Comparative genomic sequencing analysis of a region in human chromosome 11p15.3 / mouse chromosome 7 and analysis of the novel *STK33* / *Stk33* gene.

Dissertation to obtain the
Doctor's Degree in Natural Sciences

at the Biology Department
of the Johannes Gutenberg University Mainz

from Alejandro O. Mujica
born in Brighton, UK

Mainz, 2004

9. TABLE INDEX **194**

Abbreviations

Å	Armstrong
aa	Amino acid
AP	Alkaline phosphatase
ATP	Adenosine triphosphate
ATP-bd	ATP binding signature
Avg.	Average
BAC	Bacterial Artificial Chromosome
Bq	Bequerell
bp	Base pairs
BLAST	Basic Local Alignment Searching Tool
°C	Grad Celsius
cDNA	Complementary-DNA
cen	Centromere
CGAP	Cancer Genome Anatomy Project
Contigs	contiguous sequences resulting from an assembly
dATP	2'-deoxyadenosine 5'-triphosphate
dCTP	2'-deoxycytidine 5'-triphosphate
dGTP	2'-deoxyguanosine 5'-triphosphate
DIG	Digoxigenin
dTTP	2'-deoxythymidine 5'-triphosphate
DNA	Desoxiribonucleic acid
dNTP	2'-deoxyribonucleoside 5'-triphosphate
dUTP	2'-deoxyuridine 5'-triphosphate
ePK	Eukaryotic Protein Kinase
EST	Expressed Sequence Tags
FDA	Federal Drug Administration (USA)
Hsa	*Homo sapiens*
HGP	Human Genome Poject
HS	Hierarchical sequencing
HUGO	Human Genome Organization
HUSAR	Heidelberg Unix Sequence Analysis Resources
IHGSC	International Human Genome Sequencing Consortium
IPTG	Isopropil-β-D-thio-galactopyranoside
IUPAC	International Union of Pure and Applied Chemistry
Gb	Giga bases
GCG	Genomic Center Group
K	Kilo (thousands)
Kb	Kilo bases
kDa	Kilo Daltons
LB-Broth	Luria-Bertani Medium
M	Molarity
mA	Mili Ampers
Mb	Million bases
µg	Microgram
Mi	Millions
µL	Microliter
MGSC	Mouse Genome Sequencing Consortium
min	Minutes
mm	Milimeters
mM	Milimolar
Mmu	*Mus musculus*
mRNA	Messenger RNA
ms	Miliseconds
MTE	Multiple Tissue Expression array from Clonetech

MW	Molecular weight
NBT/BCIP	nitro blue tetrazolium/5-bromo-4-chloro-3-indo-lyl phosphate
NCBI	National Center for Biotechnology Information
NHGRI	National Human Genome Research Institute (USA)
nm	Nanometers
ORF	Open reading frame
PAC	Phage Artificial Chromosome
PCR	Polymerase Chain Reaction
PDB	Protein Data Bank
pI	Isoelectrical point
RACE	Rapid amplification from cDNA ends
PIP	Porcentage Identity Plot
RNA	Ribonucleic Acid
RT-PCR	Reverse Transcriptase Protein Chain Reaction
RZPD	Resource Zentrum – Primary Database (Berlin)
SDS	Sodium dodecyl sulfate
Seq.	Sequence
S_Tk	Serine/threonine kinase signature
STK33	Human serine/threonine kinase 33 gene
Stk33	Murine serine/threonine kinase 33 gene
STK33	Human serine/threonine kinase 33 protein
Stk33	Murine serine/threonine kinase 33 protein
SSC	Standar Saline-Citrat Buffer
Taq	*Thermophillus aquaticus*
tel	Telomere
TIGR	The Institute of Genome Research
UTR	Untranslated Region
WGS	Whole Genome Sequence
w/v	Weight/Volume
X-Gal	5-Brom-4-chlor-3-indolyl-ß-D-galactopyranosid
XL-PCR	eXtra Large Polymerase Chain Reaction

Our genomes can never accurately predict our futures.
But we would be more than silly if we did not use their information the fullest.
James Watson, 2001

1 Introduction

In 1953 Watson and Crick proposed the molecular structure of DNA with far reaching implications for our understanding of its function as carrier of the genetic information. More than 20 years later, there was still very little knowledge about the information stored in the DNA, because of the lack of efficient, fast and affordable techniques to determine the base sequence of longer DNA-molecules. In the late seventies, Sanger and Coulson as well as Maxam and Gilbert independently developed two different methods that allowed establishing a nucleotide sequence fast and at low cost (Sanger et al. 1977b). While the chemical method of Maxam and Gilbert is not further in use, the principle of the Sanger-technique is still the basis on which most of the current sequencing devices work. After the invention of the fast sequencing techniques, the analysis of whole genomes started immediately with the one of the bacteriophage phi X174 5,368 base pairs (bp) in length (Sanger et al. 1977a), and five years later in 1982 the genome of bacteriophage lambda with 48,502 bp was completed (Sanger et al. 1982). The way was paved for the era of genomics, and the sequencing of whole genomes including huge ones as our own became reality.

Since 1977 sequencing technologies have been improved remarkably. The first Sanger sequencing needed four tracks per sample in polyacrylamide gel electrophoresis, and the

nucleotides were radioactively labelled (Sanger et al. 1977b). Read lengths reached a few hundred bases under optimal circumstances, and the analysis of the results was performed manually. The first quantum leap of Sanger's chain-stop reaction sequencing was the use of different fluorescent dyes labelling each the four bases, reducing to one the necessary gel track for each sample, the collection and interpretation of the sequence data was feasible by computer and thus several samples could run in parallel. This breakthrough provided the necessary impulse for the development of high-throughput sequencing and thus coped with the task of having several fully sequenced genomes in a few years. Electrophoresis on thin acrylamide slab gels (0.5 mm) running in automatic sequencing machines provided enough capacity to process up to 96 samples simultaneously and produced reads of several hundred bases. Each run took just few hours. The more recent advance to become a standard is the use of very thin capillaries through which the electrophoresis runs in a matrix of linear acrylamide, replacing the function of the slab gels and letting the samples run truly separated. This and other improvements in robotics, temperature control, laser speed and sensitivity have increased the average read length up to 1,000 bp/read, reduced the run times to ca. 1.5 h for 96 samples and permit the processing of several batches continuously with much less human attention (Brown 1999; Marshall 1999). Next improvements include the manufacturing of micro-fabricated electrophoresis devices that reduce the amount of necessary fluorescent dyes even further, by controlling the sample injection volumes in the order of picoliters (Koutny et al. 2000; Mitnik et al. 2001; Paegel et al. 2003).

Sequencing technologies based on electrophoresis may be about to reach their limits. To achieve even longer and faster sequencing, new ideas and technologies may be needed. In fact, some alternative methods are already proposed which show different levels of

development. Such is the case of pyrosequencing and DNA-chip technology based sequencing (Brown 1999), and even the use of nano-devices for single-molecule sequencing (Pennisi 2002). In the meantime electrophoretically-running chain-termination sequencing remains the standard and will perhaps persist as a cost-effective solution in coming years.

This fast-paced development is the result of at least two trends of technological achievement in recent years. The first one is the improvement of sequencing technologies based on the Sanger chain-termination principle, and the second one is the rapid development of information technologies. It is perhaps not casual that the genomic era has developed in parallel with the Internet. Following the "Moore law" on the doubling time of computer performance, the cost of genomics has been dropping roughly twofold every 18 months in the last decades (Aach et al. 2001). Consequently, the amount of on-line available biological data, especially nucleic acid sequences has been growing exponentially as shown in the figure 1.1.

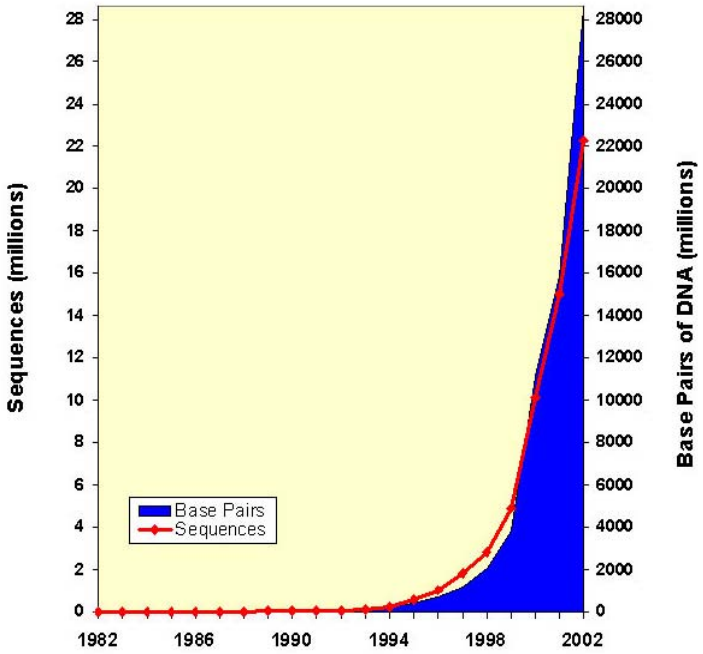


Figure 1.1: GenBank growth.
Number of sequence entries and total number of bases per year. Source: GenBank statistics: www.ncbi.nlm.nih.gov/Genbank/genbankstats.html,

1.1 Genome sequencing projects

By the end of 2002 the sequence of 119 genomes of free living organisms (i.e. excluding viruses) were completed in at least their draft version. In general the organisms whose genomes are already sequenced are classical model organisms, disease-related organisms or crop species.

In 1995 the first non-virus genome was sequenced, from the bacterium *Haemophilus influenzae*; in 1996 the first from an archaea, *Methanococcus jannaschii*; in 1997 the genome of the classic model bacteria *Escherichia coli* was completed and the first from an eukaryote *Saccharomyces cerevisiae*; in 1998, the first metazoan *Caenorhabditis elegans*; in 2000 the first insect *Drosophila melanogaster* and the first plant *Arabidopsis thaliana*; in 2001 the Human and more recently the Mouse, a model fish *Fugu rubripes* and in 2002 two sub-species of the rice *Oryza sativa*. Figure 1.2 shows a chronicle of this remarkable development until 2002 based on data from Bernal (2001), Appendix 7.1 show the whole results.

In the same issue of the journal Nature, 1 April 2004, the rat genome, and the final reports of the sequence an annotation of human chromosomes 13 and 19 were published (Dunham et al. 2004; Gibbs et al. 2004; Grimwood et al. 2004).

With the completion and analysis of the genomes from *Homo sapiens*, two species of the parasite *Plasmodium* and the vector *Anopheles gambieae*, the long standing struggle against Malaria, a disease responsible for millions of deaths per year, has received support from genomics (Hastings et al. 2002; Holt et al. 2002).

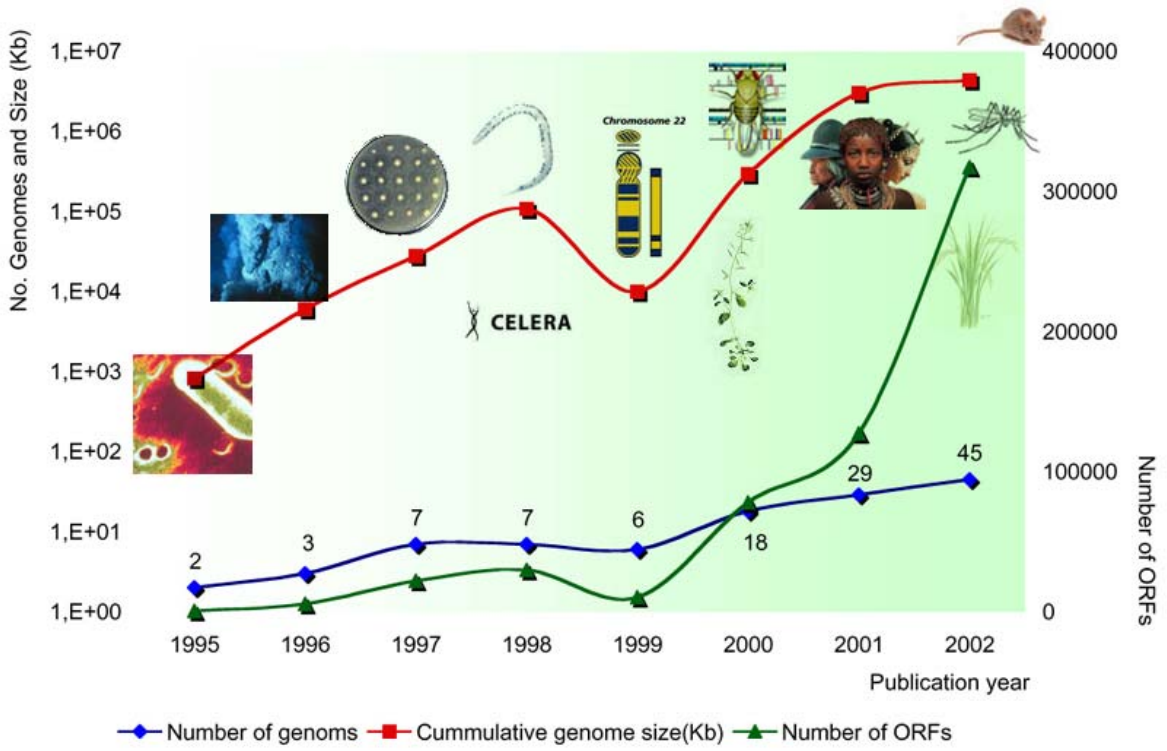


Figure 1.2: Development of eukaryotes genome sequencing from its beginning in 1995 to 2002. The curve in blue shows the number of non-viral genomes sequenced per year in an exponential scale (Y axis on the left, the number of genomes sequenced shown at each year-point of the curve). The curve in red shows the total size of genomes sequenced per year in the same exponential scale. The curve in green shows the number of ORF (~ genes) predicted in the genome projects (lineal scale, Y axis on the right). Some milestones mentioned in the text are graphically represented. Modified from GOLD[TM] Genomes OnLine Database, www.genomesonline.org/ (Bernal et al. 2001)

The international Human Genome Project (HGP) was launched in 1990 and conceived with the goal of sequencing the entire human genome in a 15 year long international effort. Despite this time frame, two independent drafts of the human genome were published simultaneously in February 2001 and the final version was announced in 2003 with the celebration of the 50th jubilee of the Watson and Crick milestone (Collins et al. 2003a). Sequencing the human genome has been regarded as one of the greatest human scientific endeavours. Originally scheduled for completion in 2005 in an open collaborative international effort, the 15 year-long period was necessary to automate the sequencing technology, to construct the genetic and physical maps for the hierarchical sequencing

strategy (also called clone-by-clone or chromosome walking) and to sequence genomes from other organisms, that should assist in the discovery of human genes and determination of their function. Funded mostly by the USA's National Human Genome Research Institute (NHGRI), the UK's Wellcome Trust and several national research funding agencies from other countries, there seemed to be work enough for sequencing centers distributed all over the world. Particularly, in the first years of sequencing technology development, even private companies were also beneficiaries from academic funding.

In May 1998, the human genome sequencing received an unexpected impulse from the private sector: the world leading manufacturer of online sequencing machines, Perkin Elmer (PE) and the director of The Institute of Genomic Research (TIGR) J.C. Venter, announced the creation of the company Celera, with the intention of sequencing the entire human genomic DNA independently and producing the first rough draft by 2001. For this, Celera declared to use the whole genome shotgun (WGS) strategy. In contrast with the hierarchical sequencing strategy used by the International Human Genome Sequencing Consortium (IHGSC), in the WGS the whole genome is sheared, the resulting fragments are cloned in plasmids, sequenced and in the end assembled by computers. To accomplish this goal, Celera set up a sequencing center with 230 networked PE PRISM 3700 Capillary-based automatic sequencing machines and the world's most powerful civilian computing center (Marshall 1999). Figure 1.3 shows a comparison between the sequencing strategies used by both teams.

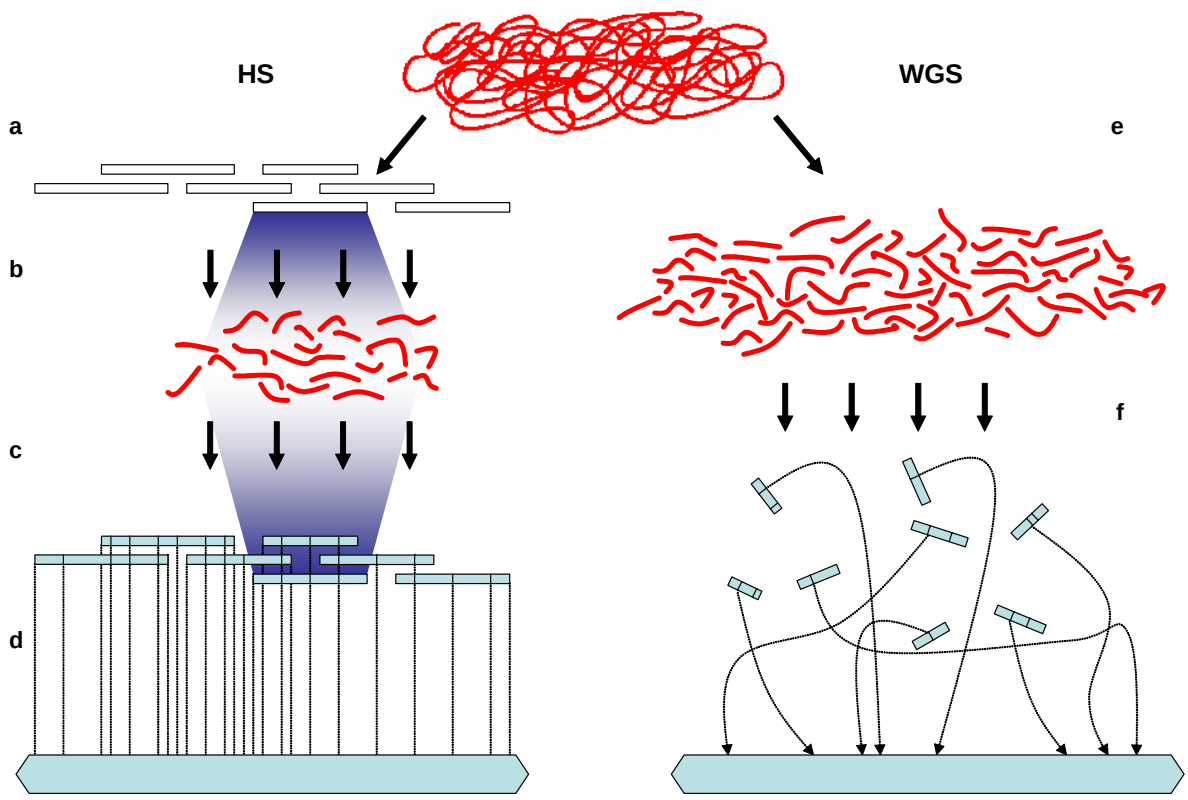


Figure 1.3: Sequencing strategies used to generate the human genome draft.
The hierarchical sequencing (HS) strategy used by the IHGSC is shown on the left and the Whole Genome Shotgun (WGS) used by Celera on the right. **a)** The first step of the HS is the construction of a map of BAC clones (i.e. positions relative to each other and to the chromosome are mapped prior to sequencing). The order of BAC clones provides the so called "tiling path". **b)** Each BAC clone is chopped in random pieces, sequenced and assembled by a computer program into a BAC contig, here represented in just one of the clones (see also Materials and methods and Results sections for detailed description of the shotgun method **c)** and **d)** The resulting sequences of the BAC clones are merged, even if they are incomplete: markers may guide the anchoring of the clones. **e.)** In the WGS the sequence the whole sequence is shredded and assembled computationally **(f)** (Lander et al. 2001; Waterston et al. 2002).

Concerned about availability of data and intellectual property rights of Celera's results on human genes, the public research funding agencies of the IHGSC increased dramatically the budget for sequencing the human genome faster than originally planned, to have available the public draft in 2001, 4 years ahead of schedule. $77 million in England and $81.6 million in the USA were awarded for an intensive 10 months work period. However, not for all the institutions involved at that time. Instead, the funding was concentrated to the Sanger Center in the UK and the three centers in the USA with the greatest sequencing capacity. Participating countries like Germany and Japan were not involved in setting this new

deadline. Their participation, and that from smaller sequencing centers in the USA, was not excluded but they were to find other funding resources to keep pace with the large sequencing centers (Pennisi 1999).

Table 1.1 Some data from human and mouse draft sequencing projects.
Compiled from (Lander et al. 2001; Venter et al. 2001; Waterston et al. 2002).

	Human		Mouse	
	IHGSC	**Celera**	**MGSC[a]**	**Celera[b]**
Estimated total size	3,289 Mb	2,910 Mb	2,493 Mb	
% Sequenced	88%		96%	
Total raw sequence	23Gb	30 Gb		
Number of reads			33.6 Mi	
Coverage	7.5	5.1	7.7	5.3
N50 length[c]	82Kb	86Kb	29Kb	14Kb
Number of PAC/BAC clones	29,298	-		
Number of sequence-clone contigs	4,884			
Number of fingerprint clone contigs	942			
Repeat content	>50%	>35%	38.55%	
Coding content	5%			
% in segmental duplications[d]	3.6% (5%)		-	
Genes from horizontal transfer from bacteria	223	-	-	-
(G+C) content	41		42	
Total CpG Islands (in thousands)	50	27 - 196		
CpG Islands after repeat masking	28,890		15,500	
Predicted protein coding genes (in thousands)	30 - 40 K	26 - 38 K	~30 K	
Kinase genes	575 (2%)	868 (3%)	2.1%	

[a]MGSC: Mouse Genome Sequencing Consortium
[b]Not published
[c]N50 length: contiguity of the sequencing assesment, defined as the largest length L such that 50% of all nucleotides are contained in contigs of size at least L. Waterston and collegues (2003) insist that is mathematically impossible to get an N50 value of 86Kb with the 5.1 WGS coverage that Celera used to sequence the human genome without having used the tiling path from the public data. The N50 results of both mouse genome drafts tend to confirm this observation.
[d]Difficult to determine when WGS is used

Attempts for finishing the human genome in collaboration between public and private efforts, as it was the case for the *Drosophila* genome, ended with rivalry (Collins et al. 2003a; Marshall 2000a). In a rather political ceremony, the completion of the draft version of the human genome was celebrated in 26. June 2000 by both teams (Marshall 2000b), and the results published simultaneously but separately in February 2001, the IHGSC's in Nature

(Lander et al. 2001) and Celera's in Science (Venter et al. 2001). Celera reached a highly polemic agreement with the journal Science (Marshall 2000c), whereby the publication of the draft paper was accepted without sending the sequences to the public databases. Access to Celera's human genome sequences is open to academic users with certain limitations and only through their Internet server. Sequences and annotation from the public consortium are freely available and have been continuously updated ever since. Table 1.1 shows some results from the draft genomes of human and mouse from both public and private efforts.

The use of a WGS strategy for sequencing the human genome was already proposed in a very detailed manner to the NHGRI before Celera adopted this method (Weber and Myers, 1997). The strategy had proven successful in sequencing genomes up to 2 million bases long, but some authors questioned whether a much bigger vertebrate genome with scores of nearly identical short and long interspersed repeats could be effectively assembled with this approach and the proposal was rejected by the NHGRI (Marshall 1999). A hybrid approach combining clone-by-clone with whole genome shotgun sequencing was used successfully for sequencing the *Drosophila* (Rubin et al. 2000) and the mouse genome (Waterston et al. 2002). The usefulness of the WGS strategy for sequencing and assembling vertebrate genomes to a level suitable for preliminary long-range genome comparisons is no longer questioned (Aparicio et al. 2002; Waterston et al. 2002), but it is still questioned whether Celera's human genome assembly relied exclusively on this kind of data alone or if they took advantage from the tiling path of the public effort (Cozzarelli 2003). This debate is still open (Adams et al. 2003; Waterston et al. 2003). It may be questioned whether or not the race for publishing the first draft of the human genome was beneficial and discussion of this is outside the scope of this work. The competition to publish the draft first may have affected the quality of the first

data shown, though this competition, resulted in two distinct results that may validate each other (Aach et al. 2001).

Sequencing genomes has become a rather routine and mostly automated work. The greater challenge now and in coming years will be to unveil the functional significance of the sequence data, in particular to identify genes involved in human diseases and to get a better understanding of the molecular principles of life. Clearly, the generation of sequence data is no longer the limiting factor, as its analysis (Malakoff 1999); Emmett, 2000; (Bohannon 2002).

1.2 Beyond the Human genome draft

Some findings of the human genome sequencing projects were rather surprising. Perhaps the most unexpected one is the fact that much fewer than predicted protein coding genes exist. Estimation for the number of human genes ranged from 50,000 to 150,000 genes (Pennisi 2000; Rubin et al. 2000) and The C. Elegans Sequencing Consortium, 1998 (1998). This high count was cut to less than half: the public human genome project estimated between 30,000 and 40,000 genes, Celera between 26,000 and 38,000, and both groups favoured final predictions closer to the lower numbers. From these numbers, still the result of a first measurement, only 11,000 are known genes, whereas the rest are computer-derived gene-predictions both de novo and made by comparison with expressed sequences in the public databases. Because of the draft nature of the sequence used for the predictions, it is conceivable that some genes escaped detection. It is also reasonable that gene-finding programs do not always succeed in finding genes from the raw data. Actually, some reports

suggest that this first calculation is underestimated. By comparing both drafts from IHGSC and Celera, it was evident that gene-predictions which in the end make-up 2/3 of the whole count, do not completely overlap with each other. If all gene-predictions from both teams were to be correct, the total number should come close to 42,000 genes (Hogenesch et al. 2001). Recent independent analyses on the draft genomes and other available resources have reached an estimate of up to 70,000 genes (Hollon, 2001). Still, there are good reasons to believe that this would be the maximum number of genes for a human being (Daly 2002). New calculations of gene numbers of other species, such as *Fugu rubripes* (Aparicio et al. 2002) and the mouse (Waterston et al. 2002) point toward a "core-number" of protein coding genes of approximately 40,000 for vertebrates.

To date, nine human chromosomes are completely sequenced to the high quality specified by the Human Genome Project. Chromosome 22 in 1999, chromosome 21 in 2000, chromosome 20 in 2001, chromosomes 14, Y, 7 and 6 in 2003 and recently chromosomes 13 and 19. The final versions of the sequence of these chromosomes permit a more realistic estimate of the total number of human genes. As an example, a re-analysis of the sequence of chromosome 22 4 years after its first publication, yielded additionally 198 potential protein coding genes, RNA genes, partial genes and pseudogenes, i.e., an increase of 43% for genes and 74% for exons. On the other hand, in this new annotation, several predicted genes were merged, some were totally removed and many predicted functional genes were reassigned as pseudogenes so that the overall protein-coding gene number has only increased by one (1) gene, from 545 to 546 (Collins et al. 2003b). This demonstrates how uncertain the prediction of genes solely on the basis of draft sequence is.

By annotating the genome of *Drosophila melanogaster* Rubin and colleges (Adams et al. 2000; Rubin et al. 2000) were already surprised that both the fly and the worm (*Caenorabditis elegans*) genomes showed only twice as many protein coding genes as the baker's yeast *Saccharomyces sereviseae*. It may result surprising that the fly with its complex body plan, multiple tissues, complex development and nervous system, has only twice as many genes as a unicellular organism. The same reasoning applies to humans, having much more complex respiratory, circulatory and digestive systems, a much more sophisticated brain and complete new traits, such as homeostasis and immune systems, but revealing only twice as many genes as *Drosophila*. Clearly, biological complexity can not be explained merely in terms of an expanded catalogue of genes in a genome, at least not by the number of protein coding genes.

In spite of the relative low number of genes, the initial sequencing and analysis of the human genome confirmed that the complexity of a vertebrate proteome, the whole set of proteins in a given cell or organism, is in line with the complexity of the body plan (Lander et al. 2001; Venter et al. 2001). How is this lack of stoichiometry between number of genes and proteins to be explained? The answer of this rather fundamental question does not lie in any new discovery, but in already known biological phenomena that may have been overlooked or underestimated. Protein diversity is increased in metazoans through several mechanisms: multiple transcription initiation sites, alternative splicing, alternative poly-adenylation signals and pre-mRNA editing, alternative splicing being maybe the most relevant (Maniatis and Tasic 2002). Though already well documented more than 20 years ago, only 5% of genes in higher eukaryotes were estimated to show alternative splicing (Modrek and Lee 2002). Current predictions, postulate that between 40 and 60% of human genes show at least one

form of alternative splicing (Brett et al. 2002; Mironov et al. 1999; Modrek and Lee 2002; Roberts and Smith 2002). A remarkable example from *Drosophila*: a single-copy gene, *Dscam*, which has the potential of producing more than twice as many protein isoforms through alternative splicing than genes exist in the whole genome of the fly (Schmucker et al. 2000). Although several examples of alternative splicing and its underlying mechanism are well characterised, less is known about the way this process is regulated (Cartegni et al. 2002; Maniatis and Tasic 2002).

1.3 Comparative sequence analysis and gene-prediction

Genome annotation, the process of extracting information from new anonymous genomic sequences relies mainly on de-novo gene-predicting algorithms and database similarity searches. However, the available bioinformatic tools do not allow extracting the genetic data thoroughly. De-novo gene-predictions frequently lead to false positives and at similar rate, true exons are overlooked (Amid 2002; Bahr 1999; Frazer et al. 2003), they are also very sensitive to assembly errors or even single-base sequencing errors (Mathe et al. 2002). Database searching may fail to detect divergent coding sequences or members of novel gene families, because of limited sequence similarity. Some genes may be missed by EST comparison because their rare transcripts are not present in the databases; after all, probably only a maximum of 17,000 genes are active in a human cell at a given time, the remaining are silent, or expressed at very low levels, such as transcription factors that may be present in less than 10 copies per cell (Jongeneel et al. 2003).

De novo gene-prediction and database searching alone are also not sufficient to detect other regions of biological significance, such as regulatory elements, RNA genes and chromosome structure specific elements (Frazer et al. 2003). Very short regulatory regions, which may be dispersed over the whole genome, are notoriously difficult to predict accurately. Even if the approximate position of the promotor region is known, the regulatory elements may be difficult to identify (Aach et al. 2001; Hardison et al. 1997). However, functionally important elements are usually thought to be evolutionary conserved because of the acting negative selection against deleterious mutations. Hence, by comparing genomic sequences from different organisms, "islands" of relative higher conservation through evolutionary constraints are one way to identify regions of biological significance. Comparative genomics is one option to increase the accuracy of de-novo gene-predictions (Mathe et al. 2002). Thus, the method of choice for an improved identification of novel genes and regulatory elements is comparative sequence analysis.

Comparative sequence analysis is possible because of the existence of Synteny: chromosomes of related species show blocks of evolutionary homologous sequence, in which the position and orientation of genes are conserved. Genomes from species, such as human and mouse that undergo divergent evolution more than 80 million years ago have accumulated many chromosomal rearrangements, yielding 342 syntenic chromosomal blocks. More than 90% of the human and mouse genomes are arranged in such conserved syntenic blocks as shown in the figure 1.4 (Gibbs et al. 2004). Even more distant species, such as *Anopheles* and *Drosophila*, which have been separated 250 millions years ago and have a faster evolutionary rate than vertebrates, show 948 shorter remaining microsyntenic clusters covering 34% of their whole genomes (Zdobnov et al. 2002). After 450 million years of

divergence between humans and the fish *Fugu rubripes*, gene shuffling and rearrangement is a major factor leading to around 900 syntenic fragments constituting 12.5% of the *Fugu* genome (Aparicio et al. 2002)

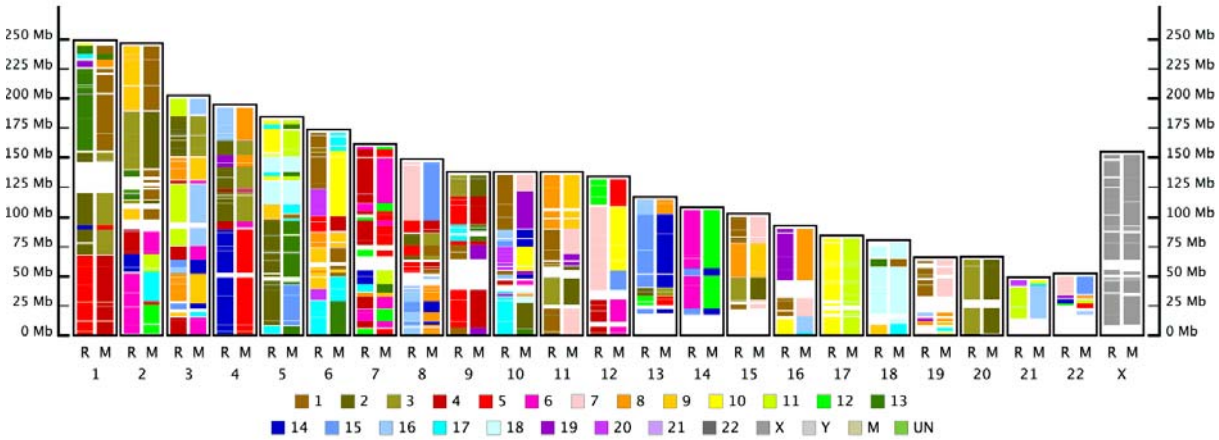


Figure 1.4: Rat-Mouse-Human Synteny Map.
Y-axis represents chromosome size in million bases. Vertical bars, aligned over the X-axis, represent the human chromosomes. Each bar contains two parallel columns representing synteny blocks with rat **(R)** and mouse **(M)** according to the colour legend beneath. Source: www.genboree.org (Gibbs et al. 2004).

The use of species of different evolutionary distance functions like a focusing device: by comparing too distant species, novel genes in both taxa may be missed, by comparing too close species, the sequences have not diverged enough to reveal conserved regions. The species selected to be compared, should deliver the right signal-to-noise ratio that aids in the detection of functionally constrained sequences (Bergman et al. 2002).

The availability of the sequence from the whole genome of model organisms from several branches of the tree of life provides a unique opportunity for studying the evolutionary history of human genes and gene families and therefore let speculate on their function and their impact on the whole body plans. Moreover, the detailed knowledge of the genomes of model organisms may shed light on evolutionary traits and may help in understanding the mechanisms of human diseases.

Human-mouse genome comparison has been used before their genomes were completely sequenced. Already in 1995, Koop recognised different “patterns” of DNA sequence conservation in the non-coding sequence of beta-globins, myosin and immunoglobulins of human and mice (Koop 1995). In 1997, Oeltjen and colleges compared the Bruton’s tyrosine kinase loci, and in this way they narrowed potential transcription factor-binding sites and putative intronic regulatory sequences. Similar analysis of this type were made in following years (Ansari-Lari et al. 1998; DeSilva et al. 2002; Oeltjen et al. 1997), and even whole chromosomes were compared with their homologous counterparts in the related species. Dehal and colleagues (2001) compared the whole human chromosome 19 with homologous regions in the mouse. A total of about 1,200 genes were predicted, among them 128 previously not-annotated genes were detected through their counterparts in the mouse genome. Vice versa, mouse chromosome 16 was also compared with its homologue fragments in human (Mural et al. 2002). They predicted 731 genes, 14 from them with apparently no human counterpart and from the corresponding human regions of conserved synteny they found 21 out of 725 with no mouse counterpart. These examples show the strength of the human-mouse genome comparison.

1.4 Aims of study

This work is part of a comparative sequencing project on human chromosome sub-region 11p15.3 and its homologous region on distal mouse chromosome 7 (see figure 1.5), addressed by the Child's Hospital and the Institute for Molecular Genetics of Mainz University since 1996 in the frame work of the German Human Genome Initiative (Amid et al. 2001; Cichutek et al. 2001). 11p15.3 is a gene-rich region of relevant medical importance. The human chromosomal region 15 on the distal region of the short arm of human chromosome 11 (Chromosome 11p15) has been associated with several defects and malignancies, such as the Beckwith-Wiedmann-Syndrome (BWS), Hemoglobinopathies, Long QT-Syndrome (Ward Romano Syndrome), Insulin-dependent Diabetes Mellitus I, Usher Syndrome 1C, T cell leukemia, Hypoparathyroidism, Nieman-Pick disease, Bardet Biedl Syndrome and Typ 2 (Nowak and Shows 1995) as well as different types of cancer in bladder, ovary, testis, breast and lung among others (Karnik et al. 1998). As shown in the figure 1.5, Chromosome 11p15 contents at least the tumor-suppressor genes of medical importance *WEE1*, *ST5* and *LMO1*. *WEE1* is the human homologue of the yeast *wee1+* gene, which encodes for a mitotic inhibitor (Igarashi et al. 1991; Watanabe et al. 1995); *St5* reduces tumorigenicity when transfected into HeLa cells (Lichy et al. 1996). *LMO1* (alternate symbols *RBT1* or *TTG1*) is a putative transcriptional regulator without DNA-binding domains involved in T-cell acute lymphoblastic leukemia in children with translocation t(11;14)(p15;q11) (Angel et al. 1993; Boehm et al. 1988; Korsmeyer 1992)

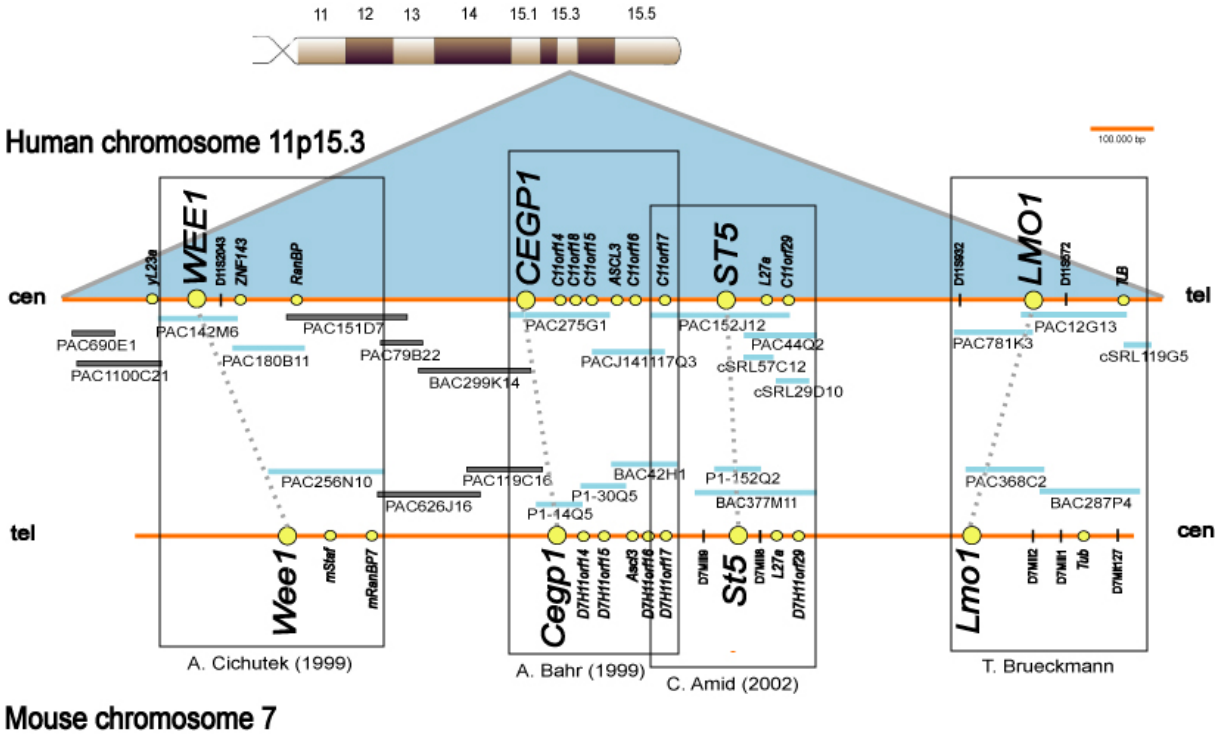


Figure 1.5: Clone-contig map of the region cooperatively analysed by the Child's Hospital and Institute for Molecular Genetics in Mainz University at the beginning of this work.
A diagram of the short arm of human chromosome 11 is shown at the top. In the middle, an expanded view of the human region showing the relative positions of PAC, BAC and Cosmid clones between genes *WEE1* and *TUB*. The homologous chromosome fragment of mouse chromosome 7 is shown below. Parallel rectangles show the different work areas of prior doctoral theses and their dates of completion. Abbreviations **cen** and **tel**, show the directions of the centromere e and telomere in each chromosome (Amid et al. 2001; Cichutek et al. 2001).

The gene *CEPG1* (alternate symbols *SCUBE2*) is expressed in vascular endothelial cells and codes for a secreted protein of the extra-cellular matrix (Löbbert, 1996; Cichutek, 1997; Bahr, 1999; (Yang et al. 2002). The order of the genes centromere e -*WEE1*-*CEPG1*-*ST5*-*LMO1*-telomere in human and telomer-*Wee1*-*Cepg1*-*St5*-*Lmo1*-centromere e in the mouse was established by in-situ hybridisation (Cichutek et al. 2001; Seipel 1996).

The primary goal of the collaborative sequencing project from the Child's Hospital and the Institute for Molecular Genetics of Mainz University was to establish the complete sequence between the *WEE1* and the *LMO1* genes and to identify genes and evolutionary conserved sequence elements in this region. The sequencing of human and murine genomic

regions around these four genes was the subject of the doctoral theses from Andrea Cichutek (Cichutek 2001), André Bahr (Bahr 1999), Clara Amid (Amid 2002) and Thomas Brueckmann (in preparation) as depicted in the figure 1.5. The analysis of the region still left gaps between the *WEE1* and *CEPG1* genes and *ST5* and *LMO1* (Amid et al. 2001; Cichutek et al. 2001).

As part of the cooperation, the objective of this work was completing the analysis in the region between *ST5* and *LMO1*. This includes screening of genomic PAC/BAC libraries to select clones that close the gap, sequence them with a hybrid shotgun-primer walking sequencing strategy, sequencing the homologous murine sequence and analyse the sequence data using comparative sequence analysis for the identification of possible novel genes and other evolutionary conserved sequence elements. As far as novel genes are identified a preliminary functional characterisation i.e., gene architecture, transcription pattern and protein product analysis, shall be done.

.

2 Materials and methods

2.1 DNA standard methods

Except when noted, standard DNA/RNA methods were performed based on (Sambrook and Russell 2001).

2.1.1 DNA preparation

Plasmid DNA preparations were performed following alkaline lysis procedures. High throughput plasmid prep for shotgung sequencing was performed with 96 block-based plasmid preparations from QIAGEN and MachereyNagel following manufacturer's protocols. Single plasmid preparations for sequencing were performed with isolation products from QIAGEN, GibcoBRL and PeqLab following manufacturer's protocols.

2.1.2 DNA quantitation

DNA concentration was quantified spectrophotometrically by meassuring its absorbance at the wavelength of 260 nm.

2.1.3 DNA digestion with endonucleases

DNA restriction was typically performed with 1 to 5 Units endonucleases per microgram DNA. Buffer and temperature were selected according to manufacturer's instructions. When using different enzymes simultaneously, OnePhorAll buffer from Pharmacia was preferred.

2.1.4 DNA precipitation

In order to concentrate, desalt or isolate from primers and unincorporated nucleotides, DNA-fragments were precipitated with 1/10 volume 3M Sodium acetate pH 5.4 and 2 volumes of absolute ethanol (or 1 volume isopropanol) at -20°C for at least 30 minutes. DNA fragments were recovered by centrifugation at maximum speed (~12,000 G) for 45 minutes. After discard of supernatant, DNA-pellet was washed with 70% ethanol and centrifuged again. After discard of supernatant, DNA-pellet was dried in a vacuum centrifuge and resuspended in TE or preferably in double distilled H_2O if DNA was to be sequenced or amplified by PCR.

2.1.5 DNA detection, analysis and isolation by Electrophoresis

According to expected size, DNA fragments or preparations were tested by electrophoresis in 0.8 to 2% w/v agarose gels. Vertical gels at intensities from 15 to 30 mA, whereas horizontal gels at 100 mA. As DNA molecular marker for the high range, phage λ genome restricted with *Hind* III was used and plasmid pF restricted with *Eco* RI and *Not* I

for the low range. DNA was visualised under UV light previous immersion of the gel in a 5 mg/ml Ethidium bromide solution. Separation and isolation of specific bands were performed by electroelution according to Sambrook 2001 or using gel elution through silica gel columns from QIAGEN, GibcoBRL or PeqLab, following instructions from the manufacturer.

2.1.6 DNA sequencing

DNA sequencing was performed by the chain-stop reaction principle using BigDyes terminators from Applied Biosystems and TE terminators from Amersham-Pharmacia in Hybaid *Omn - E* and MJ Research *PTC-200* thermal cyclers. Sequence reactions were purified by precipitation with 3M Sodium acetate and ethanol or by chromatography through sephadex G-50 (Pharmacia) moist columns in 96 blocks array from Millipore. Sequences were analysed at Genterprise Gmbh. in 377 and 3070 automatic sequencers from Applied BioSystems and MegaBACE 2000 from Amersham-Pharmacia.

2.1.7 DNA cloning

PCR-products were amplified and cloned into pGEM-T Easy (Promega) according to the manufacturer‘s instructions. 1µL of vector solution was mixed with ATP-containing buffer and insert DNA in a final 10µL ligation reaction. Vector and insert DNA were mixed in equimolar ratio. Ligations incubated several hours at 4°C or overnight at room temperature. Ligation product was desalted by sodium acetate-ethanol precipitation and mixed with 50µL of *E. coli* RR1 electrocompetent cells for transformation through

electroporation at 2.5 kV for 5 ms. Cells were recovered at 37°C for 20 minutes before plating in LB medium with ampiciline, X-Gal and IPTG. Bacterial clones were screened for the presence of insert by selection of white among blue colonies due α-complementation (Ullmann et al. 1967). Single colonies were picked and amplified overnight in LB-Broth liquid medium containing ampicilin. 100 µL aliquots of culture were preserved in 1 mL medium containing 50% glycerol at -70°C. DNA preparation and restriction was tested by gel electrophoresis and insert sequence was determined using standard primers.

2.1.8 DNA amplification by PCR

Geometrical amplification of DNA fragments was achieved by Polymerase Chain Reaction. 20-50 ng plasmid DNA / 0.5-1µg cDNA/genomic DNA was typically used as template. PCR reactions contain 100 pmol of each primer, 0.2 mM dNTPs, PCR Buffer, 1 µl Taq Polymerase and varying concentrations of $MgCl_2$ (2 to 5 mM) and H_20 up to a final volume of 50 µl. After a denaturing step of 2 min at 94 °C, 35 to 40 cycles of annealing and elongation were repeated. Annealing temperature was empirically determined for each primer pair.

2.1.9 Radioactive DNA-labelling

50 to 500 ng of Purified PCR products of known sequence were used as templates for radioactive probes. Random primed oligonucleotides kit from Roche was employed following the fabricant's protocol. To the template DNA, a mixture of DNA hexamers with buffer and non radioactive dCTPs, dGTPs and dTTPs was added. DNA was denatured by

boiling during 10 minutes before adding αP^{32} dATPs and Klenow fragment. Labelling reaction took place from 4 hours to overnight at room temperature, or two Hours at 37°C, simultaneously with the pre-hybridisation step. Typically, blots were hybridised overnight with the labelled probe and non-labelled carrier DNA at stringent temperatures depending on (G+C) content of the probe (usually 64°C for Southern-blots and 42°C + Formamide for northern-blots). Blots were washed at stringent conditions until washing puffer radiated less than 1 Bq. Autoradiography films were exposed to the washed radioactive blot with enhancer slides at -70°C. Exposition time varied from several hours to several days depending on final radiation values. Alternatively Phospho-Imager plates were exposed at room temperature from several hours to overnight. Hybridisation signals from the Phospho-Imager plates were detected with a FUJIFILM BAS-1800 Phosphor-Imager and quantified with the software Aida V. 2.0 performing manual baseline subtraction. The values were normalised to the highest signal.

When necessary, blots were stripped by boiling with washing puffer containing SDS or at 68°C with puffer containing 50% formamide.

2.1.10 Non-radioactive DNA labelling

Digoxigenin (DIG) DNA Labeling Kit from Roche was used for non-radioactive DNA hybridisation of Southern-blots containing restrictions of PAC/BAC clone candidates, in order to confirm positive clones after screening the genomic library. The manufacturer's protocol was followed. The labelling principle here is also based on random priming hybridisation followed by DNA synthesis from the Klenow fragment, but the DIG-labelled

monomer is present in a mixture of ratio 65/35 dTTP/DIG-11-dUTP. DIG-labelled DNA was detected with anti-DIG antibody conjugated with alkaline phosphatase (anti-DIG-AP) and a subsequent colour reaction with NBT/BCIP as enzymatic substrate.

2.2 Working with Genomic Libraries

Genomic libraries from the Resource Zentrum Primary Database (RZPD) in Berlin were screened for PAC/BAC clones, possibly long and with a minimal overlap to the ones already sequenced in our lab. Repeats-free PCR product from the tips of neighbour clones and an insert-free vector (pCYPAC2) aliquot were radioactive labelled, mixed together in a 10/1 Bq ratio and hybridised over-night with filters from the genomic libraries RCPI1 3-5 constructed by Ioannou and colleagues (Ioannou et al. 1994) and available from RZPD. The 10/1 mixture of insert and vector was used to become a slight background signal and in this way help orienting the signals in the array (see figure 2.1). Positive signals were selected and after determination of their coordinates, the corresponding PAC clones were ordered to the RZPD. DNA from all PAC clones were obtained with the Whitehead/MIT P1/BAC Isolation Protocol (unpublished), digested in separate reactions with *Eco* RI and *Hind III* for size determination and to reconfirm their overlap with PAC44q2 by Southern-blot analysis. The resulting restriction reactions were loaded in a 1% agarose gel and let run over-night in slow electrophoresis. Documentation of the agarose gel was used to determine the size of the PAC clones. The agarose gel was blotted over-night to transfer all DNA in a Nylon membrane for later hybridisation. The same specific PCR product used for screening the PAC clones was used as a probe to confirm the positive clones with a DIG-DNA labelling.

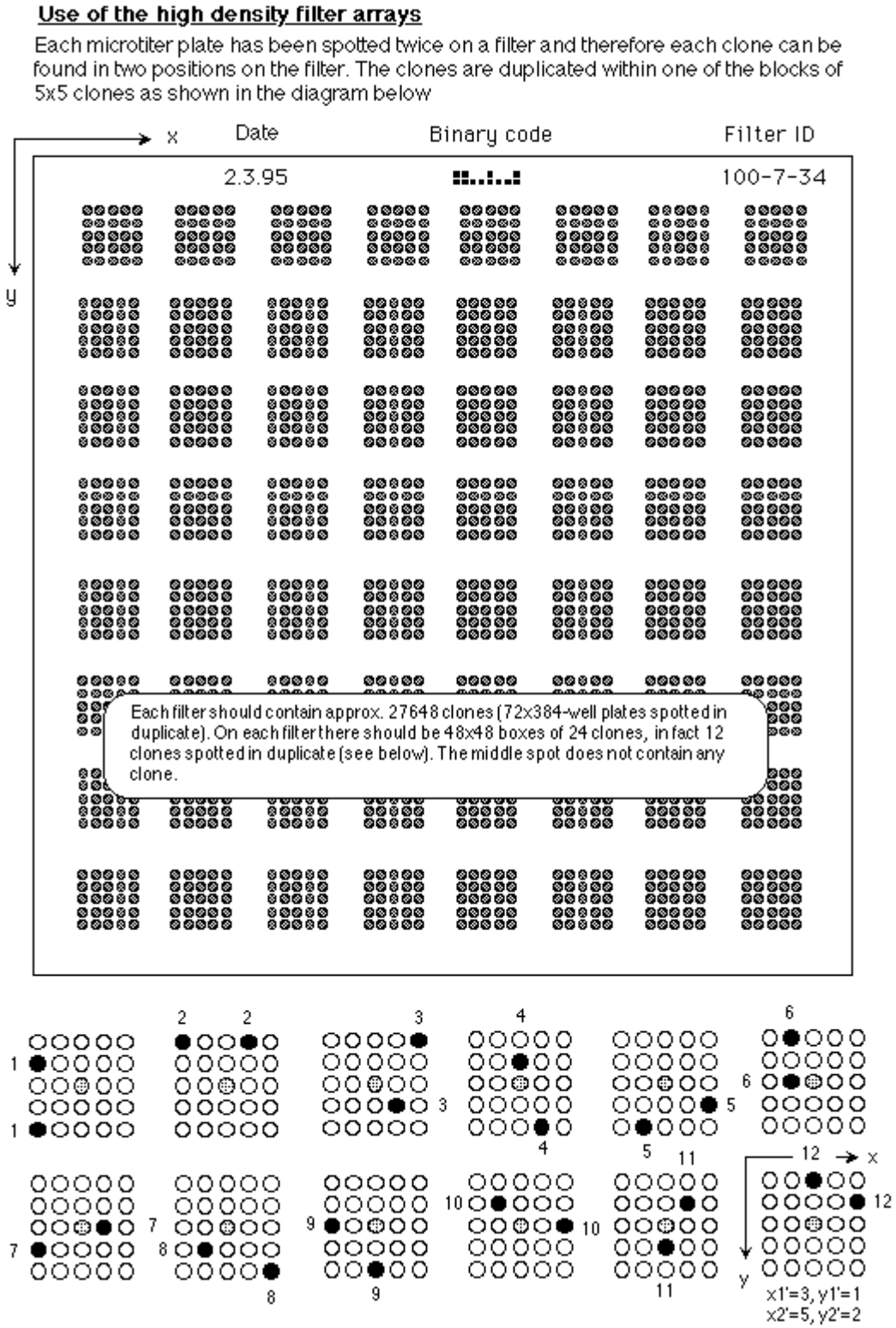


Figure 2.1: Spotting schema of RZPD filters

Positive hybridisation signal are recognised from background because of their intensity and their particular pairwise arrangement in the 5x5 block. Coordinates of the clones were determined by the following formulas: x=5(X-1) +x' and y=5(Y-1) +y', where x and y are the final coordinates of desirable clones; X and Y are the coordinates of the blocks on the filter; x' and y' are the coordinates of each clone within the 5x5 block. Source: RZPD, Berlin: www.rzpd.de

2.2.1 Genomic BAC/PAC clones used

Table 2.1 BAC/PAC clones used

Organism	Official name	Vector	Lab. Name	Size (Kb)	Final status
Mouse	BAC221D7	pBeloBAC	B4	156	Fully sequenced
Human	RPCI-4 774M4	pCYPAC2	M24	73	Confirmed
Human	RPCI-5 1013L7	pCYPAC2	L07	90	Selected/partially sequenced
Human	RPCI-5 1044O21	pCYPAC2	O21	118	Confirmed/identical to PAC44q2
Human	RPCI-5 947F14	pCYPAC2	F14	110	Confirmed
Human	RPCI-4 533N3	pCYPAC2	N03	105	Confirmed
Human	RPCI-3 508L18	pCYPAC2	L18	n.a.	Negative
Human	RPCI-3 449N9	pCYPAC2	N09	110	Confirmed

Clone BAC221D7 was produced and mapped by (Kleyn et al. 1996). RPCI Libraries produced by (Ioannou et al. 1994). Clone O21 was found to be identical to Amid's PAC44q2 and clone N03 was found to be negative (wrong coordinates).

2.3 Preparation of a Shotgun library

High molecular weight DNA from BACs/PACs was prepared according to the Whitehead protocol and then sequenced with a combination of shotgun and primer walking methods standardised in our lab by Clara Amid in her doctoral thesis (Amid 2002). BACs, PACs or any other vector with large inserts (up to 300 Kb) were mechanically shredded at random by nebulisation to guarantee the production of overlapping fragments. The DNA is shredded by letting it circulate through the tiny hole of a Nebuliser (GATC, Konstanz) at 1 bar during 45-60 seconds (the size of the resulting fragments is inversely proportional to the time the DNA-solution circulates and the pressure). The DNA fragments were size fractioned through electrophoresis in a 1% w/v preparative agarose-gel (figure 2.2); the result is a characteristic continuous spread of DNA from different sizes with no recognizable bands. Three different size ranges were selected to construct the sub-libraries, namely 0.5-

1.2 Kb, 1.2-2.5 Kb, 2.5-4.5 Kb. Accordingly, gel extraction (with QIAGEN and Gibco BRL Gel extraction Kits) was performed taking care that the DNA to be prepared did not get in contact with Ethidium bromide or UV light, using molecular weight markers as guide. Errors in the sequencing phase may be produced through the base exchange caused by these mutagenic factors.

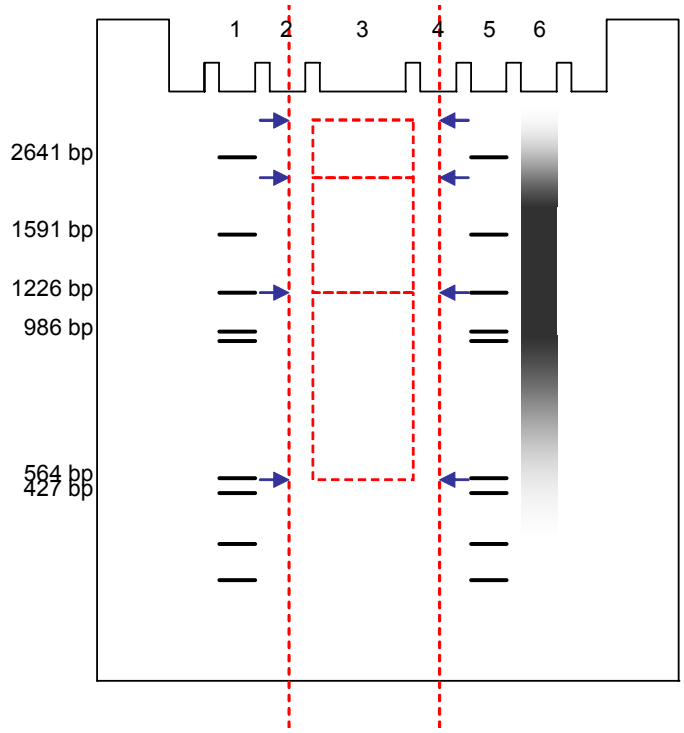


Figure 2.2: Schematic representation of a preparative agarose gel used for size fractionation.
1, 5: DNA-MW marker; **2, 4:** empty; **3** (broader lane): nebulised DNA; **6:** a sample of nebulised DNA. Middle section of the gel with the nebulised DNA was cut off along the empty lanes and kept apart. Left and right gel sections with MW marker and sample nebulised DNA were stained with ethidium bromide and marks were cut on the gel over the UV transiluminator (blue arrows). Sections were put together again over a clean surface and the marking-cuts were used to guide the excision of rectangles (red dashed boxes) for DNA gel extraction.

DNA fractions were recovered from the gel fragments by using gel elution through silica gel columns from QIAGEN, following instructions from the manufacturer. The fractions of DNA were then ligated with standard vector, in our case pUC18. In order to improve the ratio foreign DNA insertion / plasmid re-ligation, pUC18 treated with terminal phosphatase was used (Amersham). With the three sub-libraries competent *E. coli* cells were transformed and selected by ß-galactosidase disruption (Ullmann et al. 1967). White

colonies were picked and let grow in LB-Medium in 96-block format as described in the Material and Methods section. The shotgun sub-libraries were sequenced from one side (i.e. universal primer) circa 1,000 sub-clones per 100 Kb BAC/PAC clone.

2.4 RNA Methods

2.4.1 RNA isolation

Total RNA isolation from mouse tissues, was made with the guanidinium thiocyanate-phenol-chloroform extraction (Chomczynski and Sacchi 1987). mRNA was isolated from total RNA with the QIAGEN polyA+ kit, following manufacturer's instructions.

2.4.2 RT-PCR

Reverse Transcription was performed on 5 to 10 µg total-RNA with SuperScript II (Gibco BRL) following manufacturer's protocols. Poly-T primer was used to synthesise first strand cDNAs from the messenger RNAs by using their natural polyAs as priming sites. Afterwards, the first strand cDNA was used as template for PCR amplification with gene specific primers.

2.4.3 RACE

Rapid amplification of cDNA ends (RACE) from GibcoBRL was used following manufacturer's instructions. Briefly, this method produces PCR amplifications between a known region and the 5' or 3' –ends of a given RNA. In this work, 5' RACE was used to establish transcription's start points. In the 5' RACE protocol, total RNA is used as template for a cDNA first strand synthesis with a reverse transcriptase and a gene specific reverse primer. The cDNA produced corresponds to the Crick (reverse) strand of the given transcript. A poly-C tail is added using the Terminal deoxynucleotidyl transferase (TdT) to the 3' end of the cDNA. Finnally, the 5' unknown end is amplified by using a primer complementary to the poly-C tail and a nested gene specific reverse primer.

2.4.4 RNA in-situ hybridisation

T7 viral RNA-polymerase promoter adaptors were appended by PCR amplification to a gene-specific DNA-probe (See the primer sequences in table 7.4). This probe was further employed as a template for RNA in vitro synthesis. During the in-vitro transcription, the probe was labelled with digoxigenin (RNA in-vitro transcription from Roche). The probe was hybridised with frozen slides from mouse organs and detected with an anti-digoxigenin antibody coupled with alcaline phosphatase to distinguish RNA-to-RNA signals in a cell-specific manner.

The position and orientation of the adaptors determine the direction of transcription and hence the strand to be produced. Two different separate probes were produced, the RNA-probe which corresponds to the sense-strand and hence to the mRNA, and the one

corresponding to the anti-sense RNA, both labelled with Digoxigenin. The rational of the experiment, shown in the figure 2.3, is that probes with anti-sense RNA may produce RNA-RNA homoduplex with the native mRNAs present in the frozen slides, whereas the sense RNA should not produce any signal and functions as negative control of the experiment.

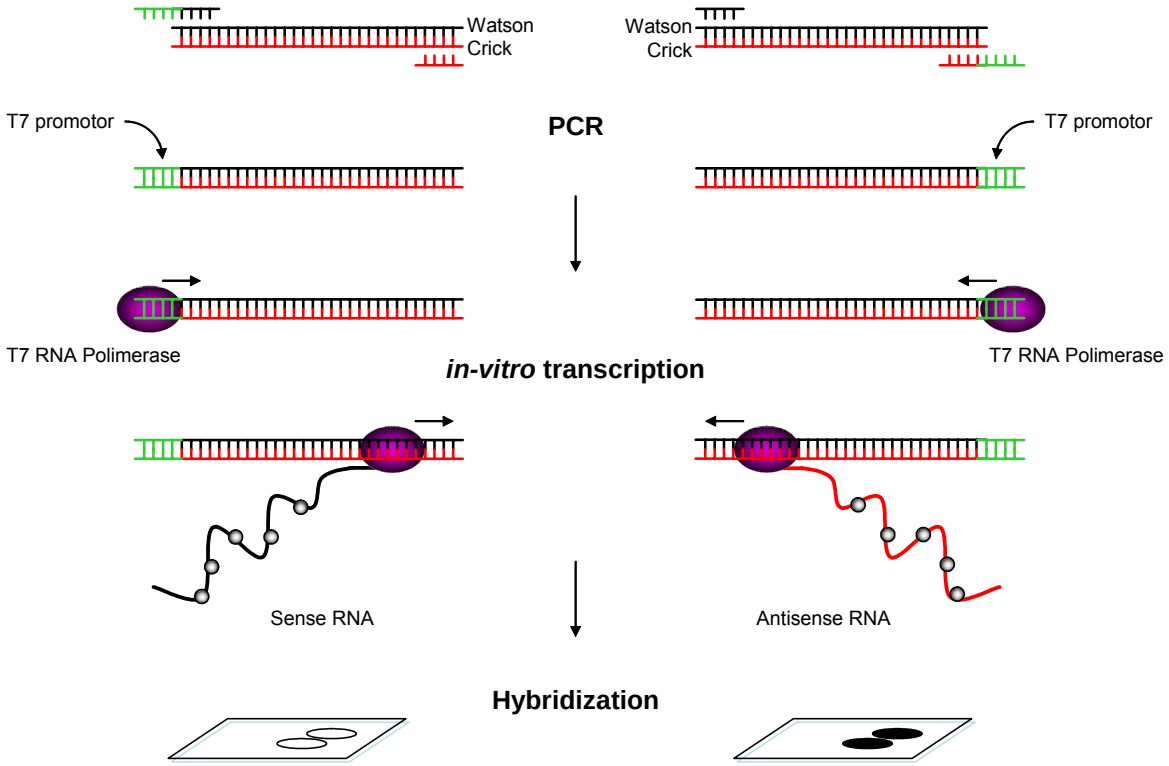


Figure 2.3: Principle of the RNA in-situ hybridisation experiments.
All RNA polymerases synthesise in the 5' - 3' direction. Accordingly, in-vitro transcribed RNA using template with T7 promotor structure appended upstream from the Watson strand, yields RNA with sequence corresponding the native mRNA; whereas, T7 promotor structure appended upstream from the Crick strand yields RNA with sequence complementary to the native mRNA.

