Technische Universität München
GSF-Forschungszentrum Neuherberg

Analysis of Molecular Events Involved in Chondrogenesis and Somitogenesis by Global Gene Expression Profiling

Matthias Wahl

Vollständiger Abdruck der von der Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften

genehmigten Dissertation.

Vorsitzender:		Univ.-Prof. Dr. A. Gierl
Prüfer der Dissertation:	1.	Hon.-Prof. Dr. R. Balling, Technische Universität Carlo-Wilhelmina zu Braunschweig
	2.	Univ.-Prof. Dr. W. Wurst

Die Dissertation wurde am 29. Oktober 2003 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 14. Januar 2004 angenommen.

Zusammenfassung

Um die molekularen Mechanismen, welche Knorpelentwicklung und Somitogenese steuern, aufzuklären, wurde eine globale quantitative Genexpressionsanalyse unter Verwendung von SAGE (Serial Analysis of Gene Expression) an der knorpelbildenden Zelllinie ATDC5 und an somitischem Gewebe, präpariert von E10.5 Mäusen, durchgeführt.

Unter insgesamt 43,656 von der murinen knorpelbildenden Zelllinie ATDC5 gewonnenen SAGE Tags (21,875 aus uninduzierten Zellen und 21,781 aus Zellen, die für 6h mit BMP4 induziert wurden) waren 139 Transkripte unterschiedlich in den beiden Bibliotheken repräsentiert ($P \leq 0.05$). 95 Tags konnten einzelnen UniGene Einträgen zugeordnet werden (77 bekannte Gene und 18 ESTs), aber überraschenderweise wurden viele davon bisher nicht mit der Differenzierung von Knorpel in Verbindung gebracht. Interessanterweise wurde von einer signifikanten Fraktion dieser Gene Gruppen physikalischer Verknüpfung gebildet.

Für die Untersuchung der Somitogenese wurden Expressionsprofile von vier verschiedenen Teilen des caudalen bereiches von E10.5 Maus Embryonen verglichen: Schwanzknospe (A), die caudalen zwei Drittel (B) und das rostrale Drittel (C) des präsomitischen Mesoderms und zwei Paare werdender Somiten. Insgesamt wurden 171,639 LongSAGE Tags (A: 21,595; B: 50,699; C: 49,732; D: 49,613) generiert, wodurch 1007 Transcripte identifiziert wurden, welche zumindest zwischen zwei Bibliotheken unterschiedlich repräsentiert waren ($P \leq 0.05$).

Alle LongSAGE Tags, die mindestens zwei Mal in dem gesamten Datensatz vorkamen, wurden mit der momentanen EnsEMBL Genom- Annotierung verglichen. Die Analyse von LongSAGE Tags ohne entsprechendem EnsEMBL Gen führte zur Identifikation von 1872 Gene, welche bisher noch nicht an das Genom annotiert wurden, aber durch einen UniGene Cluster repräsentiert sind. Zusätzlich konnten 2348 GeneScan Vorhersagen verifiziert und 547 Antisense- Gene identifiziert werden.

Eine Analyse von öffentlich zugänglichen SAGE Bibliotheken zeigte, dass Zielgene von anderen Signaltransduktionswegen als BMP4 auch Cluster im

Genom bilden. Desweiteren wurden die beobachteten Veränderungen in der Genexpression von ribosomalen Proteinen in verschiedenen Geweben unter unterschiedlichen Bedingungen untersucht. Erstaunlicherweise zeigte eine Cluster- Analyse, dass verschiedene Gewebetypen eindeutig abgegrenzte Genexpressionsprofile ribosomaler Proteine haben.

Zusammenfassend bietet die Transkriptiom- Analyse von BMP- induzierter Knorpelentwicklung sowie Somitogenese einen Einblick auf die molekularen Ereignisse, welche beide Entwicklungsprozesse steuern. Generell sind mehrere Signaltransduktionswege sowie eine vielzahl zellulärer Prozesse involviert. Weitere Studien werden zeigen, wie diese Veränderungen in der Genexpression hervorgerufen werden und wie sie in die fein abgestimmten zellulären Ereignisse von Knorpel- und Somitenentwicklung organisiert werden.

Summary

In order to better understand the molecular mechanisms that control chondrogenesis and somitogenesis, a global quantitative expression profiling using SAGE (Serial Analysis of Gene Expression) was performed on a chondrogenic cell line, ATDC5, and on somitic tissues dissected from mouse E10.5 embryos.

Among a total of 43,656 SAGE tags derived from mouse chondrogenic ATDC5 cells (21,875 from uninduced cells and 21,781 from cells induced with BMP4 for 6 h), 139 transcripts were differentially represented in the two libraries ($P \leq 0.05$). Ninety-five of them matched to single UniGene entries (77 known genes and 18 ESTs), but surprisingly, many of them have never been implicated in chondrogenic differentiation. Interestingly, a significant fraction of these genes formed physical linkage groups.

For the study of somitogenesis, the expression profiles in four different subsets of the caudal part of E10.5 embryos was compared: The tail bud (A), the caudal 2/3 (B) and the rostral 1/3 (C) of the presomitic mesoderm, and two pairs of nascent somites (D). A total of 171,639 LongSAGE tags (A: 21,595; B: 50,699; C: 49,732; D: 49,613) were generated leading to the identification of 1007 transcripts differentially represented between at least two of the libraries ($P \leq 0.05$).

The LongSAGE tags with a count of two or more were compared against the current genome annotation of EnsEMBL. The analysis of LongSAGE tags with no corresponding EnsEMBL gene lead to the identification of 1872 genes, that were not yet annotated to the genome, but represented by a UniGene cluster. Furthermore, 2348 GeneScan predictions could be verified with LongSAGE tags and 547 antisense genes were identified.

A analysis of publically available SAGE libraries showed that target genes of signaling pathways other than BMP also cluster to the genome. Furthermore, the observed changes in the expression of ribosomal protein genes in various tissues under different conditions were examined. Surprisingly, cluster analysis showed that different tissue types had distinct profiles of ribosomal protein gene expression.

In conclusion, the transcriptome analyses of both BMP-induced chondro-

genesis and somitogenesis provided an insight on the molecular events during both developmental processes. In general, multiple signaling pathways and a variety of cellular processes are involved. Further study will clarify how these changes in gene expression are brought about and are organized into the concerted cellular events of chondrogenic and somitogenic differentiation.

Acknowledgements

I would like to thank:

Prof. Rudi Balling, doctoral advisor and mentor of my thesis
for his strong support even after he moved to Braunschweig.

Dr. Kenji Imai, supervisor of my thesis
for giving me the chance to develop my projects in my own responsibility, for fruitful discussions and huge advice, and for the contribution to my work[1].

Dr. Chisa Shukunami, collaborator and host (Feb. till April, 2001)
for the contribution to my work[2] and for her great hospitality during my stay in her lab in Kyoto.

Dr. Ulrich Heinzmann, collaborator
for the contribution to my work[1], for valuable advise and for the interesting discussions during lunch.

all the people from my work group and from other groups of the Institute of Developmental Genetics and Institute of Experimental Genetics at the GSF (expecially all the people who ever participated in our joint lab-meetings) for very constructive interactions and a friendly atmosphere.

and all the members of the Department of Molecular Interaction and Tissue Engineering at the University of Kyoto for their help and advice, and for making my life in Japan very easy.

[1] especially for the participation in the dissection of the somitic tissue (3.3.1).
[2] pilot study for ATDC5 (3.2.1) and majority of northern blots in 3.2.5.3.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In the 1970s the advent of DNA cloning technologies revolutionized developmental biology. Currently a similar revolution is taking place. As summarized in Tables 1.1 and 1.2, bulk amounts of nucleotide sequence data is provided through the whole genome sequences of mouse [1] and human [2, 3], the sequencing of a large set of mouse full-length transcripts [4] and ongoing EST (expressed sequence tag) sequencing projects [5]. However, the real challenge of the so-called 'post-genomic' era will be to extract biological information on a large scale from the available sequence data. This includes annotation of genes to the genome (reviewed in [6]) and large-scale gene expression screens (reviewed in [7]), and might finally allow (in conjunction with functional data) to model biological processes (systems biology) (reviewed in [8]).

In 1995 a very powerful tool, Serial Analysis of Gene Expression (SAGE) [9], was published, which is applicable for the first two tasks. Surprisingly, SAGE was given little attention compared to its complement, DNA microarrays [10]. Both methods were published head-to-head in the same issue of Science, and are primarily methods for large-scale gene expression analysis. However, SAGE has certain advantages over microarray techniques like that no prior knowledge about gene sequences is required, that the data is quantitative and that it can be applied to limited amounts of RNA. SAGE is a sequence-based approach that identifies which genes are expressed and quantifies their level of expression. Two basic principles underlie the SAGE methodology (Figure 1.1): (A) Short sequence tags at the defined position within a transcript sequence contains sufficient information to uniquely identify a transcript; (B) the concatenation of tags in a serial fashion allows for an increased efficiency in a sequence-based analysis. With its recent improvement, LongSAGE [11], tags with the length of 21 bp are generated. Such tags are long enough to directly be assigned to the genome (Table 1.3). Therefore,

Table 1.1: Statistics of Human (build 161) and Mouse (build 123) UniGene releases

	Human	Mouse
number of sequences		
Build	161	124
mRNAs	111,183	54,936
HTC	78,46	54,473
EST, 3'reads	1,543,802	1,559,727
EST, 5'reads	2,048,480	1,617,335
EST, other/unknown	688,432	235,786
total sequences in clusters	4,407,974	3,522,257
number of clusters		
sets total	108094	88185
sets contain at least one mRNA	28,412	18,374
sets contain at least one HTC sequence	6231	29518
sets contain at least one EST	106,580	86,997
sets contain both mRNAs and ESTs	26,952	17,313
RIKEN Fantom 1 and 2[4]		
full-length mRNAs	n.a.	60,770
transcriptional units	n.a.	33,409

Statistics for number of available transcript sequences from different projects. EST: expressed sequence tag; HTC high throughput cDNA.

Table 1.2: Statistics of EnsEMBL database

	Human	Mouse
EnsEMBL Version	15.33.1	15.30.1
Genome Assembly Version	NCBI 33	NCBI 30
EnsEMBL genes[1]	24,261	24,948
EnsEMBL transcripts	32,997	32,911

Current EnsEMBL releases of annotated genomes. [1] Not including EnsEMBL EST genes.

LongSAGE is feasible to annotate expressed genes to the genome.

In order to better understand the molecular events that control chondrogenesis and somitogenesis, I performed SAGE on a chondrogenic cell line, ATDC5, and on somitic tissues dissected from mouse E10.5 embryos. Furthermore I performed *in silico* analyses of my own and publically available SAGE and LongSAGE data.

1.1 Chondrogenesis

Appendicular and axial skeletons of higher vertebrates are formed by a multistep process called endochondral bone formation (see Figure 1.2), in which cartilaginous rudiments are replaced by bone [12]. In the embryonic limb,

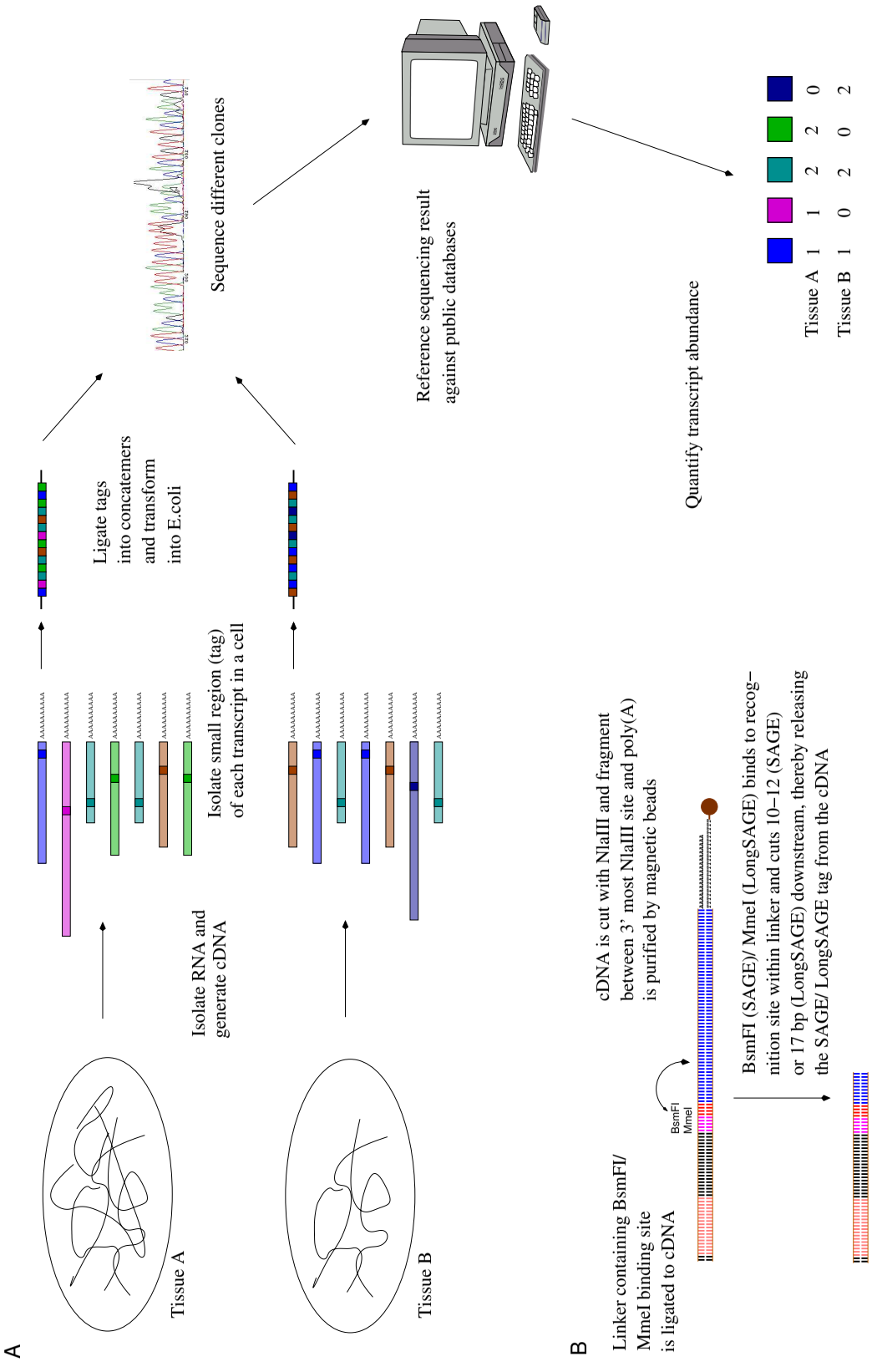


Figure 1.1: (A) Principle of SAGE. (B) Release of SAGE tag.

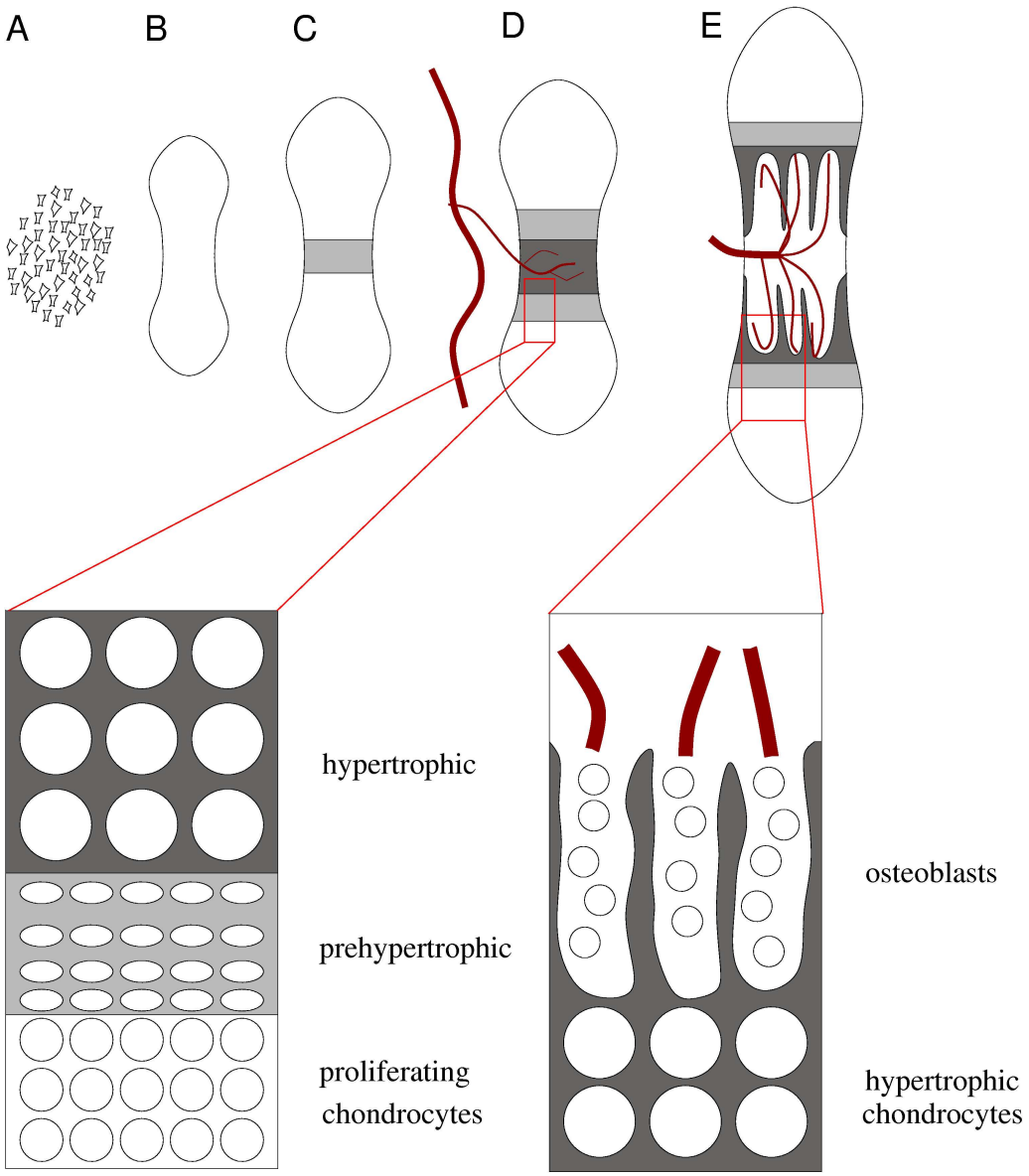


Figure 1.2: Mesenchymal cells (A) condensate and become chondrocytes prefiguring the bone rudiment (B). Primarily proliferating chondrocytes in the center stop to proliferate (C) and differentiate into hypertrophic chondrocytes surrounded by a mineralized matrix. Blood vessels invade the area of hypertrophic chondrocytes (D) and apoptotic chondrocytes are finally replaced by osteoblasts (E).

Table 1.3: Theoretical mapping of SAGE/ LongSAGE tags to genes and to genome

tag length[1] (n)	complexity[2]	tag uniqueness against all **genes**[3] [4]	tag uniqueness against **genome**[3] [5]
10	$1.0 \cdot 10^6$	98.5%	0.0%
11	$4.2 \cdot 10^6$	99.3%	0.0%
12	$1.7 \cdot 10^7$	99.8%	0.0%
13	$6.7 \cdot 10^7$	100.0%	0.0%
14	$2.7 \cdot 10^8$	100.0%	0.0%
15	$1.1 \cdot 10^9$	100.0%	0.1%
16	$1.0 \cdot 10^6$	100.0%	16.7%
17	$1.0 \cdot 10^6$	100.0%	64.0%
18	$1.0 \cdot 10^6$	100.0%	89.4%
19	$1.0 \cdot 10^6$	100.0%	97.2%
20	$1.0 \cdot 10^6$	100.0%	99.3%
21	$1.0 \cdot 10^6$	100.0%	99.8%

Tag uniqueness of SAGE (14 bp) and LongSAGE (21 bp) tags against genes and genomes. A tag must be unique to be able to unambiguously assign it to its corresponding gene. [1]Including recognition site of anchoring enzyme (*NlaIII*, 4 bp in length). [2]Possible number of different tags (with length n) consisting of all four bases (A,T,G and C): Complexity $C = 4^n$. [3]Assuming random distribution of all four bases among transcriptome/genome. [4]Assuming 30,000 genes for human and mouse genome. [5]For mouse genome (length: 2,5 Mb)

the classical experimental model used for studying chondrogenesis *in vivo*, undifferentiated mesenchymal cells of somitic or lateral plate mesoderm origin with a substantial amount of filopoda and large, featureless intracellular matrix start to condensate. Starting in the condensation centers, the cell density and cell packing increases due to targeted cell movement (not mitosis). Rounded cells with only few filopoda, which among themselves are not in direct contact, form many intercellular contacts of tight junctions and/or zona occludens type, thereby shaping the first indication of the cartilage element. At this stage, the intercellular space is extremely reduced. During midcondensation phase, the cartilage element is separated from adjacent non-chondrogenic mesenchyme by a membrane celled perichondrium. In a proximal-to-distal procedure, concomitant with the secretion of a characteristic granular and fibrillar extracellular matrix, cells flatten and progressively become separated from each other (oriented in a right angle to the long axis). These dividing cells eventually increase in size five- to tenfold and regain a round shape. The excretion of high amounts of extracellular matrix continues, however, the composition changes. Next to the proximal, most advanced chondrocytes (hypertrophic chondrocytes), calcification starts at focal sites between collagen fibrils and spreads throughout the extracellu-

lar matrix, forming longitudinal septa with cells surrounded by mineralized cartilage matrix. These areas of hypertrophic chondrocytes are invaded by blood vessels and are ultimately replaced by bone cells ([13, 14, 15, 16])

Key signaling molecules and transcription factors that control the process of endochondral bone formation have been identified in the past years. These secreted signaling molecules include members of the TGF-β superfamily, the Wnt, FGF, Hedgehog families, and parathyroid hormone related peptide (PTHrP). A number of transcription factors of the Hox, Pax, Sox, Runt-domain, Forkhead, and basic helix-loop-helix (bHLH) families are also implicated in endochondral bone formation. Interplays or regulatory feedback loops between theses signaling pathways and the coordinated actions of these transcription factors are thought to play key roles in endochondral bone formation (reviewed in [17, 18]).

Among these signaling molecules, bone morphogenetic proteins (BMPs), which are members of the TGF-β superfamily (except for BMP1), can induce and promote the formation of cartilage and bone by recruiting mesenchymal precursor cells when injected intramuscularly or subcutaneously into an ectopic site [19, 20]. Thus, BMPs are considered to play a pivotal role in cartilage and bone development. BMP signaling is received by specific serine/threonine kinase receptors, which consist of two type I receptor subunits (BMPRIA and BMPRIB) and two type II receptor subunits (BMPRII), and is transduced to the nucleus by Smad proteins (the canonical BMP-Smad pathway). BMP ligands as dimers bind to the type II receptor (RII), leading to the recruitment of the type I receptor (RI). Formation of a ternary complex consisting of ligand/RII/RI results in phosphorylation of RI by RII. The activated RI in turn phosphorylates a subgroup of Smad proteins including Smad1, Smad5, and Smad8 (collectively called receptor-regulated Smads, or R-Smads). Phosphorylated R-Smad can interact with Smad4 to form a complex, and this R-Smad/Smad4 complex translocates into the nucleus, where downstream target genes of BMP signaling are either activated or repressed by the Smad complex together with various nuclear cofactors like p300, CBP and SNP1. BMP signaling is controlled by a number of extracellular, cytoplasmic or nuclear modulators. Aside from this canonical Smad pathway, BMP signaling is also transduced via the MAP kinase pathway (reviewed by [21, 22, 23, 24]).

Bmp2 and *Bmp4* are expressed in mesenchymal cells prior to condensations and later in the perichondrium, with the highest level adjacent to prehypertrophic and hypertrophic chondrocytes. Functions of BMP2 and BMP4 during endochondral bone formation have been extensively investigated in overexpression studies, showing their involvement in the condensation phase as well as in the progression of chondrocytes to hypertrophy.

Overexpression of BMP2 or BMP4 leads to the formation of broader cartilaginous rudiments due to increased proliferation and a delay in hypertrophy of chondrocytes [25]. Similar results have been seen by overexpression of constitutively active forms of the type I receptors, BMPRIA or BMPRIB. *Bmpr1b* is highly expressed in pre-cartilagous condensations, but not in differentiating chondrocytes, while *Bmpr1a* is specifically expressed in prehypertrophic chondrocytes. Accordingly, the enhanced mitosis of mesenchymal cells prior to condensations is specifically mediated by BMPRIB, while the delay of hypertrophic differentiation is realized by overexpressing the constitutively active form of BMPRIA, but not that of BMPRIB [26]. Inhibition of BMP2/4 signaling by overexpression of Noggin [27], a potent antagonist of BMPs, a dominant-negative form of BMPRIB (but not that of BMPRIA) [26] or dominant-negative BMPRII *in vitro* [28] delays or inhibits chondrogenic condensations.

The molecular mechanism of chondrogenesis has been extensively studied in *in vitro* systems with either primary chondrocytes or cells from established chondrogenic cell lines. Mouse embryonic carcinoma-derived cell line ATDC5 provides an excellent model system to study chondrogenesis *in vitro*. In a long-term culture system, all steps of chondrogenic differentiation from the pre-condensation stage to the calcified cartilage stage can be reproduced with ATDC5 cells [29, 30]. When recombinant BMP2 or BMP4 is administered to undifferentiated ATDC5 cells at confluency, the cells synchronously and rapidly start to differentiate into chondrocytes [31]. Similarly, overexpression of a constitutively active form of BMPR-IA or BMPR-IB induced chondrogenesis of ATDC5 cells [32]. Conversely, overexpression of dominant negative forms of BMPR-IA or BMPR-IB failed to induce formation of cartilage nodules in ATDC5 cells although the formation of condensing areas was induced [32, 31].

Despite the well-established implications of BMP signaling in chondrogenic differentiation, molecular events downstream of BMP signaling are largely unknown.

1.2 Somitogenesis

The segmental nature of the body plan in vertebrate embryos is best represented by the metameric organization of somites. Somites are blocks of cells, which in a strict anterior to posterior sequence periodically bud from the rostral end of two rods of unsegmented mesoderm lying laterally to either side of the neural tube (called presomitic mesoderm in mouse and segmental plate in avians; for convenience, the abbreviation PSM is used for both) (see Fig-

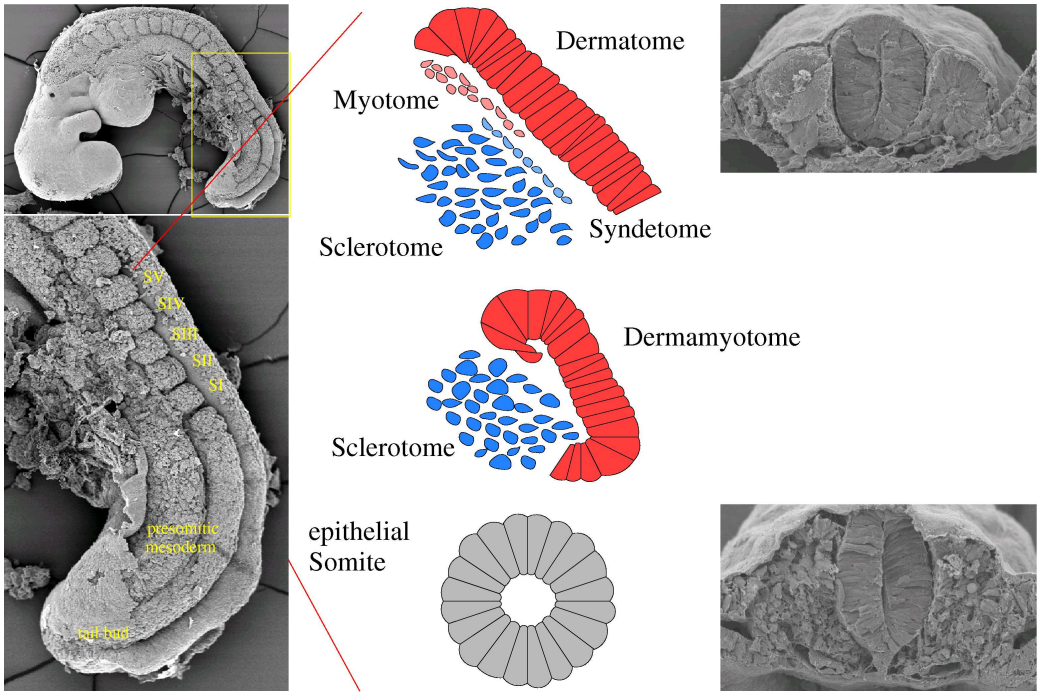


Figure 1.3: Paraxial mesodermal cells derived from primitive streak and later from tailbud form the presomitic mesoderm (PSM). At the anterior end of the PSM epithelial somites bud at a constant rate. Over time, epithelial somites differentiate into the dermatome, myotome, syndetome and sclerotome. Scanning electron microscopy pictures courtesy of Ulrich Heinzmann.

Table 1.4: Somitomeres in different species

Species	# somitomeres PSM/ segmental plate	# somites formed of cultured PSM/ segmental plate		Ref[2]
		with	without tail bud/ node	
Chick (stage 9-13)		$\geq$ 18-20	11.9 ± 1.1	[36]
Chick/ Jap. quail (stage 9-16)[1]	10.0 ± 1.5		10.0 ± 1.5	[37]
Jap. quail (stage 5-21 pairs of somites)			12.1 ± 1.9	[38]
Mouse E8.5	5.8 ± 0.2		6.9 ± 0.2	[39]
Mouse E9.5	6.1 ± 0.3	$\geq$ 12.3 ± 0.6	6.4 ± 0.2	[39]

[1]Numbers for chick and japanese quail combined. [2]Reference.

Table 1.5: Number of cells contained in potential somites and the newly segmented somites of mouse embryos

Stage	Somite	potential somites in the presomitic mesoderm						
	I	0	-I	-II	-III	-IV	-V	-VI
E8.5	286 ± 29	311 ± 23	319 ± 27	275 ± 40	222 ± 19	154 ± 15	99 ± 9	81 ± 8
E9.5	963 ± 5	1025 ± 140	882 ± 95	759 ± 68	561 ± 68	371 ± 29	374 ± 26	259 ± 26
E10.5	1203 ± 184	1209 ± 74	1086 ± 179	809 ± 99	554 ± 61	367 ± 53	270 ± 33	
E11.5	1614± 178	1420 ± 122	1370 ± 109	1243 ± 211	672 ± 112	486 ± 73	321 ± 36	

Taken from [40].

ure 1.3). Cells are constantly added to the caudal end of the PSM, initially by ingression through the primitive streak and later by cell division within the tail bud, keeping the length and cell density of the PSM relatively constant. Nascent somites, which are formed every 90 minutes in mouse (varies among species), are initially epithelial spheres, but progressively the ventral part de-epithelializes and form the sclerotome. The remaining dorso-lateral epithelial cap of cells, the dermomyotome, further becomes sub-divided into the medial myotome and the dermatome (reviewed in [33, 34]). In addition, a fourth compartment, the syndetome, which is derived from a dorsolateral domain of the early sclerotome and contains progenitor of tendons, recently was identified [35].

Even before somites form, the PSM plate is segmented into somitomeres, units composed of loose mesenchymal cells organized into squat bilaminar discs, separated by a deep transverse groove. Somitomeres have been observed in most species, including chick [41], mouse [42] and Japanese quail [37], however, as shown in Table 1.4, between species they differ in number. Within a particular species, the number of somitomeres is almost constant over different developmental stages, although its length and cell number

varies [36, 38, 39]. Explants of PSMs cultured *in vitro* showed, that after the removal of the tail bud/ node, the number of somites is almost identical to the number of somitomeres (see Table 1.4) [36, 38, 37, 39]. *In vitro* tissue culture also helped to narrow down the tissues, which are necessary for somite formation. In the absence of the primitive streak, no somites are formed, whereas 'somite centers' (tissue lateral to and slightly caudal to Hensen's node, originally thought to be essential for somite formation) or the node as well as the notochord and the neural tube are not necessary for somite formation. However, in the absence of the node, less somites are formed (as many somites as there are somitomeres in the PSM) [43, 44]. Removal of axial structures and endoderm also do not affect somite formation [36]. Contrariwise, surface ectoderm is crucial for somite formation and elongation of the PSM [45]. Reversion of the whole PSM or scrambling of the PSM within otherwise intact embryos can be compensated and somites still form [46, 47], but only if notochord and neural tube are not separated by a physical barrier [47]. However, the anterior-posterior polarity of the formed somite is not overruled and corresponds to the original orientation of a somitomere within the PSM [46]. Measurements of the mitotic activity within the tail bud/ node and the PSM showed an elevated rate of cell division within the tail bud/ node compared to the PSM [48, 40]. Obviously, the tail bud/ node is the major source of cells compensating the removal of cells everytime a new pair of somites is formed. Therefore it is not surprising, that in the absence of the tail bud/ node, no more somites are formed than there are somitomeres in the PSM. Within the PSM, cells at the same level along the A/P axis show cell synchrony [49], which is consistent with observations that cell cycle inhibitors as well as heat shock leads to periodic anomal somites every 10 hours in chick embryos, which exactly corresponds to one cell cycle [50, 49], suggesting the existence of a cyclic event underlying somitogenesis, which is linked to the cell cyclus. These observation of a cellular oscillator are in accordance with more recent molecular data. A variety of genes have been identified that show a dynamic expression pattern within the PSM, recurring everytime a new somite is being formed. This lead to the postulation of a molecular oscillator, which is established and regulated by concerted actions and regulatory interactions of multiple signaling pathways including those of Notch-Delta, Wnt and FGF (reviewed in [51, 52]).

However, the precise mechanism of how the yet identified genes and its protein interplay or function remain unclear. Furthermore, it is expected that additional yet unknown important players may be involved.

Chapter 2

Materials and methods

2.1 Molecular biology methods

2.1.1 RNA extraction

Total RNA was prepared from ATDC5 cells harvested at various time points of BMP4 induction by the single-step method of [53]. RNA from somitic tissue was extracted using the Dynabead mRNA DIRECT kit from Dynal. Somitic tissue was sonicated prior to library constraction with a microtip on a Branson Sonifier W-450 (Branson) with settings output control = 7 and duty cycle = 40%.

2.1.2 Cycle sequencing

Cycle sequencing was performed with the ABI PRISM BigDye Terminators v3.0 and v3.1 Cycle Sequencing Kit (Applied Biosystems) on an ABI PRISM 3100 Genetic Analyzer (Applied Biosystems).

2.1.3 Cell culture of ATDC5

ATDC5 cells were maintained in a medium consisting of a 1:1 mixture of DME and Ham's F-12 (DME/F12) medium (Flow Laboratories) containing 5% fetal bovine serum (JRH Biosciences), and ITS, i.e., 10 μg/ml bovine insulin (Roche), 10 μg/ml human transferrin (Roche), and 3 (10^{-8} M sodium selenite (Sigma), as previously described [29]. The inoculum size of the cells was 6 (10^4 cells/well in 6-multiwell plates (Corning Glass) at 37C under 5% CO2 in air. The medium was replaced every other day. For induction of chondrogenesis, human recombinant BMP4 (R & D Systems), diluted with

PBS containing 0.1% bovine serum albumin (Sigma), was added to ATDC5 cells at confluence.

2.1.4 Serial analysis of gene expression

SAGE and LongSAGE libraries were constructed according to the standard protocol ([9, 11] and the updated version of the protocol available at www.sagenet.org), with some modifications according to published proposals [54, 55, 56]. Instead of *NlaIII*, the isoschizomer *Hsp92II* (Promega) was used, because of its advantage with stability in storage. To eliminate the possibility of cross-contamination of ditag molecules (templates for ditag amplification) between the different libraries, a distinct (library-specific) linker/primer combination for each library construction was used.

2.1.5 GLGI

The GLGI (Generation of longer cDNA fragments from SAGE tags for gene identification) method was performed as described [57, 58] with the following modifications: cDNA was synthesized by using a biotinylated oligo(dT) oligonucleotide containing a primer-binding site as described [57], and processed according to the protocol for SAGE library construction up to the linker ligation step. Linker-ligated cDNA was used as template for amplification of fragments between SAGE tag and poly A, using a primer specific to the oligo(dT) oligonucleotide and a tag-specific primer. The tag-specific primer contains 10 (or whenever possible, 11) bases of the SAGE tag, the CATG and as many bases from the linker as necessary for getting a reasonable melting temperature. Amplified fragments were gel-purified, cloned into pCR-TOPO (Invitrogen) and sequenced.

2.1.6 Probes used for northern blot analysis and whole mount *in situ* hybridization (ATDC5)

DNA fragments as specific probes corresponding to SAGE tag genes were amplified by RT-PCR from ATDC5 RNA (either undifferentiated or induced) with specific primer pairs listed in Table 2.1, and cloned into pCR-TOPO. All cloned RT-PCR products were confirmed by sequencing. The probes for Col2a1, Bmp4 and noggin were as described [31, 59].

2.1.7 Northern blot analysis

Conditions for Northern blot analysis were as described [31].

Table 2.1: Primer pairs: ATDC5

Tag #	Gene	Forward	Reverse	Size (bp)
D2	*Vim*	CAGCAATATCAGCAGCAACG	CTGGTCCACAGACGGTGG	522
D20	*Igfbp5*	TTGAAGTGAATCCACCAAGCC	TTGGATCCTGAATCAGTTACC	521
D21	*Ptmb10*	GTCTCTTGCTGCAGCAACG	TTCACAGTGCAGCTTGTGG	328
D26	*Bgn*	AATCCATGACAACCGTATCCG	TTGTTGAAGAGGCTGATGCC	525
D31	*Actg*	CTGCATCATCTTCCTCCTAGG	AAGGCAACTAACAACCGATGG	964
D42	*Gas1*	TCCAATGGACTTGGAGAAGG	TAAGACACGGGTGCAGAGG	362
D62	*Itgp*	CAAGGATCAGCCTGTTCTTACG	GTTCCTTCCAGCTGTGAGTCG	658
D72	*Fxc1*	CTTGGAGCATGGAGCAGC	CCAAGGTCCATAGAGGAGCC	366
D74	EST	GCTAGCTTGTACAGGTTACAGGTTGGAGAA	AGATCTAATATTGAAGTCAGGCAGGTCTGT	574
U77	*Ptn*	GAGTGTGTGCGTGCCTACC	GCCTCTCTCCTCTCAGTCTGC	960
U78	*F2rl1*	TGAACATCACCACCTGTCACG	GATTGGTGCAGAGACAGACAGC	707
U81	*Cox6c*	CGTTGGTGTAGAGGACATTGG	TCATAGTTCAGGAGCGCAGG	330
U82	EST	GCTAGCCTAAATAAAGCAGAGAAGGCTTGG	ACGCGTAAGTCCAGGTCTTTGCCTATAGTG	661
U84	EST	AGTTCCTGGTGGTTCCTGTAGG	CAGCATGCAGTAGACAGAAGCC	540
U87	EST	AGAAGGATGATGAACGTGTCC	TCCAATCATCACGACTACGC	501
U88	EST	GTGTGAACTCTGACAATAGCGG	GAGCGTACAGTCTATCACCTGC	505
U90	*Sparc*	ATCCATGAGAATGAGAAGCGC	AGTCGAGAAGACAGCAAGGTCC	935
U91	*Sui-rs*	GCTAGCGGAAAAGGAATCGTATCGTATGTC	ACGCGTCTCAACCTGTTTAAATGAGGGACT	573
U92	*Osf2*	AAGAGAATGTTAACCAAGGACCTG	GTCACAATGTCTTTCTTGTTCACC	545
U94	*Fn1*	GCCGAATGTAGATGAGGAGG	AGTTGACACCGTTGTCATGG	542
U97	*Pcbp2*	GCTAGCAGAGAATTATCACTTTGGCTGGAC	ACGCGTCCTTGAATGGTATAGGCCTCTAGA	523
U98	*Hspa8*	CAAGGCTGAGGATGAGAAGC	ATCCACCTCTTCAATGGTGG	600
U99	*Ywhae*	ATAACCTGACGCTGTGGACC	AACTGTTACCAGCACCATGC	699
U100	*Hk1*	GTGAGATTGGACTCATCGTGG	GCATGATTCTGGAGAAGTGTGG	612
U103	*Tgfbi*	GGAATCTGACGTCTCCACTGC	CAGATCTCAATATGGTGCGCC	1033
U106	*Idb3*	GCCTCTTAGCCTCTTGGACG	CAGCTCTTATGCTGCCTTGG	623
U107	*Fin14*	TCTGTTCAAGCTGCGTTGAG	GTTATGGCTACACGCCAATG	967
U110	*Vcp*	ATCATTGGAGCTACCAACAGGC	CAGACTGAGGAATGGAGCAGG	680
U111	*Atp6g1*	GCTAGCCATCCAGCAGCTACTGCAGG	ACGCGTAAAAAGTGAAGGGTCCTACAACAG	474
U117	EST	AGAACAAGTTGAGGAACGGC	CCGTTCTAATCCTCCTGTGC	555
U126	*Idb2*	CAACATGAACGACTGCTACTCC	CATTCAACGTGTTCTCCTGG	421

Primer pairs used to amplify northern blot and *in situ* probes.

2.1.8 Quantitative real-time PCR

Quantitative Real-Time PCR was carried out using the LightCycler DNA Master SYBR Green I kit (Roche) on a Roche LightCycler instrument.

2.1.9 Whole mount *in situ* hybridization

Whole mount in situ hybridization was performed as described [60].

2.2 Methods for experimentation on animals

2.2.1 Preparation of mouse embryos

C57BL6 females and males were purchased through Charles River (Germany). The day in which a vaginal plug was observed for mated females was considered as day 0.5 of embryonic development (0.5 day post coitus [dpc] or E0.5).

2.2.2 Microdissection

Embryos were fixed with etched tungsten needles (0.2 mm) on plates coated with Sylgard 184 (Dow Corning). Cuts were made with surgical knifes from Surgical Specialities Corporation (Reading).

2.3 Data processing

The following paragraphs will describe the algorithms underlying the respective programs for the data analysis. All custom programs were written in Perl (practical extraction and report language) [61]. Non-standart APIs, that were implemented into the programs, are listed in Table 2.2. Existing programs integrated into the own programs are listed in Table 2.4 and external data sources are summarized in Table 2.3. For clarification, database and database tables are highlighted (`database table`), as are sequences (`ATGC`). Species are indicated by *`SPECIES`* (shortcut: *`Sp`* if original file contains Mm or Hs prefix). Version numbers and dates are denoted by *`VERSION`* and *`DATE`*. Alternation is symbolized by '|' and optional items are within brackets ([]). All schemas for the database tables are given in the appendix A.

Table 2.2: Bioinformatics Perl APIs utilized

name	version	reference
BioPerl	1.2.1	[62]
EnsEMBL	13	[63]
GO	n.a.	[64]

Additionally, a large number of non-bioinformatics Perl APIs were downloaded from the Comprehensive Perl Archive Network (www.cpan.org).

Table 2.3: Version/ date of datasets utilized

name	version/date	reference
EnsEMBL	13	[63]
GO	Aug 2003	[64]
InterPro	7.0	[65]
MGD	Sep 19th 2003	[66]
GXD	Sep 19th 2003	[67]
UniGene	Builds 161[1]/ 123[2]	[68]

[1]Human and [2]mouse.

