Computational Genomics of Regulatory Elements and Regulatory Territories

Ge Tan

Supervisor: Prof. Boris Lenhard

Institute of Clinical Sciences
Imperial College London

This dissertation is submitted for the degree of
Doctor of Philosophy

November 2017

Declaration

I hereby confirm that I am the sole author of this PhD thesis here enclosed and that I have compiled it in my own words. Parts excepted are corrections of form and content by the supervisor. I have documented all methods, data and processes truthfully. The work presented here is my own, and any contribution from collaborators has been acknowledged and referenced.

Ge Tan
November 2017

Acknowledgements

To me, the past four years in London means much more than pursuing a PhD degree. During this period, I got challenged not just in science, but also in life. Towards the end of this journey, I would like to acknowledge the people who have accompanied, helped and supported me in the last four years.

Dear Boris, thank you for giving me the opportunity to be part of Lenhard group. Without you, I probably would never have a chance of doing PhD in UK. You hired me as a Scientific Programmer initially, and made this part-time PhD programme feasible. During the past four years, you have always been supportive and understanding. Thank you for your supervision and guidance, and so many scientific ideas. I will always be honoured as a member of Lenhard group.

I also would like to thank all the members of Lenhard group for all the help, discussion and pub nights. Piotr, thank you for being a such great friend, colleague and most importantly, a flatmate. I miss the days of us tweaking the servers, talking about science, watching movies together. Dimitris, you accompanied me taking tube home together so many times and having dinner together at Foundry. Thank you for your help on CNEr and GRB target gene prediction projects. Alex, thank you for all the inspiring discussion on CNEs and input on the rotifer CNE project. Your Illustrator skills really amazed me! Nathan, thank you for being a great friend and all the help discussion on GRB. Malcolm, you are a talented boy and thank you for sharing knowledge and discussion on R. Liz, I still remember you printed out the map and pointed me the fastest way to Heathrow airport on my interview day. Thank you for always being so kind. Amit and Anja, thank you for the encouragement and tips of taking care of a baby. Dunja has always been a great neighbour and Nevena was always joking me as her boss. Thank you for all the joys.

I feel grateful to the three major airlines doing business between London and Amsterdam: KLM, BA and Easyjet. I am sure you earned quite some from me. But without you, I can never survive during the past four-year travel.

Finally, I'd like to thank my family, Yang and Suri. Dear Yang, thank you for taking care of everything at home in Utrecht when I was not there and feeding me so fat. You are always my best friend, playmate and life companion. And, apart from your creativities on cooking, you are a truly great scientist with impressive scientific achievements. Life will never be boring with you. Suri, my dear daughter. Now being with you everyday is the most rewarding thing for me. Thank you for bringing me the greatest happiness in the life.

Abstract

Whole genome comparison of metazoan genomes reveals extremely high level of noncoding conservation over tens to hundreds of base pairs across distant species. These sequences are termed as conserved noncoding elements (CNEs). Arrays of conserved noncoding elements that span the loci of developmental regulatory genes and their span defines regulatory genomic blocks (GRBs). CNEs are currently known to be involved in transcriptional regulation and development as long-range enhancers. However, no molecular mechanism can yet explain their exceptional degree of conservation.

As a first step towards the genome-wide study of these elements, I developed two R/Bioconductor packages *CNEr* and *TFBSTools*, to detect and analyse regulatory elements. Next, I designed a novel CNE detection pipeline for duplicated regions in the ameiotic *Adineta vaga* genome. Identification of CNEs in this genome suggests that the principal function of CNEs is regulation of developmental gene expression rather than copy number sensing. In addition, I performed a *de novo* genome annotation of European common carp *Cyprinus carpio*. This genome stands as an ideal candidate for comparative study of zebrafish genome. Its analysis revealed a wealth of previously undetected fish regulatory elements and their unexpectedly high level of conservation between the two genomes. Finally, I presented a computational method for the identification of GRB boundaries and prediction of the corresponding target genes under long-range regulation. The predicted target genes are implicated in developmental, transcriptional regulation and axon guidance. The disruption of regulation of these target genes is likely to cause complex diseases, including cancer. The GRB boundaries and predicted target genes are valuable resource for investigating developmental regulation and interpreting genome-wide association studies.

Table of contents

List of abbreviations

DNA ..... Deoxyribonucleic acid
bp ..... Base pair
chr ..... Chromosome
CNEs ..... Conserved Noncoding Elements
GRBs ..... Genomic Regulatory Blocks
WGD ..... Whole Genome Duplication
Mya ..... Million years ago
TADs ..... Topologically Associating Domains
MSA ..... Multiple Sequence Alignment
GO ..... Gene Ontology
KEGG ..... Kyoto Encyclopedia of Genes and Genomes
DO ..... Disease Ontology
TF ..... Transcription Factor
TFBS ..... Transcription Factor Binding Site
PFM ..... Position Frequency Matrix
PWM ..... Position Weight Matrix
ICM ..... Information Content Matrix
TFFM ..... Transcription Factor Flexible Model
CNV ..... Copy Number Variants
SD ..... Segmental Duplications
CRM ..... *cis*-Regulatory Modules

FC . . . . . Fold Change
RF . . . . . Random Forest
oob . . . . . out-of-bag
TPR . . . . . True Positive Rate
TNR . . . . . True Negative Rate
ACC . . . . . Accuracy
MDS . . . . . Multi Dimensional Scaling
GWAS . . . . . Genome-wide Association Study
SNPs . . . . . Single Nucleotide Polymorphisms

List of figures

List of tables

Chapter 1

Introduction

Deoxyribonucleic acid (DNA) molecule, present in all living organism, encodes the genetic information for cell function and development. The entire set of DNA within an organism forms the genome. To produce an assembled reference sequence of the human genome, Human Genome Project was initiated in 1990 and declared finished in 2003 (International Human Genome Sequencing Consortium, 2004). In parallel, genomes of additional model organisms became available. To identify all the functional elements in human genome, the ENCODE (Encyclopedia of DNA Elements) was launched (pilot phase in 2003 (ENCODE Project Consortium, 2004; ENCODE Project Consortium et al., 2007), full genome first phase in 2007 (ENCODE Project Consortium, 2012) and is still ongoing. ENCODE has produced an unprecedented amount of information from genome-wide, next-generation sequencing based assay, for a panel of selected human cell lines (Sloan et al., 2016). Other consortium-based efforts of similar nature have produced data for actual human tissues (Roadmap Epigenomics Consortium et al., 2015), specific processes such as hematopoiesis (Adams et al., 2012) or for model organisms (Celniker et al., 2009; Yue et al., 2014).

However, the question of what fraction of genome is subject to function still puzzles us. Genes, coding units for proteins, were once believed to make up the most of functional DNA, while the rest of DNA is not functional. In addition to these efforts to functionally characterise genomic elements, newest high throughput sequencing technologies, especially single molecule long read sequencing (PacBio and Nanopore), have made *de novo* genome assembly becomes more and more economically affordable. These developments have made it easier to sequence and assemble genomes of new species, increasing the power of computational genomics approaches. Comparison of genomes apparently shows that organismal complexity cannot be explained by the number of protein coding genes and

genome size. It is even more challenging to understand how all the cells with same genome differentiate into different tissues and functional diversity.

Comparisons of metazoan genome sequences have revealed abundant genomic elements that are extremely well conserved across large evolutionary distances, although they do not encode proteins. These segments have been termed conserved noncoding elements (CNEs); several other names are used in the literature (Kikuta et al., 2007a). Using a very strict definition, 481 CNEs were found as non-protein-coding sequences longer than 200bp with perfect identity between human, mouse and rat genomes; 97% of them were conserved in chicken and 67.3% in fish (Bejerano et al., 2004). Using less stringent criteria of sequence length and sequence similarity, many more CNEs can be found in human genome that are conserved in organisms as distant as teleost fish (Engström et al., 2008; Sandelin et al., 2004; Woolfe et al., 2005). The strong conservation of these sequences implies an important conserved biological function as a source of purifying selection. Experimental studies in transgenic animals suggest that CNEs have the ability to drive transcription of a reporter gene and act as tissue-specific enhancers during development (de la Calle-Mustienes et al., 2005; Nobrega et al., 2003; Pennacchio et al., 2006; Woolfe et al., 2005). Despite this, the molecular mechanism that requires this high degree of conservation remains unexplained: no known source of purifying selection can explain either its pattern or its extent (Harmston et al., 2013). In the experiments, CNEs and developmental genes are separated by distance up to more than a megabase, suggesting CNEs can act as enhancers from very long distances (Kikuta et al., 2007b; Pennacchio et al., 2006).

A striking property of CNEs is that they cluster around many key developmental regulatory genes. This pattern has been observed in vertebrates (Sandelin et al., 2004; Woolfe et al., 2005) and in insects (Engström et al., 2007). The regions spanned by CNE clusters have been named *Genomic Regulatory Blocks* (GRBs) (Kikuta et al., 2007b). The explanation of these observations is that an array of CNEs defines the region of regulatory inputs for the gene they regulate ("target gene"). The arrays of CNEs are kept in *cis* and in synteny to their target genes as a functional regulatory unit. Although CNEs may be hundreds of kilobases away from their target genes, and often closer to other genes ("bystander genes") or residing in introns of the latter, CNEs can regulate their target genes without affecting the bystander genes. It has also been shown that there is dense interaction between CNEs, which indicates that work in a cooperative manner to regulate the target gene (Dimitrieva and Bucher, 2012). If all these conjectures are correct, chromosomal rearrangements within

GRB should be selected against during evolution. As a result, GRBs should be detectable between two distant organisms by their synteny conservation. The GRB target genes are a set of genes with central importance in the multicellular developmental process regulation (Engström et al., 2007; Kikuta et al., 2007b; Navratilova and Becker, 2009). Despite the importance of GRB target genes, their identification still relies on a manual, semi-intuitive and non-standardised process that includes the inspection of genes in the region and checking their functional annotation and sequence properties (e.g. CpG islands). The list of target genes obtained this way has been produced for mammalian genomes only and is far from complete. For *Drosophila*, there is no list of target genes of this kind. This prevents a comprehensive genome-wide analysis of GRB structure and function. To tackle this problem, there should be a robust approach for the automated determination of span of GRBs on chromosomes and the identification of the most likely target genes. Only with these methods in place we will be able to perform efficient genome-wide exploration of GRBs, CNEs and their regulatory content. Akalin et al (Akalin et al., 2009) proposed several transcriptional and epigenetic features that may distinguish target genes from bystander genes, especially long and/or multiple CpG islands overlapping genes, larger transcription initiation regions, and specific combinations of histone modifications. The growing and increasingly informative list of features likely to be associated with GRB target genes enables us to devise computational approaches such as machine learning to predict the target genes and the regulatory domains around them genome-wide.

In this thesis, I have focused on developing new approaches for better understanding the roles of CNEs, GRBs in gene regulation. I describe the development of the methods and software tools and employ each to addressing specific biological questions.

1.1 Outline of the thesis

In the following chapters, I first present two R/Bioconductor packages developed for studying regulatory elements. Then, with these packages, I investigate noncoding conservation in several species to better understand the evolution, genomic organisation of regulatory elements and regulatory territories. Finally, I present conclusions based on the findings and discuss future directions.

Chapter 2 describes the rationale, implementation and use cases of the R/Bioconductor package *CNEr* for CNE detection and visualisation. This package fills the gap of missing

publicly available tools for large-scale detection of noncoding conservation. It also provides necessary data structures for comparative genomics studies of regions characterised by high CNE densities, which include the loci of most of the key developmental regulatory genes. Results of noncoding conservation between *Drosophila* and *Glossina* (testse fly), between two species of sea urchin, and general features of CNEs are also presented in this chapter.

Chapter 3 demonstrates the design and functionality of R/Bioconductor package *TFBSTools*. The package contains the first comprehensive R/Bioconductor toolbox for the analysis of transcription factor binding sites (TFBS) in genomic sequences and alignments, including scanning of DNA sequences and alignments with consensus motifs and matrix profiles, wrappers for *de novo* motif discovery tools, and for the visualisation of sequence logos. The package is equipped with easy access to the JASPAR database of transcription factor binding site matrix profiles, and is complemented by R/Bioconductor data packages for its each major release.

Chapter 4 investigates one possible source of extreme noncoding conservation by studying the intragenomic noncoding conservation between paralogous regions of the ameiotic, recently tetraploidised genome of the rotifer *Adineta vaga* genome. The main purpose of this analysis is to answer the question if germline and meiosis are required for the presence of selective pressure that gives rise to CNEs. A novel CNE detection pipeline is proposed to detect intragenomic noncoding conservation resulting from whole genome duplications.

Chapter 5 describes the *de novo* assembly of European common carp *Cyprinus carpio* and genome annotation. This was a collaborative project in which we played a major role in the assembly, annotation and analysis of the new genome. Due to large evolutionary distance between zebrafish and other teleost fish whose genomic sequences were available at that point, common carp as a member of the same family as zebrafish stands as a perfect candidate for comparative genomics studies of zebrafish, expanding the repertoire of regulatory elements that can be studied that way. In addition, the carp genome underwent the most recent known additional whole-genome duplication about 9 (Million years ago) Mya, giving us an unprecedented glimpse into early rediploidisation patterns, especially informative in GRBs.

Chapter 6 solves the problem of detecting the regulatory boundaries of CNEs and the corresponding target genes under transcriptional regulation. The predicted spans of regulatory territories and target genes are beneficial for the community studying developmental regulation and disease-associated genomic variation.

Chapter 7 concludes the thesis and discusses the directions of further study.

Chapter 2

CNEr: a toolkit for exploring extreme noncoding conservation

Comparative genomics has revealed noncoding regions with extremely high conservation across large evolutionary distances, termed conserved noncoding elements (CNEs). Our research group has more than a decade of experience characterising these elements and their genome distribution in different metazoan clades (Engström et al., 2007; Kikuta et al., 2007b; Sandelin et al., 2004). Most recently, we have shown that the clusters of CNEs, genomic regulatory blocks (GRBs) coincide with topologically associating domains (TADs), insulated chromosomal regions with increased internal 3D contacts, around developmentally regulated genes (Harmston et al., 2017). This has provided further hypotheses about the still elusive origin of CNEs, and has provided a comparative genomics-based method of estimating the position of TADs in genomes where chromatin conformation capture data is missing. Systematic examination of extreme noncoding conservation across different genomes at varying evolutionary distances necessitates the generation of CNE datasets under different conservation criteria and their efficient manipulation and visualisation. Despite various resources providing variously defined sets of CNEs, there is no publicly available tool to identify these elements from scratch. For that purpose I developed *CNEr*, a toolkit for large-scale identification and advanced visualisation of CNEs. Given whole genome pairwise alignments as input, our pipeline enables manipulation and screening of whole-genome alignments, storage and querying of CNEs, calculation of CNE density, as well as visualisation in horizon plots (Few, 2008), which eliminates the need for multiple CNE density tracks at different thresholds for the same pairwise genome comparison. Furthermore,

it provides efficient scalable data structures for representing paired genomic ranges, providing essential functionality for phylogenetic footprinting and related analyses, which are currently not supported by R/Bioconductor infrastructure. The *CNEr* package is freely available as part of Bioconductor (Gentleman et al., 2004).

2.1 Introduction

CNEs are a pervasive class of elements that cluster around genes with roles in development and differentiation in Metazoa (Woolfe et al., 2005). While many have been shown to act as long-range developmental enhancers (de la Calle-Mustienes et al., 2005; Pennacchio et al., 2006), the source of their extreme conservation remains unexplained. The need to maintain arrays of CNEs in *cis* to the genes they regulate has led to their spatial arrangement into clusters termed Genomic Regulatory Blocks (GRBs) (Engström et al., 2007; Kikuta et al., 2007b). The emerging role of those clusters in genome organisation is further supported by recent findings demonstrating that ancient metazoan clusters of extreme noncoding conservation coincide with TADs (Harmston et al., 2017).

It has been shown that CNEs are selectively constrained and not mutational cold spots (Drake et al., 2006), leaving open the question of the source of such high conservation (Harmston et al., 2013). Furthermore, diverse regulatory motifs have been found in subsets of those elements (Xie et al., 2007), and alternative hypotheses have been proposed which suggest that CNEs are not under positive selection as continuous stretches of DNA but might be under a different kind of evolutionary pressure that acts on the single nucleotide level (Silla et al., 2014). A recent study also provided evidence that deletion or duplication of a CNE could be deleterious to the mammalian cell, and as a consequence proposed a homology-based mechanism evaluating genome integrity as an underlying role of CNEs, in addition to their roles in gene regulation (McCole et al., 2014). All of these explanations are hypotheses that have not been tested, and which are at this stage either impossible or difficult to test.

Numerous recent studies highlight and seek to elucidate the importance of functional non-coding regions in our genomes, most recently by mainly employing the CRISPR-Cas9 based techniques to locate and dissect elements that affect gene expression and phenotype/disease - associated processes (Montalbano et al., 2017; Mumbach et al., 2017; Wright and Sanjana, 2016). Hence, it is anticipated that prioritizing target loci of interest for interrogating the

function of their regulatory context will be one of the major focuses of functional genomic studies, as has been shown in the case of the *POU5F1* locus (Diao et al., 2016) and *NF1*, *NF2* and *CUL3* genes (Sanjana et al., 2016). CNEs serve as excellent candidates for such studies. To our knowledge, there is no available software that allows for efficient identification of CNEs using user-specified thresholds and across vertebrates, invertebrates and plants. To study the evolutionary dynamics of these elements and their relationship to the genes around which they cluster, it is essential to be able to produce genome-wide sets of CNEs for a large number of species comparisons, each with multiple length and conservation thresholds.

The *CNEr* package aims to detect CNEs and visualise them along the genome under a unified framework. For performance reasons, the implementation of CNEs detection and corresponding I/O functions are primarily written as C extensions to R. We have used *CNEr* to produce sets of CNEs by scanning pairwise whole-genome net alignments with multiple reference species, each with two different window sizes and a range of minimum identity thresholds, most of which are available at http://ancora.genereg.net/downloads. In this work, I demonstrate the application of *CNEr* to the investigation of noncoding conservation between *Drosophila* and *Glossina* - the two species at the evolutionary separation not previously investigated in insects, and between two species of sea urchins which enable us to observe some properties of GRB target genes shared across Metazoa. In a previous study, we showed that more distant comparisons in Diptera failed to identify CNEs (e.g. between *Drosophila* and mosquitoes) (Engström et al., 2007). On the other hand, the conservation level across different species of the *Drosophila* genus is comparable to that across placental mammals. With *Drosophila* and *Glossina*, we wanted to explore the evolutionary distance comparable to human vs. fish in another lineage and establish whether it is the same functional class of genes that is accompanied by CNEs featuring such a deep level of conservation. In the case of sea urchin, we wanted to investigate a lineage at an intermediate distance to vertebrates - closer than insects, more distant than early branching chordates - to establish the continuum of GRBs across Metazoa. I present a series of downstream analysis of the newly identified CNEs, identifying their characteristic sequence features in invertebrates and functional classes of genes whose loci they span.

2.2 Results

2.2.1 Overview of *CNEr* workflow

CNEr provides the functionality of large-scale identification and advanced visualisation of CNEs based on our previous strategies of detecting CNEs (Engström et al., 2007; Kikuta et al., 2007b; Sandelin et al., 2004) as shown in Figure 2.1. *CNEr* scans the whole genome pairwise net alignment, which can be downloaded from UCSC or generated by *CNEr* pipeline, for conserved elements. Various quality controls of the alignments are provided. The composition of aligned bases in the alignment can be used for tuning parameters during pairwise alignment (Figure A.1). The closer species are expected to give higher rates of matched bases. The syntenic dotplot of the alignments (Figure A.2) quickly shows the syntenic region between two assemblies.

Considering the different extents of evolutionary divergence and sequence similarity between the assemblies, we typically use the identity thresholds of 70% to 100% identity over a scanning window of 30bp or 50bp. Known annotations of exons and repeats are usually compiled from various sources, such as UCSC (Kent et al., 2002) and Ensembl (Yates et al., 2016) for common genomes, and elements overlapping with these regions are discarded.

The net alignments only keep the best match for each region in the reference genome. This is not acceptable when one of the aligned genomes underwent one or more whole genome duplications, leading to legitimate deviations from 1:2 orthology for many CNEs. To eliminate the bias of the choice of reference genome in the alignment and to capture duplicated CNEs during whole genome duplication (WGD), we scan two sets of net alignments by using each of the two compared genomes as reference in turn. This strategy performs well when comparing species with different numbers of WGD, such as tetrapod vertebrates and teleost fish (Jaillon et al., 2004), or common carp (Kolder et al., 2016) and other teleost fish. In such cases, some of the identified CNEs pairs from two rounds of screening overlap on both assemblies, and hence merged into one CNE pair. As the last step, we align the CNEs back to the two respective genomes using BLAT and discard the ones with high number of hits. The remaining elements are considered to be a clean set of CNEs.

CNEr provides a quick overview of the genomic distribution of CNEs along the chromosomes. In Figure A.3, each CNE between human and mouse is plotted relative to each human chromosome (x-axis). A CNE cluster is represented as a big increase of height in y-axis with

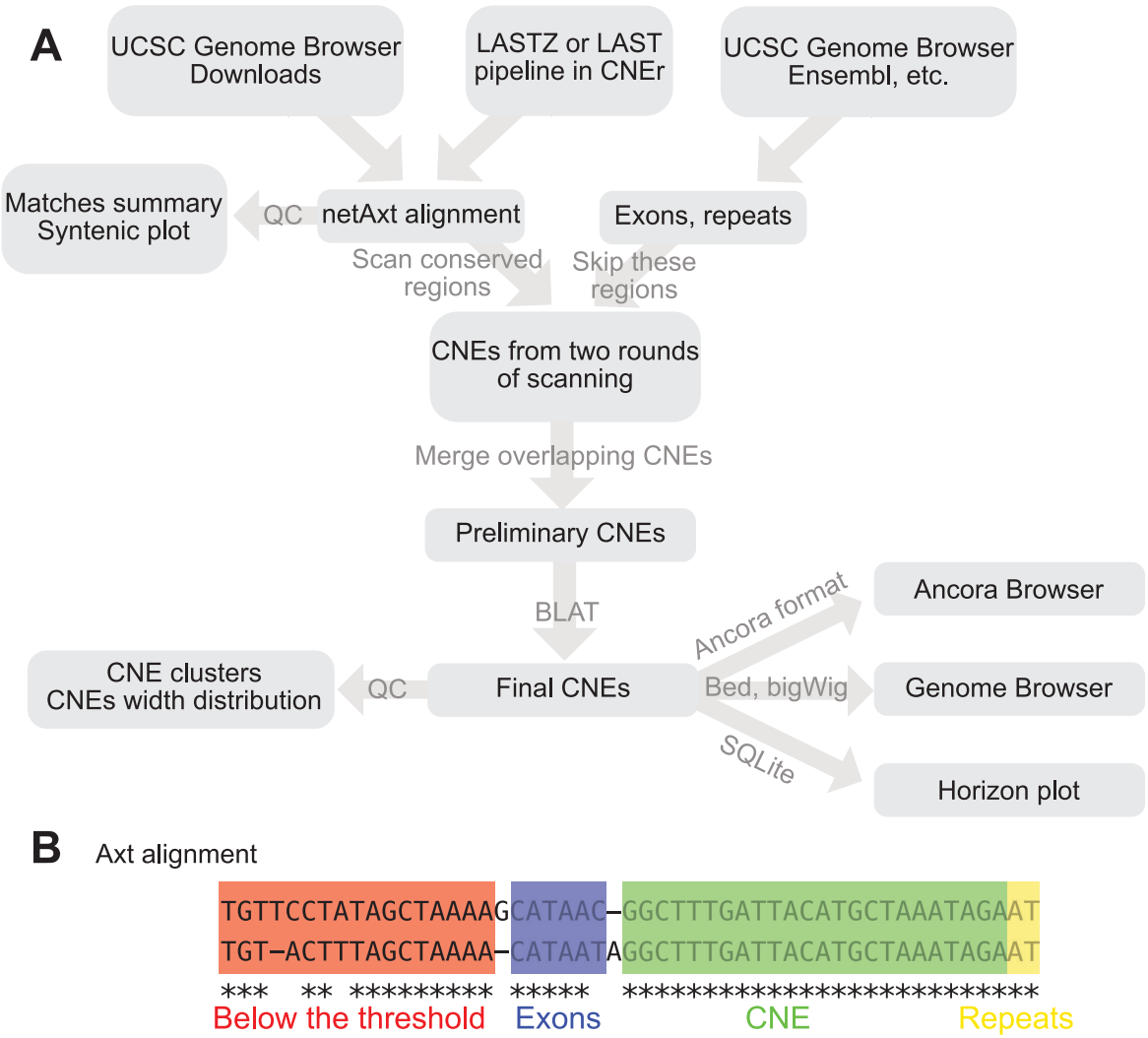

Figure 2.1 *CNEr* workflow. (A) A typical pipeline for the identification and visualisation of CNEs. (B) Illustration of scanning an alignment for CNEs. The scanning window moves along the alignment for conserved regions. The exons and repeat regions are skipped during the scanning.

small change in x-axis. For visualisation of CNEs in any genome browser, *CNEr* can export the CNE coordinates in BED file format and CNE density (measured by the percentage of area covered by CNEs within a smoothing window) in *bedGraph* and *bigWig* formats. Since running the whole pipeline of CNE detection can be time-consuming, we also implement a set of storage and query system with SQLite as backend. Based on the visualisation capability of *Gviz* package (Hahne and Ivanek, 2016), *CNEr* can produce publication-quality *horizon plots* of CNE density along with other genomic annotations (see Methods and data). Examples of the horizon plots are given in sections 2.2.3 and 2.2.4.

2.2.2 Comparison with related methods and existing CNE resources

CNEr identifies CNEs with user-defined criteria of sequence identity and minimum length of conserved sequence across organisms of interest. Our pipeline is not restricted to the identification of vertebrate conserved noncoding elements and might be utilised for retrieving invertebrate or plant CNEs. A handful of resources exist, mainly databases, which contain already pre-computed clusters of CNEs. These databases are static and mostly not updated. To mention a few, CEGA (Conserved Elements from Genomic Alignments - http://cega.ezlab.org/) provides a set of CNEs identified mainly within the vertebrate lineage (Dousse et al., 2016), while UCNEbase is a database covering 4,351 CNEs identified with stringent thresholds of sequence identity (95%-100% between human and chicken over 200 nucleotides) (Dimitrieva and Bucher, 2013). VISTA Enhancer Browser initially provided a dataset of evolutionary conserved noncoding human sequences comprised of around 170,000 noncoding sequences that are highly conserved between human and rodents (Visel et al., 2007). The most up-to-date version (9/8/2016) contains information on 2,388 *in vivo* tested elements of which 1,264 bear enhancer activity. Another resource, cneViewer (http://bioinformatics.bc.edu/chuanglab/cneViewer/), focuses again on vertebrate genomes and contains non-coding DNA elements in zebrafish that are conserved strongly with human (Persampieri et al., 2008). It mainly facilitates prioritising CNEs for experimental design and analysis. Ancora (http://ancora.genereg.net/) is a web resource containing non-exonic regions of high similarity between genomic sequences from distantly related species which also provides tools for studying the distribution of CNEs across chromosomes (Engström et al., 2008). Its main focus lies on developmental regulatory genes, their regulatory territories and their associated regulatory sequences. Available resources are summarised in Table 2.1. The ability to generate such elements in other species beyond vertebrates that is facilitated by our pipeline is very important. Towards this direction, we present the case of CNE identification between *Drosophila* and *Glossina*.

2.2.3 *CNEr* use case I: *Drosophila*:*Glossina* CNEs

In this section, we demonstrate the application of *CNEr* to the analysis of Tsetse Fly (*Glossina morsitans*) CNEs and their putative target genes. *Glossina* is the sole vector of African trypanosomiasis ("sleeping sickness"). *Glossina* has been studied due to its ability to mediate transmission of this disease during feeding on blood. It has been shown previously (Engström et al., 2007) that, while there are tens of thousands of CNEs detected across

Table 2.1 Summary of various resources of CNEs

Name	CNE definition	species	source
ANCORA	70 - 100% seq. id. over 30 or 50 bp window	Metazoa	http://ancora.genereg.net/
CEGA	Threshold-free phylogenetic modeling	vertebrates	http://cega.ezlab.org/
cneViewer	user-specified	human-zebrafish	http://bioinformatics.bc.edu/chuanglab/cneViewer/
CONDOR	65% seq. id. over 40 bp	mammalian-fugu	http://condor.nimr.mrc.ac.uk/
TFCONES	70% seq. id. over 100 bp	human-mouse	http://tfcones.fugu-sg.org
UCbase 2.0	100% seq. id. over 200bp	human-mouse-rat	http://ucbase.unimore.it
UCNEbase	>95% seq. id. over 200 bp (human - chicken)	18 vertebrate species	http://ccg.vital-it.ch/UCNEbase/
VISTA	100% seq. id. over 200 bp	human-mouse	http://enhancer.lbl.gov/
CNEr	user-specified	vertebrates, invertebrates, plants	http://bioconductor.org/packages/CNEr/

different *Drosophila* species, there are almost no highly conserved elements found between *Drosophila* and malaria mosquito *Anopheles gambiae* or other mosquitos. *Glossina* and *Drosophila* are much closer to each other than either of them is to mosquitos, having a common ancestor that has diverged around 60.3 Mya (Figure 2.2). With the newly available assembly and gene annotation of *Glossina* (International Glossina Genome Initiative, 2014), we were able to identify clusters of CNEs between these two species. The clusters correspond to a subset of clusters defined by the CNEs derived from comparisons of different *Drosophila* species. A further investigation of gene functions, which are retained or missing in *Glossina*, was carried by comparison with the *Drosophila* clusters.

A summary of CNEs detected between *Glossina* and *Drosophila* is given in Table 2.2. Unsurprisingly, many fewer CNEs are detected from the comparison of *Glossina* and *Drosophila* than between any two *Drosophila* species since *Glossina* is an outgroup to the *Drosophila*/*Sophophora* family. A closer examination of the CNE density plot in Ancora browser (Engström et al., 2008) revealed many missing clusters of CNEs relative to CNE

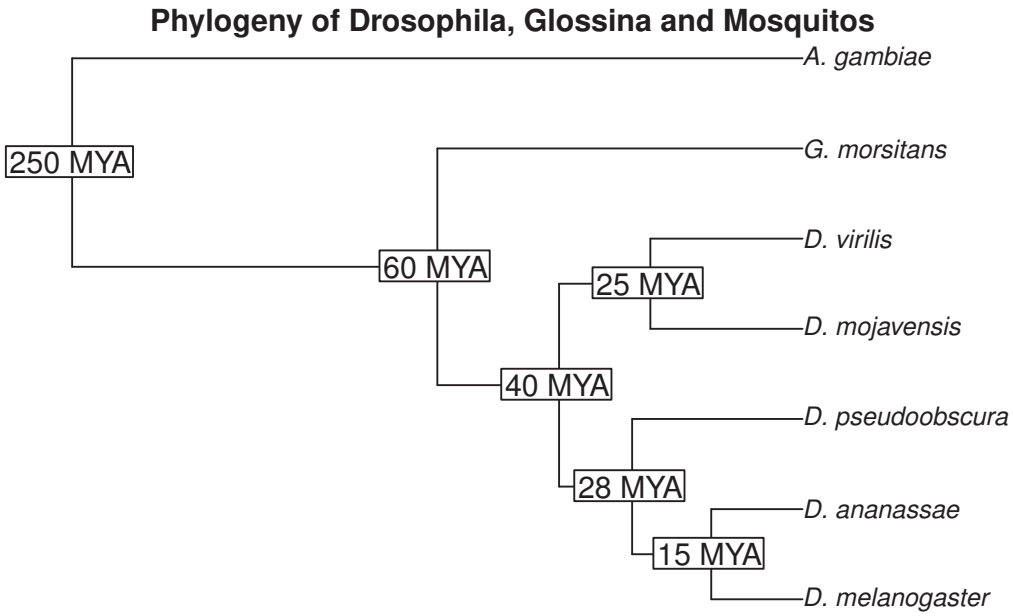

Figure 2.2 The species tree of *Drosophila*, *Glossina* and mosquitos. The phylogenetic tree is constructed based on the data on last common ancestors from TimeTree (Hedges et al., 2006). The genome of the malaria mosquito *A. gambiae* is highly divergent from *Drosophila* family and unsuitable for comparative genomics study, while *G. morsitans* is much closer.

density across *Drosophila* species, especially at a more stringent threshold. We wanted to find out if the missing and retained CNE clusters differ with respect to the functional categories of the genes they span. In the following analysis, the CNEs that are conserved for more than 70% over 30bp are considered.

The most deeply conserved vertebrate CNEs are usually associated with genes involved in transcriptional regulation or development (trans-dev) functions (Sandelin et al., 2004). Due to high divergence between *Drosophila* and *Glossina*, the regions with detectable CNE arrays tend to be of low CNE turnover, i.e. the process of sequence divergence and loss of ancestral CNEs is slow. If the same functional subset of genes is surrounded by low-turnover CNE clusters as in vertebrates, the encompassed genes will more likely be essential key developmental genes (Harmston et al., 2013). Indeed, *Drosophila* genes associated with (i.e. nearest to) *Glossina* vs. *Drosophila* CNEs are also associated with trans-dev terms (Figure 2.3A). Development, including organ development, system development, and tissue development, appears at the majority of the top Gene Ontology (GO) terms. The other highly

Table 2.2 The number of CNEs between *Glossina* and *Drosophila*, *Drosophila* family. NC, not counted due to too low threshold for close species.

Minimum identity	vs. Morsitans	vs. Ananassae	vs. Pseudoobscura	vs. Mojavensis	vs. Virilis
70% over 30 bp	9691	NC	NC	176366	204970
80% over 30 bp	3924	NC	313570	127293	146793
90% over 30 bp	1922	NC	212951	81436	92288
96% over 30 bp	813	177759	128843	47408	52134
100% over 30 bp	414	112073	76715	26972	29445
70% over 50 bp	3185	266385	248357	104476	120628
80% over 50 bp	1796	223975	177266	66063	75204
90% over 50 bp	732	142899	96994	33455	37098
96% over 50 bp	244	79631	49380	16387	17831
98% over 50 bp	150	55460	33463	10741	11548
100% over 50 bp	66	29218	17201	5250	5585

significant GO terms include biological regulation, regulation of cellular process and cell differentiation. CNE clusters can span regions of tens or hundreds of kilobases around the actual target gene, which is shorter than the equivalent spans in vertebrate genomes. This is in agreement with our observation that CNE clusters and the GRBs they define (and, by extension, the underlying TADs) expand and shrink roughly in proportion to genome size (Harmston et al., 2017). The *H15* and *mid* locus (Figure 2.4A) is one of the biggest CNE clusters retained between *Glossina* and *Drosophila*. The *H15* and *mid* genes encode the T-box family proteins involved in heart development (Reim et al., 2005). Although the CNE density between *Drosophila* and *Glossina* is much lower than that across *Drosophila* genus, it clearly marks the CNE cluster boundaries of this locus, containing 67 CNEs at the 70% identity over 30bp threshold. For the 40 largest retained CNE clusters, we provide a comprehensive list of CNE cluster coordinates, the target genes, the protein domains and the number of associated CNEs (Table A.1). As we can see, the majority of the target genes have *Homeobox*, *Forkhead* or *C2H2* Zn finger domains.

Some other regions have strong clusters of CNEs between *Drosophila* species, but the CNE cluster between *Drosophila* and *Glossina* is absent. The *ct* locus (Figure 2.4B), encoding the cut transcription factor, is one of the more extreme examples. Ct plays roles in the later stages of development, controlling axon guidance and branching in the development of nervous system, as well as in the specification of several organ structures such as Malpighian

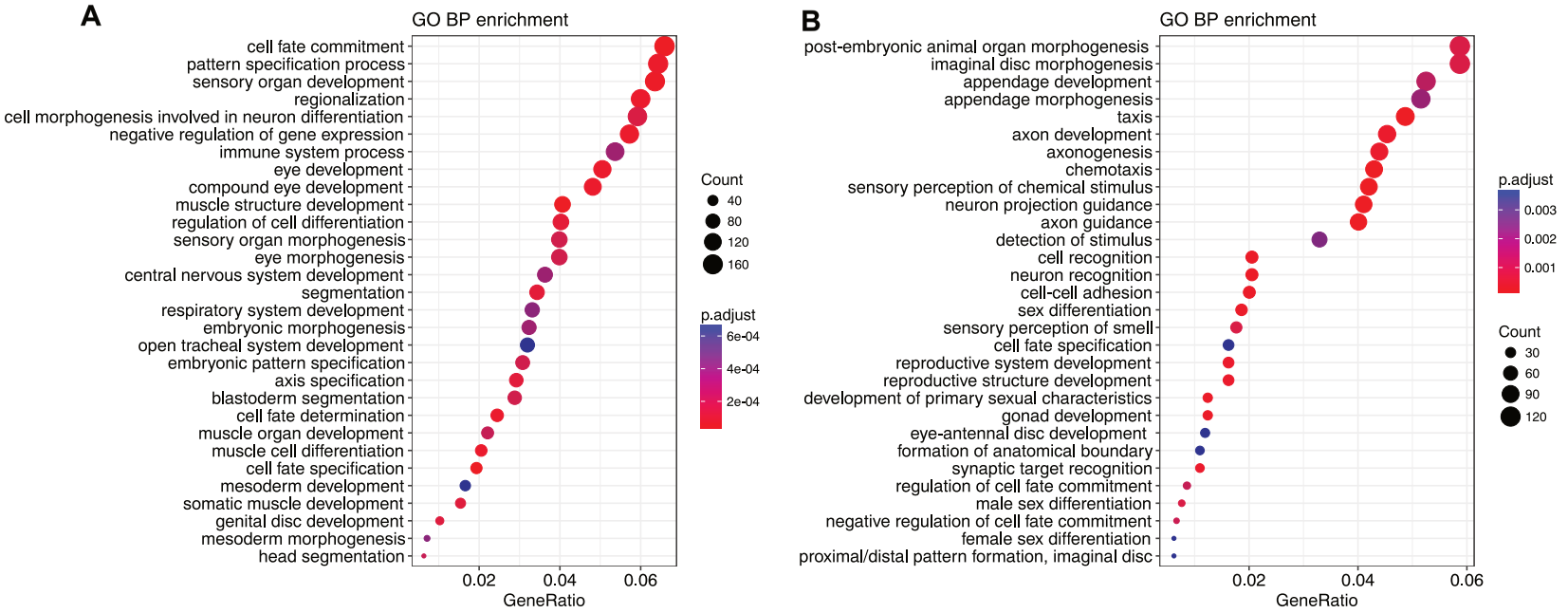

Figure 2.3 Over-represented GO Biological Process terms ranked by GeneRatio (the number of genes associated with the term in our selected genes divided by the number number of selected genes.) The p-values are adjusted by "BH" (Benjamini-Hochberg) approach (false discovery rate). The visualisation is done by *clusterProfiler* (Yu et al., 2012). (a) GO enrichment for genes nearest to *Drosophila* and *Glossina* CNEs. (b) GO enrichment for genes in the missing CNEs clusters compared between *Drosophila* and *Glossina*.