2.5 Design of synthetic peptides for antibody production

Two polyclonal antibodies against synthetic petides were produced. Regions outside conserved domains were choosen to minimize cross-reactivity of the antibody with other protein members. Human-mouse conserved regions were preferred in order to produce potential multi-species antibodies.

The potential peptides were tested by hydrophilicity by manually inspection of the sequence and analysis with the Antigen program from the GCG package. PROSITE was used to discard modification sites. Tmpred and TMHMM were used to discard transmembranal domain. SSpro2 was used to confirm external orientation. BLASTP using low stringency parameters (Expect value=1000, word size=2) was used to confirm specificity of the region chosen as epitopes. Antibodies were produced in rabits by GenMed Systems, USA.

2.6 Computers methods

2.6.1 Standard programs

Programs from the **DNAStar** package versions 4.0 and 5.0 (LaserGene) were routinely employed. **EditSeq** for editing and handling nucleic acid and amino acid sequences; **SeqMan**, for assembly of contiguous sequences and shotgun assembly; **MapDraw**, for restriction analysis of known sequences and **MegAlign**, for aligning homologous nucleic acid and amino acid sequences. Additionally, DotPlot was performed with **MegAlign** using the following parameters: 50b minimum gliding windows of at least 60% similarity. With these parameters a good signal-to-background ratio is obtained. This analysis reveals sequence conservation even when the extent is low.

The program **Sequencher** was alternatively used for assembly of contiguous sequences and shotgun assembly. **PAUP** and **Mega** were used for phylogenetical analysis. **GCG** and other programs from the **HUSAR** service of the DKFZ (Heidelberg) were used for

a variety of purposes like (G+C)-plot, antigen analysis, EST-assembly and batch BLAST-searches. (G+C)-plot was generated with a sliding window of 1000 bases with **Window** and depicted graphically with **Figplot**, both subprograms from the GCG package available online through the HUSAR interface from the DKFZ. Protein three-dimensional representation was performed with **SwissPDV-Viewer**.

Genome annotation and gene prediction was performed in the Rummage Sequence Annotation Server (Taudien et al. 2000).

2.6.2 On-line resources

Table 2.2 On-line resources

General resources	
GOLD, Genomes online database	www.genomesonline.org/
DOGS, Database of genome sizes	www.cbs.dtu.dk/databases/DOGS/
HUGO Nomenclature committee	www.gene.ucl.ac.uk/nomenclature/
IMAGE clone ordering (user account required)	www.rzpd.de
HUSAR (On-line GCG, ESTclustering, user account required)	genius.embnet.dkfz-heidelberg.de
ClinicalTrials.gov	www.clinicaltrials.gov/ct
Three-genome comparison between Rat, Mouse, and Human	www.genboree.org
BLAST searches (Altschul et al. 1997)	
at NCBI, USA	www.ncbi.nlm.nih.gov/blast
at EMBL, Europe	www.ebi.ac.uk/blast2/
at DDBJ, Japan	www.ddbj.nig.ac.jp/E-mail/homology.html
Gene annotation	
Rummage, A High-Throughput Sequence Annotation Server (user account required)	gen100.imb-jena.de/rummage/
UniGene: non-redundant set of gene-oriented clusters	www.ncbi.nlm.nih.gov/UniGene
OMIM: Online Mendelian Inheritance in Man	www.ncbi.nlm.nih.gov/Omim1
Gene Expression Atlas	expression.gnf.org
GeneCards	bioinformatics.weizmann.ac.il/cards

Genome annotation and analysis	
NCBI's genome map viewer	www.ncbi.nlm.nih.gov/mapview
Wellcome Trust, Sanger Institute Genome browser	www.ensembl.org
UCSC Genome Browser	genome.ucsc.edu/
Celera genomics (user account required)	publication.celera.com
Percent identity plot	bio.cse.psu.edu/pipmaker/
VISTA servers	gsd.lbl.gov/vista/index.shtml
Repeat Masker	
Washington	www.repeatmasker.org/
Heidelberg	woody.embl-heidelberg.de/repeatmask/
Primers evaluation and design	
Oligonucleotide Properties Calculator. Northwestern University Medical School. Chicago	www.basic.nwu.edu/biotools/oligocalc.html
Primer3. Whitehead Institute for Biomedical Research.	www-genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi
Protein analysis	
PredictProtein	www.embl-heidelberg.de/predictprotein/predictprotein.html
SWISS-MODEL An automated comparative protein modelling server	swissmodel.expasy.org
FingerPRINTScan	bioinf.man.ac.uk
Motif Scan	hits.isb-sib.ch/
SSpro, SSpro8, ACCpro	www.igb.uci.edu/tools/scratch/
Prosite	www.expasy.org/prosite/
SMART, Simple modular architecture research tool	smart.embl-heidelberg.de/
Polish bioinformatics site	bioinfo.pl/
The protein kinase resource	pkr.sdsc.edu/
Protein sub-cellular localisation	
PSORT, prediction of protein sorting signals and localisation sites in amino acid sequences	psort.nibb.ac.jp
SubLoc, Prediction of protein subcellular localisation	www.bioinfo.tsinghua.edu.cn/SubLoc/
ProtCom	www.softberry.com
PredictNLS, prediction and analysis of nuclear localisation signals	maple.bioc.columbia.edu/predictNLS
Trans membrane domains	
TMHMM, Prediction of transmembrane helices in proteins	www.cbs.dtu.dk/services/TMHMM-2.0
TMpred, Prediction of transmembrane regions and orientation	www.ch.embnet.org/software/TMPRED_form.html
DAS, Transmembrane prediction server	www.sbc.su.se/~miklos/DAS/maindas.html
Ka/Ks (dn/ds)	
SNAP HIV Program for the analysis of synonymous /non-synonymous substitutions	www.hiv.lanl.gov/content/hiv-db/SNAP/WEBSNAP/SNAP.html

2.7 Standard solutions and materials

Table 2.3 Standard solutions

Antibiotica	Ampicillin 100 µg/ml Cloranphenicol 25 µg/ml Kanamycin 25 µg/ml
Antifading / p-phenylendiamide	1mg p-phenylendiamide in 1mL 50% v/v phosphate-buffered glycerol: 1mM NaCl 1mM Na_2HPO_4 mixed until reaching pH 8 with 1mM NaCl 1mM NaH_2PO_4 mixed 1:1 with glycerol
Denaturing buffer	1.5 M NaCl 0.5 M NaOH
Dialysis buffer	0.025 M Tris-HCL 0.3 M NaCl 0.01 M Na_2EDTA
E-Buffer (Electrophoresis buffer)	0.036 M Tris-HCL 0.03 M NaH_2PO_4 x $2H_20$ 0.01 M Na_2EDTA
Ethidium bromid solution	5 µg/mL in 1X E-buffer
Hybridisation solution	7% w/v SDS 0.5 M Na_3PO_4 pH 7.2 0.001 M EDTA
Neutralisation Buffer	3 M NaCl 0.5 M Tris-HCL ph 7,5
LB-Medium	10 g Trypton 5 g Yeast extract 5 g NaCl ad 1,000 mL A. bidest
Loading buffer (for RNA gel)	500 µL Formamid (deionised) 100 µL 10X MOPS-Buffer 150 µL Formaldehyde (filtrated)
MOPS (10X)	50 mM sodim acetate 10 mM EDTA 0.2 M 3-morpholinopropanesulfonic acid, pH 7.0
PBS buffer (10X)	1.3 M NaCl 27 mM KCl 65 mM Na2HPO4

	15 mM KH2PO4 pH to 7.2
Pre-incubation medium (1XPM)	0.02% w/v ficoll 400 0.02% w/v polyvinylpyrrolidon 0.02% w/v BSA in 3X SSC
Solution 1 (Whitehead protcolol)	0.01 M Tris-HCL
Solution 2 (Whitehead protcolol)	0.2 M NaOH 1% w/v SDS
Solution 3 (Whitehead protcolol)	50 ml potassium acetate 7.5 M 23 ml acetic acid (concentrated) 127 ml H_2O
SSC (Standard saline citrate buffer)	0.15 M NaCl 0.015 M Natriumcitrat
TBE Buffer (Tris-borat electrphoresis buffer)	0.09 M Tris-HCL 0.09 M boric acid 1.25 mM Na_2EDTA
TE Buffer	10 mM Tris pH8 1 mM Na_2EDTA

E. coli strain

RR1: HB101 F^- Δ(*gpt-proA*)66 *leu*, *supE44, ara14, galK2,* Δ*(mcrC-mrr), lacY1, rpsL20, xyl-5, mtl-1, recA13,*

3 Results

3.1 Sequencing at the genomic level

3.1.1 Selection of BAC and PAC clones

The murine BAC221D7 was mapped by Kleyn (Kleyn et al. 1996) in their positional cloning of the Mouse Obesity Gene tubby, the furthermost telomere-oriented genetical marker of our ~1 Mb sequencing effort. This clone was already selected to bridge the existing gap left from clones BAC377M11 (Amid et al. 2001) and BAC282L1 (Brueckmann, in preparation) between genes *St5* and *Lmo1*. Clone BAC221D7 does not overlap with BAC377M11 at all but the positions of both clones in the map by Kleyn et al., suggested that the remaining gap between the two was short enough to be closed with a long range PCR. One primer from the telomere-oriented end of BAC377M11 and two from both ends of BAC221D7 were designed and Expand High Fidelity PCR (Roche) using mouse genome DNA as template were tested with both combinations. The resulting ~3 Kb positive PCR anchored the orientation of BAC221D7 and the product was sequenced by primer walking. The orientation of these clones may be observed in the figure 3.2(B).

The human clone PAC44q2 (Amid et al. 2001) was used as template to generate the molecular probe for screening genomic libraries in search for overlapping PACs covering

the region of interest. Primers were designed to amplify a 600b long unique and repeat-free DNA fragment by PCR, as close as possible to the end point of PAC44q2 in the telomere direction. The probe was radioactively labelled with 32P and hybridised with 11 high-density filters from the genomic libraries RCPI1 3-5 (Ioannou et al. 1994) provided by the Resource Zentrum Primary Database (RZPD) in Berlin. Six clearly positive signals were detected, the corresponding PAC clones were identified and ordered from the RZPD and DNAs was prepared according to the Whitehead Protocol. Figure 3.1 shows the *Eco* RI and *Hind III* digestions of the clones used for size determination and the Southern-blot hybridisation reconfirming overlap with PAC44q2.

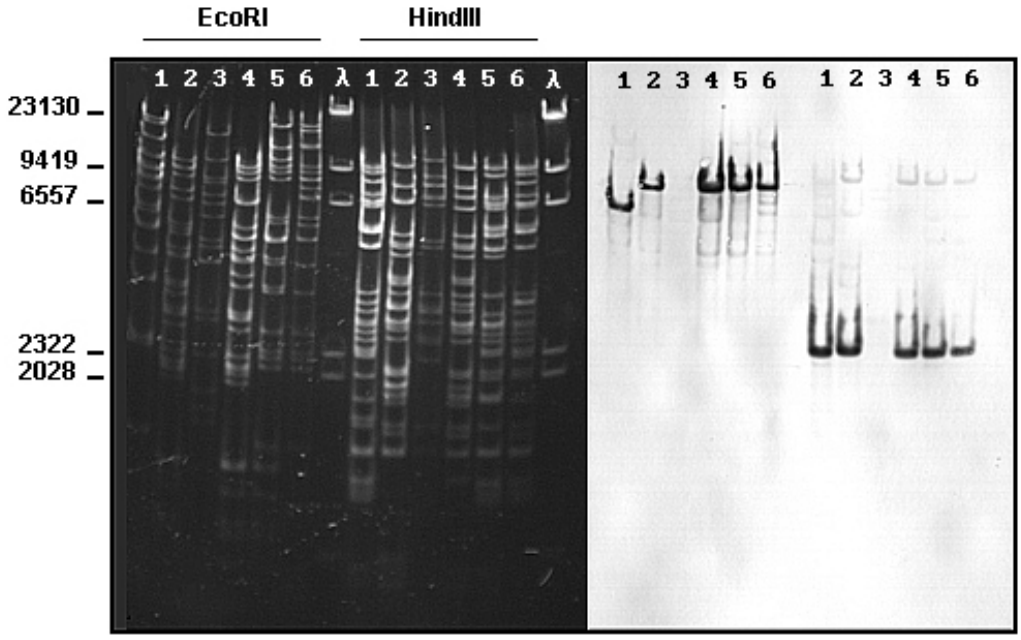


Figure 3.1: Size determination and confirmation of human PAC clones candidates.
The left panel shows the electrophoresis of *Eco* RI and *Hind III* digestions of PAC-DNA candidate clones. Resulting sizes are shown in the figure 3.2. The right panel shows the results of the Southern blot hybridisation with a PAC44Q2 specific probe labelled with biotin. The third candidate is clearly a false positive from the library screening. Resulting labelled restriction bands from digestion with *Eco* RI and *Hind III* have different sizes. Interestingly, the band from the *Eco* RI restriction of the first clone is shorter as the corresponding bands from the other clones. After end-sequenzing it was clear that this band is produced from one *Eco* RI site in the insert and the other site is in vector. All other bands are produced by sites within the insert.

The Nylon membrane with the blotted DNA was stripped to reuse it for hybridisation. A second probe was produced from the centromere e-oriented end of human PAC781k3 clone (kindly provided by T. Brueckmann) to test whether some of the candidates were large enough to close the gap with PAC44q2, unfortunately this was not the case for any of them. Also PCR experiments with the PAC44q2 and PAC781k3 specific primers were performed. Reproducibly, all confirmed positive clones produced PCR fragments with PAC44q2 specific primers, but none of them with PAC781k3 specific primers. Clearly the gap on the human was still too large to be covered with just one BAC/PAC clone or the candidates were still too short. All clones were end-sequenced to anchor their positions relative to PAC44q2. With the size information of each clone and their starting positions relative to PAC44q2, the candidate with the minimum overlap was chosen. Despite its sub- optimal length human PACL071013 clone was chosen for sequencing due its longest extension beyond PAC44q2 in telomere direction.

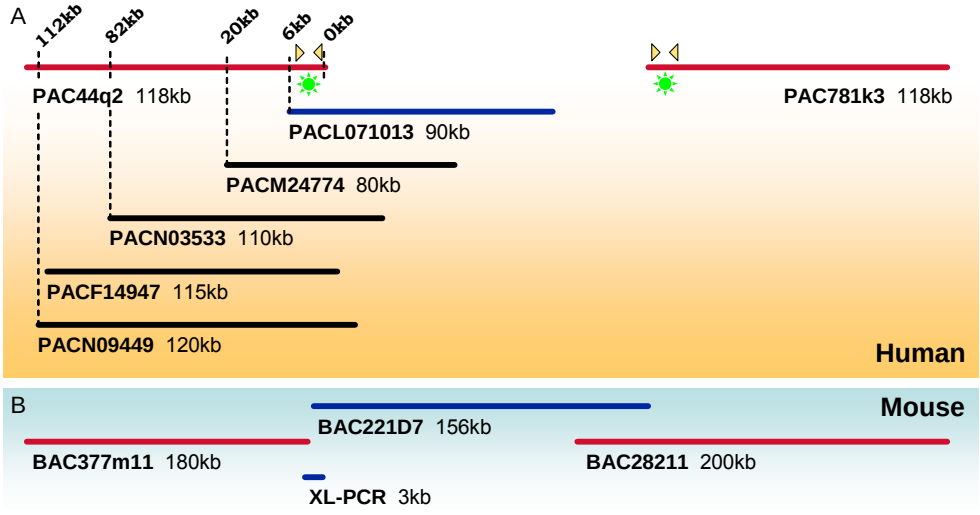


Figure 3.2: Map of the sequencing region in human and mouse.
Flanking clones from Amid's (left) and Brueckmann's (right) doctoral theses are represented in red, the blue ones were sequenced in this work and the black ones were screened but discarded for sequencing. The upper panel shows the sequencing region in human. The green stars show the relative positions of the Biotin-labelled probes used to confirm the overlap with PAC44q2 and to test overlap with PACL071013. Yellow triangles show the relative position of primers used to produce these probes by PCR. All candidates tested are shown under PAC44q2, their sizes and relative positions.

The lower panel in the figure 3.2 depicts the situation in the mouse with the XL-PCR product showing minimal overlap to BAC377M11 and BAC221D7. BAC221D7 overlaps ~30 Kb with BAC28211.

3.1.2 Sequencing strategy

Murine clone BAC221D7 and human PACL071013 were sequenced following a combined shotgun-primer walking strategy (Amid 2002; Bahr 1999) described in material and methods section. Figure 3.3 shows an evaluation of the size fragmentation (lanes 2-4) of the randomly sheared DNA of PACL071013 in the three selected ranges: 2.5-4.5 Kb, 1.2-2.5 Kb and 0.5-1.2 Kb, and their subsequent ligation (lanes 6-8) in pUC18, constituting the large, medium and short banks respectively.

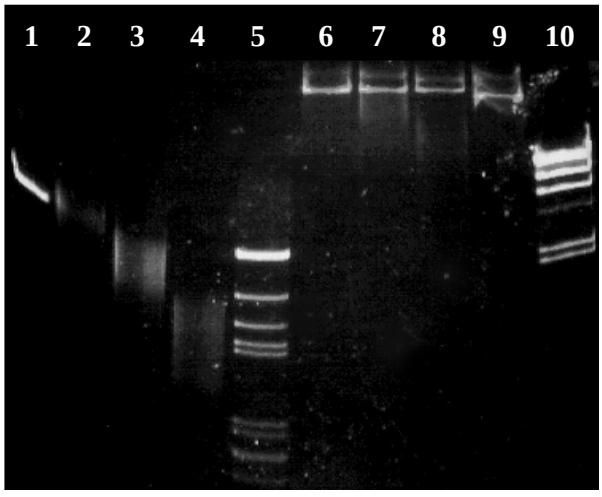


Figure 3.3: Evaluation of PAC1013L07 shreeding, separation and cloning of the fragments.
1: control linear pUC18. 2: fragments 2.5-4.5 Kb. 3: fragments 1.2-2.5Kb. 4: fragments 0.5-1.2 Kb. 5: Molecular weight marker pF digested with *Eco* RI and *HindIII*. 6, 7, 8: verification of ligation of the fragments in pUC18. 9: verification of circularisation from control primer. 10: Molecular weight marker λ *Hind* III

Typically 1000 sub-clones per 100 Kb insert size of the BAC/PAC clones were sequenced. The resulting *"single-reads"* were computer-assembled in contigs (contiguous sequences) using the program SeqMan (DNAStar, Lasergene). Those clones having the potential of closing the gaps were sequenced from both ends. The remaining gaps were then closed through primer walking extending the tips of the contigs or by direct sequencing the PCR products constructed between these tips. Also the regions with coverage in just one direction were sequenced from the other side by primer walking or reverse sequencing of selected sub-clones. A schema of the sequencing strategy is shown in the figure 3.4.

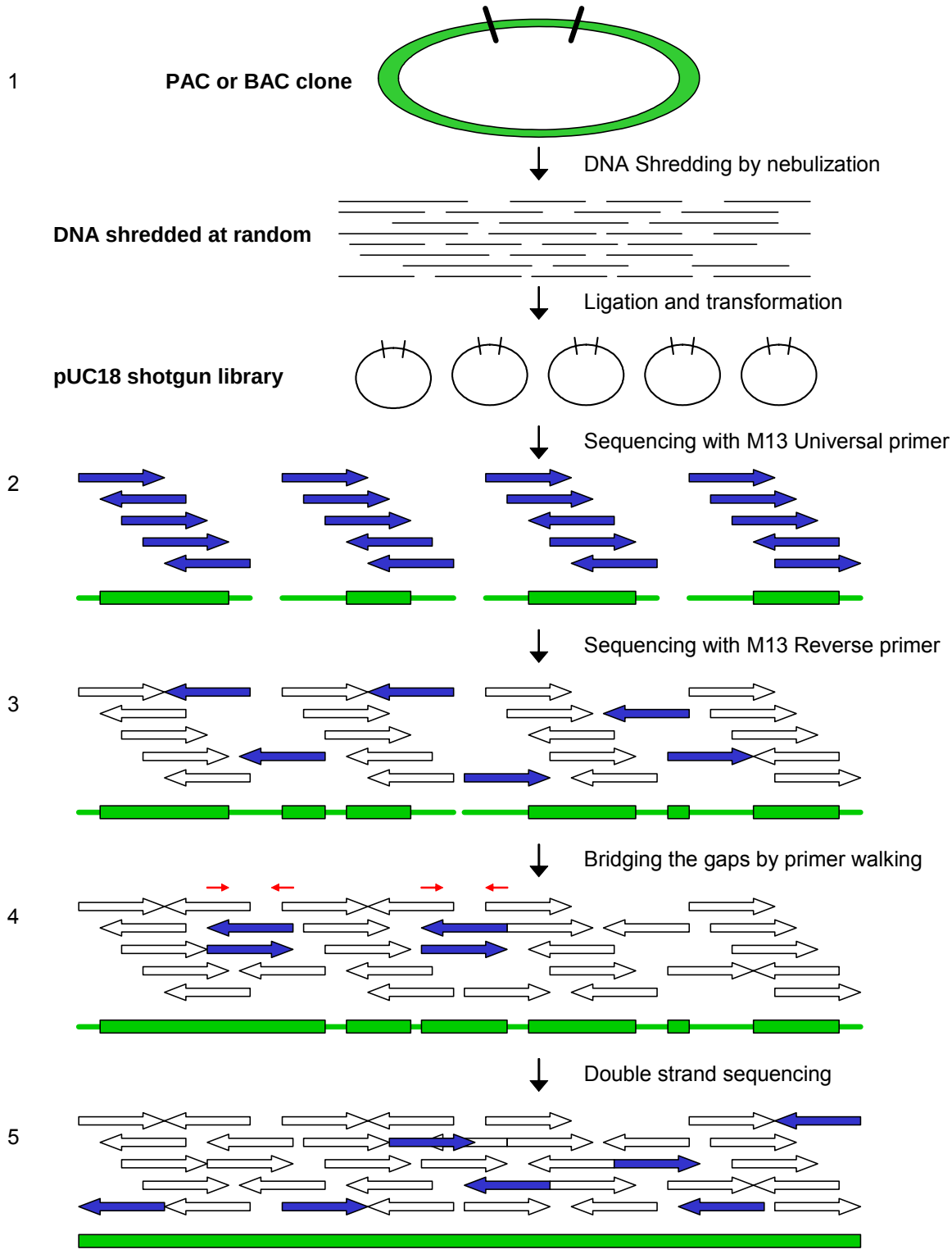


Figure 3.4: Principle of the combined shotgun-primer walking sequencing strategy.
Blue arrows represent the new reads in each phase; green bars represent coverage of the problem sequence in both strands, green lines in just one strand. **1:** Construction of the shotgun library as described in 3.2. **2:** Sequencing with M13 Universal primer (~1000 subclones/100 Kb PAC or BAC). **3.** A selection of clones are sequenced with the M13-Reverse primer. 4 and 5: Finishing face, primer walking (primers in red) and reverse sequencing of selected long reads for closing remaining gaps and completing the sequencing in both directions. (Modified from Bahr 1999).

3.1.3 Quality control of the shotgun DNA-library

Since thousands of single-reads are necessary to sequence a single BAC or PAC clone by pure random sequencing, controls in the first rounds of sequencing are necessary to guarantee the quality of a shotgun library. A maximum of sub-clones should carry DNA from the species being sequenced, the vector sequence should not be over-represented and a minimum of sequences-loss must be reached. High frequency of bacterial DNA-contamination indicates a low quality of BAC/PAC preparation. High frequency of sequence from sub-cloning vector (i.e. pUC18) results from empty plasmids. Over-representation of the BAC/PAC may result from instability of the vector.

Table 3.1: Quality assessment of the first human PAC1013L07 96-sequencing Block

	No	**%**	
Clean Single Reads	63	66	49 clearly human
Vector pCYPAC2	18	19	
Vector pUC18	7	7	
E. coli	1	1	
Heterogeneous	0	0	
Poor Data (Empty or too weak)	7	7	
	96		
	bp		
Vector Size	20,000	**18**	
Insert Size	90,000		
	110,000		

As shown in table 3.1, a preliminary survey of the sequencing from PAC1013L07 displayed low amount of *E. coli* contaminants signalling a good PAC preparation and the vector also seemed to be stable, all of them signals of a shotgun sub-library of good quality.

Quality control was also assessed by agarose electrophoresis as shown in the figure 3.5. Preparations were considered "good DNA" if both the size of the fragments and their quantity were uniform.

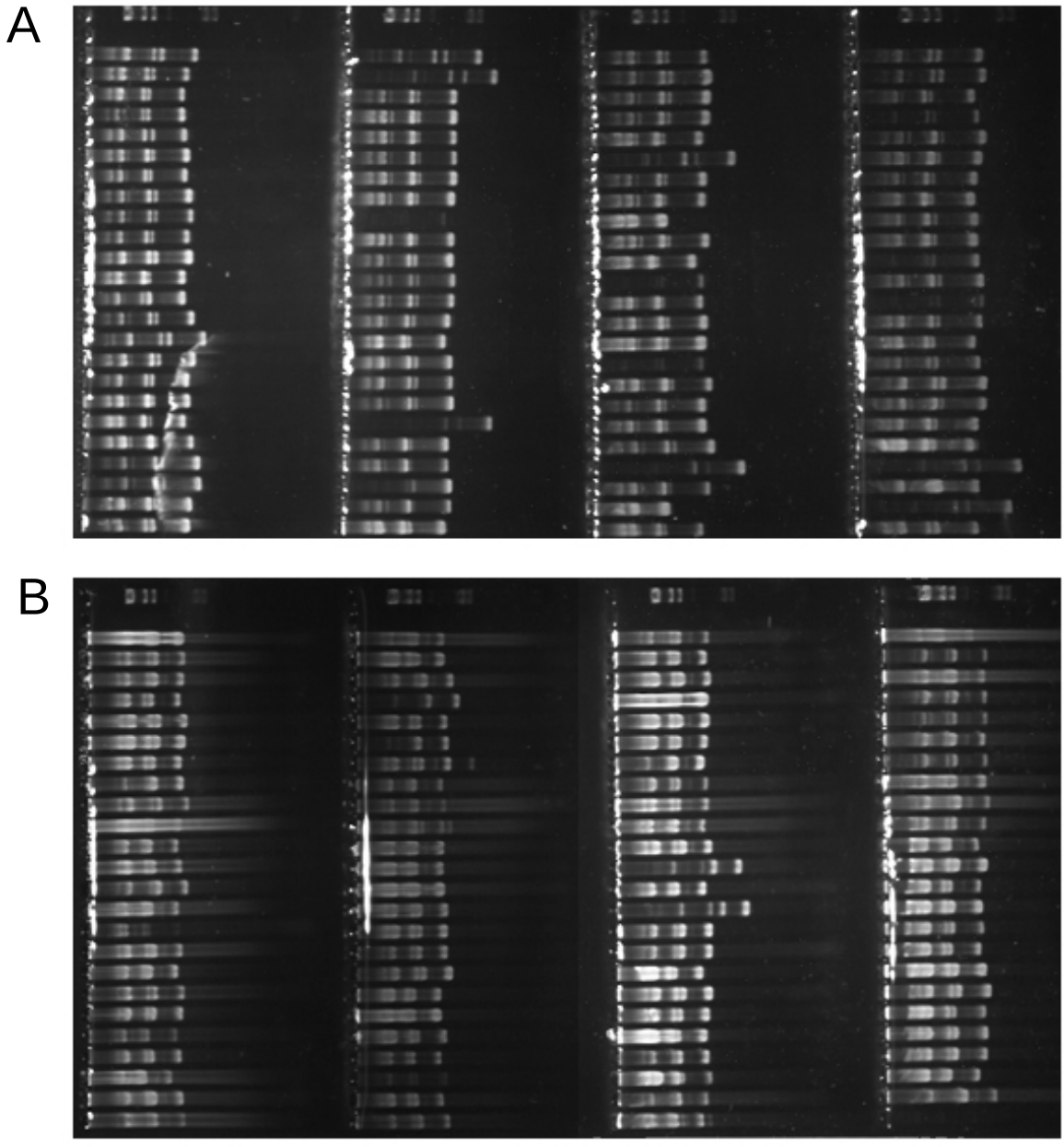


Figure 3.5: Agarose electrophoresis of two selected blocks from the sequencing of the human PAC clone PAC1013L07.

Block in A was the result of the QIAGEN R.E.A.L. Prep 96 Plasmid Kit. Block B is the result of the MACHERAY-NAGEL Nucleo Spin Multi-96 Flash Kit. In both cases 5µl plasmid prep was loaded and λ-*Hind III* DNA-MW marker was used.

3.1.4 Sequencing Statistics

Murine clone BAC221D7, together with the XL-PCR product overlapping with BAC377m11 (Amid 2002) were sequenced by assembling 2,054 single reads up to a redundancy of 6.6. This implies that, in average, every nucleotide was sequenced almost seven times. Human clone PAC1013L07 sequencing was halted in the first rounds of shotgun sequencing. Only 254 single reads were produced, from which 169 aligned together in 46 contigs with more than two reads but less than fourteen. The remaining 85 single reads did not aligned. Table 3.2 resumes some general sequencing statistics.

Table 3.2: Sequencing statistics

Clone name	Total Length	TotalSeq. Length	Number Seq.	Avg. Length	Avg. Coverage	Number of primers
BAC221D7 + XL-PCR (AJ307671)	158,201	1,051,050	2,054	512	6.6	31
PACL071013		110,843	254	436	<3.7	0

In the AJ307671 murine sequence, eight conflicts remain unsolved. The standard IUPAC codes **M** and **Y** were used for (**A** or **C**) and (**T** or **C**) respectively. Six conflicts between coordinates 88,994 and 89,031 lay in a 266-bases long simple-repeat (CACCT)n region followed by an 88-bases long C-rich low-complexity string. Two additional conflicts (**Y** = **A** or **C**) in coordinates 157,343 and 157,346 are also associated to the (CCCCT)n simple-repeat.

3.2 Computer aided sequence analysis and gene-prediction

3.2.1 Gene-prediction

As soon as the genomic sequence of the mouse, spanning clone BAC221D7 and the XL-PCR product was established, it was submitted to the EMBL/Genbank databases under the Accession Number AJ307671. The major analysis was performed using the Rummage annotation system (Taudien et al. 2000). This system combines several stand-alone programs and applications, such as exon prediction, gene-prediction, BLAST searches GC-content and CpG island estimation. All results are listed together in tables and presented graphically as shown in the figure 3.6. The most extensive prediction, from GENSCAN, resulted in three separate potential transcripts. BLASTP searches of the translated products from the last two of them produced partial matches to protein kinases in the protein databases, but none of the predicted transcripts had the potential for coding a complete kinase in its whole extension. The predicted exons were poorly represented in the mouse expression data in the form of Expressed Sequence Tags (EST) data base. However, several matching with non-contiguous human ESTs were found, strongly indicating the presence of active genes in this region of the mouse genome.

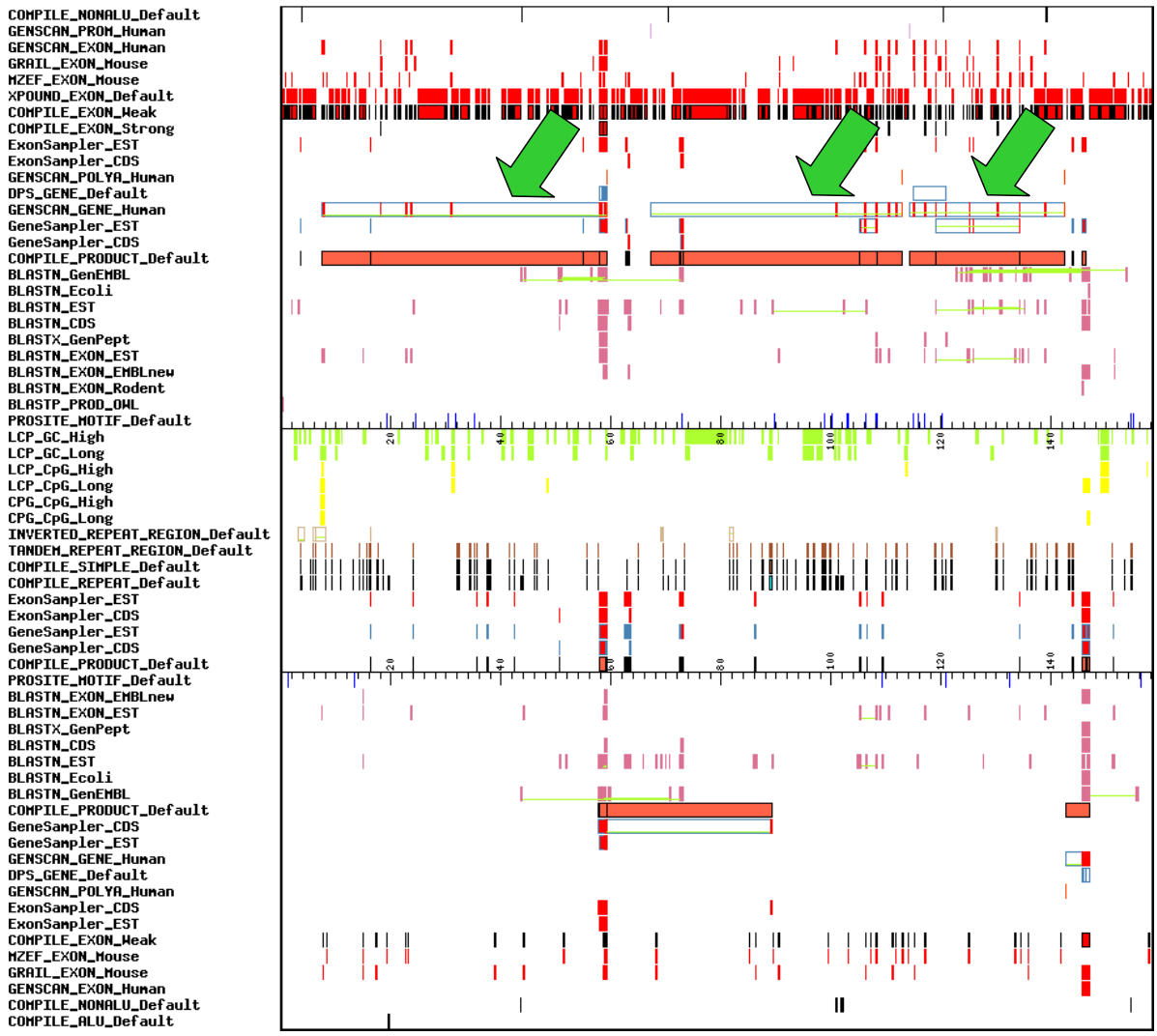


Figure 3.6: Graphic output from Rummage analysis on the mouse genomic sequence.
Upper panel shows results in centromere e direction. Middle panel shows results not necessary dependent on orientation, like (G+C) content and inverted repeat regions. Lower panel shows results in reverse complement orientation (telomere direction). Superimposed green arrows show the three different predictions found in the region.

The murine genomic sequence contained in BAC221D7 was used as scaffold for further gene-prediction in the corresponding human sequence using comparative sequence analysis. For this purpose PACL071013 was screened and partially sequenced. Also, the partial sequence of a corresponding human PAC clone from Washington University in St. Louis USA was found in the public databases under the accession number AC016718. This entry was in the first phase of sequencing, i.e. the contigs were unordered with gaps of

unknown size between them. Sequences from Washington University PAC clones were downloaded and assembled with the partial sequences from PACL071013 and PAC781k3 from our collaboration partners in the Children's Hospital of Mainz (Brueckmann, in preparation). From this analysis, it was confirmed that the corresponding sequencing from Washington University was in a much more advanced status and had the potential of covering and closing the gap still open in the clone screening presented in this work.

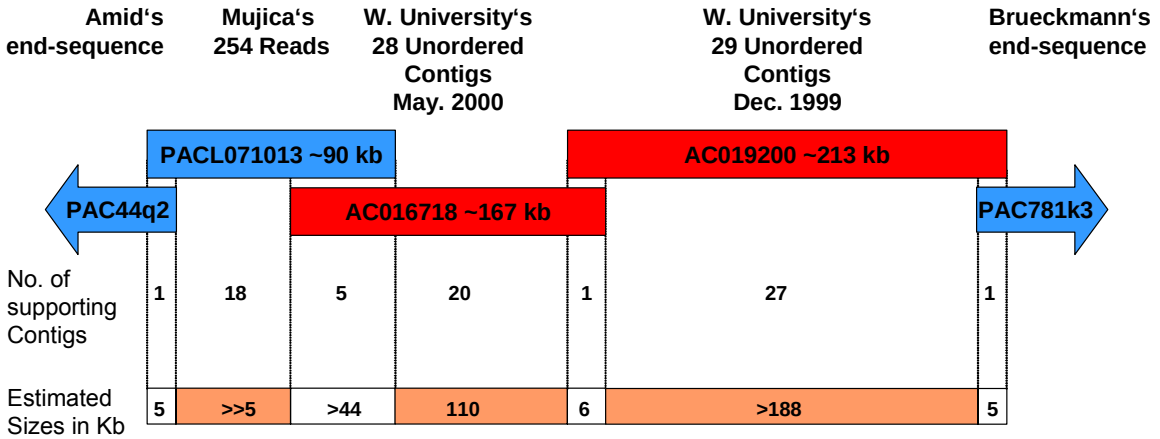


Figure 3.7: Relative positions of human clones from our sequencing project together with the competing ones from Washington University after computer-driven assembly (stand middle 2000).

Blue boxes and arrows indicate the clones being sequenced in Mainz; the red ones indicate the clones being sequenced in St. Louis. The number of contigs supporting zones with and without overlap between the clones is shown under the boxes. Estimated total sizes of these regions are indicated in the base of the figure. This numbers were still modest estimates due the incompleteness of the data. This alignment still had gaps of unknown size and position.

Being the sequence from the mouse ready and instead of competing against a centre with superior sequencing and bioinformatical facilities, the emphasis was put on the gene-prediction with the available data. The shotgun sequencing of human PACL071013 was halted and the comparative analysis between human and mouse was started immediately with the data available, despite the fact that either our sequencing effort nor the international one have had the human genomic sequence complete to that date.

The method used to perform comparative sequence analysis between our complete mouse genomic sequence and the incomplete syntenic one from the human, involved de-novo gene-prediction combined with mRNA/Genomic DotPlot analysis, sequence alignment and BLAST similarity searches against expression databases EST, in a gene-finding strategy which may be summarised as follows. The murine exons from the predicted transcripts in the Rummage analysis were joined together in-silico, in putative mRNAs and compared by DotPlot with the incomplete human AC016718 genomic sequence entry from Washington University. In this way putative mouse exons very well conserved in human were confirmed and the ones showing no signal were marked as insecure. The detected human exons were tested carefully for the presence of consensual exon/intron splicing signals at the genomic level. The filtered human exons found in this way were ordered, put together and compared by BLAST search with the human expression data in the form of Expressed Sequence Tags (ESTs). Putative human exons together with those perfectly matching EST were assembled with the SeqMan program resulting in new versions of the putative transcripts extended in both directions. Some of the mouse exons previously marked as insecure were effectively dismissed in this clustering step. With this extended but still incomplete human putative mRNAs, DotPlot was performed, this time against the mouse genomic sequence in BAC221D7. The result from this analysis confirmed the exons in the mouse that were previous correctly predicted by the Rummage annotation system and in some cases succeeded finding previously missed exons. The putative mouse mRNA at the end of one round was a filtered and extended version of the predicted starting one and it was used to begin the whole process again until no longer extension occurred. Figure 3.8 shows a schema of this method.

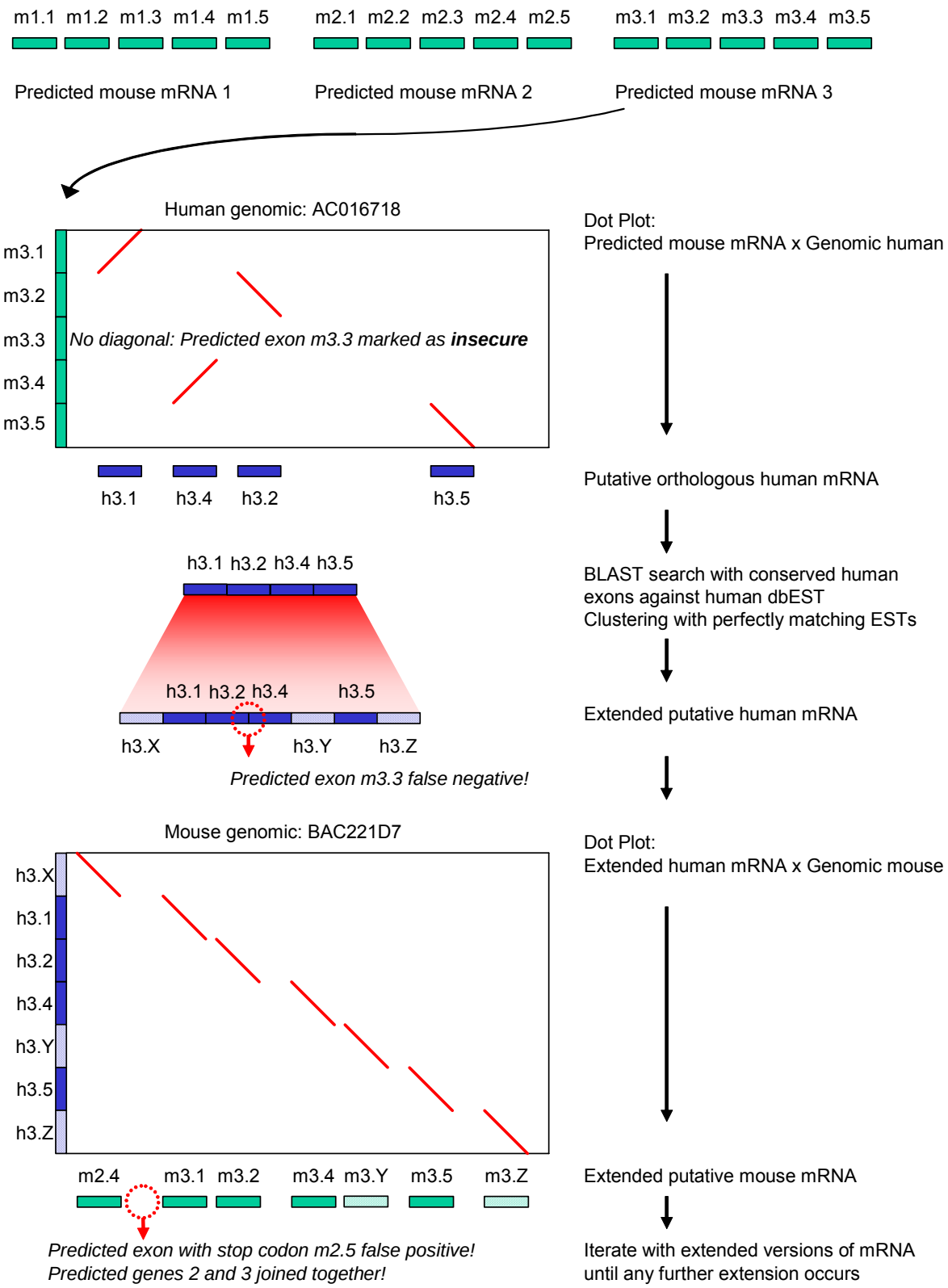


Figure 3.8: Idealisation of the method used to discover *STK33*.
Boxes in green represent exons from the mouse, in blue from human. Dashed boxes represent new incorporating exons to the prediction.

After three iterations, two transcripts predicted from Rummage were fused into one gene-prediction conserved in human and mouse, consisting of 12 coding exons, putative start codon, putative stop codon, putative poly-adenylation signals and the coding sequence for a serine/threonine kinase. The novel genes were named *STK33* for the human and *Stk33* for the mouse according to the rules of the International Nomenclature Committee of the Human Genome Organisation and a first description was published in 2001 (Mujica et al. 2001). The whole genomic structure has been completed by additional EST analysis and comparison with the finished human sequence, particularly in the 5' untranslated Region for both species, where 5 new very poorly conserved non-coding exons have been found in *STK33/Stk33* genes.

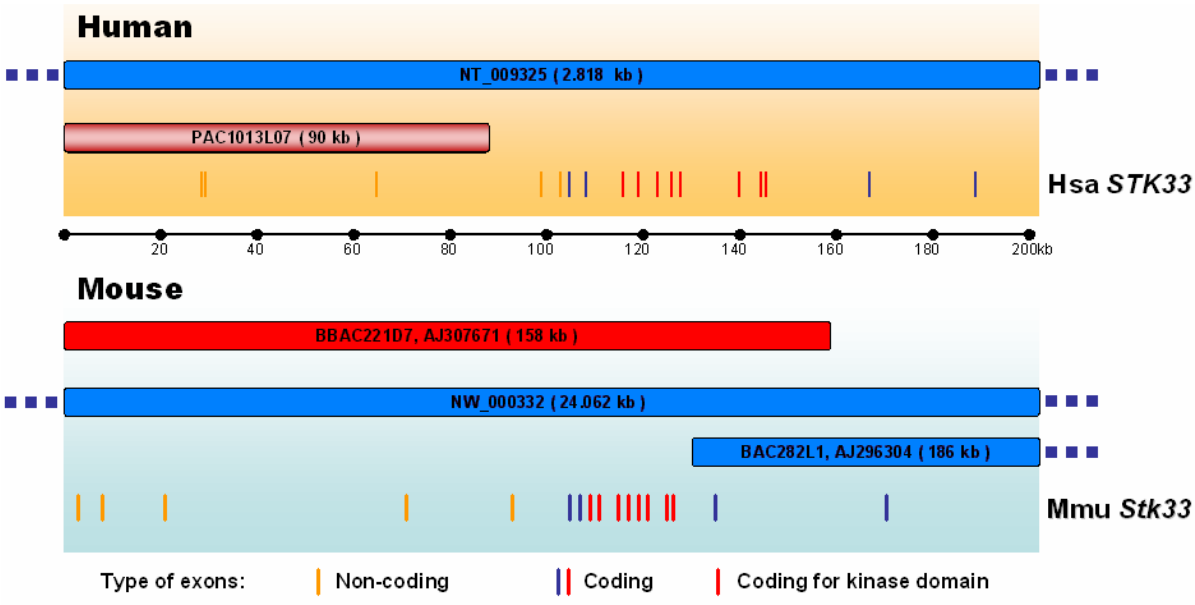


Figure 3.9: Genomic structures of Mouse *Stk33* and Human *STK33*.
Positions of *Stk33* and *STK33* exons relative to the genomic clones sequenced in this work (red boxes) and other sequences from the public database with their respective accession numbers (blue boxes). Upper panel refers to the human, lower to the mouse. Scale in kilobases is shown in the middle. Human PAC1013L07 clone was not fully sequenced (see text). Mouse BAC282L1 clone (Acc. AJ296304) was sequenced by T. Brueckmann (Cichutek, 2001). Exons are represented with vertical lines, non coding exons are with light-brown colour, coding exons with blue and red, exons coding in particular for the cannonical kinase domain are represented with red. This figure represents the situation at the moment of presenting this work. To the date of discovery of *Stk33* and *STK33*, there were no mouse genomic sequence of this region in the databases, and the corresponding fragment in the human was not in face 3, i.e. still plenty of gaps and contigs without order (See text and Mujica et al., 2001).

3.2.2 Human mouse comparison

The analysis described in this and following sections was performed with 205,255 bases in the human and homologous segment of 180,000 bases in the mouse. The latter containing the whole BAC221D7 sequence. Both sequences include the complete coding region of the novel kinase gene. The sequence starts with the downstream intergenic region from genes *C11orf29* and *D7H11orf29* (Amid 2002) and ends 5,359 bases downstream from *STK33* and 6237 bases downstream *Stk33*. Genomic sequences obtained in our labs and the equivalent ones available in the databases were aligned using the SeqMan program (Lasergene's DNAStar). A contiguous consensus sequence was obtained from the mouse; whereas the corresponding human genomic sequence still has a gap (~5 Kb estimated size) located between exons 2 and 3 from STK33.

In order to make a first assessment of sequence conservation of *STK33/Stk33* at the genomic level, sequences from mouse and human were compared by DotPlot. In this analysis, two sequences in the X and Y axis are compared using sliding windows of a given size. When within a window the two sequences display identity, a dot is plotted in the corresponding position. Thus, two identical sequences would yield a perfect diagonal row of dots. The DotPlot in the following figure shows clear diagonals around exons, in particular coding ones. Intronic regions show a lower degree of conservation with the remarkable exception of the region 46-64 Kb in the human 44-66 Kb in the mouse. This region includes exon 3 from human *STK33*, several repeats from diverse origin and a highly conserved (A+T)-rich region.

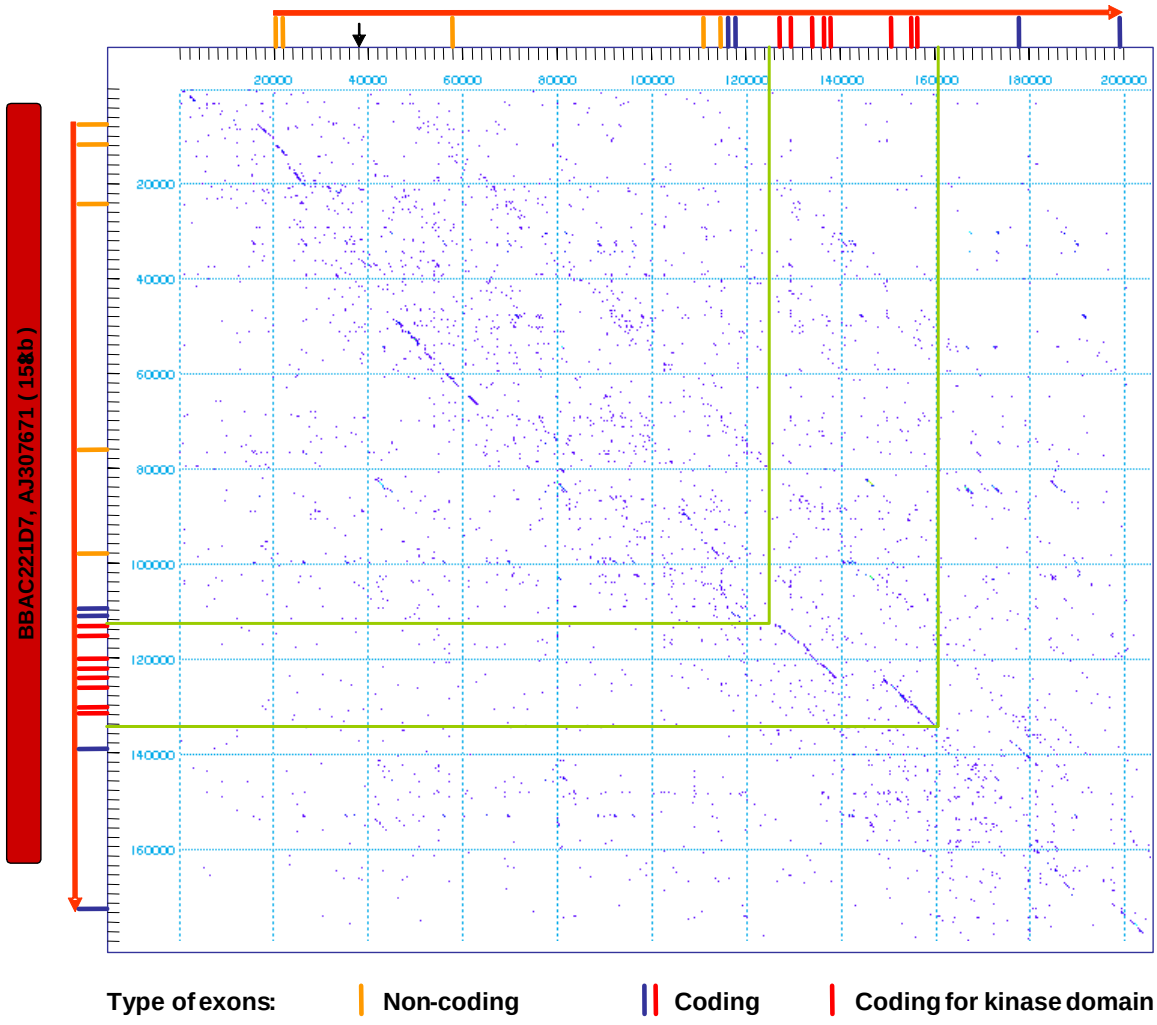


Figure 3.10: DotPlot comparison of the human (horizontal) and mouse (vertical) sequences.
The span of the *STK33/Stk33* genes is represented with red arrows parallel to their respective axis. Arrow orientation represents direction of transcription. Relative positions of the exons are represented by short lines and their coding nature is depicted by colours according to the legend. The relative position of the murine BAC clone sequenced in this work is shown parallel to the Y axis. A black vertical arrow shows the position of a gap still open in the human genome sequence. The analysis was performed with the MEGALIGN program (Lasergene) with the following parameters: 60% Similarity; 50 bases-long Windows; 50 Minimum numbers of windows.

As shown in the figure, *STK33/Stk33* DotPlot analysis shows a much clearer conservation around coding exons respect the non-coding exons here slightly shadowed in blue. Noteworthy is the interruption of the diagonal in the region highlighted with blue,

which may correspond to an indel (insertion in the human or a deletion in the mouse) in the corresponding intron sequence.

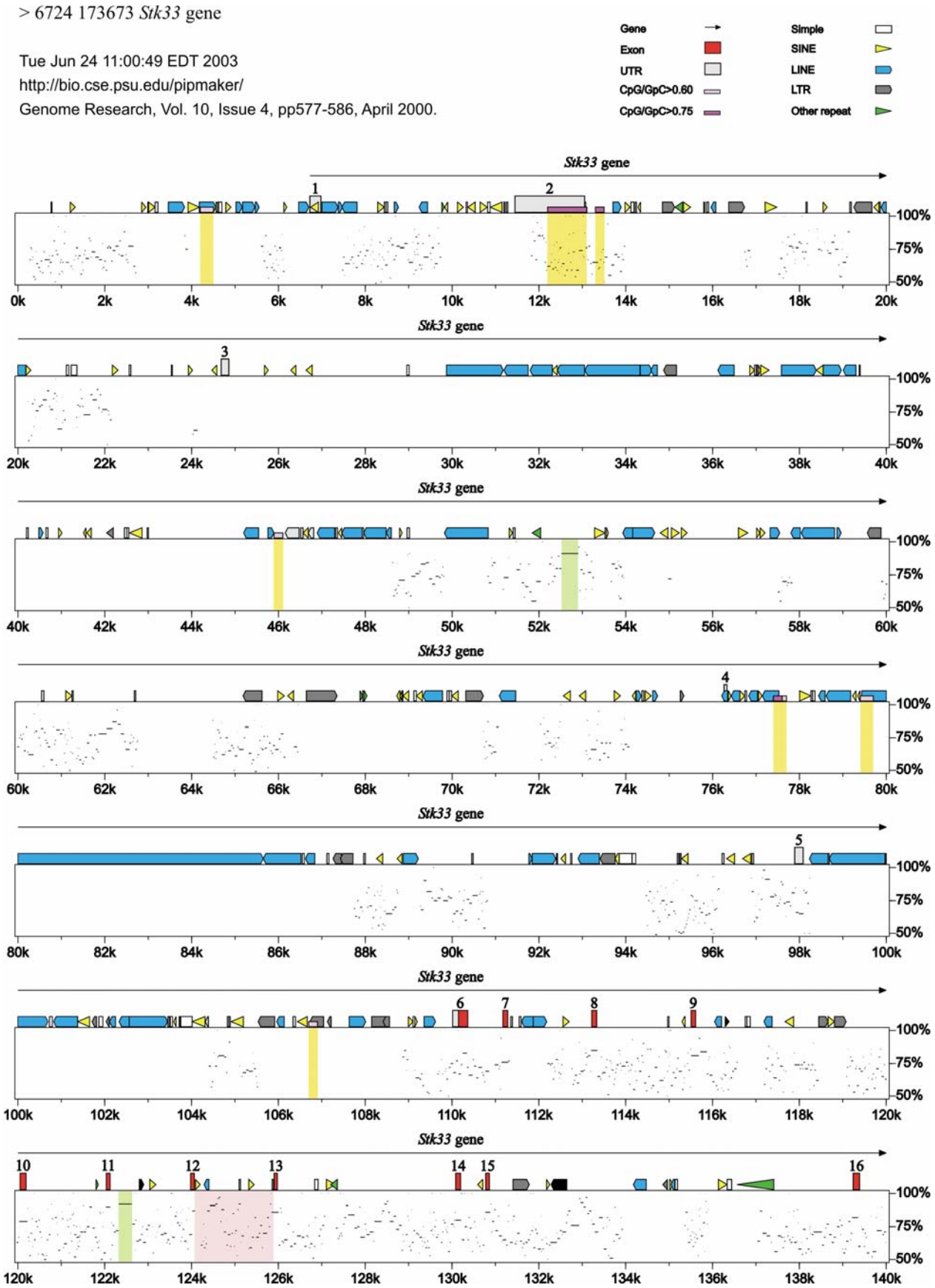
3.2.3 Percentage Identity Plot and Vista Genomic view

A more detailed view of interspecies sequence conservation, including similarity value, repeats and putative CpG islands, is provided by tools of genomic sequence analysis based on interspecies comparison, such as the Percentage Identity Plot (PipMaker), (Schwartz et al. 2000) and Vista Genomic view methods (Mayor et al. 2000). PIP is a superior tool to detect highly conserved regions in the genomes of the two species compared. The graphic result of the PIP analysis for the region under study is shown in the following pages.

The PIP program uses Repeat Masker results to display nature and orientation of the repeat sequences. As an example, the large indel observed by DotPlot between exons 12 and 13 of *STK33/Stk33* (figure 3.10), gets nicely explained by looking at the PIP analysis which follows in the next pages. In the corresponding region, starting roughly at the 140 Kb of the analysed human sequence, a very large set of LINE-type of repeats covering up to 8 Kb is observed. The homologous region in the mouse, between 124k and 126k shows a drastically lower repeat content. In this case is clear that the difference in size of this precise intron in both species lies on the insertion of line repeats in the primate linage.


> 21321 178271 STK33 gene
Tue Jan 28 20:57:01 EST 2003
Http://bio.cse.psu.edu/pipmaker/
Genome Research, Vol. 10, Issue 4, pp577-586, April 2000.
Gene
Exon
UTR
CpG/GpC>0.60
CpG/GpC>0.75
Simple
SINE
LINE
LTR
Other repeat
100%
75%
50%
0k 2k 4k 6k 8k 10k 12k 14k 16k 18k 20k
STK33 gene
1 2
20k 22k 24k 26k 28k 30k 32k 34k 36k 38k 40k
STK33 gene
3
40k 42k 44k 46k 48k 50k 52k 54k 56k 58k 60k
STK33 gene
60k 62k 64k 66k 68k 70k 72k 74k 76k 78k 80k
STK33 gene
80k 82k 84k 86k 88k 90k 92k 94k 96k 98k 100k
STK33 gene
4
5
6
7
100k 102k 104k 106k 108k 110k 112k 114k 116k 118k 120k
STK33 gene
8
9
10
11
12
120k 122k 124k 126k 128k 130k 132k 134k 136k 138k 140k

Figure 3.11: Percentage Identity Plot (PIP) of the *STK33* human genomic sequence.
(Starting in previous page). Vertical axis plots the similarities above 50% with the corresponding homologous region in the mouse, produced by BlastZ analysis. Dispersed dots and lines at different similarity values are equivalent to diagonals in a DotPlot analysis. Features along the human sequence are shown over the X-axis, according the legend on top of the figure. Relative positions and coding nature of the exons are here shown (from exon 1 to 17). Light yellow rectangles under the X axis show putative CpG islands, light green rectangles show conserved very high (A+T) regions and the large light purple rectangle shows the insertion of LINE-type of repeats between exons 12 and 13.


> 6724 173673 Stk33 gene
Tue Jun 24 11:00:49 EDT 2003
http://bio.cse.psu.edu/pipmaker/
Genome Research, Vol. 10, Issue 4, pp577-586, April 2000.
Gene
Exon
UTR
CpG/GpC>0.60
CpG/GpC>0.75
Simple
SINE
LINE
LTR
Other repeat
Stk33 gene
100%
75%
50%
0k 2k 4k 6k 8k 10k 12k 14k 16k 18k 20k
20k 22k 24k 26k 28k 30k 32k 34k 36k 38k 40k
40k 42k 44k 46k 48k 50k 52k 54k 56k 58k 60k
60k 62k 64k 66k 68k 70k 72k 74k 76k 78k 80k
80k 82k 84k 86k 88k 90k 92k 94k 96k 98k 100k
100k 102k 104k 106k 108k 110k 112k 114k 116k 118k 120k
120k 122k 124k 126k 128k 130k 132k 134k 136k 138k 140k

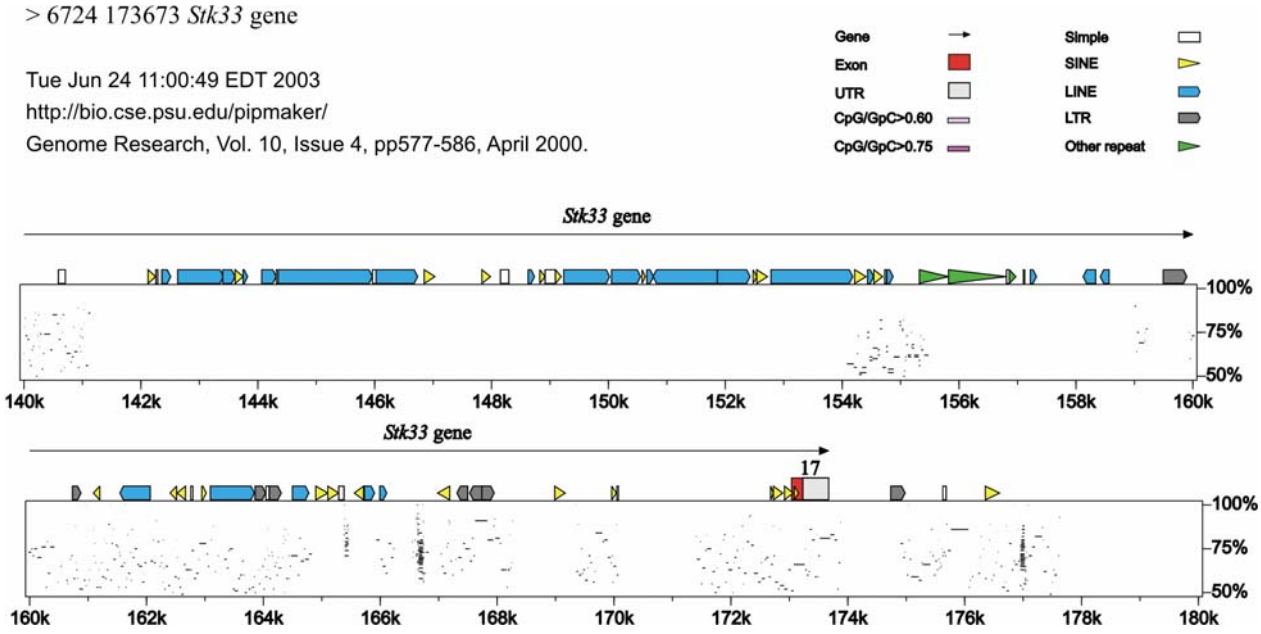


Figure 3.12: Percentage Identity Plot (PIP) of the *Stk33* mouse genomic sequence.
(Starting in previous page). Vertical axis plots the similarities above 50% with the corresponding homologous region in the human, produced by BlastZ analysis. Dispersed dots and lines at different similarity values are equivalent to diagonals in a DotPlot analysis. Features along the mouse sequence are shown over the X-axis, according the legend on top of the figure. Relative positions and coding nature of the exons are here shown (from exon 1 to 17). Light yellow rectangles under the X axis show putative CpG islands, light green rectangles show conserved very high (A+T) regions and the large light purple rectangle shows the much shorter intron between exons 12 and 13.