Table 2.4: Locally installed bioinformatics tools utilized

name	version/date	reference
blastz	May 14th, 2003	[69]
NCBI BLAST	2.2.6	[70]
MegaBLAST	2.2.6	[71]
phred	0.020425.c	[72]
InterProScan	3.2	[73]

2.3.1 SAGE tag assignment

SAGE First, a database table is constructed comprising of all sequences within a UniGene release containing the 3' end of the transcript (`3prime_unigene_`*VERSION*`|_fantom`*VERSION*, database *SPECIES*`_master`). Starting with the non-redundant UniGene sequences (file *Sp*`.seq.all`), all cDNAs annotated as being full-length (description line contains 'full-length enriched', 'complete cds' or 'RIKEN cDNA *riken_clone_id* gene') are taken, as well as 3' ESTs containing at least either a polyadenylation signal[1] or a polyadenylation tail[2]. 5' ESTs as well as partial cDNAs are only considered if they contain both polyadenylation signal and tail. Sequences with wrong orientations containing polyadenylation signal and tail[3] at the beginning of the sequence in reverse-complement orientation are changed to 5' to 3' orientation. All filtered sequences, which now can be considered as both containing the 3' end of the transcript as well as being in 5' to 3' orientation are written into the database table. Furthermore, the sequences are linked to the RIKEN Fantom2.0 Representative Transcript units (RTS) by `MegaBlast`. From all sequences[4], the 3' most `CATG` (last `CATG` before poly(A) tail) is identified and the 10 bases[5] downstream are extracted[6]. The sequence of each kind of tag together with the relative abundance[7] within a UniGene cluster for both full-length cDNAs as well as ESTs is written to another database table (`map_`*TAGLENGTH*`_unigene_`*VERSION*, database *SPECIES*`_mapping`). For the final tag assignment only entries with at least one cDNA or with a relative abundance of at least 10% of the 3' ESTs are retrieved from the table.

LongSAGE Initially, all tags are loaded in a master table, called `tags` (database *SPECIES*`_longSAGEmapping_`*VERSION*). For each tag, every ap-

[1] within 50 bp to the end of the sequence; in addition to the canonical polyadenylation signals (AATAAA and ATTAAA) [74], four additional polyadenylation signals (AATTA,AATAAT,CATAAA, AGTAAA), occurring with a frequency between 5.7 and 8.4% [75] in Human, are used.

[2] 3' ESTs from RIKEN are only deposited to GenBank if they contain a poly(A) tail, although the poly(A) tail has been removed before submission.

[3] Unless otherwise specified for the particular sequence, the correct orientation 5' to 3' (default) has to be assumed. An alignment against other members of the same UniGene cluster is not appropriate, since due to the UniGene algorithm (uses MegaBLAST[71]), the same UniGene cluster potentially also contains antisense transcripts.

[4] for RIKEN ESTs, poly(A) tail is artificially added.

[5] or 11, if 11th base is also considered.

[6] sequences without CATG are ignored.

[7] frequency of particular tag divided by total number of 3' sequences within UniGene cluster, from which the tag can be derived.

pearance in the EnsEMBL genomic sequence[8] (hereafter called 'Genome hit') is written to `genome_hits`. All EnsEMBL genes and EST genes spanning the Genome hit are stored in `genome_hit_transcripts`. The databases are linked through its unique identifers (e.g. `tags.id` = `genome_hits.tags_id`; `genome_hits.id` = `genome_hits_transcripts.id`). In addition, mappings to UniGene ('UniGene hits') are retrieved from `map_17_unigene_`*`VERSION`*[9], if the LongSAGE tag can be derived from at least one cDNA or 10% of the 3' ESTs, and written to `unigene_hits`. The representative sequence of each UniGene hit is compared against all EnsEMBL transcripts (genes and EST genes) using `BLAST` and the highest scoring hit with at least 92% percent identity over at least 250 bp[10] is taken as EnsEMBL gene/ EST gene corresponding to the particular UniGene cluster. To take the redundancy of EnsEMBL genes and EST genes into account, all other genes and EST genes, whose annotated position on the genome overlaps with the corresponding EnsEMBL gene/ EST gene and also possess a percent identity of at least 92% percent identity over at least 250 bp, are together with the corresponding gene/ EST gene written to `unigene_hit_transcripts`. Again, the tables are linked thorough its unique identifers (`tags.id` = `unigene_hits.tags_id`; `unigene_hits.id` = `unigene_hits_transcripts.id`). Genome hits and UniGene hits are merged, if at least one EnsEMBL gene or EST gene in `genome_hits_transcripts` and `unigene_hits_transcripts` are identical. It should be noted, that in this way, multiple UniGene hits could be assigned to a single Genome hit, but never multiple Genome Hits to the same UniGene hit. For each tag, its mapping Genome hits and/ or UniGene hits are written to `hits`. Multiple Genome hits, multiple UniGene hits that can not be merged to a single Genome hit as well as non-merged Genome hits and UniGene hits result in multiple 'Hits' (multiple entries in `hits`).

LongSAGE evaluation based upon ESTs aligned to the genome
For each single Genome hit, all ESTs significantly aligning to the corresponding chromosomal position (all included in EnsEMBL table `dna_align_feature` with $\geq$ 92% percent identity) are stored in `genome_hit_ests`. Next, the sequences of all hitting ESTs (temporarily stored in `genome_hit_est_sequences`) are analyzed for sequence homology using `BLAST` against all En-

[8]all possible SAGEtags (17 bp downstream of any *CATG* on both strands) in the genome are initially extracted and written to `table map_17_genome_`*`VERSION`*.

[9]generated as described for SAGE tag assignment, with the exception that tags are 17 bp long.

[10]empirical values according to [76]; due to the presence of alternative splice or polyadenylation forms, polymorphisms or sequencing errors (of both genomic and transcript sequence), the alignment of an EST to its transcript is not always 100%.

sEMBL gene and EST gene transcript sequences. According to the results the LongSAGE tags are categorized (stored in table `est_mapping`) into (1) EnsEMBL transcript(s) plus aligned EST(s), (2) EnsEMBL transcript(s) without aligned EST(s), (3) aligned EST(s) with homologous EnsEMBL transcript maximally 10000 bp upstream (3' UTR is not completely annotated to EnsEMBL), (4) aligned EST(s) with exons of homologous EnsEMBL transcript flanking Genome hit (not annotated Intron), (5) aligned EST(s) with homologous EnsEMBL transcript in opposite orientation (antisense transcript), (6) aligned EST(s) without homologous EnsEMBL transcript in vicinity (putative novel gene) as well as (7) Genome hits without any hit.

Verification of GeneScan prediction by LongSAGE tags All Genome hits without a mapping EnsEMBL gene are analyzed for cDNAs predicted by GeneScan to the corresponding chromosomal position.

2.3.2 Statistical analysis

For statistical evaluation of SAGE counts, the method described by Audic and Claverie [77, 78] was used.

2.3.3 Virtual subtraction

After loading all publically available SAGE data into separate tables within database `SAGE_data`, a `project_`*`NAME`* database table (database `SAGE_project`) containing tag and tag-per-million counts for all tags of the two ATDC5 libraries as well as the libraries to be 'subtracted' against is created (column names are preceeded by 'vs_'). Then, the tags exclusively observed in the ATDC5 libraries compared to all or to subsets of the publically available SAGE libraries are determined by selective queries against the database.

2.3.4 Link of mouse SAGE tags Mouse Genome Informatics

For Mouse, the Mouse Genome Informatics (MGI) Marker ID [66] is used as a primary ID. UniGene IDs as well as EnsEMBL genes and EST genes[11] are linked to MGI Marker ID (according to entries in the file `MRK_Sequence.rpt` (MGI) (written to table `link_mgd_`*`DATE`* in database *`SPECIES`*`_master`)).

[11] MarkerSymbol in table dbxref returns the official gene symbol; not all EnsEMBL genes and EST genes have an gene symbol associated.

2.3.5 automated annotation of LongSAGE tags

GeneOntology GeneOntoloty (GO) [64, 79] annotations are directly retrieved from a local copy of the complete core GeneOntology MySQL database. They are linked to the according genes (LongSAGE tags) through the MGI Marker ID.

InterProScan Initially, every SAGE tag to be analyzed is written to a database table (`annotation` in database *SPECIES* `_longSAGEannotation_`*VERSION*) and only tags with single hits in *SPECIES* `_longSAGEannotation_`*VERSION* are further processed. Also tags with no hit, but have only one Genome hit with an associated GeneScan prediction including the LongSAGE tag are taken. For all mappings, that have an EnsEMBL gene[12] associated, analysis results using InterProScan [73] against Pfam [80], Prosite [81] and Prints [82] that are provided through EnsEMBL are retrieved and stored into table `go_similarity`. For all other tags, the cDNA sequence(s)[13] of its corresponding gene is analyzed in all three frames, but only in sense orientation[14] by InterProScan [73] against Pfam [80], Prosite [81] and Prints [82]. The output is parsed and the database ID for each significant hit to the three databases is written into table `go_similarity`. This original database ID can now be linked to GeneOntology terms using the files *PROTEIN_DATABASE*`2go`, loaded into tables *PROTEIN_DATABASE*`2go`.

2.3.6 Chromosomal localization of differentially expressed SAGE tags

First, a database table is created, containing a non-redundant list of EnsEMBL genes, EST genes and GeneScan predictions (`cluster2chromosome_ensembl_`*VERSION*`_unigene_`*VERSION*). Because of the overlap between EnsEMBL genes and EnsEMBL EST genes[15], and since most of the EnsEMBL

[12] are annotated based upon protein sequences.

[13] the sequence is retrieved in the following with the following ranking (if not available, the next lower ranked source is used): EnsEMBL ESTgene associated with mapping: all of its transcript sequences are used; MGI ID: all GenBank entries (`sequence_EMBL`) provided through MGI (entries in `link_mgd_`*DATE* in database *SPECIES*`_master`); UniGene ID(s): representative sequence (from `sequence_unigene_unique_build`*VERSION*); GeneScan prediction: sequence of prediction.

[14] orientation is specified by LongSAGE tag; in case of multiple cDNAs for a single tag, identical protein sequences generated by translation are removed.

[15] EnsEMBL genes are solely annotated based upon protein sequences from the same or other species, and EST genes are based upon cDNA and EST sequences [63].

genes and EST genes are at least partially also predicted by GeneScan, EnsEMBL ESTgenes are only considered if they do not overlap with EnsEMBL genes, and GeneScan predictions only if no overlap to EnsEMBL genes and EST genes is detectable. Therefore the table is initialized with all EnsEMBL genes mapped to the genome. In a next step, EnsEMBL EST genes are compared against the genes in the table. If no EnsEMBL gene(s) spatially overlap with the EnsEMBL EST gene[16], the EnsEMBL EST gene is directly written to the table. Whenever an overlap exists, all overlapping transcripts of both EnsEMBL gene(s) and EnsEMBL EST gene(s), which are annotated to the same genomic strand, are compared against each other by `BLAST` (`bl2seq`). If any combination of transcripts give rise to a significant hit (1e-30 or lower over at least 250 bp[17]), both EnsEMBL gene and EnsEMBL EST gene are considered to represent the same gene. Otherwise, the EnsEMBL EST gene is written to the table. GeneScan predictions are compared to both EnsEMBL genes and EnsEMBL EST genes in the same way and only written if there is no overlap to any of them. Next, a `BLAST` database with all transcripts for the written EnsEMBL genes, EST genes and GeneScan predictions is generated(`formatdb`). It is queried (`blastn`) with all representative UniGene sequences (file *Xx*`.seq.uniq`, previously loaded into table `unigene_sequence_unique_`*VERSION*). For each UniGene cluster, the best hit is identified (at least 92% identity identity over a length of at least 250 bp[18]; a ranking of EnsEMBL gene - EnsEMBL EST gene - GeneScan prediction is applied, meaning that hits to EnsEMBL EST genes are only considered if there is no hit matching those criteria to any EnsEMBL gene, and hits to GeneScan predictions, only if no EnsEMBL gene or EST gene hits to the UniGene cluster). It should be noted, that for one entry in the database multiple hits to UniGene are possible, but a UniGene cluster is never assigned to more than one gene. Next, for every entry in `cluster2chromosome_ensembl_`*VERSION*`_unigene_`*VERSION* in ascending order on each chromosome, all possible tag sequences mapping to the associated UniGene cluster(s)[19] are retrieved from `map_10_unigene_`*VERSION*`|_fantom-`*VERSION*[20]. For each of the SAGE tags, the counts in the two libraries analyzed are retrieved and analyzed for being significantly differentially ex-

[16]start and end of both do not overlap.

[17]or, if the queried sequence is shorter (over the whole queried sequence).

[18]Although UniGene cluster and a different very homologous gene might have a percent identity higher than 92%, the score for the right gene is higher and is therefore considered.

[19]multiple UniGene cluster could represent the same gene and are thus mapped to one EnsEMBL gene, EST gene or prediction.

[20]tags mapping to more than one UniGene clusters are not considered.

pressed[21]. Whenever two or more tags corresponding to EnsEMBL gene(s), EST gene(s) and/or GeneScan prediction(s) within an interval of up to 1 Mb are significantly differentially expressed, the interval is analyzed (see). Intervals might be extended, if the next EnsEMBL gene, EST gene or prediction with significantly differentially expressed tag(s) is localized less than 1 Mb downstream. All intervals are dumped to a separate file.

2.3.7 Analysis of syntenic regions between mouse and human

Each interval, defined by the two most distance of the outermost two genes, is extended by additional 10,000 bp. For the whole sequence, all 'syntenic' blocks[22] for the other species (Human, if interval is from mouse, or vice versa) are retrieved from the EnsEMBL `ensembl-compara` database. The blocks are separately analyzed for both species by ordering all blocks and analyzing the length of non-conserved gaps in between. If not all blocks for the second species are on the same chromosome, or if a non-conserved gap in between conserved blocks is longer than 50 kb[23], the analysis is stopped. Then, both chromosomal fragments are compared against each other by `blastz` to identify conserved non-coding sequences (CNS) (at least 70 percent identity over at least 100 bp). Additionally, the percentage of genes with an orthologue in the other species (by reciprocal `BLAST` analysis of protein sequences) is measured. Conserved (within CNS of both mouse and human) and non-conserved binding sites are identified by the DNA motifs listed in Table 2.5. All features retrieved in this way are plotted to two separate pictures (one for the genomic fragment in each species) and manually evaluated to decide whether conserved synteny applies to both genomic fragments.

[21] thus, a gene would still be considered as being differentially expressed, when one transcript form (resulting in one particular SAGE tag) is statistically differentially expressed, even if the sum of all tags mapped to this particular EnsEMBL gene, EST gene or prediction are not; multiple differentially expressed transcripts for one gene are also possible.

[22] Blocks are generated by reciprocal BLAST analysis are called syntenic in the ensembl API. However, strictly speaking, without subsequent analysis the blocks itself can not be considered as syntenic, since they are too short; also, the presence (but not necessarily same order) of multiple genes genes and other features is required for conserved synteny (see [83] for a definition of 'conserved segment' and 'conserved synteny').

[23] the maximum allowed gap is rather big. However, since the genomic DNA has been masked for repeats prior to reciprocal BLAST (`blastz`) analysis, large gaps could be introduced in this way. However, break points are avoided by disallowing hits to other chromosomes or other regions within the same chromosome.

Table 2.5: Consensus sequences for downstream DNA binding sites of Shh, cJUN, TGF-β and cMYC pathways

Factor	Binding motif	Reference
Shh		
GLI1	`GACCNCCCA`	[84]
TGF-β		
Smad3	`GTCTGG`	[85]
Smad4	`GTCTMGNC`	[85]
cJUN		
AP1	`TGASTCA`	[86, 87]
CRE-binding protein1/cJUN heterodimer	`TGACGTYA`	[88]
cMYC		
cMYC/MAX heterodimer	`CACGTG`	(reviewed in [89])

2.3.8 House-keeping genes

From a `project` (see section 'virtual subtraction') database table (database `SAGE_project`) containing all SAGE libraries with more than 50,000 tags (normalized to tags-per-million values), tags are determined that are present within every single library and mean as well as standard deviation are calculated. As a measure for the smallest changes over all libraries, the ratio of mean and standard deviation is used (the ratio is indirectly proportional to the gene expression change over the libraries).

2.3.9 Clustering of ribosomal protein genes

For all mapped human ribosomal and human mitochondrial ribosomal genes mapped in [90, 91, 92] the corresponding human EnsEMBL stable ID is retrieved through EnsMart (using GenBank accession IDs provided in the paper), as well as the syntenic mouse orthologue (EnsEMBL stable ID). If the clone with the GenBank accession ID is not mapped to EnsEMBL, the sequence is manually searched by `BLAST` against the genomic sequences. If the mouse orthologue can not be automatically determined, it is manually retrieved from the syntenic mouse chromosomal fragment. All obtained mouse genes are confirmed by (if available) mapping data within the Mouse Genome Informatics database. For all transcripts of the determined EnsEMBL genes (in both species) any UniGene cluster with sequence homology of at least 92% percent identity over at least 250 bp (if one UniGene cluster hits to multiple ribosomal protein genes, only that with the highest product of percent identity and hit length is used) are taken, and all single-hit reliable SAGE tags are written to a `project` table. A dataframe with the counts for all tags mapping to every ribosomal protein gene (normalized to tags-per-million) is

applied to hierarchical clustering (euclidian distance) using R[93].

Chapter 3

Results

3.1 Tag to gene mapping

To evaluate SAGE data (in the following paragraphs, the term SAGE will be - unless otherwise specified - used for both SAGE and LongSAGE), each single tag has to be assigned to the transcript it has been experimentally extracted from (see Figure 3.1 A). Like during the experimental procedure, the 3' most `CATG` of a transcript sequence is identified and the 10 or 17 nucleotides downstream are taken as the SAGE or LongSAGE tag. It is crucial to extract the SAGE tag from the 3' most `CATG`. Using the sequence downstream of any other `CATG` would lead to the wrong assignment of a completely different tag to the particular gene. It should be noted, that if a gene is alternatively spliced or polyadenylated, different could be extracted from a single gene (see Figure 3.1 B). This assignment of a SAGE tag to a gene will subsequently be called 'mapping'.

3.1.1 Mappings against transcript sequences (UniGene)

Prior to extracting the SAGE tag from any sequence within the UniGene dataset, the sequences were parsed for containing the 3' end of the particular transcript. By an algorithm analyzing poly(A) signals and tails, explained in detail in the section Materials and methods, those sequences that do not reliably contain the 3' end of the sequence were disposed. After extracting the SAGE tags from the remaining sequences of each UniGene cluster, SAGE tags are only taken as reliable if they either occur in a full-length cDNA sequences or in at least 10% of the ESTs within one UniGene cluster. By this compromise, most of the artificial SAGE tags derived due to

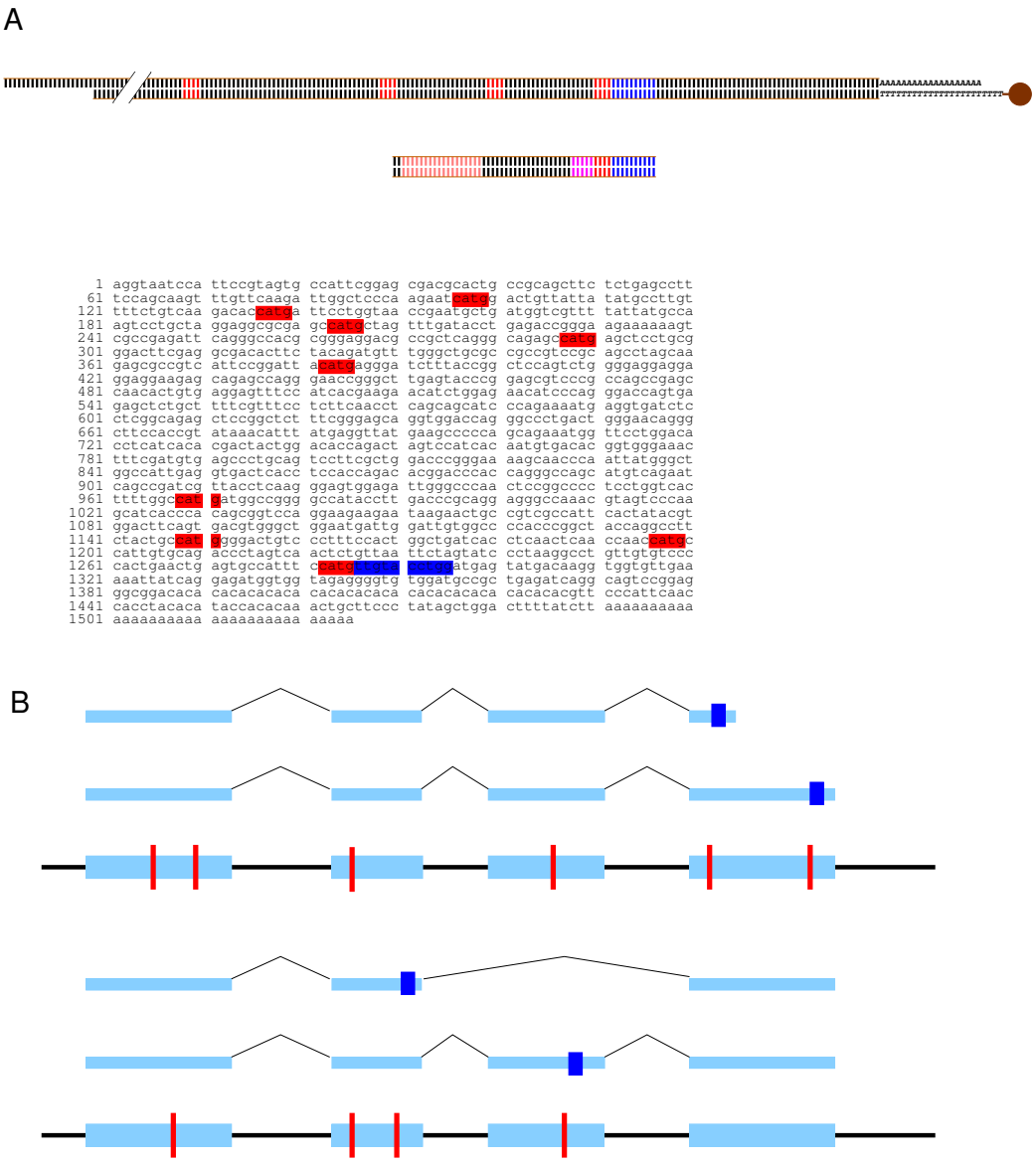


Figure 3.1: (A) Since the Linker is ligated to the 3' most *CATG* and the 10 or 17 bp downstream are excised, concatenated and finally sequenced, the last *CATG* (5'-3' orientation) before the poly(A) tail within a transcript sequence (e.g. from GenBank) is identified and the 10 to 17 bp downstream are extracted. (B) Examples for hypothetical genes with alternative transcript sequences, that lead to alternative SAGE tags due to (top) alternative polyadenylation or (bottom) alternative splicing. Red: *CATG*, blue: SAGE tag.

sequencing errors within the underlying sequence were discarded[1], whereas SAGE tags of alternative polyadenylation or splice isoforms that make up at least ten percent of all sequenced transcripts were recorded. Columns two and six in Table 3.1 summarize in how many cases a reliable SAGE tag is assigned to only one or multiple UniGene clusters. Compared to the theoretical considerations in Table 1.3, the number of SAGE tags (11% in human and 14% in mouse) that can not uniquely be assigned to a single UniGene cluster, is higher. Like expected, this numbers are reduced by the use of LongSAGE (2% in human and 3% in mouse), but still in some cases a tag can not unambiguously be assigned to a single gene. As shown in the graphs of Figure 4.1 (a), the assignment without a pre-selection for 3' sequences provided through SAGEmap [94, 68] are less useful. In this way, more SAGE (38% in human and 18% in mouse) as well as LongSAGE (55%/ 53%) tags are not grasped. Also the number of multiple hits is increased (human: 5% to 18% mouse: 11% to 18%), whereas in this respect no differences can be observed for LongSAGE tags.

3.1.2 Mappings against genomic sequence (EnsEMBL)

As shown in columns three and seven of Table 3.1, only 69% of the human and 57% of the mouse LongSAGE tags extracted from the corresponding UniGene releases were detected exactly once in the genomic sequence. This is considerably less than expected based upon theoretical considerations summarized in Table 1.3. 24 % of the human and 38% of the mouse LongSAGE tags are not found in the whole genome sequence, whereas of the remaining ones (7% in human and 5% in mouse), most hit twice to the genome. If only those hits to the genome are considered, to which a gene has been annotated through EnsEMBL[2], the number of multiple hits decreased to around 1% in both species. On the other hand, for the majority of LongSAGE tags with a single hit to the genome no gene could be reliably assigned to, leaving 71% of the human and 78% of the mouse LongSAGE tags without any mapping in this strategy.

[1] EST sequences are single-pass reads, with error rates up to 1-3% [94, 95], and are often contaminated by genomic sequences as well as by unspliced introns [95]; a sequencing error of 1% suggests that 10% of the SAGE tags extracted from EST sequences contain at least one sequence error: calculated with a binomial probability distribution, the probability of having no sequencing error (0.99 for each base) within 10 bases is $1-(0.99)^{10} = 0.096 \approx 0.1$.

[2] at least one transcript sequence of the gene has to contain the LongSAGE tag, even if it is not derived from the 3' most `CATG`: If the hit is unique to the genome, the LongSAGE tag is most likely from this particular gene, but derived from a not annotated transcript isoform.

Table 3.1: Statistics of different SAGE and LongSAGE mappings

	Human				Mouse			
	UniGene[1]	Genome[2] Raw seq.[5]	Genome[2] W. gene[6]	Combined	UniGene[3]	Genome[4] Raw seq.[5]	Genome[4] W. gene[6]	Combined
SAGE								
0 hit	0%[7]				0%[7]			
1 hit	89%				86%			
2 hits	9%				11%			
3-5 hits	2%				3%			
6-10 hits	<1%				<1%			
> 10 hits	<1%				<1%			
LongSAGE								
0 hit	0%[7]	24%	71%	0%	0%[7]	38%	78%	0%
1 hit	98%	69%	28%	89%	97%	57%	22%	94%
2 hits	2%	4%	1%	5%	2%	3%	1%	4%
3-5 hits	<0.1%	2%	<0.1%	1%	<0.1%	1%	<0.1%	1%
6-10 hits	<0.1%	<0.1%	<0.1%	<0.1%	<0.1%	<0.1%	<0.1%	<0.1%
> 10 hits	<0.1%	<0.1%	<0.1%	<0.1%	<0.1%	<0.1%	<0.1%	<0.1%

SAGE and LongSAGE tags analyzed are those extracted from the particular UniGene releases that are considered to be 'reliable' (see Materials and Methods), [7]therefore all tags must hit at least to one UniGene cluster. UniGene builds [1]161 and [3]123, Assemblies [2]NCBI33 and [4]NCBI30. [5]Raw sequence: Any hit against genome. [6]Only hits against genome with gene annotated to it.

3.1.3 Combination of UniGene and EnsEMBL mappings

In order to get the best possible tag-to-gene assignment, both previously described mappings were combined, as illustrated in Figure 3.2. With the two approaches described above, for each LongSAGE tag the corresponding UniGene cluster (hereafter called 'UniGene hit'[3]) and the accordant hit to the genome (together with the gene(s) annotated to it; 'Genome hit'[4]) were identified. To merge both mappings, EnsEMBL genes[5] linked to both UniGene hit and Genome hit will be used as the bench mark to determine whether both correspond to the same gene. Only Genome hits with a gene associated were used. For the identification of the EnsEMBL gene corresponding to a UniGene entry, all UniGene clusters were queried for sequence homology (using `BLAST`) against all EnsEMBL transcript sequences[6]. The EnsEMBL gene with the highest sequence similarity[7] to the UniGene cluster is taken as being identical. Since the EnsEMBL genome database contains redundant entries[8], all genes with an overlapping chromosomal localization to the gene identified as being identical to the UniGene hit are also considered, if its sequence similarity to the UniGene cluster is above the threshold and it also contains the LongSAGE tag. Next all Genome hits and UniGene hits for a particular LongSAGE tag were compared, and a Genome hit and a UniGene hit were taken as identical, if at least one of the associated EnsEMBL genes overlapped.

Not all LongSAGE tags were found only once within the genome and/ or derived from only one UniGene cluster. Figure 3.3 gives an overview, how

[3] in case of multiple mapping UniGene clusters, each UniGene cluster is considered as an independent UniGene hit.

[4] in case of multiple hits to the genome, each hit together with the gene(s) annotated to it is treated as a independent Genome hit.

[5] this mapping does not discriminate between EnsEMBL genes and EnsEMBL EST genes; therefore, for simplicity, the term 'gene' used in this paraph will account for both.

[6] one gene could have multiple transcripts, due to alternative splicing of polyadenylation.

[7] above a certain threshold: The threshold is critical, since a too low cutoff could result in a miss-assignment to a wrong gene with high sequence homology. But due to sequencing errors or alternative transcript forms, the alignment between UniGene sequence and EnsEMBL transcript sequence (is derived from the genomic sequence) for the same gene will not always be 100%. Since only the best hit for a UniGene cluster is used, as long as the corresponding gene is annotated to EnsEMBL, the UniGene cluster will always be assigned to its corresponding gene. Please note, that one EnsEMBL gene could be associated to multiple UniGene clusters, but every UniGene cluster is only assigned to a single EnsEMBL gene.

[8] EnsEMBL genes (genes supported by protein sequence of same or other species) and EST genes (supported by cDNA or EST transcript sequences) often overlap.

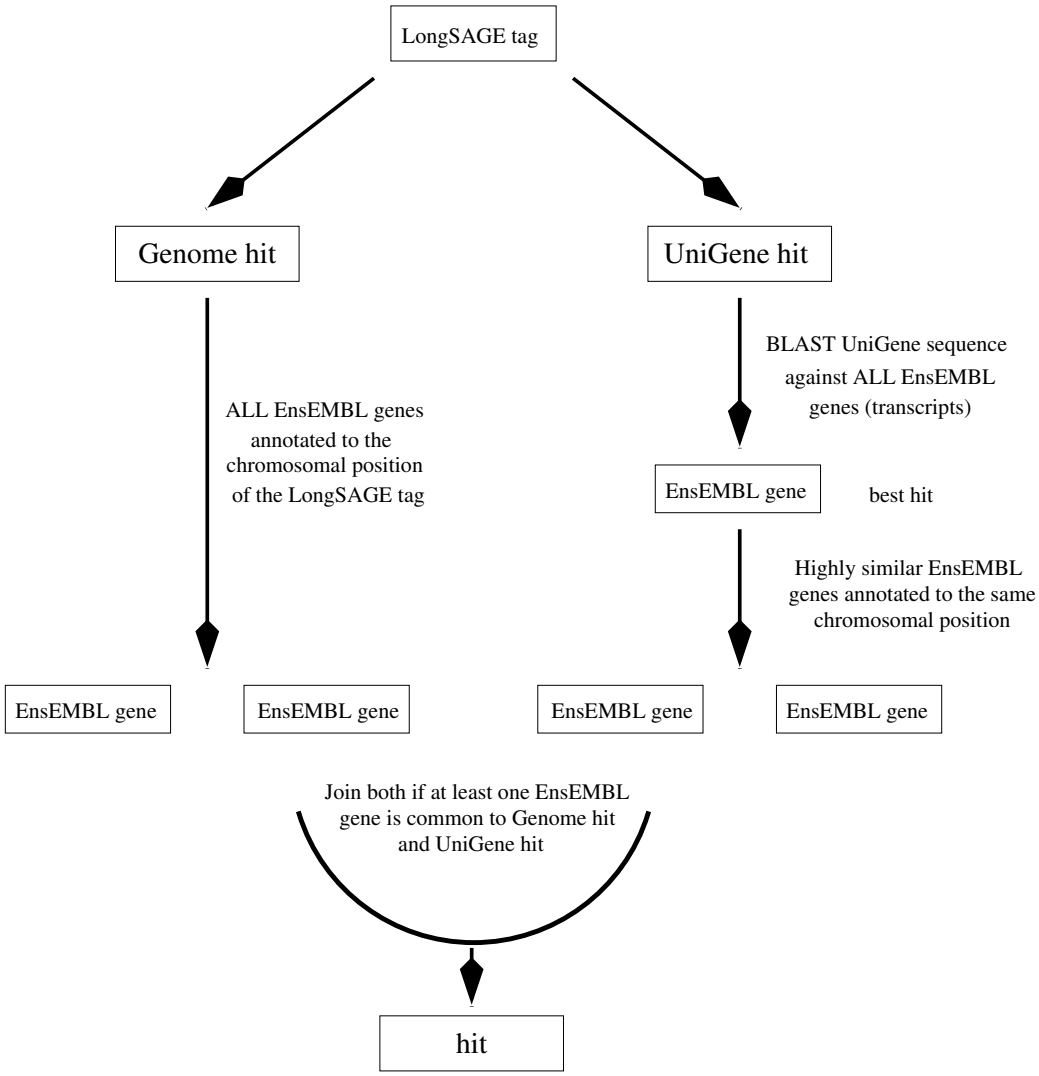


Figure 3.2: Strategy used to merge Genome hits and UniGene hits. First EnsEMBL genes associated to Genome and UniGene hits are identified. If at least one gene is common to both, hits are considered to be identical.

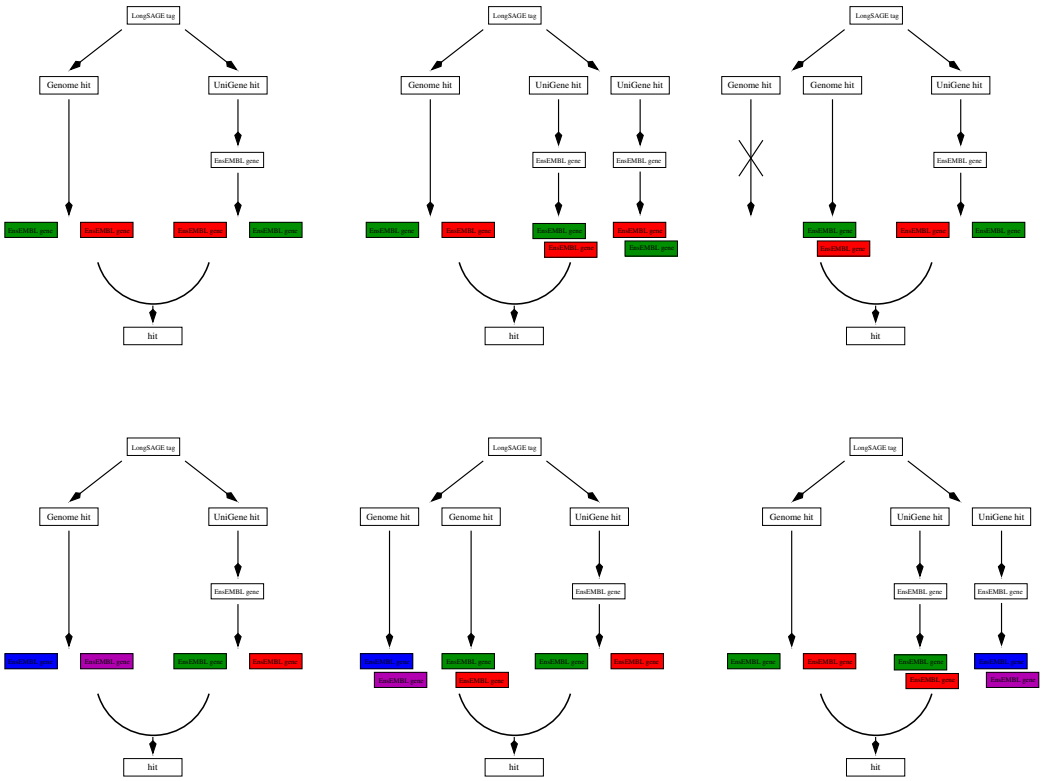


Figure 3.3: Examples showing hypothetical cases in which one LongSAGE tag gives rise to a single (top) or multiple (bottom) hits. Identical EnsEMBL genes (genes as well as EST genes; for simplicity, only EnsEMBL gene is written into the boxes) are drawn with the same color. If one LongSAGE tag is found in the genome two times or mapped to UniGene two times, two Genome hit or two UniGene hit boxes are drawn. See text for details.

Table 3.2: Genome annotation by verification of EST alignments with LongSAGE tags

	all	unique hits only
total		13127
antisense gene	805	547
new intron for existing gene	2	0
extension of 3' UTR of existing gene	371	260
novel	2595	1827

Result of genome annotation based upon combining a LongSAGE hit and ESTs aligned to the chromosomal position. All LongSAGE tags with a count of two or more in all somite LongSAGE libraries were analyzed.

multiple Genome hits and/or multiple UniGene hits were dealt with. Like LongSAGE tags with single Genome and UniGene hits, that share at least one EnsEMBL gene or EST gene (top left), LongSAGE tags with a single Genome hit and multiple UniGene hits having at least one EnsEMBL gene or EST gene in common (top middle), were treated as a single hit, since the different UniGene clusters should correspond to the same gene. Genome hits with no gene annotated to the position of the LongSAGE tags were not considered (top right). On the other hand, if none of the EnsEMBL genes linked to the Genome hit was identical to the UniGene hit derived from the LongSAGE tag (bottom left), both were taken as being different. Whenever two or more different Genome hits with an associated EnsEMBL gene were found for one LongSAGE tags (bottom middle), or if two UniGene hits were linked to different EnsEMBL genes (bottom right), all hits were treated as being independent. Not all single-hit LongSAGE tags were mapped to both to EnsEMBL and UniGene. UniGene clusters without a significant `BLAST` hit against at least one EnsEMBL gene were taken as being different from any other Genome hit. The numbers in Table 3.3 almost do not change (slightly worsen due to more multiple hits) by this strategy (columns five and nine) compared to the UniGene mappings, since only LongSAGE tags with a reliable mapping to UniGene were used. When analyzing the LongSAGE tags generated by the LongSAGE analysis of somitogenesis (Table 3.15), the number of no-hits could be decreased in all abundance classes.

3.1.4 Gene annotations based upon aligned ESTs and GeneScan predictions

Due to the fact that many unique hits of LongSAGE tags to the genome are not associated to a EnsEMBL gene, a different approach was used. As it will be discussed below, the homology-based assignment (e.g. by `BLAST`) of ESTs

Table 3.3: Genome annotation by verification of GeneScan predictions by LongSAGE tags

total	13127
UniGene hit, no Genome hit	651
no UniGene hit, no Genome hit	2348

GeneScan predictions verified by LongSAGE tags. GeneScan predictions are only listed if only a single GeneScan prediction is associated to a particular LongSAGE tag.

to a chromosomal position, and thereby to its genomic locus, is the most difficult task of genome annotation. Here, the uniqueness of the LongSAGE tag within the whole genome was taken to definitely associate ESTs to the particular chromosomal position. All ESTs containing the LongSAGE tag that could be aligned to this position on the chromosome[9] were considered as representing the gene. As shown in Table 3.2, in this way 1827 genes, that are not represented in EnsEMBL, could be annotated to the genome. Furthermore, by extending the 3' UTR, which originally was not annotated to EnsEMBL, 360 additional LongSAGE tags could be assigned to a EnsEMBL gene. In addition, 547 cases were identified, in which the LongSAGE tag proved the existence of a yet unknown antisense gene.

Furthermore, for the chromosomal positions to which a LongSAGE tag mapped, genes were predicted by GeneScan [96]. If a predicted cDNA overlapped with the LongSAGE tag, the predicted gene was considered to be a real gene (Table 3.3). For 651 cases, the newly annotated gene was represented within the UniGene dataset. In addition, 2348 novel genes not represented in the UniGene database could be annotated to the genome.

3.1.5 Assignment of SAGE tags to genomic sequence

SAGE tags are too short to be unambiguously assigned to genomic sequence (see Table 3.1). Thus an indirect approach was used and the UniGene cluster(s) mapped to every SAGE tag were located to its corresponding genomic sequence. In principle, such a link between UniGene and EnsEMBL genes is available through LocusLink [97, 68], but this is only the case for a fraction of all genes annotated through EnsEMBL (see Table 3.4). Therefore, a own algorithm was developed. Due to the exon-intron structure of eukaryotic genes, transcript sequences can not be directly queried against the genome by homology search programs (see Discussion). Thence the assignment of UniGene transcript sequences is solely generated by sequence similarity to

[9] $\geq$ 92% percent identity

Table 3.4: UniGene clusters assigned to EnsEMBL genes

	number of human UniGene clusters linked	number of mouse UniGene clusters linked
LocusLink - total	15,602	12,812
UniGene clusters mapped in this study - total	23,825	21,231
EnsEMBL genes	18,040	18,451
EnsEMBL genes with more than one UniGene cluster	3415	2267
EnsEMBL EST genes	2805	4576
EnsEMBL EST genes with more than one UniGene cluster	356	189
EnsEMBL GeneScan Predictions	798	386
EnsEMBL GeneScan Predictions with more than one UniGene cluster	30	9

Number of EnsEMBL genes, EST genes and GeneScan Predictions, to which at least one UniGene cluster was assigned. UniGene clusters were only assigned to EST genes and then to GeneScan Predictions, if they did not match to the former.

genes or predictions annotated to the genome[10]. As shown in Table 3.4, in this way more UniGene clusters could be mapped to EnsEMBL genes, EST genes and GeneScan predictions[11] than through LocusLink. Since in some cases multiple UniGene clusters could be assigned to the same gene (alternative transcripts), a total of 29,434 human and 24,464 mouse UniGene clusters were assigned to 23,853 and 21,231 chromosomal positions.

3.2 SAGE of chondrogenesis

3.2.1 Induction of chondrogenic differentiation of ATDC5 cells by BMP4 treatment

To determine a proper amount of BMP4 for inducing chondrogenic response in ATDC5 cells, ATDC5 cells were treated with different amounts of human recombinant BMP4 for 36 hours. The expression levels of *type II collagen (Col2a1)* were monitored as an early marker for the onset of chondrogenesis by Northern blot analysis (Figure 3.4 A). A range of BMP4 concentration between 100 and 200 ng/ml was sufficient to induce early chondrogenesis in ATDC5 cells with a detectable increase in *Col2a1* expression. After 76 hours, the cells induced with 200 ng/ml of BMP4 showed the typical round

[10]EnsEMBL genes as well as EnsEMBL EST genes and GeneScan predictions, for details see Materials and methods.

[11]Ranking gene - EST gene - GeneScan prediction.

