

Johannes Gutenberg – Universität Mainz



Genexpressionsanalyse der fötalen Wachstumsfuge des Rindes

Dissertation

zur Erlangung des Grades
Doktor der Naturwissenschaften

Am Fachbereich Biologie,
Der Johannes Gutenberg-Universität Mainz
Institut für Molekulargenetik,
Gentechnische Sicherheitsforschung und -beratung

vorgelegt von
Steffen Rapp
geb. am 02.09.1980 in Speyer

Mainz 2011

Dekan:

1. Berichterstatter:

2. Berichterstatter:

Tag der Promotion: 19.12.2011

Für meine Eltern,
die mich immer unterstützt haben.

Im Gedenken an meine Mutter.

*„Die Neugier steht immer an erster Stelle eines Problems,
das gelöst werden will.“*

(Galileo Galilei)

INHALT

Abbildungsverzeichnis.....V

Tabellenverzeichnis.....VII

AbkürzungsverzeichnisVIII

Einleitung..... 1

1.1 Evolution des Vertebratenskeletts 2

 Strukturelle Entwicklung..... 2

 Genetik der Skelettentwicklung..... 3

1.2 Knochen-, Knorpelentwicklung und Wachstumsfuge 5

 Chondrale Ossifikation..... 6

 Wachstumsfuge, Aufbau und Homöostase 9

1.3 Genexpressionsanalyse und neue Sequenziertechnologien.....15

 Sequenzierung der nächsten Generation 16

 Genexpressionsanalyse mittels RNA-seq..... 23

1.4 Zielsetzung25

2 Material und Methode..... 26

2.1 Material.....26

 Standard-Lösungen 26

 Kits und Enzyme 28

 Ausgangsmaterial..... 28

2.2 Molekularbiologische Standardmethoden.....29

 Isolierung von DNA 29

 Isolierung von RNA..... 30

Fällung von Nucleinsäuren.....	30
Agarose-Gelelektrophorese	31
Konzentration und Qualität von Nucleinsäuren	32
Ligationsreaktionen.....	33
Transformation mittels Elektroporation.....	34
Restriktion mit Restriktionsendonukleasen	34
polyA ⁺ -Anreicherung „NucleoTrap mRNA Mini Kit“	34
cDNA-Synthese.....	35
Polymerasekettenreaktion.....	35
Verwendete Primer	37
2.3 Programme, Skripte und WEB-Tools.....	38
Verwendete Programme.....	38
Verwendete Perl-Skripte.....	39
Verwendete WEB-Tools	41
2.4 Bioinformatische Standardmethoden	42
2.5 cDNA-Bibliotheken und EST-Sequenzierung.....	44
2.6 Auswertung der EST-Klone	45
Expressionsprofile knochen- knorpelspezifischer Gene	47
Weiterführende Analyse möglicher interessanter Transkripte	47
2.7 Nachweis spezifischer Transkripte mittels RT-PCR	48
RT-PCR mit cDNAs der Wachstumsfuge und Vergleichsgewebe	48
Semi-quantitative RT-PCR mit cDNAs der Wachstumsfuge und Vergleichsgeweben	48
2.8 Roche-454/FLX-Sequenzierung.....	49
Clustern der Roche-454/FLX-Daten	49
Blast-Analysen der Roche-454/FLX-Daten	50
Statistische Analyse der Bast-Ergebnisse.....	51

2.9	Transkriptomanalyse mittels Illumina-GAllx-Sequenzierung	52
	Vorbereitungen und cDNA-Bibliothek	52
	Analyse Vorbereitungen	52
	Hochdurchsatz-BLAST-Analysen	53
	Mapping begrenzter genomischer Bereiche mit PASS (Campagna et al. 2009)	54
	Genomweite „Mappings“ mit CLC	54
3	Ergebnisse	55
3.1	Genexpressionsanalysen mittels EST-Sequenzierung	56
	Aufbereitung der cDNA-Bank und Sequenzierung.....	56
	Charakterisierung der cDNA-Klone	58
	Stark transkribierte Gene in der fötalen Wachstumsfuge	60
	Expressionsprofile knochen-/knorpelspezifischer Gene.....	65
	Expression ribosomaler Proteine	67
3.2	Charakterisierung einzelner Sequenz-Cluster und EST-Klone.....	68
	Unbekannte, stark exprimierte ESTs.....	69
	Unbekannte, weniger stark exprimierte ESTs.....	69
	Synthänie und Interspeziesvergleich	74
	RT-PCR zum Nachweis von interessanten Transkripten in verschiedenen Geweben	76
	Analyse der Transkriptvarianten des LOC618319.....	78
3.3	Genexpressionsanalysen mittels Roche/454-Sequenzierung.....	82
	Vergleich von Assembly-Plattformen	83
	Annotation und „Gene Ontology“-Analysen mit BLAST2GO (Conesa et al. 2005)	84
	Sequenzannotation und quantitative Expressionsanalyse	86
	Hoch exprimierte uncharakterisierte Sequenzen.....	89

3.4	Genexpressionsanalysen mittels Illumina-Sequenzierung	91
	Qualitätskontrolle und Qualitätsfilter	92
	„Read Count“-Ansatz mittels Hochdurchsatz-Blast-Suchen	93
	Charakterisierung unbekannter Sequenzcluster	95
3.5	Vergleich zwischen verschiedenen Analyseverfahren	101
3.6	Genomweites „Mapping“ und differentielle Genexpression	103
4	Diskussion.....	105
4.1	Genexpressionsanalyse der fötalen Wachstumsfuge des Rindes	106
	Expressionsanalyse über „Next Generation Sequencing“	115
	Transkriptom-Analyse mittels Illumina-Sequenzierung.....	118
4.2	Analyse potentiell neuer Gene und Funktionen bei der Knochen- und Knorpelentwicklung.....	120
	LOC618319 – Gen(produkte) und mögliche Funktion	121
4.3	Next-Generation-Sequencing – Datenauswertung und Vergleichbarkeit	129
5	Zusammenfassung	134
6	Literaturverzeichnis	137
7	Anhang	161
	Verwendete Vektoren.....	161
	„Tissue Specificity Index“ - TSI	163
	Eigene verwendete Perl-Skripte.....	164
	Übersicht elektronischer Anhang	Fehler! Textmarke nicht definiert.

Abbildungsverzeichnis

Abb. 1: Schematischer Aufbau des Röhrenknochens	6
Abb. 2: Frühstadien der enchondralen Ossifikation	7
Abb. 3: Aufbau der Wachstumsfuge	10
Abb. 4: Regulation der Chondrozytendifferenzierung in der Wachstumsfuge	12
Abb. 5: Entwicklung der Kosten für Hochdurchsatzsequenzierungen	17
Abb. 6: 454/Pyrosequenzierung	18
Abb. 7: Prinzip der SOLiD™-Sequenzierung	19
Abb. 8: Sequenzierbibliothek und "Bridge Amplification" der Illumina-Technologie	21
Abb. 9: Sequenzabdeckung nach RNA- bzw. DNA-Fragmentierung	24
Abb. 10: Ergebnisse eines RNA-Seq-Experiments	24
Abb. 11: cDNA-Synthese	44
Abb. 12: Kontrolle der Plasmidpräparationen	57
Abb. 13: Verteilung unbrauchbarer Sequenzen über Platten und Fraktionen	58
Abb. 14: Gesamtübersicht über die Klonverteilung in der cDNA-Bank	59
Abb. 15: Funktionelle Gruppen – der 100 höchst exprimierten Gene	62
Abb. 16: EST-Expressionsprofil der Kollagene	65
Abb. 17: EST-Expressionsprofil weiterer ECM-Proteine	66
Abb. 18: EST-Expressionprofil der Proteoglykane	66
Abb. 19: EST-Expressionsprofil der ribosomalen Proteine	67
Abb. 20: EST-Cluster LOC618319 und LOC511905 genomische Organisation	69
Abb. 21: Contig 20 - E12P44 - genomische Organisation	71
Abb. 22: Genomische Organisation und Domänen - Contig 136	72
Abb. 23: genomischer Locus von F07P42	73
Abb. 24: Chromosomale Translokationen	74
Abb. 25: Interspeziesvergleich - <i>LOC618319</i> und F07P42	75
Abb. 26: Übersicht aller Sequenzen des Contigs LOC618319	79
Abb. 27: EST-Sequenzvarianten und genomische Zuordnung (LOC618319)	80
Abb. 28: Aminosäurealignment der möglichen Isoformen von LOC618319	80

Abb. 29: RT-PCR LOC618319	81
Abb. 30: Zusammensetzung der 454-Daten	82
Abb. 31: Verteilung der Contig-Größen	84
Abb. 32: Funktionelle Gruppierung der 454-Sequenzen	85
Abb. 33: Ablauf der Blast-Auswertungen der Roche-454/FLX-Daten	86
Abb. 34: Fötale Wachstumsfuge - Gesamt-RNA und amplifizierte polyA ⁺ -RNA	91
Abb. 35: Qualitätskontrolle und Trimmen der Illumina-Daten	92
Abb. 36: Mapping - LOC100296593 / LOC281370	97
Abb. 37: Mapping - LOC100467179 - SERF2	98
Abb. 38: Mapping LOC614348 - CH12H19orf15 homolog	99
Abb. 39: Mapping - LOC618319	100
Abb. 40: Vergleich der Blast-Ergebnisse von EST-, 454- und Illumina-Sequenzierung	102
Abb. 41: Vergleich zwischen Blast- und Mapping- Analyse am Beispiel der Proteoglykane	103
Abb. 42: Vergleich - Kollagene – Expressionsprofile	108
Abb. 43: Vergleich - Expressionsprofile der ECM-Proteine	111
Abb. 44: Signalwege der ECM – Abdeckung im EST-Projekt	112
Abb. 45: Zentrale Aspekte der enchondralen Ossifikation	119
Abb. 46: Genomische Organisation des LOC618319 bei <i>Bos taurus</i>	121
Abb. 47: Konservierte Bereiche im Genprodukt von <i>LOC618319</i>	123
Abb. 48: LOC618319 – Signalpeptid	124
Abb. 49: Möglicher schematischer Aufbau des Proteins LOC618319	125
Abb. 50: Mögliche Bindestellen von Transkriptionsfaktoren am <i>LOC618319</i>	128
Abb. 51: Vorgehensweise bei der Analyse der NGS-Daten	129
Abb. 52: Read-Abdeckung und Transkriptlängen-Fehler	133

Tabellenverzeichnis

Tab. 1: Spezifikationen der NGS-Technologien	22
Tab. 2: Standardansatz für die cDNA-Erststrangsynthese	35
Tab. 3: PCR-Ansatz	36
Tab. 4: Verwendete Primer	37
Tab. 5: Programmliste	38
Tab. 6: Verwendete - Web - Applikationen	41
Tab. 7: Blast-Optionen und Datenbanken	45
Tab. 8: Kategorien der EST-Sequenzen	46
Tab. 9: Filter für Illumina-Sequenzen	53
Tab. 10: Gefilterte Illumina-Sequenzen	53
Tab. 11: Übersicht über Transformationen und Anzahl der sequenzierten Klone	57
Tab. 12: Hoch exprimierte Gene in der fötalen Wachstumsfuge des Rindes	61
Tab. 13: Funktionelle Gruppierung der 100 höchst repräsentierten EST-Sequenzen	63
Tab. 14: Mögliche proteinkodierende EST-Sequenzen	70
Tab. 15: PCR mit cDNA verschiedener Gewebe von <i>Bos taurus</i>	77
Tab. 16: RT-PCR zum Nachweis interessanter Transkripte in verschiedenen Geweben	78
Tab. 17: Vergleich verschiedener Assembler	83
Tab. 18: Höchst exprimierte Gene durch 454-Sequenzierung	88
Tab. 19: Höchst exprimierte Gene nach Illumina-RNA-Sequenzierung	94
Tab. 20: Manuelle Annotation der unbekanntenen Treffer der Top 300	95
Tab. 21: Hoch exprimierte uncharakterisierte Genloci aus dem Illumina-Projekt	96
Tab. 22: Blast-Ergebnisse auffälliger Bereiche - LOC618319	101
Tab. 23: Top 25 exprimierte Gene nach BLAST-Analyse und „Mapping“	104
Tab. 24: "Tissue Specificity Index" - <i>Bos Taurus</i>	163

Abkürzungsverzeichnis

3'-UTR	3'-untranslatierte Region
5'-UTR	5'-untranslatierte Region
Abb.	Abbildung
ABI	"Applied Biosystems"
abs.	absolute
Amp	Ampicillin
AS	Aminosäure
BLAST	"Basic lokal alignment search tool"
BMP	"bone morphogenetic protein"
bp	Basenpaare
bzw.	beziehungsweise
cDNA	komplementäre DNA
COMP	"Cartilage oligomeric matrix protein"
CTGF	"Connective tissue growth factor"
DNA	"deoxyribonucleic acid"
DNase	Desoxyribonuklease
dNTP	2'-Deoxynukleosidtriphosphat
<i>E. coli</i>	Escherichia coli
ECM	"Extra Cellular Matrix"
EST	"expressed sequence tags"
FGF	"fibroblas growth factor"
FGP	"fetal growth plate" - fötale Wachstumsfuge des Rindes
FN1	Fibronectin 1
FTH1	"ferritin heavy polypeptide 1"
GAG	Glykosaminoglykane
HAPLN1	"Cartilage linking protein 1"
HMG	"high mobility group"
ID1	"Inhibitor of DNA binding 1"
Ihh	"indian hedgehog"
kb	Kilobasenpaare
MGP	"Matrix Gla protein"
mRNA	"messenger" Ribonukleinsäure
NCBI	"National Center for Biotechnology Information"
NIH	"National Institute of Health"
nt	Nukleotide
ORF	"open reading frame"
PCR	"polymerase chain reaction"
PTHrP	"Parathyroid Hormone related Protein"
RNA	"ribonucleic acid"
RNase	Ribonuklease
RT	Reverse Transkriptase
RT	Raumtemperatur
RT-PCR	"reverse transcriptase polymerase chain reaction"
SOLiD	"Sequencing by Oligonucleotide Ligation and Detection"
SPP1	"Secreted phosphoprotein 1"
SSR	"simple sequence repeat"
Tab.	Tabelle
Taq	Thermus aquaticus
TGF	"transforming growth factor"
U	"units" (Enzymeinheiten)
vgl.	vergleiche
Vol.	Volumen
z.B.	zum Beispiel

Einleitung

Ein grundlegender Schritt in der Evolution des Körperbaus ist die Entwicklung eines tragenden, oft nur teilweise organischen Gerüsts. Dieses Konzept ist auf verschiedenste Weise realisiert. Im Innern (Endoskelett) oder auch außen aufgelagert (Exoskelett) beeinflusst es in einem hohen Maß die Eigenschaften und das Aussehen von Lebewesen. Es hat tragende Funktion, schützt die inneren Organe und liefert Ansatzpunkte für die Muskulatur. Im Allgemeinen erlauben Endoskelette eine höhere Körpergröße, größere Beweglichkeit (Currey 1967) und besitzen bessere regenerative Fähigkeiten. Das Endoskelett ist nicht nur evolutionär, wie an der Entwicklung vieler Vertebratenspezies zu sehen ist, sondern auch innerhalb der Lebensspanne eines einzelnen Individuums höchst wandelbar und anpassungsfähig (Currey 1984). Wachstum, Funktionalität und Flexibilität sind abhängig von den hoch spezialisierten knochen- und knorpelbildenden Geweben. Die Spezialisierung von verschiedenen Geweben geht einher mit der Differenzierung und Proliferation bestimmter Zelltypen (Adams und Watt 1993). Eine der in dieser Hinsicht aktivsten Regionen ist die Wachstumsfuge. Zwischen der Epiphyse und Diaphyse der Röhrenknochen gelegen bildet sie gerade in den frühen Entwicklungsjahren mit hohen Differenzierungs- und Proliferationsraten die notwendigen Zellpopulationen für ein schnelles Knochen- und Knorpelwachstum. Mit zunehmendem Alter nimmt diese Aktivität rapide ab und die Wachstumsfuge verkümmert. Besonders interessant sind daher Genexpressionsanalysen in Wachstumsfugen früher Entwicklungsstadien. Expressionsdaten aus derart spezialisierten Geweben können wichtige Rückschlüsse auf Funktion und Zusammenspiel von Genen bei der Knochen- und Knorpelentwicklung geben. Die dazu zur Verfügung stehenden Methoden sind zahlreicher denn je. Sie reichen von bereits länger verwendeten Verfahren wie SAGE (Serial Analysis of Gene Expression, Velculescu 1999), DNA-Microarrays (Schena et al. 1995) oder EST-Sequenzierung (Adams et al. 1991; Boguski et al. 1994) bis hin zu den neuen Hochdurchsatztechniken des „Next-Generation-Sequencing“. Die weltweit bestetablierten Plattformen Roche/454 (Margulies et al. 2005, 454 Life Sciences), Illumina/Solexa Genome Analyzer (Bentley 2006, Illumina) und ABI/SOLID™ (Applied Biosystems) liefern große Mengen an Sequenzdaten in kurzer Zeit und bieten im Bereich der Expressionsanalysen die Möglichkeit umfassender Sequenzierung von Transkriptomen. Da jede Technologie ihre Vor-

und Nachteile hat ist gerade der Vergleich verschieden generierter Datensätze interessant. Die große Datenmenge macht die Auswertung bioinformatisch anspruchsvoll. Hier werden flexible und handhabbare Auswerteverfahren benötigt.

1.1 Evolution des Vertebratenskeletts

Strukturelle Entwicklung

Die Evolution des Vertebratenskeletts lässt sich sehr eindrucksvoll anhand der Taxa der frühen Vertebraten nachvollziehen. Die Agnatha besitzen anstatt eines überwiegend mineralisierten Skeletts nur wenig mehr als knorpelige Elemente die, wichtige Organe, beispielsweise das Gehirn, schützen (Bardack und Zangerl 1968). Schon die primitivsten lebenden Vertebraten, die Haie (*Selachii*) und Knochenfische (*Osteichthyes*) zeigen, die typischen vier Teile des Skelettsystems. Als ektodermale Exoskelett-Strukturen finden sich beispielsweise Zähne oder Schuppen. Wie auch die Knochen des Gesichtsschädels (Viscerocranium), entstehen sie aus Zellen der Neuralleiste. Dagegen sind die übrigen Schädelknochen (Neurocranium) sowie das Rumpf- (Achsen skelett ohne Gesichtsschädel) und das Extremitätenskelett mesodermalen Ursprungs (Donoghue und Sansom 2002; Helms und Schneider 2003; Donoghue et al. 2006).

Die ersten Vorläufer heutiger Skelettsysteme sind zweifellos in Form von Exoskeletten und dem Viscerocambrium der Agnatha, später in Form von Zähnen (Donoghue et al. 2006) zu beobachten. Die ersten Spuren eines mineralisierten Endoskeletts finden sich jedoch in Form von „Aspidin“, einer azellulären mineralisierten Matrix mit organischen Komponenten (Donoghue und Sansom 2002). Dieses vermutlich schon Kollagene enthaltende Vorläuferskelett (Donoghue et al. 2006) taucht erstmals in Form eines Dermalskeletts bei Pteraspidomorphen vor rund 400 Mio. Jahren (Sansom et al. 2005) auf und wird strukturell mit dem Cementum von Vertebratenzähnen verglichen.

Weitere Hinweise auf Vorläufer des Endoskeletts finden sich in Form Ispedin (Basaler lamellarer Knochen)-ähnlicher Strukturen der Galeaspida. Die Gruppe der Osteostraci (Cephalaspidiformes) zeigt nicht nur erstmals zelluläre Knochenstrukturen im Bereich des Dermalskeletts und des Neurocraniums (Denison 1963), sondern auch erste Hinweise auf

Resorption und Knochenerneuerung. Damit zeigen sie schon viele Merkmale des typischen Vertebratenskeletts.

Genetik der Skelettentwicklung

Im Mittelpunkt der Untersuchungen zu den jüngeren genetischen Veränderungen im evolutionären Verlauf der Skelettentwicklung stehen die Vertreter der wichtigsten beteiligten Genfamilien. Diese finden sich im strukturellen Bereich der Extrazellulären Matrix (ECM) und beinhalten Genprodukte der Kollagene, Proteoglykane und weiterer strukturgebender oder regulatorischer Proteine. Regulatorisch stehen SOX-Transkriptionsfaktoren und RUNX2 im Zentrum heutiger Hypothesen. Diese gehen im Verlauf der Skelettentwicklung von mindestens zwei Genomduplikationen aus (Huxley-Jones et al. 2009; Dehal und Boore 2005; Panopoulou und Poustka 2005). In einer raschen molekularen Evolution und Subspezialisierung (He und Zhang 2005) spiegelt sich das Potential solcher Duplikationen zu schnellen Veränderungen (Hakes et al. 2007) wider. Letztlich ist davon auszugehen, dass dies unter anderem in Trennung von Knochen- und Knorpelgewebe erkennbar wird. (Zhang und Cohn 2008).

Der Beginn der molekularen Evolution des Vertebratenskeletts liegt nach heutigen Erkenntnissen in einer sehr frühen Phase der Entwicklung der Chordaten. Als Folge der genomischen Duplikationen kam es in frühen Chordaten zunächst zu einer schnellen Expansion der Anzahl von Proteinen der ECM (Huxley-Jones et al. 2009). Von den bisher 29 verschiedenen Kollagentypen gehören 7 (11 Gene) zum fibrillären Typ. Dabei werden im Allgemeinen die Typen I, V und XXIV dem mineralisierten Knochen, die Typen II, XI und XXVII Knorpelgewebe zugeordnet (Wada et al.). Strukturell und funktionell bilden diese drei Schwestergruppen. Hierbei setzt sich Gruppe 1 aus den Typen Col1A1, Col1A2, Col5A2, Col3A1 und Col2A1 zusammen, Gruppe 2 bilden Col5A1, Col5A3, Col11A1 und Col11A2. Die dritte Gruppe beinhaltet lediglich Col24A1 und Col27A1 (Wada et al.). Insgesamt gehen die Mitglieder der drei Gruppen fibrillärer Kollagene vermutlich auf jeweils eines von drei ursprünglichen Kollagenen zurück. Noch bei den frühen Chordaten, den Tunicaten (ColA), Manteltieren (CiFCol1-3) und Cephalochordaten (BfFCol1) sind diese zu finden (Runnegar 1985; Wada et al.). Eine parallele Aufspaltung der *Hox*-Gencluster A-D und den mit ihnen gekoppelten Kollagenen der Gruppe 1 (Bailey et al. 1997; Morvan-Dubois et al. 2003)

unterstützt die Hypothese genomischer Duplikationsereignisse (Wada et al.; Zhang und Cohn 2008; Wada 2010). Die Vorläufer der Vertebratenkollagene werden schon in der Chorda dorsalis (Wada et al.; Meulemans und Bronner-Fraser 2007) und auch im pharyngialen Endoderm (Rychel und Swalla 2007) von Urchordaten und Cephalochordaten exprimiert. Neben den Kollagenen finden sich ebenfalls Vorläufer anderer struktureller Komponenten der ECM heutiger Vertebraten. So gehen beispielsweise die Mitglieder der SCPPs („secretory Ca-binding phosphoproteins“) der sowohl Osteonectin (SPARC) als auch das „Secreted Phosphoprotein 1“ (SPP1) angehören auf eine ancestrale Form eines einzelnen *SPARC*-Genes zurück (Kawasaki et al. 2004; Kawasaki und Weiss 2006).

Neben den Strukturkomponenten können bei Ur- und Cephalochordaten auch schon einige der wichtigsten regulatorisch wirkenden Proteine beobachtet werden. So besitzen frühe Chordaten schon eine Kopie eines „Hedgehog“-Vorläufers, dem in der weiteren Evolution drei funktionell unterschiedliche Hedgehog-Proteine entspringen: „Sonic“- (Shh), „Desert“- (Dhh) und „Indian-hedgehog“(Ihh) (Shimeld 1999). Nimmt Ihh bei Vertebraten eine zentrale Rolle bei der Regulation der chondralen Ossifikation ein (Minina et al. 2001), so bilden die Transkriptionsfaktoren Sox9 und Runx2 sicherlich die wichtigsten übergeordneten Faktoren der, vor allem pränatalen, Knochen- und Knorpelentwicklung. Die Untergruppe der SoxE-Transkriptionsfaktoren (Sox8, Sox9, Sox10) geht dabei auf ein einzelnes ancestrales Gen zurück. Demensprechend stammen die von Sox9 regulierten Sox5 und Sox6 von einem *SoxD*-Vorläufergen (Rychel und Swalla 2007; Hecht et al. 2008; Larroux et al. 2008). In engem Zusammenhang mit der Ihh- und Sox-Regulation stehen Transkriptionsfaktoren der Runx-Familie, speziell Runx2. In einfacher Kopienzahl im Genom von Cephalochordaten expandiert die Runx-Familie im Laufe der Evolution und enthält bei heutigen Tetrapoden drei Gene (Runx1-3). Der Runx-Vorläufer der frühen Chordaten ist zusammen mit dem Sox-Vorläufer entscheidend bei der Regulation der Expression von Strukturkomponenten der Knochen- und Knorpelvorläufer der Chorda dorsalis und dem pharyngialen Skelett beteiligt (Hecht et al. 2008; Wada 2010). Bei der genomischen Expansion, Subfunktionalisierung und Spezialisierung der Mitglieder der verschiedenen Genfamilien konzentriert sich die Regulation der Sox-Transkriptionsfaktoren auf Knorpelbildung und enchondrale Ossifikation. Runx bleibt essentiell für die Frühphase der Knorpelentwicklung, behält aber die zentrale Rolle in der Knochenbildung sowohl bei der desmalen als auch bei der Mineralisierung der

enchondralen Ossifikation. Mit der zunehmenden Differenzierung von Knochen und Knorpel in der weiteren Entwicklung der Vertebraten entstehen auch neue wichtige Strukturkomponenten die neue, vorteilhafte Eigenschaften in das Skelettsystem einbringen. Eine wesentliche Eigenschaft der Endoskelette ist die Fähigkeit zur Absorption von mechanischen Einwirkungen. Ein wichtiger Aspekt hierfür ist das Zusammenspiel von Hyaluronsäure, Aggrecan und HAPLN1 (Doerge et al. 1986; Doerge et al. 1987; Watanabe et al. 1994). Für Aggrecan, ein Proteoglykan und nicht kollagenöses Knorpelprotein, ist kein Homologes in frühen Chordaten zu finden. Dennoch sind schon hier Proteine vorhanden, die charakteristische Xlink-Domänen enthalten und auf eine Neukombination vorhandener Exons, „exon shuffling“, hindeuten (Upholt et al. 1994; Patthy 2003).

1.2 Knochen-, Knorpelentwicklung und Wachstumsfuge

Generell durchläuft die Bildung des Skelettsystems mehrere Entwicklungsphasen. In der frühesten Phase, oft als Musterbildungsphase bezeichnet, kondensieren zunächst mesenchymale Zellen in Regionen der späteren Knochenbildung und stellen schon jetzt ein vorläufiges Modell des Skelettapparates dar (DeLise et al. 2000; Pizette und Niswander 2001). In der weiteren embryonalen Entwicklung, der Morphogenese, kommt es zur Differenzierung der mesenchymalen Zellen in die drei für die Knochen- und Knorpelbildung spezifischen Zelltypen, Chondrozyten, Osteoblasten und Osteoklasten. Die eigentliche Knochenbildung kann dann auf zwei Arten stattfinden. Bei der desmalen Ossifikation entsteht der Knochen direkt aus mesenchymalem Bindegewebe. Diese mesenchymalen Zellen differenzieren dabei direkt zu knochenbildenden Osteoblasten, welche sich, ausgehend vom primären Ossifikationszentrum, strahlenförmig ausbreiten. Insgesamt werden nur wenige Skelettelemente desmal gebildet. Dazu gehören die Knochen des Schädeldachs, Teile des Gesichtsschädels (u.a. Unterkiefer; mandibula) und das Schlüsselbein (clavicula). Der Großteil des Skelettsystems wird jedoch enchondral gebildet. Hierbei bilden die zu Chondrozyten differenzierten Zellen, ausgehend von den primären und sekundären Ossifikationszentren, zunächst eine knorpelige Vorstufe des Knochens (Primordialskelett), die erst in der weiteren Entwicklung mineralisiert.

Chondrale Ossifikation

Ein klassisches Beispiel für die (en)chondrale Ossifikation ist die Bildung von Röhrenknochen. Die anfängliche Knochenbildung geschieht zunächst in den primären Ossifikationszentren und verlagert sich in der Phase des Längenwachstums in die sekundären Zentren, wie die

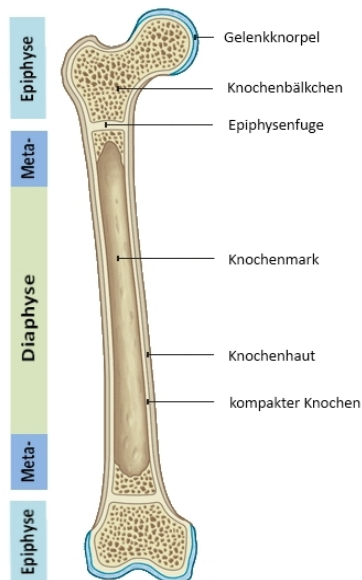


Abb. 1: Schematischer Aufbau des Röhrenknochens

Epiphysenfuge. Der in der frühen Embryonalentwicklung für die Segmentierung und Musterbildung wichtige Transkriptionsfaktor Pax1 hat auch Einfluss auf die Ausbildung der späteren Knochen- und Knorpelanlagen (Ebensperger et al. 1995). Er wirkt über die Induktion des „Sonic hedgehog proteins“ (Shh). Die daraus resultierende Expression verschiedener HMG („high mobility group containing“)-Transkriptionfaktoren und BMPs („Bone Morphogenetic Proteins“) führt zur Kondensation mesenchymaler Zellen im Bereich der späteren Knochenbildung. Zunächst liegen die mesenchymalen Zellen in einer Kollagen I- haltigen ECM kondensiert vor (Karsenty

et al. 2009). Hierbei beeinflussen die BMPs die Packungsdichte und Kondensation der Zellen (Macias et al. 1997; Barna und Niswander 2007). Ein wichtiger Antagonist zu den BMPs ist das regulativ wirkende Noggin (Pizette et al. 2001). Der Transkriptionsfaktor Sox9 („Sex determining region-type HMG box 9“) gehört zur Gruppe der HMG-Domänen-tragenden Proteine und stellt den essentiellen Faktor für die frühe Knochenentwicklung dar. Durch Sox9 aktivierte Kaskaden führen zur weiteren Kondensation und Differenzierung der mesenchymalen Zellen in Chondrozyten. Parallel ändern sich Struktur und Zusammensetzung der ECM. Dies bedeutet zunächst eine Anreicherung von Kollagen Typ II und Aggrecan (de Crombrughe et al. 2001). In der weiteren Entwicklung nimmt die Komplexität der ECM weiter zu und die in Chondrozyten differenzierten Zellen im zentralen Bereich des primären Ossifikationszentrums bilden verstärkt Strukturproteine. Hierbei handelt es sich hauptsächlich um Kollagene, Proteoglykane sowie weitere nicht-kollagenartige Glykoproteine. Die Zellen der peripheren Kondensationszone verbleiben im undifferenzierten Stadium und zeigen nach wie vor eine hohe Expression an Kollagen. Dieses so genannte Perichondrium (Knorpelhaul) beeinflusst entscheidend die Reifung der

Chondrozyten im Zentrum. Die Matrix im zentralen Bereich erfüllt in erster Linie mechanische Funktionen und dient der Anheftung der Zellen. Neben dem Schutz vor Kompression sind viele Komponenten der ECM aber auch regulatorisch an den Wachstums- und Differenzierungsprozessen beteiligt (Koziel et al. 2004). Proteoglykane wie Aggrecan, Decorin und Biglycan bilden Quervernetzungen und beeinflussen die Bildung der Kollagenfibrillen. Diese setzen sich zum größten Teil aus fibrillären und vernetzenden

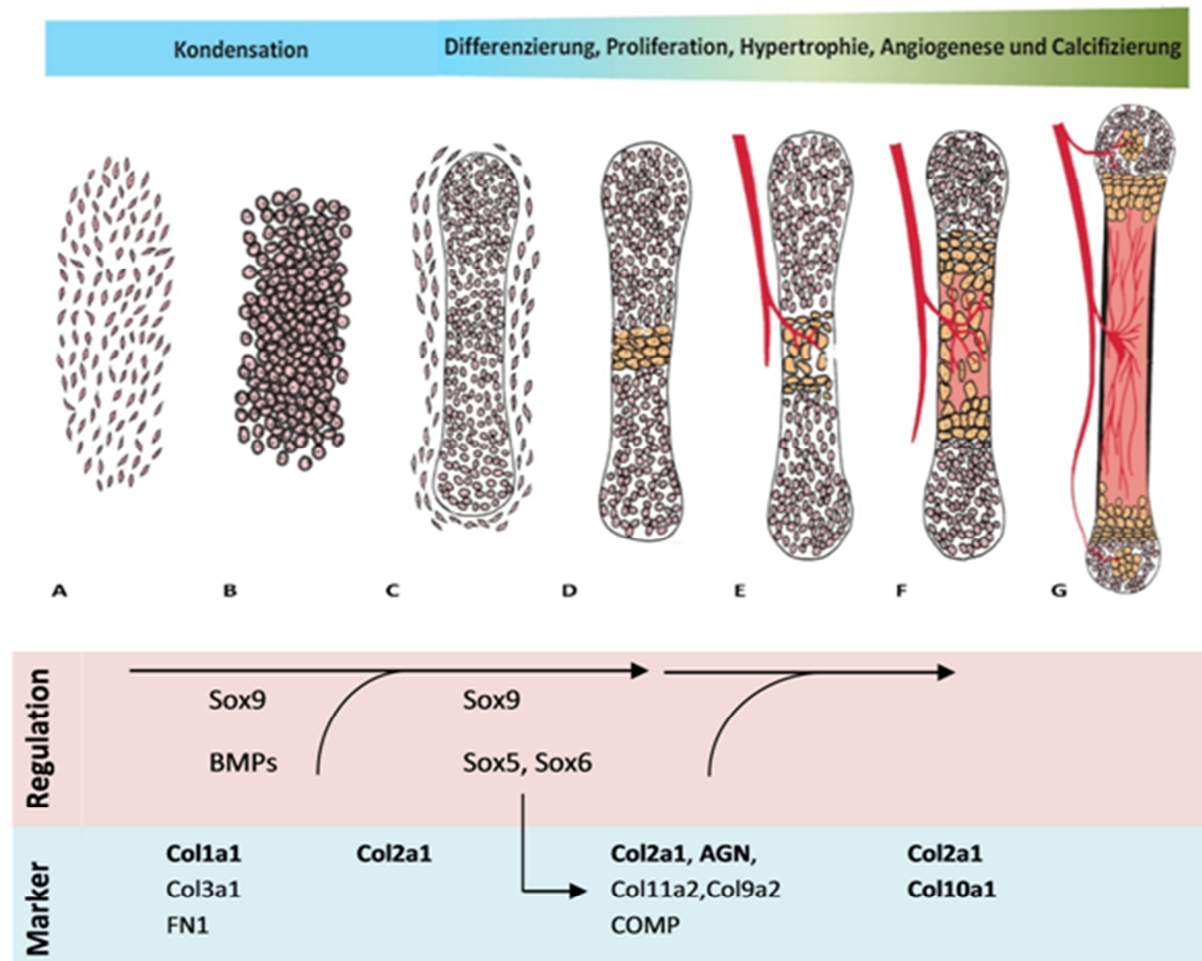


Abb. 2: Frühstadien der enchondralen Ossifikation

Gezeigt sind die Stadien der Bildung eines Röhrenknochens bis zu postnatalen Wachstumsphase und wichtige regulatorische und strukturelle Komponenten, die auch als Marker einsetzbar sind.; A-B Kondensation der mesenchymalen Stammzellen; C Bildung einer ECM und des sogenannten Primordialskeletts; D-E Beginnende Verknöcherung und Invasion der Blutgefäße; F-G Fortschreitende Verknöcherung und Aufspaltung in die sekundären Ossifikationszentren; verändert nach: Gilbert et al. 2006

Kollagenen des Typs II, IX und XI zusammen. Weitere wichtige regulatorische Elemente sind die Sox9 induzierten Transkriptionsfaktoren Sox5 und Sox6 (Lefebvre et al. 1998; Lefebvre et al. 2001). In einem frühen Stadium der Embryonalentwicklung wird transient der Transkriptionsfaktor Runx2 in geringem Maße exprimiert. Dieser nimmt eine zentrale, aber

auch sehr komplexe Rolle in der Regulation der Knochen- und Knorpelentwicklung ein. Temporär, auf geringem Niveau exprimiert, ist er essentiell für die Chondrozytenreifung (Yoshida et al. 2004), dagegen führt die konstitutive Expression im Perichondrium zu entgegengesetzter Wirkung (Hinoi et al. 2006). Nach dieser Phase hoher Proliferation verlassen die Chondrozyten schließlich den Zellzyklus und entwickeln sich über den prähypertrophen in den terminalen Zustand hypertropher Chondrozyten. Hierbei kommt es zu charakteristischen Änderungen der Morphologie und Genexpressionsmuster. Die Zellen steigern ihre Syntheseleistung und nehmen stark an Volumen zu (Hypertrophie). Die Expression vieler Chondrozyten-spezifischer Gene, insbesondere *Sox9*, wird verringert, wohingegen eine erhöhte Expression von „Vascularendothelial Growth Factor“ (VEGF) , „Hypoxia inducible Factor 1 α “ (HIF-1 α) und alkaliner Phosphatase (ALP) stattfindet (Gerber et al. 1999; Schipani et al. 2001; Erlebacher et al. 1995). Obwohl vor allem im prähypertrophen Stadium noch geringe Mengen Kollagen II gebildet werden, wird es von Kollagen X als stadiumspezifischem Markergen abgelöst. Kollagen des Typs X wird ausschließlich von Chondrozyten fortgeschrittenen Reifegrades gebildet (Linsenmayer et al. 1991; Iyama et al. 1991; Karsenty und Wagner 2002). In diesem letzten Abschnitt der Knochenbildung beeinflussen die hypertrophen Chondrozyten die chemotaktische Anziehung der Osteoblasten und Osteoklasten, die Mineralisierung der Knorpelmatrix und die Angiogenese, bis sie schließlich in Apoptose übergehen (Gerber und Ferrara 2000). Diese Vorgänge geschehen in einem engen Zusammenspiel der Zellen des zentralen Bereichs der primären Ossifikationszone und des Perichondriums (Horton 2003). Von großer Bedeutung sowohl in der Osteogenese als auch Chondrogenese ist die Bildung des „Indian Hedgehog“ Proteins (Ihh). In der Chondrogenese von den prähypertrophen Chondrozyten (Bitgood und McMahon 1995) gebildet, ist es zentral in die Kontrolle der Chondrozytenentwicklung involviert. Es fördert die Proliferation der nicht hypertrophen Chondrozyten, inhibiert allerdings auch die weitere Reifung zu hypertrophen Chondrozyten über eine negative Ihh-PTHrP („Parathyroid Hormone Related Protein“)-Rückkopplungsschleife (Vortkamp et al. 1996; St-Jacques et al. 1999; Long et al. 2001; Koziel et al. 2005). Über den Einfluss auf die Chondrozytenentwicklung hinaus ist das „Indian Hedgehog“ Protein entscheidend an der Osteoblastendifferenzierung und –proliferation beteiligt (St-Jacques et al. 1999). Diese entstehen aus mesenchymalen Zellen des Perichondriums und bilden eine perichondrale

Knochenmanschette um die Knorpelanlagen herum (Bianco et al. 1998). Von hier aus findet eine Einwanderung von Osteoblasten und Osteoklasten statt. Während die terminal differenzierten Chondrozyten in Apoptose übergehen, bauen die Osteoklasten einen Großteil der hypertrophen Knorpelmatrix ab. Die verbleibenden Fragmente dienen als Gerüst zum Aufbau der Knochenmatrix (Osteoid) durch Osteoblasten. Im Innern der dichten Knochenmanschette entsteht so eine schwammartige Knochensubstanz, die Spongiosa, in deren Hohlräumen sich das Knochenmark befindet. Mit der zunehmenden Verknöcherung im primären Ossifikationszentrum bildet sich der Knochenschaft (Diaphyse). Die Zone der Knochenbildung bzw. Chondrozytendifferenzierung und –proliferation wird in zwei Fronten aufgespalten und bilden die sekundären Ossifikationszentren. Charakteristisch für diese Wachstumsfugen sind stapelförmig, horizontal angeordnete Chondrozyten der unterschiedlichen Reifestadien. Die Knochenneubildung findet an der Front zur Diaphyse hin statt und ermöglicht das Längenwachstum des Knochens (Hunziker 1994).

Wachstumsfuge, Aufbau und Homöostase

In diesen Zonen der fortgeschrittenen postnatalen Knochenentwicklung findet man zwischen Epiphyse und primärem Ossifikationszentrum alle Stadien der chondralen Ossifikation. Bei der Teilung des primären Ossifikationszentrums durch den neu gebildeten Knochen konzentrieren sich die Chondrozyten der verschiedenen Differenzierungsstadien zur Diaphyse hin in Zonen parallel zum Knochenquerschnitt im Bereich der Knochenenden (Epiphysen). Diese Bereiche, in denen die Chondrozytendifferenzierung und schließlich Osteogenese stattfinden, bilden die Wachstumsfugen. Im Zuge der Wachstumsphase werden diese Regionen insgesamt immer dünner, die Bereiche der verschiedenen Differenzierungsstadien immer schmaler bis zu einem inaktiven Zustand am Ende der Wachstumsphase. Entscheidend für die komplexe Regulation der Proliferation und Reifung der Chondrozyten bis hin zu ihrem Absterben und damit einhergehender Verknöcherung sind sowohl in der Wachstumsfuge selbst gebildete, als auch vom Perichondrium ausgeschüttete Moleküle.

Am Beispiel der Wachstumsfuge eines 17 Tage alten Fötus (Shum und Nuckolls 2002) sind deutlich die unterschiedlichen Regionen erkennbar. Schon im Fall einer 28 Tage (postnatal)

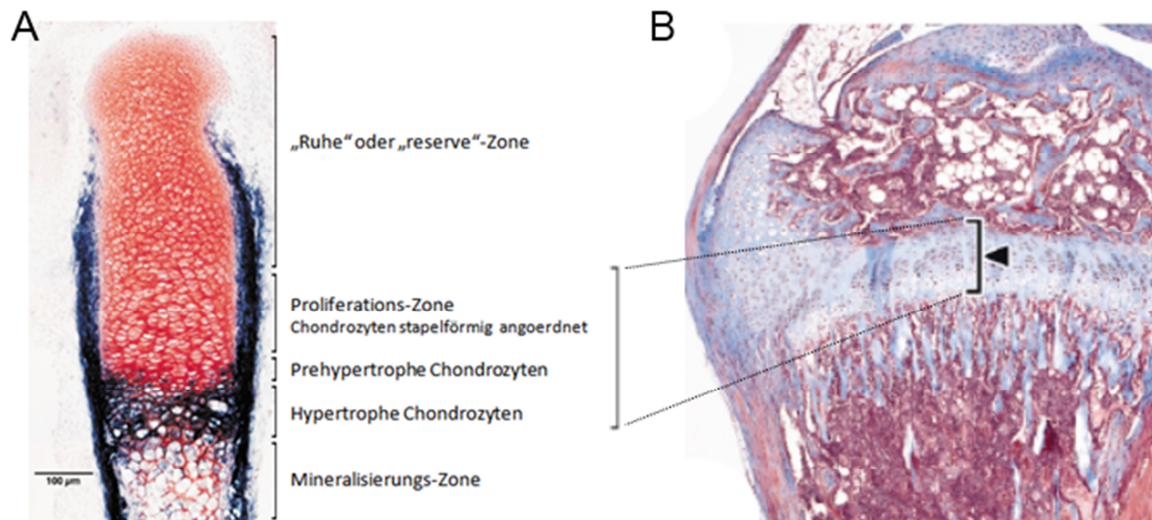


Abb. 3: Aufbau der Wachstumsfuge

A murine fötale Wachstumsfuge mit distinkten Bereichen unterschiedlicher Chondrozytendifferenzierung; B posnatale (28 Tage) murine Wachstumsfuge; verändert nach Yakar et al. 2002(rechts) Shum und Nuckolls 2002 (links)

alten Maus (Yakar et al. 2002) ist der Volumenanteil der Wachstumsfuge wesentlich reduziert. Die Proliferationsrate der Chondrozyten nimmt im Laufe der Wachstumsphase ab, bis diese letztendlich abgeschlossen ist und die Epiphyse nahezu mit dem Knochenschaft verschmilzt (Weise et al. 2001; Nilsson und Baron 2004; Emons et al. 2011). In der aktiven Wachstumsfuge bildet die erste und zugleich größte Region distal gelegen, direkt im Anschluss an den hyalinen Knorpel, die sogenannte Ruhezone. Die hier befindlichen „Reserve“-Chondrozyten gehen aus immer wieder kondensierenden Mesenchymalen Stammzellen hervor. Sie bilden ein Reservoir an Vorläuferzellen. Die Chondrozytdichte in diesem Bereich ist geringer, die Zellen sind gleichförmig und zeigen eine nur sehr geringe Teilungsaktivität (Hunziker 1994; Abad et al. 2002; Horton 2003). Analog zum zeitlichen Verlauf der Knochenbildung im primären Ossifikationszentrum, in der pränatalen Phase, beginnen die Zellen zur Metaphyse hin zu proliferieren. Dabei ordnen sie sich zunehmend in stapelähnlichen Strukturen an und exprimieren Proteine der extrazellulären Matrix (de Crombrugge et al. 2000; Ballock und O'Keefe 2003). Nach Verlassen des Zellzyklus und einsetzender Reifung wird verstärkt alkalische Phosphatase (AP-2) und Kollagen Typ X gebildet (O'Keefe et al. 1997; Lewinson et al. 1982). Außerdem findet man in dieser Region der prähypertrophen und frühen hypertrophen Chondrozyten einen zentralen Punkt in der Regulation des Knochenwachstums. Zugrunde liegt eine grundsätzliche spontane Reifung der

Chondrozyten (Ballock und O'Keefe 2003). Daher wird ein komplexes Regulationsnetzwerk benötigt. Einer der wichtigsten Mechanismen ist dabei die sogenannte „Ihh/PTHrP-Feedback-Schleife“ (Vortkamp et al. 1996; St-Jacques et al. 1999; Horton 2003; Ballock und O'Keefe 2003). Dabei produzieren (prä-)hypertrophe Chondrozyten vermehrt „Indian hedgehog“-Protein (Ihh). Dieses wirkt über die Bindung an seinen Rezeptor „Patched“ (Ptc), der vor allem auf der Oberfläche der perichondralen Zellen exprimiert wird. Die als Reaktion auf das Ihh-Signal aktivierten Transkriptionsfaktoren der Gli-Familie lösen die Expression und Sekretion des „Parathyroid Hormone related Protein“ (PTHrP) in Zellen des am Knochenende (juxta-ariikulär) liegenden Perichondriums aus. Dieses diffundiert von dort in Richtung Knorpelschicht, wobei die Konzentration in diese Richtung abnimmt. PTHrP bindet an seinen Rezeptor, der hauptsächlich in proliferierenden Chondrozyten überexprimiert wird. Dies bewirkt eine Hemmung der Proliferation und Reifung der Chondrozyten (Kronenberg et al. 1997; Karp et al. 2000). Ist die PTHrP-Konzentration gering, beginnen die Chondrozyten zu proliferieren und gehen in den hypertrophen Zustand über. Parallel zur Aktivierung der PTHrP-Expression bilden die Zellen des nahen Perichondriums den „Transforming Growth factor β “ (TGF- β) (Alvarez et al. 2001). TGF- β kann sowohl indirekt als auch direkt in die Chondrozytendifferenzierung eingreifen. Indirekt verstärkt es die PTHrP-Expression im periartikulären Bereich (Serra et al. 1999; Alvarez et al. 2001), kann aber auch direkt inhibitorisch auf Chondrozytenreifung wirken (Ballock et al. 1993; Alvarez et al. 2002). Speziell in der postnatalen Wachstumsphase könnte die regulatorische Wirkung des TGF- β entscheidend sein. Die Aktivierung der TGF- β -bedingten Kaskaden geschieht dabei über einen heteromeren Rezeptor bestehend aus zwei transmembranen „Serin/Threonin-Kinase“-Rezeptoren, TGF β RI und TGF β RII (Derynck und Zhang 2003). Analog den TGFs, zu denen neben TGF- β auch Activin oder „Bone Morphogenic Proteins“ (BMPs) gehören, bilden auch die zugehörigen Rezeptoren eine sehr heterogene Gruppe. Dabei bleibt der dimere Aufbau jedoch immer erhalten (Boesen et al. 2002; Greenwald et al. 2004; Feng und Derynck 2005). Mitglieder der TGF- β Familie spielen bei den meisten Zellpopulationen eine zentrale Rolle bei der Zellproliferation und Morphogenese. Intrazellulär kann das Signal entweder über SMAD-Faktoren oder unabhängig davon weiter gegeben werden (Derynck und Zhang 2003). Im Allgemeinen verläuft die SMAD-abhängige Signaltransduktion im Fall von TGF- β 1-3 über SMAD2 und SMAD3. Dabei bindet zunächst die phosphorylierte Form des

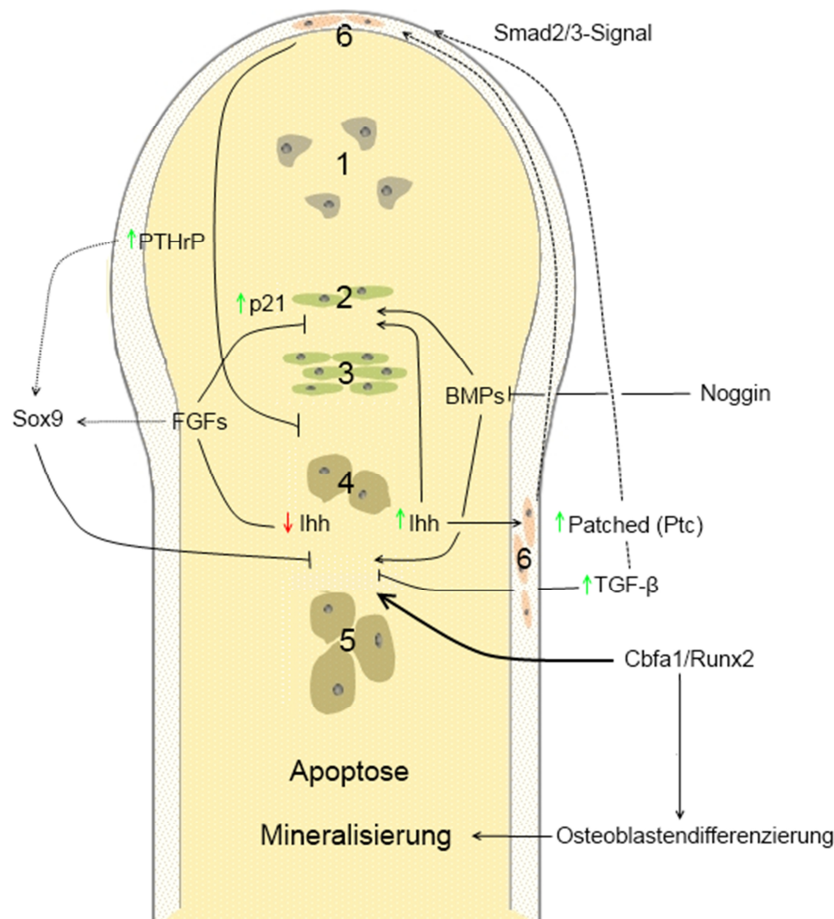


Abb. 4: Regulation der Chondrozytendifferenzierung in der Wachstumsfuge

Schematischer Aufbau eines Röhrenknochens mit für den Knochenaufbau wichtigen Zellpopulationen, deren Zusammenspiel und einen Überblick über die wichtigsten Regulationsmechanismen; den stilisierten Knochen umgebend befindet sich das Perichondrium; 1 Chondrozyten der Reservezone; 2-3 proliferierende Chondrozyten; 4 prähypertrophe und 5 hypertrophe Chondrozyten; 6 perichondrale Zellen;

rezeptorassoziierten SMAD (R-SMAD) einen SMAD-Co-Faktor (Co-Smad). Meist handelt es sich dabei um Smad4. Nach Translokation in den Zellkern wirkt dieser Komplex als Transkriptionsfaktor und steuert die Expression der Zielgene (Massagué et al. 2005; Keller et al. 2011). Speziell SMAD3-Signaltransduktion spielt eine zentrale Rolle in der Inhibition der Chondrozytenreifung (Ferguson et al. 2000; Pateder et al. 2000; Pateder et al. 2001). Unabhängig davon kann TGF- β auch direkt über MAPK-, P38-, JNK- oder Erk1/2-Regulationsmechanismen wirken (Adachi-Yamada et al. 1999; Nakamura et al. 1999; Massagué 2000; Miyazono et al. 2001). Außerdem können TGF- β -Proteine beispielsweise durch Bindung an andere Proteine der ECM zunächst inaktiv vorliegen und erst bei Spaltung durch ein drittes Enzym aktiv werden (Gleizes et al. 1997; Pedrozo et al. 1998; Horton 2003). Einen weiteren wichtigen Teil des regulatorischen Netzwerks stellen die sogenannten FGFs,

„Fibroblast Growth Factors“, dar. Die entsprechenden Rezeptoren Fgfr1-3 werden in verschiedenen Stadien und Zellpopulationen der Wachstumsfuge exprimiert. Während FGFR1 und 2 hauptsächlich in der pränatalen enchondralen Ossifikation bei der Kondensation der mesenchymalen Stammzellen eine Rolle spielen, tritt FGFR3 erst mit Beginn der Chondrogenese auf. Im Folgenden bleibt der Rezeptor mit seinen 9 möglichen Liganden während der gesamten Wachstumsphase eine wichtige Kontrollinstanz (Peters et al. 1993; Szébenyi et al. 1995; Ornitz und Marie 2002). Im Gegensatz zu FGFR1, der hauptsächlich in prähypertrophen und hypertrophen Chondrozyten exprimiert wird, findet man FGFR3 bei Chondrozyten der Ruhe- und Proliferationszone (Deng et al. 1996; Ornitz 2005). Interaktionen mit dem FGFR3 dürften zu den wichtigsten negativen Regulationsmechanismen des linearen Knochenwachstums gehören (Deng et al. 1996; Horton 2003). So konnten bei „Knockout“-Mäusen wesentlich verlängerte Knochen mit ausgedehnten Zonen von proliferierenden und hypertrophen Chondrozyten nachgewiesen werden. Bei Patienten mit Krankheitsbildern, die mit verkürzten Knochen einhergehen, konnte dagegen eine hohe FGFR3-Expression festgestellt werden (Shiang et al. 1994; Park et al. 1995; Chen et al. 1999; Vajo et al. 2000). Dabei greift FGFR3 in vielfältiger Weise in die Chondrozytenentwicklung ein. In ruhenden und proliferierenden Chondrozyten wird die Teilung direkt über die Aktivierung der Stat1-Kaskade und der Induktion des antimitotischen Faktors p21 gehemmt (Chin et al. 1996; Su et al. 1997). Über eine MAP-Kinase-Aktivierung kann die Expression von Sox9 gesteigert werden (Murakami et al. 2000). Die Expression von Ihh wird gehemmt (Naski et al. 1998). Antagonisten ähnlich können die „Bone Morphogenetic Proteins“ (BMPs) gesehen werden (Minina et al. 2002). Auch diese Proteine gehören zur TGF- β -Familie, wirken aber im Gegensatz zu TGF- β 1-3 meist über Smad1, 5 und 8 oder die inhibitorischen Faktoren Smad6 und 7 (Derynck und Zhang 2003; Massagué et al. 2005; Wan und Cao 2005). Außer über SMAD-abhängige Signaltransduktion können BMPs auch andere Transkriptionsfaktoren, wie beispielsweise JunB, JunD, ID oder Mitglieder der DLX-Familie induzieren (Xu et al. 2001; Goldring et al. 2006). Primär werden die meisten BMPs von perichondralen Zellen gebildet. Lediglich BMP-7 wird ausschließlich von proliferierenden Chondrozyten exprimiert und BMP-6 findet man nur im Bereich von hypertrophen Chondrozyten (Minina et al. 2002; Goldring et al. 2006). Im Bereich der Ruhe- und Proliferationszone erhöhen sie die Proliferationsrate. Gleichzeitig wird die terminale

Differenzierung direkt und auch indirekt durch eine erhöhte Expression von Ihh im prähypertrophen Bereich inhibiert (Minina et al. 2001). Die Überexpression des BMP-Antagonisten Noggin führt zu drastisch eingeschränktem Knochenwuchs (Zou und Niswander 1996; Capdevila und Johnson 1998; Pathi et al. 1999).

Hauptsächlich zuständig für die Regulation der Differenzierung vor der terminalen/hypertrophen Phase sind FGF(R)s und BMPs auch bei der Induktion der letzten Entwicklungsphase der Chondrozyten beteiligt. So stimuliert FGF2 zusammen mit BMP2 beispielsweise Chondrozyten zur Bildung von „Secreted Phosphoprotein 1“ (SPP1), einem Marker für hypertrophe Chondrozyten (Minina et al. 2002). Im Perichondrium und der prähypertrophen Zone exprimiert ist der Transkriptionsfaktor Runx2 essentiell für die terminale Chondrozytenentwicklung (Enomoto et al. 2000; Zheng et al. 2003). Dieser führt die Chondrozyten unter anderem im Zusammenspiel mit einem BMP-induzierten Smad1-Signal in das letzte Reifungsstadium (Leboy et al. 2001). Gleichzeitig löst er die Differenzierung von mesenchymalen Stammzellen aus, die für die Mineralisierung des Knochens benötigt werden. In diesem Stadium sezernieren die Chondrozyten verstärkt Col10A1, Annexine (II, V, VI), Strukturkomponenten wie Osteocalcin und Osteonectin (SPP1), aber auch Matrix-Metalloproteinasen wie MMP-13 (Kirsch et al. 2000a; Zheng et al. 2003; Goldring et al. 2006). Damit beginnen der Umbau der ECM und die Vorbereitung der Angiogenese, die Einwanderung der Blutgefäße und die eigentliche Verknöcherung. Metalloproteinasen bauen Teile der Knorpelmatrix ab (Mwale et al. 2002). Nun können Osteoklasten und Osteoblasten einwandern. MMP-9 und 13 spielen dabei nicht nur eine Rolle beim Abbau der ECM, sondern können auch regulatorisch wirken (D'Angelo et al. 2001). So bewirken MMPs wie beispielsweise MMP-9, das von Osteoklasten sezerniert wird, die Freisetzung von VEGF („Vascular Endothelial Growth Factor“) und sind maßgeblich an der Apoptoseinduktion der Chondrozyten beteiligt (Vu et al. 1998; Zelzer et al. 2004; Maes et al. 2004; Ortega et al. 2004). Noch bevor die Blutgefäße in den Knochen einwandern, beginnt die Mineralisierung im interzellulären Raum zwischen Stapeln hypertropher Chondrozyten (Johnstone et al. 2000). Dies geschieht durch Matrixvesikel, die von Chondrozyten oder Osteoblasten abgegeben werden. Die Anreicherung von Ca^{2+} in den Vesikeln geschieht aktiv über in der Vesikelmembran verankerte Ca-Kanäle, den Annexinen

(Kirsch et al. 2000a)Kirsch 2005 #216}. Diese sind mit verantwortlich für die Bindung an das Kollagengerüst (Typ II und X), anschließend wird das Ca^{2+} freigesetzt (Kirsch et al. 2000b).

1.3 Genexpressionsanalyse und neue Sequenziertechnologien

Für die Analyse biologischer Vorgänge auf Transkriptionsebene steht eine breite Palette von Verfahren zur Verfügung. Dabei stehen meist die Quantifizierung und der Vergleich von Expressionsstärken im Mittelpunkt. Eine der ältesten Methoden und zugleich eine der wenigen, mit der die RNA bestimmter Transkripte direkt nachgewiesen werden kann, ist die „Northern blot“-Hybridisierung (Alwine et al. 1979; Rohde und Säger 1981). Voraussetzung hierfür ist die gelelektrophoretische Auftrennung der RNA, der Transfer der RNA auf eine Membran und die Hybridisierung mit einer entsprechenden RNA- oder DNA-Sonde. Detektiert werden können dann Transkripte sowie Größenvariationen (Spleiß-Varianten). Ähnlich wie bei Sonden-Hybridisierungen in vivo ist auch eine absolute Quantifizierung bei „Northern blots“ nur bedingt möglich. Alternativ und etwas sensitiver, aber ohne die Möglichkeit Spleiß-Varianten zu erkennen, ist der „(ribo)nuclease protection assay“ (RPA) (Roth 2002). Dabei wird die zu detektierende RNA durch Hybridisierung mit einer komplementären Sonde vor dem Abbau durch eine einzelstrangspezifische Nuklease geschützt und anschließend elektrophoretisch aufgetrennt. PCR-basierende Methoden wie die (semi-)quantitative RT-PCR („reverse transcriptase PCR“, Delidow et al. 1989) und die daraus resultierende quantitative „Real time PCR“ (qPCR) können sehr schnell relativ genaue Ergebnisse (qPCR) liefern, zeigen aber ähnliche Schwächen wie „Northern-Analyse“ und RPA bei der umfassenden Analyse der Genaktivitätskomplexer Entwicklungsvorgänge, wie dies beispielsweise bei der Knochen- und Knorpelentwicklung der Fall ist. Als Alternative kommen die Hochdurchsatz-Analysemethoden wie „Expressed Sequence Tags“- (EST)-Sequenzierungsprojekte (Adams et al. 1991), „Serial Analysis of Gene Expression“ (Velculescu et al. 1995) (SAGE) und die Microarray-Technologie (Schena et al. 1995; Lipshutz et al. 1999) in Frage. Die beiden ersten Methoden benutzen zur Identifizierung und Quantifizierung der Transkripte lediglich mehr oder weniger kurze Sequenzabschnitte der in cDNA umgeschriebenen RNA. Dabei wird im EST-Verfahren versucht, möglichst Vollängen-cDNA-Klone in großem Ausmaß herzustellen und zunächst nur jeweils mit einer

Sequenzierung („Tag“) nach zu charakterisieren. Jeder „Tag“ entspricht einem einzelnen mRNA-Molekül. Die Methode erlaubt es, unbekannte Transkripte zu entdecken und bei ausreichender Anzahl an Sequenzen die Transkripte zu quantifizieren. Ähnlich verhält es sich mit der SAGE-Methode. Allerdings werden hier jeweils sehr kleine Abschnitte (10-14 bp) eines Transkripts, durch Linkersequenzen getrennt, in einem Klon kombiniert zusammengefasst. Dies spart Sequenzierungskosten und Zeit. Die Microarray-Technologie beruht auf Sonden-Hybridisierung. Deshalb müssen bei diesem Verfahren auf jeden Fall die Ziel-Sequenzen bekannt sein. Ganze Gene und Spleiß-Varianten müssen durch mehrere verschiedene Sonden abgedeckt. Unbekannte Varianten sind nur schwer oder gar nicht nachweisbar. Außerdem sind lediglich relative Quantifizierungen möglich. Obwohl die Methode durchaus mehrere potentielle Fehlerquellen beinhaltet und daher in den Anfangsjahren in Hinblick auf statistische Absicherung, Probenvorbereitung, cDNA-Synthese und Fluoreszenzfarbstoffe Entwicklungsarbeit notwendig war, sind Microarrays heute ein zuverlässiges Werkzeug. In großen Bereichen, in denen weitgehend bekannte Transkriptome der Analyse zugrunde liegen und viele Zustände zu testen sind, ist diese Technologie zur Zeit die schnellste und günstigste Methode. Mit der Markteinführung der 2. Generation der Sequenziertechnologien stehen neue Verfahren zur Verfügung, die zu einer schnellen und kostengünstigen Analyse bieten, zum anderen eine absolute Quantifizierung und bisher nie da gewesene Sequenzdaten.