The PIP analysis produces the starting points for searching evolutionary conserved regulatory regions or any further segment with potential biological meaning. For example, very highly conserved sequences in the genomic region under study, which correspond to neither coding sequence nor repeats, are of special interest (positions 49,553-50,049 and 137,132-137,617 in human and 52,386-52,921 and 122,217-122,710 in the mouse, see next figure for an alignment). These regions expand some few hundreds of bases and exhibit a remarkably high (A+T) content close to 70% but they were not recognised as low complexity repeats by RepeatMasker. The possible putative function of these conserved regions is up to now unknown.

Figure 3.13: Local alignment of (A+T)-rich regions detected with PIP and VISTA analysis.
First lines of each pair of alignments correspond to the human sequence.

Another useful tool for genome-wide analysis is the VISTA program. Instead of drawing diagonals of local similarity like in the PIP, VISTA shows a histogram of global similarity and knowing the position of the exons from the user's input, it remarks especially the positions of Conserved Non-coding Sequences (CNS). In the next page an output from the VISTA is shown, where the genomic sequence of *STK33* is compared with the homologous region in the mouse.

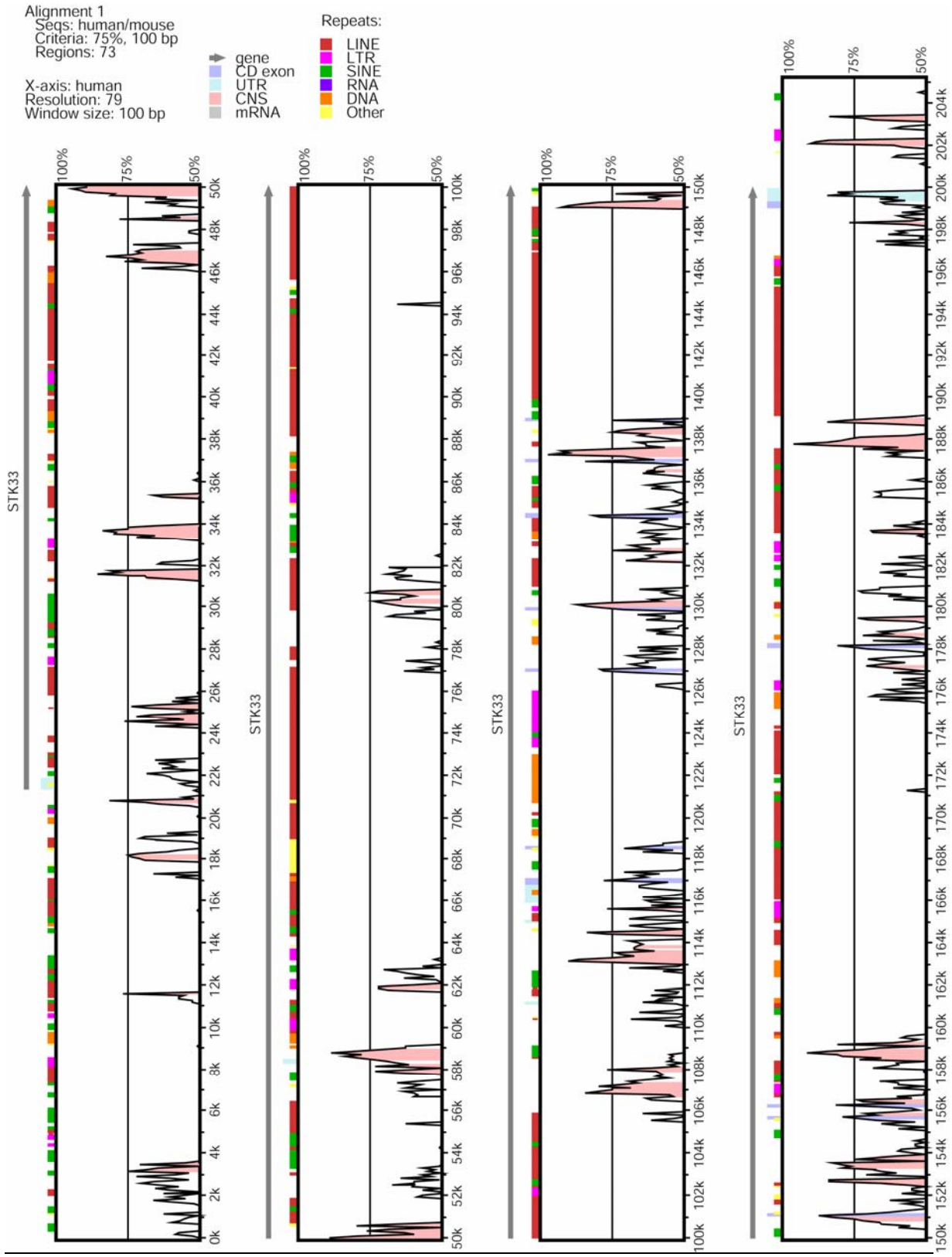


Figure 3.14: VISTA view of the genomic region of human *STK33* compared with the homologous region in the mouse.

On the y-axis the sequence identity is given in percentage.

3.2.4 (G+C) content and CpG islands

(G+C) content was obtained using the programs "windows" and "statplot" from the HUSAR- Interface and the programs GeneQuest and MapDraw from Laserge's DNAStar. The analysed human genomic region is 38.56% (G+C) and 40.89% the murine region. In the graphic display of the (G+C) distribution and in the PIP results, there are clear short regions of very high (>60%) content of (C+G), so called CpG islands, the first two of them associated with the 5'-region from exon 1 of *STK33/Stk33* genes and conforming exon 2 (Figures 3.11 and 3.12). The rest of the detected CpG islands, are clearly associated with repeats.

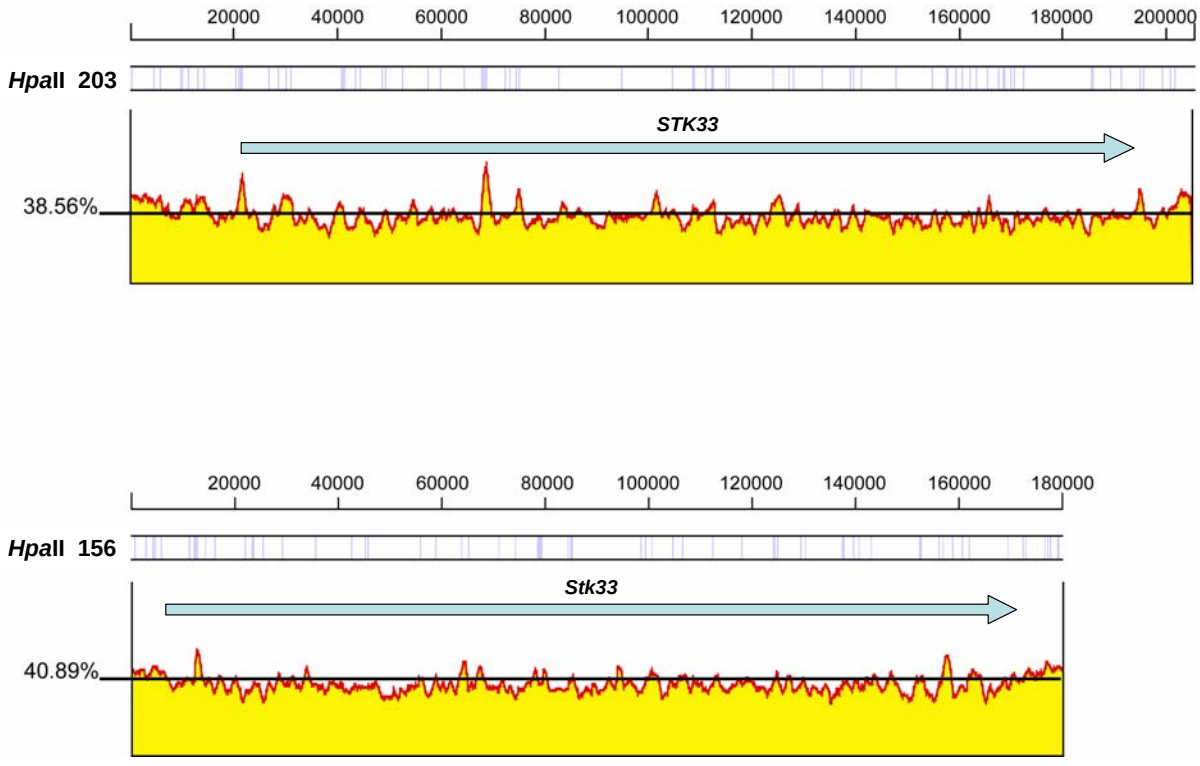


Figure 3.15: (G+C)-Plot analysis
Genomic region around human *STK33* (top) and mouse *Stk33* (bottom). *Hpa* II restriction map was obtained using the program MAPDRAW (Laserge's DNAStar). *Hpa* II restriction sites (CCGG) usually correlate with CpG-islands. (G+C) Plots were obtained using the program GeneQuest (Laserge's DNAStar) with a window size of 1000 bases. Average (G+C) content was obtained using the option "DNA statistics" from the program SeqMan (Laserge's DNAStar).

3.2.5 Repeat content

The genomic regions around human *STK33* and mouse *Stk33* were analysed with the RepeatMasker program (unpublished, see repeatmasker.genome.washington.edu) using the sensitive parameters. General results are shown in the table below.

Table 3.3: Repeats content of the genomic region around *STK33/Stk33*

	Human			Mouse		
	number of elements	**length (bp)**	**% of seq.**	**number of elements**	**length (bp)**	**% of seq.**
ALUs	68	18,348	8.94	-	-	-
B1s	-	-	-	36	4,123	2.29
B2-B4	-	-	-	62	10,136	5.63
IDs	-	-	-	3	206	0.11
MIRs	17	2,823	1.38	8	1,048	0.58
Total SINEs	**85**	**21,171**	**10.31**	**109**	**15,513**	**8.62**
LINE1	52	68,316	33.28	60	41,364	22.98
LINE2	23	7,014	3.42	9	1,655	0.92
L3/CR1	7	1,766	0.86	0	0	0.00
Total LINEs	**82**	**77,096**	**37.56**	**69**	**43,019**	**23.90**
MaLRs	16	7,353	3.58	23	6,463	3.59
ERVL	3	1,283	0.63	2	289	0.16
ERV_classI	2	1,207	0.59	2	458	0.25
ERV_classII	0	0	0.00	4	1,302	0.72
Total LTR elements	**21**	**9,843**	**4.80**	**36**	**10,418**	**5.79**
MER1_type	14	3,666	1.79	2	300	0.17
MER2_type	6	5,229	2.55	3	1,165	0.65
DNA elements	**22**	**9,131**	**4.34**	**5**	**1,465**	**0.81**
Unclassified	**1**	**1,522**	**0.74**	**3**	**1,644**	**0.91**
Total interspersed repeats		**118,763**	**57.86**		**72,059**	**40.03**
Simple repeats	**22**	**1,065**	**0.52**	**73**	**4,059**	**2.25**
Low complexity	**18**	**624**	**0.30**	**23**	**1057**	**0.59**

Neither small RNA nor satellites were found

As already observed for the whole genomes (Lander et al. 2001; Waterston et al. 2002), the human sequence contains a larger fraction of detectable repeats than the homologous one in the mouse. Noteworthy, the human genomic sequence around *STK33*

contains a 10% higher fraction (57.86% vs. 46.36%) of total interspersed repeats than the whole human genome.

3.2.6 Evidence of a transposition event of prokaryotic origin in BAC221D7

Already with partial contigs from the sequencing of murine clone BAC221D7, a first candidate match (100% identity) from the BLAST comparison to public databases was detected, showing strikingly "conservation" in a broad variety of species, such as human, plants and unicellular organisms. But a detailed look showed that this 1,329 bp long sequence was one of prokaryote origin: IS*10*, the right flanking insertion element sequence from Tn*10*, one of the classical examples of transposable elements. IS*10*-Right alone codes for a transposase and hence has transposition activity independent from Tn*10*. Eukaryotes and prokaryotes have very different transposable elements and Tn*10* clearly does not belong to those naturally to be found in a rodent's genome. A first evidence of transposition of this element is the presence of the characteristic 9-bases long direct repeat string at both ends of IS*10*, namely TGTCTAGCA. To confirm the absence of this element in the genomic sequence of the mouse, primers were designed flanking the IS*10* element in BAC221D7 and parallel PCR reactions using the BAC clone and mouse genome as template were tested. As expected, the PCR product from the BAC221D7 was ~1,338 (1,329 + 9) bp longer than the one from the mouse genome. Direct sequencing from both ends of the PCR product from the mouse genome aligned with each other and showed the string TGTCTAGCA just once. Hence, the presence of IS10 in the murine BAC221D7 is proved to be an artifact surely by the BAC library construction. Comparison of these sequences with the one of BAC221D7 showed that no genomic sequence from the mouse was lost by this insertion, since DNA

loss, though in one out of 1,000 cases, is one of the scenarios of insertion of IS*10* (Kleckner N. in (Berg and Howe 1989). The presence of this artifact-sequence in BAC221D7 between bases 145,650 and 146,978 was fully annotated in its database entry AJ307671.

3.3 The novel kinase gene *STK33*

3.3.1 Gene structure

Human *STK33* expands over more than 178,574 bases at the genomic level, whereas the murine version of the gene covers exactly 166,949 bases. The transcribed length of mouse *Stk33* is longer than human *STK33* principally due to differences in the 5' UTR: mouse *Stk33* exon 2 is remarkably long. The putative protein product has 514 aa in human and 491 aa in the mouse. Table 3.4 shows some of these general features.

Table 3.4: Sizes of *STK3* and *Stk33* genes and protein product

Lengths / Genes		Genomic (bp.)	Transcript (bp.)	Mean exon (bp.)	Mean coding. Exon (bp.)	Protein (aa.)
Hsa.	*STK33*	>178,574	3,779	222	128	514
Mmu.	*Stk33*	166,949	4,303	253	123	491

The 9% difference in size of human and mouse versions of *STK33* is in line with the general observation that the mouse genome is 14% smaller than the human genome (Waterston et al. 2002). Human *STK33* and mouse *Stk33* consists of 17 exons. Even though, the number of exons is conserved, there is a remarkable 5' UTR divergence between the two species and also variability in exon size, splicing pattern, transcription initiation and poly-adenylation signals. In the following pages, tables 3.5-6 and figures 3.16-17 depict the

exon/intron structure of human *STK33* and murine *Stk33*. The full length transcript is 3,085 bases long, with a remarkably high (A+T) content of 62%.

Table 3.5: Exon-intron structure of human *STK33* gene

Exon No.	Exon size (bp)	Intron / Exon \ Intron	Intron size (Kb)
1*	146	tgcacccgccgcc/**G**CTGTTTG...CGGCCGCG**gt**gagtggtccct	0.1
2*	233	acgcggtggaaaaac/GGGCCTCTG...TACTGTGGG**gt**aagtgtgtgggaa	>36.4
3	149	acctctttcttct**ag**/GACCTTCCT...CTGCCTAAG**gt**aacaaacatgtac	52.7
4	81	ctttttctttttct**ag**/GCAGGGCCA...CACTCCAGG**gt**aaattccaataat	3.8
5	65	gttcaataaaagt**ag**/GAGAGGACT...CAAAGCAAG**gt**aaggggggagagg	0.9
6a*	1144	cataaaaccataata/**A**TGTACTCC...AAAGATTTG**gt**aaatatcaaagat	1.4
6b	386	cttttctttctgc**ag**/CTCTCACGT...AAAGATTTG**gt**aaatatcaaagat	-
6c	136	ttgtgtatgttcc**ag**/CAAAACAAG...AAAGATTTG**gt**aaatatcaaagat	-
7	114	ttttttggtccc**ag**/CCCTCAAGA...GCTATTGAG**gt**atatagaaagcaa	8.3
8	114	tttttttttctcc**ag**/GAAATCTAT...AAAGAAAAG**gt**aaggctcattagc	2.8
9	105	tcttcattgatta**ag**/GCTGGAAGC...ACGCCAAAG**gt**aaacctctattag[d]	4.3
10	139	tatatataatgtc**ag**/AAAATGTAC...ACAATAATG**gt**aagaagagggtta	2.5
11	89	ttattttcttggt**ag**/ATATTGTAC...AACATAAAG**gt**aagattagcaatg	1.8
12	85	aatttcttctccc**ag**/GTGACTGAT...TCTATATGG**gt**aagttagtagact	10.2
13	76	ctgcttgtttggc**ag**/CCCCTGAAG...GTACATGTT**gt**aagtagcctgcta	4.5
14	113	ttgcctttgttgc**ag**/ATTACGTGG...GTGACTGTG**gt**aagtatagatgcg	0.5
15	86	ttttcttatttttt**ag**/CTAAAAGTG...TGGTTAACA**gt**aagttacaatatt	22.1
16	198	gattgttaatatt**ag**/GGCAATAAA...GAAAAACAG**gt**aggaagaatcatt	20.8
17 cod	201	acctttttatttgc**ag**/TCTACTGCT...AAACTCTAAGGTTCCCTCCAGTGTT	Stop
17 utr 1	467	ACT**AATAAA**CTTGCCATACGTATTACAGCA(**AAAAAAA**...)	polyA
17 utr 2	641	AAATC**AATAAA**AAACAGATGTTACTCAGCA(**AAAAAAA**...)	polyA

Traces of intronic sequences are shown in lowercase letters, traces of exonic sequences are shown in uppercase letters. Intron sizes and sequences were deduced from the alignments of mRNA/cDNA sequences with the corresponding supercontigs from NCBI (Accession No. NT_009325) and from Celera (Contig c11_8357730-8596998, publication.celera.com/). Conserved 3' splice acceptors (**ag**/) are shown in **bold**. Conserved 5' splice donors (**gt*) are shown in **bold**. Marked with asterisk (*) the proposed 3 transcription initiation sites are shown by the initial bold letter in exons 1, 2, 6a. EST analysis suggests that exon 6 may be present in three forms, whole (6a) and shortened through internal 3' splice acceptors (6b, 6c). The stop codon is shown here with bold TAA at the end of the coding region of exon 17. EST data also suggests that Human *STK33* have two different poly-adenylation signals indicated here by bold AATAAA in two different untranslated regions of exon 17. The position of the polyA tails (represented here in parenthesis) are shown at the end of the exon 17. See also Fig 3.16 below

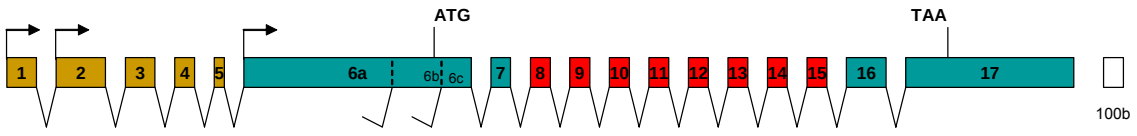


Figure 3.16: Basic exon-intron structure of human *STK33* gen.
Exons are represented with boxes, light brown the 5' UTR exons, in light blue and red the coding exons, and particularly in red those coding for the kinase domain. Exons are drawn in scale as shown at the right. Introns are extracted and not drawn in scale. The position of the start stop codons is shown as first described (Mujica et al. 2001). Arrows from exons 1, 2 and 6, show putative transcription starting points. Exon 6 shows two active alternative 3' splice acceptors.

Table 3.6: Exon-intron structure of mouse *Stk33* gene

Exon No.	Exon size (bp)	Intron / Exon \ Intron	Intron size (Kb)
1*	245	aaggcaggaggcagg/**A**CTCTTTTG...TGGGAAAAT**gt**tccagcctaatgg	4.5
2*	1601	gggaggactcacc**ag**/CATGCCTTT...TCACTCAGG**gt**aagtgtgtgtgtg	-
2a	1263	gggaggactcacc**ag**/CATGCCTTT...GCCGTTGTG**gt**gagtggcctcgca	0.3
2b	37	ccatgttcgctac**ag**/GAAACATTT...TCACTCAGG**gt**aagtgtgtgtgtg	11.6
3	175	gaattattgtctc**ag**/TTGTAGGCT...AGGCATTAG**gt**aagtattgatggg	51.4
4	66	ttatctattcatc**ag**/CTGGCAGTC...CATTCCTGT**gt**aagtatatatgtg	21.6
5	194	aatctattgtttcag/**A**CACAGACCC...GATGTCATT**gt**gggtgacactgtg	11.9
6	315	cttttctgcccac**ag**/CTCCCATGT...AAAGATTTG**gt**aaatattgaaaat	0.8
7	111	atcttttggccct**ag**/TCCATAAGA...GGTATCGAG**gt**atgcaggaagcaa	1.9
8	114	tttgtttgtttcc**ag**/GAATTCTAT...AAAGAAAAG**gt**aagactttcccgt	2.2
9	105	ttgttcatcgatt**ag**/GCTGGAAGT...TCGCCTCAG**gt**aaagttccatcag	4.4
10	139	aacctgtaatgtc**ag**/AAAATGTAT...ATAACAAGG**gt**aagacaagaactg	1.8
11	89	ttttttttcttggt**ag**/ATATAGTGC...AACATAAAG**gt**aagagcctagaga	1.9
12	88	tgtgtctgttctt**ag**/GTGACTGAT...TCTATATGG**gt**aagcgaggagaac	1.8
13	76	ctgcttgtttggc**ag**/CACCAGAGG...GTTCATTTT**gt**aagtagcttcctg	4.1
14	113	actgcttttgttt**ag**/ACTGTGTGG...GTGATTCTG**gt**gagtatggactta	0.6
15	86	ccccttttatttgc**ag**/CAAAAAATA...TGGTTGACA**gt**aagttatgacact	8.4
16	150	tgtttttaatctt**ag**/GGCAATACC...GCAAAGCAG**gt**aggagggatggct	32.3
17 cod	192	ttctttaattcat**ag**/CCCACCAAT...AGGCTC**TAA**GGTTCTGTCCAGTGCT	Stop
17 utr	444	TTACT**AATAA**AATGGCCACAGGCATTGGCAGAAGCG(**AAAAAAA...**)	polyA

Traces of intronic sequences are shown in lowercase letters; exonic sequences are shown in uppercase letters. Intron sizes and sequences were deduced from the alignments of mRNA/cDNA sequences with the corresponding BAC clones 221D7 (Mujica, AJ307671), 282L1 (Brueckmann, AJ296304) and the supercontig from NCBI (Accession No. NW_000332). Conserved 3‘ splice acceptors (**ag**/) are shown in **bold**. Conserved 5‘ splice donors (**gt**) are shown in **bold**. Marked with asterisk (*) the proposed 2 transcription initiation sites are shown by the initial bold letter in exons 1 and 2. EST analysis suggests that an internal fragment of exon 2 is spliced out in some transcripts yielding two shorter exons, here named 2a and 2b. The stop codon is shown here with bold TAA at the end of the coding region of exon 17. The unique poly-adenilation signal detected in mouse *Stk33* is indicated here by bold AATAAA in the untranslated region of exon 17. The position of the polyA tail (represented here in parenthesis) is shown at the end of the exon 17. See also Fig 3.17 below

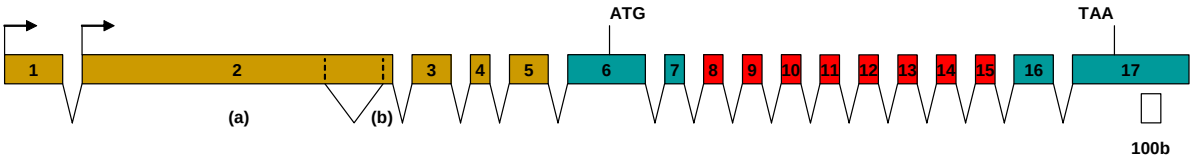


Figure 3.17: Basic exon-intron structure of mouse *Stk33* gen.
Exons are represented with boxes, light brown the 5’ UTR exons, in light blue and red the coding exons, and particularly in red those coding for the kinase domain. Exons are drawn in scale as shown at the right. Introns are extracted and not drawn in scale. Position of the start and stop codons is shown as first described (Mujica et al. 2001). Arrows from exons 1 and 2 show putative transcription starting points. Exon 2 has two intern 3' splice acceptor and 5' splice donor signals that may also be interpreted as a case of intron retention.

3.3.2 Expression analysis

By the date of its characterisation mid 2000, expression of *STK33* was only partially documented by a small number of EST entries in the databases from the following human tissues: testis (twice); uterus; lung; pooled: fetal lung, testis, B-cell (4 times); uterus leiomyosarcoma; uterus endometrium adenocarcinoma; colon tumor metastasis; carcinoid lung (5 times); pooled germ cell tumors and cervix carcinoma cell line. There were only three mouse entries and two from the pig (see figure 3.18). Several of the human EST entries were obtained in the course of the Cancer Genomic Anatomy Project (CGAP), an effort whereby ESTs are massively produced from malignant tissues extracted from human patients and model organisms.

The presence of this few perfectly matching EST entries suggested that *STK33* is an active gene. To study the expression of *STK33* experimentally, commercially available human uterus total RNA (Research Genetics) was used for RT-PCR experiments. *STK33*-specific double stranded cDNA was synthesised by Polymerase Chain Reaction with specific primers. Results are shown in the figure below.

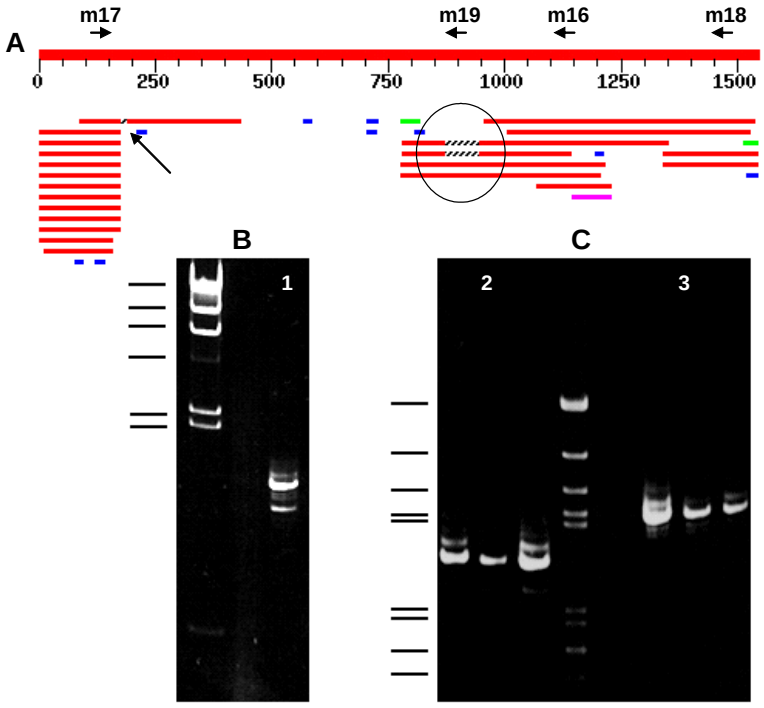
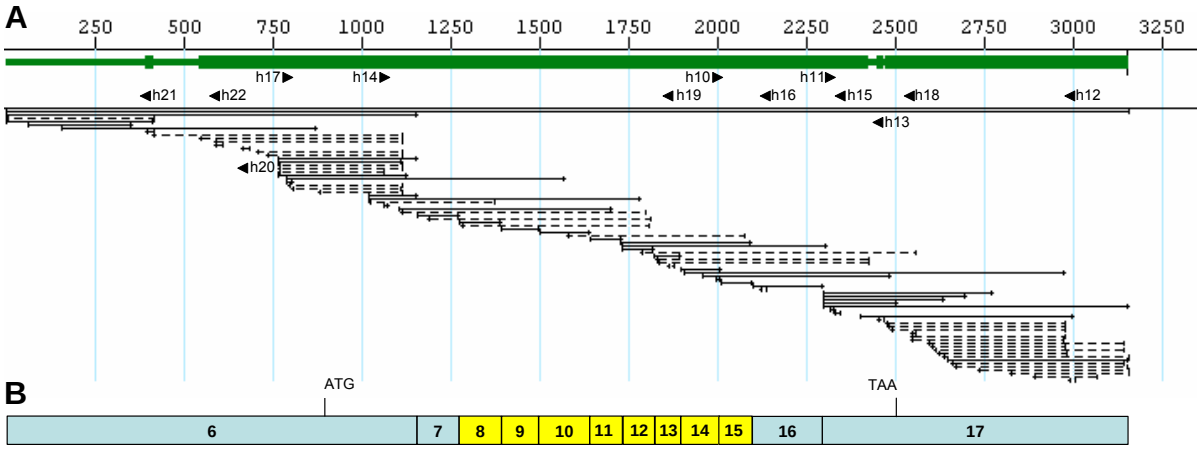


Figure 3.18: Expression analysis of *STK33*
A: Graphic output from BLAST search with putative *STK33* against all ESTs. The upper axis with coordinates represents the query sequence and its length in base pairs. The 23 significant results (see text) are represented graphically with red and purple bars. Note that matches do not cover the whole putative gene region. This may explain partially why *STK33* remained undiscovered up to that date. A diagonal arrow show the position of an adenine-rich string which may explain the sharp coincidence of the right tips from the ESTs in this area: the site served as false priming site for the poly-T primer employed to reverse amplify from mRNA. The circle shows a gap in the alignment of two entries, the first evidence of alternative splicing in *STK33*. Arrows over the axis represent the relative position of primers used for the amplification of the cDNAs in our lab. **B:** First experimental evidence of functional *STK33*. Reverse amplification from human uterus RNA with poly-T primer, PCR amplification with specific primers h17 and h18. The DNA-fragments were cloned and sequenced confirming the suspect of alternative spliced variants. All *STK33* putative exons were fully confirmed. **C:** Further tests with different reverse primers (h16 and h19) and several PCR conditions to discard unspecific amplifications due in-vitro artifacts. In all cases the cloned bands were confirmed to be alternative versions of *STK33*.

The confirmed full length sequence of the principal transcript of Human *STK33* was obtained by aligning the sequences from the databases, the expanded sequences of EST-clones obtained in the lab and the cDNA amplifications from human uterus. Rapid amplification from cDNA ends (RACE) was used to define the starting point of transcription of *STK33* in human uterus. As a result, the putative mRNA sequence from *STK33* was extended to the first base of exon number 6. A typical TATA box motif was found at position –20 in the human genomic sequence. Figure 3.19 shows the alignment of the cDNA

from human *STK33* from uterus. Figure 3.20 shows the whole mRNA sequence and the sequence of its inferred protein product.



C

AA331945	AI703417*	BE326215	AW207019	AA886440	BE396172
AI243518	AI910878*	AX056400	AW118625	AA626278	AW273934
BF679548	AW341417	AA297313	AW959747	BE885808	AA454060
AI681535	AI267885*	AI818733*	AA453640	AW268879	AI741915
AI198872	AW341578	AI818734	AI677729	AI829855	BE219074
AW129267*	AI377304			+10 cDNA reads	

Figure 3.19: Alignment of cDNA from human *STK33* gene.
A: Graphic representation of the alignment. Continuous arrows represent sequences in the Watson strand; segmented arrows represent sequences in the complementary Crick strand. Short black triangles signal the positions and orientation of the primers employed to amplify and sequence the DNA. Sequences were aligned with the program SeqManII (Lasergene's DNAStar). **B:** Relative positions of the exons represented in the transcript in scale with the upper alignment. Positions of the ATG start codon and the TAA stop show the largest ORF. Yellow boxes represent exons coding for the protein kinase domain. **C:** Sequences supporting this alignment. A list with the GenBank Accession numbers of publicly available EST is shown. Those marked with asterisk were additionally ordered at RZPD and sequenced from both sides expanding the length originally given in the databases. Additionally, 10 reads from RT-PCR sequences obtained from human uterus were necessary to close the full length sequence and resolve single-pass sequencing errors in the ESTs.

```
   1 ATGTACTCCCAATTACTTCTGGAAGTTTCTCAAAGTACTCCTTTATATATACTGCAGAGTGTATTTTTCTTCCTCCTCAACTGAGATCTTTCCAACTTGC
 101 CACCATGCAGCTGCCAATGGTCCTAGTTAAGTAAAATGCTGCCATACCTATTTTAGACTCAGGGAAAAATAGCACCCACTCATTTTTATTTTTGCTCAAT
 201 ATAAAAATGAGGATACTTATGAGGATACTTAAACTTTTAGGATTAGCTAGTTTTCTAAAAATCGAATTATTCACTCCTTTGTAAAGTATGTAATAGGAAT
 301 TTGCTCTAATAATCAATAGATTAAGGTTTAAAATTTGAAACCATAGTAATGTATGTTTAACACCAATATTTTAAGCCTTTTTAAAAACCACAACCCACAT
 401 TAAGAAATACATTTCATACTGTGATCAAGTACACACGCACACACACACTCTATACATATATGTCTGTCCAATTAAAAGTTTCACAGAAATTTCCAAGGAG
 501 GTATGCTAAATATTATCTCTTTGATTCTACTTTATTTTTAAAAAGTGGTATCAACCCACAAAATGGATTTCATAACCCACTACGCAGTTTGATAAGATGC
 601 TGTTTTAGACCATGCTTTTCACCAGTTTTGTGGTCCTATTTTGTCCTTTTCATGTCTATACAGGATGCTTCTAGTGCTAGTTGCTAGCTTTTCTCTGATT
 701 TCCAGGATGGTAATAGGTTAAGAATTTCTCTAAATGGTTATTTCTTTTCTTTCTGCAGCTCTCACGTGTGAATATGTGTCTAGTGCATCCTTAACCTGAG
 801 GACTTCACCAGTTCGAAATTACAGTTTTCACCATCAACTACCTTATCCTTTTTGGCCTGGTTTTCTTCCTCAAACAGTGGAAACATTTTTAAAGTTGCTT
 901 TTGTTGCAGAGTTAAACAA ATG GCTGATAGTGGCTTAGATAAAAAATCCACAAAATGCCCCCGACTGTTCATCTGCTTCTCAGAAAGATGTACTTTGTGTA
                          M  A  D  S  G  L  D  K  K  S  T  K  C  P  D  C  S  S  A  S  Q  K  D  V  L  C  V   27
1001 TGTTCCAGCAAAACAAGGGTTCCTCCAGTTTTGGTGGTGGAAATGTCACAGACATCAAGCATTGGTAGTGCAGAATCTTTAATTTCACTGGAGAGAAAAA
       C  S  S  K  T  R  V  P  P  V  L  V  V  E  M  S  Q  T  S  S  I  G  S  A  E  S  L  I  S  L  E  R  K   60
1101 AAGAAAAAAATATCAACAGAGATATAACCTCCAGGAAAGATTTGCCCTCAAGAACCTCAAATGTAGAGAGAAAAGCATCTCAGCAACAATGGGGTCGGGG
       K  E  K  N  I  N  R  D  I  T  S  R  K  D  L  P  S  R  T  S  N  V  E  R  K  A  S  Q  Q  Q  W  G  R   93
1201 CAACTTTACAGAAGGAAAAGTTCCTCACATAAGGATTGAGAATGGAGCTGCTATTGAGGAAATCTATACCTTTGGAAGAATATTGGGAAAAGGGAGCTTT
     G  N  F  T  E  G  K  V  P  H  I  R  I  E  N  G  A  A  I  E  E  I  Y  T  F  G  R  I  L  G  K  G  S  F  127
1301 GGAATAGTCATTGAAGCTACAGACAAGGAAACAGAAACGAAGTGGGCAATTAAAAAAGTGAACAAAGAAAAGGCTGGAAGCTCTGCTGTGAAGTTACTTG
       G  I  V  I  E  A  T  D  K  E  T  E  T  K  W  A  I  K  K  V  N  K  E  K  A  G  S  S  A  V  K  L  L   160
1401 AACGAGAGGTGAACATTCTGAAAAGTGTAAAACATGAACACATCATACATCTGGAACAAGTATTTGAAACGCCAAAGAAAATGTACCTTGTGATGGAGCT
       E  R  E  V  N  I  L  K  S  V  K  H  E  H  I  I  H  L  E  Q  V  F  E  T  P  K  K  M  Y  L  V  M  E   193
1501 TTGTGAGGATGGAGAACTCAAAGAAATTCTGGATAGGAAAGGGCATTTCTCAGAGAATGAGACAAGGTGGATCATTCAAAGTCTCGCATCAGCTATAGCA
     L  C  E  D  G  E  L  K  E  I  L  D  R  K  G  H  F  S  E  N  E  T  R  W  I  I  Q  S  L  A  S  A  I  A  227
1601 TATCTTCACAATAATGATATTGTACATAGAGATCTGAAACTGGAAAATATAATGGTTAAAAGCAGTCTTATTGATGATAACAATGAAATAAACTTAAACA
       Y  L  H  N  N  D  I  V  H  R  D  L  K  L  E  N  I  M  V  K  S  S  L  I  D  D  N  N  E  I  N  L  N   260
1701 TAAAGGTGACTGATTTTGGCTTAGCGGTGAAGAAGCAAAGTAGGAGTGAAGCCATGCTGCAGGCCACATGTGGGACTCCTATCTATATGGCCCCTGAAGT
       I  K  V  T  D  F  G  L  A  V  K  K  Q  S  R  S  E  A  M  L  Q  A  T  C  G  T  P  I  Y  M  A  P  E   293
1801 TATCAGTGCCCACGACTATAGCCAGCAGTGTGACATTTGGAGCATAGGCGTCGTAATGTACATGTTATTACGTGGAGAACCACCCTTTTTGGCAAGCTCA
     V  I  S  A  H  D  Y  S  Q  Q  C  D  I  W  S  I  G  V  V  M  Y  M  L  L  R  G  E  P  P  F  L  A  S  S  327
1901 GAAGAGAAGCTTTTTGAGTTAATAAGAAAAGGAGAACTACATTTTGAAAATGCAGTCTGGAATTCCATAAGTGACTGTGCTAAAAGTGTTTTGAAACAAC
       E  E  K  L  F  E  L  I  R  K  G  E  L  H  F  E  N  A  V  W  N  S  I  S  D  C  A  K  S  V  L  K  Q   360
2001 TTATGAAAGTAGATCCTGCTCACAGAATCACAGCTAAGGAACTACTAGATAACCAGTGGTTAACAGGCAATAAACTTTCTTCGGTGAGACCAACCAATGT
       L  M  K  V  D  P  A  H  R  I  T  A  K  E  L  L  D  N  Q  W  L  T  G  N  K  L  S  S  V  R  P  T  N   393
2101 ATTAGAGATGATGAAGGAATGGAAAAATAACCCAGAAAGTGTTGAGGAAAACACAACAGAAGAGAAGAATAAGCCGTCCACTGAAGAAAAGTTGAAAAGT
     V  L  E  M  M  K  E  W  K  N  N  P  E  S  V  E  E  N  T  T  E  E  K  N  K  P  S  T  E  E  K  L  K  S  427
2201 TACCAACCCTGGGGAAATGTCCCTGATGCCAATTACACTTCAGATGAAGAGGAGGAAAAACAGTCTACTGCTTATGAAAAGCAATTTCCTGCAACCAGTA
       Y  Q  P  W  G  N  V  P  D  A  N  Y  T  S  D  E  E  E  E  K  Q  S  T  A  Y  E  K  Q  F  P  A  T  S   460
2301 AGGACAACTTTGATATGTGCAGTTCAAGTTTCACATCTAGCAAACTCCTTCCAGCTGAAATCAAGGGAGAAATGGAGAAAACCCCTGTGACTCCAAGCCA
       K  D  N  F  D  M  C  S  S  S  F  T  S  S  K  L  L  P  A  E  I  K  G  E  M  E  K  T  P  V  T  P  S   493
2401 AGGAACAGCAACCAAGTACCCTGCTAAATCCGGCGCCCTGTCCAGAACCAAAAAGAAACTC TAA GGTTCCCTCCAGTGTTGGACAGTACAAAAACAAAGC
     Q  G  T  A  T  K  Y  P  A  K  S  G  A  L  S  R  T  K  K  K  L  .                                       514
2501 TGCTCTTGTTAGCACTTTGATGAGGGGGTAGGAGGGGAAGAAGACAGCCCTATGCTGAGCTTGTAGCCTTTTAGCTCCACAGAGCCCCGCCATGTGTTTG
2601 CACCAGCTTAAAATTGAAGCTGCTTATCTCCAAAGCAGCATAAGCTGCACATGGCATTAAAGGACAGCCACCAGTAGGCTTGGCAGTGGGCTGCAGTGGA
2701 AATCAACTCAAGATGTACACGAAGGTTTTTTAGGGGGGCAGATACCTTCAATTTAAGGCTGTGGGCACACTTGCTCATTTTTACTTCAAATTCTTATGTT
2801 TAGGCACAGCTATTTATAGGGGAAAACAAGAGGCCAAATATAGTAATGGAGGTGCCAAATAATTATGTGCACTTTGCACTAGAAGACTTTGTTAGAAAAT
2801 TACTAATAAACTTGCCATACGTATTACAGCAGAAGTGCTTCAGTCATTCACATGTGTTCGTGAGATTTTAGGTTGCTATAGATTGTTTAAGACAGCTTAT
3001 TTTAAATGTAGAAAAATAGGAGATTTTGTAACTGCTTGCCATTAACTTGCTGCTAAATTCCCAATGTATTGATTAAATC AATAAA AAACAGATGTTACTC
3101 AGCA
```

Figure 3.20: Full length sequence of the principal transcript from human *STK33* and its inferred amino-acid sequence.

Start and stop codon and poly-denylation site are shown in bold face.

To investigate the transcription of *Stk33* in the mouse, organs were prepared from Balb/C laboratory mice of the two genders, and used freshly or frozen for the extraction of total RNA. These RNAs were used as template for RT-PCR with *Stk33* specific primers. RT-PCR amplification was successful only for lung RNA, whereas no product was obtained from muscle, kidney or uterus. In further experiments this turned out to be due to technical reasons. As in human, also the mouse showed variability at the RNA level. The cDNA fragment obtained with the primer combination **m4**x**m2** displays the expected size (1,844b) according to the predicted *Stk33* sequence. The primer combination **m6**x**m2** (see figure 3.21) produced at least one additional band than the one expected of 2,202b, suggesting a diverse splicing pattern of the region included.

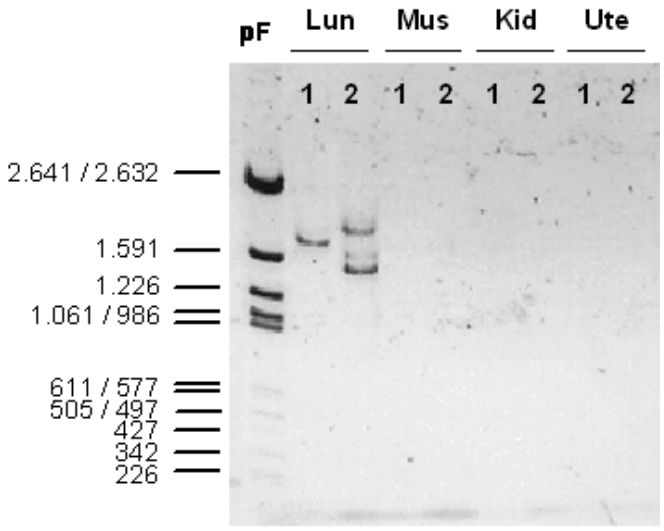


Figure 3.21: RT-PCR amplification of mouse *Stk33* from total RNA extracted from lungs, muscle, kidney and uterus.

Probes in tracks 1 were amplified with forward primer **m4** and probes in tracks 2 were amplified with the forward primer **m6**. Both forward primers were selected upstream from putative start codon. All probes were amplified with reverse primer **m2**. Reverse primer was selected downstream the putative stop codon. The relative positions of the primers are shown in the figure 3.22.

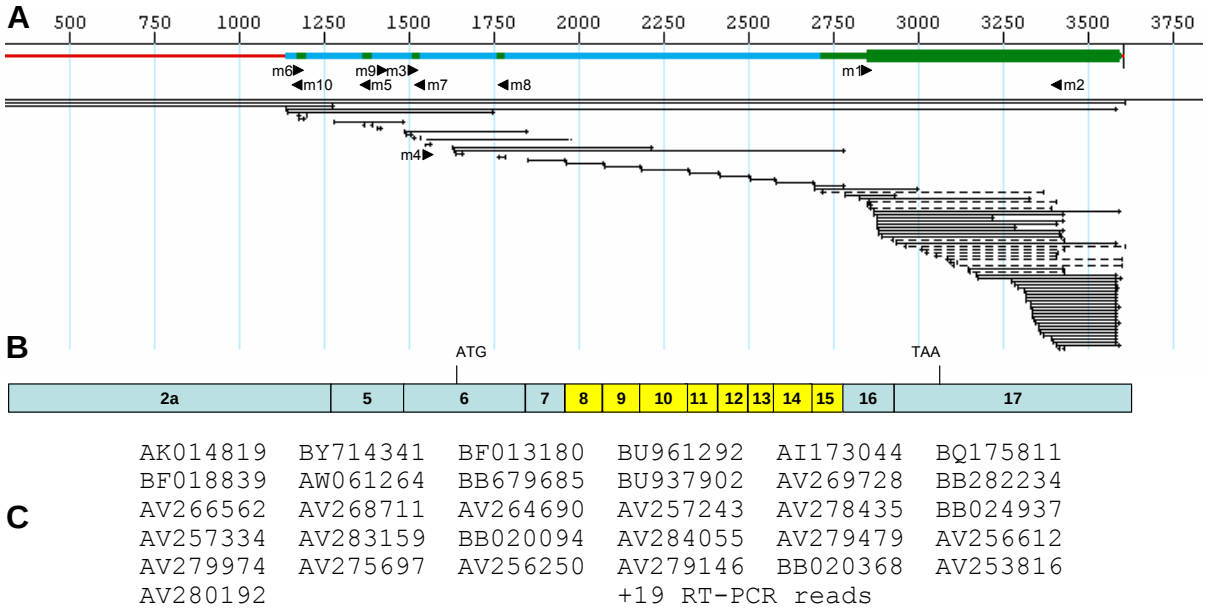


Figure 3.22: cDNA alignment of the main transcript from mouse *Stk33* gene.
The alignment was done with the program SeqManII (Lasergene's DNAStar). A: Graphic representation of the alignment. Continuous arrows represent sequences in the Watson strand; segmented lines represent sequences in the complementary Crick strand. Short black triangles signal the positions and orientation of the primers employed to amplify and sequence the DNA. B: Relative positions of the exons represented in the transcript in scale with the upper alignment. Positions of the ATG start codon and the TAA Stop codon limits the UTR from coding regions. Yellow boxes represent exons coding for the protein kinase domain. C: Sequences supporting this alignment. A list with the GenBank Accession numbers of publicly available EST is shown. Additionally, 19 reads from RT-PCR sequences obtained from human uterus were necessary to close the full length sequence and resolve single-pass sequencing errors in the ESTs.

The determination of the transcriptional start site of mouse *Stk33* has been problematic; RACE experiments did not deliver any reliable result. The 5' UTR of mouse *Stk33* is very (G+C)-rich (as shown in the PIP results in the figure 3.12) where the poly-G primer anneals prematurely without extending the transcript to its 5' end. Attempts with protocols for (G+C)-rich 5' UTRs have also failed. Analysis of the full-length cDNA clones available in the data bases suggests up to two potential transcription start points, in exon 1 and 2. This evidence is not as well supported as by human.

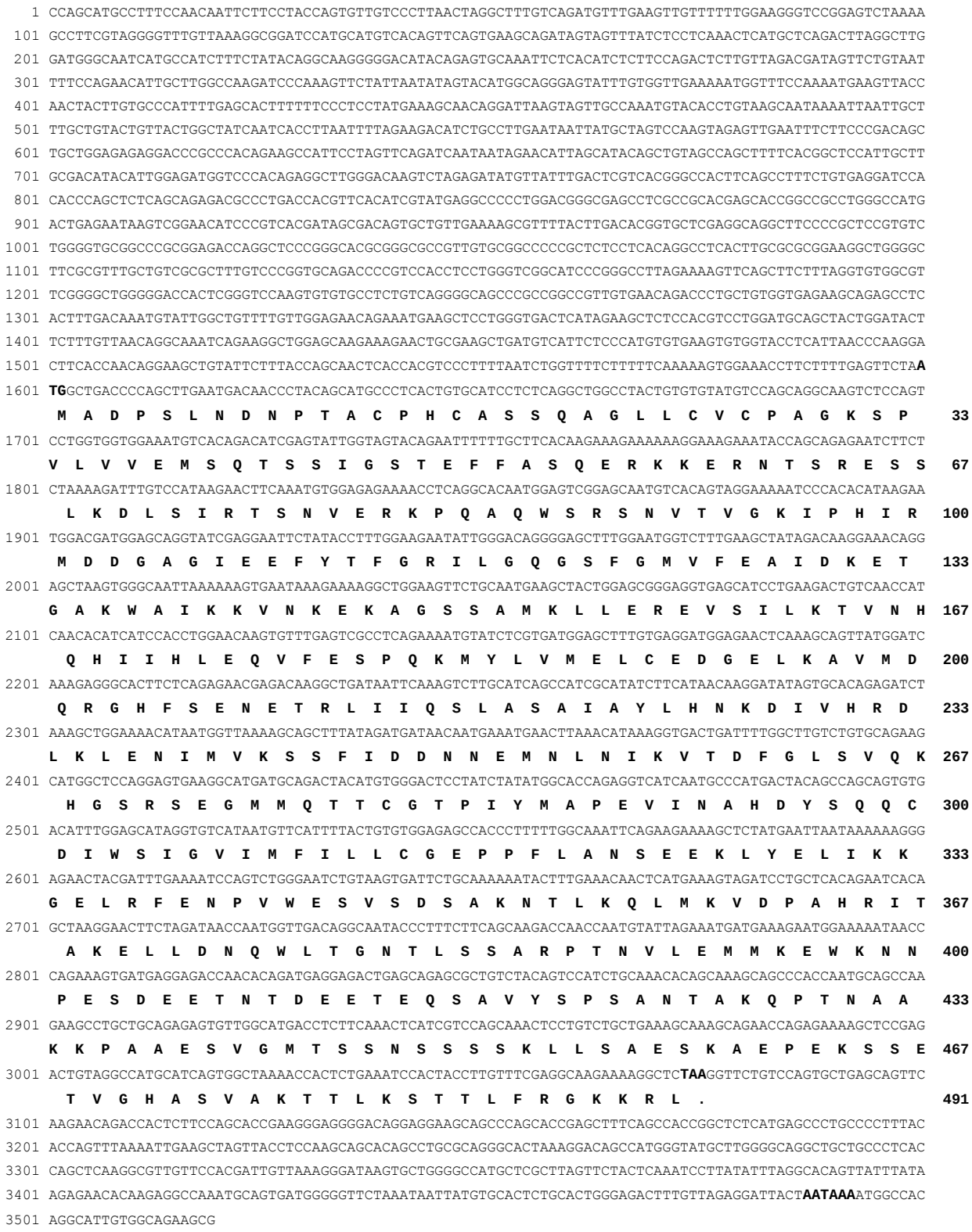


Figure 3.23: Full length sequence of the principal transcript from mouse *Stk33* and its protein product.

Start and stop codon and poly-adenylation site are shown in bold face.

a) EST pattern

STK33 and *Stk33* are represented by few entries in the database of ESTs, as shown in table 3.7 below. The EST data reflects the tissue's transcription state, the so called *Transcriptome*. The relative frequency of EST entries for a given gene is a first very rough but informative approximation of its expression pattern. Accordingly, housekeeping genes, being transcribed virtually in any kind of tissue, in every developing state and for some cases in high amounts, typically are well represented in EST databases. In contrast, genes showing differential and general low expression rates get moderate representations in only those EST projects targeting the tissues where they are active and are likely to get outnumbered from those much frequent transcripts.

Table 3.7: Number of EST entries in UniGene of some human and murine genes

Human	UniGene Cluster	EST entries	Mouse	UniGene Cluster	EST entries
STK33	**Hs.148135**	80	*Stk33*	**Mm.79075**	36
GAPDH	**Hs.169476**	16,200	*Gapdh*	**Mm.333399**	1,643
CAMKI	**Hs.512804**	139	*CamKI*	**Mm.277373**	113
ST5	**Hs.117715**	290	*St5*	**Mm.252009**	140
LMO1	**Hs.1149**	32	*Lmo1*	**Mm.12607**	47
WEE1	**Hs.249441**	183	*WEE1*	**Mm.287173**	166

STK33/Stk33: Serine/threonine kinase 33; ***GAPDH/Gapdh:*** Glyceraldehyd 3-phosphate dehydrogenase, is a housekeeping gene and is broadly used as normalisation standard of gene expression; ***CAMK1/Camk1:*** Calcium/Calmodulin dependent protein kinase I, the human kinase with higher blastp matching to *STK33;* ***ST5/St5, LMO1/Lmo1*** and ***WEE1/Wee1*** (the latter also a protein kinase-gene) are single copy genes syntenic to *STK33* which had been associated with some neoplasias (see section 1.4. of this work).

EST data obviously rely on the availability of the studied tissues, thus the absence of EST from a given tissue for a given gene, whether informative, is not a proof of the lack of transcription there. Table 3.8 shows all EST occurrences of *STK33/Stk33* according to UniGene.

Table 3.8: Number of *STK33/Stk33* EST entries in UniGene per tissue

Human *Hs.148135*	#	**Murine** *Mm.79075*	#
Reproductive system			
Testis	14	Testis	23
Uterus	1	Spermatocytes	5
Prostate	1	Round spermatides	2
Respiratory system			
Lung	1		
Lung epithelial cells	5		
Embrionic/Fetal			
Embryo, 8 weeks	1		
Fetal eyes	2		
Fetal pancreas	2		
Others			
Germinal center B cell	1	Pituitary gland	1
Medulla	1	Retina	1
Cochlea	1	Subfornical organ and postrema	1
Pooled	13	Pooled	1
Diseases			
Carcinoid (diverse types)	23		
Carcinoma cella line	1		
Primary lung cystic fibrosis epithelial cells	3		
Not specified	10		2
Total	**80**		**36**

Some general considerations should be kept in mind when EST data are discussed. ESTs are error prone since the sequencing is performed just once. Because of this, entries of multi-copy genes with few base changes may be hard to distinguish from each other. *STK33* and *Stk33* are probably single-copy genes as far as can be concluded from the analysis of the human and mouse genomes. At the nucleotide level, *STK33/Stk33* is divergent enough from other kinases so that their EST entries are unmistakeable. Another issue concerning EST analysis is the coverage of the entries. Typical EST reads are usually not “full length” because reverse transcription is rarely complete without additional protocols and because internal priming may occur if the transcripts are A-rich or the annealing conditions are not stringent enough (this issues got nicely demonstrated in the preliminary EST analysis from *STK33/Stk33* described in the figure 3.18).

b) Evidence of alternative splicing and other variability of *STK33/Stk33* transcripts

All RT-PCR experiments from *STK33/Stk33* have produced a variety of bands in most of the tissues tested. Different PCR conditions were tested including different polymerases, annealing temperatures and helping reagents, in all cases at least two different cDNA fragments resulted as depicted in the figure 3.24. PCR bands extraction, cloning and sequencing of the DNA fragments showed reproducibly that the principal varying bands were alternative versions of *STK33/Stk33* mRNAs and not unspecific PCR artifacts. The variant transcripts seem to be tissue-specific.

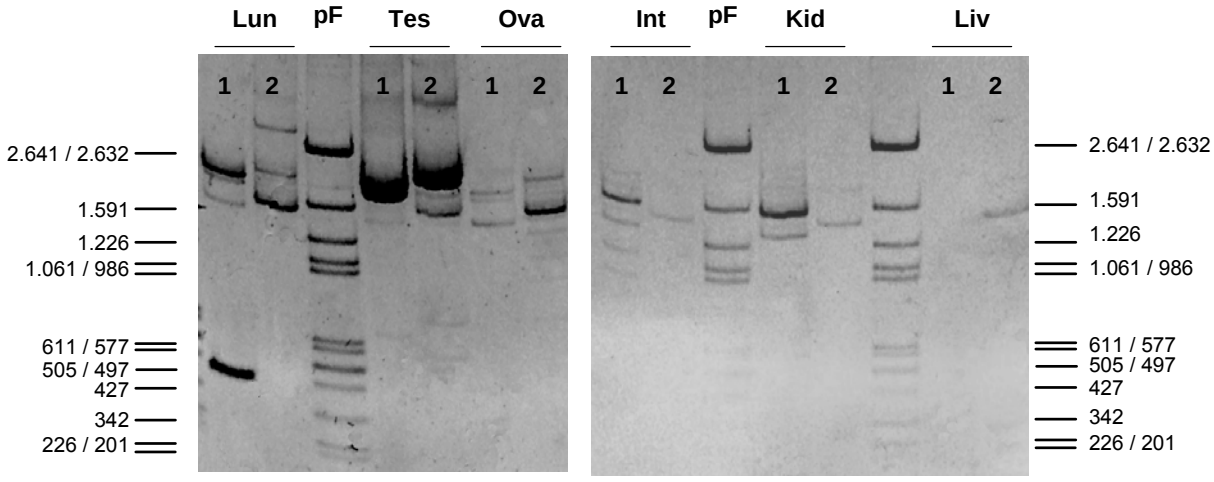


Figure 3.24: ***Stk33*** **RT-PCR amplification from polyA+ RNA extracted from mouse organs**
Lun: lungs, **Tes:** testis, **Ova:** ovary, **Int:** intestine, **Kid:** kidney and **Liv:** liver. Probes in tracks 1 were amplified with the primer combination **m4**x**m2** and probes in tracks 2 were amplified with the primer combination **m6**x**m2** (Brauksiepe 2003)

Clearly, alternative splicing does play a significant role in the expression of *STK33/Stk33* gene. However, the complexity of alternative spliced RNAs makes a more detailed analysis indispensable. Although the direction is clear, at this level it is more

cautious to talk about alternative variants instead of real alternative transcripts. A first look to these alternative versions is provided by combining the experimental data presented here and the expression data available in the databases. Results are shown in the figures 3.25 and 3.26.

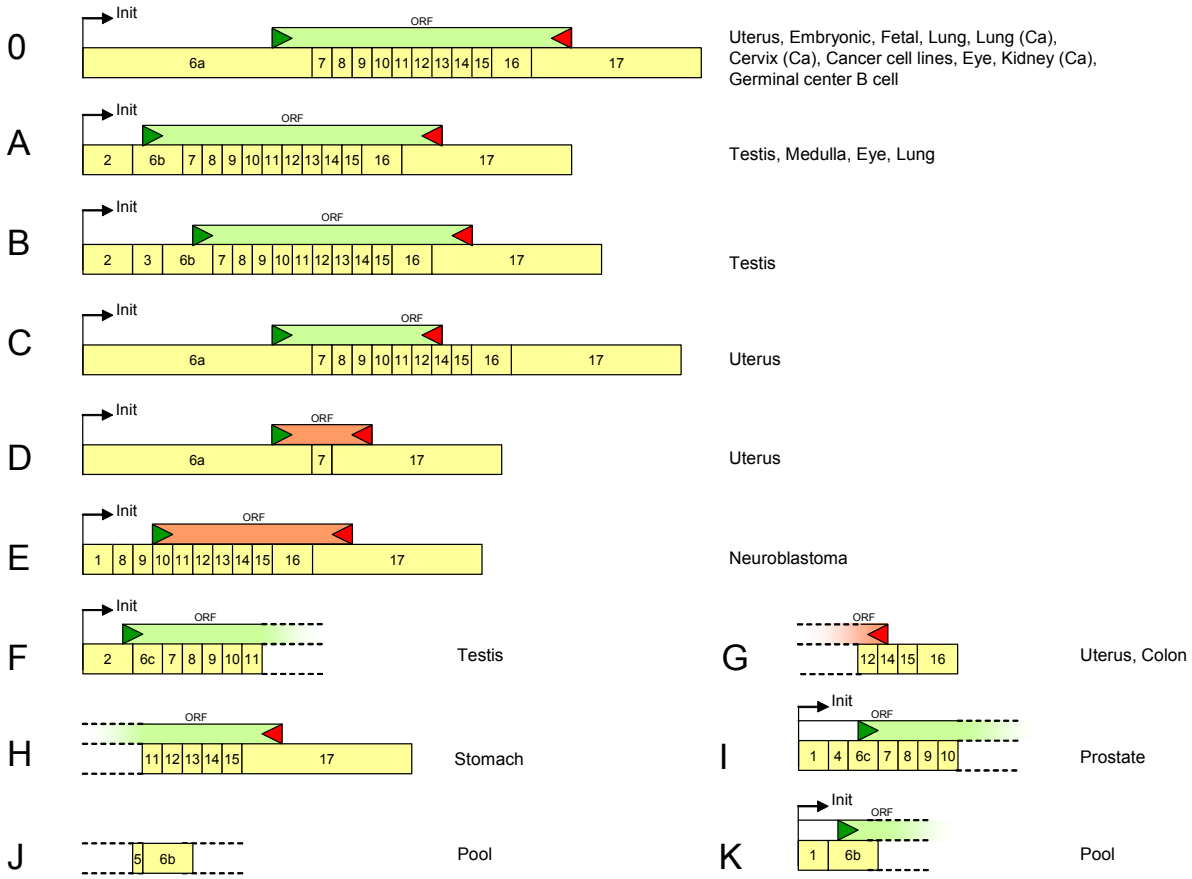


Figure 3.25: Preliminary alternative versions of human *STK33*.
Numbered boxes represent the exons. Green and red triangles over the transcripts represent the relative positions of start and stop codon respectively. The ORFs are coloured with light green or light red when they code for an intact or truncated kinase domain respectively. Right to every transcript, the tissue of origin is shown; **(Ca)** stands for malignant tissue. Putative initiation of transcription is signalled in some transcripts. Principal transcript is named here zero **(0)** for being the first transcript amplified from human uterus and clearly the most abundant in both our experiments and in the EST database. Transcripts **0** to **E** are very likely to be full length; transcripts **F** to **K** are incomplete. Transcripts **0**, **C** and **D** were obtained in the lab and transcripts **A, B, E, F, G, H, I, J** and **K** are represented in the EST database.

The observed alternative versions of *STK33/Stk33* show no dramatic differences in the coding region. Much of the variation occurs in the 5' UTR, which is not a novel observation among eukaryotic genes, or produces such truncations in the open reading frame

that, in almost all cases, a resulting active kinase is hardly imaginable. Compared with the principal transcript, the variant **A** in human *STK33* codes the same open reading frame but has a different initiation of transcription. The variant **B** has the potential of coding for a protein with the same kinase domain but with a slightly different N-terminal sequence. With a premature stop codon compared with the main transcript, variant **C** codes for a protein product with slightly shorter C-Terminal, but still the principal kinase features remain. Other transcripts code for protein products which are very likely inactive or are still incomplete. In general, the most variability is observed in the 5' UTR producing transcripts with at least three different start points of transcription.