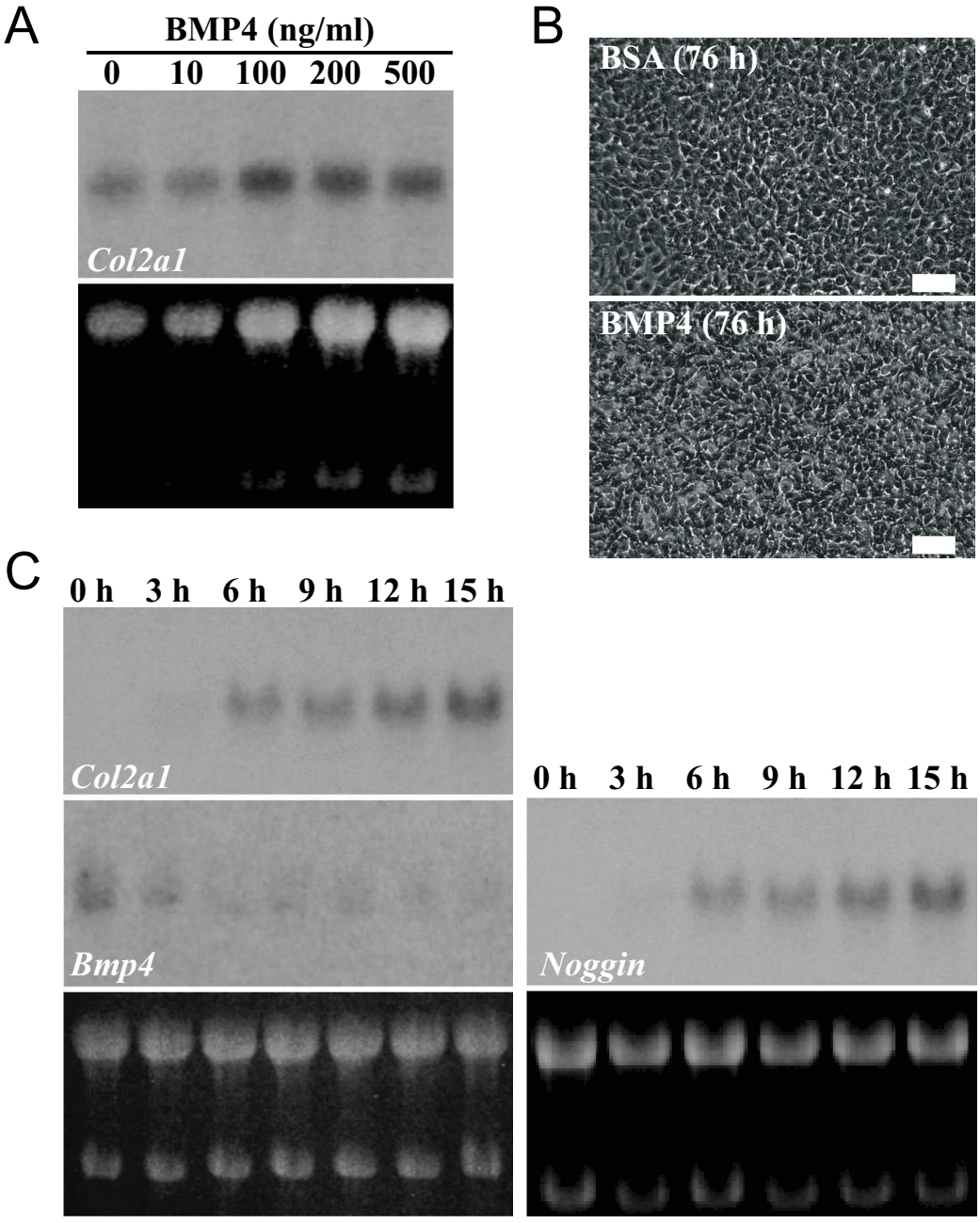


Figure 3.4: Pilot study to determine optimal conditions for BMP4 induction in ATDC5 cells. (A) Northern blot analysis of *Col2a1* expression in ATDC5 cells treated with different amounts of BMP4 for 36 hours. *Col2a1* is upregulated in the presence of exogenous BMP4 at 100 - 500 ng/ml. (B) Cellular morphology of BMP4 treated cells. Phase contrast microscopy shows that after 76 hours BMP4 induced a typical morphology of round chondrocytes all over the culture. The scale bars: 100 μm. (C) Northern blot analysis of expression of *Col2a1*, *Bmp4* and *noggin* in ATDC5 cells treated with 200 ng/ml of BMP4 for 3 - 15 hours. 0 h means RNA extracted from ATDC5 cells at confluence without BMP4 treatment. Note that upregulation of *Col2a1* and *noggin* becomes detectable in 6 hours, while downregulation of endogenous *Bmp4* is detected already in 3 hours. Ethidium-bromide stained RNA gels showing 28S and 18S rRNA are provided to control amounts of loaded RNA samples.

morphology of chondrocytes all over the culture (Figure. 3.4 B). Since the aim of this study was to identify genes involved in the early phase of chondrogenesis, a suitable time point had to be explored, that could be compared with undifferentiated ATDC5 cells. Thus, expression levels of *Col2a1* were measured over a time course of 15 hours after the BMP treatment with steps of three hours. It is known that the administration of exogenous BMP4 rapidly induces downregulation of endogenous Bmp4 and upregulation of *noggin* [31, 59]. As shown in Figure 3.4 C, endogenous *Bmp4* was already downregulated 3 hours after treatment. However, the first changes in *Noggin* and *Col2a1* expression levels became clearly detectable by Northern blot analysis after 6 hours. Therefore, gene expression profiles between uninduced ATDC5 cells at confluence (hereafter referred to as 'undifferentiated') and those induced by BMP4 at 200 ng/ml for 6 hours (referred to as 'induced') were compared in this SAGE study.

3.2.2 General Overview of SAGE libraries

As summarized in Table 3.5, a total of 43,656 tags including 21,875 tags (excluding linker tags) from the SAGE library made from undifferentiated ATDC5 cells were sequenced, and 21,781 tags from the induced ATDC5 library. The whole SAGE tags collected consisted of 17,166 kinds of SAGE tags, of which 7,064 tags (including 815 tags with a count of two or higher) were observed only in undifferentiated and 6,884 tags (including 822 with a count of two or higher) only in differentiated ATDC5 cells. In the case of tags found in common between the two libraries (3,218 kinds of tags), most tags were represented in both libraries at similar levels (see Figure 3.5), indicating the reproducibility of our SAGE analysis. Furthermore, the low frequency of linker contamination (average 1%) and the low incidence of identical ditags (average 1.6%) proved the high quality of both SAGE libraries. A complete list of all SAGE tags from this study is available online at the Gene Expression Omnibus (www.ncbi.nlm.nih.gov/geo) with the accession numbers GSM2575 and GSM2576.

3.2.3 Tag to gene assignment

In total, 8,138 kinds of tags (47%) could be reliably assigned to UniGene clusters, including matches to 4,800 named (known) genes (27%) and 3,338 ESTs (19%) (Table 3.6). The remaining tags (53%: 9,028 tags) could not be assigned to any gene (hereafter designated as 'no-hit' tags). As it has been reported [22], the frequency of no-hit tags was inversely proportional to their

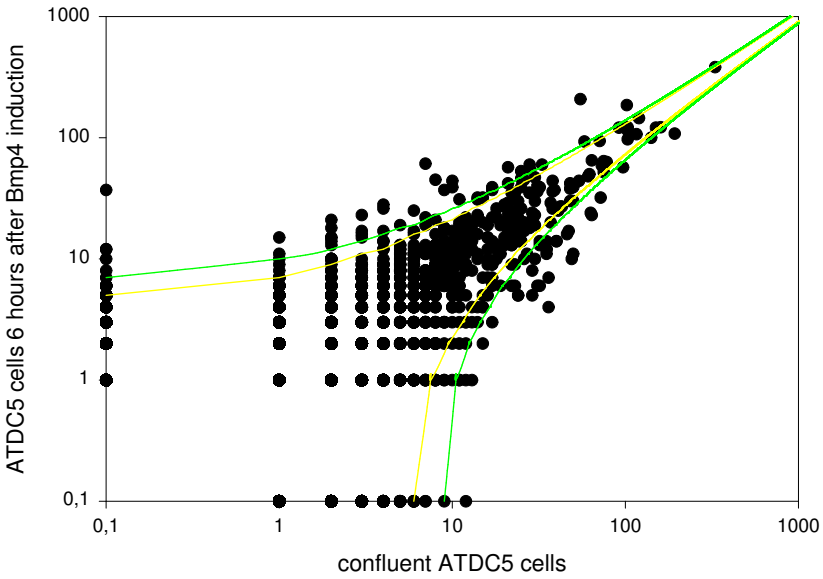


Figure 3.5: Scatter plot of tag abundance distributions in undifferentiated and BMP4-induced ATDC5 cells. Note the log scale. The numbers represent tag counts in the undifferentiated (X-axis) and BMP4-induced (Y-axis) ATDC5 libraries. Each dot simply means the presence of tags in a given X-Y coordinate, but does not provide information on how many different tags represent a single dot. Yellow and green lines indicate the confidence limits of 95% and 99%, respectively.

Table 3.5: Summary of ATDC5 SAGE libraries

	Sum[1]	Udf[2]	Def[3]
Statistics			
Total tags (excluding linker tags)	43656	21875	21781
Unique transcripts	17166	10282	10102
Transcripts observed in both libraries	3218		
Transcripts observed only in one of the two libraries		7064	6884
with a count of > 1		815	822
Quality			
Linker contamination	430 (1%)	99 (0.5%)	331 (1.5%)
Frequency of identical ditags	677 (1.6%)	446 (2.0%)	231 (1.1%)

[1]Sum, tag counts from the two ATDC5 libraries combined. [2]Udf, tag counts from the undifferentiated ATDC5 library. [3]Def, tag counts from the BMP4-induced ATDC5 library.

Table 3.6: Statistics of SAGE tag assignment sorted by abundance classes

Tag counts in the data set	Abundance in tpm[1]	Different kinds of tags	Total number of tags	Hits to named genes[2]	Hits to ESTs	No hit
> 100	$> 2{,}291$	25	5,383	19 (76%)	3 (12%)	3 (12%)
31 - 100	710 - 2,291	124	6,408	102 (82%)	9 (7%)	13 (10%)
7 - 30	160 - 687	651	7,918	497 (76%)	83 (15%)	71 (11%)
2 - 6	46 - 137	4,055	11,636	1,945 (48%)	910 (22%)	1,200 (30%)
1	23	12,311	12,311	2,237 (18%)	2,333 (19%)	7,741 (63%)
Total		17,166	43,656	4,800 (28%)	3,338 (19%)	9,028 (53%)

[1]tpm: tags-per-million. [2]Tags with multiple hits to named genes and ESTs are only counted for the former.

abundance: 63% of single-count tags, 30% of tags with a count between two and six, and about 11% of tags with a count of seven or higher.

3.2.4 Genes abundantly expressed in ATDC5 cells

The 30 most abundant tags (with tag numbers preceded by A, i.e., abundant) in the two ATDC5 libraries combined are listed in Table 3.7. When compared to other SAGE data from twelve studies in the mouse (see Table 3.9), all tags but two (tag A15 and A21) were also abundantly expressed with levels higher than 1,000 tags-per-million (tpm) (i.e., $\leq 1\%$) in one or more of these libraries. One of the exceptional tags, A15 (3,781 tpm in ATDC5 cells), corresponded to *Col1a2*, and the representation of this tag in other libraries was low (0 - 347 tpm). The other exceptional tag A21 (2,955 tpm in ATDC5 cells) did not show any match in the databases, and its representation levels in the other libraries remained modest (15 - 690 tpm). The most abundant tag in the ATDC5 library, A1 (16,312 tpm in ATDC5 cells), showed no match,

but was also highly expressed (1,397 - 8,740 tpm) in four other libraries including ones from granular cell precursors, medulloblastoma, limb buds and intraepithelial lymphocytes. This result suggests the presence of a potentially novel gene that is abundantly expressed in many different tissues. Tags A2 (multiple hits), A14 (ESTs) and A26 (*Cfl1*) were highly expressed also in all other seven libraries compared. The high expression of tags A10 (*S100a4*), A27 (*Fn1*) and A30 (*Eno1*) was only seen in one of the seven libraries: 3T3 fibroblasts (4,161 tpm), R1 ES cells (1,305 tpm) and limb buds (1,067 tpm), respectively.

Table 3.7: List of thirty most abundant tags in the ATDC5 libraries

#	Sequence	Sum[1]	Udf[2]	Def[3]	Symbol	UniGene	Description[4]
A1	ATAATACATA	712	330	382			no hit
A2	GTGGCTCACA	301	193	108			multiple hits (710)
A3	CAAACTCTCA	288	102	186	*Sparc*	35439	secreted acidic cysteine rich glycoprotein
A4	TGACCCCGGG	282	160	122			multiple hits (2)
A5	GTGAAACTAA	271	150	121	*Rps4x*	66	ribosomal protein S4, X-linked
A6	AAAAAAAAAA	265	120	145			multiple hits (195)
A7	GAATAATAAA	263	55	208	*Hspa8*	197551	heat shock 70kD protein 8
A8	GCGGCGGATG	241	141	100	*Lgals1*	43831	lectin, galactose binding, soluble 1
A9	GGCTTTGGTC	225	103	122	*Rplp1*	3158	ribosomal protein, large, P1
A10	TGCACAGTGC	223	116	107	*S100a4*	3925	S100 calcium binding protein A4
A11	TTTTATGTTT	220	101	119	EST	213020	Highly similar to RL32_HUMAN 60S ribosomal protein L32
A12	ATGTCTCAAA	216	104	112			multiple hits(3)
A13	TGGATCAGTC	213	92	121			multiple hits (2)
A14	GCTGCCCTCC	200	103	97	EST	23906	Weakly similar to retinitis pigmentosa GTPase regulator interacting protein 1; 0610005A07Rik
A15	GTTCCAAAGA	165	71	94	*Col1a2*	4482	procollagen, type I, alpha 2
A16	TGGGTTGTCT	154	97	57	*Tpt1*	254	tumor protein, translationally-controlled 1
A17	ATACTGACAT	151	58	93			no hit
A18	TAAAGAGGCC	141	78	63	*Rps26*	372	ribosomal protein S26
A19	CTAGTCTTTG	139	75	64	*Rps29*	154915	ribosomal protein S29
A20	CTAATAAAGC	132	73	59	*Fau*	4890	Finkel-Biskis-Reilly murine sarcoma virus (FBR-MuSV) ubiquitously expressed (fox derived)
A21	TAGATATAGG	129	64	65			no hit
A22	TGGTGACAAA	127	75	52			multiple hits (2)
A23	GGCAAGCCCC	112	62	50	*Rpl10a*	2424	ribosomal protein L10A
A24	GGATTTGGCT	109	61	48	*Rplp2*	14245	ribosomal protein, large P2
A25	AGGCAGACAG	104	72	32	*Eef1*a1	196614	eukaryotic translation elongation factor 1 alpha 1

table continues on following page

#	Sequence	Sum[1]	Udf[2]	Def[3]	Symbol	UniGene	Description[4]
A26	GAAGCAGGAC	98	54	44	*Cfl1*	4024	cofilin 1, non-muscle
A27	CCAACGCTTT	93	33	60	*Fn1*	193099	fibronectin 1
A28	TGTAGTGTAA	90	48	42	*Rps8*	3381	ribosomal protein S8
A29	TCAGGCTGCC	89	50	39	*Fth*	1776	ferritin heavy chain
A30	CAAAAATAAA	88	28	60	*Eno1*	90587	enolase 1, alpha non-neuron

Count values in tags-per-million. For [1]Sum, [2]Udf and [3]Def, see the corresponding footnotes in Table 3.5. [3]Description: in case of multiple hits, the number of matching genes is given in parenthesis

3.2.5 Differences between undifferentiated and BMP4-induced ATDC5 cells

3.2.5.1 Statistical analysis

By comparing hit counts of individual tags in the two ATDC5 SAGE libraries, a total of 139 tags were predicted to be differentially represented at the confidence limit 95% or higher ($P \leq 0.05$), including 74 tags at the confidence limit 99% or higher ($P \leq 0.01$). Of these differentially represented tags, 72 ($P \leq 0.05$) or 35 ($P \leq 0.01$) tags were downregulated, and 67 ($P \leq 0.05$) or 39 ($P \leq 0.01$) tags were upregulated in ATDC5 cells upon the BMP4 treatment. These tags are listed in Table 3.8 with tag numbers preceded by either D (i.e., downregulated) or U (i.e., upregulated), and they are illustrated as outliers in the scatter plot in Figure (3.5). Ninety-four of these 139 tags were assigned to 77 known (named) genes and 17 ESTs (including nine RIKEN full-length cDNA genes of undefined functions), while 27 tags matched to more than one gene (multiple hits), and 17 tags did not show any match (potentially novel genes).

Table 3.8: List of differentially expressed tags

#	Sequence	Udf[1]	Def[2]	Symbol[3]	UniGene	Description[4]
Downregulated (99 % significance)						
D1	AAGAGGCAAG	30	10			multiple hits (2)
D2	AAGGAAGAGA	29	5			multiple hits (2), including *Vim*[6,7]
D3	AAGGTGGAAG	24	5			multiple hits (2)
D4	AGAGCGAAGT	47	17	*Rpl41*	13859	ribosomal protein L41
D5	AGGCAGACAG	72	32	*Eef1a1*	196614	eukaryotic translation elongation factor 1 alpha 1
D6	AGGTCGGGTG	32	13	*Rpl13a*	13020	ribosomal protein L13a
D7	ATACTGAAGC	36	7	*Rpl13*	42578	ribosomal protein L13
D8	ATTGCTTAGA	35	11	*Rbm3*	2591	RNA binding motif protein 3

table continues on following page

#	Sequence	Udf[1]	Def[2]	Symbol[3]	UniGene	Description[4]
D9	CAAGGTGACA	64	23	*Rps2*	1129	ribosomal protein S2
D10	CAGAACCCAC	35	12	*Rps18*	42790	ribosomal protein S18
D11	CAGTCTCTCA	24	5	EST	39130	RIKEN cDNA 2210402A09 gene
D12	CCCTGGGTTC	17	3	*Ftl1*	7500	ferritin light chain 1
D13	CCGAAAGTAA	13	1	*Sdc2*	29350	syndecan 2
D14	CCTACCAAGA	29	8	*Rps20*	21938	ribosomal protein S20
D15	CCTGATCTTT	35	12			multiple hits (2)
D16	CTGAACATCT	63	24	*Arbp*	5286	acidic ribosomal phosphoprotein PO
D17	CTGTAGGTGA	49	10	*Rps23*	30011	ribosomal protein S23
D18	GACGCTGCCA	11	1	*Rpl22*	13917	ribosomal protein L22
D19	GAGACTAGCA	11	1	*Tm4sf8*	28484	transmembrane 4 superfamily member 8
D20	GCAACCTCCC	40	14			multiple hits (3), including Igfbp5[6,7]
D21	GGGGAAATCG	23	6	*Ptmb10*	3532	thymosin, beta 10
D22	GTGGCTCACA	193	108			multiple hits (562)
D23	GTGTTAACCA	22	6	EST	2050	RIKEN cDNA 2510008H07 gene
D24	TCCTTCCGAC	12	0	EST		no hit - Mus musculus similar to G protein pathway suppressor 1 $XM_1 26542$[5]
D25	TCTACAAGAA	40	14			multiple hits (2)
D26	TCTGACTTCC	35	14			multiple hits (2), including Bgn[6]
D27	TCTGGACGCG	15	2	*H2afx*	14767	H2A histone family, member X
D28	TCTTCTATGC	36	4	*Col1a2*	4482	procollagen, type I, alpha 2
D29	TCTTCTCACA	32	6	EST	30478	RIKEN cDNA 2810465O16 gene
D30	TGCTCTCCCT	12	1	EST	30016	expressed sequence C87222
D31	TGGCTCGGTC	46	16	*Actg*	196173	actin, gamma, cytoplasmic[5]
D32	TGGGTTGTCT	97	57	*Tpt1*	254	tumor protein, translationally-controlled 1
D33	TTCATTATAA	50	11	*Ptma*	19187	prothymosin alpha
D34	TTCTCCTCAG	9	0			no hit
D35	TTGGCTGCCC	31	6	*Rps14*	43778	ribosomal protein S14
Downregulated (95 % significance)						
D36	AAACCCCCAG	6	0			no hit- mitochondrial DNA[5]
D37	AAGAAAATAG	27	11	EST	22723	Mus musculus, clone IMAGE:3586350, mRNA, partial cds
D38	ACAGTGCTTG	8	1	*Ppp2cb*	7418	protein phosphatase 2a, catalytic subunit, beta isoform
D39	ACTGCTTTTC	10	2	*Ndufa7*	29513	NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 7 (14.5kD, B14.5a)
D40	AGAAGGAGGT	12	3	*Lag*	28479	leukemia-associated gene
D41	AGATCTATAC	14	4	*Rpl7*	37835	ribosomal protein L7
D42	AGGAATCCAC	26	10	*Gas1*	22701	growth arrest specific 1[7]
D43	AGGAGGACTT	14	3			tag reliably matches mitochondrial DNA
D44	ATCAGTGGCT	8	1	*Psmb4*	368	proteasome (prosome, macropain) subunit, beta type 4
D45	ATGGCTCACA	6	0			multiple hits (19)
D46	CACCACCACA	15	4			multiple hits (2)
D47	CAGAGATCCC	6	0	*Gnai2*	196464	guanine nucleotide binding protein, alpha inhibiting 2

table continues on following page

#	Sequence	Udf[1]	Def[2]	Symbol[3]	UniGene	Description[4]
D48	CCCAATGGCC	17	6	EST	219039	ESTs, Highly similar to S32604 collagen alpha 2(VI) chain
D49	CCCTCTGGAT	10	1	*S100a6*	100144	S100 calcium binding protein A6 (calcyclin)
D50	CGGCGCGGAG	12	3	*Emp3*	20829	epithelial membrane protein 3
D51	CGGGTCATAT	32	14	*Psmb1*	42197	proteasome (prosome, macropain) subunit, beta type 1
D52	CTGCTATCCG	23	9	*Rpl5*	4419	ribosomal protein L5
D53	CTGGTGGGCA	7	0	*Rab11b*	35727	RAB11B, member RAS oncogene family
D54	CTGTAAAAAA	16	5	*Cxcl12*	465	chemokine (C-X-C motif) ligand 12
D55	GAGGATTCCC	8	1			multiple hits (2)
D56	GCGGATTCTG	9	1	*Plp2*	18565	proteolipid protein 2
D57	GGCGTCATCG	9	1	*Psmb3*	21874	proteasome (prosome, macropain) subunit, beta type 3
D58	GGGAAATCGC	7	0			no hit
D59	GTTGCTGAGA	50	30	*Rpl10*	100113	ribosomal protein 10
D60	TAAAAAAAAA	37	20			multiple hits (68)
D61	TAATAAAAAT	15	5	*Hmgn1*	2756	high mobility group nucleosomal binding domain 1
D62	TCACATAAAT	6	0	*Itgp*	167842	integrin-associated protein[7]
D63	TCATCTTCAG	8	1			no hit
D64	TCATCTTTAA	22	9	*Calr*	1971	Calreticulin
D65	TCCCTTCGAC	6	0			no hit
D66	TCCTTGGGGG	8	1	*Hint*	425	histidine triad nucleotide binding protein
D67	TCGCAAGCAA	13	3	*Naca*	3746	nascent polypeptide-associated complex alpha polypeptide
D68	TCTCTCAGTC	19	6	*Anxa5*	1620	annexin A5
D69	TCTGTGCACC	11	2	*Rps11*	196538	ribosomal protein S11
D70	TCTTCTTTGG	13	4		28044	filamin-like protein[8]
D71	TCTTTGGAAC	12	2	*Mor1*	21743	malate dehydrogenase, mitochondrial
D72	TGCTGTGAAA	15	4			mulitple hits (2), including Fxc1[7]
Upregulated (99 % significance)						
U73	AAAAATCATC	10	44			no hit- mitochondrial DNA[5]
U74	AAAAGAAATA	6	25			mulitple hits (3), including EST (AK009226)e
U75	AAAATAAAAC	0	8			multiple hits (2)
U76	AAATAAAACA	2	18			multiple hits (2)
U77	AAATCCTTTC	3	23	*Ptn*	3063	Pleiotrophin
U78	AACATTAAAA	3	17			multiple hits (2), including F2rl1[6]
U79	AACATTCAAA	25	53	EST	115442	ESTs, Weakly similar to S16783 probable RNA-directed DNA polymerase [R.norvegicus]
U80	AACTTTTGTT	3	15	*Serpinh1*	22708	serine (or cysteine) proteinase inhibitor, clade H (heat shock protein 47), member 1
U81	AATATGTGTG	7	22	*Cox6c*	548	cytochrome c oxidase, subunit Vic
U82	AATTCAATTA	0	12	EST	219670	Mus musculus, Similar to eukaryotic translation initiation factor 4 gamma, 1

table continues on following page

#	Sequence	Udf[1]	Def[2]	Symbol[3]	UniGene	Description[4]
U83	AATTTCAAAA	11	31	*Rps17*	3428	ribosomal protein S17
U84	ACCTATATTG	1	15	EST	181880	RIKEN cDNA 1110007A14 gene[7]
U85	ATAATACGAA	0	12			no hit - NOVEL[5]
U86	ATACTGACAT	58	93			no hit- mitochondrial DNA[5]
U87	ATTAATCAGT	3	16	EST	46754	expressed sequence AI316867[7]
U88	ATTTGATTAG	4	28	EST	182471	RIKEN cDNA 2610524G07 gene[7]
U89	CAAAAATAAA	28	60	*Eno1*	90587	enolase 1, alpha non-neuron
U90	CAAACTCTCA	102	186	*Sparc*	35439	secreted acidic cysteine rich glycoprotein[7]
U91	CAATAAACTG	10	39	S*ui1-rs1*	13886	suppressor of initiator codon mutations, related sequence 1 (S. cerevisiae)
U92	CAATGTGGGT	2	21	*Osf2*	10681	osteoblast specific factor 2 (fasciclin I-like)[7]
U93	CCAAATAAAA	28	56	*Ldh1*	141443	lactate dehydrogenase 1, A chain
U94	CCAACGCTTT	33	60	*Fn1*	193099	fibronectin 1[7]
U95	CCAATACGAA	0	37			no hit - NOVEL[5]
U96	GAAATATATG	4	26	*Idh2*	2966	isocitrate dehydrogenase 2 (NADP+), mitochondrial
U97	GAAATGTAAG	5	19			multiple hits (2)[7], including *Pcbp2*[6]
U98	GAATAATAAA	55	208			multiple hits (2)[5], including *Hspa8*[6,7]
U99	GAATTAACAT	2	15	*Ywhae*	42972	tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, epsilon polypeptide[7]
U100	GCCCGGGAAT	17	39			multiple hits (2)[5], including Hk1[6]
U101	GCCTAATGTA	21	57	EST	209503	ESTs
U102	GGTTTCTTTT	8	45			no hit
U103	GTGCATTTGT	7	61	*Tgfbi*	14455	transforming growth factor, beta induced, 68 kDa[7]
U104	GTTTTTTAAA	4	17			multiple hits (2)
U105	TCCCCCAATG	0	10			no hit
U106	TGATGTATAT	9	37	*Idb3*	110	inhibitor of DNA binding 3[7]
U107	TGCTGTGCAT	2	14	*Fin14*	18459	fibroblast growth factor inducible 14[7]
U108	TGTTCATCTT	2	13	*Col3a1*	147387	procollagen, type III, alpha 1
U109	TTAATAAAAG	17	37	EST	391	RIKEN cDNA 9130413I22 gene
U110	TTGTAAAAGG	1	11	*Vcp*	18921	valosin containing protein
U111	TTTTTGGTGT	2	15	*Atp6g1*	29868	ATPase, H+ transporting, lysosomal (vacuolar proton pump)g
Upregulated (95 % significance)						
U112	AAAAATAAAA	4	14			multiple hits (6)
U113	AAAATGTTGT	2	10			multiple hits (3)
U114	AACATTCGCA	25	46			no hit
U115	AACTTTTAAA	0	6			multiple hits (2)
U116	AAGTAAAGCG	6	17	*Sec61g*	1164	SEC61, gamma subunit (S. cerevisiae)
U117	AATGATAAAA	2	10	EST	29363	RIKEN cDNA 2310044F10 gene[7]
U118	AATGTGAGTC	1	9	*Ywhag*	29717	3-monooxgenase/tryptophan 5-monooxgenase activation protein, gamma polypeptide

table continues on following page

#	Sequence	Udf[1]	Def[2]	Symbol[3]	UniGene	Description[4]
U119	ACAAATAAAC	15	32			multiple hits (2)
U120	ATGAGAACAG	0	6			no hit- mitochondrial DNA[5]
U121	ATGTGAATAA	1	10	*Usp5*	3571	ubiquitin specific protease 5 (isopeptidase T)
U122	ATTTGACTGG	1	8			no hit
U123	CCCTGATTTT	0	6	*Eif4g2*	525	eukaryotic translation initiation factor 4, gamma 2
U124	CTAATAAAAG	26	45			multiple hits (3)
U125	GAGTGGATTC	21	42	*Cd63*	4426	Cd63 antigen
U126	GGGAGCGAAA	2	10	*Idb2*	1466	inhibitor of DNA binding 2[7]
U127	TAAGGGAAAT	14	32	*Tpi*	4222	triosephosphate isomerase
U128	TAATAAGGTA	0	6	EST	24543	RIKEN cDNA 4833415N24 gene
U129	TATATTGATT	2	10	*Btg1*	16596	B-cell translocation gene 1, anti-proliferative
U130	TCCACAATGA	0	6			no hit
U131	TCCCAAATGA	0	6			no hit
U132	TGACAATAAA	4	14	EST	28978	RIKEN cDNA 1200013A08 gene
U133	TGCAGTGTGC	0	6	*Sara*	6698	SAR1a gene homolog (S. cerevisiae)
U134	TGGTGTAGGA	7	19	*Hspa5*	918	heat shock 70kD protein 5 (glucose-regulated protein, 78kD)
U135	TGGTTACGTA	2	10	*Bnip2*	1561	BCL2/adenovirus E1B 19 kDa-interacting protein 1, NIP2
U136	TTACAACACT	0	7	*Pla2g4a*	4186	phospholipase A2, group IVA (cytosolic, calcium-dependent)
U137	TTGATTTTTT	8	21			multiple hits (3)
U138	TTGGATAATA	0	6	EST	219678	Mus musculus, Similar to hematopoietic PBX-interacting protein
U139	TTTATTTCAT	14	31	*Shfdg1*	2469	split hand/foot deleted gene 1

For [1]Udf and [2]Def, and for [4]Description, see the corresponding footnotes in Table 3.5 and 3.7, respectively. [3]Symbol: ESTs and all RIKEN clone genes without gene names are designated as 'EST'. [5]Gene assignment by GLGI for originally no hit tags. [6]For Northern verifiation, the indicated gene was selected among multiple hits. [7]Predicted change confirmed by Northern blot. [8]No official symbol is assigned in mouse. The prefixes for Tag #, D or U, indicate predicted downregulation or upregulation, respectively.

3.2.5.2 Cloning of genes corresponding to no-hit tags

In order to identify the gene corresponding to some no-hit tags listed in Table 3.8 (indicated by footnote 5), a PCR-based cDNA cloning by the procedure of Generation of Longer cDNA Fragments (GLGI) [57, 58] was performed, with slight modifications (see Material and methods). In seven cases (tags D24, D36, U73, U85, U86, U95 and U120) the 3' fragment (between the SAGE tag and the poly A tail) of the corresponding gene could be successfully cloned. Four of them (tags D36, U73, U86 and U120) matched to mitochondrial DNA, and tag D24 corresponded to an EST (LOC209318), while the cDNA sequences for tags U85 and U95 still did not show any match. Thus, 112

of the 139 tags predicted to be differentially expressed were assigned to 77 known genes, 18 ESTs, five mitochondrial DNA and 12 potentially novel genes.

3.2.5.3 Validation by northern blot analysis

In an approach to validate the predicted differences in gene expression, the expression of selected genes was tested by Northern blot analysis. In total, 22 probes corresponding to uniquely assigned tags as well as nine probes selected from nine multiple-hit cases were selected. The blots used were generated with RNA from undifferentiated ATDC5 cells and with RNA from cells treated with BMP4 for 6 and 24 hours. In the case of unique-match tags, 16 genes showed expected differential expression as indicated by footnote 6 in Table 3.8, while in five cases (tags D21, U81, U82, U91 and U110) no change in expression levels was detected. Only in one case, tag U77 (*pleiotrophin*), the result from Northern blot analysis was discrepant from the SAGE prediction. On the other hand, in the case of multiple-hit tags, three genes (probes selected for tags D2, D20 and U98) exhibited the expected differential expression, while four genes (probes selected for tags D72, U74, U78 and U97) did not change in their expression levels and two (probes selected for tags D26 and U100) gave rise to discrepant results. Figure 3.6 shows 15 examples where we could confirm the SAGE predictions. In summary, verification by Northern blot analysis confirmed the predicted differential expression in 19 out of 31 cases (61%) in total, or in 16 out of 22 cases (73%) from the unique-match tags.

3.2.5.4 Whole mount *in situ* hybridization

The 139 genes that we have identified are potentially BMP-regulated genes. In many cases, their functional implications in chondrogenic differentiation are not known. We therefore examined expression of selected genes in E10.5 mouse embryos by whole-mount in situ hybridization (Figure 3.7). For this purpose, we chose six upregulated genes that had been verified by Northern blot analysis, including three ESTs from tags U84, U87 and U117, and three genes from tags U90 (*Sparc*), U103 (*Tgfbi*) and U111 (*Atp6g1*). Remarkably, expression of all the six genes was found in similar (i.e., overlapping or neighboring) tissues including limb buds, somitic regions, branchial arches, nasal processes and the dorsal region of the neural tube. When compared to the expression pattern of Bmp4 (Figure 3.7, A-C), the expression patterns of the six genes largely overlapped to that of Bmp4. This result suggests that these six genes are under the control of BMP signaling in these diverse tissues.

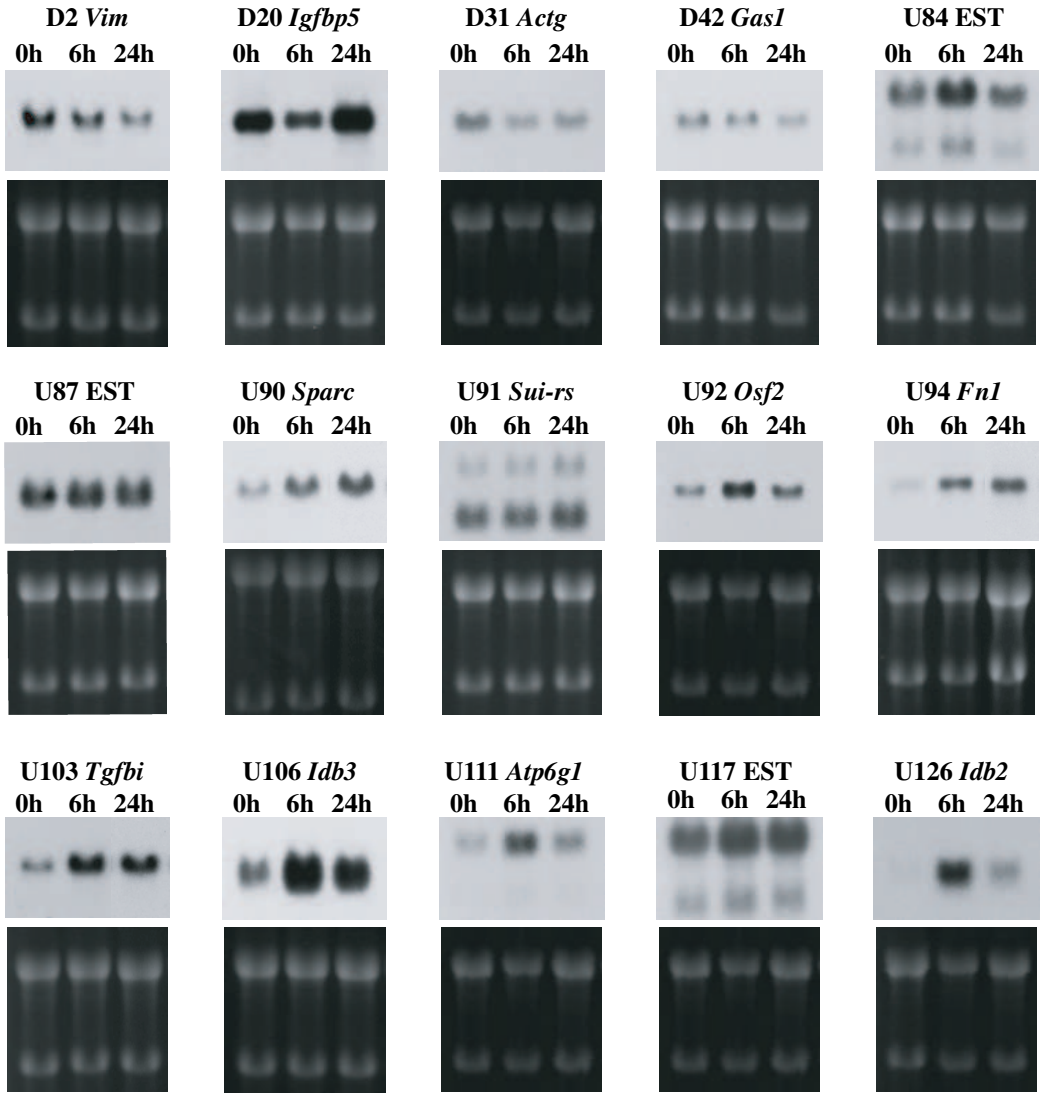


Figure 3.6: Northern blot analysis for the verification of differential expression predicted by SAGE. Expression profiles were examined at three points: uninduced (0 h) and induced for 6 or 24 hours with BMP4. The prefixes for the tag numbers, D and U, indicate predicted downregulation and upregulation, respectively. Ethidium-bromide stained RNA gels showing 28S and 18S rRNA are provided to control amounts of loaded RNA samples.

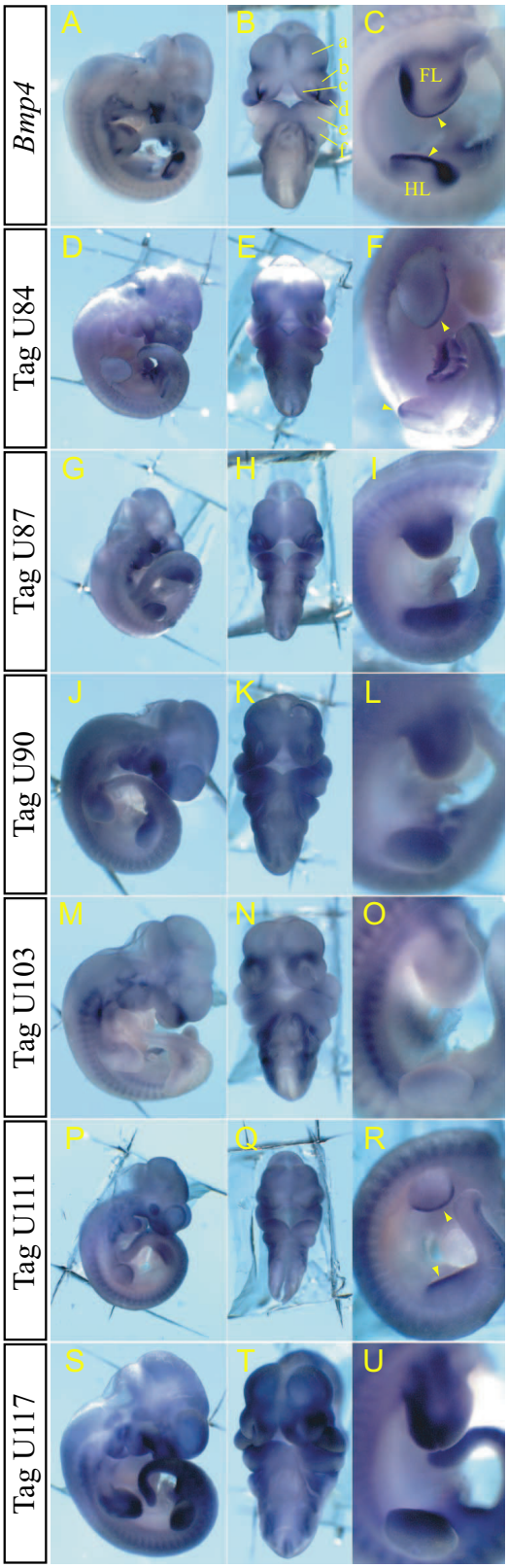


Figure 3.7: Whole-mount in situ hybridization of six selected genes (D-U) in comparison to Bmp4 (A-C) in E10.5 embryos. Upregulation of all six genes was confirmed by Northern blot (Figure 3.6). Panels A, D, G, J, M, P and S show an overview from the right lateral side, and other panels depict magnified views from either a facial-branchial region (panels B, E, H, K, N, Q and T) or a trunk region including limb buds (panels C, F, I, L, O, R and U). Arrowheads in panels C, F and R point to the apical ectodermal ridge (AER). Note three major positive domains (limb bud, nasal and branchial regions) in all six cases, as well as in the case of Bmp4. Weak expression is also seen in the trunk region within or around somites in all cases, and in the dorsal part of the neural tube in some cases (F: tag U84, R: tag U111 and U: tag U117). In limb buds, Bmp4 is expressed in the AER and also in restricted parts of limb bud mesenchyme (C). AER-specific expression is seen for tag U84 (F) and for tag U111 (R). In other cases, expression is observed in restricted parts of limb bud mesenchyme (L, O and U), or in the entire limb bud (I). Taken together, note that expression patterns of the tested six genes largely overlap to that of Bmp4. a: telencephalic vesicle, b: lateral nasal process, c: medial nasal process, d: maxillary arch, e: mandibular arch, f: hyoid arch. FL: fore limb bud, HL: hind limb bud.

3.2.6 Virtual subtraction with data from other mouse SAGE libraries

The SAGE data generated from both ATDC5 SAGE libraries combined was compared in a 'virtual subtraction' approach against all other currently available mouse SAGE libraries (listed in Table 3.9). In a total of 1,672,620 tags, collected from the publicly available SAGE data, transcripts present at 1 count per cell on average are to be detected at a probability of over 99%,based on calculation according to [98] under the assumption of 3×10^5 mRNA molecules per cell. Thus, tags exclusively observed in our ATDC5 libraries are most likely specific to our experimental system. Table 3.9 shows the numbers of ATDC5-specific tags with a count of 2 or more (excluding tags with a count of 1 to avoid potential errors from sequencing), compared either to tags from libraries grouped by the source they were derived from or to tags from all the libraries combined. In this way, 190 tags could be identified, that were not observed in any of the libraries. The most abundant tags of them, which are expressed at a count of 5 or higher, are listed in Table 3.10 with the tag numbers preceeded by S (i.e., specific). They include tags corresponding to one known gene and three ESTs, as well as three multiple-hit and 14 no-hit tags. In general, genes of unknown function (EST and no-hit tag genes) over-represented this list (17 out of 20). Interestingly, seven of them were previously listed also in Table IV as tags predicted to be differentially expressed: Tags S1, S2, S3, S5, S9, S10 and S11 in 3.10 correspond to tags U95, U85, D24, U105, U131, U130 and D65 in Table 3.8, respectively. It should be noted that all these seven tags are 'no-hit' tags and also that they are found only in either of the two ATDC5 SAGE libraries.

3.3 LongSAGE of somitogenesis

3.3.1 Tissue dissection

Initially the subsets of somitic tissue, which were taken to generate the LongSAGE libraries, were determined. As shown in Figure 3.8, a total of four cuts were made, so that known marker genes of somitogenesis would be differentially represented within the different LongSAGE libraries.

