INAUGURAL-DISSERTATION

zur

Erlangung der Doktorwürde

der

Naturwissenschaftlich-Mathematischen Gesamtfakultät

der

Ruprecht-Karls-Universität

Heidelberg

vorgelegt von

Diplom-Biologe Tim Beißbarth

aus Köln, Deutschland

Tag der mündlichen Prüfung:

Thema

Analysis of transcription factor CREM dependent gene expression during mouse spermatogenesis

Gutachter:

Prof. Dr. Günther Schütz

Prof. Dr. Richard Herrmann

Contents

1 Summary

Computational methods are getting increasingly important for the analysis of large data sets in molecular biology. The data sets analyzed in this thesis are derived from experiments measuring the changes of expression levels in response to the transcription factor CREM (cAMP Responsive Element Modulator) during mouse spermatogenesis. In the course of this analysis new computational methods were developed and used that will also be of value in other projects in Bioinformatics.

CREM belongs to a family of cAMP-responsive nuclear factors. The activator splice-isoform CREMτ is exclusively expressed at high levels in post-meiotic germ cells during mouse spermiogenesis. Mutant male mice lacking CREM expression are sterile due to lack of maturation of the germ cells.

In order to find CREM target genes the mRNA expression levels in testes of CREM-deficient mice and wild-type mice were compared using the suppression subtractive hybridization (SSH) technique as well as oligonucleotide DNA microarrays.

SSH was used to selectively amplify the differentially expressed genes. 12,000 clones, which contain sequence fragments of genes expressed stronger in wild-type as in the CREM (-/-) mutant, were analyzed by a combination of sequencing and hybridization.

Sequence analysis methods were used to characterize 956 unique sequences. Homologies to 158 known mouse genes and 99 known genes from other organisms were detected. 296 sequences show homologies to sequences of expressed sequence tags (ESTs). 199 novel sequences have been found.

The sequences not corresponding to full length genes of known function were characterized using publicly available EST data. To make EST databases useful for data analysis all of the publicly available ESTs have been grouped into clusters and methods to analyze and visualize EST data were developed.

Nylon cDNA microarrays containing the unique sequences from the CREM SSH library were constructed to determine expression levels of those sequences. Most of the sequences from the CREM SSH library are shown to be expressed in wild-type but are down-regulated in CREM deficient mice.

Statistical methods to standardize microarray expression data were developed and software was implemented to perform comparisons.

Further CREM dependent genes were detected comparing the mRNA expression levels in testes of CREM deficient mice and wild-type mice using Affymetrix oligonucleotide microarrays containing 10,000 mouse sequences. Comparison of the different techniques (SSH, nylon cDNA arrays and Affymetrix oligonucleotide microarrays) shows that the results are complementing each other.

The unique sequences from the CREM SSH library were further analyzed by determining the spermatogenic stage specific expression profiles. cDNA from prepubertal mice at certain stages of spermatogenesis were hybridized on nylon cDNA arrays. Several important functional groups of genes like transcription factors, signal transduction proteins and metabolic enzymes are shown to be coexpressed at the latest stages of spermatogenesis.

Expression profiles were arranged to find similar profile shapes and co-regulation of functionally related genes. An algorithm to arrange the profiles in an optimal linear order was developed. The linear order is constructed in a way that similar expression profiles end up close together in the linear order, i.e. the sum over all distances of neighboring profiles is minimized. This corresponds to the solution of a traveling salesman problem (TSP), which is well known in computer science. A fast algorithm that computes a heuristic solution to a TSP was adapted to be used in expression profile analysis.

2 Introduction

2.1 CREB, ATF-1 and CREM are members of a family of transcription factors

2.1.1 Transcriptional regulation by factors responsive to cyclic AMP

Coordinated gene expression programs regulate the complex processes of cell growth and differentiation. The modulation of gene expression by specific signal transduction pathways enables cells to trigger the appropriate short- and long-term adaptation programs in response to changes in the environment. These transduction pathways control the activity of transcription factors. Several transcription factors finally respond to elevated levels of cAMP by binding to a regulatory DNA sequence known as the cAMP-responsive element (CRE) and by activation of transcription. The proteins that bind to the CRE are the cAMP responsive element binding protein (**CREB**) (Hoeffler et al., 1988), the cAMP responsive element modulator (**CREM**) (Foulkes et al., 1992) and the activating transcription factor 1 (**ATF-1**) (Rehfuss et al., 1991). They form a family of transcription factors and have originally been identified as activators that respond to the cyclic AMP (cAMP)-dependent signaling pathway. Members of the CREB/ATF-1 family play important roles in the nuclear responses to a variety of external signals and are involved in physiological systems like memory and long-term potentiation (Silva et al., 1998), circadian rhythms (Foulkes et al., 1997), pituitary function (Struthers et al., 1991) and spermatogenesis (Sassone-Corsi, 1997).

CREB and CREM can be activated by a specific phosphorylation event (Figure 2.1). Their phosphorylation site (Ser133 in CREB, and Ser117 in CREM) is within a highly conserved region (Zhou et al., 1996). Ser133 and Ser117 are phosphorylated by cAMP-dependent protein kinase A (PKA). PKA is regulated by changes in intracellular cAMP levels, which are elevated following activation of adenylyl cyclase by binding of specific ligands to G-protein coupled receptors. After cAMP binds to the regulatory subunit of PKA, the catalytic subunit of PKA translocates to the nucleus, where phosphorylation of CREB and CREM occurs. Several lines of evidence now indicate that CREB

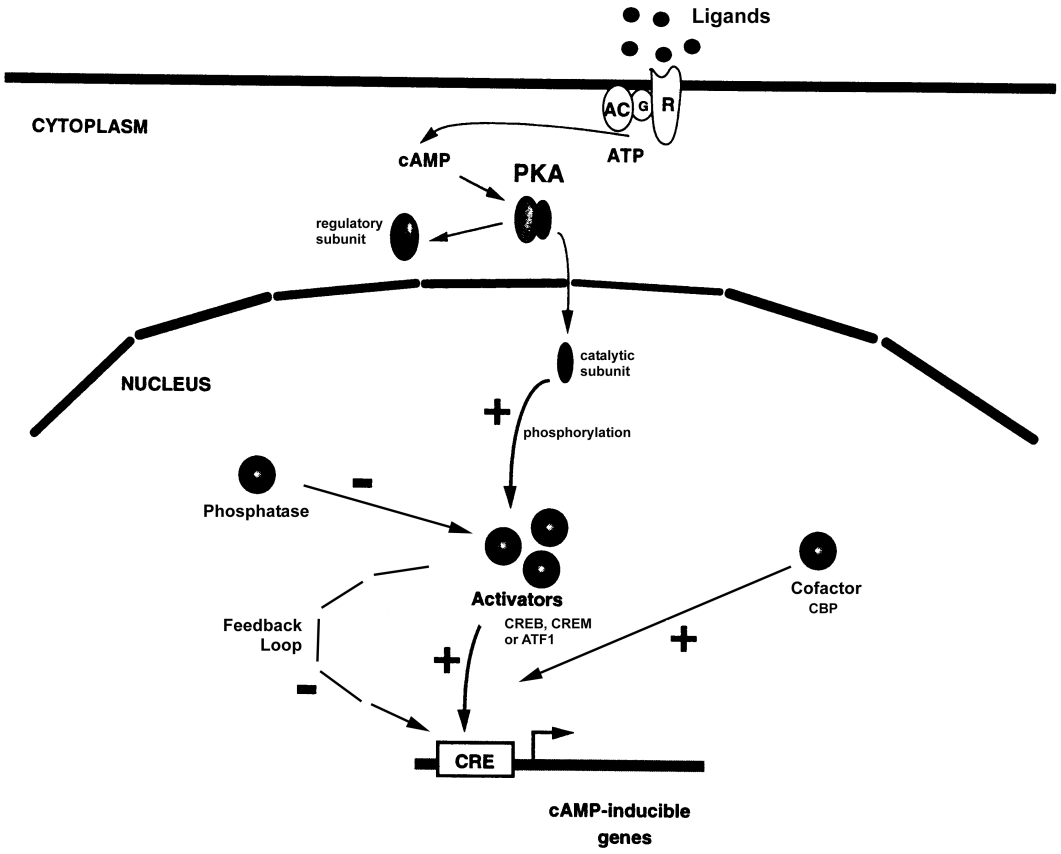


Figure 2.1: **The cAMP dependent signal transduction pathway**. Schematic representation of the route whereby ligands at the cell surface interact with membrane receptors (R), resulting in altered gene expression. Ligand binding activates G-proteins (G), which in turn stimulate the activity of the membrane-associated adenylyl cyclase (AC). This converts ATP to cAMP, causing the dissociation of the inactive tetrameric protein kinase A (PKA) complex into the active catalytic subunits and the regulatory subunits. Catalytic subunits migrate into the nucleus, where they phosphorylate and thereby activate transcriptional activators such as CREB, CREM and ATF-1. Attenuation of the activators may occur via a nuclear phosphatase. Transcriptional induction often requires interaction of the activators with CREB binding protein (CBP), a cofactor. These activators then interact with the cAMP-responsive element (CRE) found in promoters of genes responding to cAMP and activate transcription. The phosphorylated activators can also induce transcription of repressors of CRE, and by this permit a new cycle of transcriptional activation (modified from Sassone-Corsi, 2000).

and CREM can be phosphorylated by different kinases, which are activated by a variety of signals (De Cesare et al., 1999).

CREB, CREM and ATF-1 belong to the basic-domain leucine-zipper (bZip) class of proteins. They are able to form homodimers as well as heterodimers that bind to the palindromic consensus sequence TGACGTCA of CRE. CRE was originally identified in the somatostatin promoter and is present in the regulatory regions of most cAMP-responsive genes (Sassone-Corsi, 1998). cAMP-responsive factors interact with coactivators that, in turn, contact the basal transcriptional machinery. Thus, the versatility of the nuclear response is provided by the variety of signaling pathways converging on CREB and CREM, and by the diversity of interactions between these transcription factors and their coactivators.

The factors are encoded by unique genes in *Aplysia californica, Chlorohydra viridissima* and *Drosophila melanogaster* (Bartsch et al., 1998; Galliot et al., 1995; Yin et al., 1995); these probably represent evolutionary precursors of a gene that was then duplicated in higher eukaryotes.

CREB, CREM and ATF-1 encode many isoforms, which provide additional complexity to the pathways of transcriptional activation. The CREM isoforms are generated by alternative splicing, use of an alternative initiation codon or by an alternative, intronic promoter (Sassone-Corsi, 2000).

2.1.2 The transcription factor CREM

2.1.2.1 Splice isoforms of CREM

Differential transcript processing is central to the regulation of CREM expression. Control is exerted at three different levels: alternative splicing, alternative polyadenylation and alternative translation initiation (Sassone-Corsi, 2000).

Characterization of the genomic organization of the CREM gene has revealed the molecular basis for this extensive family of isoforms. The exons accurately define functional domains (Laoide et al., 1993; de Groot and Sassone-Corsi, 1993). This modular structure combined with extensive differential splicing permits the CREM gene to encode a family of transcription factors with different activation properties (Figure 2.2).

The CREM gene encodes activators as well as repressors. Functional domains are the P-box, which contains the phosphorylation site, two glutamine-rich domains (Q1 and Q2), which are thought to be important to interact with the basic transcriptional machinery, and a DNA binding domain (DBD1 or DBD2), which can form the leucine zipper by dimerization with specific other factors. By the use of an alternative promoter (P2) a transcript called ICER (inducible cAMP early repressor) is produced, which only codes for a DNA binding domain and acts as a repressor of the cAMP responsive genes.

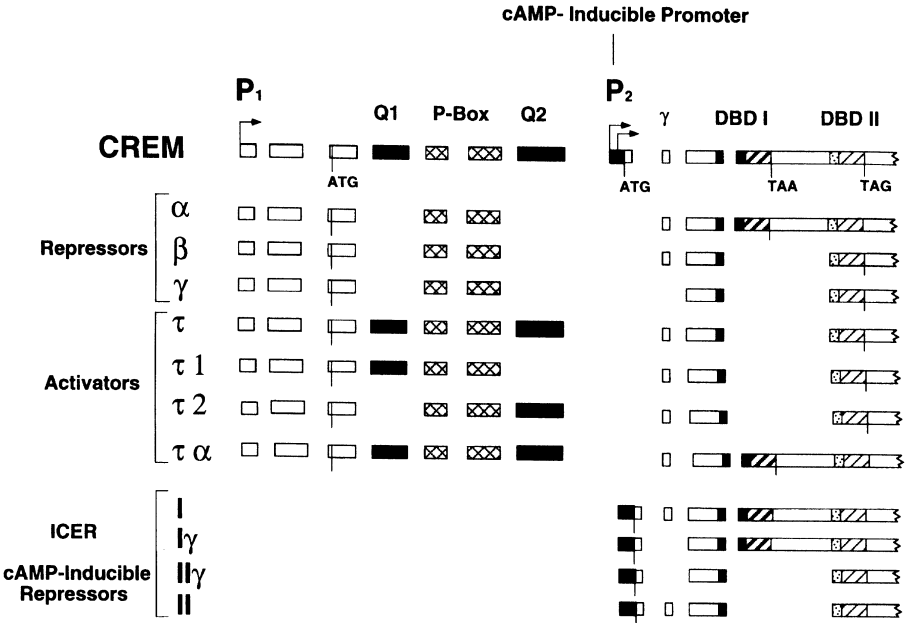


Figure 2.2: **Splice isoforms of CREM.** Schematic representation of the CREM gene. Exons encoding the glutamine-rich domains (Q1 and Q2), the P-Box, the γ domain (γ) and the two alternative DNA binding domains (DBDI and DBDII) are shown. Activators and repressors are encoded by the same gene. The various activator and repressor isoforms which have been described to date are indicated. The P1 promoter is GC-rich and directs a non-inducible pattern of expression; the P2 promoter is strongly inducible by activation of the cAMP-dependent signaling pathway (Sassone-Corsi, 2000).

The P2 promoter is strongly inducible by the cAMP-dependent signaling pathway, which leads to a feedback loop allowing a new cycle of transcriptional activation (Sassone-Corsi, 2000).

Using alternative polyadenylation sites, the CREM gene can generate transcripts bearing different numbers of AUUUA elements in the 3' untranslated regions. These elements have been demonstrated to confer mRNA instability in other genes. During spermatogenesis a more stable transcript is generated using the most proximal (5') polyadenylation site, that has only a single AUUUA element. Thus, the relative abundance of different CREM isoforms can be controlled by RNA processing (Foulkes et al., 1993).

CREMτ is an activating isoform that is uniquely and highly expressed during spermatogenesis. It contains the DNA binding domain (DBD), the P-box domain for phosphorylation dependent binding with CBP, and the Q domains for interaction with basal transcription machinery. CREM isoforms that contain only the P box and miss the Q domains behave as transcriptional repressors (Laoide et al., 1993; Foulkes et al., 1991)

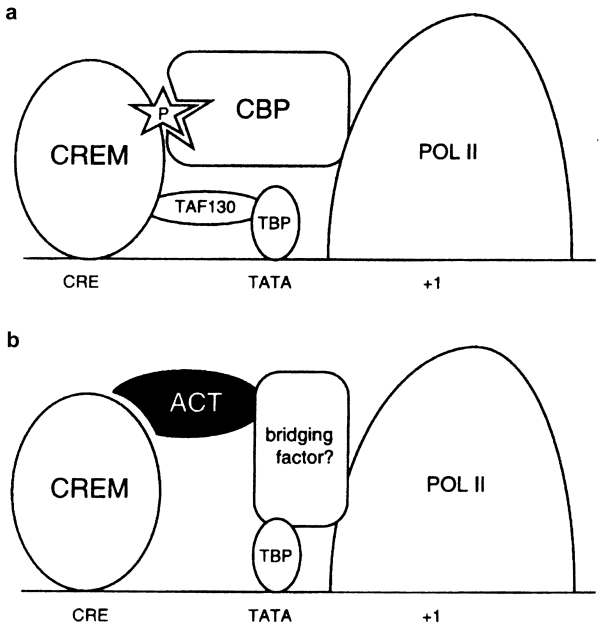


Figure 2.3: CREM-mediated transcription is promoted by interaction with different coactivators. **a** Phosphorylation of Ser117 of CREM promotes binding to CREB-binding protein (CBP) and subsequent interaction with the basal transcription machinery (TATA-binding protein TBP and Polymerase II complex) and leads to transcriptional activation. Interaction with the factor TAF130 is constitutive and is mediated by the Q2 domain of CREM. **b** Model of coactivation by activator of CREM in testis (ACT). ACT exerts its function independently of Ser117 phosphorylation and in the absence of TAF130. Thus, ACT provides an alternative activation pathway that appears to work in a signaling-independent manner. A hypothetical bridging factor is required to link ACT to the basal transcription machinery (Sassone-Corsi, 2000).

2.1.2.2 Different types of activation of the CREM protein

Besides the classical model of CREM activation by phosphorylation and binding to the coactivator CBP, new evidence indicates an alternative activation route of CREM in testis. CREMτ may be activated without phosphorylation by binding to Activator of CREM in Testis (ACT) protein (Fimia et al., 1999). The two models how CREM activates transcription are outlined in Figure 2.3.

Analysis of the phosphorylation state of CREM at various stages of spermatogenic differentiation revealed that CREM is unphosphorylated at the time it activates post-meiotic genes. Employing a genomic screen in yeast of a testis-derived cDNA library, a clone was identified that encodes the protein ACT (Fimia et al., 1999). ACT converts an inactive mutant of CREM (Ser117→Ala) into a transcriptionally active molecule both in yeast and in mammalian cells. CREM in association with ACT activates transcription independent of Ser117 phosphorylation and the binding of CBP (De Cesare et al., 1999).

2.1.2.3 Functions of CREM in different tissues

The CREM gene encodes various transcription factors which play physiological key roles within different tissues. The various functions of CREM have been analyzed using transgenic CREM (-/-) mice (Blendy et al., 1996; Nantel et al., 1996). Three abnormalities were found in CREM deficient mice: altered circadian cycle, delayed liver regeneration and impairment of spermatogenesis.

1. **Altered circadian cycle**

   CREM appears to act as a regulator of output functions of the biological clock in response to adaptive environmental changes (Foulkes et al., 1996). CREM deficient mice are hyperactive and do not show the characteristic day-night change in locomotion. The emotional state of these mice indicates a decrease in anxiety-like behavior (Maldonado et al., 1999).

2. **Delay of liver regeneration**

   The liver has a remarkable ability to regenerate in mice. As much as 70% of the liver can be surgically removed and hepatocytes will proliferate to fully regenerate the original cell mass. CREM appears to coordinate the timing of hepatocyte proliferation during the process of liver regeneration (Servillo et al., 1998).

3. **Impairment of spermatogenesis**

   Heterozygous CREM (+/-) mice display reduced fertility. The testes of these mice show a 46% reduction in the overall number of spermatozoa, a 35% decrease in the ratio of motile spermatozoa, and a twofold increase in the number of spermatozoa with aberrant structures. Homozygous transgenic CREM (-/-) males are unable to reproduce despite the normal mating behavior. Testes of these mice show a reduction of 20–25% of their weight and display a complete absence of spermatozoa. Further characterization of the testes of CREM deficient mice shows mature spermatozoa to be absent due to an arrest of spermatogenesis at the stage of round spermatids. Instead of normal differentiation the spermatids appear to enter the apoptotic cell-death pathway. In contrast, female CREM (-/-) mice are fertile.

   A number of post-meiotic germ cell-specific genes were shown to be not expressed in testes of CREM deficient mice. Among them are: protamine 1 and 2, Tp-1, MCS, outer dense fiber protein (RT7), Krox-20, Krox-24, proacrosin, and calspermin (Blendy et al., 1996; Nantel et al., 1996).

In this thesis, the role of CREM in mouse spermatogenesis is studied in more detail. This analysis is aimed to determine those genes which are activated in response to CREM in mouse testes. Many more genes were found to be down-regulated in testes of CREM deficient mice and are likely to play roles in spermatogenesis.

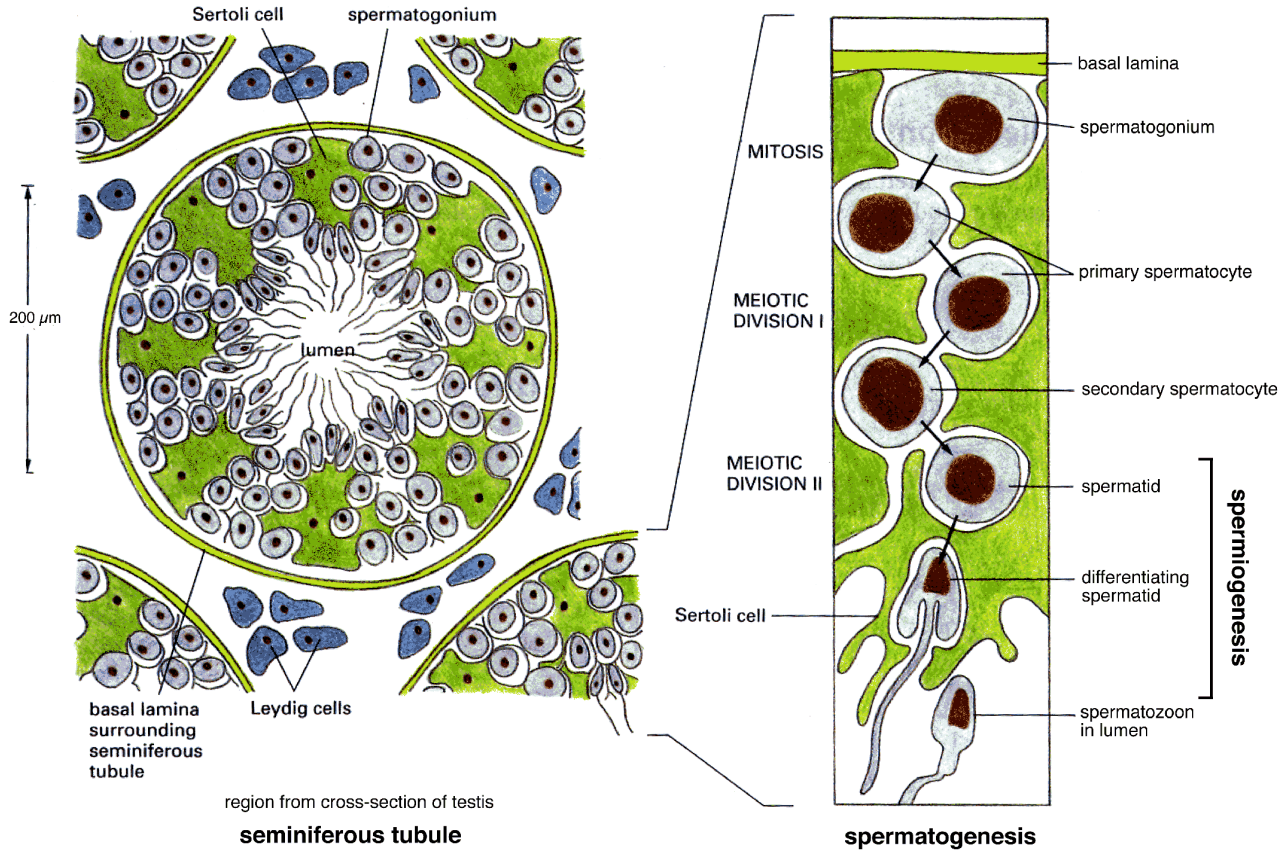


Figure 2.4: Simplified drawing of a cross-section of a seminiferous tubule in mouse testis. Spermatogonia proliferate to generate a pool of undifferentiated stem cells. Some of the spermatogonia differentiate to mature sperm. This process includes two meiotic divisions and a restructuring of the cells during spermiogenesis. The developing gametes are in intimate association with Sertoli cells (modified from Alberts *et al.*, 1994).

2.2 Overview over the course of spermatogenesis

More than one hundred million sperm cells are produced by one testis per day. This depends on a highly coordinated process ensuring efficient cellular replication and a finely tuned differentiation program. Spermatogenesis is the process in which diploid spermatogonia differentiate into mature haploid spermatozoa.

2.2.1 Stages of germ cells

Spermatogenesis takes place within the seminiferous tubules of a testis. During the entire developmental process the germ cells are in tight contact with Sertoli cells, which supply them with growth factors and nutrients. Figure 2.4 shows a schematic representation of the process. As the germ cells mature, they move from the periphery towards the lumen of the tubule until the mature spermatozoa

are released from the lumen to the collecting ducts. In mice, the entire developmental process takes 35 days (Browder et al., 1991; Sassone-Corsi, 1997).

The developing germ cells during spermatogenesis are classified into several stages that can be morphologically distinguished. Figure 2.5 shows some of these stages in a diagrammatic representation. The process is divided in a proliferative phase (which takes approximately 8 days), a meiotic phase (13 days) and a spermiogenic phase (14 days). During the proliferative phase the cells are called spermatogonia and undergo mitotic divisions. Some of these spermatogonia undergo further differentiation and divide mitotically before entering the meiotic phase. The resulting diploid cells are then called spermatocytes and undergo two meiotic divisions. Before the first meiotic division the cells are called primary spermatocytes. The prophase of the first meiotic division in primary spermatocytes is divided into leptotene, zygotene, pachytene and diplotene. After the first meiotic division the secondary spermatocytes divide again meiotically and are then called spermatids. The haploid spermatids differentiate from round spermatids to elongated spermatids to mature spermatozoa in a number of stages during the phase of spermiogenesis. This phase involves an extensive biochemical and morphological restructuring of the germ cells. The mature sperm consist of a head and a tail region. The motile tail propels the sperm to the egg and contains mitochondria. The head contains the condensed haploid DNA and the acrosome, which is important to penetrate the outer coat of the egg.

The cells in a segment of a seminiferous tubule differentiate in synchrony. Twelve cyclic stages of the seminiferous tubules have been defined, each of which contains several stages of the developing germ cells (Russel et al., 1990).

Spermatogenesis is initiated shortly after birth. In consequence, during the prepubertal period the seminifereous epithelium contains only Sertoli cells, spermatogonia, and, with increasing age, progressively more advanced stages of the developing germ cells. 3-5 days after birth the spermatogonial progenitor cells start to proliferate mitotically. At day 10 the meiotic prophase is initiated. The germ cells reach the early and late pachytene stage by days 14 and 18, respectively. Secondary spermatocytes and haploid spermatids appear in increasing numbers between days 18 and 20 (Bellve et al., 1977).

A proportion of germ cells undergoes apoptosis in the seminiferous epithelium. This number increases dramatically in some pathological conditions, including infertility due to spermatogenic arrest.

2.2.2 Hormonal control of spermatogenesis

The hypothalamic-pituitary axis evokes a cyclic hormonal control responsible for a coordinated differentiation program of the germ cells. In response to hormone stimulation, testicular cells initiate

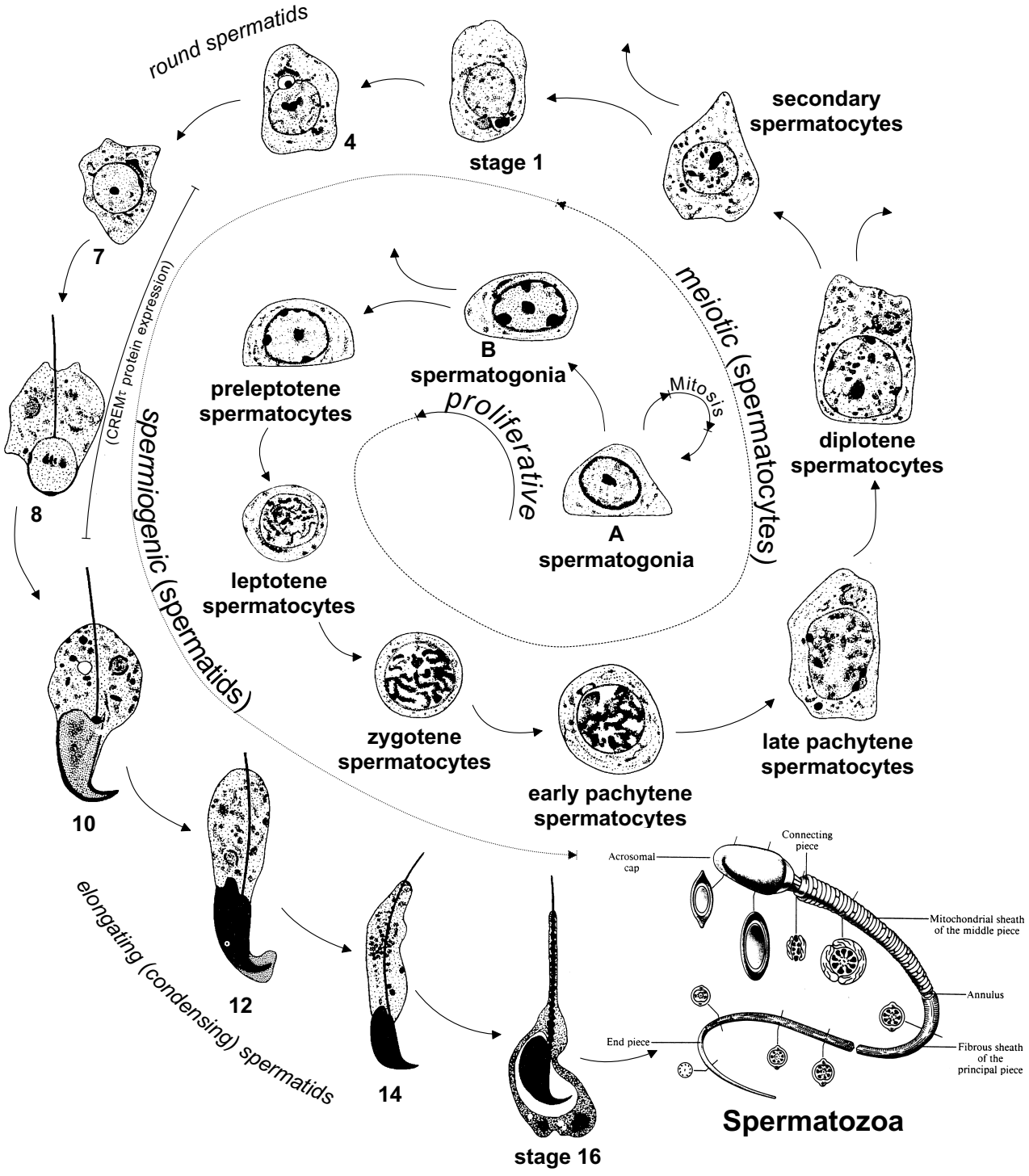


Figure 2.5: Diagrammatic representation of mouse spermatogenesis showing the three major developmental phases (*proliferative*, *meiotic* and *spermiogenic*) that are occupied by spermatogonia, spermatocytes, and spermatids.

a cascade of events inducing changes in cellular metabolism and gene expression. Proliferation and differentiation of germ cells is dependent on two hormones produced by the gonadotrophs of the anterior pituitary gland, luteinizing hormone (LH) and follicle-stimulating hormone (FSH). In the absence of these two hormones, spermatogenesis does not proceed beyond the meiotic prophase.

The developing male germ cells lack receptors for LH and FSH and it is consequently believed that they receive hormonal signals indirectly via the somatic cells present in the testis. LH and FSH receptors are located on the Leydig and Sertoli somatic cells, respectively. Stimulation of Leydig cells by LH results in the secretion of testosterone into the interstitial compartment which then diffuses into the seminiferous tubules where Sertoli and germ cells are located. Upon combined FSH and testosterone stimulation, Sertoli cells secrete peptides and other components that are required for differentiation of the germ cells (Sassone-Corsi, 1997).

2.2.3 Gene expression at various stages of germ cell differentiation

Developing germ cells respond to physiological stimuli and convert them into signals leading to differentiation. This causes modulation in gene expression and rapid changes in morphology and biochemistry of the maturating germ cells.

The most dramatic changes in gene expression happen during differentiation of round spermatids to mature sperm cells. To realize these changes during the round spermatid stage most proteins and structures are substituted by new proteins and structures characteristic for sperm. The protein degradation and synthesis machinery is very active. For example, histones enter the degradation pathway after being tagged with ubiquitin in round spermatids (Baarends et al., 1999). Protamines and transition proteins are synthesized to substitute the histones and condense the chromatin (Kistler et al., 1994; Ha et al., 1997).

Highly specialized transcriptional mechanisms ensure stringent stage-specific gene expression in the germ cells. Specific checkpoints correspond to the activation of transcription factors; these regulate gene promoters with a restricted pattern of activity specific for germ cells.

Since global transcription ceases several days before the completion of spermiogenesis (about stage 9 of spermatids) post-transcriptional control is important at the end of spermatogenesis. Thus, mRNA storage and translational activation play prominent roles in the expression of many spermatid and spermatozoan proteins that are synthesized in late stages of germ cell maturation. For example, in early spermatids, many mRNAs, such as protamine and transition protein transcripts, are translationally repressed with long poly(A) tracts and are sequestered in cytoplasmic ribonucleoprotein particles for up to a week. Translation subsequently takes place in late spermatids after the mRNAs undergo a poly(A) shortening by deadenylation. Cellular signaling pathways must control the mRNA processing (Sassone-Corsi, 1997).

2.2.4 Possible role of CREMτ in spermatogenesis

CREM, a transcriptional regulator responsive to the cAMP signaling pathway (Section 2.1), has various neuroendocrine functions and has a direct role in determining the fate of male germ cells (Sassone-Corsi, 1995). The second messenger cAMP is known to play an important role in several steps of spermatogenesis, in particular by governing the timing of post-meiotic gene activation. The hypothalamic-pituitary axis influences the CREM developmental switch. FSH appears to regulate CREM expression by alternative polyadenylation, which results in an enhancement of transcript stability and accumulation of the CREM activator protein in the germ cells (Foulkes et al., 1993). Targeted disruption of the CREM gene results in a complete block of the germ cell differentiation program at the first step of spermiogenesis (Blendy et al., 1996; Nantel et al., 1996). Early differentiation and stem cell renewal occur normally in CREM deficient mice. This is in accordance with previous observations showing that the CREM protein accumulates during the round spermatid stage and exerts its function later on (Sassone-Corsi, 1995). The stringent requirement for CREM is manifested by the lack of maturation of the germ cells and by their entering the apoptotic cell death pathway. Indeed, deletion of CREM causes a 10-fold increase in the number of apoptotic germ cells (Blendy et al., 1996; Nantel et al., 1996).

CREM constitutes an abundant transcript from the pachytene spermatocyte stage onwards (Foulkes et al., 1992). Characterization of the CREM isoform expressed in the adult testis reveals that it encodes exclusively the CREMτ activator, while in prepubertal testis only the repressor forms are detected at low levels.

Due to a translational delay the CREMτ protein is not detected in pachytene spermatocytes but in spermatids which have undergone meiosis. The CREMτ protein is restricted to round spermatids at stages 7-8 (Figure 2.5). CREM has to exert its transactivator function during only a few stages as at the later stages in elongating spermatids all transcription ceases due to the compaction of the DNA (Delmas et al., 1993; Fimia et al., 2001).

As CREMτ protein is highly abundant in haploid germ cells and maturation of these cells ceases in CREM (-/-) mice, it is critical for entrance into the last steps of differentiation of the spermatids. Several genes have been identified which are transcribed at the time of appearance of the CREM protein, and which have CRE-like sequences in their promoter regions (Montminy, 1997). Various target genes for CREM-mediated activation have been identified in post-meiotic germ cells, such as protamine, RT7 (Galliot et al., 1995), transition protein-1 (Molina et al., 1993), angiotensin-converting enzyme (Radhakrishnan et al., 1997), and calspermin (Sun and Means, 1995).

2.3 Aims and structure of this thesis

For this thesis bioinformatical methods have been developed and applied in the analysis of genes that play a role during mouse spermatogenesis.

The role of the transcription factor CREM during mouse spermatogenesis was investigated. Mice with a targeted disruption of the CREM gene were used. Male CREM deficient mice fail to form mature spermatids and several genes are known to be down-regulated in round spermatids of CREM (-/-) mice (Blendy et al., 1996; Nantel et al., 1996). It is expected that many more genes are under the direct or indirect influence of CREM. This study aims to find new genes that are transcriptionally activated dependent on CREMτ during spermatogenesis.

In order to compare expression levels of genes in wild-type versus CREM (-/-) testes different techniques as suppression subtractive hybridization (SSH) and DNA microarray hybridizations were applied. Large scale subtractive cloning of genes down-regulated in CREM deficient mice using the SSH technology was performed. The expression profiles of these genes were measured during the development of sperm in prepubertal mice by nylon cDNA microarrays. Alternatively, the transcript expression levels of 10,000 genes or expressed sequence tags (ESTs) were compared between wild-type and CREM (-/-) testes directly using oligonucleotide DNA microarrays.

The large amount of data generated in this project and in general by methods such as SSH and DNA microarray hybridization has created the necessity to develop computational methods in order to analyze the data. A large part of the work is the analysis (*in silicio genomics*) and organization of the data. The field of computational molecular biology is very new and rapidly developing. The chapter "Methods" summarizes shortly the biological methods from which the analyzed data are derived and gives an overview over the state of the art methods for bioinformatical analysis, which have been developed in the last few years.

The development of methods in computational biology constitutes an important part of this thesis. In the first part of the chapter "Results" several new developments are described which were prerequisites to analyze the data presented in this thesis. These methods can be applied in other projects as well.

The second part of the chapter "Results" shows an example of the application of these computational methods. The data generated by SSH to clone CREM dependent target genes are analyzed in detail and presented in this chapter. The biological data were generated in the group of Prof. Dr. Günther Schütz in collaboration with Dr. Igor Borisevich.

The chapter "Discussion" points out the significance of the biological findings in the CREM SSH library, the applicability of the presented computational methods in general as well as in regard to the experiences gathered in the exemplary use on the CREM SSH data set.

3 Methods

3.1 Cloning differential messages via Suppression Subtractive Hybridization

CREM dependent cDNA sequences were obtained from subtractive cloning. A scheme of the subtraction is shown in Figure 3.1.

Suppression subtractive hybridization (SSH) is an effective method for finding genes expressed in one mRNA population but reduced in another (Diatchenko et al., 1999). SSH is based primarily on a technique called suppression PCR, and combines normalization and subtraction in a single procedure. The normalization step equalizes the abundance of cDNAs within the target population and the subtraction step excludes the common sequences between the target and reference populations. After one round of subtractive hybridization the subtracted library is normalized in terms of abundance of different cDNAs. This increases the probability of obtaining low-abundance differentially expressed cDNA (Diatchenko et al., 1996; Gurskaya et al., 1996; von Stein et al., 1997; Tchernitsa et al., 1999).

For SSH the mRNAs were isolated from testes of adult wild-type and transgenic CREM (-/-) mice (Figure 3.2). Transgenic mice with a targeted disruption in the CREM gene were used. In the target vector a deletion disrupts the gene and removes the coding information for both leucine zipper domains. Leucine zipper domains are essential for dimerization and subsequent DNA binding of CREM. Therefore, in these mice no functional CREM protein is made which leads to an absence of cells in the late stages of spermatogenesis in testes (Blendy et al., 1996). The isolated mRNA was used for the cDNA synthesis and was digested with the restriction enzyme *Rsa*I which recognizes the four nucleotide sequence GTAC and releases blunt ended DNA fragments.

The SSH procedure was performed by Andreas Hörlein, Igor Borisevich and Annette Klewe-Nebenius using the PCR-Select method (Clontech). Details are summarized in the PhD thesis of Igor Borisevich (Borisevich, 2001). In order to interpret the data of the SSH that are analyzed in this thesis, a short schematic overview is given below.

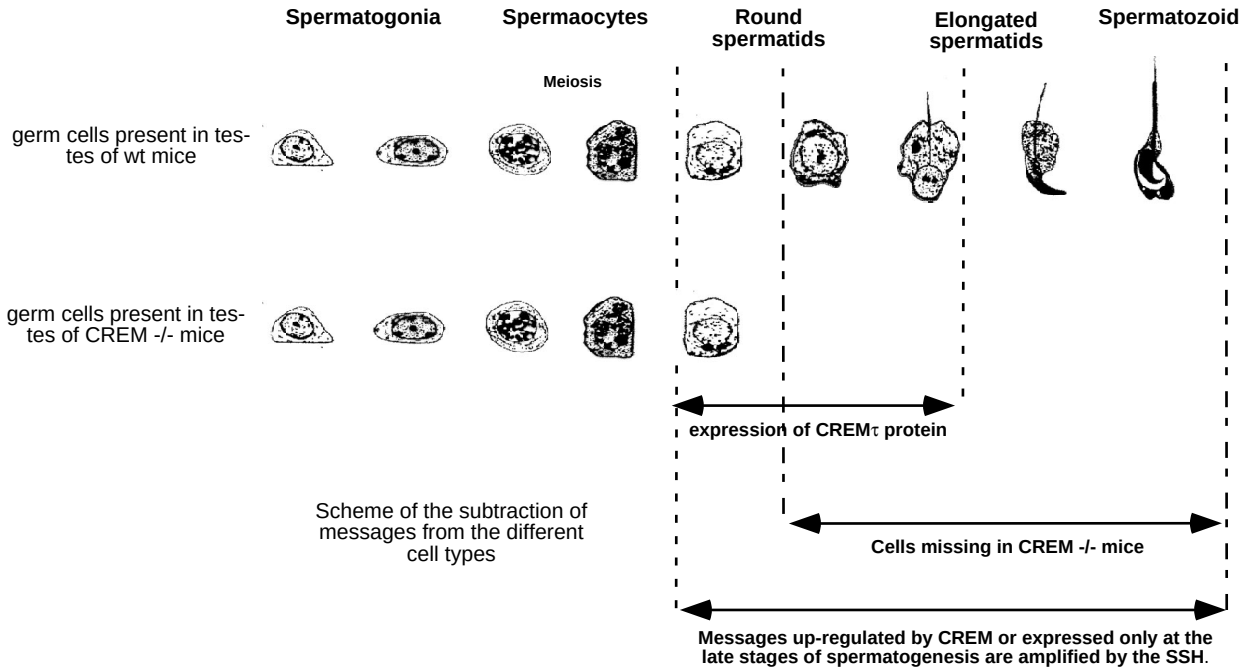


Figure 3.1: Scheme of cell types used for differential cloning of CREM target genes. By use of SSH the pool of cDNAs expressed in CREM (-/-) testes was subtracted from the pool of cDNAs expressed in wild-type testes. As a result CREM-dependent cDNAs were cloned.

3.1.1 Principle of suppression subtractive hybridization

1. **cDNA synthesis & adaptor ligation**
   First, RNA is isolated from the two types of tissues or cells being compared, then cDNA is synthesized. The cDNA in which specific transcripts (i.e. transcripts that are of much higher abundance compared to the reference) are to be found is called “tester” (here wild-type cDNA), and the reference cDNA is called “driver” (here CREM (-/-) testes cDNA). The tester and driver cDNAs are digested with a four-base-cutting restriction enzyme that yields blunt ends. The tester cDNA is then subdivided into two portions, each of which is ligated to a different ds cDNA adaptor. The ends of the adaptors lack a phosphate group, so only one strand of each adaptor attaches to the 5’ ends of the cDNAs. The driver cDNA has no adaptors.

2. **The SSH technique uses two hybridizations**
   An excess of driver is added to each sample of the tester. The samples are then heat-denatured and allowed to anneal. The tester fraction (a) is normalized, meaning concentrations of high and low abundance cDNAs become roughly equal. Normalization occurs because the reannealing process generating homohybrid cDNAs (b) is faster for more abundant molecules, due to the second order kinetics of hybridization. Furthermore, the cDNAs in the tester fraction (a) are significantly enriched for differentially expressed genes, as “common” non-target cDNAs form heterohybrids (c) with the driver.

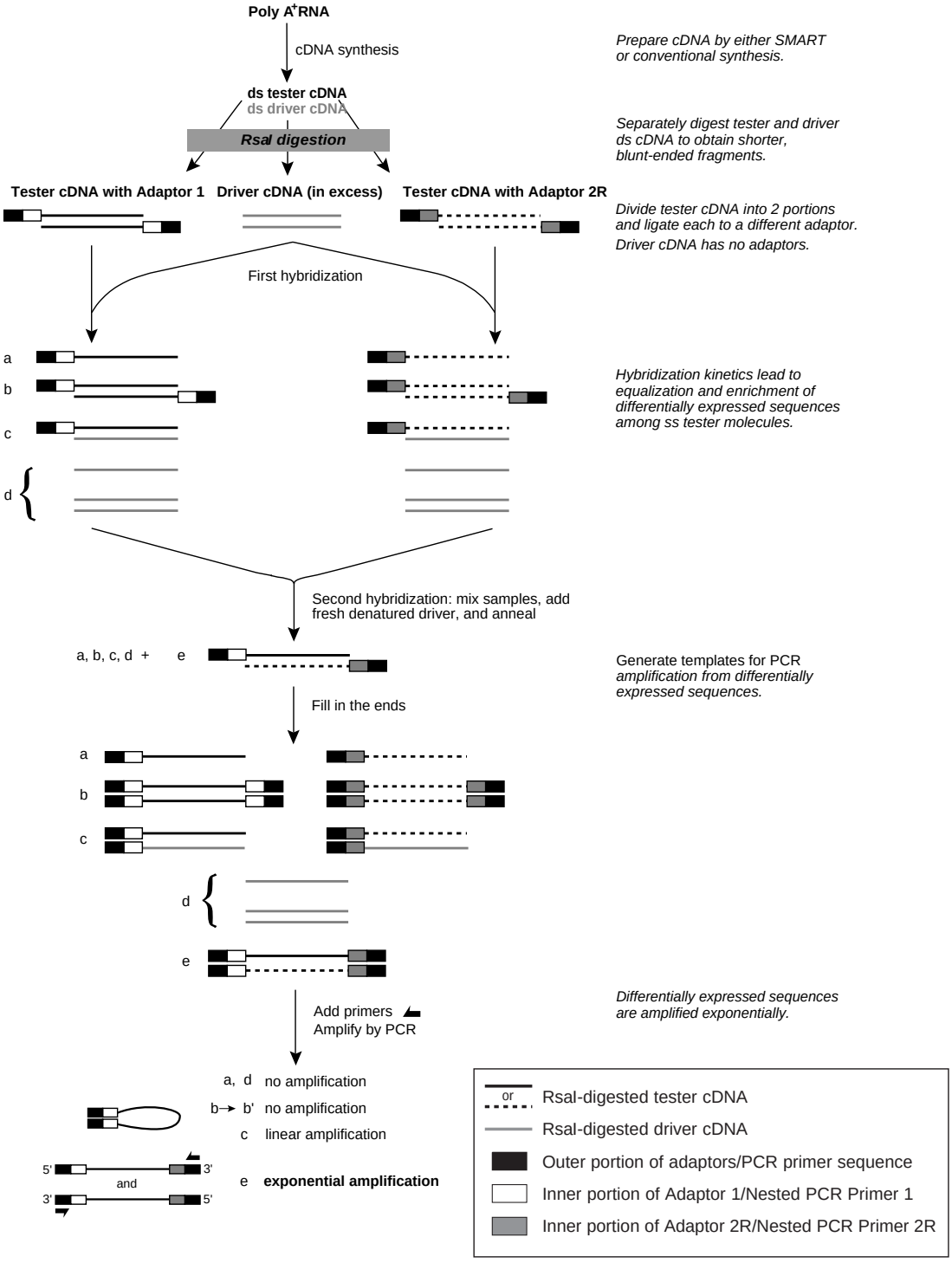


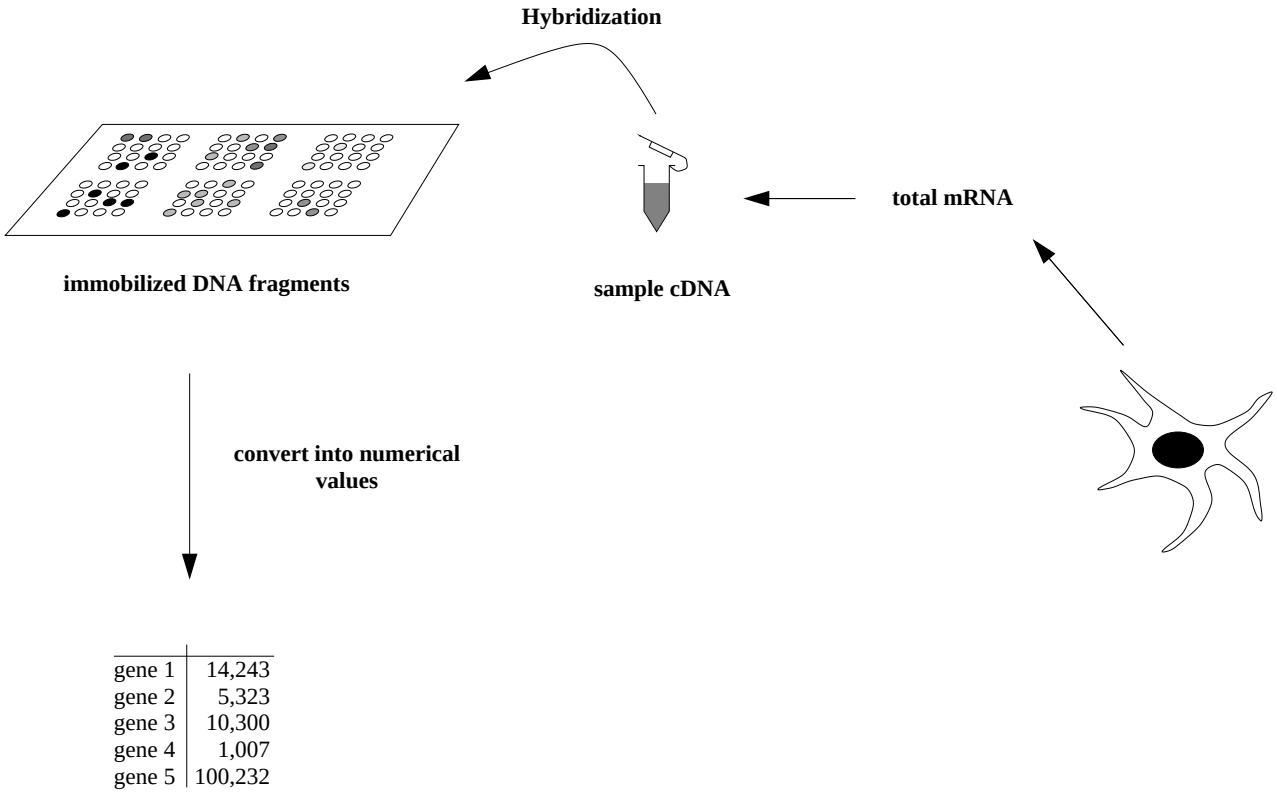
Figure 3.2: **Scheme of the SSH method.** Solid lines represent the *Rsa*I digested tester or driver cDNA. Solid boxes represent the outer part of the adaptor 1 longer strand and corresponding PCR primer P1 sequence. Shaded boxes represent the outer part of the adaptor 2 longer strand and corresponding PCR primer P2 sequence. Clear boxes represent the inner part of the adaptors and corresponding nested PCR primers PN1 and PN2. Note that after filling in the recessed 39 ends with DNA polymerase, type a, b, and c molecules having adapter 2 are also present but are not shown. From Diatchenko, 1996.

In the second hybridization, the two samples from the first hybridization are mixed together. Only the remaining normalized and subtracted tester cDNAs are able to reassociate and form (b), (c), and new (e) hybrids. Addition of a second portion of denatured driver at this stage further enriches fraction (e) for differentially expressed genes. The newly formed (e) hybrids have an important feature that distinguishes them from hybrids (b) and (c) formed during first and second hybridizations. This feature is that they have different adapter sequences at their 5'-ends. One is from sample 1 and the other is from sample 2.