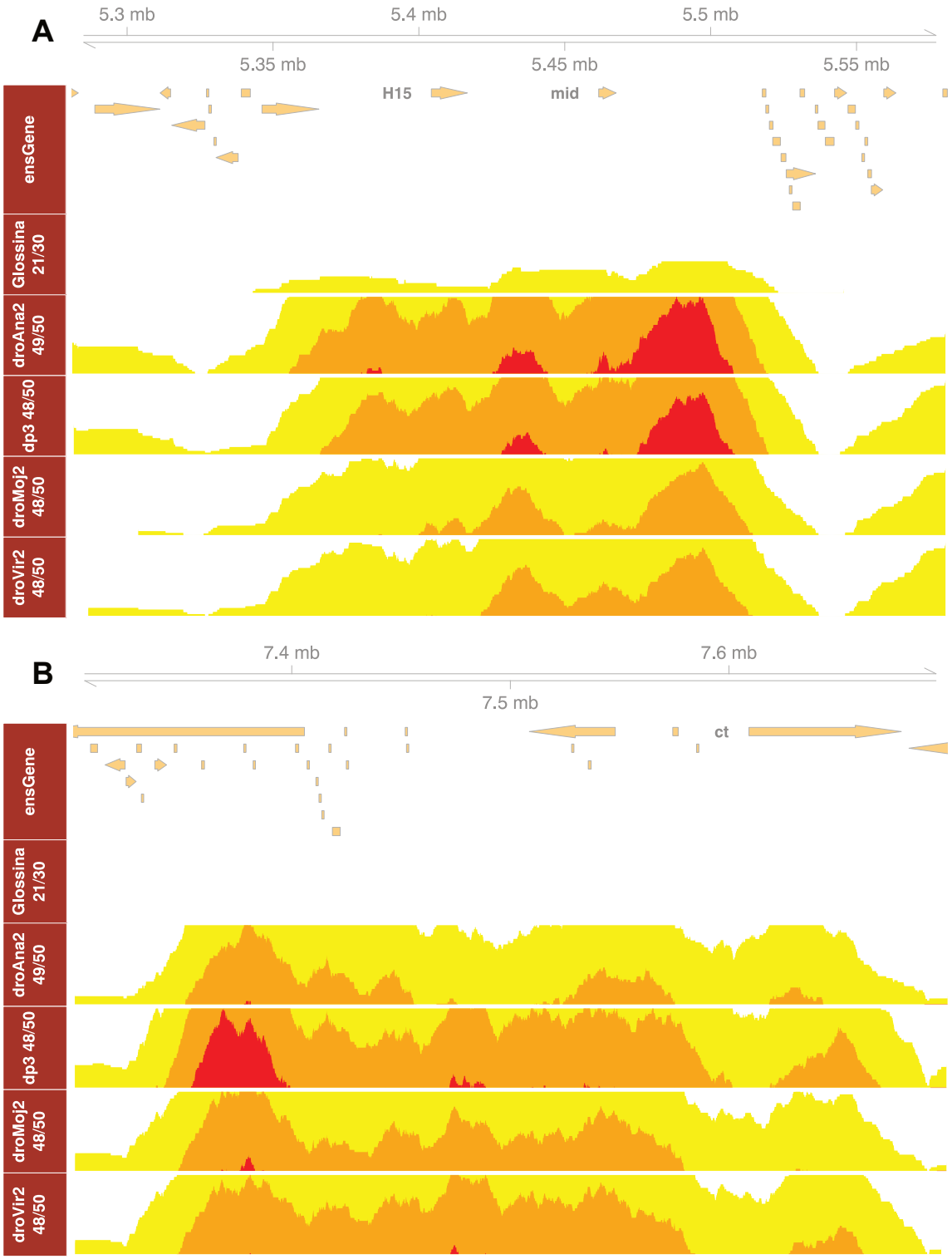

Figure 2.4 Horizon plots of CNE density at two loci containing key developmental genes. The CNE density in y range is cut into three segments and overlaid with three different colours representing the magnitude: yellow (the bottom segment), orange (the middle segment) and red (the top segment). (a) *H15* and *mid* genes are spanned by arrays of CNEs. Despite the much lower CNE density from *D. melanogaster* and *Glossina*, it reconstructs a CNE cluster boundary that is consistent with CNEs from other *Drosophila* species. (b) The CNE cluster around *ct* gene is missing in the comparison of *D. melanogaster* and *Glossina* since no CNEs are detected. It implies that this region undergoes a higher CNE turnover rate.

tubules (Nepveu, 2001). To locate the CNE clusters missing from *Drosophila* vs. *Glossina* comparison, we use the CNE clusters from *D. melanogaster* vs. *D. ananassae* as reference and compare them with the aforementioned retained CNE clusters. The genes within those missing CNE clusters are highly enriched for axon guidance and neuron development (Figure 2.3B). We then examine the CNE turnover rate (the speed of replacing old CNEs) of the 216 human genes from axon guidance (GO:0007411) term with both human and *Drosophila* as reference. The turnover rate is calculated by the reduction of the number of CNEs between two sets of CNEs. For human reference, we choose the CNEs set of human vs. mouse and human vs. zebrafish, while *D. melanogaster* vs. *D. ananassae* and *D. melanogaster* vs. *Glossina* are chosen for *Drosophila* reference. As shown in Figure 2.5, the axon guidance genes have significantly higher turnover rate than the other genes ($p < 1e-5$, Kolmogorov-Smirnov one-sided test) in both human and *Drosophila* lineages.

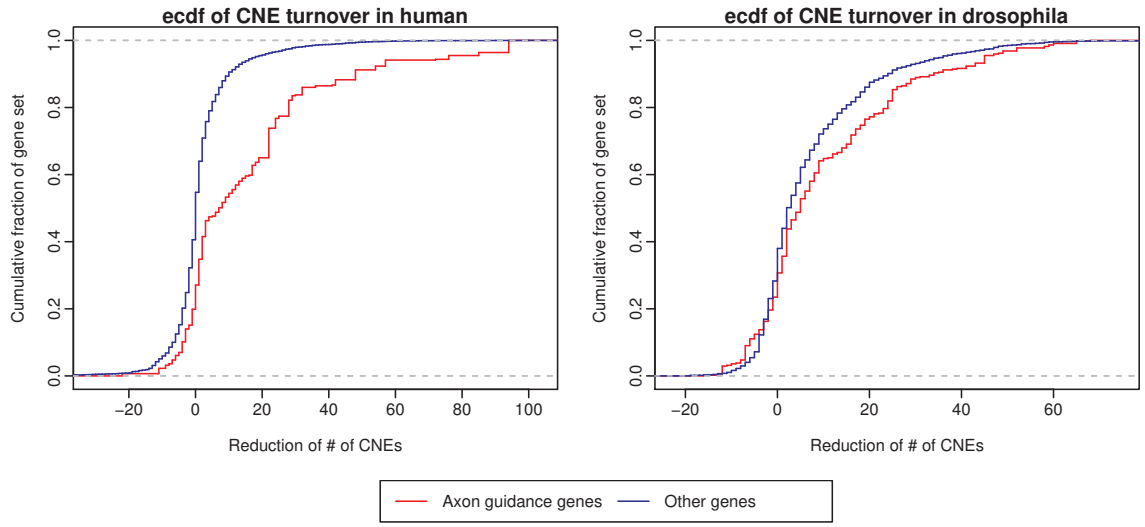

Figure 2.5 Cumulative distribution function of the changes of CNE number. For a 40kb window around each orthologous gene pair between human and *Drosophila*, we calculate the reduction of the number of CNEs for human (# of CNEs from human vs. mouse comparison minus # of CNEs from human vs. zebrafish comparison) and *Drosophila* (# of CNEs from *D. melanogaster* vs. *D. ananassae* comparison minus # of CNEs from *D. melanogaster* vs. *Glossina*) as reference. The axon guidance genes show a significantly higher degree of CNE number reduction, compared with the other genes ($p < 1e-5$, Kolmogorov-Smirnov one-sided test).

2.2.4 *CNEr* use case II: sea urchin CNEs

In this section we apply *CNEr* to the comparison of highly fragmented genome assemblies of two sea urchin species *Strongylocentrotus purpuratus* and *Lytechinus variegatus*. The purpose of this analysis is twofold. First, we want to demonstrate how well *CNEr* is able to call CNEs and their clusters in the case of highly fragmented draft genomes: the ability to perform this analysis on draft genome assemblies would show that our approach can be applied to a large number of available genomes, most of which haven't been assembled past the draft stage and are likely to remain in that state. Second, we want to ask if a third lineage, evolutionarily closer to vertebrates than insects but still lacking any shared CNEs with vertebrates, would exhibit the same patterns of noncoding conservation, thereby establishing their universal presence in Metazoa and providing an informative additional dataset for comparative studies of genomic regulatory blocks.

S. purpuratus is a popular model organism in cell and developmental biology. These two organisms have a divergence time of 50 Mya (Cameron et al., 2009) and historically moderate rates of sequence divergence, which makes them ideal for comparative genomics of regulatory elements. We identified 18,025 CNEs with threshold of 100% identity over 50 bp window. Despite the highly fragmented assemblies, we can clearly detect 808 prominent CNE clusters.

An especially interesting observation is the largest cluster we detected, at the *Meis* gene locus (Figure 2.6). The CNE density clearly marks the boundaries of CNE cluster. In Metazoa, *Meis*, one of the most well-known homeobox genes, is involved in normal development and cell differentiation. Tetrapod vertebrates have three *Meis* orthologs as a result of two rounds of whole genome duplication. The CNE cluster around *Meis2* is the largest such cluster in vertebrates (Sandelin et al., 2004). Remarkably, the cluster of CNEs around *Drosophila*'s *Meis* ortholog, *hth* (homothorax), is also the largest CNE cluster in *D. melanogaster* genome (Engström et al., 2007). It is currently unknown why the largest clusters of deeply conserved CNEs in three different metazoan lineages are found around the same gene, even though none of CNEs from one lineage has any sequence similarities to CNEs in other two. The most plausible explanation is that the ancestral *Meis* (hth) locus was already the largest such locus in the ancestral genome, and that CNE turnover lead to three separate current lineage-specific sets of CNEs.

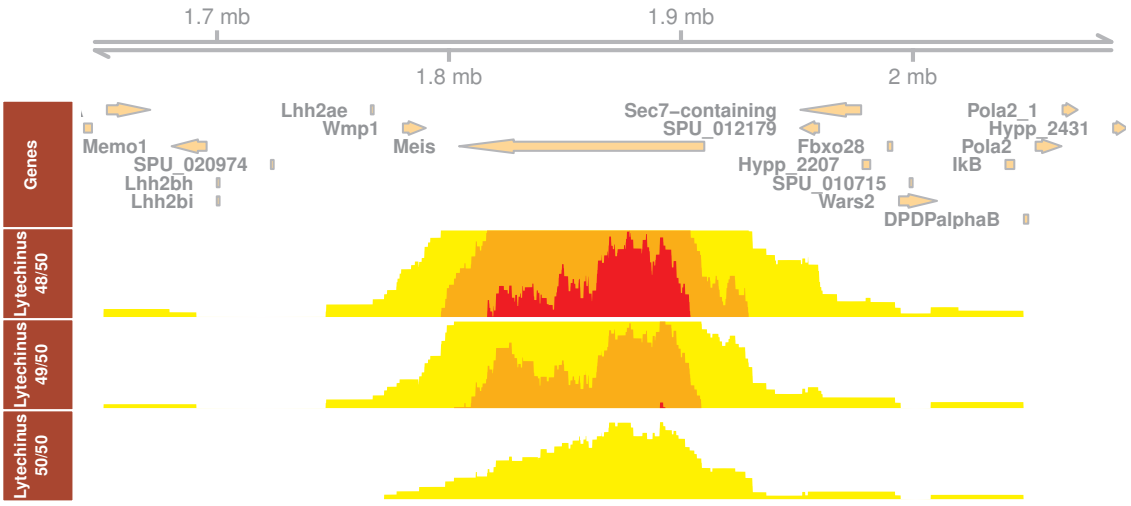

Figure 2.6 Horizon plot of CNE density at *Meis* locus on sea urchin *Strongylocentrotus purpuratus*. The threshold used for CNE identification is 100% identity over 50bp.

2.2.5 CNEs identified by *CNEr* reveals interesting sequence features characteristic of ultraconservation

It has been shown that vertebrate nonexonic CNEs are enriched in the TAATTA hexanucleotide motif, which looks like an extended recognition site for the homeodomain DNA-binding module (Chiang et al., 2008). With *CNEr*, we can easily verify the existence of TAATAA motif in CNEs of invertebrate species. In Figure A.4A, we consider CNEs identified by *CNEr* that are conserved between *D. melanogaster* and *D. virilis* over 98% for more than 50 nucleotides and plot them by increasing width using *heatmaps* package (https://bioconductor.org/packages/heatmaps/). The first two heatmaps confirm that CNEs are enriched in AT inside but exhibit a marked depletion of AT at their borders, consistent with what is known about their biology in vertebrates (Walter et al., 2005). Furthermore, TAATTA motif is enriched in insect CNEs. However, the motif seems to be extended further by flanking A/T nucleotides. When replacing A/T (W) with G/C (S), the heatmap pattern disappears. We asked whether this is a general property of CNEs in Metazoa and, using *CNEr*, proceed to the identification of CNEs that are conserved between (a) *C. elegans* and *C. briggsae* at 100% for more than 30 nucleotides (worm CNEs, see Figure A.4B), (b) *L. variegatus* and *S. purpuratus* at 100% for more than 50 nucleotides (sea urchin CNEs, see Figure A.4C). We observe that the same pattern does not hold in those cases, i.e. it appears like enrichment of CNEs in TAATTA is not a universal phenomenon but applies only to

insect and vertebrate elements. It would be interesting to investigate how and when during evolution this TAATTA-richness originated, and we believe that our pipeline is a powerful means towards this direction.

2.3 Methods and data

2.3.1 *CNEr* package implementation

CNEr is a Bioconductor package developed in R statistical environment, distributed under the GPL-2 licence for *CNEr* code, and UCSC Kent's licence for Jim Kent's C source code it builds on (Kent et al., 2002). Although *CNEr* supports compilation for both 32-bit and 64-bit systems across multiple platforms, it has limited functionality on the Windows platform due to the lack of the external sequence alignment software BLAT (Kent, 2002), which is required in the pipeline.

2.3.2 Overview of whole genome pairwise alignment

UCSC Genome Informatics (http://hgdownload.soe.ucsc.edu/downloads.html) provides the pairwise alignments between many popular species. However, there is a frequent need to produce pairwise alignments for novel genome assemblies for new species, or using specific assembly versions when they are not available from UCSC. This pipeline mostly requires external sequence aligners and UCSC Kent's utilities (Kent et al., 2002), and provides well-tested parameters for species with a varying degree of evolutionary divergence. In brief, first a sequence alignment software, LASTZ (Schwartz et al., 2003) or LAST (Kiełbasa et al., 2011), is used to find the similar regions between two repeat masked genomes. Then if two neighbouring alignments are close enough in the genome, they are joined into one fragment. During the alignment, every genomic fragment can match several others, and the longest one is kept. Finally blocks of alignments are grouped into stretches of synteny and form the so called "net" alignments in Axt format (Kent et al., 2003). *CNEr* comes with a vignette to demonstrate the whole pipeline.

2.3.3 Overview of Axt scanning algorithm

The Axt alignment scanning algorithm constitutes the central part of this package for the identification of conserved noncoding elements. Due to the massive manipulation of charac-

ters, we implemented this algorithm purely in C for performance reasons; it is available to the R environment through R's C interface. The minimal input is the Axt alignment and the ranges to filter out, i.e., the coding and/or repeat masked regions.

The Axt screening algorithm proceeds as in Algorithm 1. First, the Axt alignment is converted into a linked 'axt' data structure as implemented in Jim Kent's UCSC source code (Kent et al., 2002). The filtering ranges are encoded into a hash table, where keys are the chromosome/sequences names and values are pointers to the linked lists of coordinates ranges. We then iterate over the linked 'axt' alignments. For each alignment, we use a running window to scan the alignment with a step size of 1bp. Each base is searched against the filtering hash table and matched bases are skipped. Any segments above the identity threshold are kept. The overlapping segments are merged into larger pieces. This procedure produces a set of CNEs conserved between the two aligned genome assemblies.

```
Data: axt: Axt alignment;
filtersTarget: ranges to filter of target assembly;
filtersQuery: ranges to filter of query assembly;
W: the window size of the running window;
I: the minimal identity over the winSize;
Result: ranges of CNEs
HT ← new Hashtable from filtersTarget;
HQ ← new Hashtable from filtersQuery;
for a ∈ axt do
    initialise temp alignment t;
    for each running window w of size W in a do
        if HT contains w or HQ contains w then
            goto next w;
        end
        if identity of w > identity then
            append w to t;
        end
    end
    merge overlapping w in t;
    return t;
end
```

Algorithm 1: Scan axt alignments and identify the conserved noncoding elements

2.3.4 *CNEr* visualisation capability

PhastCons and phyloP are typically used to produce conservation scores from multiple sequence alignments, and the suitability of each depends on the application. The most important difference between the two is that the scores produced by phyloP reflect individual alignment columns, and do not consider conservation at neighbouring sites. PhyloP would be more appropriate than phastCons for evaluating signatures at individual nucleotide positions, while phastCons may be more suitable for detecting conserved elements overall since it directly models multibase elements. In comparison to ordinary conservation profiles (using phastCons and phyloP), which are available from other genome browsers, the profiles of density plots of CNEs revealed along the chromosome do not directly reflect conservation at the sequence / alignment level but display density distributions of CNEs on a larger scale. The output is qualitatively different from a sequence-based conservation plot (such as the conservation tracks in the UCSC genome browser mentioned earlier) and allows us to locate the large CNE arrays which are likely to flank some developmental regulatory regions (Engström et al., 2008).

Instead of using the standard density plot for CNE density (as implemented in e.g. Ancora browser, we introduce the horizon plot to increase the dynamic range of CNE density visualisation. The horizon plot provides a way of visualising the CNE density over several orders of magnitude and eliminates the need for multiple standard density tracks at different thresholds along the genomic coordinates. Instead, a relatively low conservation threshold is used, and multiple overlaid sections of the horizon plots will reveal peaks with different conservation density (Figure 2.4). We expand the functionality of "horizonplot" in *latticeExtra* package and integrate it into *Gviz* (Hahne and Ivanek, 2016), which is the plot engine used in *CNEr*.

2.3.5 Working with paired genomic ranges

In Bioconductor, the *GRanges* class defined in the *GenomicRanges* package (Lawrence et al., 2013) is an essential class that encodes the "start" and "end" position of ranges, as well as the chromosome identifier, strand designation and other metadata. Due to the nature of conducting pairwise comparison between species, we needed a class that stores two parallel *GRanges* classes, which represent the genomic coordinates information from each species. With direct inheritance from *Pairs* class of *S4Vectors* package, we created a *GRangePairs* class. The only restriction for these two *GRanges* objects is that they must have same

lengths. They can represent data from the same genome or different genomes. Since the CNE identification algorithm scans the conserved elements from axt alignment files, it is important to be able to manipulate the axt alignment efficiently in R. We build another Axt class on top of *GRangePairs* with shared element metadata, including Blastz score, alignment length and alignment sequence, for each pair. This *Axt* class can be especially useful for comparative genomics and phylogenetic footprinting (Tan and Lenhard, 2016). With the inheritance from *Pairs*, many common Bioconductor Vector APIs are preserved for convenient operations. More details about the specific methods defined for *GRangePairs* and *Axt* class are available in the documentation or vignette.

2.3.6 *Glossina* and sea urchin data

The *Glossina morsitans* genome assembly was obtained from Sanger Institute release December 2010 and the gene set version GmorY1.5 was acquired from VectorBase (https://www.vectorbase.org) (International Glossina Genome Initiative, 2014). This 366-Megabase *Glossina* assembly contains 13,807 scaffolds with a N50 value of 120 kb. This genome size is more than twice the size of *D. melanogaster* genome.12,308 protein-coding genes were predicted and the average gene size is almost double of that of *Drosophila*. The average exon and intron sizes are 491 bp and 1.6 kb, respectively. The whole genome pairwise alignment between *Drosophila* and *Glossina* is generated by our LASTZ pipeline with the parameter of *distance="far"*.

The sea urchin *Strongylocentrotus purpuratus* v3.1 and *Lytechinus variegatus* v2.2 genome assemblies and gene annotations were downloaded from EchinoBase (http://www.echinobase.org/Echinobase/) (Cameron et al., 2009). The number of scaffolds are 32,008 and 322,794, respectively. Due to the highly fragmented assemblies, the whole genome pairwise alignment was done with LAST pipeline with parameter of *distance="far"*.

2.4 Discussion

The advent of sequencing methodologies and the growing availability of genomes has brought the field of comparative genomics analysis into an unprecedented focus of interest. Aside from protein-coding genes, which have been the main focus of genomics and disease-associated studies so far, the role of variants lying on non-coding sequences is becoming increasingly important in a range of diseases, including cancer (Khurana et al., 2016). Since

the first report of extremely conserved regulatory elements, initially in human noncoding sequences (Duret and Bucher, 1997), the reasons for the emergence and possible roles of CNEs connected to their conservation levels are still largely unknown. Towards the direction of exploring their potential roles in genomes, there is a need to consistently produce sets of CNEs in a wide range of genomes. *CNEr* is the first freely available package in Bioconductor for large-scale identification, handling and advanced visualisation of sets of CNEs. The package incorporates functions and comes along with a detailed tutorial that allows the user to explore CNEs, going through the initial step of identifying conserved regions by scanning alignments to visualising the identified elements by horizon plots. The main algorithm that identifies the elements by scanning alignments is presented as well as several other novel classes that have been designed and implemented particularly for handling CNEs: the *Axt* class for efficient manipulation of axt alignments and the *GRangePairs* class for storing CNE coordinates from pairwise whole-genome alignments. Convenient functions to extract CNEs from GRBs and export CNE coordinates in genome browsers are also implemented and visualisation of CNEs is achieved by leveraging CNE density information. This visualisation strategy confirms our previous findings which suggest that CNE arrays cluster around genes important in development. Interesting sequence features of invertebrate CNEs are revealed by plotting via heatmaps the elements identified by *CNEr* in insects, worms and sea urchin. We applied our pipeline to the case of the Tsetse fly and sea urchin demonstrating in the latter case the efficacy of our pipeline in identifying CNEs in highly fragmented genomes.

2.5 Conclusions

In this chapter, we presented *CNEr* as an extremely efficient and flexible pipeline for CNEs identification, manipulation and visualisation. It is the only available tool that enables researchers to explore ultraconserved noncoding elements using arbitrary own criteria, in a unified framework. The package comes with detailed documentation and is addressed to a wide audience, ranging from people with little computational experience to the advanced user.

Chapter 3

TFBSTools: an R/Bioconductor package for transcription factor binding site analysis

The ability to efficiently investigate transcription factor binding sites genome-wide is central to computational studies of gene regulation. To facilitate our analyses of regulatory elements in this thesis and beyond, we developed *TFBSTools* (Tan and Lenhard, 2016), an R/Bioconductor package for the analysis and manipulation of transcription factor binding sites and their associated transcription factor profile matrices. *TFBSTools* provides a toolkit for handling TFBS profile matrices, scanning sequences and alignments including whole genomes, and querying the JASPAR database. The functionality of the package can be easily extended to include advanced statistical analysis, data visualisation and data integration.

3.1 Introduction

Transcription factor binding sites (TFBSs) on DNA play a central role in gene regulation via their sequence-specific interaction with transcription factor (TF) proteins (reviewed in Wasserman and Sandelin (2004)). Most individual TFBSs are 4-30 base-pairs (bp) wide, but are most often located in larger *cis*-regulatory regions of 50-200 bp. Analysis and identification of TFBSs is crucial for understanding the regulatory mechanisms of gene regulation. At present, the TFBS analysis functionality in R/Bioconductor (Gentleman et al., 2004) is limited and scattered across multiple packages. Here we describe the design

and functionality of the R package *TFBSTools*, which provides a unified and efficiently implemented suite of TFBS analysis tools. The package provides a number of functions for manipulating TFBS profile matrices and searching DNA sequence and pairwise alignments using them. We have ported all of the functionality of our popular TFBS Perl modules (Lenhard and Wasserman, 2002), retaining the equivalent class structure where possible, and expanded the functionality to provide efficient genome-wide analysis of TFBSs. Our implementation is tightly integrated with the existing Bioconductor core packages, enabling high-performance sequence and interval manipulation. A database interface for JASPAR2014 (Mathelier et al., 2014), JASPAR2016 (Mathelier et al., 2016), JASPAR2018 (Khan et al., 2017) and wrapper function for *de novo* motif discovery software are also provided.

3.2 Functionality of *TFBSTools*

3.2.1 Novel S4 classes defined in *TFBSTools*

To provide easy data storage, manipulation and exchange, we created several novel S4 classes (see Table 3.1 and Figure 3.1), and also defined an aggregate version of each class (e.g. *PFMatrixList*) to help manipulate sets of the corresponding objects. The design of these classes corresponds to classes in TFBS Perl modules, while remaining extensible in an object-oriented manner, adding new functionality and taking advantage of functional programming capabilities of R.

3.2.2 Operations with TFBS matrix profiles

To characterise the binding preference of a TF, the aligned sequences bound by the TF are aggregated into a position frequency matrix (PFM). From this matrix, another two matrices can be derived: position weight matrix (PWM, the most commonly used kind of position-specific scoring matrix) and information content matrix (ICM). PWM is a matrix of positional log-likelihoods normally used for sequence scanning and scoring against the motif, while ICM is mostly used in motif visualisation, e.g. for drawing sequence logos which can be easily done by the package *seqLogo* (Figure 3.1A). As a novel feature, in addition to matrix profiles, we have also implemented functionality for the manipulation of transcription factor flexible model (TFFM) profiles (Mathelier and Wasserman, 2013), which capture the dinucleotide dependence (Figure 3.1B).

Table 3.1 Novel **S4** classes defined in *TFBSTools*. There are equivalent classes in TFBS Perl module, which enables easy migration for the users of TFBS Perl modules.

Class	Description
PFMatrix	Position frequency matrix with additional information about this matrix; can be transformed to other matrices
PWMatrix	Position weight matrix with additional information about this matrix; used in sequence/alignment scan
ICMatrix	Information content matrix with additional information about this matrix; used in drawing sequence logos
XMatrix	A virtual position-specific score matrix class; all the above three objects are inherited from this virtual class
SiteSet	A container for transcription factor binding sites
SitePairSet	A container for pair of transcription factor binding sites from a pair of orthologous sequences
MotifSet	A container for storing the generated motifs identified by *de novo* motif discovery softwares

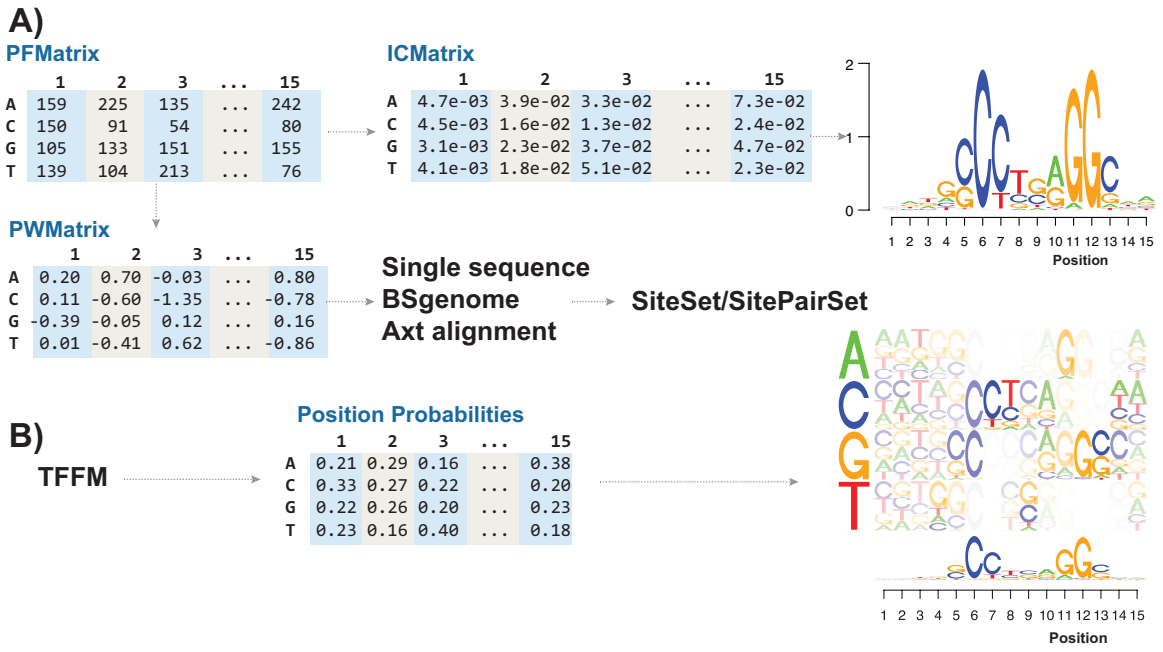

Figure 3.1 A common workflow and classes in TFBSTools. A) *PFMatrix* can be converted into *PWMatrix*, *ICMatrix*. *ICMatrix* produces the sequence logos. *PWMatrix* scans the single sequence or alignment to produce *SiteSet* object that holds transcription factor binding sites. B) *TFFM*: A virtual class for TFFM; *TFFMFirst* and *TFFMDetail* are derived from this virtual class. They can produce the position probabilities and the novel graphics representation of TFFM.

TFBSTools provides methods to perform the conversion between different types of matrices, providing a range of options and customisations. The highlights include: (a) A default pseudocount of 0.8 (Nishida et al., 2009) is used to eliminate the small or zero counts before log transformation, although a different pseudocount, or pseudocount function, for each column is possible; (b) Schneider correction, which is small sample correction to increase the uncertainty of pattern, for ICM is available (Schneider and Stephens, 1990; Schneider et al., 1986); (c) Unequal background nucleotide frequencies can also be specified.

TFBSTools additionally implements tools for comparing pairs of PFMs, or a PFM with IUPAC strings, using a modified Needleman-Wunsch algorithm (Sandelin et al., 2003). Quantification of the similarity between PFMs is commonly used for comparing a newly discovered matrix with existing matrices in the motif database, such as JASPAR, to determine whether the motif is related to known annotated motifs. .

The similarity between two PWMs can be quantified using three metrics: *normalised Euclidean distance*, *Pearson correlation* and *Kullback-Leibler divergence* (Linhart et al., 2008). Given two PWMs in probability mode, P^1 and P^2, where l is the length. $P^j_{i,b}$ is the values in column i with base b in PWM j.

1. Normalised Euclidian distance: this distance is between 0 (perfect identity) and 1 (complete dis-similarity).
   $D(P^1,P^2) = \frac{1}{\sqrt{2l}} \cdot \sum_{i=1}^{l} \sqrt{\sum_{b\in\{A,C,G,T\}}(P^1_{i,b} - P^2_{i,b})^2}$
2. Pearson correlation coefficient:
   $r(P^1,P^2) = \frac{1}{l} \cdot \sum_{i=1}^{l} \frac{\sum_{b\in\{A,C,G,T\}}(P^1_{i,b}-0.25)(P^2_{i,b}-0.25)}{\sqrt{\sum_{b\in\{A,C,G,T\}}(P^1_{i,b}-0.25)^2 \cdot \sum_{b\in\{A,C,G,T\}}(P^2_{i,b}-0.25)^2}}$
3. Kullback-Leibler divergence:
   $KL(P^1,P^2) = \frac{1}{2l} \cdot \sum_{i=1}^{l} \sum_{b\in\{A,C,G,T\}}(P^1_{i,b} \log \frac{P^1_{i,b}}{P^2_{i,b}} + P^2_{i,b} \log \frac{P^2_{i,b}}{P^1_{i,b}})$

In addition, *TFBSTools* also allows random profile generation by: (a) Sampling the posterior distribution of Dirichlet multinomial mixture models trained on all available JASPAR matrices; (b) Permutation of columns from selected PFMs. The availability of random matrices with the same statistical properties as selected profiles is particularly useful for computational/simulation studies, such as matrix comparison.

3.2.3 Sequence/alignment scanning with PWM profiles

TFBSTools includes facilities for screening potential transcription factor binding sites present in a DNA sequence (`searchSeq`), or conserved in a pairwise alignment.

When a pairwise alignment is available, it can be used to combine the TFBSs prediction with phylogenetic footprinting, which can in many cases reduce the false discovery rate whilst retaining a sufficient level of sensitivity, especially on proximal promoters (Lenhard et al., 2003). Alternatively, it can be used in combination with other data (e.g. ChIP-seq) to study the cross-species conservation properties of TF binding.

For genome-wise phylogenetic footprinting, *TFBSTools* can accept two *BSgenome* objects, and a chain file for *liftover* from one genome to another (`searchPairBSgenome`) or a novel S4 class *Axt* from our *CNEr* package (available from the Bioconductor website) for representing the axt alignments (`searchAln`). It can take up to 50 CPU hours to run `searchAln` on human-mouse pairwise alignment with the possibility of parallel computation, while `searchSeq` or `searchPairBSgenome` only needs several minutes. The computationally predicted putative TFBSs can be returned in GFF format or *GRanges* for downstream analysis.

3.2.4 JASPAR database interface

Since the release of JASPAR2014 (Mathelier et al., 2014), we have provided Bioconductor data packages, *JASPAR2014* and *JASPAR2016*, holding the profile matrices and associated metadata. To accompany the use of this data package for TFBS analysis, *TFBSTools* provides functions to enable efficient database querying and manipulation.

3.2.5 Use of *de novo* motif discovery software

TFBSTools provides wrapper functions for *de novo* motif discovery softwares. and seamlessly integrates the results back into R objects. Currently, support for MEME (Bailey et al., 2009) is implemented and reported motifs are stored in *MotifSet* object.

3.3 Conclusions and future directions

The Bioconductor *TFBSTools* package provides a full suite of TFBS analysis tools. The package allows the efficient and reproducible identification and analysis of TFBSs. In

combination with other functionality in Bioconductor, it provides a powerful way to analyse TF binding motifs on genome-wide scale. Further development will include an efficient implementation of scanning sequence/alignment with TFFM and capability of scanning multiple sequence alignment. A tutorial and additional use cases are available at Bioconductor website.

Chapter 4

The function of conserved noncoding elements: insights from the ameiotic *Adineta vaga* genome

Despite more than a decade of research on conserved noncoding elements (CNEs), the source of the extreme noncoding conservation remains unexplained. Some research proposed the model of negative selection of CNEs with mismatch in the homologous pairing. This, and several other hypotheses, require that the process on which selective pressure is exerted occur during meiosis or, more general, in the germline. The publication of a genome of anciently tetraploid genome of an ameiotically reproducing Metazoa, the bdelloid rotifer *Adineta vaga*, has given us an non-obvious opportunity to examine if the purifying selection that acts on CNEs still active in an organism which has been reproducing asexually for tens of millions of years. In this chapter, we developed a novel CNE identification approach for duplicated regions in *Adineta vaga* genome and any other genome harbouring detectable genomic traces of one or more whole-genome duplications. We find numerous intragenomically conserved CNE clusters around the developmental regulatory genes, suggesting that meiosis is not necessary for CNE conservation, and that the primary function of CNEs is developmental gene regulation.

4.1 Introduction

Metazoan comparative genomics has revealed an abundance of DNA that is both noncoding and under strong negative selective pressure. Much of this DNA is accounted for by numerous long stretches of deeply conserved sequence known as conserved noncoding elements (CNEs) (Bejerano et al., 2004; Sandelin et al., 2004; Woolfe et al., 2005). CNEs have been identified in multiple metazoan lineages and many have remained extremely well conserved over more than 400 million years of evolution. Further, many CNEs exhibit levels of conservation that even exceed those observed in protein-coding genes (Bejerano et al., 2004). Transgenic reporter assays have shown that many CNEs function as regulatory elements capable of driving complex spatiotemporal patterns of gene expression (Bhatia et al., 2014; Navratilova et al., 2009; Woolfe et al., 2005), however, no known source of selective pressure is able to completely account for their observed levels of conservation.

CNEs are essentially single copy in the haploid genome (Bejerano et al., 2004), prompting the investigation of CNE dosage sensitivity. It was found that in healthy cells, CNEs are generally depleted from copy number variants (CNVs) and segmental duplications (SDs) (Chiang et al., 2008; Derti et al., 2006; McCole et al., 2014). Further, the depletion of CNEs in *de novo* CNVs (CNVs that have passed through the germline at most once) lead some researchers to suggest that this depletion is due to rapid selection against cells with CNVs containing CNEs (McCole et al., 2014). In light of these results, it has been proposed that CNEs have a role in monitoring the copy number of the genome, potentially through the pairing of homologous CNEs followed by the initiation of apoptotic processes upon detection of mismatches or copy number changes (McCole et al., 2014). If true, this model would provide a long sought-after source of purifying selection that could explain the extreme levels of conservation seen within CNEs; however, there is no evidence for CNE pairing or interaction with mismatch sensing proteins. Also, while CNE duplication is apparently not tolerated in *cis*, in vertebrates there is a significant number of paralogous CNEs left over after each round of whole-genome duplication (Dong et al., 2010; Kikuta et al., 2007b), which does not agree straightforwardly with the hypothesis of cells' dosage sensitivity to individual CNE copy number.

Here we take advantage of the evolutionary history of the genome of the ameiotically reproducing bdelloid rotifer, *Adineta vaga*, to investigate this hypothesis. Rotifers are minute freshwater invertebrates commonly found in lakes, streams and ponds. Despite extensive observation, neither males, hermaphrodites nor meiosis have ever been reported in rotifers.

In fact, a partial assembly of the *A. vaga* genome revealed a chromosomal structure that is incompatible with pairing of homologous chromosomes (Flot et al., 2013), suggesting either that rotifers have evolved ameiotically for millions of years, or that they undergo atypical meiosis in which chromosome segregation occurs without the requirement for homologous pairing (Signorovitch et al., 2015). If the primary role of CNEs in metazoa is to identify potentially deleterious copy number changes during homologous chromosome pairing, we expect that they will be absent from the *A. vaga* genome.

Typically CNEs are detected by scanning pairwise genome alignments for runs of conserved non-coding sequence, however, in the case of *A. vaga*, this is not possible as there are no sequenced genomes from sufficiently closely related species to compare to the *A. vaga* genome. To overcome this setback we exploit the unusual structure of the *A. vaga* genome, which is made up of collinear blocks that are divisible into allelic (recent homologs) and ohnologous (ancient whole genome duplication-derived homologs) pairs (Flot et al., 2013). The ohnologous pairs are on average 75.1% identical at the nucleotide sequence level and are assumed to have evolved independently for at least 35-40 millions of years (Hur et al., 2009). In the absence of meiotic recombination, comparing these regions using pairwise alignment and subsequent CNE detection should therefore be equivalent to comparing orthologous regions between distantly related species.

In this chapter, we describe a pipeline for the identification of CNEs, within a single species, surrounding paralogous genes derived from whole genome duplication (WGD), and validate its efficacy using the zebrafish genome. We then apply this pipeline to the *A. vaga* genome and successfully identify CNEs in clusters surrounding developmental regulatory genes, suggesting that the principal function of CNEs is regulation of developmental gene expression rather than copy number sensing.

4.2 Results

4.2.1 A control analysis: CNEs from duplicated zebrafish regions

No information is available on the existence of ordinary, cross-species conserved CNEs in *A. vaga*, so we needed an independent verification that our pipeline functions as intended. Before attempting to identify CNEs in collinear regions of the *A. vaga* genome, we therefore sought to assess the validity of identifying CNEs in independently evolving homologous regions within a single species whose cross-species CNEs are already well characterised.