With the first cut (blue) at the level of the posterior neuropore, the tail tip (hereafter called tissue A), that expresses *Wnt3a* [106], is separated. The two successive cuts, between somitomeres S-III and S-IV (red) and at the border of the newly formed somite (between S-I and S0, yellow), result in two tissue pieces encompassing the caudal 1/3 (tissue B) and the rostral 2/3

Table 3.9: Virtual subtraction

Libraries	Total tags	Tags unique to ATDC5
Cerebellum[1]	20,713	2775
Hearts[2]	253,149	1083
Medullablastoma[3]	278,385	644
ES cells[4]	153,000	1349
E11.5 limb buds[5]	136,856	1184
Brain[6]	152,791	1679
Intraepithelial lymphocytes[7]	148,821	2127
$CD4^+$ T cells[8]	96,388	2063
$CD4^+$ spleen T cells[9]	83,855	2748
Dendritic cells[10]	179,202	1416
Testis[11]	143,506	1651
Mouse 3T3 fibroblasts[12]	25,954	2978
all combined	1672,620	190
common to all		624

SAGE tags with a count of 2 or more from both ATDC5 libraries combined compared to other publicly aviable SAGE libraries. [1]Cerebellum: Tags are combined from three libraries, one generated from C57Bl6/J P23 cerebrella (1 male, 1 female) (GSM2415) (unpublished) and two from P10 wildtype (GSM5050) and lurcher (GSM5051) mice (unpublished). [2]Hearts: Tags are combined from four libraries, one generated from adult C57Bl6/J heart (GSM1681) [99], one from undifferentiated P19 EC cells (GSM1682) and two from P19 cells induced for 0.5 (GSM16834) and 3.0 (GSM1684) hours [100]. [3]Medullablastoma: Four libraries, generated from primary tumors in PTCH +/- knockout mice (GSM766), primary granule cell precursors (GSM767), granular cell precursors cultured for 18 h in serum-free medium (GSM787) or in serum-free medium containing sonic hedgehog (GSM788) (unpublished CGAP libraries). [4]ES cells: Two libraries, generated from R1 ES cells (GSM56) [101] and ESF116 ES cells (GSM3829) (unpublished). [5]E11.5 limb buds: two libraries, generated from forelimbs (GSM55) and hindlimbs (GSM56) of E11.5 embryos [22]. [6]Brain: three libraries, generated from whole brains and spleens from three adult control females, four adult control males (all littermates of Ts65Dn P30 mice) and four Ts65Dn P30 mice (aviable from http://medgen.unige.ch) [102]. [7]Intraepithelial lymphocates: Two libraries, generated from T cell receptor (TCR)$\alpha\beta^+$ and (TCR)$\gamma\delta^+$ enriched intraepithelial lymphocytes (aviable from http://www.iiwe-irg.umds.ac.uk) [56]. [8]$CD4^+$ T cells: Six libraries generated from two resting Th1 (GSM3677 and GSM3679), one resting Th2 (GSM3678), T*reg* (GSM3679) and one resting T*skin* (GSM3680) $CD4^+$ T cell clones derived from spleen cells isolated from primed A1(M)×RAG-$1^{-/-}$ TCR female trangenic mouse that had been previously grafted with male tail skin, and draining lymph nodes taken on day 7 from male skin grafted anti-HY TCR transgenic female A1(M)×RAG-$1^{-/-}$ mice (GSM3680) [103]. [9]$CD4^+$ spleen T cells: Four libraries, two from $CD4^+CD25^-$ spleen cells purified from naive CBA/Ca mice without (GSM3683) and with overnight activation by solid phase andi-CD3 (GSM3685) and two from $CD4^+CD25^+$ spleen cells purified from naive CBA/Ca mice without (GSM3686) and with overnight activation by solid phase andi-CD3 (GSM3684) [104]. [10]Dendritic cells: Four libraries, generated from CBA/Ca mouse bone marrow derived cells cultured for 7 days in GM-CSF (GSM3833) or cultured for 6 days and treated with lipopolysaccharide (GSM3834), interleukin 10 (IL-10) (GSM3834) and IL-10 and LPS (GSM3835) (unpublished). [11]Testis: Two libraries, generated from testis of adult mice 60 days after treatment with busulphan (GSM5435) and from W^v/W^v E18 embryos (GSM5434)[105] [12]Mouse 3T3 fibroblasts: Single library generated from untransformed 3T3 fibroblasts (aviable from http://www.sagenet.org) (unpublished).

Table 3.10: List of ATDC5-specific tags with a count of at least 5

#	Sequence	Sum	Udf[1]	Def[2]	Symbol[3]	UniGene	Description[4]
S1	CCAATACGAA	37	0	37			no hit - NOVEL[5]
S2	ATAATACGAA	12	0	12			no hit - NOVEL[5]
S3	TCCTTCCGAC	12	120	0	EST		no hit - Mus musculus similar to G protein pathway suppressor 1 (XM_126542)[5]
S4	TGCTTATAAA	11	7	4			multiple hits (2)
S5	TCCCCCAATG	10	0	10			no hit
S6	GAGTCTGGGA	8	2	6			multiple hits (2)
S7	TGTAATGATT	7	1	6			multiple hits (2)
S8	TGCTGAGCAA	6	3	3			no hit
S9	TCCCAAATGA	6	0	6			no hit
S10	TCCACAATGA	6	0	6			no hit
S11	TCCCTTCGAC	6	6	0			no hit
S12	ATTTTTGTGA	5	1	4	*Ptgs2*	3137	prostaglandin-endoperoxide synthase 2
S13	ATTACAGAAA	5	0	5			no hit
S14	TGACAATTGA	5	0	5	EST	29,250	RIKEN cDNA 4930517K11 gene
S15	CTAACTCTCA	5	0	5	EST	213,927	ESTs
S16	TCCCTCCGAC	5	5	0			no hit
S17	ATTCTTTGAC	5	1	4			no hit
S18	TTCATGCCCT	5	0	5			no hit
S19	TATTAGCTAC	5	3	2			no hit
S20	TGGATTAATA	5	3	2			no hit

For [1]Udf, [2]Def, [3]Symbol and [4]Description, see the corresponding footnotes in Tables 3.5, 3.7 and 3.8. [5]Gene assignment by GLGI for originally no hit tags.

(tissue C) of the PSM. Tissue C contains the only expression domain of genes like *Hes5* [107], *EphA4* [108] and *Pcdh8* [109]. By the last cut, made between somites SII and SIII (tissue D, black), two pairs of the newly formed somites, which for example express *Unxc4.1* [110], *Tbx18* [111] and *Pax1* [112] are obtained. By this strategy genes will also be captured, that are expressed in two or more, but not all dissected tissues. For example *Fgf8* [106] and *T* (*brachiury* [113]) (data not shown) have its highest expression in A and B, *Mox1* and *Mox2* [114] and *Foxc1* and *Foxc2* [115] are weakly expressed in B and highly expressed in C and D. On the other hand, this approach is unable to determine genes cycling over the whole PSM, like *(*Lfng) [116], since by collecting material from a large amount of embryos, the SAGE tag counts obtained will represent an average of all states of a cycling gene. However, genes locally cycling within one subset, like *Mesp2* [117] in tissue C, at least will be detected as being differentially represented.

The tissue pieces do not exclusively consist of paraxial mesoderm (B - C) or somites (D) (defining the state of tissue A is difficult, since it in addition to surface ectoderm contains ectodermal cells in the transition phase to become mesoderm and endoderm). Mainly entoderm, notochord, neural tube and surface ectoderm are also included. Nevertheless, as seen in the transversal pictures in Figure 3.9, paraxial mesoderm and somites account for more than half of the cells within the respective parts.

In total, 268 stage-matched E10.5 mouse embryos have been processed in this way to obtain 268 parts of each of the four tissues.

3.3.2 Optimization of the SAGE protocol for tiny amounts of cells

According to Table 1.5, tissue A contains less than 1000 cells per specimen (the tail bud itself is not listed in the table, but it contains less cells than the last somitomere). In principle, generating SAGE libraries with that small amount of cells is possible with the modifications proposed by [55], but at the expense of a high percentage of linker tags, and also requiring large numbers of PCR cycles for the ditag amplification, whereby GC-rich ditags could be less efficiently amplified ([118], J.M. Elalouf, personal communication). Thus optimal conditions for both reverse transcription as well as RNA isolation were determined before the LongSAGE libraries were generated.

To test the efficiency of different reverse transcriptases from several manufacturers, first-strand cDNA was generated by the particular reverse transcriptase and the yield was quantified by Real-Time PCR (Roche LightCycler). Starting with aliquots of the same size from a identical poly(A) RNA

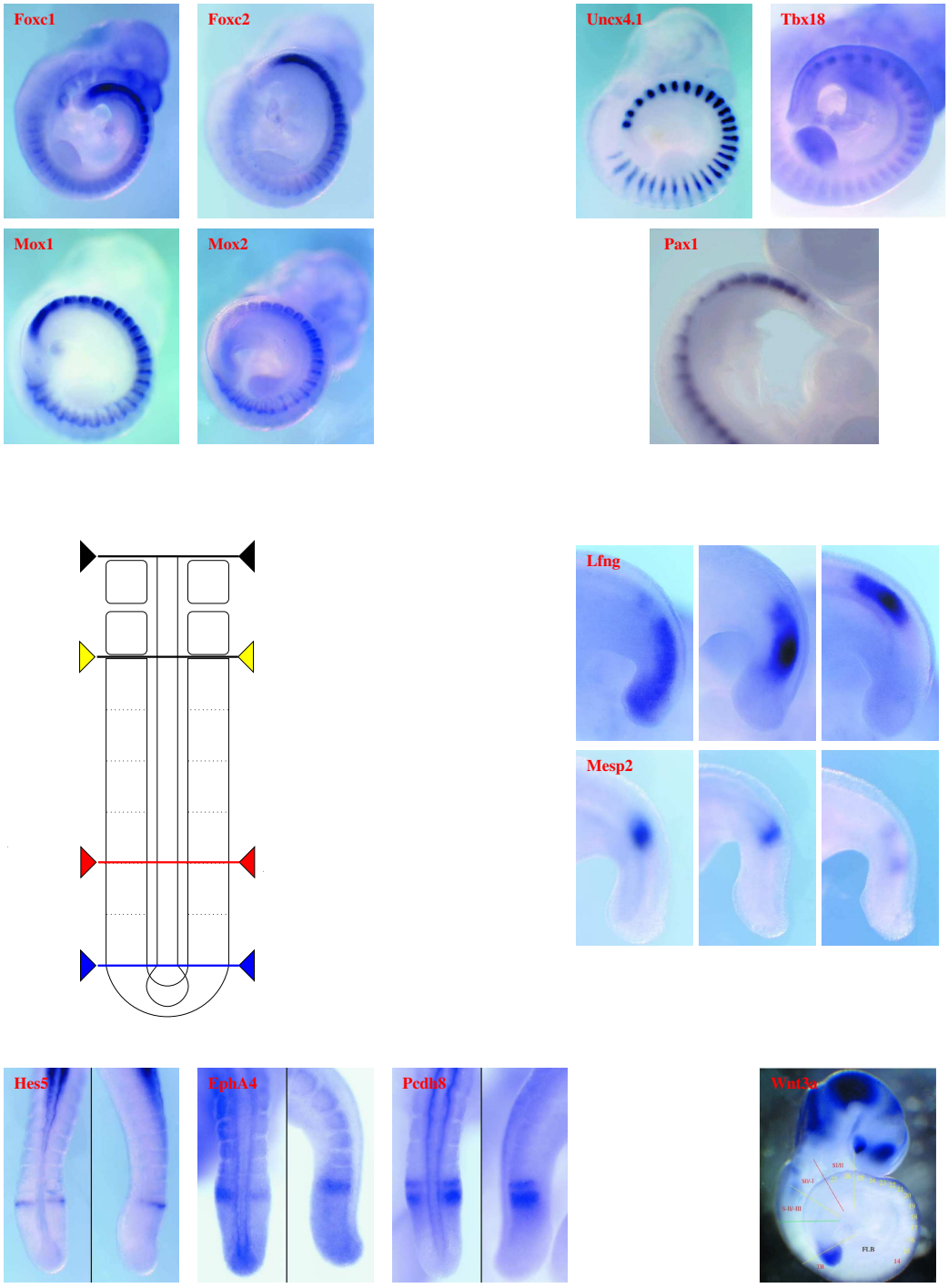


Figure 3.8: Embryos were dissected as indicated by the blue, red, yellow and black arrowheads/ lines, based upon a study of marker genes for somitogenesis. Genes with similar expression level are grouped: Top left: expressed within the rostral 2/3 of the PSM and the newly formed somites; top right: expressed within the newly formed somites; right: cycling expression over the PSM; bottom right: expressed within tail tip; bottom left: expressed within rostral 1/3 of the PSM.

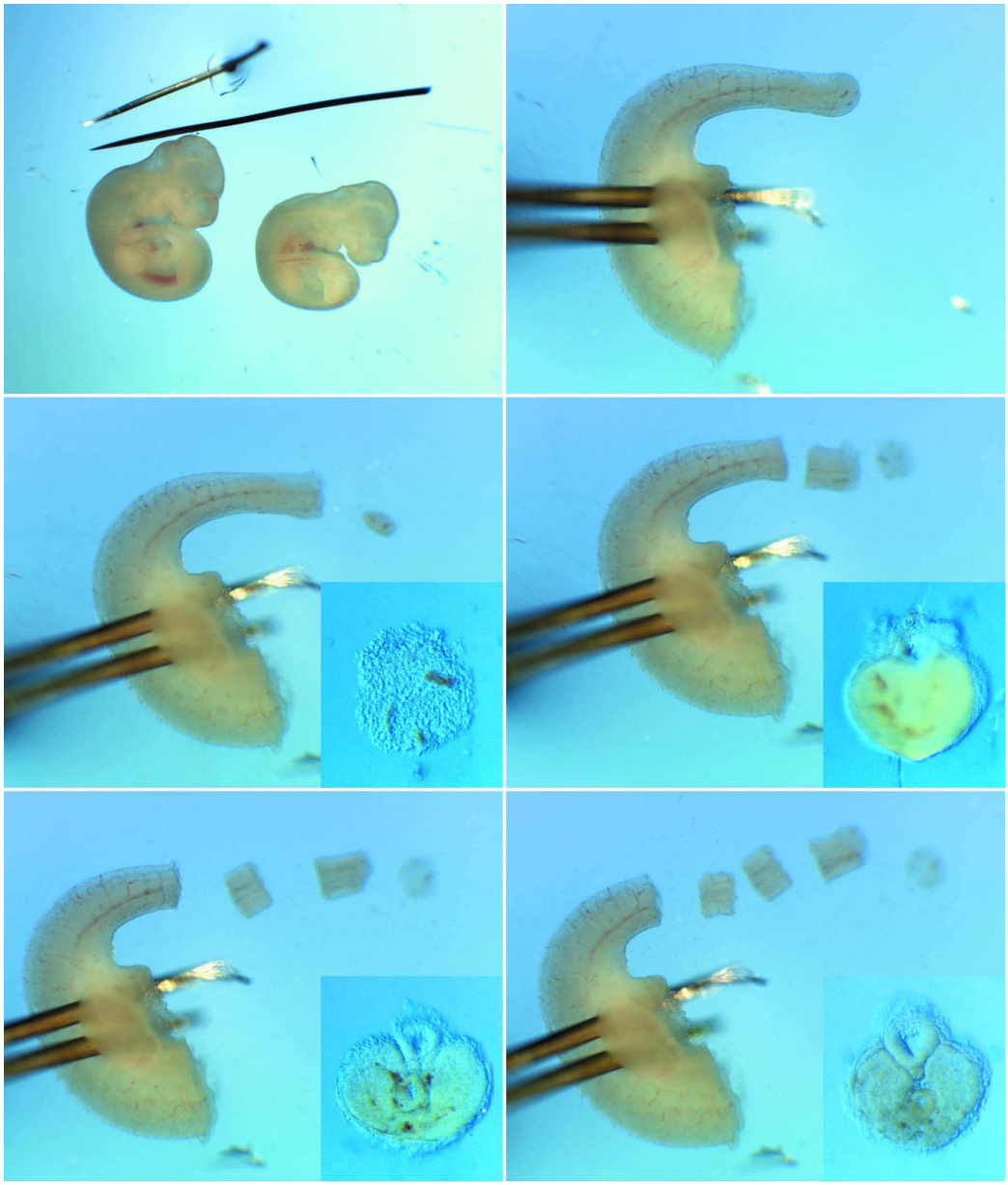


Figure 3.9: Dissection of different subsets of the tail region of E10.5 day old mouse embryos (strain C57/BL6). By four successive transversal cuts rostral to the posterior neuropore (blue arrows, resulting in part A), between the rostral 1/3 and the caudal 2/3 of the PSM (red arrows, part B), at the level between the PSM and the recently formed somite (yellow arrows, part C) and between somites SI and SII (black arrows, part D).

Table 3.11: Efficiency of different reverse transcriptases

Δcp[1] for	Roche AMV	Invitrogen SuperScript II	Qiagen Omni-Script	Qiagen Sensi-Script
Arbp	0	0.5	1.7	2.3
Hprt	0	0.6	2.6	3.1

Aliquots from identical poly(A) isolation (Qiagen mRNA Direct kit, poly(A) RNA isolated from approximately 100.000 cells per aliquot) were used for each reverse transcription (primed by oligo(dT) primers). [1]Calculated as arithmetic mean of the difference in crossing points of the same dilution between the current RT and Roche AMV. Three different dilutions were used; performed in independent duplicates. Positive values denote that the amount of cDNA synthesized by particular RT is lower by a factor of $10^{\Delta cp}$

isolation, first-strand cDNA was synthesized strictly following the manufacturers protocols. Since SAGE and LongSAGE tags are extracted from the 3' end of the transcripts, the number of transcript was quantified using two primer pairs amplifying a fragment between the third- and the second-last exons or the house-keeping gene *Hprt* (Hypoxanthine-guanine phosphoribosyltransferase; distance to 3' end of the gene[12]: 701-504 bp) and the ribosomal protein *Arbp* (60S Acidic Ribosomap protein; 704-564 bp). As shown in Table 3.11, best results were obtained for Roche Avian Myeloblastosis Virus (AMV) reverse transcriptase[13].

In order to the improve yield of RNA extractions, cell lysates were sonicated before RNA extractions. Aliquots of the same cell lysate were exposed to sonic waves for different periods of time before extracting total RNA, and the resulting amount of total RNA was measured by determining its optical density (Table 3.12). Since sonication is known to damage RNA, the integrity of the isolated RNA was successively examined by gel electrophoresis (Figure 3.10). In deed, increasing exposure time of cell lysate to sonication improved the total RNA yield, but judging from the integrity of 18S and 26S RNA, times longer than 20 seconds dramatically damaged RNA. Since the amount of tissue used for the LongSAGE study of somitogenesis was lower, a similar experiment was carried out with 100.000 cells per test. Poly(A) RNA was isolated using the Qiagen mRNA direct kit. Due to its low amount, RNA could not be measured by its optical density or by gel electrophoresis. The amount of RNA was quantified by converting the mRNA into first-strand cDNA using Roche AMV and successive quantitative Real-Time PCR (Table 3.13). Since best results were obtained for 15 seconds of sonication, this

[12]according to EnsEMBL

[13]The overall length of cDNA for Invitrogen SuperScript II RT was longest (data not shown), but since most LongSAGE tags reside near the poly(A) tail, the RT that gave rise to most copies of the 3' UTR was used.

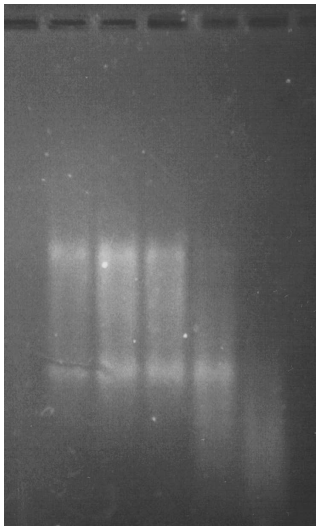


Figure 3.10: Ethidium bromide-stained gels of RNA isolated from the same tissue homogenate sonicated for different periods of time. Durations were (left to right) 0, 10, 20, 40 and 60 seconds (in pulses of 5 s; tissue sample is cooled on ice in between pulses).

Table 3.12: Effect of sonication time on total RNA yield

0 s	10 s	20 s	40 s	60 s
1	1.3	1.8	1.9	1.8

OD of the RNA extraction without sonication was taken as 1. All other ODs are given proportional to 3.10

duration was used for the subsequent construction of LongSAGE libraries.

3.3.3 General overview of SAGE libraries

For each tissue, two independent libraries were generated to minimize potential experimental bias, and for each of the (total of 8) libraries, (except for library A) around 25,000 tags were sequenced (see Table 3.14). Contamination between the different libraries is avoided by using distinct Linker pairs for each LongSAGE library construction. Even though the starting amount

Table 3.13: Effect of sonication time on poly(A) RNA yield

Δcp[1] for	0 s[2]	10 s	15 s	20 s
Arbp	0	-1.0	-2.2	-1.4
Hprt	0	-0.8	-2.1	-1.5

[1]Arithmetic mean of the difference in crossing points between specified duration of sonication and 0 s (no sonication). Three different dilutions were used; performed in independent duplicates. Negative values denote that the amount of cDNA synthesized by particular RT is lower by a factor of $10^{-\Delta cp}$. [2]In pulses of 5 s; tissue sample is cooled on ice in between pulses

Table 3.14: Summary of somite LongSAGE libraries

	Sum[1]	A[2]	B[3]	C[4]	D[5]
Statistics					
Total tags (excluding linker tags)	171,639	21,595	50,699	49,732	49,613
Unique transcripts	49,128	9,882	19,395	20,701	18,126
Transcripts observed in all libraries	2,633				
Transcripts observed only in one librariy		4,690	11,113	12,512	10,022
with a count of > 1		550	582	772	477
Quality					
Linker contamination					
Frequency of identical ditags					

[1]Sum, tag counts from the four libraries combined. SAGE libraries generated from [2]tail tip, [3]caudal 2/3 of the PSM, [4]rostral 1/3 of the PSM and [5]recently formed two pairs of somites.

of cells greatly differed between the tissues used for library construction (see Table 1.5), the same number of PCR cycles (24 + 12) was used for the ditag amplification in any of the LongSAGE libraries. In total, excluding linker tags, 171,639 LongSAGE tags were collected from all libraries combined. These correspond to 49,128 unique transcripts, of which only 2,633 are common to all four libraries. However, excluding library A, whose total tag count is less than half of the others, 4,333 tags are observed in libraries B to D. The frequency of linker contamination and duplicate ditags, as well as the fact that most of the abundant tags were represented in all libraries, proved the high quality as well as the reproducibility of the LongSAGE libraries.

3.3.4 Tag to Gene assignment

Table 3.15 summarizes the statistics for the assignment of all LongSAGE tags to the different LongSAGE mappings, sorted by abundance classes. Like for the ATDC5 SAGE data, the number of no-hit tags increased with decreasing tag abundance.

3.3.5 Genes abundantly expressed in the presomitic mesoderm and the first formed somites

Table 3.16 lists the thirty most abundant LongSAGE tags in all somitic LongSAGE libraries combined. Surprisingly, the majority (18 tags) hit multiple times to the databases. However, except for A1, A2 and A3, these multiple hits could be manually resolved due to the existence of pseudogenes (10 cases; indicated by footnote 7) or because of the inability to automatically merge UniGene and Genome hit to the same gene (footnote 8).

Table 3.15: Statistics of LongSAGE tag assignment sorted by abundance classes

Tag counts in the data set	Abundance in tpm[1]	Different kinds of tags	Total number of tags	UniGene hits	Hits against whole genome	Genome hits	Mapping[2]
> 31	>180	607	71,336	555 (91%)	585 (96%)	506 (86%)	575 (95%)
11-30	64 - 174	1478	24,686	1229 (85%)	1374 (93%)	1034 (70%)	1309 (89%)
6-10	35 - 58	1899	14,177	1371 (81%)	1674 (88%)	1128 (59%)	1489 (78%)
4-5	23 - 29	2068	9111	1305 (63%)	1714 (83%)	1031 (50%)	1451 (70%)
3	17	2133	6399	1137 (53%)	1720 (81%)	886 (42%)	1303 (61%)
2	12	4987	9974	1848 (37%)	3389 (68%)	1513 (30%)	2289 (46%)
1	6	35,956	35,956	4303 (12%)	22,522 (63%)	3913 (11%)	6086 (17%)
Total		49,128	171,639	11,748 (24%)	31304 (64%)	10011 (20%)	14502 (30%)

[1]tpm: tags-per-million. [2]Mapping: Combination of mapping against UniGene (transcript) and EnsEMBL (genome).

Table 3.16: List of thirty most abundant tags in the somite libraries

#	Sequence	Sum[1]	A[2]	B[3]	C[4]	D[5]	UniGene	Symbol	Description[5]
A1	GTGGCTCACAACCATCC	16424	15374	15247	15222	19289			multiple hits (16)
A2	GTGGCTCACAACCATCT	10353	13661	9487	8988	11166			multiple hits (59)
A3	CAAGGTGACAGGCCGCT	9910	9956	8738	9008	11993			multiple hits (23)
A4	TGGCTCGGTCACTTGGG	10103	7872	9507	10597	11187			multiple hits (2): ACTIN, CYTOPLASMIC 2 (GAMMA-ACTIN), ACTIN-LIKE
A5	GGGACTGCATTGAGAGC	7586	11484	8442	5469	7135	196718	*Hbb-b2,Hbb-bh1,Hbb-b1,Hbb-y*	HEMOGLOBIN EPSILON-Y2 CHAIN. [Source:SWISSPROT;Acc:P02104]
A6	AGCAGTCCCCTCCCTAG	6420	6900	6213	6756	6087	no hit		
A7	GAGCGTTTTGGGTCCAG	6479	5742	6430	6073	7256			multiple hits (9): CYCLOPHILIN A[7]
A8	CCTACCAAGACTTTGAG	5156	5511	5621	4042	5644			multiple hits (3): 40S RIBOSOMAL PROTEIN S20[7]
A9	TTCATTATAATCTCAAA	4842	6622	5010	5067	3668	19187	EST	prothymosin alpha
A10	CCCTGAGTCCACCCCGG	4294	3334	4280	3660	5362	297	EST	ACTIN, CYTOPLASMIC 1 (BETA-ACTIN) [Source:SWISSPROT;Acc:P02570]
A11	GCCTCCAAGGAGTAAGA	3892	2825	3925	3217	4999			multiple hits (28): GLYCERALDEHYDE 3-PHOSPHATE DEHYDROGENASE[7]
A12	GGCTTCGGTCTTTTTGA	3770	2176	4162	2694	5140			multiple hits (2): 60S ACIDIC RIBOSOMAL PROTEIN P1[7]
A13	GGAAGCCACTTTGACAG	3577	2454	3491	2896	4837			multiple hits (4): 40S RIBOSOMAL PROTEIN S27A[7]

table continues on following page

#	Sequence	Sum[1]	A[2]	B[3]	C[4]	D[5]	UniGene	Symbol	Description[5]
A14	GGCTTCCGCGAGGGTAC	3577	2315	3610	2855	4817	906	*Mdk*	MIDKINE PRECURSOR (RETINOIC ACID-INDUCED DIFFERENTIATION FACTOR). [Source:SWISSPROT;Acc:P12025]
A15	AGGCAGACAGTTGCTGT	3303	3705	3669	2835	3225			multiple hits (5): ELONGATION FACTOR 1-ALPHA 1 (EF-1-ALPHA-1) (ELONGATION FACTOR 1 A-1) (EEF1A-1) (ELONGATION FACTOR TU) (EF-TU)[7]
A16	CCCTTCTTCTCTCCCTT	3111	4677	3373	2755	2520			multiple hits (2): HEMOGLOBIN ZETA CHAIN, HEMOGLOBIN ALPHA CHAIN
A17	CGCCGCCGGCTCACCAA	3088	2454	2840	2855	3850	16423	*Rpl35*	RIBOSOMAL PROTEIN L35
A18	GGATTTGGCTTGTTTGA	3158	1852	3077	2755	4213			multiple hits (2): 60S ACIDIC RIBOSOMAL PROTEIN P2[8]
A19	GTGGCTCACAACCACCC	2907	3334	2604	2433	3507			multiple hits (2): MITOCHONDRIAL RIBOSOMAL PROTEIN L51; MITOCHONDRIAL RIBOSOMAL PROTEIN 64[7]
A20	ATGACTGATAGCAAGTC	2971	2871	3432	3056	2459	no hit		
A21	TTTAATAAAGATCATCC	2517	5140	2327	2252	1834	35830	*Hbb-b2,Hbb-bh1,Hbb-b1,Hbb-y*	HEMOGLOBIN EPSILON-Y2 CHAIN. [Source:SWISSPROT;Acc:P02104]
A22	CAAAAATAAAAGCCGCA	2628	3936	2110	2534	2681			multiple hits (5): ALPHA ENOLASE (EC 4.2.1.11) (2-PHOSPHO-D-GLYCERATE HYDRO-LYASE) (NON- NEURAL ENOLASE) (NNE) (ENOLASE 1)[7]
A23	AGATCTATACAGTCGGG	2872	2315	3314	2332	3205			multiple hits (2): 60S RIBOSOMAL PROTEIN L7[7]
A24	GTTGCTGAGAAGCGGCT	2698	3241	2466	2554	2842			multiple hits (5): QM PROTEIN (FRAGMENT), 60S RIBOSOMAL PROTEIN L10 (QM PROTEIN HOMOLOG)[7]
A25	ACCAGCTATGATCCCTC	2365	4353	2525	2131	1572	141758	*Hba-a1,Hba-x*	hemoglobin X, alpha-like embryonic chain in Hba complex
A26	AGAGCGAAGTGGCGGAA	2505	3288	2209	2252	2721			multiple hits (2): 60S RIBOSOMAL PROTEIN L41[8]
A27	GAAGCAGGACCAGTAAG	2587	2501	2387	2694	2721	4024	*Cfl1*	COFILIN, NON-MUSCLE ISOFORM. [Source:SWISSPROT;Acc:P18760]
A28	CCCGTGTGCTCATCCGC	2564	2454	2860	1910	2963			multiple hits (2): RIKEN cDNA 3010033P07 gene, RIKEN cDNA 6330437E22 gene
A29	CAGGCCACACAAGAGCC	2540	2176	2268	2332	3185	103838	*Atp5b*	ATP SYNTHASE BETA CHAIN, MITOCHONDRIAL PRECURSOR (EC 3.6.3.14). [Source:SWISSPROT;Acc:P56480]
A30	GGGGAAATCGCCAGCTT	2703	1158	2899	2855	3023			multiple hits (2): THYMOSIN, BETA 10[8]

Count values in tags-per-million. For [2]A, [3]B, [4]C and [5]D, see the corresponding footnotes in Table 3.14.[1]Sum: Tags-per-million for all libraries combined (count in all libraries, divided by total tag count, multiplied by 1,000,000. Note that tags are not sorted by Sum, but by sum of tpms of all four tissues. [6]Description: in case of multiple hits, the number of matching genes is given in parenthesis. [7]Multiple hits to genome due to pseudogenes (only single UniGene entry). [8]Hit to identical gene; automated link of Genome hit impossible due to not annotated 3' UTR.

3.3.6 Tags differentially expressed between the subsets

A total of 1007 genes were predicted to be differentially represented at the confidence limit of 95% ($P \leq 0.05$) between at least two of the four tissues analyzed. Of these, 625 tags could be reliably assigned to a single gene, whereas 173 tags matched to more than one gene, and 209 tags could not be assigned to any gene.

Table 3.17: Tags differentially expressed between libraries

	Any[1]	A[2]	B[3]	C[4]	D[5]
vs A					
vs B		301			
vs C		315	307		
vs D		350	255	255	
vs any other	1007	575	721	721	713

[1]Any of the four libraries. SAGE libraries generated from [2]tail tip, [3]caudal 2/3 of the PSM, [4]rostral 1/3 of the PSM and [5]recently formed two pairs of somites.

3.3.7 Changes to members of FGF, Wnt and Delta/ Notch signaling pathways

For known members of the FGF, Wnt and Delta/Notch signaling pathways the representation of the corresponding tags in the whole dataset was analyzed. The results are given in Table 3.18. Tags were only considered if they match only to the particular gene.

Table 3.18: Representation of tags for FGF, Wnt and Delta/ Notch signaling pathways in the dataset

A[1]	B[2]	C[3]	D[4]	Symbol	Name
FGF signaling pathway					
Ligands					
0	0	0	0	*Fgf1*	fibroblast growth factor 1
0	0	0	0	*Fgf2*	fibroblast growth factor 2
0	0	0	0	*Fgf3*	fibroblast growth factor 3
0	0	0	0	*Fgf4*	fibroblast growth factor 4
0	0	0	0	*Fgf5*	fibroblast growth factor 5
0	0	0	0	*Fgf6*	fibroblast growth factor 6
0	0	0	0	*Fgf7*	fibroblast growth factor 7
139	158	40	0	*Fgf8*	fibroblast growth factor 8
0	0	0	0	*Fgf9*	fibroblast growth factor 9
0	0	0	0	*Fgf10*	fibroblast growth factor 10
0	0	0	0	*Fgf11*	fibroblast growth factor 11
0	0	0	0	*Fgf12*	fibroblast growth factor 12
0	0	0	20	*Fgf13*	fibroblast growth factor 13
0	0	0	0	*Fgf14*	fibroblast growth factor 14
0	20	0	0	*Fgf15*	fibroblast growth factor 15

table continues on following page

A[1]	B[2]	C[3]	D[4]	Symbol	Name
0	0	0	0	*Fgf16*	fibroblast growth factor 16
0	0	0	0	*Fgf17*	fibroblast growth factor 17
46	0	0	0	*Fgf18*	fibroblast growth factor 18
0	0	0	0	*Fgf20*	fibroblast growth factor 20
0	0	0	0	*Fgf21*	fibroblast growth factor 21
0	0	0	0	*Fgf22*	fibroblast growth factor 22
0	0	0	0	*Fgf23*	fibroblast growth factor 23
Receptors					
325	119	141	60	*Fgfr1*	fibroblast growth factor receptor 1
0	0	60	0	*Fgfr2*	fibroblast growth factor receptor 2
0	0	0	0	*Fgfr3*	fibroblast growth factor receptor 3
0	0	0	0	*Fgfr4*	fibroblast growth factor receptor 4
0	20	0	0	*Fgfrl1*	fibroblast growth factor receptor-like 1
Mediators					
RAS					
0	0	40	20	*Diras1*	DIRAS family, GTP-binding RAS-like 1
0	0	0	0	*Eras*	ES cell-expressed Ras
0	20	0	0	*Ermap*	erythroblast membrane-associated protein
0	0	0	0	*Kbras1-pending*	I-kappa-B-interacting Ras-like protein 1
0	0	0	0	*Kbras2-pending*	I-kappa-B-interacting Ras-like protein 2
0	0	0	0	*Mras*	muscle and microspikes RAS
0	197	201	241	*Nras*	neuroblastoma ras oncogene
0	0	0	40	*Rasa1*	RAS p21 protein activator 1
0	0	0	0	*Rasa2*	RAS p21 protein activator 2
0	0	20	0	*Rasa3*	RAS p21 protein activator 3
139	59	241	40	*Rasal1*	RAS protein activator like 1 (GAP1 like)
0	0	0	0	*Rasd1*	RAS, dexamethasone-induced 1
0	0	0	0	*Rasgrf1*	RAS protein-specific guanine nucleotide-releasing factor 1
0	0	0	0	*Rasgrf2*	RAS protein-specific guanine nucleotide-releasing factor 2
0	0	0	0	*Rasgrp1*	RAS guanyl releasing protein 1
0	20	0	0	*Rasgrp2*	RAS, guanyl releasing protein 2
0	20	0	0	*Rasgrp4*	RAS guanyl releasing protein 4
509	394	664	585	*Rasl2-9*	RAS-like, family 2, locus 9
0	0	0	0	*Rras2*	related RAS viral (r-ras) oncogene homolog 2
MAPK					
92	20	221	60	*Mapk1*	mitogen activated protein kinase 1
92	0	0	60	*Mapk3*	mitogen activated protein kinase 3
0	0	0	0	*Mapk4*	mitogen-activated protein kinase 4
0	20	20	20	*Mapk6*	mitogen-activated protein kinase 6
46	20	0	0	*Mapk7*	mitogen-activated protein kinase 7
0	0	20	0	*Mapk8*	mitogen activated protein kinase 8
46	20	0	40	*Mapk9*	mitogen activated protein kinase 9
0	0	0	0	*Mapk10*	mitogen activated protein kinase 10
0	0	20	0	*Mapk11*	mitogen-activated protein kinase 11
0	0	0	0	*Mapk12*	mitogen-activated protein kinase 12

table continues on following page

A[1]	B[2]	C[3]	D[4]	Symbol	Name
0	0	0	20	*Mapk13*	mitogen activated protein kinase 13
46	0	60	60	*Mapk14*	mitogen activated protein kinase 14
0	0	20	20	*Mapbpip-pending*	mitogen activated protein binding protein interacting protein
139	99	20	121	*Mapk8ip*	mitogen activated protein kinase 8 interacting protein
0	0	0	0	*Mapk8ip2*	mitogen-activated protein kinase 8 interacting protein 2
0	79	20	0	*Mapk8ip3*	mitogen-activated protein kinase 8 interacting protein 3
0	0	0	0	*Mapkap1*	mitogen-activated protein kinase associated protein 1
46	39	0	120	*Mapkapk2*	MAP kinase-activated protein kinase 2
46	79	40	60	*Mapkapk5*	MAP kinase-activated protein kinase 5
93	20	60	20	*Mapkbp1*	mitogen activated protein kinase binding proten 1
MAPKK					
92	39	40	40	*Map2k1*	mitogen activated protein kinase kinase 1
93	158	80	40	*Map2k2*	mitogen activated protein kinase kinase 2
0	20	20	0	*Map2k3*	mitogen activated protein kinase kinase 3
0	20	20	60	*Map2k4*	mitogen activated protein kinase kinase 4
46	0	20	40	*Map2k5*	mitogen activated protein kinase kinase 5
0	0	20	20	*Map2k6*	mitogen activated protein kinase kinase 6
0	0	0	0	*Map2k7*	mitogen activated protein kinase kinase 7
139	0	20	0	*Map2k1ip1*	mitogen-activated protein kinase kinase 1 interacting protein 1
MAPKKK					
0	59	0	20	*Map3k1*	mitogen activated protein kinase kinase kinase 1
0	0	0	0	*Map3k2*	mitogen activated protein kinase kinase kinase 2
0	20	0	0	*Map3k3*	mitogen activated protein kinase kinase kinase 3
0	0	40	40	*Map3k4*	mitogen activated protein kinase kinase kinase 4
0	20	0	0	*Map3k5*	mitogen activated protein kinase kinase kinase 5
0	0	0	0	*Map3k6*	mitogen-activated protein kinase kinase kinase 6
0	0	0	0	*Map3k7*	mitogen activated protein kinase kinase kinase 7
0	0	0	0	*Map3k8*	mitogen activated protein kinase kinase kinase 8
0	0	0	0	*Map3k9*	mitogen-activated protein kinase kinase kinase 9
0	0	0	0	*Map3k10*	mitogen activated protein kinase kinase kinase 10
0	0	0	0	*Map3k11*	mitogen activated protein kinase kinase kinase 11
278	592	442	504	*Map3k12*	mitogen activated protein kinase kinase kinase 12

table continues on following page

A[1]	B[2]	C[3]	D[4]	Symbol	Name
0	0	0	0	*Map3k14*	mitogen-activated protein kinase kinase kinase 14
139	20	120	121	*Map3k7ip1*	mitogen-activated protein kinase kinase kinase 7 interacting protein 1
185	59	80	20	*Map3k7ip2*	mitogen-activated protein kinase kinase kinase 7 interacting protein 2
MAPKKKK					
0	0	0	0	*Map4k1*	mitogen activated protein kinase kinase kinase kinase 1
46	0	40	0	*Map4k2*	mitogen activated protein kinase kinase kinase kinase 2
0	0	0	0	*Map4k3*	mitogen-activated protein kinase kinase kinase kinase 3
0	79	20	20	*Map4k4*	mitogen-activated protein kinase kinase kinase kinase 4
0	0	20	40	*Map4k5*	mitogen-activated protein kinase kinase kinase kinase 5
46	20	20	60	*Map4k6-pending*	mitogen-activated protein kinase kinase kinase kinase 6
Wnt signaling pathway					
Ligands					
0	0	0	20	*Wnt1*	wingless-related MMTV integration site 1
0	0	0	0	*Wnt2*	wingless-related MMTV integration site 2
0	0	0	0	*Wnt2b*	wingless related MMTV integration site 2b
0	0	60	20	*Wnt3*	wingless-related MMTV integration site 3
93	79	0	0	*Wnt3a*	wingless-related MMTV integration site 3A
0	0	0	0	*Wnt4*	wingless-related MMTV integration site 4
139	158	161	40	*Wnt5a*	wingless-related MMTV integration site 5A
185	0	0	0	*Wnt5b*	wingless-related MMTV integration site 5B
0	20	20	0	*Wnt6*	wingless-related MMTV integration site 6
0	40	0	0	*Wnt7a*	wingless-related MMTV integration site 7A
0	0	20	0	*Wnt7b*	wingless-related MMTV integration site 7B
0	0	0	0	*Wnt8a*	wingless-related MMTV integration site 8A
0	0	0	0	*Wnt8b*	wingless related MMTV integration site 8b
0	0	0	0	*Wnt9a*	wingless-type MMTV integration site 9A
0	0	0	0	*Wnt9b*	wingless-type MMTV integration site 9B
0	0	0	0	*Wnt10a*	wingless related MMTV integration site 10a
0	0	0	0	*Wnt10b*	wingless related MMTV integration site 10b
0	39	20	0	*Wnt11*	wingless-related MMTV integration site 11
0	0	0	0	*Wnt16*	wingless-related MMTV integration site 16
Receptors					
0	0	0	0	*Fzd1*	frizzled homolog 1 (Drosophila)
0	20	60	40	*Fzd2*	frizzled homolog 2 (Drosophila)
0	0	0	20	*Fzd3*	frizzled homolog 3 (Drosophila)
0	0	0	0	*Fzd4*	frizzled homolog 4 (Drosophila)
0	0	0	0	*Fzd5*	frizzled homolog 5 (Drosophila)
46	0	0	0	*Fzd6*	frizzled homolog 6 (Drosophila)
46	79	20	40	*Fzd7*	frizzled homolog 7 (Drosophila)
0	0	20	0	*Fzd8*	frizzled homolog 8 (Drosophila)
0	0	0	0	*Fzd9*	frizzled homolog 9 (Drosophila)

table continues on following page

A[1]	B[2]	C[3]	D[4]	Symbol	Name
0	0	0	0	*Fzd10*	frizzled homolog 10 (Drosophila)
92	157	120	60	*Frzb*	frizzled-related protein
Co-Receptors					
46	39	201	20	*Lrp1*	low density lipoprotein receptor-related protein 1
0	0	40	0	*Lrp2*	low density lipoprotein receptor-related protein 2
0	0	0	0	*Lrp4*	low density lipoprotein receptor-related protein 4
0	20	40	0	*Lrp5*	low density lipoprotein receptor-related protein 5
0	20	0	0	*Lrp6*	low density lipoprotein receptor-related protein 6
0	0	0	0	*Lrp8*	low density lipoprotein receptor-related protein 8, apolipoprotein e receptor
46	59	101	81	*Lrp10*	low-density lipoprotein receptor-related protein 10
0	0	0	0	*Lrp1b*	low density lipoprotein-related protein 1B (deleted in tumors)
0	0	0	0	*Lrp2bp-pending*	low density lipoprotein receptor-related protein 2 binding protein
46	0	0	0	*Lrpap1*	low density lipoprotein receptor-related protein associated protein 1
Extracellular mediators					
0	20	0	0	*Dkk1*	dickkopf homolog 1 (Xenopus laevis)
0	0	0	0	*Dkk2*	dickkopf homolog 2 (Xenopus laevis)
93	40	80	81	*Dkk3*	dickkopf homolog 3 (Xenopus laevis)
0	0	0	0	*Dkk4*	dickkopf homolog 4 (Xenopus laevis)
Intracellular mediators					
93	39	80	0	*Dvl1*	dishevelled, dsh homolog 1 (Drosophila)
0	0	0	40	*Dvl2*	dishevelled 2, dsh homolog (Drosophila)
0	0	0	0	*Dvl3*	dishevelled 3, dsh homolog (Drosophila)
0	0	0	0	*Daam1*	dishevelled associated activator of morphogenesis 1
0	0	0	0	*Daam2*	dishevelled associated activator of morphogenesis 2
46	79	20	20	*Nkd1*	naked cuticle 1 homolog (Drosophila)
139	178	121	0	*Nkd2*	naked cuticle 1 homolog (Drosophila)
0	20	20	20	*Axin*	axin
46	79	20	20	*Axin2*	axin2
0	0	0	0	*Idax-pending*	inhibitor of the Dvl and Axin complex
0	0	0	0	*Gsk3b*	glycogen synthase kinase 3 beta
139	59	141	181	*Catnb*	catenin beta
0	0	0	0	*Catnbip1*	catenin beta interacting protein 1
0	20	20	40	*Dact1*	dapper homolog 1, antagonist of beta-catenin (xenopus)
231	79	341	120	*Apc*	adenomatosis polyposis coli
0	0	0	0	*Apc2*	adenomatosis polyposis coli 2
0	119	40	121	*Tcf1*	transcription factor 1
0	0	0	0	*Tcf2*	transcription factor 2

table continues on following page

A[1]	B[2]	C[3]	D[4]	Symbol	Name
0	0	0	0	*Tcf3*	transcription factor 3
139	139	100	81	*Tcf4*	transcription factor 4
0	0	0	0	*Tcf7*	transcription factor 7, T-cell specific
0	0	0	0	*Tcf7l2*	transcription factor 7-like 2, T-cell specific, HMG-box
0	0	20	60	*Tcf12*	transcription factor 12
93	79	141	222	*Tcf15*	transcription factor 15
0	0	0	0	*Tcf19*	transcription factor 19
0	0	20	0	*Tcf20*	transcription factor 20
0	0	0	0	*Tcf21*	transcription factor 21
0	0	0	0	*Tcf23*	transcription factor 23
139	217	181	161	*Tcfl1*	transcription factor-like 1
0	20	101	60	*Tcfl4*	transcription factor-like 4
0	0	0	20	*Tcfap2a*	transcription factor AP-2, alpha
139	59	261	60	*Tcfap2b*	transcription factor AP-2 beta
0	0	0	20	*Tcfap2c*	transcription factor AP-2, gamma
0	0	0	0	*Tcfap2d*	transcription factor AP-2, delta
0	20	0	0	*Tcfcp2*	transcription factor CP2
185	178	160	100	*Tcfe2a*	transcription factor E2a
0	39	0	20	*Tcfe3*	transcription factor E3
0	0	0	0	*Tcfeb*	transcription factor EB
0	0	0	0	*Tcfec*	transcription factor EC
232	158	242	161	*Lef1*	lymphoid enhancer binding factor 1
Delta/Notch signaling pathway					
Ligands					
0	39	0	40	*Dlk1*	delta-like 1 homolog (Drosophila)
139	178	80	0	*Dll1*	delta-like 1 (Drosophila)
139	276	241	0	*Dll3*	delta-like 3 (Drosophila)
0	0	20	0	*Dll4*	delta-like 4 (Drosophila)
139	59	241	40	*Jag1*	jagged 1
0	20	0	20	*Jag2*	jagged 2
Receptors					
0	139	261	81	*Notch1*	Notch gene homolog 1 (Drosophila)
0	0	0	20	*Notch2*	Notch gene homolog 2 (Drosophila)
0	0	80	0	*Notch3*	Notch gene homolog 3 (Drosophila)
0	0	0	0	*Notch4*	Notch gene homolog 4 (Drosophila)
0	0	0	0	*Dner*	delta/notch-like EGF-related receptor
Notch glycosylation					
0	39	40	0	*Lfng*	lunatic fringe gene homolog (Drosophila)
0	20	0	0	*Mfng*	manic fringe homolog (Drosophila)
0	0	0	20	*Rfng*	radical fringe gene homolog (Drosophila)
0	0	0	0	*Frcl1*	frc, fringe-like 1 (Drosophila)
Proteolytical release of notch intracellular domain (NICD)					
0	0	0	0	*Psen1*	presenilin 1
0	0	0	0	*Psen2*	presenilin 2
Mediators of notch signaling					
0	20	20	0	*Mesp1*	mesoderm posterior 1
0	0	121	0	*Mesp2*	mesoderm posterior 2

table continues on following page

A[1]	B[2]	C[3]	D[4]	Symbol	Name
NICD association/ DNA binding					
880	454	623	463	*Rbpsuh*	recombining binding protein suppressor of hairless (Drosophila)
0	0	0	0	*Rbpsuhl*	recombining binding protein suppressor of hairless-like (Drosophila)
Effectors					
0	0	0	0	*Gprk2l*	G protein-coupled receptor kinase 2, groucho gene related (Drosophila)
0	0	20	0	*Herpud1*	homocysteine-inducible, endoplasmic reticulum stress-inducible, ubiquitin-like domain member 1
0	20	20	60	*Hes1*	hairy and enhancer of split 1 (Drosophila)
0	0	0	0	*Hes2*	hairy and enhancer of split 2 (Drosophila)
0	39	20	60	*Hes3*	hairy and enhancer of split 3 (Drosophila)
0	0	0	20	*Hes5*	hairy and enhancer of split 5 (Drosophila)
185	79	60	161	*Hes6*	hairy and enhancer of split 6 (Drosophila)
46	39	0	0	*Hes7*	hairy and enhancer of split 7 (Drosophila)
0	20	40	60	*Hey1*	hairy/enhancer-of-split related with YRPW motif 1
0	20	0	0	*Hey2*	hairy/enhancer-of-split related with YRPW motif 2
0	20	0	0	*Heyl*	hairy/enhancer-of-split related with YRPW motif-like
46	20	40	81	*Tle1*	transducin-like enhancer of split 1, homolog of Drosophila E(spl)
0	20	20	20	*Tle2*	transducin-like enhancer of split 2, homolog of Drosophila E(spl)
46	20	40	101	*Tle3*	transducin-like enhancer of split 3, homolog of Drosophila E(spl)
46	0	0	0	*Tle4*	transducin-like enhancer of split 4, E(spl) homolog (Drosophila)
0	20	0	0	*Tle6*	transducin-like enhancer of split 6, homolog of Drosophila E(spl)
0	0	0	0	*Aes*	amino-terminal enhancer of split
0	0	0	20	*hr*	hairless

Count values in tags-per-million. For [1]A , [2]B, [3]C and [4]D, see the corresponding footnotes in Table 3.14.

