Noncoding Elements: Evolution and Epigenetic Regulation

Thesis by

Loqmane Seridi

Submitted in Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

King Abdullah University of Science and Technology

Computer, Electrical and Mathematical Science and Engineering

Thuwal, Makkah Province, Kingdom of Saudi Arabia

March, 2016

This dissertation of Loqmane Seridi is approved by the examination committee.

Dr. Timothy Ravasi
Associate Professor, Bioengineering

Biological and Environmental Science and Engineering

King Abdullah University of Science and Technology

Committee Chair (Dissertation Supervisor)

Dr. Valerio Orlando
Professor, Bioscience

Biological and Environmental Science and Engineering

KAUST Environmental Epigenetics Program (KEEP)

King Abdullah University of Science and Technology

Committee Member

Dr. Xin Gao
Assistant Professor, Computer Science

Computer, Electrical and Mathematical Science and Engineering

King Abdullah University of Science and Technology

Committee Member

Dr. Piero Carninci
Deputy Director of the RIKEN Center for Life Science Technologies

Director of the Division of Genomic Technologies

RIKEN

Committee Member

King Abdullah University of Science and Technology

2016

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

CAGE	Cap Analysis of Gene Expression
CNE	conserved non-coding element
CTC	consensus tag cluster
DMD	Duchenne muscular dystrophy
DPE	downstream promoter element
DRE	DNA replication element binding factor
eRNA	enhancer RNA
GO	Gene Ontology
GRB	gene regulatory block
H3K27ac	H3 lysine 27 acetylation
H3K4me1	H3 lysine 4 mono-methylation
H3K4me3	H3 lysine 4 tri-methylation
H3K9ac	H3 lysine 9 acetylation
HCNE	highly conserved noncoding element
Inr	initiator element
Inr/DPE	Inr followed by a downstream promoter element
LAD	lamina-associated domain
LINE	long interspersed repeat
LTR	long terminal repeat
MDS	multidimensional scaling
Motif1/6	Motif1 followed by Motif6
mRNA	messenger RNA
PIC	pre-initiation complex
rRNA	ribosomal RNA
S/MAR	scaffold/matrix attachment region

SNP	single nucleotide polymorphism
TATAbox/Inr	TATA box followed by an initiator element
TF	transcription factor
TFBS	transcription factor binding site
tRNA	transfer RNA
TSS	transcriptional start site
UCE	ultraconserved element
UCR	ultraconserved region

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Noncoding elements: evolution and epigenetic regulation

Loqmane Seridi

When the human genome project was completed, it revealed a surprising result. 98% of the genome did not code for protein of which more than 50% are repeats— later known as "Junk DNA". However, comparative genomics unveiled that many noncoding elements are evolutionarily constrained; thus luckily to have a role in genome stability and regulation. Though, their exact functions remained largely unknown.

Several large international consortia such as the Functional Annotation of Mammalian Genomes (FANTOM) and the Encyclopedia of DNA Elements (ENCODE) were set to understand the structure and the regulation of the genome. Specifically, these endeavors aim to measure and reveal the transcribed components and functional elements of the genome. One of the most the striking findings of these efforts is that most of the genome is transcribed, including non-conserved noncoding elements and repeat elements.

Specifically, we investigated the evolution and epigenetic properties of noncoding elements.

1. We compared genomes of evolutionarily distant species and showed the ubiquity of constrained noncoding elements in metazoa.

2. By integrating multi-omic data (such as transcriptome, nucleosome profiling, histone modifications), I conducted a comprehensive analysis of epigenetic properties (chromatin states) of conserved noncoding elements in insects. We showed that those elements have distinct and protective sequence features, undergo dynamic epigenetic regulation, and appear to be associated with the structural components of the chromatin, replication origins, and nuclear matrix.

3. I focused on the relationship between enhancers and repetitive elements. Using Cap Analysis of Gene Expression (CAGE) and RNASeq, I compiled a full catalog of active enhancers (a

class of noncoding elements) during myogenesis of human primary cells of healthy donors and donors affected by Duchenne muscular dystrophy (DMD). Comparing the two time-courses, a significant change in the epigenetic landscape in DMD was observed that lead to global dysregulation of enhancers and associated repetitive elements.

ACKNOWLEDGEMENTS

I thank my advisors Dr. Timothy Ravasi and Dr. Valerio Orlando for their guidance, help, and support. I would also like to thank Dr. Xin Gao and Dr. Piero Carninci for their advice and for being members of my thesis committee.

I thank my co-authors Yanal Ghosheh and Taewoo Ryu for their collaboration, help and fruitful discussions.

I would like to thank my parents Ahcen and Hassina, my wife Fatma and my three princesses: Rihana, Tala, and Rittal.

Last but not least, I would like to thank King Abdullah University of Science and Technology (KAUST), my friends and all the KAUST community.

Chapter 1

Introduction

The beginning of 21^{st} century witnessed the completion of the human genome project. After more than a decade of work and three billion dollars, the project revealed an exciting but unexpected result: most of the DNA is "noncoding". Unlike previous estimates of $\sim 100,000$ genes contained in the human genome, the project reported only $\sim 24,500$ genes— accounting for $\leq 5\%$ of the genome (Lander et al., 2001). Besides, repeat elements, or "junk DNA" composed $\geq 50\%$ of the human genome (Lander et al., 2001).

Rapid advances in sequencing technologies reduced the cost (time and money) and improved the quality of genome sequencing. Consequently, genomes of different species became available, and comparing the genomes became possible. Comparative genomics revealed that many noncoding elements are under purifying selection— and thus, are likely to be functional. Though, their functions remained largely unknown.

To understand genome structure and regulation, many international consortiums were created. For instance, the Functional Annotation of Mammalian Genomes (FANTOM) and the Encyclopedia of DNA Elements (ENCODE). The FANTOM datasets showed that $\geq 69\%$ of the genome is transcribed— including repeats (Carninci et al., 2005; Faulkner et al., 2009). Few years later, the ENCODE project assigned biochemical activities to $\sim 80\%$ of the non-repeat genome (Dunham et al., 2012). Although, it is still unclear whether the transcriptional and biochemical activities are indicators of function or just noise, biochemical and evolutionary approaches are complementary and their integration is essential to identify functional elements in the genome (Kellis et al., 2014).

In this thesis, we integrated evolutionary and biochemical datasets to investigate putative regulatory roles of noncoding elements. It particularly put emphasis on the epigenetic aspect of this regulation.

This thesis's scientific contributions are:

- Using comparative genomics, I showed that ultraconserved element (UCE) elements are ubiquitously present in metazoa, and they are evolutionarily constrained in a lineage-specific manner (Chapter 4).

- In this work, I conducted a comprehensive analysis of epigenetic properties of highly conserved noncoding element (HCNE) during fruitfly development by integrating multi-omics and conservation datasets (histone modifications, nucleosome profiling and transcriptome). We showed that HCNE in insects have distinct sequence features and reside mostly in heterochromatin but exhibit characteristics of active enhancers at specific stages of development (Chapter 5).

- Using CAGE and RNA-Seq, I compiled a catalog of active enhancers during myogenesis of human primary cells of healthy donors and donors affected by DMD. We showed that those enhancers are active in time specific manner. Although, they are active in both phenotypes, they are dysregulated in DMD donors. A number of enhancers overlapped repeat elements and were enriched for long terminal repeats (LTRs), suggesting an enhancer activity of some LTRs during differentiation (Chapter 6).

In addition to self-contained chapters (Chapter 4, Chapter 5 and Chapter 6) that describe the above mentioned contributions, the thesis includes: a short background chapter (Chapter 2) that introduces concepts related to aforementioned chapters; a conclusion chapter (Chapter 6); and appendices that include supplemental data related to Chapters 4 and Chapter 6.

Chapter 2

Contribution to Bioinformatics and Computer Science

Bioinformatics is a multidisciplinary field that utilizes mathematics, computer science, statistics and other disciplines to answer interesting and complex biological questions. "Bioinformatics" can be easily confused with other related fields such as "computational biology" due to the considerable overlap between these fields and their integrative nature. Here, we adopt the National Institute of Health (NIH) definition of bioinformatics; that is the "Research, development, or *application of computational tools* and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, archive, analyze, or visualize such data".

In this chapter, I will highlight my contribution to the field of bioinformatics and computer science.

2.1 Thesis contributions

I contributed to the disciplines of bioinformatics and computer science by designing and developing frameworks that combine algorithms to integrate, analyze and visualize large heterogeneous multi-omics (genomic, transcriptomic and, epigenomics) datasets and answer specific biological questions.

2.1.1 Contribution to bioinformatics

- Using comparative genomics approaches, we reported for the first time the ubiquitously of UCEs in metazoa. I showed that UCEs evolved in a lineage-specific manner (Chapter 4).
- Using integrative data approaches, I characterized epigenetic properties of HCNE during the

development of fruitfly. I reported that HCNE in insects have distinct sequence features; they reside repressive chromatin state; they may act as enhancers at specific stages of development (Chapter 5).

- Using CAGE and RNA-Seq, I annotated active enhancers in myogenesis of human primary cells of healthy donors and donors affected by DMD. I found that those enhancers are active in time specific manner in this process; their activity is dysregulated in DMD donors. I showed that enhancers overlapped repeats were enriched for LTRs, implying a possible role of LTRsassociated enhancers during differentiation (Chapter 6).

2.1.2 Contribution to computer science

- I developed an unsupervised learning-based framework to examine the evolution of ultraconserved elements. This framework used, for the first time, the Markov Cluster (MCL) algorithm plus Minimum Curvilinear Embedding (MCE) analysis in the context of UCEs (Chapter 4).

- I designed a framework to study time courses of gene expression from healthy and disease samples. This framework uses mutual information variation distance, minimum spanning tree algorithm, and multidimensional scaling. Also, it uses Non-Negative Matrix Factorization (NMF) to estimate the number of gene expression profiles and Kmeans clustering (Chapter 6).

2.2 Overview of the frameworks

2.2.1 Evolution of UCE framework (EUCEF)

Ultraconserved elements were shown to exist between pairs of related species such as mouse and human. In this project, we investigated the existence and the evolution of UCEs. To answer this question, I developed an unsupervised learning framework that begins with the identification of ultraconserved elements (UCEs) shared between pairs of distant species, and then it merges overlapping UCEs are into ultraconserved regions (UCRs). To unveil the evolutionary history of UCEs, the framework uses unsupervised learning approach by constructing a sequence-based similarity network by comparing UCEs using BLAST and cluster UCRs by Markov cluster algorithm (MCL)(Enright et al., 2002). To visualize UCR clusters the sequences are described in all possible pentamer feature spaces. The feature space is then projected on two dimension space using Minimum Curvilinear Embedding (MCE)(Cannistraci et al., 2013). The framework assigns possible functional association

to UCEs by investigating enrichment of motifs (with TFBS), and functional annotation of flanking genes. More details can be found in methods section of Chapter 4.

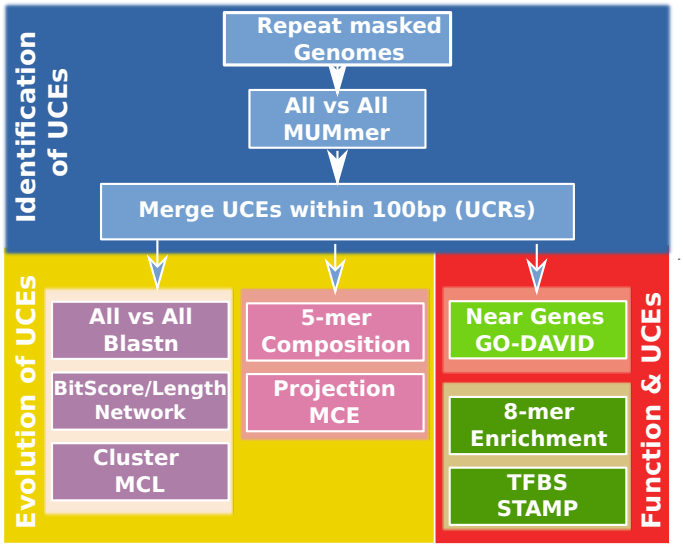

Figure 2.1: Flowchart of the EUCEF.

2.2.2 Mutual information over minimum spanning tree framework for comparing time courses (MMF)

To study the role of enhancers and repeat elements during muscle differentiation, I designed a framework that exploits mutual information based distance, minimum mapping tree, and multidimensional scaling to compare two-time course transcriptomes from samples of health and DMD affected patients. This framework was able to detect important signals that were not possible to detect when using standard distance such as Euclidean distance, correlation and so forth. In addition, I designed to a pipeline for detection and characterization of active enhancers during differentiation. Details of the framework are described in the methods section of Chapter 6.

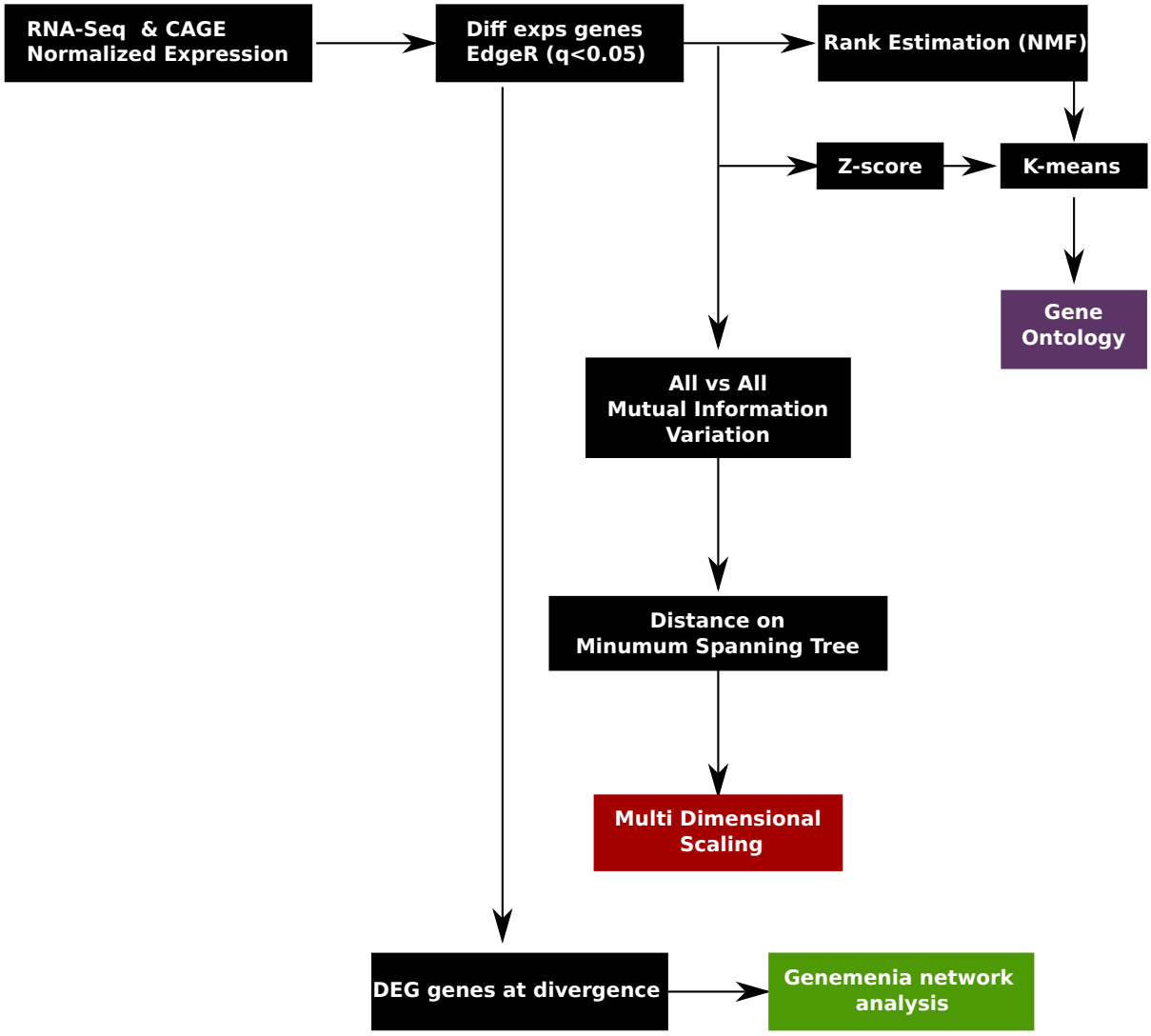

Figure 2.2: Flowchart of the MMF.

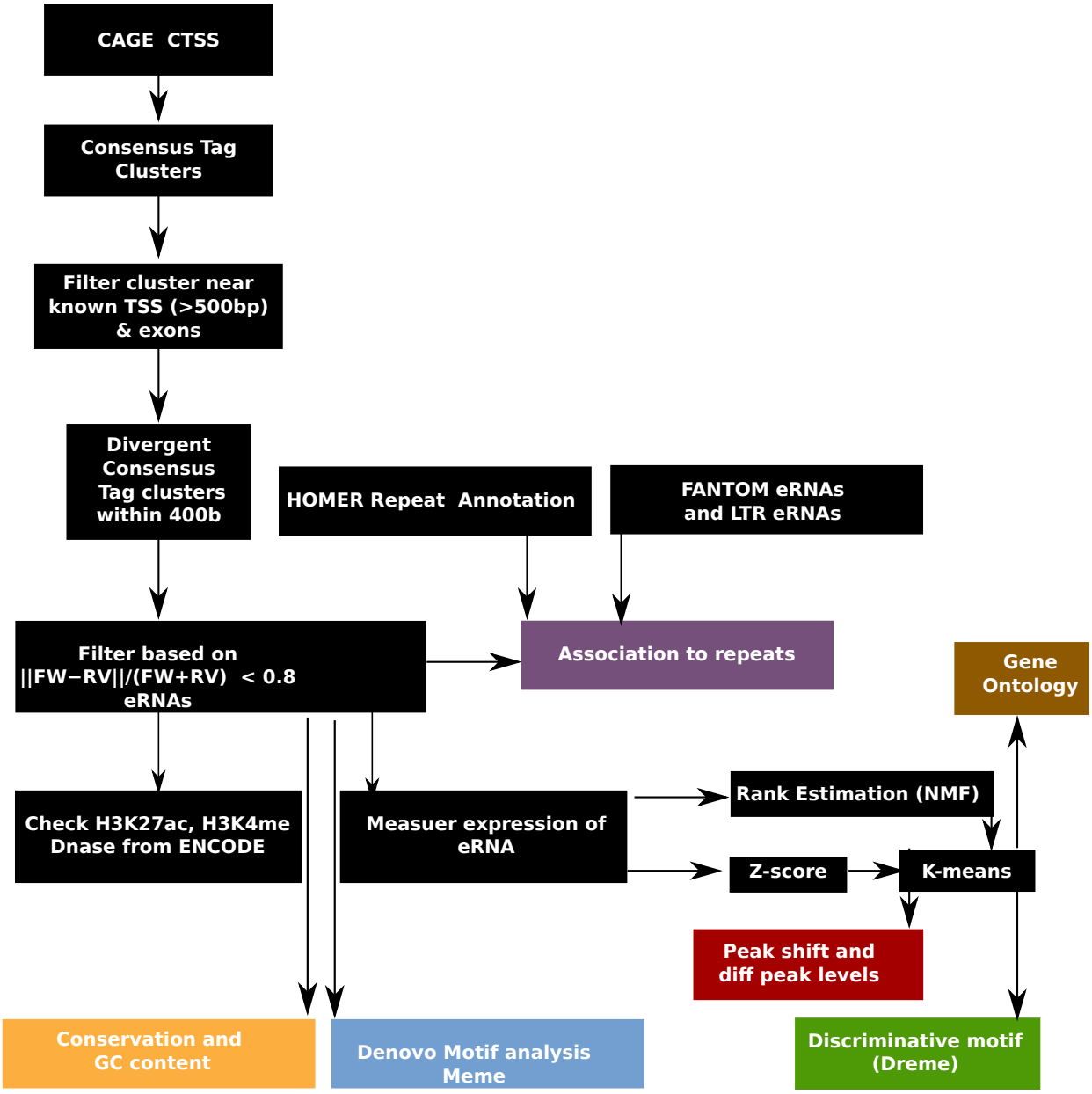

Figure 2.3: Characterization of active enhancers.

Chapter 3

Background

This chapter contains excerpts from the following publications:

- Taewoo Ryu, Loqmane Seridi and Timothy Ravasi. The evolution of ultraconserved elements with different phylogenetic origins. *BMC Evolutionary Biology* 2012, 12:236. http://www.biomedcentral.com/1471-2148/12/236. ©Ryu et al 2012.
- Loqmane Seridi, Taewoo Ryu and Timothy Ravasi. Dynamic epigenetic control of highly conserved noncoding elements. *PLOSONE*. DOI: 10.1371/journal.pone.0109326 ©Seridi et al 2014.

3.1 Regulation of gene expression

Cell identity is beyond its DNA code. Cells in multi-cellular organisms share identical DNA but are diverse and respond to stress differently (Holle and Engler, 2010). Cells type identity is defined by distinct gene expression landscape that is governed by complex regulatory networks as a result of the action of various classes of regulatory elements such as promoters, enhancer, transcription factors ...etc.

Here, we shed some light on gene expression and different elements involved in its regulation.

3.1.1 Gene and gene transcription

Gene is a broadly used term that refers to a portion DNA sequence in the genome and associated putative phenotype; yet as the complexity of genome output is revealed its definition is evolving. Indeed, since its first use in 1903 by W. Johannsen, the term has been repeatedly updated to comply with challenges imposed by new findings and significant technological advances (Gerstein et al., 2007). For consistency, we find it necessary to use one definition of a gene in this thesis. Skipping the long history of definitions (as it was well listed and discussed by Gerstein et al., 2007),

here, we adopt the definition of a gene as *"a union of genomic sequences encoding a coherent set of potentially overlapping functional products"*(Gerstein et al., 2007).

From a gene, RNA is transcribed by an enzyme called RNA-polymerase. Eukaryotic cells have three such enzymes: PolI, PolII, and PolIII that transcribe distinct RNA types. PolI transcribes ribosomal RNAs (rRNAs) and transfer RNA (tRNA). PolII transcribes messenger RNA (mRNA). PolIII transcribes one rRNA.

Transcription of RNA from DNA is achieved in three steps: initiation, strand elongation, and termination. Briefly, transcription initiation begins by recruiting and stabilizing RNA-polymerase at transcriptional start site (TSS) by pre-initiation complex (PIC). Then, the polymerase enzyme moves and reads along the DNA (from 5' to 3') while appending bases to the 3' of synthesized RNA. Finally, transcription terminates by different ways depending on the type of RNA-polymerase: PolI is stopped by a terminator factor; PolII passes through a sequence signature, the synthesized RNA is then cleaved at the signature by a cleavage complex; PolIII stops at a particular termination sequence signature (Clancy, 2008).

3.1.2 Promoters

Regions surrounding the transcripts' TSSs are called promoters. Promoters usually span 100bps to 1000bps, mostly upstream of TSS. They harbor sequence motifs that are recognizable by proteins known as transcription factors (TFs); the sequence motifs are known as transcription factor binding sites (TFBSs). In eukaryotes, regions containing TFBSs are located tens to hundreds of bases from TSS are know as promoter-proximal elements, whereas, regions that overlap TSS are known as core promoters (Figure 3.1).

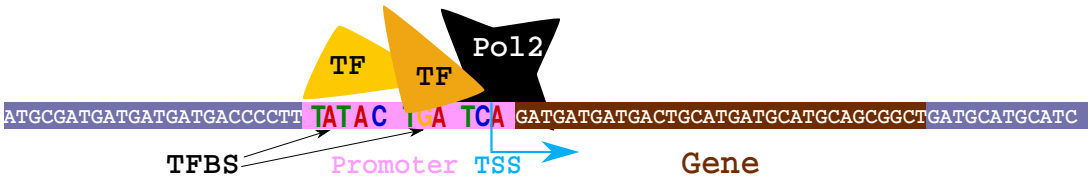

Figure 3.1: A schematic figure shows TFs binding to a promoter.

According to Carnincis paper, core promoters are classified into three major classes. *Type I promoters* associate with tissue-specific genes; they are narrow (TSS spans a short region); they are enriched for TATA box followed by an initiator element (TATAbox/Inr) motifs. *Type II promoters*

associate with genes ubiquitously expressed across tissues; they are broad (multiple TSSs span broad region); they are enriched for CpG islands in vertebrates, and enriched for Motif1 followed by Motif6 (Motif1/6) and DNA replication element binding factor (DRE) motifs in invertebrates (*Drosophila*). *Type III promoters* associate with genes expressed in particular developmental stages; they are enriched for large CpG islands (which may overlap gene body) in vertebrates, and in invertebrates (*Drosophila*) are enriched for downstream promoter element (DPE) in (Lenhard et al., 2012).

3.1.3 Enhancers

TFBSs are not unique to promoters; they are scattered throughout the genome. Some distal regions which are found kilobases from TSS and harbor TFBSs are known as enhancers. Over 400,000 enhancers have been identified in the human genome (Lam et al., 2014). Enhancers' regulation of gene expression is unconstrained by the distance nor by the orientation (they can be located upstream or downstream of their target genes)(Maston et al., 2006). When enhancers are active, they become near to their targets' promoters through DNA looping (Figure 3.2). The enhancer activity can be suppressed by blocking the promoter-enhancer interaction through regulation of 3D chromatin structure by CTCF binding proteins at loci called insulators.

Figure 3.2: A schematic figure shows enhancer-promoter interaction through DNA looping.

3.1.4 Chromatin

In eukaryotic cells, DNA wraps around octamers of conserved histone proteins (H2A, H2B, H3 and H4) to form nucleosomes. Nucleosomes pack DNA to form first level of compaction of the chromatin structure (Figure 3.3). Chromatin is a dynamic structure— it switches between many states. Chromatin states are results of combinations of biochemical changes that occur onto histone proteins by the enzymatic action of chromatin associated protein complexes. Changes include the exchange of one histone with its variances (like exchange of H2A by H2A.Z) (Zovkic et al., 2014)

and hundreds of modification of their N-terminal tails (mono- di- tri-methylation, acetylation and phosphorylation, to name a few) (Baker, 2011; Bernstein et al., 2006; Chen et al., 2012).

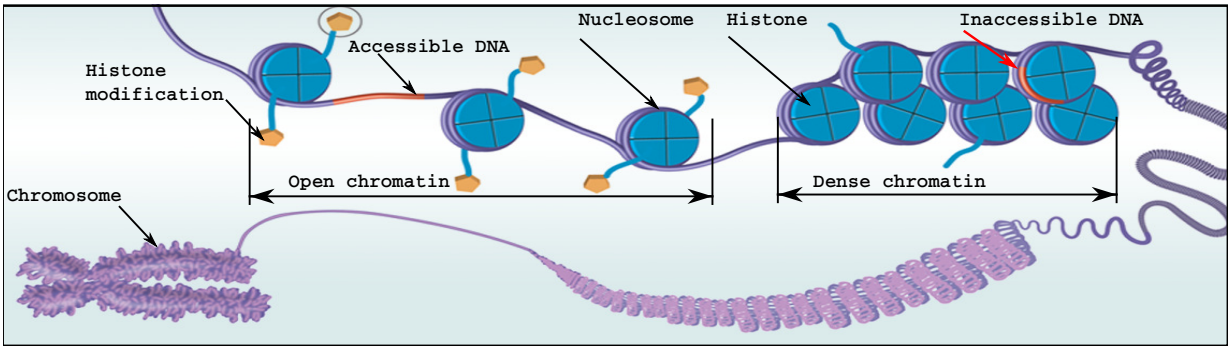

Figure 3.3: DNA wraps histones to create chromatin structure. Art reproduced from *Environmental Health Perspectives*.

For TFs to bind promoters or enhancers of their target genes, the DNA has to be accessible. DNA accessibility depends on the state of the chromatin. Though DNA accessibility can be achieved by many combinations of histone modifications, active promoters and active enhancers exhibit distinct histone modification signatures: promoters are enriched for H3 lysine 4 tri-methylation (H3K4me3) and H3 lysine 9 acetylation (H3K9ac), whereas enhancers are enriched for H3 lysine 4 mono-methylation (H3K4me1) and H3 lysine 27 acetylation (H3K27ac).

3.1.5 Enhancer RNAs

Many enhancers are transcribed as eRNAs. Most eRNAs are short capped RNAs ($\sim$ 300-400 bp), bi-directionally transcribed, unspliced, non-polyadenylated, degrade at fast rates and expressed at low levels (Lam et al., 2014; Andersson et al., 2014). However, few eRNAs are relatively long, polyadenylated and unidirectional (Lam et al., 2014). Although eRNAs expression is a sign of an enhancers' active state (eRNAs expression correlates with active enhancer marks, e.g., H3K27ac (Lam et al., 2014)), it is not clear whether eRNAs have a function (such as mediation of transcription or facilitation of chromatin looping for promoter-enhancer interaction) or they are just a byproduct of the enhancer activity (transcriptional noise) (Lam et al., 2014). A conserved eRNA upstream of the MyoD gene, core enhancer (CE), was reported to be crucial for loading the transcriptional machinery in promoters of its target gene in *trans* (MyoG)— in support of the functional role of eRNAs (Mousavi et al., 2013). However, it should be noted that CE belong to a small class of eRNAs which are long, polyadenylated and unidirectional; therefore, it is still unclear whether other classes of eRNAs are also functional.

3.2 Conserved non-coding elements

Advances in sequencing technologies and bioinformatic tools enabled genomic comparisons across species. Comparing genomes of distant species unveiled hundreds of noncoding sequences that evolve at slow rates— even slower than that of protein coding sequences— suggesting they have essential functions (Bejerano et al., 2004; Ryu et al., 2012). Many studies identified conserved sequences using different similarity criteria and comparing different sets of species. As a result, conserved noncoding elements appear under many names in the literature (Table 3.2. In this chapter, we will refer to them as conserved non-coding elements (CNEs).

Name	**Acronym**	**Criteria**	**Species against human**	**Study**
Conserved noncoding Sequences	CNS	length $\geq$ 100bp; identity $\geq$ 70%	mouse	Nobrega et al., 2003
Ultra-Conserved Elements	UCE	length$\geq$ 200bp; identity 100%	mouse	Bejerano et al., 2004
Highly Conserved noncoding Sequences	HCNS	length$\geq$100bp; identity $>$ 74%	pufferfish	Woolfe et al., 2005
Highly Conserved noncoding Elements	HCNE	length $\geq$ 50bp; P(S) $\geq$ 0.95*	mouse, dog	Lindblad-Toh et al., 2005
Long Conserved noncoding Sequences	LCNE	length $\geq$ 50bp; identity $>$ 95%	mouse	Sakuraba et al., 2008

Table 3.1: Examples of names used for CNEs. * P(S): probability of selection

3.2.1 Conserved non-coding elements are found universally in metazoan

Many CNEs were found in the human genome when compared to genomes of evolutionarily distant species that diverged 80 to 450 million years ago. These comparisons include: human-mouse (Mural et al., 2002; Bejerano et al., 2004; Nobrega et al., 2003; Sakuraba et al., 2008), human-mouse-dog (Lindblad-Toh et al., 2005), and human-pufferfish (Woolfe et al., 2005). Human CNEs are highly conserved within vertebrates (Harmston et al., 2013), but lack homologous sequences in invertebrate genomes (Woolfe et al., 2005; Vavouri et al., 2007). In addition, lineage-specific CNEs are found among insects (Glazov et al., 2005; Siepel et al., 2005), worms (Siepel et al., 2005; Vavouri et al., 2007; Kent and Zahler, 2000), yeast (Siepel et al., 2005) and plants (Kritsas et al., 2012; Burgess and Freeling, 2014).

Despite the high sequence divergence between CNEs from different clades, they share various properties: they cluster near transcriptional and developmental regulators; they are enriched in TFBSs; they have distinct sequence properties, e.g., sharp depletion of A+T at their boundaries (Vavouri et al., 2007).

3.2.2 Functions of conserved non-coding elements

CNEs are believed to modulate gene expression in *cis* (Boffelli et al., 2004). *In vivo* transgenic assays in mouse confirmed enhancer activities of 45% of 167 tested CNEs (conserved in human-pufferfish or ultra-conserved in human-mouse-rat) (Pennacchio et al., 2006). Similar assays in zebrafish and frog yielded similar conclusions (Table 3.2.2). CNEs functions are diverse. For instance, out of 13 CNEs that did not exhibit enhancer characteristics in zebrafish, 3 showed enhancer blocking function (Royo et al., 2011). Moreover, CNEs are implicated in alternative splicing regulation; CNEs flank 77% of conserved exon alternative splicing between human and mouse (Sorek and Ast, 2003) and every splicing regulator (SR protein family) alternatively spliced CNEs in human and mouse (Lareau et al., 2007). Moreover, CNEs were linked to alternative splicing regulation in *Drosophila* genus (Glazov et al., 2005)

Nevertheless, some CNEs functions are yet to be determined. For instance, alteration of a UCE sequence embedded in Dc12 enhancer— identical in human, mouse, chicken, zebrafish and pufferfish— had no impact on the function of the Dc12 enhancer. Moreover, homozygote mouse embryo knockout of a UCE yield viable mouse (Ahituv et al., 2007). Although, this may also suggest the dispensability or functional redundancy of some UCEs.

CNEs	**Assay Species**	**Number of tested CNEs**	**Number tested positive CNEs**	
human, pufferfish or UCE (human,mouse,rat)	mouse	167	74	Pennacchio et al., 2006
humane, zebrafish	zebrafish	16	10	Shin et al., 2005
human, pufferfish	zebrafish	25	23	Woolfe et al., 2005
human, vertebrates	zebrafish	35	22	Royo et al., 2011
human, mouse, rat, chicken, frog, zebrafish, and two pufferfishs	zebrafish	23	16	Allende et al., 2006
human, mouse, rat, chicken, frog, zebrafish, and two pufferfishs	frog	26	10	Allende et al., 2006

Table 3.2: Summary of *in vivo* transgenic assays on CNEs.

3.2.3 Conserved non-coding elements and disease

Many CNEs regulate expression of key developmental genes, mostly transcription regulators, in a tissue- and stage-specific manner. Mutations in CNEs sequences can alter binding of transcription factors causing misregulation of their target genes. Consequently, this may lead to developmental disorders. For example, mutations in two CNEs located in the upstream and downstream regions of SOX9 resulted in its misregulation that caused Pierre Robin syndrome (developmental defect

in newborns characterized by cleft palate, small jaw, malformed tongue, and breathing difficulties) (Benko et al., 2009). Moreover, deletion of a 5777bp locus located $\sim$ 80kb downstream of HMX1 caused its misregulation and produced rats with abnormal ears and eyes (Quina et al., 2012).

Single nucleotide polymorphisms (SNPs) within CNEs were associated to colon cancer (Lin et al., 2012) and breast cancer (Yang et al., 2008), although the detected SNPs related to breast cancer were specific to the investigated population (Catucci et al., 2009).

Chapter 4

The Evolution of Ultraconserved Elements with Different Phylogenetic Origins

This chapter was published as:

- Taewoo Ryu, Loqmane Seridi and Timothy Ravasi. The evolution of ultraconserved elements with different phylogenetic origins. *BMC Evolutionary Biology* 2012, 12:236. http://www.biomedcentral.com/1471-2148/12/236.

4.1 Background

Large numbers of DNA elements ($\geq$ 200 bp) exhibiting 100% similarity have been found to be conserved across several mammalian species (Bejerano et al., 2004; Ovcharenko, 2008). Shorter ultraconserved elements (UCEs) longer than 50 bp and 100 bp have also been identified in several insect species and plants, respectively (Glazov et al., 2005; Zheng and Zhang, 2008).

Since the discovery of UCEs, a lot of effort has been expended on elucidating their functions and to determine the reasons for their extreme conservation. UCEs are often located near genes implicated in transcription and developmental processes, splicing, and ion flow control across membranes (Bejerano et al., 2004; Ovcharenko, 2008; Papatsenko et al., 2006; Ni et al., 2007; Lareau et al., 2007). *In vivo* analysis of the embryos of transgenic mice uncovered the transcriptional enhancer activities of UCEs targeting developmental genes and TFs (Pennacchio et al., 2006; Visel et al., 2008). Depletion of UCEs among segmental duplications and copy number variations were also reported (Derti

et al., 2006). SNPs in UCEs have been linked to cancer risk, impaired TF binding, and homeobox gene regulation in the central nervous system (Yang et al., 2008; Poitras et al., 2010). Nevertheless, homozygote embryo knockout experiments in mice revealed that deletion of ultraconserved elements can yield viable mice, suggesting the dispensability or functional redundancy of UCEs (Ahituv et al., 2007).

