From
The Programme for Genomics and Bioinformatics
Department of Cell and Molecular Biology
Karolinska Institutet, Stockholm, Sweden

GENE COMPLEXES AND REGULATORY DOMAINS IN METAZOAN GENOMES

Pär Engström

Karolinska Institutet

Stockholm 2007

Published By Karolinska Institutet. Printed by Larserics Digital Print.

ISBN 978-91-7357-361-0

Watching a coast as it slips by the ship is like thinking about an enigma. There it is before you - smiling, frowning, inviting, grand, mean, insipid, or savage, and always mute with an air of whispering, Come and find out.

- Joseph Conrad, Heart of Darkness

ABSTRACT

Despite the recent massive increases in genome and transcript sequence data, including whole-genome sequences for humans and many other metazoans, our understanding of the content of these sequences is far from complete. This thesis is about making use of metazoan sequence data to detect functional genetic elements on a genome-wide scale and examine the distribution of those elements on chromosomes. Specifically, the thesis focuses on the occurrence of gene complexes, such as pairs of overlapping genes, and on chromosomal regulatory domains of importance in development and disease.

Mammalian genomes contain a larger than expected number of complex loci, in which genes on opposite strands share transcribed regions, exons and/or core promoters. We find that, in both human and mouse genomes, 25% of transcriptional units (TUs) share exon sequence with a TU on the opposite strand. The true proportion is likely to be significantly higher because transcriptomes are not fully sequenced. Intriguingly, most pairs of overlapping TUs consist of one coding and one noncoding TU. We have included a large dataset of transcript sequences from such noncoding TUs in a database of noncoding RNA (http://research.imb.uq.edu.au/RNAdb). While nearly a thousand cases of overlapping TU arrangements are conserved between human and mouse, these constitute only 17% of all detected TU overlaps, suggesting that many species-specific arrangements exist. Taking advantage of newly available CAGE tag data on transcription start site locations, we analyze bidirectional promoters and show that their divergent transcription initiation regions are broad and often separated only by a small region (<60 bp) at which overall sequence composition changes strand.

Vertebrate, insect and nematode genomes contain an abundance of highly conserved noncoding elements (HCNEs) that appear to function as enhancers for developmental regulatory genes around which they cluster. We show evidence that large blocks of conserved synteny (genomic regulatory blocks, GRBs) have been maintained, across vertebrates and across insects, to keep arrays of HCNEs intact. GRBs often contain “bystander” genes whose functions and expression patterns are unrelated to those of the presumptive target genes of HCNE enhancer activity. By analyzing the fate of duplicated genes and HCNEs after whole-genome duplication in teleosts, we show that bystander genes are indeed independent of the regulatory input of HCNE arrays. In addition, we describe differences in core promoters between target genes and bystander genes that might explain the differences in their responsiveness to long-range enhancers. We present a web resource (http://ancora.genereg.net) for exploring the distribution of HCNEs on metazoan chromosomes.

Together with other recent studies, this work challenges the canonical “colinear” model of how genes and their regulatory elements are arranged in metazoan genomes. Vertebrate and insect genomes appear to contain an abundance of nested and overlapping gene structures, giving rise to both coding and noncoding transcripts. In addition, regulatory elements controlling the expression of a gene are frequently distributed within or beyond other genes. These findings should be taken into account in future studies of regulation of gene expression and effects of genetic variation by considering the genomic neighborhood of genes and polymorphisms of interest, up to distances on the order of a million base pairs in the human genome.

PUBLICATIONS INCLUDED IN THIS THESIS

This thesis is based on the following articles, which will be referred to by their roman numerals in the text.

I. KC Pang, S Stephen, **PG Engström**, K Tajul-Arifin, W Chen, C Wahlestedt, B Lenhard, Y Hayashizaki, and JS Mattick
RNAdb – a comprehensive mammalian noncoding RNA database.
Nucleic Acids Res., 2005, **33**:D125.

II. S Katayama*, Y Tomaru*, T Kasukawa, K Waki, M Nakanishi, M Nakamura, H Nishida, CC Yap, M Suzuki, J Kawai, H Suzuki, P Carninci, Y Hayashizaki, C Wells, M Frith, T Ravasi, KC Pang, J Hallinan, J Mattick, DA Hume, L Lipovich, S Batalov, **PG Engström***, Y Mizuno*, MA Faghihi, A Sandelin, AM Chalk, S Mottagui-Tabar, Z Liang, B Lenhard, and C Wahlestedt
Antisense transcription in the mammalian transcriptome.
Science, 2005, **309**:1564.
*SK, YT, **PGE** and YM contributed equally to this work (authors are ordered by affiliation)

III. **PG Engström**, H Suzuki, N Ninomiya, A Akalin, L Sessa, G Lavorgna, A Brozzi, L Luzi, SL Tan, L Yang, G Kunarso, E Lian-Chong Ng, S Batalov, C Wahlestedt, C Kai, J Kawai, P Carninci, Y Hayashizaki, C Wells, VB Bajic, V Orlando, JF Reid, B Lenhard, and L Lipovich
Complex loci in human and mouse genomes.
PLoS Genetics, 2006, **2**:e47.

IV. H Kikuta, M Laplante, P Navratilova, AZ Komisarczuk, **PG Engström**, D Fredman, A Akalin, M Caccamo, I Sealy, K Howe, J Ghislain, G Pezeron, P Mourrain, S Ellingsen, AC Oates, C Thisse, B Thisse, I Foucher, B Adolf, A Geling, B Lenhard, and TS Becker
Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates.
Genome Res., 2007, **17**:545.

V. **PG Engström**, SJ Ho Sui, Ø Drivenes, TS Becker, and B Lenhard
Genomic regulatory blocks underlie extensive microsynteny conservation in insects.
Genome Res, in press.

VI. **PG Engström**, D Fredman, and B Lenhard
Ancora: a web resource for exploring highly conserved noncoding elements and their association with developmental regulatory genes.
Manuscript (submitted).

OTHER PUBLICATIONS

During the course of my doctoral studies, I have also contributed to the following publications.

- B Lenhard, A Sandelin, L Mendoza, **P Engström**, N Jareborg, and WW Wasserman (2003) Identification of conserved regulatory elements by comparative genome analysis. *J. Biol*. **2**:13.
- A Sandelin, W Alkema, **P Engström**, WW Wasserman, and B Lenhard (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.* **32**:D91.
- A Sandelin, P Bailey, S Bruce, **PG Engström**, JM Klos, WW Wasserman, J Ericson, and B Lenhard (2004) Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes. *BMC Genomics* **5**:99.
- S Mottagui-Tabar, MA Faghihi, Y Mizuno, **PG Engström**, B Lenhard, WW Wasserman, and C Wahlestedt (2005) Identification of functional SNPs in the 5-prime flanking sequences of human genes. *BMC Genomics* **6**:18.
- N Ståhlberg, R Merino, LH Hernandez, L Fernandez-Perez, A Sandelin, **P Engström**, P Tollet-Egnell, B Lenhard, and A Flores-Morales (2005) Exploring hepatic hormone actions using a compilation of gene expression profiles. *BMC Physiology* **5**:8.
- RIKEN Genome Exploration Research Group and Genome Science Group and the FANTOM Consortium (2005) The transcriptional landscape of the mammalian genome. *Science* **309**:1559.
- N Maeda, T Kasukawa, R Oyama, J Gough, M Frith, **PG Engström**, B Lenhard, RN Aturaliya, S Batalov, KW Beisel, CJ Bult, CF Fletcher, AR Forrest, M Furuno, D Hill, M Itoh, M Kanamori-Katayama, S Katayama, M Katoh, T Kawashima, J Quackenbush, T Ravasi, BZ Ring, K Shibata, K Sugiura, Y Takenaka, RD Teasdale, C A Wells, Y Zhu, C Kai, J Kawai, DA Hume, P Carninci, and Y Hayashizaki (2006) Transcript annotation in FANTOM3: mouse gene catalog based on physical cDNAs. *PLoS Genetics* **2**:e62.
- P Carninci, A Sandelin, B Lenhard, S Katayama, K Shimokawa, J Ponjavic, CA Semple, MS Taylor, **PG Engström**, MC Frith, AR Forrest, WB Alkema, SL Tan, C Plessy, R Kodzius, T Ravasi, T Kasukawa, S Fukuda, M Kanamori-Katayama, Y Kitazume, H Kawaji, C Kai, M Nakamura, H Konno, K Nakano, S Mottagui-Tabar, P Arner, A Chesi, S Gustincich, F Persichetti, H Suzuki, SM Grimmond, CA Wells, V Orlando, C Wahlestedt, ET Liu, M Harbers, J Kawai, VB Bajic, DA Hume, and Y Hayashizaki (2006) Genome-wide analysis of mammalian promoter architecture and evolution. *Nature Genetics* **38**:626.
- KC Pang, S Stephen, ME Dinger, **PG Engström**, B Lenhard, and JS Mattick (2007) RNAdb 2.0 – an expanded database of mammalian noncoding RNAs. *Nucleic Acids Res.* **35**:D178.
- Y Sheng, **PG Engström**, and B Lenhard (2007) Mammalian microRNA prediction through a support vector machine model of sequence and structure. *PloS ONE* **2**:e946.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

bp	Base pair
CAGE	Cap analysis of gene expression
cDNA	Complementary DNA
ChIP	Chromatin immunoprecipitation
DNA	Deoxyribonucleic acid
DPE	Downstream promoter element
DRE	DNA replication element
ENCODE	Encyclopedia of DNA elements
EST	Expressed sequence tag
FANTOM	Functional annotation of the mouse
GRB	Genomic regulatory block
HCNE	Highly conserved noncoding element
Inr	Initiator element
kb	Kilo bases or kilo base pairs
Mb	Mega bases or mega base pairs
MPSS	Massive parallel signature sequencing
mRNA	Messenger RNA
NCBI	National Center for Biotechnology Information
ncRNA	Non-protein coding RNA
nt	Nucleotide
ORESTES	Open reading frame expressed sequence tags
PET	Paired end ditag
RNA	Ribonucleic acid
SAGE	Serial analysis of gene expression
TSS	Transcription start site
TU	Transcriptional unit
UCSC	University of California Santa Cruz
UTR	Untranslated region

1 INTRODUCTION

I began my doctoral studies in 2002, one year after a draft sequence of the human genome was published [1, 2]. This draft sequence was a crucial milestone in the human genome project, which was launched in 1990 with the aim to produce a high-quality reference sequence for the human genome. Now, five years later, the human genome project has been declared finished [3], and reference genome sequences for many other animals have been published (for examples, see [4-12]). At the UCSC Genome Browser web site [13], researchers can view alignments between genome sequences of 28 vertebrate species and alignments between genome sequences of 14 insect species. Advances in sequencing technology will likely make it affordable to determine genome sequences of individual humans on a large scale in the near future [14]. Pioneering efforts in this area have recently provided draft genome sequences of two scientists who have made fundamental contributions to the understanding of genomes: James Watson, co-discoverer of the structure of DNA [15], and Craig Venter, advocate of high-throughput methods for transcriptome and genome sequencing [2, 16, 17].

Despite the recent massive increases in genome sequence data, our understanding of the content of these sequences is far from complete [18]. A major goal of *genomics* - the study of genomes - is to identify functional elements in genome sequences. If we could produce a complete parts list for a genome, we would have a better foundation for understanding how these parts function together in living organisms, and how their disruption, for example by genetic mutations, leads to disease. This thesis is about making use of sequence data for metazoans (humans and other animals) to detect functional genetic elements and examine the distribution of those elements on chromosomes. Specifically, the thesis focuses on the occurrence of gene complexes, such as pairs of overlapping genes, and on chromosomal regulatory domains of importance in development and disease.

The first two sections of this introduction describe genome-wide discovery and analysis of two types of functional elements that are of primary interest: genes and their regulatory elements. This is followed by a section on how whole-genome sequences from different species can be compared to reveal how genomes have been rearranged in evolution and how functional elements have been shuffled. The final section discusses the arrangement of genes and regulatory elements into higher-order structures on chromosomes.

1.1 GENES

Although most contemporary biologists have an intuitive understanding of what is meant by a gene, there is no precise consensus definition of this term, which has taken a variety of meanings since its conception by Wilhelm Johannsen in 1909. For a comprehensive historical review about the definition a gene, see [19]. In this thesis, the word gene is used in a broad sense: any part of a genome that is transcribed is considered to belong to a gene.

Genes can be further subdivided into those that encode proteins and those that only give rise to RNA transcripts (RNA genes). While protein-coding genes historically have been given more attention than RNA genes, recent studies have shown that non-protein coding RNAs (ncRNAs) are more abundant and diverse than previously thought and thus caused the focus to shift somewhat [20]. In particular, the discovery of microRNAs has fueled research in the field of RNA genes [21]. While microRNAs are increasingly well understood, the function of many other ncRNAs is unclear [22, 23].

1.1.1 Gene discovery

Most known transcripts have been discovered by sequencing from complementary DNA (cDNA) libraries [24-30]. Briefly, a cDNA library is constructed by purifying RNA from cells or tissues, reverse-transcribing RNA into double-stranded cDNA, and inserting cDNA into vectors that can be maintained in bacteria [31]. Selected cDNA inserts are then sequenced and, for species with assembled genomes, resulting transcript sequences are aligned to the genome sequence to reveal the locations and exon-intron structures of genes [32]. In large-scale projects, cDNAs to sequence are often selected at random from a library. As a result, more sequences are obtained for transcripts that are more abundant in the original RNA samples. To increase the rate of discovery of new genes, normalization and subtraction techniques can be applied in the library construction to reduce the difference in abundance between different transcript isoforms and remove sequences that are already known [33]. In addition, a number of techniques have been developed to control the extent to which obtained cDNAs represent full RNA transcripts, as opposed to partial transcripts. In the now common case when full mRNA sequences are desired, a primer that recognizes the poly-A tail of mRNAs is typically used in the reverse transcription reaction, and modern approaches also include a step where the cap structure at the other (5') end of mRNAs is recognized [34, 35].

Most mammalian mRNAs are several kb in size [36]. Because a sequencing run only can cover a few hundred bp, several rounds of sequencing are usually required to obtain full-length transcript sequences, making the process laborious. It is therefore common to perform only one round of sequencing at either or both ends of inserts. The products of such one-pass sequencing are called expressed sequence tags (ESTs) [16]. While ESTs rarely represent full-length transcripts (hence the name tags) and often contain low-quality sequence, they can be produced in a very high-throughput fashion and have proven highly useful for gene discovery [24-30, 37]. The first publication on the dbEST database, which was created in 1992 to store all publicly available EST sequences, reported a content of 14,556 sequences from human [38]; today the database contains more than eight million human ESTs, demonstrating the popularity of this approach. For comparison, there are only 245,820 sequences supposed to represent full-length human transcripts (cDNA sequences) in the public domain (the sequence count was obtained from the UCSC Genome Browser database [13]), and many of them are not truly full-length [39]. A number of other methods for obtaining sequence tags in even more high-throughput have been developed. These methods, which produce shorter tags (9-21 bp), include serial analysis of gene expression (SAGE) [40-42], massive parallel signature sequencing (MPSS) [43], cap-analysis of gene expression (CAGE) [44] and paired-end ditag (PET) sequencing [45]. In MPSS, the higher throughput is

achieved by cloning on microbeads instead of in bacteria, while SAGE, CAGE and PET gain efficiency by concatenating tags so that multiple tags can be sequenced in one run. The CAGE and PET methods have the advantage of capturing tags from 5'-ends of transcripts, thereby allowing accurate mapping of transcription start sites and corresponding promoters on the genome [46]. If large numbers of tags are sequenced from non-normalized libraries, they can also be used as a measure of transcript abundance: a higher proportion of tags with a particular sequence corresponds to a higher abundance of the corresponding RNA species.

