, die weitaus mehr konservierte Genombereiche, auch außerhalb der kodierenden Sequenzanteile, aufwiesen (Kap. 3.8.2).

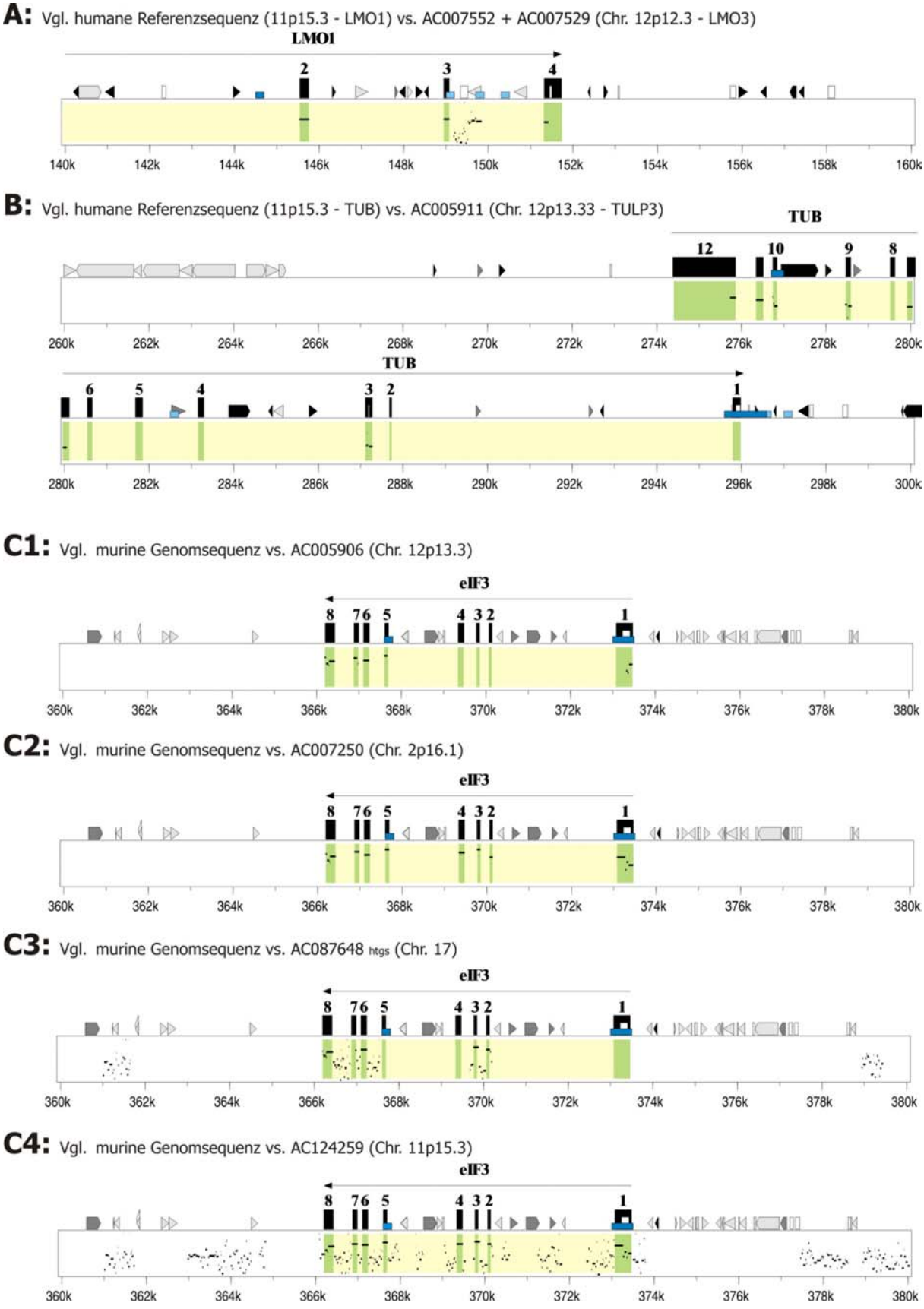


Abb. 28 PIP-Analyse mit den kodierenden Genomabschnitten der Gene LMO1, TUB und Eif3s5 und mit paralogen Genomsequenzen aus den humanen Chromosomenregionen 12p12 und 12p13, bzw.

der Chromosomenregion 2p16.1 und Chr. 17. A: Vergleich der LMO1-Genregion mit den genomischen Sequenzen Acc.-Nr. AC007552 (113.951 bp) und Acc.-Nr. AC007529 (147.239 bp) aus der Chromosomenregion 12p13.3, in der das Gen LMO3 (wird in den Datenbanken unter DAT1 geführt) lokalisiert ist. Die Exons 1a und 1b sind nicht konserviert (Homologie unter 50%) und somit nicht dargestellt. Interessanterweise ist ein Teil der Intron 3-Sequenz konserviert erhalten, der weder für LMO1 noch für LMO3 kodierende Funktion besitzt. Aufgrund der Homologie zur Alu-Familie dürfte es sich hier um einen konservierten repetitiven Bereich handeln. B: Vergleich der TUB-Genregion mit der Chromosom 12-Sequenz Acc.-Nr. AC005911 des TULP3-Gens. Konserviert in der genomischen DNA sind die Exons 3, 7 und 9 bis 12. C1: Die murine Genomsequenz zeigte für die Bereiche der Eif3-Exons ebenfalls eine Konservierung zur Chromosomregion 12p13. Die fehlende Homologie zu den Eif3-Exon 2 bis 3 konnte durch eine Deletion dieses Abschnittes in der Chromosom 12-Sequenz der Annotierung AC005906 erklärt werden. C2: Eine weitere Homologie der murinen Eif3-Exons zeigte sich zur Chromosomenregion 2p16.1. Die Genomsequenz Acc.-Nr: AC007250 zeigte zu allen Exonbereichen eine Konservierung. Die spätere Analyse dieses Chromosomenbereiches lies einen durchgehenden, intronlosen kodierenden Bereich erkennen. C3: Auffällige Homologien ließen sich auch für einen genomischen Klon (Acc.-Nr: AC087648) nachweisen, der auf Chromosom 17 kartiert worden war. Interessanterweise erstreckten sich hier die Konservierung auf die Intronbereiche zwischen den Exons 2 bis 4 und 5 bis 8. C4: Durch die neue Annotierung eines genomischen Klon (Acc.-Nr: AC124259) für die humane Region 11p15.3, ermöglichte auch die Gegenüberstellung der genomischen Maussequenz mit orthologen humanen Region. Es zeigte sich, dass die Konservierungen nicht ausgeprägter, insbesondere in den kodierenden Exonbereichen, waren als bei dem Vergleich mit den paralogen Regionen.

3.10 Vergleich der Genomorganisation

3.10.1 CpG-Analyse

Die statistische Auswertung der Basenpaarzusammensetzung ergab für die humane Genomsequenz einen GC-Gehalt von durchschnittlich 48,01%. Dieser Prozentwert entspricht der unteren Grenze der **Isochoren-Familie H2**. Hierbei handelt es sich um eine Klassifizierung von genomischen DNA-Bereichen in Bezug auf ihren durchschnittlichen GC-Basenanteils mit einer ungefähren Ausdehnung von über 300 kb (Bernardi *et al.*, 2000). Dieses Klassifizierungssystem unterscheidet insgesamt fünf verschiedene Klassen (L1, L2, H1, H2, H3) mit unterschiedlichem GC-Gehalt, wobei die relative Verteilung der oberen Klassen aus der Sicht des Gesamtgenoms immer kleiner wird. So beträgt der Anteil der Klasse H2 mit einem GC-Anteils von 48% bis 52% am Gesamtgenom nur 7,5%, insgesamt befinden sich aber 32% aller Gene in diesen genomischen Abschnitten (Zoubak *et al.*, 1996)

Der GC-Gehalt der murinen Genomsequenz lag mit 46,72% um 1,3% unter dem Wert des Menschen. Der verminderte GC-Anteil führte daher zur Eingruppierung in die nächst niedrigere Isochoren-Familie H1, die die Spanne von 44% bis 48% an GC-Nukleotiden beschreibt. Betrachtete man die genomische Architektur der Gene, so fiel auf, dass kodierende Regionen meistens durch einen erhöhten GC-Gehalt charakterisiert sind. Der transkribierte Bereich des *LMO1*-Gens besitzt z.B. im Menschen einen durchschnittlichen GC-Gehalt von 52,86%; in der Maus sind es 51,06%. Einen noch höheren Wert weist die transkribierte Sequenz des *TUB*-Gens mit 54,67% im Mensch, bzw. 51,53% in der Maus auf. Insbesondere die Promotorbereiche der Gene zeichneten sich durch das Vorhandensein von sog. CpG-Inseln aus, deren GC-Gehalt bis über 70% ansteigen konnte, wie im 5´UTR des *LMO1* Exons 1a mit 73,6% über 878 bp gezeigt werden konnte. Der Interspeziesvergleich belegte, dass dieser tendenziell erhöhte GC-Gehalt evolutiv erhalten ist und sich am deutlichsten an den konservierten CpG-Inseln im 5´UTR-Bereich der Gene zeigt. Die CpG-Inseln der beiden genomischen Consensussequenzen wurden mithilfe des Programms *CPG-FINDER* bestimmt und in der Tab. 22 zusammengefasst, bzw. in Abb. 29 grafisch im GC-Plot dargestellt. Insgesamt konnten für die humangenomische Consensussequenz 13 Sequenzbereiche mit Längen größer 100 Basenpaaren charakterisiert werden, deren GC-Gehalt über 60% liegt. Für die Maus-genomische Sequenz waren es insgesamt 7 Sequenzabschnitte mit ebenso hohem GC-Gehalt.

Eine Besonderheit in der untersuchten Region war der Bereich zwischen Nukleotid 185.000 und 263.000 der humangenomischen Consensussequenz, der sich durch eine signifikante **Absenkung des GC-Gehaltes**, bzw. durch einen respektive erhöhten AT-Gehalt auszeichnete. In diesem Abschnitt über 78 Kilobasen zwischen den Genen *LMO1* und *TUB* betrugt der durchschnittliche GC-Gehalt im Menschen nur noch 39,73% (siehe Abb. 29: „GC-arme Region"). Im Mausgenom war diese „GC-Senke" mit einem GC-Gehalt von 40,96% zwischen Nukleotid 213.00 bis 266.000 der Maus-genomischen Consensussequenz nicht ganz so deutlich ausgeprägt. Außerdem war dieser Bereichsabschnitt mit 53 kb in der Maus im Vergleich zur humanen DNA um 25 kb an

Sequenzinformation (= 32%) verkürzt. Dieser Genombereich war bereits bei der Dotplot-Analyse (vgl. Kap. 3.8.1.1) durch eine verhältnismäßig schwache Konservierung zwischen Mensch und Maus in Erscheinung getreten (siehe Abb. 24; siehe auch Kap. 4.4.2).

Tab. 25 **Ergebniszusammenfassung der CpG-Analyse.** Es wurden alle Sequenzabschnitte beider genomischer Referenzsequenzen aus Mensch und Maus zusammengestellt, deren GC-Gehalt in einem Bereich größer 100 bp über 60% lag. Zu jedem Abschnitt wurde die genaue Position innerhalb der Consensussequenz mit Angabe der Länge, des durchschnittlichen GC-Gehaltes in Prozent, des entsprechenden Qualitätswertes, der Anzahl der CpG-Dinukleotiden und der etwaigen Korrespondenz zu bekannten genkodierenden Sequenzabschnitten angegeben. Besonders signifikante Bereiche mit einem Qualitätswert über 400 wurden grau unterlegt und im nachfolgenden GC-Plot in Abb. 29 mit der hier vergebenen Nummerierung hervorgehoben. Die CpG-Inseln #6 und #7 der Maus lagen außerhalb der sequenzierten Humansequenz und konnten somit im weiteren nicht mehr komparativ untersucht werden.

	#	genomischer Bereich	Länge	GC-Gehalt	Wert	Anzahl der CpGs	Gen
Mensch	1	nt 18.812 – nt 19.021	210 bp	68,1%	133	19	?
	2	nt 36.433 – nt 36.660	228 bp	70,2%	95	17	?
	3	nt 38.500 – nt 38.642	143 bp	72,0%	64	11	?
	4	nt 45.998 – nt 46.097	100 bp	64,0%	63	9	?
	5	nt 92.113 – nt 92.472	360 bp	70,6%	109	24	?
	6	nt 94.981 – nt 95.082	102 bp	65,7%	61	9	?
	7	nt 107.196 – nt 108.073	878 bp	73,6%	1.227	113	*LMO1* Ex1a
	8	nt 112.521 – nt 113.505	985 bp	71,6%	1.159	117	*LMO1* Ex1b
	9	nt 174.946 – nt 175.085	140 bp	69,3%	113	14	
	10	nt 206.756 – nt 207.282	527 bp	66,0%	464	55	AL520122
	11	nt 225.810 – nt 226.044	235 bp	60,9%	144	21	?
	12		924 bp	74,8%	1.112	113	*TUB* Ex1c
	13	nt 296.981 – nt 297.145	164 bp	70,9%	114	15	*TUB* Ex1c
Maus	1	nt 76.162 – nt 76.278	117 bp	74,4%	105	12	?
	2	nt 142.238 – nt 142.962	725 bp	70,6%	823	82	*Lmo1* Ex1a
	3	nt 147.533 – nt 148.134	602 bp	70,6%	623	68	*Lmo1* Ex1b
	4	nt 226.403 – nt 226.766	364 bp	71,2%	465	46	AL520122 AA724522
	5	nt 296.554 – nt 297.191	638 bp	75,9%	772	78	*Tub* Ex1c
	6	nt 356.564 – nt 357.503	940 bp	67,8%	753	94	?
	7	nt 373.093 – nt 373.455	363 bp	73,8%	466	46	*Eif3* Ex1

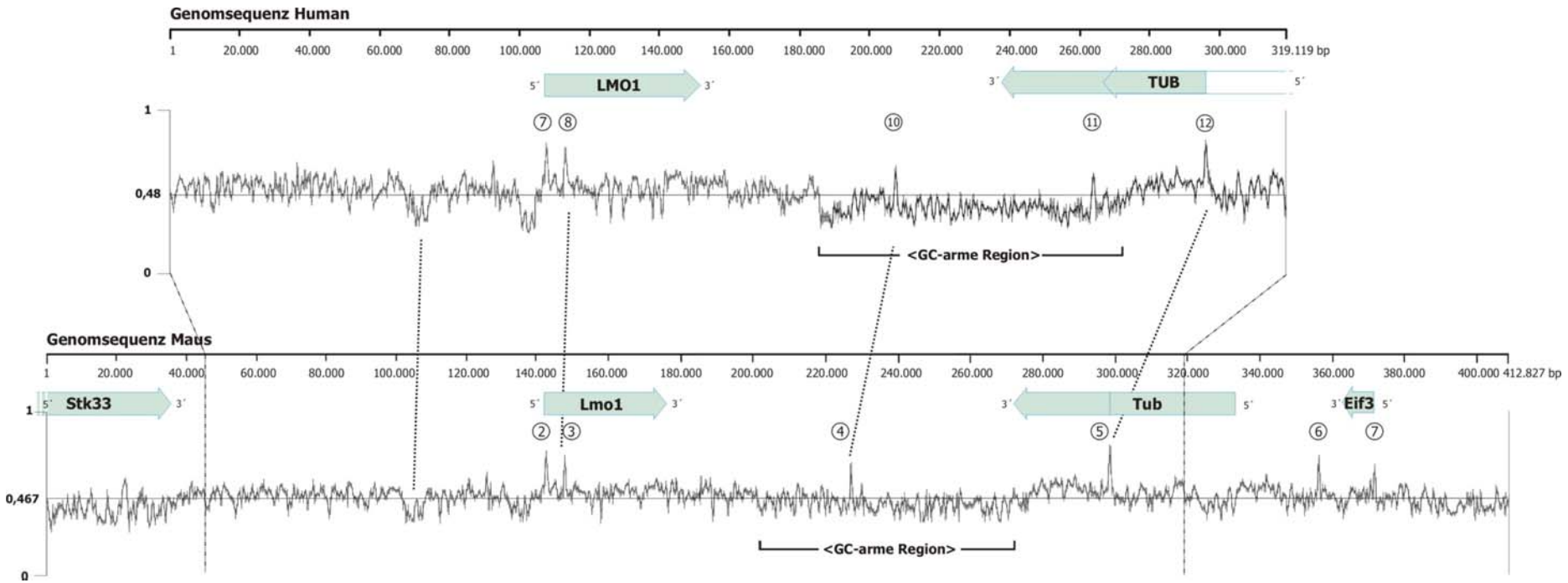


Abb. 29 **GC-Plot für die sequenzierten Bereich in beiden Spezies.** Der direkte Vergleich zeigte eine generelle Konservierung in der Verteilung der Basen Guanin und Cytosin zwischen menschlicher (oben) und Maus-genomischer Sequenz (unten). Es finden sich sowohl evolutiv erhaltene GC-arme Regionen wie gemeinsame Abschnitte mit sehr hohen GC-Gehalt (CpG-Inseln). Die nummerierten „Peaks" verweisen auf Tab. 22. Insbesondere die Promotorbereiche der Gene LMO1 und TUB weisen signifikante Spitzen im GC-Gehalt auf, die in beiden Spezies zu finden sind. So besitzt das LMO1-Gen für beide Anfangsexons im 5´Bereich eine GC-Spitze (Nummer 7/2 + 8/3). Das Gen TUB zeigt ebenfalls für das Exon 1c (Nummer 12/5) in beiden Genomen eine CpG-Insel. Die graphische Darstellung des GC-Gehaltes wurde mit dem Programm *GeneQuest* realisiert.

3.10.2 MAR-Analyse

Genomische DNA beinhaltet in ihrer Basenabfolge nicht nur funktionelle Bereiche mit proteinkodierender Information, sie besitzt auch Abschnitte, die für die Konformation und Struktur der DNA im Interphasekern wichtig sind. Bestimmte Chromatinbereiche zeichnen sich durch eine extreme Krümmung der DNA-Doppelhelix im Interphasekern aus (Vogelstein *et al.*, 1980) und bewirken eine Assoziation an die Proteine der Kernmembran. Solche als MARs („matrix attachment regions") (Cockerhill *et al.*, 1986) oder auch als SARs – für „scaffold attachment regions" – bezeichnete Regionen (Gasser *et al.*, 1986) haben einen überdurchschnittlich hohen AT-Gehalt und fallen durch bestimmte Sequenzmotive auf. Diese werden mit der transkriptionellen Regulation im Eukaryontengenom in Beziehung gebracht und spielen bei der DNA-Replikation eine besondere Rolle (McCready *et al.*, 1984). Ebenso können sie als Signalsequenz für Rekombinationsstellen im Genom, bzw. als Topoisomerase II-Erkennungsstellen angesehen werden (Cockerhill *et al.*, 1986), so dass topologische Veränderungen im DNA-Molekül in diesen Regionen eine funktionelle Bedeutung zukommt (Singh *et al.*, 1997).

In der über 300 kb langen humangenomischen Consensussequenz konnten zwei MAR-Regionen über der Signifikanzgrenze von 0,6 identifiziert werden. Der erste MAR-Bereich befand sich um die Base 70.500, etwa 36,5 kb proximal zum ersten Exon des *LMO1*-Gens in einem Cluster mit mehreren repetitiven Elementen, wie L1 und Alu. Die zweite Spitze mit einer noch deutlicheren Signifikanz von 0,91 umschrieb den Bereich zwischen den Basen Nukleotid 174.200 und 175.100 und befand sich ca. 22,4 kb distal zum letzten Exon des *LMO1*-Gens. Gefolgt wurde dieser Abschnitt von der „GC-armen Region", die sich bis in den 3´-Bereich des *TUB*-Gens erstreckte (siehe auch Abb. 30).

In der orthologen DNA-Sequenz der Maus konnten insgesamt 3 Bereiche (① nt. 104500 bis 104800: *0,66*; ② nt. 246300 bis 247200: *0,82*; ③ nt. 386200 bis 386400: *0,60*) mit einem signifikanten MAR-Potential > 0,6 charakterisiert werden. Vergleicht man die murinen MAR-Regionen mit denen des Menschen, so zeigte sich die Konservierung des ersten MARs vor dem Gen *LMO1*/*Lmo1*. Über einen Bereich von mehr als 500 bp betrug die Interspezieshomologie der beiden Genomsequenzen 60 bis 82%. Die zweite MAR-Region der Maus wies im Humangenom keinen korrespondierenden Bereich auf. Anders verhielt es sich mit der zweiten MAR-Region des Menschen. Die ersten 400 bp von nt. 174.207 bis 174.644 zeigten zur murinen Sequenz eine überdurchschnittliche Interspezieshomologie von 75%. Dieser Sequenzabschnitt fand sich aber in der MAR-Analyse der Maus, von nt. 199.293 bis 199.757 der murinen Consensussequenz nicht wieder. Die dritte MAR-Region der Maus befand sich in einem Abstand von 12,7 kb zum stromaufwärts gelegenen Bereich des *eIF3*-Gens und lag außerhalb des orthologen sequenzierten Bereichs des Menschen, so dass er nicht mit diesem verglichen werden konnte. Eine vierte weniger ausgeprägte MAR-Spitze unmittelbar vor dem 5´-Bereich des *LMO1*/*Lmo1*-Gens konnte ebenfalls in beiden Genomen identifiziert werden.

Der Vergleich der drei MAR-Positionen in der Maus zueinander zeigte einen ähnlichen Abstand von 142kb und 139 kb. In der humangenomischen Sequenz betrug die Distanz der beiden MAR-Positionen

den geringeren Abstand von 104 kb. Hervorzuheben ist, dass die identifizierten MAR-Regionen sich nicht mit den transkribierten Abschnitte der identifizierten Gene *Stk33*, *Lmo1*, *Tub* und *Eif3* überschneiden. Nicht ganz in dieses Schema passt allerdings der schwach ausgeprägte MAR im 5´-Bereich des *LMO1*/*Lmo1*-Gens, da er sich unmittelbar vor einem Promotorbereich befindet. Alle beschriebenen Ergebnisse sind in nachfolgender Abb. 30 grafisch zusammengestellt.

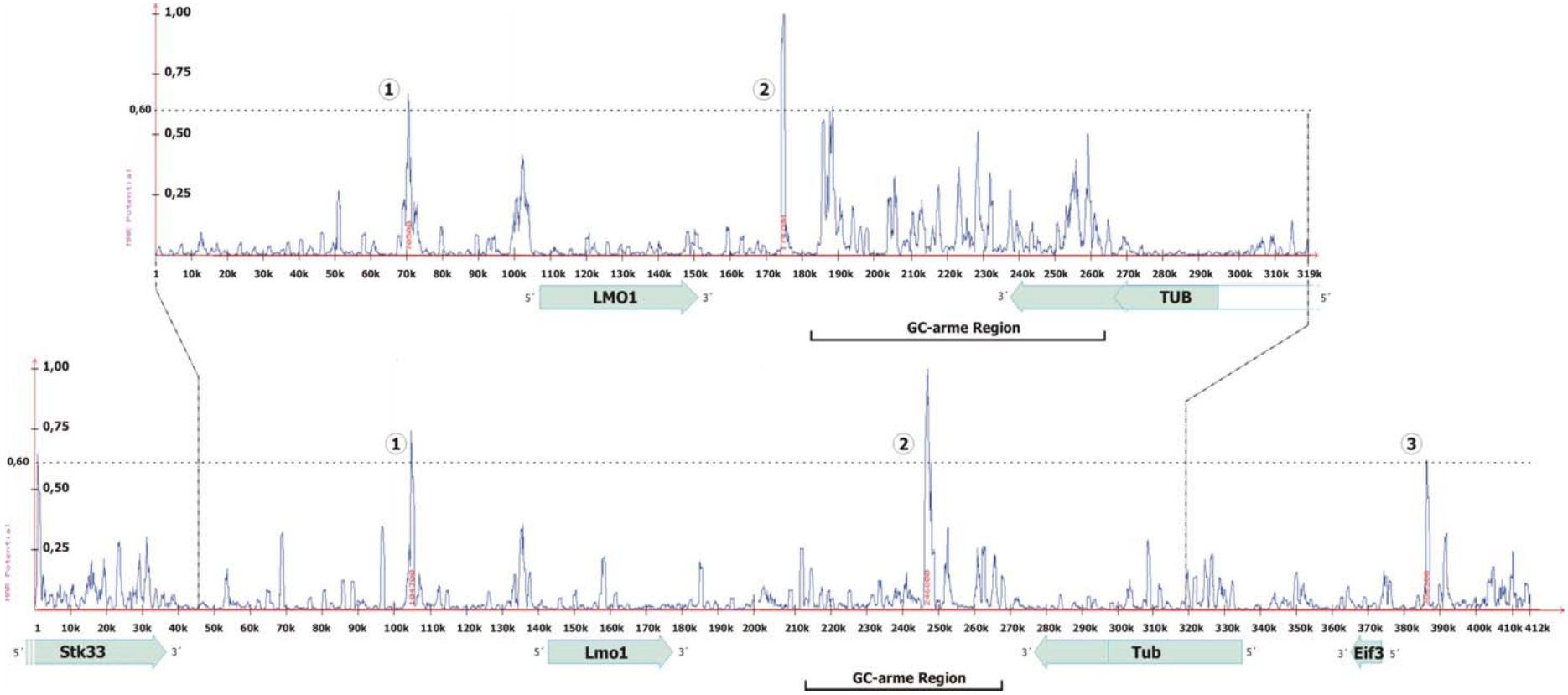


Abb. 30 **MAR-Plot der humanen und murinen DNA-Sequenz.** Über dem Signifikanzwert von 0,6 (entspricht einer Vorhersagewahrscheinlichkeit von 60%) konnten im Menschen zwei und im etwas größeren Bereich der Maus drei MAR-Regionen identifiziert werden. Während die beiden ersten MARs in beiden Genomen konserviert geblieben zu sein scheinen, konnte die beiden zweiten MARs nicht einander zugeordnet werden. Der dritte murine MAR liegt außerhalb des orthologen humanen Bereiches. Der GC-arme Region zwischen nt 185.00 und nt 263.00 im Menschen fällt deutlich als Region mit vielen „Peaks" zwischen den Signifikanzwerten 0,25 und 0,5 auf insbesondere im Humangenom auf.

3.10.3 Repetitive Elemente

Die Analyse der repetitiven Bereiche ergab, dass 36% der menschlichen DNA-Sequenz und 28% der Maus-Sequenz von repetitiven Sequenzanteilen repräsentiert wird. Im direkten Vergleich der beiden Spezies ist somit der relative repetitive Anteil im Menschen um 7,6% gegenüber der Maus angehoben. Insbesondere die kurzen (SINEs = +2,5%) und die langen (LINEs = +6,7%) interspergierten Sequenzwiederholungen sind beim Menschen signifikant erhöht. Werden alle repetitiven Sequenzen addiert, beträgt die Differenz zwischen diesen Genomregion in Mensch und Maus sogar 9,3%. Dies dürfte eine Erklärung für die Beobachtung aus der Dotplot-Analyse sein, dass der untersuchte genomische Bereich im Menschen mit 319 Kilobasen nur von 274 Kilobasen in der Maus repräsentiert wird. Eine Ausnahme von dieser Tendenz bilden nur die LTR-Elemente, die in der untersuchten Maus-genomischen Consensussequenz im Vergleich zum Humangenom um 0,5% leicht erhöht sind. Auch die einfachen Sequenzwiederholungen („Single repeats") zeigen sich im Mausgenom um 1,7% zahlreicher als im Humangenom.

Eine Zusammenfassung aller Ergebnisse ist in nachfolgender Tabelle 23 dargestellt. Die detaillierte Übersicht über Lokalisierung und Verteilung der identifizierten Elemente ist der graphischen Darstellung der PIP-Analyse (Abb. 25a/b; Kap. 3.8.1.2) zu entnehmen.

Die lokale Verteilung der repetitiven Elemente zeigte im Interspeziesvergleich, dass fast alle größeren Sequenzabschnitte ohne signifikante Konservierung, von repetitiven Sequenzen flankiert und besetzt sind. Insbesondere die humane Sequenz weist Bereiche auf, die im Vergleich zur Maus-genomischen Sequenz durch eine starke Anhäufung von „Repeat"-Elementen bis über mehrere Kilobasen an DNA geprägt sind. So zeichnet sich z.B. der humane Genomabschnitt von Nukleotid 185.445 bis Nukleotid 194.086 gegenüber der Maus-genomischen Sequenz von Nukleotid 213.662 bis 214.401 durch einen DNA-Zuwachs von nahezu acht Kilobasen aus, der im Mausgenom nicht zu finden ist. Der repetitive Anteil in diesem Abschnitt stieg dabei auf über 62%. Weitere Beispiele sind die humanen Bereiche von Nukleotid 224.914 bis 236.541 vs. Maus Nukleotid 244.207 bis 249.263 oder von Nukleotid 251.305 bis 263.378 vs. Maus 260.720 bp bis 266.198. Auch hier erhöhte sich der repetitive Anteil des 11.848 bp großen Abschnitt auf 74,9%, bzw. auf über 90% in einem Abschnitt von 12.074 bp Länge (siehe auch Abb. 31 – A bis C).

Große Cluster repetitiver Elemente, wie die drei aufgeführten Beispiele, existieren in der analysierten Sequenz nur außerhalb der transkribierten Bereiche in den Intergen-Regionen. Die Intronbereiche sind im Humangenom zwar im Durchschnitt länger, doch ist hier die Zunahme an neuen repetitiven Elemente nur vereinzelt gegeben und erstreckt sich selten über mehrere Hundert Basenbasen.

Tab. 26 **Zusammenstellung aller repetitiven Sequenzanteile** der Mensch- und Maus-genomischen Consensussequenz. Das humane Genom weist insgesamt mit über 7,6% mehr repetitive Bereiche auf als die orthologe murine Genomsequenz. Die repetitiven Elemente wurde in vier Kategorien unterteilt: SINEs = „short interspersed nuclear elements"; LINEs = „long interspersed nuclear elements"; LTRs = „long terminal repeats" und MERs = „medium reiteration frequency interspersed repeats". Weitere Subklassifizierungen wurden wie folgt abgekürzt: MIR = „Mammalian-wide interspersed repeat"; MaLR = Mammalian LTR-retrotransposon; ERV = endogenous retrovirus.

Humangenomische Sequenz: 319.119 bp
maskierte Bereiche insgesamt 115.004 bp (= 36,04%)

repetitives Element		Anzahl der Elemente	Länge [bp]	% der Gesamt-sequenz
SINEs		165	32.901	10,31
	ALUs	68	19.083	5,98
	MIRs	97	13.818	4,33
LINEs		115	56.548	17,72
	LINE1	48	36.228	11,35
	LINE2	55	17.432	5,46
	L3/CR1	12	2.888	0,90
LTR-Elemente		38	14.582	4,57
	MaLRs	24	7.989	2,50
	ERVL	6	1.749	0,55
	ERV Kl. 1	8	4.844	1,52
	ERV Kl. 2	0	0	0,00
DNA Elemente		25	6.144	1,93
	MER T1	16	3.364	1,05
	MER T2	2	1.676	0,53
Summe aller interspergierten Wiederholungen			110.175	34,52
„small RNAs"		0	0	0,00
Einfache Wiederholungen („simple repeats")		49	2.971	0,93
Bereiche niedriger Komplexität („low complexity regions")		33	1.880	0,59

Mausgenomische Sequenz: 412.827 bp
maskierte Bereiche insgesamt 117.030 bp(= 28,35%)

repetitives Element		Anzahl der Elemente	Länge [bp]	% der Gesamt-sequenz
SINEs		237	32.060	7,77
	B1s	81	9.426	2,28
	B2-B4	107	17.674	4,28
	IDs	15	957	0,23
	MIRs	34	4.003	0,97
LINEs		80	45.489	11,02
	LINE1	61	41.689	10,10
	LINE2	17	3.426	0,83
	L3/CR1	2	374	0,09
LTR-Elemente		74	20.667	5,01
	MaLRs	60	16.729	4,05
	ERVL	2	189	0,05
	ERV Kl. 1	5	959	0,23
	ERV Kl. 2	3	1.630	0,39
DNA Elemente		20	3.823	0,93
	MER T1	12	1.735	0,42
	MER T2	6	1.868	0,45
Summe aller interspergierten Wiederholungen			103.933	25,18
„small RNAs"		2	181	0,04
einfache Wiederholungen („simple repeats")		167	10.809	2,62
Bereiche niedriger Komplexität („low complexity regions")		40	2.274	0,55

Die gesonderte *REPEATMASKER*-Analyse des genomischen Abschnitts der „GC-armen Region" von Nukleotid 185.000 bis 263.000 im Menschen und von Nukleotid 213.00 bis 266.000 der Maus (Abb. 29) zeigte, dass dieser Bereich durch einen erhöhten repetitiven Anteil charakterisiert ist. Mit 30,59% repetitiver Anteil war dieser Abschnitt in der Maus um 2,24% nur leicht erhöht. Der orthologe Bereich im Humangenom hingegen wies mit 56,53% repetitiven Anteil eine Erhöhung um insgesamt 20,49% zum Consensussequenz-Durchschnitt auf. Dabei stieg vor allen die Zahl der LINE-Elemente in der humanen DNA und verdreifachte sich im Vergleich zum murinen Abschnitt (human: 34 vs. murin: 11). Die Zahl der SINE-Elemente blieb dagegen mit 39 zu 38 nahezu konstant. Auch die Anzahl der LTR- und MER-Elemente veränderte sich für diesen Abschnitt in beiden Genomen mit 13 vs. 10, bzw. 3 vs. 3 nur unwesentlich.

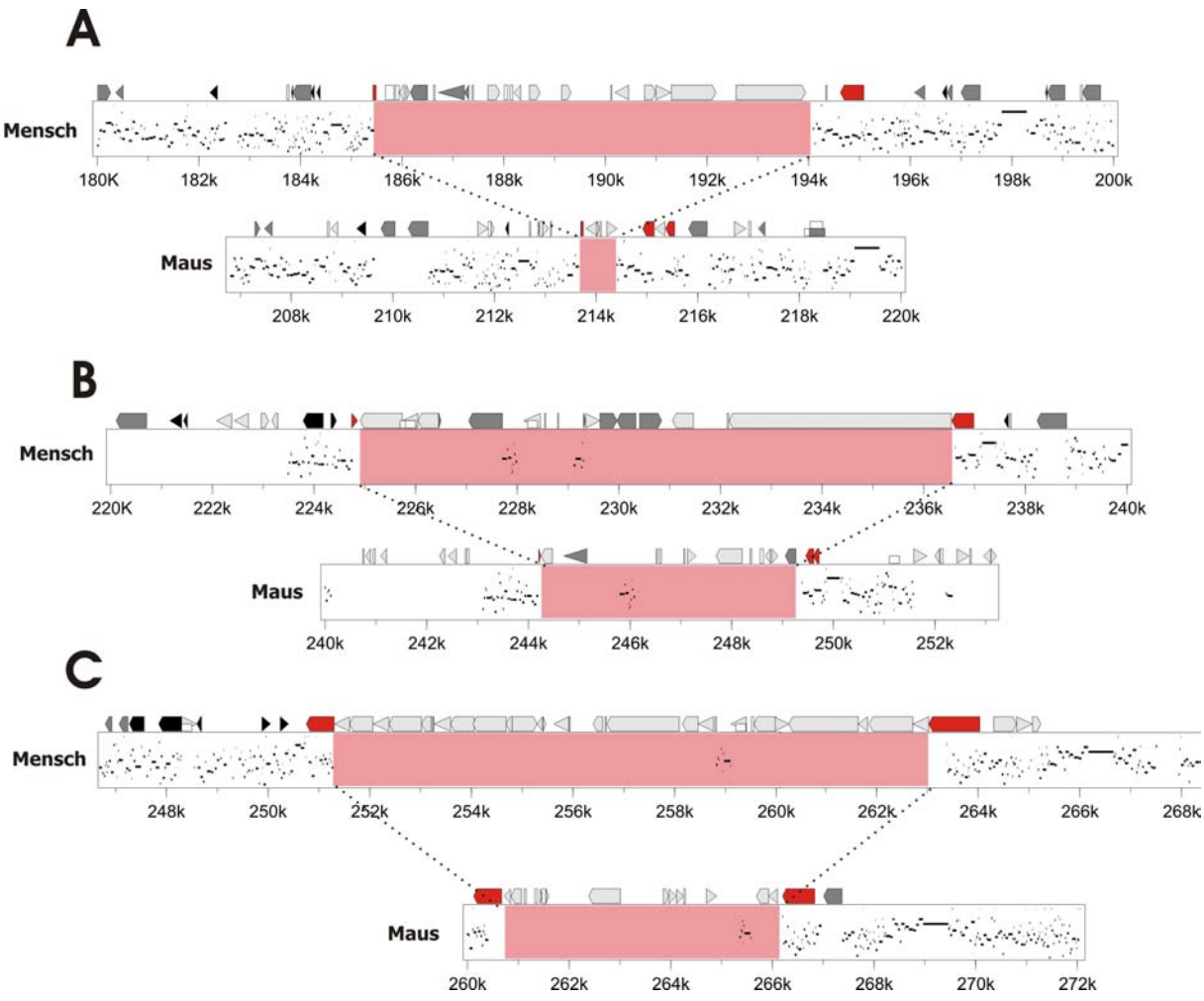


Abb. 31 **Interspeziesvergleich repetitiver Sequenzabschnitte.** Mithilfe der PIP-Analyse wurden die drei größten konservierten „Repeat"-Cluster zwischen Mensch und Maus gegenübergestellt. **A)** Vergleich eines repetitiven genomischen Bereiches etwa 33 kb distal zum LMO1-Gen, der im Humanen ca. 8,6 kb umfasst und durch 21 repetitive Elemente (8 LINEs, 3 SINEs, 1 LTR, 2 MER und 7 „simple repeats", bzw. Abschnitte geringer Komplexität) charakterisiert ist. Der orthologe Bereich der Maus weist dagegen nur 5 repetitive Elemente auf (3 SINEs und 2 „simple repeats") und ist lediglich 739 bp groß. **B)** Ein weiteres konserviertes „Repeat"-Cluster befindet sich ca. 38 kb proximal zum TUB-Gen und setzt sich im Humangenom aus 6 LINEs, 3 SINEs, 5 LTRs, 2

einfachen Sequenzwiederholungen und einer niedrig-komplexen Region zusammen. Der orthologe murine Genombereich besitzt dagegen nur 2 LINEs, 4 SINEs, ein LTR, ein MER-DNA-Element und zwei Bereiche niedriger Komplexität. **C)** Ein dritter Bereich, ca. 11 kb proximal zum TUB-Gen, scheint im Humangenom seine Sequenzlänge durch eine Vermehrung der LINE-Elementen vergrößert zu haben. So erhöhte sich in diesem Abschnitt ihre Zahl von 7 in der Maus auf insgesamt 15 im Menschen. Auch die Zahl der SINEs stieg von 5 auf 10 an. Die flankierenden SINE Elemente (Mensch: AluSx—AluSq vs. Maus: B1F—B4) mit den sich in beiden Genomen anschließenden LINE-Elementen können als konservierte Randbereiche, mögliche Reintegrationsstellen des Transpositionsereignisses darstellen und sind wie in A) und B) rot hervorgehoben. Rosa wurde der Gesamtbereich des betrachteten „Repeat"-Clusters markiert.

4 DISKUSSION

4.1 Sequenzierungsstrategien großer genomischer Bereiche und ihre Analyse

4.1.1 Die Sequenzierungsstrategie

Zu Beginn der Sequenzierungsarbeiten galt es eine effiziente Sequenzierungsstrategie zu wählen, um die zu bestimmenden chromosomalen Bereiche von ca. 250 bis 300 kb in Mensch und Maus mit den zur Verfügung stehenden laboreigenen Ressourcen und in der veranschlagten Zeit in ihrer Nukleotidbasenabfolge zu entschlüsseln. Es wurde sich für die **hierarchische „Shotgun"-Sequenzierung** entschieden, da sie in Hinblick auf den Grad der möglichen Automatisierung und den dabei generierten Datenmengen am geeignetesten erschien. Kombiniert wurde diese Strategie mit der „**Primerwalking"-Sequenzierung**, die am Ende der „Shotgun"-Sequenzierung-Phase beim Sequenz-„Finishing" zum Schließen von verbliebenen Sequenzlücken, bzw. zur Verifizierung von unsicheren Sequenzbereichen eingesetzt wurde (siehe auch Abb. 4). Die alternative Methode der **„Whole-genome-shotgun"-Sequenzierung**, die bei der Entschlüsselung kleinerer Eukaryonten-genome wie *Saccharomyces cerevisiae, Caenorhabditis elegans* und *Escherichia coli* zum Einsatz kam (Weber & Myers, 1997), bzw. bei der Sequenzierung des humanen Genoms durch Venter & Mitarbeiter (2001) angewandt wurde, schied aus, da die Aufgabenstellung *nur* die Sequenzierung eines distinkten chromosomalen Bereichs in Mensch und Maus umfasste und nicht die Entzifferung eines gesamten Genoms.

Der entscheidende Vorteil der hierarchischen „Shotgun"-Sequenzierung basierte auf der Sequenzierung von zuvor genau lokalisierten einzelnen genomischen Fragmenten. Je kleiner diese genomischen Abschnitte umschrieben sind, desto genauer kann eine exakte Zuordnung der Einzelsequenzen aus den „Shotgun"-Subklonen vorgenommen werden (Deininger, 1983). Insbesondere in repetitiven Bereichen ist der umgrenzte DNA-Abschnitt beim Assemblieren der Einzelsequenzen sehr hilfreich. Die Größe der in einem Projekt zu sequenzierenden genomischen DNA kann durch die Auswahl des Vektorsystems bestimmt werden. In dieser Arbeit wurden hauptsächlich PAC- und BAC-Klone mit 130 bis 190 kb großen genomischen DNA-Integraten subkloniert und sequenziert, da sich ihr Kartierungs- und Subklonierungsaufwand auf der einen Seite und der Assemblierungs- und „Finishing"-Aufwand bei der Ermittlung der Consensussequenz auf der anderen Seite in einem Verhältnis bewegte, bei der sowohl die vorhandene Sequenzierungskapazität wie auch die zur Verfügung stehenden Rechnerleistungen effizient genutzt werden konnte. Außerdem standen für die Suche nach Klonen aus dem zu charakterisierenden Genombereichen entsprechende Klonbibliotheken mit ausreichender Klon-Redundanz (siehe Kap. 2.19.4) zur Verfügung. Auf die Sequenzierung von Cosmid-Klonen wurde bis auf die Ausnahme im distalen Bereichen des humanen

Genomabschnitts um das Gen *LMO1* verzichtet, da die Suche und Kartierung nach geeigneten Klone über den gesamten zu charakterisierenden Bereich zu arbeitsintensiv gewesen wäre. Die Sequenzierung von YAC-Klonen wurde ebenso nicht in Betracht gezogen, da die Gefahr für chimäre Klone bestehend aus unterschiedlichen Genomabschnitten für zu hoch eingeschätzt wurde. Gute Erfahrungen und Ergebnisse, die bereits im Rahmen des Humangenomprojektes mit diesem strategischen Vorgehen erzielt worden waren (Green, 1997), bestätigten die hier gewählte Vorgehensweise.

Am Ende einer jeden BAC-, bzw. PAC-Sequenzierung fand zum Schließen der übrig geblieben Sequenzlücken die „Primerwalking"-Sequenzierung Anwendung. Aufgrund der sehr kostenintensiven Primer-Synthese wurde diese aber erst zu einem Zeitpunkt realisiert, wenn aus den angeordneten Subklonen keine auswertbaren Sequenzinformationen mehr zu ermitteln waren. Ein Kriterium für den Umfang dieser Art der Sequenzierung während des Sequenz-„Finishing" war die Qualität der Subklonierung der nebulisierten DNA-Fragmente in den Sequenzierungsvektor pUC18 und die Zahl, d.h. die Redundanz der generierten Klonsequenzen. Je höher die Redundanz der „shotgun"-sequenzierten Klone, desto weniger Primer mussten für die Verifizierung und Schließung von Sequenzlücken eingesetzt werden. Diese Verhältnismäßigkeit wurde lediglich durch einen hohen Gehalt an repetitiven Elementen, wie beispielsweise für BAC 287P4 der Fall, verschoben.

Die anfängliche Kartierung der zu untersuchenden Region und die Anordnung der lokalisierten Klone zueinander in einem Contig, ermöglichte zum einen die gezielte Auswahl von bestimmten Klonen, die nur zu einem geringen Teil über ihre flankierenden Sequenzanteile mit ihrem Nachbar überlappen, so dass das doppelte Sequenzieren der gemeinsamen Abschnitte auf ein Mindestmaß reduziert werden konnte. Zum anderen boten diese gemeinsamen Sequenzbereichsflanken die Möglichkeit, beide aufeinander folgenden Bereiche nach ihrer Sequenzierung eindeutig zueinander für die genomische Consensussequenz zu verankern. Da das Finishing der sequenzierten BAC-, bzw. PAC-Klone zeitlich versetzt hintereinander gestaffelt geschah, konnte gezielt auf die zweifache Verifizierung dieser überlappenden Randbereiche verzichtet werden und wurde auf das Finishing einer der beiden Klone beschränkt. Somit führte die standardisierte Sequenzierung vieler randomisierter Subklone und das hierarchisch abgestufte, sukzessive Zusammenführen immer größer werdender, zusammenhängender Bereiche zur Ermittlung einer durchgehenden genomischen Consensussequenz.

4.1.2 Die „High-throughput"-Sequenzierung

Um die Zahl von insgesamt 10.812 informativen Sequenzierungen aus dieser Arbeit zu generieren, d.h. all jene sequenzierten Basenabfolgen, die für die Assemblierung in der vorgegebenen Redundanz von drei Sequenzierungen bis zur verifizierten Consensussequenz nötig waren, musste insbesondere während der „Shotgun"-Phase eine gewisser Grad an Automatisierung entwickelt werden. Der etablierte Hochdurchsatz-Maßstab („high-throughput") wurde im Kern durch die Umstellung der verschiedenen Arbeitsabläufe auf das 96-Loch Mikrotiterplatten-Format erreicht. Sämtliche Schritte beginnend von der Anordnung der rekombinanten Subklone über die DNA-Präparation (Kap. 2.2), Sequenzprobenvorbereitung mit „Cycle-Sequencing" und anschließender Aufreinigung (Kap. 2.14), bis hin zum Probenauftrag auf das Polyacrylamid-Sequenziergel wurden in ein und demselben Format durchgeführt. Dies hatte zur Folge, dass bei einigen im Arbeitsablauf eingebundenen Instrumente eine Anpassung auf dieses Format vorgenommen werden musste. Dies bezog sich sowohl auf Instrumente für das „Handling", z.B. Achtkanal-Metallkapillar-Pipette (KLOEHN) zum effizienten Beladen des Sequenziergels, wie auch auf verschiedene Geräte, z.B. PCR-Geräten mit speziellem Heizblock für 96er-Mikrotiter-Platten und ein 96-Proben detektierender DNA-Sequenzierautomat (ABI PRISM 377-96) mit entsprechend optimierter Software. Ebenso mussten auftretende Schwierigkeiten bezüglich der nicht vorhandenen Uniformität in der Qualität der Probenaufbereitung gelöst werden. Insbesondere galt es Randeffekte des rechteckigen Plattenformates z.B. während des „Cycle-Sequencings" zu minimieren. Da zu Beginn der verliegenden Arbeit nur spärlich auf allgemeine Erfahrungswerte innerhalb und außerhalb des Labors zurückgegriffen werden konnte, mussten viele Optimierungsschritte selbst etabliert werden. Beispiele hierfür waren die Wahl des richtigen DNA-Aufarbeitungskits unter verschiedenen Angeboten z.B. des Herstellers QIAGEN oder die Aufreinigung der „Cycle-Sequencing"-Proben mit Hilfe des MILLIPORE Multiscreen-Systems. Die Optimierung der Arbeitsabläufe zeigte sich schließlich im Laufe der Arbeit durch das immer schnellere Abschließen der einzelnen Sequenzierprojekte von 1½ Jahren (PAC 12G13) auf fünf Monate (BAC 282L1). Einen entscheidenden Anteil an dieser zeitlichen Verkürzung hatte auch die Optimierung der im nachfolgenden Kapitel beschriebenen computergestützten Sequenzassemblierung.

Eine noch größere Steigerung der parallelen Probenbearbeitung auf das nächst höhere 384-Loch Mikrotiterplatten-Format wurde nicht in Erwägung gezogen, da hier die Miniaturisierung derart fortgeschritten ist, dass eine sichere Probenbearbeitung nur noch durch ein vollautomatisches „Handling" wie etwa mit einem Pipetierroboter effizient gewährleistet werden kann. Eine manuelle Bearbeitung von 384 Proben hätte den vordergründigen zeitlichen Gewinn durch die parallele Probenbearbeitung durch die verstärkt auftretende Fehleranfälligkeit u.a. bedingt durch die wesentlich kleineren Volumina wieder zunichte gemacht.

4.1.3 Die Sequenzassemblierung

Ein weiterer Bereich für die Automation und Optimierung war die Verwaltung und Bearbeitung der generierten Sequenzinformationen. Insgesamt wurden für die Sequenzassemblierung vier verschiedene Programm-Pakete (*SEQUENCHER*, *PHREDPHRAP*, *STADEN*, *LASERGENE*) getestet, die sich in ihrem Funktionsumfang mit zwei weiteren Programmen (*SEQUENCING-ANALYSIS-SOFTWARE*, *CHROMAS*) für das Baseneditieren und Visualisieren der Rohdaten teilweise überschnitten (vgl. Abb. 5).

Anfangs wurde ausschließlich mit den Programmen *SEQUENCING ANALYSIS- SOFTWARE* (PE BIOSYSTEMS) für das Editieren und dem *SEQUENCHER*-Programm (GENCODES) für die Contigerstellung („Assembling") gearbeitet. Das Editieren umfasste zum einen das notwendige visuelle Gegenlesen der ermittelten Nukleotidsequenz anhand des vom Sequenziergerät detektierten Elektropherogramms mit dem Korrigieren einzelner fehlinterpretierter Basen und zum anderen das Entfernen der nichtinformativen Sequenzränder am 5´- und am 3´-Ende. Die Charakterisierung der bearbeiteten Sequenz wurde über eine BlastN-Analyse vorgenommen und mit dem Entfernen etwaiger vektorieller Überhänge beendet. Da die *SEQUENCING-ANALYSIS-SOFTWARE* keine Automatisierungsmöglichkeit und das *SEQUENCHER*-Programm nur ansatzweise für eine Automatisierungsroutine konfigurierbar war, wurde auf das Software-Paket *PHREDPHRAP* (Ewing & Green, 1998) gewechselt, da dieses die obigen sehr zeitintensiven manuellen Eingriffe automatisierte. Diese Umstellung vollzog sich auch Plattform-übergreifend, so dass nicht mehr auf Macintosh-Rechnern, sondern auf eine leistungsstarke Sun-Workstation mit einem auf UNIX-basierenden Betriebssystem gearbeitet wurde. Mit diesem Schritt konnte eine Software etabliert werden, die zu den im Humangenomprojekt standardmäßig eingesetzten Programmen der Sequenzassemblierung zählte. Als Alternative wurde anfangs auch mit dem *STADEN*-Paket (Staden, 1996) gearbeitet, welches sich bezüglich seiner Programmarchitektur ähnlich dem *PHREDPHRAP* aus mehreren eigenständigen Programm-Modulen (*pregap4*, *gap4*, *vector_clip*, *trev*) zusammensetzt. Aufgrund seiner komplexeren Programmstruktur und der gewöhnungsbedürftigen Bedienungsführung fand dieses Programm allerdings keine weiterführende Anwendung.