Versions observable as well in human (**E**) as in the mouse (**A** and **B**), have the potential of producing isoforms with missing or incomplete ATP-binding motifs. The number of transcripts shown here are not representative of comparative abundance of alternative transcripts between human and mouse. With a total of 71 entries, human *STK33* is much better represented in the databases than the mouse *Stk33* with 45. On the other hand RT-PCR data from a variety of mouse organs is shown here, whereas only from human uterus was screened.

As already mentioned, *STK33/Stk33* shows also variability in the starting point of transcription. Comparison of the data produced in this work with the EST entries available in the databases, suggests that the starting point of transcription varies in a tissue-specific manner. In human, testis, uterus and prostate seem to have distinct starting points of transcription. Despite splicing variability, in testis and lungs from the mouse the starting point is at the beginning of exon 2, with, one exceptional EST from testis which starts in exon 1, together with some from other tissues. Supporting this alternative starting point

candidates is the observation that exons 1, 2 and 6 have no upstream 3' splice acceptor signal and are mutually exclusive, all transcripts observed up to now exhibit always just one of them (see figure 3.25 above). In the mouse transcripts, mutually exclusion is observed between exons 1 and 2.

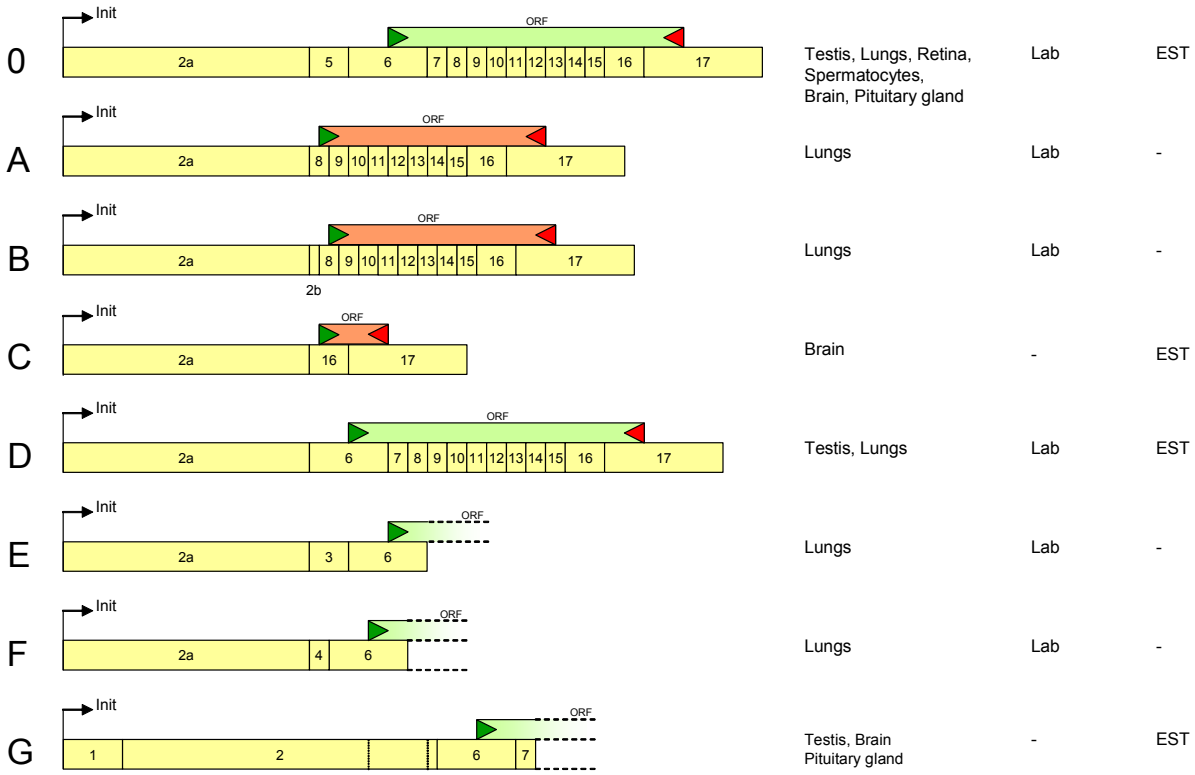


Figure 3.26: Preliminary alternative versions of mouse *Stk33*.
Numbered boxes represent the exons. Green and red triangles over the transcripts represent the relative positions of start and stop codon respectively. The ORFs are coloured with light green or light red when they code for an intact or truncated kinase domain respectively. Right to every transcript, the tissue of origin is shown. Transcripts obtained experimentally are marked with the tag **Lab**, and those supported by entries from databases are marked with **EST**. There is no experimental data confirming the start points of transcription, however exons 1 and 2a were deduced from transcripts reported as full length clones. Principal transcript is named here zero (**0**) for being the most abundant in both our experiments and in the EST database. Transcripts **0** to **D** are very likely to be full length, transcripts **E** to **G** are incomplete.

Variability is also observable at the end of the transcript. Human EST data suggests two different poly-adenylation signals (EST entries: AK122808.1 and AL833011.1). This phenomenon has been not yet detected in mice.

Also potential single-base polymorphisms are observable. Worth to mention the one at positions 1,308 and 1,309 downstream of the start codon, in exon 16, where a TG dinucleotide is present in all data from the laboratory and several ESTs (BE885808, BG186360, BM984480, etc.), whereas in two ESTs from different projects (AW959747, AA454060) the corresponding sequence reads GA, which would give rise to an exchange of two amino acids (GAU.GCC: AspAla → GAG.ACC: GluThr). The same difference is also observed between the data from the two human genome projects, supporting the interpretation as a natural polymorphism which, however, does not lay within the region coding for the catalytic domain, described in coming sections.

c) Multiple human Tissue Array

Dot blot experiments with a Multiple Tissue Array™ (Clonetech) analysis suggest that STK33 is expressed in a limited variety of tissues. The array contains a dot's matrix of immobilised cDNAs derived from normalised total RNA from different human tissues, mostly normal and some carcinogenic ones. The rational of the method is that, since the amount of material may be assumed to be regular along the sample population, hybridisation with a specific labelled DNA probe provides a pattern of expression for the gene studied and

a first semi-quantitative measure of the relative levels of its expression in all tissues evaluated.

To avoid cross-hybridisation with similar kinase genes, the exons coding for the eukaryotic protein kinase catalytic domain were excluded from the probe used. As shown in the figure 3.27, the probe is a fragment of exon 17 and contains the last 186 coding bases plus 490 bases from the 3' UTR. The uniqueness of the probe was checked by BLAST searches against all human sequences in the databases in particular with the EST subset and by RepeatMasker.

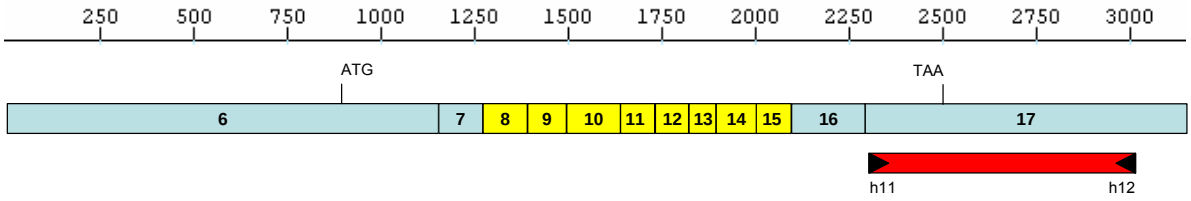


Figure 3.27: *STK33*-specific DNA probe used for hybridisation experiments.
Primers h11 and h12 were used to generate the probe. Light blue boxes depict untranslated exons and regions not coding for the kinase domain, yellow boxes show exons coding for the kinase domain and the red box represents the length and placement of the DNA probe relative to the *STK33*-transcript.

In two successive experiments, weak but reproducible signals were obtained after very long exposition times (see figure below). The strongest hybridisation signals were observed in testis, fetal lung and heart, followed by weaker signals in pituitary gland, kidney, interventricular septum, pancreas, all heart tissues, trachea, thyroid gland and uterus. Very weak hybridisation signals, were observed in amygdala, aorta, esophagus, colon ascending, colon transverse, skeletal muscle, spleen, peripheral blood leukocyte, lymph node, bone marrow, placenta, prostate, liver, salivary gland, mammary gland, leukemia, HL-

60, HeLa S3, leukemia K-562, leukemia MOLT-4, Burkitts lymphoma daudi, fetal brain, fetal liver, fetal spleen, fetal thymus. No signal at all was detectable in RNA from tissues of the nervous system, were some CaMK genes are known to be expressed in high rates (Scott and Soderling 1992).

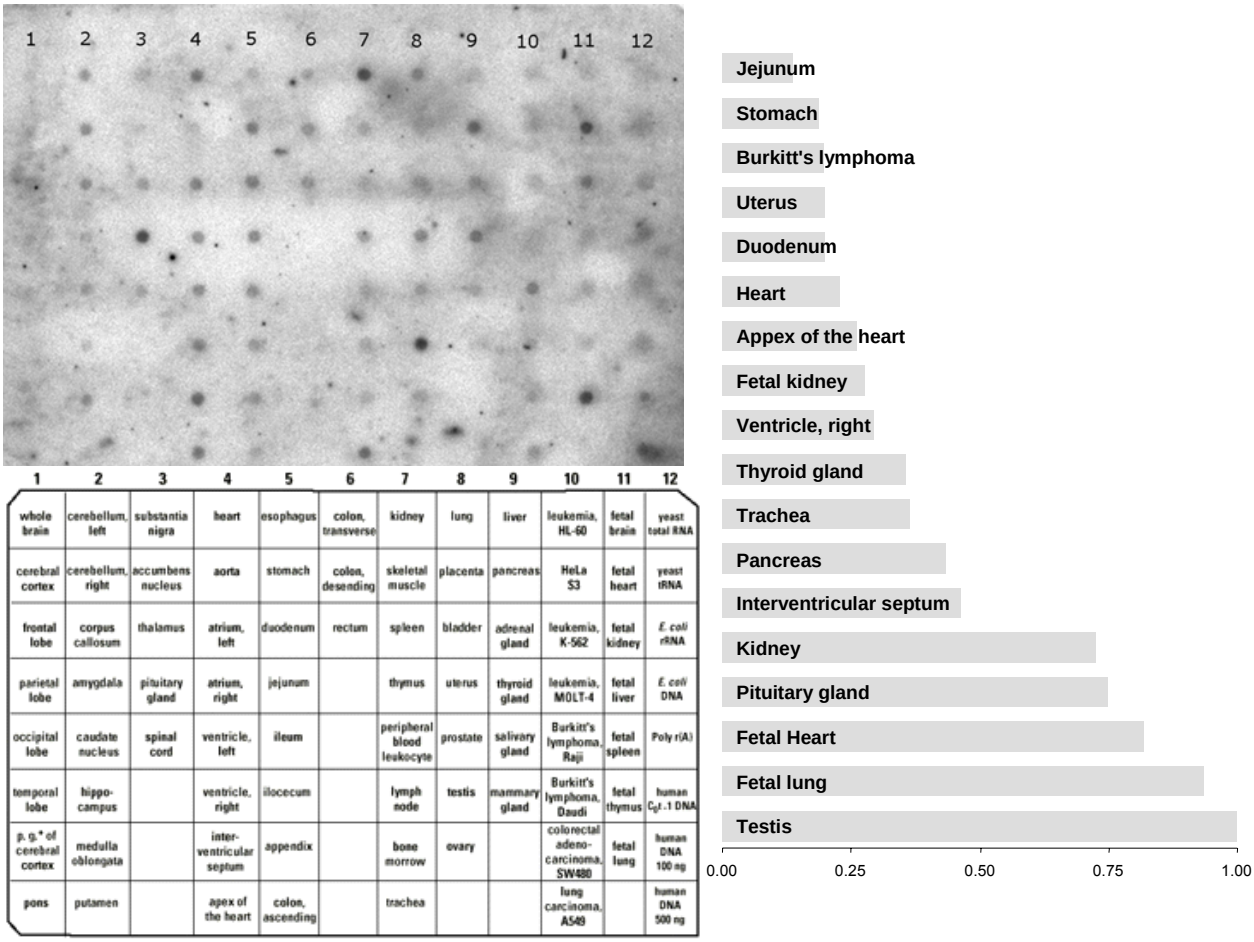


Figure 3.28: Multiple Tissue Array (MTE) from human *STK33*.
Upper left: MTE from Clonetech was hybridised with a radioactive labelled probe specific for human *STK33*. Experiment was performed in duplicate with reproducible results. **Bottom left:** tissue distribution of the MTE provided by the manufacturer. **Right:** Highest signal was detected in testis. Base line subtraction was performed and all other signals were normalised relative to the signal in testis signal quantification with Aida software.

d) Human Cancer Panel Array

To evaluate whether the expression of *STK33* is up or down regulated in human neoplasias, a Cancer Profiling Array™ (Clontech) was hybridised with the same gene specific probe and hybridisation procedures described above. In this blot, the array consists of normal, malignant and in few cases metastasis cDNAs obtained from total RNA of cancer patients. Two independent hybridisation s show reproducibly hybridisation signals in most samples from normal ovary, with a clearly much less intensity in its malignant counterparts. Lower overall signal but still a rough tendency to down regulation, is also observable in samples from tumors of other tissues, such as kidney, lung, thyroid and prostate. Very low signals in stomach, colon and rectum tissues are detected. In uterus the signal shows a mosaic-like behaviour, samples from some patients show no hybridisation at all, others show signal equally strong in normal and tumor samples, some exhibit lower signal with tumor cDNAs and at least in one patient, the contrary is the case. The panel corresponding to samples from lung was unfortunately seriously obscured by several dots of non specific hybridisation also in the second attempt of the experiment. However, also here some of the samples show slightly lower signals in the tumor cDNAs.

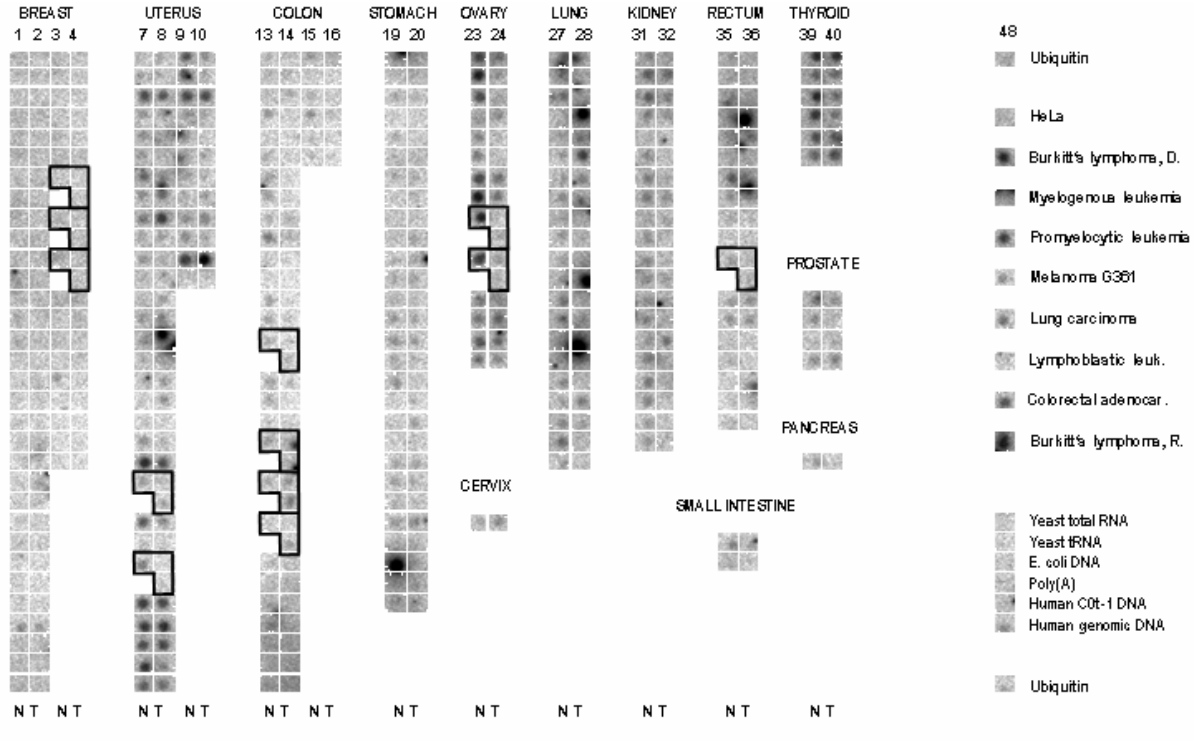


Figure 3.29: Hybridisation of the Cancer Profiling Array (Clonetech) with a *STK33*-specific probe. The Array consists of 241 pairs of normalised cDNA from tumor (**T**) and their corresponding normal (**N**) tissues from individual patients. The groups of three boxes outlined with dark lines correspond to, in counter clock wise to normal, tumor and metastasis from the same patient. Column 48 contains positive and negative controls and cDNA from some cancer cell lines. Hybridisation signals aligned with the provided manufacturer's grid, here in red, are interpreted as real signals. Darker and broader dots clearly shifted from the grid are considered traces from radioactivity not correctly removed from the media.

e) Mouse northern-blot

To analyse the expression of *Stk33* in the mouse, total RNA was isolated from mice tissues represented in the EST database: brain, lungs, liver, kidney, intestine, testis and uterus; loaded in denaturing RNA gel for electrophoresis; transferred to positively charged nylon membranes and hybridised with radioactively labelled gene-specific DNA-probe (Figure 3.30).

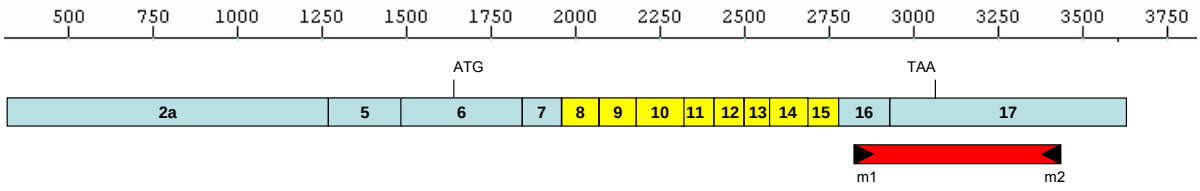


Figure 3.30: ***Stk33*-specific DNA probe used for hybridisation experiments.**
Primers **m1** and **m2** were used to generate the probe. Light blue boxes depict untranslated exons and regions not coding for the kinase domain, yellow boxes show exons coding for the kinase domain and the red box represents the length and relative position of the DNA probe relative to the *Stk33*-transcript.

Three experiments immobilizing total RNA and with varying experimental conditions have yielded negative results. Not until using relative high amounts of purified polyA+ RNA for northern analysis, a positive result can be obtained for testis RNA (Brauksipe, 2003). A faint and reproducible hybridisation signal is obtained with 12µg

polyA+ RNA from testis, though, no signal was detected from other tissues even with higher amounts of target material (20µg liver, 15µg lungs, 20µg stomach), even after long exposition time with high sensitive BioMax MS autoradiography film. Finally, two previously hybridised northern blots were stripped with formamide at 68°C and re-hybridised with a double labelling of probes specific for *Stk33* and a housekeeping gene for signal normalisation. In these last experiments, ten times more DNA than recommended was used for the hybridisation. The choice of the housekeeping gene candidate was taken with care in order to avoid overlap between both signals. For this reason, ribosomal protein genes *L19* and *L32* were selected, due their relative short mRNA sizes. In particular *L19* was favoured due its broader expression and in particular due its higher expression in testis, according to the Gene Expression Atlas (Su et al. 2002).

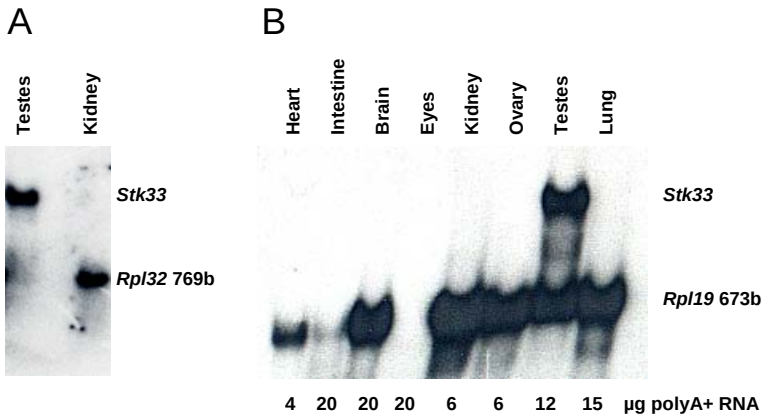


Figure 3.31: Mouse northern-blots
A: Test northern-blot with *Stk33* and *L32* probes double hybridisation. **B:** Northern-blot with *Stk33* and *L19* probes double hybridisation showing very strong signal of *Stk33* in testes, and eventually a second band supporting previous observation of alternative splicing. Remarkably, *Stk33* signal is negative in other tissues even those like brain and lung where much more target material was immobilised.

f) Mouse RNA in-situ hybridisation

To study Stk33 expression histologically, RNA in-situ hybridisation experiments were performed. Frozen sections were produced from those organs with known or expected transcription of *STK33/Stk33* based on the Multiple Tissue Expression Array and EST analysis mentioned before. Hybridisation signals were detected on liver, testis and lungs. The experiments were performed several times with varying conditions, but still further optimisation is required to fully confirm some results.

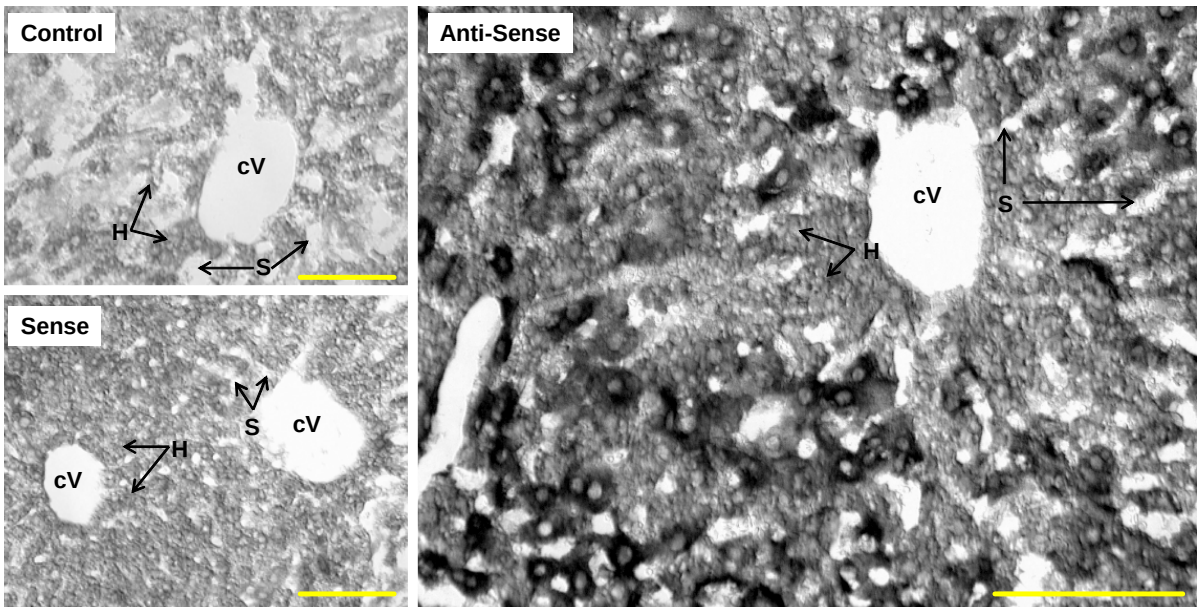


Figure 3.32: RNA *in-situ* hybridisation s of *Stk33*-specific probe with frozen sections of mouse liver. Yellow scale bar: 100µm, all pictures were taken with a 20X objective. Control and anti-sense views are shown to the left in reduced scale. Control consists of antibody detection with no previous RNA sonde. Anti-sense preparation is shown at full size to the right. **cV:** central venes; **H:** hepatocytes and **S:** sinusoid capillaries. See appendix 7.3 for a version of this images in negative colours, which confers an artificial but very illustrative three dimensional impression.

All sections show hepatic lobes with corresponding central venes (**cV**). In all three views, hepatocytes (**H**) and sinusoid capillaries (**S**) are recognizable through, respectively, dark and light zones radially arranged, converging into the central Venes (**cV**). The actual interchange between blood and liver cells occurs in the interface between hepatocytes and the sinusoid capillaries. Anti-sense probe produced signals in cells which may correspond to subsets of hepatocytes with *Stk33* differential expression. Alternatively, signals may correspond to liver-specific macrophages (or Kupffer-cells) dispersed along the sinusoid capillaries. This is supported partially by position and distribution of the signal and by the star-like form of some of the labelled cells. Kupffer-cells may build prolongations bridging one side to the other of the capillary wall.

In testis preparations, several seminal tubules sections are easily recognised separated by interstitial space. In a concentric manner in each tubule the fine surrounding basal lamina is observable, as well as the very wide germinal epithela where the actual spermatogenesis occurs in basal-luminal direction and finally the seminal lumen in the center, where the fully histologically differentiated, but still functional immature spermatozoids are released.

An example in the figure 3.33 shows strong *Stk33*-specific signal in roughly half of the basal section of the seminiferous tubuli. The one-cell-wide basal compartment of spermatogonia, separated through tight junctions of the Sertoli-cells (blood-testis barrier) from the adluminal compartment, is not clearly distinguishable in the preparations from fresh frozen tissues, and is not possible to depict in which stage of the spermatogenesis the signal is further present, but is clearly stronger at the basal region where cell division by

spermatocytes is very active, becoming fuzzy in the upper luminal section, where the meiotic spermatides are shorter and is totally absent in the late spermatides and/or spermatozoids.

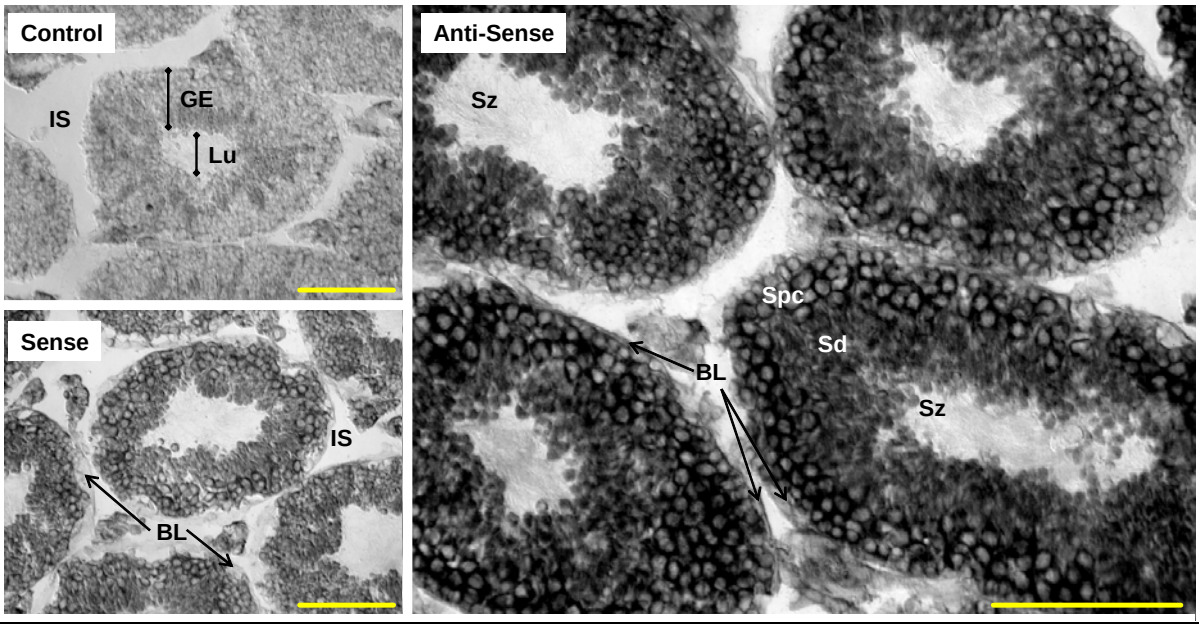


Figure 3.33: RNA *in-situ* hybridisation s of *Stk33*-specific probe with sections of frozen mouse testis. Yellow scale bar: 100µm, all pictures were taken with a 20X objective. Control and anti-sense views are shown to the left in reduced size. Control consists of antibody detection with no previous RNA hybridisation . Anti-sense preparation is shown at full size to the right. In all preparations, major features of the seminiferous tubuli are distinguishable like basal lamina (**BL**), germinal epithelium (**GE**), lumen (**Lu**) and interstitial spaces (**IS**). Particularly in sense and anti-sense preparations primary spermatocytes (**Spc**), spermatides (**Sd**) and eventually late spermatides and/or spermatozoa (**Sz**) are recognizable.

This observation is confirmed by hybridisation experiments on perfused organs and additional DNA labelling with Hoechst 33258 (example in the figure 3.34). In these preparations the histological characterisation is much easier, the tissue in the interstitial space reveals blood vessels, myoid cells are recognizable in the lamina propia, Sertoli cells nuclei become apparent and the major stages of the spermatogenesis are distinguishable. *Stk33*-specific signal is clearly restricted to the very cells of the spermatid differentiation process, from the spermatogonia to the early spermatides with a remarkable maximum of

signal in the spermatocytes. In all cases the signal encircles round nuclei, strongly supporting its association to germinal cells. Cells from the interstitial area, such as Leydig cells and vascular tissue as well as support cells from the germinal epithelia, such as Sertoli and myoid cells show no signal at all.

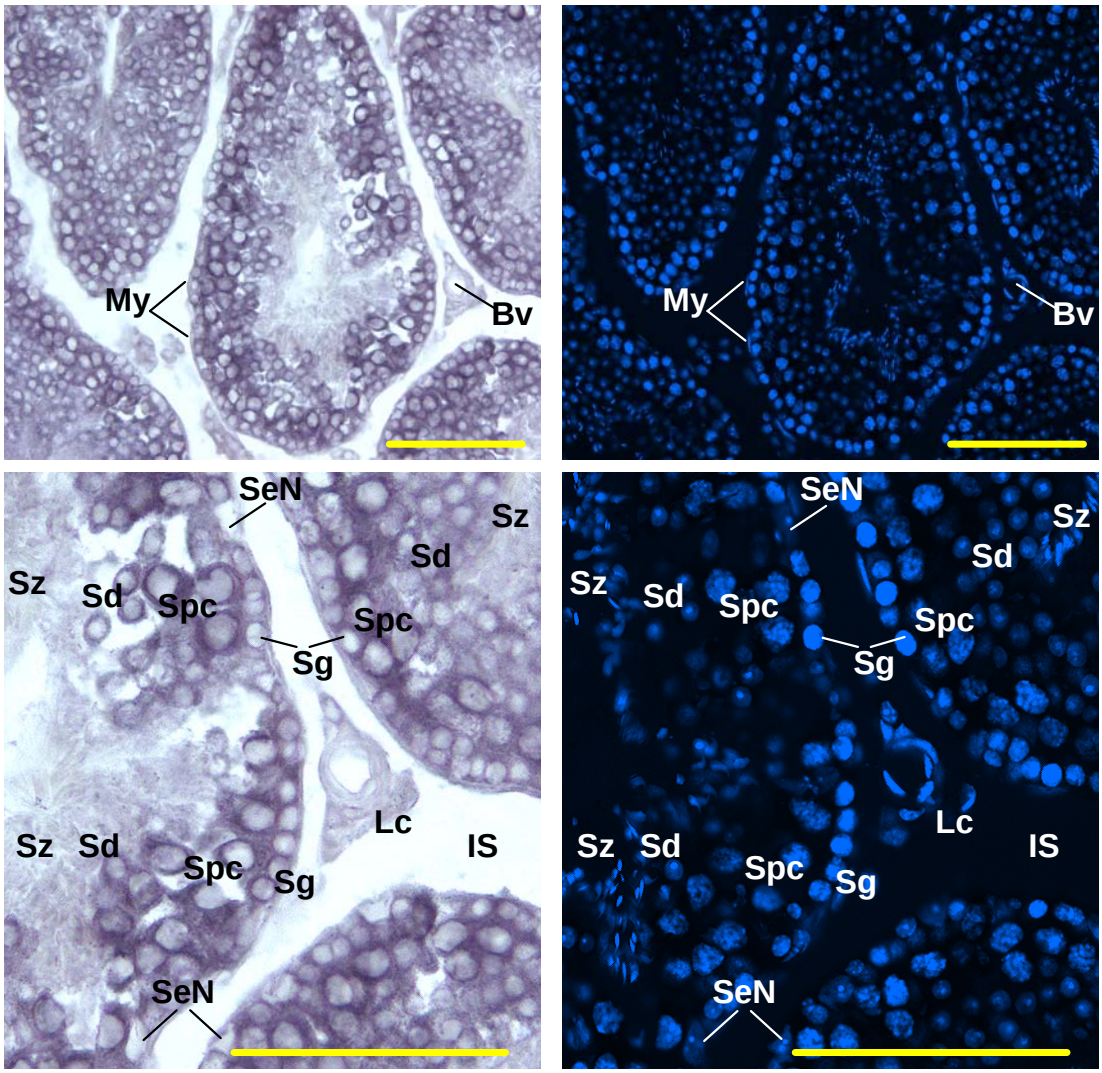


Figure 3.34: **RNA *in-situ* hybridisation s of *Stk33*-specific anti-sense sonde with sections of perfused mouse testis and nuclear staining.**

Yellow scale bar: 100µm, all pictures were taken with a 20X objective. Left panels: RNA in-situ hybridisation , right panels: nuclear-specific fluorescence staining with Hoechst 33258 for the same region. Upper panels show several sections of seminiferous tubules. **My:** Myoid cells; **Bv:** blood vessel Botton panels show an expanded view around the interstitial space (**IS**) between three seminiferous tubules. **Sg:** permatogonien; **Spc** spermatocytes; **Sd:** spermatides: **Sz:** late spermatides and/or spermatozoa; **SeN:** Sertoli cells; **Lc:** Leydig cells.

Upper panels of figure 3.34 show several sections of seminiferous tubules. Some myoid cells (**My**) are distinguishable in the basal lamina with a blood vessel (**Bv**) in the interstitial space between them. Botton panels show an expanded view around the interstitial space (**IS**) between three seminiferous tubules. Left panels present the RNA hybridisation signal as dark accumulations in some differentiation stages of the germinal epithelia. Spermatogonien (**Sg**) are organised in a one-cell width basal layer with relative big and dense nuclei; the next adluminal layer of cells correspond to the spermatocytes (**Spc**) with much bigger nuclei (with coarse chromatin) and large cytoplasm; spermatides (**Sd**) are recognised for being closer to the lumen, having shorter size and being remarkably abundant, in some cases chromatin condensation is visible; finally late spermatides and/or spermatozoa (**Sz**) are recognised through their short elongated nuclei with very condensed chromatin. Some oval or pyramidal nuclei of Sertoli cells (**SeN**) are visible mostly in basal locations with one ore more prominent nucleoli, and Leydig cells (**Lc**) in the interstitial space, in both cases with no associated *Stk33*-specific signal. The signal is clearly cytoplasmic as the light zones in the middle correlate with the nuclear staining in the right panels. The major differentiation stages are here recognizable in particular with the nuclear staining. Also the absence of the *Stk33*-specific signal in the late spermatides and/or spermatozoids in the lumen is confirmed (figure 3.35 for an example) very bright nuclear signals were obtained from the late spermatides still in the lumen of the seminiferous tubuli, with no corresponding *Stk33*-specific signal.

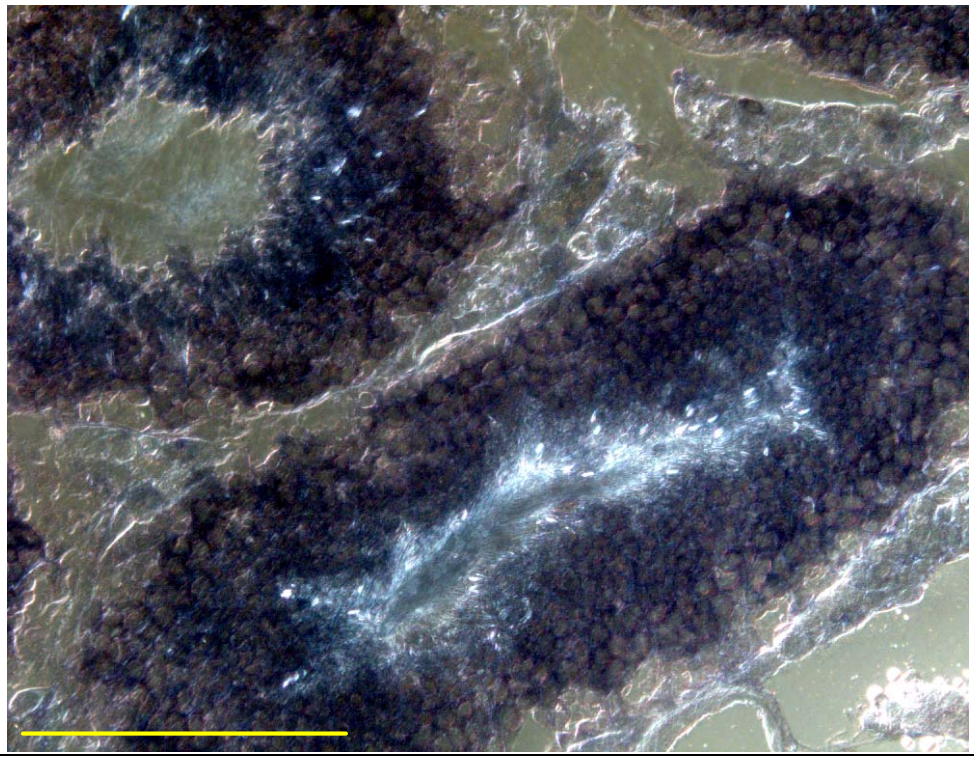


Figure 3.35: Anti-sense *in-situ* RNA hybridisation s of *Stk33*-specific sonde and Hoechst 33258 nuclear staining with a mouse testis frozen section.
Yellow scale bar: 100µm, picture was taken with a 20X objective. Combined phase contrast and fluorescent view (with false colours). Note the intensive blue staining in the lumen of one of the seminiferous tubuli and the presence of some late spermatides not yet released from the Sertoli cells.

The recognition of structures in the preparations from the lungs was particularly difficult due to the collapse of the alveoli, which are full of air in vivo. However the identification of cartilage segments at the base of the bronchi had helped by describing the surrounding regions. Cartilage constitutes an important structural element of the air conducting section of the respiratory system, i.e., trachea and bronchi, but is not present in the more flexible bronchiole and alveoli, where the actual gas exchange takes place.

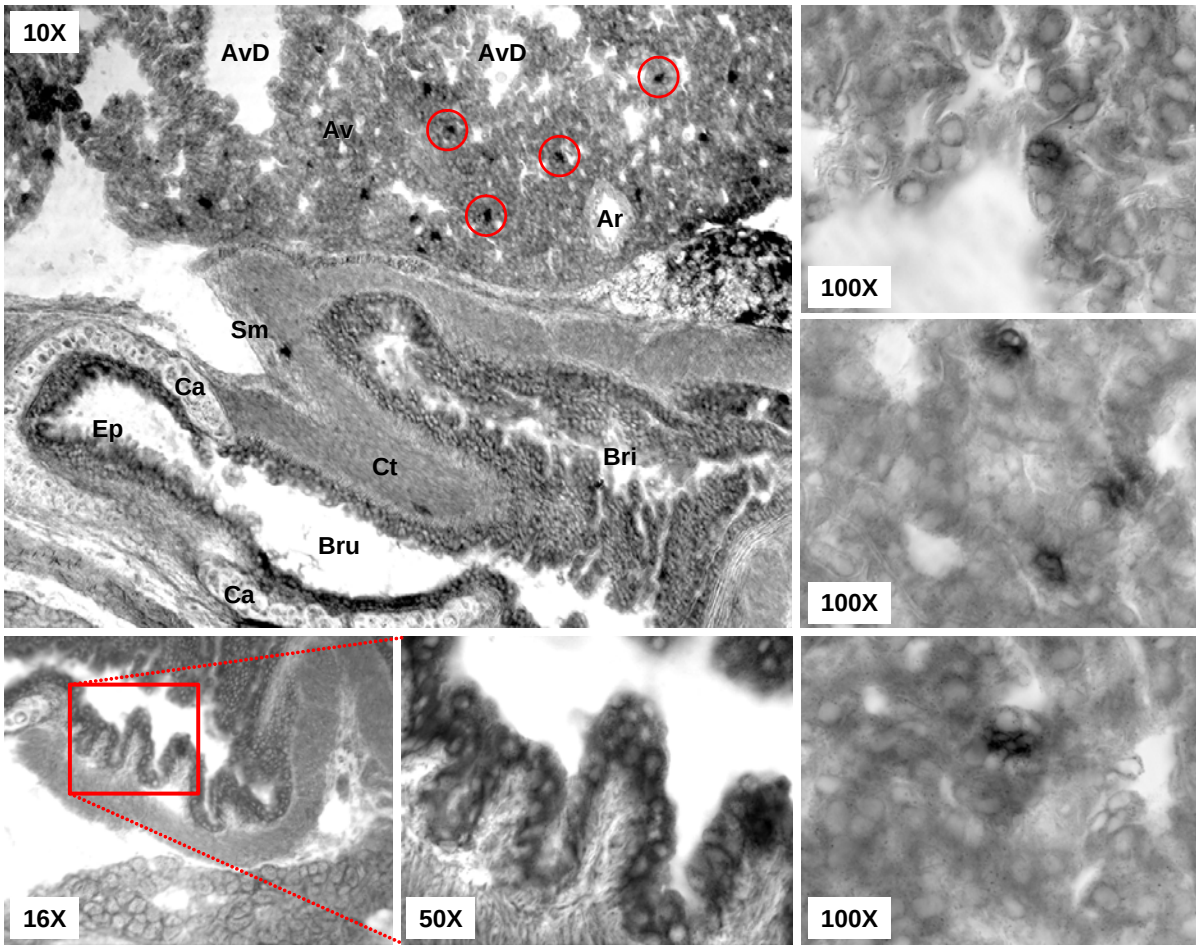


Figure 3.36: RNA *in-situ* hybridisation s of *Stk33*-specific sonde with frozen sections of mouse lungs. Objectives employed are shown in each picture. In the view with the 10X objective several structures are recognised: Cartilage (**Ca**) support the basal layer or the bronchus (**Bru**); in contrast bronchiole (**Bri**) exhibit no cartilage; other structures: Epithelia (**Ep**), smooth muscle (**Sm**), connective tissue (**Ct**), alveolar duct (**AvD**), Alveole (**Av**), Artery (**Ar**). Red circles show stark *Stk33*-specific signals dispersed in the alveole, which are documented to the right with the 100X objective, possibly alveolar macrophages. Objectives 16X and 50X, down left show details of the bronchial epithel with *Stk33*-specific signal. See appendix 7.3 for a version of this images in negative colours, which confers an artificial but very illustrative three dimensional impression.

A *Stk33*-specific signal was detected (figure 3.36) in the folded epithelia of the bronchi. Additionally, strong signals were observed in the form of isolated dispersed single cells. Nuclear staining with Hoechst 33258 revealed that these were single cells with irregularly structured nuclei typical for alveolar macrophages. Whether this interpretation holds true, has to be investigated by further experiments.

3.3.3 Protein product analysis

The *STK33/Stk33* open reading frames code for putative proteins of 57.8 and 54.5 kDa respectively. Other major predicted biochemical characteristics are shown in table 3.9.

Table 3.9: General features of putative human and mouse serine/threonine kinase 33

	Amino acids	MW kDa	pI	Charge at pH 7.0
STK33	514	57.8	6.87	-0.62
Stk33	491	54.5	6.25	-5.26

STK33 shows significant similarity (E = $5e^{-54}$) to serine/threonine kinases, in particular to members of the subfamily of calcium/calmodulin-dependent protein kinases from several eukaryotes. The best database match is with the myosin light chain kinase from *Dictyostelium discoideum* (39% identity, 58% similarity). The best similarity to any human protein was observed in the KCC1_HUMAN calcium/calmodulin-dependent protein kinase type I (Cam Kinase I) (37% identity, 55% similarity) while the similarity to the human myosin light chain kinase was much lower (28% identity 47% similarity). These data suggested that STK33 represents a novel member of the protein kinase super-family in human and mouse. Protein-domain analysis tools (Pfam, (Letunic et al. 2004); PROSITE (Falquet et al. 2002); SMART (Letunic et al. 2004) revealed a perfectly conserved serine/threonine-protein kinase catalytic domain; no other known domains seemed to be present in the protein sequence.

Analysis of the inferred amino acid sequence suggests that STK33/Stk33 has all features of a protein kinase, particularly the ATP binding domain and the phospho-transfer

active site characteristic of serine/threonine kinases. Figure 3.37 shows the span of the protein kinase domain in a human mouse alignment, the positions of the reactive signatures and potential post-translational modifications.

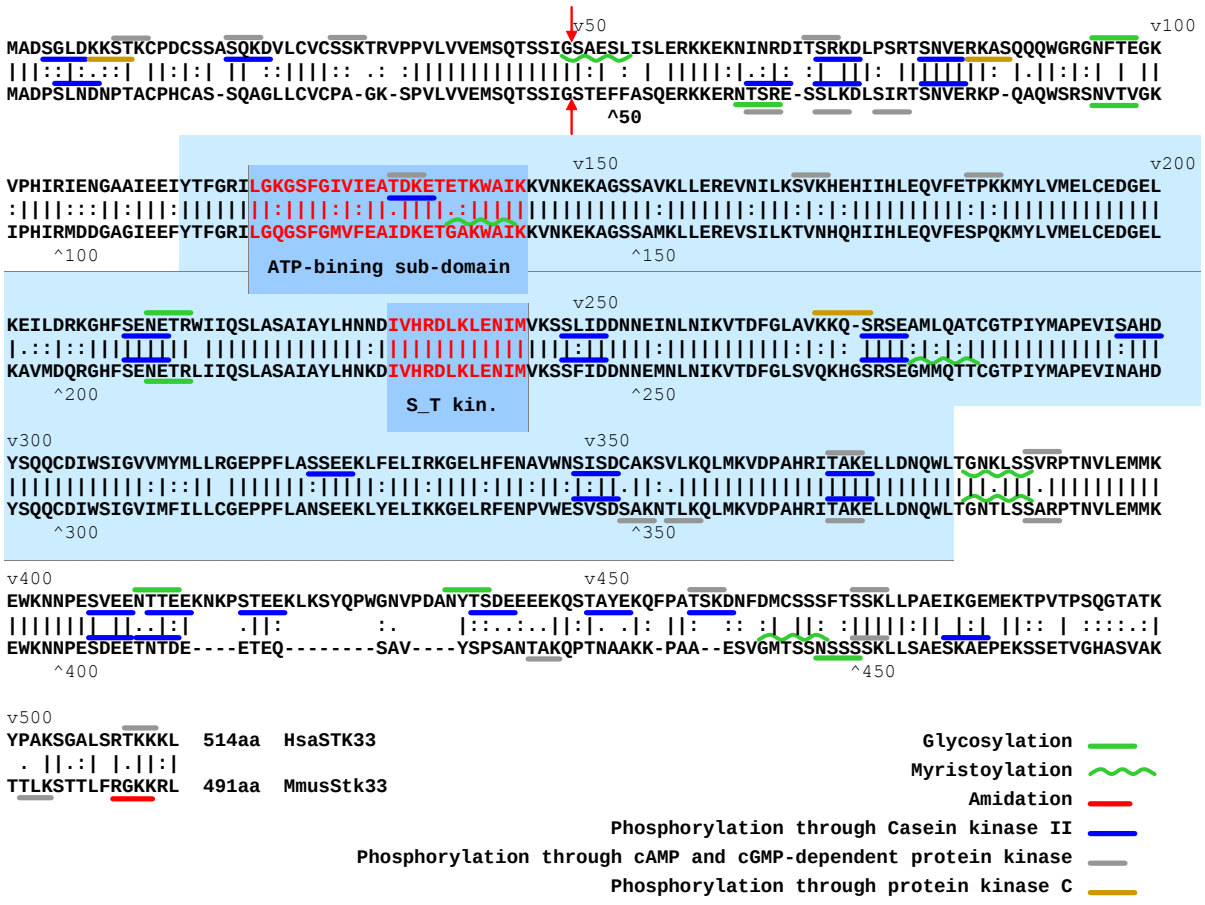


Figure 3.37: Lipman-Pearson protein alignment of human *STK33* and mouse *Stk33* deduced products and potential post-translational modifications.

Human STK33 in the first, and mouse Stk33 on the second lines as shown at the end of the alignment. Murine protein product is 23 aa shorter than the human one. The designated start methionine is the first one in the open reading frame for both human and murine sequences. The Protein kinase domain (YTFG…NQWL) is shown over a light blue background. ATP-binding site signature and Serine_Threonine phosphorylation active site are shown with red characters over darker blue boxes. Potential post-translational modifications are shown with lines of different colours (see legend). Modifications were deduced according to the PROSITE motif search (Falquet et al. 2002). Red arrows show a signal peptide cleavage site in positions 49-50 in STK33 and 46-47 in Stk33, predicted from the PSORT server (Nakai and Horton 1999). In both cases the prediction received a low statistical score. (Alignment produced with the program MegAlign from DNAStar, ktuple: 2; gap penalty: 6; gap length penalty: 8; similarity index: 67.0; gap number: 11, gap length: 25, consensus length: 516).

The canonical protein kinase subdomains are perfectly conserved in STK33/Stk33. Also most α-helices and β-sheets predicted with several methods correspond in size, number and placement to those occurring in closely and distantly related protein kinases with known 3D structures.

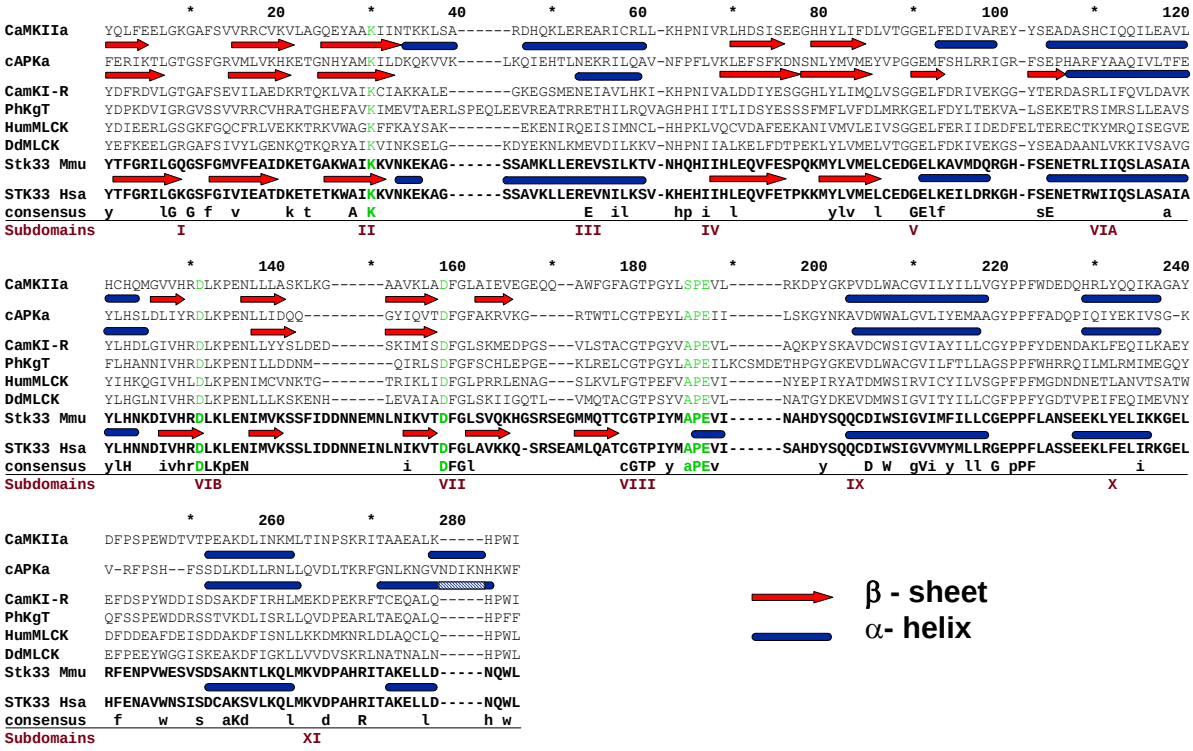


Figure 3.38: Amino-acid sequence alignment and structural features of the catalytic domain of human STK33 and mouse Stk33 with representative members of the protein kinase super-family.

Canonical protein kinase subdomains according to Hank's reference alignment and classification (Hanks and Hunter 1995) are indicated by roman numerals below the consensus string. A member of the AGC Group I of protein kinases (cAPKa) is included for comparison; all others are members of the CaMK group, with both groups belonging to the family of serine/threonine kinases. Secondary structures of cAPKa and CaMKI-R and a PSIpred (Jones, 1999) prediction for human STK33 are shown over each corresponding sequence. ß-sheets are represented by red arrows and α-helices by rounded blue boxes. PROSITE (Hofmann et al., 1999) signatures for protein kinases ATP-binding region (ATP-bd.: PS00107) and the serine/threonine protein kinases active site (S_Tk.: PS00108) are shown by dotted rectangles. For simplicity, the same naming criterias as used in the Protein Kinase Resource (www.sdsc.edu/kinases/) were taken. Full names and accession numbers are as follows: **CaMKIIa** (P11275) rat calcium/calmodulin-dependent protein kinase type II, alpha chain; **cAPKa** (P17612) human cAMP-dependent protein kinase; **CamKI-R** (1A06) calmodulin-dependent protein kinase from rat; **PhKgT** (P15735) human phosphorylase B kinase γ catalytic chain, testis/liver isoform; **HumMLCK** (Q15746): human myosin light chain kinase, smooth muscle and non-muscle isozymes; **DdMLCK** (P25323) *Dictyostelium discoideum* myosin light chain kinase. Highly conserved residues are shown in green (see a detailed discussion on the canonical protein kinase domain in section 4.4.4). Modified from (Mujica et al. 2001).

The available data on the very well-studied super-family of protein kinases make it possible to speculate about the three dimensional structure of STK33. In the 5th edition of CASP meetings (Critical Assessment of Techniques for Protein Structure Prediction), the comparative modelling between proteins sharing sequence similarity, is considered the lighter category of this biannual experiment and has accomplished very acceptable results, "with the vast majority of methods producing good models (with an r.m.s. difference <2.5Å) for targets sharing >25% identity with known structures" (Tramontano 2003). According to SWISS-MODEL, a pioneering resource for protein folding prediction, the canonical kinase folding is in STK33 conserved.

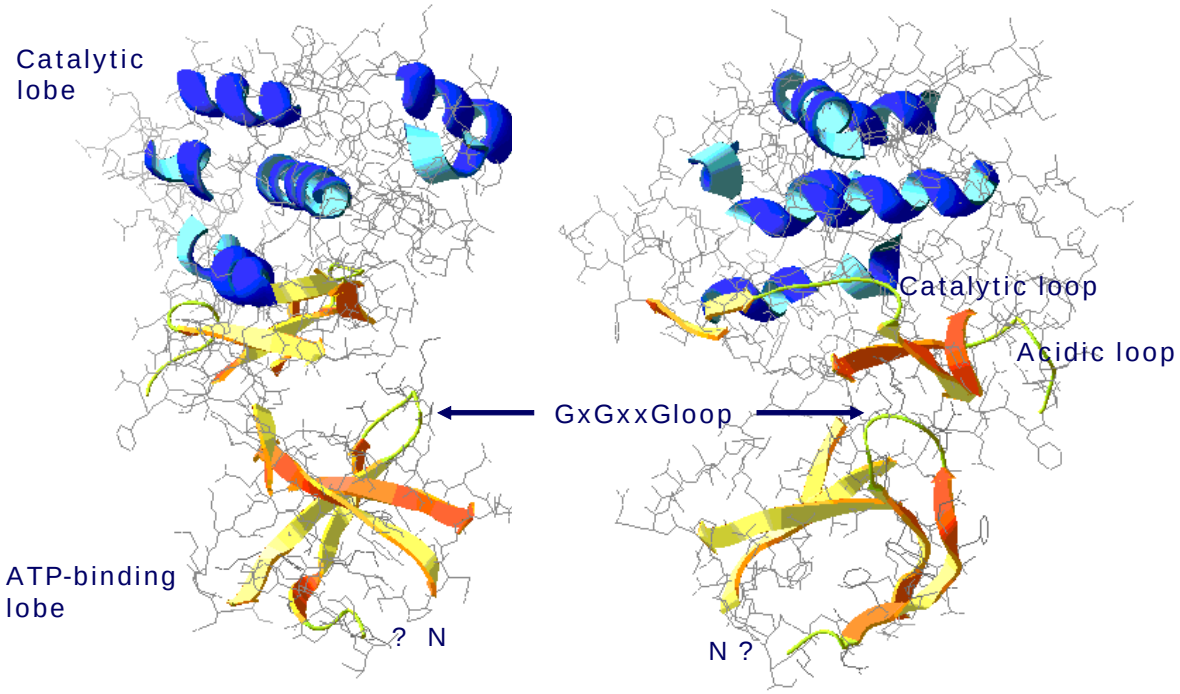


Figure 3.39: Two views of a SWISS-MODEL folding prediction for STK33 from the PredictProtein server.
The prediction is based on homology modelling by similarity with other protein kinases with already known three dimensional structures, hence only the catalytic domain of STK33 was taken in consideration. Some remarkable features are shown and Ns point to the N-terminal. A description of the STK33/Stk33 specific acidic loop is found in a coming section. Model was decorated and visualised with the Swiss PDB viewer (swissmodel.expasy.org)

a) Acidic loop

In the center of the catalytic domain of all STK33 orthologous proteins known to date, a very well conserved and distinctive stretch of highly polar acidic residues is observed, lying between the subdomains VIB and VII. This Asx-rich string is not identified as a particular domain by any resource analysis employed (PFAM, SMART, PROSITE). According to MAXHOM alignments (Sander and Schneider 1991) of hundreds of protein kinases, the alignment of all 510 human protein kinases catalogued by (Kostich et al. 2002) and the ClustalW-based alignment of all CAMK members from the human kinome, this string seems to be an specific attribute of the novel STK33 proteins.

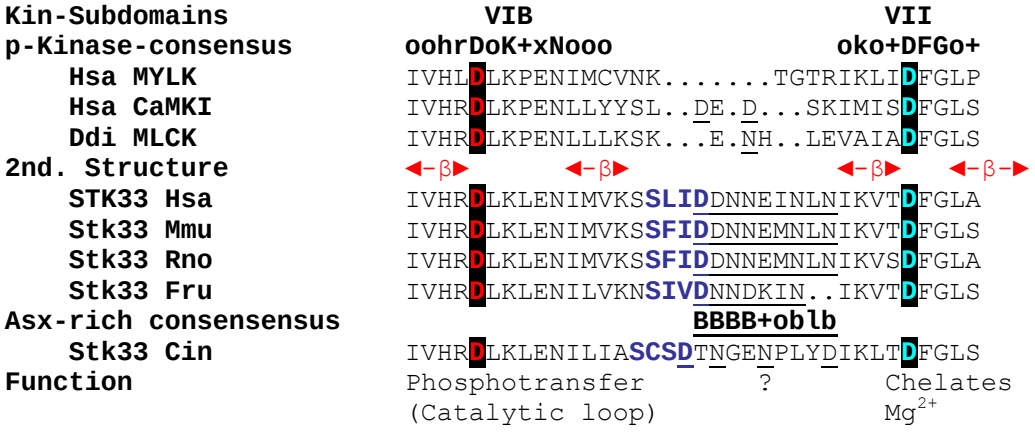


Figure 3.40: Alignment around the STK33-distinctive Asx-rich loop in the catalytic site.
Multiple alignment of STK33 and STK33-similar kinases in the region between the S/T Signature and chelating DFG motifs. Sequences from: human myosin light chain kinase (Hsa MYLK), human Ca^{2+}/calmodulin dependent kinase (HsaCaMKI), *Dictyostelium discoideum* myosin light chain kinase (DdiMLCK) and STK33 from human (Hsa), mouse (Mmu), rat (Rno), *Fugu rubripes* (Fru) and *Ciona intestinalis* (Cin). STK33 from *Ciona intestinalis* is shown separately to reflect the much higher conservation of the acidic loop in the mammals and fish entries. Catalyitic aspartate residues of canonical subdomains are shown in red and blue respectively. Novel Asx-rich STK33 distinctive loop and some similar residues of relative kinases are underlined. Conserved casein kinase II phosphorylation sites are shown in blue. Secondary structure ß-sheets are shown over the alignment in orange, they were predicted in STK33 reproducibly with all programs employed and find confirmed counterparts in the typical protein kinase folding. Protein kinase consensus and the STK33 acidic (Asx-rich) loop consensus residues are written in the single-character code under this general rules: **capitals**, invariant; **lowercase**, nearly invariant; **x**, any; **o**, non-polar; **+**, polar.

It contains no typical secondary structure building residues and extends the loop between the flanking β-sheets of domains VIB and VII, predicted in STK33 and already observed in the catalytic site of other protein kinases whose 3D structure is already established. The aspartic residue in the S/T signature is highly conserved in all kinases and is recognised as the catalytic base which intermediates the transfer of the γ-phosphate from ATP to the substrate. Very conserved as well, the aspartic residue from the DFG triplet, chelates activating Mg^{2+} ions placing the ATP molecule in the right orientation for the phosphate transfer (Hanks and Hunter 1995).

b) Outside the catalytic domain

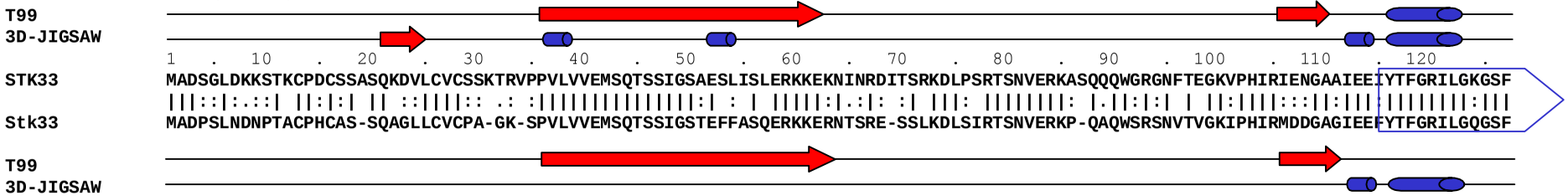
The STK33 sequence of was analysed with a "meta-server" (Ginalski et al., 2003), the new generation of protein folding prediction programs, in which the results from publicly available prediction servers are collected and compared and have been awarded in the last CASP5 experiment as effective for not only resolving the folding of proteins with sequence similarity by comparison, but also by recognizing similar folds from proteins with non-similar sequences (Tramontano 2003). These *in-silico* results, though still preliminary and requiring experimental confirmation, look promising. Despite the remarkably poor conservation of the STK33-C-terminal region between human and mouse, some topological issues are perhaps conserved. According to the 3D-JIGSAW analysis also based in homology modelling, at least three β-sheets may be formed in this region. These structures may correspond in number and size to the ones forming the C-terminal regulatory tails of some CAMKs which precisely result from the Meta-server comparison. This analysis reveals that several methods consistently align STK33 by topology with other kinases from

the CAMK group, suggesting a similar conformation for the C-terminal region and narrowing a putative Ca^{2+}/calmodulin-binding domain.

Figure 3.41: Amino acids alignment comparison of human STK33 and mouse Stk33 N-terminal and C-terminal regions outside the eukaryotic catalytic domain with a selection of results from the BIOINFO Meta-server.