While the many publicly available transcript sequences provide a rich source of information, sequences that have been produced in high-throughput must be interpreted with caution. Transcript sequence databases contain many artifacts, including truncated, chimeric and incompletely spliced sequences [47-50]. One particular problem is that many libraries were not constructed by a directional cloning procedure [31]. Sequences from such libraries therefore have no information about orientation of transcription. Examples of such sequences include 700,000 ESTs from the ORESTES project, in which cDNAs were produced using random primers to increase the representation of coding sequence in ESTs [26]. For ESTs that come from directionally cloned libraries, annotated read direction (5' or 3') usually reveals orientation with respect to the original transcript. However, this annotation can be inconsistent between libraries and sometimes incorrect, due to cloning of cDNAs in the wrong orientation or so-called lane-tracking errors [48, 49]. Reference sequence collections where many of the above-mentioned errors have been screened out are available [36, 51, 52], but they fail to capture much of the complexity of animal transcriptomes [18, 53-55].

Recently, several groups have used high-density oligonucleotide arrays to reveal transcribed regions [18, 23, 53-56]. In this approach, up to several hundred million unique oligonucleotide probes are arrayed on chips. Each probe is about 30 nt in size and designed to uniquely identify a genomic region. The chips are called tiling arrays because the probes are designed to collectively cover as much as possible of a genome, or of selected chromosomal segments if the genome is too large. Labeled nucleic acids obtained from RNA samples are hybridized to the arrays to determine regions corresponding to stable transcripts and their expression levels. While conventional expression arrays are designed to detect known transcripts, the use of tiling arrays has revealed many previously unknown exons (see below).

It is also possible to predict genes based on genome sequence alone. A number of such prediction methods have been developed for metazoan genomes [57-59]. These methods model properties of known genes, such as exon and intron sizes, codon usage, sequence composition, translational and splicing signals, repeat content and cross-species conservation, and, for comparison, properties of intergenic regions. Although many predicted exons have been validated as expressed [53, 56, 60], it is thought that the prediction methods have a considerable false-positive rate [61], and predictions are usually not trusted unless they show similarity to known genes in other species or have been confirmed by transcript sequencing or hybridization.

1.1.2 Pervasive transcription

Around the time of the publication of the draft human genome sequence, Wong et al. argued that most of the human genome is transcribed [62, 63]. It is now clear that this is this is the case for both human and mouse genomes. By aligning transcript sequences to the mouse genome sequence, the FANTOM consortium found evidence for transcription of 63% of the mouse genome sequence [30]. Recently, the ENCODE consortium used a variety of technologies to investigate transcription in selected cell lines for 44 genomic regions, which together comprise 1% of the human genome, and detected 74% of the bases in these regions as transcribed by at least two different technologies [18]. While it is not straightforward to extrapolate this result to the entire human genome, the number is similar to the earlier estimate for the mouse genome.

Although human and mouse genomes are pervasively transcribed, the fraction of these genomes that corresponds to stable transcripts appears to be much smaller, because many genes contain large introns. Remarkably, only ~2% of the human genome is covered by exons of known protein-coding genes from the commonly used reference collections at NCBI, Ensembl and UCSC, and only half of that sequence is predicted to be protein coding (the other half is predicted as 5'- and 3'-untranslated regions, UTRs) [51]. However, it now clear that these gene collections lack many exons, the majority of which may be noncoding. For example, in a large-scale tiling array experiment interrogating 380 Mb of non-repeat human genome sequence (about 30% of the genome) with polyadenylated cytosolic RNA from eight different cell lines, transcripts were detected for 16.5% of the interrogated sequence, and only a quarter of the detected expressed bases corresponded to known genes [54]. These observations are supported by several other recent studies [18, 23, 30, 53].

Gene annotations are denser in the more compact genome of the fruit fly *Drosophila melanogaster*: annotated exons cover 24% of the euchromatic genome sequence [64] (although this number is from 2002, the estimated percentage has not changed since then). Nevertheless, even this percentage is likely to be an underestimate, because a recent tiling array study reported that 30% of all genomic bp found to correspond to transcripts expressed during embryonic development were not covered by known genes, cDNAs or ESTs [55]. Based on correlation analysis of developmental expression profiles, the authors estimated that that at least 85% of the non-repeat portion of the *D. melanogaster* genome is transcribed.

1.1.3 Transcript diversity

A transcribed region can typically give rise to multiple transcript isoforms due to alternative transcription initiation, alternative splicing, and alternative polyadenylation [65, 66]. It is well established alternative transcript isoforms produced from the same region can have related functions [67]. However, many loci appear to contain a network of exon-intron structures giving rise to transcripts with potentially diverse functions [54, 68]. Protein-coding and noncoding transcripts may be transcribed from the same region [23], and many genomic regions are transcribed on both strands [69-71]. At some loci, a switch from generation of protein-coding to noncoding isoforms may be a way to downregulate protein expression. This hypothesis seems particularly plausible

for noncoding transcripts that are targets of nonsense-mediated decay [72]. However, in most cases the relation between protein-coding and noncoding isoforms emanating from the same locus is not clear [23, 68].

The ability of single loci to give rise to a rich diversity of transcripts conflicts with most biologists' notion of what a gene is, and makes it difficult to devise general rules for grouping transcripts into "genes" in a meaningful way based on sequence information alone [19]. Nevertheless, grouping of transcripts is often done in large-scale analysis to remove redundant information and answer questions about the extent of phenomena such as alternative splicing and alternative promoter usage. The FANTOM consortium defined transcriptional units (TUs) by grouping together all cDNA sequences that mapped to the same genomic strand and shared one or more bases of exon sequence [22]. More recently, the FANTOM consortium introduced the notion of a transcriptional framework, which is a grouping of transcripts that takes into account sites of splicing, polyadenylation and transcription initiation [30]. Based on experiences in the ENCODE pilot project, Gerstein et al. [19] proposed a way of grouping transcripts that emphasizes the final expressed products (protein or ncRNA) and always puts mRNAs and ncRNAs in separate groups.

1.2 GENE REGULATORY ELEMENTS

Protein-coding genes and many noncoding RNAs, including some microRNAs, are transcribed by RNA polymerase II from specific transcription start sites (TSSs) in genomes [73, 74]. TSSs are surrounded by sequence elements involved in the regulation of their transcription. These elements, which are often binding sites for transcription factors, are called *cis*-regulatory elements because they occur on the same molecule (the chromosome) as the regulated gene, as opposed to transcription factors, which are separate molecules and therefore said to act in *trans*. Gene expression is also regulated at the post-transcriptional level through *cis*-regulatory elements in transcripts, such as binding sites for splicing regulatory proteins and target sites for microRNAs [21, 75]. The summary below focuses on *cis*-elements involved in the regulation of transcription, because of their relevance for the thesis work.

1.2.1 Genome-wide discovery of transcription factor binding sites

DNA segments directly or indirectly bound by specific proteins *in vivo* can be purified by a technique called chromatin immunoprecipitation (ChIP) [76]. Because the purification is facilitated by crosslinking chromatin and its bound proteins, the recovered DNA segments represent a snapshot of bound segments at the time the crosslinking is performed. In an alternative technique, DamID, the protein of interest is fused to a DNA methyltransferase, so that DNA methylation levels are increased at sites bound by the protein [77]. Regions with increased methylation compared to a control experiment are then identified. The DamID technique thus provides a cumulative measure of chromatin binding over the time the fusion protein is expressed. Recently, these methods have been used extensively in combination with tiling arrays or large-scale sequencing to determine bound DNA segments in targeted genomic regions or even genome-wide (for examples, see [18, 78-80]).

Transcription factor binding sites can also be predicted in sequences by matching against profiles (motifs) that capture documented binding preferences of transcription factors [81]. One problem with this approach is that we do not know the binding specificity for most transcription factors, but this may soon be solved by progress in high-throughput determination of transcription factor binding specificities [82, 83]. A more severe limitation is that many false positive predictions are generated because of the notoriously low DNA binding specificity of transcription factors. By requiring predicted sites to be conserved in a genome at suitable evolutionary distance, the signal-to-noise ratio can be increased by an order of magnitude [84]. However, due to the vast noncoding content of metazoan genomes, in particular the mammalian ones, additional information about which regions are likely to contain regulatory information is required to reduce the number of false positives to a manageable level. Although individual transcription factor binding sites are hard to identify with high confidence in genome sequences, regulatory modules consisting of multiple sites can be found by searching for regions enriched for site combinations. This approach has been successfully used for genome-wide detection of functional enhancers in *D. melanogaster* by additionally requiring that site combinations be preserved in *D. pseudoobscura* [85].

1.2.2 Core promoters

Prior to initiation of transcription, the required components are assembled at a region around the TSS called the core promoter [86]. Core promoters contain *cis*-regulatory elements that are recognized by components required for transcription. These *cis*-regulatory elements may include a TATA box 28 to 34 bp upstream of the TSS, an initiator element (Inr) at the TSS, a downstream promoter element (DPE) about 30 bp downstream of the TSS, as well as other elements identified by biochemical studies and computational searches for overrepresented motifs in core promoter sequences [46, 86-89]. However, most mammalian core promoters do not contain matches to well-characterized *cis*-regulatory motifs, but instead coincide with a region enriched for CpG dinucleotides (CpG island) [46]. These core promoters typically contain multiple TSSs distributed over a region that can span more than 100 bp. CpG dinucleotides are otherwise underrepresented in the genome because they are substrates for methylation, and methylated cytosine can be converted to thymine [90]. It has been suggested that core promoters used during early embryonic development escape *de novo* methylation [90]. Accordingly, broadly expressed genes and some developmental regulatory genes tend to have core promoters with CpG islands, while tissue-specific genes tend to have core promoters without CpG islands [46, 91].

1.2.3 Proximal and long-range regulation

Cis-regulatory elements outside core promoters mediate additional control over transcription initiation. Such elements include enhancers, silencers, insulators and locus control regions [73]. The region 200 bp upstream of a TSS tends to be well conserved between human and mouse [46], and researchers often explore sequence up to 5 or even 10 kb upstream of a TSS in search of putative transcription factor binding sites. However, *cis*-regulatory elements for a gene can also be found within its boundaries.

For example, the zebrafish s*hh* gene contains enhancers in its first and second intron that appear to be responsible for its embryonic expression pattern [92].

Long-range *cis*-regulatory elements occur at larger distances from their target genes. One of the most spectacular examples to date is an enhancer 780 kb upstream of human *DACH1*, a gene implicated in regulation of several aspects of development [93]. This and several other enhancers around *DACH1* were discovered by searching for sequences conserved between human, mouse, frog, and three fish genomes (zebrafish, fugu and *Tetraodon nigroviridis*). The conserved sequences were found to possess enhancer activity by cloning each element upstream of a reporter gene and measuring reporter expression in transgenic mice, where reporters displayed spatial developmental expression patterns compatible with the endogenous expression pattern of *DACH1*. It should be noted that long-range *cis*-regulatory elements and their target genes may not always be far from each other in the cell, because chromosomes can bend to bring together elements that are distant in the linear genome sequence. In addition, undiscovered alternative TSSs may exist closer to *cis*-regulatory elements located far upstream of currently known TSSs [18, 55].

Genomic regions that contain enhancers can be detected by inserting reporter genes at random genomic locations [94]. The reporter gene should have a promoter that requires an enhancer for full activation, so that an increase in reporter expression can be detected if the insertion occurs in a location where the promoter is activated by a neighboring enhancer. Genomic locations of insertions can be determined by using primers specific to the inserted constructs to clone and sequence flanking genomic regions. This technique has been widely applied for enhancer detection in *D. melanogaster* [94], and more recently also in zebrafish [95]. It is suitable for locating genomic regions of interest based on tempo-spatial expression patterns of reporters and for making hypotheses about expression patterns of genes in the identified regions. For example, a recent insertion screen for developmental enhancers in zebrafish produced 95 transgenic lines of fish with distinct embryonic reporter expression patterns [95].

Since enhancers can be located at great distances from their target genes, and even inside neighboring genes [96], there must exist mechanisms by which enhancers specifically target certain genes in their vicinity. DNA elements called insulators play a role in this by restricting the reach of enhancers to within defined chromosomal domains [97]. It has also been demonstrated that enhancers can selectively target certain promoters [98, 99] and that this selectivity may be facilitated by the occurrence of different core promoter types [100, 101]. For example, Butler and Kadonaga carried out an enhancer detection screen by inserting constructs containing two reporter genes at random locations in the *D. melanogaster* genome [101]. While the two reporter genes were identical, they had different types of core promoters. The core promoters had motif combinations TATA-Inr and Inr-DPE, respectively. The constructs were designed so that, following insertion, either reporter gene could be selectively excised together with its associated core promoter. By analyzing pairs of fly lines with insertions in the same position but different core promoter types, the authors identified three pairs of fly lines with higher expression from the TATA-Inr promoter and one line with higher expression from the Inr-DPE promoter, suggesting that the different

core promoter types responded differently to enhancers located in the vicinity of insertion sites.

1.2.4 Highly conserved noncoding elements

Searches for enhancers by comparison of distantly related vertebrate genomes, as described for the *DACH1* locus above, have now been performed genome-wide and revealed many examples of long-range enhancers [102-106]. There are thousands of elements that are highly conserved between human and fish and do not overlap known exons [102, 107, 108]. These highly conserved noncoding elements (HCNEs) do not tend to be near known TSSs, but form broad clusters around developmental regulatory genes. Hundreds of HCNEs have been characterized as developmental enhancers by reporter gene assays in transgenic mice, frogs or zebrafish, and the list is growing rapidly [103, 109-112]. Success rates in these experiments have been high, considering that only discrete developmental time points have been explored, suggesting that most HCNEs function as enhancers. For example, Woolfe et al. [102] tested 25 human-fugu conserved HCNEs in transgenic zebrafish and detected enhancer activity of 23 on the second day of development. The elements generally showed reproducible tissue-specific activity. In a more recent study, Pennacchio et al. [106] tested 167 HCNEs in transgenic mice and detected reproducible tissue-specific enhancer activity of 75 (45%) at embryonic day 11.5. Figure 1 illustrates one of the loci investigated in this study and demonstrates how positive elements act as enhancers in distinct anatomic regions, suggesting that they function together in a modular fashion to specify the complete expression pattern of the target gene *SALL1*. In agreement with their function as enhancers, HCNEs conserved between mammals and fish tend to be enriched for matches to binding site profiles for homeodomain proteins and some other developmental transcription factors [104, 113].

The numbers and sizes of HCNEs identified in different genome-wide studies vary depending on the method used. For example, Bejerano et al. [107] identified 381 elements perfectly conserved between human, mouse and rat over at least 200 bp. These elements, which the authors termed ultraconserved elements, include 256 that did not show evidence of transcription from any matching cDNA or EST from any species. The three largest ultraconserved elements, which all exceed 700 bp in size, are located in introns of a gene neighboring the homeobox gene *ARX*. Sandelin et al. [108] identified 3583 HCNEs with a median length of 125 bp by searching for non-exonic sequences with >95% sequence identity over at least 50 bp between human and mouse, as well as evidence of conservation in fugu. They inspected HCNE densities along chromosomes and found peaks at many developmental gene loci. The largest HCNE cluster was found around the homeobox gene *MEIS2* (Figure 2). HCNEs are also abundant and associated with developmental genes in insect [114] and nematode genomes [115]. Vavouri et al. [115] found 26 groups of orthologous genes that are spatially associated with HCNEs in human, flies and worms, although the HCNE sequences are not conserved between the three phyla, suggesting that the mechanisms by which these sequences control gene expression and by which their conservation is maintained existed before the divergence of these phyla.