The zebrafish genome is well suited for this analysis as it has a high quality assembly, numerous sets of CNEs defined from multiple species comparisons, and most importantly, it underwent WGD approximately 300 Mya (Figure 4.1). CNE identification was focussed on recently duplicated regions, anchored by genes which have a 1:2 orthology mapping between human and zebrafish. The duplication of these regions is known to have arisen from the most recent zebrafish WGD (3R WGD), and thus the paralogs within them have been independently evolving for approximately 300 million years. Table 4.1 lists the number of CNEs identified from both the zebrafish self-alignment and the human-zebrafish alignment at multiple thresholds. Generally, the more stringent the threshold, the fewer CNEs we detect, however at the most stringent identity thresholds (98% and 100% over 50bp), the number of CNEs identified increases. This is due to the fragmentation of long CNEs into short stretches of highly similar sequence, as confirmed by the decreasing total CNE widths at higher thresholds. From the self-alignment, we identified 4,275 CNEs at 70% sequence identity over 50bp. This is a much smaller set of CNEs compared to the standard human-zebrafish comparison (Table 4.1), which is not surprising as the zebrafish self-alignment is limited on only the recently duplicated regions of the genome which have retained both paralogous copies of a gene, and even there most of the CNEs conserved between human and zebrafish will survive in only one copy (Dimitrieva and Bucher, 2012). The reason for this is not clear, but there are several properties of the retained duplicated ("ohnologus") clusters of CNEs that make the intragenomic CNE identification work. First, for the cluster to be retained in two copies after WGD, the target gene must be retained in two copies. Second, the GRB target genes are retained in two copies after WGD more often than most other genes, especially the ubiquitously expressed ones. The reason for this is postulated to be the ease of regulatory sub- or neofunctionalisation of the two gene copies, because the total length of regulatory sequence, divided over many dozens of CNEs and other enhancer in the GRB, is larger than the length of the genes protein-coding sequence, and individual enhancers are much more independent. Third, the divergence of regulatory elements at the two loci is stochastic, leading to slow decrease in conservation of CNEs across the two paralogous loci over time. Zebrafish 3R WGD was recent enough that more than 4000 paralogous pairs of CNEs are still present, but vertebrate genomes also contain a smaller number of clearly detectable paralogous CNEs from the two ancestral rounds of WGD (Dong et al., 2010). The span of a cluster of intragenomic CNEs will be an intersection of the spans of two paralogous GRBs, so it will most often be shorter than one or both of them.

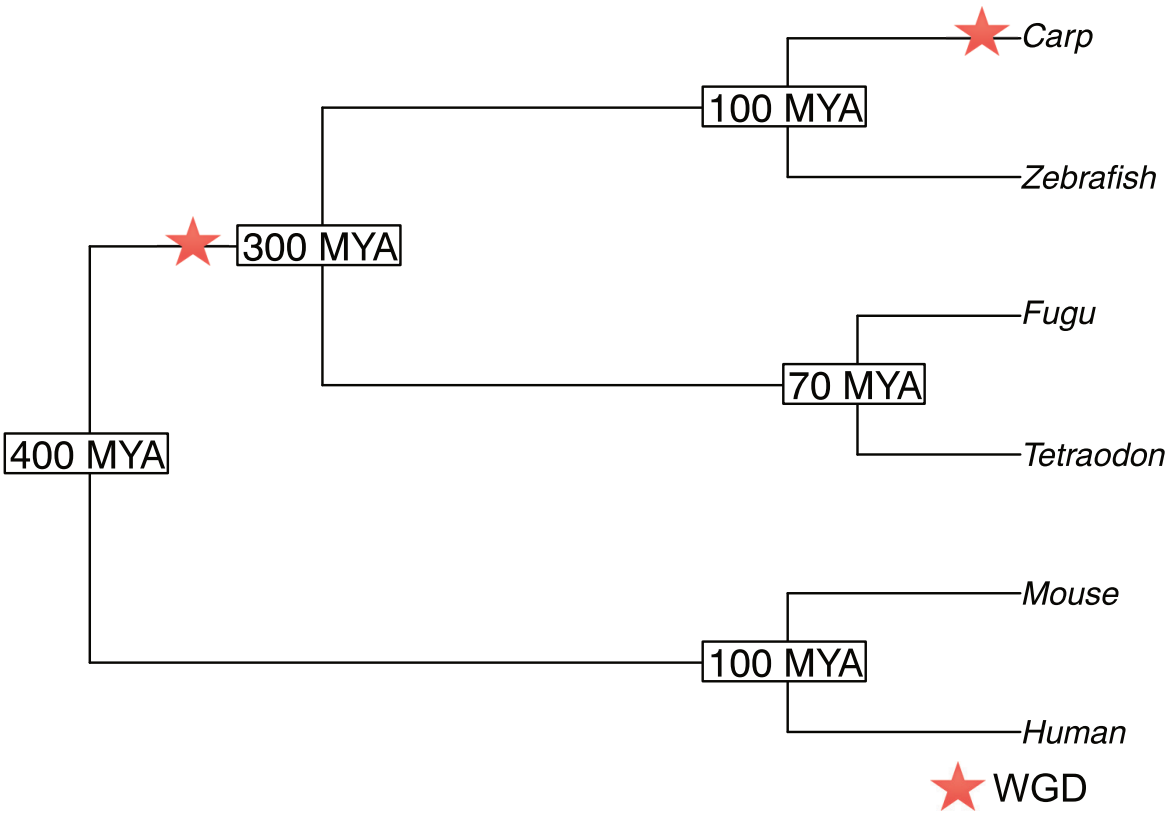

Figure 4.1 The species tree of teleost fish and human, mouse. Phylogenetic tree and the ages of the nodes are based on the data from TimeTree (Hedges et al., 2006). The red stars indicate the WGD. One is the third-round (3R) for fish lineage and one is the fourth-round (4R) for common carp lineage around 8 Mya.

To confirm the authenticity of these CNEs, we applied the standard CNE quality controls from the *CNEr* package. CNE widths are known to follow a scale-invariant property (Salerno et al., 2006), a property which the zebrafish self-alignment CNEs also exhibited with the cumulative distribution of CNE widths following a linear correlation for CNEs from length 100 to 500 (Figure B.1). Another landmark feature of CNEs is their non-random distribution across the genome, occurring in clusters surrounding developmental genes (Sandelin et al., 2004; Woolfe et al., 2005). The distribution of the zebrafish self-alignment CNEs is also not random, forming definite clusters (Figure B.2). In comparison to the rest of the genome, zebrafish CNEs have an increased AT content, with a sharp spike in GC content at both the 5' and 3' boundaries (Figure B.3A). Our zebrafish self-alignment CNEs also exhibited these features, with a particularly strong sequence composition boundary effect evident (Figure B.3B). In light of these results, we are confident that our self-alignment based CNE identification successfully yields valid CNEs.

Table 4.1 Summary of zebrafish CNEs from two approaches

Minimum identity	# of CNEs	total width of CNEs	# of zebrafish:human
70% over 50 bp	4275	617024	36649
80% over 50 bp	3330	464330	7960
90% over 50 bp	1563	307362	3209
96% over 50 bp	1108	259301	1232
98% over 50 bp	1148	243420	741
100% over 50 bp	1576	203960	311

To compare the CNEs identified from the zebrafish self-alignment to CNEs from the standard zebrafish:human pairwise comparison, we uploaded the CNEs into Ancora to visualise their genomic distribution and corresponding CNE density (Engström et al., 2008). Inspection of both sets of CNEs at the whole chromosome scale reveals that the self-alignment method is capable of recovering a number of CNE clusters identified using the standard CNE calling pipeline (Figure 4.2). On the zebrafish chromosome 1, we identified the most prominent CNE clusters and manually annotated their possible target genes. Three CNE clusters were recovered by the zebrafish self-alignment, each of which targets a developmental transcription factor gene: *uncx4.1*, *dachc* and *sox1b* (highlighted in red). In each case the density of the CNEs from the zebrafish self-alignment closely approximates the density profile of the standard zebrafish-human CNEs. The unrecovered CNE clusters (target genes in black) were not detected in the zebrafish self-alignment for various reasons. *mnx2b* has an ohnologous gene *mnx2a* on chr9, but the corresponding ortholog in human is not available from ENSEMBL. *efnb2b*, and its ohnolog *efnb2a* on chr9, have an ortholog in humans, however, no self-alignment was retrieved between these two regions. *helt*, *mab21l2*, *bnc2*, *smad1*, *pou4f2*, *hand2* were not recovered because of the lack of ohnologs in zebrafish.

Examining the CNEs identified from the self-alignment at the single locus scale further highlights the degree to which self-alignment based CNE identification recapitulates the results of the standard CNE identification pipeline. Figure 4.3A shows a 1Mb region around the human *TLE3* gene, a transcriptional co-repressor which functions in the Notch signaling pathway and regulates cell fate during development (Liu et al., 1996). This gene is encompassed by an array of CNEs identified from the standard human-zebrafish pairwise alignment. The zebrafish CNEs reside on two chromosomes, chr7 and chr18, surrounding the two corresponding ohnologs (duplicated in the last zebrafish WGD), *tle3b* and *tle3a* respectively. Figure 4.3B and Figure 4.3C show the distribution of human-zebrafish and

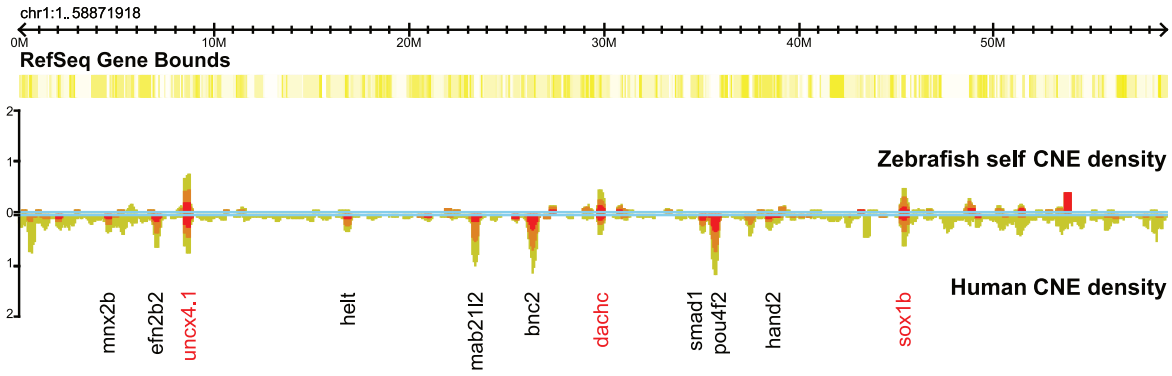

Figure 4.2 Comparison of noncoding conservation landscape from zebrafish self-alignment and standard zebrafish-human pairwise comparison on zebrafish chromosome 1. The self-alignment approach is capable of recovering at least three CNE clusters with target genes: *uncx4.1*, *dachc* and *sox1b* (in red).

zebrafish self CNEs around the two zebrafish copies of *tle3*. As we can see, our approach successfully detects CNEs from each of the two duplicated regions. These CNEs form clusters around their target genes and closely resemble the CNE density from a standard human-zebrafish pairwise comparison. The same pattern can be observed at many other loci, including known target genes *PAX6* (*pax6a*/*pax6b*) and *DACH1* (*dachc*/*dachd*).

To evaluate the ability of the self-alignment method to recover duplicated CNEs genome-wide, first a set of duplicated zebrafish CNEs were identified (defined as two zebrafish CNEs on different chromosomes which both map to the same human CNE) (Figure 4.4A). The recovery rate of the self-alignment method was then defined as the proportion of duplicated zebrafish CNEs that are also identified by the self-alignment method. Figure 4.4B shows the recovery rate of duplicated CNEs as a function of the minimal overlap required to consider two zebrafish-human CNEs duplicated. With a minimal overlap of 1bp, 58.55% of the duplicated CNEs are recovered by our approach. While the number of duplicated CNEs decreases with increasing minimal overlap, the percentage of recovered CNEs improves up to 75%. Taken together, it appears that the self-alignment method identifies a limited, but valid set of CNEs. Since our aim was to identify the presence of any CNEs in *A. vaga*, and not to create a comprehensive catalog of CNEs, the method is sufficiently sensitive to apply to the *A. vaga* genome.

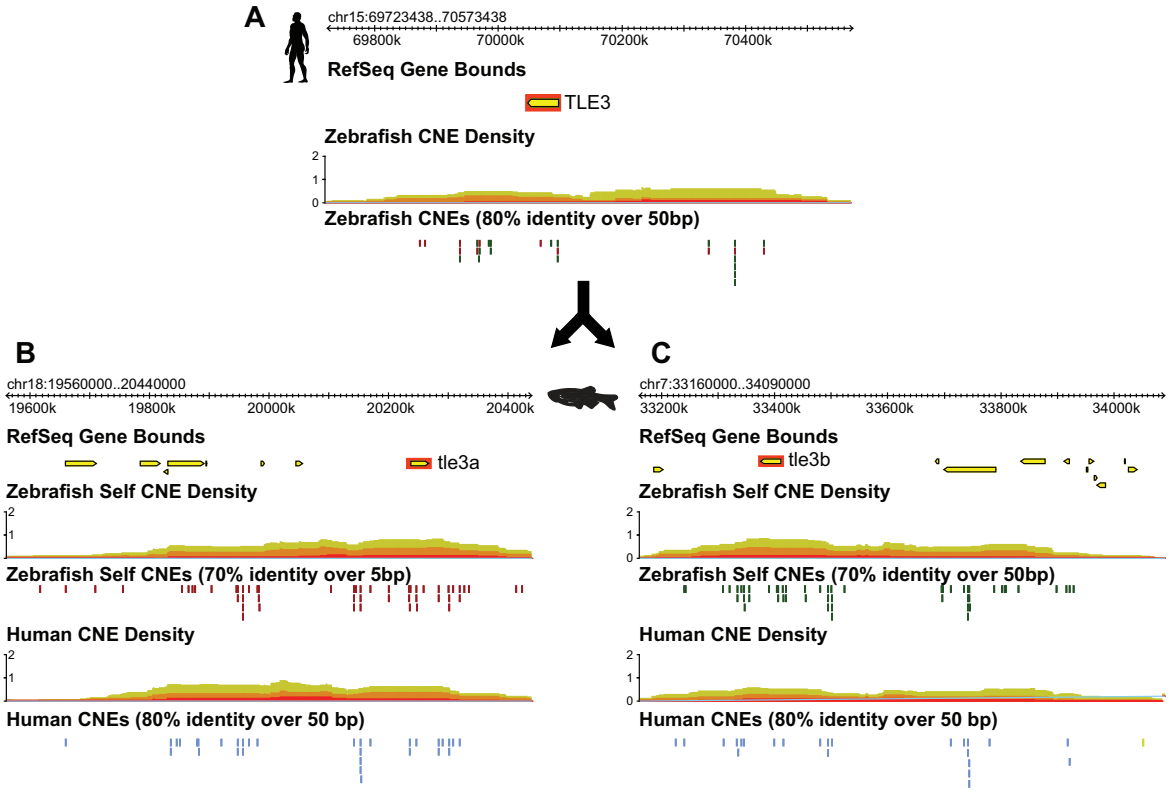

Figure 4.3 Arrays of CNEs around *TLE3* gene. (A) The CNEs around *TLE3* human gene are detected from human-zebrafish comparison. (B) and (C) *tle3a* and *tle3b* genes, on zebrafish chromosome 7 and 18, are duplicated from teleost fish WGD. The CNEs identified by from self-alignment approach recapitulates the distribution of CNEs from standard pairwise comparison of human-zebrafish.

4.2.2 CNEs from collinear regions of *A. vaga*

As discussed in (Flot et al., 2013; Signorovitch et al., 2015), *A. vaga* does not undergo homologous chromosome pairing or recombination and thus homologous loci have been evolving independently for millions of years. It is therefore possible to compare paralogous regions in *A. vaga* as one would compare orthologous regions between two diverged species. Initially, *A. vaga* collinear blocks were divided into ohnologous and allelic pairs as described in (Flot et al., 2013) and Methods and Data. We identified 740 allelic blocks and 836 ohnologous blocks, containing 14,764 genes and 12,181 genes, respectively. The median sequence identity of gene pairs within allelic blocks is much higher than those in ohnologous blocks (98.60% vs. 75.10%) (Table B.1). The median sequence identity is essential for choosing the threshold for CNE detection. We applied the whole genome self-alignment approach to *A. vaga* and produced net alignments for both allelic and ohnologous regions

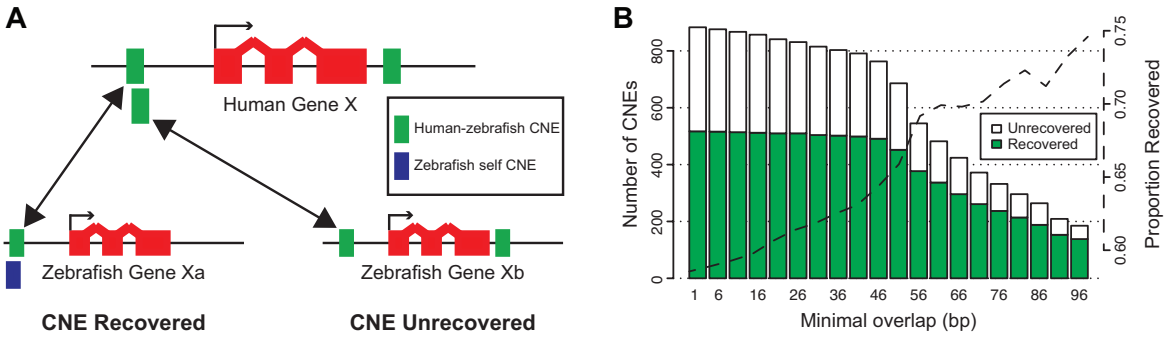

Figure 4.4 Evaluation of recovery capability of self-alignment approach. (A) Illustration of evaluation method. For the duplicated zebrafish CNEs, it is considered recovered when there is an overlapping zebrafish CNE from self-alignment. This illustration plot was conceived by me and made by Alex Nash, who is a collaborator on this Rotifer CNE project. (B) The rate of recovery improves with the increasing required minimal overlap.

(see Methods and Data). The *CNEr* pipeline was then run on each set of net alignments to produce CNEs. As shown in Table 4.2, due to the high identity between allelic regions, a very stringent threshold had to be used (100% over 250bp) to yield a reasonable set of CNEs. The majority of both the allelic and ohnologous CNEs reside in intergenic regions, despite the high gene density of the genome. For all subsequent analyses ohnologous CNEs identified at 70% over 50bp and allelic CNEs identified at 100% over 250bp were used.

Table 4.2 CNE counts for *A.vaga*. CNEs are collapsed on the A.vaga genome prior to the counting. NC, not calculated due to high identity.

	Allelic		Ohnologous	
Threshold	All	Intronic	All	Intronic
70% over 50bp	NC	NC	9,255	2,445
80% over 50bp	NC	NC	9,557	2,398
90% over 50bp	NC	NC	5,614	1,399
96% over 50bp	NC	NC	1,508	457
98% over 50bp	NC	NC	1,195	325
100% over 50bp	NC	NC	1,136	235
100% over 100bp	123,931	33,366	381	74
100% over 250bp	11,454	2,789	88	15

The *A. vaga* genome assembly is highly fragmented and as such it is difficult to assess the degree to which the identified CNEs cluster, however for the six largest scaffolds it appears that ohnologous CNEs cluster to a similar degree as the zebrafish self CNEs. On the other hand, the clustering of allelic CNEs is less obvious on some scaffolds, such as scaffold_2 and

scaffold_4 (Figure B.5). Visual inspection of loci containing key regulators of development illustrates this clustering and suggests that both the allelic and ohnologous CNEs form equivalent clusters around developmental genes to those observed at their orthologs in other genomes (Figure 4.5).

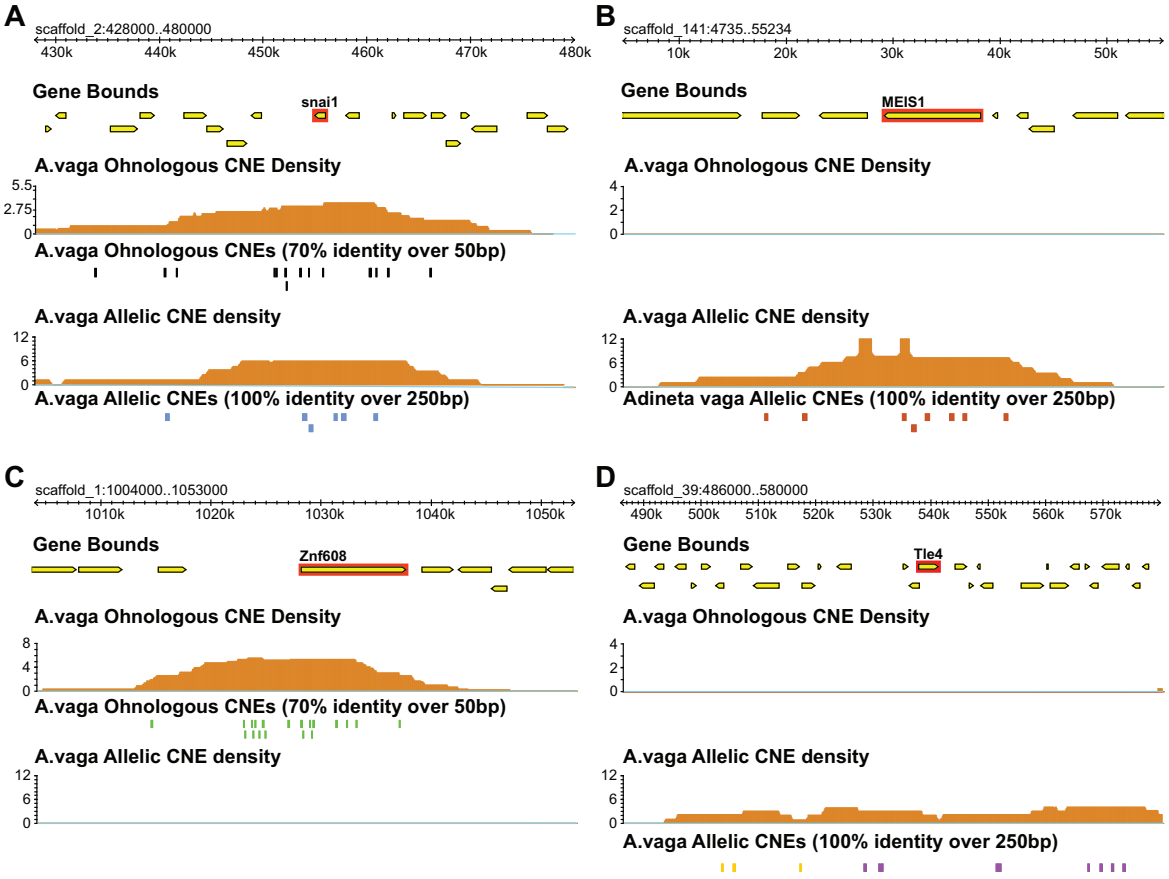

Figure 4.5 Examples of loci around some key developmental genes. (A) *snail1* has both ohologous and allelic CNEs. (B) *MEIS1* only has allelic CNEs. (C) Ohnologous CNEs are detected around *Znf608* gene. (D) Homologous gene of *TLE3*, *Tle4* has weak CNE distribution around it.

Gene ontology (GO) enrichment analysis of the genes closest to *A. vaga* CNEs highlighted numerous nominally significant developmental terms for both the ohnologous (Table 4.3) and allelic CNEs (Table 4.4). Due to difficulties in gene annotation, the total number of genes that each GO term could be assigned to was low. This limited our power to identify a robust statistical enrichment of any term, however the most significant terms conform with the previously identified preferential clustering of CNEs around developmental genes in

numerous species. Given more complete gene annotation it is likely that a similar, statistically robust enrichment would be observed in *A. vaga*.

Table 4.3 GO BP enrichment for genes around ohnologous CNEs

Term	p-value	Counts	GO ID
developmental process	3.4E-10	42/165	GO:0032502
.single-organism developmental process	4.11E-08	37/158	GO:0044767
. .multicellular organism development	6.3E-08	28/103	GO:0007275
. . .system development	1.34E-07	19/55	GO:0048731
. . . .nervous system development	1.03E-05	13/37	GO:0007399
. . . . . .brain development	6.38E-05	4/4	GO:0007420
. . . . . .neuron development	9.78E-04	5/10	GO:0048666
. . . . .sensory organ development	3.37E-05	6/9	GO:0007423
. . . . . .eye development	5.09E-07	6/6	GO:0001654
. . . . .muscle organ development	2.67E-03	3/4	GO:0007517
. . .cell differentiation	1.49E-03	12/51	GO:0030154
. .regulation of development, heterochronic	2.67E-03	3/4	GO:0040034

Table 4.4 GO BP enrichment for genes around allelic CNEs

Term	p-value	Counts	GO ID
developmental process	8.34E-07	51/165	GO:0032502
.single-organism developmental process	1.22E-06	49/158	GO:0044767
. .multicellular organism development	1.47E-07	38/103	GO:0007275
. . .system development	3.53E-06	23/55	GO:0048731
. . ..nervous system development	2.63E-04	15/37	GO:0007399
. . . . . .neuron development	1.81E-03	6/10	GO:0048666
. . . . .sensory organ development	8.36E-04	6/9	GO:0007423
. . .tissue regeneration	3.58E-03	5/8	GO:0042246
cell adhesion	2.76E-05	31/94	GO:0007155
. . .homophilic cell adhesion via plasma membrane adhesion molecu	3.10E-07	19/36	GO:0007156

4.3 Discussion

In this chapter, we proposed an approach for duplicated regions derived from WGD within a species. To validate this novel approach, we evaluate the performance on the well-studied zebrafish genome, which underwent a WGD 300 Mya. By focusing on the duplicated loci around the paralogous genes, we were able to recover more than half of the duplicated

zebrafish CNEs. The identified CNEs from self-alignment of zebrafish genome exhibits the known features of CNEs. The width of CNEs follows a power law distribution. The distribution of CNEs along the chromosome is not random, but forming clusters. There is a clear depletion of G/C at the boundaries of CNEs. All this evidence suggests we are able to identify a genuine set of CNEs. Although this set of CNEs are hardly a complete set and they tend to be more noisy than the CNEs from a standard cross-species pipeline, this approach is sufficient for our purpose of demonstrating the existence of CNEs in *A. vaga* genome.

In the *A. vaga* genome, we applied this method on two types of duplicated regions: allelic and ohnologous regions. Due to the extremely high sequence similarity of allelic regions, we had to use a very stringent threshold (100% over 250bp) for searching CNEs. We still yielded many more allelic CNEs than ohnologous CNEs. From the genomic distribution, these allelic CNEs are less convincing, compared to ohnologous CNEs. This is expected since we are comparing regions with 98.60% sequence identity, and inevitably such high sequence similarity impairs the power of comparative genomics. This level of sequence similarity (98.60%) is comparable to that of human and gorilla, for which it is very difficult to identify CNEs. This also suggests that the loss of sexual reproduction and meiosis in *A. vaga* is relatively recent. Nevertheless, we observe that both allelic and ohnologous CNEs cluster around developmental genes. This is further proved by the GO enrichment analysis of genes within CNEs clusters. A locus can have either allelic or ohnologous CNEs, or both of them. We wonder if there is any pattern in terms of the preference of allelic or ohnologous CNEs. Among the 258 and 289 GRBs identified from ohnologous and allelic CNEs, respectively, 98 of them overlap. We further performed GO enrichment for allelic and ohnologous specific GRBs. Interestingly, allelic specific GRBs are enriched in some general terms, such as metabolic process and sigalling, with few development terms. On the contrary, ohnologous specific GRBs are still enriched in development terms. This also suggests the difficulty of detecting CNEs from highly similar allelic regions. The existence of CNEs in the *A. vaga* genome, especially those between allelic copies of chromosomes, suggests that primary function of CNEs is the regulation of developmental genes rather than copy number sensing during homologous recombination.

Why are then CNEs depleted from CNVs and SDs? Duplication of a part of an array of enhancers is very likely to cause dysregulation of the target gene. Given the key developmental role of these target genes, which include most of the transcription factors which regulate development, is likely poorly tolerated. The mis-regulation of the target gene could be due

to alteration of the TAD structure at CNVs, allowing enhancers to contact genes other than their targets, or preventing correct looping to the promoters of their intended target genes.

4.4 Methods and data

4.4.1 Genomic data

The assembly and gene annotation of zebrafish (danRer10) were downloaded from UCSC. The assembly and gene annotation of *A.vaga* (version 2.0) were downloaded from http://www.genoscope.cns.fr/adineta/data/ (Flot et al., 2013). The provided gene annotation only contains the gene structure. To annotate the putative gene function, we first ran BLASTP (v2.6.0) (E-value: $1e-5$) search against manually curated UniProt/Swiss-Prot (14/07/2016 snapshot). We annotated the protein domain functions and putative GO terms with InterProScan (v5.19-58.0) (Jones et al., 2014). Prior to the whole genome alignment, *A. vaga* assembly was repeat masked with RepeatMasker (v4.0.5) (Smit et al., 2015) against all the species in Repbase (v20140131) (Jurka et al., 2005), as well as the *de novo* library constructed by RepeatModeler (v1.0.8).

4.4.2 Whole genome self-alignment and CNE detection for zebrafish

We aligned the zebrafish genome with itself using LASTZ (v1.02.00) (Schwartz et al., 2003). The most fundamental parameter during the alignment pipeline is the distance option. This parameter controls the scoring matrix and criteria of forming an alignment during the alignment procedure. Although "far" is usually chosen for species divergent than 100 Mya and the whole genome duplication of zebrafish happened 300 Mya, we found it difficult to finish the self-alignment for zebrafish within reasonable time. Setting distance option to "medium" during the alignment pipeline might miss some alignments, however, it has little impact on the most conserved sequences. We focused on the genomic regions which are 1Mb windows around the orthologous zebrafish genes from human (ENSEMBL 86) (Yates et al., 2016). The homology type of 1:2 mapping between human and zebrafish was chosen since these pairs are known to appear from the last WGD. In total, we retained 2444 pairs of paralogous zebrafish genes after filtering out close paralogous genes. The alignments within these regions were used to detect CNEs with the standard *CNEr* pipeline.

4.4.3 Collinear regions detection for *A. vaga*

The aim is to identify the homologous regions with homologous genes as anchors. In brief, we first did a all-against-all homolog searching for *A. vaga* protein sequences with BLASTP (v2.6.0) (-a 12 -e 1e-10 -b 5 -v 5 -m 8) (Camacho et al., 2009). Then the collinear regions, along with the synonymous and nonsynonymous rates of homologous genes, were estimated by the package MCScanX (Wang et al., 2012). The collinearity of each collinear region is defined as the number of collinear genes divided by the total number of genes within the collinear region. A collinear region is classified as an ohnologous region when the ratio of synonymous rate to collinearity is larger than 0.5 (Figure B.4).

4.4.4 Whole genome self-alignment and CNE detection for *A. vaga*

The *A. vaga* is highly fragmented with 38,875 scaffolds. Hence, we chose LAST (v830) (Kiełbasa et al., 2011) to align the assembly to itself. The sequence database was built with the option "-uNEAR". When a sequence is aligned to itself, the full alignment from LAST contains the trivial self-alignment and mirror-image pair. These naive alignments were filtered out before the following steps. With the previously identified ohnologous regions or allelic regions, we subset all the alignments within the regions and applied the standard chaining and netting steps. Then we conducted the standard CNE detection pipeline of *CNEr* on the produced netAxt files.

4.4.5 Data availability

- The zebrafish CNEs from self-alignment is available at Ancora browser http://ancora.genereg.net/cgi-bin/gbrowse/danRer10
- The *A.vaga* CNEs from self-alignment is available at http://ancora.genereg.net/cgi-bin/gbrowse/adiVa2

Chapter 5

Genome and regulatory elements of the european common carp (*Cyprinus carpio*)

The common carp (*Cyprinus carpio*) is one of the most importance domesticated freshwater fish, which accounts for approximately 40% of aquaculture production worldwide. Apart from its economic importance, the common carp is also highly suitable for comparative genomics studies with the animal model zebrafish (*Danio rerio*). Here, we present a *de novo* common carp genome assembly along with a high quality genome annotation. We investigate the differential gene expression pattern for duplicated genes and also provide a comprehensive set of regulator elements for zebrafish and common carp.

5.1 Introduction

Among vertebrate model organisms, zebrafish genomics has reached a high level of quality and reliability due to the substantially improved genome assembly and the identification of protein-coding and non-coding genes in the recent years. What is still missing is the annotation of *cis*-regulatory modules (CRM) comparable to that achieved for human and mouse through ENCODE and related projects. Researchers working on zebrafish have reached the stage at which such annotated CRMs are required to design reliable *cis*-regulatory experiments for the study of transcriptional regulation in embryogenesis and differentiation. Comparative genomics has proven to be a successful first approach to identify functional genomic elements, ranging from CRMs to non-coding RNAs and functional units of genome such as GRBs. However, one difficulty for zebrafish comparative genomics is that it is an out-

group relative to other teleost fish with well-characterised genomes (fugu, tetraodon, medaka and stickleback), from which it diverged around 300 Mya. This distance is larger than that between human and chicken, which is known to be too large to align most regulatory elements (Lenhard et al., 2003) other than CNEs, and reduces the predictive power of comparative approaches. A special framework was developed for detecting CNEs from phylogenetically isolated genomes (Hiller et al., 2013). Nevertheless, the availability of genomes of species more closely related to zebrafish is still required to obtain a comprehensive set of genomic elements under purifying selection. Common carp (*Cyprinus carpio*) is a commercially important species that shares last common ancestor with zebrafish about 100 Mya, making the distance comparable to that across placental mammals (Figure 4.1).

Starting from a carp assembly draft from H. Spaink lab at Leiden University, we first employed the genome annotation pipeline to identify the repeat regions and the structures of protein-coding genes. A phylogenetic analysis of *Hox* gene clusters suggests the satisfactory quality of genome annotation. With the criteria previously used for other vertebrates at comparable evolutionary distances, I detected an unexpectedly large number of CNEs between zebrafish and carp. After verifying the pairwise alignment and evolutionary rate between these two species, we conclude that CNEs in zebrafish and carp are significantly longer than those observed across equivalent evolutionary distances in tetrapod vertebrates.

5.2 Results

5.2.1 Carp genome assembly

I received the assembled draft genome and transcriptome from H. Spaink lab at Leiden University. They acquired the genomic DNA from a clonal double-haploid common carp, which was the same sample used for the first draft of assembly (Henkel et al., 2012). The assembly quality has been significantly improved using long-range sequencing information. With three paired-end libraries with various insert sizes, 4.5 million reads were yielded for each library on Illumina platform. PacBio long read platform was used to generate 6.8Gb of DNA isolated from nucleated red blood cells. All the reads were pooled together to form a *de novo* genome assembly of 1.38Gb with a N50 of 67kb in 80273 scaffolds. The carp genome has a GC content of 36.34%, which is close to the 36.53% of zebrafish, but lower than that of other sequenced teleost fishes, for example, 39.10% for tetraodon. A comparison of the quantitative assembly summary with the published Chinese Songpu strain (Xu et al., 2014) is

available in Table 5.1. Obviously, our genome size is much smaller than the reported 1.69Gb in Songpu strain. Pairwise whole genome alignment by LAST (Kiełbasa et al., 2011) reveals 1.13Gb syntenic regions between the two assemblies, which consists 82% and 67% of our assembly and the published Songpu strain, respectively.

Dataset	Songpu	Leiden (this study)	Leiden (2012)
Genome size	1.69Gb	1.38Gb	1.4-1.5 Gb
Scaffolds	2503	80273	511891
N50	1Mb	67kb	17kb
GC content	36.34%	37%	NA
Predicted genes	52610	50527	NA
Average gene span	12145bp	8316bp	NA
Predicted exons (total/per gene)	390620/7.48	387245/7.664	NA
Repeats	31%	32%	NA

Table 5.1 Comparison of genome assembly and annotation of common carp among this Wageningen strain from Leiden, the first draft of Leiden assembly, and the published Songpu strain (Xu et al., 2014)

So far, there is no indications of biological relevance to the difference in size of these two assemblies. There are several possible reasons for this discrepancy. One possible reason is the contamination of Illumina adapters in Songpu strain genome assembly, as reported in the blog of Graham Etherington (http://grahametherington.blogspot.co.uk/2014/09/why-you-should-qc-your-reads-and-your.html?m=1). Another possible explanation is that our genome is based on homozygous double-haploid carp, whereas a heterozygous individual is used in Songpu genome. Since common carp has undergone an additional, recent carp-specific WGD, the genome is expected to experience considerable genomic rearrangements during the re-diploidization, resulting in the phenotypic and genomic variations. This genomic instability is also observed in the recently published Atlantic Salmon, which has a salmonid-specific WGD in around 80 Mya (Lien et al., 2016). The sequencing technologies can also contributes to the difference. Both assemblies were sequenced primarily on Illumina short-read sequencing platform, which is certainly not optimal for *de novo* genome assembly. In this study, we applied the low coverage PacBio sequencing to improve our assembly. The explanation of the discrepancy will be more clear when we can apply the new single-molecule sequencing technology, such as Oxford Nanopore long read sequencing.

5.2.2 Genome annotation

We used a comprehensive pipeline to annotate the common carp genome. 31.8% percentage of the common carp genome is estimated to be repeats: 16.2% identified from existing other repeats library and 26.7% from constructed carp-specific repeats library. This is very close to the reported 31.3% in Songpu strain, but is significantly less than the repeat content in zebrafish (59.78%), A total of 50,527 genes (Table 5.1) were identified after integrating the evidence from *ab-initio* gene predictors, mRNA expression and homology to known proteins. The annotated genes were deposited in the NCBI database, representing a massive expansion of the previous few thousand annotated genes of common carp. This is almost double of the number of genes found in zebrafish. At the same time, the median gene body length is nearly half the size compared to zebrafish (5194bp vs. 12307bp). Both the intronic and intergenic regions are much smaller than those in zebrafish (331bp vs. 1029bp and 6296bp vs. 36181bp). This confirms that the carp genome is highly compacted, with a relatively lower content of repeats, as well as introns and intergenic regions. This also provides strong support for the fact that carp has undergone an extra WGD and that this tetraploid genome has retained a large proportion of the duplicated genes.

To further provide insight into the tetraploid nature of carp genome at the gene level after the fourth WGD and evaluate the quality of our annotation, a very common set of loci to check are the *Hox* clusters, since these well-defined large gene clusters are able to reveal the genome duplication and rearrangement events (Amores et al., 1998), and have a tendency to remain in two copies after each round of WGD. 48 *Hox* genes in zebrafish, excluding the pseudogenes, spread over 7 *Hox* clusters. The remnant of eighth cluster (*HoxDb*) has no *Hox* genes inside and can only be detected by a surviving miRNA gene (Woltering and Durston, 2006), hence it is excluded from our analysis. From our automatic annotation, 77 *Hox* genes were recovered, scattered over 37 scaffolds (Figure 5.1). For 26 out of 48 zebrafish *Hox* genes, we found two complete orthologs in the carp genome, and only two copies of *Hoxc12b* orthologs are absent in carp. Four of the predicted carp *Hox* genes are closer to the *Hox* genes in other vertebrates, such as chicken and mouse, instead of the *Hox* genes in zebrafish. In such cases we could not determine to which *Hox* cluster they belong and excluded them from Figure 5.1. When there is only one copy of ortholog found in common carp, the gene is plot as transparent on both carp clusters, because it's not possible to determine cluster it belongs to due to the fragmented assembly. To verify the orthologous gene prediction, we investigated the phylogenetic trees reconstructed for each *Hox* cluster. The paralogous genes

duplicated in the fourth WGD should be closer to each other than other *Hox* genes within the cluster if our annotation works well. In Figure 5.2, as expected, the leaves with same gene names, resulting from the fourth WGD, are generally the closest to each other in *HoxA* cluster. In the upper level, the duplicated genes from the third (teleost) WGD (with 'a' or 'b' as the last character of gene names) are also clustered together. The similar pattern can also be observed in *HoxB*, *HoxC* and *HoxD* clusters (data not shown).

