Imperial College London

Institute of Clinical Sciences

The Evolutionary Dynamics of Genomic Regulatory Blocks in Metazoan Genomes

Alexander Jolyon Nash

July 2018

Submitted in part fulfilment of the requirements for the degree of

Doctor of Philosophy of Imperial College London

Declaration

I hereby declare that this submission is my own work including all the analyses performed. To the best of my knowledge it contains no material previously published or written by another person nor material which has been accepted for any other degree of any university or other institute of higher learning, except where due acknowledgement is made in the text.

Abstract

Developmental genes require intricate control of the timing, location and magnitude of their expression. This is provided by multiple evolutionarily conserved enhancers, known as conserved non-coding elements (CNEs). CNEs cluster around their target genes, forming long syntenic arrays known as genomic regulatory blocks (GRBs). Current methods for GRB identification rely on the selection of arbitrary minimum conservation thresholds, impeding their performance in many contexts. In this thesis, I propose a novel measure of pairwise genome conservation that eliminates the need for conservation thresholds, and use this measure to study the evolutionary dynamics of GRBs in metazoa. I define sets of GRBs based on their rate of regulatory turnover – high turnover GRBs (htGRBs) and low turnover GRBs (ltGRBs) – in three independent metazoan lineages. I show that ht- and ltGRBs target functionally distinct classes of genes, and that these genes tend to be expressed during late and early development respectively, potentially contributing to their differing tolerance of regulatory turnover. Moreover, the differences between ht- and ltGRBs are consistent across all three lineages, suggesting that similar evolutionary pressures have defined the rate of turnover in these GRBs since their emergence in the metazoan ancestor. Next I identify GRBs in the extremely compact *Caenorhabditis elegans* and *Oikopleura dioica* genomes for the first time, and use these GRBs to investigate the effects of genome compaction on GRB size and composition. I show that GRB size scales proportionally with genome size and that GRBs exhibit similar enrichment and depletion of specific genomic features.

This suggests that regardless of background genome content, GRBs are under similar pressure to maintain a permissive environment for long-range gene regulation. The development of a threshold-free GRB identification method has facilitated the analysis of GRBs in both closely related species and compact genomes, providing further insights into their origin and evolution.

Acknowledgements

Firstly, I would like to thank Boris for giving me the opportunity to do this PhD. Your passion for your research is infectious and it has been an absolute pleasure to be a part of the Lenhard group. Thank you for your guidance and for always being available to discuss my ideas - no matter how daft or outlandish!

So much of this experience has been shaped by the amazing members of the Lenhard group, and I am extremely grateful to have worked alongside all of you. Nathan, thank you for putting up with my naivety and introducing me to the highs and lows of GRB analysis. Anja, thank you for always looking out for me (and the rest of the lab) throughout my PhD - you were definitely our lab mom. Ge, thank you for all the technical guidance and great CNE discussions. Malcolm and Liz, you set the PhD student bar high and forced me to improve. Thank you for the many interesting discussions and becoming good friends. Dunja, you were the best deskmate and you are the undisputed Lenhard group bake-off winner! Its been great getting to know you and I truly value our friendship. Piotr, thank you for teaching me so much about servers, sysadmin, slurm and how not to kill mamut! Dimitris, your knowledge of the CNE literature is astounding, it was great fun to write the review with you. Nevena, your drive is inspirational, thank you for being a great role model. And finally to everyone who joined me for my last few months, Damir, Nejc, Eleni and Sebastian, you have been a great addition to the group, thank you for all the laughs!

Leonie, thank you for being so supportive, positive and selfless. Your constant motivation and mentorship, particularly during the write up, made this process possible and I am immensely grateful. You are the best of the best.

James and Sasha, you guys have made the last four years so damn fun. I couldn't have asked for more ridiculous conversations or late night gaming sessions. Keep being the krabs you were born to be.

Finally, to my family, you have always stood behind me and supported me in the pursuit of my interests. Even if that meant encouraging me to move across the world. Thank you for your unconditional love and support, it means everything to me.

Contents

List of Figures

List of Tables

Abbreviations

3’ three prime, the asymmetric end of a DNA strand possessing a terminal hydroxyl group . . . 19

3D three dimensional . . . 19

5’ five prime, the asymmetric end of a DNA strand possessing a terminal phosphate group . . . 19

bp base pairs . . . 20

BP biological process . . . 85

CAGE cap analysis of gene expression . . . 20

CGI CpG island . . . 20

ChIP-seq chromatin immunoprecipitation sequencing . . . 28

ChromHMM chromatin state segmentation by hidden markov model . . . 28

CNE conserved non-coding element . . . 36

CNV copy number variation . . . 41

CpG cytosine followed by a guanine in 5’ to 3’ direction . . . 20

CPM change point modelling . . . 59

DNA deoxyribonucleic acid . . . 19

ENCODE	encyclopedia of DNA elements	28
eRNA	enhancer RNA	26
FDR	false discovery rate	118
GO	gene ontology	85
GRB	genomic regulatory block	45
hESC	embryonic stem cells	60
HMM	hidden Markov model	38
htGRB	high-turnover GRB	90
IFN-β	interferon-β	23
LINE	long interspersed nuclear element	87
ltGRB	low-turnover GRB	90
LTR	long terminal repeat	87
MF	molecular function	85
MSA	multiple sequence alignment	38
PIC	pre-initiation complex	20
RNA	ribonucleic acid	19
SINE	short interspersed nuclear element	87
SOM	self organising map	86
TAD	topologically associated domain	30

TBP TATA binding protein 20

TE transposable element 35

TF transcription factor 19

tpm tag per million 86

TSS transcription start site 20

UTR untranslated region 36

Chapter 1

Introduction

1.1 Metazoan gene regulation

Pioneering work by Jacob and Monod in the 1960s produced the first model of gene regulation. Through a series of elegant experiments in *E. coli*, they were able to show that the product of one gene was able to regulate the expression of another by binding its "operator" sequence and repressing transcription (Jacob and Monod 1961). The operator sequence is a stretch of DNA partially overlapping the promoter of an operon. The binding of a repressor to this sequence obstructs the access of RNA polymerase to the promoter, thereby preventing transcription. While this simple mode of gene regulation was identified in prokaryotes, it provides us with the foundation for understanding gene regulation in eukaryotes - that is that gene products can act in *trans* to activate or repress the transcription of other genes. These gene products are DNA binding proteins known as transcription factors (TFs). In eukaryotes, TFs can affect the expression of a gene by binding its promoter, similar to prokaryotic gene regulation, but can also bind distal *cis*-regulatory elements called enhancers. While promoters always coincide with the site of transcription initiation, enhancers can be either 5' or 3' of the gene they regulate, and are frequently very distant in linear genomic space. Eukaryotic gene regulation is further modulated by DNA accessibility. Eukaryotic DNA is packaged into chromatin by structural proteins called histones. Through post-translational modification of histones, chromatin can be either more or less compacted, thereby regulating the access of TFs to their *cis*-regulatory targets. A final layer of complexity in the regulation of eukaryotic gene expression comes from the requirement for enhancers to be in close 3D proximity with the promoter of their target gene. This imposes strict constraints on the 3D organisation of the genome, resulting in the formation of several hierarchical levels of preferential nuclear localisation and self-interaction. The presence of multiple layers of regulation provides the fine spatiotemporal control of transcription required for the development of a complex, multicellular organism. This section outlines and discusses the two main classes of eukaryotic *cis*-regulatory elements, promoters and enhancers, and their interaction with chromatin accessibility and 3D genome organisation.

1.1.1 Promoters

The promoter is the integration point of the total regulatory input for a gene, converting signals from multiple regulatory elements and TFs into differing rates of transcriptional initiation. The promoter spans the transcription start site (TSS), and positions RNA polymerase, and its associated general TFs, ensuring initiation of transcription occurs at the correct location (Figure 1.1B). Transcription initiation occurs within the 'core promoter' region, which stretches approximately 50 base pairs (bp) up- and downstream of the TSS. The core promoter sequence contains sequence motifs that are specifically bound by TFs, facilitating assembly of the pre-initiation complex (PIC). Promoters contain different combinations of TF binding motifs depending on the gene they regulate. For example, the promoters of many tissue specific genes contain the TATA box motif, while ubiquitously expressed gene's promoters do not, instead containing an greater proportion of cytosine followed by guanine (CpG) dinucleotides than expected, forming long CpG islands (CGIs) (Carninci et al. 2006; Akalin et al. 2009). In fact, metazoan RNA polymerase II promoters can be roughly separated into three main classes based on their expression patterns, sequence features, and precision of transcription initiation (Lenhard, Sandelin, and Carninci 2012), although there are likely many more unidentified classes of promoters.

Promoters can be divided into two general categories based on the precision with which they initiate transcription. In this context, the precision of transcription initiation from a promoter is defined by how frequently transcription initiation occurs at the exact same nucleotide within the core promoter. When visualising the frequency of all initiation events within a promoter, very precise promoters have a sharp profile, with the majority of transcription initiation events occurring at a single nucleotide position or within a very narrow window. Broad promoters initiate transcription at multiple positions spread over a wider area, resulting in a more dispersed profile of initiation events. These definitions are facilitated by cap analysis of gene expression (CAGE), a technique that identifies all transcription initiation events with single nucleotide resolution (Shiraki et al. 2003). In general, sharp promoters, or type I promoters, are enriched for the TATA box motif and tend to regulate tissue specific genes. Broad promoters can be further divided into two categories based on the tissue specificity of the genes they regulate and their sequence content. Type II promoters regulate ubiquitously expressed genes and contain a single short CGI. The second class of broad promoters, type III promoters, are enriched for genes that regulate multicellular development. These promoters are also CpG rich, however, unlike type II promoters, they frequently contain multiple CGIs that extend beyond the promoter into the gene body (Akalin et al. 2009). Interestingly, broad promoters have very consistently positioned 5' and 3' nucleosomes relative to the TSS, while there is no consistent positioning of the nucleosomes surrounding sharp promoters. This likely points to different mechanisms governing the site of transcription initiation between sharp and broad promoters. Sharp promoters are enriched for a strong TATA box motif, located 30 (+/- 2) nucleotides 5' of the TSS, while broad promoters do not contain any very strong positionally constrained TFBS motifs. Thus it is possible that transcription initiation in sharp promoters is very precise due to anchoring of the PIC via the binding of TATA binding protein (TBP) to the TATA motif, while in broad promoters the location of the downstream nucleosomes may dictate where initiation takes place (Haberle et al. 2014).

The division between promoter classes is still fairly rough, and there likely exist several classes of promoters which have not yet been identified, however, what is clear is that promoters have evolved from simple "on-off" switches, as originally described by Jacob and Monod, to achieve fine scale control of transcription initiation in a variety of contexts.

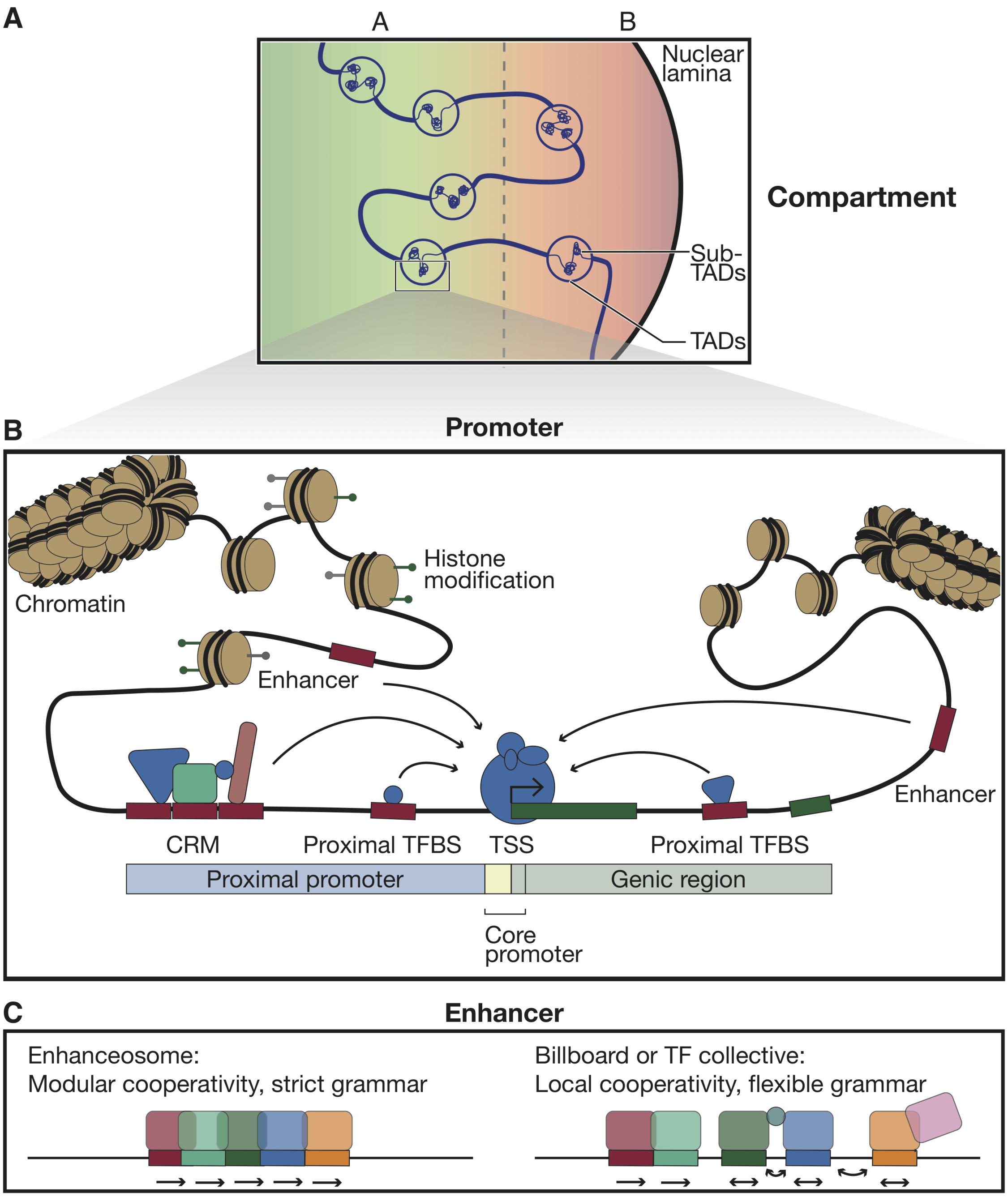

Figure 1.1: Metazoan gene regulation. (**A**) The genome can be broadly separated into two preferentially self-interacting compartments. The A compartment is generally located at the centre of the nucleus and associated with active transcription, while the B compartment is correlated with the nuclear lamina and generally transcriptionally silent. Within compartments, the genome is further divided into TADs and subTADs that define the regulatory environment of a gene. (**B**) Promoters integrate the total regulatory input required to correctly direct transcription. This occurs via the recruitment of RNA polymerase II to a gene's core promoter - the region immediately surrounding the transcription start site (TSS). This is orchestrated by the binding of transcription factors to their binding sites (TFBS) within proximal and distal *cis*-regulatory elements (enhancers).

(continued)

Figure 1.1: It is common for transcription factor binding to occur in clusters known as *cis*-regulatory modules (CRMs). (**C**) There are several proposed models for the binding of multiple transcription factors at enhancers. These models range from the strict enhanceosome model, in which the order and orientation of TFBS must be maintained for enhancer activation, to the highly flexible billboard model, in which each transcription factor within an enhancer can bind and provide signals to the promoter without maintenance of TFBS order and orientation, and without the presence of other transcription factors. Adapted from: (A) Yu and Ren 2017, (B) Lenhard, Sandelin, and Carninci 2012, (C) Lelli, Slattery, and Mann 2012.

1.1.2 Enhancers

The second major class of *cis*-regulatory elements in metazoa is enhancers. Enhancers are DNA sequences of approximately 10 - 1000bp that contain binding sites for TFs. Enhancers can be up- or downstream of their target gene, and can act at long distances from the TSS. Enhancers can be bound by a single or, more commonly, multiple TFs. The combinatorial binding of multiple TFs to an individual enhancer provides greater context-dependent control of a gene's regulation.

Transcription factor binding at enhancers

The mode by which multiple TFs cooperatively bind enhancers is the subject of much research and currently there are several competing models. The enhanceosome model posits that enhancer activity requires the binding of multiple TFs to tightly spaced, or even overlapping, TFBS with a strict order and orientation (Figure 1.1C). This model is derived from the identification of the enhancer responsible for the activation of interferon-β (IFN-β). The cooperative binding of eight TFs to the IFN-β enhancer is required for its activation (Thanos and Maniatis 1995; Panne 2008) and subsequent IFN-β expression. None of these eight TFs were able to individually activate IFN-β, and, in fact, the loss of any one of them prevented IFN-β expression (Thanos and Maniatis 1995). Structural studies and molecular dynamics simulations suggest that, despite the density of TFBS motifs in this 55bp enhancer, TF cooperativity is unlikely to be mediated by protein-protein interactions. Instead these studies propose that TF binding induces changes in the local DNA structure, facilitating subsequent TF binding to overlapping binding sites (Panne, Maniatis, and Harrison 2004; Panne, Maniatis, and Harrison 2007; Panne 2008). The IFN-β enhancer is extensively studied and very well characterised, however it is one of very few examples (Barthel et al. 2003) in which the order and orientation of TFBS is so inflexible. The strict threshold for enhancer activation, and the high stability of the enhanceosome complex (Thanos and Maniatis 1995), suggests that these types of enhancers may only be used to regulate genes that require very precise regulation and a prolonged period of expression once activated (Lelli, Slattery, and Mann 2012).

At the opposite end of the spectrum to the highly inflexible enhanceosome model is the billboard model (Figure 1.1C)(Kulkarni and Arnosti 2003; Arnosti and Kulkarni 2005). Under this model, TFBS within an enhancer can have their orientation and order changed without affecting the expression of the target gene. Further, loss of individual binding sites does not abolish gene expression, rather the gene expression pattern in a particular cellular context is altered. The flexibility of TFBS spacing and orientation is attributed to individual TFs independently recruiting different components of the transcriptional machinery, or specific cofactors, that promote transcription. This form of combinatorial TF binding has been observed in many genome-wide TF binding assays (Menoret et al. 2013; Jiang and Singh 2014), and fine-scale dissection of mammalian enhancers has shown that, in general, enhancer grammar is relatively flexible (Patwardhan et al. 2012; Smith et al. 2013).

A third model for TF binding at enhancers has been proposed that can be viewed as an intermediate between the enhanceosome and the billboard model. The TF collective model suggests that TF binding at enhancers is highly cooperative, resulting in an all-or-nothing, switch-like behaviour, similar to the enhanceosome model, but does not rely on strict ordering or orientation of the TFBS within the enhancer (Junion et al. 2012; Erceg et al. 2014). Under this model, protein-protein interactions between the TFs stabilise their cooperative binding to the enhancer to such a degree that there need not be a TFBS motif for all TFs present in the complex. Given the vast number of cellular contexts in which metazoan genes must be correctly regulated, it is likely that no single model can accurately encompass all modes of TF binding, and there likely exists a spectrum of flexibility with respect to cooperative TF binding.

Enhancer-core promoter specificity

Once an enhancer has been bound by the correct combination of TFs, and recruited the various cofactors required for its activation, it must make contact with the core promoter to influence gene expression. It is generally accepted that this occurs by looping of the DNA such that the enhancer is in close physical proximity with the promoter of the gene that it regulates (discussed in detail below). Since enhancers can be separated from their target gene by large genomic distances, and frequently by intervening genes, a long-standing question in regulatory genomics is how an active enhancer identifies its correct regulatory target. Temporarily overlooking chromatin conformation mediated mechanisms, there appears to be enhancer-core promoter specificity based purely on sequence composition. Since distinct classes of genes tend to have specific core promoter types, it is logical that their disparate motif content may result in differing affinity for specific enhancers. Indeed, using self-transcribing active regulatory region sequencing (STARR-seq) (Arnold et al. 2013), Zabidi et al. showed that enhancers had a marked preference for either a housekeeping or a developmental gene promoter in *Drosophila* (Zabidi et al. 2015). The authors selected core promoters from *ribosomal protein gene 12* and *even skipped* to represent the housekeeping and developmental category respectively, and found that of the 11,364 enhancers tested, 72% activated one class of core promoter at least twofold more than the other. Further the enhancers that specifically activated one promoter more than the other were differentially enriched for TFBS motifs. Since both housekeeping and developmental core promoters, and the enhancers that activate them, are differentially enriched for specific TFBS motifs, it is likely that the specificity between enhancers and promoter types is driven by biochemical interactions between the TFs (and their cofactors) recruited at enhancers and promoters. It has been hypothesised that different types of promoters allow the assembly of distinct PICs that are only biochemically compatible with specific TFs, resulting in the observed enhancer-core promoter specificity. This hypothesis is supported by the replacement of TBP with TBP-related factor 2 (TRF2) at the promoters of ribosomal protein genes (Wang et al. 2014).

Enhancer transcription

While the role of enhancers is to facilitate transcription initiation at core promoters, there is also pervasive transcription initiation within enhancers themselves (Kim et al. 2010; Wang et al. 2011; Andersson et al. 2014). In fact, transcription initiation at promoters and enhancers is remarkably similar - both have similar frequencies of core promoter motifs, recruit transcriptional machinery (including RNA polymerase II), initiate transcription bidirectionally and have well positioned up- and downstream nucleosomes (Core et al. 2014). Further, the level of enhancer transcription is correlated with the activity of the enhancer (Kim et al. 2010; Andersson et al. 2014) and the expression of the target gene (Cheng et al. 2015).

It is currently still unclear whether transcription from enhancers is functional or simply a fortuitous consequence of the recruitment of strong transcriptional activators to enhancers - a requirement for the activation of their target genes. The observation that knockdown of some enhancer RNAs (eRNAs) results in the downregulation of their target genes has prompted the proposal that eRNAs have a direct role in transcriptional activation (Li et al. 2013; Mousavi et al. 2013; Hsieh et al. 2014). Many eRNA functions have been hypothesised, including facilitating enhancer-promoter looping (Lai et al. 2013; Li et al. 2013; Hsieh et al. 2014), ensuring accessibility to DNA at the promoter (Mousavi et al. 2013) and recruiting RNA polymerase II (Johnson et al. 2003; Mousavi et al. 2013). However, it is possible that many of the observations relating to eRNA level and target gene activation are solely a consequence of creating a local transcriptionally permissive environment via the recruitment of strong transcriptional activators (Haberle and Stark 2018). Supporting this argument is the observation that bidirectional transcription is a pervasive feature of open chromatin regions, suggesting that this is not specific to promoters and enhancers (Young et al. 2017). Further, the authors found no evidence of purifying selection on eRNAs within the human population, challenging their functional significance (Young et al. 2017).

1.1.3 Chromatin conformation and higher-order genome organisation

Up to this point I have discussed gene regulation by enhancers and promoters without taking into consideration the effects of chromatin conformation and the 3D organisation of the genome. For an entire genome to fit inside the nucleus of each cell, it must be highly compacted. This is achieved via DNA association with histones, yielding nucleosomes, the basic repeating elements of chromatin. Chromatin can be more or less compacted, thereby regulating the access of TFs, and other proteins, to DNA. Further, regions of the genome that are in active and inactive chromatin conformations tend to separate into distinct multi-megabase scale compartments that preferentially contact other regions in the same compartment. Within these compartments smaller scale regions also preferentially interact with themselves, defining the boundaries beyond which enhancers are physically unable to target genes.

Histone modifications

The nucleosome is made up of DNA wrapped around an octamer of histone proteins. Each histone has an N-terminal domain which extends beyond the core nucleosome, referred to as the histone tail. Histone tails can be post-translationally modified (most commonly acetylated, methylated, phosphorylated, ubiquitinated, sumolated or ribosylated) at multiple residues (Suganuma and Workman 2011). The post-translational modification of histones is an important factor in the modulation of chromatin conformation, and thus the accessibility of the DNA to transcriptional initiators. Histone modifications can directly influence chromatin conformation by affecting nucleosome-nucleosome interactions (Shogren-Knaak et al. 2006; Lu et al. 2008), however they usually exert their influence by modulating the access of chromatin remodelling complexes to nucleosomes, and recruiting TFs that modify the local chromatin state (Vettese-Dadey et al. 1996; Margueron, Trojer, and Reinberg 2005). Chromatin remodelling complexes influence the accessibility of chromatin by adjusting the spacing between nucleosomes, mainly via the repositioning or eviction of nucleosomes at specific locations (reviewed in Clapier et al. 2017). This is a key step in transcription initiation, as formation of the PIC requires that the nucleosome occupying the promoter of a gene be displaced, thereby uncovering TFBS motifs and facilitating TF binding. A similar process can occur at enhancers allowing TF binding and enhancer activation (Bossen et al. 2015), although enhancer activation can also be mediated by maintaining the chromatin in an accessible state without nucleosome depletion (Iwafuchi-Doi et al. 2016).

Following more than a decade of extensive application of techniques to assay histone modifications, including chromatin immunoprecipitation sequencing (ChIP-seq) and ChIP-chip, various genomic features have been robustly associated with specific histone modifications. For example, active promoters tend to be marked with H3K4me3, while active enhancers are marked by H3K4me1 and H3K27ac. A number of genomic features and their associated histone modifications are listed in Table 1.1. These associations are so robust that histone modification data has been widely used for the annotation of *cis*-regulatory elements in multiple metazoan species. Further, histone modification data has been used to automatically categorise the genome into functional "segments" using the primary cell lines from the ENCODE project (Birney et al. 2007) via a chromatin state segmentation by hidden markov model (ChromHMM) (Ernst and Kellis 2012) and Segway algorithms (Hoffman et al. 2012). These include states annotated as active, weak, strong, poised, and repressed chromatin.

Table 1.1: Genomic features and their associated histone modifications (Adapted from Rivera and Ren 2013)

Genomic Feature	Histone Modification	References
Promoters	H3K4me3	Bernstein et al. 2005; Kim et al. 2005; Pokholok et al. 2005
Bivalent/Poised Promoter	H3K4me3/H3K27me3	Bernstein et al. 2006
Transcribed Gene Body	H3K36me3	Barski et al. 2007
Enhancer (active and poised)	H3K4me1	Heintzman et al. 2007
Poised Developmental Enhancer	H3K4me1/H3K27me3	Creyghton et al. 2010; Rada-Iglesias et al. 2011
Active Enhancer	H3K4me1/H3K27ac	Heintzman et al. 2007; Creyghton et al. 2010; Rada-Iglesias et al. 2011
Polycomb Repressed Regions	H3K27me3	Bernstein et al. 2006; Lee et al. 2006
Heterochromatin	H3K9me3	Mikkelsen et al. 2007

Compartments

The regulation of gene expression can also be affected by limiting the potential loci with which an enhancer can interact. This occurs by, or due to, the organisation of the genome in three dimensions. At the multi-megabase scale, the genome is divided into two compartments that contain loci which preferentially interact with other loci in the same compartment. These compartments are known as the A and B compartments (Figure 1.1A). Loci within the A compartment tend to be gene dense, have high gene expression and correlate with active chromatin modifications and DNA accessibility (Lieberman-Aiden et al. 2009). Further, the A and B compartment tend to localise to different regions of the nucleus, with the A compartment located in the centre and the B compartment found at the periphery of the nucleus, frequently coinciding with lamina-associated domains (Ryba et al.

2010). Approximately a third of the genome switches between compartments during stem cell differentiation (Dixon et al. 2015), and when comparing a broader range of tissues the proportion of switching regions increases to almost two thirds of the genome (Schmitt et al. 2016). Switching from the B to the A compartment is associated with increased gene expression, suggesting that sequestering genes in the B compartment may be a way to regulate their expression in cell types in which they are not required (Lin et al. 2012).

More recent analysis, using higher resolution Hi-C, has further divided compartments into five subcompartments (A1-A2 and B1-B3)(Rao et al. 2014). Each of these subcompartments is associated with a specific set of histone modifications, however it is still questionable as to whether these divisions represent any further meaningful biological differences.

Topologically associating domains

Examining 3D genome organisation at higher resolution, it is clear that within compartments there exist smaller domains that are strongly self-interacting. These are known as topologically associated domains (TADs) (Figure 1.1A). TADs have a median size of approximately 800kb in the human and mouse genome and cover the vast majority of the genome (Dixon et al. 2012). TADs have also been identified in zebrafish (Gómez-Marín et al. 2015) and *Drosophila melanogaster* (Sexton et al. 2012). There is accumulating evidence that TADs are a fundamental unit of genome organisation in metazoan genomes (Dixon, Gorkin, and Ren 2016). First, TAD boundaries tend to be cell-type invariant (Dixon et al. 2012; Dixon et al. 2015; Nora et al. 2012) and when compartment switching occurs during development, TADs switch as a whole (Dixon et al. 2015). TAD boundary positions are also highly conserved between species (Dixon et al. 2015). Further, TADs appear to be stable units of replication as the position of replication domain boundaries correlate very strongly with TAD boundaries (Pope et al. 2014).

Since enhancers must be in close physical proximity with the promoter of their target gene, TADs define the 3D space in which an enhancer may perform its function. This means that while TADs are fundamental units of genome organisation, they are also discrete, functional units of long-range gene regulation. Upon deletion of TAD boundaries, interactions between adjacent TADs are significantly increased and there is disregulation of nearby genes (Nora et al. 2012). Further, mutation or deletion of binding sites for CTCF, the protein that stabilises TAD boundaries, results in altered local chromatin structure and disregulation of nearby genes (Dowen et al. 2014; Guo et al. 2015; Narendra et al. 2015). This occurs due to the ectopic interaction of enhancers in one TAD with the promoters of genes in neighbouring TADs, illustrating the essential role TADs play in defining an enhancer's potential search space.

Enhancer-promoter looping

When examining interactions within TADs using very high resolution Hi-C or 5C data, smaller domains of self-association become apparent (Phillips-Cremins et al. 2013; Rao et al. 2014). These have been dubbed subTADs or contact domains (Figure 1.1A). subTADs are much less conserved across cell types and tend to relate to cell type specific gene expression (Berlivet et al. 2013; Phillips-Cremins et al. 2013). These fine-scale interaction structures seem to be driven by the looping of enhancers to make contact with their target promoters in the cell types in which they are active, or by the process of transcription itself. While structural TAD boundary-stabilising interactions are mediated by the architectural proteins CTCF and cohesin, subTAD interactions appear to be mediated by Mediator and cohesin, reflecting their functional regulatory nature (Phillips-Cremins et al. 2013).

1.2 The evolution of metazoan gene regulation

The emergence of distal *cis*-regulatory elements was essential for the development of the complex multicellular metazoan body plan, and their continual evolution is a major driver of the extreme phenotypic diversity of extant metazoan species. The gene repertoire of metazoan species is remarkably well conserved. Perhaps even more remarkable is the recent observation that some unicellular eukaryotes contain many of the genes essential for the complex developmental regulation that is a hallmark of metazoan species. Thus, the major driver of metazoan phenotypic diversity, and their transition from a single to multicellular lifestyle, is not changes in gene complement, but evolution of the regulation of these genes and expansion of their regulatory networks.

1.2.1 *Cis*-regulatory evolution and the origin of multicellularity

Comparison of simple metazoa and their most closely related unicellular eukaryotes has shown that unicellular eukaryotes lack distal *cis*-regulatory elements. In contrast, and perhaps surprisingly, they contain many of the genes that are important for metazoan multicellularity-related functions. These genes include cell adhesion genes crucial for cell-cell and cell-extracellular matrix interactions in the establishment of cell layers and tissues in animals (King et al. 2008; Sebé-Pedrós et al. 2010; Nichols et al. 2012; Suga et al. 2013), signal transduction genes (Manning et al. 2008; Suga et al. 2012; Suga et al. 2014), including all the intracellular components of the Hippo pathway (Sebé-Pedrós et al. 2012), and finally, many TFs such as *NF-κB*, *p53*, *RUNX* and *T-box* (Sebé-Pedrós et al. 2011; Sebé-Pedrós et al. 2013a). Further, unicellular eukaryotes undergo temporally controlled cell differentiation to distinct lifestyle phases, and this process is tightly regulated at the transcriptional level (Sebé-Pedrós et al. 2013b). The major differences between unicellular eukaryotes (with the exception of yeast) and simple metazoa is the lack of distal *cis*-regulatory elements, and repressive histone modifications such as H3K27me3 and H3K9me3 (Sebé-Pedrós et al. 2016). Both *S. cerevisiae* and *S. pombe* transitioned to unicellularity from multicellular ancestors, and thus repressive histone modifications in these species may have been acquired in their ancestral transition to multicellularity (Nagy et al. 2014). Interestingly, typeI and typeIII promoters, which control cell-type specific and developmental genes respectively, also appear to be a metazoan innovation (Sebé-Pedrós et al. 2016). Taken together these results suggest that one of the major factors contributing to the transition from a single to multicellular lifestyle, was the evolution of complex gene regulation. This includes the evolution of distal *cis*-regulatory elements to provide greater spatiotemporal control of gene expression, the addition of repressive chromatin modifications to securely regulate important developmental genes, and the diversification of promoter types to allow for more precise regulation of enhancer-promoter interactions.

1.2.2 *Cis*-regulatory evolution and metazoan phenotypic diversity

The sequencing of hundreds of metazoan genomes over the last two decades has revealed that, in general, the vast array of phenotypic and morphological diversity observed in metazoan species can not be explained by species-specific differences in gene content. In fact, the major contributor to phenotypic diversity in closely related species is changes in the timing, location and magnitude of developmental gene expression. These changes in gene regulation are the result of the independent evolution of *cis*-regulatory elements in each metazoan lineage (Wittkopp, Haerum, and Clark 2008; McManus et al. 2010). Comparative analyses have shown that while *cis*-regulatory elements tend to be evolutionarily conserved, they are under reduced selective pressure compared to protein coding regions (Asthana et al. 2007). This combined with the observation that modification of *cis*-regulatory elements has a more subtle effect on gene expression than mutation of the coding sequence (Carroll 2008; Wray 2007), suggests that changes in the *cis*-regulatory repertoire of a gene would be better tolerated than coding mutations. In this way, *cis*-regulatory elements provide the substrate upon which selection can act to fine tune developmental gene regulation. Indeed, numerous examples of variation within *cis*-regulatory elements providing evolutionary innovations have been identified (Carroll 2005; Wray 2007; Carroll 2008; Wittkopp and Kalay 2012). The modularity of TF binding within *cis*-regulatory elements, and in particular enhancers, allows altered binding of an individual TF to affect only a portion of the full regulatory input provided by that element. Further, enhancers can act independently of each other, each controlling a subset of the total expression pattern of the gene. This modular nature of TF binding and enhancer function results in reduced pleiotropic effects of changes to individual enhancers, when compared to mutations in protein coding genes. It has also been shown that some enhancers exist in pairs with highly overlapping function (Perry et al. 2010; Osterwalder et al. 2018). These enhancer pairs are hypothesised to ensure robust expression of the target gene, but it is also plausible that mutations within mostly redundant enhancers would be well tolerated, and could result in co-option of the existing enhancer to novel functions (Rebeiz et al. 2011).

1.2.3 Mechanisms of *cis*-regulatory evolution

Differences in gene expression between species are largely driven by species-specific variability in the timing, location or efficiency with which a gene's repertoire of enhancers initiate transcription. This process is chiefly governed by the binding of TFs, and their associated cofactors, at enhancers and promoters. It is conceivable then, that changes in sequence content or local chromatin environment that affect TF binding could drive evolutionary divergence.