The origin and evolution of UCEs have also been also investigated. There is evidence that some UCEs originated from retroposons and stabilized in genomes after acquiring a function that benefitted the host (Bejerano et al., 2006). Stephen et al. studied the evolution of UCEs in several vertebrate genomes and found that they were generated and expanded on a large scale during tetrapod evolution (Stephen et al., 2008). Other studies of the human genome showed that UCEs experienced strong purifying selection and were not mutational cold spots (Katzman et al., 2007; Lin et al., 2008; Sakuraba et al., 2008).

In this study, we investigated if evidence of the conservation of DNA elements could be found in primitive species, such as sponge and hydra, and if these conserved elements have similar functions as those previously reported for higher eukaryotes. We identified many UCEs across diverse phyla, including Porifera, Cnidaria, Arthropoda, Echinodermata, and Chordata, as well as a new type of short UCEs. By comparing distant species, we were able to identify new UCEs in human and fruitfly. Clustering the UCEs based on the sequence similarity unveiled lineage specificity and distinct functions outlined by protein domains of their flanking genes and DNA regulatory motifs. We concluded that each UCE group arose independently on a specific lineage and was "frozen" on the genome as a regulatory innovation after the divergence of specific taxa.

4.2 Methods

4.2.1 Data preparation

Genome sequences, gene annotation, and protein sequences were downloaded from the UCSC database for human (assembly version: hg19) and fruitfly (assembly version: dm3), and each genome project for sponge (assembly version as of 5 Aug 2010) (Srivastava et al., 2010), hydra (assembly version as of 28 Jan 2009) (Chapman et al., 2010), sea anemone (assembly version as of 26 Oct 2005) (Putnam et al., 2007), and sea urchin (assembly version as of 13 Oct 2006) (Sodergren et al., 2006).

4.2.2 Phylogenetic analysis

First, we identified single copy genes from each of six species under investigation to infer their phylogenetic relationships. This approach had been used previously in other studies to avoid the paralogy issue (Putnam et al., 2007; Roth et al., 2008; Dessimoz et al., 2006). Inparanoid was used to identify orthologs and paralogs between species pairs (Ostlund et al., 2010). Only the longest peptide was used when multiple transcripts came from the same gene. We identified 472 single-copy genes that were found to be largely involved in ribosome, spliceosome, or proteasome pathways. Gene sequences were aligned using MUSCLE (Edgar, 2004) and the evolutionary distance and phylogenetic tree were obtained using MEGA5 (Tamura et al., 2011). The phylogenetic tree reveals the overall relationship between six species, which was in agreement with the known classification of these lineages (Figure 4.1) (Sodergren et al., 2006; Ryu et al., 2011; Dunn et al., 2008).

4.2.3 Identification of ultraconserved elements

To identify UCEs for all species pairs, we masked repetitive sequences in the scaffolds of sponge, hydra, sea anemone, and sea urchin using CENSOR (Kohany et al., 2006) and tandem repeats finder (Benson, 1999). Repeat-masked chromosomes from the UCSC database were used for human and fruitfly (Karolchik et al., 2003). To identify non-gapped conserved elements between two species, we used MUMmer, which rapidly aligned long sequences and detected exact matches using the suffix tree algorithm, with the maxmatch option to compute all maximal identical matches regardless of uniqueness (Kurtz et al., 2004). Both forward and reverse complement matches were reported. Identical matches equal to or longer than 50 bp were identified, and $\geq$ 30 bp matches were also identified for incidental analysis. Identified UCEs were further masked using CENSOR and tandem repeat finder again. It should be mentioned that this stringent repeat-masking process may have deleted potential UCEs containing repetitive elements.

Two UCEs were joined if they overlapped, and this merging process was repeated until no two UCEs overlapped. Fifty base flanking sequences on both sides of merged UCEs were retrieved using the custom python script.

4.2.4 Clustering of ultraconserved elements

Merged ultraconserved elements with flanking sequences were grouped by sequence similarity. Pairwise alignment of all sequences was computed using BLASTN (Altschul et al., 1990). The score density, i.e. the BLAST bit-score divided by the alignment length, was used as the similarity mea-

sure. Sequences were clustered using the Markov cluster (MCL) algorithm (Enright et al., 2002) with default parameters (Additional file 4). In the Minimum Curvilinear Embedding (MCE) analysis (Cannistraci et al., 2010), 5-mer compositions of the sequences were used as features. In particular, we used the new singular-value-decomposition-based algorithm to implement MCE (Cannistraci et al., 2013), using the Matlab code provided on the author's website (https://sites.google.com/site/carlovittoriocannistraci/home). The embedding was performed without centering the minimum curvilinear kernel (non-centered MCE).

4.2.5 Nearby genes analysis

Flanking genes within 100 kb of the merged UCEs were obtained from all species under study. For human and fruitfly, we used the gene models from RefSeq (Pruitt et al., 2009). We used the gene models from the respective genome sequencing projects of the non-model metazoans.

Pfam domains of nearby genes were annotated using Interproscan (Zdobnov and Apweiler, 2001) for functional analysis of UCEs. For each domain in each ultraconserved region (UCR) cluster, the domain enrichment of nearby genes within 100 kb of UCRs was calculated using cumulative hypergeometric distribution:

$$P = \sum_{i=d}^{min(D,g)} \frac{(\frac{D}{i})(\frac{G-D}{g-i})}{(\frac{G}{g})} \tag{4.2.1}$$

where G is the total number of genes from the species pool in the cluster, g is the number of selected nearby genes in the species pool in the cluster, D is the number of occurrences of the domain in the species pool in the cluster, and d is the number of occurrences of the domain in the selected nearby genes in the species pool of the cluster.

Gene ontology enrichment of the nearby genes was analyzed using DAVID (Huang et al., 2009). Considering that human has the most comprehensive biological process terms and nearly nothing is annotated in non-model species, only human UCRs and their nearby genes were analyzed.

4.2.6 Motif analysis

A representative sequence of each cluster was generated using MUSCLE (Edgar, 2004) and the seqinR package in R (Charif and Lobry, 2007). To assess the statistical significance of overrepresented 8-mers, we generated a 10 kb background sequence for each cluster. The background sequence was a combination of segments chosen randomly from all genomes, and each genome contributed to the background with an amount equal to the ratio of its species in the cluster composition. A cumulative

binomial probability of observing the given number of the oligomer or more in each cluster was then computed as follows:

$$F(x|n,p) = 1 - \sum_{i=0}^{x-1} \binom{n}{i} p^i (1-p)^{n-i} \tag{4.2.2}$$

where x is the number of occurrences of the oligomer, n is the sample size, i.e. $sequence\ length - oligomer\ size + 1$, and p is the probability of observing such an oligomer in the random background sequence. Related TFs for oligomers were identified using STAMP (Mahony and Benos, 2007).

4.3 Results and discussion

4.3.1 Identification of ultraconserved elements across diverse taxa

We began our analysis by asking if there is evidence of ultraconservation in primitive species and, if so, how UCEs diverged during the process of evolution. We considered six species whose genomes were previously sequenced including sponge (*Amphimedon queenslandica*) from the phylum Porifera, hydra (*Hydra magnipapillata*) and sea anemone (*Nematostella vectensis*) from the phylum Cnidaria, sea urchin (*Strongylocentrotus purpuratus*) from the phylum Echinodermata, fruitfly (*Drosophila melanogaster*) from the phylum Arthropoda, and human (*Homo sapiens*) from the phylum Chordata. We identified UCEs ($\geq$ 50 bp) and shorter UCEs ($\geq$ 30 bp) by pairwise comparison of the whole genomic sequences across six species.

Unexpectedly, the number of identified UCEs and the size of some of them (11 UCEs $\geq$ 200 bp) were large considering the evolutionary distance between analyzed species. This result suggested the presence of UCEs in primitive species and across distant taxa (Table 4.1 and Figure 4.1). Most of the UCEs were found in hydra and sea anemone, which belong to the same phylum, Cnidaria. However, the exact reason for the predominance of UCEs in these species cannot be addressed until more genome sequences of species around this lineage become available and current genome assemblies are improved. Interestingly, the longest UCE (796 bp) was conserved in both sea anemone and human, two species that diverged approximately 892 million years ago (Hedges et al., 2006). We found that the number of UCEs and the evolutionary distance (Table 4.1 and Figure 4.1) between species are negatively correlated, an observation that is also the case for shorter UCEs.

	A. queenslandica **(sponge)**	***N. vectensis*** **(sea anemone)**	***H. magnipapillata*** **(hydra)**	***D. melanogaster*** **(fruitfly)**	***S. purpuratus*** **(sea urchin)**	***H. sapiens*** **(human)**
A. queenslandica	-	2,135	669	43	108	9
N. vectensis	5,303	-	54,732	256	5,525	10
H. magnipapillata	1,300	97,669	-	125	400	0
D. melanogaster	75	5,440	478	-	188	27
S. purpuratus	537	43,707	5,498	1,129	-	19
H. sapiens	83	381	328	415	967	-

Table 4.1: **Identification of UCEs.** Columns and rows are sorted by the phylogeny as shown in Figure 4.1. Upper and lower triangles show the numbers of 100 % identical matches $\geq$ 50 bp and $\geq$ 30 bp between two species, respectively.

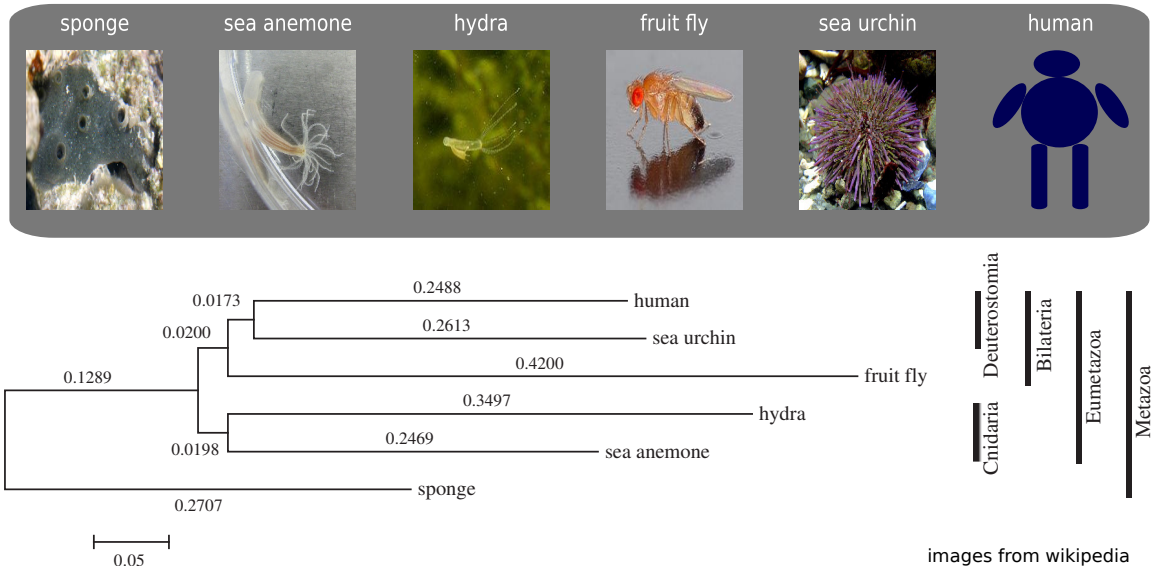

Figure 4.1: **Evolutionary relationships between analyzed species.** The JTT matrix-based method (Jones et al., 1992) is used to compute the evolutionary distances and the phylogenetic tree is constructed using the Neighbor-Joining method (Saitou and Nei, 1987). Bootstrapping values from 500 replicates are shown and selected taxon information is depicted on the right. Species abbreviations are as follows: AQ: *Amphimedon queenslandica* (sponge), DM: *Drosophila melanogaster* (fruitfly), HM: *Hydra magnipapillata* (hydra), HS: *Homo sapiens* (human), NV: *Nematostella vectensis* (sea anemone), SP: *Strongylocentrotus purpuratus* (sea urchin).

We noticed that a large number of conserved DNA elements that we identified overlapped in each species because the UCE-identification program, MUMmer, reported all maximal matches regardless of the overlap (Kurtz et al., 2004). To minimize redundancy and facilitate downstream analysis, neighboring UCEs and short UCEs in each species were joined as non-overlapping UCRs. The numbers of these non-overlapping UCRs ($\geq$ 50 bp) were 30 for sponge, 64 for fruitfly, 673 for hydra, 56 for human, 3,807 for sea anemone, and 187 for sea urchin.

4.3.2 Novel ultraconserved elements in human and fruitfly

As a benchmark for our UCE discovery pipeline, we examined how many UCEs that had been previously identified we were able to recover. Previously reported UCEs in human and fruitfly were aligned to their reference genome using Bowtie (Langmead et al., 2009) to determine their exact locations in the current genome build (hg19 and dm3, respectively). The majority of known UCEs (all 481 elements from the human-mouse-rat alignment (Bejerano et al., 2004), 23,695 out of 23,699 elements from the *D. melanogaster Drosophila pseudoobscura* alignment, and all 126 elements from the *D. melanogaster Anopheles gambiae* alignment (Glazov et al., 2005)) were successfully aligned. We then compared these elements with our UCR set. Unlike in the fruitfly where 42 out of 64 UCRs

overlapped with data reported by Glazov et al. (Glazov et al., 2005), we could not find any UCR in human that overlapped with previously reported UCEs (Bejerano et al., 2004).

To understand this incongruence, we tested if our pipeline could recover known UCEs of the human-mouse-rat alignment with the same species list and length constraint ($\geq$ 200 bp) of Bejerano et al. (Bejerano et al., 2004). Our pipeline recovered 464 out of 481 known human UCEs that are conserved both in mouse and rat. The missing 17 known UCEs overlapped with repetitive regions, and these elements could not be recovered by our pipeline, which masks repetitive elements. Furthermore, the human UCEs that were conserved in mouse and rat identified by our pipeline did not also overlap with those newly identified in this study, suggesting that our pipeline works properly. The effect of the genome assembly version used for UCE identification was also negligible as explained above. On the other hand, our stringent repeat masking reduced the number of detectable known UCEs. The numbers of known UCEs were 304, 20,602, and 83 for human-mouse-rat, *D. melanogaster D. pseudoobscura*, and *D. melanogaster A. gambiae*, respectively, when we removed known UCEs with simple and known repetitive elements by repeat-masked chromosomes (Karolchik et al., 2003), CENSOR (Kohany et al., 2006), and tandem repeat finder (Benson, 1999), the same criteria that we used in this study. However, the most important factor contributing to the identification of novel UCRs was the length constraint (50 bp for human) and species compared. To test this further, our human UCR set was divided into 50 bp sub-sequences, and then a search for these sub-sequences in the genomes of mouse and rat was conducted. Of 28 UCRs, one sub-sequence occurred in both the mouse and rat genomes with 100% similarity. On the other hand, the other 28 UCRs were not conserved in both species, suggesting that those sequences were no longer under strong selective pressure in rodents and could therefore not be identified by the traditional human-mouse-rat alignment. Indeed, large portions of identified human UCEs are positioned in less conserved loci in placental mammals (Figure 4.2), which further supports our findings of novel highly conserved DNA elements in model organisms.

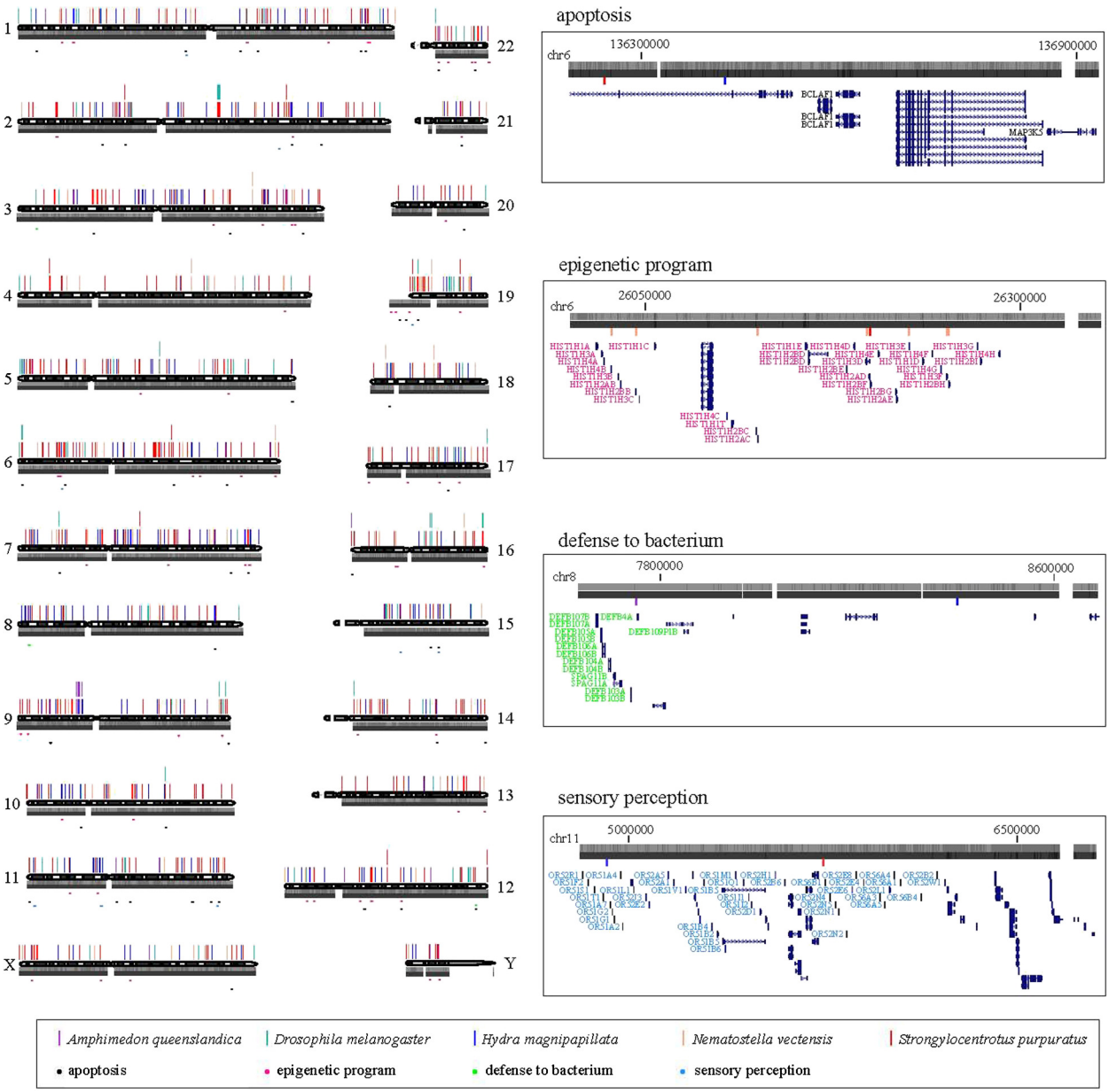

Figure 4.2: **UCEs in the human genome.** The short conserved elements ($\geq$ 30 bp) are depicted above each chromosome and the UCEs ($\geq$ 50 bp) are depicted above them. The conservation level of human DNA across primates or placental mammals is shown below each chromosome, where the darker color indicates more conservation. Species in which the human sequences are conserved and selected gene categories are labeled as indicated at the bottom. The selected region for each gene category is magnified for a detailed view in the right panel. R package quantsmooth (Oosting et al., 2009) and the UCSC custom track (Karolchik et al., 2003) were used for the plot.

4.3.3 UCR clusters arose independently

We then sought evidence for if UCRs from the same or different species share similarity. Considering the short length of UCRs and also assuming that distal regions of ultraconserved elements have higher mutation rates than proximal regions (Stephen et al., 2008; Faircloth et al., 2012; Halligan

et al., 2011), we analyzed UCRs and their 50 bp-flanking sequences. In all, 4,817 UCRs with flanking sequences from all species were clustered, and orthologous and paralogous UCRs were defined. This yielded 61 clusters, of which the largest cluster consisted of 1,168 UCRs from hydra, sea anemone, and sea urchin. Although there are large numbers of UCRs across different taxa, we found that UCRs share sequence similarities and that each cluster of UCRs has a distinct species composition. Moreover, Cnidarian UCRs show a tight association, while human UCRs are largely clustered together with those of sea urchin and/or fruitfly. Gain of essential functions for the survival of the species in ancestral sequences might contribute to the conservation of the sequence in a specific lineage (Bejerano et al., 2006). Another possible explanation would be that even if the ancestral sequences were not beneficial to the species, random sampling contributed to the elimination of other alleles and the fixation of these sequences in the downsized population, creating a new lineage, due to natural catastrophe or population migration, referred to as a "genetic drift" "population bottleneck" (Gherman et al., 2007). Although further study is required to explain the immutability of UCEs after lineage divergence and sequence fixation across a long evolutionary history, we cannot rule out this possibility. It also should be noted that the absence of UCRs in species from the same lineage does not necessarily mean that those UCRs disappeared in those species but rather that they may exist as derivative sequences by mutation (Ovcharenko, 2008; Stephen et al., 2008; Wang et al., 2009; Kim and Pritchard, 2007).

As shown in Figure 4.3A, UCR clusters are clearly separated in a Minimum Curvilinear Embedding (MCE) plot (Cannistraci et al., 2010), although species is not a good factor to distinguish UCRs (Figure 4.3B). Short UCRs ($\leq$ 30 bp) also followed a similar pattern. Interestingly, some clusters have nearly symmetric elements on the MCE plot and it turns out that they are partially reversed complementary sequences.

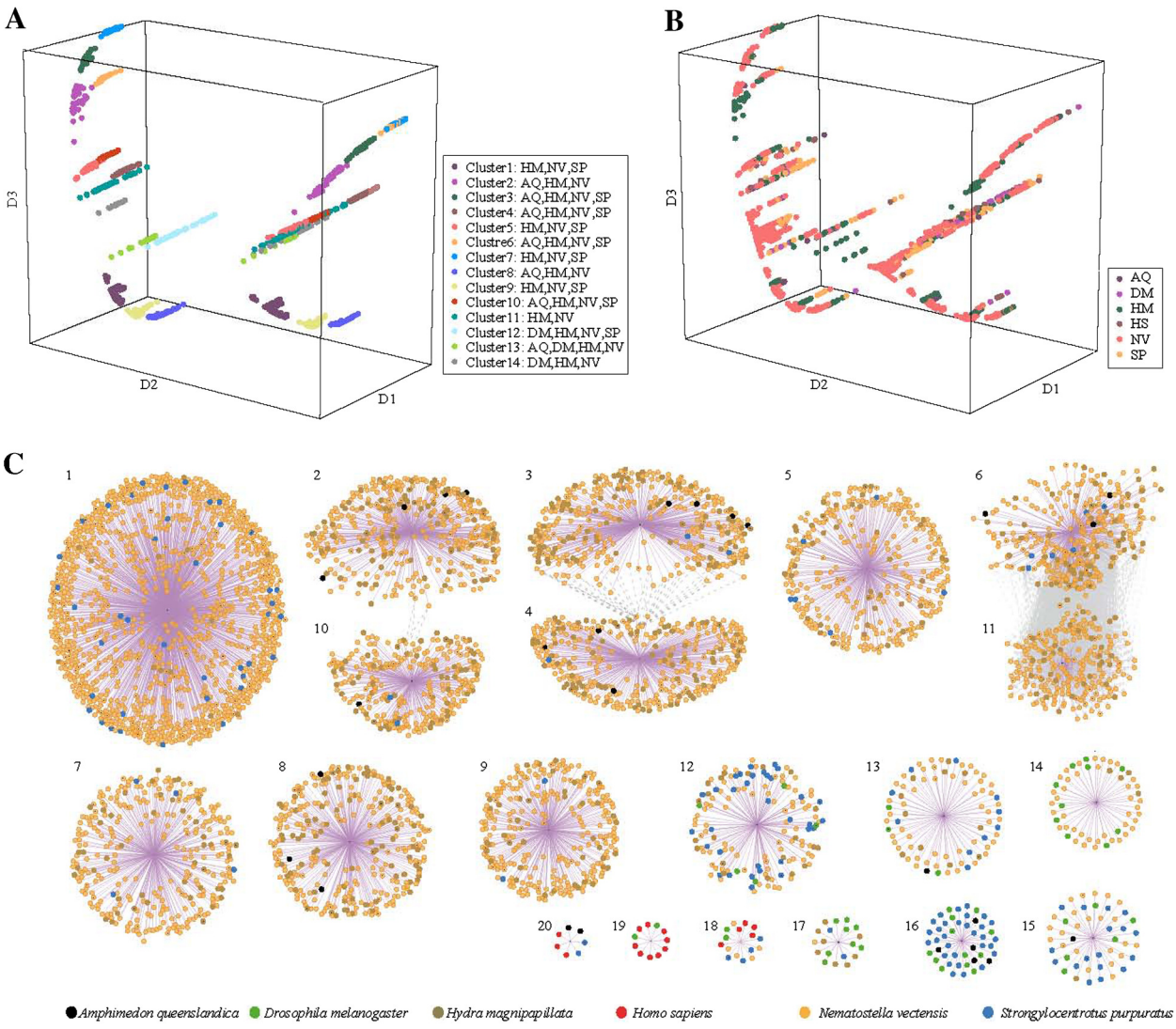

Figure 4.3: **UCR clusters arose independently during evolution. A** and **B**. The 5-mer composition of UCRs and 50 bp-flanking sequences were taken as input for the Minimum Curvilinear Embedding analysis and the top three dimensions are depicted here. Clusters and species are marked with different colors as indicated in the inset on the lower right corner. **C**. UCR cluster relationships. Each node represents a UCR with flanking sequence. The similarity between nodes in a same cluster is omitted to avoid extreme density. A cluster centroid is made instead and connected to the components to show membership within the cluster (purple lines). The gray lines show sequence similarities between nodes in different clusters. Clusters with fewer than 7 nodes are not shown. UCRs from sea anemone predominate in this figure due to the large number (3,807 among all 4,817 UCRs).

Network topology demonstrates the relationship between these UCR clusters, where some clusters are connected due to the sequence similarity between components, although most clusters do not share sequence similarity with others and have unique species composition (Figure 4.3C). Thus, the UCRs of each cluster may have their own independent origin in a specific lineage.

4.3.4 The neighboring genes of UCR have distinct functions

UCEs are often flanked by developmental genes, TFs, ion channels, or splicing factors (Papatsenko et al., 2006; Lareau et al., 2007). We investigated the functions of each cluster's nearby genes. Due to the paucity of functional annotations of genes and the short length of genome scaffolds in non-model species, we focused our analysis on the protein domains of nearby genes within 100 kb from UCRs. Neighboring genes to UCR clusters span a spectrum of statistically significant protein domains. However, each cluster is enriched with a distinct set of domains (Figure 4.4).

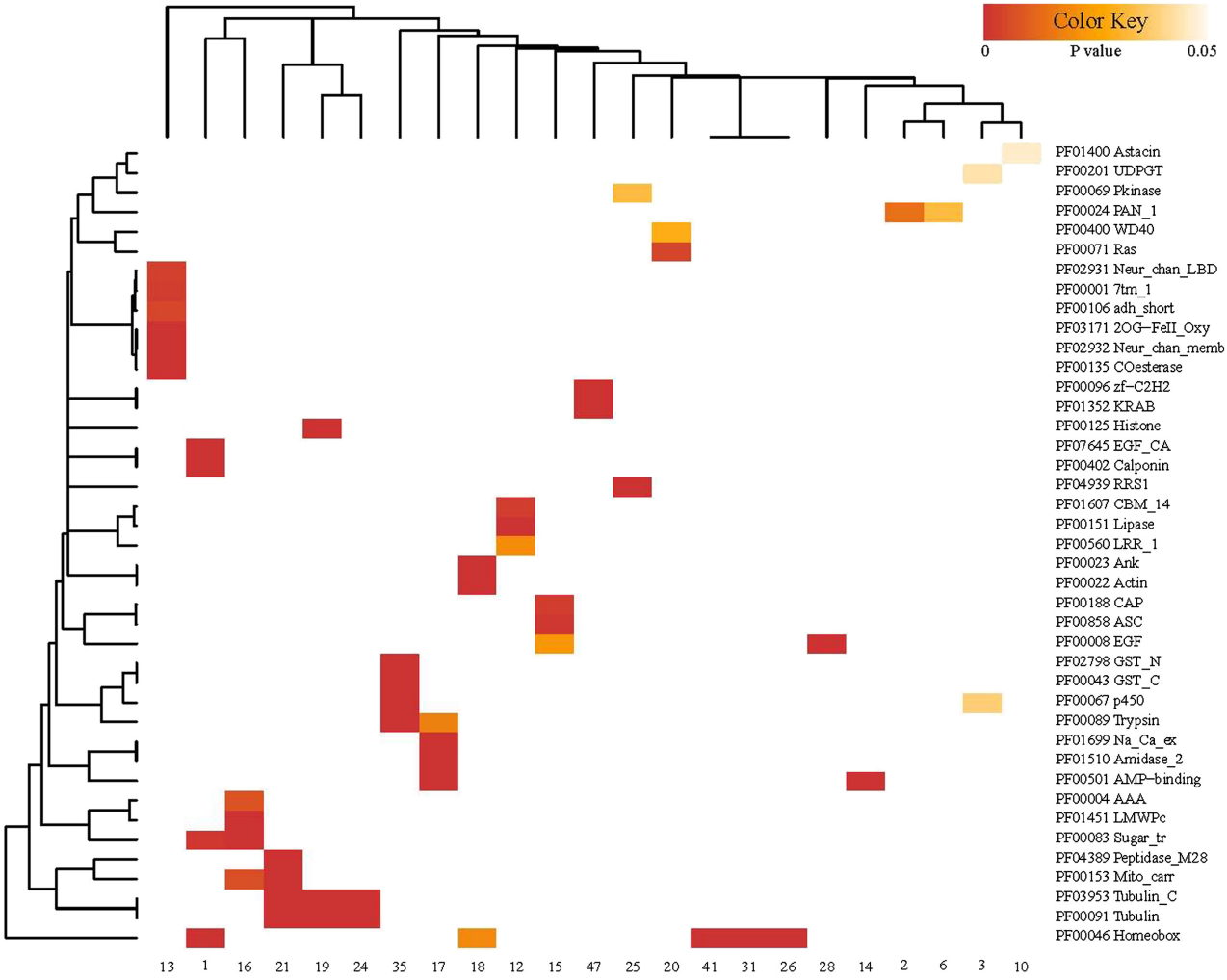

Figure 4.4: **Protein domain enrichment of UCR flanking genes.** Association of domains (rows) or clusters (columns) is depicted in dendrograms on the right and upper side of the heatmap, respectively. Only clusters having at least 10 genes were analyzed. Domains whose p-value < 0.05 in at least one cluster and that occurred in at least three nearby genes are shown on the heatmap for visualization purpose.

Ion channel and transporter domains are the predominant categories; they appear in many clusters composed of various species. Neurotransmitter-gated ion channels and sodium or calcium

ion exchanger genes are overrepresented in clusters 13, 15, and 17, whose UCRs are conserved in all species considered here but human (Figure 4.4). Cation transporters are identified in cluster 30, which consists of human and fruitfly UCRs. Sugar transporters and mitochondrial carrier domains that transport various molecules across membranes are enriched in clusters 1, 16, and 21. These observations are probably because ion channels and transporters are crucial in all living organisms for the maintenance of water, salt, and nutrient homeostasis as well as for electric signal transmission in neuronal and muscle cells (Dubyak, 2004).

The homeobox domain, part of the TFs that act during the developmental process, is enriched in five clusters. This domain is found in all six species, with three of the five enriched UCR clusters composed of UCRs from human and fruitfly, one from fruitfly and sea urchin, and the last cluster from hydra, sea anemone, and sea urchin. Fruit fly genes regulating developmental programs ranging from axis patterning to molting, such as bicoid, fushi tarazu, and ecdysone receptor, are also found in several clusters, even those without significant domains.

Histones are overrepresented in cluster 19, which consists of sea anemone and sea urchin UCRs. Evidence that chromatin-related genes flank conserved elements in human and from other studies (Lee et al., 2006; Bernstein et al., 2006) suggest that there is a liaison between conserved elements and epigenetic control mechanisms.

Detoxification domains such as cytochrome p450, UDPGT, and GST are enriched in cluster 3 and cluster 35. Cluster 3 consists of UCRs from sponge, hydra, sea anemone, and sea urchin; cluster 35 consists of UCRs from fruitfly and human. These enzymes are important to catalyzing and eliminating endogenous and exogenous substrates and therefore to providing a healthy environment for the cellular system (Ishii et al., 2005). This remarkable linkage between UCRs and detoxification mechanisms has not previously been reported to our knowledge.

Further analysis of UCRs ($\geq$ 50 bp) and short UCRs ($\geq$ 30 bp) in human reveals similar but more interesting properties in terms of nearby gene functions and species conservation . Genes acting in various developmental processes are highly enriched near the UCRs in human that are also conserved in fruitfly and sea urchin. To our surprise and contrary to previous studies, few genes related to development are enriched near the human sequences conserved in sponge, hydra, or sea anemone. Expansion of the relationship between developmental programs and UCRs in human, fruitfly and sea urchin (Figure 4.1) implies that the association of conserved sequences with the regulation of developmental genes started or expanded after the divergence of the Bilateria lineage from the metazoan stem. Our UCR clustering results bolster this hypothesis (Figure 4.4). Four out of five UCR clusters that have overrepresented homeobox domains of nearby genes come from

human, fruitfly, and sea urchin.

Interestingly, genes surrounding short UCRs are enriched with epigenetic program-related genes (Figure 4.2). Short UCRs conserved in human and in fruitfly, hydra, sea anemone, or sea urchin are located near histone gene clusters across several chromosomes. Furthermore, many important epigenetic regulators are also found near elements conserved in sponge, hydra, sea anemone, or sea urchin. These include histone demethylases (KDM3B, KDM4C, KDM5C, and KDM5D), histone acetyltransferases (EP300 and KAT7), histone deacetylases (HDAC2 and HDAC10), retinoblastoma-like protein (RBL1), polycomb ring finger oncogene (BMI1), chromodomain helicase (CHD8), and components of the chromatin remodeling complex, SWI/SNF (SMARCA2, SMARCB1, SMARCC2, and SMARCD3). Taken together with the previously suggested relationship between highly/ultra-conserved elements and epigenetic control (Stephen et al., 2008; Lee et al., 2006; Bernstein et al., 2006), our results suggest an interesting hypothesis that epigenetic control mechanisms have tight relationships with conserved DNA sequences and that they might have coevolved from metazoan ancestors rather than recently developed.

Genes implicated in apoptosis, olfactory reception, and defense mechanisms are also enriched near DNA elements conserved in sponge, hydra, or sea urchin (Figure 4.2). Our analysis suggests that genomes preserve ancestral sequences well, and these ancestral sequences might have coevolved with a diverse set of essential genes. When and how genes and conserved elements initiated their relationships remains unclear and the mechanism for such an association needs to be further elucidated. However, our analysis expands the repertoire of conserved genomic elements that are possible regulatory elements.

4.3.5 UCR are enriched with binding sites for developmental TF

The enhancer activities of UCEs have been reported by several studies (Pennacchio et al., 2006; Visel et al., 2008). To investigate the possibility that these enhancer activities were also conserved in primitive species, we identified significantly overrepresented oligomers and related TFBSs for each UCR cluster (Figure 4.5).