5.2.3 The fate of recent duplicated genes

WGD creates genomic redundancy. It is an important genomic event in the evolution of genome complexity and diversity. Despite the increased number of genes after the WGD, a re-diploidization process usually follows subsequently, leading many of the genes back to diploid state (Wolfe, 2001). In general, three scenarios of duplicated genes can happen during the re-diploidization process (Li et al., 2015): 1) one copy of the duplicated genes becomes inactive due to mutation; 2) one duplicate gains some novel function while the other copy retains the original function; 3) each of the two duplicated genes keep subset of the function of their ancestor gene. Rapid gene deletion and differential gene expression with slow functional divergence is evident in both plant (Roulin et al., 2013) and yeast (Wolfe and Shields, 1997) genomes. However, the evolution of duplicated genes and differential gene expression pattern after a recent WGD in vertebrates are not well studied because the well studied WGDs are ancient and the genomes haven been turned back to functional diploidy. Common carp emerges as an ideal candidate for this study since the carp-specific WGD happened 8.2 Mya (Xu et al., 2014) and the genome is still largely tetraploid.

We first investigated the sets of genes that rapidly return to single copy. With the assumption of the doubled number of chromosomes and genes after the WGD, we consider the genes with unique annotated gene names as the re-diploidized genes. By excluding the 5744 genes with unknown gene names from the total 50527 genes, we collect 3805 single copy genes. A GO enrichment (Table 5.2) shows the significant terms under metabolic process, especially the "tRNA aminoacylation for protein translation". Aminoacyl-tRNA synthetase genes are known to be conserved and used in rooting the universal tree of life due to the ancient gene duplication and divergence (Brown and Doolittle, 1995; O'Donoghue and Luthey-Schulten, 2003). These translation machinery genes tend to remain single copy even after a such recent WGD in common carp lineage.

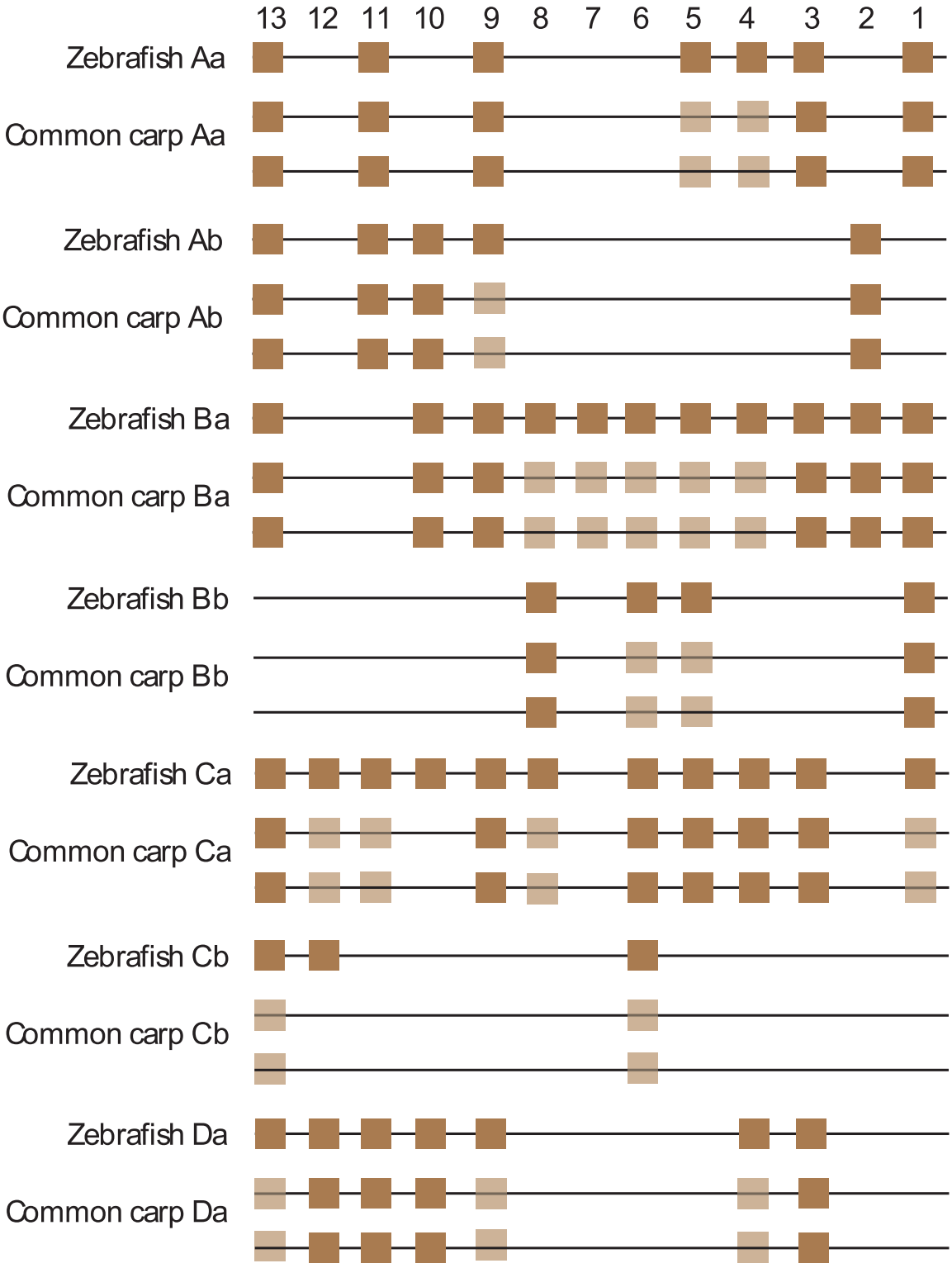

Figure 5.1 Comparison of the zebrafish and common carp *Hox* clusters. The position of the zebrafish *Hox* genes is based on the RefSeq gene annotation. Pseudogenes are not shown in this figure. Each horizontal thick line represents a cluster, and for each cluster, there are two duplicated paralogous clusters in carp. *HoxDb* cluster is not plotted as there is no *Hox* genes within it. Transparent gene represents one copy of *Hox* on either of the two clusters.

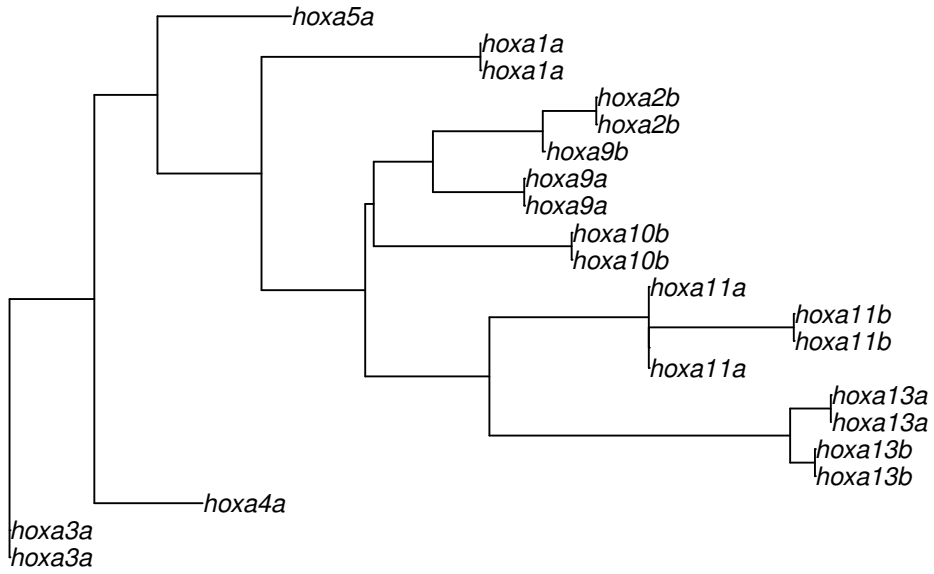

Figure 5.2 The phylogenetic trees reconstructed from genes in *HoxA* cluster. The leaves with same gene names represent the paralogs resulted from the fourth WGD. The last character 'a' or 'b' represents the paralogs produced in the third WGD

Next we studied the expression pattern of the duplicated genes. H. Spaink group sequenced the carp cDNA from 89 tissues and two embryos to generate a comprehensive resource for carp transcriptome. We retrieved the transcriptome data from them and wanted to investigate whether such ohnologs exhibit differences in overall expression pattern, which would be an indication of diverged function from the ancestor gene. 18241 ortholog groups were identified from InParanoid (Sonnhammer and Östlund, 2015) (*conf_cutoff = 0.1*) for the orthology mapping between common carp and zebrafish. Then we selected 3549 groups for 1:2 mapping between zebrafish and common carp. This group consists a majority of putative ohnologs resulting from the recent carp-specific duplication. Expression data from 15 different tissues on the ohnologous genes pairs were compared. In Figure 5.3, the number of differential expression ohnologs with an absolute fold change (FC) of 2 and 4 are shown in each tissue. As we can see, over 30% of the gene pairs show an absolute FC difference of more than 2 in all tissues analysed. Even with a FC cut off of 4 , many pairs were also differentially expressed in a majority of all organs: e.g. four gene pairs are even differentially expressed at FC 4 in all 15 tissues under consideration. Although the majority of the putative

Table 5.2 The over-represented GO biological process terms for the genes that rapidly return to single copy after the carp-specific WGD.

terms	p-values	counts
metabolic process	2.17E-04	97/878
. . . . .DNA metabolic process	3.51E-04	7/19
. . . . . .DNA replication	8.81E-04	17/95
. . . . . .DNA repair	6.65E-04	24/153
. . . . . .nucleotide-excision repair	7.06E-04	7/21
. . . . . .RNA processing	1.12E-03	17/97
. . . . . .tRNA aminoacylation for protein translation	5.5E-05	16/69
. . . . . .translation	1.62E-04	33/219
.methylation	5.04E-04	7/20
. .oxidation-reduction process	7.14E-12	124/857
protein folding	2.82E-04	26/162
flagellated sperm motility	4.56E-04	3/3

ohnologs were co-expressed in at least one tissue, these data suggest that many of them have undergone an rapid expression divergence either towards being less dominant or towards acquiring altered functions in various tissues.

5.2.4 Analysis of conserved regulatory elements in common carp

With the annotated repeat regions and gene structures, then we proceed to the regulatory elements study for zebrafish and common carp. Pairwise whole genome alignment between zebrafish and common carp was generated (see Methods and Data), we then applied *CNEr* package to identify the CNEs. Two more stringent thresholds were used in addition to the standard 70% to 100% identity over 30, 50bp thresholds. The number of CNEs detected between zebrafish and carp is strikingly high, compared with the other two sets of CNEs from zebrafish vs. tetraodon and human vs. dog, Table 5.3. The divergence time between zebrafish and carp is 128 Mya (Xu et al., 2014), while human and dog are known to be separated around 100 Mya. If we assume the fish genome evolve as fast as human and dog, slightly fewer CNEs are expected to be detected from zebrafish vs. carp, compared with human vs. dog. To find out this discrepancy, we looked further into the Axt alignment of zebrafish vs. common carp, and human vs. dog. Same conservation parameters were used during the pairwise whole genome alignment. The Axt alignments demonstrated a comparable ratio of total length of alignments over their genome size, but a slightly higher rate of matches

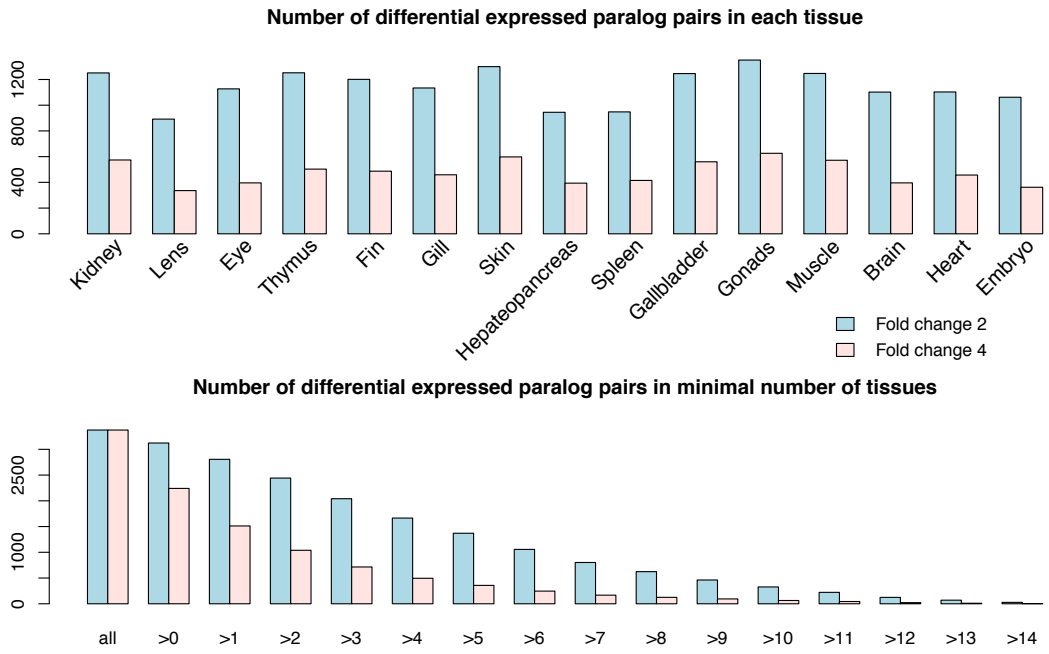

Figure 5.3 Numeric overview of differential expression levels of possible paralogous pairs. (A) Number of paralog pairs that are differentially expressed with a FC of larger than 2 in 15 separate tissues. (B) Number of possible paralog pairs filtered on the minimum number of tissues in which a differential expression of larger than FC 2 is observed. E.g. in the bottom of the scale 27 possible paralogous pairs have a FC of larger than 2 in all 15 tissue types listed in panel A. (C and D) The same analysis as in panels A and B respectively with a FC of 4 as cut-off value. With these criteria there are still 4 possible paralogous pairs that have a FC larger than 4 in all 15 tissues.

between zebrafish and common carp (Figure C.1). This implies that the fish genomes do not evolve any faster than the slowest-evolving mammalian genomes. And zebrafish and common carp is likely to be separated more recently than 128 Mya, as reported 87 Mya in (Zhao et al., 2016). It is also noteworthy that many of the zebrafish CNEs may retain two copies in common carp due to the recent WGD. For the set of 100% identity over 50bp, around 40% of zebrafish CNEs have two copies of paralogous CNEs in common carp. This property of common carp CNEs makes carp a even better candidate for detecting CNEs from duplicated regions, as described in Chapter 4. But the draft assembly we have is still too fragmented. One GRB may be chopped and scattered into several scaffolds, which makes the usage of synteny information impossible.

Table 5.3 The summary of the number of CNEs in three comparisons at the threshold 100% over 30, 50, 75 and 100 bp. NC, not counted with our usual CNE detection thresholds.

	zebrafish vs. carp	zebrafish vs. tetraodon	human vs. dog
98% over 50bp	218889	1749	190059
100% over 50bp	115627	715	96037
100% over 75bp	35699	NC	NC
100% over 100bp	14236	NC	NC

The CNE density visualisation is available on Ancora browser http://ancora.genereg.net/cgi-bin/gbrowse/danRer10. Due to the sufficient amount of CNEs, the zebrafish:carp CNEs track provides much more information than ever about the regulatory elements in zebrafish. For some key developmental genes, like *dachd* (Figure 5.4) and *sall1a*, the CNEs detected from the comparison to other vertebrates can encompass the entire gene loci and visualise the density track clearly. However, for genes like *fzd5* and *creb1b*, there are many CNEs preserved from common carp comparison and some from blind cave fish, but very few CNEs can be identified from other vertebrates, Figure 5.5. The availability of carp CNEs makes the study of CNE turnover pattern possible using zebrafish genome as reference.

5.3 Methods and data

5.3.1 Genome annotation

Genome annotation can usually be split into two phases: the first "computation" phase to identify the gene structures; the second "annotation" phase to integrate annotation data into the predicted genes (Yandell and Ence, 2012). The genome was annotated using the MAKER pipeline (v2.31.8) (Cantarel et al., 2008).

The first step in the "computation" phase is repeat masking. We applied the homology-based search with RepeatMasker (4.0.5) (http://www.repeatmasker.org) against all species from Repbase (20140131) (Jurka et al., 2005) and a *de novo* transposable element library (630 kb in 1114 repetitive sequences) for the carp genome from RepeatModeler (v1.0.8). Then we combined three approaches for gene prediction: *ab-initio* gene prediction, mRNA expression evidence-based prediction and protein homology-based prediction. Two *ab-initio* prediction programs, AUGUSTUS Keller et al. (2011) and SNAP (Korf, 2004), were used to predict the genes in the repeat-masked genome. A pre-trained zebrafish gene model for

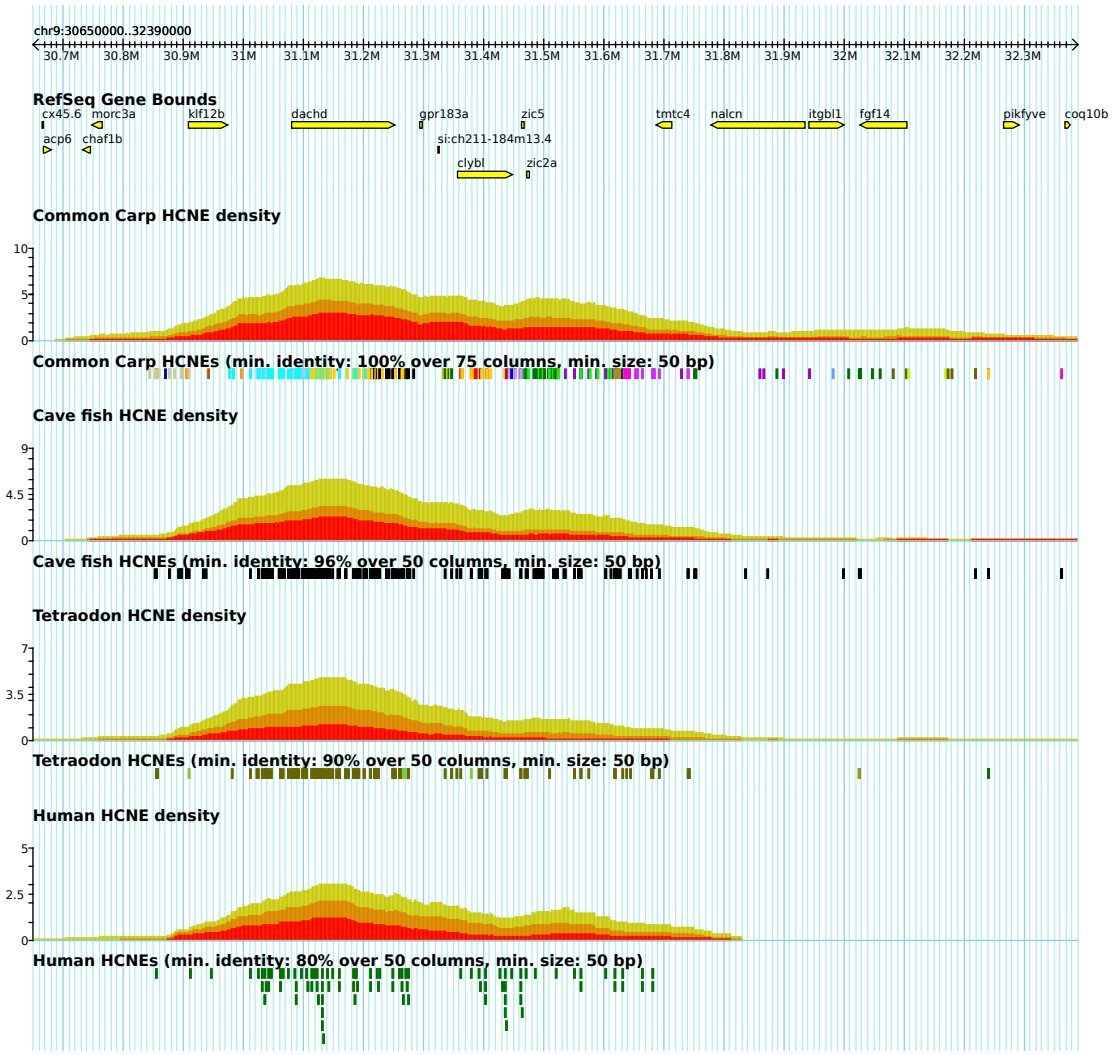

Figure 5.4 The CNE density plot around the developmental gene *dachd* of zebrafish, with the comparisons to common carp, blind cavefish, tetraodon and human. Enough CNEs are detected from each comparison to clearly visualise the CNE density trend.

AUGUSTUS was downloaded from the AUGUSTUS homepage. The gene model for SNAP was trained from a separate round of gene annotation purely on expression evidence and protein homology evidence. *Ab-initio* gene predictors produce preliminary gene models. Additional strong evidence is that (A) the region is transcribed (B) this region has homology to a known protein. With the assembled transcriptome, we aligned these sequences to the assembly with BLASTN (Camacho et al., 2009) to get the transcribed regions. Zebrafish protein sequence from Ensembl (version 77) was collected to build the database for the BLASTN step. The assembled genome sequences were aligned to the corresponding protein sequences in the database with BLASTX. Exonerate (Slater and Birney, 2005) was used to

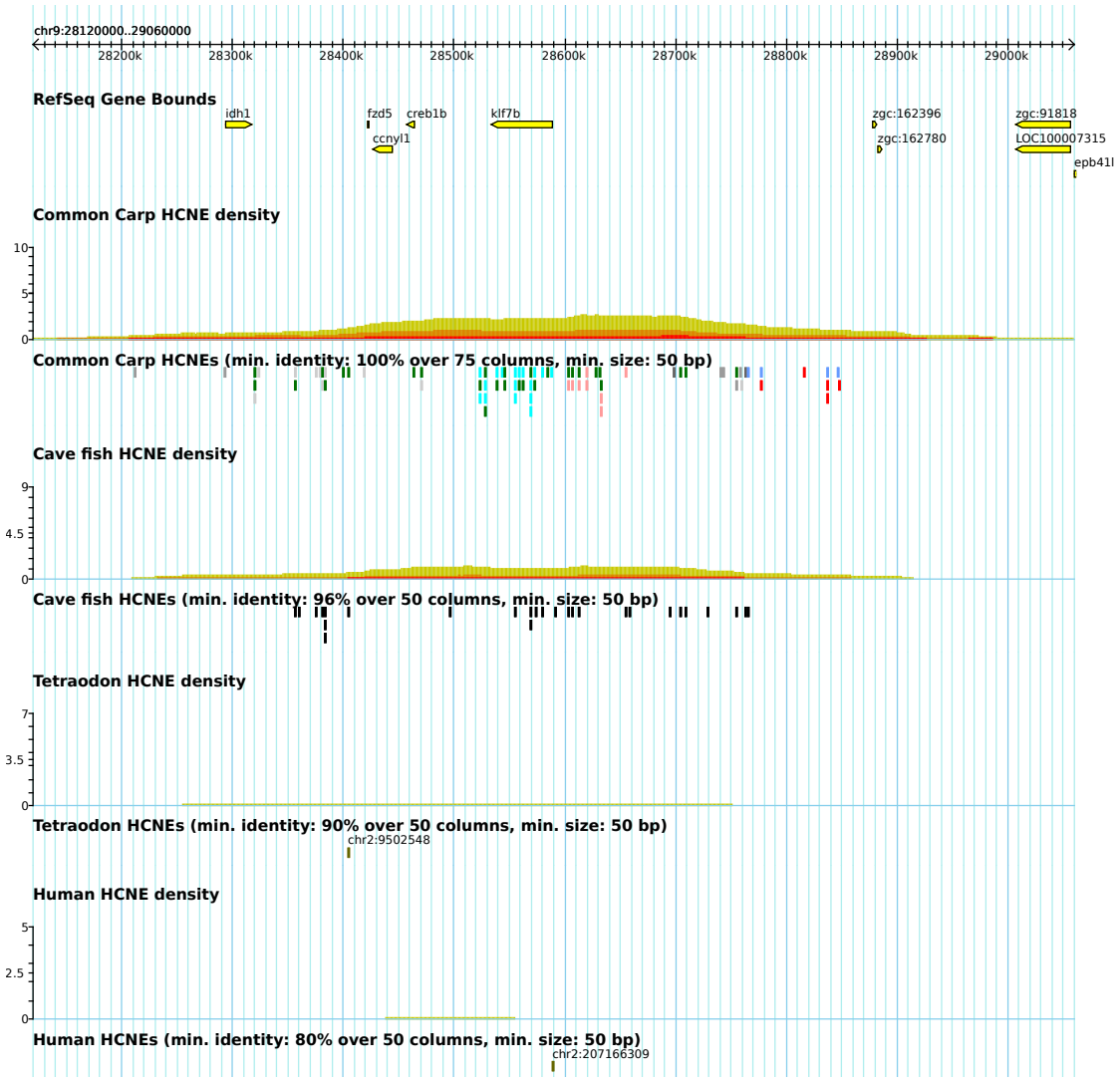

Figure 5.5 The CNE density plot around the developmental genes *fzd5* and *creb1b* of zebrafish, with the comparisons to common carp, blind cavefish, tetraodon and human. Enough CNEs detected from common carp and spotted gar clearly visualise the CNE density trend. But very few CNEs can be identified from other vertebrates.

postprocess the BLAST hits to make the alignments retain genomic order because by default BLAST will align multiple regions wherever it is possible. All these three sets of evidence were then integrated to provide the final gene model.

In the second "annotation" phase of obtaining the gene function annotations, we ran a BLASTP run (E-value: $1e-5$) of the protein sequences against UniProt/Swiss-Prot to get the putative gene functions. InterProScan (Jones et al., 2014) on predicted protein sequences was used to provide protein domain functions, putative GO terms annotation. Finally, the

whole genome annotation was summarised into a GFF file, a fasta file of transcriptome and a fasta file of protein sequences. Visualisation of the produced GFF file is available at http://jbrowse.genereg.net/index.html?data=data/mpirunAugustus/.

5.3.2 Whole genome alignment

The synteny analysis between our assembly and Songpu strain was done by the Last pipeline available from *CNEr*. The scoring matrix "near" was used due to close relationship between two assemblies.

The pairwise whole genome alignment between zebrafish and common carp was conducted with Last pipeline from *CNEr*. The scoring matrix "medium" was used.

5.3.3 Phylogenyetic analysis of *Hox* genes

To build the phylogenetic tree of *Hox* genes from one cluster, we first combined the CDS sequences for each gene. Then the DNA sequences are translated into AA sequences since AA sequences can yield better tree if the synonymous sites are saturated by multiple substitutions. MSAs were built with Mafft (v7.215) (Katoh and Standley, 2013) (*–retree 2 –maxiterate 0*). MSAs were then passed into PhyML (v3.0) (Guindon et al., 2010) to construct the phylogenetic trees.

5.4 Discussion

We provided the research community with a draft carp genome and genome annotation. One of the main aims was to assist the growing and increasingly genomics-oriented community of zebrafish researchers in predicting the functional elements from sequence comparison. Although the quality control of *Hox* gene clusters indicates the satisfactory genome annotation so far, a more assembled genome will certainly improve the genome annotation by providing stronger synteny information. We performed the CNE identification pipeline for zebrafish and carp genome. Many more and longer CNEs have been observed than between terrestrial vertebrates at comparable evolutionary distance. Importantly, carp genome acts an excellent candidate for predicting the GRB edges. In addition, the close distance between them provides information for studying GRB turnover pattern in zebrafish and may resolve the conflict

between variable turnover model (Harmston et al., 2013) and three waves of regulatory innovation models (Lowe et al., 2011).

Chapter 6

Genome-wide automated prediction of regulatory territories and target genes under complex long distance *cis*-regulation

Comparative genomics and high-throughput experimental methods like ChIP-seq have enabled efficient detection of regulatory elements in metazoan genomes. Nevertheless, the assignment of those elements to their target genes has remained a difficult task. Traditional assignment to the nearest gene, or a manual and semi-intuitive process is far from reliable, since regulatory regions can be located hundreds of kilobases away from their target genes, sometimes beyond neighbouring genes. We previously showed that arrays of conserved noncoding elements (CNEs) span the loci of developmental regulatory genes ("targets") and several other genes ("bystanders"), and define the edges of genomic regulatory blocks (GRBs) (Engström et al., 2007; Kikuta et al., 2007b). We found that the target genes that respond to input from distal regulatory elements in those regions have specific features that distinguish them from bystander genes in the locus and the genome (Akalin et al., 2009). In this study, we develop methodologies to solve two problems central to studying genomic regulatory blocks and their role in developmental long-range regulation: 1. A robust approach for the automated determination of GRB spans; 2. A machine learning based method for genome-wide detection of target genes in GRBs. The result is a comprehensive catalogue of nearly one thousand human genes likely to be regulated by long-range interactions and the

regions harbouring their corresponding *cis*-regulatory elements. The catalogue comprises a large number of genes involved in development, developmental transcriptional regulation and axon guidance. Furthermore, these genes are enriched for genes involved in complex diseases, including cancer and diabetes. The GRB spans and target genes identified in this study provide a rich resource for studying developmental regulation and disease-associated genomic variation.

6.1 Introduction

Although regulatory elements may be hundreds of kilobases away from their target genes, and sometimes much closer to other genes, specific cellular programs require that they regulate their target genes without affecting other genes in the region. We have previously shown that the necessity of retaining intact long-range spatial *cis*-relationships between enhancers and their target genes have constrained vertebrate gene order and genome organisation. As a result, those loci are the regions with the most ancient synteny conservation across vertebrate genomes (Engström et al., 2007; Kikuta et al., 2007b). These findings lead to the formulation of the genomic regulatory block (GRB) model, where one or more regulatory target genes are located in an evolutionary stable genomic domain rich in regulatory elements, along with "bystander" genes which do not respond to that regulatory input (Kikuta et al., 2007a,b; Navratilova et al., 2009).

The GRB target genes are a functionally narrow set with central importance in the regulation of multicellular developmental processes (Akalin et al., 2009; Engström et al., 2007; Kikuta et al., 2007b; Navratilova and Becker, 2009). We showed that they have specific features that distinguish them from both other genes in the neighbourhood and the rest of genome, notably long and often multiple CpG islands, large transcription initiation regions and a specific pattern of histone modifications (Akalin et al., 2009). However, until now the computational determination of specific target genes has been a manual, semi-intuitive and insufficiently well defined process, and the list of the known GRBs and their target genes has been far from complete. With the growing list of features that distinguish target genes from other genes in their neighbourhood and the rest of the genome, we can now begin to devise automated methods to catalogue a genome-wide set of all GRBs and their target genes. This is necessary as at present the experimental detection of target genes remains limited and of low-resolution. In this work, we present a computational method that uncovers an exhaustive

set of GRB target genes and the extent of the regulatory domains around them. Using the predicted GRB targets, we discover further compelling evidence for their complex regulation and elevated long-range regulatory potential.

6.2 Results

6.2.1 Automated genome-wide GRB boundaries identification

A GRB is defined as a cluster of syntenic CNEs around a key developmental target gene. The availability of well-defined GRB boundaries is a prerequisite for target gene prediction. Here we developed a CNE clustering pipeline for estimating the GRB boundaries given sets of CNEs (see Methods for details). The method uses CNE locations as their only input, so the goal of the algorithm is to identify the spans of high density of CNEs, effectively segmenting the genome into GRBs and regions outside GRBs. In addition to depending on the starting set of CNEs, the method should be optimised to minimise fragmentation of individual GRBs, to minimise the occurrence of merging of two or more adjacent GRBs, as well to balance between false positives and false negatives. For the particular purpose of target gene prediction, we wanted to generate an inclusive set of GRBs for human reference. For that reason we chose to use GRBs estimated from the CNEs in the comparisons of human and dog (hg38-canFam3, 100% identity over 50 bp), human and mouse (hg38-mm10, 98% identity over 50bp). In total, we predicted 832 GRBs, which covered a total of 1 Gbp genomic region and include 4595 protein-coding genes. Counter-intuitively, the number of predicted GRBs was somewhat lower than that predicted from human-chicken GRBs (847) in our most recent paper, where a different CNE clustering approach was used (Harmston et al., 2017). The *MEIS1* locus (Figure 6.1A) is one of the most prominent GRBs identified in the human genome. The *MEIS1* gene encodes a homeobox protein, which among many other roles acts as a transcriptional regulator of *PAX6* (another prominent GRB target gene) and is involved in vertebrate lens development (Zhang et al., 2002). The identified GRB span precisely reflects the ranges of CNEs arrays and encompasses the target gene *MEIS1* and other 6 predicted protein-coding bystander genes.

One major problem that can occur from this GRB boundaries identification approach is the merging of two adjacent GRBs. Since we solely rely on the CNE distribution and CNE density, it is hard to separate them unless we have additional information. *TOX3* and *SALL1* GRBs (Figure 6.1B) (Pennacchio et al., 2006; Royo et al., 2011) are known to be distinct

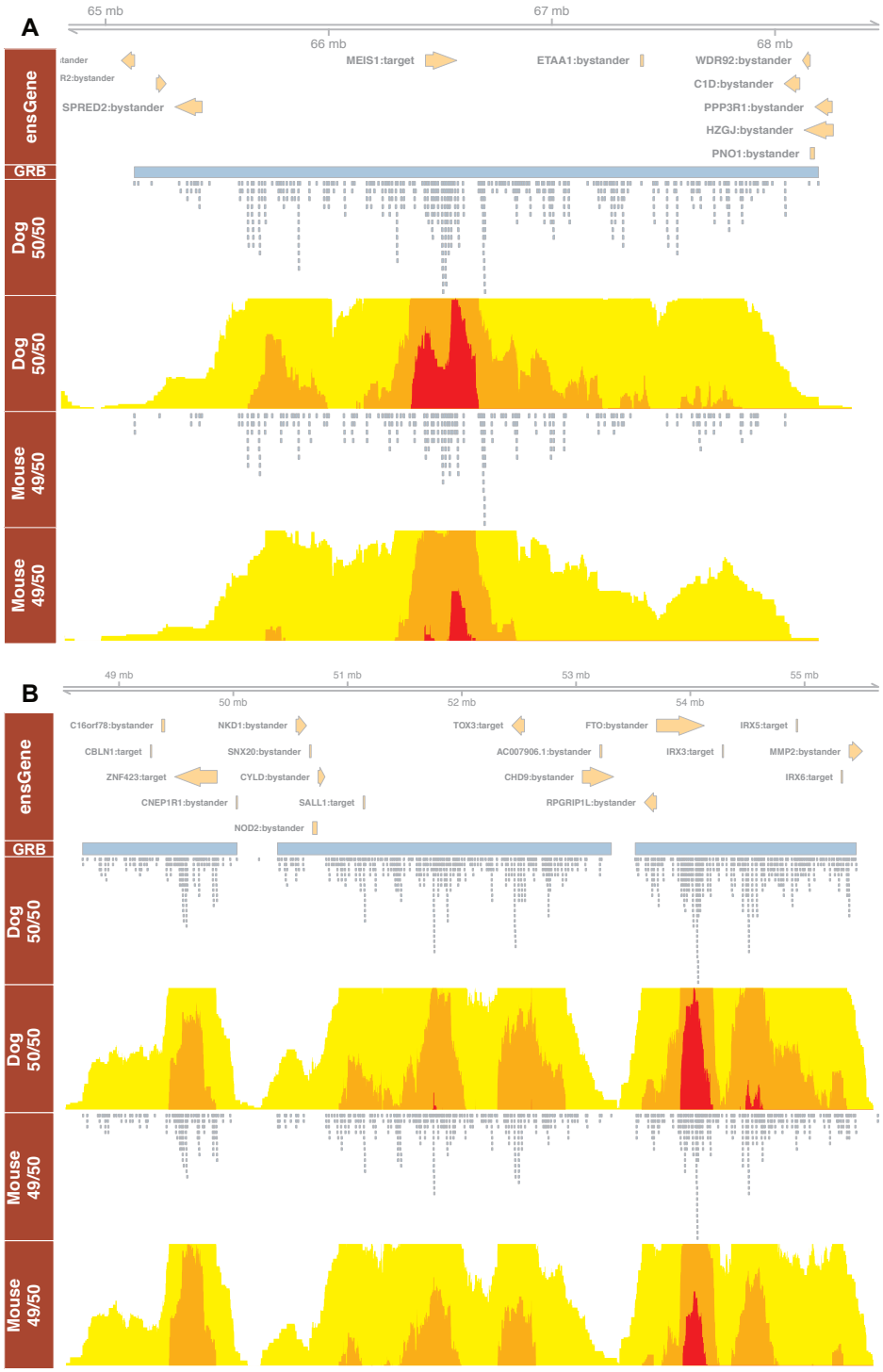

Figure 6.1 Examples of automated GRB edge identification from human-dog and human-mouse CNEs. The regions that have a higher than expected CNE density are considered as putative GRBs. (A) The well defined GRB region around the target gene *MEIS1*. (B) Three more complicated GRBs. The first GRB contains two target genes: *CBLN1* and *ZNF423*. The second GRB is actually a merged two adjacent GRBs: *SALL1* GRB and *TOX3* GRB. The third GRB contains three target genes: *IRX3*, *IRX5* and *IRX6*.

GRBs, as are the *PAX6* and *WT1* GRBs (Navratilova et al., 2009). However, it doesn't stand as a severe issue for the following target genes predictions. There are cases of multiple target genes within one GRB where the targets belong to an ancient cluster of paralogs, such as *IRX3/IRX5/IRX6* GRB in Figure 6.1B. Our target gene prediction method (see Methods) can deal with this scenario as well as the multi-target genes due to a merging GRB.

6.2.2 Accurate machine-learning based genome-wide prediction of target genes subject to long-range regulation

The method we developed to predict the regulatory target genes of GRBs genome-wide uses a random forest based machine learning approach. A random forest is an ensemble method that overcomes the overfitting problem of decision trees, while being fast, scalable and able to deal with outliers. A random forest consists of many classification trees, where the class of one input vector is determined by the classification having the most votes from all the trees in the forest (Liaw and Wiener, 2002). Our initial training dataset includes the manually annotated 259 target and 830 bystander genes taken from a previous publication (Akalin et al., 2009).

We carefully selected a set of genomic features, which were likely to be informative for target genes prediction (see Table 6.1 and Methods for more details). Some attributes were chosen based on insights reported in our group's previous publications (Akalin et al., 2009; Engström et al., 2008), like the length and number of CpG islands overlapping with the gene and CNE density across the gene body. We also introduced novel attributes like the entropy of gene's expression and gene divergence. Gene expression entropy is a measure of tissue-specificity of a gene, calculated from expression data vector representing each gene's expression in a panel of different tissue or cell types: higher entropy means a more ubiquitous expression, while lower entropy means more tissue-specific expression. Due to their multiple enhancers and expression in multiple but not all tissues, we expect most GRB target genes to have an intermediate entropy value (provided that the set of tissue/cell types represented in expression matrix is sufficiently large and diverse), while the non-developmental tissue-specific genes usually reside outside the GRBs. As a result, the gene with lowest entropy within a GRB is more likely to be the target gene. Considering gene divergence as an attribute is based on the assumption that, due to their essential role in development, target genes are under a different type of evolutionary selective pressure than bystander genes. Due to the

specific nature of random forest algorithm, even if some of the attributes are less relevant or redundant, they will not have a major impact on the classification.