3.3.8 Functional annotation of genes represente in the dataset

The GeneOntology (GO) consortium [64, 79] provides a hierarchical structured set of controlled vocabularies (ontologies), that describe gene products in terms of their associated biological processes, cellular components and molecular functions. Except for the evidence code IEA (inferred from electronic annotation), all GO associations are manually curated, therefore it can

Table 3.19: Statistics for functional annotation of differentially expressed genes

top ontologies	GO	GO without IEA[1]	InterPro-Scan	all combined
Molecular Function	289	154	323	404
Biological Process	256	133	246	334
Cellular Component	265	178	157	310
all	347	234	355	453

[1]Evidence code IEA (inferred from electronic annotation): It is the only GeneOntology source that has not been manually curated. [2]Considered only if single GeneScan Prediction associated to single or multiple Genome hits.

be considered as a high-quality data source. However, new findings within the scientific community lead to knowledge changes and updates, whose incorporation to GO always lag behind. Additionally, GO does not provide a comprehensive dataset. Due to lack of GO curators, many genes with an MGI Marker ID have no or incomplete GO associations, even though the respective data is available. Furthermore, there were many genes identified in the dataset, that are either only represented by UniGene clusters, are novel EnsEMBL genes, or are based upon GeneScan prediction, and therefore have no association in GO. Thus, every peptide[14] or cDNA sequence is analyzed by InterProScan [73] to identify known protein domains from InterPro [119] and its affiliated protein databases (PROSITE [81], PRINTS [82], Pfam [80], ProDom [120], SMART [121] and TIGRFAMs [122]). To keep the standardized vocabulary, the IDs of the particular protein database are linked to the GeneOntology terms.

As listed in Table 3.19, 234 tags (out of 623 single-hit LongSAGE tags) could be directly associated to manually curated GeneOntology annotations. This number could be increased to 453 by extending the GeneOntology dataset with the approach described above. All GeneOntology terms with ten or more tags associated are listed in Table 3.20.

Table 3.20: Functional annotation of differentially expressed genes

GO term ID	name	number
GO:0003674	**molecular_function**	404
GO:0005488	binding activity	252
GO:0046872	metal ion binding activity	30

table continues on following page

[14]available only for EnsEMBL genes (not EST genes).

GO term ID	name	number
GO:0005509	calcium ion binding activity	19
GO:0046914	transition metal ion binding activity	8
GO:0003676	nucleic acid binding activity	119
GO:0003677	DNA binding activity	62
GO:0003700	transcription factor activity	26
GO:0003723	RNA binding activity	32
GO:0003729	mRNA binding activity	5
GO:0000166	nucleotide binding activity	76
GO:0030551	cyclic nucleotide binding activity	
GO:0017076	purine nucleotide binding activity	76
GO:0030554	adenyl nucleotide binding activity	72
GO:0019001	guanyl nucleotide binding activity	16
GO:0005515	protein binding activity	50
GO:0019956	chemokine binding activity	
GO:0008092	cytoskeletal protein binding activity	16
GO:0019838	growth factor binding activity	2
GO:0005102	receptor binding activity	12
GO:0005125	cytokine activity	4
GO:0008083	growth factor activity	8
GO:0005179	hormone activity	2
GO:0003824	enzyme activity	137
GO:0004386	helicase activity	3
GO:0016787	hydrolase activity	55
GO:0016853	isomerase activity	6
GO:0016301	kinase activity	18
GO:0008478	pyridoxal kinase activity	
GO:0016874	ligase activity	9
GO:0016491	oxidoreductase activity	25
GO:0008641	small protein activating enzyme activity	3
GO:0004839	ubiquitin activating enzyme activity	3
GO:0008642	ubiquitin-like activating enzyme activity	1
GO:0016740	transferase activity	37
GO:0005554	molecular_function unknown	21
GO:0004871	signal transducer activity	37
GO:0004872	receptor activity	21
GO:0004879	ligand-dependent nuclear receptor activity	1
GO:0004888	transmembrane receptor activity	9
GO:0005102	receptor binding activity	12
GO:0005125	cytokine activity	4
GO:0001664	G-protein-coupled receptor binding activity	
GO:0008083	growth factor activity	8
GO:0005179	hormone activity	2
GO:0005198	structural molecule activity	36
GO:0005201	extracellular matrix structural constituent	1
GO:0005200	structural constituent of cytoskeleton	5
GO:0003735	structural constituent of ribosome	20
GO:0030528	transcription regulator activity	32
GO:0003700	transcription factor activity	26

table continues on following page

GO term ID	name	number
GO:0003705	RNA polymerase II transcription factor activity, enhancer binding	
GO:0005215	transporter activity	60
GO:0005386	carrier activity	17
GO:0005489	electron transporter activity	14
GO:0015075	ion transporter activity	11
GO:0008565	protein transporter activity	12
GO:0008150	**biological_process**	334
GO:0000004	biological_process unknown	24
GO:0009987	cellular process	153
GO:0007154	cell communication	46
GO:0007155	cell adhesion	7
GO:0007165	signal transduction	37
GO:0007166	cell surface receptor linked signal transduction	8
GO:0007242	intracellular signaling cascade	25
GO:0008151	cell growth and/or maintenance	105
GO:0016049	cell growth	5
GO:0019725	cell homeostasis	1
GO:0016043	cell organization and biogenesis	22
GO:0008283	cell proliferation	25
GO:0007049	cell cycle	20
GO:0006810	transport	55
GO:0015031	protein transport	17
GO:0045045	secretory pathway	2
GO:0016192	vesicle-mediated transport	4
GO:0006928	cell motility	11
GO:0016477	cell migration	5
GO:0007275	development	38
GO:0030154	cell differentiation	9
GO:0009790	embryonic development	2
GO:0009653	morphogenesis	27
GO:0007389	pattern specification	8
GO:0007582	physiological processes	291
GO:0008151	cell growth and/or maintenance	105
GO:0016049	cell growth	5
GO:0016043	cell organization and biogenesis	22
GO:0008283	cell proliferation	25
GO:0007049	cell cycle	20
GO:0042127	regulation of cell proliferation	
GO:0006810	transport	55
GO:0008152	metabolism	219
GO:0006519	amino acid and derivative metabolism	9
GO:0009058	biosynthesis	56
GO:0006118	electron transport	23
GO:0006091	energy pathways	9
GO:0006139	nucleobase, nucleoside, nucleotide and nucleic acid metabolism	80
GO:0006259	DNA metabolism	10

table continues on following page

GO term ID	name	number
GO:0009117	nucleotide metabolism	6
GO:0016070	RNA metabolism	19
GO:0006350	transcription	50
GO:0006793	phosphorus metabolism	13
GO:0006796	phosphate metabolism	13
GO:0019538	protein metabolism	90
GO:0006412	protein biosynthesis	36
GO:0030163	protein catabolism	25
GO:0009605	response to external stimulus	17
GO:0006950	response to stress	10
GO:0005575	**cellular_component**	310
GO:0005623	cell	272
GO:0005622	intracellular	211
GO:0005694	chromosome	2
GO:0005737	cytoplasm	111
GO:0000153	cytoplasmic ubiquitin ligase complex	1
GO:0016023	cytoplasmic vesicle	1
GO:0005856	cytoskeleton	23
GO:0005783	endoplasmic reticulum	13
GO:0005794	Golgi apparatus	7
GO:0005739	mitochondrion	21
GO:0005840	ribosome	21
GO:0005634	nucleus	95
GO:0016363	nuclear matrix	
GO:0005635	nuclear membrane	4
GO:0005730	nucleolus	7
GO:0005654	nucleoplasm	15
GO:0005681	spliceosome complex	5
GO:0030529	ribonucleoprotein complex	32
GO:0030532	small nuclear ribonucleoprotein complex	
GO:0005732	small nucleolar ribonucleoprotein complex	2
GO:0005681	spliceosome complex	5
GO:0000153	cytoplasmic ubiquitin ligase complex	1
GO:0016020	membrane	90
GO:0012505	endomembrane system	6
GO:0000139	Golgi membrane	1
GO:0019866	inner membrane	7
GO:0005743	mitochondrial inner membrane	7
GO:0016021	integral to membrane	64
GO:0008372	cellular_component unknown	23
GO:0005576	extracellular	60
GO:0005578	extracellular matrix	6
GO:0005615	extracellular space	58

Ontologies with ten or more differentially expressed genes (plus a few interesting hand-selected ones). Degree of Indention represents the level within the GeneOntology hirarchy. Numbers assigned for each ontology include all ontologies lower in hierarchy.

3.4 Genome-wide analysis of publically available SAGE libraries

3.4.1 Chromosomal localization of genes regulated by signaling cascades or transcription factors

In the case of four publically available SAGE library pairs (listed in Table 3.21; including the ATDC5 libraries generated in this study), the chromosomal locations of the differentially expressed genes were determined[15]. The libraries were chosen, since they are derived from cell lines induced with a single stimulus. Therefore, by comparing the libraries to the untreated control, the effect of the corresponding factors could be monitored. As summarized in Table 3.22, in all four pairs of libraries a significant[16] number of tags were physically linked with a distance of less than 1 Mb (the genes within each interval are listed in Table 3.23). In any set of libraries even the very unlikely event[17], that two of the differentially expressed were immediate neighbors, occurred multiple times. Next, the pairs were analyzed for DNA binding sites downstream of its corresponding pathways (BMP: Smad3 and Smad4 binding sites; SHH: Gli binding site; JNK2: AP1 and CRE-binding protein1/c-JUN binding sites; c-MYC: c-MYC/MAX heterodimer binding site; for sequences and references, see Materials and methods). Since all binding sites are very short and therefore statistically occur every few thousand bases, only those binding sites are considered, that are conserved between mouse and human. Thus, for each chromosomal fragment its syntenic region in the other species was retrieved in a semi-automated way. The program written for this purpose first determined the syntenic genomic fragment based upon genomic DNA alignments. Both fragments were analyzed for the existence of putative orthologous genes with high ammino acid sequence similarity as well as conserved noncoding sequences (CNS) sharing at least 70 percent identity over at least 100 basepairs. The program in parallel generated pictures displaying all features of both genomic segments. Both were only considered as being syntenic if a reasonable number of orthologous genes was identified[18] between both species. For most of the gene clusters the syntenic region could be successfully retrieved (ATDC5: 4 out of 9 cases;

[15] As described above (3.1.5; since the raw sequence data was only available for the ATDC5 libraries, in the other cases the information about the $11^{(}th$) could not be used to refine the tag-to-UniGene mappings).

[16] as compared to 1000 simulations randomly picking the same number of genes.

[17] observed on average in less than one case per library within the simulations.

[18] Differences in the order of genes were accepted, since by definition the same order of genes is required for 'conserved segments' but not for 'conserved synteny' [83].

Table 3.21: SAGE libraries generated from cells induced with single factors

Factor	tags in control	tags in induced cells	cell kind	reference
BMP4	21,875[1]	21,781[2]	embryonal carcinoma-derived cell line, mouse	in press
Shh	87,837[3]	85,510[4]	primary granule cell precursor cells, mouse	not published
JNK2	38,819[1]	40,768[2]	PC3, human	not published
c-MYC	37,047[7]	55,426[7]	HUVEC, human	[123]

Publically available SAGE libraries used for analysis. Cells were either supplied with growth factors (BMP4, Shh) or factor was overexpressed (JNK2, c-MYC). [1]GSM2575. [2]GSM2576. [3]GSM787. [4]GSM788. [5]GSM1515. [6]GSM1514. [7]available from www.biochem.mpg.de/hermeking/mycsage.html.

SHH: 40/45; JNK2: 34/42; c-MYC: 16/24). Binding sites were only considered if they lie within CNS features. As shown in Table 3.23, some, but not all clusters contained at least one conserved DNA binding site within a CNS for its corresponding downstream transcription factors.

Table 3.22: Summary of physical linkeage

Factor	# Diff. expressed		# Clusters		Immed.	Simulated	
	total	assigned	≥ 2	≥ 3	neigh.[1]	≥ 2	≥ 3
BMP4	139	120	9	3	2	3.4 ± 1.7	0.1 ± 0.3
SHH	600	263	45	11	12	26.8 ± 4.1	3.0 ± 1.7
JNK2	562	258	42	19	7	23.3 ± 4.1	2.2 ± 1.4
cMYC	421	164	24	5	4	10.3 ± 2.9	0.6 ± 0.7

Number of clusters ≥ 1 Mb with two or three differentially expressed genes upon induction compared to 1000 random simulations. [1]Immediate neighbors: No other gene lied in between two differentially expressed genes.

Table 3.23: Physical linkeage of potential factor-regulated genes

Chr.	Genes and interlocus distances	# BS[1]	
BMP4		Smad3	Smad4
5	*Mor1* - 3 (0 partial) genes [118021 bp] - EST	3	0
6	*Tpi* - 0 (0 partial) genes [607 bp] - *Usp5*	0	0
7	*Rps11* - 0 (0 partial) genes [1185 bp] - *Rpl13a* - 58 (0 partial) genes [803396 bp] - *Emp3* - 28 (0 partial) genes [930277 bp] - *Ldh1*	n.a.	n.a.

table continues on following page

Chr.	Genes and interlocus distances	# BS[1]	
10	EST[1] - 50 (0 partial) genes [866328 bp] - *Cd63*	0	0
11	*Hint* - 9 (0 partial) genes [518954 bp] - *Sparc*	n.a.	n.a.
11	EST - 18 (0 partial) genes [204121 bp] - EST	n.a.	n.a.
17	*Rab11b* - 2 (0 partial) genes [63385 bp] - *Ndufa7* - 8 (0 partial) genes [116517 bp] - *Rps18*	n.a.	n.a.
X	*Plp2* - 25 (0 partial) genes [471900 bp] - *Rbm3*	0	0
X	*Filamin-like protein* - 2 (0 partial) genes [24580 bp] - *Rpl10*[1] - 36 (0 partial) genes [972587 bp] - EST	n.a.	n.a.
SHH		Gli	
1	*Bzw1* - 11 (0 partial) genes [515240 bp] - EST	0	
1	EST - 4 (0 partial) genes [487614 bp] - *Itm2c*	3	
1	EST - 0 (0 partial) genes [9629 bp] - *2010320B01Rik*	0	
1	*1110021H02Rik* - 24 (0 partial) genes [757376 bp] - *Nhlh1*	0	
2	EST - 1 (0 partial) genes [23249 bp] - EST	0	
2	EST - 2 (0 partial) genes [119944 bp] - *2810027O19Rik*	0	
2	EST - 6 (0 partial) genes [197146 bp] - *Csen* - 2 (0 partial) genes [65325 bp] - *Mrps5*	n.a.	
2	*5730494N06Rik* - 0 (0 partial) genes [1504 bp] - *Pcna*	0	
2	*1010001H21Rik* - 5 (0 partial) genes [212270 bp] - EST	0	
4	*Rps8* - 4 (0 partial) genes [408507 bp] - *Prnpip1*	0	
4	EST - 6 (0 partial) genes [301315 bp] - *2810449C13Rik*	0	
6	EST - 1 (0 partial) genes [112031 bp] - *2410127E18Rik*	0	
6	EST - 5 (0 partial) genes [957681 bp] - *Dfna5h*	0	
7	*2410022M24Rik* - 1 (0 partial) genes [99531 bp] - *Psip2*	0	
7	EST - 3 (0 partial) genes [36959 bp] - *Pold1* - 13 (0 partial) genes [312911 bp] - EST	0	
7	*9030624J02Rik* - 12 (0 partial) genes [932601 bp] - *6330575P11Rik*	0	
8	*Ris2* - 29 (0 partial) genes [835919 bp] - *Tubb3*	0	
9	*2010004J23Rik* - 12 (0 partial) genes [924849 bp] - *Nope* - 12 (0 partial) genes [427101 bp] - *AI840980* - 4 (0 partial) genes [311308 bp] - *2810417H13Rik*	0	
9	*Ccnb2* - 3 (0 partial) genes [258783 bp] - EST	0	
9	*Mapk6* - 1 (0 partial) genes [43758 bp] - EST	0	
9	*Gnai2* - 44 (0 partial) genes [956408 bp] - *Arih2*	0	
10	*1810010L20Rik* - 20 (0 partial) genes [799916 bp] - *Nnp1*	0	
11	EST - 9 (0 partial) genes [330004 bp] - *Pold2*	n.a.	
11	*Sqstm1* - 5 (0 partial) genes [83679 bp] - *Canx* - 2 (0 partial) genes [52178 bp] - *Hnrph1*	0	
11	*Gps2* - 6 (0 partial) genes [74782 bp] - *Gabarap*	0	
11	*Zfp144* - 12 (0 partial) genes [327475 bp] - EST	1	
11	*Tk1* - 0 (0 partial) genes [23298 bp] - *Birc5*	0	

table continues on following page

Chr.	Genes and interlocus distances	# BS[1]	
11	EST - 16 (0 partial) genes [229610 bp] - EST - 37 (0 partial) genes [746840 bp] - *0610008N23Rik*	0	
12	*Bag5* - 7 (0 partial) genes [248335 bp] - *2010107E04Rik* - 9 (0 partial) genes [680907 bp] - *Siva-pending* - 2 (0 partial) genes [75605 bp] - EST	0	
13	*Trim27* - 27 (0 partial) genes [624606 bp] - EST	0	
15	EST - 1 (0 partial) genes [111671 bp] - *Pabpc1*	0	
15	*Tuba3* - 12 (0 partial) genes [467232 bp] - *Tegt*	0	
15	*Pfdn5* - 0 (0 partial) genes [231 bp] - *Myg1-pending*	1	
16	EST - 11 (0 partial) genes [347583 bp] - *Nude-pending*	0	
16	EST - 3 (0 partial) genes [62328 bp] - EST	0	
17	*Tulp4* - 10 (0 partial) genes [694793 bp] - *1110008A10Rik*	n.a.	
17	*Pkmyt1-pending* - 20 (0 partial) genes [428440 bp] - EST - 30 (0 partial) genes [549453 bp] - EST	0	
17	EST - 6 (0 partial) genes [294755 bp] - EST	1	
17	EST - 47 (0 partial) genes [854831 bp] - EST - 5 (0 partial) genes [61899 bp] - *0610011P08Rik*	3	
19	*Prdx5* - 14 (0 partial) genes [288361 bp] - *AI850305* - 20 (0 partial) genes [939980 bp] - *Men1* - 38 (0 partial) genes [734691 bp] - *1500026D16Rik* - 16 (0 partial) genes [217164 bp] - *Sart1*		
19	*Rad9* - 0 (0 partial) genes [-221 bp] - *Ppp1ca*	0	
19	*Gng3lg* - 12 (0 partial) genes [118538 bp] - *2610301D06Rik* - 12 (0 partial) genes [865131 bp] - *Fth* - 14 (0 partial) genes [547899 bp] - *2810441K11Rik*	n.a.	
19	*Ldb1* - 28 (0 partial) genes [984661 bp] - *Ina*	1	
19	*Xpnpep1* - 0 (0 partial) genes [100070 bp] - *Add3*	0	
X	*Dlgh3* - 10 (0 partial) genes [613974 bp] - *Nono* - 19 (0 partial) genes [754125 bp] - EST	n.a.	
JNK2		AP1	CRE/ cJUN[2]
1	*MRPL20* - 17 (0 partial) genes [718931 bp] - EST	0	0
1	EST - 4 (0 partial) genes [387247 bp] - *SFN* - 21 (0 partial) genes [802127 bp] - *G1P3*	0	0
1	*PTP4A2* - 12 (0 partial) genes [395486 bp] - *MLP* - 9 (0 partial) genes [439748 bp] - *YARS*	0	0
1	*MUC1* - 33 (0 partial) genes [922045 bp] - *LMNA* - 8 (0 partial) genes [169647 bp] - *CCT3* - 10 (0 partial) genes [331048 bp] - *NES*	0	0
3	*IMPDH2* - 29 (0 partial) genes [857619 bp] - *MST1R* - 6 (0 partial) genes [323099 bp] - *GNAI2*	0	0
3	EST - 13 (0 partial) genes [685228 bp] - *TKT*	0	0
3	EST - 0 (0 partial) genes [91719 bp] - *SFRS10*	4	4

table continues on following page

Chr.	Genes and interlocus distances	# BS[1]	
6	*ABCF1* - 25 (0 partial) genes [676038 bp] - *HLA-C* - 13 (0 partial) genes [263657 bp] - *BAT2* - 21 (0 partial) genes [197344 bp] - *C6orf48*	n.a.	n.a.
6	*TAPBP* - 17 (0 partial) genes [922629 bp] - *HMGA1*	n.a.	n.a
7	*FSCN1* - 7 (0 partial) genes [415594 bp] - EST	4	4
7	EST - 9 (0 partial) genes [825025 bp] - *GTF2I*	0	0
8	EST - 0 (0 partial) genes [1405 bp] - *EEF1D* - 9 (0 partial) genes [194357 bp] - EST - 26 (0 partial) genes [513832 bp] - EST - 20 (0 partial) genes [430220 bp] - *RPL8*	n.a.	n.a.
9	EST - 4 (0 partial) genes [218619 bp] - *CLTA*	1	1
10	*CUL2* - 2 (0 partial) genes [246385 bp] - EST	n.a.	n.a.
10	*SEC24C* - 2 (0 partial) genes [10858 bp] - EST - 5 (0 partial) genes [29312 bp] - *CAMK2G* - 0 (0 partial) genes [36567 bp] - *PLAU*	8	8
10	*KCNMA1* - 4 (0 partial) genes [396223 bp] - *RPS24*	43	39
11	*POLR2L* - 0 (0 partial) genes [886 bp] - *CD151* - 15 (0 partial) genes [222177 bp] - *IRF7* - 5 (0 partial) genes [77273 bp] - *HRAS* - 7 (0 partial) genes [215745 bp] - *IFITM1* - 6 (0 partial) genes [98805 bp] - EST	0	0
11	EST - 13 (0 partial) genes [497728 bp] - EST - 11 (0 partial) genes [430404 bp] - *FEN1* - 5 (0 partial) genes [167321 bp] - *FTH1*	0	0
11	EST - 39 (0 partial) genes [810507 bp] - *CFL1* - 5 (0 partial) genes [35979 bp] - *FOSL1*	5	4
12	*TPI1* - 18 (0 partial) genes [843456 bp] - *APOBEC1*	n.a.	n.a.
12	*PRKAG1* - 4 (0 partial) genes [114672 bp] - *TUBA1* - 1 (0 partial) genes [133687 bp] - EST	1	1
12	*NACA* - 23 (0 partial) genes [760893 bp] - *MARS*	n.a.	n.a.
14	EST - 16 (0 partial) genes [664542 bp] - *PSME2* - 4 (0 partial) genes [42499 bp] - *TM9SF1*	10	10
14	*RPS29* - 6 (0 partial) genes [181251 bp] - *KLHDC2*	2	2
15	*NOLA3* - 4 (0 partial) genes [145848 bp] - EST	0	0
16	*TCEB2* - 13 (0 partial) genes [243568 bp] - *TNFRSF12A*	1	1
16	*ALDOA* - 11 (0 partial) genes [213238 bp] - EST	1	1
17	*ETV4* - 0 (0 partial) genes [94055 bp] - *MEOX1*	4	4
17	EST - 9 (0 partial) genes [475002 bp] - *PHB*	0	0
17	EST - 2 (0 partial) genes [61960 bp] - *PSMC5*	0	0
17	EST - 1 (0 partial) genes [57569 bp] - EST - 34 (0 partial) genes [945000 bp] - *ITGB4* - 32 (0 partial) genes [976296 bp] - *SFRS2*	16	13
17	*LGALS3BP* - 0 (0 partial) genes [11737 bp] - EST	3	3
19	*BSG* - 8 (0 partial) genes [213918 bp] - *PTBP1* - 31 (0 partial) genes [595257 bp] - *DAZAP1*	6	5

table continues on following page

Chr.	Genes and interlocus distances	# BS[1]	
19	*DPP9* - 16 (0 partial) genes [966492 bp] - *RPL36* - 10 (0 partial) genes [203075 bp] - EST	2	1
19	*PIN1* - 4 (0 partial) genes [236607 bp] - EST - 13 (0 partial) genes [297879 bp] - *CDC37* - 9 (0 partial) genes [250788 bp] - *ILF3* - 21 (0 partial) genes [743844 bp] - *PRKCSH*	7	4
19	EST - 8 (0 partial) genes [120140 bp] - *JUNB* - 0 (0 partial) genes [3509 bp] - *PRDX2* - 14 (0 partial) genes [303058 bp] - EST - 2 (0 partial) genes [36672 bp] - EST - 16 (0 partial) genes [964621 bp] - *ASF1B*	0	0
19	*CHERP* - 23 (0 partial) genes [860539 bp] - *BST2*	n.a.	n.a.
19	*C20orf109* - 26 (0 partial) genes [940590 bp] - EST - 3 (0 partial) genes [101278 bp] - EST	0	0
19	*FXYD5* - 19 (0 partial) genes [478370 bp] - *COX6B* - 28 (0 partial) genes [481231 bp] - *CAPNS1*	0	0
19	*PPP1R15A* - 41 (0 partial) genes [679654 bp] - *NOSIP* - 18 (0 partial) genes [348646 bp] - *ATF5* - 27 (0 partial) genes [863770 bp] - EST - 22 (0 partial) genes [542912 bp] - *ETFB*	3	3
20	*NTSR1* - 18 (0 partial) genes [725528 bp] - *EEF1A2* - 19 (0 partial) genes [371531 bp] - *TPD52L2*	18	18
X	*CETN2* - 18 (0 partial) genes [804260 bp] - EST - 14 (0 partial) genes [286077 bp] - *IRAK1* - 10 (0 partial) genes [291555 bp] - *FLNA* - 24 (0 partial) genes [391037 bp] - *DKC1*	n.a.	n.a.
cMYC		cMYC/ MAX[3]	
1	*TIE* - 11 (0 partial) genes [612875 bp] - EST	0	
1	*S100A10* - 0 (0 partial) genes [38674 bp] - *S100A11*	0	
1	*S100A6* - 17 (0 partial) genes [386257 bp] - *JTB*	0	
1	*ARHGEF2* - 13 (0 partial) genes [314617 bp] - EST - 0 (0 partial) genes [13292 bp] - *CCT3*	0	
2	*TMSB10* - 12 (0 partial) genes [677731 bp] - *VAMP5*	0	
2	EST - 9 (0 partial) genes [510701 bp] - *CNNM3*	n.a.	
2	EST - 0 (0 partial) genes [-16457 bp] - EST	0	
3	EST - 28 (0 partial) genes [790454 bp] - *STAB1*	n.a.	
6	*HLA-C* - 26 (0 partial) genes [369945 bp] - *DDAH2* - 0 (0 partial) genes [324 bp] - *CLIC1* - 5 (0 partial) genes [79034 bp] - *HSPA1B*	0	
7	*FSCN1* - 18 (0 partial) genes [970797 bp] - *ZDHHC4*	0	
7	*MDH2* - 3 (0 partial) genes [235992 bp] - *HSPB1*	0	
7	EST - 3 (0 partial) genes [258997 bp] - EST	6	
8	*PLEC1* - 42 (0 partial) genes [794762 bp] - *RPL8*	n.a.	
9	*UBE2R2* - 15 (0 partial) genes [695560 bp] - *DCTN3*	n.a.	
11	*SLC25A22* - 16 (0 partial) genes [256579 bp] - *HRAS*	0	
12	*CD9* - 25 (0 partial) genes [528445 bp] - *PTMS*	0	

table continues on following page

Chr.	Genes and interlocus distances	# BS[1]
14	*APEX1* - 11 (0 partial) genes [343594 bp] - *RNASE1*	n.a.
17	*PFN1* - 14 (0 partial) genes [484389 bp] - *C1QBP*	0
17	*ICAM2* - 8 (0 partial) genes [488357 bp] - *DDX5*	6
17	*ACTG1* - 6 (0 partial) genes [175914 bp] - *MRPL12* - 2 (0 partial) genes [102313 bp] - *P4HB* - 8 (0 partial) genes [71870 bp] - *PYCR1* - 11 (0 partial) genes [142049 bp] - *FASN*	4
19	EST - 28 (0 partial) genes [584282 bp] - *GPX4* - 24 (0 partial) genes [490379 bp] - EST	5
19	*SH3GL1* - 7 (0 partial) genes [175609 bp] - *SEMA6B* - 2 (0 partial) genes [116729 bp] - *DPP9*	2
19	*CDC42EP5* - 33 (0 partial) genes [911396 bp] - *RPL28*	n.a.
X	*BGN* - 10 (0 partial) genes [283967 bp] - *SSR4*	2

Number of genes as well as length of interval between two or more differentially expressed genes is given between gene symbols. Negative values for distance means, that the genes overlap. [1]number of conserved binding sites between mouse and human. [2]CRE-binding protein 1/cJUN heterodimer. [3]cMYC/MAX heterodimer

3.4.2 House-keeping genes

All publically available human and mouse SAGE libraries with at least 50.000 tags[19] sequenced were adducted for a search for genes similarly expressed within all libraries. All together, there were only 203 (out of 69 libraries) human and 719 (out of 17 libraries) mouse SAGE tags detected in all libraries. The top ten tags of mouse and human with the smallest changes (highest mean to standard deviation ratio) are listed in Table 3.24.

3.4.3 Ribosomal protein gene expression

To construe the observed changes of gene expression of ribosomal protein genes, all publically available mouse and human SAGE libraries with at least 50,000 tags were mined. For a total of 80 human ribosomal protein genes [90, 91] and 54 human mitochondrial ribosomal genes [92], which were mapped to the genome, the corresponding human entry in EnsEMBL as well as the syntenic mouse entry were retrieved. Due to the fact, that for most ribosomal protein genes many pseudogenes exist in both genomes, this strategy was needed to eliminate falsely annotated ribosomal protein genes. For each gene the associated SAGE tag(s) were assigned through the corresponding UniGene clusters (see Materials and methods). The data was normalized by calculating the tags-per-million value of all SAGE tags corresponding to a single ribosomal protein gene.

Next, the expression profile of all human (GEO/ own identifier is preceded

[19]to minimize statistical fluctuations, libraries with less tags were not adducted.

Table 3.24: Genes with most constant expression levels over all SAGE libraries

tag	UniGene ID	Description	mean
human			
GCCTGCTGGG	2706	GPX4: Glutathione peroxidase 4 (phospholipid hydroperoxidase)	517.7±207.5
TGGAGTGGAG		multiple hits(3)	320.9±131.1
CGCTGGTTCC		multiple hits(2)	1220.5±500.3
TCACAAGCAA	32916	NACA nascent-polypeptide-associated complex alpha polypeptide	465.2±200.3
CCTATTTACT		multiple hits(2)	387.0±167.11
ACAGTGGGGA		multiple hits(3)	287.3±124.5
GCTTCCATCT		multiple hits(4)	168.1±73.0
GTGACCTCCT	433901	COX8 cytochrome c oxidase subunit VIII	446.6±193.8
CCCTGATTTT	183684	EIF4G2 eukaryotic translation initiation factor 4 gamma, 2	189.5±82.8
GTTCCCTGGC		multiple hits(2)	591.3±260.8
mouse			
GCTGCCAGGG	688	Bcl2-associated athanogene 1	235.7±50.2
CATTGCGTGG	27955	Williams-Beuren syndrome chromosome region 1 homolog (human)	300.7±79.1
CTCCTGCAGC	38055	esterase 10	222.0±60.5
GGAGGGATCA	8131	integrin linked kinase	181.9±50.5
GCTGGCAGCC		multiple hits(2)	589.8±165.3
GAGGGCATCC	20946	proteasome (prosome, macropain) 26S subunit, ATPase 3	203.3±57.2
GGGTGCGTCT	196604	angio-associated migratory protein	163.1±46.3
TGCTGCTCGT		multiple hits(2)	212.0±60.7
GGGGTGACAG	22040	expressed sequence AW556797	46.9±13.6
AACAATTTGG		multiple hits(2)	1406.1±411.4

Top ten human and mouse tags with the smallest expression changes over all SAGE libraries. Mean and standard deviations were calculated for tpm values.

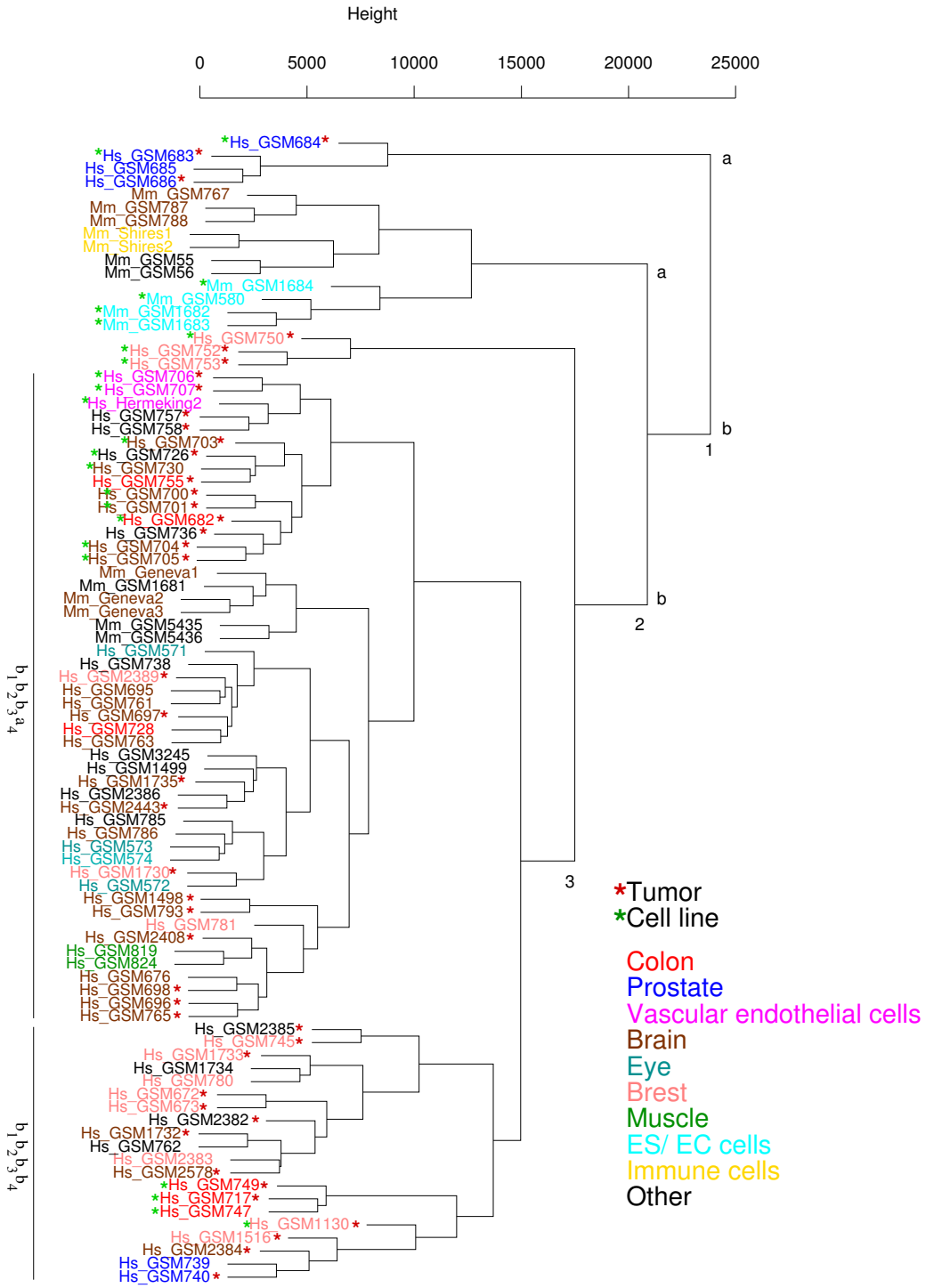


Figure 3.11: human and mouse SAGE libraries clustered (hierarchical clustering, euclidian distance) according to expression of ribosomal protein genes. All libraries derived from the ncbi Gene Expression Omnibus are indicated by the GEO accession number starting with GSM; for the remaining libraries, see Table 3.9. To label a particular clusters, each branching event is specified (a: top, b: bottom, subscript indicates the number of branching [from right to left]; as an example, two clusters are specified with the black bars).

by Hs_) and mouse (Mm_) SAGE libraries adducted to hierarchical clustering. As shown in Figure 3.11, the libraries could be separated into several distinct clusters, with the first branching point at a height of around 23,000 to 24,000. There is a tendency, that libraries generated from the same major tissue fall into the same cluster (e.g. cluster $b_1b_2b_3a_4$ consists mainly of libraries derived from brain and brain-derived [eye] tissues whereas cluster $b_1b_2b_3b_4$ contains almost no brain libraries). Furthermore, libraries generated from cell lines (indicated by green asterisk in Figure 3.11) tend to separate at a lower level from its counterparts derived from tissues. For example, within cluster $b_1b_2b_3a_4$, all cell-line derived SAGE libraries are exclusively present in sub-cluster a, but none in sub-cluster b. On the other hand, libraries derived from cancer cell lines or tissues (green asterisk) are more randomly distributed.

All mouse libraries were grouped into two separated clusters. Whereas 11 of the mouse libraries branch at the second node at a height of around 20,000, six of the mouse libraries, out of which three were derived of adult brain tissues (Mm_Geneva1 to 3), are included in the large (brain-enriched) cluster $b_1b_2b_3a_4$, and do not branch before a height of around 8000. Another interesting observation was, that most of the library pairs of the kind of uninduced vs. induced or normal vs. cancer are very similar, even though in most of the cases differentially expressed ribosomal protein genes were observed between both libraries.