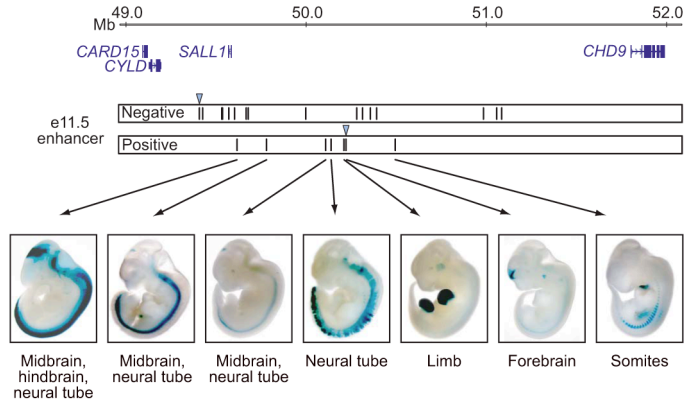

Figure 1. HCNEs with enhancer activity around the human *SALL1* gene. Figure from Pennacchio et al. [106]. The middle tracks depict human fragments that were tested in a transgenic mouse enhancer assay, and their classification as either 'negative' or 'positive' refers to their enhancer activity at embryonic day 11.5. All elements tested were conserved in the fugu genome, and two of these elements were also defined as ultraconserved (denoted by arrowheads). The bottom panel indicates the positive enhancer activities captured in transgenic mice. Reprinted by permission from Macmillan Publishers Ltd: *Nature* **444**:499-502, copyright 2006.

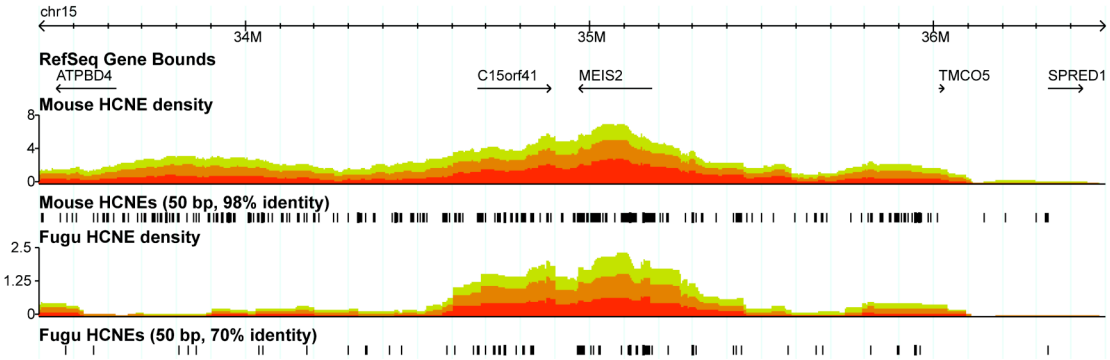

Figure 2. HCNE locations and densities around the human *MEIS2* gene. Screenshot from the Ancora genome browser (Paper VI). *MEIS2* is embedded in an array of HCNEs detected by comparison with mouse and fugu genomes. Overlaid density plots show densities of HCNEs detected at similarity thresholds of 95% (yellow), 98% (orange) and 100% (red) in the mouse comparison and similarity thresholds of 70%, 80% and 90% in the fugu comparison. Density values indicate the percentage of HCNE sequence in a 300 kb window.

Curiously, the conservation level of many HCNEs exceeds what is thought to be required for preservation of transcription factor binding sites, which typically can tolerate mutations in accordance with binding preferences of corresponding transcription factors [81]. It is therefore likely that additional mechanisms, which remain to be understood, act to preserve HCNEs in evolution. To investigate whether ultraconserved elements may be mutational cold spots, Katzman et al. [116] sequenced 315 ultraconserved elements from 72 individuals. They found that polymorphisms do exist in the ultraconserved elements, but that the derived alleles are of low frequency in comparison to derived alleles in protein-coding sequence, suggesting that the ultraconserved elements experience negative selection to a greater degree than protein-coding sequence does.

Comparisons of distant genomes appear to be particularly well suited for revealing *cis*-elements that regulate embryonic development, an intricately regulated process with many similarities even between distantly related species, such as human and fish [117].

However, comparisons of distantly related genomes miss the potentially large number of elements that have appeared or disappeared following the ancient speciation events. Identification of evolutionarily constrained elements between closely related genomes is more challenging because there has been less time for neutrally evolving sequence to diverge, resulting in a lower signal-to-noise ratio. Published approaches to this problem involve comparison of multiple closely related genomes and comparison of likelihood estimates for observed sequence similarity under phylogenetic models of constrained versus neutral sequence evolution [105, 118].

1.3 GENOME EVOLUTION

One incentive for sequencing a large number of metazoan genomes is to further the understanding of our own genome by identifying genomic features that have been conserved in evolution and therefore are likely to be functional. As described in the previous section, identification of genomic regions that are well conserved greatly aids detection of *cis*-regulatory elements. Genomes can also be compared on a larger scale, to determine which genetic elements have been kept in proximity in evolution, and which ones have been separated. Based on such comparisons, hypotheses can be made about selective pressures acting to keep certain elements in proximity and the possible functional relations between those elements.

1.3.1 Genome rearrangements

Chromosomes rearrange through a variety of mechanisms: segments can be deleted, inverted, duplicated, exchanged (translocated) between chromosomes and moved (transposed) to other locations [119]. Signs of ancient whole-genome duplications are also visible in extant genomes. For example, there is evidence that two rounds of whole-genome duplication occurred early in the evolution of vertebrates [120]. Another whole-genome duplication occurred in the lineage leading to teleost fish, which includes the now sequenced fugu [7], *Tetraodon* [9], medaka [12], zebrafish (The Sanger Institute, unpublished) and stickleback (The Broad Institute, unpublished). Whole-genome duplications are typically followed by loss of the majority of gene duplicates, because the additional copies are redundant [121]. Explanations for why some genes survive in duplicate include subfunctionalization, where the copies take on different subsets of the function of the single original gene, and neofunctionalization, where one or both copies evolve new functions [122].

1.3.2 Methods for comparing whole genome sequences

To directly compare genome sequences, it is necessary to align them. The goal of an alignment strategy is typically to match segments that derive from the same ancestral segment. Such segments are called homologs (or homologous segments) and can be further classified as orthologs or paralogs [123]. In evolution, homologous segments can be created by speciation, which results in orthologs. Homologous segments can also be created by duplication in the genome of one species, which results in paralogs. In whole-genome alignments, one often wants to maximize the number of matches between orthologous segments, and minimize the number of matches between paralogous segments, because the goal is to match segments that were the same at the

time of speciation. Alignment of whole genome sequences is challenging because of the large size of the sequences involved (the human genome contains about 3×10^9 bp), the occurrence of lineage-specific insertions, the many duplications that have persisted in evolution, and the generally low similarity between noncoding sequences of distantly related genomes [6, 124].

In the first comparison of the human and mouse draft genome sequences, two whole-genome alignment strategies were used [6]. In one of these approaches, orthologous landmarks corresponding to perfect matches of 40 bp were determined. These are useful for studying large-scale rearrangements, but do not reveal the details of how genomes have diverged. In the other approach, more sensitive local alignments were independently created with the program BLASTZ, designed to align even neutrally evolving human and mouse sequences [125]. One difficulty with interpreting such sensitive local alignment matches is to distinguish orthologous from paralogous matches. In a later study addressing this problem, Kent et al. [124] described two algorithms, implemented in the programs axtChain and chainNet, to post-process BLASTZ alignments. AxtChain uses a new scoring scheme for alignment gaps to combine local matches into larger "chained alignments" that can contain simultaneous gaps in both sequences. The chained alignments can thus span regions of highly diverged sequence or independent insertions in both lineages. To make rearrangements apparent, axtChain only chains together matches that occur in the same order and orientation in the two sequences. The chainNet program further processes the chained alignments by selecting, for each position in each of the genomes, the best match to the other genome. The result is two sets of "nets", one created from the perspective of each genome. In selecting best matches, higher-scoring chained alignments are given priority, on the assumption that long matches of high similarity are more likely to be orthologous matches. The advantage of chaining together matches before creating the nets is that much larger alignments form the basis for deciding which matches are best. The authors of these programs have since applied them to a variety of pairwise genome comparisons and made the results available for visualization and download on the UCSC Genome Browser website [13]. For example, the site currently provides pairwise alignments between the human genome and the genomes of 19 other vertebrates.

An alternative alignment strategy underlies the VISTA portal for comparative genomics [126]. In this approach, the program Shuffle-LAGAN is used to select local matches as a starting point for more sensitive alignments. By modeling rearrangements explicitly in the selection of local matches, Shuffle-LAGAN performs the tasks of both axtChain and chainNet in a unified framework. Alignments produced with this approach are available for browsing and download at the VISTA portal, but the set of alignments provided is currently not as extensive and up-to-date as the chained and net alignments in the UCSC Genome Browser.

1.3.3 Synteny blocks

Two genes or other features that are located on the same chromosome are said be in synteny (literally "same thread") [127]. If orthologs of the genes in a different species also share chromosome, the genes are said to be in conserved synteny between the

species. The terms “conserved linkage” and “conserved microsynteny” can be used to denote the preservation of proximity between genes on the same chromosome in evolution [127, 128]. In this thesis, I refer to regions that contain genes in conserved microsynteny or that otherwise have been largely maintained in evolution as “synteny blocks”. This is in agreement with recent studies [124, 129], although other terms and definitions have been used earlier in the field [6, 130].

Synteny blocks were investigated before whole genome sequences were available. In a study from 1984, Nadeau and Taylor [130] used linkage maps and cytogenetic data to compare the chromosomal locations of homologous human and mouse genes. Based on the limited data available, they detected thirteen synteny blocks and estimated that there are about 180 human-mouse synteny blocks in total. The length distribution of the thirteen synteny blocks is compatible with a model where autosomal rearrangements that have been fixed in evolution are randomly distributed within the genomes, suggesting that there is no reason to assume that long synteny blocks have been protected from rearrangements in evolution. The initial comparison the mouse and human genome sequence detected 342 synteny blocks, with a length distribution compatible with the random breakage model of Nadeau and Taylor [6]. However, this model has recently been challenged by several studies based on more detailed sequence-based maps of synteny blocks [124, 131]. These studies have revealed the existence of many short synteny blocks that were not detected by earlier approaches. The short blocks are clustered in regions between longer blocks, suggesting the existence of fragile regions more prone to breakage and/or that selection has acted to maintain the longer blocks intact. Importantly, Pevzner and Tesler [131] argued that the apparent existence of many short blocks can not be explained by lower genome assembly quality or lower alignability of those regions, because even if only large blocks (> 1 Mb) are considered, any rearrangement scenario transforming the mouse genome into the human genome would require multiple breaks in some of the regions between the large blocks. They invented the term “breakpoint reuse” for referring to the occurrence of multiple breaks in the same region between two large blocks and estimated that at least 190 breakpoint reuses have occurred in the divergence of human and mouse genomes. While this controversial result has been questioned [132-134], deviance from the random breakage model has also been observed in comparisons of multiple mammalian species [135] and multiple insect species [136].

1.4 GENE COMPLEXES AND REGULATORY DOMAINS

Irrespective of whether large synteny blocks have been maintained by natural selection or not, it is clear that genes in metazoan genomes are not randomly distributed [137]. There are many clusters of functionally and structurally related genes that have arisen by tandem duplication, i.e. the duplication of a chromosomal segment into two adjacently situated copies [138-140]. At the well-studied Hox loci, which contain clusters of paralogous genes encoding key regulators of embryonic development, the established gene order appears to be required for proper sequential activation of the genes in the course of development [140, 141]. Another well-studied example are the globin loci, where genes also tend to be arranged in the order of their temporal activation [140]. As described below, animal genomes also contain an abundance of gene complexes of more mysterious nature.

1.4.1 Clusters of coexpressed or functionally related genes

A tendency for genes with similar expression profiles to occur in clusters on chromosomes has been found in the genomes of *Caenorhabditis elegans* [142], *D. melanogaster* [143], mouse and human [144-146]. In nematodes, this phenomenon can be largely explained by the existence of operons, i.e. gene clusters that are transcribed as multi-gene transcripts [142]. This explanation may not apply to the other genomes, in which operons are considered to be rare, although a recent study suggests that numerous operons exist in *D. melanogaster* [147].

At least two studies have further classified coexpression clusters in the human genome. Based on SAGE expression data for different tissues, Lercher et al. [144] found statistically significant spatial clusters of coexpressed human genes to span up to 350 kb in size. They argued that these broad clusters of coexpressed genes are best explained by a tendency for genes expressed in most tissues (housekeeping genes) to cluster rather than a tendency for tissue-specific genes to cluster. The authors did, however, see a significant trend for tight (<100 kb) clusters of coexpressed genes even after correcting for clustering of housekeeping genes. This result - that housekeeping genes and otherwise coexpressed genes might be independently clustered in the human genome - was corroborated by Singer et al. [146], who used expression data from microarrays instead of SAGE. These authors further investigated the conservation of observed clusters in the mouse genome, and found that both housekeeping gene clusters and other clusters of coexpressed genes tend be in conserved synteny in mouse compared to a null model where chromosomal breakpoints are randomly distributed between genes, suggesting that the clusters are maintained by natural selection.

The functional significance of the coexpression clusters remains poorly understood. For example, the coexpression clusters observed in *D. melanogaster* did not tend to contain genes with similar functions according to their Gene Ontology annotation [143]. Other studies have described large-scale clustering of functionally related genes in animal genomes, but the relation of those clusters to the coexpression clusters has, to my knowledge, not been investigated. Lee and Sonnhammer [148] analyzed the distribution of genes in the genomes of five different species - human, *D. melanogaster*, *C. elegans*, *Arabidopsis thaliana* and *Saccharomyces cervisiae* - and showed that, in each of these genomes, genes involved in the same pathway tend to be less dispersed than expected if genes were randomly ordered. It is not clear how this observation relates to the coexpression clusters found by others, because the clusters observed by Lee and Sonnhammer may be very diffuse. Petkov et al. [149, 150] analyzed linkage in inbred mouse strains and found that some allelic combinations occur more often than expected by chance among genes in the same chromosomal region. They found several hundred such regions across the mouse genome and argued that these regions contain clusters of functionally related genes. In support of this, some of the identified regions were enriched for genes annotated to the same pathway.

1.4.2 Bidirectional promoters

One possible mechanistic explanation for the occurrence of coexpression clusters is that the clustered genes share *cis*-regulatory elements. Intriguingly, the human genome contains over a thousand pairs of neighboring genes that are transcribed in opposite orientation and are so closely spaced that it seems likely that they share upstream regulatory elements [151-153]. The upstream regions of such genes are therefore called bidirectional promoters. Investigators surveying the human genome for bidirectional promoters have typically searched for divergently transcribed genes with their TSSs separated by less than 1 kb, and found that the distribution of distances between TSSs has a peak between 100 and 300 bp [151-153]. Experiments measuring the activity of truncated bidirectional promoters to drive expression in a reporter gene assay indicate that the elements required to drive expression in one direction are either are nested or overlapping with those required to drive transcription in the other direction [153, 154]. It seems that these arrangements are not readily disentangled in evolution, because most bidirectionally promoted pairs found in human appear to have mouse orthologs in the same arrangement [153]. Because of the relatively limited number of chromosomal rearrangements that have occurred since the divergence of human and mouse, however, this observation should be confirmed in more distant or multiple-species comparisons. Many human gene pairs in bidirectionally promoted arrangement appear to be conserved in order and orientation in the chicken or fugu genome [155, 156], but it has not been determined to what extent the spacing between paired genes is conserved. Such an analysis is currently difficult to carry out because of the lack of data on TSS locations in those genomes.