Table 6.1 The attributes used in random forest model

Attributes used in Random Forest Classification	Reference
Total CpG island length overlapping with the gene	(Akalin et al., 2009)
Total CpG island length to the gene length ratio	(Akalin et al., 2009)
Number of CpG islands overlapping with the gene	(Akalin et al., 2009)
Number of CpG islands to the gene length ratio	(Akalin et al., 2009)
Human-dog 100% CNE density overlapping with the gene	(Engström et al., 2007; Kikuta et al., 2007a)
Human-mouse 98% CNE density overlapping with the gene	(Engström et al., 2007; Kikuta et al., 2007a)
Human-chicken 96% CNE density overlapping with the gene	(Engström et al., 2007; Kikuta et al., 2007a)
Human-spotted gar 80% CNE density overlapping with the gene	(Engström et al., 2007; Kikuta et al., 2007a)
Human-frog 80% CNE density overlapping with the gene	(Engström et al., 2007; Kikuta et al., 2007a)
Human-zebrafish 70% CNE density overlapping with the gene	(Engström et al., 2007; Kikuta et al., 2007a)
The %identity of human-dog homolog gene	
The %identity of human-mouse homolog gene	
The %identity of human-chicken homolog gene	
The %identity of human-spotted gar homolog gene	
The %identity of human-frog homolog gene	
The %identity of human-zebrafish homolog gene	
The gene expression entropy measurement from RNA-Seq experiment EMTAB513	
The gene expression entropy measurement from RNA-Seq experiment EMTAB1733	
The gene expression entropy measurement from Gtex project	

We built the random forest model based on the 19 attributes in Table 6.1. The random forest model assigns a vote of being either target or bystander gene for each gene. When a gene has higher target vote than bystander vote, this gene will be called as target gene, and vice versa. A total out-of-bag (oob) error rate of 8.18% is achieved, with 3% and 24% error rate for bystanders and targets, respectively. In particular, some of the previously confirmed target genes are classified as bystanders by the random forest model. A closer inspection of

these mis-classified genes revealed that most of them have the highest target vote in their respective GRB. A target gene can have a small, but highest, absolute target vote in the GRB, and the classifier can still declare this gene as a bystander in that GRB. The attributes of genes across GRBs are not directly comparable, as well as the vote values. Hence, a further processing of the vote values from the model is required. We normalised the absolute target vote score within each GRB (see Methods) to make it comparable across GRBs. Then we chose a cutoff score of 0.6 as optimal to maximise the accuracy rate for training set (Figure D.1). This final model predicted 98% of the training set accurately. This gave us 1161 potential targets and 3434 bystanders in total (Table D.1).

A useful feature of the random forest model is the availability of the measure of predictive importance of each of the attributes (Figure D.2). The importance of each attribute is evaluated with two measures "MeanDecreaseAccuracy" and "MeanDecreaseGini". The former is calculated from permuting oob data for each predictor variable and evaluating the prediction error. The latter is calculated as the Gini decrease for each variable over all trees in the forest. A higher value in both measures means a higher impact on the classifier. As we can see, the CpG islands length, the entropy of gene's expression, and CNE densities have the highest importance in this random forest model. These observations are consistent with the published observations regarding the properties of GRBs and target genes (Akalin et al., 2009; Kikuta et al., 2007b).

To test the robustness of our attributes used in the prediction, we employed an unsupervised random forest clustering. With the proximity matrix obtained from unsupervised learning, in Figure D.3, we observed the published targets clustered together with the top three dimensions, simply based on the information of the attributes. This confirms that the attributes are informative and that the supervised method avoids overfitting.

6.2.3 GRB target genes are involved in transcription/development and associated with complex diseases

Previous enhancer trapping experiments and manual curation of GRB targets (Ellingsen et al., 2005; Kikuta et al., 2007b) indicated that most developmental regulators are GRB targets. These genes seem to be recruited into the role of encoding the entire structural classes of proteins early on in Metazoan evolution. Thus, the predicted target genes are expected to be related to GO terms associated with development and transcriptional activity. As expected (Akalin et al., 2009; Engström et al., 2007; Kikuta et al., 2007a), the predicted target genes

are significantly enriched in biological process GO terms relating to development, regulation of transcription, axon development and cell fate commitment (Figure 6.2A). Meanwhile, no significant terms emerge from the predicted bystander genes. This further confirms that our model can distinguish target genes from bystander genes. We also performed KEGG pathway enrichment analysis for the predicted target genes (see Methods for details). The enriched pathways are mainly related to various types of cancers and diseases, such as "breast cancer" and "maturity onset diabetes of the young", as well as "axon guidance" and "signalling pathways" (Figure 6.2B). Following this, we also performed a disease ontology enrichment analysis for target genes (see Methods for details). This analysis also revealed enrichment in genes involved in various cancers and diseases such as "developmental disorder of mental health" and "acute leukemia" from our predicted targets (see Figure D.4). No enrichment is observed from genes classified as bystanders.

Following general functional annotation enrichment analysis, we studied the CNE turnover pattern of these target genes. We clustered target genes based on their maximal CNE densities across the whole gene body. The CNEs are derived from the same pairwise genome comparisons used in random forest model (human vs. dog, mouse, chicken, spotted gar, frog and zebrafish, see "Methods" for details). Three major clusters by the hierarchical clustering reveals the non-coding conservation depth of target genes (See Figure 6.3 for clusters and their GO Biological Process enrichment). The cluster with high densities of CNEs in all species (green in Figure 6.3) was highly enriched for genes involved in embryonic development, pattern specification and regionalization (See Figure 6.4 A for example gene *HMX2* and its CNE densities). Another cluster (red in Figure 6.3), showing a steeper decrease in CNE density across the species, is enriched for neural development, heart development and axon development terms. The last cluster (blue in Figure 6.3) has lower CNE density in dog and mouse, and very few CNEs observed in other species. This enrichment for this cluster is much less significant and is associated with dendritic spine development, signaling and neuronal action. *ISL1* and *RELN* genes are examples of predicted target genes from the last two clusters, respectively (Figure 6.4 B and C). *ISL1* encodes a transcription factor and is involved in cardiac cell fate determination (Bu et al., 2009) and axon regeneration. *RELN* plays important roles in brain development and schizophrenia (Fehér et al., 2015).

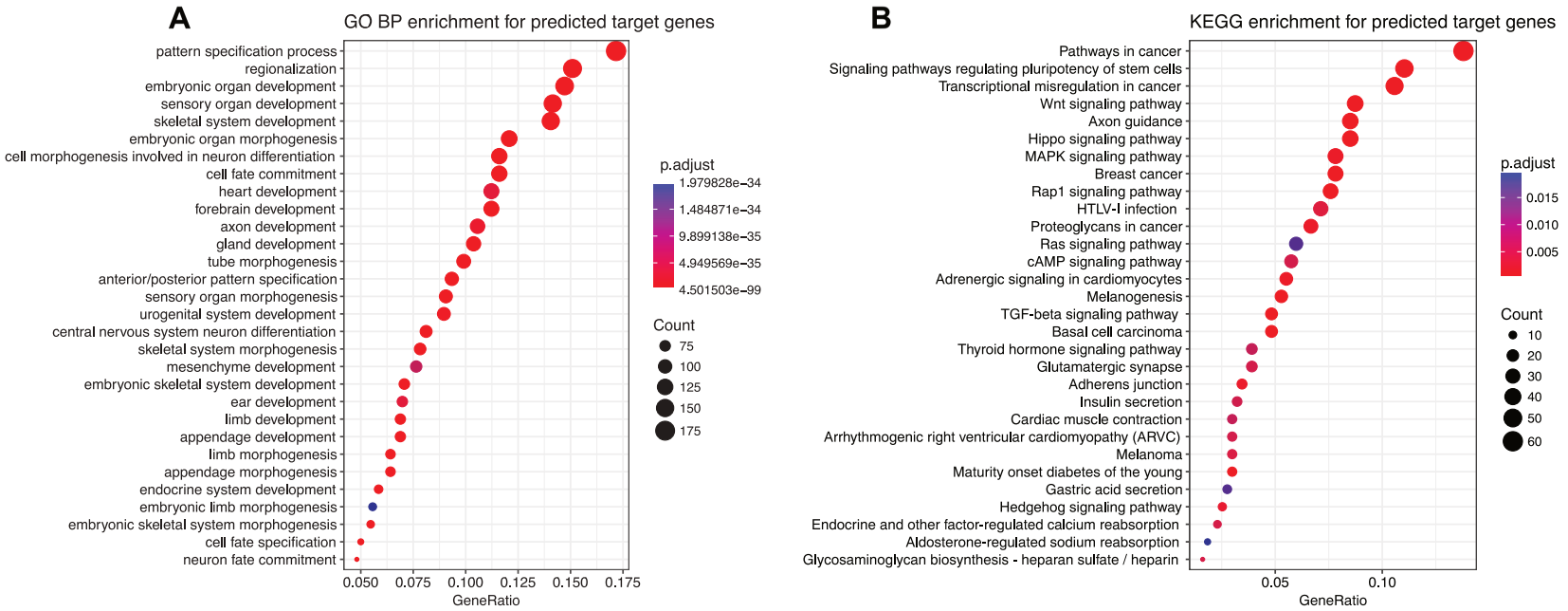

Figure 6.2 Over-represented GO Biological Process terms and KEGG pathways for predicted target genes, ranked by GeneRatio (The number of genes associated with the term in our selected genes divided by the number number of selected genes.) The p-values are adjusted by "BH" approach. The visualisation is done by clusterProfiler (Yu et al., 2012). (A) GO Biological Process terms enrichment. Target genes were significantly enriched in GO terms relating to organ, embryonic, nervous system development. No enrichment are observed from bystander genes. (B) KEGG pathways enrichment. Many terms are related to cancer and complex diseases, signaling pathways. No enrichment are observed from bystander genes.

Figure 6.3 Clustering and heatmap of predicted targets based on CNE densities over 6 species. Three major clusters of target genes are indicated by colours. The green cluster is deeply conserved and include genes involved in development and regionalization. Blue and red clusters exhibit a decreasing CNE densities over 6 species. They are involved with neuron development, axon development and signaling.

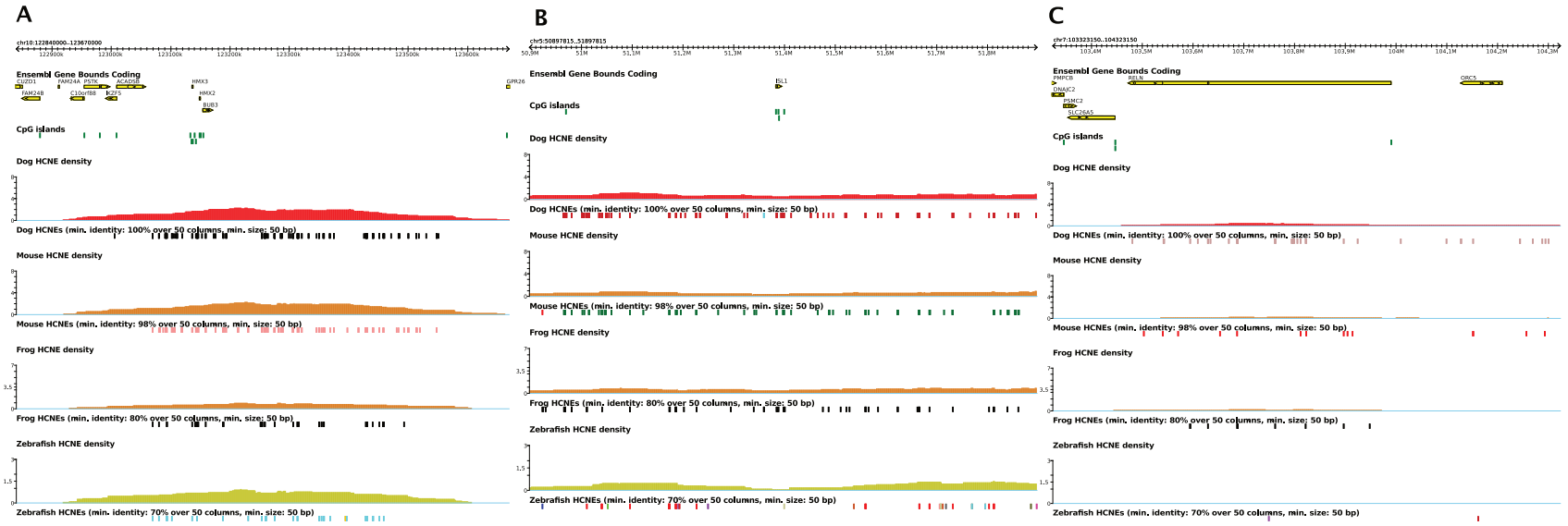

Figure 6.4 Examples of target genes with high and low CNE densities over the species. (A) Predicted target genes *HMX2* has high CNE densities over all four species, representing the green class in Figure 4. (B) and (C) *ISL1* and *RELN* have decreasing CNE density over four species, representing the red and blue clusters, respectively.

6.2.4 Substitution rates of targets and bystanders

The evolutionary rate can be inferred and distinguished based on substitution rates of synonymous (dS) and non-synonymous sites (dN). The substitutions on non-synonymous sites are under purifying selection, while the substitutions on synonymous sites are assumed not to be. Thus, the dS value is expected to reflect the neutral evolutionary rate of genes (although there are exceptions (Dong et al., 2010)). (Castillo-Davis and Hartl, 2002) showed that, in *C. elegans*, dN shows no significant difference between genes expressed before embryogenesis, genes after embryogenesis and nonmodulated genes, but there is significant difference for dS values. Among the early-expressed genes are many transcription factors, including an overwhelming majority of homeobox proteins and are highly likely to be target genes, while the nonmodulated genes are more likely to be bystander genes. If this pattern is applicable to human, we may detect a more informative signal from synonymous sites than non-synonymous sites, by separating the substitution rates of synonymous and non-synonymous sites. Surprisingly, in Figure 6.5, besides the smaller dS values for target genes than bystander genes ($p < 0.01$, Wilcoxon two-sided test), we also observed significantly smaller dN and dN/dS values for target genes ($p < 2.2e-16$, Wilcoxon two-sided test). In addition, all the genes within GRBs tend to have both smaller dN and dS values than genes outside the GRBs (Figure 6.5). All this evidence suggests a slower evolutionary rate for target genes than bystander genes, and as a whole set, they also evolve more slowly than genes outside the GRBs. It is a reasonable observation, considering GRB as a synteny lock-in of regulatory input, target genes and bystander genes. This lock-in can only be escaped via re-diploidization and subfunctionalization processes after WGD. A common scenario of a possible fate of a GRB after WGD is one copy of the bystander gene accumulating mutations and gradually getting decomposed and lost (Dong et al., 2009; Kikuta et al., 2007b). In this case, we should observe more duplications of target genes than bystanders. As shown in Figure D.5, the predicted target genes demonstrate significantly higher duplication levels than bystander genes and the other genes outside GRBs ($p < 0.01$, Wilcoxon test, two-sided). This further confirms our approach towards predicting target genes.

6.3 Discussion

We proposed a computational method for genome-wide identification of regulatory territories and the target genes under long-range *cis*-regulation. From the more comprehensive list of

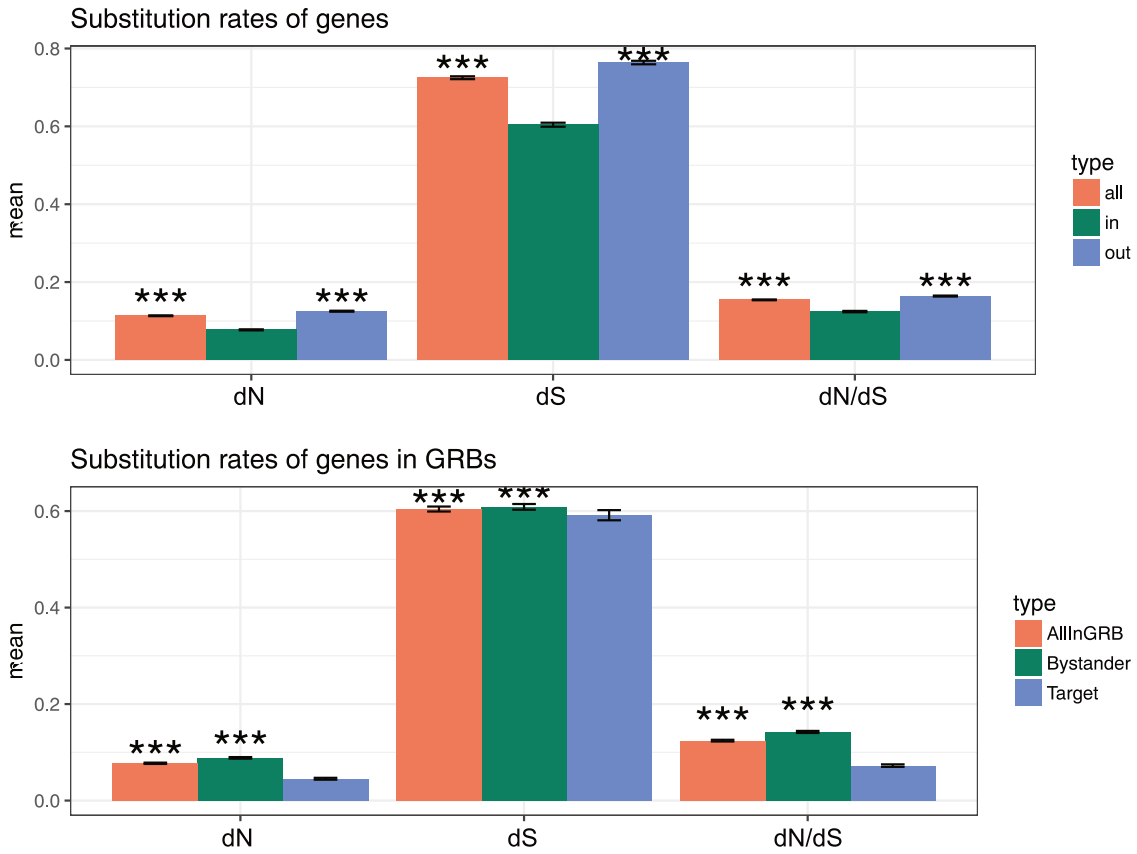

Figure 6.5 Comparison of substitution rates on synonymous (dS) and non-synonymous sites (dN) for different gene categories. (A) Compared with bystander genes and all the genes within GRBs, the dN and dN/dS values in target genes are significantly smaller ($p < 0.01$, Wilcoxon test, two-sided). (B) dN, dS and dN/dS values of the genes within GRBs are also significantly smaller than the genes outside GRBs. All this evidence suggests a slower evolutionary rate for target genes than bystander genes, and as a whole set, they also evolve more slowly than genes outside the GRBs.

target genes and bystander genes, we have obtained more insight into the properties of specific target genes. In addition to the previously proposed attributes of CpG islands and CNE density, we also find target genes are highly associated with gene entropy and gene divergence. The predicted target genes are over-represented in development and transcriptional regulation, while the predicted bystander genes have no enrichment for specific functions. KEGG pathway enrichment also indicates that target genes are also involved in cancer and other complex diseases. The GRB boundaries and predicted target genes are especially helpful for the interpretation of genome-wide association study (GWAS). Some disease-associated Single nucleotide polymorphisms (SNPs) from GWAS might be assigned to wrong genes,

for instance, bystanders from our prediction. SNPs, in the first intron of *FTO* gene, are associated with type 2 diabetes and obesity. The CNEs found in the first intron of *FTO* are more likely to target the developmental transcriptional factor *IRX3* gene (Ragvin et al., 2010; Smemo et al., 2014).

Investigation of noncoding conservation along the target genes reveals at least two types of target genes. The genes with most deeply conserved CNE densities down to fish are involved in organ development and transcriptional regulation. The genes with a decreasing CNE density along further distance play important roles in heart development and axon development. This is consistent with the findings of identifying a set of weakly conserved enhancers involved in heart development (Blow et al., 2010).

6.4 Methods

6.4.1 Detecting CNEs and estimating the edges of GRBs

The CNEs used in this study were identified using our Bioconductor package *CNEr*. For human-dog (hg38-canFam3) and human-mouse (hg38-mm10), we downloaded the pairwise alignments from UCSC Genome Browser (Kent et al., 2002). We identified the CNEs with thresholds of 100% and 98% over 50bp for dog and mouse, respectively. The CNE densities were calculated with a smoothing window of 300kb. The edges of GRBs were estimated from the CNE density. In brief, the genome is sliced into segments where the CNE density is above the expected CNE density. Within the segment, any CNEs violating syntenic order are discarded. Then some post-processing steps are applied to shrink the boundaries to the location of CNEs and discard the GRBs not encompassing any protein coding genes. A detailed vignette and implementation could be found in the *CNEr* package.

6.4.2 Feature extraction for random forests

We downloaded Ensembl gene boundaries, percentages of identity of homologous genes from Ensembl 79 with *biomaRt* (Durinck et al., 2005). We obtained CNE densities for human-dog (hg38-canFam3, 100% identity over 50bp), human-mouse (hg38-mm10, 98% identity over 50bp), human-chicken (hg38-galGal4, 96% identity over 50bp), human-spotted gar (hg38-LepOcu1, 80% identity over 50bp), human-frog (hg38-xenTro3, 80% identity over 50bp) and human-zebrafish (hg38-danRer7, 70% identity over 50bp) with 300kb smoothing

windows from *CNEr* package. We extracted the maximal values of CNE density overlapping with each gene as the CNE density of the gene. We downloaded CpG island locations from the UCSC table browser, and intersected them with the Ensembl gene boundaries. The total length, the number, and the normalised values by gene length as the CpG-related attributes for each gene. The gene entropy value is calculated as $H(x) = -\sum_i P(x_i) \log_2 P(x_i)$, where $P(x_i)$ is the proportion of expression value (FPKM) from tissue i. Higher entropy means more ubiquitous expression while lower entropy means more tissue-specific expression. Two RNA-Seq experiments of human tissues are available from Expression Atlas: E-MTAB-513 (Derrien et al., 2012) and E-MTAB-1733 (Fagerberg et al., 2014). The sample E-MTAB-513 is a RNA-Seq of human individual tissues and mixtures of 16 tissues from Illumina Body Map. The sample E-MTAB-1733 is a RNA-Seq of coding RNA from tissue samples of 95 human individuals representing 27 different tissues. The gene expression values of GTEx project (GTEx Consortium, 2013) were downloaded from UCSC. For many genes, one or more attributes are missing, which makes it unclassifiable with the machine learning method. To overcome this limit, we filled the missing value by the rough imputation of missing values by medians.

6.4.3 Target gene prediction using random forests

We used the R package *randomForest* (v4.6-12) (Liaw and Wiener, 2002) to train and predict the target genes. Random forests is an ensemble classifier that consists of many decision trees and predicts the class of an instance by collecting the vote from each decision tree. The input for each decision tree is a subset of the training dataset with replacement. At each node, the best attribute from the random subset of attributes is chosen. We constructed 500 trees for each forest. We also incorporated the information that most of the target genes had the highest score in their GRB, regardless of the score cutoff used, in our target detection pipeline as well. We normalised the scores of the genes in each GRB by subtracting the prediction score from the minimum prediction score in the GRB and dividing the result by the maximum score in GRB subtracted from minimum score in the GRB (see below for the equation, RF stands for "Random Forest").

$$RFScore_{normalised} = \frac{RFScore - min(RFScoresinGRB)}{max(RFScoresinGRB) - min(RFScoresinGRB)}$$

As a result, the gene with maximal RF score will have a normalised value of 1 and will always be selected as the target gene. Then we optimised a cut-off for this normalised score by maximising percentage of prediction accuracy for the training set.

6.4.4 Clustering of target genes based on CNE densities over species

We clustered predicted target genes based on their associated CNE densities generated from pairwise alignments of human to six different species. We used a distance based on the Manhattan distance between CNE densities in each species to produce the distance matrix between predicted targets. Following this, we used a hierarchical clustering with complete linkage method on the distance matrix. Three major clusters were produced from this analysis, and were displayed as heatmaps using the `heatmap.2` function from *gplots* package.

6.4.5 Functional annotation terms enrichment: GO, KEGG and DO

GO and KEGG term enrichment analysis for predicted target genes was done using the Bioconductor package *clusterProfiler* (Yu et al., 2012). The hypergeometric test was used to check for over-representation. We tested GO and KEGG term over-representation for all predicted target genes using the whole Ensembl protein coding gene set as a background, and considering a cutoff of adjusted P-values 0.05 with the "BH" method for false discovery rate (Benjamini and Hochberg, 1995). Concerning disease ontology enrichment analysis, we utilised the DOSE package (Yu et al., 2015) with a FDR cutoff of 0.05. When we tested for GO term over-representation of target gene clusters based on their CNE density, we used all the genes within GRBs as background.

Chapter 7

Conclusions and discussion

Despite CNEs having been discovered for more than a decade, their function is still not fully clear. The underlying mechanism of maintaining extreme conservation remains a mystery. This thesis focused on development of computational tools and approaches for studying non-coding conservation in metazoan genomes. The tools and resources presented here could shed light on understanding non-coding conservation.

CNEr is a high performance package for CNE detection and visualisation. It provides a comprehensive pipeline from whole genome pairwise alignment of any two species to detection and visualisation of CNEs. This packages also fills the gap of missing infrastructure for easy manipulation of comparative genomics alignment data. The comparative analysis of the *Glossina* and sea urchin genomes demonstrate the usage of *CNEr*. *TFBSTools* bridges JASPAR database and TFBS analysis on a genome-wide scale. Ever since the initial release on Bioconductor, it becomes extremely popular.

In Chapter 4, we developed a specific CNE detection pipeline from duplicated regions in a ameiotic *A. vaga* genome. Although CNEs from this pipeline tend to be noisy and incomplete, we gained first and clear clues of noncoding conservation in a genome without chromosome pairing. This rules out the possibility of CNE's role in monitoring copy number of genome. The identified CNEs in such ancient genome imply the function in regulation of development.

Chapter 5 described *de novo* assembly and gene annotation of European common carp *Cyprinus carpio*. This new assembly should become the primary choice for comparative genomics study of zebrafish. With the massive amount of new discovered zebrafish CNEs, further experimental verification of enhancer potential can be designed. Due to the extra

WGD in carp lineage, carp is also perfect for studying the fate of GRB evolution after a recent WGD.

The methods for GRB boundary and target genes prediction presented in Chapter 6 are especially useful as tools to support any other studies of regulatory territories. Although we focused on human genome in this chapter, this GRB boundary prediction can be easily extended to other species. For instance, we predicted the GRBs boundaries in zebrafish and amphioxus to compare the number of enhancers within GRBs (data not shown in this thesis). The predicted target genes were used in other projects in our group, such as targets genes associated with SNPs.

7.1 Future directions

While we have detected CNEs and shown their functions in several species, answers to many questions remain unclear. CNEs and GRBs have been shown to be prevalent in many lineages, but how ancient is the origin of CNEs and the associated transcriptional regulation? CNEs were first identified in vertebrates, then later in insects (Engström et al., 2007), worms (Vavouri et al., 2007). We also identified CNEs and GRBs in early diverging metazoan phyla amphioxus (data not shown in this thesis) and *A. vaga* (Chapter 4). Our former group member, Slavica Dimitrieva, showed preliminary results of the equivalent noncoding conservation in *Aspergillus* and *Candida*. Even in *Dictyostelium* genus, we managed to detect thousands of CNEs and they cluster around genes that are involved in sorocarp development (data not shown in this thesis). If these are equivalent to Metazoan CNEs, it suggests the existence of such elements even in an organism with a special life cycle of conversion between unicellular and multicellular phase. Together with the evidence of CNEs in plants (Burgess and Freeling, 2014; Haudry et al., 2013), the origin of CNEs may date back to the very root of tree of life. Studying how CNEs regulate the genes across different lineages will provide a full picture of CNE's function. And *CNEr* was built to facilitate this kind of study.

The source of the extreme noncoding conservation still puzzles us. We ruled out the explanation of copy number sensing during allele recombination in Chapter 4 by detecting noncoding conservation in *A. vaga*. However, the sets of identified CNEs, especially allelic CNEs, are found to be noisy, compared to the CNEs generated from a standard pairwise comparison. It is desired to look into further the genomic distribution difference of allelic and ohnologous CNEs, and the associated distinct gene functions. Certainly it heavily hinges

on the quality of the genome assembly and gene annotation. The *A. vaga* genome assembly is very fragmented, reducing its power for comparative genomics analysis. Short scaffolds fragment the GRB and make studying GRBs evolution more difficult. Even though we managed to annotate the gene function and protein structure, a more thorough and curated gene annotation is in demand for this genome. Common carp has the same problem of fragmented assembly. The planned enhanced assembly with Nanopore technology will make possible the study of GRB evolution after a recent WGD.

In Chapter 6, we developed an automated approach to determine GRB boundaries and predict the target genes. While the GRB boundaries estimation approach can be easily extended to other species, prediction of target genes is still limited to human. This is due to lack of available attributes of random forests model in other species, such as gene expression entropy from multiple tissues and CpG information. The current workaround is to use the ortholog mapping from human target genes. A dedicated solution is needed for other species. To further verify the sets of predicted target genes, chromosome conformation capture technologies, such as Hi-C, can be useful by revealing more chromatin interactions between target genes and CNEs.

In Chapter 6, we also gained first insights into the noncoding conservation changes for the target genes. However, a more systematic investigation of CNE turnover pattern within GRB is desired. A very common phenomenon is that the number of detectable CNEs between two species decreases significantly with increasing evolutionary distance. Although mutation rate variation has already been shown in vertebrate genomes (Ellegren et al., 2003), this still cannot explain the different pattern and rate of divergence along a GRB and across different GRBs (Kim and Pritchard, 2007; Lee et al., 2011). A CNE turnover model was proposed in (Harmston et al., 2013) that *cis*-regulatory elements in the common ancestor of two lineages continue to be lost during evolution while new elements are recruited. Still, the mechanisms governing the mutation accumulation and turnover pattern of CNEs are unknown. It will be worth investigating the distribution of CNEs within GRBs between species at various evolutionary distances to learn more about the patterns of divergence and turnover. We expect that the study of genes and CNEs in high turnover regions and low turnover regions will identify distinct functions of genes regulated by GRBs with different turnover rates and explain the associated evolutionary patterns.

References

Adams, D., Altucci, L., Antonarakis, S. E., Ballesteros, J., Beck, S., Bird, A., Bock, C., Boehm, B., Campo, E., Caricasole, A., Dahl, F., Dermitzakis, E. T., Enver, T., Esteller, M., Estivill, X., Ferguson-Smith, A., Fitzgibbon, J., Flicek, P., Giehl, C., Graf, T., Grosveld, F., Guigo, R., Gut, I., Helin, K., Jarvius, J., Küppers, R., Lehrach, H., Lengauer, T., Lernmark, A., Leslie, D., Loeffler, M., Macintyre, E., Mai, A., Martens, J. H. A., Minucci, S., Ouwehand, W. H., Pelicci, P. G., Pendeville, H., Porse, B., Rakyan, V., Reik, W., Schrappe, M., Schübeler, D., Seifert, M., Siebert, R., Simmons, D., Soranzo, N., Spicuglia, S., Stratton, M., Stunnenberg, H. G., Tanay, A., Torrents, D., Valencia, A., Vellenga, E., Vingron, M., Walter, J., and Willcocks, S. (2012). BLUEPRINT to decode the epigenetic signature written in blood. *Nat. Biotechnol.*, 30(3):224–226.

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

Amores, A., Force, A., Yan, Y. L., Joly, L., Amemiya, C., Fritz, A., Ho, R. K., Langeland, J., Prince, V., Wang, Y. L., Westerfield, M., Ekker, M., and Postlethwait, J. H. (1998). Zebrafish hox clusters and vertebrate genome evolution. *Science*, 282(5394):1711–1714.

Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., Ren, J., Li, W. W., and Noble, W. S. (2009). MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.*, 37(Web Server issue):W202–208.

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

Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300.

Bhatia, S., Monahan, J., Ravi, V., Gautier, P., Murdoch, E., Brenner, S., van Heyningen, V., Venkatesh, B., and Kleinjan, D. A. (2014). A survey of ancient conserved non-coding elements in the PAX6 locus reveals a landscape of interdigitated cis-regulatory archipelagos. *Dev. Biol.*, 387(2):214–228.

Blow, M. J., McCulley, D. J., Li, Z., Zhang, T., Akiyama, J. A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F., Afzal, V., Bristow, J., Ren, B., Black, B. L., Rubin, E. M., Visel, A., and Pennacchio, L. A. (2010). ChIP-Seq identification of weakly conserved heart enhancers. *Nat. Genet.*, 42(9):806–810.

Brown, J. R. and Doolittle, W. F. (1995). Root of the universal tree of life based on ancient aminoacyl-tRNA synthetase gene duplications. *Proc. Natl. Acad. Sci. U.S.A.*, 92(7):2441–2445.

Bu, L., Jiang, X., Martin-Puig, S., Caron, L., Zhu, S., Shao, Y., Roberts, D. J., Huang, P. L., Domian, I. J., and Chien, K. R. (2009). Human ISL1 heart progenitors generate diverse multipotent cardiovascular cell lineages. *Nature*, 460(7251):113–117.

Burgess, D. and Freeling, M. (2014). The Most Deeply Conserved Noncoding Sequences in Plants Serve Similar Functions to Those in Vertebrates Despite Large Differences in Evolutionary Rates. *Plant Cell*.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T. L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics*, 10:421.

Cameron, R. A., Samanta, M., Yuan, A., He, D., and Davidson, E. (2009). SpBase: the sea urchin genome database and web site. *Nucl. Acids Res.*, 37(suppl 1):D750–D754.

Cantarel, B. L., Korf, I., Robb, S. M. C., Parra, G., Ross, E., Moore, B., Holt, C., Sánchez Alvarado, A., and Yandell, M. (2008). MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.*, 18(1):188–196.

Castillo-Davis, C. I. and Hartl, D. L. (2002). Genome evolution and developmental constraint in Caenorhabditis elegans. *Mol. Biol. Evol.*, 19(5):728–735.

Celniker, S. E., Dillon, L. A. L., Gerstein, M. B., Gunsalus, K. C., Henikoff, S., Karpen, G. H., Kellis, M., Lai, E. C., Lieb, J. D., MacAlpine, D. M., Micklem, G., Piano, F., Snyder, M., Stein, L., White, K. P., Waterston, R. H., and modENCODE Consortium (2009). Unlocking the secrets of the genome. *Nature*, 459(7249):927–930.

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

de la Calle-Mustienes, E., Feijóo, C. G., Manzanares, M., Tena, J. J., Rodríguez-Seguel, E., Letizia, A., Allende, M. L., and Gómez-Skarmeta, J. L. (2005). A functional survey of the enhancer activity of conserved non-coding sequences from vertebrate Iroquois cluster gene deserts. *Genome Res.*, 15(8):1061–1072.

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D. G., Lagarde, J., Veeravalli, L., Ruan, X., Ruan, Y., Lassmann, T., Carninci, P., Brown, J. B., Lipovich, L., Gonzalez, J. M., Thomas, M., Davis, C. A., Shiekhattar, R., Gingeras, T. R., Hubbard, T. J., Notredame, C., Harrow, J., and Guigó, R. (2012). The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res.*, 22(9):1775–1789.

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

Diao, Y., Li, B., Meng, Z., Jung, I., Lee, A. Y., Dixon, J., Maliskova, L., Guan, K.-L., Shen, Y., and Ren, B. (2016). A new class of temporarily phenotypic enhancers identified by CRISPR/Cas9-mediated genetic screening. *Genome Res.*, 26(3):397–405.

Dimitrieva, S. and Bucher, P. (2012). Genomic context analysis reveals dense interaction network between vertebrate ultraconserved non-coding elements. *Bioinformatics*, 28(18):i395–i401.

Dimitrieva, S. and Bucher, P. (2013). UCNEbase–a database of ultraconserved non-coding elements and genomic regulatory blocks. *Nucleic Acids Res.*, 41(Database issue):D101–109.

Dong, X., Fredman, D., and Lenhard, B. (2009). Synorth: exploring the evolution of synteny and long-range regulatory interactions in vertebrate genomes. *Genome Biol.*, 10(8):R86.

Dong, X., Navratilova, P., Fredman, D., Drivenes, O., Becker, T. S., and Lenhard, B. (2010). Exonic remnants of whole-genome duplication reveal cis-regulatory function of coding exons. *Nucleic Acids Res.*, 38(4):1071–1085.

Dousse, A., Junier, T., and Zdobnov, E. M. (2016). CEGA–a catalog of conserved elements from genomic alignments. *Nucleic Acids Res.*, 44(D1):D96–100.

Drake, J. A., Bird, C., Nemesh, J., Thomas, D. J., Newton-Cheh, C., Reymond, A., Excoffier, L., Attar, H., Antonarakis, S. E., Dermitzakis, E. T., and Hirschhorn, J. N. (2006). Conserved noncoding sequences are selectively constrained and not mutation cold spots. *Nat Genet*, 38(2):223–227.

Duret, L. and Bucher, P. (1997). Searching for regulatory elements in human noncoding sequences. *Curr. Opin. Struct. Biol.*, 7(3):399–406.

Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., and Huber, W. (2005). BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, 21(16):3439–3440.

Ellegren, H., Smith, N. G. C., and Webster, M. T. (2003). Mutation rate variation in the mammalian genome. *Curr. Opin. Genet. Dev.*, 13(6):562–568.

Ellingsen, S., Laplante, M. A., König, M., Kikuta, H., Furmanek, T., Hoivik, E. A., and Becker, T. S. (2005). Large-scale enhancer detection in the zebrafish genome. *Development*, 132(17):3799–3811.

ENCODE Project Consortium (2004). The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*, 306(5696):636–640.

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74.