Sequenzierung der nächsten Generation

Mit der Markteinführung der 454/Pyrosequenzier-Technologie (Roche) beginnt die Zeit der sogenannten Sequenzierung der 2. (nächsten) Generation. Unter einer Vielzahl von neuen Ansätzen setzten sich die Plattformen der Firmen Illumina (Genome Analyzer), Roche (454/FLX) und Applied Biosystems (SOLID) durch. Gestecktes Ziel für die nähere Zukunft ist hierbei die Sequenzierung ganzer Genome zu einem Preis von nicht mehr als \$1000. Das auf Daten des „National Institute of Health“ basierende Diagramm verdeutlicht nochmals die Preisentwicklung der letzten Jahre und die enorme technische Weiterentwicklung die sich darin widerspiegelt. Angefangen mit der Pyrosequenzierung, umgesetzt in der 454/GFX

Technologie der Firma Roche, tragen immer neue Technologien zu den stetig sinkenden Sequenzierkosten bei (Abb. 5).

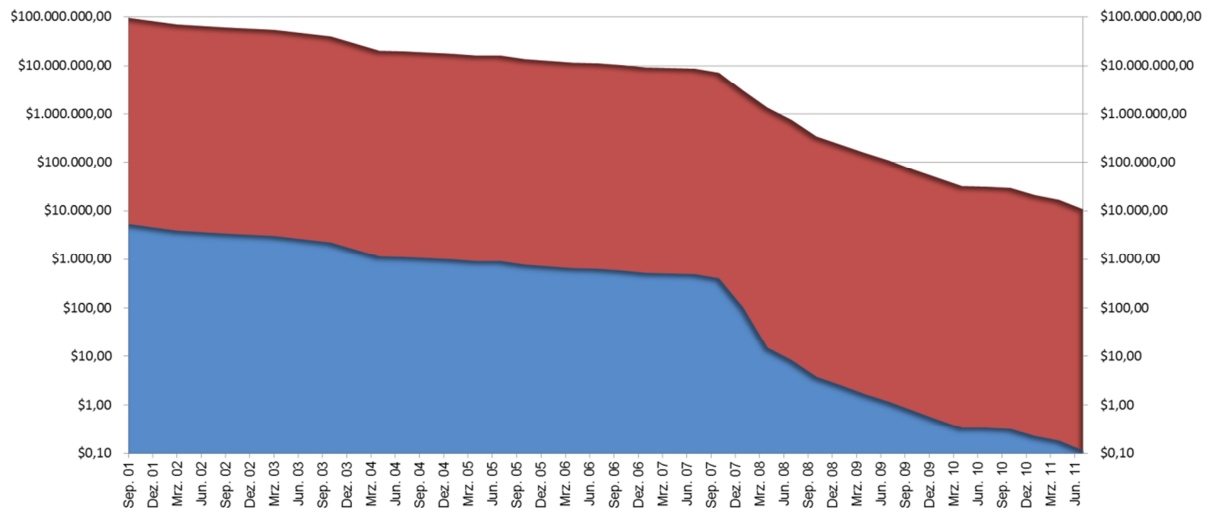


Abb. 5: Entwicklung der Kosten für Hochdurchsatzsequenzierungen
Dargestellt ist die Preisentwicklung der letzten 10 Jahre für die Sequenzierung von einem menschlichen Genom (rot) und pro Megabase (blau)
(Quelle: NIH)

Bei der Pyrosequenzierung wird bei der Synthese des komplementären DNA-Stranges nicht direkt das eingebaute Nukleotid, sondern das dabei abgespaltene Pyrophosphat detektiert (Ronaghi et al. 1996). Dies geschieht mittels einer einfachen Luciferase-Lichtreaktion. Da nicht zwischen dem Einbau verschiedener Nukleotide unterschieden werden kann, ist es notwendig, den Einbau verschiedener Nukleotide zeitlich zu trennen. Daher umfasst jeder Sequenzierzyklus vier Phasen, in denen jeweils ein anderes Nukleotid verwendet wird. Zur Anreicherung (Amplifikation) und räumlichen Trennung werden die DNA-Fragmente zunächst mit Adaptoren versehen und im Anschluss an die Oberfläche von sehr kleinen Kugelstrukturen, so genannten „Beads“, gebunden). Sequenzierung längerer homopolymerer Bereiche problematisch ist. Ebenfalls kritisch ist die Sequenzierung polyadenylierter Moleküle. Aufgrund der oftmals Oligo-d(T) gepimten cDNA-Synthese enthalten viele Fragmente unterschiedlich lange polyT-Bereiche. Dabei darf pro Kugel lediglich ein DNA-Molekül binden. Danach erfolgt die klonale Amplifikation der DNA-Fragmente in abgeschlossenen Mikroreaktionsräumen einer Emulsions-PCR (Dressman et al. 2003). Anschließend werden die mit DNA beladenen „Beads“ in mikroskopisch kleinen Vertiefungen einer sogenannten „Nanowell“-Platte verteilt. Während der Sequenzierung

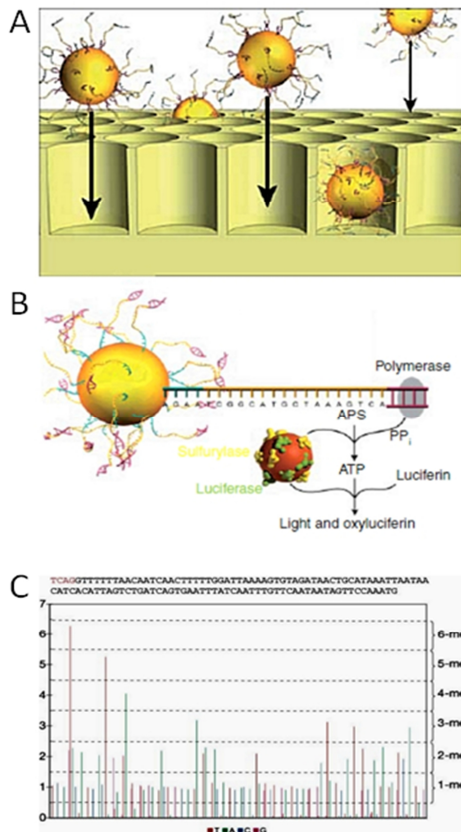


Abb. 6: 454/Pyrosequenzierung
 A „bead“-gekoppelte DNA-Fragmente und 454-Picotiterplatte; B Lichtreaktion bei Nukleotideinbau; C Pyrogramm; (verändert nach Owen-Hughes und Engholm 2007)

werden Lichtblitze beim Einbau der Nukleotide erfasst und in eine zusammenhängende Nukleotidsequenz zusammengefasst. Das in Abb. 6C dargestellte „Flow“- oder „Pyro“-Gramm gibt die Intensität der Lichtemission einer bestimmten Koordinate auf der Nanowell-Platte wieder. Die Stärke der Intensität entspricht der Zahl der eingebauten dNTPs bis zu einem gewissen Grad. Mit zunehmender Zahl des gleichen eingebauten Nukleotids wird die Korrelation allerdings unsicherer, so dass die bei der Sequenzierung kann dies zu sehr starken Lichtemissionen führen und dadurch weite Bereiche der Nano-/Picotiterplatte für viele Zyklen unbrauchbar machen. Daher müssen besondere Strategien angewandt werden, um zu lange polyA-Bereiche zu vermeiden. Eine von mehreren Möglichkeiten ist die Verwendung von speziellen Oligo-d(T)-Primern, die am 3' Ende noch zusätzlich „Ankerbasen“ (5' TTTTTTVN 3') enthalten und somit

direkt am Übergang zur Polyadenylierung binden. Kombiniert wird diese Methode oft mit einer zusätzlichen Restriktionsschnittstelle (Ng et al. 2006; Ng et al. 2007). Des Weiteren ist es möglich, modifizierte Oligo-d(T)Primer einzusetzen. Die nicht ausschließlich aus Thyminen bestehenden Primer bringen in gewissen Abständen andere Nukleotide ein. So werden homopolymere Bereiche von ausschließlich A/T kurz gehalten (Beldade et al. 2006). Derzeit können mit Hilfe der „Titanium“-Chemie Leseweiten bis zu 800 nt erreicht werden.

Mit einem gänzlich anderen Ansatz stellt sich die SOLiD™ („Sequencing by Oligo Ligation and Detection“)-Technologie der Firma Applied Biosystems/ „Life technologies“ dar. Ähnlich dem 454/FLX-System geschieht die Amplifikation der mit Adaptoren versehenen Fragmente über emPCR an magnetische „Beads“. Diese werden, auf der Oberfläche eines Glaträgers verteilt, immobilisiert. Bei der Sequenzierung werden degenerierte, fluoreszenzmarkierte

Oligonukleotide, meist Oktamere, eingesetzt. Dabei sind definierte Basen an festen Positionen des Oktamers (1, 2 oder auch 4, 5) durch die Fluoreszenzmarkierung des Oligonukleotids in einem Vier-Farbensystem kodiert (Abb. 7A). Ausgehend vom Sequenzierprimer an der Position n hybridisiert das nächste passende Oktamer und wird über eine Ligase mit dem bestehenden Zweitstrang verknüpft, bevor die

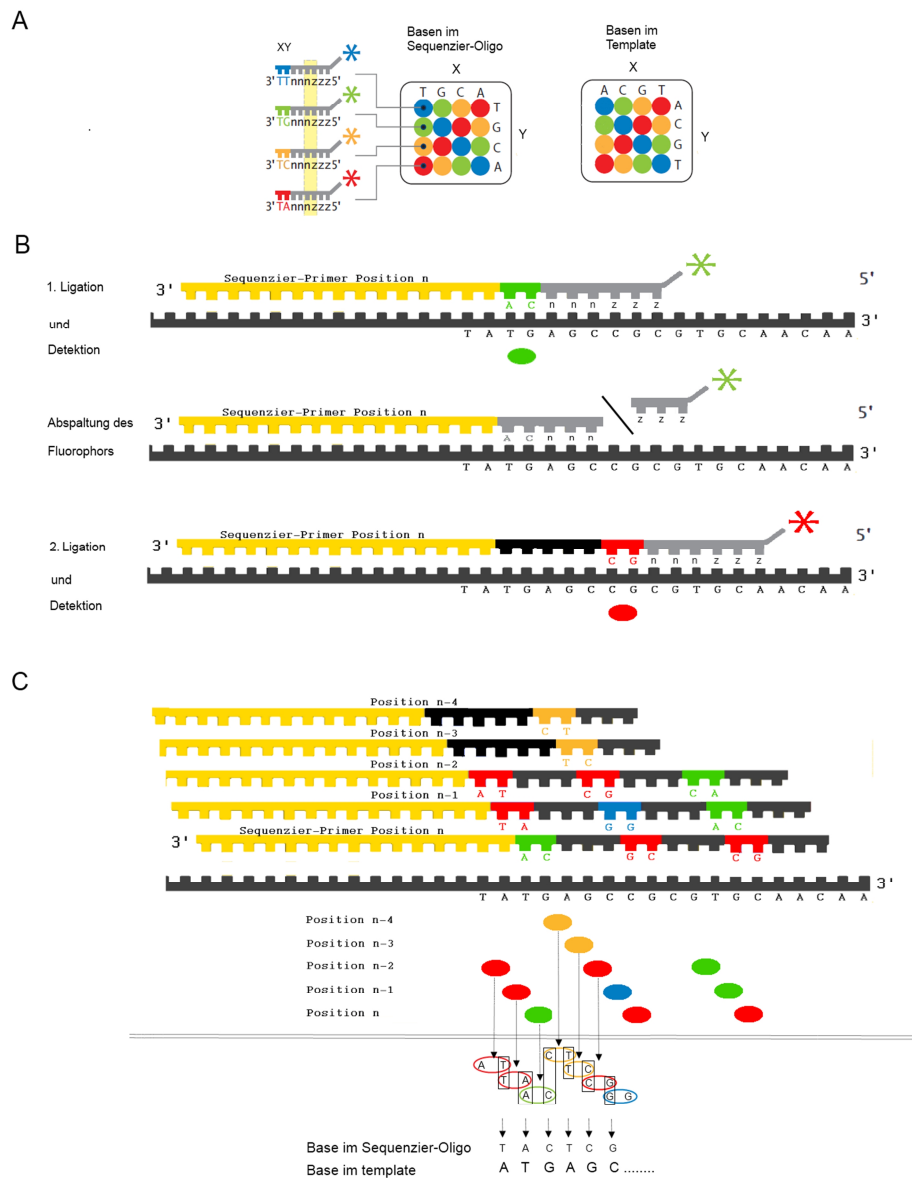


Abb. 7: Prinzip der SOLiD™-Sequenzierung

A Dekodierungskarte für die Fluoreszenzmarkierung der Oligonukleotide (verändert nach: Mardis 2008a) B Schematischer Ablauf über 5 Runden der Sequenzierung; grau (unten) Templateausschnitt; Sequenzierprimer (gelb) und Oligonukleotide in der jeweiligen Farbmarkierung; C mehrere Runden Ligations-Runden mit verschiedenen Start-Primern führen zu einer überlappenden Sequenzierung die das „Base calling“ erst ermöglicht.

Fluoreszenzdetektion beginnt. Nicht verlängerte Moleküle werden durch Dephosphorylierung für weitere Ligationsreaktionen blockiert. Verlängerte Stränge besitzen nach Abspaltung der drei terminalen Inosine und dem Fluorophor (Abb. 7A zzz*) wieder eine freie 5'Phosphatgruppe. Diese wird im nächsten Zyklus der Sequenzierung um ein weiteres Oligonukleotid verlängert.

Nach Abschluss der ersten Runde, ausgehend vom Primer an der Position n , ist noch keine Dekodierung möglich. Für jede Fluoreszenzmarkierung existieren vier Kombinationen für die Nukleotide an Position 1 und 2 des Oktamers. Außerdem ist die Sequenz nur lückenhaft abgedeckt. Um immer überlappende Sequenzinformationen über die gesamte Sequenz zu erhalten, werden weitere vier Runden sequenziert. Dabei wird die Startposition um jeweils eine Position zum 5'-Ende hin verschoben. Durch dieses System wird jede Base zweimal sequenziert. Fehler bei der Sequenzierung werden sofort deutlich und können direkt aus der Analyse entfernt werden. Die daraus folgende sehr geringe Fehlerrate macht die Methode besonders interessant für die Detektion von SNPs. Nachteile sind die kurzen Leseweiten von 50-75 nt und die sehr aufwändige Bioinformatik.

Mit bis zu 50 Gb Sequenzinformation pro Tag liegt die Illumina Plattform noch über den Spezifikationen des SOLiDTM-Systems und weit über denen der Roche-454/FLX-Systeme.

Dennoch dürften der Genome Analyzer Iix und HiSeq 2000TM mit Leseweiten von bis zu 150 nt und moderaten Fehlerraten zur Zeit weltweit den größten Anteil an großen Sequenzier-Projekten haben. Grundlage für diese Technologie ist, wie bei der 454-Technologie, die Neusynthese von DNA. Ein wesentlicher Unterschied ist die Methode der Proben Amplifikation und Separation. Wird bei 454/FLX- und SOLiD auf die Emulsions-PCR und Picotiterplatten zurückgegriffen, so findet die spezifische Amplifikation bei den Illumina-Systemen erst nach dem Immobilisieren der DNA-Fragmente auf der Oberfläche des Reaktionsraums, der „Flow Cell“, statt. Für die sogenannte „Bridge Amplification“ (Fedurco et al. 2006) werden Fragmente benötigt, die an den Enden unterschiedliche Adapter tragen.

Dies wird durch spezielle Y-Adapter (Abb. 8 A) gewährleistet. Die beiden Stränge bilden über einen kleinen Bereich homologer Sequenzen einen Doppelstrang, während der größere Teil verschieden ist und kein Doppelstrang ausbilden kann. Werden diese nun mit den doppelsträngigen DNA-Molekülen ligiert, wird jeder Einzelstrang für sich die beiden

unterschiedlichen Adaptersequenzen tragen. Durch wenige Runden einer PCR mit adapterspezifischen Primern werden diese Fragmente noch weiter angereichert. Nach der Denaturierung hybridisieren die linearen Fragmente mit beiden Enden der Adapter an komplementäre Oligonukleotide auf der „Flow Cell“-Oberfläche (Abb. 8B). Dabei wird auf einen statistischen Abstand zwischen einzelnen Molekülen geachtet, damit Amplifikations-Cluster verschiedener Fragmente sich nicht überschneiden. Bei der „Bridge-Amplification“ folgt in mehreren Runden auf eine Synthesephase, in der der Zweitstrang synthetisiert wird, eine Denaturierung. Da jeweils an das andere auf der Oberfläche der „Flow Cell“ verankerte Oligonukleotid angeknüpft wird, ist der Zweitstrang nun auch kovalent an die Oberfläche

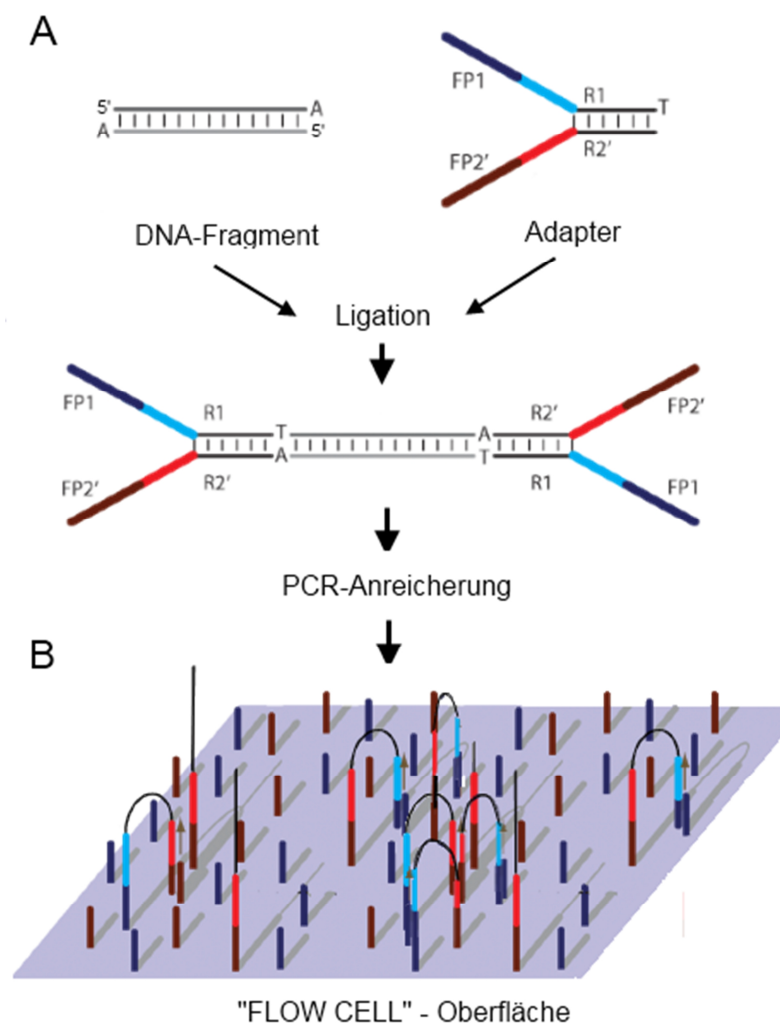


Abb. 8: Sequenzierbibliothek und "Bridge Amplification" der Illumina-Technologie

A DNA-Fragment und Adapter; Adapter mit komplementären Sequenzen R1 und R2, unterschiedlichen „Flow Cell“-Bindesequenzen (FP1 und FP2') sowie zu den Sequenzierprimern („read1“ und „read2“) komplementäre Sequenzen (rot und blau); Konstrukt nach Adapterligation: B „Flow Cell“-Oberfläche mit gebundenen DNA-Fragmenten im Laufe der „Bridge-Amplifikation“ (verändert nach Kozarewa und Turner 2011)

gebunden und kann im nächsten Amplifikationsschritt mit dem anderen Fragmentende an ein freies Oligo binden. Von dessen 3'-Ende kann nun eine neue Zweitstrangsynthese beginnen kann. So entstehen Cluster mit tausenden Kopien eines DNA-Fragments. Im Bereich der verankerten Oligonukleotide befinden sich adapterspezifische Schnittstellen mit deren Hilfe alle Moleküle für die Sequenzierung einheitlich ausgerichtet werden. Die Amplifikation direkt in der „Flow Cell“ und ein robustes Indexsystem (im Adapter enthalten) minimieren die Gefahr von Kontaminationen und gehören zu den eindeutigen Vorteilen dieser Technologie. Letztlich erfolgt die Sequenzierung ausgehend von einem im Adapterbereich bindenden Primer nach dem Prinzip der Neusynthese von DNA (Sanger et al. 1977). Verwendet werden ebenfalls fluoreszenzmarkierte Nukleotide in einem Vier-Farben-System. Allerdings befinden sich die Markierungen an einer reversiblen 3'-Schutzgruppe. Diese verhindert den Einbau von mehreren Nukleotiden in einem Sequenzier-Zyklus. Nach der Detektion wird die Schutzgruppe mit dem Fluoreszenzfarbstoff chemisch abgespalten und der nächste Zyklus kann beginnen. So entstehen keine Probleme mit homopolymeren Sequenzabschnitten. Durch chemische Einflüsse und Laserbestrahlung bei der Detektion werden die Cluster in geringem Umfang ausgedünnt und eventuell beschädigt, wodurch die Signalintensitäten und die Qualität der Sequenzierung sinken. Zurzeit sind auf dieser Plattform einfache Leseweiten von 100 nt (HiSeq 2000™) bis 150 nt (GAIIx) möglich mit bis zu 600 Gb pro Lauf. Dies entspricht einem Analyse-Lauf mit zwei „Flow Cells“ (HiSeq 2000™), der ca. 11 Tage dauert. Tab. 1 fasst nochmals kurz die wichtigsten Eckdaten der weitverbreitetsten NGS-Technologien zusammen.

Tab. 1: Spezifikationen der NGS-Technologien

Quellen: Mardis 2008a, Mardis 2008b, Zhou et al. 2010, (Illumina), (454 Life Sciences)

Plattform	Methode	Amplifikation	Leseweiten [bp]	Fehlerraten [%]	Datenmenge [Gb/Tag]
Roche 454	Pyrosequenzierung	emPCR	> 800	0,5-1,5	0,8-1,2
Illumina GAIIx / HiSeq™	reversible Terminatoren	clusterPCR	50 -150	0,2-2	1,5-50
ABI SOLiD™	Ligation	emPCR	50 - 75	< 0,1	0,8-20,0

Solches Leistungspotential (Tab. 1) bei sinkenden Kosten ermöglicht neue Strategien für eine umfassende genomische, RNA oder regulatorische Analyse. Die Möglichkeiten reichen

von de novo Sequenzierung ganzer Genome (Edwards 2006; Imelfort und Edwards 2009; Jarvie und Harkins 2008), kompletter Transkriptome bis hin zur genomweiten Detektion aller SNPs (Nickerson et al. 2001) eines Genoms. Darüber hinaus sind NGS-Technologien auch eingesetzt worden zur Aufklärung von „copy number variations“ (Xie und Tammi 2009) oder der Sequenzierung von small RNAs (Girard et al. 2006) oder Amplikons, die aus vielen verschiedenen Genomen generiert wurden (Binladen et al. 2007; Dahl et al. 2007; McPherson 2009; Mardis 2008b).

Genexpressionsanalyse mittels RNA-seq

Bei der Genexpressionsanalyse wird untersucht, welche Gene in einem bestimmten Gewebe unter welchen Bedingungen wie stark transkribiert werden. Gegebenenfalls ist von Interesse, welche Spleißvarianten bevorzugt entstehen oder ob alternative Polyadenylierung stattfindet.

Die mit Hilfe der neuen Sequenziertechnologien generierten enormen Datenmengen (Stein 2010) bestehen im Falle der SOLiD™- und der Illumina-Technologie hauptsächlich aus kurzen Sequenz-Fragmenten und stellen ganz neue Anforderungen an die computertechnische Ausstattung und Analysealgorithmen. Für die Analyse von Transkriptom-Daten gibt es eine Reihe von Möglichkeiten: 1. Assemblierung und Vergleich mit Datenbanken. 2. Direkter Vergleich der Rohdaten mit den Datenbanken in Anlehnung traditioneller „tag count“-Ansätze wie der SAGE (Velculescu et al. 1995) oder der EST-Sequenzierung („read count“-Ansatz). Erweitert auf ganze sequenzierte Genome oder Teilbereiche, bietet sich das „Mapping“ an (Mortazavi et al. 2008). Für die Vorbereitung der Analyse wie auch die Bewertung der Ergebnisse müssen neben den Sequenzlängen und Besonderheiten bei der Sequenzierung weitere Fehlerquellen wie Amplifikationsartefakte aufgrund unterschiedlicher GC-Gehalte oder beispielsweise Artefakte als Folge der cDNA-Synthese berücksichtigt werden. Zum Beispiel wird die Verteilung der cDNA-Fragmente durch eine Oligo-d(T) gepripte cDNA-Synthese in Richtung 3'-Ende verschoben (Abb. 9 aus Wang et al.). Auch „random“-Hexamer-gepripte cDNA (Hansen et al. 2010; Li et al. 2010) oder die Basenzusammensetzung der Zielsequenz (Dohm et al. 2008) können Einfluss auf die Sequenzabdeckung haben. Auch durch die für die Sequenzierung notwendige Fragmentierung entstehen in Abhängigkeit von der Größe des Ursprungs-Moleküls mehr

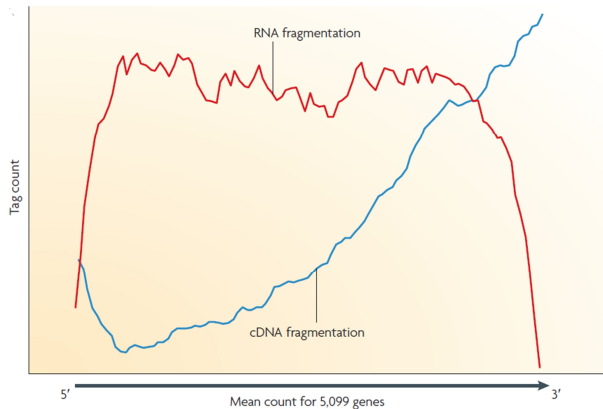


Abb. 9: Sequenzabdeckung nach RNA- bzw. DNA-Fragmentierung

Gezeigt ist die Sequenzabdeckung über 5099 Hefegene nach Sequenzierung und vorheriger cDNA-Fragmentierung (blau) bzw. RNA-Fragmentierung (rot); Quelle: Wang et al. 2009

oder weniger Fragmente, was bei quantitativen Analysen mittels „read counts“ zu berücksichtigen ist (Mortazavi et al. 2008).

Altbewährte Algorithmen wie Smith-Waterman (Smith et al. 1981; Waterman 1984), BLAT(Kent 2002) und BLAST (Altschul et al. 1990) sind nur noch begrenzt einsetzbar (Horner et al. 2010).

Fast alle heutigen Assemblierungs- oder „Mapping“-Programme beruhen auf k-mer Indices in Hash-Tabellen und nutzen De

Bruijn-Graphen (de Bruijn 1946) oder Eulerpfad (Pevzner et al. 2001; Horner et al. 2010; Alkan et al. 2011). Diese zeichnen sich hauptsächlich durch Schnelligkeit aus, benötigen aber meist viel Arbeitsspeicher und sind wenig flexibel. Vor allem bei bisher wenig erforschten Genomen werden hier schnell Grenzen erreicht.

Nach dem „Mapping“ der Sequenzen mit einem entsprechenden Referenzgenom erhält man im Idealfall wie in Abb. 10 dargestellt, den Sequenzen zugeordnete Koordinaten. Die Sequenzen können demnach Exons zugeordnet werden, Spleißen wird durch eine „geteilte“ Lage von Sequenzen gezeigt und die Detektion potentieller neuer Exons ist ebenfalls möglich.

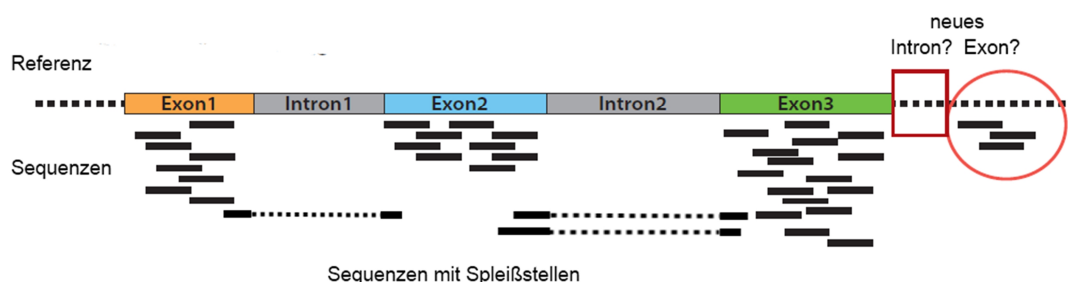


Abb. 10: Ergebnisse eines RNA-Seq-Experiments

Dargestellt ist die Lage der Sequenzen relativ zur Referenz; die Sequenzen liegen meist im Bereich der Exons 1-3; Sequenzen, die Spleißstellen enthalten, sind über das Intron hinweg mit Punkten verbunden
 verändert nach: (Morozova et al. 2009)

1.4 Zielsetzung

Die Aufklärung der molekulargenetischen Grundlagen bei der Knochen- und Knorpelentwicklung stellt eine wichtige Forschung für viele wissenschaftliche und wirtschaftliche Institutionen dar. Das Rind ist nicht nur Nahrungsgrundlage vieler Millionen Menschen und eines der am längsten domestizierten Tiere, sondern auch evolutionsbiologisch und wirtschaftlich ein besonders interessantes Forschungsobjekt. Die hohe genetische Ähnlichkeit zum Menschen (Elsik et al. 2009) und die allgegenwärtige Verfügbarkeit boviner Gewebe und Organe machen das Rind zu einem Modellorganismus für den Menschen und seine Erkrankungen. Die nahezu unbegrenzte Verfügbarkeit von Geweben und Organen eröffnet Möglichkeiten der Analyse von Material, das beim Menschen nur in minimalen Mengen und nur aus klinischen Quellen zur Verfügung stünde.

Mit Gewebe aus der Wachstumsfuge des Rindes steht ein besonders interessantes Material für Genexpressionsanalysen im Zusammenhang mit der Entwicklung von Knochen und Knorpel zur Verfügung. Mit der vorliegenden Arbeit soll eine umfassende Analyse aller an der Knorpel- und Knochenbildung beteiligten Gene durchgeführt werden. Die Transkriptomanalyse soll sowohl im Rahmen eines EST-Projekts als auch mittels NGS-Technologien erfolgen. Dabei soll die Expression bekannter knochen- und knorpelspezifisch exprimierter Gene analysiert und mit den Ergebnissen ähnlicher Untersuchungen an Mensch und Maus verglichen werden. Besonderer Fokus soll auf Gene gerichtet werden, die in der Wachstumsfuge stark exprimiert, aber bisher uncharakterisiert sind. Diese sollen in Hinblick auf eine mögliche Rolle bei der Knochen- und Knorpelentwicklung gezielt detaillierter analysiert und auf ihre mögliche Funktion hin untersucht werden. Dazu gehören auch die Untersuchung der genomischen Organisation, Interspeziesvergleiche zur Überprüfung evolutionärer Zusammenhänge, Identifizierung funktioneller Domänen der abgeleiteten Proteine und mögliche Proteinmodifikationen.

Die redundante Sequenzierung von cDNA aus der fötalen Wachstumsfuge mit Hilfe der 454/Roche- und der Solexa/Illumina-Plattform dient vor allem zur Überprüfung der EST-Ergebnisse und selbstverständlich zur extremen Erweiterung und Festigung der Datenbasis. Nicht zuletzt sollen Aussagen über die Vergleichbarkeit der verschiedenen Verfahren und ihrer Aussagekraft getroffen werden.

2 Material und Methode

2.1 Material

Standard-Lösungen

Ampicillin-Stammlösung:	10mg/l vollentsalztes Wasser
DEPC-Wasser:	0,1 % (v/v) Diethyl-Pyro-Carbonat Inaktiviert durch Autoklavieren
Dialysepuffer (10 x):	3 M Natriumchlorid, pH 7,4 0,25 mM Tris-HCl 0,1 M EDTA-Dinatriumsalz (Dihydrat)
E-Puffer (10 x):	0,36 M Tris 0,3 M Natriumdihydrogenphosphat (Dihydrat) 0,1 M EDTA-Dinatriumsalz (Dihydrat)
Ethidiumbromid-Färbelösung:	1 ml Ethidiumbromid-Stammlösung (DNA) ad 1 l 1x E-Puffer
Ethidiumbromid-Stammlösung:	0,5% (w/v) Ethidiumbromid in 1x Elektrophoresepuffer
Guanidinthiocyanatlösung:	4 M Guanidinthiocyanat 25 mM Natrium-Citrat, pH 7,0 0,5% Sarcosyl-N-Laurylsarcosin 0,1 M 2-Mercaptoethanol
Glycerin-Kultur-Stock:	50% 2 x LB-Medium 50% Glycerin
Ladepuffer 1 (DNA):	4 M Harnstoff 0,1 M EDTA-Dinatriumsalz (Dihydrat) 50% Saccharose 0,1 % (w/v) Bromphenolblau
Ladepuffer 2 (DNA):	1x TBE-Puffer 30% Saccharose
Ladepuffer (RNA):	500 µl Formamid (deionisiert) 100 µl 10x MOPS-Puffer 150 µl Formaldehyd (filtriert)

LB- Medium:	5 g Natriumchlorid 5 g Hefe-Extrakt 10 g Trypton/Pepton ad 1000 ml vollentsalztes Wasser pH 7,0
MOPS-Puffer (10 x):	0,2 M 3-(N-Morpholin)-Propan-Sulfonsäure 0,05 M Natriumacetat 0,01 M EDTA-Dinatriumsalz (Dihydrat)pH 7,0
Proteinase K:	1x SSC 0,1% (w/v) Proteinase K
RNA-Färbelösung:	0,25 % Bromphenolblau 0,25 % Xylencyanol FF 1 mM EDTA-Dinatriumsalz (Dihydrat) 50% (v/v) Glycerin pH 8,0
Standard-Salz-Citrat-Puffer (20x):	3 M Natriumchlorid (SSC 20x) 0,3 M Natriumcitrat, pH 7,0 TBST (10x): 100 mM Tris-HCl, pH 8,0 1,4 M Natriumchlorid 1% Tween 20
Tris-Borat-EDTA-Puffer (20 x): (TBE-Puffer)	432 g Tris 222,4 g Borsäure 18,6 g EDTA-Dinatriumsalz (Dihydrat) ad 2000 ml vollentsalztes Wasser
X-Gal-Platten :	13 g Agar-Agar 10 ml Ampicillin (10 mg/ml) 0,0476 g IPTG 0,1 g X-Gal in 1 ml N-N, Dimethylformamid ad 1000 ml 1x LB-Medium
Solution I (autoklaviert)	10 mM EDTA pH 8 (3,7224 g/l) (+ 7,5 mg RNase A / 100 ml Solution I)
Solution II:	0,2 M NaOH, 2% SDS (0,4 g NaOH in 45 ml VE- H ₂ O lösen + 5 ml 20% SDS)
Solution III:	für 200 ml: 50 ml 7,5 M Kaliumacetat, 23 ml Eisessig, 127 ml H ₂ O (7,5 M Kaliumacetat: 184 g pro 250 ml H ₂ O)

Kits und Enzyme

RNAeasy Mini Kit	Qiagen, Hilden
Oligotex® mRNA Midi-Kit	Qiagen, Hilden
NucleoTrap® mRNA Midi-Kit	Macherey-Nagel, Düren
SuperScript™ Plasmid System	Invitrogen, Karlsruhe
Plasmid-Cloning with Gateway™ Technology	Invitrogen, Karlsruhe
CloneJET™ PCR Cloning Kit	Fermentas, St. Leon-Rot
MessageAmp™ II aRNA Amplification Kit	Ambion/Applied Biosystems, Austin, Texas
SuperScriptII™ reverse Transcriptase	Invitrogen, Karlsruhe
SuperScriptIII™ reverse Transcriptase	Invitrogen, Karlsruhe
T4-DNA-Ligase	Fermentas, St. Leon-Rot
DNase I	Roche, Fermentas
<i>EcoRI</i>	Roche, Fermentas, NEB
<i>HindIII</i>	Roche, Fermentas, NEB
<i>NotI</i>	Roche, Fermentas, NEB
<i>Sall</i>	Roche, Fermentas, NEB

AusgangsmaterialRNA der fötalen Wachstumsfuge von *Bos taurus*

Die Gesamt-RNA der bovinen fötalen Wachstumsfuge (FGP) wurde von der Arbeitsgruppe Experimentelle Medizin I (Bindegewebforschung), in Kooperation mit Herrn Prof. Dr. A. Winterpacht, Universität Erlangen, zur Verfügung gestellt. Die RNA wurde nach der Methode von Weisser et al. 2002 isoliert.

Vergleichsgewebe aus adulten Rindern

Das Gewebematerial aus Herz, Leber, Lunge, Nieren und Thymus wurde aus drei Jahre alten Rindern entnommen und vom „Schlachthof Alzey „ (Robert-Bosch-Str. 23-25, 55232 Alzey) zur Verfügung gestellt. Die Proben wurden umgehend nach Entnahme auf Trockeneis verbracht und bis zur Verwendung bei -80°C gelagert.

2.2 Molekularbiologische Standardmethoden

Isolierung von DNA

Präparation von Plasmiden im 96er Format

Die Anzucht der Klone erfolgte in ca. 1150 µl LB-Medium in abgedeckten („Airpore Tapesheets“/Abgene) 96er Deepwell-Platten (Abgene) für 18-24 h bei 37°C. Die Isolationen erfolgten nach dem Prinzip der alkalischen Lyse (Birnboim und Doly 1979; Sambrook et al. 1989). Die Bakterien wurden zunächst abzentrifugiert, die Bakterienpellets in 335 µl Lösung I (+ RNase A) gelöst und zur Lyse mit 665 µl Lösung II versetzt. Nach 10 min Inkubation auf Eis erfolgte die Neutralisierung durch Zugabe von je 500 µl Lösung III. Renaturierte genomische DNA und Zelltrümmer wurden nach weiteren 10 min bei 3000 g abzentrifugiert.

Zur Fällung wurde der klare Überstand mit ca. 1 ml kaltem Isopropanol versetzt und 30 min bei 4°C und 3000 g zentrifugiert. Die Pellets wurden in 70 µl, bei sehr kleinen Pellets in 50 µl HPLC-Wasser, gelöst und nochmals über 96er Multiscreen™-HV-Filterplatten der Fa. Millipore mit etwa 30 mg Sephadex™ G-50 Fine der Fa. Amersham aufgereinigt.

Plasmidpräparationen mit Hilfe des E.Z.N.A Plasmid Mini Kits

Zur Plasmid-Präparation wurde das „E.Z.N.A. Plasmid-Miniprep-Kit II der Firma PeqLab verwendet. Die Präparationen wurden nach Herstellerangaben mit jeweils 1,5 ml einer 3 ml-Übernachtskultur (LB-Medium; 0,1 mg/ml Ampicillin) des Bakterienklons durchgeführt. Abweichend vom Hersteller-Protokoll wurde die DNA mit HPLC-Wasser von der Silica-Säule eluiert. Dabei wurde das Volumen je nach Folgeapplikation zwischen 20 und 50 µl variiert.

Isolierung genomischer DNA

Die genomische DNA für Vergleichsexperimente wurde nach dem Prinzip der Alkalischen Lyse nach einem Protokoll Blin und Stafford 1976 in Sambrook et al. 1989 durchgeführt.

Isolierung von RNA

Guanidinthiocyanat-Methode (GTC)

Im Wesentlichen beruht das hier angewendete Verfahren zur Isolierung von RNA auf der GTC-Methode nach Chomczynski und Sacchi 1987. Der Gewebeaufschluss erfolgte durch Mörser und Pistill unter flüssigem Stickstoff. Pro 100 mg eingesetztem Gewebe wurde das Material in 2 ml Guanidinthiocyanatlösung aufgenommen. Es folgten zwei bis drei Phenol-Chloroform/Isoamylalkohol-Extraktionen mit jeweils $\frac{1}{10}$ Volumen 2M NaAcetat (pH 4,1), 1 Volumen tris-gesättigtem (pH 7,2) Phenol und $\frac{1}{5}$ Volumen Chloroform/Isoamylalkohol (24:1). Nach Fällung mit 1 Volumen Isopropanol abs.(1 h bei -20°C) wurde das Pellet in $\frac{3}{10}$ Volumen Guanidinthiocyanatlösung resuspendiert, gefolgt von einer weiteren Fällung durch Zugabe von 1 Volumen Isopropanol. Anschließend wurde das Pellet mit 75% Ethanol gewaschen, in einer Vakuum-Zentrifuge (Eppendorf „Concentrator 5301“) bei 30°C getrocknet und in DEPC-Wasser gelöst.

RNA-Isolierung: „RNeasy Mini Kit“ (Qiagen)

Alternativ zur GTC-Methode wurde für die Isolierung von Gesamt-RNA aus verschiedenen Rindergeweben RNeasy Mini Kit“ der Fa. Qiagen verwendet. Die DNase-Behandlung zur Entfernung genomischer DNA wurde direkt auf der Silika-Membran durchgeführt. Die Elution der RNA von der Säule erfolgte zweimal mit je 50 μl RNase-freiem H_2O . Die Konzentration der RNA wurde mit dem „NanoDrop“ (siehe 2.8.1) gemessen und die Probe anschließend gefällt und bei -80°C gelagert.

Fällung von Nukleinsäuren

DNA-Fällung

Wahlweise wurden Fällungen von DNA durch Zugabe von $\frac{1}{10}$ Vol. 10x Dialysepuffer oder 3M Na-Acetate (Sambrook et al. 1989) und 2,5 Vol. Ethanol (abs.) durchgeführt. Nach Zentrifugation (mind. 15 min, 21000 g) und Waschen mit 70 % Ethanol wurde das Pellet in HPLC- H_2O gelöst.

Fällung von RNA: LiCl-Fällung

Zur Fällung der RNA wurde die Lösung mit 1/5 Vol. 4M LiCl, gut gemischt und mit 2,5 Vol. Ethanol versetzt. Die Proben wurden mindestens 2 h, meist über Nacht bei -20° C gefällt. Anschließend wurde 30 min bei 21.000 g und 4° C zentrifugiert. Der Überstand wurde abgenommen und verworfen und das Pellet mit 70% Ethanol (DEPC-H₂O) gewaschen. Dazu wurde nochmals 10 min bei 21000 g und RT zentrifugiert. Der Überstand wurde verworfen und das Pellet getrocknet. Das trockene RNA-Pellet wurde im gewünschten Volumen RNase-freiem H₂O (Macherey & Nagel oder DEPC-H₂O) gelöst.

Fällung von RNA: NaAc-Fällung

Um nachfolgende enzymatische Reaktionen nicht zu beeinflussen, wurde alternativ zu LiCl 1/10 Volumen 3M Na-Acetate pH 4,1 (Sambrook et al. 1989) verwendet. Nach Zugabe von 2,5 Vol. Ethanol (abs.) wurde zwischen 30 min und 16 h auf -20°C gefällt. Nach Zentrifugation (mind. 15 min, 21000 g) und Waschen mit 70 % Ethanol wurde das Pellet in DEPC (2fach autoklaviert) oder RNase-freiem H₂O (Macherey & Nagel) gelöst.

Agarose-Gelelektrophorese

DNA-Agarose-Gelelektrophorese

In Anlehnung an das Protokoll von Sambrook et al. wurden eigene Protokolle zur Auftrennung von DNA verwendet. Für DNA-Fragmente, einzelne Plasmidisolierungen und PCR-Produkte wurden vertikale Elektrophoresekammern der Firma GENTERprise verwendet. Als Gel- und Laufpuffer diente 1x TBE-Puffer. Die Agarosekonzentration wurde an die Fragmentgrößen angepasst und lag zwischen 1%-2%. Als Molekulargewichtsstandard wurde DNA des Phagen λ , restringiert mit *Hind III* sowie der „100bp-Ladder plus“ und 50bp-Marker der Fa. Fermentas (s. Anhang) verwendet. Die Elektrophoresen erfolgten bei Raumtemperatur und einer Stromstärke von 140 mA. Nach anschließendem Färben im Ethidiumbromidbad und Waschen in VE-Wasser erfolgte die Detektion auf dem Transilluminator bei einer Anregungswellenlänge von 312 nm.

RNA-Agarose-Gelelektrophorese

RNA wurde in 1,2%igen Agarosegelen, mit 5,55% Formaldehyd elektrophoretisch aufgetrennt. Sowohl für die Gele als auch als Laufpuffer wurde 1xMOPS-Puffer eingesetzt. Vor dem Auftragen wurden die Proben mit drei Volumen Ladepuffer und 0,5 µl Ethidiumbromid-Stammlösung versetzt und 10 min bei 65°C inkubiert. Anschließend wurden die Proben nach Zugabe von ½ Volumen RNA-Färbelösung auf das Gel aufgetragen. Die Elektrophoresen wurden in vertikalen Elektrophoresekammern der Firma GENterprise bei einer Stromstärke von 150 mA durchgeführt. Die Detektion und Dokumentation erfolgte mit Hilfe der „Kodak 1D 3.0“-Software.

Konzentration und Qualität von Nukleinsäuren

Neben der klassischen Gelelektrophorese wurden außerdem NanoDrop und Bioanalyzer eingesetzt.

NanoDrop

Die photometrische Konzentrationsbestimmung erfolgte mit Hilfe des NanoDrop ND-1000 UV/VIS-Spektralphotometers (PeqLab, Erlangen). Dabei wurden sowohl 1µl als auch 2µl DNA- oder RNA-Lösung verwendet.

Bioanalyzer

In Anlehnung an die Qualitätskontrolle bei Microarray-Experimenten (Harrington et al. 2009) wurde der Bioanalyzer 2100 der Firma Agilent hauptsächlich für die Qualitätskontrolle von RNA-Proben verwendet. Dabei wurde für Gesamt-RNA oder größere RNA-Mengen der RNA 6000 Nano-, für kleinere Mengen der Pico-Assay verwendet. Die RNA-Qualität wurde über die „RNA Integrity Number“ (RIN, Schroeder et al. 2006) festgehalten.

Für die Messung sehr geringer Mengen an DNA wurde der „High Sensitivity Chip“ eingesetzt.

Ligationsreaktionen

Für die Klonierung von PCR-Produkten aus genomischen oder cDNA-Amplifikationen wurden drei Vektorsysteme verwendet: pGEM®-T easy- (Invitrogen), pJet1.2- CloneJET™ (Fermentas) und institutseigener pUC19/“blunt“ (Vektorkarten im Anhang). Dabei wurden die PCR-Produkte sowohl direkt als auch nach Gelextraktion eingesetzt.

pGEM®-T easy - Vektorsystem

2,5 µl PCR-Produkt in 5 µl Ligationsansätzen eingesetzt:

PCR-Produkt	1-2,5 µl
pGEM®-T easy (5 ng/µl)	1 µl
10 x T4-DNA-Ligase-Puffer	0,5 µl
T4-DNA-Ligase (3 U/µl)	1 µl
Nuklease freies H ₂ O	x µl
<hr/> Gesamtvolumen	<hr/> 5 µl

Die Ansätze wurden stets ü.N. bei RT inkubiert. Vor der Transformation wurden die Ligationsansätze zur Entfernung störender Salze gefällt und in 5-10 µl HPLC-H₂O gelöst.

pJet1.2 –Vektorsystem - CloneJET™(Fermentas)

Die Klonierungen wurden nach Herstellerangaben durchgeführt.

PCR-Produkt	1-7 µl
2X Reaction Buffer	10 µl
DNA Blunting Enzyme	1 µl
Nuklease freies H ₂ O	x µl
<hr/> Gesamtvolumen	<hr/> 18 µl

Nachdem Mischen und 5 minütiger Inkubation bei 70 °C wurden die Ansätze auf Eis herunter gekühlt und dann mit 1 µl T4 DNA Ligase und 1 µl pJET1.2/blunt Cloning Vector (50 ng/µl) komplettiert. Nach mindestens 5 min Inkubation bei RT wurden die Ansätze entweder direkt in *E.coli DH10B* transformiert oder bei -20°C gelagert.

Transformation mittels Elektroporation

Es wurde jeweils 1 µl des größenfraktionierten Ligationsgemischs boviner cDNA aus fötaler Wachstumsfuge und dem Vektor pSPORT1 in *E.coli DH10B* elektrokompetente Zellen transformiert (PeQLab oder Invitrogen) und anschließend für 45 min in 1 ml LB Medium bei 37°C inkubiert. Für weitere Klonierungen wurden alternativ die Vektoren pGEM®-T easy (Invitrogen), pUC18 oder pUC19 verwendet.

Restriktion mit Restriktionsendonukleasen

Der Verdau von DNA wurde typischerweise mit 1-5 Units Restriktionsendonuklease pro Mikrogramm DNA durchgeführt. Dabei wurden sowohl der vom Hersteller empfohlene Puffer als auch die vorgeschriebene Inkubationstemperatur verwendet.

Speziell für Testrestriktionen von Plasmiden mit klonierten PCR, RT-PCR oder cDNA-Integraten wurden in der Regel 5 µl Isolat eingesetzt. Für die Restriktion von pGEM®-T easy-Konstrukten wurde ein einfacher Verdau mit *EcoRI*, für pUC19 und pSPORT1 (EST-Bibliothek) wurden Doppelrestriktionen mit *EcoRI* und *HindIII* durchgeführt.

Standard-Restriktionsansatz

Plasmid-DNA	5 µl
10 x Restriktionspuffer	2 µl
<i>EcoRI</i>	1 µl
<i>HindIII</i>	1 µl
HPLC-H ₂ O	X µl
Gesamtvolumen	20 µl

Die Ansätze wurden mindestens für 60 min, in der Regel 3h, bei 37° C inkubiert

polyA⁺-Anreicherung „NucleoTrap mRNA Mini Kit“

Für die Anreicherung der polyA⁺-RNA wurde das auf Oligo-d(T)-Latexbeads basierende Kit „NucleoTrap mRNA Mini Kit“ der Fa. Macherey-Nagel verwendet. Die Anreicherung wurde sowohl im Säulen- wie auch im „Batch“-Verfahren gemäß Herstellerangaben durchgeführt.

Zur Elution wurde 65°C warmes DEPC- (2fach autoklaviert) oder RNase-freies H₂O (Quiagen oder Macharey&Nagel) verwendet.

cDNA-Synthese

Die cDNA-Synthese erfolgte mit der „SUPERSCRIPT™ III“- Reversen Transkriptase der Fa. Invitrogen. In der Regel wurden zwischen 1-2 µg Gesamt-RNA eingesetzt. Die Erststrangsynthese (Verma et al. 1972; Kacian et al. 1972) erfolgte gemäß Herstellerangaben. Für die Verwendung bei der RT-PCR wurden außer Oligo-(dT)₁₈-Primern auch „Random Hexamere“ verwendet.

Tab. 2: Standardansatz für die cDNA-Erststrangsynthese

Gesamt-RNA	1 – 1,2 µg
dNTPs (10 pmol/µl)	1 µl
Oligo(dT) ₁₈ -Primer (0,5 µg/µl)	1 µl
„Random Hexamer“- Primer (0,1 µg/µl)	0,5 µl
RNase-freies H ₂ O	X µl
	13 µl
5 – 10 min bei 65° C und 1 min auf Eis bis RT erreicht wurde	
5 x „First-Strand“ - Puffer	4 µl
DTT (0,1 mM)	1 µl
RNase Out (optional)	1 µl
Superscript™ III (200 U/µl)	1 µl
Gesamtvolumen	20 µl

Die Zweitstrangsynthese wurde lediglich für die Herstellung der Sequenzierbibliotheken verwendet.

Polymerasekettenreaktion

Ansatz

Als Matrize für PCRs wurde einzelsträngige cDNA, 1:10 verdünnt eingesetzt, in schwierigen Fällen unverdünnt, eingesetzt. Es wurde sowohl die Polymerase der Firma Sigma als auch die „GoTaq“-Polymerase (Promega, Madison) verwendet.

Tab. 3: PCR-Ansatz

cDNA	1 µl
dNTPs (10 mM each)	2 µl
MgCl ₂ (25 mM)	2 µl
5 x GoTaq-Puffer	10 µl
Primer 1 (10 pM)	1 µl
Primer 2 (10 pM)	1 µl
GoTaq-Polymerase (5 U/µl)	0,3-0,5 µl
HPLC-H ₂ O	x µl
Gesamtvolumen	50 µl

Programme

Die PCR-Reaktionen wurden überwiegend in einem GeneAmp® 9700 PCR-System (Applied Biosystems) durchgeführt. Alternativ stand ein Peltier Thermal-Cycler (BioRad, München) zur Verfügung.

Standardprogramm:

95°C	5:00 min	
50°C	0:30 min	
72°C	1:20 min	40x
95°C	0:30 min	
72°C	15:00 min	
8°C	∞	

Dieses Programm wurde vor allem bei den PCR-Nachweisen auf cDNA angepasst. Dabei lag die Annealing-Temperatur für RT-PCR-Reaktionen bei 48°C (Standard48) und wurde nur bei unspezifischen Amplifikaten erhöht. Die Amplifikation besonders schwieriger Matrizen konnte unter anderem in Verbindung mit einer verringerten Elongationstemperatur (68°C) erreicht werden. Die Zyklenzahl wurde ebenfalls im Extremfall auf bis zu 45 Zyklen erhöht.

Alternativ wurden auch sogenannte „Touchdown“-PCRs nach (Don et al. 1991) durchgeführt. Dabei wurde die Annealing-Temperatur in den ersten 10 Zyklen schrittweise um 1°C pro Zyklus erniedrigt (55°C-46°C) und 30 Zyklen auf 45°C belassen.

Verwendete Primer

Die verwendeten Primer wurden über die Firma biomers.net GmbH (Söflinger Straße 100 D-89077 Ulm) bezogen, in HPLC-H₂O gelöst und im Stock in einer Konzentration von 100 pmol/μl bei -20°C gelagert.

Tab. 4: Verwendete Primer

Primer Bez.			Sequenz (5'→3')	bp	T _m (Biomers)
Bt_1	GAPDH	for	aacaccctcaagattgtcag	20	46
Bt_2	GAPDH	rev	catgtggaagtcaggagatt	20	46
Bt_3	Col2A1	for	gtaagtgggcaagactatg	20	46
Bt_4	Col2A1	rev	cagaggtgtttgacacagaa	20	45
Bt_5	MGP	for	ggcagctctgtgttatgaat	20	47
Bt_6	MGP	rev	atgctactgcaagggaatac	20	46
Bt_7	511905	for	catgagagatagaacagcac	20	41
Bt_8	511905	rev	gactaaaggaacgtcgata	19	42
Bt_9	618319	for	gactccacaaacagaataag	20	41
Bt_10	618319	rev	aggatggcgtcctgtctg	18	50
Bt_11	ID1	for	agtcggtgatcattgtaata	20	42
Bt_12	ID1	rev	agaggaataagtgtctctctg	20	43
Bt_13	E12P44	for	gtcggtaggatctctttac	19	46
Bt_14	E12P44	rev	ggaggacttcaggacat	18	45
Bt_15	D04P33	for	ggatcaaatgtacggataat	20	43
Bt_16	D04P33	rev	cgtagcttggatcctcta	18	42
Bt_17	F07P42	for	ctgattcaggaagtcaaga	19	41
Bt_18	F07P42	rev	ctagacagatgctgttgaag	20	42
Bt_19	B10P30	for	attatttatttctgcctcct	20	42
Bt_20	B10P30	rev	acaatacagcctctaacaaa	20	42
Bt_21	E12P44_2	for	gatctctttacggcctctggact	23	55
Bt_22	E12P44_2	rev	gatagactggacctgctgtgacc	23	54
Bt_23	D04P33_2	for	aatgtacggataataactaaagattc	25	46
Bt_24	D04P33_2	rev	agcacttatctagaagactgtaaat	25	45
Bt_25	F07P42_2	for	ggtagcgagtctgattcaggaagt	24	55
Bt_26	F07P42_2	rev	gatgcctagacagatgctgttgaa	24	55
Bt_27	B10P30_2	for	ctttcatacatcttattgttcctga	25	49
Bt_28	B10P30_2	rev	gacagtacattatggaaggacttg	24	49
Bt_29	D04P33_3	for	gattcaaatacttaggcaga	20	41
Bt_30	D04P33_3	rev	cagacttgaaacatatcgac	20	42
Bt_31	F07P42_3	for	ggagtgaaacctcaaaaat	19	43
Bt_32	F07P42_3	rev	aacttagccctcttccag	18	43
Bt_33	B10P30_3	for	acattaatttctgaagggtca	20	41
Bt_34	B10P30_3	rev	cctacagagtaaaggacag	20	41
Bt_35	618319_2	for	ggatggaagtgggtctg	17	44
Bt_36	618319_2	rev	gtcgcgctgagaaaagt	17	45
Bt_37	511905_2	for	cggagaatgagagagaag	18	41
Bt_38	511905_2	rev	accatatcatatccctctgt	20	42

Primer Bez.			Sequenz (5'->3')	bp	T _m (Biomers)
Bt_39	618319_3	for	gaagacccccgtggagagc	18	52
Bt_40	618319_3	rev	ggcccctttattcgttcag	20	55
Bt_41	LECT 1	for	gcattctatacagcagattc	20	41
Bt_42	LECT 1	rev	GAGTATCAATGGGAAGTTG	19	41
Bt_43	CTGF	for	CTGAGCGAGTTGTGTACC	18	43
Bt_44	CTGF	rev	ACTCCACTCTGTGGTCTG	18	41
Bt_45	TPT1	for	AGTAGGACAGAGGGTAACAT	20	42
Bt_46	TPT1	rev	CAGTTATGACGACAGGTG	18	40
Bt_47	LAMR1	for	CTTCACTAACCAGATCCAG	19	41
Bt_48	LAMR1	rev	CAAGTGTCTGGAAGAAGAG	19	41
Bt_49	ATXN10	for	AACTATCTTCCAGAGAGTCC	20	41
Bt_50	ATXN10	rev	GAGGTTTTCTCCAGCTC	18	43

2.3 Programme, Skripte und WEB-Tools

Verwendete Programme

Tab. 5: Programmliste

Programm/Skript	Quelle
Blastall	NCBI ftp://ftp.ncbi.nih.gov
Blastclient	NCBI ftp://ftp.ncbi.nih.gov
ESTAnnotator	HUSAR Bioinformatics Lab, DKFZ http://genome.dkfz-heidelberg.de Hotz-Wagenblatt et al. 2003
<u>Visualisierung</u>	
Artemis	The Wellcome Trust Sanger Institute http://www.sanger.ac.uk Rutherford et al. 2000
GenePalette	http://www.genepalette.org/ (Rebeiz und Posakony 2004)
<u>Gene ontology / Metagenomics</u>	
Blast2GO	Centro de Investigación Príncipe Felipe Valencia http://www.blast2go.org Conesa et al. 2005
GOEast	Institute of Genetics and Developmental Biology http://omicslab.genetics.ac.cn Zheng und Wang 2008
MEGAN	Universität Tübingen http://ab.inf.uni-tuebingen.de Huson et al. 2007

Programm/Skript	Quelle
GenMAPP-CS	http://www.genmapp.org (Dahlquist et al. 2002) (Salomonis et al. 2007)
<u>Assembly und Mapping</u>	
PASS	CRIBI, University of Padova http://pass.cribi.unipd.it Campagna et al. 2009
Bowtie	Johns Hopkins University, USA http://bowtie-bio.sourceforge.net Langmead et al. 2009
NGEN2	DNASTAR, Inc., USA http://www.dnastar.com
Newbler	Roche Diagnostics Corporation http://www.my454.com/
MIRA	(Chevreux et al. 2004) http://chevreux.org/projects_mira.html
<u>Programmpakete</u>	
Lasergene® v7.01, v8.13 v8.15,	DNASTAR, Inc., USA http://www.dnastar.com
CLC genomic workbench	CLC bio A/S, Dänemark http://www.clcbio.com
Genomatix Genome Analyzer (GGA)	Genomatix Software GmbH http://www.genomatix.de

Verwendete Perl-Skripte

qseq2fastq.pl (modifiziert nach Benjamin Rieger / IMSB)

Umformung von qseq-Dateien in fastq-Dateien. Dabei wird auf den Illumina-internen „Chastity“-Filter und Sequenzen mit nicht mehr als 3 ungenauen Basen („N“) gefiltert.

fasta_extract.pl (Benjamin Rieger / IMSB)

Extrahiert Sequenzen anhand einer in einer Text-Datei angegebenen Liste mit Sequenznamen die entsprechenden Sequenzen aus einer FASTA-Datei. Dabei sind nur genaue Übereinstimmungen zulässig.

count_fasta.pl (Joseph Fass / Brad Sickle, University of California)

Das Skript extrahiert eine Längenverteilung der Sequenzen in einer FASTA-Datei in 100er-Schritten. Darüber hinaus errechnet es N25, N50, und N75-Werte, die eine bessere Beurteilung der Effizienz der Assemblierung erlauben.

fastq_extract.pl (Benjamin Rieger / IMSB)

In diesem Fall wird nach dem gleichen Prinzip extrahiert, wie das fasta_extract-Skript Sequenzen aus seiner FASTQ-Datei. Zusätzlich sind auch genaue Übereinstimmungen mit Teilen der Sequenznamen möglich.

Blaster.pl (Benjamin Rieger / IMSB)

Bei Blaster.pl handelt es sich um ein Skript zum automatischen splitten von FASTA-Dateien in bis zu 1000 Dateien, die jeweils als separater Job an ein „High Performance Cluster“ (HPC) zur lokalen Blastanalyse mittels „blastall“ (www.ncbi.nlm.nih.gov) übergeben werden können. Dieses Skript wurde in verschiedenen Versionen verwendet und kontinuierlich weiter entwickelt. Über Blaster können außerdem die gesplitteten Ergebnisse wieder vereinigt werden und abgebrochene Suchen wieder aufgenommen werden.

Bowtier1_0.pl (Benjamin Rieger / IMSB)

Analog zum Blaster handelt es sich um ein Skript zum Splitten der Sequenzen in mehrere Dateien und einer automatisierten, parallelisierten Bearbeitung auf dem Rechner-Cluster.

batchblast_extractor.pl (Pirooznia et al. 2008)

Das auf Perl basierende Skript extrahiert die für eine Annotation relevanten Daten aus einer Blastx-Ausgabedatei und bringt diese in Tabellenformat. Das Skript fand lediglich bei den Auswertungen der 454-Daten Verwendung, bis eigene Skripte etabliert waren.

blast_reduce.pl (Benjamin Rieger / IMSB)

Bei diesem Skript handelt es sich um eine Vorstufe zum „Bort“-Skript. Es reduziert Textausgabe-Dateien auf lokalen Blast-Suchen uauf spezifische Einträge wie Trefferbeschreibung und E-value.

Bortv1.3.-v2.3. (Benjamin Rieger / IMSB)

Bei Bort („Blast Output Refinement Tool“) handelt es sich um eine Weiterentwicklung der ebenfalls verwendeten Skripte `blast_parser.pl` und `blast_parser_interactive.pl`. Ziel der Entwicklung war die Extraktion der wichtigen Daten einer Blast-Ausgabedatei und die Quantifizierung der Treffer. Die Skripte wurden für alle „Read-Count“-Ansätze auf Blast-Basis verwendet.