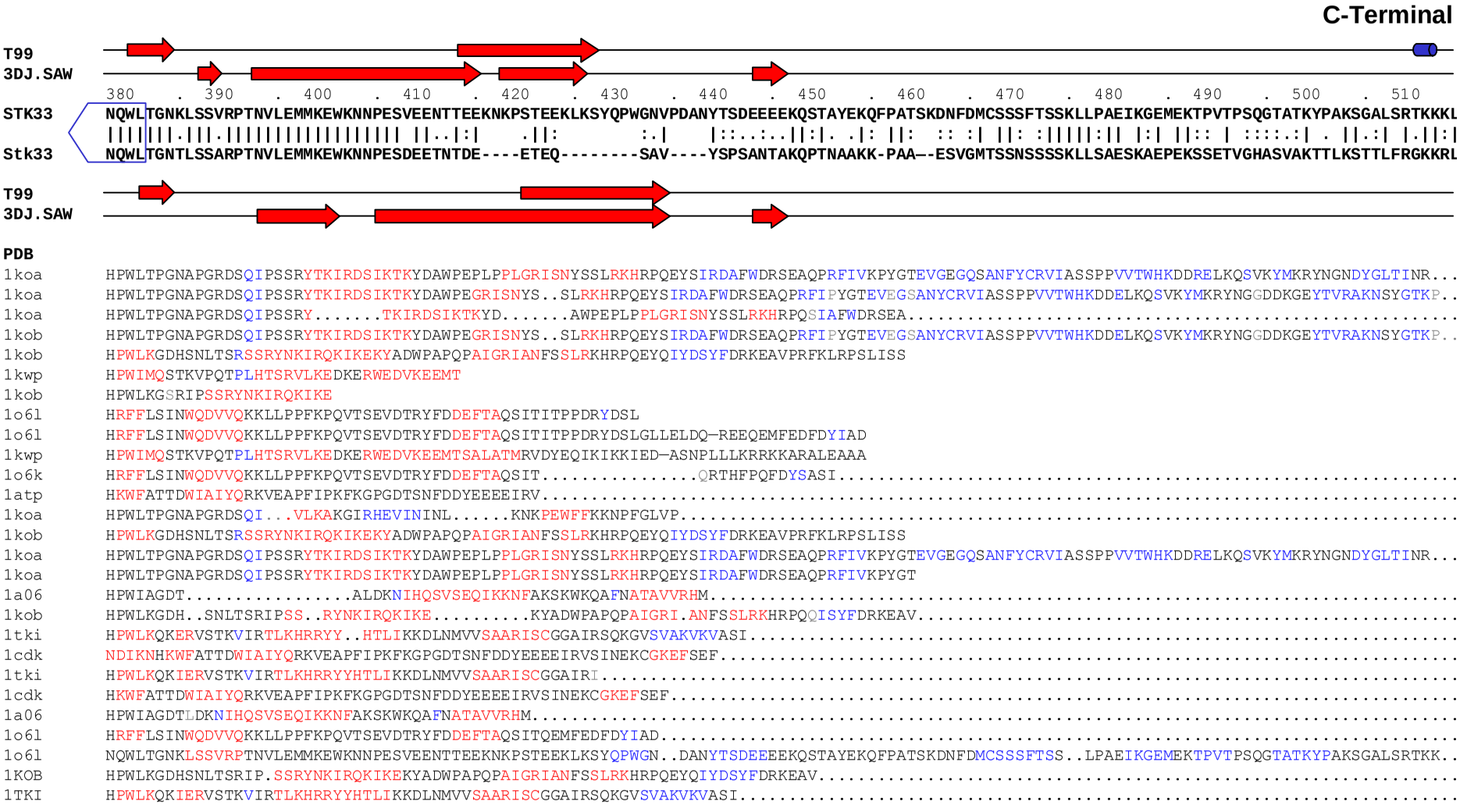
(In the next two pages) T99 and 3D-JIGSAW predictions of secondary structure are placed over the human and under mouse amino-acid sequence, red arrows depict ß-sheets, blue barrels depict α-helixes. Light blue block-arrows show the start and end point of the kinase catalytic domain. Several Meta server results for STK33 and Stk33. Alternative prediction are suppported by sequence conservation with protein with known 3D structures at the PDB (Protein Database). **1A06**, calcium-calmodulin dependent kinase from *Rattus norvegicus*; **1KOA**, twitchin kinase from *Caenorabditis elegans*; **1KOB**, twitchin kinase from *Aplysia californica*; **1KWP**, map kinase activated protein kinase 2 from human; **1O6L**, activated Akt/protein kinase B from human; **1ATP**, c-AMP-dependent protein kinase from mouse; **1TKI**, titin kinase from human heart; **1CDK**, cAMP-dependent protein kinase from *Sus scrofa* heart; **1GZN**, protein kinase Akt-2 from human; **2CPK**, c-AMP-dependent protein kinase from mouse; **1B6C**, type I TGF beta receptor from human; **1CMK**, myristylated cAMP-dependent protein kinase from human; **1APM**, c-AMP-Dependent Protein Kinase from mouse; **1FMK**, tyrosine-protein kinase SRC from human. Amino-acid residues in blue represent α-helices, in red represent β-sheets. Boxes expandding the secondary structures of some PDB entries (**1KOB**, **1AO6**, **1TKI**) signal the Ca^{2+}/calmodulin binding domain already confirmed in the literarture by structural data. bioinfo.pl/meta

N-Terminal



PDB

```
1KOA ..........................................................RKRRRGYDVDEQGKIVRGKGTVSSNYDNYVFDIWKQYYPQPVEIKHDHVLDHYDIHEELGTGAF
1KOA ..................................................INDYDKFYEDIWKKYVPQPVE.....................................VKQGSVYDYYDILEELGSGAF
1cmk ..........................................................................XGNAAAAKKGSEQESVKEFLAKAKEDFLKKWENPAQNTAHLDQFERIKTLGTGSF
1kob ..................................................................INDYDKFYEDIWKK......YVPQPVEVKQGSVYDYYDILEELGSGAF
1koa ..................................................................GYDVDEQGKIVRGKGTVSSNYDNYVFDIWKQYYPQPVEIKHDHVLDHYDIHEELGTGAF
1kob ..................................................................SKVRGKYDGPKINDYDKFYEDIWKKYVPQPVEVKQGSVYDYYDILEELGSGAF
1CMK ..................................................................................EQESVKEFLAKAKEDFLKKWENPLDQFERIKTLGTGSF
1KOA ..............................................RKRRRGYDVDEQGKRGKGTVSSNYDNYVFDIW-------KQYYPQPVEIKHDHVLDHYDIHEELGTGAF
1KOA .........PELCPACAHPKA...CCRCVANYTGNGRQCVDGNRTIGTGLVTDTPAMTEED.............GMPGSVAGVHYRANVQGW....IWKQYYPQPVEIKHDHVLDHYDIHEELGTGAF
1gzn ..................................................................................................KVTMNDFDYLKLLGKGTF
1cdk ..................................................................................................LDQFERIKTLGTGSF
1kob ..........................................................................SKVRGKYDGPKINDYDKFYEDIWKKYVPQPVEVKQGSVYDYYDILEELGSGAF
1KOA ..........................................................................GYDVDEQGKIVRGKGTVSSNYDNYVFDIWKQYYPQPVEIKHDHVLDHYDIHEELGTGAF
2cpk ..................................................................AAAAKKGSEQESVKEFLAKAKEDFLKKWENPAQNTDQFERIKTLGTGSF
1b6c ..............................................................DPSLDRPFISEGTTLKDLIYDMTTSGSGSGLPLLVQRTIARTIVLQESIGKGRF
1koa ..................................................................RKRRRGYDVDEQGKIVRGKGTVSSNYDNYVFDIWKQYYPQPVEIKHDHVLDHYDIHEELGTGAF
1cmk ..........................................................................XGNAAAAKKGSEQESVKEFLAKAKEDFLKKWENPAQNTAHLDQFERIKTLGTGSF
1kob ..................................................................INDYDKFYEDIWKK......YVPQPVEVKQGSVYDYYDILEELGSGAF
1koa ..................................................................................YDNYVFDIWKQYYPQPVEIKHD.HVLDHYDIHEELGTGAF
1apm_ ..................................................................................SEQESVKEFLAKAKEDFPSQNTAQL.DQFDRIKTLGTGSF
1cdk ..................................................................KGSEQESVKEFLAK........AKEDFLKKWENPAQNTAHLDQFERIKTLGTGSF
1cdk ..................................................................................KGSEQESVKEFLAKAKEDFLKKWENPLDQFERIKTLGTGSF
1fmk APDSIQAEEWYFGKITRRESERLLLNAENPRSETTKGAYCLSVSDFDNAKGLNVKHYKIRKLDSGGFYITSRTQFNSLQSKHADGLCHRLTTVCPTSKPQTQGLAKDAWEIPRESLRLEVKLGQGCF
1kob ..................................................INDYDKFYEDIWKKYVPQPVE.....................................VKQGSVYDYYDILEELGSGAF
```


C-Terminal
T99
3DJ.SAW
380 . 390 . 400 . 410 . 420 . 430 . 440 . 450 . 460 . 470 . 480 . 490 . 500 . 510
STK33 NQWLTGNKLSSVRPTNVLEMMKEWKNNPESVEENTTEEKNKPSTEEKLKSYQPWGNVPDANYTSDEEEEKQSTAYEKQFPATSKDNFDMCSSSFTSSKLLPAEIKGEMEKTPVTPSQGTATKYPAKSGALSRTKKKL
Stk33 NQWLTGNTLSSARPTNVLEMMKEWKNNPESDEETNTDE----ETEQ--------SAV----YSPSANTAKQPTNAAKK-PAA--ESVGMTSSNSSSSKLLSAESKAEPEKSSETVGHASVAKTTLKSTTLFRGKKRL
T99
3DJ.SAW
PDB
1koa HPWLTPGNAPGRDSQIPSSRYTKIRDSIKTKYDAWPEPLPPLGRISNYSSLRKHRPQEYSIRDAFWDRSEAQPRFIVKPYGTEVGEGQSANFYCRVIASSPPVVTWHKDDRELKQSVKYMKRYNGNDYGLTINR...
1koa HPWLTPGNAPGRDSQIPSSRYTKIRDSIKTKYDAWPEGRISNYS..SLRKHRPQEYSIRDAFWDRSEAQPRFIPYGTEVEGSANYCRVIASSPPVVTWHKDDELKQSVKYMKRYNGGDDKGEYTVRAKNSYGTKP..
1koa HPWLTPGNAPGRDSQIPSSRY.......TKIRDSIKTKYD.......AWPEPLPPLGRISNYSSLRKHRPQSIAFWDRSEA.....
1kob HPWLTPGNAPGRDSQIPSSRYTKIRDSIKTKYDAWPEGRISNYS..SLRKHRPQEYSIRDAFWDRSEAQPRFIPYGTEVEGSANYCRVIASSPPVVTWHKDDELKQSVKYMKRYNGGDDKGEYTVRAKNSYGTKP..
1kob HPWLKGDHSNLTSRSSRYNKIRQKIKEKYADWPAPQPAIGRIANFSSLRKHRPQEYQIYDSYFDRKEAVPRFKLRPSLISS
1kwp HPWIMQSTKVPQTPLHTSRVLKEDKERWEDVKEEMT
1kob HPWLKGSRIPSSRYNKIRQKIKE
1o6l HRFFLSINWQDVVQKKLLPPFKPQVTSEVDTRYFDDEFTAQSITITPPDRYDSL
1o6l HRFFLSINWQDVVQKKLLPPFKPQVTSEVDTRYFDDEFTAQSITITPPDRYDSLGLLELDQ—REEQEMFEDFDYIAD
1kwp HPWIMQSTKVPQTPLHTSRVLKEDKERWEDVKEEMTSALATMRVDYEQIKIKKIED—ASNPLLLKRRKKARALEAAA
1o6k HRFFLSINWQDVVQKKLLPPFKPQVTSEVDTRYFDDEFTAQSIT.....QRTHFPQFDYSASI.....
1atp HKWFATTDWIAIYQRKVEAPFIPKFKGPGDTSNFDDYEEEEIRV.....
1koa HPWLTPGNAPGRDSQI...VLKAKGIRHEVININL.....KNKPEWFFKKNPFGLVP.....
1kob HPWLKGDHSNLTSRSSRYNKIRQKIKEKYADWPAPQPAIGRIANFSSLRKHRPQEYQIYDSYFDRKEAVPRFKLRPSLISS
1koa HPWLTPGNAPGRDSQIPSSRYTKIRDSIKTKYDAWPEPLPPLGRISNYSSLRKHRPQEYSIRDAFWDRSEAQPRFIVKPYGTEVGEGQSANFYCRVIASSPPVVTWHKDDRELKQSVKYMKRYNGNDYGLTINR...
1koa HPWLTPGNAPGRDSQIPSSRYTKIRDSIKTKYDAWPEPLPPLGRISNYSSLRKHRPQEYSIRDAFWDRSEAQPRFIVKPYGT
1a06 HPWIAGDT.....ALDKNIHQSVSEQIKKNFAKSKWKQAFNATAVVRHM.....
1kob HPWLKGDH..SNLTSRIPSS..RYNKIRQKIKE.....KYADWPAPQPAIGRI.ANFSSLRKHRPQQISYFDRKEAV.....
1tki HPWLKQKERVSTKVIRTLKHRRYY..HTLIKKDLNMVVSAARISCGGAIRSQKGVSVAKVKVASI.....
1cdk NDIKNHKWFATTDWIAIYQRKVEAPFIPKFKGPGDTSNFDDYEEEEIRVSINEKCGKEFSEF.....
1tki HPWLKQKIERVSTKVIRTLKHRRYYHTLIKKDLNMVVSAARISCGGAIRI.....
1cdk HKWFATTDWIAIYQRKVEAPFIPKFKGPGDTSNFDDYEEEEIRVSINEKCGKEFSEF.....
1a06 HPWIAGDTLDKNIHQSVSEQIKKNFAKSKWKQAFNATAVVRHM.....
1o6l HRFFLSINWQDVVQKKLLPPFKPQVTSEVDTRYFDDEFTAQSITQEMFEDFDYIAD.....
1o6l NQWLTGNKLSSVRPTNVLEMMKEWKNNPESVEENTTEEKNKPSTEEKLKSYQPWGN..DANYTSDEEEEKQSTAYEKQFPATSKDNFDMCSSSFTSS..LPAEIKGEMEKTPVTPSQGTATKYPAKSGALSRTKK..
1KOB HPWLKGDHSNLTSRIP.SSRYNKIRQKIKEKYADWPAPQPAIGRIANFSSLRKHRPQEYQIYDSYFDRKEAV.....
1TKI HPWLKQKIERVSTKVIRTLKHRRYYHTLIKKDLNMVVSAARISCGGAIRSQKGVSVAKVKVASI.....

c) Subcellular localisation of STK33

According to their general features, STK33/Stk33 are likely soluble proteins. PSORT analysis (Horton and Nakai 1997) was employed to explore the protein sequence for signals that point to its sub-cellular localisation. This algorithm makes an extensive use of the signals already known in the literature that determine transport and localisation of the proteins into the different cellular compartments. This includes evaluation of the amino-acid composition, search for signal peptide sequences, transmembrane segments and their topologies, recognition of whether a protein exhibits transport signals to the mitochondria, nucleus, peroxisome, endoplasmic reticulum or vesicles and other miscellaneous analyses. No salient signal from the ones analysed is found STK33/Stk33 except from conserved di-lysine motifs (Thr-Lys-Lys-Lys in human and Gly-Lys-Lys-Arg in the mouse), which are recognised as putative endoplasmatic reticulum membrane retention signals (Teasdale and Jackson 1996). PSORT generates an overall report with the probabilities of the studied protein to belong to a given sub-cellular compartment.

Table 3.10: Subcellular localisation of STK33/Stk33 according to PSORT analysis

	Hsa STK33 Score in %	**Mmu Stk33**
Nuclear	56.5	39.1
Mitochondrial	21.7	17.4
Cytoplasmatic	17.4	13.0
Cytoskeletal	4.3	-
Golgi	-	13.0
Plasma membrane	-	8.7
Vesicles of secretory system	-	4.3
Extracellular, including cell wall	-	4.3

Together with PSORT, other programs SubLoc (Hua and Sun 2001), ProtCom (www.softberry.com) also reproducibly predict a nuclear localisation for both STK33/Stk33. However, there is no known NLS, nuclear localisation signal (Cokol et al. 2000) detected in the protein sequence of both proteins. These observations are not necessarily inconsistent, since proteins without NLS may be co-transported with another protein that has one (Hicks and Raikhel 1995). In particular for some protein kinases NLS-independent nuclear transport has been detected (Schmalz et al. 1998).

STK33/Stk33 does not exhibit transmembrane domains according to TMHMM (Krogh et al. 2001) which has been regarded as the best predictor of transmembranal helices by independent evaluators (Moller et al. 2001). This result gets confirmed by alternative analysis with PSORT, TMpred, and DAS.

3.3.4 Phylogenetical analysis

The similarity of STK33 to members of the CAMK group like MLCK, CamKI, PhK, detected by BLASTP searches, first suggested it may belong to this group of protein kinases. To test this hypothesis, maximum parsimony trees using the PAUP software package (Swofford 1991) were calculated based on the catalytic domain of representative members from all other groups of kinases, using the same parameters and criteria Hanks and Hunter employed for their phylogenetical analysis (Hanks and Hunter 1995). Reproducibly, all trees clustered STK33 in the CAMK kinase group with moderately high bootstrapping support. Figure 3.42 shows an example of these trees.

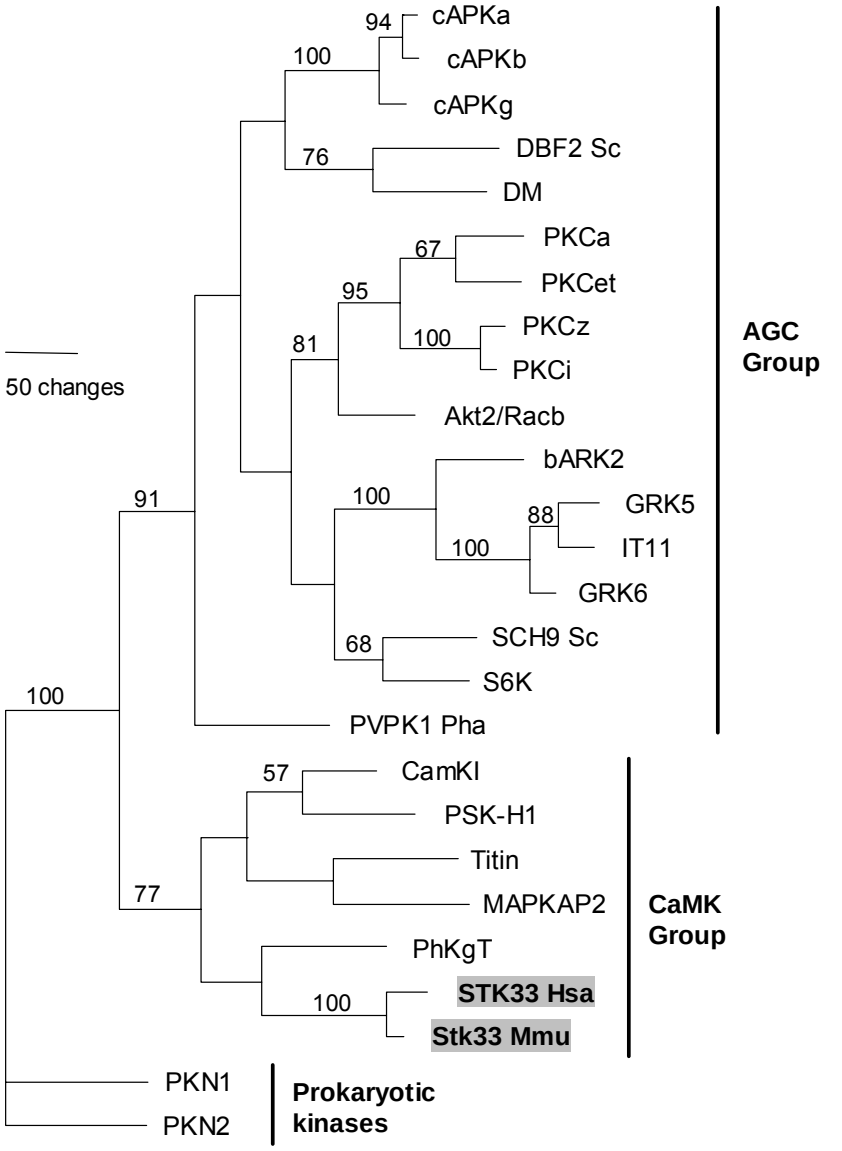


Figure 3.42: Example minimum-length parsimony tree based on a mul-tiple alignment of the protein kinase catalytic domain.
STK33 and mouse Stk33 (here shaded) cluster within the CaMK group of protein kinases. For comparison, members of the AGC Group were added. Two prokaryotic proteins were used as outgroup. Bootstrap values (100 repetitions) are placed above the corresponding branches. The tree is based on the alignment of the catalytic domain sequences available at the Protein Kinase Resource (Smith 1999) and the names of the taxa were taken accordingly (Hanks and Hunter 1995).
Full names and accession numbers are as follows: **cAPKa** (P17612) human camp-dependent protein kinase, alpha-catalytic subunit; **cAPKb** (P22694) human camp-dependent protein kinase, beta-catalytic subunit; **cAPKg** (P22612) human camp-dependent protein kinase, gamma-catalytic subunit; **DBF2 Sc** (P22204) *Saccharomyces cerevisiae* cell cycle protein kinase; **DM** (B49364) human protein kinase (EC 2.7.1.37), myotonic dystrophy-associated ; **PKCa** (P17252) human protein kinase C, alpha type; **PKCet** (P24723) human protein kinase C, eta type; **PKCz** (Q05513) human protein kinase C, zeta type; **PKCi** (P41743) human protein kinase c, iota type; **Akt2/Racb** (P31751) human rac-beta serine/threonine protein kinase (rac-pk-beta) (protein kinase akt-2) (protein kinase b, beta) (pKb beta); **bARK2** (P35626) human beta-adrenergic receptor kinase 2 (beta-ark-2) (g-protein coupled receptor kinase 3); **GRK5** (P34947) human G protein-coupled receptor kinase 5; **IT11** (P32298) human G protein-coupled receptor kinase 4 (GRK4); **GRK6** (P43250) human G protein-coupled receptor kinase; **SCH9 Sc** (P11792) *Saccharomyces cerevisiae* camp-dependent protein kinase; **S6K** (P23443) human ribosomal protein s6 kinase (p70-s6k); **PVPK1 Pha** (P15792) *Phaseolus vulgaris* protein kinase; **CamKI** (XP_002911) human calcium/calmodulin-dependent protein kinase I; **PSK-H1** (I38138) protein-serine kinase (EC 2.7.1.-) PSK-H1; **Titin** (S20898) human titin; **MAPKAP2** (S39793) human MAPK-activated protein kinase 2; **PhKgT** (P15735) human phosphorylase b kinase gamma catalytic chain, testis/liver isoform (phk-gamma-t) (psk-c3) (phosphorylase kinase gamma subunit); **STK33 hum** (CAC29064) human serine/threonine kinase 33; **STK33 ms** (CAC39171) mouse serine/threonine kinase 33; **PKN1** (P33973) *Myxococcus xanthus* serine/threonine-protein kinase; **PKN2** (S21533) *Myxococcus xanthus* protein kinase (Mujica et al. 2001).

Another interesting question is the phylogenetical origin of STK33. Orthologous of STK33 are detectable only in the genomes of vertebrates and urochordates, showing a strikingly conservation of the catalytic domain, whereas, no orthologous genes were found in *Sacharomyces cereviseae*, *Caenorhabditis elegans*, and *Drosophila melanogaster*.

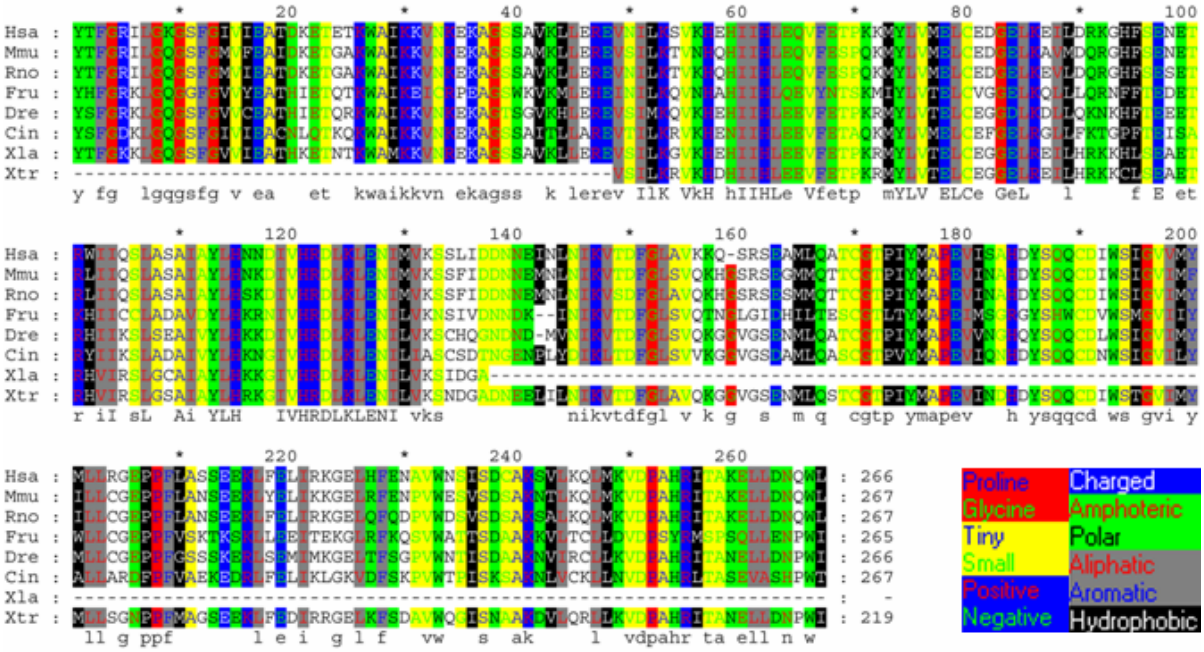


Figure 3.43: Alignment of the catalytic domains of STK33 in several organisms.
Hsa: *Homo sapiens*, **Mmu**: *Mus musculus*, **Rno**: *Rattus norvegicus*, **Fru**: *Fugu rubripes*, **Dre**: *Danio rerio*, **Cin**: *Ciona intestinalis*. **Xla**: *Xenopus laevis*, **Xtr**: *Xenopus tropicalis*. *Xenopus* sequences derived out of incomplete EST entries. The alignment was obtained with ClustalW and decorated physiochemically with GeneDoc according to the legend at the right-lower corner.

4 Discussion

In this work the human-mouse comparative sequencing analysis of a genomic region flanked by genes *C11orf29/D7h11orf29* and *LMO1/Lmo1* is described. In following sections, the genomic organisation of the region under study is evaluated in comparison with both neighbouring regions presented in previous works in our institute and with the respective whole genome projects already available in the literature and databases. Finally, some aspects of the preliminary analysis of the *STK33/Stk33* genes found in these regions and their protein product are discussed in detail.

4.1 Genomic organisation

The whole genomic region studied in our institute spans roughly 730 kilobases in human chromosome 11p15.3 and the corresponding 620 kilobases-long syntenic segment in mouse chromosome 7E3, between the homologues gene pairs *CEPG1/Cepg1* and *STK33/Stk33* as depicted in the following figure.

A total of 11 protein-coding genes were found in human and 10 in the mouse (Amid et al. 2001), which corresponds, as shown in table 4.1, to a slightly higher gene-density than observed in the whole genomes. On the other hand, the analysed lengths are slightly longer than the minimum genome from the prokaryote model *Mycoplasma genitalium* (580

kilobases) with around 500 genes (Fraser et al. 1995) and hence a 50-times higher gene density. This reflects the fundamental difference in compaction between prokaryote and eukaryote genomes.

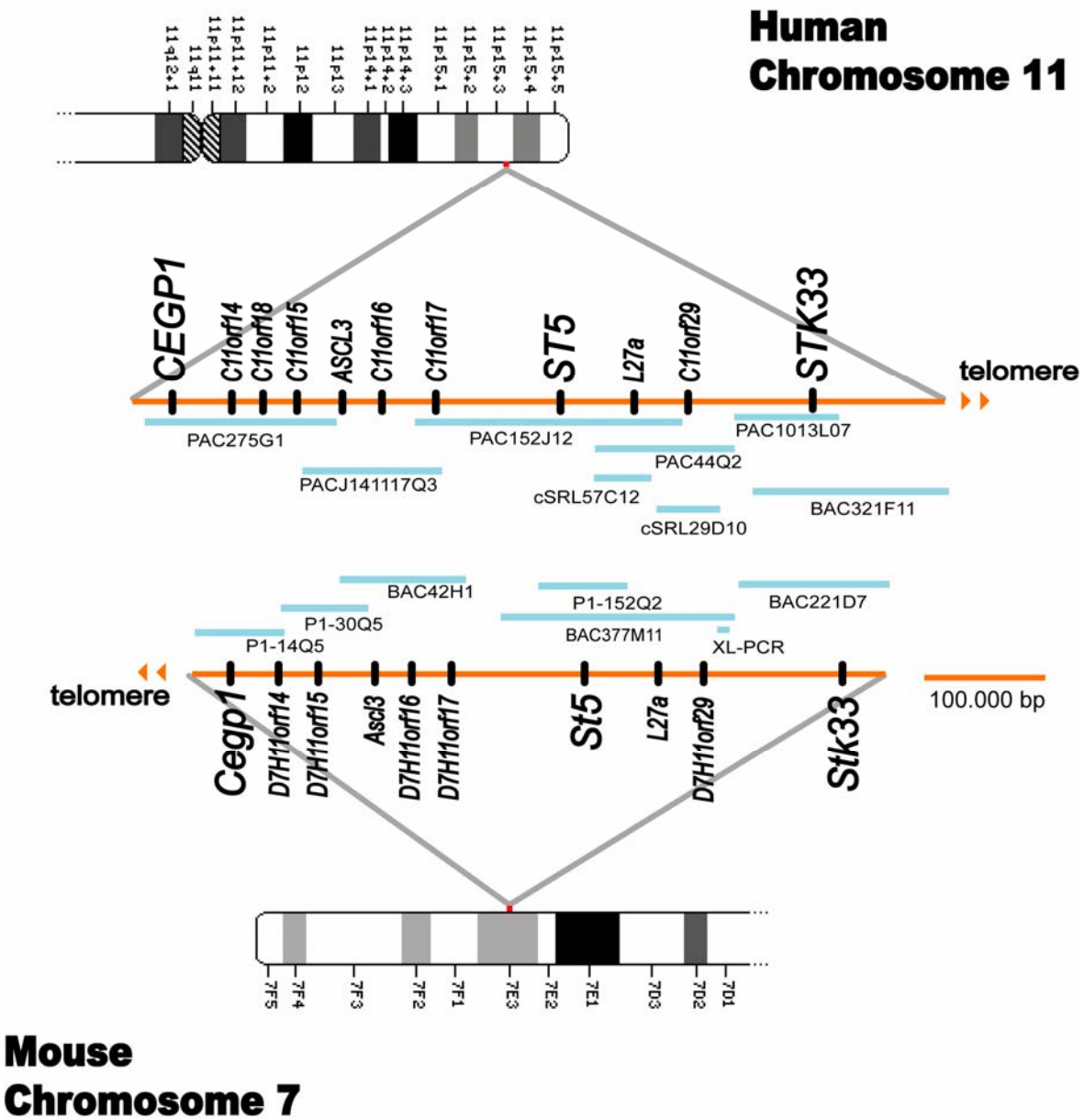


Figure 4.1: Clone-contig map of the region analysed by the Institute for Molecular Genetics in Mainz University.

Diagram of the short arm of human chromosome 11 is shown at the top showing the relative positions of PAC, BAC and Cosmid clones between *CEPG1* and *STK33*. The homologous chromosome fragment of mouse chromosome 7 is shown below. The directions to the respective telomeres in each chromosome are shown. The flanking regions, not shown here, were analysed at the Children`s Hospital of Mainz University. (Amid et al. 2001; Cichutek et al. 2001) and NCBI's MapViewer www.ncbi.nlm.nih.gov/mapview/

Table 4.1: Gene density and repeats percentage from several genomes of model organisms compared with the area under study in our institute

	Genes*	Base pairs	Ensembl release	Gene per Mb	% intersp. repeats	Source†
Myc. genitalium	484	580,074	n.a.	834,4	n.a.	Fraser
Sac. cerevisiae	6,335	13,478,000	n.a.	470.0	3.4	Mewes
Drosophila	18,289	128,343,463	20.3a.1	142.5	12	FlyBase
Fugu rubripes	38,510	329,140,338	20.2b.1	117.0	3	Aparicio
Rat	30,232	2,571,104,688	20.3b.1	11.8	40	Gibbs
Mouse	34,076	3,167,465,022	20.32b.1	10.8	39	Waterston
Human	31,609	3,201,762,515	20.34c.1	9.9	46	Waterston
Mouse Mz Molgen	**10**	**620,000**	**n.a.**	**16.1**	**30**	**Amid**
Human Mz Molgen	**11**	**730,000**	**n.a.**	**15.1**	**44**	**Amid**

*Number of genes and total base pairs according to Ensembl (www.ensembl.org).
†(Fraser et al. 1995); (Mewes et al. 1997) (mips.gsf.de/proj/yeast/tables/inventy.html); flybase.org/; (Amid et al. 2001; Aparicio et al. 2002; Gibbs et al. 2004; Waterston et al. 2002).

The majority of prokaryotic genome corresponds to protein coding sequences (89% in *E. coli*), whereas the contrary (~2% in human) is the case for eukaryotes (Alberts 2002). Prokaryotic genomes are compact because they have virtually no interspersed repeats, their genes are arranged very closely to each other with very short intergenic regions in-between and they lack introns. In line with their higher complexity, eukaryotic genomes have much more genes than prokaryotic ones, and they are also several orders of magnitude bigger in terms of base pairs. Among eukaryotes there is some correlation between genome size and organism complexity, but remarkable exceptions occur. Rice contents ca. 20,000 more protein coding genes than humans, but its genome is 6.7 times smaller (Yu et al. 2002). The genome of some unicellular eukaryotes like the *Amoeba* is up to 200 times bigger than the human genome (Alberts 2002) (*Amoeba dubia:* 670 billion bases, according to **DOGS**, the Database Of Genome Sizes www.cbs.dtu.dk/databases/DOGS/). This discrepancy between genome size and complexity of an organism has been termed the C-value paradox (Mirsky and Ris 1951).

Eukaryotic genomes show a mosaic pattern of global distribution of gene density, (G+C) content, CpG islands and repeats content (Koop 1995; Zoubak et al. 1996). This mosaic pattern has been confirmed at the base pair level in the numerous whole genome sequencing publications (1998; Gibbs et al. 2004; Lander et al. 2001; Rubin et al. 2000; Venter et al. 2001; Waterston et al. 2002) and has been found to be consistent with the *"random breakage"* model from Nadeau and Taylor (Nadeau and Taylor 1984). According to this model, the current landscape of eukaryotic genomes reflects the multiple chromosomal rearrangements that have been occurred during evolution.

In some extent and in spite of their short size, the genomic regions between genes *CEPG1* and *STK33* and its syntenic one in the mouse exhibit this mosaic pattern, as discussed below.

4.1.1 (G+C) content and CpG islands

With a mean of 43% in human and 44% in mouse, the whole genomic region under study in our institute, has a slightly higher (G+C) content than the average of the respective whole genomes: 41% human and 42% in the mouse (Lander et al. 2001; Waterston et al. 2002). This is not unusual, since along a given genome even larger fluctuations of the (G+C) content are well known. The draft mouse genome sequence has proved to have fewer of this fluctuations (see figure 4.2) and that even at the chromosomal level, the (G+C) content tends to be more uniform than in humans (Waterston et al. 2002).

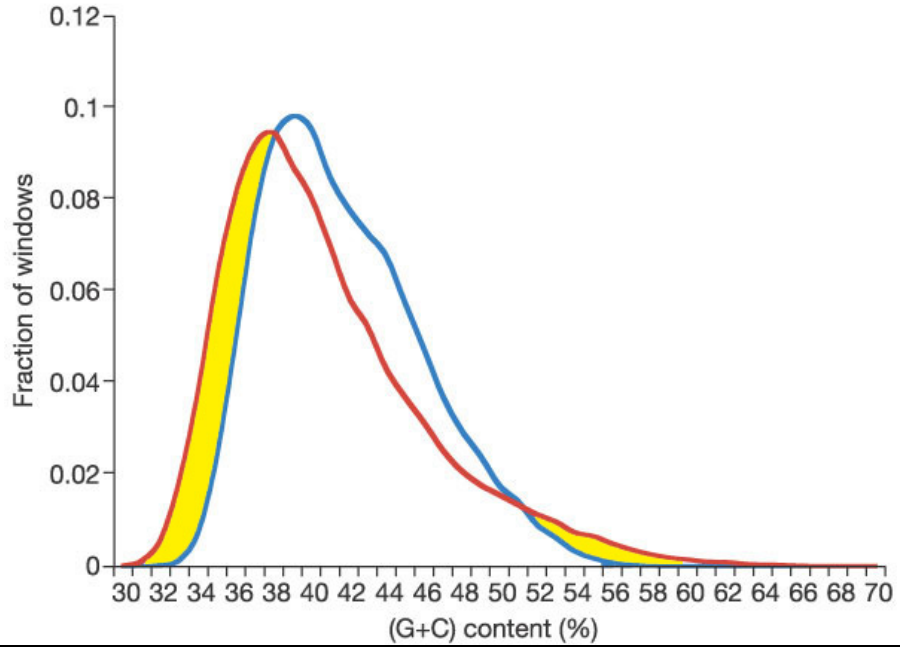


Figure 4.2: Distribution of (G+C) content in the mouse (blue) and human (red) genomes.
The distribution was determined using windows of 20-Kb. Mouse has a higher mean (G+C) content than human (42% vs. 41%) but human has a larger fraction of windows (shadowed here in yellow) with high or low (G+C) content. (Waterston et al. 2002).

The genomic regions analysed in our institute well represent the fluctuations in (G+C) content in the human and mouse genomes, as it may be observed in the figure 4.3. The overall (G+C) content of murine clone BAC221D7, which contains all but one of *Stk33* exons is 40%. The (G+C) content drops sharply 6% in the human between *C11orf29* and *STK33* and 5% in the mouse between their homologues *D7h11orf29* and *Stk33* (Amid et al. 2001), to rise again 9% in the human between *STK33* and *LMO1* and 7% in the mouse between *Stk33* and *Lmo1* (Cichutek et al. 2001). Much larger shifts are observable in the human genome, where in regions of less than 300 Kb a difference of up to 33% in (G+C) is observable (Lander et al. 2001).

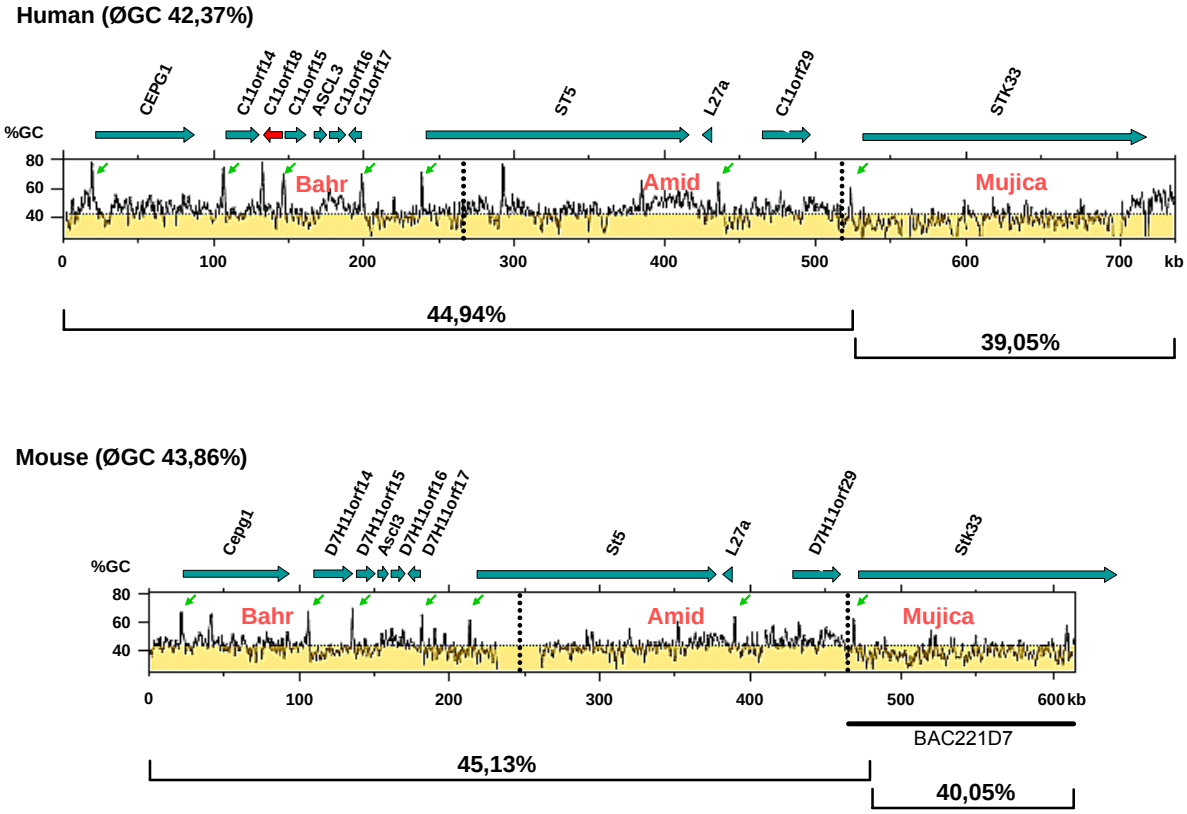


Figure 4.3: (G+C)-plot of BAC221D7 in context.
Arrows show the relative positions and directions of putative transcripts described in the doctoral theses from A. Bahr and C. Amid, and the present work. Green arrows point to the CpG islands very likely associated with the promotor structure of most of the genes found in the region. Modified from (Amid et al. 2001).

STK33/Stk33 (A+T)-rich sub-regions may correspond to the previously postulated *isochores* which are defined as regions of "homogeneous" (G+C) content. Five types of isochores were originally proposed according to low (L) or high (H) (G+C) content: **L1** <38%, **L2** 38-42%, **H1** 42-47%, **H2** 47-52% and **H3** >52% (Cuny et al. 1981). The region between genes *CEPG/Cepg1* and *C11orf29/D7h11orf29* would correspond to the H1 isochore type, the region of *STK33/Stk33* to the L2 type and the region around *LMO1/Lmo1* and *TUB/Tub* to the H2 type (Cichutek et al. 2001).

In the draft version of the human genome, Lander and colleagues (2001), questioned the concept of isochores. They found substantial variation in (G+C) content within different regions of the human genome and at different spatial scales, but the term *"long-range variation in (G+C) content"* was preferred, since the central definition of homogeneity in the isochore's model is inconsistent with the binomial test, the statistical assessment they used (Lander et al. 2001). However, recently an alternative statistical method, based on analysis of variance, confirm the isochores as "fairly homogeneous" as they were actually described originally for more than 20 years (Li et al. 2003), instead of "strictly homogeneous" as they were interpreted in the human genome draft paper. In his response (Bernardi 2001), warned that the definitions of strict isochores or classic isochores in recent literature, including the draft of the human genome, may be misleading.

Using DNA microarray data from the Gene Expression Atlas (Su et al. 2002) from 7,708 human and 6,078 mouse well annotated genes, Vinogradov confirmed the observation that housekeeping genes are in general slightly (G+C) richer than genes expressed in a tissue-specific manner (Bernardi 1995; Vinogradov 2003). In particular, Vinogradov found in his study a tendency to relatively low (G+C) content in genes specific to germ-line tissues (ovary and testis). Results on the (G+C) content of *STK33* and its expression pattern discussed in this work are in agreement with these observations.

Regions of very high (>60%) content of (C+G), so called CpG islands, have been associated with the promotor structure of eukaryotic genes, particularly with those showing differential expression. CpG islands may be target of cytosine-methylation and play an important role in the regulation of the associated genes. The *STK33/Stk33* genes are

associated with CpG islands conserved in human and mouse. This is the case for 14 of the 21 genes analysed by our group.

4.1.2 Repeat content

Repeat content is one major reason for the C-value paradox. At least 46% of the human genome consists of repeats, whereas less than 5% consists of coding sequence. A comparison of the repeat content in the analysed regions with the whole genomes shows a significant difference. In both human and mouse, the percentage of interspersed repeats of the genomic region under study is higher than in the whole genome, but in both organisms SINEs repeats are at least 50% under-represented and LINEs are up to three times over-represented as illustrated in the following table.

Table 4.2: Content of repetitive elements in the genomic region around *STK33/Stk33*

	Human		**Murine**	
	Whole genome*	***STK33***	**Whole Genome***	***Stk33***
SINEs	20.99	10.31	19.20	8.62
LINEs	13.64	37.56	8.22	23.90
LTR elements	8.55	4.80	9.87	5.79
DNA elements	3.03	4.45	0.88	0.81
Unclassified	0.15	0.74	0.38	0.91
Total interspersed	**46.36**	**57.86**	**38.55**	**40.03**
Simple repeats	0.87	0.52	2.27	2.25

*Taken from Waterston et al., 2002, table 5.

These differences correlate with the low (G+C) content in the region. As for back as 1982, Meunier-Rotival and colleagues observed a direct correlation between (G+C) content and the density of certain types of repeats (Meunier-Rotival et al. 1982). LINE-repeats are more frequent in (A+T)-rich regions, whereas the SINE-repeats are more frequent in (G+C)-

rich regions. This correlation is demonstrated in the data of the draft human genome shown in the figure 4.4 and is also observed in our analysed region as shown in table 4.3.

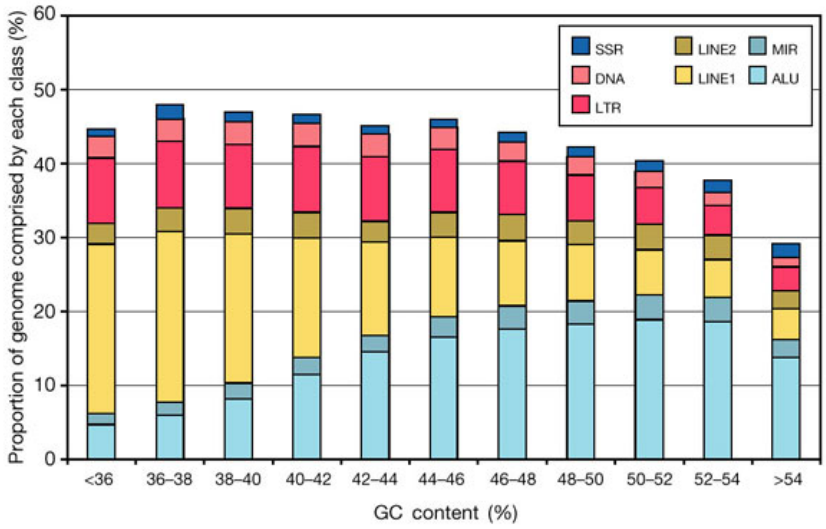


Figure 4.4: Distribution of repeats relative to (G+C) content in the human genome. (Lander et al. 2001).

Table 4.3: Content of repetitive elements and (G+C) content in the sequence under study in our institute

	Around *CEPG1*		Around *ST5*		Around *STK33*	
	Human	**Mouse**	**Human**	**Mouse**	**Human**	**Mouse**
SINEs	19.17	18.30	17.21	12.64	10.31	8.62
LINEs	16.72	5.92	10.95	3.90	37.56	23.90
Others	2.45	7.15	6.06	3.49	9.99	7.51
T. interspersed	**38.34**	**31.37**	**34.22**	**20.03**	**57.86**	**40.03**
(G+C) Content	**44.05**	**44.54**	**45.99**	**45.83**	**38.56**	**40.89**

4.1.3 Genomic sequence conservation

As shown in the figure 3.10 in the results section, DotPlot diagonals suggest sequence conservation around the coding exons of *STK33* and *Stk33* but the conservation in

intergenic regions is remarkably less evident than in the neighbour regions analysed in previous works (Amid 2002; Bahr 1999) as shown in the following figure.

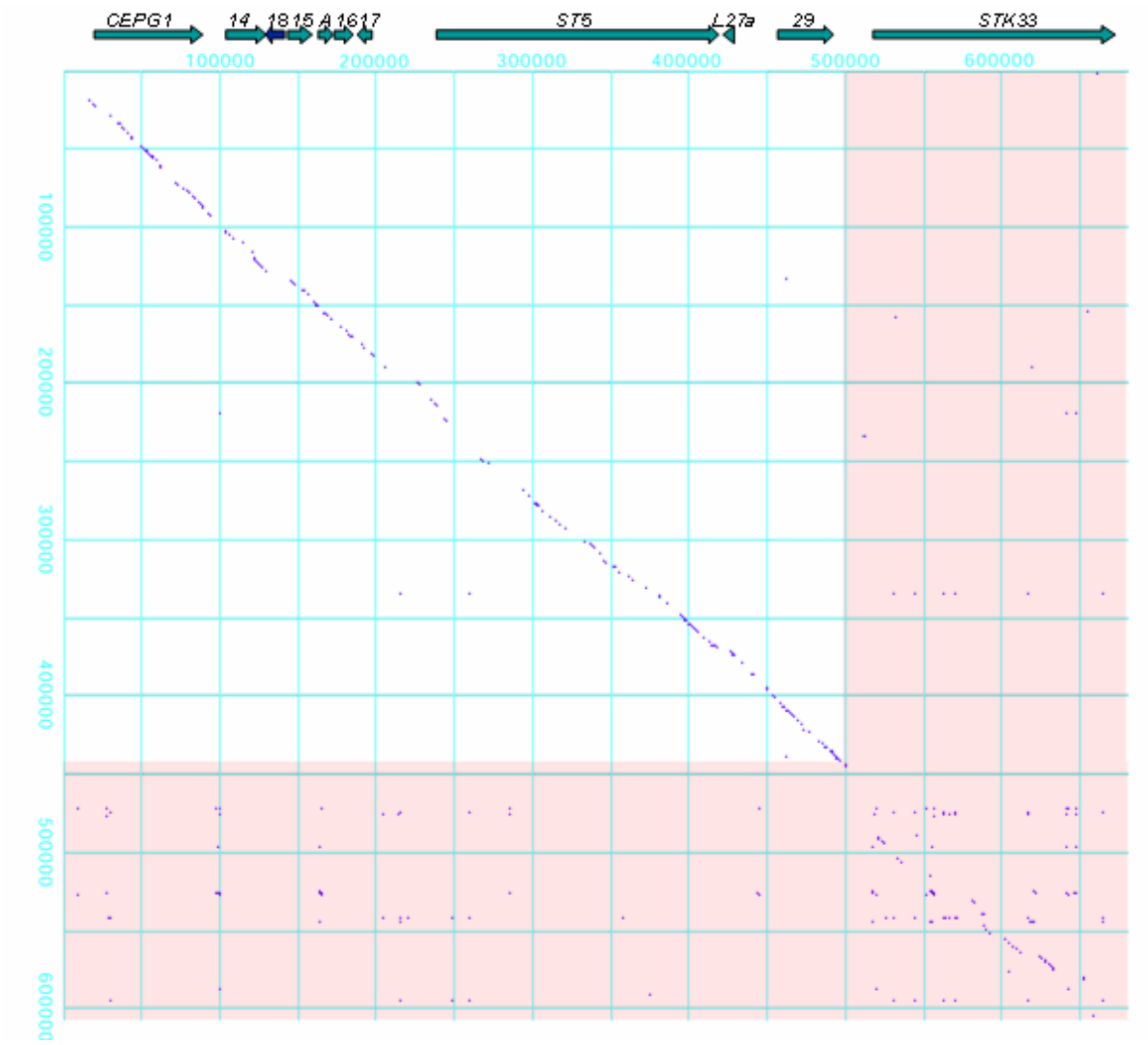


Figure 4.5: DotPlot of the genomic regions of human and mouse sequenced in our lab.
Arrows show the relative positions and directions of putative transcripts described in the doctoral theses of A. Bahr, C. Amid and the present work. Genes with numbers correspond to ORF named according to HUGO nomenclature rules for genes of unknown function (14: ***C11orf14***, etc.) and A: ***ASCL3***. The plot was generated with the program MegaAlign (DNAStar) with the following parameters: Percentage: 60, Window 50; Min. Quality: 50.

4.1.4 Synteny and genomic mosaic pattern

By comparing the whole sequence of the human chromosome 19 with homologous regions of the mouse, Dehal and colleagues (2001) were able to identify 128 novel previously unannotated human genes. Furthermore they found out that genes and even entire clusters of genes for olfactory receptors may have undergone duplication in the mouse or even more likely loss in the human, which is in line with our poor sense of smelling relative to other mammals. They also noted primate-specific gene loss of pheromone receptors. Through the exact characterisation of the boundaries of the syntenic clusters, they provided insight into the chromosome evolution in mammals: most of the boundaries observed were associated with gene family expansion. The role of recombination between repeats on chromosomal rearrangements was also clearly observed, through the presence of tandem organised L1 elements and retrovirus-associated sequences at these syntenic boundaries.

In their comparison of mouse chromosome 16 with its homologous fragments in human chromosomes, Mural and colleagues (2002) compared the syntenic boundaries with the sharp transitions in distribution of gene density, (G+C) content, CpG islands and repeats content. In some cases a strong correlation was observable suggesting that these boundaries signal the positions of the chromosomal rearrangements occurred by the divergence of the primate and murine lineages. On the other hand, those cases where no correlation is detectable suggest that the syntenic block resembles the ancestral pattern of rearrangement prior to the mouse-primate divergence.

Altogether, the strikingly lack of human-mouse sequence conservation, the drop in (G+C) content and the difference in repeat content compared with the prior studies in the

region, let suggest that *STK33/Stk33* constitutes a chromosomal fragment whose rearrangement (i.e. translocation) occurred before the rodent-primate radiation i.e. before the rearrangement that resulted in the whole synteny block of chromosome 11p14.3 - 11p15.5 in human and the distal region of mouse chromosome 7 and rat chromosome 1 took place (See figure 4.6). To test this hypothesis, comparative genomic analysis must be performed using the region surrounding *STK33* in murine and human with those of vertebrates, such as dog, fish, frog and chicken, but also with those of urochordates like *Ciona* and more generally, deuterostomes, for example *Amphioxus*. Such a comparison could eventually help to further elucidate the phylogeny of *STK33* among the super-family of protein kinases.

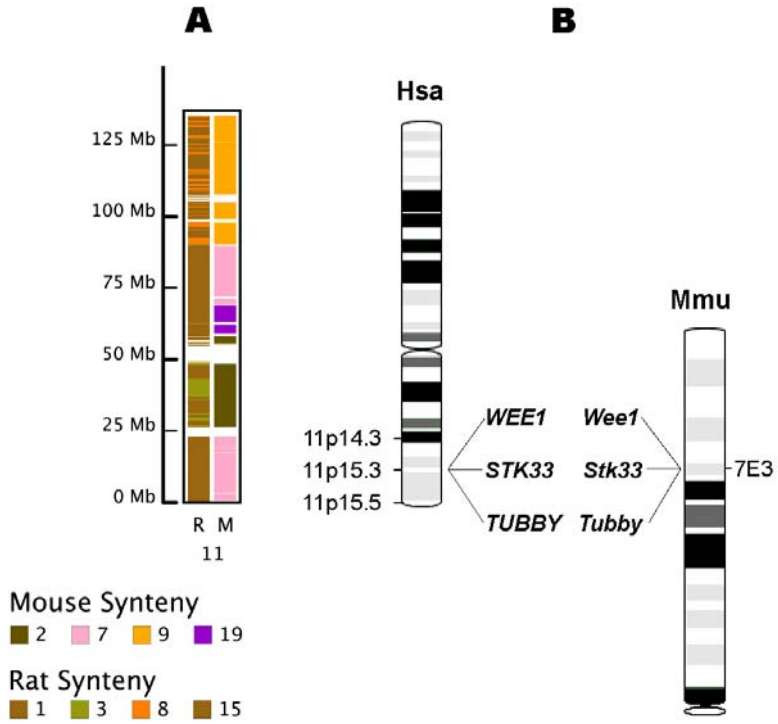


Figure 4.6: Human, mouse and rat synteny relative to human chromosome 11.
A: Synteny blocks of human chromosome 11 in mouse and rat according to the colour legend in the bottom-left corner. **B.** Human chromosome 11 and mouse chromosome 7 aligned around the syntenic block between 11p14.3 and 11p15.5 regions. Please note that chromosomes are represented with the centromeres pointing down. Modified from www.genboree.org (Kalafus et al. 2004).

4.3 Genomic annotation

The difficulties of using de-novo gene-prediction programs have already been thoroughly discussed and assessed (Amid 2002; Bahr 1999; Frazer et al. 2003). In the present study, none of the gene-prediction programs used (e.g. GENSCAN, GRAIL) succeeded in predicting the *Stk33* gene within the murine genomic sequence. Remarkably, the core exons coding for the kinase domain were separated in two predicted transcripts, excluding the chance of prediction of any active kinase gene. Some real exons were also missed in the de-novo predictions. Furthermore, the sequencing of the *STK33* cDNA revealed some obvious single-pass sequencing errors of the EST, but also a deletion of one base in the coding exon 9 of the sequences contained in the public databases (Acc. No. AC016718). This point is critical, since the resulting frame shift would make the prediction of a correct ORF nearly impossible.

The discovery of *STK33/Stk33* constitutes a good example of the strength of the comparative sequencing approach even by using incomplete public data from the human genome. The gene has received RefSeq status (NM_030906, XM_110633 (Pruitt et al. 2003), is listed by UniGene (*Hs.148135, Mm.79075* (Wheeler et al. 2003) and MapView as a confirmed gene model (Figure 4.7).

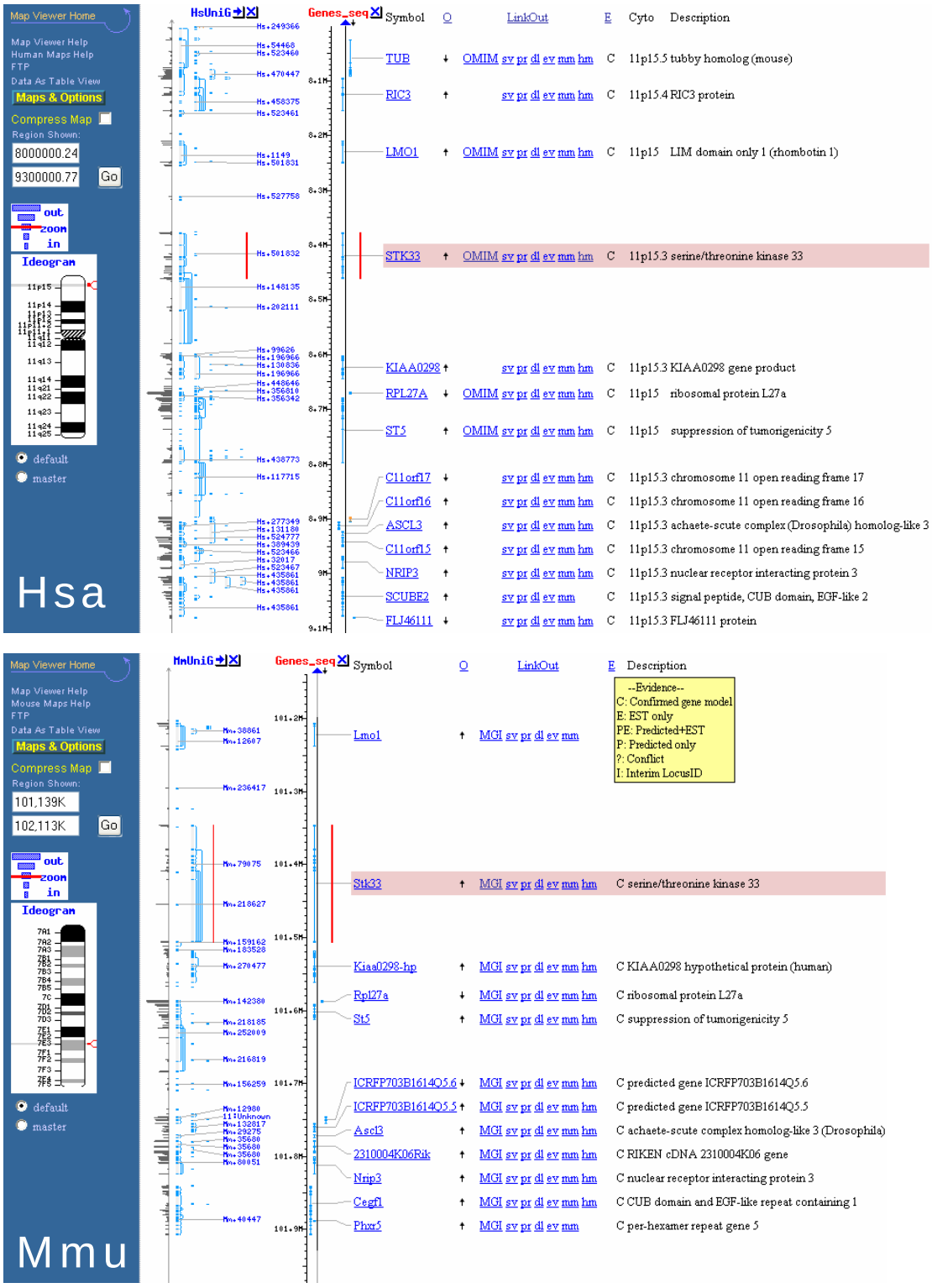


Figure 4.7: ***STK33/Stk33* in genome MapViewer from NCBI**
NCBI's Genome Map Viewer (www.ncbi.nlm.nih.gov/mapview/), showing human *STK33* on the top and murine *Stk33* on the bottom.

Gene order is also a critical annotation result which relies on a correct computer assembly of the sequences and obviously on the completeness of the data. In the draft version of the human genome, the relative orientation of genes the *TUB*, *RANBP7* and *WEE1* are correct, but *ST5*, *RPL27A* and *LMO1* are inverted as shown in the figure 4.8. The right orientation centromere e -*WEE1-RANBP7-ST5-RPL27A-LMO1-TUB*-telomere in human was established in our works (Amid et al. 2001; Cichutek et al. 2001), with a combination of chromosomal in-situ hybridisation and sequencing of the whole region. The wrong order of genes in the human genome is due the presence of gaps in the assembly of the first draft version.