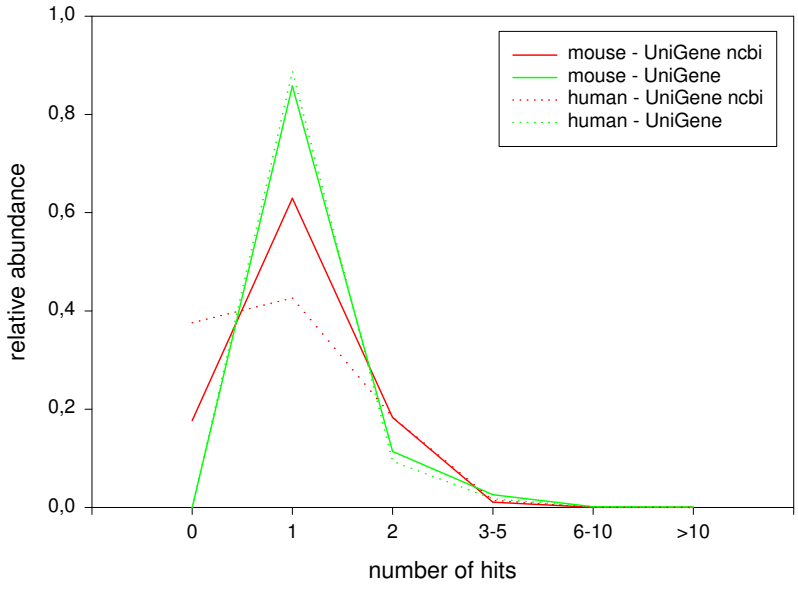
Chapter 4

Discussion

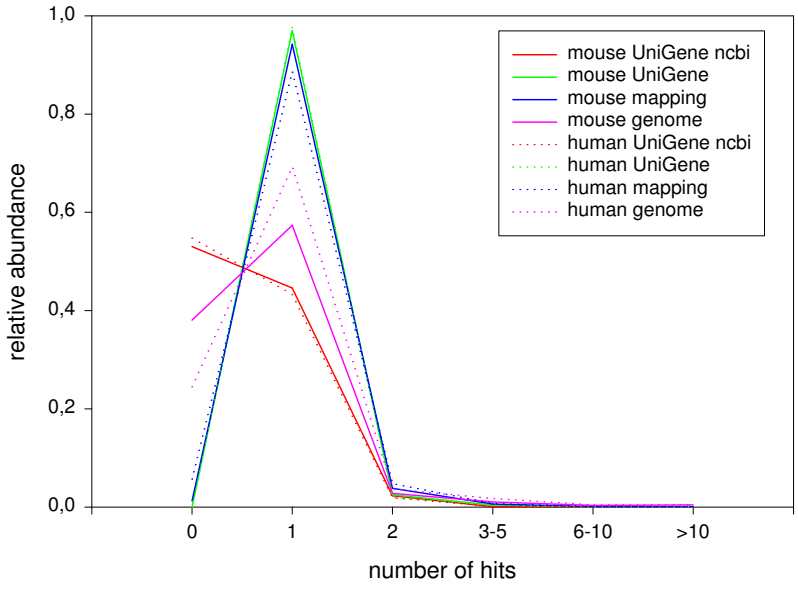
4.1 SAGE mapping

Compared to the commonly used SAGEmap database [94, 68], the algorithms developed in this work significantly improved the mapping of a SAGE tag to its corresponding gene via UniGene, by decreasing both no-hit and multiple-hit cases (Fig 4.1). This is mainly due to the fact, that for the SAGEmap project no pre-selection for 3' sequences is made. The critical point is, that due to alternative splicing or alternative polyadenylation multiple SAGE tags are possible for a single gene/ UniGene cluster. Based upon EST data mining, at least 59 % of Human and 33 % of Mouse genes have alternative splice forms [124] and 28.6 % of human genes show alternative polyadenylation [125]. Therefore even the large set of full-length cDNA sequences in mouse (60,777 RIKEN full-length clones for mouse [4]) does not cover all genes and its associated transcripts. In deed, a global survey showed that for 44% of human and 22% of both mouse and rat genes have different SAGE tags due to alternative polyadenylation [126]. The difference becomes clear, if a hypothetical gene is assumed, which has a rare transcript with alternative polyadenylation, present in only 15% of all transcripts. Statistically this transcript should be represented in 15% of the 3' reads, but is indistinguishable in the 5' reads. Therefore, without a pre-selection of 3' sequences, the difference could only be detected in 7.5% of all sequences. Since SAGEmap similarly discards the 10% of the most infrequent tags for each UniGene cluster [94], this particular SAGEtag would be left out in the SAGEmap mappings, but would be detected in the algorithm proposed in this work, since its abundance in the 3' reads lies over 10%.

In addition, for the first time, a strategy for assigning LongSAGE tags to the genome as well as for combining both genome associations with UniGene



(a) SAGE mapping



(b) LongSAGE mapping

Figure 4.1: Relative abundance of zero, one, two three to five, six to ten and larger ten hits per SAGE and LongSAGE tag for different mappings. Mapping: Combination of mapping against UniGene (transcript) and EnsEMBL (genome).

mappings was described in this work. Surprisingly, a large fraction of reliable LongSAGE tags derived from the UniGene releases as well as many of the experimentally extracted LongSAGE tags in this study (similar tendency for human LongSAGE tags, M. Souquet, personal communication) can not be assigned to the mouse or human genome. This could be explained in that even at the current state of genome sequencing the assembled sequence is not complete. Still there are gaps within the assembled sequence and also not all sequence contigs could be mapped to the assembled sequence, mainly in pericentromeric and subtelomeric regions [127]. Furthermore, even a 7-fold coverage by shotgun sequencing, some parts of the genome are covered only with one sequencing read, and are therefore more prone to sequencing errors. Since the human genome sequencing is more advanced and since the murine Y chromosome has not yet been sequenced, it is reasonable that the number of no-hits of mouse LongSAGE tags is higher than for human. On the other hand, the mouse genome was sequenced from the same strain (C57/BL6) as the LongSAGE tags, and therefore polymorphisms, that could account for no-hits to the human genome, can be excluded.

That several LongSAGE tags could not be assigned to its corresponding EnsEMBL gene is due to an artefact in the EnsEMBL gene annotation pipeline [63][1]. As a result of the exon-intron structure of eukaryotic genes, alignments of transcript sequences to its genomic counterpart contain gaps of sometimes up to several tens of kilobases, whereby the score of the alignment often is lower than that to other homologous genes. Therefore genes are only annotated to the genome either based upon homology to other species or by aligning ESTs to computational gene predictions [63, 2, 1]. However, untranscribed regions of genes are less conserved than coding regions (between human and mouse: 84.7% for coding regions compared to 74.7% for 3' UTRs [1]), and gene prediction programs have shortcomings in determining UTRs [96]. Thus, most of the EnsEMBL genes lack at least parts of the 3' UTR, in which the LongSAGE tag could reside. Furthermore pseudogenes, which share a very high sequence homology to its active counterparts but are not described [128], interfere with the correct identification of the genomic locus of a gene. Therefore it is estimated that 76% (of a total of 2,700) of the annotated mouse genes lacking a equivalent in a syntenic human interval and 30% (of a total of 5,143) of genes, that are members of local gene clusters and lack a reciprocal best match in the human genome, correspond to pseudogenes [1].

[1] the EnsEMBL genome annotations contain a redundant set of genes, which are solely annotated based upon protein sequences from the same or other species, and EST genes, which are based upon GeneScan predictions verified by cDNA and EST alignments.

Furthermore two very powerful strategies to annotate the genome based upon LongSAGE tags were proposed in this study. As described above, transcript sequences hardly can be assigned to the genome based upon sequence homology programs. But if the chromosomal position was uniquely defined by the LongSAGE tag, all ESTs aligned to this positions could be used to annotate the gene. In addition to 1827 newly annotated genes in this approach, 2348 additional genes could be annotated to the genome based upon GeneScan prediction supported by LongSAGE tags. GeneScan prediction are a very powerful tool in identifying cDNA sequences of known genes, but more than two-thirds of GeneScan prediction to Human and Mouse genome without any additional proof are false-positives [129]. Interestingly, also 546 antisense genes were identified.

4.2 SAGE of Chondrogenesis

This is the first report on a transcriptome analysis in early chondrogenesis triggered by BMP signaling. By investigating a total of 43,656 tags derived from both the uninduced and induced SAGE libraries, 17,166 different transcripts were identified. Of these, 139 transcripts are predicted to be differentially expressed upon BMP4 induction. The success rate of verification for the changes predicted by SAGE is 73% in the case of unique tag-to-gene assignment, and is comparable to those in other SAGE studies [130, 102, 131]. These differentially expressed transcripts can be regarded as candidates for genes regulated by BMP signaling. Indeed, whole-mount *in situ* hybridization analysis of selected genes has revealed, that their expression patterns are largely overlapping to that of *Bmp4* in mouse embryos at E10.5 (Figure 3.7), suggesting that these genes are under the control of BMP signaling. Table 4.1 lists the 77 known genes out of the 139 predicted genes, sorted by functional groups, based on our own literature search for their functions (for details and used references, see Table B.1 in the appendix). The products of the listed genes function in a variety of cellular processes including transcriptional regulation, protein metabolism (biosynthesis, folding, transport and degradation), vesicle functions (mineralization and transport), general metabolism, cell signaling, and cell adhesion. In some cases, differential expression during chondrogenesis is expected or reasonable. *Fn1* (tag U94) and *Sparc* (tag U90) are upregulated in ATDC5 cells treated with BMP4, consistent with the previous reports that their expression is enhanced in areas undergoing chondrogenesis [132, 133]. Downregulation of *Actg* (tag D31) and *Vim* (tag D2) is reasonable since alteration of cellular morphology is closely associated with the changes in the cytoskeletal organization during

Table 4.1: Functional classification of predicted genes

Transcriptional regulation		**Signaling**		**Vesicles**	
DNA binding		**Extracellular**		**Vesicle mineralization**	
D64	Calr	D54	Cxcl12	D68	Anxa5
D67	Nca	U77	Ptn	**Vesicle transport**	
Inhibitor of DNA binding		**Transmembrane**		D53	Rab11b
U106	Idb3	D13	Sdc2	D68	Anxa5
U126	Idb2	D19	Tm4sf8	U110	Vcp
DNA packaging		D42	Gas1	U133	Sara
D33	Ptma	D50	Emp3	U139	Shfdg1
D61	Hmgn1	D62	Itgp	**Metabolism/homeostasis**	
		U125	Cd63	D12	Ftl1
Protein metabilism		**Intracellular**		D39	Ndufa7
Protein synthesis		D32	Tpt1	D71	Mor1
	ribosomal proteins*	D38	Ppp2cb	U81	Cox6c
D5	Eef1a1	D40	Lag	U89	Eno1
U91	Sui-rs1	D47	Gnai2	U93	Ldh1
U123	Eif4g2	D49	S100a6	U96	Idh2
Protein folding		D64	Calr	U111	Atp6g1
U80	Serpinh1	D66	Hint	U127	Tpi
U134	Hspa5	D68	Anxa5	U136	Pla2g4a
Protein transport/ sorting		U99	Ywhae		
D67	Nca	U118	Ywhag	**Structural**	
U116	Sec61g	U129	Btg1	**Components of ECM**	
Protein degradation		U135	Bnip2	D28	Col1a2
D38	Ppp2cb	U136	Pla2g4a	U90	Sparc
D44	Psmb4			U92	Osf2
D51	Psmb1	**Others or unknown**		U94	Fn1
D57	Psmb3	D8	Rbm3	U103	Tgfbi
U110	Vcp	D27	H2afx	U108	Col3a1
U116	Sec61g	D56	Plp2	**Cell adhesion**	
U121	Usp5	D66	Hint	D13	Sdc2
		D70	filamin-like protein	**Cytoskeleton**	
		U107	Fin14	D31	Actg
				D21	Ptmb10
				Associated with cytoskeleton	
				D32	Tpt1
				D40	Lag

Tag numbers correspond to those in Table 3.8. The prefixes for Tag numbers, D or U, indicate predicted downregulation or upregulation, respectively, as in Table 3.8.

the early stage of chondrogenesis [134]. *Tgfbi* (tag U103) is known to be induced by TGF-β signaling [135], and the product of this gene is implicated in cartilage formation as a collagen-binding protein, with a high level of expression in prehypertrophic chondrocytes [136]. It has been reported that stimulation of cells by TGF-β/BMP factors leads to induction of *Idb* genes including *Idb2* (tag U126) and *Idb3* (tag U106) as direct targets [137, 138]. *Osf2* (tag U92) has been shown to be a TGFβ/BMP-inducible gene mediated by upregulation of twist in osteoblastic cell line [139]. Downregulation of *Gas1* (tag D42), encoding a Wnt-inducible, Shh-binding protein [140], is consistent with the recent finding that this downregulation is required for chondrogenic precursor cells to be recruited into forming cartilage nodules [141].

Remarkably, in the rest of the cases, the listed genes have never been implicated either in chondrogenic differentiation or in BMP signaling. Thus, our SAGE analysis may provide new and wider insights into molecular and cellular mechanisms that are controlled by BMP signaling during chondrogenesis. In general, genes in one or related functional groups are not regulated only in one direction. Rather, it appears that some components in the same functional group are upregulated, while others are downregulated. As a net effect, function of a complex or a pathway may be modulated by BMP signaling. The most remarkable case is seen in the ribosomal protein genes (14 downregulated and one upregulated, see Table 3.8). Changes in expression of ribosomal protein genes have also been reported in other global expression studies. For example, overexpression of nMYC results in a characteristic change in expression of a number of ribosomal protein genes, thereby leading to a notion that nMYC controls ribosome biogenesis and protein synthesis [142]. Another discrete change in expression of a subset of ribosomal protein genes is reported in the brain transcriptome analysis of Down syndrome model mice [102]. Indeed, in a survey on public gene expression databases, it could be realized that in many cases expression levels of ribosomal protein genes are, more or less, changed. Thus, modulation of the ribosome function might be a general strategy to control cellular statuses in a variety of biological contexts. Alternatively, but not exclusively, the changes in expression of ribosomal protein genes may reflect extra-ribosomal functions of some of them [143], as discussed by Chrast and colleagues [102]. In any case, it is intriguing to examine whether BMP signaling exerts its effects in part via modulation of protein biosynthesis. It should be noted in this context that a translation elongation factor, *Eef1a1* (tag D5), and a translation initiation factor, *Eif4g2* (tag U123), are included in our list. In gastrulating Xenopus embryos, BMP4 upregulates a translation initiation factor, *eIF-4aIII*, and this upregulation is causally related to epidermal induction and inhibition of

neural fate [144]. It has also been shown that the action of BMP signaling in the inhibition of neurogenesis is due to selective proteolysis of Mash1 [145]. In this regard, it is interesting that our list contains seven genes encoding products that are implicated in protein degradation (Table 4.1). Specific proteasomal degradation of the Smad1/5 proteins via activation of *Smurf1* by BMP signaling plays a critical role in the dorsoventral patterning of Xenopus embryos [146]. Thus, our results suggest the possibility that the effects of BMP signaling on chondrogenic differentiation is also exerted via control on protein synthesis on one hand and via (specific) protein degradation on the other hand. The genes listed in the group of 'Signaling' in Table 4.1 are implicated in a variety of signaling pathways, but no obvious connection to BMP signaling can be made under the current status of knowledge. The predicted differential expression in these genes might reflect the fact that BMP signaling interact with or modulate other signaling pathways, like those of Wnt, Ca2+/Calmodulin, Erk-MAPK and JAK-STAT (reviewed in [21, 23]).

In addition to the potentially BMP-regulated genes, our SAGE analysis identified 190 ATDC5-specific transcripts as compared to seven other mouse SAGE libraries by virtual subtraction (Table 3.10 and 3.10). As already exemplified with the top 20 of the ATDC5-specific transcripts listed in Table 3.10, the majority of them are 'no-hit' tags. These 'no-hit' tags may represent novel genes that preferentially function in chondrogenic cells.

In conclusion, the present transcriptome analysis of BMP-induced chondrogenesis has provided several lines of new, unexpected findings. The results suggest that BMP signaling affects diverse cellular functions within a short period of time by controlling expression of a number of known and uncharacterized genes, as well as potentially novel genes. Further study will clarify how these concomitant changes in gene expression are brought about and are organized into the concerted cellular event of chondrogenic differentiation.

4.3 LongSAGE of somitogenesis

This study presents for the first time a comprehensive analysis of somitogenesis by analyzing different subsets of tissues involved in the process. Within a total number of 171,639 LongSAGE tags derived from the tail bud (tissue A), the posterior 2/3 of the PSM (B), the anterior 1/3 of the PSM (C) and the two pairs of nascent somites (D), 1007 transcripts were identified that show statistically different expression profiles between at least two of the libraries. Since the whole dataset generated from the LongSAGE study is too complex to be discussed in full detail, the following paragraphs will only deal with two different aspects. First it will be analyzed, how our knowledge on

the roles of the FGF, Wnt and Delta/ Notch signaling pathways could be extended by the LongSAGE prediction. Afterwards it will be tried to associate the observed morphological and cellular changes during the transition of mesenchymal presomitic mesoderm to epithelial somites with gene expression changes. The discussion will not be restricted to differentially expressed genes, but also those genes, that can be detected at a reasonable level in the dataset (count of three or higher[2]) will be utilized. The inclusion of the latter genes is valid, since unlike microarrays, SAGE counts do not interfere with background noise, but supply quantitative expression data. Therefore those genes are in deed expressed in the tissues analyzed, although no definite predictions can be made, whether they are differentially expressed within the subsets.

FGF, Wnt and Notch signaling pathways In vertebrates there are at least 22 fibroblast growth factors (FGFs) and four FGF receptors, for which several alternatively spliced isoforms are known (reviewed in [147]). In concordance with the SAGE data, *Fgf8* is expressed in a anterior-posterior gradient with the highest expression in the caudal tail tip [106]. The fine-tuning of this *Fgf8* gradient is crucial, since ectopic FGF8 leads to the absence of or to smaller somites, whereas lack of FGF8 results in larger somites. It is believed, that the high levels of FGF8 in the caudal part of the PSM keeps cells in a immature state, whereas once cells reach a position below a FGF8 threshold at the cranial part of the PSM, the segmentation program can be launched [148]. Among the FGF receptors, only *Fgfr1* is expressed in the presomitic mesoderm [149, 150], which could be reproduced with the SAGE analysis. Lack of *Fgfr1* impairs the formation of somites [151], and a hypermorphic allele of *Fgfr1* causes smaller somites [152]. Since other FGFs and FGF receptors are either not detected in the LongSAGE data or its expression level is statistically not evaluable, they are therefore either not expressed or at a lover level than *Fgf8* or *Fgfr1*, suggesting they play no or a minor role during somitogenesis. The downstream events of the FGF signaling cascade during somitogenesis are little known. As FGF receptors are receptor tyrosine kinases, FGF signaling is implicated to the RAS-MAPK cascade, albeit other signaling pathways are known to mediate FGF-dependent events (reviewed in [153, 154]). There is experimental evidence that the RAS-MAPK pathway is activated during the formation of somites, since antibodies against double

[2]At the PCR step of the library construction or during sequencing, errors could occur. However, it is extremely unlikely, that exactly the same base substitution occurs three times in independent tags. The sequences must not be derived from the originally same tag, since the PCR and sequencing steps are after ditag ligation, and duplicate ditags are discarded

phosphorylated forms of Mapk1 and 3 (synonyms: Erk 2 and 1) stain the posterior PSM of zebrafish [155]. In the mouse, however, only sporadic signal is observed in newly forming somites [156]. The specificity of the signaling by different FGFs and FGF receptors or through other receptor tyrosine kinases is thought to be mediated by the use of different components of the RAS-MAPK pathway. In this respect it is interesting to observe that only a subset of *RAS*, *MAP*, *MAPK*, *MAPKK*, *MAPKKK* and *MAPKKK* could be detected.

WNT signaling is closely connected to the function of FGF8 during somitogenesis. Although other modes of WNT signaling are known, for the canonical pathway Wnt proteins bind to its receptors, Frizzled and LRP5/6 (low-density-lipoprotein-receptor like protein 5 or 6), which in turns activate Dishevelled. Active (phosphorylated) Dishevelled leads to an inactivation of GSK3, thereby preventing β-catenin degradation. Stabilized β-catenin enters the nucleus and in cooperation with TCF/Lef protein activates the expression of WNT target genes. Furthermore several WNT signaling antagonists exist, like Dickkopf, which interacts with LRP, or Naked cutikule, an antagonist of Dishevelled (reviewed in [157, 158]). Like *Fgf8*, *Wnt3a* is also expressed in the tail tip, but the expression in the PSM is restricted to the very caudal part [159], explaining why in addition to library A *Wnt3a* is also represented in library B. The absence of *Wnt3a* transcripts also prevents the formation of somites [160], which can be explained by the fact that FGF8 acts downstream of Wnt3a [161]. However, despite of the assumed cranial to caudal gradient of Wnt3a protein, its action is thought to take effect in regular pulses, since the inhibitor *Axin2* of canonical Wnt signaling is dynamically expressed throughout the PSM. These pulses of Wnt/β-catenin signaling in turn controls oscillations of the Delta/Notch signaling pathway. A constitutive missexpression of *Axin2* throughout the whole PSM strongly disturbs the formation of somites [161]. Since the concept that WNT signaling is involved in somitogenesis is rather new, little is known how WNT signal is received and mediated in the PSM and nascent somites. For several of the *Frizzleds*, *Lrps* and *Dickkopfs*, no knock-outs are avilable. And among those, for which null alleles exist, somite phenotypes are not apparent or have been not been analyzed in detail (like for *Fzd3*, *Fzd4* and *Fzd5* [162, 163, 164]), or mice die prior to somite formation (*Lrp1* and *Lrp6* [165, 166]). With the exception of *Lrp1* (embryos die around implantation stage [165]), for *Fzd2*, *Fzd7*, *Lrp10* and *Dkk3* no null mutants have been published. But according to its abundance in the LongSAGE data, exactly these genes might be the mediators of the canonical WNT signaling pathway during somite formation.

The largest set of cycling genes are involved in Delta/Notch signaling. In general, upon binding of Notch ligands Delta or Jagged (Serrate in drosophila),

which are membrane-bound proteins expressed on neighboring cells, the intracellular domain of Notch (NIc) is cleaved by a γ-secretase complex (including Presenilin and Nicastrin). NIc is translocated to the nucleus, where it binds to a transcriptional repressor, recombining binding protein suppressor of hairless (Suppressor of Hairless in drosophila), facilitating the expression of target genes of Notch signaling such as the HES (Enhancer of Split in drosophila) basic helix loop helix family of transcriptional regulators. The Fringe proteins modulate Notch signaling by glycosylating Notch, resulting in a altered specificity of Notch ligands to Notch itself (reviewed in [167, 168, 169]). In agreement with the LongSAGE predictions, *Notch1* [170] and its ligands *Dll1* [171] and *Dll3* [172] are expressed in the PSM, and the cycling gene *Lfng* is also detected as expected [116]. In the absence of each of the four genes, boundaries between somites are not formed and anterior-posterior patterning of formed somites is disturbed [173, 174, 175, 176]. A lack of the ligand *Jag1*, expressed in the tail bud and throughout the PSM except for the anterior halves of S0 and S-I [177] does not result in an obvious somite phenotype [178]. Surprisingly, *Psen1* is not observed, although its absence affects (but not completely inhibits) somite boundary formation and disturbs A/P compartment determination in formed somites [179, 180]. But this could be explained in that expression level of *Psen1* is low and therefore can not be detected at this scale of LongSAGE analysis. Lack of *RBP-J*κ, which is ubiquitously expressed in the tissues assayed, delays the formation of somites, but (however poorly) segmented somites with a correct anterior-posterior polarity can be observed [181]. Out of the four effectors of the HES family, which could be detected in the LongSAGE data, *Hes1* and *Hes7* cycle within the PSM in the course of the generation period of one somite [182, 183], whereas *Hes3* and *Hes5* show a static expression pattern inside the PSM [184, 107]. But only for knock-out alleles of *Hes7* a somite phenotype, irregular epithelial somites and especially for cranially located somites a anterior-posterior patterning defect, has been reported [183, 185]. For targeted disruption of *Hes1*, *Hes3*, *Hes5* alone and double knock-outs of *Hes1* and *Hes3* as well as *Hes1* and *Hes5* no defects during somitogenesis are reported [186, 187, 188]. One very interesting aspect is the measured expression of *Mesp1* and *Mesp2*, two mediators of Notch signaling. Both show cycling expression between S-1 and the cranial half of S0 [189, 117]. While in the absence of *Mesp2* no segmented somites are observed and the anterior-posterior polarity in the fragmental somites formed is disturbed, lack of *Mesp1* does not cause a somite phenotype, albeight both are functionally redundant, since a knock-in of *Mesp1* cDNA into the locus of *Mesp2* largely rescues the *Mesp2* phenotype [190]. Knowing from the LongSAGE data, that the expression level of *Mesp2* is higher than that of *Mesp*, and since the res-

cue of the *Mesp2* phenotype by *Mesp1* is dose-dependent, this can be easily explained just by the higher amount of *Mesp2*. New within the dataset is the observed expression of different *Tle* genes (reviewed in [191]), which have not yet been analyzed in detail for its function during somitogenesis. The knockout of *Tle1* did not show embryonic defects [192], however this is not surprising, since the LongSAGE data suggests that the similarly expressed paralogs *Tle2* and *Tle3* could compensate for its function.

In summary, the LongSAGE predictions for elements of all tree pathways is in accordance with previous reports. Based upon the measured expression level, candidates mediating downstream events of the particular signaling pathways during somitogenesis could be proposed.

mesenchyme-to-epithelium transformation This transition is not a abrupt change, but a progressive shift of initially loose, unoriented mesenchymal cells, that become compacted at the anterior part of the PSM and in the periphery start to become organized in an epithelial-like structure. Thereby epithelial structures on the dorsal and ventral aspects of the future somites precede the final separation of the nascent somites by epithelia [193]. Time-lapse analysis of chick embryos showed that this separation is not a simple conversion in fate of the particular cells, but a highly dynamic event including tissue separation, cell movement, and selective integration of cells into the anterior and posterior somite borders [194]. The epithelial morphogenesis, which is associated with the assembly of a basement membrane (basal lamina plus an associated layer of reticulin fibers), can be monitored by the expression of basal lamina constituents *Laminin*, *Fibronectin* and *Collagens* [195], whose protein expression (shown for Laminin and Fibronectin) in deed co-localizes at the basal lamina of forming epithelia, despite each of them on its own is not required for the aggregation of PSM cells *in vitro* [193]. Among the *Laminins*, which form glycoprotein heterotrimers of α, β and γ chains (reviewed in [196]), the α subunits *Lama1* and *Lama5* as well as the γ subunit *Lamc1* are similarly expressed in the PSM and the recently formed somites, with a higher expression peak of *Lama5* in the nascent somites. Though, targeted disruption of most of the Laminin did not yet revel its function during somitogenesis (reviewed in [196]). *Fn1* however, which interestingly is no longer observed in tissue D (in contrast, protein could be detected in nascent somites [193]), is essential for somite formation, since in knock-out mice no somites are formed [197]. Even though it has been shown that collagen synthesis inhibitors severely interfered with the formation of somites [198], detailed roles of collagens during somitogenesis are not known (reviewed in [199]). In the case of *Col4a1* (and other *type IV Col-*

lagens, which are expressed at around the detection level) and *Col18a1*, its integration in basement membranes is known (reviewed in [195]). For the remaining *Col2a1*, *Col5a1*, *Col5a2*, *Col8a1*, *Col11a1*, *Col13a1* and *Col19a1* it is intriguing to analyze its function during the process of somitogenesis. Integrins predominantly function as transmembrane receptors for ECM proteins including Laminin, Fibronectin and Collagen, albeight participation in direct cell-cell adhesion is known for some integrins (reviewed in [200, 201]). Both alpha and beta subunits of heterodimeric integrins are detected, namely *Itga2b*, *Itga3*, *Itga5*, *Itgav*, *Itgb1*, *Itgb2*, *Itgb4* and *Itgb5*. The statistically differentially expressed ones (*Itga2b*, *Itgb2* and *Itgb4*) have its highest expression in tissue C, like for the integrin receptor *Vcam1* (reviewed in [200]), suggesting its importance for epithelial cells. Concordant, single- knock-outs for many integrins showed its requisite for basement membrane ECM organization and anchoring to epithelia (reviewed in [202]). Blocking beta 1 class of integrins by monoclonal antibodies results in lateral translocation of PSM and somites, but segmentation itself is not affected [203].

Among cell-cell adhesion receptors, the epithelial marker *Cdh1* (old name: *E-cadherin*) (reviewed in [204]) as well as *Cdh2* (*N-cadherin*), *Ncam1*, *Pcdh8* (*Papc*) and *Cdh11* are expressed, though (except for *Pcdh8*: highest expression of tissue C) statistical changes can not be predicted. Inhibition of Cdh2 and N-CAM prevented aggregation of dissociated PSM and of somite cells *in vitro* [193], and a secreted from of Ncam as well as lack of Cdh2 *in vivo* results in small irregular somites [205, 206, 207]. Double homozygous null mutants for *Cdh2* and *Cad11* as well as antibodies against Pcdh8 epithelia between somites are almost absent, although segmentation still takes place [208, 209]. Furthermore, additional *Cadherins* (*Cdh3*, *Cdh5* and *Cdh11*), *CAMs* (*Mcam* and *Pecam*) and *Protocadherins* (*Pcdh10* and *Pcdhga11*) are detected. *Pcdhga11*, the only other Protocadherin included in the dataset, might be a good candidate cross-reacting against the antibody against Pdch8, since Pdch8 knock-outs showed no phenotype [210]. Additionally 86 (9 differentially expressed) additional cell-adhesion genes (GO:0007155), including 9 uncharacterized ESTs, are recorded in this study. Aside from being candidates mediating epithelial-mesenchymal transition, these genes moreover could facilitating anterior-posterior sub-division of nascent somites, in case they would only be expressed in one compartment of the nascent somites. Yet no cell adhesion molecules accountable for this events have been shown (discussed in [208]).

In conclusion, the LongSAGE data suggests, that the conversion of mesenchymal presomitic mesoderm to epithelial somites requires a concerted action of many often redundant cell adhesion molecules and components of the ECM. Among the large families of molecules, only a small subset is ex-

pressed at high levels, suggesting that only a small subset within each family is involved in the process. However, within most of the families more than one gene is expressed, which could compensate the function of its paralog when that is inactivated, explaining why knock-outs for many cell adhesion molecules did not result in defects during somite formation. Interestingly most discussed genes have its expression peak within tissue C, indicating that the most dramatic changes at the molecular level occur at the anterior end of the PSM. In the same tissue components of the cytoskeleton, the tubulins *Tuba8* and *Tubgcp2* become significantly upregulated. Expression levels of actins itself do not change, but the Rho GTPase activating and interacting genes *Arhgap8* and *Rhiip3* become upregulated. Reorganization of actin and myosin fibers, mediated by Rho GTPases, catalysators of filament polymerization, is a typical sign of cell movement (reviewed in [211, 212]). Similarily the dynein motors *Dnahc11* and *Dnchc11* as well as the myosin motor *Myh2* are most abundant in tissue C. Furthermore there are 11 additional differentially expressed genes associated to the GeneOntology terms 'cell motility' (GO:0006928) or 'cell migration' (GO:0016477) in the dataset, all together suggesting that, as observed by above mentioned time-lapse analysis, cell movements or even cell migration plays a major role during the transition from PSM to epithelial somites.

4.4 Genome-wide analysis of publically available SAGE libraries

physical linkage The finding that 18 - 46 % (BMP4: 18 %; SHH: 41 %; JNK2: 46 %; cMYC: 34 %) of the genes predicted to be differentially expressed in all four induction systems analyzed are physically linked is of great interest. Even though simulations showed that some of the clusters are random, due to high gene densities in the respective chromosomal domain (see Table 4.2), the number of clusters is approximately twice as high compared to the mean of the simulations and clusters with three ore more genes or genes next to each other are highly overrepresented[3]. This observation fits in a line of earlier reports also describing that based upon data mining in a large collection of human SAGE data highly expressed genes cluster to certain gene-rich chromosomal domains [75]. Furthermore, studies based upon EST and microarray data analysis detected that a significant fraction of tissue-specific genes are in close vicinity on the genome of the lower vertebrates *C.*

[3]It should be noted, that more genes within these clusters could be differentially expressed; its expression level is just too low to be addressed at this scale of SAGE

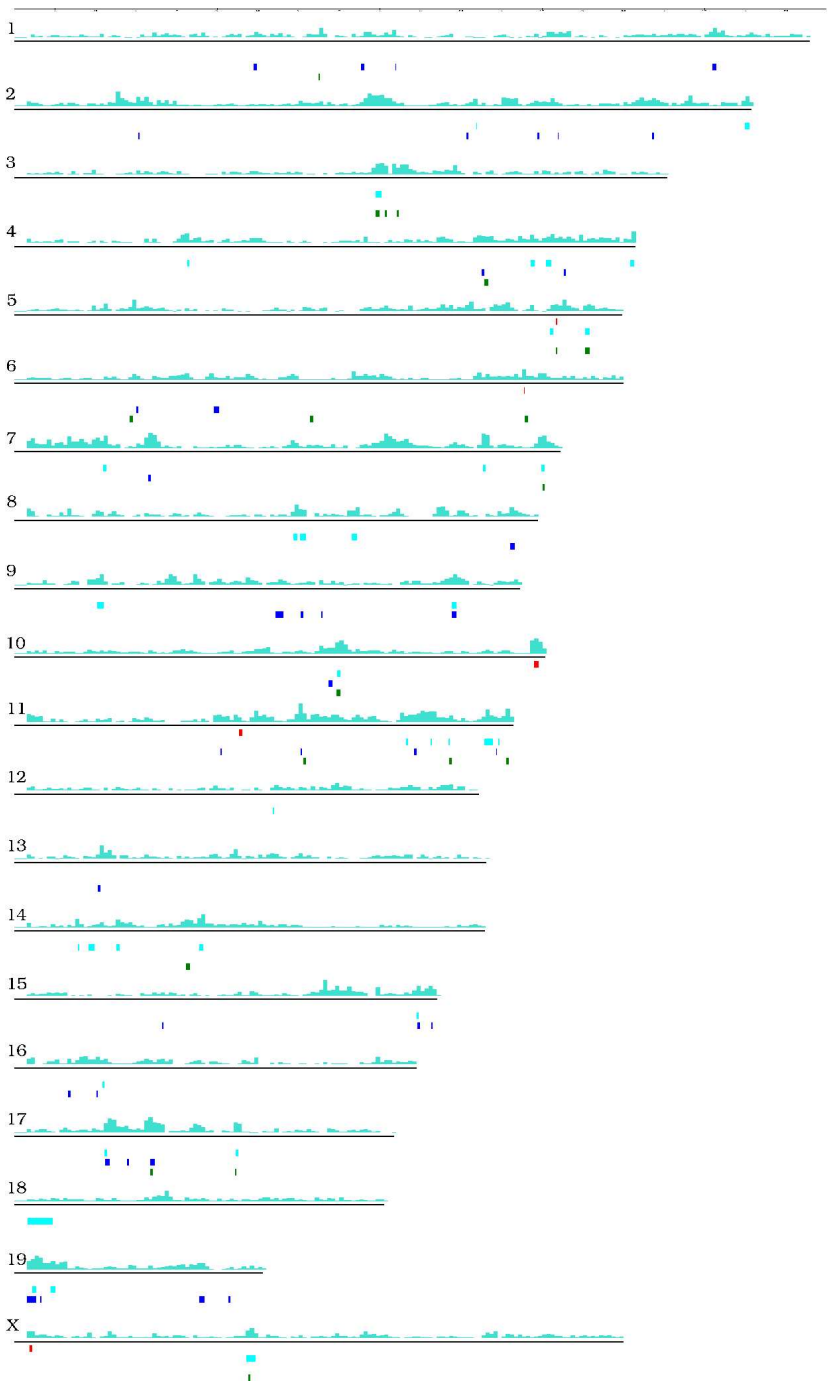


Figure 4.2: Clusters containing two or more differentially expressed genes within 1 Mb plotted to each single mouse chromosome. Library-specific colored boxes indicate the area the clusters cover. Black vertical bars represent chromosomes (names left to respective bar). Turquoise histograms on top of chromosomes display gene density. Mouse fragments were directly used, and for Human fragments, its corresponding syntenic region was used. Red: BMP4, cyan: JNK2, blue: SHH, green: cMYC.

elegans [213] and *D. melanogaster* [214, 215][4]. Although there are striking examples, conserved transcription factor binding sites of the downstream effectors of the TGFβ (SMAD3 and 4 [85]), SHH (GLI [84]), JNK (AP1 [86, 87] and CRE-binding protein 1/c-Jun heterodimers [88]) signaling pathways and cMYC (cMYC/MAX heterodimer (reviewed in [89])) itself within conserved noncoding sequences between mouse and human are observed, but not significantly enriched in the respective clusters, suggesting that most of the genes are not direct targets. In deed, by inhibiting protein synthesis in cultured ATDC5 cells, changes for some of the differentially expressed genes were no longer observed upon BMP induction (data not shown). The data indicates that there might be a common mechanism for transcriptional control of gene expression in many different biological contexts. Solely based upon these observations based gene expression data, it is difficult to speculate the nature of this mechanism. It has been suggested, that these observations are due to changes in the chromatin structure [214]. Since chromatin in its most packed form is inaccessible for transcription factors and RNA polymerases, it has to be unpacked in an ATP-dependent manner before being transcribed (reviewed in [216]). Also histone modification enhances chromatin recruitment complexes and thereby boost transcription, as shown for the IFN-β promotor [217]. But the complexes yet identified mediating mucleosome remodeling prior to transcriptional activation all act locally on single genes [218]. Therefore a recruitment of these to a certain chromosomal domain can not account for affecting gene expression for multiple neighboring genes. Another reason to argue against selective chromosomal 'opening' and 'closing' at the particular chromosomal domains is, that this way all genes within such a cluster should be either up- or downregulated. However, the focus in this study was on differentially expressed genes independent of the direction. In deed, in many of the linkage groups, gene both up- and downregulated genes were included.

House-keeping genes To be applicable as a house-keeping gene, a gene should be expressed at a medium level, and its expression should be constant within any tissue. Surprisingly, there were only a few genes that could be detected within all SAGE libraries. However, since SAGE is a tag-sampling experiment, low abundant genes might not be represented within the libraries, although they were expressed. This also explains why there were more com-

[4]However, in all three studies differences between tissues are not statistically significant, due to low numbers of ESTs or the wrong assumption, that microarray data are quantitative, and furthermore no statistical evaluation of the microarray data is discussed in the publication of [214].

mon mouse tags than human tags, since the number of human SAGE libraries was larger.

Taken together, the results argue against a true house-keeping gene. For both human and mouse, the most constant genes still vary between 21% (mouse) to 40% (human). Interestingly, genes commonly used as 'house-keeping genes', like *Gapdh* or *Hprt*, are not included in the list.

Ribosomal gene expression In agreement with the observations made during the SAGE analysis of chondrogenesis in ATDC5 cells, there have been several additional reports from SAGE analyses pointing out strong changes in ribosomal protein gene expression upon induction, like by nerve growth factor [219] or nMYC [142]. Similar results have been obtained in a own survey on SAGE library pairs of normal compared to cancer tissues, which is in accordance to non-SAGE publications showing differential ribosomal protein expression between normal and cancer cell lines[220] and tissues [221] on the protein level.

When all publically available SAGE libraries were adducted to cluster analysis according to its expression profile of ribosomal protein genes, very distinct sets of libraries could be identified. Unexpectedly, those libraries derived from exactly the same tissue or cell line were clustered together, even if cultured cells with and without induction by an ectopic stimulus were compared or if libraries were generated with tissue from cancer patients and corresponding tissue from wildtype controls. Since most of the analyzed SAGE libraries were generated one after another in a semi-automated fashion in the lab of Greg Riggins at Duke University with the same people specialized on a single step of the SAGE library construction (personal communication), it can be excluded that this phenomenon is due to experimental artefacts introduced by a particular experimenter. Some outliers appeared in clusters that mainly comprised of libraries generated from another tissue. This could be interpreted in that the nature of the cell type within the particular tissue also influences the expression of ribosomal protein genes. For example, the prostate SAGE libraries HS_GSM683 to 686, generated from epithelium and stroma are very distinct from the prostate SAGE libraries Hs_GSM739 to 740, which were made from microdissected adenocarcinoma and its wild-type equivalent. Interestingly, this phenomenon is even conserved between mouse and human, since three mouse brain SAGE libraries fall within the large cluster consisting mainly of brain libraries. However, there was no correlation of sex and age of the individual (from whom/ which the tissue was extracted from) with particular clusters.

These observation suggest that every cell type has an characteristic state

of ribosomal protein gene expression. Since the main function of ribosomal protein genes is to contribute to ribosomes, which in turn carry out protein biosynthesis (reviewed in [222]), this difference could lead to a different global level of protein synthesis. Recently, the 'ribosomal filter hypothesis' has been proposed by Mauro et al [223]. They suggest, that ribosomal proteins itself mediate interactions between mRNAs and components of the translation machinery, thereby selectively regulating protein synthesis. Therefore, the cell-/ tissue-specific ribosomal protein gene expression observed in this study could be to some extend responsible for the synthesis of cell-/ tissue-specific proteins. The fact, that SAGE libraries constructed from cell lines clearly separated from those generated from tissues could imply that cell lines can only in part substitute for *in vivo* data.

4.5 Outlook

All experimental results obtained as well as the bioinformatical analyses accomplished in this study are based upon gene expression data. Such a approach is very powerful, since all genes are included and a large number of biological states can be assayed. But it should only be considered as a screening approach. All observation will have to be analyzed by functional studies.

Appendix A

Database tables

Table A.1: tables in database *SPECIES*_master

Field	Type	Null	Key	Default	Extra
table 3prime_unigene_build*VERSION*_fantom*VERSION*					
id	int(10)		PRI	NULL	auto_increment
genbank_id	varchar(15)	YES	MUL	NULL	
riken_clone_id	varchar(25)	YES		NULL	
image_clone_id	varchar(25)	YES		NULL	
sequence	text	YES		NULL	
unigene_id	int(7)	YES	MUL	NULL	
riken_rts	varchar(10)	YES		NULL	
riken_fantom2	int(7)	YES		NULL	
sequence_source	enum('unknown', 'cDNA_full-length', 'cDNA_partial', 'cDNA_unknown' ,'EST_5prime', 'EST_3prime', 'EST_unknown')	YES		NULL	
polyA_tail	enum('no', 'yes')			no	
polyA_signal	enum('no', 'yes')			no	
strand	enum('+','-')	YES		NULL	
table description_unigene_build*VERSION*|_fantom*VERSION*					
id	varchar(15)	YES		NULL	
description	text	YES		NULL	
table gxd_*DATE*					
id	int(10)		PRI	NULL	auto_increment
mgd_id	int(8)	YES		NULL	
symbol	varchar(255)				
assay_type	varchar(100)				
result_detail	int(8)			0	
mutation	varchar(255)	YES		NULL	

table continues on following page

Field	Type	Null	Key	Default	Extra
age	varchar(50)	YES		NULL	
structure	varchar(255)	YES		NULL	
detected	enum('yes', 'no')	YES		NULL	
table `link_mgd_`*DATE*					
id	int(10)		PRI	NULL	auto_increment
mgd_id	int(8)		UNI	0	
symbol	varchar(255)	YES		NULL	
name	text	YES		NULL	
cm_position	varchar(8)	YES		NULL	
chromosome	char(2)	YES		NULL	
genbamk_id	text	YES		NULL	
unigene_id	varchar(255)	YES		NULL	
refseq_id	varchar(255)	YES		NULL	
table `link_unigene_`*VERSION*`_enstrans_`*VERSION*					
id	int(10)		PRI	NULL	auto_increment
unigene_id	varchar(15)	YES		NULL	
ensembl_gene_stable_id	varchar(25)	YES		NULL	
ensembl_transcript_stable_id	varchar(25)	YES		NULL	
percent_identity	double(3,1)	YES		NULL	
hit_length	int(5)	YES		NULL	
table `sequence_unigene[_unique]_build`*VERSION*`|_fantom`*VERSION*					
id	varchar(15)	YES		NULL	
sequence	text	YES		NULL	

Data from the respective sources was directly loaded into tables of *SPECIES* `_master` without being processed.