3. **Selective amplification**
   The two sequences allow preferential amplification of the subtracted and normalized fraction (e) using PCR and a pair of primers, P1 and P2, which correspond to the outer part of the adapter 1 and 2, respectively. To accomplish this selective amplification, an extension reaction is performed to fill in the sticky ends of the molecules for primer annealing prior to initiating the PCR procedure. Type (b) molecules contain long inverted repeats on the ends and form stable "panhandle-like" structures after each denaturation-annealing PCR step. The resulting "panhandle-like" structure cannot serve as a template for exponential PCR, because intramolecular annealing of longer adapter sequences is both highly favored and more stable than intermolecular annealing of the much shorter PCR primers. This is the suppression PCR effect. Furthermore, type (a) and (d) molecules do not contain primer binding sites, and type (c) molecules can be amplified only at a linear rate. Only type (e) molecules have different adapter sequence at their ends which allows them to be exponentially amplified using PCR. The mathematical model and calculations describing the process of forming of fraction (e) as well as the rate of enrichment is described by Gurskaya (1996).

3.2 DNA microarrays for large scale expression profiling

Microarray technology provides insight into the transcriptional state of the cell (transcriptome), measuring mRNA levels for thousands of genes at once. Ongoing progress in sequencing promises to yield complete gene sets mounted on microarrays for many organisms of interest in the near future. It enables searching for genes that are expressed specifically under certain conditions. Data coming from different technologies of microarrays that were generated in the laboratory of Prof. Günther Schütz are analyzed in this thesis. A brief overview over these technologies is described below. Microarray experiments produce large amounts of data demanding computational methods for their analysis. Known methods are summarized below which are commonly applied to compare data from different experiments or to compare profiles of several genes. Further methods for the analysis of microarray data were developed for this thesis that improve or build a basis for these described methods.



gene 1	14,243
gene 2	5,323
gene 3	10,300
gene 4	1,007
gene 5	100,232

Figure 3.3: Schematic layout of a radioactive hybridization to a DNA microarray.

3.2.1 Methods to determine the expression levels of thousands of genes simultaneously

For simultaneous determination of transcript levels of a large number of genes representative sequences are immobilized and then cDNA samples under study are hybridized to the immobilized genes. The detection of the hybridization signal may rely on radioactive (Friemert et al., 1989) or fluorescent labeling (Shalon et al., 1996). The amount of radioactivity or fluorescence is the indicator of the amount of RNA present.

A grid of such immobilized gene sequences is referred to as DNA microarray. Parts of a cDNA sequence or a set of oligos for each gene are immobilized ("spotted") on a glass or nylon membrane.

cDNA species are often spotted in duplicate on the membrane. These spots are denoted "primary" and "secondary" spot. Nylon filters that are used for hybridization with radioactively labeled samples were better established when the experimental part of this thesis was begun and they do not require specialized hardware for the read-out of the signal (Lennon and Lehrach, 1991). A scheme of DNA microarray hybridization can be seen in Figure 3.3. The details on the spotting and hybridization procedure of cDNA microarrays that were analyzed in this thesis are described in Borisevich (2000).

oligonucleotide arrays contain a number of short nucleotide sequences (e.g. 20 sequences of 25

bps) representing the same cDNA sequence. Further, for each oligonucleotide sequence, control sequences with one base difference are added to the array to check for cross hybridization. The mean over the differences of the measured intensities of match and mismatch oligonucleotides for each cDNA is used as an indicator for its expression. oligonucleotide arrays containing 10,000 mouse genes were used that are commercially available from Affymetrix (Santa Clara, Ca, USA). The hybridization procedure has been carried out according to the Affymetrix manual instructions.

DNA microarray hybridization involves the following steps. mRNA is prepared from cells growing under specific experimental conditions. For each condition, the prepared mRNA is processed separately, performing reverse transcription with radioactively or fluorescently labeled nucleotides and hybridization onto an array. Spots on the membrane are referred to as “probe”, and the sample hybridized to the membrane-bound array as “target”, according to *Nature Genet.* **21** (Suppl.), 1999, p. 1. The target cDNA is derived from the total RNA or from poly(A)$^+$ RNA prepared from a biological sample, referred to as “mRNA pool”.

The resulting signals on the autoradiograph of each hybridization have to be quantified. Commercial software is available for detecting spots and quantifying their intensity. Typically, such software will generate a table of signal intensities assigned to the individual spots.

The data of a multiconditional or time-course experiment may be represented in a table where rows represent genes, columns represent various samples such as tissues or experimental conditions, and numbers in each cell characterize the expression level of the particular gene in the particular sample. To obtain knowledge about the underlying biological process these matrices have to be analyzed further (Vingron and Hoheisel, 1999; Brazma and Vilo, 2000).

3.2.2 Finding differentially expressed genes in pairs of conditions

Microarray-based gene expression measurements do not result in absolute amounts of mRNA counts per cell in the sample. The numbers are relative. Only the expression levels of the same gene in different samples can be compared. Comparison of the results from different hybridizations requires **standardization**. Because of different background intensities, different labeling efficiencies or differing exposure times, two (or more) hybridization experiments are not readily comparable without prior standardization. The experimental variance of a measurement may depend on the intensity of the signal and could vary from spot to spot due to specificity of the sequence and cross-hybridization of homologous sequences (Claverie, 1999).

Various methods for standardization have been suggested (Chen et al., 1997b; Piétu et al., 1996; Richmond et al., 1999). Some of these procedures require separate standardization for different regions on the nylon membrane. Other methods show that sequences spotted with the same pin of

the spotting robot behave similar and standardization procedures are designed to correct for such effects (Schuchhardt et al., 2000; Dudoit et al., 2000b).

In this thesis methods are provided to deal automatically with questions of additive and multiplicative distortion (Section 4.1.2). These methods are based on a physical model and have been successfully applied to several hundred hybridizations. Using arrays with representatives for almost all genes of a genome (*genome-wide arrays*), a subpopulation of hybridization intensities across the array can be modeled by log-normal distribution. This distribution can be used to determine a threshold of reliability for these intensities (Beißbarth et al., 2000).

Other attempts have been made to construct mathematical models for the variability of experimental ratios dependent on the measured intensity distribution (Chen et al., 1997b; Newton et al., 2000). Methods for analysis of variance (ANOVA) can be applied to standardize microarray data, correct for potential confounding effects and provide estimates of changes in gene expression at the same time (Kerr et al., 2000).

3.2.3 Detecting correlations between different genes or between different experimental conditions

Given a gene expression matrix two different types of correlations can be searched for:

1. Correlations between genes (i.e. rows in the table of expression intensities). This is of interest as genes that behave similar in their expression profile might also share functional properties (DeRisi et al., 1997; Chu et al., 1998; Spellman et al., 1998; Holstege et al., 1998; Iyer et al., 1999).

2. Differences between experimental conditions (i.e. columns) (Alizadeh et al., 2000; Golub et al., 1999).

The methods to explore these data can be distinguished in supervised and unsupervised methods. Figure 3.4 shows a conceptual illustration.

The unsupervised approach is more an exploratory way to visualize the data. A typical example of unsupervised data analysis is expression profile clustering to find groups of co-regulated genes or related experimental conditions (samples). The goal of clustering is to group together objects (genes or samples) with similar properties.

The supervised approach assumes that for some (or all) profiles we have additional information, such as functional classes for the genes, or diseased/normal states attributed to the samples. We can

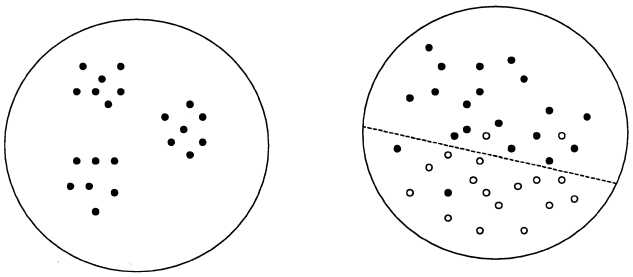


Figure 3.4: *Unsupervised* and *supervised* data analysis. In the unsupervised case (left) data points are given in n-dimensional space ($n = 2$ in the example) and points with similar features are grouped together. For instance, there are three natural clusters in the example, each consisting of data points close to each other in a sense of Euclidean distance. A clustering algorithm should identify these clusters. In the supervised case (right), the objects are labeled (e.g. filled and unfilled points in the example), and the task is to find a set of classification rules allowing us to discriminate between these points as precisely as possible. For instance, the dotted line in the drawing discriminates most of the points correctly, allowing us to predict their "labels" - filled or unfilled - by their position above or below the dotted line.

view this additional information as labels attached to the rows or columns. Having this information, a typical task is to build a classifier to predict the labels from the expression profile.

Methods for supervised and unsupervised analysis that have been applied to expression data are summarized below. In all of these methods a way to measure the similarity (or distance) between the gene- or experiment-profiles being compared are needed.

3.2.3.1 Distance measure

Gene- or experiment-profiles (rows or columns in the matrix) can be regarded as points in n-dimensional space, where n being either the number of samples for gene comparison or the number of genes for sample comparison. There is a choice of different distance measures to take depending on the features of similarity one wants to observe. It is an unsolved problem how to choose the best distance measure. The most commonly known distance is the so-called Euclidean distance. Used very often in the analysis of expression profiles is a distance derived from the Pearson correlation coefficient, which is related to the angle between the two n-dimensional vectors. Euclidean and correlation distance measures are related, if the length of the n-dimensional vectors is normalized to 1. In this thesis another distance measure is defined which is based on the comparison of two probability distributions given by the "relative entropy" (Section 4.1.2.4).

3.2.3.2 Unsupervised analysis

Clustering is not a new technique, many algorithms have been developed for it and many of the algorithms have been applied to analyze expression data. The most commonly used algorithms are described below. These algorithms have different requirements on the parameters that have to be set by the user and different limitations in the number of profiles that can be clustered due to different computation time. It is, however, unclear which of the algorithms works best for expression profile clustering.

The hierarchical clustering method was the first clustering method to be applied to gene expression analysis (Eisen et al., 1998). Hierarchical clustering works either by iteratively joining the two closest clusters starting from singleton clusters (Eisen et al., 1998) or by iteratively partitioning clusters starting with the complete set (Alon et al., 1999). After each join of two clusters, the distances between all the other clusters and a new joined cluster are recalculated. The complete linkage, average linkage, and single linkage methods use maximum, average, and minimum distances between the members of two clusters to calculate the new distances. Note that in order to obtain a particular partitioning into clusters, the threshold distance should be chosen by independent means (typically by the user).

In K-means clustering the desired number of clusters K has to be chosen a priori. After the initial partitioning of the vector space into K parts, the algorithm calculates the center points for each cluster and adjusts the partition so that each vector is assigned to the cluster the center of which is the closest. This is repeated iteratively until either the partitioning stabilizes or the given number of iterations is exceeded (Tavazoie et al., 1999).

Another clustering method that has been used for clustering expression profiles is called self-organizing maps (Tamayo et al., 1999). More recently new algorithms have been developed specifically for gene expression profile clustering, for instance an algorithm based on finding approximate cliques in graphs (Ben-Dor and Yakhini, 1999; Sharan and Shamir, 2000).

Clustering of experimental conditions has been combined with clustering of genes to identify which genes are the most important for a particular experimental condition (Alon et al., 1999; Alizadeh et al., 2000). Specialized algorithms have been developed to cluster rows and columns simultaneously (Cheng and Church, 2000).

Clustering of expression profiles can be viewed as a reduction of the dimensionality of the system. Other methods for dimension reduction have been applied. Singular value decomposition, also called principal components analysis, may be used for dimension reduction from the “gene x sample” space to the reduced “eigengene x eigensample” space. The data can be screened for those principal components that contribute most to the total variance (Hilsenbeck et al., 1999; Alter et al., 2000). Correspondence analysis is a method for data projection, closely related to principal components

analysis, simultaneously visualizing both experiments and genes as well as the associations between these two categories of variables (Fellenberg et al., 2001).

In this thesis a one dimensional projection of gene expression profiles is constructed by a linear order of the gene expression profiles which minimizes the overall distances. This order can be computed by a “traveling salesman problem” (Section 4.1.2.4) and is compared to the linear order which can be obtained from hierarchical clustering.

3.2.3.3 Supervised analysis

The goal of supervised expression data analysis is to construct classifiers such as linear discriminants, decision trees or support vector machines (SVM), which assign predefined classes to given expression profiles. For instance, if a classifier can be constructed based on gene expression profiles that is able to distinguish between two different, but morphologically closely related tumor tissues, such a classifier can be used for diagnostics. Moreover, if such a classifier is based on a set of relatively simple rules, it can help to understand the underlying mechanisms involved in each tumor. Typically, such classifiers are trained on a subset of data with given classification and tested on another subset with known classification. After assessing the quality of the prediction they can be applied to data with unknown classification.

A comparative study of several discrimination methods in the context of cancer classification based on filtered sets of genes was performed by Sandrine Dudoit (2000). Support vector machines have been applied for the classification of genes with respect to functional properties (Brown et al., 2000) and also for the classification of cancer tissue samples (Furey et al., 2000). Other methods for the prediction of cancer types such as specialized clustering algorithms (Golub et al., 1999; Ben-Dor et al., 2000) or Bayesian regression (West et al., 2000; Spang et al., 2000) have been applied.

Note that when classifying samples, we are confronted with a problem that there are many more attributes (genes) than objects (samples, i.e. experimental conditions) that we are trying to classify. This makes it always possible to find a perfect discriminator if we are not careful in restricting the complexity of the permitted classifiers. To avoid this problem one has to use classifiers as simple as possible, compromising between simplicity and classification accuracy.

3.3 Construction of a microarray containing the sequences of the CREM SSH library

In order to investigate the expression profiles of CREM dependent genes DNA microarrays were constructed that contain representatives for all unique sequences found in the CREM SSH library.

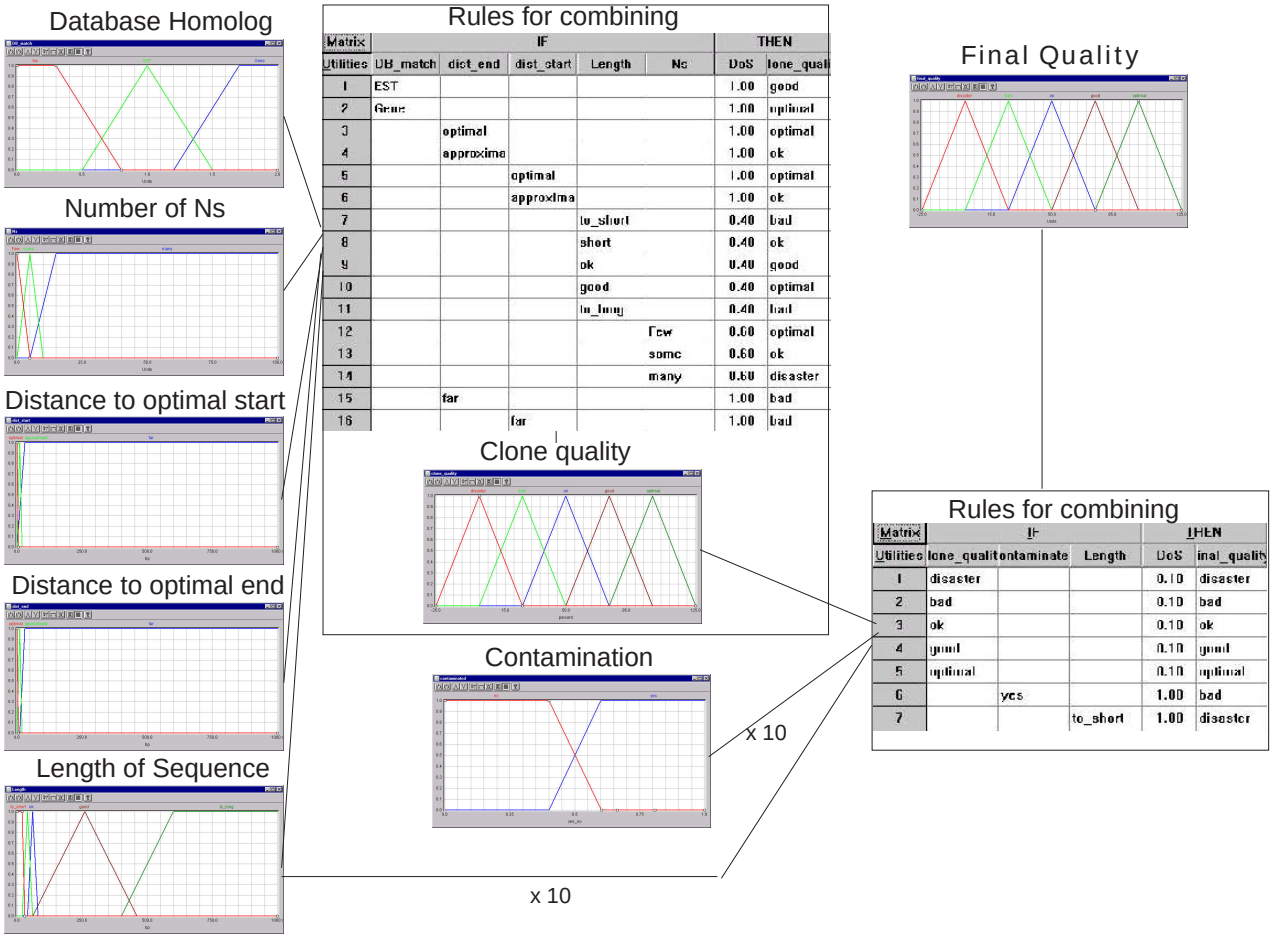


Figure 3.5: **Combination of variables in the Program FuzzyTech.** Several parameters were combined by a rule-set to get a score for the quality of each clone. Some of the parameters were combined in a second rule-set with a higher weight. These parameters usually give very strong criteria for the neglection of a clone. Finally, for each sequence cluster the clone with the highest quality score was selected.

There are 956 unique sequences represented in the CREM SSH library. These sequences represent clusters of sequenced clones which were found in the CREM SSH library. 952 of those clusters that are longer than 30 bp were used to select representative clones. For 879 of the clusters one or two clones were used in the microarray. Further control sequences were added, for example to judge the quality of the hybridization results and to enable comparisons between independent hybridizations.

3.3.1 Selection of representative clones from the CREM SSH library

The sequences of the clones of the CREM SSH library are redundant. For each of the sequence clusters a representative clone was selected and spotted on a microarray.

For the clusters which contain several clones there is a choice of which clone to pick. To choose a

clone that is a good representative, several rules were set up that characterize a “good” clone. These criteria were combined by a system called *fuzzy logic*. This is a mathematical theory that makes it possible to formulate a discrete set of rules with discrete input values and which results in one discrete output value that combines the formulated rules with a given set of weights (Zadeh, 1965). To implement the system the program FuzzyTech MCU-C (Inform GmbH, Aachen) was used. This program makes it possible to set up a rule-system via a graphical interface (Altrock, 1995). The exact definitions of the criteria and a C program for clone selection are available at a web-site[1].

Figure 3.5 shows how the combination of criteria is defined in the FuzzyTech Program. The criteria for the clone quality were combined rather intuitively and were improved during the processing. Variables were defined that represent certain criteria for the clones. These criteria were weighted and combined. A short overview over the criteria that were used and an indication of their weight is summarized in the list below. Criteria with the label *weak* influence the overall score only little, criteria with a label *strong* influence the overall score very much and usually outvote other criteria.

Criteria	**Description**	**Weight**
Database match	Clones with homolog to a database sequences are preferred as they represent known sequences.	weak
Number of *N*s	percent of unrecognized nucleotides in the sequence. If the sequencing quality is good the sequence contains few *N*s.	weak
Distance to optimal start and end position	As the clones in a cluster represent *Rsa*I-fragments they should all start or end at the same position. If there are enough clones mean start and end position are computed, and clones which start and end close to the the mean (optimal) positions are more secure.	weak
Length	For good hybridization results clones should have an equal length distribution. Clones were chosen preferentially with a length around an optimum of 200-300 bp, i.e. clones will get a higher score the closer their length is to the optimum.	weak
	However, clones with a length lower than 30 bp were never used on the microarray.	very strong
Contamination	Sequences which contain *Rsa*I sites or primers were rated very bad as they may represent concatemers of several genes.	strong

For each cluster the clone with the highest score was chosen and spotted on the DNA microarray if the clone was available. Some clones however were not accessible any more, due to technical problems. In these cases the next best clone was chosen. For some clusters no clone could be accessed any more and no representative sequence was spotted on the microarray.

[1] http://www.dkfz.de/tbi/crem

3.3.2 Selection of hybridization controls

Several controls were added to the array to check the hybridization process and to be able to standardize and interpret the results. The following types of controls were spotted:

- **Empty control spots**: Spots without DNA can be used to determine the background intensity measured on the nylon membrane.
- **Heterologous DNA**: Hybridization signals might be the result of unspecific DNA-DNA interaction. Therefore spots were added to the microarray that contain DNA from other organisms that contain sequences that should not be present in the hybridization samples. Plasmid DNA and salmon sperm DNA were used.
- **Homologous DNA**: Sequences with a known degree of homology were added to the array to check for cross-hybridization. This can be used to adjust the hybridization conditions in a way to minimize cross-hybridization. To find such sequences the AXELDB database of frog clones (Pollet et al., 2000) was screened for homologies to the clones from the CREM SSH library. Frog clones without any homology were chosen as well as some clones that display 80-90% identity over a stretch of more than 100 bp. 12 frog clones were chosen of which 5 show a homology to clones from CREM SSH.
- **House-keeping Genes**: To compare data from DNA microarrays which were hybridized with cDNA from different sources standardization is necessary. As most of the genes from the CREM SSH library are expected to be differently expressed under the experimental conditions an additional set of non-differentially expressed genes was spotted on the DNA microarrays in order to be able to standardize the hybridization data. The number of non-differentially expressed genes must be sufficient for statistically significant standardization. We searched for potential house-keeping genes in the literature. A list of human house-keeping genes in the analysis of human cancer was published by de Risi *et al.* (1996). Another list of genes observed to be frequently non-differentially expressed genes in human was provided by Dr. Bernhard Korn. The GeneNest EST databases were used to search for homologous genes in mouse, and the representative clones were ordered from the resource center of the German Human Genome Project (RZPD, Berlin). These clones were partially sequenced by one sequencing run, and the sequences were screened for homology to known genes. 54 of these clones in fact represented potentially non-differentially expressed genes and were used for spotting. They belong to different functional classes, such as metabolic enzymes, transcription and translation factors.
- **Differentially expressed controls**: The differentially expressed control clones spotted on the filters represent genes described in literature as CREM-dependent. These genes are ex-

pressed specifically during the post-meiotic stages of spermatogenesis. Among those are the angiotensin converting enzyme (ACE) and the spermatid nuclear transition protein 1 (TP1). These genes are known to be direct CREM targets (Goraya et al., 1995; Zhou et al., 1996). Protamine 1 is expressed specifically in round spermatids. All three genes are not expressed in CREM-deficient mice (Blendy et al., 1996). The sequences for those genes were either cloned from RT-PCR products, or EST clones homologous to these genes were ordered from the resource center of the German Human Genome Project (RZPD, Berlin).

3.4 Mining information from databases of genomic or EST sequences

Public databases were used to retrieve sequence information and gene annotations of sequences corresponding to the sequences found in CREM SSH library. In the last decade increasing amounts of sequence information have become available. Databases which collect comprehensively DNA sequence data are the EMBL nucleotide sequence database, the GeneBank database and the DNA Database of Japan DDBJ (Stoesser et al., 2001). Each of this databases collects all DNA sequences published. However, the sequence information in these databases is highly redundant, the annotation of the sequences is variable and often incomplete or inconsistent.

Many databases aim to provide fewer sequences but better annotation of the sequences. These can be divided into databases that collect and annotate the complete genome of one organism and databases that collect information of all expressed sequences or known genes for different species. For example, the ENSEMBL project collects information about the human genome and adds automatic annotation to it (Birney et al., 2001). The Swiss-Prot database is a collection of well characterized proteins (Bairoch and Apweiler, 2000). The annotation of Swiss-Prot is however not always up to date as the proteins have to be manually added and annotated. The GeneCards database attempts to collect information about all known human genes automatically from other databases (Rebhan et al., 1998).

A large portion of the sequences in public databases are so-called ESTs. Expressed sequence tags (ESTs) are short pieces of an expressed cDNA sequence. Many EST sequences are generated in large scale sequencing projects. For the generation of ESTs the RNAs are extracted out of a certain tissue or cell type, reversely transcribed and cut into pieces. Mostly poly(A)$^+$ purified sequences are used. Part of the cDNA is sequenced by one sequencing run from one side or by two sequencing runs from both sides of the sequence. ESTs provide a good resource to get sequence information of yet unknown genes. They were introduced by Adams *et al.* (1991). Although ESTs may be of low sequence quality they are useful for detecting new genes, determining the genomic structure of

a gene (exon-intron boundaries, alternative splicing) (Mironov et al., 1999) or for expression studies (Schmitt et al., 1999).

Because EST sequence information is highly redundant a single gene may be covered by many ESTs each representing different parts of that gene. In order to simplify the analysis of specific genes several efforts have been made to cluster sequences belonging to the same gene resulting in so-called gene indices. Some commonly used gene indices are UNIGENE (Schuler, 1997) at NCBI (National Center for Biotechnology Information), TIGR (The Institute for Genomic Research) gene indices (Adams et al., 1995) and STACK (Burke et al., 1998) at SANBI (South African National Bioinformatics Institute). In particular UNIGENE and the TIGR gene indices differ mainly in the clustering strategy used and in the presentation of cluster related information (Bouck et al., 1999). We developed GeneNest, a software and database for automated generation and visualization of gene indices (Haas et al., 2000).

Moreover many local similarities can be found between functionally or evolutionary related proteins. Further clustering of protein sequences can be performed to obtain information about relations between proteins. A database that uses local similarities between proteins to cluster proteins into families is the SYSTERS database (Krause et al., 2000). Other approaches first split the proteins into domains. The Pfam database uses manually aligned domains to scan other proteins for the appearance of these domains (Bateman et al., 2000).

3.5 Software used

The following software packages were used. All programs were run on a Sun Ultra 5 workstation under the Solaris operating system, except the ABI software which requires a Macintosh Computer and the AIS and FuzzyTech software which were used on a Pentium PC under Windows NT.

1. **ABI Prism**: ABI sequencing software for lane tracking and base calling. Perkin Elmer Applied Biosystems.

2. **ACEDB** (Version 4.7): Database system specialized for biological information developed by Jean Tierry-Mieg and Richard Durbin. Available at http://www.acedb.org/.

3. **AIS** (3.0, Array Vision Module): Software for image analysis of phosphoimager data. Automatically detects spots of the array and calculates intensity values. Imaging Research, Ontario, Canada

4. **Apache** (Version 1.3): Free World Wide Web information server. This is the source for documents and presentations in the WWW. Documents are stored in the **HTML** format (Hypertext

Markup Language, Version 4.0 specifications). Programs are invoked on the server side via the common gateway interface (**CGI**). The software is available at http://www.apache.org.

5. **BLAST** (NCBI Toolbox): Fast algorithm to search with a protein or DNA query sequence against a database of sequences. Available at http://www.ncbi.nlm.nih.gov/BLAST/.

6. **FuzzyTech** (4.13 MCU-C Edition): Program to graphically set up a *fuzzy logic* rule system. Inform GmbH, Aachen

7. **Matlab** (Version 5.3): Interpreted programming environment especially suited for matrix calculations. Furthermore, the **Statistical Toolbox**, a package for statistical analysis, was used. MathWorks Inc., MA, U.S.A.

8. **Staden Package** (Version 4.4): Package containing programs for the assembly of DNA sequences. Available at http://www.mrc-lmb.cam.ac.uk:80/pubseq/.

9. **Perl** (Version 5.004): Perl is an interpreted high-level programming language developed by Larry Wall. Available at http://www.perl.com/.

4 Results

4.1 Developments in Bioinformatics

4.1.1 Use of expressed sequence tags to generate gene index databases

Many of the sequences in the CREM SSH library could not be linked to a gene of known function. These sequences represent expressed sequence tags of unknown genes, i.e. the sequences in the CREM SSH library are ESTs. The sequences of the CREM SSH library were compared to the ESTs in public databases, as the sequences in public databases are often longer and of higher sequence quality.

ESTs provide a good resource to get sequence information of yet unknown genes. Public databases contain millions of EST sequences. To make efficient use of the EST data the redundancy was reduced by clustering. Each cluster of ESTs can be represented by a single consensus sequence and ideally constitutes the complete and unique sequence for a gene. Therefore, clustered EST databases are often referred to as gene indices. We developed an automatic procedure to generate gene indices. Methods to store, visualize and search these gene indices were developed and are described below.

Databases of clustered ESTs for mouse and man which have been generated were used in the analysis of sequences in the CREM SSH library.

4.1.1.1 Generation of gene indices

The generation of the GeneNest gene indices starts either with a database of sequences extracted from the EMBL database (in the case of mouse), or from a UNIGENE database of already clustered ESTs (in the case of man) (Figure 4.1). All sequences are subjected to clipping based on an extensive quality check. The repeats, vector sequences as well as low quality regions, that were detected during the quality clipping, are masked. Similarities between the cleaned-up sequences are then determined

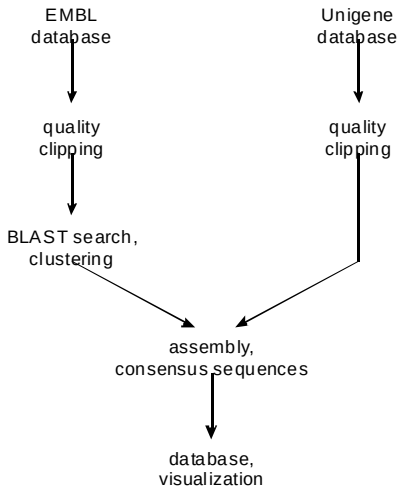


Figure 4.1: **Generation of GeneNest indices.** Two alternative methods were used. The human EST database is based on the clustering from the UNIGENE project. For the mouse ESTs the EST data were extracted from the EMBL database and a clustering algorithm based on BLAST searches was used.

using BLAST (Altschul et al., 1990). Applying the BLAST program, each of the sequences is used as a query against all the other sequences. The sequences are clustered together when there is a near perfect match extending over at least half of the shorter sequence. Sequences in a cluster are assembled in order to determine their relative positions and to obtain a representative consensus sequence. A cluster may be split into several contigs each reflecting a group of sequences sharing global similarity. Splitting is often caused by alternative splicing, ESTs derived from nuclear RNA or other artifacts like chimeric sequences. In a final step, a web site presenting all these data is generated automatically.

4.1.1.2 Update of the gene index databases

The databases of sequences, ESTs and known genes are still rapidly growing and changing. This requires regular updates of the clustered databases. However, adding new sequences might also affect the clustering: previous clusters can either be joined or fall apart. In order to keep the information transparent, links are added in the old version of the database to the corresponding clusters in newer versions of the database. For each of the sequences in the old cluster, it is tested, in which new cluster this sequence appears. This information is summarized and annotated in the old cluster.

4.1.1.3 Querying of gene indices and storage of the data

To be able to use the gene index, methods were developed to search the database and to access the information. A text file format was designed to store and access the data in a quick and portable

```
ID          Mm_TBI2156
NEWCLUSTER  Mm2:MM_TBI4041,MM_TBI4066
CONTIGS     4
CONTIG      1
NEWCONTIG   Mm2:Mm_TBI4041.1
LENGTH      4712
SEQUENCES   8
GENE        ACC=J04947; POS=1; LEN=4038; DIR=0; INFO="Angiotensin-converting enzyme mRNA";
GENE        ACC=J04946; POS=1; LEN=3191; DIR=0; INFO="Angiotensin-converting enzyme mRNA";
SEQUENCE    ACC=J04947; POS=695; LEN=4018; DIR=0;
SEQUENCE    ACC=AA895502; LID=284; CLONE=1297351; END=5'; POS=1877; LEN=571; DIR=0;
SEQUENCE    ACC=AA146321; LID=285; CLONE=602381; END=5'; POS=2541; LEN=537; DIR=0;
SEQUENCE    ACC=W82778; LID=217; CLONE=404004; END=5'; POS=2575; LEN=509; DIR=0;
SEQUENCE    ACC=W82801; LID=217; CLONE=404347; END=5'; POS=2750; LEN=442; DIR=0;
SEQUENCE    ACC=AA871429; LID=564; CLONE=1096232; END=5'; POS=2831; LEN=326; DIR=0;
ORF         FRAME=1; START=1; LEN=2943; INFO="Mm_TBI2156.1_orf1.1 1-2943 1-2943";
ORF         FRAME=1; START=3556; LEN=291; INFO="Mm_TBI2156.1_orf1.2 3556-3846 3556-3846";
PROTSIM     ACC=P408; ... INFO="mammalian peptidyl-dipeptidase A superfamily";
CLUSTERSIM  ACC=Mm_TBI2156.2; START=941; END=3063; EVALUE=0.0; SUBJ_LENGTH=3127; ...
CLUSTERSIM  ACC=Mm_TBI2156.2; START=4; END=146; EVALUE=0.0; ...
CONSENSUS   GAATT ... GAATTC
//
```

Figure 4.2: **Data format to store the information of an EST cluster.** Each cluster has a unique ID. The position of the cluster in new versions of the database is indicated. A cluster can fall into several contigs, for each sequence in the cluster the position in the assembly is stored. Further information about open reading frames (ORFs), the homology to known protein families in the SYSTERS database, and possible local homologies to other clusters is stored.

manner. An example EST cluster in this data format is shown in Figure 4.2.

To quickly access the information about a cluster in the text file, several index files are generated. The start position of each cluster in the text file is stored. Lists of all the sequence accession numbers, the clone identifiers or keywords are stored in alphabetical order and can quickly be searched and used to find the corresponding cluster. A database to perform BLAST searches against the consensus sequences was constructed. The methods to search and access the data as well as the generation of the web presentation are implemented in the programming language “Perl”.

4.1.1.4 Visualization of gene index data

The GeneNest EST clusters are visualized in a web based interface. The visualization displays the data of the clustering and the EST data and is linked to external databases for further analysis of the sequences. An interface was developed that allows to search for specific clusters and the results are visualized.

The view of a contig of similar sequences is visualized in Figure 4.3. All sequences are sketched by an arrow indicating the direction of this sequence. Additionally, the type (mRNA/gene or EST) or annotated direction of a sequence is reflected by different colors. If two ESTs are derived from the same clone they are connected by a grey line indicating the putative clone sequence. Specific

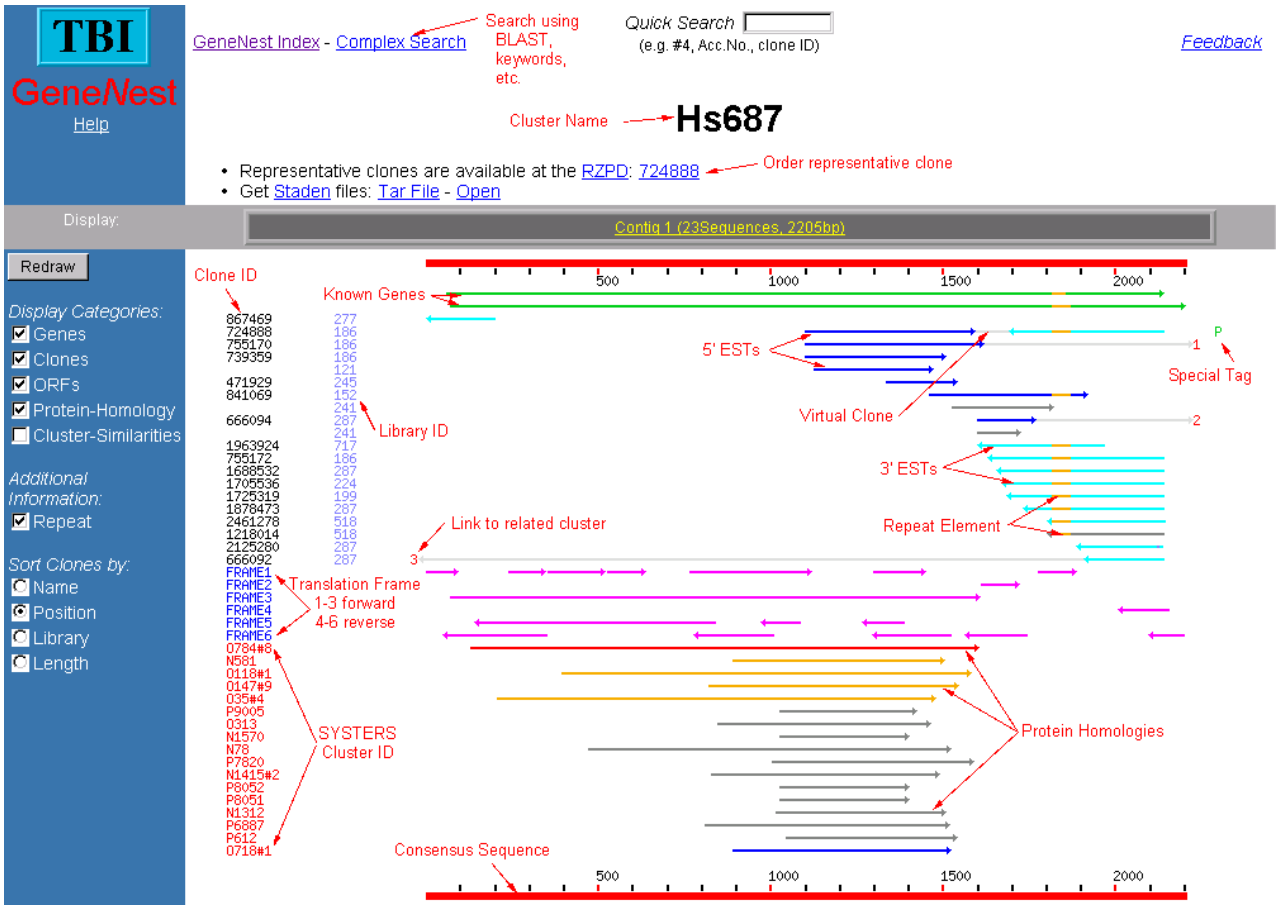


Figure 4.3: **GeneNest visualization.** A graphical overview of the clusters is presented through an interface in the world wide web. The assembly information about each contig of a cluster can be displayed. Further information about the coding regions, homologies to protein families and other clusters is visualized. The information is linked to other databases, such as EMBL, GeneCards and SYSTERS.

features like repeats or poly(A) signals are also color coded. As far as possible, clone identifiers are directly linked to institutions where these clones can be ordered. Clones labeled by a *P* are part of a non-redundant clone set available at the Resource Center of the German Human Genome Project (RZPD). Each contig is represented by a single consensus sequence summarizing the sequence content of this contig. As the consensus sequence should reflect a single mRNA it is likely to find at least a part of the coding region for the gene in an open reading frame (ORF). Putative ORFs longer than 30 bases are symbolized by arrows providing the predicted amino acid sequence. In order to determine local similarities between clusters or contigs, sequence homologies to other contigs/clusters are displayed. Information about potential function or membership to a family of proteins is provided by precomputed protein homologies to the SYSTERS protein consensus sequences (Krause et al., 2000). These homologies are marked by an arrow where the respective color indicates the degree of sequence similarity.

All items visualized can be accessed interactively by clicking on the appropriate symbol thus linking to more detailed information, databases or related institutes. In the case of contigs composed of a large number of sequences the visualization of features can be turned on or off, optionally allowing the user to focus only on part of the data.

4.1.2 Analysis of cDNA microarray expression data

DNA microarray hybridization is a new technique to determine expression profiles of thousands of genes at the same time.

Differentially expressed genes should be detectable by comparisons of hybridization data from pairs or series of experiments provided that the signal intensities are related to the proportion of the complementary mRNA in the mRNA pool. However, differences in signal intensities might not only be due to true expression changes but also to experimental variabilities, which are often in the same range as the differences one expects to occur by differential expression. To find differentially expressed genes or to do pattern analysis careful correction for influences on the experiment, like incorporation of radioactive label or exposure time, is needed. Methods for standardization of microarray expression were developed and are described below.

These methods were tailored for hybridization of radioactively labeled targets to DNA arrays spotted on nylon membranes. They allow to compare the intensity values of several hybridization experiments. Apart from the application in the CREM project these methods were tested on several hundred hybridizations with data produced on arrays of mouse, man, yeast and arabidopsis genes. The same standardization procedures could also be successfully applied to data generated by hybridization of fluorescently labeled targets to Affymetrix oligonucleotide arrays.

4.1.2.1 Extracting numerical data from an image

The amount of radioactivity on the membrane was measured by means of a phosphoimager and converted into gray levels in an image. Gray levels are supposed to be linearly related to the amount of radioactivity on the filter over the range of measured numbers. Every spot of the array needs to be recognized and assigned to its position in the array, *i.e.* to the corresponding clone number. For each of the spots in the array an intensity value needs to be assigned. Due to the large number of spots only automatic or semiautomatic procedures are suitable for this task. Images were analyzed by the AIS program (Array Vision, Imaging Research, Ontario, Canada) installed on a PC. Values for all spots were saved in text files.

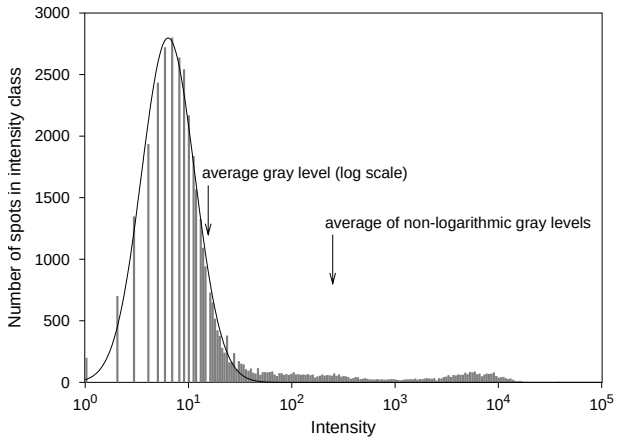


Figure 4.4: Histogram of gray level distribution across an array of 17,280 EST clones (Human UNIGENE Collection) hybridized to a complex target. A normal probability distribution was fitted by an iterative non-linear regression method and is shown overlaid.

4.1.2.2 Separating expressed and non expressed genes of one microarray hybridization

There are several sources of background signal for nylon filter hybridization: the nylon filter by itself absorbs some radioactivity during hybridization. Another source of background is the imaging plate, which gradually absorbs background irradiation from the environment. Phosphoimager screens are particularly sensitive and will make non-zero background intensity visible. The background may be inhomogeneously distributed, in which case it is referred to as “local background”. The AIS ArrayVision array analysis program allows to choose various methods of background correction. In the analyses performed here, background was evaluated locally around small areas on the array and subtracted individually from the corresponding spots. Furthermore, the 5% quantile, which corresponds to the number where 5% of the data have smaller values, was subtracted as background, if this number was positive.

Besides subtracting the background intensity of the nylon membrane it is necessary to determine a threshold of reliability, which is used to filter the genes that can be trusted to have a reliable expression level in the hybridization experiment. The values of *empty spots* and spotted *heterologous* DNA controls can be used to control a threshold of reliability, which marks the values we want to trust in later analysis.

Generally, in a genome-wide array, for most spots there will be no complementary mRNA species in the complex target because only a comparatively small number of genes is expressed in the biological sample under investigation. Yet, such spots may display a marked signal intensity in the hybridization. Looking at the histogram of the gray level distribution for all spots on an array (Figure 4.4), we frequently observe two populations of spots. On a logarithmic scale, the gray levels in the low-intensity region follow a normal distribution comprising for most arrays more than 80% of

all genes. This means that the majority of signals roughly follows a log-normal distribution typical for data that are centered at a minimum near zero and cannot extend below zero (Sachs, 1984). The logarithms of the remaining intensities extend to the right of the Gaussian curve; this gives the histogram a much heavier tail than expected from random data alone. The population, that is within a normal distribution of intensities, represents the set of those spots that do not have a complement in the target or where the number of transcripts is below the detection limit. As the distribution of the signals for this population, mainly due to unspecific interaction with target DNA fragments, and the distribution of expression levels of all genes overlap, we are not able to distinguish lowly expressed genes from genes which are not expressed.

For most experiments done on genome-wide arrays a normal distribution function can be fitted to the values of the low-intensity class by an algorithm which starts by fitting a Gauss normal distribution function to the histogram by means of non-linear regression. To distinguish between "noise" that needs to be modeled and signal that does not obey the normal distribution, the data set used for fitting is then iteratively reduced by truncating above a certain threshold. This threshold is calculated by adding one standard deviation to the mean of the calculated distribution. This is followed by a new fitting based on the reduced data set. The iteration stops when the mean of the distribution stays constant. This fitted normal distribution provides a rational approach for the identification of those genes which are actually expressed. Obviously, even for higher intensities there is still a positive probability that such a signal might be due to chance. This probability is related to the area under the normal distribution above a certain threshold.

4.1.2.3 Standardization procedure to compare pairs of experiments

When comparing two data sets originating from different hybridization experiments, two types of systematic differences can be noticed: one type is background, which has an additive influence with respect to the measured intensities. The other systematic difference is a constant multiplicative factor between the intensities of genes of two hybridizations, probably due to different labeling rates of the complex probe used for hybridization or to unequal exposure times of the filters.

A good way to visualize the comparison of two hybridization experiments is a scatter plot (Figure 4.5). The intensity of every spot in Experiment 1 is plotted against its intensity in Experiment 2. It is appropriate to use logarithmic scale because one is interested in intensity ratios rather than absolute differences. The plot can be subdivided into regions with different interpretations: data points in the lower left corner represent genes which are inactive or are expressed only at a low level. In most experiments these constitute the vast majority of signals. In the upper left and lower right corner, points correspond to genes which are only expressed in one experiment and not (or not discernibly) in the other. The intensity ratios calculated for these regions are poorly reproducible. Genes in the channel around the diagonal are detected as expressed in both experiments. The farther

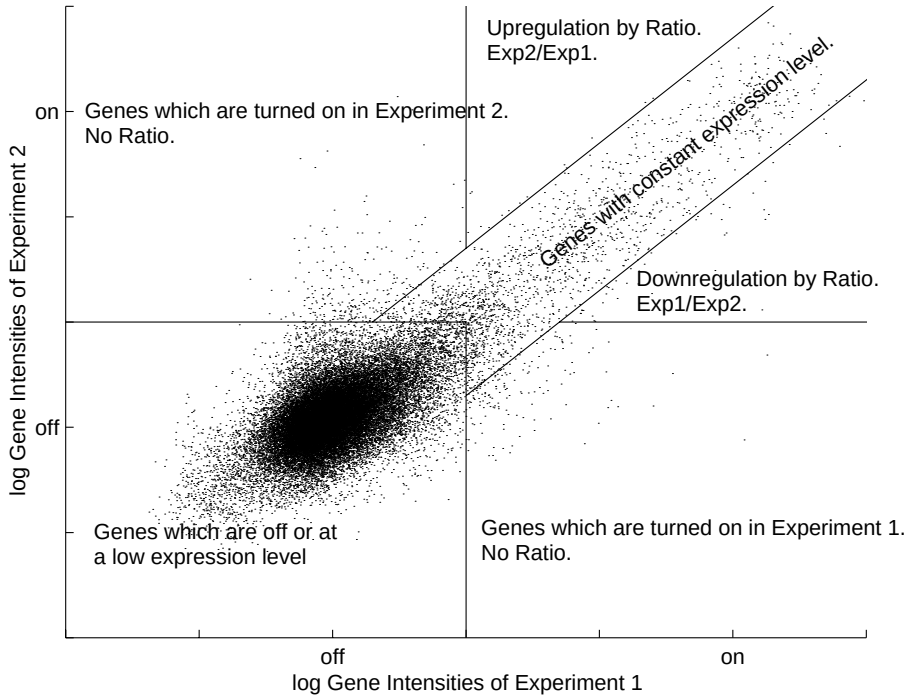


Figure 4.5: Scheme of a scatter plot using logarithmic scale. The hybridization intensities of one experiment are plotted on the ordinate while the intensities of a second hybridization experiment are plotted on the abscissa. Each point in the graph represents the intensities of a sequence measured in the two experiments. A schematic representation of different regions in the plot is drawn in the figure, and interpretations for the different regions are given. (Note: The example data plotted here were derived from two hybridization experiments on an array containing 18,432 mouse ESTs to targets derived from mouse thymus with (ordinate) or without (abscissa) stimulation by dexamethason.)

a gene is away from the diagonal, the higher its intensity ratio between the measurements in the two experiments.

First the effects of background are eliminated by subtracting an additive constant, or offset. If the image analysis software does reliably measure background intensity, these values are used for correction. Otherwise, the offset can be robustly and rapidly estimated by taking the 5% quantile of intensity values in either data set and subtracting it from all corresponding intensity values. The influence of background on data sets is outlined in Figure 4.6. The scatter plot of two experiments with highly different background is distorted (Figure 4.6c) to yield an arc-shaped cloud of points around the identity line when plotted on a logarithmic scale. This distortion is corrected by subtraction of an offset (Figure 4.6d). After background correction, one frequently observes values less than or equal to zero. To avoid numerical problems with these values when taking logarithms, they are replaced by a small positive number.

Intensity values close to background raise additional problems since they usually display an unfa-

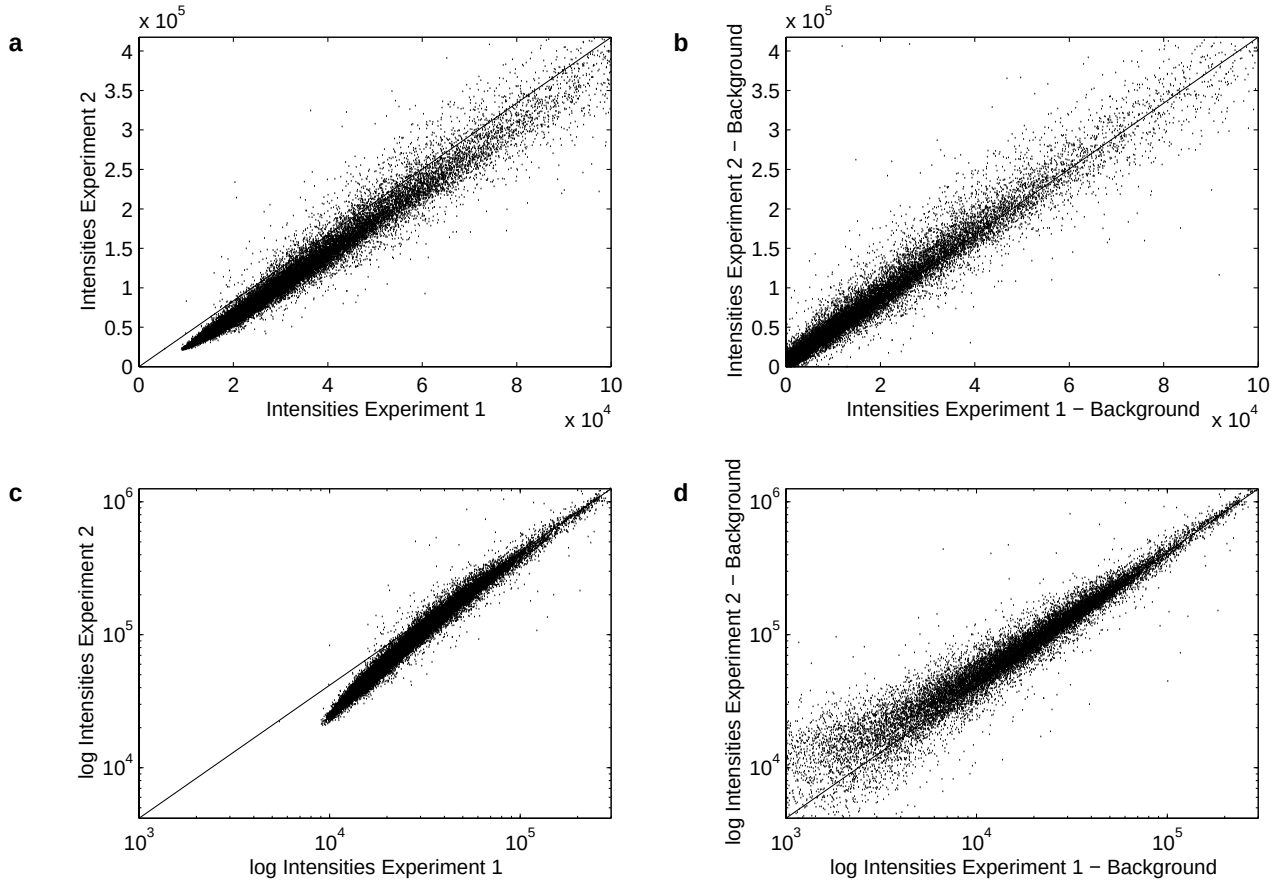


Figure 4.6: Influence of correction for background on expression profile data. Two hybridizations with an oligonucleotide recognizing all spots are compared, carried out on an array of 13,824 Arabidopsis ESTs at differing temperatures. The hybridizations show highly different background intensity. **a**, **c** scatter plots of raw data with linear or log scale, respectively. **b**, **d** scatter plots of background corrected data with linear or log scale, respectively. The raw data **c** in log scale display an arc shape which has been corrected by subtracting the 5% quantile (i.e. 5% of all values are smaller than the 5% quantile) **b**, **d**.

vorable signal-to-noise ratio, leading to highly unreliable intensity values. Ratios formed with such values can get very high numbers even when there is no significant difference in the expression levels of the corresponding genes. An intensity threshold is defined in order to exclude these spots from being marked as “differential”. The following procedure is used to determine a rough estimate of this threshold based on the comparison of two experiments. If only high-intensity values are included in the computation, the linear correlation coefficient of intensities in Experiment 1 relative to intensities in Experiment 2 will increase when more and more intensity values are added to the analysis. Including lower intensity values, the point at which the linear correlation coefficient starts to decrease is chosen as the threshold.