`pass_extract_multiple_genomic_regions_spliced_0_9.pl`

Eigenes Skript zur Extraktion von Mapping-Daten aus PASS-Mappings. Dabei werden sowohl ein „default“-Mapping über die gesamte „Read“-Länge benötigt als auch ein „spliced“-Mapping. Dabei wird die Abdeckung jeder einzelnen Basenposition in multiplen Referenzsequenzen ausgegeben. Da Spleißstellen-überspannende „Reads“ seltener vorkommen, werden in der Standardeinstellung diese „Reads“ mit einer 10-fachen Verstärkung hochgerechnet.

Verwendete WEB-Tools

Tab. 6: Verwendete - Web - Applikationen

Genome Browser		
UCSC	Genome Browser	(Karolchik et al. 2003)
Genome Browser		(Karolchik et al. 2011)
ENSEMBLE	Genome Browser	EMBL - EBI
Genome Browser		Wellcome Trust Sanger Institute
VISTA	Genome Browser	
Genome Browser		
Multiple Alignments Alignments und Promotorvorhersage		
zPicture	Multiples Sequenz-Alignment und Promotorvorhersage (rVISTA2)	(Ovcharenko et al. 2004)
mVISTA	Multiples Sequenz-Alignment	(Brudno et al. 2003a)
rVISTA v1 und v2	Multiples Sequenz-Alignment und Promotorvorhersage	(Loots et al. 2002)

European Bioinformatics Institute		
ClustalW2	Multiples Sequenz-Alignment	(Larkin et al. 2007)
InterProScan	Domänensuche	(Quevillon et al. 2005) (Zdobnov und Apweiler 2001)
Center for Biological Sequence Analysis, Technical University of Denmark		
NetAcet 1.0 Server	Acetylierung	(Kiemer et al. 2005; Liu und Lin 2004)
NetPhos 2.0 Server	Phosphorylierung	(Blom et al. 1999)
NetPhosK 1.0 Server	Phosphorylierung	(Blom et al. 2004)
NetNGlyc 1.0 Server	N-Glykosylierung	R. Gupta, E. Jung and S. Brunak, 2004
NetOGlyc 3.1 Server	O-Glykosylierung	(Julenius et al. 2005)
Pathway- und GO-Analysen		
WikiPathways	Pathway-Visualisierung	(Pico et al. 2008)
KEGG-Pathways	Pathway-Analyse	(Hua et al. 2008)
DAVID (NIH)	GO-Analyse	(Huang DW et al. 2009)

2.4 Bioinformatische Standardmethoden

Standard-Sequenzanalyse-Prozeduren

Für alle grundlegenden Sequenzanalysen wie Restriktionskartierungen, Planungen von Klonierungen, Dot-Plot-Analysen oder auch die Standard-Assemblierung kleiner Mengen an Sequenzen wurden die Programmpakete DNASTAR/Lasergene v7.0-8.15 oder die CLC Bio-Genomics Workbench verwendet.

„Basic Local Alignment Search Tool“ – BLAST

Für den Vergleich von Nukleotid- und Proteinsequenzen mit verschiedenen Datenbanken wurde das „Basic Local Alignment Search Tool“ verwendet (BLAST®, <http://blast.ncbi.nlm.nih.gov/Blast.cgi>, National Library of Medicine, Altschul et al. 1990)

Automatisierte Blast-Suchen

Für automatisierte Blast-Analysen wurden im Bedarfsfall die in den Programmen Blast2GO, Lasergene (DNASTAR und CLC workbench oder ESTAnnotator ((Hotz-Wagenblatt et al. 2003) enthaltenen Blast-Optionen genutzt. Hochdurchsatz-Blast-Suchen wurden überwiegend lokal

mit Hilfe des lokalen „blast“-Programms (National Library of Medicine) durchgeführt. Dabei wurden jeweils die aktuellen Datenbanken unter <ftp://ftp.ncbi.nih.gov/blast/db/> verwendet.

Darstellung genomischer Loci

Für die Darstellung genomischer Strukturen, wie beispielsweise die genomische Organisation von Transkripten oder die Lage von Promotoren, wurde überwiegend das Programm GenePalette (Rebeiz und Posakony 2004) verwendet. Neben reinen FASTA-Dateien wurden auch annotierte Sequenzen im gbk-Format des NCBI „MapViewers“ oder verschiedener Genom Browser verwendet.

Normalisierung auf RPKM (Mortazavi et al. 2008)

Die Normalisierungsmethode wurde versuchsweise zur Kompensation von etwaigen systematischen Fehlern aufgrund ungleichmäßiger „Read“-Verteilungen sowohl bei der 454/FLX-Sequenzierung als auch Illumina-Sequenzierung durchgeführt.

$$RPKM = \frac{\textit{Gesamt} - (\textit{Exon}) - \textit{reads}}{\textit{mapped reads (millions)} * \textit{exon length (kb)}}$$

2.5 cDNA-Bibliotheken und EST-Sequenzierung

Im Rahmen der Diplomarbeit von Frau Tatjana Kölzer (IMSB, Mainz) wurde die RNA polyA⁺ angereichert („Oligotex mRNA Midi Kit“; Qiagen GmbH) und eine cDNA Bibliothek

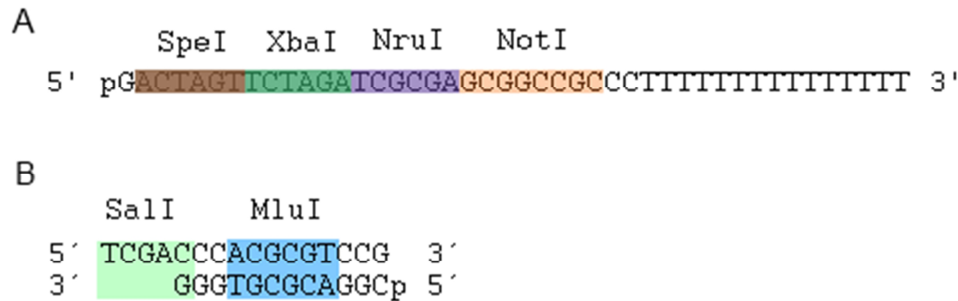


Abb. 11: cDNA-Synthese

A Oligo-dT-Primer mit NotI-Schnittstelle für die Erststrangsynthese; B SalI-Adapter für 5' gerichtete Klonierung

hergestellt. Dafür wurde das „Superscript™ II“-System (Invitrogen) verwendet. Um eine gerichtete Klonierung in den Vektor pSPORT1 (Invitrogen) zu gewährleisten, wurde die Erststrangsynthese (Verma et al. 1972; Kacian et al. 1972) mit Hilfe eines modifizierten Oligo-d(T)-Primer durchgeführt. Nach der Zweitstrangsynthese wurde ein SalI-Adapter ligiert und die Fragmente nach *NotI*-Restriktion gerichtet in den Vektor pSPORT1 kloniert. In der Regel liegt daher das 5' gelegene Ende des Fragments an der *SalI*-Schnittstelle. Vor der Ligation wurden die cDNA-Fragmente zunächst über „Sephacryl S-500 Spin Columns“ (Stratagene, Amsterdam) aufgereinigt und in 4 Größenfraktionen geteilt. Mitunter wurden Reste der *NotI*-Restriktion und *SalI*-Adaptoren entfernt, um deren Ligation in den Vektor zu verhindern. Fragmente kleiner als 500 bp wurden nicht berücksichtigt. Die Fraktionen wurden vom hohen zum niederen Molekulargewicht mit 1-4 gekennzeichnet.

Nach Transformation in *E.coli* DH10B wurden die Ansätze komplett auf LB-Agar-Platten ausplattiert und über Nacht bei 37°C inkubiert. Einzelkolonien wurden in Flüssigkultur (LB-Medium/Amp) im 96er-Format vermehrt, die Plasmide isoliert und die Ausbeute über ein Agarosegel überprüft. In Einzelfällen wurden Testrestriktionen zur Kontrolle der Integratgrößen durchgeführt. Anschließend wurden jeweils ca. 500 ng der Plasmide mit 1 µl T7-Primer (10 pmol/µl) versetzt und durch die Firma GENterprise (Mainz) mittels Kapillarsequenzierer sequenziert.

2.6 Auswertung der EST-Klone

Die Sequenzen wurden im Vorfeld der Analysen zunächst qualitativ beurteilt (Chromas Lite) und die Vektorsequenzen entfernt (ESTAnnotator, VectorScan NCBI). Sequenzen schlechter Qualität reiner Vektor oder sehr kurze Integrate wurden als solche gekennzeichnet und aussortiert.

Für die Annotation wurden Blast-Suchen gegen verschiedene Datenbanken durchgeführt. Dabei wurden sowohl Einzelsequenzen als auch Konsensus-Sequenzen geclusterter Daten verwendet.

Manuelle Annotation mittels Blast-Suchen

Für die manuelle Auswertung der EST-Klone sowie die Einzelanalysen von Nukleotid- oder Proteinsequenzen wurde das „Basic Local Alignment Search Tool“ (BLAST®, <http://blast.ncbi.nlm.nih.gov/Blast.cgi>, National Library of Medicine; Altschul et al. 1990) verwendet. Dabei wurden im Wesentlichen folgende Optionen mit ihren Grundeinstellungen verwendet:

Tab. 7: Blast-Optionen und Datenbanken

Option	Datenbanken
blastn	„Nucleotide collection“ (nt/nr) „Expressed sequence tags“ (est) „Reference RNA sequences“ (refseq_rna)
blastx	„Non-redundant protein sequences“ (nr)
tblastx	„Nucleotide collection“ (nt/nr)

Im Einzelfall wurden verschiedene Parameter angepasst. Dabei wurde bei nicht genomischen Blast-Suchen mit Sequenzen aus Sanger-Sequenzierungen ein maximaler E-value von $1e-30$ angenommen. Diese Grenze wurde bei Bedarf angepasst. Die Daten zum Blast-Treffer, die „Accession“-Nummer und der E-Value wurden tabellarisch erfasst und teilweise durch genomische Blast-Suchen ergänzt.

Automatisierte Blast-Suchen und Sequenzannotation

Ergänzend zur manuellen Identifikation der EST-Sequenzen konnte für einen kurzen Zeitraum das ESTAnnotator-Modul (Hotz-Wagenblatt et al. 2003) des HUSAR-Servers (DKFZ) genutzt werden. Für Suchen über den DKFZ-Server wurden die Unigene-Datenbanken der Spezies *Bos taurus*, *Homo sapiens*, *Mus musculus* sowie die Proteindatenbank Swissprot und die allgemeine EST-Datenbank verwendet. Die Informationen wurden ebenfalls in die tabellarische Sequenzannotation übernommen.

Tabellarische Erfassung und quantitative Auswertung der EST-Sequenzen

Die Ergebnisse wurden in einer Tabelle zusammengefasst. Wenn möglich wurden beim Erfassen der Blast-Ergebnisse bereits die offiziellen Gen-Namen extrahiert. Neben Informationen über den besten Blast-Treffer und zugehörige E-values wurden weitere relevante Daten gespeichert. Wenn möglich wurden bekannten Genen über einen Zahlen/Farb-Code in Großgruppen eingeteilt:

Tab. 8: Kategorien der EST-Sequenzen

Kategorie	Farbe	Bezeichnung
2	gelb	zu kurz / Vektor
4	rosa	schlechte Sequenz
5	rot	nicht genau zuzuordnen / überarbeiten
8	weiss	bekanntes Gen
10	hellgrün	Gen mit Verbindung zur Kn/Kn-Entwicklung
11	dunkelgrün	Gen mit potentieller Verbindung zur Kn/Kn-Entwicklung
12	weiss	bekanntes Gen / ribosomales Protein

Mehrfach vorkommende Gene wurden manuell aufsummiert (elektronischer Anhang). Klone mit keinen oder zu kleinen Integraten sowie Sequenzen mit Übereinstimmungen zu ribosomaler RNA wurden von der Analyse ausgenommen.

Expressionsprofile knochen- knorpelspezifischer Gene

Die Genauswahl und Gruppierung der speziellen Expressionsprofile wurde auf Grundlage verschiedener die Knochen- und Knorpelentwicklung betreffende Veröffentlichungen getroffen (Zhang et al. 2003; Jung et al. 2004; Tagariello et al. 2005).

Weiterführende Analyse möglicher interessanter Transkripte

Funktionelle Charakterisierung unbekannter EST-Sequenzen

Alle uncharakterisierten 226 Sequenzen wurden nochmals geclustert (SeqMan 8.15, Lasergene/DNASTAR) und mit den Konsensus-Sequenzen gesonderte blastn- und blastx-Suchen sowie automatisierte Analysen über BLAST2GO (Conesa et al. 2005; Götz et al. 2008) auf funktionelle Domänen oder Signalpeptide durchgeführt (InterProScan, European Bioinformatics Institute). Dabei wurden folgenden Datenbanken durchsucht:

BlastProDom	HMMTigr	SignalPHMM
FPrintScan	ProfileScan	TMHMM
HMMPPIR	HAMAP	HMMPanther
HMMPfam	PatternScan	Gene3D
HMMSmart	SuperFamily	

Charakterisierung von EST-Clustern und Einzelklonen

Die genomische Organisation bestimmter Gene, Contigs oder auch einzelner EST-Klone wurde zunächst durch Blast-Suchen mit Standard-Einstellungen über das „MapView“-Programm des NCBI (Wheeler et al. 2001) gegen entsprechende Genome durchgeführt. Genomische Abschnitte und Annotationen wurden mit Hilfe des „Ensemble Genome Browser“ (Hubbard et al. 2002) oder direkt über den „MapView“ extrahiert. Um Exon-/Intronstrukturen aufzuklären, wurden in Einzelfällen manuell Alignments der Sequenzen mit der genomischen Referenz durchgeführt und zusätzlich auf GT/GA-Introngrenzen untersucht.

Genomische Analysen und Interspezies-Vergleich

Homologie und Interspeziesvergleiche wurden mit den Alignment-Funktionen, mVISTA (Brudno et al. 2003a; Brudno et al. 2003b) und rVISTA (Loots et al. 2002) des „VISTA Genome Browser“ (Frazer et al. 2004) oder zPicture (Ovcharenko et al. 2004) durchgeführt. Für detaillierte Analyse und die Zuordnung von Spleiß-Stellen wurde das „SeqMan“-Modul der DNASTAR Lasergene[®]™-Software verwendet

2.7 Nachweis spezifischer Transkripte mittels RT-PCR

Der Nachweis spezifischer Transkripte in RNA aus fötaler Wachstumsfuge des Rindes aber auch Referenzgeweben aus adulten Tieren wurde sowohl qualitativ als auch semi-quantitativ durchgeführt.

RT-PCR mit cDNAs der Wachstumsfuge und Vergleichsgewebe

Für die qualitativen Nachweise wurde zunächst einzelsträngige cDNA synthetisiert, die als Template für eine PCR mit genspezifischen Primern diente (Tab. 4). Dabei wurde darauf geachtet, Exon-überspannende Amplifikate zu erzeugen.

Die PCR-Nachweise wurden wie unter 2.2 beschrieben durchgeführt. Dabei wurde für jeden Nachweis immer eine Reaktion mit GAPDH- oder MGP-Primern als Positivkontrolle mit durchgeführt. Jeweils 1/10 des Reaktionsvolumens wurde dann über Agarose-Gelelektrophorese aufgetrennt, mit Ethidiumbromid gefärbt, und dokumentiert.

Semi-quantitative RT-PCR mit cDNAs der Wachstumsfuge und Vergleichsgeweben

Für den semi-quantitativen Nachweis wurde ebenfalls einzelsträngige cDNA und auch die gleichen Primer wie in der qualitativen RT-PCR verwendet.

Im Idealfall wird ein externes standardisiertes Molekül (in die Probe und die zugehörigen Primer) mit in die Amplifikation gegeben. Außerdem wird eine Sättigung der PCR-Reaktion vermieden. Dadurch kann jede einzelne Amplifikation in Relation zum Standard gesetzt

werden und erlaubt eine ungefähre quantitative Aussage im Vergleich zur Amplifikation verschiedener Zielsequenzen (Serazin-Leroy et al. 1998).

Im konkreten Fall wurde als interner Standard ein endogenes Transkript, GAPDH, gewählt. Die Intensitäten der Amplifikat-Banden und GAPDH wurden mit der „Kodak 1D 3.0“-Software erfasst und die Werte in Relation gesetzt.

2.8 Roche-454/FLX-Sequenzierung

Für das Erstellen der cDNA-Bibliothek zur Sequenzierung mittels 454/FLX-Technologie wurden ca. 5 µg polyA⁺-angereicherte RNA aus der fötalen Wachstumsfuge des Rindes dem Institut für Immungenetik in Kaiserslautern überlassen. Da der 1. Versuch einer Sequenzierung fehlschlug, wurde die verbleibende cDNA über WGA („whole genome amplification“) amplifiziert und anschließend sequenziert.

Die 38492 Sequenzen wurden zunächst nach durchschnittlicher Leseweite und Qualität charakterisiert. Dazu wurden die Sequenzen in SeqMan (v8.15; DNASTAR – Lasergene) geladen, die Sequenzlängen exportiert und in Microsoft Excel ausgewertet.

Clustern der Roche-454/FLX-Daten

Mit Hilfe der Programme Newbler (Roche), MIRA2 (Chevreux et al. 2004), SeqMan™ v7.1 (DNASTAR, Lasergene®) und NGEN™ (DNASTAR, Lasergene®) erfolgte die Assemblierung der Daten. Dabei wurden im Allgemeinen die für 454-Sequenzen vorgeschlagenen Parameter verwendet.

Für MIRA2 wurden sowohl die Optionen „EST accurate“ als auch „EST normal“ mit Standardeinstellungen verwendet. Es wurden nur die unter der „EST normal“ Option entstandenen Contigs weiter verwendet.

Die Assemblierung in SeqMan v7.1 erfolgte in Datenpaketen zu 5000 Sequenzen mit einer minimalen Sequenzlänge von 50 bp und einer minimalen Übereinstimmung von 85 %. Auch für die Assemblierung in NGEN wurden 85 % statt 80 % Übereinstimmung verwendet.

Zum Vergleich der verschiedenen Programme wurden die Gesamtzahl der assemblierten Sequenzen, die maximale Größe der Contigs und die Contig-Größenverteilung ermittelt. Für Letzteres wie auch die Ausgabe von N50- und N75-Werten wurde das Skript „count_fasta“ verwendet. Die Ausgaben wurden verglichen und bewertet. Die weitere Analyse erfolgte mit den NGEN-Sequenzclustern.

Blast-Analysen der Roche-454/FLX-Daten

Geclusterte Daten

In einem ersten Schritt wurden die geclusterten Daten (8376 Konsensus-Sequenzen) über den Blastn-Algorithmus mit der „Sequence Reference collection –RNA“ (refseq_rna) und mit Hilfe des Blastx-Algorithmus mit der Proteindatenbank (nr) verglichen. Dazu wurden automatisierte Netz-Blastsuchen („blastclient“; www.ncbi.nlm.nih.gov) und lokale Blast-Suchen über „blastall“ (www.ncbi.nlm.nih.gov) durchgeführt. Für Blastx-Suchen wurde als „Threshold“ ein E-Wert von $1 e^{-05}$ und bei Blastn-Suchen von $1 e^{-10}$ gewählt.

Die Blastx-Ergebnisse wurden mit Hilfe des „Batch Blast Extractors“ (Pirooznia et al. 2008) aus den Text-Ausgabe-Dateien der mit „blastall“ durchgeführten lokalen Blast-Suchen gegen die „Non-redundant protein sequences“ (nr)- Datenbank extrahiert und in tabellarische Form gebracht. Die Blast-Ergebnisse jedes einzelnen Contigs wurden mit der Information über die „Read“-Anzahl im jeweiligen Contig kombiniert. Die offiziellen Gen-Namen wurden wenn möglich aus dem Blast-Treffer extrahiert. Im Fall hoch exprimierter Gene wurde bei fehlendem Gen-Name manuell nachgebessert.

Die Blastn-Ergebnisse wurden mit Hilfe des „blast_parsers“ (IMSB/Mainz) extrahiert. Dieser verarbeitet die „Read“-Information direkt und gibt zum jeweiligen Contig und Blast-Treffer auch die quantitative Information. Auch hier wurden die offiziellen Gen-Namen extrahiert.

Einzelsequenzen

Parallel zur Annotation der geclusterten Sequenzen wurden auch die nicht assemblierten Daten untersucht. Die Ergebnisse wurden mit Hilfe der Bort1.3-Plattform gefiltert und über

die Beschreibungen der Blast-Treffer zusammengefasst und quantifiziert. Auch hier galten die Grenzwerte für signifikante Treffer von $1 e^{-05}$ (Blastx) und $1 e^{-10}$ (Blastn).

Statistische Analyse der Blast-Ergebnisse

Da eine durch das kommerzielle Sequenzier-Unternehmen verschuldete offensichtliche Kontamination der Sequenzen mit Sequenzen aus einem Wein-Projekt vorlag, wurden die Blast-Ergebnisse statistisch nach ihren Spezies-Treffern ausgewertet. Dazu wurde das Programm MEGAN (Huson et al. 2007) verwendet, das normalerweise für Metagenetische Untersuchungen verwendet wird.

Blast2GO-Analysen (Conesa et al. 2005)

Die Java-Plattform Blast2GO ist nicht für die Auswertung von Einzelsequenzen aus NGS-Projekten geeignet. Daher wurde im Wesentlichen mit geclusterten Daten bzw. deren Konsensussequenzen gearbeitet.

Als Grundlage für die Analyse dienen dem Programm die zu annotierenden Sequenzen im FASTA-Format. Damit sind automatisierte Blast-Suchen möglich. Darüber hinaus können auch Blastx-Ergebnisse der entsprechenden FASTA-Datei eingeladen werden. Dies bietet den Vorteil die mitunter sehr langwierigen Blast-Suchen lokal und unter Ausnutzung besserer Hardware durchführen zu können.

Entfernen von Kontaminationen

Die *Vitis*-Kontaminationen konnten nur anhand ihrer Übereinstimmungen nach den Blast-Suchen identifiziert werden. Sequenzen mit Übereinstimmungen zu pflanzlichen Taxa wurden aus der Analyse entfernt.

Automatisierte GO-Annotation

Da die manuelle funktionelle Charakterisierung mitunter sehr langwierig ist und sich meist nur für kleine, spezielle Sequenzgruppen (beispielsweise die höchst exprimierten Gene)

praktikabel ist, wurde die GO-Annotation mit Hilfe dieses Programms durchgeführt. Für die Blast-Suchen wurde dabei ein „Threshold“ von 1×10^{-6} gewählt. D

Detektion möglicher neuer funktioneller Transkripte

Da mit dem Programm auch automatisiert die InterProScan-Datenbanken durchsucht werden können wurde das Programm auch zur oberflächlichen Suche nach Hinweisen auf Funktionalität bei möglichen neuen Transkripten verwendet.

2.9 Transkriptomanalyse mittels Illumina-GAllx-Sequenzierung

Vorbereitungen und cDNA-Bibliothek

Um ausreichend Material für die Sequenzierung zu erhalten, wurden zunächst aus ca. 2 µg Gesamt-RNA aus der fötalen Wachstumsfuge polyA⁺-angereichert. 100 ng der ca. 300 ng PolyA⁺-RNA wurde eine Lineare Amplifikation mit Hilfe des „MessageAmpTM II aRNA Amplification Kit“ (Ambion/Applied Biosystems, Austin, Texas) durchgeführt. Die Amplifikation wurde gemäß den Herstellerangaben durchgeführt. Die Elution der aRNA erfolgte in 200 µl RNase-freiem H₂O.

Für die Herstellung der cDNA-Bibliothek wurden ca. 5 µg der amplifizierten RNA an das „Cologne Center for Genomics“ (CCG,) geschickt. Dort

Analyse Vorbereitungen

Blast-Analysen und PASS-„Mapping“

Für die Blast-Analysen wurden die Rohdaten im qseq-Format mittels qseq2fastq.pl in fastq-Dateien umgeschrieben, dabei kam der „Chastity“-Filter zum Tragen. Außerdem wurde keine Sequenz mit mehr als drei unbekanntenen Basen („N“) weiter verwendet. Im Weiteren wurde gegen die repetitiven Elemente mittels PASS oder Bowtie gemappt und so entsprechende Sequenzen entfernt.

Tab. 9: Filter für Illumina-Sequenzen

Filter	Kriterium/Datenbank	Skript
Qualitätsfilter	(„chastity“ / max. 3 „N“-Positionen)	qseq2fastq.pl
Repetitive	(Annotation Bt. Build 4.1) aus gbk-Dateien extrahiert	Bowtier1_0.pl
rRNA	(euk. rRNA gem. „SILVIA“-Datenbank)	pass_auto_rRNA.

Dabei wurden die in folgender Tabelle verzeichneten Sequenzen aus der Analyse entfernt:

Tab. 10: Gefilterte Illumina-Sequenzen

Probe	Rohdaten	Filter		verbleibend
		Qualität	Repetitive	
FGP s_7_1	37.742.382	5.202.291	6.314.561	26.225.530
FGP s_7_2	37.742.235	5.282.968	5.201.115	27.258.152

Hochdurchsatz-BLAST-Analysen

Für die Blast-Analysen der Illumina-Sequenzen aus RNA der fötalen Wachstumsfuge des Rindes wurden die Daten zunächst wie oben beschrieben gefiltert (Qualität / Repetitive). Sowohl „Read 1“ (s_7_1) als auch „Read 2“ (/s_7_2“) wurden dann über das Perl-Skript „Blaster1.0-1.2.pl“ in 1000 Dateien und über das LSF-System des HPC-Clusters der Universität-Mainz übergeben. Hoch parallelisiert wurden dort Lokale Blastn-Suchen gegen die „Reference RNA sequences“ (refseq_rna)-Datenbank durchgeführt. mit Hilfe des „blastall“-Programms (NCBI) durchgeführt. Die Resultate jeder einzelnen Teildatei wurden anschließend über „Blaster“ wieder zusammengeführt und mit Bort1.3-2.3 ausgewertet. Dabei werden gleiche Blast-Treffer zusammengefasst und aufsummiert. Sequenzen die ein und dasselbe Gen repräsentieren, allerdings unterschiedliche Blast-Treffer zeigen, werden nicht zusammengefasst und verbleiben redundant in der Quantifizierung.

Extrahieren der offiziellen Gen-Namen und Zusammenfassen von Einträgen

Mit Hilfe von MS Excel wurden die Gen-Namen aus den Blast-Treffern extrahiert. War dies nicht möglich wurden Listen mit „Accession“-Nummern in Convertern wie beispielsweise in DAVID (NIH) enthalten versucht umzuwandeln. Einträge mit gleichen Gen-bezeichnungen wurden zusammengefasst.

Mapping begrenzter genomischer Bereiche mit PASS (Campagna et al. 2009)

Für die PASS-„Mappings“ wurden ebenso wie für die Blast-Analysen die gefilterten Daten verwendet. Als Referenzen wurden FASTA-Dateien der Referenzregionen verwendet. Parallel zur FASTA-Datei wurden auch exakt passende gbk-Dateien mit der aktuellen Annotation generiert. Für das mapping wurde sowohl die „default“- als auch die „spliced“-Option verwendet. Dabei war für eine positive Zuordnung eine minimale Übereinstimmung von 96% notwendig. Durch Lücken getrennte „Reads“ mussten zusätzlich insgesamt wieder eine Länge von 90 nt ergeben.

Für die grafische Darstellung wurden dann die mit den PASS-Informationen (pass_extract_multiple_genomic_regions_spliced_0_9.pl) generierten Diagramme mit den „Read“-Abdeckungen mit den mittels Genepalette erstellten Annotationskarten der betreffenden Region kombiniert.

Genomweite „Mappings“ mit CLC

Für die genomweiten „Mappings“ wurden die Rohdaten in die CLC-Workbench geladen und automatisch gefiltert und getrimmt. Dazu wurde maximal ein „N“ und ein minimaler Phred-Score von 24 erlaubt. Die Daten wurden anschließend mit Hilfe der RNA-Seq-Option gegen die annotierten chromosomalen Sequenzen von *Bos taurus* gemappt.

3 Ergebnisse

Als Basis zur Aufklärung spezifischer Genexpression bei der Knochen- und Knorpelentwicklung wurde RNA aus der Wachstumsfuge sowohl über herkömmliche EST-Sequenzierung als auch mit Hilfe der neuen FLX- (Roche) und Illumina-Technologie GAllx analysiert. Da mit zunehmendem Alter die Aktivität der Wachstumsfuge nachlässt, wurde fötales Rindergewebe verwendet. Hierbei ist der Vorteil eines hochaktiven Gewebes kombiniert mit höherer Ausbeute und besserer Verfügbarkeit gegenüber beispielsweise murinem oder menschlichem Gewebe. Die Schwerpunkte der Arbeit liegen dabei in der Auswertung der drei Datensätze, dem Einstieg in neue Sequenziertechnologien und der damit verbundenen Datenauswertung sowie dem Vergleich der Verfahren.

Zunächst wurden von der insgesamt ca. 10000 Klone umfassenden cDNA-Bank der bovinen fötalen Wachstumsfuge 4032 Klone ansequenziert. Diese wurden bioinformatisch ausgewertet. Dabei konnten Übereinstimmungen zu nahezu 1000 bereits bekannten Genen gefunden werden. Besonderes Augenmerk wurde bei der Sequenzanalyse auf hoch exprimierte oder unbekannte Sequenzen bzw. solche mit bereits bekanntem Bezug zu Knochen- und Knorpelentwicklung gelegt. Beim Vergleich der Expressionsprofile konnten vor allem bei den Kollagenen, Proteoglykane und anderer ECM-Proteine große Übereinstimmungen, aber auch kleinere Abweichungen, zu bereits existierenden Datensätzen gefunden werden. Ausgesuchte Transkripte wurden sowohl qualitativ als auch semi-quantitativ im Vergleich zu anderen Gewebetypen, mittels RT-PCR nachgewiesen.

Neben dem klassischen EST-Ansatz wurden zwei NGS-Analysen durchgeführt.

1. Die Analyse mittels Roche/454 FLX, die zu einem relativ kleinen Datensatz von 38492 Sequenzen (FLX/Roche) geführt hat. Diese FLX-Daten wurden ähnlich wie die EST-Daten ausgewertet und klassifiziert. Die in 8376 Contigs zusammengefassten Sequenzen repräsentieren ungefähr 3000 bekannte Gene und liefern diesbezüglich schon rund dreimal mehr Informationen im Vergleich zum EST-Projekt. Trotz weitgehender Übereinstimmungen in den Expressionsprofilen knochen- und knorpelspezifischer Gene, konnten sowohl systematische Abweichungen als große Unterschiede bei der Expressionsstärke einzelner Transkripte festgestellt werden (3.5). Hierfür konnten mögliche Gründe identifiziert und

mögliche Methoden zur Angleichung beider Datensätze entwickelt werden. Außerdem konnten Vor- und Nachteile der beiden Analyseansätze aufgezeigt werden.

2. Die Sequenzierung des Transkriptoms mittels Illumina Plattform GAIIx. Dabei konnte ein Datensatz mit ca. 75 Millionen Sequenzen generiert werden. Dieser wurde hinsichtlich knochen- und knorpelspezifischer Genexpression untersucht. Dabei wurden verschiedene Verfahren sowohl zum Filtern als auch zur Analyse der Daten getestet, angepasst und optimiert. Die Ergebnisse von BLAST-Suchen im Hochdurchsatz sowie Cluster- und Mapping-Ansätze lieferten neue Hinweise auf eventuell bei der Knochenentwicklung beteiligte Gene. Darüber hinaus bestand die Möglichkeit eines Vergleichs, der Daten aus der fötalen Wachstumsfuge mit Transkriptom-Daten aus mesenchymalen Stammzellen des Rindes, vor der Differenzierung zu Chondrozyten.

3.1 Genexpressionsanalysen mittels EST-Sequenzierung

Aufbereitung der cDNA-Bank und Sequenzierung

Die cDNA der fötalen Wachstumsfuge lag zunächst in drei Größenfraktionen vor. Diese wurden jeweils in einer Ligationsreaktion in den Vektor pSPORT1 kloniert. Über spätere Restriktionsanalysen konnten die ungefähren Integratgrößen wie folgt abgeschätzt werden.

Fraktion 1	900 bp – 3500 bp
Fraktion 2	700 bp – 2000 bp
Fraktion 3	300 bp – 800 bp

Mit diesen drei Ligationsansätzen wurden insgesamt 15 Transformationen mit *E.coli DH10B* durchgeführt. Aus 50 Plasmidpräparationen konnten 4965 integrathaltige Plasmide gewonnen werden. Davon wurden 4032 sequenziert (Tab. 11). Dies entsprach ungefähr der Hälfte des verfügbaren Materials.

Tab. 11: Übersicht über Transformationen und Anzahl der sequenzierten Klone

In der linken Spalte aufgeführt sind die Bezeichnungen der Transformationen sowie die entsprechende Größenfraktion der cDNA; dann von links nach rechts, die Anzahl der resultierenden Klone, die Kennung der 96-Well-Platten und die Anzahl letztendlich sequenzierter Klone

Transformation	Anzahl der Klone	96er-Platten Nr.	sequenzierte Klone
T1F1 (Fraktion 1)	99	14	96
T2F1 (Fraktion 1)	111	13	96
T3F1 (Fraktion 1)	98	15	96
T4F1 (Fraktion 1)	12	--	--
T5F1 (Fraktion 1)	67	49	67
T6F1 (Fraktion 1)	141	49,50	29
T1F2 (Fraktion 2)	877	1,2,3,4,5,6,7,8,9	864
T2F2 (Fraktion 2)	193	10,11	96
T3F2 (Fraktion 2)	--	--	--
T4F2 (Fraktion 2)	--	--	--
T5F2 (Fraktion 2)	--	--	--
T6F2 (Fraktion 2)	107	12	--
T7F2 (Fraktion 2)	301	19,20,21	288
T8F2 (Fraktion 2)	997	22,23,24,25,26, 27,28,29,30,31	768
T9F2 (Fraktion 2)	1561	32,33,34,35,36,37,38,39, 40,41,42,43,44,45,46,47	1344
T10F2 (Fraktion 2)	99	48	96
T1F3 (Fraktion 3)	302	16,17,18	288

Vor der Sequenzierung wurden Ausbeute und Qualität der Plasmidpräparationen durch Agarose-Gelelektrophorese überprüft. Im Folgenden ist exemplarisch eine Stichprobe aus einer erfolgreichen Isolation (Abb. 12).

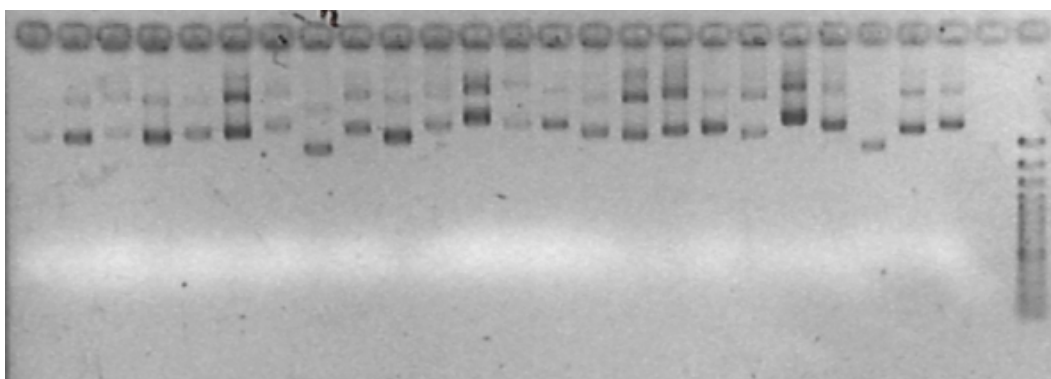


Abb. 12: Kontrolle der Plasmidpräparationen

Aufgetragen sind jeweils 5 µl aus 50 µl; als Molekulargewichtsstandard (M) wurde der 100 bp Plus Marker (Fermentas) verwendet

Wie auch aus Tab. 11 hervorgeht, wurden demnach bevorzugt Klone der Fraktionen 1 und 2 verwendet. Der Anteil an leeren Klonen lag bei allen drei Ligationen um 10 %.

Charakterisierung der cDNA-Klone

Zur Analyse der 4032 cDNA-Klone wurden diese zunächst ausgehend von der T7-Promotorsequenz des Vektors (5'-Ende des Transkripts) sequenziert. Aufgrund schlechter Sequenzqualität wurden bereits davon im Vorfeld 404 Sequenzen (10 %) von allen weiteren Betrachtungen ausgeschlossen. Auf alle anderen Sequenzen wurden sowohl manuelle Blast-Suchen (blastn, blastx, tblastx, blastn genomisch gegen Rind) als auch automatisierte Verfahren mit Hilfe des „Blastclient“ (Netblastservice, NCBI) und des ESTAnnotators (Hotz-Wagenblatt et al. 2003) angewendet. Dabei wurde eine relativ große Anzahl von Sequenzen mit Übereinstimmungen zu rRNA und eine geringere Anzahl von Klonen ohne oder mit sehr kurzen Integraten, meist poly(A), identifiziert (Abb. 13 und Abb. 14). Diese wurden ebenfalls von weiteren Analysen weitestgehend ausgeschlossen. In der folgenden Übersicht sind die

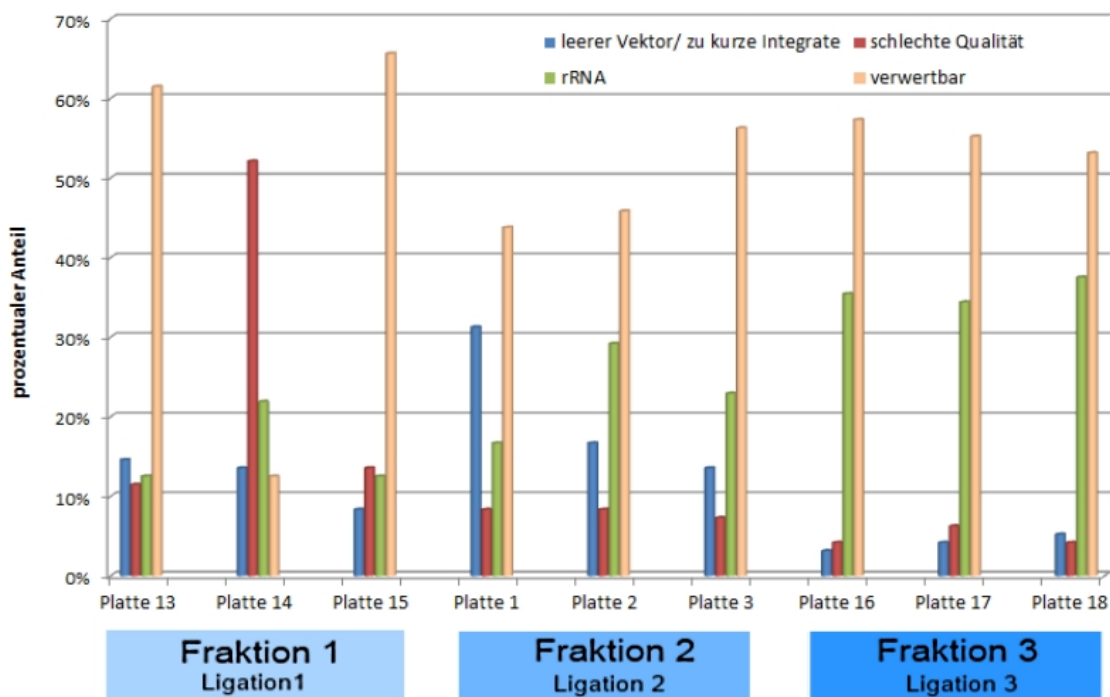


Abb. 13: Verteilung unbrauchbarer Sequenzen über Platten und Fraktionen
 Dargestellt sind die prozentualen Anteile von Sequenzen mit schlechter Qualität (rot), leeren Vektoren/zu kurze Integraten (blau) und rRNA-Sequenzen; von je drei Platten aus jeder Fraktion dargestellt

Anteile dieser als nicht brauchbar klassifizierten Sequenzen bei jeweils drei Platten aus jeder Fraktion aufgeschlüsselt. Obwohl nur eine Auswahl von Platten dargestellt ist, bestätigen die Sequenzierungen im Wesentlichen die Ergebnisse der Testrestriktionen. Die Anteile zu kleiner Integrate (meist poly(A)) oder leerer Vektoren liegen bei Fraktion 1 und 2 im Schnitt bei 10%. Ausnahme stellt lediglich Platte 1 mit 31,3% dar. Aus Klonen der 3. Fraktion konnten hingegen mehr verwertbare Sequenzen gewonnen werden (90 %). Mit durchschnittlich über 30 % (35,8 %) sind allerdings auch wesentlich mehr rRNA-Sequenzen enthalten.

Von 4032 ansequenzierten Klonen waren 2242 Klone (55,6 %) informativ und können in 2 Kategorien eingeteilt werden. 1. 419 Sequenzen (10,57 %) ohne eindeutige

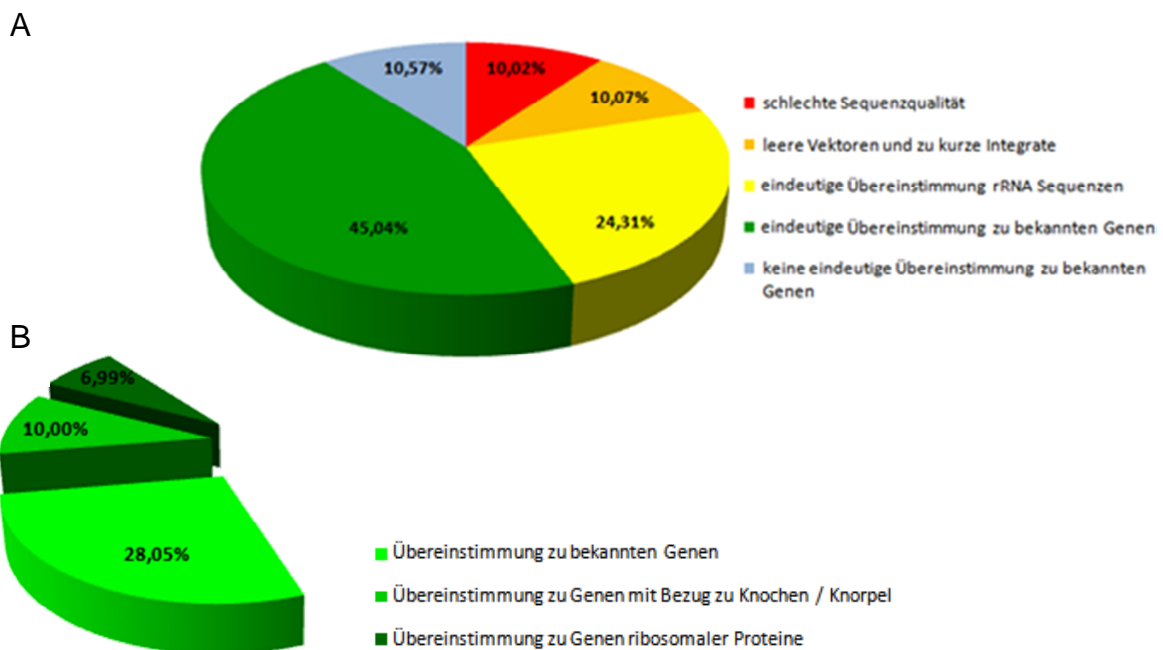


Abb. 14: Gesamtübersicht über die Klonverteilung in der cDNA-Bank (nach T7-Sequenzierung von 4032 Klonen); A zeigt die Einteilung der Klone in fünf Hauptkategorien, 1 (■) nicht auswertbare Sequenzen aufgrund schlechter Sequenzqualität, 2 (■) Klone mit leeren Vektoren oder zu kurzen Integraten, 3 (■) Klone mit rRNA-Integraten, 4 (■) eindeutig bekannten Genen zugeordnete Sequenzen und 5 (■) solche mit keiner Übereinstimmung oder erheblichen Abweichungen zu bekannten Gensequenzen; B Unterteilung der eindeutig identifizierten Sequenzen; ribosomale Proteine (■), Bezug zu Knochen- und Knorpelentwicklung (■) und Übrige (■)

Übereinstimmungen in den relevanten Datenbanken, mithin vermutlich "neue Gene". 2. Sequenzen mit eindeutigen Übereinstimmungen zu bekannten Genen (1816 Sequenzen, 45,04%), darunter 289 (7 %) Sequenzen, die für ribosomale Proteine kodieren und 403 Sequenzen (10 %) von Genen, die bekanntermaßen eine Rolle bei der Knochen- und

Knorpelentwicklung (Abb. 14B) spielen. Die zugrunde liegenden Ergebnisse der Datenbanksuche sind aus der Tabelle „EST-Projekt-BLAST“ des elektronischen Anhangs zu entnehmen.

Insgesamt waren 44,4% der T7-Sequenzierungen nicht weiter für eine Untersuchung der Genexpression verwendbar. 404 Sequenzen (10,02%) wurden aufgrund fehlgeschlagener Sequenzierung, 406 (10,07%) Klone wegen fehlender bzw. zu kleiner Integrate nicht weiter betrachtet. Darüber hinaus zeigten 980 Klone Übereinstimmungen zu rRNA und wurden ebenfalls aus den Analysen genommen (Abb. 14).

Stark transkribierte Gene in der fötalen Wachstumsfuge

EST-Projekte bzw. die Sequenzierung von cDNA-Bibliotheken sind wichtige Verfahren für Expressionsanalysen. Generell geben sie einen guten Überblick über die Genexpressionsmuster in einem Gewebe (Adams et al. 1991). Dabei müssen die nachzuweisenden Sequenzen nicht unbedingt bekannt sein. Ein weiterer Vorteil ist, dass die cDNA-Klone für weitere Analysen physisch zur Verfügung stehen. Da jeder Klon im günstigsten Fall ein RNA-Molekül repräsentiert, lässt sich durch Vergleich der Anzahl von EST-Sequenzen eine Aussage über die relative Stärke der Expression machen. Zu diesem Zweck wurden verschiedene Profile aus dem vorliegenden Datensatz erstellt. Besonders interessant sind die überrepräsentierten und demnach potentiell sehr stark exprimierten Sequenzen. Darüber hinaus sind Gruppen von Genen für Proteine interessant, die spezifisch an der Knochen- und Knorpelbildung beteiligt sind. Dies sind vor allem die Gene für Kollagene und andere Proteine der extrazellulären Matrix (ECM), wie beispielsweise die Proteoglykane.

Die nachfolgende Tabelle (Tab. 12) enthält die 25 höchst exprimierten Gene mit der Anzahl der identifizierten Sequenzen, den zugehörigen besten Treffern zu einem Datenbank-Eintrag und den dazugehörigen Datenbankverweis („accession number“). Die vollständige Genliste mit jeweiliger Sequenzanzahl ist im elektronischen Anhang in der Tabelle „EST-Projekt-“ abgelegt.

Tab. 12: Hoch exprimierte Gene in der fötalen Wachstumsfuge des Rindes
 Aufgelistet sind die 25 meist vertretenen Sequenzen aus dem EST-Projekt; neben Gensymbol bzw. -locus ist jeweils das entsprechende Protein und der Datenbankeintrag mit der besten Übereinstimmung aufgeführt; hervorgehoben sind Genprodukte mit eindeutigem Bezug zur Knochen- oder Knorpelentwicklung

Gen/Name	Protein	bester Datenbanktreffer	Anzahl der EST
<i>COL2A1</i>	Kollagen Type II alpha 1	<u>X02420</u>	72
<i>SPP1</i>	Secreted phosphoprotein 1	<u>NM_174187</u>	52
<i>COL10A1</i>	Kollagen type X alpha 1	<u>X53556</u>	48
<i>COL9A1</i>	Kollagen type IX alpha 1	<u>XM_582924</u>	28
<i>FTH1</i>	ferritin heavy polypeptide 1	<u>AY911329</u>	26
<i>MGP</i>	Matrix Gla protein	<u>AY911361</u>	25
<i>COL11A2</i>	Kollagen type XI alpha 2	<u>XM_593458</u>	24
<i>LECT1</i>	chondromodulin I	<u>NM_174269</u>	23
<i>LOC618319</i>	ASCL830	<u>XM_870650</u>	23
<i>COMP</i>	Cartilage oligomeric matrix protein	<u>XM_591255</u>	20
<i>LOC514954</i>	hyp. Protein	<u>XM_611557</u>	20
<i>HAPLN1</i>	Cartilage linking protein 1	<u>NM_174288</u>	15
<i>COL11A1</i>	Kollagen type XI alpha 1	<u>XM_587946</u>	14
<i>EEF1A1</i>	elongationfactor 1 alpha 1	<u>AB060107</u>	14
<i>RPL23</i>	ribosomalprotein L23	<u>BC102556</u>	14
<i>CTGF</i>	Connective tissue growth factor	<u>BC113279</u>	12
<i>YBX1</i>	nuclease sensitive element binding protein 1	<u>NM_174815</u>	12
<i>RPS2</i>	ribosomal protein S2	<u>BC102227</u>	12
<i>FN1</i>	fibronectin 1	<u>XM_873679</u>	11
<i>RPL27</i>	Ribosomal protein L27	<u>BC102313</u>	11
<i>RPLP1</i>	60S acidic ribosomal protein P1	<u>BC102695</u>	11
<i>RPS15a</i>	Ribosomal protein S15a	<u>BC108231</u>	11
<i>ID1</i>	Inhibitor of DNA binding 1	<u>XM_585780</u>	11
<i>CCDC80</i>	CCDC80 protein	<u>XM_593217</u>	10
<i>CD109</i>	150 kDa TGF-beta-1-binding protein	<u>XM_613667</u>	10

Auf den ersten Blick wird deutlich, dass ein auffällig hoher Anteil an knochen- und knorpelspezifischen Genen unter den häufigsten EST zu finden ist. Die meisten dieser Gene kodieren für Strukturproteine.

Einen wesentlichen Anteil bilden die Kollagene, COL10A1, COL9A1, COL11A2 und, besonders hervorzuheben, COL2A1, das mit 72 EST-Sequenzen die häufigste Sequenz darstellt. Außerdem konnten Sequenzen für weitere wichtige Bestandteile der extrazellulären Matrix als „hoch exprimiert“ klassifiziert werden. Hierzu gehören sowohl das „secreted phosphoprotein 1“ (SPP1) als auch das „matrix Gla protein“ (MGP), das „cartilage oligomeric matrix protein“ (COMP) und das „cartilage linking protein“ (HAPLN1). Lediglich CTGF („connective tissue growth factor“) und Chondromodulin 1 (LECT1) fallen als Wachstums- und Differenzierungsfaktoren aus der Gruppe der Strukturkomponenten heraus. Gene aus der Gruppe der Proteoglykane konnten unter den häufigsten ESTs nicht gefunden werden. Wie in vielen EST-Projekten konnten jedoch etliche Übereinstimmungen zu Genen für Proteine gefunden werden, die an der Proteinbiosynthese beteiligt sind. Hierbei ist neben einer Vielzahl der Gene für ribosomale Proteine vor allem das Gen für den Elongationsfaktor EEF1A1 durch eine große Zahl an Sequenzen repräsentiert.

Besonders hervorzuheben und interessant sind die als LOC618319 und LOC514954 gekennzeichneten Sequenzcluster. Es handelt sich um zwei bisher nicht näher charakterisierte Genprodukte, die eine auffällig starke Expression in der Wachstumsfuge zeigen. Sie wurden auf eine mögliche knochen- und knorpelbezogene Funktion überprüft.

Die 100 höchst repräsentierten Sequenzen und ihre funktionelle Einordnung

Für die funktionelle Charakterisierung der 100 häufigsten Gene wurden Vergleiche mit sogenannten „Gene Ontology“-Datenbanken durchgeführt. Wie in Abbildung Abb. 15

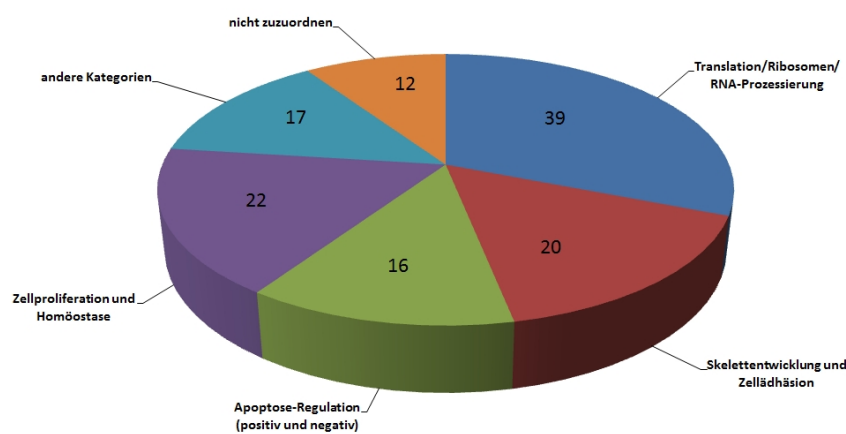


Abb. 15: Funktionelle Gruppen – der 100 höchst exprimierten Gene

dargestellt, stellen die an der Translation beteiligten Proteine, insbesondere die ribosomalen Proteine, die größte Gruppe unter den 100 höchst exprimierten Genen dar. In einer sehr heterogenen Gruppe sind solche Genprodukte zusammengefasst, die an Steuerung der Zellproliferation und der Homöostase beteiligt sind. Diese 22 Gene beinhalten IGFs, TGF- β bis hin zu Translations-Initiations-Faktoren und TNF-Rezeptoren. Die nächst kleinere Gruppe, mit 20 Genen, kann der Skelettentwicklung zugeordnet werden. Enthalten sind Strukturelemente der extrazellulären Matrix, wie beispielsweise Kollagene und bestimmte Proteoglykane, aber auch regulatorische Elemente wie TGF- β , CTGF, LECT1 und SPP1. Damit fallen 1/5 der 100 höchst exprimierten Gene in diese Rubrik.

Tab. 13: Funktionelle Gruppierung der 100 höchst repräsentierten EST-Sequenzen
 Von 100 Genen konnten 19 nicht eingeordnet werden. Die Gruppen der restlichen 81 sind hier dargestellt

DAVID_ID	Gen-Symbol	Beschreibung / Protein	ESTs
Translation/Ribosomen			
821677	<i>EE1A1</i>	eukaryotic translation elongation factor 1 alpha 1	14
780689	<i>RPL23</i>	ribosomal protein L23	14
816191	<i>RPS3</i>	ribosomal protein S3	12
784332	<i>RPL27</i>	ribosomal protein L27	11
788428	<i>RPS15a</i>	ribosomal protein S15a	11
794792	<i>RPLP1</i>	ribosomal protein, large, P1	11
779090	<i>RPL15</i>	ribosomal protein L15	10
782709	<i>RPL39</i>	ribosomal protein L39	9
803377	<i>RPL35</i>	ribosomal protein L35	8
781689	<i>RPS20</i>	ribosomal protein S20	8
799716	<i>RPL10</i>	ribosomal protein L10	7
790090	<i>RPL37</i>	ribosomal protein L37	7
825605	<i>RPL7</i>	ribosomal protein L7	7
776216	<i>RPS2</i>	ribosomal protein S2	7
789083	<i>RPS27</i>	ribosomal protein S27	7
810885	<i>RPL11</i>	ribosomal protein L11	6
822027	<i>RPL13</i>	ribosomal protein L13	6
795791	<i>RPL6</i>	ribosomal protein L6	6
797877	<i>RPS5</i>	ribosomal protein S5	6
796227	<i>RPLP0</i>	ribosomal protein, large, P0	6
793126	<i>EEF2</i>	eukaryotic translation elongation factor 2	5
783997	<i>RPL17</i>	ribosomal protein L17	5
793235	<i>RPL21</i>	ribosomal protein L21	5
778568	<i>RPL9</i>	ribosomal protein L9	5
805465	<i>RPS17</i>	ribosomal protein S17	5
816407	<i>RPS24</i>	ribosomal protein S24	5
781935	<i>RPS25</i>	ribosomal protein S25	5
775912	<i>RPSA</i>	ribosomal protein SA	5
773179	<i>EIF5A</i>	eukaryotic translation initiation factor 5A	4
776035	<i>RPL18</i>	ribosomal protein L18	4
819045	<i>RPL41</i>	ribosomal protein L41	4
786754	<i>RPL4</i>	ribosomal protein L4	4
808990	<i>RPS10</i>	ribosomal protein S10	4
786137	<i>RPS11</i>	ribosomal protein S11	4

DAVID_ID	Gen-Symbol	Beschreibung / Protein	ESTs
799749	<i>RPS19</i>	ribosomal protein S19	4
822433	<i>RPS29</i>	ribosomal protein S29	4
808508	<i>RPS6</i>	ribosomal protein S6	4
773042	<i>RPS7</i>	ribosomal protein S7	4
785199	<i>RPL36</i>	ribosomal protein L36	3
Skelettentwicklung und Zelladhäsion			
787944	<i>COL2A1</i>	collagen, type II, alpha 1	66
772843	<i>SPP1</i>	secreted phosphoprotein 1	52
803916	<i>COL10A1</i>	collagen, type X, alpha 1	48
826638	<i>COL9A1</i>	collagen, type IX, alpha 1	28
814112	<i>MGP</i>	matrix Gla protein	25
803177	<i>COL11A2</i>	collagen, type XI, alpha 2	24
808868	<i>LECT1</i>	leukocyte cell derived chemotaxin 1	23
807878	<i>COMP</i>	cartilage oligomeric matrix protein	20
826674	<i>HAPLN1</i>	hyaluronan and proteoglycan link protein 1	15
777404	<i>COL11A1</i>	collagen, type XI, alpha 1	14
775692	<i>CTGF</i>	connective tissue growth factor	12
800291	<i>FN1</i>	fibronectin 1	11
788090	<i>COL9A2</i>	collagen, type IX, alpha 2	10
816980	<i>SPARC</i>	secreted protein, acidic, cysteine-rich (osteonectin)	8
803057	<i>IGF2</i>	insulin-like growth factor 2 (somatomedin A)	5
825450	<i>SBDS</i>	ribosome maturation protein SBDS	4
794775	<i>TGFB</i>	transforming growth factor, beta 1	4
812851	<i>SPON1</i>	spondin 1, extracellular matrix protein	4
802704	<i>AGC1</i>	aggrecan	3
Apoptoseregulation (positiv und negativ)			
807878	<i>COMP</i>	cartilage oligomeric matrix protein	20
816191	<i>RPS3</i>	ribosomal protein S3 pseudogene 3	12
775480	<i>TIMP3</i>	TIMP metalloproteinase inhibitor 3	8
810885	<i>RPL11</i>	ribosomal protein L11	6
782724	<i>TPT1</i>	similar to tumor protein, translationally-controlled 1	6
793567	<i>SCAND1</i>	SCAN domain containing 1	5
803057	<i>IGF2</i>	insulin-like growth factor 2 (somatomedin A)	5
806903	<i>SOD2</i>	superoxide dismutase 2	5
773179	<i>EIF5A</i>	eukaryotic translation initiation factor 5A	4
808508	<i>RPS6</i>	ribosomal protein S6 pseudogene 25	4
794775	<i>TGFB</i>	transforming growth factor, beta 1	4
814301	<i>TNFRSF6B</i>	tumor necrosis factor receptor superfamily, member 6b	4
798664	<i>BAG1</i>	BCL2-associated athanogene	3
795442	<i>BIRC3</i>	baculoviral IAP repeat-containing 3	3
804233	<i>APOE</i>	hypothetical LOC100129500	3
Zellproliferation und Homöostase			
787944	<i>COL2A1</i>	collagen, type II, alpha 1	66
826638	<i>COL9A1</i>	collagen, type IX, alpha 1	28
809515	<i>FTH1</i>	ferritin, heavy polypeptide 1	26
803177	<i>COL11A2</i>	collagen, type XI, alpha 2	24
788428	<i>RPS15a</i>	ribosomal protein S15a pseudogene 17	11
816980	<i>SPARC</i>	secreted protein, acidic, cysteine-rich (osteonectin)	8
793494	<i>FTL</i>	similar to ferritin, light polypeptide	6
782724	<i>TPT1</i>	similar to tumor protein, translationally-controlled 1	6
793567	<i>SCAND1</i>	SCAN domain containing 1	5
803057	<i>IGF2</i>	insulin-like growth factor 2 (somatomedin A)	5
806903	<i>SOD2</i>	superoxide dismutase 2, mitochondrial	5
805465	<i>RPS17</i>	ribosomal protein S17	5
816407	<i>RPS24</i>	ribosomal protein S24	5
797736	<i>IL11</i>	interleukin 11	4
824572	<i>SSR1</i>	signal sequence receptor, alpha	4
794775	<i>TGFB</i>	transforming growth factor, beta 1	4
799749	<i>RPS19</i>	ribosomal protein S19 pseudogene 3	4

DAVID_ID	Gen-Symbol	Beschreibung / Protein	ESTs
808508	RPS6	ribosomal protein S6 pseudogene 25	4
814301	TNFRSF6B	tumor necrosis factor receptor superfamily, member 6b	4
786688	CTBP2	C-terminal binding protein 2	3
804233	APOE	hypothetical LOC100129500	3
773179	EIF5A	eukaryotic translation initiation factor 5A	2
Sonstige			
773105	CD109	CD109 molecule	10
774707	H19	H19, imprinted maternally expressed transcript	8
777021	PHOSPHO1	phosphatase, orphan 1	8
799015	COI9A3	collagen, type IX, alpha 3	7
787202	HDLBP	high density lipoprotein binding protein	5
788983	CEBPD	CCAAT/enhancer binding protein (C/EBP), delta	4
780333	G0S2	G0/G1switch 2	4
777269	RCN1	reticulocalbin 1, EF-hand calcium binding domain	4
810749	RCN2	reticulocalbin 2, EF-hand calcium binding domain	4
817974	ANGPTL2	angiopoietin-like 2	3
782629	CSNK1G2	casein kinase 1, gamma 2	3
823969	CKAP4	cytoskeleton-associated protein 4	3

Expressionsprofile knochen-/knorpelspezifischer Gene

Naturgemäß sind Gene für Kollagene in der Wachstumsfuge stark exprimiert. In Abbildung Abb. 16 sind die Häufigkeiten aller gefundenen Transkripte dieser Gruppe dargestellt. Insgesamt konnten 211 ESTs als Kollagene identifiziert werden. Davon entfallen alleine 187 auf die bereits genannten Typen II alpha 1, X alpha 1, IX alpha 1, XI alpha 1 und alpha 2. Mit

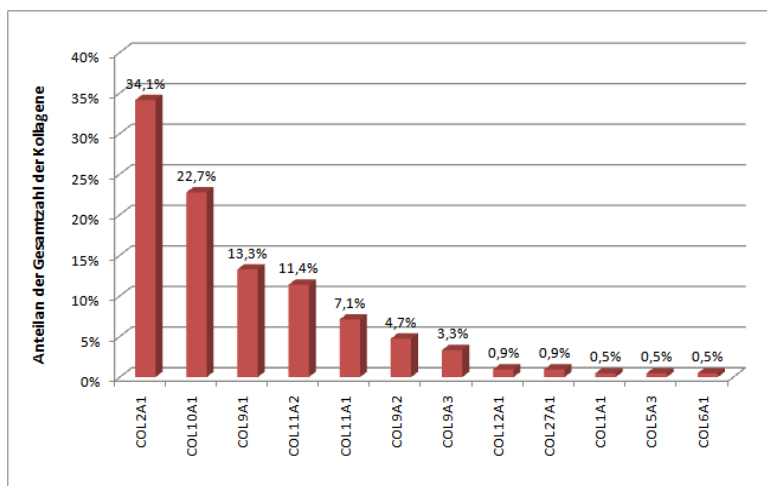
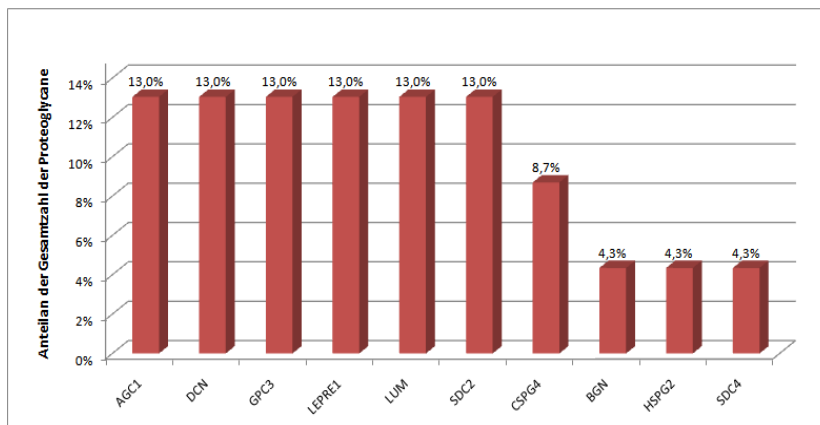


Abb. 16: EST-Expressionsprofil der Kollagene
 Dargestellt ist der relative Anteil an allen gefundenen Kollagen-Sequenzen; n=211

einer wesentlich geringeren Anzahl von ESTs konnten *col9A2* (10 ESTs) und *col9A3* (7 ESTs) nachgewiesen werden. Neben den Kollagenen vom Typ XII alpha1 und XXVII alpha 1, mit je zwei Kopien, und den Typen I alpha 1, V alpha 3, VI alpha 1, mit je einer Kopie, konnten keine weiteren Kollagene erfasst werden. Auf

Proteinebene bilden die Kollagene vom Typ IX und XI Heterotrimere, wobei jede Kette einmal vorkommt. Die Verhältnisse zwischen den nachgewiesenen ESTs der Kollagene vom

Typ XI alpha 1 zu alpha 2 beträgt 15:24. Bei Kollagen IX konnte ein Verhältnis von 28:10:7



(IXA1:IXA2:IXA3) festgestellt werden.