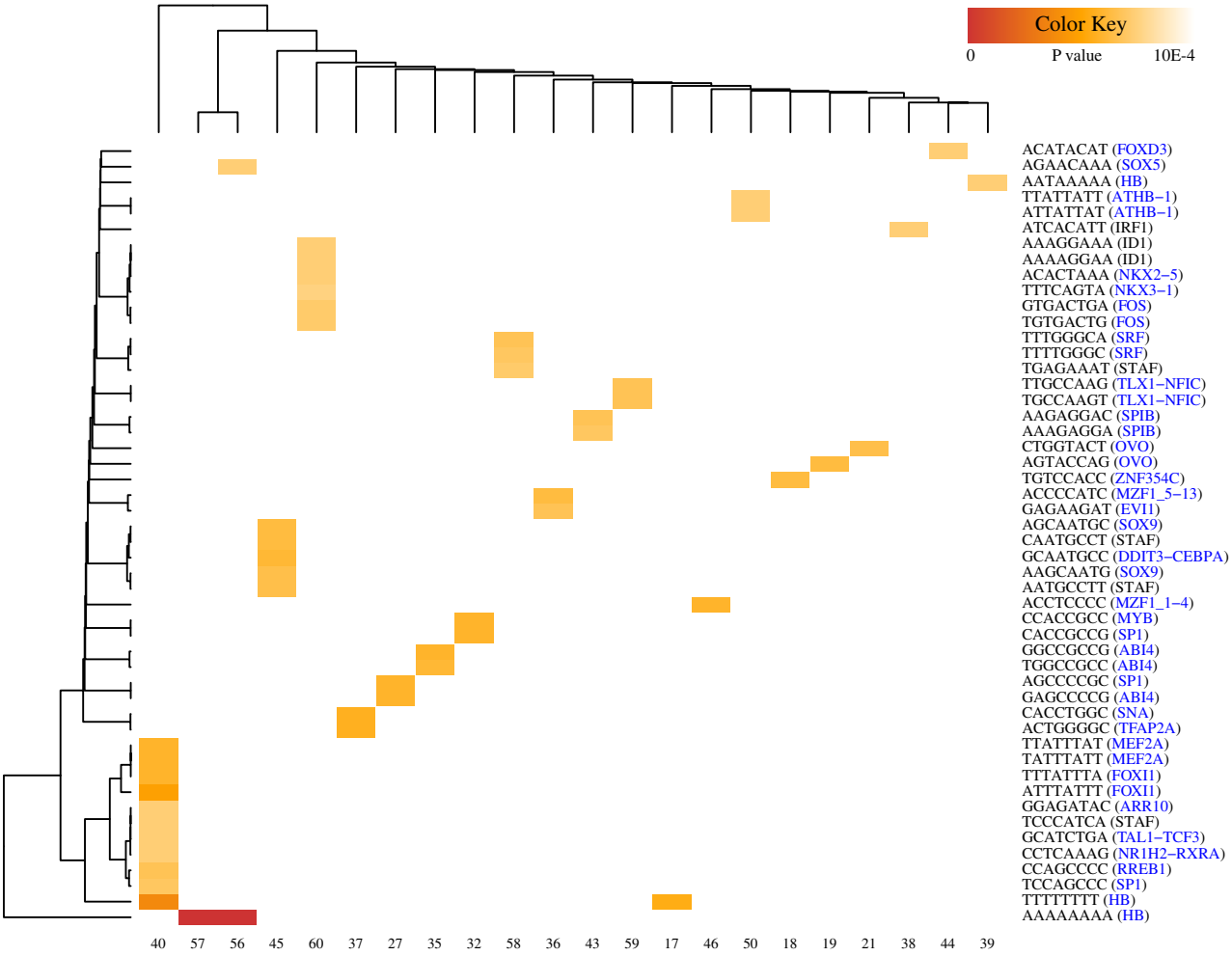

Figure 4.5: **Oligomer enrichment on UCRs.** Each cluster (column) shows distinct TFBS patterns. Association of 8-mers and clusters is depicted in dendrograms. Only 8-mers with p-value$<1 \times 10^{-4}$ are shown by the heatmap. The most related TFBSs of each 8-mer are shown in the brackets on the right side. TFs with developmental functions indicated by NCBI Gene (NCB) or GeneCards (Safran et al., 2010) are colored blue.

Among 31 TFs that had significant 8-mer matches, 28 were implicated in developmental processes and many were homeobox TFs. Binding sites of homeobox TFs on UCEs near the developmental genes in higher eukaryotes have been identified (Chiang et al., 2008; Lampe et al., 2008; Rodelsperger et al., 2009), although our clustering results identified various nearby gene categories that were not limited to developmental genes. Prevalent occurrence of developmental TFBSs regardless of cluster and species may be an indication that extensive binding of developmental TFs on UCEs existed in metazoan ancestors and these TFs regulated various nearby genes to coordinate developmental functions. These may have contributed to the strong selective pressure on UCEs that function as regulatory sequences.

4.4 Conclusions

Genomes are dynamic entities and are under selective evolutionary pressure from mutation and fixation. Beneficial or neutral mutations in the ancestors of specific lineages are maintained in the population and vertically transferred to descendants (Gogarten and Townsend, 2005). However, these dynamic and selective pressures are not applied uniformly across the whole genome (Katzman et al., 2007; McLean and Bejerano, 2008; Lander et al., 2001). Deleterious mutations in essential regions are corrected in a population (Stephen et al., 2008; Katzman et al., 2007). Sequence conservation thus implies that the function of the sequence is essential. Despite controversy about the indispensability of ultraconserved elements (Ahituv et al., 2007; Gross, 2007), much work has demonstrated various vital functions of such elements (Papatsenko et al., 2006; Ni et al., 2007; Pennacchio et al., 2006; Visel et al., 2008; Derti et al., 2006).

As more genomes from various taxa are being sequenced, the opportunity to understand genome conservation and usage increases. Here, we compared genome sequences ranging from primitive aquatic to higher terrestrial species and described for the first time a number of novel UCEs present in primitive species as well as previously uncharacterized UCEs in human and fruitfly. We observed that UCEs cluster by sequence similarity and each cluster has distinct patterns of species composition. These UCEs also exhibited specific biases toward the function of nearby genes and oligomer compositions of the UCE sequences, suggesting that each group of UCEs was generated in the common ancestors of specific lineages and fixed during the evolution of descendants. Although a more detailed functional analysis of UCEs cannot currently be conducted due to the nature of the short draft sequences and because gene functions of non-model species have been less studied, our analysis suggests that UCEs harbor important sequence features, such as binding sites of developmental TFs to coordinate the expression of essential genes, which is why they were readily conserved over the long course of evolution.

Chapter 5

Dynamic Epigenetic Control of Highly Conserved Noncoding Elements

This chapter was published as:

- Loqmane Seridi, Taewoo Ryu and Timothy Ravasi. Dynamic epigenetic control of highly conserved noncoding elements. *PLOSONE*. DOI: 10.1371/journal.pone.0109326 ©Seridi et al 2014.

5.1 Introduction

Genomic DNA is subject to diverse mutations caused by chemicals, replication errors, and mobile genetic elements. Coding sequences are generally under higher selective pressure than noncoding sequences, due to the essential roles that proteins play in the cell (Pang et al., 2006). However, some noncoding regions show extreme conservation (even more than coding sequences) over very long evolutionary timeframes (Bejerano et al., 2004; Ryu et al., 2012). These extremely conserved sequences are found universally in multicellular eukaryotes of the animal and plant kingdoms (Ryu et al., 2012; Kritsas et al., 2012), indicating that such sequences have essential functions.

Indeed, many highly conserved noncoding elements (HCNEs), noncoding loci that maintain high level of similarity across different species, function as developmental enhancers (Pennacchio et al., 2006), enhancer-blocking insulators (Royo et al., 2011), and regulators of splicing (De Grassi et al., 2010) and RNA editing (Daniel et al., 2012). Mutations in HCNEs have been associated

with various diseases, including cancers and neurodevelopmental disorders (Yang et al., 2008; Lin et al., 2012; Martinez et al., 2010). However, these functions of HCNEs reflect the activities of the DNA-interacting proteins that bind to very short and degenerate DNA sequences within them (Wasserman and Sandelin, 2004), and are insufficient to explain the invariability of HCNEs ($\geq$ 200 bp) during evolution. Even the enhanceosome model, which requires a strict array of TFBSs over a long sequence (Harmston et al., 2013), does not explain the observed sequence conservation between the TFBSs. We therefore speculated that HCNEs are subjected to a higher level of protection against mutations compared to other sequences.

In eukaryotic cells, histones pack DNA into nucleosomes to form the chromatin structure; this protects DNA from damage, and offers an additional layer of regulation via the chemical modification of histones (Bernstein et al., 2006; Chen et al., 2012). A few previous studies have suggested that HCNEs may be associated with epigenetic control mechanisms. For instance, an analysis of mammalian stem cells found epigenetic modifications of some HCNEs, such as the bivalent methylation, H3K27me3 + H3K4me3 (Bernstein et al., 2006; Akalin et al., 2009; Xiao et al., 2012). In addition, 11% of mammalian HCNEs co-occur with scaffold/matrix attachment regions (S/MARs), which have been implicated in the structural organization and remodeling of chromatin (Glazko et al., 2003). To understand the properties underlying the extreme conservation of HCNEs, however, we need data from a systematic analysis of their potential epigenetic regulation.

Here, we performed an integrative analysis of the epigenetic properties and regulations of HCNEs throughout the development of *Drosophila melanogaster*. Our results indicate the following: HCNEs intrinsically favor stable nucleosome occupancy at the sequence level; HCNEs reside within nucleosome-enriched and mononucleosome- and H3.3-depleted regions in S2 cells; the chromatin regions around HCNEs undergo significant changes in epigenetic modification during development, and such changes are correlated with the transcription levels of flanking genes; most HCNEs fire later in replication, however some serve as early replication origins; and HCNEs are significantly associated with lamina-associated domains (LADs). Our results collectively indicate that HCNEs are under special evolutionary control at the levels of chromatin and nuclear structural organization.

5.2 Materials and Methods

5.2.1 Extraction of HCNEs and generation of background sequences

We downloaded the DM3 compilation of the D. melanogaster genome (Apr. 2006, BDGP Release 5) and extracted highly conserved elements using the ECE algorithm (Tseng and Tompa, 2009) from the phastCons score tracks of 14 insect genomes, as obtained from the UCSC Genome Browser (Siepel et al., 2005; Fujita et al., 2011). We set the minimum length and the phastCons conservation score to 200 bp and 0.95, respectively. We trimmed highly conserved elements overlapping coding regions based on the R5.46 genome annotation from FlyBase (Marygold et al., 2013), and filtered out elements shorter than 200 bp. We used BEDTools (Quinlan and Hall, 2010) to extract HCNE sequences and 10,000 random noncoding sequences having the same length distribution and the same distance distribution to the nearest gene or exon, compared to the intergenic and intronic HCNEs, respectively.

5.2.2 Analysis of HCNE sequence properties

To investigate possible TFBSs within HCNEs, we identified overrepresented heptamers in HCNEs sequences and queried the top 50 against FlyReg motifs (Bergman et al., 2005) using STAMP (Mahony and Benos, 2007). To identify overrepresented heptamers, we computed the binomial probability of whether their observed frequency in HCNEs is higher than expected by chance. We estimated the expected frequency of a heptamer by averaging its frequencies over 10,000 datasets sampled from random noncoding backgrounds (described above); each dataset contains same number of sequences as HCNEs.

5.2.3 Identification of HCNE-proximal genes

We downloaded a list of 1,321 genes showing conserved microsynteny in the *Drosophila* genus, along with their microsyntenic blocks mapping (Sahagun and Ranz, 2012). We used FlyMine (Lyne et al., 2007) to obtain the gene coordinates, and found that two of them were absent from current genome annotation. We determined the coordinates of the GRBs based on the coordinates of genes located at the boundaries of the microsyntenic blocks. For downstream analysis, we selected all genes within 50 kb from GRB boundaries plus genes within 50 kb from boundaries of HCNEs that were not near a GRB. To determine the promoter type of the HCNE-proximal genes, we download McPromoter (Ohler, 2006) predictions from the current genome release

(http://tools.genome.duke.edu/generegulation/McPromoter006/mcpromoter.rel5.thres0.03.gff), using a stringent threshold of 0.03. Each gene was assigned the promoter type prediction found within 500 bp upstream of the TSS and a minimum of 500 bp and the length of 5'UTR downstream of the TSS. When multiple predictions were made for a given gene, we chose the one with the highest score. The 5'UTR coordinates were obtained from FlyMine (Lyne et al., 2007), which was also used to compute the protein domains and assess GO enrichment.

5.2.4 Epigenomic data

We obtained the preprocessed nucleosome occupancy data (profiling by genome-wide tiling array) deposited by Henikoff et al. (Henikoff et al., 2009) from the NCBI GEO database (GSE13217): the nucleosome densities were the averages obtained from GSM333835, GSM333840, and GSM333844; the mononucleosome data were the averages obtained from GSM333837 and GSM333841; the H3.3 occupancies were obtained from GSM333869; and the H3.3dN occupancies were taken from GSM333870. The data for the six studied histone modification markers [H3 lysine 4 mono- and tri-methylation (H3K4me1 and H3K4me3), H3 lysine 9 acetylation and tri-methylation (H3K9ac and H3K9me3) and H3 lysine 27 acetylation and tri-methylation (H3K27ac and H3K27me3)], CBP binding, and replication time were obtained from the modENCODE project. These data span 12 developmental stages, including six embryonic stages, three larval stages, and the pupae, adult male, and adult female. The data were downloaded from the online server (ftp://data.modencode.org/D.melanogaster/Histone-Modification/ChIP-seq/computed-peaks_gff3/). Some developmental time points were missing for some of the histone markers. For example, the larva 3, adult male and adult female stages were missing data for H3K9me3 and the adult female stage was missing data for H3K27me3. We obtained ChIP-Seq peaks for CBPs throughout the same developmental stages, with the exceptions of the embryonic 8 to 12 h and larva 2 stages, which were missing. Missing data were ignored in our analysis. The histone modification and CBP binding data can be found in the modENCODE depository under the following IDs (DCCids): modENCODE_862, modENCODE_863, modENCODE_854, modENCODE_856, modENCODE_857 modENCODE_855, modENCODE_859, modENCODE_860, modENCODE_861, and modENCODE_858. The data for our genome-wide replication timing characterization and the early origin of replication peaks can be found under the following DCCids: modENCODE_668, modENCODE_66 and modENCODE_670 (for replication timing) and modENCODE_3441 (for the early origin of replication peaks). Due to the lack of raw data for many of the studied epigenomic modifications (which is required for normalization across conditions), we

analyzed the frequencies of the histone modification ChIP-Seq peaks on HCNEs rather than the enrichment levels of these modifications. We considered a marker present in an HCNE if the overlap between the peak and HCNE covered at least 10% of the HCNE length.

5.2.5 Gene expression data and analysis

We downloaded the normalized read counts of *D. melanogaster* in reads per kilobase per million (RPKM) for 30 developmental stages and 28 tissues, as compiled in FlyBase. The utilized data are available at FlyBase (ftp://ftp.flybase.net/releases/FB2014_02/precomputed_files/genes/gene_rpkm_report_fb_2014_02.tsv.gz). We changed values greater than 100 to 100 (i.e., they were considered to be very highly expressed). To assess the relationship between histone modification at HCNEs and the transcriptional activity of proximal genes, we downloaded files of aligned reads from the modENCODE ftp server (ftp://data.modencode.org/D.melanogaster/mRNA/RNA-seq/alignment_sam/). We computed normalized read counts (in RPKM) for each gene. We computed the Shannon entropy for the expression of each gene using the following formula:

$$Entropy = -\sum_{i=1}^{n} p_i \times log_n(p_i) \tag{5.2.1}$$

where n is the number of conditions, and

$$p_i = \frac{k_i}{\sum_{j=1}^{n} k_j} \tag{5.2.2}$$

where k_i is the gene expression level at condition i. We used n as the base of the log in order to keep the value between 0 and 1. Genes with uniform distribution of expression had entropy of 1, while those expressed under only one condition had entropy of 0.

5.3 Results

5.3.1 HCNEs in the *D. melanogaster* genome

Using a minimum average conservation score of 0.95 across 14 insect species that diverged 2.3 to 366 million years ago (Hedges et al., 2006), we identified 1,456 HCNEs $\geq$ 200 bp in the *D. melanogaster* genome. Their level of conservation was greater than that of the protein-coding sequences in this

genome (Figure 5.1). More than half of the HCNEs (56.94%) were intergenic, while the rest were intronic.

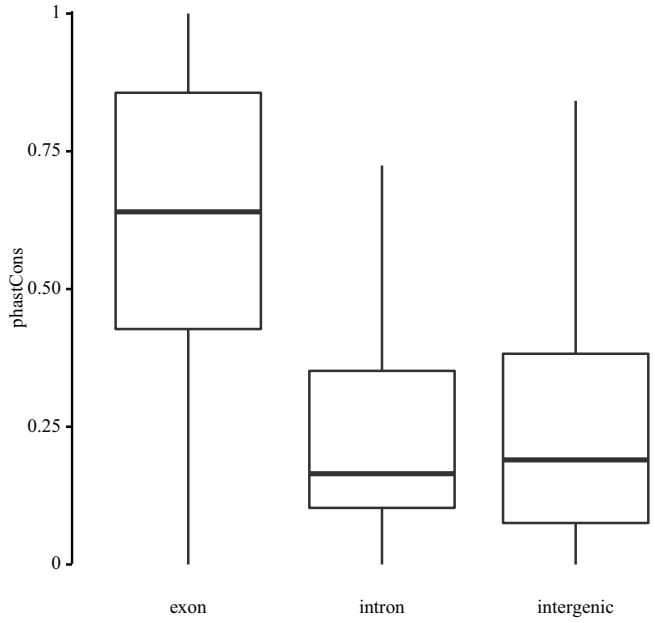

Figure 5.1: PhastCons score distributions for exons, intergenic regions, and introns

The identified HCNEs displayed distinct sequence properties. They had higher GC contents (Figure 5.2A) (P=2.29e-29; P-values were obtained using the Mann-Whitney test throughout this chapter, unless otherwise specified) compared to the random noncoding sequences. Moreover, the frequency of A and T nucleotides was found to drop sharply at the boundaries of HCNEs and increase smoothly in the surrounding regions (Figure 5.2B), in a pattern that is conserved across different lineages (including insects) (Walter et al., 2005; Vavouri et al., 2007). Interestingly, the central regions of the HCNEs were slightly GC-poor, similar to the sequences of short conserved elements (Kenigsberg and Tanay, 2013) (data not shown).

To examine the intrinsic biological roles of the conserved regions, we queried the 50 most over-represented heptamer against FlyReg motifs (Bergman et al., 2005) using STAMP (Mahony and Benos, 2007) and detected many putative TFBSs related to developmental TFs (Figure 5.2C and Table A.1). This result is consistent with previous reports showing that HCNEs harbor binding sites for developmental TFs (Ryu et al., 2012; Visel et al., 2008).

Gene regulatory blocks (GRBs) are broad genomic regions of conserved synteny that harbor dense distributions of HCNE loci, developmental and regulatory genes. GRBs are thought to have emerged due to evolutionary pressure to maintain HCNEs and their target genes in *cis*, both in

vertebrates and in insects (Engstrom et al., 2007; Kikuta et al., 2007). To assess the amount of overlap between the identified HCNEs and GRBs, we determined the GRB boundaries from the coordinates of 1,319 genes that show conserved microsynteny across the *Drosophila* genus (Sahagun and Ranz, 2012) and span regions of dense HCNE loci (von Grotthuss et al., 2010). We found that 51.44% of the identified HCNEs reside inside 110 GRBs, and an additional 7.35% lie within 50 kb of a GRB boundary.

HCNEs located inside GRBs are thought to regulate genes that show conserved microsynteny and are characterized by involvement in "regulation of transcription" and "multicellular-organismal development" (Engstrom et al., 2007). Of them, 95% contain an initiator element (Inr) motif in their core promoter, consisting of Inr only, Inr followed by a downstream promoter element (Inr/DPE), or a TATAbox/Inr (Engstrom et al., 2007). However, 49% of insect GRBs do not contain any gene that satisfies these characteristics, suggesting that the criteria are either insufficient to characterize HCNE targets, or that regulation by HCNEs is not restricted to Inr-motif promoters (Sahagun and Ranz, 2012). Moreover, around 41% of the identified HCNEs were not associated with any GRB. Therefore, for downstream analysis, we included all genes lying within 50 kb from the boundary of a GRB or a HCNE (for HCNEs that were not located in close proximity to a GRB). These are henceforth referred to as "HCNE-proximal" genes. In this study, we focused on genes that had trustworthy promoter-type predictions available; *Drosophila* core promoters are classified into five types: Inr-only, Inr/DPE, TATAbox/Inr, Motif1/6 (as described in (Ohler, 2006)) and DRE core promoters (Engstrom et al., 2007; Ohler, 2006).

We identified 7,291 HCNE-proximal genes (approximately 39% of all annotated genes in *Drosophila* genome), 2,612 of which had reliable predictions available for their core promoter types. The distribution of core promoter types among the HCNE-proximal genes was significantly different from expected (P=3.94e-5 by Chi-square homogeneity test; Figure 5.2D). The HCNE-proximal genes were prominently enriched in genes with Inr-motif promoters and depleted in genes Motif1/6 and DRE promoters.

Consistent with the previously detected differences in Gene Ontology (GO) enrichment between genes of distinct promoter types (Engstrom et al., 2007), the HCNE-proximal genes with Inr-only promoters were predominantly enriched in biological processes related to regulation and development, such as "regulation of transcription, DNA-dependent" and "system development". Many GO terms related to developmental and cell adhesion processes were enriched among genes with Inr-only and Inr/DPE promoters. Meanwhile the genes with Motif1/6 and DRE promoters tended to be involved in general processes such as "metabolic process" and "cellular process" (Figure 5.2E

and Table A.2). Interestingly, we detected some previously unreported differences that may reflect updates in the GO annotations. Most notably, genes with TATAbox/Inr promoter were enriched in terms related to developmental processes, such as "cell fate specification" which was also enriched among genes with Inr-only promoter, and "mesodermal morphogenesis" and "cuticle development", which were not enriched in the genes of other promoter types (Figure 5.2E and Table A.2). Protein domain analysis suggests similar results for genes with Inr only and Inr/DPE promoters: genes with Inr-only promoter were enriched in homeobox protein domains; genes with Inr/DPE promoter were enriched immunoglobulin protein domains. However, genes with TATAbox/Inr promoter were enriched in protein domains of unknown function and no protein domains were enriched among genes with Motif1/6 and DRE promoters (Table A.3).

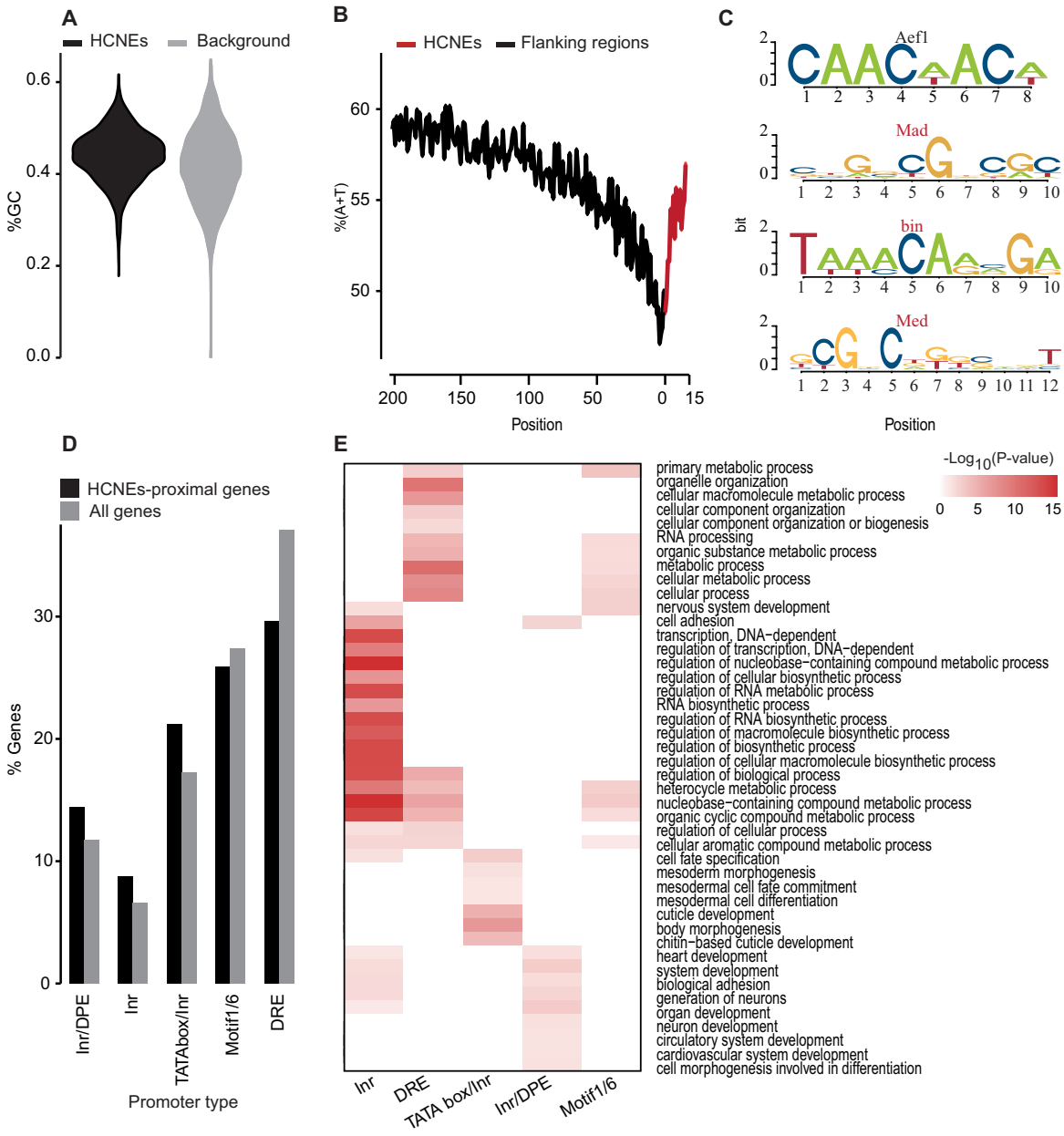

Figure 5.2: **Genomic properties of HCNEs.**(A) Violin plot illustrating that HCNEs have a higher GC content versus 10,000 random sequences. (B) Per-base A+T frequencies within 200 bp of HCNE-flanking regions and 15 bp of HCNEs aligned at their boundaries (region on downstream boundaries were reversed). The faded colors show actual frequencies, while the sharp colors represent smoothed frequencies. (C) Sequence logos of four TFBSs that were among the most significant matches to overrepresented heptamers; red marks TFs reported to be involved in developmental processes; for complete list of TFBSs matches to top 50 overrepresented heptamers refer to Table A.1. (D) Bar plot showing the fraction of core promoter predictions among the HCNE-proximal genes. HCNE-proximal genes are enriched in the Inr-motif and depleted of Motif1/6 and DRE core promoters, to a higher degree than expected by chance. (E) GO enrichment among HCNE-proximal genes grouped by their core promoter type. The top 10 significantly enriched terms (Holm-Bonferroni adjusted P-values < 0.05) are shown for each promoter-type group. Colors represent $-log_{10}$ (Holm-Bonferroni adjusted P-values).

5.3.2 Nucleosome landscape in the proximity of HCNEs

Nucleosome occupancy and positioning is intimately related to the regulation and protection of genetic material (Chen et al., 2012). Nucleosomes occupy coding sequences more highly than intergenic regions (Westenberger et al., 2009). Regions 150 bp upstream of TSSs, which typically harbor many TFBSs, are generally depleted of nucleosomes (Spitz and Furlong, 2012). Enrichment of nucleosomes in a promoter region is negatively correlated with gene expression (Westenberger et al., 2009).

To assess the intrinsic information embedded in HCNEs, we analyzed their nucleosome occupancy in the *D. melanogaster* embryonic S2 cell line (Henikoff et al., 2009). We observed that the nucleosome density was higher in HCNEs compared to their flanking regions (Figure 5.3A) and slightly lower at the center of HCNEs. This pattern was similar to that previously reported for short HCNEs (Kenigsberg and Tanay, 2013). In addition, mononucleosomes were depleted at the centers of HCNEs, enriched at their borders and immediate flanking sequences, and showed smooth decreases along their distal flanking regions (Figure 5.3B).

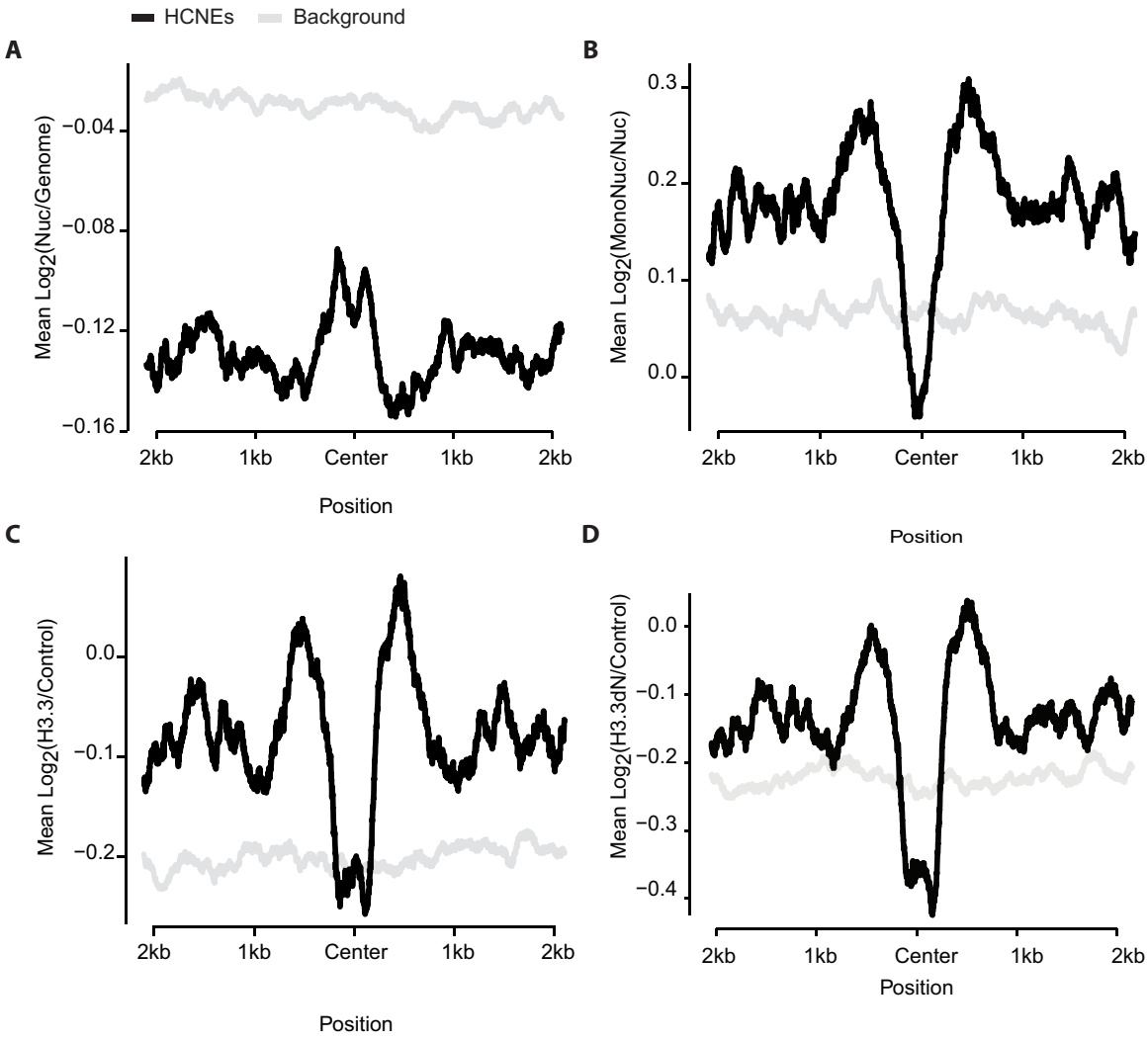

Figure 5.3: **Nucleosome landscape at HCNEs.**(A) Mean nucleosome density of sequences aligned with respect to their centers. Nucleosomes are enriched at the centers of HCNEs compared to the flanking regions. (B) Mononucleosome enrichment was calculated from sequences aligned as described in (A). Mononucleosomes are depleted in HCNEs compared to flanking regions. (C) H3.3 enrichment, calculated from sequences aligned as described in (A). H3.3 is depleted at HCNEs compared to the flanking regions. (D) Same as (C) but for H3.3dN.

H3.3, which is a non-canonical histone that replaces H3.1 during chromatin-disrupting processes, such as transcriptional regulation (Ahmad and Henikoff, 2002; De Koning et al., 2007; Stroud et al., 2012; Wirbelauer et al., 2005), has been shown to be important for fruit fly fertility and mammalian development (Bonnefoy et al., 2007; Bush et al., 2013). Therefore, we investigated the occupancy of H3.3 and H3.3dN (an N-terminal-region-lacking H3.3 that undergoes replication-independent incorporation into chromatin) (De Koning et al., 2007). Similar to the pattern observed for mononucleosomes, H3.3 was depleted at the center of HCNEs, while being enriched at their borders and immediate flanking regions (Figure 5.3C and D).

Taken together, these findings indicate that HCNEs are characterized by a high nucleosome density, a low mononucleosome density, and a low H3.3 density in the S2 cell line. Thus, HCNEs appear to exist in a more compact chromatin environment compared to their flanking regions.

5.3.3 Dynamic regulation of histone modification at HCNEs

Chemical modifications of histones can determine the state of chromatin and regulate gene expression (Bernstein et al., 2006). Since HCNEs are believed to regulate developmental genes, we questioned whether their chromatin state might change during development. We tested six histone modification markers: H3 lysine 4 mono- and tri-methylation (H3K4me1 and H3K4me3), H3 lysine 9 acetylation and tri-methylation (H3K9ac and H3K9me3), and H3 lysine 27 acetylation and tri-methylation (H3K27ac and H3K27me3). In addition, we examined nejire (a CREB-binding protein; CBP) for any association with the regulation of transcription in our system.

H3K9me3 and H3K27me3, which have been associated with the repressed state of chromatin (Negre et al., 2011), were found to cover 19-45% of HCNEs throughout development (Table 5.1). This indicates that many HCNEs maintain a repressive chromatin state during the development of *D. melanogaster*. However, H3K9me3 and H3K27me3 had broad peaks that spanned many kb while the HCNEs covered only small parts of these regions (Figure 5.4). This suggests that the repressed chromatin state is a property of the regions that harbor HCNEs, and does not appear to be specific to HCNEs.

Although the histone modifications normally associated with active chromatin and CBP were relatively depleted among HCNEs (Table 5.1), they demonstrated significant stage-specific patterns. For example, H3K4me1 and H3K9ac were predominantly enriched during the embryonic stages, whereas the number of HCNEs with H3K27ac increased during the second larval stage and CBP was more abundant during later developmental stages (Table 5.1). Unlike the peaks seen for the repressive markers (see above), the peaks of these active markers were narrow, and thus appeared to reflect the activities of HCNEs rather than their surrounding regions (Figure 5.4).

	Embryonic 0-4 hours	**Embryonic 4-8 hours**	**Embryonic 8-12 hours**	**Embryonic 12-16 hours**	**Embryonic 16-20 hours**	**Embryonic 20-24 hours**	**Larva 1**	**Larva 2**	**Larva 3**	**Pupa**	**Adult Male**	**Adult Female**
CBP	0.1	1.9	-	3.8	0	0.1	0.4	-	1.6	5.0	4.5	3.6
H3K4me1	7.0	0.7	9.3	25.3	14.1	2.8	0.3	3.4	0.5	3.2	0.1	1.3
H3K4me3	1.4	0.3	1.9	5.2	4.1	0	0.6	1.3	0	0.6	0.4	
H3K9ac	5.6	2.7	5.8	3.4	5.1	7.6	0	1.3	1.4	0.3	0.3	0.8
H3K9me3	19.7	19.6	23.3	30.2	28.7	26.1	25.5	30.0	-	21.9	-	-
H3K27ac	5.5	3.8	4.7	10.9	6.3	4.4	0.4	23.2	3.8	2.1	0	2.7
H3K27me3	41.8	25.0	29.1	21.9	41.5	40.7	31.7	49.7	31.0	45.4	23.1	-

Table 5.1: **Dynamic histone modification at HCNEs during *Drosophila* development.** Table shows the percentage of HCNEs that were positive for any of the six analyzed histone modification markers or CBP binding across 12 developmental stages. A dash (-) indicates missing data. An HCNE is considered to have a given marker when at least 10% of its length overlapped with the marker. We observed a prominent presence of H3K27me3 and H3K9me3 among the HCNEs throughout development. Markers associated with the active chromatin state (H3K27ac, H3K9ac, and H3K4me3) displayed stage-specific patterns.