The next step is to find the systematic factor of change. Therefore, a set of genes is needed which

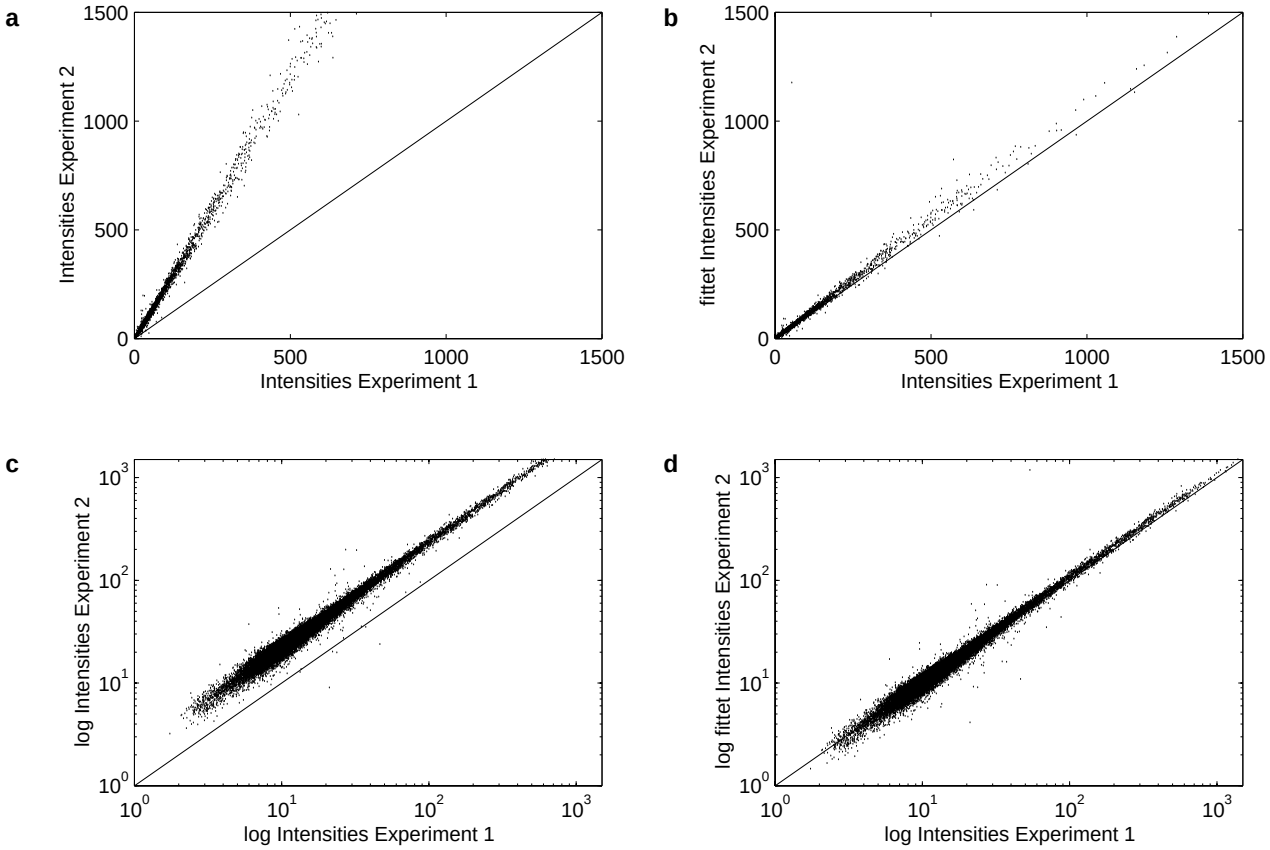


Figure 4.7: Standardization of two experiments with highly different average intensity level. The data are from one hybridization of a complex target to a mouse array, with different exposure times to the phosphoimager screen. **a**, **c** Data before standardization (linear or log scale, resp.). **b**, **d** Standardized data.

we believe should have an equal or similar expression level in the experiments we want to compare. Using genome-wide arrays, it is a good assumption that the expression level, and hence the signal intensity for most spots does not change when comparing closely related experiments. When comparing more distantly related experiments or using arrays which are biased towards a selection of genes where a lot of changes are expected this assumption does not hold true, for example, using an array with genes selected from the CREM SSH library, it is expected that most genes show changes of expression. To compare these kinds of experiments one needs to be able to define a set of control genes, where no changes are expected. If a set of house-keeping genes can be defined, these are used to adjust the intensity values. These genes are believed to be expressed constitutively at a constant level, independent of the conditions of the experiment. Another method relies on externally added controls, *i.e.* heterologous DNA spotted on the filter that hybridizes with a complementary sample added to the complex target.

The intensities of the experiments have to be adjusted such that the ratio of intensities for these control genes becomes 1 (Figure 4.7). To estimate the factor of change between experiments, we

compute the arithmetic mean of the logarithmic differences for the set of control genes. The intensities of, *e.g.* Experiment 2, can then be adjusted to be on the same scale as the other experiment by subtracting this mean from all intensities of Experiment 2:

$$\ln e_{2,k} - \frac{\sum_{i=1}^{n}(\ln e_{2,i} - \ln e_{1,i})}{n}$$

for each intensity $e_{2,k}$ ($k = 1, \ldots, n$). In this equation, $e_{2,\cdot}$ refers to the intensity data of Set 2, $e_{1,\cdot}$ to those of Set 1, and n is the number of spots on the filter. To make the results less sensitive to outliers, the arithmetic mean may be replaced by the median. Genes below an intensity threshold, which display a considerable variance in the intensity ratio, should not be included in the calculation of the mean or the median.

The statistical analysis routines described here have been realized in MATLAB 5.3 (MathWorks Inc.) and are available through a web-based interface[1].

4.1.2.4 Expression profiling with series of experiments

In the comparison of several data sets, *e.g.* a time course, a concentration series or a collection of mutants, standardization is equally required. A standard has to be defined when a control condition has been repeatedly analyzed. In this case a virtual standard that is obtained by taking the gene-wise median of intensity values across all replicates of a hybridization under the control condition is used. All values, including those of the control condition hybridizations, must then be standardized to this virtual standard. The values of repeated experiments are standardized as described before the values were combined using the median.

Distances between expression profiles

Gene expression profiles can be compared based on their shape. Genes which have a similar pattern of up- and down-regulation might also have similarities in their regulation or share functional properties. To be able to compare expression profiles a measure of similarity or distance is needed. A distance measure known in statistics as the **relative entropy** was adapted and used here. The relative entropy is a measure for the probability that values of two sample distribution are taken from the same distribution (Haussler and Opper, 1997).

First the measured expression intensities have to be adjusted to reflect the profile shapes. As expression measurements are relative the absolute intensity is not relevant for comparison. We observed that several fragments of the same gene display similar profile shapes but at very different absolute intensity levels. To make the measured intensity independent of the absolute level the expression

[1] http://www.dkfz.de/tbi/services/matlab2web/webdiffs

profiles are **normalized**, i.e. the profile is divided by the sum of the profile. Logarithmic intensities of the expression data are being used. Given a vector $G = g_1, \ldots, g_n$ of standardized intensities for the expression levels of a gene the normalized vector is calculated by

$$\texttt{for } k = 1 \ldots n: \; g_k \leftarrow \frac{\ln g_k}{\sum_{i=1}^{n} \ln g_i}$$

Next the relative entropy can be calculated for each pair of expression profiles, which can be regarded as two discrete probability distributions. For two discrete probability distributions $P = (p_1, \ldots, p_n)$ and $Q = (q_1, \ldots, q_n)$, the relative entropy between P and Q is defined by

$$D(P \parallel Q) = \sum_{i=1}^{n} (p_i \ln \frac{p_i}{q_i})$$

The relative entropy is not symmetric, i.e. the value calculated from P depending on Q is different from the number calculated from Q depending on P. To be able to use the relative entropy as a distance measure the sum of the relative entropies of P and Q to the center of P and Q is calculated. The **symmetrized relative entropy** can be calculated by

$$D(PQ) = D(P \parallel \frac{P+Q}{2}) + D(Q \parallel \frac{P+Q}{2}) = \sum_{i=1}^{n} (p_i \ln \frac{2p_i}{p_i+q_i} + q_i \ln \frac{2q_i}{p_i+q_i})$$

The symmetrized relative entropy has been calculated between each pair of genes.

Sorting the gene expression profiles by minimizing the distances of neighbors

As an alternative for clustering gene expression profiles an algorithm has been developed that rearranges the genes in a linear order. The linear order is constructed in a way so that the sum of all pairwise distances of neighboring genes is approximately minimized. The natural way to accomplish this is to convert the task into a so called *Traveling Salesman Problem* (TSP).

The TSP is a classical problem in computer science. Given a number of cities and a table of distances the task is to find a shortest round trip for the salesman through all the cities. This can be easily converted into the problem of finding a *traveling salesman tour*, where the salesman starts in one city and visits each city once. To do this a virtual city is added to the distance matrix which has a distance of 0 to all other cities. After solving the TSP this city is chosen as the start city. It is always shorter to go from one city directly to another instead of going through a third city because of the triangle inequality, and therefore each city is only visited once.

Although the TSP is a frequently appearing problem, no efficient algorithm is known to solve the problem. The computational time needed to solve it exactly increases exponentially with the number

of cities. It belongs to the class of so called NP-hard problems. Characteristic for this class is that once a solution is found it can be verified in polynomial time but there is an exponential number of possible solutions. Up to date it could not be proven that no algorithms exist to solve this kind of problems in polynomial time, but also no polynomial algorithms could be given (Wegener, 1999).

As it is impossible to compute an exponential time algorithm for several hundred genes, we used a heuristic to get a close to optimal solution. For the TSP several heuristics exist that are likely to give good solutions for well behaving inputs. The *simulated annealing* method is a fast and powerful method.

The algorithm to solve a TSP by simulated annealing can be outlined as follows:

1. **Configuration**: The genes are numbered $i = 1, \ldots, N$. A configuration is a permutation of the number $1, \ldots, N$, interpreted as the order of genes.
2. **Rearrangements**: The moves consist of two types: (a) A section of the path is removed and then replaced with the same genes running in the opposite order; or (b) a section of the path is removed and then replaced in between two genes on another, randomly chosen, part of the path.
3. **Objective Function**: Minimize the total length of the journey.
4. **Annealing**: Generate random rearrangements. Choose a rearrangement as current solution if its path length is shorter then the old one, or with a given random probability depending on the amount of change and the current temperature. With lower temperature only smaller changes will be allowed. Start with a high temperature T, proceed downward in multiplicative steps each amounting to a 10 percent decrease in T. Hold each new value of T constant for, e.g., 100 reconfigurations, or for 10 successful reconfigurations, whichever comes first.

The algorithm was implemented in the C programming language based on the algorithm sketched in “Numerical Recipes in C” (Press et al., 1992). It was compiled as a Matlab function using the Matlab *mex* compiler and used for calculations from Matlab. The simulated annealing program was run 100 times on each data-set with different starting configurations to obtain the best result. This leads to a one dimensional projection of the data and to an intuitive order in which to visually inspect the resulting data.

4.2 Analysis of expression data from a CREM SSH library

In order to identify the genes differentially expressed in CREM-deficient mice we constructed a subtracted and normalized library using the SSH technology (Section 3.1). The mRNAs from testes of CREM (-/-) mice were subtracted from mRNAs of wild-type mice. The resulting cDNA pool should be enriched with sequences which are down-regulated or missing in the CREM-deficient testes.

The clones of the CREM SSH library were analyzed either by sequencing or by hybridization to nylon arrays where the complete set of 12,000 clones has been spotted. About 12,000 clones were picked from the CREM SSH library and analyzed by sequencing or subjected to redundancy reduction by hybridization. Clones found to be redundant in sequencing were hybridized with high-density filters that contain spots for all clones from the CREM SSH library. The hybridized clones were identified and excluded from the further sequencing. A total of 3400 clones were sequenced. The other clones hybridized to at least one of the selected “probes”.

The “probes” are clone sequences or corresponding full length sequences of a gene that were selected after a sequence was shown to appear redundantly during sequencing of the CREM SSH library. 128 sequences were selected as probes, labeled radioactively and hybridized to nylon arrays where the complete set of 12,000 clones was spotted. For example, a special oligonucleotide sequence complementary to the empty cloning site of the plasmid vector, that was used to clone the CREM SSH library sequences, was used as a hybridization probe to identify clones that contain empty vectors. 490 clones hybridize to this oligonucleotide sequence and were excluded from sequencing. The most abundant probe tested was GAPD-S which hybridizes to 751 clones. Another very abundant probe corresponds to one fragment of the α-sarcoglycan gene that hybridizes to 603 clones.

4.2.1 Analysis of DNA sequences from the CREM SSH library

Sequences resulting from SSH of CREM (-/-) testes cDNA from wild-type testes cDNA were analyzed. In the SSH protocol the cDNA sequences are cleaved with the restriction enzyme *Rsa*I which recognizes the four base pair sequence GTAC and cuts fragments with blunt ends. The resulting fragments are ligated into vector plasmids (pBS) and cloned into bacteria (E. coli). Many of the clones from the CREM SSH library were partially sequenced. Sequencing was performed with one run of sequencing starting from one end of the *Rsa*I-fragment only. Some sequence analysis procedures were adapted to analyze the sequences from the CREM SSH library.

Table 4.1: **Sequencing Statistics**. The types of contamination found in the preprocessing steps of the sequences in the CREM SSH library are summarized and the number (#) of sequences found in each step is given.

Statistics on sequences from CREM SSH library	Number of sequences
Sequenced clones	3400
Sequences that contain 1 *Rsa*I (GTAC) site were split	219
Sequences that contain 2 *Rsa*I sites were split 2x	11
Sequences that contain 3 *Rsa*I sites were split 3x	1
Primer was cut of at the end of sequences	643
Sequences with primer in the middle were split	22
Sequences that consist only of primer were thrown away	152
Resulting sequence fragments used for clustering	3197
Sequence clusters - *Rsa*I-fragments	956
Sequence clusters longer 30 bp	952

4.2.1.1 Processing of DNA sequences from the sequencer

The sequences represent parts of differentially expressed mRNAs. The sequencing was performed using the Sanger method and the sequences were run on an ABI 377 sequencing machine (Perkin Elmer Applied Biosystems). The ABI sequencing software (ABI Prism) was used to read the sequences from the gel image, assign the lanes for the individual sequences and assign the bases from the four color traces. The traces and gel images were also checked manually by visual inspection to correct errors during this procedure.

The resulting sequences contain a piece of the sequence from the vector at the beginning, followed by the sequence of the inserted *Rsa*I-fragment, optionally followed by another piece of vector sequence. Towards the end of the sequence read sequencing errors usually increase. A cutoff was set manually by visual inspection of the sequence traces. In some cases the sequence ends or the quality of the sequence decreases before the end of the *Rsa*I-fragment. In these cases the sequence of the beginning of the *Rsa*I-fragment is analyzed as far as it could be read. In average, 300-500 bases could be read by one sequencing run. Vector sequences as well as low quality sequence were removed. To automate the processing of the sequences some simple programs were written in the Perl programming language.

In the next step, all sequences were compared to vector sequences, that might have been artificially cloned, and for sequences of the PCR primers that were used in the SSH, that were frequently ligated to the *Rsa*I-fragments. Sequences of vector or PCR primers were removed, if they were near the end of the fragment. If the vector or PCR primer sequences or *Rsa*I restriction sites (GTAC) were found in the middle of the sequenced fragment, the fragments were split into several sequences before they are used in further analysis. It is likely that the presence of *Rsa*I restriction sites or PCR primer in the middle of the fragment is the result of an artificial ligation of several *Rsa*I-fragments which form

Table 4.2: **Clustering Statistics**. Assembly of clones from the CREM SSH library. The number of clusters that contain certain numbers of sequences are summarized.

Number of clusters	Number of sequences in cluster
559	1
138	2
40	3
39	4
35	5
85	6-11
35	12-17
14	18-23
10	26-57
1	61

concatemers. Statistics on the sequences and preprocessing steps are summarized in Table 4.1.

4.2.1.2 Clusters of sequences in the CREM SSH library

Most of the sequenced DNA fragments represent the same *Rsa*I-fragment and are identical except for sequencing errors. In order to reduce the redundancy of the sequence fragments the preprocessed sequences were clustered.

The preprocessed sequences resulting from the sequencing were assembled using the programs of the Staden package (Staden et al., 2000). During the assembly procedures overlapping sequences are detected and joined to a so called ”contig”. The Staden package software provides a simple assembly routine (Bonfield et al., 1995). Moreover, the Staden package offers a nice graphical interface. The sequences were assembled iteratively with decreasing stringency. Finally, contigs were searched for homologies and manually screened whether further joins between contigs were possible.

The assembly resulted in 956 contigs of overlapping or identical sequence that represent unique *Rsa*I-fragments in the CREM SSH library. The results of this clustering are shown in Table 4.2. Most of the clusters contain only one or two sequences. From the distribution of the picked clones we can conclude that in spite of the high number of clones that were picked it is unlikely that we have found representatives for all sequences that can be found in the subtractive cDNA library.

4.2.1.3 Finding homologies in databases of known sequences

The sequences resulting from the *Rsa*I-fragments were used to search against several databases of known sequences to determine known genes. Database search was performed using the BLAST (Basic local alignment search tool) program (Altschul et al., 1997). The EMBL nucleotide database,

Table 4.3: **Statistics on database homologies in the CREM SSH library.** For each category the number (#) of *Rsa*I-fragments that show homology to these database sequences is given as well as the total number of database sequences found. For the ESTs and novel sequences the "real" number of genes is unknown as the sequences do not reflect full length genes and several of these sequences could correspond to the same gene.

Type of homology found	# of homologous *Rsa*I-fragments	# of genes found
Known mouse genes	255	158
Homologous other genes	133	99
Highly homolog mouse ESTs	303	236
Homologous other ESTs	66	60
Novel sequences	199	199

the Swiss-Prot protein database and the EST consensus databases of mouse and man from GeneNest (described in Section 4.1.1) were searched. For each *Rsa*I-fragment the gene that shows the best homology was annotated. To determine the best homology several criteria were used. The gene should be similar over the complete length of the fragment with a high percentage of identity. Preferentially hits to known mouse genes were annotated. When none of these could be found, known genes from other organisms or ESTs were used. If several *Rsa*I-fragment sequences show homology to the same gene, the scores were combined for this gene. In cases where sequence fragments show homology to more then one gene, those genes are preferred that show homology to several fragments. Sequences of *Rsa*I-fragments that show only a weak homology to a gene as a result of bad sequence quality can often still be assigned to a gene with a lower stringency, if that gene is already found to be represented in the library. Therefore, the number of sequences that cannot be assigned to a gene because of poor sequencing quality is reduced. The database sequences found were assembled with the *Rsa*I-fragments using the Staden package (Staden et al., 2000).

The statistics of the database homologies found are shown in Table 4.3. The annotations fall into several categories: **Known mouse genes** for complete mouse cDNAs with a known function that are highly homologous with identity of over 90% for most of the fragment. **Homologous other genes** are related genes from other species or similar genes in mouse that show a similarity of more than 80%. **ESTs** are gene fragments or full length cDNAs of unknown function. The GeneNest database was searched for homologous mouse sequences with more than 90% identity, or for human sequences with more than 80% identity. The consensus sequences of the EST provide longer sequence information representing a larger part of the gene. Information about open reading frames or protein families is annotated for the EST consensus sequences.

Often several *Rsa*I-fragments are part of the same gene. A homologous gene to the human casein kinase is represented by eight *Rsa*I-fragments. Most genes, however, are only found by one *Rsa*I-fragment (426 database sequences) or two fragments (78 sequences). Further *Rsa*I-fragments might belong to the same genes but were not picked from the CREM SSH library.

Table 4.4: Functional classification of homologous genes found in the CREM SSH library.

Functional category of genes	# of genes found
Signal Transduction	50
Metabolic Enzymes	34
Protein Degradation	19
Transcription Factor	18
Sperm Structure	14
Cytoskeleton + Motility	14
Translation	14
Intracellular Transport	13
Membrane Transport	7
Molecular Chaperone	7
Signal Transmission	7
Protein Modification	5
Histones + HMGs	4
Cell Cycle Regulator	4
RNA Modification	3
Cell Junction	2
DNA Repair	2
Axon Guidance	2
Mitosis	1
Meiosis	1
Energy Metabolism	1
Energy Transduction	1
Erythrocyte Membrane	1
Protein Transport	1
Viral Protein	1
not classified / unknown	31

A list of all the known genes found with homologies to the CREM SSH library is shown in Appendix A. The complete list of all the ESTs homologous to the CREM SSH library and the list of all completely novel sequences is presented on the web site[2] (Section 4.2.4).

4.2.1.4 Ontology of the genes

There are 257 homologous known genes found in the CREM SSH library. To be able to make better use of this list we functionally classified these genes (Table 4.4). The functional categories were derived from database information or from literature. The functional annotations of the corresponding entries to the genes in the Swiss-Prot database were used. For genes where the Swiss-Prot annotations were not sufficient or no Swiss-Prot entry was found, keyword search in the PubMed literature

[2] http://www.dkfz.de/tbi/crem

```
?Clone  Comment Homology UNIQUE ?LongText
        Positive  Positive_probe ?Probe XREF Hybridizes_to ?Grid
                  Pos_probe_strong ?Probe XREF Hybridizes_strong ?Grid
                  Pos_probe_weak ?Probe XREF Hybridizes_weak ?Grid
        Sequence    ?Sequence XREF Clone
        RSAfragment ?RSAfragment XREF Clones
        Length  Seq_length UNIQUE Int // bp
        Gridded ?Grid
        Probed ?Probe

?Probe  Source  Clone UNIQUE ?Clone XREF Probed
                Sequence UNIQUE ?Sequence XREF Probed
                Oligo UNIQUE ?Oligo XREF Probed
        Length UNIQUE Int
        Positive  Hybridizes_to ?Clone XREF Positive_probe ?Grid
                  Hybridizes_strong ?Clone XREF Pos_probe_strong ?Grid
                  Hybridizes_weak ?Clone XREF Pos_probe_weak ?Grid
        Grid      ?Grid XREF Hybridization_Results

?RSAfragment Information Clone_Number UNIQUE Int
                         Clones ?Clone XREF RSAfragment
             Sequence ?Sequence XREF RSAfragment
             Database_Info ?Database_Info
             Seq_length UNIQUE Int
```

Figure 4.8: Example classes from the ACEDB database model to store the information about the clones from the CREM SSH library, the results of hybridizations with selected probes, and the information about the *Rsa*I-fragments.

database was performed.

4.2.2 Integration of sequencing and hybridization data in a database

The results of the sequencing and the data of the hybridizations that were used for redundancy reduction were combined in a special database system to have an overview over the clones from the CREM SSH library and to determine the state in the analysis of each clone during the work.

The ACEDB (*A Caenorhabditis elegans database*) software, which was originally designed for genome sequencing projects (Eeckman and Durbin, 1995) in C. elegans, yeast, arabidopsis, several human chromosomes and *Xenopus laevis* (Pollet et al., 2000), was used. For this project some specialized models for storing the data were designed. Structures for storing the information about the clones in the library, the hybridization probes and the unique *Rsa*I-fragments are necessary. The definitions how these structures are stored are given in the ACEDB specific format in Figure 4.8.

Further structures to store database homologies, sequence information and literature were used. The complete definitions of the database models that were used as well as a text version of the complete database are available in the world wide web[3]. The database with the information about the CREM SSH library is called ”ACREMSLDB”. An overview of the user interface is shown in Figure 4.9.

[3] http://www.dkfz.de/tbi/crem

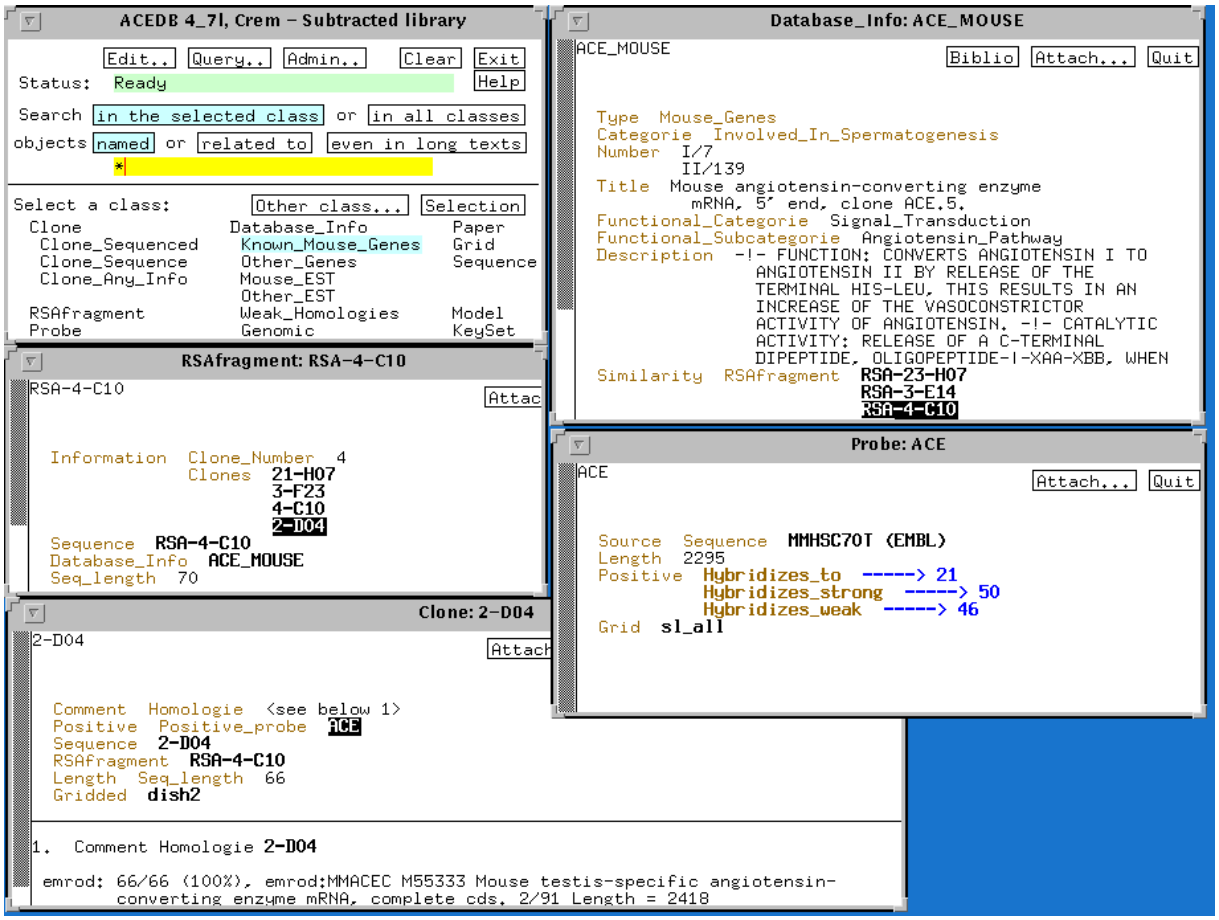


Figure 4.9: User interface to the ACREMSLDB database. Information about the genes, the *Rsa*I-fragments, the clones and the hybridization results is cross-linked. Some information about the gene “ACE” is displayed here.

4.2.3 Expression analysis of CREM-dependent sequences

Expression analysis on DNA microarrays was used to confirm the differential expression of sequences found in the CREM SSH library, to find additional sequences expressed dependent on CREM in mouse testis, and to study the pattern of expression of CREM-dependent genes during the course of spermatogenesis. Two different techniques of DNA arrays were applied: nylon cDNA arrays containing selected sequences from the CREM SSH library and commercially available oligonucleotide arrays from Affymetrix containing representative oligonucleotides for 10,000 mouse sequences.

4.2.3.1 Comparison of wild-type versus CREM (-/-) testes

Measuring the expression levels on nylon microarrays containing sequences of the CREM SSH library

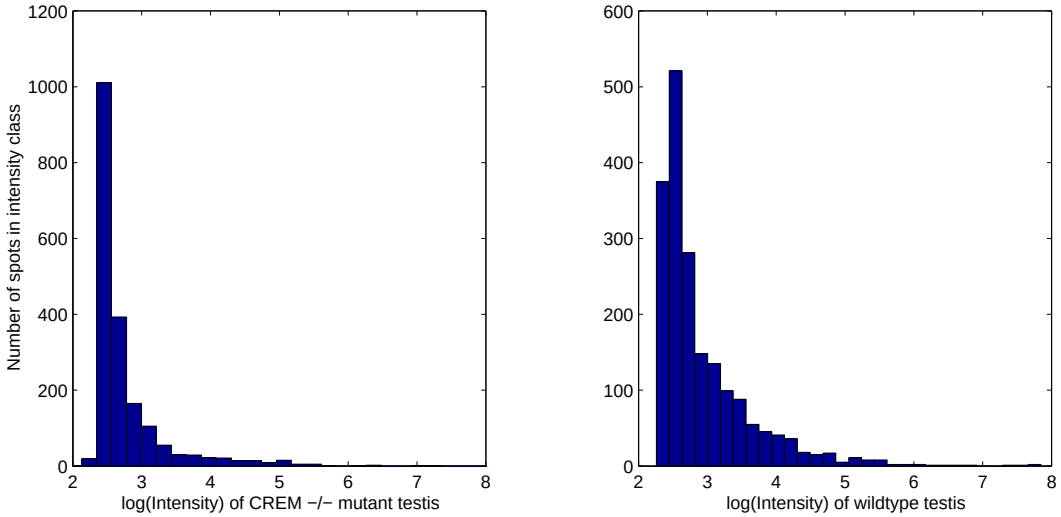


Figure 4.10: Histograms of the distribution of measured intensity levels for two hybridization experiments on wild-type testes and CREM (-/-) mutant testes.

The expression levels of the genes found after SSH were measured by nylon cDNA arrays. Wild-type mouse testes cDNA as well as CREM (-/-) testes cDNA were hybridized to the microarray containing CREM SSH library sequences.

The histograms in Figure 4.10 show the distribution of all measured expression values in wild-type and in CREM (-/-) testes. The logarithmic intensity is plotted against the count of sequences in each intensity class. Most clones show only a hybridization signal close to background. As the number of spots is comparatively low compared to whole genome arrays, the normal distribution of background values could not be estimated as described in Section 4.1.2, but a cutoff value was set manually at an absolute intensity of 15 by looking at the distribution of empty control spots. The histogram of the intensities in wild-type testes shows a higher number of expressed genes than the histogram of intensities in CREM (-/-) testes.

Figure 4.11 shows a comparison of the expression levels measured in wild-type testes and CREM (-/-) testes. Most of the points show a higher expression in wild-type mice in comparison to the CREM deficient mice. This means that most sequences show differential expression, as was expected, as the measured sequences are resulting from SSH. Almost half of the spots show only intensities at a background level in all measurements. For the lowly expressed sequences it is not meaningful to estimate a factor of differential expression. Of the sequences which show significant intensities most sequences are down-regulated more than 3-fold in CREM-deficient testes compared to wild-type testes (319 of 452). No sequence is up-regulated in CREM-deficient mice.

Measuring the expression levels on Affymetrix oligonucleotide microarrays

To find additional genes expressed in a CREM dependent manner the expression of a large number of sequences was measured with a different technique on Affymetrix oligonucleotide arrays. The

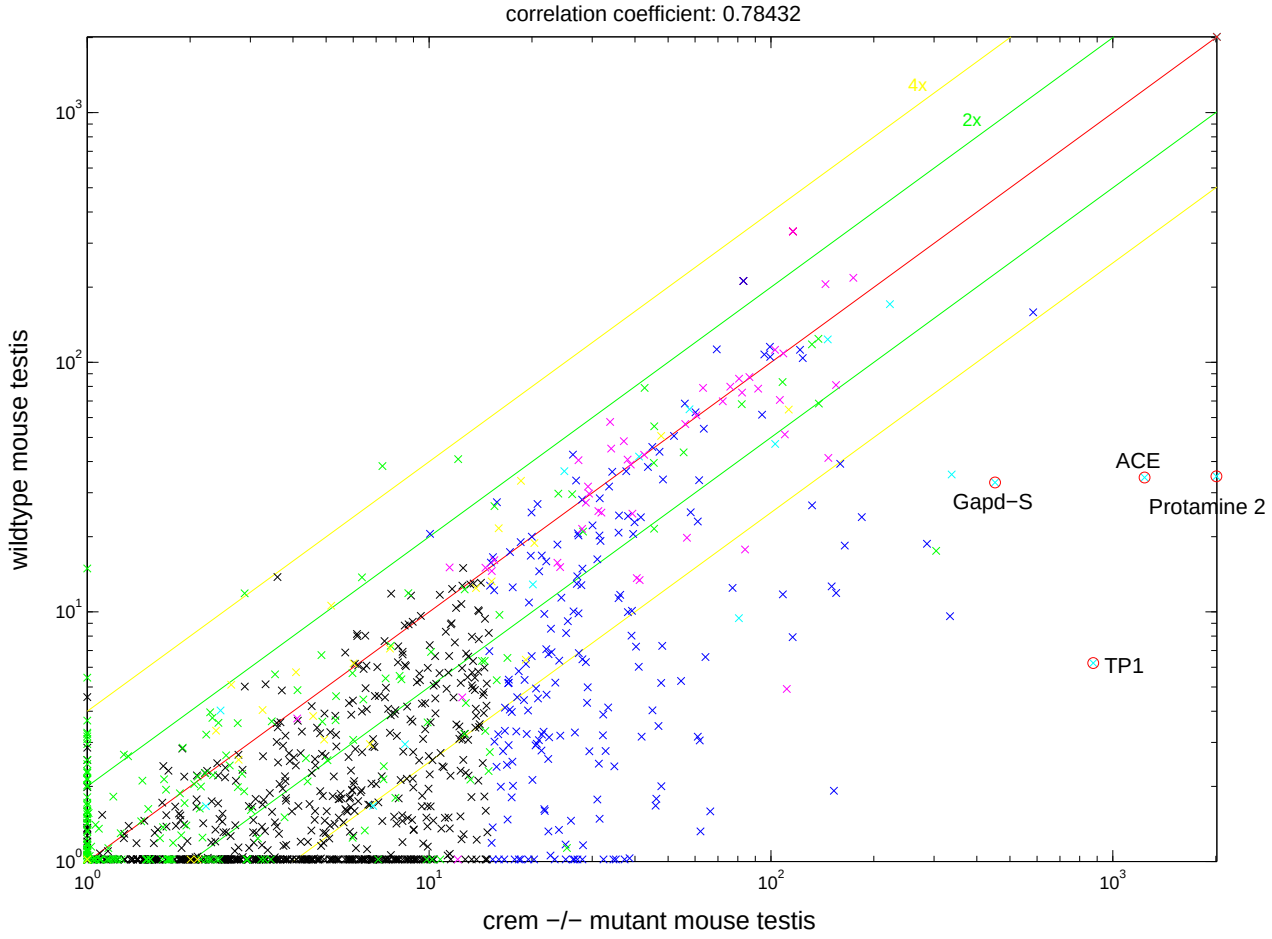


Figure 4.11: **Comparison of expression levels in wild-type testes versus CREM (-/-) mutant testes.** Scatter plot of all measured expression intensities. The values represent a median of three reproductions of experiments. Measurements for positions with empty spots or heterologous DNA spotted are indicated by green crosses, house-keeping genes used to standardize the median of differences to the centerline of the plot are colored in magenta, spots representing genes with known differential expression are cyan colored. The name for a few genes with known function in spermatogenesis is given besides the corresponding points. Among them are the known CREM target genes ACE, TP1 and Protamine and the testis specific isoform of glyceraldehyde 3-phosphatedehydrogenase (Gapd-S). Black crosses represent spots that have an expression level below 15 in both conditions, i.e. are expressed close to background level. The red line indicates genes expressed at the same level in wild-type and mutant, the green line indicates a factor of difference of 2-fold, the yellow line a factor 3-fold.

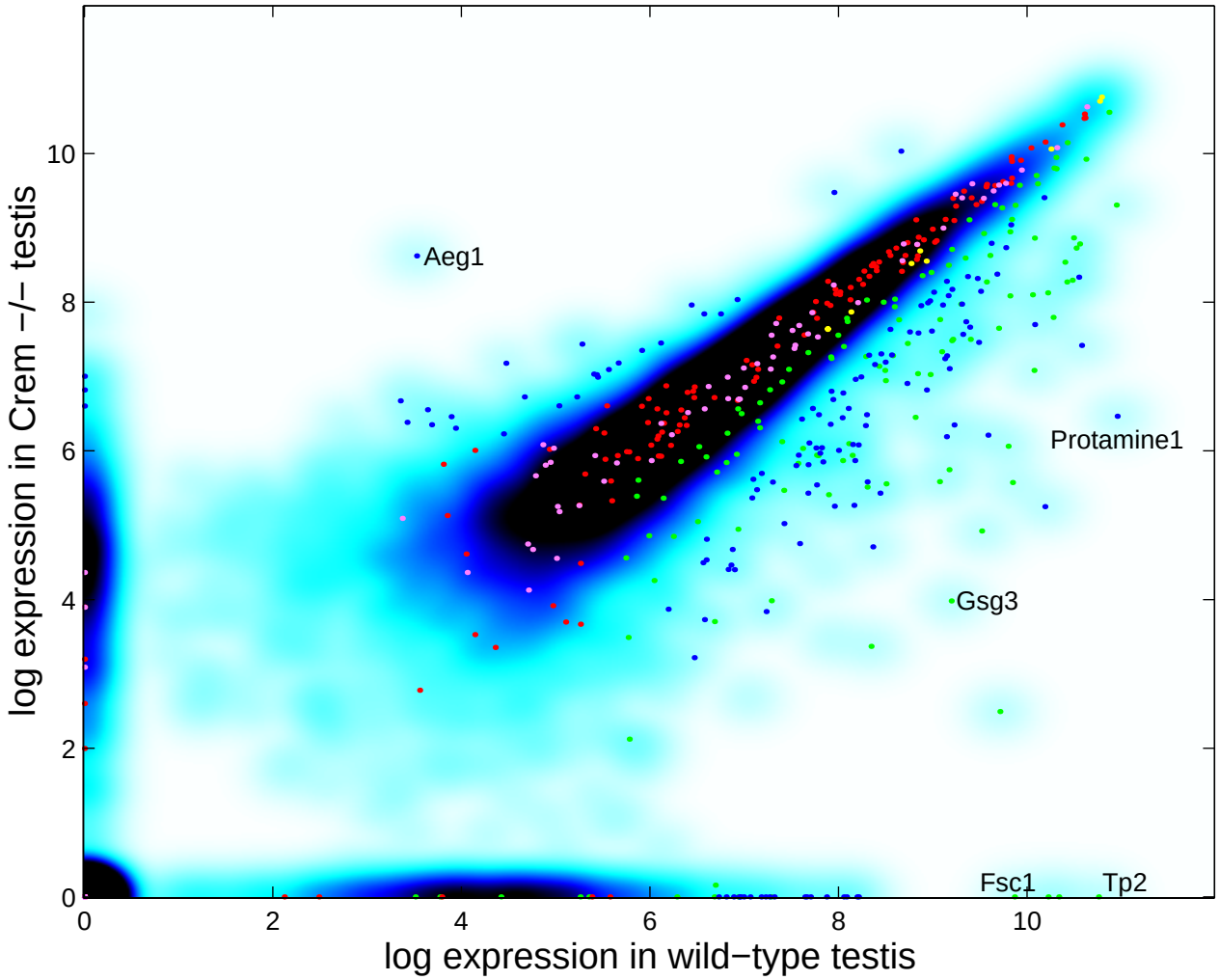


Figure 4.12: Comparison of the median expression levels in CREM (-/-) testes and wild-type testes measured on an Affymetrix array containing 10,000 mouse sequences. The densities of the points in a scatter plot are visualized. Additionally selected individual points are drawn: Green Dots represent sequences of genes found to be differential in the CREM SSH library as well as on the Affymetrix array. Dots with blue color represent genes that are found to be differentially expressed on the Affymetrix array but which are not in the CREM SSH library, provided that they are in the regions of low density in the plot, their expression level is higher than 400, and the difference of expression is more than a factor of 2. Red dots represent sequences on the Affymetrix array that are homologous to sequences in the CREM SSH library, which appear to be equally expressed or have an insecure expression measurement on the Affymetrix array as well as on nylon arrays constructed from clones of the CREM SSH library. Pink dots represent sequences which were found in the CREM SSH library but show non differential expression on the Affymetrix arrays as well as on the nylon arrays. The yellow dots represent “spiked” control sequences on the Affymetrix array with equal expression.

Table 4.5: Comparison of the hybridization results of wild-type versus CREM (-/-) testes cDNA on Affymetrix arrays containing 10,000 mouse sequences and nylon microarrays containing selected clones of the CREM SSH library. The qualitative results of differential expression for those sequences in the SSH library that show homology to sequences spotted on the Affymetrix array are compared. The numbers where the results of the Affymetrix and nylon array hybridizations are in agreement are colored green. Where Affymetrix and nylon array hybridizations show opposite results the numbers are colored in red. By both techniques many of the investigated sequences show no reproducible result or are expressed close to background level, and therefore no clear statement on differential expression can be made.

Measurement on	**Affymetrix mouse array**		
nylon SSH library array	*down-regulated in mutant*	*no measurement or lowly expressed*	*not differential*
down-regulated in mutant	84	25	14
no measurement or lowly expressed	20	48	43
not differential	1	4	8

expression levels of mRNAs expressed in CREM (-/-) testes and in wild-type testes were measured each by 3 independent hybridizations to an Affymetrix array of approximately 10,000 mouse genes (U74A). For each replicate the mRNAs of the testes of 3-4 mice were pooled.

As a result of the Affymetrix hybridizations, 102 of the Affymetrix sequences were found to be significantly and reproducibly down-regulated in CREM-deficient testes in comparison to wild-type testes. Results were counted only when they exhibited a factor of differential expression of more than 2-fold and with an expression intensity of more than 400 in at least one of the experiments. Furthermore, genes were excluded where results were irreproducible, and those which were called “absent” by the Affymetrix software in all measurements. 61 of these sequences correspond to known genes and 41 sequences to ESTs. 60 of these genes or EST sequences were also found in the CREM SSH library. Furthermore, 24 sequences were found to be expressed stronger in CREM (-/-) mutant testes than in wild-type testes. 18 of these sequences correspond to known genes and 8 to ESTs.

Figure 4.12 shows a comparison of the median expression levels in CREM (-/-) testes and wild-type testes. The density of points in a scatter plot of the median logarithmic hybridization intensities is converted into a smoothed map and shown as a false color representation. Individual points in the plot represent measurements in the array that correspond to several groups of interesting genes:

1. Spots with differential expression measured with Affymetrix chips where the corresponding genes were also found by SSH. (green)

2. Genes found by SSH with marginally differential expression measured with Affymetrix chips. (green)

3. Genes which appear on the Affymetrix chip to be expressed stronger in wild-type in comparison to CREM-deficient testes with a factor of more than 2-fold. (blue)

4. Genes which appear on the Affymetrix chip to be expressed in CREM (-/-) mutant testes and up-regulated in comparison to wild-type testes with a factor of more than 2-fold. (blue)

5. Genes found by SSH with differential expression indicated on nylon CREM SSH arrays and non-differential expression indicated on Affymetrix chips. (red)

6. Genes found by SSH with unclear result on Affymetrix chips and on nylon CREM SSH arrays as well. (red)

7. Genes found by SSH with non-differential expression. (pink)

Table 4.5 summarizes the comparison of the results of the Affymetrix oligonucleotide array hybridization with the results of the SSH and the nylon cDNA array hybridizations. 246 of the genes or ESTs found in the CREM SSH library correspond to 342 measurements for genes on the Affymetrix array. Most of the genes for which the results could be compared show agreeing results. For many of the sequences there is no reproducible measurement in one or the other technology.

The complete list of the summarized results of the Affymetrix array hybridizations is available as Appendix B.

4.2.3.2 Expression profiles of sequences found in the CREM SSH library in testes of prepubertal mice

To perform a more detailed analysis of the expression patterns of the genes found in the CREM SSH library, we analyzed the expression profiles of these sequences in young mice. cDNA prepared from testes of young mice between 9 and 27 days old were used for hybridization on the high-density filters containing the CREM SSH sequences. In these mice the first round of spermatogenesis is synchronized and spermatids at specific stages can be found at distinct time-points after birth (see Section 2.2).

The expression profile experiments result in a table, where each row represent a gene and each column represents an experimental condition for a specific stage of mouse spermatogenesis. The cells contain the logarithmic expression intensities, which is a median over several hybridization experiments for the same condition. The intensities for each hybridization were standardized according to values of house-keeping genes as described in Section 4.1.2.

In Figure 4.13 the expression profiles of 46 house-keeping genes are shown. After adjusting the intensity levels the house-keeping genes stay at constant level over the time-course.

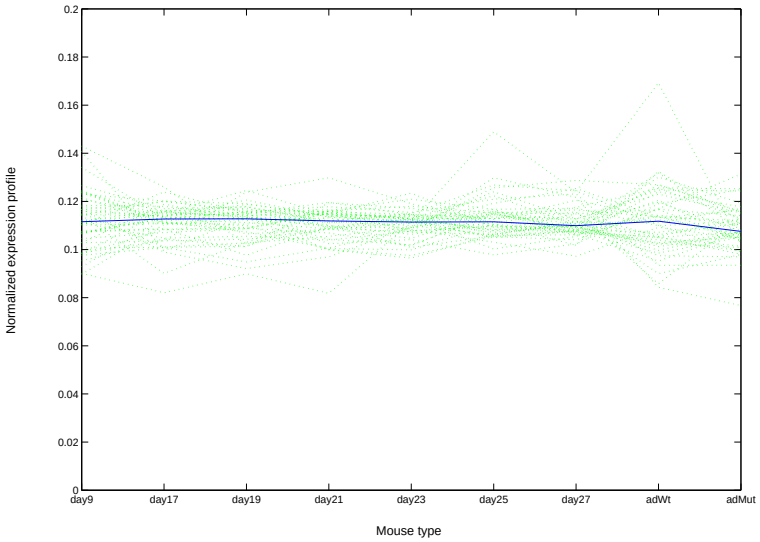


Figure 4.13: Expression profiles of 46 house-keeping genes. The measured intensity values have been standardized so that the median of the house-keeping genes is constant. The values shown here are divided by the sum of the profile, so that their sum equals one. The individual house-keeping genes are drawn in green, the median line is drawn in blue.

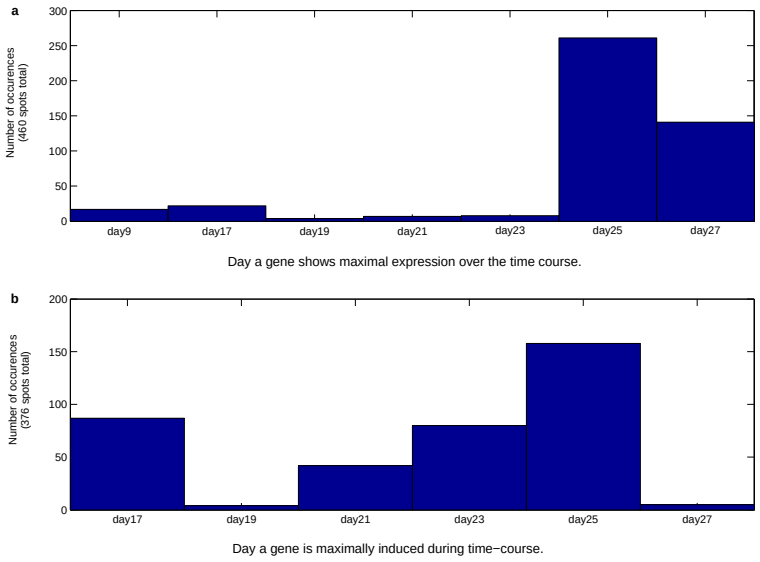


Figure 4.14: Statistics on expression profiles. **a** shows a histogram of the count of genes which have their maximal expression levels in the time-course in prepubertal mice for each day in the time-course. Only genes are counted that have a significant expression level (i.e. absolute intensity > 15) in at least one measurement (total 460 spots). **b** shows a histogram of the count of genes which are maximally induced on a specific day of the time-course. Only those genes are regarded, whose expression is induced during the time-course by more than 3-fold (total 376 spots).

Measuring the expression profiles of the sequences in the CREM SSH library, we found that most genes are down-regulated in CREM (-/-) mice and start to be expressed after meiosis at day 21 or later. From published results it is expected, that on day 21 the CREMτ protein starts to be expressed as well. Meanwhile the mRNA of CREMτ is induced after day 9 and constantly expressed over the time-course in our measurements.

The results of the expression profiles can be classified: 530 of 994 sequences show an expression level of more than 15 in at least one condition. For the other sequences no clear statement about differential expression can be made. During the analysis of the time-course in young mice the expression levels of the cDNAs of most genes appear to be up-regulated in testes (376 of 460). Almost all sequences show a maximum of their expression level at day 25 or day 27 (Figure 4.14a). The curves of the expression profiles show the maximal induction of genes between day 21 and day 25 and the profiles stay at a constant level at day 27. A smaller portion of the genes are up-regulated between day 9 and day 17 (Figure 4.14b).

The expression profiles of 227 sequences, that show a positive expression under at least one experimental condition and that correspond to known genes, are visualized as a color coded table in Figure 4.15a. The rows of the table are sorted in a way that the sum of the pairwise distances is minimized. This means that sequences with similar expression profiles should be close to each other in the linear order. This corresponds to the solution of a traveling salesman problem (see Section 4.1.2.4).