In *Drosophila* species, the binding locations of developmental TFs are highly conserved, and a linear relationship between quantitative changes in binding intensity and evolutionary distance was observed (Bradley et al. 2010; He et al. 2011; Paris et al. 2013). In mammals changes in TF binding intensity around target genes are also clearly correlated with changes in the gene's expression pattern, however the exact location of TF binding is less well conserved (Villar, Flicek, and Odom 2014). This suggests that upon loss of TF binding, compensatory gain must occur in the vicinity to maintain the transcriptional output of the target gene (Kunarso et al. 2010). Despite the reduction in the conservation of TF binding between mammalian species, there is still a strong correlation between the proportion of overlapping binding events and the evolutionary distance of the comparison (Villar, Flicek, and Odom 2014).

A substantial proportion of TF binding differences between species can be explained by differences in the underlying sequence bound by the TFs (Zheng et al. 2011). These sequence changes can be derived from point mutations, insertions and deletions, or genomic rearrangements. Since TFs bind short recognition sequences, it is possible for neutral sequence changes to generate weak transcription factor binding sites that can have an effect on gene expression (Stone and Wray 2001). Accumulation of further sequence changes due to increased accessibility of the region, accompanied by changes in local chromatin environment, could then result in the evolution of these weak enhancers into mature, constrained enhancers (Emera et al. 2016). The effect of sequence changes on divergent TF binding is also supported by the observation that there is an enrichment of TFBS motif-disrupting mutations in differentially bound loci within the human population (Kasowski et al. 2010; Reddy et al. 2012).

TFBS motifs can also be introduced by the integration of transposable elements (TEs). TEs can contain sequences that are similar to the recognition site of a TF, therefore only requiring a few mutations to become strong binding sites (Johnson et al. 2006). Further, some TEs, such as endogenous retroviral sequence 1 (ERV1), already contain strong TFBS motifs, and their expansion throughout the genome may have recruited new genes the the regulatory network of that TF (Wang et al. 2007). Repeat expansion may play a more important role in the generation of TF binding sites for TFs that have longer binding motifs, as these motifs are less likely to be generated by the accumulation of point mutations (Stone and Wray 2001).

While sequence change is clearly an important factor in the evolution of TF binding at *cis*-regulatory elements, it must be noted that a large proportion of divergent TF binding events can not be explained by sequence differences (Villar, Flicek, and Odom 2014). It may be that the divergent binding at these sites is mediated by another mechanism, such as chromatin conformation (Degner et al. 2012; Shibata et al. 2012), or that we do not yet understand all the ways sequence can affect TF binding. In support of the latter, a study in which TF binding in a mouse strain containing a segregating copy of human chromosome 21 was compared to TF binding in human liver samples, found that up to 85-92% of TF binding events on human chromosome 21 were identical between the mouse and human contexts. Further, the gene expression profiles for genes on human chromosome 21 were highly correlated in both contexts (Wilson et al. 2008). This study suggests that sequence alone is sufficient to direct highly conserved TF binding.

1.3 Extreme non-coding conservation

The comparison of vertebrate genomes has identified numerous stretches of non-coding DNA that are deeply conserved across vertebrates, known as conserved non-coding elements (CNEs). A handful of these elements were first identified in the 1980s by comparing mammalian and avian introns and untranslated regions (UTRs) (Yaffe et al. 1985; Lemaire, Heilig, and Mandel 1988; Hraba-Renevey and Kress 1988; Kajimoto and Rotwein 1991; Rouault et al. 1993), however their pervasive presence in vertebrate genomes only became apparent upon systematic comparison of multiple vertebrate non-coding genomes (Bejerano et al. 2004; Sandelin et al. 2004; Woolfe et al. 2005). Depending on the method used, it is possible to identify several thousand CNEs that have remained all but unchanged over 450 million years of evolution. In fact, many of these elements display levels of sequence conservation well beyond what is observed in protein coding sequences (De Silva, Nichols, and Elgar 2014; Polychronopoulos et al. 2017). This level of non-coding conservation is unparalleled in the rest of the genome, and prompted a period of intensive research seeking to explain this phenomenon.

1.3.1 CNE identification

Since their initial discovery, numerous methods for CNE identification have been developed. At the most basic level, these methods can be divided into alignment-based and alignment-free methods.

Alignment-based CNE identification

Alignment-based CNE identification methods aim to identify stretches of highly conserved non-coding regions in either pairwise or multiple whole-genome alignments.

In the case of pairwise genome comparisons, whole-genome alignments are generated using one of many tools, however BLASTZ/LASTZ (Schwartz et al. 2003; Harris 2007) or LAST (Kiełbasa et al. 2011) are generally favoured. CNEs are then identified by scanning the alignments for non-coding regions which pass a predetermined conservation threshold, such as 90% sequence identity over 50bp (Dubchak et al. 2000; Bejerano et al. 2004; Sandelin et al. 2004). The conservation threshold used for a species comparison is usually selected based on the evolutionary distance between the two species. The selection of this threshold can be somewhat arbitrary, leading Babarinde and Saitou to propose an approach for the systematic selection of CNE identification thresholds. The authors used the sequence divergence at protein coding genes to define a threshold for CNE identification (Babarinde and Saitou 2013). A more stringent approach was also proposed in which the sequence divergence at protein coding genes was calculated after the exclusion of all third codon positions, and only non-coding regions that exhibit divergence below the mean divergence of protein coding sequences were defined as CNEs. This approach assumes that the rate of sequence divergence in protein coding regions is proportional to that of all sequences under negative selection. It is difficult to assess the validity of this assumption as coding and non-coding regions would be under different selective pressures based on their function. It is also unclear whether using this method results in CNE sets that are biologically different from those defined using an arbitrary but stringent threshold.

CNE identification using pairwise alignments suffers from poor power to detect short stretches of conservation. This can be mitigated by comparing multiple genomes, allowing for estimation of conservation at the resolution of single base pairs. Whole-genome multiple sequence alignments (MSAs) are most commonly generated using MULTIZ, an extension of LASTZ adapted for MSA (Blanchette et al. 2004). Several methods exist for the identification of constrained elements from MSAs, but two of the most commonly used are phastCons (Siepel et al. 2005) and genomic evolutionary rate profiling (GERP) (Cooper et al. 2005). PhastCons detects CNEs from multiple sequence alignments using a two-state hidden Markov model (HMM) to estimate the probability that each base pair belongs to a conserved element. These base-by-base conservation scores are then used to predict whole elements. GERP first builds a phylogenetic tree for the species used, based on the neutral substitution rates, and then identifies elements that exhibit fewer substitutions than expected. Candidate elements are then scored based on the magnitude of their substitution deficit, or how many "rejected substitutions" they contain (Cooper et al. 2005). Neither of these methods set a minimum sequence length and are therefore able to detect shorter CNEs than pairwise alignment-based methods. The disadvantage of using MSA-based approaches is that since a region must be conserved in multiple species in the alignment to be identified, they lack power to identify conserved elements that are turning over more rapidly and independently in different lineages.

Alignment-free CNE identification

Alignment-free methods avoid some of the problems associated with whole genome alignment, such as computational complexity and aligning highly fragmented assemblies. Most alignment free methods are only alignment free in that there is no requirement for whole-genome alignment. These methods use local alignment tools, such a BLAST (Altschul et al. 1990), to perform homology searches on repeat- and coding sequence-masked genomes (Babarinde and Saitou 2016). An alternative alignment-free approach was proposed by Warnefors et al. in which all possible unique k-mers in the reference genome are mapped to the genome of the species of interest using a short read aligner. Overlapping hits are then merged into longer elements, yielding CNEs (Warnefors et al. 2016). This approach can overcome some potential errors in MSA-based CNE detection, such as failing to identify CNEs due to gap insertion at ambiguous positions of the alignment, or due to CNEs being split over alignment blocks. However, it also has an increased false positive rate as a result of poor handling of sequences that occur in multiple copies in the genome due to duplication or assembly errors (Warnefors et al. 2016).

1.3.2 Sequence properties of CNEs

The majority of CNE studies have been performed in mammalian (Bejerano et al. 2004; Sandelin et al. 2004; Woolfe et al. 2005; Davies, Tsagkogeorga, and Rossiter 2014) or vertebrate genomes (Walter et al. 2005; Kikuta et al. 2007a; Lee et al. 2011; Davies, Tsagkogeorga, and Rossiter 2014), however CNEs have also been identified in arthropods (Glazov et al. 2005; Siepel et al. 2005; Engström et al. 2007), nematodes (Siepel et al. 2005; Vavouri et al. 2007), and other metazoan genomes

(Clarke et al. 2012; Doglio et al. 2013; Irvine 2013; Sanges et al. 2013). Identification of CNEs in such a broad range of species has enabled the characterisation of their shared properties, as well as some lineage-specific features of each set.

Comparing CNEs to genomic background sequence, it is clear that CNEs are strongly enriched in adenine and thymine (AT) relative to their flanking sequence. Beyond the general enrichment, there is also a sharp increase in AT content at CNE boundaries and a sharp decrease in AT content in the boundaries of sequences flanking CNEs (Walter et al. 2005). This effect is particularly pronounced in genomes with a high overall GC content (Vavouri et al. 2007). While this observation is still unexplained, it has been suggested that it may be due to a role for CNEs in nucleosome positioning or higher-order chromatin structure (Chiang et al. 2008).

Consistent with the high AT content of CNEs, it has been shown that over one third of human CNEs contain a TAATTA motif, which contains the core recognition motif of homeodomain DNA binding proteins (Chiang et al. 2008). In fact, there is ample evidence that CNEs are enriched in TF binding sites (Abnizova et al. 2007; Viturawong et al. 2013; Warnefors et al. 2016). One of the earliest proposed explanations for the extreme conservation observed at CNEs was that they contain multiple overlapping TF binding sites, similar to the enhanceosome model for TF binding at enhancers (Levy, Hannenhalli, and Workman 2001; Loots et al. 2002). However, there is no evidence that CNEs are more frequently bound by TFs, or bound by more TFs, than enhancers that do not exhibit such deep evolutionary conservation. Further, given the general promiscuity of TF binding, and the rapid turnover of TF binding sites between species, it does not seem likely that overlapping TF binding would be sufficient to explain the degree to which CNEs are conserved (Schmidt et al. 2010).

1.3.3 Biological functions of CNEs

The most noticeable characteristic of CNEs, which is shared between all species, is their nonrandom distribution across the genome. CNEs tend to occur in clusters that frequently span gene deserts and loci of genes encoding developmental transcription factors (Bejerano et al. 2004; Sandelin et al. 2004; Woolfe et al. 2005; Plessy et al. 2005; Engström et al. 2007; Kikuta et al. 2007a), and genes involved cell-cell communication (Vavouri et al. 2007). While CNEs identified independently in different clades do not share sequence conservation, they tend to cluster around these same functional subsets of genes (Engström et al. 2007; Vavouri et al. 2007). This observation, combined with the abundance of TF binding sites within CNEs, suggested that CNEs act as enhancers, regulating early development in metazoa. Indeed, transgenic reporter assays have shown that the vast majority of tested CNEs are capable of driving complex spatiotemporal patterns of gene expression (Kimura-Yoshida 2004; Woolfe et al. 2005; McEwen et al. 2006; Pennacchio et al. 2006; Navratilova et al. 2009; Bhatia et al. 2014; Parker et al. 2014; Spieler et al. 2014). The deep evolutionary conservation of CNEs suggests that they are crucial to embryonic development, however this assumption was initially questioned when it was shown that the deletion of large clusters of CNEs yielded viable, fertile mice with no deleterious phenotypes (Ahituv et al. 2007). This counter-intuitive finding remained a contentious issue in the field, until recently, when Dickel et al. conducted similar deletion experiments accompanied with deep phenotyping (Dickel et al. 2018). The authors showed that while CNE deletion yielded fertile, viable mice, fine-scale phenotyping revealed that these mice had neurological and growth abnormalities (Dickel et al. 2018). The authors argued that while these phenotypes appear mild in a lab based setting, in the wild they may have substantial fitness consequences. Further evidence of the crucial role CNEs play in development comes from the medical field. There are numerous examples of mutations within CNEs causing congenital defects in humans (reviewed in Polychronopoulos et al. 2017).

The observation that CNEs are essentially single copy in the haploid genome led researchers to ask whether they are dosage sensitive (Bejerano et al. 2004). To investigate this possibility, several high quality copy number variation (CNV) datasets from humans and other model organisms were used to assess the overlap of CNEs and CNVs (Derti et al. 2006; Chiang et al. 2008; McCole et al. 2014). CNEs are generally depleted from CNVs and segmental duplications in healthy cells, and this depletion is most likely due to rapid selection against cells with CNVs containing CNEs. Moreover, CNEs are depleted in *de novo* CNVs that have passed through germline meiotic processes at most once. These results led to the suggestion that CNEs play a role in monitoring the copy number of the genome, potentially through pairing of homologous CNEs during meiosis, followed by the initiation of apoptotic processes upon detection of mismatches or copy number changes. Interestingly, the same study found that CNEs show no depletion, and in some cases even enrichment, in cancer-specific CNVs; although it is difficult to disentangle whether the CNE copy number changes in cancerous cells drive disease or are simply the consequence of widespread genome instability associated with cancer (McCole et al. 2014). This hypothesis is attractive as it provides another potential source of selection against CNE sequence changes beyond their role as enhancers, however evidence of CNE pairing or interaction with mismatch sensing proteins is required for its validation.

1.3.4 Evolutionary dynamics of CNEs

The prevalence of CNEs in many species from many different kingdoms of life, and their clustering around equivalent classes of genes, suggests that CNEs are an ancient innovation of multicellular organisms. Given that one of the key factors driving the transition from a single- to multicellular lifestyle in eukaryotes was the evolution of distal *cis*-regulatory elements (Sebé-Pedrós et al. 2016), it is plausible that the first CNEs originate from these crucial distal regulators of spatiotemporal gene expression. Subsequent recruitment of CNEs to regulate new genes may then have contributed to increasing developmental complexity and evolution of highly specialised anatomical structures.

Emergence and recruitment of CNEs

The lack of sequence similarity (beyond AT content) between CNEs identified within a genome is likely due to recruitment of CNEs from a diverse array of genomic features. There are examples of CNEs that have been recruited from introns (Bejerano et al. 2004; Glazov et al. 2005; Siepel et al. 2005), TEs (Bejerano et al. 2006; Lowe, Bejerano, and Haussler 2007), ancient repeats (Kamal, Xie, and Lander 2006), and exons (Lampe et al. 2008; Dong et al. 2009). Interestingly, some exons can be recruited to enhancer activity, thereby serving both a protein coding and a regulatory role (Birnbaum et al. 2012; Ritter et al. 2012). The broad range of elements from which CNEs are recruited suggests that any sequence that is within range of a gene under long-range regulation can come under extreme purifying selection after acquiring regulatory function. Neutral sequence probably acquires regulatory function via neutral mutation towards TFBS motifs, as described in section 1.2.3.

CNE recruitment has been continual throughout vertebrate evolution, with the rate of recruitment varying between lineages (Wang et al. 2009). Primates appear to have recruited CNEs particularly rapidly (Babarinde and Saitou 2013), whereas since the divergence of tetrapods and teleosts, tetrapod CNEs appear to have been evolving very slowly (Stephen et al. 2008). Beyond vertebrates, there is evidence that CNEs have been recruited in clusters around equivalent classes of genes in more anatomically simple phyla such as Porifera and Cnidaria (Ryu, Seridi, and Ravasi 2012). CNEs could not be identified between these phyla, leading the authors to conclude that they were recruited independently in each phylum, however a more parsimonious explanation would be that both the source of purifying selection, and CNE-based gene regulation were already in place in the last common ancestor of all phyla studied. The presence of CNEs at functionally equivalent genes in some of the earliest diverging metazoan species lends further support to the idea that CNE-based regulation is at least as ancient as the urmetazoan ancestor.

Mutation and loss of CNEs

Mutation and deletion of CNEs has occurred throughout the vertebrate lineage despite the extraordinary levels of purifying selection required to maintain them. Such changes in CNEs are likely to be highly deleterious, but many have been tolerated and underlie lineage-specific traits. For example, CNE losses are hypothesised to explain penile spine loss (McLean et al. 2011) and foot digit shortening in humans (Indjeian et al. 2016). In snakes, limblessness is associated with partial and total loss of CNEs that control limb development genes (Sagai et al. 2004; Infante et al. 2015; Kvon et al. 2016; Leal and Cohn 2016). Substitutions in a CNE nearby the *SHH* gene are thought to be sufficient to yield snakes with vestigial limbs, while total deletion of the same CNE, combined with other changes, results in snakes with total limb loss (Kvon et al. 2016). Interestingly, hundreds of CNEs have been identified that have been lost independently in multiple species (Hiller, Schaar, and Bejerano 2012). Further, it has been shown that independent loss of CNEs drives convergent morphological adaptations. For example, elbow structure modifications in dolphins and manatees are thought to be due to independent loss of a CNE near the *EGR2* gene (Marcovitz, Jia, and Bejerano 2016).

CNEs can also undergo bursts of lineage-specific positive selection. This phenomenon has been extensively studied in humans, identifying thousands of so-called human accelerated regions. (Sur and Taipale 2016; Pollard et al. 2006; Bird et al. 2007; Hubisz and Pollard 2014; Gittelman et al. 2015; Dong et al. 2016). These elements are defined by high conservation within mammals, and frequently beyond, but exhibit rapid divergence in humans. Transgenic reporter assays have shown that the equivalent human and chimpanzee elements drive divergent expression, suggesting that these accelerated elements may underlie many human-specific traits including brain size and bipedalism (Prabhakar et al. 2008; Boyd et al. 2015).

Overall these results show that while CNEs are under extreme selective pressure, they remain dynamic within vertebrate genomes, and their mutation, loss or gain can have dramatic effects on the phenotype of an organism.

1.4 Genomic regulatory blocks

The requirement for CNEs to remain in *cis* with the gene that they regulate has constrained metazoan genome evolution, resulting in syntenic arrays of CNEs that span developmental regulators (Goode et al. 2005; Kikuta et al. 2007a; Engström et al. 2007; Akalin et al. 2009). These arrays, known as genomic regulatory blocks (GRBs), form functional units of long-range regulation, with their constituent CNEs jointly regulating a single target gene (Akalin et al. 2009) (Figure 1.2A). GRBs are present in all sampled metazoan genomes, and as such represent an ancient and important feature of animal development. In this section I will discuss current methods for GRB identification and the defining characteristics of GRBs, based on the limited number of genomes sampled thus far.

1.4.1 GRB identification

Generally speaking, GRBs are regions of high CNE density, however accurate identification of their boundaries is non-trivial. In close species comparisons, high background conservation results in "noisy" CNE density signal across the genome, while in distant comparisons there is a chance that GRBs will be truncated or split due to the erosion of CNE conservation. Several approaches for GRB identification have previously been used, and they can be divided into three main groups based on their methodology: those that rely on the relatively crude merging of adjacent CNEs, those that use of an HMM to segment the genome into regions

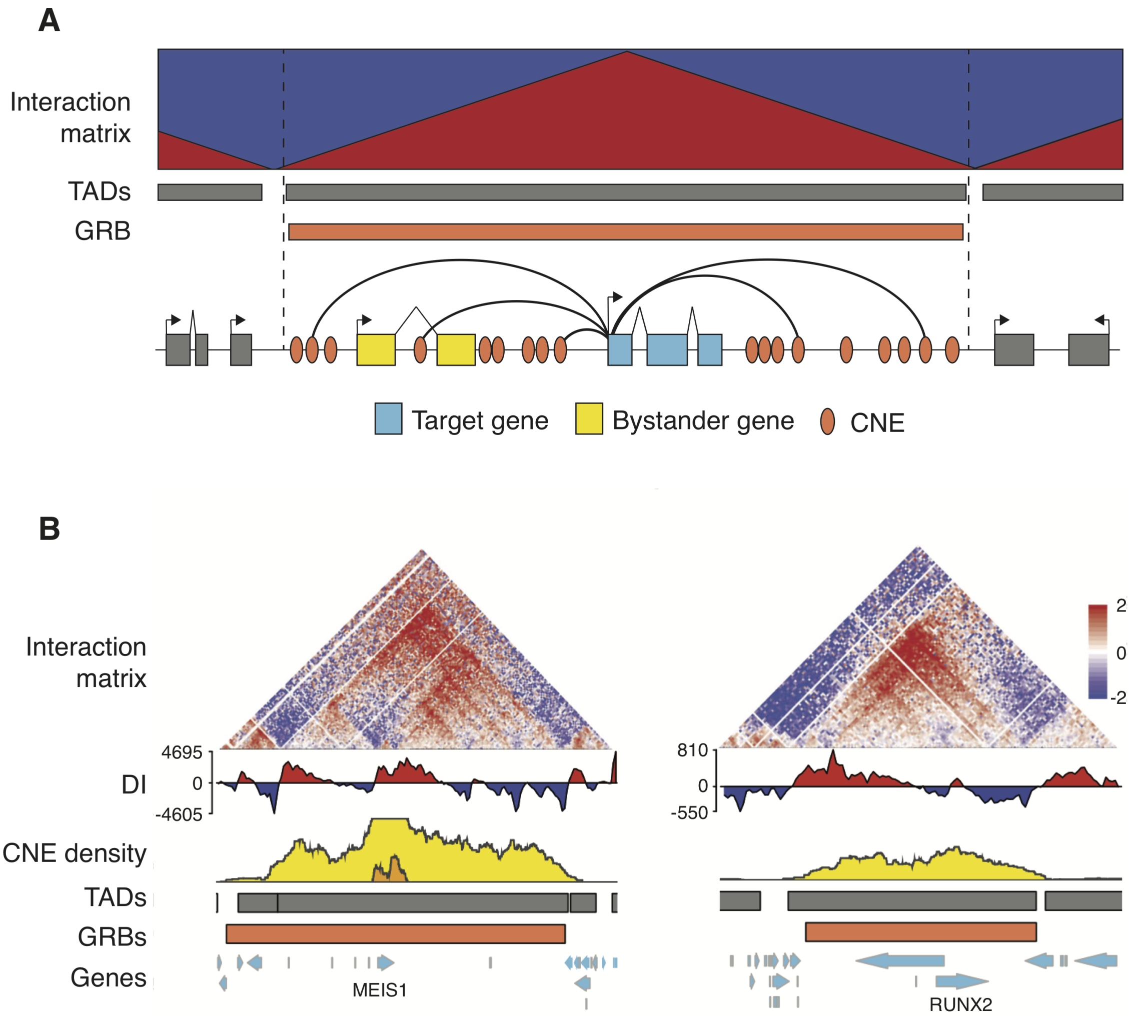

Figure 1.2: The GRB model of gene regulation. (**A**) GRBs contain syntenic arrays of CNEs that each contribute to the regulation of a single developmental target gene. GRBs can also contain bystander genes which are unresponsive to regulation by CNEs due to differences in their promoter architecture. GRBs coincide with TADs, indicating that there is a strong link between genome organisation and long-range regulation. (**B**) The *MEIS1* and (**C**) *RUNX2* GRBs. Adapted from Harmston et al. 2017.

of high and low conservation, and those based on identification of regions with a greater observed CNE density than expected under a null model of even distribution throughout the genome.

Dimitrieva and Bucher identified what they called ultraconserved genomic regulatory blocks (UGRBs) by merging all neighbouring CNEs that were separated by less than 0.5Mb in both human and chicken (Dimitrieva and Bucher 2013). Akalin et al., applied a similar method, but instead of merging CNEs, net alignments which were within a specified genomic distance of each other were merged (Akalin et al. 2009). The authors used a cut off of 450 kb in the human genome and 150 kb in zebrafish. These approaches are simple and effective, however the predicted GRB boundaries will be highly sensitive to lineage specific CNE mutation, loss or gain.

Recently, Harmston et al. developed a more generalisable approach based on the CNE density profile across the genome (Harmston et al. 2017). The authors implemented an unsupervised two state HMM that splits the genome in high and low CNE density regions. The high density regions that were within a predefined genomic distance of each other were then merged forming blocks of high conservation. The merging step continues iteratively until the gaps between all adjacent blocks are greater that a specified quantile of the widths of gaps between all adjacent CNEs. This method was applied to a range of species comparisons, and the identified GRBs were generally robust to the species comparison and CNE identification thresholds used. While this method has been highly successful in the past, it requires repeated rounds of parameter tuning and visual inspection to yield reliable GRBs. Further, the current implementation fails to run on less complete genome assemblies with many unassembled scaffolds (Nash, unpublished observation).

A fourth method is implemented in the `CNEr` package available for `R` (Tan 2015). This approach predicts GRBs based on the observed CNE density at a region compared to the expected density if CNEs were evenly distributed across the genome. GRBs are defined wherever the observed to expected ratio exceeds a predefined cut-off. This is followed by a post-processing step in which the GRB boundaries are shrunk to the location of the closest CNE to the boundary.

1.4.2 GRB target genes

Under the GRB model of gene regulation, each of the CNEs within a GRB independently contribute to the overall expression pattern of a single target gene (Kikuta et al. 2007a; Kikuta et al. 2007b; Akalin et al. 2009). The CNEs within a GRB are frequently separated from their target gene by intervening genes that are unresponsive to enhancer-based regulation, known as bystander genes (Figure 1.2A). Bystander genes often contain CNEs within their introns, thereby explaining their conserved synteny with the target gene (Kikuta et al. 2007a; Kikuta et al. 2007b; Dong, Fredman, and Lenhard 2009). In general, GRB target genes tend to be developmental TFs, while bystander genes are frequently ubiquitously expressed (Akalin et al. 2009). An initial set of target and bystander genes were first characterised in a study by Akalin et al. in which the authors defined target genes as those TFs that occur within a CNE density peak in both the human and zebrafish genome (Akalin et al. 2009). All other genes were then classed as bystander genes. This study showed that target and bystander genes differ in a number of ways, but most crucially in their promoter structure. GRB target genes tend to have a broader transcriptional initiation profile, and their promoters are spanned by multiple long CpG islands that often extended into the gene body. In contrast, GRB bystanders have consistently broad transcription initiation, but narrower than most GRB targets. Further, their promoters were generally spanned by a single short CpG island that did not extend past the promoter region. These differences in promoter structure may explain the difference in responsiveness to CNE-based regulation (as described in 1.1.1) (Akalin et al. 2009).

Recently, a PhD student in the Lenhard group developed a method for automatic annotation of GRB target and bystander genes (Tan 2018). The method makes use of a random forest model trained on the features of the original 259 target and 830 bystander genes annotated by Akalin et al. to predict target genes in all GRBs (Akalin et al. 2009). The random forest was trained on 19 informative features of GRB target genes, the most important of which were the number and size of CpG islands overlapping the gene, the tissue specificity of the gene's expression and the CNE densities surrounding the gene in multiple species comparisons. This method has significantly expanded our set of target and bystander predictions, and can now be applied to multiple genomes to further study the properties of GRB target genes.

1.4.3 GRBs and genome organisation

Interestingly, GRBs provide a link between developmental gene regulation and genome organisation. A recent study by Harmston et al. found that GRB boundaries strongly coincide with TAD boundaries in both vertebrates and invertebrates (Figure 1.2B)(Harmston et al. 2017). Further, the authors identified a set of features that define TADs that do and do not coincide with a GRB (GRB-TADS and nonGRB-TADs). GRB-TADs tend to be larger than nonGRB-TADs and are more strongly self interacting. Further, GRB-TADs span gene sparse regions, while nonGRB-TADs span regions of high gene density. It has been proposed that the difference in interaction strength within GRB-TADs and nonGRB-TADs may reflect the absence of a need for stable 3D structure in regions of the genome that do not contain genes under long-range regulation, however this hypothesis requires testing using newly available high resolution Hi-C data (Harmston et al. 2017).

When TADs switch compartments, they do so as a whole. Given the concordance between GRBs and TADs, compartment switching provides the ideal system in which to investigate the effects of different regulatory contexts on the expression of GRB target and bystander genes. To assess the expression patterns of target and bystander genes within GRB-TADs, Harmston et al. examined several loci for evidence of co-regulation. The authors showed that the dynamic range of gene expression of GRB target genes, in multiple contexts, was much greater than that of bystander genes, and interestingly, when GRB-TADs switch compartments between cell types, only the GRB target gene exhibits strongly correlated changes in expression (Harmston et al. 2017). This finding supports the hypothesis that GRB target and bystander genes are independently regulated (Akalin et al. 2009), however this analysis was only performed on a handful of identifiable loci, and must be repeated with Hi-C data from diverse cell types and conditions to draw solid conclusions.

1.5 Aims of this thesis

This thesis focuses on the evolutionary dynamics of GRBs, hoping to provide further evidence that they are an ancient and crucial feature of metazoan gene regulation. To study the dynamics of GRB evolution, it is necessary to identify GRBs in a broad range of species comparisons and genomes (Figure 1.3). This requires robust methods for calculating genome conservation and identifying GRB boundaries that are comparable between multiple contexts.

Thus, the first aim of this thesis is to develop a measure of pairwise genome conservation that is not reliant on the strict minimum identity and length thresholds that are used in CNE identification. Identification of CNEs in closely related species, using pairwise genome comparisons, has the greatest ability to identify lineage-specific conserved gene regulation, however these comparisons suffer from the need to select extremely stringent conservation criteria, including defining minimum lengths for CNEs that exceed 400bp. In Chapter 2 I define a novel measure of pairwise genome conservation that implicitly takes into account the background conservation of the species compared, thereby reducing the requirement for arbitrary threshold selection. Further, in this chapter I define a novel approach to accurately identify GRB boundaries that relies on statistical change point modelling to identify changes in the mean and variance of conservation scores across the genome.

Having established a robust set of methods for GRB identification, in Chapter 3 I use these methods to identify and characterise GRBs that exhibit deep and shallow conservation in three independent metazoan lineages. I show that these GRBs share many features between lineages, despite independent evolution for hundreds of millions of years, that may explain their relative rates of divergence.

Finally, in Chapter 4, I use these novel methods to identify GRBs in extremely compact genomes for the first time, and investigate the effects of genome compaction on the identified GRBs. I also examine the relationship between genome size and GRB size and composition, and show that GRB size tends to scale proportionally with genome size and GRBs are enriched and depleted for similar genomic features.

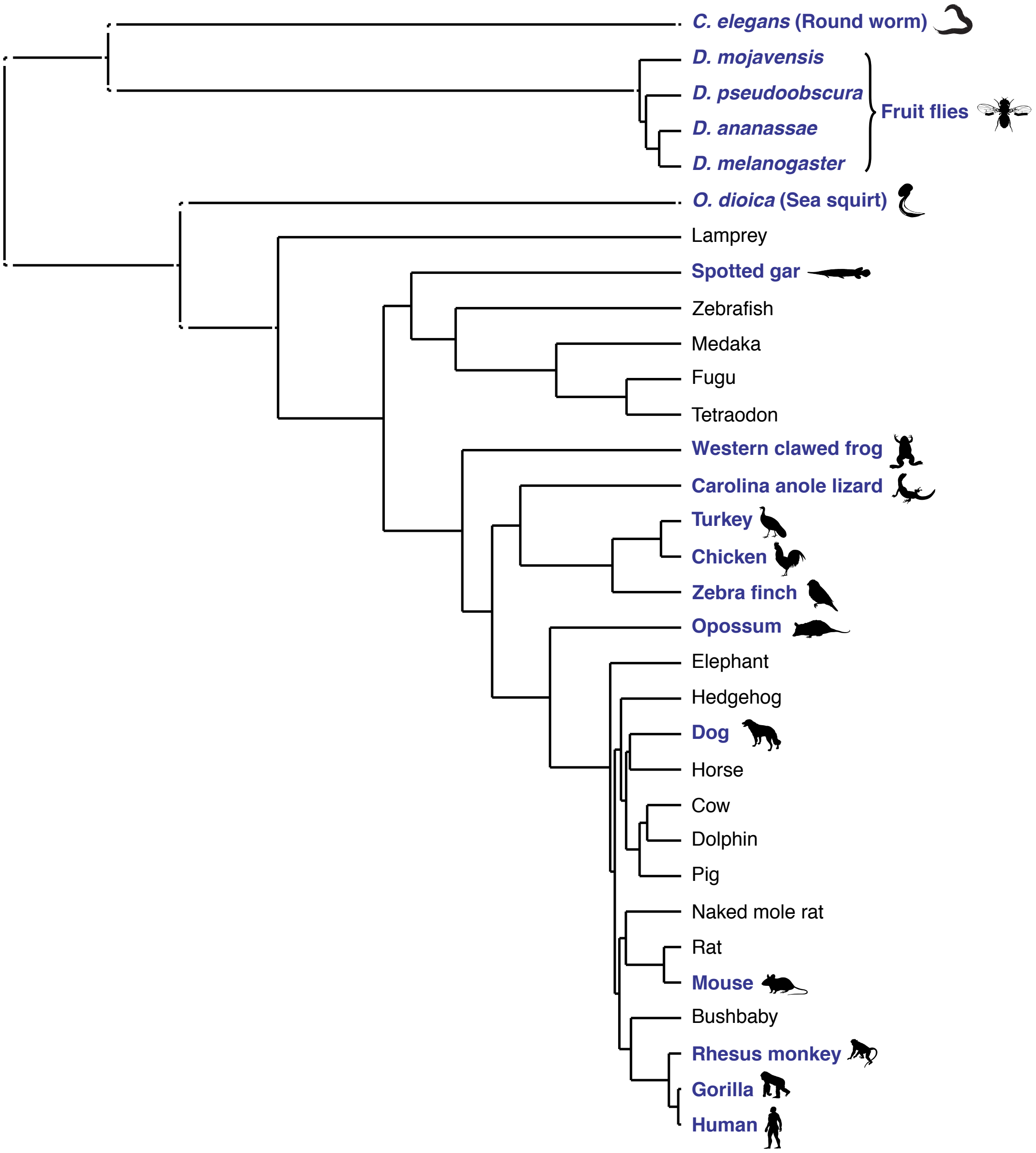

Figure 1.3: Evolutionary relationship of species studied in this thesis. The species studied in this thesis are highlighted in blue and accompanied by their outline (obtained from phylopic.org). Branches within the phylogeny depicted as solid lines are based on evolutionary distance as estimated from multiple sequence alignment, while those represented as dashed lines are approximate relationships based on estimated divergence times.

Chapter 2

A kurtosis-based measure of conservation

2.1 Introduction

Studying the dynamics of GRB evolution and the functional relationship between GRBs and TADs relies on robust methods to identify GRBs across a wide range of evolutionary timescales. Currently, CNE identification, and therefore GRB identification, hinges tightly on the selection of a conservation threshold at which a conserved region is defined as a CNE. For the species comparisons used in Harmston et al. 2017, the thresholds used ranged from 98% sequence identity over 50bp for the human - dog comparison (separated by 96 million years) to 70% identity over 30bp for human - spotted gar (separated by 435 million years). The reduction in the stringency of the thresholds was required to account for the fact that CNEs also diverge over time, albeit slowly. While the boundaries of predicted GRBs were robust to the CNE identification threshold used for relatively distant genome comparisons, in closely related species the approach breaks down because the neighbouring neutrally evolving sequence has not diverged enough to be able to non-arbitrarily define CNE identification thresholds. Due to the resulting increasing average length of conserved sequences between closely related species, it is often necessary to choose very long thresholds for the minimal CNE length ($>$ 400bp), thereby casting doubt on the biological relevance of comparing the distribution of such elements with those identified in distant comparisons. In this chapter I address this problem by defining and exploring a threshold-free measure of pairwise sequence conservation based on the kurtosis of the distribution of the lengths of all sequences perfectly conserved between two genomes.