Ein weiterer essentieller Bestandteil der ECM sind die Proteoglykane. Sie bilden durch ihre Interaktion mit anderen Proteoglykanen und Kollagenen nicht nur eine

Abb. 18: EST-Expressionprofil der Proteoglykane
Relative Anzahl der gefundenen Proteoglykane dargestellt; n=23

wichtige Strukturkomponente, sondern sind auch am Transport von Molekülen und Ionen in der Matrix beteiligt. Im Rahmen der EST-Sequenzierung wurden lediglich 10 verschiedene Proteoglykane, und diese in geringer Kopienzahl nachgewiesen (Abb. 18). Dabei entfielen je 3 EST-Sequenzen auf Aggrecan (AGC1), Decorin (DCN), Glypican 3 (GPC3), „leucine- and proline-enriched proteoglycan 1“ (LEPRE1), Lumican (LUM) und Syndecan 2 (SDC2). „chondroitin sulfat proteoglycan NG2“ (CSPG4) konnte zweimal, Biglycan (BGN), „basement membrane-specific heparan sulfat proteoglycan core protein“ (HSPG2) und Syndecan 4 (SDC4) lediglich einmal nachgewiesen werden.

Die als restliche ECM-Proteine zusammengefassten

Bestandteile der extrazellulären Matrix umfassen eine Vielzahl struktureller Komponenten,

Proteine mit katalytischer oder regulatorischer Aktivität sowie n=164

solche mit Transporteigenschaften. Im Folgenden sind diese, sofern nachgewiesen, gemäß ihrer relativen Häufigkeit aufgeschlüsselt. Insgesamt konnten 164 Sequenzen dieser Gruppe zugeordnet werden. Die stärkste Expression zeigt SPP1 mit 52 ESTs, dies entspricht 32,1% aller gefundenen ECM Protein-Sequenzen. MGP (25 ESTs), COMP (20 ESTs) und LECT1 (20

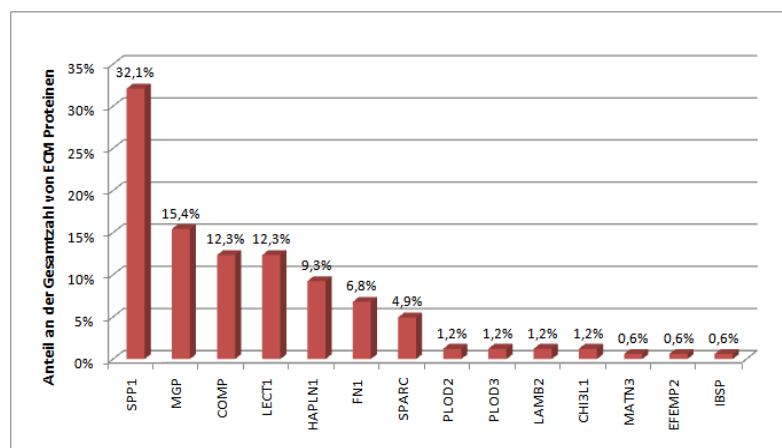


Abb. 17: EST-Expressionsprofil weiterer ECM-Proteine
Relative Anzahl der gefundenen weiterer ECM-Proteine dargestellt; n=164

ESTs) machen zusammen wiederum mehr als ein Drittel aller ECM-ESTs aus. Des Weiteren konnten 11 Sequenzen Fibronectin (FN1) und 8 Sequenzen Osteonectin (SPARC) zugeordnet werden. Mit lediglich 2 ESTs konnten PLOD2 und 3 (Prokollagen-lysine, 2-oxoglutarate-5-dioxygenase 2) sowie LAMB2 („Laminin subunit beta-2“) und CHI3L1 (Chitinase-3-like protein 1) nachgewiesen werden. Als einzelne EST-Sequenzen sind Matrilin 3 (MATN3), EFEMP2 und IBSP („Bone sialoprotein 2“) relativ selten exprimiert. Diese eher heterogen zusammengesetzte Gruppe enthält neben wichtigen matrix- und strukturbildenden Mitgliedern auch wichtige regulatorische (LECT1 und SPARC) oder modifizierende Proteine (PLOD2 und 3).

Ausführliche Tabellen mit allen erhobenen Daten befinden sich im elektronischen Anhang.

Expression ribosomaler Proteine

Immer wieder beschrieben ist die differentielle Expression ribosomaler Proteine in unterschiedlichen Geweben und Entwicklungsphasen (Kirby et al. 1995; Sim et al. 2010). Um Hinweise auf eine knochen- und knorpelspezifische Expression zu erhalten wurden die ESTs

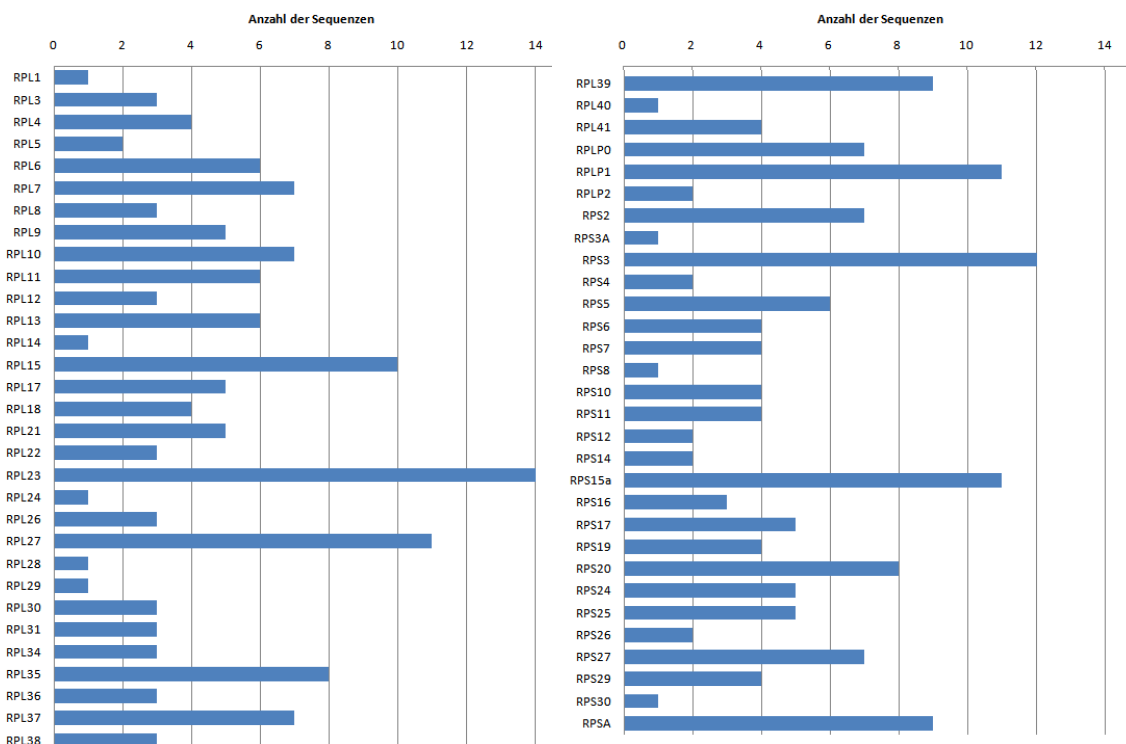


Abb. 19: EST-Expressionsprofil der ribosomalen Proteine n=289; neben den Proteinbezeichnungen ist die Anzahl der EST-Sequenzen dargestellt

für ribosomale Proteine separat betrachtet und das Profil, soweit möglich, mit Hilfe des „EST Profile Viewer“ (NCBI, UniGene) mit den Datenbanken verglichen. Insgesamt konnten 289 Sequenzen 61 verschiedenen ribosomalen Proteinen zugeordnet werden (Abb. 19).

Obwohl ESTs einiger ribosomaler Proteine in großer Zahl nachgewiesen werden konnten, bewegt sich der Großteil zwischen 1 und 4 Sequenzen. Gene für ribosomale Proteine werden daher als eher weniger stark exprimiert eingestuft. Etwas stärker vertreten sind die schon in Tabelle Tab. 12 genannten RPL23 (14 ESTs), RPS3 (12 ESTs), RPL27 (11 ESTs), RPLP1 (11 ESTs) und RPS15a (11 ESTs). Darüber hinaus fallen lediglich RPL39 und RPSA mit je 9 ESTs sowie RPL35 und RPS20 mit je 8 ESTs auf. Um eine erhöhte Expression erkennen zu können, wurden die hier festgestellten ESTs auf eine Million EST normalisiert und dann mit der generellen Häufigkeit RP-ESTs in EST-Datenbanken verglichen. Dabei zeigte sich keine größere Häufigkeit der Expression von ribosomalen Protein-Genen in der Wachstumsfuge des Rinderknochens.

3.2 Charakterisierung einzelner Sequenz-Cluster und EST-Klone

Während der vorangegangenen Analysen konnten zwei Gruppen von Sequenzen nicht eindeutig einer Funktion oder einem Gen zugeordnet werden:

1. Die hoch exprimierten Loci LOC618319 (24 ESTs), LOC514954 (20 ESTs) und LOC511905 mit 11 EST-Sequenzen.
2. 226 Sequenzen, die weder eindeutige Datenbankübereinstimmungen aufwiesen noch besonders häufig auftraten.

Um in den EST-Daten enthaltene neue Transkripte oder Transkriptvarianten zu identifizieren, wurden die assemblierten EST-Daten dieser beiden Gruppen für genomische Vergleiche genutzt. Dabei wurde besonders auf die typischen Exon-/Intronstrukturen geachtet. Durch den Vergleich der abgeleiteten Proteinsequenzen mit verschiedenen Proteindatenbanken sollten mögliche Domänen oder Signalpeptide identifiziert werden, die Rückschlüsse auf eine mögliche Funktion zulassen.

Unbekannte, stark exprimierte ESTs

Im genomischen Vergleich konnte LOC511905 eindeutig dem Locus des Gens *H19* auf Chromosom 29 von *Bos taurus* zugeordnet werden. Dabei stimmt die in der Datenbank hinterlegte Exon-/Intronstruktur mit der der ESTs überein (Abb. 20 B). Da es sich bei *H19* um ein nicht proteincodierendes Gen handelt, konnten auch keine Homologien in den Proteindatenbanken gefunden werden.

Im Gegensatz zu LOC511905 konnten die EST-Sequenzen für LOC514954 keinem bestimmten

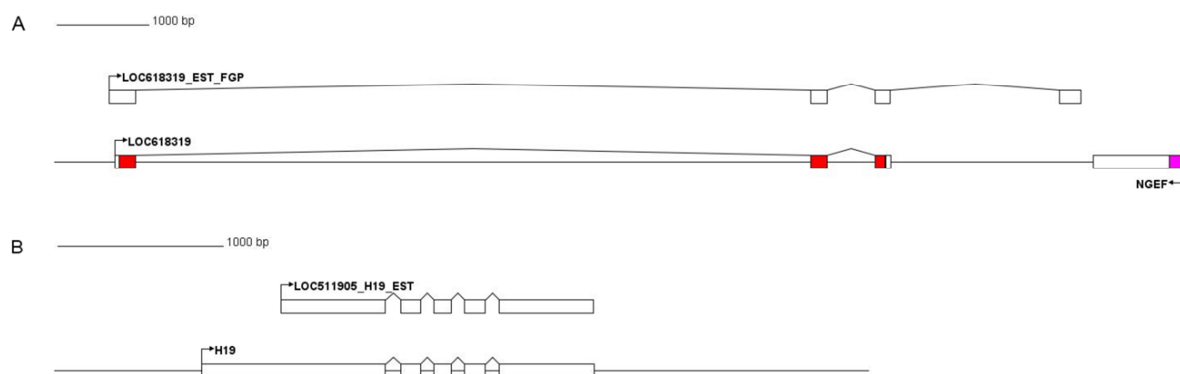


Abb. 20: EST-Cluster LOC618319 und LOC511905 genomische Organisation
 Verglichen wird die in der Datenbank hinterlegte Annotatin am LOC61319 (A) und LOC511905 (B) mit den gefundenen Transkripten

genomischen Locus zugeordnet werden. Sie zeigen genomweite multiple Übereinstimmungen auf verschiedensten Chromosomen. Diese bei ribosomalen Proteinen verbreitete Eigenschaft unterstützt die Identifikation als ribosomales Protein L41 über Datenbanksuchen. Dieses liegt in mehreren Kopien, unter anderem in Form von Pseudogenen, im Genom vor.

Die EST-Sequenzen am LOC618319 zeigen im Vergleich zum gleichnamigen genomischen Locus auf Chromosom 3 von *Bos taurus* vier Exon-Bereiche, von denen bisher nur drei durch die Datenbanken unterstützt werden (Abb. 20 A). Zudem konnten mit Hilfe der Proteindatenbanken sowohl Transmembrandomänen als auch ein mögliches Signalpeptid identifiziert werden.

Unbekannte, weniger stark exprimierte ESTs

Um die Anzahl der zu untersuchenden Sequenzen weiter einzugrenzen, wurden die 226 EST-Klone, die bisher nicht identifiziert werden konnten, nochmals untereinander geclustert und

die genomischen Loci näher betrachtet und automatisierte Domänensuchen durchgeführt (Blast2GO, InterProScan, EBI). Dabei konnten nochmals 80 Sequenzen, die bei der vorangegangenen Annotation nicht zugeordnet werden konnten, eindeutig bekannten Genen zugeordnet werden. Dabei fällt auf, dass die Übereinstimmungen meist im äußersten UTR-Bereich zu finden sind. Möglicherweise ist dies auch ein Grund für die fehlgeschlagene Identifikation in den vorherigen Analysen. Darüber hinaus wurden zwei weitere Contigs als Vektor und acht als ribosomale RNA identifiziert.

In den verbleibenden 122 Konsensus-Sequenzen wurden zunächst Hinweise auf proteinkodierende Bereiche gesucht. Dabei wurde auf eine Polyadenylierung, die Exon-/Intronstruktur, partielle Übereinstimmungen zu bekannten Proteinen oder Domänen geachtet. Neben einigen repetitiven Elementen konnten bei einer Vielzahl von Sequenzen lediglich einzelne genomische Übereinstimmungen gefunden werden. 104 Sequenzen zeigten keines der oben genannten Kriterien und wurden daher nicht näher untersucht. Lediglich 18 EST-Cluster oder Einzelsequenzen wurden genauer überprüft (Tab. 14).

Tab. 14: Mögliche proteinkodierende EST-Sequenzen

Contig	Klon	Accession Nr.	Beschreibung	E-value
52	B01_T9F2_P35	AL441886.18	Human DNA sequence from clone RP11-287G8 on chromosome 10 Contains the 5'end of gene KIAA1010, the 3'end of gene FLJ40792	3,00E-17
136	D06_T1F3_P14	NM_001110094.1	Bos taurus hypothetical protein LOC782657 (LOC782657)	0
36	E04_T9F2_P40	AK021932.1	Homo sapiens cDNA FLJ11870 fis, clone HEMBA1007051	3,00E-20
160	E09_T1F2_P5	NM_001045957.1	Bos taurus chromosome 17 open reading frame 61 ortholog (C19H17orf61)	2,00E-17
85	E10_T8F2_P30	XM_002818443.1	PREDICTED: Pongo abelii hypothetical protein LOC100455770 (LOC100455770), mRNA	2,00E-08
98	F12_T82_P28	XM_002685421.1	PREDICTED: Bos taurus hypothetical protein LOC100336997 (LOC100336997)	0
166	G03_T1F2_P1_P2_S05	HM045018.1	Bos taurus breed Heck cattle mitochondrion, complete genome	e-172
90	G04_T8F2_P25		Kein signifikanter BLAST-Teffer	
129	G05_T1F3_P16	EZ467942.1	TSA: Mustela putorius furo Ferret_c11503, complete sequence	1,00E-22
99	G05_T82_P28	XM_002688250.1	PREDICTED: Bos taurus hypothetical LOC785528 (CWH43)	5,00E-05
162	G10_T2F1_P13	FP102657.5	Pig DNA sequence from clone CH242-272B23 on chromosome X, complete sequence	6,00E-18
70	H11_T9F2_P32	AC150871.6	Bos taurus BAC CH240-316E7 (Children's Hospital Oakland Research Institute Bovine BAC Library complete sequence	2,00E-09
68	H11_T9F2_P33	NM_001135578.1	Bos taurus hypothetical protein LOC100140337 (LOC100140337)	0
34	H11_T9F2_P41	AC122190.4	Mus musculus BAC clone RP23-41J14 from 12, complete sequence	3,00E-05
89	C07_T8F2_P25	BC134702.1	Bos taurus cDNA clone IMAGE:7944277	0
13	E07_T1F2_P6	AL391294.16	Human DNA sequence from clone RP11-196G4 on chromosome X Contains the 3' end of the NSBP1 gene for nucleosomal binding protein 1 and a CpG island, complete sequence	2,00E-07
39	E12_T9F2_P38	CU633489.14	Pig DNA sequence from clone CH242-15J9 on chromosome 5, complete sequence	4,00E-22
20	E12_T9F2_P44	XM_002700877.1	PREDICTED: Bos taurus hypothetical protein LOC100335754 (LOC100335754)	0

Hierbei stellt Contig 13 mit 5 Sequenzen das stärkste Cluster dar. Dennoch konnten keine Exon/Intron-artigen Strukturen im genomischen Vergleich festgestellt werden. Auch zeigten sich nur geringe Evidenzen für Homologien zu funktionellen Proteinen. Aufgrund fehlender Polyadenylierung wurden die Klone zunächst nicht weiter analysiert.

Contig 89 mit 4 Sequenzen hingegen zeigt eindeutig eine Polyadenylierung mit vorherigem Polyadenylierungssignal. Durch die Sequenzierung mit dem entgegengesetzten Primer konnte keine zusätzliche Information gewonnen werden. Der Contig überspannt keine Introns und zeigt multiple Treffer zur genomischen Sequenz. Dies bestätigt sich auch in anderen Spezies wie Mensch, Affe und Maus. Übereinstimmungen auf Proteinebene finden sich lediglich zu hypothetischen Proteinen.

Contig 20, im Weiteren mit E12P44 bezeichnet, beinhaltet neben E12_T9F2_P44 noch 2 weitere kürzere Sequenzen. Dabei gibt es Hinweise auf eine Polyadenylierung und im genomischen Vergleich sind klare Exon-/Intronstrukturen zu erkennen. Obwohl in Abb. 21 noch nicht annotiert, unterstützen weitere Datenbankeinträge diese Struktur. Ein mögliches

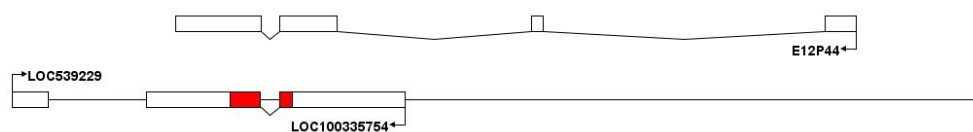


Abb. 21: Contig 20 - E12P44 - genomische Organisation

Produkt aus dem relativ kurzen ORF zeigt kaum Übereinstimmungen zu bekannten Proteinen und lässt keine Beurteilung der Relevanz für das Knochenwachstum zu.

Der aus zwei EST-Sequenzen bestehende Contig 68 konnte einem Locus auf Chromosom 23 von *Bos taurus* zugeordnet werden. Dabei ist die Übereinstimmung auf den 3'UTR-Bereich von *LOC100140337* beschränkt. Darüber hinaus gibt es Hinweise auf ein Signalpeptid und eine überlappende Transmembrandomäne im Anfangsbereich des 255 Nukleotide langen ORF. Weitere Informationen oder Indizien für eine wachstumsfugenassoziierte Funktion lassen sich nicht finden.

Der Klon D06P14 konnte in der weiteren Analyse mit zwei weiteren, schon identifizierten EST-Sequenzen geclustert werden (Contig 136). Wie Abb. 22 B zeigt, deckt die Konsensus-Sequenz den Großteil von *LOC782657* auf Chromosom 9 ab. Der ORF über die vier

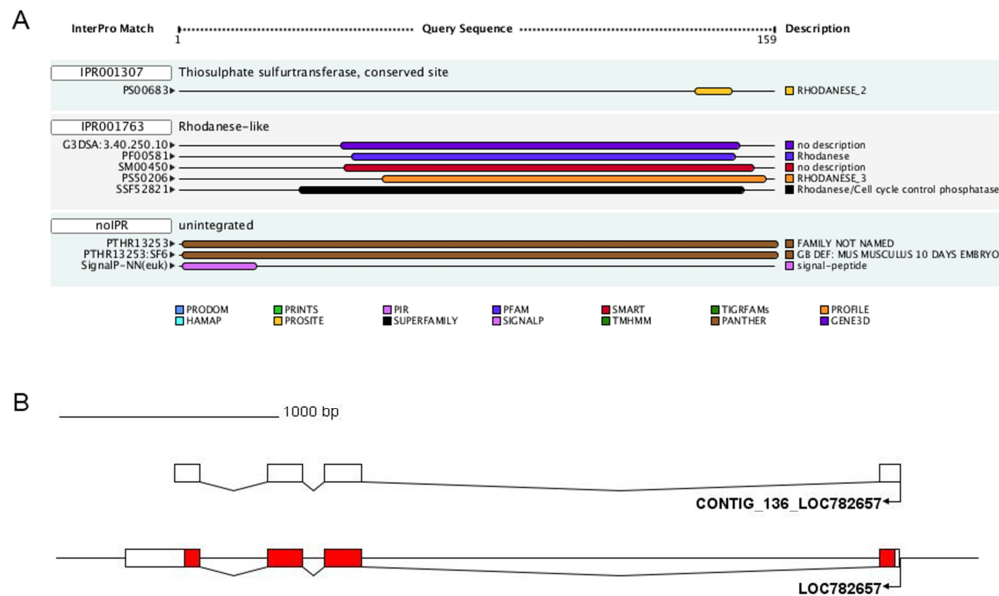


Abb. 22: Genomische Organisation und Domänen - Contig 136
 A Ergebnisse der InterPro-Suchen nach funktionellen Domänen (© European Bioinformatics Institute); B Exon-/Intronstruktur am Locus LOC782657; der offene Leserahmen ist rot markiert.

Exonbereiche ist vollständig vorhanden. Lediglich ein Teil des UTR von Exon 4 fehlt. Zudem finden sich Hinweise auf eine frühe Polyadenylierung noch vor dem annotierten Transkriptende. Die vorhandenen Domänen (Abb. 22 A) weisen das Genprodukt als ein Mitglied der Rhodanese-Superfamilie aus. Rhodanesen gehören zu den Sulfurtransferasen. Diese katalysieren den Schwefeltransfer von Thiosulfaten auf Cyanid (WESTLEY 1959; Bordo und Bork 2002).

Alle weiteren untersuchten Sequenzen sind lediglich durch einen einzigen EST-Klon abgedeckt. In den meisten Fällen handelt es sich um einzelne genomische Übereinstimmungen ohne erkennbare Exon-/Intronstruktur.

Contig 166 beispielsweise lieferte über die mitochondriale Zuordnung hinaus keine weiteren Erkenntnisse. Klon E12P38 konnte eindeutig der Gen-Region für das Carrierprotein SLC2A13 zugeordnet werden und grenzt direkt an ein Exon. Allerdings sind keine Hinweise auf Spleißen oder Polyadenylierung vorhanden. Außerdem zeigt sich der betreffende

genomische Bereich mit ca. 74 % sehr A/T-reich. Dies erhöht die Möglichkeit eines „Fehlprimings“ bei PolyA⁺-Aufreinigung und cDNA-Synthese mit Oligo-d(T)-Nukleotiden.

Contig 160 liegt ebenfalls angrenzend zu einem Exon im Bereich des Gens *CTAGE* (Chromosom 21), zeigt aber im Gegensatz zu E12P38 eindeutige Hinweise auf eine Polyadenylierung und besitzt auch ein mögliches Polyadenylierungssignal. Zudem können 42 Nukleotide am 3'-Ende nicht dem Locus *CTAGE* zugeordnet werden. Diese zeigen neben kurzen multiplen Übereinstimmungen zu mehreren genomischen Loci einen eindeutigen Treffer zum 3'UTR-Bereich von Chr19H17orf61. Darüber hinaus konnten keine Übereinstimmungen zu Chr19H17orf61 gefunden werden.

B01P35 zeigt auch eindeutig eine Polyadenylierung. Diese ist genomisch nicht vorhanden. Der Klon konnte im Intronbereich des „*Telomeric repeat binding factor (NIMA-interacting) 1*“ (*TERF1*) lokalisiert werden. Eine Übereinstimmung mit dem kodierenden Bereich konnte nicht festgestellt werden. Der Vergleich mit Proteindatenbanken ergab lediglich partielle Homologien zu reversen Transkriptasen. Dem zufolge handelt es sich mit großer Wahrscheinlichkeit um den Teil eines Transposons.

Auf Chromosom 16 gelegen, zeigt die genomische Organisation des cDNA-Klons F07P42 eine eindeutige Exon-/Intronstruktur, die über GT/AG-Übergänge abgesichert ist. Von den

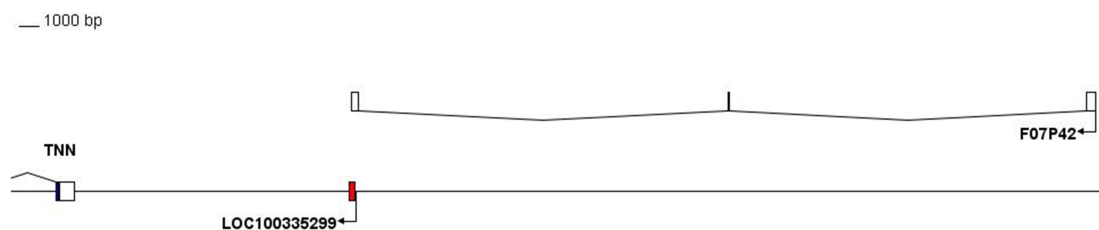


Abb. 23: genomischer Locus von F07P42
In der Nachbarschaft zu TNN („Tenascin“) findet sich der annotierte ORF von LOC100335299 (rot), welcher sich teilweise mit dem vermutlichen Locus des EST-Klons F07P42 überschneidet

möglichen offenen Leserahmen zeigen lediglich die ersten 500 Nukleotide im Leserahmen +1 der cDNA-Sequenz geringe Homologien zu einigen unbekannt Proteinen. Im 3'-Bereich von Base 667 bis 744 gibt es weitaus bessere und häufigere Übereinstimmungen zu etlichen unbekannt Proteinen und cDNA-Klonen. Darüber hinaus liegt hier eine potentielle Transmembrandomäne vor. Diese Region überschneidet sich mit dem bereits annotierten LOC100335299. Gegen einen zusammenhängenden ORF sprechen drei zwischen den beiden einzelnen ORFs liegende Stoppcodons.

Synthänie und Interspeziesvergleich

In ausgewählten Fällen wurden Vergleiche der genomischen Sequenz zwischen verschiedenen Spezies angestellt. Ziel solcher Analysen ist das Auffinden eventuell konservierter Bereiche die Rückschlüsse auf zusätzliche Exonbereiche oder Promotorstrukturen zulassen. Dabei standen die beiden Loci, *LOC618319* und *F07P42*, wegen ihrer besonderen Bedeutung im Vordergrund.

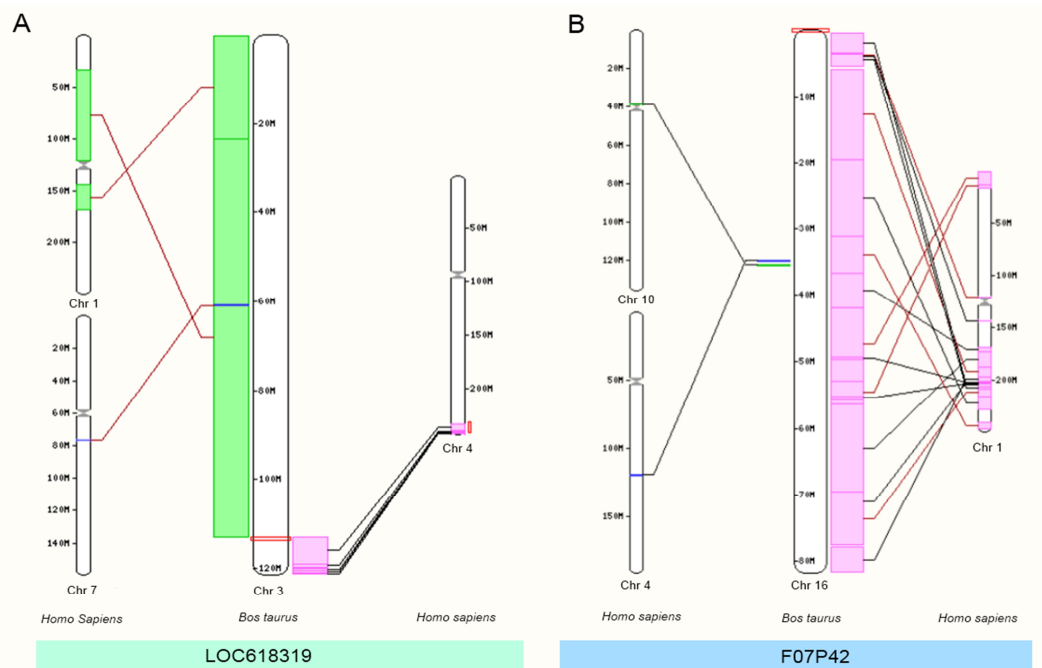


Abb. 24: Chromosomale Translokationen
 Dargestellt ist die Lage der synthänen Bereiche der genomische Loci von *LOC618319* und *F07P42* bei *Bos taurus* (jeweils mittig) und *Homo sapiens*

Sind im Fall von *LOC618319* in vielen Spezies bis hin zu *Gallus gallus* sehr gute bis moderate Homologien in den transkribierten Bereichen festzustellen, so zeigt *F07P42* lediglich zu einem einzigen Locus auf Chromosom 1 des Menschen (Abb. 24) eine hohe Übereinstimmung. Diese beschränkt sich allerdings nicht auf mögliche transkribierte Bereiche, sondern erstreckt sich nahezu über die gesamte Region. Die Übereinstimmungen zu Sequenzen aus dem Schimpansen, Hund oder Maus zeigen bestenfalls marginale Homologien (Abb. 25 B). Daher lassen sich im Fall von *F07P42* keine potentiell konservierten Strukturen ableiten.

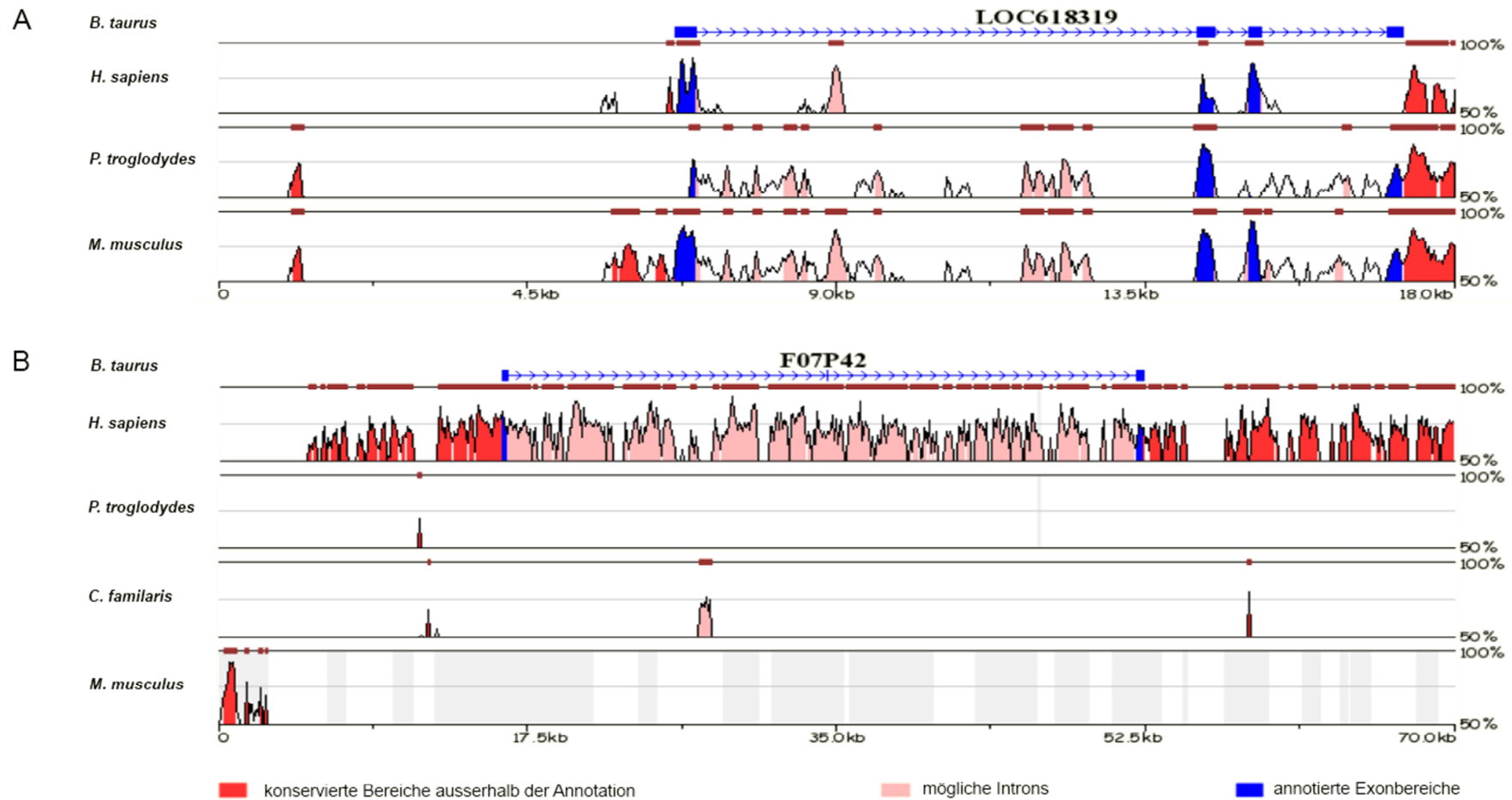


Abb. 25: Interspeziesvergleich - *LOC618319* und F07P42

A Vergleich der genomischen Region von *LOC618319* mit dem homologen Bereich aus Mensch, Schimpanse und Maus; die Histogramme zeigen die Sequenzübereinstimmung ab einem Wert von 50 %; B Vergleich der homologen genomischen Sequenzen zu F07P42 aus Rind mit den Regionen aus Mensch, Schimpanse Hund und Maus

Im Gegensatz dazu sind bei LOC618319 deutlich stärker konservierte Exonbereiche zu erkennen (Abb. 25 A). Diese variieren leicht zwischen den einzelnen Spezies (*Homo sapiens*, *Pan troglodydes*, *Mus musculus*), bestätigen aber alle in der Datenbank annotierten Exonbereiche und zusätzlich das im EST-Projekt gefundene Exon 4. Wobei letzteres im Menschen auf Nukleotidebene nicht konserviert vorzuliegen scheint.

Darüber hinaus finden sich weitere moderat konservierte Regionen in den beiden größeren der drei Intronbereiche. Interessant ist auch der 5' vom kodierenden Bereich liegende Mikrosatellit mit seiner (GA)_n/(GT)_n-SSR-Struktur. Dieser tritt im Menschen nur rudimentär und in der Maus wesentlich ausgeprägter gegenüber der Rinder-Sequenz auf.

RT-PCR zum Nachweis von interessanten Transkripten in verschiedenen Geweben

Aufgrund der vorherigen Analysen der EST-Sequenzen wurden verschiedene Gene bzw. EST-Sequenzen zum Nachweis über PCR auf cDNA verschiedener Gewebe ausgewählt. Dazu wurde *GAPDH* als Positivkontrolle eingesetzt, *COL2A1* und *MGP* als bekannte hoch exprimierte Gene der Knochen- und Knorpelentwicklung sowie *ID1* und *TPT1* als hoch exprimierte bekannte Gene ohne speziellen Bezug zu Knochen- oder Knorpelgewebe. Dabei stand außerdem der Nachweis der in der Wachstumsfuge hoch exprimierten Loci LOC618319 und LOC511905/H19 sowie der niedrig exprimierten Transkripte der cDNA-Klone E12P44 und F07P42 im Mittelpunkt. Letztere zeigten bei den bioinformatischen Analysen neue Exonbereiche (E12P44) oder möglicherweise ein neues, noch nicht annotiertes Gen (F07P42). Für die PCR wurde cDNA aus der fötalen Wachstumsfuge sowie Herz, Leber, Lunge und Niere aus adulten Tieren verwendet. Dabei wurde RNA von guter bis sehr guter Qualität (RIN>8) eingesetzt.

Bei der PCR mit Exon überspannenden Primern konnten in allen Geweben Transkripte von *GAPDH* detektiert werden. Knochen- und Knorpel-assoziierte Gene wie *COL2A1* oder *MGP* konnten in großer Menge nur in der cDNA der Wachstumsfuge nachgewiesen werden. Allerdings ergaben sich auch schwache positive Nachweise in Herz und Lunge (*MGP*) bzw.

Tab. 15: PCR mit cDNA verschiedener Gewebe von *Bos taurus*

Verglichen werden die Größen der Amplifikate nach Software unterstützter Auswertung der Gelelektrophorese

Amplikon	Primer	erwartete Größe	nachgewiesene Transkripte in				
			Wachstumsfuge	Herz	Leber	Lunge	Niere
GAPDH	Bt_1/2	693	700 bp	700 bp	700 bp	710 bp	700 bp
COL2A1	Bt_3/4	519	520 bp	NN	NN	NN	490 bp
MGP	Bt_5/6	330	310 bp	320 bp	NN	310 bp	NN
ID1	Bt_11/12	290	300 bp	260 bp	240 bp	240 bp	NN
E12P44	Bt_21/22	349	339 bp	290 bp	290 bp	310 bp	350 bp
F07P42	Bt_31/32	384	380 bp	NN	NN	NN	NN
511905	Bt_37/38	566	580 bp	NN	NN	NN	NN
618319	Bt_39/40	247	290 bp	NN	NN	NN	NN
TPT1	Bt_45/46	471	330 bp	330 bp	330bp	240 bp	200 bp

Niere (COL2A1) (vgl. Tab. 15). Dies gilt auch für Transkripte homolog zu „Inhibitor of DNA binding 1“ (ID1) und „Translationally-controlled tumor protein“ (TPT1). Diese konnten aus allen Geweben außer der Niere amplifiziert werden. Transkripte homolog zu dem cDNA-Klon E12P44, im EST-Projekt mit lediglich einer Sequenz vertreten, konnte in allen Geweben nachgewiesen werden. Die beiden in der Wachstumsfuge hoch exprimierten Gene *LOC511905* und *LOC618319* zeigten nur für die cDNA aus der fötalen Wachstumsfuge spezifische Amplifikate.

Semi-quantitative-RT-PCR

Als Alternative zur quantitativen RT-PCR wurden außerdem semi-quantitativ die Expressionslevel in Relation zu GAPDH bestimmt. Dabei diente COL2A1 als höchst exprimiertes Gen im EST-Projekt als Positivkontrolle. Des Weiteren sollte die Expression des ebenfalls hoch exprimierten ID1 überprüft werden. Dazu wurden parallel zu den jeweils EST-spezifischen Primern zusätzlich solche für GAPDH verwendet. Der jeweils aus den spezifischen Bandenintensitäten gebildete Koeffizient ist in der folgenden Tabelle aufgeführt. Qualitativ konnten die Ergebnisse aus den vorhergehenden RT-PCR bestätigt werden. Tendenziell zeigen analog zu den EST-Auswertungen COL2A1 und LOC618319 eine hohe und wachstumsfugenspezifische Expression.

Tab. 16: RT-PCR zum Nachweis interessanter Transkripte in verschiedenen Geweben

Enthalten sind jeweils die Koeffizienten aus der Stärke der amplifizierten Bande des jeweiligen Transkripts und der Bandenstärke der parallel durchgeführten Amplifikation von *GAPDH*; nicht nachweisbare Transkripte sind mit „NN“ gekennzeichnet

Amplikon	Primer	Amplifikate in Relation zu GAPDH				
		Wachstumsfuge	Herz	Leber	Lunge	Niere
COL2A1	Bt_3/4	1,5	NN	NN	NN	0,4
ID1	Bt_11/12	0,7	0,3	1,5	0,8	NN
E12P44	Bt_21/22	1,1	0,9	0,4	0,9	1,1
F07P42	Bt_31/32	0,4	NN	NN	NN	NN
LOC618319	Bt_39/40	1,5	NN	NN	NN	NN

Auch die Ergebnisse von F07P42 und E12P44 lassen sich noch mit der Anzahl der EST-Klone vergleichen. Auch hier zeigt F07P42 eine wachstumsfugenspezifische niedrige Expression. E12P44 kann außerdem noch in anderen Geweben nachgewiesen werden und scheint ebenso stark in der Niere exprimiert zu werden. Im Unterschied zu der Anzahl der EST-Klone zeigt ID1 eine wesentlich geringere Expression als beispielsweise E12P44, aber auch als LOC618319 und COL2A1. Die größte Menge an PCR-Produkt konnte in der Leber nachgewiesen werden.

Analyse der Transkriptvarianten des LOC618319

Das mit dem genomischen Locus 618319 übereinstimmende EST-Cluster enthält 23 Sequenzen und stellt damit eine der meistvertretenen Sequenzen des Projekts dar (vgl. Tab. 12, Seite 61). Da, wie bereits gezeigt wurde, ein großer Anteil der hoch exprimierten Gene an der Knochen- und Knorpelbildung beteiligt ist, wäre dies auch für LOC618319 zu vermuten. Obwohl das theoretische Gesamt-Transkript aus allen 23 kombinierten Sequenzen ca. 870 bp umfasst, werden die letzten 220 bp lediglich durch eine Sequenz repräsentiert und in allen 23 Fällen folgt eine erkennbare Polyadenylierung. Diese erfolgt wenige Basen nach dem ebenfalls überall enthaltenen Polyadenylierungssignal (vgl. Abb. 26).

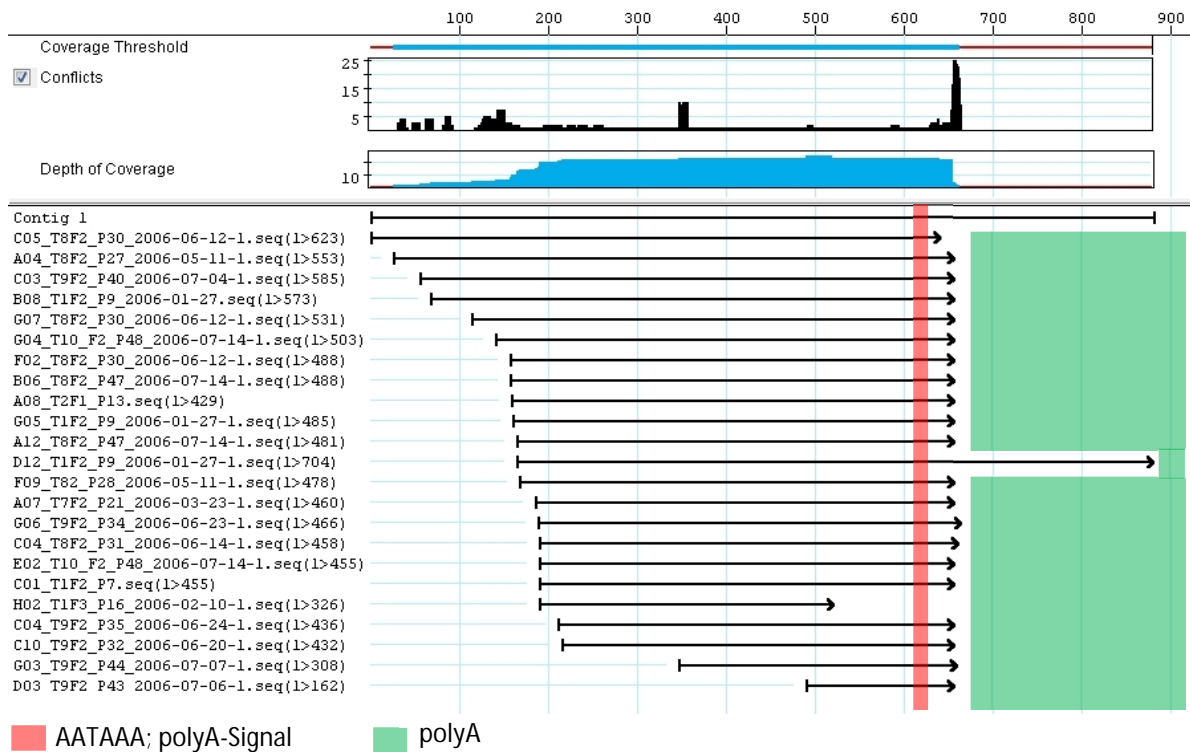


Abb. 26: Übersicht aller Sequenzen des Contigs LOC618319
 Schematische Übersicht der Anordnung aller den Locus betreffenden Sequenzen aus dem EST-Projekt; markiert sind das Polyadenylierungssignal und die Polyadenylierung an sich

Durch die gesamte Konsensus-Sequenz der ESTs zieht sich ein durchgehender offener Leserahmen (ORF), der lediglich durch zwei aufeinanderfolgende Stopp-Codons beendet wird. Dies trifft auch für die Sequenz des längeren cDNA-Klons D12P9 zu. Allerdings setzt sich stromabwärts des Stopps der theoretische ORF fort.

In zwei der 23 Sequenzen konnten jeweils kleinere Deletionen nachgewiesen werden. In Klon A04_P27 fehlen 63 bp von Position 118 bis 181, in Klon A08_P13 sind 57 Nukleotide deletiert (Position 339-396). Der Leserahmen verschiebt sich in beiden Fällen durch die Deletion nicht. Exon-Intron-Grenzen gemäß der GT/AG-Regel sind offensichtlich nicht vorhanden. In der nachfolgenden Abbildung ist die genomische Organisation der vier bisher detektierten Varianten nochmals dargestellt. Alle LOC618319-EST-Klone lassen sich einem Bereich von ca. 10 kbp auf Chromosom 3 von *Bos taurus* in direkter Nachbarschaft von GIGYF2 und NGEF zuordnen. In der Grundform liegt eine Organisation in drei Exonbereiche vor.

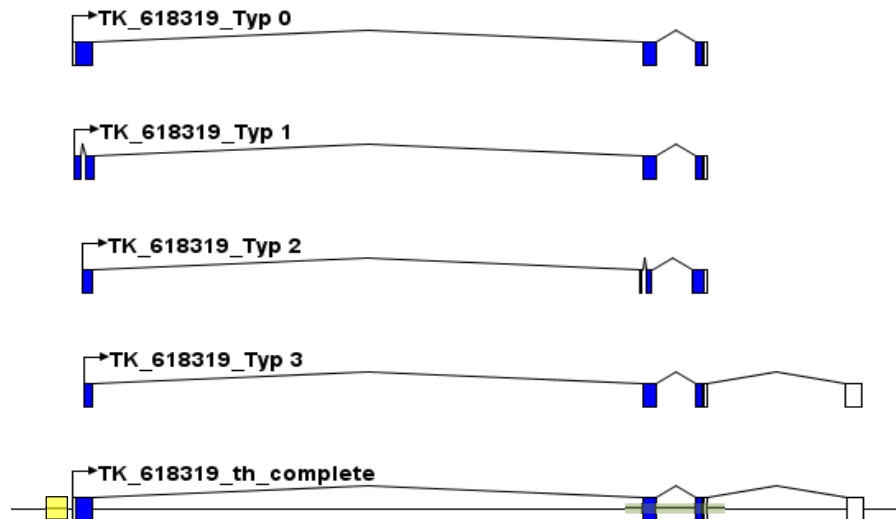


Abb. 27: EST-Sequenzvarianten und genomische Zuordnung (LOC618319)
 TK_618319_Typ0-3 zeigen die vier durch EST-Sequenzen unterstützte Transkriptvarianten in ihrer genomischen Organisation; TK_618319_th_complete zeigt die aus der Konsensussequenz abgeleitete Struktur des Genlocus; gelb: SSR; grün: CpG-Region; blau: vermutete kodierende Exons

Nur wenige Nukleotide vor dem Beginn der kodierenden Sequenz (Abb. 27) befindet sich ein Cluster von Mikrosatellitensequenzen („simple sequence repeats“ – SSR) vom Typ (TG)_n/(GA)_n. Obwohl im 5'-Bereich nicht ausgeprägt vorhanden, liegen sowohl Exon 2 als auch Exon 3 in einer CpG-Region (Abb. 27; grün). Der offene Leserahmen (Pos. 21-194) erlaubt die Synthese eines 173 Aminosäuren langen Polypeptids. Zwei Varianten unterscheiden sich lediglich in einem Aminosäureaustausch an Position 115 des AS-Alignments (Abb. 28). Die Deletionen der Typen 1 und 2 würden zu einer Verkürzung des Polypeptides um 21, bzw. 19 Aminosäuren führen, sofern diese Varianten translatiert würden.

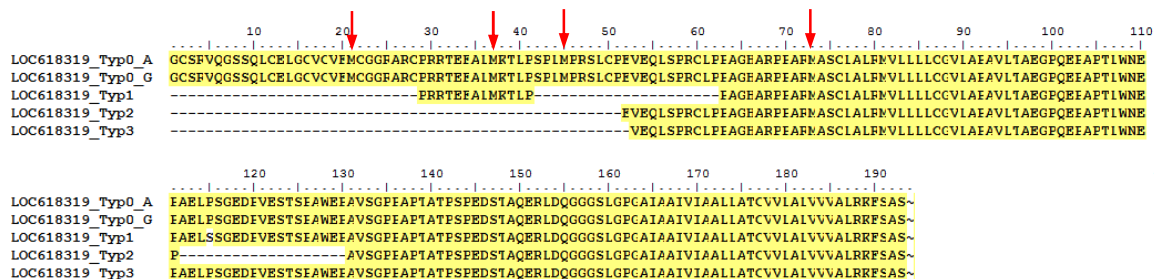


Abb. 28: Aminosäurealignment der möglichen Isoformen von LOC618319

Der sich theoretisch hinter dem Stopp-Codon fortsetzende Leserahmen der Variante Typ 3 zeigt bei Blast-Suchen keine signifikanten Übereinstimmungen.

LOC618319 zeigt neben der wachstumsfugenspezifischen Expression gleich mehrere Besonderheiten. Selbst mit angepassten PCR-Bedingungen und Strategien, wie die Verwendung verschiedener Polymerasen oder Gradienten-PCR und „Touch-down“-PCR, kam es über die erwartete Fragmentlänge hinaus zu weiteren Amplifikaten. Abb. 29 zeigt ein typisches Ergebnis. Diese konnten mittels Gelextraktion separiert, allerdings nicht sequenziert oder kloniert werden. Die Sequenzierungen von PCR-Produkten aus dem Bereich der erwarteten Größe zeigen allerdings inkonsistente Sequenzabschnitte mit Deletionen variabler Größe. Ähnliches lässt sich bei den Amplifikaten mit F07P42-spezifischen Primern beobachten.

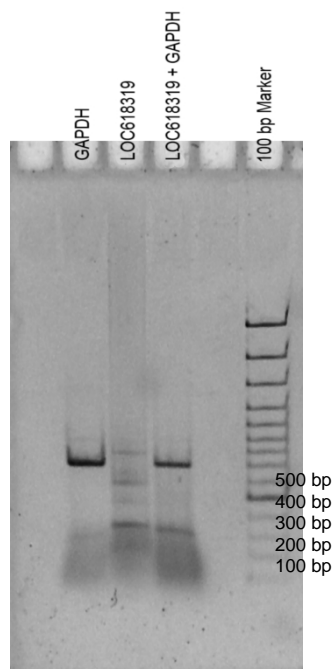


Abb. 29: RT-PCR LOC618319

Gelelektrophoretische Auftrennung im 1.2% Agarosegel der PCR Produkte auf cDNA aus der fötalen Wachstumfuge; für die Amplifikation von *GAPDH* wurden die Primer Bt_1 und Bt_2, für LOC618319 Bt_39 und Bt_40 verwendet

3.3 Genexpressionsanalysen mittels Roche/454-Sequenzierung

Für die 454-Sequenzierung wurden zunächst aus ca. 100 µg Gesamt-RNA der bovinen fötalen Wachstumsfuge 12 µg polyA⁺-RNA angereichert. Daraus ergaben sich 5 µg doppelsträngige cDNA. Diese wurden im Institut für Immungenetik (Kaiserslautern) zur Sequenzierung mit Hilfe der 454/FLX-Technologie verwendet. Dabei wurden 38492 Sequenzen mit einer

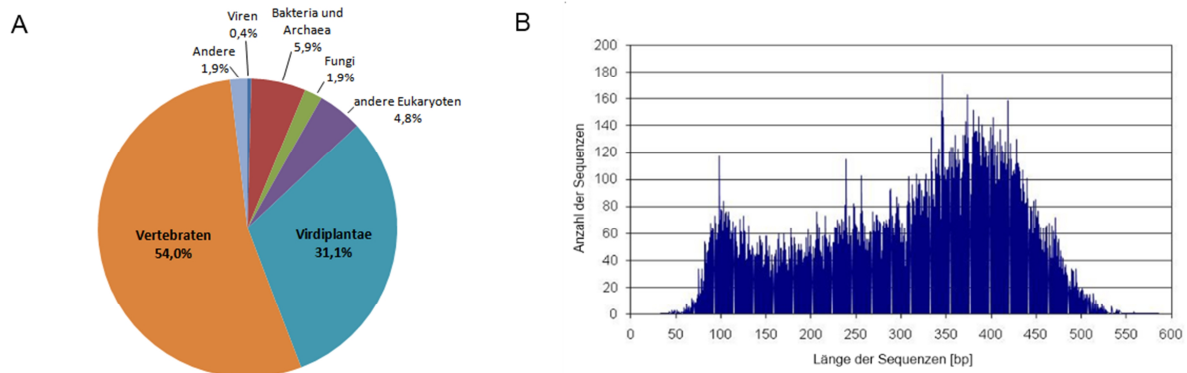


Abb. 30: Zusammensetzung der 454-Daten

A Statistische Auswertung (MEGAN; Huson et al. 2007) der getroffenen Taxa nach einer „blastn“-Suche gegen die „refseq“-RNA-Datenbank. Die Graphik spiegelt die Verteilung von 25021 der 38492 Sequenzen wieder, die signifikante „blast“-Treffer erhielten; B Verteilung der Sequenzlängen n=38492

mittleren Länge von 340 bp (Abb. 30) generiert. Durch erste BLAST-Analysen konnte eine Kontamination mit pflanzlichen Sequenzen von ungefähr 31 % (Abb. 30 A) nachgewiesen werden. In der die Sequenzierung durchgeführten Institution wurde parallel ein Sequenzierprojekt mit *Vitis vinifera* durchgeführt. Die Kontamination geht mit hoher Wahrscheinlichkeit hierauf zurück.

Im Weiteren wurden die Daten sowohl geclustert als auch ungeclustert analysiert. Für die Assemblierung größerer Datenmengen wurden drei verschiedene Programme eingesetzt. Der geclusterte wie auch der ungeclusterte Datensatz wurde über Hochdurchsatz-Blast-Analysen annotiert und funktionell charakterisiert. Verschiedene unbekannte, stark exprimierte Sequenz-Gruppen wurden näher untersucht und auf eine Relevanz bei der Knochenentwicklung überprüft.

Vergleich von Assembly-Plattformen

Als eine alternative Möglichkeit zur Annotation jeder einzelnen Sequenz wurde eine vorherige Assemblierung der Daten durchgeführt. Außerdem lassen sich so funktionelle Bereiche besser identifizieren. Hierzu wurden verschiedene zur Verfügung stehende Programme verglichen. Als Grundlage für die Assemblierung mit Newbler (454 Life Sciences Corporation, 2.0.00.20), MIRA V2.9.46 (Chevreux et al. 2004), und NGEN2 (DNASTAR Inc., Lasergene®) dienten alle Sequenzen mit einer Mindestlänge von 100 bp. Dies entsprach 31497 von 38492 Sequenzen. Mit optimierten Parametereinstellungen konnten zwischen 88,2 % (NGEN) und 82,2 % (Newbler) der Sequenzen geclustert werden. Dabei konnten mit allen drei Programmen sehr ähnliche N50- und N75- Werte erzielt werden (Tab. 17). Diese

Tab. 17: Vergleich verschiedener Assembler

	Anzahl der Contigs	Anzahl der Sequenzen	prozentualer Anteil	max. Länge [bp]	N50	N75
Newbler	2657	25902	82,2 %	3525	419	359
MIRA	4060	27079	85,9 %	3699	440	362
NGEN	3903	27780	88,2 %	4392	443	366

geben jeweils die durchschnittliche Sequenzlänge aller Konsensus-Sequenzen innerhalb der größten 50 % bzw. 75 % der Contigs wieder und dienen als Qualitätskriterium bei Assemblierungen. Auch bei der Menge der insgesamt geclusterten Daten wurden nur geringe Unterschiede festgestellt. Gravierendere Unterschiede treten lediglich im Vergleich zwischen Newbler und den beiden anderen Assembler-Programmen auf. Bei ungefähr gleichem Datenvolumen ergibt das Clustern mit Hilfe des Newblers nur 2657 Contigs, ca. 1400 weniger als mit den beiden anderen Plattformen. Daraus resultieren wesentlich ungenauere Assemblierungen und eine höhere Fehlerrate. Da auch ein höherer Anteil kleinerer Contigs entstand (Abb. 31), wurde das Programm Newbler für nicht geeignet eingestuft und nicht weiter verwendet.

Qualitativ konnten sowohl mit MIRA als auch mit NGEN gute Ergebnisse erzielt werden. Allerdings waren die Vorarbeiten und die Rechenzeit für MIRA-Assemblierungen etwas umfangreicher. Außerdem konnten mit NGEN ca. 700 Sequenzen mehr geclustert und geringfügig größere Contigs generiert werden (Abb. 31).

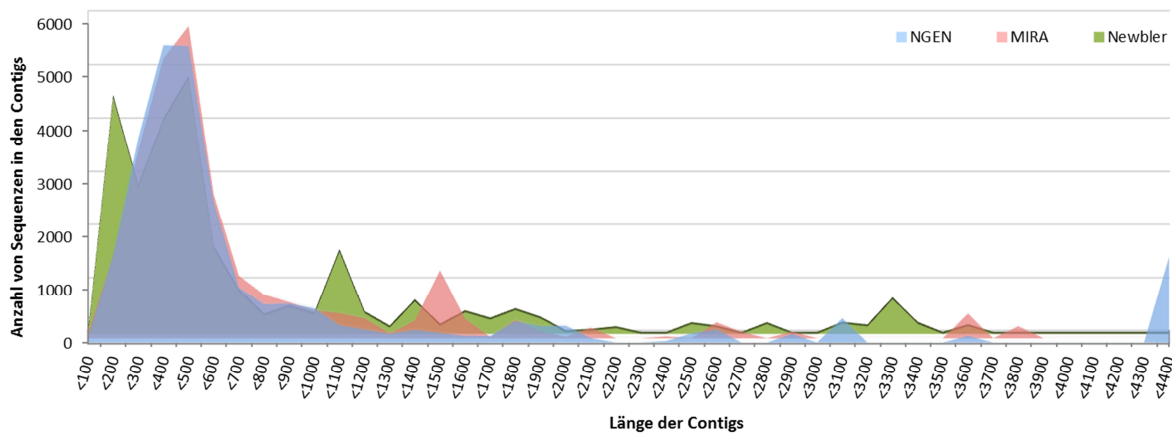


Abb. 31: Verteilung der Contig-Größen
 Dargestellt ist die Größenverteilung nach Assemblierung der 454-Daten mit drei verschiedenen Assemblern; NGEN2 (DNASTAR Lasergene), blau ; MIRA, rot; Newbler (Roche), grün

Annotation und „Gene Ontology“-Analysen mit BLAST2GO (Conesa et al. 2005)

Durch Assemblierung von 38492 Sequenzen mit DNASTAR Lasergene® NGEN2 wurden 3903 Cluster mit mindestens 2 Sequenzen und 4473 Contigs als Einzelsequenzen gebildet. Diese 8376 Contigs enthielten 38150 von 38492 Sequenzen. Bei den verbleibenden 342 Sequenzen handelte es sich um Homopolymerabfolgen oder simple Repeatstrukturen. Die Cluster wurden zunächst automatisiert über die Java Applikation BLAST2GO (Conesa et al. 2005) annotiert und funktionell gruppiert. Dies erfolgte über automatische Blastx-Suchen gegen die Swissprot-Datenbank. Die eigentliche quantitative Analyse erfolgte anschließend sowohl auf Grundlage der Cluster als auch der einzelnen Sequenzen im Hochdurchsatz-Blast gegen Proteindatenbanken bzw. die ref_seq-RNA-Datenbank.

Anteil der Kontamination mit *Vitis vinifera*-Sequenzen

Für die BLAST2GO-Analysen wurden von den 8376 Contigs zunächst 875 (7351 Sequenzen) als Verunreinigungen identifiziert und entfernt. Dies entspricht mit ca. 19,1 % einem geringeren Anteil im Vergleich zur statistischen Auswertung der Blastn-Suchen über MEGAN (s.o.). Von den verbleibenden 4319 Contigs (31141 Sequenzen) konnten 2981 (9085 FLX-Sequenzen) keinem Datenbankeintrag zugeordnet werden. 220 Cluster (930 FLX-Sequenzen) konnten nur unsicher zugeordnet werden. Damit zeigten 28477 „Reads“ (73,9%) eine

Übereinstimmung mit einem Datenbankeintrag. Nach dem Entfernen eines Großteils der Kontaminationen verblieben 21126 Sequenzen (54,8%) in der Analyse.