A gene that is up-regulated on day 21 from which the expression is well characterized in literature is the outer dense fiber protein of sperm tails (Odf 1) (Carrera et al., 1994; Chen et al., 1997a). The expression of Odf 1 gradually increases between day 21 and day 25. Odf 2, the diazepam binding inhibitor-like protein, a phosphoglucomutase-like gene, and hexokinase show very similar expression profiles (Figure 4.15g).

Many sequences in the CREM SSH library start to be expressed at day 23 and reach the maximum of expression at day 25. The known direct CREMτ targets ACE and TP 1 (Zhou et al., 1996; Kessler et al., 1998) show corresponding expression profiles (Figure 4.15f). Sequences expressed similarly to CREM target genes are of particular interest as their co-expression might also be an indication for similar regulation. Therefore, sequences were sorted based on the distance calculated from a mean profile of ACE and TP 1. The result is shown in Figure 4.16, the closest 27 genes are shown. Protamine 2, long-chain fatty acyl-CoA synthetase, a number of other genes, EST clones and yet uncharacterized sequences cloned in the SSH are among them.

The latest genes are up-regulated on day 25. Among these is the Krox20-gene (Chavrier et al., 1989). Most measured sequences reach a maximal expression at day 25 and stay at the same level at day 27 as well (Figure 4.15e). The period from day 21 to day 25 may be referred as a period of mRNA collection in spermiogenesis.

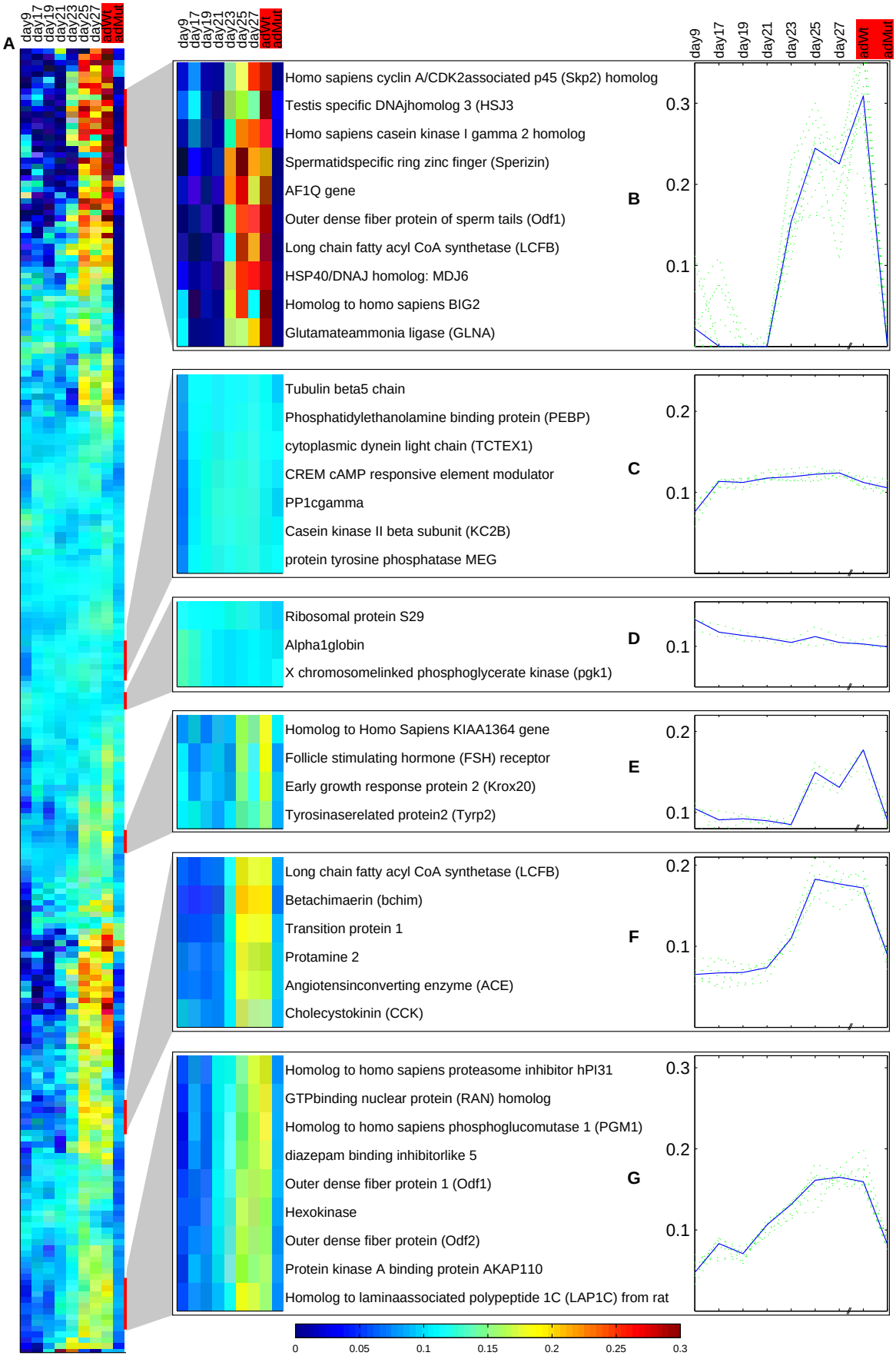


Figure 4.15: Expression profiles of 227 sequences corresponding to known genes. The logarithmic expression levels were divided by the row-sum and are shown in false color representation. **a** 227 profiles; **b**, **c**, **d**, **e**, **f**, **g** enlarged regions with interesting genes as well as diagrams with the median profiles of these genes.

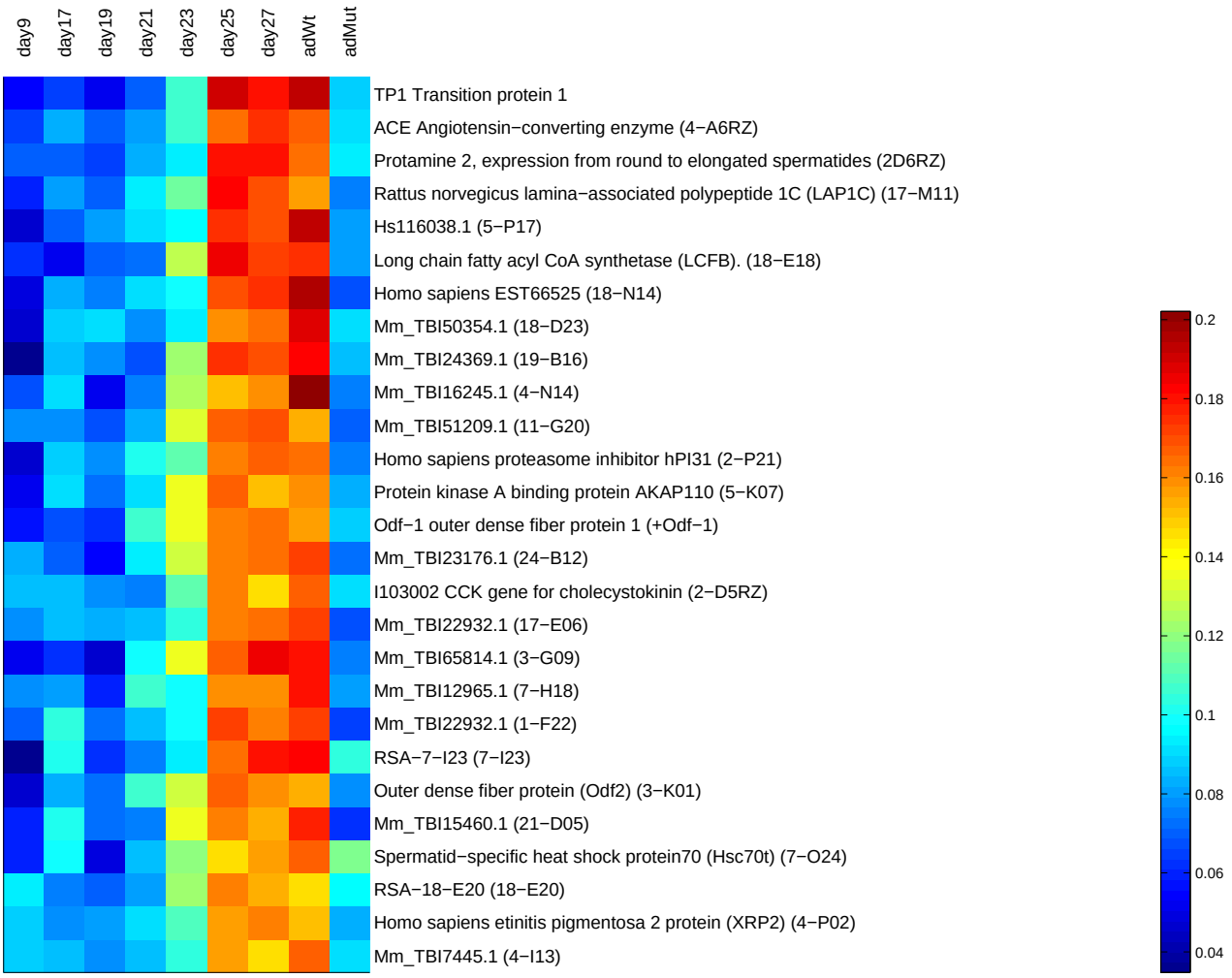


Figure 4.16: CREM target like expression profile. The mean of the expression profiles of an ACE and a TP1 cDNA have been used to find similar expression profiles. The displayed genes are arranged from top to bottom with increasing distances calculated.

Some of the sequences cloned in the CREM SSH library do not show a differential expression between the CREM (-/-) mutant testes and wild-type testes in the hybridization experiments. A group of these genes is expressed at the same level at all stages of spermatogenesis. Another group of genes is not expressed at early, spermatogonia stage (day 9) of spermatogenesis and expressed at constant level at later stages (days from 17 to 27). This type of expression is shown by Tctex1, tubulin and protein phosphatase 1γ (Figure 4.15c). A third group of genes is highly expressed at early spermatogenesis and down-regulated at later stages. The phosphoglycerate kinase 1, the ribosomal protein S29 and the sulfated glycoprotein 2 show this kind of expression (Figure 4.15d).

The table shown here represents only some of the measured expression profile data for the sequences from the CREM SSH library. In Appendix C a larger dataset including 492 sequences that show a significant expression under at least one experimental condition together with the homology information is shown.

4.2.4 Availability of results via a web based interface

It is not possible to present the complete data set used for the analysis in this document. Therefore, they are presented in a searchable manner via the world wide web. An advantage of the web presentation is, that the data can be easily linked to other databases. The sequences and the hybridization results as well as the assembly information and the complete list of found database homologies is presented these supplementary web pages.[4]

The information that is presented on the web-site is shortly discussed below:

1. **Information about the CREM SSH library** The complete datasets of the CREM SSH library and the analysis of the sequences is displayed.

   - **Clones in SSH Library.** All clones that have been picked from the CREM SSH library are listed. The clone names are derived from the number of the plate and plate position where they are stored. The clones are presented in two lists. *Sequenced clones* lists all 3400 clones with sequence information. For those clones the sequence and the information about the cluster the sequence is in is displayed. 128 sequences were used to hybridize against all clones of the SSH library. *Clones analyzed by hybridization to selected probes* are summarized in tables. Clones that hybridize with more than one probe are indicated.
   - ***Rsa*I-Fragments (clusters of sequenced clones) found.** The information about the redundancy in the library and the resulting unique sequence clusters is summarized in the table of *Rsa*I fragments. The consensus sequences as well as the information which clones belong to the cluster are shown. The *Rsa*I fragment have been analyzed for homologies in several databases. The best hits are annotated.
   - **Homologies to known Genes or EST.** The full lists which genes and EST sequences were found in the library can be displayed. The list of the known genes found in the CREM SSH library is also shown in Appendix A. In the web presentation also a list of homologous ESTs and the list of completely novel sequences is provided. The gene names are linked to the corresponding Swiss-Prot entries or to the GeneNest visualization. A short title and some literature references are displayed. Several database sequences can correspond to a gene entry. Also many genes are represented by several *Rsa*I-fragments.

2. **Use BLAST to check for the occurrence of sequences in the CREM SSH library.** The occurrence of a sequence or homologies to other sequences can be checked with BLAST. A blast-server has been set up with a web-based interface to search against the CREM SSH library.

[4] http://www.dkfz.de/tbi/crem: To access the site please contact G.Schuetz@dkfz.de

3. **Expression analysis of sequences in the CREM SSH library.** The results of the analysis of expression levels of the sequences from the CREM SSH library and information about the co-expression of genes is summarized.

   - **Scatter Plot of differences between wild-type testes and CREM (-/-) mutant testes.** In the comparison of expression levels of genes in wild-type mice testes and testes of mutant mice differentially expressed genes can be visualized. In the scatter plot it can be seen that most genes picked in the CREM SSH library show differential expression. The genes are visualized in the web based interface of a software for scatter plot based analysis in an interactive way.
   - **Expression profiles in testes of young mice** The expression levels of genes in the CREM SSH library were measured during the development of young mice. The information about the expression profiles is summarized in several tables. In each table the genes were arranged by similarity of the profiles in a linear order. The list of *all clones* extends in a printed version over several hundred pages. It contains all controls except empty spots, as well as all measurements of clones which are expressed close to background level under all measured conditions. To make the data more clear and to focus on the important information several smaller sublists are provided. The list of *clones expressed under at least one condition* gives a subset with all sequences that provide useful expression information. This table is also presented in Appendix B. The list of *clones with homology to known genes* is shorter and focuses on correlations between the known genes. This table is the same as presented in Figure 4.15 but with all the gene names and functional categories annotated. The expression profiles of known genes were further sorted by the functional category of the gene. This table allows to directly compare the profiles of functionally related genes or different *Rsa*I-fragments of the same gene. This is useful to detect co-regulations of complete functional networks.

4. **Expression analysis on Affymetrix arrays with 10,000 mouse genes.** The results of the analysis of expression levels of 10,000 mouse sequences in wild-type testes and CREM (-/-) mutant testes measured on Affymetrix oligonucleotide arrays.

   - **Scatter Plot of differences between wild-type testes and CREM (-/-) mutant testes.** In the comparison of expression levels of genes in wild-type mice testes and testes of mutant mice differentially expressed genes can be visualized in an interactive way.
   - **Differences measured in the expression levels between wild-type testes and CREM (-/-) mutant testes**. Summary of all measured expression levels of genes found to be differentially expressed with a factor of more than 2-fold and expressed in at least one measurement with an intensity of more than 400. The numbers of all measurements are

given as well as information about the corresponding gene and the spotted sequence. The genes are annotated with a functional description from the Swiss-Prot database and a gene ontology present in the Mouse-Genome-Database (MGD).

- **Comparison of measured differences of CREM SSH sequences measured on Affymetrix and Nylon arrays**. All measured expression levels are shown for sequences that show homology to sequences found in the CREM SSH library.
- **Summary of Affymetrix results**. Information about the genes found to be differentially expressed between wild-type testes and CREM (-/-) mutant testes is summarized as well as the comparison of the results for the sequences that were found in the CREM SSH library that are also present on the Affymetrix arrays.

5. **Developments of methods and programs**. Links to description of the methods and source code or web-interfaces of software that was developed are linked here.

   - **Access to clustered EST databases used for data analysis**. The data of the GeneNest databases can be accessed, visualized and searched via a web-based interface.
   - **ACEDB database and database models used to organize the data of the CREM SSH library**. The database models used to define a database of the CREM SSH data are provided as well as a text-dump of the complete database. This can be used to install the database locally.
   - **Web interface for standardization procedures on expression data**. Links to a web interface, where the standardization procedures described in this thesis can be applied. Pairs of experiments can be compared and the comparison visualized as an interactive scatter plot.
   - **Algorithm to estimate the optimal linear order of gene expression profiles**. Source code of an algorithm used to sort a set of gene expression profiles into an optimal linear order by solving a *traveling salesman problem*. The algorithm is programmed in the C programming language and can be compiled in Matlab using the mex function.
   - **Selection method and criteria to chose optimal clones**. To chose optimal clones from the CREM SSH library for array construction several parameters were combined using the FuzzyTech program. The file containing the exact definitions of criteria as well as the generated C program for clone selection are made available.

5 Discussion

5.1 Analysis of CREM dependent genes

956 different sequences were found using SSH. They represent genes likely to be expressed CREM dependently and probably have functions in the late stages of sperm development. This dependency may be of different nature: the cloned genes may be either direct CREMτ target genes or the expression of these genes may not be detectable in CREM deficient mice due to secondary events caused by the absence of CREM. Particularly, genes which are specifically expressed at the late stages of spermatogenesis might be found in the CREM SSH library, as cells at these stages are undergoing apoptosis and subsequently are lost in the testes of CREM (-/-) mice.

Expression measurements by DNA microarrays mostly confirm the result obtained by SSH. As expected, most of the cloned genes appear to be differentially expressed in the comparison of expression levels in wild-type testes and in CREM (-/-) testes. However, some genes were cloned that do not show a differential expression, many of which are very abundant in testis. In these cases it is likely that the normalization procedure of the SSH was not sufficient to subtract these transcripts. The time-course expression profiling study also revealed that most of the identified genes are expressed post-meiotically. All of the mRNAs that were down-regulated in the CREM (-/-) mutant mice showed specific expression at the late stages of spermatogenesis.

We compared the SSH results with screening of a set of 10,000 mouse genes on Affymetrix oligonucleotide arrays. This adds a further verification with a different technique for the differential expression of many genes we have already found in the CREM SSH library, as well as indication for more genes down-regulated and also genes up-regulated in the CREM (-/-) mutant. The pro-apoptotic genes CPP32 apoptotic protease and the apoptotic signal-regulating kinase were found to be up-regulated in CREM (-/-) mutant mice. The anti-apoptotic transcripts of Bcl2a1d and Bazf are expressed in wild-type, but are absent in CREM deficient mice.

The expression profiles of most of the genes found in the CREM SSH library show some common features. Most of these genes are not expressed until day 19. After this day the mouse testis yields

round spermatids, which develop to elongating spermatids and mature sperm. CREM protein expression is expected to begin at about day 21 (Foulkes et al., 1992). On day 21 and day 23 only few genes are up-regulated. Most genes are maximally expressed at day 25. The amounts of most mRNAs remain the same on day 27 as on day 25. From the point of view of gene expression, the round spermatid stage may be called the stage of collection of mRNAs necessary for sperm development.

Many of the genes that encode proteins belonging to different sperm structures and responsible for different functions in sperm were found to be expressed post-meiotically and down-regulated in CREM (-/-) mice.

Expression profiles for several sperm structure proteins are shown in Figure 5.1a. Outer dense fiber proteins 1 and 2 (Odf1 and 2), fiber sheath component 1 (Fsc1 or AKAP82) are components of the sperm tail. The protein calicin is a component of sperm head structure calix.

Several proteins are involved in DNA compaction during sperm maturation. For instance, RCC1 (Regulator of Chromatin Condensation) may play a role in the entry of round spermatides into the chromatin condensation phase. Another probable regulator of DNA compaction is the protein kinase SRPK2, which is able to phosphorylate protamine 1 and by this initiate the substitution of histones (Kuroyanagi et al., 1998). TP1, TP2 and protamine 2 substitute histones in sperm.

A large group of genes found in the CREM SSH library encode signal transduction proteins. The genes for AKAP110, AKAP82 (Fsc1), D-AKAP1 and protein kinase A regulatory subunit II participate in cAMP mediated signal transduction (Banky et al., 1998) (Figure 5.1b). AKAPs (A-Kinase Anchoring Proteins) are responsible for compartmentalization of cAMP mediated signaling by anchoring protein kinase A to specific organelles. These genes are expressed at the late stages of spermatogenesis and are not found in the CREM deficient mice. The proteins are localized in different sperm organelles and may participate in processes of sperm capacitation and sperm-egg fusion (Visconti et al., 1997).

The genes encoding metabolic enzymes represent the largest functional group in the CREM SSH library. There are enzymes involved in purine, glycerolipid, glutamate metabolism and different other metabolic pathways. The genes encoding enzymes performing all steps from glucose to 1,3-diphosphoglycerate during glycolysis are found to be expressed post-meiotically and are down-regulated in CREM (-/-) mice (Figure 5.1c). These enzymes are important to produce the energy in sperm that is necessary for tail movement.

Several groups of co-regulated genes belonging to the same functional categories were found. However, some clones corresponding to different *Rsa*I-fragments of the same gene show differences in the profiles. This can be explained by the fact that many of the sequences show only weak hybridization signals, possibly due to different length of the sequence. These profile shapes are more difficult to determine accurately and they display steep induction, if the gene moves from an expression level

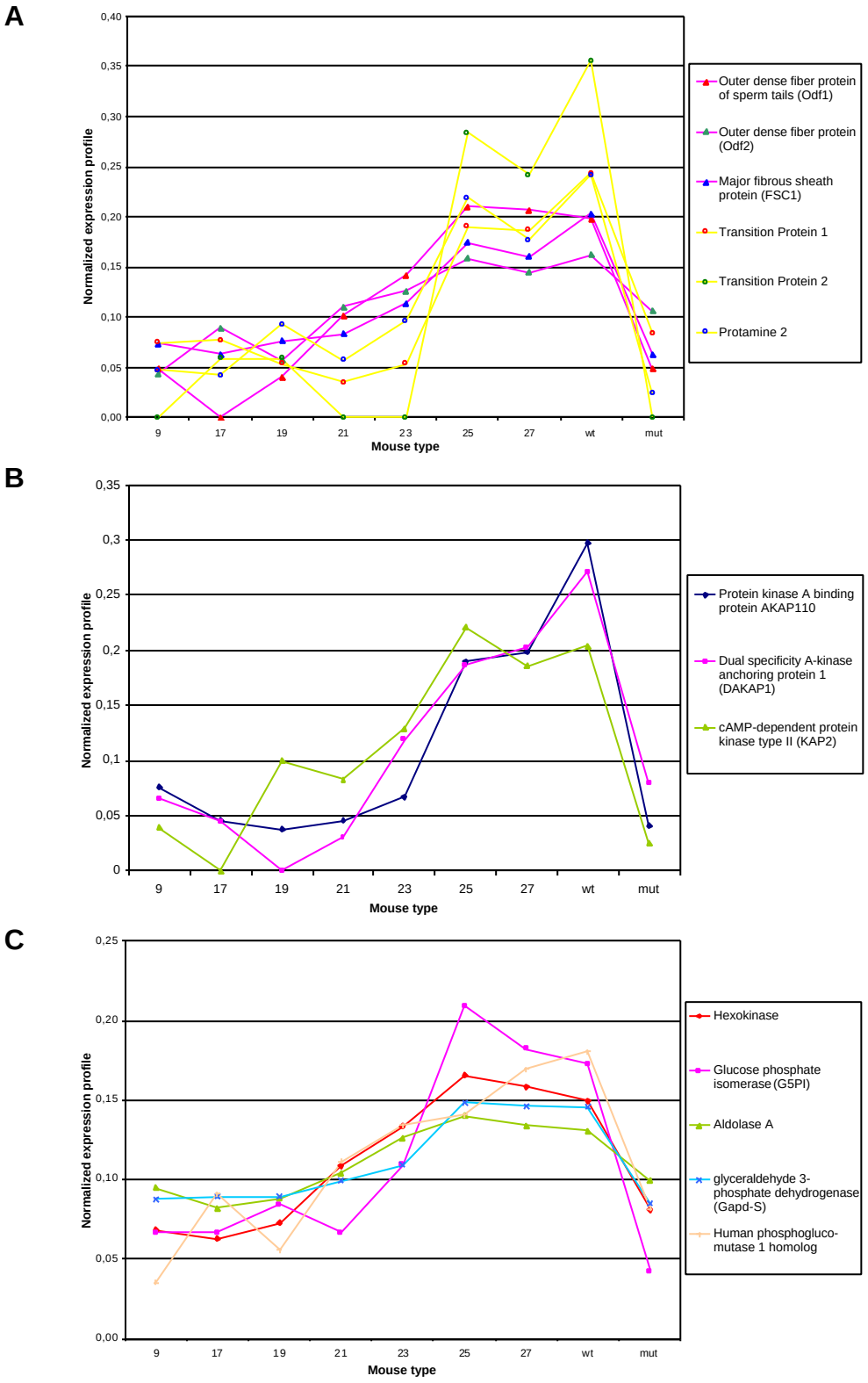


Figure 5.1: Expression profiles of functionally related groups. **a** proteins important for sperm structure. **b** genes important in cAMP mediated signal transduction. **c** enzymes involved in different steps during glycolysis.

close to background level to a more accurately detectable expression level. However, qualitatively their results are usually in agreement with the results of other *Rsa*I-fragments of the same gene.

For most genes with known post-meiotic expression profiles, for example proacrosin, protamine, Tp-1, Krox-20 and the outer dense fiber protein (RT7) (Blendy et al., 1996; Nantel et al., 1996), we could nicely confirm previous studies about their expression. In addition, for many known genes we found novel information about the post-meiotic expression. Some of these genes encode different components of sperm structures or may be involved in different processes from sperm maturation to the sperm capacitation and, finally, to the acrosome reaction. It is likely that many of the genes found to be co-expressed in the differentiation of spermatids are functionally related. To characterize complete functional networks responsible for these specific sperm structures more detailed studies will be necessary.

All this data is qualitatively in agreement with northern blot hybridization data published previously (Blendy et al., 1996). The reliability of this large scale analysis appears to be high. Only one clone shows a result conflicting with published data. STAT4 is shown in literature to be expressed in late sperm stages only (Herrada and Wolgemuth, 1997). However, our results indicate two peaks in the expression, one in early spermatides and one at the late stages. In this cases it is necessary to verify the expression profile by classical methods.

Of the 956 unique sequences found in the CREM SSH library, 568 represent information about yet uncharacterized genes. Of these novel sequences, 369 show homology to 296 EST clones in the database, therefore a clone with longer sequence is easily available. Most of these sequences are expressed post-meiotically and dependent on CREM. The novel sequences as well as the ESTs provide interesting candidates for new genes with functions in sperm. Cloning of the full-length cDNA followed by sequence analysis would be necessary to learn more about these genes.

This study does not provide the information about cell specificity of gene expression; we only used total testis poly(A)$^+$ RNA. In future studies the microarray hybridization with mRNA from fractionated testis cells or *in situ* hybridizations will display the cell specificity of gene expression.

Taken together, our results demonstrate the CREM-dependent and post-meiotic expression of many known genes, as well as of novel sequences. The exact role of CREM in their expression can now be investigated further. It will be interesting to determine, which of these genes are direct CREM targets and which are influenced by secondary events. The role of products of these genes in post-meiotic stages of spermatogenesis may be the object of future studies as well. The sequences from the CREM SSH library can be useful for the detection of abnormalities in post-meiotic gene expression in infertile patients.

5.2 Comparison of the methods used for the identification of differentially expressed genes

In order to identify the mRNAs expressed in a CREMτ dependent manner we used suppression subtractive hybridization (SSH) as well as Affymetrix oligonucleotide DNA microarrays. SSH is an efficient method to clone differentially expressed genes for expression profiling experiments. The advantage of SSH compared to the use of mouse cDNA arrays is that by SSH novel sequences can be found. In fact in the CREM SSH library we found 255 of 956 unique sequences that would not have been available in public EST databases at the state of December 2000. It is, however, likely that in the near future DNA chips with sequences representing all mouse genes will be available. The SSH approach has limitations. We arbitrarily chose to pick a number of 12,000 clones based on the capacity we were able to analyze. The data analysis revealed that this number of clones is unlikely to contain all differentially expressed sequences in the CREM SSH library, because most of the sequences were found only once and many of the found genes were represented only by one *Rsa*I-fragment. It is however a considerable effort to analyze such large number of clones using sequencing. Another disadvantage is that by SSH no quantitative information about expression differences is obtained.

We therefore analyzed the expression levels of the genes in the CREM SSH library in more detail on cDNA nylon microarrays. This proofed that most of the sequences in the CREM SSH library show differential expression. We have noticed that hybridization experiments to be compared are best performed with the same filter, or with filters from the same production batch, because we observed significant differences with filters from different batches, often rendering the experiments incomparable.

As an additional technique we used Affymetrix oligonucleotide arrays to screen for additional genes with differential expression between wild-type testes and CREM (-/-) testes. The lists of genes that have been found to be differentially expressed with Affymetrix oligonucleotide arrays and by SSH are overlapping, but a large fraction appears only with one or the other technique. The results where both techniques provide comparable numbers are qualitatively largely in agreement, but can vary significantly on a quantitative level. The techniques appear to be complementing each other. Only few of the sequences show results clearly opposing each other. Opposing results might be due to families of closely related genes or alternative splice isoforms that show different measurements depending on features of the spotted sequence. These results would have to be checked with another method to make a clear statement about the expression of these genes.

Expression levels of lowly expressed genes are hard to measure by DNA microarray hybridization both by nylon cDNA arrays and by Affymetrix oligonucleotide arrays. For genes that display a signal close to background level the calculated ratios get very high and tend to be not good indicators of

differential expression (Section 4.1.2). The enrichment of particular clones by the SSH procedure provides additional evidence of differential expression of the corresponding genes. For future studies the resulting cDNA pool as obtained by SSH could be analyzed more easily by hybridization with DNA chips containing representatives for all genes. The factor of enrichment by SSH seems to provide an indication about the differential expression of a gene. This would be an efficient method to find lowly expressed target genes.

5.3 Computational methods to analyze gene expression profiles

The technique of expression profiling by means of hybridization to DNA microarrays offers a new tool to investigate the expression levels of thousands of genes at the same time. But comparison of hybridization experiments is hampered by the fact that differences in signal intensities might not only represent true expression changes but also to experimental variabilities, which are often in the same range as the differences one expects to occur by differential expression. Corrections for several methodological influences on the measured intensities, like incorporation of radioactive label or exposure time were developed for nylon cDNA microarrays.

The most basic data processing techniques, background correction and linear transformation of the data have been described in this thesis (Section 4.1.2). There is, however, discrepancy in the literature about the methods to find an adjusting factor for standardization. Chen *et al.* (1997) use a ratio distribution function derived for two normally distributed data sets with a constant factor of variance to standardize by an iterative procedure. Piétu *et al.* (1996) standardize by subtracting the mean of the logarithmic data and dividing by their standard deviation. This implies that the intensities follow a normal distribution. This holds true only for a subpopulation of spots as was shown in this thesis and thus should not be used for standardization. Richmond *et al.* (1999) calculate the relative percentage of total signal as a means of standardization. We observe that intensity values on an array do not in their entirety follow a log-normal distribution. This distribution may be used to define the background on one array but is not suited to standardize for comparison between hybridizations.

Standardization is a prerequisite to a more thorough study of the data in order to reveal the inherent information. Comparison between pairs of condition usually results in long lists of potentially differentially expressed genes. Long lists, however, do not necessarily lead to biological understanding. Studies with series of experiments, *e.g.* with a time course, a concentration series or different tumor stages provide additional information. To partition these long lists into smaller, more easy to analyze portions, frequently clustering of expression profiles is used. Clustering of gene expression profiles aims to find groups of co-expressed and therefore possibly co-regulated or functionally

related genes.

However, clustering has a number of disadvantages. It is often not possible to detect distinct clusters in the dataset or to find the exact number of distinguishable groups. Therefore, in hierarchical clustering a tree is constructed. This tree can be cut at different levels. However, deeper levels of the tree tend to be of lower significance. To avoid the problem of defining borders between clusters the tree is frequently transformed into a linear order by which to visually inspect the data (Eisen and Brown, 1999). This procedure can be misleading because there is not a unique linear order resulting from hierarchical clustering. For a tree there are 2^{N-1} linear orders of the leaves possible. Neighbors in this order might be very far apart in the tree. Frequently mistakes are made because of this, and the tree structure is frequently ignored or not even visualized in presentations. A recent report from Z. Bar-Joseph *et al* (2001) presents an algorithm to arrange the leaves of a hierarchical clustering tree optimally.

In this thesis a different approach to detect genes with similar expression profiles is presented. We have used a novel approach to visualize the expression data in a linear order by sorting them in a way that similar expression profiles are arranged close to each other. We have approximated the solution of a traveling salesman problem to get such a linear representation. Thus we offer a more appropriate mathematical solution to a frequently used presentation of expression data.

To gain further biological understanding from expression data it is important to link it to other sources of information. One source may be the information about known regulatory pathways. At the moment our aim is to present and order the data in ways that make it possible to visually inspect the data and detect interesting correlations between genes. An interesting task for the future is to correlate clusters of gene expression with sequence patterns in promoter regions. This is easier for yeast, as typically yeast promoters are relatively short (Brazma et al., 1998; van Helden et al., 1998; Vilo et al., 2000). Further understanding and availability of gene promoters is necessary to be able to accomplish this task in mouse or man.

Other more ambitious attempts try to reconstruct functional networks based directly on the expression data (Chen et al., 1999; Friedman et al., 2000). This is clearly a hard problem. The current data are extremely noisy. Moreover, mRNA expression data only give a partial picture that does not reflect key events such as translation and protein (in)activation.

5.4 Perspectives for Bioinformatics

In this context it is interesting that on February 15th, 2001, a first, nearly complete draft of the human genome was published in a special issue of *Nature*. At the same date another draft by the company Celera was announced in *Science*. This opens up new perspectives for the analysis of

genes of a complete genome. It can be expected that the mouse genome will be available soon, too. The usefulness of sequence data depends, however, very much on their quality of annotation and accessibility.

A next important step will be to identify all of the genes. Some methods predict the structure of the gene, i.e. the translated regions, based on the nucleotide composition (Burge and Karlin, 1997). This gives, however, only an estimate of the exact exon/intron borders and the structure of the gene. Data of expressed sequence tags (ESTs) can be very useful to get further insight into the gene structure. Fast algorithms exist to align EST sequences to a genomic sequence (Florea et al., 1998). Databases of clustered EST datasets like the one described in this thesis will be very useful to detect the genes. An ongoing project uses the EST data to predict alternative splice variants that can be found as an expressed transcript (E. Coward, M. Vingron, *unpublished*).

The GeneNest database can be used as a general tool for sequence analysis apart from its use for the CREM project. Currently, the GeneNest database[1] comprises gene indices of man (based on UNIGENE), mouse, *Arabidopsis thaliana* and *Zebrafish*. GeneNest combines properties of both UNIGENE and the TIGR gene indices. Similar to UNIGENE, clusters represent sequences related to one gene. However, GeneNest clusters are solely based on sequence homologies, while links between clusters containing sequences of the same clone are visualized. Because of the high rate of misannotation in public databases this strategy avoids clustering of unrelated sequences, still presenting all sequence relationships to the user. Contigs which are generated by GeneNest often reflect single transcripts in a similar way as clusters of the TIGR gene indices. The comparison of contigs provides insight into the genomic structure of the gene represented. The comprehensive database of consensus sequences summarizing every putative transcript can be used as an efficient tool to search for homologies to private sequences. This way multiple searching of sequence databases, often containing only fragments of transcripts, can be avoided. The interactive interface of GeneNest together with its compact presentation of cluster/gene related data minimizes the manual interaction for the user (S. Haas, T. Beißbarth *et al.*, 2000).

Analysis of data sets similar to the data presented here will be dramatically simplified as soon as a list of all genes is available. At the moment the analysis of sequences is made difficult by low quality of sequence databases and bad or inconsistent annotation.

Another interesting aspect about the availability of genomic data is the analysis of gene regulatory elements. In databases like TRANSFAC[2] several hundred regulatory elements that constitute binding sites for transcription factors are summarized (Wingender et al., 2001). However, gene regulation in eukaryotes is very complex and regulatory sequences can be far apart from the coding sequence. As the regulatory elements are very short sequences, there is a high chance to also find them by chance.

[1] http://www.dkfz.de/tbi/services/GeneNest/index
[2] http://transfac.gbf.de

To be able to make more reliable predictions of a given binding site, several regulatory elements should be found in a short stretch of sequence. These elements should be correlated with clusters of similarly expressed genes. Further analysis of regulatory regions is necessary in order to build predictors.

5.5 Accomplishments of this thesis

This thesis provides an extensive study of CREM dependent genes during mouse spermatogenesis. Many known genes were analyzed for their expression and possible role during spermatogenesis. Many new genes were found that are expressed specifically and CREM dependently at the late stages of spermatogenesis and therefore are likely to play roles in the final development of sperm.

We have picked 12,000 clones by the suppression subtractive hybridization and detected 956 unique sequences. These sequences show homologies to 158 known mouse genes, 99 other known genes and 296 ESTs. 199 sequences are novel. The expression profiles show that most of these clones are down-regulated in CREM deficient mice and expressed at the post-meiotic stages of spermatogenesis. Among these sequences several important groups of functionally related genes were found to be co-expressed post-meiotically and CREM dependently. We found novel information about the expression of hundreds of known genes, ESTs and novel sequences. Many of those sequences might be useful for the detection of abnormalities in testes of infertile males. All the data from the project was summarized in a database which is accessible from the world wide web[3], and a manuscript describing the biological results of this work is in preparation (Beißbarth, Borisevich, et al.).

During the process of analyzing this data several new methods in Bioinformatics were developed and have been applied in many other projects as well.

For large scale expression analysis on DNA microarrays it is an essential step to be able to compare data sets. The expression levels of thousands of genes can be measured on DNA microarrays in parallel. To be able to analyze the data presented in this thesis a strategy comparing the numbers from several hybridization experiments was developed and software was implemented for pairwise comparison of high density nylon array hybridizations. This method can be used more generally (Beißbarth et al., 2000) and an interface for comparing experiments is accessible in the world wide web[4].

Analysis of gene expression data usually aims to find co-regulated genes by clustering the gene expression profiles (Brazma and Vilo, 2000). Hierarchical clustering methods are frequently used to arrange expression profiles in a linear order according to their similarity (Eisen and Brown, 1999).

[3] http://www.dkfz.de/tbi/crem: For access information please contact G.Schuetz@dkfz.de
[4] http://wwww.dkfz.de/tbi/services/matlab2web/webdiffs

We rather use an algorithm to approximate the linear order with the minimal sum of the distances between all neighbors, corresponding to a traveling salesman problem.

EST sequences provide an important source of information about genes. To analyze the sequences of the CREM SSH library EST data were used. To make efficient use of EST data a searchable interface for clustered EST databases as well as a software for visual presentation and structures for efficient storage and handling of the data was developed. This tool is publicly available via a web interface[5] (Haas, Beißbarth et al., 2000).

[5] http://www.dkfz.de/tbi/services/GeneNest/index

6 Acknowledgments

This work represents the outcome of extensive collaborations and I would like to thank all people that helped me in this thesis. People involved in parts of the work or who developed methods that were used are cited accordingly. The work was largely performed in the groups for “Molecular Biology of the cell I” and “Theoretical Bioinformatics” at the German Cancer Research Center in Heidelberg. I want to thank all members and former members of the groups “Theoretical Bioinformatics”, “Molecular Biology of the Cell I” and “Functional Genome Analyzis” for the nice working atmosphere.

Especially, I want to thank the supervisors of this work Prof. Dr. Martin Vingron and Prof. Dr. Günther Schütz for their help. I thank Prof. Dr. Richard Herrmann for co-correcting this thesis.

The experimental parts on the SSH were carried out by Dr. Igor Borisevich, Dr. Andreas Hörlein, and Annette Klewe-Nebenius. Ralf Klären helped us with the sequencing. The spotting and hybridization of the nylon CREM SSH cDNA microarrays was performed by Dr. Igor Borisevich. The methods how these data were obtained and some of the biological interpretations are presented in more detail in his PhD thesis (Borisevich, 2001). The Affymetrix oligonucleotide array hybridizations were performed by Dr. Mark Kenzelmann and Ralf Klären. I want to thank these people for their invaluable contributions.

In the field of expression analysis I was able to work with data generated in the groups of Prof. Dr. Annemarie Poustka, as well as Dr. Joerg Hoheisel. The following people did experimental work and helped me a lot in discussions: Dr. Rosa Arribas, Maja Vujic, Dr. Marcel Scheideler, Dr. Nicole Hauser, Dr. Judith Boer, Dr. Ekkehard Werner and Dr. Soeren Eichhorst. On the Bioinformatics side I had a lot of interesting discussions with Kurt Fellenberg and Dr. Benedikt Brors. Also thanks to Dr. Alvis Brazma and Dr. Jaak Vilo from the European Bioinformatics Institute. I have learned a lot of mathematics from Dr. Rainer Spang and Dr. Tobias Mueller. I have learned a lot about transcription factors from Dr. Erich Greiner. Bernd Binder ported my program code for a expression analysis software to a web base interface.

For setting up a database of expressed sequence tags, which was essential for the data analysis performed here I worked together with Dr. Stefan Haas and Dr. Eric Rivals. A lot of helpful input

came from Dr. Bernhard Korn. I am also very thankful to Antje Krause, who locally installed many of the sequence databases that were used here.

For setting up an ACEDB database system I had very helpful discussions with Dr. Nicolas Pollet and with Dr. Anja Kuehl. I also thank Dr. Karl-Heinz Glatting for some helpful advices about sequence analysis procedures.

The work relied on our permanently running computer systems, which were setup and maintained by Heiko Schmidt and Karl-Heinz Gross.

Further thanks to Dr. Erich Greiner, Dr. Benedikt Brors, Dr. Stefan Haas, Dr. Rosa Arribas, Dr. Mark Kenzelmann, Dr. Wolfgang Schmid, Dr. Wolfgang Huber for critically reading and correcting the manuscript of this thesis. Thanks to Daniela Thomas for reading the manuscript and for moral support.

7 Abbreviations

ACE - angiotensin converting enzyme.
ACT - activator of CREM in testis.
AEG1 - acidic epididymal glycoprotein.
AKAP - A kinase anchoring protein.
Ala - Alanin.
ATF - activating transcription factor.
ATP - adenosine 5’-triphosphate.
BLAST -basic local alignment search tool.
bp - base pairs.
bZip - basic leucine zipper.
cAMP - cyclic adenosin monophosphate
CBP - CREB binding protein.
cDNA - complementary DNA.
CRE - cAMP-response element.
CREB - cAMP responcive element binding protein.
CREM - cAMP responcive element modulator.
DBD - DNA binding domain.
DNA - deoxyribonucleic acid.
EST - expressed sequence tag.
FSH - follicle stimulating hormone.
Gapd-S - testis specific glyceraldehyde 3-phosphatedehydrogenase.
ICER - inducible cAMP early repressor.
LH - luteinizing hormone.
MCS - mitochondrial capsule selenoprotein.
mRNA - messenger RNA.
ODF - outer dense fiber protein.
ORF - open reading frame.
PCR - polymerase chain reaction.

PKA - protein kinase A.
Pol - Polymerase.
RNA - ribonucleic acid.
SSH - suppression subtractive hybridisation.
Ser - serin.
TAF - transcription activation factor.
TBP - TATA-box binding protein.
TP - transition protein.
TSP - treveling salesman problem
wt - wild-type.

Bibliography

Adams, M. D., Kelley, J. M., Gocayne, J. D., Dubnick, M., Polymeropoulos, M. H., Xiao, H., Merril, C. R., Wu, A., Olde, B., Moreno, R. F., Kerlavage, A. R., McCombie, W. R., and Venter, J. C. (1991). Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*, 252:1651–1656.

Adams, M. D., Kerlavage, A. R., Fleischmann, R. D., Fuldner, R. A., Bult, C. J., Lee, N. H., Kirkness, E. F., Weinstock, K. G., Gocayne, J. D., White, O., and a. l. et (1995). Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature*, 377(6547 Suppl):3–174.

Alberts, B., Bray, D., J.Lewis, M.Raff, Roberts, K., and Watson, J., editors (1994). *Molecular Biology of the Cell*. Garland Publishing, Inc., 3 edition.

Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., Powell, J. I., Yang, L., Marti, G. E., Moore, T., r. Hudson J, J., Lu, L., Lewis, D. B., Tibshirani, R., Sherlock, G., Chan, W. C., Greiner, T. C., Weisenburger, D. D., Armitage, J. O., Warnke, R., Staudt, L. M., and a. l. et (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403(6769):503–11.

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., and Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA*, 96:6745–6750.

Alter, O., Brown, P. O., and Botstein, D. (2000). Singular value decomposition for genome-wide expression data processing and modeling. *Proc Natl Acad Sci U S A*, 97(18):10101–6.

Altrock, v. (1995). *Fuzzy Logic*, chapter 1-3. R. Oldenbourg, München.

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

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, 25(17):3389–402.

Baarends, W. M., Hoogerbrugge, J. W., Roest, H. P., Ooms, M., Vreeburg, J., Hoeijmakers, J. H., and Grootegoed, J. A. (1999). Histone ubiquitination and chromatin remodeling in mouse spermatogenesis. *Dev Biol*, 207(2):322–33.

Bairoch, A. and Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res*, 28(1):45–8.

Banky, P., Huang, L. J., and Taylor, S. S. (1998). Dimerization/docking domain of the type Ialpha regulatory subunit of cAMP-dependent protein kinase. Requirements for dimerization and docking are distinct but overlapping. *J Biol Chem*, 273(52):35048–55.

Bar-Joseph, Z., Gifford, D. K., and Jaakkola, T. S. (2001). Fast Optimal Leaf Ordering for Hierarchical Clustering. In *ISMB 01: Proceedings of the Ninth International Conference on Intelligent Systems for Molecular Biology*.

Bartsch, D., Casadio, A., Karl, K. A., Serodio, P., and Kandel, E. R. (1998). CREB1 encodes a nuclear activator, a repressor, and a cytoplasmic modulator that form a regulatory unit critical for long-term facilitation. *Cell*, 95(2):211–23.

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

Beißbarth, T., Fellenberg, K., Brors, B., Arribas-Prat, R., Boer, J. M., Hauser, N. C., Scheideler, M., Hoheisel, J. D., Schütz, G., Poustka, A., and Vingron, M. (2000). Processing and quality control of DNA array hybridization data. *Bioinformatics*, 16(11):1014–22.

Bellve, A. R., Cavicchia, J. C., Millette, C. F., O'Brien, D. A., Bhatnagar, Y. M., and Dym, M. (1977). Spermatogenic cells of the prepuberal mouse. Isolation and morphological characterization. *J Cell Biol*, 74(1):68–85.

Ben-Dor, A., Bruhn, L., Friedman, N., Nachman, I., Schummer, M., and Yakhini, Z. (2000). Tissue Classification and Gene Expression Profiles. In Shamir, R., Miyano, S., Istrail, S., Pevzner, P., and Waterman, M., editors, *Proceedings of the 4th annual conference on computational molecular biology (RECOMB 00)*, pages 54–64. ACM Press.

Ben-Dor, A. and Yakhini, Z. (1999). Clustering Gene Expression Patterns. In Istrail, S., Pevzner, P., and Waterman, M., editors, *Proceedings of the 3rd annual conference on computational molecular biology (RECOMB 99)*, pages 33–42. ACM Press.

Birney, E., Bateman, A., Clamp, M. E., and Hubbard, T. J. (2001). Mining the draft human genome. *Nature*, 409(6822):827–8.

Blendy, J. A., Kaestner, K. H., Weinbauer, G. F., Nieschlag, E., and Schütz, G. (1996). Severe impairment of spermatogenesis in mice lacking the CREM, gene. *Nature*, 380:162–165.

Bonfield, J. K., f. Smith, K., and Staden, R. (1995). A new DNA sequence assembly program. *Nucleic Acids Res*, 23(24):4992–9.

Borisevich, I. (2001). *Analysis of transcription factor CREM target gene expression during mouse spermatogenesis*. PhD thesis, Universitaet Heidelberg, DKFZ, Im Neuenheimer Feld 280, 69120 Heidelberg.

Bouck, J., Yu, W., Gibbs, R., and Worley, K. (1999). Comparison of gene indexing databases. *Trends Genet*, 15(4):159–62.

Brazma, A., Jonassen, I., Vilo, J., and Ukkonen, E. (1998). Predicting gene regulatory elements in silico on a genomic scale. *Genome Res*, 8(11):1202–15.

Brazma, A. and Vilo, J. (2000). Gene expression data analysis. *FEBS Lett*, 480(1):17–24.

Browder, L. W., Erickson, C. A., and Jeffery, W. R. (1991). *Developmental biology*, volume 1. Saunders College Publishing, third edition edition.

Brown, M. P., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C. W., Furey, T. S., r. Ares M, J., and Haussler, D. (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc Natl Acad Sci U S A*, 97(1):262–7.

Burge, C. and Karlin, S. (1997). Prediction of complete gene structures in human genomic dna. *J Mol Biol*, 268(1):78–94. (eng).

Burke, J., Wang, H., Hide, W., and Davison, D. B. (1998). Alternative gene form discovery and candidate gene selection from gene indexing projects. *Genome Res*, 8(3):276–90.

Carrera, A., Gerton, G. L., and Moss, S. B. (1994). The major fibrous sheath polypeptide of mouse sperm: structural and functional similarities to the A-kinase anchoring proteins. *Dev Biol*, 165(1):272–84.

Chavrier, P., Janssen-Timmen, U., Mattei, M. G., Zerial, M., Bravo, R., and Charnay, P. (1989). Structure, chromosome location, and expression of the mouse zinc finger gene Krox-20: multiple gene products and coregulation with the proto-oncogene c-fos. *Mol Cell Biol*, 9(2):787–97.

Chen, Q., Lin, R. Y., and Rubin, C. S. (1997a). Organelle-specific targeting of protein kinase AII (PKAII). Molecular and in situ characterization of murine A kinase anchor proteins that recruit regulatory subunits of PKAII to the cytoplasmic surface of mitochondria. *J Biol Chem*, 272(24):15247–57.

Chen, T., Filkov, V., and Skiena, S. S. (1999). Identifying regulatory networks from experimental data. In Istrail, S., Pevzner, P., and Waterman, M., editors, *Proceedings of the 3rd annual conference on computational molecular biology (RECOMB)*, pages 94–103. ACM Press.

Chen, Y., Dougherty, E. R., and Bittner, M. (1997b). Ratio-based decisions and the quantitative analysis of cDNA microarray images. *J. Biomed. Opt.*, 2:364–374.

Cheng, Y. and Church, G. M. (2000). Biclustering of Expression Data. In Altman, R., Baily, T. L., Bourne, P., Gribskov, M., Lengauer, T., an Lynn F.T. Eyck, I. N. S., and Weissig, H., editors, *ISMB 2000: Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*, pages 307–316, 93–103.

Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P. O., and Herskowitz, I. (1998). The transcriptional program of sporulation in budding yeast [published erratum appears in Science 1998 Nov 20;282(5393):1421]. *Science*, 282(5389):699–705.

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

De Cesare, D., Fimia, G. M., and Sassone-Corsi, P. (1999). Signaling routes to CREM and CREB: plasticity in transcriptional activation. *Trends Biochem Sci*, 24(7):281–5.

de Groot, R. P. and Sassone-Corsi, P. (1993). Hormonal control of gene expression: multiplicity and versatility of cyclic adenosine 3',5'-monophosphate-responsive nuclear regulators [published erratum appears in Mol Endocrinol 1993 Apr;7(4):603]. *Mol Endocrinol*, 7(2):145–53.

Delmas, V., van der Hoorn, F., Mellstrom, B., Jegou, B., and Sassone-Corsi, P. (1993). Induction of CREM activator proteins in spermatids: down-stream targets and implications for haploid germ cell differentiation. *Mol Endocrinol*, 7(11):1502–14.

DeRisi, J., Penland, L., Brown, P. O., Bittner, M. L., Meltzer, P. S., Ray, M., Chen, Y., Su, Y. A., and Trent, J. M. (1996). Use of a cDNA microarray to analyse gene expression patterns in human cancer [see comments]. *Nat Genet*, 14(4):457–60.

DeRisi, J. L., Iyer, V. R., and Brown, P. O. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, 278(5338):680–6.

Diatchenko, L., Lau, Y. F., Campbell, A. P., Chenchik, A., Moqadam, F., Huang, B., Lukyanov, S., Lukyanov, K., Gurskaya, N., Sverdlov, E. D., and Siebert, P. D. (1996). Suppression subtractive hybridization: a method for generating differentially regulated or tissue-specific cDNA probes and libraries. *Proc Natl Acad Sci U S A*, 93(12):6025–30.