Table A.2: tables in database *SPECIES* `_longSAGEmapping_`*VERSION*

Field	Type	Null	Key	Default	Extra
est_mapping					
id	int(10)		PRI	NULL	auto_increment
genome_hits_id	int(10)			0	
tagseq	varchar(17)	YES	MUL	NULL	
rep_ensembl_stable_id	varchar(25)	YES		NULL	

table continues on following page

Field	Type	Null	Key	Default	Extra
sequence	ENUM('genome_hit_transcripts AND genome_hit_ests', 'genome_hit_transcripts ONLY', 'antisense transcript', 'antisense transcript upstream', 'antisense transcript downstream', 'genome_hit_transcripts NEW INTRON', 'genome_hit_transcripts NEW 3prime UTR', 'NOVEL')	YES		NULL	
distance	INT(6)			0	
unigene_ids	varchar(255)	YES		NULL	
mgd_ids	varchar(255)	YES		NULL	
description	text	YES		NULL	
table genome_hit_est_sequences					
id	int(10)		PRI	NULL	auto_increment
genbank_acc	varchar(40)				
ensembl_gene_stable_id	varchar(25)	YES		NULL	
ensembl_transcript_stable_id	varchar(25)	YES		NULL	
percent_identity	double(3,1)	YES		NULL	
hit_length	int(5)	YES		NULL	
table genome_hit_est_sequences					
id	int(10)		PRI	NULL	auto_increment
genbank_acc	varchar(40)				
sequence	text				
table genome_hit_ests					
id	int(10)		PRI	NULL	auto_increment
genome_hits_id	int(10)			0	
genbank_acc	varchar(40)				
table genome_hit_transcriptss					
id	int(10)		PRI	NULL	auto_increment
genome_hits_id	int(10)			0	
ensembl_stable_id	varchar(25)	YES		NULL	
table genome_hits					
id	int(10)		PRI	NULL	auto_increment
tags_id	int(10)			0	
chromosome	char(2)	YES		NULL	
start	int(9)	YES		NULL	
end	int(9)	YES		NULL	
strand	int(1)	YES		NULL	
table hits					
id	int(10)		PRI	NULL	auto_increment
tags_id	int(10)			0	

table continues on following page

Field	Type	Null	Key	Default	Extra
genome_hits_id	int(10)			0	
unigene_hits_id	varchar(255)			0	
rep_ensembl_stable_id	varchar(25)	YES		NULL	
table map_17_genome_*VERSION*					
id	int(10)		PRI	NULL	auto_increment
tag_sequence	char(17)	YES	MUL	NULL	
chromosome	char(2)	YES		NULL	
start	int(9)	YES		NULL	
end	int(9)	YES		NULL	
strand	int(1)	YES		NULL	
table map_17_unigene_*VERSION*					
id	int(10)		PRI	NULL	auto_increment
unigene_id	varchar(15)	YES		NULL	
tag_sequence	varchar(17)	YES	MUL	NULL	
full_length	double(3,2)	YES		NULL	
3prime	double(3,2)	YES		NULL	
table mapping					
id	int(10)		PRI	NULL	auto_increment
hits_id	int(10)			0	
tagseq	varchar(17)	YES	MUL	NULL	
genome_hit	enum('yes', 'no')	YES		NULL	
unigene_ids	varchar(255)	YES		NULL	
rep_ensembl_stable_id	varchar(25)	YES		NULL	
mgd_ids	varchar(255)	YES		NULL	
description	text	YES		NULL	
table tags					
id	int(10)		PRI	NULL	auto_increment
tagseq	char(17)	YES	MUL	NULL	
status	enum('new', 'hits', 'hits analyzed')	YES		NULL	
table unigene_hit_transcripts					
id	int(10)		PRI	NULL	auto_increment
genome_hits_id	int(10)			0	
ensembl_stable_id	varchar(25)	YES		NULL	
table unigene_hits					
id	int(10)		PRI	NULL	auto_increment
tags_id	int(10)			0	
unigene_id	varchar(15)	YES		NULL	
full_length	double(3,2)	YES		NULL	
3prime	double(3,2)	YES		NULL	
table xref_mgd					
id	int(10)		PRI	NULL	auto_increment
hits_id	int(10)			0	
mgd_id	int(8)	YES		NULL	

All tables within *SPECIES*_longSAGEmapping_*VERSION* are linked to the master table (`tags`) by its primary id. Characters preceeding any _id denote the name of the table this particular number is the id for (e.g. tags.id = genome_hits.tags_id).

Table A.3: tables in database *SPECIES*_SAGEmapping

Field	Type	Null	Key	Default	Extra
table map_*TAGLENGTH*_unigene_*VERSION*					
id	int(10)		PRI	NULL	auto_increment
unigene_id	varchar(15)	YES		NULL	
tag_sequence	varchar(17)	YES	MUL	NULL	
full_length	double(3,2)	YES		NULL	
3prime	double(3,2)	YES		NULL	

Database with tables for non-LongSAGE mappings.

Table A.4: tables in database *SPECIES*_tag2genome

Field	Type	Null	Key	Default	Extra
table cluster2chromosome_ensembl_*VERSION*_unigene_*VERSION*					
id	int(10)	PRI		NULL	auto_increment
ensembl_id	varchar(25)	YES		NULL	
unigene_ids	varchar(100)	YES		NULL	
riken_rts	varchar(15)	YES		NULL	
riken_repclone_id	varchar(15)	YES		NULL	
number_exons	int(3)	YES		NULL	
start	int(9)	YES		NULL	
end	int(9)	YES		NULL	
chromosome	char(2)	YES		NULL	
sequence	text	YES		NULL	
symbol	varchar(30)	YES		NULL	
description	text	YES		NULL	

For each non-redundant transcript entry in EnsEMBL, all associated sources were written to a single database table.

Table A.5: tables in database SAGE_data

Field	Type	Null	Key	Default	Extra
table master					
id	int(10)	PRI		NULL	auto_increment
sample_id	varchar(100)	YES		NULL	
titel	varchar(255)	YES		NULL	
anchor	enum('NlaIII', 'Sau3A')	YES		NULL	
taglength	int(3)	YES		NULL	
single_count_tags	enum('yes', 'no')	YES		NULL	
organism	enum('Hs', 'Mm', 'Rn', 'At')	YES		NULL	
source	varchar(255)	YES		NULL	

table continues on following page

Field	Type	Null	Key	Default	Extra
description	text	YES		NULL	
count	int(8)	YES		NULL	
author	varchar(255)	YES		NULL	
institute	varchar(255)	YES		NULL	
table *SAMPLE_ID*					
id	int(10)	PRI		NULL	auto_increment
tag	char(*TAGLENGTH*)				
count	int(7)			0	

Any SAGE and LongSAGE data used is initially written to seperate ***SAMPLE_ID*** tables, and its attributes are stored in table **master** (master.sample_id: Name of the ***SAMPLE_ID*** **table**

Table A.6: tables in database `SAGE_project`

Field	Type	Null	Key	Default	Extra
table master					
id	int(10)		PRI	NULL	auto_increment
project_name	varchar(100)		UNI		
status	enum('new', 'processed')	YES		NULL	
mapping_database	varchar(40)	YES		NULL	
ensembl_database	varchar(40)	YES		NULL	
table *PROJECT_ID*					
id	int(10)		PRI	NULL	auto_increment
data_id	int(10)			0	
group_name	varchar(255)				
kind	enum('data', 'virtual_subtraction')	YES		NULL	
table project_*PROJECT_ID*					
id	int(10)	PRI		NULL	auto_increment
tag	char(*TAGLENGTH*)				
DATA	int(4)			0	
*DATA*_vs	double(5,2)			0	
				for each set of DATA SAGE libraries	
*DATA*_vs_*DATA*	double(5,2)			0	
				for each combination of DATA	
data_tpm	double(5,2)			0	
data_short_tpm	double(5,2)			0	
vs_*VS*	double(5,2)			0	
				for each set of VS SAGE libraries	
factor_*VS*	double(5,2)			0	
				for each set of VS SAGE libraries	

For each analysis including more than one SAGE library, **project** tables were generated. **master**: General attributes of analysis; *PROJECT_ID*: Libraries used for analysis; **project_*PROJECT_ID***: Results of analysis.

Table A.7: tables in database *SPECIES*_longSAGEannotation_*VERSION*

Field	Type	Null	Key	Default	Extra
table annotation					
id	int(10)		PRI	NULL	auto_increment
tagseq	char(17)		UNI		
sequence_source	enum('new', 'no hit', 'multiple hits', 'no hit multiple predictionTranscripts', 'EnsEMBL protein', 'EnsEMBL estgene', 'EnsEMBL predictionTranscript', 'MGD', 'UniGene')	YES		NULL	
table go_similarity					
id	int(10)		PRI	NULL	auto_increment
annotation_id	int(10)			0	
prot_db	enum('Coils', 'Family', 'Lowcompl', 'Pfam', 'Prints', 'Profile', 'Prosite', 'Sigp', 'Superfamily', 'Transmembrane' ,'blastp')	YES		NULL	
prot_db_id	varchar(40)				
go_id	varchar(10)	YES		NULL	
table *PROTEIN_DATABASE*2go					
prot_db_id	varchar(40)				
go_id	varchar(10)	YES		NULL	

go_similarity holds the final result of annotation for any tag loaded to **annotation** (annotation.id = go_similarity.annotation_id). Tables ***PROTEIN_DATABASE*2go** contain link of *PROTEIN_DATABASE* ID to GO ID.

Appendix B

ATDC5

Table B.1: Detailed annotation for genes differentially expressed between the two ATDC5 libraries

#	Symbol	Function	References
Transriptional regulation			
DNA binding			
D64	*Calr*	interacts with DNA-binding domain of glucocorticoid receptor (prevents it from binding to glucocorticoid response element)	[224, 225]
D67	*Nca*	stabilizes AP-1 complex formed by c-Jun homodimer on target sequence	[226, 227]
inhibition of DNA binding			
U106	*Idb3*	negative regulator of basic helix-loop-helix transcription factors; cell cycle regulation; arrest of myotube differentiation	[228, 229] (reviews)
U126	*Idb2*	negative regulator of basic helix-loop-helix transcription factors	[228] (review)
DNA packaging			
U33	*Ptma*	interacts with histones (H1- binding); interacts with CBP (CREB-binding protein); stimulates AP1 and NK-kB-dependent transcription	[230]
D61	*Hmgn1*	reduces compactness of chromatin fiber and enhances transcription from chromatin templates	[231] (review)
Protein metabolism			
protein synthesis			
ribosomal proteins			

table continues on following page

#	Symbol	Function	References
D5	*Eef1a1*	translation elongation	[232]
U91	*Sui-rs1*	homolog to Sui1 (S. cerevisiae); translation initiation factor interacting with eIF2 (eukaryotic inititiation factor 2)	[233, 234]
U123	*Eif4g2*	homolog to Eif4G; inhibits cap-dependent and cap-independent translation	[235]
protein folding			
U80	*Serpinh1*	heat-shock protein; collagen biosynthesis	[236]
U134	*Hspa5*	heat-shock protein	[237]
protein transport/sorting			
D67	*Nca*	signal-sequene specific sorting and translocation	[226, 227]
U116	*Sec61g*	gamma subunit of Sec61/Sec complex; co- and post-translational transport of proteins into ER; integration of membrane proteins	[238] (review)
protein degradation			
D38	*Ppp2cb*	catalytic subunit of protein phosphatase 2 (beta isoform)	[239] (review)
D44	*Psmb4*	beta subunit 20S core proteasome (of 26S proteasome)	[240] (review)
D51	*Psmb1*	beta subunit 20S core proteasome (of 26S proteasome)	[240] (review)
D57	*Psmb3*	beta subunit 20S core proteasome (of 26S proteasome)	[240] (review)
U110	*Vcp*	involved degradation of ubiquitin-fusion proteins and proteasome-dependent cleavage of ER membrane proteins	[241]
U116	*Sec61g*	gamma subunit of Sec61/ Sec complex; retrograde transport of misfolded proteins from ER lumen to cytosol for degradation	[242] (review)
U121	*Usp5*	disassembly of polyubiquitin changes after ubiquitin-dependent protein degradation (26 S proteasome)	[243], [240] (review)
Vesicles			
Vesicle mineralization			
D68	*Anxa5*	mediates influx of Ca2+ into (mineralizing) vesicles	[244] (review)
Vesicle transport			
D53	*Rab11b*	associates with Myosin Vb (motor for vesicle movement on F-actin); regulates plasma membrane recycling	[245] (review)
D68	*Anxa5*	anchors vesicles to ECM (binds type II and X collagen)	[244] (review)

table continues on following page

#	Symbol	Function	References
U98	*Hspa8*	part of complex removing clathrin from coated vesicles	[246] (and refs therein)
U110	*Vcp*	associated with clathrin and Hspa8 (tag D98) on coated vesicles	[247]
U133	*Sara*	involved in COPII coated vesicle transport from ER; COPII assembly and disassembly are regulated by Sar1 cycle	[248] (review)
U139	*Shfdg1*	function in association with exocyst complex in yeast	[249, 250]
metabolism/ homeostasis			
D12	*Ftl1*	ion metabolism/ transport	[251] (review)
D39	*Ndufa7*	subunit of NADH:ubiquitone oxidoreductase complex	[252]
D71	*Mor1*	malat dehydrogenase; participates in malate-aspartate shuttle	
U81	*Cox6c*	respiratory chain complex IV	
U89	*Eno1*	glycolysis	[253]
U93	*Ldh1*	glycolysis	[254]
U96	*Idh2*	tricarboxylic acid cycle	[255]
U111	*Atp6g1*	proton transport, ATP biosynthesis	
U127	*Tpi*	triose phosphate isomerase (glycolysis/ gluconeogenesis, fetty acid biosynthesis)	
U136	*Pla2g4a*	lipid degradation	[256] (review)
signaling extracellular			
D54	*Cxcl12*	Receptor in Lymphocytes: CXCR4: activates G-coupled phosphoinositide 3-kinase activates ERK1/2	[257]
U77	*Ptn*	heparin-binding cytokine with multiple functions	[258]
transmembrane proteins (integral membrane protein)			
D13	*Sdc2*	regulates signaling of HS binding growth factors and cell-cell signaling (e.g. modifies signals generated by integrin-mediated cell adhesion	[259] (Review)
D19	*Tm4sf8*	signal transduction (members of family; no functional data about gene itself	[260]
D42	*Gas1*	block of cell proliferation in vitro (entry to S phase); no growth inhibition observed in Gas1 -/- mice; antagonist of shh (somites)	[140], [261]
D50	*Emp3*	cell-cell signaling	[262]
D62	*Itgp*	integrin-mediated signaling	[263] (review)
U125	*Cd63*	signal transduction (members of family; no functional data about gene itself	[264]

table continues on following page

#	Symbol	Function	References
intracellular			
D32	*Tpt1*	anti-apoptotoc (could bind and scavenge Ca2+ released in response to apoptotic stimul)	[265, 266]
D38	*Ppp2cb*	Wng-beta-catenin signaling; Protein phosphatase 2A inhibits Wnt signaling	[267, 268]
D40	*Lag*	blocks (phosphorylated) cell cycle (arrest at G2/M interphase)	[269, 270]
D47	*Gnai2*	Adenylate cyclase inhibitor	[271]
D49	*S100a6*	Ca2+ (and Zn2+) binding protein of EF-hand type: involved in ca2+ - dependent signal transduction	[272] (review)
D64	*Calr*	integrin-mediated calcium signaling	[224, 225]
D66	*Hint*		[273], [274] (review)
D68	*Anxa5*	inhibitor of phospholipase A2 and protein kinase C	[244] (review)
U99	*Ywhae*	multiple signaling pathways	[275] (review)
U118	*Ywhag*	Ca2+ - binding; inhibitor of phospholipase A2 and protein kinase C	[276]
U129	*Btg1*	cell-cycle: anti-proliferative	[277, 278]
U135	*Bnip2*	anti-apoptotic	[279]
U136	*Pla2g4a*	calcium-dependent phosphatase	[256]
structural components of ECM			
D28	*Col1a2*	one of two alpha chains of type I collagen (heterotrimer of two alpha1 and one alpha2 chains); most abundant collagen, major structural protein of the ECM of bone, skin and tendon	[280]
U90	*Sparc*	bone formation (decreased in k.o. mice)	[281]
U92	*Osf2*	supports cell spreading and attachment in vivo	[282, 283]
U94	*Fn1*	heparin-binding; involved in limb precartilage condensations	[284]
U103	*Tgfbi*	integrin ligand	[285, 135]
U108	*Col3a1*	essential for normal collagen I fibrillogenesis (in variety of organs)	[286]
transmembrane cell adhesion molecules			
D13	*Sdc2*	heparin-binding	[259] (review)
cytoskeleton			
D21	*Ptmb10*	G-actin sequestering peptin: prevents spontaneous nucleation	[287]
D31	*Actg*	cytoskeletal gamma-actin	[288]

table continues on following page

#	Symbol	Function	References
associated with cytoskeleton			
D32	*Tpt1*	transiently associates with microtubules during cell cycle	[268, 266]
D40	*Lag*	microtubule-destabilizing factor	[269, 270]
Others or unknown			
D8	*Rbm3*		[289]
D27	*H2afx*	repair of double-strand breaks	[290]
D56	*Plp2*	ion channel	[291]
D66	*Hint*	Nucleotide-binding and diadenosine polyphosphate hydrolase	[273, 274]
D70	*filamin-like protein*	unknown	
U107	*Fin14*	unknown	[292]

Detailed annotation with referenes for the genes listed in Table 4.1

Bibliography

[1] R.H. Waterston, 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. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420(6915):520–62, 2002.

[2] E.S. Lander, 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, C. Raymond, 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. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.

[3] J.C. Venter, 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. The sequence of the human genome. *Science*, 291(5507):1304–51, 2001.

[4] Y. Okazaki, M. Furuno, T. Kasukawa, J. Adachi, H. Bono, S. Kondo, I. Nikaido, N. Osato, R. Saito, H. Suzuki, I. Yamanaka, H. Kiyosawa, K. Yagi, Y. Tomaru, Y. Hasegawa, A. Nogami, C. Schonbach, T. Gojobori, R. Baldarelli, D.P. Hill, C. Bult, D.A. Hume, J. Quackenbush, L.M. Schriml, A. Kanapin, H. Matsuda, S. Batalov, K.W. Beisel, J.A. Blake, D. Bradt, V. Brusic, C. Chothia, L.E. Corbani, S. Cousins, E. Dalla, T.A. Dragani, C.F. Fletcher, A. Forrest, K.S. Frazer, T. Gaasterland, M. Gariboldi, C. Gissi, A. Godzik, J. Gough, S. Grimmond, S. Gustincich, N. Hirokawa, I.J. Jackson, E.D. Jarvis, A. Kanai, H. Kawaji, Y. Kawasawa, R.M. Kedzierski, B.L. King, A. Konagaya, I.V. Kurochkin, Y. Lee, B. Lenhard, P.A. Lyons, D.R. Maglott, L. Maltais, L. Marchionni, L. McKenzie, H. Miki, T. Nagashima, K. Numata, T. Okido, W.J. Pavan, G. Pertea, G. Pesole, N. Petrovsky, R. Pillai, J.U. Pontius, D. Qi, S. Ramachandran, T. Ravasi, J.C. Reed, D.J. Reed, J. Reid, B.Z. Ring, M. Ringwald, A. Sandelin, C. Schneider, C.A. Semple, M. Setou, K. Shimada, R. Sultana, Y. Takenaka, M.S. Taylor, R.D. Teasdale, M. Tomita, R. Verardo, L. Wagner, C. Wahlestedt, Y. Wang, Y. Watanabe, C. Wells, L.G. Wilming, A. Wynshaw-Boris, M. Yanagisawa, I. Yang, L. Yang, Z. Yuan, M. Zavolan, Y. Zhu, A. Zimmer, P. Carninci, N. Hayatsu, T. Hirozane-Kishikawa, H. Konno, M. Nakamura, N. Sakazume, K. Sato, T. Shiraki, K. Waki, J. Kawai, K. Aizawa,

T. Arakawa, S. Fukuda, A. Hara, W. Hashizume, K. Imotani, Y. Ishii, M. Itoh, I. Kagawa, A. Miyazaki, K. Sakai, D. Sasaki, K. Shibata, A. Shinagawa, A. Yasunishi, M. Yoshino, R. Waterston, E.S. Lander, J. Rogers, E. Birney, and Y. Hayashizaki. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature*, 420(6915):563–73, 2002.

[5] M.S. Boguski and G.D. Schuler. ESTablishing a human transcript map. *Nat Genet*, 10(4):369–71, 1995.

[6] L. Stein. Genome annotation: from sequence to biology. *Nat Rev Genet*, 2(7):493–503, 2001.

[7] M. Kanehisa and P. Bork. Bioinformatics in the post-sequence era. *Nat Genet*, 33 Suppl:305–10, 2003.

[8] H. Kitano. Systems biology: a brief overview. *Science*, 295(5560):1662–4, 2002.

[9] V.E. Velculescu, L. Zhang, B. Vogelstein, and K.W. Kinzler. Serial analysis of gene expression. *Science*, 270(5235):484–7, 1995.

[10] M. Schena, D. Shalon, R.W. Davis, and P.O. Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270(5235):467–70, 1995.

[11] S. Saha, A.B. Sparks, C. Rago, V. Akmaev, C.J. Wang, B. Vogelstein, K.W. Kinzler, and V.E. Velculescu. Using the transcriptome to annotate the genome. *Nat Biotechnol*, 20(5):508–12, 2002.

[12] A. Erlebacher, E.H. Filvaroff, S.E. Gitelman, and R. Derynck. Toward a molecular understanding of skeletal development. *Cell*, 80(3):371–8, 1995.

[13] P.V. Thorogood and J.R. Hinchliffe. An analysis of the condensation process during chondrogenesis in the embryonic chick hind limb. *J Embryol Exp Morphol*, 33(3):581–606, 1975.

[14] D.A. Ede. *Cellular condensations and chondrogenesis*, chapter 4, pages 143–185. CRC Press, Boca Raton, 1991.

[15] M.D. Olson and F.N. Low. The fine structure of developing cartilage in the chick embryo. *Am J Anat*, 131(2):197–215, 1971.

[16] B.K. Hall and T. Miyake. All for one and one for all: condensations and the initiation of skeletal development. *Bioessays*, 22(2):138–47, 2000.

[17] B. de Crombrugghe, V. Lefebvre, and K. Nakashima. Regulatory mechanisms in the pathways of cartilage and bone formation. *Curr Opin Cell Biol*, 13(6):721–7, 2001.

[18] A.M. DeLise, L. Fischer, and R.S. Tuan. Cellular interactions and signaling in cartilage development. *Osteoarthritis Cartilage*, 8(5):309–34, 2000.

[19] M.R. Urist. Bone: formation by autoinduction. *Science*, 150(698):893–9, 1965.

[20] J.M. Wozney, V. Rosen, A.J. Celeste, L.M. Mitsock, M.J. Whitters, R.W. Kriz, R.M. Hewick, and E.A. Wang. Novel regulators of bone formation: molecular clones and activities. *Science*, 242(4885):1528–34, 1988.

[21] J. Massague. How cells read TGF-beta signals. *Nat Rev Mol Cell Biol*, 1(3):169–78, 2000.

[22] K. Miyazono, K. Kusanagi, and H. Inoue. Divergence and convergence of TGF-beta/BMP signaling. *J Cell Physiol*, 187(3):265–76, 2001.

[23] A. von Bubnoff and K.W. Cho. Intracellular BMP signaling regulation in vertebrates: pathway or network? *Dev Biol*, 239(1):1–14, 2001.

[24] Y. Shi and J. Massague. Mechanisms of TGF-beta signaling from cell membrane to the nucleus. *Cell*, 113(6):685–700, 2003.

[25] D. Duprez, E.J. Bell, M.K. Richardson, C.W. Archer, L. Wolpert, P.M. Brickell, and P.H. Francis-West. Overexpression of BMP-2 and BMP-4 alters the size and shape of developing skeletal elements in the chick limb. *Mech Dev*, 57(2):145–57, 1996.

[26] H. Zou, R. Wieser, J. Massague, and L. Niswander. Distinct roles of type I bone morphogenetic protein receptors in the formation and differentiation of cartilage. *Genes Dev*, 11(17):2191–203, 1997.

[27] S. Pizette and L. Niswander. BMPs are required at two steps of limb chondrogenesis: formation of prechondrogenic condensations and their differentiation into chondrocytes. *Dev Biol*, 219(2):237–49, 2000.

[28] M. Enomoto-Iwamoto, M. Iwamoto, Y. Mukudai, Y. Kawakami, T. Nohno, Y. Higuchi, S. Takemoto, H. Ohuchi, S. Noji, and K. Kurisu. Bone morphogenetic protein signaling is required for maintenance of differentiated phenotype, control of proliferation, and hypertrophy in chondrocytes. *J Cell Biol*, 140(2):409–18, 1998.

[29] C. Shukunami, C. Shigeno, T. Atsumi, K. Ishizeki, F. Suzuki, and Y. Hiraki. Chondrogenic differentiation of clonal mouse embryonic cell line ATDC5 in vitro: differentiation-dependent gene expression of parathyroid hormone (PTH)/PTH-related peptide receptor. *J Cell Biol*, 133(2):457–68, 1996.

[30] C. Shukunami, K. Ishizeki, T. Atsumi, Y. Ohta, F. Suzuki, and Y. Hiraki. Cellular hypertrophy and calcification of embryonal carcinoma-derived chondrogenic cell line ATDC5 in vitro. *J Bone Miner Res*, 12(8):1174–88, 1997.

[31] C. Shukunami, H. Akiyama, T. Nakamura, and Y. Hiraki. Requirement of autocrine signaling by bone morphogenetic protein-4 for chondrogenic differentiation of ATDC5 cells. *FEBS Lett*, 469(1):83–7, 2000.

[32] M. Fujii, K. Takeda, T. Imamura, H. Aoki, T.K. Sampath, S. Enomoto, M. Kawabata, M. Kato, H. Ichijo, and K. Miyazono. Roles of bone morphogenetic protein type I receptors and Smad proteins in osteoblast and chondroblast differentiation. *Mol Biol Cell*, 10(11):3801–13, 1999.

[33] A. Gossler and M. Hrabe de Angelis. Somitogenesis. *Curr Top Dev Biol*, 38:225–87, 1998.

[34] P.P. Tam, D. Goldman, A. Camus, and G.C. Schoenwolf. Early events of somitogenesis in higher vertebrates: allocation of precursor cells during gastrulation and the organization of a meristic pattern in the paraxial mesoderm. *Curr Top Dev Biol*, 47:1–32, 2000.

[35] A.E. Brent, R. Schweitzer, and C.J. Tabin. A somitic compartment of tendon progenitors. *Cell*, 113(2):235–48, 2003.

[36] D.S. Packard, Jr. The influence of axial structures on chick somite formation. *Dev Biol*, 53(1):36–48, 1976.

[37] D.S. Packard, Jr and S. Meier. An experimental study of the somitomeric organization of the avian segmental plate. *Dev Biol*, 97(1):191–202, 1983.

[38] D.S. Packard, Jr. Somitogenesis in cultured embryos of the Japanese quail, Coturnix coturnix japonica. *Am J Anat*, 158(1):83–91, 1980.

[39] P.P. Tam. A study of the pattern of prospective somites in the presomitic mesoderm of mouse embryos. *J Embryol Exp Morphol*, 92:269–85, 1986.

[40] P.P. Tam and R.S. Beddington. *The metameric organization of the presomitic mesoderm and somite specification in the mouse embryo*, volume 111 of *NATO ASI Series: Somites in Developing Embryos*, pages 17–36. Plenum Press, New York and London, 1986.

[41] S. Meier. Development of the chick embryo mesoblast. Formation of the embryonic axis and establishment of the metameric pattern. *Dev Biol*, 73(1):24–45, 1979.

[42] P.P. Tam and S. Meier. The establishment of a somitomeric pattern in the mesoderm of the gastrulating mouse embryo. *Am J Anat*, 164(3):209–25, 1982.

[43] R. Bellairs. The Development of Somites in the Chick Embryo. *J. Embryol. exp. Morph.*, 11(4):697–714, 1963.

[44] C.D. Stern and R. Bellairs. The roles of node regression and elongation of the area pellucida in the formation of somites in avian embryos. *J Embryol Exp Morphol*, 81:75–92, 1984.

[45] K.M. Correia and R.A. Conlon. Surface ectoderm is necessary for the morphogenesis of somites. *Mech Dev*, 91(1-2):19–30, 2000.

[46] R.J. Keynes and C.D. Stern. Segmentation in the vertebrate nervous system. *Nature*, 310(5980):786–9, 1984.

[47] D.S. Packard, Jr, R.Z. Zheng, and D.C. Turner. Somite pattern regulation in the avian segmental plate mesoderm. *Development*, 117(2):779–91, 1993.

[48] C.D. Stern and R. Bellairs. Mitotic activity during somite segmentation in the early chick embryo. *Anat Embryol (Berl)*, 169(1):97–102, 1984.

[49] D.R. Primmett, W.E. Norris, G.J. Carlson, R.J. Keynes, and C.D. Stern. Periodic segmental anomalies induced by heat shock in the chick embryo are associated with the cell cycle. *Development*, 105(1):119–30, 1989.

[50] D.R. Primmett, C.D. Stern, and R.J. Keynes. Heat shock causes repeated segmental anomalies in the chick embryo. *Development*, 104(2):331–9, 1988.

[51] Y. Saga and H. Takeda. The making of the somite: molecular events in vertebrate segmentation. *Nat Rev Genet*, 2(11):835–45, 2001.

[52] O. Pourquie. The segmentation clock: converting embryonic time into spatial pattern. *Science*, 301(5631):328–30, 2003.

[53] P. Chomczynski and N. Sacchi. Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. *Anal Biochem*, 162(1):156–9, 1987.

[54] M. Kenzelmann and K. Muhlemann. Substantially enhanced cloning efficiency of SAGE (Serial Analysis of Gene Expression) by adding a heating step to the original protocol. *Nucleic Acids Res*, 27(3):917–8, 1999.

[55] B. Virlon, L. Cheval, J.M. Buhler, E. Billon, A. Doucet, and J.M. Elalouf. Serial microanalysis of renal transcriptomes. *Proc Natl Acad Sci U S A*, 96(26):15286–91, 1999.

[56] J. Shires, E. Theodoridis, and A.C. Hayday. Biological insights into TCRgammadelta+ and TCRalphabeta+ intraepithelial lymphocytes provided by serial analysis of gene expression (SAGE). *Immunity*, 15(3):419–34, 2001.

[57] J.J. Chen, J.D. Rowley, and S.M. Wang. Generation of longer cDNA fragments from serial analysis of gene expression tags for gene identification. *Proc Natl Acad Sci U S A*, 97(1):349–53, 2000.

[58] S. Lee, T. Clark, J. Chen, G. Zhou, L.R. Scott, J.D. Rowley, and S.M. Wang. Correct identification of genes from serial analysis of gene expression tag sequences. *Genomics*, 79(4):598–602, 2002.

[59] H. Ito, H. Akiyama, C. Shigeno, and T. Nakamura. Noggin and bone morphogenetic protein-4 coordinately regulate the progression of chondrogenic differentiation in mouse clonal EC cells, ATDC5. *Biochem Biophys Res Commun*, 260(1):240–4, 1999.

[60] A. Neubuser, H. Koseki, and R. Balling. Characterization and developmental expression of Pax9, a paired-box-containing gene related to Pax1. *Dev Biol*, 170(2):701–16, 1995.

[61] L. Wall, T. Christiansen, and J. Orwant. *Programming Perl, 3rd Edition.* O'Reilly and associates, Sebastopol (Ca US), 2000.

[62] J.E. Stajich, D. Block, K. Boulez, S.E. Brenner, S.A. Chervitz, C. Dagdigian, G. Fuellen, J.G. Gilbert, I. Korf, H. Lapp, H. Lehvaslaiho, C. Matsalla, C.J. Mungall, B.I. Osborne, M.R. Pocock, P. Schattner, M. Senger, L.D. Stein, E. Stupka, M.D. Wilkinson, and E. Birney. The Bioperl toolkit: Perl modules for the life sciences. *Genome Res*, 12(10):1611–8, 2002.

[63] T. Hubbard, D. Barker, E. Birney, G. Cameron, Y. Chen, L. Clark, T. Cox, J. Cuff, V. Curwen, T. Down, R. Durbin, E. Eyras, J. Gilbert, M. Hammond, L. Huminiecki, A. Kasprzyk, H. Lehvaslaiho, P. Lijnzaad, C. Melsopp, E. Mongin, R. Pettett, M. Pocock, S. Potter, A. Rust, E. Schmidt, S. Searle, G. Slater, J. Smith, W. Spooner, A. Stabenau, J. Stalker, E. Stupka, A. Ureta-Vidal, I. Vastrik, and M. Clamp. The Ensembl genome database project. *Nucleic Acids Res*, 30(1):38–41, 2002.

[64] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 25(1):25–9, 2000.

[65] R. Apweiler, T.K. Attwood, A. Bairoch, A. Bateman, E. Birney, M. Biswas, P. Bucher, L. Cerutti, F. Corpet, M.D. Croning, R. Durbin, L. Falquet, W. Fleischmann, J. Gouzy, H. Hermjakob, N. Hulo, I. Jonassen, D. Kahn, A. Kanapin, Y. Karavidopoulou, R. Lopez, B. Marx, N.J. Mulder, T.M. Oinn, M. Pagni, F. Servant, C.J. Sigrist, and E.M. Zdobnov. The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res*, 29(1):37–40, 2001.

[66] J.A. Blake, J.E. Richardson, C.J. Bult, J.A. Kadin, and J.T. Eppig. MGD: the Mouse Genome Database. *Nucleic Acids Res*, 31(1):193–5, 2003.

[67] M. Ringwald, J.T. Eppig, D.A. Begley, J.P. Corradi, I.J. McCright, T.F. Hayamizu, D.P. Hill, J.A. Kadin, and J.E. Richardson. The Mouse Gene Expression Database (GXD). *Nucleic Acids Res*, 29(1):98–101, 2001.

[68] D.L. Wheeler, 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. Database resources of the National Center for Biotechnology. *Nucleic Acids Res*, 31(1):28–33, 2003.

[69] S. Schwartz, W.J. Kent, A. Smit, Z. Zhang, R. Baertsch, R.C. Hardison, D. Haussler, and W. Miller. Human-mouse alignments with BLASTZ. *Genome Res*, 13(1):103–7, 2003.

[70] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. Basic local alignment search tool. *J Mol Biol*, 215(3):403–10, 1990.

[71] Z. Zhang, S. Schwartz, L. Wagner, and W. Miller. A greedy algorithm for aligning DNA sequences. *J Comput Biol*, 7(1-2):203–14, 2000.

[72] B. Ewing and P. Green. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res*, 8(3):186–94, 1998.

[73] E.M. Zdobnov and R. Apweiler. InterProScan–an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*, 17(9):847–8, 2001.

[74] E. Beaudoing, S. Freier, J.R. Wyatt, J.M. Claverie, and D. Gautheret. Patterns of variant polyadenylation signal usage in human genes. *Genome Res*, 10(7):1001–10, 2000.

[75] H. Caron, B. van Schaik, M. van der Mee, F. Baas, G. Riggins, P. van Sluis, M.C. Hermus, R. van Asperen, K. Boon, P.A. Voute, S. Heisterkamp, A. van Kampen, and R. Versteeg. The human transcriptome map: clustering of highly expressed genes in chromosomal domains. *Science*, 291(5507):1289–92, 2001.

[76] Z. Kan, E.C. Rouchka, W.R. Gish, and D.J. States. Gene structure prediction and alternative splicing analysis using genomically aligned ESTs. *Genome Res*, 11(5):889–900, 2001.

[77] S. Audic and J.M. Claverie. The significance of digital gene expression profiles. *Genome Res*, 7(10):986–95, 1997.

[78] J.M. Claverie. Computational methods for the identification of differential and coordinated gene expression. *Hum Mol Genet*, 8(10):1821–32, 1999.

[79] D.P. Hill, J.A. Blake, J.E. Richardson, and M. Ringwald. Extension and integration of the gene ontology (GO): combining GO vocabularies with external vocabularies. *Genome Res*, 12(12):1982–91, 2002.

[80] A. Bateman, E. Birney, R. Durbin, S.R. Eddy, K.L. Howe, and E.L. Sonnhammer. The Pfam protein families database. *Nucleic Acids Res*, 28(1):263–6, 2000.

[81] L. Falquet, M. Pagni, P. Bucher, N. Hulo, C.J. Sigrist, K. Hofmann, and A. Bairoch. The PROSITE database, its status in 2002. *Nucleic Acids Res*, 30(1):235–8, 2002.

[82] T.K. Attwood, P. Bradley, D.R. Flower, A. Gaulton, N. Maudling, A.L. Mitchell, G. Moulton, A. Nordle, K. Paine, P. Taylor, A. Uddin, and C. Zygouri. PRINTS and its automatic supplement, prePRINTS. *Nucleic Acids Res*, 31(1):400–2, 2003.

[83] S.G. Gregory, M. Sekhon, J. Schein, S. Zhao, K. Osoegawa, C.E. Scott, R.S. Evans, P.W. Burridge, T.V. Cox, C.A. Fox, R.D. Hutton, I.R. Mullenger, K.J. Phillips, J. Smith, J. Stalker, G.J. Threadgold, E. Birney, K. Wylie, A. Chinwalla, J. Wallis, L. Hillier, J. Carter, T. Gaige, S. Jaeger, C. Kremitzki, D. Layman, J. Maas, R. McGrane, K. Mead, R. Walker, S. Jones, M. Smith, J. Asano, I. Bosdet, S. Chan, S. Chittaranjan, R. Chiu, C. Fjell, D. Fuhrmann, N. Girn, C. Gray, R. Guin, L. Hsiao, M. Krzywinski, R. Kutsche, S.S. Lee, C. Mathewson, C. McLeavy, S. Messervier, S. Ness, P. Pandoh, A.L. Prabhu, P. Saeedi, D. Smailus, L. Spence, J. Stott, S. Taylor, W. Terpstra, M. Tsai, J. Vardy, N. Wye, G. Yang, S. Shatsman, B. Ayodeji, K. Geer, G. Tsegaye, A. Shvartsbeyn, E. Gebregeorgis, M. Krol, D. Russell, L. Overton, J.A. Malek, M. Holmes, M. Heaney, J. Shetty, T. Feldblyum, W.C. Nierman, J.J. Catanese, T. Hubbard, R.H. Waterston, J. Rogers, P.J. de Jong, C.M. Fraser, M. Marra, J.D. McPherson, and D.R. Bentley. A physical map of the mouse genome. *Nature*, 418(6899):743–50, 2002.

[84] K.W. Kinzler and B. Vogelstein. The GLI gene encodes a nuclear protein which binds specific sequences in the human genome. *Mol Cell Biol*, 10(2):634–42, 1990.

[85] D.U. Kloos, C. Choi, and E. Wingender. The TGF-beta–Smad network: introducing bioinformatic tools. *Trends Genet*, 18(2):96–103, 2002.

[86] W. Lee, P. Mitchell, and R. Tjian. Purified transcription factor AP-1 interacts with TPA-inducible enhancer elements. *Cell*, 49(6):741–52, 1987.

[87] W. Lee, A. Haslinger, M. Karin, and R. Tjian. Activation of transcription by two factors that bind promoter and enhancer sequences of the human metallothionein gene and SV40. *Nature*, 325(6102):368–72, 1987.

[88] D.M. Benbrook and N.C. Jones. Different binding specificities and transactivation of variant CRE's by CREB complexes. *Nucleic Acids Res*, 22(8):1463–9, 1994.

[89] C.V. Dang. c-Myc target genes involved in cell growth, apoptosis, and metabolism. *Mol Cell Biol*, 19(1):1–11, 1999.

[90] T. Uechi, T. Tanaka, and N. Kenmochi. A complete map of the human ribosomal protein genes: assignment of 80 genes to the cytogenetic map and implications for human disorders. *Genomics*, 72(3):223–30, 2001.

[91] M. Yoshihama, T. Uechi, S. Asakawa, K. Kawasaki, S. Kato, S. Higa, N. Maeda, S. Minoshima, T. Tanaka, N. Shimizu, and N. Kenmochi. The human ribosomal protein genes: sequencing and comparative analysis of 73 genes. *Genome Res*, 12(3):379–90, 2002.

[92] N. Kenmochi, T. Suzuki, T. Uechi, M. Magoori, M. Kuniba, S. Higa, K. Watanabe, and T. Tanaka. The human mitochondrial ribosomal protein genes: mapping of 54 genes to the chromosomes and implications for human disorders. *Genomics*, 77(1-2):65–70, 2001.

[93] Ross Ihaka and Robert Gentleman. R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5(3):299–314, 1996.

[94] A.E. Lash, C.M. Tolstoshev, L. Wagner, G.D. Schuler, R.L. Strausberg, G.J. Riggins, and S.F. Altschul. SAGEmap: a public gene expression resource. *Genome Res*, 10(7):1051–60, 2000.

[95] R. Sorek and H.M. Safer. A novel algorithm for computational identification of contaminated EST libraries. *Nucleic Acids Res*, 31(3):1067–74, 2003.

[96] S. Rogic, A.K. Mackworth, and F.B. Ouellette. Evaluation of gene-finding programs on mammalian sequences. *Genome Res*, 11(5):817–32, 2001.

[97] D.R. Maglott, K.S. Katz, H. Sicotte, and K.D. Pruitt. NCBI's LocusLink and RefSeq. *Nucleic Acids Res*, 28(1):126–8, 2000.

[98] J. Jendrisak, R.A. Young, and J.D. Engel. Cloning cDNA into lambda gt10 and lambda gt11. *Methods Enzymol*, 152:359–71, 1987.

[99] S.V. Anisimov, K.V. Tarasov, M.D. Stern, E.G. Lakatta, and K.R. Boheler. A quantitative and validated SAGE transcriptome reference for adult mouse heart. *Genomics*, 80(2):213–22, 2002.

[100] S.V. Anisimov, K.V. Tarasov, D. Riordon, A.M. Wobus, and K.R. Boheler. SAGE identification of differentiation responsive genes in P19 embryonic cells induced to form cardiomyocytes in vitro. *Mech Dev*, 117(1-2):25–74, 2002.

[101] S.V. Anisimov, K.V. Tarasov, D. Tweedie, M.D. Stern, A.M. Wobus, and K.R. Boheler. SAGE identification of gene transcripts with profiles unique to pluripotent mouse R1 embryonic stem cells. *Genomics*, 79(2):169–76, 2002.

[102] R. Chrast, H.S. Scott, M.P. Papasavvas, C. Rossier, E.S. Antonarakis, C. Barras, M.T. Davisson, C. Schmidt, X. Estivill, M. Dierssen, M. Pritchard, and S.E. Antonarakis. The mouse brain transcriptome by SAGE: differences in gene expression between P30 brains of the partial trisomy 16 mouse model of Down syndrome (Ts65Dn) and normals. *Genome Res*, 10(12):2006–21, 2000.

[103] D. Zelenika, E. Adams, S. Humm, L. Graca, S. Thompson, S.P. Cobbold, and H. Waldmann. Regulatory T cells overexpress a subset of Th2 gene transcripts. *J Immunol*, 168(3):1069–79, 2002.

[104] L. Graca, S. Thompson, C.Y. Lin, E. Adams, S.P. Cobbold, and H. Waldmann. Both CD4(+)CD25(+) and CD4(+)CD25(-) regulatory cells mediate dominant transplantation tolerance. *J Immunol*, 168(11):5558–65, 2002.

[105] P.J. O'Shaughnessy, L. Fleming, P.J. Baker, G. Jackson, and H. Johnston. Identification of Developmentally-Regulated Genes in the Somatic Cells of the Mouse Testis Using Serial Analysis of Gene Expression. *Biol Reprod*, 2003.

[106] P.H. Crossley and G.R. Martin. The mouse Fgf8 gene encodes a family of polypeptides and is expressed in regions that direct outgrowth and patterning in the developing embryo. *Development*, 121(2):439–51, 1995.

[107] L. Pissarra, D. Henrique, and A. Duarte. Expression of hes6, a new member of the Hairy/Enhancer-of-split family, in mouse development. *Mech Dev*, 95(1-2):275–8, 2000.

[108] M.A. Nieto, P. Gilardi-Hebenstreit, P. Charnay, and D.G. Wilkinson. A receptor protein tyrosine kinase implicated in the segmental patterning of the hindbrain and mesoderm. *Development*, 116(4):1137–50, 1992.

[109] S.H. Kim, W.C. Jen, E.M. De Robertis, and C. Kintner. The protocadherin PAPC establishes segmental boundaries during somitogenesis in xenopus embryos. *Curr Biol*, 10(14):821–30, 2000.

[110] A. Mansouri and P. Gruss. Pax3 and Pax7 are expressed in commissural neurons and restrict ventral neuronal identity in the spinal cord. *Mech Dev*, 78(1-2):171–8, 1998.

[111] F. Kraus, B. Haenig, and A. Kispert. Cloning and expression analysis of the mouse T-box gene Tbx18. *Mech Dev*, 100(1):83–6, 2001.