Funktionelle Charakterisierung des kompletten identifizierbaren Datensatzes

Die BLAST2GO-Plattform bietet mehrere Möglichkeiten, die funktionell charakterisierten Sequenzen in funktionelle Gruppen zusammen zu fassen. Die nachfolgende Abbildung gibt einen Überblick über die Anteile der funktionellen Gruppen an den Gesamtsequenzen (ohne

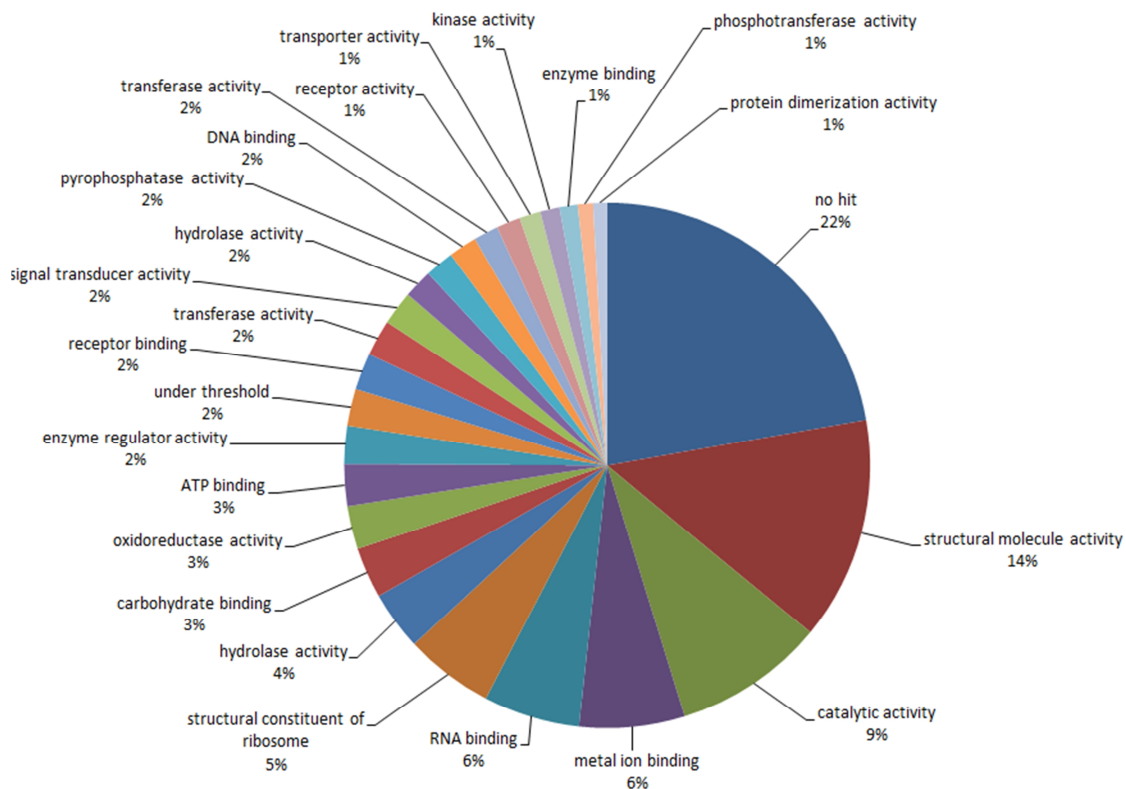


Abb. 32: Funktionelle Gruppierung der 454-Sequenzen
 Ergebnisse der Blastx-Datenbanksuchen gegen die Nr-Proteindatenbank (NCBI) GO-annotiert und funktionell geclustert mittels Blast2GO

Kontamination). Bei der nachbearbeiteten automatischen Annotation und funktionellen Gruppierung verbleibt ein kleiner redundanter Teil. Neben den 22 % nicht charakterisierten Sequenzen (fehlende Übereinstimmung mit Datenbankeinträgen) entfallen besonders viele Sequenzen auf strukturelle Komponenten, katalytische Enzyme, Ionen- und RNA-bindende Proteine. Auch ribosomenassoziierte Proteine (~5 %) haben einen nicht unwesentlichen Anteil. Alle anderen Gruppen haben nur Anteile < 5 %.

Auch hier ist bei der näheren Betrachtung der 100 höchst repräsentierten Gene (hier nicht gezeigt) eine deutliche Verlagerung in Richtung ribosomaler Proteine, Ca²⁺-Stoffwechsel und ECM- Organisation zu beobachten.

Sequenzannotation und quantitative Expressionsanalyse

Meist sind frei zugängliche automatisierte Annotationsverfahren und Analyseprogramme (Bsp. BLAST2GO) nur für die Analyse einer relativ geringen Anzahl von Sequenzen geeignet. Um die Zahl der Sequenzen zu reduzieren ist daher ein vorgeschaltetes Clustern in den meisten Fällen unumgänglich. Um dennoch eine Analyse auf Basis der Einzelsequenzen durchführen zu können, wurde in einer Kombination aus dem frei zugänglichen lokalen Blast-Algorithmus „blastall“ (NCBI), Rechencluster-Parallelisierung und angepassten Perl- und Python-Skripten eine Möglichkeit zum Hochdurchsatz-Blast und anschließender quantitativer Analyse der Einzelsequenzen geschaffen.

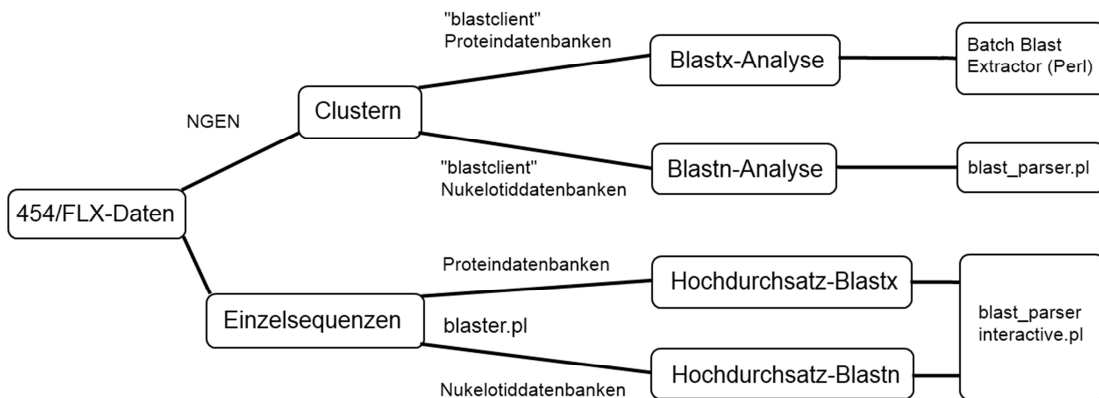


Abb. 33: Ablauf der Blast-Auswertungen der Roche-454/FLX-Daten
 Dargestellt ist der schematische Ablauf der Blast-Analysen der 454-FLX-Daten.

Blast-Analyse der 454/FLX-Daten

Bei den Blast-Suchen mit den geclusterten Sequenzen konnten durch Vergleich mit der Proteindatenbank des NCBI (blastx, nr) 19987 Sequenzen (52 %) funktionell zugeordnet werden. Diese enthielten keine größere Mengen an Kontaminationen mehr und mindestens 40 % Übereinstimmung bzw. einen E-value kleiner $1 * e^{-005}$.

Analog konnten durch Blastn-Suchen gegen die refseq-RNA-Datenbank zu 18679 Sequenzen (48,5%) Übereinstimmungen mit einem E-value besser als $1 * e^{-010}$ gefunden werden. Unter gleichen Voraussetzungen konnten beim Vergleich (blastn) der Einzelsequenzen mit der refseq-RNA-Datenbank 18807 Sequenzen (48,8%) zugeordnet werden.

Für die weitere Auswertung wurden die Datensätze in Excel überführt, redundante Treffer weitgehend zusammengefasst und einheitlich benannt (offizielles Gen-Symbol). Obwohl die blastx-Suchen tendenziell etwas mehr Treffer ergaben, wurde mit den blastn-Suchen weiter gearbeitet. Der Grund ist folgender: Obwohl mehr Übereinstimmungen gefunden wurden, konnte bei Stichproben eine höhere Rate falsch positiver Zuordnungen festgestellt werden. Die Konvertierung der verschiedenen „Accession“-Nummern der Proteindatenbanken auf einheitliche Genbezeichnungen war wesentlich problematischer. Dadurch führte die quantitative Analyse der Blastx-Suchen zu einer höheren Redundanz. (s.e. Anhang: 454_Projekt.xlsx)

Hoch exprimierte Gene und Vergleich mit der EST-Sequenzierung

Die Sequenzcluster wurden schließlich entsprechend ihrer Häufigkeit („read counts“) sortiert (Tab. 18). Wie erwartet konnten auch mittels 454-Sequenzierung viele hoch exprimierte knochen-/knorpelspezifische Gene identifiziert werden. Unter den 25 am häufigsten exprimierten Sequenzen finden sich 15 Gene wieder, die auch bei der Auswertung der ESTs zu den höchst exprimierten Genen gehören (Tab. 18 A). Darüber hinaus ist auch ein Großteil der verbleibenden 10 Gene im EST-Projekt als „stark exprimiert“ eingestuft worden.

Größere Unterschiede zwischen dem EST-Ansatz und der Roche-454/FLX-Sequenzierung lassen sich bei Aggrecan (*ACAN*) und „Lysyl oxidase-like 4“ (*LOXL4*) feststellen. Diese sind im EST-Projekt unterrepräsentiert und jeweils nur mit 3 Sequenzen vertreten. Noch extremer zeigte sich der Unterschied hinsichtlich LOC618319. Hierfür konnten nur drei 454-Sequenzen gefunden werden, die darüber hinaus nur das verlängerte 3'-Ende der Variante Typ 3 repräsentieren.

Der Vergleich zwischen der Analyse mit geclusterten Daten einerseits und Einzelsequenzen andererseits zeigte nur geringfügige Unterschiede. So zeigten weniger als 10 % (101) der 1137 verglichenen Gene mehr als 10 % Abweichung zwischen den Ergebnissen („read

counts“) aus der Analyse geclusterter Daten und denen der Einzelsequenzen. Dabei weisen die Ergebnisse der Einzelsequenzanalyse fast immer weniger „read counts“ im Vergleich zur Contig-Analyse auf. Eine Korrelation zu den theoretischen Transkriptgrößen konnte nicht festgestellt werden.

Normalisierung auf „reads per kilobase transkript million reads“ (RPKM)

Viele Verfahren generieren aufgrund der Fragmentierung und der kürzeren Sequenzlängen mit steigender Fragmentlänge theoretisch auch eine größere Zahl an Sequenzen. Dies führt zu einem systematischen Fehler, der durch die mittlerweile weit verbreitete Normalisierung auf „reads per Kilobase Transkript per million (mappable) reads“ (RPKM) versucht wird zu kompensieren. Hierdurch kann sich das Ranking von Genen nach ihrer Expressionsstärke im Vergleich zu den absoluten „read counts“ verschieben (Tab. 18 B). Nach Auswertung entsprechender RPKM-Berechnung stimmen nur noch 11 der 25 stark exprimierten Gene mit denen aus dem EST-Projekt überein. Die Verschiebung, die sich aufgrund der Berechnungsweise ergibt, ist insgesamt jedoch eher marginal und verändert das Gesamtbild nicht wesentlich.

Tab. 18: Höchst exprimierte Gene durch 454-Sequenzierung

Aufgelistet sind die 25 höchst repräsentierten Gene nach blastn-Hochdurchsatzanalyse; A Absolute Sequenzanzahl im Vergleich zu geclusterten Sequenzen mit geblasteten Einzelsequenzen, die Reihenfolge ergibt sich aus den gemittelten „read counts“ aus Blastx- und Blastn-suchen B Verändertes Ranking nach Normalisierung auf RPKM;

grau hinterlegt sind die auch schon im EST-Projekt unter den Top25 vorkommende Gene, rot markiert sind besonders große Abweichungen zwischen absoluten Werten und RPKM

A	Gen	Sequenzen Cluster	Sequenzen Reads	B	Gen	Länge	RPKM Cluster	RPKM Reads
	<i>COL2A1</i>	1861	1877		<i>COL2A1</i>	5109	9463	9545
	<i>COL10A1</i>	519	523		<i>RPLP1</i>	342	5393	5014
	<i>FN1</i>	364	369		<i>COL10A1</i>	3049	4422	4456
	<i>COL11A1</i>	253	255		<i>MGP</i>	573	4262	4217
	<i>COMP</i>	225	227		<i>SPP1</i>	1331	4216	4236
	<i>SPP1</i>	216	217		<i>RPS15a</i>	468	4219	4108
	<i>COL9A2</i>	198	204		<i>RPS20</i>	477	3595	3540
	<i>COL11A2</i>	219	167		<i>RPS18</i>	541	3458	3458
	<i>ACAN</i>	175	192		<i>CTGF</i>	1047	3499	3052
	<i>LOXL4</i>	157	134		<i>FTH1</i>	925	3314	3174
	<i>CTGF</i>	141	123		<i>RPL35A</i>	448	3131	3131

Gen	Sequenzen Cluster	Sequenzen Reads	Gen	Länge	RPKM Cluster	RPKM Reads
<i>COL9A3</i>	135	127	<i>RPS19</i>	509	3011	2909
<i>EEF1A1</i>	147	112	<i>RPS15</i>	489	2922	2869
<i>HAPLN1</i>	132	109	<i>RPS12</i>	464	2800	2800
<i>FTH1</i>	118	113	<i>RPS24</i>	539	2796	2796
<i>SPARC</i>	114	102	<i>RPL41</i>	641	3729	1581
<i>COL9A1</i>	95	117	<i>COMP</i>	2461	2375	2396
<i>IGF2</i>	104	101	<i>RPL35</i>	929	2265	2237
<i>MGP</i>	94	93	<i>LOC786430</i>	557	2192	2192
<i>LECT1</i>	89	87	<i>RPL32</i>	506	2208	2002
<i>RPL35</i>	81	80	<i>TPT1</i>	516	2719	1359
<i>COL6A1</i>	82	78	<i>RPL30</i>	467	2058	1947
<i>LOC100300443</i>	142	18	<i>RPS3A</i>	865	1982	1982
<i>TIMP3</i>	80	79	<i>LECT1</i>	1167	1981	1937
<i>RPS15a</i>	76	74	<i>RPL24</i>	551	1933	1933

Hoch exprimierte uncharakterisierte Sequenzen

Für weitere Betrachtungen wurden aus der Liste der 100 am stärksten exprimierten Sequenzen solche ausgewählt, die mit Blast-Treffern zwar erfasst, aber nicht eindeutig zugeordnet werden konnten und damit Gene unbekannter Funktion darstellen könnten. Von den acht zeigen drei große Sequenzhomologien zu 28s ribosomaler RNA (LOC100300443, LOC100337158, LOC100471656). Im Genom von *Bos taurus* sind diese Sequenzen offenbar repetitiv und auf vielen Chromosomen verteilt vorhanden.

Vorherige Blast-Annotation	Ergebnisse der weiteren Analysen
LOC100300443	Ähnlichkeit zu Sequenzen ribosomaler RNA (28s)
LOC100337158	Ähnlichkeit zu Sequenzen ribosomaler RNA (28s)
LOC100471656	Ähnlichkeit zu Sequenzen ribosomaler RNA (28s)
LOC448984	eindeutig rRNA
LOC781186	ribosomales Protein S15a
LOC100293593	humane mitochondriale Cytochrome C Oxidase Untereinheit 2
CpipJ_CPIJ012185	keine genomische Entsprechung /Vektorübereinstimmungen
MGC165939	Sehr geringe Übereinstimmungen zu hypothetischen Proteinen / Vektorübereinstimmungen

Die Sequenz LOC448984 ist dagegen definitiv homolog zu 28s rRNA. Die Sequenzen von LOC781186 und LOC100293593 konnten bekannten Genen zugeordnet werden und CpipJ_CPIJ012185 bzw. MGC165939 zeigen große Homologien zu Vektorsequenzen.

Blast2GO-Analyse und nicht identifizierte Sequenz-Cluster

Alternativ zur manuellen Annotation wurde für die Auswertung der geclusterten Daten das Java-Programm Blast2GO eingesetzt. Dabei wurden die 8376 Konsensus-Sequenzen aus der Assemblierung mit NGEN zunächst mit Hilfe automatisierter Blastx-Suchen gegen die Proteindatenbanken (nr) untersucht. Interessant für die Suchen nach Genen mit bisher unbekanntem Bezug zu Knochen- oder Knorpelentwicklung waren die 50 höchst repräsentierten nicht identifizierten Cluster. Für die generelle Suche nach bisher unbekanntem Genen wurden die 62 längsten Konsensus-Sequenzen (>500 bp), ohne Übereinstimmung bei der Blastx-Suche, näher betrachtet.

Durch manuelle Blastn-Suchen wurde der überwiegende Teil der 50 höchstrepräsentierten unbekanntem Contigs als pflanzliche Sequenzen identifiziert. Diese müssen daher der *Vitis*-Kontamination zugeordnet werden. Lediglich vier Konsensus-Cluster-Sequenzen zeigten Homologien zu genomischen Loci des Rindergenoms. Trotz fehlender Blastx-Übereinstimmungen wurden alle vier Sequenzen im direkten genomischen Vergleich als bekannte Gene identifiziert.

Bei den 62 größten Contigs ergaben sich fast ausschließlich Übereinstimmungen zu bovinen Sequenzen. Obwohl eine Vielzahl dieser Contigs auch über Blastn-Suchen nicht zugeordnet werden konnte, machte der Vergleich mit den genomischen Sequenzen in vielen Fällen eine eindeutige Zuordnung zu UTR-Bereichen bekannter Gene möglich. Wiederum vier Contigs zeigten nur einzelne genomische Übereinstimmungen, die keine Genregion betrafen oder nicht mit Exons überlappten.

Insgesamt gestaltete sich die Analyse möglicher unbekannter Gene mit Hilfe der 454/FLX-Sequenzierung als schwierig. Dazu trägt nicht unwesentlich die Kontamination mit *Vitis*-Sequenzdaten bei. Diese konnten im Vorfeld der Analysen nur unvollständig entfernt

werden und tauchten gerade innerhalb der schlecht identifizierbaren Sequenzen gehäuft auf.

3.4 Genexpressionsanalysen mittels Illumina-Sequenzierung

Für die Hochdurchsatz-DNA-Sequenzierung mit Hilfe der Illumina Genome Analyzer II-Technologie (GAIIx) wurde die PolyA⁺-RNA aus Gesamt-RNA der fötalen Wachstumstufe des Rindes zunächst angereichert, und anschließend 100 ng mittels linearer Amplifikation vermehrt. Die verwendete Menge liegt damit um das zehnfache über der minimalen Menge, 10 ng Ausgangs-RNA, bei der konsistente Amplifikationen möglich sind (Sengupta et al. 2010). Es kann davon ausgegangen werden, dass der dadurch eingebrachte Fehler minimal

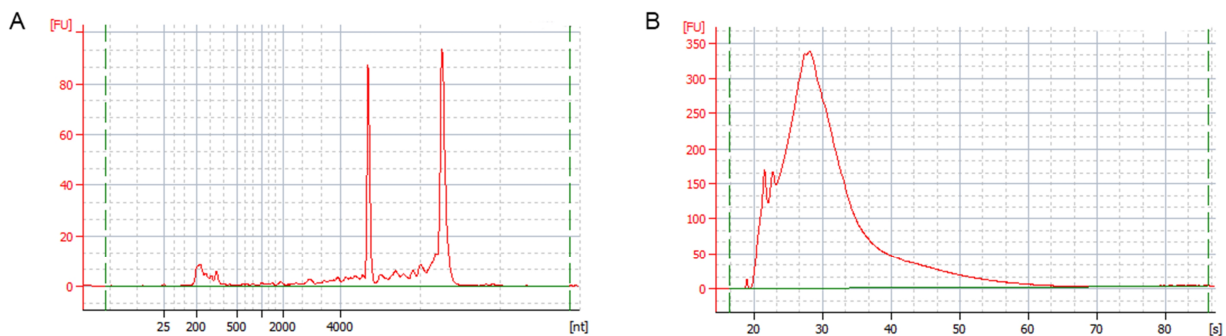


Abb. 34: Fötale Wachstumstufe - Gesamt-RNA und amplifizierte polyA⁺-RNA
 A Bioanalyzer-Fluorogramme von Gesamt-RNA und B linear amplifizierter polyA⁺-RNA (1:10) der bovinen fötalen Wachstumstufe RIN ges.RNA = 7,1

ist. Die linear amplifizierte RNA zeigt erwartungsgemäß eine völlig andere Größenverteilung als die Ausgangs-RNA. Von der linearamplifizierten RNA (Abb. 34) wurden 5 µg im „Cologne Center for Genomics“ (Köln) für das Erstellen einer entsprechenden cDNA-Bibliothek verwendet und mittels eines Illumina Genome Analyzer IIx sequenziert. Dabei wurden „paired end“-Sequenzen von jeweils 100 nt Länge generiert. Die Gesamtzahl der Sequenz-„Reads“ betrug 75.484.617. Diese wurden vor der weiteren Analyse mit verschiedenen Methoden aufgearbeitet und gefiltert.

Qualitätskontrolle und Qualitätsfilter

Vor den Blast-Analysen wurden anfänglich die Daten unter Verwendung des Illumina internen „Chastity“-Filters und einer maximalen Anzahl von 3 nicht eindeutigen Basen („N“) gefiltert. Dabei wurden rund 10 Mio. Sequenzen (10.485.259) entfernt. Dadurch konnte die

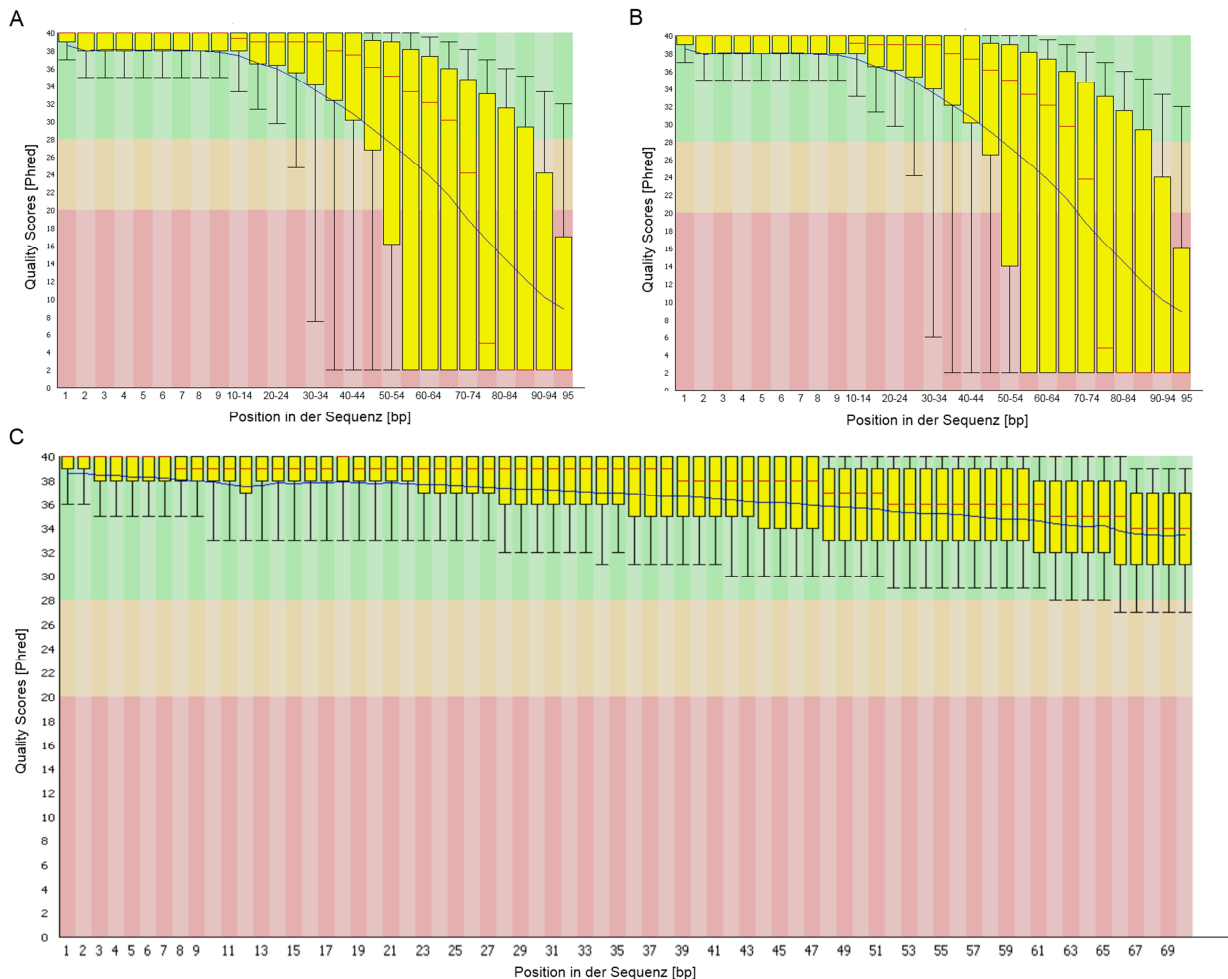


Abb. 35: Qualitätskontrolle und Trimmen der Illumina-Daten

Qualitätsdiagramme der Illumina-Daten mittels fastqQC (www.bioinformatics.bbsrc.ac.uk), mit grün ist der Bereich mit sehr guter Qualität und Phred-Scores über 28 gekennzeichnet, gelb der Bereich mittlerer Qualität (Phred 20 – 28) rot Nukleotidpositionen mit durchschnittlich schlechteren Phred-Werten als 20

A Rohdaten mit einer Länge von 100 nt; B gefilterte Daten (Perl) mit 100 nt Länge, Kriterien: „Chastity“-Filter und keine Sequenzen mit mehr als drei N (nicht eindeutige Basen); C stringentes Trimmen: grundsätzlich wurden die ersten 5 und die letzten 25 Basen entfernt; parallel wurden Teile der Sequenzen unter einem Phred-Wert von 20 entfernt

Fehlerrate jedoch nur geringfügig verbessert werden. Der Vergleich der Qualitätsdiagramme von den Rohdaten (Abb. 35 A) mit den verbliebenen ca. 65 Mio. Sequenzen (Abb. 35 B) zeigt nahezu identische Verteilungen der Qualitätswerte. Für Blast-Suchen sind kleinere Fehler nicht relevant. Für genaues Mappen und Clustern sind auch kleinere Fehler sehr störend.

Deshalb wurden die Filterprozeduren modifiziert. Betrachtungen der einzelnen Sequenzen zeigten schlechtere durchschnittliche Qualitätswerte in den ersten 5 und den letzten 25 Positionen. Diese wurden aus den Sequenzen entfernt. Die übrigen 70 Nukleotidpositionen wurden dynamisch nach einem Phred-Wert über 30 zugeschnitten und durften maximal eine unidentifizierte Base („N“) enthalten. Durch diese Maßnahmen konnte die Fehlerrate der verwendeten Sequenzen erheblich reduziert werden (Abb. 35 C).

„Read Count“-Ansatz mittels Hochdurchsatz-Blast-Suchen

Für die Auswertung von exprimierten Sequenzen auf Basis ihrer Häufigkeit wird, in Anlehnung an klassische „Tag Count“-Ansätze der EST- oder SAGE –Sequenzierungen, in NGS-Projekten oft der sogenannte „Read Count“-Ansatz verwendet.

Vor der quantitativen Auswertung war es nötig, bekannte repetitive Sequenzen und rRNA-Verunreinigungen zu entfernen. Die qualitätsgefilterten Daten wurden wiederum durch „Mapping“ (PASS/Bowtie) gegen die extrahierten (*.gbk-Datei) repetitiven Sequenzen des Rinder-Genoms von diesen befreit. Danach verblieben noch 53.483.642 Sequenzen, die zur Annotation und quantitativen Auswertung mit Hilfe parallelisierter Blastn-Suchen gegen die „Refseq RNA“-Datenbank eingesetzt wurden. Alle Treffer mit jeweils gleichen Datenbankeinträgen wurden aufsummiert und in Gruppen zusammengefasst. Insgesamt konnten 39.283.917 (73,5 %) in 44714 Gruppen eingeordnet werden.

Bei der quantitativen Auswertung wurden sowohl die nicht normalisierten Sequenzhäufigkeiten als auch die auf RPKM normalisierte Zahl der Sequenzen bestimmt (Tab. 19). Unter den 25 am häufigsten gefundenen Transkripten sind deutliche Unterschiede zwischen den unnormalisierten und den RPKM-normalisierten Daten erkennbar. Durch diese Berechnungsweise fallen viele der typischen Knochen- und Knorpel-spezifisch exprimierten Gene aus der Riege der höchst exprimierten Gene heraus. Die ribosomalen Proteine treten, auch im Vergleich zur 454-Sequenzierung, noch stärker hervor. Dies zeigt auch die funktionelle Charakterisierung über DAVID (NIH) der 100 höchst exprimierten Gene beider Auswertungen. So steigt der Anteil ribosomaler und solcher an der Translation beteiligter Proteine von 50 % auf 65 % nach Normalisierung an. Auffällig ist der

Rang von SAA3, welches im Zusammenhang mit Kollagenasen auftritt und möglicherweise ein Substrat zum Abbau durch Metalloproteinasen (Mitchell et al. 1991; Mitchell et al. 1993) darstellt.

Tab. 19: Höchst exprimierte Gene nach Illumina-RNA-Sequenzierung
 A absolut quantifiziert; aufgelistet sind die 25 Gene mit den meisten Treffern; B auf RPKM normalisierte Expressionsstärke: aufgelistet sind wiederum die 25 höchstexprimierten Gene; farblich hervorgehoben sind die knochen-/knorpelspezifisch exprimierten Gene

A	GEN	Sequenzen	B	GEN	Exonlänge	RPKM
	<i>COL2A1</i>	1842121		<i>RPLP1</i>	342	21870
	<i>COL10A1</i>	1115403		<i>SPP1</i>	1331	18380
	<i>SPP1</i>	961024		<i>MGP</i>	573	16748
	<i>HAPLN1</i>	761019		<i>RPL31</i>	433	13152
	<i>COL11A1</i>	616748		<i>RPS21</i>	353	12566
	<i>COL9A1</i>	414067		<i>RPS12</i>	464	11756
	<i>FN1</i>	388086		<i>RPS17</i>	493	11690
	<i>MGP</i>	376993		<i>RPL11</i>	604	11319
	<i>FTH1</i>	309433		<i>RPS15</i>	489	10563
	<i>RPLP1</i>	293828		<i>RPS8</i>	690	10526
	<i>RPS8</i>	285319		<i>RPS20</i>	477	10510
	<i>RPL11</i>	268584		<i>HAPLN1</i>	1939	9991
	<i>CTGF</i>	263038		<i>SAA3</i>	546	9716
	<i>RPS2</i>	253158		<i>COL10A1</i>	3049	9312
	<i>SPARC</i>	242182		<i>RPS10</i>	583	9200
	<i>COL9A3</i>	232382		<i>RPS24</i>	539	9180
	<i>RPSA</i>	226439		<i>COL2A1</i>	5109	9178
	<i>RPS17</i>	226396		<i>RPL23</i>	495	9059
	<i>RPL31</i>	223720		<i>FTH1</i>	925	8515
	<i>RPL4</i>	216861		<i>RPS27A</i>	468	8212
	<i>RPS12</i>	214282		<i>RPL13A</i>	663	8027
	<i>RPS10</i>	210697		<i>RPS7</i>	582	7937
	<i>RPL13A</i>	209075		<i>RPL30</i>	467	7433
	<i>SAA3</i>	208401		<i>RPS2</i>	940	6856
	<i>RPS15</i>	202910		<i>RPL19</i>	309	6663

Im EST-Projekt nur durch eine einzige Sequenz repräsentiert, konnte SAA3 in der 454 Sequenzierung 21-mal nachgewiesen werden und steht damit an Position 102.

Charakterisierung unbekannter Sequenzcluster

Unter den 300 höchst repräsentierten Sequenzgruppen konnten 22 nur mit besonderen Verfahren zugeordnet werden. Für 14 Loci konnten funktionelle Zuordnungen getroffen werden indem homologe Loci anderer Spezies zur Klassifizierung herangezogen wurden. Auffällig war LOC100438158, der eindeutig dem ribosomalen Protein RPL41 zugeordnet werden konnte. RPL41 mit seiner Vielzahl an Pseudogenen erscheint auch im EST- und im 454-Projekt hoch exprimiert. Ähnlich verhält es sich mit LOC10015759, das CCDC80 entspricht. Ebenfalls interessant ist ein Fusionsprotein aus FIBU Domäne und dem ribosomalen Protein S30 (Michiels et al. 1993). Es gibt Hinweise, dass dieses Protein mit Ubiquitin-ähnlichen Anteilen als Tumorsuppressor wirkt und an der Apoptosekontrolle beteiligt ist (Pickard et al. 2010; Pickard et al. 2011; Rossman et al. 2003). Neben der Strukturkomponente Tubulin alpha 1B finden sich mit Testin und Thymosin beta 4 weitere an Zelladhäsion und Zytoskelett beteiligten Proteine.

Tab. 20: Manuelle Annotation der unbekanntenen Treffer der Top 300

Locus	Gen	Beschreibung
LOC100438158	<i>RPL41</i>	Ribosomal protein L41
LOC100295976	<i>TBA1B</i>	Tubulin alpha 1B
LOC100300995		Cytochrome c oxidase subunit III-like
LOC785455	<i>TMSB4X</i>	Thymosin beta 4, X-linked
LOC100360647	<i>FAU</i>	Ubiquitin-like protein fubi and ribosomal protein S30-like
LOC100469688	<i>PCBP2</i>	Poly(rC)-binding protein 2-like
LOC100295449	<i>DSS1</i>	26S proteasome complex subunit
LOC100335186	<i>TCEAL8</i>	Transcription elongation factor A protein-like 8
LOC789240	<i>TES</i>	Testin
LOC100297716	<i>TMSB4X</i>	Thymosin beta 4, X-linked
LOC281370	<i>UBB</i>	Poly-ubiquitin-B Ubiquitin
LOC100296593	<i>UBB, LOC281370</i>	Poly-ubiquitin-B Ubiquitin
LOC788696	<i>RAN</i>	GTP-binding nuclear protein Ran
LOC100157590	<i>CCDC80</i>	CCDC80 / similar to MGC159613 protein

Die restlichen acht Cluster konnten wiederum keiner eindeutigen Funktion zugeordnet werden. Unter diesen 8 Loci befinden sich unter anderem zwei bereits im EST-Projekt auffällige Loci, LOC100335754 und LOC618319. LOC100335754 wurde sowohl im EST- als

auch im Illumina-Projekt im Bereich der Positionen 110-120 der höchst repräsentierten Gene gefunden.

Tab. 21: Hoch exprimierte uncharakterisierte Genloci aus dem Illumina-Projekt

Locus	Übereinstimmung	Beschreibung
MGC127538	LOC100137875	MGC127538 Protein
LOC100335754	3 ESTs - E12P44	
LOC100467971		Bos taurus Chr 2; <i>SERF3</i> ; <i>MIR1282</i> ; <i>SERINC4</i>
LOC100335527		<i>MEG3</i> non protein coding
LOC100469233		Bos taurus Chr 5; <i>ZC3H10</i>
C12H13orf15		hypothetical protein LOC614348
LOC618319	23 ESTs	<i>ASCL830</i>
LOC787568		Uncharacterized protein / Tubulin alpha-1C chain

Genomische Mappings mit Hilfe von PASS (Campagna et al. 2009)

Aufgrund der anfänglich fehlenden Möglichkeiten eine große Anzahl von kurzen Sequenzen mit einer Referenz zu assemblieren, wurde nach alternativen Analyseverfahren für die Untersuchung von interessanten Sequenzbereichen gesucht. Dabei wurde auf effizientere und „schlankere“ Algorithmen zurückgegriffen. Mit Hilfe der „Mapping“-Programme Bowtie (Langmead et al. 2009) und PASS (Campagna et al. 2009) lassen sich NGS-Sequenzen den übereinstimmenden Koordinaten einer Referenzsequenz zuordnen. Allerdings wird diese Information nicht in Alignment-Dateien gespeichert sondern in Tabellenform (GFF/GFF3).

Für die Untersuchung der unbekanntem Loci aus der Blast-Analyse der Illumina-Daten wurde eigens ein Perl-Skript programmiert, das die „Mapping“-Informationen des Programms PASS ausliest und die „Read“-Abdeckung jeder einzelnen Basenposition errechnet (Anhang: `pass_extract_multiple_genomic_regions_spliced_0_9`). Diese in einer Text-Datei abgelegten Informationen können mittels einfachster Plot-Programme oder Tabellenkalkulationen graphisch dargestellt werden (Abb. 36). Dabei werden sowohl „Reads“ mit zusammenhängenden Treffern („default-Mapping“) als auch potentielle Spleiß-Stellen-überspannende „Reads“ („spliced-Mapping“) in die Analyse mit einbezogen. Die Methode eignet sich daher in erster Linie zum Nachweis von Exon-/Introngrenzen und der Bestimmung der Abdeckung einzelner Exons. Dadurch bietet sich auch die Möglichkeit zum Vergleich verschiedener Datensätze und der Detektion differenzieller Expression von ganzen

Transkripten, oder Transkriptvarianten. Darüber hinaus können die erfolgreich gemappten Sequenzen einer Region leicht aus der Gesamtmenge der Daten herausgezogen werden und sind damit auch für konventionelle Assemblierungsprogramme handhabbar.

Wie aus dem Histogramm der Abb. 36 deutlich hervorgeht, konnte in der Region des untersuchten LOC100296593 lediglich die Expression von LOC281370, einem Polyubiquitin

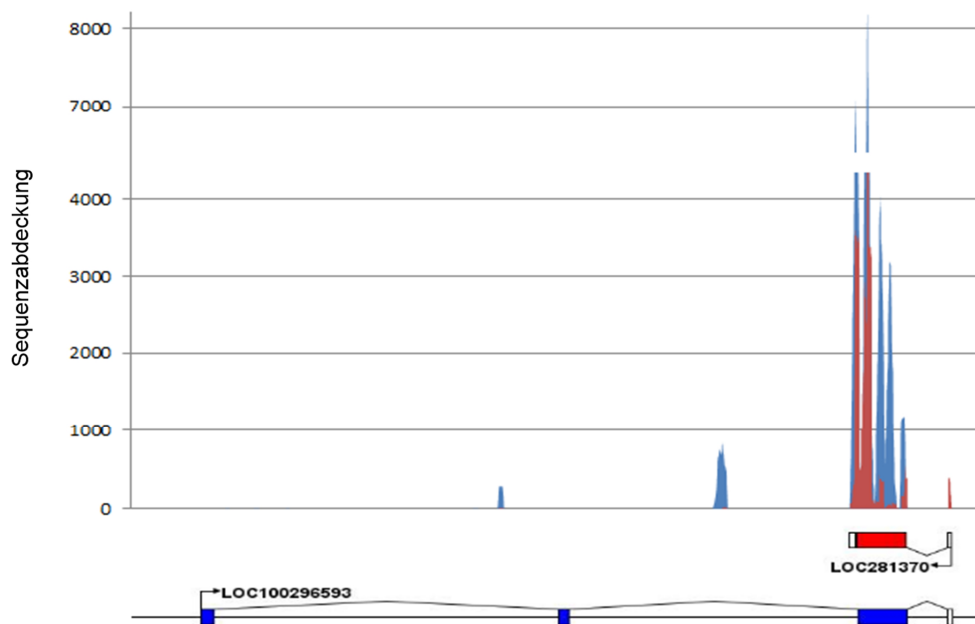


Abb. 36: Mapping - LOC100296593 / LOC281370

Das Diagramm zeigt die Abdeckung der genomischen Region von LOC100296593 und LOC281370 mit Sequenzen aus der Illumina-Sequenzierung; rote Peaks stellen mit Hilfe der „spliced“-Option gemappte „Reads“ dar (Abdeckung 10fach verstärkt), blau dagegen sind zusammenhängende Treffer unterhalb der x-Achse bekannte, in der Datenbank zu findende Transkript-Annotationen dargestellt; rot bzw. blau sind mögliche offene Leserahmen markiert

nachgewiesen werden. Des Weiteren sind Peaks weiter 5' gelegen zu erkennen (blau). Diese werden nicht von potentiell Exon/Exon-überspannenden Sequenzen (rot) unterstützt.

In der Region des Rinder-Chromosoms 21, die den *LOC100467179* enthält (Abb. 37), sind zwei Transkripte mit Exon-/Intronübergängen nachgewiesen worden. Bei einem handelt es sich mit *SERF2* offensichtlich um ein Homologes zu *LOC100467179* des Pandas. Deutlich sind die hoch abgedeckten Exon-/Introngrenzen zu erkennen. Allerdings zeigt schon das nächste 5' gelegene Exon eine nur sehr geringe Zahl an Transkripten. Nicht annotierte

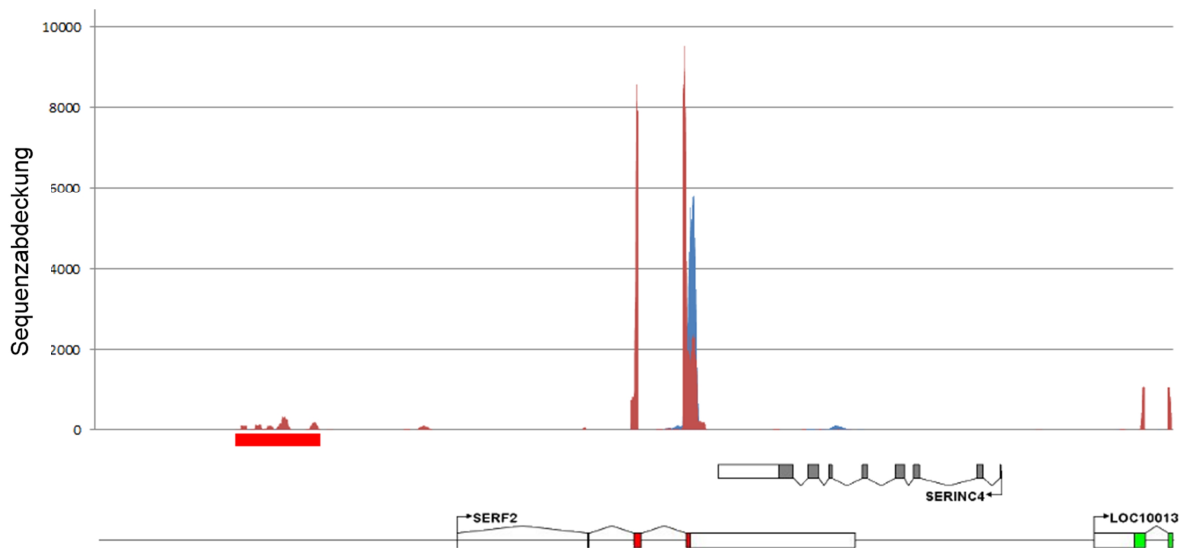


Abb. 37: Mapping - LOC100467179 - SERF2

Dargestellt ist die „Read“-Abdeckung der genomischen Region von LOC100467179 mit Sequenzen aus der Illumina-Sequenzierung; rote Peaks stellen mit Hilfe der „spliced“-Option gemappte „Reads“ dar (Abdeckung 10fach verstärkt), blau dagegen sind zusammenhängende Treffer; Unterhalb der x-Achse bekannte, in der Datenbank zu findende Transkript-Annotationen dargestellt; rot bzw. grau und grün sind mögliche offene Leserahmen markiert

neue Exonbereiche sind nicht zu vermuten. Weiter 5' von SERF2 gelegen, zeigen einige Reads unzusammenhängende Treffer. In dieser Region konnten kleinere Tandemrepeats und eine Übereinstimmung zu diversen reversen Transkriptasen bzw. LINE-Elementen festgestellt werden. Auch die geclusterten Daten unterstützen lediglich Exon 2, das folgende Intron, dann Exon 3 und 4 mit dem dazwischen liegenden Intron. Deutlich, aber wesentlich weniger stark, sind auch die Exonränder des „Huntingtin interacting protein“ (*LOC100138230*, CDS grün dargestellt) zu erkennen.

Beim „Mappen“ gegen den menschlichen *LOC614348* (*CH12H19orf15*) homologen Region bei *Bos taurus* werden wesentlich mehr Exonbereiche unterstützt. Bisher sind zwei verschiedene Isoformen annotiert (Abb. 38). Deutlich werden die Exongrenzen durch die Illumina-Sequenzen abgedeckt. Die vier höchsten Peaks markieren ausschließlich Exons der ersten Variante. Tendenziell nimmt die Abdeckung von 3' nach 5' ab. Der kleine Peak (rot) im Intronbereich von Variante 1 könnte ein Hinweis auf ein sehr kleines Exon darstellen. Ebenso die Regionen mit blau hervorgehobenen „Peaks“, sie markieren Regionen in denen Sequenzen ununterbrochen über ihre gesamte Länge mit der genomischen Sequenz übereinstimmen. Diese Region wird allerdings nicht durch potentiell

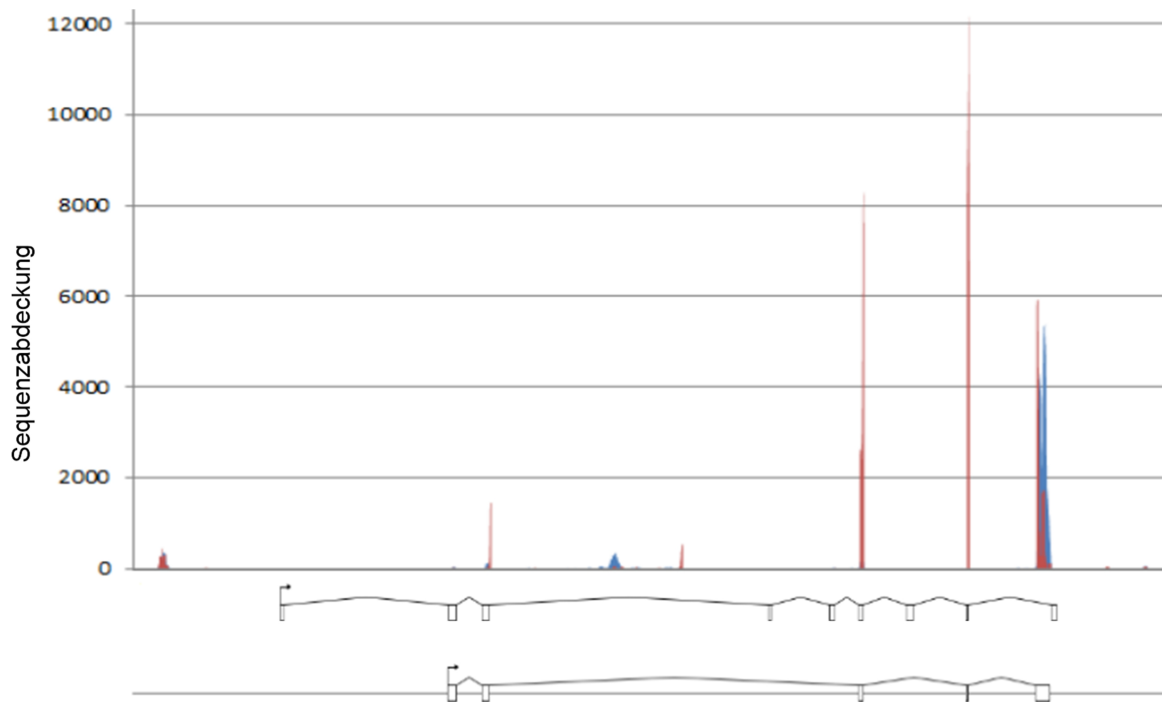


Abb. 38: Mapping LOC614348 - CH12H19orf15 homolog
 Abdeckung der Region um LOC614348 durch die gemapten Illumina-Sequenzen; rote Peaks stellen mit Hilfe der „spliced“-Option gemappte „Reads“ dar (Abdeckung 10fach verstärkt), blau dagegen sind zusammenhängende Treffer; Unterhalb der x-Achse bekannte, in der Datenbank zu findende Transkript-Annotationen dargestellt;

gespleißte Transkripte unterstützt. Es konnte keine längere Sequenz aus den geclusterten Daten identifiziert werden, die Exonbereiche unterstützt, die ausschließlich der zweiten Isoform zugeordnet werden.

Die Gene *LOC100335754* (E12P44) und *LOC618319* wurden bereits im EST-Projekt identifiziert und hinsichtlich ihrer genomischen Struktur analysiert. Assemblierung und Mapping der Illumina Daten bestätigen im Wesentlichen diese Ergebnisse. Im Fall von *LOC618319* werden alle drei bisher annotierten Exons bestätigt. Das vierte, in direkter Nachbarschaft zu *NGEF* gelegene Exon wird mit einer geringeren Abdeckung ebenfalls unterstützt. Neben diesen Exonbereichen, die sich mit den konservierten Regionen decken (Abb. 39; unten, Vergleich mit *Mus musculus*), sind weitere Regionen mit potentiell gespleißten Transkripten (rot) identifiziert worden. Allerdings konnte weder durch die zusammengesetzten Daten noch die EST-Klone oder 454-Sequenzen eine Verbindung zu Transkripten von *LOC618319* geknüpft werden. Auffällig bei diesen Peaks ist eine

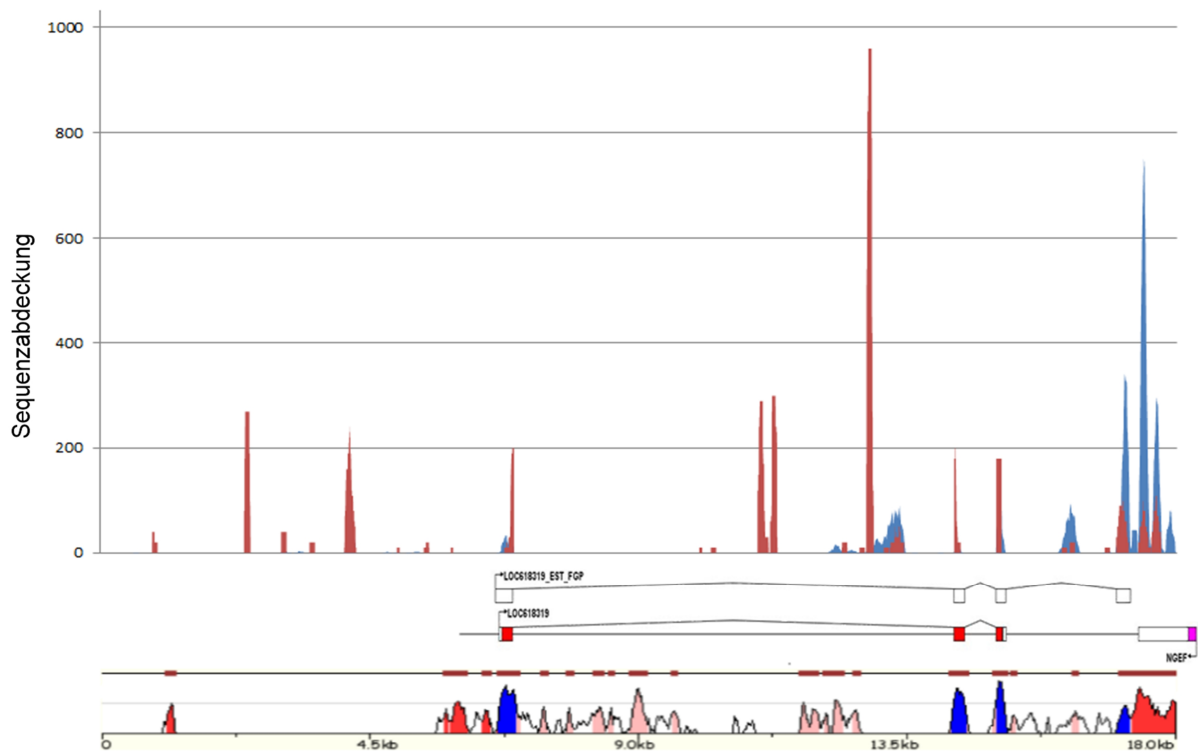


Abb. 39: Mapping - LOC618319

Abdeckung der Region um LOC618319 durch die gemappten Illumina-Sequenzen; rote Peaks stellen mit Hilfe der „spliced“-Option gemappte „Reads“ dar (Abdeckung 10fach verstärkt), blau dagegen sind zusammenhängende Treffer; Unterhalb der x-Achse befindet sich die zugehörige Transkript-Annotationen; rot ist der mögliche offene Leserahmen markiert; zusätzlich ist darunter der Vergleich zur homologen Region im Maus-Genom dargestellt

gute Abdeckung mit „gespleißten“ Sequenzen, bei fehlender bzw. sehr geringer Abdeckung mit durchgehend treffenden Sequenzen. Eine direkte Korrelation mit konservierten Bereichen ist nicht zu erkennen. Die Regionen wurden daher mittels blastx-Suchen mit Proteindatenbanken abgeglichen.

Insgesamt wurden 5 Regionen näher untersucht. Dabei konnten signifikante Übereinstimmungen zu verschiedenen Proteinen identifiziert werden. Im Folgenden werden diese kurz zusammengefasst. Neben den häufig vertretenen *PAOX*, *SHD*, *CCNT1*, *ZNF679* und *SETMAR* finden sich allerdings auch Hinweise auf Übereinstimmungen mit Transposasen oder reverser Transkriptasen. Alle Regionen werden durch viele EST-Sequenzen unterstützt. Wobei lediglich die ersten 150-170 bp von Region 5 potentiell proteincodierend sind.

Tab. 22: Blast-Ergebnisse auffälliger Bereiche - LOC618319

	Start	Ende	Blast-Ergebnisse
1	2415	2524	PAOX; SHD; CCNT1; ZNF679
2	3037	3212	histone-lysine N-methyltransferase SETMAR; transposase-like protein
3	4086	4290	PAOX; SHD; CCNT1; ZNF679
4	10999	11333	GPC5; PP1; reverse transcriptase-like
5	12704	13481	PAOX; SHD; CCNT1; ZNF679

3.5 Vergleich zwischen verschiedenen Analyseverfahren

In einzelnen Fällen wurden schon parallel zur Auswertung der jeweiligen Projekte Vergleiche mit den anderen Methoden angestellt. Dazu wurden die Expressionswerte nach den Blast-Auswertungen der verschiedenen knochen- und knorpelspezifischen Gengruppen, der Kollagene, Proteoglykane und ECM-Proteine, miteinander verglichen.

Obwohl Expressionsniveaus vieler Gene unabhängig von der Analysemethode ähnlich sind, gibt es auch starke Abweichungen. Besonders gut stimmen die Ergebnisse der EST-Sequenzierung und der Illumina-Sequenzierung überein. Weniger gut ist die Übereinstimmung mit den 454/FLX-Sequenzen. Oftmals liegen die auf jeweils 1 mio. Sequenzen normalisierten Werte bei den knochen- und knorpelassoziierten Genen für die 454-Sequenzierung über denen der beiden anderen Ansätze (Abb. 40). Seltener wurden signifikant weniger Treffer verzeichnet.

Bei den Kollagenen sind es vor allem *COL2A1*, *COL6A1* und *COL9A2*, die in der 454-Sequenzierung überproportional stark vertreten sind. Gleiches gilt für *ACAN*, *BGN*, *FMOD* und *GPC1* bei den Proteoglykanen, sowie *FN1* und *COMP*. Lediglich *COL9A1* und *SPP1* zeigen eine entgegengesetzte Tendenz. Beim Vergleich der Illumina-Daten mit den anderen Datensätzen fällt lediglich die äußerst geringe Expressionsstärke von *COL11A2* auf. Diese tritt sowohl bei den auf 1 Million Sequenzen normalisierten Daten als auch beim relativen Anteil an der Gruppe der Kollagene auf. Die Expressionslevel der niedrig exprimierten Gene, wie beispielsweise in der Gruppe der ECM-Proteinen, zeigen oftmals etwas höhere

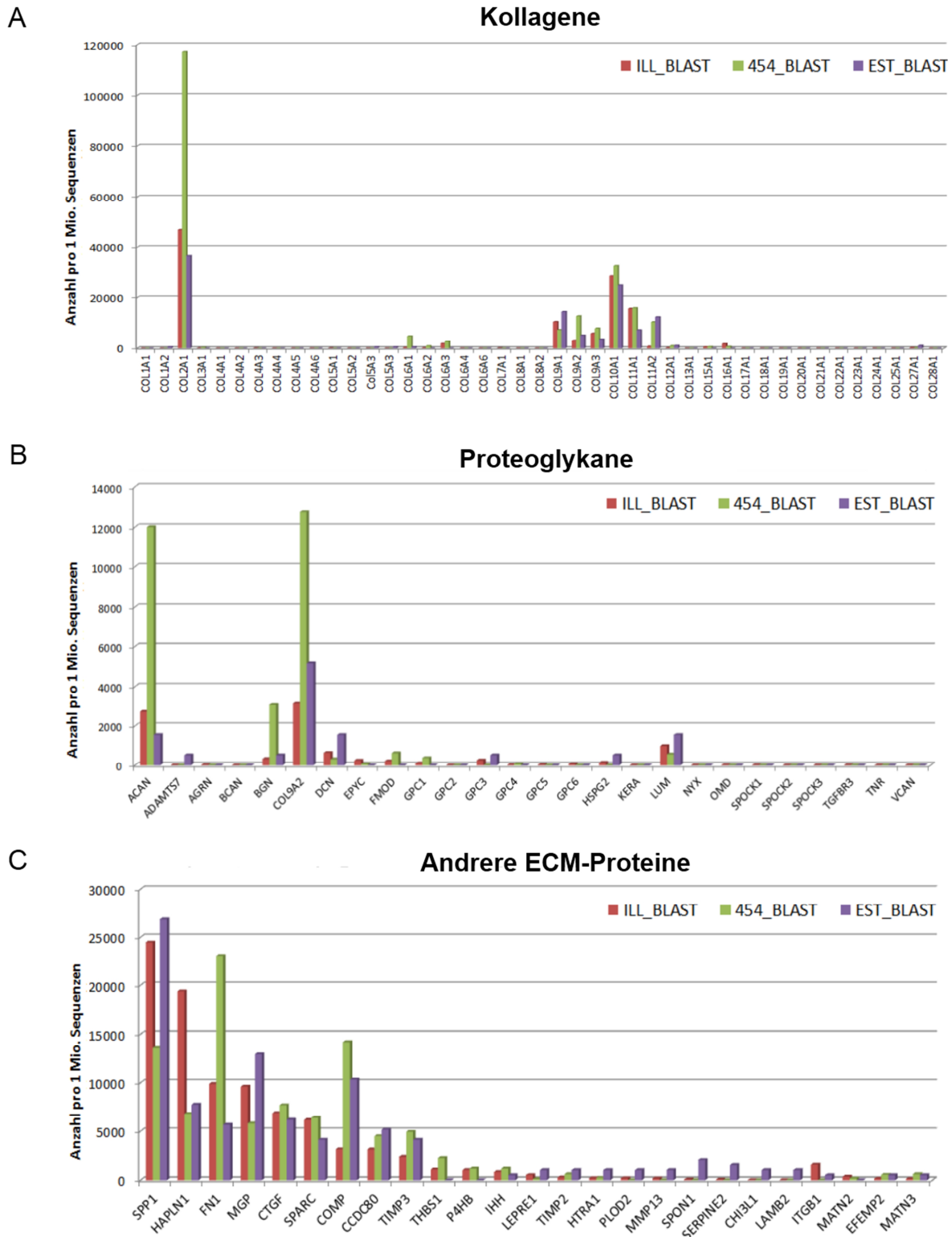


Abb. 40: Vergleich der Blast-Ergebnisse von EST-, 454- und Illumina-Sequenzierung
 Dargestellt sind die Expressionsprofile der drei großen Protein-Gruppen mit besonderer Relevanz bei der Knochen und Knorpelentwicklung; dabei werden die Ergebnisse aus den drei verschiedenen RNA-Sequenzierungen und deren Blast-basierenden Auswertungen miteinander verglichen; rot = Illumina-Sequenzierung; grün = 454/FLX-Sequenzierung; lila = EST-Sequenzierung; die Expressionswerte wurden jeweils auf eine Million analysierte Sequenzen umgerechnet

Expressionsraten im EST-Projekt. Dabei ist aber zu berücksichtigen dass durch die Angleichung auf je 1 Million Sequenzen bei der relativ geringen Sequenzanzahl im EST-Projekt der Fehler wesentlich größer ist.

Eine mögliche Abhängigkeit der Unterschiede in Bezug auf die Transkriptlänge wurde überprüft. Eine solche konnte nicht festgestellt werden. Daher konnte auch durch eine Normalisierung auf „reads per kilo base transcript per million mappable reads“ (RPKM) keine wesentliche Angleichung der Datensätze bewirkt werden. Im Gegenteil, Vergleichswerte werden wesentlich schlechter.

3.6 Genomweites „Mapping“ und differentielle Genexpression

Für die quantitative Analyse von NGS-RNA-Sequenzierungen bietet sich das genomweite Mappen an und ist besonders gut bei Organismen mit sequenzierten und gut annotierten Genomen anzuwenden. Hierzu wurden teilweise die Computerprogramme PASS und Bowtie verwendet. Da beide Programme nur begrenzt für ein genomweites Mapping einsetzbar waren, wurde als Ergänzung bzw. Alternative eine kommerzielle Lösung gewählt. Mit Hilfe

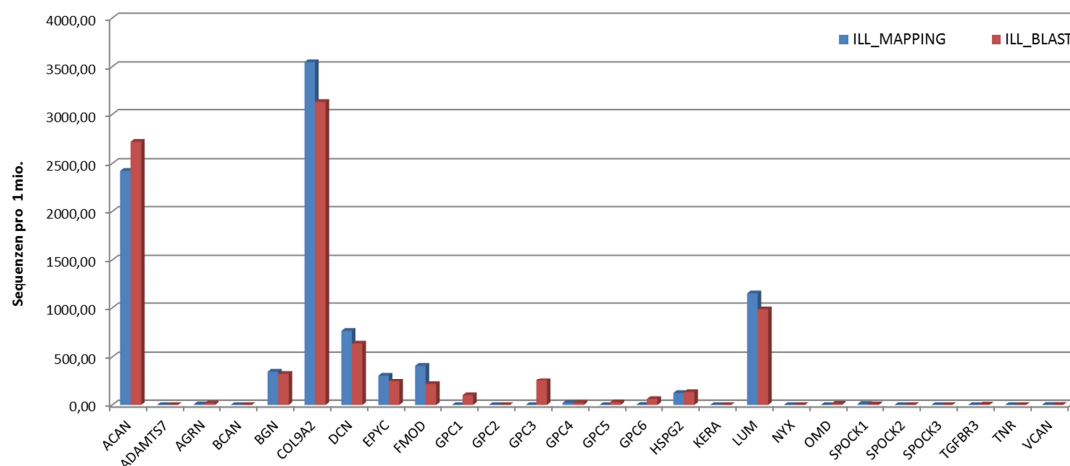


Abb. 41: Vergleich zwischen Blast- und Mapping- Analyse am Beispiel der Proteoglykane
 Dargestellt ist der direkte Vergleich von Expressionstärken (auf 1 Million „Reads“ normalisiert) von Blast-Analyse zu Mapping-Ansatz

der CLC Genomics Workbench wurden die Daten zunächst qualitätsgefiltert (siehe auch Abb. 35C) und gegen die entsprechenden GBK-Dateien (Gene Bank, NCBI) des *Bos taurus*-Genoms gemappt. „Mapping“ und Blast-Suchen führen aber erkennbar zu sehr ähnlichen Ergebnissen (Abb. 41). Dies deckt sich auch mit den Beobachtungen bei weiterführenden Vergleichen.

Dennoch gibt es einzelne aber extreme Unterschiede. Zum Beispiel, wird Glypican 3 (GPC3) im Mapping fast nicht detektiert, während das gleiche Gen durch Blast-Suchen fast 300-fach erscheint. Ähnliches gilt für COL10A1, welches nach Blast-Analyse mit mehreren 100000 Reads vertreten ist und im „Mapping“ nahezu verschwindet. Dies wird besonders beim Vergleich der 25 höchst repräsentierten Gene nach Mapping mit den Ergebnissen der Blast-Analysen deutlich (Tab. 23).