Diatchenko, L., Lukyanov, S., Lau, Y. F., and Siebert, P. D. (1999). Suppression subtractive hybridization: a versatile method for identifying differentially expressed genes. *Methods Enzymol*, 303:349–80.

Dudoit, S., Fridlyand, J., and Speed, T. P. (2000a). Comparison of discrimination methods for the Classification of tumors using gene expression data. Technical Report 576, Dept. of Statistics, University of California at Berkeley, Berkeley, CA.

Dudoit, S., Yang, Y., Callow, M., and Speed, T. P. (2000b). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Technical Report 578, Dept. of Statistics, University of California at Berkeley, Berkeley, CA.

Eeckman, F. H. and Durbin, R. (1995). ACeDB and macace. *Methods Cell Biol*, 48:583–605.

Eisen, M. B. and Brown, P. O. (1999). DNA arrays for analysis of gene expression. *Methods Enzymol*, 303:179–205.

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A*, 95(25):14863–8.

Fellenberg, K., Hauser, N. C., Brors, B., Neutzner, A., Hoheisel, J. D., and Vingron, M. (2001). Correspondence anaylsis applied to microarray data. *Proc Natl Acad Sci USA*, 98:10781–10786.

Fimia, G. M., De Cesare, D., and Sassone-Corsi, P. (1999). CBP-independent activation of CREM and CREB by the LIM-only protein ACT. *Nature*, 398(6723):165–9.

Fimia, G. M., Morlon, A., Macho, B., Cesare, D. D., and Sassone-Corsi, P. (2001). Transcriptional cascades during spermatogenesis; pivotal role of CREM and ACT. *Mol Cell Endocrinol*, 179:17–23.

Florea, L., Hartzell, G., Zhang, Z., Rubin, G. M., and Miller, W. (1998). A computer program for aligning a cdna sequence with a genomic dna sequence. *Genome Res*, 8(9):967–74. (eng).

Foulkes, N. S., Borjigin, J., Snyder, S. H., and Sassone-Corsi, P. (1997). Rhythmic transcription: the molecular basis of circadian melatonin synthesis. *Trends Neurosci*, 20(10):487–92.

Foulkes, N. S., Borrelli, E., and Sassone-Corsi, P. (1991). CREM gene: use of alternative DNA-binding domains generates multiple antagonists of cAMP-induced transcription. *Cell*, 64(4):739–49.

Foulkes, N. S., Duval, G., and Sassone-Corsi, P. (1996). Adaptive inducibility of CREM as transcriptional memory of circadian rhythms. *Nature*, 381(6577):83–5.

Foulkes, N. S., Mellstrom, B., Benusiglio, E., and Sassone-Corsi, P. (1992). Developmental switch of CREM function during spermatogenesis: from antagonist to activator. *Nature*, 355(6355):80–4.

Foulkes, N. S., Schlotter, F., Pevet, P., and Sassone-Corsi, P. (1993). Pituitary hormone FSH directs the CREM functional switch during spermatogenesis. *Nature*, 362(6417):264–7.

Friedman, N., Linial, M., Nachman, I., and Pe'er, D. (2000). Using Bayesian networks to analyze expression data. In Shamir, R., Miyano, S., Istrail, S., Pevzner, P., and Waterman, M., editors, *Proceedings of the 4th annual conference on computational molecular biology (RECOMB 00)*, pages 127–135. ACM Press.

Friemert, C., Erfle, V., and Strauss, G. (1989). Preparation of radiolabeled cDNA probes with high specific activity for rapid screening of gene expression. *Methods Mol. Cell. Biol.*, 1:143–153.

Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M., and Haussler, D. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16(10).

Galliot, B., Welschof, M., Schuckert, O., Hoffmeister, S., and Schaller, H. C. (1995). The cAMP response element binding protein is involved in hydra regeneration. *Development*, 121(4):1205–16.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Holler, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537.

Goraya, T. Y., Kessler, S. P., Stanton, P., Hanson, R. W., and Sen, G. C. (1995). The cyclic AMP response elements of the genes for angiotensin converting enzyme and phosphoenolpyruvate carboxykinase (GTP) can mediate transcriptional activation by CREM tau and CREM alpha. *J Biol Chem*, 270(32):19078–85.

Gurskaya, N. G., Diatchenko, L., Chenchik, A., Siebert, P. D., Khaspekov, G. L., Lukyanov, K. A., Vagner, L. L., Ermolaeva, O. D., Lukyanov, S. A., and Sverdlov, E. D. (1996). Equalizing cDNA subtraction based on selective suppression of polymerase chain reaction: cloning of Jurkat cell

transcripts induced by phytohemaglutinin and phorbol 12-myristate 13-acetate. *Anal Biochem*, 240(1):90–7.

Ha, H., van, W. A., and Hecht, N. B. (1997). Tissue-specific protein-DNA interactions of the mouse protamine 2 gene promoter. *J Cell Biochem*, 64(1):94–105.

Haas, S., Beissbarth, T., Rivals, E., Krause, A., and Vingron, M. (2000). GeneNest: automated generation and visualization of gene indices. *Trends Genet.*, 16(11):521–523.

Haussler, D. and Opper, M. (1997). Mutual Information, Metric Entropy, and Cumulative Relative Entropy Risk. *Annals of Statistics*, 25(6).

Herrada, G. and Wolgemuth, D. J. (1997). The mouse transcription factor Stat4 is expressed in haploid male germ cells and is present in the perinuclear theca of spermatozoa. *J Cell Sci*, 110(Pt 14):1543–53.

Hilsenbeck, S. G., Friedrichs, W. E., Schiff, R., O'Conell, P., Hansen, R. K., Osborne, C. K., and Fuqua, S. A. W. (1999). Statistical analysis of array expression data as applied to the problem of tamoxifen resistance. *J. Natl. Canc. Inst.*, 91:453–459.

Hoeffler, J. P., Meyer, T. E., Yun, Y., Jameson, J. L., and Habener, J. F. (1988). Cyclic AMP-responsive DNA-binding protein: structure based on a cloned placental cDNA. *Science*, 242(4884):1430–3.

Holstege, F. C. P., Jennings, E. G., Wyrick, J. J., Lee, T. I., Hengartner, C. J., Green, M. J., Golub, T. R., Lander, E. S., and Young, R. A. (1998). Dissecting the regulatory circuitry of a eukaryotic genome. *Cell*, 95:717–728.

Iyer, V. R., Eisen, M. B., Ross, D. T., Schuler, G., Moore, T., Lee, J. C. F., Trent, J. M., Staudt, L. M., r. Hudson J, J., Boguski, M. S., Lashkari, D., Shalon, D., Botstein, D., and Brown, P. O. (1999). The transcriptional program in the response of human fibroblasts to serum. *Science*, 283(5398):83–7.

Kerr, K., Martin, M., and Churchill, G. (2000). Analysis of Variance for Gene Expression Microarray Data. *Journal of Computational Biology*.

Kessler, S. P., Rowe, T. M., Blendy, J. A., Erickson, R. P., and Sen, G. C. (1998). A cyclic AMP response element in the angiotensin-converting enzyme gene and the transcription factor CREM are required for transcription of the mRNA for the testicular isozyme. *J Biol Chem*, 273(16):9971–5.

Kistler, M. K., Sassone-Corsi, P., and Kistler, W. S. (1994). Identification of a functional cyclic adenosine 3',5'-monophosphate response element in the 5'-flanking region of the gene for transition protein 1 (TP1), a basic chromosomal protein of mammalian spermatids. *Biol Reprod*, 51(6):1322–9.

Krause, A., Stoye, J., and Vingron, M. (2000). The SYSTERS protein sequence cluster set. *Nucleic Acids Res*, 28(1):270–2.

Kuroyanagi, N., Onogi, H., Wakabayashi, T., and Hagiwara, M. (1998). Novel SR-protein-specific kinase, SRPK2, disassembles nuclear speckles. *Biochem Biophys Res Commun*, 242(2):357–64.

Laoide, B. M., Foulkes, N. S., Schlotter, F., and Sassone-Corsi, P. (1993). The functional versatility of CREM is determined by its modular structure. *Embo J*, 12(3):1179–91.

Lennon, G. G. and Lehrach, H. (1991). Hybridization analyses of arrayed cDNA libraries. *Trends Genet.*, 7:314–317.

Maldonado, R., Smadja, C., Mazucchelli, C., and Sassone-Corsi, P. (1999). Altered emotional and locomotor responses in mice deficient in the transcription factor CREM. *Proc Natl Acad Sci U S A*, 96(24):14094–9.

Mironov, A. A., Fickett, J. W., and Gelfand, M. S. (1999). Frequent alternative splicing of human genes. *Genome Res*, 9(12):1288–93.

Molina, C. A., Foulkes, N. S., Lalli, E., and Sassone-Corsi, P. (1993). Inducibility and negative autoregulation of CREM: an alternative promoter directs the expression of ICER, an early response repressor. *Cell*, 75(5):875–86.

Montminy, M. (1997). Transcriptional regulation by cyclic AMP. *Annu Rev Biochem*, 66:807–22.

Nantel, F., Monaco, L., Foulkes, N. S., Masquilier, D., LeMeur, M., Henriksen, K., Dierich, A., Parvinen, M., and Sassone-Corsi, P. (1996). Spermiogenesis deficiency and germ-cell apoptosis in CREM-mutant mice. *Nature*, 380(6570):159–62.

Newton, M., Kendziorski, C., Richmond, C., Blattner, F., and Tsui, K. (2000). On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology*.

Piétu, G., Alibert, O., Guichard, V., Lamy, B., Bois, F., Leroy, E., Mariage-Samson, R., Houlgatte, R., Soularue, P., and Auffray, C. (1996). Novel gene transcripts preferentially expressed in human muscles revealed by quantitative hybridization of a high density cDNA array". *Genome Res.*, 6:492–503.

Pollet, N., Schmidt, H. A., Gawantka, V., Vingron, M., and Niehrs, C. (2000). Axeldb: a Xenopus laevis database focusing on gene expression. *Nucleic Acids Res*, 28(1):139–40.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (1992). *Numerical Recipes in C*, chapter 10.9, pages 444–455. CAMBRIDGE UNIVERSITY PRESS.

Radhakrishnan, I., Perez-Alvarado, G. C., Parker, D., Dyson, H. J., Montminy, M. R., and Wright, P. E. (1997). Solution structure of the KIX domain of CBP bound to the transactivation domain of CREB: a model for activator:coactivator interactions. *Cell*, 91(6):741–52.

Rebhan, M., Chalifa-Caspi, V., Prilusky, J., and Lancet, D. (1998). GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support. *Bioinformatics*, 14(8):656–64.

Rehfuss, R. P., Walton, K. M., Loriaux, M. M., and Goodman, R. H. (1991). The cAMP-regulated enhancer-binding protein ATF-1 activates transcription in response to cAMP-dependent protein kinase A. *J Biol Chem*, 266(28):18431–4.

Richmond, C. S., Glasner, J. D., Mau, R., Jin, H., and Blattner, F. R. (1999). Genome-wide expression profiling in *Escherichia coli* K-12. *Nucl. Acids Res.*, 27:3821–3835.

Russel, L. D., Ettlin, R. A., Hikim, A. P. S., and Clegg, E. D. (1990). *Hisstological and histopatological evaluation of the testis*. Cache River Press, 1 edition.

Sachs, L. (1984). *Applied Statistics*, pages 107–110. Springer Verl., Berlin, Heidelberg, New York, Tokio, 2nd edition.

Sassone-Corsi, P. (1995). Transcription factors responsive to cAMP. *Annu Rev Cell Dev Biol*, 11:355–77.

Sassone-Corsi, P. (1997). Transcriptional checkpoints determining the fate of male germ cells. *Cell*, 88(2):163–6.

Sassone-Corsi, P. (1998). CREM: a master-switch governing male germ cells differentiation and apoptosis. *Semin Cell Dev Biol*, 9(4):475–82.

Sassone-Corsi, P. (2000). CREM: a master-switch regulating the balance between differentiation and apoptosis in male germ cells. *Mol Reprod Dev*, 56(2 Suppl):228–9.

Schmitt, A. O., Specht, T., Beckmann, G., Dahl, E., Pilarsky, C. P., Hinzmann, B., and Rosenthal, A. (1999). Exhaustive mining of EST libraries for genes differentially expressed in normal and tumour tissues. *Nucleic Acids Res*, 27(21):4251–60.

Schuchhardt, J., Beule, D., Malik, A., Wolski, E., Eickhoff, H., Lehrach, H., and Herzel, H. (2000). Normalization strategies for cDNA microarrays. *Nucleic Acids Res*, 28(10):E47.

Schuler, G. D. (1997). Pieces of the puzzle: expressed sequence tags and the catalog of human genes. *J Mol Med*, 75(10):694–8.

Servillo, G., Della Fazia, M. A., and Sassone-Corsi, P. (1998). Transcription factor CREM coordinates the timing of hepatocyte proliferation in the regenerating liver. *Genes Dev*, 12(23):3639–43.

Shalon, D., Smith, S. J., and Brown, P. O. (1996). A DNA microarray system for analyzing complex DNA samples using two- color fluorescent probe hybridization. *Genome Res*, 6(7):639–45.

Sharan, R. and Shamir, R. (2000). CLICK: A Clustering Algorithm with Application to Gene Expression Analysis. In Altman, R., Baily, T. L., Bourne, P., Gribskov, M., Lengauer, T., an Lynn F.T. Eyck, I. N. S., and Weissig, H., editors, *ISMB 2000: Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*, pages 307–316, Tel-Aviv, Israel.

Silva, A. J., Kogan, J. H., Frankland, P. W., and Kida, S. (1998). CREB and memory. *Annu Rev Neurosci*, 21:127–48.

Spang, R., Zuzan, H., West, M., Nevins, J., Blanchette, C., and Marks, J. R. (2000). Prediction and uncertainty in the analysis of gene expression profiles. Technical Report 00-31, Institute of Statistics, Duke University, Durham, North Carolina.

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., and Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol Biol Cell*, 9(12):3273–97.

Staden, R., Beal, K. F., and Bonfield, J. K. (2000). The Staden package, 1998. *Methods Mol Biol*, 132:115–30.

Stoesser, G., Baker, W., van den Broek, A., Camon, E., Garcia-Pastor, M., Kanz, C., Kulikova, T., Lombard, V., Lopez, R., Parkinson, H., Redaschi, N., Sterk, P., Stoehr, P., and Tuli, M. A. (2001). The EMBL nucleotide sequence database. *Nucleic Acids Res*, 29(1):17–21.

Struthers, R. S., Vale, W. W., Arias, C., Sawchenko, P. E., and Montminy, M. R. (1991). Somatotroph hypoplasia and dwarfism in transgenic mice expressing a non-phosphorylatable CREB mutant. *Nature*, 350(6319):622–4.

Sun, Z. and Means, A. R. (1995). An intron facilitates activation of the calspermin gene by the testis- specific transcription factor CREM tau [published erratum appears in 1995 Nov 24;270(47):28494]. *J Biol Chem*, 270(36):20962–7.

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E. S., and Golub, T. R. (1999). Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. USA*, 96:2907–2912.

Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J., and Church, G. M. (1999). Systematic determination of genetic network architecture. *Nature Genet.*, 22:281–285.

Tchernitsa, O. I., Zuber, J., Sers, C., Brinckmann, R., Britsch, S. K., Adams, V., and Schafer, R. (1999). Gene expression profiling of fibroblasts resistant toward oncogene- mediated transformation reveals preferential transcription of negative growth regulators. *Oncogene*, 18(39):5448–54.

van Helden, J., Andre, B., and Collado-Vides, J. (1998). Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J Mol Biol*, 281(5):827–42.

Vilo, J., Brazma, A., Jonassen, I., Robinson, A., and Ukkonen, E. (2000). Mining for putative regulatory elements in the yeast genome using gene expression data [In Process Citation]. *ISMB*, 8:384–94.

Vingron, M. and Hoheisel, J. (1999). Computational aspects of expression data. *J. Mol. Med.*, 77:3–7.

Visconti, P. E., Johnson, L. R., Oyaski, M., Fornes, M., Moss, S. B., Gerton, G. L., and Kopf, G. S. (1997). Regulation, localization, and anchoring of protein kinase A subunits during mouse sperm capacitation. *Dev Biol*, 192(2):351–63.

von Stein, O. D., Thies, W. G., and Hofmann, M. (1997). A high throughput screening for rarely transcribed differentially expressed genes. *Nucleic Acids Res*, 25(13):2598–602.

Wegener, I. (1999). *Theoretische Informatik*. Teubner, Stgt., 2nd edition edition.

West, M., Nevins, J. R., Marks, J. R., Spang, R., Blanchette, C., and Zuzan, H. (2000). DNA microarray data analysis and regression modeling for genetic expression profiling. Technical Report 00-15, Institute of Statistics, Duke University, Durham, North Carolina.

Wingender, E., Chen, X., Fricke, E., Geffers, R., Hehl, R., Liebich, I., Krull, M., Matys, V., Michael, H., Ohnhauser, R., Pruss, M., Schacherer, F., Thiele, S., and Urbach, S. (2001). The TRANSFAC system on gene expression regulation. *Nucleic Acids Res*, 29(1):281–3.

Yin, J. C., Wallach, J. S., Wilder, E. L., Klingensmith, J., Dang, D., Perrimon, N., Zhou, H., Tully, T., and Quinn, W. G. (1995). A Drosophila CREB/CREM homolog encodes multiple isoforms, including a cyclic AMP-dependent protein kinase-responsive transcriptional activator and antagonist. *Mol Cell Biol*, 15(9):5123–30.

Zadeh, L. (1965). Fuzzy sets. *Information and Control*, 8(3):338–353.

Zhou, Y., Sun, Z., Means, A. R., Sassone-Corsi, P., and Bernstein, K. E. (1996). cAMP-response element modulator tau is a positive regulator of testis angiotensin converting enzyme transcription. *Proc Natl Acad Sci U S A*, 93(22):12262–6.

Appendix A: Homologies to known Genes in the CREM SSH library

The clones in the CREM SSH library contain 956 unique sequences (*Rsa*I−fragments). These sequences were analyzed by searching for homologies to known genes in the **EMBL** and **SwissProt** databases. Homologous ESTs are summarized using the GeneNest indices. Database searches were performed using the BLAST2 software, the sequences of the found genes together with the sequences of the *Rsa*I−fragments were assembled using the STADEN software.

The detected homologies to known genes are summarized in the table below. They fall into the categories **Known Mouse Genes** (indicated with yellow background) and **Homologous Other Genes** (indicated with blue background). The number of matching *Rsa*I−fragments is indicated in the column labeled **#**. The found genes were functionally classified, the list is sorted by the functional category.

Database Title	Title	Functional Category	#
DPY3_MOUSE	Dihydropyrimidinase−like 3 (Ulip)	Axon Guidance	2
SM4C_MOUSE	SEMAPHORIN 4C PRECURSOR (SM4C)	Axon Guidance	1
GNRP_MOUSE	RAS protein−specific guanine nucleotide−releasing factor 1 (GNRP)	Cell Cycle Regulator	1
RCC_HUMAN	Human REGULATOR OF CHROMOSOME CONDENSATION (RCC1)	Cell Cycle Regulator	1
SKP2	Homo sapiens cyclin AVCDK2−associated p45 (Skp2)	Cell Cycle Regulator	2
RAD21	DOUBLE−STRAND−BREAK REPAIR PROTEIN RAD21 HOMOLOG (PW29)	Cell Cycle Regulator	2
PLAK_MOUSE	Plakoglobin (PLAK)	Cell Junction	1
SPK_HUMAN	Homo sapiens Symplekin	Cell Junction	1
HIAT1	Tetracycline transporter−like protein (Hiat1)	Crossmembrane Transport	1
SGRP23	Storage granule protein 23 (SGRP23)	Crossmembrane Transport	1
KIAA0245	Human KIAA0245 gene	Crossmembrane Transport	1
mCAT2	Cationic amino acid transporter (mCAT2)	Crossmembrane Transport	4
AF155660	Homo sapiens mitochondrial solute carrier	Crossmembrane Transport	1
ATP1A4	Sodium/Potassium−Transporting ATPase alpha 4 subtunit (ATP1A4)	Crossmembrane Transport	3
ATP1B3	Sodium/Potassium−Transporting ATPase beta 3 subtunit (ATP1B3)	Crossmembrane Transport	3
SGCA	Alpha−sarcoglycan gene	Cytoskeleton+Motility	3
VimC1	Vimentin−binding fragment (VimC1)	Cytoskeleton+Motility	1
WDR1_MOUSE	WD−REPEAT PROTEIN 1 (Wdr1)	Cytoskeleton+Motility	1
tACTIN2	Testis specific actin (tActin2)	Cytoskeleton+Motility	5
GELS_MOUSE	Actin−Depolymerizing factor Gelsolin	Cytoskeleton+Motility	1
MM17324	N−retinoic acid−regulated gene\/profilinII homolog	Cytoskeleton+Motility	1
DYL1_HUMAN	Human cytoplasmic dynein light chain 1 (DYL1)	Cytoskeleton+Motility	1
DYLX_MOUSE	CYTOPLASMIC DYNEIN LIGHT CHAIN (TCTEX−1)	Cytoskeleton+Motility	1
LAP1C	Rattus norvegicus lamina−associated polypeptide 1C (LAP1C)	Cytoskeleton+Motility	3
KIF9	Kinesin like protein 9 (Kif9)	Cytoskeleton+Motility	1
IPP	Homo sapiens actin−binding protein (IPP)	Cytoskeleton+Motility	1
ACTG2	Smooth muscle gamma actin (ACTG2)	Cytoskeleton+Motility	2
TBA1_MOUSE	Alpha−tubulin gene (M−alpha−1)	Cytoskeleton+Motility	3
TBB5_Mouse	Tubulin beta−5 chain	Cytoskeleton+Motility	1
MSH3_MOUSE	DNA MISMATCH REPAIR PROTEIN (MSH3)	DNA Repair	1
Tankyrase	Homo sapiens TRF1−interacting ankyrin−related ADP−ribose polymerase (Tankyrase)	DNA Repair	1
DBI5_MOUSE	DIAZEPAM BINDING INHIBITOR−LIKE 5	Energy Metabolism	3
KCRB_HUMAN	Homo sapiens creatine kinase B (KCRB)	Energy Transduction	1
AF044312	Erythrocyte protein band 4.1−like 2	Erythrocyte Membrane	1
H2A1_MOUSE	Histone H2A.1	Histones+HMGs	1
HMG1_MOUSE	High mobility group 1 protein (HMG−1)	Histones+HMGs	1
BH5_MOUSE	PEANUT−LIKE PROTEIN 2 (BH5)	Histones+HMGs	2
H3A	Histone H3.3A	Histones+HMGs	1
RAB6	GTP BINDING PROTEIN ASSOCIATED PROTEIN 1 (Rab6)	Intracellular Transport	2
AF039023	Homo sapiens Ran−GTP binding protein	Intracellular Transport	1
RAN_MOUSE	GTP−BINDING NUCLEAR PROTEIN (RAN)	Intracellular Transport	3
Ap−4	Homo sapiens AP−4 adaptor complex beta4 subunit	Intracellular Transport	1

Appendix A

Database Title	Title	Functional Category	#
PXR2a	Homo sapiens peroxisomal targeting signal 1 receptor−like Gene, PXR2a	Intracellular Transport	1
AP47_MOUSE	Clathrin−associated protein (AP47)	Intracellular Transport	1
AP50_HUMAN	Clathrin−associated AP−2 complex AP50 subunit	Intracellular Transport	1
SRPR_HUMAN	Signal recognition particlereceptor alpha subunit (SRPR)	Intracellular Transport	1
Syntaxin−16A	Homo sapiens syntaxin−16A	Intracellular Transport	1
SCA1_HUMAN	Homo sapiens secretory carrier membrane protein (SCAMP1)	Intracellular Transport	3
E1−like	Homo sapiens E1−like protein	Intracellular Transport	1
COPE_MOUSE	COATOMER EPSILON SUBUNIT (COPE)	Intracellular Transport	1
KAP3A	KINESIN−ASSOCIATED PROTEIN 3 (KAP3)	Intracellular Transport	1
STAG3	nuclear protein stromal antigen 3 (SA3)	Meiosis	1
C11A_HUMAN	Human cholesterol side−chain cleavage enzyme P450sccK	Metabolic Enzymes	1
CAOQ_RAT	Rattus norvegicus Pristanoyl−CoA Oxidase	Metabolic Enzymes	2
DHQV_HUMAN	Human quinone oxidoreductase (NQO2)	Metabolic Enzymes	1
F26H_HUMAN	Homo sapiens 6−phosphofructo−2−kinase\/fructose−2,6−bisphosphatase (F26H)	Metabolic Enzymes	1
FPPS_HUMAN	Human farnesyl pyrophosphate synthetase (FPPS)	Metabolic Enzymes	1
FTDH_RAT	Homo sapiens 10−formyltetrahydrofolate dehydrogenase (FTDH).	Metabolic Enzymes	1
GLNA_MOUSE	Glutamate−ammonia ligase (GLNA)	Metabolic Enzymes	4
GLPK_MOUSE	Mus musculus glycerol kinase (Gyk)	Metabolic Enzymes	1
GPDM_MOUSE	Glycerol−3−phosphate dehydrogenase (GDPM)	Metabolic Enzymes	1
GPX4	Phospholipid hydroperoxide glutathione peroxidase (GPX4)	Metabolic Enzymes	2
HO2_MOUSE	Heme oxygenase 2a (HO−2a)	Metabolic Enzymes	1
KDGH_MESAU	Cricetinae gen. sp. diacylglycerol kinase delta	Metabolic Enzymes	1
KPR1_HUMAN	Homo sapiens phosphoribosyl pyrophosphate synthetasesubunit I	Metabolic Enzymes	1
ODO1_HUMAN	Human 2−oxoglutarate dehydrogenase	Metabolic Enzymes	1
ODPT_MOUSE	Pyruvate dehydrogenase (pdha−2)	Metabolic Enzymes	1
PLCB_HUMAN	Homo sapiens lysophosphatidic acid acyltransferase	Metabolic Enzymes	3
SERA_MOUSE	Rattus norvegicus D−3−phosphoglycerate dehydrogenase	Metabolic Enzymes	1
ATPA_MOUSE	ATP synthase alpha subunit (ATPA)	Metabolic Enzymes	1
ATPO_HUMAN	Homo sapiens ATP synthase	Metabolic Enzymes	1
ATPR_MOUSE	Mitochondrial ATP synthase coupling factor 6 (ATPR)	Metabolic Enzymes	1
ASSY_MOUSE	Argininosuccinate synthetase (Ass)	Metabolic Enzymes	2
LCFB_MOUSE	Long chain fatty acyl CoA synthetase (LCFB).	Metabolic Enzymes	5
F263_RAT	Rat testis fructose−6−phophate, 2−kinase:fructose−2, 6−bisphosphatase	Metabolic Enzymes	1
ALFA_MOUSE	Aldolase A	Metabolic Enzymes	2
G3PT_MOUSE	Testis−specific isoform of Glyceraldehyde 3−phosphatedehydrogenase (Gapd−S)	Metabolic Enzymes	3
G6PI_MOUSE	Glucose phosphate isomerase (G5PI)	Metabolic Enzymes	2
HXK1_MOUSE	Hexokinase	Metabolic Enzymes	2
KPY2_MOUSE	Pyrubate kinase M	Metabolic Enzymes	2
LDHM_MOUSE	Lactate dehydrogenase A−4 (LDH−A)	Metabolic Enzymes	1
PGMU_HUMAN	Human phosphoglucomutase 1 (PGM1)	Metabolic Enzymes	1
DPM2_RAT	DOLICHOL PHOSPHATE−MANNOSE BIOSYNTHESIS REGULATORY PROTEIN (DPM2)	Metabolic Enzymes	1
IMD2_MOUSE	INOSINE−5'−MONOPHOSPHATE DEHYDROGENASE 2 (IMD2)	Metabolic Enzymes	2
MAN2_MOUSE	Alpha−mannosidase II (MAN2)	Metabolic Enzymes	2
Aoh1	Aldehyde oxidase homolog−1 (Aoh1)	Metabolic Enzymes	1
HSU63743	Homo sapiens mitotic centromere−associated kinesin	Mitosis	2
BiP	IMMUNOGLOBULIN HEAVY CHAIN BINDING PROTEIN (BIP)	Molecular Chaperone	2
HS72_MOUSE	Heat−shock−like protein (HSP70.2)	Molecular Chaperone	2
HS7T_MOUSE	Spermatid−specific heat shock protein70 (Hsc70t)	Molecular Chaperone	3
HSJ3_MOUSE	Testis specific DNAj−homolog 3 (HSJ3)	Molecular Chaperone	2
MDJ6	HSP40/DNAJ homolog: MDJ6	Molecular Chaperone	3
RNDNAJ−Like	Rattus norvegicus dnaj−like protein	Molecular Chaperone	1
TCPG_MOUSE	chaperonin subunit 3 gamma (Cct3)	Molecular Chaperone	1
CATH_MOUSE	Mus musculus cathepsin H prepropeptide (ctsH)	Protein Degradation	1
HIP2	UBIQUITIN−CONJUGATING ENZYME E2 (HIP2)	Protein Degradation	1
UBPP_MOUSE	Ubiquitin−specific processing protease (Usp25)	Protein Degradation	1

Appendix A

Database Title	Title	Functional Category	#
MEPD_RAT	Rat metalloendopeptidase	Protein Degradation	1
CATD_MOUSE	Cathepsin D	Protein Degradation	1
HPI31	Homo sapiens proteasome inhibitor hPI31	Protein Degradation	2
CRES_MOUSE	Cystatin−related epididymal spermatogenic protein (Cres)	Protein Degradation	1
ICAL_HUMAN	Homo sapiens testis calpastatin	Protein Degradation	1
ICAL_RAT	Rat calpastatin	Protein Degradation	1
ACRO_MOUSE	Acrosin gene	Protein Degradation	1
100K_RAT	Rattus norvegicus 100 kDa protein	Protein Degradation	1
PRS7_HUMAN	Human 26S PROTEASE REGULATORY SUBUNIT 7 (PRS7)	Protein Degradation	1
PSA6_HUMAN	Homo sapiens PROTEASOME SUBUNIT ALPHA TYPE 6 (PSMA6)	Protein Degradation	1
UB5B_HUMAN	Ubiquitin conjugating enzyme (ubc4)	Protein Degradation	2
UBA1_MOUSE	Ubiquitin activating enzyme E1 (UBA1)	Protein Degradation	2
UBC3_HUMAN	Homo sapiens ubiquitin conjugating enzyme (UBC3)	Protein Degradation	2
UBH1	Deubiquitinating enzyme (UBH1)	Protein Degradation	1
Ubp41	Ubiquitin−specific protease UBP41 (Ubp41)	Protein Degradation	2
WWP2	Homo sapiens Nedd−4−like ubiquitin−protein ligase (WWP2)	Protein Degradation	1
SIAT7	ALPHA−N−ACETYLGALACTOSAMINIDE ALPHA−2,6−SIALYLTRANSFERASE (Siat7b)	Protein Modification	1
POMT1	Homo sapiens protein O−mannosyl−transferase 1 (POMT1)	Protein Modification	2
MPPB_HUMAN	Homo sapiens mitochondrial processing peptidase beta−subunit (MPPB)	Protein Modification	1
TGLC_CHICK	Chicken transglutaminase	Protein Modification	2
SPC1_HUMAN	Homo sapiens MICROSOMAL SIGNAL PEPTIDASE 25	Protein Modification	1
S61A_CANFA	Canis familiaris PROTEIN TRANSPORT PROTEIN SEC61	Protein Transport	1
PTB	Homo sapiens neural polypyrimidine tract binding protein (PTB)	RNA Modification	1
DDX3_HUMAN	Homo sapiens dead box, X isoform (DBX)	RNA Modification	1
AF083383	Homo sapiens 38 kDa splicing factor	RNA Modification	1
ANX2_MOUSE	ANNEXIN II	Signal Transduction	1
FRT1_MOUSE	Proto−oncogene Frat1	Signal Transduction	1
HGS	HGF−regulated tyrosine kinase substrate	Signal Transduction	1
IFR2_HUMAN	Homo sapiens INTERFERON−RELATED DEVELOPMENTAL REGULATOR 2 (IFR2)	Signal Transduction	1
IQGA	IQ motif containing GTPase activating protein 1 (Iqgap1)	Signal Transduction	2
KICE_MOUSE	CholineVethanolamine kinase	Signal Transduction	1
KLK8_MOUSE	GLANDULAR KALLIKREIN K8	Signal Transduction	1
LFC_MOUSE	lymphoid blast crisis−like 1 (LFC)	Signal Transduction	1
MEG1_MOUSE	MEIOSIS EXPRESSED PROTEIN 1 (MEG1)	Signal Transduction	1
NAP4	Homo sapiens Nck, Ash and phospholipase C gamma−bindingprotein (NAP4)	Signal Transduction	1
PEBP_MOUSE	Phosphatidylethanolamine binding protein (PEBP)	Signal Transduction	1
RETL2	Rattus norvegicus GDNF−Receptor beta (RETL2)	Signal Transduction	1
SRPK2	SERINE/ARGININE−RICH PROTEIN SPECIFIC KINASE 2 (SRPK2)	Signal Transduction	3
TSSKS1	Testis specific serine kinase substrate (TsskS1)	Signal Transduction	1
ACE_MOUSE	Angiotensin−converting enzyme (ACE)	Signal Transduction	3
HIPK1	Homeodomain−interacting protein kinase 1 (HIPK1)	Signal Transduction	3
CHIO_HUMAN	Homo sapiens beta2−chimaerin	Signal Transduction	4
RalGPS1A	Homo sapiens Ral guanine nucleotide exchange factor (RalGPS1A)	Signal Transduction	1
BIG2	Homo sapiens brefeldin A−inhibited guanine nucleotide−exchange protein 2	Signal Transduction	1
HIPK3	Homeodomain−interacting protein kinase 3	Signal Transduction	1
IDE_HUMAN	Insulin−degrading enzyme (IDE)	Signal Transduction	1
hook1	Homo sapiens hook1 protein	Signal Transduction	1
AF015811	Putative lysophosphatidic acid acyltransferase	Signal Transduction	1
GRINA	NMDA receptor glutamate−binding subunit (GRINA)	Signal Transduction	1
HPS1_HUMAN	Human pHS1−2 homologous to membrane receptor proteins	Signal Transduction	1
Nfkbia	Nuclear factor of kappa light chain gene enhancer in B−cells inhibitor (Nfkbia)	Signal Transduction	1
TDXN_MOUSE	THIOREDOXIN PEROXIDASE AO372	Signal Transduction	1
AKAP110	Protein kinase A binding protein AKAP110	Signal Transduction	3
D−AKAP1	Dual specificity A−kinase anchoring protein 1 (D−AKAP1)	Signal Transduction	4
KAP2_MOUSE	cAMP−dependent protein kinase type II (KAP2)	Signal Transduction	4

Appendix A

Database Title	Title	Functional Category	#
ATM	Human phosphatidylinositol 3−kinase homolog (ATM)	Signal Transduction	1
I5P1_HUMAN	Homo sapiens InsP3 5−phosphatase.	Signal Transduction	1
INPP_MOUSE	Inositol polyphosphate 1−phosphatase (INPP)	Signal Transduction	1
PDK1	Phosphoinositide−dependent protein kinase PDK1	Signal Transduction	1
pi4K230	Homo sapiens phosphatidylinositol 4−kinase 230 (pi4K230)	Signal Transduction	1
PTPLB	Protein tyrosine phosphatase−like protein PTPLB (Ptplb)	Signal Transduction	1
HU−PP−1	Homo sapiens PROTEIN−TYROSINE PHOSPHATASE (BM−008)	Signal Transduction	1
Dyrk1B	Homo sapiens protein kinase Dyrk1B	Signal Transduction	1
KC12_HUMAN	Homo sapiens casein kinase I gamma 2	Signal Transduction	8
KC2B_HUMAN	Casein kinase II beta subunit (KC2B)	Signal Transduction	1
KPCD_MOUSE	Protein kinase C delta (KPCD)	Signal Transduction	5
KPT1_MOUSE	SERINE/THREONINE−PROTEIN KINASE PCTAIRE−1	Signal Transduction	1
MAPK6	MITOGEN−ACTIVATED PROTEIN KINASE 6 (MAPK6)	Signal Transduction	1
Prkmk7	Mitogen−activated protein kinase kinase 7 (PRKMK7)	Signal Transduction	2
TESK1	Testis−specific protein kinase 1 (Tesk1)	Signal Transduction	1
PP1G_MOUSE	Protein phosphatase type 1 (PP1G)	Signal Transduction	1
ppx	Protein phosphatase X homolog (PPX)	Signal Transduction	1
MmLATS2	Warts/lats−like kinase (MmLATS2)	Signal Transduction	2
SYNGAPA	Rattus norvegicus synaptic ras GTPase−activating protein	Signal Transduction	2
MMP40GPRT	G protein−coupled receptor (P40GPRT)	Signal Transduction	1
GRIP	Rattus norvegicus AMPA receptor interacting protein GRIP	Signal Transmission	1
NTTA_MOUSE	Retinal taurine transporter (mTAUT)	Signal Transmission	1
RNU67140	Rattus norvegicus PSD−95\/SAP90−associated protein−4	Signal Transmission	1
SNAA_HUMAN	Homo sapiens ALPHA−SOLUBLE NSF ATTACHMENT PROTEIN (SNAP−ALPHA)	Signal Transmission	1
CCBB_HUMAN	Human neuronal DHP−sensitive, voltage−dependent, calcium channel beta−2 subunit (CCBB)	Signal Transmission	1
CCAG_RAT	Rattus norvegicus low voltage T−type calcium channel alpha−1g subunit (CCAG)	Signal Transmission	1
POR1_MOUSE	Voltage dependent anion channel 1 (POR1)	Signal Transmission	1
DDC8	Testis−specific protein (DDC8)	Sperm Structure	1
Gsg3	F−ACTIN CAPPING PROTEIN ALPHA−3 (Gsg3)	Sperm Structure	4
CALI_HUMAN	Homo sapiens Calicin (CALI)	Sperm Structure	4
HSP2_MOUSE	Protamine 2 (mP2)	Sperm Structure	3
STP1_MOUSE	SPERMATID NUCLEAR TRANSITION PROTEIN 1 (STP−1)	Sperm Structure	1
STP2_MOUSE	Transition protein 2 (TP2)	Sperm Structure	1
FSC1_MOUSE	Major fibrous sheath protein (FSC1, AKAP82)	Sperm Structure	5
FibrousheathinII	Homo sapiens fibrousheathin II	Sperm Structure	1
ODFP_MOUSE	Outer dense fiber protein of sperm tails (Odf1)	Sperm Structure	4
Odf2	Outer dense fiber protein (Odf2)	Sperm Structure	2
ADAM4	A disintegrin and metalloprotease domain 4 (ADAM4)	Sperm Structure	2
ADAM5	A disintegrin and metalloprotease domain 5 (ADAM 5)	Sperm Structure	1
TPX1	Mouse testis−specific protein (TPX−1)	Sperm Structure	1
TCX2_MOUSE	T COMPLEX TESTIS−SPECIFIC PROTEIN 2 (TCTEX−2)	Sperm Structure	1
Mlark	RNA BINDING MOTIF PROTEIN 4 (Mlark)	Transcription Factor	4
OSF−6	Oxidative stress−induced protein (OSF−6)	Transcription Factor	1
PLAGL2	Homo sapiens zinc finger protein PLAGL2 (PLAGL2)	Transcription Factor	1
Rnf4	RING FINGER PROTEIN 4 (RNF4)	Transcription Factor	1
SOX5_MOUSE	Testis Sox−5	Transcription Factor	1
SOX6_MOUSE	TRANSCRIPTION FACTOR SOX−LZ	Transcription Factor	1
STA4_MOUSE	Mus musculus BALB\/c gamma interferon activation site−bindingprotein (STAT4)	Transcription Factor	1
Tctex−3	Mus musculus Tctex−3 mRNA, complete cds.	Transcription Factor	1
Tex27	Tex27	Transcription Factor	2
YB1_MOUSE	Y−box binding protein	Transcription Factor	1
ZN76_HUMAN	Human zinc−finger protein 76	Transcription Factor	1
Zik1	Zinc finger protein interacting with K protein 1 (Zik1)	Transcription Factor	1
MOF	Homo sapiens histone acetyltransferase (MOF)	Transcription Factor	1
UBQLN3	Homo sapiens ubiquilin 3 (UBQLN3)	Transcription Factor	2

Appendix A

Database Title	Title	Functional Category	#
RTR	RETINOID RECEPTOR−RELATED TESTIS SPECIFIC RECEPTOR (RTR)	Transcription Factor	3
CNBP_MOUSE	CELLULAR NUCLEIC ACID BINDING PROTEIN (CNBP)	Transcription Factor	2
T2FA_HUMAN	TRANSCRIPTION INITIATION FACTOR RAP74 (T2FA)	Transcription Factor	1
TF2D_MOUSE	TRANSCRIPTION INITIATION FACTOR TFIID	Transcription Factor	1
28SRNA	Homo sapiens 28S ribosomal RNA gene	Translation	2
MMRNA1	45S ribosomal RNA.	Translation	1
RL28_MOUSE	Ribosomal protein L28	Translation	1
RL38	Rat ribosomal protein L38	Translation	1
RL8_HUMAN	Ribosomal protein L8 (RPL8)	Translation	1
RLA0_MOUSE	Acidic ribosomal phosophoprotein PO	Translation	1
RS17_CRIGR	Ribosomal protein S17	Translation	1
RS24_HUMAN	Ribosomal protein S24	Translation	2
Rps29	Ribosomal protein S29	Translation	1
UbA52	Ubiquitin/60S ribosomal fusion protein	Translation	1
EIF−4G	Homo sapiens eukaryotic protein synthesis initiation factor (EIF−4G)	Translation	1
EF1G_HUMAN	Homo sapiens elongation factor 1−gamma homolog	Translation	2
IF33_HUMAN	Homo sapiens translation initiation factor eIF3 p40 subunit	Translation	1
SYC_HUMAN	Human cysteinyl−tRNA synthetase	Translation	1
EB2	APC−binding protein (EB2)	Viral Protein	1
ENV1_MOUSE	Repeat. Endogenous murine leukemia virus clone E1.	Viral Protein	1
MDMHTLE2	Repeat. MHC TL−associated endogenous retrovirus TLev1	Viral Protein	1
PRTC_MOUSE	Repeat. Anticoagulant protein C gene (PRTC)	Blood Coagulation	1
AF1Q_MOUSE	AF1Q gene	?	4
Ariadne	Ariadne protein	?	1
CLUS_MOUSE	SULFATED GLYCOPROTEIN 2 (Clusterin)	?	1
Cdyl	Testis−specific chromodomain Y−like protein (Cdyl)	?	1
DAL1	Homo sapiens putative lung tumor suppressor (DAL1)	?	1
DS1_HUMAN	Homo sapiens ICT1 (alias DS−1)	?	1
Dp1l1	Polyposis locus protein 1−like 1	?	1
Fhl4	LIM−protein (Fhl4)	?	1
G100_HUMAN	Human Mr 110,000 antigen	?	1
Gcap3	Granule cell antiserum positive 3 (Gcap3)	?	1
HS06631	Human H326	?	1
HS19878	Human transmembrane protein (HS19878)	?	1
Herp	Homo sapiens stress protein Herp	?	1
IDN3	Homo sapiens IDN3 gene	?	1
LAS1_MOUSE	LIM AND SH3 DOMAIN PROTEIN 1 (LASP−1)	?	1
MEN1_MOUSE	Menin (Men1) gene	?	1
MMLEUPS	Mouse leukosialin pseudogene (CD 43)	?	1
NIPIL	NIPI−like protein (NIPIL)	?	1
NY−CO−7	Homo sapiens antigen NY−CO−7 (NY−CO−7)	?	1
PLRG1	Pleiotropic regulator 1 (PLRG1)	?	1
RNTMDCIV	Rattus norvegicus tMDC IV protein	?	3
SKD3_MOUSE	Suppressor of K+ transport defect 3 (SKD3)	?	1
SUR4_MOUSE	Surfeit locus protein 4 (SUR4)	?	2
UNR_RAT	Rat unr protein with unknown function	?	1
VMD2_MOUSE	Bestrophin homolog (VMD2)	?	2
XRP2	Homo sapiens etinitis pigmentosa 2 protein (XRP2)	?	1
cp151	Rattus norvegicus cp151	?	1
Sperizin	Spermatid−specific ring zinc finger (Sperizin)	?	1

Appendix B: Summary of expression measurements in wild−type and CREM −/− testes measured on Affymetrix oligonucleotide arrays

The expression levels of mRNAs expressed in CREM (−/−) testes and in wild−type testes were measured each by 3 independent reproductions on **Affymetrix** oligonucleotide microarrays containing about 10,000 mouse genes (array *U74A*). For each reproduction the mRNAs of the testes of 3−4 mice were pooled.

The summarized results are shown in the list below. Results of measurements are grouped in one cell, if there are several representatives for one gene based on the accession number of a corresponding EST cluster from the GeneNest database. The list is divided into several sub−tables, which were derived from categories that were assigned based on the comparison of the results from the Affymetrix arrays and the cDNA nylon arrays. Each sub−table is sorted by the maximal fold difference of the gene measured on the Affymetrix array.

The columns of the table contain the following information:

1. *EST cluster*: Database accession number from the GeneNest database.
2. *Affymetrix ID*: Affymetrix identifier. The text color of the Affymetrix ID indicates the expression on the Affymetrix array as well as on the CREM SSH array.
   - Green: Genes found to be differential in the CREM SSH library, as well as on the Affymetrix array.
   - Light green: Genes found in the CREM SSH library and differential on the Affymetrix array, but with no measurement or not differential on the nylon array.
   - Blue: Genes that are found to be differentially expressed on the Affymetrix array: which are not in the CREM SSH library, are in the regions of low density in the plot, their expression level is higher than 400, and the difference of expression is more than a factor of 2.
   - Red: Sequences on the Affymetrix array, that are homologous to sequences in the CREM SSH library, which appear to be equally expressed.
   - Light Red: Sequences, which have an insecure expression measurement on the Affymetrix array, as well as on nylon arrays constructed from clones of the CREM SSH library.
   - Pink: Sequences, which were found in the CREM SSH library, that appear non differentially expressed on the Affymetrix arrays, but differentially expressed on the nylon arrays.
3. *Affy. Diff.* summarizes the measured factors of difference of the expression levels between CREM−deficient mice and wild−type mice.
   The median fold change is highlighted in false color representation. The numbers, which are printed with yellow text color, represent measurements where both values are in the same range as background. Numbers printed in green represent measurements, which are in the range of background in one condition and expressed in the other. Their absolute value is not reliable.
4. *SSH ID*: This column has an entry, if the gene is present in the CREM SSH library.
5. *Nylon Diff.*: For the genes, present in the SSH library, the ratios between CREM (−/−) testes and wild−type testes are given, that were measured on a nylon cDNA array for *Rsa*I−fragments of that gene.
   See description for column 'Affy. Diff.'.
6. *Title*: A short description of the gene is provided in the *Title* column.
7. *Description*: Short description on the function of the gene, mainly derived from the entries in the SwissProt database or from literature.