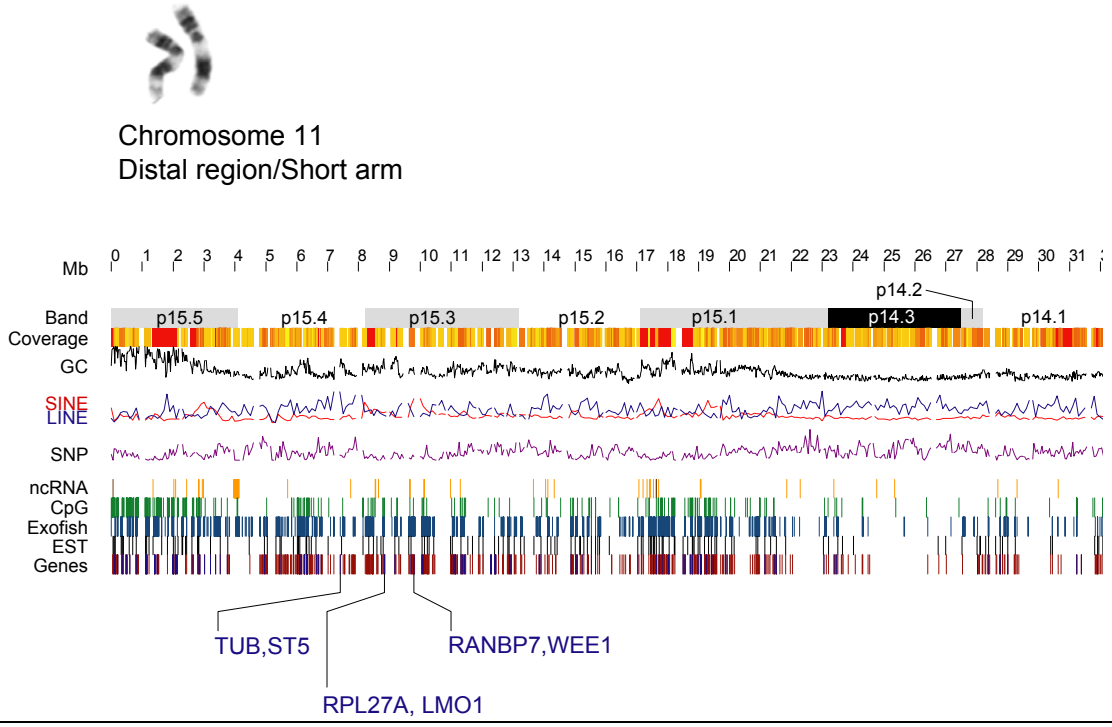


Figure 4.8: **Major genomic features from human chromosome 11 (distal region of the short arm) as published in the draft of the human genome and relative positions of the gene region sequenced in Mainz.**
Note the correlation of gene density with (G+C) content in region p14.3. Modified from (Lander et al. 2001)

4.3.1 IS10

The IS*10*-R bacterial mobile element sequence present in the murine bacterial clone BAC221D7 is most likely explicable through a transposition event during the BAC library preparation. Such entries, wrongly included within eukaryotic sequences, constitute a serious database contamination that may cause problems to further investigators. Actually, a databases screen with the sequence of IS*10*-R showed several matches of eukaryote origin. From the twenty best matches, sixteen were from human origin and only seven of the total were correctly annotated as being mobile elements from prokaryote origin. Indeed, it has been found that IS10 is one of the most frequent contaminant of DNA databases with a not negligible frequency of up to one contamination every 1,000 clones (Kovarik et al. 2001).

The presence of these well-known contaminant sequences from prokaryote origin in eukaryote genomic databases has been used to check the accuracy of sequencing projects. Hill and colleagues (Hill et al. 2000) used the presence of the IS*10*-R element in eukaryotic genomic sequences in the databases to confirm that in general the large-scale sequencing projects reach and even surpass the goal of less than one sequencing error in 10,000 bp.

The kind of PCR and re-sequencing tests directly from the genome performed in this work to confirm the absence of IS*10* in the murine genome, were important for testing one of the most striking results from the Human Genome Project. Lander and colleagues, in their draft publication of the Human Genome found that "*hundreds of human genes appear likely to have resulted from horizontal transfer from bacteria at some point in the vertebrate*

lineage. Dozens of genes appear to have been derived from transposable elements". If true, this would imply that horizontal transfer from bacteria breaks all natural barriers to reach the germ cells and may become fixed in the population much more frequently than expected. However, later experiments showed that only 28 genes of the original 113 listed potential bacteria-to-vertebrate transfers were confirmed by PCR in the human genome (Genereux and Logsdon 2003).

The hierarchical sequencing strategy employed by the human genome draft authors (Lander et al. 2001) and in this work implies the intermediate step of producing large genomic DNA libraries. During the propagation of the libraries in bacteria, transposition from mobile elements into the eukaryotic DNA does occur (Kovarik et al. 2001). This *in vitro* event may also happen in the whole genome shotgun as well but just in one sub-clone at a time and by chance in different areas of the genome. These artifact-reads could be very likely filtered and excluded from the alignment.

4.4 The novel *STK33/Stk33* gene

Serine/threonine kinase 33 gene is a novel gene that has not been described previously (Mujica et al. 2001). *STK33* is highly conserved in human and mouse, with 84% coding nucleotide sequence identity and 85% amino-acid identity of the inferred protein products. High local similarities (>75%) were observed in exons 3 to 10, which encode the catalytic kinase domain. Nuclear import of kinases without known NLS, have been reported (Schmalz et al. 1998). The absence of a typical NLS on the one hand, and the computer aided prediction of *STK33* being a nuclear protein on the other hand, leaves the question of

where it is localised in the cell still unanswered. Experiments are underway to analyse the subcellular localisation of *STK33*.

4.4.1 *STK33/Stk33* have low and differential expression

Northern analyses demonstrated that *STK33* is expressed differentially at low levels in various tissues in mouse and human. These results are in agreement with a relatively low coverage of *STK33/Stk33* in the EST databases. Very long exposition times were necessary to get signals in both Multiple Tissue Array™ and Cancer Panel Array™. This indicates that *STK33/Stk33* is transcribed at a rather low level.

The much more sensitive RT-PCR analysis shows a broader, but still not ubiquitous expression of *STK33/Stk33,* showing expression in tissues negative in the Northern analysis with alternative splicing evidence. In some of the tissues screened only shorter transcripts where detectable. This may correspond to a very low basal expression level.

The preliminary experimental data presented were supported by the literature and publicly available resources. Su and colleagues (2002) used Affymetrix technology to evaluate the global expression analysis of the human and mouse transcriptomes on a large scale and produced what they called the Gene Expression Atlas. They designed their DNA-chips based on the annotated genes found in UniGene, as well as from uncharacterised EST-clusters. DNA-chips from the characterised genes were hybridised with probes from forty-five tissues, whereas those with unknown ESTs were analysed with only probes from ten tissues. In their analysis *Stk33* fell under the uncharacterised EST-clusters and was

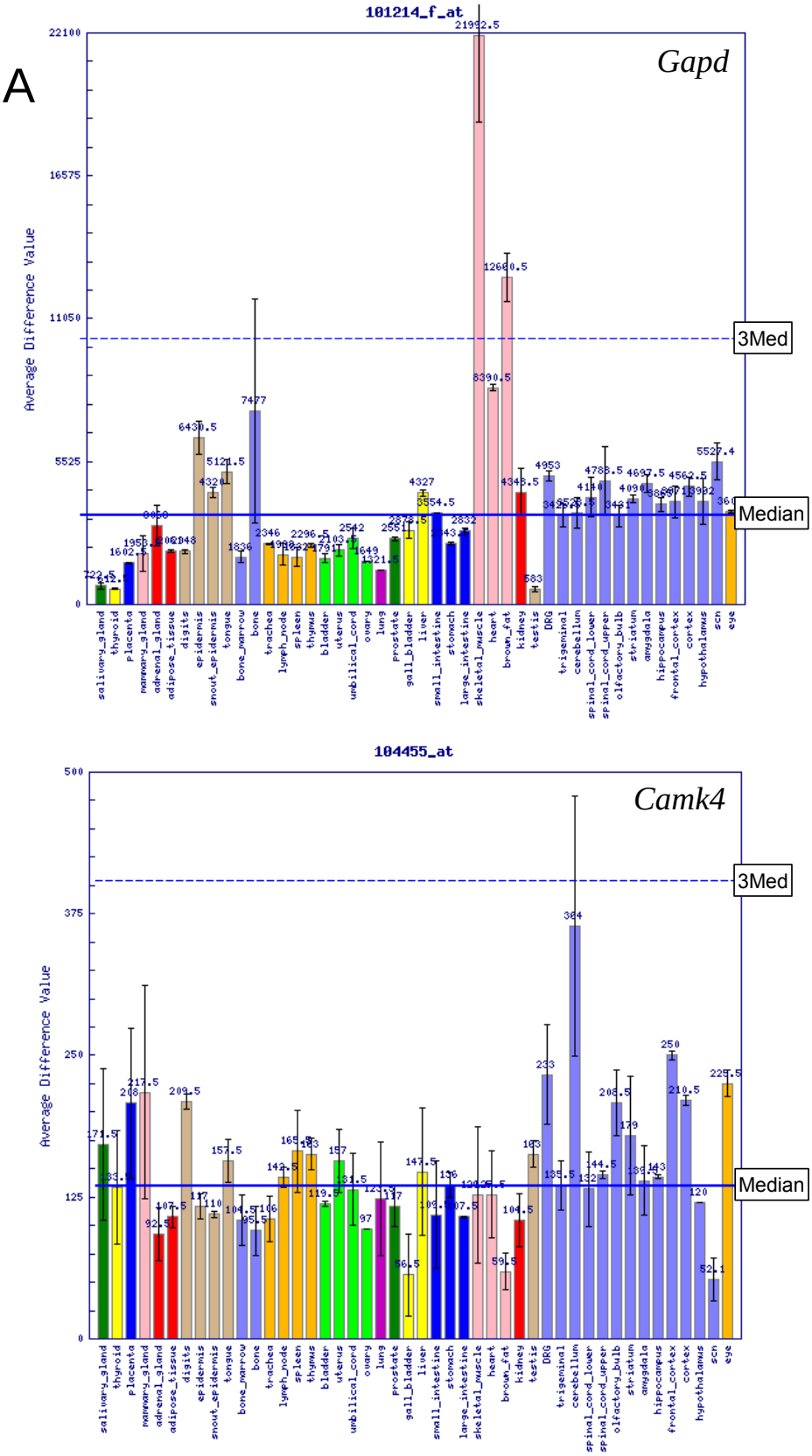
represented by the AW061264 entry. For comparison, in the figure 4.9 in the following two pages, the expression level of four genes as measured by affimetrix technology is shown. The genes are *Gapd*, glyceraldehyd 3-phosphate dehydrogenase, a housekeeping gene widely used as a normalisation standard of gene expression; *Camk4*, calcium/caldmodulin dependent protein kinase IV; *Stk33*; and *Phkg2*, phosphorylase kinase, testis-liver, γ-2, a kinase, also classified among the CAMK group and whose human homologous protein product clustered close to the *STK33* product in phylogenetic analysis (Mujica et al. 2001).

Notably, according to these results *Stk33* is predominantly expressed in testis, approximately twenty times higher than the median value for the ten tissues tested. The authors considered a gene as *expressed* with signals higher than the arbitrary value of 200, which in their results for *Stk33* was only the case for testis. This is in accordance with the signals obtained for testis in Northern analysis and RNA in-situ hybridisation experiments in this work, as well as the major representation of testis in the mouse EST data set.

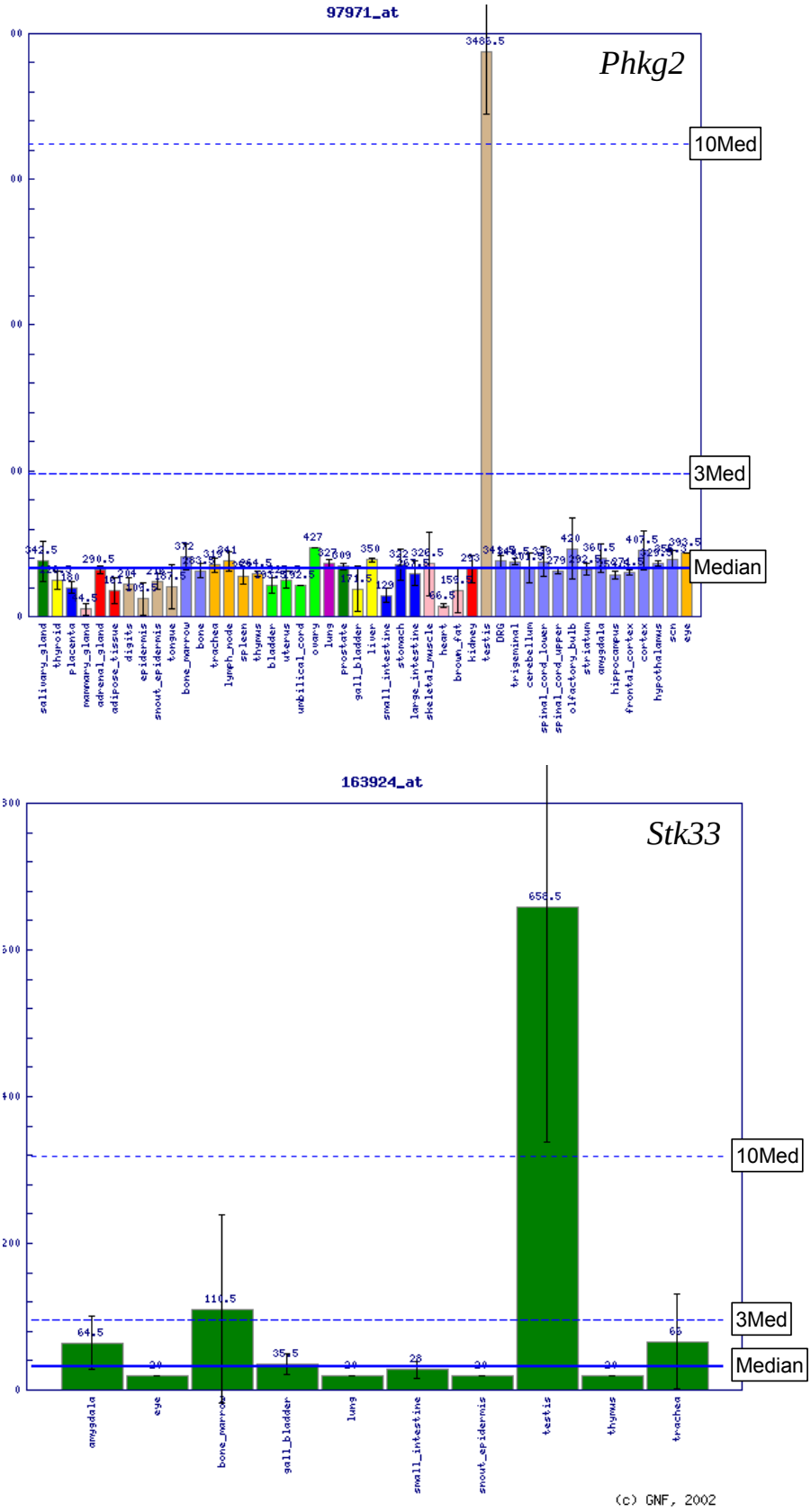
The maximum expression signal of *Stk33* is one order of magnitude lower than the maximum signal observed of *Phkg2* (which also had its maximum in testis) and two orders of magnitude lower than *Gapd*. *Camk4* also shows a very low level of expression, but the signals suggest a much broader tissue distribution, underlining *Stk33*'s testis specificity. *Phkg2* was originally described as testis-specific (Hanks 1989), but was subsequently associated with hepatic disorders (Burwinkel et al. 1998; Maichele et al. 1996).

Figure 4.9: Affymetrix expression assessment for A: *Gapdh, CamK4*; B: *Stk33, Phkg2*.
(In the next two pages) Data from the Gene Expression Atlas at the Genomics Institute of the Novartis Research Foundation. *Stk33* is represented in this analysis by the EST entry AW061264 and under the set of uncharacterised genes. Source: expression.gnf.org (Su et al. 2002)

A


101214_f_at
Gapd
Average Difference Value
3Med
Median
104455_at
Camk4
Average Difference Value
3Med
Median

B


97971_at
Phkg2
10Med
3Med
Median
3488.5
163924_at
Stk33
10Med
3Med
Median
658.5
110.5
64.5
35.5
(c) GNF, 2002

The second-highest *Stk33*-expression is observed in bone marrow, which may be in line with the observation of RNA in-situ signals in macrophages from the lung and liver. Notably the expression in the lungs is among the lowest in the data set. If the RNA in-situ signals are actually associated with a biological function, this should occur with very low levels of *Stk33* expression. Very low levels of expression in *Stk33* were also observed in parallel analysis from the GeneCards survey (Rebhan 1997), a database where the relative abundance of EST and other resources in the public database are used to compare the expression of genes (Figure 4.10).

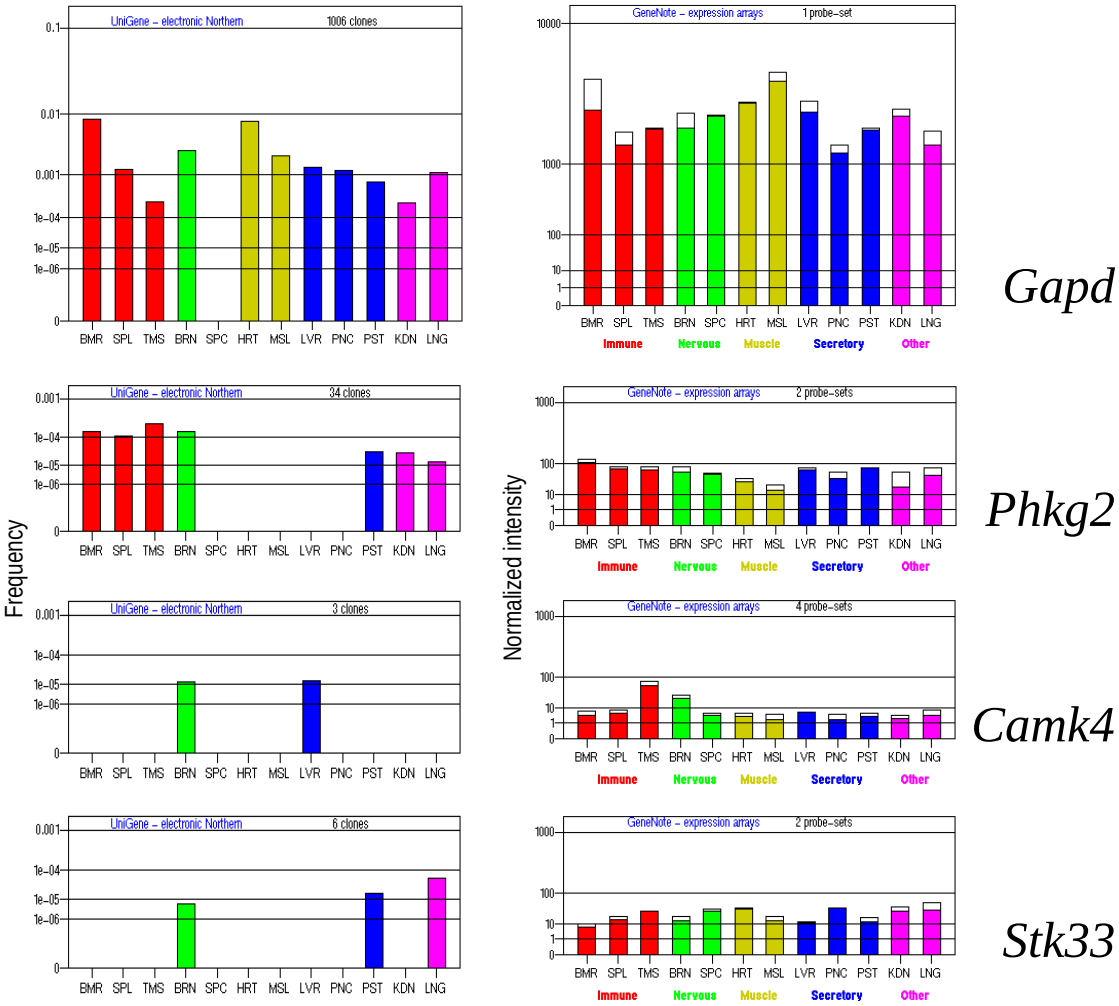


Figure 4.10: GeneCards expression assessment for *Gapd*, *Phkg2*, *Camk4* and *Stk33*.
Data from the GeneCards resource (Rebhan 1997).

4.4.2 RNA in-situ Hybridisation

The result of RNA in-situ hybridisation experiments with tissue sections from testis and lung, correlate with the MTE hybridisation experiments, where signals precisely from testis and fetal lung yield the strongest signals. There is also a good correlation with the EST pattern, since entries from lung and testis, above all, are very well represented in the public databases. The association of *Stk33* expression with germinal rather than with supporting cells like Sertoli or Leyding cells, is confirmed by several EST entries from spermatocytes and round spermatides from mice (see table 3.7 and UniGene cluster Mm.79075). It might be a weak hybridisation signal in the first basal layer of cells made up of the spermatogonia. Spermatogenesis first described by Leblond and Clermont (Leblond and Clermont 1952), and (Holstein et al. 2003) for a recent review) starts with a first round of mitosis, in which one spermatogonia daughter cell stays as stem cell reservoir in the basal compartment, while the other enters the adluminal layer and goes through the multiplication phase, which results in the spermatocytes. After meiosis, spermatocytes turn into spermatides. *Stk33*-expression as seen by in-situ hybridisation seems to have a maximum in the dividing spermatocytes and early (round) spermatides. Spermatides become differentiated (spermiogenesis) in late spermatides which are clearly signal-free. They are released (spermiation) as spermatozoids in the lumen of the seminiferous tubule and get transported through peristaltic activity of the peritubular myofibroblasts, which also show no hybridisation to *Stk33* probes, in the rete testis.

The expression pattern in the lung is not so easy to establish. Hybridisation is observed in two different cell types, first in respiratory epithelium in the bronchi and second in alveolar macrophages. However, RNA in-situ hybridisation experiments with blood smears did not yield any *Stk33*-specific signals in cells from the immune system. On the other hand, the epithelial signal is also supported by several specific EST entries from lung epithelial cells, whereas just one entry from the germinal B cell is the only representation from the immune system. It cannot be excluded, however, that both signals are positive.

Experiments with the kidney, the organ which showed the fifth identity MTE signal, did not produce satisfactory results: the background was too high to identify any specific signal. In addition, there was no clear difference between anti-sense, sense and even control RNA from prokaryotic origin. This may have been due to the action of RNA-binding proteins in the preparations or unspecific binding to the tissue. In addition, tissue-specific endogenous alkaline-phosphatase activity may produce misleading cell-specific staining (Weiss et al., 1988). For this reason, the RNA in-situ hybridisation results in the liver may be also misleading. Indeed, elevated concentration of alkaline-phosphatase, produced in the liver from the hepatocytes, is a common diagnostic marker in blood tests for injury to any part of the biliary tree (Mahl 1998).

Some signal was also observed (data not shown), albeit for a confusing nature, in sections of the retina and intestine. There are several retina EST entries, particularly from human embryos, but *STK33/Stk33* expression in these tissues and its histological localisation remain to be confirmed.

4.4.3 Alternative splicing

It is known that alternative splicing plays a major role in the function of several kinases (Brodbeck et al. 2001; Burgess and Reiner 2002; Chalfant et al. 1995; Wansink et al. 2003). Even alternative transcripts coding for apparently inactive proteins may play a major regulatory role. The Lck tyrosine kinase isoform lacking exon 7 (LckΔ7) and hence the ATP-binding domain, retains some phosphorylation capabilities, and may function as cell-signalling regulator of the native isoform (Germani et al. 2003).

STK33/Stk33 versions show differential 5' UTR, with fewer or no changes in the coding sequence. These transcripts could eventually experience differential regulation at the mRNA level, with important functional implications. Alternative start points of transcription may be under the regulation of different transcription factors in different tissues. Different protein N-terminal may yield to distinct transport to sub-cellular compartments. The genes for protein kinases *CDC2L1* and *CDC2L2* (also named p58 and p34 or *PITSLRE* A and B) are nearly identical tumor suppressor genes likely involved in apoptosis, located closely together tail-to-tail on chromosome 1p36, a region frequently deleted during late stages of tumorigenesis and associated with enhanced metastasic potential (Lahti et al. 1995); (Gururajan et al. 1998). Through alternative splicing, *CDC2L1* and *CDC2L2* code for a variety of isoforms that vary in the length and sequence of their N-terminal regions, retainig the protein kinase catalytic and C-terminal domains, in a similar way as observed in *STK33* potential transcripts.

The variability of *STK33/Stk33* transcripts was confirmed by immunodetection of at least two Stk33 protein isoforms in the mouse. Synthetic peptides were designed as epitopes for antibody production in regions outside the kinase catalytic domain in order to avoid unspecific signals with homologue proteins and possibly in human-mouse conserved sites.

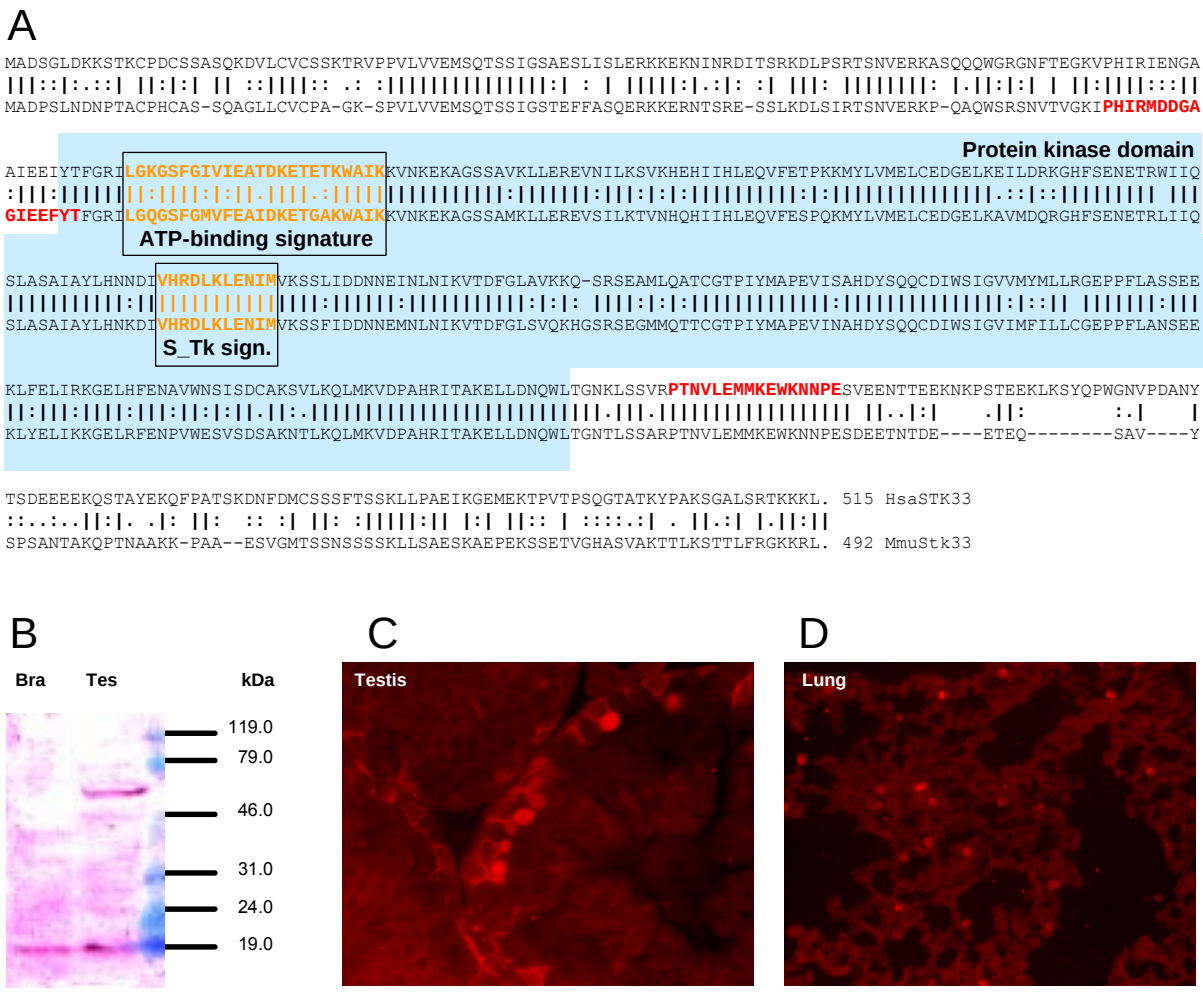


Figure 4.11: Peptides designed for antibody production and first immunodetection with mouse proteins extracts and tissue sections

A Peptides used for immunisation are shown in red. **B** First test with antibody against peptide 1 with protein extracts from mouse brain (Bra) and testis (Tes). **C** and **D:** Immunostainig of mouse testis and lungs preparations with antibody against peptide 1 (**B, C and D** kindly provided by Brauksiepe, unpublished results).

The C-terminal peptide (**PTNVLEMMKEWKNNPE**) is totally conserved between human and mouse and hence very likely to function also with material of human origin. The antibody against the N-terminal peptide (**PHIRMDDGAGIEEFYT**) is not fully conserved but still may work in human, due the similarity of this position in both species.

Immunodetection with rabbit antibody against Stk33 the peptide 1 in protein extracts of testis shows a band correlating with the molecular weight of the putative native isoform (54.5 kDa). As shown in the figure 4.11, no band of this size is detected in brain extract, which is in line with the lack of signal with cDNA from the brain in the human Multiple Tissue Expression array. A second band of an isoform with much lower molecular weight, present in both tissues, correlates strikingly with the putative protein product of human transcript variant D (19.5 kDa), which remarkably lacks all eight exons coding for the kinase domain but still conserves the reading frame and may correspond to the mouse variants C, E, F or G (see figure 3.26). It is conceivable a functional STK33 isoform that lacking phosphorylation activity but retainig other domain for protein-to-protein interaction in the N- and C-terminal, may function as repressor of the native isoform activity (see figure 4.16 for a model).

4.4.4 On protein phosphorylation

The phosphorylation of the amino-acid side chains of a protein typically produces major conformational arrangements. This post-translational modification is catalysed by protein kinases and may lead to changes in the activity of the phosphorylated protein by

exposing or hiding its active site, or it may influence the way the protein interacts with its ligands. The phosphate group is provided by an ATP and the amount of free energy released in the breakage of its phosphate-phosphate bond makes this process essentially unidirectional. However, the phosphate may be removed by dephosphorylation by protein phosphatases. The protein phosphorylation state of a cell is determined by the balance between kinases and phosphatases. Frequency, specificity and activity of kinases and phosphatases are very diverse and together mediate most of the signal transduction in eukaryotic cells. They play a major role in the regulation of fundamental cellular processes, such us DNA replication, metabolic pathways and cell growth, differentiation, proliferation and cell death (Alberts 2002).

There are 518 known active kinases in the human *"kinome"*, the catalogue of all protein kinases occurring in *Homo sapiens* (Manning et al. 2002b). The draft sequence of the human genome had more predicted: 575 (Lander et al. 2001) and 868 (Venter et al. 2001). It is a well-studied protein family but even after years of targeted cloning experiments, 15% of its members remain undescribed, probably due to a restricted level of expression (Manning et al. 2002b).

The classification of the family (Hanks and Hunter 1995) is based on similarity of the kinase domain, which comprises around 270 amino acids. Kinases that phosphorylate serine or threonine residues are the most frequent, followed by tyrosine kinases and a group of atypical kinases with confirmed biochemical kinase activity but very remote similarity to the typical kinase domain. Crystal structure data from representative protein kinases have shown a common general folding consisting of two major lobes (Taylor et al. 1995). The N-

terminal ATP-binding lobe is mainly made up of ß-sheet structures is an ATP-binding motif which is unique to protein kinases. The larger COOH-terminal lobe principally structured with α- helices confers the substrate specificity and is where the actual phosphotransfer takes place.

Hanks and Quinn (1991) reported the conserved features in the protein sequence of several protein kinases. Hanks and Hunter (1995) described in detail the twelve canonical subdomains of protein kinases. The following description of their general features is based mainly on this fundamental review. The subdomains I-IV conform the N-terminal ATP-binding lobe, whereas subdomains VIA-XI conform the C-terminal, lobe and subdomain V functions as a pivot between the two lobes. The cleft between the lobes is the site of catalysis.

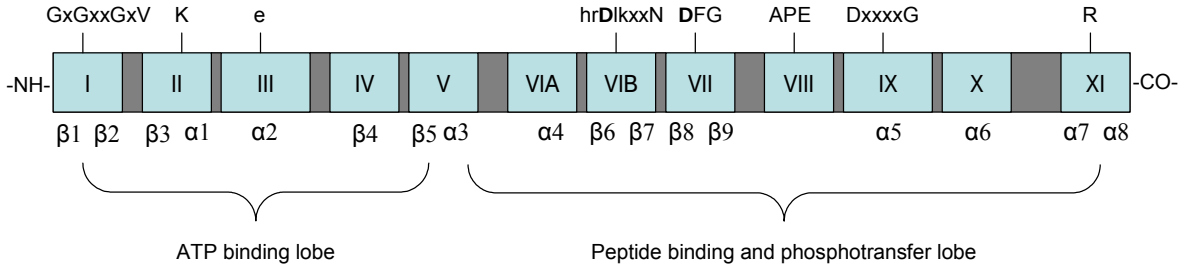


Figure 4.12: Canonical protein kinase domains.
Light boxes represent the subdomains with the roman numerals originally proposed by Hanks and Hunter, which are widely adopted in the literature. Over each block the invariant residues are shown with the single-letter amino-acid code in capitals, the nearly invariant with lowercase letters and x represents any residue. The spacers between the subdomains (represented here by dark grey segments) may vary greatly between different protein kinases and generally build loops towards the surface of the protein. These loops also vary greatly between different kinases and play a major role in the distinguishable interactions with other proteins. Typical secondary structure motifs are depicted under each corresponding subdomain. Adapted from Hanks 2003; Hanks and Hunter 1995 and Knigton et al., 1991.

Subdomain I consists of two ß-sheets and contains the nearly invariable string Gly-x-Gly-x-x-Gly-x-Val. The first glycine of the consensus is the residue number 8 of the typical protein kinase domain. This secondary structure motif acts like a mobile clamp that

folds over the non- transferable phosphates of the ATP. **Subdomain II** contains a ß-sheet which is followed in many but not all kinases by an α-helix. In this subdomain lysine is an invariant residue which forms a salt bridge with glutamic, or similar residue, in a central position of a large α-helix from **subdomain III**. **Subdomain IV** consists of a hydrophobic ß-sheet and has no invariant residue. All these four subdomains fit and orient the ATP in the right position for the phosphate transfer in two ways: they provide a hydrophobic pocket where the adenine ring docks, and interact with the α- and β- phosphates by means of basic residues.

Subdomain V connects the two lobes of the typical protein kinase fold and consists of a very hydrophobic ß-sheet and a short α-helix. It contributes to the hydrophobic pocket around the ATP's adenine ring, but ion-pair interactions between polar residues of this domain with residues in the near vicinity of the peptide substrate have also been observed. **Subdomain VIA** is made up of a large α-helix and has no invariant residues. A structural role for this domain is proposed, since any interaction with neither the ATP nor the substrate has been observed. **Subdomain VIB** has two ß-sheets and the fold between them is called the catalytic loop, since it contains the catalytic aspartate, which functions as a proton acceptor from the hydroxyl group of the substrate. The consensus-string His-Arg-Asp-Leu-Lys-x-x-Asn characterises the protein-serine/threonine kinases, whereas arginine takes the place of lysine in the protein-tyrosine kinases. Several functions have been associated with the residues in this subdomain, such as neutralisation of the charge of the γ-phosphate during the transfer, chelation of the Mg^{2+} ions which bridge the α- and β- phosphates of the ATP, hydrogen bond formation with the ATP's ribose, and ion-pair forming with residues on the substrate. **Subdomain VII** also has two ß-sheets with a functionally important loop in

between. The consensus triplet Asp-Phe-Gly is highly conserved among protein kinases. It helps to orient the γ-phosphate for transfer by chelating activation Mg^{2+} ions, which also bridge the α- and β-phosphates. The highly conserved Ala-Pro-Glu characterises the **subdomain VIII,** which is believed to play a major role in both substrate recognition and protein activation. In a mechanism proposed for some protein kinases, activation occurs by phosphorylation of this subdomain, which becomes the proper orientation towards the substrate. In the unphosphorylated state it blocks sterically the catalytic cleft. **Subdomain IX** contains a large α-helix and a highly conserved aspartate, which stabilises the catalytic loop by hydrogen bonding. Specific interactions have been also detected between residues in this subdomain and the peptide substrate. **Subdomains X** and **XI** exhibit less conservation among protein kinases; they are made up of the last three α-helices and their function other than structural is not clear.

Protein kinases of the CaMK group tend to prefer substrates with basic residues, and the primary sequence around the phosphorylation site play a major role in substrate recognition (Hanks and Hunter 1995; Soderling 1999). Acidic residues are present in some CaMK kinases between sub-domains VIB and VII, but in STK33 they constitute a distinctively acidic loop (see figure 3.40). *Dictyostelium discoideum* myosin light chain kinase (KMLC_DICDI, P25323), the kinase with highest similarity to STK33, has one such acidic residue (Asn-143); and human Ca^{2+}/calmodulin dependant kinase (CAMK1, NP_003647), the human kinase most similar to STK33, has a string of three Asp153-Glu154-Asp155. The accumulation of Asx and acidic residues may conceivably have been an evolutionary trend in STK33 evolution, as this development is observable in protein alignment of several species (figure 3.43).

STK33 acidic loop forms a loop close to the actual components of the active site. It may determine substrate recognition in a very specific manner. Asx and Glx residues are frequent constituents of much bigger cation-binding domains like the EF-hand. N-terminal from the acidic loop a putative Casein kinase II phosphorylation site is observed that is also conserved in all species evaluated (shown in blue in the figure 3.40). Perhaps, a model of activation/inhibition of STK33, as shown in the figure below, may include such post-translational modification and the acidic loop.

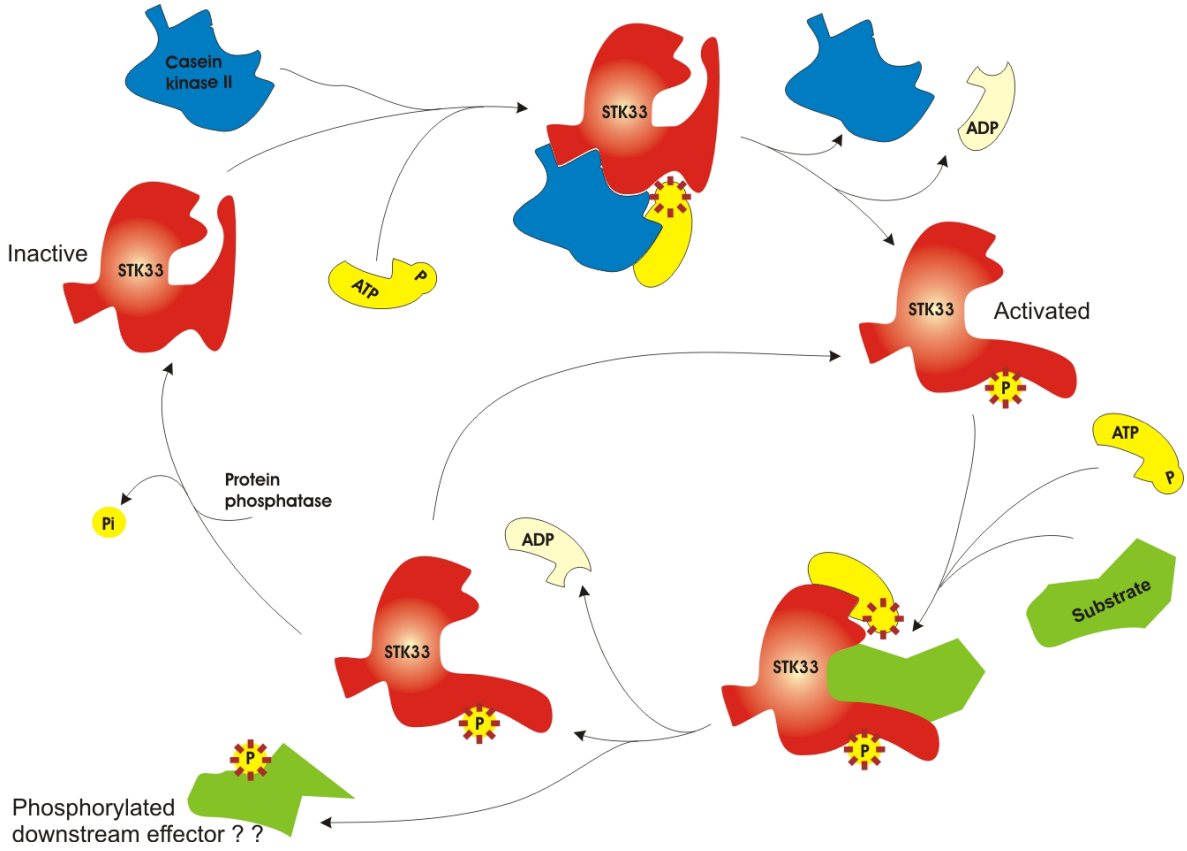


Figure 4.13: **Hypothetical model of activation of STK33 through phosphorylation close to the STK33-specific acidic loop.**

4.4.5 Outside the catalytic domain

Members of the CAMK group of kinases, to which the STK33 shows the highest similarity, are known to be regulated by a C-terminal Ca^{2+}/calmodulin binding catalytic tail, first documented by Knighton and colleagues in 1992. Wilmanns and colleagues in 2000 generalised the model of activation for several homologous kinases from the group. According to this model, the kinase is inactive (auto-inhibitory form) through its own regulatory tail that blocks the active site to the substrate. As a consequence of a calcium signal, four Ca^{2+} ions bind to the closed conformation of the EF-hand-type calmodulin protein. This Ca^{2+}/calmodulin complex in turn binds to the CAMkinase C-terminal regulatory tail, releases it from the catalytic cleft of the kinase and, by so doing, activates the enzyme (Figure 4.14).

Though well characterised, the Ca^{2+}/calmodulin binding is not recognised by protein resources (Manning et al. 2002b) due to the lack of sequence conservation, *"they are usually too divergent to be detected by significant similarity"* (Wilmann et al. 2000). Even among orthologous genes, the similarity drops abruptly outside the catalytic domain (see figure 3.37 and table 4.4). But despite this lack of conserved amino-acid sequence between the C-terminal regions of divergent kinases, regulatory tails of some CAMK proteins, display a similar topology (Wilmann et al. 2000).

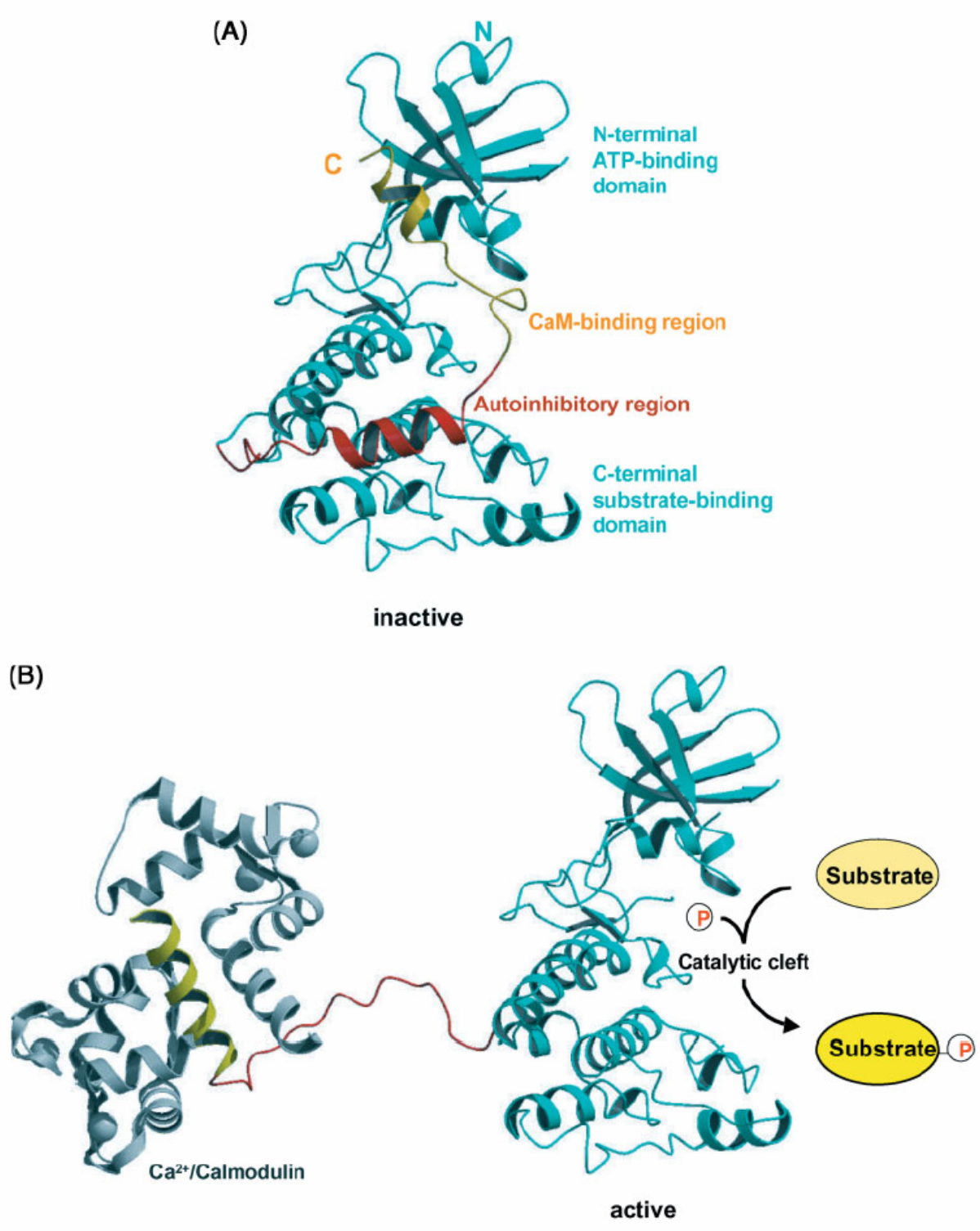


Figure 4.14: Model of Ca^{2+}/Calmodulin activation of CaM kinaseI.
A: Auto-inhibitory C-terminal region of CaM kinaseI blocks the catalytic cleft. **B:** Ca^{2+}/CaM binds the C-terminal regulatory region putting it away from the catalytic cleft and leading to the activation of the CaM kinaseI (Ikura et al. 2002)

The results of STK33/Stk33 analysis with the MetaServer (figure 3.41) do indeed suggest a nearly conserved pattern of folding in the C-terminal region, which may correspond to a motif binding the Ca^{2+}/calmodulin activator, or of the other calmodulin-similar Ca^{2+}-binding proteins recently discovered (Haeseleer et al. 2002). Hence, a mode of activation of STK33/Stk33 through Ca^{2+}/calmodulin, as shown in the figure 4.15, is feasible.

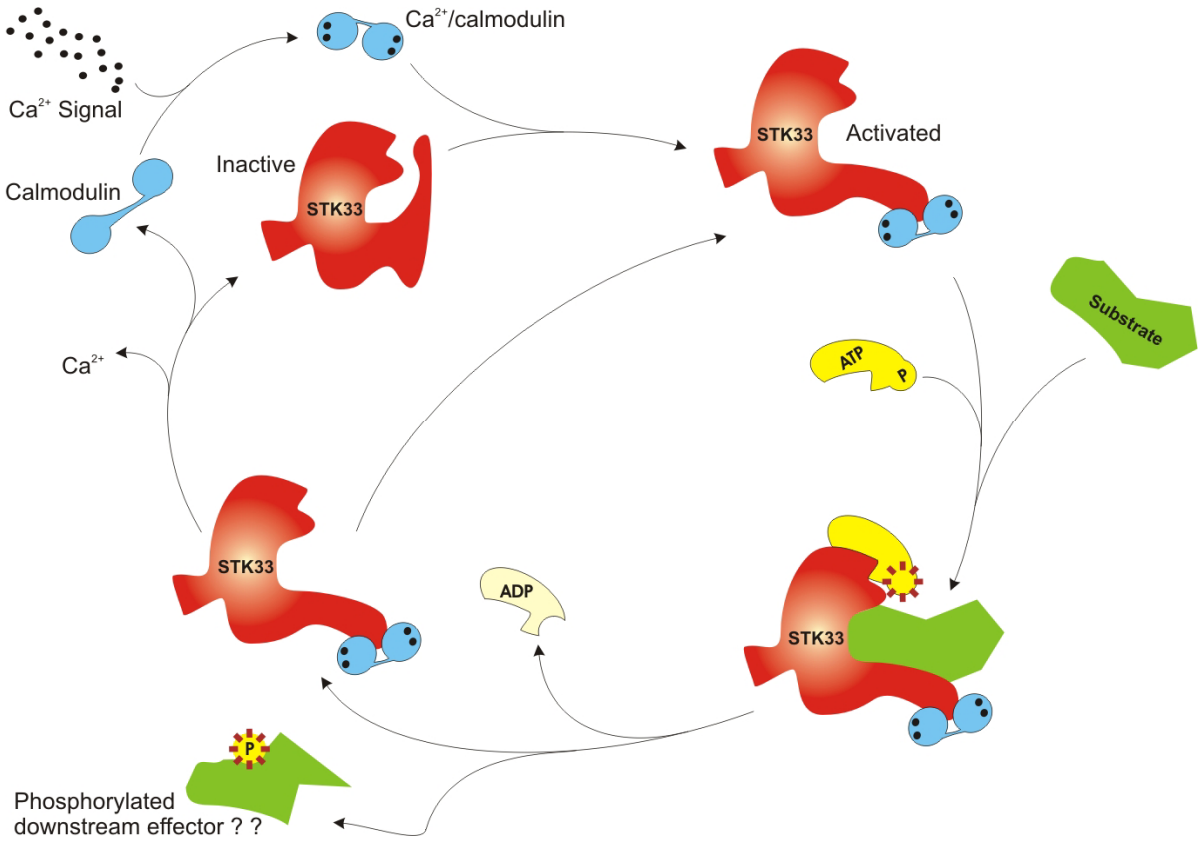


Figure 4.15: Hypothetical alternative model of activation of STK33 through Ca^{2+}/calmodulin. Phylogenetic analysis and protein folding prediction let hypothesise a regulation of STK33 through Ca^{2+}/calmodulin.

4.4.6 A model for STK33 function

In vivo, several factors may achieve a limited expression and activity of *STK33* in space and time. As already discussed, regulation through posttranslational modifications and interactions with cation-binding EF-hand proteins may be involved in triggering or inhibiting STK33 activity.

The results in of this work let postulate alternative models at the RNA level. Very short transcripts were detected by RT-PCR in several tissues which are also confirmed by EST data. Despite the splicing of the majority of the core exons coding for the kinase domain, these transcripts still retain the reading frame. If these short transcripts are translated at all, they produce protein isoforms which very likely lack phosphorylation activity. N- and C- terminal regions outside the catalytic domain, perhaps, harbour specific motifs that determine STK33 interaction with its environment. Isoforms having this motifs but with no phosphorylation activity may function as repressors. This is a very well known scenario in alternative splicing, and perhaps is also the case for STK33.

STK33/Stk33 seems to be expressed in tissues with active cell division, and low or no expression elsewere. Repressor isoforms may play an important role in those tissues where *STK33/Stk33* should not be active. In order to test this, a relative quantification of diverse transcripts in different tissues is necessary. Figure 4.16 resumes this and all models of regulation of *STK33/Stk33* expression and activity that may be proposed based on the data presented in this work.

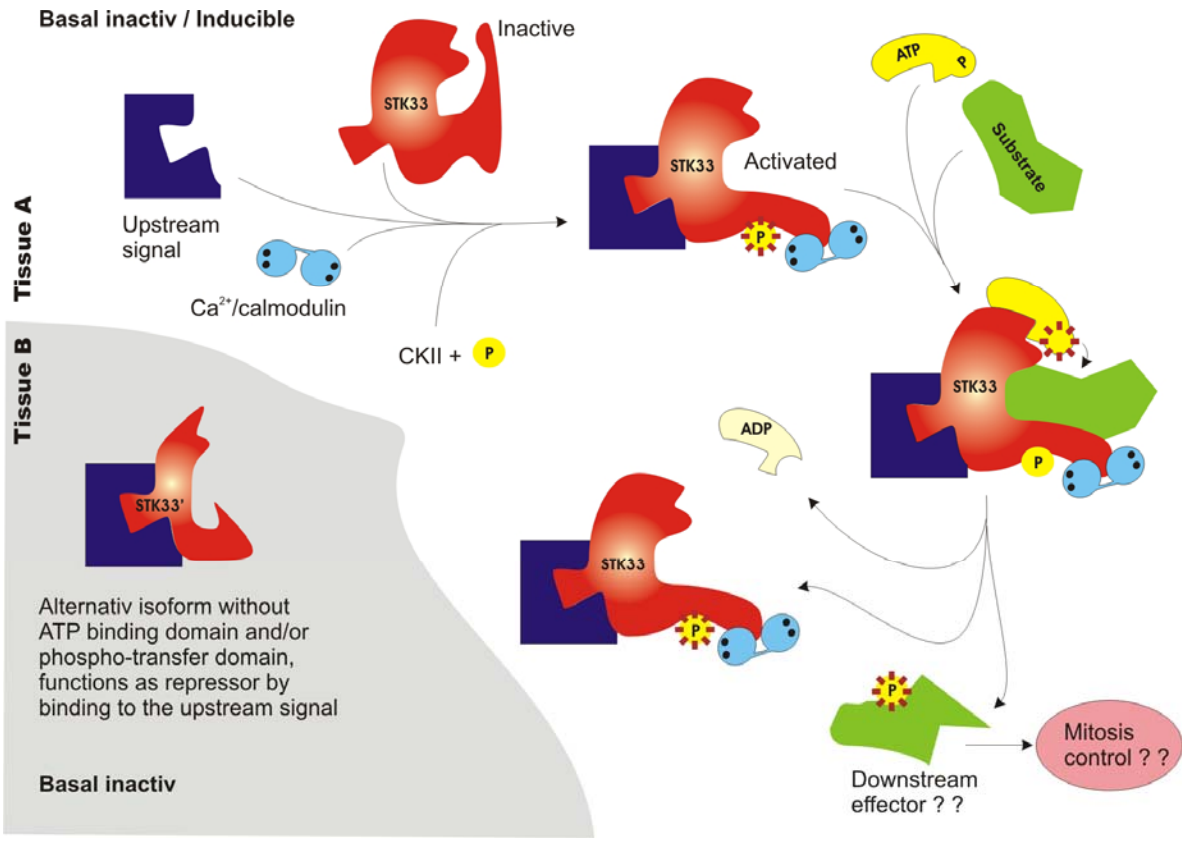


Figure 4.16: Model of basal inactivity in non mitotic tissues through alternative splicing.

4.4.6 Phylogeny and classification of STK33

Phylogenetic analysis has been employed to classify the super-family of eukaryotic protein kinases ever since it became clear that it was a very diverse protein group (Hanks et al. 1988). Roughly half of the protein kinases (Manning et al. 2002b) exhibit a complex structure, in which the phosphorylation domain (historically named **ePK**: eukaryotic protein kinase domain) is in combination with known additional domains, as shown in the figure 4.17, conferring particular function, protein-interaction pattern or sub-cellular localisation. Hence, in order to be effective, a classification of protein kinases based on molecular phylogeny must be constrained to the very core of the protein phosphorylation domain. The 250 to 300 amino-acid residues long catalytic domain was first detected by sequence

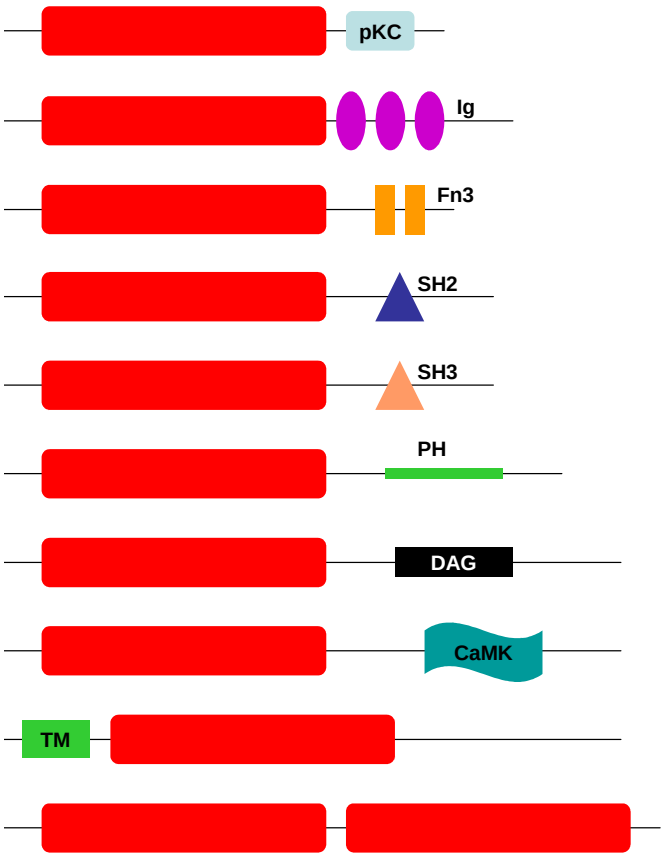


Figure 4.17: A sample of domain diversity in multi-domain kinases

Red Bars represent the relative position of the protein kinase domain. Other domains are represented as follows: **pKC,** Accessory from protein kinase C; **Ig,** Immunoglobulin; **Fn3,** Fibronectin type III; **SH2,** phosphotyrosine binding; **SH3,** prolin-rich motifs binding; **PH**, phospholipid binding; **DAG,** Diacylglycerol binding; **CaMK,** Ca^{+2}/calmodulin binding; **TM**, transmembrane region. This list does not pretend to be exhaustive, till now, up to 83 additional domains are detectable in the protein kinases. Modified from Krupa and Srinivasan 2003 and Maninnig *et al.*, 2002b.

conservation and subsequently confirmed by assay of truncated enzymes (Hanks et al., 1988). Further comparison of additional sub-domains, together with a knowledge of biological function has been useful in resolving minor classification conflicts within a group (Manning et al. 2002a). Current classification of the human kinome is based on a hierarchy of groups, families and subfamilies (Manning et al. 2002b), based mainly on the Hanks and Hunter canonical classification (Hanks et al. 1988); (Hanks and Hunter 1995). Kinases that phosphorylate serine and threonine, the most frequent among the super-family, are divided into the following groups: AGC, CAMK, CK1, CMGC, Other and STE. The metazoan-specific tyrosine kinases are distinguishable in the groups TK, TKL and RGC (table 4.4).

Table 4.4: Groups of the typical eukaryotic protein kinases and their respective families

Type	Group	Families
Serine/ Threonine kinases	AGC	Cyclic nucleotide dependent (PKA, PKG) and protein kinase C (PKC)
	CAMK	Calcium/calmodulin-dependent protein kinases
	CK1	Casein kinases
	CMGC	Cyclin-dependent kinases, MAP kinases, Glycogen synthase 3, Cdk-like kinases
	Other	Other relative kinases
	STE	Mitogen activated and STE kinases
Tyrosine kinases	TK	Typical tyrosine kinase
	TKL	Tyrosinekinase–like kinases
	RGC	Receptor guanylate cyclase group

(Hanks and Hunter 1995; Manning et al. 2002a; Manning et al. 2002b)

Finally, atypical protein kinases are mostly constituted by single families of proteins with experimentally confirmed phosphorylation activity but with little or no sequence similarity with the canonical protein kinase domain. Typical CAMK proteins exhibit a C-terminal regulatory calcium/calmodulin binding domain, as shown for the canonical CaMKI (figure 4.14). However, not all members of the CAMK group are activated by the calcium/calmodulin complex (Hanks and Hunter 1995). The presence of a C-terminal calcium/calmodulin binding domain in STK33 is still to be confirmed, as discussed

previously. Additional 3D data would be necessary to identify such a C-terminal regulatory tail unequivocally (Wilmann et al. 2000).

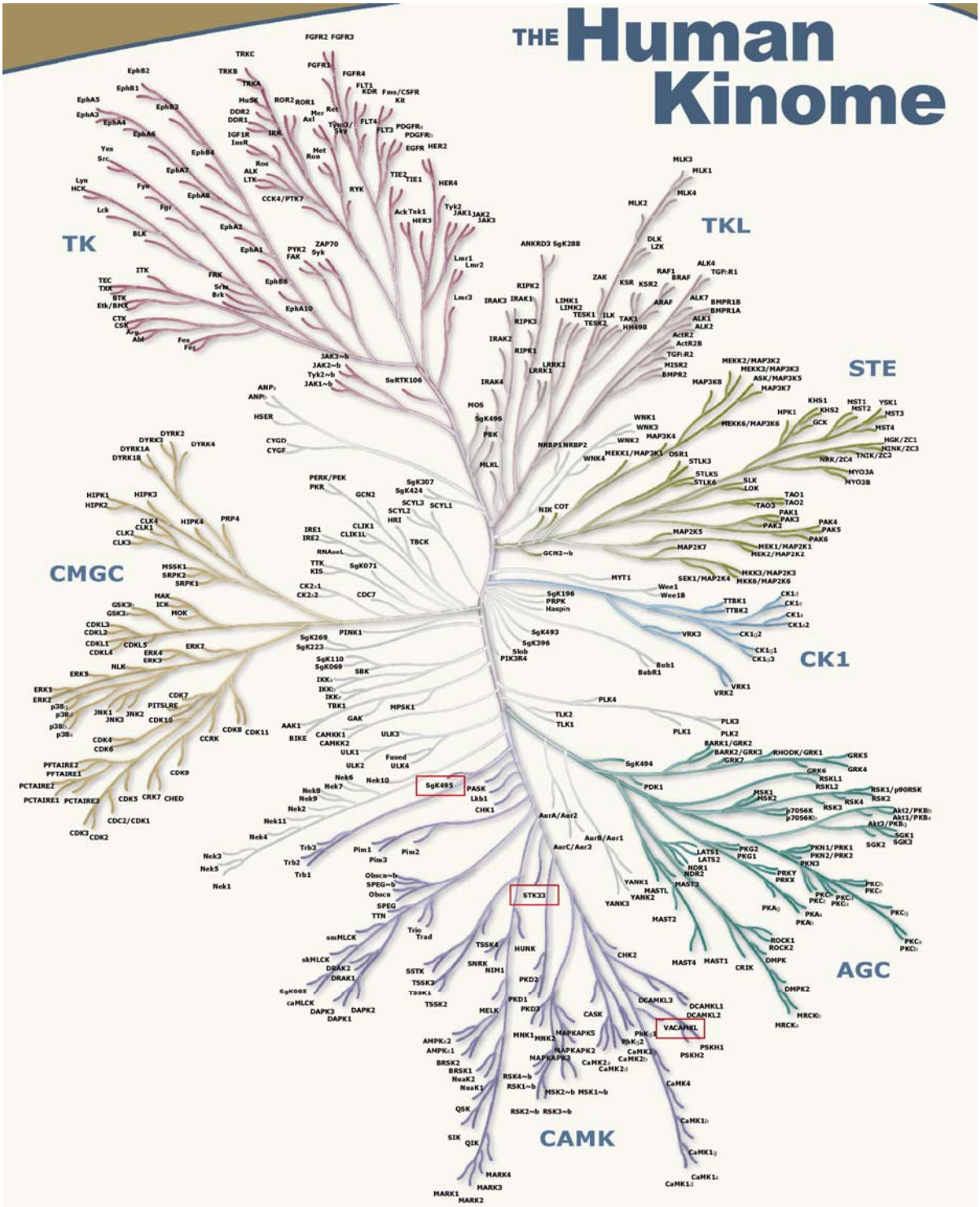


Figure 4.18: Human kinome phylogenetic tree (Manning et al. 2002b).

The classification of STK33 as established in (Mujica et al. 2001) and shown in the figure 3.41, is in agreement with the human kinome phylogeny published last year (Manning et al. 2002b). STK33 is placed among the CAMK protein kinases, particularly within a family they termed CAMK-Unique together with the novel kinase SgK495 and the VACAMKL which is annotated in the database but has not been described in the literature (Supplementary data from (Manning et al. 2002b). These three CAMK members do not cluster together and exhibit relatively low similarity with each other (see figure 4.18), but they do have in common that they diverge alone from their respective principal branches and hence do not form any distinguishable monophyletic group with neighbour members.