The set of genes transcribed from bidirectional promoters appears to be enriched for genes with housekeeping functions – for example, more than 5-fold enrichment for genes annotated to be involved in DNA repair has been found. Most bidirectional promoters (70-80%) overlap a CpG island [151-153] and certain transcription factor binding sites preferentially occur in bidirectional promoters [154]. These observations are consistent with a common mode of regulation across the entire set of bidirectional promoters, and the genes expressed from these promoters may be represented in some of the clusters of coexpressed housekeeping genes discussed in the previous section.

Are genes transcribed from the same bidirectional promoter coregulated? There are several examples of pairs of functionally related and coordinately expressed genes that share a bidirectional promoter (reviewed in [153]). Recent studies have investigated the correlation within bidirectionally promoted gene pairs on a larger scale. Takai and Jones [152] used SAGE and EST data to analyze the expression of genes transcribed from bidirectional promoters on human chromosomes 20, 21 and 22, and did not see any trend for genes sharing a bidirectional promoter to be coexpressed compared to several control sets. On the other hand, other investigators have used gene expression data from different microarray studies to analyze genome-wide sets of bidirectional promoters and noted a trend for genes sharing a bidirectional promoter to be coexpressed more often than randomly paired genes [153, 154, 156]. However, the positive correlations between expression profiles were only significant for a minority of bidirectional promoters (17%-36% of promoters at $P < 0.05$, without correction for multiple testing) and two of the studies [153, 156] noted that a smaller subset of the

pairs have significant negative correlations between their expression profiles. In summary, while some bidirectional promoters appear to underlie coregulation of associated genes, it is not clear to what extent this is a general phenomenon.

1.4.3 Overlapping genes

Metazoan genomes also contain numerous loci where oppositely transcribed genes overlap [49, 54, 69-71, 157-160]. Overlapping genes may share exon sequence or just transcribed regions. Genome-wide surveys for overlapping and oppositely transcribed genes have typically focused on pairs of overlapping genes that share exon sequence (*cis*-antisense pairs) [49, 54, 69-71, 157-160]. One reason behind the interest in *cis*-antisense pairs is that they give rise to mature transcripts that are partially complementary (antisense transcripts) and therefore may form duplexes in the cell [161, 162]. Duplex formation could serve to regulate transcript stability or other aspects of gene expression. Duplexes can also be formed by partially complementary transcripts emanating from different loci (*trans*-antisense pairs); microRNAs could be considered to belong to this category [21]. Based on cDNA and EST data, Chen et al. [71] estimated that up to 22% of human genes are involved in *cis*-antisense pairs. Results presented in this thesis (Papers II and III) and data from tiling array experiments [54] suggest that *cis*-antisense pairs are even more common in the human genome. Surveys of transcript sequence data for other organisms indicate that *cis*-antisense pairs are abundant in most vertebrate genomes and in fly genomes, and that a few hundred also exist in nematode genomes [64, 69, 159, 160].

A number of regulatory mechanisms have been proposed for *cis*-antisense pairs [161, 162]. As mentioned above, complementary transcripts may form duplexes that impact transcript stability. RNA duplexes can also be targets for editing by the ADAR enzyme. Duplex formation could occur before maturation of one or both transcripts and thereby affect splicing. Transcripts from *cis*-antisense pairs may also interact with DNA to mediate DNA methylation or histone modifications that result in silencing or monoallelic inactivation. In addition, transcription of *cis*-antisense pairs may be affected by collision of RNA polymerases that transcribe opposite strands. However, few *cis*-antisense pairs have been studied in detail and evidence for the most of the mechanisms listed above is limited [161, 162]. One relatively well understood example relates to inactivation of one X chromosome in the female embryo of placental mammals [163]. X chromosome inactivation involves accumulation of noncoding transcripts from the *Xist* gene along the X chromosome designated for silencing. *Xist* overlaps another noncoding gene, *Tsix*, the transcription of which blocks *Xist* expression from the same chromosome. *Tsix* transcription thereby determines which of the two X chromosomes will be silenced. Another intriguing example from mammals involves the thyroid hormone receptor gene *Thra* and its overlapping gene *Nr1d1* [164]. *Thra* gives rise to two transcript isoforms, of which one encodes a functional receptor (Trα1), while the other encodes an antagonistic protein (Trα2) that does not bind thyroid hormone. Trα2 transcripts contain regions complementary to *Nr1d1* transcripts, but Trα1 transcripts do not. Reconstruction of *Thra* splicing *in vitro* and overexpression experiments in cell lines suggest that transcripts from *Nr1d1* inhibit maturation of the Trα2 transcript isoform, thereby altering the ratio of Trα1 to Trα2 transcripts. Alterations in this ratio may affect the response to thyroid hormone.

Several studies have explored genome-wide properties of *cis*-antisense pairs in search of trends that could imply a regulatory role of *cis*-antisense overlaps in general. Neeman et al. [165] compared human cDNA sequences to the genome sequence and recorded differences indicative of RNA editing. The great majority of editing events detected by this strategy were confined to expressed Alu repeats and the incidence of editing was not significantly different between *cis*-antisense pairs and other genes, suggesting that antisense regulation does not, in general, occur through RNA editing. A role in imprinting might be more plausible for *cis*-antisense pairs, as they appear to be enriched for known or predicted imprinted genes in both human and mouse genomes ([159] and Paper II). In addition, an analysis of SAGE tags showed that pairs of human complementary transcripts have been detected in the same tissue slightly more often than expected by chance, lending some support to mechanisms that involve coexpression of transcripts from opposite strands [166]. Interestingly, the coexpressed pairs appeared to be conserved in the mouse genome more often than non-coexpressed pairs.

1.4.4 Chromosomal regulatory domains

Several findings mentioned above imply a widespread existence of chromosomal domains that contain multiple genes regulated in a similar manner and/or contain multiple dispersed *cis*-regulatory elements that together specify the expression program of a single gene. Some additional observations that support a regulatory domain architecture of metazoan chromosomes are outlined here.

Lercher et al. [167] showed evidence that clusters of broadly expressed human genes are located in chromosomal regions corresponding to negative or very light staining Giemsa bands, which indicate an open chromatin structure. These regions also tend to have an elevated density of known genes [1]. Similar observations were made by Gilbert et al. [168], who determined regions of open chromatin in the human genome and found them to correspond regions of high gene density. A recent report describes measurements of expression levels of reporter constructs inserted into regions of the human genome where most genes are either highly or weakly expressed [169]. The expression levels of the reporter constructs correlated with the overall endogenous expression in the region, but were not as well explained by the activity of the immediate neighboring genes only. The domains of high expression identified in this study appear to correspond to gene-dense regions with open chromatin and broadly expressed genes found by others [144, 167, 168]. The human genome contains many other regions that, in contrast, are gene-poor, enriched for HCNEs and correspond to blocks of conserved synteny [170, 171]. It has been proposed these synteny blocks have been maintained to preserve linkage of long-range enhancers with their target genes [170-172].

Another example highlighting the occurrence of regulatory domains comes from genome-wide mapping of binding sites for Polycomb group proteins in human and *D. melanogaster*. These proteins constitute a family of transcriptional repressors with important roles in development of mammals and flies [173]. ChIP experiments have shown that Polycomb Repressive Complex 2 is distributed over much of the length of

key developmental regulatory genes in human embryonic stem cells [79]. In *D. melanogaster* embryonic cell lines, Polycomb was found by DamID to bind large regions also coinciding with developmental genes, but often large enough to span multiple genes [80]. It was mentioned above that CpG islands are common at mammalian core promoters. However, some mammalian developmental regulatory genes overlap CpG islands over much of their length; these CpG islands are highly conserved among mammals and correlated with binding sites for Polycomb Repressive Complex 2 [91], indicating a link between gene silencing and sequence evolution in these regions.

2 PRESENT INVESTIGATION

The main aim of the present investigation has been to answer a number of open questions about the organization of genes and regulatory elements in animals:

- How abundant are gene overlaps, bidirectional promoters and large regulatory domains in animal genomes?
- To what extent are these structures maintained in evolution?
- Do gene overlaps underlie some regulatory process, e.g. natural antisense regulation?

To this end, we have made use of the extensive transcript and genome sequence data that has become available in recent years. In summary, we have found that mammalian genomes contain an abundance of nested and overlapping gene structures, giving rise to both coding and noncoding transcripts, but only a minority of these arrangements appear to be well conserved between human and mouse. We have also found that *cis*-regulatory elements controlling the expression of developmental regulatory genes are frequently distributed within or beyond other genes, and this arrangement underlies maintenance of large synteny blocks in evolution. The following sections give a more detailed summary of the work. Since several of the underlying papers are the result of collaborations where several authors have contributed to a large extent, the summary below emphasizes those aspects of the papers to which I have contributed the most.

2.1 COMPLEX GENE ARRANGEMENTS IN HUMAN AND MOUSE GENOMES

To explore the incidence, conservation and other properties of *cis*-antisense pairs and bidirectional promoters, we focused on human and mouse, for which complete genome sequences and extensive transcript sequence data were available.

2.1.1 Detection of *cis*-antisense pairs from sequence data (Papers I-III)

When we took up an interest in natural antisense transcripts, it had already been reported that *cis*-antisense pairs are common in mammalian genomes [49, 69, 70, 157, 158]. However, no large-scale assessment of the evolutionary conservation of these arrangements had been performed. In addition, because the analyses had been based on cDNA sequences [69], reference collections of known mRNAs [157, 158] or EST data filtered in ways that excluded many 5'-ESTs [49, 70], we suspected that many gene overlaps had not been detected. We therefore constructed a pipeline for extracting reliable transcript sequences from cDNA and EST data, and designed the pipeline so that it would be applicable to both human and mouse data.

A major obstacle in screening for overlapping genes is the presence of reversal artifacts in cDNA and EST databases (see section 1.1.1 above). In determining the correct orientation of transcripts sequences, our pipeline looks for poly-A tails and polyadenylation signals in transcript sequences and splice signals in genome sequences. It also considers read direction of annotated ESTs, but only for cDNA libraries for which read direction annotations are estimated to be reliable. ESTs from the same cDNA clone are treated together; this allows orientation of both 5'- and 3'-ESTs from

the same cDNA clone based on features identified from only one of the ESTs (for example, a poly-A tail in a 3'-EST). The transcript sequences that passed through our filtering procedure were combined into TUs based on their genomic mappings, and *cis*-antisense pairs were identified by searching for oppositely transcribed TUs with that overlapped by at least 20 bp within exons (others have used a similar criterion [69, 70]).

The first version of our pipeline is outlined in a supplement to Paper I (included in this thesis) and the final version is described in detail in Paper III, where its accuracy is also assessed by using orientation-specific RT-PCR to investigate the expression of complementary transcripts corresponding to a random sample of 20 *cis*-antisense pairs. We were able to amplify transcripts from both strands for 16 of he 20 *cis*–antisense pairs, suggesting that at least 80% of the pairs found by the pipeline are expressed from both strands. This may well be an underestimate, as some of the transcripts might be expressed at low levels and some primers might not have worked. Indeed, simulations estimated that the pipeline correctly determines the orientation of 99.8% of the unspliced cDNAs and 99.8% of the unspliced ESTs that are not rejected by the pipeline. It is expected that spliced sequences are processed with even higher accuracy.

2.1.2 Noncoding natural antisense transcripts for RNAdb (Paper I)

We first applied the pipeline to the cDNA and EST data publicly available in 2004 (Paper I). This was done as part of an effort to construct a database of mammalian ncRNAs. The incentive for building this database was the rising awareness of roles for ncRNAs in gene regulation [174]. While databases of canonical infrastructural RNAs (e.g. transfer RNAs and ribosomal RNAs) already existed, we wanted to build a database focused on RNAs with putative regulatory functions and RNAs classified as noncoding by computational screens, but which had not been further characterized. Such a database would constitute a platform for further computational and experimental studies.

For this purpose, we were specifically interested in identifying noncoding natural antisense transcripts. Kiyosawa et al. [69] had demonstrated the existence of many such transcripts in mouse, but other studies [49, 70, 157, 158] had focused on mRNAs or had not discriminated between coding and noncoding antisense transcripts. To identify a set of high-confidence natural antisense ncRNAs, we applied stringent criteria for screening out transcripts that might encode protein or represent noncoding fragments of protein-coding transcripts. While lack of a gold-standard set of long noncoding RNAs precluded a rigorous assessment of the screening method, we designed the criteria to match those an annotator would consider when classifying a transcript as coding or noncoding.

The first release of the database contained 668 and 624 mouse putative natural antisense ncRNAs forming 579 and 571 distinct TUs, respectively. These counts were lower than the number of previously reported putative natural antisense ncRNAs from mouse (870 *cis*-antisense pairs containing one or two noncoding transcripts)[69], reflecting our more stringent filtering. Other sets of known and putative ncRNAs were gathered by collaborators at the University of Queensland. The resulting database,

RNAdb, can be accessed at http://research.imb.uq.edu.au/RNAdb. The current release of RNAdb [175] contains an updated set of putative antisense ncRNAs - 1068 from human and 1615 from mouse, forming 919 and 1395 TUs respectively - from the final pipeline described in Paper III.

2.1.3 Prevalence and conservation of *cis*-antisense pairs (Papers II and III)

We continued the research on *cis*-antisense transcripts while participating in the FANTOM consortium coordinated by RIKEN, Japan to annotate and analyze novel transcript sequence data produced by RIKEN for the FANTOM3 project. Our pipeline for determining transcript sequence orientation was adopted into the cDNA and EST processing pipeline in the FANTOM3 project, and therefore underlies the analysis in Paper II, as well as Paper III and other related papers [30, 113, 176].

We found 29% of all mouse TUs inferred from cDNA to share transcribed sequence with a cDNA mapped to the other genomic strand, and 19% to also share exon sequence (Paper II). When we built TUs from both cDNA and ESTs sequences, the latter number increased to 25%; we obtained the same percentage for human (Paper III). The total *cis*-antisense pair counts (6141 for human and 5248 for mouse) from with these cDNA- and EST-based TUs were twice as high as previously reported [69-71]. Since sequencing of the human and mouse transcriptomes has not yet reached saturation, we attempted to use available sequences to estimate the true proportion of TUs that are involved in *cis*-antisense pairs. Three different sampling methods each gave an estimate of about 40% for both human and mouse, even though the amount and average quality of cDNA and EST sequences differs between the two organisms. It should be noted that this is an estimate of the percentage of TUs sharing *exon sequence* with a TU on the opposite strand. The percentage of TUs with evidence for *transcription* on both strands appears to be even higher, even if one does not extrapolate beyond currently available data: collaborators incorporated short sequence tags (CAGE and PET) into the analysis and found that, depending on the stringency of the analysis, up to 72% of mouse TUs inferred from cDNA had evidence of transcription on both strands (Paper II).

Other studies published around the same time found that more human *cis*-antisense pairs consisted of genes overlapping tail-to-tail (convergently transcribed) than genes overlapping head-to-head (divergently transcribed) [70, 177]. On the other hand, we found a roughly equal proportion of each of these two classes (Paper III). This result was consistent between human and mouse. The result was also consistent when we used only mouse cDNA data (Paper II), indicating that it is not due to artifactual ESTs slipping through our pipeline. There is a more likely explanation for the disagreement between studies: considering that the other studies (like us) used 3'-end proximal sequence features to verify sequence orientation, but (unlike us) did not associate 5'- and 3'-ESTs originating from the same cDNA clone, it is likely that the other studies were biased against inclusion of 5'-ESTs and therefore failed to detect many head-to-head overlaps.