Karl Pearson defined kurtosis as a measure of how flat or peaked the top of a symmetric distribution is (Pearson 1905). The kurtosis of a distribution is actually more influenced by scores in the tails of the distribution than the centre of the distribution, and thus distributions with a high kurtosis can be considered "fat-tailed" (DeCarlo 1997). Kurtosis has also been defined as the "location- and scale-free movement of probability mass from the shoulders of a distribution into its centre and tails" (Balanda and Macgillivray 1988).

I use kurtosis to measure the effect of the number of extreme observations on the distribution of the lengths of runs of perfect sequence identity between two genomes. I show that this measure is highly correlated with CNE density and can be effectively used to predict high quality GRBs for the species comparisons used in Harmston et al. 2017. Further, I use this kurtosis-based measure to predict GRBs between human and non-human primates and show that it is superior to CNE density at these short evolutionary distances.The ability of my method to detect GRBs across close evolutionary distances, without the requirement for arbitrary conversation thresholds, will enable the study of GRB turnover and the detection of recent lineage-specific changes in gross GRB structure.

2.2 Methods

2.2.1 Pairwise Genome Alignment

Pairwise genome alignments from human (hg19) to all species used in this analysis were retrieved from the UCSC Genome Browser (Kent et al. 2002) with the exception of human to spotted gar (LepOcu1).

The human to spotted gar alignment was produced using LASTZ (Harris 2007). The HoxD55 nucleotide substitution matrix was used for penalising alignment mismatches and all other parameters were set to the default. This alignment was generated by Ge Tan.

The species used were chosen based on their genome assembly quality and to facilitate testing the kurtosis-based conservation measure at multiple evolutionary time scales. While the spotted gar genome assembly is not of the same quality as the other species used, it holds a key position in the tree of life, having diverged from teleost fish shortly before they underwent whole genome duplication (Hoegg et al. 2004; Amores et al. 2011). Conservation analysis between human and spotted gar therefore avoids the potential complexities introduced by a genome duplication event and its aftermath.

2.2.2 CNE Identification

CNEs were identified by scanning the pairwise net whole-genome alignments for regions of high identity over a defined length. The alignments were filtered for known repeat regions and exons prior to scanning. Each net alignment was scanned twice, using each species in turn as a reference. The regions identified by each scan were then merged. The merged regions were aligned to the human genome using BLAT (Kent 2002), and any regions which mapped to more than four sites in the genome were removed as potentially unannotated repeats. CNE density across the genome was calculated by running a 300kb sliding window across the genome, in 1kb increments, and calculating the number of CNEs in each window. CNE identification was performed using the `CNEr` package in `R` (Tan 2015).

The minimum length and identity thresholds for CNE identification must be adjusted for each species comparison due to the continuous divergence of CNEs since the last common ancestor of the two species being compared. The identification thresholds used for each species comparison are listed in Table 2.1.

Table 2.1: Thresholds used for CNE identification in the human genome

Query Species (Genome Assembly)	Minimum Identity (%)	Minimum Length (bp)	Divergence Time (million years)
Gorilla (gorGor3)	100	400; 600	8.6
Rhesus monkey (rheMac3)	99.3; 100	150	29.4
Dog (canFam3)	80; 96; 100	50	96
Opossum (monDom5)	80; 96; 100	50	159
Chicken (galGal4)	80; 90; 98	50	312
Spotted gar (LepOcu1)	70; 80; 96.6	30	435

2.2.3 CNE-based GRB Identification

CNE-dense regions of the genome were identified using an unsupervised two-state HMM which partitions the genome into high and low CNE density regions (as described in Harmston et al. 2017). In brief, the genome was segmented into high- and low-density regions, and those CNEs within the high-density regions, which were separated by less than a predefined genomic distance, were merged to form blocks. This merging continues until the gaps between blocks are greater than a specified quantile of the widths of gaps between all adjacent CNEs. The quantile was set for each species comparison based on the how well the predicted GRB boundaries recapitulated a set of known GRB boundaries. Human - rhesus monkey and human - gorilla GRBs were generated for this project, while human - opossum GRBs were previously generated by Nathan Harmston (Harmston et al. 2017).

2.2.4 Genome-wide Kurtosis Calculation

For each species comparison, the kurtosis of the distribution of the lengths of all identical sequences was calculated in bins across the genome. Initially, all runs of 100% sequence identity were extracted from the pairwise whole-genome alignment and filtered for annotated repeats and exonic sequences. The genome was then divided into 30kb bins and the lengths of all runs of identity within each bin were calculated. 30kb was selected as a window size as this is the window size previously used for CNE density calculation, thereby maximising the comparability of the two approaches. The kurtosis of the distribution of lengths in each bin was then calculated as follows:

$$R(F) = \frac{\Big(q_{0.99}(F) - q_{0.01}(F)\Big)}{G_{50}}$$

where F is the distribution of the lengths of runs of perfect sequence identity in a bin, and G_{50} is the range of the middle 50% of the distribution of lengths of all runs of identity from all bins; calculated as follows:

$$G_{50} = q_{0.75}(J) - q_{0.25}(J)$$

where J is the distribution of the lengths of runs of perfect sequence identity across the whole genome.

For each bin, $R(F)$ is a ratio of the range of 99% of all lengths of runs of identical sequence, in a bin, to the range of 50% of all lengths of runs of identity for the whole genome. In practice it measures the number, and extremity, of long runs of perfect identity, in each bin, compared to the background conservation for the whole genome. This is an adaptation of the robust kurtosis measure proposed in Ruppert 1987.

2.2.5 Correlation of Kurtosis and CNE density

Maximum kurtosis and CNE density were calculated in 90kb windows across the genome, with 1000 windows derived from previously defined human - opossum GRBs and 1000 from non-GRB regions. This was performed for human to dog, chicken and spotted gar comparisons at each CNE identification threshold listed in Table 2.1. The Spearman's correlation between maximum scores in each window was then calculated. For the purpose of visualisation, a linear model was fitted to the data for each comparison at each CNE identification threshold.

2.2.6 Kurtosis-based GRB Identification

Kurtosis-based GRBs were generated by using the change point modelling (CPM) approach to identify change points in the binned kurtosis data, indicating a shift to higher mean kurtosis values (Ross 2015). Under this framework, kurtosis values in bins across the genome are treated as a series of n independent observations $x_1, ..., x_n$. The assumption that all observations derived from a genomic window are identically distributed, according to an undefined distribution F_0, can then be tested by choosing between the following hypotheses:

$$H_0 : X_i \sim F_0(x; \theta_0), i = 1, ..., n,$$

$$H_1 : X_i \sim \begin{cases} F_0(x; \theta_0), i = 1, 2..., k, \\ F_1(x; \theta_1), i = k + 1, k + 2, ..., n, \end{cases}$$

where θ_i represent the unknown parameters of each distribution. In this scenario the two distributions F_0 and F_1 represent the distribution of values coming from non-GRB and GRB regions of the genome respectively. The presence of a change point can be tested using a two-sampled Mann-Whitney test and the null hypothesis rejected if the test statistic exceeds a predefined cut-off. For a series of observations $x_1, ..., x_t$ the test statistic is calculated at every x_k, for $1 < k < t$, and the maximum test statistic obtained for all values of k is used. As successive observations are made (successive windows along the genome), the test statistic is calculated again at every x_k, but now for $1 < k < t + 1$. If no significant change point is detected, the next observation, x_{t+2}, is received and the testing is performed again on $x_1, ..., x_{t+2}$. However if a change is detected at x_k, the process begins again with x_{k+1} as the first observation in the new series of observations to be tested. For further details refer to Ross 2015. This analysis was performed using the `cpm` package in `R`, and the ARL0 parameter was set to 370. This is the least stringent ARL0 value implemented in the package and favours detection of more potential change points at the risk of including more false positives. Greater sensitivity combined with a merging step (described below) was preferred to stringent change point detection which potentially misses GRB boundaries.

Once significant change points in the binned kurtosis values have been identified, these are treated as potential GRB boundaries. The mean kurtosis within each range is then calculated, and adjacent ranges are merged if the mean kurtosis in both is above a specified percentile of all binned kurtosis values. The percentile used was determined empirically based on the predicted GRBs ability to recapitulate known GRB boundaries. For all species comparisons used, the quantile used was 0.7.

2.2.7 Hi-C Directionality Index within GRBs

hESC and IMR90 Hi-C data were obtained from the Gene Expression Omnibus (GEO Accession: GSE35156) and processed as described in Harmston et al. 2017. In brief, reads were aligned to hg19 using bowtie (Langmead et al. 2009) and aligned reads were binned into 40kb bins. The directionality index of each bin was then calculated as follows:

$$DI = \left(\frac{B-A}{|B-A|}\right)\left(\frac{(A-E)^2}{E} - \frac{(B-E)^2}{E}\right)$$

where A is the number of reads that map from a given 40kb bin to the 2Mb region upstream of the bin, B is the number of reads that map from a given bin to the 2Mb region downstream of the bin, and E is the expected number of reads mapping up and downstream of the bin. Under the null hypothesis $E = \frac{(A+B)}{2}$. This method was first proposed in Dixon et al. 2012. Nathan Harmston processed all Hi-C datasets used and produced the corresponding directionality index data.

To visualise how well kurtosis-based GRBs recapitulate TAD boundaries, heatmaps of genomic windows centred around GRBs and ordered from largest to smallest GRBs were produced. Each window was then divided into 500 bins and the average DI within each bin was plotted.

2.2.8 Data visualisation

All plots were produced using a combination of the `ggplot2` and `genomation` packages in `R` (Akalin et al. 2015).

2.3 Results

2.3.1 CNE identification

The first step in this analysis was to identify CNEs between human and a range of species chosen to represent distinct vertebrate lineages. For each species comparison I used multiple CNE identification thresholds to facilitate a comprehensive comparison between the proposed kurtosis-based conservation measure and CNE density. The results of CNE identification are presented in Table 2.2.

Table 2.2: CNE sets generated in Chapter 2

Query Species (Genome Assembly)	Identification Threshold	Number of CNEs	Mean Width (bp)	Divergence Time (million years)
Gorilla (gorGor3)	100% over 400bp	105,187	488.3	8.6
..	100% over 600bp	10,922	721.8	8.6
Rhesus monkey (rheMac3)	99.3% over 150bp	148,757	208.8	29.4
..	100% over 150bp	48,601	199.9	29.4
Dog (canFam3)	80% over 50bp	3,763,684	104	96
..	96% over 50bp	312,516	95	96
..	100% over 50bp	99,079	77.5	96
Opossum (monDom5)	80% over 50bp	280,065	120.1	159
..	96% over 50bp	53,743	101.4	159
..	100% over 50bp	21,694	80.8	159
Chicken (galGal4)	80% over 50bp	67,325	149.1	312
..	90% over 50bp	39,400	132.4	312
..	98% over 50bp	18,296	102.6	312
Spotted gar (LepOcu1)	70% over 30bp	60,345	77.4	435
..	80% over 30bp	33,172	81.5	435
..	96.6% over 30bp	11,898	54.6	435

As expected, the number of CNEs identified for each species comparison decreases as the stringency of the threshold increases. The mean width of the elements identified also decreases as the minimum required identity is increased. In general, the stringency of the threshold used for CNE identification is reduced as the evolutionary distance between the species compared increases. This is to account for the continual sequence divergence, in conserved regions, during the time that the two genomes have been evolving independently. This is due to increasing difficulty in identifying conserved regions as the amount of time the two genomes have been evolving independently increases. The effect of this sequence divergence is clearly discernible from the number of CNEs identified in dog, opossum, chicken and spotted gar at 80% identity over 50bp (30bp in spotted gar). The divergence time between human and each of these species ranges from 96 - 435 million years, and with the increasing time so the number of CNEs identified drops from 3,763,684 to just 33,172. I produced CNE density tracks for each of these CNE sets for comparison to kurtosis-based conservation and visualisation.

2.3.2 Comparing kurtosis-based conservation to CNE density

Next, for each of the same species comparisons, I calculated the kurtosis of the distribution of the lengths of perfectly conserved sequences in bins across the genome. Figure 2.1A shows the distribution of kurtosis values across the genome for each comparison. The distributions are very similar across species comparisons, illustrating that the results of the method are comparable for multiple evolutionary distances. In the closer species comparisons (gorilla to opossum) the distributions of kurtosis values are centred on 4.5. This is the kurtosis of the negative binomial distribution NB(1, 0.222) (after dropping all zero observations), and therefore in the majority of bins, 99% of identical sequences are shorter than 18bp. As the evolutionary distance of the comparison increases, so the number of bins containing a value of zero increases, and the median kurtosis value drops. The increasing number of zero bins is due to an increasing numbers of bins that do not contain any alignable sequence. This trend shows that, as expected, with increasing evolutionary distance there will be larger portions of the genome that are unalignable due to continual sequence divergence. It is also striking that the range of the kurtosis values increases with evolutionary distance. This trend reflects the potential for more extreme outliers relative to the genomic background in more distant comparisons. These extreme outliers are the CNEs that we normally identify using traditional CNE identification approaches described in section 2.2.2. Another notable feature of the distributions is the increased dispersion with increasing evolutionary distance. This is most likely due to the increased variability in the number and length of runs of sequence identity from bin to bin in the more distant comparisons.

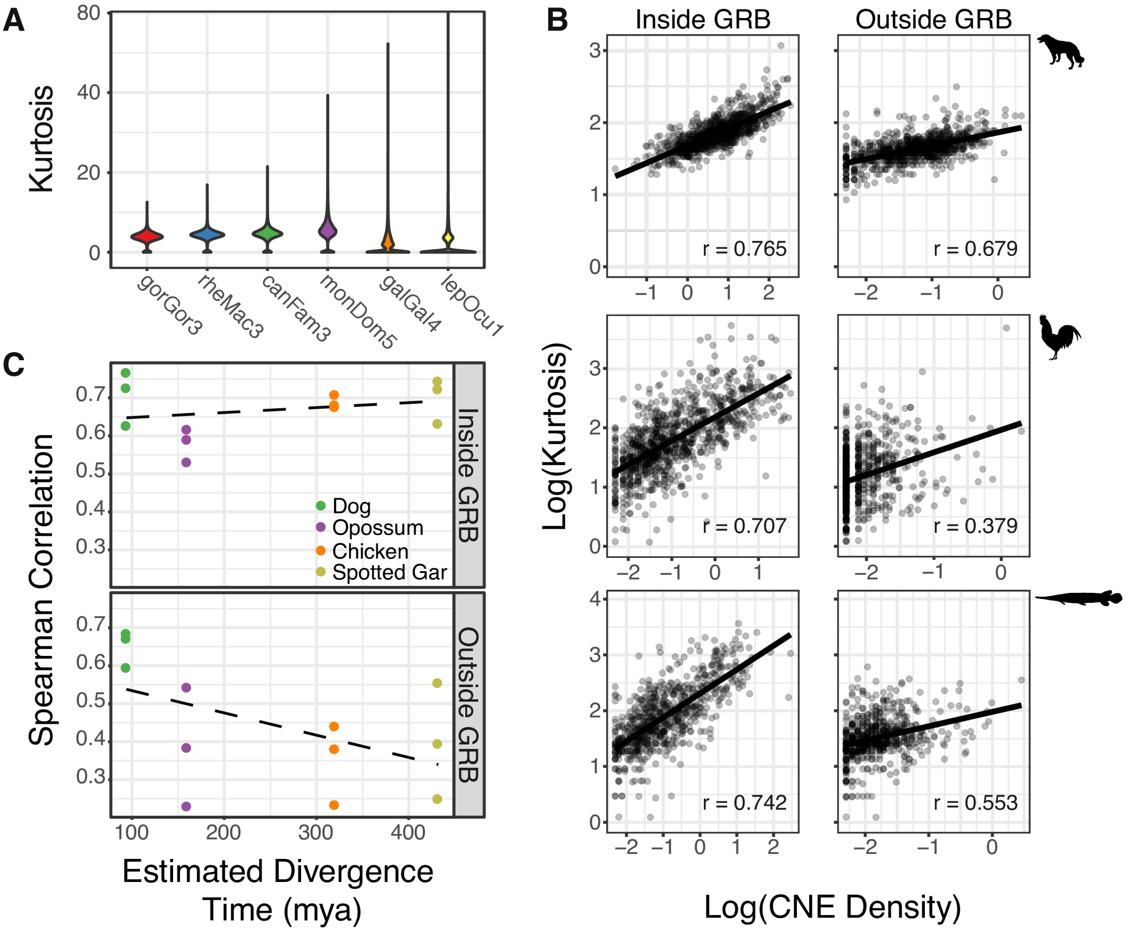

Figure 2.1: Kurtosis and CNE density are highly correlated.
(**A**) The distribution of binned kurtosis values across the human genome for all species comparisons. (**B**) The correlation between CNE density and kurtosis inside and outside of CNE-based GRBs for the human to dog, chicken and spotted gar species comparisons. (**C**) The correlation of CNE density and kurtosis, inside and outside of GRBs, as a function of evolutionary distance. Each of the three points for each species represents the correlation between kurtosis values and CNE-density at a separate CNE conservation threshold. CNE density and kurtosis are highly correlated within GRBs, regardless of the evolutionary distance of the comparison. Outside of GRBs there is a decreasing linear relationship between the correlation and the evolutionary distance of the comparison.

To compare kurtosis and CNE density across the genome, I sampled random 90kb windows from previously defined CNE-based human - opossum GRBs, and non-GRB regions of the genome. I then calculated the maximum kurtosis and

CNE density in each window. I repeated this using CNE density calculated at multiple thresholds for each species comparison. Next, for each species comparison I calculated the Spearman's correlation coefficient between kurtosis and CNE density for each species comparison inside and outside of GRBs. There is a strong correlation between kurtosis and CNE density, and this correlation is greater within GRBs than outside GRBs (Figure 2.1B). This trend is confirmed in Figure 2.1C, which shows the Spearman's correlation coefficient between kurtosis and CNE density, calculated for all CNE identification thresholds used for each species comparison. It is striking that regardless of the evolutionary distance of the comparison, kurtosis and CNE density values are similarly correlated within GRBs, whereas outside of GRBs it appears that the correlation drops with increasing evolutionary distance. The reduced correlation outside of GRBs may be caused by multiple properties of kurtosis and CNE density:

1. Outside of GRBs, CNE density is consistently either zero or close to zero, while kurtosis fluctuates around 4.5 from bin to bin, thereby reducing the correlation. Within GRBs, both the CNE density and kurtosis will be high in the majority of bins.

2. Outside of GRBs, there may be stretches of identical non-coding sequence which are shorter than the minimum length of the threshold used for calling CNEs, and are therefore not identified. However, these stretches will still result in distributions with relatively high kurtosis, but low CNE density. Within GRBs there are many identifiable CNEs and thus both CNE density and kurtosis will be high, resulting in a stronger correlation.

Overall, the consistency of the distribution of kurtosis values for species comparisons spanning vastly different evolutionary timescales, and its high correlation with CNE density in conserved regions of the genome, suggest that the kurtosis of the lengths of runs of sequence identity can be used as an effective threshold-free proxy for sequence conservation.

2.3.3 Kurtosis-based GRB identification in moderately to distantly related species

In the past, GRB identification has succeeded for moderate to distant evolutionary comparisons because the CNE density across the genome forms discrete peaks that are easily distinguished from the genomic background (Engström et al. 2007; Kikuta et al. 2007a; Akalin et al. 2009; Harmston et al. 2017). To test how well the kurtosis-based measure of conservation can discriminate highly conserved regions of the genome from non-conserved regions, I used binned kurtosis values to identify GRBs from human to moderately and distantly related species which have previously been used for CNE-based GRB prediction (Harmston et al. 2017). I used the CPM framework (described in Ross 2015) to identify shifts in the mean and variance of kurtosis across the genome, and treated these change points as potential GRB boundaries (detailed description in section 2.2.6). I then assigned the regions between these boundaries as either high or low kurtosis, based on their mean kurtosis value, and defined a final set of GRBs. This was performed for human to dog, opossum, chicken and spotted gar and the number and size of GRBs identified for each comparison are presented in Table 2.3

Table 2.3: Kurtosis-based GRBs identified in the human genome

Query Species	Number of GRBs	Average Width (kb)
Dog	559	1,233.1
Opossum	487	1,195
Chicken	426	978.7
Spotted gar	400	804.8

The number of GRBs identified in each comparison is similar, but there is a slight decrease in total GRBs as the evolutionary distance of the comparison increases. The average width of the identified GRBs also decreases with increasing evolutionary distance. The decreasing average width reflects the erosion of sequence conservation over time at the boundaries of GRBs, making accurate prediction of true GRB boundaries difficult over large evolutionary distances. This effect is also observed in GRBs identified using CNE density and has been previously described (Harmston et al. 2017). The decreasing number of GRBs may be due to the identification of relatively rapidly evolving GRBs in the closer comparisons that are not identifiable in the more distant comparisons. The differences in turnover rate of GRBs will be addressed in detail in Chapter 3.

As an initial assessment of the quality of the identified GRBs, I visualised CNE density within genomic windows centred on the kurtosis-based GRBs for each species comparison (Figure 2.2). GRBs were ordered by width, and thus any feature which is enriched within GRBs forms a characteristic funnel pattern. From these heatmaps, it is immediately apparent that there is a very strong enrichment of CNE density within kurtosis-based GRBs, and that this enrichment is robust to the CNE identification threshold used. Interestingly, as the stringency of the CNE identification threshold is increased, there are an increasing number of GRBs that contain no enrichment for CNE density. This likely reflects the ability of the kurtosis-based measure to identify runs of non-coding identity that fail to pass the more stringent CNE identification thresholds.

Previously, it has been shown that the borders of GRBs predicted using CNE density are robust to the species comparison used to predict the GRBs (Harmston et al. 2017). To investigate how robust the kurtosis-based GRB boundaries are, I plotted a heatmap for each set of GRBs in which genomic windows were centred on GRBs and coloured based on the number of sets in which each region of the window is found inside a GRB (Figure 2.3). The boundaries of kurtosis-based GRBs are very similar between species comparisons. Each funnel in Figure 2.3 is overwhelmingly either orange or red, indicating that these regions were predicted to be within a GRB in three or four of the species comparisons. For the human - dog and human - opossum GRBs, there is a clear white space immediately outside the

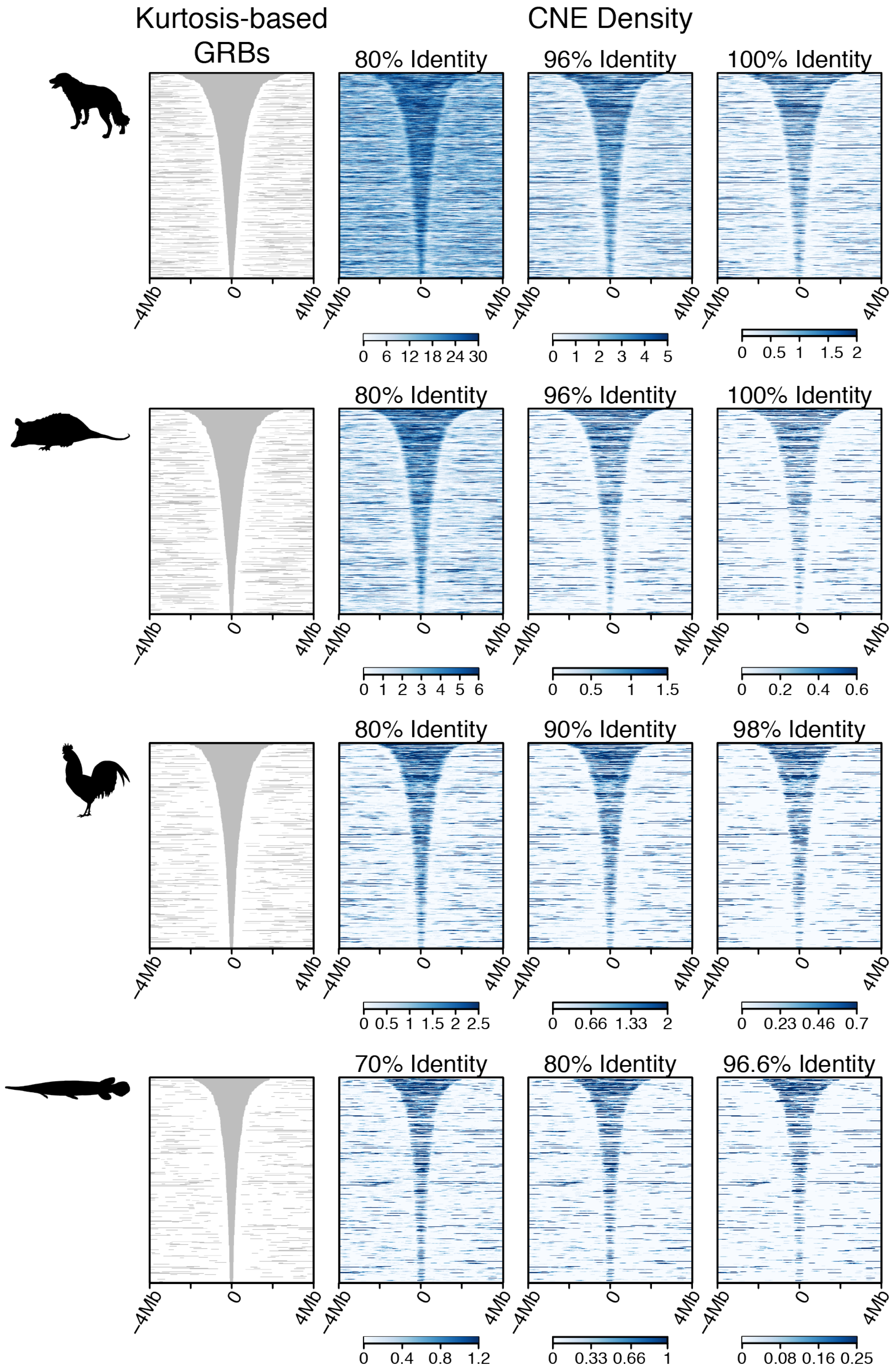

Figure 2.2: CNE density across kurtosis-based GRBs. GRBs were predicted from human to dog, opossum, chicken and spotted gar using the kurtosis-based measure of conservation. The heatmaps in the first column show the extent of the predicted GRBs, while the next three columns show the density of CNEs in each genomic window. Above each heatmap is the minimum sequence identity threshold of the CNEs visualised in that heatmap. For each species comparison the kurtosis-based GRBs are CNE-dense at every CNE identification threshold used.

funnel indicating that the majority of predicted GRB boundaries are consistent for these comparisons. The increased amount of blue and yellow immediately outside of the human - chicken and human - spotted gar GRB boundaries suggests that, for these species comparisons, the predicted GRBs are slightly narrower than for the closer species comparisons. This is also reflected in the mean width of GRBs identified for these species comparisons (Table 2.3). There also appears to be a relationship between the evolutionary distance of the comparison and the proportion of the GRBs identified in all four of the species comparisons, with almost all of the human - spotted gar GRBs being identified in all four sets. This observation reflects the effects of regulatory turnover on our ability to successfully identify GRBs (discussed in detail in Chapter 3).

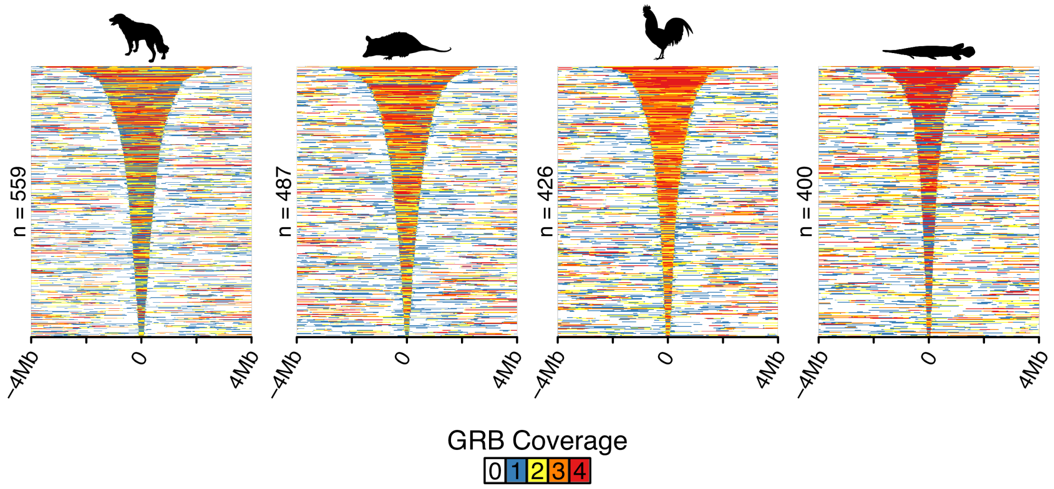

Figure 2.3: Overlap of all kurtosis-based GRB sets. Genomic windows were centred on kurtosis-based GRBs (identified from human to dog, opossum, chicken, and spotted gar), and regions within the windows were coloured based on the number of sets in which they were predicted to be within a GRB. The number of GRBs identified for each comparison is shown on the side of the appropriate heatmap.

Since the kurtosis-based GRBs are CNE-dense, there should be significant overlap between these sets of GRBs and those predicted using CNE density. To investigate this I plotted genomic windows centred on the predicted GRBs, with regions coloured by whether they appear within a GRB in both methods, or are unique to either one. I did this for GRBs derived from each method and species comparison in turn (Figure 2.4). There is a very high degree of overlap between the kurtosis-based and CNE-based GRBs, with the entirety of almost every kurtosis-based GRB overlapping a CNE-based GRB in the human to dog, opossum and chicken comparisons. There is a significantly higher proportion of unique GRBs in the human to spotted gar comparison than any of the other comparisons, further highlighting the increased difficulty in accurately defining GRBs at large evolutionary distances. These results are further summarised in Table S1. Another characteristic of the kurtosis-based GRBs visible from these heatmaps is that, in general, they are narrower than the CNE-based GRBs. This is evidenced by a green stripe outside of the red funnels in Figure 2.4A, and is most apparent in the human to chicken GRBs. Looking at the CNE-based GRBs in Figure 2.4B, it is clear that there are many CNE-based GRBs that are not identified as GRBs using the kurtosis-based measure, and that this proportion increases with the evolutionary distance of the comparison.

These results show that kurtosis-based GRB prediction successfully identifies a CNE-dense set of GRBs which appear to be a high-confidence subset of the previously defined CNE-based GRBs. The boundaries of these GRBs are robust to the species comparison used to predict them, but are on average narrower than the CNE-based boundaries.

To further evaluate the accuracy of the kurtosis-based GRB boundaries, I took advantage of the fact that GRB boundaries frequently coincide with TAD boundaries (Harmston et al. 2017) by plotting the Hi-C DI from hESC and IMR90 cells within the same GRB-containing genomic windows as in (Figure 2.5). In these plots the intensity of red and blue in a region show the frequency with which this region interacts with downstream and upstream loci respectively. Visualised this way, TADs appear as a span of red followed by a span of blue. For GRBs defined from human to dog, opossum and chicken, there is a very clear funnel present in the DI heatmaps in both cell types (Figure 2.5). The funnels have a well defined

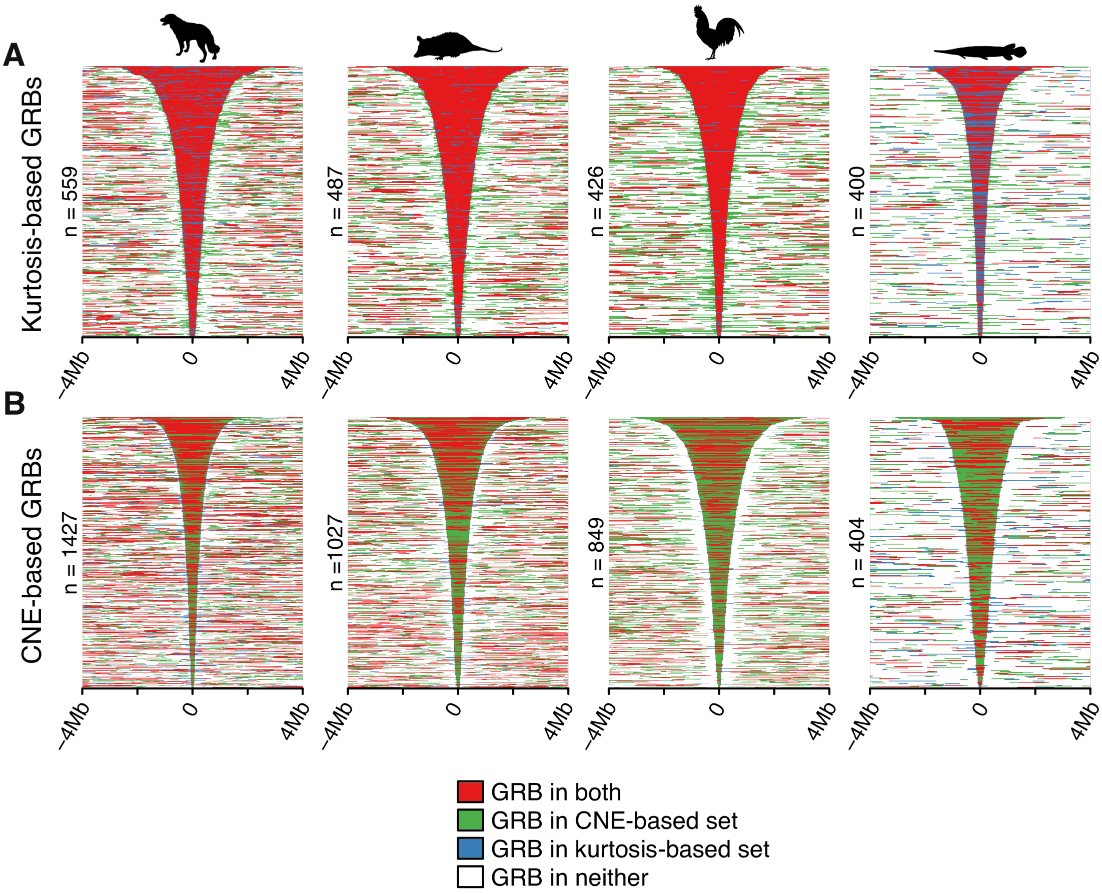

Figure 2.4: Overlap of kurtosis-based and CNE-based GRBs. For GRBs defined from human to dog, opossum, chicken and spotted gar, genomic windows were centred on the predicted GRBs and the regions within the windows were coloured by whether they were predicted to occur within a GRB by either the kurtosis-based method, CNE-based method, or both. (**A**) Kurtosis-based GRBs. (**B**) CNE-based GRBs.

red boundary followed by a well defined blue boundary, indicating that the GRBs coincide well with TADs. There is no visible funnel in the human to spotted gar GRBs, with only a hint of a funnel visible in the very largest GRBs, many of which are also the most strongly conserved. While these GRBs clearly do not coincide with TADs, it is possible that at these evolutionary distances, the kurtosis-based conservation measure is only identifying the core, highly conserved regions of each GRB and thus underestimating their true extent. Based on the concordance between the kurtosis-based GRB predictions and the CNE density for all species comparisons, it is likely that CNE-based GRB prediction will suffer from the same problem.