Das Programm-Paket *PHREDPHRAP* setzte sich in der verwendeten Version aus insgesamt vier verschiedenen Modulen, den Unterprogrammen *PHRED*, *CROSS_MATCH*, *PHRAP* und *CONSED* zusammen, die den gesamten Arbeitsablauf vom Import neuer Rohsequenzdaten bis zur fertig assemblierten Consensussequenz abdeckten. Durch die direkte Kopplung dieser vier Programme untereinander konnte die gesamte Einheit nach entsprechender Konfiguration mit projektrelevanten Zusatzinformationen (v.a. Angabe des Klonierungs- und Subklonierungsvektors) selbstständig arbeiten. Manuell wurde auf die Sequenzinformationen nur noch am Ende des Protokolls mit Hilfe des Programms *CONSED* zugegriffen und gegebenenfalls im Contig editiert. Im Einzelnen durchliefen die importierten Sequenzinformationen eine mehrstufige Computeranalyse, die außer der Automatisierung eine Reihe von weiteren Vorteilen mit sich brachte. Bevor neue Sequenz-Daten einem *PHREDPHRAP*-Projekt zugeführt wurden, mussten die Dateien gemäß der Nomenklaturkonvention mit einem

entsprechenden Suffix versehen werden, welches dem Programm Informationen über die verwendete Chemie (Dye-Primer, bzw. Dye-Terminatoren) und den eingesetzten Primer (Subklonierungsprimer „forward" oder „revers", bzw. „Walking-Primer") der generierten Sequenz mitteilt. Dieser Zusatz hatte, wie sich bei den Arbeiten herausstellte, einen großen Einfluss auf die korrekte Assemblierung der Sequenzdaten und verhinderte „Alignment"-Fehler. Nach Einlesen der importierten Sequenzen wurden sie im Programm *PHRED* (Ewing *et al.*, 1998) einem nochmaligen „Base-Calling"-Algorithmus unterzogen. In diesem Schritt wurden die Primärinformationen, d.h. die detektierten Fluoreszenzsignale des markierten DNA-Moleküls im Chromatogramm, neu ausgelesen und in ihre Nukleotidbasenabfolge übersetzt. Statistische Untersuchungen belegten, dass die Fehlerrate des *PHRED*-Base-Callings um 40% bis 50% niedriger liegt als bei dem zu Beginn eingesetzten *DATA-COLLECTING-SOFTWARE* der ABI-Software (Ewing *et al.*, 1998). Gleichzeitig wurde jede Base mit einem Qualitätswert („quality score") von 4 bis 60 versehen, der eine Aussage darüber gibt, mit welcher Sicherheit - entsprechend der Qualität des detektierten Fluoreszenzsignals - diese Nukleotid-bestimmung vorgenommen werden konnte. Dieses „Quality-Valuation" oder „-Scoring" stellte einen entscheidenden Vorteil zum *SEQUENCHER*-Programm dar, da beim späteren Assemblieren diese Qualitätswerte zum Ermitteln der einzelnen Basen für die Consensussequenz berücksichtigt wurden. Das Ergebnis von *PHRED*, bestehend aus der Basenabfolge mit den entsprechenden Qualitätswerten, wurde am Ende einer jeder Sequenzanalyse in einer gesonderten PHD-Datei zusammengefasst. Diese PHD-Textdatei wurde im nachfolgenden mit Hilfe des Programms *CROSS_MATCH* auf Homologien zu vorgegebenen Referenzsequenzen hin untersucht. Das waren Sequenzen des Klonierungsvektors (PAC-, BAC- oder Cosmid-Vektor) und des Sequenziervektors (pUC18) und die Sequenz des *E. coli*-Genoms, mit der vereinzelt klonierte bakteriengenomische Kontaminationen identifiziert wurden. Das Programm lieferte als „Output" eine um die Vektorsequenz maskierte Version der gelesenen Basensequenz. Dieser Punkt ist außer der Automatisierungsfunktion als Vorteil gegenüber der irreversiblen Option des Vektor-„Clippings" des *SEQUENCHER*-Programms zu werten, da hier vorhandenen Informationen nicht weggeschnitten und dadurch gelöscht, sondern nur ausgeblendet und jederzeit einsehbar waren. Die durch *CROSS_MATCH* bis auf die informative genomische Sequenz maskierte DNA wurden danach mit Hilfe des Programms *PHRAP* (= „phragment assembly program" oder „phil's revised assembly program") mit allen anderen Sequenzen des Projektes verglichen und zu Contigs anhand ihrer überlappenden Sequenzbereiche zusammengesetzt. Dieser Assemblierungs-schritt ist die Kernfunktion des gesamten Programmpakets, da hier in einem fortgeschrittene Projekt mehrere Tausend Sequenzen (z.B. BAC 287P4 = 2.955 informative Einzelsequenzen) miteinander verglichen werden. Konnten alle vorbereitenden Sequenzbearbeitungen bis zu diesem Schritt noch manuell vorgenommen werden, so war dieser „Alignment"-Schritt ohne die computergestützte Hilfe nicht mehr zu realisieren. Mit dem *SEQUENCHER*-Programm durchgeführt, war dieser Schritt der rechenzeitintensivste Teil, der im fortgeschrittenen Projekt mehrer Stunden (bis zu 8h) dauern konnte. Auch in diesem Gesichtspunkt erwies sich das *PHRAP*-Programm als wesentliche Verbesserung, da selbst Projekte mit über 2.000 Einzelsequenzen in weniger als 30 Minuten komplett neu assembliert

werden konnten. Der extrem leistungsfähige Programm-Algorithmus erlaubte es, nach dem Import neuer Sequenzen das gesamte Projekt jedes Mal neu zu berechnen und alle Einträge unabhängig von zuvor gebildeten Contigs miteinander zu vergleichen. Dies hatte den entscheidenden Vorteil, dass etwaige Fehl-„Alignments" aufgrund schlechter Sequenzinformation nach Hinzufügen von neuen informativeren Sequenzen automatisch aufgelöst wurden. Das anfängliche Problem falsch zusammengesetzter Contigs in Projekten des *SEQUENCHER*-Programm in hoch repetitiven Genomabschnitten, konnte so vermieden werden. Im Unterschied zu *PHRAP* wird die gebildete Consensussequenz im *SEQUENCHER*-Programm über alle Projektaktualisierungen hinweg beibehalten, sofern die gebildeten Contigs nicht absichtlich aufgelöst und neu berechnet werden. Ein weiterer entscheidender Vorteil, ist die Ermittlung der Consensussequenz mit Hilfe der Qualitätswerte der einzelnen Basen. Verfährt das *SEQUENCHER*-Programm in der verwendeten Version 3.1 nach dem statistischen Prinzip der Mehrheit, so entscheidet *PHRAP* an jeder Sequenzposition nach der Base mit dem höchsten „quality score". Das Ergebnis war auch die korrekte Abbildung einer genomischen Sequenz in Abschnitten, die durch mehrere qualitativ schlechte und nur einer guten Sequenz gebildet wurden. Dieser Qualitätswert wurde als Referenz für die Base in der Consensussequenz übernommen und konnte für ihre graphische Darstellung mit dem Programm *CONSED* verwendet werden. Liefen diese Bearbeitungsschritte nach Start des *PHREDPHRAP*-Programms autonom ohne Eingriff des Anwenders und ohne Zwischendarstellung ab, so erfolgte die Visualisierung und Korrekturmöglichkeit der berechneten Consensussequenz im letzten Programmmodul *CONSED*. *CONSED* stellte für den Editor eine graphische Oberfläche dar, die mit einer Palette von verschiedenen Programm-Optionen verknüpft war. Sie erlaubte es sowohl Zugriff auf die Primärdaten zu nehmen, wie sie vom Sequenziergerät generiert worden waren, wie auch darüber hinaus die qualitative Bewertung der Basen zu verändern. Für diesen abschließenden Schritt des Sequenz-„Finishing" standen sowohl mehrere Möglichkeiten der graphisch-farblichen Darstellung wie auch die Möglichkeit Sequenzbereiche direkt auf ihre Homologie hin zu vergleichen. So konnte eine schnelle visuelle Einschätzung anhand der Farbkodierung (Basenqualität, Fehlpaarung, editierte oder maskierte Base, etc.), getroffen und gegebenenfalls manuell korrigiert werden. Als besonders hilfreich erwies sich die Möglichkeit, direkt aus bestimmten Bereichen der Consensussequenz heraus, Primersequenzen generieren zu lassen. Musste dieser Schritt des „Primerwalking" am Ende eines jeden Sequenzierprojekts, um verbliebene Contiglücken zu schließen oder Bereiche mit schlechter Qualität zu verifizieren, mit dem *SEQUENCHER*-Programm sehr arbeitsintensiv über das Herauskopieren des Primer-kodierenden Sequenzabschnitts in ein Internet-basiertes Primer-Design-Programm realisiert werden, konnte dieser Schritt komfortabel direkt im *CONSED*-Programm ausgeführt werden. Ein zusätzlicher Vorteil bestand darüber hinaus in dem Homologieabgleich der berechneten Primersequenz zur Consensussequenz des Gesamtprojektes, um so die Möglichkeit des Fehl-„primings" zu minimieren. Nach Auswahl einer geeigneten Primersequenz, die nach vorheriger Eingabe verschiedener Parameter (z.B. Primerlänge, Schmelztemperatur) berechnet wurde, erfolgte eine automatische Bezeichnung des Primers und eine Markierung an der entsprechenden Stelle in der Consensussequenz. Für den Aspekt der Daten-

sicherung wurden alle manuellen Veränderungen innerhalb des Projektes in einem „Log-File" mitprotokolliert, so dass nach einem etwaigen Systemabsturz, das Projekt exakt bis zu der Stelle rekonstruiert werden konnte. Diese Funktion konnte bei Sequencher nur in Form eines in gewissen Zeitabständen automatisch erfolgendes Abspeichern des gesamten Projektes erzielt werden, wodurch es aber bei größeren Projekten zu erheblichen zeitlichen Verzögerungen von mehreren Minuten kam.
Eine andere Möglichkeit der Sequenzassemblierung hätte die Verwendung des ebenfalls modular aufgebauten *LASERGENE*-Paketes (DNASTAR, USA) bedeutet, das mit den Programm-Modulen *EDITSEQ* und *SEQMANII* einen ähnlichen Funktionsumfang abdeckte wie *PHREDPHRAP* oder *STADEN*. Da diese Programm-Module aber ebenfalls einer größeren Einflussnahme durch den Editor bedurften und die Software nicht für eine leistungsstarke Workstation zur Verfügung stand, wurden auch dieses Programm-Paket nicht über eine kurze Einarbeitung hinaus für die Sequenz-Assemblierung angewendet.

4.1.4 Die Sequenzanalyse

Nach Verifizierung der generierten Genomsequenzen der sequenzierten Klone wurde eine umfassende Analyse der DNA-Sequenzinformation durchgeführt. Diese Analyse sollte nicht nur auf die transkribierten Bereiche der bekannten Gene beschränkt sein, sondern sich auch auf die Intergen-Regionen erstrecken und putativ regulativ wichtige Sequenzabschnitte identifizieren. Auch hierbei wurde eine mehrstufige Strategie angewendet, um eine möglichst umfassende Auswertung zu erreichen. Zum einen wurde die Charakterisierung der genomischen DNA über die Homologiesuche nach verschiedenen Datenbankeinträgen vorgenommen. Hierbei diente die BlastN- und BlastX-Analyse zum Vergleich mit bekannten Genen und annotierten EST- und cDNA-Sequenzen. Dies führte zur genauen Bestimmung der Exon-Intron-Grenzen der Gene *LMO1*|*Lmo1*, *TUB*|*Tub* und *Eif3* und zu deren neuen Spleißvarianten. Gleichzeitig konnten eine ganze Reihe von EST-Sequenzen der genomischen Sequenz zugeordnet werden (siehe Tab. 10). Ebenfalls über die Homologie wurden auch repetitive Motive (z.B. Alu-, LINE-Sequenzen) mit Hilfe der vier verschiedenen Programmen *CENSOR*, *REPEATMASKER*, *SST* und *XNUN* (siehe Kap. 3.5.1 und 3.10.3) identifiziert und für die Maskierung der Genomsequenz in nachfolgenden Analysen gespeichert. In einem zweiten parallel verlaufenden Schritt wurde über verschiedene Exonvorhersage-Programme (*GENSCAN, GRAIL2, MZEF, XPOUND*) putativ kodierende Bereiche berechnet. Da die Ergebnisse der verschiedenen Programm-Algorithmen meist sehr verschieden und somit nur von bedingter Aussagekraft waren, wurde durch eine Zusammenfassung der unterschiedlichen Programmergebnisse eine sicherere Vorhersage erreicht. Insbesondere die Zusammenstellung aller mehrfach, von mindestens drei Programmen errechneten Sequenzabschnitte, führte zu erheblichen Reduzierung der angesprochenen Bereiche. Durch diese Vorgehensweise konnten die insgesamt 229 putativen Exonbereiche der Humansequenz, bzw. die 527 Exonabschnitte der Maussequenz auf 24, bzw. 49 eingeschränkt werden (Tab. 11). Außerdem wurde der Informationsgehalt dieser vorhergesagten Exonabschnitte über eine gesonderte Homologiesuche

mit BlastN- und BlastX-Analyse versucht mit bereits bekannten Datenbankinformationen zu korrelieren.

Bezogen sich diese beiden ersten Schritte auf die Identifizierung von genkodierenden Abschnitten in der genomischen DNA, so wurde in einem dritten Schritt mithilfe einer statistischen Analyse der Basenpaarzusammensetzung CpG-Inseln und AT-reiche Sequenzbereiche charakterisiert. Ebenso wurde nach bestimmten Sequenzmotiven gesucht, die Hinweise auf die Existenz von Promotoren und Polyadenylierungssignalen (*PROSCAN*, *GENSCAN*) geben können. Insgesamt konnten auf diese Weise 48 unterschiedliche humane Promotorbereiche angegeben werden, von denen sich allerdings nur zwei Promotoren durch beide Programme vorhersagen ließen. Das gleiche Bild zeichnete sich auch bei der Analyse der Maussequenz ab, von insgesamt 41 errechneten Promotorstellen wurde nur eine Stelle durch beide Algorithmen vorhergesagt.

Um diese verschiedenen, einzeln sehr zeitintensiven Analyseschritte in einer effizienten Art und Weise durchzuführen, wurde mit Hilfe der Programm-Routine *RUMMAGE* gearbeitet, die alle obigen bioinformatischen „Tools" in einen automatisierten Prozess zusammenführte und die Ergebnisse der untersuchten genomischen DNA in einer komplexen Grafik aus Balken und Sequenzmarkierungen darstellte (siehe Abb. 6). Gleichzeitig speicherte dieses Programm die verschiedenen Einzelergebnisse in verschiedene Datenbanken zwischen, so dass diese Resultate für weitere Analyseschritte separat auswertbar waren. Somit bot das Programm *RUMMAGE* die ideale Grundlage für eine komplexe Sequenzanalyse. Die Problematik in der Ergebnisbewertung war die eindimensionale Betrachtungsweise der Resultate, die immer nur Einzelbefunde aus einem Genom darstellten. Viele der Ergebnisse konnten sich gegenseitig nicht bestätigen, so dass die Qualität der Aussage nicht ohne experimentelle Versuche beurteilt werden konnte. Somit war die vierte Ebene der Analyse, basierend auf dem komparativen Ansatz der vorliegenden Arbeit, zwei homologe Genombereiche verschiedener Spezies parallel zu sequenzieren und zu untersuchen, ein unverzichtbarer Schritt in der Gesamtanalyse. Dadurch konnte zu vielen Ergebnissen der evolutive Aspekt der Sequenzkonservierung eingebracht werden, was als Indikator zur Richtigkeit der Vorhersage gewertet wurde.

Für den genomischen Interspezies-Sequenzvergleich standen zwei unterschiedliche Betrachtungsweisen zur Verfügung. Wurde in der Dotplot-Analyse, die zwei zu vergleichende Sequenzen in einen zweidimensionalen Koordinatensystem gegenübergestellt, vor allem die Konservierung eines Sequenzabschnittes über einen zuvor festgelegten Schwellenwert in Form einer Diagonalen dargestellt, so zeigt die PIP-Analyse (Schwartz *et al.*, 2000) die Konservierung zur Referenzsequenz in differenzierterer Form mit dem Grad der Homologie von 50 – 100%. Einschränkend für diese Art der Analyse ist die nicht mögliche Darstellung von syntän angeordneten Sequenzkonservierungen in zwei unterschiedlichen Spezies, da die Lokalität der Homologie der zweiten Sequenz zur Referenzsequenz unberücksichtigt bleibt. Somit haben beide Analysen ihre Stärken, die methodenspezifische Aussagen zulassen. Nur in der Dotplot-Analyse fallen Insertionen oder Deletionen durch Stufen in der Diagonale

auf, bzw. gibt die Neigung der Diagonalen Auskunft über die Verteilung der Basen und signalisiert welche Sequenz kürzer, bzw. länger ist.

So konnte bereits durch diese Betrachtungsweise gezeigt werden, dass in der untersuchten Region die Maus-genomische Sequenz um durchschnittlich 14% kürzer ist als der homologe humane Genomabschnitt (vgl. Kap. 3.8.1.1). Ein Mittelwert der auch für die chromosomale Region 11p15.5 (Engemann *et al.*, 2000) und das gesamte Mausgenom zutrifft (Waterston *et al.*, 2002). Engemann und Mitarbeiter verglichen zwischen den Genen *CARS*|*Cars* und *KCNQ1*|*Kcnq1* eine 473 kb großen nicht-kodierenden humanen genomischen Bereich mit dem homologen nur 409 kb großen murinen Genomabschnitt. Im Verhältnis zueinander wies die Maussequenz eine Verkürzung von 13,5% auf. Auch Waterston und Mitarbeiter (2002) kamen nach Abschluss des Mausgenomprojekt bei ihrer globalen Analyse beider Genome auf ein Ergebnis von 13,8% (2,5 Gb Maus zu 2,9 Gb Mensch), um das das Mausgenom kleiner ist als das des Menschen.

Ein anderer Vorteil der PIP-Analyse war hingegen die Möglichkeit, beliebig viele zusätzliche Sequenzinformationen und -merkmale in die graphische Darstellung des Plots zu implementieren, so dass eine umfassende Analyse in sehr verdichteter Form graphisch visualisierbar war. In einer einzigen Grafik konnten dadurch sowohl die quantitative Auswertung des Homologiegrades wie auch putativ funktionelle Aspekte der Referenzsequenz skizziert werden. So wurden in Abb. 25 in der PIP-Analyse der sequenzierten Humansequenz nicht nur die Homologie zwischen 50% und 100% zur Maus-genomischen DNA mit den bekannten Exonsequenzen dargestellt, sondern auch gleichzeitig die Ergebnisse aus der *REPEATMASKER*-Analyse, der GC-Analyse, das kumulierte Ergebnis der Exonvorhersage-Programme, der EST-Homologiesuche und der Promotor-Analyse durch *PROSCAN* und *GENSCAN* und putative Poly-A-Bereiche eingezeichnet. Dabei bewährte sich die PIP-Analyse nicht nur als nützliches Hilfsmittel im Interspeziesvergleich von orthologen Sequenzbereichen; auch beim Vergleich der paralogen Chromosomenabschnitte innerhalb des Humangenoms lieferte sie interessante Informationen, wie bei der Gegenüberstellung der genomischen Bereiche auf Chromosom 12 der Gene *TULP3*, *LMO3* und *ΨEIF3S5* (siehe Kap. 3.9). Die PIP-Analyse ermöglichte zudem eine detaillierte basengenaue Betrachtung einzelner Subregionen in der Grafik, da mit Hilfe eines extra abgespeicherten Vergleichsprotokolls bis auf die Position der Einzelbase genau die Sequenzhomologie der PIP-Analyse dokumentiert und auswertbar war.

4.2 Vergleichende Sequenzanalyse proteinkodierender Genomabschnitte

Die in dieser Arbeit untersuchten proteinkodierenden DNA-Abschnitte zeigten im Interspezies-Vergleich zwischen Mensch und Maus mit meist über 90% Homologie eine äußerst hohe Konservierung der Nukleotidsequenz. Diese Ähnlichkeit zwischen den Spezies beschränkte sich nicht nur auf die Basenpaarabfolge und die Größe der Exons, sondern fand sich auch in der nahezu gleichen Genarchitektur mit im Verhältnis gleichgroßen Introns wieder. In der sequenzierten humangenomischen Sequenz konnten insgesamt 21 verschiedene Exons isoliert und mit denen der Maus verglichen werden. Fünf Exons kodierten für das Gen *LMO1* und 16 Exons für das Gen *TUB*. Da sich die hohe Sequenzkonservierung auch teilweise in den nichtkodierenden Intronbereichen nachweisen ließ, dürften auch diese DNA-Abschnitte funktionell eine sehr wichtige Rolle spielen. Insbesondere für regulative Vorgänge dürften sie von besonderer Bedeutung sein. Weiterführende Untersuchungen zeigten, dass die beiden Gene *LMO1* und *TUB* zu zwei sehr konservierten Genfamilien zählen, die bereits früh in der Evolution entstandenen und im Stoffwechsel der höheren Lebewesen von zentraler Bedeutung sein dürften.

4.2.1 Das LMO1-Gen

Für das Gen *LMO1*/*Lmo1* konnten insgesamt 6 verschiedene Exons beschrieben werden, die posttranskriptionell zu drei unterschiedlichen mRNAs miteinander verknüpft werden. Dem alternativen Spleißen unterliegt in beiden Spezies Mensch und Maus nur das erste Exon, das von zwei unterschiedlichen genomischen Bereichen kodiert wird; die Exons zwei bis vier sind in allen Spleißvarianten identisch. Kodiert der erste Bereich nur für das Exon 1a, so konnten für den zweiten genomischen Bereich die Exons 1b und 1c nachgewiesen werden. Dabei handelt es sich bei der Spleißform mit Exon 1c um eine im 3´-Bereich des ersten Exons verkürzte Exon 1b-Sequenz. Die Nutzung einer im Exon 1b internen Spleißdonorstelle führt zu einem vorzeitigen Beenden der Exonsequenz (vgl. Kap. 3.6.1.1). Trotz der unterschiedlichen Loci der Exons 1a und 1b/1c, die voneinander ca. 4,5 kb auf der genomischen DNA entfernt liegen, sind die translatierten Bereich bis auf einen einzigen Basenpaaraustausch miteinander identisch. Ein Genmuster, das sich in beiden Spezies exakt erhalten hat. Auch die verkürzte Spleißform C, deren Startcodon durch die Verkürzung der ersten Exonsequenz erst im Exon 2 zu finden ist, ist in beiden Spezies vorzufinden. Reguliert werden diese drei unterschiedlichen Genvarianten durch zwei unterschiedliche Promotoren, die mit der durchgeführten PromotorScan-Analyse nachgewiesen werden konnten. Dieses Ergebnis ist auch bei Boehm & Mitarbeitern (1988 und 1990) beschrieben. Auch sie wiesen den dualen Genstart in Mensch und Maus nach. Studien von Boehm & Mitarbeitern (1991) konnten in der Maus zeigen, dass es sich um zwei eigenständige, voneinander unabhängig aktivierbare Promotoren handelt, deren die Regulation entwicklungsspezifisch gesteuert werden kann. Sie konnten für die beiden unterschiedlichen Promotoren erste Hinweise für eine entwicklungsspezifisch unterschiedliche Aktivität

im sich entwickelten Gehirn liefern. Die Transkripte des ersten Promotors, bestehend aus dem Exon 1a, schienen ein Maximum für die Tage 12 bis 15 des Mausembryo zu haben, während für den zweiten Promotor vor Exon 1b ein Aktivitätshöhepunkt am Tag 18 dokumentiert werden konnte. Somit ist der Organismus in der Lage zwei unterschiedliche mRNA-Transkripte, kodierend für ein nahezu identisches Protein, unabhängig voneinander zu exprimieren. Ein Mechanismus, der auch bei anderen LMO-Genfamilienmitgliedern beschrieben werden konnte (siehe Kap. 4.3.1.1)

Trotz unterschiedlicher Regulation dürfte die Genfunktion für alle drei Spleißformen ähnlich, bzw. identisch sein, da die kodierenden Basen für die Genprodukt-charakteristische LIM-Domäne in allen Spleißformen ab dem Exon 2 konserviert zu finden sind. Die LIM-Domäne beschreibt ein Motiv, das aus zwei hintereinander geschalteten Kopien der Sequenzabfolge C-X_2-C-X_{17-19}-H-X_2-C-X_2-C-X_2-C-X_{7-11}-(C-)-X_8-C besteht. Die Abkürzung LIM leitet sich aus dem Akronym der drei Gene *lin-11,* ein in die Blastozysten-Entwicklung involviertes C. elegans-Gen (Freyd *et al.*, 1990), *Isl-1*, ein Insulin-Gen-Enhancer-bindendes Protein der Ratte (Karlsson *et al.*, 1990) und *mec-3*, welches bei der Neuronendifferenzierung von C. elegans eine Rolle spielt (Freyd *et al.*, 1990) ab. Jede LIM-Domäne ist für sich in der Lage ein Zink-Ion zu binden (Michelsen *et al.*, 1993; Archer *et al.*, 1994) und verleiht dem Protein die Fähigkeit mit anderen Proteinen zu assoziieren und zu interagieren (Schmeichel *et al.*, 1994). Da diesen LIM-Proteinen keinerlei DNA-bindenden Eigenschaften nachgewiesen werden konnten, zählen sie zu einer eigenen Klasse von transkriptionellen Regulatoren in der Zelle, die ihre Gen-Funktion über die intermolekulare kompetitve Bindung mit anderen DNA-bindenden Transkriptionsfaktoren ausüben.

Die eigene Analyse der transkribierten Nukleotidsequenz zeigte, dass alle Codons für dieses Motiv im identischen Teil der Spleißvarianten vorhanden sind und von den Exons 2 bis 4 kodiert werden. Ein Bereich der auf Proteinsequenzebene bei Mensch und Maus eine identische Aminosäureabfolge vorweist (siehe Abb. 33). Auch der Interspeziesvergleich mit *Fugu* zeigte interessanterweise für die genomische Nukleotidsequenz eine Konservierung genau für diese Exonbereiche (siehe Abb. 27a). Die beiden alternative gespleißten Bereiche der Exons 1a und 1b/1c waren in der PIP-Analyse der Genombereiche von Mensch und *Fugu* von nicht hervorgehoben (Kap. 3.8.2.1). Auffällig war indes die Homologie im Promotorbereich des Exons 1a, während der transkribierte Bereich des Exons 1a selbst keine Konservierung aufwies. Außerdem zeigten sich zwei evolutiv äußerst konservierte Bereiche im ersten großen Intron von *LMO1*, in Abschnitten die auch bei der komparativen Analyse zwischen Mensch und Maus mit hoher Homologie aufgefallen waren. Da die Homologie dieser beiden Abschnitte mit ca. 80% (siehe Tab. 20) sich in der gleichen Größenordnung bewegte wie der Promotorbereich im 5´UTR vor Exon 1a, könnte für diese Intronbereiche ebenfalls eine *LMO1*-Gen-assoziierte regulative Funktion postuliert werden.

Welche physiologische Rolle dieses evolutiv hoch konservierte *LMO1*-Gen im Organismus besitzt, lässt sich bisher nur grob beschreiben, bzw. indirekt über die deskriptive Namensgebung der weiteren Synonyme für dieses Gen erschließen. *LMO1* wird als putatives T-Zell-Onkogen angesehen und wurde im Rahmen von Untersuchungen der chromosomalen Translokation t(11;14) (p15;q11) einer T-Zell

akuten lymphoblastischen Leukämie (T-ALL) durch Takasaki & Mitarbeiter (1987) entdeckt. McGuire & Mitarbeiter (1989 und 1991) wiesen dem auch als *TTG-1* (T-Zell Translokations Gen-1) bezeichneten tumorassoziierten Gen aber nicht Thymus oder lymphoide Zelllinien als eigentliche Expressionsorte aus - in Bezug auf das Krankheitsbild der T-ALL war dies naheliegend - sondern identifizierten Zellen des zentralen Nervensystems. Durch *in situ*-Hybridisierung konnte verdeutlicht werden, dass Transkripte in den Rhombomeren, in pyramidalen Neuronen des sich entwickelten fötalen Gehirns der Maus zu finden sind (Boehm *et al.*, 1990 + 1991), wodurch das Gen einen weiteren Namen bekam und als Rhombotin 1 (*RBTN-1*) bezeichnet wurde. Insbesondere zeigte sich die Expression im cerebralen Neocortex, im Di- und Mesencephalon, im Thalamus, Hypothalamus, Cerebellum und im Auge (Hinks *et al.*, 1997). Die anfänglich nicht nachzuweisende Expression in Thymus und T-Zellen konnte erst nach sehr langer Exposition von Northern-Hybridisierungen gezeigt werden. Ebenso zeigten die Organe Lunge, Niere, Uterus, Plazenta ein schwaches Vorkommen des *LMO1*-Genprodukts, welches im Vergleich zur Expression im Gehirn mengenmäßig aber vernachlässigt werden kann (Boehm *et al.*, 1991).

Die vorliegenden Daten zeigen somit, dass *LMO1* über seine proteinbindende Eigenschaft der LIM-Domäne mit anderen DNA-bindenden Transkriptionsfaktoren interagiert und regulative Funktion bei der Transkription überwiegend in neuronalen Zellen und auch in Zellen des Immunsystems beteiligt ist. Das diese Regulation insbesondere in Immunzellen äußerst spezifisch gesteuert werden kann, zeigen Studien von Herbolt & Mitarbeitern (2000), die eine *LMO1*-Expression in T-Lymphozyten und eine Steuerung der T-Zell-Proliferation und -Differenzierung in Zusammenhang mit SCL-(stem cell leukemia)-Transkriptionsfaktoren dokumentieren konnten. Der SCL-Komplex wirkt regulativ auf T-Zell-spezifische Gene, die wichtig für die Synthese der T-Zell-Rezeptoren (TZR) und das TZR-Signalgebung sind. Ebenso ist der *LMO1-SCL*-Komplex bei der Expression der invarianten α-Kette pTα und beim Rearrangement der TZRβ-Kette beteiligt (Herbolt *et al.*, 2000). Allerdings ist in dieser Funktion *LMO1* nicht alleine involviert. Als zweiten nukleären Partner wurde ein weiteres Genfamilienmitglied, das *LMO2*-Gen beschrieben. Wie im nachfolgenden Kapitel ausgeführt, zeichnen sich die Mitglieder der LMO-Genfamilie durch eine ganze Reihe von gemeinsamen genomischen Aspekten aus, die auf eine ähnliche Funktionalität im Organismus schließen lassen.

4.2.1.1 Die LMO-Genfamilie

Der in dieser Arbeit durchgeführte Homologievergleich führte sehr rasch zu weiteren Transkripten, die die *LMO1*-Gen typische tandemartige Doppelanordnung der LIM-Domäne besitzen. Insgesamt zeigten sich drei weitere Familienmitglieder (*LMO2*, *LMO3* und *LMO4*), die sich nicht nur auf Proteinebene, sondern auch auf mRNA-Ebene in vielen Aspekten sehr ähneln. Sowohl die Zahl der Exons, das Spleißmuster, wie auch die Größe des offenen Leserahmens – von 146 (*LMO3*) bis 165 Aminosäuren (*LMO4*) - zeigen die enge Verwandtschaft der vier Gene untereinander an. Auffällig ist, dass die Spleißstellen in allen vier Genen insbesondere zwischen den beiden LIM-Domänen 1 und 2 konserviert vorhanden sind. Der höchste Homologiegrad mit ca. 90% auf Aminosäureebene findet sich zwischen

den Genen *LMO1/Lmo1* und *LMO3/Lmo3* (Foroni *et al*., 1992). Auch die gleiche Genarchitektur, bestehend aus jeweils fünf Exons mit alternativ gespleißten 5´-Bereich, ist bei beiden Genen nachzuweisen. Ebenso wie das *LMO1* besitzt auch *LMO3* zwei alternativ genutzte erste Exons. Interessanterweise wies das *LMO3*/*Lmo3* aber nur die um 11 Aminosäuren verkürzte Variante in Homologie zum *LMO1c/Lmo1c* auf (vgl. Abb. 32). Dies kann als Hinweis dafür gedeutet werden, dass die beiden LIM-Domänen die hauptsächliche Funktion der Gene bestimmen und der aminoterminale Bereich der Proteinsequenz mehr von regulativer Bedeutung ist. Dies individuellere Sequenzstruktur könnte dann mit dem veränderten Expressionsmuster von *LMO1* und *LMO3* in Verbindung gebracht werden.

Tab. 27 Tabellarische Betrachtung der Gemeinsamkeiten und Unterschiede der vier beschriebenen LMO-Gene. Alle vier Gene weisen neben der tandemartig duplizierten LIM-Domäne in beiden Spezies einen sehr ähnlichen genomischen Aufbau bestehend aus 5, bzw. 6 verschiedenen Exons auf, deren 5´-Bereich immer in zwei alternativen Spleißvarianten vorliegt. **Abk.:** altn. Ex = alternative Exonsequenzen; AA = Aminosäuren; Fkt. = bekannte Genfunktion.

Gen	Zahl der Exons	Länge		Homologie zum LMO1 auf Proteinebene	Fkt.	Locus	
		cDNA/ transkribierter Bereich	AA			Mensch	Maus
LMO1	5/ 2 altn. Ex 1	1,2/1,4 kb/ 39,5 kb	156	./.	Akute T-Zell-Leukämie (T-ALL), Proto-Oncogen	11p15.3	7
LMO2	6/ Ex 1 + 2 können fehlen, sind nicht kodierend	1,7 kb/ 33,7 kb	158	60%	Wichtig für die Hämato-poese, bindet an basisches helix-loop-helix-Protein TAL-1	11p13	2
LMO3	5/ 2 altn. Ex 1	1,5/2,7/4,3 kb	146	89%	Proto-Oncogen, CNS-Entstehung	12p13	6
LMO4	6/ 2 altn. Ex 1 sind nicht kodierend	17,1 kb	165	55%	Proto-Oncogen, überexprimiert in Brustkrebs	1p22.3	3

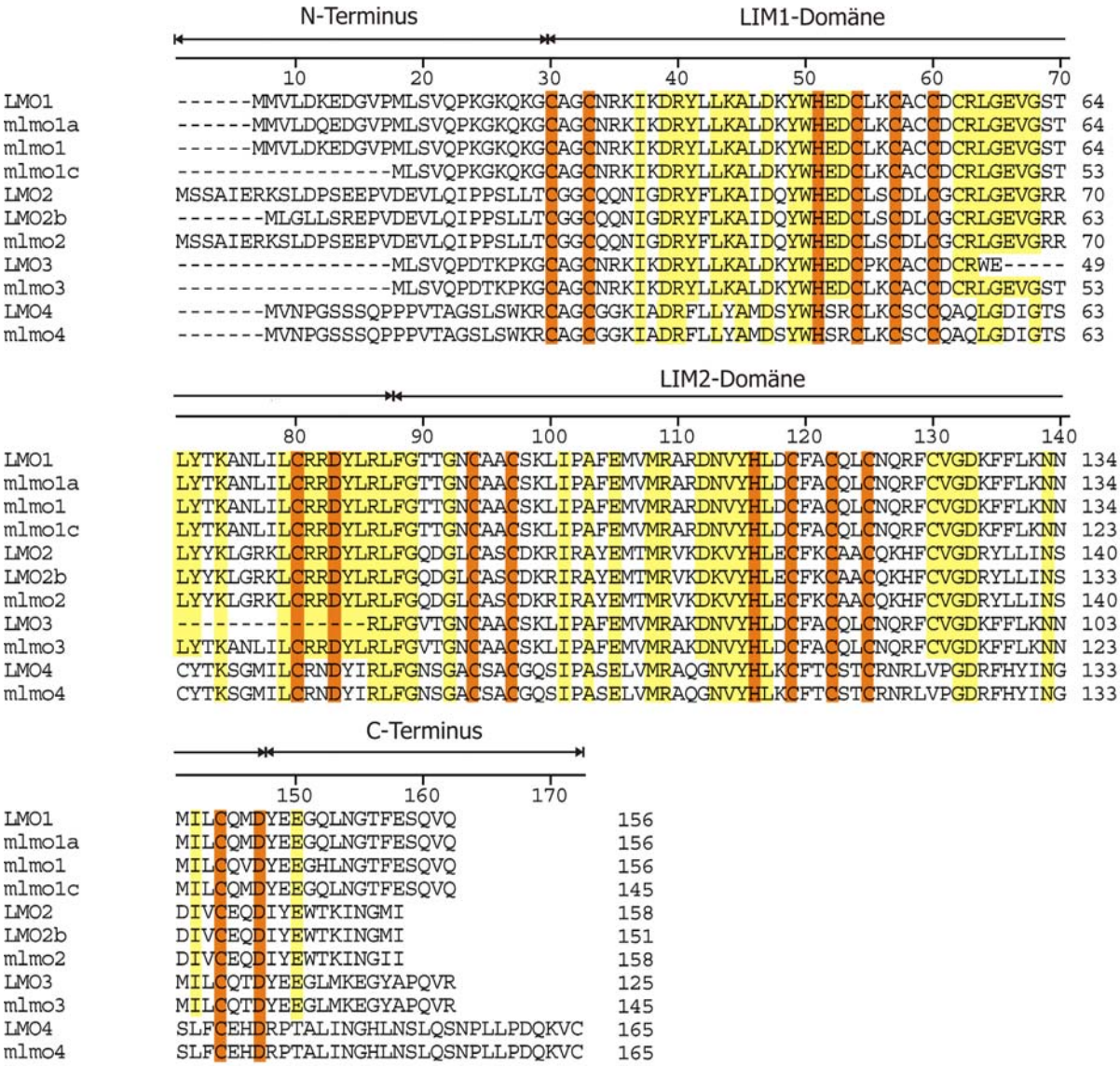


Abb. 32 Homologie- und Interspeziesvergleich der Proteinsequenzen der vier humanen und murinen LMO-Gene (LMO1/Lmo1 bis LMO4/Lmo4) zusammen mit den drei in dieser Arbeit beschriebenen LMO1-Spleißvarianten (siehe Abb. 17 und Tab. 13) (aufgrund der Proteinsequenzgleichheit sind beispielhaft nur die drei murinen Sequenzen = mlmo 1a/1b/1c (siehe Tab. 14) dargestellt). Außerdem wurde eine weitere kurze LMO2-Spleißvariante (LMO2b) des Menschen dargestellt. Die konservierten Aminosäuren der beiden LIM-Domänen sind farbig gelb hervorgehoben. Die Zink-bindenden Cysteine sind orange markiert. Die humanen LMO-Proteinsequenzen wurden nach Tse et al. (1999) dargestellt.

Ein Ergebnis des paralogen Sequenzvergleiches mit Homologie im Intron 3-Bereich der beiden LMO-Gene auf Chromosom 11 und 12, dürfte funktionell nicht relevant sein, da diesem konservierten Sequenzabschnitt ein repetitives Element zugeordnet werden konnte (Kap. 3.9.1) und dieser Bereich im Interspeziesvergleich zwischen Mensch und Maus nicht mehr auffällig war. Vielmehr könnte es sich bei dieser Konservierung um ein Relikt aus der gemeinsamen Evolution der beiden chromosomalen Bereiche auf Chromosom 11 und 12 handeln, die aus einer einstigen Genduplikation im Rahmen eines chromosomalen Rearrangements aus einem gemeinsamen Vorläufergen entstanden sein könnten (Endo *et al.*, 1997)(Kap. 4.5.2).

Ähnlich den Genen *LMO1* und *LMO3* zeigen auch die Gene *LMO2* und *LMO4* in beiden Spezies einen höheren Verwandtschaftsgrad zueinander. Beide Gene besitzen Spleißvarianten, deren erste beiden Exons keine proteinkodierende Funktion besitzen. Beide Gene bauen sich aus 6 Exons auf, wobei ebenfalls zwei unterschiedliche erste Exons beschrieben wurden. Auch *LMO2* ist im Besitz von zwei unterschiedlichen Promotoren; einem vor Exon 1 und einem vor Exon 3 (Royer-Pokora *et al.*, 1995). Der zweite Promotor führt dabei zu einem Transkript, dem die beiden ersten Exons 1 und 2 fehlen. Wahrscheinlich wird die sehr differenzierte Genexpression, die räumlich und zeitlich in leicht versetzten Fenstern stattfindet, durch diese unterschiedlichen 5´-Bereiche geprägt. Die Betrachtung des *LMO4*|*Lmo4*-Gens zeigte zudem, dass das vierte LMO-Familienmitglied mit der geringsten Homologie zu *LMO1* am weitesten verwandt mit allen anderen LMOs zu sein scheint (siehe Abb. 33). Dies zeigt sich auch in dem eher ubiquitären Expressionsmuster von *LMO4*, das schwach ausgeprägt nur *LMO2* zeigt, und eher untypisch für die anderen LMO-Gene ist. *LMO1* und *LMO3* zeigen eine sehr differenzielle Expression und lassen sich in der frühen Embryogenese und hauptsächlich im Gehirn nachweisen (Foroni *et al.*, 1992).

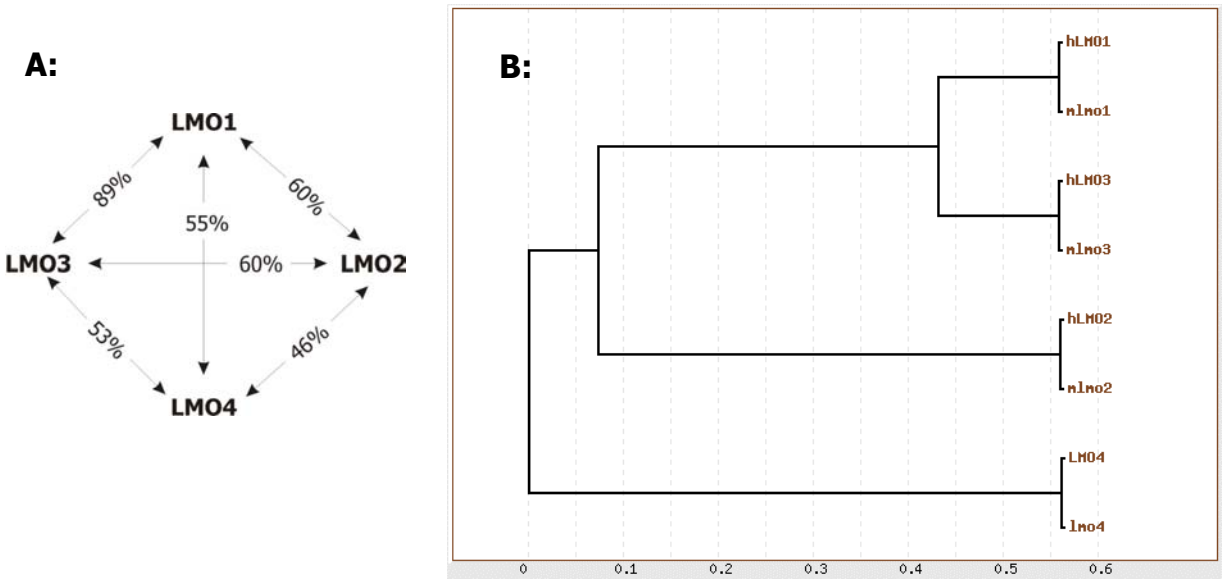


Abb. 33 **Aminosäuresequenzhomologien der vier humanen LMO-Gene. A:** Das Diagramm zeigt den Prozentanteil der identischen Aminosäuren der verschiedenen LMO-Gene untereinander. Als Referenz dienten die humanen Proteinsequenzen mit der Acc.-Nr. AAH69673 (= *LMO1*), AAH42426 (= *LMO2*), NP_061110 (= *LMO3*) und P61968 (= *LMO4*). **B:** Die vergleichende Analyse der humanen und murinen Aminosäuresequenzen führte zu folgenden genealogischen Beziehungen. Während *LMO1*|*Lmo1* und *LMO3*|*Lmo3* am engsten miteinander verwandt zu sein scheinen, ist *LMO4*|*Lmo4* am weitesten von den anderen LMO-Gene entfernt. Die Homologie der Proteinsequenz im Interspeziesvergleich zwischen Mensch und Maus zeigte bei *LMO3*|*Lmo3* und *LMO4*|*Lmo4* eine 100%ige Sequenzübereinstimmung; bei *LMO2*|*Lmo2* betrug der Unterschied eine, bzw. bei *LMO1*|*Lmo1* zwei Aminosäuren.

Allen chromosomalen Loci der LMO-Gene ist gemein, dass sie sich in Regionen befinden, die mit der Tumorentstehung (Lymphome, Leukämien, Hirn, Leber, Niere und Lunge) in Zusammenhang gebracht wurden und dort entweder deletiert, bzw. sich in Bruchpunktregionen von Translokationsereignissen

befinden (Tse et al., 1999). Eine weitere Übereinstimmung ist die für *LMO1* schon angesprochene Assoziation mit der T-Zell-Differenzierung. Sowohl *LMO1* (Boehm et al., 1988), als auch *LMO2* (Rabbitts et al., 1999) und *LMO4* (Tse *et al.*, 1999) zeigen eine Beteiligung an der Ätiologie von Leukämien, die für alle vier Gene gilt. Somit dürften alle Gene zumindest in diesem Aspekt als Proto-Oncogene angesehen werden.

Kenny & Mitarbeiter (1998) diskutieren daher den gemeinsamen Ursprung von allen LMO-Genen, die sich zusammen mit den *LHX*-Genen (LIM-Homeobox-Gene) aus einem gemeinsamen primordialen Vorläufer entwickelt haben könnten. Sie begründen dies mit der hohen Sequenzhomologie untereinander und den gemeinsamen funktionellen Eigenschaften, wie ihre nukleäre Lokalisation, ihre hochaffine Interaktion mit NLIs (Nuclear LIM Interactors) und die Verzahnung mit dem Transkriptionskomplex. In *Drosophila* konnte z.B. bisher nur ein *dLMO*-Gen nachgewiesen werden (Zhu *et la.*, 1995), welches als ursprüngliches Modell für die Funktionsweise der humanen und murinen LMOs angesehen werde kann. Auffällig ist bei *Drosophila*, dass die genomische Organisation der Exon-Intron-Grenzen im konservierten Muster von vier Exons mit den gleichen Spleißstellen innerhalb der kodierenden Sequenz wie beim *LMO1*/*Lmo1* zu finden ist. Auch der in dieser Arbeit durchgeführte Vergleich mit Genomsequenzen aus *Fugu* (Kap. 3.8.2.1) zeigte die hohe Konservierung des kodierenden Bereiches für die beiden Homöodomänen des LIM-Motivs. Im Gegensatz zu *Drosophila*, scheint *Fugu* allerdings schon im Besitz von mehreren *LMO*-Genen zu sein, da insgesamt drei unterschiedliche Genomsequenzen (Abb. 26) Homologie zum *LMO1* des Menschen zeigten und in der PIP-Analyse (Abb. 27a) durch Mehrfachbanden im Bereich der Exons auffielen. Diese Konstanz in der Genarchitektur scheint demnach zwingend im funktionellen Zusammenhang mit der physiologischen Rolle dieser Gene im Organismus zu stehen, so dass die Funktion als transkriptioneller Regulatoren vor allem während der Embryogenese als zentrale biologische Rolle angesehen werden kann. Durch die besondere Eigenschaft der Protein-Protein-Interaktion stellen die LMOs wahrscheinlich universelle Cofaktoren dar, die der Organismus gewebsspezifisch regulieren und so die Differenzierung bestimmter Zellen steuern kann.

Der in dieser Arbeit vorgenommene genomische Vergleich der *LMO1*-Genregion lieferte insbesondere für die kodierenden Sequenzbereiche konservierte Sequenzmotive, die sowohl bei der Maus wie auch bei *Fugu* zu finden waren, als auch Homologien zu allen weiteren Mitgliedern der Genfamilie. An bestimmten Sequenzbereichen wie etwa den Promotorregionen konnte gezeigt werden, dass diese zwar im Interspeziesvergleich konserviert waren, aber nicht mehr im Vergleich mit den paralogen Genregionen, d.h. zu anderen LMO-Familienmitglieder. Dies lässt den Schluss zu, dass es sich bei den LMO-Genen um ein evolutiv relativ alte Genfamilie handeln muss, die schon relativ früh in der Evolution durch Duplikationsereignisse eines primordialen Vorläufergens, ähnlich z.B. dem *Drosophila*-LMO-Gen, entstanden sein könnte und die sich danach bezüglich ihrer Regulation rasch weiterentwickelt und differenziert haben müssen, um etwa Dosiseffekte von ein und demselben Genprodukt zu vermeiden. Die so entstandenen LMO-Familienmitglieder behielten zwar ihre prinzipielle Genfunktion als DNA-bindende Transkriptionsfaktoren bei, spezialisierten sich aber

bezüglich ihres Expressionsorts und Zeitpunkts. Wahrscheinlich aufgrund ihrer wichtigen neuen Stellung im Organismus der höheren Lebewesen blieben sie während der darauffolgenden Evolution sehr konserviert. Dies würde die Interspezieskonservierung im regulativ relevanten Bereich der einzelnen Familienmitglieder erklären und für die Diversität dieser regulativen Bereiche untereinander sprechen. Somit dürfte die weitere Charakterisierung vor allem der Promotorbereiche unter dem Aspekt der Regulierung durch Enhancer oder anderen Mediatoren in Bezug auf deren genomisches Umfeld insbesondere in konservierten nicht-kodierenden DNA-Bereichen zu interessanten Ergebnissen führen. Unter Umständen lassen sich Regulationsstrukturen beschreiben, die weiterführend dann auch mit der Entstehung von Leukämien diskutiert werden können. Dies setzt Untersuchungen voraus, die erst mit Kenntnis der genomische Sequenz durch diese Arbeit erfolgreich durchgeführt werden können.

Weitere Familiemitglieder, die in Aufbau und Organisation der LMO-Gene entsprechen, sind bis heute keine bekannt. Zwar findet sich als Datenbankeintrag für die Chromosomenregion Xp11.23 das Gen *LMO6* (Acc.-Nr: NM_006150), doch weicht es von seiner Struktur durch drei konsekutive LIM-Domänen von den übrigen vier LMOs ab. Auch ist die mit 616 Aminosäuren lange Proteinsequenz, kodiert von 9 Exons, untypisch im Vergleich zu den anderen LMO-Genen. Putilina & Mitarbeiter (1998) charakterisierten ein weiteres, als *LMO7* bezeichnetes Gen, das nur in seinem C-terminalen Ende eine LIM-Domäne aufweist und auf Chromosom 13q12 kartiert werden konnte (Kurihara *et al.*, 2002). Da das Transkript zusätzlich im N-terminalen Abschnitt eine PDZ-Domäne (= PSD-95, *Discs-large*, ZO-1)-Domänen, die ebenfalls für Protein-Interaktionen und für die Bindung mit Rezeptoren verantwortlich ist und eine Assoziation zum Zytoskelett nahe legt (Fanning *et al.*, 1999), ist die Nomenklatur des Gens streng genommen nicht ganz korrekt, da es sich hier nicht mehr nur um ein „*LIM-domaine only*"-Gen handelt, das keine zusätzlichen Motive besitzt.

4.2.2 Das TUB-Gen

Das humane *TUB*-Gen konnte im Rahmen dieser Arbeit durch die Sequenzierung seines genomischen Bereiches in die chromosomale Bande 11p15.3 kartiert werden. Eine Lokalisierung, die die publizierten zytogenetischen Daten korrigierte, welche das Gen in die chromosomalen Bande 11p15.5 verwiesen (Mapping Information des TUB-UniGene Clusters Hs.54468 unter <http://www.ncbi.nlm.nih.gov/UniGene/>). Ebenso zeigte die Kartierung, dass sich das humane *TUB* in einer Entfernung von 268 kb distal zum STS-Marker D11S932 befindet und nicht wie durch North & Mitarbeitern (1997) beschrieben proximal zu diesem STS-Marker. Auch für das murine Genom konnte durch die genomische Sequenzierung der *Tub*-Genlocus, der durch Noben-Trauth & Mitarbeiter (1996) grob mit der Umgebung des STS-Markers D7Mit219 auf Chromosom 7 angegeben wurde, exakt bestimmt und mit einer Entfernung von 65.147 bp zwischen D7Mit219 und dem 3´-Ende des Gens angegeben werden.

Insgesamt wurden für das *TUB*/*Tub*-Gen im Rahmen dieser Arbeit 16 verschiedene Exonsequenzen in beiden Spezies Mensch und Maus bestimmt. Jeweils zwei Exons waren dabei in beiden Genomen

speziesspezifisch, d.h. für das humane *TUB*-Gen konnten zwei Exonbereiche charakterisiert werden, die kein nachweisbares murines Homolog hatten und *vice versa*. Insgesamt zeigten sich in beiden Spezies vier unterschiedliche Spleißvarianten, die sich im ihrem 5´-Bereich, bezüglich der ersten zwei Exons unterschieden oder im 3´-Bereich different waren. Eine weitere Spleißvariante, betraf das alternative Spleißen des internen Exons 5. Für jeweils eine Spleißvariante aus den 5´-Bereich, bzw. 3´-Bereich konnte im Genom der anderen Spezies kein entsprechendes Homolog charakterisiert werden (Vgl. Abb. 19).

4.2.2.1 Verschiedene Spleißvarianten des TUB/Tub-Gens

Mit Hilfe der cDNA-Sequenzen Acc.-Nr. NM_003320 (Kleyn *et al.*, 1996) und Acc.-Nr. U82467 (North *et al.*, 1997) ließen sich die Spleißvarianten A und C des Menschen bestätigen. Der Sequenzunterschied betraf den Genanfang, der bei Spleißvariante A aus einem Exon (1c), bzw. bei Spleißvariante C aus zwei Exons (1a + 1b) besteht. Die Folge ist ein unterschiedlicher Translationsstart, da beide Anfangs Exons im Besitz eines Startcodons sind. Bei der Spleißvariante C werden insgesamt 68 Codons des ORFs von den Exons 1a und 1b kodiert. Bei der Spleißvariante A umfasst der offene Leserahmen des Exons 1c nur 13 Codons, so dass die beiden unterschiedlichen mRNAs, die für 561 Aminosäuren (Variante B), bzw. 506 Aminosäuren (Variante A) kodieren, im N-terminalen Bereich um 55 Aminosäuren differieren. Erst ab dem Exon 2 sind beide Spleißvarianten in ihrem offenen Leserahmen wieder identisch.

Im Interspeziesvergleich mit der Maus zeigten sich die cDNA-Sequenzen Acc.-Nr. U52433 (Noben-Trauth *et al.*, 1996) und Acc.-Nr. U54643 (Kleyn *et al.*, 1996) als die murinen Homologen zur humanen Spleißvariante A. Lediglich ein unterschiedlich langer 3´-UTR unterschied die beiden cDNAs voneinander. Das murine Homolog zur humanen Variante B konnte nicht als cDNA-Sequenz aus den Datenbanken, bzw. der Literatur entnommen werden. Erst nach der in der vorliegenden Arbeit erfolgten Sequenzierung der Maus-genomischen Sequenz erlaubte der Interspeziesvergleich die Auswahl von zwei DNA-Bereichen, die mit einer Homologie von 92% zu den humanen Exons 1a und 1b auffielen. Durch Amplifikation dieser Abschnitte über RT-PCR konnten die beiden neuen Exonsequenzen verifiziert werden. Bei Betrachtung der genomischen Architektur dieser Spleißvariante B/b in beiden Spezies stellte sich heraus, dass diese Exons, die ihrerseits nur durch eine Intronsequenz von 616 bp (Mensch), bzw. 598 bp (Maus) getrennt sind, über ein großes zweites Intron von der übrigen Gensequenz distanziert sind. Aufgrund der enormen Größe konnte diese Intronsequenz nur im Mausgenom mit 47.009 bp basengenau dargestellt werden. Im Humangenom lagen die Exonsequenzen 1a und 1b außerhalb der sequenzierten Klone PAC12G13 und cSRL119g5 und kamen erst im Anschlussklon cSRL82e3 zum liegen (Vgl. Abb. 10), der im Rahmen dieser Arbeit lediglich kartierten werden konnte. Somit konnte die Größe der Intron 2-Sequenz im Menschen nur in Referenz zum Mausgenom auf ca. 50 kb geschätzt werden.

Als dritte alternative Spleißvariante für den Gen-5´-Bereich konnte für die Maus mit Hilfe der cDNA-Sequenz Acc.-Nr. U52824 (Noben-Trauth *et al.*, 1996) eine mRNA charakterisiert werden, die ebenfalls aus zwei Anfangsexons (1d und 1e) besteht, die aber nur in einer Entfernung von 2.488 bp zum Exon 2 liegen. Weder die EST-Datenbanksuche, noch der durchgeführte Interspeziesvergleich mit der orthologen Genomsequenz des Menschen, der keine signifikanten Homologien zum erwarteten humanen DNA-Bereich erbrachte, ließen auf die Existenz einer homologen humanen Spleißvariante C schließen. Lediglich zwei Bereiche mit 66% und 68% Homologie konnten für die humane Genomsequenz eingegrenzt werden (Vgl. Abb. 20). Allerdings befanden sich diese beiden Bereiche nicht in der erwartenden Entfernung von ca. 2,5 kb bis 2,7 kb, sondern waren in einem kürzeren Abstand von 2.250 bp zu finden. Auch der Versuch deren Existenz auch auf cDNA-Ebene zu bestätigen, führte zu keinem Ergebnis. Trotzdem kann das Vorhandensein dieser Spleißvariante im humanen Transkriptom nicht ausgeschlossen werden. Da der offene Leserahmen dieser Spleißvariante erst in Exon 3 beginnt, repräsentieren diese beiden Exons den 5´-UTR. Es wäre denkbar, dass durch die fehlende kodierende Funktion der Exons 1d und 1e diese sich während der getrennten Evolution von Mensch und Maus so stark verändert haben, dass sie über ihre Homologie nicht mehr deutlich zu identifizieren sind. Das negative RT-PCR-Ergebnis könnte durch eine gewebsspezifischen Expression dieser Spleißvariante z.B. ausschließlich in Testis - dieses war das Ursprungsgewebe aus dem cDNA-Sequenz Acc.-Nr. U52824 isoliert worden war (Noben-Trauth *et al.*, 1996) - erklärt werden, da für die RT-PCR zum Nachweis des humanen Homologs nur cDNA aus den Geweben Niere und Gehirn zur Verfügung stand.