Consistent with the earlier study of Kiyosawa et al. [69], which was based on fewer cDNA sequences, we found that most *cis*-antisense pairs appear to consist of one

coding and one noncoding TU, suggesting that the role of these noncoding TUs might be to regulate mRNA expression from the opposite strand (Paper II). In agreement with a study published around the same time as ours [71], we found that some functional categories of genes (in particular genes encoding intracellular and/or catalytic proteins) more frequently share exon sequence with transcripts from the other strand. However, this result should be interpreted with caution because it may be affected by biased coverage of the human and mouse transcriptomes in current databases, as well as differences in gene length between functional gene categories.

We found a striking agreement between our human and mouse datasets regarding proportions of TUs involved in *cis*-antisense as well as the structural properties of *cis*-antisense overlaps (Paper III). We also detected nearly a thousand *cis*-antisense pairs that are conserved between human and mouse - close to three times more than previously reported [177]. While numerous, the conserved pairs constitute a minority (17%) of pairs found in each species. By a sampling approach, we estimated that only ~25% of the pairs discovered in human are conserved in mouse, indicating flexibility of antisense gene arrangements in evolution.

At more than 1400 human loci and 1100 mouse loci, we observed that gene overlaps and/or bidirectional promoters occur between three or more neighboring TUs, so that they appear "chained" together (Paper II and III). To our knowledge, only a few such examples had been described previously [178]. While the functional significance of these structures is unclear, their existence is intriguing, since bidirectional promoters can mediate coregulation and gene overlaps might also underlie regulatory mechanisms (see section 1.4.3 and experimental work by collaborators in Paper II). The tendency for genes in *cis*-antisense pairs to be coexpressed across tissues is consistent with a regulatory potential of the overlaps ([166] and Paper III and work by collaborators in Paper II).

2.1.4 Properties of bidirectional promoters (Paper III)

Using the TUs constructed from human and mouse cDNAs and ESTs, we found 10% and 9% of human and mouse TUs to be in bidirectionally promoted arrangement (Paper III), consistent with earlier work [151-153]. For a detailed analysis, we assembled a dataset of 766 putative bidirectional promoters that were well supported by mouse CAGE tag data. Compared to a control set of unidirectional promoters, the bidirectional promoter TCs showed a markedly larger dispersion of CAGE-determined TSS locations. Consistent with this finding, nearly all (94%) bidirectional promoters were associated with CpG islands. In addition, CpG islands at bidirectional promoters were significantly larger than CpG islands at unidirectional promoters (median CpG island sizes of 760 and 557 bp, respectively). While it was known from before that many CpG island promoters contain large transcriptional initiation regions [86], here we provided genome-wide evidence for this phenomenon at bidirectional promoters. To confirm that CAGE tag distributions can reliably indicate the sizes of transcriptional initiation regions, we used quantitative real-time PCR to measure expression levels of transcripts for the genes *Ddx49* and *Cope*, which share a bidirectional promoter. Expression levels obtained with primers for different positions within the initiation regions and downstream matched the observed distribution of CAGE tags.

We further showed that the two divergently oriented TSS regions of a bidirectional promoter rarely overlapped, but were often closely spaced: the distribution of distances between paired TSSs peaked between 0 and 60 bp. This suggests that earlier studies had overestimated the distance between start sites [151-153] and shows the added value of the deeper sampling of TSS locations provided by CAGE. Since even the CAGE data does not cover all existing TSSs, it is possible that the typical distance between TSS regions is even smaller than our results indicated. By aligning the entire set of bidirectional promoters, we showed that their overall sequence composition changes at the midpoint between the divergently oriented TSS regions, so that each side has more guanines than cytosines on the sense strand. This mirror-image sequence composition might be a result of transcription-coupled DNA repair in germline cells [179], and could be useful for prediction of bidirectional promoters in genome sequences [59].

2.2 GENOMIC REGULATORY BLOCKS IN VERTEBRATE AND INSECT GENOMES

As described in section 1.2.4, we and others had found vertebrate and insect genomes to contain highly conserved noncoding elements (HCNEs) that tend to cluster in large arrays spanning the loci of developmental regulatory genes. In Papers IV and V, we explore the conservation of these arrays, and ask how they relate to synteny blocks and gene order conservation. In Paper V, we also examine differences between core promoters that may underlie enhancer-promoter specificity. In Paper VI, we present a web resource for exploring HCNEs and their association with developmental regulatory genes.

2.2.1 Genomic regulatory blocks in vertebrates (Paper IV)

Previous studies had found that many large synteny blocks are conserved among mammals, and suggested that these blocks may have been maintained by negative selection (see section 1.3.3). We wished to investigate whether the existence of large synteny blocks might be explained by a pressure to maintain HCNE arrays, many of which span multiple genes. To this end, we chose to investigate synteny blocks that, like numerous HCNEs, are conserved between human and fish. We therefore identified HCNEs and synteny blocks conserved between human and zebrafish genomes. Because we were interested in rearrangements of both coding and noncoding sequence, we defined synteny blocks based on direct genome sequence comparisons (BLASTZ net alignments). By comparing the distributions of synteny block spans for different functional categories of genes, we showed that genes encoding developmental transcriptional regulators tend to be surrounded by larger regions of conserved microsynteny than other functional categories of genes ($P < 10^{-6}$). In addition, we examined the 100 largest synteny blocks and detected a putative developmental regulatory gene and associated HCNEs in almost every one of them. We introduced the term genomic regulatory blocks (GRBs) for HCNE-rich regions maintained in evolution.

This work was done in collaboration with the group of Tom Becker at the Sars Centre for Marine Molecular Biology in Bergen. They had carried out a retroviral enhancer

detection screen in zebrafish and obtained a number fish lines where insertions far from developmentally regulated genes nevertheless showed developmental expression patterns [95]. Several of these insertions were located in large synteny blocks where HCNEs (like the insertions) are distributed within and around developmental regulatory genes, even beyond neighboring unrelated genes and in their introns. This finding suggests that regulatory information encoded in HCNEs can be embedded in large areas that include multiple additional genes around the genes targeted by that regulatory information. For several synteny blocks, we analyzed the fate of duplicated genes and HCNEs after whole-genome duplication in teleosts, and showed that the unrelated genes – which we call "bystander genes" - are indeed independent of the regulatory input of HCNE arrays. The conclusion from this work is that bystander genes have been kept in proximity to HCNE target genes because regulatory information for target genes is contained beyond bystander genes or in their introns.

2.2.2 Genomic regulatory blocks in insects (Paper V)

As mentioned in section 1.2.4, HCNEs are also abundant and associated with developmental regulatory genes in fly genomes [114]. In addition, a comparison of draft genome sequences for 12 insects had indicated that the distribution of insect synteny block lengths is incompatible with the random breakage model of chromosome evolution [136]. To investigate whether HCNE arrays may explain the occurrence of large synteny blocks also in insects, we carried out a genome-wide analysis of HCNEs and synteny blocks conserved among five *Drosophila* species, of which the most distantly related are estimated to have diverged about 40 million years ago [180]. We identified peaks of HCNE density along chromosomes, and found that that such peaks tend to be centrally located in large synteny blocks containing multiple genes. This observation can not be explained by lower alignment quality close to synteny breaks, because we found that protein-coding sequence that aligns in a reciprocal-best manner between *D. melanogaster* and each of the four other investigated fly genomes tends to be concentrated near synteny breaks. These findings strongly suggest that large regions containing multiple genes have maintained microsynteny in order to preserve arrays of HCNEs. Hence, GRBs also exist in *Drosophila*.

We found evidence that some of these GRBs have been maintained even between flies and mosquitoes, which are estimated to have diverged about 250 million years ago [128]. Although few noncoding elements are highly conserved between *Drosophila* and the malaria mosquito *Anopheles gambiae*, we could show that *A. gambiae* regions orthologous to *Drosophila* GRBs contain an equivalent distribution of noncoding elements highly conserved in the yellow-fever mosquito *Aëdes aegypti* and coincide with regions of ancient microsynteny between *Drosophila* and mosquitos. We estimated fly-mosquito synteny blocks genome-wide, and demonstrated that genes associated with HCNEs in *Drosophila* tend to be located in large fly-mosquito synteny blocks.

At several GRBs, we observed a striking correspondence between boundaries of synteny blocks, HCNE arrays and Polycomb binding domains determined by DamID [80]. These examples indicate that synteny blocks, HCNE arrays and Polycomb binding regions can independently pinpoint the same large regulatory domains in insect

genomes, suggesting that they reveal different aspects of the same evolutionarily conserved regulatory mechanism.

The insect GRBs also contain unrelated genes, probably in a similar way to bystander genes in vertebrate GRBs. In Paper IV, enhancer detection experiments in zebrafish demonstrated that regulatory information for target genes is present within and beyond bystander genes. Since enhancer detection has been performed extensively in *Drosophila*, we searched for examples of such insertions near bystander genes in the literature. The most striking example we found was an insertion in the 5'-UTR of *out at first* [99]. This insertion replicates part of the expression pattern of *decapentaplegic*, a developmental regulatory gene located 33 kb away. Both genes are in the same synteny block, which contains multiple HCNEs that have been characterized as long-range enhancers for *dpp* [99].

2.2.3 Enhancer-promoter specificity (Paper V)

It is unknown how enhancer activity is specifically directed towards certain genes at HCNE-spanned loci. To investigate the possibility that enhancers in HCNE arrays may target specific genes within "striking distance" on the basis of their core promoter architecture (see section 1.2.3), we classified *D. melanogaster* genes by core promoter type. For the classification, we made use of genome-wide core promoter predictions from the program McPromoter, which predicts promoters in fly genome sequences and classifies the predictions into five categories based on motif content and sequence composition [181]. We then used Gene Ontology annotation [182] to analyze core promoter assignments for different functional categories of genes. Of all genes that were annotated as developmental transcriptional regulators, located in a HCNE-dense region and assigned a core promoter prediction, we found that 95% had a core promoter containing an Inr motif. For comparison, only 39% all protein-coding genes assigned a core promoter prediction had a prediction with an Inr motif. Further analysis revealed that genes in the "TATA/Inr" core promoter class tend to have tissue-specific functions, while other genes with Inr-containing core promoters (classes "Inr/DPE" and "Inr-only") tend to be associated with development. Genes in the remaining two classes predicted by McPromoter ("DRE" and "Motif 1/6") tend to have housekeeping functions. Based on these results, we speculated that it is the Inr-type of core promoters without TATA boxes that are most likely to respond to long-range regulation.

2.2.4 A web resource for exploring HCNEs in metazoan genomes (Paper VI)

Despite a rising interest in HCNEs in the genomics and evo-devo community, there has been a lack of resources that provide information about HCNEs. To fill this gap, we built the web resource Ancora, which is available at http://ancora.genereg.net. Ancora contains a genome browser designed for exploring the distribution of HCNEs on metazoan chromosomes. The browser is currently set up to show the genomes of human, mouse, zebrafish and *D. melanogaster*. To put HCNEs in context, the browser also shows gene models, synteny blocks, CpG islands and other selected annotation tracks.

In addition to HCNE locations, the browser also shows densities of HCNEs along chromosomes. Such HCNE density plots highlight regions that harbor large HCNE arrays and thus are likely to contain key developmental regulatory genes and correspond to regulatory domains ([10, 11, 108] and Papers IV and V). Unlike conservation profiles, which can be seen in several other genome browsers [13, 52, 126, 170], HCNE density plots do not directly reflect conservation on sequence level; instead, they show density distributions of HCNEs on a larger scale. The result is qualitatively different: it clearly reveals chromosomal regions of extensive noncoding conservation and points to approximate extent of GRBs, as well as the most likely target gene(s) within those regions. We built the Ancora genome browser using the GBrowse software [183], which we extended with plugins and custom glyphs designed to visualize HCNE data in the most informative manner and to efficiently plot HCNE densities along entire chromosomes. For example, the user can activate an option that separates HCNE density plots based on chromosome in the other genome. The result is an overview of how HCNE-dense regions have been partitioned over different chromosomes in evolution. Based on the assumption that fundamental regulatory domains have been maintained in evolution ([171, 172] and Papers IV and V), the displayed separation of HCNE-dense regions across chromosomes should correspond to a separation of distinct regulatory domains.

HCNE locations and densities are available for download. In addition, we aimed to make it as easy as possible for users to visualize this data in other genome browsers. The downloadable data files can be directly used as custom tracks in the UCSC Genome Browser [13]. Ancora also includes a service that allows users to view much of the HCNE data in Ensembl [52] through the distributed annotation system (DAS) protocol for sharing sequence annotations.

3 PERSPECTIVES

The results presented here describe an abundance of overlapping genes and genes sharing promoters in the human and mouse genomes. Numerous regions contain chains of multiple genes associated in this manner. This thesis also shows that, in both vertebrate and insect genomes, regulatory elements for single genes are frequently dispersed within and beyond other genes. Together with other recent studies in human, mouse and Drosophila, these findings challenge the canonical "colinear" model of how genes and their regulatory elements are arranged in metazoan genomes [184]. To a computational biologist, an appealing comparison was made by Gerstein et al. [19], who likened genome sequences to a computer program that has been written in a sloppy manner, e.g. by using many GOTO statements to make jumps in the code. This notion is, of course, also compatible with the process of evolution, which has little respect for style as long as the end result is successful. Although functionally related elements can be distant in the sequence, they may be closer together in the chromatin structure. Emerging methods for determining which DNA segments are associated in the cell will likely provide insights into this [185].

Antisense transcription is no longer a curiosity. Work in this thesis shows that a quarter of all human and mouse TUs share exon sequence with a TU on the other strand. It is clear that this is an underestimate: our extrapolation beyond currently available cDNA and EST data (Paper III), evidence from CAGE and PET sequences (Paper II) and evidence from tiling array experiments [18, 54] independently suggest that the actual fraction of genes that share exons with genes on the other strand is much higher. If these arrangements indeed have regulatory implications, as has been proposed [162], most genes could be affected by transcription from the other strand. However, the evidence for a regulatory role of mammalian antisense transcripts is limited to a few experimentally studied cases [162]. The abundance of gene overlaps calls for large-scale functional screens to investigate their relevance and single out promising cases. If *cis*-antisense overlaps imply regulation at the post-transcriptional level, i.e. through hybridization of complementary transcripts, it might be possible to perturb such interactions by knocking down each transcript with small interfering RNA (siRNA), as was done in Paper II. Large-scale siRNA knockdown screens are now feasible [186]. Can bioinformatics be of any help in this regard? Conservation is often a good indicator of function [187]. Since most *cis*-antisense pairs do not appear to be well conserved (Paper III), conservation might be an efficient filter for highlighting functionally important *cis*-antisense pairs. In Paper III, we identified close to a thousand *cis*-antisense pairs conserved between human and mouse. These may be a good starting point for large-scale knockdown screens. In addition, we noted in Paper III that a few pairs of transcripts originating from opposite strands of the same locus exhibit coordinated expression profiles and high levels of conservation in their overlapping regions. These might be suitable targets for smaller-scale experiments, again under the hypothesis that conservation could indicate function in the context of gene overlaps.

One study noted that complementary transcripts from *cis*-antisense pairs tend to be coexpressed across tissues, but show inverse expression profiles within tissues [166]. This observation is related to a problem with applying gene expression analysis to *cis*-

antisense pairs and to gene clusters in general. *Cis*-antisense overlaps are proposed to have regulatory implications at the single-cell level, but virtually all available expression data is averaged over many cells, often including a mix of different cell types. Transcriptome profiling in single cells is possible, but technically challenging and unavailable to most research groups at the time [188].