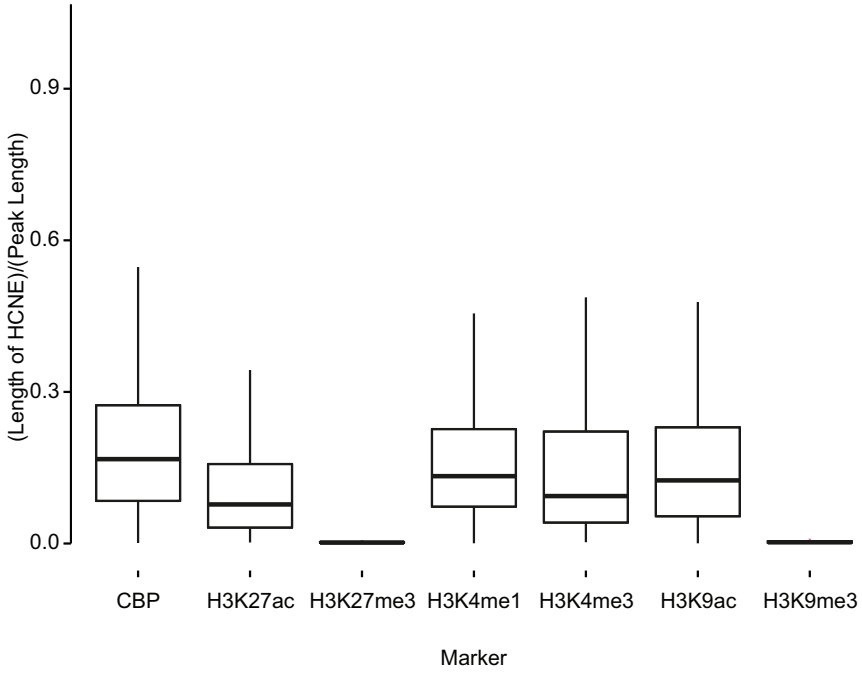

Figure 5.4: Boxplot illustrates HCNE coverage for various histone modifications and CBP peaks

5.3.4 Association between the transcriptional activities of HCNE-proximal genes and histone modification at HCNEs

We next examined the transcriptional activities of HCNE-proximal genes during 30 stages of development and across 28 different tissues. We first assessed their stage and tissue specificities with Shannon entropy (see Methods), and found that the stage and tissue specificities of HCNE-proximal gene expression followed a bimodal distribution (Figure 5.5A and B). This indicates that HCNEs are flanked by both stage/tissue-specific genes and those that are uniformly expressed across different stages and tissues. Moreover, we noted that genes with Inr-motif promoters exhibited higher degrees of stage- and tissue-specific expression compared to those with Motif1/6 and DRE promoters (Figure 5.5A and B, Table A.4). This is consistent with our observations regarding the functional enrichment of genes of distinct promoter types (see above).

We next explored the expression profiles of the HCNE-proximal genes that showed stage/tissue-specific expression based on a stringent cutoff inferred from their bimodal distributions (1-entropy 0.25 and 0.4 for stage and tissue specificity, respectively). Clusters of stage-specific genes with Inr-motif promoters showed high levels of expression across most developmental stages (Figure 5.5C). In contrast, genes with Motif1/6 and DRE promoters were predominantly expressed during later

developmental stages. Interestingly, many of the selected genes with Motif1/6 or DRE promoters were male-specific (Figure 5.5C). Unlike the stage-specific genes, tissue-specific genes of different promoter types showed similar expression profiles across tissues. Interestingly, we observed large clusters of genes with Motif1/6 and DRE promoters expressed at high levels in ovary and testis (Figure 5.6).

Histone modifications around genes are often correlated with their transcriptional regulation. Thus, we studied whether histone modification and/or CBP binding at HCNEs could be associated with the transcription levels of nearby genes. For all promoter types, we grouped the stage-specific HCNE-proximal genes by the presence or absence of active markers (H3K4me1, H3K4me2, H3K9ac, H3K27ac, or CBP) at the nearest HCNE (A marker is considered present at HCNEs if at least 10% of HCNE length overlap with marker peak). Interestingly, genes with Inr only, TATA box/Inr and DRE promoters near active HCNEs showed higher expression levels than those near inactive HCNEs (P=1.27e-11, P=2.10e-2, and P=7.64e-7, respectively; Figure 5.5D). This suggests that HCNEs and their histone modifications could be involved in the transcriptional regulation of adjacent genes.

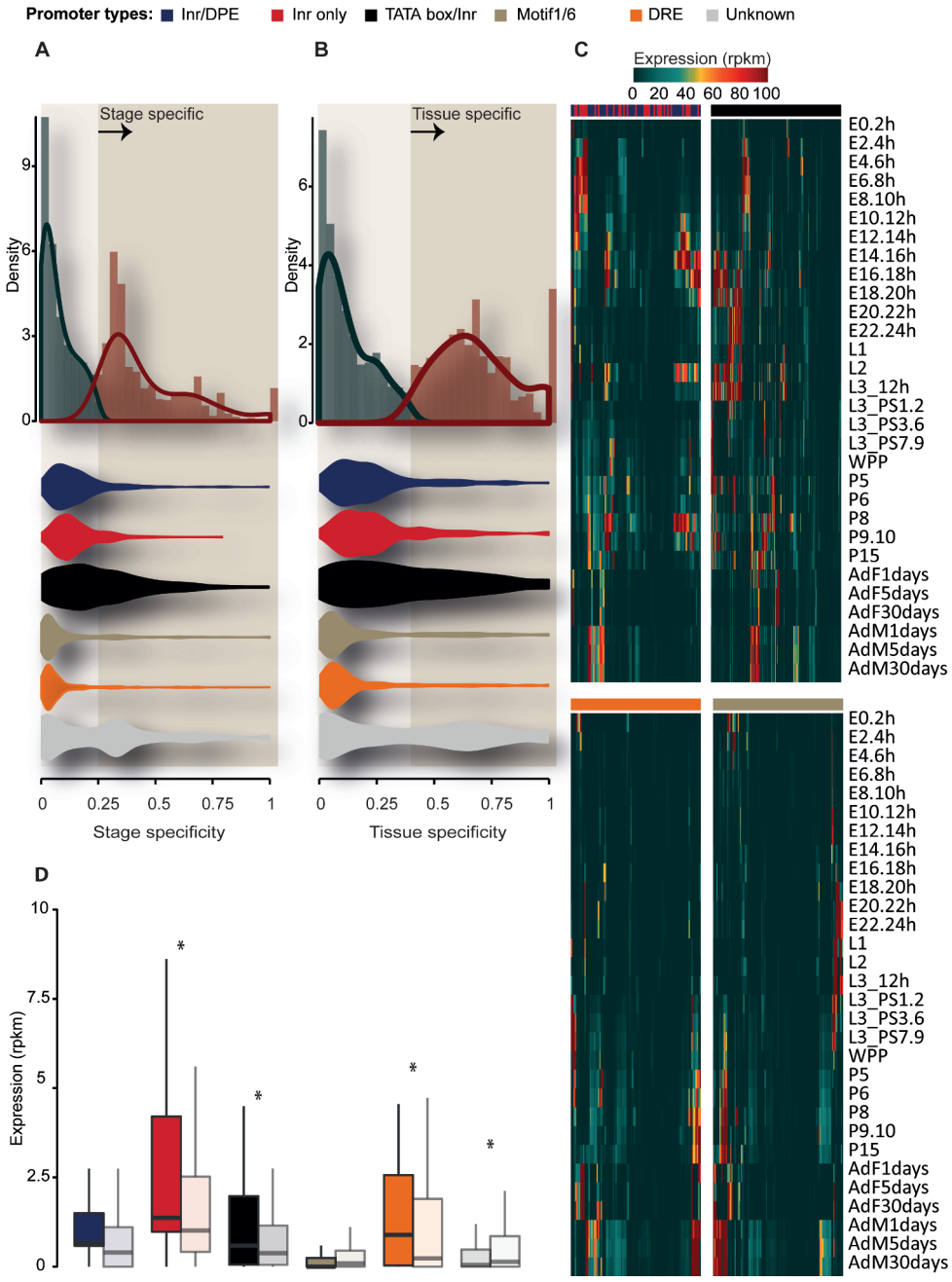

Figure 5.5: **Transcriptional activity of HCNE-proximal genes and their associations with histone modifications of HCNEs.**(A) (upper) Histogram showing the bimodal distribution of stage specificity amongst HCNE-proximal genes, measured as 1-entropy; 0 indicates that genes were expressed evenly across different stages, while 1 indicates that genes were expressed during only one stage. (lower) Violin plots showing the stage specificity of HCNE-proximal genes grouped by their core promoter type. Genes with Inr-motif promoters are more stage-specific than genes of the other core promoter types. (B) Same as in (A), but assessing tissue specificity. (C) Heatmaps illustrating the expression levels of stage-specific HCNE-proximal genes across 30 developmental stages (from FlyBase); E, L, P, AdF and AdM refer to Embryonic, Larva, Pupa, Adult Female and Adult Male stages. Genes were grouped by their promoter type (Color key for promoter type is shown on the top of each Heatmap; Inr only and Inr/DRE are grouped together for visualization purposes), and expression values greater than 100 were rounded to 100. Complete linkage hierarchical clustering is performed with Euclidean distance as the distance metric. Clusters of genes with Inr-motif promoters exhibit high levels of expression throughout development, whereas the genes having other promoter types are predominantly expressed during the later stages. (D) Boxplot showing differences in the expression levels of HCNE-proximal genes grouped by the presence (sharp color) or absence (faded color) of active markers at the nearest HCNEs. Expression levels were examined for all 12 developmental stages (from modEncode). Symbol "*" indicates < 0.05.

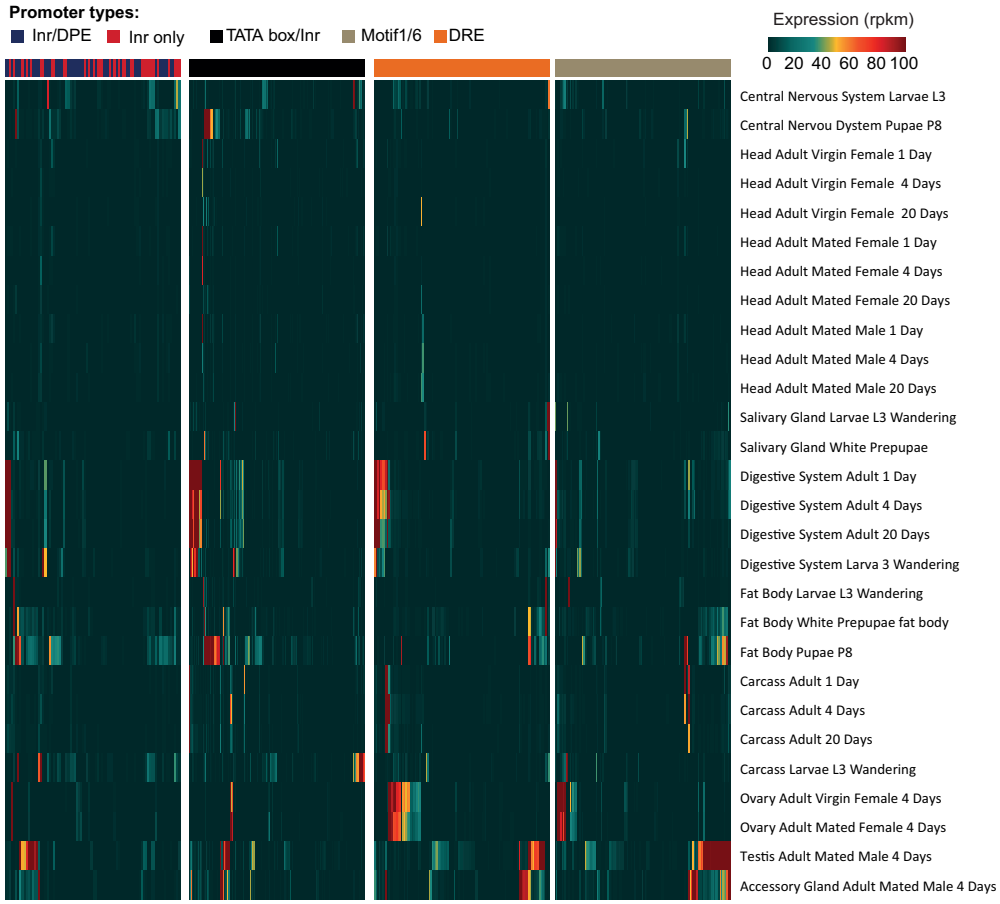

Figure 5.6: Heatmaps illustrating the expression levels of tissue-specific HCNE proximal genes across 28 different tissues

5.3.5 Some HCNEs are initiated early in replication

DNA replication is a highly accurate process that ensures the correct transmission of genetic information to daughter cells. The location and temporal order of replication is conserved in several yeast species (Muller and Nieduszynski, 2012), but replication origins do not seem to be strongly conserved among higher eukaryotes (Kritsas et al., 2012). Here, we examined the replication timing of HCNEs in three *D. melanogaster*-derived cell lines: ML-DmBG3-c2 (Bg3), Kc-167 (Kc), and S2-DRSC (S2) cells. We found that HCNEs fire later during replication compared to other genomic loci in Bg3 (P=8.9e-9), Kc (P=1.4e-9), and S2 (P=5.45e-21) cells (Figure 5.7A). This result is consistent with the reported late replication of repressed and Polycomb-associated heterochromatic regions enriched

in HCNEs (Filion et al., 2010). However, we found a subset of HCNEs that serve as early replication origins, at percentages higher than those expected by chance: 10.09% (P=7.7e-9), 11.20% (P=1.4e-3) and 7.42% (P=0.02) for Bg3, Kc and S2 cells, respectively (P-values computed by Fisher's exact test). Only 26.03% of these early-replication HCNEs were common to all three cell lines (Figure 5.7B), indicating that the activities of HCNEs as early replication origins are cell-line-specific.

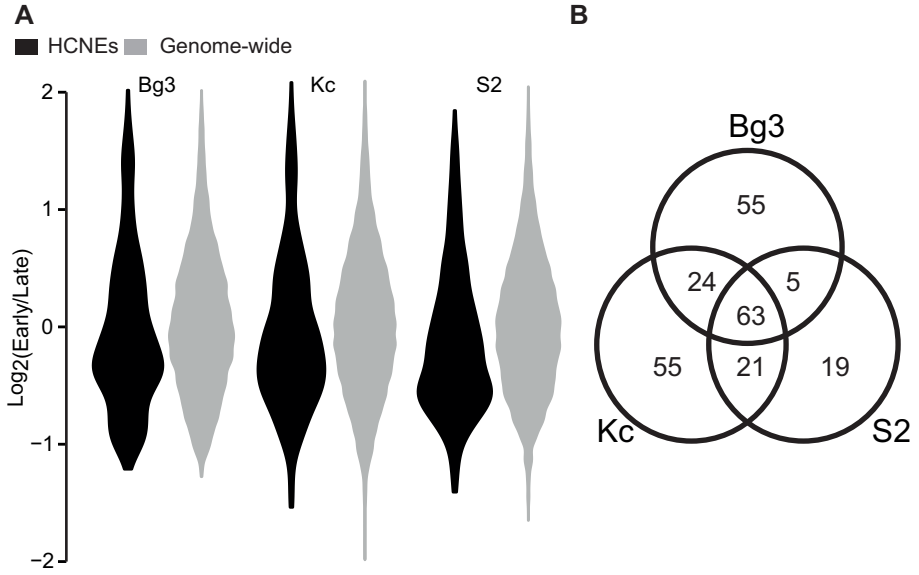

Figure 5.7: **Some HCNEs fire early in replication.**(A) Violin plot displaying the distributions of replication timing at HCNEs and other genomic regions in three cell lines. HCNEs fire later in replication compared to all genomic regions across all cell lines. (B) Venn diagram demonstrating the number of HCNEs within the early replication origin peaks identified in the three cell lines. Approximately 10% of HCNEs are associated with early replication origins; of them, only 26% are common to all three cell-lines, indicating that some HCNEs undergo cell-specific initiation of early replication.

5.3.6 Association of HCNEs with nuclear structures

The nuclear lamina, which is an important part of the nuclear structure, functions in important cellular processes of metazoan cells, including chromatin organization and DNA replication (Prokocimer et al., 2009). Several studies have identified lamina-associated domains (LADs) as genomic elements that are capable of mediating the association between the genome and the structural framework of the nucleus (Guelen et al., 2008; van Bemmel et al., 2010). We questioned whether the identified HCNEs could be associated with LADs. We downloaded the positions of LADs in the *D. melanogaster* genome (van Bemmel et al., 2010) and converted the coordinates to those of the current genome release using the FlyBase conversion tool (van Bemmel et al., 2010). Consistent with the previous report that lamin protein binding is enhanced along HCNE-enriched repressed chromatin (Filion et al., 2010), we found that 872 HCNEs were located within LADs (P=3.39e-32 by Fisher's exact

test). Our results therefore suggest that HCNEs can associate with structural components of the nucleus, potentially contributing to their evolutionary selection.

5.3.7 Correlations among distinct properties of HCNEs

To gain new insights into the overall properties of HCNEs, we performed cluster analysis (complete-linkage hierarchical clustering) using the identified HCNEs and the examined genomic and epigenomic features (Figure 5.8). To reduce the complexity of this analysis, we summed the number of histone modifications and CBP bindings across the various developmental stages.

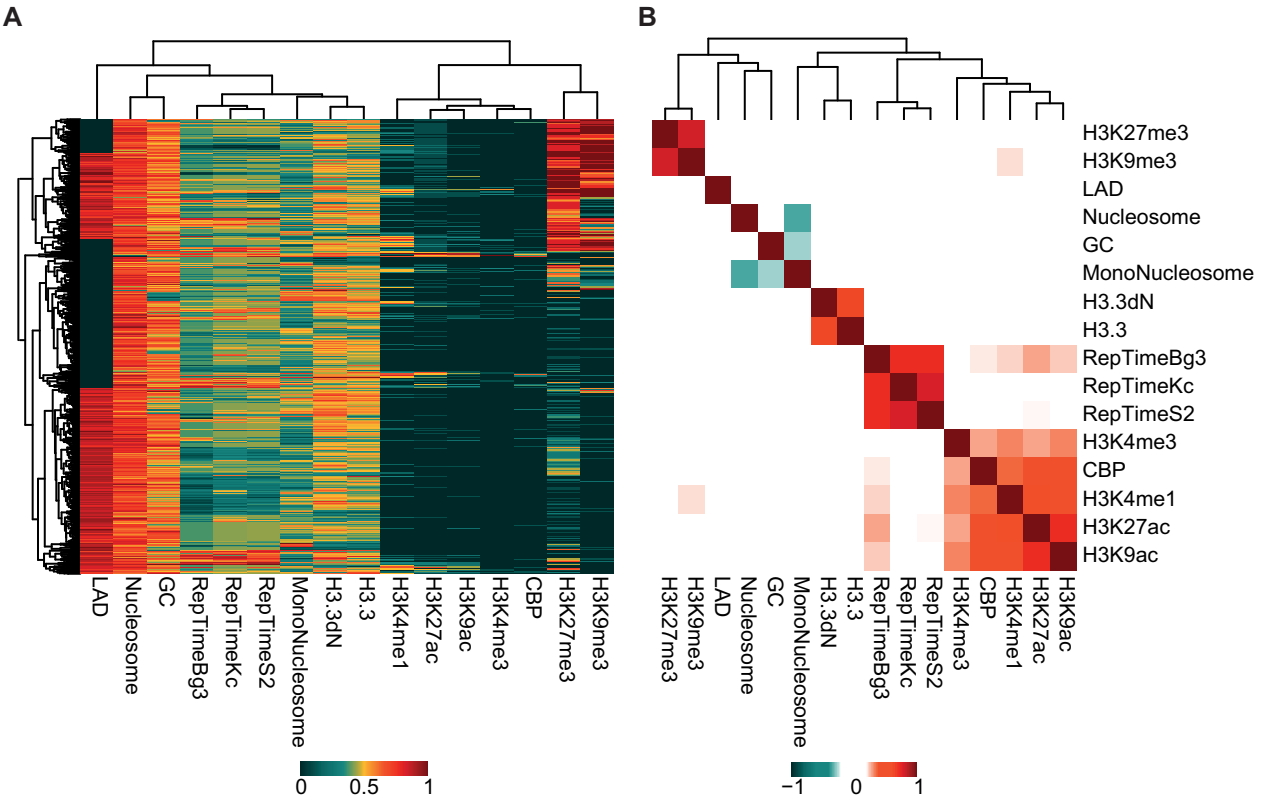

Figure 5.8: **Genomic and epigenomic properties of HCNEs** Heatmap demonstrating the clustering of HCNEs based on the studied features (see below). (B) Heatmap showing correlation between features. The studied features include: the levels of nucleosomes, mononucleosomes, H3.3 and H3.3dN; the summed occurrence of each histone modification and the CBP binding for each HCNE during development; the replication timing in the three studied cell lines (columns labeled as RepTimeBg3, RepTimeS2 and RepTimeKc); and the LAD scores. To facilitate visualization, the values of each feature were scaled to between 0 and 1 using the equation $\frac{value-min}{max-min}$, where max and min were the maximum and minimum values of each feature, respectively; complete linkage hierarchical clustering is performed with Euclidean distance as the distance metric.

A number of HCNE clusters were observed. The members of the most prominent cluster were tri-methylated on H3 during development (H3K9me3 and H3K27me3; dark red cells in the bottom

and middle clusters of Figures 5.8A). As these markers have been associated with transcriptional repression (Negre et al., 2011), our results may indicate that these HCNEs are silenced throughout most of development. Although the activation markers (H3K4me1, H3K4me3, H3K9ac, H3K27ac, and CBP) (Negre et al., 2011) were generally depleted on HCNEs, we observed a few small clusters of HCNEs that exhibited various combinations of activation modifications throughout development. This suggests that although HCNEs were typically maintained in a repressed state in the studied cell lines and during most developmental stages, they may be activated in specific cell types or for short periods during development. Similar patterns have been observed among some developmental genes (Dolecki et al., 1988).

Other HCNEs were found to be associated with early replication origins (Figures 5.8A, dark red cells in columns 4, 5 and 6). This suggests that replication may be tightly connected with the mechanisms underlying the genomic conservation of some HCNEs.

Our correlation analysis further revealed the following relationships at HCNE (Figures 5.8B): a negative correlation between mononucleosome levels and GC content/nucleosome occupancy; a positive correlation between H3.3 and H3.3dN; a positive correlation between the repression markers, H3K9me3 and H3K27me3; positive correlations among the activation markers, CBP, H3K9ac, H3K4me1 and H3K27ac; and positive correlations among replication origins. Weak positive correlations were also observed between replication markers and activation markers (Figure 5.8B). These correlations between the different features indicate that they all act on roughly the same subsets of HCNEs. However, this held true only between features from a given cell line or developmental stage. HCNEs appeared to be mostly repressed in S2 cells, while being active during some of the studied developmental stages of the fly. Thus, nucleosome occupancy and other chromatin features from the S2 cell line do not appear to correlate with the histone modification marker data from the various developmental stages.

5.4 Discussion

5.4.1 Local trends in GC and nucleosome density provide insights into DNA conservation

In HCNEs, elevated GC levels and associations with developmental genes appear to be universal beyond the kingdom level (Kritsas et al., 2012; Kenigsberg and Tanay, 2013). Recent studies have suggested that compensatory evolution may exist, as divergent sequences have been found to main-

tain their GC contents and nucleosomal organizations in yeast (Kenigsberg et al., 2010). Short conserved elements (30-70 bp) in *D. melanogaster* also showed similar patterns (Kenigsberg and Tanay, 2013). Local GC contents have been found to exert strong effects on the flexible organization of nucleosome spacing, with AT-rich sequences serving as repelling elements and forcing the nucleosome to position on GC-rich areas that favor nucleosome binding. Nucleosomes suppress C to T, G to T, and A to T mutations by about 2-fold in yeast by reducing the exposure of naked DNA (Chen et al., 2012). These observations are consistent with our findings that HCNEs are GC- and nucleosome-rich, and are demarcated by AT- and mononucleosome-rich sequences (Figures 5.2 and 5.3). The nucleosome data used in this study were obtained from S2 cells, which originated from a late embryonic macrophage-like lineage (Schneider, 1972). Our results suggest that nucleosomes remain dormant on HCNEs, at least in S2 cells. These results may be cell-specific, however, as nucleosome occupancy can differ in various situations, including during development (Teif et al., 2012).

5.4.2 Histone modifications suggest that HCNE may play a regulatory role

Chromatin is not a static entity. Chemical and thermal fluctuations around chromatin can denature the DNA (i.e., the so-called "DNA breathing") (Chen et al., 2012), and dynamic changes can occur via histone modifications. Most of the previous studies of epigenetic regulation have focused on genes and promoter regions with well-defined locations and properties. The epigenetic regulation mechanisms of other locations, such as distal enhancers, are not yet well understood because it is difficult to identify such elements and their long-distance relationships with target genes. Here, we examined possible epigenetic control mechanisms at HCNEs, and identified their dominant histone modifications, which included H3K9me3 and H3K27me3 (Table 5.1). Histone modification is often associated with the transcriptional activation or repression of nearby target genes. Consistent with this, many HCNE-proximal genes, most notably genes with Inr-motif and DRE promoters, showed higher expression levels when their associated HCNEs were in the active chromatin state (Figure 5.5D). This suggests that the chromatin modification of HCNEs may modulate the transcription of those genes, many of which are developmental regulators. *Cis*-regulation can control distant genes, but it is generally difficult to discriminate between target and bystander genes due to the lack of comprehensive transcriptomic and epigenetic data (Akalin et al., 2009) and the complications caused by high gene density, such as that found in the *D. melanogaster* genome.

5.4.3 Replication timing at HCNEs

The orchestrated and properly timed initiation of replication from multiple origins during cell division is essential for the vertical transfer of genetic materials. Thus, most species maintain common genomic and epigenomic features at their replication origins. Replication origins are GC-rich, but their nascent strands are AT-rich, allowing the DNA to open easily (Cayrou et al., 2011). In addition, they are lineage-specific, and do not appear to be related to the conservation level of DNA in higher eukaryotes (Kritsas et al., 2012; Rowntree and Lee, 2006). Replication origins fire at different times, and early replication origins tend to allow for fewer mutations in cancer cells (Lawrence et al., 2013). Our results indicate that most HCNEs initiate late in replication. However, about 10% of HCNEs co-localize with early replication start sites, at least in *D. melanogaster*, and thus should be under strong negative selection. Future work is needed to identify the molecular mechanism(s) through which the origin-recognition complex recognizes these specific sequences.

5.4.4 Crosstalk between HCNEs and the nuclear architecture

Previous studies have shown that HCNEs, including ultraconserved elements, are not mutational cold spots (Kritsas et al., 2012; Katzman et al., 2007; Sakuraba et al., 2008), suggesting that they are likely to have intrinsically important functions and be under strong selection pressures. However, it is unlikely that every base of a $\geq$ 200 bp HCNE plays a regulatory role, suggesting other vital functions are likely to be involved in the negative selection of these elements. Studies have shown that genomic elements and nuclear structures frequently undergo crosstalk and dynamic regulation. The nuclear scaffold/matrix provides both a mechanical anchor and distinct territories for genomic elements and proteins during various processes, such as replication and transcription (Linnemann et al., 2009). Although we found some overlap between HCNEs and S/MARs, the amount of overlap was not statistically significant (data not shown). This could reflect the generally poor identification of S/MARs in insects; only a very small number of S/MARs have been experimentally verified, and the computational prediction of such regions has been far from accurate due to their sequence diversity (Frisch et al., 2002). However, overlap between HCNEs and S/MARs has been reported in other lineages (Glazko et al., 2003; Tetko et al., 2006), and these elements have some common characteristics (e.g., associations with developmental genes), seeming to indicate that they may interact. LADs have been relatively well characterized in the *D. melanogaster* genome (van Bemmel et al., 2010), which is $\sim$ 40% covered by LADs of varying size (7 $\sim$ 700 kb). LADs are closely related to S/MARs and significantly overlap with S/MARs in human cells (Linnemann et al., 2009). Lamin

B1, which is the primary component of the nuclear lamina, binds to matrix-attachment regions (Luderus et al., 1992). Lamin also binds histones and is involved in chromatin remodeling, DNA replication, apoptosis, and early development (Prokocimer et al., 2009). Thus, lamin could arguably link the genome to nuclear structures. We observed a striking overlap between LADs and HCNEs indicating possible intimate relationship between the nuclear lamina and mechanisms of genomic conservation. The LAD data used in the present study were obtained from *Drosophila* Kc cells (van Bemmel et al., 2010). The composition of the nuclear lamina changes during development; for example, lamin B1 predominates in the early chicken embryo and decreases thereafter, whereas lamin A shows the opposite expression pattern (Lehner et al., 1987). In the future, it will be interesting to examine the interaction between conserved elements and various lamin proteins during development and in different cell types. Lamins are exclusive to metazoan cells, and are not detected in plant cells (reviewed in (Prokocimer et al., 2009; Meier, 2007)). Thus, although HCNEs have similar properties in animals and plants, it is likely that different nuclear proteins may be involved in their structural associations with the nuclear matrix in these two systems. Because HCNEs do not represent mutational cold spots (Kritsas et al., 2012; Katzman et al., 2007; Sakuraba et al., 2008), any mutations in these elements must be repaired. Very little is known regarding the binding preferences and recruitment mechanisms of DNA-repair proteins. A few lines of evidence have indicated that DNA repair-related proteins show weak sequence preferences in other species (Tracy et al., 1997; Andersen et al., 1985). However, additional detailed molecular studies will be required to assess the repair mechanisms that may be responsible for suppressing mutations in HCNEs. We also need future studies of nucleosome territories (Cremer and Cremer, 2001) to understand how these repair proteins gain easy access to the mutated region. Multifunctional and highly conserved chromatin-related proteins could be considered as candidate regulators for this mechanism. One such protein is Heterochromatin Protein 1 (HP1), which binds to the nuclear envelope, histones, replication origins, and DNA-damage-response proteins (Cayrou et al., 2011; Lachner et al., 2001; Prasanth et al., 2010; Dinant and Luijsterburg, 2009).

Chapter 6

Epigenetic Dysregulation of Human Myogenesis Affects Time Regulated eRNA and Associated Transposable Element Expression

This chapter will be submitted for publication.

- Loqmane Seridi, Yanal Ghosheh, Beatrice Bodega, Gregorio Alanis-lobato, Timothy Ravasi and Valerio Orlando. Profiling Enhancer RNAs during Myogenesis. .

6.1 Introduction

Transcriptional regulation is a complex process that involves the interaction of transcription factors, promoters, enhancers, noncoding RNAs, transposable elements and chromatin states. To understand the transcriptional regulome, spatiotemporal measurements of its components is necessary. Myogenesis is a favorable model system to study transcriptional regulation because factors driving the process are well known and evolutionary conserved (Buckingham and Vincent, 2009; Cesana et al., 2011). However, most time-course studies of myogenesis are limited to few time points and cell lines (Giordani and Puri, 2013). Here, using RNASeq and CAGE, we deep sequenced a high-resolution time-course of myogenesis transcriptome from human primary cells of healthy donors and donors affected by Duchenne muscular dystrophy an X-linked disease that causes muscle degener-

ation. We compiled a catalog of coding and noncoding RNAs, promoters, enhancers, and active transposable elements that are activated in a time-regulated manner during cell differentiation of cultured myoblasts to myotubes. Comparative analysis of of healthy with DMD samples revealed a global dysregulation of coding and noncoding genes, enhancers, and transposable elements possibly due to the epigenetic defect in HDAC2 pathway characteristic of the disease (Colussi et al., 2008). Finally, our analysis revealed high correlation between enhancers and transposable elements activities.

6.2 Methods

6.2.1 Relationship between donors

To explore the differences between the six donors. First, we computed the complexity-invariant distances (Batista et al., 2014) between the expression profile of each transcript in one donor against its profile in another. Then, a distance matrix, between all donors, was constructed by averaging the complexity-invariant distances of all transcripts for every pair of comparison. Last, we projected the matrix onto a two-dimensional space using multidimensional scaling (MDS). Appendix B.1c shows that healthy donors 1 cluster with healthy donor 2 and DMD donors 1 cluster DMD donor 3; healthy donor 3 and DMD donor 2 are outliers of their corresponding groups. We confirmed this observation by looking at the expression profile of genes key to myogenesis: MYH2, MYH3, MYH8, MYOD1, and MYOG (Appendix B.1d).

6.2.2 Differential expression, clustering and functional annotation

We conducted DEG analysis on RNASeq data using edgeR v3.6.8 (Robinson et al., 2010); we used data from healthy donors 1 and 2 and DMD donors 1 and 3 as replicates and qvalue cutoff of 0.05. To cluster dynamic genes and eRNAs. First, we averaged gene/eRNA expression over used donors. Second, we scale expression across time (z-score). Last, we clustered scaled gene/eRNA expression using kmeans. Functional analysis was conducted using GOstats v1.7.4 (Falcon and Gentleman, 2007) for genes and GREAT (McLean et al., 2010) for genomic loci eRNAs.

6.2.3 Relationship between time points and phenotypes

To understand how myogenesis progresses through time, and how it differs between the two phenotypes. First we define a condition as a combination of time and phenotype, e.g., healthy time 0,

DMD time 0, healthy time 1, etc. Second, using RNASeq data of healthy donors 1 and 2, we determined genes implicated in myogenesis: genes that were differentially expressed between any two time points during differentiation (qval$<$ 0.05) using edgeR v3.6.8 (Robinson et al., 2010). Third, based on the expression values of these genes (mean expression of consistent donors; done using RNASeq and CAGE independently), we computed information variation (IV) distance between all possible pairs of conditions; $IV(X,Y) = H(X) + H(Y) - 2\ I(X,Y)$, where H is entropy and I is mutual information that were computed using "infotheo" R package (Meyer, 2014). Fourth, we constructed minimum spanning tree (MST) over the IV matrix and calculated distances between conditions over MST. Last, we projected the MST distances on two-dimensional space using MDS.

6.2.4 Identification of putative active enhancers

We identified active enhancers during myogenesis as described by (Andersson et al., 2014). First, we filtering: consensus tag clusters (CTCs) within 500bp from TSSs of annotated genes or novel transcriptome assemblies, CTCs less than 200bp from exon boundaries, and CTCs overlapping CTCs from opposite strands. Second, we identified the centers of divergent CTCs that are at most 400bp apart. We measured bi-directionality score around those centers as $||FW - RV||/(FW + RV)$, where FW is expression of forward transcript (downstream the center), and RV expression of reverse transcript (upstream the center). Loci with a bi-directionality score of at most 0.8 in at least one sample were taken as putative active enhancers

6.2.5 Association to repeat elements

We downloaded the repeat element annotation of repeatmasker (Benson, 1999) from HOMER (Heinz et al., 2010); the annotation file includes, for every repeat element, its divergence score from, and coverage of, its corresponding consensus full length. We used BEDTools (Quinlan and Hall, 2010) to determine whether eRNA overlap a repeat element. eRNA is an eRNA-overlapping repeat when its center is within repeat body. The distance, however, was measured from eRNA boundary (center +200bp).

6.3 Results and discussion

6.3.1 Sequencing time-course transcriptome of myogenesis

Human primary myoblasts from three healthy donors (two females and one male) and threeDMD patient donors (all male) (Telethon Biobank Besta Institute Milan, Italy) were cultured and differentiated to myotubes *in vitro*; as expected cells from DMD donors differentiated at a slower rate and generated weaker myotube density (Figure 6.1a). We determined the differentiation time-course transcriptome using RNASeq and CAGE (nine time points over twelve days of differentiation). The sequencing produced over 2.5 billion reads, of which $> 87\%$ mapped to the genome (Figure 6.1a). $29,794$ genes (from RefSeq annotations and novel assemblies) were expressed based on RNASeq, 72% of them were confirmed by CAGE (Figure 6.1b). Gene expression measured by RNASeq and CAGE were highly correlated (minimum Pearson > 0.8; Figure 6.1c and Appendix B.1 a-b). We excluded data of two donors (one healthy and one DMD) form analysis specific to myogenesis because they showed inconsistent gene expression profiles with other donors of the same phenotype (Appendix B.1c-d). Throughout the manuscript, only consistent donors were used unless otherwise mentioned.

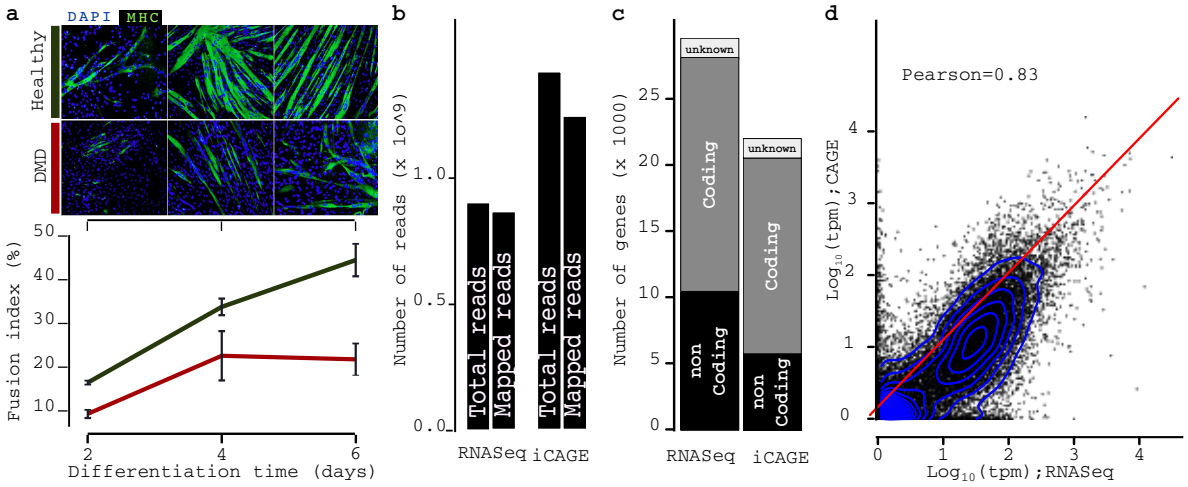

Figure 6.1: **a)** (top) shows fluorescence microscopy images taken from 3 time points of differentiation process of health donor and DMD donor (bottom) line plot showing amount of cells that fused at each time point (error bar +/- std); as expected healthy donors normally differentiated whereas DMD donors differentiated at much slower rate. **b)** barplot showing the number of total reads (after removing bad quality reads) and the number of mapped reads to the genome from 54 libraries. **c)** barplot shows the number of expressed transcripts during differentiation using RNASeq and CAGE. **d)** scatterplot shows the amount of correlation between sequencing technologies used; a high correlation between RNASeq and CAGE.