ENCODE Project Consortium, Birney, E., Stamatoyannopoulos, J. A., Dutta, A., Guigó, R., Gingeras, T. R., Margulies, E. H., Weng, Z., Snyder, M., Dermitzakis, E. T., Thurman, R. E., Kuehn, M. S., Taylor, C. M., Neph, S., Koch, C. M., Asthana, S., Malhotra, A., Adzhubei, I., Greenbaum, J. A., Andrews, R. M., Flicek, P., Boyle, P. J., Cao, H., Carter, N. P., Clelland, G. K., Davis, S., Day, N., Dhami, P., Dillon, S. C., Dorschner, M. O., Fiegler, H., Giresi, P. G., Goldy, J., Hawrylycz, M., Haydock, A., Humbert, R., James, K. D., Johnson, B. E., Johnson, E. M., Frum, T. T., Rosenzweig, E. R., Karnani, N., Lee, K., Lefebvre, G. C., Navas, P. A., Neri, F., Parker, S. C. J., Sabo, P. J., Sandstrom, R., Shafer, A., Vetrie, D., Weaver, M., Wilcox, S., Yu, M., Collins, F. S., Dekker, J., Lieb, J. D., Tullius, T. D., Crawford, G. E., Sunyaev, S., Noble, W. S., Dunham, I., Denoeud, F., Reymond, A., Kapranov, P., Rozowsky, J., Zheng, D., Castelo, R., Frankish, A., Harrow, J., Ghosh, S., Sandelin, A., Hofacker, I. L., Baertsch, R., Keefe, D., Dike, S., Cheng, J., Hirsch, H. A., Sekinger, E. A., Lagarde, J., Abril, J. F., Shahab, A., Flamm, C., Fried, C., Hackermüller, J., Hertel, J., Lindemeyer, M., Missal, K., Tanzer, A., Washietl, S., Korbel, J., Emanuelsson, O., Pedersen, J. S., Holroyd, N., Taylor, R., Swarbreck, D., Matthews, N., Dickson, M. C., Thomas, D. J., Weirauch, M. T., Gilbert, J., Drenkow, J., Bell, I., Zhao, X., Srinivasan, K. G., Sung, W.-K., Ooi, H. S., Chiu, K. P., Foissac, S., Alioto, T., Brent, M., Pachter, L., Tress, M. L., Valencia, A., Choo, S. W., Choo, C. Y., Ucla, C., Manzano, C., Wyss, C., Cheung, E., Clark, T. G., Brown, J. B., Ganesh, M., Patel, S., Tammana, H., Chrast, J., Henrichsen, C. N., Kai, C., Kawai, J., Nagalakshmi, U., Wu, J., Lian, Z., Lian, J., Newburger, P., Zhang, X., Bickel, P., Mattick, J. S., Carninci, P., Hayashizaki, Y., Weissman, S., Hubbard, T., Myers, R. M., Rogers, J., Stadler, P. F., Lowe, T. M., Wei, C.-L., Ruan, Y., Struhl, K., Gerstein, M., Antonarakis, S. E., Fu, Y., Green, E. D., Karaöz, U., Siepel, A., Taylor, J., Liefer, L. A., Wetterstrand, K. A., Good, P. J., Feingold, E. A., Guyer, M. S., Cooper, G. M., Asimenos, G., Dewey, C. N., Hou, M., Nikolaev, S., Montoya-Burgos, J. I., Löytynoja, A., Whelan, S., Pardi, F., Massingham, T., Huang, H., Zhang, N. R., Holmes, I., Mullikin, J. C., Ureta-Vidal, A., Paten, B., Seringhaus, M., Church, D., Rosenbloom, K., Kent, W. J., Stone, E. A., NISC Comparative Sequencing Program, Baylor College of Medicine Human Genome Sequencing Center, Washington University Genome Sequencing Center, Broad Institute, Children's Hospital Oakland Research Institute, Batzoglou, S., Goldman, N., Hardison, R. C., Haussler, D., Miller, W., Sidow, A., Trinklein, N. D., Zhang, Z. D., Barrera, L., Stuart, R., King, D. C., Ameur, A., Enroth, S., Bieda, M. C., Kim, J., Bhinge, A. A., Jiang, N., Liu, J., Yao, F., Vega, V. B., Lee, C. W. H., Ng, P., Shahab, A., Yang, A., Moqtaderi, Z., Zhu, Z., Xu, X., Squazzo, S., Oberley, M. J., Inman, D., Singer, M. A., Richmond, T. A., Munn, K. J., Rada-Iglesias, A., Wallerman, O., Komorowski, J., Fowler, J. C., Couttet, P., Bruce, A. W., Dovey, O. M., Ellis, P. D., Langford, C. F., Nix, D. A., Euskirchen, G., Hartman, S., Urban, A. E., Kraus, P., Van Calcar, S., Heintzman, N., Kim, T. H., Wang, K., Qu, C., Hon, G., Luna, R., Glass, C. K., Rosenfeld, M. G., Aldred, S. F., Cooper, S. J., Halees, A., Lin, J. M., Shulha, H. P., Zhang, X., Xu, M., Haidar, J. N. S., Yu, Y., Ruan, Y., Iyer, V. R., Green, R. D., Wadelius, C., Farnham, P. J., Ren, B., Harte, R. A., Hinrichs, A. S., Trumbower, H., Clawson, H., Hillman-Jackson, J., Zweig, A. S., Smith, K., Thakkapallayil, A., Barber, G., Kuhn, R. M., Karolchik, D., Armengol, L., Bird, C. P., de Bakker, P. I. W., Kern, A. D., Lopez-Bigas, N., Martin, J. D., Stranger, B. E., Woodroffe, A., Davydov, E., Dimas, A., Eyras, E., Hallgrímsdóttir, I. B., Huppert, J., Zody, M. C., Abecasis, G. R., Estivill, X., Bouffard, G. G., Guan, X., Hansen, N. F., Idol, J. R., Maduro, V. V. B., Maskeri, B., McDowell, J. C., Park, M., Thomas, P. J., Young, A. C., Blakesley, R. W., Muzny, D. M., Sodergren, E., Wheeler, D. A., Worley, K. C., Jiang, H., Weinstock, G. M., Gibbs, R. A., Graves, T., Fulton, R., Mardis, E. R., Wilson, R. K., Clamp, M., Cuff, J., Gnerre, S., Jaffe, D. B., Chang, J. L., Lindblad-Toh, K., Lander, E. S., Koriabine, M., Nefedov, M., Osoegawa, K., Yoshinaga, Y., Zhu, B., and de Jong, P. J. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447(7146):799–816.

Engström, P. G., Fredman, D., and Lenhard, B. (2008). Ancora: a web resource for exploring highly conserved noncoding elements and their association with developmental regulatory genes. *Genome Biol.*, 9(2):R34.

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

Fagerberg, L., Hallström, B. M., Oksvold, P., Kampf, C., Djureinovic, D., Odeberg, J., Habuka, M., Tahmasebpoor, S., Danielsson, A., Edlund, K., Asplund, A., Sjöstedt, E., Lundberg, E., Szigyarto, C. A.-K., Skogs, M., Takanen, J. O., Berling, H., Tegel, H., Mulder, J., Nilsson, P., Schwenk, J. M., Lindskog, C., Danielsson, F., Mardinoglu, A., Sivertsson, A., von Feilitzen, K., Forsberg, M., Zwahlen, M., Olsson, I., Navani, S., Huss, M., Nielsen, J., Ponten, F., and Uhlén, M. (2014). Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Mol. Cell Proteomics*, 13(2):397–406.

Fehér, A., Juhász, A., Pákáski, M., Kálmán, J., and Janka, Z. (2015). Genetic analysis of the RELN gene: Gender specific association with Alzheimer's disease. *Psychiatry Res*, 230(2):716–718.

Few, S. (2008). Time on the Horizon. *Visual Business Intelligence Newsletter*.

Flot, J.-F., Hespeels, B., Li, X., Noel, B., Arkhipova, I., Danchin, E. G. J., Hejnol, A., Henrissat, B., Koszul, R., Aury, J.-M., Barbe, V., Barthélémy, R.-M., Bast, J., Bazykin, G. A., Chabrol, O., Couloux, A., Da Rocha, M., Da Silva, C., Gladyshev, E., Gouret, P., Hallatschek, O., Hecox-Lea, B., Labadie, K., Lejeune, B., Piskurek, O., Poulain, J., Rodriguez, F., Ryan, J. F., Vakhrusheva, O. A., Wajnberg, E., Wirth, B., Yushenova, I., Kellis, M., Kondrashov, A. S., Mark Welch, D. B., Pontarotti, P., Weissenbach, J., Wincker, P., Jaillon, O., and Van Doninck, K. (2013). Genomic evidence for ameiotic evolution in the bdelloid rotifer Adineta vaga. *Nature*, 500(7463):453–457.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A. J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J. Y. H., and Zhang, J. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, 5(10):R80.

GTEx Consortium (2013). The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.*, 45(6):580–585.

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.*, 59(3):307–321.

Hahne, F. and Ivanek, R. (2016). Visualizing Genomic Data Using Gviz and Bioconductor. *Methods Mol. Biol.*, 1418:335–351.

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

Harmston, N., Ing-Simmons, E., Tan, G., Perry, M., Merkenschlager, M., and Lenhard, B. (2017). Topologically associating domains are ancient features that coincide with Metazoan clusters of extreme noncoding conservation. *Nat Commun*, 8(1):441.

Haudry, A., Platts, A. E., Vello, E., Hoen, D. R., Leclercq, M., Williamson, R. J., Forczek, E., Joly-Lopez, Z., Steffen, J. G., Hazzouri, K. M., Dewar, K., Stinchcombe, J. R., Schoen, D. J., Wang, X., Schmutz, J., Town, C. D., Edger, P. P., Pires, J. C., Schumaker, K. S., Jarvis, D. E., Mandáková, T., Lysak, M. A., van den Bergh, E., Schranz, M. E., Harrison, P. M., Moses, A. M., Bureau, T. E., Wright, S. I., and Blanchette, M. (2013). An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. *Nat Genet*, 45(8):891–898.

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

Henkel, C. V., Dirks, R. P., Jansen, H. J., Forlenza, M., Wiegertjes, G. F., Howe, K., van den Thillart, G. E. E. J. M., and Spaink, H. P. (2012). Comparison of the exomes of common carp (Cyprinus carpio) and zebrafish (Danio rerio). *Zebrafish*, 9(2):59–67.

Hiller, M., Agarwal, S., Notwell, J. H., Parikh, R., Guturu, H., Wenger, A. M., and Bejerano, G. (2013). Computational methods to detect conserved non-genic elements in phylogenetically isolated genomes: application to zebrafish. *Nucleic Acids Res.*, 41(15):e151.

Hur, J. H., Van Doninck, K., Mandigo, M. L., and Meselson, M. (2009). Degenerate tetraploidy was established before bdelloid rotifer families diverged. *Mol. Biol. Evol.*, 26(2):375–383.

International Glossina Genome Initiative (2014). Genome sequence of the tsetse fly (Glossina morsitans): vector of African trypanosomiasis. *Science*, 344(6182):380–386.

International Human Genome Sequencing Consortium (2004). Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–945.

Jaillon, O., Aury, J.-M., Brunet, F., Petit, J.-L., Stange-Thomann, N., Mauceli, E., Bouneau, L., Fischer, C., Ozouf-Costaz, C., Bernot, A., Nicaud, S., Jaffe, D., Fisher, S., Lutfalla, G., Dossat, C., Segurens, B., Dasilva, C., Salanoubat, M., Levy, M., Boudet, N., Castellano, S., Anthouard, V., Jubin, C., Castelli, V., Katinka, M., Vacherie, B., Biémont, C., Skalli, Z., Cattolico, L., Poulain, J., De Berardinis, V., Cruaud, C., Duprat, S., Brottier, P., Coutanceau, J.-P., Gouzy, J., Parra, G., Lardier, G., Chapple, C., McKernan, K. J., McEwan, P., Bosak, S., Kellis, M., Volff, J.-N., Guigó, R., Zody, M. C., Mesirov, J., Lindblad-Toh, K., Birren, B., Nusbaum, C., Kahn, D., Robinson-Rechavi, M., Laudet, V., Schachter, V., Quétier, F., Saurin, W., Scarpelli, C., Wincker, P., Lander, E. S., Weissenbach, J., and Roest Crollius, H. (2004). Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. *Nature*, 431(7011):946–957.

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A. F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S.-Y., Lopez, R., and Hunter, S. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics*, 30(9):1236–1240.

Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.*, 110(1-4):462–467.

Katoh, K. and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.*, 30(4):772–780.

Keller, O., Kollmar, M., Stanke, M., and Waack, S. (2011). A novel hybrid gene prediction method employing protein multiple sequence alignments. *Bioinformatics*, 27(6):757–763.

Kent, W. J. (2002). BLAT–the BLAST-like alignment tool. *Genome Res.*, 12(4):656–664.

Kent, W. J., Baertsch, R., Hinrichs, A., Miller, W., and Haussler, D. (2003). Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc. Natl. Acad. Sci. U.S.A.*, 100(20):11484–11489.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., and Haussler, D. (2002). The human genome browser at UCSC. *Genome Res.*, 12(6):996–1006.

Khan, A., Fornes, O., Stigliani, A., Gheorghe, M., Castro-Mondragon, J. A., van der Lee, R., Bessy, A., Chèneby, J., Kulkarni, S. R., Tan, G., Baranasic, D., Arenillas, D. J., Sandelin, A., Vandepoele, K., Lenhard, B., Ballester, B., Wasserman, W. W., Parcy, F., and Mathelier, A. (2017). JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res.*

Khurana, E., Fu, Y., Chakravarty, D., Demichelis, F., Rubin, M. A., and Gerstein, M. (2016). Role of non-coding sequence variants in cancer. *Nat Rev Genet*, 17(2):93–108.

Kiełbasa, S. M., Wan, R., Sato, K., Horton, P., and Frith, M. C. (2011). Adaptive seeds tame genomic sequence comparison. *Genome Res.*, 21(3):487–493.

Kikuta, H., Fredman, D., Rinkwitz, S., Lenhard, B., and Becker, T. S. (2007a). Retroviral enhancer detection insertions in zebrafish combined with comparative genomics reveal genomic regulatory blocks - a fundamental feature of vertebrate genomes. *Genome Biol.*, 8 Suppl 1:S4.

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

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

Kolder, I. C. R. M., van der Plas-Duivesteijn, S. J., Tan, G., Wiegertjes, G. F., Forlenza, M., Guler, A. T., Travin, D. Y., Nakao, M., Moritomo, T., Irnazarow, I., den Dunnen, J. T., Anvar, S. Y., Jansen, H. J., Dirks, R. P., Palmblad, M., Lenhard, B., Henkel, C. V., and Spaink, H. P. (2016). A full-body transcriptome and proteome resource for the European common carp. *BMC Genomics*, 17:701.

Korf, I. (2004). Gene finding in novel genomes. *BMC Bioinformatics*, 5:59.

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M. T., and Carey, V. J. (2013). Software for computing and annotating genomic ranges. *PLoS Comput. Biol.*, 9(8):e1003118.

Lee, A. P., Kerk, S. Y., Tan, Y. Y., Brenner, S., and Venkatesh, B. (2011). Ancient vertebrate conserved noncoding elements have been evolving rapidly in teleost fishes. *Mol. Biol. Evol.*, 28(3):1205–1215.

Lenhard, B., Sandelin, A., Mendoza, L., Engström, P., Jareborg, N., and Wasserman, W. W. (2003). Identification of conserved regulatory elements by comparative genome analysis. *J. Biol.*, 2(2):13.

Lenhard, B. and Wasserman, W. W. (2002). TFBS: Computational framework for transcription factor binding site analysis. *Bioinformatics*, 18(8):1135–1136.

Li, J.-T., Hou, G.-Y., Kong, X.-F., Li, C.-Y., Zeng, J.-M., Li, H.-D., Xiao, G.-B., Li, X.-M., and Sun, X.-W. (2015). The fate of recent duplicated genes following a fourth-round whole genome duplication in a tetraploid fish, common carp (Cyprinus carpio). *Scientific Reports*, 5:8199.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomforest. *R News*, 2(3):18–22.

Lien, S., Koop, B. F., Sandve, S. R., Miller, J. R., Kent, M. P., Nome, T., Hvidsten, T. R., Leong, J. S., Minkley, D. R., Zimin, A., Grammes, F., Grove, H., Gjuvsland, A., Walenz, B., Hermansen, R. A., von Schalburg, K., Rondeau, E. B., Di Genova, A., Samy, J. K. A., Olav Vik, J., Vigeland, M. D., Caler, L., Grimholt, U., Jentoft, S., Inge Våge, D., de Jong, P., Moen, T., Baranski, M., Palti, Y., Smith, D. R., Yorke, J. A., Nederbragt, A. J., Tooming-Klunderud, A., Jakobsen, K. S., Jiang, X., Fan, D., Hu, Y., Liberles, D. A., Vidal, R., Iturra, P., Jones, S. J. M., Jonassen, I., Maass, A., Omholt, S. W., and Davidson, W. S. (2016). The Atlantic salmon genome provides insights into rediploidization. *Nature*, advance online publication.

Linhart, C., Halperin, Y., and Shamir, R. (2008). Transcription factor and microRNA motif discovery: the Amadeus platform and a compendium of metazoan target sets. *Genome Res.*, 18(7):1180–1189.

Liu, Y., Dehni, G., Purcell, K. J., Sokolow, J., Carcangiu, M. L., Artavanis-Tsakonas, S., and Stifani, S. (1996). Epithelial expression and chromosomal location of human TLE genes: implications for notch signaling and neoplasia. *Genomics*, 31(1):58–64.

Lowe, C. B., Kellis, M., Siepel, A., Raney, B. J., Clamp, M., Salama, S. R., Kingsley, D. M., Lindblad-Toh, K., and Haussler, D. (2011). Three periods of regulatory innovation during vertebrate evolution. *Science*, 333(6045):1019–1024.

Mathelier, A., Fornes, O., Arenillas, D. J., Chen, C.-y., Denay, G., Lee, J., Shi, W., Shyr, C., Tan, G., Worsley-Hunt, R., Zhang, A. W., Parcy, F., Lenhard, B., Sandelin, A., and Wasserman, W. W. (2016). JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucl. Acids Res.*, 44(D1):D110–D115.

Mathelier, A. and Wasserman, W. W. (2013). The next generation of transcription factor binding site prediction. *PLoS Comput. Biol.*, 9(9):e1003214.

Mathelier, A., Zhao, X., Zhang, A. W., Parcy, F., Worsley-Hunt, R., Arenillas, D. J., Buchman, S., Chen, C.-y., Chou, A., Ienasescu, H., Lim, J., Shyr, C., Tan, G., Zhou, M., Lenhard, B., Sandelin, A., and Wasserman, W. W. (2014). JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, 42(Database issue):D142–147.

McCole, R. B., Fonseka, C. Y., Koren, A., and Wu, C.-T. (2014). Abnormal dosage of ultraconserved elements is highly disfavored in healthy cells but not cancer cells. *PLoS Genet.*, 10(10):e1004646.

Montalbano, A., Canver, M. C., and Sanjana, N. E. (2017). High-Throughput Approaches to Pinpoint Function within the Noncoding Genome. *Mol. Cell*, 68(1):44–59.

Mumbach, M. R., Satpathy, A. T., Boyle, E. A., Dai, C., Gowen, B. G., Cho, S. W., Nguyen, M. L., Rubin, A. J., Granja, J. M., Kazane, K. R., Wei, Y., Nguyen, T., Greenside, P. G., Corces, M. R., Tycko, J., Simeonov, D. R., Suliman, N., Li, R., Xu, J., Flynn, R. A., Kundaje, A., Khavari, P. A., Marson, A., Corn, J. E., Quertermous, T., Greenleaf, W. J., and Chang, H. Y. (2017). Enhancer connectome in primary human cells identifies target genes of disease-associated DNA elements. *Nat. Genet.*

Navratilova, P. and Becker, T. S. (2009). Genomic regulatory blocks in vertebrates and implications in human disease. *Brief Funct Genomic Proteomic*, 8(4):333–342.

Navratilova, P., Fredman, D., Hawkins, T. A., Turner, K., Lenhard, B., and Becker, T. S. (2009). Systematic human/zebrafish comparative identification of cis-regulatory activity around vertebrate developmental transcription factor genes. *Dev. Biol.*, 327(2):526–540.

Nepveu, A. (2001). Role of the multifunctional CDP/Cut/Cux homeodomain transcription factor in regulating differentiation, cell growth and development. *Gene*, 270(1-2):1–15.

Nishida, K., Frith, M. C., and Nakai, K. (2009). Pseudocounts for transcription factor binding sites. *Nucleic Acids Res.*, 37(3):939–944.

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

O'Donoghue, P. and Luthey-Schulten, Z. (2003). On the evolution of structure in aminoacyl-tRNA synthetases. *Microbiol. Mol. Biol. Rev.*, 67(4):550–573.

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

Persampieri, J., Ritter, D. I., Lees, D., Lehoczky, J., Li, Q., Guo, S., and Chuang, J. H. (2008). cneViewer: a database of conserved non-coding elements for studies of tissue-specific gene regulation. *Bioinformatics*, 24(20):2418–2419.

Ragvin, A., Moro, E., Fredman, D., Navratilova, P., Drivenes, o., Engström, P. G., Alonso, M. E., de la Calle Mustienes, E., Gómez Skarmeta, J. L., Tavares, M. J., Casares, F., Manzanares, M., van Heyningen, V., Molven, A., Njolstad, P. R., Argenton, F., Lenhard, B., and Becker, T. S. (2010). Long-range gene regulation links genomic type 2 diabetes and obesity risk regions to HHEX, SOX4, and IRX3. *Proc. Natl. Acad. Sci. U.S.A.*, 107(2):775–780.

Reim, I., Mohler, J. P., and Frasch, M. (2005). Tbx20-related genes, mid and H15, are required for tinman expression, proper patterning, and normal differentiation of cardioblasts in Drosophila. *Mech. Dev.*, 122(9):1056–1069.

Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., Amin, V., Whitaker, J. W., Schultz, M. D., Ward, L. D., Sarkar, A., Quon, G., Sandstrom, R. S., Eaton, M. L., Wu, Y.-C., Pfenning, A. R., Wang, X., Claussnitzer, M., Liu, Y., Coarfa, C., Harris, R. A., Shoresh, N., Epstein, C. B., Gjoneska, E., Leung, D., Xie, W., Hawkins, R. D., Lister, R., Hong, C., Gascard, P., Mungall, A. J., Moore, R., Chuah, E., Tam, A., Canfield, T. K., Hansen, R. S., Kaul, R., Sabo, P. J., Bansal, M. S., Carles, A., Dixon, J. R., Farh, K.-H., Feizi, S., Karlic, R., Kim, A.-R., Kulkarni, A., Li, D., Lowdon, R., Elliott, G., Mercer, T. R., Neph, S. J., Onuchic, V., Polak, P., Rajagopal, N., Ray, P., Sallari, R. C., Siebenthall, K. T., Sinnott-Armstrong, N. A., Stevens, M., Thurman, R. E., Wu, J., Zhang, B., Zhou, X., Beaudet, A. E., Boyer, L. A., De Jager, P. L., Farnham, P. J., Fisher, S. J., Haussler, D., Jones, S. J. M., Li, W., Marra, M. A., McManus, M. T., Sunyaev, S., Thomson, J. A., Tlsty, T. D., Tsai, L.-H., Wang, W., Waterland, R. A., Zhang, M. Q., Chadwick, L. H., Bernstein, B. E., Costello, J. F., Ecker, J. R., Hirst, M., Meissner, A., Milosavljevic, A., Ren, B., Stamatoyannopoulos, J. A., Wang, T., and Kellis, M. (2015). Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330.

Roulin, A., Auer, P. L., Libault, M., Schlueter, J., Farmer, A., May, G., Stacey, G., Doerge, R. W., and Jackson, S. A. (2013). The fate of duplicated genes in a polyploid plant genome. *Plant J.*, 73(1):143–153.

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

Salerno, W., Havlak, P., and Miller, J. (2006). Scale-invariant structure of strongly conserved sequence in genomic intersections and alignments. *Proc. Natl. Acad. Sci. U.S.A.*, 103(35):13121–13125.

Sandelin, A., Bailey, P., Bruce, S., Engström, P. G., Klos, J. M., Wasserman, W. W., Ericson, J., and Lenhard, B. (2004). Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes. *BMC Genomics*, 5(1):99.

Sandelin, A., Höglund, A., Lenhard, B., and Wasserman, W. W. (2003). Integrated analysis of yeast regulatory sequences for biologically linked clusters of genes. *Funct. Integr. Genomics*, 3(3):125–134.

Sanjana, N. E., Wright, J., Zheng, K., Shalem, O., Fontanillas, P., Joung, J., Cheng, C., Regev, A., and Zhang, F. (2016). High-resolution interrogation of functional elements in the noncoding genome. *Science*, 353(6307):1545–1549.

Schneider, T. D. and Stephens, R. M. (1990). Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, 18(20):6097–6100.

Schneider, T. D., Stormo, G. D., Gold, L., and Ehrenfeucht, A. (1986). Information content of binding sites on nucleotide sequences. *J. Mol. Biol.*, 188(3):415–431.

Schwartz, S., Kent, W. J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R. C., Haussler, D., and Miller, W. (2003). Human-mouse alignments with BLASTZ. *Genome Res.*, 13(1):103–107.

Signorovitch, A., Hur, J., Gladyshev, E., and Meselson, M. (2015). Allele Sharing and Evidence for Sexuality in a Mitochondrial Clade of Bdelloid Rotifers. *Genetics*, 200(2):581–590.

Silla, T., Kepp, K., Tai, E. S., Goh, L., Davila, S., Catela Ivkovic, T., Calin, G. A., and Voorhoeve, P. M. (2014). Allele frequencies of variants in ultra conserved elements identify selective pressure on transcription factor binding. *PLoS ONE*, 9(11):e110692.

Slater, G. S. C. and Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, 6:31.

Sloan, C. A., Chan, E. T., Davidson, J. M., Malladi, V. S., Strattan, J. S., Hitz, B. C., Gabdank, I., Narayanan, A. K., Ho, M., Lee, B. T., Rowe, L. D., Dreszer, T. R., Roe, G., Podduturi, N. R., Tanaka, F., Hong, E. L., and Cherry, J. M. (2016). ENCODE data at the ENCODE portal. *Nucleic Acids Res.*, 44(D1):D726–732.

Smemo, S., Tena, J. J., Kim, K.-H., Gamazon, E. R., Sakabe, N. J., Gómez-Marín, C., Aneas, I., Credidio, F. L., Sobreira, D. R., Wasserman, N. F., Lee, J. H., Puviindran, V., Tam, D., Shen, M., Son, J. E., Vakili, N. A., Sung, H.-K., Naranjo, S., Acemel, R. D., Manzanares, M., Nagy, A., Cox, N. J., Hui, C.-C., Gomez-Skarmeta, J. L., and Nóbrega, M. A. (2014). Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature*, 507(7492):371–375.

Smit, A., Hubley, R., and Green, P. (2015). Repeatmasker open-4.0. 2013–2015. *Institute for Systems Biology. http://repeatmasker. org*.

Sonnhammer, E. L. L. and Östlund, G. (2015). InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. *Nucleic Acids Res.*, 43(Database issue):D234–239.

Tan, G. and Lenhard, B. (2016). TFBSTools: an R/bioconductor package for transcription factor binding site analysis. *Bioinformatics*, 32(10):1555–1556.

Vavouri, T., Walter, K., Gilks, W. R., Lehner, B., and Elgar, G. (2007). 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.

Visel, A., Minovitsky, S., Dubchak, I., and Pennacchio, L. A. (2007). VISTA Enhancer Browser–a database of tissue-specific human enhancers. *Nucleic Acids Res.*, 35(Database issue):D88–92.

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

Wang, Y., Tang, H., Debarry, J. D., Tan, X., Li, J., Wang, X., Lee, T.-h., Jin, H., Marler, B., Guo, H., Kissinger, J. C., and Paterson, A. H. (2012). MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.*, 40(7):e49.

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

Wolfe, K. H. (2001). Yesterday's polyploids and the mystery of diploidization. *Nat. Rev. Genet.*, 2(5):333–341.

Wolfe, K. H. and Shields, D. C. (1997). Molecular evidence for an ancient duplication of the entire yeast genome. *Nature*, 387(6634):708–713.

Woltering, J. M. and Durston, A. J. (2006). The zebrafish hoxDb cluster has been reduced to a single microRNA. *Nat. Genet.*, 38(6):601–602.

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

Wright, J. B. and Sanjana, N. E. (2016). CRISPR Screens to Discover Functional Noncoding Elements. *Trends in Genetics*, 32(9):526–529.

Xie, X., Mikkelsen, T. S., Gnirke, A., Lindblad-Toh, K., Kellis, M., and Lander, E. S. (2007). Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites. *Proc Natl Acad Sci U S A*, 104(17):7145–7150.

Xu, P., Zhang, X., Wang, X., Li, J., Liu, G., Kuang, Y., Xu, J., Zheng, X., Ren, L., Wang, G., Zhang, Y., Huo, L., Zhao, Z., Cao, D., Lu, C., Li, C., Zhou, Y., Liu, Z., Fan, Z., Shan, G., Li, X., Wu, S., Song, L., Hou, G., Jiang, Y., Jeney, Z., Yu, D., Wang, L., Shao, C., Song, L., Sun, J., Ji, P., Wang, J., Li, Q., Xu, L., Sun, F., Feng, J., Wang, C., Wang, S., Wang, B., Li, Y., Zhu, Y., Xue, W., Zhao, L., Wang, J., Gu, Y., Lv, W., Wu, K., Xiao, J., Wu, J., Zhang, Z., Yu, J., and Sun, X. (2014). Genome sequence and genetic diversity of the common carp, Cyprinus carpio. *Nat. Genet.*, 46.

Yandell, M. and Ence, D. (2012). A beginner's guide to eukaryotic genome annotation. *Nat. Rev. Genet.*, 13(5):329–342.

Yates, A., Akanni, W., Amode, M. R., Barrell, D., Billis, K., Carvalho-Silva, D., Cummins, C., Clapham, P., Fitzgerald, S., Gil, L., Girón, C. G., Gordon, L., Hourlier, T., Hunt, S. E., Janacek, S. H., Johnson, N., Juettemann, T., Keenan, S., Lavidas, I., Martin, F. J., Maurel, T., McLaren, W., Murphy, D. N., Nag, R., Nuhn, M., Parker, A., Patricio, M., Pignatelli, M., Rahtz, M., Riat, H. S., Sheppard, D., Taylor, K., Thormann, A., Vullo, A., Wilder, S. P., Zadissa, A., Birney, E., Harrow, J., Muffato, M., Perry, E., Ruffier, M., Spudich, G., Trevanion, S. J., Cunningham, F., Aken, B. L., Zerbino, D. R., and Flicek, P. (2016). Ensembl 2016. *Nucleic Acids Res.*, 44(D1):D710–716.

Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS*, 16(5):284–287.

Yu, G., Wang, L.-G., Yan, G.-R., and He, Q.-Y. (2015). DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis. *Bioinformatics*, 31(4):608–609.

Yue, F., Cheng, Y., Breschi, A., Vierstra, J., Wu, W., Ryba, T., Sandstrom, R., Ma, Z., Davis, C., Pope, B. D., Shen, Y., Pervouchine, D. D., Djebali, S., Thurman, R. E., Kaul, R., Rynes, E., Kirilusha, A., Marinov, G. K., Williams, B. A., Trout, D., Amrhein, H., Fisher-Aylor, K., Antoshechkin, I., DeSalvo, G., See, L.-H., Fastuca, M., Drenkow, J., Zaleski, C., Dobin, A., Prieto, P., Lagarde, J., Bussotti, G., Tanzer, A., Denas, O., Li, K., Bender, M. A., Zhang, M., Byron, R., Groudine, M. T., McCleary, D., Pham, L., Ye, Z., Kuan, S., Edsall, L., Wu, Y.-C., Rasmussen, M. D., Bansal, M. S., Kellis, M., Keller, C. A., Morrissey, C. S., Mishra, T., Jain, D., Dogan, N., Harris, R. S., Cayting, P., Kawli, T., Boyle, A. P., Euskirchen, G., Kundaje, A., Lin, S., Lin, Y., Jansen, C., Malladi, V. S., Cline, M. S., Erickson, D. T., Kirkup, V. M., Learned, K., Sloan, C. A., Rosenbloom, K. R., Lacerda de Sousa, B., Beal, K., Pignatelli, M., Flicek, P., Lian, J., Kahveci, T., Lee, D., Kent, W. J., Ramalho Santos, M., Herrero, J., Notredame, C., Johnson, A., Vong, S., Lee, K., Bates, D., Neri, F., Diegel, M., Canfield, T., Sabo, P. J., Wilken, M. S., Reh, T. A., Giste, E., Shafer, A., Kutyavin, T., Haugen, E., Dunn, D., Reynolds, A. P., Neph, S., Humbert, R., Hansen, R. S., De Bruijn, M., Selleri, L., Rudensky, A., Josefowicz, S., Samstein, R., Eichler, E. E., Orkin, S. H., Levasseur, D., Papayannopoulou, T., Chang, K.-H., Skoultchi, A., Gosh, S., Disteche, C., Treuting, P., Wang, Y., Weiss, M. J., Blobel, G. A., Cao, X., Zhong, S., Wang, T., Good, P. J., Lowdon, R. F., Adams, L. B., Zhou, X.-Q., Pazin, M. J., Feingold, E. A., Wold, B., Taylor, J., Mortazavi, A., Weissman, S. M., Stamatoyannopoulos, J. A., Snyder, M. P., Guigo, R., Gingeras, T. R., Gilbert, D. M., Hardison, R. C., Beer, M. A., Ren, B., and Mouse ENCODE Consortium (2014). A comparative encyclopedia of DNA elements in the mouse genome. *Nature*, 515(7527):355–364.

Zhang, X., Friedman, A., Heaney, S., Purcell, P., and Maas, R. L. (2002). Meis homeoproteins directly regulate Pax6 during vertebrate lens morphogenesis. *Genes Dev.*, 16(16):2097–2107.

Zhao, J., Xu, D., Zhao, K., Diogo, R., Yang, J., and Peng, Z. (2016). The origin and divergence of Gobioninae fishes (Teleostei: Cyprinidae) based on complete mitochondrial genome sequences. *J. Appl. Ichthyol.*, 32(1):32–39.

Appendix A

Appendix for Chapter 2

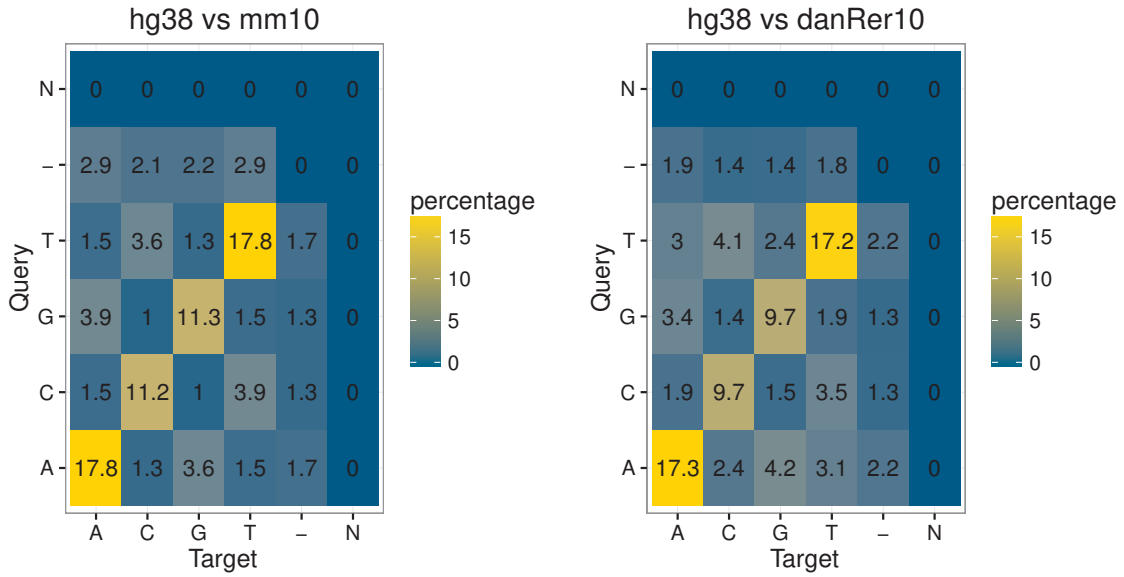

Figure A.1 The percentage of matched bases in the Axt alignment. The left panel is the alignment from hg38 to mm10. The right panel is the alignment from hg38 to danRer10.

Coordinates	Targets	Type	# of CNEs
chr3R:16625157-16857938	Ubx, abd-A	Homeobox	90
chr3R:6,610,959-7,088,732	pb, lab, zen, zen2, ftz, Antp..	Homeobox,	73
chr3R:17542354-17683793	Hmx	Homeobox	73
chr2L:5326653-5523912	H15, mid	T-box binding domain	70

chr3R:10391341-10644417	hth	Homeobox	69
chr2L:14,345,689..14,611,960	noc, elB	C2H2 Zn finger	64
chr2L:11362748-11495734	salr, salm	C2H2 Zn finger	46
chr2R:20879784-21024512	hbn, opt	Homeobox	41
chr3R:12269437-12354168	svp	Nuclear receptor gene	40
chr3R:28496608-28603525	fkh	Forkhead	39
chr3R:8125198-8242389	grn	ZnF_GATA	34
chr3R:29486423-29596589	Dr	Homeobox	32
chr3R:21497278-21572822	slou	Homeobox	32
chr3L:13386527-13433948	sens	C2H2 Zn finger	29
chr2R:25035237-25126556	gsb-n, gsb	Homeobox	26
chr3L:6088108-6166450	l(3)mbn	Other	24
chr3R:13899600-13969817	ems	Homeobox	22
chr3R:16381020-16436909	ss	PAS domain	22
chr3R:25986074-26081891	gro	Other	21
chr3L:21436650-21543885	sim	PAS domain	21
chr3R:17,989,000-18,159,999	htl	Other	34
chr3R:8931589-8981499	osk	Other	14
chr2L:18845907-18890674	tup	Homeodomain	14
chr2R:7984667-8102167	Optix	homeodomain	14
chr2R:14748124-14798975	kn	Transcription factor COE1 helix-loop-helix domain	14
chr3L:21573335-21641140	TfAP-2	Transcription factor AP-2	14
chr3L:4058773-4098012	dib	Other	13
chr2L:3771696-3847958	slp1, slp2	Forkhead	13
chr2R:14,420,258..14,466,590	Oaz	C2H2 Zn finger	13
chr3L:6788580-6819508	vvl	Homeobox	13
chr3L:4136835-4182890	nab	NAB conserved region	12

chr3R:30893999-30955611	Ptx1	Homeobox	12
chr3L:13849016-13929440	dysc	Other	12
chr2R:14168606-14234658	Sox15	HMG-box	11
chr2R:22855288-22877666	fd59A	Forkhead	11
chr3L:1925049-1954572	Dbx	Homeobox	11
chr2L:22019262-22034092	tio	C2H2 Zn finger	11
chr3R:6653795-6702034	lab	Homeobox	10
chr3R:21427794-21453961	lbe	Homeobox	10
chr2R:17800554-17848639	grh	CP2 transcription factor	10

Table A.1 A list of the most prominent CNE clusters detected between *Drosophila* and *Glossina*.

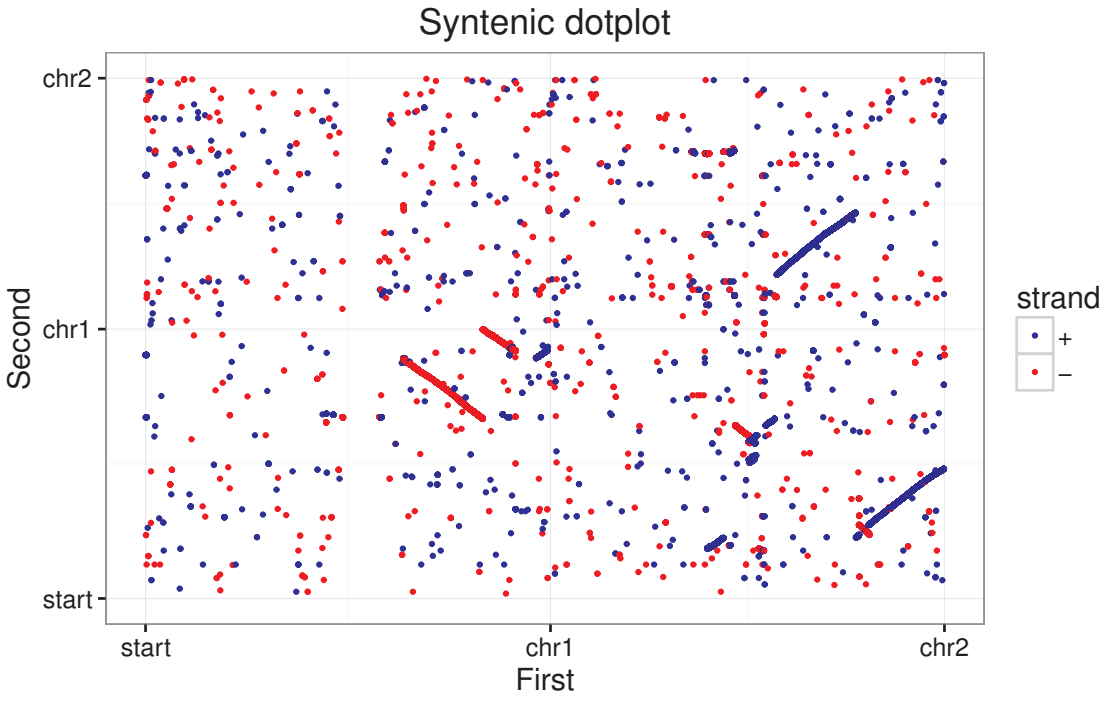

Figure A.2 The plot of alignment blocks between chr1, chr2 of human and chr1, chr2 of mouse. This plot is mostly used for tuning the parameters during whole genome pairwise alignment to get better alignments. It can also show ancient duplications for the alignment of a sequence against itself.