In Paper III, we showed that mouse genes transcribed from bidirectional promoters tend to have broad transcription initiation regions. A parallel study demonstrated that this applies to genes transcribed from CpG-island promoters in general [46]. Even though these findings were anticipated before CAGE tags were available [189], many promoter studies still assume that TSSs are isolated and do not occur in broad clusters. For example, in the recent analysis of 1% of the human genome by the ENCODE consortium, the occurrence of broad TSS regions was recognized, but only a single representative TSS was chosen from each initiation region for further study [18]. *Cis*-regulatory sequences were found to be symmetrically distributed around the representative TSSs, but it is unclear how this symmetric distribution relates to the distribution of additional TSSs around the representative ones. A change in how gene models are represented in major genome browsers might increase the awareness that broad initiation regions are the rule rather than the exception.

Papers IV and V in this thesis demonstrate that GRBs exist in both vertebrates and insects. Others have shown that HCNEs are associated with orthologous developmental regulatory genes also in nematodes [115]. These observations suggest that HCNE arrays have shaped genome evolution across metazoans. However, the existence of HCNEs remains to be explored in several branches of the metazoan tree; our preliminary results indicate their existence in e.g. *Ciona* species. It also remains to be shown whether HCNEs arrays underlie synteny conservation outside vertebrates and insects. Intriguingly, extensive synteny conservation has been observed between human and the sea anemone *Nematostella vectensis*, a member of the oldest eumetazoan phylum [190].

As mentioned in section 1.2.4, it is a mystery why HCNEs are so highly conserved, because transcription factor binding sites can usually tolerate mutations to some extent [81]. Given the apparent importance of HCNEs in regulation of development, one possibility could be that a mechanism has evolved that keeps entire HCNEs preserved in order to guarantee that their transcription factor binding site content is maintained (i.e. this mechanism would maintain a higher level of conservation than actually needed). Another possibility could be that HCNEs are involved in interactions between homologous chromosomes [191]. If HCNEs are required to be virtually identical between sister chromatids, that would presumably slow down their divergence. However, both of these hypotheses are speculative and remain to be explored. The former hypothesis would have to be reconciled with observations of negative selection on ultraconserved sequences [116]. The comparison of orthologous fly and mosquito GRBs in Paper V demonstrates a difference in how HCNEs and protein-coding sequences diverge. Up to a certain evolutionary distance, HCNE sequences that are more similar than protein-coding sequences abound, but at larger distances orthologous HCNEs cannot be aligned, while orthologous protein-coding sequences can. The availability of a draft genome sequence for lamprey [192], as well as sequencing of

other species that are more distant from human than are teleost fish, will make it possible to further explore how HCNEs have diverged in vertebrate evolution.

A recent study where four ultraconserved elements were deleted in mice has cast some doubt on the importance of HCNEs [193]. Surprisingly, no phenotype was observed in the engineered mice. Since the HCNEs that were removed are proposed to regulate gene expression in the developing nervous system, it is possible that phenotypes were subtle. Indeed, disruption of other long-range *cis*-regulatory elements and breakpoints within GRBs, even if far from the target gene, can result in target gene dysregulation and severe phenotypes ([171, 172] and Paper IV). Many single nucleotide polymorphisms that affect gene regulation appear to exist in the human population [194], and some of these can be expected to affect long-range regulatory elements [172]. Association studies may therefore benefit from considering the genomic neighborhood of identified polymorphisms, up to distances on the order of a million base pairs in the human genome [93, 96]. This recommendation is motivated by the existence of many GRBs in the human genome (Papers IV and VI), but other, less well understood aspects of interleaved genome organization revealed by this and other work [19, 184] also suggest that individual functional genetic elements should be considered in their genomic context.

4 ACKNOWLEDGEMENTS

During the past five years, I have had the honor to work with many fantastic people. The list below is not exhaustive.

First and foremost: many thanks to Boris Lenhard for taking me on as your first PhD student and providing challenging project plans, brilliant ideas and excellent guidance. Your deep knowledge and true dedication has been a great source of inspiration. I am particularly grateful for the exposure to an exceptionally broad range of research topics in genomics and bioinformatics and for your confidence in me throughout.

Thanks to Bengt Persson for taking the responsibility as co-supervisor, and always being friendly and helpful.

Participating in the FANTOM consortium has been an extraordinary experience. I am grateful to Boris Lenhard and Claes Wahlestedt for encouraging my participation, to Yoshihide Hayashizaki, Piero Carninci and Harukazu Suzuki for welcoming me into the consortium, and to the many people who I have collaborated with in related and resulting projects – including Leonard Lipovich, John Mattick, Martin Frith, Ken Pang, Stuart Stephen, Christine Wells, Shintaro Katayama, Shinji Kondo, Hidenori Kiyosawa, Noriko Ninomiya, Takeya Kasukawa, Norihiro Maeda, Vlad Bajic, James Reid, Alessandro Brozzi, Lucilla Luzi and Valerio Orlando.

Past and present group members: Albin, Wynand, Bill, Ying, Kairi, David, Xianjun, Jan Christian and Chirag, as well as short-term members Johan, Christian, Ezzat, Sara, Grigori, Nina, Joao and Sandra. Thanks to all of you for creating a cheerful and friendly environment, providing valuable feedback and for your understanding in hectic periods! Many thanks to Christian, Ezzat and Altuna for enjoyable lunches at KS Pizza, Zen Café and elsewhere; to Albin, Altuna, Ying and David for collaborations on various projects, many stimulating discussions and memorable adventures in faraway land.

Thanks to the many others I have had the privilege to work with during this time, including: Tom Becker, Hiroshi Kikuta, Øyvind Drivenes, Pavla Navratilova and Anna Komisarczuk at Sars; Claes Wahlestedt, Salim Mottagui-Tabar, Yosuke Mizuno and Mohammad Ali Faghihi at CGB; Malin Andersen and Jacob Odeberg at KTH; Wyeth Wasserman, Shannan Ho Sui and Dave Arenillas at UBC. Thanks in particular to Wyeth for first introducing me to bioinformatics together with Albin and Johan, and encouraging me to venture into PhD studies!

Thanks also to everyone else who made CGB a pleasant and stimulating research environment, and everyone who has contributed to make CBU a great place to work during my time there. I did try to list all of you but the list became ridiculously long!

Thanks to family and friends for support on other levels during this time. Special thanks to Ola, Erik, Johan, Marko and Pendo for occasionally abducting me from work and providing perspectives on life.

Roses to Therese for encouragement, patience, laughs and love.

5 REFERENCES

1. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W *et al*: **Initial sequencing and analysis of the human genome**. *Nature* 2001, **409**:860-921.
2. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA *et al*: **The sequence of the human genome**. *Science* 2001, **291**:1304-1351.
3. **2003 Release: International Consortium Completes Human Genome Project** [http://www.genome.gov/11006929]
4. The C. elegans Sequencing Consortium: **Genome sequence of the nematode C. elegans: a platform for investigating biology**. *Science* 1998, **282**:2012-2018.
5. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF *et al*: **The genome sequence of Drosophila melanogaster**. *Science* 2000, **287**:2185-2195.
6. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P *et al*: **Initial sequencing and comparative analysis of the mouse genome**. *Nature* 2002, **420**:520-562.
7. Aparicio S, Chapman J, Stupka E, Putnam N, Chia JM, Dehal P, Christoffels A, Rash S, Hoon S, Smit A *et al*: **Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes**. *Science* 2002, **297**:1301-1310.
8. International Chicken Genome Sequencing Consortium: **Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution**. *Nature* 2004, **432**:695-716.
9. Jaillon O, Aury JM, Brunet F, Petit JL, Stange-Thomann N, Mauceli E, Bouneau L, Fischer C, Ozouf-Costaz C, Bernot A *et al*: **Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype**. *Nature* 2004, **431**:946-957.
10. Lindblad-Toh K, Wade CM, Mikkelsen TS, Karlsson EK, Jaffe DB, Kamal M, Clamp M, Chang JL, Kulbokas EJ, 3rd, Zody MC *et al*: **Genome sequence, comparative analysis and haplotype structure of the domestic dog**. *Nature* 2005, **438**:803-819.
11. Mikkelsen TS, Wakefield MJ, Aken B, Amemiya CT, Chang JL, Duke S, Garber M, Gentles AJ, Goodstadt L, Heger A *et al*: **Genome of the marsupial Monodelphis domestica reveals innovation in non-coding sequences**. *Nature* 2007, **447**:167-177.
12. Kasahara M, Naruse K, Sasaki S, Nakatani Y, Qu W, Ahsan B, Yamada T, Nagayasu Y, Doi K, Kasai Y *et al*: **The medaka draft genome and insights into vertebrate genome evolution**. *Nature* 2007, **447**:714-719.
13. Kuhn RM, Karolchik D, Zweig AS, Trumbower H, Thomas DJ, Thakkapallayil A, Sugnet CW, Stanke M, Smith KE, Siepel A *et al*: **The UCSC genome browser database: update 2007**. *Nucleic Acids Res* 2007, **35**:D668-673.
14. Church GM: **The personal genome project**. *Mol Syst Biol* 2005, **1**:2005 0030.
15. Watson JD, Crick FH: **Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid**. *Nature* 1953, **171**:737-738.
16. Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF *et al*: **Complementary DNA sequencing: expressed sequence tags and human genome project**. *Science* 1991, **252**:1651-1656.
17. Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, Axelrod N, Huang J, Kirkness EF, Denisov G *et al*: **The Diploid Genome Sequence of an Individual Human**. *PLoS Biol* 2007, **5**:e254.
18. Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, Thurman RE *et al*: **Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project**. *Nature* 2007, **447**:799-816.
19. Gerstein MB, Bruce C, Rozowsky JS, Zheng D, Du J, Korbel JO, Emanuelsson O, Zhang ZD, Weissman S, Snyder M: **What is a gene, post-ENCODE? History and updated definition**. *Genome Res* 2007, **17**:669-681.
20. Mattick JS, Makunin IV: **Non-coding RNA**. *Hum Mol Genet* 2006, **15 Spec No 1**:R17-29.
21. Jackson RJ, Standart N: **How do microRNAs regulate gene expression?** *Sci STKE* 2007, **2007**:re1.

22. Okazaki Y, Furuno M, Kasukawa T, Adachi J, Bono H, Kondo S, Nikaido I, Osato N, Saito R, Suzuki H *et al*: **Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs**. *Nature* 2002, **420**:563-573.
23. Kapranov P, Cheng J, Dike S, Nix DA, Duttagupta R, Willingham AT, Stadler PF, Hertel J, Hackermuller J, Hofacker IL *et al*: **RNA maps reveal new RNA classes and a possible function for pervasive transcription**. *Science* 2007, **316**:1484-1488.
24. Marra M, Hillier L, Kucaba T, Allen M, Barstead R, Beck C, Blistain A, Bonaldo M, Bowers Y, Bowles L *et al*: **An encyclopedia of mouse genes**. *Nat Genet* 1999, **21**:191-194.
25. Williamson AR: **The Merck Gene Index project**. *Drug Discov Today* 1999, **4**:115-122.
26. Camargo AA, Samaia HP, Dias-Neto E, Simao DF, Migotto IA, Briones MR, Costa FF, Nagai MA, Verjovski-Almeida S, Zago MA *et al*: **The contribution of 700,000 ORF sequence tags to the definition of the human transcriptome**. *Proc Natl Acad Sci U S A* 2001, **98**:12103-12108.
27. Brentani H, Caballero OL, Camargo AA, da Silva AM, da Silva WA, Jr., Dias Neto E, Grivet M, Gruber A, Guimaraes PE, Hide W *et al*: **The generation and utilization of a cancer-oriented representation of the human transcriptome by using expressed sequence tags**. *Proc Natl Acad Sci U S A* 2003, **100**:13418-13423.
28. Imanishi T, Itoh T, Suzuki Y, O'Donovan C, Fukuchi S, Koyanagi KO, Barrero RA, Tamura T, Yamaguchi-Kabata Y, Tanino M *et al*: **Integrative annotation of 21,037 human genes validated by full-length cDNA clones**. *PLoS Biol* 2004, **2**:e162.
29. Gerhard DS, Wagner L, Feingold EA, Shenmen CM, Grouse LH, Schuler G, Klein SL, Old S, Rasooly R, Good P *et al*: **The status, quality, and expansion of the NIH full-length cDNA project: the Mammalian Gene Collection (MGC)**. *Genome Res* 2004, **14**:2121-2127.
30. Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C *et al*: **The transcriptional landscape of the mammalian genome**. *Science* 2005, **309**:1559-1563.
31. Adams MD, Soares MB, Kerlavage AR, Fields C, Venter JC: **Rapid cDNA sequencing (expressed sequence tags) from a directionally cloned human infant brain cDNA library**. *Nat Genet* 1993, **4**:373-380.
32. Kent WJ: **BLAT--the BLAST-like alignment tool**. *Genome Res* 2002, **12**:656-664.
33. Bonaldo MF, Lennon G, Soares MB: **Normalization and subtraction: two approaches to facilitate gene discovery**. *Genome Res* 1996, **6**:791-806.
34. Carninci P, Kvam C, Kitamura A, Ohsumi T, Okazaki Y, Itoh M, Kamiya M, Shibata K, Sasaki N, Izawa M *et al*: **High-efficiency full-length cDNA cloning by biotinylated CAP trapper**. *Genomics* 1996, **37**:327-336.
35. Suzuki Y, Yoshitomo-Nakagawa K, Maruyama K, Suyama A, Sugano S: **Construction and characterization of a full length-enriched and a 5'-end-enriched cDNA library**. *Gene* 1997, **200**:149-156.
36. Pruitt KD, Tatusova T, Maglott DR: **NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins**. *Nucleic Acids Res* 2007, **35**:D61-65.
37. Schuler GD: **Pieces of the puzzle: expressed sequence tags and the catalog of human genes**. *J Mol Med* 1997, **75**:694-698.
38. Boguski MS, Lowe TM, Tolstoshev CM: **dbEST--database for "expressed sequence tags"**. *Nat Genet* 1993, **4**:332-333.
39. Bashiardes S, Lovett M: **cDNA detection and analysis**. *Curr Opin Chem Biol* 2001, **5**:15-20.
40. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW: **Serial analysis of gene expression**. *Science* 1995, **270**:484-487.
41. Saha S, Sparks AB, Rago C, Akmaev V, Wang CJ, Vogelstein B, Kinzler KW, Velculescu VE: **Using the transcriptome to annotate the genome**. *Nat Biotechnol* 2002, **20**:508-512.
42. Wei CL, Ng P, Chiu KP, Wong CH, Ang CC, Lipovich L, Liu ET, Ruan Y: **5' Long serial analysis of gene expression (LongSAGE) and 3' LongSAGE for transcriptome characterization and genome annotation**. *Proc Natl Acad Sci U S A* 2004, **101**:11701-11706.
43. Brenner S, Johnson M, Bridgham J, Golda G, Lloyd DH, Johnson D, Luo S, McCurdy S, Foy M, Ewan M *et al*: **Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays**. *Nat Biotechnol* 2000, **18**:630-634.
44. Shiraki T, Kondo S, Katayama S, Waki K, Kasukawa T, Kawaji H, Kodzius R, Watahiki A, Nakamura M, Arakawa T *et al*: **Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage**. *Proc Natl Acad Sci U S A* 2003, **100**:15776-15781.