Tab. 23: Top 25 exprimierte Gene nach BLAST-Analyse und „Mapping“
 A Blast-Analyse der Illumina-Daten (unnormalisiert.); B genomisches Mapping der Illumina-Daten (unnormalisiert); grün markiert sind alle Gene die auch im EST-Projekt unter den 25 höchst exprimierten lagen

A	GEN	Sequenzen	B	GEN	Sequenzen
	COL2A1	1842121		COL2A1	2253073
	COL10A1	1115403		SPP1	975956
	SPP1	961024		MGP	699073
	HAPLN1	761019		FTH1	670399
	COL11A1	616748		COL11A1	624921
	COL9A1	414067		HAPLN1	484292
	FN1	388086		EEF1A1	416714
	MGP	376993		LOC100335480	402228
	FTH1	309433		FN1	379921
	RPLP1	293828		RPS3A	365408
	RPS8	285319		RPS8	324614
	RPL11	268584		RPL4	320283
	CTGF	263038		RPL9	302615
	RPS2	253158		RPL11	294408
	SPARC	242182		RPL21	278236
	COL9A3	232382		RPS24	270652
	RPSA	226439		RPS2	263504
	RPS17	226396		TPT1	261185
	RPL31	223720		RPS20	241752
	RPL4	216861		SPARC	240458
	RPS12	214282		RPS10	238058
	RPS10	210697		RPS17	225342
	RPL13A	209075		RPL31	224329
	SAA3	208401		RPSA	212509
	RPS15	202910		RPS12	205628

Wiederum finden sich einige knochen- /knorpelassoziierte Gene im hohen Expressionsbereich. Die ribosomalen Proteine treten noch weiter in den Vordergrund. Auch hier führte die Normalisierung auf RPKM zu einer wesentlich schlechteren Vergleichbarkeit.

4 Diskussion

Die molekularen Mechanismen der Regulation der Genexpression sind vielfältig und komplex. Sie stellen die Grundlage für die evolutionäre Entwicklung ganzer Spezies bzw. die Entwicklung des einzelnen Individuums dar. Sie steuern die Ausbildung und Funktion hoch spezialisierter Zell- und Gewebeverbände, aber auch deren Anpassung an kurzfristige Einflüsse. Dies ist grundsätzlich abhängig von der genetischen Ausstattung des jeweiligen Organismus. Darüber hinaus gibt es eine Vielzahl regulatorischer Ebenen die die Stärke der Expression, aber auch die Struktur, und Funktion der Genprodukte wie der RNA oder letztendlich die der Proteine beeinflussen. Dies können die chromosomale Organisation, die Modifikationen der DNA und des Chromatins, die Bildung der RNA oder die posttranslationalen Modifikationen bei Proteinen sein.

Ein zentraler Punkt dieser komplexen Vorgänge ist die Expression und Prozessierung der RNA. Dies kann untersucht werden, indem die Struktur und Zahl von Transkripten aus verschiedenen Geweben und Zuständen bestimmt wird. Der Rückschluss von spezifischen RNA-Ausprägungen auf einen Phänotyp kann wertvolle Hinweise auf die Funktion und das Zusammenspiel von Genen liefern. Hierfür stehen heute eine Vielzahl an Strategien und Analyseverfahren zur Verfügung. Dabei stellen ESTs („Expressed Sequence Tags“), bei denen Teile von cDNA-Klonen sequenziert werden und damit ein Abbild der exprimierten Gene wiedergeben, ein wichtiges Instrument zur Expressionsanalyse dar (Adams et al. 1991; Velculescu et al. 1999). Geht man davon aus, dass pro Zelle mindestens 300000 verschiedene mRNA-Moleküle vorliegen, würde eine komplette Abdeckung aller in einem Zell- oder Gewebetyp exprimierten Gene die Sequenzierung von weit über einer halben Millionen ESTs erforderlich machen. Dies lag bisher außerhalb der praktischen Möglichkeiten. Mit der Entwicklung der sogenannten „Next-Generation-Sequencing“-Plattformen können nun wesentlich mehr Sequenzen in kürzerer Zeit und wesentlich geringeren Investitionskosten generiert werden. Dies macht die Analyse kompletter Transkriptome möglich. Ein Ziel dieser Arbeit war die Analyse des Transkriptoms eines äußerst komplexen Gewebes, nämlich der „Wachstumsfuge“. Diese ist zentral an der Bildung und dem Erhalt von Knochen- und Knorpelgewebe beteiligt. Dabei wurden bereits sowohl

etablierte Methoden zur Expressionsanalyse mit den neuen Techniken kombiniert und verglichen. Mit den modernen, heute zur Verfügung stehenden Methoden ist es möglich, die Aktivität aller Gene in einem solchen Gewebe zu bestimmen und damit ein besseres Verständnis der zugrunde liegenden Mechanismen bei der Knochen- und Knorpelentwicklung zu erhalten.

4.1 Genexpressionsanalyse der fötalen Wachstumsfuge des Rindes

Mit fötalem Material aus der bovinen Wachstumsfuge stand ein hoch spezialisiertes Gewebe der Knochen- und Knorpelentwicklung zur Verfügung. Das in einem frühen Stadium der Entwicklung hoch aktive Gewebe verspricht eine besonders hohe Expression knochen- bzw. knorpelspezifischer Gene. Diese wurde zunächst in einem EST-Projekt untersucht.

Als Grundlage für die Genexpressionsanalyse wurde zunächst eine cDNA-Bank erstellt. Von dieser cDNA-Bank wurden im Rahmen des EST-Projekts ca. 4000 Klone analysiert, charakterisiert und in Gruppen eingeteilt. Der auffällig hohe Anteil von über 25 % ribosomaler RNA sowie nicht auswertbarer Sequenzen und Leervektoren ergab lediglich 2242 (55,6 %) auswertbare EST-Sequenzen. Obwohl diese Menge bei weitem nicht für eine vollständige Expressionsanalyse ausreicht, können auch kleinere Datenmengen schon einen guten Überblick über das Expressionsmuster geben. Mit zusammen ca. 20 % Sequenzen schlechter Qualität und leeren Vektoren liegt dieses Projekt über dem Durchschnitt im Vergleich zu anderen Projekten vergleichbarer Gewebe (ca. 6 % (Kwon et al. 2008; Tagariello et al. 2005)). Der hohe Anteil ribosomaler RNA kann auf eine weniger effiziente Anreicherung der polyA⁺-RNA zurückzuführen sein. Oftmals verbleibt gerade bei Aufreinigungen mit oligo(dT)-gekoppelten „Beads“ im „Batch“-Verfahren ein hoher Anteil rRNA in der Probe. Wiederholungen der Aufreinigung oder Säulenverfahren können allgemein die Verluste erhöhen, aber auch die rRNA deutlich reduzieren.

Von den 2242 verwertbaren Sequenzen konnten 1816 ca. 870 bekannten Genen zugeordnet werden. Die eindeutige Zuordnung zu funktionellen Gruppen ist schwierig. Erwartungsgemäß konnte aber ein Großteil der eindeutigen Gene allgemein exprimierten funktionellen Gruppen des Stoffwechsels, Transports, der RNA-Prozessierung oder

Proteinbiosynthese zugeordnet werden. Diese finden sich oft ubiquitär auch in anderen Geweben hoch exprimiert wieder (Adams et al. 1995). Ein Maß hierfür kann der sogenannte „Tissue Specificity Index“ (HUSAR, Husar Bioinformatics Lab, Molecular Biophysics, Deutsches Krebsforschungszentrum) sein. Dieser beinhaltet die Anzahl der Gewebe für ein bestimmtes Gen, in denen dies schon durch ESTs nachgewiesen wurde. So konnte beispielsweise der „*Eucaryotic Elongation Factor 1 alpha 1*“ (*EE1A1*), mit einem TSI von 21 (max. 21) an Position 17 der höchst exprimierten Gene, mit 14 ESTs, nachgewiesen werden. Auch das „*ferritin, heavy polypeptide 1*“ (*FTH1*) mit einem TSI von 17 oder das ribosomale Protein S2 (*RPS2*, TSI 20) wird in der fötalen Wachstumsfuge hoch exprimiert (Tab. 12). Eine der größten Gruppen, die ubiquitär exprimiert werden, sind die ribosomalen Proteine. Allein 282 Sequenzen verteilt auf 66 verschiedene Gene (ca. 7,6 %) zeigten Übereinstimmungen zu ribosomalen Proteinen (15,5 % der identifizierten Sequenzen) und bestätigen damit eine realistische EST-Verteilung. Für die hohe Spezifität des Gewebes spricht der sehr hohe Anteil von Genen mit Verbindungen zur Knochen- und Knorpelentwicklung. Insgesamt konnten ihnen 48 Gene mit 415 Sequenzen zugeordnet werden. Damit vereinigt diese Gruppe mit ca. 5,6 % der identifizierten Gene fast 23 % der Sequenzen. Darüber hinaus sind allein 20 der 100 höchst repräsentierten Gene mit der Skelettentwicklung und Zelladhäsion assoziiert (Abb. 15).

Unter den knochen- und knorpelspezifischen Genen sind die Kollagene mit insgesamt 211 ESTs am meisten vertreten. Diese sind zusammen mit Glykosaminoglykanen, Proteoglykanen und einer Vielzahl weiterer nicht kollagenartiger Proteine Bestandteil der so genannten extrazellulären Matrix (ECM) (Hay 1981). Sie nimmt den wesentlichen Volumenanteil am knochenbildenden Gewebe ein (Ballock und O'Keefe 2003). Eine der zentralen Aufgaben der ECM ist die Bildung eines stabilen Gerüsts, das auch als Matrix für die Zelladhäsion dient. Dies geschieht hauptsächlich über miteinander vernetzte fibrilläre und Fibrillen-assoziierte Kollagene des Typs II, IX und XI. Dabei ist der wesentliche Bestandteil Kollagen Typ II (Eyre et al. 2002; Hay 1981). Zusätzlich zur Vernetzung untereinander können weitere Strukturen bei der Verknüpfung durch Kollagen des Typs X oder durch Proteoglykane entstehen. Diese Zusammenhänge spiegeln sich im Expressionsprofil der Kollagene wider. Kollagen Typ II wird als Hauptbestandteil der ECM am stärksten exprimiert. Seine Expression beginnt schon früh

in der Knochenentwicklung und wird sowohl von Chondrozyten aller Reifestadien als auch den perichondralen Zellen gebildet. Die Expression nimmt mit zunehmendem Alter und Reifung der Chondrozyten zu (Zhang et al. 2003; Pogue et al. 2004; Tagariello et al. 2005). Das aus den Untereinheiten alpha 1, 2 und 3 bestehende Heterotrimer Kollagen Typ IX findet sich in einem Verhältnis 28:10:7 auch in der fötalen Wachstumsfuge wieder. Auch dies deckt sich mit der Beobachtung eines erhöhten Gehalts von Kollagen IX in fötalem Gelenkknorpel, der anteilig um das ca. 10-fache steigen kann (Bruckner et al. 1988; Har-el et al. 1992). Kollagene dieses Typs bilden selbst keine Fibrillen, stabilisieren jedoch das fibrilläre Netzwerk, in dem sie auf der Oberfläche der Knorpelfibrillen kovalent gebunden werden (Bruckner et al. 1988). Für den Vergleich mit entsprechenden EST-Projekten mit menschlichem Material stehen sowohl die Arbeiten von Zhang et al. (2003) und Tagariello et al. (2005) zur Verfügung. Diese beinhalten Expressionsanalysen von humanem fötalen

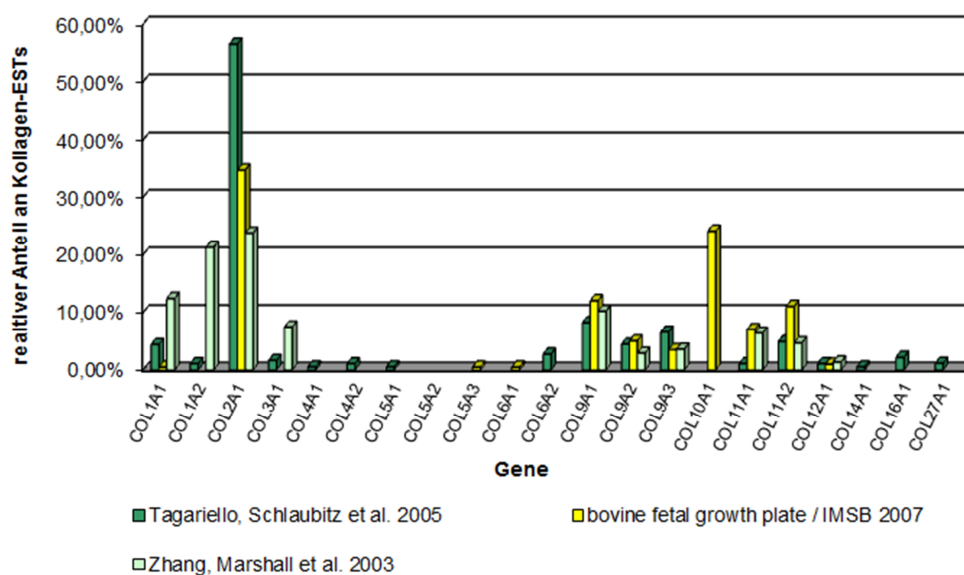


Abb. 42: Vergleich - Kollagene – Expressionsprofile
Vergleich der Expressionsprofile von humanem Knorpel- und Wachstumsfugengewebe und der fötalen Wachstumsfuge des Rindes

Knorpel aus der 8.-12. Schwangerschaftswoche (Zhang et al. 2003) und humanem Knorpel aus der 20. Schwangerschaftswoche bis zu einem Alter von 2 Jahren (Tagariello et al. 2005). Dabei sind einige Unterschiede zu erkennen (Abb. 42). Am deutlichsten treten diese Unterschiede in Bezug auf das frühe Expressionsprofil von Zhang et al. 2003 auf. Dabei wird *COL2A1* deutlich geringer exprimiert, wohingegen *COL1A1*, *COL1A2* und *COL3A1* wesentlich größere Anteile an der gesamten Kollagenexpression zeigen. Die Expression von *COL1A1* und

COL1A2 wird meist von kondensierten mesenchymalen Stammzellen im Frühstadium der embryonalen Skelettbildung exprimiert und unterstreicht die Herkunft der Proben aus einem frühen Entwicklungsstadium (Mundlos et al. 1990; Gilbert 2006). Über dies hinaus wird eine hohe Expression von Kollagen des Typs I durch eigene Transkriptom-Daten aus adulten MSCs unterstützt. Das ebenfalls höher repräsentierte Kollagen vom Typ III alpha 1 wird auch eher in der Frühphase exprimiert, ist mit Kollagen Typ I assoziiert und oftmals koexprimiert (Masuda et al. 1994). Neben Kollagen IX, das sich in allen untersuchten Stadien ähnlich hoch exprimiert nachweisen ließ, konnte in der bovinen Wachstumsfuge eine nochmals gesteigerte Expression von Kollagen Typ XI, vor allem des Typs alpha 2 festgestellt werden. Kollagene vom Typ XI sind Heterotrimere [$\alpha 1$ (XI) $\alpha 2$ (XI) $\alpha 3$ (XI)] und werden von den drei Genen *COL11A1* und *COL11A2* und *COL2A1* codiert. Dabei werden die $\alpha 1$ (II)-Kette und die $\alpha 3$ (XI)-Kette von einem Gen kodiert (*COL2A1*) und unterscheiden sich lediglich in der posttranslationalen Modifikation (Eyre et al. 1987). Im Innern der Fibrillen gelegen hat Kollagen XI einen Einfluss auf deren Bildung und kontrolliert den Durchmesser wie auch die räumliche Organisation der Knorpelfibrillen (Blaschke et al. 2000). Der Anteil an den exprimierten Kollagenen von Typ II liegt im Vergleich zum älteren Knorpelmaterial (Tagariello et al. 2005) etwas niedriger. Alles in allem lässt sich das bearbeitete Material zwischen die beiden verglichenen Entwicklungsstadien einordnen. Obwohl der sehr hohe Anteil von Kollagen Typ X (ca. 22%) eine spätere Entwicklungsphase nahe legt, kann dies auch durch einen höheren Anteil terminal differenzierter Chondrozyten erklärt werden. Da die Probennahme in diesem Stadium sehr schwierig ist, ist davon auszugehen, dass es sich bei diesem um einen geringfügigen Unterschied in der Probenzusammensetzung handelt.

Neben der großen Stabilität, die die ECM vermittelt, wird vor allem in einem späteren Entwicklungsstadium, wenn vermehrt mechanische Kräfte auftreten, auch eine hohe Elastizität benötigt. Diese wird durch eine weitere wichtige funktionelle Gruppe der ECM, die Proteoglykane, vermittelt. Mit ihren gebundenen Glycosaminoglykanen (GAG) bilden sie extrazellulär Quervernetzungen zwischen den Strukturkomponenten der ECM, wie beispielsweise den Kollagenen. Außerdem halten sie durch ihre Polysaccharid-Seitenketten einen hohen osmotischen Druck aufrecht, der für einen erhöhten Wasseranteil (80 %) im Knorpel und damit eine hohe Elastizität sorgt (Zhu et al. 1996). Darüber hinaus gibt es

Gruppen von Proteoglykanen, die intrazellulär oder auch an der Zelloberfläche essentiell für die Gewebeorganisation, die Regulierung, das Wachstum und die Differenzierung von Zellen (Iozzo 1998; Iozzo und Murdoch 1996) sind. Hier konnten nur 23 ESTs aus der fötalen Wachstumsfuge des Rindes den Proteoglykanen zugeordnet werden. Dabei sind die Gene für Aggrecan (*AGC1*; *ACAN*), Decorin (*DCN*), Glypican 3 (*GPC3*), Leprekan (*LEPRE*) Lumican (*LUM*) und Syndecan 2 (*SDC2*) mit je 3 ESTs am häufigsten vertreten. Chondroitinsulfat Proteoglykan 4 (*CSPG4*), Biglycan (*BGN*), Perlecan (*HSPG2*) und Syndecan 4 (*SDC4*) sind nur mit zwei bzw. einer EST-Sequenz unterstützt. Dabei teilen sich diese in zwei Gruppen auf, strukturelle und regulatorische Proteoglykane. Aggrecan, mit ca. 210 kDa das größte Proteoglykan, bindet an Hyaluronat (Mörgelin et al. 1988) und kann über eine Lectin-ähnliche Struktur Ca^{2+} -abhängig verschiedene Kohlenhydrate binden (Saleque et al. 1993). Dabei sorgt es mit einer hohen fixierten Ladungsdichte für eine hohe Elastizität. Decorin, Biglycan, Leprekan, Lumican und Perlecan sind ebenfalls in die Strukturbildung der ECM involviert und beeinflussen in hohem Maß die Bildung und Struktur der Kollagenfibrillen (Vogel et al. 1984; Vogel und Trotter 1987). Anders als Aggrecan sind diese neben Knochen oder Knorpel auch in anderen Bindegeweben zu finden (Pringle und Dodd 1990; Fleischmajer et al. 1991; Blochberger et al. 1992). Außerdem konnte eine mögliche Bindung von Wachstumsfaktoren und die Interaktion mit Oberflächenrezeptoren, wie EGFR oder IGF-IR bei Decorin (Iozzo et al. 1999) oder TLR2/4 bei Biglycan (Schaefer et al. 2005) bzw. Lumican, nachgewiesen werden (Iozzo und Schaefer 2010). Gerade der Einfluss auf den TGF- β -Regulationsmechanismus (Yamaguchi et al. 1990) und den IGF-IR zeigen den möglichen Einfluss der Zusammensetzung der ECM auch in der Knochen- und Knorpelentwicklung auf die Zellproliferation und -differenzierung. Dazu kommen noch kleine transmembrane Heparansulfat-Proteoglykane wie Glypican-3 oder Syndecan-2, die entscheidend in die Regulation der Zellproliferation und oftmals in embryonale und maligne Entwicklungen involviert sind (Saikali und Sinnott 2000; Grisar et al. 2001; Song et al. 2005; Fukumoto und Levin 2005). Speziell SDC2 zeigt eindeutige Einflüsse auf das TGF- β -Regulationsnetzwerk oder die Assemblierung der ECM, die auch in der Knochenbildung eine zentrale Rolle spielen (Klass et al. 2000; Chen et al. 2004). Die Expression dieser Gene spiegelt das Anfangsstadium einer stabilen ECM mit hauptsächlich in Knorpel- oder Knochengewebe vorkommenden Proteoglykanen und die Spezifität des Probenmaterials wider. Eine eher niedrigere

Expression von Biglycan (1 von 2242 Sequenzen) und Aggrecan (3 von 2242 Sequenzen) gegenüber der Expression bei Tagariello et al. 2005 (BGN 9/3647 und ACAN 6/3647) spricht ebenfalls für ein frühes Entwicklungsstadium und unterstützt das Kollagen-Expressionsprofil.

Im Gegensatz zu den Proteoglykanen, die mit insgesamt 23 ESTs nur schwach exprimiert werden, fallen bei den stark exprimierten Genen neben den Kollagenen noch andere ECM-Proteine auf. Besonders hoch exprimiert finden sich *SPP1*, *MGP*, *LECT1*, *COMP* und *HAPLN1* (Tab. 12). Damit sind durch Osteopontin (*SPP1*), „Matrix Gla Protein“ (*MGP*) und „Cartilage oligomeric Matrix Protein“ (*COMP*) gleich drei wichtige Proteine der Regulation von Knochenmineralisierung und Blutgefäßeinwanderung vertreten (Canfield et al. 2002). Osteopontin, ein nicht kollagenöses 60 kDa großes Phosphoprotein (Senger et al. 1979), findet sich in diversen Geweben und ist in unterschiedliche Mechanismen involviert (Sodek et al. 2000; Denhardt et al. 2001). Im knochenbildenden Gewebe tritt es vermehrt auf (Butler WT 1996) und wird der Zone hypertropher Chondrozyten bzw. Mineralisierung zugeordnet (Belluoccio et al. 2008). Dabei bindet es an transmembrane Integrine, fördert die Zelladhäsion und wirkt inhibitorisch auf den Mineralisierungsprozess (Boskey et al. 2002).

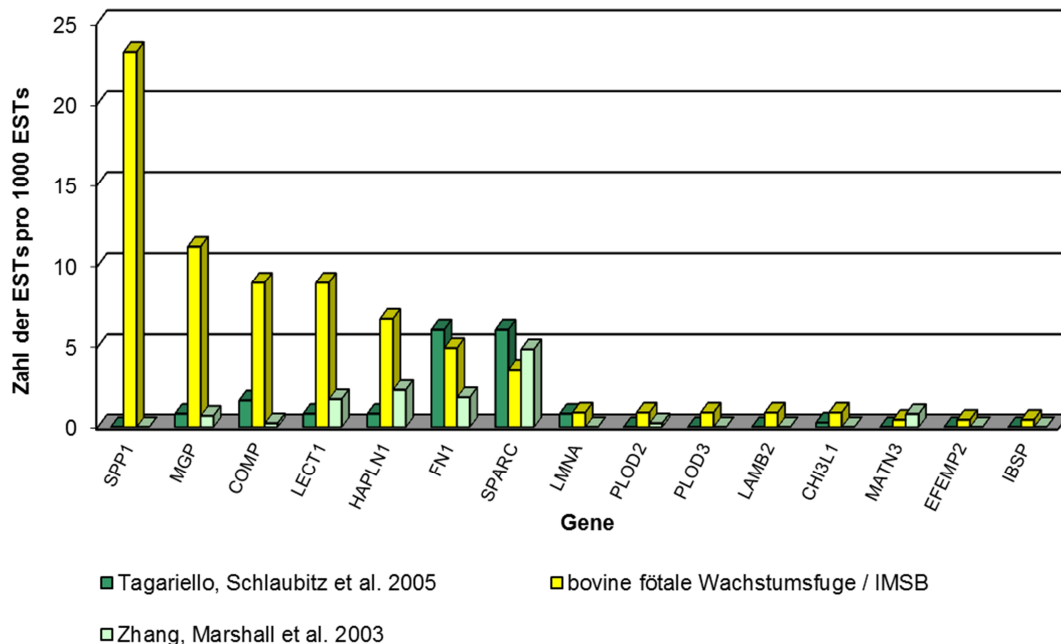


Abb. 43: Vergleich - Expressionsprofile der ECM-Proteine
Vergleich der Expressionsprofile von humanem Knorpel- und Wachstumsfugengewebe und der fötalen Wachstumsfuge des Rindes

Ähnlich verhält es sich mit MGP. Auch für dieses Ca^{2+} -bindende Protein konnte eine inhibitorische Wirkung bei der Mineralisierung im Knochenbereich, aber auch der Blutgefäße nachgewiesen werden. Dabei beeinflusst es nicht nur die Zell-Zell-Interaktionen und die Bindung zu anderen Strukturelementen wie Fibronectin, sondern auch Signaltransduktionskaskaden der „Transforming growth Factor“-Familie, wie BMP2 (Zebboudj et al. 2002; Boström et al. 2004).

Allgemein konnten im EST-Projekt nur wenige regulatorische Elemente der Knochen- und Knorpelentwicklung identifiziert werden. Dies dürfte im Wesentlichen auf die geringe Abdeckung des Transkriptoms durch EST-Sequenzen zurück zuführen sein. Neben TGF- β konnten essentielle Komponenten bei der Interaktion mit Transmembranproteinen wie Integrinen oder Proteoglykanen in Form von Kollagenen, Fibronectin und Osteopontin

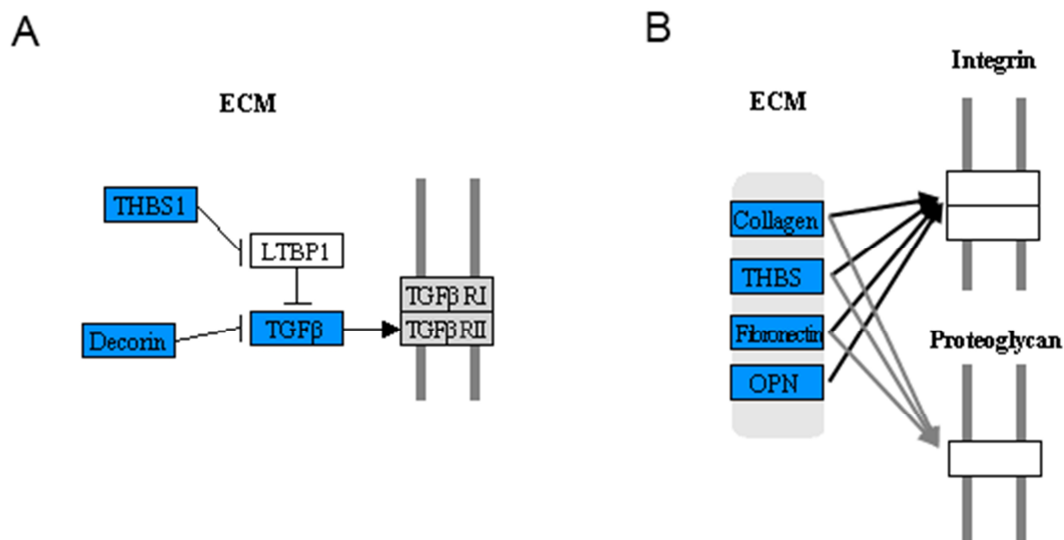


Abb. 44: Signalwege der ECM – Abdeckung im EST-Projekt
Blau hinterlegt sind die in den 100 höchstexprimierten EST-Sequenzen nachgewiesenen Gene mit Beteiligung an regulatorischen Signalwegen in der ECM
Verändert nach KEGG: Kyoto Encyclopedia of Genes and Genomes, Copyright Kanehisa Laboratories

(OPN/SPP1) unter den 100 am stärksten exprimierten EST-Sequenzen der Wachstumsfuge identifiziert werden. Sowohl in die negative Regulation der TGF- β -Kaskade als auch der Interaktion mit Integrinen und Proteoglykanen spielt das „cartilage oligomeric matrix protein“ (COMP) eine wichtige Rolle. Darüber hinaus ist das Thrombospondin COMP dazu in der Lage Ca^{2+} -Ionen zu binden und ist verstärkt an der Strukturbildung beteiligt aber auch in die Zelladhäsion und Fibrillenorganisation involviert. COMP ist assoziiert mit über 70

pathogenen Phänotypen im Bereich des Skelettsystems (Hecht et al. 2005; Merritt et al. 2006). Im fötalen Gewebe wird es dem basophilen Bereich zwischen hypertropher- und Mineralisierungszone zugeordnet. Diese Region ist besonders reich an Proteoglykanen und anderen Strukturkomponenten (Hedbom et al. 1992). Wie in Abb. 43 zusammengefasst dargestellt ist, sind die beschriebenen Gene nicht nur in der fötalen Wachstumsfuge laut EST-Daten hoch exprimiert, sondern treten auch verglichen mit den humanen Expressionsprofilen deutlich höher exprimiert auf. Es ist davon auszugehen, dass die in dieser Arbeit untersuchten Proben einen wesentlich höheren Anteil der territorialen Matrixbereiche von reifen Chondrozyten enthielten. Gerade in diesem Bereich ist die Regulation der Mineralisierung besonders wichtig. Besonders die Expression von inhibitorischen Faktoren wie SPP1, MGP und COMP, die gleichzeitig an der Bildung einer stabilen Knorpelmatrix beteiligt sind, könnte eine fortschreitende Mineralisierung zur Epiphyse hin verhindern. Ob dies zusätzlich durch das embryonale Stadium beeinflusst ist, lässt sich über den Vergleich mit anderen Projekten nicht erschließen. Grundsätzlich ist die Knochenmineralisierung sowohl im fötalen als auch im frühen postnatalen Stadium noch unvollständig und die Knochen sehr flexibel. Dies bietet einen gewissen Schutz vor Frakturen.

Des Weiteren zeigen auch FN1, SPARC sowie LECT1 und HAPLN1 eine starke Expression. Letztere sind wiederum Strukturkomponenten, die hauptsächlich mit der hypertrophen Zone assoziiert sind (Watanabe und Yamada 1999; Amil et al. 2004). Bei FN1 und SPARC handelt es sich vor allem in der embryonalen Phase ubiquitär exprimierte Gene, die in der Quervernetzung von Proteoglykanen und Kollagenen sowie der Zelladhäsion eine wichtige Rolle spielen.

Darüber hinaus konnten weitere interessante hoch exprimierte Gene identifiziert werden. Zu diesen gehörten neben „Inhibitor of DNA-binding 1“ (ID1), „Coiled-coil domain-containing protein 80“ (CCDC80) und „150 kDa TGF-beta-1-binding protein“ (CD109) auch H19 und PHOSPHO1.

Davon spielen sowohl ID1 als auch CD109 in den TGF- β -Kaskaden eine Rolle. ID1, ein Helix-Loop-Helix-Protein, wird nicht nur in der Wachstumsfuge exprimiert, sondern ist bisher in einer Vielzahl von Geweben nachgewiesen worden. Dies trifft vor allem für embryonale und

maligne Gewebe zu. Dort hat es Einfluss auf die Differenzierung, die es inhibiert und die Zellproliferation, die es fördert (Ying et al. 2003). In großem Maße dürfte dies mit der Fähigkeit zur Dimerisierung mit anderen HLH-Transkriptionsfaktoren zusammenhängen. Diese werden durch die Bindung in ihrer Wirkung neutralisiert (Langlands et al. 1997). Obwohl ID1 eine übergeordnete Bedeutung zukommt, kann davon ausgegangen werden, dass es auch zentral in die Knochen- und Knorpelentwicklung eingreift. So wurden beispielsweise BMP-Gene schon als direkte Zielgene für die ID-Regulation identifiziert (Hollnagel et al. 1999). Bei dem bisher weniger untersuchten CD109 handelt es sich um ein Oberflächenantigen aus der Familie der alpha 2 Makroglobuline (Solomon et al. 2004). Bisher wurde nur die Expression in diversen Tumorgeweben (Hashimoto et al. 2004) und die Interaktion mit dem TGF- β -Rezeptor bei Keratinozyten näher untersucht (Finnson et al. 2006). Eine spezifische Rolle in Knochen und Knorpel oder deren Bildung kann daraus nicht direkt abgeleitet, aber auch nicht ausgeschlossen werden. Gerade dass CD109 bisher in keinem anderen mit fötalem, knochenbildendem Gewebe durchgeführtem Projekt in Erscheinung trat, ist ungewöhnlich.

Obwohl CCDC80 in vielen Geweben nachgewiesen wurde, fehlt bisher eine eindeutige Zuordnung zur Wachstumsfuge. Hinweise auf eine mögliche Funktion in diesem Bereich sind die verminderte Expression bei der osteogenen Differenzierung von MSC (Liu et al. 2004) und die Möglichkeit zur Bindung von Fibronectin. Aus dem Vergleich mit Daten aus *in vitro* in Osteoblasten differenzierten MSCs geht eindeutig eine Abnahme der Expression um mindestens den Faktor 2 hervor. Der Expressionslevel in der Wachstumsfuge liegt mehr als 20-fach höher als in den undifferenzierten MSCs. Demnach ist anzunehmen, dass CCDC80 im Bereich der Wachstumsfuge und speziell der hypertrophen und Mineralisierungszone an der Zelladhäsion und Bildung der ECM-Struktur beteiligt ist. Ähnlich verhält es sich mit der „Phosphoethanolamine/phosphocholine phosphatase“ (PHOSP1). Die Dehalogenase ist nicht nur durch seine genomische Lage in der Nähe zum HOXB-Cluster und COL1A1 mit der Knochenbildung assoziiert, sondern konnte auch schon stark exprimiert in der Mineralisierungszone nachgewiesen werden (Houston et al. 1999; Houston et al. 2004). Dies kann in dieser Arbeit bestätigt werden. Dabei stellt es vermutlich die Phosphatgruppe für die Bildung von Hydroxyapatit zur Verfügung.

Besonders interessant ist die hohe Expression des nicht proteinkodierenden des „Imprinting“ unterliegenden Gens H19. Dieses wird im Menschen lediglich vom maternalen Allel transkribiert (Bartolomei et al. 1991; Brannan et al. 1990). Es scheint im Zusammenhang mit dem geprägten Locus IGF2 zu stehen, der paternal exprimiert wird. In Verbindung mit dem ubiquitären Vorkommen, vor allem in embryonalen und Tumorgewebe, wurde die Funktion eines Tumorsuppressors angenommen (Ayesh et al. 2002). Bei Expressionsuntersuchungen der Wachstumsfuge trat H19 bei Belluoccio et al. (2008) in der Proliferationszone hervor. Demnach könnte die H19-Expression durchaus eine Rolle bei der Proliferationsrate der Chondrozyten spielen.

Obwohl im EST-Projekt mit 2242 verwertbaren Sequenzen nur ca. 900 Gene nachgewiesen werden konnten, zeigen die Expressionsprofile dennoch das Abbild eines hoch aktiven, spezifischen knochenbildenden Gewebes. Es konnten Gene der grundlegenden zellulären Mechanismen ebenso wie spezifische Gene der Knochen- und Knorpelentwicklung oder der Embryonalentwicklung nachgewiesen werden. Allerdings konnten über die weniger stark exprimierten Gene, wie beispielsweise Transkriptionsfaktoren, nur wenig neue Informationen gewonnen werden.

Expressionsanalyse über „Next Generation Sequencing“

Um eine höhere Abdeckung zu erzielen, die EST-Daten abzusichern und zu überprüfen wurden zwei verschiedene Techniken des „Next-Generation-Sequencing“ eingesetzt. Dabei blieben die generierten Datenmengen bei der 454-FLX-Technologie, mit „nur“ 38492 Sequenzen, weit hinter den Erwartungen zurück. Dennoch mussten für die Analyse dieser doch umfangreichen Datensätze neue Strategien zum Auswerten etabliert werden. Die Daten wurden sowohl geclustert als auch über Hochdurchsatzblast-Analysen identifiziert und quantitativ ausgewertet (s. S. 82). Dabei konnten über Auswertung der geclusterten Daten 2629 verschiedene Gene und über Einzelblast-Suchen 2990 Gene zugeordnet werden. So konnten trotz der relativ kleinen Ausbeute 3-mal so viele Gene wie im EST-Projekt abgedeckt werden. Bei Betrachtung des allgemeinen Expressionsprofils und der stark

repräsentierten Gene ist auch hier eine knochen-/knorpelspezifische Expression festgestellt worden.

Neben den charakteristischen Kollagenen Typ II, IX, X und XI werden auch viele Knorpel-spezifische Proteoglykane exprimiert. Besonders stark repräsentiert sind Aggrecan, Biglycan Decorin und Lumican. Analog zur EST-Analyse lassen sich diese gut in den arbeitsbezogenen Zusammenhang setzen. Außerdem treten beim absoluten Quantifizieren eine Reihe weiterer hoch repräsentierter Gene auf, die sich im EST-Projekt etwas geringer exprimiert darstellten (Tab. 18). Dazu gehören die drei Untereinheiten des Kollagens Typ IX und Kollagen Typ VI alpha 1 ebenso wie Aggrecan, „insulin-like growth factor 2“ (IGF2), „lysyl oxidase-like 4“ (LOXL4) und „TIMP metalloproteinase inhibitor 3“ (TIMP3). Dabei lassen sich die Kollagene vom Typ XI ähnlich wie im EST-Projekt gut mit der frühen Knochenbildung in Verbindung bringen. Der hohe Anteil von Kollagen des Typs VI alpha 1 ist eher ungewöhnlich. Obwohl Kollagen VI in der extrazellulären Matrix vieler Gewebe exprimiert wird, so findet es sich doch hauptsächlich in Muskeln und Fettkörpern. Dort bildet das Kollagen perlenschnurartige Fibrillen. Es wird auch im Zusammenhang mit Knorpelgewebe beschrieben (Hagiwara et al. 1993) und zeigt nachweislich eine Interaktion mit Decorin (Bidanset et al. 1992), ist aber mit einem durchschnittlichen Anteil von ungefähr 0,5 % der Kollagene im Knorpel eher unterrepräsentiert (Swoboda et al.).

Besser fügen sich die Lysyl-Oxidase, LOXL4, und der Wachstumsfaktor IGF2 in den Knochen- und Knorpel-spezifischen Zusammenhang. Obwohl es in jüngster Zeit eher in Verbindung mit Kopf- und Nackentumoren beschrieben wird, konnte LOXL4 sowohl in murinen Chondrozyten-Zellkulturen als auch in den Bereichen der hypertrophen Chondrozyten und Mineralisierung nachgewiesen werden. Dort ist es in die Modulation der Kollagenfibrillen involviert (Ito et al. 2001).

TIMP3, das sich auch in den ESTs recht hoch repräsentiert wiederfindet, liegt nicht spezifisch im ECM-Knorpelgewebe vor. Als Inhibitor von Metalloproteinasen, die zentral am Um- und Abbau der ECM beteiligt sind, ist eine Beteiligung an Vorgängen in der Wachstumsfuge denkbar.

Der insulinabhängige Wachstumsfaktor IGF2 ist vor allem an frühen Entwicklungsprozessen beteiligt (Szabó und Mann 1995; Constância et al. 2002). Des Weiteren wurde eine

Expression in der Proliferationszone der murinen Wachstumsfuge nachgewiesen (Belluoccio et al. 2008). In Zusammenhang mit verschiedenen malignen Gewebeveränderungen ist IGF2 eine Funktion bei der Regulation der Zellproliferation zuzuordnen. Wie bereits erwähnt, steht die Expression von IGF2 in Verbindung mit der von H19. Beide liegen in derselben genetisch imprinteten Kontrollregion. Dabei fällt auf, dass im Gegensatz zum EST-Projekt, in dem das Verhältnis von H19 zu IGF2 bei 2:1 liegt, die 454 Sequenzierung ein Verhältnis von ungefähr 1:8 ergibt. Hier finden sich erste Hinweise auf einen eventuell systematischen Fehler.

Im Weiteren konnten im Vergleich der Expressionsprofile der Knochen-/Knorpel-relevanten Gruppen mit denen des EST-Projekts noch weitere gravierende Unterschiede festgestellt werden (Abb. 40). Diese sind zum Teil widersprüchlich. So lagen die Anteile von Kollagen Typ II, VI und IX alpha 2 im 454-Projekt wesentlich höher. Auch Aggrecan und Biglycan erschienen wesentlich stärker exprimiert, wie auch Fibronectin und COMP. Im Gegensatz dazu sind SPP1 und MGP unterrepräsentiert (Abb. 40). Der hohe Anteil von COL2A1, ACAN und BGN würde auf eine spätere Entwicklungsphase oder eine Zone mit Chondrozyten höheren Reifegrades schließen lassen. Die übersteigerte Expression von FN1 und COMP zeigt Parallelen zu reinem Knorpelgewebe (Kwon et al. 2008). Dieses steht wiederum nicht in Einklang mit einer verminderten MGP-Expression.

Demnach zeigt die 454-Sequenzierung eine grundsätzliche Abweichung, die das gesamte Expressionsprofil gegenüber dem der ESTs scheinbar etwas verschiebt. Im Einzelfall können diese Abweichungen gravierend sein. Dennoch ist das Profil eindeutig Knochen-/Knorpel-spezifisch. Hierfür gibt es mehrere Erklärungsmöglichkeiten: Aufgrund der großen Menge benötigter mRNA wurden 12 polyA⁺-Aufreinigungen durchgeführt und vereinigt. Daher musste Gesamt-RNA verschiedener Chondrozyten-Isolationen verwendet werden. Es ist nicht auszuschließen, dass unterschiedliche Extraktionen in der Zusammensetzung variieren. Mögliche Ursachen für die abweichende Expression von Kollagenen und ECM-Proteinen wären eine nur teilweise Isolierung des Wachstumsfugengewebes oder große Anteile („Kontaminationen“) von hyalinem Gelenkknorpel.

Darüber hinaus darf nicht außer Acht gelassen werden, dass die durch ein kommerzielles Unternehmen durchgeführte Sequenzierung zunächst nicht gelang. Aufgrund zu geringer

Probenmengen wurde vor einem zweiten Versuch vom Institut für Immunologie und Genetik (Kaiserslautern) eine WGA („whole genome amplification“) durchgeführt. Bei der Amplifikation, bei der die zugrunde liegende Molekülzahl eine große Rolle spielt, kann das Expressionsprofil bei entsprechend geringen Ausgangsmengen verfälscht werden.

Hinzu kommt die Möglichkeit einer selektiven Amplifikation in Abhängigkeit der Größe oder der Basenzusammensetzung. Auch die Auswirkungen der Kontamination der cDNA-Bibliothek mit Material von *Vitis vinifera* bei der Sequenzierung sind schwer abzuschätzen.

Ein bekannter Effekt bei der Analyse von kurzen NGS Sequenzen ist eine gewisse Abhängigkeit der Read-Anzahl von der Länge des kompletten Transkripts. Um dies zu kompensieren, wurde zunächst einheitlich auf eine Transkriptlänge von 1000 bp normalisiert. Im weiteren Verlauf wurde die mittlerweile weit verbreitete Einheit RPKM verwendet. Der Vergleich im Ranking zwischen den höchst exprimierten Genen durch 454- und EST-Sequenzierung zeigt eine Annäherung der Datensätze. Dennoch bleiben Unterschiede, die im Einzelfall groß sein können.

Insgesamt konnte der 454-Datensatz zur Etablierung der Auswerteverfahren von NGS-Datensätzen und der Identifikation von Transkripten mit Einfluss auf die Knochen- und Knorpelbildung verwendet werden. Dennoch blieben sowohl die Abdeckung des Transkriptoms wie auch die Abdeckung der einzelnen Transkripte und die Vergleichbarkeit mit den EST-Daten hinter den Erwartungen zurück.

Transkriptom-Analyse mittels Illumina-Sequenzierung

Mit der Weiterentwicklung der Illumina Genome Analyzer- Technologie bot sich die Möglichkeit, diese auch für die Untersuchung von Genexpression mittels cDNA-Sequenzierung einzusetzen. Dazu wurde linear amplifizierte mRNA aus der bovinen Wachstumsfuge in einer „Lane“ auf einem Illumina Genome Analyzer Iix in Form von 100 bp langen „paired ends“ sequenziert. Die enormen Datenmengen und kurzen Leseweiten machten die Etablierung und Entwicklung neuer Methoden für die Analyse notwendig. Dazu wurden verschiedene Qualitätskriterien und -filter getestet und in Anlehnung an die Auswertung der 454 Daten wurde die Hochdurchsatz-Blast-Analyse optimiert. Als Alternative

oder Ergänzung zur Blast-Annotation wurden verschiedene „Mapping“-Verfahren zu genomweiten-, aber auch Einzel-Loci-Analysen etabliert.

Über die Blast-Analysen konnten dabei ca. 44700 Gencluster identifiziert werden. Diese enthalten aus technischen Gründen noch redundante Einträge. Grundsätzlich kann festgestellt werden, dass die Ergebnisse der Illumina-Sequenzierung wesentlich höhere Übereinstimmungen mit den Ergebnissen der EST-Sequenzierung zeigen. Aber auch hier kommt es in Einzelfällen zu größeren Unterschieden, die in diesem Fall möglicherweise auf Einflüsse der Transkriptlänge und/oder der Basenzusammensetzung zurückzuführen sein könnten. Eine generelle Normalisierung der Daten auf RPKM-Werte führte im Vergleich zum EST-Projekt zu wesentlich größeren Abweichungen mit zum Teil offensichtlich falschen Ergebnissen und wurde für diese Datensätze nicht weiter verwendet. Über die Ergebnisse des EST-Projekts hinaus konnten aufgrund der höheren Abdeckung in der Illumina-Sequenzierung auch regulatorische Zusammenhänge aufgeklärt werden. Dabei ist zu

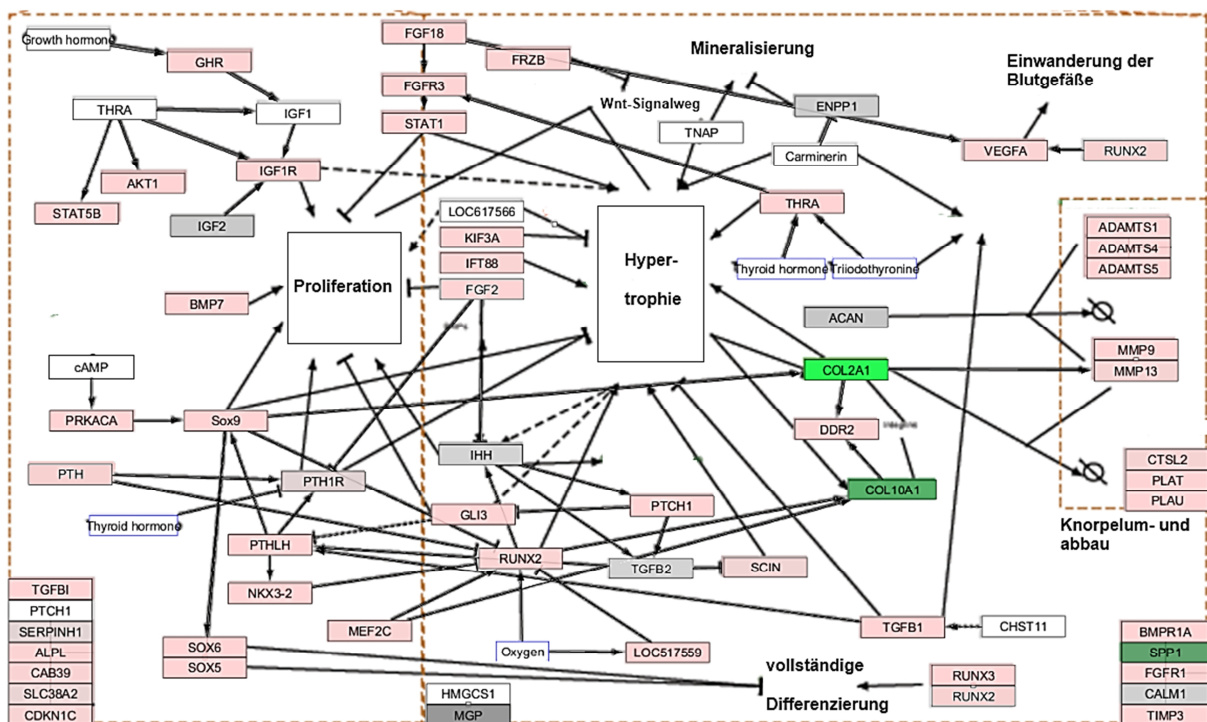


Abb. 45: Zentrale Aspekte der enchondralen Ossifikation

Dargestellt sind die Zusammenhänge der in der Wachstumsfuge exprimierten Gene; die Expression reicht von sehr stark (hell grün - dunkelgrün) über (dunkelgrau – hellgrau) mittlere Expression bis hin zu (rosa) geringe Expression und (weiss) – keine oder minimale Expression bis 10 „Reads“ (erstellt mit Hilfe von WikiPathways und GenMAPP-CS)

erkennen, dass nahezu alle bekannten an der enchondralen Ossifikation beteiligten Gene exprimiert werden. Obwohl eine leichte Tendenz zu einer verstärkten Expression der hypertrophen Zusammenhänge besteht, sind auch zentrale Gene anderer Stadien höher exprimiert (Abb. 45).

4.2 Analyse potentiell neuer Gene und Funktionen bei der Knochen- und Knorpelentwicklung

Über die Analyse hochspezifischer Gewebe besteht die Möglichkeit, die Expression eines bestimmten Gens in Verbindung mit dem Phänotyp oder, wie hier, mit dem physiologischen Vorgang zu bringen. Im Rahmen der Analyse der RNA aus der fötalen Wachstumsfuge wurden cDNA-Bibliotheken über EST-, 454/FLX- und Illumina-Sequenzierung ausgewertet. Nach der Annotation der Sequenzcluster oder Einzelsequenzen wurden uncharakterisierte Sequenzen oder Sequenz-Cluster näher untersucht. Im 454/FLX waren dies lediglich acht, von denen aber keines einem transkribierten Locus zugeordnet werden konnte. Besonders erschwert wurde die Suche nach neuen Transkripten durch die Kontamination mit *Vitis*-Sequenzen. Da nur Sequenzen mit pflanzlichen Übereinstimmungen gefiltert wurden, verblieben Kontaminationen ohne signifikante Blast-Treffer im Analysepool. Dies war zu erwarten, da gerade die Sequenzinformation von Pflanzen in den Datenbanken noch größere Lücken aufweist.

Im EST-Projekt wurden insgesamt 20 uncharakterisierte Sequenzen untersucht (Tab. 14). Diese beinhalteten die besonders interessanten Sequenzen von LOC618319 und F07P42. Über weitere bioinformatische Auswertungen konnten mehrere Cluster und Einzelklone einem Gen oder einer Funktion zugeordnet werden. So konnte bei LOC782657 eine Rhodanase-Domäne identifiziert werden und der Klon E12P38 als „solute carrier family 2“ (SLC2A13). Neben mitochondrialen Zuordnungen und Treffern zu reversen Transkriptasen, was auf ein repetitives Element hindeutet, konnten einige Sequenzen aus Mangel an Sequenzinformation überhaupt nicht zugeordnet werden. Contig 68 und 89 zeigen eindeutige Hinweise auf mRNA-Strukturen wie ein Polyadenylierungssignal und Polyadenylierung.

Leider konnte die Datenlage auch mit Unterstützung der Illumina-Sequenzierung nicht verbessert werden. Da diese EST-Sequenzen weniger stark exprimiert waren, wird nicht unbedingt von einem Knochen-/Knorpel-spezifischen Zusammenhang ausgegangen.

Ähnlich verhält sich die Analyse des Klon F07P42. Bei diesem Klon konnten zwar Exon-Intron-ähnliche Strukturen nachgewiesen werden, allerdings konnten weder funktionelle Einheiten noch zusätzliche Exonbereiche identifiziert werden. Auffällig ist die durchgehend große Homologie im betreffenden genomischen Bereich zwischen Mensch und Rind. Darüber hinaus sind keine signifikanten Übereinstimmungen nachweisbar.

LOC618319 – Gen(produkte) und mögliche Funktion

Als wesentlich vielversprechender erwies sich LOC618319. Dieser wurde im EST-Projekt mit 23 Sequenzen abgedeckt und war somit sehr hoch exprimiert. Spätere „Mappings“ mit den Illumina-Daten bestätigten die starke Expression des Locus ebenso wie das im 3'UTR

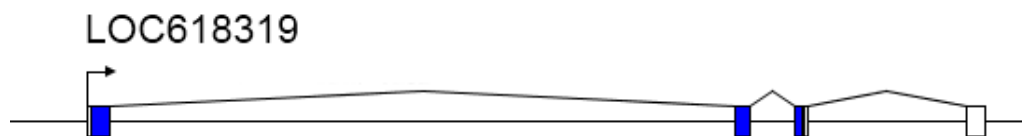


Abb. 46: Genomische Organisation des LOC618319 bei *Bos taurus*

liegende zusätzliche Exon 4 (3.4, Abb. 39). Dabei stellt sich Exon 4, obwohl 3' gelegen, analog zu den EST-Daten, als unterrepräsentiert dar. Daher kann davon ausgegangen werden, dass es sich um eine alternative und seltenere Variante handelt. Die 454-Sequenzierung unterstützt lediglich den Übergang von Exon 3 zu Exon 4. Das sich daraus ergebende modifizierte genomische Modell des LOC618319 ist nochmals in der folgenden Abbildung dargestellt. Der offene Leserahmen endet in Exon 3. Die Sequenz aus dem EST-Projekt konnte mit Hilfe der Illumina-Daten verifiziert werden und beinhaltet zwei aufeinanderfolgende Translationsstopps. Obwohl der Leserahmen durch das neue Exon 4 nicht verlängert wird, kann ein wesentlich veränderter und verlängerter 3'UTR vielfältige regulatorische Auswirkungen haben. So kann beispielsweise ein alternatives Polyadenylierungssignal eine veränderte Polyadenylierung bedingen und damit die Stabilität der mRNA direkt beeinflussen (Mitchell und Tollervey 2000, 2001). Darüber hinaus können

enthaltene Signalsequenzen Änderungen der Lokalisation bewirken (Jansen 2001). Allerdings konnten mittels verschiedener Vorhersageprogramme keine Hinweise auf eine solche 3' gelegene Signalsequenz gefunden werden. Eine weitere Möglichkeit ist die Beeinflussung der Stabilität verschiedener Transkripte sowie die Geschwindigkeit und Termination der Translation durch die Bindung sogenannter UTR-bindender Proteine (Copeland et al. 2000). Die Relevanz dieser Mechanismen im Fall von *LOC618319* lässt sich nicht abschätzen.

Für eine Knochen- und Knorpel-assoziierte Expression und Funktion spricht zunächst die hohe Expression in der Wachstumsfuge, die sowohl über die ESTs als auch die Illumina-Daten belegt ist. Das etwas niedrige Ranking im Illumina-Datensatz ist vermutlich auf das geringe Molekulargewicht des Transkripts zurückzuführen, das daher etwas unterrepräsentiert sein könnte. Obwohl das EST-Profil des homologen Gens beim Menschen den Nachweis in mehreren Geweben dokumentiert, konnte *LOC618319* bei RT-PCR-Vergleichen zwischen der Wachstumsfuge und verschiedenen Referenzgeweben, zusätzlich nur aus Herz amplifiziert werden. Die Menge lag nach semi-quantitativer Bestimmung allerdings deutlich unter dem Ergebnis der Wachstumsfuge. Ergänzend konnte durch den Vergleich zwischen fötaler Wachstumsfuge und einer MSC-Zellkultur eine um ein Vielfaches gesteigerte Expression im knochenbildenden Gewebe gegenüber seinen Vorläuferzellen festgestellt werden. Zusätzlich zur Relevanz für die frühe Knochen- und Knorpelentwicklung ist auch eine darüber hinausgehende Funktion möglich.

Bei Vergleichen zwischen den entsprechenden genomischen Regionen verschiedener Spezies finden sich auf Nukleotidebene lediglich im näheren evolutionären Umfeld hohe Übereinstimmungen. So zeigen beispielsweise homologe Sequenzen bei Mensch, Maus und Schimpanse zwischen 60-70% Übereinstimmungen in den Exon-Bereichen. Diese Werte fallen bei *Canis* oder *Gallus* schnell unter 50%.

Auf Aminosäureebene ist das Genprodukt bis hin zum Frosch nachweisbar konserviert. Obwohl in der Nukleotidsequenz einige Sequenzvariationen zwischen verschiedenen detektierten Sequenzen vorliegen, kann unabhängig davon ein möglicher Aufbau des Proteins mit drei funktionellen Einheiten abgeleitet werden. Dabei ist am N-Terminus ein

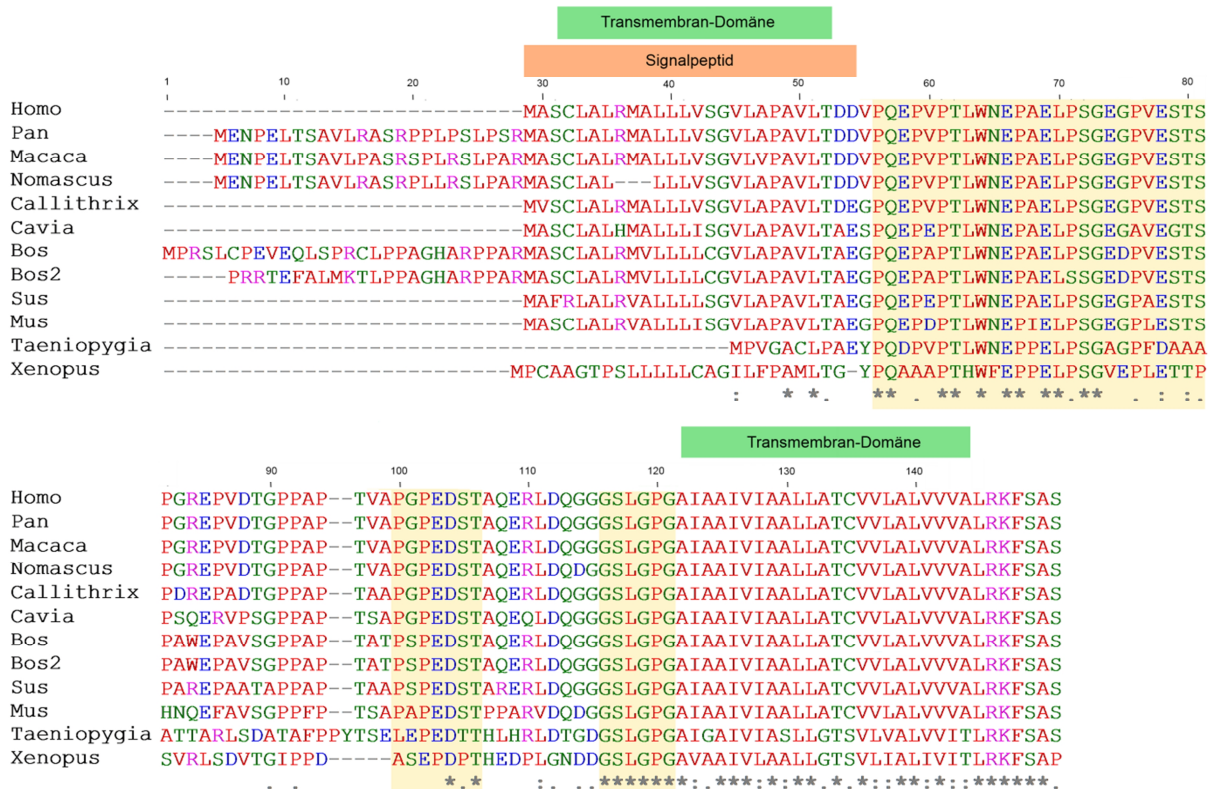


Abb. 47: Konservierte Bereiche im Genprodukt von *LOC618319*

Dargestellt ist ein ClustalW-Alignment (ClustalW2, EBI) von Proteinsequenzen 10 verschiedener Spezies, verglichen mit 2 möglichen Isoformen aus dem EST-Projekt der fötalen Wachstumsfuge des Rindes; hoch konservierte AS-Positionen sind mit * markiert; stark konservierte Bereiche sind nochmals farblich hervorgehoben (beige); oberhalb des Alignments sind nochmals die identifizierten Domänen markiert

potentielles Signalpeptid mit überlappender Transmembrandomäne zu erkennen. Bei der Untersuchung der N-terminalen Region des von der Nukleotidsequenz abgeleiteten Proteins mit den Vorhersageprogrammen SignalP (v3.0 und v4.0; Bendtsen et al. 2004; Petersen et al. 2011) und NetNES (v1.1; La Cour et al. 2004) wurden verschiedene Translationsstartpunkte angenommen (Abb. 48). Dabei steigt die Evidenz für ein Signalpeptid vom Beginn der Proteinsequenz über die möglichen Translationsstartpunkte bis hin zum vierten Methionin deutlich an. Daher muss diese Position als der wahrscheinlichste Translationsstart angenommen werden. Blast-Analysen und Recherchen in Signalpeptid-Datenbanken (Frank und Sippl 2008) zeigen Übereinstimmungen zu Transmembranproteinen der verschiedensten Kategorien. Außerdem bestätigt sich eine große Ähnlichkeit zum Signalpeptid von Phosphohippolin („FXYP domain-containing ion transport regulator 6 precursor“), der bei der Regulation von Ionenkanälen eine Rolle spielt.

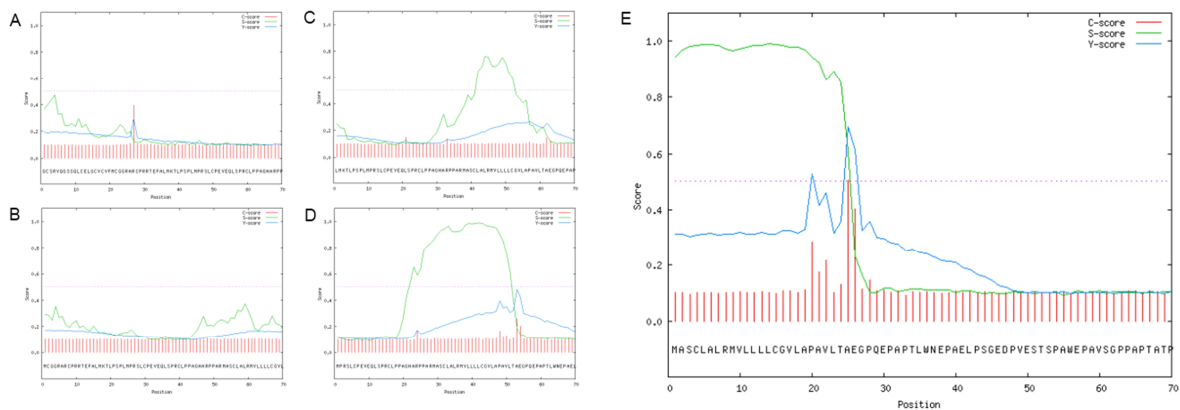


Abb. 48: LOC618319 – Signalpeptid

A-E zeigen mittels SignalP generierte Plots zur Vorhersage von Signalpeptiden, ausgehend von verschiedenen Translationsstarts; A nutzt die am weitesten 3' gelegene EST-Sequenz; B-E die jeweils eines der vier folgenden Methionine

Dementsprechend zeigt auch die N-terminal gelegene Transmembrandomäne (CLALRMVLLLLCGVLAPAVLT) Ähnlichkeiten zum „FXD domain-containing ion transport regulator 6 precursor“ (NP_001029571.1). Darüber hinaus konnten Übereinstimmungen zu einer möglichen Farnesyltransferase („putative protoheme IX farnesyltransferase“, XP_002411071.1) gefunden werden. Eine Aussage über eine mögliche Funktion des Proteins ist aufgrund dieser Tatsachen nicht möglich. Sie unterstützen lediglich die Eigenschaft des Proteins als Transmembranprotein. Dies bestätigt sich auch durch ein zweites Transmembranmotiv (AIAAIVIAALLATCVVLALVVVA), das nicht in Form des Signalpeptids abgespalten würde. Im Proteinalignment der Homologen verschiedener Spezies stellt sich die Domäne stark konserviert dar (Abb. 47). Allerdings sind Übereinstimmungen zu bereits charakterisierten Proteinen weniger signifikant.