In the human kinases phylogeny Kostich and collaborators (2002) classified STK33 in the CAMK group also like a monophyletic cluster within the neighbour members of the group (see figure 4.19).

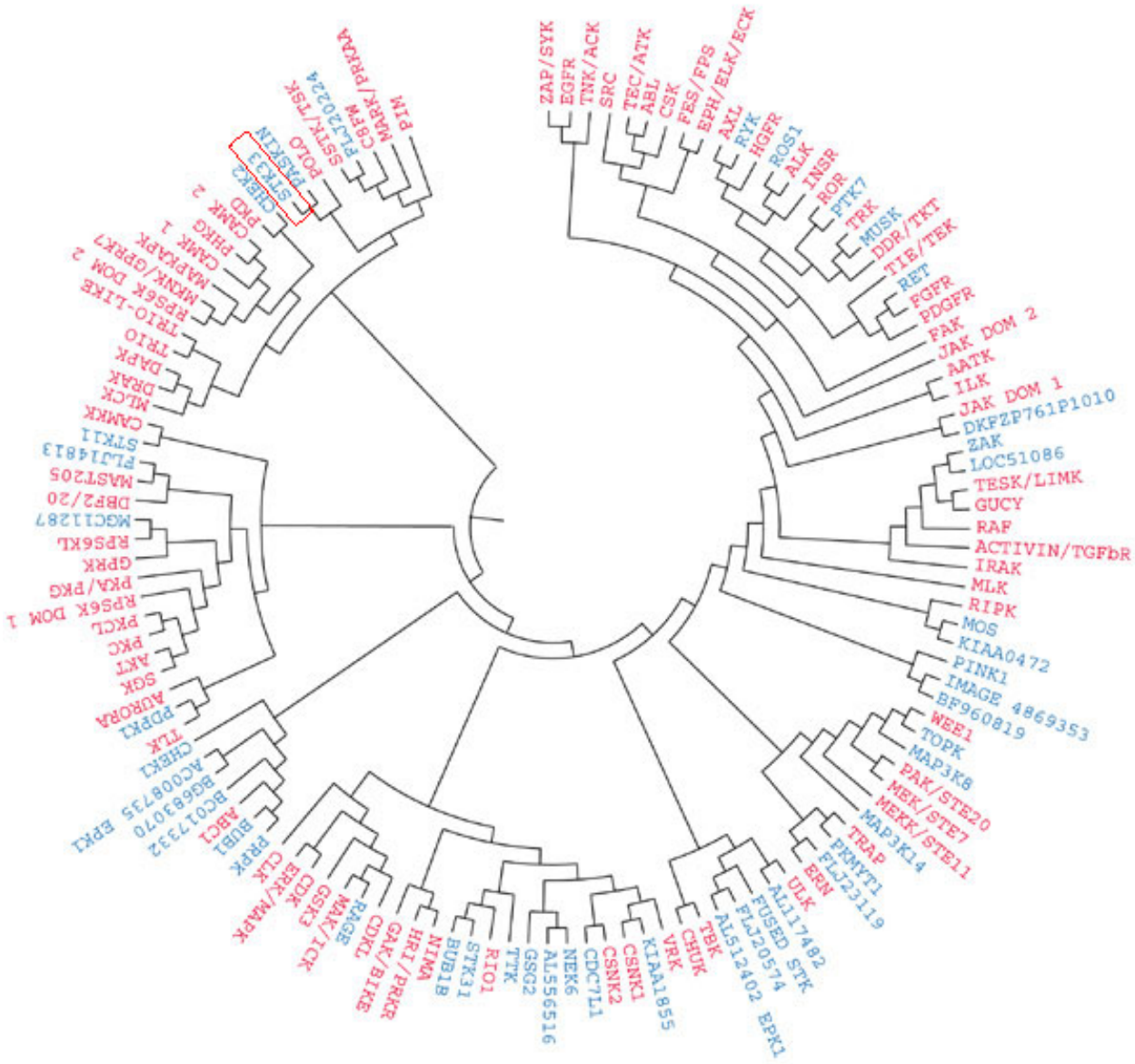
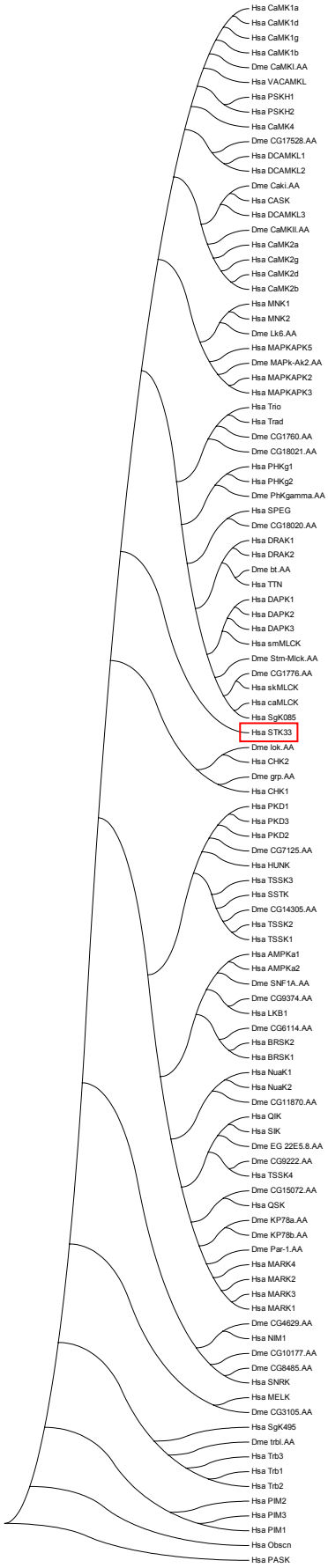


Figure 4.19: Dendogram of human protein Kinases
Red terminal nodes represent collapsed branches from a phenogram that groups their whole catalogue of 510 human kinases. Blue terminal nodes represent singletons, members of the phenogram which *"do not cleanly fall into a cluster"* (Kostich et al. 2002).

Phylogenetic analysis and similarity searches in public databases have not yet conduced to clear orthologous of STK33 in yeast, worm or fly. Considering the advanced state of annotation and validation of these three model genomes, it is very unlikely that the STK33 gene version in these organisms remain underrepresented in the genomic scaffold. The well-studied kinomes of yeast, fly and worm (Manning et al. 2002a) were used to align



all protein kinases belonging to the CAMK group of these model organisms with all human CAMKs. These alignments were then used to draw phylogenetical trees by Neighbour-Joining and Maximum-parsimony methods. Consistently, known orthologues grouped together, whereas STK33 produced long single branches or singletons in most analysis performed.

This conclusion is in line with the fact that vertebrates have twice the number of CAMK protein kinases than flies and worms. Figure 4.20 clearly shows how some genes of this group have experienced several duplications in the primate lineage compared with the insects, resulting in several isoforms in the human for each fly gene. This is the case for CamKI and CamKII, which both have just one member in *Drosophila* and four in the human (Manning et al. 2002b). In fact, humans have more than twice as many CAMKs as yeast, worms, and flies as clearly shown in table 4.5.

Figure 4.20: Example Neighbour-Joining tree topology of CAMK kinases from the Human and *Drosophila melanogaster*.
Protein sequences were obtained from www.kinase.com and aligned with the program ClustalW. Neighbour-Joining tree was calculated with the MEGA program with the pair-weise gap deletion criteria.

Table 4.5: Typical eukaryotic protein kinases in some model organisms and humans

Group	Yeast	Worm	Fly	Human
AGC	17	30	30	63
CAMK	21	46	32	74
CK1	4	85	10	12
CMGC	21	49	33	61
Other	38	67	45	83
STE	14	25	18	47
Tyrosine kinase	0	90	32	90
Tyrosinekinase–like	0	15	17	43
RGC	0	27	6	5

Taken from (Manning et al. 2002b)

The most distantly related *STK33* orthologous known up to now, is found in *Ciona intestinalis*, an ascidian model organism member of the urochordates (Dehal et al. 2002). This, together with the absence of orthologues in the genomes of fungi, nematodes and insects, indicate that *STK33* may be a *"novel invention"* in the chordate lineage, and its functional role could be specific to the typical features of the chordate body plan. Comparative genomics may help to resolve this issue by reconstructing the phylogeny of *STK33*, identifying the orthologous genes in model organisms and developing experiments to determine its cellular function.

In order to estimate the rate of evolution of *STK33/Stk33*, the ratio of nonsynonymous substitutions to synonymous substitutions in human and mouse was calculated using the SNAP server, which incorporates the methods of Nei and Gojobori (1986) and the statistics of Ota and Nei (1994). The number of observed synonymous substitutions was Sd=128.17 and the number of nonsynonymous substitutions was Sn=170.83. Calculated over the potential substitutions codons, the resulting proportions were ps=0.42 synonymous substitutions per synonymous position and pn=0.15 nonsynonymous substitutions per nonsynonymous codon. These values were corrected for

multiple substitutions, according to Jukes and Cantor (1969), resulting in ds=0.62 for synonymous substitutions and dn=0.16 for nonsynonymous substitutions and a dn/ds ratio of 0.26. This value suggests that *STK33* has been under purifying selection. Table 4.6 shows the dn/ds ratios for each exon and protein region.

Table 4.6: dn/ds of *STK33/Stk33* exons in human and mouse

Exon	ePK sub domains	Features	Indels (aa)	Identity (aa) %	dn	ds	dn/ds
6*			4	54.4	0.26	0.66	0.39
7			1	92.9	0.20	0.50	0.40
8	I, II	ATP binding		81.6	0.08	0.15	0.53
9	III, IV			80.0	0.09	1.17	0.08
10	V, VIA			84.8	0.08	0.67	0.12
11	VIB	Phosphorilation, Acidic loop		93.3	0.10	0.21	0.48
12	VII, VIII		1	73.3	0.10	0.20	0.50
13	IX			85.2	0.20	0.06	3.33
14	X			75.0	0.20	0.25	0.80
15	XI			93.1	0.10	0.09	1.11
16			16	47.8	0.23	0.53	0.43
17*			3	39.7	0.44	1.17	0.38

*Only coding regions of exons 6 and 17 were used in this analysis. Eukaryote protein kinase (**ePK**) domain are restricted to exons 8-15 according to Hanks and Hunter (1995). Similarity at the amino-acid level and nucleic acid alignment for dn/ds calculation based on the protein alignment shown in the figure 3.37.

A detailed look at the dn/ds ratios for each exon shows some interesting fluctuations. For example, exon 13 and in much lesser extent exon 15, which code for the last subdomains N-terminal to the catalytic domain, exhibit dn/ds > 1, suggesting that these regions are under positive selection, although the identity at the amino-acid level coded by these exons is still relatively high. Exons 9 and 10, which code for the intermediate domains between the ATP-binding and the phosphorylation regions exhibit particular low dn/ds (0.08 and 0.12 respectively), suggesting that very few amino-acid substitutions are tolerated here. This is

not surprising, since subdomain V bridges the two major lobes, the ATP binding lobe and the substrate recognition/phosphorylation lobe, and hence plays a role of paramount importance in the conservation of the canonical kinase folding. Exons outside the catalytic domain showing significantly less identity at the amino-acid level (exon 6: 54.4%, exon 16: 47.8% and exon 17: 39.7%) and considerable numbers of indels (exon 6: 4 gaps, exon 16: 16 gaps and exon 17: 3 gaps), but still exhibit dn/ds ratios pointing to a purifying selection. Clearly, some regions of the catalytic domain are under positive selection and others do show a remarkable variability in human-mouse comparison at the sequence level. In even more distantly related organisms, regions outside the catalytic domain are dificult to recognise by simple similarity search. But overall, it may be argued that STK33 function has been already stabilised after an ancient gene duplication event (Ohno 1970) and hence the gene is under purifying selection.

4.4.7 On *STK33/Stk33* function and its medical significance

"*The reversible phosphorylation of proteins regulates almost all aspects of cell life, while abnormal phosphorylation is a cause or consequence of many diseases*" (Cohen 2001). Synthetic kinase inhibitors, which target aberrant kinases involved in diseases, have found their way out of clinic trials to become registered drugs. Perhaps the most remarkable example is imatinib mesylate (Gleevec, Novartis Pharmaceuticals, Basel Switzerland), a designer kinase inhibitor which selectively blocks the ATP-binding site of the aberrant activated tyrosine protein kinase Bcr-Abl produced in patients carrying the Philadelphia Chromosome, a translocation between chromosomes 9 and 22 (t(q;22)(q34;q11)) (Nowell

and Hungerford 1960; Rowley 1973). This chromosome is present in 90% of patients with chronic myelogenous leukemia and some persons with acute leukemia (Kurzrock et al. 2003) for a review). 94% of patients (n = 553) taking Gleevec, exhibited complete hematologic response and 69% exhibited complete cytogenetic response in a clinical trial of phase III (Sawyers et al. 2002).

Similarly, several major diseases including some types of cancer, diabetes and chronic inflamatory diseases have been associated with abnormal phosphorylation (Cohen 2001), and several different inhibitors targeting kinases and phosphatases are in various phases of research and trial (Courtneidge and Plowman 1998; Sachsenmaier 2001; Shapiro and Harper 1999). A query to the current Clinical Trial database of the National Institute of Health in the USA, produced 128 trials matching the words [kinase AND cancer]; 28 of them in phase III, the last step before approval by the Food and Drug Administration (FDA) of the USA.

Members of major group tyrosine kinases have been frequently found to act as *oncogenes,* genes "causing" cancer; while serine/threonine kinases are frequently *tumor suppresors,* genes whose loss of function is associated with increased risk of cancer (Sachsenmaier 2001). Cyclin-dependent kinases (a family of serine/threonine kinases) are obvious oncogene candidates owing to their roles in initiation, progression and completion of cell division (Shapiro and Harper 1999). Members of the Aurora family are involved with the centrosomes and mitotic spindle in dividing cells and their overexpression seems to be associated with many primary tumors (Bischoff et al. 1998; Stenoien et al. 2003).

Several human diseases, including cancer predisposition disorders, have been mapped to the chromosome region 11p15.3 (Bepler and Koehler 1995; Redeker et al. 1995). Regions of loss of heterozygocity in chromosome 11p14-15 have been associated with ovarian cancer (Wang 2002) and lung cancer (Kohno and Yokota 1999). Tumor suppressor genes have been already identified in these regions; although a role of STK33 in these or other neoplasias can not be discarded and should be investigated.

The strong expression of *STK33* in testis and the observation of down-regulation in ovarian cancer may support the idea that the protein product plays a role in gamete development. Related members of the the CAMK family of serine/threonine kinases have been associated with spermatogenesis like the *CamKII* (Guo et al., 2004) and *CamKIV* (Blaeser et al. 2001). *CamKIV* has also shown to be associated with epithelial ovarian cancer (Takai et al. 2002) and mice defective in *CamKIV* has leaded to reduced fertility in females (Wu et al. 2000a) and impaired spermiogenesis in males (Wu et al. 2000b). *CamKII* isoform δ is down-regulated in both human and mouse tumor cells (Tombes et al. 1999). Interesting also, the case of *PhKgT* gene (phosphorylase kinase, testis/liver, gamma-2; PHKG2), which codes for the closest relative of STK33 in our phylogenetic analysis. *PhKgT* was originally found to be mainly expressed in testes (Hanks 1989) but has subsequently been shown to be associated with hepatic disorders (Burwinkel et al. 1998; Maichele et al. 1996). DAP-kinases are serine/threonine kinases involved in apoptosis that phosphoprylate myosin light chains in a calmodulin-dependent way and are associated with the cytoeskeleton (Krebs 1998).

A recent study on human heart malformations using DNA-micro arrays, shows that *STK33* is among the down-regulated genes in patients with Tetralogy of Fallot, a nonfatal congenital heart condition in which bottom ventricles are not fully separated through the septum and the pulmonary artery valve is narrowed, causing a partial mixing of venal and arterial blood and a decrease of blood flow to the lungs (Kaynak et al. 2003). This observation is consistent with the results shown in this work. *STK33*-expression in adult heart, both in human and mouse were low, but the third highest signal in the MTE array (section 3.3.2.c, figure 3.28) corresponds to foetal heart. *STK33* could be involved in the normal development of heart and other organs in embryonic and foetal stadiums.

A role of *STK33/Stk33* in cell division is feasible, based on the expression pattern in normal tissues, the down-regulation in cancer from certain tissues, as well as from the structural and phylogenetical analysis shown in this work. Although further experiments are required, it is attractive to speculate about a function of STK33 in carcinogenesis, since the chromosomal region in which the gene is located, is known to harbour human suppressor genes. Even in the less optimistic scenario, where STK33 down-regulation is a consequence and not a cause of cancer, for example through genomic instability of chromosomal region 11p15, and is not associated with the risk of disease through inactivation (i.e., is not a tumor suppressor gene) STK33 may still be useful as a marker for diagnosis.

5 Summary

The comparative genomic sequence analysis of a region in human chromosome 11p15.3 and its homologous segment in mouse chromosome 7 between *ST5* and *LMO1* genes has been performed. 158,201 bases were sequenced in the mouse and compared with the syntenic region in human, partially available in the public databases. The analysed region exhibits the typical eukaryotic genomic structure and compared with the close neighbouring regions, strikingly reflexes the mosaic pattern distribution of (G+C) and repeats content despites its relative short size.

Within this region the novel gene *STK33* was discovered (*Stk33* in the mouse), that codes for a serine/threonine kinase. The finding of this gene constitutes an excellent example of the strength of the comparative sequencing approach. Poor gene-predictions in the mouse genomic sequence were corrected and improved by the comparison with the unordered data from the human genomic sequence publicly available. Phylogenetical analysis suggests that STK33 belongs to the calcium/calmodulin-dependent protein kinases group and seems to be a novelty in the chordate lineage. The gene, as a whole, seems to evolve under purifying selection whereas some regions appear to be under strong positive selection. Both human and mouse versions of *serine/threonine kinase 33*, consists of seventeen exons highly conserved in the coding regions, particularly in those coding for the core protein kinase domain. Also the exon/intron structure in the coding regions of the gene is conserved between human and mouse.

The existence and functionality of the gene is supported by the presence of entries in the EST databases and was *in vivo* fully confirmed by isolating specific transcripts from human uterus total RNA and from several mouse tissues. Strong evidence for alternative splicing was found, which may result in tissue-specific starting points of transcription and in some extent, different protein N-termini. RT-PCR and hybridisation experiments suggest that *STK33/Stk33* is differentially expressed in a few tissues and in relative low levels. *STK33* has been shown to be reproducibly down-regulated in tumor tissues, particularly in ovarian tumors. RNA in-situ hybridisation experiments using mouse *Stk33*-specific probes showed expression in dividing cells from lung and germinal epithelium and possibly also in macrophages from kidney and lungs. Preliminary experimentation with antibodies designed in this work, performed in parallel to the preparation of this manuscript, seems to confirm this expression pattern.

The fact that the chromosomal region 11p15 in which *STK33* is located may be associated with several human diseases including tumor development, suggest further investigation is necessary to establish the role of *STK33* in human health.

Vergleichende Genomanalyse einer Region des menschlichen Chromosoms 11p15.3/Maus-Chromosoms 7 und Analyse des neuen *STK33/Stk33* Gens.

In der vorliegenden Arbeit wurde eine vergleichende Genomanalyse der humanen Chromosomenregion 11p15.3 mit dem homologen Bereich des murinen Chromosoms 7 zwischen den Genen *ST5* und *LMO1* durchgeführt. 158.201 Basenpaare der Maus wurden sequenziert und mit dem syntänen Bereich im Mensch-Genom verglichen, der zum Teil selbst sequenziert wurde und zum Teil in den Datenbanken verfügbar war. Die analysierten Bereiche weisen die typischen strukturellen Eigenschaften eines eukaryotischen Genoms auf, z.B. im Hinblick auf (G+C)-Gehalt und repetitive Sequenzen.

Innerhalb dieser Region wurde ein für eine Serin/Threonin-Proteinkinase kodierendes neues Gen gefunden, das nach den Regeln des Human Genome Organisation Nomenclature Committee *STK33* in Mensch und *Stk33* in Maus genannt wurde. Der Nachweis dieses Gens zeigt die Effektivität der vergleichenden Genomanalyse-Strategie. Schwache Genvorhersage-Ergebnisse in der genomischen Sequenz der Maus wurden durch vergleichende Analyse mit der unvollständigen Sequenz des menschlichen Genoms korrigiert. Daraus konnten die vollständigen *STK33/Stk33*-Transkripte erschlossen werden , die durch Laborexperimente bestätigt wurden. Phylogenetische Analysen zeigen, dass STK33/Stk33 zu der Calcium/Calmodulin-abhängigen Protein-Kinasen Gruppe gehört und eventuell ein Novum der Chordaten ist. Das Gen scheint unter reinigender Selektion, ein Bereich sogar unter starker positiver Selektion zu evolvieren. Beide Versionen der Serin/Threonin-Kinase-33 in Mensch und Maus bestehen aus siebzehn Exonen, die hoch konserviert in den kodierenden Bereichen vorliegen, allen voran die für die Kinase-Domäne.

Auch die Exon/Intron-Struktur ist in den kodierenden Bereichen zwischen Mensch und Maus gut konserviert.

Der Nachweis der Expression des Gens wird durch Einträge in den EST-Datenbanken unterstützt und wurde *in vivo* durch Isolierung spezifischer Transkripte aus Gesamt-RNA des menschlichen Uterus und aus mehreren murinen-Geweben komplett bestätigt. Es wurden starke Hinweise auf differentielles Spleißen gefunden, was eventuell für gewebsspezifische Transkriptions-Startpunkte und verschiedene Protein-N-termini sprechen könnte. RT-PCR und Hybridisierungs-Experimente beweisen, dass *STK33/Stk33* in nur wenigen Geweben und in relativ geringen Mengen exprimiert wird. *STK33* ist in einigen Tumor-Geweben herunterreguliert, insbesondere in Ovarial-Tumoren. RNA *in situ* Hybridisierungs-Experimente mit Maus-spezifischen Sonden zeigen, dass *Stk33* in teilenden Zellen aus Lungengewebe und des Germinalepithels aktiv ist und möglicherweise auch in Nieren- und Lungen-Makrophagen. Erste Tests mit Antikörpern, die in dieser Arbeit entworfen wurden, bestätigen dieses Expressionsmuster.

Die Tatsache, dass der chromosomale Bereich 11p15, wo *STK33* lokalisiert ist, mit mehreren menschlichen Krankheiten inklusive Tumorentwicklung in Verbindung gebracht wird, zeigt deutlich, dass weitere Untersuchungen nötig sind, um den Zusammenhang von *STK33* mit der menschlichen Gesundheit zu klären.

6 Bibliography

Aach, J., M.L. Bulyk, G.M. Church, J. Comander, A. Derti, and J. Shendure. 2001. Computational comparison of two draft sequences of the human genome. *Nature* **409:** 856-859.

Adams, M.D. S.E. Celniker R.A. Holt C.A. Evans J.D. Gocayne P.G. Amanatides S.E. Scherer P.W. Li R.A. Hoskins R.F. Galle R.A. George S.E. Lewis S. Richards M. Ashburner S.N. Henderson G.G. Sutton J.R. Wortman M.D. Yandell Q. Zhang L.X. Chen R.C. Brandon Y.H. Rogers R.G. Blazej M. Champe B.D. Pfeiffer K.H. Wan C. Doyle E.G. Baxter G. Helt C.R. Nelson G.L. Gabor J.F. Abril A. Agbayani H.J. An C. Andrews-Pfannkoch D. Baldwin R.M. Ballew A. Basu J. Baxendale L. Bayraktaroglu E.M. Beasley K.Y. Beeson P.V. Benos B.P. Berman D. Bhandari S. Bolshakov D. Borkova M.R. Botchan J. Bouck P. Brokstein P. Brottier K.C. Burtis D.A. Busam H. Butler E. Cadieu A. Center I. Chandra J.M. Cherry S. Cawley C. Dahlke L.B. Davenport P. Davies B. de Pablos A. Delcher Z. Deng A.D. Mays I. Dew S.M. Dietz K. Dodson L.E. Doup M. Downes S. Dugan-Rocha B.C. Dunkov P. Dunn K.J. Durbin C.C. Evangelista C. Ferraz S. Ferriera W. Fleischmann C. Fosler A.E. Gabrielian N.S. Garg W.M. Gelbart K. Glasser A. Glodek F. Gong J.H. Gorrell Z. Gu P. Guan M. Harris N.L. Harris D. Harvey T.J. Heiman J.R. Hernandez J. Houck D. Hostin K.A. Houston T.J. Howland M.H. Wei C. Ibegwam M. Jalali F. Kalush G.H. Karpen Z. Ke J.A. Kennison K.A. Ketchum B.E. Kimmel C.D. Kodira C. Kraft S. Kravitz D. Kulp Z. Lai P. Lasko Y. Lei A.A. Levitsky J. Li Z. Li Y. Liang X. Lin X. Liu B. Mattei T.C. McIntosh M.P. McLeod D. McPherson G. Merkulov N.V. Milshina C. Mobarry J. Morris A. Moshrefi S.M. Mount M. Moy B. Murphy L. Murphy D.M. Muzny D.L. Nelson D.R. Nelson K.A. Nelson K. Nixon D.R. Nusskern J.M. Pacleb M. Palazzolo G.S. Pittman S. Pan J. Pollard V. Puri M.G. Reese K. Reinert K. Remington R.D. Saunders F. Scheeler H. Shen B.C. Shue I. Siden-Kiamos M. Simpson M.P. Skupski T. Smith E. Spier A.C. Spradling M. Stapleton R. Strong E. Sun R. Svirskas C. Tector R. Turner E. Venter A.H. Wang X. Wang Z.Y. Wang D.A. Wassarman G.M. Weinstock J. Weissenbach S.M. Williams WoodageT K.C. Worley D. Wu S. Yang Q.A. Yao J. Ye R.F. Yeh J.S. Zaveri M. Zhan G. Zhang Q. Zhao L. Zheng X.H. Zheng F.N. Zhong W. Zhong X. Zhou S. Zhu X. Zhu H.O. Smith R.A. Gibbs E.W. Myers G.M. Rubin and J.C. Venter. 2000. The genome sequence of *Drosophila melanogaster*. *Science* **287:** 2185-2195.

Adams, M.D., G.G. Sutton, H.O. Smith, E.W. Myers, and J.C. Venter. 2003. The independence of our genome assemblies. *PNAS* **100:** 3025-3026.

Alberts, B. 2002. *Molecular biology of the cell*. Garland Science, New York.

Altschul, S.F., T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25:** 3389-3402.

Amid, C. 2002. Vergleichende Genomanalyse: Die Primärstrukture der Chromosomenregion 11p15.3 des Menschen und des homologen Abschnitts der Maus. In *Fachbereich Biologie, Institut für Molekulargenetik*, pp. 189. Johannes-Gutenberg-Universität Mainz, Mainz.

Amid, C., A. Bahr, A. Mujica, N. Sampson, S.E. Bikar, A. Winterpacht, B. Zabel, T. Hankeln, and E.R. Schmidt. 2001. Comparative genomic sequencing reveals a strikingly similar architecture of a conserved syntenic region on human chromosome 11p15.3 (including gene *ST5*) and mouse chromosome 7. *Cytogenet Cell Genet* **93:** 284-290.

Angel, J.M., J.L. Moore, A. Pelphrey, and E.R. Richie. 1993. The mouse homolog of the rhombotin (*Ttg-1*) gene maps on chromosome 7 distal to the beta-globin (*Hbb*) locus. *Mamm Genome* **4:** 281-282.

Ansari-Lari, M.A., J.C. Oeltjen, S. Schwartz, Z. Zhang, D.M. Muzny, J. Lu, J.H. Gorrell, A.C. Chinault, J.W. Belmont, W. Miller, and R.A. Gibbs. 1998. Comparative sequence analysis of a gene-rich cluster at human chromosome 12p13 and its syntenic region in mouse chromosome 6. *Genome Res* **8:** 29-40.

Aparicio, S., J. Chapman, E. Stupka, N. Putnam, J.M. Chia, P. Dehal, A. Christoffels, S. Rash, S. Hoon, A. Smit, M.D. Gelpke, J. Roach, T. Oh, I.Y. Ho, M. Wong, C. Detter, F. Verhoef, P. Predki, A. Tay, S. Lucas, P. Richardson, S.F. Smith, M.S. Clark, Y.J. Edwards, N. Doggett, A. Zharkikh, S.V. Tavtigian, D. Pruss, M. Barnstead, C. Evans, H. Baden, J. Powell, G. Glusman, L. Rowen, L. Hood, Y.H. Tan, G. Elgar, T. Hawkins, B. Venkatesh, D. Rokhsar, and S. Brenner. 2002. Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science* **297:** 1301-1310.

Bahr, A. 1999. Die molekulare Struktur der Chromosomenregion 11p15.3 des Menschen und des homologen Abschnitts der Maus: Nukleotidsequenz, neue Gene und Interspeziesvergleich. In *Fachbereich Biologie, Institut für Molekulargenetik*, pp. 179. Johannes-Gutenberg-Universität Mainz, Mainz.

Bepler, G. and A. Koehler. 1995. Multiple chromosomal aberrations and 11p allelotyping in lung cancer cell lines. *Cancer Genet Cytogenet* **84:** 39-45.

Berg, D.E. and M.M. Howe. 1989. *Mobile DNA*. American Society for Microbiology, Washington, D.C.

Bergman, C.M., B.D. Pfeiffer, D.E. Rincon-Limas, R.A. Hoskins, A. Gnirke, C.J. Mungall, A.M. Wang, B. Kronmiller, J. Pacleb, S. Park, M. Stapleton, K. Wan, R.A. George, P.J. de Jong, J. Botas, G.M. Rubin, and S.E. Celniker. 2002. Assessing the impact of comparative genomic sequence data on the functional annotation of the *Drosophila* genome. *Genome Biol* **3:** RESEARCH0086.

Bernal, A., U. Ear, and N. Kyrpides. 2001. Genomes OnLine Database (GOLD): a monitor of genome projects world-wide. *Nucleic Acids Res* **29:** 126-127.

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

Bernardi, G. 2001. Misunderstandings about isochores. Part 1. *Gene* **276:** 3-13.

Bischoff, J.R., L. Anderson, Y. Zhu, K. Mossie, L. Ng, B. Souza, B. Schryver, P. Flanagan, F. Clairvoyant, C. Ginther, C.S. Chan, M. Novotny, D.J. Slamon, and G.D. Plowman. 1998. A homologue of *Drosophila* aurora kinase is oncogenic and amplified in human colorectal cancers. *Embo J* **17:** 3052-3065.

Blaeser, F., J. Toppari, M. Heikinheimo, W. Yan, M. Wallace, N. Ho, and T.A. Chatila. 2001. CaMKIV/Gr is dispensable for spermatogenesis and CREM-regulated transcription in male germ cells. *Am J Physiol Endocrinol Metab* **281:** E931-937.

Boehm, T., R. Baer, I. Lavenir, A. Forster, J.J. Waters, E. Nacheva, and T.H. Rabbitts. 1988. The mechanism of chromosomal translocation t(11;14) involving the T-cell receptor C delta locus on human chromosome 14q11 and a transcribed region of chromosome 11p15. *Embo J* **7:** 385-394.

Bohannon, J. 2002. Bioinformatics. The human genome in 3D, at your fingertips. *Science* **298:** 737.

Brauksiepe, B. 2003. Expressionsanalyse des neuen Gens *Stk33* in der Maus. In *Fachbereich Biologie, Institut für Molekulargenetik*, pp. 167. Johannes-Gutenberg-Universität Mainz, Mainz.

Brett, D., H. Pospisil, J. Valcarcel, J. Reich, and P. Bork. 2002. Alternative splicing and genome complexity. *Nat Genet* **30:** 29-30.

Brodbeck, D., M.M. Hill, and B.A. Hemmings. 2001. Two splice variants of protein kinase B gamma have different regulatory capacity depending on the presence or absence of the regulatory phosphorylation site serine 472 in the carboxyl-terminal hydrophobic domain. *J Biol Chem* **276:** 29550-29558.

Brown, T.A. 1999. *Genomes*. Bios Scientific Publishers : Wiley-Liss, New York.

Burgess, H.A. and O. Reiner. 2002. Alternative splice variants of doublecortin-like kinase are differentially expressed and have different kinase activities. *J Biol Chem* **277:** 17696-17705.

Burwinkel, B., S. Shiomi, A. Al Zaben, and M.W. Kilimann. 1998. Liver glycogenosis due to phosphorylase kinase deficiency: PHKG2 gene structure and mutations associated with cirrhosis. *Hum Mol Genet* **7:** 149-154.

C. elegans Sequencing Consortium, The. 1998. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282:** 2012-2018.

Cartegni, L., S.L. Chew, and A.R. Krainer. 2002. Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nat Rev Genet* **3:** 285-298.

Chalfant, C.E., H. Mischak, J.E. Watson, B.C. Winkler, J. Goodnight, R.V. Farese, and D.R. Cooper. 1995. Regulation of alternative splicing of protein kinase C beta by insulin. *J Biol Chem* **270:** 13326-13332.

Chomczynski, P. and N. Sacchi. 1987. Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. *Anal Biochem* **162:** 156-159.

Cichutek, A. 2001. Vergleichende Sequenzierung und Analyse eines ca. 250 kb großen Bereiches der humanen Chromsomenregion 11p15.3 und der homologen Region der Maus. In *Fachbereich Biologie, Molekulargenetisches Labor der Kinderklinik*, pp. 148. Johannes-Gutenberg-Universität Mainz, Mainz.

Cichutek, A., T. Brueckmann, B. Seipel, H. Hauser, S. Schlaubitz, D. Prawitt, T. Hankeln, E.R. Schmidt, A. Winterpacht, and B.U. Zabel. 2001. Comparative architectural aspects of regions of conserved synteny on human chromosome 11p15.3 and mouse chromosome 7 (including genes WEE1 and LMO1). *Cytogenet Cell Genet* **93:** 277-283.

Cohen, P. 2001. The role of protein phosphorylation in human health and disease. The Sir Hans Krebs Medal Lecture. *Eur J Biochem* **268:** 5001-5010.

Cokol, M., R. Nair, and B. Rost. 2000. Finding nuclear localization signals. *EMBO Rep* **1:** 411-415.

Collins, F.S., M. Morgan, and A. Patrinos. 2003a. The Human Genome Project: lessons from large-scale biology. *Science* **300:** 286-290.

Collins, J.E., M.E. Goward, C.G. Cole, L.J. Smink, E.J. Huckle, S. Knowles, J.M. Bye, D.M. Beare, and I. Dunham. 2003b. Reevaluating human gene annotation: a second-generation analysis of chromosome 22. *Genome Res* **13:** 27-36.

Courtneidge, S.A. and G.D. Plowman. 1998. The discovery and validation of new drug targets in cancer. *Curr Opin Biotechnol* **9:** 632-636.

Cozzarelli, N.R. 2003. Revisiting the independence of the publicly and privately funded drafts of the human genome. *Proc Natl Acad Sci U S A* **100:** 3021.

Cuny, G., P. Soriano, G. Macaya, and G. Bernardi. 1981. The major components of the mouse and human genomes. 1. Preparation, basic properties and compositional heterogeneity. *Eur J Biochem* **115:** 227-233.

Daly, M.J. 2002. Estimating the human gene count. *Cell* **109:** 283-284.

Dehal, P., P. Predki, A.S. Olsen, A. Kobayashi, P. Folta, S. Lucas, M. Land, A. Terry, C.L. Ecale Zhou, S. Rash, Q. Zhang, L. Gordon, J. Kim, C. Elkin, M.J. Pollard, P. Richardson, D. Rokhsar, E. Uberbacher, T. Hawkins, E. Branscomb, and L. Stubbs. 2001. Human chromosome 19 and related regions in mouse: conservative and lineage-specific evolution. *Science* **293:** 104-111.

Dehal, P., Y. Satou, R.K. Campbell, J. Chapman, B. Degnan, A. De Tomaso, B. Davidson, A. Di Gregorio, M. Gelpke, D.M. Goodstein, N. Harafuji, K.E. Hastings, I. Ho, K. Hotta, W. Huang, T. Kawashima, P. Lemaire, D. Martinez, I.A. Meinertzhagen, S. Necula, M. Nonaka, N. Putnam, S. Rash, H. Saiga, M. Satake, A. Terry, L. Yamada, H.G. Wang, S. Awazu, K. Azumi, J. Boore, M. Branno, S. Chin-Bow, R. DeSantis, S. Doyle, P. Francino, D.N. Keys, S. Haga, H. Hayashi, K. Hino, K.S. Imai, K. Inaba, S. Kano, K. Kobayashi, M. Kobayashi, B.I. Lee, K.W. Makabe, C. Manohar, G. Matassi, M. Medina, Y. Mochizuki, S. Mount, T. Morishita, S. Miura, A. Nakayama, S. Nishizaka, H. Nomoto, F. Ohta, K. Oishi, I. Rigoutsos, M. Sano, A. Sasaki, Y. Sasakura, E. Shoguchi, T. Shin-i, A. Spagnuolo, D. Stainier, M.M. Suzuki, O. Tassy, N. Takatori, M. Tokuoka, K. Yagi, F. Yoshizaki, S. Wada, C. Zhang, P.D. Hyatt, F. Larimer, C. Detter, N. Doggett, T. Glavina, T. Hawkins, P. Richardson, S. Lucas, Y. Kohara, M. Levine, N. Satoh, and D.S. Rokhsar. 2002. The draft genome of *Ciona intestinalis*: insights into chordate and vertebrate origins. *Science* **298:** 2157-2167.

DeSilva, U., L. Elnitski, J.R. Idol, J.L. Doyle, W. Gan, J.W. Thomas, S. Schwartz, N.L. Dietrich, S.M. Beckstrom-Sternberg, J.C. McDowell, R.W. Blakesley, G.G. Bouffard, P.J. Thomas, J.W. Touchman, W. Miller, and E.D. Green. 2002. Generation and comparative analysis of approximately 3.3 Mb of mouse genomic sequence orthologous to the region of human chromosome 7q11.23 implicated in Williams syndrome. *Genome Res* **12:** 3-15.

Dunham, A. L.H. Matthews J. Burton J.L. Ashurst K.L. Howe K.J. Ashcroft D.M. Beare D.C. Burford S.E. Hunt S. Griffiths-Jones M.C. Jones S.J. Keenan K. Oliver C.E. Scott R. Ainscough J.P. Almeida K.D. Ambrose D.T. Andrews R.I.S. Ashwell A.K. Babbage C.L. Bagguley J. Bailey R. Bannerjee K.F. Barlow K. Bates H. Beasley C.P. Bird S. Bray-Allen A.J. Brown J.Y. Brown W. Burrill C. Carder N.P. Carter J.C. Chapman M.E. Clamp S.Y. Clark G. Clarke C.M. Clee S.C.M. Clegg V. Cobley J.E. Collins N. Corby G.J. Coville P. Deloukas P. Dhami I. Dunham M. Dunn M.E. Earthrowl A.G. Ellington L. Faulkner A.G. Frankish J. Frankland L. French P. Garner J. Garnett J.G.R. Gilbert C.J. Gilson J. Ghori D.V. Grafham S.M. Gribble C. Griffiths R.E. Hall S. Hammond J.L. Harley E.A. Hart P.D. Heath P.J. Howden E.J. Huckle P.J. Hunt A.R. Hunt C. Johnson D. Johnson M. Kay A.M. Kimberley A. King G.K. Laird C.J. Langford S. Lawlor D.A. Leongamornlert D.M. Lloyd C. Lloyd J.E. Loveland J. Lovell S. Martin M. Mashreghi-Mohammadi S.J. McLaren A. McMurray S. Milne M.J.F. Moore T. Nickerson S.A. Palmer A.V. Pearce A.I. Peck S. Pelan B. Phillimore K.M. Porter C.M. Rice S. Searle H.K. Sehra R. Shownkeen C.D. Skuce M. Smith C.A. Steward N. Sycamore J. Tester D.W. Thomas A. Tracey A. Tromans B. Tubby M. Wall J.M. Wallis A.P. West S.L. Whitehead D.L. Willey L. Wilming P.W. Wray M.W. Wright L. Young A. Coulson R. Durbin T. Hubbard J.E. Sulston S. Beck D.R. Bentley J. Rogers and M.T. Ross. 2004. The DNA sequence and analysis of human chromosome 13. *Nature* **428:** 522-528.

Falquet, L., M. Pagni, P. Bucher, N. Hulo, C.J. Sigrist, K. Hofmann, and A. Bairoch. 2002. The PROSITE database, its status in 2002. *Nucleic Acids Res* **30:** 235-238.

Fraser, C.M., J.D. Gocayne, O. White, M.D. Adams, R.A. Clayton, R.D. Fleischmann, C.J. Bult, A.R. Kerlavage, G. Sutton, J.M. Kelley, and et al. 1995. The minimal gene complement of *Mycoplasma genitalium*. *Science* **270:** 397-403.

Frazer, K.A., L. Elnitski, D.M. Church, I. Dubchak, and R.C. Hardison. 2003. Cross-species sequence comparisons: a review of methods and available resources. *Genome Res* **13:** 1-12.

Genereux, D.P. and J.M. Logsdon, Jr. 2003. Much ado about bacteria-to-vertebrate lateral gene transfer. *Trends Genet* **19:** 191-195.

Germani, A., S. Malherbe, and E. Rouer. 2003. The exon 7-spliced Lck isoform in T lymphocytes: a potential regulator of p56lck signaling pathways. *Biochem Biophys Res Commun* **301:** 680-685.

Gibbs, R.A. G.M. Weinstock M.L. Metzker D.M. Muzny E.J. Sodergren S. Scherer G. Scott D. Steffen K.C. Worley P.E. Burch G. Okwuonu S. Hines L. Lewis C. DeRamo O. Delgado S. Dugan-Rocha G. Miner M. Morgan A. Hawes R. Gill Celera R.A. Holt M.D. Adams P.G. Amanatides H. Baden-Tillson M. Barnstead S. Chin C.A. Evans S. Ferriera C. Fosler A. Glodek Z. Gu D. Jennings C.L. Kraft T. Nguyen C.M. Pfannkoch C. Sitter G.G. Sutton J.C. Venter T. Woodage D. Smith H.M. Lee E. Gustafson P. Cahill A. Kana L. Doucette-Stamm K. Weinstock K. Fechtel R.B. Weiss D.M. Dunn E.D. Green R.W. Blakesley G.G. Bouffard P.J. De Jong K. Osoegawa B. Zhu M. Marra J. Schein I. Bosdet C. Fjell S. Jones M. Krzywinski C. Mathewson A. Siddiqui N. Wye J. McPherson S. Zhao C.M. Fraser J. Shetty S. Shatsman K. Geer Y. Chen S. Abramzon W.C. Nierman P.H. Havlak R. Chen K.J. Durbin A. Egan Y. Ren X.Z. Song B. Li Y. Liu X. Qin S. Cawley A.J. Cooney L.M. D'Souza K. Martin J.Q. Wu M.L. Gonzalez-Garay A.R. Jackson K.J. Kalafus M.P. McLeod A. Milosavljevic D. Virk A. Volkov D.A. Wheeler Z. Zhang J.A. Bailey E.E. Eichler E. Tuzun E. Birney E. Mongin A.

Ureta-Vidal C. Woodwark E. Zdobnov P. Bork M. Suyama D. Torrents M. Alexandersson B.J. Trask J.M. Young H. Huang H. Wang H. Xing S. Daniels D. Gietzen J. Schmidt K. Stevens U. Vitt J. Wingrove F. Camara M. Mar Alba J.F. Abril R. Guigo A. Smit I. Dubchak E.M. Rubin O. Couronne A. Poliakov N. Hubner D. Ganten C. Goesele O. Hummel T. Kreitler Y.A. Lee J. Monti H. Schulz H. Zimdahl H. Himmelbauer H. Lehrach H.J. Jacob S. Bromberg J. Gullings-Handley M.I. Jensen-Seaman A.E. Kwitek J. Lazar D. Pasko P.J. Tonellato S. Twigger C.P. Ponting J.M. Duarte S. Rice L. Goodstadt S.A. Beatson R.D. Emes E.E. Winter C. Webber P. Brandt G. Nyakatura M. Adetobi F. Chiaromonte L. Elnitski P. Eswara R.C. Hardison M. Hou D. Kolbe K. Makova W. Miller A. Nekrutenko C. Riemer S. Schwartz J. Taylor S. Yang Y. Zhang K. Lindpaintner T.D. Andrews M. Caccamo M. Clamp L. Clarke V. Curwen R. Durbin E. Eyras S.M. Searle G.M. Cooper S. Batzoglou M. Brudno A. Sidow E.A. Stone B.A. Payseur G. Bourque C. Lopez-Otin X.S. Puente K. Chakrabarti S. Chatterji C. Dewey L. Pachter N. Bray V.B. Yap A. Caspi G. Tesler P.A. Pevzner D. Haussler K.M. Roskin R. Baertsch H. Clawson T.S. Furey A.S. Hinrichs D. Karolchik W.J. Kent K.R. Rosenbloom H. Trumbower M. Weirauch D.N. Cooper P.D. Stenson B. Ma M. Brent M. Arumugam D. Shteynberg R.R. Copley M.S. Taylor H. Riethman U. Mudunuri J. Peterson M. Guyer A. Felsenfeld S. Old S. Mockrin and F. Collins. 2004. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428:** 493-521.

Grimwood, J., L.A. Gordon, A. Olsen, A. Terry, J. Schmutz, J. Lamerdin, U. Hellsten, D. Goodstein, O. Couronne, M. Tran-Gyamfi, A. Aerts, M. Altherr, L. Ashworth, E. Bajorek, S. Black, E. Branscomb, S. Caenepeel, A. Carrano, C. Caoile, Y.M. Chan, M. Christensen, C.A. Cleland, A. Copeland, E. Dalin, P. Dehal, M. Denys, J.C. Detter, J. Escobar, D. Flowers, D. Fotopulos, C. Garcia, A.M. Georgescu, T. Glavina, M. Gomez, E. Gonzales, M. Groza, N. Hammon, T. Hawkins, L. Haydu, I. Ho, W. Huang, S. Israni, J. Jett, K. Kadner, H. Kimball, A. Kobayashi, V. Larionov, S.H. Leem, F. Lopez, Y. Lou, S. Lowry, S. Malfatti, D. Martinez, P. McCready, C. Medina, J. Morgan, K. Nelson, M. Nolan, I. Ovcharenko, S. Pitluck, M. Pollard, A.P. Popkie, P. Predki, G. Quan, L. Ramirez, S. Rash, J. Retterer, A. Rodriguez, S. Rogers, A. Salamov, A. Salazar, X. She, D. Smith, T. Slezak, V. Solovyev, N. Thayer, H. Tice, M. Tsai, A. Ustaszewska, N. Vo, M. Wagner, J. Wheeler, K. Wu, G. Xie, J. Yang, I. Dubchak, T.S. Furey, P. DeJong, M. Dickson, D. Gordon, E.E. Eichler, L.A. Pennacchio, P. Richardson, L. Stubbs, D.S. Rokhsar, R.M. Myers, E.M. Rubin, and S.M. Lucas. 2004. The DNA sequence and biology of human chromosome 19. *Nature* **428:** 529-535.

Gururajan, R., J.M. Lahti, J. Grenet, J. Easton, I. Gruber, P.F. Ambros, and V.J. Kidd. 1998. Duplication of a genomic region containing the Cdc2L1-2 and MMP21-22 genes on human chromosome 1p36.3 and their linkage to D1Z2. *Genome Res* **8:** 929-939.

Haeseleer, F., Y. Imanishi, I. Sokal, S. Filipek, and K. Palczewski. 2002. Calcium-binding proteins: intracellular sensors from the calmodulin superfamily. *Biochem Biophys Res Commun* **290:** 615-623.

Hanks, S.K. 1989. Messenger ribonucleic acid encoding an apparent isoform of phosphorylase kinase catalytic subunit is abundant in the adult testis. *Mol Endocrinol* **3:** 110-116.

Hanks, S.K. 2003. Genomic analysis of the eukaryotic protein kinase superfamily: a perspective. *Genome Biol* **4:** 111.

Hanks, S.K. and T. Hunter. 1995. Protein kinases 6. The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification. *Faseb J* **9:** 576-596.

Hanks, S.K. and A.M. Quinn. 1991. Protein kinase catalytic domain sequence database: identification of conserved features of primary structure and classification of family members. *Methods Enzymol* **200:** 38-62.

Hanks, S.K., A.M. Quinn, and T. Hunter. 1988. The protein kinase family: conserved features and deduced phylogeny of the catalytic domains. *Science* **241:** 42-52.

Hardison, R.C., J. Oeltjen, and W. Miller. 1997. Long human-mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome. *Genome Res* **7:** 959-966.

Hastings, I.M., P.G. Bray, and S.A. Ward. 2002. Parasitology. A requiem for chloroquine. *Science* **298:** 74-75.

Hicks, G.R. and N.V. Raikhel. 1995. Protein import into the nucleus: an integrated view. *Annu Rev Cell Dev Biol* **11:** 155-188.

Hill, F., C. Gemund, V. Benes, W. Ansorge, and T.J. Gibson. 2000. An estimate of large-scale sequencing accuracy. *EMBO Rep* **1:** 29-31.

Hogenesch, J.B., K.A. Ching, S. Batalov, A.I. Su, J.R. Walker, Y. Zhou, S.A. Kay, P.G. Schultz, and M.P. Cooke. 2001. A comparison of the Celera and Ensembl predicted gene sets reveals little overlap in novel genes. *Cell* **106:** 413-415.

Holstein, A.F., W. Schulze, and M. Davidoff. 2003. Understanding spermatogenesis is a prerequisite for treatment. *Reprod Biol Endocrinol* **1:** 107.

Holt, R.A. G.M. Subramanian A. Halpern G.G. Sutton R. Charlab D.R. Nusskern P. Wincker A.G. Clark J.M. Ribeiro R. Wides S.L. Salzberg B. Loftus M. Yandell W.H. Majoros D.B. Rusch Z. Lai C.L. Kraft J.F. Abril V. Anthouard P. Arensburger P.W. Atkinson H. Baden V. de Berardinis D. Baldwin V. Benes J. Biedler C. Blass R. Bolanos D. Boscus M. Barnstead S. Cai A. Center K. Chaturverdi G.K. Christophides M.A. Chrystal M. Clamp A. Cravchik V. Curwen A. Dana A. Delcher I. Dew C.A. Evans M. Flanigan A. Grundschober-Freimoser L. Friedli Z. Gu P. Guan R. Guigo M.E. Hillenmeyer S.L. Hladun J.R. Hogan Y.S. Hong J. Hoover O. Jaillon Z. Ke C. Kodira E. Kokoza A. Koutsos I. Letunic A. Levitsky Y. Liang J.J. Lin N.F. Lobo J.R. Lopez J.A. Malek T.C. McIntosh S. Meister J. Miller C. Mobarry E. Mongin S.D. Murphy D.A. O'Brochta C. Pfannkoch R. Qi M.A. Regier K. Remington H. Shao M.V. Sharakhova C.D. Sitter J. Shetty T.J. Smith R. Strong J. Sun D. Thomasova L.Q. Ton P. Topalis Z. Tu M.F. Unger B. Walenz A. Wang J. Wang M. Wang X. Wang K.J. Woodford J.R. Wortman M. Wu A. Yao E.M. Zdobnov H. Zhang Q. Zhao S. Zhao S.C. Zhu I. Zhimulev M. Coluzzi A. della Torre C.W. Roth C. Louis F. Kalush R.J. Mural E.W. Myers M.D. Adams H.O. Smith S. Broder M.J. Gardner C.M. Fraser E. Birney P. Bork P.T. Brey J.C. Venter J. Weissenbach F.C. Kafatos F.H. Collins and S.L. Hoffman. 2002. The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science* **298:** 129-149.

Horton, P. and K. Nakai. 1997. Better prediction of protein cellular localization sites with the k nearest neighbors classifier. *Proc Int Conf Intell Syst Mol Biol* **5:** 147-152.

Hua, S. and Z. Sun. 2001. Support vector machine approach for protein subcellular localization prediction. *Bioinformatics* **17:** 721-728.

Igarashi, M., A. Nagata, S. Jinno, K. Suto, and H. Okayama. 1991. *Wee1(+)-like* gene in human cells. *Nature* **353:** 80-83.

Ikura, M., M. Osawa, and J.B. Ames. 2002. The role of calcium-binding proteins in the control of transcription: structure to function. *Bioessays* **24:** 625-636.

Ioannou, P.A., C.T. Amemiya, J. Garnes, P.M. Kroisel, H. Shizuya, C. Chen, M.A. Batzer, and P.J. de Jong. 1994. A new bacteriophage P1-derived vector for the propagation of large human DNA fragments. *Nat Genet* **6:** 84-89.

Jongeneel, C.V., C. Iseli, B.J. Stevenson, G.J. Riggins, A. Lal, A. Mackay, R.A. Harris, M.J. O'Hare, A.M. Neville, A.J. Simpson, and R.L. Strausberg. 2003. Comprehensive sampling of gene expression in human cell lines with massively parallel signature sequencing. *Proc Natl Acad Sci U S A* **100:** 4702-4705.

Kalafus, K.J., A.R. Jackson, and A. Milosavljevic. 2004. Pash: efficient genome-scale sequence anchoring by Positional Hashing. *Genome Res* **14:** 672-678.

Karnik, P., M. Paris, B.R. Williams, G. Casey, J. Crowe, and P. Chen. 1998. Two distinct tumor suppressor loci within chromosome 11p15 implicated in breast cancer progression and metastasis. *Hum Mol Genet* **7:** 895-903.

Kaynak, B., A. von Heydebreck, S. Mebus, D. Seelow, S. Hennig, J. Vogel, H.P. Sperling, R. Pregla, V. Alexi-Meskishvili, R. Hetzer, P.E. Lange, M. Vingron, H. Lehrach, and S. Sperling. 2003. Genome-wide array analysis of normal and malformed human hearts. *Circulation* **107:** 2467-2474.

Kleyn, P.W., W. Fan, S.G. Kovats, J.J. Lee, J.C. Pulido, Y. Wu, L.R. Berkemeier, D.J. Misumi, L. Holmgren, O. Charlat, E.A. Woolf, O. Tayber, T. Brody, P. Shu, F. Hawkins, B. Kennedy, L. Baldini, C. Ebeling, G.D. Alperin, J. Deeds, N.D. Lakey, J. Culpepper, H. Chen, M.A. Glucksmann-Kuis, K.J. Moore, and et al. 1996. Identification and characterization of the mouse obesity gene *tubby*: a member of a novel gene family. *Cell* **85:** 281-290.

Knighton, D.R., R.B. Pearson, J.M. Sowadski, A.R. Means, L.F. Ten Eyck, S.S. Taylor, and B.E. Kemp. 1992. Structural basis of the intrasteric regulation of myosin light chain kinases. *Science* **258:** 130-135.

Kohno, T. and J. Yokota. 1999. How many tumor suppressor genes are involved in human lung carcinogenesis? *Carcinogenesis* **20:** 1403-1410.

Koop, B.F. 1995. Human and rodent DNA sequence comparisons: a mosaic model of genomic evolution. *Trends Genet* **11:** 367-371.

Korsmeyer, S.J. 1992. Chromosomal translocations in lymphoid malignancies reveal novel proto-oncogenes. *Annu Rev Immunol* **10:** 785-807.

Kostich, M., J. English, V. Madison, F. Gheyas, L. Wang, P. Qiu, J. Greene, and T.M. Laz. 2002. Human members of the eukaryotic protein kinase family. *Genome Biol* **3:** RESEARCH0043.

Koutny, L., D. Schmalzing, O. Salas-Solano, S. El-Difrawy, A. Adourian, S. Buonocore, K. Abbey, P. McEwan, P. Matsudaira, and D. Ehrlich. 2000. Eight hundred-base sequencing in a microfabricated electrophoretic device. *Anal Chem* **72:** 3388-3391.

Kovarik, A., M.A. Matzke, A.J. Matzke, and B. Koulakova. 2001. Transposition of IS*10* from the host *Escherichia coli* genome to a plasmid may lead to cloning artefacts. *Mol Genet Genomics* **266:** 216-222.

Krebs, J. 1998. The role of calcium in apoptosis. *Biometals* **11:** 375-382.

Krogh, A., B. Larsson, G. von Heijne, and E.L. Sonnhammer. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* **305:** 567-580.

Kurzrock, R., H.M. Kantarjian, B.J. Druker, and M. Talpaz. 2003. Philadelphia chromosome-positive leukemias: from basic mechanisms to molecular therapeutics. *Ann Intern Med* **138:** 819-830.

Lahti, J.M., J. Xiang, and V.J. Kidd. 1995. The PITSLRE protein kinase family. *Prog Cell Cycle Res* **1:** 329-338.

Lander, E.S. L.M. Linton B. Birren C. Nusbaum M.C. Zody J. Baldwin K. Devon K. Dewar M. Doyle W. FitzHugh R. Funke D. Gage K. Harris A. Heaford J. Howland L. Kann J. Lehoczky R. LeVine P. McEwan K. McKernan J. Meldrim J.P. Mesirov C. Miranda W. Morris J. Naylor C. Raymond M. Rosetti R. Santos A. Sheridan C. Sougnez N. Stange-Thomann N. Stojanovic A. Subramanian D. Wyman J. Rogers J. Sulston R. Ainscough S. Beck D. Bentley J. Burton C. Clee N. Carter A. Coulson R. Deadman P. Deloukas A. Dunham I. Dunham R. Durbin L. French D. Grafham S. Gregory T. Hubbard S. Humphray A. Hunt M. Jones C. Lloyd A. McMurray L. Matthews S. Mercer S. Milne J.C. Mullikin A. Mungall R. Plumb M. Ross R. Shownkeen S. Sims R.H. Waterston R.K. Wilson L.W. Hillier J.D. McPherson M.A. Marra E.R. Mardis L.A. Fulton A.T. Chinwalla K.H. Pepin W.R. Gish S.L. Chissoe M.C. Wendl K.D. Delehaunty T.L. Miner A. Delehaunty J.B. Kramer L.L. Cook R.S. Fulton D.L. Johnson P.J. Minx S.W. Clifton T. Hawkins E. Branscomb P. Predki P. Richardson S. Wenning T. Slezak N. Doggett J.F. Cheng A. Olsen S. Lucas C. Elkin E. Uberbacher M. Frazier R.A. Gibbs D.M. Muzny S.E. Scherer J.B. Bouck E.J. Sodergren K.C. Worley C.M. Rives J.H. Gorrell M.L. Metzker S.L. Naylor R.S. Kucherlapati D.L. Nelson G.M. Weinstock Y. Sakaki A. Fujiyama M. Hattori T. Yada A. Toyoda T. Itoh C. Kawagoe H. Watanabe Y. Totoki T. Taylor J. Weissenbach R. Heilig W. Saurin F. Artiguenave P. Brottier T. Bruls E. Pelletier C. Robert P. Wincker D.R. Smith L. Doucette-Stamm M. Rubenfield K. Weinstock H.M. Lee J. Dubois A. Rosenthal M. Platzer G. Nyakatura S. Taudien A. Rump H. Yang J. Yu J. Wang G. Huang J. Gu L. Hood L. Rowen A. Madan S. Qin R.W. Davis N.A. Federspiel A.P. Abola M.J. Proctor R.M. Myers J. Schmutz M. Dickson J. Grimwood D.R. Cox M.V. Olson R. Kaul N. Shimizu K. Kawasaki S. Minoshima G.A. Evans M. Athanasiou R. Schultz B.A. Roe F. Chen H. Pan J. Ramser H. Lehrach R. Reinhardt W.R. McCombie M. de la Bastide N. Dedhia H. Blocker K. Hornischer G. Nordsiek R. Agarwala L. Aravind J.A. Bailey A. Bateman S. Batzoglou E. Birney P. Bork D.G. Brown C.B. Burge L. Cerutti H.C. Chen D. Church M. Clamp R.R. Copley T. Doerks S.R. Eddy E.E. Eichler T.S. Furey J. Galagan J.G. Gilbert C. Harmon Y. Hayashizaki D. Haussler H. Hermjakob K. Hokamp W. Jang L.S. Johnson T.A. Jones S. Kasif A. Kaspryzk S. Kennedy W.J. Kent P. Kitts E.V. Koonin I. Korf D. Kulp D. Lancet T.M. Lowe A. McLysaght T. Mikkelsen J.V. Moran N. Mulder V.J. Pollara C.P. Ponting G. Schuler J. Schultz G. Slater A.F. Smit E. Stupka J. Szustakowski D. Thierry-Mieg J. Thierry-Mieg L. Wagner J. Wallis R. Wheeler A. Williams Y.I. Wolf K.H. Wolfe S.P. Yang R.F. Yeh F. Collins M.S. Guyer J. Peterson A. Felsenfeld K.A. Wetterstrand A. Patrinos M.J. Morgan J. Szustakowki P. de Jong J.J. Catanese K. Osoegawa H. Shizuya S. Choi and Y.J. Chen. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860-921.

Leblond, C.P. and Y. Clermont. 1952. Spermiogenesis of rat, mouse, hamster and guinea pig as revealed by the periodic acid-fuchsin sulfurous acid technique. *Am J Anat* **90:** 167-215.

Letunic, I., R.R. Copley, S. Schmidt, F.D. Ciccarelli, T. Doerks, J. Schultz, C.P. Ponting, and P. Bork. 2004. SMART 4.0: towards genomic data integration. *Nucleic Acids Res* **32 Database issue:** D142-144.

Li, W., P. Bernaola-Galvan, P. Carpena, and J.L. Oliver. 2003. Isochores merit the prefix 'iso'. *Comput Biol Chem* **27:** 5-10.

Lichy, J.H., M. Majidi, J. Elbaum, and M.M. Tsai. 1996. Differential expression of the human *ST5* gene in HeLa-fibroblast hybrid cell lines mediated by *YY1*: evidence that *YY1* plays a part in tumor suppression. *Nucleic Acids Res* **24:** 4700-4708.

Mahl, T.C. 1998. Approach to the patient with abnormal liver tests. *Lippincotts Prim Care Pract* **2:** 379-389.

Maichele, A.J., B. Burwinkel, I. Maire, O. Sovik, and M.W. Kilimann. 1996. Mutations in the testis/liver isoform of the phosphorylase kinase gamma subunit (PHKG2) cause autosomal liver glycogenosis in the gsd rat and in humans. *Nat Genet* **14:** 337-340.

Malakoff, D. 1999. NIH urged to fund centers to merge computing and biology. *Science* **284:** 1742.

Maniatis, T. and B. Tasic. 2002. Alternative pre-mRNA splicing and proteome expansion in metazoans. *Nature* **418:** 236-243.

Manning, G., G.D. Plowman, T. Hunter, and S. Sudarsanam. 2002a. Evolution of protein kinase signaling from yeast to man. *Trends Biochem Sci* **27:** 514-520.

Manning, G., D.B. Whyte, R. Martinez, T. Hunter, and S. Sudarsanam. 2002b. The protein kinase complement of the human genome. *Science* **298:** 1912-1934.

Marshall, E. 1999. A high-stakes gamble on genome sequencing. *Science* **284:** 1906-1909.

Marshall, E. 2000a. Genome sequencing. Talks of public-private deal end in acrimony. *Science* **287:** 1723-1725.

Marshall, E. 2000b. Human genome. Rival genome sequencers celebrate a milestone together. *Science* **288:** 2294-2295.

Marshall, E. 2000c. Human genome. Storm erupts over terms for publishing Celera's sequence. *Science* **290:** 2042-2043.

Mathe, C., M.F. Sagot, T. Schiex, and P. Rouze. 2002. Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Res* **30:** 4103-4117.

Mayor, C., M. Brudno, J.R. Schwartz, A. Poliakov, E.M. Rubin, K.A. Frazer, L.S. Pachter, and I. Dubchak. 2000. VISTA : visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics* **16:** 1046-1047.

Meunier-Rotival, M., P. Soriano, G. Cuny, F. Strauss, and G. Bernardi. 1982. Sequence organization and genomic distribution of the major family of interspersed repeats of mouse DNA. *Proc Natl Acad Sci U S A* **79:** 355-359.