6.3.2 Healthy and DMD differentiation diverge at day 2

To understand how the differentiation progresses throughout time and across phenotype, we first identified 3,830 dynamic genes implicated in myogenesis (genes differentially expressed between any two time points from healthy donors; $qvalue < 0.05$): 3,717 protein-coding genes (392 TFs) and 113 noncoding genes. Using these genes, we computed the mutual information distance between all time points and projected them by multidimensional scaling (MDS). The projection unveiled ordered trajectories of differentiation for both phenotypes. Although the differentiation of both phenotypes started and progressed similarly, they diverged after day two, henceforth denoted as divergence pivot; this divergence increased rapidly over time (Figure 6.2a and Appendix 6.3a).

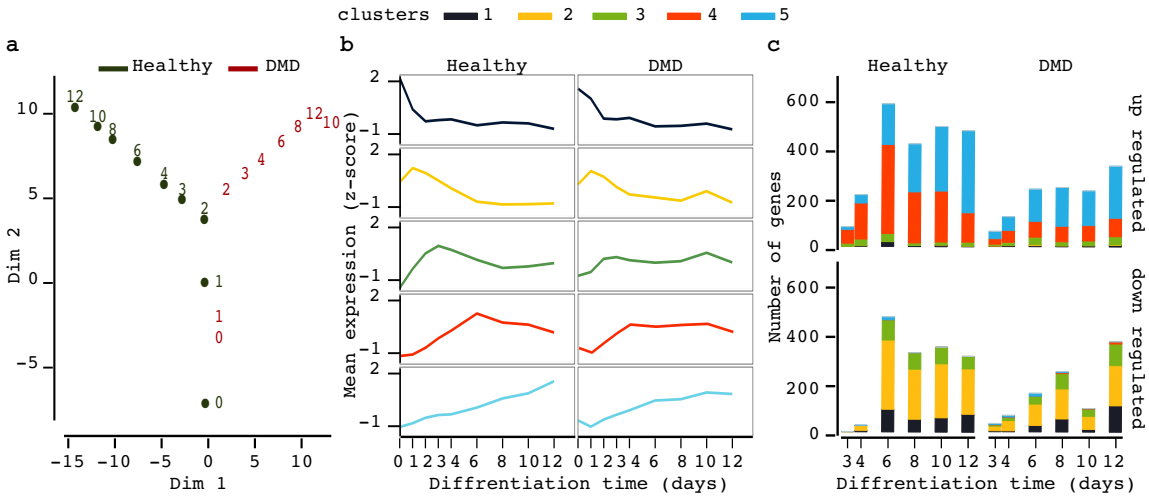

Figure 6.2: **a)** scatterplot shows MDS projection of RNASeq time-course data; projection unveiled bifurcation of differentiation programs **b)**lineplot demonstrates that dynamic genes group into 5 expression profiles; the profiles are inconsistent between healthy and DMD donors especially after day 2. **c)** barplot showing number of differentially expressed genes compared to healthy day 2; clear difference between number of differently up/down-regulated genes in healthy and DMD— consistent with observation from (a) and (b)

The dynamic genes clustered into five waves based on their expression profile in healthy donors (Figure 6.2a and Appendix 6.3b). The clusters were enriched in biological processes consistent with the current knowledge about myogenesis (Bentzinger et al., 2012) (Appendix 6.3c): cluster 1 is enriched in genes associated with proliferation; clusters 2 and 3, are enriched in genes involved in adhesion and fusion stages; cluster 4 is enriched in genes involved in early muscle structure development; cluster 5 is enriched in genes involved in development. Except for cluster1, the clusters exhibit different expression profiles between healthy and DMD donors— most notably after day two (Figure 6.2b-c). This result is consistent with our earlier projection results and suggests that proliferation stage ended similarly for both healthy and DMD donors.

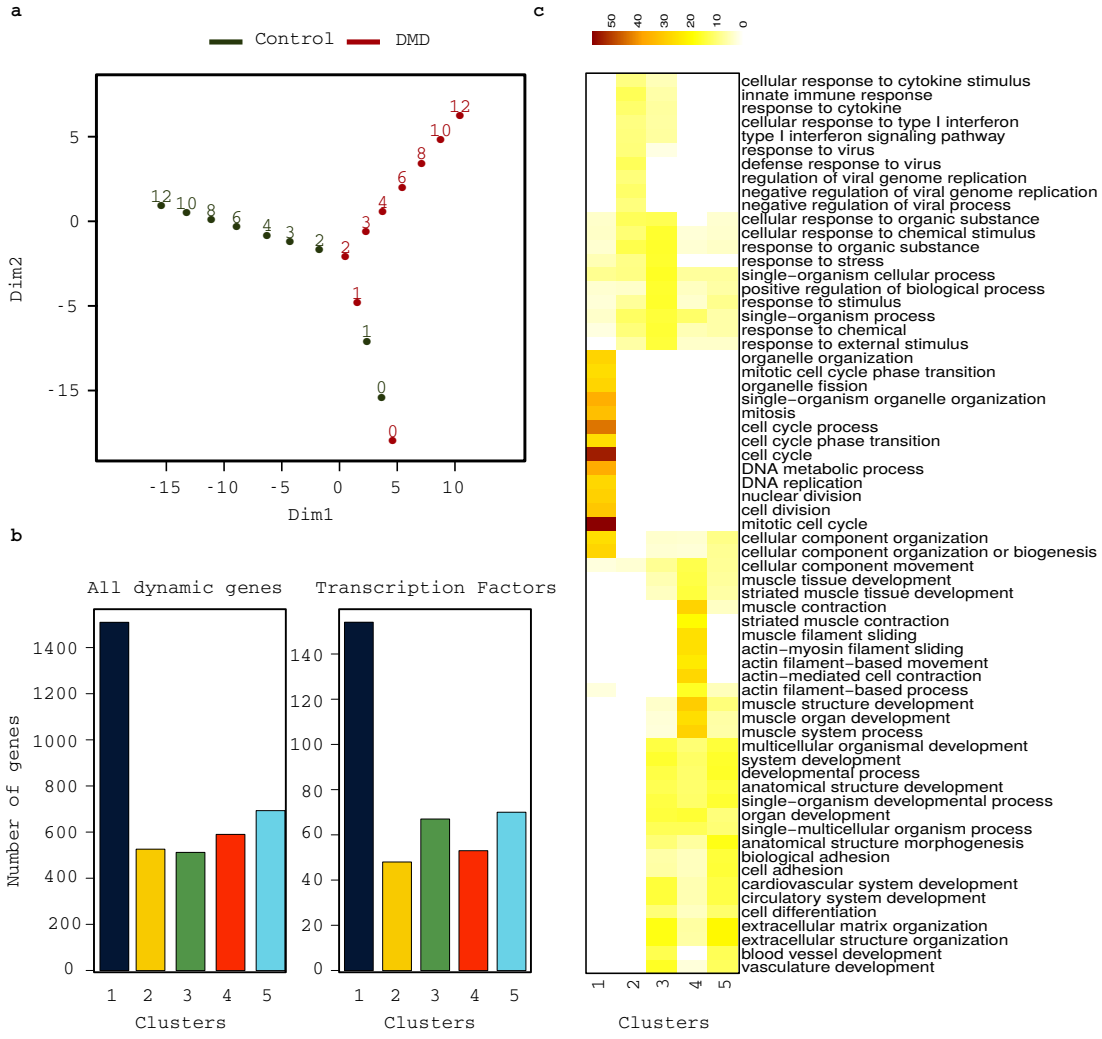

Figure 6.3: **a)** scatterplot shows MDS projection of CAGE time-course data; a bifurcation of differentiation program between cells from health and DMD donors at day 2. **b)** barplot shows the number of genes in each cluster (left) and transcription factors (right). **c)** heatmap displays the p-value of top 10 GO in enriched in every cluster. Every cluster associates with a distinct biological process.

To find the genes possibly initiating the divergence between the two differentiation trajectories, we retrieved differentially expressed genes in the day immediately following the divergence pivot (Figure 6.1c). Many key regulators of myogenesis exclusively upregulated in healthy cells such as MYOG, ID3, MEF2C, MYH2, NOTCH3, and SMYD1. Furthermore, we observed in DMD a down-regulation of the interferon pathway and noncoding RNA MALAT1.

6.3.3 Profiling active eRNAs during myogenesis

eRNA control expression of genes crucial to specific cellular conditions such as development, differentiation and response to stress (Lam et al., 2014). However, their spatiotemporal dynamics during

differentiation are not well understood (Arner et al., 2015). Putative enhancers are marked by bidirectional transcription that is detectable by CAGE (Lam et al., 2014; Andersson et al., 2014). Using 12,708,419 CTSS from all CAGE libraries, we identified 4,132 putative active enhancers which are: bi-directionally transcribed (Figure 6.4a); distant from TSS of known genes (RefSeq) and novel assemblies (Figure 6.4b); 35% intergenic; and 36% found in FANTOMs enhancer database (Andersson et al., 2014). These putative enhancers exhibit strong occupancy of chromatin-associated markers of enhancers (H3K4me1, H3K27ac, and DHS) in myoblast (HSMM) and myotubes (HSMMt) cell lines (Figure 6.4c-e and Appendix 6.5a-b); ChIP-Seq data from ENCODE (Dunham et al., 2012). Using MEME (Bailey and Elkan, 1994), de novo motif finder, we detected many motifs of general transcription factors enriched in eRNA sequences (Appendix 6.6).

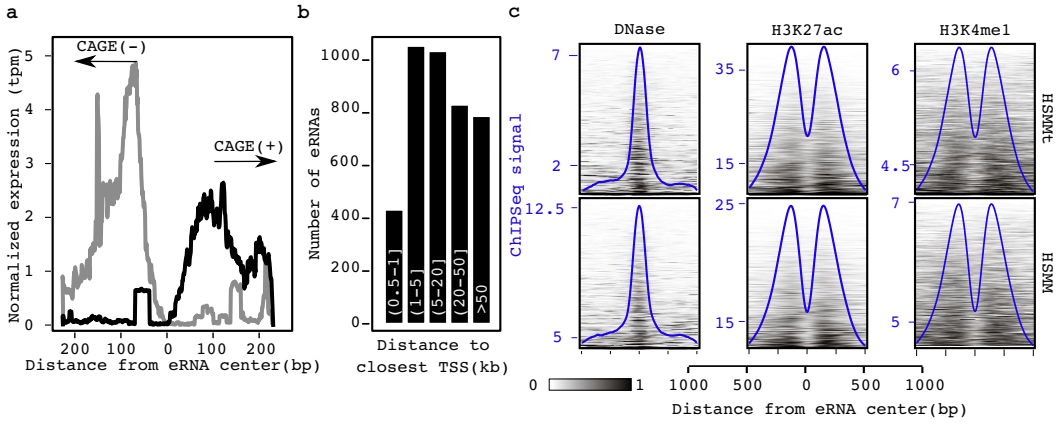

Figure 6.4: **a)** mean CAGE signal shows bidirectional transcription of identified eRNAs; the signal is an aggregate of all samples. **b)** histogram shows that more than 89% of eRNAs are at least 1kb from closest TSS. **c)**heatmaps (rows are eRNAs and columns are genomic position relative to eRNA centers) and mean signal (blue) showing that eRNAs have characteristic features of regulatory loci: DHS and depletion of H3K4me1 and H3K27ac at eRNA centers compare to flanking regions; for visualisation purpose, the signal of heatmap was scaled by setting outliers values (values $\geq 98^{th}$ percentile) to the value of 98^{th} percentile. All values were further scaled to the [0, 1] range by dividing them by the maximum. The signal of DHS, H3K4me1 and H3K27ac on muscle cell lines HSMMt and HSMM were obtained from ENCODE repository.

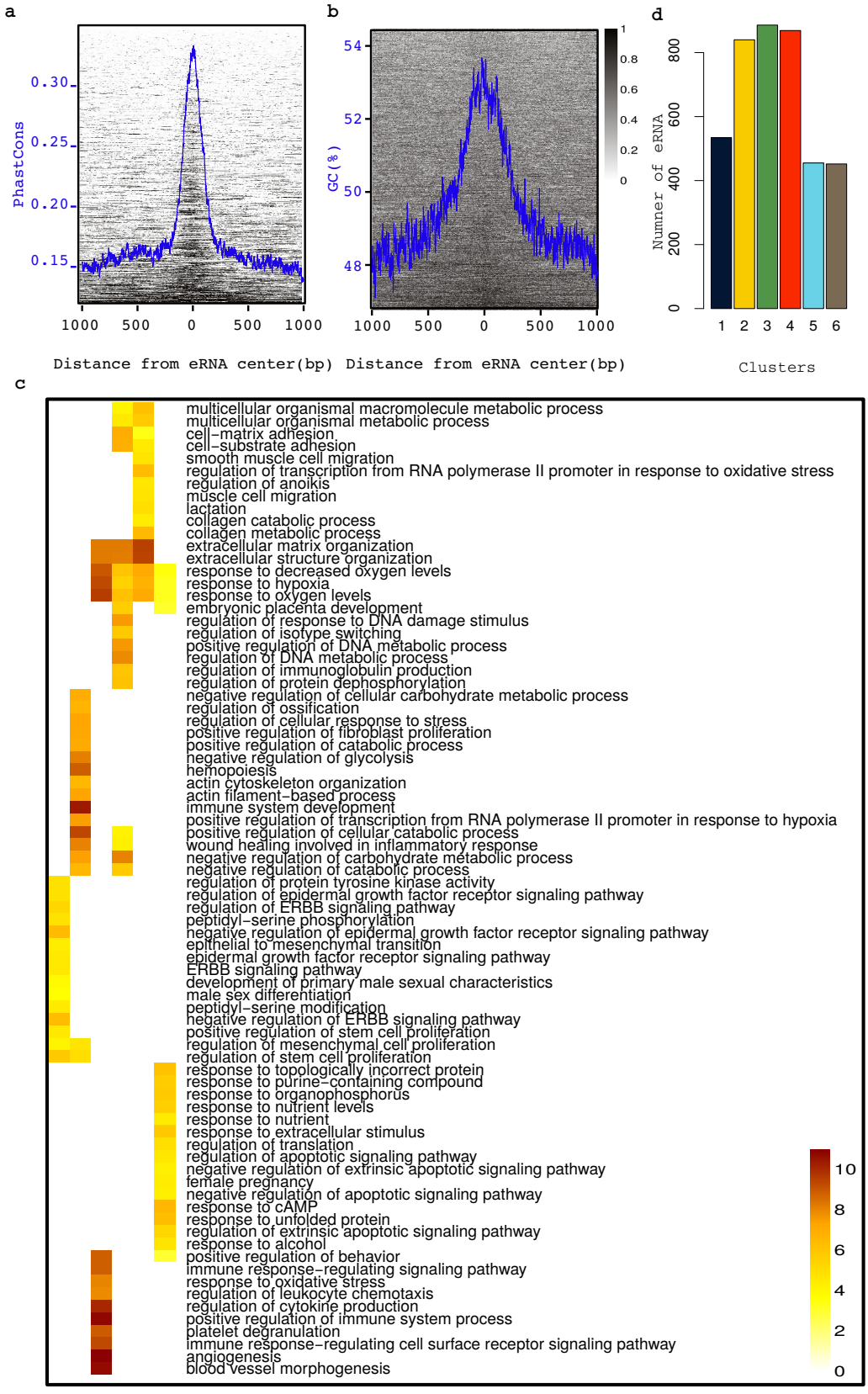

Figure 6.5: **a)** heatmap (rows are eRNAs, and columns are genomic position relative to eRNA centers) and mean signal (blue) shows that eRNAs are conserved at eRNA centers compared to flanking regions. **b)** same like (a) but shows higher GC content. **c)** heatmap displays the p-value of top 10 GO in enriched in every cluster. Every cluster is associated with distinct biological process (GO enrichment was computed using GREAT (McLean et al., 2010)); for visualization purpose, signal of heatmap was scaled by setting outliers values (values $\geq 98^{th}$ percentile) to the value of $\geq 98^{th}$ percentile. All values were further scaled to the [0, 1] range by dividing them by the maximum.

MOTIFS	SEQ-LOGO	E-VALUE	NUMBER OF MATCHES	BEST MATCH	P-VALUE OF BEST MATCH
MOTIF 1		1.6e-056	22	IRF4	2.65e-05
MOTIF 2		8.5e-064	12	IRF4	1.56e-04
MOTIF 3		3.0e-056	30	ZN148	1.96e-08
MOTIF 4		2.7e-033	21	MAZ	1.31e-07
MOTIF 5		9.9e-043	13	FOSL1	9.42e-07
MOTIF 6		9.4e-005	33	WT1	5.32e-08
MOTIF 7		4.4e-011	0	--	--
MOTIF 8		4.1e-004	9	MAZ	2.68e-05
MOTIF 9		9.8e-006	0	--	--
MOTIF 10		1.1e-003	0	--	--

Figure 6.6: Sequence logo of 10 most enriched motifs within eRNAs and their TF best match.

6.3.4 Degenerate activities of eRNAs in DMD

eRNAs expression was time specific (Figure 6.7a-b). They clustered into six waves each flanked by genes implicated in distinct biological processes (Appendix 6.5c). 69% of eRNAs correlated with at least one gene within 500kb (average 2.17 eRNA per gene and 1.37 gene per eRNA). The wave-like expression profile of eRNA clusters suggests possible distinct TFs controlling the expression of each cluster. However, discriminative motif analysis using DREME (Bailey, 2011) yielded no discriminative motifs, except for cluster 4 (Appendix 6.6). Moreover, while most eRNAs were active in both phenotypes, their expression profiles were inconsistent. Except for eRNAs of cluster 1, eRNAs from the same cluster peaked at different times (Figure 6.7c) and to lower expression level in DMD compared to healthy (Figure 6.7d). The lack of discriminative motifs between clusters and the asynchronous expression of eRNAs from the same cluster in DMD indicate that eRNA clusters are unlikely to be controlled by specific master regulator. Thus, we hypothesize that the degenerate activity of eRNAs in DMD may be due to a major change in the epigenetic landscape caused by the hyperactivity of HDACs (Colussi et al., 2008), degenerate activity of SMYD1, or other factors in DMD.

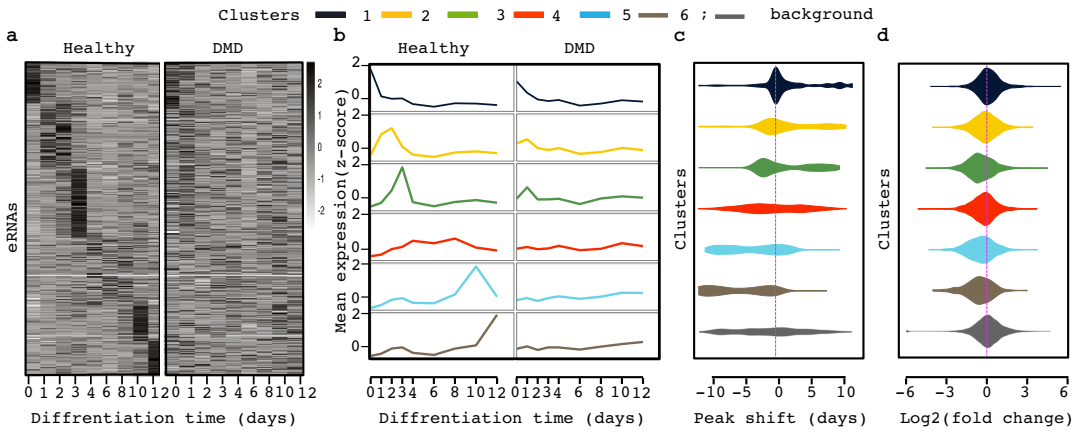

Figure 6.7: **a)** heatmap exhibits expression profile of eRNAs during differentiation; eRNAs are transcribed in time specific manner, however, distinct from control and DMD samples (healthy and DMD are clustered independently). **b)** lineplot showing expression profile of eRNAs illustrating that eRNA transcription comes in waves during muscle differentiation in control samples and the signal is lost in DMD (clustering based on healthy samples). **c)** violin plot illustrating that eRNAs from different waves exhibit an asynchronous shift in DMD samples. **d)** violin plots showing the distribution of fold change of expression between DMD and control at their peaks; except for the first wave, eRNA tend to peak to lower expression levels in DMD samples.

6.3.5 eRNA associate to transposon insertion sites

Recent reports indicate a role for retrotransposon ncRNA in stem cell maintenance (Fort et al., 2014). Data from our laboratory indicate a positive correlation between L1 and SINE retrotransposon activity and correct transcriptional activation of myogenic cell program (Bodega et al, submitted). Notably, L1 expression is under the control of nNOS HDAC pathway and is aberrantly repressed in DMD. While eRNAs are depleted in repeat elements (Andersson et al., 2014), 83% of eRNA are within $100 - 1500$ bases of repeat DNA (Figure 6.8a); thus we sought to investigate the association between eRNA and repeats. Repeat element composition around eRNAs is complex— no repeat family was enriched (Figure 6.8a). Further, most repeats flanking eRNAs are truncated (Figure 6.8c-d) and are unlikely to be active transposon. Recently, DNA double stranded break regions were shown to share common properties with eRNAs such as bidirectional transcription of small RNAs (Pefanis et al., 2015). We speculate that eRNA boundaries are insertion target sites of TEs. These insertions may regulate eRNA activities. This speculation needs to be further investigated.

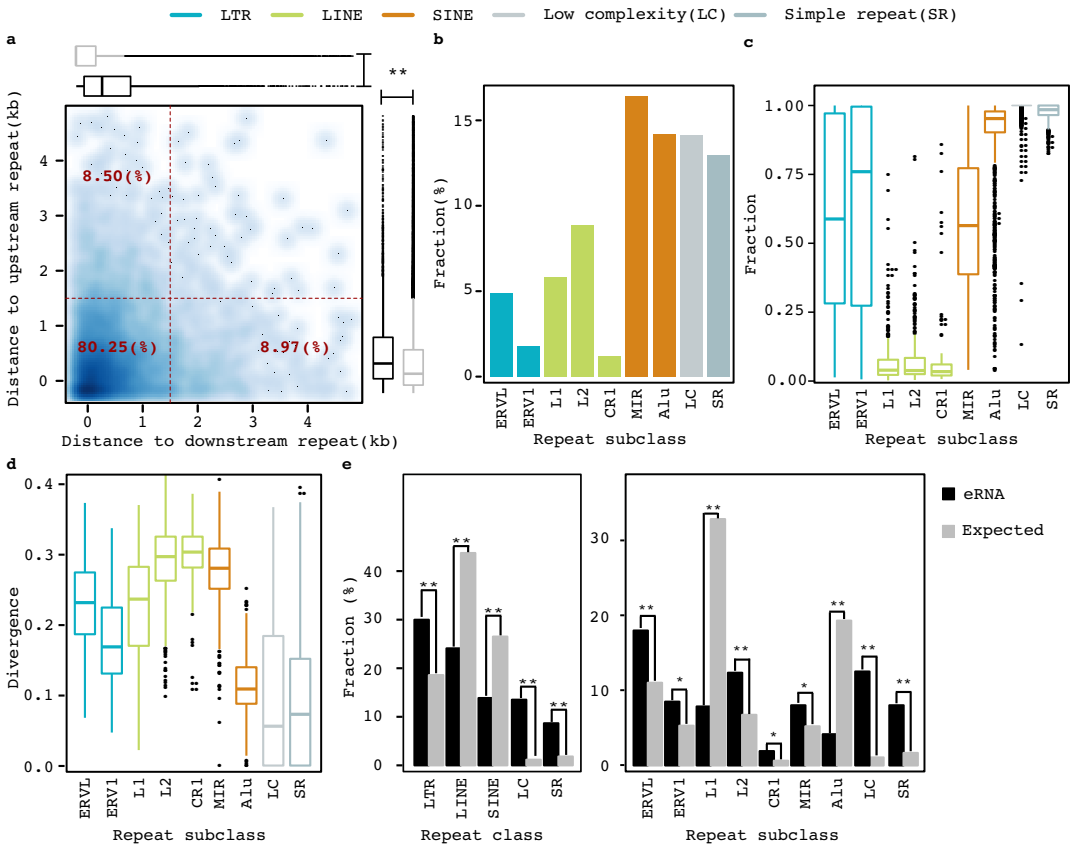

Figure 6.8: **a)** density scatterplot shows that most eRNAs are within 1.5kb from repeat element, though, most of them do not overlap any repeat element (it would be expected by chance that most eRNA overlap repeat elements — boxplot on sides of scatterplot). **b)** barplot shows the number of most abundant repeat elements around eRNAs. **c)** boxplot shows fraction of length of alignment of repeat element over the length of their corresponding full-length consensus; with the exception of simple and low complexity repeats, repeat elements around eRNAs cover only small a fraction of their full length consensus. **d)** boxplot shows the divergence of repeat elements from their corresponding full-length consensus. **e)** barplot shows number of eRNAs within repeat elements at class level (left panel) and subclass level (left panel); although eRNAs are depleted in repeats, eRNA overlapping repeats are enriched in LTRs; (*) indicate p-value < 0.05, (**) indicate p-value $< 1e-05$.

6.3.6 eRNAs overlapping repeats are enriched in LTRs

14% of eRNAs overlap repeat elements. Surprisingly, eRNAs overlapping repeats are enriched in LTRs and depleted in long interspersed repeats (LINEs) and LINEs (Figure 6.8d); the enrichment is dominated by ERVL and ERV1 subclasses of LTR. Fort et al., 2014, identified 217 LTR-associated eRNA (annotated by FANTOM) in human stem cells; those eRNAs are suggested to maintain cell pluripotency. We found 173 LTR-associated eRNAs, 63 of them were previously identified by FANTOM; only five were part of Fort et al., 2014 list (Figure 6.9a). We sought to determine whether the 63 LTR-associated eRNAs were related to stem cells. We found those eRNAs expressed in many cell types (average 153 cells per eRNA) including stem cells (Figure 6.9b-c). The remaining 110

LTR-associated eRNAs were exclusively expressed during myogenesis. We conclude that LTR and perhaps L1/SINE evolve subsets of distal regulatory elements that correlate with various functions including the maintenance of cell pluripotency and differentiation.

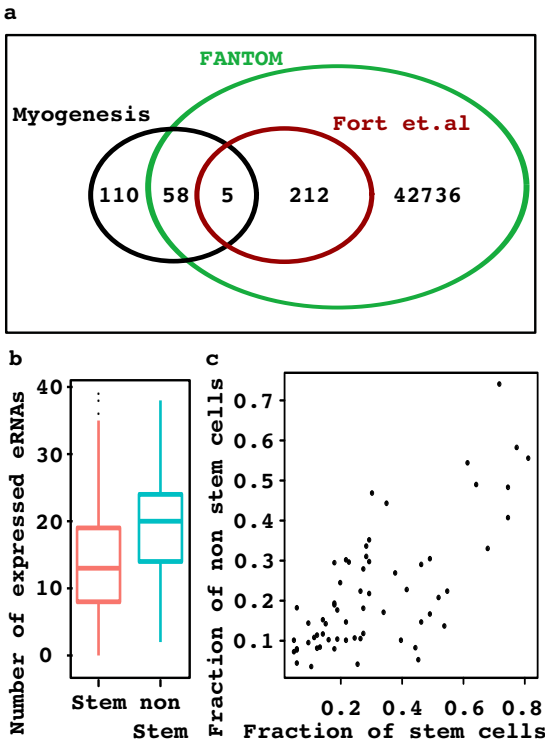

Figure 6.9: **a)** venn diagram showing small overlap between identified LTR-associated eRNAs and LTR-associated eRNAs reported by (Fort et al., 2014). **b)** boxplot shows that LTR-associated eRNAs are expressed in many somatic cells. **c)**LTR-associated eRNAs express similarly across stem and somatic cells.

Chapter 7

Concluding Remarks

7.1 Summary

This work provides new insights on the properties of noncoding elements in the genome. First, we investigated the existence of ultraconserved elements (UCEs) between distant species (spanning $\sim$ 892 million years of evolution): sponge, hydra, sea anemone, sea urchin, fruit-fly, human. Unexpectedly, we found a prevalence of UCEs which are maintained in the genomes in a lineage-specific manner. In addition, the newly identified UCEs showed some characteristics of regulatory element such as enrichment of transcription factor binding sites (TFBSs) and proximity to important developmental gene clusters. Then, we conducted the first comprehensive analysis on the epigenetic characteristics of highly conserved noncoding elements (HCNEs) in insects. In addition to confirming and generalising some previous observations about sequence features, e.g., enrichment of GC within HCNE and enrichment of TFBSs of transcription factors (TFs) implicated in development. We showed that HCNEs resided mostly in compact chromatin. Although the identified HCNEs did not overlap annotated enhancers, they also exhibit histone modifications associated with active enhancers at specific developmental stages. This study also showed that some HCNEs served as replication origin and associated to lamina-associated domains (LADs), suggesting possible alternative roles of HCNE. Finally, because most investigated noncoding elements shows enhancer-like characteristics, we sought to learn more about enhancers. Thus, we examined the activities of enhancers during human myogenesis. In this analysis, we generate two-time course datasets from healthy donors and donors affected with DMD using CAGE and RNA-Seq. The study annotated active, bi-directionally transcribed, enhancers specific to human myogenesis. Interestingly, those enhancers were highly stage-specific. Although they were active in both phenotypes (healthy and DMD), their activity was degenerate in DMD. Our analysis suggests that chromatin remodelers such as HDAC2 and SMYD1

may be implicated in their misregulation. Furthermore, we found a high correlation between active enhancers and repeat elements. This correlation, combined with observations from other studies, suggest that enhancers may be insertion sites of transposons— though this needs to be investigated further.

7.2 Perspective future

Although a limited number of transposable elements are capable of integrating into the genome (Cowley and Oakey, 2013), transposable element insertions make up most of the genome (de Koning et al., 2011). This indicates high activity of transposable elements throughout evolution. Recent studies showed dynamic transcription of full-length and truncated repeats in somatic cells in different contexts, e.g., development and stress (Fort et al., 2014). However, the role of repeat transcription and retrotransposition in somatic cells remains unclear.

Here, I will state some of my thoughts and speculations, based on my observation mainly from the work reported in the last chapter of this thesis, on the possible role of somatic retrotransposition in link to enhancer activities.

Our analysis of eRNAs during myogenesis proposes non-random spatial organization of eRNAs and repeat elements; eRNAs are surrounded by repeat element within 200bp on average. This exciting organization may suggest two interesting opinions: enhancers are retrotransposition hot spots or enhancers are promoters of repeat containing noncoding RNAs. The first hypothesis is supported partially by reported enrichment of double strand DNA break makers at eRNA loci (Pefanis et al., 2015). The second hypothesis is backed by the fact that some repeat in some noncoding RNAs serve as functional domains (for example, SINEUPs that regulate translation). However, these need much more investigation.

The other hypothesis is that repeats around enhancers are markers of enhancer regions, and this organization occurred to provide similar services (enhancement of expression of some target genes) from distinct genomic loci ("available" enhancers) imposed by constraints of dynamic chromatin structure. This hypothesis is based on the facts that: enhancers elements are active in time specific manner; eRNA life is much shorter than that of its target; eRNAs of different time points are enriched with similar motifs; eRNA clusters lack differential motifs between them.

Finally, although I find all these hypotheses attractive, all of them seem to suffer from the chicken and egg problem.

REFERENCES

Ncbi gene. `http://www.ncbi.nlm.nih.gov/gene/`. webcite.

N. Ahituv, Y. Zhu, A. Visel, A. Holt, V. Afzal, L. A. Pennacchio, and E. M. Rubin. Deletion of ultraconserved elements yields viable mice. *PLoS Biol.*, 5(9):e234, Sep 2007.

K. Ahmad and S. Henikoff. The histone variant H3.3 marks active chromatin by replication-independent nucleosome assembly. *Mol. Cell*, 9(6):1191–1200, Jun 2002.

A. Akalin, D. Fredman, E. Arner, X. Dong, J. C. Bryne, H. Suzuki, C. O. Daub, Y. Hayashizaki, and B. Lenhard. Transcriptional features of genomic regulatory blocks. *Genome Biol.*, 10(4):R38, 2009.

M. L. Allende, M. Manzanares, J. J. Tena, C. G. Feijoo, and J. L. Gomez-Skarmeta. Cracking the genome's second code: enhancer detection by combined phylogenetic footprinting and transgenic fish and frog embryos. *Methods*, 39(3):212–219, Jul 2006.

S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J. Mol. Biol.*, 215(3):403–410, Oct 1990.

A. H. Andersen, E. Gocke, B. J. Bonven, O. F. Nielsen, and O. Westergaard. Topoisomerase I has a strong binding preference for a conserved hexadecameric sequence in the promoter region of the rRNA gene from Tetrahymena pyriformis. *Nucleic Acids Res.*, 13(5):1543–1557, Mar 1985.