Eine vierte Spleißform (Variante B in Abb. 19) wurde durch die *TUB*/*Tub*-mRNA mit fehlendem Exon 5 beschrieben. Bereits Kleyn & Mitarbeitern (1996) veröffentlichten diese alternative Genvariante, die sie u.a. bei den Mäusestämmen *Mus spretus* und *Mus castaneous* vorfanden. Trotz der daraus nach Translation des ORFs resultierenden Verkürzung des Proteins um 56 interne Aminosäuren - es fehlen die Aminosäuren 153-208 des ORFs aus Acc-.Nr. U54643 - zeigten die Tiere keinerlei phänotypische Auffälligkeiten. Dadurch wurden diese beiden Spleißvarianten als allele Formen des *Tub*-Gens interpretiert. Interessanterweise fiel Exon 5 auch im Interspeziesvergleich dieser Arbeit mit Genomsequenzen von *Fugu rubripes* durch seine nicht vorhandene Homologie auf, wohingegen alle weiteren Exons mit Ausnahme des ersten konserviert geblieben sind (Vgl. Abb. 27b). Dies lässt die Hypothese zu, dass der durch Exon 5 kodierte Proteinbereich nicht essentiell für die physiologische Funktion von *TUB* ist und dass es sich bei dieser Spleißvariante um die evolutiv ältere Form des Gens handeln dürfte. In diese Interpretationsweise würde ferner das Ergebnis des Homologievergleichs mit der paralogen Genomregion auf Chromosom 12p13 mit dem Gen *TULP3* passen (Vgl. Abb. 28, B). Auch bei dieser Analyse konnte keine Konservierung zum Exon 5 festgestellt werden. Die angrenzenden Exons 4 und 5 zeigten ebenfalls keine Homologien zum transkribierten genomischen Abschnitt des *TULP3*-Gens (Acc.-Nr. AC005911), was auf einen noch größeren nicht essentiellen Bereich deuten könnte. Alle publizierten *TUB*/*Tub*-mRNA-Sequenzen wiesen das gleiche Genende bei Exon 12 auf; lediglich die Länge des 3´-UTRs variierte in den Einträgen. RT-PCR-Experimente und Homologievergleiche mit Einträgen aus EST-Datenbanken konnten im Rahmen der vorliegenden Arbeit

aber einen neuen 3´-Bereich des humanen *TUBs* mit zwei zusätzlichen Exonsequenzen (Exon 13 und Exon 14) beschreiben, die in ca. 32 kb Entfernung zum Exon 11 von der genomischen DNA kodiert werden. Gleichzeitig durchgeführte Experimente, diese Spleißform auch für die Maus nachzuweisen, führten zu keinem Ergebnis. Die Besonderheit dieser *TUB*-Spleißvariante (Variante D in Abb. 19) ist das Fehlen des Exons 12 und dass sich der ORF von Exon 11 nach Exon 13 – unter Auslassung des Exon 12 – über ein großes Intron bis zum dortigen Stoppcodon erstreckt. Exon 14 kodiert lediglich für den 3´-UTR des Gens. Dies hat zur Folge, dass nach Translation dieser Spleißvariante dem TUB-Protein die letzten 44 Aminosäuren fehlen und durch 10 alternative neue ersetzt sind. Welche Auswirkung diese Veränderung auf Transkriptebene für die Wirksamkeit und Funktion des translatierten Proteins haben könnte, kann z.B. mit dem Phänotyp der dem Gen namensgebenden Tubby-Mausmutante mit Adipositas und neurosensorischen Defekten verglichen werden, und zum anderen in Bezug zu bereits bekannten funktionellen Eigenschaften des TUB-Proteins als Transkriptiosregulator interpretiert werden (siehe Kap. 4.2.2.2). Genomisch auffällig für das *TUB*|*Tub*-Gen mit seinen verschiedenen alternativen Spleißformen ist die enorme Größe des transkribierten, exonkodierenden Genombereiches, der im humanen Genom eine Distanz von über 100 kb überspannt und erst nach Sequenzierung mehrer Klone (PAC 12G13, cSRL199g5 und cSRL82e3) bestimmt werden konnte. Für das murine Genom konnte „nur" ein Bereich von 58 kb beschrieben werden, da die humane 3´-Genvariante D sich in der Maus nicht nachweisen ließ. Insbesondere die ersten und letzten Exons zeichneten sich durch sehr große Intronsequenzen und die dadurch hervorgerufene räumlich Trennung zum zentralen Genbereich aus. Es zeigte sich eine neue Genarchitektur, die für das *TUB*|*Tub*-Gen in dieser Form noch nicht beschrieben war und erst mit Vorliegen der Genomsequenz aus dieser Arbeit ersichtlich wurde.

4.2.2.2 Mögliche funktionelle Aspekte der TUB/Tub-Speißvarianten

Das murine *Tub*-Gen wurde bekannt durch den prominenten Phänotyp der Tubby-Mäuse, die sich durch eine ausgeprägte Fettleibigkeit (Adipositas) (*Coleman & Eicher*, 1990) und verschiedene neurodegenerative Erscheinungen der Retina und der Cochlear des Innerohres auszeichneten (Ohlemiller *et al.*, 1995). Je nach Stärke der Ausprägung können im fortgeschrittenen Stadium diese Veränderungen bis zur Erblindung und zur Taubheit der betroffenen Mäuse führen. In ursächlichen Zusammenhang mit diesem Phänotyp steht eine Mutation, die durch einen Einzelbasenaustausch im 3´-Bereich des *Tub*-Gens charakterisiert wird (Noben-Trauth *et al.*, 1996). Es zeigte sich an der Spleißdonorstelle des vorletzten Exons 11 eine Transversion von G nach T, die zu einem Spleißdefekt und zu einem Nichtentfernen des letzten Introns führt. Stattdessen zieht sich der offene Leserahmen des *Tub*-Gens von Exon 11 in die Intronsequenz bis zum nächsten dortigen Stoppcodon hinein. Das Resultat ist nach Translation dieses mutierten ORFs ein trunkiertes Protein, welches anstelle der ursprünglichen letzten 44 Aminosäuren 19 alternative, intronkodierte Aminosäuren besitzt. Weitere Veränderung auf genomischer Ebene sind in den betroffenen Mäusen nicht beschrieben. Dieser Zusammenhang zwischen veränderten Transkript und veränderter Proteinfunktion ist insbesondere für

die Bedeutung der alternativen Spleißform D mit den neuen 3´-Exons 13 und 14 sehr interessant, da dieser Genvariante ebenfalls der 3´-kodierende Bereich des Exons 12 fehlt.

Die offensichtlich hohe biologische Bedeutsamkeit des 3´-Genbereichs zeigte sich bereits im Interspezies-Homologievergleich des *TUB*|*Tub*-Gens zwischen Mensch und Maus. Die höchste Basenpaar-Konservierung war im kodierenden Genabschnitt für den C-Terminus zu verzeichnen. Auch im Vergleich mit *Fugu* waren die Exons 6 bis 12 mit einer Homologie zwischen 65% und 85% konserviert geblieben (Vgl. Tab. 20). Selbst im paralogen Vergleich mit dem TUB-Verwandten *TULP3* ist eine Homologie zu den Exons 7 bis 12 zu erkennen (Vgl. Abb. 28). Diese evolutive Konservierung auf genomischer Ebene über die Speziesgrenze hinweg findet sich auch auf Aminosäureebene wieder und charakterisiert die sog. Tubby-Domäne, die den C-terminalen Bereich von AS 243 bis 505 umfasst. Gel-shift-Experimente durch Boggon & Mitarbeitern (1999) konnten diesem Bereich die Fähigkeit der Bindung von doppelsträngiger DNA zuweisen. Es scheint, dass das Tubby-Protein dabei die DNA nicht über unspezifische elektrostatische Interaktion bindet, sondern eine sequenzspezifische Verbindung eingeht. Diese Basenspezifität dürfte einen funktionellen Charakter besitzen, der zum einen eine Erklärung für die starke Interspezieskonservierung wäre und zum anderen auch die Einzigartigkeit der Tubby-Domäne begründen könnte, da sie keiner der bekannten DNA-bindenden Proteinklassen mit „Helix-turn-helix"- bzw. „B-zip"-Motiv ähnelt. Eine weitere Interpretations-möglichkeit bietet die modellhafte Rekonstruktion der 3D-Struktur der C-terminalen Domäne durch Boggon & Mitarbeitern (1999). Sie entwarfen ein Modell, in dem die Domäne aus sechs Alpha-Helices und zwölf Beta-Faltblatt-Strukturen besteht, die sich zu einem Zylinder formen, der durch alternierend ausgerichtete Faltblätter gebildet wird (siehe Abb. 33A). Der hydrophobe Innenbereich dieses Zylinders wird durch eine zentrale Helix (in Abb. 12: H12, die letzte Alpha-Helix des C-Terminus) ausgefüllt, die sich axial in den Innenraum das Zylinders erstreckt. Betrachtet man sich nun die verkürzte Sequenz des mutierten *Tub*-Gens, bzw. die Sequenz der im Rahmen dieser Arbeit beschriebenen neuen 3´-Genvariante mit den Exons 13 und 14, so ist genau dieser integrale Helix-Bereich des Strukturmodells von den Veränderungen betroffen. Da die Sequenz ab AS 462 eine andere Aminosäuren-Zusammensetzung bekommt, dürfte es nicht mehr zur Ausbildung der letzten beiden Faltblätter 11 und 12 und der sich anschließenden integralen Helix H12 kommen. Zu welchen physiologischen Folgen diese Proteinverkürzung führen kann, kann zumindest im Fall der mutierten *Tub*-Variante am Phänotyp der Mausmutante skizziert werden. Histologische Studien zeigten, dass degenerative Erscheinungen die Ursache des komplexen Veränderungen in den Tubby-Mäusen sind und das apoptotische Vorgänge in den betroffenen Zellen, vor allem in Auge und Ohr, zum Verlust der Zellfunktion führen (Ohlemiller *et al.*, 1995). Für die charakteristische Fettleibigkeit der betroffenen Mäuse wird ein Absterben von Appetit-regulierenden neuronalen Zellen im Hypothalamus diskutiert, in Bereichen also, die die Nahrungsaufnahme regulieren (Nishina *et al.*, 1998).

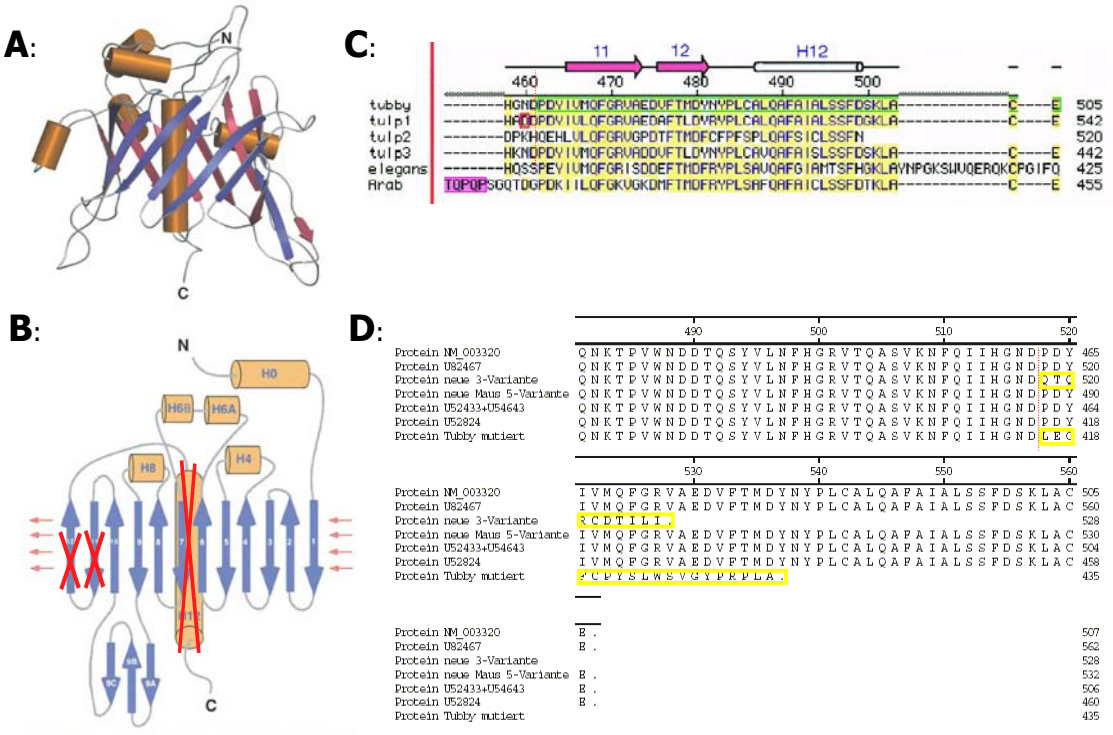


Abb. 34 Verkürzung der TUB-Proteinsequenz im COOH-Bereich und die daraus resultierende mögliche Veränderung der 3D-Proteinstruktur. A: Räumliche Darstellung des TUB-Proteins, bestehend aus sechs Alpha-Helices und zwölf Beta-Faltblättern, die sich zylindrisch in alternierender Ausrichtung um die axiale Alpha-Helix H12 gruppieren (nach Boggon *et al.*,1999). **B:** Dieselbe Proteinstruktur in einer zweidimensionalen Darstellung. Die roten Kreuze markieren die beiden Faltblätter 11 und 12, und die Helix H12, die bei den verkürzten Gensequenzen (unter D:) nicht mehr existieren. **C:** Dieser C-terminale Bereich, ist bei allen TULP-Genfamilienmitgliedern und den Homologen aus C.elegans (=elegans) und Arabidopsis (=Arab) weitestgehend konserviert erhalten. **D:** Bei der neuen humanen 3´-TUB-Variante und der mutierten Genform ist dieser terminale Bereich stark verkürzt und durch eine andere Aminosäuresequenz ausgetauscht. Die roten Pfeile verweisen auf die Exon-Intron-Grenze zwischen den Exons 11 und 12. Die Sequenzen Acc.-Nr. NM_003320, U82467 und „neue 3´-Variante" stellen humane Proteinsequenzen dar; die Sequenzen „neue 5´Variante, Acc.-Nr. U52433/U54643, U52824 und das „mutierte Tubby" zeigen die murine Aminosäuresequenz.

Der 3´-Bereich des *TUB*-Gens zeigte zum 5´-Ende im Interspeziesvergleich von Mensch und Maus zwar eine ähnlich hohe Konservierung von im Durchschnitt 91%, doch ändert sich dieses Verhältnis deutlich im Vergleich mit *TULP3*, dem im C-terminalen Bereich nächsten Verwandten des *TUB*-Gens. Mit Ausnahme von Exon 3 lässt sich in den übrigen Exonbereichen (Ex1-2 und Ex4-6) in der PIP-Blot-Analyse keine Konservierung mehr darstellen (Vgl. Abb. 28). Diese Beobachtung wird auch von Boggon & Mitarbeitern (1999) im Vergleich mit den Aminosäuresequenzen der anderen TUB-Genfamilienmitgliedern *TULP1* und *TULP2* geteilt. Boggon kennzeichnete den Abschnitt von AS 1 bis AS 242 durch eine geringe Tendenz zur Ausbildung von Sekundärstrukturen und als einen Abschnitt mit niedriger Komplexität in der Aminosäureabfolge. Ein Merkmal, das typisch zu sein scheint für die Transaktivierungsdomäne vieler Transkriptionsfaktoren (Triezenberg, 1995). Weiterführende Studien konnten die These belegen, dass der N-terminale Bereich des Tubby-Proteins eine Transkription

aktivierende Eigenschaft hat. Boggon & Mitarbeiter (1999) konnten mit Hilfe eines Fusionsproteins, bestehend aus den 242 Aminosäuren der NH_2-terminalen Tubby-Region und der DNA-bindenden Domäne von GAL4 zeigen, dass der N-terminale Tubby-Anteil die Transkription eines CAT-Reportergens („Chloramphenicol-acetyltransferase") um das über 20-fache steigerte. Interessanterweise zeigte das gleiche Experiment mit der alternativ gespleißten Tubby-Variante ohne das Exon 5 eine kaum messbare Steigerung der Transkriptionsrate. Es zeigte sich, dass die Funktion dieses alternativen Spleißens in der Modulation der Transkriptionsaktivierung liegt. Der Sequenzabschnitt, der durch das Exon 5 kodiert wird, wies zudem Sequenzelemente auf, die ähnlich denen sind, die in Glutamin-reichen Transkriptionsaktivatoren gefunden wurden, wie das Cyclische-Adenosin-3´,5´-Monophosphat-Response-Element, Sp1 oder Oct-2. Auch die Häufung von Serin und Threonin-Resten im gesamten NH_2-Region könnte als typisches Merkmal vieler Aktivierungsdomänen gewertet werden (Triezenberg, 1995). Das alternative Spleißen im 5´-Bereich mit unterschiedlichen Anfangsexons dürfte somit ein zusätzlicher Regulationsmechanismus für die Modulation der Transkriptionsaktivierung der potentiellen Zielgene darstellen.

Auffällig im Interspeziesvergleich der TUB-Proteinsequenzen war die 100%ige Konservierung der Exon 3-kodierten Aminosäuresequenz (siehe Abb. 35). Ein Abschnitt, der wie bereits oben erwähnt, selbst im Vergleich mit dem paralogen Familienmitglied *TULP3* aufgefallen war (Vgl. Abb. 28). Die Motivsuche für diesen konservierten Proteinbereich zeigte das Motivmuster für zwei nukleäre Lokalisationssignale (NLS) K^{39}KKR und P^{56}RSRRAR, die charakteristisch sind für im Zellkern funktionell aktive Gene (Garcia-Bustos *et al.*, 1991). Auch dieses „*In-silico*"-Ergebnis konnte experimentell durch Immunfluoreszenz-markierte Proben mit polyklonalen Antikörpern gegen den C-terminalen Tubby-Bereich und durch Protein-Immunoblots von subzellulären Kernfraktionen, die das Tubby-Protein hauptsächlich in der Zellkernfraktion nachwiesen, bestätigt werden (Boggon *et al.*, 1999).

Insgesamt zeichnet sich das Tubby-Protein durch seine zweiteilige Proteinfunktion mit DNA-bindender Domäne auf der C-terminalen Seite und der Transkriptions-modulierenden Domäne auf der N-terminalen Seite aus. Allerdings ist der Kontext zum sich ausprägenden Phänotyp mit Adipositas und neurodegenerativen Veränderungen der Tubby-Mäuse noch nicht geklärt. Vor allem sind weder die Zielgene des Tubby-Proteins noch der TULP-Proteine bekannt. Erst Studien von Santagata & Mitarbeitern (2001) geben einen Hinweis zu möglichen Interaktionspartnern der Tubby-Familie. Mit verschiedenen GFP-Tubby-Fusionsproteinen (GFP=„green-**f**luorescence-protein") konnte demonstriert werden, dass Tubby mit seinem C-terminalen Ende in der Plasmamembran durch Bindung an das Phosphatidyl-4,5-bisphosphat verankert ist (siehe Abb. 36).

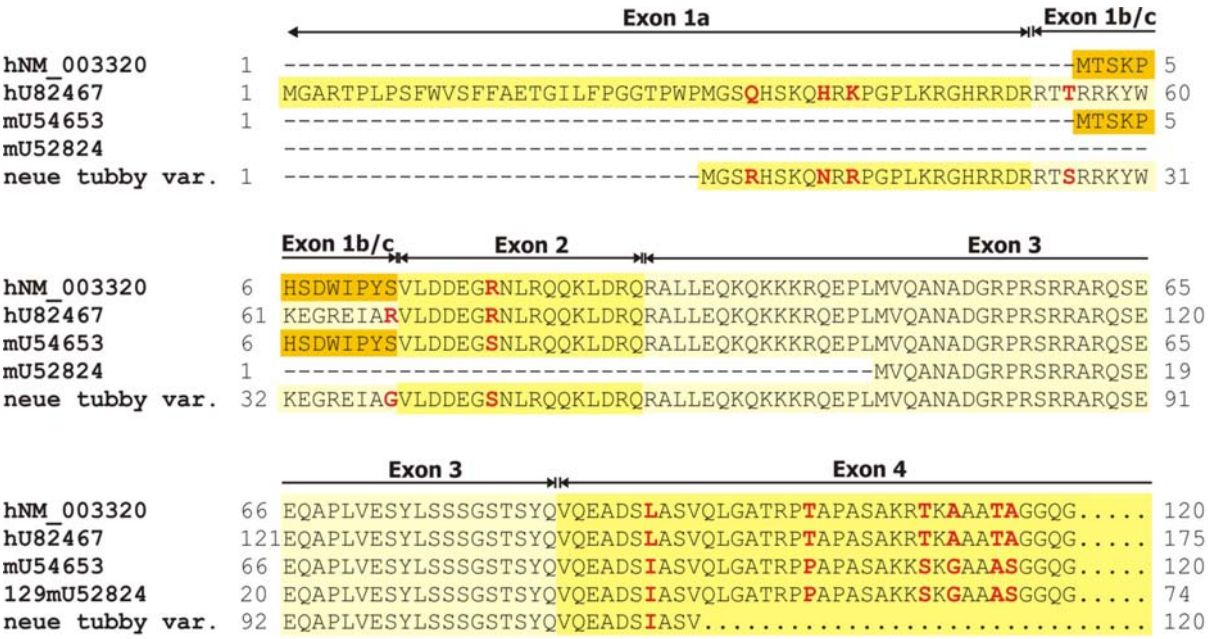


Abb. 35 **N-terminale Proteinsequenzen der verschiedenen TUB/Tub-Genvarianten.** Für den 5´-Bereich des TUB-Gens existieren drei verschiedene Proteinsequenzen. Kodiert werden diese alterierenden Bereiche von den ersten drei Exons 1a bis 1c. Erst ab Exon 2 ist die Sequenz bis auf die murine mU54653, deren ORF erst im Exon 3 beginnt, gleich. Die neue murine 5´-Genvariante beginnt erst am zweiten Startcodon, so dass insgesamt 29 Aminosäuren fehlen. Im Vergleich zur humanen Genvariante weist die murine Peptidsequenz an vier Stellen alternative Aminosäuren auf, die in der Grafik rot hervorgehoben wurden. Es fällt auf, dass insbesondere der kodierende Abschnitt des Exons 3 in beiden Spezies zu 100% konserviert geblieben ist. Der orange hervorgehoben Bereich der Genvarianten hNM_003320 und mU54653 wird von Exon 1c kodiert. Der erste Buchstabe vor der Acc.-Nr: gibt die jeweilige Spezies an: h=human, m=murin.

Es scheint als nachfolgender Effektor der G-Protein-gekoppelten Rezeptoren (GPCR) zu fungieren, die Signale auf das G_q, einer Subklasse des Gα-Proteins, weiterleiten. Gα_q setzt durch die Phospholipase C-β-vermittelte Hydrolyse des Phosphatidyl-4,5-bisphosphats das Tubby-Protein frei, was zu einer Translokation in den Nukleus führt. Dieser Mechanismus kann durch Aktivierung der Gα_q-gekoppelten Rezeptoren wie den Serotonin-Rezeptors-$5HT_{2C}$ induziert werden. Tubby scheint so die direkte Verbindung zwischen GPCR und der Regulation der Genexpression bestimmter Zielgene darzustellen. GFP-Tubby-Fusionsproteine, die nur aus der N-terminalen Domäne bestanden, zeigten dagegen die eindeutige Korrelation zum Nukleus als Aufenthaltsort. Diese Lokalisierung kann durch den Nachweis der bereits erwähnten NLS, kodiert in Exon 3 der Tubby-Domäne, in Einklang gebracht werden (Santagata *et al.*, 2001); einem Proteinsequenzbereich, der von allen aufgezeigten alternativen 5´-Spleißvarianten kodiert wird. Somit besitzt das Tubby-Protein mit der dualen Lokalisierung in Plasmamembran und Zellkern eine Eigenschaft, die auch bei verschiedenen anderen Transkriptionsfaktoren wie SREBP (Baeuerle *et al.*, 1996), NF-κB (Brown *et al.*, 2000), SMADS (Horvath *et al.*, 1997), STATs (Crabtree *et al.*, 1994) und N-FAT (Kretzschmar *et al.*, 2000) bekannt geworden ist.

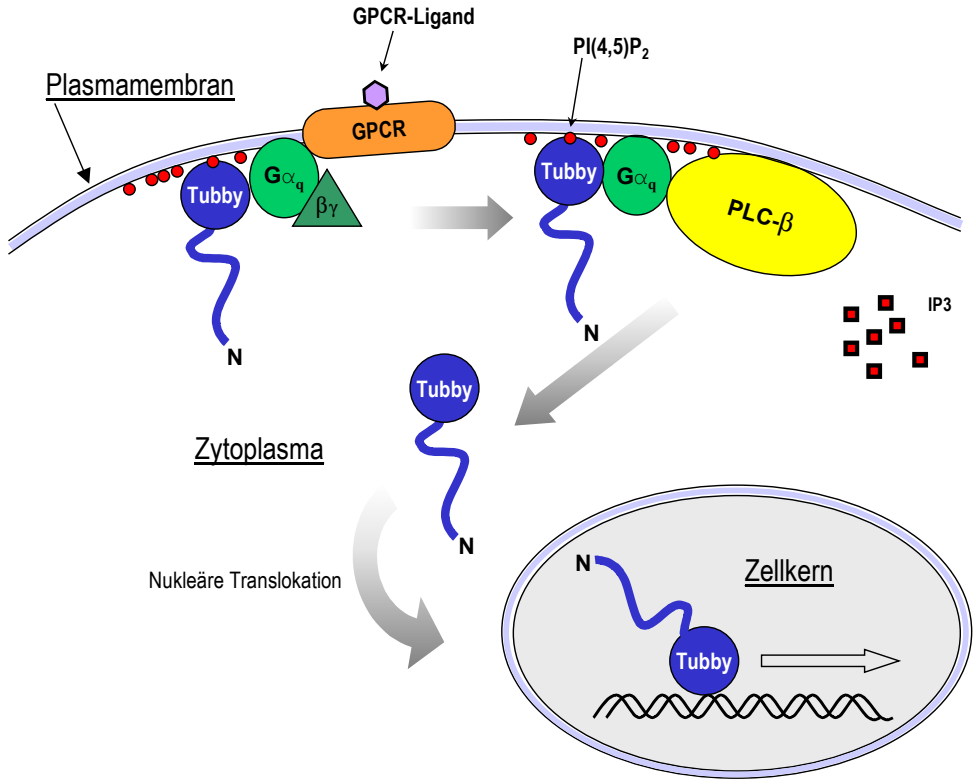


Abb. 36 Schematisches Modell der möglichen Tubby-Protein-vermittelten Signaltransduktion von Gαq. Die Aktivierung von Gαq zu Gαq* durch Interaktion mit dem G-Protein-gekoppelten Rezeptor (GPCR)-Liganten-Komplex führt zu einer Aktivierung der Phospholipase C-β (PLC-β) mit anschließender Hydrolyse der phosphorylierten Inositol-Lipiden (IP3) in der unmittelbaren Umgebung des Komplexes. Dies führt zur Dissoziation des Tubby-Proteins von der Plasmamembran ins Cytosol. Tubby transloziert daraufhin, geführt durch die „Nuclear-Localization-Seuenzen" (NLS) der N-terminalen Domäne, in den Nukleus und interagiert dort mit Zielsequenzen auf der DNA. Tubby scheint so die direkte Verbindung zwischen GPCR und der Regulation der Genexpression darzustellen. Inwieweit andere nukleäre Faktoren für diese Interaktion des Tubby-Proteins eine Rolle spielen ist noch ungeklärt.

Ein Erklärungsversuch für die sich ausbildende Fettsucht der mutationstragenden Mäuse und der Beteiligung des Tubby-Gens wurde durch Koritschoner & Mitarbeiter (2001) unternommen. Sie stellten einen Zusammenhang zwischen dem Tyroidhormon-Rezeptor (TR) und der *TUB*-Genexpression fest. Mit Hilfe von „Diffential-Display"-Analysen auf der Suche nach T3 (Trijodthreonin)- regulierten Genen konnte *TUB* als potentielles Zielgen des TR identifizierte werden. Es konnte gezeigt werden, dass *TUB* eines der wenigen Gene ist, die unter Thyroidhormon-Kontrolle im adulten Gehirn stehen und nicht wie die meisten beschriebenen Target-Gene nur in der frühen postnatalen Entwicklungsperiode von Thyroidhormonen gesteuert werden. Dies gibt nun eine neue Argumentationsgrundlage für die Ursache der beobachteten Adipositas, da T3 nachweislich an der Regulation des Energiehaushaltes im

Organismus beteiligt ist. Weitere Indizien für diese Annahme wären die Gegebenheiten, dass TR und *Tub* in verschiedenen Regionen des Gehirns koexprimiert werden (Mellström *et al.*, 1991; Kleyn *et al.*, 1996), dass $TR^{-/-}$ und $Tub^{-/-}$ Maus-Mutanten einen sich überlappenden Phänotyp aufweisen - so zeigen auch $TR\beta^{-/-}$ Mäuse sensorische Defekte im Hören und Sehen (Forrest *et al.*, 1996), und dass Defekte im T3-Metabolismus sich ihrerseits oft in den betroffenen Mäusen in Fettleibigkeit äußern. Diese Beobachtungen unterstützen die These, dass der Thyroidhormon-Rezeptor und *Tub* miteinander interagieren oder Teil desselben biochemischen Synthesewegs sind.

Alle aufgeführten Ergebnisse stammen aus funktionellen Studien am Mausmodell und als einzige murine genomische Sequenzveränderung ist die beschriebene Mutation am Übergang von Exon 11 zum Intron 11 bekannt. Für vergleichbare Syndrome beim Menschen konnte allerdings weder diese noch eine andere genomischen Veränderungen beschrieben und als Auslöser für die recht ähnlichen Krankheitsbilder verantwortlich gemacht werden. Sämtliche Patientenkollektive mit komplexen Erkrankungen wie Usher Typ 1c (Heckenlively *et al.*, 1995), Alstrom (Alstrom *et al.*, 1953) und Bardet-Biedl (Leppert *et al.*, 1994), die sowohl durch eine charakteristische Fettleibigkeit, wie auch durch Degenerationserscheinungen ihrer sensorischen Zellen in Auge und Ohr auffallen, besitzen keine bekannten genetischen Veränderungen in ihrem *TUB*-Gen. Zumindest nicht in den bereits publizierten Spleißvarianten des *TUB*-Gens. Auch wenn die murine Mutation nicht relevant für die humanen Krankheitsbilder zu sein scheint, so könnten andere noch unbekannte genetische Veränderung, z. B. in den neuen in dieser Arbeit erstmals beschriebenen Spleißvarianten, mit der Ätiologie dieses Phänotypus zusammenhängen. Mit den vorliegenden Resultaten ist nun die Grundlage für weitere gegebenenfalls diagnostisch relevante Sequenzbereiche geschaffen, um etwaige Mutationen in den neuen Genbereichen durch Homologievergleiche aufzuspüren.

4.2.2.3 Die TUB-Genfamilie

Die Tubby-Familie scheint eine alte hoch konservierte Genfamilie zu sein, da entsprechende Homologe in fast allen in dieser Arbeit untersuchten Spezies gefunden wurden. Allein die BlastX-Analyse des genomischen *TUB*-Genbereiches brachte Homologien zu Proteinsequenzen von mehr als 14 unterschiedlicher Spezies zutage. Darunter befanden sich nicht nur die Vertebraten wie Maus, Ratte, Huhn, Pferd, *Xenopus*, *Danio* (Zebrafisch) und *Oncorhynchus* (Forelle), sondern auch Arthropoden wie *Drosophila* und Nematoden wie *Caenorhabditis elegans.* Ebenso konnten Homologien zu Sequenzen aus Vertretern des Pflanzenreichs wie Reis, Mais, *Arabidopsis*, *Cicer arietinum* (Kichererbse) und *Lemna paucicostata* (Wasserlinse; dies war die einzige annotierte Datenbanksequenz dieser Spezies überhaupt!) angesprochen werden. Eine phylogenetische Verwandtschaftsbeziehung wurde in Abb. 37 mit allen Spezies vorgenommen, deren vollständige TUB-Proteinsequenz den Datenbanken zu entnehmen waren. Alle weiteren oben erwähnten Arten wiesen nur unvollständige Sequenzabschnitte auf. Es zeigte sich die hohe verwandtschaftliche Beziehung innerhalb der Mammalier Maus, Ratte und Mensch und mit etwas Abstand zum Nematoden *C. elegans*. Die TUB-Proteinsequenzen der vier

pflanzlichen Vertreter *Oryza sativa* (Reis), *Arabidopsis thaliana*, *Lemna paucicostata* (Wasserlinse) und *Cicer arietinum* (Kichererbse) bilden dagegen eine eigene Linie. Auch der Unterschied in der botanischen Systematik zwischen Monocotyledoneae und Dikotyledoneae zeigte sich in dem größeren verwandtschaftlichen Abstand zwischen Reis (Einkeimblättrige Art) und den drei Spezies *Arabidopsis*, *Lemna* und *Cicer* (Zweikeimblättrige Arten). Die TUB-Proteinsequenz des Arthropoden *Drosophila melanogaster* und des Huhns (*Gallus gallus*) wiesen interessanterweise die größten Sequenzunterschiede auf und spiegeln nicht mehr wie oben die systematischen Verwandtschaftsverhältnisse wider. Homologien zu einzelligen Organismen konnten in diesem Zusammenhang nicht gezeigt werden.

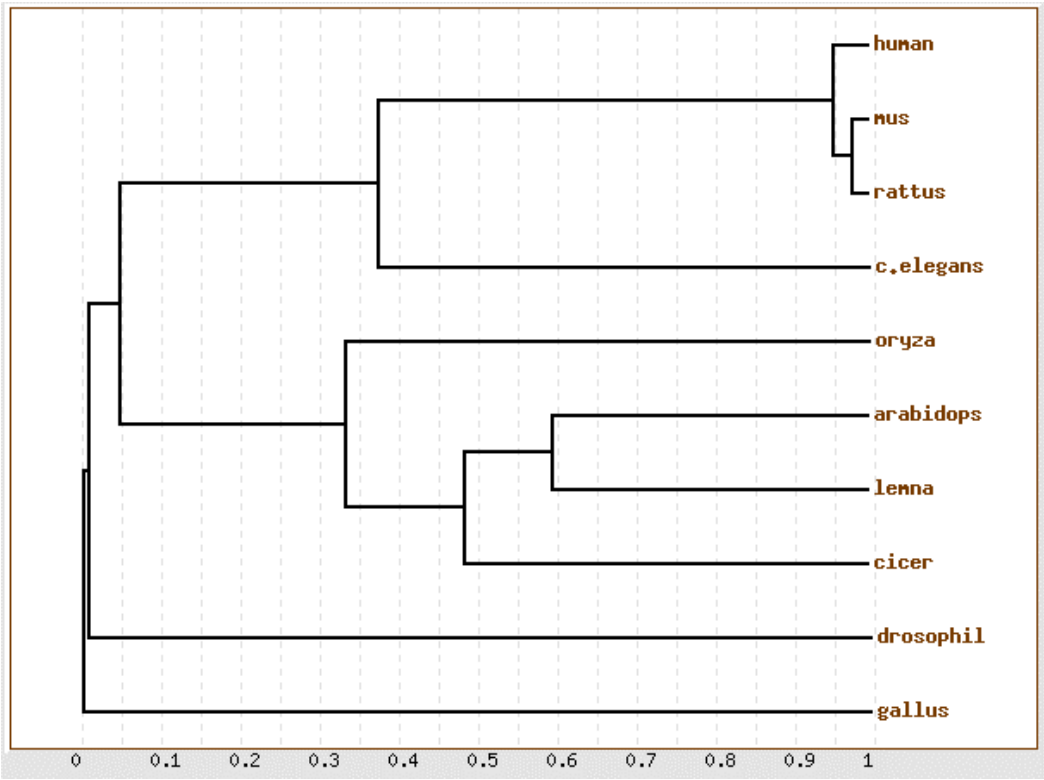


Abb. 37 **Phylogenetischer Stammbaum der TUB-Proteinsequenzen** (erstellt mit dem Programm TreeTop <www.genebee.msu.su/genebee.html>). Die TUB-Sequenzen von Maus (Acc. Nr. P50586) und Ratte (Acc. Nr. NP_037209) zeigen sehr hohe Homologie zur Proteinvariante des Menschen (Acc. Nr. P50607). Die Ähnlichkeit zum Nematoden C.elegans (Acc. Nr. CAB61010) fällt hingegen deutlich ab. Die TUB-Proteine der pflanzlichen Spezies Oryza sativa (Reis) (Acc. Nr. NP_916202), Arabidopsis thaliana (Acc. Nr. NP_849894), Lemna paucicostata (Wasserlinse) (Acc. Nr. BAA82866) und Cicer arietinum (Kichererbse) (Acc. Nr. CAB88665) bilden dagegen eine eigene, in sich divergentere Linie, wobei der Reis als einziger Vertreter der Monocotyledoneae in dieser Zuordnung die meisten Unterschiede aufweist. Die TUB-Proteinsequenzen von Drosophila (Acc. Nr. NP_730825) und dem Huhn (gallus) (Acc. Nr. NP_990776) weisen sie größten Unterschiede auf und sind nicht mehr in Einklang mit der systematischen Zuordnung zu bringen.

Die starke Konservierung des TUB-Proteins und seiner Gensequenz zeigte sich nicht nur im Interspeziesvergleich, sondern auch innerhalb der Genfamilie der „Tubby-ähnlichen"-Gene TULP (TULP= **TU**B-**L**ike-**P**rotein). Diese Ähnlichkeit bezieht sich vor allem auf die drei Gene *TULP1*, *TULP2* und *TULP3*. Mit 74% identischen Aminosäuren zeigt das *TULP1* die größte Ähnlichkeit zum *TUB*-Gen

(siehe Abb. 38A). Die 2.116 bp umfassende und für 542 AS kodierende *TULP1*-Gensequenz konnte dabei auf Chromosom 6p21.3 kartiert werden (North *et al.*, 1997) und wird exklusiv in retinalem Gewebe exprimiert. Mit ebenfalls einer hohen Homologie von 68% konnte – wie bereits das Beispiel *LMO3* zeigte – auch auf dem zu Chromosom 11 paralogen Chromosom 12 ein TULP-Gen identifiziert werden (Nishina *et al.*, 1998). Das humane *TULP3* kartiert in die Chromosomenregion 12p13 und kodiert mit einem offenen Leserahmen für 442 Aminosäuren. Es hat damit die kürzeste Proteinsequenz unter den TULP-Genfamilienmitgliedern. Das vierte Familienmitglied *TULP2* besteht aus einer 1.733 bp langen cDNA, kodierte für 520 AS und konnte über „Radiation Hybrid Mapping" der Chromosomenregion 19q13.1 zugeordnet werden (North *et al.*, 1997). Allen Familienmitgliedern ist ein hochkonservierter C-Terminus über 250 Aminosäuren gemein, dem für das *TUB*-Gen eine DNA-bindende Aktivität nachgewiesen werden konnte. Ebenfalls allen vier TUB/TULP-Genen gemeinsam ist bei Mensch und Maus der Expressionsort in der Retina des Auge. Da *TULP1* ausschließlich in der Retina nachgewiesen werde konnte, wird es als Kandidatengen für die Erkrankung *Retinitis pigmentosa* angesehen, einem degenerativen Prozess der Augennetzhautgefäße, der bis zur Erblindung führen kann (Banerjee *et al.*, 1998). Interessanterweise ist die genetische Ursache im *TULP1*-Gen, ähnlich wie beim *TUB*-Phänotyp, auch eine Spleißmutation. Dies legt die Vermutung nahe, dass die Störung im Spleißen, wie im Phänotyp von *TUB,* degenerative Erscheinungen für die genexprimierenden Zellen zur Folge hat. Ob dieser mutationsbedingte Phänotyp, der sich im Auge als apoptotischer Prozess von Photorezeptorzellen äußert, typisch für alle TUB/TULP-Gene ist, versuchten Ikeda & Mitarbeiter (1999) zu klären. Die Untersuchungen zeigten allerdings, dass jedes der vier Gene ein eigenes zellspezifisches Expressionsmuster im Auge aufweist und dass die apoptotischen Prozesse bei *Tub* und *Tulp1* in der Maus jeweils nur als Ergebnis unterschiedlicher Degenerationserscheinungen zu werten sind. Interessant ist in diesem Zusammenhang auch, dass für das Gen *TULP3*/*Tulp3*, d. h. für den chromosomalen Locus 12p13, bzw. der Telomer-Region des Chromosoms 6 der Maus derzeit keine Erkrankungen bekannt sind. Lediglich Ikeda & Mitarbeiter (2001) konnten mit Hilfe von *Tulp3*-knockout-Mäusen zeigen, dass es während der Embryogenese der Mäuse zu Neuralrohrdefekten kommt, die durch neuroepitheliale Apoptose hervorgerufen werden und dass *Tulp3* essentiell für die Embryonalentwicklung ist. Somit scheinen alle TUB/TULP-Familienmitglieder mit ihren Genfunktionen an ganz zentralen Schnittstellen im Organismus zu stehen. Dies würde dann auch das ubiquitäre Vorkommen dieser Gene bei fast allen untersuchten Organismen erklären.

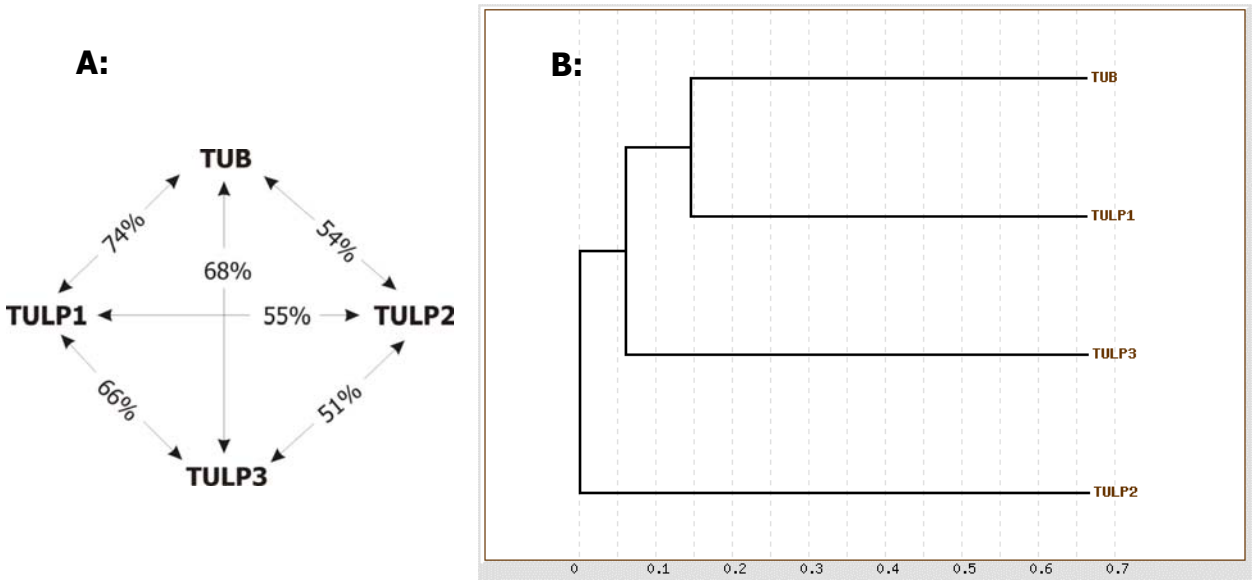


Abb. 38 **Aminosäuresequenzvergleich der vier humanen TUB-Gene. A:** Das Diagramm zeigt den Prozentanteil der identischen Aminosäuren der drei TULP-Gene 1 bis 3 mit dem TUB-Protein. Als Referenz dienten die Proteinsequenzen Acc.Nr. NM_003320 (*TUB*), AAH32714 (*TULP1*), NP_003314 (*TULP2*) und NP_003315 (*TULP3*). **B:** Die vergleichende Analyse der humanen und murinen Aminosäuresequenzen führte zur dargestellten genealogischen Beziehungen. Während *TUB* und *TULP1* am engsten miteinander verwandt zu sein scheinen, ist *TULP2* am weitesten von den anderen TULP-Genen entfernt.

4.2.3 Der eukaryontische Translationsinitiationsfaktor 3

Am proximalen Ende der in der vorliegenden Arbeit sequenzierten genomischen Maussequenz konnten über die Homologie zum humanen Gen *EIF3-p47* (Acc.-Nr. U94855; Asano *et al*, 1997) acht unbekannte Exons identifiziert werden und zu einem neuen murinen Transkript zusammengeführt werden. Es handelt sich dabei um die 47 kDa Untereinheit des eukaryontischen Translationsinitiations-faktors 3 (eIF3). Als Multiproteinkomplex besteht dieser aus verschiedenen löslichen Proteinen und findet seine Funktion bei der Initiation der eukaryontischen Proteinbiosynthese.

Mit ca. 600 kDa stellt *EIF3* unter den eukaryontischen Initiationsfaktoren (eIF´s) das Protein mit dem größten Molekulargewicht dar. *EIF3* bindet ohne die Anwesenheit von anderen Initiationsfaktoren an die 40 S ribosomale Untereinheit und bewirkt, dass die 40 S und die 60 S ribosomalen Untereinheiten im dissoziierten Zustand verbleiben. Ebenso spielt er eine wichtige Rolle bei der Bildung des 40 S Initiationskomplexes mit dem ternären Komplex aus eIF2·GTP·Met-tRNA$_i$ und bei der Bindung dieser beiden an die mRNA (Benne *et al*., 1978). *EIF3* interagiert spezifisch mit *EIF4G*, der größten Untereinheit des mRNA-„Cap-binding"-Proteinkomplexes *eIF4*, und mit dem Protein *EIF4B*, welches die RNA-Bindung herstellt (Méthot *et al*., 1996). Der *eIF3*-Komplex besteht seinerseits aus mindestens zehn nichtidentischen Untereinheiten, die nach ihrer Masse in der SDS-PAGE mit p170, p116, p110, p66, p48, p47, p44, p40, p36 und p35 benannt worden sind (Hershey *et al*., 1996; Johnson *et al*., 1997; Méthot *et al*., 1997; Asano *et al*., 1997; Block *et al*., 1998). Darüber hinaus zeigt *EIF3* bei EST-

Homologiesuchen zu verschiedenen Klonen aus Hefe, *C. elegans,* Maus und weiteren Spezies Homologien. Allein für die neu beschriebene murine Untereinheit p47 konnten bei der Homologiesuche in den EST-Datenbanken über 550 Treffer aus den verschiedensten Geweben verzeichnet werden. Somit dürfte es sich um ein im Organismus ubiquitär exprimiertes Gen handeln, was aufgrund der ihm zugeschriebenen Funktion der Translationsinitiation auch verständlich wäre.

Die hohe Interspezieskonservierung auf Proteinebene zwischen Mensch und Maus beschränkt sich mit 90,9% nicht nur auf die Untereinheit p47 (siehe Abb. 36,A). Wie bei ORFs anderer homologer Gene zeigen auch die anderen Untereinheiten des eIF3s eine starke evolutive Konservierung. So beträgt die Proteinsequenzübereinstimmung für die RNA-bindende Untereinheit *eIF3-p44* zwischen Mensch und Maus 97% und zwischen dem humanen *eIF3-p44* und dem *eIF3-p33* von *S. cerevisiae*, dem homologen Protein der Hefe, sind es noch 33% Übereinstimmung und 42% Ähnlichkeit (Block *et al.*, 1998). Das zum murinen *Eif3-p47* putativ homologe Gen "D2013.7" aus *C. elegans* (Asano *et al.*, 1998) zeigt noch eine Übereinstimmung von 35,5% der Aminosäuren (siehe Abb. 38, B).

A: **Murines eif3-p47 vs. humanes EIF3-p47**
90.9% Identität auf 363 Residuen Überlappung; Score: 1639.0; Häufigkeit der Lücken: 1.1%

```
Maus            1 MASPAVPANVPPATAAAAPAPVVTAAPASAPTPSTPAPTPAATPAASPAPVSSDPAVAAP
Mensch          1 MATPAVPASAPPATPAPVPAAVPASAPASVPAP-TPAPAAAPVPAAAPAS-SSDPAAASA
                  ** *****  **** *  ** *   **** * * ****  *  *** **  ***** *
                  MATPAVPVSAPPATPTPVPAAAPAS----VPAP-TPAPAAAPVPAAAPAS-SSDPAAAAA

Maus           61 --AAPGQTPASAPAPAQTPAPSQPGPALPGPFPGGRVVRLHPVILASIVDSYERRNEGAA
Mensch         59 TTAAPGQTPASAQAPAQTPAPALPGPALPGPFPGGRVVRLHPVILASIVDSYERRNEGAA
                    ********** ********  ************************************
                  ATAAPGQTPASAQAPAQTPAPALPGPALPGPFPGGRVVRLHPVILASIVDSYERRNEGAA

Maus           119 RVIGTLLGTVDKHSVEVTNCFSVPHNESEDEVAVDMEFAKNMYELHKKVSPNELILGWYA
Mensch         119 RVIGTLLGTVDKHSVEVTNCFSVPHNESEDEVAVDMEFAKNMYELHKKVSPNELILGWYA
                   ************************************************************
                  RVIGTLLGTVDKHSVEVTNCFSVPHNESEDEVAVDMEFAKNMYELHKKVSPNELILGWYA

Maus           179 TGHDITEHSVLIHEYYSREAPNPIHLTVDTGLQHGRMSIKAYVSTLMGVPGRTMGVMFTP
Mensch         179 TGHDITEHSVLIHEYYSREAPNPIHLTVDTSLQNGRMSIKAYVSTLMGVPGRTMGVMFTP
                   ***************************** ** *************************
                  TGHDITEHSVLIHEYYSREAPNPIHLTVDTSLQNGRMSIKAYVSTLMGVPGRTMGVMFTP

Maus           239 LTVKYAYYDTERIGVDLIMKTCFSPNRVIGLSSDLQQVGGASARIQDALSTVLQYAEDVL
Mensch         239 LTVKYAYYDTERIGVELIMKTCFSPNRVIGLLSDLQQVGGASARIQDALSTVLQYAEDVL
                   ************** *************** ****************************
                  LTVKYAYYDTERIGVDLIMKTCFSPNRVIGLSSDLQQVGGASARIQDALSTVLQYAEDVL

Maus           299 SGKVSADNTVGRFLMSLVNQVPKIVPDDFETMLNSNINDLLMVTYLANLTQSQIALNEKL
Mensch         299 SGKVSADNTVGRFLMSLVNRVPKIVPDDFETMLNSNINDLLMVTYLANLTQSQIALNEKL
                   ****************** ****************************************
                  SGKVSADNTVGRFLMSLVNQVPKIVPDDFETMLNSNINDLLMVTYLANLTQSQIALNEKL

Maus           359 VNL
Mensch         359 VNL
                   ***
                  VNL
```

B: **Murines eif3-p47 vs. C.elegans D2013.7**

35.5% Identität auf 287 Residuen Überlappung; Score: 371.0; Häufigkeit der Lücken: 7.7%

```
Maus,          96 VRLHPVILASIVDSYERR------NEGAARVIGTLLGTVDKHSVEVTNCFSVPHNESEDE
C_elegans,      7 VNVHPGVYMNVVDTHMRRTKSSAKNTGQEKCMGTLMGYYEKGSIQVTNCFAIPFNESNDD
                  *  **      **   **      * *     *** *   * *  *****  * *** *

Maus,         150 VAVDMEFAKNMYELHKKVSPNELILGWYATGHDITEHSVLIHEYY---------SREAPN
C_elegans,     67 LEIDDQFNQQMISALKKTSPNEQPVGWFLTTSDITSSCLIYHDYYVRVITEASARRESPP
                     *  *   *    ** ****   **  *  ***      * **          ** *

Maus,         201 PIHLTVDT---GLQHGRMSIKAYVSTLMGVPGRTMG--VMFTPLTVKYAYYDTERIGVDL
C_elegans,    127 IVVLTIDTTFSGDMSKRMPVRAYLRSKAGIPGAAGPHCAIFNPLRVELAAFPGELVAMQL
                     ** **   *    **   **     * **        * ** *  *    *     *

Maus,         256 IMKTCFSPNRVIGLSSDLQQVGGASARIQDALSTVLQYAEDV-LSGKVSADNTVGRFLMS
C_elegans,    187 IEKALDSRRREATLESGLEQLETSTAQMIEWLERMLHYVEDVNKNGEKPGDAQIGRQLMD
                  * *   *  *   * * * *     *     *   * * ***   *    *   ** **

Maus,         315 LVNQVPK-IVPDDFETMLNSNINDLLMVTYLANLTQSQIALNEKLVN
C_elegans,    247 IVTASSNNMQPEKLDTLVKNTLRDYVMVSYLAKLTQTQLQVHERLVS
                   *        *    *       *  ** *** *** *    * **
```

Abb. 39 **A: Vergleich der Proteinsequenzen des neuen murinen Gens Eif3-p47 mit dem putativen humanen Homolog EIF3-p47** (Acc.-Nr. U94855). Insgesamt beträgt die Konservierung mehr als 90% über die gesamte Proteinsequenz. Insbesondere der C-terminale Bereich zeigte die wenigsten Basenaustausche **B:** Proteinsequenzvergleich des neuen murinen Gens Eif3-p47 mit dem putativ homologen Gen D2013.7 von C. elegans (Acc.-Nr. Z47808). Im Vergleich zur Maus ist die Proteinsequenz des Nematoden im Bereich von AS 24 bis AS 165 um 20 Aminosäuren länger.

Des weiteren konnten für das Gen *Eif3-p47* auch Homologien zur eIF3-Untereinheit p40 festgestellt werden, höchstwahrscheinlich ein Zeichen dafür, dass diese Gene phylogenetisch einen gemeinsamen Ursprung haben könnten, was auch durch die Sequenzübereinstimmung von knapp 30% über die gesamte Proteinsequenz unterstützt wird. Eine ähnliche Homologie konnte zu keinem der anderen Untereinheiten des eIF3 festgestellt werden. Insbesondere der N-terminale Proteinbereich ab der Aminosäure 96 scheint besonders in sechs Teilbereichen konstant geblieben zu sein (Asano *et al*, 1997). Innerhalb dieser sechs Abschnitte waren 38,9% der Aminosäuren identisch (Vgl. Abb. 37) und weisen Homologien zum Motiv der Proteinfamilie Mov-34 auf. Diese Proteinfamilie setzt sich aus insgesamt 19 eukaryontischen Genprodukten verschiedener Spezies (z. B. *Drosophila melanogaster*, C. elegans, *Arabidopsis thaliana*, *S. cerevisiae*) zusammen und fungiert als Regulatoren von Transkriptionsfaktoren (Asano *et al*, 1997). Das humane Homolog zum namensgebenden murinen *Mov-34* zeigt seinerseits 96% Homologie zur S12 Untereinheit der 26S Proteosomen der Erythrozyten (Dubiel *et al*., 1995). Somit besitzen alle fünf hier untersuchten Gene – *eif3-p47*, *EIF3-p47*, *EIF3-p40*, *mov-34* und *26S Untereinheit S12* – Homologien zu dieser konservierten Mov-34-Domäne. Des weiteren zeigte dieser Sequenzbereich auch Homologien zur Untereinheit (H) des Transkriptionsfaktor II (TFIIH), einer von sieben Hilfsfaktoren der RNA-Polymerase II am Transkriptionsinitiationsprozess (Aravind & Ponting, 1998). Es scheint sich daher um universelle Motivstrukturen zu handeln, die wahrscheinlich für regulative Prozesse während der Transkription essenziell sind. Das Mov-34-Motiv, dessen genaue Funktion noch nicht aufgeklärt ist, zeigte darüber hinaus auch Ähnlichkeit mit der JAB1/MPN-Domäne (Jun activation domain-binding protein 1)/(Mpr1, Pad1 N-terminal), ein Motiv des

Bindeproteins 1 der Aktivierungsdomäne der Jun-Kinase, die in die Regulation der Transkription involviert ist. Da beide Motive fast vollständig zueinander überlappen (siehe Abb. 40), könnte es sich auch um ein und die selbe Domänenfamilie oder zumindest um nahe Verwandte handeln, die ähnliche Funktionen als transkriptionelle Regulatoren in der Zelle wahrnehmen. Interessant für das funktionelle Verständnis ist die koregulatorische Wirkung dieser Gene und der Zusammenhang, dass das *JAB1* als Koaktivator im überexprimierten Zustand bereits mit der Entstehung von Ovarialkrebs (Sui *et al.*, 2001) und Neuroblastomen (Shen *et al.*, 2000) diskutiert wurde. Dieser Zusammenhang könnte als Hinweis gewertet werden, dass auch das Gen *EIF3-p47*/*eif3-p47* möglicherweise bei der Entstehung von tumorassoziierten Krankheitsbildern mit beteiligt ist. Somit könnte auch das Gen *EIF3-p47* – neben *LMO1* und *TUB* – ein weiterer Kandidat für ein tumorassoziiertes Gen in der untersuchten chromosomalen Region sein.