Taken together, the concordance between kurtosis-based GRB predictions

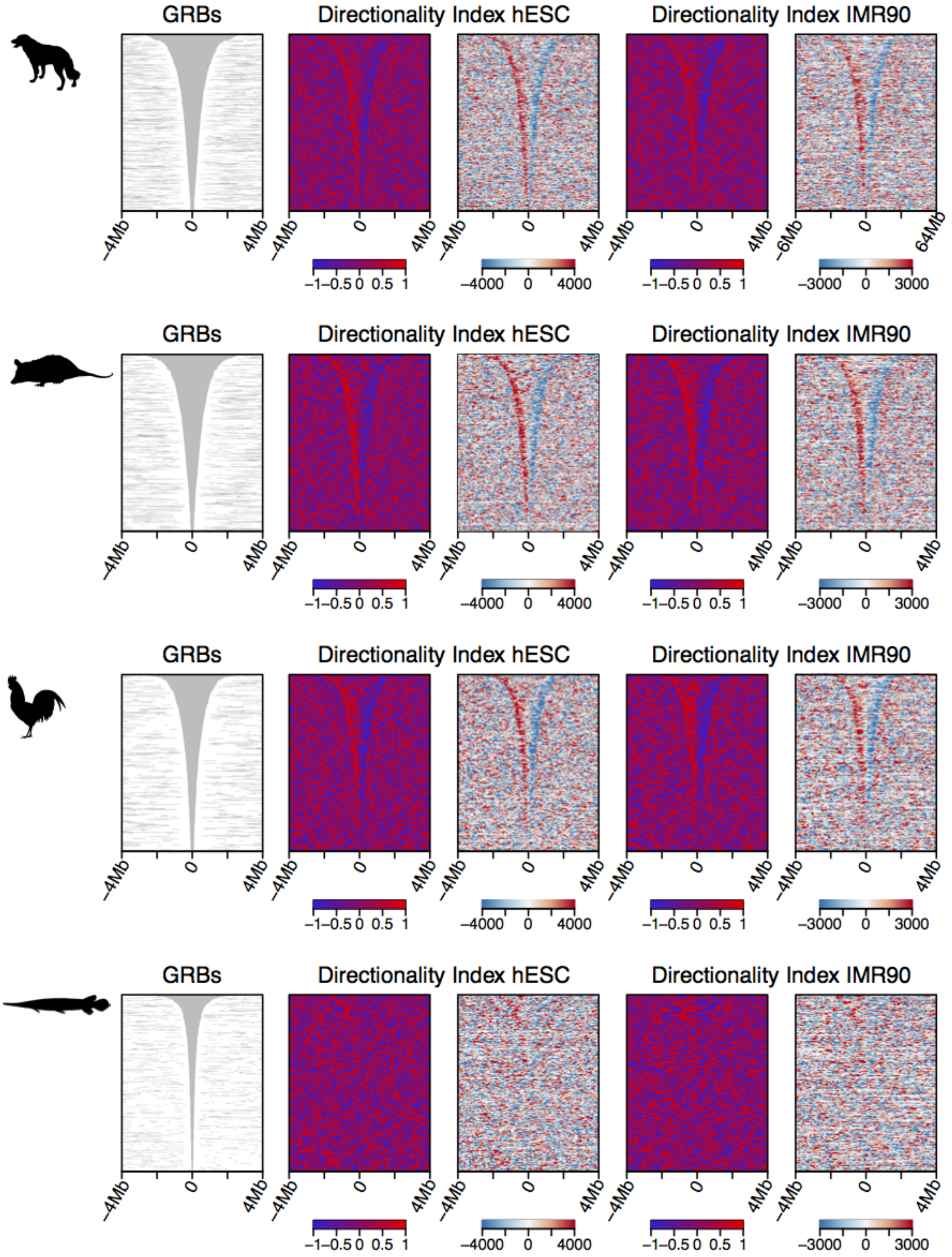

Figure 2.5: Hi-C directionality index within kurtosis-based GRBs. The heatmaps in the first column show the extent of the kurtosis-based GRBs for each species comparison, while the next 4 columns show the Hi-C directionality index, derived from hESC and IMR90 cells, in each genomic window. For each cell type, the first column shows the binarised DI, representing the direction of the interaction bias, while the second column shows both the direction and the strength of the interaction bias. Strongly positive (red) values indicate that a region preferentially interacts with downstream regions, while strongly negative (blue) values indicate a preference for upstream interactions.

and CNE density, the high degree of overlap between kurtosis-based and CNE-based GRBs, and the strong correlation between GRB and TAD boundaries, suggests that kurtosis-based conservation can be used to accurately predict high quality GRBs.

2.3.4 Kurtosis-based GRB identification in non-human primates

CNE identification thresholds necessitate the implementation of an arbitrary cut-off for what is defined as a CNE and what is not. At the edge of the threshold, a single mismatch in two aligned sequences is sufficient for an otherwise highly conserved region to be declared non-conserved. In the context of GRB identification, this is seldom a problem for evolutionarily distant species comparisons, but at shorter evolutionary timescales it becomes increasingly difficult to determine how long a stretch of perfect sequence identity should be for the region to be declared a CNE. This is the context in which kurtosis-based conservation may see the most utility. By its nature, kurtosis-based conservation takes into account the background level of conservation for a particular species comparison, and only defines those regions with unexpectedly long runs of identity as highly conserved.

I predicted GRBs for human to two non-human primates, the rhesus-monkey and the gorilla, to test the limits of kurtosis-based conservation for GRB detection. Humans and rhesus monkeys (referred to as rhesus for the rest of the chapter) diverged approximately 30 million years ago, while humans and gorillas diverged only 8.6 million years ago. Using kurtosis-based conservation, I predicted 523 human - rhesus GRBs (mean width = 1,279.9kb) and 483 human - gorilla GRBs (mean width = 1,242.9kb). This is a similar number of GRBs to the sets identified to more distant vertebrates, with a slightly larger average width, suggesting that the method can predict sensible GRBs even at such short evolutionary timescales.

To assess the quality of these GRBs, I plotted CNE density and Hi-C DI across genomic windows centred on the kurtosis-based GRB predictions, as previ-

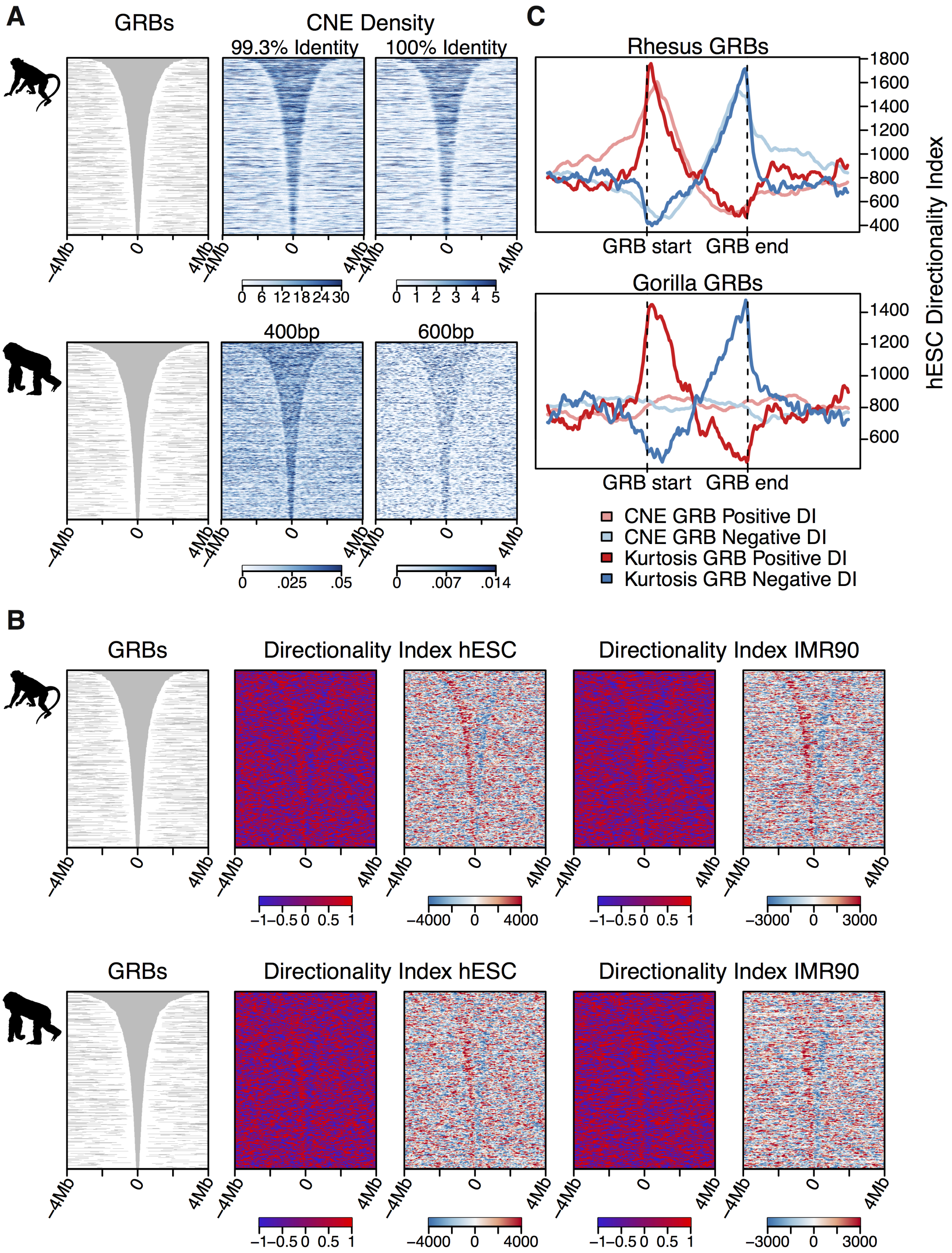

Figure 2.6: Kurtosis-based GRBs in primates. Kurtosis-based GRBs were identified from human to rhesus monkey and gorilla. Grey heatmaps represent the extent of the predicted GRBs. (**A**) CNE density and (**B**) Hi-C DI within genomic windows centred on predicted GRBs. (**C**) Average Hi-C DI strength within kurtosis- and CNE-based GRBs for human - rhesus monkey and human - gorilla.

ously described (Figure 2.6A-CB). For the human-rhesus GRBs there is a strong enrichment of CNE density within the predicted GRBs, indicating that for this species comparison kurtosis is a good proxy for conserved non-coding conservation. For the human to gorilla comparison there is also a visible CNE density enrichment within the predicted GRBs, but the strength of the enrichment is much reduced. The average mismatch rate between human and gorilla is only 1.75%, and therefore it is very surprising that there is any CNE density enrichment within the kurtosis-based GRBs (Scally et al. 2012). This result is strong evidence that kurtosis-based conservation can identify highly conserved regions of the genome. Examining the DI heatmaps, it is clear that the rhesus GRBs have a visible funnel, although it is not as strong as in the more distant comparisons. The largest rhesus GRBs have the weakest correspondence with the DI, and appear to span multiple TADs. These are probably physically close GRBs that have been merged by the GRB prediction. This may also account for the increased mean GRB width in this set. Separating adjacent synteny blocks using sequence conservation alone is a known difficulty in GRB prediction (Harmston et al. 2017), and kurtosis-based GRB prediction appears to suffer from the same issue. For the human-gorilla GRBs a similar issue is visible. The largest third of GRBs display no visible funnel in the DI heatmaps, however there is a noisy funnel visible in the rest of the GRBs. Overall these results suggest that kurtosis-based conservation can identify signatures of non-coding conservation in very closely related species, but that GRB boundary prediction in these comparisons is less precise than in the more distant comparisons.

Next I compared the kurtosis-based GRBs to CNE-based GRBs predicted as described in Harmston et al. (Harmston et al. 2017). I predicted CNE-based GRBs in human to rhesus and gorilla using a two-state HMM that partitions the genome into conserved and non-conserved regions (details in Section 2.2.3), and plotted the average Hi-C DI across the predicted GRBs from both sets (Figure 2.6C). The CNE-based GRB prediction yielded 744 human to rhesus GRBs with a mean width of 482.9kb and 2220 human to gorilla GRBs with a mean width of 504,4kb. The number of GRBs identified in human-rhesus is greater than for the other species comparisons used so far, but not exceedingly so. For the human-gorilla comparison, however, there were an unreasonable number of GRBs predicted. In Figure 2.6C, the average Hi-C DI is plotted across the predicted GRBs from both sets. We can clearly see that for the human - rhesus comparison the kurtosis-based GRBs have a much stronger peak of the positive and negative DI, at their starts and ends respectively, than the CNE-based GRBs. There is also a much sharper boundary effect in the kurtosis-based GRBs, with the peaks of DI spreading well beyond the boundaries of the CNE-based GRBs. In the human - gorilla comparison the kurtosis-based GRBs boundaries also coincide with peaks in the positive and negative DI, while the CNE-based GRBs show no enrichment of DI score at either boundary.

These results conclusively demonstrate that the kurtosis-based conservation measure can identify highly conserved regions of the genome, even in very closely related species, and that kurtosis-based GRB predictions recapitulate TAD boundaries better than the CNE-based GRB predictions at these evolutionary timescales.

2.4 Discussion

In this chapter I have defined a novel measure of pairwise sequence conservation based on the kurtosis of the distribution of the lengths of sequences perfectly conserved between two genomes. I have shown that the kurtosis-based measure is highly correlated with CNE density and can be used to generate high quality GRB predictions for moderate to distant species comparisons. I have also shown that kurtosis-based GRB prediction far outperforms CNE-based GRB prediction in closely related species. The identification of GRB-like structures between human and gorilla is a surprising result as previously it has been impossible to define conserved regulatory domains between such closely related species. Humans and gorillas share over 98% of their genome sequence, and so to be able to use sequence conservation to define regulatory regions that coincide with TADs is strong testament to my method's ability to account for the general background conservation between two genomes.

Most importantly, unlike CNE-based conservation analysis, my method works without needing the definition of a minimum length or sequence identity threshold required for a sequence to be considered conserved. Having a threshold-free approach for measuring conservation allows us to directly compare the results of species comparisons spanning a range of evolutionary distances. This feature, combined with the success in identifying GRB-like structures in extremely closely related species opens up the possibility of systematically investigating the evolutionary dynamics of GRBs in multiple closely related metazoan lineages, potentially yielding a greater understanding of the origin and evolution of long-range gene regulation in metazoan genomes.

Further, my method may have utility in the analysis of GRB developmental gene regulation in species that have undergone extreme genome compaction such as the puffer fish, Tetraodon nigroviridis, and the sea squirt, Oikopleura dioica. The tiny size of these genomes makes it very difficult to define the minimum length a stretch of conserved sequence should be to be considered a conserved element, and as described above, comparing the results of this analysis with those performed in larger genomes is problematic. My method may provide the ability to accurately define GRB boundaries in compact genomes and therefore deliver insights into the effects of genome compaction of long-range gene regulation.

Chapter 3

Regulatory turnover in metazoan GRBs

3.1 Introduction

GRBs, or clusters of CNEs, have been identified at functionally similar classes of genes in many metazoan lineages, including mammals, teleost fish, insects and nematodes (Akalin et al. 2009; Engström et al. 2007; Kikuta et al. 2007a; Vavouri et al. 2007). The similarity in the genomic distribution of CNEs in multiple phyletic groups suggests that they perform a common function in each group, however, the CNEs identified in different phlya share no sequence conservation. This lack of non-coding sequence conservation between phyla has previously been attributed to lineage-specific CNE recruitment. Vavouri *et al.* posited that the presence of unrelated CNEs around similar developmental genes in humans, *D. melanogaster* and *C. elegans* can be explained by the parallel evolution of long-range regulation of key developmental regulatory genes, resulting in "distinct yet parallel sets of CNEs" (Vavouri et al. 2007). Lowe *et al.* used a multiple sequence alignment of 40 vertebrate genomes to infer the evolutionary time point at which human CNEs came under selection, concluding that CNEs were recruited around specific functional classes of genes during three distinct periods of vertebrate evolution (Lowe et al. 2011). The first period was reported to have occurred from the last common vertebrate ancestor until approximately 300 million years ago, with CNEs being preferentially recruited around transcription factors and the developmental genes that they regulate. The second period ranged from approximately 300 million years ago until 100 million years ago, and CNEs were mostly recruited at cell signalling genes. The final wave was reported to have occurred in placental mammals less than 100 million years ago, with recruitment of CNEs taking place predominantly at post-translational modification genes involved in intracellular signalling (Lowe et al. 2011).

Contrarily, it been proposed that GRB-like gene regulation is an ancient feature of metazoan gene regulation, and that the lack of sequence conservation between CNEs identified in different lineages is due to gradual and continual CNE turnover within GRBs (Harmston, Baresic, and Lenhard 2013). This model is supported by the observations that conservation of gene expression between species is not dependent on conservation of its enhancers (Fisher et al. 2006; Hare et al. 2008), that small sequence changes are capable of driving both *de novo* gain and loss of enhancer function (Eichenlaub and Ettwiller 2011), and that there is pervasive turnover of functional non-coding sequences, occurring at different rates, within mammalian genomes (Meader, Ponting, and Lunter 2010). Under this model, given enough time, all CNEs would accumulate enough sequence changes that they would no longer be identifiable across lineages based on sequence conservation alone. Further, if sequence turnover is occurring at different rates around distinct functional classes of genes, this model can also explain what appears to be the three distinct periods of CNE recruitment described by Lowe *et al.* (Lowe et al. 2011).

In this chapter I assess the validity of the turnover model by comparing high- and low-turnover GRBs identified in three distinct metazoan lineages. Assuming that the turnover model is correct, GRBs have been undergoing sequence turnover at different rates since their initial recruitment, and therefore, there should be many shared features between GRBs which are high- or low-turnover in each lineage respectively. For this analysis I make use of the kurtosis-based measure of conservation, described in Chapter 2, as it is threshold-free and implicitly accounts for the background conservation of the species compared. This is particularly useful when predicting GRBs in closely related species, thereby providing better power to discriminate between the most and least deeply conserved GRBs. I find that similar classes of genes are regulated by high- and low-turnover GRBs in each lineage, and that these groups of GRBs are distinct from each other with respect to timing of target-gene expression during development, epigenetic state, and repeat content.

3.2 Methods

3.2.1 Species used in Chapter 3

For this chapter, GRBs were identified between a reference species and a number of query species in three separate phylogenies. Throughout the chapter the species are referred to by their genome assembly abbreviations. Table 3.1 lists the species used, their common name and their genome assembly abbreviations.

Table 3.1: Species used in Chapter 3

Species	Common name	Genome Assembly
Homo sapiens	Human	hg19
Macaca mulatta	Rhesus monkey	rheMac3
Canis familiaris	Dog	canFam3
Monodelphis domestica	Grey short-tailed opossum	monDom5
Gallus gallus	Chicken	galGal4
Meleagris gallopavo	Wild turkey	melGal1
Taeniopygia guttata	Zebra finch	taeGut2
Anolis carolinensis	Carolina anole lizard	anoCar2
Xenopus tropicalis	Western clawed frog	xenTro3
Drosophila melanogaster	Fruit fly	dm6
Drosophila ananassae	Fruit fly	droAna2
Drosophila pseudoobscura	Fruit fly	dp2
Drosophila mojavensis	Fruit fly	droMoj2

3.2.2 Pairwise genome alignment

Pairwise genome alignments from each of the three reference species (hg19, galGal4, and dm6) to all species used in this analysis were either retrieved from the UCSC genome browser (Kent et al. 2002) or generated using LASTZ (Harris 2007). The source of all alignments used in this chapter is listed in Table 3.2.

The three reference species were selected to be representative of three vastly different metazoan lineages, and therefore facilitate analysis of the conservation of the regulatory dynamics of GRBs. The species used in each phylogeny were selected such that each phylogeny spanned approximately equivalent evolutionary distance from the last common ancestor at the root of each tree. These species were also

Table 3.2: Pairwise alignments from the hg19, galGal4, and dm6 genomes

Reference Species	**Query Species**	**Divergence Time (million years)**	**Source**	**Generated By**
Human (hg19)	Rhesus monkey (rheMac3)	29.4	UCSC	UCSC
..	Dog (canFam3)	96	UCSC	UCSC
..	Opossum (monDom5)	159	UCSC	UCSC
..	Chicken (galGal4)	312	UCSC	UCSC
Chicken (galGal4)	Turkey (melGal1)	37	Lenhard group	Alexander Nash
..	Zebra finch (taeGut2)	98	UCSC	UCSC
..	Lizard (anoCar2)	280	Lenhard group	Ge Tan
..	Frog (xenTro3)	352	Lenhard group	Ge Tan
D. melanogaster (dm6)	*D. ananassae* (droAna2)	34	Lenhard group	Ge Tan
..	*D. pseudoobscura* (dp2)	37	Lenhard group	Ge Tan
..	*D. mojavensis* (droMoj2)	50	Lenhard group	Ge Tan

selected so that each comparison to the reference species spanned a unique evolutionary time in that phylogeny. While the *Drosophila* species diverged much more recently, due to their short generation time the evolutionary distance between the species is similar. For example, it has been estimated that the degree of divergence between humans and chickens and *D. melanogaster* and *D. pseudoobscura* is approximately equal (Stark et al. 2007), as measured by substitutions per fourfold degenerate site.

3.2.3 Kurtosis-based conservation

For each phylogeny, pairwise kurtosis-based conservation was calculated between the reference species and all other species in the phylogeny. This was performed as described in Section 2.2.4. In brief, the reference genome was binned and for each bin the lengths of sequences which were perfectly conserved between the reference and the query genome were extracted from the pairwise alignment. The kurtosis of the distribution of these lengths was then calculated. Bin sizes of 20kb were used for the two vertebrate reference species (hg19 and galGal4), while for dm6, 4kb bins were used. Different bin sizes were used to account for the differences in genome size between the vertebrates and dm6.

3.2.4 Kurtosis-based GRB prediction

For each phylogeny, kurtosis-based GRBs were predicted between the reference species and all other species in the phylogeny. GRBs were predicted as described in Section 2.2.6. In brief, the CPM framework was used to identify change points in the mean and variance of the binned kurtosis-based conservation values across the genome (Ross 2015). These change points were treated as potential GRB boundaries. Adjacent windows separated by a potential boundary were then merged if both windows had a mean kurtosis greater than a predefined quantile of the distribution of all kurtosis values across the genome. The merged windows were used as the final set of GRBs.

The quantile used for each reference species was determined by visual inspection of how well the GRB predictions recapitulated the boundaries of known GRBs. For hg19 the quantile used was 0.8, while for galGal4 and dm6, 0.7 was used.

3.2.5 Identification of high- and low-turnover GRBs

To identify a set of high- and low-turnover GRBs, the GRBs predicted between each reference species and its most closely related species were filtered for GRBs which were supported by identification in at least one other species in the phylogeny. This ensured that the analysis was performed on a robust set of GRBs by excluding any spuriously identified GRBs. Next, the kurtosis values calculated for each species comparison were quantile-normalised and for each bin in the remaining GRBs, the value for each species comparison was summed. GRBs were then ranked based on their maximum summed kurtosis and the top 20% of GRBs were defined as low-turnover GRBs, while the bottom 20% were defined as high-turnover. This was performed for each phylogeny in turn.

3.2.6 CNE density in high- and low-turnover GRBs

CNEs were identified between the reference species and each of the other species in the phylogeny. The CNE identification thresholds used for each species comparison are listed in Table 3.3. CNE density across each reference genome was calculated by running a sliding window across the genome and counting the number of CNEs in each window. For hg19 and galGal4 a 300 kb window was used, while for dm6 a 40 kb window was used to account for the differences in genome size. CNE identification was performed using the `CNEr` package in `R`. All heatmaps were produced using the `genomation` package in `R` (Akalin et al. 2015).

Table 3.3: CNE sets used for visualisation of CNE density within GRBs

Reference Species (Genome Assembly)	Query Species (Genome Assembly)	Minimum Identity (%)	Minimum Length (bp)	Divergence Time (million years)
Human (hg19)	Rhesus monkey (rheMac3)	99.3	150	29.4
..	Dog (canFam3)	98	50	96
..	Opossum (monDom5)	90	50	159
..	Chicken (galGal4)	80	50	312
Chicken (galGal4)	Turkey (melGal1)	100	150	37
..	Zebra finch (taeGut2)	100	50	98
..	Lizard (anoCar2)	80	50	280
..	Frog (xenTro3)	80	50	352
D. melanogaster (dm6)	*D. ananassae* (droAna2)	98	50	34
..	*D. pseudoobscura* (dp2)	98	50	37
..	*D. mojavensis* (droMoj2)	96	50	50

3.2.7 GRB boundary stability and TAD comparisons

To assess how consistently the boundaries of high- and low-turnover GRBs are predicted, I calculated the distance from the boundaries of each GRB predicted in the most closely related to the nearest boundary predicted in each other species in the phylogeny. These distances were then presented as cumulative distributions where each line represents the distance from the initial set of boundaries to the boundaries identified in that particular species. This was performed on each phylogeny in turn.

To compare the ability of high- and low-turnover GRBs to recapitulate

TAD boundaries, the distance from high- and low-turnover GRB boundaries to the nearest TAD boundary was calculated. Hi-C data for hg19 was retrieved and processed as described in Section 2.2.7. The dm6 Hi-C data was produced by Sexton *et al.* and retrieved from the GEO (GEO Accession: GSM849422)(Sexton et al. 2012). The data was processed using the same pipeline described in Section 2.2.7 and Harmston et al. 2017. All Hi-C data was retrieved and processed by Nathan Harmston.

3.2.8 High- and Low-turnover GRB gene ontology enrichment

All genes contained within high- or low-turnover GRBs were tested for enrichment of biological process (BP) and molecular function (MF) gene ontology (GO) terms against a background of all annotated protein coding and micro-RNA genes in the genome. The top 10 most enriched terms from high- and low-turnover GRBs for each phylogeny were then plotted together to illustrate the overlap of enriched terms between phylogenies. GO enrichment was performed using the `GOstats` package in `R` (Falcon and Gentleman 2007).

3.2.9 GRB target-gene expression in development

Genes within hg19 high- and low-turnover GRBs were filtered for genes predicted to be the targets of GRB regulation. GRB target-gene prediction was performed by Ge Tan using a random forest based machine learning approach (Tan 2018). In brief, the random forest was trained on a manually annotated set of 259 target and 830 bystander genes (Akalin et al. 2009). Random forest predictions were based on a set of 19 informative features of GRB target genes, the most important of which were the number and size of CpG islands overlapping the gene, the tissue specificity of the gene's expression and the CNE densities surrounding the gene in multiple species comparisons.

As GRB target gene predictions were only available for the human genome, ENSEMBL's ortholog predictions were used to identify the mm9 high- and low-turnover GRB target gene orthologs. This was performed using the `biomaRt` package for `R`.

GRB target gene expression during embryonic development was then analysed using the FANTOM5 CAGE mouse developmental time course (Forrest et al. 2014). The data was downloaded, processed and normalized using the standard pipeline detailed in the `CAGEr` package in `R` (Haberle et al. 2015). Promoters were identified across all developmental samples with a minimum requirement that each promoter have a normalized tag per million (tpm) value of greater than 1 in all samples to be included. A self organising map (SOM) was then used to split consensus tag clusters into groups based on their expression dynamics during mouse embryonic development. The enrichment or depletion of high- and low-turnover GRB target genes in groups corresponding to early development, late development and constant stable expression was then tested using a Fisher's exact test, using the distribution of all GRB target genes between clusters as the expected distribution.

3.2.10 Chromatin modifications at high- and low-turnover GRBs

Human and *D. melanogaster* chromatin modification data was downloaded from the Roadmap Epigenomics (Kundaje et al. 2015) and modENCODE (Roy et al. 2010) repositories respectively. Details of the datasets used are listed in Table 3.4.

The Roadmap Epigenomics data was downloaded as raw ChIP-seq read density across the genome. The modENCODE data was downloaded in a processed format in which the raw ChIP-seq read density had been normalised against input and presented as log of the fold change over input. In both cases, the average ChIP-

Table 3.4: Publicly available chromatin modification data used in Chapter 3

Data Type	Sample	Genome Assembly	Source	GEO
H3K27me3 ChIP-seq	Fetal brain	hg19	Roadmap Epigenomics	GSM806937
H3K4me1 ChIP-seq	Fetal brain	hg19	Roadmap Epigenomics	GSM806934
H3K4me3 ChIP-seq	Fetal brain	hg19	Roadmap Epigenomics	GSM669624
H3K27me3 ChIP-seq	20-24hr Embryo	dm6	modENCODE	GSM439443
H3K27ac ChIP-seq	20-24hr Embryo	dm6	modENCODE	GSM401423
H3K4me1 ChIP-seq	20-24hr Embryo	dm6	modENCODE	GSM439464
H3K4me3 ChIP-seq	20-24hr Embryo	dm6	modENCODE	GSM400673

seq signal for each chromatin modification (raw or normalised) was visualised across high- and low-turnover GRBs.

3.2.11 Repeat content of high- and low-turnover GRBs

RepeatMasker-annotated repeats were downloaded for the hg19 genome (*RepeatMasker Open-4.0.*). The average coverage of long interspersed nuclear elements (LINEs), short interspersed nuclear elements (SINEs) and long terminal repeats (LTRs), the three most prominent repeat families in human genome, was calculated in bins across ht- and ltGRBs and visualised as a metaplot. The difference between SINE density within ht- and ltGRBs was then quantified as the number of elements per kb in each GRB.

To assess the activity of repeats in the human fetal brain, repeats were overlapped with H3K27ac peaks retrieved from Yan et al. (Yan et al. 2016, GEO Accession:GSE63634). H3K27ac consensus peaks were identified by taking the intersection of peaks identified independently in each replicate. Those repeats that overlapped an H3K27ac consensus peak were considered active. The proportion of active repeats in both high- and low-turnover GRBs were compared to the proportion of active repeats in non-GRB regions.

3.3 Results

3.3.1 GRB identification in three distinct metazoan lineages

To assess the evolutionary dynamics of GRBs that have been evolving independently for millions of years, I identified GRBs in three reference species from distinct metazoan lineages. In each reference species, GRBs were predicted using genome-wide kurtosis-based conservation (described in Chapter 2) for multiple species comparisons. The species used were selected such that, within a phylogeny, each species comparison spans a unique evolutionary distance, and between phylogenies the species comparisons span comparable distances (Figure 3.1A). I selected hg19, galGal4 and dm6 as reference species as these species have well assembled genomes and abundant publicly available functional genomics data, allowing for further characterisation of the identified GRBs. The number, and average width of the GRBs identified for each species comparison are listed in Table 3.5.

Table 3.5: Kurtosis-based GRBs identified in hg19, galGal4, and dm6

Reference Species	Query Species	Number of GRBs	Average Width (kb)
hg19	rheMac3	501	763.2
..	canFam3	522	826.4
..	monDom5	444	895.6
..	galGal4	363	738.6
galGal4	melGal1	429	495.2
..	taeGut2	370	556.3
..	anoCar2	298	558.7
..	xenTro3	251	528.8
dm6	droAna2	250	130.6
..	dp2	221	125.7
..	droMoj2	223	126.1

In Table 3.5, the query species are ordered by increasing evolutionary distance from the reference species, and it is clear that there is a general trend towards identification of a greater number of GRBs in the shortest evolutionary comparisons. Under the turnover model this result is intuitive as the most recently diverged species will have had the least time to accumulate sequence changes in their GRBs, thereby

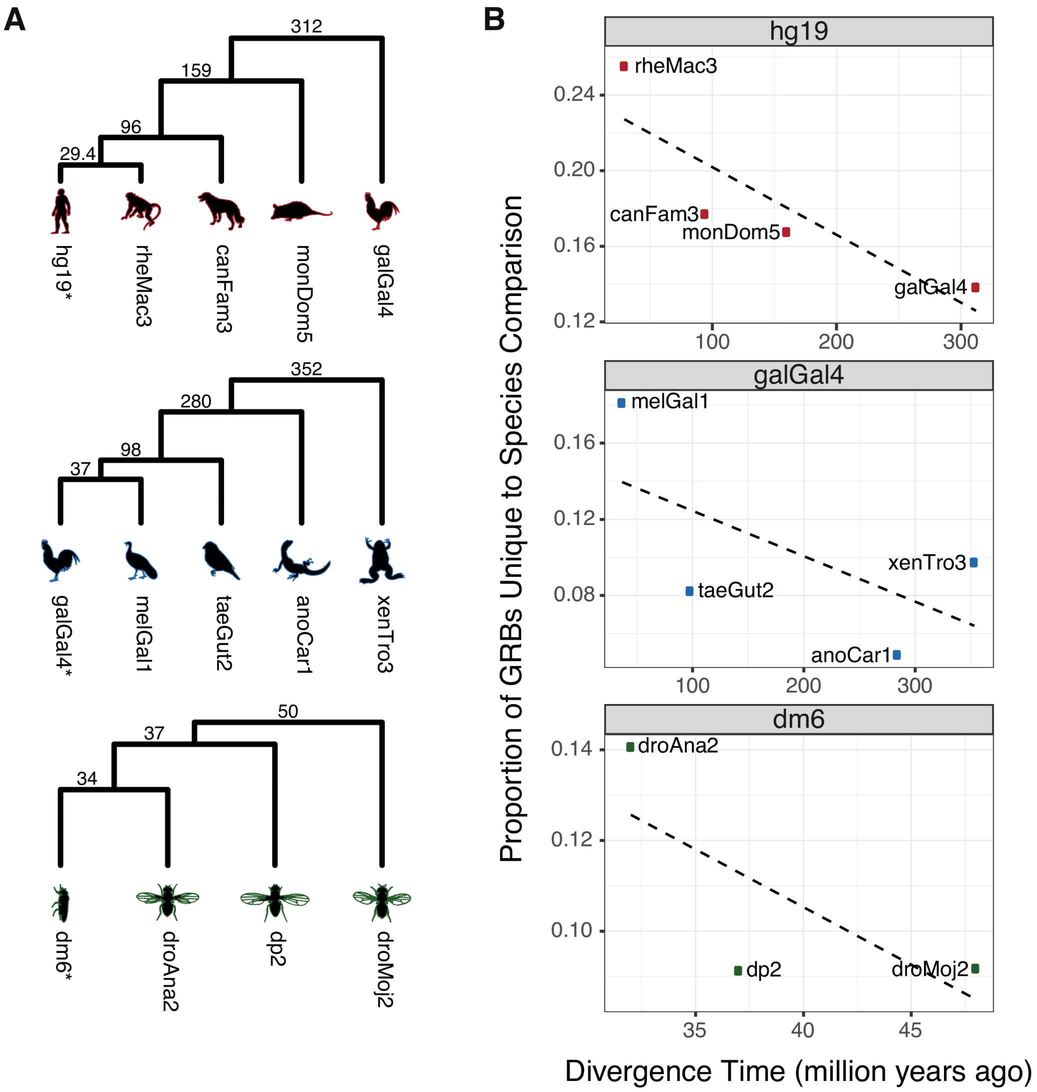

Figure 3.1: GRB identification in distinct metazoan lineages. (**A**) GRBs were identified for multiple species comparisons to three reference species (marked with an asterisk) from distinct metazoan lineages. For each species, the time since divergence from the reference species is shown in millions of years above the appropriate branch point. (**B**) The proportion of GRBs unique to a specific species comparison is related to the evolutionary distance of the comparison. GRB sets identified between closely related species tend to have a greater proportion of unique GRBs.

facilitating their identification.