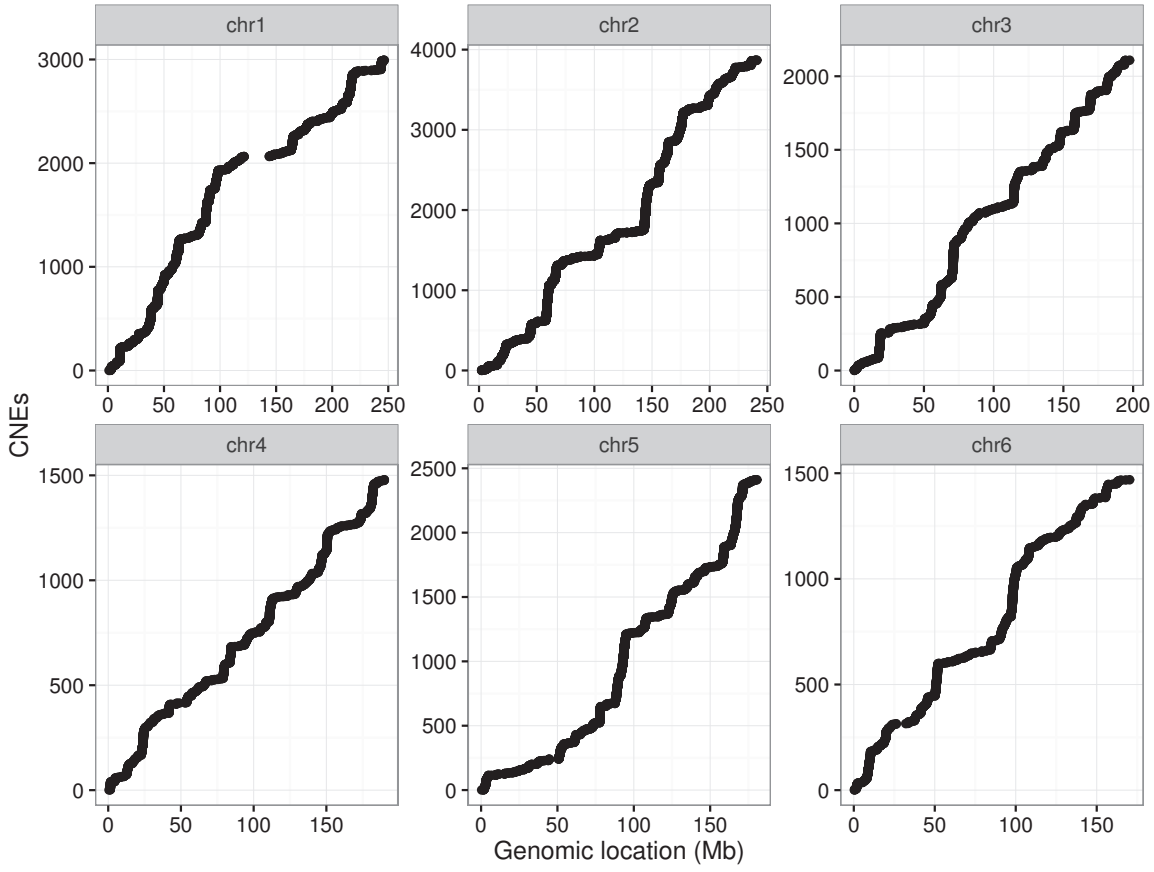

Figure A.3 The distribution of human vs. mouse CNEs along the 6 biggest chromosomes in human genome. Each CNE is plotted as a dot with the position in chromosome as x-axis. A sharp increase in y-axis represents a CNE cluster.

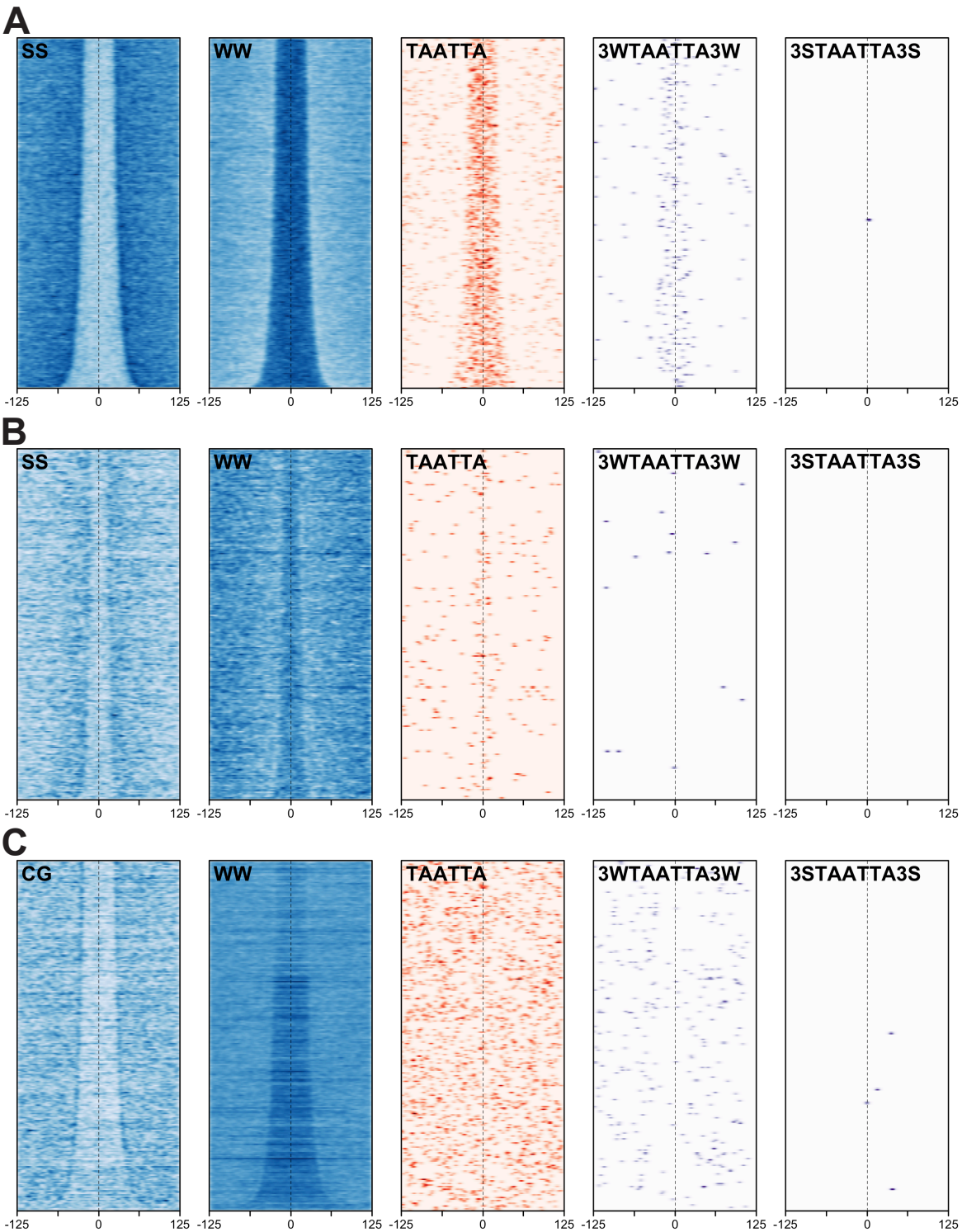

Figure A.4 Sequence patterns of CNEs in different lineages. (A) *D. melanogaster* and *D. virilis* (B) *C. elegans* and *C. briggsae* (C) *L. variegatus* and *S. purpuratus*. These plots were produced by Dimitris Polychronopoulos, who is a collaborator on this CNEr project.

Appendix B

Appendix for Chapter 4

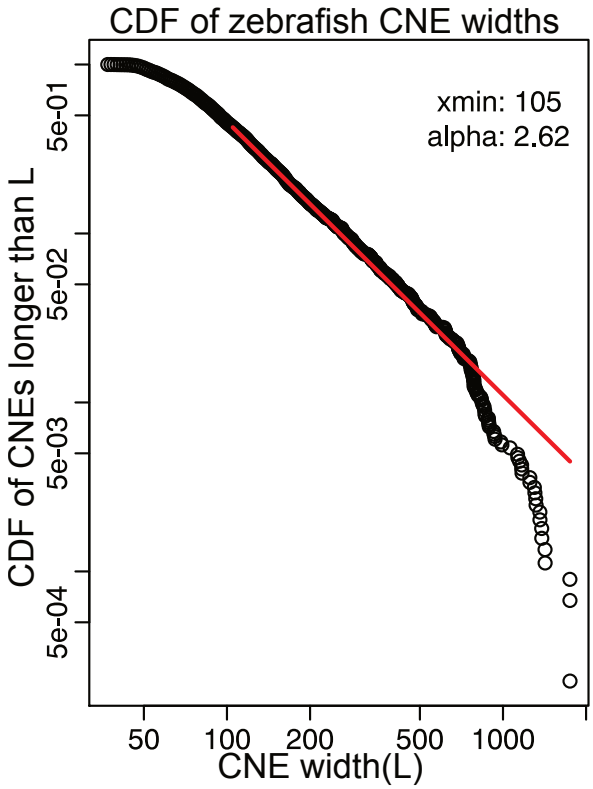

Figure B.1 A power-law distribution of the lengths of zebrafish CNEs from self-alignment. In this log-log plot, the CDF of CNEs longer than L follows a linear relation with the CNE width L between 100 and 500.

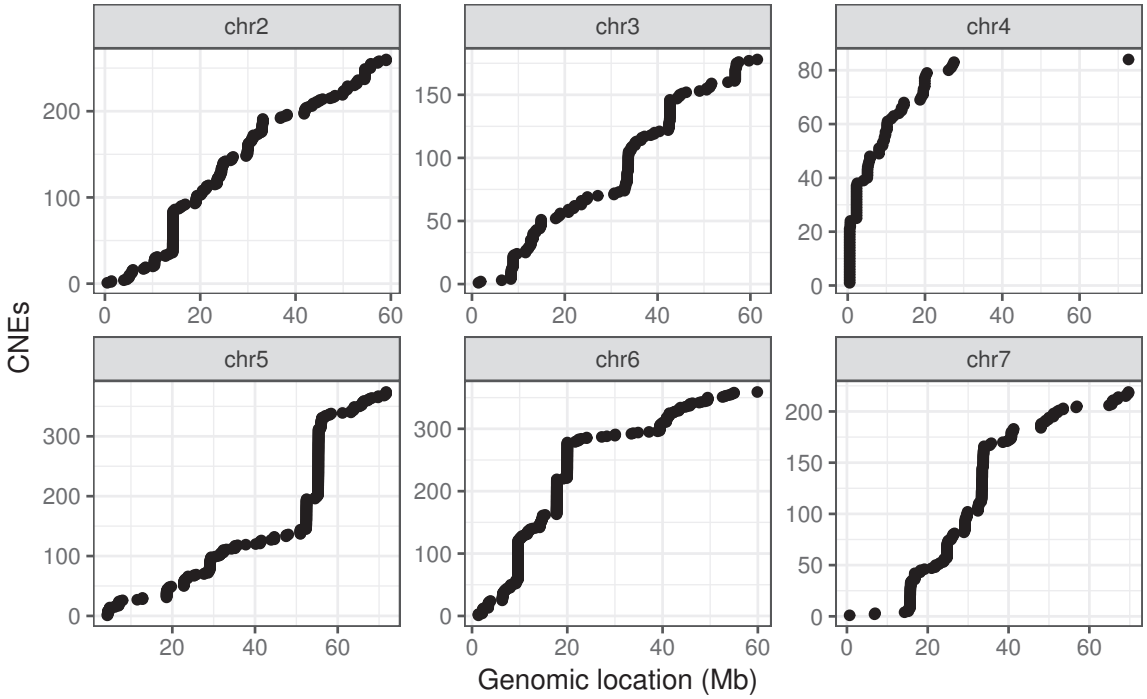

Figure B.2 The distribution of CNEs from self-alignments along the 6 biggest chromosomes in zebrafish genome. Each CNE is plotted as a dot with the position in chromosome as x-axis. A sharp increase in y-axis represents a CNE cluster.

Table B.1 Allelic and ohnologous collinear blocks

	# of collinear blocks	# of genes	total length	median identity at DNA level
allelic	740	14764	66 Mb	98.6%
ohnologous	836	12181	90 Mb	75.1%

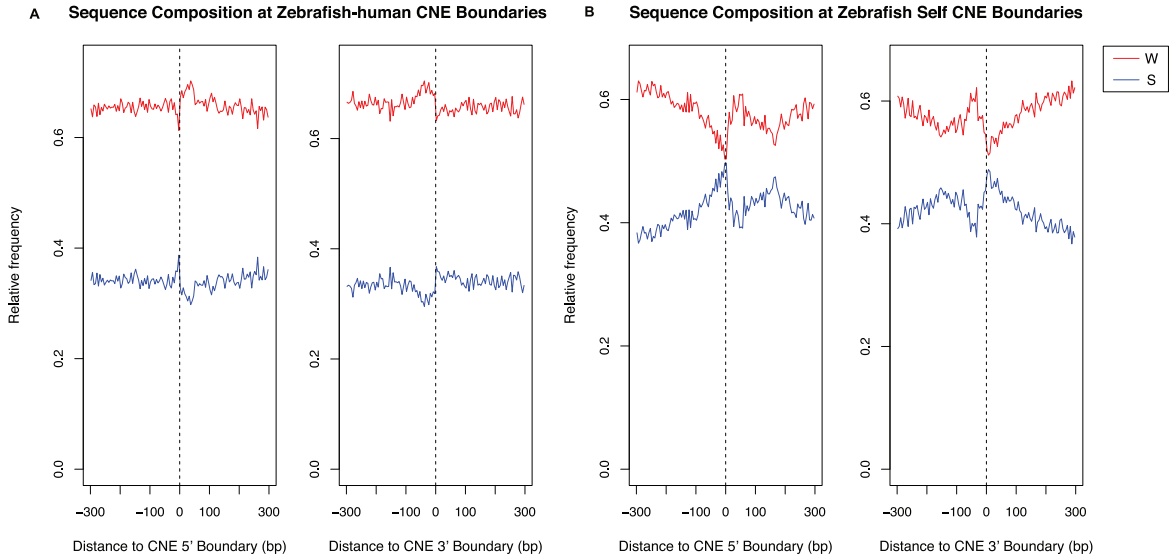

Figure B.3 Sequence composition at CNE boundaries. (A) The zebrafish-human CNEs from standard pairwise comparison has a depletion of G/C content at 5' and 3' bounadires. (B) The zebraifhs CNEs from self-alignment exhibits the same pattern. This plot was made by Alex Nash, who is a collaborator on this Rotifer CNE project.

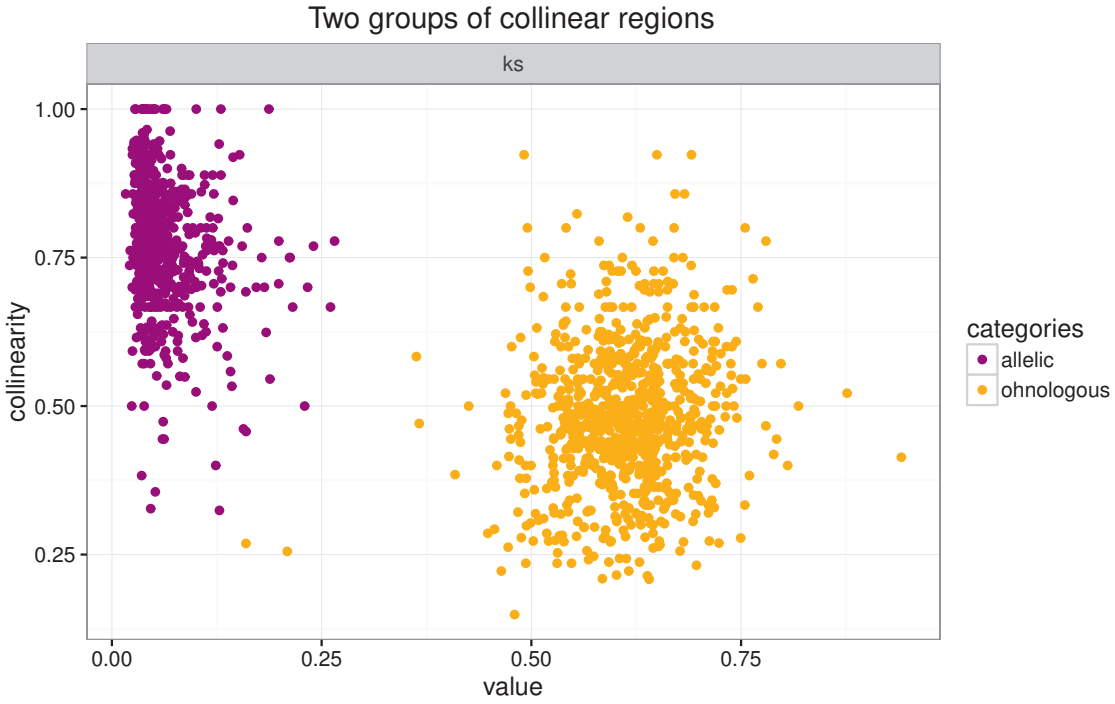

Figure B.4 Two groups of collinear regions. The ratio of synonymous rate to collinearity is used to classify these two regions. The orange ohnologous regions has a ratio > 0.5, and the rest are purple allelic regions.

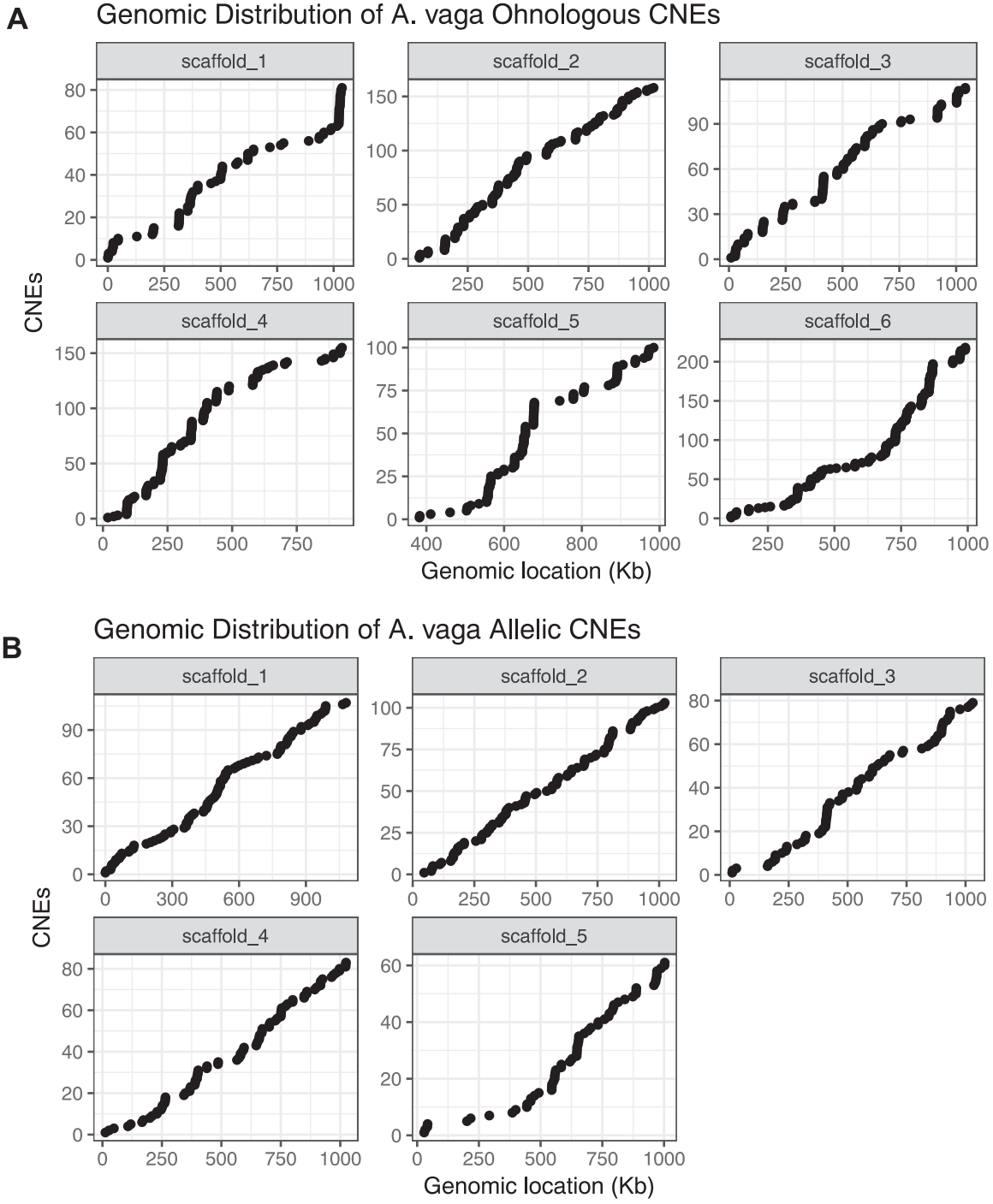

Figure B.5 The distribution of rotifer CNEs from two types of collinear regions along the 6 biggest scaffolds. Each CNE is plotted as a dot with the position in scaffold as x-axis. A sharp increase in y-axis represents a CNE cluster. (A) Ohnologous CNEs, 70% identity over 50bp. (B) Allelic CNEs, 100% identity over 250bp. scaffold_6 is not shown because there is no CNEs detected on this scaffold.

Appendix C

Appendix for Chapter 5

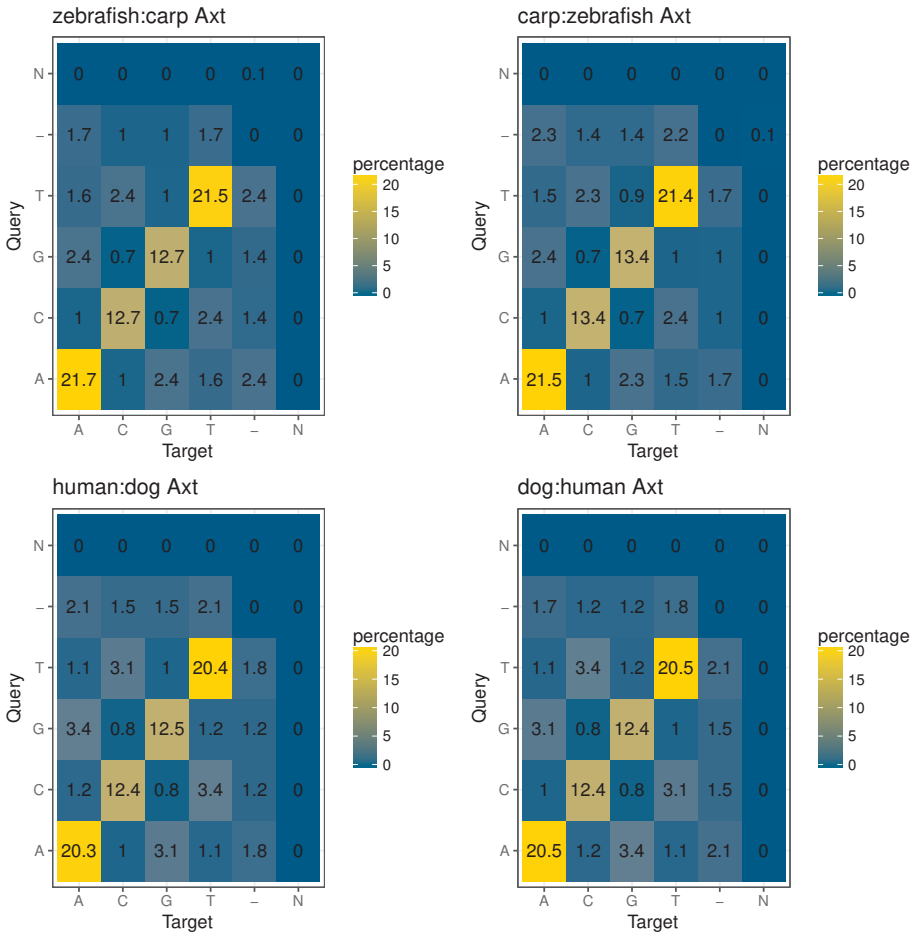

Figure C.1 Alignment matches comparison between zebrafish vs. carp and human vs. dog. The identity rate of zebrafish vs. carp is slightly higher than human vs. dog.

Appendix D

Appendix for Chapter 6

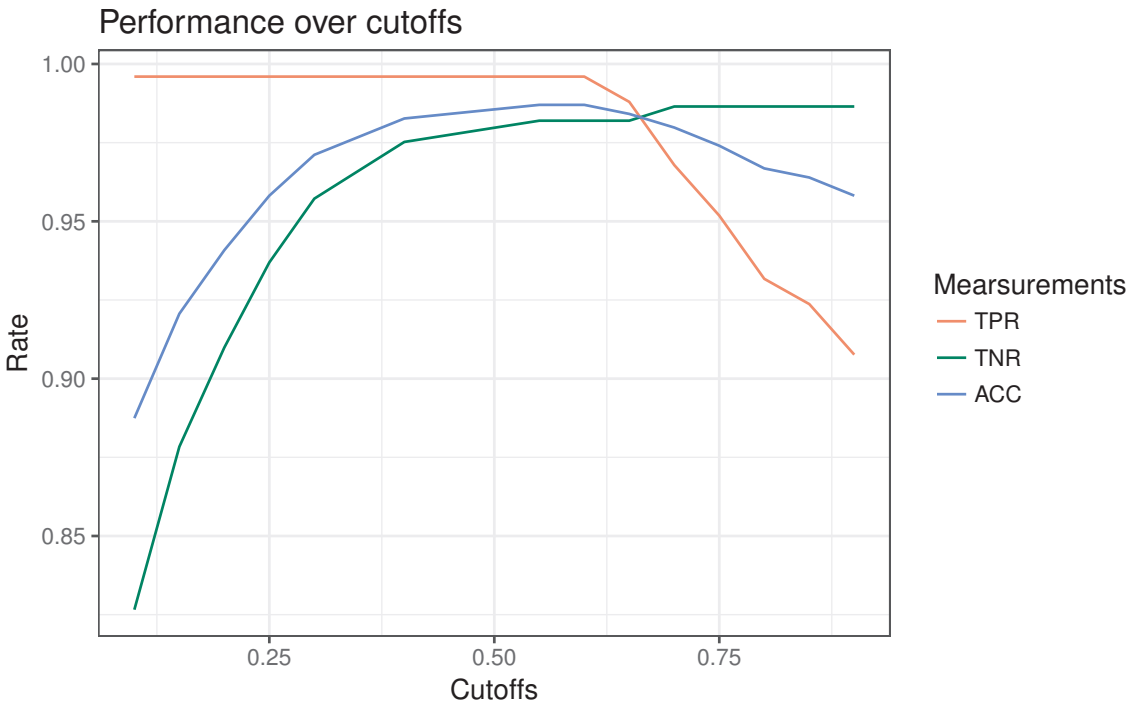

Figure D.1 The performance of random forest model over various cutoffs. A cutoff of 0.6 is chosen to achieve the highest accuracy. (TPR: true positive rate; TNR: true negative rate; ACC: accuracy).

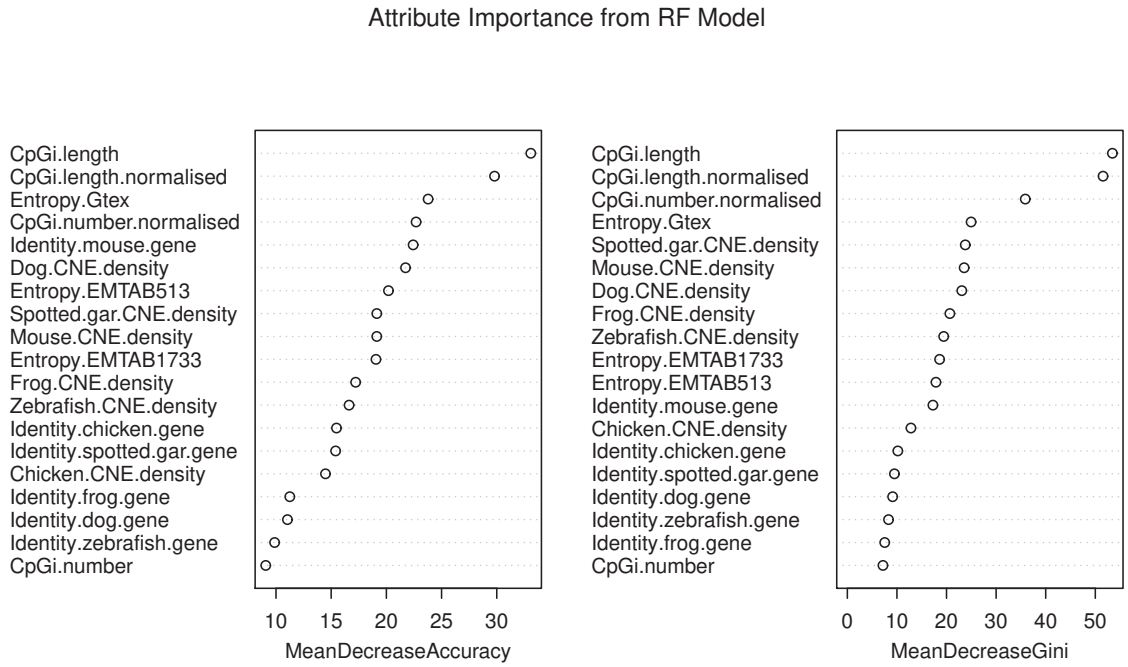

Figure D.2 Attribute importances for RF model. Importances are shown as the mean decrease in accuracy and Gini index. The most important attributes have higher mean decrease in both index values. In both cases, the attributes relating to CpG islands, gene entropy measurements have high predictive importance.

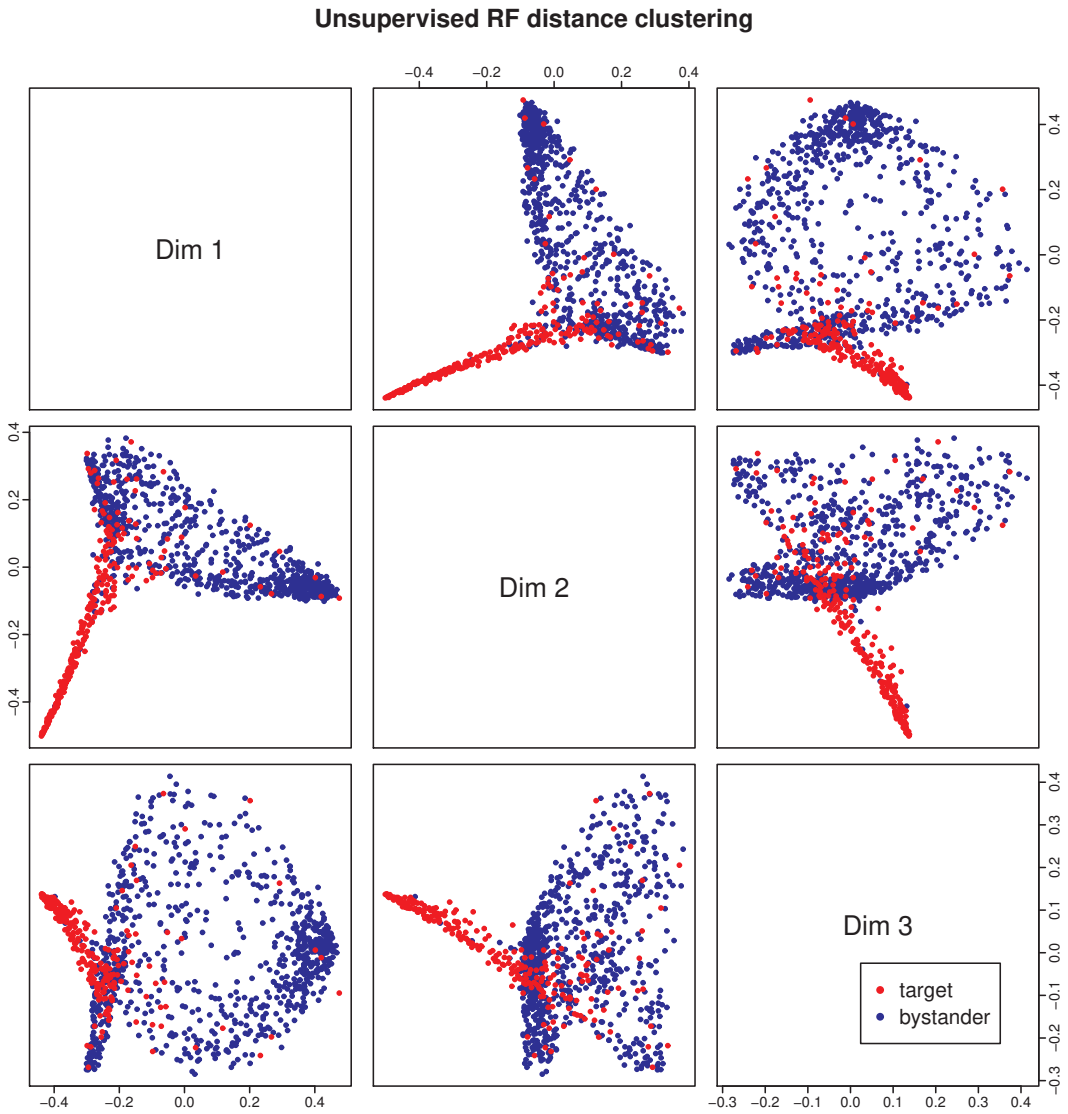

Figure D.3 Multi-dimensional scaling (MDS) of distances from unsupervised random forest model. First three dimensions are shown in the plot. The training targets are depicted with red dots, bystanders with blue dots.

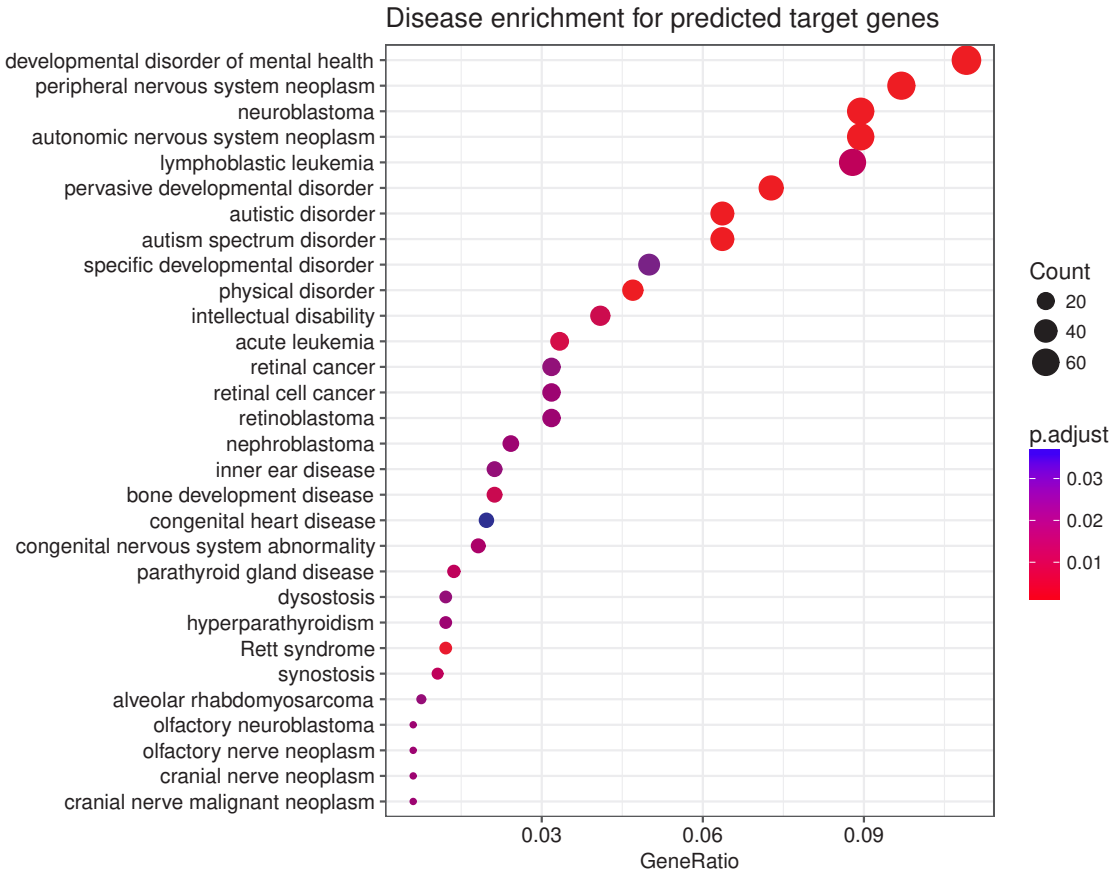

Figure D.4 Over-represented Disease ontology terms for predicted target genes, ranked by GeneRatio.