Grundsätzlich ist die komplette Aminosäuresequenz innerhalb der Säugetiere sehr gut konserviert. Dennoch sind kleinere Bereiche besser erhalten (Abb. 47). Da nicht nur Domänen wichtig für die Funktion eines Proteins sind, wurde das abgeleitete Protein auf mögliche postranslationale Modifikationen untersucht. Dabei wurden besonders die hoch konservierten Regionen auf mögliche Stellen für Phosphorylierung, Acetylierung oder Glykosylierung untersucht. Dabei konnten keine signifikanten Hinweise auf eine Acetylierung festgestellt werden. Als mögliche Phosphorylierungsstellen mit einem Score-Wert von über 0,7 konnten die Serine an Position 72, 81, 101, 105, 149 der *Bos taurus*-Sequenz (Abb. 47)

identifiziert werden. Hervorzuheben sind die zwei, auf die Transmembrandomäne folgenden konservierten Serine, von denen eines eine der möglichen Phosphorylierungsstellen darstellt. Hinweise für die mögliche Phosphorylierung an einem Threonin-Rest konnten lediglich an Position 99 gefunden werden und sind ausschließlich auf die AS-Sequenzen aus *Bos taurus* beschränkt.

Für eine Glykosylierung ergeben sich zwei Möglichkeiten. Zum einen die Bindung des Oligosaccharids über eine N-glykosidische Bindung, dies würde im Endoplasmatischen Reticulum geschehen, oder über eine o-glykosidische Bindung an Serin, Threonin, Hydroxyprolin oder Hydroxylysin. Letzteres geschieht im Golgi-Apparat. Für eine N-

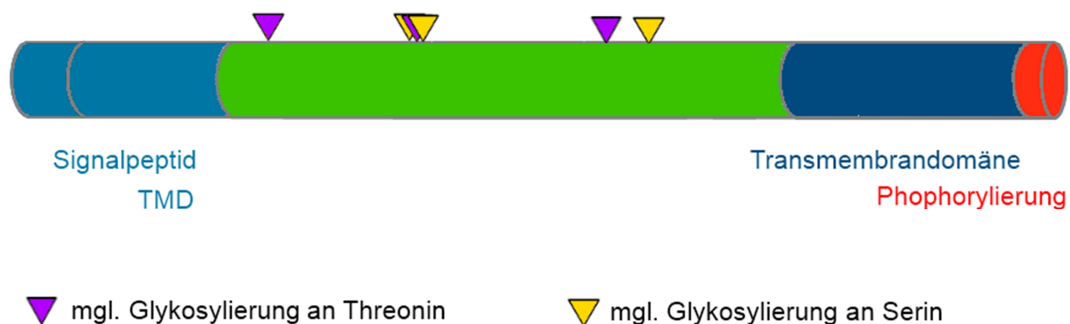


Abb. 49: Möglicher schematischer Aufbau des Proteins LOC618319

Glykosylierung konnten keine signifikanten Hinweise gefunden werden. Allerdings konnten gleich mehrere mögliche o-Glykosylierungsstellen identifiziert werden. Dabei handelt es sich um die bei allen untersuchten Mammalia konservierten Positionen 79-81 sowie 97 und 105, die Threonine 62 und 104 (Position 106 im Alignment) sowie Serin₇₂. Die Positionen 89, 99 und 101 sind lediglich in der *Bos taurus*-Sequenz vorhanden.

Geht man von einer Glykosylierung des Proteins aus, geschieht dies im Fall der o-Glykosylierung zunächst am rauhen endoplasmatischen Retikulum und abschließend im Golgi-Apparat. Dabei stellt die GalNAc α -O-Ser/Thr-Glykosylierung die häufigste Form im Tierreich dar (Pathak et al. 1988; Hanisch 2001). Für eine Translokation in das raue endoplasmatische Retikulum und damit eine posttranslationale Modifikation spricht das vorhandene Signalpeptid. Der hydrophobe und als Transmembrandomäne identifizierte Teil könnte dabei eine Rolle bei der Verankerung in der Membran spielen. Für die Glykosylierung gibt es

unterschiedliche Möglichkeiten. Dabei ist die o-Glykosylierung bei Proteinen der Knochen- und Knorpelentwicklung an mehreren Beispielen belegt (Christensen et al. 2005; Olstad et al. 1992). Speziell bei bovinen Chondrozyten konnte die Fähigkeit zur Glykosylierung von Proteinen mit Typ „Core 1“-Glycan (Gal β 1-3GalNAc α Ser/Thr) nachgewiesen werden. Humane Chondrozyten synthetisierten außerdem Typ „Core 2“-o- (GlcNAc β 1-6(Gal β 1-3)GalNAc α Ser/Thr) (Yang et al. 2007). Darüber hinaus sind auch sulfatierte C6 und/oder C4 der GalNAc möglich, eine Modifikation, die gerade bei glykosylierten Proteinen der ECM von Knochen- und Knorpelgewebe auftritt. Die modifizierten Proteine werden als Chondroitinsulfat-Proteoglykane bezeichnet (Hassell et al. 1986). Gleichzeitig konnte ein Zusammenhang zwischen TGF- β und der Glykosyltransferase-Aktivität nachgewiesen werden (Yang et al. 2007). Die weiter C-terminal gelegene Transmembrandomäne bietet auch nach der posttranslationalen Modifikation und Abspaltung des Signalpeptids die Möglichkeit einer Membranverankerung. Vorstellbar ist daher die Verankerung in der Zellmembran, mit nach außen gerichtetem, glykosylierten Rest-Peptid. Der zytoplasmatische Teil wäre auf wenige Aminosäuren mit einer konservierten Phosphorylierungsstelle beschränkt. Damit wäre LOC618319 in die Gruppe der membranständigen Proteoglykane einzuordnen.

Die Funktionen membranständiger Proteoglykane sind nicht zwangsläufig ausschließlich auf die Knochen- und Knorpelbildung beschränkt. Viele stellen wichtige Bestandteile von Mechanismen der Zelladhäsion, Signaltransduktion oder Strukturbildung in einer Vielzahl von Geweben und Entwicklungsstadien dar (Perrimon und Bernfield 2001). Dabei würde es sich beim Produkt von LOC618319 um ein sehr kleines Proteoglykan handeln. Es zeigt aber keinerlei Ähnlichkeit mit anderen kleinen Proteoglykanen wie CD44 oder der Syndecan-Familie. Außerdem fehlt die bei Transmembran-Proteoglykanen vorhandene PDZ-Domäne, die die Interaktion mit anderen im Zytoplasma befindlichen Proteinen erlaubt (Hung und Sheng 2002; Couchman 2010).

Daher muss das identifizierte Gen eventuell einer neuen, abweichenden Subpopulation zugeordnet werden. Eine direkte Signaltransduktions-Funktion ist aufgrund des kleinen zytoplasmatischen Anteils unwahrscheinlich. Möglich wäre eine Funktion bei der Zelladhäsion oder als Co-Rezeptor. Die Gewebeexpression in den EST-Profilen von Mensch und Maus zeigt ein eher ubiquitäres Vorkommen. Dabei sind die embryonalen Stadien

unterrepräsentiert. Dies macht das Zusammenspiel mit einem Wachstumsfaktor, bzw. den Rezeptoren von Wachstumsfaktoren, unwahrscheinlich.

Für die Regulation der Expression lassen sich aufgrund der identifizierten Promotorstrukturen mögliche Zusammenhänge festhalten:

Zunächst konnten keine Kernpromotorstrukturen wie beispielsweise eine TATA-Box identifiziert werden (Abb. 50). Die geschätzte Zahl der Gene, bei denen keine TATA-Box im Promotor zu finden ist, schwankt stark, kann aber bis zu 80 % betragen (Anish et al. 2009; Sandelin et al. 2007). Im Fall von LOC618319 könnte ein XCP-Element („X core promoter element“), ein eher selten vorkommendes, TBP-abhängiges, Kern-Promoter-Element diese Funktion übernehmen. Allerdings zeigen TATA-Box-lose Promotoren oft variable Transkriptionsstartpunkte. Dies könnte in dem vorliegenden Fall Einfluss auf die Bindung weiterer Transkriptionsfaktoren haben. Denn es lassen sich nicht nur im Bereich des XCPE mögliche interessante Bindestellen identifizieren, sondern auch ca. 50 bp downstream im Bereich der Übereinstimmung mit den EST-Sequenzen. Dort finden sich beispielsweise Bindestellen für STAT1 und SMAD4 (Abb. 50). Ebenso wie in der XCPE-Region finden sich hier Sp1-Bindestellen. So wäre ein Einfluss von TGF- β 2 während der Embryonalentwicklung über die MAP-Kinase-Kaskade denkbar. SMAD4 ist über den TGF- β -Weg direkt mit der Chondrocytendifferenzierung assoziiert und greift zentral in die Abläufe der BMP wie auch in die ID-Regulation ein. Dabei sind die „Inhibitor of DNA-binding“-Proteine (ID) wesentlich an der Osteoblasten-Differenzierung und der Mineralisierung der Wachstumszone beteiligt.

4.3 Next-Generation-Sequencing – Datenauswertung und Vergleichbarkeit

Voraussetzung für die Auswertung der NGS-Daten war die Etablierung adäquater Analyseverfahren. Diese sollen im Folgenden nochmals kurz zusammengefasst dargestellt werden und kritische Punkte näher beleuchtet werden. In den folgenden Ablaufschemata sind die Vorgehensweisen für die beiden NGS-Datensätze nochmals dargestellt. Für die

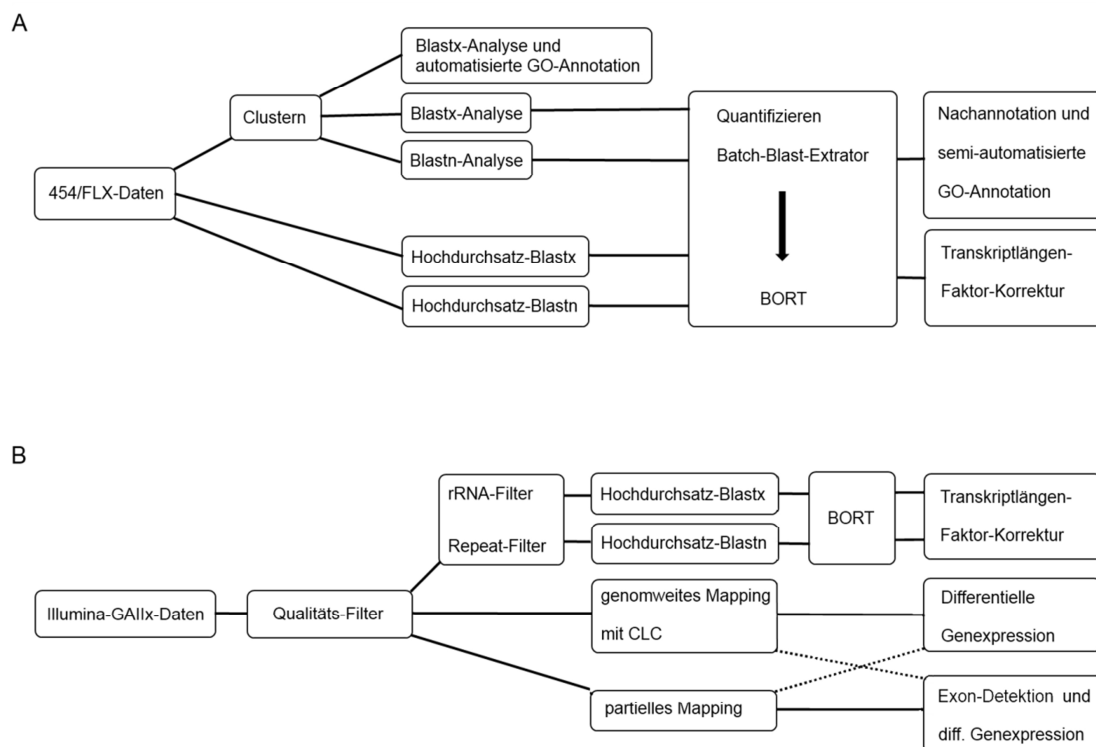


Abb. 51: Vorgehensweise bei der Analyse der NGS-Daten

A Vorgehen bei der Auswertung der 454/FLX-Daten; B Vorgehen bei der Auswertung der Illumina-Daten

Analyse der FLX-Daten wurden die Konsensus-Sequenzen der geclusterten Daten als auch Einzelsequenzen eingesetzt. Für die Assemblierung der Daten wurden verschiedene Plattformen verglichen. Durch den Einsatz neuer Algorithmen in Form der Assembler NGEN2 (DNSTAR Lasergene) und MIRA konnte die benötigte Zeit für das Clustern der Daten mit der zur Verfügung stehenden Hardware (38492 Sequenzen) von ca. 96 Stunden auf unter 24 h verkürzt werden. Die so generierten Konsensus-Sequenzen wurden sowohl für Blastx- als auch Blastn-Suchen eingesetzt. Dabei ist ein wesentlicher Bestandteil der automatisierten Auswertung die Extraktion wichtiger Informationen aus den umfangreichen Text-Datensätzen der Blast-Ergebnisse. Erste Daten wurden mit Hilfe des „Batch Blast Extractors“ (Pirooznia et al. 2008) verarbeitet. Aufgrund seiner eingeschränkten Einsatzmöglichkeiten

wurde die Plattform BORT („Blast Output Refinement Tool“, Benjamin Rieger, Hanno Schmidt / IMSB) etabliert. Mit Hilfe dieses umfangreichen Perl-Skripts können unter anderem Daten aus Blast-Ergebnissen extrahiert und gleiche Einträge aufsummiert, d.h. quantifiziert werden. Problematisch ist dabei, dass dies nur auf Ebene der genauen Blast-Treffer geschieht. Durch verschiedene Treffer von verschiedenen Sequenzen zu ein und demselben Gen entstehen redundante Positionen. In dieser Hinsicht wäre ein Zusammenführen der Einträge auf eine eindeutige Gen-Bezeichnung wie die Gene-ID oder das offizielle Gen-Symbol angebracht. Parallel dazu wurden verschiedene automatisierte Annotationsprogramme getestet. Das auf Blastx-Suchen basierende Java-Programm Blast2GO (Conesa et al. 2005; Götz et al. 2008) extrahiert aus Blastx-Suchen die notwendigen Informationen wie E-value und Trefferbeschreibung. Eine automatische Zuordnung von „Gene Ontology“-Gruppen zu den identifizierten Treffern ermöglicht die Bildung funktioneller Gruppen. Dabei ist zu beachten, dass nicht alle Sequenzen über Blastx-Suchen gegen Proteindatenbanken identifiziert werden können. Meist geschieht dies aufgrund eines zu niedrigen E-values. Größere Probleme könnten dabei ausgedehnte UTR-Bereiche sein, die in den Proteindatenbanken nicht vorhanden sind. Ergänzend dazu wurden daher Blastn-Suchen durchgeführt. Diese erlauben allerdings keine Auswertung im Programm Blast2GO. Alternativ wurden daher die Informationen über eigene Skripte extrahiert und die „Read“-Anzahl zu einzelnen Genen quantitativ erfasst. Hierbei bewährte sich eine einheitliche Benennung nach offiziellen Gen-Symbolen. Tabstopp-getrennte Listen von Gen-Symbolen mit oder ohne zugeordnete Expressionswerte können von den meisten weiterführenden Analyse-Tools interpretiert werden. Für die funktionelle Analyse solcher Expressionprofile stehen eine Vielzahl freier oder kommerzieller Anwendungen zur Verfügung. Diese eignen sich ähnlich des hier verwendeten DAVID, „Database for Annotation, Visualization and Integrated Discovery“ (Dennis et al. 2003; Huang et al. 2007; Huang et al. 2009), meist nur für die Annotation einer begrenzten Anzahl an Genen. Das Einordnen von exprimierten Genen in physiologische Vorgänge einschließlich der quantitativen Information kann mit Java-Applikationen wie GenMAPP-CS (Gladstone Institutes) durchgeführt werden. Die Erfahrungen aus der Analyse der 454-Sequenzierung wurden als Grundlage für die Auswertung der Illumina-Daten verwendet.

Allerdings erforderte die Analyse der ca. 75 Millionen Illumina- Reads in mancher Hinsicht eine andere Herangehensweise. Nicht alle der präanalytisch durchgeführten Filtermethoden können im Nachhinein als effektiv und praktikabel bezeichnet werden. Eine stringente Selektion nach Phred-Score (>28) und maximal einer zugelassenen unbekannte Basen erscheint gerade bei nachfolgenden Assemblierungsschritten sinnvoll. Bei „Read Count“-Ansätzen spielen kleinere Fehler eine geringere Rolle und es muss abgewägt werden, ob der relativ große Verlust an quantitativer Information notwendig ist. Ähnlich verhält es sich mit dem Filtern von Daten gegen spezifische Datenbanken mit unbrauchbaren Sequenzen, wie repetitive Elemente oder rRNA. Der Verlust einer großen Anzahl an Sequenzen deutet auf eine zu geringe Stringenz bei den Filtervorgängen hin. Die Gefahr spezifische Sequenzgruppen mit Relevanz zur Fragestellung zu entfernen steigt. Bei rRNA-Kontaminationen kann vielmehr schon bei der Probenaufbereitung auf eine adäquate PolyA⁺-Aufreinigung geachtet werden. Dieser Anteil war im bearbeiteten Datensatz ohnehin nur gering. Anders beim Filtern gegen alle im bovinen Genom enthaltenen repetitiven Sequenzen. Hier wurde ein sehr hoher Anteil (ca. 40 %) gefiltert. Aufgrund der Möglichkeit von Annotationsfehlern oder überlappenden Genbereichen sollte die Auswahl der Sequenzen in der eigens erstellten Datenbank nochmals überprüft werden.

Analog zu den 454-Daten wurden die Illumina-Daten im Hochdurchsatz-Blast sowohl gegen RNA-Datenbanken als auch Proteindatenbanken analysiert. Auch hier konnten Blastx – Analysen geringfügig weniger Sequenzen zugeordnet werden. Eine Clusteranalyse war zu diesem Zeitpunkt nicht möglich. Das Etablieren einer zuverlässigen Assemblierungs-Strategie muss allerdings eines der nächsten Ziele sein.

Alternativ wurden verschiedene „Mapping“-Ansätze verfolgt. Dabei werden die Sequenzen nicht untereinander verglichen sondern mit einer Referenz (Mortazavi et al. 2008). Diese kann einzelne Gene bis hin zu ganzen Genomen enthalten. Wichtig bei einer solchen Analyse ist die Zuordnung der Koordinaten zu funktionellen Einheiten, sprich eine Annotation. Die in dieser Arbeit verwendeten Verfahren beruhen zum einen auf einer kommerziellen Lösung, der „Genomics-Workbench“ der Firma CLC Bio und dem frei zugänglichen Programm PASS (Campagna et al. 2009). Mit der CLC-Plattform ist schnelles Ressourcen-schonendes Mapping gegen annotierte Genome im GBK-Format in relativ kurzer Zeit möglich. Damit ist sie vor

allem für eine schnelle Analyse mehrerer Datensätze, also für die Detektion von differentieller Genexpression besonders geeignet.

Alternativ wurde auf Grundlage der immer ähnlich aufgebauten Textausgaben frei zugänglicher Mapper, wie Bowtie, BWA oder eben auch PASS, ein eigenes, auf Perl-Skripten basierendes, Analyseverfahren entwickelt. Bisher wurde dies genutzt um Illumina-Sequenzen mit relativ kurzen genomischen Bereichen oder Transkripten zu vergleichen. Eine Weiterentwicklung sieht die Erweiterung dieser Bereiche und die Nutzung von Annotations-Information vor.

Gerade die Vielfalt möglicher Techniken zur Analyse kann es notwendig machen verschiedene Datensätze miteinander vergleichen zu müssen. Wie in schon vielen Studien und auch in dieser Arbeit gezeigt, ist dies nicht ohne weiteres möglich. Daher muss gerade für die differenzielle Genexpressionsanalyse zunächst eine einheitliche Probenbehandlung angestrebt werden. Sollte dies nicht möglich sein, sind mitunter umfangreiche Normalisierungsstrategien zu verfolgen.

Im konkreten Fall wurde dabei zunächst eine Normalisierung auf einheitliche Bibliotheksgrößen durchgeführt. Wie klar gezeigt werden konnte, ist dies jedoch nicht ausreichend. Gerade bei den neuen Sequenzieretechnologien können mehrere Faktoren die Anzahl generierter Sequenzen für ein Gen beeinflussen. Im Rahmen der einfachen Anpassungen stehen dabei die Transkriptlänge und der GC-Gehalt im Vordergrund (Dohm et al. 2008; Oshlack und Wakefield 2009). In der vorliegenden Arbeit konnte mit der Anpassung der Expressionswerte (später RPKM) aus der 454-Sequenzierung eine Annäherung an die EST-Daten erzielt werden. Im Gegensatz dazu zeigte das Expressionsprofil der Illumina-Daten wesentlich größere Abweichungen. Eine mögliche Ursache konnte durch die Untersuchung der Abdeckung einzelner Gene auf genomischer Ebene identifiziert werden. Wie aus der nachfolgenden Abbildung deutlich wird, zeigt die 454-Sequenzierung (Abb. 52 C) eine relativ gleichmäßige Abdeckung des Gens über alle Exonbereiche. Dagegen zeigt die Illumina-Sequenzierung (Abb. 52 B) einen deutlichen Gradienten in der Abdeckung von 3' zu 5' entsprechend der erwarteten Verteilung der cDNA-Fragmente (Abb. 52 A). Da die Normalisierung auf RPKM einen proportionalen Anstieg kompensiert, ergibt die auf die 454-Daten angewendete Normalisierung gute Ergebnisse. Da die Read-Abdeckung bei der

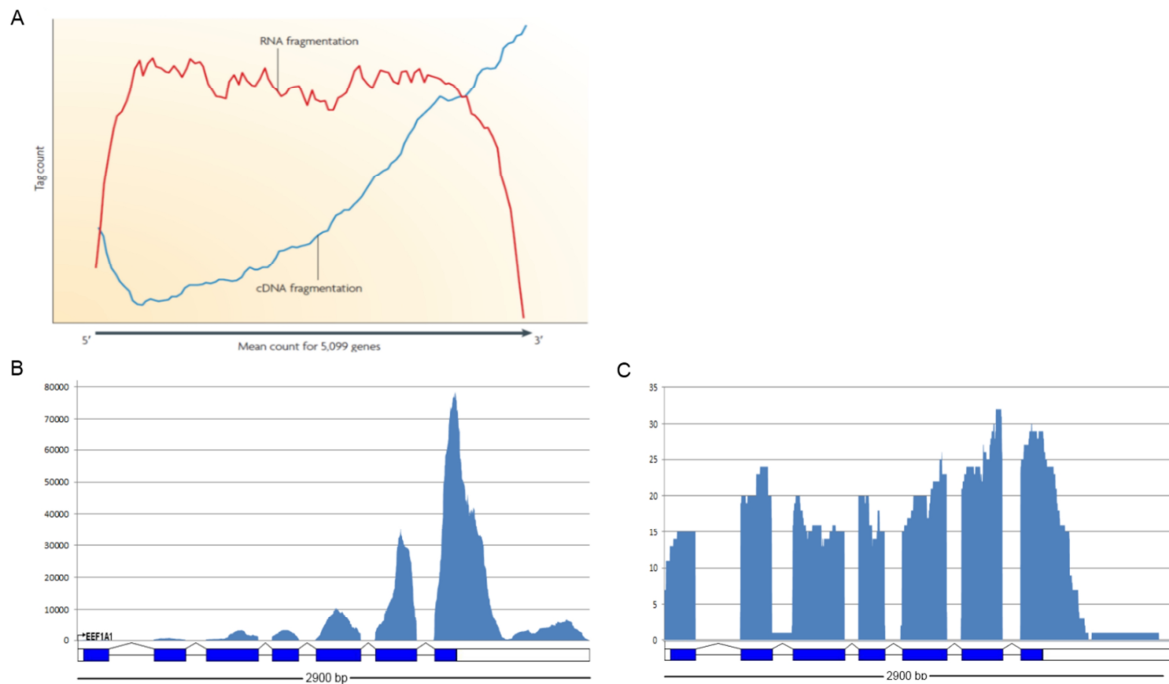


Abb. 52: Read-Abdeckung und Transkriptlängen-Fehler

A Größenverteilung von RNA und cDNA-Fragmenten nach Wang et al. 2009; B Abdeckung des genomischen EE1A1-Locus durch die Illumina-Daten; C Abdeckung des genomischen EE1A1-Locus durch die 454-Daten

Illumina-Sequenzierung nahezu exponentiell abnimmt, werden die Expressionswerte zu sehr nach oben korrigiert. Dies geht soweit, dass die absoluten Werte im Vergleich mit den EST-Daten bessere Übereinstimmungen liefern als die RPKM-normalisierten.

Neben der Entwicklung umfangreicher Formeln zum Angleichen der Datensätze konnte durch das Priming der cDNA-Synthese mit zufälligen Hexamer-Primern schon eine wesentlich gleichmäßigere Verteilung der Fragmente erreicht werden. Obwohl so die Problematik wesentlich abgemildert wurde, muss gerade beim Vergleich verschieden generierter Datensätze auf eine geeignete Normalisierung geachtet werden.

Daher muss vor allem bei der Analyse differentieller Genexpression, wie bei vielen anderen Methoden auch, auf eine einheitliche Behandlung der Proben und auch der Daten geachtet werden. Vor dem Vergleich verschieden generierter Datensätze empfiehlt es sich die „Read“-Verteilungen über die Transkriptlänge zu begutachten und eine geeignete Normalisierungsmethode zu wählen. Dies wird umso bedeutungsvoller je kleiner die untersuchten Expressionsstärken sind.

5 Zusammenfassung

Erkrankungen des Skelettsapparats wie beispielsweise die Osteoporose oder Arthrose gehören neben den Herz-Kreislaufkrankungen und Tumoren zu den häufigsten Erkrankungen des Menschen. Ein besseres Verständnis der Bildung und des Erhalts von Knochen- oder Knorpelgewebe ist deshalb von besonderer Bedeutung. Viele bisherige Ansätze zur Identifizierung hierfür relevanter Gene, deren Produkte und Interaktionen, beruhen auf der Untersuchung pathologischer Situationen. Daher ist die Funktion vieler Gene nur im Zusammenhang mit Krankheiten beschrieben. Untersuchungen, die die Genaktivität bei der Normalentwicklung von knochen- und knorpelbildenden Geweben zum Ziel haben sind dagegen weit weniger oft durchgeführt worden.

Eines der entwicklungsphysiologisch interessantesten Gewebe ist die Epiphysenfuge der Röhrenknochen. In dieser sogenannten Wachstumsfuge ist insbesondere beim fötalen Gewebe eine sehr hohe Aktivität derjenigen Gene zu erwarten, die an der Knochen- und Knorpelbildung beteiligt sind. In der vorliegenden Arbeit wurde daher aus der Epiphysenfuge von Kälberknochen RNA isoliert und eine cDNA-Bibliothek konstruiert. Von dieser wurden ca. 4000 Klone im Rahmen eines klassischen EST-Projekts sequenziert. Durch die Analyse konnte ein ungefähr 900 Gene umfassendes Expressionsprofil erstellt werden und viele Transkripte für Komponenten der regulatorischen und strukturbildenden Bestandteile der Knochen- und Knorpelentwicklung identifiziert werden. Zu diesen gehören in erster Linie die Kollagene, allen voran Kollagen II alpha 1, das mit Abstand höchst exprimierte Gen in der fötalen Wachstumsfuge. Nach den ribosomalen Proteinen stellen die Kollagene mit ca. 10 % aller auswertbaren Sequenzen die zweitgrößte Gengruppe im erstellten Expressionsprofil dar. Ebenfalls erwartet stark vertreten zeigten sich viele wichtige Komponenten der ECM wie beispielsweise SPP1, MGP oder COMP. Neben diesen typischen Genen für knochen- und knorpelbildende Gewebe sind auch deutlich Bestandteile embryonaler Entwicklungsprozesse vertreten. Hierzu gehören beispielsweise bekannte Wachstumsfaktoren der IGF- oder TGF-Familie aber auch weniger gut untersuchte Gen-Produkte wie H19, TPT1, CD109 oder CCDC80. Diese können nicht ausschließlich der frühen embryonalen Knochen- und Knorpelentwicklung zugeordnet werden, treten allerdings mitunter in ähnlichen Projekten

verstärkt auf. Aufgrund der geringen Abdeckung konnten im EST-Projekt niedrig exprimierte regulatorische Elemente, wie Transkriptionsfaktoren oder auch Proteoglykane nur in sehr geringer Kopienzahl gefunden werden. Allerdings förderte die EST-Analyse darüber hinaus mehrere interessante, bisher nicht bekannte Transkripte zutage, die detaillierter untersucht wurden. Dazu gehören Transkripte, die dem LOC618319 zugeordnet werden konnten. Diese stellen mit 23 ESTs zudem eine der höchst exprimierten Sequenzen dar. Neben den bisher beschriebenen drei Exonbereichen konnte ein weiteres Exon im 3'-UTR identifiziert werden. Im abgeleiteten Protein, das mindestens 121 AS lang ist, wurden ein Signalpeptid und eine Transmembrandomäne nachgewiesen. In Verbindung mit einer möglichen Glykosylierung ist das Genprodukt in die Gruppe der Proteoglykane einzuordnen. Leicht abweichend von den typischen Strukturen knochen- und knorpelspezifischer Proteoglykane ist eine mögliche Funktion dieses Genprodukts bei der Interaktion mit Integrinen und der Zell-Zellinteraktion, aber auch bei der Signaltransduktion denkbar.

EST-Sequenzierungen von ca. 4000 cDNA-Klonen können aber in der Regel nur einen Bruchteil der möglichen Transkripte des untersuchten Gewebes abdecken. Mit den neuen Sequenzieretechnologien des „Next Generation Sequencing“ bestehen völlig neue Möglichkeiten komplette Transkriptome mit sehr hoher Abdeckung zu sequenzieren und zu analysieren. Zur Unterstützung der EST-Daten und zur deutlichen Verbreiterung der Datenbasis wurde das Transkriptom der bovinen fötalen Wachstumsfuge sowohl mit Hilfe der Roche-454/FLX- als auch der Illumina-Solexa-Technologie sequenziert. Dabei wurden Verfahren zur allgemeinen Handhabung, der Qualitätskontrolle, dem „Clustern“, der Annotation und quantitativen Auswertung von großen Mengen an Sequenzdaten etabliert. Im Zuge dessen wurden an die 40000 454/FLX- und 75 Millionen Illumina-Sequenzen analysiert. Aufgrund dieser großen Datenmengen konnte die Zahl identifizierter Gene nach Hochdurchsatz-Blast-Analysen mit den 454/FLX-Sequenzen auf über 2200 gesteigert werden. Die Auswertung der Illumina-Daten ergab mehr als 40000 unterschiedliche Sequenz-Cluster, die vermutlich über 25000 Gene repräsentieren. Durch die verbesserte Datenlage konnten vor allem mit Hilfe der Illumina-Sequenzen die Ergebnisse des EST-Projekts besser abgesichert werden und darüber hinaus eine umfangreichere Analyse, vor allem der regulatorischen Elemente bei der Knochen- und Knorpelentwicklung, durchgeführt werden.

Auch hier ergibt sich das Bild eines hoch aktiven knochen- und knorpelbildenden Gewebes. Deutlich konnte eine vergleichbar hohe Expression zentraler regulatorischer Elemente wie IHH, TGF- β , IGF2, MGP SPP1, und PTH1R nachgewiesen werden. Insgesamt konnte beim Vergleich zwischen den im klassischen „Read-Count“-Ansatz durchgeführten NGS-Projekten mit dem erstellten EST-Expressionsprofil gute Übereinstimmungen gezeigt werden. Dies trifft vor allem für den Illumina-Datensatz zu. Abweichungen zwischen den einzelnen Methoden konnten nicht in allen Fällen methodisch erklärt werden. In einigen Fällen sind Korrelationen zwischen Transkriptlänge und „Read“-Verteilung zu erkennen. Obwohl schon simple Methoden wie die Normierung auf RPKM („reads per kilo base transkript per million mappable reads“) eine Verbesserung der Interpretation ermöglichen, konnten messtechnisch durch die Art der Sequenzierung bedingte systematische Fehler nicht immer ausgeräumt werden. Besonders wichtig ist daher die geeignete Normalisierung der Daten beim Vergleich verschieden generierter Datensätze.

Die hier diskutierten Ergebnisse aus den verschiedenen Analysen zeigen die neuen Sequenzieretechnologien als gute Ergänzung und potentiellen Ersatz für etablierte Methoden zur Genexpressionsanalyse.

6 Literaturverzeichnis

- 454 Life Sciences. Hg. v. © 1996-2010 Roche Diagnostics Corporation. Online verfügbar unter <http://www.454.com>.
- Abad, Veronica; Meyers, Jodi L.; Weise, Martina; Gafni, Rachel I.; Barnes, Kevin M.; Nilsson, Ola et al. (2002): The role of the resting zone in growth plate chondrogenesis. In: *Endocrinology* 143 (5), S. 1851–1857.
- Adachi-Yamada, T.; Nakamura, M.; Irie, K.; Tomoyasu, Y.; Sano, Y.; Mori, E. et al. (1999): p38 mitogen-activated protein kinase can be involved in transforming growth factor beta superfamily signal transduction in Drosophila wing morphogenesis. In: *Mol. Cell. Biol* 19 (3), S. 2322–2329.
- Adams, J. C.; Watt, F. M. (1993): Regulation of development and differentiation by the extracellular matrix. In: *Development* 117 (4), S. 1183–1198.
- Adams, M. D.; Kerlavage, A. R.; Fleischmann, R. D.; Fuldner, R. A.; Bult, C. J.; Lee, N. H. et al. (1995): Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. In: *Nature* 377 (6547 Suppl), S. 3–174.
- Adams, M. D.; Kelley, J. M.; Gocayne, J. D.; Dubnick, M.; Polymeropoulos, M. H.; Xiao, H. et al. (1991): Complementary DNA sequencing: expressed sequence tags and human genome project. In: *Science* 252 (5013), S. 1651–1656.
- Alkan, Can; Sajjadian, Saba; Eichler, Evan E. (2011): Limitations of next-generation genome sequence assembly. In: *Nat. Methods* 8 (1), S. 61–65.
- Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. (1990): Basic local alignment search tool. In: *J. Mol. Biol* 215 (3), S. 403–410.
- Alvarez, J.; Horton, J.; Sohn, P.; Serra, R. (2001): The perichondrium plays an important role in mediating the effects of TGF-beta1 on endochondral bone formation. In: *Dev. Dyn* 221 (3), S. 311–321.
- Alvarez, Jesus; Sohn, Philip; Zeng, Xin; Doetschman, Thomas; Robbins, David J.; Serra, Rosa (2002): TGFbeta2 mediates the effects of hedgehog on hypertrophic differentiation and PTHrP expression. In: *Development* 129 (8), S. 1913–1924.
- Alwine, J. C.; Kemp, D. J.; Parker, B. A.; Reiser, J.; Renart, J.; Stark, G. R.; Wahl, G. M. (1979): Detection of specific RNAs or specific fragments of DNA by fractionation in gels and transfer to diazobenzoyloxymethyl paper. In: *Meth. Enzymol* 68, S. 220–242.
- Amil, Benito; Fernandez-Fuente, Marta; Molinos, Ines; Rodriguez, Julian; Carbajo-Pérez, Eduardo; Garcia, Enrique et al. (2004): Chondromodulin-I expression in the growth plate of young uremic rats. In: *Kidney Int* 66 (1), S. 51–59.
- Anish, Ramakrishnan; Hossain, Mohammad B.; Jacobson, Raymond H.; Takada, Shinako (2009): Characterization of transcription from TATA-less promoters: identification of a

- new core promoter element XCPE2 and analysis of factor requirements. In: *PLoS ONE* 4 (4), S. e5103.
- Applied Biosystems. Hg. v. © 2010 Life Technologies. Online verfügbar unter <http://www.appliedbiosystems.com/>.
- Ayesh, Suhail; Matouk, Imad; Schneider, Tamar; Ohana, Patricia; Laster, Morris; Al-Sharef, Wasif et al. (2002): Possible physiological role of H19 RNA. In: *Mol. Carcinog* 35 (2), S. 63–74.
- Bailey, W. J.; Kim, J.; Wagner, G. P.; Ruddle, F. H. (1997): Phylogenetic reconstruction of vertebrate Hox cluster duplications. In: *Mol. Biol. Evol* 14 (8), S. 843–853.
- Ballock, R. T.; Heydemann, A.; Wakefield, L. M.; Flanders, K. C.; Roberts, A. B.; Sporn, M. B. (1993): TGF-beta 1 prevents hypertrophy of epiphyseal chondrocytes: regulation of gene expression for cartilage matrix proteins and metalloproteases. In: *Dev. Biol* 158 (2), S. 414–429.
- Ballock, R. Tracy; O'Keefe, Regis J. (2003): The biology of the growth plate. In: *J Bone Joint Surg Am* 85-A (4), S. 715–726.
- Bardack, D.; Zangerl, R. (1968): First fossil lamprey: a record from the Pennsylvanian of Illinois. In: *Science* 162 (859), S. 1265–1267.
- Barna, Maria; Niswander, Lee (2007): Visualization of cartilage formation: insight into cellular properties of skeletal progenitors and chondrodysplasia syndromes. In: *Dev. Cell* 12 (6), S. 931–941. Online verfügbar unter doi:10.1016/j.devcel.2007.04.016.
- Bartolomei, M. S.; Zemel, S.; Tilghman, S. M. (1991): Parental imprinting of the mouse H19 gene. In: *Nature* 351 (6322), S. 153–155.
- Beldade, Patrícia; Rudd, Stephen; Gruber, Jonathan D.; Long, Anthony D. (2006): A wing expressed sequence tag resource for *Bicyclus anynana* butterflies, an evo-devo model. In: *BMC Genomics* 7, S. 130.
- Belluoccio, Daniele; Bernardo, Bianca C.; Rowley, Lynn; Bateman, John F. (2008): A microarray approach for comparative expression profiling of the discrete maturation zones of mouse growth plate cartilage. In: *Biochim. Biophys. Acta* 1779 (5), S. 330–340.
- Bendtsen, Jannick Dyrlov; Nielsen, Henrik; Heijne, Gunnar von; Brunak, Søren (2004): Improved prediction of signal peptides: SignalP 3.0. In: *J. Mol. Biol* 340 (4), S. 783–795.
- Bentley, David R. (2006): Whole-genome re-sequencing. In: *Curr. Opin. Genet. Dev.* 16 (6), S. 545–552. Online verfügbar unter doi:10.1016/j.gde.2006.10.009.
- Bianco, P.; Cancedda, F. D.; Riminucci, M.; Cancedda, R. (1998): Bone formation via cartilage models: the "borderline" chondrocyte. In: *Matrix Biol.* 17 (3), S. 185–192.
- Bidanset, D. J.; Guidry, C.; Rosenberg, L. C.; Choi, H. U.; Timpl, R.; Hook, M. (1992): Binding of the proteoglycan decorin to collagen type VI. In: *J. Biol. Chem* 267 (8), S. 5250–5256.
- Binladen, Jonas; Gilbert, M. Thomas P.; Bollback, Jonathan P.; Panitz, Frank; Bendixen, Christian; Nielsen, Rasmus; Willerslev, Eske (2007): The use of coded PCR primers enables

- high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing. In: *PLoS ONE* 2 (2), S. e197.
- Birnboim, H. C.; Doly, J. (1979): A rapid alkaline extraction procedure for screening recombinant plasmid DNA. In: *Nucleic Acids Res* 7 (6), S. 1513–1523.
- Bitgood, M. J.; McMahon, A. P. (1995): Hedgehog and Bmp genes are coexpressed at many diverse sites of cell-cell interaction in the mouse embryo. In: *Dev. Biol.* 172 (1), S. 126–138. Online verfügbar unter doi:10.1006/dbio.1995.0010.
- Blaschke, U. K.; Eikenberry, E. F.; Hulmes, D. J.; Galla, H. J.; Bruckner, P. (2000): Collagen XI nucleates self-assembly and limits lateral growth of cartilage fibrils. In: *J. Biol. Chem* 275 (14), S. 10370–10378.
- Blin, N.; Stafford, D. W. (1976): A general method for isolation of high molecular weight DNA from eukaryotes. In: *Nucleic Acids Res* 3 (9), S. 2303–2308.
- Blochberger, T. C.; Vergnes, J. P.; Hempel, J.; Hassell, J. R. (1992): cDNA to chick lumican (corneal keratan sulfate proteoglycan) reveals homology to the small interstitial proteoglycan gene family and expression in muscle and intestine. In: *J. Biol. Chem* 267 (1), S. 347–352.
- Blom, N.; Gammeltoft, S.; Brunak, S. (1999): Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. In: *J. Mol. Biol* 294 (5), S. 1351–1362.
- Blom, Nikolaj; Sicheritz-Pontén, Thomas; Gupta, Ramneek; Gammeltoft, Steen; Brunak, Søren (2004): Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. In: *Proteomics* 4 (6), S. 1633–1649.
- Boesen, Christian C.; Motyka, Shawn A.; Patamawenu, Apisit; Sun, Peter D. (2002): Crystallization and preliminary crystallographic studies of human TGF-beta type II receptor ligand-binding domain. In: *Acta Crystallogr. D Biol. Crystallogr* 58 (Pt 7), S. 1214–1216.
- Boguski, M. S.; Tolstoshev, C. M.; Bassett, D. E. (1994): Gene discovery in dbEST. In: *Science* 265 (5181), S. 1993–1994.
- Bordo, Domenico; Bork, Peer (2002): The rhodanese/Cdc25 phosphatase superfamily. Sequence-structure-function relations. In: *EMBO Rep* 3 (8), S. 741–746.
- Boskey, A. L.; Spevak, L.; Paschalis, E.; Doty, S. B.; McKee, M. D. (2002): Osteopontin deficiency increases mineral content and mineral crystallinity in mouse bone. In: *Calcif. Tissue Int* 71 (2), S. 145–154.
- Boström, Kristina; Zebboudj, Amina F.; Yao, Yucheng; Lin, Than S.; Torres, Alejandra (2004): Matrix GLA protein stimulates VEGF expression through increased transforming growth factor-beta1 activity in endothelial cells. In: *J. Biol. Chem* 279 (51), S. 52904–52913.
- Brannan, C. I.; Dees, E. C.; Ingram, R. S.; Tilghman, S. M. (1990): The product of the H19 gene may function as an RNA. In: *Mol. Cell. Biol* 10 (1), S. 28–36.

- Bruckner, P.; Mendler, M.; Steinmann, B.; Huber, S.; Winterhalter, K. H. (1988): The structure of human collagen type IX and its organization in fetal and infant cartilage fibrils. In: *J. Biol. Chem* 263 (32), S. 16911–16917.
- Brudno, Michael; Do, Chuong B.; Cooper, Gregory M.; Kim, Michael F.; Davydov, Eugene; Green, Eric D. et al. (2003a): LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. In: *Genome Res* 13 (4), S. 721–731.
- Brudno, Michael; Malde, Sanket; Poliakov, Alexander; Do, Chuong B.; Couronne, Olivier; Dubchak, Inna; Batzoglou, Serafim (2003b): Glocal alignment: finding rearrangements during alignment. In: *Bioinformatics* 19 Suppl 1, S. i54-62.
- Bruijn, N. G. de (1946): A Combinatorial Problem. In: *Koninklijke Nederlandse Akademie v. Wetenschappen* (49), S. 758–764.
- Butler WT, Ridall AL McKee MD (1996): Osteopontin. In: Principles of bone biology.
- Campagna, Davide; Albiero, Alessandro; Bilardi, Alessandra; Caniato, Elisa; Forcato, Claudio; Manavski, Svetlin et al. (2009): PASS: a program to align short sequences. In: *Bioinformatics* 25 (7), S. 967–968.
- Canfield, A. E.; Farrington, C.; Dziobon, M. D.; Boot-Handford, R. P.; Heagerty, A. M.; Kumar, S. N.; Roberts, I. S. D. (2002): The involvement of matrix glycoproteins in vascular calcification and fibrosis: an immunohistochemical study. In: *J. Pathol* 196 (2), S. 228–234.
- Capdevila, J.; Johnson, R. L. (1998): Endogenous and ectopic expression of noggin suggests a conserved mechanism for regulation of BMP function during limb and somite patterning. In: *Dev. Biol* 197 (2), S. 205–217.
- Chen, L.; Adar, R.; Yang, X.; Monsonogo, E. O.; Li, C.; Hauschka, P. V. et al. (1999): Gly369Cys mutation in mouse FGFR3 causes achondroplasia by affecting both chondrogenesis and osteogenesis. In: *J. Clin. Invest* 104 (11), S. 1517–1525.
- Chen, Ligong; Klass, Carmen; Woods, Anne (2004): Syndecan-2 regulates transforming growth factor-beta signaling. In: *J. Biol. Chem* 279 (16), S. 15715–15718.
- Chevreur, Bastien; Pfisterer, Thomas; Drescher, Bernd; Driesel, Albert J.; Müller, Werner E. G.; Wetter, Thomas; Suhai, Sándor (2004): Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. In: *Genome Res.* 14 (6), S. 1147–1159. Online verfügbar unter doi:10.1101/gr.1917404.
- Chin, Y. E.; Kitagawa, M.; Su, W. C.; You, Z. H.; Iwamoto, Y.; Fu, X. Y. (1996): Cell growth arrest and induction of cyclin-dependent kinase inhibitor p21 WAF1/CIP1 mediated by STAT1. In: *Science* 272 (5262), S. 719–722.
- Chomczynski, P.; Sacchi, N. (1987): Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. In: *Anal. Biochem* 162 (1), S. 156–159.
- Christensen, Brian; Nielsen, Mette S.; Haselmann, Kim F.; Petersen, Torben E.; Sørensen, Esben S. (2005): Post-translationally modified residues of native human osteopontin are located in clusters: identification of 36 phosphorylation and five O-glycosylation sites and their biological implications. In: *Biochem. J* 390 (Pt 1), S. 285–292.

- Conesa, Ana; Götz, Stefan; García-Gómez, Juan Miguel; Terol, Javier; Talón, Manuel; Robles, Montserrat (2005): Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. In: *Bioinformatics* 21 (18), S. 3674–3676.
- Constância, Miguel; Hemberger, Myriam; Hughes, Jennifer; Dean, Wendy; Ferguson-Smith, Anne; Fundele, Reinald et al. (2002): Placental-specific IGF-II is a major modulator of placental and fetal growth. In: *Nature* 417 (6892), S. 945–948.
- Copeland, P. R.; Fletcher, J. E.; Carlson, B. A.; Hatfield, D. L.; Driscoll, D. M. (2000): A novel RNA binding protein, SBP2, is required for the translation of mammalian selenoprotein mRNAs. In: *EMBO J* 19 (2), S. 306–314.
- Couchman, John R. (2010): Transmembrane signaling proteoglycans. In: *Annu. Rev. Cell Dev. Biol* 26, S. 89–114.
- Crombrugge, B. de; Lefebvre, V.; Behringer, R. R.; Bi, W.; Murakami, S.; Huang, W. (2000): Transcriptional mechanisms of chondrocyte differentiation. In: *Matrix Biol.* 19 (5), S. 389–394.
- Crombrugge, B. de; Lefebvre, V.; Nakashima, K. (2001): Regulatory mechanisms in the pathways of cartilage and bone formation. In: *Curr. Opin. Cell Biol.* 13 (6), S. 721–727.
- Currey, J. D. (1967): The failure of exoskeletons and endoskeletons. In: *J. Morphol.* 123 (1), S. 1–16. Online verfügbar unter doi:10.1002/jmor.1051230102.
- Currey, J. D. (1984): What should bones be designed to do? In: *Calcif. Tissue Int.* 36 Suppl 1, S. S7-10.
- Dahl, Fredrik; Stenberg, Johan; Fredriksson, Simon; Welch, Katrina; Zhang, Michael; Nilsson, Mats et al. (2007): Multigene amplification and massively parallel sequencing for cancer mutation discovery. In: *Proc. Natl. Acad. Sci. U.S.A* 104 (22), S. 9387–9392.
- Dahlquist, Kam D.; Salomonis, Nathan; Vranizan, Karen; Lawlor, Steven C.; Conklin, Bruce R. (2002): GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. In: *Nat. Genet* 31 (1), S. 19–20.
- D'Angelo, M.; Billings, P. C.; Pacifici, M.; Leboy, P. S.; Kirsch, T. (2001): Authentic matrix vesicles contain active metalloproteases (MMP). a role for matrix vesicle-associated MMP-13 in activation of transforming growth factor-beta. In: *J. Biol. Chem* 276 (14), S. 11347–11353.
- Dehal, Paramvir; Boore, Jeffrey L. (2005): Two rounds of whole genome duplication in the ancestral vertebrate. In: *PLoS Biol* 3 (10), S. e314.
- Delidow, B. C.; Peluso, J. J.; White, B. A. (1989): Quantitative measurement of mRNAs by polymerase chain reaction. In: *Gene Anal. Tech* 6 (6), S. 120–124.
- DeLise, A. M.; Fischer, L.; Tuan, R. S. (2000): Cellular interactions and signaling in cartilage development. In: *Osteoarthr. Cartil.* 8 (5), S. 309–334. Online verfügbar unter doi:10.1053/joca.1999.0306.
- Deng, C.; Wynshaw-Boris, A.; Zhou, F.; Kuo, A.; Leder, P. (1996): Fibroblast growth factor receptor 3 is a negative regulator of bone growth. In: *Cell* 84 (6), S. 911–921.

- Denhardt, D. T.; Noda, M.; O'Regan, A. W.; Pavlin, D.; Berman, J. S. (2001): Osteopontin as a means to cope with environmental insults: regulation of inflammation, tissue remodeling, and cell survival. In: *J. Clin. Invest* 107 (9), S. 1055–1061.
- Denison, R. H. (1963): The early history of the vertebrate calcified skeleton. In: *Clin. Orthop. Relat. Res* 31, S. 141–152.
- Dennis, Glynn; Sherman, Brad T.; Hosack, Douglas A.; Yang, Jun; Gao, Wei; Lane, H. Clifford; Lempicki, Richard A. (2003): DAVID: Database for Annotation, Visualization, and Integrated Discovery. In: *Genome Biol* 4 (5), S. P3.
- Derynck, Rik; Zhang, Ying E. (2003): Smad-dependent and Smad-independent pathways in TGF-beta family signalling. In: *Nature* 425 (6958), S. 577–584.
- Doerge, K.; Sasaki, M.; Horigan, E.; Hassell, J. R.; Yamada, Y. (1987): Complete primary structure of the rat cartilage proteoglycan core protein deduced from cDNA clones. In: *J. Biol. Chem* 262 (36), S. 17757–17767.
- Doerge, K.; Hassell, J. R.; Caterson, B.; Yamada, Y. (1986): Link protein cDNA sequence reveals a tandemly repeated protein structure. In: *Proc. Natl. Acad. Sci. U.S.A* 83 (11), S. 3761–3765.
- Dohm, Juliane C.; Lottaz, Claudio; Borodina, Tatiana; Himmelbauer, Heinz (2008): Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. In: *Nucleic Acids Res* 36 (16), S. e105.
- Don, R. H.; Cox, P. T.; Wainwright, B. J.; Baker, K.; Mattick, J. S. (1991): 'Touchdown' PCR to circumvent spurious priming during gene amplification. In: *Nucleic Acids Res* 19 (14), S. 4008.
- Donoghue, Philip C. J.; Sansom, Ivan J. (2002): Origin and early evolution of vertebrate skeletonization. In: *Microsc. Res. Tech.* 59 (5), S. 352–372. Online verfügbar unter doi:10.1002/jemt.10217.
- Donoghue, Philip Conrad James; Sansom, Ivan James; Downs, Jason Philip (2006): Early evolution of vertebrate skeletal tissues and cellular interactions, and the canalization of skeletal development. In: *J. Exp. Zool. B Mol. Dev. Evol.* 306 (3), S. 278–294. Online verfügbar unter doi:10.1002/jez.b.21090.
- Dressman, Devin; Yan, Hai; Traverso, Giovanni; Kinzler, Kenneth W.; Vogelstein, Bert (2003): Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. In: *Proc. Natl. Acad. Sci. U.S.A* 100 (15), S. 8817–8822.
- Ebensperger, C.; Wilting, J.; Brand-Saberi, B.; Mizutani, Y.; Christ, B.; Balling, R.; Koseki, H. (1995): Pax-1, a regulator of sclerotome development is induced by notochord and floor plate signals in avian embryos. In: *Anat. Embryol.* 191 (4), S. 297–310.
- Edwards, Jeremy (2006): Small genomes and big science. In: *Genome Biol* 7 (3), S. 308.
- Elsik, Christine G.; Tellam, Ross L.; Worley, Kim C.; Gibbs, Richard A.; Muzny, Donna M.; Weinstock, George M. et al. (2009): The genome sequence of taurine cattle: a window to

- ruminant biology and evolution. In: *Science* 324 (5926), S. 522–528. Online verfügbar unter doi:10.1126/science.1169588.
- Emons, Joyce; Chagin, Andrei S.; Sävendahl, Lars; Karperien, Marcel; Wit, Jan M. (2011): Mechanisms of growth plate maturation and epiphyseal fusion. In: *Horm Res Paediatr* 75 (6), S. 383–391.
- Enomoto, H.; Enomoto-Iwamoto, M.; Iwamoto, M.; Nomura, S.; Himeno, M.; Kitamura, Y. et al. (2000): Cbfa1 is a positive regulatory factor in chondrocyte maturation. In: *J. Biol. Chem* 275 (12), S. 8695–8702.
- Erlebacher, A.; Filvaroff, E. H.; Gitelman, S. E.; Derynck, R. (1995): Toward a molecular understanding of skeletal development. In: *Cell* 80 (3), S. 371–378.
- Eyre, D. R.; Apon, S.; Wu, J. J.; Ericsson, L. H.; Walsh, K. A. (1987): Collagen type IX: evidence for covalent linkages to type II collagen in cartilage. In: *FEBS Lett* 220 (2), S. 337–341.
- Eyre, D. R.; Wu, J. J.; Fernandes, R. J.; Pietka, T. A.; Weis, M. A. (2002): Recent developments in cartilage research: matrix biology of the collagen II/IX/XI heterofibril network. In: *Biochem. Soc. Trans* 30 (Pt 6), S. 893–899.
- Fedurco, Milan; Romieu, Anthony; Williams, Scott; Lawrence, Isabelle; Turcatti, Gerardo (2006): BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. In: *Nucleic Acids Res* 34 (3), S. e22.
- Feng, Xin-Hua; Derynck, Rik (2005): Specificity and versatility in tgf-beta signaling through Smads. In: *Annu. Rev. Cell Dev. Biol* 21, S. 659–693.
- Ferguson, C. M.; Schwarz, E. M.; Reynolds, P. R.; Puzas, J. E.; Rosier, R. N.; O'Keefe, R. J. (2000): Smad2 and 3 mediate transforming growth factor-beta1-induced inhibition of chondrocyte maturation. In: *Endocrinology* 141 (12), S. 4728–4735.
- Finnsen, Kenneth W.; Tam, Betty Y. Y.; Liu, Kai; Marcoux, Anne; Lepage, Pierre; Roy, Stephane et al. (2006): Identification of CD109 as part of the TGF-beta receptor system in human keratinocytes. In: *FASEB J* 20 (9), S. 1525–1527.
- Fleischmajer, R.; Fisher, L. W.; MacDonald, E. D.; Jacobs, L.; Perlish, J. S.; Termine, J. D. (1991): Decorin interacts with fibrillar collagen of embryonic and adult human skin. In: *J. Struct. Biol* 106 (1), S. 82–90.
- Frank, Karl; Sippl, Manfred J. (2008): High-performance signal peptide prediction based on sequence alignment techniques. In: *Bioinformatics* 24 (19), S. 2172–2176.
- Frazer, Kelly A.; Pachter, Lior; Poliakov, Alexander; Rubin, Edward M.; Dubchak, Inna (2004): VISTA: computational tools for comparative genomics. In: *Nucleic Acids Res* 32 (Web Server issue), S. W273-9.
- Fukumoto, Takahiro; Levin, Michael (2005): Asymmetric expression of Syndecan-2 in early chick embryogenesis. In: *Gene Expr. Patterns* 5 (4), S. 525–528.
- Gerber, H. P.; Vu, T. H.; Ryan, A. M.; Kowalski, J.; Werb, Z.; Ferrara, N. (1999): VEGF couples hypertrophic cartilage remodeling, ossification and angiogenesis during endochondral

- bone formation. In: *Nat. Med.* 5 (6), S. 623–628. Online verfügbar unter doi:10.1038/9467.
- Gerber, H. P.; Ferrara, N. (2000): Angiogenesis and bone growth. In: *Trends Cardiovasc. Med.* 10 (5), S. 223–228.
- Gilbert, Scott F. (2006): *Developmental biology*. 8. Aufl. Sunderland, Mass: Sinauer. Online verfügbar unter <http://www.worldcat.org/oclc/314717877>.
- Girard, Angélique; Sachidanandam, Ravi; Hannon, Gregory J.; Carmell, Michelle A. (2006): A germline-specific class of small RNAs binds mammalian Piwi proteins. In: *Nature* 442 (7099), S. 199–202.
- Gleizes, P. E.; Munger, J. S.; Nunes, I.; Harpel, J. G.; Mazzieri, R.; Noguera, I.; Rifkin, D. B. (1997): TGF-beta latency: biological significance and mechanisms of activation. In: *Stem Cells* 15 (3), S. 190–197.
- Goldring, Mary B.; Tsuchimochi, Kaneyuki; Ijiri, Kosei (2006): The control of chondrogenesis. In: *J. Cell. Biochem* 97 (1), S. 33–44.
- Götz, Stefan; García-Gómez, Juan Miguel; Terol, Javier; Williams, Tim D.; Nagaraj, Shivashankar H.; Nueda, María José et al. (2008): High-throughput functional annotation and data mining with the Blast2GO suite. In: *Nucleic Acids Res* 36 (10), S. 3420–3435.
- Greenwald, Jason; Vega, Mark E.; Allendorph, George P.; Fischer, Wolfgang H.; Vale, Wylie; Choe, Senyon (2004): A flexible activin explains the membrane-dependent cooperative assembly of TGF-beta family receptors. In: *Mol. Cell* 15 (3), S. 485–489.
- Grisaru, S.; Cano-Gauci, D.; Tee, J.; Filmus, J.; Rosenblum, N. D. (2001): Glypican-3 modulates BMP- and FGF-mediated effects during renal branching morphogenesis. In: *Dev. Biol* 231 (1), S. 31–46.
- Hagiwara, H.; Schröter-Kermani, C.; Merker, H. J. (1993): Localization of collagen type VI in articular cartilage of young and adult mice. In: *Cell Tissue Res* 272 (1), S. 155–160.
- Hakes, Luke; Pinney, John W.; Lovell, Simon C.; Oliver, Stephen G.; Robertson, David L. (2007): All duplicates are not equal: the difference between small-scale and genome duplication. In: *Genome Biol* 8 (10), S. R209.
- Hanisch, F. G. (2001): O-glycosylation of the mucin type. In: *Biol. Chem* 382 (2), S. 143–149.
- Hansen, Kasper D.; Brenner, Steven E.; Dudoit, Sandrine (2010): Biases in Illumina transcriptome sequencing caused by random hexamer priming. In: *Nucleic Acids Res* 38 (12), S. e131.
- Har-el, R.; Sharma, Y. D.; Aguilera, A.; Ueyama, N.; Wu, J. J.; Eyre, D. R. et al. (1992): Cloning and developmental expression of the alpha 3 chain of chicken type IX collagen. In: *J. Biol. Chem* 267 (14), S. 10070–10076.
- Harrington, Christina A.; Winther, Michael; Garred, Michelle M. (2009): Use of bioanalyzer electropherograms for quality control and target evaluation in microarray expression profiling studies of ocular tissues. In: *J Ocul Biol Dis Infor* 2 (4), S. 243–249.

- Hashimoto, Mizuo; Ichihara, Masatoshi; Watanabe, Tsuyoshi; Kawai, Kumi; Koshikawa, Katsumi; Yuasa, Norihiro et al. (2004): Expression of CD109 in human cancer. In: *Oncogene* 23 (20), S. 3716–3720.
- Hassell, J. R.; Kimura, J. H.; Hascall, V. C. (1986): Proteoglycan core protein families. In: *Annu. Rev. Biochem* 55, S. 539–567.
- Hay, E. D. (1981): Extracellular matrix. In: *J. Cell Biol* 91 (3 Pt 2), S. 205s–223s.
- He, Xionglei; Zhang, Jianzhi (2005): Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. In: *Genetics* 169 (2), S. 1157–1164.
- Hecht, Jacqueline T.; Hayes, Elizabeth; Haynes, Richard; Cole, William G. (2005): COMP mutations, chondrocyte function and cartilage matrix. In: *Matrix Biol* 23 (8), S. 525–533.
- Hecht, Jochen; Stricker, Sigmar; Wiecha, Ulrike; Stiege, Asita; Panopoulou, Georgia; Podsiadlowski, Lars et al. (2008): Evolution of a core gene network for skeletogenesis in chordates. In: *PLoS Genet* 4 (3), S. e1000025.
- Hedbom, E.; Antonsson, P.; Hjerpe, A.; Aeschlimann, D.; Paulsson, M.; Rosa-Pimentel, E. et al. (1992): Cartilage matrix proteins. An acidic oligomeric protein (COMP) detected only in cartilage. In: *J. Biol. Chem* 267 (9), S. 6132–6136.
- Helms, J. A.; Schneider, R. A. (2003): Cranial skeletal biology. In: *Nature* 423 (6937), S. 326–331. Online verfügbar unter doi:10.1038/nature01656.
- Hinoi, Eiichi; Bialek, Peter; Chen, You-Tzung; Rached, Marie-Therese; Groner, Yoram; Behringer, Richard R. et al. (2006): Runx2 inhibits chondrocyte proliferation and hypertrophy through its expression in the perichondrium. In: *Genes Dev.* 20 (21), S. 2937–2942. Online verfügbar unter doi:10.1101/gad.1482906.
- Hollnagel, A.; Oehlmann, V.; Heymer, J.; Rütger, U.; Nordheim, A. (1999): Id genes are direct targets of bone morphogenetic protein induction in embryonic stem cells. In: *J. Biol. Chem* 274 (28), S. 19838–19845.
- Horner, David Stephen; Pavesi, Giulio; Castrignanò, Tiziana; Meo, Paolo D'Onorio de; Liuni, Sabino; Sammeth, Michael et al. (2010): Bioinformatics approaches for genomics and post genomics applications of next-generation sequencing. In: *Brief. Bioinformatics* 11 (2), S. 181–197.
- Horton, William A. (2003): Skeletal development: insights from targeting the mouse genome. In: *Lancet* 362 (9383), S. 560–569. Online verfügbar unter doi:10.1016/S0140-6736(03)14119-0.
- Hotz-Wagenblatt, Agnes; Hankeln, Thomas; Ernst, Peter; Glatting, Karl-Heinz; Schmidt, Erwin R.; Suhai, Sándor (2003): ESTAnnotator: A tool for high throughput EST annotation. In: *Nucleic Acids Res.* 31 (13), S. 3716–3719.
- Houston, B.; Seawright, E.; Jefferies, D.; Hoogland, E.; Lester, D.; Whitehead, C.; Farquharson, C. (1999): Identification and cloning of a novel phosphatase expressed at

- high levels in differentiating growth plate chondrocytes. In: *Biochim. Biophys. Acta* 1448 (3), S. 500–506.
- Houston, Brian; Stewart, Alan J.; Farquharson, Colin (2004): PHOSPHO1-A novel phosphatase specifically expressed at sites of mineralisation in bone and cartilage. In: *Bone* 34 (4), S. 629–637.
- Hua, Lin; Zheng, Weiyang; Liu, Hong; Lin, Hui; Gao, Lei (2008): [Relativity of gene expression and co-regulated gene patterns in feature KEGG pathways]. In: *Sheng Wu Gong Cheng Xue Bao* 24 (9), S. 1643–1648.
- Huang, Da Wei; Sherman, Brad T.; Tan, Qina; Collins, Jack R.; Alvord, W. Gregory; Roayaei, Jean et al. (2007): The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. In: *Genome Biol* 8 (9), S. R183.
- Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. *Nature Protoc.* 2009;4(1):44-57. [PubMed] (2009).
- Huang, Da Wei; Sherman, Brad T.; Zheng, Xin; Yang, Jun; Imamichi, Tomozumi; Stephens, Robert; Lempicki, Richard A. (2009): Extracting biological meaning from large gene lists with DAVID. In: *Curr Protoc Bioinformatics* Chapter 13, S. Unit 13.11.
- Hubbard, T.; Barker, D.; Birney, E.; Cameron, G.; Chen, Y.; Clark, L. et al. (2002): The Ensembl genome database project. In: *Nucleic Acids Res* 30 (1), S. 38–41.
- Hung, Albert Y.; Sheng, Morgan (2002): PDZ domains: structural modules for protein complex assembly. In: *J. Biol. Chem* 277 (8), S. 5699–5702.
- Hunziker, E. B. (1994): Mechanism of longitudinal bone growth and its regulation by growth plate chondrocytes. In: *Microsc. Res. Tech.* 28 (6), S. 505–519. Online verfügbar unter doi:10.1002/jemt.1070280606.
- Huson, Daniel H.; Auch, Alexander F.; Qi, Ji; Schuster, Stephan C. (2007): MEGAN analysis of metagenomic data. In: *Genome Res* 17 (3), S. 377–386.
- Huxley-Jones, Julie; Pinney, John W.; Archer, John; Robertson, David L.; Boot-Handford, Raymond P. (2009): Back to basics--how the evolution of the extracellular matrix underpinned vertebrate evolution. In: *Int J Exp Pathol* 90 (2), S. 95–100.
- Illumina, Inc. Hg. v. Inc © 2010 Illumina. Online verfügbar unter <http://www.illumina.com/>.
- Imelfort, Michael; Edwards, David (2009): De novo sequencing of plant genomes using second-generation technologies. In: *Brief. Bioinformatics* 10 (6), S. 609–618.
- Iozzo, R. V. (1998): Matrix proteoglycans: from molecular design to cellular function. In: *Annu. Rev. Biochem* 67, S. 609–652.
- Iozzo, R. V.; Murdoch, A. D. (1996): Proteoglycans of the extracellular environment: clues from the gene and protein side offer novel perspectives in molecular diversity and function. In: *FASEB J* 10 (5), S. 598–614.