The average size of the GRBs identified in each reference species is fairly consistent for each individual species comparison, however we see that the average size of the GRBs varies significantly between reference species. This is consistent with previous observations that the average size of GRBs identified in a species is approximately proportional to its genome size (Harmston et al. 2017).

Lastly, for each phylogeny I calculated the number of GRBs uniquely identified in each species comparison as a proportion of the total number of GRBs identified for that comparison. The results for each reference species are presented in Figure 3.1B. From this figure it is clear that there is an inverse relationship between the evolutionary distance of the species comparison and the number of GRBs that are uniquely identified in that comparison. The increase in the number of GRBs identified and the proportion of uniquely identified GRBs with decreasing evolutionary distance appears to be continuous, suggesting that sequence turnover within GRBs is continual and that CNEs have not been recruited in distinct bursts.

3.3.2 Identification of high- and low-turnover GRBs

To compare the features of GRBs that exhibit similar levels of conservation in independent lineages, I identified shallowly and deeply conserved GRBs in each phylogeny. Under the assumption that loss of conservation within a GRB is, in the majority of cases, due to turnover within regulatory elements, I define these as high- and low-turnover GRBs respectively. I assessed the depth of conservation of each GRB identified in the closest species comparison by summing the kurtosis-based conservation score from all species comparisons across its length. GRBs were then ranked based on their maximum summed kurtosis value, and the bottom 20% of GRBs were defined as high-turnover GRBs (htGRBs), while the top 20% were defined as low-turnover GRBs (ltGRBs). This was performed for each phylogeny separately. To confirm that each set of GRBs exhibited the expected conservation patterns, I visualised CNE density for each species comparison in genomic windows centred on ht- and ltGRBs respectively (Figure 3.2 and Figure S1).

From these figures we can see that, as expected, there is a clear enrichment of CNEs within ltGRBs regardless of the distance of the species comparison, while in htGRBs, the enrichment is lost in the more distant species comparisons. This is summarised in the form of a metaplot in Figure 3.3. Under the turnover model this difference in the depth of conservation of these two sets of GRBs would be explained by differing rates of sequence turnover within each set. These sets of high- and low-turnover GRBs are by no means exhaustive, as by using a quantile of a continuum of conservation values to group GRBs will inevitably exclude some GRBs that have similar conservation patterns but fall just outside the included quantile. However, for the purpose of this analysis, selecting the two extremes of a distribution of turnover rates should facilitate identification of potential shared characteristics or mechanisms which might explain the difference in rate of turnover at each set of GRBs.

3.3.3 High-turnover GRB boundary positions are less conserved than low-turnover GRBs

A clear difference between htGRBs and ltGRBs is their average width, visible as narrower funnels in Figure 3.2. In all three reference species, htGRBs are significantly narrower than ltGRBs (Figure 3.4A; hg19 $p = 2.29\text{x}10^{-13}$; galGal4 $p = 6.99\text{x}10^{-6}$; dm6 $p = 3.79\text{x}10^{-6}$). This may be due to the most deeply conserved GRBs targeting genes whose expression pattern is the most tightly regulated, or function in a larger number of contexts. These genes would require numerous enhancers to ensure the robust regulation of their expression, thereby extending the size of the GRB, as we observe here. This may also be due to non-coding conservation degrading more rapidly at the boundaries of htGRBs than ltGRBs. If conserved enhancers at the edges of GRBs gain mutations, thus reducing the conservation of

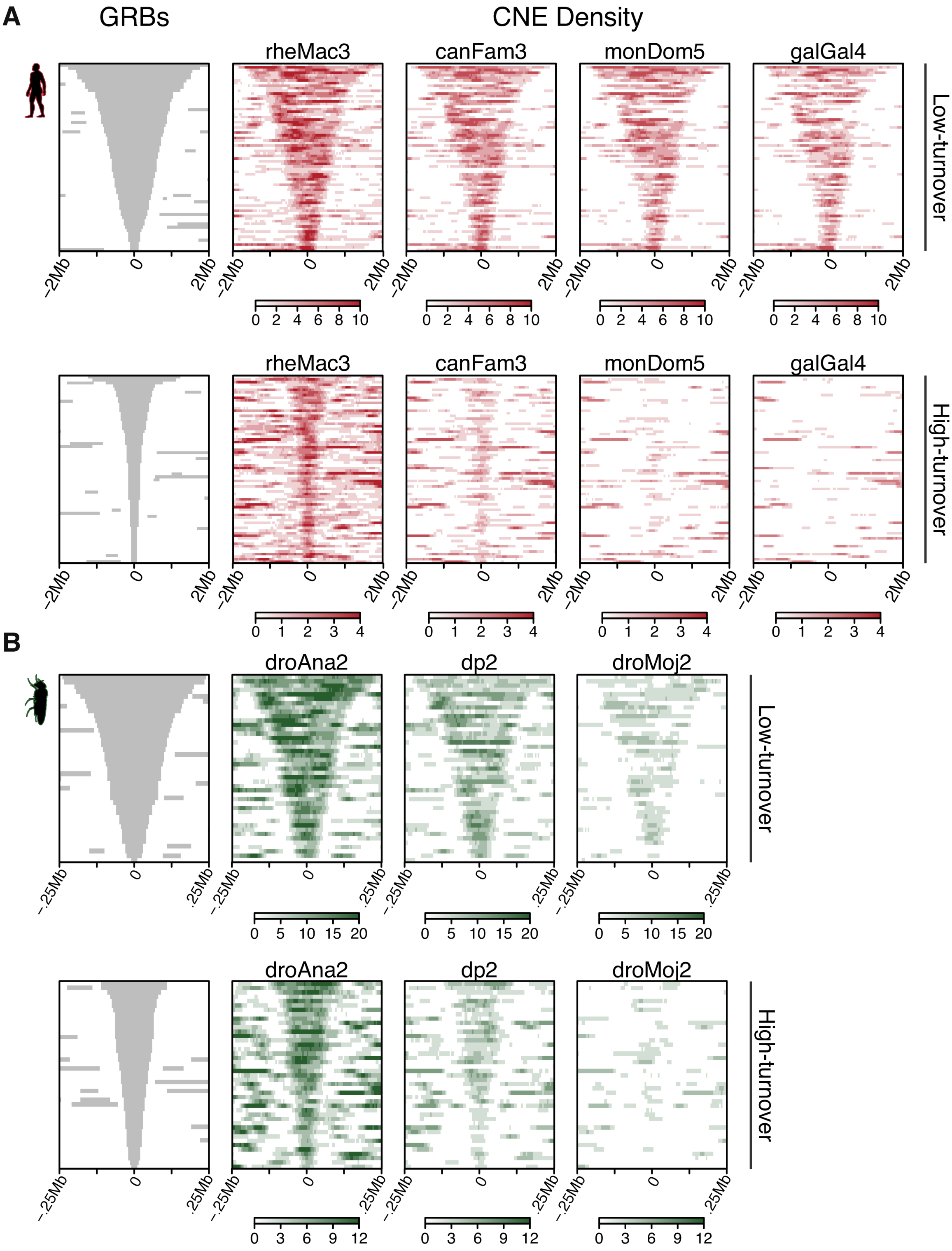

Figure 3.2: CNE Density in high- and low-turnover GRBs. CNE density for each of the species comparisons from hg19 (**A**) and dm6 (**B**) was plotted in high- and low-turnover GRBs. The grey funnels represent the extent of the GRBs, while the coloured plots show the CNE density in the same genomic windows. In the ltGRBs, there is visible enrichment of CNEs within the GRBs for all species comparisons, while in the htGRBs, this enrichment is only visible in the most closely related species comparisons and appears to decrease with increasing distance of the comparison (from left to right).

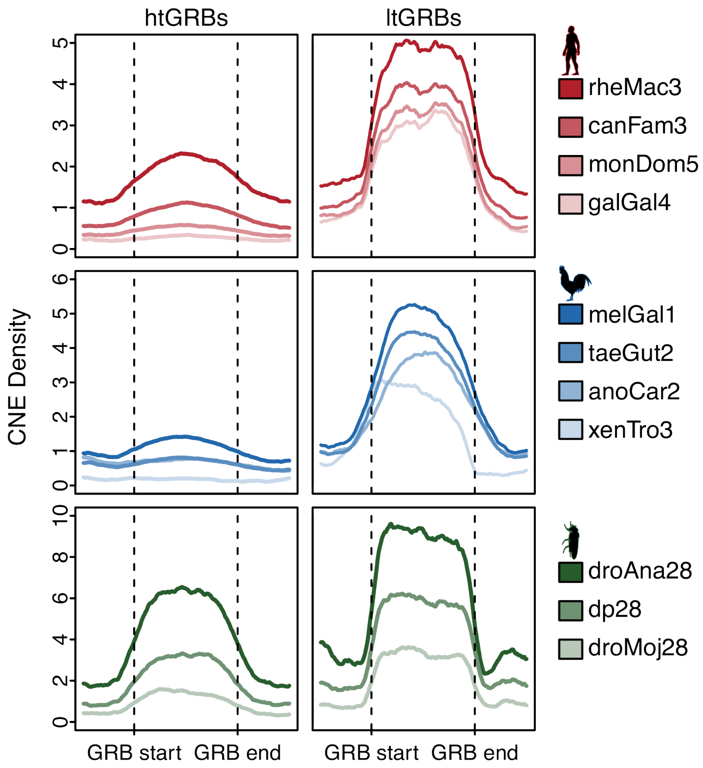

Figure 3.3: Average CNE Density in high- and low-turnover GRBs. The average CNE density in high- and low-turnover GRBs in each reference species (hg19, galGal4 and dm6). Each line represents the density of CNEs identified between the reference species and each of the species in its phylogeny. Low-turnover GRBs exhibit high CNE density for all species comparisons, while in high-turnover GRBs there is a loss of CNE density in the more distant comparisons.

that region, then GRB detection will truncate the GRB, resulting in shorter GRBs on average. Further, if their is a loss of sequence conservation in the centre of the GRB, it may be identified as two separate shorter GRBs.

If the effect is primarily due to the degradation of non-coding sequence conservation at the edges of htGRBs, then the boundaries should be less consistently predicted over multiple species comparisons. To assess this assumption I calculated the distance from GRB boundaries predicted in the closest species, for each phylogeny, to the nearest boundary predicted in each of the other species comparisons. Figure 3.4B shows these results as cumulative distributions of distances for ht- and ltGRBs. In this figure, we can see that in htGRBs, there is a trend towards increasing difference between boundary predictions as the distance of the evolutionary comparison increases, while the boundaries of ltGRBs are consistently predicted across all evolutionary timescales. These results indicate that GRB boundary prediction over multiple evolutionary timescales is less consistent in htGRBs than ltGRBs, supporting the hypothesis there is increased degradation of sequence conservation at htGRB boundaries, or that htGRB boundaries are more able to shift through evolution than ltGRB boundaries.

3.3.4 High- and low-turnover GRB boundaries are equally predictive of TAD boundaries

While degradation of non-coding conservation in distant genomes would strongly influence GRB boundary prediction in more distant comparisons, this effect should be reduced in the most closely related species comparisons. To assess the effects of sequence degradation on htGRB identification in closely related species, I compared how well the boundaries of ht- and ltGRBs coincide with TAD boundaries (Figure 3.4C). For this analysis publicly available hg19 and dm6 Hi-C data were used to identify a set of TADs for each species. I then compared the concordance of GRB and TAD boundaries by visualising the distance from each GRB boundary to its closest TAD boundary. From these figures it appears that ht- and ltGRBs coincide equally well with TADs, suggesting that htGRBs defined in close species comparisons are as accurately predicted as the ltGRBs.

Given the dramatic difference in the widths of ht- and ltGRBs, it would be expected that the TADs overlapping these two groups would also differ in width. However, when comparing TADs overlapping htGRBs (htTADs) and TADs overlapping ltGRBs (ltTADs), I find that the difference in widths is minimal. In hg19, htTADs are significantly narrower than ltTADs (Figure 3.4D; hg19 $p = 0.021$), however the difference in width is small. In dm6 htTADs tend to be narrower than ltTADs, but the difference is not statistically significant. Together, these results suggest that over short evolutionary distances htGRBs correspond as well with TADs

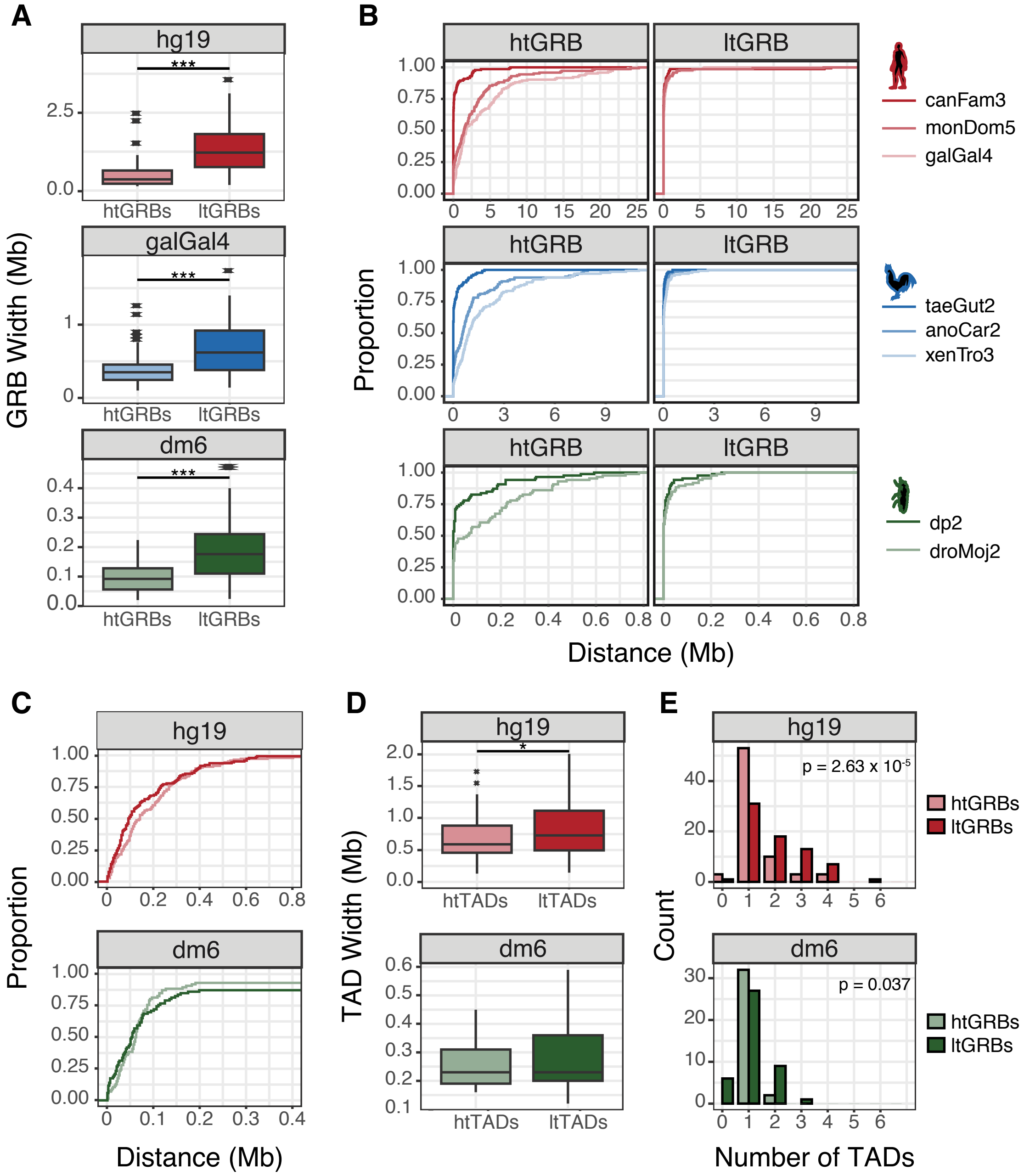

Figure 3.4: High- and low-turnover GRB width and boundary prediction. (**A**) High- and low-turnover GRB widths in each of the reference species used. (**B**) Cumulative distributions of the distance from the boundaries of GRBs identified in the closest species to the reference, in each phylogeny, to the nearest boundary identified in each of the other species comparisons. (**C**) Cumulative distributions of the distance from hg19 and dm6 GRB boundaries to the nearest TAD boundary. (**D**) The widths of TADs overlapping ht- and ltGRBs in hg19 and dm6. (**E**) Distributions of the number of TADs overlapped by ht- and ltGRBs.

as ltGRBs, and these TADs tend to be narrower than TADs overlapping ltGRBs, however differences in ht- and ltTAD widths are not sufficient to explain the striking difference in ht- and ltGRB widths.

Previously it has been noted that some of the largest and most strongly conserved GRBs are actually two or more GRBs in such close physical proximity that it is impossible to separate them using sequence conservation alone (Harmston et al. 2017). This is another potential explanation for the differences in ht- and ltGRB widths, as ltGRBs span multiple TADs significantly more often than htGRBs (Figure 3.4E; hg19 p = 2.63x10^{-5}; dm6 p = 0.037).

Taken together these results suggest that while htGRB boundary are less conserved over multiple species comparisons, in close species comparisons, htGRBs still correspond as well with TADs as ltGRBs. The TADs which coincide with htGRBs appear to be narrower in general than those which coincide with ltGRBs, supporting the theory that ltGRB target genes have a larger array of enhancers to ensure their robust spatiotemporal expression. However, the large difference in the average widths of htGRBs and ltGRBs is likely due to merging of highly conserved neighbouring GRBs in the ltGRB set.

3.3.5 High- and low-turnover GRBs target distinct subsets of genes

The next step in characterising high- and low-turnover GRBs was to test whether they regulate specific classes of genes, and whether this is consistent across phylogenies. For each phylogeny, I retrieved the co-ordinates of all protein coding, microRNA and long non-coding RNA genes in each reference genome. Next, I subset these genes into those that overlap htGRBs and ltGRBs. I performed GO enrichment analysis on the high- low-turnover sets separately in each reference species. When examining the ten most enriched GO terms from each reference species, it is clear that htGRBs and ltGRBs regulate functionally distinct subsets of genes

(Figure 3.5). Genes in ltGRBs are strongly enriched for BP terms related to the regulation of transcription, RNA biosynthetic processes and pattern specification in embryogenesis, while genes in htGRBs are enriched for cell adhesion, nervous system development and general multicellular organism development terms. These differences are also reflected in the MF terms, with ltGRB genes enriched for DNA binding and transcription factor activity terms, and htGRBs enriched for cell-surface receptor related terms such as calcium ion binding, transmembrane receptor protein tyrosine kinase activity and neurotrophin receptor binding (Figure S2).

Most importantly, the differences in enrichment seen in the high- and low-turnover sets are largely consistent across all three reference species. This indicates that similar classes of genes are undergoing similar rates of regulatory turnover in three distantly related phylogenies, and that the GRBs we identify as either deeply or shallowly conserved in each of the three phylogenies are functionally equivalent. It is likely that the rate of regulatory turnover within a GRB may be directly influenced by the function of the gene it regulates. This phenomenon is illustrated in Figure 3.6. Shown are the orthologous regulatory landscapes surrounding the hg19, galGal4 and dm6 equivalents of the *MEIS2* and *CDH6* genes. *MEIS2*, a transcription factor crucial to the regulation of early development, is located within a ltGRB in all three phylogenies, and the GRB boundaries predicted from each species comparison are highly similar regardless of the distance of the comparison. In contrast, *CDH6*, a cell membrane glycoprotein that mediates homophilic cell-adhesion during kidney development, occurs within a htGRB in all three phylogenies. It is clear that in all three phylogenies the non-coding conservation surrounding *CDH6* is only maintained between closely related species, and robust GRBs are only predicted from each reference species to its most closely related species.

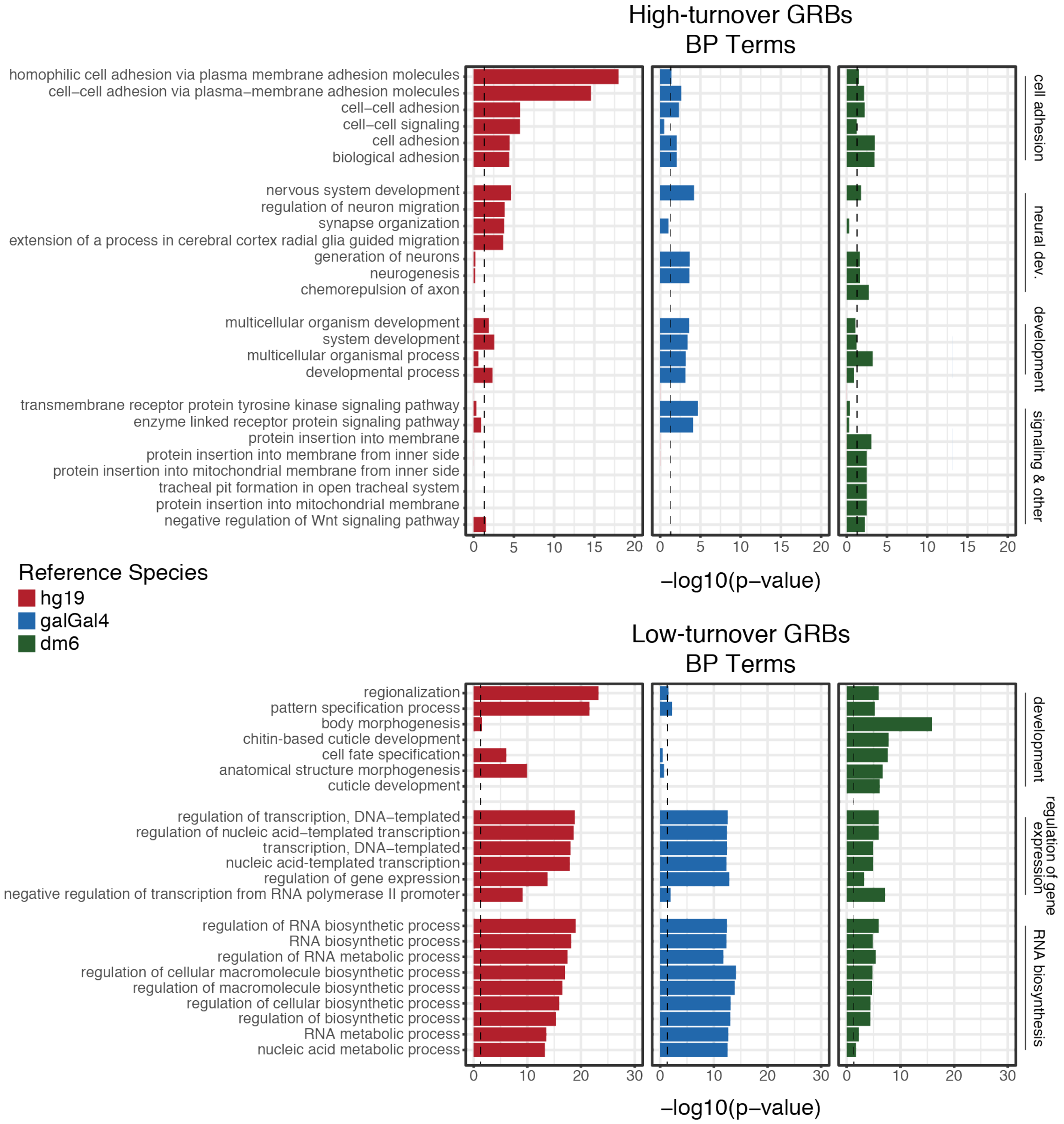

Figure 3.5: High- and low-turnover GRB target gene GO enrichment. Gene ontology (GO) enrichment analysis was performed for the target genes of ht- and ltGRBs. The -log10 p-values for the top 10 enriched BP terms from each reference species are shown here. Biological process GO terms indicate that ht- and ltGRBs are enriched for distinct subsets of developmental genes, and that the enrichment is consistent across the three reference species used.

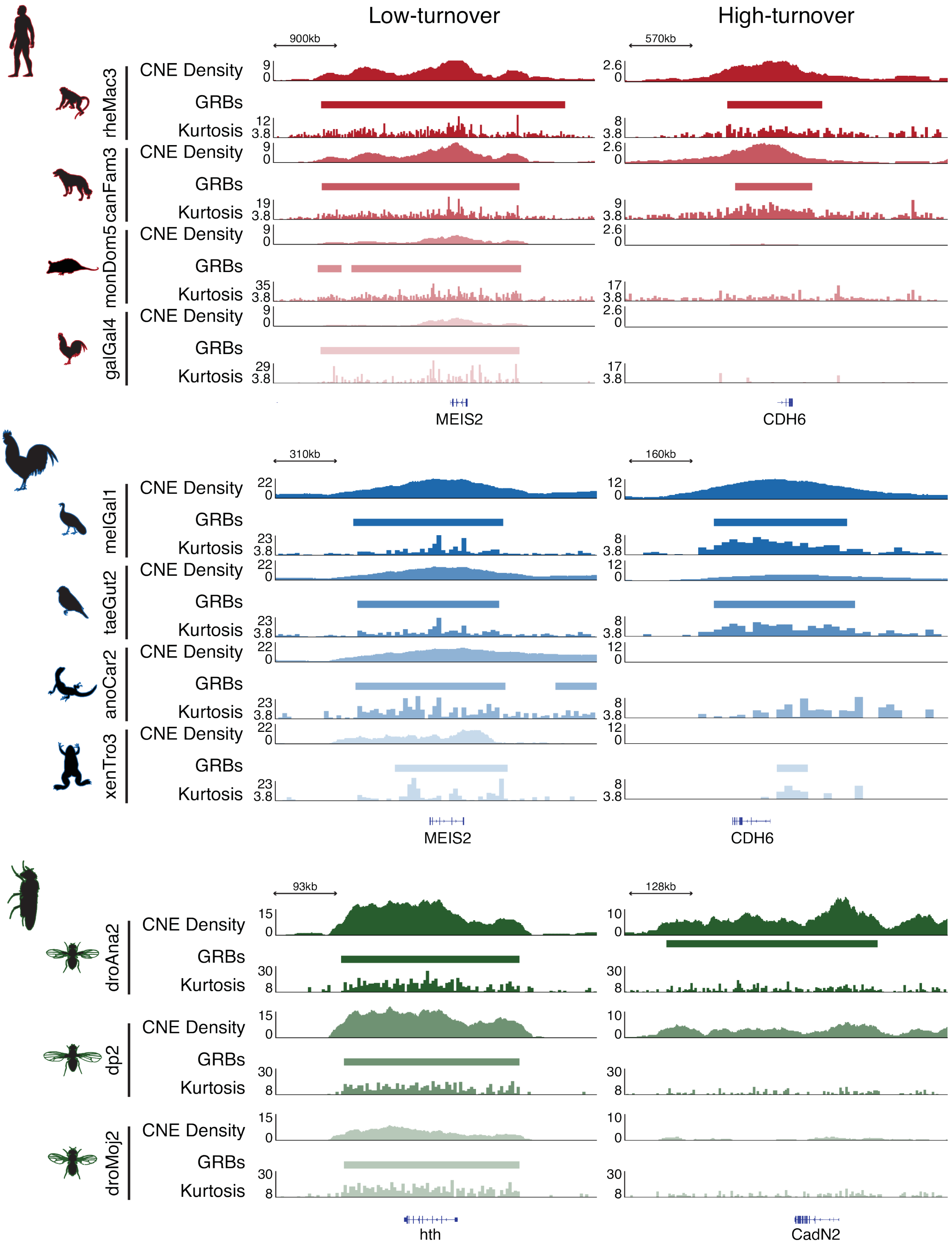

Figure 3.6: Patterns of regulatory turnover are conserved across distinct phylogenies. CNE density, kurtosis-based GRB predictions and kurtosis-based conservation are shown around each reference species ortholog of a hg19 ltGRB target gene (*MEIS2*) and htGRB target gene (*CDH6*). Conservation patterns around each class of target gene are very similar in all three phylogenies. Note that for clarity, only the predicted GRB target genes are shown in this figure.

3.3.6 High- and low-turnover GRB target gene expression during development

Since the genes regulated by ht- and ltGRBs appear to tolerate turnover within their regulatory landscapes differently, I sought to further characterise them and identify shared properties that may explain the differing GRB conservation patterns. I used ENSEMBL's ortholog predictions to map hg19 GRB target genes to their mouse (mm9) orthologs and then assessed their expression dynamics during mouse embryonic development using CAGE data from the FANTOM5 consortium (Forrest et al. 2014). The FANTOM5 mouse developmental time course begins at embryonic day 11, spanning 9 time-points until embryonic day 18. Using a SOM, I clustered all mouse genes based on their dynamics through development (Figure 3.7A). SOMs tend to place the largest and most distinct clusters at the corners of the grid, and in this case the three most distinct clusters represent early expressed genes with decreasing expression as development progresses, late expressed genes which start low and increase their expression as development progresses, and genes expressed stably throughout development. The fourth corner of the SOM contains genes which don't show any consistent pattern across development, and I therefore excluded this cluster from further analysis.

To assess whether ht- and ltGRB target genes display different expression dynamics during development, I subset all genes overlapping ht- and ltGRBs based on whether they were predicted to be targets of GRB regulation (see Section 3.2.9). Next I performed a Fisher's exact test for enrichment or depletion of high- and low-turnover target genes in the early, late and stable expression clusters derived from the SOM. The expected number of high- and low-turnover genes in each cluster was based on the distribution of all GRB target genes between clusters (Figure 3.7B). High- and low-turnover GRB target genes are significantly differently distributed between all clusters ($p < 0.0005$). Low-turnover target genes are significantly enriched in the early expression cluster ($p < 6.38e-08$) and significantly depleted in the

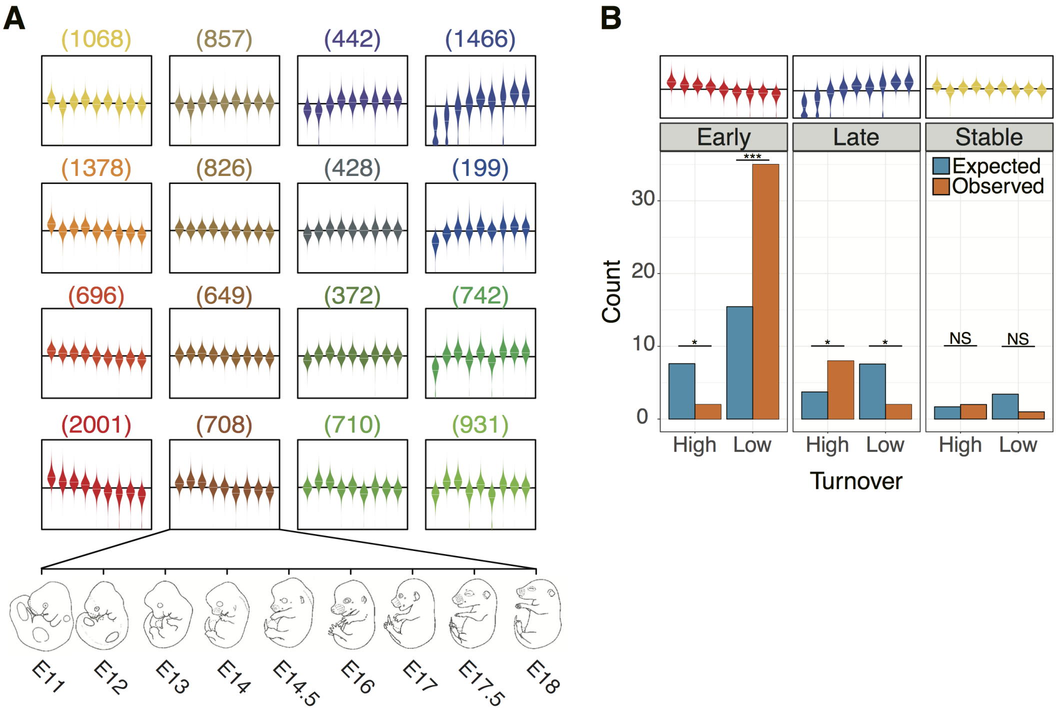

Figure 3.7: High- and low-turnover GRB target gene expression during development. (**A**) Gene expression profiles obtained by self-organising map clustering of CAGE signal at mouse promoters during development. Each box represents a cluster and contains a beanplot representing the relative expression of the genes, in each cluster, at each developmental time point. The developmental time points used, which serve as the x-axis for all boxes, are show on the bottom. (**B**) The observed and expected number of ht- and ltGRB target genes in the early, late and stably expressed gene expression clusters.

late expression cluster ($p < 0.024$), while high-turnover target genes are depleted in the early cluster ($p < 0.0195$) and significantly enriched in the late expression cluster ($p < 0.026$). Neither high- nor low-turnover target genes have any significant enrichment or depletion in the stably expressed cluster.

These results are in agreement with the functional classes of high- and low-turnover GRB genes identified in Section 3.3.5. Low-turnover genes are enriched for transcription factors involved in pattern specification in the developing embryo, which occurs early in development. High-turnover genes are enriched for genes involved cell-cell adhesion in neural development. These are processes which are specific to distinct tissues in the developing embryo, and therefore occur later in development than the initial establishment of the embryonic body plan.

Altering the regulation of a gene expressed early in development is likely to have greater pleiotropic effects than genes expressed later in development. Any detrimental change in the regulation of an early expressed gene will have an impact on the cells in which the gene is expressed, as well as all of the tissues which are derived from those cells. In this scenario it is likely that selection against regulatory change in ltGRBs is greater than in htGRBs due to the difference in function and timing of expression of the genes that they regulate.

3.3.7 High- and low-turnover GRBs occur in distinct chromatin states

It has been shown that GRBs tend to be maintained in a silent or polycomb repressed chromatin state in adult tissues (Harmston et al. 2017). This is due to the vast majority of GRBs being active during development and subsequently silenced in mature tissues. Given that there were differences in the developmental expression and function of genes targeted by ht- and ltGRBs, I next sought to identify concomitant differences in chromatin state at ht- and ltGRBs.

To this end, I plotted the average coverage of H3K27me3, H3K4me1 and H3K4me3 across ht- and ltGRBs using hg19 fetal brain chromatin modification data from the NIH Roadmap Epigenomics Consortium (Kundaje et al. 2015). Since htGRB target genes are expressed late in development and enriched in neural developmental GO terms, fetal brain samples provide an excellent tissue in which to observe differences in ht- and ltGRB chromatin state. ltGRBs exhibit a clear and strong enrichment of H3K27me3 and H3K4me1, but no enrichment of H4K4me3 (Figure 3.8A). H3K27me3 enrichment indicates that as expected, ltGRBs are polycomb repressed. Further, the combination of H3K27me3 and H3K4me1 is usually associated with poised enhancers, but can also mark enhancers that have been inactivated following a period of activity during development (Bonn et al. 2012). The human chromatin modification data used in this chapter was generated from 17 week old embryos and, given that the majority of ltGRB target genes are expressed early in development, we would expect that at this stage of development most ltGRBs would be inactive rather than poised. Unlike ltGRBs, htGRBs are not polycomb repressed in the fetal brain. There is no enrichment of H3K27me3 inside htGRBs, and in fact there appears to be a slight dip in coverage within their boundaries. Similar to the ltGRBs, htGRBs are enriched for H3K4me1, but to a greater extent. There is also a sharp increase in H3K4me3 at htGRB boundaries and at their centre, suggesting that they contain both active enhancers and promoters. Taken together these results suggest that htGRBs are more active in the human fetal brain, while ltGRBs are polycomb repressed and inactive. This result is in agreement with the GO results, which identified an enrichment in neural development genes within htGRBs.