[112] U. Deutsch, G.R. Dressler, and P. Gruss. Pax 1, a member of a paired box homologous murine gene family, is expressed in segmented structures during development. *Cell*, 53(4):617–25, 1988.

[113] V. Wilson, L. Manson, W.C. Skarnes, and R.S. Beddington. The T gene is necessary for normal mesodermal morphogenetic cell movements during gastrulation. *Development*, 121(3):877–86, 1995.

[114] A.F. Candia, J. Hu, J. Crosby, P.A. Lalley, D. Noden, J.H. Nadeau, and C.V. Wright. Mox-1 and Mox-2 define a novel homeobox gene subfamily and are differentially expressed during early mesodermal patterning in mouse embryos. *Development*, 116(4):1123–36, 1992.

[115] G.E. Winnier, L. Hargett, and B.L. Hogan. The winged helix transcription factor MFH1 is required for proliferation and patterning of paraxial mesoderm in the mouse embryo. *Genes Dev*, 11(7):926–40, 1997.

[116] S.H. Johnston, C. Rauskolb, R. Wilson, B. Prabhakaran, K.D. Irvine, and T.F. Vogt. A family of mammalian Fringe genes implicated in boundary determination and the Notch pathway. *Development*, 124(11):2245–54, 1997.

[117] Y. Saga, N. Hata, H. Koseki, and M.M. Taketo. Mesp2: a novel mouse gene expressed in the presegmented mesoderm and essential for segmentation initiation. *Genes Dev*, 11(14):1827–39, 1997.

[118] D.G. Spinella, A.K. Bernardino, A.C. Redding, P. Koutz, Y. Wei, E.K. Pratt, K.K. Myers, G. Chappell, S. Gerken, and S.J. McConnell. Tandem arrayed ligation of expressed sequence tags (TALEST): a new method for generating global gene expression profiles. *Nucleic Acids Res*, 27(18):e22, 1999.

[119] N.J. Mulder, R. Apweiler, T.K. Attwood, A. Bairoch, D. Barrell, A. Bateman, D. Binns, M. Biswas, P. Bradley, P. Bork, P. Bucher, R.R. Copley, E. Courcelle, U. Das, R. Durbin, L. Falquet, W. Fleischmann, S. Griffiths-Jones, D. Haft, N. Harte, N. Hulo, D. Kahn, A. Kanapin, M. Krestyaninova, R. Lopez, I. Letunic, D. Lonsdale, V. Silventoinen, S.E. Orchard, M. Pagni, D. Peyruc, C.P. Ponting, J.D. Selengut, F. Servant, C.J. Sigrist, R. Vaughan, and E.M. Zdobnov. The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res*, 31(1):315–8, 2003.

[120] F. Corpet, J. Gouzy, and D. Kahn. Recent improvements of the ProDom database of protein domain families. *Nucleic Acids Res*, 27(1):263–7, 1999.

[121] I. Letunic, L. Goodstadt, N.J. Dickens, T. Doerks, J. Schultz, R. Mott, F. Ciccarelli, R.R. Copley, C.P. Ponting, and P. Bork. Recent improvements to the SMART domain-based sequence annotation resource. *Nucleic Acids Res*, 30(1):242–4, 2002.

[122] D.H. Haft, B.J. Loftus, D.L. Richardson, F. Yang, J.A. Eisen, I.T. Paulsen, and O. White. TIGRFAMs: a protein family resource for the functional identification of proteins. *Nucleic Acids Res*, 29(1):41–3, 2001.

[123] A. Menssen and H. Hermeking. Characterization of the c-MYC-regulated transcriptome by SAGE: identification and analysis of c-MYC target genes. *Proc Natl Acad Sci U S A*, 99(9):6274–9, 2002.

[124] D. Brett, H. Pospisil, J. Valcarcel, J. Reich, and P. Bork. Alternative splicing and genome complexity. *Nat Genet*, 30(1):29–30, 2002.

[125] E. Beaudoing and D. Gautheret. Identification of alternate polyadenylation sites and analysis of their tissue distribution using EST data. *Genome Res*, 11(9):1520–6, 2001.

[126] E. Pauws, A.H. van Kampen, S.A. van de Graaf, J.J. de Vijlder, and C. Ris-Stalpers. Heterogeneity in polyadenylation cleavage sites in mammalian mRNA sequences: implications for SAGE analysis. *Nucleic Acids Res*, 29(8):1690–4, 2001.

[127] J.A. Bailey, A.M. Yavor, H.F. Massa, B.J. Trask, and E.E. Eichler. Segmental duplications: organization and impact within the current human genome project assembly. *Genome Res*, 11(6):1005–17, 2001.

[128] A.J. Mighell, N.R. Smith, P.A. Robinson, and A.F. Markham. Vertebrate pseudogenes. *FEBS Lett*, 468(2-3):109–14, 2000.

[129] M. Alexandersson, S. Cawley, and L. Pachter. SLAM: cross-species gene finding and alignment with a generalized pair hidden Markov model. *Genome Res*, 13(3):496–502, 2003.

[130] V. de Waard, B.M. van den Berg, J. Veken, R. Schultz-Heienbrok, H. Pannekoek, and A.J. van Zonneveld. Serial analysis of gene expression to assess the endothelial cell response to an atherogenic stimulus. *Gene*, 226(1):1–8, 1999.

[131] M. Kenzelmann and K. Muhlemann. Transcriptome analysis of fibroblast cells immediate-early after human cytomegalovirus infection. *J Mol Biol*, 304(5):741–51, 2000.

[132] W.M. Kulyk, W.B. Upholt, and R.A. Kosher. Fibronectin gene expression during limb cartilage differentiation. *Development*, 106(3):449–55, 1989.

[133] H. Sage, R.B. Vernon, J. Decker, S. Funk, and M.L. Iruela-Arispe. Distribution of the calcium-binding protein SPARC in tissues of embryonic and adult mice. *J Histochem Cytochem*, 37(6):819–29, 1989.

[134] N.C. Zanetti and M. Solursh. Induction of chondrogenesis in limb mesenchymal cultures by disruption of the actin cytoskeleton. *J Cell Biol*, 99(1 Pt 1):115–23, 1984.

[135] J. Skonier, M. Neubauer, L. Madisen, K. Bennett, G.D. Plowman, and A.F. Purchio. cDNA cloning and sequence analysis of beta ig-h3, a novel gene induced in a human adenocarcinoma cell line after treatment with transforming growth factor-beta. *DNA Cell Biol*, 11(7):511–22, 1992.

[136] K. Hashimoto, M. Noshiro, S. Ohno, T. Kawamoto, H. Satakeda, Y. Akagawa, K. Nakashima, A. Okimura, H. Ishida, T. Okamoto, H. Pan, M. Shen, W. Yan, and Y. Kato. Characterization of a cartilage-derived 66-kDa protein (RGD-CAP/beta ig-h3) that binds to collagen. *Biochim Biophys Acta*, 1355(3):303–14, 1997.

[137] T. Ogata, J.M. Wozney, R. Benezra, and M. Noda. Bone morphogenetic protein 2 transiently enhances expression of a gene, Id (inhibitor of differentiation), encoding a helix-loop-helix molecule in osteoblast-like cells. *Proc Natl Acad Sci U S A*, 90(19):9219–22, 1993.

[138] A. Hollnagel, V. Oehlmann, J. Heymer, U. Ruther, and A. Nordheim. Id genes are direct targets of bone morphogenetic protein induction in embryonic stem cells. *J Biol Chem*, 274(28):19838–45, 1999.

[139] A. Oshima, H. Tanabe, T. Yan, G.N. Lowe, C.A. Glackin, and A. Kudo. A novel mechanism for the regulation of osteoblast differentiation: transcription of periostin, a member of the fasciclin I family, is regulated by the bHLH transcription factor, twist. *J Cell Biochem*, 86(4):792–804, 2002.

[140] C.S. Lee, L. Buttitta, and C.M. Fan. Evidence that the WNT-inducible growth arrest-specific gene 1 encodes an antagonist of sonic hedgehog signaling in the somite. *Proc Natl Acad Sci U S A*, 98(20):11347–52, 2001.

[141] K.K. Lee, A.K. Leung, M.K. Tang, D.Q. Cai, C. Schneider, C. Brancolini, and P.H. Chow. Functions of the growth arrest specific 1 gene in the development of the mouse embryo. *Dev Biol*, 234(1):188–203, 2001.

[142] K. Boon, H.N. Caron, R. van Asperen, L. Valentijn, M.C. Hermus, P. van Sluis, I. Roobeek, I. Weis, P.A. Voute, M. Schwab, and R. Versteeg. N-myc enhances the expression of a large set of genes functioning in ribosome biogenesis and protein synthesis. *EMBO J*, 20(6):1383–93, 2001.

[143] I.G. Wool. Extraribosomal functions of ribosomal proteins. *Trends Biochem Sci*, 21(5):164–5, 1996.

[144] D.C. Weinstein, E. Honore, and A. Hemmati-Brivanlou. Epidermal induction and inhibition of neural fate by translation initiation factor 4AIII. *Development*, 124(21):4235–42, 1997.

[145] J. Shou, P.C. Rim, and A.L. Calof. BMPs inhibit neurogenesis by a mechanism involving degradation of a transcription factor. *Nat Neurosci*, 2(4):339–45, 1999.

[146] H. Zhu, P. Kavsak, S. Abdollah, J.L. Wrana, and G.H. Thomsen. A SMAD ubiquitin ligase targets the BMP pathway and affects embryonic pattern formation. *Nature*, 400(6745):687–93, 1999.

[147] D.M. Ornitz and N. Itoh. Fibroblast growth factors. *Genome Biol*, 2(3):REVIEWS3005, 2001.

[148] J. Dubrulle, M.J. McGrew, and O. Pourquie. FGF signaling controls somite boundary position and regulates segmentation clock control of spatiotemporal Hox gene activation. *Cell*, 106(2):219–32, 2001.

[149] A. Orr-Urtreger, D. Givol, A. Yayon, Y. Yarden, and P. Lonai. Developmental expression of two murine fibroblast growth factor receptors, flg and bek. *Development*, 113(4):1419–34, 1991.

[150] G. Patstone, E.B. Pasquale, and P.A. Maher. Different members of the fibroblast growth factor receptor family are specific to distinct cell types in the developing chicken embryo. *Dev Biol*, 155(1):107–23, 1993.

[151] B.G. Ciruna, L. Schwartz, K. Harpal, T.P. Yamaguchi, and J. Rossant. Chimeric analysis of fibroblast growth factor receptor-1 (Fgfr1) function: a role for FGFR1 in morphogenetic movement through the primitive streak. *Development*, 124(14):2829–41, 1997.

[152] J. Partanen, L. Schwartz, and J. Rossant. Opposite phenotypes of hypomorphic and Y766 phosphorylation site mutations reveal a function for Fgfr1 in anteroposterior patterning of mouse embryos. *Genes Dev*, 12(15):2332–44, 1998.

[153] P. Klint and L. Claesson-Welsh. Signal transduction by fibroblast growth factor receptors. *Front Biosci*, 4:D165–77, 1999.

[154] B. Boilly, A.S. Vercoutter-Edouart, H. Hondermarck, V. Nurcombe, and X. Le Bourhis. FGF signals for cell proliferation and migration through different pathways. *Cytokine Growth Factor Rev*, 11(4):295–302, 2000.

[155] A. Sawada, M. Shinya, Y.J. Jiang, A. Kawakami, A. Kuroiwa, and H. Takeda. Fgf/MAPK signalling is a crucial positional cue in somite boundary formation. *Development*, 128(23):4873–80, 2001.

[156] L.B. Corson, Y. Yamanaka, K.M. Lai, and J. Rossant. Spatial and temporal patterns of ERK signaling during mouse embryogenesis. *Development*, 130(19):4527–37, 2003.

[157] A. Wodarz and R. Nusse. Mechanisms of Wnt signaling in development. *Annu Rev Cell Dev Biol*, 14:59–88, 1998.

[158] M. Povelones and R. Nusse. Wnt signalling sees spots. *Nat Cell Biol*, 4(11):E249–50, 2002.

[159] S. Takada, K.L. Stark, M.J. Shea, G. Vassileva, J.A. McMahon, and A.P. McMahon. Wnt-3a regulates somite and tailbud formation in the mouse embryo. *Genes Dev*, 8(2):174–89, 1994.

[160] T.L. Greco, S. Takada, M.M. Newhouse, J.A. McMahon, A.P. McMahon, and S.A. Camper. Analysis of the vestigial tail mutation demonstrates that Wnt-3a gene dosage regulates mouse axial development. *Genes Dev*, 10(3):313–24, 1996.

[161] A. Aulehla, C. Wehrle, B. Brand-Saberi, R. Kemler, A. Gossler, B. Kanzler, and B.G. Herrmann. Wnt3a plays a major role in the segmentation clock controlling somitogenesis. *Dev Cell*, 4(3):395–406, 2003.

[162] Y. Wang, N. Thekdi, P.M. Smallwood, J.P. Macke, and J. Nathans. Frizzled-3 is required for the development of major fiber tracts in the rostral CNS. *J Neurosci*, 22(19):8563–73, 2002.

[163] Y. Wang, D. Huso, H. Cahill, D. Ryugo, and J. Nathans. Progressive cerebellar, auditory, and esophageal dysfunction caused by targeted disruption of the frizzled-4 gene. *J Neurosci*, 21(13):4761–71, 2001.

[164] T. Ishikawa, Y. Tamai, A.M. Zorn, H. Yoshida, M.F. Seldin, S. Nishikawa, and M.M. Taketo. Mouse Wnt receptor gene Fzd5 is essential for yolk sac and placental angiogenesis. *Development*, 128(1):25–33, 2001.

[165] J. Herz, D.E. Clouthier, and R.E. Hammer. LDL receptor-related protein internalizes and degrades uPA-PAI-1 complexes and is essential for embryo implantation. *Cell*, 71(3):411–21, 1992.

[166] K.I. Pinson, J. Brennan, S. Monkley, B.J. Avery, and W.C. Skarnes. An LDL-receptor-related protein mediates Wnt signalling in mice. *Nature*, 407(6803):535–8, 2000.

[167] M. Baron. An overview of the Notch signalling pathway. *Semin Cell Dev Biol*, 14(2):113–9, 2003.

[168] T. Iso, L. Kedes, and Y. Hamamori. HES and HERP families: multiple effectors of the Notch signaling pathway. *J Cell Physiol*, 194(3):237–55, 2003.

[169] N.J. Justice and Y.N. Jan. Variations on the Notch pathway in neural development. *Curr Opin Neurobiol*, 12(1):64–70, 2002.

[170] F.F. Del Amo, D.E. Smith, P.J. Swiatek, M. Gendron-Maguire, R.J. Greenspan, A.P. McMahon, and T. Gridley. Expression pattern of Motch, a mouse homolog of Drosophila Notch, suggests an important role in early postimplantation mouse development. *Development*, 115(3):737–44, 1992.

[171] B. Bettenhausen, M. Hrabe de Angelis, D. Simon, J.L. Guenet, and A. Gossler. Transient and restricted expression during mouse embryogenesis of Dll1, a murine gene closely related to Drosophila Delta. *Development*, 121(8):2407–18, 1995.

[172] S.L. Dunwoodie, D. Henrique, S.M. Harrison, and R.S. Beddington. Mouse Dll3: a novel divergent Delta gene which may complement the function of other Delta homologues during early pattern formation in the mouse embryo. *Development*, 124(16):3065–76, 1997.

[173] R.A. Conlon, A.G. Reaume, and J. Rossant. Notch1 is required for the coordinate segmentation of somites. *Development*, 121(5):1533–45, 1995.

[174] M. Hrabe de Angelis, J. McIntyre, 2nd, and A. Gossler. Maintenance of somite borders in mice requires the Delta homologue DII1. *Nature*, 386(6626):717–21, 1997.

[175] K. Kusumi, E.S. Sun, A.W. Kerrebrock, R.T. Bronson, D.C. Chi, M.S. Bulotsky, J.B. Spencer, B.W. Birren, W.N. Frankel, and E.S. Lander. The mouse pudgy mutation disrupts Delta homologue Dll3 and initiation of early somite boundaries. *Nat Genet*, 19(3):274–8, 1998.

[176] N. Zhang and T. Gridley. Defects in somite formation in lunatic fringe-deficient mice. *Nature*, 394(6691):374–7, 1998.

[177] T.A. Mitsiadis, D. Henrique, I. Thesleff, and U. Lendahl. Mouse Serrate-1 (Jagged-1): expression in the developing tooth is regulated by epithelial-mesenchymal interactions and fibroblast growth factor-4. *Development*, 124(8):1473–83, 1997.

[178] Y. Xue, X. Gao, C.E. Lindsell, C.R. Norton, B. Chang, C. Hicks, M. Gendron-Maguire, E.B. Rand, G. Weinmaster, and T. Gridley. Embryonic lethality and vascular defects in mice lacking the Notch ligand Jagged1. *Hum Mol Genet*, 8(5):723–30, 1999.

[179] J. Shen, R.T. Bronson, D.F. Chen, W. Xia, D.J. Selkoe, and S. Tonegawa. Skeletal and CNS defects in Presenilin-1-deficient mice. *Cell*, 89(4):629–39, 1997.

[180] K. Koizumi, M. Nakajima, S. Yuasa, Y. Saga, T. Sakai, T. Kuriyama, T. Shirasawa, and H. Koseki. The role of presenilin 1 during somite segmentation. *Development*, 128(8):1391–402, 2001.

[181] C. Oka, T. Nakano, A. Wakeham, J.L. de la Pompa, C. Mori, T. Sakai, S. Okazaki, M. Kawaichi, K. Shiota, T.W. Mak, and T. Honjo. Disruption of the mouse RBP-J kappa gene results in early embryonic death. *Development*, 121(10):3291–301, 1995.

[182] C. Jouve, I. Palmeirim, D. Henrique, J. Beckers, A. Gossler, D. Ish-Horowicz, and O. Pourquie. Notch signalling is required for cyclic expression of the hairy-like gene HES1 in the presomitic mesoderm. *Development*, 127(7):1421–9, 2000.

[183] Y. Bessho, R. Sakata, S. Komatsu, K. Shiota, S. Yamada, and R. Kageyama. Dynamic expression and essential functions of Hes7 in somite segmentation. *Genes Dev*, 15(20):2642–7, 2001.

[184] C.G. Lobe. Expression of the helix-loop-helix factor, Hes3, during embryo development suggests a role in early midbrain-hindbrain patterning. *Mech Dev*, 62(2):227–37, 1997.

[185] Y. Bessho, H. Hirata, Y. Masamizu, and R. Kageyama. Periodic repression by the bHLH factor Hes7 is an essential mechanism for the somite segmentation clock. *Genes Dev*, 17(12):1451–6, 2003.

[186] M. Ishibashi, S.L. Ang, K. Shiota, S. Nakanishi, R. Kageyama, and F. Guillemot. Targeted disruption of mammalian hairy and Enhancer of split homolog-1 (HES-1) leads to up-regulation of neural helix-loop-helix factors, premature neurogenesis, and severe neural tube defects. *Genes Dev*, 9(24):3136–48, 1995.

[187] H. Hirata, K. Tomita, Y. Bessho, and R. Kageyama. Hes1 and Hes3 regulate maintenance of the isthmic organizer and development of the mid/hindbrain. *EMBO J*, 20(16):4454–66, 2001.

[188] T. Ohtsuka, M. Ishibashi, G. Gradwohl, S. Nakanishi, F. Guillemot, and R. Kageyama. Hes1 and Hes5 as notch effectors in mammalian neuronal differentiation. *EMBO J*, 18(8):2196–207, 1999.

[189] Y. Saga, N. Hata, S. Kobayashi, T. Magnuson, M.F. Seldin, and M.M. Taketo. MesP1: a novel basic helix-loop-helix protein expressed in the nascent mesodermal cells during mouse gastrulation. *Development*, 122(9):2769–78, 1996.

[190] Y. Saga. Genetic rescue of segmentation defect in MesP2-deficient mice by MesP1 gene replacement. *Mech Dev*, 75(1-2):53–66, 1998.

[191] G. Chen and A.J. Courey. Groucho/TLE family proteins and transcriptional repression. *Gene*, 249(1-2):1–16, 2000.

[192] M. Mallo, M. Gendron-Maguire, M.L. Harbison, and T. Gridley. Protein characterization and targeted disruption of Grg, a mouse gene related to the groucho transcript of the Drosophila Enhancer of split complex. *Dev Dyn*, 204(3):338–47, 1995.

[193] J.L. Duband, S. Dufour, K. Hatta, M. Takeichi, G.M. Edelman, and J.P. Thiery. Adhesion molecules during somitogenesis in the avian embryo. *J Cell Biol*, 104(5):1361–74, 1987.

[194] P.M. Kulesa and S.E. Fraser. Cell dynamics during somite boundary formation revealed by time-lapse analysis. *Science*, 298(5595):991–5, 2002.

[195] A.C. Erickson and J.R. Couchman. Still more complexity in mammalian basement membranes. *J Histochem Cytochem*, 48(10):1291–306, 2000.

[196] S. Li, D. Edgar, R. Fassler, W. Wadsworth, and P.D. Yurchenco. The role of laminin in embryonic cell polarization and tissue organization. *Dev Cell*, 4(5):613–24, 2003.

[197] E.L. George, E.N. Georges-Labouesse, R.S. Patel-King, H. Rayburn, and R.O. Hynes. Defects in mesoderm, neural tube and vascular development in mouse embryos lacking fibronectin. *Development*, 119(4):1079–91, 1993.

[198] E.A. Chernoff, D.O. Clarke, J.L. Wallace-Evers, L.P. Hungate-Muegge, and R.C. Smith. The effects of collagen synthesis inhibitory drugs on somitogenesis and myogenin expression in cultured chick and mouse embryos. *Tissue Cell*, 33(1):97–110, 2001.

[199] R.O. Hynes. Targeted mutations in cell adhesion genes: what have we learned from them? *Dev Biol*, 180(2):402–12, 1996.

[200] R.O. Hynes. Cell adhesion: old and new questions. *Trends Cell Biol*, 9(12):M33–7, 1999.

[201] A.M. Belkin and M.A. Stepp. Integrins as receptors for laminins. *Microsc Res Tech*, 51(3):280–301, 2000.

[202] D. Sheppard. In vivo functions of integrins: lessons from null mutations in mice. *Matrix Biol*, 19(3):203–9, 2000.

[203] C.J. Drake and C.D. Little. Integrins play an essential role in somite adhesion to the embryonic axis. *Dev Biol*, 143(2):418–21, 1991.

[204] J.P. Thiery. Epithelial-mesenchymal transitions in tumour progression. *Nat Rev Cancer*, 2(6):442–54, 2002.

[205] J.E. Rabinowitz, U. Rutishauser, and T. Magnuson. Targeted mutation of Ncam to produce a secreted molecule results in a dominant embryonic lethality. *Proc Natl Acad Sci U S A*, 93(13):6421–4, 1996.

[206] K.K. Linask, C. Ludwig, M.D. Han, X. Liu, G.L. Radice, and K.A. Knudsen. N-cadherin/catenin-mediated morphoregulation of somite formation. *Dev Biol*, 202(1):85–102, 1998.

[207] G.L. Radice, H. Rayburn, H. Matsunami, K.A. Knudsen, M. Takeichi, and R.O. Hynes. Developmental defects in mouse embryos lacking N-cadherin. *Dev Biol*, 181(1):64–78, 1997.

[208] K. Horikawa, G. Radice, M. Takeichi, and O. Chisaka. Adhesive subdivisions intrinsic to the epithelial somites. *Dev Biol*, 215(2):182–9, 1999.

[209] J. Rhee, Y. Takahashi, Y. Saga, J. Wilson-Rawls, and A. Rawls. The protocadherin papc is involved in the organization of the epithelium along the segmental border during mouse somitogenesis. *Dev Biol*, 254(2):248–61, 2003.

[210] A. Yamamoto, C. Kemp, D. Bachiller, D. Geissert, and E.M. De Robertis. Mouse paraxial protocadherin is expressed in trunk mesoderm and is not essential for mouse development. *Genesis*, 27(2):49–57, 2000.

[211] T. Wittmann and C.M. Waterman-Storer. Cell motility: can Rho GTPases and microtubules point the way? *J Cell Sci*, 114(Pt 21):3795–803, 2001.

[212] O.C. Rodriguez, A.W. Schaefer, C.A. Mandato, P. Forscher, W.M. Bement, and C.M. Waterman-Storer. Conserved microtubule-actin interactions in cell movement and morphogenesis. *Nat Cell Biol*, 5(7):599–609, 2003.

[213] P.J. Roy, J.M. Stuart, J. Lund, and S.K. Kim. Chromosomal clustering of muscle-expressed genes in Caenorhabditis elegans. *Nature*, 418(6901):975–9, 2002.

[214] P.T. Spellman and G.M. Rubin. Evidence for large domains of similarly expressed genes in the Drosophila genome. *J Biol*, 1(1):5, 2002.

[215] A.M. Boutanaev, A.I. Kalmykova, Y.Y. Shevelyov, and D.I. Nurminsky. Large clusters of co-expressed genes in the Drosophila genome. *Nature*, 420(6916):666–9, 2002.

[216] G. Felsenfeld and M. Groudine. Controlling the double helix. *Nature*, 421(6921):448–53, 2003.

[217] T. Agalioti, S. Lomvardas, B. Parekh, J. Yie, T. Maniatis, and D. Thanos. Ordered recruitment of chromatin modifying and general transcription factors to the IFN-beta promoter. *Cell*, 103(4):667–78, 2000.

[218] H. Boeger, J. Griesenbeck, J.S. Strattan, and R.D. Kornberg. Nucleosomes unfold completely at a transcriptionally active promoter. *Mol Cell*, 11(6):1587–98, 2003.

[219] J.M. Angelastro, B. Torocsik, and L.A. Greene. Nerve growth factor selectively regulates expression of transcripts encoding ribosomal proteins. *BMC Neurosci*, 3(1):3, 2002.

[220] D. Nadano, G. Ishihara, C. Aoki, T. Yoshinaka, S. Irie, and T.A. Sato. Preparation and characterization of antibodies against human ribosomal proteins: heterogeneous expression of S11 and S30 in a panel of human cancer cell lines. *Jpn J Cancer Res*, 91(8):802–10, 2000.

[221] H. Kasai, D. Nadano, E. Hidaka, K. Higuchi, M. Kawakubo, T.A. Sato, and J. Nakayama. Differential expression of ribosomal proteins in human normal and neoplastic colorectum. *J Histochem Cytochem*, 51(5):567–74, 2003.

[222] A. Fatica and D. Tollervey. Making ribosomes. *Curr Opin Cell Biol*, 14(3):313–8, 2002.

[223] V.P. Mauro and G.M. Edelman. The ribosome filter hypothesis. *Proc Natl Acad Sci U S A*, 99(19):12031–6, 2002.

[224] M.G. Coppolino, M.J. Woodside, N. Demaurex, S. Grinstein, R. St-Arnaud, and S. Dedhar. Calreticulin is essential for integrin-mediated calcium signalling and cell adhesion. *Nature*, 386(6627):843–7, 1997.

[225] K. Burns, B. Duggan, E.A. Atkinson, K.S. Famulski, M. Nemer, R.C. Bleackley, and M. Michalak. Modulation of gene expression by calreticulin binding to the glucocorticoid receptor. *Nature*, 367(6462):476–80, 1994.

[226] B. Wiedmann, H. Sakai, T.A. Davis, and M. Wiedmann. A protein complex required for signal-sequence-specific sorting and translocation. *Nature*, 370(6489):434–40, 1994.

[227] A. Moreau, W.V. Yotov, F.H. Glorieux, and R. St-Arnaud. Bone-specific expression of the alpha chain of the nascent polypeptide-associated complex, a coactivator potentiating c-Jun-mediated transcription. *Mol Cell Biol*, 18(3):1312–21, 1998.

[228] Y. Yokota. Id and development. *Oncogene*, 20(58):8290–8, 2001.

[229] J.D. Norton, R.W. Deed, G. Craggs, and F. Sablitzky. Id helix-loop-helix proteins in cell growth and differentiation. *Trends Cell Biol*, 8(2):58–65, 1998.

[230] Z. Karetsou, A. Kretsovali, C. Murphy, O. Tsolas, and T. Papamarcaki. Prothymosin alpha interacts with the CREB-binding protein and potentiates transcription. *EMBO Rep*, 3(4):361–6, 2002.

[231] M. Bustin. Chromatin unfolding and activation by HMGN(*) chromosomal proteins. *Trends Biochem Sci*, 26(7):431–7, 2001.

[232] C. Bischoff, S. Kahns, A. Lund, H.F. Jorgensen, M. Praestegaard, B.F. Clark, and H. Leffers. The human elongation factor 1 A-2 gene (EEF1A2): complete sequence and characterization of gene structure and promoter activity. *Genomics*, 68(1):63–70, 2000.

[233] R. Purohit, D. McCormick, and J. Dyson. Multiple translation initiation factor Sui1 related sequences in mammalian genomes. *Mamm Genome*, 7(1):79–80, 1996.

[234] H. Yoon and T.F. Donahue. Control of translation initiation in Saccharomyces cerevisiae. *Mol Microbiol*, 6(11):1413–9, 1992.

[235] S. Yamanaka, K.S. Poksay, K.S. Arnold, and T.L. Innerarity. A novel translational repressor mRNA is edited extensively in livers containing tumors caused by the transgene expression of the apoB mRNA-editing enzyme. *Genes Dev*, 11(3):321–33, 1997.

[236] N. Nagai, M. Hosokawa, S. Itohara, E. Adachi, T. Matsushita, N. Hosokawa, and K. Nagata. Embryonic lethality of molecular chaperone hsp47 knockout mice is associated with defects in collagen biosynthesis. *J Cell Biol*, 150(6):1499–506, 2000.

[237] L.M. Hendershot, V.A. Valentine, A.S. Lee, S.W. Morris, and D.N. Shapiro. Localization of the gene encoding human BiP/GRP78, the endoplasmic reticulum cognate of the HSP70 family, to chromosome 9q34. *Genomics*, 20(2):281–4, 1994.

[238] K.E. Matlack, W. Mothes, and T.A. Rapoport. Protein translocation: tunnel vision. *Cell*, 92(3):381–90, 1998.

[239] S. Wera and B.A. Hemmings. Serine/threonine protein phosphatases. *Biochem J*, 311 (Pt 1):17–29, 1995.

[240] R.C. Conaway, C.S. Brower, and J.W. Conaway. Emerging roles of ubiquitin in transcription regulation. *Science*, 296(5571):1254–8, 2002.

[241] Y. Ye, H.H. Meyer, and T.A. Rapoport. The AAA ATPase Cdc48/p97 and its partners transport proteins from the ER into the cytosol. *Nature*, 414(6864):652–6, 2001.

[242] K. Romisch. Surfing the Sec61 channel: bidirectional protein translocation across the ER membrane. *J Cell Sci*, 112 (Pt 23):4185–91, 1999.

[243] K.D. Wilkinson, V.L. Tashayev, L.B. O'Connor, C.N. Larsen, E. Kasperek, and C.M. Pickart. Metabolism of the polyubiquitin degradation signal: structure, mechanism, and role of isopeptidase T. *Biochemistry*, 34(44):14535–46, 1995.

[244] T.E. Hawkins, C.J. Merrifield, and S.E. Moss. Calcium signaling and annexins. *Cell Biochem Biophys*, 33(3):275–96, 2000.

[245] J.A. Hammer, 3rd and X.S. Wu. Rabs grab motors: defining the connections between Rab GTPases and motor proteins. *Curr Opin Cell Biol*, 14(1):69–75, 2002.

[246] C.R. Hunt, A.J. Parsian, P.C. Goswami, and C.A. Kozak. Characterization and expression of the mouse Hsc70 gene. *Biochim Biophys Acta*, 1444(3):315–25, 1999.

[247] I.T. Pleasure, M.M. Black, and J.H. Keen. Valosin-containing protein, VCP, is a ubiquitous clathrin-binding protein. *Nature*, 365(6445):459–62, 1993.

[248] Y. Takai, T. Sasaki, and T. Matozaki. Small GTP-binding proteins. *Physiol Rev*, 81(1):153–208, 2001.

[249] M.A. Crackower, S.W. Scherer, J.M. Rommens, C.C. Hui, P. Poorkaj, S. Soder, J.M. Cobben, L. Hudgins, J.P. Evans, and L.C. Tsui. Characterization of the split hand/split foot malformation locus SHFM1 at 7q21.3-q22.1 and analysis of a candidate gene for its expression during limb development. *Hum Mol Genet*, 5(5):571–9, 1996.

[250] J. Jantti, J. Lahdenranta, V.M. Olkkonen, H. Soderlund, and S. Keranen. SEM1, a homologue of the split hand/split foot malformation candidate gene Dss1, regulates exocytosis and pseudohyphal differentiation in yeast. *Proc Natl Acad Sci U S A*, 96(3):909–14, 1999.

[251] P.M. Harrison and P. Arosio. The ferritins: molecular properties, iron storage function and cellular regulation. *Biochim Biophys Acta*, 1275(3):161–203, 1996.

[252] J.L. Loeffen, R.H. Triepels, L.P. van den Heuvel, M. Schuelke, C.A. Buskens, R.J. Smeets, J.M. Trijbels, and J.A. Smeitink. cDNA of eight nuclear encoded subunits of NADH:ubiquinone oxidoreductase: human complex I cDNA characterization completed. *Biochem Biophys Res Commun*, 253(2):415–22, 1998.

[253] C. Couldrey, M.B. Carlton, J. Ferrier, W.H. Colledge, and M.J. Evans. Disruption of murine alpha-enolase by a retroviral gene trap results in early embryonic lethality. *Dev Dyn*, 212(2):284–92, 1998.

[254] R. Sandulache, W. Pretsch, B. Chatterjee, W. Gimbel, J. Graw, and J. Favor. Molecular analysis of four lactate dehydrogenase-A mutants in the mouse. *Mamm Genome*, 5(12):777–80, 1994.

[255] S.H. Jo, M.K. Son, H.J. Koh, S.M. Lee, I.H. Song, Y.O. Kim, Y.S. Lee, K.S. Jeong, W.B. Kim, J.W. Park, B.J. Song, T.L. Huh, and T.L. Huhe. Control of mitochondrial redox balance and cellular defense against oxidative damage by mitochondrial NADP+-dependent isocitrate dehydrogenase. *J Biol Chem*, 276(19):16168–76, 2001.

[256] J.V. Bonventre. The 85-kD cytosolic phospholipase A2 knockout mouse: a new tool for physiology and cell biology. *J Am Soc Nephrol*, 10(2):404–12, 1999.

[257] Y. Sotsios, G.C. Whittaker, J. Westwick, and S.G. Ward. The CXC chemokine stromal cell-derived factor activates a Gi-coupled phosphoinositide 3-kinase in T lymphocytes. *J Immunol*, 163(11):5954–63, 1999.

[258] T.F. Deuel, N. Zhang, H.J. Yeh, I. Silos-Santiago, and Z.Y. Wang. Pleiotrophin: a cytokine with diverse functions and a novel signaling pathway. *Arch Biochem Biophys*, 397(2):162–71, 2002.

[259] A.C. Rapraeger. Syndecan-regulated receptor signaling. *J Cell Biol*, 149(5):995–8, 2000.

[260] S.C. Todd, V.S. Doctor, and S. Levy. Sequences and expression of six new members of the tetraspanin/TM4SF family. *Biochim Biophys Acta*, 1399(1):101–4, 1998.

[261] J.L. Mullor and A. Ruiz i Altaba. Growth, hedgehog and the price of GAS. *Bioessays*, 24(1):22–6, 2002.

[262] V. Taylor, A.A. Welcher, A.E. Program, and U. Suter. Epithelial membrane protein-1, peripheral myelin protein 22, and lens membrane protein 20 define a novel gene family. *J Biol Chem*, 270(48):28824–33, 1995.

[263] P. Jiang, C.F. Lagenaur, and V. Narayanan. Integrin-associated protein is a ligand for the P84 neural adhesion molecule. *J Biol Chem*, 274(2):559–62, 1999.

[264] M.J. Metzelaar, P.L. Wijngaard, P.J. Peters, J.J. Sixma, H.K. Nieuwenhuis, and H.C. Clevers. CD63 antigen. A novel lysosomal membrane glycoprotein, cloned by a screening procedure for intracellular antigens in eukaryotic cells. *J Biol Chem*, 266(5):3239–45, 1991.

[265] X. Li, H.J. Yost, D.M. Virshup, and J.M. Seeling. Protein phosphatase 2A and its B56 regulatory subunit inhibit Wnt signaling in Xenopus. *EMBO J*, 20(15):4122–31, 2001.

[266] Y. Gachet, S. Tournier, M. Lee, A. Lazaris-Karatzas, T. Poulton, and U.A. Bommer. The growth-related, translationally controlled protein P23 has properties of a tubulin binding protein and associates transiently with microtubules during the cell cycle. *J Cell Sci*, 112 (Pt 8):1257–71, 1999.

[267] J.M. Seeling, J.R. Miller, R. Gil, R.T. Moon, R. White, and D.M. Virshup. Regulation of beta-catenin signaling by the B56 subunit of protein phosphatase 2A. *Science*, 283(5410):2089–91, 1999.

[268] F. Li, D. Zhang, and K. Fujise. Characterization of fortilin, a novel antiapoptotic protein. *J Biol Chem*, 276(50):47542–9, 2001.

[269] N. Larsson, H. Melander, U. Marklund, O. Osterman, and M. Gullberg. G2/M transition requires multisite phosphorylation of oncoprotein 18 by two distinct protein kinase systems. *J Biol Chem*, 270(23):14175–83, 1995.

[270] L.D. Belmont and T.J. Mitchison. Identification of a protein that interacts with tubulin dimers and increases the catastrophe rate of microtubules. *Cell*, 84(4):623–31, 1996.

[271] R. Taussig, J.A. Iniguez-Lluhi, and A.G. Gilman. Inhibition of adenylyl cyclase by Gi alpha. *Science*, 261(5118):218–21, 1993.

[272] R. Donato. S100: a multigenic family of calcium-modulated proteins of the EF-hand type with intracellular and extracellular functional roles. *Int J Biochem Cell Biol*, 33(7):637–68, 2001.

[273] P.M. Brzoska, H. Chen, N.A. Levin, W.L. Kuo, C. Collins, K.K. Fu, J.W. Gray, and M.F. Christman. Cloning, mapping, and in vivo localization of a human member of the PKCI-1 protein family (PRKCNH1). *Genomics*, 36(1):151–6, 1996.

[274] C. Brenner, P. Bieganowski, H.C. Pace, and K. Huebner. The histidine triad superfamily of nucleotide-binding proteins. *J Cell Physiol*, 181(2):179–87, 1999.

[275] V. Baldin. 14-3-3 proteins and growth control. *Prog Cell Cycle Res*, 4:49–60, 2000.

[276] M.V. Autieri and C.J. Carbone. 14-3-3Gamma interacts with and is phosphorylated by multiple protein kinase C isoforms in PDGF-stimulated human vascular smooth muscle cells. *DNA Cell Biol*, 18(7):555–64, 1999.

[277] J.P. Rouault, R. Rimokh, C. Tessa, G. Paranhos, M. Ffrench, L. Duret, M. Garoccio, D. Germain, J. Samarut, and J.P. Magaud. BTG1, a member of a new family of antiproliferative genes. *EMBO J*, 11(4):1663–70, 1992.

[278] W.J. Lin, J.D. Gary, M.C. Yang, S. Clarke, and H.R. Herschman. The mammalian immediate-early TIS21 protein and the leukemia-associated BTG1 protein interact with a protein-arginine N-methyltransferase. *J Biol Chem*, 271(25):15034–44, 1996.

[279] J.M. Boyd, S. Malstrom, T. Subramanian, L.K. Venkatesh, U. Schaeper, B. Elangovan, C. D'Sa-Eipper, and G. Chinnadurai. Adenovirus E1B 19 kDa and Bcl-2 proteins interact with a common set of cellular proteins. *Cell*, 79(2):341–51, 1994.

[280] A. Forlino, F.D. Porter, E.J. Lee, H. Westphal, and J.C. Marini. Use of the Cre/lox recombination system to develop a non-lethal knock-in murine model for osteogenesis imperfecta with an alpha1(I) G349C substitution. Variability in phenotype in BrtlIV mice. *J Biol Chem*, 274(53):37923–31, 1999.

[281] A.M. Delany, M. Amling, M. Priemel, C. Howe, R. Baron, and E. Canalis. Osteopenia and decreased bone formation in osteonectin-deficient mice. *J Clin Invest*, 105(7):915–23, 2000.

[282] A. Kruzynska-Frejtag, M. Machnicki, R. Rogers, R.R. Markwald, and S.J. Conway. Periostin (an osteoblast-specific factor) is expressed within the embryonic mouse heart during valve formation. *Mech Dev*, 103(1-2):183–8, 2001.

[283] K. Horiuchi, N. Amizuka, S. Takeshita, H. Takamatsu, M. Katsuura, H. Ozawa, Y. Toyama, L.F. Bonewald, and A. Kudo. Identification and characterization of a novel protein, periostin, with restricted expression to periosteum and periodontal ligament and increased expression by transforming growth factor beta. *J Bone Miner Res*, 14(7):1239–49, 1999.

[284] S.A. Downie and S.A. Newman. Different roles for fibronectin in the generation of fore and hind limb precartilage condensations. *Dev Biol*, 172(2):519–30, 1995.

[285] J. Skonier, K. Bennett, V. Rothwell, S. Kosowski, G. Plowman, P. Wallace, S. Edelhoff, C. Disteche, M. Neubauer, H. Marquardt, and et al. beta ig-h3: a transforming growth factor-beta-responsive gene encoding a secreted protein that inhibits cell attachment in vitro and suppresses the growth of CHO cells in nude mice. *DNA Cell Biol*, 13(6):571–84, 1994.

[286] X. Liu, H. Wu, M. Byrne, S. Krane, and R. Jaenisch. Type III collagen is crucial for collagen I fibrillogenesis and for normal cardiovascular development. *Proc Natl Acad Sci U S A*, 94(5):1852–6, 1997.

[287] H. Chen, B.W. Bernstein, and J.R. Bamburg. Regulating actin-filament dynamics in vivo. *Trends Biochem Sci*, 25(1):19–23, 2000.

[288] B. Peter, Y.M. Man, C.E. Begg, I. Gall, and D.P. Leader. Mouse cytoskeletal gamma-actin: analysis and implications of the structure of cloned cDNA and processed pseudogenes. *J Mol Biol*, 203(3):665–75, 1988.

[289] S. Danno, H. Nishiyama, H. Higashitsuji, H. Yokoi, J.H. Xue, K. Itoh, T. Matsuda, and J. Fujita. Increased transcript level of RBM3, a member of the glycine-rich RNA-binding protein family, in human cells in response to cold stress. *Biochem Biophys Res Commun*, 236(3):804–7, 1997.

[290] A. Celeste, S. Petersen, P.J. Romanienko, O. Fernandez-Capetillo, H.T. Chen, O.A. Sedelnikova, B. Reina-San-Martin, V. Coppola, E. Meffre, M.J. Difilippantonio, C. Redon, D.R. Pilch, A. Olaru, M. Eckhaus, R.D. Camerini-Otero, L. Tessarollo, F. Livak, K. Manova, W.M. Bonner, M.C. Nussenzweig, and A. Nussenzweig. Genomic instability in mice lacking histone H2AX. *Science*, 296(5569):922–7, 2002.

[291] G.E. Breitwieser, J.C. McLenithan, J.F. Cortese, J.M. Shields, M.M. Oliva, J.L. Majewski, C.E. Machamer, and V.W. Yang. Colonic epithelium-enriched protein A4 is a proteolipid that exhibits ion channel characteristics. *Am J Physiol*, 272(3 Pt 1):C957–65, 1997.

[292] M.A. Guthridge, M. Seldin, and C. Basilico. Induction of expression of growth-related genes by FGF-4 in mouse fibroblasts. *Oncogene*, 12(6):1267–78, 1996.