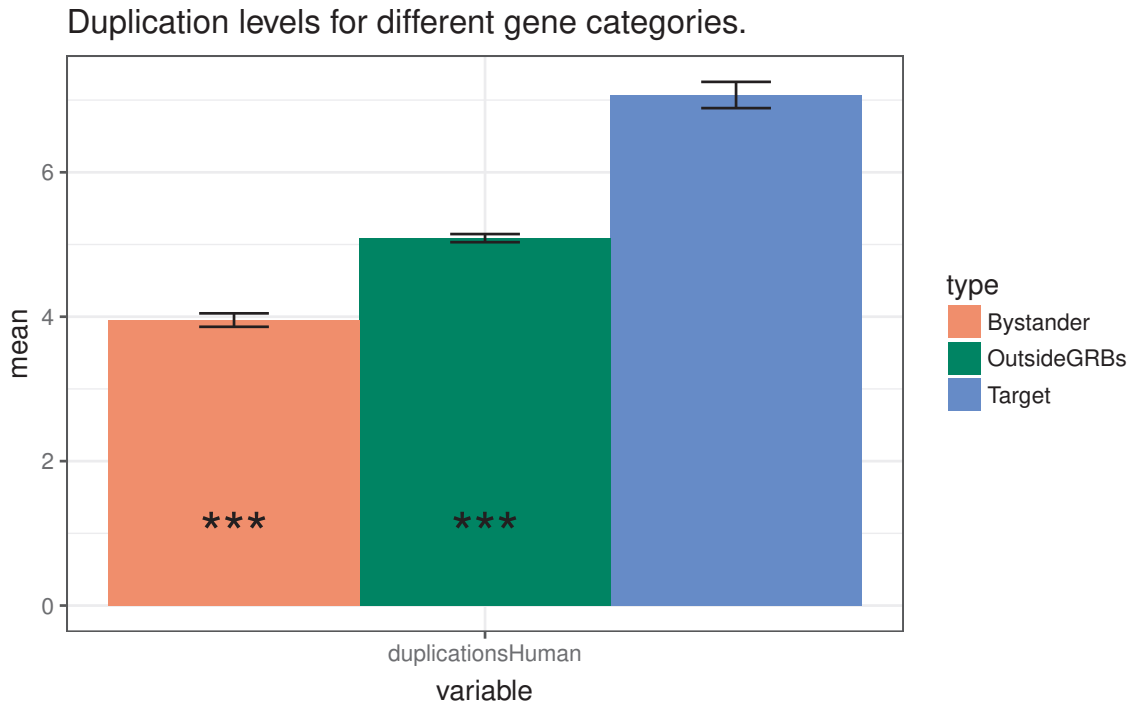

Figure D.5 Duplication levels for predicted target genes, bystander genes and genes outside GRBs. (***: $p < 0.01$, Wilcoxon test, two-sided)

gene	original vote	normalised vote	gene	original vote	normalised vote	gene	original vote	normalised vote
ARX	1.00	1.00	EGR2	0.42	0.65	YPEL5	0.14	1.00
HOXA6	1.00	1.00	ESX1	0.42	1.00	TBCC	0.14	0.81
DLX2	1.00	1.00	SLC12A5	0.42	0.98	DAGLA	0.14	0.62
HOXD11	1.00	1.00	EFNB2	0.42	1.00	COL2A1	0.14	1.00
NKX2-1	1.00	1.00	GFRA1	0.42	1.00	FMN2	0.14	1.00
SIX3	1.00	1.00	CTNNA2	0.42	1.00	ATP2B2	0.14	0.61
NKX6-1	1.00	1.00	CTIF	0.42	1.00	DHRS3	0.14	1.00
IRX5	1.00	1.00	CELF2	0.41	1.00	CDH10	0.13	1.00
HMX2	1.00	1.00	CRYBA2	0.41	0.71	MYL9	0.13	1.00
TLX1	1.00	1.00	LINGO1	0.41	1.00	CORO2B	0.13	1.00
FOXA2	1.00	1.00	ASIC3	0.41	0.74	SEMA7A	0.13	1.00
EVX2	1.00	1.00	CILP2	0.41	1.00	LDB2	0.13	1.00
POU3F3	1.00	1.00	MYOD1	0.41	1.00	DCC	0.13	1.00
ZIC2	1.00	1.00	LPHN3	0.41	1.00	LRRTM3	0.13	1.00
EVX1	1.00	1.00	ACTG1	0.41	1.00	C12orf57	0.13	0.71
LBX1	1.00	1.00	FOXL2	0.40	1.00	ATXN1	0.13	0.81
DLX1	1.00	1.00	MYCN	0.40	1.00	FAM84A	0.13	1.00
FOXG1	1.00	1.00	BAMBI	0.40	1.00	TBC1D22A	0.13	1.00
SOX1	1.00	1.00	BCOR	0.40	1.00	UBE2E1	0.13	1.00
HMX3	1.00	1.00	AGAP3	0.40	0.71	TTC30A	0.13	0.72
DLX5	0.99	1.00	DHH	0.40	0.74	TRHDE	0.13	1.00
HOXA5	0.99	0.99	ZSWIM6	0.39	1.00	NDP	0.13	1.00
OTX1	0.99	1.00	SFN	0.39	1.00	MYL3	0.13	1.00
TFAP2A	0.99	1.00	FSCN2	0.39	0.96	C7orf55	0.13	1.00
FEZF2	0.99	1.00	ID2	0.39	1.00	GJB1	0.13	1.00
GBX2	0.99	1.00	NHLH2	0.39	1.00	CEBPB	0.13	1.00
HOXB9	0.99	1.00	GLIS3	0.38	1.00	SLC16A11	0.13	0.66
PHOX2B	0.99	1.00	SPRY2	0.38	1.00	CADM2	0.13	1.00
HOXB8	0.99	1.00	PPARGC1A	0.38	1.00	CHRM2	0.13	1.00
TBX2	0.99	1.00	EGR4	0.38	0.69	FOXO4	0.13	1.00
SOX21	0.99	1.00	C1QL4	0.38	0.71	C17orf58	0.13	1.00
PITX2	0.99	1.00	KLHL29	0.38	1.00	RP11-599B13.6	0.13	0.66
HOXC5	0.99	1.00	FOXL2NB	0.38	0.94	RP5-850E9.3	0.13	0.82
GATA3	0.99	1.00	PCDH17	0.38	1.00	HSP90AB1	0.12	1.00
SIX2	0.99	0.99	TSSK6	0.38	0.91	KLF3	0.12	1.00
MAB21L2	0.99	1.00	NKX2-5	0.38	1.00	CELF1	0.12	0.68

PAX6	0.99	1.00	MAML3	0.38	1.00	IL13	0.12	1.00
HOXD3	0.99	0.99	KCNC1	0.37	0.91	NKAIN2	0.12	1.00
HOXD8	0.99	0.99	UNC5A	0.37	1.00	FOXI3	0.12	1.00
SOX2	0.99	1.00	PDZRN3	0.37	1.00	AC104057.1	0.12	1.00
HOXC6	0.99	1.00	ZNF618	0.37	1.00	MKNK1	0.12	1.00
HOXA9	0.99	0.99	CDK5	0.37	0.65	RAB11A	0.12	1.00
HOXC11	0.99	0.99	MEX3B	0.36	1.00	NDST4	0.12	1.00
BARHL2	0.99	1.00	RHOV	0.36	1.00	PIANP	0.12	0.65
POU4F2	0.99	1.00	CXXC5	0.36	1.00	NLGN1	0.12	1.00
OTP	0.99	1.00	SMAD7	0.36	0.80	DMBX1	0.12	1.00
POU3F2	0.99	1.00	HYAL2	0.35	1.00	CDC73	0.12	1.00
HOXD9	0.98	0.98	LYPD1	0.35	1.00	DYNC2LI1	0.12	1.00
HOXD10	0.98	0.98	DDN	0.35	0.66	ETV6	0.12	1.00
SOX3	0.98	1.00	COMMD3-BMI1	0.35	1.00	SGCD	0.12	1.00
NKX2-8	0.98	0.98	DUSP6	0.35	1.00	SPERT	0.12	1.00
HOXD12	0.98	0.98	GRM8	0.35	0.71	TMEM256	0.12	0.63
HOXC9	0.98	0.99	AC006486.9	0.35	1.00	RP1-4G17.5	0.12	0.63
HOXA1	0.98	0.98	GNAT1	0.35	0.98	SRCIN1	0.12	1.00
HOXB5	0.98	0.99	PIPOX	0.35	1.00	C17orf100	0.12	1.00
TSHZ3	0.98	1.00	NLK	0.35	1.00	FSTL4	0.12	0.82
NR2F1	0.98	1.00	INHBB	0.35	1.00	AQP2	0.12	0.74
SALL1	0.98	1.00	TCF21	0.34	1.00	WSCD1	0.12	0.98
HOXA3	0.98	0.98	PPP3CA	0.34	1.00	FRAT2	0.12	0.80
PAX2	0.98	0.98	SNCB	0.34	0.92	EXT1	0.12	1.00
HOXC4	0.98	0.99	ZMIZ1	0.34	1.00	MED12	0.12	0.91
PAX9	0.98	0.98	AXIN2	0.34	0.71	TTC36	0.11	1.00
BCL11A	0.98	1.00	XXcos-LUCA11.5	0.34	0.97	ADRBK1	0.11	1.00
MAB21L1	0.98	1.00	WNT3	0.34	1.00	AP1S2	0.11	1.00
HOXB7	0.98	0.98	CTNNB1	0.34	1.00	FLRT2	0.11	1.00
EBF3	0.97	1.00	FGF13	0.34	1.00	CNR1	0.11	1.00
ZIC1	0.97	1.00	CYP26B1	0.34	0.61	BAI2	0.11	1.00
POU3F1	0.97	1.00	DSCAML1	0.34	1.00	ANKRD60	0.11	1.00
HOXC12	0.97	0.98	MYO18A	0.34	0.97	IQCJ-SCHIP1	0.11	1.00
SP8	0.97	1.00	SND1	0.34	0.68	SREBF1	0.11	0.62
FOXD3	0.97	1.00	CDH23	0.33	1.00	PITPNM3	0.11	0.91
NR2E1	0.97	1.00	BAHCC1	0.33	0.80	TNRC6B	0.11	1.00

BARHL1	0.97	1.00	DLX3	0.33	1.00	RAB5B	0.11	0.69
VAX1	0.97	1.00	RASGRP2	0.33	1.00	TBX18	0.11	1.00
PTF1A	0.97	1.00	TENM4	0.33	1.00	STK26	0.11	1.00
EMX2	0.97	1.00	CETN1	0.33	1.00	CAPRIN1	0.11	1.00
SIX6	0.97	1.00	NNAT	0.33	1.00	NLGN3	0.11	0.85
SP9	0.97	1.00	KCNK4	0.33	1.00	ARID1B	0.11	1.00
NKX2-2	0.97	0.97	QKI	0.32	1.00	WNT2	0.11	1.00
BCL11B	0.97	1.00	TBX15	0.32	1.00	PPP2CA	0.11	0.74
FEZF1	0.97	1.00	MIB1	0.32	1.00	DCAF8	0.11	0.75
FOXB1	0.97	1.00	GLRA2	0.32	1.00	TMEM151B	0.11	0.87
HOXA13	0.97	0.97	FGF12	0.32	1.00	WWP2	0.11	1.00
ESRRG	0.97	1.00	EFNA3	0.32	1.00	RP11-508N12.4	0.11	0.69
SIM1	0.96	1.00	NKX2-6	0.32	1.00	TASP1	0.11	0.68
HOXC13	0.96	0.97	CPLX2	0.32	1.00	MID1	0.11	1.00
SHOX2	0.96	1.00	PCDH1	0.32	1.00	OLFML3	0.11	1.00
GSX1	0.96	1.00	FZD5	0.32	1.00	RAP2C	0.11	0.96
FGF8	0.96	0.96	GREM1	0.32	1.00	TSC22D3	0.11	0.65
ONECUT1	0.96	1.00	NTM	0.32	0.71	BMP5	0.10	1.00
ZIC4	0.96	0.99	EPHB1	0.31	1.00	HSD17B12	0.10	1.00
HOXC8	0.96	0.97	KCNA3	0.31	1.00	MPP7	0.10	1.00
FOXP2	0.96	1.00	FOXO6	0.31	1.00	CACNA2D1	0.10	1.00
HOXD13	0.96	0.96	HAND1	0.31	1.00	GRIN2A	0.10	1.00
ZIC3	0.96	1.00	NPBWR2	0.31	1.00	FAM78B	0.10	1.00
IRX6	0.96	0.96	C1QL1	0.31	1.00	GNAS	0.10	0.91
UNCX	0.96	1.00	SPRY1	0.31	1.00	PLS3	0.10	1.00
OTX2	0.96	1.00	FGF9	0.31	1.00	BCL9	0.10	1.00
SOX14	0.96	1.00	OLFM3	0.30	1.00	KCNQ4	0.10	1.00
HOXA7	0.96	0.96	CITED2	0.30	1.00	FILIP1	0.10	1.00
HOXD1	0.96	0.96	LKAAEAR1	0.30	0.97	HIPK1	0.10	0.88
MEIS1	0.96	1.00	HDAC2	0.30	1.00	MSANTD4	0.10	1.00
TLX3	0.96	1.00	RP11-540D14.8	0.30	0.94	FAM64A	0.10	0.82
TFAP2B	0.95	1.00	MSI2	0.30	1.00	TAOK3	0.10	1.00
TBX4	0.95	0.96	ENHO	0.30	1.00	MKX	0.10	0.95
SIX1	0.95	0.98	MPPED1	0.30	1.00	GJA1	0.10	1.00
PAX3	0.95	1.00	PLXNA4	0.30	1.00	TNNI1	0.10	1.00
ZIC5	0.95	0.96	HNF1B	0.30	1.00	MYH7	0.10	1.00
IRX1	0.95	1.00	GPC4	0.30	1.00	BBX	0.10	1.00
ZNF536	0.95	0.97	CACNA2D2	0.29	0.83	CAPN13	0.10	1.00

HOXA10	0.95	0.95	SEMA3B	0.29	0.83	PLAG1	0.10	1.00
ZNF503	0.95	1.00	ALX4	0.29	1.00	RPL37A	0.10	1.00
ZEB2	0.95	1.00	NRXN2	0.29	0.87	CLCN4	0.10	0.90
NR2F2	0.95	1.00	GPR85	0.29	1.00	CRYAB	0.10	1.00
NR4A2	0.95	1.00	NRXN1	0.29	1.00	SKP1	0.10	0.64
ID4	0.95	1.00	DPYD	0.29	1.00	ANKRD17	0.10	1.00
HOXB6	0.95	0.95	PRR23C	0.29	0.73	ADAMTSL1	0.10	1.00
SOX6	0.95	1.00	HS3ST5	0.29	0.97	DPF3	0.10	1.00
NKX6-2	0.95	1.00	PHF12	0.29	0.84	TMEM255A	0.09	1.00
GSX2	0.95	1.00	RUNX3	0.29	1.00	TMEM63B	0.09	0.75
DBX1	0.95	1.00	CLTC	0.29	1.00	CSNK1G1	0.09	1.00
PROX1	0.95	1.00	ALDH1A2	0.29	1.00	BNC1	0.09	1.00
MAFA	0.95	1.00	ZNRF4	0.29	1.00	ISLR	0.09	0.69
CASZ1	0.94	1.00	HS3ST2	0.29	1.00	IL2RG	0.09	0.68
INSM1	0.94	0.95	FTHL17	0.28	1.00	SCN3A	0.09	0.62
HOXA11	0.94	0.94	NRG2	0.28	0.79	TMEM178B	0.09	1.00
PRDM6	0.94	1.00	PRMT8	0.28	1.00	TTC17	0.09	0.86
NPAS3	0.94	1.00	GSE1	0.28	1.00	CDH6	0.09	1.00
LMO1	0.94	1.00	FKBP2	0.28	0.85	ZFP36L2	0.09	0.69
BHLHE22	0.94	1.00	CHRM4	0.28	1.00	KALRN	0.09	1.00
HOXB1	0.94	0.94	TEAD1	0.28	1.00	HS3ST4	0.09	1.00
IRX3	0.94	0.94	NRP1	0.28	1.00	SHISA9	0.09	1.00
PDX1	0.94	0.97	CXCR4	0.28	1.00	ATP1B4	0.09	0.93
HOXD4	0.94	0.94	CNTFR	0.28	0.93	SYT16	0.09	1.00
SP5	0.94	1.00	ASTN2	0.28	1.00	YPEL4	0.09	1.00
LHX2	0.93	1.00	ROBO2	0.28	1.00	KDM2A	0.09	0.77
PBX3	0.93	1.00	DNM1	0.28	1.00	RTN4RL2	0.09	1.00
TOX	0.93	1.00	B3GALT1	0.27	1.00	TRIB2	0.09	1.00
HOXB3	0.93	0.93	EGR1	0.27	1.00	GALNT18	0.09	1.00
HOXC10	0.93	0.94	GRIK4	0.27	1.00	NEDD1	0.09	1.00
HOXA10-HOXA9	0.93	0.93	PRR23B	0.27	0.68	FAM49B	0.09	1.00
PAX7	0.93	1.00	TEX40	0.27	0.82	EDIL3	0.09	1.00
ZBTB16	0.92	1.00	DLX4	0.27	0.81	CXorf65	0.09	0.62
FOXA1	0.92	0.92	CITED1	0.27	1.00	ZBTB10	0.09	1.00
MNX1	0.92	1.00	SLITRK2	0.27	1.00	LAMP2	0.08	0.88
OLIG3	0.92	1.00	DIO3	0.27	1.00	PHACTR1	0.08	1.00
GSC	0.92	1.00	SPOP	0.27	0.80	NPR3	0.08	1.00
PAX5	0.92	1.00	SST	0.27	1.00	ZNF385B	0.08	1.00
LMX1B	0.92	0.98	CTXN2	0.27	0.61	GRIA1	0.08	1.00

HOXA4	0.92	0.91	MDGA1	0.26	1.00	MEGF11	0.08	0.67
OLIG2	0.92	1.00	PRRX1	0.26	1.00	MACROD2	0.08	1.00
LHX3	0.91	1.00	NTRK3	0.26	1.00	LUC7L2	0.08	0.63
MECOM	0.91	1.00	ENC1	0.26	1.00	GLRB	0.08	1.00
NKX2-4	0.91	0.91	IER5L	0.26	1.00	DIAPH2	0.08	1.00
ASCL1	0.91	1.00	GRIK5	0.26	0.72	PRKCB	0.08	1.00
HELT	0.91	1.00	RND3	0.26	1.00	RBPJ	0.08	1.00
SKOR1	0.91	1.00	CDK5R1	0.26	1.00	HSPB2	0.08	0.83
TBR1	0.90	1.00	PRICKLE1	0.26	1.00	CLP1	0.08	0.91
PITX3	0.90	0.90	ARHGEF1	0.26	0.71	PTPRCAP	0.08	0.70
HHEX	0.90	1.00	MSX2	0.26	1.00	PCDH18	0.08	1.00
DLX6	0.90	0.91	GDF5OS	0.26	1.00	PTCHD4	0.08	1.00
NOG	0.90	1.00	STK3	0.25	1.00	SEMA3A	0.08	1.00
BSX	0.90	1.00	RPS23	0.25	1.00	CDKN1B	0.08	1.00
DMRTA2	0.90	1.00	HIVEP2	0.25	1.00	GRIK1	0.08	1.00
NR5A2	0.90	1.00	HMGA2	0.25	1.00	GPR139	0.08	1.00
NEUROD2	0.90	1.00	DOCK1	0.25	1.00	RS1	0.07	1.00
PRDM16	0.89	1.00	RCOR2	0.25	0.76	ELF2	0.07	0.94
EN2	0.89	0.96	OTUD6A	0.25	1.00	CNOT2	0.07	1.00
PTCH1	0.89	1.00	RABAC1	0.25	0.68	GIN1	0.07	1.00
IRX4	0.89	0.93	UNC5C	0.25	1.00	KCNV1	0.07	1.00
HOXB2	0.89	0.89	POU2F1	0.25	1.00	IGDCC3	0.07	1.00
LHX1	0.89	1.00	CDH2	0.25	1.00	TMEM167A	0.07	0.83
SHH	0.88	0.96	KCND3	0.25	1.00	ZNF609	0.07	0.76
IRX2	0.88	0.92	IKZF2	0.24	1.00	C11orf31	0.07	0.82
SATB2	0.88	1.00	NGFR	0.24	0.73	FOSL2	0.07	1.00
LMO4	0.88	1.00	CALM2	0.24	1.00	FGF20	0.07	1.00
EN1	0.88	1.00	DEDD2	0.24	0.66	FLRT3	0.07	0.84
PITX1	0.88	1.00	GRIK2	0.24	1.00	NEGR1	0.07	1.00
CUX2	0.88	1.00	SMOC1	0.24	1.00	CARNS1	0.07	0.61
ISL2	0.87	1.00	KIRREL3	0.24	1.00	MAGEB10	0.07	1.00
ZFHX3	0.87	1.00	DCAF12L1	0.24	1.00	RIMS1	0.07	1.00
AC009336.19	0.87	0.87	TRIM29	0.24	0.88	RAPGEF2	0.07	1.00
ZFPM2	0.86	1.00	TNNC1	0.24	1.00	DOCK4	0.07	1.00
LHX5	0.86	1.00	SORCS3	0.24	1.00	BTG1	0.07	1.00
FOXP1	0.86	1.00	OSR2	0.24	0.94	HS3ST3A1	0.07	1.00
NR5A1	0.86	0.92	RGMA	0.24	1.00	EPHA6	0.07	1.00
VSX2	0.86	1.00	BCORL1	0.24	1.00	PRMT5	0.07	1.00
FZD2	0.86	1.00	TUSC2	0.24	0.66	TMPO	0.07	0.69
ATOH1	0.85	1.00	OPRL1	0.24	0.76	TCP11	0.07	1.00

TBX3	0.85	0.99	KCNS2	0.24	0.93	TIMM10	0.07	0.75
SATB1	0.85	1.00	TSPAN5	0.24	1.00	BAI3	0.07	1.00
HOXB4	0.85	0.86	BRINP1	0.23	0.84	SFRP2	0.07	1.00
MEIS2	0.85	1.00	STAG2	0.23	1.00	SCUBE3	0.07	1.00
NFIA	0.85	1.00	MAST4	0.23	1.00	ANAPC10	0.07	1.00
JAG1	0.85	1.00	ID3	0.23	1.00	MRPL33	0.07	0.94
RFX4	0.85	1.00	MYB	0.23	1.00	PTHLH	0.06	1.00
EBF1	0.85	1.00	MSI1	0.23	1.00	NAV1	0.06	0.64
TFAP2D	0.84	0.88	VWC2L	0.23	1.00	CHMP2B	0.06	1.00
GLI3	0.84	1.00	ROMO1	0.23	0.89	IL17RD	0.06	1.00
TLE3	0.84	1.00	CACNA1E	0.23	1.00	ARHGEF3	0.06	1.00
TSHZ1	0.84	1.00	ATP1A3	0.23	0.61	ST6GALNAC6	0.06	1.00
TOX3	0.84	0.85	PVRL1	0.23	0.84	SUPT4H1	0.06	1.00
MYF6	0.84	1.00	EPHB2	0.23	0.98	EPHA3	0.06	1.00
BNC2	0.83	1.00	CASK	0.23	1.00	HACE1	0.06	1.00
FOXF1	0.83	1.00	EOMES	0.23	1.00	EPHA5	0.06	1.00
ONECUT2	0.83	1.00	ATOH7	0.23	1.00	PIK3R1	0.06	1.00
HLX	0.83	1.00	PRMT6	0.23	1.00	C9orf3	0.06	1.00
RUNX1T1	0.83	1.00	WNT11	0.23	1.00	ATP2A2	0.06	1.00
SMAD6	0.83	0.91	EIF3E	0.23	1.00	SYN3	0.06	1.00
PRDM8	0.83	1.00	FGF1	0.23	0.69	RP11-298I3.5	0.06	0.91
ZNF703	0.82	1.00	XKR6	0.23	1.00	BARX2	0.06	1.00
PSD	0.82	0.82	FAT4	0.23	1.00	NTRK2	0.06	1.00
WNT5A	0.82	1.00	MAGI1	0.22	1.00	NCAM1	0.06	1.00
CDX2	0.82	0.85	SEMA6A	0.22	1.00	CT62	0.06	1.00
ZFHX4	0.82	1.00	DMRTB1	0.22	1.00	CKMT1B	0.06	1.00
TWIST1	0.81	1.00	GLIS1	0.22	1.00	DAAM1	0.06	1.00
HOXA2	0.81	0.81	GRID1	0.22	1.00	P2RX3	0.06	0.64
FZD10	0.81	1.00	FAM19A1	0.22	0.98	RBBP5	0.06	1.00
FOXP4	0.81	1.00	WNT7B	0.22	1.00	ATXN7L1	0.06	1.00
NFIX	0.80	1.00	HSPB1	0.22	1.00	GDNF	0.06	1.00
HES1	0.80	1.00	CYR61	0.22	1.00	CD5L	0.05	1.00
ZNF521	0.80	1.00	BMI1	0.22	0.61	REM2	0.05	0.82
SOX5	0.80	1.00	DUSP21	0.22	1.00	CAMK2D	0.05	1.00
POU6F2	0.80	1.00	KCND2	0.22	1.00	JRKL	0.05	1.00
NEUROD6	0.80	1.00	ALX1	0.21	1.00	TAF2	0.05	1.00
ISL1	0.80	1.00	HS6ST3	0.21	1.00	JADE1	0.05	1.00
PCDH8	0.80	1.00	RELN	0.21	1.00	LGR6	0.05	1.00
WT1	0.80	0.81	PTPRT	0.21	1.00	BAHD1	0.05	1.00

PAX1	0.79	0.80	GRM4	0.21	1.00	CDH13	0.05	1.00
OSR1	0.79	1.00	ID1	0.21	1.00	CCDC82	0.05	0.93
KLHL14	0.79	1.00	CACNG2	0.21	1.00	TMEM178A	0.05	1.00
RP11-834C11.12	0.79	0.80	KLHL34	0.21	1.00	FGFR2	0.05	1.00
FZD1	0.79	1.00	SMC1A	0.21	1.00	NEDD4	0.05	1.00
POU4F3	0.78	1.00	CREB1	0.21	0.66	NEIL3	0.05	1.00
LHX6	0.78	1.00	RORB	0.21	1.00	ACVR1	0.05	1.00
SMAD2	0.78	1.00	PRSS56	0.21	1.00	COL8A1	0.05	1.00
NKX3-2	0.78	1.00	TRPS1	0.21	1.00	HSF5	0.05	0.77
HIC1	0.78	1.00	KCNA1	0.21	1.00	UBE2T	0.05	0.92
PBX1	0.78	1.00	IL1RAPL1	0.21	1.00	MEDAG	0.05	1.00
MAF	0.77	1.00	KCNA2	0.21	0.62	LGR5	0.05	1.00
SALL3	0.77	1.00	PHLDA2	0.21	1.00	NDNF	0.05	1.00
FIGN	0.77	1.00	LIN7C	0.21	1.00	SIRPA	0.05	1.00
PRDM12	0.77	1.00	LHX8	0.21	1.00	AJUBA	0.05	0.70
SIX4	0.77	0.79	LRRC10B	0.21	1.00	PTPRR	0.05	0.62
GATA2	0.76	1.00	CAPN6	0.20	1.00	LRFN2	0.05	1.00
LHX9	0.76	1.00	DMC1	0.20	1.00	BRE	0.05	0.64
PAX8	0.76	1.00	CNTN4	0.20	1.00	FAM46A	0.04	1.00
RAX	0.76	1.00	NR0B1	0.20	0.98	MEF2D	0.04	1.00
TCF7L2	0.76	1.00	TRIAP1	0.20	0.88	KNSTRN	0.04	0.69
BMP4	0.75	1.00	SLC17A6	0.20	1.00	PI15	0.04	1.00
FOXF2	0.75	1.00	BMP2	0.20	1.00	PKP4	0.04	1.00
SETBP1	0.75	1.00	SAMD5	0.20	1.00	TMPRSS15	0.04	1.00
RBFOX1	0.75	1.00	ADAMTS6	0.20	1.00	ATP1A1	0.04	1.00
CBLN1	0.75	1.00	PCDH10	0.20	1.00	TEX26	0.04	0.88
ETV1	0.75	1.00	OTUB1	0.20	0.61	C4orf22	0.04	1.00
EYA1	0.74	1.00	SPAG4	0.20	0.77	CAPZB	0.04	1.00
AUTS2	0.74	1.00	GRIA3	0.20	1.00	DNAJB9	0.04	1.00
POU3F4	0.74	1.00	NHS	0.20	1.00	DLC1	0.04	1.00
IKZF5	0.74	0.72	MDK	0.20	0.70	CENPW	0.04	1.00
POU4F1	0.73	1.00	CTB-55O6.8	0.20	1.00	FGFR1	0.04	1.00
PRDM13	0.73	1.00	ALPP	0.20	0.93	NRCAM	0.04	0.75
FOXC2	0.73	0.87	ECEL1	0.20	0.93	NPY	0.04	1.00
FOXN3	0.72	1.00	PCDH9	0.20	1.00	ARHGAP40	0.04	1.00
SLC18A3	0.72	1.00	VPS13B	0.19	0.76	PNPLA8	0.04	0.75
FOXB2	0.72	1.00	PLXNC1	0.19	1.00	PHF2	0.04	1.00
SFTA3	0.72	0.72	ARHGAP36	0.19	1.00	EYA2	0.04	1.00

DACH2	0.72	1.00	TENM1	0.19	0.81	DMGDH	0.04	1.00
CDCA7	0.72	0.74	CDKN1C	0.19	0.91	PTGFRN	0.04	0.86
DMAP1	0.72	1.00	GABARAP	0.19	1.00	ARL8A	0.04	0.73
NKX1-2	0.72	1.00	NRXN3	0.19	1.00	WHSC1L1	0.04	0.94
SOX4	0.72	1.00	CLDN7	0.19	0.99	PTPRJ	0.04	1.00
TCF12	0.72	1.00	RTP2	0.19	0.71	FOXO1	0.04	1.00
INSM2	0.72	0.71	MLLT3	0.19	1.00	ALCAM	0.04	1.00
EBF2	0.72	1.00	NOL4	0.19	1.00	CLDN11	0.04	1.00
CYP26A1	0.71	0.78	ACRBP	0.19	1.00	PHIP	0.04	1.00
NEUROD1	0.71	1.00	KIAA1161	0.19	0.62	RALGAPB	0.04	0.82
TENM3	0.71	1.00	SRGAP3	0.19	1.00	HAS2	0.04	1.00
NOVA1	0.71	0.71	NOL4L	0.19	1.00	LRRTM4	0.04	1.00
SMAD3	0.71	0.78	SYT1	0.18	1.00	SERTAD2	0.04	1.00
KCTD15	0.71	1.00	MEOX2	0.18	1.00	LHFPL1	0.04	1.00
RP11-546B8.6	0.71	0.74	MXI1	0.18	1.00	AJAP1	0.04	1.00
TLE4	0.70	1.00	FAM219A	0.18	0.61	SSR3	0.03	1.00
ESRRB	0.70	1.00	YWHAG	0.18	0.84	TNR	0.03	1.00
ST18	0.70	1.00	TRAPPC1	0.18	0.96	SPOCK1	0.03	1.00
MAFB	0.70	1.00	MAMLD1	0.18	1.00	AC022431.2	0.03	1.00
NKX1-1	0.69	1.00	KHDRBS3	0.18	1.00	MED23	0.03	1.00
LMO3	0.69	1.00	RBMS3	0.18	0.78	SGOL1	0.03	1.00
OLIG1	0.69	0.75	FGF16	0.18	1.00	RGS8	0.03	1.00
FOXC1	0.69	0.91	SLC6A1	0.18	1.00	GPR112	0.03	1.00
ZNF423	0.69	0.92	G0S2	0.18	1.00	TIPARP	0.03	0.94
SOBP	0.69	1.00	PTPRK	0.18	1.00	MMD	0.03	1.00
BAZ2B	0.69	1.00	PACSIN3	0.18	1.00	CDC42EP3	0.03	1.00
ZNF827	0.69	0.70	ALOXE3	0.18	0.93	FSHR	0.03	1.00
DRGX	0.69	0.95	FGF23	0.18	0.61	XYLT1	0.03	1.00
CELF4	0.68	1.00	CAMK2A	0.17	1.00	IRS1	0.03	1.00
TSHZ2	0.68	1.00	HNRNPH3	0.17	0.76	ZNF438	0.03	1.00
NCOA2	0.68	1.00	CD276	0.17	1.00	COL4A6	0.03	1.00
EFNA5	0.68	1.00	KCTD10	0.17	1.00	PEG10	0.03	1.00
FOXD2	0.68	1.00	ASTN1	0.17	1.00	FYN	0.03	1.00
FGF10	0.67	1.00	PKDCC	0.17	1.00	AEBP2	0.03	1.00
NEUROG2	0.67	1.00	SOST	0.17	1.00	FAM160B1	0.03	1.00
MLLT10	0.67	1.00	RYR2	0.17	1.00	CAST	0.03	1.00
FOXD1	0.67	1.00	YBX2	0.17	0.90	SULF2	0.03	0.65
MSX1	0.67	1.00	PPARG	0.17	1.00	FHL1	0.02	0.75
SP3	0.67	0.69	CDK5RAP2	0.17	0.62	SH3PXD2A	0.02	1.00

PRICKLE2	0.67	1.00	GLRA1	0.17	1.00	RFTN1	0.02	1.00
DICER1	0.66	0.71	SERP2	0.17	1.00	RAPGEF5	0.02	1.00
GAS1	0.66	1.00	CUX1	0.17	1.00	STEAP2	0.02	1.00
TCF4	0.66	0.79	DGKG	0.17	1.00	RP11-310N16.1	0.02	0.80
ZNF644	0.65	0.64	SERTAD4	0.17	1.00	CBR4	0.02	1.00
ZBTB18	0.65	1.00	GFRA2	0.17	1.00	C11orf74	0.02	1.00
FGF5	0.65	0.78	SLITRK1	0.17	1.00	GBE1	0.02	1.00
SMAD1	0.64	0.65	HES7	0.17	0.89	PMEPA1	0.02	1.00
ARID5B	0.64	1.00	IRF2BP2	0.17	1.00	PIGA	0.02	1.00
ZNF507	0.64	0.65	CBLN4	0.17	1.00	DIRC3	0.02	1.00
FOXL1	0.64	0.76	ATG5	0.17	1.00	DNAH11	0.02	0.70
NRIP1	0.63	1.00	NCKAP1	0.17	1.00	ATP10B	0.02	1.00
LHX4	0.63	1.00	GNMT	0.17	1.00	EHF	0.02	1.00
PIAS1	0.63	0.69	PRPS1	0.16	1.00	ASB11	0.02	0.90
CHD7	0.63	0.60	KCNJ4	0.16	0.80	OTOL1	0.02	1.00
ATF2	0.62	0.62	CXXC4	0.16	1.00	TLL2	0.02	1.00
OVCA2	0.62	0.80	ZNF608	0.16	1.00	SORBS1	0.02	1.00
HAND2	0.62	1.00	KCNQ1	0.16	0.74	HIVEP1	0.02	1.00
SOX9	0.61	1.00	RAI1	0.16	1.00	NEURL1	0.02	0.64
STAT6	0.61	1.00	ENOX1	0.16	1.00	MAP1B	0.02	1.00
EGR3	0.61	1.00	ADM	0.16	1.00	TSPAN2	0.02	1.00
CHST8	0.60	0.85	NKX6-3	0.16	1.00	KIF2B	0.02	1.00
RUNX2	0.60	1.00	SFTPC	0.16	0.95	TERF1	0.02	1.00
DMRT2	0.60	1.00	LINGO2	0.16	1.00	LRR1	0.02	1.00
NEUROG1	0.59	0.67	ACTL10	0.16	1.00	CMKLR1	0.02	1.00
FZD8	0.59	1.00	C6orf226	0.16	0.98	RPS29	0.02	1.00
FLI1	0.59	1.00	C17orf61-PLSCR3	0.16	0.84	ADAM22	0.01	1.00
LPHN2	0.58	1.00	COPS7A	0.16	0.86	TM9SF3	0.01	0.86
NFIB	0.58	1.00	AQP5	0.16	1.00	CDH9	0.01	1.00
EPHA7	0.58	1.00	TTC30B	0.16	1.00	GADL1	0.01	1.00
IHH	0.58	1.00	RPS26	0.16	1.00	TMEM207	0.01	1.00
LYL1	0.57	0.72	TMEM117	0.16	1.00	SMAP1	0.01	1.00
GDF6	0.57	1.00	ISM1	0.16	1.00	SLC10A2	0.01	1.00
ZBTB20	0.57	1.00	KLF4	0.16	1.00	ADAMTS19	0.01	1.00
SIM2	0.56	1.00	FJX1	0.16	1.00	RGS21	0.01	1.00
FOXE3	0.56	0.82	CDC42	0.15	0.66	OGFRL1	0.01	0.75
GBX1	0.55	1.00	RBM24	0.15	1.00	FOXN2	0.01	1.00
CAMTA1	0.55	1.00	KCNA5	0.15	0.74	CHRDL2	0.01	1.00

IKZF1	0.55	1.00	NIPBL	0.15	1.00	POLD3	0.01	1.00
EMX1	0.54	1.00	SRGAP2C	0.15	1.00	MAN2A1	0.01	1.00
NXPH1	0.54	1.00	TMEM132E	0.15	1.00	PVRL3	0.01	1.00
WNT1	0.54	1.00	C19orf67	0.15	0.79	PABPC4L	0.01	1.00
MN1	0.53	1.00	SRGAP2	0.15	1.00	MAN1A1	0.01	1.00
GMDS	0.53	0.70	CDH20	0.15	1.00	IL21	0.01	1.00
AKT3	0.53	0.77	NRN1	0.15	1.00	NRSN1	0.01	1.00
FAM181B	0.52	1.00	EIF2B5	0.15	1.00	CADM3	0.01	1.00
DAB1	0.52	0.61	MAGEF1	0.15	1.00	PRELID2	0.01	1.00
CADM1	0.52	1.00	SCRT2	0.15	1.00	SLC39A10	0.01	1.00
COX8C	0.52	1.00	MID2	0.15	0.91	RP11-166N6.3	0.01	1.00
ELAVL2	0.52	1.00	KHDRBS2	0.15	1.00	ARHGAP44	0.00	1.00
TBX1	0.52	1.00	HNRNPR	0.15	0.64	LMCD1	0.00	1.00
RUNX1	0.51	1.00	MLF2	0.15	0.79	RP11-58C22.1	0.00	1.00
HEY1	0.51	1.00	FAM222A	0.15	0.84	AC108925.1	0.00	1.00
LEF1	0.50	1.00	AFF2	0.15	1.00	AHR	0.00	1.00
FZD7	0.50	1.00	ALPPL2	0.15	0.70	TYRP1	0.00	1.00
FEV	0.50	0.87	FAT3	0.15	1.00	SMC2	0.00	1.00
RBFOX3	0.50	1.00	BCL2L1	0.15	0.70	C3orf56	0.00	1.00
MSC	0.50	0.67	NAA38	0.15	0.77	AC104981.1	0.00	1.00
SSBP3	0.50	1.00	LCOR	0.15	1.00			
PHOX2A	0.49	1.00	SRGAP2B	0.15	1.00			
RTN4RL1	0.49	0.63	MAGEB1	0.15	0.71			
DACH1	0.49	1.00	HMGB3	0.15	0.79			
LRRC4	0.49	1.00	DYNLL1	0.15	0.62			
TLE1	0.48	0.68	SCN2A	0.15	1.00			
NR4A3	0.48	1.00	ZNF462	0.15	0.94			
NKX2-3	0.48	1.00	DKK2	0.15	1.00			
FAM53A	0.48	0.68	KCNJ3	0.15	1.00			
GRIK3	0.48	1.00	GABRB1	0.15	1.00			
PCDH7	0.48	1.00	BDNF	0.15	0.70			
CTD-2535L24.2	0.48	1.00	CSNK1E	0.15	0.71			
VGLL2	0.47	1.00	ATP1A2	0.14	1.00			
PTPRD	0.47	1.00	ASAP3	0.14	0.62			
CASC10	0.47	0.70	EFEMP1	0.14	1.00			
NTN1	0.47	1.00	TMEM160	0.14	1.00			
CDK5R2	0.46	0.80	MBNL1	0.14	1.00			

MEF2C	0.46	1.00	EPHB3	0.14	0.94
CA10	0.46	1.00	PLCB4	0.14	1.00
CHST2	0.46	1.00	RAB33A	0.14	0.60
DMRT3	0.45	0.60	TMEM88	0.14	0.74
PCDH19	0.45	1.00	FAM168A	0.14	1.00
SMARCD3	0.45	0.80	TMEFF2	0.14	1.00
AHDC1	0.45	1.00	EIF4A1	0.14	0.73
GFI1	0.45	1.00	FRAT1	0.14	0.95
OPCML	0.45	1.00	Mar.03	0.14	1.00
RARB	0.44	1.00	ZFP36L1	0.14	1.00
TENM2	0.44	1.00	VDAC1	0.14	1.00
SOX17	0.44	1.00	RP11-542C16.2	0.14	0.73
PKNOX2	0.44	1.00	FBXW7	0.14	1.00
NRP2	0.43	1.00	BHLHE40	0.14	1.00
SEMA6D	0.43	1.00	LRP4	0.14	0.76
CDH11	0.43	1.00	TRIB1	0.14	1.00
CDH22	0.43	1.00	LPAR5	0.14	0.74
C1QL3	0.43	1.00	SCN8A	0.14	1.00
FST	0.43	1.00	CPEB4	0.14	1.00

Table D.1 A list of 1161 predicted target genes with original vote from random forests model and normalised vote. The gene with low original vote is considered less convincing even though the normalised vote is 1, when this gene is the only gene with vote value within that GRB.