To test whether these results generalise to other species, I retrieved dm6 chromatin modification data from the modENCODE consortium (Roy et al. 2010). Unfortunately the developmental samples available from modENCODE are not divided by tissue, so I could not directly compare the developing brains of the two organisms. Instead I retrieved chromatin modification data derived from embryos 20-24hr post fertilisation. This is the final stage prior to hatching and the subsequent larval stages of Drosophila development. Drosophila larvae hatch with a complete central nervous system, and therefore at least some of the signal observed in the whole organism will be contributed by the developing nervous system. Further, this is a late stage of the first *Drosophila* developmental phase, thereby maximising the chances that we can observe active htGRBs. As with the hg19 GRBs, I plotted the H3K27me3, H3K27ac, H3K4me1 and H3K4me3 signal across dm6 ht- and ltGRBs (Figure 3.8B). In this case, the coverage is presented as the log fold enrichment over input.

Similar to hg19 ltGRBs, dm6 ltGRBs show the characteristic enrichment of H3K27me3 and concomitant depletion of H3K27ac. htGRBs, however, are not en-

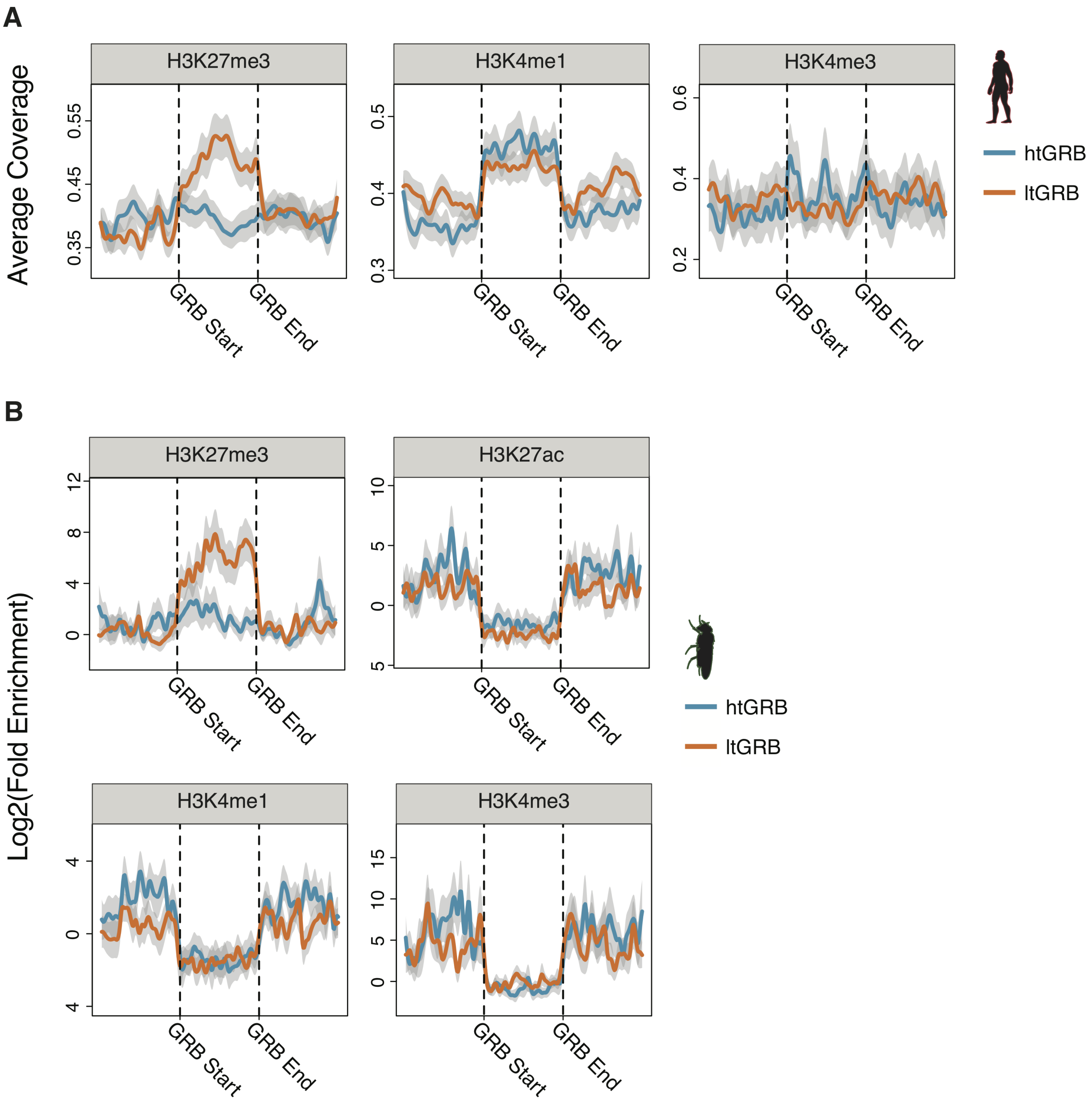

Figure 3.8: Chromatin modifications at high- and low-turnover GRBs. (**A**) Average coverage of human fetal brain H3K27me3, H3K4me1 and H3K4me3 ChIP-seq reads across ht- and lt-GRBs. (**B**) Log2 fold enrichment of *D. melanogaster* H3K27me3, H3K27ac, H3K4me1 and H3K4me3 ChIP-seq signal over input across ht- and ltGRBs.

riched for H3K27me3, mirroring what we see in the hg19 htGRBs. There is very little difference between dm6 ht- and ltGRBs with respect to H3K4me1 and H3K4me3 enrichment. Both are relatively depleted for H3K4me1, and have no specific enrichment of H3K4me3. Interestingly, there is a strong enrichment of H3K4me3 just outside the boundaries of dm6 GRBs. It has been previously reported that *Drosophila* TADs are flanked by clusters of housekeeping genes (Sexton et al. 2012). Given the strong concordance between GRBs and TADs, H3K4me3 enrichment at GRB boundaries is likely due to clusters of active housekeeping genes. The depletion of H3K4me1 in both ht- and ltGRBs in dm6 is unexpected, but could be explained by both sets of GRBs being inactive in the majority of tissues, thereby diluting the signal from active tissues. These results suggest that in 20-24h post fertilisation *Drosophila* embryos, target genes in ltGRBs are polycomb repressed, while target genes in htGRBs may occur within open chromatin. Inclusion of more chromatin modification data may help to clarify the chromatin state at htGRBs.

Taken together, these results show that ht- and ltGRBs are active in different tissues and time points during development. From the time points sampled, it appears that htGRBs are more likely to be active in late neural development than ltGRBs in humans. In *Drosophila* htGRBs are potentially in open chromatin regions in the final stage before hatching while ltGRBs are polycomb repressed. These results are consistent with the results of the GO enrichment and the developmental expression analysis, identifying a tendency towards later expression in htGRBs and a potential role in neural development.

3.3.8 High-turnover GRBs are more likely to contain active repeat elements that low-turnover GRBs

GRBs are known to be depleted of repeat elements, including SINEs, LINEs, and LTRs, presumably because repeat element insertion may disrupt the regulation of the GRB target gene, and is therefore not tolerated (Harmston et al.

2017). However, repeat elements have been shown to play an instrumental role in the evolution of gene regulation in general (Thornburg, Gotea, and Makałowski 2006; Feschotte and Pritham 2007; Feschotte 2008; Chuong, Elde, and Feschotte 2017), and LINEs and SINEs in particular have contributed to the evolution of mammalian brain development (Sasaki et al. 2008; Singer et al. 2010). Given the role of htGRBs in neural development, and their evidence of activity in the human fetal brain, I compared the repeat content of ht- and ltGRBs, hypothesising that repeats may be better tolerated in htGRBs and may contribute to their regulatory turnover.

To this end, I downloaded all RepeatMasker annotated hg19 repeats and visualised the density of SINEs, LINEs, and LTRs across ht- and ltGRBs (Figure 3.9A). ltGRBs exhibit the expected patterns, with visible depletion of all three repeat families within the boundaries of GRBs. htGRBs, however, while depleted of LINEs and LTRs, are not visibly depleted of SINEs. Quantified as SINEs per 10kb, ltGRBs are significantly less SINE dense than htGRBs (Figure 3.9B, p=0.0492).

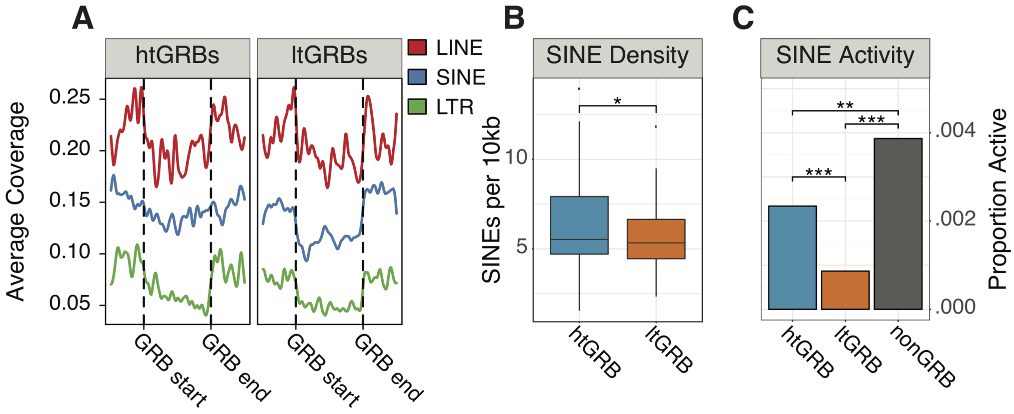

Figure 3.9: Repeats in high- and low-turnover GRBs. (**A**) Average repeat coverage across ht- and ltGRBs. (**B**) SINEs per 10kb in ht- and ltGRBs. (**C**) The proportion of SINEs which are active (as determined by H3K27ac) in htGRB, ltGRB and nonGRB regions

While htGRBs are more SINE dense than ltGRBs, SINE elements are generally silenced in the human genome, and thus their presence alone is not sufficient evidence that they have an effect on target gene regulation in these GRBs (Rhee et al. 2002; Carnell and Goodman 2003). To assess whether SINEs in htGRBs are more likely to be functional in development than SINEs in ltGRBs, I determined the proportion of SINEs which are marked by H3K27ac in the fetal human brain for SINEs within htGRB, ltGRB and nonGRB regions of the genome, using this a proxy for activity (Figure 3.9C). In general only a tiny proportion of SINEs in the human genome are active in the fetal human brain, however SINEs in both ht- and ltGRBs are significantly less likely to be active than SINEs in the rest of the genome (htGRB: $p = 1.37\text{x}10^{-5}$; ltGRB: $p < 2.2\text{x}10^{-16}$). This likely reflects silencing of SINEs within GRBs, as they have the potential to act as regulatory elements and alter the expression of GRB target genes. Further, SINEs in htGRBs are significantly more likely to be active than SINEs in ltGRBs ($p = 3.73\text{x}10^{-7}$).

These results suggests that htGRBs are not only more likely to tolerate SINE insertions, but that SINEs within htGRBs are more likely to be active during neural development and may potentially be exapted into regulatory function within these GRBs.

3.4 Discussion

In this chapter I have successfully identified GRBs that exhibit deep and shallow evolutionary conservation in three distinct metazoan lineages using kurtosis-based conservation. I define these as ht- and ltGRBs and, in support of the turnover model, show that they share many characteristics between lineages. Firstly, ht- and ltGRBs target distinct subsets of genes - ltGRBs mostly regulate developmental transcription factors, while htGRBs tend to regulate developmental genes involved in cell adhesion and neural development. Further, ht- and ltGRBs consistently regulate the same classes of genes regardless of the phylogeny in which they were identified. Next I showed that ht- and ltGRB target genes are expressed at different stages of development, with ltGRB targets expressed early and htGRB targets expressed late. Additionally, in late development, ht- and ltGRBs are maintained in different chromatin states. In both hg19 and dm6, htGRBs lack polycomb repression and appear to be active while ltGRBs are polycomb repressed and silent. Finally, I show that htGRBs in hg19 are more likely to contain active SINEs, potentially contributing to the regulatory turnover in these regions.

Lowe et al. proposed that there have been three distinct periods of regulatory innovation during vertebrate evolution (Lowe et al. 2011). They posit that a first wave of innovation occurred at developmental transcription factors, a second at cell-signalling genes and a third, mammalian specific wave, occurred at post translational modification genes. Interestingly, these are the same classes of genes I identify in my turnover analysis, with the ltGRBs targeting Lowe et al.'s first wave genes, and the htGRBs targeting their second and third wave genes. Given that htGRBs identified in a mammalian, vertebrate and invertebrate lineage all target highly similar classes of genes, a more parsimonious explanation than independent, lineage-specific recruitment of enhancers can be provided by the turnover model. It is more plausible that these genes were under this form of long-range regulation in the metazoan ancestor, and that due to continual sequence turnover it is not possible to identify GRBs by sequence comparison alone for large evolutionary distances. Thus the classes of genes proposed to have been the subject of more recent regulatory innovation, as identified by Lowe et al., are those undergoing more rapid regulatory sequence turnover, thereby appearing to be recently recruited in a specific lineage. While these results support the turnover model, it is also true that lineage-specific CNE recruitment does occur (Wang et al. 2009). It is possible that recruitment is more likely to take place within htGRBs as these regions are more permissive environments for regulatory changes in general. Thus validation of the turnover model would not invalidate the conclusions made by Lowe et al., rather recent lineage-specific recruitment of CNEs in htGRBs would be continuous with the turnover model.

To understand why certain classes of genes tolerate sequence turnover in their regulatory regions better than others, I characterised ht- and ltGRB target genes with respect to timing of expression during development, chromatin state, and repeat content. The general picture that emerged is that ltGRBs are expressed early in development, are polycomb repressed in neural tissues in humans and late in development in *D. melanogaster*, and are significantly depleted of SINE elements. In contrast htGRBs appear to be expressed late in development, are not polycomb repressed in either human neural tissue or late *D. melanogaster* development, and are not significantly depleted of SINE elements. The timing of expression of ht- and ltGRB target genes could potentially explain many of the trends I observe. It is plausible that perturbation of the gene expression pattern of genes expressed in the early stages of development (via sequence changes to their regulatory elements) is more strongly selected against than those expressed later in development. Any alteration of the expression pattern of a very early gene is more likely to have pleiotropic effects on the cell expressing the gene, but also all those cell populations derived from that cell. This effect could also explain the increased tolerance of SINE insertions in htGRBs, and their increased overall activity in developing neural tissues. It has been repeatedly shown that transposable elements can be exapted into regulatory function (Girard and Freeling 1999; Deininger et al. 2003; Mikkelsen et al. 2007; Chuong, Elde, and Feschotte 2017), and in particular Alu elements (SINEs) have been shown to resemble enhancers and may be able to function as proto-enhancers (Su et al. 2014). Further, knockdown of enhancer SINEs in mouse cortical neurons induces defects of both cortical radial migration *in vivo* and activity-dependent dendritogenesis *in vitro* (Policarpi et al. 2017). Therefore, ltGRBs would be less likely to tolerate insertion of potential new regulatory sequences as the viability of the developing embryo may be more sensitive to misregulation of early expressed ltGRB target genes. In the future it would be interesting to directly investigate this hypothesis by comparing the number of distinct cell types/populations which are derived from progenitor cell that express ht- and ltGRBs respectively.

Overall, the identification of GRBs that have similar conservation patterns and target similar classes of genes, in three independent metazoan lineages, is strong support for the turnover model. The most parsimonious explanation for these similarities is that GRBs are an ancient feature of metazoan gene regulation, and that since the initial evolution of distantly acting *cis*-regulatory elements, CNEs and GRBs have been under strong negative selective pressure due to their role in multicellular development. While selection against sequence changes in these regions is incredibly strong, it is likely that no sequence is totally indispensable, and thus given enough time these regions will turnover to the point that they are no longer identifiable by sequence conservation alone.

Chapter 4

GRBs in compact genomes

4.1 Introduction

The relationship between GRB and genome size was first investigated by comparing the size of opossum, human, mouse, chicken, spotted gar and *D. melanogaster* GRBs to their respective genome sizes (Harmston et al. 2017). The authors observed that GRB size scaled with genome size, and that the size of GRBs identified in one species was highly predictive of the size of their orthologous GRBs in another species. These results suggest that GRBs expand and contract at a comparable rate to the genome, although none of the species studied had undergone a high degree of genome contraction. Under the assumption that genome expansion or contraction does not alter the selective pressure on the regulation of a gene under GRB regulation, GRB expansion and contraction should occur by changes in the spacing between CNEs rather than sweeping gain or loss events.

While this is an intuitive assumption, it is at odds with a 2006 study that found that the distance between pairs of adjacent CNEs is highly conserved between species (Sun, Skogerbø, and Chen 2006). The authors compared the distance between pairs of CNEs in the human genome to the distance between their orthologs in the mouse, rat and dog genome and found that the distance between CNEs was significantly better conserved than the distance between pairs of exons and genes. However, when the analysis was extended to include non-mammalian vertebrates, a significant proportion of the CNE pairs were more closely spaced in the non-mammalian vertebrates than the human genome, with the remainder maintaining a conserved distance. Further, the smaller the genome, the more pronounced the effect, with human to fugu and tetraodon comparisons showing that approximately half of all CNE pairs were much more closely spaced in the fugu and tetraodon genomes than in the human genome. Interestingly, the authors note that the subset of CNE pairs that maintain a conserved spacing in the more distant comparisons are those that were already relatively close together (< 40kb). The bimodality of these results may imply that there is a minimum spacing required between CNEs for them to function correctly, as was shown for the stripe enhancers in *D. melanogaster* (Small, Arnosti, and Levine 1993). This could explain why, in the context of genome compaction in fugu and tetraodon, the distance between closely spaced CNEs was relatively well conserved, while distantly spaced CNE pairs are found much closer together post genome compaction.

If there is indeed a minimum required distance between CNEs, GRBs should also be limited in their minimum size. Thus far GRBs have only been identified in large vertebrate genomes and the *Drosophila melanogaster* genome. While it appears that GRBs in the *Drosophila* genome have contracted proportionally with genome size, GRBs need to be identified in more genomes that have undergone rapid compaction to discover whether this is *Drosophila*-specific or a general trend of genome compaction.

Two species that satisfy this criterion are *Caenorhabditis elegans* and *Oikopleura dioica*. *C. elegans* is a hermaphroditic nematode with a 97Mb genome (The C. elegans Sequencing Consortium 1998). Genome compaction in *C. elegans* is thought to have occurred due to the transition from outcrossing to self-fertilisation, as all self-fertilising *Caenorhabditis* species have significantly smaller genomes than their outcrossing relatives (Fierst et al. 2015). *O. dioica* is a pelagic tunicate with a 70Mb genome, the smallest metazoan genome sequenced to date (Denoeud et al. 2010). *O. dioica* are extremely fast evolving, perhaps contributing to the genome plasticity required for such extreme compaction (Berna and Alvarez-Valin 2014), however the reason for this compaction is unknown.

In this chapter I identify GRBs in the highly compacted *Caenorhabditis elegans* and *Oikopleura dioica* genomes for the first time, and analyse the effects of genome compaction on GRBs. Initially I show that these GRBs are functionally equivalent to GRBs identified in larger genomes with respect to chromatin modifications and target gene enrichment. Next, I analyse the relationship between genome and GRB size and find that GRB size scales proportionally with genome size, even in highly compacted genomes. Further, I show that while GRB and genome composition vary widely between the species analysed, GRBs appear to be under similar constraints with respect to sequence feature composition. This is the first time GRBs have been identified in such highly compact genomes, and their functional equivalence with previously identified GRBs strengthens the argument that GRBs are an ancient feature of metazoan gene regulation.

4.2 Methods

4.2.1 Species used in Chapter 4

To facilitate a systematic and unbiased comparison of the properties of GRBs in multiple genomes of varying sizes and compositions, GRBs were identified in each of the species listed in Table 4.1. GRBs were identified using the kurtosis-based GRB identification pipeline first outlined in Chapter 2. Throughout this chapter species will be referred to by their genome assembly identifiers.

Table 4.1: Species used in Chapter 4

Species	Common name	Genome Assembly
Homo sapiens	Human	hg19
Mus musculus	Mouse	mm10
Danio rerio	Zebrafish	danRer10
Gallus gallus	Chicken	galGal4
Tetraodon nigroviridis	Green spotted puffer fish	tetNig2
Drosophila melanogaster	Fruit fly	dm6
Caenorhabditis elegans	Roundworm	ce10
Oikopleura dioica	Sea squirt	OikDioicaNorway

4.2.2 Pairwise genome alignment

All pairwise genome alignments used in this analysis were either retrieved from the UCSC genome browser or generated using LASTZ (Harris 2007), as described in section 2.2.1. The source of all alignments used in this chapter is listed in Table 4.2.

Table 4.2: Pairwise alignments used in Chapter 4

Reference Species	Query Species	Source	Generated By
Human (hg19)	Mouse (mm10)	UCSC	UCSC
Mouse (mm10)	Human (hg19)	UCSC	UCSC
Zebrafish (danRer10)	Blind Cave Fish (AstMex102)	Lenhard group	Ge Tan
Chicken (galGal4)	Zebra finch (taeGut2)	UCSC	UCSC
Green spotted puffer fish (tetNig2)	Japanese puffer fish (fr3)	UCSC	UCSC
Fruit fly (dm6)	Fruit fly (droMoj2)	Lenhard group	Ge Tan
Roundworm (ce10)	Roundworm (cb3)	Lenhard group	Ge Tan
Sea squirt (OikDioicaNorway)	Sea squirt (OikDioicaJapan)	Lenhard group	Ge Tan

4.2.3 CNE identification

CNEs were identified in ce10 and OikDioicaNorway to facilitate a comparison between CNE-based and kurtosis-based GRBs. CNEs were identified as described in section 2.2.2. Briefly, pairwise net whole-genome alignments were scanned for regions of high sequence identity over a predefined length. Regions that passed the minimum sequence identity and length criteria were filtered such that those overlapping repeat regions or exons were excluded. To prevent the inclusion of unannotated repeats, the remaining regions were aligned to the reference genome, using BLAT, and those which mapped to more than four loci were removed (Kent 2002). The remaining regions constitute the final set of CNEs. CNE identification was performed using the `CNEr` package in `R`. The parameters used for both species are listed in Table 4.3.

CNE density across the genome was calculated by sliding a window across the genome in 1kb increments and counting the number of CNEs in each window. A 20kb and 10kb window was used for ce10 and OikDioicaNorway respectively.

Table 4.3: Parameters used for CNE identification in ce10 and OikDioicaNorway

Parameter	ce10	OikDioicaNorway
Minimum Identity (%)	96.6	98
Minimum Length (bp)	30	50
Smoothing Window Size (kb)	20	10
Smoothing Step Size (kb)	1	1

4.2.4 Kurtosis-based conservation calculation

Kurtosis-based conservation was calculated as described in section 2.2.4 for all species comparisons listed in Table 4.2. In summary, the reference genome was divided into bins and for each bin all sequences perfectly conserved between the reference and the query genome were extracted from the pairwise alignment. The kurtosis of the distribution of the lengths of these sequences was then calculated. The bin sizes were selected such that the range of sizes used in each species was the same relative to the size of that species' genome. The bins used for each species were approximately 1/150000, 1/100000, 1/75000 of the genome size for that species. The bin sizes used for each reference species are listed in Table 4.4.

Table 4.4: Bin sizes used for kurtosis-based conservation calculation

Reference Species	Query Species	Bin Size (kb)
Human (hg19)	Mouse (mm10)	20; 30; 40
Mouse (mm10)	Human (hg19)	18; 27; 36
Zebrafish (danRer10)	Blind Cave Fish (AstMex102)	9; 14; 18
Chicken (galGal4)	Zebra finch (taeGut2)	7; 10; 14
Green spotted puffer fish (tetNig2)	Japanese puffer fish (fr3)	2.4; 3.6; 4.8
Fruit fly (dm6)	Fruit fly (droMoj2)	1; 1.4, 1.9
Roundworm (ce10)	Roundworm (cb3)	0.7; 1; 1.3
Sea squirt (OikDioicaNorway)	Sea squirt (OikDioicaJapan)	0.5; 0.7; 0.9

4.2.5 GRB identification

For ce10 and OikDioicaNorway, the CPM framework was used to identify GRBs from both CNE density and kurtosis-based conservation. For the remainder of the species comparisons listed in Table 4.4, only kurtosis-based conservation was used. The GRB prediction method is described in detail in section 2.2.6. Briefly, change points in the mean and variance of the input signal (either CNE density or kurtosis-based conservation) were identified across the genome. These change points were then treated as potential GRB boundaries. Adjacent windows, on either side of a potential GRB boundary, were then merged if both windows had a mean signal greater than a predefined quantile of the distribution of all signal values across the genome. The merged windows were then used as the final set of GRBs.

Multiple quantiles were used for GRB calling to ensure the identified trends relate to GRB properties and genome size and are not caused by the thresholding applied. GRBs were called with merging occurring when the mean signal in both adjacent windows was above 70%, 80% and 90% of all signal values across the genome.

4.2.6 Chromatin modifications in GRBs

Publicly available chromatin modification data was used to assess the quality of the predicted GRBs in ce10 and OikDioicaNorway. The datasets used are listed in Table 4.5. Processed signal tracks were retrieved for all datasets.

Table 4.5: Publicly available chromatin modification data used in Chapter 2

Data type	Sample	Genome assembly	Citation	GEO accession
H3K27me3 ChIP-chip	Ovary	OikDioicaNorway	Navratilova et al. 2017	GSE78915
H3K27ac ChIP-chip	Ovary	OikDioicaNorway	Navratilova et al. 2017	GSE78915
H3K4me1 ChIP-chip	Ovary	OikDioicaNorway	Navratilova et al. 2017	GSE78915
H3K4me3 ChIP-chip	Ovary	OikDioicaNorway	Navratilova et al. 2017	GSE78915
H3K27me3 ChIP-seq	Early Embryo	ce10	No citation provided	GSE49738
H3K27ac ChIP-seq	Early Embryo	ce10	No citation provided	GSE49734
H3K4me1 ChIP-seq	Early Embryo	ce10	No citation provided	GSE50262
H3K4me3 ChIP-seq	Early Embryo	ce10	No citation provided	GSE49739
H3K27me3 ChIP-seq	Larvae L3	ce10	No citation provided	GSE49724
H3K27ac ChIP-seq	Larvae L3	ce10	No citation provided	modENCODE
H3K4me1 ChIP-seq	Larvae L3	ce10	No citation provided	GSE49206
H3K4me3 ChIP-seq	Larvae L3	ce10	No citation provided	GSE28770
H3K27me3 ChIP-seq	Young Adult	ce10	No citation provided	GSE50314
H3K27ac ChIP-seq	Young Adult	ce10	No citation provided	modENCODE
H3K4me1 ChIP-seq	Young Adult	ce10	No citation provided	GSE50312/GSE50287

To assess the chromatin state of predicted GRBs, heatmaps were produced in which each row is a genomic window centred on a predicted GRB, and the intensity of the colour represents the enrichment of a particular chromatin modification in that genomic location. The rows are ordered by the width of the predicted GRBs, and thus, in these figures, any chromatin modification enriched inside the GRBs forms a funnel shape. This was performed for all chromatin modifications available for each species.

For all metaplots, the predicted GRBs were extended to 3 times their width and the resulting windows were divided into 300 bins. For each bin, the average ChIP-chip or ChIP-seq signal was then calculated. For OikDioicaNorway the ChIP-chip enrichment is presented as log2 fold enrichment over input DNA, while for ce10 the ChIP-seq data is presented as raw read coverage. K-means clustering of these windows was used to separate ce10 GRBs into those which displayed active chromatin modifications and those which did not.

4.2.7 Gene ontology enrichment analysis

GO enrichment analysis was performed on all genes within predicted GRBs, compared to a universe of all annotated genes. The resulting p-values were false discovery rate (FDR) corrected and the top 15 most enriched BP and MF terms in each species were visualised. For OikDioicaNorway, gene to GO term mappings were retrieved from OikoBase, a curated Oikopleura database (Danks et al. 2013). GO enrichment analysis was performed for ce10 and OikDioicaNorway separately using the `GOstats` package in `R`.

4.2.8 GRB size and composition analysis

GRBs were predicted from kurtosis-based conservation, calculated in bins across the genome, for all species listed in Table 4.1. For each species, the bin sizes used correspond to approximately 1/150000, 1/100000 and 1/75000 of the genome size. Further, in the GRB identification step, three quantiles were used for merging of neighbouring windows. Thus, for each species 9 sets of GRBs were generated. To examine the relationship between GRB size and genome size, the mean GRB width was determined for each set and plotted against genome size. GRB sets identified using the same merging quantile were analysed together, and a linear model was fitted based on the mean widths at each bin size for each species. Similarly, the proportion of the genome covered by each GRB set was calculated and visualised against genome size.

To investigate the composition of GRBs, the locations of all exons, introns and repeat regions were retrieved for each genome. The exon and intron coordinates were retrieved from ENSEMBL for all species except OikDioicaNorway (Zerbino et al. 2018). The OikDioicaNorway gene annotation was retrieved from OikoBase (Danks et al. 2013). The repeat coordinates were retrieved from the RepeatMasker (*RepeatMasker Open-4.0.*) database for all species except for OikDioicaNorway and tetNig2. RepeatMasker generated tetNig2 repeats were downloaded from the UCSC genome browser, while the OikDioicaNorway repeats were specifically generated for this analysis using RepeatMasker with default settings. For each species, the remaining sequence which overlapped neither exons, introns nor repeats was designated intergenic sequence.

To identify the enrichment or depletion of specific features within GRBs compared to the rest of the genome, a chi-squared test was performed for the observed abundance of each feature within GRBs versus an expected abundance based on the proportion of the genome covered by that feature. The average width of exons, introns and repeats within GRBs was also compared to the average width in the rest of the genome. For this comparison, the mean width of a particular class of element within GRBs was visualised as a proportion of the mean width in the rest of the genome. For clarity of visualisation, the log2 of these proportions was taken. A similar process was followed to visualise the average density of each class of element within GRBs compared to the genome.

4.3 Results

4.3.1 CNEs within compact genomes occur in clusters

To evaluate the viability of GRB detection in highly compacted genomes, I first assessed the occurrence and distribution of CNEs in the ce10 and OikDioicaNorway genomes. In the ce10 genome, I identified CNEs (96.6% identical over 30bp) by pairwise comparison with *Caenorhabditis briggsae*, while in the OikDioicaNorway genome, I identified CNEs (98% identical over 50bp) by pairwise comparison with another *Oikopleura dioica* strain, OikDioicaJapan. *O. dioica* is extremely fast evolving, making it possible to detect signatures of extreme non-coding conservation between strains of the same species (Denoeud et al. 2010). CNE detection yielded 8105 ce10 CNEs with a mean width of 38.77bp and 40275 OikDioicaNorway CNEs with a mean width of 81.62bp. The much larger number of CNEs identified in OikDioicaNorway is due to the short evolutionary distance of the species comparison. It would be preferable to use a more distantly related species for CNE identification, unfortunately the only other sequenced *Oikopleura* genomes are those of *O. albicans* and *O. vanhoeffeni*, both of which are too distantly related to *O. dioica* for CNE detection (unpublished data).

Next, I visualised the CNE distribution across the genomes of both species (Figure 4.1). In these figures we can see the location of all CNEs on the three largest chromosomes (or scaffolds in the case of OikDiocaNorway) for both species. In Figure 4.1A and B it appears that the CNEs are quite evenly spread across the chromosome, as evidenced by the even diagonal line formed by the CNE locations. This can be caused by high background conservation or the absence of CNE clustering, however in this case it is due to the compact nature of these two genomes. In Figure 4.1C we can see the CNE distribution, but across a 500kb region of scaffold_2 in OikDioicaNorway and chrV in ce10. In this figure it is clear that the distribution of CNEs is made up of very small clusters (vertical lines) interspersed with sparsely populated regions. There appear to be very short gaps between clusters of CNEs, thus giving the impression that the genome is evenly covered by CNEs as seen in Figure 4.1A-B.

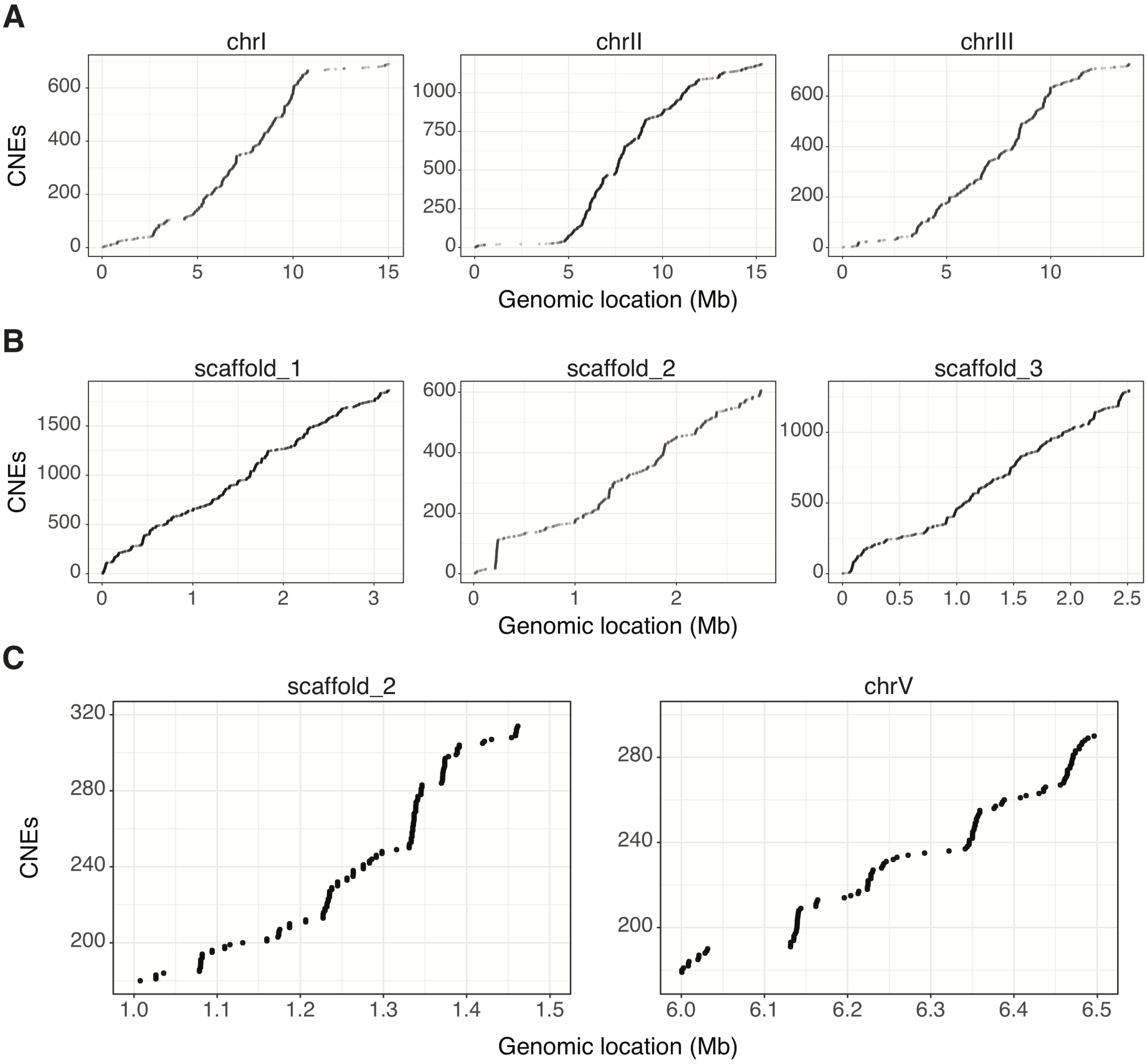

Figure 4.1: CNE distribution in the ce10 and OikDioicaNorway genomes. (**A**) The distribution of CNEs across the three largest ce10 chromosomes. The x-axis represents the genomic location, while the y is the CNE index. (**B**) The distribution of CNEs across the three largest OikDioicaNorway scaffolds. (**C**) The CNE distribution across a 500kb region of scaffold_2 in OikDioicaNorway and chrV in ce10.