R. Andersson, C. Gebhard, I. Miguel-Escalada, I. Hoof, J. Bornholdt, M. Boyd, Y. Chen, X. Zhao, C. Schmidl, T. Suzuki, E. Ntini, E. Arner, E. Valen, K. Li, L. Schwarzfischer, D. Glatz, J. Raithel, B. Lilje, N. Rapin, F. O. Bagger, M. J?rgensen, P. R. Andersen, N. Bertin, O. Rackham, A. M. Burroughs, J. K. Baillie, Y. Ishizu, Y. Shimizu, E. Furuhata, S. Maeda, Y. Negishi, C. J. Mungall, T. F. Meehan, T. Lassmann, M. Itoh, H. Kawaji, N. Kondo, J. Kawai, A. Lennartsson, C. O. Daub, P. Heutink, D. A. Hume, T. H. Jensen, H. Suzuki, Y. Hayashizaki, F. Muller, A. R. Forrest, P. Carninci, M. Rehli, A. Sandelin, A. R. Forrest, H. Kawaji, M. Rehli, J. K. Baillie, M. J. de Hoon, V. Haberle, T. Lassmann, I. V. Kulakovskiy, M. Lizio, M. Itoh, R. Andersson, C. J. Mungall, T. F. Meehan, S. Schmeier, N. Bertin, M. J?rgensen, E. Dimont, E. Arner, C. Schmid, U. Schaefer, Y. A. Medvedeva, C. Plessy, M. Vitezic, J. Severin, C. A. Semple, Y. Ishizu, R. S. Young, M. Francescatto, I. Alam, D. Albanese, G. M. Altschuler, T. Arakawa, J. A. Archer, P. Arner, M. Babina, S. Rennie, P. J. Balwierz, A. G. Beckhouse, S. Pradhan-Bhatt, J. A. Blake, A. Blumenthal, B. Bodega, A. Bonetti, J. Briggs, F. Brombacher, A. M. Burroughs, A. Califano, C. V. Cannistracti, D. Carbajo, Y. Chen, M. Chierici, Y. Ciani, H. C. Clevers, E. Dalla, C. A. Davis, M. Detmar, A. D. Diehl, T. Dohi, F. Drabl?s, A. S. Edge, M. Edinger, K. Ekwall, M. Endoh, H. Enomoto, M. Fagiolini, L. Fairbairn, H. Fang, M. C. Farach-Carson,

G. J. Faulkner, A. V. Favorov, M. E. Fisher, M. C. Frith, R. Fujita, S. Fukuda, C. Furlanello, M. Furuno, J. Furusawa, T. B. Geijtenbeek, A. P. Gibson, T. Gingeras, D. Goldowitz, J. Gough, S. Guhl, R. Guler, S. Gustincich, T. J. Ha, M. Hamaguchi, M. Hara, M. Harbers, J. Harshbarger, A. Hasegawa, Y. Hasegawa, T. Hashimoto, M. Herlyn, K. J. Hitchens, S. J. Ho Sui, O. M. Hofman, I. Hoof, F. Hori, L. Huminiecki, K. Iida, T. Ikawa, B. R. Jankovic, H. Jia, A. Joshi, G. Jurman, B. Kaczkowski, C. Kai, K. Kaida, A. Kaiho, K. Kajiyama, M. Kanamori-Katayama, A. S. Kasianov, T. Kasukawa, S. Katayama, S. Kato, S. Kawaguchi, H. Kawamoto, Y. I. Kawamura, T. Kawashima, J. S. Kempfle, T. J. Kenna, J. Kere, L. M. Khachigian, T. Kitamura, S. P. Klinken, A. J. Knox, M. Kojima, S. Kojima, N. Kondo, H. Koseki, S. Koyasu, S. Krampitz, A. Kubosaki, A. T. Kwon, J. F. Laros, W. Lee, A. Lennartsson, K. Li, B. Lilje, L. Lipovich, A. Mackay-Sim, R. Manabe, J. C. Mar, B. Marchand, A. Mathelier, N. Mejhert, A. Meynert, Y. Mizuno, D. A. de Lima Morais, H. Morikawa, M. Morimoto, K. Moro, E. Motakis, H. Motohashi, C. L. Mummery, M. Murata, S. Nagao-Sato, Y. Nakachi, F. Nakahara, T. Nakamura, Y. Nakamura, K. Nakazato, E. van Nimwegen, N. Ninomiya, H. Nishiyori, S. Noma, T. Nozaki, S. Ogishima, N. Ohkura, H. Ohmiya, H. Ohno, M. Onshima, M. Okada-Hatakeyama, Y. Okazaki, V. Orlando, D. A. Ovchinnikov, A. Pain, R. Passier, M. Patrikakis, H. Persson, S. Piazza, J. G. Prendergast, O. J. Rackham, J. A. Ramilowski, M. Rashid, T. Ravasi, P. Rizzu, M. Roncador, S. Roy, M. B. Rye, E. Saijyo, A. Sajantila, A. Saka, S. Sakaguchi, M. Sakai, H. Sato, H. Satoh, S. Savvi, A. Saxena, C. Schneider, E. A. Schultes, G. G. Schultz-Tanzil, A. Schwegmann, T. Sengstag, G. Sheng, H. Shimoji, Y. Shimoni, J. W. Shin, C. Simon, D. Sugiyama, T. Sugiyama, M. Suzuki, N. Suzuki, R. K. Swoboda, P. A. 't Hoen, M. Tagami, N. Takahashi, J. Takai, H. Tanaka, H. Tatsukawa, Z. Tatum, M. Thompson, H. Toyoda, T. Toyodo, E. Valen, M. van de Wetering, L. M. van den Berg, R. Verardo, D. Vijayan, I. E. Vorontsov, W. W. Wasserman, S. Watanabe, C. A. Wells, L. N. Winteringham, E. Wolvetang, E. J. Wood, Y. Yamaguchi, M. Yamamoto, M. Yoneda, Y. Yonekura, S. Yoshida, S. E. Zabierowski, P. G. Zhang, X. Zhao, S. Zucchelli, K. M. Summers, H. Suzuki, C. O. Daub, J. Kawai, P. Heutink, W. Hide, T. C. Freeman, B. Lenhard, V. B. Bajic, M. S. Taylor, V. J. Makeev, A. Sandelin, D. A. Hume, P. Carninci, and Y. Hayashizaki. An atlas of active enhancers across human cell types and tissues. *Nature*, 507(7493):455–461, Mar 2014.

E. Arner, C. O. Daub, K. Vitting-Seerup, R. Andersson, B. Lilje, F. Drabl?s, A. Lennartsson, M. Ronnerblad, O. Hrydziuszko, M. Vitezic, T. C. Freeman, A. M. Alhendi, P. Arner, R. Axton, J. K. Baillie, A. Beckhouse, B. Bodega, J. Briggs, F. Brombacher, M. Davis, M. Detmar, A. Ehrlund, M. Endoh, A. Eslami, M. Fagiolini, L. Fairbairn, G. J. Faulkner, C. Ferrai, M. E. Fisher, L. Forrester, D. Goldowitz, R. Guler, T. Ha, M. Hara, M. Herlyn, T. Ikawa, C. Kai, H. Kawamoto, L. M. Khachigian, S. P. Klinken, S. Kojima, H. Koseki, S. Klein, N. Mejhert, K. Miyaguchi, Y. Mizuno, M. Morimoto, K. J. Morris, C. Mummery, Y. Nakachi, S. Ogishima, M. Okada-Hatakeyama, Y. Okazaki, V. Orlando, D. Ovchinnikov, R. Passier, M. Patrikakis, A. Pombo, X. Y. Qin, S. Roy, H. Sato, S. Savvi, A. Saxena, A. Schwegmann, D. Sugiyama, R. Swoboda, H. Tanaka, A. Tomoiu, L. N. Winteringham, E. Wolvetang, C. Yanagi-Mizuochi, M. Yoneda, S. Zabierowski, P. Zhang, I. Abugessaisa, N. Bertin, A. D. Diehl, S. Fukuda, M. Furuno, J. Harshbarger, A. Hasegawa, F. Hori, S. Ishikawa-Kato, Y. Ishizu, M. Itoh, T. Kawashima, M. Kojima, N. Kondo, M. Lizio, T. F. Meehan, C. J. Mungall, M. Murata, H. Nishiyori-Sueki, S. Sahin, S. Nagao-Sato, J. Severin, M. J. de Hoon, J. Kawai, T. Kasukawa, T. Lassmann, H. Suzuki, H. Kawaji, K. M. Summers, C. Wells, D. A. Hume, A. R. Forrest, A. Sandelin, P. Carninci, and

Y. Hayashizaki. Gene regulation. Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. *Science*, 347(6225):1010–1014, Feb 2015.

T. L. Bailey. DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics*, 27 (12):1653–1659, Jun 2011.

T. L. Bailey and C. Elkan. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol*, 2:28–36, 1994.

M. Baker. Making sense of chromatin states. *Nat. Methods*, 8(9):717–722, Sep 2011.

GustavoE Batista, EamonnJ Keogh, ObenMoses Tataw, and VinículusM de Souza. CID: an efficient complexity-invariant distance for time series. 28(3):634–669, 2014. doi: 10.1007/s10618-013-0312-3. URL `http://dx.doi.org/10.1007/s10618-013-0312-3`.

G. Bejerano, M. Pheasant, I. Makunin, S. Stephen, W. J. Kent, J. S. Mattick, and D. Haussler. Ultraconserved elements in the human genome. *Science*, 304(5675):1321–1325, May 2004.

G. Bejerano, C. B. Lowe, N. Ahituv, B. King, A. Siepel, S. R. Salama, E. M. Rubin, W. J. Kent, and D. Haussler. A distal enhancer and an ultraconserved exon are derived from a novel retroposon. *Nature*, 441(7089):87–90, May 2006.

S. Benko, J. A. Fantes, J. Amiel, D. J. Kleinjan, S. Thomas, J. Ramsay, N. Jamshidi, A. Essafi, S. Heaney, C. T. Gordon, D. McBride, C. Golzio, M. Fisher, P. Perry, V. Abadie, C. Ayuso, M. Holder-Espinasse, N. Kilpatrick, M. M. Lees, A. Picard, I. K. Temple, P. Thomas, M. P. Vazquez, M. Vekemans, H. Roest Crollius, N. D. Hastie, A. Munnich, H. C. Etchevers, A. Pelet, P. G. Farlie, D. R. Fitzpatrick, and S. Lyonnet. Highly conserved non-coding elements on either side of SOX9 associated with Pierre Robin sequence. *Nat. Genet.*, 41(3):359–364, Mar 2009.

G. Benson. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.*, 27 (2):573–580, Jan 1999.

C. F. Bentzinger, Y. X. Wang, and M. A. Rudnicki. Building muscle: molecular regulation of myogenesis. *Cold Spring Harb Perspect Biol*, 4(2), Feb 2012.

C. M. Bergman, J. W. Carlson, and S. E. Celniker. Drosophila DNase I footprint database: a systematic genome annotation of transcription factor binding sites in the fruitfly, Drosophila melanogaster. *Bioinformatics*, 21(8):1747–1749, Apr 2005.

B. E. Bernstein, T. S. Mikkelsen, X. Xie, M. Kamal, D. J. Huebert, J. Cuff, B. Fry, A. Meissner, M. Wernig, K. Plath, R. Jaenisch, A. Wagschal, R. Feil, S. L. Schreiber, and E. S. Lander. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell*, 125 (2):315–326, Apr 2006.

D. Boffelli, M. A. Nobrega, and E. M. Rubin. Comparative genomics at the vertebrate extremes. *Nat. Rev. Genet.*, 5(6):456–465, Jun 2004.

E. Bonnefoy, G. A. Orsi, P. Couble, and B. Loppin. The essential role of Drosophila HIRA for de novo assembly of paternal chromatin at fertilization. *PLoS Genet.*, 3(10):1991–2006, Oct 2007.

M. Buckingham and S. D. Vincent. Distinct and dynamic myogenic populations in the vertebrate embryo. *Curr. Opin. Genet. Dev.*, 19(5):444–453, Oct 2009.

Diane Burgess and Michael Freeling. The Most Deeply Conserved Noncoding Sequences in Plants Serve Similar Functions to Those in Vertebrates Despite Large Differences in Evolutionary Rates. *The Plant Cell Online*, pages tpc.113.121905+, March 2014. ISSN 1532-298X. doi: 10.1105/tpc.113.121905. URL `http://dx.doi.org/10.1105/tpc.113.121905`.

K. M. Bush, B. T. Yuen, B. L. Barrilleaux, J. W. Riggs, H. O'Geen, R. F. Cotterman, and P. S. Knoepfler. Endogenous mammalian histone H3.3 exhibits chromatin-related functions during development. *Epigenetics Chromatin*, 6(1):7, 2013.

C. V. Cannistraci, T. Ravasi, F. M. Montevecchi, T. Ideker, and M. Alessio. Nonlinear dimension reduction and clustering by Minimum Curvilinearity unfold neuropathic pain and tissue embryological classes. *Bioinformatics*, 26(18):i531–539, Sep 2010.

C. V. Cannistraci, G. Alanis-Lobato, and T. Ravasi. Minimum curvilinearity to enhance topological prediction of protein interactions by network embedding. *Bioinformatics*, 29(13):199–209, Jul 2013.

P. Carninci, T. Kasukawa, S. Katayama, J. Gough, M. C. Frith, N. Maeda, R. Oyama, T. Ravasi, B. Lenhard, C. Wells, R. Kodzius, K. Shimokawa, V. B. Bajic, S. E. Brenner, S. Batalov, A. R. Forrest, M. Zavolan, M. J. Davis, L. G. Wilming, V. Aidinis, J. E. Allen, A. Ambesi-Impiombato, R. Apweiler, R. N. Aturaliya, T. L. Bailey, M. Bansal, L. Baxter, K. W. Beisel, T. Bersano, H. Bono, A. M. Chalk, K. P. Chiu, V. Choudhary, A. Christoffels, D. R. Clutterbuck, M. L. Crowe, E. Dalla, B. P. Dalrymple, B. de Bono, G. Della Gatta, D. di Bernardo, T. Down, P. Engstrom, M. Fagiolini, G. Faulkner, C. F. Fletcher, T. Fukushima, M. Furuno, S. Futaki, M. Gariboldi, P. Georgii-Hemming, T. R. Gingeras, T. Gojobori, R. E. Green, S. Gustincich, M. Harbers, Y. Hayashi, T. K. Hensch, N. Hirokawa, D. Hill, L. Huminiecki, M. Iacono, K. Ikeo, A. Iwama, T. Ishikawa, M. Jakt, A. Kanapin, M. Katoh, Y. Kawasawa, J. Kelso, H. Kitamura, H. Kitano, G. Kollias, S. P. Krishnan, A. Kruger, S. K. Kummerfeld, I. V. Kurochkin, L. F. Lareau, D. Lazarevic, L. Lipovich, J. Liu, S. Liuni, S. McWilliam, M. Madan Babu, M. Madera, L. Marchionni, H. Matsuda, S. Matsuzawa, H. Miki, F. Mignone, S. Miyake, K. Morris, S. Mottagui-Tabar, N. Mulder, N. Nakano, H. Nakauchi, P. Ng, R. Nilsson, S. Nishiguchi, S. Nishikawa, F. Nori, O. Ohara, Y. Okazaki, V. Orlando, K. C. Pang, W. J. Pavan, G. Pavesi, G. Pesole, N. Petrovsky, S. Piazza, J. Reed, J. F. Reid, B. Z. Ring, M. Ringwald, B. Rost, Y. Ruan, S. L. Salzberg, A. Sandelin, C. Schneider, C. Schonbach, K. Sekiguchi, C. A. Semple, S. Seno, L. Sessa, Y. Sheng, Y. Shibata, H. Shimada, K. Shimada, D. Silva, B. Sinclair, S. Sperling, E. Stupka, K. Sugiura, R. Sultana, Y. Takenaka, K. Taki, K. Tammoja, S. L. Tan, S. Tang, M. S. Taylor, J. Tegner, S. A. Teichmann, H. R. Ueda, E. van Nimwegen, R. Verardo, C. L. Wei, K. Yagi, H. Yamanishi, E. Zabarovsky, S. Zhu, A. Zimmer, W. Hide, C. Bult, S. M. Grimmond, R. D. Teasdale, E. T. Liu, V. Brusic, J. Quackenbush, C. Wahlestedt, J. S. Mattick, D. A. Hume, C. Kai, D. Sasaki, Y. Tomaru, S. Fukuda, M. Kanamori-Katayama, M. Suzuki, J. Aoki, T. Arakawa, J. Iida, K. Imamura, M. Itoh, T. Kato, H. Kawaji, N. Kawagashira, T. Kawashima, M. Kojima, S. Kondo, H. Konno, K. Nakano, N. Ninomiya, T. Nishio, M. Okada, C. Plessy, K. Shibata, T. Shiraki, S. Suzuki, M. Tagami, K. Waki, A. Watahiki, Y. Okamura-Oho, H. Suzuki, J. Kawai, and

Y. Hayashizaki. The transcriptional landscape of the mammalian genome. *Science*, 309(5740): 1559–1563, Sep 2005.

I. Catucci, P. Verderio, S. Pizzamiglio, S. Manoukian, B. Peissel, M. Barile, L. Tizzoni, L. Bernard, F. Ravagnani, L. Galastri, M. A. Pierotti, P. Radice, and P. Peterlongo. SNPs in ultraconserved elements and familial breast cancer risk. *Carcinogenesis*, 30(3):544–545, Mar 2009.

C. Cayrou, P. Coulombe, A. Vigneron, S. Stanojcic, O. Ganier, I. Peiffer, E. Rivals, A. Puy, S. Laurent-Chabalier, R. Desprat, and M. Mechali. Genome-scale analysis of metazoan replication origins reveals their organization in specific but flexible sites defined by conserved features. *Genome Res.*, 21(9):1438–1449, Sep 2011.

M. Cesana, D. Cacchiarelli, I. Legnini, T. Santini, O. Sthandier, M. Chinappi, A. Tramontano, and I. Bozzoni. A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA. *Cell*, 147(2):358–369, Oct 2011.

J. A. Chapman, E. F. Kirkness, O. Simakov, S. E. Hampson, T. Mitros, T. Weinmaier, T. Rattei, P. G. Balasubramanian, J. Borman, D. Busam, K. Disbennett, C. Pfannkoch, N. Sumin, G. G. Sutton, L. D. Viswanathan, B. Walenz, D. M. Goodstein, U. Hellsten, T. Kawashima, S. E. Prochnik, N. H. Putnam, S. Shu, B. Blumberg, C. E. Dana, L. Gee, D. F. Kibler, L. Law, D. Lindgens, D. E. Martinez, J. Peng, P. A. Wigge, B. Bertulat, C. Guder, Y. Nakamura, S. Ozbek, H. Watanabe, K. Khalturin, G. Hemmrich, A. Franke, R. Augustin, S. Fraune, E. Hayakawa, S. Hayakawa, M. Hirose, J. S. Hwang, K. Ikeo, C. Nishimiya-Fujisawa, A. Ogura, T. Takahashi, P. R. Steinmetz, X. Zhang, R. Aufschnaiter, M. K. Eder, A. K. Gorny, W. Salvenmoser, A. M. Heimberg, B. M. Wheeler, K. J. Peterson, A. Bottger, P. Tischler, A. Wolf, T. Gojobori, K. A. Remington, R. L. Strausberg, J. C. Venter, U. Technau, B. Hobmayer, T. C. Bosch, T. W. Holstein, T. Fujisawa, H. R. Bode, C. N. David, D. S. Rokhsar, and R. E. Steele. The dynamic genome of Hydra. *Nature*, 464(7288):592–596, Mar 2010.

D. Charif and J.R. Lobry. SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In U. Bastolla, M. Porto, H.E. Roman, and M. Vendruscolo, editors, *Structural approaches to sequence evolution: Molecules, networks, populations*, Biological and Medical Physics, Biomedical Engineering, pages 207–232. Springer Verlag, New York, 2007. ISBN : 978-3-540-35305-8.

X. Chen, Z. Chen, H. Chen, Z. Su, J. Yang, F. Lin, S. Shi, and X. He. Nucleosomes suppress spontaneous mutations base-specifically in eukaryotes. *Science*, 335(6073):1235–1238, Mar 2012.

C. W. Chiang, A. Derti, D. Schwartz, M. F. Chou, J. N. Hirschhorn, and C. T. Wu. Ultraconserved elements: analyses of dosage sensitivity, motifs and boundaries. *Genetics*, 180(4):2277–2293, Dec 2008.

S. Clancy. DNA transcription. *Nature Education.*, 1(1):41, 2008.

C. Colussi, C. Mozzetta, A. Gurtner, B. Illi, J. Rosati, S. Straino, G. Ragone, M. Pescatori, G. Zaccagnini, A. Antonini, G. Minetti, F. Martelli, G. Piaggio, P. Gallinari, C. Steinkuhler, C. Steinkulher, E. Clementi, C. Dell'Aversana, L. Altucci, A. Mai, M. C. Capogrossi, P. L. Puri, and C. Gaetano. HDAC2 blockade by nitric oxide and histone deacetylase inhibitors reveals a common target

in Duchenne muscular dystrophy treatment. *Proc. Natl. Acad. Sci. U.S.A.*, 105(49):19183–19187, Dec 2008.

M. Cowley and R. J. Oakey. Transposable elements re-wire and fine-tune the transcriptome. *PLoS Genet.*, 9(1):e1003234, 2013.

T. Cremer and C. Cremer. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nat. Rev. Genet.*, 2(4):292–301, Apr 2001.

C. Daniel, M. T. Ven?, Y. Ekdahl, J. Kjems, and M. Ohman. A distant cis acting intronic element induces site-selective RNA editing. *Nucleic Acids Res.*, 40(19):9876–9886, Oct 2012.

A. De Grassi, C. Segala, F. Iannelli, S. Volorio, L. Bertario, P. Radice, L. Bernard, and F. D. Ciccarelli. Ultradeep sequencing of a human ultraconserved region reveals somatic and constitutional genomic instability. *PLoS Biol.*, 8(1):e1000275, Jan 2010.

A. P. de Koning, W. Gu, T. A. Castoe, M. A. Batzer, and D. D. Pollock. Repetitive elements may comprise over two-thirds of the human genome. *PLoS Genet.*, 7(12):e1002384, Dec 2011.

L. De Koning, A. Corpet, J. E. Haber, and G. Almouzni. Histone chaperones: an escort network regulating histone traffic. *Nat. Struct. Mol. Biol.*, 14(11):997–1007, Nov 2007.

A. Derti, F. P. Roth, G. M. Church, and C. T. Wu. Mammalian ultraconserved elements are strongly depleted among segmental duplications and copy number variants. *Nat. Genet.*, 38(10):1216–1220, Oct 2006.

C. Dessimoz, B. Boeckmann, A. C. Roth, and G. H. Gonnet. Detecting non-orthology in the COGs database and other approaches grouping orthologs using genome-specific best hits. *Nucleic Acids Res.*, 34(11):3309–3316, 2006.

C. Dinant and M. S. Luijsterburg. The emerging role of HP1 in the DNA damage response. *Mol. Cell. Biol.*, 29(24):6335–6340, Dec 2009.

G. J. Dolecki, G. Wang, and T. Humphreys. Stage- and tissue-specific expression of two homeo box genes in sea urchin embryos and adults. *Nucleic Acids Res.*, 16(24):11543–11558, Dec 1988.

G. R. Dubyak. Ion homeostasis, channels, and transporters: an update on cellular mechanisms. *Adv Physiol Educ*, 28(1-4):143–154, Dec 2004.

I. Dunham, A. Kundaje, S. F. Aldred, P. J. Collins, C. A. Davis, F. Doyle, C. B. Epstein, S. Frietze, J. Harrow, R. Kaul, J. Khatun, B. R. Lajoie, S. G. Landt, B. K. Lee, F. Pauli, K. R. Rosenbloom, P. Sabo, A. Safi, A. Sanyal, N. Shoresh, J. M. Simon, L. Song, N. D. Trinklein, R. C. Altshuler, E. Birney, J. B. Brown, C. Cheng, S. Djebali, X. Dong, I. Dunham, J. Ernst, T. S. Furey, M. Gerstein, B. Giardine, M. Greven, R. C. Hardison, R. S. Harris, J. Herrero, M. M. Hoffman, S. Iyer, M. Kellis, J. Khatun, P. Kheradpour, A. Kundaje, T. Lassmann, Q. Li, X. Lin, G. K. Marinov, A. Merkel, A. Mortazavi, S. C. Parker, T. E. Reddy, J. Rozowsky, F. Schlesinger, R. E. Thurman, J. Wang, L. D. Ward, T. W. Whitfield, S. P. Wilder, W. Wu, H. S. Xi, K. Y. Yip, J. Zhuang, M. J. Pazin, R. F. Lowdon, L. A. Dillon, L. B. Adams, C. J. Kelly, J. Zhang,

J. R. Wexler, E. D. Green, P. J. Good, E. A. Feingold, B. E. Bernstein, E. Birney, G. E. Crawford, J. Dekker, L. Elnitski, P. J. Farnham, M. Gerstein, M. C. Giddings, T. R. Gingeras, E. D. Green, R. Guigo, R. C. Hardison, T. J. Hubbard, M. Kellis, W. Kent, J. D. Lieb, E. H. Margulies, R. M. Myers, M. Snyder, J. A. Stamatoyannopoulos, S. A. Tenenbaum, Z. Weng, K. P. White, B. Wold, J. Khatun, Y. Yu, J. Wrobel, B. A. Risk, H. P. Gunawardena, H. C. Kuiper, C. W. Maier, L. Xie, X. Chen, M. C. Giddings, B. E. Bernstein, C. B. Epstein, N. Shoresh, J. Ernst, P. Kheradpour, T. S. Mikkelsen, S. Gillespie, A. Goren, O. Ram, X. Zhang, L. Wang, R. Issner, M. J. Coyne, T. Durham, M. Ku, T. Truong, L. D. Ward, R. C. Altshuler, M. L. Eaton, M. Kellis, S. Djebali, C. A. Davis, A. Merkel, A. Dobin, T. Lassmann, A. Mortazavi, A. Tanzer, J. Lagarde, W. Lin, F. Schlesinger, C. Xue, G. K. Marinov, J. Khatun, B. A. Williams, C. Zaleski, J. Rozowsky, M. Roder, F. Kokocinski, R. F. Abdelhamid, T. Alioto, I. Antoshechkin, M. T. Baer, P. Batut, I. Bell, K. Bell, S. Chakrabortty, X. Chen, J. Chrast, J. Curado, T. Derrien, J. Drenkow, E. Dumais, J. Dumais, R. Duttagupta, M. Fastuca, K. Fejes-Toth, P. Ferreira, S. Foissac, M. J. Fullwood, H. Gao, D. Gonzalez, A. Gordon, H. P. Gunawardena, C. Howald, S. Jha, R. Johnson, P. Kapranov, B. King, C. Kingswood, G. Li, O. J. Luo, E. Park, J. B. Preall, K. Presaud, P. Ribeca, B. A. Risk, D. Robyr, X. Ruan, M. Sammeth, K. S. Sandhu, L. Schaeffer, L. H. See, A. Shahab, J. Skancke, A. M. Suzuki, H. Takahashi, H. Tilgner, D. Trout, N. Walters, H. Wang, J. Wrobel, Y. Yu, Y. Hayashizaki, J. Harrow, M. Gerstein, T. J. Hubbard, A. Reymond, S. E. Antonarakis, G. J. Hannon, M. C. Giddings, Y. Ruan, B. Wold, P. Carninci, R. Guigo, T. R. Gingeras, K. R. Rosenbloom, C. A. Sloan, K. Learned, V. S. Malladi, M. C. Wong, G. P. Barber, M. S. Cline, T. R. Dreszer, S. G. Heitner, D. Karolchik, W. Kent, V. M. Kirkup, L. R. Meyer, J. C. Long, M. Maddren, B. J. Raney, T. S. Furey, L. Song, L. L. Grasfeder, P. G. Giresi, B. K. Lee, A. Battenhouse, N. C. Sheffield, J. M. Simon, K. A. Showers, A. Safi, D. London, A. A. Bhinge, C. Shestak, M. R. Schaner, S. K. Kim, Z. Z. Zhang, P. A. Mieczkowski, J. O. Mieczkowska, Z. Liu, R. M. McDaniell, Y. Ni, N. U. Rashid, M. J. Kim, S. Adar, Z. Zhang, T. Wang, D. Winter, D. Keefe, E. Birney, V. R. Iyer, J. D. Lieb, G. E. Crawford, G. Li, K. S. Sandhu, M. Zheng, P. Wang, O. J. Luo, A. Shahab, M. J. Fullwood, X. Ruan, Y. Ruan, R. M. Myers, F. Pauli, B. A. Williams, J. Gertz, G. K. Marinov, T. E. Reddy, J. Vielmetter, E. Partridge, D. Trout, K. E. Varley, C. Gasper, A. Bansal, S. Pepke, P. Jain, H. Amrhein, K. M. Bowling, M. Anaya, M. K. Cross, B. King, M. A. Muratet, I. Antoshechkin, K. M. Newberry, K. McCue, A. S. Nesmith, K. I. Fisher-Aylor, B. Pusey, G. DeSalvo, S. L. Parker, S. Balasubramanian, N. S. Davis, S. K. Meadows, T. Eggleston, C. Gunter, J. Newberry, S. E. Levy, D. M. Absher, A. Mortazavi, W. H. Wong, B. Wold, M. J. Blow, A. Visel, L. A. Pennachio, L. Elnitski, E. H. Margulies, S. C. Parker, H. M. Petrykowska, A. Abyzov, B. Aken, D. Barrell, G. Barson, A. Berry, A. Bignell, V. Boychenko, G. Bussotti, J. Chrast, C. Davidson, T. Derrien, G. Despacio-Reyes, M. Diekhans, I. Ezkurdia, A. Frankish, J. Gilbert, J. M. Gonzalez, E. Griffiths, R. Harte, D. A. Hendrix, C. Howald, T. Hunt, I. Jungreis, M. Kay, E. Khurana, F. Kokocinski, J. Leng, M. F. Lin, J. Loveland, Z. Lu, D. Manthravadi, M. Mariotti, J. Mudge, G. Mukherjee, C. Notredame, B. Pei, J. M. Rodriguez, G. Saunders, A. Sboner, S. Searle, C. Sisu, C. Snow, C. Steward, A. Tanzer, E. Tapanari, M. L. Tress, M. J. van Baren, N. Walters, S. Washietl, L. Wilming, A. Zadissa, Z. Zhang, M. Brent, D. Haussler, M. Kellis, A. Valencia, M. Gerstein, A. Reymond, R. Guigo, J. Harrow, T. J. Hubbard, S. G. Landt, S. Frietze, A. Abyzov, N. Addleman, R. P. Alexander, R. K. Auerbach, S. Balasubramanian, K. Bettinger, N. Bhardwaj, A. P. Boyle, A. R. Cao, P. Cayt-

ing, A. Charos, Y. Cheng, C. Cheng, C. Eastman, G. Euskirchen, J. D. Fleming, F. Grubert, L. Habegger, M. Hariharan, A. Harmanci, S. Iyengar, V. X. Jin, K. J. Karczewski, M. Kasowski, P. Lacroute, H. Lam, N. Lamarre-Vincent, J. Leng, J. Lian, M. Lindahl-Allen, R. Min, B. Miotto, H. Monahan, Z. Moqtaderi, X. J. Mu, H. O'Geen, Z. Ouyang, D. Patacsil, B. Pei, D. Raha, L. Ramirez, B. Reed, J. Rozowsky, A. Sboner, M. Shi, C. Sisu, T. Slifer, H. Witt, L. Wu, X. Xu, K. K. Yan, X. Yang, K. Y. Yip, Z. Zhang, K. Struhl, S. M. Weissman, M. Gerstein, P. J. Farnham, M. Snyder, S. A. Tenenbaum, L. O. Penalva, F. Doyle, S. Karmakar, S. G. Landt, R. R. Bhanvadia, A. Choudhury, M. Domanus, L. Ma, J. Moran, D. Patacsil, T. Slifer, A. Victorsen, X. Yang, M. Snyder, T. Auer, L. Centanin, M. Eichenlaub, F. Gruhl, S. Heermann, B. Hoeckendorf, D. Inoue, T. Kellner, S. Kirchmaier, C. Mueller, R. Reinhardt, L. Schertel, S. Schneider, R. Sinn, B. Wittbrodt, J. Wittbrodt, Z. Weng, T. W. Whitfield, J. Wang, P. J. Collins, S. F. Aldred, N. D. Trinklein, E. C. Partridge, R. M. Myers, J. Dekker, G. Jain, B. R. Lajoie, A. Sanyal, G. Balasundaram, D. L. Bates, R. Byron, T. K. Canfield, M. J. Diegel, D. Dunn, A. K. Ebersol, T. Frum, K. Garg, E. Gist, R. Hansen, L. Boatman, E. Haugen, R. Humbert, G. Jain, A. K. Johnson, E. M. Johnson, T. V. Kutyavin, B. R. Lajoie, K. Lee, D. Lotakis, M. T. Maurano, S. J. Neph, F. V. Neri, E. D. Nguyen, H. Qu, A. P. Reynolds, V. Roach, E. Rynes, P. Sabo, M. E. Sanchez, R. S. Sandstrom, A. Sanyal, A. O. Shafer, A. B. Stergachis, S. Thomas, R. E. Thurman, B. Vernot, J. Vierstra, S. Vong, H. Wang, M. A. Weaver, Y. Yan, M. Zhang, J. M. Akey, M. Bender, M. O. Dorschner, M. Groudine, M. J. MacCoss, P. Navas, G. Stamatoyannopoulos, R. Kaul, J. Dekker, J. A. Stamatoyannopoulos, I. Dunham, K. Beal, A. Brazma, P. Flicek, J. Herrero, N. Johnson, D. Keefe, M. Lukk, N. M. Luscombe, D. Sobral, J. M. Vaquerizas, S. P. Wilder, S. Batzoglou, A. Sidow, N. Hussami, S. Kyriazopoulou-Panagiotopoulou, M. W. Libbrecht, M. A. Schaub, A. Kundaje, R. C. Hardison, W. Miller, B. Giardine, R. S. Harris, W. Wu, P. J. Bickel, B. Banfai, N. P. Boley, J. B. Brown, H. Huang, Q. Li, J. J. Li, W. S. Noble, J. A. Bilmes, O. J. Buske, M. M. Hoffman, A. D. Sahu, P. V. Kharchenko, P. J. Park, D. Baker, J. Taylor, Z. Weng, S. Iyer, X. Dong, M. Greven, X. Lin, J. Wang, H. S. Xi, J. Zhuang, M. Gerstein, R. P. Alexander, S. Balasubramanian, C. Cheng, A. Harmanci, L. Lochovsky, R. Min, X. J. Mu, J. Rozowsky, K. K. Yan, K. Y. Yip, and E. Birney. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, Sep 2012.

C. W. Dunn, A. Hejnol, D. Q. Matus, K. Pang, W. E. Browne, S. A. Smith, E. Seaver, G. W. Rouse, M. Obst, G. D. Edgecombe, M. V. S?rensen, S. H. Haddock, A. Schmidt-Rhaesa, A. Okusu, R. M. Kristensen, W. C. Wheeler, M. Q. Martindale, and G. Giribet. Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature*, 452(7188):745–749, Apr 2008.

R. C. Edgar. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, 5:113, Aug 2004.

P. G. Engstrom, S. J. Ho Sui, O. Drivenes, T. S. Becker, and B. Lenhard. Genomic regulatory blocks underlie extensive microsynteny conservation in insects. *Genome Res.*, 17(12):1898–1908, Dec 2007.

A. J. Enright, S. Van Dongen, and C. A. Ouzounis. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.*, 30(7):1575–1584, Apr 2002.

B. C. Faircloth, J. E. McCormack, N. G. Crawford, M. G. Harvey, R. T. Brumfield, and T. C.

Glenn. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Syst. Biol.*, 61(5):717–726, Oct 2012.

S. Falcon and R. Gentleman. Using GOstats to test gene lists for GO term association. *Bioinformatics*, 23(2):257–258, Jan 2007.

G. J. Faulkner, Y. Kimura, C. O. Daub, S. Wani, C. Plessy, K. M. Irvine, K. Schroder, N. Cloonan, A. L. Steptoe, T. Lassmann, K. Waki, N. Hornig, T. Arakawa, H. Takahashi, J. Kawai, A. R. Forrest, H. Suzuki, Y. Hayashizaki, D. A. Hume, V. Orlando, S. M. Grimmond, and P. Carninci. The regulated retrotransposon transcriptome of mammalian cells. *Nat. Genet.*, 41(5):563–571, May 2009.

G. J. Filion, J. G. van Bemmel, U. Braunschweig, W. Talhout, J. Kind, L. D. Ward, W. Brugman, I. J. de Castro, R. M. Kerkhoven, H. J. Bussemaker, and B. van Steensel. Systematic protein location mapping reveals five principal chromatin types in Drosophila cells. *Cell*, 143(2):212–224, Oct 2010.

A. Fort, K. Hashimoto, D. Yamada, M. Salimullah, C. A. Keya, A. Saxena, A. Bonetti, I. Voineagu, N. Bertin, A. Kratz, Y. Noro, C. H. Wong, M. de Hoon, R. Andersson, A. Sandelin, H. Suzuki, C. L. Wei, H. Koseki, Y. Hasegawa, A. R. Forrest, and P. Carninci. Deep transcriptome profiling of mammalian stem cells supports a regulatory role for retrotransposons in pluripotency maintenance. *Nat. Genet.*, 46(6):558–566, Jun 2014.

M. Frisch, K. Frech, A. Klingenhoff, K. Cartharius, I. Liebich, and T. Werner. In silico prediction of scaffold/matrix attachment regions in large genomic sequences. *Genome Res.*, 12(2):349–354, Feb 2002.

P. A. Fujita, B. Rhead, A. S. Zweig, A. S. Hinrichs, D. Karolchik, M. S. Cline, M. Goldman, G. P. Barber, H. Clawson, A. Coelho, M. Diekhans, T. R. Dreszer, B. M. Giardine, R. A. Harte, J. Hillman-Jackson, F. Hsu, V. Kirkup, R. M. Kuhn, K. Learned, C. H. Li, L. R. Meyer, A. Pohl, B. J. Raney, K. R. Rosenbloom, K. E. Smith, D. Haussler, and W. J. Kent. The UCSC Genome Browser database: update 2011. *Nucleic Acids Res.*, 39(Database issue):D876–882, Jan 2011.

M. B. Gerstein, C. Bruce, J. S. Rozowsky, D. Zheng, J. Du, J. O. Korbel, O. Emanuelsson, Z. D. Zhang, S. Weissman, and M. Snyder. What is a gene, post-ENCODE? History and updated definition. *Genome Res.*, 17(6):669–681, Jun 2007.

A. Gherman, P. E. Chen, T. M. Teslovich, P. Stankiewicz, M. Withers, C. S. Kashuk, A. Chakravarti, J. R. Lupski, D. J. Cutler, and N. Katsanis. Population bottlenecks as a potential major shaping force of human genome architecture. *PLoS Genet.*, 3(7):e119, Jul 2007.

L. Giordani and P. L. Puri. Epigenetic control of skeletal muscle regeneration: Integrating genetic determinants and environmental changes. *FEBS J.*, 280(17):4014–4025, Sep 2013.

G. V. Glazko, E. V. Koonin, I. B. Rogozin, and S. A. Shabalina. A significant fraction of conserved noncoding DNA in human and mouse consists of predicted matrix attachment regions. *Trends Genet.*, 19(3):119–124, Mar 2003.

E. A. Glazov, M. Pheasant, E. A. McGraw, G. Bejerano, and J. S. Mattick. Ultraconserved elements in insect genomes: a highly conserved intronic sequence implicated in the control of homothorax mRNA splicing. *Genome Res.*, 15(6):800–808, Jun 2005.