A:

```
                        I              II          III          IV           V          VI

Maus eif3p47   VRLHPVILASIVDS 10 VIGTLLG 8 VTNCFSVPH 29 ILGWYATGHD 22 PIHLTVDT 10 KAYVST
Human eIF3p47  VRLHPVILASIVDS 10 VIyTLLG 8 VTNCFSVPH 29 ILGWYATGHD 22 PIHLTVDT 10 KAYVST
Human eIF3p40  VQIDGLVVLKIIKH 10 VQyVLLG 8 ITNCFPFPQ 29 HVGWYQS.TY 22 SVVLIYDP  9 KAYRLT
Maus mov-34    VVVHPLVLLSVVDH 12 VVGVLLG 8 VSNSFAVPF 32 IVGWYHTGPK 18 SVLVIIDV  9 EAYISV
Human 26S S12  VVVHPLVLLSVVDH 12 VVGVLLG 8 VSNSFAVPF 32 IVGWYHTGPK 18 SVLVIIDV  9 EAYISV
```

B:

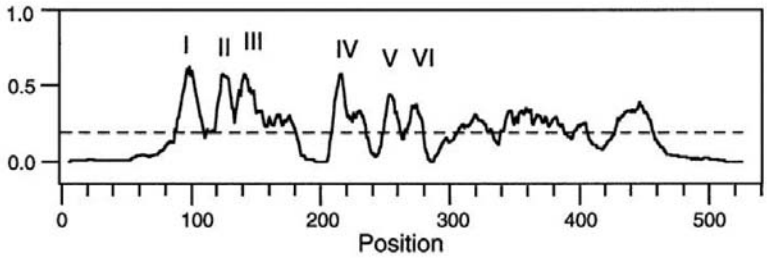


Abb. 40 **Direkter Vergleich der konservierten Proteinabschnitte** des N-terminalen Bereichs der fünf Gene eif3-p47, eIF3p47, eIF3p40, mov-34 und 26S-S12. **A:** Allen Proteinen gemeinsam ist die Konservierung der Aminosäuren innerhalb der sechs Abschnitte I bis IV, die typisch sind für alle mov-34 Familienmitglieder. Gelb unterlegt wurden die Aminosäuren, die sich in allen Sequenzbeispielen wiederfinden. Gleiche Aminosäuren zwischen eIF3-p47 und eIF3-p40 wurden rot hervorgehoben. Die Ziffern zwischen den römisch nummerierten Abschnitten geben die ausgelassenen Aminosäuren an. **B:** Die Konservierung der sechs Bereichsabschnitte I bis IV zeichnet sich deutlich als Spitzen im Homologie-Plot über den gesamte Proteinsequenzbereich ab (nach Asano *et al.*, 1997).

Eif3-p47-Protein

JAB/MPN-Motiv

```
                Länge der konservierten Domäne = 135 Residuen,  98.5% angeordnet
                Score = 98.5 bits (245), Wahrscheinlichkeit = 2e-21
JAB/MPN    95   VVRLHPVILASIVDSYERRNEGAARVIGTLLGTVDKHSVEVTNCFSVP-HNESEDEVAVD  153
eif3-p47    1   QVKVHPLVPLKILKHAER--TGPEEVCGVLLGKSNKDSPRVTECFAVPNEPQDDVVQEYP   58

JAB/MPN   154   MEFAKNMYELHKKVSPNELILGWYATG----HDITEHSVLIHEYYSREAPNPIHLTVDTG   209
eif3-p47   59   EDYSHLMDEELKATEKDLEIVGWYHSHPDESPWPSEVDVATHESYQAPWPISVVLGVDPI   118

JAB/MPN   210   LQ-HGRMSIKAYVST   223
eif3-p47  119   RSFSGRLSLRAFRLT   133
```

Mov-34-Motiv

```
                Länge der konservierten Domäne = 103 Residuen, 100.0% angeordnet
                Score = 78.9 bits (194), Wahrscheinlichkeit = 2e-15
eif3-p47   97   RLHPVILASIVDSYERRNEGAARVIGTLLGTVDKHSV-EVTNCFSVPHNESEDEVAVDME  155
Mov34       1   KIHPLVLLKILDHARRGGPSAEEVMGLLLGKVEGDVVIEVTNVFALPQSESSDDVDAVDL   60

eif3-p47  156   FAKNMYELHKKVSPNELILGWYATGHDI----TEHSVLIHEYYSREAPN   200
Mov34      61   DQEYMM------SMLEEVVGWYHSHPGPGCWLSEVDVHTQFLYQRYHPE   103
```

Abb. 41 Die beiden konservierten Domänen JAB/MPN und Mov-34 in der Aminosäuresequenz des Gens *eif3-p47*. Die Aminosäuren 95 bis 223 weisen Homologie zum JAB/MPN-Domäne auf, die typisch ist für das Jun-Kinase-Aktivierungsdomäne-Bindeprotein und für proteosomale Untereinheiten. Ebenso zeigte dieser Bereich in etwas verkürztem Sequenzumfang Homologie zum Motiv der Mov-34-Familie, die charakteristisch für die eIF3-Untereinheiten und für Regulatoren der Transkriptionsfaktoren ist. Grau unterlegt wurden die Motivabschnitte I bis IV aus vorheriger Abb. 37. Der Motivvergleich zeigte, dass die Konservierung zu *JAB/MPN* und *Mov-34* sich über diese sechs Bereichsabschnitte nach Asano *et al.* (1997) hinweg erstreckt. Rot wurden alle identischen, blau alle ähnlichen und schwarz alle anderen Aminosäuren ohne Übereinstimmung markiert.

4.2.3.1 EIF3-47-Pseudogene

Die in der vorliegenden Arbeit durchgeführten Homologievergleiche zur Bestimmung des chromosomalen Locus des putativen homologen humanen Gens *EIF3-p47* führten anfangs nicht auf Chromosom 11, sondern zu einer Sequenz, die für die Chromosomenregion **2p16.1** kartiert wurde. Der Vergleich mit der humanen Genomsequenz (Acc.-Nr: AC007250) zeigte in der PIP-Analyse eine Konservierung zu allen acht murinen Exon mit einer Übereinstimmung der Basenpaare von 76% (Exon 1) bis 93% (Exon 5). Die genaue Analyse des kodierenden Abschnitts zur genomischen Chromosom 2-DNA zeigte allerdings, dass das dortige „*EIF3-p47*" aus einer einzigen prozessierten Sequenz besteht und nicht von Intronsequenzen unterbrochen wird. Die Vermutung, dass es sich bei dieser Exonarchitektur, um ein prozessiertes Pseudogen handelt, ließ sich nach Analyse der flankierenden Genbereiche bestätigen: Als Retrotranspositions-typische direkte Sequenzwiederholungen am Genanfang und am Genende im Anschluss an das Poly-A-Ende (Vanin *et al.*, 1985) konnte die 15 bp-Sequenz „AAA GTA AAG CTT ATT" identifiziert werden, die sich beiderseits 10 bp vor dem Startcodon und 34 bp nach dem Poly-A-Signal wiederfindet (Vgl. Abb. 41). Da der gesamte offene Leserahmen erhalten ist, und das Gen somit seine proteinkodierende Kapazität nicht verloren hat,

könnte es sich um ein transkriptionell aktives Pseudogen handeln. Durch RT-PCR-, bzw. Northern-Analysen müsste allerdings diese Vermutung erst noch bestätigt werden.

```
137   CCA TTT GAA CGT TTG AAA GTA AAG CTT ATT CTC GAC AAG ATG GCC   181
                                                        M   A      2

182   ACA CCG GCG GTA CCA GCA AGT GCT CCT CCG GCC ACG CCA GCC CCA   226
  3    T   P   A   V   P   A   S   A   P   P   A   T   P   A   P    17

....   :   :   :   :   :   :   :   :   :   :   :   :   :   :   :   ....

1217  ACA CAG TCA CAG ATT GCC CTC AAT GAA AAA CTT GTA AAC CTG TGA   1261
 348   T   Q   S   Q   I   A   L   N   E   K   L   V   N   L STP    361

1262  ATG GAC CCC AAG CAG TAC ACT TGC TGG TCT AGG TAT TAA CCC CAG   1306

1307  GAC TCA GAA GTG AAG GAG AAA TGG GTT TTT TGT GGT CTT GAG TCA   1351

1352  CAC TGA GAT AGT CAG TTG TGT GTG ACT CTA ATA AAC GGA GCC TAC   1396

1397  CTT TTG TAA ATT AAA AAA AAA AAA AAA GTA AAG CTT ATT GTA AAA   1441
```

Abb. 42 **Sequenzausschnitte des humanen Pseudogens *ΨeIF3-p47* auf Chromosom 2p16.1.** Dargestellt ist der Genanfang und das Genende unter Hervorhebung des Start- und Stoppcodons (fett), des offenen Leserahmens (gelb), des Poly-A-Signals (fett, unterstrichen) und der 15 bp-langen direkten Sequenzwiederholung (rot). Der mittlere Sequenzbereich wurde ausgelassen und nur durch Doppelpunkte angedeutet.

Komparative Untersuchungen der *Eif3-p47*-kodierenden mausgenomischen DNA erbrachten auch Homologien zum humanen Chromosom 12, auf das bereits die Genfamilienmitglieder von *LMO3* und *TULP3* kartiert werden konnten. Die für die untersuchte Chromosom 11p15-Region als paralog anzusehende Chromosomenregion **12q13.3** (siehe auch Kap. 4.6.2) wies einen unter der Acc.-Nr: AC005906 annotierten sequenzierten Genombereich auf, der zum murinen *Eif3-p47* homologe Abschnitte besaß, die in der PIP-Analyse als konserviert dargestellt werden konnten (Vgl. Abb. 28, C1). Die Ähnlichkeit erstreckte sich allerdings nicht über die gesamte kodierende Sequenz wie beim obigen Beispiel auf Chromosom 2, sondern dehnte sich nur auf den 5´-Bereich von Exon 1 und auf die Exons 5 bis 8 aus (siehe Abb. 42). Für die Intronbereiche waren in der genomischen Sequenz keine Homologien zu finden. Die genomische Betrachtung dieser homologen Region mit der mRNA-Sequenz des murinen *Eif3-p47* -Gens, bzw. der Vergleich mit dem humanen Pseudogen auf Chromosom 2 (siehe Abb. 42) zeigte, dass es sich hier um ein weiteres Pseudogen handelt. Die ehemals kodierenden Sequenzbereiche wiesen allerdings im Vergleich zum *ΨEIF3-p47* auf Chromosom 2 zahlreiche Basenaustausche, bzw. Deletionen auf, die sich teils über bis zu einem Duzend Basen erstreckten (siehe Abb. 44). Auch die direkten Sequenzwiederholungen wiesen mehrere alternative Nukleotide auf, so dass es sich zum einen um ein eigenständiges Transpositionsereignis und zum anderen, aufgrund der Deletionen, um eine wesentlich älteres Integrationsereignis handeln dürfte. Auch die phylogenetische Betrachtung aller vier diskutierten Gene (humane *EIF3-p47*, murine *Eif3-p47*, *ΨEIF3-p47-Chr. 2* und *ΨEIF3-p47-Chr. 12*) zeigte, wie in Abb. 45 dargestellt, dass das

Gen *ΨEIF3-p47-Chr. 12* selbst unter Beachtung des deletierten internen Pseudogenbereiches, am weitesten von allen anderen verwandtschaftlich entfernt ist.

A: 12p13-*ΨEIF3-p47*- 5´-Anfang: homolog zum murinen Exon 1:

```
      CCA TGG TCC TCC CAG GCT TTC TTT CTC CAC AAG ATG GCC ACA CCG
1                                                  M   A   T   P    4

      GCG GTA CCA GCA AGT GCT CCT CCA GCC ACG CCA GCC CCA GCC CCG
5      A   V   P   A   S   A   P   P   A   T   P   A   P   A   P   19

      GCA GCG GTC CCA GCC TCA GTT CCA GCA CCA ATG CCA GCA CCG GCT
20     A   A   V   P   A   S   V   P   A   P   M   P   A   P   A   34

      GCG GCT CTG GTT CCC ACT TTC CCT TCT TGA TGC ATT CCC TGG GGC
35     A   A   L   V   P                                       39
```

B: 12p13-*ΨEIF3-p47*-3´-Ende: homolog zu den murinen Exons 5 bis 8:

```
      GAG GAG CAC TTG CTG GTA CCG CCG GTG TGG CCA TCT TGT GGA GAA
       E   E   H   L   L   V   P   P   V   W   P   S   C   G   E

      AGA AAG CCT GGG AGG ACC ATG GCA GTG ATG TTC ACA CCT CTG ACA
228    R   K   P   G   R   T   M   A   V   M   F   T   P   L   T   240

....   :   :   :   :   :   :   :   :   :   :   :   :   :   :   ....

      TCA CAG ATT GCC CTC AGT GAA AAA CTT GTA AAC CTG TGA ATG GAG
350    S   Q   I   A   L   S   E   K   L   V   N   L               361

      CCC AAG CAG TAC GCC TGC TGG TCT AGG TCT TAA CCC CAG GAC TCA

      GAA GTG AAG GAG AAA TGG GTT TTT CGT GGT CTT GAG TCA CAC TGA

      GAC AGT CAA CTG TGT GTG ACT CTA ATA AAC ATG GCC TAC TTT TTG

      TAA ATT AAA AAA AAA AAG AAG GGA AAG TGG ATG TAA TGT GGG AGA
```

Abb. 43 Konservierte Sequenzausschnitte des humanen Pseudogens *ΨEIF3-p47* auf Chromosom 12p13.3. Dargestellt ist die Nukleotid- und die kodierte Aminosäuresequenz des 5´-Bereichs von AS 1 bis AS 39 (unterstrichen) des ORFs (A.) und der 3´-Bereich von AS 228 bis AS 361 (B.). Start- und Stoppcodons wurden fett hervorgehoben und der ORF wurde gelb unterlegt; Poly-A-Signal (fett, unterstrichen) und der 15 bp-langen direkten Sequenzwiederholung (rot) wurden entsprechend hervorgehoben. Der mittlere Sequenzbereich des 3´-Endes wurde ausgelassen und nur durch Doppelpunkte angedeutet.

Nicht geklärt werden konnte die ungewöhnliche genomische Anordnung der beiden homologen Abschnitte. Die Zuordnung von Genanfang und Genende ergab eine „Kopf-an-Kopf"-Anordung, so dass die kodierenden Basen sowohl auf dem Hinstrang, wie auch auf den Gegenstrang zum liegen kamen. Theoretisch wäre diese Anordnung nur durch eine Zurückfaltung und Integration in umgekehrter Ausrichtung in den stromaufwärtsgelegenen 5´-Bereich des Pseudogens zu erklären (Vgl. Abb. 43). Da diese Konstruktion als zu unwahrscheinlich gewertet wurde, dürfte es sich bei dem beobachteten Phänomen wahrscheinlich um ein falsches „Alignment" der genomischen Sequenz und somit um ein Sequenzierungsartefakt handeln.

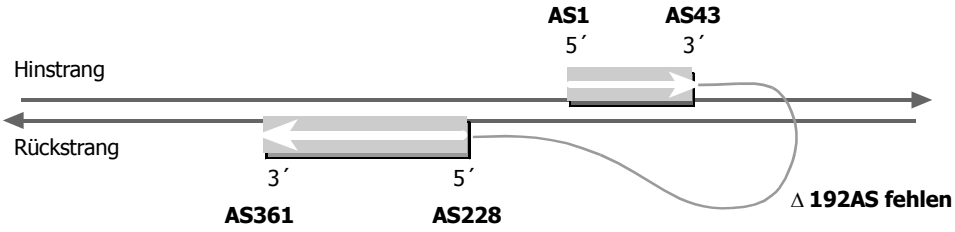


Abb. 44 Grafische Darstellung der genomischen Lage des *ΨeIF3-p47*-Gens auf Chromosom 12p13.3. Der 3´-Bereich von AS 228 bis AS 361 liegt in invertierter Ausrichtung stromaufwärts des 5´-Bereiches von AS 1 bis AS 43. Insgesamt fehlen 192 Aminosäuren aus dem internen Bereich. Diese Anordnung kann lediglich durch eine Zurückfaltung des 3´-Bereiches, bzw. durch ein Sequenzierungsartefakt in der genomischen Sequenz erklärt werden.

Die Existenz von Pseudogenen in der eIF-Genfamilie wurde bereits von Gao & Mitarbeitern (1998) für das humane Gen *EIF4E* beschrieben. Genau wie das putative *ΨEIF3-p47* kartiert auch *ΨEIF4E* in die Chromosomregion 12p13 und weist keinen vollständigen offenen Leserahmen mehr auf. Das funktionelle Gen *EIF4E*, welches auf Chromosom 4 kartiert, besteht seinerseits aus sieben Exons und wird aus einem über 50 kb großen genomischen Bereich transkribiert. Als „Cap-bindung"-Protein ist es Teil des *EIF4*-Faktors, der aus der ATP-abhängigen RNA-Helikase *EIF4A*, dem RNA-bindenden Protein *EIF4B* und dem Protein *EIF4G* besteht, welches Bindestelle für die eben aufgeführten Protein-Untereinheiten und für den Proteinkomplex *EIF3* aufweist. Auch für das Gen *EIF4E* wurde ein nur in vier Nukleotiden unterschiedliches Pseudogen (*ΨEIF4E2*) mit vollständigem Leserahmen charakterisiert, das zusammen mit dem Gen *EIF4E1* differentiell in verschiedenen Zelllinien exprimiert wird (Gao *et al.*, 1998). Dieser Befund unterstützt die Annahme, dass es sich auch bei dem auf Chromosom 2 befindlichen *ΨEIF3-p47*, um ein transkriptionell aktives Pseudogen handeln könnte.

Die definitive Bestätigung, dass das humane Homologe zum murinen *eif3-p47*-Gen auch auf dem humanen Chromosom 11 in der Region 11p15.3 existiert, konnte am Ende der schriftlichen Ausarbeitung durch einen neuen Sequenzeintrag gezeigt werden. Die Annotierung des Klons RP11-236J17 mit der Acc.-Nr. AC116456 beinhaltet 146.502 bp und umfasste damit den gesamten transkribierten Bereich von 9,13 kb des orthologen humanen Gens *EIF3-p47* (Acc.-Nr. O00303).

12p13-*ΨEIF3-p47*-3´-UTR im Vergleich mit dem **2p16-*ΨEIF3-p47*-3´-UTR**

```
2p16_1          ATG GAC CCC AAG CAG TAC ACT TGC TGG TCT AGG TAT TAA CCC CAG
12p13_3         ATG GAG CCC AAG CAG TAC GCC TGC TGG TCT AGG TCT TAA CCC CAG

2p16_1          GAC TCA GAC TCA GAA GTG AAG GAG AAA TGG GTT TTT TGT GGT CTT GAG TCA
12p13_3         --- --- --- --- GAA GTG AAG GAG AAA TGG GTT TTT CGT GGT CTT GAG TCA

2p16_1          CAC TGA CAC TGA GAT AGT CAG TTG TGT GTG ACT CTA ATA AAC GGA GCC TAC
12p13_3         GAC AGT C-- --A -A- --- C-- -TG TGT GTG ACT CTA ATA AAC ATG GCC TAC

2p16_1          TTT TTG CTT TTG TAA ATT AAA AAA AAA AAA AAA GTA AAG CTT ATT GTA AAA
12p13_3         --- --- CTT TTG TAA ATT AAA AAA AAA AAG AAG GGA AAG TGG AT- GTA ATGT
```

Abb. 45 **Nukleotidsequenzvergleich der 3´-UTRs des humanen Pseudogens *ΨEIF3-p47* auf Chr. 12p13.3 mit dem *ΨEIF3-p47* auf Chr. 2p16.1.** Es zeigte sich, dass der Pseudogensequenz mehrer Basen fehlen, bzw. mehrere Nukleotide (blau) ausgetauscht sind. Fett unterstrichen wurde das Polyadenylierungssignal und rot hervorgehoben die direkte Sequenzwiederholung des Transpositionsereignisses. Unterschiedliche Basen dieses „directed repeat" wurden eingekästelt.

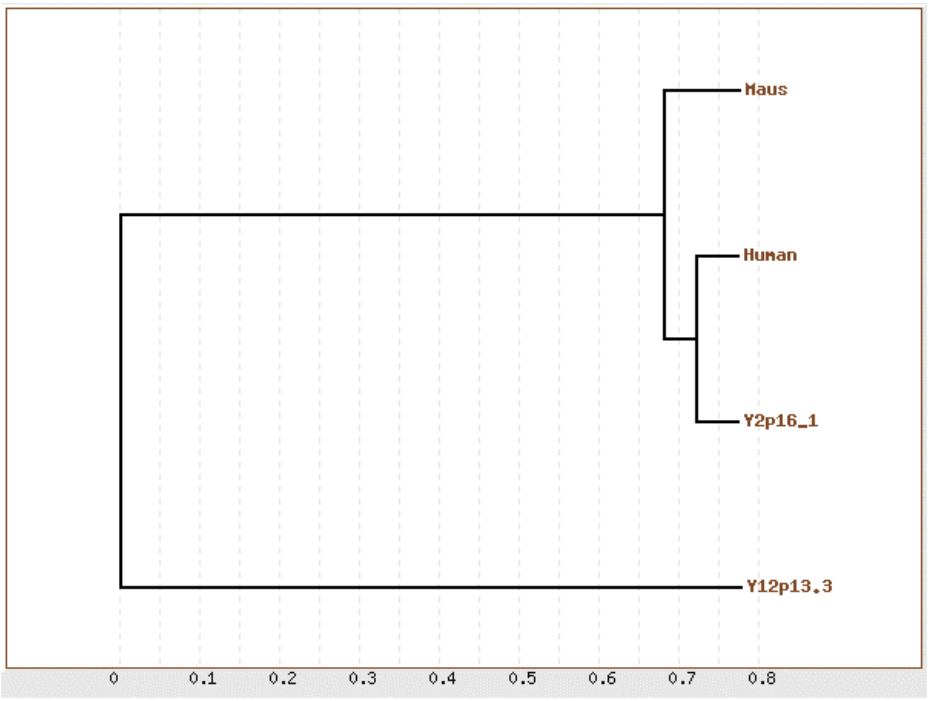


Abb. 46: **Phylogenetischer Vergleich der Proteinsequenzen** zwischen dem neuen murinen Eif3p47 auf Maus-Chromosom 7, dem orthologen humanen Gen EIF3-p47 auf Chromosom 11p15.3 und den beiden humanen Pseudogenen aus den Chromosomenregionen 2p16.1 und 12p13.3. Die Homologie der beiden humanen Pseudogenvarianten zueinander beträgt zwar im N-terminalen Bereich über 43 AS 60% und für den C-terminalen Bereich über 134 AS 85,8%, doch fehlte der Pseudogensequenz des chromosomalen Lokus 12p13.3 der integrale Bereich der Proteinsequenz, so dass die evolutive Distanz in der Grafik zu weit auseinander dargestellt wird.

4.3 Vergleichende Sequenzanalyse der Intergenregionen

4.3.1 Analyse der Nukleotidzusammensetzung

Für die globale Betrachtung der sequenzierten Gesamtregion ist vor allem die G+C- bzw. die A+T-Basenpaare-Verteilung von besonderem Interesse, da GC-reiche Genomregionen mit einer sehr hohen Dichte an genkodierenden Bereichen korreliert zu sein scheinen (Zoubak *et al.*, 1996). Bezugnehmend auf diesen Zusammenhang wurde versucht, verschiedene genomische Bereiche nach ihrem GC-Gehalt zu unterteilen und mit dem fünfstufigen Isochoren-System (L1, L2, H1, H2 und H3) nach Bernardi & Mitarbeitern (1985a) zu klassifizieren. In einem Betrachtungsrahmen von jeweils ca. 300 kb genomischer DNA, der nochmals in 20 kb-Fenster unterteilt wird, findet die Berechnung des GC-Gehalts statt und wird für den Gesamtabschnitt statistisch gemittelt. Die Regionen des Genoms mit einem durchschnittlichen GC-Gehalt von mehr als 52% werden in der GC-reichsten Klasse H3 (H = „heavy") zusammengefasst. Diese Regionen beinhalten insgesamt 17mal so viele Gene wie die GC-ärmsten Regionen L1 und L2 (L = „light") mit einem GC-Gehalt unter 38%, bzw. von 38-42% (Zoubak *et al.*, 1996). Insgesamt werden 63% des gesamten humanen Genoms von den GC-armen Isochoren-Klassen L1 und L2 repräsentiert und nur 4,7% von der Klasse H3 (Bernardi, 2000).

Der in dieser Arbeit untersuchte humane Genombereich wies einen durchschnittlichen GC-Gehalt von insgesamt 48% auf und konnte demnach zum Isochoren-Typ H2 (GC-Gehalt von 47 bis 52%) gezählt werden. Zu diesem Isochoren-Typ gehören insgesamt nur 7,6% aller Genombereiche des Menschen, die aber über ein Viertel (26,1%) aller Gene besitzen (Zoubak *et al.*, 1996). Somit kann die untersuchte Region nach obiger Korrelation von GC-Gehalt und Gendichte zu den Gen-reichen Abschnitten im Genom gezählt werden. Die Erweiterung des Sequenzanalysebereiches auf die gesamte 1,6 Megabasen große Region (vgl. Abb. 47, Kap. 4.5.1) ließ eine leichte Abnahme des GC-Gehaltes in Richtung Centromer verzeichnen. Der durchschnittliche GC-Gehalt der Region mit den Genen *ST5* und *CEGF1* lag bei knapp 45% (Amid & Bahr *et al.*, 2001), und reduzierte sich zum proximalen Ende der Megabase auf einen Wert von 42,7% (Cichutek & Brückmann *et al.*, 2001). Beide Bereiche zählen somit zu der Isochoren-Klasse H1, der einen durchschnittlichen GC-Gehalt von 42-47% aufweist und bereits ein Viertel (24,3%) des Genoms repräsentiert (Bernardi, 2000). Für diesen genomischen Untersuchungsbereich ließ sich demnach ein sukzessiver Anstieg des GC-Gehalts in Richtung Telomer hin verzeichnen.

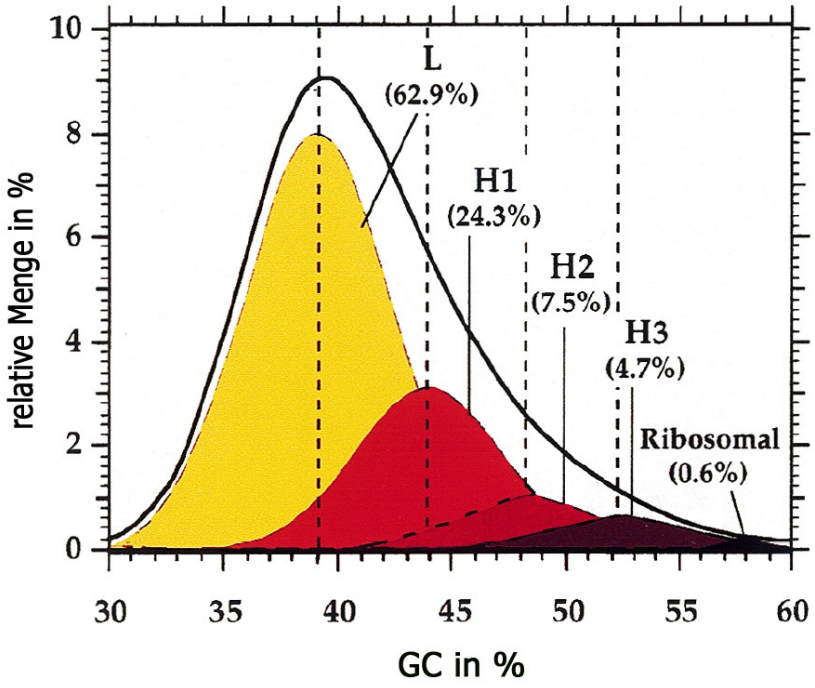


Abb. 47 **Schematische Isochorenverteilung im humanen Genom** (nach Bernardi, 2000). Der GC-Gehalt der Isochoren bewegt sich im Rahmen zwischen 30% und 60% Es zeigt sich eine mosaikartige Organisation der Isochoren im Genom; Bereiche mit ähnlich hohem relativen GC-Gehalt erstrecken sich über Bereiche von im Durchschnitt 300kb. Der Anteil der GC-armen Bereiche (L-Isochoren) im Genom beträgt im Mittel 63%, der an GC-reicheren z.B. H2-Isochroren beträgt nur noch 7,5% (Durchschnittswerte in Klammern). Satelliten DNA wurde in diesem Zusammenhang nicht berücksichtigt.

Der Vergleich des relativen GC-Gehalts des Menschen mit dem im syntänen Bereich der Maus, ergab ein im Durchschnitt uneinheitliches Verhältnis. Beträgt die Interspezies-Differenz in der Region um die Gene *TUB*|*Tub* und *LMO1*|*Lmo1* durchschnittlich +1,3%, d.h. der GC-Gehalt im Menschen ist höher als bei der Maus, so kehrt sich das Verhältnis in den angrenzenden proximalen Abschnitten um die Gene *STK33*|*Stk33* (39,05% vs. 40,05% -> -1,0%), *ST5*|*St5* - *CEGF1*|*Cegf1* (44,9% vs. 45,1% -> -0,2%) und *WEE1*|*Wee1* (42,7% vs. 42,8%) um. In diesen Bereichen liegt der GC-Gehalt in der Maus um 1,0 % bis 0,1% höher als beim Menschen und entspricht dem Genom-typischen Verhältnis beider Spezies, da der Vergleich beider vollständigen Genome von Mensch und Maus gezeigt hat, dass im Mittel der GC-Gehalt der Maus um einen Prozentpunkt über dem des Menschen liegt (Waterston *et al.*, 2002). Eine Erklärung für die GC-Gehaltsverschiebung könnte im hier sequenzierten Bereich mit der Zunahme an repetitiven Elementen in Richtung Centromer gegeben werden (siehe Kap. 4.3.2), da insbesondere LINE-Elemente, die sich durch einen relativ hohen AT-Gehalt auszeichnen, verstärkt auftreten.

Insgesamt besitzt das Mausgenom bei genauer Analyse in der sequenzierten Region dieser Arbeit die Tendenz zu geringeren Schwankungen und zu einem ausgeglicheneren GC-Niveau als beim Mensch, was sich grafisch innerhalb der zugrundegelegten Sequenzintervallen von nicht-überlappenden 500 bp-Abschnitten demonstrieren ließ (vergleiche Abb. 29). Trotz ähnlicher Werte in der

genomischen Verteilung und Dichte der orthologen Gene zeigt das gesamte Mausgenom somit moderatere Schwankungen im GC-Gehalt. Noch geringere Schwankungen im Verteilungsmuster des GC-Gehaltes zeigte das Genom von *Fugu*. Schwankte der GC-Gehalt beim Menschen innerhalb eines 100 kb-Intervalls im Durchschnitt zwischen 34% und 53%, so engte sich bei *Fugu* der Schwankungsbereich auf 44,1% bis 47,5% ein (Aparicio *et al.*, 2002). Tendenziell ist somit in kleineren Genomen mit einer höheren relativen Gendichte auch eine ausgeglichenere GC-Gehalt-Verteilung zu verzeichnen.

Untersuchungen von Saccone & Mitarbeiter (1993) fanden früh heraus, dass die mosaikartige Isochoren-Unterteilung auch mit der cytogenetischen Bandierung der Giemsa-gefärbten Metaphase-Chromosomen in Zusammenhang steht. So können vor allem die unter dem Mikroskop dunkel erscheinenden Banden der Metaphase-Chromosomen mit den GC-armen Isochoren-Klassen L1 und L2 assoziiert werden. Die hellen R-Banden zeigen dagegen eine Korrelation mit den Isochoren H1 und H2. Die Telomer-ständigen T-Banden erwiesen sich am GC-reichsten und setzten sich vor allem aus den Isochoren H2 und H3 zusammen. Diese Zuordnung findet auch für den hier untersuchten Bereich, der in der hellen Chromosomenbande 11p15.3 (vergleiche Abb. 1) lokalisiert ist, ihre Bestätigung.

Auffällig für die untersuchte Region dieser Arbeit und auch für den erweiterten Betrachtungsbereich der 1,6 Megabasenregion sind zwei signifikante grafische Senken in der Darstellung des durchschnittlichen GC-Gehalts. Diese „GC-Senken" befinden sich im 3´-Bereich des *TUB*|*Tub*-Gens und in der Region um das *STK33*|*Stk33*-Gen und dessen 5´-Bereich. Am proximalen Ende des *TUB*|*Tub*-Gens sinkt der durchschnittliche GC-Gehalt auf einer Länge von 78 kb beim Menschen auf 39,73%, bzw. von 53 kb bei der Maus, auf respektive 40,96% ab. In der zweiten Region um das Gen *STK33*|*Stk33* sinkt der GC-Gehalt ebenfalls auf durchschnittlich 39,05% beim Menschen, bzw. auf 40,05% bei der Maus. Interessanterweise zeigen diese beiden Bereiche genau den durchschnittlichen GC-Unterschied von 1% auf, der typisch ist für die beide Genome Mensch und Maus. Die genaue Betrachtung der GC-Senke proximal zum *TUB*|*Tub*-Gen zeigte zudem, dass sich dieser Bereich im Interspeziesvergleich nicht nur durch eine schwache Konservierung, sondern auch durch einen hohen Gehalt an repetitiven Elementen, insbesondere an LINE-Elementen auszeichnet (vgl. Abb. 25b und Abb. 29). Diese konservierten Cluster an repetitiven Elementen innerhalb von AT-reichen Subregionen wurden auch von Engemann & Mitarbeitern (2000) in der weiter chromosomal distal gelegenen Region 11p15.5 beobachtet. Solchen Bereichen wurde dabei eine besondere Rolle für subchromosomale Effekte am Rande von Gengruppen zugeschrieben, die z.B. für epigenetische Kontrollmechanismen von Wichtigkeit sein können und bei der Transkriptionssteuerung von benachbarten Genen mitwirken sollen (Smit, 1999; Bailey *et al.*, 2000).

Eine grundsätzliche Sequenzähnlichkeit zwischen Mensch und Maus zeigte sich auch in der Konservierung der **CpG-Inseln**, kurze CpG-Dinukleotid-Sequenzwiederholungen mit einem sehr hohen GC-Basenanteil, die insbesondere im 5´-Bereich der Gene zu finden sind (Larsen *et al.*, 1992, Bird *et al.*, 1985). So sind die markanten CpG-Inseln der beiden alternativen ersten Exons des Gens *LMO1*|*Lmo1* genauso in beiden Spezies zu finden wie die CpG-Insel in Höhe von *TUB*|*Tub*-Exon 1c

(vgl. Abb. 29). Auch die CpG-Insel innerhalb der GC-Senke (Tab. 22: #10 [Mensch]/#4 [Maus]) zwischen den Genen *LMO1*|*Lmo1* und *TUB*|*Tub* ist in beiden Genomen präsent. Da dieser Abschnitt auch Homologie zu einer cDNA-Sequenz besitzt, könnte dies der Startbereich eines noch unbekannten Gens sein. Auch die CpG-Insel #2 [Mensch], bzw. #1 [Maus] (Tab. 22) findet sich in beiden Genomen wieder. Distal wird dieser Abschnitt gefolgt von einem 380 bp-langen Bereich, der eine Konservierung von 96% aufweist. Trotzdem ließen sich zu diesem sehr konservierten Abschnitt bei der Homologie-Analyse keine Ähnlichkeiten zu exprimierten DNA-Sequenzen ermitteln. So dürfte es sich auch hier, um einen putativen 5´-Bereich eines weiteren unbekannten Gens handeln.

Vergleicht man bei diesen beiden CpG-Inseln die Anzahl der Dinukleotid-Wiederholungen, so fällt auf, dass die Zahl der CpG-Dinukleotide im Mausgenom gegenüber der Humansequenz reduziert ist. Dieses Phänomen konnte mit alle anderen CpG-Inseln bestätigt werden und zeigte sich im GC-Plot an einer verminderten Höhe der GC-„Peaks" (siehe Tab. 22). Sowohl Antequera & Mitarbeiter (1993) wie auch im Gesamtvergleich der beiden Genome Mensch/Maus zeigen diesen speziesspezifischen Unterschied, der sich grafisch in einer ausgeglicheneren Linienführung darstellen lässt (Waterston *et al.*, 2002). Ursache für diesen Schwund könnte die Tatsache sein, dass sich die meisten Cytosine der CpG-Inseln in einem methylierten Zustand befinden. Diese methylierten CpGs können in der Zelle relativ leicht durch spontane Desaminierung des 5-Methylcytosins in ein Thymidin umgewandelt werden, so dass das CpG-Dinukleotid in ein TpG-Dinuleotid mutiert. Folgt danach z. B. ein Replikationsschritt, so tauscht sich an dieser Stelle das G·C-Basenpaar dauerhaft durch ein A·T-Basenpaar aus (Bird, 1980). Diese spontane Methylierung könnte der Grund dafür sein, dass die CpG-Dinukleotide in der globalen Betrachtung des Genoms sowohl beim Menschen wie auch bei der Maus statistisch unterrepräsentiert sind (IHGSC, 2001). Warum allerdings das murine Genom streckenweise einen niedrigeren CpG-Gehalt aufweist als der Mensch, kann dieser Mechanismus, der in beiden Speziesgenomen ähnlich abgelaufen sein dürfte, nicht erklären. Vielmehr scheinen bestimmte Genombereiche stärker diesen Veränderungen unterworfen gewesen zu sein als andere. Gleichzeitig zeigte die CpG-Analyse auch, dass dieselben Programm-Parameter der Software CPG-FINDER für die humane Sequenz mehr CpG-Inseln berechnen als für die orthologe Maussequenz. Diese Beobachtung bestätigt sich nicht nur in der angrenzenden Region mit den Genen *STK33*|*Stk33*, *ST5*|*St5*, *CEGF1*|*Cegf1*, in der sich von 10 CpG-Inseln nur 6 auch bei der Maus wiederfinden (Amid & Bahr *et al.*, 2001). Auch der globale Genomvergleich beider Spezies zeigt diesen speziesspezifischen Unterschied. Generell weist demnach das Mausgenom weniger CpG-Inseln auf als das humane Genom, obwohl es einen um durchschnittlich einen Prozent höheren globalen GC-Gehalt besitzt (Antequera & Bird, 1993; Waterston *et al.*, 2002). Eine Erklärung hierfür könnte sowohl eine höhere Substitutionsrate durch Desaminierung speziell in den 5´-Bereichen der Gene sein oder aber auch auf einen Algorithmusfehler des verwendeten Programms CPG-FINDER hinweisen, da die CpG-Spitzen sich weniger stark vor dem insgesamt höheren G·C-Basenpaar-Hintergrund diskriminieren lassen (Adams & Eason, 1984). Zusammenfassend scheinen insbesondere die funktionellen CpG-Inseln evolutiv stärker konserviert geblieben zu sein. Ein Ergebnis, das auch durch den hier analysierten Bereich mit seinen bekannten Genen eine Bestätigung findet und in Kap. 4.4.1 nochmals detaillierter diskutiert wird.

4.3.2 Analyse der interspergiert repetitiver Elemente

Ebenso wie genkodierende Bereiche und die unterschiedliche Verteilung des GC-Gehalt in der genomischen DNA funktionelle Informationen über einen bestimmten genomischen DNA-Abschnitt geben, so liefern auch repetitive Sequenzen – die im Menschen immerhin über 50% des gesamten DNA-Materials stellen - in der Art und Weise ihres Vorkommens und der Verteilung Hinweise zu ihrer Entstehung. Im Interspezies-Vergleich der verschiedenen repetitiven Klassen können Aussagen über die Dynamik von z.B. transposablen Elementen gemacht werden, und die Transpositionsrate erlaubt Hypothesen über die evolutive Vergangenheit und die Dynamik eines Genomabschnitts. So bieten die repetitiven Elemente einen interessanten Aspekt in der komparativen Analyse von DNA-Sequenzen.

Die Auswertung der Dotplot-Analyse (Kap. 3.8.1.1) der genomischen Consensussequenzen von Mensch und Maus zeigte, dass sich in der Grafik hinter den Lücken der fast durchgehenden Diagonalen des syntänen Bereichs immer eine zahlenmäßige Anhäufung von zusätzlichen repetitiven Elementen in der Humansequenz verbarg (Vgl. Abb. 31). Die humane Sequenz wies im Vergleich zur homologen Maussequenz im Durchschnitt 6,7% mehr repetitive DNA-Bereiche auf. Dieser Wert entspricht in Annäherung dem Verhältnis, das auch im Vergleich der Gesamtgenome beider Spezies zu verzeichnen war. Hier beträgt der Unterschied 7,8%, die das humane Genom mehr an repetitiven Bereichen besitzt als die Maus (Waterston *et al.*, 2002). Dass dieser Unterschiedswert keine statische Größe zu sein scheint, spiegeln die Werte der angrenzenden Bereiche um die Gene *ST5-CEGF1*|*St5-Cegf1*, *WEE1*|*Wee1* und *CARS-ASCL2*|*Cars-Ascl2* wider, deren Differenzen zwischen 13,7% und 3,2% schwanken. Als signifikant dürfte aber die tendenzielle Abnahme des interspergiert repetitiven Anteils der genomischen DNA im betrachteten Ausschnitt der Region *WEE1*|*Wee1* bis *TUB*|*Tub* und *CARS-ASCL2*|*Cars-Ascl2* in Richtung Telomer sein, innerhalb der der durchschnittliche repetitive Prozentsatz sich nahezu halbiert. Dieser Rückgang ist in sehr ähnlichem Umfang auch für den syntänen murinen Genombereich zu beobachten. Das diese Abnahme spezifisch für den telomeren Bereich eines Chromosoms, bzw. typisch für Chromosom 11, respektive Mauschromosom 7 wäre, konnte durch die Literatur nicht bestätigt werden. Dagegen spricht auch die Lokalisation der beiden Regionen *Wee1*|*Tub* und *Cars*|*Ascl2*, die auf Mauschromosom 7 nicht in einem durchgehenden chromosomalen Syntäniebereich liegen.

Eine Zusammenstellung des repetitiven Anteils in der sequenzierten Gesamtregion im Vergleich zum Gesamtgenom ist in nachfolgender Tabelle 24 zusammengefasst.

Tab. 28 Zusammenfassung der REPEATMASKER-Analysen verschiedener Chromosomenregionen. Gegenübergestellt ist der relative Anteil an interspergiert repetitiven Elementen unterteilt in die verschiedenen Repeat-Klassen der jeweils orthologen Genombereiche von Mensch und Maus. Die Angaben für die Region WEE1/RanBP7 nach Cichutek & Brückmann *et al.*, 2001; CEGF1/ST5 nach Amid & Bahr *et al.*, 2001; der Region CARS/ASCL2 nach Engemann *et al.*, 2000; Angaben der Spalte Genom nach Waterston *et al*, 2002. In Feldern mit „k.A." lagen keine Angaben vor.

	11p15.3			11p15.5	Genom
		Region WEE1/RanBP7	**Region CEGF1/ST5**	**Region LMO1/TUB**	**Region CARS/ASCL2**	
Mensch		34,38%	21,40%	17,72%	k.A.	20,99%
	SINEs	14,58%	15,33%	10,31%	k.A.	13,64%
	LTRs	2,37%	3,49%	4,58%	k.A.	8,55%
	DNA Elem.	2,74%	2,97%	1,93%	k.A.	3,03%
	Total nterspergierte Sequenzen	52,6%	43,79%	34,53%	25,8%	46,36%
Maus	LINEs	27,37%	12,93%	11,13%	k.A.	19,20%
	SINEs	2,37%	11,24%	7,62%	k.A.	8,22%
	LTRs	8,40%	4,89%	4,59%	k.A.	9,87%
	DNA Elem.	0,93%	0,72%	0,84%	k.A.	0,88%
	Total nterspergierte Sequenzen	41,87%	30,12%	27,82%	22,6%	38,55%
	Differenz des repetitiven Anteils	10,73%	13,67%	6,71%	3,2%	7,81%
		Maus-Chr. 7 F2			**Maus-Chr. 7 F5**	

Die Analyse zeigte, dass auch innerhalb der Teilregionen erhebliche Schwankungen zu verzeichnen sind, die sich nicht auf eine zufällige Mehrverteilung der repetitiven Elemente zurückführen ließen. Vielmehr waren „Cluster" zu beobachten, die auf eine vermehrte Integration von bestimmten Klassen an repetitiven Elementen schließen lassen. Insbesondere der Bereich mit verminderten GC-Gehalt zwischen Nukleotid 185.000 und 263.000 der humanen Consensussequenz und zwischen Nukleotid 213.00 bis 266.000 der murinen Consensussequenz zeigte eine sehr starke Zunahme an repetitiven Elementen, die sich größtenteils durch eine Vermehrung der transposablen Elemente zusammensetzt. Die transposablen Elemente können in vier verschiedene Klassen unterteilt werden, die sich in beiden Speziesgenomen nachweisen ließen, allerdings mit speziesspezifischen Unterschieden.

Die bedeutendste Klasse an interspergiert repetitiven Elementen waren die **LINE-Elemente** („long interspersed nucleotide element"), die allein 21% der gesamten genomischen DNA im Menschen und 19,2% des Mausgenoms stellen (Waterston *et al.*, 2002). LINE-Elemente zeichnen sich durch die Besonderheit aus, dass sie sich durch ihre Transposase-Aktivität selbstständig im Genom vermehren können. Mit einer Länge von 6 kb, einem internen Polymerase II-Promotor und zwei offenen Leserahmen sind sie in der Lage über reverse Transposition – nach Translation und reverse Transkription der LINE-RNA – in neue Bereiche des Genoms mithilfe von 7-20 bp langen Zielsequenz-Duplikationen („target site duplication") zu integrieren. Da die reverse Transkription sich vom 3´-Ende

der LINE-RNA vollzieht und in den seltensten Fällen nur bis zum vollständigen 5´-Ende des Transkripts gelangt, finden sich meistens trunkierte und nicht mehr funktionelle LINE-Fragmente im Genom. So beträgt die durchschnittliche Länge des LINE1-Elements (L1) – die größte Fraktion der LINE-Elemente – ca. 900 bp im humanen Genom (IHGSC, 2001). Für den hier untersuchten Humanbereich konnte sogar nur eine durchschnittliche Länge von 692 bp ermittelt werden. Allerdings wies der hoch repetitive Bereich der „GC-Senke" (Vgl. Kap. 3.10.1) mit 940 bp pro L1 einen leicht höheren Wert als im Gesamtgenomdurchschnitt auf. Da es sich bei allen L1-Elementen nur noch um Fragmente eines höchstwahrscheinlich gemeinsamen Vorläufers handelt, dürfte von den allermeisten L1-Elementen kein Transposase-Aktivität mehr ausgehen, da ihre offenen Leserahmen nicht mehr vollständig vorhanden sind. Lediglich ein L1-Element im Bereich der „GC-Senke" mit einer Länge von 4.376 bp (nt 232.166 bis nt 236.541 der humanen Sequenz) und einer Homologie von 97,4% wies noch einen vollständigen ORF auf. Da dieses Elemente exklusive im Humangenom zu finden war, müsste es sich um ein vergleichsweise junges Integrationsereignis handeln (Smit, 1996).

Für die L1-Elemente im Mausgenom zeigte sich ein ähnliches Bild; die durchschnittliche Länge beträgt im Gesamtgenom 776 bp/L1 (Waterston *et al.*, 2002). Auch dieser Durchschnittswert wurde von den L1-Elementen des hier untersuchten Genomabschnitts mit 683 bp/L1 unterschritten. Ebenso wie im Humangenom dürften auch bei der Maus die meisten L1-Elemente nicht mehr aktiv sein. Diese Vermutung wird auch durch die Analyse des Mausgenoms bestätigt, die in über 660.000 L1-Kopien nur 12 Kopien finden konnte, die beide ORFs in intaktem Zustand beinhalteten (Waterston *et al.*, 2002).

Die kopienreichste Klasse der interspergiert repetitiven Elemente bilden die **SINE-Elemente** („short interspersed nucleotide element"), deren Zahl im Humangenom mit über 1,5 Millionen angegeben wird und 10,6% der gesamten Genomsequenz ausmachen (IHGSC, 2001). Auch im murinen Genom konnten 1,498 Millionen dieser etwa 100 bis 400 bp langen und einen Polymerase III-Promotor beinhaltenden Elemente gezählt werden. Werden die SINES im humanen Genom hauptsächlich durch die aktiven Alu-Elemente (ca. 70% aller SINE-Elemente) und den selteneren MIR-, bzw. MIR3-Elementen repräsentiert, so teilt sich die Klasse der SINES bei der Maus in vier Gruppierungen, bestehend aus B1-, B2, ID- und B4-Elementen, auf. Die humanen Alu-Elemente und die murinen B1-Elemente scheinen sich dabei von einem gemeinsamen Vorläufer einer 7SL RNA abzuleiten (Quentin *et al.*, 1994). B2-Elemente lassen sich auf eine tRNA-abgeleitete Promotor-Region und ID-Elemente auf das neuronal exprimiertes RNA-Gen *BC1* zurückführen. B4-Elemente scheinen sich aus der Fusion von B1- und ID-Elementen rekrutiert zu haben (Serdobova & Kramerov, 1998). Da alle SINE-Elemente keine autonome Transposase-Aktivität besitzen, haben sie sich im Laufe der Evolution nur in Verbindung mit der Transposition der LINE-Elemente vermehren können. Man findet daher oft SINES mit gemeinsamen 3´-Ende eines benachbarten LINE-Elements. Im Interspezies-Vergleich liegt die Zahl der SINE-Elemente im Genom der Maus höher als im Genom des Menschen. Da aber die durchschnittliche Größe der murinen SINES mit 135 bp/Element kleiner ist als mit 230 bp/humanen Element ist der prozentuale Anteil am Gesamtgenom geringer. Dieser Unterschied konnte auch

anhand der hier vorliegenden Genom-Sequenzen bestätigt werden, wo 7,7% der Maus-DNA und 10,3% der Human-DNA aus SINE-Elementen bestehen.

Die dritte Klasse der transposablen Elemente bilden die **LTR-Elemente** („long terminal direct repeats), die sich von Vertebraten-spezifischen, retroviralen Retrotransposons ableiten lassen und daher auch mit der Abkürzung „ERV" für „endogenous retroviral-like element" bezeichnet werden (Grifford & Tristem, 2003). Als eigenständige autonome Elemente setzen sie sich aus einem Protease-kodierenden (*gag*) und einem die Reverse Transkriptase-kodierenden (*pol*) Gen, einer RNase H und einer Integrase zusammen. Typisch sind die terminalen langen Sequenzwiederholungen über die sich die Transposition vollzieht. Unterteilt werden die LTR-Elemente in drei unterschiedliche Klasse (I bis III), die in Mensch und Maus eine äußerst unterschiedliche Verteilung in beiden Genomen zeigen und wahrscheinlich eine verschiedenartige Vergangenheit durchlebt haben müssen. So zählen zur Klasse III etwa 80% aller bekannten LTR-Elemente. Insbesondere die nicht-autonomen MaLRs („mammalian LTR-Retrotransposon") dieser Klasse waren im Mausgenom besonders erfolgreich und sind mit 388.000 Kopien (= 4,8%) (Waterston *et al.*, 2002) stärker vertreten als im Menschen mit seinen 240.000 Kopien (= 3,8%) (IHGSC, 2001). Auch die Klasse II ERVs der Maus weisen eine ca. zehnmal dichtere Anordnung als im Menschen auf. In den hier untersuchten Genomsequenzen konnten somit auch nur drei Klasse II ERVs bei der Maus bestimmt werden, die in der Maus-Consensussequenz bei nt 33.852 bis nt 34.062, nt 205.652 bis nt 206.093 und nt 397.491 bis nt 398.467 lagen. Welche Rolle dieser ERV Klasse II-Elementen zukommt, zeigt die Beobachtung, dass in 15% aller spontanen Mausmutanten ein Allel nachgewiesen werden kann, das mit IAP („intracisternal-A particles") oder ETn („early-transposons")-Insertionen assoziiert ist; zwei aktiven Gruppen innerhalb dieser ERV-Klasse II. Das umgekehrte Verhältnis zeigen die Klasse I ERVs, die im gesamten Humangenom mehr als viermal so häufig anzutreffen sind, wie im Mausgenom. Auch in den hier analysierte Genomsequenzen waren 1,52% der humanen DNA und nur 0,25% der murinen DNA zu dieser ERV-Klasse I zählend. Warum dieser Unterschied besteht, ist zur Zeit noch völlig ungeklärt.

Insgesamt zeigte sich, dass die hier untersuchten genomischen Bereiche etwa nur die Hälfte an LTR-Elementen aufwiesen, wie es im gesamtgenomischen Durchschnitt der Fall ist. Im erweiterten Betrachtungsfokus der 1,6 Megabasenregion war für das humane Genom in Richtung Centromer eine Abnahme um fast die Hälfte von 4,58% auf 2,37% zu verzeichnen. Diese Tendenz war in syntänen Maus-genomischen Abschnitt interessanterweise genau invertiert; hier war für den Bereich von *Tub* nach *Wee1* eine Zunahme von 4,59% auf 8,40% festzustellen (Cichutek & Brückmann *et al.*, 2001).
Als vierte Klasse der interspergiert repetitiven Elemente wurden die **DNA-Transposons** untersucht, die mithilfe ihrer Transposase-Aktivität und kurzen terminal invertierten Sequenzwiederholungen sich durch einen simplen „cut & paste"-Mechanismus durch Exzision und Reintegration im Genom „bewegen" können. Diese muss sich nicht unbedingt nur replikativ vollziehen, es kann dadurch auch zur Duplikation von ganzen Elementbereichen kommen, wenn die entstandene Lücke über das Schwesterchromatid von der Zelle repariert werden kann (Engels *et al.*, 1990). Ebenso werden ganze Chromosomen-Rearrangements als Resultat der Aktivität von DNA-Transposons diskutiert (Lim &

Simmons, 1994). Unterteilt werden die DNA-Transposons beim Menschen in sieben Hauptgruppen; für die Maus werden nur vier Gruppen unterschieden (Smit *et al.*, 1996). In dieser Arbeit wurde sich im Rahmen der REPEATMASKER-Analyse auf zwei Untergruppen, den MER T1 und MER T2 („medium reiteration frequency interspersed repeats"), beschränkt. Da DNA-Transposons zur Translation der kodierenden Information ihrer DNA ins Zytoplasma gelangen müssen, kann das Enzym nach Rückkehr in den Nukleus meist nicht mehr zwischen aktiven und inaktiven Elementen unterscheiden, so dass es in vielen Fällen zu einer Anhäufung von inaktiven Kopien im Genom kommt. Dadurch wird die Transposition so ineffektiv, dass die Expansion innerhalb eines Genoms ganz zum Erliegen kommt. Der Hauptverbreitungsweg der DNA-Transposons ist daher hauptsächlich der horizontale Transfer in andere Genome (Robertson & Lampe, 1995). Dies ist ein elementarer Unterschied zu der Verbreitungsstrategie der LINE- und SINE-Elemente, die exklusiv auf den vertikalen Transfer innerhalb eines Genoms beschränkt bleiben (Smit *et al.*, 1996). Die quantitative Betrachtung des DNA-Transposonanteils in beiden Genomen zeigte im Vergleich zur Maus ein erhöhtes Vorkommen im Menschen. Mit insgesamt 1,93% lag der Wert um einen Prozentpunkt unter dem Durchschnitt des Humangenoms. Erst die sich anschließenden Bereiche in Richtung *WEE1*-Gen zeigten eine Annährung an den globalen Mittelwert von 3%. Der Gehalt an DNA-Transposons in der untersuchten Maus-Sequenz unterschied sich – wie auch in der erweiterten Betrachtung der 1,6 Megabasen-Region – mit 0,88% nur unwesentlich vom Mittelwert des murinen Gesamtgenoms.