These results suggest that GRBs may exist within these small genomes, but that they are much reduced in size compared to species in which we have previously identified GRBs. Further, the spaces between CNE clusters also appear to have been greatly reduced upon genome compaction.

4.3.2 GRB identification in compact genomes

As this was the first attempt at GRB identification in such compact genomes, I predicted GRBs using both CNE- and kurtosis-based conservation and compared the quality of the predicted GRBs identified using both methods. To investigate general features of compact GRBs that are independent of GRB identification parameters, multiple GRB merging quantiles and, for kurtosis-based conservation, bin sizes for conservation calculation, were used for GRB identification. The results of GRB prediction are presented in Figure 4.2 (Detailed results in Table S2).

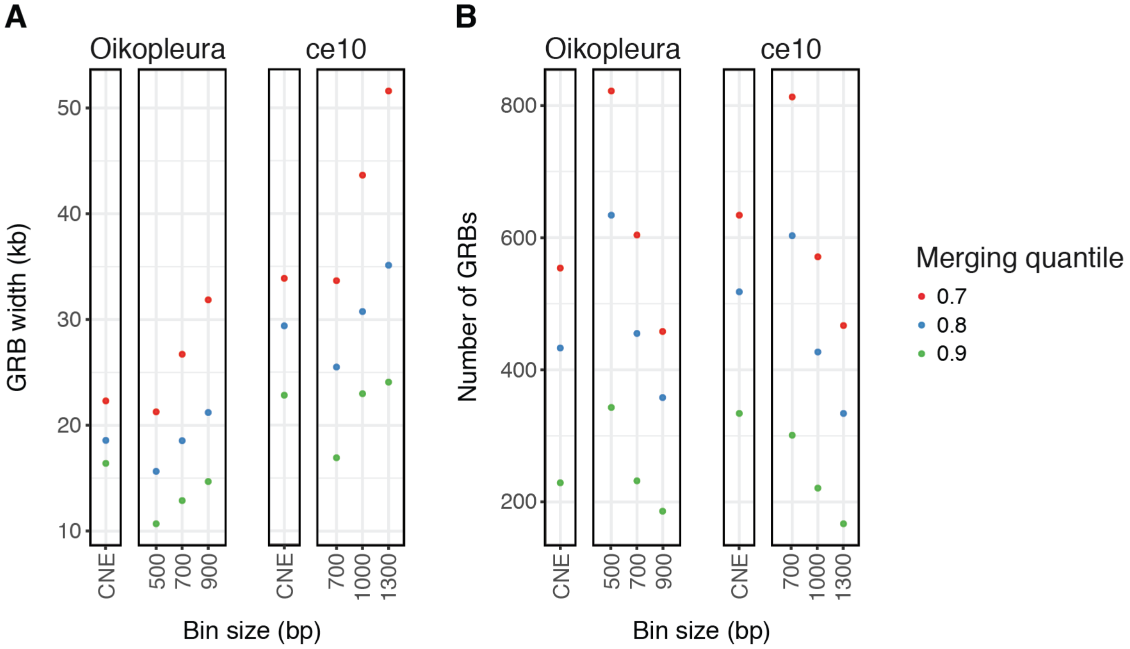

Figure 4.2: The properties of GRB sets identified in OikDioicaNorway and ce10. The mean width (**A**) and number (**B**) of GRBs identified using either CNE- or kurtosis-based GRB identification. GRB sets are split by GRB merging quantile, and in the case of kurtosis-based GRB sets, bin size used in the GRB identification pipeline. As there is no bin size paramter used in the identification of CNE-based GRBs, instead of a bin size they are marked with "CNE" on the x-axis.

From this figure we can see that, in general, CNE-based and kurtosis-based GRB identification predict a similar number of putative GRBs with similar mean widths. Overall, I identify between 167 and 813 putative GRBs in ce10 with mean widths ranging from 23kb to 51.6kb. In OikDioicaNorway I identify between 186 and 822 putative GRBs with mean widths ranging from 14.7kb to 31.9kb. At the most stringent merging quantiles and largest bin sizes, the number of putative GRBs identified is low compared to sets previously identified in invertebrates and mammals (Harmston et al. 2017), however the majority of parameter combinations predict a similar number of GRBs to past studies. In both species all parameter combinations yielded extremely narrow GRBs, suggesting they have undergone a similar degree of compaction to the genome. This is consistent with the results presented by Harmston et al. (Harmston et al. 2017).

It appears that the merging quantile and bin size have a larger influence on both the mean width and number of putative GRBs identified than the method used. In general, when using the same merging quantile, the width of GRBs identified using kurtosis-based conservation increase with increasing bin size, while the number of GRBs identified shows the opposite trend. This could be due to an increased chance that a narrower bin may contain no, or very few, runs of perfect sequence identity. These breaks in the continuity of high-kurtosis within a GRB may then result in the fragmentation of large GRBs into multiple smaller GRBs

Based on visual inspection of the concordance between GRB sets and CNE-density or kurtosis-based conservation, I selected one CNE-based set and one kurtosis-based set of putative GRBs to use for the remainder of the analysis. For ce10 the CNE-based set used was predicted using a merging quantile of 0.8 and the kurtosis-based set used was predicted on kurtosis-based conservation calculated in 1kb bins with a merging quantile of 0.7. For OikDioicaNorway the CNE-based set was predicted using a merging quantile of 0.8 and the kurtosis-based set used was predicted on kurtosis-based conservation calculated in 500bp bins using a merging quantile of 0.7. While for the remainder of the chapter I will present results based on these sets of putative GRBs, the other sets produced very similar results for all analyses.

4.3.3 CNE-based and kurtosis-based GRBs in compact genomes are highly concordant

Next, I assessed the degree of overlap between CNE-based and kurtosis-based GRB predictions to confirm that both methods were identifying biologically similar sets of candidate GRBs. To this end, I visualised the enrichment of CNE density within kurtosis-based GRBs and vice versa (Figure 4.3A). In these figures, each row is a genomic window centred on a GRB, with the putative GRBs ordered by width. The intensity of the colour represents the CNE density or kurtosis-based conservation respectively. From this figure it is clear that in both species there is an enrichment of CNE density within kurtosis-based GRBs, and also an enrichment of high kurtosis-based conservation scores within CNE-based GRBs. This is compelling evidence that both techniques are identifying strong conservation in similar genomic loci.

To further investigate the concordance between CNE-based and kurtosis-based GRB predictions, I plotted genomic windows centred on putative GRBs and coloured regions based on whether they were predicted to be in a GRB in either the CNE-based set, the kurtosis-based set, both sets or neither set. This was performed for GRB predictions derived from each method in turn in both species (Figure 4.3B). There is significant overlap of the putative GRBs identified by the two methods, as shown by the strong enrichment of red within the funnels in these figures. To quantify this observation, the Jaccard coefficient for the two sets was calculated for each species. In this context, the Jaccard coefficient measures the overlap of two sets of ranges by calculating the intersect of the two sets divided by their union. In ce10, the Jaccard coefficient for CNE-based and kurtosis-based GRBs is 0.47, while for OikDioicaNorway it is 0.41. In both species, the GRB predictions derived from each

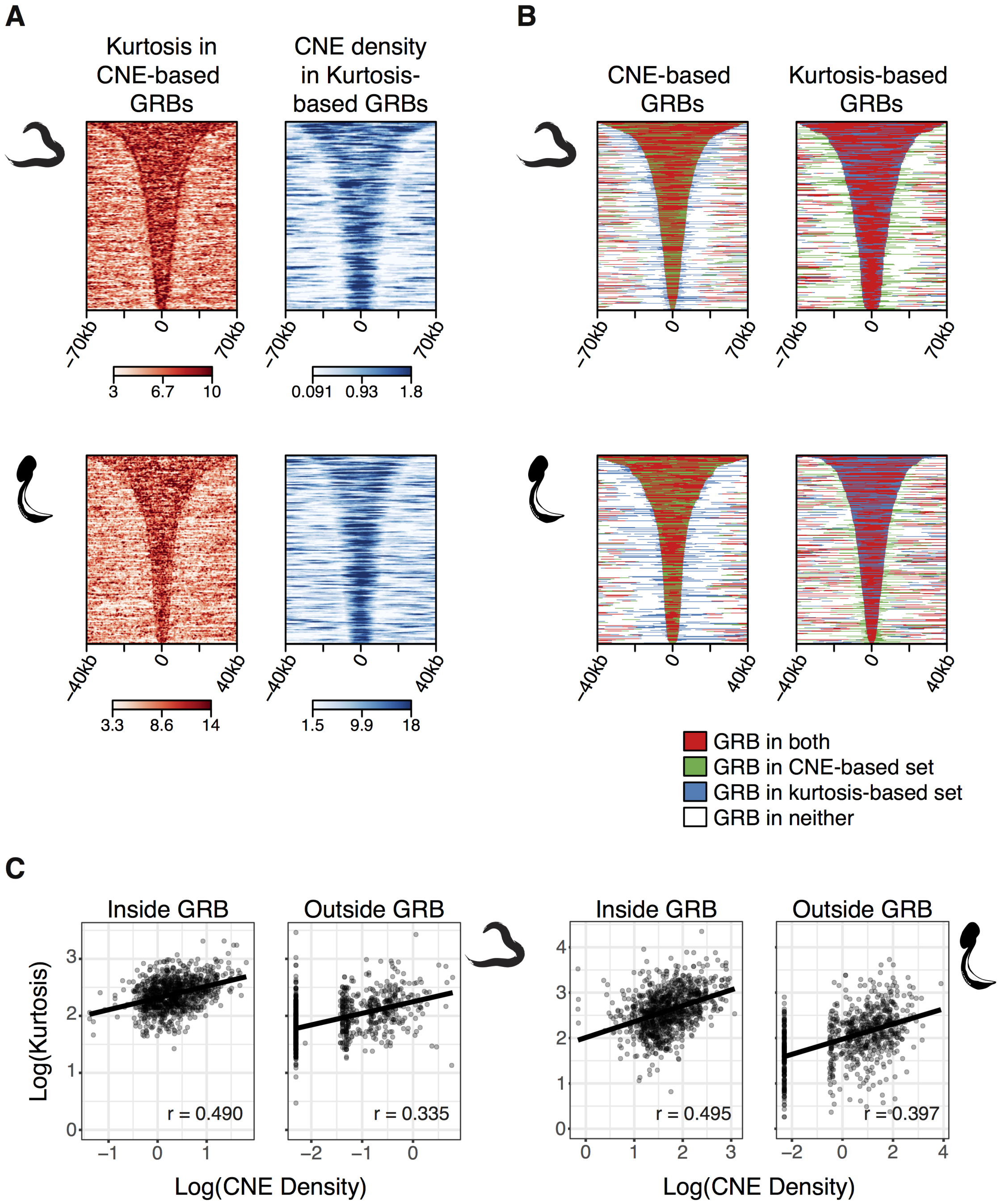

Figure 4.3: Comparing CNE-based and kurtosis-based GRB predictions. (**A**) For both ce10 (above) and OikDioicaNorway (below), CNE density was visualised within genomic windows centred on kurtosis-based GRB predictions. Conversely, kurtosis-based conservation was visualised in genomic windows centred on CNE-based GRB predictions. Both measures are enriched within their counterpart's predicted GRB, highlighting the overlap of the GRB predictions derived from each method. (**B**) Genomic windows, centred on either CNE-based or kurtosis-based GRB predictions, were coloured by whether they fell within predicted kurtosis-based GRBs, CNE-based GRBs, both, or neither. (**C**) The correlation between maximum CNE density and kurtosis-based conservation in 1000 random windows sampled from GRB and non-GRB windows in both species.

method are very similar, but are slightly less overlapping that GRBs identified using the two methods in human, in Chapter 2 on page 64 (human to opossum, Jaccard = 0.52). Further when examining the correlation between kurtosis-based conservation and CNE density in 1000 random windows, derived from GRB and non-GRB loci across the genome (Figure 4.3C), it is clear that the correlation is much lower than in the human genome (Figure 2.1C). This could be due to difficulty in assigning an appropriate minimum length for CNE density in such highly compacted genomes.

While these results provide clear evidence that, in both species, the two sets derived from each method are largely overlapping, it is also true that they are not completely interchangeable. The degree of overlap of GRB predictions derived from the two methods suggests that they are genuine GRBs, however biological validation is required.

4.3.4 Chromatin modifications in *C. elegans* and *O. dioica* GRBs

GRBs tend to be broadly marked with H3K27me3 in adult tissues (Harmston et al. 2017), and in Chapter 3 I showed that several chromatin modifications exhibit either enrichment or depletion that is clearly delineated by GRB boundaries. Here, I use these known features to simultaneously assess the quality of the CNE- and kurtosis-based GRB predictions, and to characterise the chromatin features of the ce10 and OikDioicaNorway GRBs.

For OikDioicaNorway, I retrieved publicly available ovary H3K27ac, H3K27me3, H3K4me3 and H3K4me1 ChIP-chip data (Navratilova et al. 2017) and visualised their enrichment within genomic windows centred on both CNE-based and kurtosis-based GRB predictions (Figure 4.4A). I also extended putative GRBs to 3 times their width, binned the resulting windows and plotted the average ChIP-chip enrichment within each bin for each chromatin modification (Figure 4.4B). From these figures we can see that in both sets of GRBs are enriched for

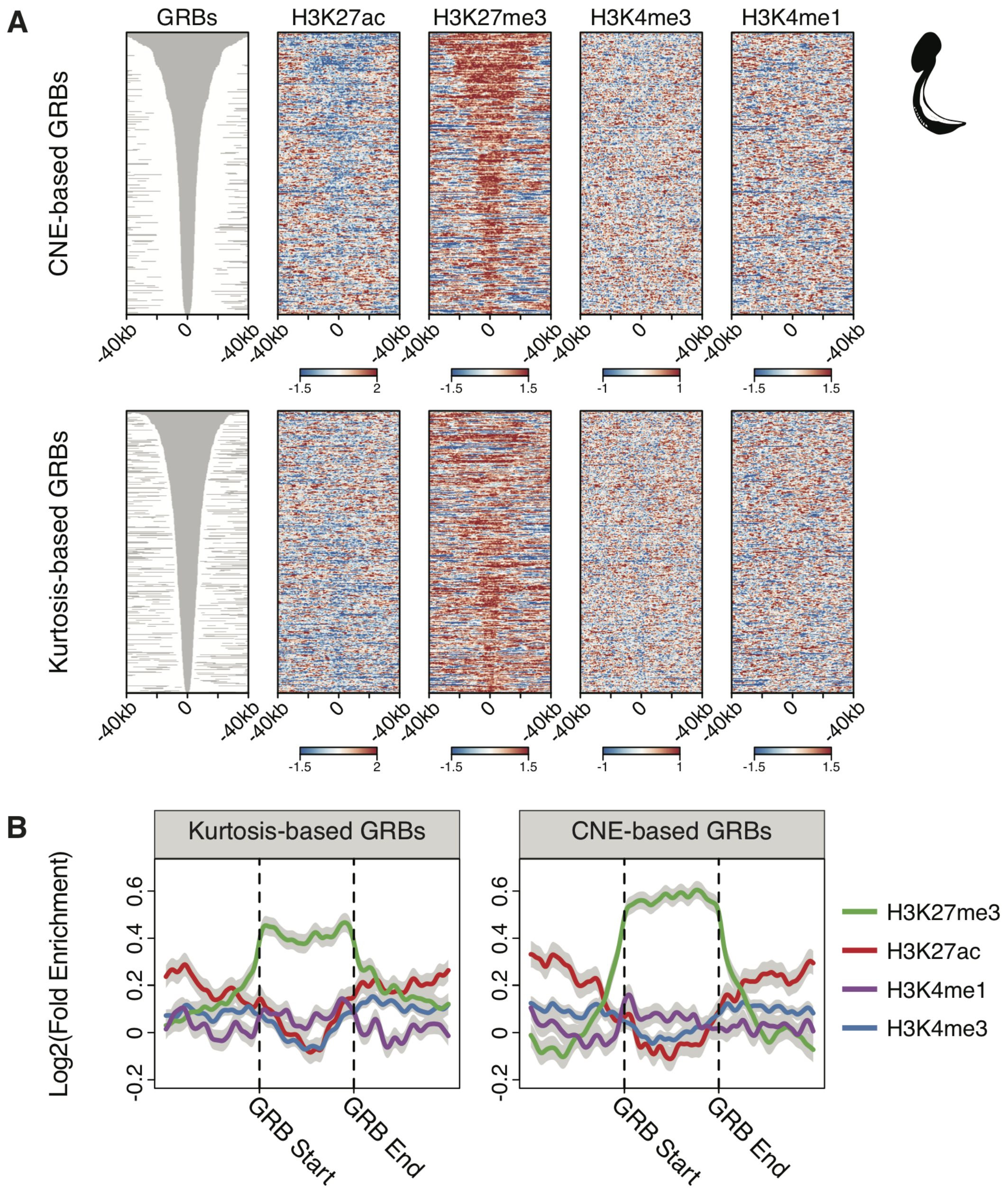

Figure 4.4: Chromatin modifications inside Oikopleura kurtosis- and CNE-based GRBs. (**A**) Heatmaps showing the enrichment of OikDioicaNorway ovary derived, chromatin modification, ChIP-chip data in genomic windows centred on either CNE-based or kurtosis-based GRBs. The grey heatmap shows the extent of the GRBs. (**B**) Metaplots showing the average enrichment of each chromatin modification in CNE-based and kurtosis-based GRBs.

H3K27me3. This is similar to what we observe in other GRBs, and expected as the ovary is a terminally differentiated, highly specialised tissue. GRB target genes are generally expressed in development, and thus GRBs should be silenced in this context. When comparing GRB predictions, we can see that H3K27me3 more clearly demarcates the span of the CNE-based GRBs than the kurtosis-based GRBs. There is also a visible depletion of H3K27ac in the CNE-based GRBs that is not apparent in the kurtosis-based GRBs (Figure 4.4A). These results suggest that, in OikDioicaNorway, the CNE-based GRB boundaries are more reliably estimated than the kurtosis-based GRBs. Therefore for the rest of the analysis I will use this set as my OikDioicaNorway GRBs.

For ce10, I retrieved L3 larval H3K27ac, H3K27me, H3K4me3 and H3K4me1 ChIP-seq data from modENCODE (Gerstein et al. 2010). As with OikDioicaNorway, I visualised the coverage of each chromatin modification within genomic windows centred on both CNE-based and kurtosis-based GRB predictions in turn (Figure 4.5). In contrast to OikDioicaNorway, in ce10 there is a clear enrichment of H3K27ac and H3K4me1 inside both CNE-based and kurtosis-based GRBs. H3K27ac and H3K4me1 in combination indicate the presence of active enhancers, however it is unusual to see such broad domains covered by both modifications together. In L3 larvae, there are still several developing tissues, in particular the gonadal tissues (Kimble and Hirsh 1979), and therefore the enrichment of active marks may be driven by a subset of GRBs that are still active in the L3 stage. Indeed, when clustering GRBs based on their chromatin state, approximately one third of GRBs are enriched in H3K27ac and H3K4me1 and depleted of H3K27me3 (Figure 4.6). The remainder do not appear to be strongly enriched for any modification, however there is a weak enrichment of H3K27me3, within these GRBs. This suggests that these GRBs are polycomb repressed and silent.

When examining the meta-profiles of the clusters derived from ce10 CNE-based and kurtosis-based GRBs, it is obvious that there is a much sharper boundary effect on the chromatin modifications in the kurtosis-based GRBs than the CNE-based GRBs (Figure 4.5B and Figure 4.6B). This implies that the kurtosis-based

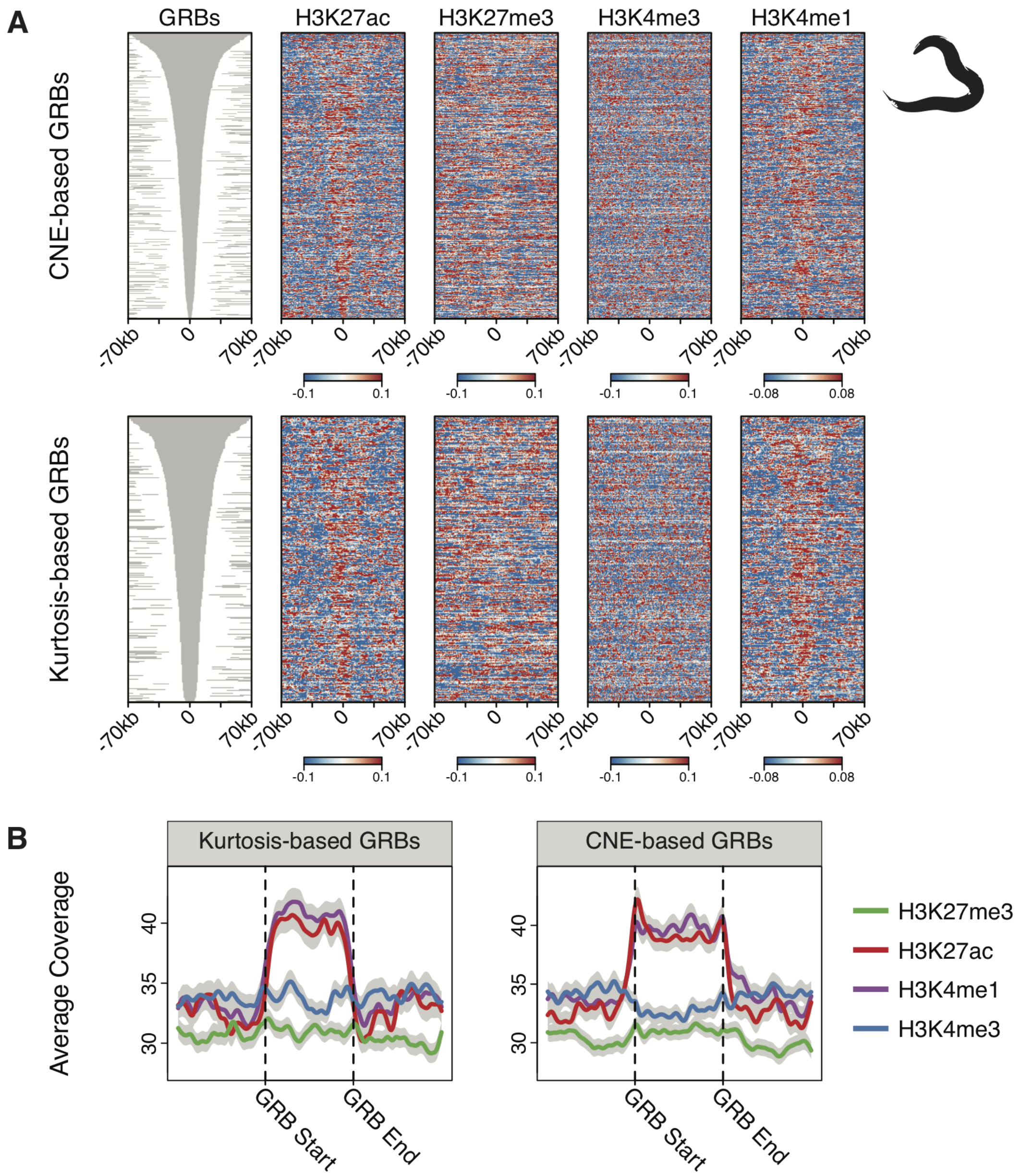

Figure 4.5: Chromatin modifications inside *C. elegans* kurtosis- and CNE-based GRBs. (**A**) Heatmaps showing the coverage of ce10 L3 larvae derived, chromatin modification ChIP-seq data in genomic windows centred on either CNE-based or kurtosis-based GRBs. The grey heatmap shows the extent of the GRBs. (**B**) Metaplots showing the average coverage of each chromatin modification in CNE-based and kurtosis-based GRBs.

GRBs predict the boundaries of GRBs better than the CNE-based set, and thus for the remainder of this analysis I use the kurtosis-based GRBs as my ce10 GRBs.

Taken together, the strong enrichment and depletion of specific chromatin modifications within the predicted ce10 and OikDioicaNorway GRBs is convincing

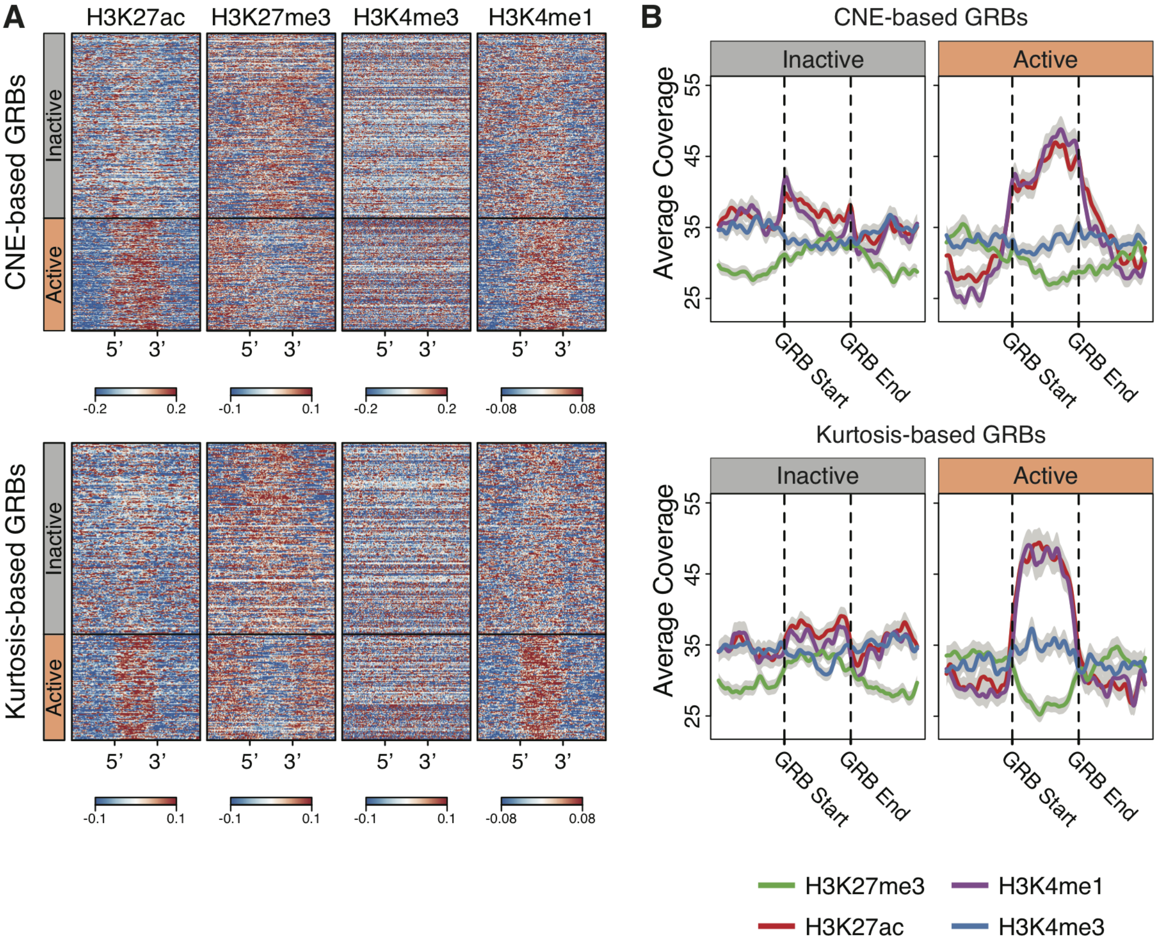

Figure 4.6: A subset of *C. elegans* GRBs appear to be active. (**A**) Heatmaps of chromatin modification enrichment within CNE-based and kurtosis-based GRBs. GRBs were extended to three times their width and the resulting windows were divided into bins. The windows were then clustered based on their histone modification patterns. The x-axis indicates the position of the 5' and 3' boundary of the GRBs. For both sets of GRBs approximately one third of the GRBs are covered by active chromatin modifications. (**B**) Metaplots showing the average coverage of chromatin modifications within the two clusters of GRBs defined in the heatmaps.

evidence that GRB detection successfully identified biologically significant regions of high non-coding conservation in highly compacted genomes for the first time. Further, given the similarities between these GRBs and those that have been detected in non-compacted genomes, it is likely that they are functionally equivalent.

4.3.5 Gene ontology enrichment analysis of *C. elegans* and *O. dioica* GRB genes

To further assess the functional equivalence of GRBs in compact genomes and those in larger genomes, I performed GO analysis on all genes within ce10 and OikDioicaNorway GRBs. Similar to Chapter 3, GRB genes were compared to a universe of all annotated genes. Figure 4.7 shows the top 15 most enriched BP and MF terms for GRB genes in ce10 and OikDioicaNorway.

In Figure 4.7A we can see that OikDioicaNorway GRB genes are very strongly enriched for BP terms related to the regulation of transcription such as "nucleic acid-templated transcription", "regulation of RNA biosynthetic process" and "positive regulation of transcription from RNA polymerase II promoter". These are terms frequently associated with transcription factors, and indeed the only MF terms enriched in OikDioicaNorway GRB genes are "DNA binding transcription factor activity" and "sequence-specific DNA binding" confirming that these are transcription factor genes. The OikDioicaNorway GRBs are also strongly enriched for general developmental BP terms such as "multicellular organism development" and "animal organ morphogenesis". These results confirm that there is an over representation of developmental transcription factor genes within OikDioicaNorway GRBs. This is in line with what has previously been described for GRB target genes, thereby confirming both the quality of the predicted OikDioicaNorway GRBs, and their similarity to GRBs identified in larger genomes.

In ce10, the picture is slightly different. Looking at the ce10 enriched MF terms, there is still an enrichment of transcription factor related terms such

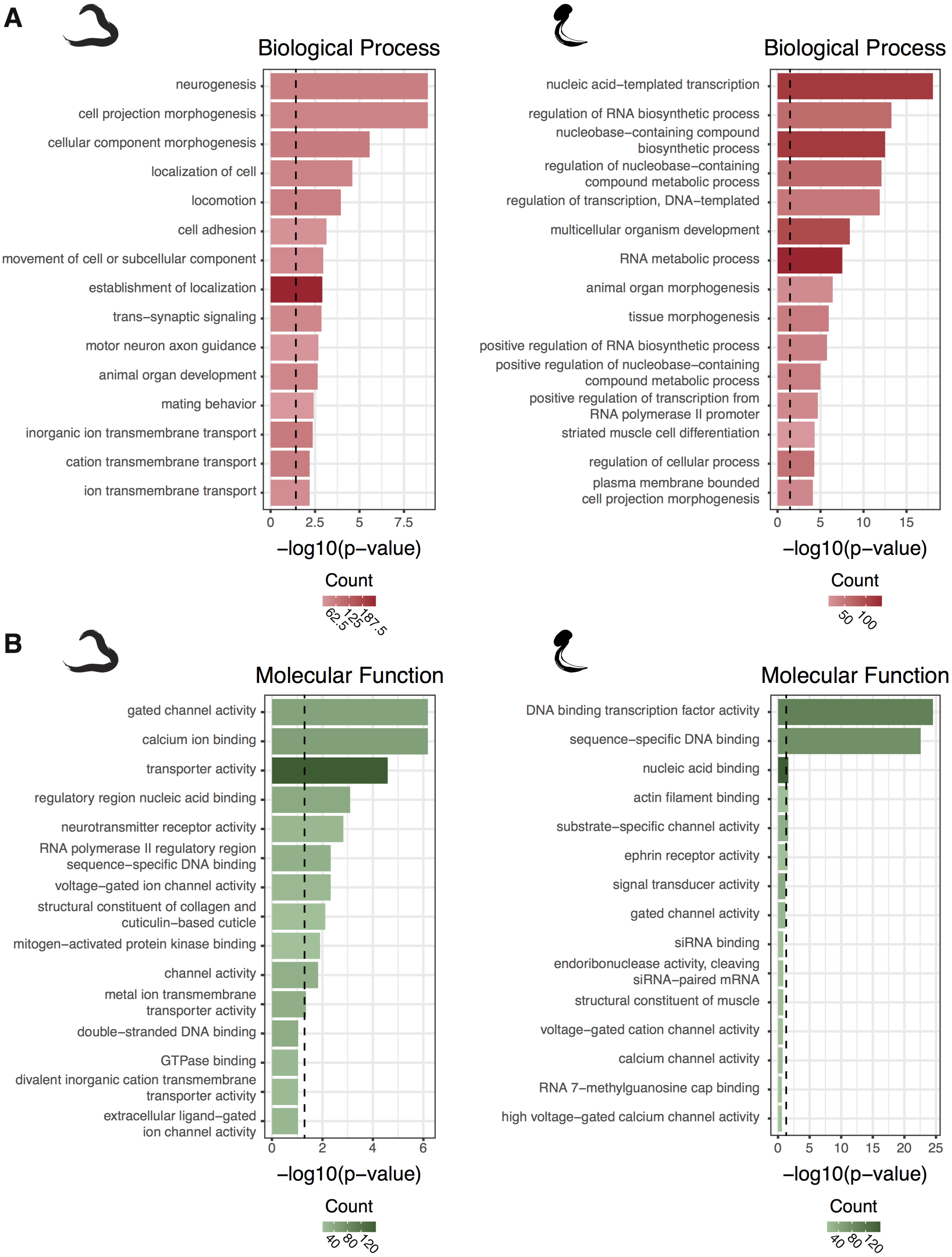

Figure 4.7: GRB gene ontology enrichment analysis. Gene ontology (GO) enrichment analysis was performed for GRB genes from ce10 (left) and OikDioicaNorway (right) GRBs. The multiple test corrected -log10 p-values for the top 15 enriched (**A**) biological process and (**B**) molecular function terms from each species are shown. In each case the dashed line represents the threshold of statistical significance.

as "regulatory region nucleic acid binding" and "RNA polymerase II regulatory region sequence-specific DNA binding", however, the most enriched and abundant terms are related to cell-cell communication. Further, when examining the enriched BP terms there is a strong enrichment for terms related to neural development, cell adhesion and cell-cell signalling. This is unexpected, as the majority of the terms are far more specific than the developmental terms usually observed for GRB genes. Overall, the enriched BP and MF terms suggest that GRBs in ce10 tend to regulate a mixture of developmental transcription factors and components of cell signalling pathways involved in neurogenesis, axon guidance and general cell adhesion and communication. These results could be influenced by the overall tissue composition of *C. elegans*. Of the 959 somatic cells in an adult animal, 302 are neurons, and therefore perhaps the enrichment in neural developmental terms is due to the complexity of the ce10 nervous system relative to the overall complexity of the animal (White 1988). If the majority of genes that require GRB-like regulation are related to neural development, it is likely that GO analysis of GRB genes would identify enriched neural development terms. While the most enriched GO terms in ce10 GRB genes are more specific than what we expect, they are all still children of the terms frequently associated with GRBs.