EST cluster	Affymetrix ID	Affy. Diff.	SSH ID	Nylon Diff.	Title	Description
Differential expression measured with Affymetrix chips for genes found by SSH (factor > 2)						
MM_TBI2107.28	101534_at	47088	STP2_MOUSE	65.2	M60254:transition protein 2 (TP2) gene	STP2_MOUSE. function: in the elongating spermatids of mammals, the conversion of nucleosomal chromatin to the compact, nonnucleosomal form found in the sperm nucleus is associated with the appearance of a small set of basic chromosomal transition proteins. subcellular location: nuclear. similarity: strong, to other mammalian spermatid nuclear transition proteins 2. DNA compaction
MM_TBI3005.1	102945_at 102946_r_at	145 34508	G3PT_MOUSE	26.5 8.0 12.2 2.0 0.8	U09964:ICR/Swiss glyceraldehyde 3−phosphate dehydrogenase (Gapd−S) gene	G3PT_MOUSE. function: may play an important role in regulating the switch between different pathways for energy production during spermiogenesis and in the spermatozoon. catalytic activity: d−glyceraldehyde 3−phosphate + orthophosphate + nad(+) = 1,3−diphosphateglycerate + nadh. pathway: first step in the second phase of glycolysis. subunit: homotetramer (by similarity). subcellular location: cytoplasmic (by similarity). tissue specificity: testis−specific. developmental stage: first expressed at day 20 in post−meiotic germ cells. levels increase until day 24 and then remain constant during maturity. similarity: belongs to the glyceraldehyde 3−phosphate dehydrogenase family. Glycolysis and Gluconeogenesis
MM_TBI4286.2	97367_at 97368_at 97369_g_at	9.77 6.45 30863	D−AKAP1	6.5 32.6 2.5 3.3	U95145:S−AKAP84 mRNA	PKA pathway

Appendix B

EST cluster	Affymetrix ID	Affy. Diff.	SSH ID	Nylon Diff.	Title	Description
MM_TBI3455.1	102857_at	19295	FSC1_MOUSE	1.2 8.2 61.9 2.7 2.8 26.5 5.8	U10341:Fibrous sheath component 1	AKA4_MOUSE. function: major structural component of sperm fibrous sheath. tissue specificity: spermatid. developmental stage: post−meiotic phase of spermatogenesis. caution: it is uncertain whether met−1 or met−10 is the initiator. 1. PKA pathway
MM_TBI3697.1	93942_at 95197_f_at	1459 0.14	INPP_MOUSE	0.1	U27295:Inositol polyphosphate−1−phosphatase	INPP_MOUSE. catalytic activity: d−myo−inositol 1,4−bisphosphate + h(2)o = d−myo−inositol 4−phosphate + orthophosphate. enzyme regulation: inhibited by li(+) (by similarity). pathway: phosphatidyl inositol signaling pathway. subunit: monomer (by similarity). similarity: belongs to the inositol monophosphatase family.
MM_TBI24369.1	103632_at	1364	Mm_TBI24369.1	36.5 8.7	AI036958: cDNA	
MM_TBI2813.1	92448_s_at 92449_at	3.58 1028	RETL2		AF002701:Glial cell line derived neurotrophic factor family receptor alpha 2	NRTR_MOUSE. function: receptor for neurturin. mediates the nrtn−induced autophosphorylation and activation of the ret receptor. also able to mediate gdnf signaling through the ret tyrosine kinase receptor. subcellular location: attached to the membrane by a gpi−anchor (by similarity). alternative products: 2 isoforms; a long form (shown here) and a short form; are produced by alternative splicing. tissue specificity: neurons of the superior cervical and dorsal root ganglia, and adult brain and testis. low level in the spleen and in the adrenal. similarity: belongs to the gdnfr family. Receptor for neurturin (NTN) is potent trophic factors for motoneurons. Activating the RET tyrosine kinase in the presence of Gfra2.
MM_TBI3717.1	93631_at 93632_g_at	799 10.6	LFC_MOUSE	4.1	X95761:new−Rhobin	LFC_MOUSE. tissue specificity: ubiquitous, with the exception of liver tissue. levels are high in hemopoietic tissues (thymus, spleen, bone marrow) as well as in kidney and lung. similarity: contains 1 dbl−homology domain (dh). similarity: contains 1 ph domain. similarity: contains 1 zinc−dependent phorbol−ester and dag binding domain. similarity: to human nucleotide exchange protein lbc. Guanine nucleotide exchange factor for Rac Lfc localizes to microtubules and mediates the activation of Rac signaling, JNK activation and actin cytoskeletal changes (J Biol Chem 1999 Jan 22;274(4):2279−85).
MM_TBI5267.1	102301_at	184	Gsg3	18.2 4.9	AB026984:Gsg3 gene for actin capping protein alpha	CAZ3_MOUSE. function: f−actin capping proteins bind in a ca(2+)−independent manner to the fast growing ends of actin filaments (barbed end) thereby blocking the exchange of subunits at these ends. unlike other capping proteins (such as gelsolin and severin), these proteins do not sever actin filaments. may play a role in the morphogenesis of spermatid. subunit: heterodimer of an alpha and a beta subunit (by similarity). tissue specificity: exclusively expressed in the testis. developmental stage: expressed in 24−day−old and adult testis, but not in 4−, 10− and 16−day−old testis. similarity: belongs to the f−actin capping protein alpha subunit family. Actin−capping protein The transcript contains a putative cAMP−responsive motif (CREM) upstream of the initiation codon in the DNA sequence and is expressed postmeiotically, first appearing between 20 and 30 days of postnatal development accumulates asymmetrically in the cytoplasm of round spermatids coincident with the position of the developing acrosome, may have an important role in determining the final shape of mature sperm heads. (Mol Reprod Dev 1998 Jan;49(1):81−91).
MM_TBI9171.1	97096_at	99.2	KAP2_MOUSE	9.0 25.5 23.5 8.9	J02935:cAMP−dependent protein kinase type II regulatory subunit mRNA	KAP2_MOUSE. function: type ii regulatory chains mediate membrane association by binding to anchoring proteins, including the map2 kinase. subunit: the inactive form of the enzyme is composed of two regulatory chains and two catalytic chains. activation by camp produces two active catalytic monomers and a regulatory dimer that binds four camp molecules. tissue specificity: four types of regulatory chains are found: i−alpha, i−beta, ii−alpha, and ii−beta. their expression varies among tissues and is in some cases constitutive and in others inducible. ptm: phosphorylated by the activated catalytic chain. similarity: contains 2 cyclic nucleotide−binding domains. similarity: belongs to the camp−dependent kinase regulatory chain family. PKA pathway
MM_TBI10943.1	103645_at	71.6	MDJ6	2.6 44.5 1.5 17.7	"AB028856: mDj6"	? Cloning paper: Cell Stress Chaperones 2000 Apr;5(2):98−112
MM_TBI667.1	92455_at	42.0	DDC8	9.1	Y09878:testis−specific protein, DDC8	? RNase protection assays indicate DDC8 to be expressed during the postmeiotic stages of spermatogenesis and database searches using both nucleotide and amino acid sequences show DDC8 to have similarities to structural, cytoskeletal and associated proteins (Mol Hum Reprod 1997 Mar;3(3):215−21).
MM_TBI117421.1	94852_at 98243_f_at 98244_r_at 99498_at	0.68 0.05 0.59 32.6	GLNA_MOUSE	4.3 5.4 22.0 7.8	M60803:Glutamine synthetase	
MM_TBI4407.1	103245_at	30.7	CRES_MOUSE	21.4	S49926:Cystatin related epididymal specific	CRES_MOUSE. function: performs a specialized role during sperm development and maturation. subcellular location: secreted. tissue specificity: proximal caput region of the epididymis. lower expression in the testis. within the testis it is localized to the elongating spermatids, whereas within the epididymis it is exclusively synthesized by the proximal caput epithelium. induction: testicular factors or hormones other than androgens present in the testicular fluid may be involved in the regulation of cres gene expression. similarity: belongs to the cystatin family. Cysteine protease inhibitor
MM_TBI930.1	95585_at 97046_f_at	27.2 0.86	PRTC_MOUSE	6.3	AF034569:anticoagulant protein C gene	PRTC_MOUSE. function: protein c is a vitamin k−dependent serine protease that regulates blood coagulation by inactivating factors va and viiia in the presence of calcium ions and phospholipids. catalytic activity: degradation of blood coagulation factors va and viiia. subunit: synthesized as a single chain precursor, which is cleaved into a light chain and a heavy chain held together by a disulfide bond. the enzyme is then activated by thrombin, which cleaves a tetradecapeptide from the amino end of the heavy chain; this reaction, which ours at the surface of endothelial cells, is strongly promoted by thrombomodulin. tissue specificity: plasma; synthesized in the liver. ptm: the vitamin k−dependent, enzymatic carboxylation of some glu residues allows the modified protein to bind calcium. miscellaneous: calcium also binds, with stronger affinity to another site, beyond the gla domain. this gla−independent binding site is necessary for the recognition of the thrombin−thrombomodulin complex. similarity: contains 2 egf−like domains. similarity: belongs to peptidase family also known as the trypsin family.

Appendix B

EST cluster	Affymetrix ID	Affy. Diff.	SSH ID	Nylon Diff.	Title	Description
MM_TBI1943.1	102714_at	21.8	HS7T_MOUSE	16.5 7.4 13.3 3.4	L27086:Heat shock protein cognate 70, testis	HS7T_MOUSE. tissue specificity: expressed in spermatids. developmental stage: specifically expressed in postmeiotic phases of spermatogenesis. similarity: belongs to the heat shock protein 70 family.
MM_TBI153.1	100624_i_at 100625_r_at	19.7 1.79	STP1_MOUSE	77.1 64.7	X12521:Transition protein 1	STP1_MOUSE. function: in the elongating spermatids of mammals, the conversion of nucleosomal chromatin to the compact, nonnucleosomal form found in the sperm nucleus is associated with the appearance of a small set of basic chromosomal transition proteins. subcellular location: nuclear. tissue specificity: testis. similarity: strong, to other mammalian spermatid nuclear transition proteins 1.
MM_TBI311.4	101245_f_at 94079_at	5.97 19.0	BH5_MOUSE	4.2	"X61452: H5"	Recent work suggests novel functions for septins in vesicle trafficking, oncogenesis and compartmentalization of the plasma membrane. Given the ability of the septins to bind GTP and phosphatidylinositol 4,5−bisphosphate in a mutually exclusive manner, these proteins might be crucial elements for the spatial and/or temporal control of diverse cellular functions. J Cell Sci 2001 Mar;114(Pt 5):839−44 GTP and phosphatidylinositol 4,5−bisphosphate binding protein function: involved in cytokinesis (potential).
MM_TBI18355.1	102937_at	16.2	Mm_TBI18355.1	4.4	AA636300: cDNA	
MM_TBI15244.1	99055_at	12.0	SIAT7		X94000:Gal beta−1,3−GalNAc−specific GalNAc alpha−2,6−sialyltransferase gene	PATHWAY: GLYCOSYLATION. Sialylation of glycoproteins and glycolipids The cytoplasmic droplet of epididymal spermatozoa contains the Golgi/TGN glycosylating activities (including alpha−2,6−sialyltransferase) in the saccules may berelated to plasma membrane modifications which occur during epididymal sperm maturation.J Cell Biol 1993 Nov;123(4):809−21
MM_TBI21228.1	96237_at	9.09	Mm_TBI21228.1	15.9	AI118905:uc15f04.r1 cDNA	
MM_TBI10954.1	98310_at	8.88	Sperizin	8.0	AB016984:Sperizin	? Transcription of the sperizin gene became detectable at day 23, exclusively in the round spermatid (Genomics 1999 Apr 1;57(1):94−101).
MM_TBI10132.1	98033_at	8.79	Mm_TBI6748.1	1.0 10.5	AA710132:v cDNA	
MM_TBI13432.1	96134_at	8.59	Dp1l1	5.9	AA755260: cDNA	
MM_TBI2299.1	100573_f_at 100574_f_at 95846_f_at	12.0 8.12 1.57	G6PI_MOUSE	1.7 10.4	L09104:Glucose phosphate isomerase 1 complex	G6PI_MOUSE. function: neuroleukin is a neurotrophic factor for spinal and sensory neurons. catalytic activity: glucose 6−phosphate = fructose 6−phosphate. pathway: involved in glycolysis and in gluconeogenesis. subunit: homodimer. subcellular location: cytoplasmic. similarity: belongs to the gpi family.
MM_TBI3481.1	99110_at 99111_at 99251_f_at	12.02 7.75 1.96	SKD3_MOUSE	2.8	U09874:SKD3 mRNA	SKD3_MOUSE. function: may function as a regulatory atpase and be related to secretion/protein trafficking process. tissue specificity: present in a wide variety of tissues, is abundant in mouse heart, skeletal muscle and kidney, and is most abundant in testis. similarity: contains 4 ank repeats. similarity: belongs to the clpa/clpb family.
MM_TBI1097.1	99036_s_at 99037_at	7.47 33.44	Prkmk7	1.0 1.8	AB005654:Mitogen−activated protein kinase kinase 7	
MM_TBI1855.4	100217_at 98984_f_at	2.62 7.32	GPDM_MOUSE	1.7	D50430:glycerol−3−phosphate dehydrogenase	GPDM_MOUSE. catalytic activity: sn−glycerol 3−phosphate + aeptor = glycerone phosphate + reduced aeptor. cofactor: fad. enzyme regulation: calcium−binding enhance the activity of the enzyme. pathway: glycerol metabolism. subcellular location: mitochondrial. inner−membrane. similarity: belongs to the fad−dependent glycerol−3−phosphate dehydrogenase family. similarity: contains 2 ef−hand calcium−binding domains. one of which seems to be non functional. Glycerol metabolism
MM_TBI5905.1	92733_at	7.28	ADAM4	4.3 5.1	U22058:A disintegrin and metalloprotease domain (ADAM) 4	Protease
MM_TBI2160.3	102026_s_at 102027_s_at	1.60 7.01	KICE_MOUSE	4.4	AB011000:choline/ethanolamine kinase	KICE_MOUSE. catalytic activity: atp + choline = adp + o−phosphocholine. catalytic activity: atp + ethanolamine = adp + o− phosphoethanolamine. tissue specificity: expressed ubiquitously with the highest level in testis. similarity: belongs to the choline/ethanolamine kinases family. Glycerolipid metabolism Choline/ethanolamine kinase expressed ubiquitously with the highest level in testis (Biochim Biophys Acta 1998 Jul 31;1393(1):179−85).
MM_TBI12464.1	99138_at	7.00	RCC_HUMAN	11.8	AA756292: cDNA	Regulator of chromosome condensation RCC1, a gene reportedly involved in regulating onset of mammalian chromosome condensation (Proc Natl Acad Sci U S A 1990 Nov;87(21):8617−21).
MM_TBI4905.1	94624_at	6.69	TSSKS1	4.3	AF025310:tssk−1 and tssk−2 kinase substrate	? 65−kD protein phosphorylated by both kinases.
MM_TBI18247.1	104136_at	6.33	RNU67140	3.7	AI840413: cDNA	
MM_TBI19001.1	103680_at 93757_at	6.20 1.09	SRPR_HUMAN	8.4	AA683850: cDNA	function: this integral membrane protein ensures, in conjunction with srp, the correct targeting of the nascent secretory proteins to the endoplasmic reticulum membrane system. targeting of nascent secretory proteins to endoplasmic reticulum
MM_TBI8299.1	95561_at	6.05	Mm_TBI8299.1	6.6	AW120867: cDNA	
MM_TBI480.1	101968_at	6.00	ODFP_MOUSE	31.3 33.7 18.6 2.8 30.9	X79446:Outer dense fiber of sperm tails 1	ODFP_MOUSE. function: component of the outer dense fibers (odf) of spermatozoa. odf are filamentous structures located on the outside of the axoneme in the midpiece and principal piece of the mammalian sperm tail and may help to maintain the passive elastic structures and elastic recoil of the sperm tail. domain: the c−terminal contains many c−x−p repeats.
MM_TBI16279.1	96284_at	5.91	KC12_HUMAN	2.7 7.2 7.9 10.0 3.0 13.7 14.4 1.9 4.5		Protein kinase

Appendix B

EST cluster	Affymetrix ID	Affy. Diff.	SSH ID	Nylon Diff.	Title	Description
MM_TBI3450.1	102994_at	5.81	STA4_MOUSE	1.3	U06923:Signal transducer and activator of transcription 4 (STAT4)	STA4_MOUSE. function: carries out a dual function: signal transduction and activation of transcription. subunit: forms a homodimer or a heterodimer with a related family member (by similarity). subcellular location: nuclear; translocated into the nucleus in response to phosphorylation. ptm: tyrosine phosphorylated. serine phosphorylation is also required for maximal transcriptional activity (by similarity). similarity: belongs to the stat family of transcription factors. similarity: contains 1 sh2 domain. Cytokines signaling pathways Stat4 mRNA has been detected in spleen, heart, brain, peripheral blood cells, and testis (Cytogenet Cell Genet 1997;77(3−4):207−10).
MM_TBI1863.8	103926_at 92843_r_at	5.67 0.28	EIF−4G		AV380793: cDNA	
MM_TBI18267.1	94048_at	5.65	UBC3_HUMAN	13.1 5.9		Function: catalyses the covalent attechment of UBIQUITIN to other proteins. Ubiquitin−protein ligase
MM_TBI1106.1	101857_at	5.57	SRPK2	4 12.0 4.5	AB006036:Serine/arginine−rich protein specific kinase 2 (SRPK2)	Protamine 1 phosphorylation SRPK2 is able to phosphorylate protamine 1 (Mech Dev. 2000 Dec;99(1−2):51−64.)
MM_TBI122004.17	103613_at 93115_at 98677_f_at	5.54 1.16 0.79	ALFA_MOUSE	2.5 1.8 1.5 1.1	AA144642: cDNA	
MM_TBI13448.1	96620_at	5.16	Mm_TBI13448.1	12.7	"D87325: Gsg1"	?
MM_TBI1591.5	100334_f_at 102693_f_at 94716_f_at	5.10 1.45 2.51	KLK8_MOUSE	3.4	M17962:Epidermal growth factor binding protein type C, kallikrein 1	KLK9_MOUSE. function: glandular kallikreins cleave met−lys and arg−ser bonds in kininogen to release lys−bradykinin. catalytic activity: preferential cleavage of arg−|−xaa bonds in small molecule substrates. highly selective action to release kallidin (lysyl−bradykinin) from kininogen involves hydrolysis of met−|−xaa or leu−|−xaa. similarity: belongs to peptidase family also known as the trypsin family. kallikrein subfamily. cleavage of kininogen to release lys−bradykinin
MM_TBI4076.1	99579_at	5.09	ATP1B3	9.5 3.6	U59761:ATPase, Na+/K+ beta 3 subunit	ATND_MOUSE. function: this is the non−catalytic component of the active enzyme, which catalyzes the hydrolysis of atp coupled with the exchange of na and k ions across the plasma membrane. the exact function of this glycoprotein is not known. subunit: composed of three subunits: alpha (catalytic), beta and gamma. subcellular location: type ii membrane protein. tissue specificity: widely expressed. miscellaneous: the beta subunit seems to be encoded by a multigene family. each different subunit may have specialized functions. similarity: belongs to the na+/k+ and h+ atpases beta chain family.
MM_TBI2716.1	101593_at 96072_at	4.89 1.15	LDHM_MOUSE	3.2	"Y00309:LDH−A"	
MM_TBI1402.1	102063_at	4.72	PDK1	4.1	AF079535:3−phosphoinositide dependent protein kinase−1	
MM_TBI4337.1	100546_at 100547_at	1.00 4.09	Mlark	1.1 6.9 3.8 4.5	"X94344: Neosin"	
MM_TBI2181.1	92547_at 92548_g_at	3.75 0.77	HIP2	3.0	"AB011081:huntingtin interacting protein−2"	MGI:Hip2.
MM_TBI2896.1	100578_at	3.70	IMD2_MOUSE	5.0 5.2	M33934: inosine−5−monophosphate dehydrogenase 2	IMD2_MOUSE. function: imp is the rate limiting enzyme in the de novo synthesis of guanine nucleotides and therefore is involved in the regulation of cell growth. it may also have a role in the development of malignancy and the growth progression of some tumors. catalytic activity: inosine 5''−phosphate + nad(+) + h(2)o = xanthosine 5''−phosphate + nadh. pathway: first reaction unique to gmp biosynthesis. subunit: homotetramer. similarity: to other eukaryotic and prokaryotic impdh and to gmp reductase. similarity: contains 2 cbs domains. Guanine synthesis
MM_TBI3071.1	104530_s_at 104531_at	192.3 1 3.66	KPCD_MOUSE KPCD_MOUSE_2	4.2 4.9 55.2 10.4 9.3	X60304:Protein kinase C, delta	KPCD_MOUSE. function: this is calcium−independent, phospholipid−dependent, serine− and threonine−specific enzyme. function: pkc is activated by diacylglycerol which in turn phosphorylates a range of cellular proteins. pkc also serves as the receptor for phorbol esters, a class of tumor promoters. similarity: contains 2 zinc−dependent phorbol−ester and dag binding domains. similarity: contains 1 c2 domain. similarity: belongs to the ser/thr family of protein kinases. pkc subfamily. PKC
MM_TBI116795.1	104682_at	3.64	TBA1_MOUSE	4.2 1.0 3.8	"AJ245923: alpha−tubulin 8"	Tubulin
MM_TBI6183.163	93087_r_at	3.63	UBPP_MOUSE			Ubiquitin pathway
MM_TBI1593.1	103188_f_at 92616_at	1.85 3.61	UBA1_MOUSE	1.7	"D10576: ubiquitin activating enzyme E1"	Ubiquitin pathway
MM_TBI12682.1	96952_at	3.37	PSA6_HUMAN		"U60288: proteasome subunit iota"	PSA6_MOUSE. function: the proteasome is a multicatalytic proteinase complexwhich is characterized by its ability to cleave peptides with arg,phe, tyr, leu, and glu adjacent to the leaving group at neutral orslightly basic ph. the proteasome has an atp−dependent proteolyticactivity.pathway: involved in an atp/ubiquitin−dependent non−lysosomalproteolytic pathway.subunit: the proteasome is composed of at least 15 non identicalsubunits which form a highly ordered ring−shaped structure.subcellular location: cytoplasmic and nuclear.similarity: belongs to peptidase family t1a; also known as theproteasome a−type family. ubiquitin pathway
MM_TBI16413.1	103908_at	2.77	Mm_TBI16413.1		AW121857: cDNA	
MM_TBI1791.1	96066_s_at 99289_f_at	2.65 0.98	KPY2_MOUSE	8.1 7.6	"X97047: M2−type pyruvate kinase"	
MM_TBI1075.1	94897_at	2.39	GPX4	3.2	"AF044056: phospholipid hydroperoxide glutathione peroxidase"	GSHH_MOUSE. function: could play a major role in protecting mammals from thetoxicity of ingested lipid hydroperoxides.catalytic activity: 2 glutathione + h(2)o(2) = oxidizedglutathione + 2 h(2)o.cofactor: selenocysteine. the active−site selenocysteine isencoded by the opal codon, uga.subcellular location: mitochondrial and cytoplasmic.alternative products: a single nuclear gene produces both formsby use of alternative initiation codons in the same reading frame.tissue specificity: present primarily in testis.similarity: belongs to the glutathione peroxidase family.
MM_TBI62.1	97480_f_at 97481_r_at	1.68 2.06	HSJ3_MOUSE	3.0 11.1 6.8	"U95607: testis specific DNAj−homolog"	

Appendix B

EST cluster	Affymetrix ID	Affy. Diff.	SSH ID	Nylon Diff.	Title	Description
MM_TBI2741.1	100626_at	2.02	Odf2	10.1 1.9	AF034105:outer dense fiber 2 (Odf2) mRNA	
MM_TBI17458.1	95402_at	2.17	SRPR_HUMAN	8.4		
MM_TBI5447.1	102302_at	0.41	Gsg3	18.2 4.9	AB026984:Gsg3 gene for actin capping protein alpha	CAZ3_MOUSE. function: f−actin capping proteins bind in a ca(2+)−independent manner to the fast growing ends of actin filaments (barbed end) thereby blocking the exchange of subunits at these ends. unlike other capping proteins (such as gelsolin and severin), these proteins do not sever actin filaments. may play a role in the morphogenesis of spermatid. subunit: heterodimer of an alpha and a beta subunit (by similarity). tissue specificity: exclusively expressed in the testis. developmental stage: expressed in 24−day−old and adult testis, but not in 4−, 10− and 16−day−old testis. similarity: belongs to the f−actin capping protein alpha subunit family.
Genes found by SSH with marginally differential expression measured with Affymetrix chips (factor < 2)						
MM_TBI11759.1	103603_at	689	Mm_TBI11759.1	2.9	AI841401: cDNA	
MM_TBI2156.1	101427_s_at 101428_at	537 0.35	ACE_MOUSE	4.4 29.3 1.9 3.1	J04946:Angiotensin converting enzyme	ACE_MOUSE. function: converts angiotensin i to angiotensin ii by release of the terminal his−leu, this results in an increase of the vasoconstrictor activity of angiotensin. catalytic activity: release of a c−terminal dipeptide, oligopeptide−|−xaa−xbb, when xaa is not pro, and xbb is neither asp nor glu. converts angiotensin i to angiotensin ii. cofactor: binds two zinc ions (by similarity). subcellular location: type i membrane protein. alternative products: the testicular angiotensin−converting enzyme is transcribed from the same gene as the somatic isoform, probably from an alternative start site. similarity: belongs to peptidase family m2 (zinc metalloprotease).
MM_TBI19171.1	95500_at	4.29	Mm_TBI19317.1_2			
MM_TBI107822.1	92695_at	4.05	FRT1_MOUSE	8.3	U58974:Frequently rearranged in advanced T−cell lymphomas	FRT1_MOUSE. function: may play a role in tumor progression and collaborate with pim1 and myc in lymphomagenesis. may bind gsk−3 and prevent gsk−3−dependent phosphorylation. tissue specificity: highly expressed in testis. lower level of expression in spleen, thymus and brain. developmental stage: expressed at low levels during embryonic development. disease: activation contributes to progression of mouse t−cell lymphomas. similarity: belongs to the gsk−3−binding protein family.
MM_TBI9264.1	104121_at	2.69	PLAK_MOUSE	5.6	M90365:Junction plakoglobin	PLAK_MOUSE. function: common junctional plaque protein. the membrane−associated plaques are architectural elements in an important strategic position to influence the arrangement and function of both the cytoskeleton and the cells within the tissue. the presence of plakoglobin in both the desmosomes and in the intermediate junctions suggests that it plays a central role in the structure and function of submembranous plaques. subunit: homodimer. subcellular location: cytoplasmic in a soluble and membrane− associated form. similarity: belongs to the beta−catenin family. similarity: contains at least 9 arm repeats.
MM_TBI14258.2	94507_at	2.62	LCFB_MOUSE	1.4 7.2 5.1 43.6 3.2	"U15977:long chain fatty acyl CoA synthetase"	
MM_TBI17635.148	93764_at	2.40	Mm_TBI22271.1	3.8		
MM_TBI13307.2	101221_at	2.34	Mm_TBI13307.1	11.0	C76746: cDNA	
MM_TBI12956.1	100224_f_at 100651_f_at 99098_at	2.17 0.89 1.92	FPPS_HUMAN	8.6	AV371705: cDNA	
MM_TBI42596.1	98311_at	2.12	cp151	2.5	AB029919:STAP sperm tail associated protein	
MM_TBI1053.2	101500_at	2.10	AF044312		"AF044312:protein 4.1G"	
MM_TBI23464.1	98892_at	1.97	Y188_HUMAN	5.9 0.4	"AF180471:Kiaa0188"	
MM_TBI651.4	100590_at 100591_g_at 100592_at	1.56 1.00 1.95	DERP2	7.1 5.9	AI929971:ul60a06.y1 cDNA	
MM_TBI9601.3	95057_at	1.93	Herp			
MM_TBI108749.1	98887_at	1.92	SNAA_HUMAN	3.0		
MM_TBI1571.1	93207_at	1.85	ACRO_MOUSE	2.8	D00754:Preproacrosin	ACRO_MOUSE. function: acrosin is the major protease of mammalian spermatozoa. it is a serine protease of trypsin−like cleavage specificity, it is synthesized in a zymogen form, proacrosin and stored in the acrosome. catalytic activity: hydrolysis of arg− and lys−bonds; preferential cleavage arg−xaa >> lys−lys >> lys−xaa. subunit: heavy chain (catalytic) and a light chain linked by two disulfide bonds. similarity: belongs to peptidase family also known as the trypsin family.
MM_TBI7735.1	95497_at	1.79	Mm_TBI7735.1	5.3		
MM_TBI1369.1	92821_at	1.77	Ubp41	2.7 3.1	"AF079565:ubiquitin−specific protease UBP41"	UBP2_MOUSE. catalytic activity: ubiquitin c−terminal thiolester + h(2)o =ubiquitin + a thiol.similarity: belongs to peptidase family also known asfamily 2 of ubiquitin carboxyl−terminal hydrolases.
MM_TBI119037.1	102584_at	1.74	Dyrk1B	5.8	Y18280:protein kinase Dyrk1B	DYRB_MOUSE. subcellular location: nuclear. ptm: phosphorylated by map kinase (by similarity). similarity: belongs to the ser/thr family of protein kinases. mnb/dyrk subfamily.
MM_TBI10948.1	101864_at	1.73	Mm_TBI10948.1	3.5	"AF113520:actin−like−7−beta protein"	
MM_TBI11379.1	98937_at	1.70	Mm_TBI11379.1			
MM_TBI3295.1	97294_at	1.68	Mm_TBI6503.1	2.9		
MM_TBI15648.6	98064_at	1.64	Mm_TBI15648.1	7.3	AW125378: cDNA	
MM_TBI3107.1	99542_at	1.63	ODPT_MOUSE	6.8	M76728:Pyruvate dehydrogenase E1alpha−like	ODPT_MOUSE. function: the pyruvate dehydrogenase complex catalyzes the overall conversion of pyruvate to acetyl−coa & co(2). it contains multiple copies of three enzymatic components: pyruvate dehydrogenase (e1), dihydrolipoamide acetyltransferase (e2) & lipoamide dehydrogenase (e3). catalytic activity: pyruvate + lipoamide = s−acetyl−dihydro− lipoamide + co(2). cofactor: thiamine pyrophosphate. enzyme regulation: e1 activity is regulated by phosphorylation (inactivation) and dephosphorylation (activation) of the alpha subunit. subunit: tetramer of two alpha and two beta subunits. tissue specificity: testis.

Appendix B

EST cluster	Affymetrix ID	Affy. Diff.	SSH ID	Nylon Diff.	Title	Description
MM_TBI1020.1	102033_at	1.62	TESK1	14.3	AB003494:Testis specific protein kinase 1	
MM_TBI1541.2	95563_at	1.60	Ariadne	2.1	AJ130977:Ariadne protein, partial	
MM_TBI1578.1	100569_at	1.60	ANX2_MOUSE	48.9	M14044:calpactin I heavy chain (p36) mRNA	ANX2_MOUSE. function: calcium−regulated membrane−binding protein whose affinity for calcium is greatly enhanced by anionic phospholipids. it binds two calcium ions with high affinity. subunit: tetramer of 2 light chains (p10 proteins) and 2 heavy chains (p36 proteins). subcellular location: in the lamina beneath the plasma membrane. domain: contains four homologous repeats with a consensus sequence common to all annexin proteins. a pair of these repeats may form one binding site for calcium and phospholipid. miscellaneous: it may cross−link plasma membrane phospholipids with actin and the cytoskeleton and be involved with exocytosis. similarity: belongs to the annexin family.
MM_TBI858.1	101062_at	1.60	HO2_MOUSE	3.1	AF054670:Heme oxygenase (decycling) 2	HO2_MOUSE. function: heme oxygenase cleaves the heme ring at the alpha methene bridge to form biliverdin. biliverdin is subsequently converted to bilirubin by biliverdin reductase. under physiological conditions, the activity of heme oxygenase is highest in the spleen, where senescent erythrocytes are sequestrated and destroyed. function: heme oxygenase 2 could be implicated in the production of carbon monoxide in brain where it could act as a neurotransmitter. catalytic activity: heme + 3 ah(2) + o(2) = biliverdin + fe(2+) + co + 3 a + 3 h(2)o. subcellular location: microsomal. similarity: belongs to the heme oxygenase family. similarity: contains 2 heme regulatory motifs (hrm).
MM_TBI49.4	99160_s_at 99161_at	1.59 1.40	GRINA		AW227647: cDNA	
MM_TBI1501.1	102245_at 102246_g_at	1.08 1.58	STAG3	11.1	AJ005678:nuclear protein stag3	
MM_TBI2412.1	100170_at 93127_at	38.74 1.55	SERA_MOUSE	25.3	"L21027:D−3−PHOSPHOGLYCERATE DEHYDROGENASE"	SERA_MOUSE. catalytic activity: 3−PHOSPHOGLYCERATE + NAD(+) = 3−PHOSPHOHYDROXYPYRUVATE + NADH. SUBUNIT: HOMOTETRAMER (by similarity). Similarity: belongs to the D−ISOMER specific 2−HYDROXYACID DEHYDROGENASES family.
MM_TBI19484.1	104528_at	1.51	Mm_TBI21319.1	3.3	"AF119498:C11orf5"	
MM_TBI13646.1	98930_at	1.44	COPE_MOUSE	1.5	"U89427:epsilon−COP"	COPE_MOUSE. function: the coatomer is a cytosolic protein complex that bindsto dilysine motifs and reversibly associates with golgi non−clathrin−coated vesicles, which further mediate biosyntheticprotein transport from the er, via the golgi up to the trans golginetwork. coatomer complex is required for budding from golgimembranes, and is essential for the retrograde golgi−to−ertransport of dilysine−tagged proteins. in mammals, the coatomercan only be recruited by membranes associated to adp−ribosylationfactors (arfs), which are small gtp−binding proteins; the complexalso influences the golgi structural integrity, as well as theprocessing, activity, and endocytic recycling of ldl receptors (bysimilarity).subunit: oligomeric complex that consists of at least the alpha,beta, beta''''''', gamma, delta, epsilon and zeta subunits.subcellular location: the coatomer is cytoplasmic or polymerizedon the cytoplasmic side of the golgi, as well as on thevesicles/buds originating from it (by similarity).similarity: belongs to the cope family.
MM_TBI3989.1	102451_f_at 104368_at	1.00 1.42	EB2	2.7	U51204:APC−binding protein EB2 mRNA	
MM_TBI13713.1	97875_at	1.42	G100_HUMAN	2.6	"AF225959:adhesion regulating molecule ARM−1"	
MM_TBI22930.1	103563_at	1.41	Mm_TBI22431.1	1.2 2.7	AW125713: cDNA	
MM_TBI10330.3	96298_f_at	1.37	Mm_TBI23636.1	10.9	"AF020185:protein inhibitor of nitric oxide synthase"	DYL1_HUMAN. function: may be involved in some aspects of dynein−relatedintracellular transport and motility. may play a role in changingor maintaining the spatial distribution of cytoskeletalstructures.function: binds and inhibits the catalytic activity of neuronalnitric oxide synthase.subunit: consists of at least two heavy chains and a number ofintermediate and light chains.subcellular location: cytoplasmic.tissue specificity: ubiquitous.similarity: belongs to the dynein light chain family.
MM_TBI107269.1	103138_f_at 92825_at	1.00 1.33	TPX1	2.9	"M25533:testis−specific protein (Tpx−1)"	
MM_TBI2582.1	94269_at	1.32	RAB6	3.1 6.7	"AF120162:prenylated RAB acceptor 1"	
MM_TBI8135.1	94861_at	1.31	Mm_TBI8135.1			
MM_TBI9483.2	96313_at	83.0	GNRP_MOUSE		"U55232:Grf1 guanine nucleotide−releasing factor 1"	
MM_TBI9925.1	101413_at	43.47	100K_RAT	1.8	AI847142: cDNA	
MM_TBI489.1	102634_at	3.26	UBH1	1.0	AF022792:deubiquitinating enzyme (UBH1) mRNA	
MM_TBI2705.6	101239_f_at	210.99	HSP2_MOUSE	31.0 27.3 20.2	AV261930: cDNA	
MM_TBI2183.1	94143_at	1.59	CHIO_HUMAN	13.6 8.9 10.7 4.5	X02801:Glial fibrillary acidic protein	GFAP_MOUSE. function: gfap, a class−iii intermediate filament, is a cell−specific marker that, during the development of the central nervous system, distinguishes astrocytes from other glial cells. similarity: belongs to the intermediate filament family.
MM_TBI49359.1	93208_at	1.00	ACRO_MOUSE	2.8	D00754:Preproacrosin	ACRO_MOUSE. function: acrosin is the major protease of mammalian spermatozoa. it is a serine protease of trypsin−like cleavage specificity, it is synthesized in a zymogen form, proacrosin and stored in the acrosome. catalytic activity: hydrolysis of arg− and lys−bonds; preferential cleavage arg−xaa >> lys−lys >> lys−xaa. subunit: heavy chain (catalytic) and a light chain linked by two disulfide bonds. similarity: belongs to peptidase family also known as the trypsin family.
Genes down−regulated in CREM −/− mice found only with Affymetrix chips (factor > 2)						
AE000663	99400_at	3683			AE000663:TCR beta locus from bases 1 to 250611 (section 1 of 3) of the complete sequence	

Appendix B

EST cluster	Affymetrix ID	Affy. Diff.	SSH ID	Nylon Diff.	Title	Description
Z47352	100545_at 93242_at	3625 23.5			Z47352:Protamine 3	Function: protamines substitutes for histones in the chromatin of sperm during the haploid phase of spermatogenesis. they compact sperm dna into a highly condensed, stable and inactive complex. subunit: cross−linked by interchain disulfide bonds around the dna−helix (by similarity). subcellular location: nuclear. tissue specificity: testis.
MM_TBI106647.1	92966_g_at	3249			J04698:Acetylcholine receptor epsilon	ACHE_MOUSE. function: after binding acetylcholine, the achr responds by an extensive change in conformation that affects all subunits and leads to opening of an ion−conducting channel across the plasma membrane. subunit: pentamer of two alpha chains, and one each of the beta, delta, and gamma (in immature muscle) or epsilon (in mature muscle) chains. subcellular location: integral membrane protein. similarity: belongs to the ligand−gated ionic channels family. Receptor
MM_TBI9674.1	97287_at 97288_at	2216 3152			"AF220100: PDZK1"	1. J Biol Chem. 2001 Mar 23;276(12):9206−13.
MM_TBI6251.1	100058_at	2625			AW047776: cDNA	
MM_TBI17048.1	93769_at	2127				
MM_TBI6410.1	97704_at	2088			AA414990:vc50a01.r1 cDNA	
MM_TBI8587.1	104737_at	1508			AA689670:vs03b05.r1 cDNA	
MM_TBI981.1	99410_at	1432			AF039213:pH sensitive maxi K+ channel (Slo3) gene, Slo3−1 allele	Ion channel
MM_TBI10957.1	99137_at	1375			AB016275:Oaz−t, ornithine decarboxylase antizyme 3	
MM_TBI118799.1	99384_at	1167			M13945:pim−1 protein; pim−1 protein kinase	PIM1_MOUSE. catalytic activity: atp + a protein = adp + a phosphoprotein. disease: frequently activated by provirus insertion in murine leukemia virus−induced t−cell lymphomas. similarity: belongs to the ser/thr family of protein kinases. Ser/thr family of protein kinases
MM_TBI5078.1	93162_f_at	1093			AF045953:unknown protein gene	?
AF030522	93846_at	1053			AF030522:Stannin	SNN_MOUSE. function: plays a role in the toxic effects of organotins. tissue specificity: high level of expression in spleen, followed by brain and kidney. induction: by trimethyltin (tmt), a trialkyltin compound which is a potent neurotoxic agent that selectively damages specific brain regions. 1. Inhibitor of apoptosis Expressed during tumor necrosis factor−alfa (Blood. 1999 May 15;93(10):3418−31.) and trialkyltin (Toxicol Pathol. 2000 Jan−Feb;28(1): 43−53.) induced APOPTOSIS
MM_TBI18998.1	98092_at	1037			"AF263458: onzin"	?
MM_TBI4202.1	100212_f_at	979			AV374868: cDNA	
MM_TBI16452.1	97937_at	906			"AF079852: intestinal−enriched Kruppel−like factor IKLF"	KLF5_MOUSE. function: transcription factor that binds to gc box promoterelements. activates the transcription of these genes.subcellular location: nuclear.tissue specificity: highest expression in digestive track.similarity: belongs to the krueppel family of c2h2−type zinc−finger proteins. zinc−finger transcription factor binding caat/gt box j biol chem. 2001 mar 9;276(10):6897−900.
MM_TBI9175.1	103492_at	835			AF077738:metallocarboxypeptidase CPX−1 mRNA	Carboxypeptidases: Enzymes (particularly of pancreas) that remove the C−terminal amino acid from a protein or peptide. Carboxypeptidase A, (EC 3.4.17.1) will remove any amino acid; carboxypeptidase B (EC 3.4.17.2) is specific for terminal lysine or arginine.(Cell Biology Dictionary) Carboxypeptidase E (CPE) family
MM_TBI6276.1	96735_at	140			"AB031550: pctp−L"	
MM_TBI2195.2	99589_f_at	89.2			X07625:Protamine 1	HSP1_MOUSE. function: protamines substitutes for histones in the chromatin of sperm during the haploid phase of spermatogenesis. they compact sperm dna into a highly condensed, stable and inactive complex. subunit: cross−linked by interchain disulfide bonds around the dna−helix (by similarity). subcellular location: nuclear. tissue specificity: testis. Sperm DNA compaction
MM_TBI9097.1	104593_at	38.6			AI849396: cDNA	
MM_TBI2746.1	97760_at	29.8			M21041:Microtubule−associated protein 2	MAP2_MOUSE. function: the exact function of map2 is unknown but maps may stabilize the microtubules against depolymerization. they also seem to have a stiffening effect on microtubules. similarity: contains 3 tau/map repeats. Microtubule Stabilization and Stiffening (?) −
MM_TBI2371.1	102330_at	29.1			D86382:Iba1 (ionized calcium binding adapter molecule 1)	
MM_TBI169.1	92936_at	25.8			"X14943: neuronal cell surface protein F3"	Probable : F3 promotes remodelling of neurosecretory terminals (Exp Physiol 2000 Mar;85 Spec No:187S−196S).
MM_TBI1625.1	101859_at	19.1			AB010100:aquaporin 7	AQP7_MOUSE. function: forms a channel for water and glycerol. subcellular location: integral membrane protein. similarity: belongs to the transmembrane channel mip family. Water channel AQP7 contributes to the volume reduction of spermatids, since this water channel protein is localized on the plasma membrane covering the condensing cytoplasmic mass of the elongated spermatid, and since the seminiferous tubule fluid is hypertonic(Cell Tissue Res 1999 Feb;295(2):279−85).
MM_TBI21853.1	103343_at	18.1			AI845815: cDNA	
MM_TBI3467.1	98603_s_at	17.7			U20857:RAN1 homolog (Fug1)	RGP1_MOUSE. function: gtpase activator for the nuclear ras−related regulatory protein ran, converting it to the putatively inactive gdp−bound state. required for postimplantation development. subunit: homodimer. forms a tight complex in association with ranbp2 and the ubiquitin−conjugating enzyme e2 (ubc9) (by similarity). subcellular location: cytoplasmic (by similarity). ptm: seems to be converted to a 20 kda heavier form by conjugation with a small ubiquitin−like protein ubl1 (sumo−1). similarity: contains ? leucine−rich repeats (lrr). similarity: to fungal rna1. The 3.5−kb transcript present in all tissues and highly expressed in brain, thymus and testis, we found a second transcript of 2.8 kb resulting from a distinct 3′ UTR in testis.
MM_TBI5648.1	94178_at	17.2			AB010919:membrane cofactor protein (CD46)	Complement regulator (protector against Immune cytolysis) complement (C) regulatory proteins CD59, MCP, decay−accelerating factor (DAF), present on the acrosomal region of condensing spermatids(Immunology 1994 Mar;81(3):452−61)

Appendix B

EST cluster	Affymetrix ID	Affy. Diff.	SSH ID	Nylon Diff.	Title	Description
MM_TBI18227.1	104207_at	17.0			AI430272:mf46f12.y1 cDNA	
MM_TBI3626.1	95471_at	15.2			"U22399: p57KIP2"	Cyclin−dependent kinase (CDK) inhibitor p57Kip2 is cyclin−dependent kinase (CDK) inhibitor p57Kip2 (Anat Embryol (Berl). 2001 Feb;203(2):77−87.)
MM_TBI8749.1	93143_at	14.9			AI844196: cDNA	
MM_TBI2179.1	92642_at	12.1			"M25944: carbonic anhydrase II (CAII)"	Function: reversible hydration of carbon dioxide. Reversible hydrotation of carbon dioxide (H2CO3 = CO2 + H2O)
MM_TBI100107.1	102846_at	11.7			AF019926:Testis specific serine/ threonine kinase 2	SIMILARITY: TO THE SER/THR FAMILY OF PROTEIN KINASES. Serine/ threonine kinase
MM_TBI5341.1	101150_at	11.3			L10427:ets−related protein 71 (ER71) mRNA	ETV2_MOUSE. function: binds to dna sequences containing the consensus pentanucleotide 5''−cgga[at]−3''. subcellular location: nuclear. tissue specificity: testis. similarity: belongs to the ets family. ETS−family transcription factor ER71 expression is restricted to testis(Genes Dev 1992 Dec;6(12B):2502−12).
MM_TBI6183.13	102153_at	11.0			M36690:germline heavy−chain gene V region	
MM_TBI123917.1	97007_at	11.0			AJ245454:sperm motility kinase 2, (Smok2(tw5) gene) strain t−haplotype tw5	similarity: to the SER/THR family of protein kinases. Protein_Kinase Smok is expressed late during spermiogenesis.
MM_TBI2762.1	96020_at	10.2			"M22531: complement C1q B chain"	Complement
MM_TBI3469.1	103083_at	10.0			U69543:Lipase, hormone sensitive	Triglycerid hydrolysis
MM_TBI20956.1	93409_at	9.43			AA139057:mr04g08.r1 cDNA	
AW122677	103432_at	9.26			AW122677: cDNA	
MM_TBI2846.1	100891_at	9.13			M88463:seleno−protein gene	MCS_MOUSE. function: structural protein of the sperm mitochondrial capsule. important for the maintenance and stabilization of the crescent structure of the sperm mitochondria. subcellular location: keratinous mitochondrial capsule. tissue specificity: testis. developmental stage: late meiotic and early haploid cells. Structural protein of the sperm mitochondrial capsule the mitochondrial capsule.
MM_TBI11153.1	99924_at	9.02			AW121845: cDNA	
MM_TBI15863.1	97404_at	8.42			AW048244: cDNA	
MM_TBI1627.1	93711_at	8.27			D12713:Secretory protein SEC23 related gene (MSEC66)	S23A_MOUSE. function: component of the copii coat, that covers er−derived vesicles involved in transport from the endoplasmic reticulum to the golgi apparatus. copii acts in the cytoplasm to promote the transport of secretory, plasma membrane, and vacuolar proteins from the endoplasmic reticulum to the golgi complex (by similarity). subunit: copii is composed of at least five proteins: the sec23/24 complex, the sec13/31 complex and sar1. subcellular location: cytoplasmic, in the ribosome−free transitional face of the er and associated vesicles. tissue specificity: high levels in brain and fibroblasts. similarity: belongs to the sec23/sec24 family. sec23 subfamily. caution: this is a conceptual translations, many probable frameshifts were corrected to produce a protein similar to the human homologs. Export from ER protein export from the ER.(Mol Biol Cell 1996 Oct;7(10):1535−46)
MM_TBI48834.1	92691_at	7.91			AW048155: cDNA	
MM_TBI27954.1	100876_at	7.89			AI841076: cDNA	
MM_TBI3219.1	98589_at	7.87			M93275:Adipose differentiation related protein	ADFP_MOUSE. function: may be involved in development and maintenance of adipose tissue. subcellular location: membrane−associated. tissue specificity: adipose tissue specific. expressed abundantly and preferentially in fat pads. induction: by dexamethasone. similarity: belongs to the peripilin family. Long chain fatty acids transport ADRP function as a saturable transport component for long chain fatty acids (J Biol Chem 1999 Jun 11;274(24):16825−30).
MM_TBI2213.1	102858_at	7.86			L02241:protein kinase inhibitor (testicular isoform) mRNA	IPKB_MOUSE. function: extremely potent competitive inhibitor of camp−dependent protein kinase activity, this protein interacts with the catalytic subunit of the enzyme after the camp−induced dissociation of its regulatory chains. alternative products: 2 isoforms; 1 and 2 (shown here); are produced by alternative splicing. similarity: belongs to the pki family. Inhibitor of PKA
MM_TBI15958.1	104537_at	7.33			AW048828: cDNA	
MM_TBI8966.1	97999_at	7.26			AI838661: cDNA	
MM_TBI11607.1	98929_at	6.99				
MM_TBI13884.1	92810_at	6.61				
AF109905	93197_at 97894_at	1313 6.49			"major histocompatibility locus class III regions"	
MM_TBI16474.1	103625_at	6.45			AA797556: cDNA	
MM_TBI4083.2	100293_at	6.25			AA268823: cDNA	
MM_TBI781.1	93358_at	6.18				
MM_TBI2228.1	94189_at	6.11			AB011665:BAZF (Bcl6 homolog)	
MM_TBI108685.1	100171_f_at 93834_at	5.97 2.40			"M25487: histone 2b protein (His2b)"	Histone
MM_TBI4439.1	101889_s_at	5.93			U53228:RAR−related orphan receptor alpha	ROR1_MOUSE. Nuclear hormone receptor
U96746	93495_at	5.76			"U96746: antioxidant enzyme AOE372"	Function: regulates the activation of NF−KAPPA−B in the cytosol by a modulation of I−KAPPA−B−ALPHA Antioxidant Prx4 was restricted to membranes of the acrosomal vesicle of the elongated spermatid and was not detected in spermatozoon.
MM_TBI4220.2	103960_at	5.72			U73941:Rap2 interacting protein 8 (RPIP8) mRNA	
MM_TBI13422.1	102052_at	5.62			AA871791:vq41a10.r1 cDNA	
MM_TBI3683.1	98484_at	5.59			U37186:Islet cell autoantigen 1, 69 kDa	?