Eine interessante Analyse erlaubten Repeat-Cluster innerhalb der konservierten Genomsequenz, die im Laufe der getrennten Evolution im humanen Genom im Vergleich zur Maus eine starke Expansion, d.h. ein Zugewinn von interspergiert repetitiver DNA erfahren haben. Der direkte Interspezies-Vergleich von drei exemplarisch ausgewählten Bereiche A, B und C (Siehe Abb. 31, Kap. 3.10.3) machte deutlich, dass vor allem die LINE-Elemente in diesen Abschnitten ihre Zahl verdoppelt bis verdreifacht haben. Die Zahl der SINE-Elemente blieb dagegen bis auf den Bereich C meist konstant. In Bereich C hat sich wahrscheinlich eine Verdoppelung der SINES im Menschen ereignet. Die Detailanalyse zeigte, dass in Bereich B und C die Leserichtung der LINEs für alle Elemente einheitlich konserviert geblieben ist. Bleibt in Bereich A der relative prozentuale Anteil an repetitiver DNA mit über 60% trotz der Verzehnfachung der Sequenzlänge nahezu gleich, so verdoppelte sich der relative repetitive Sequenzgehalt in den Bereichen B und C. Im Bereichsabschnitt C stieg der Anteil an interspergiert repetitiver DNA im Menschen sogar bis auf über 90% an. Eine ähnliche Tendenz war auch für den großen Bereich mit vermindertem GC-Gehalt zu verzeichnen. Auch hier kam es zu einer Anhäufung an repetitiver DNA, die überwiegend auf eine Verdreifachung an LINE-Elementen zurückzuführen war. SINE-, LTR- und MER-Elemente (DNA-Transposons) blieben in ihrer absoluten Anzahl nahezu konstant, nur ihre Größe pro Element war im Humangenom länger. Waren die SINE im Menschen durchschnittlich 263 bp lang, so konnte im Mausgenom eine durchschnittliche Länge von nur 135 bp ermittelt werden. Dies hatte zur Folge, das die Differenz von 25 kb zwischen der humanen und murinen Sequenz ausschließlich auf die zusätzliche Integration von interspergiert repetitiven Elementen im Menschen zurückgeführt werden konnte. Zeigte der Mausbereich im Vergleich zur gesamten Sequenz nur eine Zunahme des interspergiert repetitiven Anteils von 28,35% auf 30,59%,

so erhöhte sich dieser Wert für den humanen Abschnitt von 34,53% auf 57,53%. Der Gesamtanteil an interspergierten repetitiven Elementen lag hier mit durchschnittlich 34,5% im humanen Sequenzabschnitt um 11,8 Prozentpunkte unter dem Wert für das Gesamtgenom mit 46,3% (Waterston *et al*, 2002). Auch der repetitive Anteil der murinen Genomsequenz zeigte im Verhältnis zum Gesamtgenom mit 10,7 Prozentpunkten eine ähnliche Reduzierung. Die Betrachtung innerhalb der 1,6 Megabasen-Gesamtregion des Menschen zeigte tendenziell wie der syntäne Mausbereich auch eine Zunahme des interspergiert-repetitiven Anteils in Richtung Centromer. Ein weiterer interessanter Punkt war die Frage, in wie weit der repetitive Anteil der DNA den GC-Gehalt beieinflußt, bzw. wie sehr der Anteil an interspergiert repetitiven Elementen in Beziehung zu der Isochoreneinteilung steht. Die globale Betrachtung des Gesamtgenoms zeigte, dass in GC-reichen Abschnitten vor allem Alu-Elemente (SINES) anzutreffen sind und in H2-Isochoren ihr Maximum mit 18,7% erreichen (Pavlicek *et al*., 2001). LINE-Elemente dagegen sind eher in den GC-armen Isochoren L2 und L1 zu finden. Ihr relativer prozentualer Anteil ist in den der GC-reichsten Isochoren-Klasse H3 mit 4,4% am niedrigsten und mit 22,5% in L1 am höchsten. Tendenziell kann diese Aussage mit den Ergebnissen dieser Arbeit und unter Hinzunahme der Betrachtung der 1,6 Megabasenregion entsprochen werden. Doch differieren die prozentualen Werte der Genomanalyse von dem hier untersuchten Sequenzbereichen um einiges. Der in dieser Arbeit sequenzierte humane und in die Isochoren-Klasse H2 eingestufte Genombereich zeigte einen Alu-Sequenzanteil von nur 10,3% (vgl. Tab. 23). Für das Gesamtgenom wird ein wesentlich höherer Wert von 18,7% angegeben (Pavlicek *et al*., 2001). Auch der hohe LINE-Anteil von 17,7% weicht deutlich vom Genomwert von 7,1% für diese Isochoren-Klasse H2 ab.

In wie weit die repetitiven Elemente wirklich mit Ihrer DNA den GC-Gehalt beeinflussen, sollte der Vergleich des GC-Gehalts ohne und mit der um die repetitiven Bereiche maskierten Sequenz demonstrieren. Es zeigte sich dabei, dass für den Bereich der „GC-Senke" der GC-Gehalt inklusive der repetitiven Elemente (repetitiver Sequenzanteil: 61,85%) sich nur unwesentlich von 39,73% um etwa einen halben Prozentpunkt auf 39,18% nach Maskierung der repetitiven Anteile erniedrigte. Das gleiche Verhältnis konnte auch für die orthologe Maus-genomische Sequenz dieser Region erbracht werden. Der GC-Gehalt von 41,19% mit repetitiven Elementen veränderte sich nur auf 40,6% nach Maskierung ohne die repetitiven Anteile. Genau diese Beobachtung wurde auch Pavlicwk & Mitarbeiter (2001) dokumentiert. Somit trägt der Anteil an interspergiert repetitiven Elementen nicht maßgeblich an einer etwaigen Veränderung des GC-Gehaltes bei.

4.4 Vergleichende Sequenzanalyse insbesondere konservierter Bereiche ohne proteinkodierende Funktion

Die vergleichende Genomanalyse zwischen Mensch und Maus zeigte in der vorliegenden Arbeit, dass sich die Bereiche der hochkonservierten Sequenzabschnitte nicht nur auf die genkodierenden Gebiete beschränken, sondern dass ebenso signifikante Homologien zu CpG-Inseln in den 5´-Bereichen der Gene, gefunden werden konnten. Darüberhinaus existieren insbesondere in den Intronbereichen der beiden Gene *LMO1*|*Lmo1* und *TUB*|*Tub* eine Reihe von stark konservierten, nicht proteinkodierenden Abschnitte, denen keine eindeutige Funktion zugeordnet werden konnte. Da aber nicht davon auszugehen ist, dass sich diese Konservierung rein zufällig über die 60 bis 80 Millionen Jahre der unterschiedlichen Evolution von Mensch und Maus erhalten haben, bleibt die Frage, welche funktionellen Aspekte hinter diesen konservierten Abschnitten stecken könnten. Da dieses Problem im Rahmen der vorliegenden Arbeit nicht mehr auf experimentellem Wege, wie z. B. durch Expressionsanalysen nachgegangen werden konnte, sollen hier einige mögliche Funktionen diskutiert werden.

4.4.1 Konservierte CpG-Inseln

Für den in dieser Arbeit untersuchten humanen Genombereich konnten insgesamt 13 CpG-Inseln identifizierte werden, die bis auf eine Ausnahme sich auch alle im Mausgenom wiederfanden. Aufgrund der in der Maus reduzierten Gesamtzahl an CpG-Dinukleotiden waren auf dem ersten Blick viele nicht mehr als CpG-Inseln mithilfe des Programms CPG-FINDER zu identifizieren. Aber aufgrund der Sequenzkonservierung von über 50% war ihr Bereich trotzdem durch die PIP-Blot-Analyse anzusprechen. Lediglich für die CpG-Insel #11 der humanen Sequenz (Tab. 22) konnte kein murines Pendant ermittelt werden. Die genaue Interspeziesanalyse zeigte, dass der im Mausgenom fehlende CpG-Inselbereich im humanen Genom von repetitiven Elementen flankiert wird, so dass dies als Hinweis auf ein evolutionär jüngeres Integrationsereignis gedeutet werden könnte. Von den übrigen konservierten CpG-Inseln konnten vier den beiden Genen *LMO1* und *TUB* zugeordnet werden, deren beide alternativen 5´-Bereiche jeweils durch eine CpG-Insel (#7, #8, #12, #13) markiert wurden. Die weiteren acht CpG-Insel stellen Genomabschnitt dar, deren etwaige Funktion mit bioinformatischen Hilfsmitteln nicht hinreichend geklärt werden konnte. Die lokale Assoziation der Inseln #2 und #10 mit cDNA-Sequenzen, könnte als erstes Indiz für die Existenz auf die 5´-Bereiche zweier unbekannter neuer Gene gesehen werden. Dass es sich bei diesen beiden CpG-Inseln um genassoziierte Bereiche handelt, konnte auch durch die vergleichende CpG-Analyse des Maus-genomischen Abschnitts unterstützt werden, da die homologen Mausbereiche ebenfalls als CpG-Inseln identifiziert werden konnten. Für die konservierten Bereiche der humanen CpG-Inseln #1 und #5 gab die

Promotorvorhersage mit dem Programm PROSCAN II einen Hinweis, dass es sich hier um putative Promotoren aus unbekannten 5´-Bereichen neuer exprimierter Genomabschnitte handeln könnte. Den CpG-Inseln #3, #4 und #6 konnten mit Hilfe der bioinformatischen Analyse keine weiteren Aspekte zugeordnet werden. Auffällig war allerdings die Konservierung der CpG-Sequenzbereiche #3 und #4, deren Interspezieshomologie zwischen 60 und 80% lag. Auch das Vorhandensein eines noch stärker konservierten Sequenzbereiches in unmittelbarer Nähe, wie bei #4 mit 97% über 230 bp unterstrich die These für einen unbekannten Genstart. Lediglich CpG-Insel #6 zeigte keine Sequenzkonservierung zum murinen Genom, befand sich aber in nur 2,5 kb Distanz von CpG-Insel #5, in deren gemeinsame Mitte sich ein mit 98% konservierter Bereich über 530 bp befindet. Daher könnte auch diese CpG-Inseln (#6) mit einem kodierenden Sequenzbereich in Zusammenhang stehen. Nur die vergleichende Genomanalyse erlaubte es in der vorliegenden Arbeit alle untersuchten CpG-Inseln mit putativ exprimierten Sequenzbereichen in Verbindung zu bringen, und die in der Literatur vertretene Ansicht, dass CpG-Inseln immer mit Genen assoziiert sind, zu bestätigen (Aissani & Bernardi, 1991; Gardiner-Garden & Frommer, 1987).

4.4.2 Konservierte AT-reiche MAR-Regionen

DNA-Moleküle tragen in ihrer Sequenzabfolge nicht nur die Information für die Aminosäureabfolge von Proteinen oder kodieren Einheiten, die für die regulative Steuerung der exprimierten Bereiche zuständig sind (Promotoren, Enhancer; siehe auch Kap. 4.3.3), die DNA-Sequenz kann auch in ihrer Basenpaarzusammensetzung die dreidimensionale Struktur des Chromatins im Interphasekern vorgeben (Vogelstein *et al.*, 1980). Untersuchungen zeigten, dass kurze AT-reiche Sequenzbereiche die Fähigkeit besitzen das DNA-Molekül in Schleifen zu legen und es an die Kernmembran zu binden. Die als **MARs** („matrix attachment regions") bezeichneten Regionen (Cockerhill *et al.*, 1986) konnten auch in den hier untersuchten Genomsequenzen mit Hilfe des Programms MAR-FINDER vorhergesagt werden. Der Interspezies-Homologievergleich zeigte, dass bestimmte MARs auch in ihrer Position über die Evolution zwischen Mensch und Maus konserviert geblieben sind. So konnte MAR #1 des Menschen mit MAR #1 der Maus durch eine Sequenzhomologie von 70% gleichgesetzt werden. Auch die schwach ausgebildete MAR-Region im 5´-Bereich des *LMO1*/*Lmo1*-Gens fand sich als konservierter Bereich in beiden Genomen wieder. Der zweite MAR (#2) der humanen Sequenz war an der sequenzhomologen Stelle im Mausgenom nicht zu identifizieren, zeigte aber eine Sequenzkonservierung in diesem Bereich von über 75%. Der zweite MAR der Maus-genomischen Sequenz liegt in einem Abschnitt der keinerlei Homologie zum humanen Genom aufzeigt, aber in die durchschnittlich sehr AT-reiche Region der „GC-Senke" fällt. Untersucht man die Sequenzabstände der vorhergesagten MARs im sequenzierten Bereich dieser Arbeit, so fällt auf, dass der Abstand mit durchschnittlich 140 kb in der Maus höher liegt als mit 104 kb im Menschen. Da für die humangenomische Sequenz allerdings nur zwei MARs detektiert werden konnten, sollte diesem Wert kein all zu repräsentativer Charakter beigemessen werden. Der Vergleich mit anderen komparativ untersuchten Regionen, wie z.B. die Region 11p15.5 mit den Genen *CARS*/*Cars*-*NAP1L4*/*Nap1l4*-

KCNQ1/*Kcnq1* zeigte, dass sich dieser Wert nicht wesentlich von den dortigen MAR-Abständen unterscheidet. Für diese humane Region wurde eine durchschnittliche MAR-Entfernung von 91 kb ermittelt (Engemann *et al*, 2000); die homologe murine Region zeigte mit Distanzen zwischen 75 und 100 kb ebenfalls einen Mittelwert von 90,8 kb. Somit weisen beide Spezies sehr ähnliche Abstände auf. Allerdings konnte eine Interspezies-Konservierung in der Region 11p15.5 der MARs wie für den oben beschriebenen MAR #1 nicht festgestellt werden. Die hiesige Konservierung dürfte daher eher eine Ausnahme darstellen und nicht die Regel bedeuten. Allerdings beschreiben auch Greally & Mitarbeiter (1999) für die Chromosomenregion 15q11-q13, respektive Maus 7C eine konservierte Anordnung der orthologen MAR-Loci. Eine Erklärung für den deutlichen Unterschied im Abstand der MARs im Mausgenom könnte z. B. an der unterschiedlichen topografischen Lage der beiden untersuchten Regionen liegen. Beide humanen Regionen des Menschen – 11p15.3 und 11p15.5 – liegen am telomeren Ende des Chromosoms 11 in einer Entfernung von ungefähr 5,3 Mb zueinander (http://www.ensembl.org/). Auch die syntäne Mausregion der humanen Bande 11p15.5 liegt am telomeren Ende F5 des Mauschromosoms 7, so dass die Gemeinsamkeit der chromosomalen Lage ein Kriterium für den ähnlichen Abstand der MARs in beiden Spezies sein könnte. Im Unterschied dazu befindet sich der orthologe Abschnitt der Chromosomenbande 11p15.3 wesentlich weiter proximal in der Subregion F2 der Maus in knapp 35 MB Entfernung zur telomeren Region F5 (http://www.ensembl.org/) (siehe Abb. 42); etwaige Telomer-spezifische Effekte hätten hier keine Auswirkungen mehr. Beobachtungen zeigen, dass eine hohe Dichte der MARs insbesondere in Bereichen zu finden ist, die eine aktive Rolle bei der cis-Regulation von Genen spielen und die mit regulativen Sequenzen wie Enhancer oder Repressoren kolokalisiert sind (Forrester *et al*., 1994). Ebenso scheint eine besonders hohe MAR-Dichte mit der Eigenschaft zur Heterochromatin-Bildung der DNA assoziiert zu sein (Strissel *et al*., 1996). Da die hier gemessenen MAR-Abstände relativ weit auseinander liegen, dürften diese funktionellen, regulativen Zusammenhänge nicht zwingend für diese Region im Vordergrund stehen. Obwohl sie ein interessanter Erklärungsansatz für die regulative Funktion der hoch konservierten Sequenzbereiche rund um die MAR-Positionen ohne genkodierenden Information wären.

Ein weiterer Unterschied der in dieser Arbeit sequenzierten Region zur Chromosomenregion 11p15.5 ist die Tatsache, dass die identifizierten MAR-Regionen des hiesigen Bereiches sich nicht in den transkribierten Bereichen der bekannten Gene befinden. Dagegen sind im untersuchten Abschnitt der Region 11p15.5 MARs in den transkribierten Genbereichen von *CARS, NAP1L4* und *KCN1* zu finden (Engemann *et al*., 2000). Diese unmittelbare Nähe zu exprimierten Genen könnte in Zusammenhang mit der direkten Regulation des jeweiligen Gens stehen, da eine Assoziation des genkodierenden Abschnitts mit der Kernmatrix eine Transkription dieses Gens aufgrund der räumlichen Veränderungen nicht mehr zulässt.

Auch wenn MARs nur relativ diffus zu bestimmen sind, werden ihnen Funktionen zugeschrieben, die weit über die krümmungsgebenden Eigenschaften der DNA und die Anheftung an die Kernmatrix hinausgehen und für die hier untersuchte Region für weitere Untersuchungen von großem Interesse sein könnten. Es gibt Hinweise darauf, dass MARs auch eine wichtige Rolle bei der Verstärkung der

Genexpression spielen (Whitelaw *et al.*, 2000) und als cis-agierende Regulatoren die Chromatinstruktur verändern können und somit indirekt die Transkription von bestimmten Genbereichen verstärken (Jenuwein *et al.*, 1997). Auch gibt es Anzeichen dafür, dass die gekrümmte DNA der MARs am Mechanismus der Transposition beteiligt ist (Yamamura & Nomura, 2001) und die Eigenschaft besitzt, mit dem Histonkomplex H1 zu assoziieren (Davie, 1996). Außerdem scheinen eine besonders hohe Dichte vor allem in Imprinting-Zentren, wie z.B. auf Chromosom 15q11-q13, vorzukommen (Greally *et al.*, 1999) und durch ihre Bindung an DNA-bindende Proteine oder Multiprotein-Komplexen in Prozessen wie Replikation, DNA-Reparatur und Spleißen involviert zu sein (Hibino, 2000; Hancock, 2000). Somit liefern die in dieser Arbeit bestimmten MAR-Positionen einen idealen Ausgangspunkt für weitere Analysen zur Charakterisierung der MARs und zur Klärung des regulativen Potentials der vielen hochkonservierten Sequenzabschnitte in dieser Region.

4.4.3 Konservierte DNA-Sequenzen als mögliche Abschnitte für RNA-Transkripte

Gibt der erhöhte GC-Gehalt einer konservierten Region immerhin noch den Aspekt für eine Promotor-assoziierte Funktion vor oder zeigen MARs Bereiche auf, denen besondere Kontrollaufgaben in der Regulation von transkribierten Bereichen zukommen könnten, so sind streng konservierte Bereiche, die durch keiner dieser Merkmale charakterisiert werden ausser ihrer hohen Homologie zum Vergleichsgenom, ohne weitere experimentelle Untersuchen nur äußerst spekulativ zu beurteilen.

Eine Region mit sehr hoch konservierten Sequenzabschnitten, denen keine funktionellen Aspekte zugeordnet werden konnten, findet sich z. B. im ersten Intron des *LMO1*/*Lmo1*-Gens. Hier konnten sieben Sequenzbereiche angesprochen werden (siehe Tab. 21), deren Sequenzkonservierung zwischen 85% und 99% über mehrere hundert Basenpaare (zw. 180 bp und 724 bp) deutlich über dem ebenfalls konservierten Sequenzumfeld lag (siehe auch PIP-Analyse Abb. 25a). Da in diesen konservierten Abschnitten kein durchgehender offener Leserahmen zu finden war, kann davon ausgegangen werden, dass es sich nicht um proteinkodierende Sequenzbereiche handelt. Möglich wäre aber die kodierende Funktion für ein oder mehrere RNA-Transkripte, die keiner Translation in eine Polypeptidsequenz unterliegen und somit keinen durchgehenden Leserahmen aufweisen müssen. Diese **ncRNAs** („non-coding RNA") sind intensiver Gegenstand der aktuellen Forschung und werden mittlerweile in mehrere Familien je nach ihrer zugewiesenen Funktion unterteilt werden. Mit Hilfe der komparativen Genomanalyse, die momentan die einzige Möglichkeit darstellt, ncRNAs anzusprechen, da Vorhersage-Algorithmen noch keine verlässlichen Daten zur Verfügung stellen, werden immer weitere Verteter dieser Transkriptklasse entdeckt (Eddy, 1999). Vergleicht man die mögliche Verbreitung mit den bereits vorliegenden Ergebnissen für das Hefegenom, das aus insgesamt ca. 6.000 Genen besteht und für das über 700 RNA-kodierende Regionen charakterisiert werden konnten (Goffeau *et al.*, 1996), so dürften im Humangenom die größte Zahl der ncRNAs noch unentdeckt geblieben sein.

Unterteilt werden die ncRNA-kodierenden Regionen je nach Art und Funktion ihres Wirkungsbereiches. Am bekanntesten sind die **rRNA**- (ribosomale RNA: z.B. 16S, 25S, 5,8S) , die **tRNA**- (transfer-RNA) und die **snRNA**-kodierenden Regionen (small nuclear RNA: z.B. U1, U2, U4, U5 und U6), die die Komponenten des Spleißosoms bilden. Hierzu kommt eine Reihe von ncRNAs mit unterschiedlichsten Aufgaben, wie katalytische RNAs und „signal recognition particle"-RNAs, die sich an der Translokation von Proteinen über das E R beteiligen (Bovia & Strub, 1996). Ein anderer sehr wichtiger Funktionsbereich der ncRNAs ist die Aufgabe der X-Dosis-Kompensierung. Hier ist z.B. das *Xist*-Gen zu erwähnen, welches das inaktivierte X-Chromosom der Säuger schützt und im inaktiven Zustand hält. *Xist* selbst wird dabei über das Antisense-ncRNA *Tsix* reguliert (Panning *et al.*, 1998). Weitere Antisense-ncRNAs agieren als „Riboregulatoren" und steuern auf der Ebene der Translationsinitiation, in dem sie das Startcodon AUG durch Anlagerung in diesem Bereich blockieren. Eine andere große Gruppe der ncRNA sind die **snoRNAs** (small nucleolar RNA), die im Nukleolus gewebsspezifisch an posttranskriptionellen Prozessen und an der Modifikation der ribosomalen RNA involviert sind und zu Hunderten im Genom vorkommen sollen (Tollervey & Kiss, 1997, Hüttenhofer *et al.*, 2001). Bezugnehmend auf die Funktion wird die Population der snoRNAs nochmals in „C/D box antisense snoRNAs", deren Aufgabe unter anderem die Steuerung der RNA-Ribose-Methylierung ist, und in „H/ACA snoRNAs" unterteilt, die die spezifische Pseudouridylierung von ribosomalen RNAs steuern.

Als interessant erweisen sich diese snoRNAs auch im Zusammenhang mit dieser Arbeit, da die meisten bekannten snoRNAs bisher in Intronsequenzen der prä-mRNAs von Haushaltsgenen gefunden wurden (Weinstein & Seitz, 1999). Ein Aspekt, der von den eingangs beschriebenen konservierten Sequenzabschnitten im ersten Intron des *LMO1*/*Lmo1*-Gens erfüllt werden würde. So beinhaltet beispielsweise das humane *gas5*-Gen insgesamt 10 snoRNAs und das humane *UHG* acht snoRNAs. Diese relativ kurzen zwischen 100 und 300 bp umfassenden Gene sind aber keinesfalls repräsentativ für die durchschnittliche Länge aller ncRNAs. Es gibt ebenso ncRNA, die mehrere tausend Basenpaare umfassen, wie z.B. das humane *H19* mit 2.313 bp oder das humane *Xist* mit 16,5 kb und vermutlich als stabile Transkripte in der Zelle vorliegen (Erdmann *et al.*, 1999). Diese langen ncRNAs werden nach bisherigem Wissensstand ebenso wie ihre proteinkodierenden Pendants gespleißt, und besitzen eine 5´- und 3´-UTR mit einem polyadenylierten Ende. Ein solches Poly-A-Signal konnte auch für den ersten konservierten Bereich innerhalb des *LMO1*/*Lmo1*-Introns für den Abschnitt nt 115.400 bis nt 115.405 mit Hilfe der GENSCAN-Analyse detektiert werden, so dass es sich hierbei durchaus um einen 3´-Bereich einer unbekannten snoRNA handeln könnte.

Die Entdeckung von neuen ncRNA-Transkripten ist vor allem daher so wichtig, weil ncRNAs eigene biochemische Aufgaben erfüllen. Sie können sowohl mit der DNA wie auch mit RNA-bindenden Proteinen interagieren und Ribonucleoprotein-Komplexe (RNPs) bilden. Solche RNPs wurden z. B. in diversen zellulären Kompartimenten, wie dem Nukleolus, bei der Steuerung der dendritischen Translation beschrieben (Tiedge *et al.*, 1993). Hier kommt es an postsynaptischen Orten zu einer lokalen Translation von dentritischer mRNA, die bei der Modulation von synaptischen Strukturen, wie sie z.B. bei Prozessen des Lernens und des Gedächtnisses eine Rolle spielen, beteiligt sind (Tiedge &

Brosius, 1996). Sie können als Gen-Regulatoren, wie „Silencer", oder auch als abiotische (z. B. *adapt33*, *hsr*), wie biotische Stress-Signale (z. B. *His-1*, *ENOD20*) fungieren (Erdmann *et al.*, 1999). Diese subtile Rolle in anscheinend vielen wichtigen Regulationsmechanismen machen die ncRNAs zu vielversprechenden Kandidaten für verschiedene humane Erkrankungen, denen eine Dysregulation zugrunde liegt. So wird z. B. das humane Analogon der murinen BC200-RNA unter bestimmten Bedingungen nur in verschiedenen Karzinomen wie Brust-, Gebärmutterhals-, Speiseröhren- oder Lungenkrebs exprimiert. Im benachbarten gesunden Gewebe lässt sich keine Expression der analogen BC200-RNA nachweisen (Chen *et al.*, 1997). Verschiedene gehirnspezifische snmRNAs („small non-messenger RNA"), ein weiterer Begriff für die ncRNAs, werden z. B. auf Chromosom 15q11-13 exprimiert, in einer Region, die mit dem Prader-Willi-Syndrom (PWS) in Verbindung steht (Cavaille *et al.*, 2000). Somit scheinen die RNA-Transkripte eine immer größere Rolle für medizinische Krankheitbilder zu bekommen. Vor allem für Syndrome, deren genetische Kopplung zu einem genomischen Abschnitt zwar bekannt ist, wo aber noch keine entsprechenden Krankheitsgene identifiziert und charakterisiert werden konnten. Dieser Zusammenhang wäre auch ein Ausblick für die hiesige BWSCR3-Region, für die bisher noch keine genetische Veränderung als hinreichende Ursache für die Ausbildung des komplexen Krankheitsbildes des BWS-Syndroms verantwortlich gemacht werden konnte (siehe auch Kap. 4.6.1).

4.4.4 Konservierte DNA-Sequenzen als mögliche Promotoren oder Enhancer

Einen wichtigen Beitrag liefert die komparative Sequenzanalyse auch für die Identifizierung von genregulatorischen Sequenzeinheiten wie Promotor- oder Enhancer-Sequenzen. Da experimentelle Untersuchungen zur Verifizierung von möglichen regulativen Sequenzabschnitten äußerst aufwendig sind, wird versucht, mithilfe von Computer-Algorithmen diese **Promotor-Elemente** vorherzusagen. Es hat sich gezeigt, dass diese Analyse insbesondere ohne weitere Aspekte wie benachbarte exprimierte Sequenzbereiche sehr schwierig ist und dass durchschnittlich eine zu große Zahl an möglichen Promotor-Bereichen errechnet wird. So identifizierte die PROSCAN-Analyse für die 319 kb humangenomische Sequenz insgesamt 38 mögliche RNA-Polymerase-II-Promotoren, was einem Durchschnitt von einer Promotor-Sequenz alle 8,4 kb entspricht. Ein nicht ganz so hohen Wert wurde mit 12,5 kb pro Promotor für die 412 kb der murinen Genomsequenz ermittelt. Literaturangaben zufolge liegt der Durchschnitt bei einer Promotor-Einheit alle 30 bis 40 kb (Antequera & Bird, 1993). Somit dürfte diese Zahl nur sehr bedingt der Wirklichkeit entsprechen. Anscheinend reellere Ergebnisse lieferte die GENSCAN-Promotor-Analyse, die in ihrem Vorhersage-Algorithmus auch die vorhergesagten Exonsequenzen mit einbezieht. Dadurch ließen sich Werte von 26,6 kb pro Promotor-Sequenz im Menschen, bzw. von 45 kb für die Maus-genomische Sequenz ermitteln. Die Zahl der verifizierten Gentranskripte für den untersuchten humanen Sequenzabschnitt lag aber nur bei zwei Genen (*LMO1* und *TUB*), bzw. bei drei Genen (*Lmo1*, *Tub* und *Eif3*) in der Maus-Sequenz. Es ist deshalb davon auszugehen, dass es sich bei einem Teil der putativen Promotoren um kryptische Initiationsstellen handeln dürfte, selbst wenn in diesem Bereich noch unbekannte Transkripte

lokalisiert sein sollten, die *in vivo* ohne funktionelle Bedeutung sind und daher als Falsch-Positive gedeutet werden können. Berücksichtigt werden müssen hierbei allerdings die zusätzlichen Promotoren der alternativen Spleißvarianten. So besitzt beispielsweise sowohl das Exon 1a wie auch das Exon 1b des *LMO1*/*Lmo1*-Gens jeweils seinen eigenen Promotor (McGuire *et al.*, 1989). Die genaue Betrachtung der PIP-Analyse, in der die Ergebnisse der GENSCAN-Promotor-Vorhersage grafisch eingezeichnet wurden, zeigte für das *LMO1*-Gen, dass keine der beiden Promotoren durch das Programm vorhergesagt wurde (siehe Abb. 25a). Dies legt nahe, dass durch die verwendeten Algorithmen nicht nur zu viele falsche Promotoren vorhergesagt wurden, sondern dass einige Promotorbereiche erst gar nicht angezeigt wurden. Eine Schwierigkeit der computergestützten Vorhersage ist in diesem Zusammenhang, dass Promotoren ausschließlich stromaufwärts – „upstream" zur Initiationsstelle eines Gens erwartet werden. Der klassische Promotor besteht dabei aus einer etwa 30 bp stromaufwärts gelegenen TATA-Box mit der Consensussequenz „TATAAA" (Hahn *et al.*, 1989), so wie bei dem in dieser Arbeit neu entdeckten murinen Gen *Eif3* gezeigt werden konnte (siehe Abb. 21, Kap. 3.7.1). Es existieren aber auch TATA-Box-freie Promotoren, deren exakte Position der Transkriptionsinitiation durch ein anderes Basis-Element dem „Initiator" (Inr) kontrolliert wird (Smale, 1997). Zudem gibt es Promotor-Elemente, die stromabwärts – „downstream" lokalisiert sein können. So konnte z. B. ein in Mensch und *Drosophila* konserviertes sieben Nukleotide umfassendes sog. „Downstream"-Promotor-Element (DPE) 30 bp stromabwärts vom Transkriptionsstartpunkt als TATA-Box analoges Element beschrieben werden (Burke & Kadonaga, 1997).

Eine ganz andere Einschränkung der Promotor-Vorhersageprogramme ist die Tatsache, dass nur die putativen RNA-Polymerase-II-Bindungsstellen der proteinkodierenden Transkripte berechnet werden und dass die Promotoren der RNA-Polymerase-I, welche vorzugsweise rRNA-kodierende Gene transkribiert und Promotoren der RNA-Polymerase-III, die für die tRNA- und andere snmRNA-Transkription zuständig sind, nicht erfassen (Huet *et al.*, 1982). So können zusätzliche Indizien für den 5´-Bereich von z. B. unbekannten ncRNA-Transkripten mithilfe der Vorhersage-Algorithmen nicht erbracht werden. Die Analyse der vorhergesagten Promotor-Bereiche zeigte, dass ohne den komparativen Genomvergleich und der dadurch gewonnenen Information über die Konserviertheit eines bestimmten Genomabschnitts die alleinigen Ergebnisse der Promotor-Vorhersage keinen ausreichenden Hinweise für ein weiteres experimentelles Vorgehen auf der Suche nach neuen Transkripten geben.

Eine weitere Ursache für die schlechte Vorhersagbarkeit der Promotoren, dürfte auch in der Art und Weise der Sequenzanalyse liegen, wie sie mit Hilfe der Computer durchgeführt wird. Grundlage für die Promotorvorhersage ist immer die Einzelstrangsequenz des zu untersuchenden Abschnitts. Übergeordnete Strukturen des Chromatins, wie die räumliche Konformation der DNA innerhalb des Zellkerns, werden nicht berücksichtigt, sind aber *in vivo* wichtige Kriterien für die Erkennung der Transkriptionsinitiationsstellen innerhalb der Zelle. Auch hier scheint der komparative Ansatz einen entscheidenden Vorteil zu bringen.

Für die Transkriptionsaktivierung ist nicht nur der „Core"-Promotor als Bindungsstelle für die RNA-Polymerase und die Transkriptionsfaktoren wichtig. Es hat sich gezeigt, dass für die effiziente Transkription *in vivo* zusätzlich noch kurze regulatorische Elemente anwesend sein müssen (Gottesfeld *et al.*, 1997). Diese **proximalen** oder **Enhancer-Elemente** können in verschiedenen Distanzen von einigen Duzend Basen bis zu mehreren Kilobasen Entfernung stromauf- oder abwärts zum Transkriptionsstartpunkt liegen (Fassler & Gussin, 1996). Beide Typen von Enhancer stellen Bindestellen für Proteine – insbesondere Transkriptionsfaktoren – dar, die den Grad der Transkription des „Core"-Promotors erhöhen oder auch erniedrigen und so zur Transkriptionssteuerung beitragen. Da es Tausende von verschiedenen Transkriptionsfaktoren gibt – es wird vermutet, dass es einige Prozent aller im Genom kodierter Proteine sind, die mit den regulativen Elementen interagieren können – stellt sich die Frage, ob diese Bindestellen im Laufe der Evolution der verschiedenen Spezies erhalten geblieben sind, oder ob jede Art ihr ganz persönliches Arsenal an regulatorischen Elementen entwickelt hat.

Die Antwort auf diese Frage dürfte eher das erste Szenario der Konservierung sein, da in dieser Arbeit ausgeprägte Konservierungen der flankierenden, nicht-translatierten Randbereiche bei einigen Exons bestimmt werden konnten, die nicht nur im Vergleich zwischen Mensch und Maus, sondern auch bei der komparativen Analyse zwischen Mensch und *Fugu* auffielen. So belegten die Ergebnisse dieser Arbeit, dass z. B. der 5´-UTR des LMO1-Exons 1a und der 5´-Intronbereich des LMO1-Exons 2 in allen drei Spezies konserviert geblieben ist. Weitere Beispiele sind die Sequenzbereiche der Introns 6, 10 und 11 des *TUB*-Gens (vgl. Abb. 25 + 27). Auch diese Abschnitte blieben in allen drei Spezies relativ stabil. Ähnliche Ergebnisse wurden für einen Abschnitt aus dem humanen *HOXA*-Clusters beschrieben, dort konnten insbesondere im 5´-UTR Bereich der Gene *HOXA9* und *HOXA7*, konservierte nicht-kodierende Bereiche (CNS = „conserved non-coding sequences") zwischen Mensch, Maus und *Fugu* identifiziert werden (Hardison, 2000). Dass diese CNS eine sehr wichtige und weitreichende Funktion zu haben scheinen, konnte ebenfalls für die Cytokin-Gene *IL-4*, *IL-13* und *IL-5* auf Chromosom 5q31 belegt werden. Hier führte die experimentell herbeigeführte Deletion nur eines von 6 konservierten nichtkodierenden Bereiche mit einer Länge von 400 bp zu einer zwei- bis dreifachen Expressions-reduzierung aller drei Interleukin-Gene, die sich gemeinsam über einen Bereich von 120 kb erstrecken und unter dem Einfluss dieses entfernt lokalisierten CNS-Bereichs standen (Loots *et al.*, 2000). Dies zeigt, dass CNSs ihren Effekt auf weite Genombereiche ausüben können und nicht nur in unmittelbarer Nähe zu ihrem Zielgen liegen müssen. Somit können Mutationen oder Deletionen auch in CNSs eine mögliche Ursache und Auslöser von Krankheitsbilder z. B. in Fällen sein, deren betreffende Kandidatengene bisher keinerlei genetische Veränderungen aufwiesen. Interessant ist in diesem Zusammenhang die Größe der CNS-Bereiche von durchschnittlich ca. 400 bp. Auch die meisten konservierten Sequenzabschnitte zwischen Mensch und *Fugu*, die außerhalb der bekannten Genbereiche dieser Arbeit lagen, weisen Längen zwischen 320 und 440 bp auf (vgl. Tab. 20, Kap. 3.8.2: die konservierten Abschnitte #1, #4, #5, #10 und #11). Sollte diese Länge charakteristisch sein und einen repräsentativen Charakter für Intergen-CNSs haben, dann wäre diese Analogie ein hilfreiches Indiz für die wichtige funktionelle Bedeutung dieser konservierten Sequenzabschnitte.

Ein weiterer Aspekt für die mögliche Funktion der CNS wäre ihre Eigenschaft als konservierte Proteinbindesequenzen für bestimmte regulatorische Elemente, ähnlich den Bindestellen für den CCCTC-Bindungsfaktor (CTCF), der spezifisch an sog. Isolator- („insulator") und „Silencer"-Sequenzen bindet (Ohlsson *et al.*, 2001). So wurden für die Chromosomenregion 11p15.5 mehrere dieser CTS („CTCF-target-sequence") für die Gene *H19*, *IGF2* und *KCNQ1* als negative Regulatoren mit Enhancer-Blockierungsaktivität beschrieben (Du *et la.*, 2003). Da diese speziellen Elemente auch im Zusammenhang mit der Entstehung von familiären Erkrankungen und von Krebs diskutiert werden (Zhou *et al.*, 2004), wäre auch diese Analogie ein lohneswerter Ansatzpunkt für die weitergehende Analyse nach etwaigen Regulationsfunktionen dieser konservierten Bereiche.

4.5 Vergleichende Analyse unterschiedlicher Chromosomenregionen

4.5.1 Der syntäne Gesamtbereich der Chromosomenregion 11p15.3 des Menschen mit dem Chromosom 7 der Maus

Mit Hilfe der in dieser Arbeit sequenzierten und charakterisierten 319 kb an humangenomischer Sequenz konnte ein zuvor unbekannter Genomabschnitt in der chromosomalen Bande 11p15.3 charakterisiert werden. Gleichzeitig fügte sich dieser Bereich auch an das distale Ende eines chromosomalen Abschnitts, der im Rahmen der Promotionsarbeiten von Cichutek (2001), Amid (2002), Bahr (2000) und Mujica (2002) sequenziert worden war. Dadurch war es möglich einen Contig zu erstellen, der sich insgesamt über einen Bereich von ungefähr 1,65 Mb an humangenomischer DNA erstreckte. Es konnte somit ein lückenloser Contig konstruiert werden, der sich minimal aus insgesamt 17 Klonen (13 PACs, 3 BACs, 1 Cosmid) zusammensetzt. Da die genomische Sequenz zwischen den sequenzierten Bereichen um die Gene *RANBP* und *CEGP1*, repräsentiert durch die drei Klone PAC151D7, PAC79B22 und BAC299K14, bis zum Ende der Arbeit nicht vorlag, konnte die Größe dieses Bereiches nur anhand der PFGE-bestimmten Integratlängen abgeschätzt werden (Cichutek & Brückmann *et al.*, 2001).

Für die syntäne Region auf Chromosom 7 der Maus konnten durch die vorliegende Arbeit insgesamt 412 kb an neuer DNA-Sequenz zu einem Contig hinzugefügt werden, der in der Subregion F2 auf Maus-Chromosoms 7 insgesamt einen Bereich von ca. 1,46 Mb umfasst. Durch die hier vorliegenden Ergebnisse und denen der vier oben erwähnten Promotionsarbeiten war es nunmehr möglich, auch den orthologen Bereich der murinen Genomsequenz in einem Minimalcontig aus insgesamt 11 Klonen (5 PACs, 4 BACs und 2 P1-Klone) unter Verbleib von zwei Lücken darzustellen. Die DNA-Sequenz in der Contiglücke zwischen den Klonen BAC377M11 und PAC221D7 wurde durch eine Primerwalking-Sequenzierstrategie geschlossen (Alejo Mujica, unveröffentlicht). Die zweite Lücke mit geschätzten 30 kb Umfang blieb aufgrund fehlender proteinkodierender DNA unsequenziert (Amid & Bahr *et al.*, 2001). Auch hier wurde der nicht sequenzierte Bereich der beiden Klone PAC626J16 und PAC119C16 zwischen den Genen *RanBP* und *Cegp1* näherungsweise über den Interspeziesvergleich geschätzt. Eine grafische Gegenüberstellung beider charakterisierter Genombereiche mit Contigkarte sämtlicher informativer Klone und Gene ist in nachfolgender Abbildung 41 zusammengestellt.

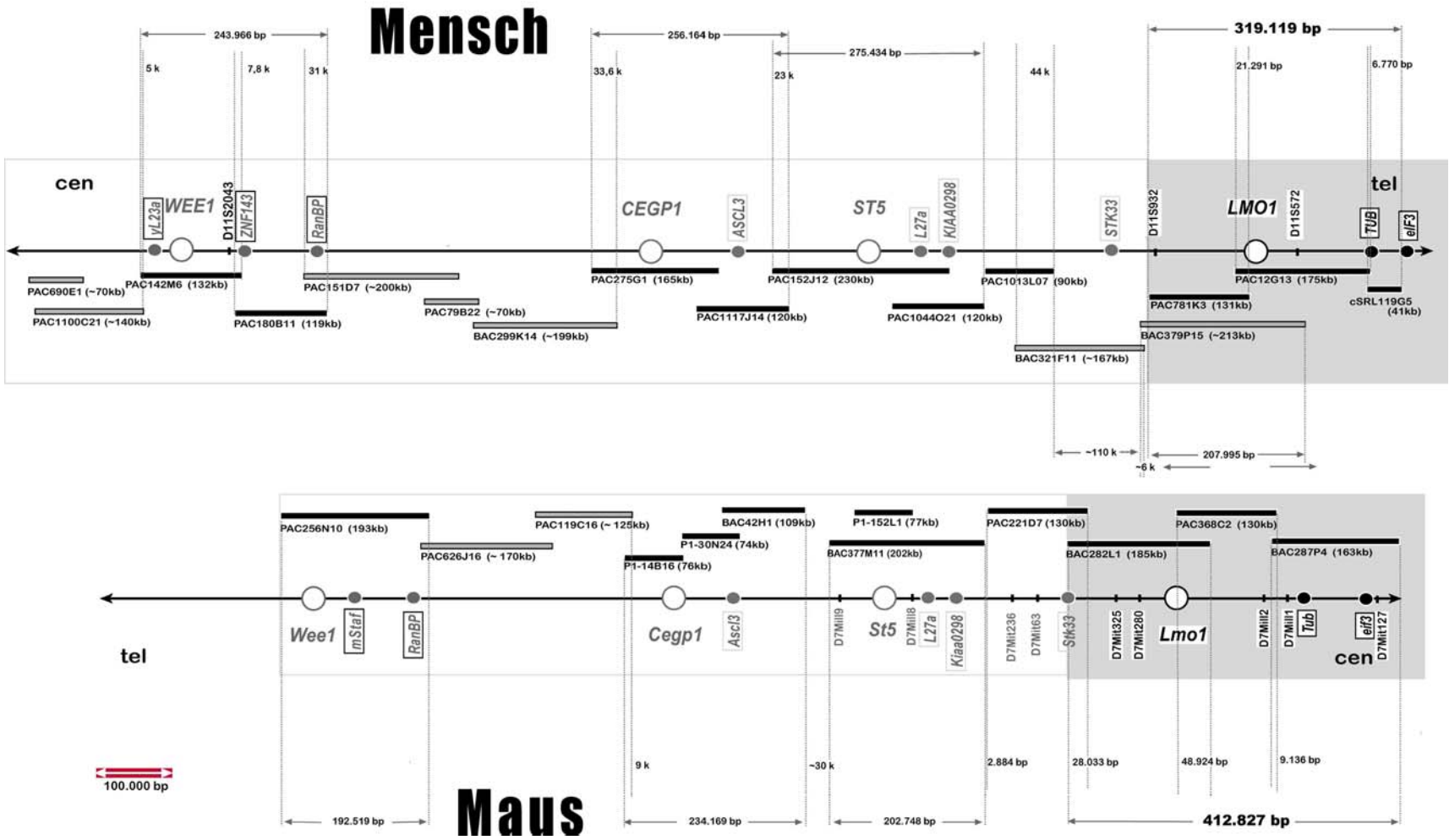


Abb. 48 **Contigkarte der syntänen Region zwischen den Genen *WEE1/ Wee1* und *TUB/ Tub*.** Die Zusammenstellung sämtlicher Daten aus Amid & Bahr et al. (2001) und Cichutek & Brückmann et al. (2001) führten zu obiger Contigkarte. Es konnte für die humangenomische Sequenz ein durchgehender Contig generiert werden, der über 1,65 Mb genomische Sequenz repräsentiert. Der Contig für den syntänen Maus-genomischen Bereich über 1,46 Mb weist zwei Lücken auf. Die Contiglücke zwischen den Klonen BAC377M11 und PAC221D7 konnte durch eine Primerwalking-Sequenzierstrategie geschlossen werden (Alejo Mujica, unveröffentlicht), die zweite Lücke mit geschätzten 30 kb Umfang blieb unsequenziert (Amid & Bahr et al., 2001). Insgesamt konnten in dieser Region in der humanen Sequenz 17 Gene und für die murine Sequenz 15 Gene kartiert werden.

Die vergleichende Analyse im erweiterten Betrachtungsfokus der insgesamt 1,65 Mb umfassenden Genomsequenz zeigte, dass sich sowohl der hohe Grad der Konservierung wie auch die Syntänie der Gene, nicht nur in dem in dieser Arbeit sequenzierten Bereich evolutiv erhalten haben, sondern sich über die gesamte Ausdehnung der 1,65 Mb erstrecken. Auch die relative Verkürzung der Maus-genomischen DNA im Vergleich zum Menschen um ca. 14% im Abschnitt der Gene *Lmo1* und *Tub*, konnte für die Gesamtregion bestätigt werden und entspricht somit genau dem Mittelwert, der auch für das Verhältnis des gesamten euchromatischen Genomanteils zwischen Mensch und Maus errechnet wurde (Waterston *et al.*, 2002). Mit einer Gendichte von 10,3 identifizierten Genen pro Megabase in beiden Genomen rangiert die untersuchte Region etwas unter dem errechneten Mittelwert für das gesamte Chromosom 11, der bei 13,45 Genen pro Megabase angegeben wird. (1.937 bekannte Gene auf 144 Mb genomischer DNA; siehe Kap. 1.4; http://www.ncbi.nlm.nih.gov/genome/guide/ HsChr11.shtml)

Dass dieser 1,65 Mb große Abschnitt nicht die einzige syntäne Region zwischen Mensch-Chromosom 11 und Maus-Chromosom 7 ist, zeigt die orthologe Region der chromosomalen Bande 11p15.5, die im chromosomalen Abschnitt 7 F5 der Maus ihr Pendant findet (Engemann *et al.*, 2000; Onyango *et al.*, 2000). Auch in diesem Abschnitt ist die strukturelle Organisation in beiden Spezies nahezu gleich geblieben und der Interspeziesvergleich zeigt viele Parallelen zur hier untersuchten Region, so dass bereits in den vorangegangenen Kapitel mehrfach auf diese Region verwiesen wurde (siehe Kap. 4.4.1, 4.4.2, 4.5.2). Insgesamt können für das humane Chromosom 11 fünf (http://www.sanger.ac.uk/Projects/M_musculus/publications/fpcmap-2002/syndata/hm.11.7.html#), bzw. sechs syntäne Bereiche (http://www.ensembl.org/Homo_sapiens/syntenyview?species=Mus_musculus&chr=11&loc=10496168) auf Maus-Chromosom 7 in Homologie gesetzt werden. Der Unterschied resultiert dabei aus der zusätzlichen Diskriminierung der syntänen Region auf Chromosomenbande 11p15.1 in zwei Teilregionen (http://www.ensembl.org) zu den Maus-Chromosomenbanden 7 F4 und 7 F5. Insgesamt findet sich fast der gesamte distale Bereich des kurzen Arms von Chromosom 11, mit dem Chromosomenabschnitt der Bandenregion 11p15 auf dem murinen Chromosom 7 wieder, allerdings unterteilt in drei verschieden Abschnitten (B4, F2 und F5). Als zweiten großen Abschnitt sind die Banden 11q13.3 bis 11q14.3 zu nennen, die ihre orthologen Bereiche in den murinen Banden E1 bis F1 und F5 haben (Details siehe Abb. 48, bzw. Tab. 25). Andere syntäne Bereichsabschnitte des humanen Chromosoms 11 zum Mausgenom finden sich zu den murinen Chromosomen 2, 19 und 9.

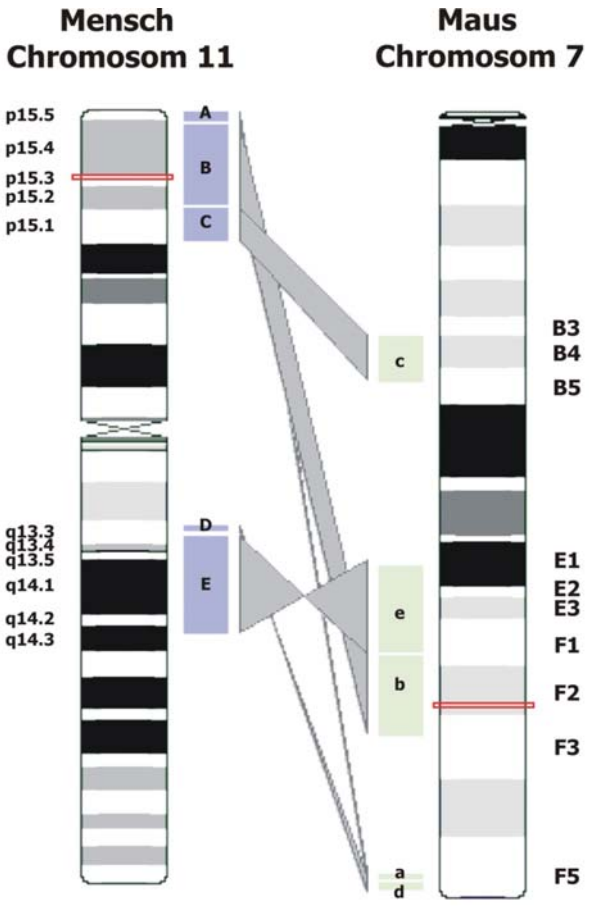


Abb. 49 Vergleich der bekannten syntänen Bereiche zwischen dem humanen Chromosom 11 und dem murinen Chromosom 7. Insgesamt konnten aufgrund der vorliegenden Sequenzinformationen aus beiden Genomprojekten fünf orthologe Bereiche (A bis F) identifiziert werden, innerhalb derer die Architektur der Abfolge der kodierenden Gensequenzen in beiden Spezies erhalten geblieben ist. Als Vorlage diente eine modifizierte Grafik unter: http://www.sanger.ac.uk/Projects/M_musculus/ Am Beispiel des kurzen Arms von Chromosom 11 zeigt sich, dass nahezu die gesamte Region 11p15 auf Chromosom 7 der Maus, allerdings in unterschiedlichen Regionen, wiederzufinden ist. Ein zweites Syntänie-Cluster auf Chromosom 11 ist zwischen den Banden 11q13.3 und 11q14.3 lokalisiert. Die orthologen Bereiche der Maus befinden sich dazu in der Regionen E1 bis F1 und in der Telomer-ständigen Bande F5. Im Vergleich zum humanen Chromosom sind die Syntäniebereiche D und E auf dem murinen Chromosom invertiert abgebildet. Rot eingezeichnet ist die in dieser Arbeit charakterisierte Region 11p15.3 mit ihrem Syntäniebereich in der murinen Bande F2.

Syntänie-bereich	Region auf Chromosom 11	Region auf Chromosom 7
A/a	0 Mb – 2,4 Mb	142,2 Mb – 143,7 Mb
B/b	2,5 Mb – 18 Mb	101,5 Mb – 116,9 Mb
C/c	18,1 Mb – 25 Mb	41,7 Mb – 51,1 Mb
D/d	77,4 Mb – 79,1 Mb	145,9 Mb – 143,8 Mb
E/e	79,2 Mb – 98,1 Mb	101,4 Mb – 84,7 Mb

Tab. 29 Auflistung der verschiedenen Syntäniebereiche aus obiger Grafik unter Angabe ihrer physikalischen Position innerhalb der chromosomalen Referenz-sequenz (http://www.sanger.ac.uk).

Der rückläufigen DNA-Sequenzfolge der Syntäniebereiche *d* und *e* ist zu entnehmen, dass diese Abschnitte auf dem Maus-Chromosom 7 in invertierter Orientierung abgebildet sind.

4.5.2 Vergleich mit Chromosom 11 paralogen Regionen

Die Analyse der Homologien, insbesondere die Auswertung der Blast-Analyse der kodierenden Abschnitte des neu entdeckten murinen Translationsinitiationsfaktors *Eif3* zeigte die Assoziation zur chromosomalen Region 12p13, für die ein putatives Pseudogen zum *eIF3S5* identifiziert werden

konnte (siehe Kap. 4.2.4). Dass dieser Bezug zum Chromosom 12 des Menschen kein Einzelergebnis zu sein scheint, belegt die Tatsache, dass sich auch verwandte Gene von *LMO1* und *TUB* auf Chromosom 12 wiederfinden. So wurde das *LMO3*-Gen – ein Mitglied der LIM-Genfamilie wie das Gen *LMO1* – ebenfalls in die Chromosomenregion 12p13 lokalisiert (Boehm *et al.*, 1991). Auch für das *TULP3*-Gen – ein Mitglied der Tubby-Genfamilie – ist als chromosomaler Locus die Bande 12p13 bekannt. Die Liste der paralogen Gene ließe sich diesbezüglich noch weiter fortführen. Auch dem Gen *PTH*, welches weiter proximal auf der chromosomalen Bande 11p15.3 lokalisiert ist, konnte das Gen *PTHLH* in der Chromosomenregion 12p12.1-p11.2 gegenübergestellt werden (Mangin *et al.*, 1989). Interessant scheint es, dass sich diese Verwandtschaft zwischen den Chromosomen 11 und 12 nicht nur auf die beiden Banden 11p15.3 und 12p13-p12 beschränkt, sondern dass eine ähnliche Assoziation auch zwischen der distalen Chromosomregion 11p15.5 und dem Abschnitt der chromosomalen Region 12q21-24 besteht (Prawitt, 1999) Als Beispiele können hier die Gene *IGF2*, *ASCL2*, *ORCTL2* und *NAP1L4* des Lokus 11p15.5 und die zu ihnen paralogen Pendants *IGF1*, *ASCL1*, *PHCA1* und *hNRP* auf Chromosom 12q genannt werden. Evolutiv betrachtet, könnte diese Verwandtschaft der beiden Chromosomen durch eine frühe Chromosomenduplikation erklärt werden, als sich dieser Bereich aus einem gemeinsamen Vorläuferchromosom duplizierte und nach einigen sekundären Genarrangements mit Inversionen und Translokationen, in den heutigen Zustand entwickelte. Für das relativ frühe Eintreten eines solchen Ereignisses in der Evolution spricht ein ähnlicher Zustand im Genom der Maus, der sehr ähnliche Homologieverhältnisse zwischen den Chromosomen 7 und 6 aufzeigt. Darüber hinaus besitzt das murine Chromosom 6 ähnlich große syntäne Bereiche mit dem humanen Chromosom 12, wie Mauschromosom 7 mit dem humanen Chromosom 11. Auch bei Chromosom 12 findet sich fast der gesamte kurze Arm in den paralogen Regionen F1 bis G3 auf Mauschromosom 6 wieder (siehe Abb. 49) (http://www.ensembl.org/Mus_musculus/syntenyview).

Einer solchen verwandtschaftlichen Beziehung könnte eine interchromosomale Duplikation eines großen genomischen Bereiches zugrunde liegen, wie sie bereits für die beiden Chromosomenregionen 6p21.3 und 9q33-34 und ihre orthologen Regionen auf den murinen Chromosomen 2 und 17 beschrieben wurde (Kasahara *et al.*, 1996). Hier kam es zu einer Verdopplung von Genen, die für verschiedene Untereinheiten der Proteosomen-bildenden Proteasen kodieren und für die Prozessierung von Peptiden zuständig sind, die über MHC-Klasse I Komplexe (MHC = „major-histocompatibility complex") bestimmter Immunzellen präsentiert werden sollen; ein Vorgang, dem eine sehr große Bedeutung bei Immunreaktionen zukommt.