J. P. Gogarten and J. P. Townsend. Horizontal gene transfer, genome innovation and evolution. *Nat. Rev. Microbiol.*, 3(9):679–687, Sep 2005.

L. Gross. Are "ultraconserved" genetic elements really indispensable? *PLoS Biol.*, 5(9):e253, Sep 2007.

L. Guelen, L. Pagie, E. Brasset, W. Meuleman, M. B. Faza, W. Talhout, B. H. Eussen, A. de Klein, L. Wessels, W. de Laat, and B. van Steensel. Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature*, 453(7197):948–951, Jun 2008.

D. L. Halligan, F. Oliver, J. Guthrie, K. C. Stemshorn, B. Harr, and P. D. Keightley. Positive and negative selection in murine ultraconserved noncoding elements. *Mol. Biol. Evol.*, 28(9): 2651–2660, Sep 2011.

N. Harmston, A. Baresic, and B. Lenhard. The mystery of extreme non-coding conservation. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.*, 368(1632):20130021, Dec 2013.

S. B. Hedges, J. Dudley, and S. Kumar. TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics*, 22(23):2971–2972, Dec 2006.

S. Heinz, C. Benner, N. Spann, E. Bertolino, Y. C. Lin, P. Laslo, J. X. Cheng, C. Murre, H. Singh, and C. K. Glass. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell*, 38(4):576–589, May 2010.

S. Henikoff, J. G. Henikoff, A. Sakai, G. B. Loeb, and K. Ahmad. Genome-wide profiling of salt fractions maps physical properties of chromatin. *Genome Res.*, 19(3):460–469, Mar 2009.

A. W. Holle and A. J. Engler. Cell rheology: Stressed-out stem cells. *Nat Mater*, 9(1):4–6, Jan 2010.

d. a. W. Huang, B. T. Sherman, and R. A. Lempicki. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*, 4(1):44–57, 2009.

Y. Ishii, S. Takeda, H. Yamada, and K. Oguri. Functional protein-protein interaction of drug metabolizing enzymes. *Front. Biosci.*, 10:887–895, Jan 2005.

D. T. Jones, W. R. Taylor, and J. M. Thornton. The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.*, 8(3):275–282, Jun 1992.

D. Karolchik, R. Baertsch, M. Diekhans, T. S. Furey, A. Hinrichs, Y. T. Lu, K. M. Roskin, M. Schwartz, C. W. Sugnet, D. J. Thomas, R. J. Weber, D. Haussler, and W. J. Kent. The UCSC Genome Browser Database. *Nucleic Acids Res.*, 31(1):51–54, Jan 2003.

S. Katzman, A. D. Kern, G. Bejerano, G. Fewell, L. Fulton, R. K. Wilson, S. R. Salama, and D. Haussler. Human genome ultraconserved elements are ultraselected. *Science*, 317(5840):915, Aug 2007.

M. Kellis, B. Wold, M. P. Snyder, B. E. Bernstein, A. Kundaje, G. K. Marinov, L. D. Ward, E. Birney, G. E. Crawford, J. Dekker, I. Dunham, L. L. Elnitski, P. J. Farnham, E. A. Feingold, M. Gerstein, M. C. Giddings, D. M. Gilbert, T. R. Gingeras, E. D. Green, R. Guigo, T. Hubbard, J. Kent, J. D. Lieb, R. M. Myers, M. J. Pazin, B. Ren, J. A. Stamatoyannopoulos, Z. Weng, K. P. White, and R. C. Hardison. Defining functional DNA elements in the human genome. *Proc. Natl. Acad. Sci. U.S.A.*, 111(17):6131–6138, Apr 2014.

E. Kenigsberg and A. Tanay. Drosophila functional elements are embedded in structurally constrained sequences. *PLoS Genet.*, 9(5):e1003512, May 2013.

E. Kenigsberg, A. Bar, E. Segal, and A. Tanay. Widespread compensatory evolution conserves DNA-encoded nucleosome organization in yeast. *PLoS Comput. Biol.*, 6(12):e1001039, 2010.

W. J. Kent and A. M. Zahler. Conservation, regulation, synteny, and introns in a large-scale C. briggsae-C. elegans genomic alignment. *Genome Res.*, 10(8):1115–1125, Aug 2000.

H. Kikuta, M. Laplante, P. Navratilova, A. Z. Komisarczuk, P. G. Engstrom, D. Fredman, A. Akalin, M. Caccamo, I. Sealy, K. Howe, J. Ghislain, G. Pezeron, P. Mourrain, S. Ellingsen, A. C. Oates, C. Thisse, B. Thisse, I. Foucher, B. Adolf, A. Geling, B. Lenhard, and T. S. Becker. Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. *Genome Res.*, 17(5):545–555, May 2007.

S. Y. Kim and J. K. Pritchard. Adaptive evolution of conserved noncoding elements in mammals. *PLoS Genet.*, 3(9):1572–1586, Sep 2007.

O. Kohany, A. J. Gentles, L. Hankus, and J. Jurka. Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. *BMC Bioinformatics*, 7:474, 2006.

K. Kritsas, S. E. Wuest, D. Hupalo, A. D. Kern, T. Wicker, and U. Grossniklaus. Computational analysis and characterization of UCE-like elements (ULEs) in plant genomes. *Genome Res.*, 22 (12):2455–2466, Dec 2012.

S. Kurtz, A. Phillippy, A. L. Delcher, M. Smoot, M. Shumway, C. Antonescu, and S. L. Salzberg. Versatile and open software for comparing large genomes. *Genome Biol.*, 5(2):R12, 2004.

M. Lachner, D. O'Carroll, S. Rea, K. Mechtler, and T. Jenuwein. Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins. *Nature*, 410(6824):116–120, Mar 2001.

M. T. Lam, W. Li, M. G. Rosenfeld, and C. K. Glass. Enhancer RNAs and regulated transcriptional programs. *Trends Biochem. Sci.*, 39(4):170–182, Apr 2014.

X. Lampe, O. A. Samad, A. Guiguen, C. Matis, S. Remacle, J. J. Picard, F. M. Rijli, and R. Rezsohazy. An ultraconserved Hox-Pbx responsive element resides in the coding sequence of Hoxa2 and is active in rhombomere 4. *Nucleic Acids Res.*, 36(10):3214–3225, Jun 2008.

E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda,

W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R. A. Gibbs, D. M. Muzny, S. E. Scherer, J. B. Bouck, E. J. Sodergren, K. C. Worley, C. M. Rives, J. H. Gorrell, M. L. Metzker, S. L. Naylor, R. S. Kucherlapati, D. L. Nelson, G. M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, D. R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H. M. Lee, J. Dubois, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R. W. Davis, N. A. Federspiel, A. P. Abola, M. J. Proctor, R. M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. R. Cox, M. V. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G. A. Evans, M. Athanasiou, R. Schultz, B. A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W. R. McCombie, M. de la Bastide, N. Dedhia, H. Blocker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D. G. Brown, C. B. Burge, L. Cerutti, H. C. Chen, D. Church, M. Clamp, R. R. Copley, T. Doerks, S. R. Eddy, E. E. Eichler, T. S. Furey, J. Galagan, J. G. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. S. Johnson, T. A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W. J. Kent, P. Kitts, E. V. Koonin, I. Korf, D. Kulp, D. Lancet, T. M. Lowe, A. McLysaght, T. Mikkelsen, J. V. Moran, N. Mulder, V. J. Pollara, C. P. Ponting, G. Schuler, J. Schultz, G. Slater, A. F. Smit, E. Stupka, J. Szustakowski, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y. I. Wolf, K. H. Wolfe, S. P. Yang, R. F. Yeh, F. Collins, M. S. Guyer, J. Peterson, A. Felsenfeld, K. A. Wetterstrand, A. Patrinos, M. J. Morgan, P. de Jong, J. J. Catanese, K. Osoegawa, H. Shizuya, S. Choi, Y. J. Chen, and J. Szustakowki. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, Feb 2001.

B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, 10(3):R25, 2009.

L. F. Lareau, M. Inada, R. E. Green, J. C. Wengrod, and S. E. Brenner. Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements. *Nature*, 446(7138): 926–929, Apr 2007.

M. S. Lawrence, P. Stojanov, P. Polak, G. V. Kryukov, K. Cibulskis, A. Sivachenko, S. L. Carter, C. Stewart, C. H. Mermel, S. A. Roberts, A. Kiezun, P. S. Hammerman, A. McKenna, Y. Drier, L. Zou, A. H. Ramos, T. J. Pugh, N. Stransky, E. Helman, J. Kim, C. Sougnez, L. Ambrogio, E. Nickerson, E. Shefler, M. L. Cortes, D. Auclair, G. Saksena, D. Voet, M. Noble, D. DiCara,

P. Lin, L. Lichtenstein, D. I. Heiman, T. Fennell, M. Imielinski, B. Hernandez, E. Hodis, S. Baca, A. M. Dulak, J. Lohr, D. A. Landau, C. J. Wu, J. Melendez-Zajgla, A. Hidalgo-Miranda, A. Koren, S. A. McCarroll, J. Mora, R. S. Lee, B. Crompton, R. Onofrio, M. Parkin, W. Winckler, K. Ardlie, S. B. Gabriel, C. W. Roberts, J. A. Biegel, K. Stegmaier, A. J. Bass, L. A. Garraway, M. Meyerson, T. R. Golub, D. A. Gordenin, S. Sunyaev, E. S. Lander, and G. Getz. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, 499(7457):214–218, Jul 2013.

T. I. Lee, R. G. Jenner, L. A. Boyer, M. G. Guenther, S. S. Levine, R. M. Kumar, B. Chevalier, S. E. Johnstone, M. F. Cole, K. Isono, H. Koseki, T. Fuchikami, K. Abe, H. L. Murray, J. P. Zucker, B. Yuan, G. W. Bell, E. Herbolsheimer, N. M. Hannett, K. Sun, D. T. Odom, A. P. Otte, T. L. Volkert, D. P. Bartel, D. A. Melton, D. K. Gifford, R. Jaenisch, and R. A. Young. Control of developmental regulators by Polycomb in human embryonic stem cells. *Cell*, 125(2):301–313, Apr 2006.

C. F. Lehner, R. Stick, H. M. Eppenberger, and E. A. Nigg. Differential expression of nuclear lamin proteins during chicken development. *J. Cell Biol.*, 105(1):577–587, Jul 1987.

B. Lenhard, A. Sandelin, and P. Carninci. Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nat. Rev. Genet.*, 13(4):233–245, Apr 2012.

M. Lin, C. Eng, E. T. Hawk, M. Huang, J. Lin, J. Gu, L. M. Ellis, and X. Wu. Identification of polymorphisms in ultraconserved elements associated with clinical outcomes in locally advanced colorectal adenocarcinoma. *Cancer*, 118(24):6188–6198, Dec 2012.

Z. Lin, H. Ma, and M. Nei. Ultraconserved coding regions outside the homeobox of mammalian Hox genes. *BMC Evol. Biol.*, 8:260, 2008.

K. Lindblad-Toh, C. M. Wade, T. S. Mikkelsen, E. K. Karlsson, D. B. Jaffe, M. Kamal, M. Clamp, J. L. Chang, E. J. Kulbokas, M. C. Zody, E. Mauceli, X. Xie, M. Breen, R. K. Wayne, E. A. Ostrander, C. P. Ponting, F. Galibert, D. R. Smith, P. J. DeJong, E. Kirkness, P. Alvarez, T. Biagi, W. Brockman, J. Butler, C. W. Chin, A. Cook, J. Cuff, M. J. Daly, D. DeCaprio, S. Gnerre, M. Grabherr, M. Kellis, M. Kleber, C. Bardeleben, L. Goodstadt, A. Heger, C. Hitte, L. Kim, K. P. Koepfli, H. G. Parker, J. P. Pollinger, S. M. Searle, N. B. Sutter, R. Thomas, C. Webber, J. Baldwin, A. Abebe, A. Abouelleil, L. Aftuck, M. Ait-Zahra, T. Aldredge, N. Allen, P. An, S. Anderson, C. Antoine, H. Arachchi, A. Aslam, L. Ayotte, P. Bachantsang, A. Barry, T. Bayul, M. Benamara, A. Berlin, D. Bessette, B. Blitshteyn, T. Bloom, J. Blye, L. Boguslavskiy, C. Bonnet, B. Boukhgalter, A. Brown, P. Cahill, N. Calixte, J. Camarata, Y. Cheshatsang, J. Chu, M. Citroen, A. Collymore, P. Cooke, T. Dawoe, R. Daza, K. Decktor, S. DeGray, N. Dhargay, K. Dooley, K. Dooley, P. Dorje, K. Dorjee, L. Dorris, N. Duffey, A. Dupes, O. Egbiremolen, R. Elong, J. Falk, A. Farina, S. Faro, D. Ferguson, P. Ferreira, S. Fisher, M. FitzGerald, K. Foley, C. Foley, A. Franke, D. Friedrich, D. Gage, M. Garber, G. Gearin, G. Giannoukos, T. Goode, A. Goyette, J. Graham, E. Grandbois, K. Gyaltsen, N. Hafez, D. Hagopian, B. Hagos, J. Hall, C. Healy, R. Hegarty, T. Honan, A. Horn, N. Houde, L. Hughes, L. Hunnicutt, M. Husby, B. Jester, C. Jones, A. Kamat, B. Kanga, C. Kells, D. Khazanovich, A. C. Kieu, P. Kisner, M. Kumar, K. Lance, T. Landers, M. Lara, W. Lee, J. P. Leger, N. Lennon, L. Leuper, S. LeVine, J. Liu, X. Liu, Y. Lokyitsang, T. Lokyitsang, A. Lui, J. Macdonald, J. Major, R. Marabella, K. Maru, C. Matthews, S. McDonough, T. Mehta, J. Meldrim, A. Melnikov, L. Meneus,

A. Mihalev, T. Mihova, K. Miller, R. Mittelman, V. Mlenga, L. Mulrain, G. Munson, A. Navidi, J. Naylor, T. Nguyen, N. Nguyen, C. Nguyen, T. Nguyen, R. Nicol, N. Norbu, C. Norbu, N. Novod, T. Nyima, P. Olandt, B. O'Neill, K. O'Neill, S. Osman, L. Oyono, C. Patti, D. Perrin, P. Phunkhang, F. Pierre, M. Priest, A. Rachupka, S. Raghuraman, R. Rameau, V. Ray, C. Raymond, F. Rege, C. Rise, J. Rogers, P. Rogov, J. Sahalie, S. Settipalli, T. Sharpe, T. Shea, M. Sheehan, N. Sherpa, J. Shi, D. Shih, J. Sloan, C. Smith, T. Sparrow, J. Stalker, N. Stange-Thomann, S. Stavropoulos, C. Stone, S. Stone, S. Sykes, P. Tchuinga, P. Tenzing, S. Tesfaye, D. Thoulutsang, Y. Thoulutsang, K. Topham, I. Topping, T. Tsamla, H. Vassiliev, V. Venkataraman, A. Vo, T. Wangchuk, T. Wangdi, M. Weiand, J. Wilkinson, A. Wilson, S. Yadav, S. Yang, X. Yang, G. Young, Q. Yu, J. Zainoun, L. Zembek, A. Zimmer, and E. S. Lander. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature*, 438(7069): 803–819, Dec 2005.

A. K. Linnemann, A. E. Platts, and S. A. Krawetz. Differential nuclear scaffold/matrix attachment marks expressed genes. *Hum. Mol. Genet.*, 18(4):645–654, Feb 2009.

M. E. Luderus, A. de Graaf, E. Mattia, J. L. den Blaauwen, M. A. Grande, L. de Jong, and R. van Driel. Binding of matrix attachment regions to lamin B1. *Cell*, 70(6):949–959, Sep 1992.

R. Lyne, R. Smith, K. Rutherford, M. Wakeling, A. Varley, F. Guillier, H. Janssens, W. Ji, P. Mclaren, P. North, D. Rana, T. Riley, J. Sullivan, X. Watkins, M. Woodbridge, K. Lilley, S. Russell, M. Ashburner, K. Mizuguchi, and G. Micklem. FlyMine: an integrated database for Drosophila and Anopheles genomics. *Genome Biol.*, 8(7):R129, 2007.

S. Mahony and P. V. Benos. STAMP: a web tool for exploring DNA-binding motif similarities. *Nucleic Acids Res.*, 35(Web Server issue):W253–258, Jul 2007.

F. Martinez, S. Monfort, M. Rosello, S. Oltra, D. Blesa, R. Quiroga, S. Mayo, and C. Orellana. Enrichment of ultraconserved elements among genomic imbalances causing mental delay and congenital anomalies. *BMC Med Genomics*, 3:54, 2010.

S. J. Marygold, P. C. Leyland, R. L. Seal, J. L. Goodman, J. Thurmond, V. B. Strelets, R. J. Wilson, W. Gelbart, K. Broll, L. Crosby, G. d. Santos, D. Emmert, K. Falls, L. Gramates, B. Matthews, S. Russo, A. Schroeder, S. S. Pierre, P. Zhou, M. Zytkovicz, N. H. Brown, B. Adryan, H. Attrill, M. Costa, H. Field, S. Marygold, P. McQuilton, G. Millburn, L. Ponting, D. Osumi-Sutherland, R. Stefancsik, S. Tweedie, T. Kaufman, K. Matthews, J. Goodman, G. Grumbling, V. Strelets, J. Thurmond, J. D. Wong, M. Werner-Washburne, R. Cripps, and H. Platero. FlyBase: improvements to the bibliography. *Nucleic Acids Res.*, 41(Database issue):D751–757, Jan 2013.

G. A. Maston, S. K. Evans, and M. R. Green. Transcriptional regulatory elements in the human genome. *Annu Rev Genomics Hum Genet*, 7:29–59, 2006.

C. McLean and G. Bejerano. Dispensability of mammalian DNA. *Genome Res.*, 18(11):1743–1751, Nov 2008.

C. Y. McLean, D. Bristor, M. Hiller, S. L. Clarke, B. T. Schaar, C. B. Lowe, A. M. Wenger, and G. Bejerano. GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.*, 28(5):495–501, May 2010.

I. Meier. Composition of the plant nuclear envelope: theme and variations. *J. Exp. Bot.*, 58(1): 27–34, 2007.

Patrick E. Meyer. *infotheo: Information-Theoretic Measures*, 2014. URL `http://CRAN.R-project.org/package=infotheo`. R package version 1.2.0.

K. Mousavi, H. Zare, S. Dell'orso, L. Grontved, G. Gutierrez-Cruz, A. Derfoul, G. L. Hager, and V. Sartorelli. eRNAs promote transcription by establishing chromatin accessibility at defined genomic loci. *Mol. Cell*, 51(5):606–617, Sep 2013.

C. A. Muller and C. A. Nieduszynski. Conservation of replication timing reveals global and local regulation of replication origin activity. *Genome Res.*, 22(10):1953–1962, Oct 2012.

R. J. Mural, M. D. Adams, E. W. Myers, H. O. Smith, G. L. Miklos, R. Wides, A. Halpern, P. W. Li, G. G. Sutton, J. Nadeau, S. L. Salzberg, R. A. Holt, C. D. Kodira, F. Lu, L. Chen, Z. Deng, C. C. Evangelista, W. Gan, T. J. Heiman, J. Li, Z. Li, G. V. Merkulov, N. V. Milshina, A. K. Naik, R. Qi, B. C. Shue, A. Wang, J. Wang, X. Wang, X. Yan, J. Ye, S. Yooseph, Q. Zhao, L. Zheng, S. C. Zhu, K. Biddick, R. Bolanos, A. L. Delcher, I. M. Dew, D. Fasulo, M. J. Flanigan, D. H. Huson, S. A. Kravitz, J. R. Miller, C. M. Mobarry, K. Reinert, K. A. Remington, Q. Zhang, X. H. Zheng, D. R. Nusskern, Z. Lai, Y. Lei, W. Zhong, A. Yao, P. Guan, R. R. Ji, Z. Gu, Z. Y. Wang, F. Zhong, C. Xiao, C. C. Chiang, M. Yandell, J. R. Wortman, P. G. Amanatides, S. L. Hladun, E. C. Pratts, J. E. Johnson, K. L. Dodson, K. J. Woodford, C. A. Evans, B. Gropman, D. B. Rusch, E. Venter, M. Wang, T. J. Smith, J. T. Houck, D. E. Tompkins, C. Haynes, D. Jacob, S. H. Chin, D. R. Allen, C. E. Dahlke, R. Sanders, K. Li, X. Liu, A. A. Levitsky, W. H. Majoros, Q. Chen, A. C. Xia, J. R. Lopez, M. T. Donnelly, M. H. Newman, A. Glodek, C. L. Kraft, M. Nodell, F. Ali, H. J. An, D. Baldwin-Pitts, K. Y. Beeson, S. Cai, M. Carnes, A. Carver, P. M. Caulk, A. Center, Y. H. Chen, M. L. Cheng, M. D. Coyne, M. Crowder, S. Danaher, L. B. Davenport, R. Desilets, S. M. Dietz, L. Doup, P. Dullaghan, S. Ferriera, C. R. Fosler, H. C. Gire, A. Gluecksmann, J. D. Gocayne, J. Gray, B. Hart, J. Haynes, J. Hoover, T. Howland, C. Ibegwam, M. Jalali, D. Johns, L. Kline, D. S. Ma, S. MacCawley, A. Magoon, F. Mann, D. May, T. C. McIntosh, S. Mehta, L. Moy, M. C. Moy, B. J. Murphy, S. D. Murphy, K. A. Nelson, Z. Nuri, K. A. Parker, A. C. Prudhomme, V. N. Puri, H. Qureshi, J. C. Raley, M. S. Reardon, M. A. Regier, Y. H. Rogers, D. L. Romblad, J. Schutz, J. L. Scott, R. Scott, C. D. Sitter, M. Smallwood, A. C. Sprague, E. Stewart, R. V. Strong, E. Suh, K. Sylvester, R. Thomas, N. N. Tint, C. Tsonis, G. Wang, G. Wang, M. S. Williams, S. M. Williams, S. M. Windsor, K. Wolfe, M. M. Wu, J. Zaveri, K. Chaturvedi, A. E. Gabrielian, Z. Ke, J. Sun, G. Subramanian, J. C. Venter, C. M. Pfannkoch, M. Barnstead, and L. D. Stephenson. A comparison of whole-genome shotgun-derived mouse chromosome 16 and the human genome. *Science*, 296(5573):1661–1671, May 2002.

N. Negre, C. D. Brown, L. Ma, C. A. Bristow, S. W. Miller, U. Wagner, P. Kheradpour, M. L. Eaton, P. Loriaux, R. Sealfon, Z. Li, H. Ishii, R. F. Spokony, J. Chen, L. Hwang, C. Cheng, R. P. Auburn, M. B. Davis, M. Domanus, P. K. Shah, C. A. Morrison, J. Zieba, S. Suchy, L. Senderowicz, A. Victorsen, N. A. Bild, A. J. Grundstad, D. Hanley, D. M. MacAlpine, M. Mannervik, K. Venken, H. Bellen, R. White, M. Gerstein, S. Russell, R. L. Grossman, B. Ren, J. W. Posakony, M. Kellis, and K. P. White. A cis-regulatory map of the Drosophila genome. *Nature*, 471(7339):527–531, Mar 2011.

J. Z. Ni, L. Grate, J. P. Donohue, C. Preston, N. Nobida, G. O'Brien, L. Shiue, T. A. Clark, J. E. Blume, and M. Ares. Ultraconserved elements are associated with homeostatic control of splicing regulators by alternative splicing and nonsense-mediated decay. *Genes Dev.*, 21(6):708–718, Mar 2007.

M. A. Nobrega, I. Ovcharenko, V. Afzal, and E. M. Rubin. Scanning human gene deserts for long-range enhancers. *Science*, 302(5644):413, Oct 2003.

U. Ohler. Identification of core promoter modules in Drosophila and their application in accurate transcription start site prediction. *Nucleic Acids Res.*, 34(20):5943–5950, 2006.

Jan Oosting, Paul Eilers, and Renee Menezes. *quantsmooth: Quantile smoothing and genomic visualization of array data*, r package version 1.12.0 edition, 2009.

G. Ostlund, T. Schmitt, K. Forslund, T. Kostler, D. N. Messina, S. Roopra, O. Frings, and E. L. Sonnhammer. InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Res.*, 2010.

I. Ovcharenko. Widespread ultraconservation divergence in primates. *Mol. Biol. Evol.*, 25(8):1668–1676, Aug 2008.

K. C. Pang, M. C. Frith, and J. S. Mattick. Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. *Trends Genet.*, 22(1):1–5, Jan 2006.

D. Papatsenko, A. Kislyuk, M. Levine, and I. Dubchak. Conservation patterns in different functional sequence categories of divergent Drosophila species. *Genomics*, 88(4):431–442, Oct 2006.

E. Pefanis, J. Wang, G. Rothschild, J. Lim, D. Kazadi, J. Sun, A. Federation, J. Chao, O. Elliott, Z. P. Liu, A. N. Economides, J. E. Bradner, R. Rabadan, and U. Basu. RNA exosome-regulated long non-coding RNA transcription controls super-enhancer activity. *Cell*, 161(4):774–789, May 2015.

L. A. Pennacchio, N. Ahituv, A. M. Moses, S. Prabhakar, M. A. Nobrega, M. Shoukry, S. Minovitsky, I. Dubchak, A. Holt, K. D. Lewis, I. Plajzer-Frick, J. Akiyama, S. De Val, V. Afzal, B. L. Black, O. Couronne, M. B. Eisen, A. Visel, and E. M. Rubin. In vivo enhancer analysis of human conserved non-coding sequences. *Nature*, 444(7118):499–502, Nov 2006.

L. Poitras, M. Yu, C. Lesage-Pelletier, R. B. Macdonald, J. P. Gagne, G. Hatch, I. Kelly, S. P. Hamilton, J. L. Rubenstein, G. G. Poirier, and M. Ekker. An SNP in an ultraconserved regulatory element affects Dlx5/Dlx6 regulation in the forebrain. *Development*, 137(18):3089–3097, Sep 2010.

S. G. Prasanth, Z. Shen, K. V. Prasanth, and B. Stillman. Human origin recognition complex is essential for HP1 binding to chromatin and heterochromatin organization. *Proc. Natl. Acad. Sci. U.S.A.*, 107(34):15093–15098, Aug 2010.

M. Prokocimer, M. Davidovich, M. Nissim-Rafinia, N. Wiesel-Motiuk, D. Z. Bar, R. Barkan, E. Meshorer, and Y. Gruenbaum. Nuclear lamins: key regulators of nuclear structure and activities. *J. Cell. Mol. Med.*, 13(6):1059–1085, Jun 2009.

K. D. Pruitt, T. Tatusova, W. Klimke, and D. R. Maglott. NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic Acids Res.*, 37(Database issue):D32–36, Jan 2009.

N. H. Putnam, M. Srivastava, U. Hellsten, B. Dirks, J. Chapman, A. Salamov, A. Terry, H. Shapiro, E. Lindquist, V. V. Kapitonov, J. Jurka, G. Genikhovich, I. V. Grigoriev, S. M. Lucas, R. E. Steele, J. R. Finnerty, U. Technau, M. Q. Martindale, and D. S. Rokhsar. Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *Science*, 317(5834):86–94, Jul 2007.

L. A. Quina, T. Kuramoto, D. V. Luquetti, T. C. Cox, T. Serikawa, and E. E. Turner. Deletion of a conserved regulatory element required for Hmx1 expression in craniofacial mesenchyme in the dumbo rat: a newly identified cause of congenital ear malformation. *Dis Model Mech*, 5(6): 812–822, Nov 2012.

A. R. Quinlan and I. M. Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, Mar 2010.

M. D. Robinson, D. J. McCarthy, and G. K. Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, Jan 2010.

C. Rodelsperger, S. Kohler, M. H. Schulz, T. Manke, S. Bauer, and P. N. Robinson. Short ultraconserved promoter regions delineate a class of preferentially expressed alternatively spliced transcripts. *Genomics*, 94(5):308–316, Nov 2009.

A. C. Roth, G. H. Gonnet, and C. Dessimoz. Algorithm of OMA for large-scale orthology inference. *BMC Bioinformatics*, 9:518, 2008.

R. K. Rowntree and J. T. Lee. Mapping of DNA replication origins to noncoding genes of the X-inactivation center. *Mol. Cell. Biol.*, 26(10):3707–3717, May 2006.

J. L. Royo, C. Hidalgo, Y. Roncero, M. A. Seda, A. Akalin, B. Lenhard, F. Casares, and J. L. Gomez-Skarmeta. Dissecting the transcriptional regulatory properties of human chromosome 16 highly conserved non-coding regions. *PLoS ONE*, 6(9):e24824, 2011.

T. Ryu, C. H. Mavromatis, T. Bayer, C. R. Voolstra, and T. Ravasi. Unexpected complexity of the reef-building coral Acropora millepora transcription factor network. *BMC Syst Biol*, 5:58, 2011.

T. Ryu, L. Seridi, and T. Ravasi. The evolution of ultraconserved elements with different phylogenetic origins. *BMC Evol. Biol.*, 12:236, 2012.

M. Safran, I. Dalah, J. Alexander, N. Rosen, T. Iny Stein, M. Shmoish, N. Nativ, I. Bahir, T. Doniger, H. Krug, A. Sirota-Madi, T. Olender, Y. Golan, G. Stelzer, A. Harel, and D. Lancet. GeneCards Version 3: the human gene integrator. *Database (Oxford)*, 2010:baq020, 2010.

V. Sahagun and J. M. Ranz. Characterization of genomic regulatory domains conserved across the genus Drosophila. *Genome Biol Evol*, 4(10):1054–1060, 2012.

N. Saitou and M. Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, 4(4):406–425, Jul 1987.

Y. Sakuraba, T. Kimura, H. Masuya, H. Noguchi, H. Sezutsu, K. R. Takahasi, A. Toyoda, R. Fukumura, T. Murata, Y. Sakaki, M. Yamamura, S. Wakana, T. Noda, T. Shiroishi, and Y. Gondo. Identification and characterization of new long conserved noncoding sequences in vertebrates. *Mamm. Genome*, 19(10-12):703–712, 2008.

I. Schneider. Cell lines derived from late embryonic stages of Drosophila melanogaster. *J Embryol Exp Morphol*, 27(2):353–365, Apr 1972.

J. T. Shin, J. R. Priest, I. Ovcharenko, A. Ronco, R. K. Moore, C. G. Burns, and C. A. MacRae. Human-zebrafish non-coding conserved elements act in vivo to regulate transcription. *Nucleic Acids Res.*, 33(17):5437–5445, 2005.

A. Siepel, G. Bejerano, J. S. Pedersen, A. S. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L. W. Hillier, S. Richards, G. M. Weinstock, R. K. Wilson, R. A. Gibbs, W. J. Kent, W. Miller, and D. Haussler. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, 15(8):1034–1050, Aug 2005.

E. Sodergren, G. M. Weinstock, E. H. Davidson, R. A. Cameron, R. A. Gibbs, R. C. Angerer, L. M. Angerer, M. I. Arnone, D. R. Burgess, R. D. Burke, J. A. Coffman, M. Dean, M. R. Elphick, C. A. Ettensohn, K. R. Foltz, A. Hamdoun, R. O. Hynes, W. H. Klein, W. Marzluff, D. R. McClay, R. L. Morris, A. Mushegian, J. P. Rast, L. C. Smith, M. C. Thorndyke, V. D. Vacquier, G. M. Wessel, G. Wray, L. Zhang, C. G. Elsik, O. Ermolaeva, W. Hlavina, G. Hofmann, P. Kitts, M. J. Landrum, A. J. Mackey, D. Maglott, G. Panopoulou, A. J. Poustka, K. Pruitt, V. Sapojnikov, X. Song, A. Souvorov, V. Solovyev, Z. Wei, C. A. Whittaker, K. Worley, K. J. Durbin, Y. Shen, O. Fedrigo, D. Garfield, R. Haygood, A. Primus, R. Satija, T. Severson, M. L. Gonzalez-Garay, A. R. Jackson, A. Milosavljevic, M. Tong, C. E. Killian, B. T. Livingston, F. H. Wilt, N. Adams, R. Belle, S. Carbonneau, R. Cheung, P. Cormier, B. Cosson, J. Croce, A. Fernandez-Guerra, A. M. Geneviere, M. Goel, H. Kelkar, J. Morales, O. Mulner-Lorillon, A. J. Robertson, J. V. Goldstone, B. Cole, D. Epel, B. Gold, M. E. Hahn, M. Howard-Ashby, M. Scally, J. J. Stegeman, E. L. Allgood, J. Cool, K. M. Judkins, S. S. McCafferty, A. M. Musante, R. A. Obar, A. P. Rawson, B. J. Rossetti, I. R. Gibbons, M. P. Hoffman, A. Leone, S. Istrail, S. C. Materna, M. P. Samanta, V. Stolc, W. Tongprasit, Q. Tu, K. F. Bergeron, B. P. Brandhorst, J. Whittle, K. Berney, D. J. Bottjer, C. Calestani, K. Peterson, E. Chow, Q. A. Yuan, E. Elhaik, D. Graur, J. T. Reese, I. Bosdet, S. Heesun, M. A. Marra, J. Schein, M. K. Anderson, V. Brockton, K. M. Buckley, A. H. Cohen, S. D. Fugmann, T. Hibino, M. Loza-Coll, A. J. Majeske, C. Messier, S. V. Nair, Z. Pancer, D. P. Terwilliger, C. Agca, E. Arboleda, N. Chen, A. M. Churcher, F. Hallbook, G. W. Humphrey, M. M. Idris, T. Kiyama, S. Liang, D. Mellott, X. Mu, G. Murray, R. P. Olinski, F. Raible, M. Rowe, J. S. Taylor, K. Tessmar-Raible, D. Wang, K. H. Wilson, S. Yaguchi, T. Gaasterland, B. E. Galindo, H. J. Gunaratne, C. Juliano, M. Kinukawa, G. W. Moy, A. T. Neill, M. Nomura, M. Raisch, A. Reade, M. M. Roux, J. L. Song, Y. H. Su, I. K. Townley, E. Voronina, J. L. Wong, G. Amore, M. Branno, E. R. Brown, V. Cavalieri, V. Duboc, L. Duloquin, C. Flytzanis, C. Gache, F. Lapraz, T. Lepage, A. Locascio, P. Martinez, G. Matassi, V. Matranga, R. Range, F. Rizzo, E. Rottinger, W. Beane, C. Bradham, C. Byrum, T. Glenn, S. Hussain, G. Manning, E. Miranda, R. Thomason, K. Walton, A. Wikramanayke, S. Y. Wu, R. Xu, C. T. Brown, L. Chen, R. F. Gray, P. Y. Lee, J. Nam, P. Oliveri, J. Smith, D. Muzny, S. Bell, J. Chacko, A. Cree, S. Curry, C. Davis, H. Dinh,

S. Dugan-Rocha, J. Fowler, R. Gill, C. Hamilton, J. Hernandez, S. Hines, J. Hume, L. Jackson, A. Jolivet, C. Kovar, S. Lee, L. Lewis, G. Miner, M. Morgan, L. V. Nazareth, G. Okwuonu, D. Parker, L. L. Pu, R. Thorn, and R. Wright. The genome of the sea urchin Strongylocentrotus purpuratus. *Science*, 314(5801):941–952, Nov 2006.

R. Sorek and G. Ast. Intronic sequences flanking alternatively spliced exons are conserved between human and mouse. *Genome Res.*, 13(7):1631–1637, Jul 2003.

F. Spitz and E. E. Furlong. Transcription factors: from enhancer binding to developmental control. *Nat. Rev. Genet.*, 13(9):613–626, Sep 2012.

M. Srivastava, O. Simakov, J. Chapman, B. Fahey, M. E. Gauthier, T. Mitros, G. S. Richards, C. Conaco, M. Dacre, U. Hellsten, C. Larroux, N. H. Putnam, M. Stanke, M. Adamska, A. Darling, S. M. Degnan, T. H. Oakley, D. C. Plachetzki, Y. Zhai, M. Adamski, A. Calcino, S. F. Cummins, D. M. Goodstein, C. Harris, D. J. Jackson, S. P. Leys, S. Shu, B. J. Woodcroft, M. Vervoort, K. S. Kosik, G. Manning, B. M. Degnan, and D. S. Rokhsar. The Amphimedon queenslandica genome and the evolution of animal complexity. *Nature*, 466(7307):720–726, Aug 2010.

S. Stephen, M. Pheasant, I. V. Makunin, and J. S. Mattick. Large-scale appearance of ultraconserved elements in tetrapod genomes and slowdown of the molecular clock. *Mol. Biol. Evol.*, 25(2):402–408, Feb 2008.