- Iozzo, R. V.; Moscatello, D. K.; McQuillan, D. J.; Eichstetter, I. (1999): Decorin is a biological ligand for the epidermal growth factor receptor. In: *J. Biol. Chem* 274 (8), S. 4489–4492.
- Iozzo, Renato V.; Schaefer, Liliana (2010): Proteoglycans in health and disease: novel regulatory signaling mechanisms evoked by the small leucine-rich proteoglycans. In: *FEBS J* 277 (19), S. 3864–3875.
- Ito, H.; Akiyama, H.; Iguchi, H.; Iyama, K.; Miyamoto, M.; Ohsawa, K.; Nakamura, T. (2001): Molecular cloning and biological activity of a novel lysyl oxidase-related gene expressed in cartilage. In: *J. Biol. Chem* 276 (26), S. 24023–24029.
- Iyama, K.; Ninomiya, Y.; Olsen, B. R.; Linsenmayer, T. F.; Trelstad, R. L.; Hayashi, M. (1991): Spatiotemporal pattern of type X collagen gene expression and collagen deposition in embryonic chick vertebrae undergoing endochondral ossification. In: *Anat. Rec.* 229 (4), S. 462–472. Online verfügbar unter doi:10.1002/ar.1092290405.
- Jansen, R. P. (2001): mRNA localization: message on the move. In: *Nat. Rev. Mol. Cell Biol* 2 (4), S. 247–256.
- Jarvie, Thomas; Harkins, Timothy (2008): De novo assembly and genomic structural variation analysis with genome sequencer FLX 3K long-tag paired end reads. In: *BioTechniques* 44 (6), S. 829–831.
- Johnstone, E. W.; Leane, P. B.; Kolesik, P.; Byers, S.; Foster, B. K. (2000): Spatial arrangement of physeal cartilage chondrocytes and the structure of the primary spongiosa. In: *J Orthop Sci* 5 (3), S. 294–301.
- Julenius, Karin; Mølgaard, Anne; Gupta, Ramneek; Brunak, Søren (2005): Prediction, conservation analysis, and structural characterization of mammalian mucin-type O-glycosylation sites. In: *Glycobiology* 15 (2), S. 153–164.
- Jung, Youn-Kwan; Jeong, Jae-Hwan; Ryoo, Hyun-Mo; Kim, Hyun-Nam; Kim, Yeon-Ju; Park, Eui-Kyun et al. (2004): Gene expression profile of human chondrocyte HCS-2/8 cell line by EST sequencing analysis. In: *Gene* 330, S. 85–92.
- Kacian, D. L.; Mills, D. R.; Kramer, F. R.; Spiegelman, S. (1972): A replicating RNA molecule suitable for a detailed analysis of extracellular evolution and replication. In: *Proc. Natl. Acad. Sci. U.S.A* 69 (10), S. 3038–3042.
- Karolchik, D.; Baertsch, R.; Diekhans, M.; Furey, T. S.; Hinrichs, A.; Lu, Y. T. et al. (2003): The UCSC Genome Browser Database. In: *Nucleic Acids Res* 31 (1), S. 51–54.
- Karolchik, Donna; Hinrichs, Angie S.; Kent, W. James (2011): The UCSC Genome Browser. In: *Curr Protoc Hum Genet* Chapter 18, S. Unit18.6.
- Karp, S. J.; Schipani, E.; St-Jacques, B.; Hunzelman, J.; Kronenberg, H.; McMahon, A. P. (2000): Indian hedgehog coordinates endochondral bone growth and morphogenesis via parathyroid hormone related-protein-dependent and -independent pathways. In: *Development* 127 (3), S. 543–548.
- Karsenty, Gerard; Wagner, Erwin F. (2002): Reaching a genetic and molecular understanding of skeletal development. In: *Dev. Cell* 2 (4), S. 389–406.

- Karsenty, Gerard; Kronenberg, Henry M.; Settembre, Carmine (2009): Genetic control of bone formation. In: *Annu. Rev. Cell Dev. Biol.* 25, S. 629–648. Online verfügbar unter doi:10.1146/annurev.cellbio.042308.113308.
- Kawasaki, Kazuhiko; Weiss, Kenneth M. (2006): Evolutionary genetics of vertebrate tissue mineralization: the origin and evolution of the secretory calcium-binding phosphoprotein family. In: *J. Exp. Zool. B Mol. Dev. Evol* 306 (3), S. 295–316.
- Kawasaki, Kazuhiko; Suzuki, Tohru; Weiss, Kenneth M. (2004): Genetic basis for the evolution of vertebrate mineralized tissue. In: *Proc. Natl. Acad. Sci. U.S.A.* 101 (31), S. 11356–11361. Online verfügbar unter doi:10.1073/pnas.0404279101.
- Keller, Bettina; Yang, Tao; Chen, Yuqing; Munivez, Elda; Bertin, Terry; Zabel, Bernhard; Lee, Brendan (2011): Interaction of TGF β and BMP signaling pathways during chondrogenesis. In: *PLoS ONE* 6 (1), S. e16421.
- Kent, W. James (2002): BLAT--the BLAST-like alignment tool. In: *Genome Res* 12 (4), S. 656–664.
- Kiemer, Lars; Bendtsen, Jannick Dyrlov; Blom, Nikolaj (2005): NetAcet: prediction of N-terminal acetylation sites. In: *Bioinformatics* 21 (7), S. 1269–1270.
- Kirby, M. L.; Cheng, G.; Stadt, H.; Hunter, G. (1995): Differential expression of the L10 ribosomal protein during heart development. In: *Biochem. Biophys. Res. Commun.* 212 (2), S. 461–465.
- Kirsch, T.; Swoboda, B.; Nah, H. (2000a): Activation of annexin II and V expression, terminal differentiation, mineralization and apoptosis in human osteoarthritic cartilage. In: *Osteoarthr. Cartil* 8 (4), S. 294–302.
- Kirsch, T.; Harrison, G.; Golub, E. E.; Nah, H. D. (2000b): The roles of annexins and types II and X collagen in matrix vesicle-mediated mineralization of growth plate cartilage. In: *J. Biol. Chem* 275 (45), S. 35577–35583.
- Klass, C. M.; Couchman, J. R.; Woods, A. (2000): Control of extracellular matrix assembly by syndecan-2 proteoglycan. In: *J. Cell. Sci* 113 (Pt 3), S. 493–506.
- Kozarewa, Iwanka; Turner, Daniel J. (2011): Amplification-free library preparation for paired-end Illumina sequencing. In: *Methods Mol. Biol* 733, S. 257–266.
- Koziel, Lydia; Kunath, Melanie; Kelly, Olivia G.; Vortkamp, Andrea (2004): Ext1-dependent heparan sulfate regulates the range of Ihh signaling during endochondral ossification. In: *Dev. Cell* 6 (6), S. 801–813. Online verfügbar unter doi:10.1016/j.devcel.2004.05.009.
- Koziel, Lydia; Wuelling, Manuela; Schneider, Sabine; Vortkamp, Andrea (2005): Gli3 acts as a repressor downstream of Ihh in regulating two distinct steps of chondrocyte differentiation. In: *Development* 132 (23), S. 5249–5260. Online verfügbar unter doi:10.1242/dev.02097.
- Kronenberg, H. M.; Lee, K.; Lanske, B.; Segre, G. V. (1997): Parathyroid hormone-related protein and Indian hedgehog control the pace of cartilage differentiation. In: *J. Endocrinol* 154 Suppl, S. S39-45.

- Kwon, Hyuck Joon; Akimoto, Hidetoshi; Ohmiya, Yoshihiro; Honma, Kenichi; Yasuda, Kazunori (2008): Gene expression profile of rabbit cartilage by expressed sequence tag analysis. In: *Gene* 424 (1-2), S. 147–152.
- La Cour, Tanja; Kiemer, Lars; Mølgaard, Anne; Gupta, Ramneek; Skriver, Karen; Brunak, Søren (2004): Analysis and prediction of leucine-rich nuclear export signals. In: *Protein Eng. Des. Sel* 17 (6), S. 527–536.
- Langlands, K.; Yin, X.; Anand, G.; Prochownik, E. V. (1997): Differential interactions of Id proteins with basic-helix-loop-helix transcription factors. In: *J. Biol. Chem* 272 (32), S. 19785–19793.
- Langmead, Ben; Trapnell, Cole; Pop, Mihai; Salzberg, Steven L. (2009): Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. In: *Genome Biol* 10 (3), S. R25.
- Larkin, M. A.; Blackshields, G.; Brown, N. P.; Chenna, R.; McGettigan, P. A.; McWilliam, H. et al. (2007): Clustal W and Clustal X version 2.0. In: *Bioinformatics* 23 (21), S. 2947–2948.
- Larroux, Claire; Luke, Graham N.; Koopman, Peter; Rokhsar, Daniel S.; Shimeld, Sebastian M.; Degnan, Bernard M. (2008): Genesis and expansion of metazoan transcription factor gene classes. In: *Mol. Biol. Evol* 25 (5), S. 980–996.
- Leboy, P.; Grasso-Knight, G.; D'Angelo, M.; Volk, S. W.; Lian, J. V.; Drissi, H. et al. (2001): Smad-Runx interactions during chondrocyte maturation. In: *J Bone Joint Surg Am* 83-A Suppl 1 (Pt 1), S. S15-22.
- Lefebvre, V.; Li, P.; Crombrughe, B. de (1998): A new long form of Sox5 (L-Sox5), Sox6 and Sox9 are coexpressed in chondrogenesis and cooperatively activate the type II collagen gene. In: *EMBO J.* 17 (19), S. 5718–5733. Online verfügbar unter doi:10.1093/emboj/17.19.5718.
- Lefebvre, V.; Behringer, R. R.; Crombrughe, B. de (2001): L-Sox5, Sox6 and Sox9 control essential steps of the chondrocyte differentiation pathway. In: *Osteoarthr. Cartil.* 9 Suppl A, S. S69-75.
- Lewinson, D.; Toister, Z.; Silbermann, M. (1982): Quantitative and distributional changes in the activity of alkaline phosphatase during the maturation of cartilage. In: *J. Histochem. Cytochem* 30 (3), S. 261–269.
- Li, Jun; Jiang, Hui; Wong, Wing Hung (2010): Modeling non-uniformity in short-read rates in RNA-Seq data. In: *Genome Biol* 11 (5), S. R50.
- Linsenmayer, T. F.; Chen, Q. A.; Gibney, E.; Gordon, M. K.; Marchant, J. K.; Mayne, R.; Schmid, T. M. (1991): Collagen types IX and X in the developing chick tibiotarsus: analyses of mRNAs and proteins. In: *Development* 111 (1), S. 191–196.
- Lipshutz, R. J.; Fodor, S. P.; Gingeras, T. R.; Lockhart, D. J. (1999): High density synthetic oligonucleotide arrays. In: *Nat. Genet* 21 (1 Suppl), S. 20–24.
- Liu, Yi; Monticone, Massimiliano; Tonachini, Laura; Mastrogiacomo, Maddalena; Marigo, Valeria; Cancedda, Ranieri; Castagnola, Patrizio (2004): URB expression in human bone

- marrow stromal cells and during mouse development. In: *Biochem. Biophys. Res. Commun* 322 (2), S. 497–507.
- Liu, Ying; Lin, Yuanlie (2004): A novel method for N-terminal acetylation prediction. In: *Genomics Proteomics Bioinformatics* 2 (4), S. 253–255.
- Long, F.; Zhang, X. M.; Karp, S.; Yang, Y.; McMahon, A. P. (2001): Genetic manipulation of hedgehog signaling in the endochondral skeleton reveals a direct role in the regulation of chondrocyte proliferation. In: *Development* 128 (24), S. 5099–5108.
- Loots, Gabriela G.; Ovcharenko, Ivan; Pachter, Lior; Dubchak, Inna; Rubin, Edward M. (2002): rVista for comparative sequence-based discovery of functional transcription factor binding sites. In: *Genome Res* 12 (5), S. 832–839.
- Macias, D.; Gañan, Y.; Sampath, T. K.; Piedra, M. E.; Ros, M. A.; Hurle, J. M. (1997): Role of BMP-2 and OP-1 (BMP-7) in programmed cell death and skeletogenesis during chick limb development. In: *Development* 124 (6), S. 1109–1117.
- Maes, Christa; Stockmans, Ingrid; Moermans, Karen; van Looveren, Riet; Smets, Nico; Carmeliet, Peter et al. (2004): Soluble VEGF isoforms are essential for establishing epiphyseal vascularization and regulating chondrocyte development and survival. In: *J. Clin. Invest* 113 (2), S. 188–199.
- Mardis, Elaine R. (2008a): Next-generation DNA sequencing methods. In: *Annu Rev Genomics Hum Genet* 9, S. 387–402. Online verfügbar unter doi:10.1146/annurev.genom.9.081307.164359.
- Mardis, Elaine R. (2008b): The impact of next-generation sequencing technology on genetics. In: *Trends Genet.* 24 (3), S. 133–141. Online verfügbar unter doi:10.1016/j.tig.2007.12.007.
- Margulies, Marcel; Egholm, Michael; Altman, William E.; Attiya, Said; Bader, Joel S.; Bemben, Lisa A. et al. (2005): Genome sequencing in microfabricated high-density picolitre reactors. In: *Nature* 437 (7057), S. 376–380. Online verfügbar unter doi:10.1038/nature03959.
- Massagué, J. (2000): How cells read TGF-beta signals. In: *Nat. Rev. Mol. Cell Biol* 1 (3), S. 169–178.
- Massagué, Joan; Seoane, Joan; Wotton, David (2005): Smad transcription factors. In: *Genes Dev* 19 (23), S. 2783–2810.
- Masuda, H.; Fukumoto, M.; Hirayoshi, K.; Nagata, K. (1994): Coexpression of the collagen-binding stress protein HSP47 gene and the alpha 1(I) and alpha 1(III) collagen genes in carbon tetrachloride-induced rat liver fibrosis. In: *J. Clin. Invest* 94 (6), S. 2481–2488.
- McPherson, John D. (2009): Next-generation gap. In: *Nat. Methods* 6 (11 Suppl), S. S2–5.
- Merritt, Thomas M.; Alcorn, Joseph L.; Haynes, Richard; Hecht, Jacqueline T. (2006): Expression of mutant cartilage oligomeric matrix protein in human chondrocytes induces the pseudoachondroplasia phenotype. In: *J. Orthop. Res* 24 (4), S. 700–707.

- Meulemans, Daniel; Bronner-Fraser, Marianne (2007): Insights from amphioxus into the evolution of vertebrate cartilage. In: *PLoS ONE* 2 (8), S. e787.
- Michiels, L.; van der Rauwelaert, E.; van Hasselt, F.; Kas, K.; Merregaert, J. (1993): *fau* cDNA encodes a ubiquitin-like-S30 fusion protein and is expressed as an antisense sequence in the Finkel-Biskis-Reilly murine sarcoma virus. In: *Oncogene* 8 (9), S. 2537–2546.
- Minina, E.; Wenzel, H. M.; Kreschel, C.; Karp, S.; Gaffield, W.; McMahon, A. P.; Vortkamp, A. (2001): BMP and *Ihh*/PTHrP signaling interact to coordinate chondrocyte proliferation and differentiation. In: *Development* 128 (22), S. 4523–4534.
- Minina, Eleonora; Kreschel, Conny; Naski, Michael C.; Ornitz, David M.; Vortkamp, Andrea (2002): Interaction of FGF, *Ihh*/Pthlh, and BMP signaling integrates chondrocyte proliferation and hypertrophic differentiation. In: *Dev. Cell* 3 (3), S. 439–449.
- Mitchell, P.; Tollervey, D. (2001): mRNA turnover. In: *Curr. Opin. Cell Biol* 13 (3), S. 320–325.
- Mitchell, P.; Tollervey, D. (2000): mRNA stability in eukaryotes. In: *Curr. Opin. Genet. Dev* 10 (2), S. 193–198.
- Mitchell, T. I.; Jeffrey, J. J.; Palmiter, R. D.; Brinckerhoff, C. E. (1993): The acute phase reactant serum amyloid A (SAA3) is a novel substrate for degradation by the metalloproteinases collagenase and stromelysin. In: *Biochim. Biophys. Acta* 1156 (3), S. 245–254.
- Mitchell, T. I.; Coon, C. I.; Brinckerhoff, C. E. (1991): Serum amyloid A (SAA3) produced by rabbit synovial fibroblasts treated with phorbol esters or interleukin 1 induces synthesis of collagenase and is neutralized with specific antiserum. In: *J. Clin. Invest* 87 (4), S. 1177–1185.
- Miyazono, K.; Kusanagi, K.; Inoue, H. (2001): Divergence and convergence of TGF-beta/BMP signaling. In: *J. Cell. Physiol* 187 (3), S. 265–276.
- Mörgelin, M.; Paulsson, M.; Hardingham, T. E.; Heinegård, D.; Engel, J. (1988): Cartilage proteoglycans. Assembly with hyaluronate and link protein as studied by electron microscopy. In: *Biochem. J* 253 (1), S. 175–185.
- Morozova, Olena; Hirst, Martin; Marra, Marco A. (2009): Applications of new sequencing technologies for transcriptome analysis. In: *Annu Rev Genomics Hum Genet* 10, S. 135–151.
- Mortazavi, Ali; Williams, Brian A.; McCue, Kenneth; Schaeffer, Lorian; Wold, Barbara (2008): Mapping and quantifying mammalian transcriptomes by RNA-Seq. In: *Nat. Methods* 5 (7), S. 621–628.
- Morvan-Dubois, Ghislaine; Le Guellec, Dominique; Garrone, Robert; Zylberberg, Louise; Bonnaud, Laure (2003): Phylogenetic analysis of vertebrate fibrillar collagen locates the position of zebrafish alpha3(I) and suggests an evolutionary link between collagen alpha chains and hox clusters. In: *J. Mol. Evol* 57 (5), S. 501–514.

- Mundlos, S.; Engel, H.; Michel-Behnke, I.; Zabel, B. (1990): Distribution of type I and type II collagen gene expression during the development of human long bones. In: *Bone* 11 (4), S. 275–279.
- Murakami, S.; Kan, M.; McKeehan, W. L.; Crombrughe, B. de (2000): Up-regulation of the chondrogenic Sox9 gene by fibroblast growth factors is mediated by the mitogen-activated protein kinase pathway. In: *Proc. Natl. Acad. Sci. U.S.A* 97 (3), S. 1113–1118.
- Mwale, Fackson; Tchetina, Elena; Wu, C. William; Poole, A. Robin (2002): The assembly and remodeling of the extracellular matrix in the growth plate in relationship to mineral deposition and cellular hypertrophy: an in situ study of collagens II and IX and proteoglycan. In: *J. Bone Miner. Res* 17 (2), S. 275–283.
- Nakamura, K.; Shirai, T.; Morishita, S.; Uchida, S.; Saeki-Miura, K.; Makishima, F. (1999): p38 mitogen-activated protein kinase functionally contributes to chondrogenesis induced by growth/differentiation factor-5 in ATDC5 cells. In: *Exp. Cell Res* 250 (2), S. 351–363.
- Naski, M. C.; Colvin, J. S.; Coffin, J. D.; Ornitz, D. M. (1998): Repression of hedgehog signaling and BMP4 expression in growth plate cartilage by fibroblast growth factor receptor 3. In: *Development* 125 (24), S. 4977–4988.
- Ng, Patrick; Tan, Jack J. S.; Ooi, Hong Sain; Lee, Yen Ling; Chiu, Kuo Ping; Fullwood, Melissa J. et al. (2006): Multiplex sequencing of paired-end ditags (MS-PET): a strategy for the ultra-high-throughput analysis of transcriptomes and genomes. In: *Nucleic Acids Res* 34 (12), S. e84.
- Ng, Patrick; Wei, Chia-Lin; Ruan, Yijun (2007): Paired-end diTagging for transcriptome and genome analysis. In: *Curr Protoc Mol Biol* Chapter 21, S. Unit 21.12.
- Nickerson, D. A.; Kolker, N.; Taylor, S. L.; Rieder, M. J. (2001): Sequence-based detection of single nucleotide polymorphisms. In: *Methods Mol. Biol* 175, S. 29–35.
- Nilsson, Ola; Baron, Jeffrey (2004): Fundamental limits on longitudinal bone growth: growth plate senescence and epiphyseal fusion. In: *Trends Endocrinol. Metab* 15 (8), S. 370–374.
- O'Keefe, R. J.; Loveys, L. S.; Hicks, D. G.; Reynolds, P. R.; Crabb, I. D.; Puzas, J. E.; Rosier, R. N. (1997): Differential regulation of type-II and type-X collagen synthesis by parathyroid hormone-related protein in chick growth-plate chondrocytes. In: *J. Orthop. Res* 15 (2), S. 162–174.
- Olstad, O. K.; Reppe, S.; Gabrielsen, O. S.; Hartmanis, M.; Blingsmo, O. R.; Gautvik, V. T. et al. (1992): Isolation and characterization of two biologically active O-glycosylated forms of human parathyroid hormone produced in *Saccharomyces cerevisiae*. Identification of a new motif for O-glycosylation. In: *Eur. J. Biochem* 205 (1), S. 311–319.
- Ornitz, David M. (2005): FGF signaling in the developing endochondral skeleton. In: *Cytokine Growth Factor Rev* 16 (2), S. 205–213.
- Ornitz, David M.; Marie, Pierre J. (2002): FGF signaling pathways in endochondral and intramembranous bone development and human genetic disease. In: *Genes Dev* 16 (12), S. 1446–1465.

- Ortega, Nathalie; Behonick, Danielle J.; Werb, Zena (2004): Matrix remodeling during endochondral ossification. In: *Trends Cell Biol* 14 (2), S. 86–93.
- Oshlack, Alicia; Wakefield, Matthew J. (2009): Transcript length bias in RNA-seq data confounds systems biology. In: *Biol. Direct* 4, S. 14.
- Ovcharenko, Ivan; Loots, Gabriela G.; Hardison, Ross C.; Miller, Webb; Stubbs, Lisa (2004): zPicture: dynamic alignment and visualization tool for analyzing conservation profiles. In: *Genome Res* 14 (3), S. 472–477.
- Owen-Hughes, Tom; Engeholm, Maik (2007): Pyrosequencing positions nucleosomes precisely. In: *Genome Biol.* 8 (6), S. 217. Online verfügbar unter doi:10.1186/gb-2007-8-6-217.
- Panopoulou, Georgia; Poustka, Albert J. (2005): Timing and mechanism of ancient vertebrate genome duplications -- the adventure of a hypothesis. In: *Trends Genet* 21 (10), S. 559–567.
- Park, W. J.; Bellus, G. A.; Jabs, E. W. (1995): Mutations in fibroblast growth factor receptors: phenotypic consequences during eukaryotic development. In: *Am. J. Hum. Genet* 57 (4), S. 748–754.
- Pateder, D. B.; Rosier, R. N.; Schwarz, E. M.; Reynolds, P. R.; Puzas, J. E.; D'Souza, M.; O'Keefe, R. J. (2000): PTHrP expression in chondrocytes, regulation by TGF-beta, and interactions between epiphyseal and growth plate chondrocytes. In: *Exp. Cell Res* 256 (2), S. 555–562.
- Pateder, D. B.; Ferguson, C. M.; Ionescu, A. M.; Schwarz, E. M.; Rosier, R. N.; Puzas, J. E.; O'Keefe, R. J. (2001): PTHrP expression in chick sternal chondrocytes is regulated by TGF-beta through Smad-mediated signaling. In: *J. Cell. Physiol* 188 (3), S. 343–351.
- Pathak, R. K.; Merkle, R. K.; Cummings, R. D.; Goldstein, J. L.; Brown, M. S.; Anderson, R. G. (1988): Immunocytochemical localization of mutant low density lipoprotein receptors that fail to reach the Golgi complex. In: *J. Cell Biol* 106 (6), S. 1831–1841.
- Pathi, S.; Rutenberg, J. B.; Johnson, R. L.; Vortkamp, A. (1999): Interaction of Ihh and BMP/Noggin signaling during cartilage differentiation. In: *Dev. Biol* 209 (2), S. 239–253.
- Patthy, László (2003): Modular assembly of genes and the evolution of new functions. In: *Genetica* 118 (2-3), S. 217–231.
- Pedrozo, H. A.; Schwartz, Z.; Gomez, R.; Ornoy, A.; Xin-Sheng, W.; Dallas, S. L. et al. (1998): Growth plate chondrocytes store latent transforming growth factor (TGF)-beta 1 in their matrix through latent TGF-beta 1 binding protein-1. In: *J. Cell. Physiol* 177 (2), S. 343–354.
- Perrimon, N.; Bernfield, M. (2001): Cellular functions of proteoglycans--an overview. In: *Semin. Cell Dev. Biol* 12 (2), S. 65–67.
- Peters, K.; Ornitz, D.; Werner, S.; Williams, L. (1993): Unique expression pattern of the FGF receptor 3 gene during mouse organogenesis. In: *Dev. Biol* 155 (2), S. 423–430.

- Petersen, Thomas Nordahl; Brunak, Søren; Heijne, Gunnar von; Nielsen, Henrik (2011): SignalP 4.0: discriminating signal peptides from transmembrane regions. In: *Nat. Methods* 8 (10), S. 785–786.
- Pevzner, P. A.; Tang, H.; Waterman, M. S. (2001): An Eulerian path approach to DNA fragment assembly. In: *Proc. Natl. Acad. Sci. U.S.A* 98 (17), S. 9748–9753.
- Pickard, Mark R.; Mourtada-Maarabouni, Mirna; Williams, Gwyn T. (2011): Candidate tumour suppressor Fau regulates apoptosis in human cells: an essential role for Bcl-G. In: *Biochim. Biophys. Acta* 1812 (9), S. 1146–1153.
- Pickard, Mark R.; Edwards, Sandra E.; Cooper, Colin S.; Williams, Gwyn T. (2010): Apoptosis regulators Fau and Bcl-G are down-regulated in prostate cancer. In: *Prostate* 70 (14), S. 1513–1523.
- Pico, Alexander R.; Kelder, Thomas; van Iersel, Martijn P.; Hanspers, Kristina; Conklin, Bruce R.; Evelo, Chris (2008): WikiPathways: Pathway Editing for the People. In: *PLoS Biol* 6 (7), S. e184.
- Pirooznia, Mehdi; Perkins, Edward J.; Deng, Youping (2008): Batch Blast Extractor: an automated blastx parser application. In: *BMC Genomics* 9 Suppl 2, S. S10.
- Pizette, S.; Niswander, L. (2001): Early steps in limb patterning and chondrogenesis. In: *Novartis Found. Symp.* 232, S. 23-36; discussion 36-46.
- Pizette, S.; Abate-Shen, C.; Niswander, L. (2001): BMP controls proximodistal outgrowth, via induction of the apical ectodermal ridge, and dorsoventral patterning in the vertebrate limb. In: *Development* 128 (22), S. 4463–4474.
- Pogue, Robert; Sebald, Eiman; King, Lily; Kronstadt, Erik; Krakow, Deborah; Cohn, Daniel H. (2004): A transcriptional profile of human fetal cartilage. In: *Matrix Biol* 23 (5), S. 299–307.
- Pringle, G. A.; Dodd, C. M. (1990): Immunoelectron microscopic localization of the core protein of decorin near the d and e bands of tendon collagen fibrils by use of monoclonal antibodies. In: *J. Histochem. Cytochem* 38 (10), S. 1405–1411.
- Quevillon, E.; Silventoinen, V.; Pillai, S.; Harte, N.; Mulder, N.; Apweiler, R.; Lopez, R. (2005): InterProScan: protein domains identifier. In: *Nucleic Acids Res* 33 (Web Server issue), S. W116-20.
- Rebeiz, Mark; Posakony, James W. (2004): GenePalette: a universal software tool for genome sequence visualization and analysis. In: *Dev. Biol* 271 (2), S. 431–438.
- Rohde, W.; Sängler, H. L. (1981): Detection of complementary RNA intermediates of viroid replication by Northern blot hybridization. In: *Biosci. Rep* 1 (4), S. 327–336.
- Ronaghi, M.; Karamohamed, S.; Pettersson, B.; Uhlén, M.; Nyrén, P. (1996): Real-time DNA sequencing using detection of pyrophosphate release. In: *Anal. Biochem* 242 (1), S. 84–89.
- Rossmann, Toby G.; Visalli, Melissa A.; Komissarova, Elena V. (2003): fau and its ubiquitin-like domain (FUBI) transforms human osteogenic sarcoma (HOS) cells to anchorage-independence. In: *Oncogene* 22 (12), S. 1817–1821.

- Roth, Charles M. (2002): Quantifying gene expression. In: *Curr Issues Mol Biol* 4 (3), S. 93–100.
- Runnegar, B. (1985): Collagen gene construction and evolution. In: *J. Mol. Evol* 22 (2), S. 141–149.
- Rutherford, K.; Parkhill, J.; Crook, J.; Horsnell, T.; Rice, P.; Rajandream, M. A.; Barrell, B. (2000): Artemis: sequence visualization and annotation. In: *Bioinformatics* 16 (10), S. 944–945.
- Rychel, Amanda L.; Swalla, Billie J. (2007): Development and evolution of chordate cartilage. In: *J. Exp. Zool. B Mol. Dev. Evol* 308 (3), S. 325–335.
- Saikali, Z.; Sinnett, D. (2000): Expression of glypican 3 (GPC3) in embryonal tumors. In: *Int. J. Cancer* 89 (5), S. 418–422.
- Saleque, S.; Ruiz, N.; Drickamer, K. (1993): Expression and characterization of a carbohydrate-binding fragment of rat aggrecan. In: *Glycobiology* 3 (2), S. 185–190.
- Salomonis, Nathan; Hanspers, Kristina; Zambon, Alexander C.; Vranizan, Karen; Lawlor, Steven C.; Dahlquist, Kam D. et al. (2007): GenMAPP 2: new features and resources for pathway analysis. In: *BMC Bioinformatics* 8, S. 217.
- Sambrook, J.; Fritsch, E. F.; Maniatis, T. (1989): Molecular cloning. A laboratory manual. 2nd. New York: Cold Spring Harbor Laboratory Press. Online verfügbar unter <http://www.worldcat.org/oclc/472864750>.
- Sandelin, Albin; Carninci, Piero; Lenhard, Boris; Ponjavic, Jasmina; Hayashizaki, Yoshihide; Hume, David A. (2007): Mammalian RNA polymerase II core promoters: insights from genome-wide studies. In: *Nat. Rev. Genet* 8 (6), S. 424–436.
- Sanger, F.; Nicklen, S.; Coulson, A. R. (1977): DNA sequencing with chain-terminating inhibitors. In: *Proc. Natl. Acad. Sci. U.S.A* 74 (12), S. 5463–5467.
- Sansom, Ivan J.; Donoghue, Philip C. J.; Albanesi, Guillermo (2005): Histology and affinity of the earliest armoured vertebrate. In: *Biol. Lett.* 1 (4), S. 446–449. Online verfügbar unter doi:10.1098/rsbl.2005.0349.
- Schaefer, Liliana; Babelova, Andrea; Kiss, Eva; Hausser, Heinz-J; Baliova, Martina; Krzyzankova, Miroslava et al. (2005): The matrix component biglycan is proinflammatory and signals through Toll-like receptors 4 and 2 in macrophages. In: *J. Clin. Invest* 115 (8), S. 2223–2233.
- Schena, M.; Shalon, D.; Davis, R. W.; Brown, P. O. (1995): Quantitative monitoring of gene expression patterns with a complementary DNA microarray. In: *Science* 270 (5235), S. 467–470.
- Schipani, E.; Ryan, H. E.; Didrickson, S.; Kobayashi, T.; Knight, M.; Johnson, R. S. (2001): Hypoxia in cartilage: HIF-1alpha is essential for chondrocyte growth arrest and survival. In: *Genes Dev.* 15 (21), S. 2865–2876. Online verfügbar unter doi:10.1101/gad.934301.

- Schroeder, Andreas; Mueller, Odilo; Stocker, Susanne; Salowsky, Ruediger; Leiber, Michael; Gassmann, Marcus et al. (2006): The RIN: an RNA integrity number for assigning integrity values to RNA measurements. In: *BMC Mol. Biol* 7, S. 3.
- Senger, D. R.; Wirth, D. F.; Hynes, R. O. (1979): Transformed mammalian cells secrete specific proteins and phosphoproteins. In: *Cell* 16 (4), S. 885–893.
- Sengupta, Srikumar; Ruotti, Victor; Bolin, Jennifer; Elwell, Angela; Hernandez, Alvaro; Thomson, James; Stewart, Ron (2010): Highly consistent, fully representative mRNA-Seq libraries from ten nanograms of total RNA. In: *BioTechniques* 49 (6), S. 898–904.
- Serazin-Leroy, V.; Denis-Henriot, D.; Morot, M.; Mazancourt, P. de; Giudicelli, Y. (1998): Semi-quantitative RT-PCR for comparison of mRNAs in cells with different amounts of housekeeping gene transcripts. In: *Mol. Cell. Probes* 12 (5), S. 283–291.
- Serra, R.; Karaplis, A.; Sohn, P. (1999): Parathyroid hormone-related peptide (PTHrP)-dependent and -independent effects of transforming growth factor beta (TGF-beta) on endochondral bone formation. In: *J. Cell Biol* 145 (4), S. 783–794.
- Shiang, R.; Thompson, L. M.; Zhu, Y. Z.; Church, D. M.; Fielder, T. J.; Bocian, M. et al. (1994): Mutations in the transmembrane domain of FGFR3 cause the most common genetic form of dwarfism, achondroplasia. In: *Cell* 78 (2), S. 335–342.
- Shimeld, S. M. (1999): The evolution of the hedgehog gene family in chordates: insights from amphioxus hedgehog. In: *Dev. Genes Evol* 209 (1), S. 40–47.
- Shum, Lillian; Nuckolls, Glen (2002): The life cycle of chondrocytes in the developing skeleton. In: *Arthritis Res.* 4 (2), S. 94–106.
- Sim, Edmund Ui Hang; Ang, Chow Hiang; Ng, Ching Ching; Lee, Choon Weng; Narayanan, Kumaran (2010): Differential expression of a subset of ribosomal protein genes in cell lines derived from human nasopharyngeal epithelium. In: *J. Hum. Genet.* 55 (2), S. 118–120. Online verfügbar unter doi:10.1038/jhg.2009.124.
- Smith, T. F.; Waterman, M. S.; Fitch, W. M. (1981): Comparative biosequence metrics. In: *J. Mol. Evol* 18 (1), S. 38–46.
- Sodek, J.; Ganss, B.; McKee, M. D. (2000): Osteopontin. In: *Crit. Rev. Oral Biol. Med* 11 (3), S. 279–303.
- Solomon, Keith R.; Sharma, Parul; Chan, Melvin; Morrison, Paul T.; Finberg, Robert W. (2004): CD109 represents a novel branch of the alpha2-macroglobulin/complement gene family. In: *Gene* 327 (2), S. 171–183.
- Song, Howard H.; Shi, Wen; Xiang, Yun-Yan; Filmus, Jorge (2005): The loss of glypican-3 induces alterations in Wnt signaling. In: *J. Biol. Chem* 280 (3), S. 2116–2125.
- Stein, Lincoln D. (2010): The case for cloud computing in genome informatics. In: *Genome Biol* 11 (5), S. 207.
- St-Jacques, B.; Hammerschmidt, M.; McMahon, A. P. (1999): Indian hedgehog signaling regulates proliferation and differentiation of chondrocytes and is essential for bone formation. In: *Genes Dev.* 13 (16), S. 2072–2086.

- Su, W. C.; Kitagawa, M.; Xue, N.; Xie, B.; Garofalo, S.; Cho, J. et al. (1997): Activation of Stat1 by mutant fibroblast growth-factor receptor in thanatophoric dysplasia type II dwarfism. In: *Nature* 386 (6622), S. 288–292.
- Swoboda, B.; Pullig, O.; Kladny, B.; Pfander, D.; Weseloh, G.: Kollagen Typ VI-Gehalt im gesunden und arthrotischen humanen Kniegelenkknorpel. In: *Z Orthop Ihre Grenzgeb* 137 (6), S. 540–544.
- Szabó, P. E.; Mann, J. R. (1995): Allele-specific expression and total expression levels of imprinted genes during early mouse development: implications for imprinting mechanisms. In: *Genes Dev* 9 (24), S. 3097–3108.
- Szebenyi, G.; Savage, M. P.; Olwin, B. B.; Fallon, J. F. (1995): Changes in the expression of fibroblast growth factor receptors mark distinct stages of chondrogenesis in vitro and during chick limb skeletal patterning. In: *Dev. Dyn* 204 (4), S. 446–456.
- Tagariello, Andreas; Schlaubitz, Silke; Hankeln, Thomas; Mohrmann, Gerrit; Stelzer, Christiane; Schweizer, Anja et al. (2005): Expression profiling of human fetal growth plate cartilage by EST sequencing. In: *Matrix Biol* 24 (8), S. 530–538.
- Upholt, W. B.; Chandrasekaran, L.; Tanzer, M. L. (1994): Molecular cloning and analysis of the protein modules of aggrecans. In: *EXS* 70, S. 37–52.
- Vajo, Z.; Francomano, C. A.; Wilkin, D. J. (2000): The molecular and genetic basis of fibroblast growth factor receptor 3 disorders: the achondroplasia family of skeletal dysplasias, Muenke craniosynostosis, and Crouzon syndrome with acanthosis nigricans. In: *Endocr. Rev* 21 (1), S. 23–39.
- Velculescu, V. E. (1999): Essay: Amersham Pharmacia Biotech & Science prize. Tantalizing transcriptomes--SAGE and its use in global gene expression analysis. In: *Science* 286 (5444), S. 1491–1492.
- Velculescu, V. E.; Zhang, L.; Vogelstein, B.; Kinzler, K. W. (1995): Serial analysis of gene expression. In: *Science* 270 (5235), S. 484–487.
- Velculescu, V. E.; Madden, S. L.; Zhang, L.; Lash, A. E.; Yu, J.; Rago, C. et al. (1999): Analysis of human transcriptomes. In: *Nat. Genet.* 23 (4), S. 387–388. Online verfügbar unter doi:10.1038/70487.
- Verma, I. M.; Temple, G. F.; Fan, H.; Baltimore, D. (1972): In vitro synthesis of DNA complementary to rabbit reticulocyte 10S RNA. In: *Nature New Biol* 235 (58), S. 163–167.
- Vogel, K. G.; Paulsson, M.; Heinegård, D. (1984): Specific inhibition of type I and type II collagen fibrillogenesis by the small proteoglycan of tendon. In: *Biochem. J* 223 (3), S. 587–597.
- Vogel, K. G.; Trotter, J. A. (1987): The effect of proteoglycans on the morphology of collagen fibrils formed in vitro. In: *Coll. Relat. Res* 7 (2), S. 105–114.
- Vortkamp, A.; Lee, K.; Lanske, B.; Segre, G. V.; Kronenberg, H. M.; Tabin, C. J. (1996): Regulation of rate of cartilage differentiation by Indian hedgehog and PTH-related protein. In: *Science* 273 (5275), S. 613–622.

- Vu, T. H.; Shipley, J. M.; Bergers, G.; Berger, J. E.; Helms, J. A.; Hanahan, D. et al. (1998): MMP-9/gelatinase B is a key regulator of growth plate angiogenesis and apoptosis of hypertrophic chondrocytes. In: *Cell* 93 (3), S. 411–422.
- Wada, Hiroshi (2010): Origin and genetic evolution of the vertebrate skeleton. In: *Zool. Sci* 27 (2), S. 119–123.
- Wada, Hiroshi; Okuyama, Makiko; Satoh, Nori; Zhang, Shicui: Molecular evolution of fibrillar collagen in chordates, with implications for the evolution of vertebrate skeletons and chordate phylogeny. In: *Evol. Dev* 8 (4), S. 370–377.
- Wan, Mei; Cao, Xu (2005): BMP signaling in skeletal development. In: *Biochem. Biophys. Res. Commun* 328 (3), S. 651–657.
- Wang, Zhong; Gerstein, Mark; Snyder, Michael (2009): RNA-Seq: a revolutionary tool for transcriptomics. In: *Nat. Rev. Genet* 10 (1), S. 57–63.
- Watanabe, H.; Yamada, Y. (1999): Mice lacking link protein develop dwarfism and craniofacial abnormalities. In: *Nat. Genet* 21 (2), S. 225–229.
- Watanabe, H.; Kimata, K.; Line, S.; Strong, D.; Gao, L. Y.; Kozak, C. A.; Yamada, Y. (1994): Mouse cartilage matrix deficiency (cmd) caused by a 7 bp deletion in the aggrecan gene. In: *Nat. Genet* 7 (2), S. 154–157.
- Waterman, M. S. (1984): Efficient sequence alignment algorithms. In: *J. Theor. Biol* 108 (3), S. 333–337.
- Weise, M.; De-Levi, S.; Barnes, K. M.; Gafni, R. I.; Abad, V.; Baron, J. (2001): Effects of estrogen on growth plate senescence and epiphyseal fusion. In: *Proc. Natl. Acad. Sci. U.S.A* 98 (12), S. 6871–6876.
- Weisser, Jürgen; Riemer, Silvia; Schmidl, Martina; Suva, Larry J.; Pöschl, Ernst; Bräuer, Rolf; Mark, Klaus von der (2002): Four distinct chondrocyte populations in the fetal bovine growth plate: highest expression levels of PTH/PTHrP receptor, Indian hedgehog, and MMP-13 in hypertrophic chondrocytes and their suppression by PTH (1-34) and PTHrP (1-40). In: *Exp. Cell Res* 279 (1), S. 1–13.
- WESTLEY, J. (1959): A comparison of some biochemical properties of the beef liver and beef kidney rhodanases. In: *J. Biol. Chem* 234 (7), S. 1857–1860.
- Wheeler, D. L.; Church, D. M.; Lash, A. E.; Leipe, D. D.; Madden, T. L.; Pontius, J. U. et al. (2001): Database resources of the National Center for Biotechnology Information. In: *Nucleic Acids Res* 29 (1), S. 11–16.
- Xie, Chao; Tammi, Martti T. (2009): CNV-seq, a new method to detect copy number variation using high-throughput sequencing. In: *BMC Bioinformatics* 10, S. 80.
- Xu, S. C.; Harris, M. A.; Rubenstein, J. L.; Mundy, G. R.; Harris, S. E. (2001): Bone morphogenetic protein-2 (BMP-2) signaling to the Col2alpha1 gene in chondroblasts requires the homeobox gene Dlx-2. In: *DNA Cell Biol* 20 (6), S. 359–365.

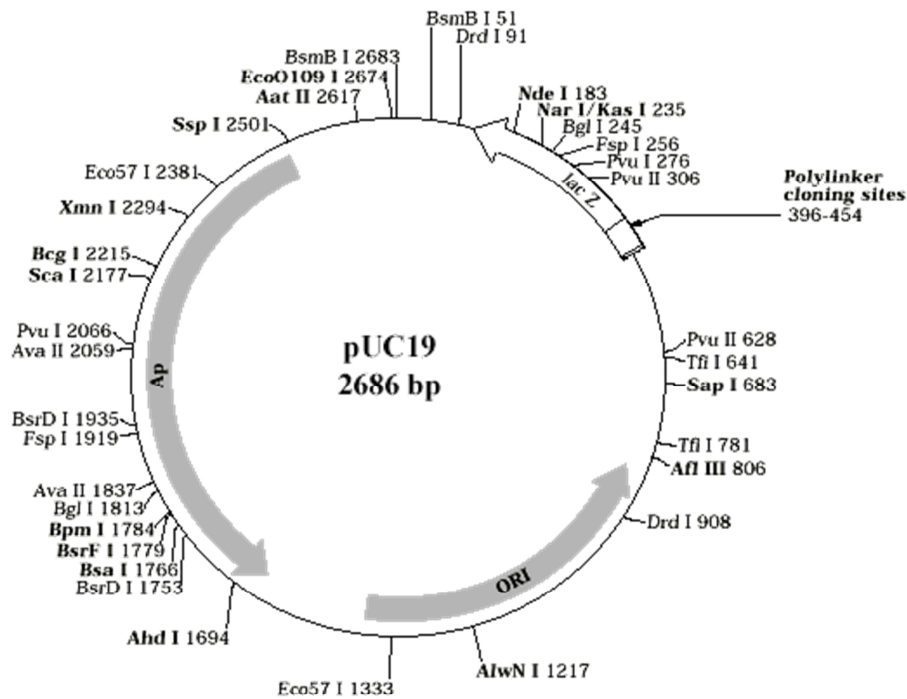
- Yakar, Shoshana; Rosen, Clifford J.; Beamer, Wesley G.; Ackert-Bicknell, Cheryl L.; Wu, Yiping; Liu, Jun-Li et al. (2002): Circulating levels of IGF-1 directly regulate bone growth and density. In: *J. Clin. Invest* 110 (6), S. 771–781.
- Yamaguchi, Y.; Mann, D. M.; Ruoslahti, E. (1990): Negative regulation of transforming growth factor-beta by the proteoglycan decorin. In: *Nature* 346 (6281), S. 281–284.
- Yang, X.; Yip, J.; Anastasiades, T.; Harrison, M.; Brockhausen, I. (2007): The action of TNFalpha and TGFbeta include specific alterations of the glycosylation of bovine and human chondrocytes. In: *Biochim. Biophys. Acta* 1773 (2), S. 264–272.
- Ying, Qi Long; Nichols, Jennifer; Chambers, Ian; Smith, Austin (2003): BMP induction of Id proteins suppresses differentiation and sustains embryonic stem cell self-renewal in collaboration with STAT3. In: *Cell* 115 (3), S. 281–292.
- Yoshida, Carolina A.; Yamamoto, Hiromitsu; Fujita, Takashi; Furuichi, Tatsuya; Ito, Kosei; Inoue, Ken-ichi et al. (2004): Runx2 and Runx3 are essential for chondrocyte maturation, and Runx2 regulates limb growth through induction of Indian hedgehog. In: *Genes Dev.* 18 (8), S. 952–963. Online verfügbar unter doi:10.1101/gad.1174704.
- Zdobnov, E. M.; Apweiler, R. (2001): InterProScan--an integration platform for the signature-recognition methods in InterPro. In: *Bioinformatics* 17 (9), S. 847–848.
- Zebboudj, Amina F.; Imura, Minori; Boström, Kristina (2002): Matrix GLA protein, a regulatory protein for bone morphogenetic protein-2. In: *J. Biol. Chem* 277 (6), S. 4388–4394.
- Zelzer, Elazar; Mamluk, Roni; Ferrara, Napoleone; Johnson, Randall S.; Schipani, Ernestina; Olsen, Bjorn R. (2004): VEGFA is necessary for chondrocyte survival during bone development. In: *Development* 131 (9), S. 2161–2171.
- Zhang, GuangJun; Cohn, Martin J. (2008): Genome duplication and the origin of the vertebrate skeleton. In: *Curr. Opin. Genet. Dev* 18 (4), S. 387–393.
- Zhang, H.; Marshall, K. W.; Tang, H.; Hwang, D. M.; Lee, M.; Liew, C. C. (2003): Profiling genes expressed in human fetal cartilage using 13,155 expressed sequence tags. In: *Osteoarthr. Cartil* 11 (5), S. 309–319.
- Zheng, Qi; Wang, Xiu-Jie (2008): GOEAST: a web-based software toolkit for Gene Ontology enrichment analysis. In: *Nucleic Acids Res* 36 (Web Server issue), S. W358-63.
- Zheng, Qiping; Zhou, Guang; Morello, Roy; Chen, Yuqing; Garcia-Rojas, Xavier; Lee, Brendan (2003): Type X collagen gene regulation by Runx2 contributes directly to its hypertrophic chondrocyte-specific expression in vivo. In: *J. Cell Biol* 162 (5), S. 833–842.
- Zhou, XiaoGuang; Ren, LuFeng; Li, YunTao; Zhang, Meng; Yu, YuDe; Yu, Jun (2010): The next-generation sequencing technology: a technology review and future perspective. In: *Sci China Life Sci* 53 (1), S. 44–57. Online verfügbar unter doi:10.1007/s11427-010-0023-6.
- Zhu, W.; Iatridis, J. C.; Hlibczuk, V.; Ratcliffe, A.; Mow, V. C. (1996): Determination of collagen-proteoglycan interactions in vitro. In: *J Biomech* 29 (6), S. 773–783.

Zou, H.; Niswander, L. (1996): Requirement for BMP signaling in interdigital apoptosis and scale formation. In: *Science* 272 (5262), S. 738–741.

7 Anhang

Verwendete Vektoren

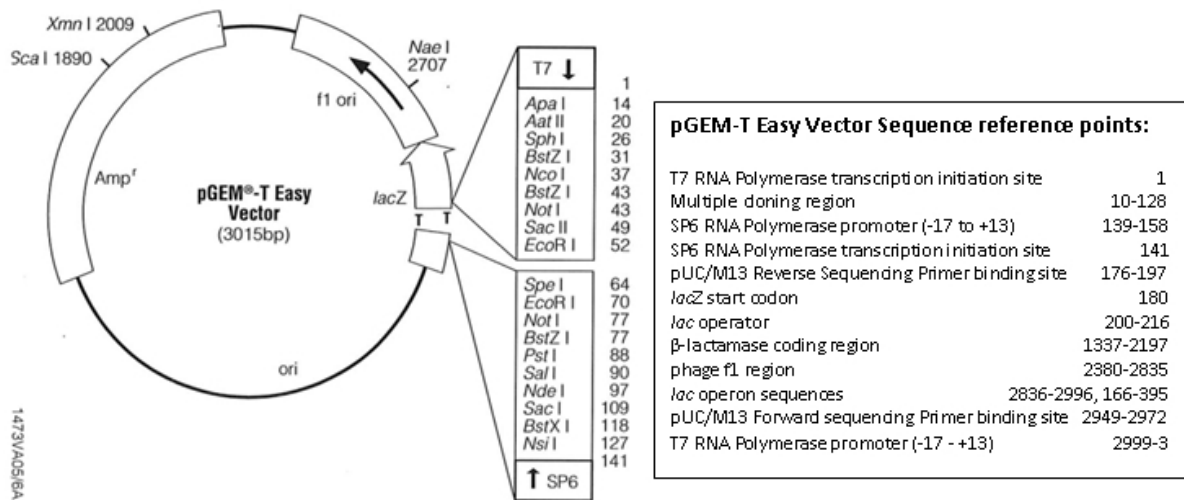
Vektor-Karte pUC19 (NEB)



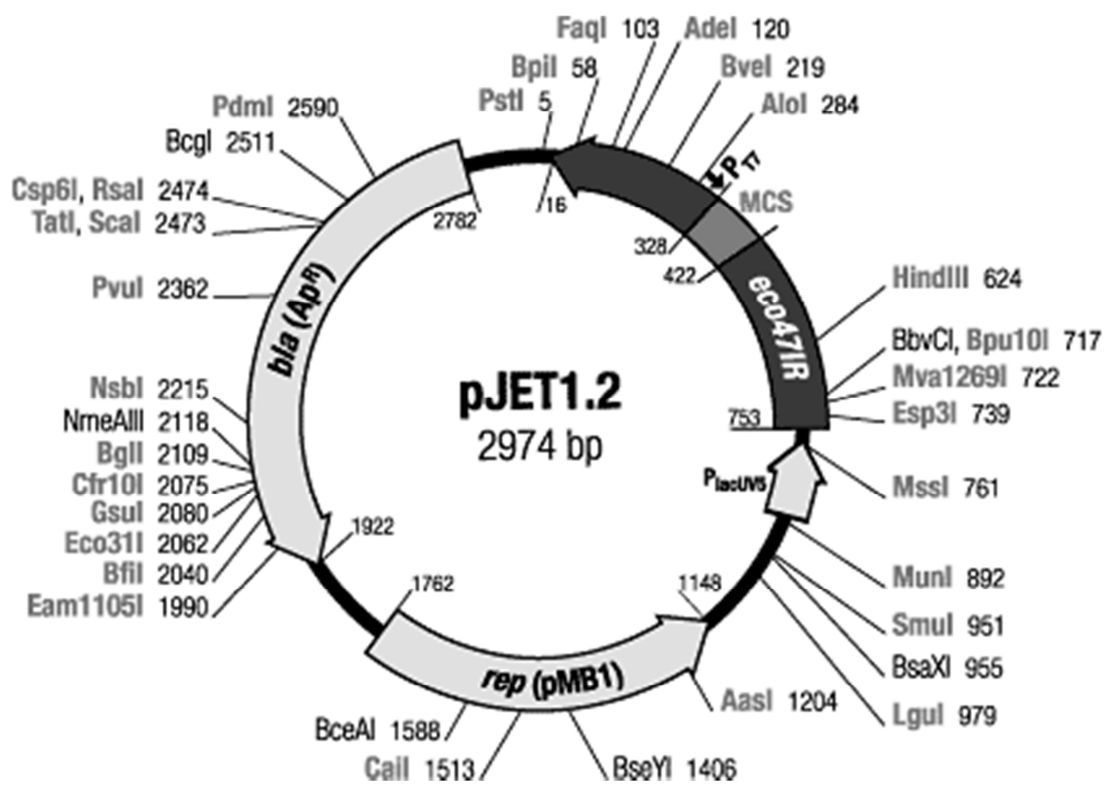
polylinker region



Vektor-Karte pGEM-T easy (Invitrogen)



Vektor-Karte pJet1.2 (Fermentas)



„Tissue Specificity Index“ - TSITab. 24: "Tissue Specificity Index" - *Bos Taurus*Aufgelistet sind lediglich Werte über 14; *Quelle: HUSAR, Bioinformatics Lab (DKFZ)*

UNIGENE	SYMBOL	BESCHREIBUNG	TSI
Bt.8462	RPLP0	Ribosomal protein, large, P0	21
Bt.63960	EEF1A1	Eukaryotic translation elongation factor 1 alpha 1	21
Bt.23522	ITM2B	Integral membrane protein 2B	20
Bt.48814	RPS2	Ribosomal protein S2	20
Bt.12309	HSPA8	Heat shock 70kDa protein 8	19
Bt.3003	RPL7L1	Ribosomal protein L7-like 1	18
Bt.64619	GNAS	GNAS complex locus	18
Bt.64714	COL3A1	Collagen, type III, alpha 1	18
Bt.64965	RPL5	Transcribed locus, strongly similar to NP_989912.1 ribosomal protein L5	18
Bt.5321	FTH1	Ferritin, heavy polypeptide 1	17
Bt.5535	YWHAB	Tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein,	17
Bt.13580	RPL7A	Ribosomal protein L7a	17
Bt.14068	RPS3A	Ribosomal protein S3A	17
Bt.37560	PGK1	Phosphoglycerate kinase 1	17
Bt.38667	ACTG1	Actin, gamma 1	17
Bt.48850	SPARC	Secreted protein, acidic, cysteine-rich (osteonectin)	17
Bt.48893	RPL4	Ribosomal protein L4	17
Bt.2832	MRCL3	Myosin regulatory light chain MRCL3	16
Bt.2981	GNB2L1	Guanine nucleotide binding protein (G protein), beta polypeptide 2-like 1	16
Bt.4070	SNX1	Sorting nexin 1	16
Bt.4880	SLC25A3	Solute carrier family 25 (mitochondrial carrier; phosphate carrier), member 3	16
Bt.5471	EIF4G2	Eukaryotic translation initiation factor 4 gamma, 2	16
Bt.7776	TPT1	Tumor protein, translationally-controlled 1	16
Bt.9042	EIF5	Eukaryotic translation initiation factor 5	16
Bt.13194	SSR1	Signal sequence receptor, alpha	16
Bt.16500	HDAC1	Histone deacetylase 1	16
Bt.49333	SFRS10	Splicing factor, arginine/serine-rich 10 (transformer 2 homolog, Drosophila)	16
Bt.59089	PBP	Phosphatidylethanolamine binding protein	16
Bt.87389	GAPDH	Glyceraldehyde-3-phosphate dehydrogenase	16
Bt.946	AK2	Adenylate kinase 2	15
Bt.1766	ARPC1A	Actin related protein 2/3 complex, subunit 1A, 41kDa	15
Bt.3599	LOC780990	Hypothetical LOC780990	15
Bt.4141	COPE	Coatomer protein complex, subunit epsilon	15
Bt.4704	CCT7	Chaperonin containing TCP1, subunit 7 (eta)	15
Bt.7736	LDHB	Lactate dehydrogenase B	15
Bt.11843	RPL24	Ribosomal protein L24	15
Bt.12504	CLU	Clusterin	15
Bt.15319	PGAM1	Phosphoglycerate mutase 1 (brain)	15
Bt.22248	RPL13	Ribosomal protein L13	15
Bt.23179		Transcribed locus, strongly similar to NP_034610.1 heat shock protein 1, alpha	15
Bt.24017	PSMC2	Proteasome (prosome, macropain) 26S subunit, ATPase, 2	15
Bt.48917	RPL9	Ribosomal protein L9	15
Bt.53243	TUBA1A	Tubulin, alpha 1a	15
Bt.53719	TEGT	Testis enhanced gene transcript	15
Bt.59450	RPS16	Ribosomal protein S16	15
Bt.64721	HSP90AB1	Heat shock 90kDa protein 1, beta	15

Eigene verwendete Perl-Skripte

pass_extract_multiple_genomic_regions_spliced_0_9

```

#!/usr/bin/perl
#####
#
# Steffen Rapp; Institut für Molekulargenetik; Universität Mainz
# pass_extract_multiple_genomic_regions_spliced_0_9
# extrahiert aus zwei durchgeführten PASS-mappings ("default" und "spliced") die "Read-Coverage"
#
#####
$| = 1;
my $default_file;
my $spliced_file;
unless (-e $ARGV[0])
{
    print "$ARGV[0] : there is no such file!\n";
    exit;
}
else
{
    $default_file = $ARGV[0];
}
unless (-e $ARGV[1])
{
    print "$ARGV[1] : there is no such file!\n";
    exit;
}
else
{
    $spliced_file = $ARGV[1];
}
print "$default_file\n";
print "$spliced_file\n";
my $output = $default_file."_extracted_list.txt";
open (DAT, "$default_file");
open (OUT, ">", $output);
my %hash;
my %hash2;
my %hash3;
my %NEWKEY;
my %NEWKEY2;
my @schluessel;
my @schluessel2;
my @region;
my $f1="";
my $f2="";
my $z1="";
my $g1="";
my $id=1;
@schluessel2=1..2914;
while (<DAT>)
{
    chomp ($_);
    if ($_ =~ /^(S+)\t(S+)\t(S+)\t(S+)\t(S+).*Hits=*(\d+)*$/)
    {
        $g1=$1;
        if ($g1 ne $z1 && !($g1 =~ @region))
        {

```



```

$NEWKEY{$g1}=$id;
$NEWKEY2{$id}=$g1;
push (@region,$g1);
foreach $schluessel2 (@schluessel2)
{
    $schluessel2="$schluessel2\t$NEWKEY{$g1}";
    $hash{$schluessel2}=0;
    $hash3{$schluessel2}=0;
    #print "$schluessel2\t$hash{$schluessel2}\n";
}
$id=$id+1;
}
$z1=$1;
}

$e6=1;
$e1=$1;
$e2=$2;
$e3=$3;
$e4=$4;
$e5=$5;
$e6=$6;
@schluessel=$e4..$e5;

if ($6 > 2)
{
    foreach $schluessel (@schluessel)
    {
        if ($NEWKEY{$e1})
        {
            $schluessel="$schluessel\t$NEWKEY{$e1}";
        }
        if ($hash2{$schluessel})
        {
            $hash2{$schluessel}=$hash2{$schluessel}+1;
        }
        else
        {
            $hash2{$schluessel}=1;
        }
    }
}
else
{
    foreach $schluessel (@schluessel)
    {
        if ($NEWKEY{$e1})
        {
            $schluessel="$schluessel\t$NEWKEY{$e1}";
        }
        $hash{$schluessel}=$hash{$schluessel}+1;
    }
}
}
open (DATA2, "$spliced_file");
my @schluessel3;
while (<DATA2>)
{
    chomp ($_);
    if ($_ =~ /^(S+)\t(S+)\t(S+)\t(S+)\t(S+).*Hits=(\d+)*$/)
    {
        $e1=$1;

```

```
$e2=$2;
$e3=$3;
$e4=$4;
$e5=$5;
@schluessel3=$e4..$e5;
foreach $schluessel3 (@schluessel3)
{
    if ($NEWKEY{$e1})
    {
        $schluessel3="$schluessel3\t$NEWKEY{$e1}";
    }
    $hash3{$schluessel3}=$hash3{$schluessel3}+10;
}
}
for my $key (sort {$a <=> $b} keys (%hash))
{
    if ($key =~ /\(\\S+)\(\\S+)\/)
    {
        print OUT "$NEWKEY2{$2}\t$1\t$hash{$key}\t$hash3{$key}\n";
    }
}
end;
```

Danksagung