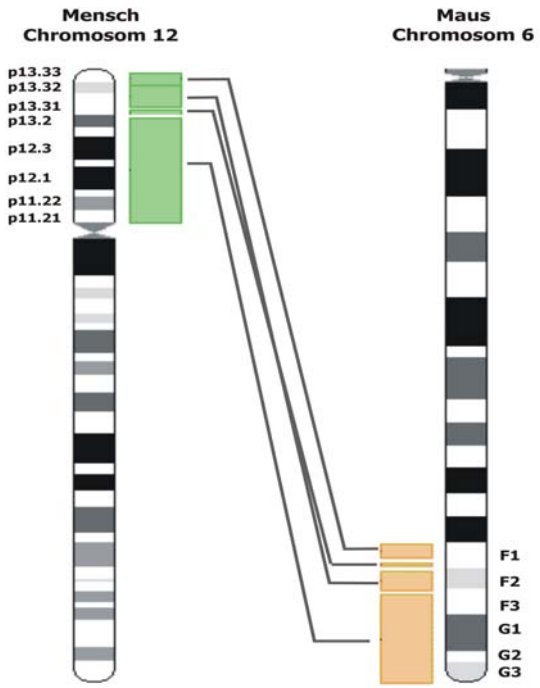


Abb. 50 Syntänie-Verhältnisse zwischen dem humanen Chromosom 12 und dem murinen Chromosom 6. Modifiziert nach den Vorgaben von <http://www.ensembl.org/Mus_musculus/syntenyview>. Die genetische Information des kurzen Arms von Chromosom 12 findet sich größtenteils in der orthologen Abschnitten F1 bis G3 auf Mauschromosom 6 wieder.

Die orthologen Gene dieser beiden Abschnitte konnten in der Maus innerhalb der entsprechenden paralogen Abschnitte auf Chromosom 2 und 17 gefunden werden. Zur evolutiven Entstehung dieser Aufteilung wurde eine Duplikation vorgeschlagen, die sich schon vor der Trennung der beiden Stammlinien zwischen Primaten und Nagern ereignet haben müsste. Endo & Mitarbeiter (1997) griffen diese Hypothese auf und formulierten für diesen Bereich, bestehend aus 12 paralogen Genpaaren, zwei unabhängige intrachromosomale Duplikationsereignisse, die während ihrer Evolution weiteren chromosomalen Rearrangements unterlagen, wie regionale Duplikationen, Translokationen und Inversionen. Es konnte für diese Region gezeigt werde, dass die relative Anordnung der Gene auf den paralogen Abschnitten invertiert oder nur noch in Teilabschnitten konserviert vorliegt.

Dass großräumige Umstrukturierungen innerhalb eines Genoms während der Speziesentwicklung keine Ausnahmeerscheinungen sind, zeigen auch andere Organismen wie z. B. das Genom von *Arabidopsis thaliana*. Für die Evolution dieses Organismus wird aufgrund der heutigen Genomstruktur von einer anfänglichen Genomduplikation ausgegangen, der einerseits anscheinend ein massiver Genverlust, aber andererseits eine lokale Genduplikation folgte (Blanc *et al.*, 2000). Weite Bereiche des fünf Chromosomen umfassenden Genoms weisen zueinander eine hohe Homologie auf, in denen Zahl, Reihenfolge und Orientierung der Gene konserviert geblieben sind.

Welche Auswirkungen solche Genduplikationen innerhalb eines Genom für den Organismus haben, bzw. welche evolutiven Vorteile daraus erwachsen können, fasste Sidow (1996) zusammen. Aufgrund

der vorhanden Redundanz an proteinkodierender Erbinformation innerhalb der Gengruppen solcher paraloger Chromosomenabschnitte, unterliegt der duplizierte Bereich einem geringen Selektionsdruck, da die kodierende Sequenz des ursprünglichen Bereichs weiterhin die Genfunktion unverändert erfüllen kann. Somit ereignen sich in den duplizierten Bereichen relativ frei Mutationen, die, wenn sie keinen nachteiligen Effekt für den Organismus darstellen, das Potential für neue Genfunktionen erschließen können. Ein anderer wichtiger Aspekt, der durch die Verdopplung der genkodierenden Abschnitte auftritt, ist die Redundanz der Genprodukte und die dadurch notwendig werdende Regulation der Gendosis. Nach Sidow (1996) kann die Evolution diesem Umstand in zwei unterschiedlichen Weisen begegnen. Entweder es ereignet sich eine Sequenzveränderung im duplizierten Gen, die zu einen Genprodukt mit modifizierter Funktion führt, oder es kommt zu Veränderungen in der Regulation des duplizierten Gens, z. B. in einem cis-regulierenden Bereich, der eine Modifikation in der Expression des duplizierten Gens zur Folge hat. Als cis-regulierende Bereiche können Transkriptionsbindungsstellen, Modifikatorbindungsstellen oder auch Sequenzbereiche angesehen werden, die die monoallelische Expression bedingen, wie z. B. Imprinting-Zentren. Da die paralogen Bereiche aufgrund ihrer hohen Sequenzhomologie bevorzugte Regionen für Rekombinationsvorgänge sind, ereignen sich in diesen Abschnitten meist sekundär weitere Mikrodeletionen, Mikroduplikationen oder Inversionen, die zu genomischen Rearrangement-Ereignissen führen (Eichler, 2001; Reiter *et al.*, 1998). Diese Veränderungen können als ein dynamischer Prozess angesehen werden, der mit einer Häufigkeit von einer auf 1.000 Geburten beobachtet wird und der oftmals klinische Relevanz bekommt, da viele genetische Erkrankungen z. B. mit segmentalen Duplikationen assoziiert sind (Lupski *et al.*, 1998; Ji *et al.*, 2000). Ein Beispiel wäre das Charcot-Marie-Tooth-Syndrom 1A (CMT1A), eine Form der neuronalen Muskelatrophie, die auf einer Tandemduplikation einer 1,5 Mb großen Region auf Chromosom 17p12 herrührt (Chance & Fischbeck, 1994). Darüber hinaus bewirkt der Mechanismus der Rearrangements, dass sich die paralogen Gene und damit die gesamte Region durch die Evolution genomisch verändern. Als Merkmale hierfür zeigen sich dann die paralogen Gene in andersartiger Abfolge und Orientierung. Eine einhergehende funktionelle Veränderung von paralogen Genombereichen während der Evolution zeigt auch die Beobachtungen, dass sich die Kopplung zu einem Krankheitsbild meist immer nur auf eine der beiden paraloge Chromosomenabschnitte bezieht. So zeigen interessanterweise die im nachfolgenden Kap. 4.6 für die Chromosomenregion 11p15.3 beschriebenen kongenitalen Erkrankungen keine Assoziation zum Chromosom 12, auch wenn z. B. das Gen *TULP3* ein sehr ähnliches Expressionsmuster zu *TUB* aufweist und ähnliche Mutationen im hochkonservierten carboxyterminalen Bereich des Gens angenommen werden (Nishina *et al.*, 1998).

4.6 Die Kopplung der Chromosomenregion 11p15.3 zu kongenitalen Erkrankungen

Zytogenetische Untersuchen haben gezeigt, dass die in dieser Arbeit sequenzierte und untersuchte Chromosomenregion mit mehreren kongenitalen Krankheitsbildern gekoppelt zu sein scheint. Es handelt sich dabei um angeborene Erkrankungen, die meist einen sehr komplexen Phänotyp aufweisen und vermutlich durch polykausale genetische Störungen verursacht werden. Somit dürfte nicht nur eine einzige Mutation, bzw. ein einziges verändertes Gen für die Entstehung des Krankheitsbildes verantwortlich sein, sondern es könnte auch eine Störung auf genetisch regulativer Ebene in Betracht kommen, deren genaue Ursache bisher noch nicht bekannt ist.

4.6.1 Die Kopplung zum Beckwith-Wiedemann-Syndrom (BWS)

Der gesamte terminale Bereich des kurzen Armes von Chromosom 11 zeigt eine Kopplung zum komplexen Phänotyp des Beckwith-Wiedemann-Syndroms (BWS) (OMIM: #130650). Charakterisiert durch verschiedenste Anomalien im Wachstum zeigen betroffene Neugeborene vor allem die drei prominenten Merkmale Exomphalos (Nabelschnurbruch), Makroglossie und Gigantismus, so dass diese Erkrankung auch oft als EMG-Syndrom abgekürzt wird. Ein weiteres Merkmal dieses genetisch bedingten Fehlbildungssyndroms ist die hohe Prädisposition zur embryonalen Tumorbildung und dabei besonders zu Wilms-Tumoren, die über 40% aller diagnostizierten frühkindlichen Malignome der betroffenen Patienten ausmachen (Wiedemann, 1983).

Die molekulare Ätiologie von BWS scheint innerhalb der Chromosomenregion 11p15 in drei unterschiedlichen Bereichen lokalisiert zu sein, die bei den betroffenen Patienten z. T. durch chromosomale Anormalitäten auffielen. Diese als „BWS-kritische Regionen" (BWSCR) bezeichneten Abschnitte befinden sich zum einen in der terminalen Chromosomenbande 11p15.5 (= BWSCR1) in einem ca. 300 kb großen Intervall 200 kb bis 300 kb proximal zum Gen *IGF2* (Hoovers *et al.*, 1995) und zum anderen in ca. fünf Megabasen Entfernung weiter centromerwärts in der chromosomalen Bande 11p15.4 (= BWSCR2). Der dritte Bereich BWSCR3, ca. sieben Megabasen proximal zum BWSCR1, kartiert genau in den chromosomalen Abschnitt, der durch die vorliegende Arbeit sequenziert und untersucht wurde. Redeker & Mitarbeiter (1994) konnten mithilfe von FISH-Analysen zeigen, dass sich ein BWS-spezifischer Translokationsbruchpunkt in unmittelbarer Nähe zu den Genen *WEE1, ST5* und *LMO1,* bzw. nahe der STS-Markers D11S572 und D11S738 befindet. Auch wenn die postulierte Lage der Gene zum damaligen Zeitpunkt noch nicht der wahren Anordnung entsprach (siehe auch Abb. 2, Kap. 1.5), so zeigte doch der Bezug zum Gen *LMO1* und dem ca. 12 kb weiter distal gelegenen Marker D11S572 eine Korrelation zu diesem chromosomalen Abschnitt. Allerdings konnten für die möglichen Kandidatengene *WEE1*, *ST5* und *LMO1*, deren Funktion als Tumor-suppressor angesehen werden kann, bei BWS-Patienten keine genetischen Veränderungen wie Mutationen oder Genexpressionsunterschiede festgestellt werden. Bis heute sind auch keine anderen

potentiellen Kandidatengene in dieser BWSCR3-Region bekannt, so dass für die Ausprägung des BWS-Phänotyps noch eine andere, unbekannte Ursache vorliegen dürfte. Das Vorhandensein der exakten DNA-Sequenz aus dieser Region durch diese Arbeit ist somit eine wesentliche Grundlage für weitere Untersuchungen in diesem Fragenkomplex. Auch wenn cis-agierende, regulative DNA-Sequenzen, wie beispielsweise Enhancer, aufgrund der relativ großen Distanz zu Genen aus BWSCR1 als nicht wahrscheinlich angesehen werden können, so wären doch Regulationseinheiten durch unbekannte RNA- oder Antisense-Gene innerhalb der BWSCR3 durchaus vorstellbar. Da in der Vergangenheit keine, bzw. nur sehr unvollständige Sequenzinformationen über die Intergenbereiche vorlagen, war eine detaillierte Analyse bezüglich solcher Fragestellungen nicht möglich. Auf Grundlage der nun bekannten Genomsequenzen könnte eine Funktionsanalyse der zahlreichen konservierten genomischen DNA-Bereiche außerhalb der bekannten transkribierten Abschnitte zu Erkenntnissen führen, die neue Erklärungsansätze für die BWS-Erkrankung liefern.

Alders & Mitarbeiter (2000) konnten z. B. für den chromosomalen Bereich BWSCR2 zeigen, dass nicht wie im BWSCR1 ein Verlust des genomischen Imprintings und der daraus resultierenden Überproduktion des *IGF2*-Genproduktes mit Ursache für die Ausbildung von BWS ist (Mannens *et al.*, 1994), sondern dass auch die Unterbrechung zweier neuer Zink-Finger-Gene (*ZNF214* und *ZNF215*) an der Ätiologie von BWS beteiligt ist. Durch den Bruchpunkt entsteht nicht nur ein anormales alternatives Spleißens des *ZNS215*, gleichzeitig resultiert auch die Störung eines weiteren neuen Antisense-Gens von *ZNF214*. Da BWS-Patienten mit chromosomalen Veränderungen in BWSCR2 oder BWSCR3 einen unterschiedlichen Phänotyp aufweisen, scheinen weitere genetische Auslöser für das Krankheitsbild verantwortlich zu sein. BWSCR2 ist charakterisiert durch zwei Bruchpunkte, die immer mit einer Hemihypertrophie und der Prädisposition zu frühkindlichen Tumoren assoziiert sind. Patienten mit chromosomalen Bruchpunkten in BWSCR1 und BWSCR3 zeigen dagegen keine Ausprägung der Hemihypertrophie (wenn man von Fällen mit Mosaikbefunden absieht). Sie weisen aber alle anderen Merkmale in ausgeprägterer Form auf, so dass eine größere Ähnlichkeit in der Art und Weise der funktionellen Störung angenommen werden kann. Auch wenn für die hier untersuchte Region des BWSCR3 noch keine Hinweise auf epigenetische Regulationsmechanismen gefunden werden konnten, wäre grundsätzlich eine Störung der Gendosis, bzw. eine Veränderung in der Regulation der Expression eines bestimmten Gens vorstellbar. Diese Ursache würde dann nicht durch Mutationen in der kodierenden Gensequenz auffallen, sondern müsste in den teils sehr konservierten Intergen-Bereichen zu suchen sein, die erst durch Vorliegen dieser Arbeit analysiert werden können.

4.6.2 Die Kopplung zu genetisch bedingte Adipositas („Obesity")

Übergewicht ist nicht immer nur die Folge eines verhaltensbedingten, übermäßigen Konsums an Kalorien. Fettleibigkeit oder Adipositas kann auch Folge einer genetischen Disposition in der Regulation des eigenen Körpergewichts sein und tritt als krankheitsbedingtes Merkmal in einigen komplexen Syndromen, wie etwa dem Prader-Willi- (OMIM# 176270), dem Cohen- (OMIM# 216550), dem Alstrom- (OMIM# 203800), dem Bardet-Biedl- (OMIM# 209900) oder dem Borjeson-Forssmann-

Lehmann Syndrom (OMIM# 301900) auf. Da Adipositas eine komplexe Erkrankung mit multigenetischer Vererbung in Interaktion mit bestimmten Umweltfaktoren ist, für die nur wenige gekoppelte Mutationen bekannt sind, war es schwer über positionelles Klonieren im Menschen potentielle Kandidatengene zu identifizieren. Stattdessen wurden die meisten Erkenntnisse am Modellorganismus der Maus herausgefunden. So konnten über die Syntänie zwischen Maus und Mensch auch potentielle Kandidatengene im Menschen identifiziert werden. Mit dieser Vorgehensweise wurden z. B. die „Obesity"-Gene A^{Y} („agouti"), *db* („diabetes"), *ob* („obese"), *fat* und das in dieser Arbeit genauer charakterisierte *tub* („tubby) im Menschen bestimmt. Da eine Störung des *Tub*-Gens der Maus mit einer übermäßigen Zunahme des Körpergewichtes und Körperfettmasse korreliert ist und diese Chromosomenregion über QTL-Analysen („quantitative trait loci"; siehe Kap. 1.1) der Adipositas zugewiesen werden konnte (Taylor *et al.*, 1996), steht das homologe Gen *TUB* des Menschen ebenfalls in diesem funktionellen Zusammenhang und ist an der Regulation der Nahrungsaufnahme, des Körpergewichtes, bzw. des Grundumsatzes beteiligt. Bisher ist allerdings die in der Maus beschriebene Spleißstellen-Mutation im 3´-Bereich des *Tub*-Gens, welche zu einem trunkierten Protein und zu einem offensichtlichen Funktionsverlust des Genprodukts führt, beim Menschen nicht nachgewiesen worden. Es scheint daher als wahrscheinlich, dass im humanen Gen eine andere genetische Ursache für den Phänotyp verantwortlich ist. Neben der Körpergewichtszunahme weisen die erkrankten Tubby-Mäuse auch neurosensorische Defekte in Auge und Ohr auf. Merkmale, die im Zusammenhang mit den oben beschriebenen komplexen Syndromen in Erscheinung treten und auf eine pleiotrope Funktionsweise des *TUB* hinweisen könnten.

Recht gut beschrieben werden konnte bisher der Mechanismus der Gewichtszunahme über die Expressionssteuerung des Leptin-Gens, das dem murinen Obese-Gen *ob* entspricht und dem Leptin-Rezeptorgen, welches dem Gen *db* der Maus homolog ist. Da die genetischen Veränderungen des Leptin-Regulationssystem einige Parallelen zum Tubby-Protein zeigen, soll dieser Mechanismus hier näher erläutert werden, um im direkten Vergleich die vorhandenen Ähnlichkeiten zu betrachten. Denn noch immer ist der Auslöser für den Phänotyp Adipositas, verursacht durch die Mutation im 3´-Bereich des *Tub*-Gen, nur für die Maus bekannt. Für das humane *TUB*-Gen konnten bislang keine derartige oder andersartige genetische Veränderungen festgestellt werden.

Zhang und Mitarbeiter (1994) zeigten an Mäusen, dass die Expression des Leptin-Gens vom Ernährungszustand der Tiere abhängig ist. Beim Fasten nimmt die Expression ab und bei Fütterung wieder zu. Man stellte fest, dass das Körpergewicht, die Größe der Fettzellen (Adipozyten) und die gespeicherte Fettmenge mit der Expression des *ob*-Gens korreliert ist und dass bei ob/ob-Mäuse die Adipositas auf einer Punktmutation im *ob*-Gen beruht. Diese Mutation führt durch Abbruch der Translation dazu, dass keine Synthese des *ob*-Genprodukts mehr stattfindet. Für das Leptin-Rezeptorgen, von Tartaglia & Mitarbeitern (1995) bei der Maus kloniert, konnten unterschiedliche Rezeptorvarianten nachgewiesen werden, die durch alternatives Spleißen entstehen. Die Leptin-Rezeptoren setzen sich aus einem Leptin-bindenden extrazellulären Anteil, einer Transmembrandomäne und einer zytoplasmatischen Domäne zusammen, die sich durch einen unterschiedlich langen intrazellulären Bereich unterschieden (Tartaglia, 1997). Durch alternatives Spleißen ist vor allem der

intrazelluläre Bereich betroffen, der verschieden lang ist und bei der Maus in fünf Spleißvarianten vorkommt, welche im C-terminalen kodierenden Exon variabel sind. Wird das Leptin-Gen fast nur in den Adipozyten exprimiert, so findet die Expression des Leptin-Rezeptors ausschließlich in hypothalamischen Neuronen statt (Elmquist *et al.*, 1998). Nach Sekretion durch die Adipozyten und nach Transport an ein Trägerprotein ins Gehirn wird das Leptin dort im Hypothalamus gebunden. Funktioniert die Signalkette ins Gehirn, vermittelt Leptin durch die Bindung an seinen hypothalamischen Rezeptor ein Absinken des Neuropeptids Y (NPY) und ein Anstieg des Corticotropin-Releasing-Hormons (CRH), das die Nahrungsaufnahme bremst und über längere Zeit zu einer Abnahme des Adipozytenvolumens und zu einem Schrumpfen der Körperfettmasse führt. Mutationen in einem der beiden Gene *ob*, bzw. *db*, die zu defekten Leptin-Molekülen, bzw. Leptin-Rezeptoren führen, würden den Regelkreis unterbrechen und zu einer Fehlsteuerung von *NPY* und *CRH* führen, was wiederum eine ungebremste Nahrungsaufnahme und demzufolge eine Gewichtszunahme zur Folge hätte.

Parallelen zwischen dem Leptin-Rezeptor und dem *TUB*-Gen finden sich nicht nur im gemeinsamen Expressionsort dem Hypothalamus, sondern auch in den unterschiedlichen Spleißvarianten, denen verschiedene Funktionen zugeschrieben werden. So fehlt beispielsweise der Leptin-Rezeptor-Variante *OB-Re* die Transmembrandomäne, so dass es sich im Gegensatz zum eigentlich langen membranständigen Rezeptor, um einen löslichen Rezeptor handelt (Lee *et al.*, 1996). Ähnlich wie bei der Tubby-Protein-vermittelten Signaltransduktion durch Interaktion mit dem G-Protein-gekoppelten Rezeptor von der Zellmembran zum Zellkern (siehe Kap. 4.3.3.2) führt auch die Signalweiterleitung durch die Leptin-Rezeptor-Bindung zu einer Transkriptionssteuerung im Zellkern. Die Aggregation der Leptin-Rezeptor-Heterodimere führt zur Aktivierung der Rezeptor-assoziierten Tyrokinase, durch welche der Rezeptor selbst sowie zytoplasmatische Transkriptionsfaktoren, die sog. STATs (Signaltransduktoren und –aktivatoren der Transkription) phosphoryliert werden (Darnell, 1997). Nach Dimerisierung und Wanderung zum Zellkern, kann dort eine Initiation der Transkription spezifischer Gene stattfinden. Somit könnten beide „Pathways" – die Tubby-Protein-vermittelte und die Leptin-Rezeptor-vermittelte Signaltransduktion alternative Wege beschreiben, um die Transkription bestimmter Gene in den Neuronen des Hypothalamus zu steuern; in Kerngebieten also, die in die Regulation des Körpergewichtes involviert sind (Nakashima *et al*, 1997).

Genauso wie sich eine hypothalamische Steuerung aus den unterschiedlichen Spleißvarianten des Leptin-Rezeptors, die eine unterschiedliche Proteinlänge zur Folge haben, ableiten lässt, so könnten die in dieser Arbeit neu beschriebenen Spleißformen des *TUB*-Gens ebenfalls für neue Aufgaben stehen. Da für die neu beschriebenen Varianten keine Mutationsuntersuchungen vorliegen, wären genetische Veränderungen in diesen Bereichen mit einem Funktionsverlust dieser Genvariante und des daraus resultierenden Phänotyps der Adipositas durchaus denkbar. Auch genetische Veränderungen in regulativ wichtigen DNA-Sequenzen, vor allem in den sehr streng konservierten Intergen-Bereichen, wie jenseits des 3´-UTR der kodierenden *TUB*-Gensequenz wären in diesem Zusammenhang vorstellbar. Aufgrund der fehlenden genomischen DNA-Information konnten diesbezüglich allerdings noch keine Untersuchungen unternommen werden.

4.6.3 Die Kopplung zu allelotypisierten Lungenkarzinomen

Die in dieser Arbeit untersuchte Chromosomenregion wird auch mit der Pathogenese von Lungenkarzinomen in Verbindung gebracht, die in den USA unter der männlichen Bevölkerung zu den häufigsten Neoplasien zählen (Boring *et al.*, 1993). Chromosomale Studien von Bepler & Koehler (1995) zeigten, dass Lungenkarzinom-Zelllinien durch multiple chromosomale Aberrationen geprägt sind, die u.a. auch in der Telomer-Region von Chromosom 11p zu finden sind. Insbesondere fiel der chromosomale Abschnitt um das *ST5*-Gen und dem STS-Marker D11S932 auf, der in einigen der untersuchten Zelllinien heterozygot, bzw. monoallelisch deletiert zu sein schien (Bepler & Garcia-Blanca, 1994). Hierbei stellte D11S932 eine Sonde dar, für die ein Verlust der Homozygotie nachgewiesen werden konnte. Zu Beginn der Arbeit waren nur diese beiden Marker (*ST5* + D11S932) für diesen Bereich getestet. Außerdem lag der komplette Abschnitt in nur sehr grob kartierter Form vor. So zählte man den chromosomalen Locus der Marker *ST5* und D11S932 noch in die Bande 11p15.5 – eine Kartierung, die durch die vorliegende Arbeit korrigiert werden konnte. Der Marker D11S932 ist nach den hier vorliegenden Ergebnissen nur 100 kb proximal zum *LMO1*-Gen lokalisiert. Eine Distanz, die sich auf physikalischen Karten nur sehr schwer diskriminieren lässt. Demnach könnte nicht nur das exemplarisch getestete *ST5* ein Kandidatengen sein, sondern durchaus auch das putative Onkogen *LMO1*, dessen Tumorassoziation in Zusammenhang mit Leukämien bereits nachgewiesen wurde. Es könnten aber auch unbekannte Gene, bzw. regulatorische Sequenzen eine Rolle spielen, die aufgrund der fehlenden Genomsequenzinformationen bisher noch nicht charakterisiert wurden.

Dass die Chromosomenregion 11p15 durch den Funktionsverlust eines Tumorsuppressorgens mit der Ätiologie von Krebserkrankungen in Zusammenhang steht, belegen auch Untersuchungen an „Nicht-kleinzelligen Lungenkarzinomen", die Deletionen für diesen Bereich beschreiben (Michelland *et al.*, 1999). Des weiteren konnten Studien am Typus des „Kleinzelligen Lungenkarzinoms" zeigen, dass am Gen *TSG101* mit seinem Lokus auf 11p15 nicht Mutationen, sondern das Auftreten unterschiedlicher Spleißvarianten mit der Entstehung der Lungenkrebserkrankung in Zusammenhang stehen (Oh *et al.*, 1998). Dies gibt einen Hinweis darauf, dass nicht unbedingt nur Mutationen als Auslöser für einen krankhaften Phänotyp verantwortlich sein müssen, sondern dass auch ein Spleißdefekt Ursache für die Entartung von Zellen und für das Entstehung von Erkrankungen sein kann. Diese Gegebenheit könnte vor dem Hintergrund der neu beschriebenen Spleißvarianten der Gene *LMO1*/*Lmo1* und *TUB*/*Tub* sehr interessant sein, weil zum einen bisher keine Vorstellungen über die etwaigen unterschiedlichen Funktionen der alternativen Spleißprodukte existierten und zum anderen mögliche Spleißdefekte in den alternativen Spleißvarianten nicht untersucht werden konnten. Somit könnte auch für diesen Aspekt eine Funktionsanalyse der verschiedenen Spleißvarianten und ihre Überprüfung in den neoplastischen Zellen zu neuen Erkenntnissen bei der Klärung der genetischen Ursachen für die in dieser Region kartieren Erkrankungen liefern.

4.7 Anwendung von Genomsequenzierungsergebnissen

4.7.1 Stärken der komparativen Analyse

Die vergleichende Analyse zweier, bzw. mehrerer Genomsequenzen unterschiedlicher Spezies hat in der vorliegenden Arbeit gezeigt, dass der komparative Ansatz aufgrund der vielen evolutionsbedingten physiologischen, anatomischen und metabolischen Gemeinsamkeiten auf genomischer und genetischer Ebene zahlreiche stark konservierte Sequenzbereiche erkennen lässt. Diese Konservierung wurde als ein Spiegelbild für wichtige biologische Informationen gewertet, die für beide Spezies lebensnotwendig sind. Würde diesen homologen Sequenzbereichen keine biologische Funktion zugrunde liegen, hätten sie sich in der Evolution durch die eigenständige Dynamik je nach Spezies unterschiedlich stark verändern müssen. Aufgrund dieses Prinzips ist es möglich, durch den bloßen Vergleich zweier Genomsequenzen informative DNA-Bereiche anzusprechen, selbst wenn deren Funktion in beiden Organismen noch unbekannt ist. Somit liefert der komparative Ansatz eine sehr effiziente Möglichkeit, aus der großen Masse der nicht-kodierenden und nicht-informativen DNA höherer Eukaryonten potentiell funktionelle Bereiche herauszufiltern, um sie dann gezielt in weiteren experimentelle Analysen (Expressionsstudien, RNAi) zu untersuchen. Für diesen methodischen Schritt waren früher vor dem Bekanntwerden der genomischen Sequenzen sehr zeitintensive Untersuchungen wie Positionelles Klonieren, Kandidatengenansatz oder die Exon-Amplifikation nötig (siehe Kap. 1.1), bzw. konnten zuvor gar nicht gezielt identifiziert werden (z. B. regulative Sequenzen), da informative DNA-Bereiche erst durch eine computergestützte Auswertung näher charakterisiert werden können, wenn entsprechende Referenzinformationen (cDNA-Sequenzen, Motive) vorliegen. Wie diese Arbeit zeigen konnte, lieferte der komparative Ansatz eine Vielzahl von validen Hinweise, deren genaue Charakterisierung und Beurteilung allerdings sukzessive in weiteren Analysen überprüft und spezifiziert werden muss. Somit bilden die Ergebnisse der vergleichenden Genomanalyse dieser Arbeit eine entscheidende Grundlage für alle weiteren genetischen Forschungsarbeiten in dieser Region.

Eine wichtige Vorraussetzung für eine aussagekräftige Interspeziesanalyse ist die Existenz möglichst vieler bekannter Genomsequenzen. Eine Rahmenbedingung, die durch die Beendigung immer weiterer Genomprojekte stetig verbessert wird. So dürfte eine erweiterte Sequenzanalyse mit einem alternativen Organismus sicherlich viele in dieser Arbeit identifizierte konservierte Sequenzbereiche in der humanen Genomsequenz bestätigen. Andererseits würden je nach Verwandtschaftsgrad der gewählten Spezies auch unbekannte konservierte Bereiche hinzukommen, deren Existenz als neue Bausteine im genomischen „Puzzle" der funktionell interessanten DNA-Bereiche gezählt werden können.

Prinzipiell ist die Möglichkeit eine Genomsequenz mit mehr als einem Organismus zu vergleichen für die inhaltliche Auswertung der Homologien sehr wichtig und notwendig. In der vorliegenden Arbeit zeigte sich, dass die Fülle an konservierten Sequenzbereichen im Mensch-Maus-Vergleich für nachfolgende detaillierte Funktionsanalysen erst einmal eingeschränkt werden musste. Zwar

ermöglichte die Kombination mehrerer unterschiedlicher Computer-Analysen (z. B. Exonvorhersage, Motiv-Homologiesuche, Promotor-Analyse) einen konservierten Sequenzbereich nach bestimmten Kriterien wie z. B. seiner proteinkodierenden Information zu untersuchen, doch muss dessen genaue vorhergesagte biologische Funktion in weiteren Laborexperimenten verifiziert werden. Als vorteilhaft für die Reduzierung von konservierten Sequenzbereichen hat sich ein zweiter Interspeziesvergleich mit einem noch entfernter verwandten Organismus, wie z. B. *Fugu*, herausgestellt. Aufgrund der getrennten Evolution von geschätzten 365 Mio. Jahren (Hedges & Kumar, 2002) wiesen die meisten Sequenzbereiche eine solch große Diversität auf, dass sie in Computerprogrammen wie z. B. der PIP-Analyse grafisch nicht mehr hervorgehoben wurden (Vgl. Abb. 27). Als konserviert zeigten sich in deutlich reduzierter Zahl nur die genkodierenden oder darüber hinaus informationstragenden Sequenzbereiche. Eine weitere Alternative wäre beispielsweise der Sequenzvergleich mit dem Nematoden *C. elegans* gewesen, da die Genetik dieser Spezies als Modellorganismus bereits gut verstanden ist und es für Expressionsanalysen ein Reihe von etablierten Methoden gibt, wie z.B. *in vivo* Genexpressionsstudien mit GFP-markierten Proteinen („Green Fluorescent Protein"). Da allerdings keine größeren zusammenhängenden Genomsequenzen für einen komparativen Vergleich zur Verfügung standen, fand diese Analyse keine Verwendung.

Eine bezüglich ihrer Auswertung sozusagen komplementäre Vorgehensweise wäre der Vergleich mit einem dem Menschen noch näher verwandten Organismus wie z. B. dem Schimpansen (*Pan troglodytes*) gewesen, dessen Genom zum Menschen eine Homologie von 98,8% aufweist (Fujiyama *et al.*, 2002). In einem solchen Interspeziesvergleich wären dann nicht mehr die vielen Konservierungen relevant, sondern vielmehr die genetischen Unterschiede. Die Existenz solcher Unterschiede, insbesondere in kodierenden DNA-Abschnitten, ist besonders für die spezifische Funktion der Expressionsprodukte interessant. Genetische Variationen und Unterschiede könnten dann konkrete Hinweise für die Fähigkeiten geben, die als typisch menschlich gelten, wie z. B. unseren mentalen und linguistischen Leistungen. So konnten beispielsweise Lai & Mitarbeiter (2001) zeigen, dass eine Punktmutation im alternativ gespleißten Gen *FOXP2* bei menschlichen Patienten zu schweren Sprach- und Sprechstörungen führt. Auch der Interspeziesvergleich der Proteinsequenz des orthologen Schimpansen-Gens zeigte lediglich zwei Aminosäureaustausche (Enard *et al.*, 2002). Selbst das homologe Gen der Maus besitzt mit insgesamt drei nur einen einzigen Aminosäureaustausch mehr. Diese marginalen speziesspezifischen Unterschiede scheinen allerdings so große sekundäre Strukturveränderungen im Proteins zu bewirken, dass die Funktion im Kontext der Sprachentwicklung nur in der humancharakteristischen Proteinsequenz verwirklicht werden konnte. Abweichungen von dieser Proteinsequenzabfolge führen sowohl beim Menschen in der krankhaften Form, wie auch beim Tier zu Veränderungen in der sprachlichen Ausdrucksfähigkeit.

Ein anderer Vorteil des Interspeziesvergleichs mit einem sehr nahen menschlichen Verwandten ist die Möglichkeit, besonders dynamische DNA-Sequenzbereiche komparativ zu untersuchen, die im Vergleich mit der Maus zu keinen sinnvollen Ergebnissen führen würden. Solche besonderen Bereiche stellen die hypervariablen Regionen im menschliche Genom dar, wie sie beispielsweise im 3´-Bereich des *apoB*-Gens (Knott *et al.*, 1986) oder in der Umgebung der Immunglobulin- und T-Zell-Rezeptor-

Gene zu finden sind. Als Orte mit hohen Mutationsraten und vielen chromosomalen Rearrangements kamen sie bisher für eine vergleichende Interspeziesanalyse aufgrund ihrer hohen Heterogenität nicht in Betracht (Cyranoski, 2002).

Auch bei der Ursachenerforschung vieler menschlicher Erkrankungen ist die komparative Analyse mit einem sehr nahen menschlichen Verwandten ein aufschlussreicher Ansatz, da einige Krankheitsverläufe und die Anfälligkeit überhaupt Krankheitssymptome zu entwickeln, zwischen Mensch und Tier sehr unterschiedlich verlaufen können. So leiden z. B. HIV-infizierte Schimpansen nicht an den AIDS-Symptomen wie sie der Mensch ausbildet (Varki, 2000). Als Ursache werden Unterschiede zwischen Mensch und Affe in den betreffenden Genen, wie etwa den MHC-Genen („major-histocompatibility complex) diskutiert (De Groot *et al.*, 2002). Ein anderes Beispiel ist die Infektion an Malaria durch den Erreger *Plasmodium falciparum*. Selbst hohe Titer des Erregers führen bei Schimpansen nicht zu einem Ausbruch der Krankheit (Ollomo *et al.*, 1997). Es scheint, dass die speziesspezifische Konstitution des Immunsystems der Schimpansen ihnen eine gewisse Immunität gegenüber diesem Parasit verleiht. Vereinzelte genetische Unterschiede könnten hier entscheidend für die Vitalität des Organismus und für die Ausprägung von Krankheitssymptomen sein.

Ein ganz anderer, aber methodisch entscheidender Vorteil der komparativen Sequenzanalyse ist die parallele Identifizierung der konservierten und somit funktionell relevanten Sequenzbereiche in *beiden* Vergleichsspezies; so wie in der vorliegenden Arbeit Abschnitte sowohl im Menschen als auch in der Maus, bzw. in *Fugu* identifiziert wurden. Dies hat den praktischen Nutzen, dass der konservierte DNA-Abschnitt auch gleichzeitig im tierischen Vergleichsorganismus bekannt ist und für nachfolgende experimentelle Studien handhabbar und manipulierbar wird. Hier bietet insbesondere die Maus als Modellsystem, wie bereits in der Einleitung (Kap. 1.3) beschrieben, verschiedenste experimentelle Möglichkeiten und erlaubt die Untersuchung komplexer Fragestellungen, die in Zusammenhang mit genetisch prädisponierten humanen Erkrankungen stehen. Da für fast alle humanen Gene murine Gegenspieler beschrieben werden konnten, bzw. für viele komplexe Erkrankungen gut charakterisierte Mäusestämme zur Verfügung stehen, die humangenetische Erkrankungen widerspiegeln, erlaubt die Maus auch eine Erforschung systemischer und multifaktorieller Störungen, die z. B. die Embryonalentwicklung oder Stoffwechselstörungen betreffen. Ergebnisse der vergleichenden Sequenzanalyse aus Maus-Mensch können somit letztlich für alle relevanten Fragestellungen wichtige Aspekte liefern, die zu einem tieferen grundlegenden Verständnis führen sowohl für die genetischen Ursachen von krankhaften Veränderungen wie Krebs oder kardiovaskuläre Erkrankungen, wie auch für komplexe Störungen, die das Verhalten, Lernen und Erinnern bis hin zu psychischen Erkrankungen betreffen (Boguski, 2002). Gleichzeitig bietet der Organismus der Maus auch die Möglichkeit, neue therapeutische und pharmakologische Ansätze experimentell „*in vivo*" zu testen und so Ergebnisse der grundlagenorientierten Forschung in die angewandte Forschung umzusetzen.

Die enorme Menge an genetischen Sequenzdaten und Ergebnissen, die aus den unterschiedlichen Genomprojekten generiert wurden, bekommen also erst durch ihre komparative Analyse und die dadurch erzielte biologische Vernetzung einen qualitativen, funktionellen Wert, der weitaus

aussagekräftiger ist, als er bei der eingeschränkten Betrachtung innerhalb einer Art wäre. Komplex strukturierte Datenbanken wie z. B. der „Ensembl Genome Browser", der die Daten von mittlerweile 12 verschiedenen Spezies miteinander in Beziehung setzt und vernetzt (Hubbard *et al.*, 2002) (http://www.ensembl.org/) bilden dabei eine Grundlage „*in silico*", die sämtliche zukünftige genetische Fragestellungen und Forschungsansätze „*in vitro*" und „*in vivo*" beeinflussen und mit zusätzlichen Daten und wichtigen Zusammenhängen komplettieren wird. Somit dürften auch die in dieser Arbeit generierten genomischen Sequenzinformationen und die Ergebnisse aus den durchgeführten Analysen konstruktiver Ausgangspunkt für weitere Studien sein, um das noch unvollständige kodierende und regulative Potential der sequenzierten chromosomalen Region 11p15.3 detailliert zu durchleuchten.

4.7.2 Biomedizinische Aspekte der komparativen Sequenzanalyse

Trotz aller beschriebenen Vorteile der komparativen Interspeziesanalyse sind der Vergleichbarkeit aber auch speziesspezifische und individuenspezifische Grenzen gesetzt, die die all zu mechanistische Vorstellung von Lebensabläufen und die generelle Übertragbarkeit von genetischen Dispositionen zu biologischen Abläufen relativieren. Diese Grenzen sind wahrscheinlich prinzipieller Natur und zeigen eine generelle Problematik auf, die auf das richtungsweisende Potential der vererbten DNA-Informationen und den sich daraus ableitenden individuellen Phänotyp beziehen. Letztlich ist auch das einzelne Genom eines jeden Organismus kein statisches Gebilde, sondern vielmehr ein höchst dynamisches Objekt, das durch vielerlei individuelle Umwelteinflüsse über die Lebenszeit geprägt und verändert werden kann. Ergebnisse eines genomischen Vergleichs können daher immer nur Momentaufnahmen einer bestimmten genetischen Konstitution sein, die in erster Näherung die Grundlage für das bestehende Leben ist. Diese Betrachtungsweise ist insbesondere im biomedizinische Zusammenhang und bei der Beurteilung des medizinischen Potentials der Ergebnisse aus Sequenzierprojekten von Bedeutung. Die Funktionalität eines gegebenen Gens wird nicht nur durch die Konstitution seiner Allele, sondern auch durch die möglichen Einflüsse aus seiner direkten und indirekten genomischen Umgebung geprägt (Baird, 2001). Diese endogenen Einflussfaktoren reichen vom genomischen Imprinting, über Heteroplasmie (Mutationen im Genom der Mitochondrien) (Erkrankung: Leigh´s Syndrom; Dahl, 1998) bis hin zu der veränderten Zahl von Trinukleotid-Wiederholungen (Erkrankung: Morbus Huntington; Brinkman *et al.*, 1997), die selbst mono-genetischen Erkrankungen eine sehr große Varianzbreite verleihen können (Weatherall, 1999). D. h., die Ausprägung, die Stärke und der Zeitpunkt des Auftretens eines krankhaften Phänotyps werden nicht ausschließlich durch die Mutationen an einem chromosomalen Locus bestimmt, sondern ebenso von diesen genetischen Modifikatoren abhängen. Zu den komplexen endogenen Einflüssen wirken zusätzlich die umweltbedingten, exogenen Faktoren, die sich aus den individuellen Lebensumständen, wie Ernährung, körperliche Aktivität oder z. B. Alkohol- oder Zigarettenkonsum zusammensetzen. Als gut untersuchtes Beispiel für dieses komplexe Zusammenspiel kann die β-Thalassämie, eine autosomal-dominante Erkrankung mit hämolytischer Anämie, genannt werden. Hier wird die Schwere der Erkrankung durch mehrere sehr unterschiedliche Faktoren, wie der individuellen Konstitution der

Allele und der „zusammengesetzten" Heterozygothie („compound heterozygosity" = verschiedene Allel-spezifische Mutationen), aber auch die postnatale Produktion von fötalem Hämoglobin und die Interaktion mit dem entfernt liegenden Globin-Gencluster beeinflusst (Weatherall, 1999). Gleichzeitig spielen zu diesem komplexen genetischen Umfeld auch Umwelt- und klimatische Faktoren eine wichtige Rolle. Somit ist häufig nicht eine Mutation im proteinkodierenden Gen die „treibende Kraft" bei der Entstehung einer Erkrankung, sondern vielmehr die individuellen epigenetischen und exogenen Faktoren. Auch der sozioökonomische Status des Einzelnen kann entscheidend für die Ausprägung von bestimmten Erkrankungen wie Herzleiden, Darm- oder Brustkrebs sein (Lynch *et al.*, 1998; Marmot *et al*, 1975).

Dieser Aspekt ist insbesondere für die Krankheitsprognose von genetisch prädisponierten Erkrankungen relevant, die durch molekulargenetische diagnostische Methoden schon sehr frühzeitig festgestellt werden können. Aber gerade das komplexe Ursachengefüge der meisten allgemeinen Zivilisationserkrankungen, wie z. B. *Diabetes mellitus* oder die arteriosklerotischen Herzerkrankungen, lassen eine Vorhersage der bevorstehenden Pathogenese auf Basis der genetischen Disposition nur sehr schwer zu. Somit dürfte die Realisierbarkeit von zuverlässigen diagnostischen Methoden, die mehr als nur einen Wahrscheinlichkeitsprozentsatz liefern und die angestrebten therapeutischen und gentherapeutischen Maßnahmen nicht in absehbarer Zeit, insbesondere für die meisten Volkserkrankungen, zu entwickeln sein (Karanjawal & Collins, 1999). Trotz dieser Schwierigkeiten bietet die komparative Genomanalyse aber neue und einzigartige Möglichkeiten, der Aufklärung dieser sehr komplizierten biologischen und genetischen Zusammenhänge und Vernetzungen in jedem lebenden Organismus sukzessive näher zu kommen. Niemand hätte sich vor 50 Jahren bei der Strukturaufklärung des DNA-Moleküls vorstellt, dass nach nur einem halben Jahrhundert die Basenpaarstruktur von ganzen Genomen entschlüsselt sein wird und welches Potential die Genetik für die Lebenswissenschaften bekommt. Auch wenn sich manche Erwartungen insbesondere für therapeutische Ansätze noch nicht erfüllt haben, so ist dies lediglich das Zeichen für die enorme Komplexität von genetischen Systemen, die sich bereits in sehr „einfachen" Lebensformen offenbart. Durch den direkten Vergleich der Nukleotidsequenz-Informationen zeigt sich, dass das Leben auf der genetischen Betrachtungsebene sich wesentlich ähnlicher ist, als es die vielen unterschiedlichen Erscheinungsformen des Lebens vermuten lassen. Für die molekulargenetische Forschung bildet die Entschlüsslung der Genome des Menschen und verschiedener Modellorganismen das Rückgrat zu einem neuen molekularen Gesamtverständnis der belebten Natur. Die hohe Komplexität und Dynamik der durch die DNA-Doppelhelix verpackten Informationen übersteigt alle bisher etablierten Vorstellungen. Die Funktionsweise dieser gemeinsamen „untersten" Ebene allen Lebens gänzlich zu verstehen wird daher noch für viele Jahrzehnte ein höchst spannendes Gebiet der Forschung mit vielen Überraschungen bleiben. Ein kleiner Bruchteil zur Klärung dieser großen Aufgabe wurde durch die vorliegende Arbeit bereitgestellt.

5 ZUSAMMENFASSUNG

Ziel der vorliegenden Arbeit war die vergleichende Sequenzierung und nachfolgende Analyse eines syntänen chromosomalen Abschnitts auf dem kurzen Arm des Chromosoms 11 in der Region 11p15.3 und dem orthologen Genomabschnitt der Maus auf Chromosom 7 F2. Die im Rahmen dieser Arbeit durchgeführte Kartierung der beiden chromosomalen Bereiche in Mensch und Maus ermöglichte durch die identifizierten Klone die Erstellung einer genomischen Karte insgesamt über eine Megabase, die im Kooperationssequenzierprojekt der Universitäts-Kinderklinik und dem Institut für Molekulargenetik charakterisiert wurde. Mit Hilfe von 28 PAC- bzw. Cosmid-Klonen konnten in dieser Arbeit 383 kb an chromosomaler DNA des Menschen und mit 6 BAC- bzw. PAC-Klonen 412 kb an muriner DNA dargestellt werden. Dies ermöglichte erstmals die Festlegung der Reihenfolge der in diesem Abschnitt enthaltenen Gene und die genaue Kartierung von 8 STS-Markern des Menschen, bzw. 4 STS-Sonden der Maus. Es zeigte sich dabei, dass die chromosomale Orientierung telomer-/centromerwärts des orthologen Bereichs in der Maus im Vergleich zum Menschen invertiert vorliegt. Die mit Hilfe von drei Klonen realisierte Sequenzierung von 319.119 bp an zusammenhängender genomischer Human-DNA ermöglichte die genaue Lokalisation und Strukturaufklärung der Gene *LMO1*, ein putatives Tumorsuppressorgen, das mit der Entstehung von Leukämien assoziiert ist, und *TUB*, ein Transkriptionsmodulator, das in die Fettstoffwechselregulation involviert ist. Für das murine Genom wurden 412.827 bp durch Sequenzierung von drei Klonen an neuer DNA-Sequenz generiert, die einen Bereich zwischen den Genen *Stk33* und *Eif3* beschreiben. Die parallele Bearbeitung beider chromosomaler Genombereiche ermöglichte die umfassende komparative Analyse nach kodierenden, funktionellen und strukturgebenden Sequenzabschnitten in beiden Spezies. Es konnten dabei für beide Organismen die Exon-Intron-Strukturen der Gene *LMO1*|*Lmo1* und *TUB*|*Tub* geklärt, inklusive vier neuer Exons und zwei neuer speziesspezifischer Spleißvarianten für *TUB*|*Tub* beschrieben werden. Die Identifizierung neuer Spleißvarianten der physiologisch wichtigen Gene *LMO1*|*Lmo1* und *TUB*|*Tub* ermöglicht neue Erklärungsansätze für die Regulation, Funktion und Proteinstruktur und für ihre Rolle bei der Entstehung der assoziierten Erkrankungen. Für die etwas größere Genomsequenz der Maus konnte zudem das neue Gen *Eif3* in seiner Exon-Intron-Struktur und die beiden letzten Exons 11 und 12 des Gens *Stk33* kartiert und charakterisiert werden. Die umfangreiche Sequenzanalyse beider sequenzierter Genombereiche ergab für den Abschnitt des Menschen insgesamt 229 potentielle Exonsequenzen und für den Bereich der Maus 527 mögliche Exonbereiche. Davon konnten beim Menschen 21 Exons und bei der Maus 31 Exons als exprimierte Bereiche experimentell mittels RT-PCR, bzw. durch cDNA-Sequenzen verifiziert und den oben genannten und noch nicht weiter Genen zugeordnet werden. Mittels des Interspeziesvergleiches war darüber hinaus auch eine Analyse der nichtkodierenden Intergen-Bereiche möglich. So konnten im ersten Intron des *LMO1*|*Lmo1* sieben Sequenzbereiche mit Konservierungen um die 90% bestimmt werden. Auch die Charakterisierung von Promotor- und putativ regulatorischen Sequenzabschnitten wurde mit Hilfe unterschiedlicher bioinformatischer Analyse-Tools durchgeführt Der Interspeziesvergleich beider Genomsequenzen

insgesamt zeigte, dass die DNA im untersuchten Bereich in beiden Organismen über die Evolution hinweg relativ konserviert geblieben ist, ohne größere DNA-Verluste oder DNA-Insertionen aufzuweisen. Allerdings ist die DNA des Menschen um ca. 14% länger als die Sequenz der Maus, was sich größtenteils durch eine verstärkte Anhäufung an repetitiven Elementen im menschlichen Genom erklären lässt. Über weite Sequenzbereiche weist die DNA aber Homologien von mehr als 65% auf. Gerade die Betrachtung der Genomorganisation zeigte prinzipielle Gemeinsamkeiten, die aber meist mit graduellen Unterschieden zur Ausprägung kommen. So weist ein knapp 80 kb großer Bereich proximal zum humanen *TUB*-Gen einen deutlich erhöhten AT-Gehalt auf, der im murinen Genom nur in verkürzter Version und schwächer ausgeprägt in Erscheinung tritt. Die zusätzliche Vergleichsanalyse mit noch einer weiteren Spezies, den orthologen Genomabschnitten von *Fugu* zeigte, dass es sich bei den untersuchten Genen *LMO1* und *TUB* um sehr konservierte und evolutiv alte Gene handelt, deren genomische Organisation sich darüber hinaus auch bei den paralogen Genfamilienmitglieder innerhalb derselben Spezies wiederfindet. Insgesamt konnte durch die Kartierung, Sequenzierung und Analyse eine umfassende Datenbasis generiert werden, die als Ausgangspunkt für alle weiteren Untersuchungen und Fragestellungen bezüglich der Region und bezüglich der in ihr kodierten Bereiche Verwendung finden kann.

6 LITERATURVERZEICHNIS

- A -

1. AALTONEN,J., BJORSES,P., SANDKUIJL,L., PERHEENTUPA,J., PELTONEN,L., 1994. "An autosomal locus causing autoimmune disease: autoimmune polyglandular disease type I assigned to chromosome 21", *Nat.Genet.*, 8, S. 83-87.

2. ADACHI,K., KATSUYAMA,M., SONG,S., OKA,T., 2000. "Genomic organization, chromosomal mapping and promoter analysis of the mouse selenocysteine tRNA gene transcription-activating factor (mStaf) gene", *Biochem.J.*, 346 Pt 1, S. 45-51.

3. ADAM,G.I., 2001. "The development of pharmacogenomic models to predict drug response", *Curr.Opin.Drug Discov.Devel.*, 4, S. 296-300.

4. ADAMS,M.D., CELNIKER,S.E., HOLT,R.A., EVANS,C.A., GOCAYNE,J.D., AMANATIDES,P.G., SCHERER,S.E., LI,P.W., HOSKINS,R.A., GALLE,R.F., GEORGE,R.A., LEWIS,S.E., RICHARDS,S., ASHBURNER,M., HENDERSON,S.N., SUTTON,G.G., WORTMAN,J.R., YANDELL,M.D., ZHANG,Q., CHEN,L.X., BRANDON,R.C., ROGERS,Y.H., BLAZEJ,R.G., CHAMPE,M., PFEIFFER,B.D., WAN,K.H., DOYLE,C., BAXTER,E.G., HELT,G., NELSON,C.R., GABOR,G.L., ABRIL,J.F., AGBAYANI,A., AN,H.J., ANDREWS-PFANNKOCH,C., BALDWIN,D., BALLEW,R.M., BASU,A., BAXENDALE,J., BAYRAKTAROGLU,L., BEASLEY,E.M., BEESON,K.Y., BENOS,P.V., BERMAN,B.P., BHANDARI,D., BOLSHAKOV,S., BORKOVA,D., BOTCHAN,M.R., BOUCK,J., BROKSTEIN,P., BROTTIER,P., BURTIS,K.C., BUSAM,D.A., BUTLER,H., CADIEU,E., CENTER,A., CHANDRA,I., CHERRY,J.M., CAWLEY,S., DAHLKE,C., DAVENPORT,L.B., DAVIES,P., DE PABLOS,B., DELCHER,A., DENG,Z., MAYS,A.D., DEW,I., DIETZ,S.M., DODSON,K., DOUP,L.E., DOWNES,M., DUGAN-ROCHA,S., DUNKOV,B.C., DUNN,P., DURBIN,K.J., EVANGELISTA,C.C., FERRAZ,C., FERRIERA,S., FLEISCHMANN,W., FOSLER,C., GABRIELIAN,A.E., GARG,N.S., GELBART,W.M., GLASSER,K., GLODEK,A., GONG,F., GORRELL,J.H., GU,Z., GUAN,P., HARRIS,M., HARRIS,N.L., HARVEY,D., HEIMAN,T.J., HERNANDEZ,J.R., HOUCK,J., HOSTIN,D., HOUSTON,K.A., HOWLAND,T.J., WEI,M.H., IBEGWAM,C., JALALI,M., KALUSH,F., KARPEN,G.H., KE,Z., KENNISON,J.A., KETCHUM,K.A., KIMMEL,B.E., KODIRA,C.D., KRAFT,C., KRAVITZ,S., KULP,D., LAI,Z., LASKO,P., LEI,Y., LEVITSKY,A.A., LI,J., LI,Z., LIANG,Y., LIN,X., LIU,X., MATTEI,B., MCINTOSH,T.C., MCLEOD,M.P., MCPHERSON,D., MERKULOV,G., MILSHINA,N.V., MOBARRY,C., MORRIS,J., MOSHREFI,A., MOUNT,S.M., MOY,M., MURPHY,B., MURPHY,L., MUZNY,D.M., NELSON,D.L., NELSON,D.R., NELSON,K.A., NIXON,K., NUSSKERN,D.R., PACLEB,J.M., PALAZZOLO,M., PITTMAN,G.S., PAN,S., POLLARD,J., PURI,V., REESE,M.G., REINERT,K., REMINGTON,K., SAUNDERS,R.D., SCHEELER,F., SHEN,H., SHUE,B.C., SIDEN-KIAMOS,I., SIMPSON,M., SKUPSKI,M.P., SMITH,T., SPIER,E., SPRADLING,A.C., STAPLETON,M., STRONG,R., SUN,E., SVIRSKAS,R., TECTOR,C., TURNER,R., VENTER,E., WANG,A.H., WANG,X., WANG,Z.Y., WASSARMAN,D.A., WEINSTOCK,G.M., WEISSENBACH,J., WILLIAMS,S.M., WOODAGET, WORLEY,K.C., WU,D., YANG,S., YAO,Q.A., YE,J., YEH,R.F., ZAVERI,J.S., ZHAN,M., ZHANG,G., ZHAO,Q., ZHENG,L., ZHENG,X.H., ZHONG,F.N., ZHONG,W., ZHOU,X., ZHU,S., ZHU,X., SMITH,H.O., GIBBS,R.A., MYERS,E.W., RUBIN,G.M., VENTER,J.C., 2000. "The genome sequence of Drosophila melanogaster", *Science*, 287, S. 2185-2195.

5. ADAMS,R.L., EASON,R., 1984. "Increased G + C content of DNA stabilizes methyl CpG dinucleotides", *Nucleic Acids Res.*, 12, S. 5869-5877.

6. AINSCOUGH,J.F., JOHN,R.M., SURANI,M.A., 1998. "Mechanism of imprinting on mouse distal chromosome 7", *Genet.Res.*, 72, S. 237-245.