45. Ng P, Wei CL, Sung WK, Chiu KP, Lipovich L, Ang CC, Gupta S, Shahab A, Ridwan A, Wong CH *et al*: **Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation**. *Nat Methods* 2005, **2**:105-111.
46. Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple CA, Taylor MS, Engstrom PG, Frith MC *et al*: **Genome-wide analysis of mammalian promoter architecture and evolution**. *Nat Genet* 2006, **38**:626-635.
47. Kozak M: **Interpreting cDNA sequences: some insights from studies on translation**. *Mamm Genome* 1996, **7**:563-574.
48. Aaronson JS, Eckman B, Blevins RA, Borkowski JA, Myerson J, Imran S, Elliston KO: **Toward the development of a gene index to the human genome: an assessment of the nature of high-throughput EST sequence data**. *Genome Res* 1996, **6**:829-845.
49. Shendure J, Church GM: **Computational discovery of sense-antisense transcription in the human and mouse genomes**. *Genome Biol* 2002, **3**:RESEARCH0044.
50. Sorek R, Safer HM: **A novel algorithm for computational identification of contaminated EST libraries**. *Nucleic Acids Res* 2003, **31**:1067-1074.
51. Hsu F, Kent WJ, Clawson H, Kuhn RM, Diekhans M, Haussler D: **The UCSC Known Genes**. *Bioinformatics* 2006, **22**:1036-1046.
52. Hubbard TJ, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T *et al*: **Ensembl 2007**. *Nucleic Acids Res* 2007, **35**:D610-617.
53. Bertone P, Stolc V, Royce TE, Rozowsky JS, Urban AE, Zhu X, Rinn JL, Tongprasit W, Samanta M, Weissman S *et al*: **Global identification of human transcribed sequences with genome tiling arrays**. *Science* 2004, **306**:2242-2246.
54. Cheng J, Kapranov P, Drenkow J, Dike S, Brubaker S, Patel S, Long J, Stern D, Tammana H, Helt G *et al*: **Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution**. *Science* 2005, **308**:1149-1154.
55. Manak JR, Dike S, Sementchenko V, Kapranov P, Biemar F, Long J, Cheng J, Bell I, Ghosh S, Piccolboni A *et al*: **Biological function of unannotated transcription during the early development of Drosophila melanogaster**. *Nat Genet* 2006, **38**:1151-1158.
56. Stolc V, Gauhar Z, Mason C, Halasz G, van Batenburg MF, Rifkin SA, Hua S, Herreman T, Tongprasit W, Barbano PE *et al*: **A gene expression map for the euchromatic genome of Drosophila melanogaster**. *Science* 2004, **306**:655-660.
57. Burge C, Karlin S: **Prediction of complete gene structures in human genomic DNA**. *J Mol Biol* 1997, **268**:78-94.
58. Korf I, Flicek P, Duan D, Brent MR: **Integrating genomic homology into gene structure prediction**. *Bioinformatics* 2001, **17 Suppl 1**:S140-148.
59. Glusman G, Qin S, El-Gewely MR, Siegel AF, Roach JC, Hood L, Smit AF: **A third approach to gene prediction suggests thousands of additional human transcribed regions**. *PLoS Comput Biol* 2006, **2**:e18.
60. Wei C, Lamesch P, Arumugam M, Rosenberg J, Hu P, Vidal M, Brent MR: **Closing in on the C. elegans ORFeome by cloning TWINSCAN predictions**. *Genome Res* 2005, **15**:577-582.
61. Guigo R, Flicek P, Abril JF, Reymond A, Lagarde J, Denoeud F, Antonarakis S, Ashburner M, Bajic VB, Birney E *et al*: **EGASP: the human ENCODE Genome Annotation Assessment Project**. *Genome Biol* 2006, **7 Suppl 1**:S2 1-31.
62. Wong GK, Passey DA, Huang Y, Yang Z, Yu J: **Is "junk" DNA mostly intron DNA?** *Genome Res* 2000, **10**:1672-1678.
63. Wong GK, Passey DA, Yu J: **Most of the human genome is transcribed**. *Genome Res* 2001, **11**:1975-1977.
64. Misra S, Crosby MA, Mungall CJ, Matthews BB, Campbell KS, Hradecky P, Huang Y, Kaminker JS, Millburn GH, Prochnik SE *et al*: **Annotation of the Drosophila melanogaster euchromatic genome: a systematic review**. *Genome Biol* 2002, **3**:RESEARCH0083.
65. Iseli C, Stevenson BJ, de Souza SJ, Samaia HB, Camargo AA, Buetow KH, Strausberg RL, Simpson AJ, Bucher P, Jongeneel CV: **Long-range heterogeneity at the 3' ends of human mRNAs**. *Genome Res* 2002, **12**:1068-1074.
66. Zavolan M, Kondo S, Schonbach C, Adachi J, Hume DA, Hayashizaki Y, Gaasterland T: **Impact of alternative initiation, splicing, and termination on the diversity of the mRNA transcripts encoded by the mouse transcriptome**. *Genome Res* 2003, **13**:1290-1300.
67. Stamm S, Ben-Ari S, Rafalska I, Tang Y, Zhang Z, Toiber D, Thanaraj TA, Soreq H: **Function of alternative splicing**. *Gene* 2005, **344**:1-20.

68. Kapranov P, Drenkow J, Cheng J, Long J, Helt G, Dike S, Gingeras TR: **Examples of the complex architecture of the human transcriptome revealed by RACE and high-density tiling arrays**. *Genome Res* 2005, **15**:987-997.
69. Kiyosawa H, Yamanaka I, Osato N, Kondo S, Hayashizaki Y: **Antisense transcripts with FANTOM2 clone set and their implications for gene regulation**. *Genome Res* 2003, **13**:1324-1334.
70. Yelin R, Dahary D, Sorek R, Levanon EY, Goldstein O, Shoshan A, Diber A, Biton S, Tamir Y, Khosravi R *et al*: **Widespread occurrence of antisense transcription in the human genome**. *Nat Biotechnol* 2003, **21**:379-386.
71. Chen J, Sun M, Kent WJ, Huang X, Xie H, Wang W, Zhou G, Shi RZ, Rowley JD: **Over 20% of human transcripts might form sense-antisense pairs**. *Nucleic Acids Res* 2004, **32**:4812-4820.
72. Lewis BP, Green RE, Brenner SE: **Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans**. *Proc Natl Acad Sci U S A* 2003, **100**:189-192.
73. Maston GA, Evans SK, Green MR: **Transcriptional Regulatory Elements in the Human Genome**. *Annu Rev Genomics Hum Genet* 2006, **7**:29-59.
74. Lee Y, Kim M, Han J, Yeom KH, Lee S, Baek SH, Kim VN: **MicroRNA genes are transcribed by RNA polymerase II**. *Embo J* 2004, **23**:4051-4060.
75. Sharma S, Black DL: **Maps, codes, and sequence elements: can we predict the protein output from an alternatively spliced locus?** *Neuron* 2006, **52**:574-576.
76. Kim TH, Ren B: **Genome-Wide Analysis of Protein-DNA Interactions**. *Annu Rev Genomics Hum Genet* 2006, **7**:81-102.
77. Greil F, Moorman C, van Steensel B: **DamID: mapping of in vivo protein-genome interactions using tethered DNA adenine methyltransferase**. *Methods Enzymol* 2006, **410**:342-359.
78. Impey S, McCorkle SR, Cha-Molstad H, Dwyer JM, Yochum GS, Boss JM, McWeeney S, Dunn JJ, Mandel G, Goodman RH: **Defining the CREB regulon: a genome-wide analysis of transcription factor regulatory regions**. *Cell* 2004, **119**:1041-1054.
79. Lee TI, Jenner RG, Boyer LA, Guenther MG, Levine SS, Kumar RM, Chevalier B, Johnstone SE, Cole MF, Isono K *et al*: **Control of developmental regulators by Polycomb in human embryonic stem cells**. *Cell* 2006, **125**:301-313.
80. Tolhuis B, de Wit E, Muijrers I, Teunissen H, Talhout W, van Steensel B, van Lohuizen M: **Genome-wide profiling of PRC1 and PRC2 Polycomb chromatin binding in Drosophila melanogaster**. *Nat Genet* 2006, **38**:694-699.
81. Wasserman WW, Sandelin A: **Applied bioinformatics for the identification of regulatory elements**. *Nat Rev Genet* 2004, **5**:276-287.
82. Berger MF, Philippakis AA, Qureshi AM, He FS, Estep PW, 3rd, Bulyk ML: **Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities**. *Nat Biotechnol* 2006, **24**:1429-1435.
83. Maerkl SJ, Quake SR: **A systems approach to measuring the binding energy landscapes of transcription factors**. *Science* 2007, **315**:233-237.
84. Lenhard B, Sandelin A, Mendoza L, Engstrom P, Jareborg N, Wasserman WW: **Identification of conserved regulatory elements by comparative genome analysis**. *J Biol* 2003, **2**:13.
85. Berman BP, Pfeiffer BD, Laverty TR, Salzberg SL, Rubin GM, Eisen MB, Celniker SE: **Computational identification of developmental enhancers: conservation and function of transcription factor binding-site clusters in Drosophila melanogaster and Drosophila pseudoobscura**. *Genome Biol* 2004, **5**:R61.
86. Smale ST, Kadonaga JT: **The RNA polymerase II core promoter**. *Annu Rev Biochem* 2003, **72**:449-479.
87. Ohler U, Liao GC, Niemann H, Rubin GM: **Computational analysis of core promoters in the Drosophila genome**. *Genome Biol* 2002, **3**:RESEARCH0087.
88. Ponjavic J, Lenhard B, Kai C, Kawai J, Carninci P, Hayashizaki Y, Sandelin A: **Transcriptional and structural impact of TATA-initiation site spacing in mammalian core promoters**. *Genome Biol* 2006, **7**:R78.
89. Xi H, Yu Y, Fu Y, Foley J, Halees A, Weng Z: **Analysis of overrepresented motifs in human core promoters reveals dual regulatory roles of YY1**. *Genome Res* 2007, **17**:798-806.
90. Bird A: **DNA methylation patterns and epigenetic memory**. *Genes Dev* 2002, **16**:6-21.
91. Tanay A, O'Donnell AH, Damelin M, Bestor TH: **Hyperconserved CpG domains underlie Polycomb-binding sites**. *Proc Natl Acad Sci U S A* 2007, **104**:5521-5526.
92. Muller F, Chang B, Albert S, Fischer N, Tora L, Strahle U: **Intronic enhancers control expression of zebrafish sonic hedgehog in floor plate and notochord**. *Development* 1999, **126**:2103-2116.

93. Nobrega MA, Ovcharenko I, Afzal V, Rubin EM: **Scanning human gene deserts for long-range enhancers**. *Science* 2003, **302**:413.
94. Bellen HJ: **Ten years of enhancer detection: lessons from the fly**. *Plant Cell* 1999, **11**:2271-2281.
95. Ellingsen S, Laplante MA, Konig M, Kikuta H, Furmanek T, Hoivik EA, Becker TS: **Large-scale enhancer detection in the zebrafish genome**. *Development* 2005, **132**:3799-3811.
96. Lettice LA, Heaney SJ, Purdie LA, Li L, de Beer P, Oostra BA, Goode D, Elgar G, Hill RE, de Graaff E: **A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly**. *Hum Mol Genet* 2003, **12**:1725-1735.
97. Valenzuela L, Kamakaka RT: **Chromatin insulators**. *Annu Rev Genet* 2006, **40**:107-138.
98. Li X, Noll M: **Compatibility between enhancers and promoters determines the transcriptional specificity of gooseberry and gooseberry neuro in the Drosophila embryo**. *Embo J* 1994, **13**:400-406.
99. Merli C, Bergstrom DE, Cygan JA, Blackman RK: **Promoter specificity mediates the independent regulation of neighboring genes**. *Genes Dev* 1996, **10**:1260-1270.
100. Ohtsuki S, Levine M, Cai HN: **Different core promoters possess distinct regulatory activities in the Drosophila embryo**. *Genes Dev* 1998, **12**:547-556.
101. Butler JE, Kadonaga JT: **Enhancer-promoter specificity mediated by DPE or TATA core promoter motifs**. *Genes Dev* 2001, **15**:2515-2519.
102. Woolfe A, Goodson M, Goode DK, Snell P, McEwen GK, Vavouri T, Smith SF, North P, Callaway H, Kelly K *et al*: **Highly conserved non-coding sequences are associated with vertebrate development**. *PLoS Biol* 2005, **3**:e7.
103. Shin JT, Priest JR, Ovcharenko I, Ronco A, Moore RK, Burns CG, MacRae CA: **Human-zebrafish non-coding conserved elements act in vivo to regulate transcription**. *Nucleic Acids Res* 2005, **33**:5437-5445.
104. Bailey PJ, Klos JM, Andersson E, Karlen M, Kallstrom M, Ponjavic J, Muhr J, Lenhard B, Sandelin A, Ericson J: **A global genomic transcriptional code associated with CNS-expressed genes**. *Exp Cell Res* 2006, **312**:3108-3119.
105. Prabhakar S, Poulin F, Shoukry M, Afzal V, Rubin EM, Couronne O, Pennacchio LA: **Close sequence comparisons are sufficient to identify human cis-regulatory elements**. *Genome Res* 2006, **16**:855-863.
106. Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, Shoukry M, Minovitsky S, Dubchak I, Holt A, Lewis KD *et al*: **In vivo enhancer analysis of human conserved non-coding sequences**. *Nature* 2006, **444**:499-502.
107. Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D: **Ultraconserved elements in the human genome**. *Science* 2004, **304**:1321-1325.
108. Sandelin A, Bailey P, Bruce S, Engstrom PG, Klos JM, Wasserman WW, Ericson J, Lenhard B: **Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes**. *BMC Genomics* 2004, **5**:99.
109. Kimura-Yoshida C, Kitajima K, Oda-Ishii I, Tian E, Suzuki M, Yamamoto M, Suzuki T, Kobayashi M, Aizawa S, Matsuo I: **Characterization of the pufferfish Otx2 cis-regulators reveals evolutionarily conserved genetic mechanisms for vertebrate head specification**. *Development* 2004, **131**:57-71.
110. de la Calle-Mustienes E, Feijoo CG, Manzanares M, Tena JJ, Rodriguez-Seguel E, Letizia A, Allende ML, Gomez-Skarmeta JL: **A functional survey of the enhancer activity of conserved non-coding sequences from vertebrate Iroquois cluster gene deserts**. *Genome Res* 2005, **15**:1061-1072.
111. Visel A, Minovitsky S, Dubchak I, Pennacchio LA: **VISTA Enhancer Browser--a database of tissue-specific human enhancers**. *Nucleic Acids Res* 2007, **35**:D88-92.
112. Woolfe A, Goode DK, Cooke J, Callaway H, Smith S, Snell P, McEwen GK, Elgar G: **CONDOR: a database resource of developmentally associated conserved non-coding elements**. *BMC Dev Biol* 2007, **7**:100.
113. Sheng Y, Engstrom PG, Lenhard B: **Mammalian MicroRNA Prediction through a Support Vector Machine Model of Sequence and Structure**. *PLoS ONE* 2007, **2**:e946.
114. Glazov EA, Pheasant M, McGraw EA, Bejerano G, Mattick JS: **Ultraconserved elements in insect genomes: a highly conserved intronic sequence implicated in the control of homothorax mRNA splicing**. *Genome Res* 2005, **15**:800-808.
115. Vavouri T, Walter K, Gilks WR, Lehner B, Elgar G: **Parallel evolution of conserved non-coding elements that target a common set of developmental regulatory genes from worms to humans**. *Genome Biol* 2007, **8**:R15.