Appendix B

EST cluster	Affymetrix ID	Affy. Diff.	SSH ID	Nylon Diff.	Title	Description
MM_TBI1238.1	103806_at	5.57			AF064984:Low density lipoprotein−related protein 5	
MM_TBI2364.1	103434_at	5.54			AF001871:guanine nucleotide exchange factor and integrin binding protein homolog GRP1 mRNA	Function: promotes guanine−nucleotide exchange on ARF1. promotes the activation of ARF through replacement of GDP with GTP. Guanyl−nucleotide exchange factor GRP1, an ADP−ribosylation factor (ARF)−guanine nucleotide exchange protein regulated by phosphatidylinositol 3,4,5−trisphosphate (Methods Enzymol. 2001;329:279−89.).
MM_TBI18404.2	103320_at	5.22			"AF218069: galectin−8"	
MM_TBI14574.1	95442_at	5.02			"clone:2−72"	
MM_TBI15576.1	104439_at	4.91			"AF172275: FUS2"	?
MM_TBI5958.1	92743_at	4.86			U33958:Sperm adhesion molecule (PH−20)	HYA1_MOUSE. function: involved in sperm−egg adhesion. upon fertilization sperm must first penetrate a layer of cumulus cells that surrounds the egg before reaching the zona pellucida. the cumulus cells are embedded in a matrix containing hyaluronic acid which is formed prior to ovulation. this protein aids in penetrating the layer of cumulus cells by digesting hyaluronic acid. catalytic activity: random hydrolysis of 1,4−linkages between n−acetyl−beta−d−glucosamine and d−glucuronate residues in hyaluronate. subcellular location: attached to the membrane by a gpi−anchor. similarity: belongs to family 56 of glycosyl hydrolases. Glycosyl hydrolases
MM_TBI9066.1	96184_at	4.73			AI835242: cDNA	
MM_TBI11493.2	101944_at 101945_g_at 101946_at	7.59 20.2 4.63			U89352:lysophospholipase I mRNA	LysophospholipaseAn enzyme that catalyzes the hydrolysis of a single fatty acid ester bond in lysoglycerophosphatidates with the formation of glyceryl phosphatidates and a fatty acid (The On−line Medical Dictionary). Lysophospholipase
MM_TBI6111.1	99324_at	4.41			U73378:enteropeptidase mRNA	ENTK_MOUSE. function: responsible for initiating activation of pancreatic proteolytic proenzymes (trypsin, chymotrypsin and carboxypeptidase a). it catalyzes the conversion of trypsinogen to trypsin which in turn activates other proenzymes including chymotrypsinogen, procarboxypeptidases, and proelastases (by similarity). catalytic activity: selective cleavage of 6−lys−|−ile−7 bond in trypsinogen. subunit: heterodimer of a catalytic (light) chain and a multidomain (heavy) chain linked by a disulfide bond (by similarity). subcellular location: type ii membrane protein (probable). ptm: the chains are derived from a single precursor that is cleaved by a trypsin−like protease (by similarity). similarity: contains 2 ldl−receptor class a domains. similarity: contains 2 cub domains. similarity: contains 1 sea domain. similarity: contains 1 srcr domain. similarity: contains 1 mam domain. similarity: belongs to peptidase family also known as the trypsin family.
MM_TBI4421.1	101312_at	4.37			D10651:Glutamate receptor channel subunit epsilon 2	NME2_MOUSE. function: nmda receptor subtype of glutamate−gated ion channels possesses high calcium permeability and voltage−dependent sensitivity to magnesium and is mediated by glycine. subunit: heterodimer of an epsilon subunit and a zeta subunit. subcellular location: integral membrane protein. similarity: belongs to the ligand−gated ionic channels family. Receptor channel
AF068865	102746_at	3.95			AF068865:Delta−like 3 (Dll3) gene, alternative splice products	DLL3_MOUSE. function: inhibits primary neurogenesis. may be required to divert neurons along a specific differentiation pathway. play a role in the formation of somite boundaries during segmentation of the paraxial mesoderm. subunit: can bind and activate notch−1 or another notch receptor (probable). subcellular location: type i membrane protein (probable). alternative products: 2 isoforms; 1 and 2 (shown here); are produced by alternative splicing. tissue specificity: predominantly expressed in the neuroectoderm and paraxial mesoderm during embryogenesis. domain: the delta−serrate−lag2 (dsl) domain is required for binding to the notch receptor. disease: a truncating mutation in dll3 is the cause of the pudgy (pu) phenotype. pudgy mice exhibit patterning defects at the earliest stages of somitogenesis. adult pudgy mice present severe vertebral and rib deformities. similarity: contains 6 egf−like domains. similarity: belongs to the delta/serrate/jagged family. Ligand
MM_TBI1172.1	95060_at	3.90			"AF058054: monocarboxylate transporter 2"	MOT2_MOUSE. function: proton−linked monocarboxylate transporter. catalyzes therapid transport across the plasma membrane of manymonocarboxylates such as lactate, pyruvate, branched−chain oxoacids derived from leucine, valine and isoleucine, and the ketonebodies acetoacetate, beta−hydroxybutyrate and acetate. mct2 is ahigh affinity pyruvate transporter.subcellular location: integral membrane protein. plasma membrane(by similarity).similarity: belongs to the slc16 family of transporters. monocarboxylate transporter
MM_TBI590.1	103088_at	3.79			X94310:L1−like protein	
MM_TBI11390.1	99529_f_at	3.68			AB025011:Trif−d	Transcription repressor (?)
MM_TBI769.1	100092_at	3.68			Z22593:fibrillarin mRNA	FBRL_MOUSE. function: fibrillarin is a component of a nucleolar small nuclear ribonucleoprotein particle thought to participate in the first step in processing preribosomal rna. it is associated with the u3, u8 and u13 small nuclear rnas. subcellular location: nuclear; fibrillar region of the nucleolus. ptm: by homology to other fibrillarins, some or all of the n−terminal arginines are n,n−dimethylated (dma). similarity: belongs to the fibrillarin family.
MM_TBI18301.1	95576_at	3.67			AA007891: cDNA	
MM_TBI182.2	97271_at	3.57				
MM_TBI656.5	93994_at	3.55				
MM_TBI12753.1	95490_at	3.52				
MM_TBI4064.1	98848_at	3.48			U58889:SH3−containing protein SH3P3, vinexin alpha	Cytoskeleton reorganisation
MM_TBI13376.1	94850_at	3.47			"AJ238894: acyl−CoA thioesterase"	AC48_MOUSE. function: active on long chain acyl−coas.subcellular location: mitochondrial.tissue specificity: ubiquitous.similarity: belongs to the acyl coenzyme a hydrolase family.
MM_TBI1535.5	97462_at 97463_g_at	3.39 2.37				
MM_TBI20033.4	99169_at	3.36			AW122165: cDNA	

Appendix B

EST cluster	Affymetrix ID	Affy. Diff.	SSH ID	Nylon Diff.	Title	Description
MM_TBI918.1	101936_at 101937_s_at	3.06 3.26			AF005423:cdc2/CDC28−like kinase 4 (Clk4) mRNA, partial cds	CLK4_MOUSE. function: phosphorylates serine− and arginine−rich (sr) proteinsof the spliceosomal complex may be a constituent of a network ofregulatory mechanisms that enable sr proteins to control rnasplicing. phosphorylates serines, threonines and tyrosines.subcellular location: nuclear (by similarity).ptm: autophosphorylates on all three type of residues.similarity: belongs to the ser/thr family of protein kinases.lammer subfamily. splicing regulation
MM_TBI1520.13	96215_f_at	3.17			AI153421: cDNA	
MM_TBI3846.1	95133_at	3.17			"U38940: asparagine synthetase"	1. Alanine and aspartate metabolism
AC002397	103993_at	3.17			AC002397:chromosome 6 BAC−284H12	
MM_TBI5479.1	101315_at	3.14	TBA1_MOUSE	4.2 1.0 3.8	M19413:testicular alpha tubulin mRNA	
MM_TBI272.1	103360_at	3.13			U01840:Testis−specific serine/threonine kinase	Serine/threonine kinase tssk−1 and tssk−2, expressed exclusively in spermatids undergoing spermiogenesis (Mech Dev 2000 May;93(1−2):175−7).
MM_TBI6828.1	97512_at	2.93			AW226650:um58g10.y1 cDNA	
MM_TBI12665.1	95011_at	2.83			AI853090: cDNA	
MM_TBI4008.1	98544_at	2.81			U53514:guanylate kinase (gmk) mRNA	KGUA_MOUSE. function: essential for recycling gmp and indhrectly, cgmp. catalytic activity: atp + gmp = adp + gdp. subunit: monomer (by similarity). similarity: belongs to the guanylate kinase family. Purine metabolism
MM_TBI10187.2	93853_at	2.80			AA763918:vv48f08.r1 cDNA	
MM_TBI2714.1	101388_at	2.78			M17299:testis−specific Phosphoglycerate kinase 2 (PGK−2)	PGK2_MOUSE. catalytic activity: atp + 3−phospho−d−glycerate = adp + 3−phospho−d−glyceroyl phosphate. pathway: second step in the second phase of glycolysis. subunit: monomer. similarity: belongs to the phosphoglycerate kinase family. 1. Glycolysis / Gluconeogenesis The results reveal that the majority of PGK−2 mRNA activity of round spermatids was present in the polysomal fraction while the relatively less abundant PGK−2 mRNA of pachytene primary spermatocytes was present in the nonpolysomal fraction (Dev Biol 1983 Aug;98(2):392−9).
MM_TBI8680.1	97377_at	2.59			"AF208663: coilin p80"	?
MM_TBI12840.1	104343_f_at	2.59			AI845798: cDNA	
MM_TBI3199.4	95486_at	2.49				
MM_TBI2141.1	102364_at	2.29			J04509:Jun proto−oncogene related gene d1 (JUN−D)	JUND_MOUSE. subunit: binds dna as a dimer (by similarity). subcellular location: nuclear. tissue specificity: brain and kidney. similarity: belongs to the bzip family. jun subfamily. Transcription factor
MM_TBI25769.3	94891_s_at	2.26			"M27983: male−enhanced antigen (Mea)"	Golgi structural protein In situ hybridization analysis suggested that the Mea−2 gene is expressed in spermatids during spermatogenesis as already shown by Mea−1, suggesting that Mea−2 gene product as well as Mea−1 have also some role for spermatogenesis (DNA Seq 1997;7(2):71−82).
MM_TBI651.14	96661_at	2.19				
Mm_TBI2361.1	93869_s_at	6			hemopoietic−specific early response protein (A1)	BFL1_MOUSE. function: retards apoptosis induced by il−3 deprivation. may function in the response of hemopoietic cells to external signals and in maintaining endothelial survival during infection. subcellular location: intracellular. tissue specificity: expressed in hemopoietic tissues, including bone marrow, spleen and thymus. induction: by granulocyte−macrophage colony−stimulating factor and lps in macrophages.
Genes up−regulated in CREM−deficient testes found only with Affymetrix chip (factor > 2)						
MM_TBI11125.1	96473_f_at	0.33			"AF155142.1:mixed lineage kinase 3 (Mlk3)"	
MM_TBI3524.19	101676_at	0.33			U13705:Glutathione peroxidase 3	GSHP_MOUSE. function: protects cells and enzymes from oxidative damage, by catalysing the reduction of hydrogen peroxide, lipid peroxides and organic hydroperoxide, by glutathione. catalytic activity: 2 glutathione + h(2)o(2) = oxidized glutathione + 2 h(2)o. cofactor: selenocysteine. the active−site selenocysteine is encoded by the opal codon, uga. subunit: homotetramer. subcellular location: extracellular. tissue specificity: secreted into the plasma. similarity: belongs to the glutathione peroxidase family. 1. Antioxidant
AI850438	98344_f_at	0.28			AI850438: cDNA	
MM_TBI1820.1	92983_at 92984_g_at	0.12 0.26			D49393:protein tyrosine phosphatase	Protein tyrosine phosphatase (cytoplasmic)
MM_TBI16674.1	93568_i_at 93569_f_at	0.25 0.22				
MM_TBI459.1	92969_at	0.24			X76505:Neurotrophic tyrosine kinase, receptor, type 3	DDR2_MOUSE. catalytic activity: atp + a protein tyrosine = adp + protein tyrosine phosphate. subcellular location: type i membrane protein. alternative products: different transcripts are derived from one gene. tissue specificity: widely expressed; high levels in skeletal muscle, heart, cns, and kidney; less in other tissues. the major 10 kda transcript is expressed in high levels in heart and lung, less in brain and testis. similarity: to other protein−tyrosine kinases in the catalytic domain. similarity: contains 1 f5/8 type c domain. similarity: belongs to the insulin receptor family of tyrosine− protein kinases.
MM_TBI2038.1	104606_at	0.22			M55561:phosphatidylinositol−linked antigen (pB7)	CD52_MOUSE. function: may play a role in carrying and orienting carbohydrate, as well as having a more specific role. subcellular location: attached to the membrane by a gpi−anchor. tissue specificity: expressed on lymphohematopoietic tissues, including thymus, spleen, and bone marrow, but not in liver, kidney, and brain. ? Here we show that the antigen is also expressed at a high level in the male reproductive system, being found in the epididymis, seminal vesicle, seminal plasma and on the surface of mature (but not testicular) spermatozoa (J Reprod Immunol 1993 Mar;23(2):189−205).
MM_TBI12376.1	95740_at	0.22				
MM_TBI27411.1	103308_at	0.22			AI450597:mq87e05.x1 cDNA	
MM_TBI17080.1	100281_at	0.22			"AF186095: germ cell−less 1 protein (Gcl−1)"	

Appendix B

EST cluster	Affymetrix ID	Affy. Diff.	SSH ID	Nylon Diff.	Title	Description
MM_TBI26536.1	103460_at	0.21			AI849939: cDNA	
MM_TBI265.2	101130_at	0.21			X58251:Procollagen, type I, alpha 2	CA21_MOUSE. function: type i collagen is a member of group i collagen (fibrillar forming collagen). subunit: trimers of one alpha 2(i) and two alpha 1(i) chains. tissue specificity: forms the fibrils of tendon, ligaments and bones. in bones the fibrils are mineralized with calcium hydroxyapatite. ptm: prolines at the third position of the tripeptide repeating unit (g−x−y) are hydroxylated in some or all of the chains. Extracellular matrix structural protein
MM_TBI4892.1	99842_at	0.20			AB000636:Procollagen, type XIX, alpha 1	Extracellular matrix structural protein Col19a1 transcripts can be detected as early as 11 days of gestation and in all embryonic tissues, except the liver, of an 18−day postcoitum mouse. In contrast, only a few adult tissues, brain, eye, and testis, seem to accumulate Col19a1 mRNA.
MM_TBI42307.1	92274_at	0.20			AB019118:decidualin, prolactin−like protein J	Hormone
MM_TBI7021.1	93193_at	0.17			X15643:Adrenergic receptor, beta 2	B2AR_MOUSE. function: beta−adrenergic receptors mediate the catecholamine− induced activation of adenylate cyclase through the action of g proteins. the beta−2−adrenergic receptor binds epinephrine with an approximately 30−fold greater affinity than it does norepinephrine. subcellular location: integral membrane protein. ptm: homologous desensitization of the receptor is mediated by its phosphorylation by beta−adrenergic receptor kinase. similarity: belongs to family 1 of g−protein coupled receptors.
MM_TBI4259.2	99552_at	0.13			U79550:Slug zinc finger protein	SLUG_MOUSE. function: transcriptional repressor. involved in the generation and migration of neural crest cells. subcellular location: nuclear (probable). similarity: belongs to the snail family of zinc finger proteins. Transcriptional repressor
MM_TBI4068.1	98582_at	0.09			U58988:Homogentisate 1, 2−dioxygenase (HGO)	HGD_MOUSE. catalytic activity: homogentisate + o(2) = 4−maleylacetoacetate. cofactor: iron. pathway: catabolism of tyrosine; third step, catabolism of phenylalanine; fourth step. subunit: homotrimer (probable). disease: defects in hgd are the cause of alkaptonuria (aku), an autosomal recessive error of metabolism. aku is characterized by an increase in the level of homogentisic acid. similarity: belongs to the homogentisate dioxygenase family. Catabolism of tyrosine
MM_TBI2405.1	92759_at	0.08			U43298:Laminin, beta 3 (Lamb3)	LMB3_MOUSE. function: binding to cells via a high affinity receptor, laminin is thought to mediate the attachment, migration, & organization of cells into tissues during embryonic development by interacting with other extracellular matrix components. subunit: laminin is a complex glycoprotein, consisting of three different polypeptide chains (alpha, beta, gamma), which are bound to each other by disulfide bonds into a cross−shaped molecule comprising one long & three short arms with globules at each end. the beta−3 chain is a subunit of laminin−5 (epiligrin/kalinin/ nicein). subcellular location: extracellular. tissue specificity: found in the basement membranes (major component). domain: the alpha−helical domains i and ii are thought to interact with other laminin chains to form a coiled coil structure. domain: domain vi is globular. similarity: contains 1 laminin n−terminal domain (domain vi). similarity: contains 6 laminin egf−like domains. Extracellular matrix structural protein the base of the Sertoli cells is in contact with the basement membrane matrix, in which the laminins constitute the major noncollagenous components. Antilaminin antibody cause the lesions included thickening of the limiting membrane, infolding in the basal lamina, deposits of immune complexes coincident with sloughing of pachytene spermatocytes and spermatids, and vacuolization of the Sertoli cells.(Biol Reprod 2000 Jun;62(6):1505−14).
MM_TBI4564.1	97764_at 97765_g_at	0.04 0.07			U15443:C−Ros proto−oncogene	Tyrosine kinase receptor Male homozygous transgenic c−ros knockout mice are sterile by natural mating, lack a part of their epididymis, and the epididymal sperm exhibit tail angulation in vivo and in vitro . The infertility of c−ros knockout male mice can be explained by the sperm's inability to enter the oviduct, as a result of their bent tails forming the entangled sperm mass and their compromised flagellar vigor within the uterus.(Biol Reprod 2000 Aug;63(2):612−8).
MM_TBI14707.3	94514_s_at	0.07				
MM_TBI108770.1	96657_at	0.05			"spermidine/spermine N1−acetyltransferase (SSAT)"	
MM_TBI12109.1	94330_at	0.05			AA710564: cDNA	
MM_TBI4478.1	93122_at	0.01			M92849: acidic epididymal glycoprotein (Aeg−1)	AEG1_MOUSE. function: this protein is supposed to help spermatozoa undergo functional maturation while they move from the testis to the ductus deferens. subcellular location: stored in secretory granules of granular convoluted tubules cells. tissue specificity: mainly found in the cauda epididymis where it is synthesized by the principal cells and secreted into the lumen. binds to the heads of spermatozoa. also expressed in the submandibular gland. developmental stage: exponential increase between days 25 and 30 after birth. induction: this protein is androgen−dependent. similarity: belongs to a family that groups mammalian scp/insects ag3/fungi sc7/sc14 and plants pr−1. Sperm−egg fusion AEG in that it is an epididymal secretory glycoprotein that binds to the postacrosomal region of the sperm head and involved in the fusion of the sperm and egg plasma membranes.(Genomics 1996 Mar 15;32(3):367−74).
MM_TBI3725.2	99942_s_at	0.001			U28932:smooth muscle calponin gene	CLP1_MOUSE. function: thin filament−associated protein that is implicated in the regulation and modulation of smooth muscle contraction. it is capable of binding to actin, calmodulin, troponin c and tropomyosin. the interaction of calponin with actin inhibits the actomyosin mg−atpase activity. alternative products: 2 isoforms; alpha (shown here) and beta; are produced by alternative splicing. tissue specificity: smooth muscle, and tissues containing significant amounts of smooth muscle. similarity: belongs to the calponin family.
MM_TBI3419.1	92266_at	0.001			U03421:Interleukin 11	IL11_MOUSE. function: directly stimulates the proliferation of hematopoietic stem cells and megakaryocyte progenitor cells and induces megakaryocyte maturation resulting in increased platelet production. subcellular location: secreted. Growth factor (ligand) In testis, IL−11 mRNA is expressed in round spermatids at stage VI−IX seminiferous tubules. Administration of IL−11 in vivo accelerates recovery of spermatogenesis after cytotoxic therapy. (J Cell Physiol 1996 Aug;168(2):362−72)
MM_TBI9326.1	103314_at	0.001			AW046158:UI−M−BH1−akw−a−11−0−UI.s1 cDNA	

Appendix B

EST cluster	Affymetrix ID	Affy. Diff.	SSH ID	Nylon Diff.	Title	Description
Mm_TBI1204.1	99855_at	0.3			apoptosis signal−regulating kinase 1	M3K5_MOUSE. function: phosphorylates and activates two different subgroups of map kinase kinases, mkk4/sek1 and mkk3/mapkk6 (or mkk6), which in turn activated stress−activated protein kinase (sapk, also known as jnk; c−jun amino−terminal kinase) and p38 subgroups of map kinases, respectively. overexpression induces apoptotic cell death (by similarity). tissue specificity: expressed in the various mouse adult tissues including heart, brain, lung, liver and kidney. similarity: belongs to the ser/thr family of protein kinases. map kinase kinase kinase subfamily.
Mm_TBI11975.1	98437_at	0.3			cysteine protease CPP32	ICE3_MOUSE. function: involved in the activation cascade of caspases responsible for apoptosis execution. at the onset of apoptosis it proteolytically cleaves poly(adp−ribose) polymerase (parp) at a 216−asp−|−gly−217 bond. cleaves and activates sterol regulatory element binding proteins (srebps) between the basic helix−loop−helix leucine zipper domain and the membrane attachment domain. cleaves and activates caspase−6, −7 and −9 (by similarity). cleaves il−1 beta between an asp and an ala, releasing the mature cytokine which is involved in a variety of inflammatory processes. subunit: heterodimer of a 17 kda (p17) and a 12 kda (p12) subunit (by similarity). subcellular location: cytoplasmic. tissue specificity: highest expression in spleen, lung, liver, kidney and heart. lower expression in brain, skeletal muscle and testis. ptm: cleavage by granzyme b, caspase−6, −8 and −10 generates the two active subunits. additional processing of the propeptides is likely due to the autocatalytic activity of the activated protease. active heterodimers between the small subunit of caspase−7 protease and the large subunit of cpp32 also our and vice versa (by similarity). similarity: belongs to peptidase family also known as the caspase family.
Genes found by SSH with differential expression indicated on nylon CREM SSH arrays and non−differential expression indicated on Affymetrix chips						
MM_TBI4260.1	95631_at	1.05	ppx	2.1	AF088911:protein phosphatase X (Ppx) mRNA	
MM_TBI17635.14	103429_i_at 94889_at 95282_at 95836_r_at	0.40 0.91 1.01 1.00	28SRNA	16.4 2.5	"AF115503: VAMP−associated protein"	
MM_TBI1850.1	100113_s_at	0.96	KAP3A	8.1 13.4	D50367:KAP3B	
MM_TBI10215.1	92623_at	0.91	UNR_RAT	1.8	"L19607:unr"	
MM_TBI2157.1	93750_at 99212_i_at	0.90 0.78	GELS_MOUSE	11.3	AV369888:AV369888 cDNA	
MM_TBI8963.1	96855_at	0.89	SPC1_HUMAN	3.6		
MM_TBI16848.1	104039_at 95507_at	1.79 0.88	KPR1_HUMAN	3.1	"AB025048:Sid6061p"	
MM_TBI13720.1	94264_at	0.87	Mm_TBI7118.1	4.5		
MM_TBI2868.1	97521_at	0.84	ASSY_MOUSE	2.1 2.8	M31690:Arginosuccinate synthetase 1	ASSY_MOUSE. catalytic activity: atp + l−citrulline + l−aspartate = amp + pyrophosphate + l−argininosuinate. pathway: urea cycle, penultimate step of the arginine biosynthetic pathway. subunit: homotetramer. similarity: belongs to the argininosuinate synthase family.
MM_TBI5878.3	101955_at 103126_f_at	0.84 0.74	BiP	1.4 2.6	AJ002387:BiP	GR78_MOUSE. function: probably plays a role in facilitating the assembly of multimeric protein complexes inside the er. subcellular location: endoplasmic reticulum lumen. similarity: belongs to the heat shock protein 70 family.
MM_TBI2205.1	100753_at 99755_i_at	0.81 0.02	ATPA_MOUSE	1.7 7.0	L01062:ATP synthase alpha subunit	ATPA_MOUSE. function: produces atp from adp in the presence of a proton gradient across the membrane. the alpha chain is a regulatory subunit. subunit: f−type atpases have 2 components, cf(1) − the catalytic core − and cf(0) − the membrane proton channel. cf(1) has five subunits: alpha(3), beta(3), gamma(1), delta(1), epsilon(1). cf(0) has three main subunits: a, b and c. subcellular location: mitochondrial inner membrane. similarity: belongs to the atpase alpha/beta chains family.
MM_TBI1019.2	95503_at	0.76	AF1Q_MOUSE	6.6 6.2 16.1 25.0	"U95498:AF1q"	AF1Q_MOUSE.
MM_TBI10117.1	96879_at	0.75	ODO1_HUMAN	17.5	"U02971:2−oxoglutarate dehydrogenase E1 component"	
MM_TBI2157.2	99213_f_at	1.00	GELS_MOUSE	11.3	AV369888: cDNA	
MM_TBI12931.11	100996_at	0.96	Mm_TBI87042.1 AF015811	7.7	AF015811:putative lysophosphatidic acid acyltransferase mRNA	
MM_TBI4559.1	92696_at	1.29	RTR	20.2 5.9 2.2	U09563:orphan receptor RTR mRNA	NR61_MOUSE. function: orphan receptor. subcellular location: nuclear (probable). tissue specificity: testis−specific. similarity: belongs to the nuclear hormone receptors family. nr6 subfamily.
MM_TBI1219.3	93082_at	1.28	Mm_TBI16245.1	23.5		
MM_TBI13437.1	96338_at	1.27	Mm_TBI13437.1	5.2		
MM_TBI121609.1	94016_at 99117_at	1.19 1.27	TBB5_Mouse	2.2	AW050256: cDNA	
MM_TBI9549.2	93011_at	1.23	Mm_TBI28224.1	8.4	"AW123904:cDNA"	
MM_TBI4160.1	94837_at	1.19	NIPIL	3.7	"U67328:NIPI−like protein"	
MM_TBI13618.1	97388_at	1.16	Mm_TBI13618.1	2.1	AW124130: cDNA	
MM_TBI17449.37	100727_at	1.16	RL28_MOUSE	11.8	X74856:L28 ribosomal protein L28	RL28_MOUSE. similarity: belongs to the l28e family of ribosomal proteins.
MM_TBI651.11	96267_at 96774_at	1.15 1.13	Mm_TBI8803.1	2.6	AW047139: cDNA	
MM_TBI4158.2	98163_f_at	1.15	AF039023	4.4	AV298789: cDNA	
MM_TBI14600.1	95058_f_at	1.13	MmTBI14600.1	3.6		
MM_TBI20930.2	95715_at	1.12	TGLC_CHICK	3.6 2.5		

Appendix B

EST cluster	Affymetrix ID	Affy. Diff.	SSH ID	Nylon Diff.	Title	Description
MM_TBI685.1	103656_at	1.12	MMP40GPRT	16.0	Y16518:G protein−coupled receptor, P40GPRT	
MM_TBI107813.2	99335_at	1.12	HXK1_MOUSE	7.1 2.9	J05277:Hexokinase 1	HXK1_MOUSE. catalytic activity: atp + d−hexose = adp + d−hexose 6−phosphate. enzyme regulation: hexokinase is an allosteric enzyme inhibited by its product glc−6−p. pathway: first step of several metabolic pathways. subunit: monomer. subcellular location: bound to the outer mitochondrial membrane. its hydrophobic n−terminal sequence may be involved in membrane binding. tissue specificity: in rapidly growing tumor cells exhibiting high glucose catabolic rates, this hexokinase is markedly elevated. miscellaneous: in vertebrates there are four major glucose− phosphorylating isoenzymes, designated hexokinase i, ii, iii and iv (glucokinase). similarity: the n− and c−terminal halves of this hexokinase show extensive sequence similarity to each other. the catalytic activity is associated with the c−terminus while regulatory function is associated with the n−terminus. similarity: belongs to the hexokinase family.
MM_TBI345.1	96105_r_at	1.12	MEG1_MOUSE	2.3	"X64455:meg1"	
MM_TBI16824.1	96176_at	1.10	Mm_TBI16824.1	12.7	AI847916: cDNA	
MM_TBI301.1	104628_at	1.04	MAN2_MOUSE	23.9 1.4	X61172:alpha−mannosidase II	MAN2_MOUSE. function: catalyzes the first committed step in the biosynthesis of complex n−glycans. it controls conversion of high mannose to complex n−glycans; the final hydrolytic step in the n−glycan maturation pathway. catalytic activity: hydrolysis of the terminal 1,3− and 1,6−linked alpha−d−mannose residues in the mannosyl− oligosaharide man(5)(glcnac)(3). pathway: glycosylation. subunit: homodimer, disulfide linked. subcellular location: type ii membrane protein. golgi. tissue specificity: all tissues, mostly in adrenal and thymus. similarity: belongs to family 38 of glycosyl hydrolases.
MM_TBI8372.1	96241_at 96242_at	1.00 1.01	Mm_TBI8372.1	4.2		
MM_TBI2562.3	93095_at	1.01	HMG1_MOUSE	2.3	"X80457:HMG1 high mobility group protein"	
MM_TBI106873.1	102174_f_at 93124_at	0.30 0.01	PEBP_MOUSE	1.7	"U43206:phosphatidylethanolamine binding protein"	
MM_TBI1351.1	103262_at	1.00	HIPK1	5.5 7.9	AF077658:homeodomain−interacting protein kinase 1 mRNA	
MM_TBI2122.93	99202_at	0.92	EF1G_HUMAN	4.8	AV221082: cDNA	
MM_TBI9185.1	97321_at	0.82	Mm_TBI9185.1	3.3	AW124201: cDNA	
MM_TBI3745.1	98139_at	0.40	POR1_MOUSE	42.1	U30840:Voltage−dependent anion channel 1	POR1_MOUSE. function: forms a channel through the mitochondrial outer membrane and also the plasma membrane. the channel allows diffusion of small hydrophilic molecules; it adopts an open conformation at low or zero membrane potential and a closed conformation at potentials above 30−40 mv. the open state has a weak anion selectivity whereas the closed state is cation−selective. subcellular location: mitochondrial vdac1 (mt−vdac1) in outer membrane of mitochondria and plasmalemmal vdac1 (pl−vdac1) in plasma membrane. alternative products: 2 isoforms; pl−vdac1 (shown here) and mt− are produced by alternative splicing. tissue specificity: high levels of expression detected in heart, kidney, brain, and skeletal muscle. not expressed in testis. domain: consists mainly of membrane−spanning sided beta−sheets. similarity: belongs to the eukaryotic mitochondrial porin family.
L38424	DapX−5_at	0.18	ACTG2	2.0 2.9		
Genes found by SSH with unclear result on Affymetrix chips and on Nylon CREM SSH arrays						
MM_TBI16864.1	93655_at 93656_g_at	1.35 0.78	Mm_TBI16864.1	0.5	X95316:USF1 (exons 2 to 10)	
MM_TBI17365.3	99128_at	1.29	ATPO_HUMAN	1.6	AI849767: cDNA	
MM_TBI2757.2	93797_g_at	1.28	ATP1A4	5.2 7.8 2.3		
MM_TBI3602.1	99982_at	1.28	Nfkbia		U19799:IkB−beta mRNA	
MM_TBI7528.1	95707_at	1.24	Mm_TBI7528.1	1.8		
MM_TBI11270.5	97882_at 97883_s_at	1.24 2.86	S61A_CANFA	1.5	"AF145253:Sec61 alpha isoform 1"	
MM_TBI12651.1	104025_at	1.23	MEPD_RAT	1.8	AW047185: cDNA	
MM_TBI19574.1	99173_s_at 99174_r_at	1.21 2.17	HS19878	2.4	"AJ400622:tomoregulin−1"	
MM_TBI10593.3	100037_at 100038_at	1.18 8.35	Mm_TBI10593.1	2.3	AI648005:uk39a11.x1 cDNA	
MM_TBI10971.1	95760_at	1.18	Mm_TBI10971.1	3.9 1.4		
MM_TBI8860.1	103717_at 93345_at	1.17 3.10	WWP2	0.2	AA921411:vz37e01.r1 cDNA	
MM_TBI12888.3	99796_f_at	1.16	MSH3_MOUSE	0.9	"M80360:Rep−3 protein"	
MM_TBI174.45	101213_at 102473_at	1.14 1.00	RLA0_MOUSE	1.0	X15267:acidic ribosomal phosophoprotein PO	RLA0_MOUSE. function: ribosomal protein p0 is the functional equivalent of e.coli protein l10. subunit: p0 forms a pentameric complex by interaction with dimers of p1 and p2. similarity: belongs to the l10p family of ribosomal proteins.
MM_TBI14428.2	96531_at	1.14	Mm_TBI39875.1		AA153773:mr77h06.r1 cDNA	
MM_TBI3653.1	99474_at	1.13	ADAM5	5.2	U22059:A disintegrin and metalloprotease domain (ADAM) 5	
MM_TBI3018.1	94248_at 95233_f_at	1.09 1.00	AP47_MOUSE	5.5	"AF139394:Ap1m1"	
MM_TBI14425.2	97229_at	1.08	Mm_TBI14425.1	2.3	"AF238866:LNR42"	
MM_TBI7150.1	97395_at	1.08	Mm_TBI7150.1	2.9	AW122465:UI−M−BH2.2−aou−c−10−0−UI.s1 cDNA	
MM_TBI119385.1	95755_at	1.04	YB1_MOUSE		"D14485:dbpA murine homologue"	

Appendix B

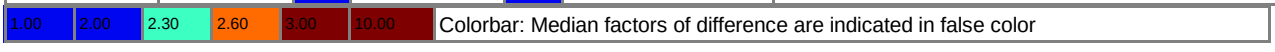
EST cluster	Affymetrix ID	Affy. Diff.	SSH ID	Nylon Diff.	Title	Description
MM_TBI17333.1	96015_at	1.03	Mm_TBI14724.1	7.6	"AF109377:ldlBp"	
MM_TBI9239.1	103416_at	0.95	MAPK6	2.2	"AF132850:extracellular signal−regulated kinase 3"	
MM_TBI787.2	100910_at	0.89	MOF		M14689:surfeit locus surfeit 3 protein gene	
MM_TBI12726.1	99150_at	0.88	DS1_HUMAN		AI844357: cDNA	
MM_TBI7673.3	104321_at	0.80	Mm_TBI7673.1	8.4	AW060175: cDNA	
MM_TBI2858.1	100666_r_at 96542_at	1.00 0.75	SUR4_MOUSE	1.6	M62606:Surfeit gene 4	SUR4_MOUSE. subcellular location: integral membrane protein. endoplasmic reticulum. similarity: belongs to the surf4 family.
MM_TBI16521.2	95494_at	0.71	Mm_TBI16521.1			
MM_TBI16165.1	96228_at	0.67	C11A_HUMAN	1.6	"AF195119:Cyp11a cytochrome P450 side chain cleavage enzyme 11a1"	
MM_TBI9345.1	94953_at	0.61	Mm_TBI22741.1	1.7	"AF079974:Rac GTPase−activating protein"	
MM_TBI1528.1	100861_at 92726_at	0.02 0.42	SOX6_MOUSE	4.8	AJ010605:SRY−box containing gene 6	
MM_TBI3444.1	94834_at	0.34	CATH_MOUSE		"U06119:cathepsin H prepropeptide"	
MM_TBI121619.1	103280_at	0.16	CCAG_RAT	1.0	AF051947:T−type calcium channel alpha−1 subunit mRNA, partial cds	CCAH_MOUSE.
MM_TBI108609.4	101503_at	44.38	DPY3_MOUSE	9.9	X87817:Ulip protein	DPY3_MOUSE. subcellular location: cytoplasmic. similarity: belongs to the dehydropyrimidinase family.
MM_TBI321.1	95062_at	4.92	ICAL_RAT ICAL_HUMAN	2.1	"AB026997:calpastatin"	
MM_TBI9242.2	93930_at	263.9	LAS1_MOUSE	2.3	U58882:LIM and SH3 protein 1	LAS1_MOUSE. similarity: contains 1 lim domain. the lim domain binds 2 zinc ions. similarity: contains 1 sh3 domain.
MM_TBI629.1	98329_at	218.1	9−K13_#0	3.0	X98847:6−phosphofructo−2−kinase/fructose−2,6−bisphosphatase clone 2kbC5	F262_MOUSE. function: synthesis and degradation of fructose 2,6−bisphosphate. catalytic activity: atp + d−fructose 6−phosphate = adp + d−fructose 2,6−bisphosphate. catalytic activity: d−fructose 2,6−bisphosphate + h(2)o = d−fructose 6−phosphate + orthophosphate. enzyme regulation: the most important regulatory mechanism of these opposing activities is by phosphorylation and dephosphorylation of the enzyme (by similarity). subunit: homodimer (by similarity). tissue specificity: highest levels in kidney; also found in heart, brain, spleen, lung, liver, skeletal muscle, and testis. similarity: in the c−terminal section; belongs to the phosphoglycerate mutase family.
MM_TBI2439.1	104048_at	1.30	SYC_HUMAN	1.0	"AB015589:cysteinyl−tRNA synthetase"	
MM_TBI17195.2	100909_at 101986_at	1.00 1.00	MOF		M14689:surfeit locus surfeit 3 protein gene	
MM_TBI2001.1	98585_at	1.00	HIAT1	1.0	D88315:tetracycline transporter−like protein	
MM_TBI1306.1	102753_at	1.00	MEN1_MOUSE	1.0	AB023401:MEN1 menin	MEN1_MOUSE. function: not known.subcellular location: nuclear (by similarity).tissue specificity: ubiquitous. expressed at high level in testisand cns.
MM_TBI6651.1	100268_at 97267_at	1.00 1.00	Mm_TBI17174.1		AI844771:cDNA	
MM_TBI6091.1	101193_at	1.00	Zik1	3.5	U69133:Zik1 mRNA	
MM_TBI2243.1	103847_at	1.00	MMUNKNM	11.7 3.9	L04848:(clone BALB13N) pseudogene mRNA	
MM_TBI6434.1	96323_at	1.00	Mm_TBI6434.1	3.1	"AF145288:Lkb1"	
MM_TBI2147.3	100135_at	0.81	Mm_TBI9177.1	1.0	AI048434: cDNA	
MM_TBI6712.1	94019_at	0.79	Mm_TBI6712.1	0.6		
MM_TBI279.1	96075_at	0.75	WDR1_MOUSE	2.7	"AF020055:Wdr1 protein"	WDR1_MOUSE. function: induces disassembly of actin filaments in conjunctionwith adf/cofilin family proteins (by similarity).similarity: contains 11 wd repeats (trp−asp domains).similarity: belongs to the aip1 family of wd−repeat proteins.
MM_TBI25038.1	93114_at	0.69	Mm_TBI12024.1		"AI843947:cDNA"	
MM_TBI11278.6	96848_at	0.68	Mm_TBI11278.1			
MM_TBI1160.3	102954_at	0.13	SOX5_MOUSE	8.7	AJ010604:transcription factor L−Sox5	
Genes found by SSH with non−differential expression						
MM_TBI2495.1	102455_at 98523_at	1.16 0.93	Rps29	1.2	L31609:(clone mcori−1ck9) S29 ribosomal protein mRNA	RS29_HUMAN. similarity: belongs to the s14p family of ribosomal proteins.
MM_TBI2798.1	99134_at	1.08	TCX2_MOUSE	1.0	U21673:T−complex−associated testis expressed 3	TCX2_MOUSE. function: candidate for involvement in male sterility. subcellular location: membrane associated. alternative products: 2 isoforms; a long form (shown here) and a short form; are produced by alternative splicing. the long form is more abundant by a factor of more than ten. tissue specificity: found exclusively on the surface of sperm tail. it is stored in cytoplasmic granules during spermatogenesis. disease: could be involved in transmission ratio distortion (trd) in mouse t−haplotype which causes male sterility.

Appendix B

EST cluster	Affymetrix ID	Affy. Diff.	SSH ID	Nylon Diff.	Title	Description
MM_TBI1932.1	101482_at	1.01	PP1G_MOUSE	1.1 1.5	D85137:PP1gamma	PP1G_MOUSE. function: protein phosphatase 1 (pp1) is essential for cell division, it participates in the regulation of glycogen metabolism, muscle contractility and protein synthesis. involved in regulation of ionic conductances and long−term synaptic plasticity. may play an important role in dephosphorylating substrates such as the postsynaptic density−associated ca/calmodulin dependent protein kinase ii. catalytic activity: a phosphoprotein + h(2)o = a protein + orthophosphate (this enzyme is serine/threonine specific). subunit: pp1 comprises a catalytic subunit, pp1−alpha, −beta or gamma, which is folded into its native form by inhibitor 2 and glycogen synthetase kinase 3, and then complexed to different targeting subunits. the g subunit binds pp1 to glycogen and the m subunit to myosin. interacts with neurabin−i and neurabin−ii. subcellular location: cytoplasmic. alternative products: 2 isoforms; gamma−1 (shown here) and gamma− 2; are produced by alternative splicing. similarity: belongs to the ppp family of phosphatases. pp−1 subfamily.
MM_TBI2738.1	99816_at	0.99	HS72_MOUSE	1.5 1.1	M20567:Heat shock protein, 70 kDa 2	HS72_MOUSE. function: in cooperation with other chaperones, hsp70s stabilize preexistent proteins against aggregation and mediate the folding of newly translated polypeptides in the cytosol as well as within organelles. these chaperones participate in all these processes through their ability to recognize nonnative conformations of other proteins. they bind extended peptide segments with a net hydrophobic character exposed by polypeptides during translation and membrane translocation, or following stress−induced damage. developmental stage: specifically expressed in prophage stage of meiosis. similarity: belongs to the heat shock protein 70 family.
MM_TBI216.1	96099_at	0.96	KC2B_HUMAN	0.9	"X80685:gMCK2−beta protein kinase"	
MM_TBI18073.2	94007_at	0.94	Mm_TBI492.2	1.3		
MM_TBI4244.1	98128_at	0.93	ATPR_MOUSE	1.1	U77128:mitochondrial ATP synthase coupling factor 6 mRNA, nuclear gene encoding mitochondrial protein	ATPR_MOUSE. function: this is one of the chains of the nonenzymatic component (cf(0) subunit) of the mitochondrial atpase complex. f6 seems to be part of the stalk that links cf(0) to cf(1). subunit: f−type atpases have 2 components, cf(1) − the catalytic core − and cf(0) − the membrane proton channel. cf(0) seems to have nine subunits: a, b, c, d, e, f, g, f6 and 8 (or a6l).
MM_TBI1231.1	103395_at 97467_at	0.92 1.00	SGCA	3.5 0.7	AF019564:adhalin mRNA	
MM_TBI59581.1	98524_f_at	0.82	Rps29	1.2	L31609:(clone mcori−1ck9) S29 ribosomal protein mRNA	RS29_HUMAN. similarity: belongs to the s14p family of ribosomal proteins.
MM_TBI1863.61	101854_r_at 93921_at 93922_g_at	0.99 1.13 1.26	EIF−4G		AI573601: cDNA	
MM_TBI4091.1	95448_at	1.17	PRS7_HUMAN	2.9	"U61283:MSS1"	
MM_TBI10395.3	99599_s_at 99600_at	1.15 0.68	Gcap3	1.5	AW210320: cDNA	
MM_TBI5605.4	92490_at	1.07	KIF9	0.6	AJ132889:kinesin like protein 9	
MM_TBI3370.1	93567_at	1.06	MM17324	0.5	"AJ272203:profilin II"	
MM_TBI8723.1	94478_at	1.05	Mm_TBI24250.1	4.9		
MM_TBI2409.1	101249_at 98153_at 98446_s_at	1.04 0.97 1.02	TCPG_MOUSE	4.3	L20509:Chaperonin subunit 3 (gamma)	TCPG_MOUSE. function: molecular chaperone; assist the folding of proteins upon atp hydrolysis. known to play a role, in vitro, in the folding of actin and tubulin. subunit: hetero−oligomeric complex of about 850 to 900 kda that forms two stacked rings, 12 to 16 nm in diameter. subcellular location: cytoplasmic. ptm: the n−terminus is blocked. similarity: belongs to the tcp−1 chaperonin family.
MM_TBI28589.1	100126_at	1.02	Mm_TBI23479.1	3.1	"AF230805: NF−YC−like"	
MM_TBI2350.1	94052_at 94614_at	1.02 0.80	DPM2_RAT	0.8	"AB013360:DPM2"	DPM2_MOUSE. function: regulates the biosynthesis of dolichol phosphate−mannose. essential for the er localization and stable expressionof dpm1.subunit: interacts with dpm1.subcellular location: integral membrane protein. endoplasmicreticulum.similarity: belongs to the dpm2 family.
MM_TBI1714.26	102129_at 99590_at	1.00 1.01	RS17_CRIGR	3.1	D25213:rpS17 ribosomal protein S17	RS17_CRIGR. similarity: belongs to the s17e family of ribosomal proteins.
MM_TBI12683.1	98084_at	1.00	Hs9552.3		AI849834: cDNA	
MM_TBI1844.1	101448_at	1.00	HGS	6.6	D50050:HGF−regulated tyrosine kinase substrate	
MM_TBI214.1	93810_at	0.98	CATD_MOUSE	4.4	"X68378:cathepsin d"	
MM_TBI6847.1	95573_at	0.98	Mm_TBI6847.1	2.3	AW122821: cDNA	
MM_TBI392.3	100093_at	0.98	KPT1_MOUSE		X69025:PCTAIRE−motif protein kinase 1	KPT1_MOUSE. function: may play a role in signal transduction cascades in terminally differentiated cells. alternative products: 2 isoforms; a long form (shown here) and a short form; are produced by alternative splicing. tissue specificity: ubiquitous with highest levels in testis and brain, with longer form predominant in all tissues except the testis. similarity: belongs to the ser/thr family of protein kinases. cdc2/cdkx subfamily.
MM_TBI9878.2	93488_at	0.96	Mm_TBI16333.1	8.7	"AF148321:serine racemase"	
MM_TBI14008.1	100561_at 93850_at	0.96 0.99	IQGA	1.0 1.0	"AF240630:IQ motif containing GTPase activating protein 1"	
MM_TBI6506.1	104126_at	0.94	Mm_TBI6506.1	2.9	AI854864: cDNA	
MM_TBI288.1	94405_at	0.93	NTTA_MOUSE	5.1	"AF020194:retinal taurine transporter"	NTTA_MOUSE. function: required for the uptake of taurine.subcellular location: integral membrane protein.tissue specificity: retinal.similarity: belongs to the sodium:neurotransmitter symporterfamily (snf).
MM_TBI1574.1	99950_at	0.93	TF2D_MOUSE	2.7	D01034:TATA box binding protein	TF2D_MOUSE. function: general factor that plays a major role in the activation of eukaryotic genes transcribed by rna polymerase ii. tfiid binds specifically to the tata box promoter element which lies close to the position of transcription initiation. subunit: binds dna as a monomer. subcellular location: nuclear. similarity: the c−terminal 180 residues are extremely well conserved in all eukaryotic tfiid. similarity: weak, with bacterial polymerase sigma−factors.
MM_TBI1353.1	103233_at	0.92	HIPK3	1.6	AF077660:homeodomain−interacting protein kinase 3 mRNA	

Appendix B

EST cluster	Affymetrix ID	Affy. Diff.	SSH ID	Nylon Diff.	Title	Description
MM_TBI2834.11	93121_at	0.92	RS24_HUMAN	3.0	"X60289:ribosomal protein S24"	
MM_TBI1519.1	101542_f_at 101785_f_at 93200_f_at 93309_at	0.49 1.00 0.92 0.97	DDX3_HUMAN	2.1	L25126:D−E−A−D (aspartate−glutamate−alanine−aspartate) box polypeptide 3	DDX3_MOUSE. function: putative atp−dependent rna helicase. it may play a role in translational activation of mrna in the oocyte and early embryo. tissue specificity: developmentally regulated. developmental stage: expressed in oocytes. ubiquitously found in 9 days post−conception embryo, at later stages it is restricted to brain and kidney. similarity: belongs to the "dead" box family helicase. ddx3 subfamily.
MM_TBI17638.3	102137_f_at 93839_at	0.81 0.91	Mm_TBI17638.1		AI845856: cDNA	
MM_TBI8568.1	98061_at	0.91	Mm_TBI8568.1	1.0	AI841192: cDNA	
MM_TBI6693.1	99988_at	0.90	Mm_TBI6693.1	2.5	AW122115: cDNA	
MM_TBI2217.1	103387_at	0.89	Tctex−3	6.6	AB011550:T−complex testis−expressed 3	PHF1_MOUSE. subcellular location: nuclear (potential).tissue specificity: testis−specific.similarity: contains 2 phd zinc−finger domains.
MM_TBI18746.1	94477_at	0.89	MMRNA1			
MM_TBI19657.1	94005_at	0.88	MPPB_HUMAN	2.2		
MM_TBI4162.8	103164_i_at 103165_f_at 96575_at	1.00 1.00 0.88	RL8_HUMAN	1.0	U67771:ribosomal protein L8 (RPL8) mRNA	RL8_HUMAN. subcellular location: cytoplasmic. similarity: belongs to the l2p family of ribosomal proteins.
MM_TBI16530.2	93993_at	0.87	Mm_TBI16530.1	4.4		
MM_TBI15247.3	94088_at	0.87	PTB	1.0	"AF095718:RRM−type RNA−binding protein brPTB"	
MM_TBI2340.6	101087_r_at 101088_f_at	0.07 0.87	CNBP_MOUSE	7.5 2.6	X63866:cellular nucleic acid binding protein	CNBP_MOUSE. function: single stranded dna−binding protein, with specificity to the sterol regulatory element (sre). cnbp is involved in sterol− mediated repression. subcellular location: cytoplasmic, also present in endoplasmic reticulum. tissue specificity: present in all tissues examined. similarity: to s.pombe byr3 and to retroviral nucleic acid binding proteins (nbp).
MM_TBI3692.1	103123_f_at 93362_at 93363_at	0.04 0.86 1.30	AP50_HUMAN	1.0	"AF001913: mu2"	AP50_HUMAN.
MM_TBI6330.1	97311_at	0.86	Mm_TBI6330.1	2.5		
MM_TBI21966.1	102234_at 95406_at	0.50 0.86	Mm_TBI21966.1		AW047207: cDNA	
MM_TBI27287.26	99592_f_at	0.85	RS17_CRIGR	3.1	D25213:rpS17 ribosomal protein S17	RS17_CRIGR. similarity: belongs to the s17e family of ribosomal proteins.
MM_TBI6546.1	102600_f_at 95286_at	0.34 0.84	CLUS_MOUSE	1.0	D14077:Clusterin	CLUS_MOUSE. function: not yet clear, it is known to be expressed in a variety of tissues and it seems to be able to bind to cells, membranes, and hydrophobic proteins. it has been associated with programmed cell death. subunit: antiparallel disulfide−linked heterodimer. tissue specificity: most abundant in stomach, liver, brain, and testis, with intermediate levels in heart, ovary, and kidney. ptm: extensively glycosylated with sulfated n−linked carbohydrates. similarity: belongs to the clusterin family.
MM_TBI2725.1	100380_at	0.83	H3A	8.2	X91866: H3.3A histone	
MM_TBI1054.1	101064_at	0.78	PLRG1	1.6	AF044334:pleiotropic regulator 1 (PLRG1) mRNA	
MM_TBI4098.1	103190_f_at 93069_at	0.67 0.71	UB5B_HUMAN	3.5 1.4	"U62483:ubiquitin conjugating enzyme"	
MM_TBI11380.1	95677_at	0.68	AF083383	1.9		
MM_TBI2049.1	101995_at	0.66	OSF−6	0.3	U40930:oxidative stress−induced protein mRNA	
MM_TBI9242.1	93793_at	0.49	Mm_TBI9242.1	4.8		
MM_TBI18327.2	102191_at 94493_at	2.73 0.48	Mm_TBI914.2	3.5 0.8	"AF095905:CPETR2"	CLD3_MOUSE. function: component of tight junction (tj) strands.subcellular location: integral membrane protein.similarity: belongs to the claudin family.
MM_TBI2792.2	100249_f_at 93516_at	1.00 4.09	DYLX_MOUSE	1.2	"tctex−1"	
MM_TBI12888.22	99627_r_at	0.98	RL38	1.3	AA638667:vo56a09.r1 cDNA	
MM_TBI119410	100776_at	0.57	MmTBI8810.1	1.0	AI117463:ub88c08.r1 cDNA	



Appendix C: Expression Profiles of expressed CREM SSH library clones

The expression profiles of selected clones, spotted from the CREM SSH library on nylon cDNA arrays, are shown in the table below. Only those genes are listed with measured intensity values significantly higher than background (> 15) in at least one of the conditions.

The columns in the table represent experimental conditions, i.e. testes of young mice (between 9−27 day old) as well as adult mice, wild−type and CREM (−/−) mutant, were used for hybridization. The measured radiation intensities were standardized based on a pool of spotted housekeeping genes (see Section Results−Standardization). The numbers provided in the table reflect medians of the intensity values of 2−4 repetitions.