7. AISSANI,B., BERNARDI,G., 1991. "CpG islands, genes and isochores in the genomes of vertebrates", *Gene*, 106, S. 185-195.

8. AISSANI,B., BERNARDI,G., 1991. "CpG islands: features and distribution in the genomes of vertebrates", *Gene*, 106, S. 173-183.

9. AISSANI,B., D'ONOFRIO,G., MOUCHIROUD,D., GARDINER,K., GAUTIER,C., BERNARDI,G., 1991. "The compositional properties of human genes", *J.Mol.Evol.*, 32, S. 493-503.

10. ALDERS,M., RYAN,A., HODGES,M., BLIEK,J., FEINBERG,A.P., PRIVITERA,O., WESTERVELD,A., LITTLE,P.F., MANNENS,M., 2000. "Disruption of a novel imprinted zinc-finger gene, ZNF215, in Beckwith-Wiedemann syndrome", *Am.J.Hum.Genet.*, 66, S. 1473-1484.

11. ALTSCHUL,S.F., MADDEN,T.L., SCHAFFER,A.A., ZHANG,J., ZHANG,Z., MILLER,W., LIPMAN,D.J., 1997. "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", *Nucleic Acids Res.*, 25, S. 3389-3402.

12. AMID,C., BAHR,A., MUJICA,A., SAMPSON,N., BIKAR,S.E., WINTERPACHT,A., ZABEL,B., HANKELN,T., SCHMIDT,E.R., 2001. "Comparative genomic sequencing reveals a strikingly similar architecture of a conserved syntenic region on human chromosome 11p15.3 (including gene ST5) and mouse chromosome 7", *Cytogenet.Cell Genet.*, 93, S. 284-290.

13. ANGEL,J.M., MOORE,J.L., PELPHREY,A., RICHIE,E.R., 1993. "The mouse homolog of the rhombotin (Ttg-1) gene maps on chromosome 7 distal to the beta-globin (Hbb) locus", *Mamm.Genome*, 4, S. 281-282.

14. ANSARI-LARI,M.A., SHEN,Y., MUZNY,D.M., LEE,W., GIBBS,R.A., 1997. "Large-scale sequencing in human chromosome 12p13: experimental and computational gene structure determination", *Genome Res.*, 7, S. 268-280.

15. ANSARI-LARI,M.A., OELTJEN,J.C., SCHWARTZ,S., ZHANG,Z., MUZNY,D.M., LU,J., GORRELL,J.H., CHINAULT,A.C., BELMONT,J.W., MILLER,W., GIBBS,R.A., 1998. "Comparative sequence analysis of a gene-rich cluster at human chromosome 12p13 and its syntenic region in mouse chromosome 6", *Genome Res.*, 8, S. 29-40.

16. ANTEQUERA,F., BIRD,A., 1993. "Number of CpG islands and genes in human and mouse", *Proc.Natl.Acad.Sci.U.S.A* , 90, S. 11995-11999.

17. ANTONARAKIS,S.E., KAZAZIAN,H.H., TUDDENHAM,E.G., 1995. "Molecular etiology of factor VIII deficiency in hemophilia A", *Hum.Mutat.*, 5, S. 1-22.

18. APARICIO,S., CHAPMAN,J., STUPKA,E., PUTNAM,N., CHIA,J.M., DEHAL,P., CHRISTOFFELS,A., RASH,S., HOON,S., SMIT,A., GELPKE,M.D., ROACH,J., OH,T., HO,I.Y., WONG,M., DETTER,C., VERHOEF,F., PREDKI,P., TAY,A., LUCAS,S., RICHARDSON,P., SMITH,S.F., CLARK,M.S., EDWARDS,Y.J., DOGGETT,N., ZHARKIKH,A., TAVTIGIAN,S.V., PRUSS,D., BARNSTEAD,M., EVANS,C., BADEN,H., POWELL,J., GLUSMAN,G., ROWEN,L., HOOD,L., TAN,Y.H., ELGAR,G., HAWKINS,T., VENKATESH,B., ROKHSAR,D., BRENNER,S., 2002. "Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes", *Science*, 297, S. 1301-1310.

19. ARAVIND,L., PONTING,C.P., 1998. "Homologues of 26S proteasome subunits are regulators of transcription and translation", *Protein Sci.*, 7, S. 1250-1254.

20. ARCHER,V.E., BRETON,J., SANCHEZ-GARCIA,I., OSADA,H., FORSTER,A., THOMSON,A.J., RABBITTS,T.H., 1994. "Cysteine-rich LIM domains of LIM-homeodomain and LIM-only proteins contain zinc but not iron", *Proc.Natl.Acad.Sci.U.S.A*, 91, S. 316-320.

21. ARMES,N., GILLEY,J., FRIED,M., 1997. "The comparative genomic structure and sequence of the surfeit gene homologs in the puffer fish Fugu rubripes and their association with CpG-rich islands", *Genome Res.*, 7, S. 1138-1152.

22.ASANO,K., VORNLOCHER,H.P., RICHTER-COOK,N.J., MERRICK,W.C., HINNEBUSCH,A.G., HERSHEY,J.W., 1997. "Structure of cDNAs encoding human eukaryotic initiation factor 3 subunits. Possible roles in RNA binding and macromolecular assembly", *J.Biol.Chem.*, 272, S. 27042-27052.

- B -

23. BAILEY,J.A., CARREL,L., CHAKRAVARTI,A., EICHLER,E.E., 2000. "Molecular evidence for a relationship between LINE-1 elements and X chromosome inactivation: the Lyon repeat hypothesis", *Proc.Natl.Acad.Sci.U.S.A*, 97, S. 6634-6639.

24. BANERJEE,P., KLEYN,P.W., KNOWLES,J.A., LEWIS,C.A., ROSS,B.M., PARANO,E., KOVATS,S.G., LEE,J.J., PENCHASZADEH,G.K., OTT,J., JACOBSON,S.G., GILLIAM,T.C., 1998. "TULP1 mutation in two extended Dominican kindreds with autosomal recessive retinitis pigmentosa", *Nat.Genet.* , 18, S. 177-179.

25. BANFI,S., BORSANI,G., BULFONE,A., BALLABIO,A., 1997. "Drosophila-related expressed sequences", *Hum.Mol.Genet.*, 6, S. 1745-1753.

26. BARLOW,D.P., 1995. "Gametic imprinting in mammals", *Science*, 270, S. 1610-1613.

27. BEPLER,G., GARCIA-BLANCO,M.A., 1994. "Three tumor-suppressor regions on chromosome 11p identified by high-resolution deletion mapping in human non-small-cell lung cancer", *Proc.Natl.Acad.Sci.U.S.A*, 91, S. 5513-5517.

28. BEPLER,G., KOEHLER,A., 1995. "Multiple chromosomal aberrations and 11p allelotyping in lung cancer cell lines", *Cancer Genet.Cytogenet.*, 84, S. 39-45.

29. BERNARDI,G., OLOFSSON,B., FILIPSKI,J., ZERIAL,M., SALINAS,J., CUNY,G., MEUNIER-ROTIVAL,M., RODIER,F., 1985. "The mosaic genome of warm-blooded vertebrates", *Science*, 228, S. 953-958.

30. BERNARDI,G., 1985. "The organization of the vertebrate genome and the problem of the CpG shortage", *Prog.Clin.Biol.Res.*, 198, S. 3-10.

31. BERNARDI,G., BERNARDI,G., 1985. "Codon usage and genome composition", *J.Mol.Evol.*, 22, S. 363-365.

32. BERNARDI,G., 1995. "The human genome: organization and evolutionary history", *Annu.Rev.Genet.*, 29, S. 445-476.

33. BERNARDI,G., 2000. "Isochores and the evolutionary genomics of vertebrates", *Gene*, 241, S. 3-17.

34. BIRD,A., TAGGART,M., FROMMER,M., MILLER,O.J., MACLEOD,D., 1985. "A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA", *Cell*, 40, S. 91-99.

35. BIRD,A.P., 1980. "DNA methylation and the frequency of CpG in animal DNA", *Nucleic Acids Res.*, 8, S. 1499-1504.

36. BIRD,A.P., TAGGART,M.H., 1980. "Variable patterns of total DNA and rDNA methylation in animals", *Nucleic Acids Res.*, 8, S. 1485-1497.

37. BIRNBOIM,H.C., DOLY,J., 1979. "A rapid alkaline extraction procedure for screening recombinant plasmid DNA", *Nucleic Acids Res.*, 7, S. 1513-1523.

38. BLANC,G., BARAKAT,A., GUYOT,R., COOKE,R., DELSENY,M., 2000. "Extensive duplication and reshuffling in the Arabidopsis genome", *Plant Cell*, 12, S. 1093-1101.

39. BLATTNER,F.R., PLUNKETT,G., III, BLOCH,C.A., PERNA,N.T., BURLAND,V., RILEY,M., COLLADO-VIDES,J., GLASNER,J.D., RODE,C.K., MAYHEW,G.F., GREGOR,J., DAVIS,N.W., KIRKPATRICK,H.A., GOEDEN,M.A., ROSE,D.J., MAU,B., SHAO,Y., 1997. "The complete genome sequence of Escherichia coli K-12", *Science*, 277, S. 1453-1474.

40. BLOCK,K.L., VORNLOCHER,H.P., HERSHEY,J.W., 1998. "Characterization of cDNAs encoding the p44 and p35 subunits of human translation initiation factor eIF3", *J.Biol.Chem.*, 273, S. 31901-31908.

41. BOEHM,T., LAVENIR,I., FORSTER,A., WADEY,R.B., COWELL,J.K., HARBOTT,J., LAMPERT,F., WATERS,J., SHERRINGTON,P., COUILLIN,P., ., 1988. "The T-ALL specific t(11;14)(p13;q11) translocation breakpoint cluster region is located near to the Wilms' tumour predisposition locus", *Oncogene*, 3, S. 691-695.

42. BOEHM,T., BAER,R., LAVENIR,I., FORSTER,A., WATERS,J.J., NACHEVA,E., RABBITTS,T.H., 1988. "The mechanism of chromosomal translocation t(11;14) involving the T- cell receptor C delta locus on human chromosome 14q11 and a transcribed region of chromosome 11p15", *EMBO J.* , 7, S. 385-394.

43. BOEHM,T., FORONI,L., KENNEDY,M., RABBITTS,T.H., 1990. "The rhombotin gene belongs to a class of transcriptional regulators with a potential novel protein dimerisation motif", *Oncogene*, 5, S. 1103-1105.

44. BOEHM,T., GREENBERG,J.M., BULUWELA,L., LAVENIR,I., FORSTER,A., RABBITTS,T.H., 1990. "An unusual structure of a putative T cell oncogene which allows production of similar proteins from distinct mRNAs", *EMBO J.*, 9, S. 857-868.

45. BOEHM,T., FORONI,L., KANEKO,Y., PERUTZ,M.F., RABBITTS,T.H., 1991. "The rhombotin family of cysteine-rich LIM-domain oncogenes: distinct members are involved in T-cell translocations to human chromosomes 11p15 and 11p13", *Proc.Natl.Acad.Sci.U.S.A*, 88, S. 4367-4371.

46. BOGGON,T.J., SHAN,W.S., SANTAGATA,S., MYERS,S.C., SHAPIRO,L., 1999. "Implication of tubby proteins as transcription factors by structure- based functional analysis", *Science*, 286, S. 2119-2125.

47. BOGUSKI,M.S., 2002. "Comparative genomics: the mouse that roared", *Nature*, 420, S. 515-516.

48. BORING,C.C., SQUIRES,T.S., TONG,T., 1993. "Cancer statistics, 1993", *CA Cancer J.Clin.*, 43, S. 7-26.

49. BORK,P., BECKMANN,G., 1993. "The CUB domain. A widespread module in developmentally regulated proteins", *J.Mol.Biol.*, 231, S. 539-545.

50. BRINKMAN,R.R., MEZEI,M.M., THEILMANN,J., ALMQVIST,E., HAYDEN,M.R., 1997. "The likelihood of being affected with Huntington disease by a particular age, for a specific CAG size", *Am.J.Hum.Genet.*, 60, S. 1202-1210.

51. BUCKLER,A.J., CHANG,D.D., GRAW,S.L., BROOK,J.D., HABER,D.A., SHARP,P.A., HOUSMAN,D.E., 1991. "Exon amplification: a strategy to isolate mammalian genes based on RNA splicing", *Proc.Natl.Acad.Sci.U.S.A*, 88, S. 4005-4009.

52. BULT,C.J., WHITE,O., OLSEN,G.J., ZHOU,L., FLEISCHMANN,R.D., SUTTON,G.G., BLAKE,J.A., FITZGERALD,L.M., CLAYTON,R.A., GOCAYNE,J.D., KERLAVAGE,A.R., DOUGHERTY,B.A., TOMB,J.F., ADAMS,M.D., REICH,C.I., OVERBEEK,R., KIRKNESS,E.F., WEINSTOCK,K.G., MERRICK,J.M., GLODEK,A., SCOTT,J.L., GEOGHAGEN,N.S., VENTER,J.C., 1996. "Complete genome sequence of the methanogenic archaeon, Methanococcus jannaschii", *Science*, 273, S. 1058-1073.

53. BURGE,C., KARLIN,S., 1997. "Prediction of complete gene structures in human genomic DNA", *J.Mol.Biol.*, 268, S. 78-94.

54. BURKE,T.W., KADONAGA,J.T., 1997. "The downstream core promoter element, DPE, is conserved from Drosophila to humans and is recognized by TAFII60 of Drosophila", *Genes Dev.*, 11, S. 3020-3031.

55. BURN,J., 1999. "Closing time for CATCH22", *J.Med.Genet.*, 36, S. 737-738.

- C -

56. CANTLEY,L.C., 2001. "Transcription. Translocating tubby", *Science*, 292, S. 2019-2021.

57. CAPECCHI,M., 1990. "Gene targeting. How efficient can you get?", *Nature*, 348, S. 109.

58. CAPECCHI,M., 1990. "Gene targeting: tapping the cellular telephone", *Nature*, 344, S. 105.

59. CAPECCHI,M.R., 1989. "Altering the genome by homologous recombination", *Science*, 244, S. 1288-1292.

60. CARVER,E.A., STUBBS,L., 1997. "Zooming in on the human-mouse comparative map: genome conservation re- examined on a high-resolution scale", *Genome Res.*, 7, S. 1123-1137.

61. CAVAILLE,J., BUITING,K., KIEFMANN,M., LALANDE,M., BRANNAN,C.I., HORSTHEMKE,B., BACHELLERIE,J.P., BROSIUS,J., HUTTENHOFER,A., 2000. "Identification of brain-specific and imprinted small nucleolar RNA genes exhibiting an unusual genomic organization", *Proc.Natl.Acad.Sci.U.S.A*, 97, S. 14311-14316.

62. CHANCE,P.F., FISCHBECK,K.H., 1994. "Molecular genetics of Charcot-Marie-Tooth disease and related neuropathies", *Hum.Mol.Genet.*, 3 Spec No, S. 1503-1507.

63. CHEN,W., BOCKER,W., BROSIUS,J., TIEDGE,H., 1997. "Expression of neural BC200 RNA in human tumours", *J.Pathol.*, 183, S. 345-351.

64. CHENG,S., FOCKLER,C., BARNES,W.M., HIGUCHI,R., 1994. "Effective amplification of long targets from cloned inserts and human genomic DNA", *Proc.Natl.Acad.Sci.U.S.A*, 91, S. 5695-5699.

65. CHOMCZYNSKI,P., SACCHI,N., 1987. "Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction", *Anal.Biochem.*, 162, S. 156-159.

66. CHUMAKOV,I., RIGAULT,P., GUILLOU,S., OUGEN,P., BILLAUT,A., GUASCONI,G., GERVY,P., LEGALL,I., SOULARUE,P., GRINAS,L., ., 1992. "Continuum of overlapping clones spanning the entire human chromosome 21q", *Nature*, 359, S. 380-387.

67. CHUNG,W.K., GOLDBERG-BERMAN,J., POWER-KEHOE,L., LEIBEL,R.L., 1996. "Molecular mapping of the tubby (tub) mutation on mouse chromosome 7", *Genomics*, 32, S. 210-217.

68. CHURCH,G.M., GILBERT,W., 1984. "Genomic sequencing", *Proc.Natl.Acad.Sci.U.S.A*, 81, S. 1991-1995.

69. CICHUTEK,A., BRUECKMANN,T., SEIPEL,B., HAUSER,H., SCHLAUBITZ,S., PRAWITT,D., HANKELN,T., SCHMIDT,E.R., WINTERPACHT,A., ZABEL,B.U., 2001. "Comparative architectural aspects of regions of conserved synteny on human chromosome 11p15.3 and mouse chromosome 7 (including genes WEE1 and LMO1)", *Cytogenet.Cell Genet.*, 93, S. 277-283.

70. COCKERILL,P.N., GARRARD,W.T., 1986. "Chromosomal loop anchorage of the kappa immunoglobulin gene occurs next to the enhancer in a region containing topoisomerase II sites", *Cell*, 44, S. 273-282.

71. COLEMAN,D.L., EICHER,E.M., 1990. "Fat (fat) and tubby (tub): two autosomal recessive mutations causing obesity syndromes in the mouse", *J.Hered.*, 81, S. 424-427.

72. CONSTANCIA,M., PICKARD,B., KELSEY,G., REIK,W., 1998. "Imprinting mechanisms", *Genome Res.*, 8, S. 881-900.

73. COUTELLE,O., NYAKATURA,G., TAUDIEN,S., ELGAR,G., BRENNER,S., PLATZER,M., DRESCHER,B., JOUET,M., KENWRICK,S., ROSENTHAL,A., 1998. "The neural cell adhesion molecule L1: genomic organisation and differential splicing is conserved between man and the pufferfish Fugu", *Gene*, 208, S. 7-15.

74. CROSS,S.H., LEE,M., CLARK,V.H., CRAIG,J.M., BIRD,A.P., BICKMORE,W.A., 1997. "The chromosomal distribution of CpG islands in the mouse: evidence for genome scrambling in the rodent lineage", *Genomics*, 40, S. 454-461.

75. CYRANOSKI,D., 2002. "Almost human..", *Nature*, 418, S. 910-912.

- D -

76. D'ONOFRIO,G., MOUCHIROUD,D., AISSANI,B., GAUTIER,C., BERNARDI,G., 1991. "Correlations between the compositional properties of human genes, codon usage, and amino acid composition of proteins", *J.Mol.Evol.*, 32, S. 504-510.

77. DAHL,H.H., 1998. "Getting to the nucleus of mitochondrial disorders: identification of respiratory chain-enzyme genes causing Leigh syndrome", *Am.J.Hum.Genet.*, 63, S. 1594-1597.

78. DARNELL,J.E., JR., 1997. "STATs and gene regulation", *Science*, 277, S. 1630-1635.

79. DAVIE,J.R., 1996. "Histone modifications, chromatin structure, and the nuclear matrix", *J.Cell Biochem.*, 62, S. 149-157.

80. DE GROOT,N.G., OTTING,N., DOXIADIS,G.G., BALLA-JHAGJHOORSINGH,S.S., HEENEY,J.L., VAN ROOD,J.J., GAGNEUX,P., BONTROP,R.E., 2002. "Evidence for an ancient selective sweep in the MHC class I gene repertoire of chimpanzees", *Proc.Natl.Acad.Sci.U.S.A*, 99, S. 11748-11753.

81. DE SARIO,A., AISSANI,B., BERNARDI,G., 1991. "Compositional properties of telomeric regions from human chromosomes", *FEBS Lett.*, 295, S. 22-26.

82. DE SARIO,A., GEIGL,E.M., PALMIERI,G., D'URSO,M., BERNARDI,G., 1996. "A compositional map of human chromosome band Xq28", *Proc.Natl.Acad.Sci.U.S.A*, 93, S. 1298-1302.

83. DEININGER,P.L., 1983. "Random subcloning of sonicated DNA: application to shotgun DNA sequence analysis", *Anal.Biochem.*, 129, S. 216-223.

84. DHELLIN,O., MAESTRE,J., HEIDMANN,T., 1997. "Functional differences between the human LINE retrotransposon and retroviral reverse transcriptases for in vivo mRNA reverse transcription", *EMBO J.*, 16, S. 6590-6602.

85. DICKSON,D., 1999. "Gene estimate rises as US and UK discuss freedom of access", *Nature*, 401, S. 311.

86. DU,M., BEATTY,L.G., ZHOU,W., LEW,J., SCHOENHERR,C., WEKSBERG,R., SADOWSKI,P.D., 2003. "Insulator and silencer sequences in the imprinted region of human chromosome 11p15.5", *Hum.Mol.Genet.*, 12, S. 1927-1939.

87. DUBIEL,W., FERRELL,K., DUMDEY,R., STANDERA,S., PREHN,S., RECHSTEINER,M., 1995. "Molecular cloning and expression of subunit 12: a non-MCP and non-ATPase subunit of the 26 S protease", *FEBS Lett.*, 363, S. 97-100.

88. DUNHAM,I., SHIMIZU,N., ROE,B.A., CHISSOE,S., HUNT,A.R., COLLINS,J.E., BRUSKIEWICH,R., BEARE,D.M., CLAMP,M., SMINK,L.J., AINSCOUGH,R., ALMEIDA,J.P., BABBAGE,A., BAGGULEY,C., BAILEY,J., BARLOW,K., BATES,K.N., BEASLEY,O., BIRD,C.P., BLAKEY,S., BRIDGEMAN,A.M., BUCK,D., BURGESS,J., BURRILL,W.D., O'BRIEN,K.P., ., 1999. "The DNA sequence of human chromosome 22", *Nature*, 402, S. 489-495.

89. DURET,L., MOUCHIROUD,D., GAUTIER,C., 1995. "Statistical analysis of vertebrate sequences reveals that long genes are scarce in GC-rich isochores", *J.Mol.Evol.*, 40, S. 308-317.

- E -

90. EDDY,S.R., 1999. "Noncoding RNA genes", *Curr.Opin.Genet.Dev.*, 9, S. 695-699.

91. ELGAR,G., SANDFORD,R., APARICIO,S., MACRAE,A., VENKATESH,B., BRENNER,S., 1996. "Small is beautiful: comparative genomics with the pufferfish (Fugu rubripes)", *Trends Genet.*, 12, S. 145-150.

92. ELGAR,G., 1996. "Quality not quantity: the pufferfish genome", *Hum.Mol.Genet.*, 5 Spec No, S. 1437-1442.

93. ELGAR,G., CLARK,M.S., MEEK,S., SMITH,S., WARNER,S., EDWARDS,Y.J., BOUCHIREB,N., COTTAGE,A., YEO,G.S., UMRANIA,Y., WILLIAMS,G., BRENNER,S., 1999. "Generation and analysis of 25 Mb of genomic DNA from the pufferfish Fugu rubripes by sequence scanning", *Genome Res.*, 9, S. 960-971.

94. ELMQUIST,J.K., MARATOS-FLIER,E., SAPER,C.B., FLIER,J.S., 1998. "Unraveling the central nervous system pathways underlying responses to leptin", *Nat.Neurosci.*, 1, S. 445-450.

95. ENARD,W., PRZEWORSKI,M., FISHER,S.E., LAI,C.S., WIEBE,V., KITANO,T., MONACO,A.P., PAABO,S., 2002. "Molecular evolution of FOXP2, a gene involved in speech and language", *Nature*, 418, S. 869-872.

96. ENDO,T., IMANISHI,T., GOJOBORI,T., INOKO,H., 1997. "Evolutionary significance of intra-genome duplications on human chromosomes", *Gene*, 205, S. 19-27.

97. ENGEMANN,S., STRODICKE,M., PAULSEN,M., FRANCK,O., REINHARDT,R., LANE,N., REIK,W., WALTER,J., 2000. "Sequence and functional comparison in the Beckwith-Wiedemann region: implications for a novel imprinting centre and extended imprinting", *Hum.Mol.Genet.*, 9, S. 2691-2706.

98. EWING,B., GREEN,P., 1998. "Base-calling of automated sequencer traces using phred. II. Error probabilities", *Genome Res.*, 8, S. 186-194.

99. EWING,B., HILLIER,L., WENDL,M.C., GREEN,P., 1998. "Base-calling of automated sequencer traces using phred. I. Accuracy assessment", *Genome Res.*, 8, S. 175-185.

100. EWING,B., GREEN,P., 2000. "Analysis of expressed sequence tags indicates 35,000 human genes", *Nat.Genet.*, 25, S. 232-234.

101. EYSTEINSSON,T., JONASSON,F., JONSSON,V., BIRD,A.C., 1998. "Helicoidal peripapillary chorioretinal degeneration: electrophysiology and psychophysics in 17 patients", *Br.J.Ophthalmol.*, 82, S. 280-285.

- F -

102. FANNING,A.S., ANDERSON,J.M., 1999. "Protein modules as organizers of membrane structure", *Curr.Opin.Cell Biol.*, 11, S. 432-439.

103. FASMAN,K.H., LETOVSKY,S.I., LI,P., COTTINGHAM,R.W., KINGSBURY,D.T., 1997. "The GDB Human Genome Database Anno 1997", *Nucleic Acids Res.*, 25, S. 72-81.

104. FASSLER,J.S., GUSSIN,G.N., 1996. "Promoters and basal transcription machinery in eubacteria and eukaryotes: concepts, definitions, and analogies", *Methods Enzymol.*, 273, S. 3-29.

105. FEINBERG,A.P., VOGELSTEIN,B., 1983. "A technique for radiolabeling DNA restriction endonuclease fragments to high specific activity", *Anal.Biochem.*, 132, S. 6-13.

106. FEINBERG,A.P., 1999. "Imprinting of a genomic domain of 11p15 and loss of imprinting in cancer: an introduction", *Cancer Res.*, 59, S. 1743s-1746s.

107. FIELDS,S., SONG,O., 1989. "A novel genetic system to detect protein-protein interactions", *Nature*, 340, S. 245-246.

108. FLEISCHMANN,R.D., ADAMS,M.D., WHITE,O., CLAYTON,R.A., KIRKNESS,E.F., KERLAVAGE,A.R., BULT,C.J., TOMB,J.F., DOUGHERTY,B.A., MERRICK,J.M., ., 1995. "Whole-genome random sequencing and assembly of Haemophilus influenzae Rd", *Science*, 269, S. 496-512.

109. FLIER,J.S., MARATOS-FLIER,E., 1998. "Obesity and the hypothalamus: novel peptides for new pathways", *Cell*, 92, S. 437-440.

110. FOOTE,S., VOLLRATH,D., HILTON,A., PAGE,D.C., 1992. "The human Y chromosome: overlapping DNA clones spanning the euchromatic region", *Science*, 258, S. 60-66.

111. FORONI,L., BOEHM,T., WHITE,L., FORSTER,A., SHERRINGTON,P., LIAO,X.B., BRANNAN,C.I., JENKINS,N.A., COPELAND,N.G., RABBITTS,T.H., 1992. "The rhombotin gene family encode related LIM-domain proteins whose differing expression suggests multiple roles in mouse development", *J.Mol.Biol.*, 226, S. 747-761.

112. FORRESTER,W.C., VAN GENDEREN,C., JENUWEIN,T., GROSSCHEDL,R., 1994. "Dependence of enhancer-mediated transcription of the immunoglobulin mu gene on nuclear matrix attachment regions", *Science*, 265, S. 1221-1225.

113. FRASER,C.M., GOCAYNE,J.D., WHITE,O., ADAMS,M.D., CLAYTON,R.A., FLEISCHMANN,R.D., BULT,C.J., KERLAVAGE,A.R., SUTTON,G., KELLEY,J.M., ., 1995. "The minimal gene complement of Mycoplasma genitalium", *Science*, 270, S. 397-403.

114. FREYD,G., KIM,S.K., HORVITZ,H.R., 1990. "Novel cysteine-rich motif and homeodomain in the product of the Caenorhabditis elegans cell lineage gene lin-11", *Nature*, 344, S. 876-879.

115. FUJIYAMA,A., WATANABE,H., TOYODA,A., TAYLOR,T.D., ITOH,T., TSAI,S.F., PARK,H.S., YASPO,M.L., LEHRACH,H., CHEN,Z., FU,G., SAITOU,N., OSOEGAWA,K., DE JONG,P.J., SUTO,Y., HATTORI,M., SAKAKI,Y., 2002. "Construction and analysis of a human-chimpanzee comparative clone map", *Science*, 295, S. 131-134.

- G -

116. GARCIA-BUSTOS,J., HEITMAN,J., HALL,M.N., 1991. "Nuclear protein localization", *Biochim.Biophys.Acta*, 1071, S. 83-101.

117. GARDINER-GARDEN,M., FROMMER,M., 1987. "CpG islands in vertebrate genomes", *J.Mol.Biol.*, 196, S. 261-282.

118. GARDINER,K., 1996. "Base composition and gene distribution: critical patterns in mammalian genome organization", *Trends Genet.*, 12, S. 519-524.

119. GASSER,S.M., LAEMMLI,U.K., 1986. "Cohabitation of scaffold binding regions with upstream/enhancer elements of three developmentally regulated genes of D. melanogaster", *Cell*, 46, S. 521-530.

120. GIFFORD,R., TRISTEM,M., 2003. "The evolution, distribution and diversity of endogenous retroviruses", *Virus Genes*, 26, S. 291-315.

121. GOFFEAU,A., BARRELL,B.G., BUSSEY,H., DAVIS,R.W., DUJON,B., FELDMANN,H., GALIBERT,F., HOHEISEL,J.D., JACQ,C., JOHNSTON,M., LOUIS,E.J., MEWES,H.W., MURAKAMI,Y., PHILIPPSEN,P., TETTELIN,H., OLIVER,S.G., 1996. "Life with 6000 genes", *Science*, 274, S. 546, 563-546, 567.

122. GOODIER,J.L., OSTERTAG,E.M., DU,K., KAZAZIAN,H.H., JR., 2001. "A novel active L1 retrotransposon subfamily in the mouse", *Genome Res.*, 11, S. 1677-1685.

123. GOTTESFELD,J.M., NEELY,L., TRAUGER,J.W., BAIRD,E.E., DERVAN,P.B., 1997. "Regulation of gene expression by small molecules", *Nature*, 387, S. 202-205.

124. GREALLY,J.M., GUINNESS,M.E., MCGRATH,J., ZEMEL,S., 1997. "Matrix-attachment regions in the mouse chromosome 7F imprinted domain", *Mamm.Genome*, 8, S. 805-810.

125. GREALLY,J.M., GRAY,T.A., GABRIEL,J.M., SONG,L., ZEMEL,S., NICHOLLS,R.D., 1999. "Conserved characteristics of heterochromatin-forming DNA at the 15q11-q13 imprinting center", *Proc.Natl.Acad.Sci.U.S.A*, 96, S. 14430-14435.

126. GREEN,P., 1997. "Against a whole-genome shotgun", *Genome Res.*, 7, S. 410-417.

127. GRIMMOND,S., LARDER,R., VAN HATEREN,N., SIGGERS,P., MORSE,S., HACKER,T., ARKELL,R., GREENFIELD,A., 2001. "Expression of a novel mammalian epidermal growth factor-related gene during mouse neural development", *Mech.Dev.*, 102, S. 209-211.

128. GRUNSTEIN,M., WALLIS,J., 1979. "Colony hybridization", *Methods Enzymol.*, 68, S. 379-389.

129. GRUTZ,G.G., BUCHER,K., LAVENIR,I., LARSON,T., LARSON,R., RABBITTS,T.H., 1998. "The oncogenic T cell LIM-protein Lmo2 forms part of a DNA-binding complex specifically in immature T cells", *EMBO J.*, 17, S. 4594-4605.

- H -

130. HAGSTROM,S.A., NORTH,M.A., NISHINA,P.L., BERSON,E.L., DRYJA,T.P., 1998. "Recessive mutations in the gene encoding the tubby-like protein TULP1 in patients with retinitis pigmentosa", *Nat.Genet.*, 18, S. 174-176.

131. HAHN,S., BURATOWSKI,S., SHARP,P.A., GUARENTE,L., 1989. "Yeast TATA-binding protein TFIID binds to TATA elements with both consensus and nonconsensus DNA sequences", *Proc.Natl.Acad.Sci.U.S.A*, 86, S. 5718-5722.

132. HANCOCK,R., 2000. "A new look at the nuclear matrix", *Chromosoma*, 109, S. 219-225.

133. HARDISON,R.C., OELTJEN,J., MILLER,W., 1997. "Long human-mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome", *Genome Res.*, 7, S. 959-966.

134. HATTORI,M., FUJIYAMA,A., TAYLOR,T.D., WATANABE,H., YADA,T., PARK,H.S., TOYODA,A., ISHII,K., TOTOKI,Y., CHOI,D.K., GRONER,Y., SOEDA,E., OHKI,M., TAKAGI,T., SAKAKI,Y., TAUDIEN,S., BLECHSCHMIDT,K., POLLEY,A., MENZEL,U., DELABAR,J., KUMPF,K., LEHMANN,R., PATTERSON,D., REICHWALD,K., RUMP,A., SCHILLHABEL,M., SCHUDY,A., ZIMMERMANN,W., ROSENTHAL,A., KUDOH,J., SCHIBUYA,K., KAWASAKI,K., ASAKAWA,S., SHINTANI,A., SASAKI,T., NAGAMINE,K., MITSUYAMA,S., ANTONARAKIS,S.E., MINOSHIMA,S., SHIMIZU,N., NORDSIEK,G., HORNISCHER,K., BRANT,P., SCHARFE,M., SCHON,O., DESARIO,A., REICHELT,J., KAUER,G., BLOCKER,H., RAMSER,J., BECK,A., KLAGES,S., HENNIG,S., RIESSELMANN,L., DAGAND,E., HAAF,T., WEHRMEYER,S., BORZYM,K., GARDINER,K., NIZETIC,D., FRANCIS,F., LEHRACH,H., REINHARDT,R., YASPO,M.L., 2000. "The DNA sequence of human chromosome 21", *Nature*, 405, S. 311-319.

135. HAYASHI,K. (2003) "*PCR-SSCP: a simple and sensitive method for detection of mutations in the genomic DNA*" **1**: S.34-38

136. HE,W., IKEDA,S., BRONSON,R.T., YAN,G., NISHINA,P.M., NORTH,M.A., NAGGERT,J.K., 2000. "GFP-tagged expression and immunohistochemical studies to determine the subcellular localization of the tubby gene family members", *Brain Res.Mol.Brain Res.*, 81, S. 109-117.

137. HECKENLIVELY,J.R., CHANG,B., ERWAY,L.C., PENG,C., HAWES,N.L., HAGEMAN,G.S., RODERICK,T.H., 1995. "Mouse model for Usher syndrome: linkage mapping suggests homology to Usher type I reported at human chromosome 11p15", *Proc.Natl.Acad.Sci.U.S.A*, 92, S. 11100-11104.

138. HEDGES,S.B., KUMAR,S., 2002. "Genomics. Vertebrate genomes compared", *Science*, 297, S. 1283-1285.

139. HENGARTNER,M.O., HORVITZ,H.R., 1994. "Programmed cell death in Caenorhabditis elegans", *Curr.Opin.Genet.Dev.*, 4, S. 581-586.

140. HERBLOT,S., STEFF,A.M., HUGO,P., APLAN,P.D., HOANG,T., 2000. "SCL and LMO1 alter thymocyte differentiation: inhibition of E2A-HEB function and pre-T alpha chain expression", *Nat.Immunol.* , 1, S. 138-144.

141. HERSHEY,J.W., ASANO,K., NARANDA,T., VORNLOCHER,H.P., HANACHI,P., MERRICK,W.C., 1996. "Conservation and diversity in the structure of translation initiation factor EIF3 from humans and yeast", *Biochimie*, 78, S. 903-907.

142. HIBINO,Y., 2000. "[Functional arrangement of genomic DNA and structure of nuclear matrix]", *Yakugaku Zasshi*, 120, S. 520-533.

143. HIGGINS,M.J., SMILINICH,N.J., SAIT,S., KOENIG,A., PONGRATZ,J., GESSLER,M., RICHARD,C.W., III, JAMES,M.R., SANFORD,J.P., KIM,B.W., ., 1994. "An ordered NotI fragment map of human chromosome band 11p15", *Genomics*, 23, S. 211-222.

144. HIMMELREICH,R., HILBERT,H., PLAGENS,H., PIRKL,E., LI,B.C., HERRMANN,R., 1996. "Complete sequence analysis of the genome of the bacterium Mycoplasma pneumoniae", *Nucleic Acids Res.*, 24, S. 4420-4449.

145. HINKS,G.L., SHAH,B., FRENCH,S.J., CAMPOS,L.S., STALEY,K., HUGHES,J., SOFRONIEW,M.V., 1997. "Expression of LIM protein genes Lmo1, Lmo2, and Lmo3 in adult mouse hippocampus and other forebrain regions: differential regulation by seizure activity", *J.Neurosci.*, 17, S. 5549-5559.

146. HOOVERS,J.M., KALIKIN,L.M., JOHNSON,L.A., ALDERS,M., REDEKER,B., LAW,D.J., BLIEK,J., STEENMAN,M., BENEDICT,M., WIEGANT,J., ., 1995. "Multiple genetic loci within 11p15 defined by Beckwith-Wiedemann syndrome rearrangement breakpoints and subchromosomal transferable fragments", *Proc.Natl.Acad.Sci.U.S.A*, 92, S. 12456-12460.

147. HRABE DE ANGELIS,M.H., FLASWINKEL,H., FUCHS,H., RATHKOLB,B., SOEWARTO,D., MARSCHALL,S., HEFFNER,S., PARGENT,W., WUENSCH,K., JUNG,M., REIS,A., RICHTER,T., ALESSANDRINI,F., JAKOB,T., FUCHS,E., KOLB,H., KREMMER,E., SCHAEBLE,K., ROLLINSKI,B., ROSCHER,A., PETERS,C., MEITINGER,T., STROM,T., STECKLER,T., HOLSBOER,F., KLOPSTOCK,T., GEKELER,F., SCHINDEWOLF,C., JUNG,T., AVRAHAM,K., BEHRENDT,H., RING,J., ZIMMER,A., SCHUGHART,K., PFEFFER,K., WOLF,E., BALLING,R., 2000. "Genome-wide, large-scale production of mutant mice by ENU mutagenesis", *Nat.Genet.*, 25, S. 444-447.

148. HU,R.J., LEE,M.P., JOHNSON,L.A., FEINBERG,A.P., 1996. "A novel human homologue of yeast nucleosome assembly protein, 65 kb centromeric to the p57KIP2 gene, is biallelically expressed in fetal and adult tissues", *Hum.Mol.Genet.*, 5, S. 1743-1748.

149. HUBBARD,T., BARKER,D., BIRNEY,E., CAMERON,G., CHEN,Y., CLARK,L., COX,T., CUFF,J., CURWEN,V., DOWN,T., DURBIN,R., EYRAS,E., GILBERT,J., HAMMOND,M., HUMINIECKI,L., KASPRZYK,A., LEHVASLAIHO,H., LIJNZAAD,P., MELSOPP,C., MONGIN,E., PETTETT,R., POCOCK,M., POTTER,S., RUST,A., SCHMIDT,E., SEARLE,S., SLATER,G., SMITH,J., SPOONER,W., STABENAU,A., STALKER,J., STUPKA,E., URETA-VIDAL,A., VASTRIK,I., CLAMP,M., 2002. "The Ensembl genome database project", *Nucleic Acids Res.*, 30, S. 38-41.

150. HUBNER,N., GANTEN,D., 1995. "Genetics in arterial hypertension--clinical and experimental aspects", *Herz*, 20, S. 309-314.

151. HUET,J., SENTENAC,A., FROMAGEOT,P., 1982. "Spot-immunodetection of conserved determinants in eukaryotic RNA polymerases. Study with antibodies to yeast RNA polymerases subunits", *J.Biol.Chem.*, 257, S. 2613-2618.

- I -

152. IGARASHI,M., NAGATA,A., JINNO,S., SUTO,K., OKAYAMA,H., 1991. "Wee1(+)-like gene in human cells", *Nature*, 353, S. 80-83.

153. IKEDA,A., ZHENG,Q.Y., ROSENSTIEL,P., MADDATU,T., ZUBERI,A.R., ROOPENIAN,D.C., NORTH,M.A., NAGGERT,J.K., JOHNSON,K.R., NISHINA,P.M., 1999. "Genetic modification of hearing in tubby mice: evidence for the existence of a major gene (moth1) which protects tubby mice from hearing loss", *Hum.Mol.Genet.*, 8, S. 1761-1767.

154. IKEDA,A., IKEDA,S., GRIDLEY,T., NISHINA,P.M., NAGGERT,J.K., 2001. "Neural tube defects and neuroepithelial cell death in Tulp3 knockout mice", *Hum.Mol.Genet.*, 10, S. 1325-1334.

155. IKEDA,S., HE,W., IKEDA,A., NAGGERT,J.K., NORTH,M.A., NISHINA,P.M., 1999. "Cell-specific expression of tubby gene family members (tub, Tulp1,2, and 3) in the retina", *Invest Ophthalmol.Vis.Sci.*, 40, S. 2706-2712.

156. IKEDA,S., SHIVA,N., IKEDA,A., SMITH,R.S., NUSINOWITZ,S., YAN,G., LIN,T.R., CHU,S., HECKENLIVELY,J.R., NORTH,M.A., NAGGERT,J.K., NISHINA,P.M., DUYAO,M.P., 2000. "Retinal degeneration but not obesity is observed in null mutants of the tubby-like protein 1 gene", *Hum.Mol.Genet.*, 9, S. 155-163.

157. IOANNOU,P.A., AMEMIYA,C.T., GARNES,J., KROISEL,P.M., SHIZUYA,H., CHEN,C., BATZER,M.A., DE JONG,P.J., 1994. "A new bacteriophage P1-derived vector for the propagation of large human DNA fragments", *Nat.Genet.*, 6, S. 84-89.

- J -

158. JAENISCH,R., BIRD,A., 2003. "Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals", *Nat.Genet.*, 33 Suppl, S. 245-254.

159. JAMES,M.R., RICHARD,C.W., III, SCHOTT,J.J., YOUSRY,C., CLARK,K., BELL,J., TERWILLIGER,J.D., HAZAN,J., DUBAY,C., VIGNAL,A., ., 1994. "A radiation hybrid map of 506 STS markers spanning human chromosome 11", *Nat.Genet.*, 8, S. 70-76.

160. JENUWEIN,T., FORRESTER,W.C., FERNANDEZ-HERRERO,L.A., LAIBLE,G., DULL,M., GROSSCHEDL,R., 1997. "Extension of chromatin accessibility by nuclear matrix attachment regions", *Nature*, 385, S. 269-272.

161. JI,Y., EICHLER,E.E., SCHWARTZ,S., NICHOLLS,R.D., 2000. "Structure of chromosomal duplicons and their role in mediating human genomic disorders", *Genome Res.*, 10, S. 597-610.

162. JI,Y., REBERT,N.A., JOSLIN,J.M., HIGGINS,M.J., SCHULTZ,R.A., NICHOLLS,R.D., 2000. "Structure of the highly conserved HERC2 gene and of multiple partially duplicated paralogs in human", *Genome Res.*, 10, S. 319-329.

163. JOHNSON,K.R., MERRICK,W.C., ZOLL,W.L., ZHU,Y., 1997. "Identification of cDNA clones for the large subunit of eukaryotic translation initiation factor 3. Comparison of homologues from human, Nicotiana tabacum, Caenorhabditis elegans, and Saccharomyces cerevisiae", *J.Biol.Chem.*, 272, S. 7106-7113.

164. JONES,D.H., WINISTORFER,S.C., 1992. "Sequence specific generation of a DNA panhandle permits PCR amplification of unknown flanking DNA", *Nucleic Acids Res.*, 20, S. 595-600.

165. JONES,J.M., MEISLER,M.H., SELDIN,M.F., LEE,B.K., EICHER,E.M., 1992. "Localization of insulin-2 (Ins-2) and the obesity mutant tubby (tub) to distinct regions of mouse chromosome 7", *Genomics*, 14, S. 197-199.

166. JURKA,J., KLONOWSKI,P., DAGMAN,V., PELTON,P., 1996. "CENSOR--a program for identification and elimination of repetitive elements from DNA sequences", *Comput.Chem.*, 20, S. 119-121.

- K -

167. KANEKO,T., SATO,S., KOTANI,H., TANAKA,A., ASAMIZU,E., NAKAMURA,Y., MIYAJIMA,N., HIROSAWA,M., SUGIURA,M., SASAMOTO,S., KIMURA,T., HOSOUCHI,T., MATSUNO,A., MURAKI,A., NAKAZAKI,N., NARUO,K., OKUMURA,S., SHIMPO,S., TAKEUCHI,C., WADA,T., WATANABE,A., YAMADA,M., YASUDA,M., TABATA,S., 1996. "Sequence analysis of the genome of the unicellular cyanobacterium Synechocystis sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions", *DNA Res.*, 3, S. 109-136.

168. KAPELLER,R., MORIARTY,A., STRAUSS,A., STUBDAL,H., THERIAULT,K., SIEBERT,E., CHICKERING,T., MORGENSTERN,J.P., TARTAGLIA,L.A., LILLIE,J., 1999. "Tyrosine phosphorylation of tub and its association with Src homology 2 domain-containing proteins implicate tub in intracellular signaling by insulin", *J.Biol.Chem.*, 274, S. 24980-24986.

169. KARLSSON,O., THOR,S., NORBERG,T., OHLSSON,H., EDLUND,T., 1990. "Insulin gene enhancer binding protein Isl-1 is a member of a novel class of proteins containing both a homeo- and a Cys-His domain", *Nature*, 344, S. 879-882.

170. KASAHARA,M., HAYASHI,M., TANAKA,K., INOKO,H., SUGAYA,K., IKEMURA,T., ISHIBASHI,T., 1996. "Chromosomal localization of the proteasome Z subunit gene reveals an ancient chromosomal duplication involving the major histocompatibility complex", *Proc.Natl.Acad.Sci.U.S.A*, 93, S. 9096-9101.

171. KENNY,D.A., JURATA,L.W., SAGA,Y., GILL,G.N., 1998. "Identification and characterization of LMO4, an LMO gene with a novel pattern of expression during embryogenesis", *Proc.Natl.Acad.Sci.U.S.A*, 95, S. 11257-11262.

172. KIM,U.J., BIRREN,B.W., SLEPAK,T., MANCINO,V., BOYSEN,C., KANG,H.L., SIMON,M.I., SHIZUYA,H., 1996. "Construction and characterization of a human bacterial artificial chromosome library", *Genomics*, 34, S. 213-218.

173. KISS,T., FILIPOWICZ,W., 1995. "Exonucleolytic processing of small nucleolar RNAs from pre-mRNA introns", *Genes Dev.*, 9, S. 1411-1424.

174. KLEYN,P.W., FAN,W., KOVATS,S.G., LEE,J.J., PULIDO,J.C., WU,Y., BERKEMEIER,L.R., MISUMI,D.J., HOLMGREN,L., CHARLAT,O., WOOLF,E.A., TAYBER,O., BRODY,T., SHU,P., HAWKINS,F., KENNEDY,B., BALDINI,L., EBELING,C., ALPERIN,G.D., DEEDS,J., LAKEY,N.D., CULPEPPER,J., CHEN,H., GLUCKSMANN-KUIS,M.A., MOORE,K.J., ., 1996. "Identification and characterization of the mouse obesity gene tubby: a member of a novel gene family", *Cell*, 85, S. 281-290.

175. KNOTT,T.J., WALLIS,S.C., PEASE,R.J., POWELL,L.M., SCOTT,J., 1986. "A hypervariable region 3' to the human apolipoprotein B gene", *Nucleic Acids Res.*, 14, S. 9215-9216.

176. KOENIG,M., HOFFMAN,E.P., BERTELSON,C.J., MONACO,A.P., FEENER,C., KUNKEL,L.M., 1987. "Complete cloning of the Duchenne muscular dystrophy (DMD) cDNA and preliminary genomic organization of the DMD gene in normal and affected individuals", *Cell*, 50, S. 509-517.

177. KOOP,B.F., HOOD,L., 1994. "Striking sequence similarity over almost 100 kilobases of human and mouse T-cell receptor DNA", *Nat.Genet.*, 7, S. 48-53.

178. KOOP,B.F., 1995. "Human and rodent DNA sequence comparisons: a mosaic model of genomic evolution", *Trends Genet.*, 11, S. 367-371.

179. KORITSCHONER,N.P., ALVAREZ-DOLADO,M., KURZ,S.M., HEIKENWALDER,M.F., HACKER,C., VOGEL,F., MUNOZ,A., ZENKE,M., 2001. "Thyroid hormone regulates the obesity gene tub", *EMBO Rep.*, 2, S. 499-504.

180. KOZAK,M., 1991. "An analysis of vertebrate mRNA sequences: intimations of translational control", *J.Cell Biol.*, 115, S. 887-903.

181. KOZAK,M., 1999. "Initiation of translation in prokaryotes and eukaryotes", *Gene*, 234, S. 187-208.

182. KRUGER,W.D., COX,D.R., 1995. "A yeast assay for functional detection of mutations in the human cystathionine beta-synthase gene", *Hum.Mol.Genet.*, 4, S. 1155-1161.

183. KUHN,R., SCHWENK,F., AGUET,M., RAJEWSKY,K., 1995. "Inducible gene targeting in mice", *Science*, 269, S. 1427-1429.

184. KUWABARA,P.E., COULSON,A., 2000. "RNAi--prospects for a general technique for determining gene function", *Parasitol.Today*, 16, S. 347-349.

- L -

185. LAI,C.S., FISHER,S.E., HURST,J.A., VARGHA-KHADEM,F., MONACO,A.P., 2001. "A forkhead-domain gene is mutated in a severe speech and language disorder", *Nature*, 413, S. 519-523.

186. LANDSCHULZ,W.H., JOHNSON,P.F., MCKNIGHT,S.L., 1988. "The leucine zipper: a hypothetical structure common to a new class of DNA binding proteins", *Science*, 240, S. 1759-1764.

187. LARSEN,F., GUNDERSEN,G., LOPEZ,R., PRYDZ,H., 1992. "CpG islands as gene markers in the human genome", *Genomics*, 13, S. 1095-1107.

188. LARSON,R.C., LAVENIR,I., LARSON,T.A., BAER,R., WARREN,A.J., WADMAN,I., NOTTAGE,K., RABBITTS,T.H., 1996. "Protein dimerization between Lmo2 (Rbtn2) and Tal1 alters thymocyte development and potentiates T cell tumorigenesis in transgenic mice", *EMBO J.*, 15, S. 1021-1027.

189. LEE,G.H., PROENCA,R., MONTEZ,J.M., CARROLL,K.M., DARVISHZADEH,J.G., LEE,J.I., FRIEDMAN,J.M., 1996. "Abnormal splicing of the leptin receptor in diabetic mice", *Nature*, 379, S. 632-635.

190. LEE,L.G., CONNELL,C.R., WOO,S.L., CHENG,R.D., MCARDLE,B.F., FULLER,C.W., HALLORAN,N.D., WILSON,R.K., 1992. "DNA sequencing with dye-labeled terminators and T7 DNA polymerase: effect of dyes and dNTPs on incorporation of dye-terminators and probability analysis of termination fragments", *Nucleic Acids Res.*, 20, S. 2471-2483.

191. LEE,M.P., BRANDENBURG,S., LANDES,G.M., ADAMS,M., MILLER,G., FEINBERG,A.P., 1999. "Two novel genes in the center of the 11p15 imprinted domain escape genomic imprinting", *Hum.Mol.Genet.*, 8, S. 683-690.

192. LEMIEUX,N., DUTRILLAUX,B., VIEGAS-PEQUIGNOT,E., 1992. "A simple method for simultaneous R- or G-banding and fluorescence in situ hybridization of small single-copy genes", *Cytogenet.Cell Genet.*, 59, S. 311-312.

193. LENNON,G., AUFFRAY,C., POLYMEROPOULOS,M., SOARES,M.B., 1996. "The I.M.A.G.E. Consortium: an integrated molecular analysis of genomes and their expression", *Genomics*, 33, S. 151-152.

194. LEPPERT,M., BAIRD,L., ANDERSON,K.L., OTTERUD,B., LUPSKI,J.R., LEWIS,R.A., 1994. "Bardet-Biedl syndrome is linked to DNA markers on chromosome 11q and is genetically heterogeneous", *Nat.Genet.*, 7, S. 108-112.

195. LI,E., BEARD,C., JAENISCH,R., 1993. "Role for DNA methylation in genomic imprinting", *Nature*, 366, S. 362-365.

196. LI,W.H., HIDE,W.A., GRAUR,D., 1992. "Origin of rodents and guinea-pigs", *Nature*, 359, S. 277-278.

197. LICHTER,P., CREMER,T., BORDEN,J., MANUELIDIS,L., WARD,D.C., 1988. "Delineation of individual human chromosomes in metaphase and interphase cells by in situ suppression hybridization using recombinant DNA libraries", *Hum.Genet.*, 80, S. 224-234.

198. LICHY,J.H., MODI,W.S., SEUANEZ,H.N., HOWLEY,P.M., 1992. "Identification of a human chromosome 11 gene which is differentially regulated in tumorigenic and nontumorigenic somatic cell hybrids of HeLa cells", *Cell Growth Differ.*, 3, S. 541-548.

199. LICHY,J.H., MAJIDI,M., ELBAUM,J., TSAI,M.M., 1996. "Differential expression of the human ST5 gene in HeLa-fibroblast hybrid cell lines mediated by YY1: evidence that YY1 plays a part in tumor suppression", *Nucleic Acids Res.*, 24, S. 4700-4708.

200. LOOTS,G.G., LOCKSLEY,R.M., BLANKESPOOR,C.M., WANG,Z.E., MILLER,W., RUBIN,E.M., FRAZER,K.A., 2000. "Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons", *Science*, 288, S. 136-140.

201. LUPSKI,J.R., 1998. "Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits", *Trends Genet.*, 14, S. 417-422.

202. LYNCH,J.W., EVERSON,S.A., KAPLAN,G.A., SALONEN,R., SALONEN,J.T., 1998. "Does low socioeconomic status potentiate the effects of heightened cardiovascular responses to stress on the progression of carotid atherosclerosis?", *Am.J.Public Health*, 88, S. 389-394.

- M -

203. MAGENIS,R.E., MASLEN,C.L., SMITH,L., ALLEN,L., SAKAI,L.Y., 1991. "Localization of the fibrillin (FBN) gene to chromosome 15, band q21.1", *Genomics*, 11, S. 346-351.

204. MAKALOWSKI,W., ZHANG,J., BOGUSKI,M.S., 1996. "Comparative analysis of 1196 orthologous mouse and human full-length mRNA and protein sequences", *Genome Res.*, 6, S. 846-857.

205. MAKALOWSKI,W., BOGUSKI,M.S., 1998. "Evolutionary parameters of the transcribed mammalian genome: an analysis of 2,820 orthologous rodent and human sequences", *Proc.Natl.Acad.Sci.U.S.A*, 95, S. 9407-9412.

206. MALLON,A.M., PLATZER,M., BATE,R., GLOECKNER,G., BOTCHERBY,M.R., NORDSIEK,G., STRIVENS,M.A., KIOSCHIS,P., DANGEL,A., CUNNINGHAM,D., STRAW,R.N., WESTON,P., GILBERT,M., FERNANDO,S., GOODALL,K., HUNTER,G., GREYSTRONG,J.S., CLARKE,D., KIMBERLEY,C., GOERDES,M., BLECHSCHMIDT,K., RUMP,A., HINZMANN,B., MUNDY,C.R., MILLER,W., POUSTKA,A., HERMAN,G.E., RHODES,M., DENNY,P., ROSENTHAL,A., BROWN,S.D., 2000. "Comparative genome sequence analysis of the Bpa/Str region in mouse and Man", *Genome Res.*, 10, S. 758-775.

207. MANGIN,M., WEBB,A.C., DREYER,B.E., POSILLICO,J.T., IKEDA,K., WEIR,E.C., STEWART,A.F., BANDER,N.H., MILSTONE,L., BARTON,D.E., ., 1988. "Identification of a cDNA encoding a parathyroid hormone-like peptide from a human tumor associated with humoral hypercalcemia of malignancy", *Proc.Natl.Acad.Sci.U.S.A*, 85, S. 597-601.