116. Katzman S, Kern AD, Bejerano G, Fewell G, Fulton L, Wilson RK, Salama SR, Haussler D: **Human genome ultraconserved elements are ultraselected**. *Science* 2007, **317**:915.
117. Gilbert SF: **Developmental biology**, 6th edn. Sunderland, Massachusetts: Sinauer Associates; 2000.
118. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S *et al*: **Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes**. *Genome Res* 2005, **15**:1034-1050.
119. Eichler EE, Sankoff D: **Structural dynamics of eukaryotic chromosome evolution**. *Science* 2003, **301**:793-797.
120. Dehal P, Boore JL: **Two rounds of whole genome duplication in the ancestral vertebrate**. *PLoS Biol* 2005, **3**:e314.
121. Wolfe KH: **Yesterday's polyploids and the mystery of diploidization**. *Nat Rev Genet* 2001, **2**:333-341.
122. Presgraves DC: **Evolutionary genomics: new genes for new jobs**. *Curr Biol* 2005, **15**:R52-53.
123. Fitch WM: **Homology a personal view on some of the problems**. *Trends Genet* 2000, **16**:227-231.
124. Kent WJ, Baertsch R, Hinrichs A, Miller W, Haussler D: **Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes**. *Proc Natl Acad Sci U S A* 2003, **100**:11484-11489.
125. Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, Haussler D, Miller W: **Human-mouse alignments with BLASTZ**. *Genome Res* 2003, **13**:103-107.
126. Brudno M, Poliakov A, Minovitsky S, Ratnere I, Dubchak I: **Multiple whole genome alignments and novel biomedical applications at the VISTA portal**. *Nucleic Acids Res* 2007, **35**:W669-674.
127. Maltais LJ, Blake JA, Eppig JT, Davisson MT: **Rules and guidelines for mouse gene nomenclature: a condensed version. International Committee on Standardized Genetic Nomenclature for Mice**. *Genomics* 1997, **45**:471-476.
128. Zdobnov EM, von Mering C, Letunic I, Torrents D, Suyama M, Copley RR, Christophides GK, Thomasova D, Holt RA, Subramanian GM *et al*: **Comparative genome and proteome analysis of Anopheles gambiae and Drosophila melanogaster**. *Science* 2002, **298**:149-159.
129. Pevzner P, Tesler G: **Genome rearrangements in mammalian evolution: lessons from human and mouse genomes**. *Genome Res* 2003, **13**:37-45.
130. Nadeau JH, Taylor BA: **Lengths of chromosomal segments conserved since divergence of man and mouse**. *Proc Natl Acad Sci U S A* 1984, **81**:814-818.
131. Pevzner P, Tesler G: **Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution**. *Proc Natl Acad Sci U S A* 2003, **100**:7672-7677.
132. Sankoff D, Trinh P: **Chromosomal breakpoint reuse in genome sequence rearrangement**. *J Comput Biol* 2005, **12**:812-821.
133. Peng Q, Pevzner PA, Tesler G: **The fragile breakage versus random breakage models of chromosome evolution**. *PLoS Comput Biol* 2006, **2**:e14.
134. Sankoff D: **The signal in the genomes**. *PLoS Comput Biol* 2006, **2**:e35.
135. Murphy WJ, Larkin DM, Everts-van der Wind A, Bourque G, Tesler G, Auvil L, Beever JE, Chowdhary BP, Galibert F, Gatzke L *et al*: **Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps**. *Science* 2005, **309**:613-617.
136. Zdobnov EM, Bork P: **Quantification of insect genome divergence**. *Trends Genet* 2007, **23**:16-20.
137. Hurst LD, Pal C, Lercher MJ: **The evolutionary dynamics of eukaryotic gene order**. *Nat Rev Genet* 2004, **5**:299-310.
138. Niimura Y, Nei M: **Evolution of olfactory receptor genes in the human genome**. *Proc Natl Acad Sci U S A* 2003, **100**:12235-12240.
139. Huntley S, Baggott DM, Hamilton AT, Tran-Gyamfi M, Yang S, Kim J, Gordon L, Branscomb E, Stubbs L: **A comprehensive catalog of human KRAB-associated zinc finger genes: insights into the evolutionary history of a large family of transcriptional repressors**. *Genome Res* 2006, **16**:669-677.
140. Sproul D, Gilbert N, Bickmore WA: **The role of chromatin structure in regulating the expression of clustered genes**. *Nat Rev Genet* 2005, **6**:775-781.
141. Kmita M, Duboule D: **Organizing axes in time and space; 25 years of colinear tinkering**. *Science* 2003, **301**:331-333.
142. Lercher MJ, Blumenthal T, Hurst LD: **Coexpression of neighboring genes in Caenorhabditis elegans is mostly due to operons and duplicate genes**. *Genome Res* 2003, **13**:238-243.
143. Spellman PT, Rubin GM: **Evidence for large domains of similarly expressed genes in the Drosophila genome**. *J Biol* 2002, **1**:5.
144. Lercher MJ, Urrutia AO, Hurst LD: **Clustering of housekeeping genes provides a unified model of gene order in the human genome**. *Nat Genet* 2002, **31**:180-183.

145. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G *et al*: **A gene atlas of the mouse and human protein-encoding transcriptomes**. *Proc Natl Acad Sci U S A* 2004, **101**:6062-6067.
146. Singer GA, Lloyd AT, Huminiecki LB, Wolfe KH: **Clusters of co-expressed genes in mammalian genomes are conserved by natural selection**. *Mol Biol Evol* 2005, **22**:767-775.
147. Ben-Shahar Y, Nannapaneni K, Casavant TL, Scheetz TE, Welsh MJ: **Eukaryotic operon-like transcription of functionally related genes in Drosophila**. *Proc Natl Acad Sci U S A* 2007, **104**:222-227.
148. Lee JM, Sonnhammer EL: **Genomic gene clustering analysis of pathways in eukaryotes**. *Genome Res* 2003, **13**:875-882.
149. Petkov PM, Graber JH, Churchill GA, DiPetrillo K, King BL, Paigen K: **Evidence of a large-scale functional organization of mammalian chromosomes**. *PLoS Genet* 2005, **1**:e33.
150. Petkov PM, Graber JH, Churchill GA, DiPetrillo K, King BL, Paigen K: **Evidence of a large-scale functional organization of Mammalian chromosomes**. *PLoS Biol* 2007, **5**:e127; author reply e128.
151. Adachi N, Lieber MR: **Bidirectional gene organization: a common architectural feature of the human genome**. *Cell* 2002, **109**:807-809.
152. Takai D, Jones PA: **Origins of bidirectional promoters: computational analyses of intergenic distance in the human genome**. *Mol Biol Evol* 2004, **21**:463-467.
153. Trinklein ND, Aldred SF, Hartman SJ, Schroeder DI, Otillar RP, Myers RM: **An abundance of bidirectional promoters in the human genome**. *Genome Res* 2004, **14**:62-66.
154. Lin JM, Collins PJ, Trinklein ND, Fu Y, Xi H, Myers RM, Weng Z: **Transcription factor binding and modified histones in human bidirectional promoters**. *Genome Res* 2007, **17**:818-827.
155. Koyanagi KO, Hagiwara M, Itoh T, Gojobori T, Imanishi T: **Comparative genomics of bidirectional gene pairs and its implications for the evolution of a transcriptional regulation system**. *Gene* 2005, **353**:169-176.
156. Li YY, Yu H, Guo ZM, Guo TQ, Tu K, Li YX: **Systematic analysis of head-to-head gene organization: evolutionary conservation and potential biological relevance**. *PLoS Comput Biol* 2006, **2**:e74.
157. Fahey ME, Moore TF, Higgins DG: **Overlapping antisense transcription in the human genome**. *Comp Funct Genomics* 2002, **3**:244-253.
158. Lehner B, Williams G, Campbell RD, Sanderson CM: **Antisense transcripts in the human genome**. *Trends Genet* 2002, **18**:63-65.
159. Zhang Y, Liu XS, Liu QR, Wei L: **Genome-wide in silico identification and analysis of cis natural antisense transcripts (cis-NATs) in ten species**. *Nucleic Acids Res* 2006, **34**:3465-3475.
160. Sun M, Hurst LD, Carmichael GG, Chen J: **Evidence for variation in abundance of antisense transcripts between multicellular animals but no relationship between antisense transcriptionand organismic complexity**. *Genome Res* 2006, **16**:922-933.
161. Vanhee-Brossollet C, Vaquero C: **Do natural antisense transcripts make sense in eukaryotes?** *Gene* 1998, **211**:1-9.
162. Werner A, Berdal A: **Natural antisense transcripts: sound or silence?** *Physiol Genomics* 2005, **23**:125-131.
163. Anguera MC, Sun BK, Xu N, Lee JT: **X-chromosome kiss and tell: how the Xs go their separate ways**. *Cold Spring Harb Symp Quant Biol* 2006, **71**:429-437.
164. Hastings ML, Ingle HA, Lazar MA, Munroe SH: **Post-transcriptional regulation of thyroid hormone receptor expression by cis-acting sequences and a naturally occurring antisense RNA**. *J Biol Chem* 2000, **275**:11507-11513.
165. Neeman Y, Dahary D, Levanon EY, Sorek R, Eisenberg E: **Is there any sense in antisense editing?** *Trends Genet* 2005, **21**:544-547.
166. Chen J, Sun M, Hurst LD, Carmichael GG, Rowley JD: **Genome-wide analysis of coordinate expression and evolution of human cis-encoded sense-antisense transcripts**. *Trends Genet* 2005, **21**:326-329.
167. Lercher MJ, Urrutia AO, Pavlicek A, Hurst LD: **A unification of mosaic structures in the human genome**. *Hum Mol Genet* 2003, **12**:2411-2415.
168. Gilbert N, Boyle S, Fiegler H, Woodfine K, Carter NP, Bickmore WA: **Chromatin architecture of the human genome: gene-rich domains are enriched in open chromatin fibers**. *Cell* 2004, **118**:555-566.
169. Gierman HJ, Indemans MH, Koster J, Goetze S, Seppen J, Geerts D, van Driel R, Versteeg R: **Domain-wide regulation of gene expression in the human genome**. *Genome Res* 2007, **17**:1286-1295.

170. Ovcharenko I, Loots GG, Nobrega MA, Hardison RC, Miller W, Stubbs L: **Evolution and functional classification of vertebrate gene deserts**. *Genome Res* 2005, **15**:137-145.
171. Ahituv N, Prabhakar S, Poulin F, Rubin EM, Couronne O: **Mapping cis-regulatory domains in the human genome using multi-species conservation of synteny**. *Hum Mol Genet* 2005, **14**:3057-3063.
172. Kleinjan DA, van Heyningen V: **Long-range control of gene expression: emerging mechanisms and disruption in disease**. *Am J Hum Genet* 2005, **76**:8-32.
173. Ringrose L, Paro R: **Epigenetic regulation of cellular memory by the Polycomb and Trithorax group proteins**. *Annu Rev Genet* 2004, **38**:413-443.
174. Mattick JS: **RNA regulation: a new genetics?** *Nat Rev Genet* 2004, **5**:316-323.
175. Pang KC, Stephen S, Dinger ME, Engstrom PG, Lenhard B, Mattick JS: **RNAdb 2.0--an expanded database of mammalian non-coding RNAs**. *Nucleic Acids Res* 2007, **35**:D178-182.
176. Maeda N, Kasukawa T, Oyama R, Gough J, Frith M, Engstrom PG, Lenhard B, Aturaliya RN, Batalov S, Beisel KW *et al*: **Transcript annotation in FANTOM3: mouse gene catalog based on physical cDNAs**. *PLoS Genet* 2006, **2**:e62.
177. Sun M, Hurst LD, Carmichael GG, Chen J: **Evidence for a preferential targeting of 3'-UTRs by cis-encoded natural antisense transcripts**. *Nucleic Acids Res* 2005, **33**:5533-5543.
178. Veeramachaneni V, Makalowski W, Galdzicki M, Sood R, Makalowska I: **Mammalian overlapping genes: the comparative perspective**. *Genome Res* 2004, **14**:280-286.
179. Green P, Ewing B, Miller W, Thomas PJ, Green ED: **Transcription-associated mutational asymmetry in mammalian evolution**. *Nat Genet* 2003, **33**:514-517.
180. Crosby MA, Goodman JL, Strelets VB, Zhang P, Gelbart WM: **FlyBase: genomes by the dozen**. *Nucleic Acids Res* 2007, **35**:D486-491.
181. Ohler U: **Identification of core promoter modules in Drosophila and their application in accurate transcription start site prediction**. *Nucleic Acids Res* 2006, **34**:5943-5950.
182. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT *et al*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium**. *Nat Genet* 2000, **25**:25-29.
183. Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A *et al*: **The generic genome browser: a building block for a model organism system database**. *Genome Res* 2002, **12**:1599-1610.
184. Kapranov P, Willingham AT, Gingeras TR: **Genome-wide transcription and the implications for genomic organization**. *Nat Rev Genet* 2007, **8**:413-423.
185. Dostie J, Richmond TA, Arnaout RA, Selzer RR, Lee WL, Honan TA, Rubio ED, Krumm A, Lamb J, Nusbaum C *et al*: **Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements**. *Genome Res* 2006, **16**:1299-1309.
186. Fuchs F, Boutros M: **Cellular phenotyping by RNAi**. *Brief Funct Genomic Proteomic* 2006, **5**:52-56.
187. Ureta-Vidal A, Ettwiller L, Birney E: **Comparative genomics: genome-wide analysis in metazoan eukaryotes**. *Nat Rev Genet* 2003, **4**:251-262.
188. Nygaard V, Hovig E: **Options available for profiling small samples: a review of sample amplification technology when combined with microarray profiling**. *Nucleic Acids Res* 2006, **34**:996-1014.
189. Smale ST: **Core promoters: active contributors to combinatorial gene regulation**. *Genes Dev* 2001, **15**:2503-2508.
190. Putnam NH, Srivastava M, Hellsten U, Dirks B, Chapman J, Salamov A, Terry A, Shapiro H, Lindquist E, Kapitonov VV *et al*: **Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization**. *Science* 2007, **317**:86-94.
191. Duncan IW: **Transvection effects in Drosophila**. *Annu Rev Genet* 2002, **36**:521-556.
192. **GSC: Petromyzon marinus** [http://genome.wustl.edu/genome.cgi?GENOME=Petromyzon%20marinus]
193. Ahituv N, Zhu Y, Visel A, Holt A, Afzal V, Pennacchio LA, Rubin EM: **Deletion of ultraconserved elements yields viable mice**. *PLoS Biol* 2007, **5**:e234.
194. Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, Beazley C, Ingle CE, Dunning M, Flicek P, Koller D *et al*: **Population genomics of human gene expression**. *Nat Genet* 2007, **39**:1217-1224.

November 1, 2007

Gene Complexes and Regulatory Domains in Metazoan Genomes
Pär Engström

Corrections

Reference errors

p. 6, last paragraph, line 4. The statement about conservation of upstream regions should refer to Taylor et al. 2006, *PLoS Genet* **2**:e30 instead of reference 46.

p. 25, paragraph 1, line 6. Instead of reference 170 (Ovcharenko et al. 2005), the sentence should refer to Ovcharenko et al. 2004, *Nucleic Acids Res* **32**:W280

p. 27, paragraph 2, line 4. Reference 189 (Smale 2001) has been included by mistake. The sentence referring to this paper should instead use reference 86 (Smale and Kadonaga 2003).

Typographical errors

p. 4, last paragraph, line 3. “established alternative” should be “established that alternative”.

p. 14, paragraph 1, line 11. “are either are nested” should be “are either nested”

p. 19, paragraph 1, line 4. “TUs with that” should be “TUs that”

p. 21, paragraph 2, line 2. “*cis*-antisense as” should be “*cis*-antisense pairs as”

p. 26, paragraph 1, line 6. “Drosophila” should be in italics