H. Stroud, S. Otero, B. Desvoyes, E. Ramirez-Parra, S. E. Jacobsen, and C. Gutierrez. Genome-wide analysis of histone H3.1 and H3.3 variants in Arabidopsis thaliana. *Proc. Natl. Acad. Sci. U.S.A.*, 109(14):5370–5375, Apr 2012.

K. Tamura, D. Peterson, N. Peterson, G. Stecher, M. Nei, and S. Kumar. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.*, 28(10):2731–2739, Oct 2011.

V. B. Teif, Y. Vainshtein, M. Caudron-Herger, J. P. Mallm, C. Marth, T. Hofer, and K. Rippe. Genome-wide nucleosome positioning during embryonic stem cell development. *Nat. Struct. Mol. Biol.*, 19(11):1185–1192, Nov 2012.

I. V. Tetko, G. Haberer, S. Rudd, B. Meyers, H. W. Mewes, and K. F. Mayer. Spatiotemporal expression control correlates with intragenic scaffold matrix attachment regions (S/MARs) in Arabidopsis thaliana. *PLoS Comput. Biol.*, 2(3):e21, Mar 2006.

R. B. Tracy, J. K. Baumohl, and S. C. Kowalczykowski. The preference for GT-rich DNA by the yeast Rad51 protein defines a set of universal pairing sequences. *Genes Dev.*, 11(24):3423–3431, Dec 1997.

H. H. Tseng and M. Tompa. Algorithms for locating extremely conserved elements in multiple sequence alignments. *BMC Bioinformatics*, 10:432, 2009.

J. G. van Bemmel, L. Pagie, U. Braunschweig, W. Brugman, W. Meuleman, R. M. Kerkhoven, and B. van Steensel. The insulator protein SU(HW) fine-tunes nuclear lamina interactions of the Drosophila genome. *PLoS ONE*, 5(11):e15013, 2010.

T. Vavouri, K. Walter, W. R. Gilks, B. Lehner, and G. Elgar. Parallel evolution of conserved non-coding elements that target a common set of developmental regulatory genes from worms to humans. *Genome Biol.*, 8(2):R15, 2007.

A. Visel, S. Prabhakar, J. A. Akiyama, M. Shoukry, K. D. Lewis, A. Holt, I. Plajzer-Frick, V. Afzal, E. M. Rubin, and L. A. Pennacchio. Ultraconservation identifies a small subset of extremely constrained developmental enhancers. *Nat. Genet.*, 40(2):158–160, Feb 2008.

M. von Grotthuss, M. Ashburner, and J. M. Ranz. Fragile regions and not functional constraints predominate in shaping gene organization in the genus Drosophila. *Genome Res.*, 20(8):1084–1096, Aug 2010.

K. Walter, I. Abnizova, G. Elgar, and W. R. Gilks. Striking nucleotide frequency pattern at the borders of highly conserved vertebrate non-coding sequences. *Trends Genet.*, 21(8):436–440, Aug 2005.

J. Wang, A. P. Lee, R. Kodzius, S. Brenner, and B. Venkatesh. Large number of ultraconserved elements were already present in the jawed vertebrate ancestor. *Mol. Biol. Evol.*, 26(3):487–490, Mar 2009.

W. W. Wasserman and A. Sandelin. Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.*, 5(4):276–287, Apr 2004.

S. J. Westenberger, L. Cui, N. Dharia, E. Winzeler, and L. Cui. Genome-wide nucleosome mapping of Plasmodium falciparum reveals histone-rich coding and histone-poor intergenic regions and chromatin remodeling of core and subtelomeric genes. *BMC Genomics*, 10:610, 2009.

C. Wirbelauer, O. Bell, and D. Schubeler. Variant histone H3.3 is deposited at sites of nucleosomal displacement throughout transcribed genes while active histone modifications show a promoter-proximal bias. *Genes Dev.*, 19(15):1761–1766, Aug 2005.

A. Woolfe, M. Goodson, D. K. Goode, P. Snell, G. K. McEwen, T. Vavouri, S. F. Smith, P. North, H. Callaway, K. Kelly, K. Walter, I. Abnizova, W. Gilks, Y. J. Edwards, J. E. Cooke, and G. Elgar. Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol.*, 3(1):e7, Jan 2005.

S. Xiao, D. Xie, X. Cao, P. Yu, X. Xing, C. C. Chen, M. Musselman, M. Xie, F. D. West, H. A. Lewin, T. Wang, and S. Zhong. Comparative epigenomic annotation of regulatory DNA. *Cell*, 149(6):1381–1392, Jun 2012.

R. Yang, B. Frank, K. Hemminki, C. R. Bartram, B. Wappenschmidt, C. Sutter, M. Kiechle, P. Bugert, R. K. Schmutzler, N. Arnold, B. H. Weber, D. Niederacher, A. Meindl, and B. Burwinkel. SNPs in ultraconserved elements and familial breast cancer risk. *Carcinogenesis*, 29(2): 351–355, Feb 2008.

E. M. Zdobnov and R. Apweiler. InterProScan–an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*, 17(9):847–848, Sep 2001.

W. X. Zheng and C. T. Zhang. Ultraconserved elements between the genomes of the plants Arabidopsis thaliana and rice. *J. Biomol. Struct. Dyn.*, 26(1):1–8, Aug 2008.

I. B. Zovkic, B. S. Paulukaitis, J. J. Day, D. M. Etikala, and J. D. Sweatt. Histone H2A.Z subunit exchange controls consolidation of recent and remote memory. *Nature*, 515(7528):582–586, Nov 2014.

A Appendix of Dynamic epigenetic control of highly conserved noncoding elements

Table A.1: Table mapping the top 50 overrepresented heptamers to known transcription factor binding sites.

Heptamer	**Counts**	**Expected Counts**	**Dispersion Index**	**Binomial P-value**	**Best match**	**Match E-value**	**Developmental TF**
CAACTGC	183	7.27×10^{1}	1.19×10^{-1}	4.33×10^{-27}	vnd	1.42×10^{-4}	TRUE
ACACACA	601	3.78×10^{2}	1.16×10^{-1}	3.76×10^{-26}	Top2	1.34×10^{-4}	
AGATACA	150	6.94×10^{1}	1.18×10^{-1}	7.84×10^{-17}	Top2	2.15×10^{-4}	
TGCAACA	196	8.41×10^{1}	10.00×10^{-2}	4.22×10^{-25}	suHw	2.89×10^{-4}	
CTGCAAC	140	5.76×10^{1}	1.18×10^{-1}	8.00×10^{-20}	suHw	2.89×10^{-4}	
GGCAGCA	204	7.88×10^{1}	1.34×10^{-1}	1.41×10^{-31}	shn	8.29×10^{-3}	TRUE
CAACGGC	106	3.80×10^{1}	7.51×10^{-2}	3.46×10^{-19}	pho	3.62×10^{-4}	
CAGTTGC	164	6.40×10^{1}	1.27×10^{-1}	3.57×10^{-25}	ovo	7.48×10^{-4}	TRUE
CAGATAC	140	5.16×10^{1}	1.44×10^{-1}	6.75×10^{-24}	ovo	1.42×10^{-3}	TRUE
CAGCAGC	631	1.43×10^{2}	1.78×10^{-1}	3.59×10^{-196}	Med	6.48×10^{-3}	TRUE
CAGCGGC	174	4.68×10^{1}	9.93×10^{-2}	1.49×10^{-45}	Med	5.34×10^{-3}	TRUE
CGGCAGC	167	5.18×10^{1}	1.08×10^{-1}	1.99×10^{-36}	Mad	8.87×10^{-4}	TRUE
CGTCGTC	102	3.44×10^{1}	2.46×10^{-1}	3.03×10^{-20}	Mad	1.48×10^{-3}	TRUE
CGACGAC	124	4.82×10^{1}	1.10×10^{-1}	1.51×10^{-19}	Mad	8.10×10^{-3}	TRUE
CGACGTC	67	1.85×10^{1}	1.79×10^{-1}	1.12×10^{-17}	Mad	1.49×10^{-3}	TRUE
CGGCGAC	82	2.42×10^{1}	1.33×10^{-1}	9.62×10^{-20}	ftz-f1	3.26×10^{-5}	TRUE

Heptamer	Occurences	Expected Occurences	Dispersion Index	Binomial P-value	Best Match	Match E-value	Developmental TF
CACACAC	596	3.58×10^{2}	1.26×10^{-1}	1.30×10^{-30}	ey	6.64×10^{-4}	TRUE
ACAGCAA	261	9.44×10^{1}	1.03×10^{-1}	1.25×10^{-44}	Epi74EF	8.30×10^{-5}	
AGCGGCA	176	6.02×10^{1}	8.33×10^{-2}	2.73×10^{-33}	brk	3.92×10^{-4}	TRUE
GCGGCAA	166	6.03×10^{1}	8.66×10^{-2}	1.09×10^{-28}	brk	3.35×10^{-3}	TRUE
GGCGGCA	111	3.93×10^{1}	8.27×10^{-2}	1.99×10^{-20}	brk	7.54×10^{-5}	TRUE
ACGACGA	107	3.45×10^{1}	1.99×10^{-1}	1.44×10^{-22}	bin	1.09×10^{-3}	TRUE
AACAACG	111	4.17×10^{1}	1.34×10^{-1}	1.21×10^{-18}	bin	1.89×10^{-7}	TRUE
ACAACGA	109	4.29×10^{1}	9.89×10^{-2}	5.44×10^{-17}	bin	1.89×10^{-7}	TRUE
AGCAGCG	124	5.26×10^{1}	1.14×10^{-1}	1.03×10^{-16}	bin	1.14×10^{-2}	TRUE
AGCAGCA	594	1.57×10^{2}	1.43×10^{-1}	6.32×10^{-155}	Aef1	1.53×10^{-2}	
CAGCAAC	444	9.97×10^{1}	9.94×10^{-2}	3.39×10^{-140}	Aef1	5.02×10^{-6}	
CAACAAC	516	1.48×10^{2}	1.26×10^{-1}	2.50×10^{-122}	Aef1	4.88×10^{-11}	
AGCAACA	451	1.21×10^{2}	8.18×10^{-2}	8.19×10^{-116}	Aef1	1.24×10^{-5}	
CAACAGC	314	7.67×10^{1}	9.07×10^{-2}	7.71×10^{-91}	Aef1	5.20×10^{-6}	
AACAGCA	322	8.96×10^{1}	1.07×10^{-1}	1.09×10^{-79}	Aef1	1.24×10^{-5}	
AACAACA	517	2.02×10^{2}	1.03×10^{-1}	7.37×10^{-76}	Aef1	1.79×10^{-10}	
ACAACAA	499	1.93×10^{2}	1.11×10^{-1}	1.04×10^{-74}	Aef1	4.12×10^{-7}	
GCAACAA	408	1.54×10^{2}	8.08×10^{-2}	4.36×10^{-64}	Aef1	4.12×10^{-7}	
GCAGCAA	362	1.31×10^{2}	8.63×10^{-2}	1.68×10^{-61}	Aef1	2.02×10^{-3}	
CTGTTGC	211	5.68×10^{1}	1.59×10^{-1}	6.25×10^{-55}	Aef1	3.51×10^{-4}	
CGACAAC	145	3.08×10^{1}	2.02×10^{-1}	2.17×10^{-49}	Aef1	5.02×10^{-6}	
ATGTTGC	206	6.82×10^{1}	1.42×10^{-1}	1.07×10^{-40}	Aef1	3.50×10^{-4}	
CGGCAAC	145	4.02×10^{1}	1.09×10^{-1}	8.91×10^{-37}	Aef1	9.09×10^{-3}	
ACGACAA	133	3.49×10^{1}	1.46×10^{-1}	3.89×10^{-36}	Aef1	2.02×10^{-3}	
GGCAACA	213	8.85×10^{1}	1.04×10^{-1}	7.00×10^{-29}	Aef1	3.51×10^{-4}	
AACGACA	106	3.31×10^{1}	1.55×10^{-1}	2.08×10^{-23}	Aef1	1.85×10^{-5}	
ACAACAG	146	5.79×10^{1}	1.19×10^{-1}	5.95×10^{-22}	Aef1	1.12×10^{-6}	
CAACATC	130	4.91×10^{1}	1.52×10^{-1}	2.12×10^{-21}	Aef1	5.02×10^{-6}	
CAACGAC	102	3.32×10^{1}	1.43×10^{-1}	2.47×10^{-21}	Aef1	7.63×10^{-6}	
ACAGCAG	162	7.07×10^{1}	1.07×10^{-1}	2.75×10^{-20}	Aef1	3.72×10^{-3}	

Heptamer	Occurences	Expected Occurences	Dispersion Index	Binomial P-value	Best Match	Match E-value	Developmental TF
AACATCA	127	5.21×10^{1}	1.84×10^{-1}	3.56×10^{-18}	Aef1	1.24×10^{-5}	
CAACATG	113	4.54×10^{1}	1.54×10^{-1}	6.63×10^{-17}	Aef1	3.51×10^{-4}	
CAACAAA	301	1.78×10^{2}	6.67×10^{-2}	7.28×10^{-17}	Aef1	2.12×10^{-7}	
CTGCTGC	530	1.36×10^{2}	1.87×10^{-1}	2.75×10^{-143}	Adf1	1.20×10^{-2}	

Table A.2: Table listing the Gene Ontology enrichment (biological processes only) among the HCNE-proximal genes.

Promoter Type	GO term	Holm-Bonferroni adjusted P-value
Inr-only	regulation of transcription, DNA-templated	1.19×10^{-16}
Inr-only	regulation of RNA biosynthetic process	1.19×10^{-16}
Inr-only	transcription, DNA-templated	2.98×10^{-15}
Inr-only	RNA biosynthetic process	3.45×10^{-15}
Inr-only	regulation of RNA metabolic process	5.34×10^{-15}
Inr-only	regulation of macromolecule biosynthetic process	1.32×10^{-14}
Inr-only	regulation of cellular macromolecule biosynthetic process	1.32×10^{-14}
Inr-only	regulation of cellular biosynthetic process	2.39×10^{-14}
Inr-only	regulation of biosynthetic process	2.50×10^{-14}
Inr-only	nucleobase-containing compound biosynthetic process	4.36×10^{-13}
Inr-only	regulation of nitrogen compound metabolic process	5.82×10^{-13}
Inr-only	regulation of nucleobase-containing compound metabolic process	6.00×10^{-13}
Inr-only	cellular nitrogen compound biosynthetic process	9.03×10^{-13}
Inr-only	organic cyclic compound biosynthetic process	1.39×10^{-12}
Inr-only	aromatic compound biosynthetic process	1.45×10^{-12}
Inr-only	heterocycle biosynthetic process	1.63×10^{-12}
Inr-only	regulation of transcription from RNA polymerase II promoter	1.80×10^{-12}
Inr-only	leg disc development	5.41×10^{-12}
Inr-only	regulation of gene expression	7.96×10^{-12}
Inr-only	imaginal disc development	6.30×10^{-11}

Promoter Type	**GO term**	**Holm-Bonferroni adjusted P-value**
Inr-only	organ development	6.44×10^{-11}
Inr-only	regionalization	7.21×10^{-11}
Inr-only	system development	1.11×10^{-10}
Inr-only	transcription from RNA polymerase II promoter	1.41×10^{-10}
Inr-only	imaginal disc-derived appendage development	2.87×10^{-10}
Inr-only	appendage development	3.49×10^{-10}
Inr-only	regulation of primary metabolic process	4.43×10^{-10}
Inr-only	anatomical structure development	6.12×10^{-10}
Inr-only	anatomical structure morphogenesis	6.34×10^{-10}
Inr-only	pattern specification process	9.91×10^{-10}
Inr-only	regulation of macromolecule metabolic process	1.07×10^{-9}
Inr-only	imaginal disc-derived appendage morphogenesis	1.38×10^{-9}
Inr-only	appendage morphogenesis	1.68×10^{-9}
Inr-only	regulation of cellular metabolic process	4.35×10^{-9}
Inr-only	post-embryonic appendage morphogenesis	5.08×10^{-9}
Inr-only	imaginal disc morphogenesis	7.32×10^{-9}
Inr-only	post-embryonic organ morphogenesis	7.32×10^{-9}
Inr-only	post-embryonic organ development	7.91×10^{-9}
Inr-only	organ morphogenesis	8.24×10^{-9}
Inr-only	proximal/distal pattern formation	1.51×10^{-8}
Inr-only	leg disc pattern formation	1.94×10^{-8}
Inr-only	RNA metabolic process	2.10×10^{-8}
Inr-only	single-organism developmental process	3.25×10^{-8}
Inr-only	imaginal disc pattern formation	3.89×10^{-8}
Inr-only	biological regulation	4.41×10^{-8}
Inr-only	tissue development	6.66×10^{-8}
Inr-only	instar larval or pupal development	6.81×10^{-8}
Inr-only	regulation of metabolic process	7.07×10^{-8}
Inr-only	developmental process	7.40×10^{-8}
Inr-only	post-embryonic development	7.53×10^{-8}

Promoter Type	**GO term**	**Holm-Bonferroni adjusted P-value**
Inr-only	cell fate specification	8.12×10^{-8}
Inr-only	regulation of cellular process	8.52×10^{-8}
Inr-only	proximal/distal pattern formation, imaginal disc	1.00×10^{-7}
Inr-only	negative regulation of transcription from RNA polymerase II promoter	1.02×10^{-7}
Inr-only	post-embryonic morphogenesis	1.24×10^{-7}
Inr-only	regulation of biological process	1.32×10^{-7}
Inr-only	instar larval or pupal morphogenesis	3.36×10^{-7}
Inr-only	leg disc proximal/distal pattern formation	3.51×10^{-7}
Inr-only	multicellular organismal development	6.13×10^{-7}
Inr-only	metamorphosis	1.15×10^{-6}
Inr-only	cell fate commitment	1.41×10^{-6}
Inr-only	single-multicellular organism process	2.02×10^{-6}
Inr-only	leg disc morphogenesis	2.56×10^{-6}
Inr-only	negative regulation of transcription, DNA-templated	2.71×10^{-6}
Inr-only	positive regulation of transcription, DNA-templated	3.82×10^{-6}
Inr-only	positive regulation of nitrogen compound metabolic process	3.82×10^{-6}
Inr-only	positive regulation of nucleobase-containing compound metabolic process	4.34×10^{-6}
Inr-only	negative regulation of RNA metabolic process	4.59×10^{-6}
Inr-only	segment specification	5.08×10^{-6}
Inr-only	positive regulation of gene expression	6.59×10^{-6}
Inr-only	positive regulation of biosynthetic process	6.99×10^{-6}
Inr-only	positive regulation of cellular biosynthetic process	6.99×10^{-6}
Inr-only	negative regulation of nucleobase-containing compound metabolic process	1.07×10^{-5}
Inr-only	positive regulation of RNA metabolic process	1.12×10^{-5}
Inr-only	negative regulation of nitrogen compound metabolic process	1.25×10^{-5}
Inr-only	positive regulation of macromolecule biosynthetic process	2.52×10^{-5}
Inr-only	imaginal disc-derived leg morphogenesis	2.75×10^{-5}

Promoter Type	**GO term**	**Holm-Bonferroni adjusted P-value**
Inr-only	negative regulation of gene expression	3.06×10^{-5}
Inr-only	nucleic acid metabolic process	3.15×10^{-5}
Inr-only	positive regulation of transcription from RNA polymerase II promoter	3.64×10^{-5}
Inr-only	head segmentation	3.86×10^{-5}
Inr-only	wing disc development	3.96×10^{-5}
Inr-only	negative regulation of macromolecule biosynthetic process	4.81×10^{-5}
Inr-only	negative regulation of cellular macromolecule biosynthetic process	4.81×10^{-5}
Inr-only	imaginal disc-derived wing morphogenesis	5.13×10^{-5}
Inr-only	imaginal disc-derived wing vein specification	5.60×10^{-5}
Inr-only	negative regulation of biosynthetic process	6.77×10^{-5}
Inr-only	negative regulation of cellular biosynthetic process	6.77×10^{-5}
Inr-only	wing disc morphogenesis	6.98×10^{-5}
Inr-only	head development	7.08×10^{-5}
Inr-only	locomotion	1.49×10^{-4}
Inr-only	negative regulation of macromolecule metabolic process	2.10×10^{-4}
Inr-only	negative regulation of cellular metabolic process	2.17×10^{-4}
Inr-only	embryo development	2.32×10^{-4}
Inr-only	positive regulation of biological process	2.57×10^{-4}
Inr-only	segmentation	2.86×10^{-4}
Inr-only	positive regulation of cellular process	3.79×10^{-4}
Inr-only	cellular macromolecule biosynthetic process	6.42×10^{-4}
Inr-only	multicellular organismal process	6.69×10^{-4}
Inr-only	macromolecule biosynthetic process	7.31×10^{-4}
Inr-only	negative regulation of metabolic process	7.73×10^{-4}
Inr-only	positive regulation of cellular metabolic process	7.84×10^{-4}
Inr-only	cell migration	8.35×10^{-4}
Inr-only	formation of anatomical boundary	1.13×10^{-3}
Inr-only	positive regulation of macromolecule metabolic process	1.14×10^{-3}
Inr-only	specification of segmental identity, head	1.16×10^{-3}

Promoter Type	**GO term**	**Holm-Bonferroni adjusted P-value**
Inr-only	positive regulation of metabolic process	1.24×10^{-3}
Inr-only	cell adhesion	1.73×10^{-3}
Inr-only	cell motility	1.80×10^{-3}
Inr-only	central nervous system development	1.98×10^{-3}
Inr-only	heart development	2.18×10^{-3}
Inr-only	cell-cell adhesion	2.42×10^{-3}
Inr-only	cardiovascular system development	2.95×10^{-3}
Inr-only	circulatory system development	2.95×10^{-3}
Inr-only	localization of cell	3.01×10^{-3}
Inr-only	cell development	3.81×10^{-3}
Inr-only	cellular nitrogen compound metabolic process	3.84×10^{-3}
Inr-only	nucleobase-containing compound metabolic process	3.98×10^{-3}
Inr-only	biological adhesion	4.19×10^{-3}
Inr-only	cellular component movement	4.20×10^{-3}
Inr-only	salivary gland boundary specification	4.41×10^{-3}
Inr-only	respiratory system development	4.68×10^{-3}
Inr-only	organic substance biosynthetic process	5.10×10^{-3}
Inr-only	epithelium development	5.59×10^{-3}
Inr-only	cellular aromatic compound metabolic process	5.85×10^{-3}
Inr-only	ectoderm development	5.89×10^{-3}
Inr-only	nervous system development	6.26×10^{-3}
Inr-only	negative regulation of cellular process	7.29×10^{-3}
Inr-only	heterocycle metabolic process	7.81×10^{-3}
Inr-only	organic cyclic compound metabolic process	7.84×10^{-3}
Inr-only	cellular biosynthetic process	8.63×10^{-3}
Inr-only	single-organism process	8.89×10^{-3}
Inr-only	open tracheal system development	1.10×10^{-2}
Inr-only	biosynthetic process	1.15×10^{-2}
Inr-only	muscle structure development	1.23×10^{-2}
Inr-only	organ formation	1.39×10^{-2}

Promoter Type	**GO term**	**Holm-Bonferroni adjusted P-value**
Inr-only	cellular developmental process	1.46×10^{-2}
Inr-only	tissue morphogenesis	1.51×10^{-2}
Inr-only	generation of neurons	1.67×10^{-2}
Inr-only	cell differentiation	1.94×10^{-2}
Inr-only	mesoderm development	1.98×10^{-2}
Inr-only	genital disc pattern formation	2.12×10^{-2}
Inr-only	genital disc anterior/posterior pattern formation	2.12×10^{-2}
Inr-only	negative regulation of biological process	2.38×10^{-2}
Inr-only	neuron differentiation	2.99×10^{-2}
Inr-only	posterior head segmentation	3.50×10^{-2}
Inr-only	anterior head segmentation	3.78×10^{-2}
Inr-only	anterior head development	3.78×10^{-2}
Inr-only	compound eye development	4.19×10^{-2}
Inr-only	anatomical structure formation involved in morphogenesis	4.59×10^{-2}
Inr-only	formation of organ boundary	4.65×10^{-2}
Inr-only	homophilic cell adhesion	4.69×10^{-2}
Inr-only	regulation of cell fate specification	4.69×10^{-2}
Inr-only	anterior/posterior pattern specification, imaginal disc	4.69×10^{-2}
Inr-only	trunk segmentation	4.69×10^{-2}
Inr/DEP	organ development	2.38×10^{-4}
Inr/DEP	system development	1.42×10^{-3}
Inr/DEP	cell adhesion	4.56×10^{-3}
Inr/DEP	biological adhesion	1.28×10^{-2}
Inr/DEP	generation of neurons	1.45×10^{-2}
Inr/DEP	response to chemical	2.82×10^{-2}
Inr/DEP	imaginal disc development	3.35×10^{-2}
Inr/DEP	synaptic target attraction	3.39×10^{-2}
Inr/DEP	heart development	4.57×10^{-2}
Inr/DEP	response to alcohol	4.57×10^{-2}

Promoter Type	**GO term**	**Holm-Bonferroni adjusted P-value**
TATA box/Inr	cuticle development	2.08×10^{-11}
TATA box/Inr	chitin-based cuticle development	1.10×10^{-10}
TATA box/Inr	body morphogenesis	5.39×10^{-8}
TATA box/Inr	cell fate specification	2.03×10^{-3}
TATA box/Inr	mesodermal cell fate commitment	3.93×10^{-2}
TATA box/Inr	mesodermal cell differentiation	3.93×10^{-2}
TATA box/Inr	mesoderm morphogenesis	4.45×10^{-2}
DRE	cellular metabolic process	1.35×10^{-11}
DRE	cellular macromolecule metabolic process	4.78×10^{-10}
DRE	cellular process	3.13×10^{-9}
DRE	cellular component organization or biogenesis	2.60×10^{-8}
DRE	primary metabolic process	1.53×10^{-7}
DRE	cellular component organization	3.99×10^{-7}
DRE	cellular nitrogen compound metabolic process	5.73×10^{-7}
DRE	metabolic process	9.96×10^{-7}
DRE	regulation of biological process	7.34×10^{-6}
DRE	regulation of cellular process	8.47×10^{-6}
DRE	biological regulation	8.89×10^{-6}
DRE	mitotic cell cycle	1.11×10^{-5}
DRE	organic substance metabolic process	1.13×10^{-5}
DRE	single-organism organelle organization	1.20×10^{-5}
DRE	cell cycle	2.16×10^{-5}
DRE	cellular localization	3.74×10^{-5}

Promoter Type	**GO term**	**Holm-Bonferroni adjusted P-value**
DRE	heterocycle metabolic process	4.18×10^{-5}
DRE	macromolecule metabolic process	5.26×10^{-5}
DRE	nitrogen compound metabolic process	6.34×10^{-5}
DRE	nucleic acid metabolic process	8.17×10^{-5}
DRE	nucleobase-containing compound metabolic process	8.46×10^{-5}
DRE	RNA processing	1.09×10^{-4}
DRE	cellular protein metabolic process	1.32×10^{-4}
DRE	RNA metabolic process	2.39×10^{-4}
DRE	cellular aromatic compound metabolic process	2.59×10^{-4}
DRE	organic cyclic compound metabolic process	2.66×10^{-4}
DRE	organelle organization	3.48×10^{-4}
DRE	mRNA metabolic process	3.92×10^{-4}
DRE	cellular protein modification process	4.33×10^{-4}
DRE	protein modification process	4.33×10^{-4}
DRE	establishment of localization in cell	5.86×10^{-4}
DRE	macromolecule modification	7.63×10^{-4}
DRE	neurogenesis	8.99×10^{-3}
DRE	cell differentiation	1.32×10^{-2}
DRE	cell cycle process	1.32×10^{-2}
DRE	regulation of metabolic process	2.55×10^{-2}
DRE	mRNA processing	2.89×10^{-2}
DRE	regulation of cellular metabolic process	3.26×10^{-2}
DRE	cellular developmental process	3.33×10^{-2}
DRE	protein modification by small protein conjugation or removal	4.88×10^{-2}
Motif1/6	organic cyclic compound metabolic process	7.16×10^{-5}
Motif1/6	cellular metabolic process	8.95×10^{-5}
Motif1/6	primary metabolic process	1.18×10^{-4}
Motif1/6	nucleobase-containing compound metabolic process	1.33×10^{-4}
Motif1/6	cellular aromatic compound metabolic process	1.36×10^{-4}
Motif1/6	organic substance metabolic process	2.01×10^{-4}

Promoter Type	**GO term**	**Holm-Bonferroni adjusted P-value**
Motif1/6	metabolic process	7.90×10^{-4}
Motif1/6	heterocycle metabolic process	8.04×10^{-4}
Motif1/6	cellular process	8.33×10^{-4}
Motif1/6	cellular nitrogen compound metabolic process	1.68×10^{-3}
Motif1/6	single-organism organelle organization	1.90×10^{-3}
Motif1/6	cellular macromolecule metabolic process	1.99×10^{-3}
Motif1/6	RNA processing	2.08×10^{-3}
Motif1/6	gene expression	2.30×10^{-3}
Motif1/6	RNA metabolic process	2.37×10^{-3}
Motif1/6	nervous system development	2.86×10^{-3}
Motif1/6	nucleic acid metabolic process	4.58×10^{-3}
Motif1/6	neurogenesis	1.10×10^{-2}
Motif1/6	mRNA metabolic process	2.01×10^{-2}
Motif1/6	nitrogen compound metabolic process	2.92×10^{-2}
Motif1/6	cellular component organization or biogenesis	4.47×10^{-2}

Table A.3: Table listing the Protein domain enrichment among the HCNE-proximal genes.

Promoter Type	**Protein domain**	**Holm-Bonferroni adjusted P-value**
Inr-only	Homeodomain-like	3.00×10^{-11}
Inr-only	Homeobox domain	8.61×10^{-10}
Inr-only	Homeobox, conserved site	3.58×10^{-8}
Inr-only	Zinc finger, C2H2	2.18×10^{-6}
Inr-only	Zinc finger C2H2-type/integrase DNA-binding domain	3.67×10^{-6}
Inr-only	Zinc finger, C2H2-like	1.46×10^{-5}
Inr-only	Immunoglobulin-like domain	3.98×10^{-3}
Inr-only	Immunoglobulin-like fold	1.24×10^{-2}
Inr/DPE	Immunoglobulin-like domain	1.93×10^{-11}

Promoter Type	**Protein domain**	**Holm-Bonferroni adjusted P-value**
Inr/DPE	Immunoglobulin subtype 2	1.44×10^{-10}
Inr/DPE	Immunoglobulin subtype	3.44×10^{-10}
Inr/DPE	Immunoglobulin-like fold	1.30×10^{-8}
Inr/DPE	CD80-like, immunoglobulin C2-set	3.39×10^{-6}
Inr/DPE	Immunoglobulin I-set	1.68×10^{-5}
Inr/DPE	Immunoglobulin V-set domain	1.02×10^{-3}
Inr/DPE	Leucine-rich repeat, typical subtype	2.50×10^{-3}
TATAbox/Inr	Domain of unknown function DUF243	9.24×10^{-7}
TATAbox/Inr	Protein of unknown function DUF1676	6.64×10^{-5}
TATAbox/Inr	Insect cuticle protein	3.38×10^{-4}
TATAbox/Inr	GYR motif	2.32×10^{-2}

Table A.4: **Table listing P-values associated with Figures 5.5A-B.** P-values obtained by comparing the stage (lower triangular) and tissue (upper triangular) specificity distributions between genes of different core promoter types.

Promoter Type	**Inr/DPE**	**Inr-only**	**TATAbox/Inr**	**Motif1/6**	**DRE**	**Unknown**
Inr/DPE	-	4.76×10^{-2}	8.89×10^{-14}	1.73×10^{-14}	2.97×10^{-18}	2.65×10^{-13}
Inr-only	2.37×10^{-1}	-	1.31×10^{-6}	5.33×10^{-16}	4.12×10^{-20}	9.97×10^{-6}
TATAbox/Inr	1.65×10^{-18}	9.58×10^{-12}	-	1.34E-36	7.73×10^{-48}	4.45×10^{-1}
Motif1/6	1.98×10^{-19}	1.47×10^{-17}	1.21×10^{-47}	-	4.05×10^{-1}	1.23×10^{-45}
DRE	1.01×10^{-30}	8.59×10^{-27}	5.06×10^{-69}	5.69×10^{-1}	-	2.00×10^{-62}
Unknown	1.00×10^{-13}	1.07×10^{-7}	4.56×10^{-1}	6.21×10^{-51}	$1.89 \times 10^{-8}0$	-

B Appendix of Dynamic eRNA

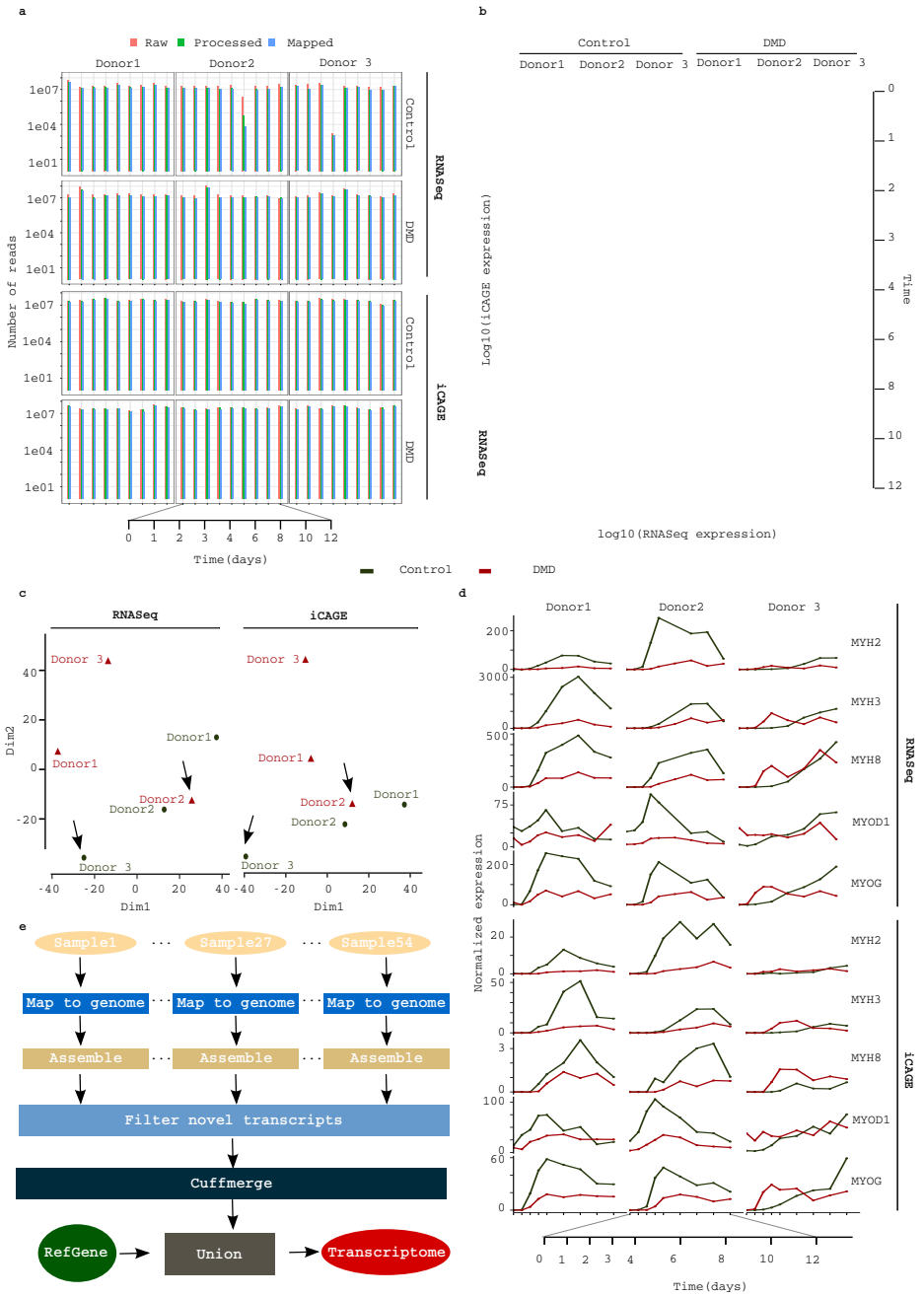

Figure B.1: **a)** b barplot shows the size of each library and number of mapped reads. **b)** scatterplot shows a high correlation between RNASeq and iCAGE in all libraries. **c)** projection of different donors based on gene expression reveal inconsistency of expression of donor 3 with donor 1 and 2 in healthy samples and donor 2 with donor 1 and donor 3 in DMD samples. **d)**expression profile of key myogenic genes confirms observation noted in (c). **d)**procedure taken to generate transcriptome profile of myogenesis.