208. MANNENS,M., SLATER,R.M., HEYTING,C., BLIEK,J., DE KRAKER,J., COAD,N., PAGTER-HOLTHUIZEN,P., PEARSON,P.L., 1988. "Molecular nature of genetic changes resulting in loss of heterozygosity of chromosome 11 in Wilms' tumours", *Hum.Genet.*, 81, S. 41-48.

209. MANNENS,M., HOOVERS,J.M., REDEKER,E., VERJAAL,M., FEINBERG,A.P., LITTLE,P., BOAVIDA,M., COAD,N., STEENMAN,M., BLIEK,J., ., 1994. "Parental imprinting of human chromosome region 11p15.3-pter involved in the Beckwith-Wiedemann syndrome and various human neoplasia", *Eur.J.Hum.Genet.*, 2, S. 3-23.

210. MARMOT,M.G., SYME,S.L., KAGAN,A., KATO,H., COHEN,J.B., BELSKY,J., 1975. "Epidemiologic studies of coronary heart disease and stroke in Japanese men living in Japan, Hawaii and California: prevalence of coronary and hypertensive heart disease and associated risk factors", *Am.J.Epidemiol.*, 102, S. 514-525.

211. MATSUO,K., CLAY,O., TAKAHASHI,T., SILKE,J., SCHAFFNER,W., 1993. "Evidence for erosion of mouse CpG islands during mammalian evolution", *Somat.Cell Mol.Genet.*, 19, S. 543-555.

212. MCCREADY,S.J., COOK,P.R., 1984. "Lesions induced in DNA by ultraviolet light are repaired at the nuclear cage", *J.Cell Sci.*, 70, S. 189-196.

213. MCGUIRE,E.A., HOCKETT,R.D., POLLOCK,K.M., BARTHOLDI,M.F., O'BRIEN,S.J., KORSMEYER,S.J., 1989. "The t(11;14)(p15;q11) in a T-cell acute lymphoblastic leukemia cell line activates multiple transcripts, including Ttg-1, a gene encoding a potential zinc finger protein", *Mol.Cell Biol.*, 9, S. 2124-2132.

214. MCKUSICK,V.A., 1995. "Reviews in molecular medicine", *Medicine (Baltimore)*, 74, S. 301-304.

215. METHOT,N., ROM,E., OLSEN,H., SONENBERG,N., 1997. "The human homologue of the yeast Prt1 protein is an integral part of the eukaryotic initiation factor 3 complex and interacts with p170", *J.Biol.Chem.*, 272, S. 1110-1116.

216. MICHELLAND,S., GAZZERI,S., BRAMBILLA,E., ROBERT-NICOUD,M., 1999. "Comparison of chromosomal imbalances in neuroendocrine and non-small-cell lung carcinomas", *Cancer Genet.Cytogenet.*, 114, S. 22-30.

217. MIGHELL,A.J., SMITH,N.R., ROBINSON,P.A., MARKHAM,A.F., 2000. "Vertebrate pseudogenes", *FEBS Lett.*, 468, S. 109-114.

218. MIRONOV,A.A., FICKETT,J.W., GELFAND,M.S., 1999. "Frequent alternative splicing of human genes", *Genome Res.*, 9, S. 1288-1293.

219. MITSUYA,K., MEGURO,M., LEE,M.P., KATOH,M., SCHULZ,T.C., KUGOH,H., YOSHIDA,M.A., NIIKAWA,N., FEINBERG,A.P., OSHIMURA,M., 1999. "LIT1, an imprinted antisense RNA in the human KvLQT1 locus identified by screening for differentially expressed transcripts using monochromosomal hybrids", *Hum.Mol.Genet.*, 8, S. 1209-1217.

220. MONTAGUTELLI,X., 2000. "Effect of the genetic background on the phenotype of mouse mutations", *J.Am.Soc.Nephrol.*, 11 Suppl 16, S. S101-S105.

221. MOORE,A., 2001. "Of mice and Mendel. The predicted rise in the use of knock-out and transgenic mice should cause us to reflect on our justification for the use of animals in research", *EMBO Rep.*, 2, S. 554-558.

222. MOUCHIROUD,D., GAUTIER,C., 1990. "Codon usage changes and sequence dissimilarity between human and rat", *J.Mol.Evol.*, 31, S. 81-91.

223. MOUCHIROUD,D., D'ONOFRIO,G., AISSANI,B., MACAYA,G., GAUTIER,C., BERNARDI,G., 1991. "The distribution of genes in the human genome", *Gene*, 100, S. 181-187.

224. MUJICA,A.O., HANKELN,T., SCHMIDT,E.R., 2001. "A novel serine/threonine kinase gene, STK33, on human chromosome 11p15.3", *Gene*, 280, S. 175-181.

225. MURRE,C., MCCAW,P.S., VAESSIN,H., CAUDY,M., JAN,L.Y., JAN,Y.N., CABRERA,C.V., BUSKIN,J.N., HAUSCHKA,S.D., LASSAR,A.B., ., 1989. "Interactions between heterologous helix-loop-helix proteins generate complexes that bind specifically to a common DNA sequence", *Cell*, 58, S. 537-544.

226. MUSHEGIAN,A.R., GAREY,J.R., MARTIN,J., LIU,L.X., 1998. "Large-scale taxonomic profiling of eukaryotic model organisms: a comparison of orthologous proteins encoded by the human, fly, nematode, and yeast genomes", *Genome Res.*, 8, S. 590-598.

227. MYERS,E.W., SUTTON,G.G., DELCHER,A.L., DEW,I.M., FASULO,D.P., FLANIGAN,M.J., KRAVITZ,S.A., MOBARRY,C.M., REINERT,K.H., REMINGTON,K.A., ANSON,E.L., BOLANOS,R.A., CHOU,H.H., JORDAN,C.M., HALPERN,A.L., LONARDI,S., BEASLEY,E.M., BRANDON,R.C., CHEN,L., DUNN,P.J., LAI,Z., LIANG,Y., NUSSKERN,D.R., ZHAN,M., ZHANG,Q., ZHENG,X., RUBIN,G.M., ADAMS,M.D., VENTER,J.C., 2000. "A whole-genome assembly of Drosophila", *Science*, 287, S. 2196-2204.

- N -

228. NABOZNY,G.H., BAISCH,J.M., CHENG,S., COSGROVE,D., GRIFFITHS,M.M., LUTHRA,H.S., DAVID,C.S., 1996. "HLA-DQ8 transgenic mice are highly susceptible to collagen-induced arthritis: a novel model for human polyarthritis", *J.Exp.Med.*, 183, S. 27-37.

229. NADEAU,J.H., 2001. "Modifier genes in mice and humans", *Nat.Rev.Genet.*, 2, S. 165-174.

230. NAKASHIMA,K., NARAZAKI,M., TAGA,T., 1997. "Leptin receptor (OB-R) oligomerizes with itself but not with its closely related cytokine signal transducer gp130", *FEBS Lett.*, 403, S. 79-82.

231. NEUMANN,B., KUBICKA,P., BARLOW,D.P., 1995. "Characteristics of imprinted genes", *Nat.Genet.*, 9, S. 12-13.

232. NEYROUD,N., TESSON,F., DENJOY,I., LEIBOVICI,M., DONGER,C., BARHANIN,J., FAURE,S., GARY,F., COUMEL,P., PETIT,C., SCHWARTZ,K., GUICHENEY,P., 1997. "A novel mutation in the potassium channel gene KVLQT1 causes the Jervell and Lange-Nielsen cardioauditory syndrome", *Nat.Genet.*, 15, S. 186-189.

233. NISHINA,P.M., NORTH,M.A., IKEDA,A., YAN,Y., NAGGERT,J.K., 1998. "Molecular characterization of a novel tubby gene family member, TULP3, in mouse and humans", *Genomics*, 54, S. 215-220.

234. NOBEN-TRAUTH,K., NAGGERT,J.K., NORTH,M.A., NISHINA,P.M., 1996. "A candidate gene for the mouse mutation tubby", *Nature*, 380, S. 534-538.

235. NOLAN,P.M., PETERS,J., STRIVENS,M., ROGERS,D., HAGAN,J., SPURR,N., GRAY,I.C., VIZOR,L., BROOKER,D., WHITEHILL,E., WASHBOURNE,R., HOUGH,T., GREENAWAY,S., HEWITT,M., LIU,X., MCCORMACK,S., PICKFORD,K., SELLEY,R., WELLS,C., TYMOWSKA-LALANNE,Z., ROBY,P., GLENISTER,P., THORNTON,C., THAUNG,C., STEVENSON,J.A., ARKELL,R., MBURU,P., HARDISTY,R., KIERNAN,A., ERVEN,A., STEEL,K.P., VOEGELING,S., GUENET,J.L., NICKOLS,C., SADRI,R., NASSE,M., ISAACS,A., DAVIES,K., BROWNE,M., FISHER,E.M., MARTIN,J., RASTAN,S., BROWN,S.D., HUNTER,J., 2000. "A systematic, genome-wide, phenotype-driven mutagenesis programme for gene function studies in the mouse", *Nat.Genet.*, 25, S. 440-443.

236. NORMAN,R.A., BOGARDUS,C., RAVUSSIN,E., 1995. "Linkage between obesity and a marker near the tumor necrosis factor- alpha locus in Pima Indians", *J.Clin.Invest*, 96, S. 158-162.

237. NORTH,M.A., NAGGERT,J.K., YAN,Y., NOBEN-TRAUTH,K., NISHINA,P.M., 1997. "Molecular characterization of TUB, TULP1, and TULP2, members of the novel tubby gene family and their possible relation to ocular diseases", *Proc.Natl.Acad.Sci.U.S.A*, 94, S. 3128-3133.

- O -

238. OELTJEN,J.C., MALLEY,T.M., MUZNY,D.M., MILLER,W., GIBBS,R.A., BELMONT,J.W., 1997. "Large-scale comparative sequence analysis of the human and murine Bruton's tyrosine kinase loci reveals conserved regulatory domains", *Genome Res.*, 7, S. 315-329.

239. OH,Y., PROCTOR,M.L., FAN,Y.H., SU,L.K., HONG,W.K., FONG,K.M., SEKIDO,Y.S., GAZDAR,A.F., MINNA,J.D., MAO,L., 1998. "TSG101 is not mutated in lung cancer but a shortened transcript is frequently expressed in small cell lung cancer", *Oncogene*, 17, S. 1141-1148.

240. OHLEMILLER,K.K., HUGHES,R.M., MOSINGER-OGILVIE,J., SPECK,J.D., GROSOF,D.H., SILVERMAN,M.S., 1995. "Cochlear and retinal degeneration in the tubby mouse", *Neuroreport*, 6, S. 845-849.

241. OHLEMILLER,K.K., HUGHES,R.M., LETT,J.M., OGILVIE,J.M., SPECK,J.D., WRIGHT,J.S., FADDIS,B.T., 1997. "Progression of cochlear and retinal degeneration in the tubby (rd5) mouse", *Audiol.Neurootol.*, 2, S. 175-185.

242. OHLSSON,R., RENKAWITZ,R., LOBANENKOV,V., 2001. "CTCF is a uniquely versatile transcription regulator linked to epigenetics and disease", *Trends Genet.*, 17, S. 520-527.

243. OHTSUKA,M., MAKINO,S., YODA,K., WADA,H., NARUSE,K., MITANI,H., SHIMA,A., OZATO,K., KIMURA,M., INOKO,H., 1999. "Construction of a linkage map of the medaka (Oryzias latipes) and mapping of the Da mutant locus defective in dorsoventral patterning", *Genome Res.*, 9, S. 1277-1287.

244. OLLOMO,B., KARCH,S., BUREAU,P., ELISSA,N., GEORGES,A.J., MILLET,P., 1997. "Lack of malaria parasite transmission between apes and humans in Gabon", *Am.J.Trop.Med.Hyg.*, 56, S. 440-445.

245. ONYANGO,P., MILLER,W., LEHOCZKY,J., LEUNG,C.T., BIRREN,B., WHEELAN,S., DEWAR,K., FEINBERG,A.P., 2000. "Sequence and comparative analysis of the mouse 1-megabase region orthologous to the human 11p15 imprinted domain", *Genome Res.*, 10, S. 1697-1710.

- P -

246. PARANJAPE,S.M., KAMAKAKA,R.T., KADONAGA,J.T., 1994. "Role of chromatin structure in the regulation of transcription by RNA polymerase II", *Annu.Rev.Biochem.*, 63, S. 265-297.

247. PAULSEN,M., EL MAARRI,O., ENGEMANN,S., STRODICKE,M., FRANCK,O., DAVIES,K., REINHARDT,R., REIK,W., WALTER,J., 2000. "Sequence conservation and variability of imprinting in the Beckwith-Wiedemann syndrome gene cluster in human and mouse", *Hum.Mol.Genet.*, 9, S. 1829-1841.

248. PAULSEN,M., FERGUSON-SMITH,A.C., 2001. "DNA methylation in genomic imprinting, development, and disease", *J.Pathol.*, 195, S. 97-110.

249. PEDERSEN,A.G., BALDI,P., CHAUVIN,Y., BRUNAK,S., 1999. "The biology of eukaryotic promoter prediction--a review", *Comput.Chem.*, 23, S. 191-207.

250. PERNA,N.T., PLUNKETT,G., III, BURLAND,V., MAU,B., GLASNER,J.D., ROSE,D.J., MAYHEW,G.F., EVANS,P.S., GREGOR,J., KIRKPATRICK,H.A., POSFAI,G., HACKETT,J., KLINK,S., BOUTIN,A., SHAO,Y., MILLER,L., GROTBECK,E.J., DAVIS,N.W., LIM,A., DIMALANTA,E.T., POTAMOUSIS,K.D., APODACA,J., ANANTHARAMAN,T.S., LIN,J., YEN,G., SCHWARTZ,D.C., WELCH,R.A., BLATTNER,F.R., 2001. "Genome sequence of enterohaemorrhagic Escherichia coli O157:H7", *Nature*, 409, S. 529-533.

251. PETTENATI,M.J., HAINES,J.L., HIGGINS,R.R., WAPPNER,R.S., PALMER,C.G., WEAVER,D.D., 1986. "Wiedemann-Beckwith syndrome: presentation of clinical and cytogenetic data on 22 new cases and review of the literature", *Hum.Genet.*, 74, S. 143-154.

252. PRESTRIDGE,D.S., 1995. "Predicting Pol II promoter sequences using transcription factor binding sites", *J.Mol.Biol.*, 249, S. 923-932.

253. PUTILINA,T., JAWORSKI,C., GENTLEMAN,S., MCDONALD,B., KADIRI,M., WONG,P., 1998. "Analysis of a human cDNA containing a tissue-specific alternatively spliced LIM domain", *Biochem.Biophys.Res.Commun.*, 252, S. 433-439.

- R -

254. RABBITTS,T. (2002) "*Rhom-3, a gene which is highly related to rhombotin and located on chromosome 12p13*"

255. RABBITTS,T.H., AXELSON,H., FORSTER,A., GRUTZ,G., LAVENIR,I., LARSON,R., OSADA,H., VALGE-ARCHER,V., WADMAN,I., WARREN,A., 1997. "Chromosomal translocations and leukaemia: a role for LMO2 in T cell acute leukaemia, in transcription and in erythropoiesis", *Leukemia*, 11 Suppl 3, S. 271-272.

256. RABBITTS,T.H., BUCHER,K., CHUNG,G., GRUTZ,G., WARREN,A., YAMADA,Y., 1999. "The effect of chromosomal translocations in acute leukemias: the LMO2 paradigm in transcription and development", *Cancer Res.*, 59, S. 1794s-1798s.

257. RAJEWSKY,K., GU,H., KUHN,R., BETZ,U.A., MULLER,W., ROES,J., SCHWENK,F., 1996. "Conditional gene targeting", *J.Clin.Invest*, 98, S. 600-603.

258. REDEKER,E., HOOVERS,J.M., ALDERS,M., VAN MOORSEL,C.J., IVENS,A.C., GREGORY,S., KALIKIN,L., BLIEK,J., DE GALAN,L., VAN DEN,B.R., ., 1994. "An integrated physical map of 210 markers assigned to the short arm of human chromosome 11", *Genomics*, 21, S. 538-550.

259. REDEKER,E., ALDERS,M., HOOVERS,J.M., RICHARD,C.W., III, WESTERVELD,A., MANNENS,M., 1995. "Physical mapping of 3 candidate tumor suppressor genes relative to Beckwith-Wiedemann syndrome associated chromosomal breakpoints at 11p15.3", *Cytogenet.Cell Genet.*, 68, S. 222-225.

260. REDEKER,V., TOULLEC,J.Y., VINH,J., ROSSIER,J., SOYEZ,D., 1998. "Combination of peptide profiling by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry and immunodetection on single glands or cells", *Anal.Chem.*, 70, S. 1805-1811.

261. REIK,W., ALLEN,N.D., 1994. "Genomic imprinting. Imprinting with and without methylation", *Curr.Biol.*, 4, S. 145-147.

262. ROBERTSON,H.M., LAMPE,D.J., 1995. "Recent horizontal transfer of a mariner transposable element among and between Diptera and Neuroptera", *Mol.Biol.Evol.*, 12, S. 850-862.

263. ROYER-POKORA,B., LOOS,U., LUDWIG,W.D., 1991. "TTG-2, a new gene encoding a cysteine-rich protein with the LIM motif, is overexpressed in acute T-cell leukaemia with the t(11;14)(p13;q11)", *Oncogene*, 6, S. 1887-1893.

264. ROYER-POKORA,B., ROGERS,M., ZHU,T.H., SCHNEIDER,S., LOOS,U., BOLITZ,U., 1995. "The TTG-2/RBTN2 T cell oncogene encodes two alternative transcripts from two promoters: the distal promoter is removed by most 11p13 translocations in acute T cell leukaemia's (T-ALL)", *Oncogene*, 10, S. 1353-1360.

265. RUBIN,G.M., YANDELL,M.D., WORTMAN,J.R., GABOR MIKLOS,G.L., NELSON,C.R., HARIHARAN,I.K., FORTINI,M.E., LI,P.W., APWEILER,R., FLEISCHMANN,W., CHERRY,J.M., HENIKOFF,S., SKUPSKI,M.P., MISRA,S., ASHBURNER,M., BIRNEY,E., BOGUSKI,M.S., BRODY,T., BROKSTEIN,P., CELNIKER,S.E., CHERVITZ,S.A., COATES,D., CRAVCHIK,A., GABRIELIAN,A., GALLE,R.F., GELBART,W.M., GEORGE,R.A., GOLDSTEIN,L.S., GONG,F., GUAN,P., HARRIS,N.L., HAY,B.A., HOSKINS,R.A., LI,J., LI,Z., HYNES,R.O., JONES,S.J., KUEHL,P.M., LEMAITRE,B., LITTLETON,J.T., MORRISON,D.K., MUNGALL,C., O'FARRELL,P.H., PICKERAL,O.K., SHUE,C., VOSSHALL,L.B., ZHANG,J., ZHAO,Q., ZHENG,X.H., LEWIS,S., 2000. "Comparative genomics of the eukaryotes", *Science*, 287, S. 2204-2215.

- S -

266. SACCONE,S., DE SARIO,A., DELLA,V.G., BERNARDI,G., 1992. "The highest gene concentrations in the human genome are in telomeric bands of metaphase chromosomes", *Proc.Natl.Acad.Sci.U.S.A*, 89, S. 4913-4917.

267. SACCONE,S., DE SARIO,A., WIEGANT,J., RAAP,A.K., DELLA,V.G., BERNARDI,G., 1993. "Correlations between isochores and chromosomal bands in the human genome", *Proc.Natl.Acad.Sci.U.S.A*, 90, S. 11929-11933.

268. SAHLY,I., GOGAT,K., KOBETZ,A., MARCHANT,D., MENASCHE,M., CASTEL,M., REVAH,F., DUFIER,J., GUERRE-MILLO,M., ABITBOL,M.M., 1998. "Prominent neuronal-specific tub gene expression in cellular targets of tubby mice mutation", *Hum.Mol.Genet.*, 7, S. 1437-1447.

269. SAIKI,R.K., GELFAND,D.H., STOFFEL,S., SCHARF,S.J., HIGUCHI,R., HORN,G.T., MULLIS,K.B., ERLICH,H.A., 1988. "Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase", *Science*, 239, S. 487-491.

270. SAMBROOK,J.; FRITSCH,E.F.; MANIATIS,T. (1989) "*Molecular Cloning: A Laboratory Manual.*"

271. SANCHEZ-GARCIA,I., RABBITTS,T.H., 1994. "The LIM domain: a new structural motif found in zinc-finger-like proteins", *Trends Genet.*, 10, S. 315-320.

272. SANGER,F., NICKLEN,S., COULSON,A.R., 1977. "DNA sequencing with chain-terminating inhibitors", *Proc.Natl.Acad.Sci.U.S.A*, 74, S. 5463-5467.

273. SANTAGATA,S., BOGGON,T.J., BAIRD,C.L., GOMEZ,C.A., ZHAO,J., SHAN,W.S., MYSZKA,D.G., SHAPIRO,L., 2001. "G-protein signaling through tubby proteins", *Science*, 292, S. 2041-2050.

274. SCHMEICHEL,K.L., BECKERLE,M.C., 1994. "The LIM domain is a modular protein-binding interface", *Cell*, 79, S. 211-219.

275. SCHOTLAND,H.M., ELDRIDGE,R., SOMMER,S.S., MALAWAR,M., 1992. "Neurofibromatosis 1 and osseous fibrous dysplasia in a family", *Am.J.Med.Genet.*, 43, S. 815-822.

276. SCHULER,G.D., 1997. "Pieces of the puzzle: expressed sequence tags and the catalog of human genes", *J.Mol.Med.*, 75, S. 694-698.

277. SCHWARTZ,S., ZHANG,Z., FRAZER,K.A., SMIT,A., RIEMER,C., BOUCK,J., GIBBS,R., HARDISON,R., MILLER,W., 2000. "PipMaker--a web server for aligning two genomic DNA sequences", *Genome Res.*, 10, S. 577-586.

278. SEIZINGER,B.R., MARTUZA,R.L., GUSELLA,J.F., 1986. "Loss of genes on chromosome 22 in tumorigenesis of human acoustic neuroma", *Nature*, 322, S. 644-647.

279. SHARP,P.A., 1994. "Split genes and RNA splicing", *Cell*, 77, S. 805-815.

280. SHEN,L., TSUCHIDA,R., MIYAUCHI,J., SAEKI,M., HONNA,T., TSUNEMATSU,Y., KATO,J., MIZUTANI,S., 2000. "Differentiation-associated expression and intracellular localization of cyclin-dependent kinase inhibitor p27KIP1 and c-Jun co-activator JAB1 in neuroblastoma", *Int.J.Oncol.*, 17, S. 749-754.

281. SHENG,Y., MANCINO,V., BIRREN,B., 1995. "Transformation of Escherichia coli with large DNA molecules by electroporation", *Nucleic Acids Res.*, 23, S. 1990-1996.

282. SHERRY,S.T., WARD,M., SIROTKIN,K., 1999. "dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation", *Genome Res.*, 9, S. 677-679.

283. SHI,J., CAI,W., CHEN,X., YING,K., ZHANG,K., XIE,Y., 2001. "Identification of dopamine responsive mRNAs in glial cells by suppression subtractive hybridization", *Brain Res.*, 910, S. 29-37.

284. SIDOW,A., 1996. "Gen(om)e duplications in the evolution of early vertebrates", *Curr.Opin.Genet.Dev.*, 6, S. 715-722.

285. SIEGFRIED,Z., EDEN,S., MENDELSOHN,M., FENG,X., TSUBERI,B.Z., CEDAR,H., 1999. "DNA methylation represses transcription in vivo", *Nat.Genet.*, 22, S. 203-206.

286. SMALE,S.T., 1997. "Transcription initiation from TATA-less promoters within eukaryotic protein-coding genes", *Biochim.Biophys.Acta*, 1351, S. 73-88.

287. SMIT,A.F., 1996. "The origin of interspersed repeats in the human genome", *Curr.Opin.Genet.Dev.*, 6, S. 743-748.

288. SOLOVYEV,V.V., SALAMOV,A.A., LAWRENCE,C.B., 1994. "Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames", *Nucleic Acids Res.*, 22, S. 5156-5163.

289. SONNHAMMER,E.L., EDDY,S.R., BIRNEY,E., BATEMAN,A., DURBIN,R., 1998. "Pfam: multiple sequence alignments and HMM-profiles of protein domains", *Nucleic Acids Res.*, 26, S. 320-322.

290. SOUTHERN,E.M., 1975. "Detection of specific sequences among DNA fragments separated by gel electrophoresis", *J.Mol.Biol.*, 98, S. 503-517.

291. SQUIRE,J.A., LI,M., PERLIKOWSKI,S., FEI,Y.L., BAYANI,J., ZHANG,Z.M., WEKSBERG,R., 2000. "Alterations of H19 imprinting and IGF2 replication timing are infrequent in Beckwith-Wiedemann syndrome", *Genomics*, 65, S. 234-242.

292. STADEN,R., 1996. "The Staden sequence analysis package", *Mol.Biotechnol.*, 5, S. 233-241.

293. STRISSEL,P.L., ESPINOSA,R., III, ROWLEY,J.D., SWIFT,H., 1996. "Scaffold attachment regions in centromere-associated DNA", *Chromosoma*, 105, S. 122-133.

294. STUBDAL,H., LYNCH,C.A., MORIARTY,A., FANG,Q., CHICKERING,T., DEEDS,J.D., FAIRCHILD-HUNTRESS,V., CHARLAT,O., DUNMORE,J.H., KLEYN,P., HUSZAR,D., KAPELLER,R., 2000. "Targeted deletion of the tub mouse obesity gene reveals that tubby is a loss-of-function mutation", *Mol.Cell Biol.*, 20, S. 878-882.

295. SUI,L., DONG,Y., OHNO,M., WATANABE,Y., SUGIMOTO,K., TAI,Y., TOKUDA,M., 2001. "Jab1 expression is associated with inverse expression of p27(kip1) and poor prognosis in epithelial ovarian tumors", *Clin.Cancer Res.*, 7, S. 4130-4135.

- T -

296. TARTAGLIA,L.A., DEMBSKI,M., WENG,X., DENG,N., CULPEPPER,J., DEVOS,R., RICHARDS,G.J., CAMPFIELD,L.A., CLARK,F.T., DEEDS,J., ., 1995. "Identification and expression cloning of a leptin receptor, OB-R", *Cell*, 83, S. 1263-1271.

297. TARTAGLIA,L.A., 1997. "The leptin receptor", *J.Biol.Chem.*, 272, S. 6093-6096.

298. TAUDIEN,S., RUMP,A., PLATZER,M., DRESCHER,B., SCHATTEVOY,R., GLOECKNER,G., DETTE,M., BAUMGART,C., WEBER,J., MENZEL,U., ROSENTHAL,A., 2000. "RUMMAGE--a high-throughput sequence annotation system", *Trends Genet.*, 16, S. 519-520.

299. TAYLOR,B.A., PHILLIPS,S.J., 1996. "Detection of obesity QTLs on mouse chromosomes 1 and 7 by selective DNA pooling", *Genomics*, 34, S. 389-398.

300. THE ARABIDOPSIS GENOME INITIATIVE, 2000. "Analysis of the genome sequence of the flowering plant Arabidopsis thaliana", *Nature*, 408, S. 796-815.

301. THE HUNTINGTON'S DISEASE COLLABORATIVE RESEARCH GROUP, 1993. "A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes.", *Cell*, 72, S. 971-983.

302. THOMAS,A., SKOLNICK,M.H., 1994. "A probabilistic model for detecting coding regions in DNA sequences", *IMA J.Math.Appl.Med.Biol.*, 11, S. 149-160.

303. THOMAS,K.R., CAPECCHI,M.R., 1986. "Introduction of homologous DNA sequences into mammalian cells induces mutations in the cognate gene", *Nature*, 324, S. 34-38.

304. TIEDGE,H., CHEN,W., BROSIUS,J., 1993. "Primary structure, neural-specific expression, and dendritic location of human BC200 RNA", *J.Neurosci.*, 13, S. 2382-2390.

305. TIEDGE,H., BROSIUS,J., 1996. "Translational machinery in dendrites of hippocampal neurons in culture", *J.Neurosci.*, 16, S. 7171-7181.

306. TOLLERVEY,D., KISS,T., 1997. "Function and synthesis of small nucleolar RNAs", *Curr.Opin.Cell Biol.*, 9, S. 337-342.

307. TRIEZENBERG,S.J., 1995. "Structure and function of transcriptional activation domains", *Curr.Opin.Genet.Dev.*, 5, S. 190-196.

308. TSE,E., GRUTZ,G., GARNER,A.A., RAMSEY,Y., CARTER,N.P., COPELAND,N., GILBERT,D.J., JENKINS,N.A., AGULNICK,A., FORSTER,A., RABBITTS,T.H., 1999. "Characterization of the Lmo4 gene encoding a LIM-only protein: genomic organization and comparative chromosomal mapping", *Mamm.Genome*, 10, S. 1089-1094.

- V -

309. VAN,H., V, LITTLE,P.F., 1995. "Report of the fourth international workshop on human chromosome 11 mapping 1994", *Cytogenet.Cell Genet.*, 69, S. 127-158.

310. VAN,H., V, LITTLE,P.F., 1995. "Report of the fourth international workshop on human chromosome 11 mapping 1994", *Cytogenet.Cell Genet.*, 69, S. 127-158.

311. VANIN,E.F., 1985. "Processed pseudogenes: characteristics and evolution", *Annu.Rev.Genet.*, 19, S. 253-272.

312. VARKI,A., 2000. "A chimpanzee genome project is a biomedical imperative", *Genome Res.*, 10, S. 1065-1070.

313. VENKATESH,B., GILLIGAN,P., BRENNER,S., 2000. "Fugu: a compact vertebrate reference genome", *FEBS Lett.*, 476, S. 3-7.

314. VENTER,J.C., ADAMS,M.D., MYERS,E.W., LI,P.W., MURAL,R.J., SUTTON,G.G., SMITH,H.O., YANDELL,M., EVANS,C.A., HOLT,R.A., GOCAYNE,J.D., AMANATIDES,P., BALLEW,R.M., HUSON,D.H., WORTMAN,J.R., ZHANG,Q., KODIRA,C.D., ZHENG,X.H., CHEN,L., SKUPSKI,M., SUBRAMANIAN,G., THOMAS,P.D., ZHANG,J., GABOR MIKLOS,G.L., NELSON,C., BRODER,S., CLARK,A.G., NADEAU,J., MCKUSICK,V.A., ZINDER,N., LEVINE,A.J., ROBERTS,R.J., SIMON,M., SLAYMAN,C., HUNKAPILLER,M., BOLANOS,R., DELCHER,A., DEW,I., FASULO,D., FLANIGAN,M., FLOREA,L., HALPERN,A., HANNENHALLI,S., KRAVITZ,S., LEVY,S., MOBARRY,C., REINERT,K., REMINGTON,K., ABU-THREIDEH,J., BEASLEY,E., BIDDICK,K., BONAZZI,V., BRANDON,R., CARGILL,M., CHANDRAMOULISWARAN,I., CHARLAB,R., CHATURVEDI,K., DENG,Z., DI,F., V, DUNN,P., EILBECK,K., EVANGELISTA,C., GABRIELIAN,A.E.,

GAN,W., GE,W., GONG,F., GU,Z., GUAN,P., HEIMAN,T.J., HIGGINS,M.E., JI,R.R., KE,Z., KETCHUM,K.A., LAI,Z., LEI,Y., LI,Z., LI,J., LIANG,Y., LIN,X., LU,F., MERKULOV,G.V., MILSHINA,N., MOORE,H.M., NAIK,A.K., NARAYAN,V.A., NEELAM,B., NUSSKERN,D., RUSCH,D.B., SALZBERG,S., SHAO,W., SHUE,B., SUN,J., WANG,Z., WANG,A., WANG,X., WANG,J., WEI,M., WIDES,R., XIAO,C., YAN,C., YAO,A., YE,J., ZHAN,M., ZHANG,W., ZHANG,H., ZHAO,Q., ZHENG,L., ZHONG,F., ZHONG,W., ZHU,S., ZHAO,S., GILBERT,D., BAUMHUETER,S., SPIER,G., CARTER,C., CRAVCHIK,A., WOODAGE,T., ALI,F., AN,H., AWE,A., BALDWIN,D., BADEN,H., BARNSTEAD,M., BARROW,I., BEESON,K., BUSAM,D., CARVER,A., CENTER,A., CHENG,M.L., CURRY,L., DANAHER,S., DAVENPORT,L., DESILETS,R., DIETZ,S., DODSON,K., DOUP,L., FERRIERA,S., GARG,N., GLUECKSMANN,A., HART,B., HAYNES,J., HAYNES,C., HEINER,C., HLADUN,S., HOSTIN,D., HOUCK,J., HOWLAND,T., IBEGWAM,C., JOHNSON,J., KALUSH,F., KLINE,L., KODURU,S., LOVE,A., MANN,F., MAY,D., MCCAWLEY,S., MCINTOSH,T., MCMULLEN,I., MOY,M., MOY,L., MURPHY,B., NELSON,K., PFANNKOCH,C., PRATTS,E., PURI,V., QURESHI,H., REARDON,M., RODRIGUEZ,R., ROGERS,Y.H., ROMBLAD,D., RUHFEL,B., SCOTT,R., SITTER,C., SMALLWOOD,M., STEWART,E., STRONG,R., SUH,E., THOMAS,R., TINT,N.N., TSE,S., VECH,C., WANG,G., WETTER,J., WILLIAMS,S., WILLIAMS,M., WINDSOR,S., WINN-DEEN,E., WOLFE,K., ZAVERI,J., ZAVERI,K., ABRIL,J.F., GUIGO,R., CAMPBELL,M.J., SJOLANDER,K.V., KARLAK,B., KEJARIWAL,A., MI,H., LAZAREVA,B., HATTON,T., NARECHANIA,A., DIEMER,K., MURUGANUJAN,A., GUO,N., SATO,S., BAFNA,V., ISTRAIL,S., LIPPERT,R., SCHWARTZ,R., WALENZ,B., YOOSEPH,S., ALLEN,D., BASU,A., BAXENDALE,J., BLICK,L., CAMINHA,M., CARNES-STINE,J., CAULK,P., CHIANG,Y.H., COYNE,M., DAHLKE,C., MAYS,A., DOMBROSKI,M., DONNELLY,M., ELY,D., ESPARHAM,S., FOSLER,C., GIRE,H., GLANOWSKI,S., GLASSER,K., GLODEK,A., GOROKHOV,M., GRAHAM,K., GROPMAN,B., HARRIS,M., HEIL,J., HENDERSON,S., HOOVER,J., JENNINGS,D., JORDAN,C., JORDAN,J., KASHA,J., KAGAN,L., KRAFT,C., LEVITSKY,A., LEWIS,M., LIU,X., LOPEZ,J., MA,D., MAJOROS,W., MCDANIEL,J., MURPHY,S., NEWMAN,M., NGUYEN,T., NGUYEN,N., NODELL,M., 2001. "The sequence of the human genome", *Science*, 291, S. 1304-1351.

315. VIEGAS-PEQUIGNOT,E., MALFOY,B., SABATIER,L., DUTRILLAUX,B., 1987. "Different reactivity of Z-DNA antibodies with human chromosomes modified by actinomycin D and 5-bromodeoxyuridine", *Hum.Genet.*, 75, S. 114-119.

316. VINOGRADOV,A.E., 1998. "Genome size and GC-percent in vertebrates as determined by flow cytometry: the triangular relationship", *Cytometry*, 31, S. 100-109.

317. VISVADER,J.E., VENTER,D., HAHM,K., SANTAMARIA,M., SUM,E.Y., O'REILLY,L., WHITE,D., WILLIAMS,R., ARMES,J., LINDEMAN,G.J., 2001. "The LIM domain gene LMO4 inhibits differentiation of mammary epithelial cells in vitro and is overexpressed in breast cancer", *Proc.Natl.Acad.Sci.U.S.A*, 98, S. 14452-14457.

318. VOGELSTEIN,B., PARDOLL,D.M., COFFEY,D.S., 1980. "Supercoiled loops and eucaryotic DNA replicaton", *Cell*, 22, S. 79-85.

- W -

319. WANG,Q., CURRAN,M.E., SPLAWSKI,I., BURN,T.C., MILLHOLLAND,J.M., VANRAAY,T.J., SHEN,J., TIMOTHY,K.W., VINCENT,G.M., DE JAGER,T., SCHWARTZ,P.J., TOUBIN,J.A., MOSS,A.J., ATKINSON,D.L., LANDES,G.M., CONNORS,T.D., KEATING,M.T., 1996. "Positional cloning of a novel potassium channel gene: KVLQT1 mutations cause cardiac arrhythmias", *Nat.Genet.*, 12, S. 17-23.

320. WARREN,S.T., ZHANG,F.P., SUTCLIFFE,J.S., PETERS,J.F., 1988. "Strategy for molecular cloning of the fragile X site DNA", *Am.J.Med.Genet.*, 30, S. 613-623.

321. WATANABE,N., BROOME,M., HUNTER,T., 1995. "Regulation of the human WEE1Hu CDK tyrosine 15-kinase during the cell cycle", *EMBO J.*, 14, S. 1878-1891.

322. WATERSTON,R.H., LINDBLAD-TOH,K., BIRNEY,E., ROGERS,J., ABRIL,J.F., AGARWAL,P., AGARWALA,R., AINSCOUGH,R., ALEXANDERSSON,M., AN,P., ANTONARAKIS,S.E., ATTWOOD,J., BAERTSCH,R., BAILEY,J., BARLOW,K., BECK,S., BERRY,E., BIRREN,B., BLOOM,T., BORK,P., BOTCHERBY,M., BRAY,N., BRENT,M.R., BROWN,D.G., BROWN,S.D., BULT,C., BURTON,J., BUTLER,J., CAMPBELL,R.D., CARNINCI,P., CAWLEY,S., CHIAROMONTE,F., CHINWALLA,A.T., CHURCH,D.M., CLAMP,M., CLEE,C., COLLINS,F.S., COOK,L.L., COPLEY,R.R., COULSON,A., COURONNE,O., CUFF,J., CURWEN,V., CUTTS,T., DALY,M., DAVID,R., DAVIES,J., DELEHAUNTY,K.D., DERI,J., DERMITZAKIS,E.T., DEWEY,C., DICKENS,N.J., DIEKHANS,M., DODGE,S., DUBCHAK,I., DUNN,D.M., EDDY,S.R., ELNITSKI,L., EMES,R.D., ESWARA,P., EYRAS,E., FELSENFELD,A., FEWELL,G.A., FLICEK,P., FOLEY,K., FRANKEL,W.N., FULTON,L.A., FULTON,R.S., FUREY,T.S., GAGE,D., GIBBS,R.A., GLUSMAN,G., GNERRE,S.,

GOLDMAN,N., GOODSTADT,L., GRAFHAM,D., GRAVES,T.A., GREEN,E.D., GREGORY,S., GUIGO,R., GUYER,M., HARDISON,R.C., HAUSSLER,D., HAYASHIZAKI,Y., HILLIER,L.W., HINRICHS,A., HLAVINA,W., HOLZER,T., HSU,F., HUA,A., HUBBARD,T., HUNT,A., JACKSON,I., JAFFE,D.B., JOHNSON,L.S., JONES,M., JONES,T.A., JOY,A., KAMAL,M., KARLSSON,E.K., KAROLCHIK,D., KASPRZYK,A., KAWAI,J., KEIBLER,E., KELLS,C., KENT,W.J., KIRBY,A., KOLBE,D.L., KORF,I., KUCHERLAPATI,R.S., KULBOKAS,E.J., KULP,D., LANDERS,T., LEGER,J.P., LEONARD,S., LETUNIC,I., LEVINE,R., LI,J., LI,M., LLOYD,C., LUCAS,S., MA,B., MAGLOTT,D.R., MARDIS,E.R., MATTHEWS,L., MAUCELI,E., MAYER,J.H., MCCARTHY,M., MCCOMBIE,W.R., MCLAREN,S., MCLAY,K., MCPHERSON,J.D., MELDRIM,J., MEREDITH,B., MESIROV,J.P., MILLER,W., MINER,T.L., MONGIN,E., MONTGOMERY,K.T., MORGAN,M., MOTT,R., MULLIKIN,J.C., MUZNY,D.M., NASH,W.E., NELSON,J.O., NHAN,M.N., NICOL,R., NING,Z., NUSBAUM,C., O'CONNOR,M.J., OKAZAKI,Y., OLIVER,K., OVERTON-LARTY,E., PACHTER,L., PARRA,G., PEPIN,K.H., PETERSON,J., PEVZNER,P., PLUMB,R., POHL,C.S., POLIAKOV,A., PONCE,T.C., PONTING,C.P., POTTER,S., QUAIL,M., REYMOND,A., ROE,B.A., ROSKIN,K.M., RUBIN,E.M., RUST,A.G., SANTOS,R., SAPOJNIKOV,V., SCHULTZ,B., SCHULTZ,J., SCHWARTZ,M.S., SCHWARTZ,S., SCOTT,C., SEAMAN,S., SEARLE,S., SHARPE,T., SHERIDAN,A., SHOWNKEEN,R., SIMS,S., SINGER,J.B., SLATER,G., SMIT,A., SMITH,D.R., SPENCER,B., STABENAU,A., STANGE-THOMANN,N., SUGNET,C., SUYAMA,M., TESLER,G., THOMPSON,J., TORRENTS,D., TREVASKIS,E., TROMP,J., UCLA,C., URETA-VIDAL,A., VINSON,J.P., VON NIEDERHAUSERN,A.C., WADE,C.M., WALL,M., WEBER,R.J., WEISS,R.B., WENDL,M.C., WEST,A.P., WETTERSTRAND,K., WHEELER,R., WHELAN,S., WIERZBOWSKI,J., WILLEY,D., WILLIAMS,S., WILSON,R.K., WINTER,E., WORLEY,K.C., WYMAN,D., YANG,S., YANG,S.P., ZDOBNOV,E.M., ZODY,M.C., LANDER,E.S., 2002. "Initial sequencing and comparative analysis of the mouse genome", *Nature*, 420, S. 520-562.

323. WATSON,J.D., CRICK,F.H., 1974. "Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. G.D. Watson and F.H.C. Crick. Published in Nature, number 4356 April 25, 1953", *Nature*, 248, S. 765.

324. WAY,J.C., CHALFIE,M., 1988. "mec-3, a homeobox-containing gene that specifies differentiation of the touch receptor neurons in C. elegans", *Cell*, 54, S. 5-16.

325. WEATHERALL,D., 1999. "From genotype to phenotype: genetics and medical practice in the new millennium", *Philos.Trans.R.Soc.Lond B Biol.Sci.*, 354, S. 1995-2010.

326. WEBER,J.L., MYERS,E.W., 1997. "Human whole-genome shotgun sequencing", *Genome Res.*, 7, S. 401-409.

327. WEINSTEIN,L.B., STEITZ,J.A., 1999. "Guided tours: from precursor snoRNA to functional snoRNP", *Curr.Opin.Cell Biol.*, 11, S. 378-384.

328. WHITE,M.A., 1996. "The yeast two-hybrid system: forward and reverse", *Proc.Natl.Acad.Sci.U.S.A*, 93, S. 10001-10003.

329. WHITELAW,C.B., GROLLI,S., ACCORNERO,P., DONOFRIO,G., FARINI,E., WEBSTER,J., 2000. "Matrix attachment region regulates basal beta-lactoglobulin transgene expression", *Gene*, 244, S. 73-80.

- X -

330. XU,Y., MURAL,R.J., UBERBACHER,E.C., 1994. "Constructing gene models from accurately predicted exons: an application of dynamic programming", *Comput.Appl.Biosci.*, 10, S. 613-623.

- Y -

331. YAMAMURA,J., NOMURA,K., 2001. "Analysis of sequence-dependent curvature in matrix attachment regions", *FEBS Lett.*, 489, S. 166-170.

332. YATSUKI,H., WATANABE,H., HATTORI,M., JOH,K., SOEJIMA,H., KOMODA,H., XIN,Z., ZHU,X., HIGASHIMOTO,K., NISHIMURA,M., KURATOMI,S., SASAKI,H., SAKAKI,Y., MUKAI,T., 2000. "Sequence-based structural features between Kvlqt1 and Tapa1 on mouse chromosome 7F4/F5 corresponding to the Beckwith-Wiedemann syndrome region on human 11p15.5: long-stretches of unusually well conserved intronic sequences of kvlqt1 between mouse and human", *DNA Res.*, 7, S. 195-206.

- Z -

333. ZAWEL,L., REINBERG,D., 1993. "Initiation of transcription by RNA polymerase II: a multi-step process", *Prog.Nucleic Acid Res.Mol.Biol.*, 44, S. 67-108.

334. ZHANG,M.Q., 1997. "Identification of protein coding regions in the human genome by quadratic discriminant analysis", *Proc.Natl.Acad.Sci.U.S.A*, 94, S. 565-568.

335. ZHANG,Y., PROENCA,R., MAFFEI,M., BARONE,M., LEOPOLD,L., FRIEDMAN,J.M., 1994. "Positional cloning of the mouse obese gene and its human homologue", *Nature*, 372, S. 425-432.

336. ZHOU,X.L., WERELIUS,B., LINDBLOM,A., 2004. "A screen for germline mutations in the gene encoding CCCTC-binding factor (CTCF) in familial non-BRCA1/BRCA2 breast cancer", *Breast Cancer Res.*, 6, S. R187-R190.

337. ZHU,T.H., BODEM,J., KEPPEL,E., PARO,R., ROYER-POKORA,B., 1995. "A single ancestral gene of the human LIM domain oncogene family LMO in Drosophila: characterization of the Drosophila Dlmo gene", *Oncogene*, 11, S. 1283-1290.

338. ZOUBAK,S., CLAY,O., BERNARDI,G., 1996. "The gene distribution of the human genome", *Gene*, 174, S. 95-102.

7 DANKSAGUNG

Das Anfertigen und Zusammenschreiben einer Dissertation ist ein Unterfangen, das nicht nur eine Menge Zeit in Anspruch nimmt, sondern an dem auch eine Vielzahl von Mensch mittel oder unmittelbar ihren Anteil haben. All diesen Menschen gebührt an dieser Stelle mein herzlicher Dank.

An erster Stelle möchte ich mich bei Herrn Prof. Dr. [redacted] bedanken für die Überlassung des Promotionsthemas und die Bereitstellung des Arbeitsplatzes mit all seiner modernen technischen Ausstattung, ohne die eine Bearbeitung des Themas schlichtweg unmöglich gewesen wäre.

Ein weiteres großes Dankeschön gilt Herrn Prof. Dr. [redacted] der die ersten beiden Jahre meine Arbeit fachlich betreute und für mich immer ein offenes Ohr hatte.

Mein Dank richtet sich weiter an Dr. [redacted], der die fachliche Betreuung der Arbeit nach dem Ausscheiden von [redacted] im nahtlosen Wechsel und ebenso hohem Niveau übernahm und mir bis zum Schluss engagiert zur Verfügung stand. Ich danke ihm für die immerwährende Gesprächs- und Diskussionsbereitschaft und seiner fachlichen Kompetenz bei der Beurteilung von manchmal eher „unlogischen" Versuchsergebnisse.

Des weiteren geht mein Dank an [redacted] und [redacted], die mir bei der Sequenzierung der vielen Tausend Proben und bei der Etablierung und der Routine des „High-Throughput-Sequencings" tatkräftig geholfen haben. An dieser Stelle gilt mein Dank natürlich auch an Dr. [redacted] und [redacted] die mir beim Aufbau der Computer-Infrastruktur und bei vielen Rechner- und Software-relevanten Problemen kompetent zur Seite standen und auf viele Fragen und Komplikationen eine Lösung parat hatten. Ein dickes Dankeschön geht auch an Dr. [redacted]; mit ihr als Doktorandin und Sequenzierprojekt-Partnerin teilten wir uns über die gesamte Zeit das Sequenzierlabor. Danke für den immer offenen Austausch von Tipps und Erfahrungen und für die Diskussions- und Kooperationsbereitschaft bei der gemeinsamen Nutzung der Sequenzierfacilities. Dank gebührt auch [redacted], die im Rahmen ihrer Diplomarbeit mit vielen Sequenz-informationen bei der Charakterisierung des telomerwärts-gelegenen Sequenzbereiches mithalf.

Nicht vergessen werden darf natürlich auch des gesamte Team der Arbeitgruppe [redacted] das mir über die gesamte Zeit helfend bei den vielen kleinen und größeren Problemen des Laboralltages zur Seite stand. Mein Dank geht dabei an Dr. [redacted], Dr. [redacted], Dr. [redacted], Dr. [redacted] [redacted] [redacted] und [redacted].

Darüber hinaus möchte ich mich natürlich auch ganz herzlich bei [redacted] bedanken, die während der „never ending story" meiner Promotion es nie an Verständnis, Zuspruch und Geduld hat mangeln lassen und die mich in ihrer liebevollen Weise mental darin sehr bestärkt hat, diese Arbeit schließlich zu beenden.

8 ANHANG

8.1 Publikationen

Brueckmann,T. und Cichutek, A.; Seipel, B.; Hauser, H.; Schlaubitz, S.; Prawitt, D.; Hankeln, T.; Schmidt, E.R.; Winterpacht, A.; Zabel, B.U.:„Comparative architectural aspects of regions of conserved synteny on human chromosome 11p15.3 and mouse chromosome 7 (including genes WEE1 and LMO1)" *Cytogent. Cell Genet.* 93:277-283 (2001)

8.2 Poster

Brueckmann, T.; Schlaubitz, S.; Hankeln, T.; Winterpacht, A.; Schmidt, E.R.; Zabel, B.
„Comparative genomic sequencing of the TUB gene region in man and mouse - Identification of putative regulatory elements", 12. Jahrestagung, Deutsche Gesellschaft für Humangenetik, Lübeck, 22.-25.03.2000, Med Genetik, 12:90, 2000

Zabel B, Amid C, Bahr A, Bikar S, **Brückmann T**, Cichutek A, Mujica A, Sampson N, Schlaubitz S, Hankeln T, Winterpacht A, Schmidt ER: *„Identification of genes and putative gene regulatory sequences by comparative sequencing between man and mouse.*" DHGP-Tagung: German Human Genome Project - Implications, Progress, and the Future, München 28.-30.11.1999

Winterpacht, A.; Cichutek, A.; **Brückmann, T.**; Bahr, A.; Amid, C.; Bikar, S.; Hankeln, T.; Zabel, B.; Schmidt, E.R., *„Identification of genes and putativ regulatory sequences by comparative sequencing between man and mouse*", 11. Jahrestagung, Deutsche Gesellschaft für Humangenetik, Nürnberg, 24.-27.03.1999, Med Genetik, 11:222-223,1999

B. Zabel, A. Bahr, C. Amid, S. Bikar, **T. Brückmann**, A. Cichutek, B. Seipel, A. Winterpacht, T. Hankeln, E.R. Schmidt, *„Comparative Sequencing of a 1 Mb Region in Man (Chromosome 11p15) and Mouse (Chromosome 7)",* Progress Report 1996-1998, German Human Genome Project

Zabel B, Bahr A, Amid C, Bikar S, **Brückmann T**, Cichutek A, Després S, Seipel B, Winterpacht A, Hankeln T, Schmidt ER, *„Comparative sequencing of a 1 Mb region in man (chromosome 11p15) and mouse (chromosome 7)";* 6th International Workshop on Human Chromosome 11 - Mapping, Sequencing, Disease Genes. Nizza, Frankreich, 02.-05.05.1998

A. Bahr, C. Amid, S.E. Bikar, **T. Brückmann**, A. Cichutek, T. Hankeln, B. Seipel, E.R. Schmidt, A. Winterpacht, B. Zabel, *„Comparative Genomic Sequencing of A 1MB Syntenic Region in Man (HsaC11p15) and Mouse (MmuC7)",* Human Genome Meeting 1998, Turin, Italien, 28.-30.03.1998

Winterpacht A, Amid C, Bahr A, **Brückmann T**, Cichutek A, Hankeln T, Schmidt ER, Seipel B, Zabel B:
„Analysis of a 1 Mb region on chromosome 11p15.3 by comparative sequencing between man and mouse" 10. Jahrestagung, Deutsche Gesellschaft für Humangenetik, Jena, 25.-28.03.1998, Med Genetik 10:113,1998

Schmidt ER, Amid C, Bahr A, **Brückmann T**, Cichutek A, Hankeln T, Seipel B, Winterpacht A, Zabel B:
„Comparative genomics: sequencing of a 1 Mb syntenic region (HSAC11p15/MmuC7) in mouse and man."
2nd International Beutenberg Symposium - Genome Analysis: Strategies, Medical and Industrial Applications, Jena, 11.-13.12.1997

Schmidt ER, Bahr A, Amid C, **Brückmann T**, Cichutek A, Seipel B, Winterpacht A, Hankeln T, Zabel B:
„Comparative sequencing of a 1 Mb region in man (chromosome 11o15) and mouse (chromosome 7)."
Fifth Annual Meeting of the Society for Molecular Biology and Evolution, Garmisch-Partenkirchen, 01.-04.06.1997

Zabel B, Bahr A, Amid C, **Brückmann T**, Cichutek A, Seipel B, Winterpacht A, Hankeln T, Schmidt ER:
„Comparative sequencing of a 1 Mb region in man (chromosome 11p15) and mouse (chromosome 7)."Human Genome Meeting 1997, Toronto, 06.-08.03.1997

Zabel B, Löbbert R, Prawitt D, Seipel B, Germayer S, **Brückmann T**, Cichutek A, Munroe DJ, Pelletier J, Housman DE, Winterpachr A, *„Chromosome region 11p15: genomic analysis, transcript mapping, gene identification, comparative sequencing*", Fifth International Chromosome 11 Workshop, Niagara-On-The-Lake, Canada, 12.-16.5.1996, Cytogenet Cell Genet 74:56,1996

Prawitt D, Munroe DJ, Pelletier J, Löbbert R, Bric E, Fricke G, Higgins M, Shows TB, Housman DE, **Brückmann T**, Winterpacht A, Zabel B, *„Analysis of a region in 11p15.5 that is homozygously deleted in Wilms tumors*" 8. Jahrestagung, Deutsche Gesellschaft für Humangenetik, Göttingen, 6.-9.3.1996; Med Genetik 8:38,1996

8.3 Patente

WO 02/078630 A3 • PCT/US2/09473

Titel: "*TUB 3´-Variant*"; Patentklasse: **A61K**; Erfinder: Prawitt, D.; Zabel, B.; **Brueckmann, T**.; Schmidt, E.R.; Winterpacht, A.; Hankeln T.; Anmelder: Johannes Gutenberg-Universität, Mainz; Priorität: 30. März 2001; Anmeldedatum: 28. März 2002, Internationales Publikationsdatum: 10. Oktober 2002

WO 02/083707 A1 • PCT/US2/09474

Titel: "*TUB Antisense Construct*", Int. Patentklasse: **C07H21/04**; Erfinder: Prawitt, D.; Zabel, B.; **Brueckmann, T**.; Schmidt, E.R.; Winterpacht, A.; Hankeln T.; Anmelder: Johannes Gutenberg-Universität, Mainz; Priorität: 30. März 2001; Anmeldedatum: 28. März 2002; Internationales Publikationsdatum: 24. Oktober 2002

US 2004/0242513 A1

Titel: "*TUB Antisense Construct*", Int. Patentklasse: **A61K 48/00**; Erfinder: Prawitt, D.; Zabel, B.; **Brueckmann, T**.; Schmidt, E.R.; Winterpacht, A.; Hankeln T.; Anmelder: Johannes Gutenberg-Universität; Anmeldedatum: 07. Mai 2004; Veröffentlichungsdatum: 02 Dezember 2004

Ich versichere, die vorliegende Arbeit selbständig und nur mit den angegebenen Hilfsmitteln angefertigt zu haben.

Mainz, den 14. März 2005