Mewes, H.W., K. Albermann, K. Heumann, S. Liebl, and F. Pfeiffer. 1997. MIPS: a database for protein sequences, homology data and yeast genome information. *Nucleic Acids Res* **25:** 28-30.

Mironov, A.A., J.W. Fickett, and M.S. Gelfand. 1999. Frequent alternative splicing of human genes. *Genome Res* **9:** 1288-1293.

Mirsky, A.E. and H. Ris. 1951. The desoxyribonucleic acid content of animal cells and its evolutionary significance. *J Gen Physiol* **34:** 451-462.

Mitnik, L., M. Novotny, C. Felten, S. Buonocore, L. Koutny, and D. Schmalzing. 2001. Recent advances in DNA sequencing by capillary and microdevice electrophoresis. *Electrophoresis* **22:** 4104-4117.

Modrek, B. and C. Lee. 2002. A genomic view of alternative splicing. *Nat Genet* **30:** 13-19.

Moller, S., M.D. Croning, and R. Apweiler. 2001. Evaluation of methods for the prediction of membrane spanning regions. *Bioinformatics* **17:** 646-653.

Mujica, A.O., T. Hankeln, and E.R. Schmidt. 2001. A novel serine/threonine kinase gene, *STK33*, on human chromosome 11p15.3. *Gene* **280:** 175-181.

Mural, R.J. M.D. Adams E.W. Myers H.O. Smith G.L. Miklos R. Wides A. Halpern P.W. Li G.G. Sutton J. Nadeau S.L. Salzberg R.A. Holt C.D. Kodira F. Lu L. Chen Z. Deng C.C. Evangelista W. Gan T.J. Heiman J. Li Z. Li G.V. Merkulov N.V. Milshina A.K. Naik R. Qi B.C. Shue A. Wang J. Wang X. Wang X. Yan J. Ye S. Yooseph Q. Zhao L. Zheng S.C. Zhu K. Biddick R. Bolanos A.L. Delcher I.M. Dew D. Fasulo M.J. Flanigan D.H. Huson S.A. Kravitz J.R. Miller C.M. Mobarry K. Reinert K.A. Remington Q. Zhang X.H. Zheng D.R. Nusskern Z. Lai Y. Lei W. Zhong A. Yao P. Guan R.R. Ji Z. Gu Z.Y. Wang F. Zhong C. Xiao C.C. Chiang M. Yandell J.R. Wortman P.G. Amanatides S.L. Hladun E.C. Pratts J.E. Johnson K.L. Dodson K.J. Woodford C.A. Evans B. Gropman D.B. Rusch E. Venter M. Wang T.J. Smith J.T. Houck D.E. Tompkins C. Haynes D. Jacob S.H. Chin D.R. Allen C.E. Dahlke R. Sanders K. Li X. Liu A.A. Levitsky W.H. Majoros Q. Chen A.C. Xia J.R. Lopez M.T. Donnelly M.H. Newman A. Glodek C.L. Kraft M. Nodell F. Ali H.J. An D. Baldwin-Pitts K.Y. Beeson S. Cai M. Carnes A. Carver P.M. Caulk A. Center Y.H. Chen M.L. Cheng M.D. Coyne M. Crowder S. Danaher L.B. Davenport R. Desilets S.M. Dietz L. Doup P. Dullaghan S. Ferriera C.R. Fosler H.C. Gire A. Gluecksmann J.D. Gocayne J. Gray B. Hart J. Haynes J. Hoover T. Howland C. Ibegwam M. Jalali D. Johns L. Kline D.S. Ma S. MacCawley A. Magoon F. Mann D. May T.C. McIntosh S. Mehta L. Moy M.C. Moy B.J. Murphy S.D. Murphy K.A. Nelson Z. Nuri K.A. Parker A.C. Prudhomme V.N. Puri H. Qureshi J.C. Raley M.S. Reardon M.A. Regier Y.H. Rogers D.L. Romblad J. Schutz J.L. Scott R. Scott C.D. Sitter M. Smallwood A.C. Sprague E. Stewart R.V. Strong E. Suh K. Sylvester R. Thomas N.N. Tint C. Tsonis G. Wang M.S. Williams S.M. Williams S.M. Windsor K. Wolfe M.M. Wu J. Zaveri K. Chaturvedi A.E. Gabrielian Z. Ke J. Sun G. Subramanian J.C. Venter C.M. Pfannkoch M. Barnstead and L.D. Stephenson. 2002. A comparison of whole-genome shotgun-derived mouse chromosome 16 and the human genome. *Science* **296:** 1661-1671.

Nadeau, J.H. and B.A. Taylor. 1984. Lengths of chromosomal segments conserved since divergence of man and mouse. *Proc Natl Acad Sci U S A* **81:** 814-818.

Nakai, K. and P. Horton. 1999. PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem Sci* **24:** 34-36.

Nei, M. and T. Gojobori. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol* **3:** 418-426.

Nowak, N.J. and T.B. Shows. 1995. Genetics of chromosome 11: loci for pediatric and adult malignancies, developmental disorders, and other diseases. *Cancer Invest* **13:** 646-659.

Nowell, P.C. and D.A. Hungerford. 1960. Chromosome studies on normal and leukemic human leukocytes. *J Natl Cancer Inst* **25:** 85-109.

Oeltjen, J.C., T.M. Malley, D.M. Muzny, W. Miller, R.A. Gibbs, and J.W. Belmont. 1997. Large-scale comparative sequence analysis of the human and murine Bruton's tyrosine kinase loci reveals conserved regulatory domains. *Genome Res* **7:** 315-329.

Ohno, S. 1970. *Evolution by gene duplication*. Allen & Unwin; Springer-Verlag, London, New York.

Ota, T. and M. Nei. 1994. Variance and covariances of the numbers of synonymous and nonsynonymous substitutions per site. *Mol Biol Evol* **11:** 613-619.

Paegel, B.M., R.G. Blazej, and R.A. Mathies. 2003. Microfluidic devices for DNA sequencing: sample preparation and electrophoretic analysis. *Curr Opin Biotechnol* **14:** 42-50.

Pennisi, E. 1999. Academic sequencers challenge Celera in a sprint to the finish. *Science* **283:** 1822-1823.

Pennisi, E. 2000. Human Genome Project. And the gene number is...? *Science* **288:** 1146-1147.

Pennisi, E. 2002. The Institute for Genomic Research meeting. Gene researchers hunt bargains, fixer-uppers. *Science* **298:** 735-736.

Pruitt, K.D., T. Tatusova, and D.R. Maglott. 2003. NCBI Reference Sequence project: update and current status. *Nucleic Acids Res* **31:** 34-37.

Rebhan, M., Chalifa-Caspi, V., Prilusky, J., Lancet, D. 1997. GeneCards: encyclopedia for genes, proteins and diseases. Weizmann Institute of Science, Bioinformatics Unit and Genome Center. Rehovot, Israel.

Redeker, E., M. Alders, J.M. Hoovers, C.W. Richard, 3rd, A. Westerveld, and M. Mannens. 1995. Physical mapping of 3 candidate tumor suppressor genes relative to Beckwith-Wiedemann syndrome associated chromosomal breakpoints at 11p15.3. *Cytogenet Cell Genet* **68:** 222-225.

Roberts, G.C. and C.W. Smith. 2002. Alternative splicing: combinatorial output from the genome. *Curr Opin Chem Biol* **6:** 375-383.

Rowley, J.D. 1973. Letter: A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining. *Nature* **243:** 290-293.

Rubin, G.M., M.D. Yandell, J.R. Wortman, G.L. Gabor Miklos, C.R. Nelson, I.K. Hariharan, M.E. Fortini, P.W. Li, R. Apweiler, W. Fleischmann, J.M. Cherry, S. Henikoff, M.P. Skupski, S. Misra, M. Ashburner, E. Birney, M.S. Boguski, T. Brody, P. Brokstein, S.E. Celniker, S.A. Chervitz, D. Coates, A. Cravchik, A. Gabrielian, R.F. Galle, W.M. Gelbart, R.A. George, L.S. Goldstein, F. Gong, P. Guan, N.L. Harris, B.A. Hay, R.A. Hoskins, J. Li, Z. Li, R.O. Hynes, S.J. Jones, P.M. Kuehl, B. Lemaitre, J.T. Littleton, D.K. Morrison, C. Mungall, P.H. O'Farrell, O.K. Pickeral, C. Shue, L.B.

Vosshall, J. Zhang, Q. Zhao, X.H. Zheng, and S. Lewis. 2000. Comparative genomics of the eukaryotes. *Science* **287:** 2204-2215.

Sachsenmaier, C. 2001. Targeting protein kinases for tumor therapy. *Onkologie* **24:** 346-355.

Sambrook, J. and D.W. Russell. 2001. *Molecular cloning : a laboratory manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.

Sander, C. and R. Schneider. 1991. Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins* **9:** 56-68.

Sanger, F., G.M. Air, B.G. Barrell, N.L. Brown, A.R. Coulson, C.A. Fiddes, C.A. Hutchison, P.M. Slocombe, and M. Smith. 1977a. Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* **265:** 687-695.

Sanger, F., A.R. Coulson, G.F. Hong, D.F. Hill, and G.B. Petersen. 1982. Nucleotide sequence of bacteriophage lambda DNA. *J Mol Biol* **162:** 729-773.

Sanger, F., S. Nicklen, and A.R. Coulson. 1977b. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **74:** 5463-5467.

Sawyers, C.L., A. Hochhaus, E. Feldman, J.M. Goldman, C.B. Miller, O.G. Ottmann, C.A. Schiffer, M. Talpaz, F. Guilhot, M.W. Deininger, T. Fischer, S.G. O'Brien, R.M. Stone, C.B. Gambacorti-Passerini, N.H. Russell, J.J. Reiffers, T.C. Shea, B. Chapuis, S. Coutre, S. Tura, E. Morra, R.A. Larson, A. Saven, C. Peschel, A. Gratwohl, F. Mandelli, M. Ben-Am, I. Gathmann, R. Capdeville, R.L. Paquette, and B.J. Druker. 2002. Imatinib induces hematologic and cytogenetic responses in patients with chronic myelogenous leukemia in myeloid blast crisis: results of a phase II study. *Blood* **99:** 3530-3539.

Schmalz, D., F. Hucho, and K. Buchner. 1998. Nuclear import of protein kinase C occurs by a mechanism distinct from the mechanism used by proteins with a classical nuclear localization signal. *J Cell Sci* **111 (Pt 13):** 1823-1830.

Schmucker, D., J.C. Clemens, H. Shu, C.A. Worby, J. Xiao, M. Muda, J.E. Dixon, and S.L. Zipursky. 2000. *Drosophila* Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. *Cell* **101:** 671-684.

Schwartz, S., Z. Zhang, K.A. Frazer, A. Smit, C. Riemer, J. Bouck, R. Gibbs, R. Hardison, and W. Miller. 2000. PipMaker--a web server for aligning two genomic DNA sequences. *Genome Res* **10:** 577-586.

Scott, J.D. and T.R. Soderling. 1992. Serine/threonine protein kinases. *Curr Opin Neurobiol* **2:** 289-295.

Seipel, B. 1996. Vergleichende cytogenetische Analyse der humanen Chromosomregion 11p15 und der entsprechenden Syntäniegruppe auf dem murinen Chromosom 7 mit Hilfe der Fluoreszenz-in situ-Hybridizierung (FISH). In *Fachbereich Biologie und Kinderklinik*. Johannes-Gutenberg-Universität Mainz.

Shapiro, G.I. and J.W. Harper. 1999. Anticancer drug targets: cell cycle and checkpoint control. *J Clin Invest* **104:** 1645-1653.

Smith, C.M. 1999. The protein kinase resource and other bioinformation resources. *Prog Biophys Mol Biol* **71:** 525-533.

Soderling, T.R. 1999. The Ca-calmodulin-dependent protein kinase cascade. *Trends Biochem Sci* **24:** 232-236.

Stenoien, D.L., S. Sen, M.A. Mancini, and B.R. Brinkley. 2003. Dynamic association of a tumor amplified kinase, Aurora-A, with the centrosome and mitotic spindle. *Cell Motil Cytoskeleton* **55:** 134-146.

Su, A.I., M.P. Cooke, K.A. Ching, Y. Hakak, J.R. Walker, T. Wiltshire, A.P. Orth, R.G. Vega, L.M. Sapinoso, A. Moqrich, A. Patapoutian, G.M. Hampton, P.G. Schultz, and J.B. Hogenesch. 2002. Large-scale analysis of the human and mouse transcriptomes. *Proc Natl Acad Sci U S A* **99:** 4465-4470.

Swofford, D.L. 1991. PAUP: Phylogenetic Analysis Using Parsimony. Computer program distributed by Illinois Natural History Survey, Champaign, Illinois.

Takai, N., T. Miyazaki, M. Nishida, K. Nasu, and I. Miyakawa. 2002. *Ca(2+)/calmodulin-dependent protein kinase IV* expression in epithelial ovarian cancer. *Cancer Lett* **183:** 185-193.

Taudien, S., A. Rump, M. Platzer, B. Drescher, R. Schattevoy, G. Gloeckner, M. Dette, C. Baumgart, J. Weber, U. Menzel, and A. Rosenthal. 2000. RUMMAGE--a high-throughput sequence annotation system. *Trends Genet* **16:** 519-520.

Taylor, S.S., E. Radzio-Andzelm, and T. Hunter. 1995. How do protein kinases discriminate between serine/threonine and tyrosine? Structural insights from the insulin receptor protein-tyrosine kinase. *Faseb J* **9:** 1255-1266.

Teasdale, R.D. and M.R. Jackson. 1996. Signal-mediated sorting of membrane proteins between the endoplasmic reticulum and the golgi apparatus. *Annu Rev Cell Dev Biol* **12:** 27-54.

Tombes, R.M., R.B. Mikkelsen, W.D. Jarvis, and S. Grant. 1999. Downregulation of delta *CaM kinase II* in human tumor cells. *Biochim Biophys Acta* **1452:** 1-11.

Tramontano, A. 2003. Of men and machines. *Nat Struct Biol* **10:** 87-90.

Ullmann, A., F. Jacob, and J. Monod. 1967. Characterization by in vitro complementation of a peptide corresponding to an operator-proximal segment of the beta-galactosidase structural gene of Escherichia coli. *J Mol Biol* **24:** 339-343.

Venter, J.C. M.D. Adams E.W. Myers P.W. Li R.J. Mural G.G. Sutton H.O. Smith M. Yandell C.A. Evans R.A. Holt J.D. Gocayne P. Amanatides R.M. Ballew D.H. Huson J.R. Wortman Q. Zhang C.D. Kodira X.H. Zheng L. Chen M. Skupski G. Subramanian P.D. Thomas J. Zhang G.L. Gabor Miklos C. Nelson S. Broder A.G. Clark J. Nadeau V.A. McKusick N. Zinder A.J. Levine R.J. Roberts M. Simon C. Slayman M. Hunkapiller R. Bolanos A. Delcher I. Dew D. Fasulo M. Flanigan L. Florea A. Halpern S. Hannenhalli S. Kravitz S. Levy C. Mobarry K. Reinert K. Remington J. Abu-Threideh E. Beasley K. Biddick V. Bonazzi R. Brandon M. Cargill I. Chandramouliswaran R. Charlab K. Chaturvedi Z. Deng V. Di Francesco P. Dunn K. Eilbeck C. Evangelista A.E. Gabrielian W. Gan W. Ge F. Gong Z. Gu P. Guan T.J. Heiman M.E. Higgins R.R. Ji Z. Ke K.A. Ketchum Z. Lai Y. Lei Z. Li J. Li Y. Liang X. Lin F. Lu G.V. Merkulov N. Milshina H.M. Moore A.K. Naik V.A. Narayan B. Neelam D. Nusskern D.B. Rusch S. Salzberg W. Shao B. Shue J. Sun Z. Wang A. Wang X. Wang J. Wang M. Wei R. Wides C.

Xiao C. Yan A. Yao J. Ye M. Zhan W. Zhang H. Zhang Q. Zhao L. Zheng F. Zhong W. Zhong S. Zhu S. Zhao D. Gilbert S. Baumhueter G. Spier C. Carter A. Cravchik T. Woodage F. Ali H. An A. Awe D. Baldwin H. Baden M. Barnstead I. Barrow K. Beeson D. Busam A. Carver A. Center M.L. Cheng L. Curry S. Danaher L. Davenport R. Desilets S. Dietz K. Dodson L. Doup S. Ferriera N. Garg A. Gluecksmann B. Hart J. Haynes C. Haynes C. Heiner S. Hladun D. Hostin J. Houck T. Howland C. Ibegwam J. Johnson F. Kalush L. Kline S. Koduru A. Love F. Mann D. May S. McCawley T. McIntosh I. McMullen M. Moy L. Moy B. Murphy K. Nelson C. Pfannkoch E. Pratts V. Puri H. Qureshi M. Reardon R. Rodriguez Y.H. Rogers D. Romblad B. Ruhfel R. Scott C. Sitter M. Smallwood E. Stewart R. Strong E. Suh R. Thomas N.N. Tint S. Tse C. Vech G. Wang J. Wetter S. Williams M. Williams S. Windsor E. Winn-Deen K. Wolfe J. Zaveri K. Zaveri J.F. Abril R. Guigo M.J. Campbell K.V. Sjolander B. Karlak A. Kejariwal H. Mi B. Lazareva T. Hatton A. Narechania K. Diemer A. Muruganujan N. Guo S. Sato V. Bafna S. Istrail R. Lippert R. Schwartz B. Walenz S. Yooseph D. Allen A. Basu J. Baxendale L. Blick M. Caminha J. Carnes-Stine P. Caulk Y.H. Chiang M. Coyne C. Dahlke A. Mays M. Dombroski M. Donnelly D. Ely S. Esparham C. Fosler H. Gire S. Glanowski K. Glasser A. Glodek M. Gorokhov K. Graham B. Gropman M. Harris J. Heil S. Henderson J. Hoover D. Jennings C. Jordan J. Jordan J. Kasha L. Kagan C. Kraft A. Levitsky M. Lewis X. Liu J. Lopez D. Ma W. Majoros J. McDaniel S. Murphy M. Newman T. Nguyen N. Nguyen M. Nodell S. Pan J. Peck M. Peterson W. Rowe R. Sanders J. Scott M. Simpson T. Smith A. Sprague T. Stockwell R. Turner E. Venter M. Wang M. Wen D. Wu M. Wu A. Xia A. Zandieh and X. Zhu. 2001. The sequence of the human genome. *Science* **291:** 1304-1351.

Vinogradov, A.E. 2003. Isochores and tissue-specificity. *Nucleic Acids Res* **31:** 5212-5220.

Wang, N. 2002. Cytogenetics and molecular genetics of ovarian cancer. *Am J Med Genet* **115:** 157-163.

Wansink, D.G., R.E. van Herpen, M.M. Coerwinkel-Driessen, P.J. Groenen, B.A. Hemmings, and B. Wieringa. 2003. Alternative splicing controls myotonic dystrophy protein kinase structure, enzymatic activity, and subcellular localization. *Mol Cell Biol* **23:** 5489-5501.

Watanabe, N., M. Broome, and T. Hunter. 1995. Regulation of the human WEE1Hu CDK tyrosine 15-kinase during the cell cycle. *Embo J* **14:** 1878-1891.

Waterston, R.H., E.S. Lander, and J.E. Sulston. 2003. More on the sequencing of the human genome. *Proc Natl Acad Sci U S A* **100:** 3022-3024; author reply 3025-3026.

Waterston, R.H. K. Lindblad-Toh E. Birney J. Rogers J.F. Abril P. Agarwal R. Agarwala R. Ainscough M. Alexandersson P. An S.E. Antonarakis J. Attwood R. Baertsch J. Bailey K. Barlow S. Beck E. Berry B. Birren T. Bloom P. Bork M. Botcherby N. Bray M.R. Brent D.G. Brown S.D. Brown C. Bult J. Burton J. Butler R.D. Campbell P. Carninci S. Cawley F. Chiaromonte A.T. Chinwalla D.M. Church M. Clamp C. Clee F.S. Collins L.L. Cook R.R. Copley A. Coulson O. Couronne J. Cuff V. Curwen T. Cutts M. Daly R. David J. Davies K.D. Delehaunty J. Deri E.T. Dermitzakis C. Dewey N.J. Dickens M. Diekhans S. Dodge I. Dubchak D.M. Dunn S.R. Eddy L. Elnitski R.D. Emes P. Eswara E. Eyras A. Felsenfeld G.A. Fewell P. Flicek K. Foley W.N. Frankel L.A. Fulton R.S. Fulton T.S. Furey D. Gage R.A. Gibbs G. Glusman S. Gnerre N. Goldman L. Goodstadt D. Grafham T.A. Graves E.D. Green S. Gregory R. Guigo M. Guyer R.C. Hardison D. Haussler Y. Hayashizaki L.W. Hillier A. Hinrichs W. Hlavina T. Holzer F. Hsu A. Hua T. Hubbard A. Hunt I. Jackson D.B. Jaffe L.S. Johnson M. Jones T.A. Jones A. Joy M. Kamal E.K. Karlsson D. Karolchik A. Kasprzyk J. Kawai E. Keibler C. Kells W.J. Kent A. Kirby D.L. Kolbe I. Korf R.S. Kucherlapati E.J. Kulbokas D. Kulp T. Landers J.P. Leger S. Leonard I. Letunic R. Levine J. Li M. Li C. Lloyd S. Lucas B. Ma D.R. Maglott E.R. Mardis L. Matthews E. Mauceli J.H. Mayer M. McCarthy W.R. McCombie S. McLaren K. McLay J.D. McPherson J. Meldrim B. Meredith J.P. Mesirov W. Miller T.L. Miner E. Mongin K.T. Montgomery M. Morgan R. Mott J.C. Mullikin D.M. Muzny W.E. Nash J.O. Nelson M.N. Nhan R. Nicol Z. Ning

C. Nusbaum M.J. O'Connor Y. Okazaki K. Oliver E. Overton-Larty L. Pachter G. Parra K.H. Pepin J. Peterson P. Pevzner R. Plumb C.S. Pohl A. Poliakov T.C. Ponce C.P. Ponting S. Potter M. Quail A. Reymond B.A. Roe K.M. Roskin E.M. Rubin A.G. Rust R. Santos V. Sapojnikov B. Schultz J. Schultz M.S. Schwartz S. Schwartz C. Scott S. Seaman S. Searle T. Sharpe A. Sheridan R. Shownkeen S. Sims J.B. Singer G. Slater A. Smit D.R. Smith B. Spencer A. Stabenau N. Stange-Thomann C. Sugnet M. Suyama G. Tesler J. Thompson D. Torrents E. Trevaskis J. Tromp C. Ucla A. Ureta-Vidal J.P. Vinson A.C. Von Niederhausern C.M. Wade M. Wall R.J. Weber R.B. Weiss M.C. Wendl A.P. West K. Wetterstrand R. Wheeler S. Whelan J. Wierzbowski D. Willey S. Williams R.K. Wilson E. Winter K.C. Worley D. Wyman S. Yang S.P. Yang E.M. Zdobnov M.C. Zody and E.S. Lander. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520-562.

Wheeler, D.L., D.M. Church, S. Federhen, A.E. Lash, T.L. Madden, J.U. Pontius, G.D. Schuler, L.M. Schriml, E. Sequeira, T.A. Tatusova, and L. Wagner. 2003. Database resources of the National Center for Biotechnology. *Nucleic Acids Res* **31:** 28-33.

Wilmann, M., M. Gautel, and O. Mayans. 2000. Activation of calcium/calmodulin regulated kinases. *Cell Mol Biol (Noisy-le-grand)* **46:** 883-894.

Wu, J.Y., I.J. Gonzalez-Robayna, J.S. Richards, and A.R. Means. 2000a. Female Fertility Is Reduced in Mice Lacking Ca2+/Calmodulin-Dependent Protein Kinase IV. *Endocrinology* **141:** 4777-4783.

Wu, J.Y., T.J. Ribar, D.E. Cummings, K.A. Burton, G.S. McKnight, and A.R. Means. 2000b. Spermiogenesis and exchange of basic nuclear proteins are impaired in male germ cells lacking Camk4. *Nat Genet* **25:** 448-452.

Yang, R.B., C.K. Ng, S.M. Wasserman, S.D. Colman, S. Shenoy, F. Mehraban, L.G. Komuves, J.E. Tomlinson, and J.N. Topper. 2002. Identification of a novel family of cell-surface proteins expressed in human vascular endothelium. *J Biol Chem* **277:** 46364-46373.

Yu, J., S. Hu, J. Wang, G.K. Wong, S. Li, B. Liu, Y. Deng, L. Dai, Y. Zhou, X. Zhang, M. Cao, J. Liu, J. Sun, J. Tang, Y. Chen, X. Huang, W. Lin, C. Ye, W. Tong, L. Cong, J. Geng, Y. Han, L. Li, W. Li, G. Hu, J. Li, Z. Liu, Q. Qi, T. Li, X. Wang, H. Lu, T. Wu, M. Zhu, P. Ni, H. Han, W. Dong, X. Ren, X. Feng, P. Cui, X. Li, H. Wang, X. Xu, W. Zhai, Z. Xu, J. Zhang, S. He, J. Xu, K. Zhang, X. Zheng, J. Dong, W. Zeng, L. Tao, J. Ye, J. Tan, X. Chen, J. He, D. Liu, W. Tian, C. Tian, H. Xia, Q. Bao, G. Li, H. Gao, T. Cao, W. Zhao, P. Li, W. Chen, Y. Zhang, J. Hu, S. Liu, J. Yang, G. Zhang, Y. Xiong, Z. Li, L. Mao, C. Zhou, Z. Zhu, R. Chen, B. Hao, W. Zheng, S. Chen, W. Guo, M. Tao, L. Zhu, L. Yuan, and H. Yang. 2002. A draft sequence of the rice genome (*Oryza sativa L. ssp. indica*). *Science* **296:** 79-92.

Zdobnov, E.M., C. von Mering, I. Letunic, D. Torrents, M. Suyama, R.R. Copley, G.K. Christophides, D. Thomasova, R.A. Holt, G.M. Subramanian, H.M. Mueller, G. Dimopoulos, J.H. Law, M.A. Wells, E. Birney, R. Charlab, A.L. Halpern, E. Kokoza, C.L. Kraft, Z. Lai, S. Lewis, C. Louis, C. Barillas-Mury, D. Nusskern, G.M. Rubin, S.L. Salzberg, G.G. Sutton, P. Topalis, R. Wides, P. Wincker, M. Yandell, F.H. Collins, J. Ribeiro, W.M. Gelbart, F.C. Kafatos, and P. Bork. 2002. Comparative genome and proteome analysis of *Anopheles gambiae* and *Drosophila melanogaster*. *Science* **298:** 149-159.

Zoubak, S., O. Clay, and G. Bernardi. 1996. The gene distribution of the human genome. *Gene* **174:** 95-102.

7 Appendix

7.1 Published genome projects

Table 7.1: Sequenced genomes between 1995 and 2002

Organism Name	Kingdom	Publication Date	Size (Kb)	Predicted ORFs
Mycoplasma genitalium G-37	Bacteria	1995	580	468
Haemophilus influenzae KW20	Bacteria	1995	1830	1850
			2410	2318
Mycoplasma pneumoniae M129	Bacteria	1996	816	677
Methanococcus jannaschii DSM 2661	Archaea	1996	1664	1750
Synechocystis sp. PCC6803	Bacteria	1996	3573	3168
			6053	5595
Borrelia burgdorferi B31	Bacteria	1997	1230	1256
Archaeoglobus fulgidus DSM4304	Archaea	1997	2178	2493
Bacillus subtilis 168	Bacteria	1997	4214	4099
Methanobacterium thermoautotrophicum delta H	Archaea	1997	1751	1918
Escherichia coli K12- MG1655	Bacteria	1997	4639	4289
Helicobacter pylori 26695	Bacteria	1997	1667	1590
Saccharomyces cerevisiae S288C	Eukarya	1997	12069	6294
			27748	21939
Caenorhabditis elegans	Eukarya	1998	97000	19099
Rickettsia prowazekii Madrid E	Bacteria	1998	1111	834
Chlamydia trachomatis D/UW-3/CX (serovar D)	Bacteria	1998	1042	896
Treponema pallidum subsp. pallidum Nichols	Bacteria	1998	1138	1041
Mycobacterium tuberculosis H37Rv (lab strain)	Bacteria	1998	4411	4402
Pyrococcus horikoshii(shinkaj) OT3	Archaea	1998	1738	1979
Aquifex aeolicus VF5	Bacteria	1998	1551	1544
			107991	29795
Deinococcus radiodurans R1	Bacteria	1999	3284	3187
Thermotoga maritima MSB8	Bacteria	1999	1860	1877
Aeropyrum pernix K1	Archaea	1999	1669	2620
Chlamydophila pneumoniae CWL029	Bacteria	1999	1230	1052
Leishmania major Friedlin Chromosome 1	Eukarya	1999	257	79
Helicobacter pylori J99	Bacteria	1999	1643	1495
			9943	10310

Thermoplasma volcanium GSS1	Archaea	2000	1584	1524
Arabidopsis thaliana	Eukarya	2000	115428	25498
Mesorhizobium loti MAFF303099	Bacteria	2000	7596	6752
Halobacterium sp. NRC-1	Archaea	2000	2014	2058
Ureaplasma urealyticum (parvum) serovar 3	Bacteria	2000	751	650
Pseudomonas aeruginosa PAO1	Bacteria	2000	6264	5570
Thermoplasma acidophilum DSM 1728	Archaea	2000	1564	1478
Buchnera aphidicola AP (Acyrthosiphon pisum)	Bacteria	2000	640	564
Vibrio cholerae serotype O1, Biotype El Tor, strain N16961	Bacteria	2000	4000	3885
Xylella fastidiosa CVC 8.1.b clone 9.a.5.c	Bacteria	2000	2679	2904
Chlamydophila pneumoniae J138	Bacteria	2000	1228	1070
Bacillus halodurans C-125	Bacteria	2000	4202	4066
Neisseria meningitidis Z2491 (serogroup A)	Bacteria	2000	2184	2121
Drosophila melanogaster	Eukarya	2000	137000	14100
Chlamydia trachomatis MoPn / Nigg	Bacteria	2000	1069	924
Chlamydia pneumoniae AR39	Bacteria	2000	1229	1052
Neisseria meningitidis MC58 (serogroup B)	Bacteria	2000	2272	2158
Campylobacter jejuni subsp. jejuni NCTC 11168	Bacteria	2000	1641	1654
			293345	78028

Agrobacterium tumefaciens C58-DuPont	Bacteria	2001	4915	5402
Agrobacterium tumefaciens C58-Cereon	Bacteria	2001	4915	5299
Encephalitozoon cuniculi	Eukarya	2001	2500	1997
Nostoc (Anabaena) sp. PCC 7120	Bacteria	2001	6413	5366
Listeria monocytogenes EGD-e	Bacteria	2001	2944	2855
Listeria innocua Clip11262, rhamnose-negative	Bacteria	2001	3011	2981
Salmonella typhimurium,LT2 SGSC1412	Bacteria	2001	4857	4597
Salmonella typhi CT18	Bacteria	2001	4809	4600
Streptococcus pneumoniae R6	Bacteria	2001	2038	2043
Yersinia pestis CO-92 (Biovar Orientalis)	Bacteria	2001	4653	4012
Mycobacterium tuberculosis CDC1551	Bacteria	2001	4403	4187
Rickettsia conorii Malish 7	Bacteria	2001	1268	1374
Sulfolobus tokodaii 7	Archaea	2001	2694	2826
Clostridium acetobutylicum ATCC 824D	Bacteria	2001	4100	4927
Sinorhizobium meliloti 1021	Bacteria	2001	6690	6205
Streptococcus pneumoniae TIGR4 ATCC-BAA-334	Bacteria	2001	2160	2094
Sulfolobus solfataricus P2	Archaea	2001	2992	2977
Caulobacter crescentus CB15	Bacteria	2001	4016	3737
Mycoplasma pulmonis UAB CTIP	Bacteria	2001	963	782
Lactococcus lactis subsp. lactis IL1403	Bacteria	2001	2365	2266
Guillardia theta	Eukarya	2001	551	464
Staphylococcus aureus Mu50 (VRSA)	Bacteria	2001	2878	2697
Staphylococcus aureus N315 (MRSA)	Bacteria	2001	2813	2594
Streptococcus pyogenes SF370 (M1)	Bacteria	2001	1852	1696
Pasteurella multocida Pm70	Bacteria	2001	2250	2014
Escherichia coli O157:H7. Sakai	Bacteria	2001	5594	5448
Mycobacterium leprae TN	Bacteria	2001	3268	1604
Homo sapiens	Eukarya	2001	2900000	35000
Escherichia coli O157:H7 EDL933	Bacteria	2001	4100	5283
			2996012	127327

Mus musculus C57BL/6J	Eukarya	2002	2500000	30000
Escherichia coli UPEC-CFT073	Bacteria	2002	5231	
Pseudomonas putida KT2440	Bacteria	2002	6100	
Mycoplasma penetrans HF-2	Bacteria	2002	1358	1039
Corynebacterium efficiens YS-314T	Bacteria	2002	3140	3031
Bifidobacterium longum NCC2705	Bacteria	2002	2256	1813
Streptococcus mutans UA159	Bacteria	2002	2030	1963
Glossina (Wigglesworthia) brevipalpis P-endosymbiont	Bacteria	2002	697	621
Leptospira interrogans serovar lai 56601	Bacteria	2002	4691	4728
Shigella flexneri, 2a 301	Bacteria	2002	4607	4434
Shewanella oneidensis (putrefaciens) MR-1 ATCC700550	Bacteria	2002	4969	4931
Anopheles gambiae	Eukarya	2002	278000	14000
Plasmodium yoelii yoelii 17XNL	Eukarya	2002	23100	5878
Plasmodium falciparum 3D7	Eukarya	2002	22900	5268
Brucella melitensis biovar suis 1330	Bacteria	2002	3310	3388
Streptococcus agalactiae NEM316	Bacteria	2002	2211	2118
Oceanobacillus iheyensis HTE831	Bacteria	2002	3630	3496
Streptococcus agalactiae 2603V/R	Bacteria	2002	2160	
Thermosynechococcus elongatus BP-1	Bacteria	2002	2600	2477
Yersinia pestis KIM5 P12 (Biovar Mediaevalis)	Bacteria	2002	4600	4198
Streptococcus pyogenes MGAS315	Bacteria	2002	1900	1865
Methanosarcina mazei Go1 (DSMZ 3647)	Archaea	2002	4096	3371
Chlorobium tepidum TLS	Bacteria	2002	2154	2288
Dictyostelium discoideum Chromosome 2	Eukarya	2002	8000	2799
Buchnera aphidicola SG (Schizaphis graminum)	Bacteria	2002	641	545
Staphylococcus aureus subsp. aureus MW2	Bacteria	2002	2820	2632
Xanthomonas axonopodis pv. citri 306	Bacteria	2002	5273	4386
Xanthomonas campestris pv.campestris ATCC 33913	Bacteria	2002	5076	4182
Streptomyces coelicolor A3(2)	Bacteria	2002	8667	7825
Thermoanaerobacter tengcongensis MB4T	Bacteria	2002	2689	2588
Fusobacterium nucleatum ATCC 25586	Bacteria	2002	2170	2067
Methanosarcina acetivorans C2A	Archaea	2002	5751	4540
Oryza sativa L. ssp. indica	Eukarya	2002	466000	50000
Oryza sativa ssp. japonica	Eukarya	2002	466000	50000
Streptococcus pyogenes MGAS8232	Bacteria	2002	1895	1889
Methanopyrus kandleri AV19	Archaea	2002	1694	1691
Corynebacterium glutamicum ATCC-13032	Bacteria	2002	3309	3040
Schizosaccharomyces pombe	Eukarya	2002	14000	4824
Pyrococcus abyssi GE5	Archaea	2002	1765	1765
Pyrococcus furiosus DSM 3638	Archaea	2002	1908	2065
Ralstonia solanacearum GMI1000	Bacteria	2002	5810	5120
Clostridium perfringens 13	Bacteria	2002	3031	2660
Pyrobaculum aerophilum IM2	Archaea	2002	2222	2587
Brucella melitensis 16M	Bacteria	2002	3294	3197
Ciona intestinalis	Eukarya	2002	120000	16000
Fugu rubripes	Eukarya	2002	365000	40000

7.2 Primers

Table 7.2: Primes used for mouse clone BAC221D7 sequencing project

Primer	Sequence	Use
B4p310a	TGAAGACATGACCTGGAG	Gaps closure
B4p310b	CAGTTGCTACGTACTGAG	Gaps closure
B4p357a	GTGGTTCATTTCTTATGCAC	Gaps closure
B4p357b	TCTGAAGTCTCAGTTGAG	Gaps closure
B4p302a	CATCACTGGAAAGAGAGG	Gaps closure
B4p302b	AGGCAACACTGGAGATAG	Gaps closure
B4p376a	CTTGTATGTTCATAGGTATC	Gaps closure
B4p376b	AGACAGCCACAAGAACAG	Gaps closure
B4p374a	TTTCAGACAGTCACAGCC	Gaps closure
B4p374b	TGACTGATTTCATTACCC	Gaps closure
B4p305a	CTATTTGAGTTCCTGTCC	Gaps closure
B4p305b	TCCACCTGATTCTGAGGT	Gaps closure
B4p317a	GACAATGCTACTGCTAGC	Gaps closure
B4p317b	CCTGTATGTGGTATTGCC	Gaps closure
B4p315a	GATCTGGTTGAAAATCCC	Gaps closure
B4p315b	TTGGATGCTGATTATGGG	Gaps closure
B4p294a	TAGAGGAACATGACCTCC	Gaps closure
B4p294b	CTGGATGGTTTTAGCTCC	Gaps closure
B4p363a	TTAGCCAGAGCAATTCGC	Gaps closure
B4p363b	CAGAGTAATAGTGGCTTC	Gaps closure
B4p373a	AGCTATTCCAGCTACTAC	Gaps closure
317xl	CAAGTTGGTACACACTAAGTCCTCTTCCTG	XL-PCR
373xl	GTCTAACAGGTTTCTAATCTGTGTACTTGG	XL-PCR
377xl	TTAGGTTCTGGTTGTGGAGATGCTTCCTTC	XL-PCR
I10a	AAGCTGAAAGACTGATAT	Is10 Test
I10b	AAAGAATGTGTACTCTGC	Is10 Test
B4p27	CAACACACCACTTTCACC	Finishing
B4p28	CTTCAACACACCACTTTC	Finishing
B4p29	TGCCTTCTACTCCTCTTG	Finishing
B4p30	TAGTGAGTCTAGCTAAGG	Finishing
B4p31	GAGATTACAGAGGATTGC	Finishing

Table 7.3: Primes used for human PAC library screening and *STK33* specific primers

Primer	Sequence	Use
H1	TGCAGGTATAGCTTTCAG	Probe for human PAC library screening
H2	CCCTCTTTAAGATCTCCA	Probe for human PAC library screening
H3	CATCTTGTAGCACAGACT	Probe for human PAC library screening
H4	TTACGTAAACGGGACAAC	Probe for human PAC library screening
H5	TATCTATGGCTCTTTGGG	Probe for human PAC library screening
H6	AAGGTCTCCTAAGACTAC	Probe for human PAC library screening
H7	AATGCATTTGCCCTGTTG	Probe for human PAC library screening
H8	GGAATGATGAGCCTAGAC	Probe for human PAC library screening
H9	ACAACTCACTGCTGGCAC	Probe for human PAC library screening
(*STK33* specific primers)		
H10	TTCCATAAGTGACTGTGC	Probe for expression experiments
H11	AAGCAATTTCCTGCAACC	Probe for expression experiments
H12	CATGTGAATGACTGAAGC	Probe for expression experiments
H13	GATTTAGCAGGGTACTTG	Probe for expression experiments
H14	ACAGACATCAAGCATTGG	RT-PCR from Uterus
H15	TTGTCCTTACTGGTTGCA	RT-PCR from Uterus
H16	TACATTGGTTGGTCTCAC	RT-PCR from Uterus + Exon 13 Splicing test
H17	TAGTGCATCCTTAACCTG	RT-PCR from Uterus + Exon 13 Splicing test
H18	CAAAGTGCTAACAAGAGC	RT-PCR from Uterus + Exon 13 Splicing test
H19	CTATGCTCCAAATGTCAC	RT-PCR from Uterus + Exon 13 Splicing test
H20	GCACTAGAAGCATCCTGTATAGAC	RACE
H21	TGTATTTCTTAATGTGGGTTGTGG	RACE
H22	ACAGCATCTTATCAAACTGCGTAG	RACE

Table 7.4: Mouse *Stk33* specific primers

Primer	Sequence	Use
M1	AACCCAGAAAGTGATGAG	MmuSTK33 probe forward
M2	TAGAACTAAGCGAGCATG	MmuSTK33 probe reverse
M3	ATGTGTGAAGTGTGGTAC	MmuSTK33 Exon 06 UTR forward
M4	TCTTTACCAGCAACTCAC	MmuSTK33 Exon 06 UTR forward
M5	ATGCAGCTACTGGATACT	MmuSTK33 Exon 05 UTR forward
M6	GGCCTTAGAAAAGTTCAG	MmuSTK33 Exon 05 UTR forward
M7	TGGTGAAGTCCTTGGGTTAATGAGG	GSP1 RACE Exon 06
M8	AATTCTGTACTACCAATACTCG	GSP1 RACE Exon 06
M9	AGAGCTTCTATGAGTCACCCAGGAG	GSP2 RACE Exon 05
M10	GAAGCTGAACTTTTCTAAGGCCCGG	GSP3 RACE Exon 04
M11	CAGAGATGCATAATACGACTCACTATAGGGAGAAACCCAGAAAGTGATGAG	Oligo with T7 Promoter derivated from AMcp1
M12	CAGAGATGCATAATACGACTCACTATAGGGAGATAGAACTAAGCGAGCATG	Oligo with T7 Promoter derivated from AMcp2

Table 7.5: Standard primers

Primer	Sequence	Use
M13-Forward	GTAAAACGACGGCCAGT	Shotgun and IMAGE clones seq.
M13-Reverse	CAGGAAACAGCTATGAC	Shotgun and IMAGE clones seq.
T7	TAATACGACTCACTATAGGG	IMAGE clones sequencing
T3	AATTAACCCTCACTAAAGGG	IMAGE clones sequencing
sp6	ACCTTATGTATCATACACAT	IMAGE clones sequencing
AUAP	GGCCACGCGTCGACTAGTAC	RACE AUAP
3'AP	GGCCACGCGTCGACTAGTACTTTTTTTTTTTTTTTT	RACE AP

7.3 Negative versions of the RNA in-situ hybridisation

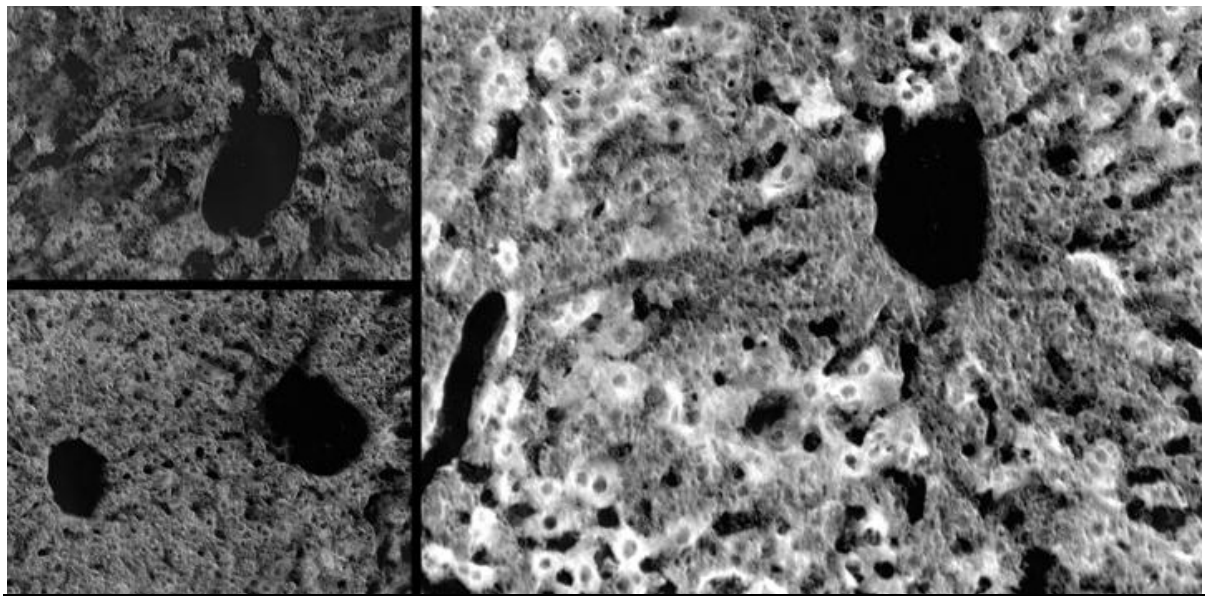


Figure 7.1: Negative visualisation of figure 3.31
Control, sense and anti-sense *in-situ* RNA hybridisation s of *Stk33*-specific sonde with frozen slides of mouse liver. Note the sinusoid capillars as dark cavities in this visualisation. *Stk33*-specific signal appears as very intensive white regions.

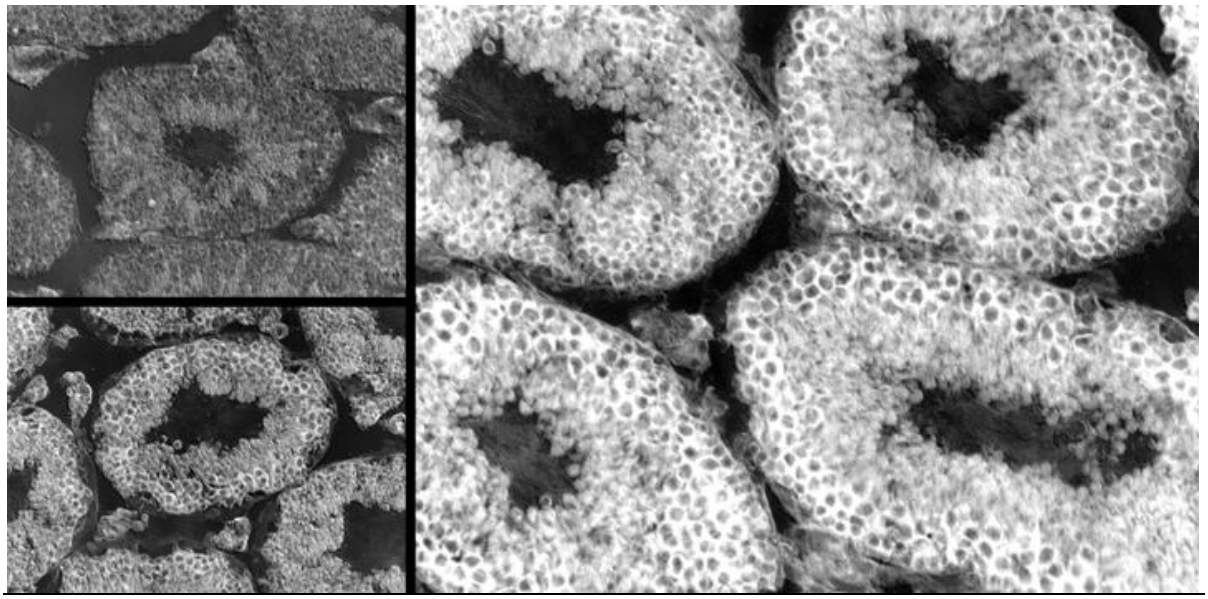


Figure 7.2: Negative visualisation of figure 3.32.
Control, sense and anti-sense *in-situ* RNA hybridisation s of *Stk33*-specific sonde with frozen slides of mouse testis. *Stk33*-specific signal appears as very intensive white regions.

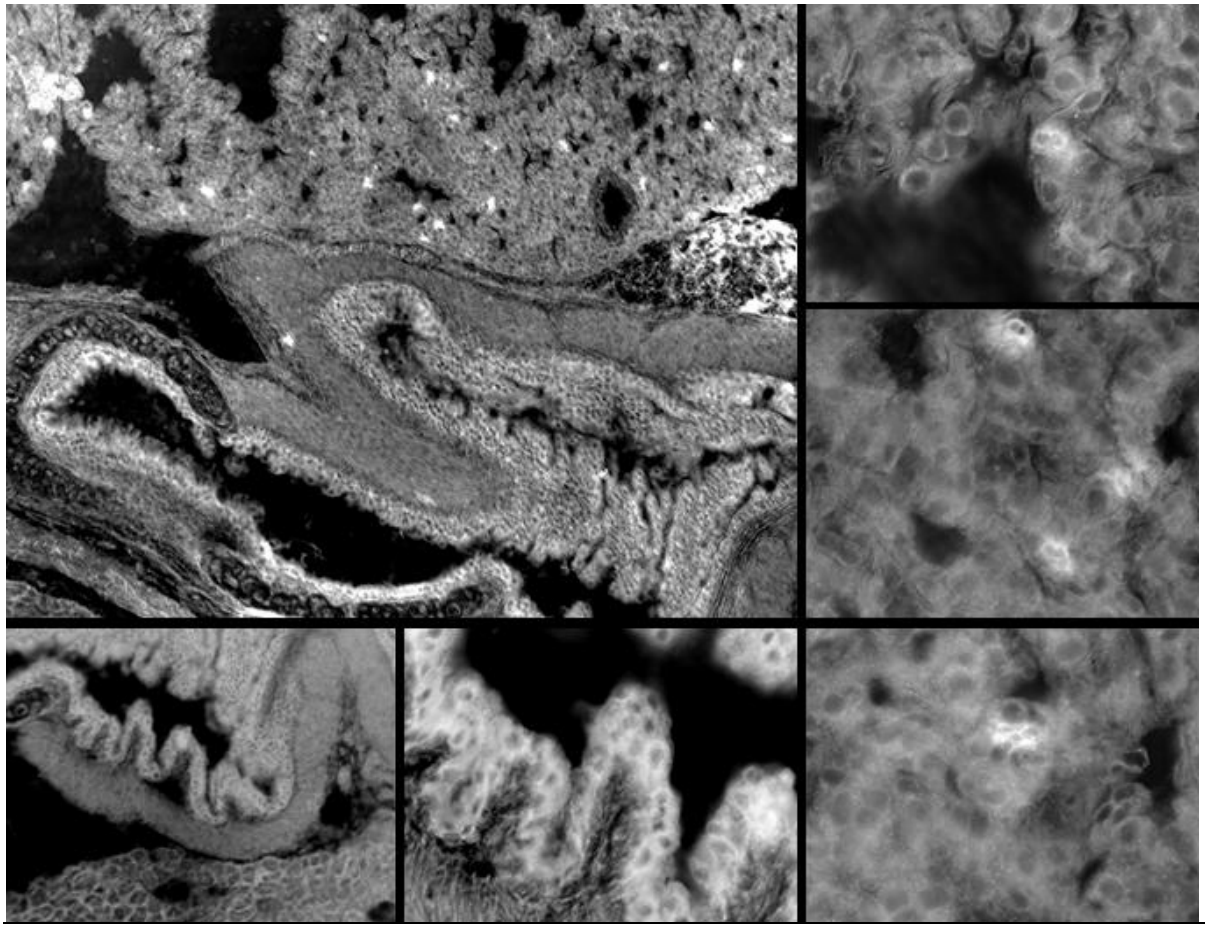


Figure 7.3: Negative visualisation of figure 3.35
Control, sense and anti-sense *in-situ* RNA hybridisation s of *Stk33*-specific sonde with frozen slides of mouse lungs. *Stk33*-specific signal appears as very intensive white regions.

7.4 Human/Mouse Alignment of *STK33* coding sequence

```
                 *        20         *        40         *        60
H.STK33 ATGGCTGATAGTGGCTTAGATAAAAAATCCACAAAATGCCCCGACTGTTCATCTGCTTCT :   60
M.Stk33 ATGGCTGACCCCAGCTTGAATGACAACCCTACAGCATGCCCTCACTGTGCATCC---TCT :   57
        ATGGCTGA     GCTT  AT A AA  C ACA  ATGCCC  ACTGT CATC    TCT

                 *        80         *       100         *       120
H.STK33 CAGAAAGATGTACTTTGTGTATGTTCCAGCAAAACAAGGGTTCCTCCAGTTTTGGTGGTG :  120
M.Stk33 CAGGCTGGCCTACTGTGTGTATGTCCAGCA---GGC---AAGTCTCCAGTCCTGGTGGTG :  111
        CAG   G   TACT TGTGTATGT C                CTCCAGT  TGGTGGTG

                 *       140         *       160         *       180
H.STK33 GAAATGTCACAGACATCAAGCATTGGTAGTGCAGAATCTTTAATTTCACTGGAGAGAAAA :  180
M.Stk33 GAAATGTCACAGACATCGAGTATTGGTAGTACAGAATTTTTTGCTTCACAAGAAAGAAAA :  171
        GAAATGTCACAGACATC AG ATTGGTAGT CAGAAT TTT   TTCAC  GA AGAAAA

                 *       200         *       220Ex06¬    *       240
H.STK33 AAAGAAAAAAATATCAACAGAGATATAACCTCCAGGAAAGATTTGCCCTCAAGAACCTCA :  240
M.Stk33 AAGGAAAGAAATACCAGCAGAGAA---TCTTCTCTAAAAGATTTGTCCATAAGAACTTCA :  228
        AA GAAA AAATA CA CAGAGA     C TC    AAAGATTTG CC  AAGAAC TCA

                 *       260         *       280         *       300
H.STK33 AATGTAGAGAGAAAAGCATCTCAGCAACAATGGGGTCGGGGCAACTTTACAGAAGGAAAA :  300
M.Stk33 AATGTGGAGAGAAAA---CCTCAGGCACAATGGAGTCGGAGCAATGTCACAGTAGGAAAA :  285
        AATGT GAGAGAAAA    CTCAG  ACAATGG GTCGG GCAA  T ACAG AGGAAAA

                 *       320         *    Ex07¬          *       360
H.STK33 GTTCCTCACATAAGGATTGAGAATGGAGCTGCTATTGAGGAAATCTATACCTTTGGAAGA :  360
M.Stk33 ATCCCACACATAAGAATGGACGATGGAGCAGGTATCGAGGAATTCTATACCTTTGGAAGA :  345
         T CC CACATAAG AT GA  ATGGAGC G TAT GAGGAA TCTATACCTTTGGAAGA

                 *       380         *       400         *       420
H.STK33 ATATTGGGAAAAGGGAGCTTTGGAATAGTCATTGAAGCTACAGACAAGGAAACAGAAACG :  420
M.Stk33 ATATTGGGACAGGGGAGCTTTGGAATGGTCTTTGAAGCTATAGACAAGGAAACAGGAGCT :  405
        ATATTGGGA A GGGAGCTTTGGAAT GTC TTGAAGCTA AGACAAGGAAACAG A C

                 *       440        Ex08¬    460         *       480
H.STK33 AAGTGGGCAATTAAAAAAGTGAACAAAGAAAAGGCTGGAAGCTCTGCTGTGAAGTTACTT :  480
M.Stk33 AAGTGGGCAATTAAAAAAGTGAATAAAGAAAAGGCTGGAAGTTCTGCAATGAAGCTACTG :  465
        AAGTGGGCAATTAAAAAAGTGAA AAAGAAAAGGCTGGAAG TCTGC  TGAAG TACT

                 *       500         *       520         *       540
H.STK33 GAACGAGAGGTGAACATTCTGAAAAGTGTAAAACATGAACACATCATACATCTGGAACAA :  540
M.Stk33 GAGCGGGAGGTGAGCATCCTGAAGACTGTCAACCATCAACACATCATCCACCTGGAACAA :  525
        GA CG GAGGTGA CAT CTGAA A TGT AA CAT AACACATCAT CA CTGGAACAA

                 *   Ex09¬           *       580         *       600
H.STK33 GTATTTGAAACGCCAAAGAAAATGTACCTTGTGATGGAGCTTTGTGAGGATGGAGAACTC :  600
M.Stk33 GTGTTTGAGTCGCCTCAGAAAATGTATCTCGTGATGGAGCTTTGTGAGGATGGAGAACTC :  585
        GT TTTGA  CGCC  AGAAAATGTA CT GTGATGGAGCTTTGTGAGGATGGAGAACTC
```

```
                *       620         *       640         *       660
H.STK33 AAAGAAATTCTGGATAGGAAAGGGCATTTCTCAGAGAATGAGACAAGGTGGATCATTCAA :  660
M.Stk33 AAAGCAGTTATGGATCAAAGAGGGCACTTCTCAGAGAACGAGACAAGGCTGATAATTCAA :  645
        AAAG A TT TGGAT   A AGGGCA TTCTCAGAGAA GAGACAAGG  GAT ATTCAA

                *       680         *  Ex10¬700         *       720
H.STK33 AGTCTCGCATCAGCTATAGCATATCTTCACAATAATGATATTGTACATAGAGATCTGAAA :  720
M.Stk33 AGTCTTGCATCAGCCATCGCATATCTTCATAACAAGGATATAGTGCACAGAGATCTAAAG :  705
        AGTCT GCATCAGC AT GCATATCTTCA AA AA GATAT GT CA AGAGATCT AA

                *       740         *       760         *       780
H.STK33 CTGGAAAATATAATGGTTAAAAGCAGTCTTATTGATGATAACAATGAAATAAACTTAAAC :  780
M.Stk33 CTGGAAAACATAATGGTTAAAAGCAGCTTTATAGATGATAACAATGAAATGAACTTAAAC :  765
        CTGGAAAA ATAATGGTTAAAAGCAG  TTAT GATGATAACAATGAAAT AACTTAAAC

         Ex11¬  *       800         *       820         *       840
H.STK33 ATAAAGGTGACTGATTTTGGCTTAGCGGTGAAGAAGCAAA---GTAGGAGTGAAGCCATG :  837
M.Stk33 ATAAAGGTGACTGATTTTGGCTTGTCTGTGCAGAAGCATGGCTCCAGGAGTGAAGGCATG :  825
        ATAAAGGTGACTGATTTTGGCTT  C GTG AGAAGCA       AGGAGTGAAG CATG

                *       860       Ex12¬   880         *       900
H.STK33 CTGCAGGCCACATGTGGGACTCCTATCTATATGGCCCCTGAAGTTATCAGTGCCCACGAC :  897
M.Stk33 ATGCAGACTACATGTGGGACTCCTATCTATATGGCACCAGAGGTCATCAATGCCCATGAC :  885
         TGCAG C ACATGTGGGACTCCTATCTATATGGC CC GA GT ATCA TGCCCA GAC

                *       920         *       940     Ex13¬       960
H.STK33 TATAGCCAGCAGTGTGACATTTGGAGCATAGGCGTCGTAATGTACATGTTATTACGTGGA :  957
M.Stk33 TACAGCCAGCAGTGTGACATTTGGAGCATAGGTGTCATAATGTTCATTTTACTGTGTGGA :  945
        TA AGCCAGCAGTGTGACATTTGGAGCATAGG GTC TAATGT CAT TTA T  GTGGA

                *       980         *      1000         *      1020
H.STK33 GAACCACCCTTTTTGGCAAGCTCAGAAGAGAAGCTTTTTGAGTTAATAAGAAAAGGAGAA : 1017
M.Stk33 GAGCCACCCTTTTTGGCAAATTCAGAAGAAAAGCTCTATGAATTAATAAAAAAGGGAGAA : 1005
        GA CCACCCTTTTTGGCAA  TCAGAAGA AAGCT T TGA TTAATAA AAA GGAGAA

                *      1040         *       Ex14¬      *      1080
H.STK33 CTACATTTTGAAAATGCAGTCTGGAATTCCATAAGTGACTGTGCTAAAAGTGTTTTGAAA : 1077
M.Stk33 CTACGATTTGAAAATCCAGTCTGGGAATCTGTAAGTGATTCTGCAAAAAATACTTTGAAA : 1065
        CTAC  TTTGAAAAT CAGTCTGG A TC  TAAGTGA T TGC AAAA T  TTTGAAA

                *      1100         *      1120         *      1140
H.STK33 CAACTTATGAAAGTAGATCCTGCTCACAGAATCACAGCTAAGGAACTACTAGATAACCAG : 1137
M.Stk33 CAACTCATGAAAGTAGATCCTGCTCACAGAATCACAGCTAAGGAACTTCTAGATAACCAA : 1125
        CAACT ATGAAAGTAGATCCTGCTCACAGAATCACAGCTAAGGAACT CTAGATAACCA

            Ex15¬*      1160         *      1180         *      1200
H.STK33 TGGTTAACAGGCAATAAACTTTCTTCGGTGAGACCAACCAATGTATTAGAGATGATGAAG : 1197
M.Stk33 TGGTTGACAGGCAATACCCTTTCTTCAGCAAGACCAACCAATGTATTAGAAATGATGAAA : 1185
        TGGTT ACAGGCAATA  CTTTCTTC G  AGACCAACCAATGTATTAGA ATGATGAA

                *      1220         *      1240         *      1260
H.STK33 GAATGGAAAAATAACCCAGAAAGTGTTGAGGAAAACACAACAGAAGAGAAGAATAAGCCG : 1257
M.Stk33 GAATGGAAAAATAACCCAGAAAGTGATGAGGAGACCAACACAGATGAG------------ : 1233
        GAATGGAAAAATAACCCAGAAAGTG TGAGGA A CA  ACAGA GAG

                *      1280         *      1300         *      1320
H.STK33 TCCACTGAAGAAAAGTTGAAAAGTTACCAACCCTGGGGAAATGTCCCTGATGCCAATTAC : 1317
M.Stk33 GAGACTGAGCAG----------------------AGCGCTGTC------------TAC : 1257
           ACTGA  A                         G   TGTC            TAC
```

```
                 *      1340  Ex16¬  *      1360         *      1380
H.STK33 ACTTCAGATGAAGAGGAGGAAAAACAGTCTACTGCTTATGAAAAGCAATTTCCTGCAACC : 1377
M.Stk33 AGTCCATCTGCAAACACAGCAAAGCAGCCCACCAATGCAGCCAAGAAG---CCTGCTGCA : 1314
        A T CA  TG A A    G AAA CAG C AC   T   G  AAG A    CCTGC  C

                 *      1400         *      1420         *      1440
H.STK33 AGTAAGGACAACTTTGATATGTGCAGTTCAAGTTTCACATCTAGCAAACTCCTTCCAGCT : 1437
M.Stk33 ------GAGAGTGTTGGCATGACCTCTTCAAACTCATCGTCCAGCAAACTCCTGTCTGCT : 1368
              GA A   TTG  ATG  C  TTCAA  T   C TC AGCAAACTCCT  C GCT

                 *      1460         *      1480         *      1500
H.STK33 GAAATCAAGGGAGAAATGGAGAAAACCCCTGTGACTCCAAGCCAAGGAACAGCAACCAAG : 1497
M.Stk33 GAAAGCAAAGCAGAACCAGAGAAAAGCTCCGAGACTGTAGGCCATGCATCAGTGGCTAAA : 1428
        GAAA CAA G AGAA   GAGAAAA C C G GACT  A GCCA G A CAG   C AA

                 *      1520         *      1540
H.STK33 TACCCTGCTAAATCCGGCGCCCTGTCCAGAACCAAAAAGAAACTC TAA : 1545
M.Stk33 ACCACTCTGAAATCCACTACCTTGTTTCGAGGCAAGAAAAGGCTC TAA : 1476
        ACC CTCT AAATCC    CC TGT   GA  CAA AA A  CTC TAA
```

Figure 7.4: Nucleic acid alignment of human and mouse *STK33* coding sequences.
The alignment was manually adapted to the protein alignment shown in the figure 3.35 for dn/ds calculation purposes. Start and stop codons are shown in bold face and exons limits in red.

8. Figure index

9. Table index