Each row in the table represents a gene or clone from the CREM SSH library, the gene names were annotated by homology. The genes were classified into functional groups and the functional groups are visualized in the table by the color of the title of the gene, i.e.:
Axon Guidance, Cell Junction, Cytoskeleton+Motility, Sperm Structure, Crossmembrane Transport, Intracellular Transport, Metabolic Enzymes, Energy Metabolism, Energy Transduction, Erythrocyte Membrane, Signal Transduction, Signal Transmission, Cell Cycle Regulator, Meiosis, Mitosis, DNA Repair, Histones+HMGs, Transcription Factor, RNA Modification, Translation, Molecular Chaperone, Protein Degradation, Protein Modification, Protein Transport

The rows are sorted according to the shape of the profiles. The sum of distances of all neighboring rows was minimized by estimating the solution to a "Traveling Salesman Problem". The shape of the profile is visualized by the background colors in the table. The color code indicates the logarithmic expression level divided by the row−sum, i.e. the total sum of the profile, for each gene:

0.00	0.05	0.10	0.15	0.20	0.25	0.30	0.50	1.00	Colorbar: Normalized Expression Levels on 0−1 scale are color coded

9	17	19	21	23	25	27	wt	mut	Gene	Title	Clone
1	1	1	3	1	1	4	16	1	UBC3_HUMAN	Homo sapiens ubiquitin conjugating enzyme (UBC3)	9−H22
1	3	1	2	4	1	4	15	11		RSA−27−M02	27−M02
1	5	2	5	4	5	16	10	4	T2FA_HUMAN	TRANSCRIPTION INITIATION FACTOR RAP74 (T2FA)	10−E09
2	5	4	6	9	9	26	40	5	RAN_MOUSE	GTP−BINDING NUCLEAR PROTEIN (RAN)	2−I02
2	6	3	9	14	16	28	35	5	PGMU_HUMAN	Human phosphoglucomutase 1 (PGM1)	18−L06
2	4	3	8	7	12	13	20	1	AV258083	testis cDNA clone:4922501M16	7−B09
2	3	2	4	5	8	7	17	2	HIPK1	Homeodomain−interacting protein kinase 1 (HIPK1)	1−K15
2	5	2	7	8	13	9	18	10	Odf2	Outer dense fiber protein (Odf2)	3−D13
2	3	1	7	10	18	6	10	1	KPCD_MOUSE	Protein kinase C delta (KPCD)	7−A19
3	2	2	10	10	22	11	31	3	Mm_TBI78124.1	Mm_TBI78124.1	24−H19
3	3	3	14	14	59	44	72	1	Mm_TBI84804.1	Mm_TBI84804.1	24−H17
5	3	3	7	12	46	39	24	13	Mm_TBI13459.1	Mm_TBI13459.1	19−E07
4	3	4	8	18	65	52	46	3		RSA−5−L20	5−L20
4	3	5	7	25	63	46	15	1			16−H08
2	2	2	5	13	20	15	24	1	MAN2_MOUSE	Alpha−mannosidase II (MAN2)	21−B14
3	5	5	11	39	49	40	26	1			22−L09
4	12	9	24	48	143	116	99	10	Odf2	Outer dense fiber protein (Odf2)	
3	16	7	46	87	186	241	398	19	DBI5_MOUSE	DIAZEPAM BINDING INHIBITOR−LIKE 5	
2	5	5	11	18	84	89	67	3		RSA−13−C05	
3	5	6	7	8	47	43	71	6	Hs116038.1	Hs116038.1	
2	4	4	5	6	21	24	36	3	EST66525	Homo sapiens EST66525	
3	5	4	6	9	33	26	20	4	LAP1C	Rattus norvegicus lamina−associated polypeptide 1C (LAP1C)	
3	5	5	4	6	24	25	8	2		RSA−13−O08_#0	
3	7	7	6	8	33	37	63	7	Mm_TBI50354.1	Mm_TBI50354.1	
3	6	7	4	9	16	20	34	3	RCC_HUMAN	Human REGULATOR OF CHROMOSOME CONDENSATION (RCC1)	
5	19	14	12	24	86	87	102	9	RL28_MOUSE	Ribosomal protein L28	
3	10	8	12	16	44	47	54	3	Mm_TBI87042.1	Mm_TBI87042.1	
4	11	8	16	20	75	88	84	8	HPI31	Homo sapiens proteasome inhibitor hPI31	

Appendix C

9	17	19	21	23	25	27	wt	mut	Gene	Title	Clone
6	21	23	44	44	153	140	93	48	MmTBI16020.1	Mm_TBI16020.1	
13	44	40	74	109	193	223	161	30	RAN_MOUSE	Ran GTPase525 bp long incert RTPCR−cloned from testis	
10	44	42	72	124	207	157	87	29	HSJ3_MOUSE	Testis specific DNAj−homolog 3 (HSJ3)	29−O10
10	90	129	383	666	710	568	415	252	PGK2	Phosphoglycerate kinase 2 (pgk−2)	+2−A5RZ
10	83	94	173	185	206	145	68	71	KC2B_HUMAN	Casein kinase II beta subunit (KC2B)	17−K18
8	33	39	50	45	56	54	44	34	Mm_TBI492.2	Tex189	3−K11
11	47	54	69	103	166	103	93	45	Mm_TBI13618.1	Mm_TBI13618.1	23−A06
27	144	147	245	333	332	410	244	140	CREM_MOUSE	P171233 CREM cAMP responsive element modulator	+1−D5RZ
94	2348	2296	5020	7138	7292	6400	2527	2440	TCX2_MOUSE	T complex testis−specific protein 2 (Tctex2), expression: from leptotene to late spermatides	+1−D3RZ
41	287	285	473	347	419	451	173	114	PP1G_MOUSE	PP1c−gamma RTPCR−cloned from testis	PP1cg
79	977	1003	977	811	914	992	73	101	SGCA	Alpha−sarcoglycan gene	18−G05
16	194	175	162	127	137	165	26	25	TBA1_MOUSE	Alpha−tubulin gene (M−alpha−1)	12−N02
23	177	144	187	161	164	194	210	173	DYLX_MOUSE	CYTOPLASMIC DYNEIN LIGHT CHAIN (TCTEX−1)	17−A16
45	160	180	164	153	193	200	142	82	PEBP_MOUSE	Phosphatidylethanolamine binding protein (PEBP)	19−P14
47	153	181	167	166	149	242	169	77	TBB5_Mouse	Tubulin beta−5 chain	22−A01
147	747	662	356	411	237	490	171	62		mice DNA 1.8 mkg/mkl	mice
42	95	58	54	63	76	88	154	89	CALU_MOUSE	Mus musculus calumenin mRNA, complete cds.	4−C5RZ
91	98	84	80	79	78	68	108	77		EST clone 5−B2RZ	5−B2RZ
313	238	169	115	92	76	92	85	73	Q9QWJ3	Mouse alpha−1−globin mRNA, 5’ end. Length = 373	3−D5RZ
92	78	63	61	56	49	43	57	54		EST clone 4−B3RZ	4−B3RZ
65	65	66	67	65	42	65	71	93	CACYBP	U97327 Calcyclin binding protein (CACYBP)	4−C3RZ
136	145	177	189	159	129	170	97	97		EST clone 4−D3RZ	4−D3RZ
763	687	912	1163	758	828	1027	126	221		mice DNA 1.8 mkg/mkl	mice
116	170	145	215	254	251	195	352	60		HSPANCAN Homo sapiens pancreatic tumor−related protein	4−B1RZ
62	94	120	131	147	151	137	64	59	ATPR_MOUSE	Mitochondrial ATP synthase coupling factor 6 (ATPR)	
109	158	250	203	272	233	281	219	82	HS7T_MOUS	D85732 Hsc70t spermatid−specific heat shock protein 70,	
96	74	164	150	135	210	182	121	81	TRAF6	mTRAF6 PCR product	
118	152	172	136	126	282	204	430	62	ATPA_MOUSE	ATP SYNTHASE ALPHA CHAIN, MITOCHONDRIAL PRECURSOR	
90	89	94	95	91	241	182	355	58	AT91_MOUSE	ATP synthase, H+ transporting, mitochondrial F0 complex, subunit c	
124	45	68	75	70	150	100	53	56	AP1_MOUSE	Jun oncogene TRANSCRIPTION FACTOR AP−1	
116	81	87	83	70	135	85	56	45	Rps29	Ribosomal protein S29	
85	61	52	43	26	71	52	89	32		RSA−8−P21	
44	37	28	24	23	40	40	60	44	RRAS_MOUSE	Mouse R−ras mRNA, complete cds. Length = 949	
30	18	18	19	21	36	35	74	24	ACE_MOUSE	Angiotensin−converting enzyme (ACE)	
44	42	41	31	40	64	54	105	35		HSU60800 Human semaphorin (CD100)	
67	88	71	66	69	109	88	136	50	BAG1_MOUSE	BAG−1 RTPCR	BAG−1
37	53	56	48	60	89	71	92	32	Mm_TBI14951.1	Mm_TBI14951.1	13−P13
44	46	50	46	53	78	70	169	88		EST clone 5−A3RZ	5−A3RZ
70	64	73	65	85	118	112	84	53	IMD1_MOUSE	Mouse IMP dehydrogenase mRNA, complete cds.	3−C1RZ
104	204	190	208	243	441	378	237	176		EST clone 4−C4RZ	4−C4RZ
73	128	125	230	246	192	268	211	145	MY88_MOUSE	MyD88 PCR product	MyD88
60	99	142	214	244	281	261	235	195		Homo sapiens chromosome 17, clone HCIT75G16,	5−D1RZ
73	129	126	176	319	948	549	1530	209	DUS2_MOUSE	Trosine−threonine dual specificity phosphatase PAC−1	4−D6RZ
102	110	95	94	99	416	354	881	36	MTDC_MOUSE	NAD−dependent methylenetetrahydrofolate dehydrogenase−methenyltetrahydrofolate cyclohydrolase	5−C1RZ
54	42	47	40	72	180	191	468	66	THIK	3−ketoacyl−CoA thiolase peroxisomal B precursor	1−B1RZ
42	40	47	33	68	96	91	166	42	PGK1_MOUSE	B161139 PHOSPHOGLYCERATE KINASE 1	1−D1RZ
21	16	16	20	29	42	54	63	15	Mm_TBI20710.1	Mm_TBI20710.1	21−A02
13	10	11	11	17	18	25	28	9	MmLATS2	Warts/lats−like kinase (MmLATS2)	22−P24
12	11	10	11	19	21	28	39	17	SPK_HUMAN	Homo sapiens Symplekin	23−D11
11	16	10	12	18	24	21	33	17		EST clone 4−C6RZ	4−C6RZ
13	14	15	15	24	25	19	5	7	KKIT_MOUSE	Kit oncogene stem cell growth factor receptor precursor	+2−C6aR
15	20	23	18	20	21	30	17	1	28SRNA	Homo sapiens 28S ribosomal RNA gene	6−O05
13	26	25	18	19	22	27	11	5	Mm_TBI14425.1	Mm_TBI14425.1	18−I11
11	48	37	24	16	28	39	58	44	STA4_MOUSE	BALB/c gamma interferon activation site−binding protein	7−I05

9	17	19	21	23	25	27	wt	mut	Gene	Title	Clone
7	20	17	8	8	16	20	5	1	Mm_TBI474.5	Mm_TBI474.5	3−K02
7	12	11	7	5	18	21	58	15	LHR	LH receptor RTPCR−cloned	LHrec
5	7	10	5	4	15	19	45	7	HU−PP−1	Homo sapiens PROTEIN−TYROSINE PHOSPHATASE	3−J03
6	9	7	5	6	43	33	28	2	Cdyl	Testis−specific chromodomain Y−like protein (Cdyl)	15−G03
4	5	6	3	6	30	26	45	3	MMP40GPRT	G protein−coupled receptor (P40GPRT)	15−M23
3	13	9	3	9	69	43	43	5	Mm_TBI28224.1	Mus musculus cDNA clone 482300	30−C24
3	3	4	2	3	17	15	14	2	Mm_TBI75907.1	Mm_TBI75907.1	11−E11
3	7	7	2	2	26	14	22	1		RSA−29−G11	29−G11
7	6	7	3	2	30	24	23	5	AV271345	testis cDNA clone:4931408C20	20−K13
14	9	9	5	7	55	46	33	5	Hs97726.1	Hs97726.1	11−F23
11	7	7	5	7	24	25	36	10		RSA−22−A06	22−A06
14	6	8	6	11	54	33	48	3	HS7T_MOUSE	Spermatid−specific heat shock protein70 (Hsc70t)	7−G01
7	3	5	3	8	33	13	11	5	Lfc1	Lfc1 RTPCR−cloned	Lfc1
13	6	5	3	9	49	24	25	1	SERA_MOUSE	Rattus norvegicus D−3−phosphoglycerate dehydrogenase	12−P18
13	10	8	2	9	108	29	48	10	PIM1_MOUSE	Pim−1 RTPCR−cloned	
5	3	3	2	3	10	11	18	3	Mm_TBI8139.1	Mm_TBI8139.1	
3	2	3	1	4	10	13	39	5	Mm_TBI84802.1	Mm_TBI84802.1	
3	3	3	2	10	15	10	17	3	FSC1_MOUSE	Major fibrous sheath protein (FSC1)	
2	3	2	1	7	19	15	24	1	BB017947	testis cDNA clone 4930571K03	
1	3	1	1	6	10	21	33	1	SKP2	Homo sapiens cyclin A/CDK2−associated p45 (Skp2)	
2	2	2	1	5	22	20	30	1	Mm_TBI13980.1	Mm_TBI13980.1	
4	3	2	1	2	20	16	21	1	Mm_TBI64114.1	Mm_TBI64114.1	
7	6	3	1	1	45	47	351	5	STP1_MOUSE	SPERMATID NUCLEAR TRANSITION PROTEIN 1 (STP−1)	
1	2	2	1	1	7	8	24	1	AF1Q_MOUSE	AF1Q gene	
1	2	2	1	1	28	17	65	1	STP2_MOUSE	Transition protein 2 (TP2)	
2	1	6	1	1	21	8	28	1	BB012379	testis cDNA clone:4930448I04,	
1	1	2	1	1	4	2	18	2	Hs152818.2	Hs152818.2	2−K17
1	1	5	2	2	12	10	17	2		RSA−6−K05	6−K05
1	1	2	2	3	11	6	17	1	Mm_TBI13849.1	Mm_TBI13849.1	4−G07
1	1	2	3	14	25	36	23	1		RSA−32−F10	32−F10
1	1	1	2	8	16	11	33	1	D−AKAP1	Dual specificity A−kinase anchoring protein 1 (D−AKAP1)	2−M19
1	1	1	2	3	4	3	25	1	RTR	Orphan receptor RTR	9−I04
1	1	1	1	5	10	5	30	1		RSA−10−J12	10−J12
1	1	1	1	8	14	9	16	1	Hs136317.1	Hs136317.1	2−H13
1	1	1	1	8	21	15	12	1		RSA−6−O21	6−O21
1	1	1	2	6	22	20	61	3		RSA−17−N19	17−N19
1	1	1	2	3	11	17	8	1	Dp1l1	Polyposis locus protein 1−like 1	8−P01
1	1	1	3	6	23	18	19	2	Mm_TBI47943.1	Mm_TBI47943.1	3−A12
1	1	1	16	31	102	101	22	1		RSA−3−H18	3−H18
1	1	1	6	11	9	17	23	8	Mm_TBI47943.1	Mm_TBI47943.1	3−N17
1	1	1	27	4	5	4	1	1		RSA−4−P22	4−P22
1	1	1	25	8	40	20	6	1		RSA−2−P20	2−P20
1	1	1	4	4	13	16	12	1	Mm_TBI22387.1	Mm_TBI22387.1	16−N15
1	1	1	4	2	12	12	27	16	GPDM_MOUSE	Glycerol−3−phosphate dehydrogenase (GDPM)	4−E21
1	1	1	3	2	43	23	44	2	Mm_TBI10958.1	Mm_TBI10958.1	
1	1	1	2	1	14	18	65	1	ANX2_MOUSE	ANNEXIN II	
1	1	1	1	1	11	15	19	2	CHIO_HUMAN	Homo sapiens beta2−chimaerin	
1	1	1	1	1	17	16	11	1		RSA−3−C23	
1	1	1	1	1	8	7	22	1	Mm_TBI23178.1	Mm_TBI23178.1	
1	1	1	1	2	19	15	22	1	Mm_TBI64097.1	Mm_TBI64097.1	
1	1	1	1	2	8	5	16	2	GLNA_MOUSE	Glutamate−ammonia ligase (GLNA)	
1	1	1	1	5	13	14	18	1	MDJ6	HSP40/DNAJ homolog: MDJ6	
3	1	2	1	13	53	33	79	2	ODFP_MOUSE	Outer dense fiber protein of sperm tails (Odf1)	
2	1	2	1	3	9	12	46	3	KIAA0231	Human KIAA0231 gene	
7	1	2	1	4	23	13	16	2	DDC8	Testis−specific protein (DDC8)	
3	1	1	2	5	12	10	16	2	Mlark	RNA BINDING MOTIF PROTEIN 4 (Mlark)	

Appendix C

9	17	19	21	23	25	27	wt	mut	Gene	Title	Clone
5	1	5	5	14	55	36	30	1	KAP2_MOUSE	cAMP−dependent protein kinase type II (KAP2)	
4	1	5	3	6	31	31	23	3	Mm_TBI64087.1	Mm_TBI64087.1	
1	1	3	2	2	9	7	24	1	KAP2_MOUSE	cAMP−dependent protein kinase type II (KAP2)	
2	1	4	3	6	47	24	13	1		RSA−2−A08	6−N01
2	1	4	11	17	26	13	17	2	FRT1_MOUSE	Proto−oncogene Frat1	6−P16
2	1	3	3	12	16	10	12	5		RSA−29−I09	29−I09
3	1	3	4	9	33	14	6	1	HS1187176	Homo sapiens cDNA clone IMAGE:704776	3−M08
3	2	5	5	14	29	13	13	3		RSA−26−M05	26−M05
3	2	5	2	8	27	13	17	1	Mm_TBI42599.1	Mm_TBI42599.1	21−G19
3	3	4	3	6	31	20	17	2	G6PI_MOUSE	Glucose phosphate isomerase (G5PI)	32−E10
3	3	6	4	9	48	37	44	1	MDJ6	HSP40/DNAJ homolog: MDJ6	3−G03
4	5	7	8	14	101	57	37	3		RSA−13−D05	13−D05
4	5	5	7	15	87	50	33	3		RSA−12−N03	12−N03
4	3	4	4	14	46	36	38	5	LCFB_MOUSE	Long chain fatty acyl CoA synthetase (LCFB).	18−E18
4	3	3	4	15	121	74	68	5	CHIO_HUMAN	beta−chimaerin nucleotides from 156 to 1070 of x69462 RTPCR−cloned	b−chim
8	13	8	16	72	1906	1307	2077	32	STP1_MOUSE	TRANSITION PROTEIN 1	+TP1
29	28	22	51	91	5693	5492	2461	90	HSP2_MOUSE	Protamine 2, expression from round to elongated spermatides	+2−D6RZ
19	47	25	38	134	1715	2786	1971	67	ACE_MOUSE	Angiotensin−converting enzyme	+4−A6RZ
4	7	5	7	12	86	80	106	2		RSA−7−O13_#0	7−O13
3	7	4	7	9	58	60	129	3		RSA−7−O13_#0	7−O13
3	6	3	6	9	82	59	63	2	Hs98947.1	Hs98947.1	9−O24
2	3	2	3	4	18	21	31	3	Mm_TBI23636.1	Mm_TBI23636.1	6−I19
1	2	1	2	3	6	7	18	6	LCFB_MOUSE	Long chain fatty acyl CoA synthetase (LCFB).	9−K22
1	2	2	4	8	29	23	47	1	Mm_TBI84692.1	Mm_TBI84692.1	18−K01
1	4	4	7	18	47	67	117	3	Mm_TBI23176.1	Mm_TBI23176.1	1−C13
1	7	5	7	14	135	207	74	11	Mm_TBI22634.1	Mm_TBI22634.1	29−H22
1	4	4	5	23	57	47	33	2	Gsg3	F−ACTIN CAPPING PROTEIN ALPHA−3 (Gsg3)	4−A20
2	4	4	6	26	58	54	39	1	AV258666	testis cDNA clone:4930401F20	13−C03
2	3	4	5	28	42	26	24	4	Mm_TBI42599.1	Mm_TBI42599.1	19−G11
2	3	5	4	16	30	29	80	3	FSC1_MOUSE	Major fibrous sheath protein (FSC1)	18−C11
2	7	6	5	16	55	48	63	7	Mm_TBI24369.1	Mm_TBI24369.1	19−B16
3	7	6	7	21	50	37	54	2	Mm_TBI23177.1	Mm_TBI23177.1	8−P04
3	9	6	9	25	55	37	46	7	AKAP110	Protein kinase A binding protein AKAP110	5−K07
4	12	6	6	27	53	44	74	5	Mm_TBI15460.1	Mm_TBI15460.1	21−D05
6	10	4	7	23	45	53	159	7	Mm_TBI16245.1	Mm_TBI16245.1	4−N14
4	11	3	8	19	37	47	61	18	HS7T_MOUSE	Spermatid−specific heat shock protein70 (Hsc70t)	7−O24
6	16	4	15	28	43	45	20	3	HS7T_MOUSE	Spermatid−specific heat shock protein70 (Hsc70t)	4−J03
8	14	8	27	52	139	111	62	17		RSA−5−F14	5−F14
12	17	14	95	313	978	1175	852	46	ODFP_MOUSE	Odf−1 outer dense fiber protein 1	+Odf−1
7	6	8	22	45	112	91	71	10	HXK1_MOUSE	Hexokinase	10−N16
10	8	5	26	50	115	69	78	5	Hs136317.1	Hs136317.1	4−P17
8	6	4	11	27	57	63	76	6	Mm_TBI23176.1	Mm_TBI23176.1	24−B12
6	6	5	7	21	43	48	33	5	Mm_TBI51209.1	Mm_TBI51209.1	11−G20
6	5	6	5	15	27	23	28	3		RSA−20−L17	20−L17
9	5	6	7	15	33	21	10	1		RSA−7−A01	3−D10
12	6	4	6	14	35	32	26	6	CALI_HUMAN	Homo sapiens Calicin (CALI)	3−E24
7	5	4	6	14	33	29	24	8		RSA−18−E20	18−E20
13	6	6	12	29	58	67	53	7	tACTIN2	Testis specific actin (tActin2)	23−H06
9	7	5	9	13	31	37	38	13	HPI31	Homo sapiens proteasome inhibitor hPI31	3−J05
10	11	8	12	20	49	55	97	12	Mm_TBI23176.1	Mm_TBI23176.1	8−O17
12	12	9	11	19	35	39	40	9	Mm_TBI32099.1	Mm_TBI32099.1	1−A02
10	11	9	8	20	40	47	63	5	Mm_TBI42622.1	Mm_TBI42622.1	16−I17
13	10	10	9	29	43	37	52	8	AI789539	testis cDNA clone 1745780	12−A09
19	12	13	15	43	57	53	32	9	Hs211912.1	Hs211912.1	
16	11	13	21	40	60	50	45	18	ALFA_MOUSE	Aldolase A	

9	17	19	21	23	25	27	wt	mut	Gene	Title	Clone
9	13	11	20	27	42	38	37	5	DERP2	Mm_TBI14440.2	
6	12	8	12	24	27	26	34	8	Hs106292.1	Hs106292.1	
6	10	9	12	14	21	19	17	4	Mm_TBI7118.1	Mm_TBI7118.1	
9	10	11	10	16	30	25	28	3	Mm_TBI93317.1	Mm_TBI93317.1	
8	10	8	9	10	20	19	30	10	FSC1_MOUSE	Major fibrous sheath protein (FSC1)	
8	7	6	6	7	12	14	28	7	Aoh1	Aldehyde oxidase homolog−1 (Aoh1)	
11	6	10	6	8	16	16	32	7		RSA−23−G15	
13	10	12	11	13	30	19	15	4	Mm_TBI36421.1	Mm_TBI36421.1	
10	8	8	8	12	19	15	13	5	G100_HUMAN	Human Mr 110,000 antigen	
18	12	9	15	16	39	14	13	7	MEPD_RAT	Rat metalloendopeptidase	
9	7	6	10	13	34	18	19	4	GLNA_MOUSE	Glutamate−ammonia ligase (GLNA)	
7	6	5	7	9	19	14	8	2	MMUNKNM	Mm_TBI2242.1	
10	9	7	9	13	33	24	24	6	Mm_TBI13386.1	Mm_TBI13386.1	
12	6	7	8	8	22	19	18	7	EB2	APC−binding protein (EB2)	
34	21	24	21	18	92	60	360	22	KROX−20	Krox20	
56	53	44	25	43	239	162	450	31	TYRP2	MMTYRP2 Tyrosinase−related protein−2 (Tyrp2)	5−C2RZ
21	23	20	8	21	59	30	193	43		Progesteron bind. prot.	ProgBP
12	9	20	6	14	49	18	47	3	KAP3A	KINESIN−ASSOCIATED PROTEIN 3 (KAP3)	8−O15
7	8	9	4	8	16	16	13	6	HSC1LE111	Homo sapiens cDNA clone c−1le11	19−F22
8	8	10	4	5	10	14	43	5		RSA−21−P13	21−P13
8	12	8	5	7	13	14	29	10	Mm_TBI6503.1	Mm_TBI6503.1	21−J23
12	12	9	7	8	18	15	14	6	Mm_TBI6330.1	Mm_TBI6330.1	17−O12
40	25	24	17	14	33	30	50	21	Mm_TBI77934.1	Mm_TBI77934.1	22−N17
26	16	21	11	11	19	26	24	22	MYC_MOUSE	mouse homolog of MYC PROTO−ONCOGENE PROTEIN	+1−C1RZ
29	32	26	8	14	23	26	39	26	COPE_MOUSE	COATOMER EPSILON SUBUNIT (COPE)	1−M21
12	15	17	7	11	13	12	18	11	Transferrin	Transferrin	Transf
14	25	20	12	14	19	16	30	10	RAB6	GTP BINDING PROTEIN ASSOCIATED PROTEIN 1 (Rab6)	
23	24	16	13	10	22	16	35	16	Ariadne	Ariadne protein	
22	20	19	18	12	21	19	19	14	RL38	Rat ribosomal protein L38	
59	41	42	31	20	31	26	32	30	Mm_TBI28775.1	Mm_TBI28775.1	
120	66	44	44	36	54	36	47	58	PGK1_MOUSE	X chromosome−linked phosphoglycerate kinase (pgk−1),	
48	51	32	23	23	42	28	56	67	DUS2_MOUSE	MM09268 Tyrosine−threonine dual specificity phosphatase PAC−1	
48	46	39	36	41	61	34	65	62		no data	
24	28	38	33	32	71	34	24	12	KIAA0788	Homo sapiens KIAA0788 protein	
25	38	27	23	34	58	41	117	62	RL3_MOUSE	MMJ1PRO J1 protein, yeast ribosomal protein L3 homologue	
15	22	22	22	24	45	41	29	16	UNR_RAT	Rat unr protein with unknown function	
12	15	15	16	20	35	33	14	7	Mm_TBI11804.1	Mm_TBI11804.1	
11	15	12	11	18	42	36	32	5	Mm_TBI8299.1	Mm_TBI8299.1	10−L21
18	29	17	17	46	77	66	80	35		RSA−8−P21	7−D16
28	42	26	45	118	157	186	231	107	ACET_MOUSE	ANGIOTENSIN−CONVERTING ENZYME PRECURSOR	+1−B4RZ
40	42	38	47	96	345	292	210	26	G3PT_MOUSE	Gapd−S fragment RTPCR−cloned	+GAPD−S
83	53	61	110	207	1163	1166	771	63	G3PT_MOUSE	testis glyceraldehyde 3−phosphate dehydrogenase (Gapd−S)	+5−A4RZ
29	29	22	19	83	580	309	758	38	CCKN_MOUSE	I103002 CCK gene for cholecystokinin	2−D5RZ
11	9	12	17	29	101	95	42	2	mCAT2	Cationic amino acid transporter (mCAT2)	4−J11
8	6	6	8	13	38	43	35	7	XRP2	Homo sapiens etinitis pigmentosa 2 protein (XRP2)	4−P02
6	7	7	7	12	45	48	60	5	Mm_TBI22932.1	Mm_TBI22932.1	17−E06
4	6	6	5	12	29	27	21	4	BB016512	testis cDNA clone:4930562A21,	20−O09
5	6	8	4	12	30	29	41	4	Mm_TBI15783.1	Mm_TBI15783.1	10−I10
6	7	9	6	12	39	34	28	6	Mlark	RNA BINDING MOTIF PROTEIN 4 (Mlark)	20−N17
6	9	9	7	16	30	41	43	6	MMLEUPS	Mouse leukosialin pseudogene (CD 43)	20−I22
10	18	16	16	28	61	94	76	17	EST215554	Rat EST215554	20−I21
6	11	12	13	18	38	40	20	6	Mm_TBI9185.1	Mm_TBI9185.1	20−E12
5	7	9	10	12	32	36	38	4	Mm_TBI42575.1	Mm_TBI42575.1	7−H15
4	9	6	7	10	20	24	23	6	MmTBI14600.1	Mm_TBI14600.2	
7	13	9	7	11	36	51	75	11	Hs132975.1	Hs132975.1	
14	14	11	9	11	48	51	43	13	AI789539	testis cDNA clone 1745780	

Appendix C

9	17	19	21	23	25	27	wt	mut	Gene	Title	Clone
9	9	6	7	8	34	26	31	11	ACRO_MOUSE	Acrosin gene	
7	6	5	6	9	28	22	35	7	Mm_TBI7445.1	Mm_TBI7445.1	
5	7	5	6	10	28	16	30	2	ODO1_HUMAN	Human 2−oxoglutarate dehydrogenase	
4	4	5	5	8	23	10	6	1	PLAK_MOUSE	Plakoglobin (PLAK)	
4	6	4	4	9	21	13	16	1	GELS_MOUSE	Actin−Depolymerizing factor Gelsolin	
4	6	3	2	8	21	13	11	1	KC12_HUMAN	Homo sapiens casein kinase I γ 2	
7	5	3	3	9	25	6	10	3		RSA−23−A05_#0	
6	3	3	3	13	15	6	3	1	VimC1	Vimentin−binding fragment (VimC1)	
6	2	3	3	12	33	10	30	3	Mm_TBI10958.1	Mm_TBI10958.1	
7	2	2	4	10	20	13	13	4	KIAA0878	Homo sapiens KIAA0878 protein	
5	2	3	5	12	24	17	15	2		RSA−10−E22	
6	3	2	7	11	18	16	16	3	Mm_TBI86656.1	Mm_TBI86656.1	
5	3	2	5	7	14	16	10			RSA−20−L15	
6	4	3	4	10	21	14	27	3	Mm_TBI64114.1	Mm_TBI64114.1	
4	4	2	3	7	24	14	17	2	Mm_TBI51812.1	Mm_TBI51812.1	
2	6	1	2	5	17	19	52	5		RSA−18−A07	
4	9	2	6	9	23	15	11	2	Mm_TBI42937.1	Mm_TBI42937.1	
3	7	2	6	10	19	22	36	4	Mm_TBI13443.1	Mm_TBI13443.1	
2	5	2	6	13	25	35	28	2	Hs189105.1	Hs189105.1	
3	8	3	8	25	94	105	98	5	BB015390	testis cDNA clone 4930553L18	
3	5	3	8	25	58	56	94	2		RSA−2−H02	
3	4	3	9	21	42	63	56	5	Mm_TBI65814.1	Mm_TBI65814.1	
3	5	2	10	26	133	97	77	4	Mm_TBI23177.1	Mm_TBI23177.1	
2	3	2	7	22	81	36	86	34	BB013591	testis cDNA clone:4930469L17,	
3	2		11	27	53	47	31	3	Mm_TBI22300.1	Mm_TBI22300.1	
2	2	1	4	5	19	12	21	1	Y195_HUMAN	Human KIAA0195 gene	
3	4	1	7	6	63	77	55	1	KPCD_MOUSE	Protein kinase C δ (KPCD)	
4	4	1	6	2	20	18	30	4	AK002017	Homo sapiens cDNA FLJ11155	
3	5	1	4	3	15	15	19	4	PLCB_HUMAN	Homo sapiens lysophosphatidic acid acyltransferase	
3	3		6	8	7	9	20	3		RSA−3−E05	3−E05
3	3	1	12	38	44	37	36	3	Mm_TBI86656.1	Mm_TBI86656.1	29−L22
2	3	1	12	46	58	85	91	1	Mm_TBI65814.1	Mm_TBI65814.1	2−H19
2	4	1	5	33	93	88	75	2	Mm_TBI23178.1	Mm_TBI23178.1	7−K03
1	3	1	5	6	22	17	17	4	CHIO_HUMAN	Homo sapiens beta2−chimaerin	3−J07
1	2	1	4	6	15	8	4	2	KC12_HUMAN	Homo sapiens casein kinase I gamma 2	1−J12
1	2	1	5	15	33	15	71	1	AI575885	Rattus norvegicus cDNA clone UI−R−G0−ut−d−01−0−UI	9−K04
1	1	1	3	10	15	9	16	2	Mm_TBI58338.1	Mm_TBI58338.1	6−C10
1	2	1	3	10	22	19	22	2		RSA−14−K15	14−K15
1	2	1	2	9	11	11	18	2	Mm_TBI6748.1	Mm_TBI6748.1	3−D12
1	3	1	2	5	21	36	64	2	Mm_TBI23178.1	Mm_TBI23178.1	7−K03
1	3	1	3	6	24	28	26	1		RSA−16−O04	16−O04
1	4	2	4	13	44	33	10	1		RSA−4−P05_#0	4−P05
1	3	3	3	17	35	29	66	1		RSA−8−P15	8−P15
1	5	4	7	14	45	34	31	1	Mm_TBI42622.1	Mm_TBI42622.1	16−I17
1	4	3	5	15	23	17	13	2	Mm_TBI13386.1	Mm_TBI13386.1	4−A13
1	6	3	4	8	19	9	15	4	BB012305	Testis cDNA clone:4930448B06	7−J15
1	5	3	6	16	29	33	81	60	Mm_TBI84692.1	Mm_TBI84692.1	21−G24
1	11	12	20	40	57	93	46	6		RSA−14−J17	14−J17
2	9	11	21	40	64	69	44	5	Mm_TBI84992.1	Mm_TBI84992.1	14−C07
2	4	7	8	15	21	25	12	1	BB518056	heart cDNA clone:D830027J05	14−N24
2	3	5	5	16	8	11	10	3	ATP1B3	Sodium/Potassium−Transporting ATPase β 3 (ATP1B3)	4−O17
3	3	6	12	13	28	29	17	4	KC12_HUMAN	Homo sapiens casein kinase I gamma 2	4−L13
2	3	6	7	6	21	15	11	1	Mm_TBI87042.1	Mm_TBI87042.1	6−G13
3	3	6	5	7	17	15	16	1	SRPK2	serine/arginine−rich protein specific kinase 2 (SRPK2)	13−J07
5	4	7	6	10	11	15	38	15	BB012329	testis cDNA clone:4930448D20	
2	4	4	4	5	6	8	16	8	ACE_MOUSE	Angiotensin−converting enzyme (ACE)	

9	17	19	21	23	25	27	wt	mut	Gene	Title	Clone
5	10	13	10	12	10	16	28	13	ASSY_MOUSE	Argininosuccinate synthetase (Ass)	
7	17	11	14	15	13	24	29	16		RSA−22−B18	
13	56	32	26	33	20	52	7	15		mice DNA 1:25	
7	31	24	32	28	9	43	5	6		mice DNA 1:25	
2	7	19	16	18	8	24	3	13		RSA−3−O20	
2	9	10	9	19	14	17	35	8	AF039023	Homo sapiens Ran−GTP binding protein	
2	16	11	20	28	30	34	49	16	SNAA_HUMAN	Homo sapiens alpha−soluble NSF attachment protein (SNAP−ALPHA)	
2	19	13	33	34	33	42	13	11	Mm_TBI20749.1	Mm_TBI20749.1	
1	28	23	38	45	42	51	68	23	Mm_TBI13480.1	Mm_TBI13480.1	
1	52	34	36	27	33	31	40	36	HS72_MOUSE	Heat−shock−like protein (HSP70.2)	
1	9	7	12	7	13	17	5	2		RSA−11−P14	
1	11	11	10	8	12	17	15	5			13−J13
1	8	5	5	8	8	8	32	14	MEG1_MOUSE	MEIOSIS EXPRESSED PROTEIN 1 (MEG1)	1−J07
1	12	10	14	17	21	18	10	7		RSA−10−O13_#0	10−O13
1	15	12	22	37	56	51	111	16	HSJ3_MOUSE	Testis specific DNAj−homolog 3 (HSJ3)	14−L03
1	3	8	4	8	20	16	39	4	KAP2_MOUSE	cAMP−dependent protein kinase type II (KAP2)	8−J21
1	3	5	2	5	16	7	10	4	BE192705	cDNA clone R3TAE89,	11−K11
1	2	2	2	2	6	6	20	1	BB014624	testis cDNA clone 4930483L20	10−L18
1	3	2	2	2	12	10	19	1	Mm_TBI21228.1	Mm_TBI21228.1	14−K08
2	7	4	5	1	30	20	45	4	PLCB_HUMAN	Homo sapiens lysophosphatidic acid acyltransferase	4−L09
9	7	4	14	3	112	28	94	3	HSP2_MOUSE	Protamine 2 (mP2)	6−M17
6	6	4	5	2	23	12	19	2	KIAA0940	Homo sapiens KIAA0940	20−I03
8	9	6	6	3	27	22	18	4	Mm_TBI18355.1	Mm_TBI18355.1	15−P17
11	8	7	6	4	29	23	19	6		RSA−19−G05	19−G05
15	7	7	6	3	13	14	25	3	FPPS_HUMAN	Human farnesyl pyrophosphate synthetase (FPPS)	21−O22
30	12	14	11	7	87	37	81	5	TES2_MOUSE	Testin	Testin
12	5	7	7	6	36	18	52	6	FSHR	FSH receptor RTPCR−cloned	FSHrec
8	3	7	7	7	21	13	35	6	AK000209	Homo sapiens cDNA FLJ20202	22−A04
6	5	10	9	9	28	19	29	8	Mm_TBI10966.1	Mm_TBI10966.1	9−J15
9	11	13	13	19	58	28	51	8		RSA−8−C01	8−C01
10	15	17	19	29	64	26	17	5	SPC1_HUMAN	Homo sapiens MICROSOMAL SIGNAL PEPTIDASE 25	14−C01
7	11	11	12	15	31	22	12	3	Mm_TBI29791.1	Mm_TBI29791.1	19−K03
6	9	8	8	11	24	16	12	2	RTR	Orphan receptor RTR	12−O17
4	7	5	4	5	16	10	7	1	IMD2_MOUSE	INOSINE−5'−monophosphate dehydrogenase 2 (IMD2)	15−F16
5	9	7	8	8	16	12	29	8	NIPIL	NIPI−like protein (NIPIL)	18−O23
7	11	11	10	12	21	11	13	5	KCRB_HUMAN	Homo sapiens creatine kinase B (KCRB)	23−K22
7	10	17	12	10	23	15	13	5	Mm_TBI24381.1	Mm_TBI24381.1	8−F19
10	27	16	15	11	35	21	29	12		RSA−8−A20_#0	8−A20
21	50	44	34	23	49	42	80	26	KPR1_HUMAN	Human phosphoribosyl pyrophosphate synthetasesubunit I	20−B03
16	27	30	30	21	32	24	22	19	ATF1_MOUSE	ATF1	ATF1
38	50	48	41	40	49	33	42	58		HS985343 Human STS WI−15071.	5−B5RZ
33	54	42	40	34	38	46	40	15	Mm_TBI8803.1	Mm_TBI8803.1	13−M17
42	62	49	26	30	38	40	68	43		EST clone 4−D1RZ	4−D1RZ
31	41	32	21	22	21	28	56	21	FSC1_MOUSE	Major fibrous sheath protein (FSC1)	8−P20
48	45	35	27	26	16	27	39	18	RSP4_MOUSE	Laminin receptor RTPCR	
55	44	49	28	16	15	42	54	55	MmTBI8810.1	Mm_TBI8810.2	
45	28	21	24	15	19	23	30	14		RSA−21−C16	
48	28	19	16	10	17	17	35	15	HMG1_MOUSE	High mobility group 1 protein (HMG−1)	
33	20	12	11	7	10	7	12	12	CLUS_MOUSE	SULFATED GLYCOPROTEIN 2 (Clusterin)	
14	18	19	6	6	7	4	18	2	UBQLN3	Homo sapiens ubiquilin 3 (UBQLN3)	
7	17	13	11	11	10	6	6	2	CREM_MOUSE	CREM cAMP responsive element modulator	
43	66	90	57	54	51	16	18	8	KRAF_MSV36	RAF serine/threonine−protein kinase transforming protein	
18	42	41	50	44	45	30	14	8	ATPA_MOUSE	ATP synthase alpha subunit (ATPA)	
10	36	27	31	36	24	22	24	14	p18ink4	p18ink4 RTPCR−cloned	
28	59	65	48	74	66	47	32	2	SYA_MOUSE	Alanyl−tRNA synthetase	
21	41	52	58	80	89	46	30	11	BiP	immunoglobulin heavy chain binding protein (BIP)	11−I03

Appendix C

9	17	19	21	23	25	27	wt	mut	Gene	Title	Clone
18	33	45	51	47	73	49	24	21	PP1G_MOUSE	Protein phosphatase type 1 (PP1G)	24−I17
22	64	49	52	36	86	58	84	37		Est clone	3−B3RZ
13	23	22	31	30	44	35	38	21	Mm_TBI58367.1	Mm_TBI58367.1	4−C23
17	38	35	50	74	67	80	38	52	MMPTPMEG	MMPTPMEG PTP MEG, protein tyrosine phosphatase	5−B3RZ
34	40	35	55	107	92	127	111	65	ACET_MOUSE	ANGIOTENSIN−CONVERTING ENZYME PRECURSOR	+2−A6RZ
17	16	21	32	32	36	50	60	28	Mm_TBI11045.1	Mm_TBI11045.1	8−A03
16	10	13	17	15	20	29	43	18	Hs104930.1	Hs104930.1	7−P06
17	12	17	17	15	26	44	50	17	HIP2	UBIQUITIN−CONJUGATING ENZYME E2 (HIP2)	23−E16
57	58	55	47	34	86	129	120	78	Mm_TBI68705.1	Mm_TBI68705.1	1−N03
17	35	31	25	31	36	74	129	52	Mm_TBI3547.1	Mm_TBI3547.1	7−G24
14	43	46	43	31	72	92	31	11	ACTG2	Smooth muscle gamma actin (ACTG2)	18−F11
4	12	11	13	8	18	26	35	3	STAG3	nuclear protein stromal antigen 3 (SA3)	7−K02
4	32	32	24	17	40	54	14	3	Mm_TBI11079.1	Mm_TBI11079.1	18−F19
4	17	19	13	13	25	20	10	1	RNTMDCIV	Rattus norvegicus tMDC IV protein	19−K20
4	33	22	15	12	27	26	61	31	G3PT_MOUSE	Testis Glyceraldehyde 3−phosphatedehydrogenase (Gapd−S)	9−G06
2	22	18	12	11	16	19	19	13	Mm_TBI16028.1	Mm_TBI16028.1	9−L06
2	10	8	10	11	16	17	79	6	Mm_TBI16824.1	Mm_TBI16824.1	10−L20
2	8	5	10	4	11	17	13	5	Mm_TBI22431.1	Mm_TBI22431.1	11−P17
3	12	7	15	12	30	22	45	7	AV258083	testis cDNA clone:4922501M16	7−B09
3	14	10	13	15	26	25	25	3	KIAA0370	Homo sapiens KIAA0370 gene	3−N10
3	18	11	17	15	19	21	21	4	IMD2_MOUSE	inosine−5'−monophosphate dehydrogenase 2 (IMD2)	8−L05
5	37	33	36	43	43	55	51	32	Mm_TBI95030.1	Mm_TBI95030.1	18−P23
4	36	31	37	31	39	43	39	19	Mm_TBI16028.1	Mm_TBI16028.1	6−A15
4	60	83	80	66	94	62	39	46	Mm_TBI914.2	Mm_TBI914.2	6−K04
4	40	38	60	71	78	62	45	32	Mm_TBI10971.1	Mm_TBI10971.1	29−M17
4	18	17	40	45	41	28	23	21	AI326289	Stratagene mouse testis cDNA clone IMAGE:514598	23−N01
5	20	20	24	43	47	44	21	14	Mm_TBI20069.1	Mm_TBI20069.1	19−L04
3	15	15	13	17	24	16	68	24	ODFP_MOUSE	Outer dense fiber protein of sperm tails (Odf1)	5−O18
3	13	10	8	18	28	16	52	12		RSA−7−E22	7−E22
4	11	8	7	21	37	20	17			RSA−13−N04	13−N04
5	16	9	7	15	30	23	74	20	Mm_TBI22271.1	Mm_TBI22271.1	18−K19
4	10	14	9	16	27	21	20	1	ZN76_HUMAN	Human zinc−finger protein 76	10−B02
5	13	14	13	15	21	22	22	6	POMT1	Homo sapiens protein O−mannosyl−transferase 1 (POMT1)	
7	19	14	17	15	22	24	27	6	Mm_TBI8372.1	Mm_TBI8372.1	
5	16	10	11	9	16	12	27	11	TGLC_CHICK	Chicken transglutaminase	
6	12	12	11	10	10	12	27	12	ppx	Protein phosphatase X homolog (PPX)	
7	17	12	9	12	9	14	20	13	CBPA	Mouse mast cell carboxypeptidase A mRNA	
14	17	11	11	19	13	19	28	18	Mm_TBI10966.1	Mm_TBI10966.1	
15	11	11	10	12	13	14	19	12	MmTBI16947.2	Mm_TBI16947.2	
18	9	10	9	13	11	19	11	2	EF1G_HUMAN	Homo sapiens elongation factor 1−gamma homolog	
9	7	5	7	10	10	11	24	8	Mm_TBI11759.1	Mm_TBI11759.1	
9	8	4	6	13	10	16	23	10	KIAA1140	Homo sapiens KIAA1140 protein	
8	9	4	5	12	14	14	39	13	Mm_TBI84992.1	Mm_TBI84992.1	
11	9	5	7	22	25	12	11	2	Mm_TBI7735.1	Mm_TBI7735.1	
8	8	7	7	18	27	15	34	2		RSA−17−I21	
5	7	8	8	22	27	18	8		Tankyrase	H. TRF1−interacting ankyrin−related ADP−ribose polymerase (Tankyrase)	
5	7	6	8	19	41	21	15	3	Mm_TBI24381.1	Mm_TBI24381.1	
4	5	4	6	15	26	21	42	3	Mm_TBI8139.1	Mm_TBI8139.1	
6	9	5	10	20	46	37	32	4	Mm_TBI23178.1	Mm_TBI23178.1	
4	7	4	7	21	22	15	17	6		RSA−29−J11	
3	5	4	7	13	30	9	11	3		RSA−29−O22	
3	8	5	8	14	29	13	9	3	RS24_HUMAN	Ribosomal protein S24	
3	8	5	11	19	24	17	31	3	Mm_TBI49847.1	Mm_TBI49847.1	
3	7	4	12	9	25	12	15	8	H2A1_MOUSE	Histone H2A.1	
6	10	7	18	15	38	29	23	8	MmTBI16933.1	MmTBI16933.1	
7	8	5	14	13	18	23	72	34	BB012305	Testis cDNA clone:4930448B06	

9	17	19	21	23	25	27	wt	mut	Gene	Title	Clone
11	10	6	13	12	23	40	41	5	PTPLB	Protein tyrosine phosphatase−like protein PTPLB (Ptplb)	
10	13	9	11	10	35	38	33	5	Hs16492.3	Hs16492.3	
14	16	16	19	12	104	108	99	22	HPI31	Homo sapiens proteasome inhibitor hPI31	
6	3	4	7	5	18	27	23	3		RSA−21−M06	
3	2	3	4	3	12	16	22	2		RSA−14−B03	
3	3	5	3	3	13	14	48	6		RSA−4−J04	
3	4	6	5	4	28	16	5	3	Prkmk7	Mitogen−activated protein kinase kinase 7 (PRKMK7)	
3	3	4	4	4	26	11	11	2	BB359245	male corpus striatum cDNA clone:C030037D09	
4	4	3	6	6	30	17	27	1	G3PT_MOUSE	Testis Glyceraldehyde 3−phosphatedehydrogenase (Gapd−S)	
4	4	3	6	5	13	13	18	4	Mm_TBI12965.1	Mm_TBI12965.1	
5	7	4	8	6	31	23	47	10	KIAA1364	Homo Sapiens KIAA1364 protein	
4	7	4	5	7	28	23	28	4	Mm_TBI22932.1	Mm_TBI22932.1	1−F22
2	6	4	4	4	15	18	26	2	HS1202587	Homo sapiens cDNA clone 726980	18−B10
2	7	3	4	6	23	30	31	7		RSA−7−I23	7−I23
2	4	2	2	6	20	16	50	4	SBBI03	Homo sapiens hypothetical SBBI03 protein	11−P11
2	5	3	4	14	51	43	41	2	CALI_HUMAN	Homo sapiens Calicin (CALI)	3−D20
2	5	2	2	14	24	12	10	2		RSA−24−D09	24−D09
2	7	2	2	12	37	34	53	3		RSA−3−E13	3−E13
2	5	3	1	6	8	10	31	10	HO2_MOUSE	Heme oxygenase 2a (HO−2a)	5−I17
2	5	6	2	10	15	12	10	3	Mlark	RNA BINDING MOTIF PROTEIN 4 (Mlark)	15−F10
1	2	4	2	7	24	26	58	1	Mm_TBI20699.1	Mm_TBI20699.1	13−K01
1	3	4	2	8	20	17	9	1		RSA−11−P22	11−P22
1	3	6	1	3	5	15	18	5	Mm_TBI51297.1	Mm_TBI51297.1	8−M19
1	3	2	1	3	12	14	20	1	Mm_TBI84692.1	Mm_TBI84692.1	12−K09
1	2	2	1	10	9	17	12	1	Mm_TBI91543.1	Mm_TBI91543.1	
1	2	2	1	7	31	41	18	1	KIAA1053	Homo sapiens KIAA1053 protein	
1	1	4	1	5	37	25	51	1	Mm_TBI22634.1	Mm_TBI22634.1	
1	1	4	1	5	13	10	23	1	HSP2_MOUSE	Protamine 2 (mP2)	
1	1	2	1	3	9	11	57	2	Mm_TBI24369.1	Mm_TBI24369.1	
1	1	1	1	7	13	14	30	3	CHIO_HUMAN	Homo sapiens β2−chimaerin	
1	1	1	1	4	9	9	31	1	ODFP_MOUSE	Outer dense fiber protein of sperm tails (Odf1)	
1	1	1	1	4	22	15	16	1	Mm_TBI13385.1	Mm_TBI13385.1	
1	1	1	1	3	12	13	20	1	Mm_TBI85000.1	Mm_TBI85000.1	
1	1	1	1	3	11	8	44	1	LCFB_MOUSE	Long chain fatty acyl CoA synthetase (LCFB).	
1	1	1	1	8	39	21	9	1		RSA−4−A22	
1	1	1	1	5	15	7	7	1		RSA−7−H14	
1	1	1	1	4	14	5	16	1		RSA−3−N13	
1	1	1	1	11	21	8	25	1	AF1Q_MOUSE	AF1Q gene	
1	1	1	1	10	20	10	8	1	Sperizin	Spermatid−specific ring zinc finger (Sperizin)	
1	1	1	1	5	10	5	16	1		RSA−14−P10_#0	
1	2	1	1	4	11	10	17	1	KC12_HUMAN	Homo sapiens casein kinase I γ 2	
3	5	1	2	6	58	49	178	38	AI575885	Rattus norvegicus cDNA clone UI−R−G0−ut−d−01−0−UI	
3	4	1	1	5	16	11	25	1		RSA−18−P12	
2	3	1	1	4	4	4	15	1	HSJ3_MOUSE	Testis specific DNAj−homolog 3 (HSJ3)	
5	3	1	1	2	5	11	18	7	D−AKAP1	Dual specificity A−kinase anchoring protein 1 (D−AKAP1)	
1	2	1	1	1	4	5	30	5	DBI5_MOUSE	DIAZEPAM BINDING INHIBITOR−LIKE 5	
1	2	1	1	1	4	5	25	1	DBI5_MOUSE	DIAZEPAM BINDING INHIBITOR−LIKE 5	
1	5	1	1	1	1	2	22	10		RSA−3−O17_0	
2	5	2	2	1	11	19	18	1		RSA−15−E05	
2	2	1	1	1	2	2	42	1	POR1_MOUSE	Voltage dependent anion channel 1 (POR1)	
3	1	1	4	1	12	12	20	1		RSA−15−J21	
5	1	1	2	1	8	6	16	2	ATP1B3	Sodium/Potassium−Transporting ATPase beta 3 subtunit (ATP1B3)	
2	1	1	1	1	4	5	17	1	AKAP110	Protein kinase A binding protein AKAP110	7−J03
2	1	1	2	2	20	14	12	1	FSC1_MOUSE	Major fibrous sheath protein (FSC1)	17−K23
2	1	1	1	5	12	11	28	2	Mm_TBI13385.1	Mm_TBI13385.1	6−G12
2	1	1	2	6	15	10	19	1	BB020196	testis cDNA clone:4930597N20	3−D16

Appendix C

9	17	19	21	23	25	27	wt	mut	Gene	Title	Clone
2	1	1	4	9	14	11	20	12		RSA−6−I14	6−I14
3	1	1	5	5	7	5	18	5	Hs10043.3	Hs10043.3	4−L03
4	1	1	12	10	79	33	62	1	FSC1_MOUSE	Major fibrous sheath protein (FSC1)	3−E08
3	1	1	5	7	29	12	16	2	KAP2_MOUSE	cAMP−dependent protein kinase type II (KAP2)	19−I12
3	1	1	10	25	117	143	89	3	ODFP_MOUSE	Outer dense fiber protein of sperm tails (Odf1)	2−P22
3	1	1	3	7	20	9	13	1		RSA−7−E22	7−E22
2	2	1	4	10	45	37	73	1		RSA−7−A01	7−A01
1	1	1	2	2	3	6	15	3	SRPK2	serine/arginine−rich protein specific kinase 2 (SRPK2)	4−I03
1	2	1	4	12	44	43	51	2	BB015390	testis cDNA clone 4930553L18	6−M22
1	2	1	4	13	29	22	12	2	tACTIN2	Testis specific actin (tActin2)	2−G04
1	2	1	2	9	11	11	32	2	UBQLN3	Homo sapiens ubiquilin 3 (UBQLN3)	2−A15
3	1	1	1	6	22	18	43	3		RSA−6−C12	6−C12
4	1	1	1	6	18	22	32	6	Mm_TBI23832.1	Mm_TBI23832.1	17−G06
2	1	1	1	3	8	8	15	1	Mm_TBI20710.1	Mm_TBI20710.1	21−A02
5	1	1	1	6	23	16	20	1	Mm_TBI85000.1	Mm_TBI85000.1	3−B03
3	1	1	1	4	4	5	22	1	GLNA_MOUSE	Glutamate−ammonia ligase (GLNA)	31−B19
2	1	1	1	4	7	2	19	1	BIG2	Homo sapiens brefeldin A−inhibited guanine nucleotide−exchange protein 2	13−A22
1	1	1	1	3	4	1	22	1	Hs189105.1	Hs189105.1	2−P11
1	1	3	70	11	2	1	2	2	CCAG_RAT	Rattus norvegicus low voltage T−type calcium channel α−1γ subunit (CCAG)	24−H21

0.00	0.05	0.10	0.15	0.20	0.25	0.30	0.50	1.00	Colorbar: Normalized Expression Levels on 0−1 scale are color coded