In summary, both OikDioicaNorway and ce10 appear to target genes typically associated with GRBs, thereby confirming their functional equivalence with previously identified GRBs. Compared to previously defined GRB target genes, ce10 GRBs are more frequently associated with neural development and cell communication genes. This enrichment in axon guidance and cell adhesion GO terms in the general population of ce10 GRB genes is interesting as these are the terms enriched in high-turnover GRBs identified in three independent phylogenies in Chapter 3.

4.3.6 GRB size is proportional to genome size

To assess the effect of genome compaction on GRB size, and to identify the overall relationship between GRB and genome size, I identified GRBs for all the species comparisons listed in Table 4.4. The number and size of GRBs identified in all species are listed in Table S3. Similar to Chapter 3, overall the mean GRB width increases and total number of predicted GRBs decreases as bin sizes for kurtosis calculation are increased. There is also a strong trend towards decreasing GRB number and width with increasing merging quantile. This is due to the increased stringency in the GRB prediction method, requiring regions to have higher non-coding conservation to be predicted as a GRB. To assess the relationship between GRB and genome size, I plotted the mean width of GRBs predicted for each bin size, split by GRB merging quantile, for each species. From Figure 4.8A, it is clear that regardless of the bin size used in the kurtosis-based conservation calculation, or the merging quantile used in the GRB prediction, there is a very strong linear relationship between GRB and genome size. This result shows that GRBs have indeed undergone either compaction or expansion at a comparable rate to genome compaction or expansion. Further, identification of such compact, functional GRBs in the OikDioicaNorway genome suggests that either there is no maximal level of compaction a GRB can achieve before becoming nonfunctional, or we have not yet identified GRBs in compact enough genomes to identify such a limit.

The ancestral metazoan genome underwent significant innovation with respect to gene content, predominantly acquiring novel nucleic acid binding, transcription factor and cell signalling genes - genes frequently found within GRBs (Paps and Holland 2018). Further, the evolution of long-range gene regulation was likely one of the major drivers of the transition to multicellularity (Sebé-Pedrós et al. 2016). These results suggest that GRB-like gene regulation first emerged in the metazoan ancestor and therefore it is plausible that there are a core set of GRBs present in all metazoa. In support of this, the proportion of the genome covered by GRBs, at the

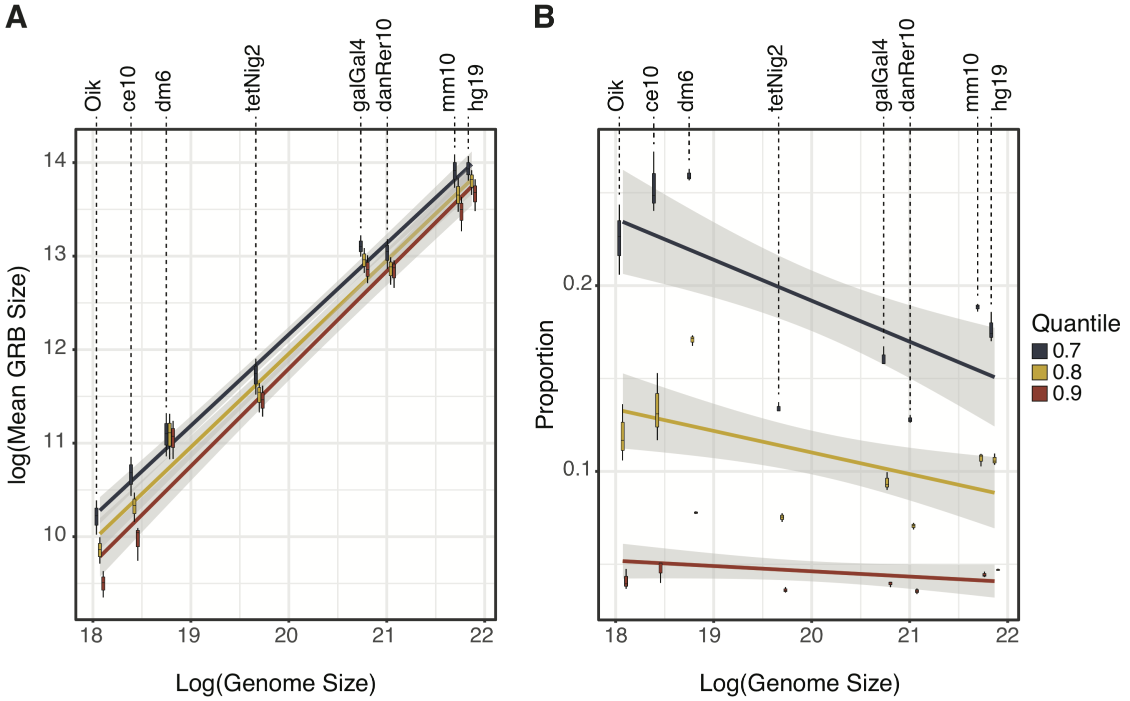

Figure 4.8: GRB size and proportion of the genome covered by GRBs vs genome size. (**A**) The log of the mean GRB size from GRB sets predicted using kurtosis-based conservation calculated in three bin sizes plotted against log genome size. GRB sets were stratified by the merging quantile used in GRB prediction. A linear model was fit to the data for each merging quantile separately, and the grey region around the fitted line represents the 95% confidence interval of the model fit. (**B**) For all GRB sets identified at a GRB merging quantile, the proportion of the genome covered by the predicted GRBs was plotted against the log of the size of the genome from which they were derived. Again, a linear model was fit to the data for each merging quantile.

most stringent GRB prediction thresholds, is approximately equal in all genomes (Figure 4.8B). Further, the most stringent GRB prediction also identifies approximately the same number of GRBs in all species, despite the total number ranging widely at the less stringent thresholds (Figure S3). As the stringency of the merging quantile is reduced, the proportion of the genome covered by GRBs increases in all species. This effect is most apparent in the compact genomes, likely due to reduced spacing between adjacent GRBs in compact genomes.

4.3.7 GRB composition is under similar selective pressure in all species

GRBs tend to be distinct from the genome with respect to gene density, repeat content and intron size (Engström et al. 2007; Akalin et al. 2009). Each of the genomes used in this analysis have undergone species-specific gene duplication and loss events, repeat integration and deletion, and general expansion and contraction. While it is clear from Figure 4.8A that GRBs expand and contract proportionally with the genome in which they are identified, it is unknown whether this expansion and contraction occurs via the proportional gain or loss of the same genomic features as the rest of the genome; or whether GRBs tend to maintain a specific environment which is favourable for long-range gene regulation. To address this question, I retrieved, or generated, gene and repeat annotation for all species and compared the composition of GRBs and the genome with respect to exons, introns, repeats and intergenic regions.

Figure 4.9A shows the proportion of the genome and GRBs covered by each genomic feature for each species. Genome composition varies substantially between species. With decreasing genome size there is a trend towards an increase in the proportion of exonic sequence. The large genomes contain a greater proportion of repeats, with the exception of the chicken genome which is known to have a relatively low repeat content (Wicker et al. 2005). The intergenic portion of the genome does

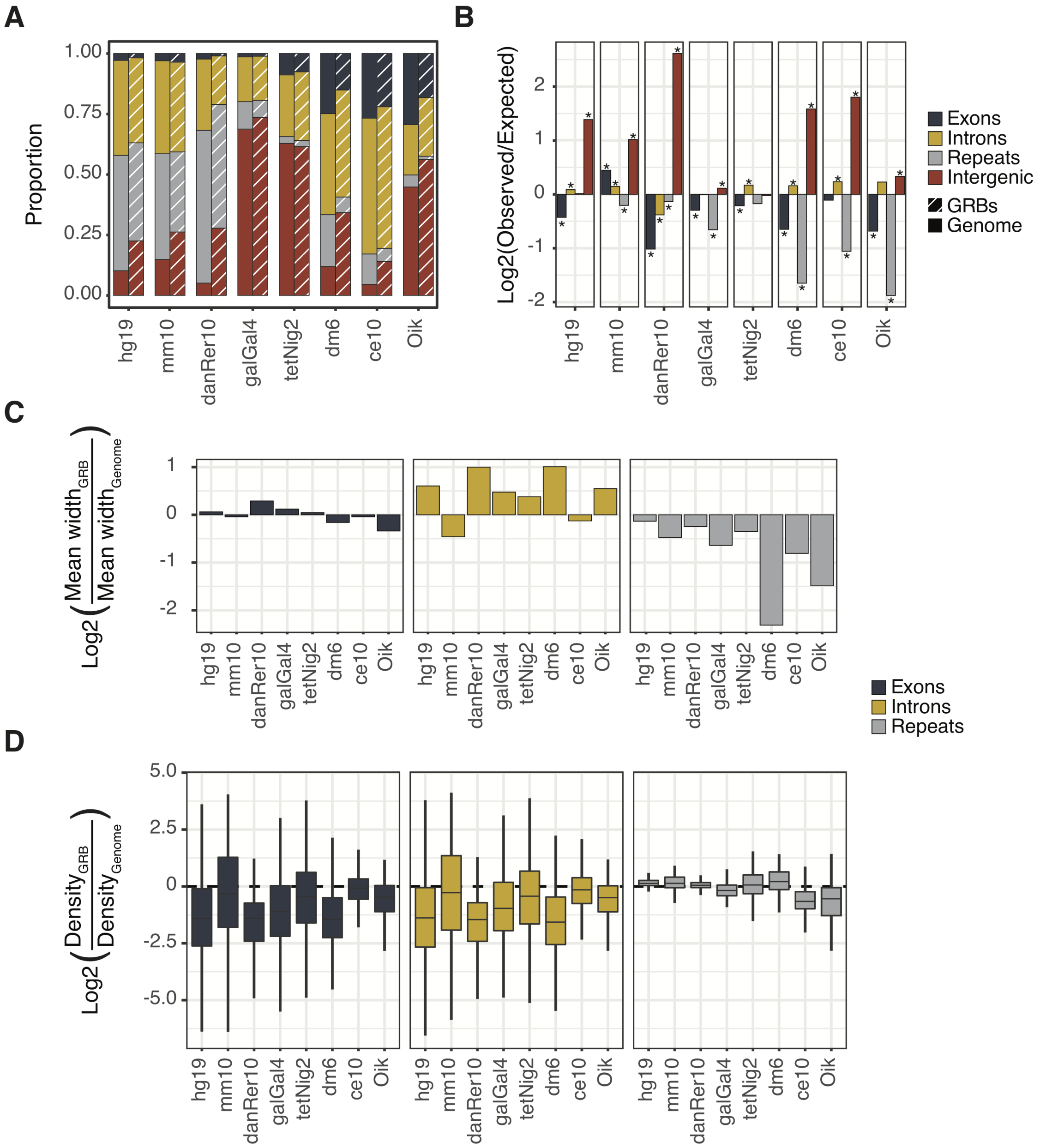

Figure 4.9: GRB and genome composition. (**A**) Each species exonic, intronic, repeat and intergenic content, presented as a proportion of the total genome or GRB size. Species are ordered from left to right by decreasing genome size. (**B**) The enrichment of each genomic feature within GRBs presented as log2 of the observed content divided by the expected content. For each species, the expected GRB content of each feature was based on the whole-genome content. The stars above each bar represent whether the enrichment or depletion was statistically significant as defined by a Chi-squared test. The mean width (**C**) and density (**D**) of exons, introns and repeats within GRBs compared to the genome average.

not show a clear trend relative to genome size. In general it is expected that the compact genomes should contain less intergenic space relative to the larger genomes, however this is not the case in the OikDioicNorway and tetNig2 genomes. This may be due to incomplete annotation in these genomes, increasing the proportion that appears to be unannotated intergenic space. When comparing GRB and genome content, it is clear that in each species GRB content is significantly different to genome content. This suggests that there is selective pressure to maintain a specific regulatory environment within GRBs, and therefore GRB contraction and expansion may occur by specific gain and loss of particular genomic features, independent of the general pattern observed in the rest of the genome.

To better assess the differences between a species genome and GRB content, I directly visualised the enrichment or depletion of specific genomic features in GRBs compared to the genomic background (Figure 4.9B). In general GRBs are depleted of exonic and repeat sequence, and enriched for intergenic and intronic sequence. Further, it appears that GRBs in compact genomes are more depleted of repeat sequence than larger genomes. Examining the mean width of exons, introns and repeats within GRBs compared to the genome-wide mean (Figure 4.9C), it is clear that, in general, repeats within GRBs are smaller than the rest of the genome, while introns tend to be longer within GRBs. To complete the picture of GRB vs genome content, I also calculated the density of exons, introns and repeats in each GRB and compared it to the density within the genome (Figure 4.9D). GRBs tend to contain fewer exons and introns than the rest of the genome, however in all but the most compact genomes, the repeat density is very similar to the genome average.

The reduced density of exons and introns within GRBs is expected, as GRBs tend to occur within gene deserts, frequently only containing the gene that is subject to regulation by that GRB. The increased mean width of introns within GRBs has been previously reported within *D. melanogaster* (Engström et al. 2007), however here we can see that the this phenomenon is observable in all but two of the species analysed. In general the enrichment of intergenic and intronic space within GRBs is due to these regions containing the arrays of conserved enhancers that constitute the GRB. The depletion of repeat content is likely due to the potential for repeat insertion to disrupt existing enhancers or to act as enhancers themselves, thereby causing misregulation of the target gene. This phenomenon might also explain the observation that GRBs in the larger genomes tend to have a similar density of repeats to the rest of the genome, but shorter repeats. Insertion of a short element is less likely to disrupt the spacing and interaction between existing enhancers within the GRB. The trend towards stronger depletion of repeats in the more compact genomes is likely due to the already reduced space which can be occupied by enhancers. This would increase the chance that repeat insertion could disrupt the regulation of the target gene.

It is clear that despite the differences in GRB composition between species, compared to their genomic background composition, GRBs are similarly enriched and depleted of specific genomic features. This suggests that GRBs are functionally similar and under common constraints to maintain a favourable environment for long-range regulation in all of the analysed genomes, regardless of genome size or composition.

4.4 Discussion

In this chapter I have successfully identified GRBs in *C. elegans* and *O. dioica* for the first time. In both species, chromatin modification data shows that the predicted GRBs are biologically distinct from background genomic sequence. In *O. dioica* ovaries these GRBs are broadly marked by H3K27me3, showing that they are polycomb repressed in this tissue. This agrees with previous studies on chromatin state at GRBs (Akalin et al. 2009; Harmston et al. 2017). Surprisingly, the predicted *C. elegans* GRBs are not strongly marked by H3K27me3 at the L3 larval stage, instead a subset of the GRBs are broadly marked by H3K27ac and H3K4me1 - modifications that, in combination, suggest the presence of active enhancers. It is possible that this subset of GRBs are still active during L3 larval stage, directing the remainder of the developmental trajectory of *C. elegans* larvae. Interestingly, the subset of *C. elegans* GRBs that do not show active chromatin marks at the L3 stage also don't show an enrichment of H3K27me3. As these GRBs are presumably no longer active at this stage, they should be repressed. It is possible that at this stage these GRBs have already been compacted into heterochromatin domains. The inclusion of H3K9me3, a modification found specifically at heterochromatin, would make it possible to investigate this hypothesis. Overall the chromatin modification data shows that in both species, GRBs are distinct from background genomic sequence and thus likely biologically relevant.

As further evidence of the quality of the predicted GRBs, and their equivalency with previously identified GRBs, I showed that *C. elegans* and *O. dioica* GRBs are enriched for developmental genes. *O. dioica* GRBs are enriched for developmental transcription factors, genes which are typically associated with GRBs (Kikuta et al. 2007a; Engström et al. 2007; Akalin et al. 2009), while in *C. elegans* GRBs there is an enrichment of developmental genes, but not a strong enrichment of transcription factors. *C. elegans* GRB genes were most enriched for cell signalling, cell adhesion and neural development terms. Interestingly, these are the gene families I found to be enriched in the htGRB target genes in Chapter 3 This may be due to *C. elegans* having simpler, more deterministic development, resulting in a greater relative proportion of cell-cell communication and axon guidance genes under long-range regulation than early patterning TFs. The identification of GRBs, that appear to be functionally equivalent to vertebrate GRBs, in two more non-vertebrate genomes, provides further support for the hypothesis that GRBs are ancient feature of metazoan gene regulation, rather than a vertebrate innovation resulting from multiple rounds of vertebrate specific CNE recruitment (Lowe et al. 2011). The identification of clusters of non-coding conservation around the same classes of genes described in Lowe et al. in two species which diverged from vertebrates between 600-800 million years ago suggests that GRBs, or GRB-like structures, were already in place in the bilaterian ancestor.

Using these, and other newly identified GRBs, I was able to show that GRB size scales proportionally with genome size. This is in agreement with the results presented by Harmston et al., but expands this preliminary analysis to include the smallest known metazoan genome. Even in the *O. dioica* genome, the relationship between GRB size and genome size remained proportional, with the average GRB size being approximately 25kb. These GRBs are on average narrower than the gaps between pairs of consecutive CNEs that showed conserved spacing between human and tetraodon in Sun et al., suggesting that either there is no minimum distance allowed between consecutive CNEs in GRBs, or that we are yet to identify compact enough GRBs to identify this limit. In the future, a direct analysis of *O. dioica* CNEs may yield further insights into the minimum spacing between CNEs.

In this chapter I also showed that when using the most stringent GRB identification parameters, approximately equal proportions of the genome are covered by GRBs in all species. GRB calling using stringent parameters results in identification of only the most highly conserved regions of the genome, and thus it is tempting to assume that the GRBs identified using these stringent parameters are a core set of essential GRBs identified in all species. Paps et al. found that there is a set of 25 homology groups (or gene families) that were innovated in the metazoan ancestor and remain in the genomes of almost all metazoa today. These essential animal homology groups are enriched for genes frequently associated with GRBs, and therefore it would be interesting to analyse how many of these genes fall in GRBs identified using stringent parameters, and whether perhaps this equal proportion of the genome covered is due to the need to appropriately regulate these essential genes in all genomes.

Finally, in this chapter I showed that regardless of the overall composition of the genome or the GRBs identified in that genome, GRBs tend to be enriched and depleted of the same genomic features. GRBs are enriched for intergenic and intronic sequence, and depleted of exonic and repeat sequence. The enrichment of intergenic and intronic sequence within GRBs is due to reduced gene density and longer introns within GRBs. This is a logical result of GRBs containing multiple conserved enhancers within the intronic and intergenic sequences surrounding the target gene. The depletion of exonic sequence is also due to the general reduction of gene density within GRBs, again likely due to the requirement for increased enhancer-harbouring intergenic space. The depletion of repeat sequence within GRBs appears to be a combination of a reduction in the number and width of repeat elements within GRBs. This may be due to the potential for repeat insertion within a GRB to affect the regulation of the target gene by either disrupting an existing enhancer, or acquiring enhancer activity itself. These results are in agreement with previous studies on GRB content (Engström et al. 2007; Akalin et al. 2009; Harmston et al. 2017), but extends this analysis to extremely compact genomes. In contrast to previous analyses, in this analysis I did not observe a strong reduction of repeat density in the larger genomes analysed. This may be to the treatment of all repeat classes together rather than splitting them by repeat class. Different classes of repeats have different regulatory potential, and thus would differently tolerated within GRBs.

For example, Alu elements in primates resemble enhancers and require only slight modification to acquire regulatory function (Su et al. 2014), while low complexity repeats are less likely to acquire regulatory function. Future analysis of repeat content of GRBs, split by class would likely resolve these disagreements. Overall, the identification of similar selective pressure on GRBs derived from such diverse genomes (in size and content) is yet again support for the hypothesis that GRBs are ancient features of metazoan gene regulation.

Chapter 5

Discussion

In this thesis I have used robust GRB identification methods to analyse the evolutionary dynamics of GRBs in metazoan genomes. First I presented a novel method for calculating pairwise genome conservation and applied a rigorous statistical framework to the problem of GRB boundary identification. Using these newly developed techniques I have explored the evolutionary dynamics of GRBs in distinct metazoan lineages and genomic contexts. In Chapter 2, I outline a novel kurtosis-based measure of genome conservation and show that it performs as well as CNE-based GRB identification in moderate to distantly related species comparisons, but far outperforms CNE-based GRB identification in closely related species. In Chapter 3, I apply this method to identify GRBs for multiple species comparisons in three independent metazoan lineages. I then define deeply and shallowly conserved GRBs in each lineage and show that they share many features between lineages, supporting the hypothesis that GRBs are an ancient feature of metazoan gene regulation. Finally, in Chapter 4 I identify GRBs in the extremely compact *Caenorhabditis elegans* and *Oikopleura dioica* genomes for the first time. I show that these GRBs are functionally equivalent to GRBs in larger metazoan genomes, and go on to assess the impacts of genome compaction on GRB size and composition. In this chapter I will summarise the main results of this thesis and discuss their implications. I will also consider future work that would further our understanding of the origin and evolution of GRBs, and their relationship with 3D genome organisation.

5.1 Kurtosis-based GRB identification

In Chapter 2, I defined a novel kurtosis-based measure of pairwise genome conservation. Previously used CNE-based measures of conservation rely on the selection of arbitrary thresholds for the minimum length and sequence identity required for a region to be defined as conserved. The kurtosis-based approach identifies all stretches of perfectly conserved sequence in bins across the genome, and effectively measures the contribution of extremely long stretches to the distribution in each bin. This measure implicitly takes into account the background conservation of the species comparison, because for a bin to have high kurtosis, it must contain many long runs of perfect conservation, relative to the general genome-wide distribution of all runs of perfect identity. Most importantly, this allows for the identification of highly conserved regions of the genome without the need for arbitrary threshold selection. In the remainder of the chapter I showed that GRB identification using kurtosis-based conservation successfully identifies high quality GRBs in moderately distant species comparisons, and that in close species comparisons it outperforms CNE-based GRB identification. This is likely due to the difficulty in selecting an appropriate threshold for CNE identification in very close species comparisons. For example, in the human to gorilla comparison it was necessary to define a minimum length of 400bp for CNE identification, however the relevance of excluding perfectly conserved sequences of 399bp is dubious at best. The kurtosis-based measure accounts for all lengths of perfect conservation, resulting in far better estimates of human - gorilla GRBs. This method also has great utility in compact genomes (as demonstrated in Chapter 4), as the selection of the minimum length threshold in a genome of 79Mb compared to a genome of 3Gb poses its own problems. Either selecting a minimum size based on the genome size, or selecting the same minimum size regardless of genome size, makes comparison of the results difficult to interpret. Using kurtosis-based conservation it is possible to more systematically compare highly conserved regions in vastly different genomic contexts.

While the kurtosis-based measure is a great improvement on CNE-based approaches for very close species comparisons, it has its own set of potential limitations. During the conservation calculation and subsequent GRB identification pipeline, there is still the need to select two parameters - the bin size in which kurtosis is calculated, and the merging quantile used for merging of adjacent ranges during GRB identification. Both thresholds can affect the size and number of GRBs identified, as shown in Chapter 4. Small bins for kurtosis calculation can result in GRB fragmentation due to the increased chance of a bin containing no stretches of perfectly conserved sequence. This is of particular importance in distant species comparisons due to the reduced conservation between the species. The merging quantile used affects the size and number of GRBs identified, as increasing stringency results in identification of increasingly strongly conserved GRB regions. While this is a limitation, it should also be possible to systematically evaluate the parameter selection to define the most robust set of GRBs for a species comparison. This could perhaps be achieved through GRB identification using multiple combinations of parameters followed by identification of the most consistently identified boundaries. A second potential limitation of the kurtosis-based method is that it does not identify the exact locations of long stretches of perfectly conserved sequence, however if analysis downstream of GRB identification requires this information, the general CNE identification pipeline can be used to define a set of conserved sequences.

Overall, kurtosis-based GRB identification works well and identifies reliable GRBs for species comparisons spanning most evolutionary distances. Further, the implicit compensation for the background conservation of the species compared allows for more reliable and robust comparisons of GRBs identified in multiple species. While it is unlikely that this approach will completely replace CNE-based analyses, it is a powerful addition to the current methodology, especially in the analysis of closely related species or compact genomes.

5.2 Regulatory turnover within GRBs

In Chapter 3, I used kurtosis-based GRB identification to define GRBs for multiple species comparisons in three distinct metazoan lineages. Next I identified subsets of GRBs that exhibit either deep or shallow conservation within each lineage, defining these as low- and high-turnover GRBs, respectively. I showed that ht- and ltGRBs regulate different functional classes of developmental genes, with ltGRBs generally targeting developmental transcription factors and htGRBs targeting genes involved in cell-cell communication and neural development, such as cell adhesion molecules and axon guidance genes. Further, very similar genes were enriched in the respective sets when comparing the results from each lineage. These results suggest that not only are similar classes of genes regulated by GRBs in each metazoan lineage, but also that the similar classes of genes are subject to similar rates of sequence turnover within their regulatory regions in each lineage. This is strong support for the turnover model originally proposed by Harmston et al. (Harmston, Baresic, and Lenhard 2013). The turnover model states that GRB-like gene regulation likely evolved in the metazoan ancestor, and that CNEs have been gradually accumulating sequence changes, since their initial recruitment, albeit extremely slowly. Under this model, the lack of sequence conservation between lineages at typical GRB target genes would be explained by complete turnover of the CNEs in these regions, making it impossible to identify GRBs by sequence conservation alone. Identification of GRBs at similar functional classes of genes, which exhibit similar conservation patterns within each lineage provides strong evidence for this model, as this is a far more parsimonious explanation than independent evolution of GRB-based regulation at these loci in each lineage.

In the remainder of the chapter I characterised human ht- and ltGRBs with respect to the timing of their expression in development, their chromatin state in late neural development, and their repeat content. htGRBs tend to be active later in development than ltGRBs, are enriched for histone modifications associated with active enhancers in the fetal brain, and are more likely to contain active SINEs than ltGRBs. Taken together, these results show that htGRBs are more likely to be active during late neural development than ltGRBs. The timing of expression during development may explain the differences in the conservation patterns between ht- and ltGRBs. Changes in the expression pattern of genes expressed early in development will have far greater pleiotropic effects than those expressed later in development, simply due to differences in the number of tissues and cell types derived from the cells expressing those genes. Therefore it is plausible that ltGRB target gene expression is under stronger selection than that of htGRB target genes.

This analysis has a few important limitations that should be noted. Firstly, this is by no means an exhaustive identification of ht- and ltGRBs. The two sets were identified by belonging to either the bottom or top 20% of GRBs, ranked by kurtosis-based conservation summed across all species comparisons. As a result there are likely many GRBs excluded from this analysis that show similar conservation patterns to those included. Further, the division of GRBs into two classes is in itself artificial, as the rate of turnover within GRBs is likely continuous. However, the emphasis here was on identifying the features of GRBs that may explain these differing rates of turnover, and therefore selecting extreme examples from two tails of a distribution facilitated a strong comparison. A second shortcoming of this analysis is the use of imperfect target gene prediction in the analysis of the timing of GRB target gene expression during development. GRB target gene prediction is currently only available for human GRBs, and the predictions remain largely experimentally unverified. Target gene predictions were used for this analysis because when attempting to identify enrichment for different expression dynamics through development in each group of GRBs, the signal from ubiquitously expressed bystander genes overpowered the signal from GRB targets. The lack of target gene prediction in other species, combined with difficulty in assigning orthologs over large evolutionary distances (~650 million years), also prevented replication of the results observed for human GRB targets for either chicken or fruitfly GRB target genes. It would be of great interest to repeat this analysis after applying GRB target gene prediction to both other species.

Overall, this analysis successfully identified ht- and ltGRBs in three independent species, and strengthened support for the GRB turnover model. Further, it provides a hypothesis for the origin of differing rates of turnover within GRBs, namely that decreased pleiotropy of htGRB target genes results in decreased selection against regulatory changes, and a more permissive environment for sequence turnover and enhancer recruitment.

5.3 GRBs in compact genomes

In Chapter 4, the final analysis chapter in this thesis, I identified GRBs in the extremely compact *Caenorhabditis elegans* and *Oikopleura dioica* genomes. I showed that these GRBs are enriched for similar histone modifications, and regulate similar genes to GRBs identified in larger genomes, thereby confirming their functional equivalence. This is the first time GRBs have been identified in both genomes, and the identification of GRBs in species from two more metazoan lineages (nematodes and tunicates respectively) provides further support for the ancient origin of GRB-based gene regulation. The identification of such compact GRBs in these genomes may guide future Hi-C analysis in these species. Previous attempts at TAD prediction in *C. elegans* only identified TAD-like structures on the X chromosome, with very little detectable 3D structure identified on the autosomes (Crane et al. 2015). Given the general concordance between GRBs and TADs, it may be that deeper sequencing is required in *C. elegans* to provide the resolution required to identify such fine-scale structures.

To assess the effects of genome size on GRB size and content, I identified kurtosis-based GRBs in multiple metazoan genomes, using multiple combinations of GRB identification parameters. I showed that there is a strong linear relationship between genome and GRB size, suggesting that GRB size scales proportionally with genome size. This analysis confirms preliminary observations by Harmston et al. (Harmston et al. 2017), but extends the analysis of this trend to extremely compact genomes. It has been suggested that the spacing between adjacent CNEs is highly conserved between species, however this observation is most apparent when comparing genomes that are of a similar size (Sun, Skogerbø, and Chen 2006). When comparing the distance between orthologous human and tetraodon CNEs, only CNEs that were already relatively closely spaced exhibited conserved spacing between the two species (Sun, Skogerbø, and Chen 2006). This result suggests that there may be a minimum space between CNEs for them to function correctly. In this analysis I found no evidence for a minimum GRB size. This could be due to no such requirement existing, or due to not analysing GRB size in small enough genomes.

In Chapter 4, I also showed that using the most stringent GRB identification parameters in all species resulted in GRB predictions that covered a similar proportion of each genome. It is tempting to speculate that this is due to the presence of a core set of developmental transcription factors in all metazoan genomes that are extremely tightly regulated, resulting in the presence of highly conserved GRBs at these genes in all species. This hypothesis is supported by the identification of a number of gene families that evolved in the metazoan ancestor, and remain in all metazoan genomes to this day (Paps and Holland 2018), many of which perform similar functions to the GRB target genes.

Finally, I showed that GRB composition (with respect to exonic, intronic, intergenic and repeat content) is significantly different from the background genome composition. Further, I showed that while a species' GRB composition is more similar to its genome composition than to other species' GRB composition, the differences between GRBs and their genome of origin are due to enrichment and depletion of similar genomic features. This suggests that GRBs experience similar selective pressure to maintain a permissive environment for long-range gene regulation. In general GRBs are depleted of exonic and repeat sequence and enriched in intergenic and intronic sequence. The depletion in exonic sequence is due to the general gene sparsity within GRBs (Kikuta et al. 2007a; Engström et al. 2007; Akalin et al. 2009; Harmston et al. 2017), while the depletion of repeats is likely due to the ability of repeat elements to acquire enhancer function, and the potential for repeat insertion to disrupt existing enhancers. The enrichment of intergenic and intronic sequence is due to the presence of CNEs within the introns and intergenic space in GRBs, making them larger than average, especially in compact genomes.

5.4 Future directions

In this thesis I have identified GRBs in numerous metazoan species, many for the first time. This has certainly cemented the assertion that GRB-like gene regulation is a pervasive feature of metazoan gene regulation, however there are still many outstanding questions regarding the origin and evolution of GRBs.

The observation that GRBs and TADs coincide in both vertebrates and invertebrates (Harmston et al. 2017) shows that TADs act as functional units of both 3D genome organisation and long-range gene regulation. However, it is not known whether TADs evolved to insulate GRBs, or GRBs expanded to the boundaries of TADs after their evolution. Recently it has been suggested that long-range gene regulation by distal *cis*-regulatory elements is a metazoan innovation that contributed to the transition from a uni- to a multicellular lifestyle (Sebé-Pedrós et al. 2016). This is derived from two key observations. First, the cnidarian species *Nematostella vectensis* has distal *cis*-regulatory elements that are present in similar genomic features to bilatarian genomes (Schwaiger et al. 2014), and second, the unicellular eukaryote with the largest known gene repertoire for transcriptional regulation, *Capsaspora owczarzaki*, lacks any evidence of distal regulation (Sebé-Pedrós et al. 2016). Interestingly, CTCF, the architectural protein that stabilises TADs, does not have an identifiable ortholog outside of bilaterian species (Acemel, Maeso, and Gómez-Skarmeta 2017), implying that long-range gene regulation preceded TAD formation, or that TADs evolved before CTCF. It is possible that prior to the evolution of CTCF, and the formation of strongly insulated TADs, that long-range gene regulation was mediated by cohesin and other general transcription factors - similar to enhancer-promoter looping, or subTADs in vertebrates (Phillips-Cremins et al. 2013). Given then, that CNEs have been identified in very basic species at the root of metazoa, including poriferan and cnidarian species (Ryu, Seridi, and Ravasi 2012), it is tempting to speculate that GRBs preceded TADs, and that TADs evolved to provide better insulation between neighbouring GRBs. Physical separation of the target genes of neighbouring GRBs would provide a more permissive environment for regulatory innovation, as it would reduce the potential for ectopic interaction of newly evolved enhancers with target genes in other GRBs, and also reduce the potential impact of regulatory innovation by limiting the interaction of newly recruited enhancers to the promoters of a single developmental gene. It is also possible that GRBs preceded TADs, but that the GRBs in these genomes were very small relative to the genome size. Evolution of TADs may have facilitated enhancer-promoter interactions over larger genomic distances, thereby enabling the expansion of GRBs to the boundaries of TADs.

To test these hypotheses I would identify CNEs and GRBs in the sea anenome, *Nematostella vectensis*, by comparison to other sequenced cnidarian genomes, including the stony coral, *Stylophora pistillata* (Voolstra et al. 2017), the complex coral, *Acropora digitifera* (Shinzato et al. 2011) and the fresh-water polyp *Hydra magnipapillata* (Chapman et al. 2010). Depending on the distance of the species comparison, this could be performed using a combination of traditional CNE- and kurtosis-based approaches. Next I would analyse the size, composition and chromatin state of the identified GRBs, and determine whether cnidarian GRBs appear to be functionally equivalent to GRBs in more complex organisms. *Nematostella vectensis* is an ideal candidate for this analysis as there is publicly available chromatin modification data for this species, and previously it has been successfully cultivated in the lab (Sebé-Pedrós et al. 2018). This is essential, as the final data required for this analysis would be *Nematostella vectensis* Hi-C data. Hi-C data would be essential to determine whether the lack of CTCF in cnidaria is coupled with a lack of TADs, or perhaps whether TAD formation in cnidaria is mediated by different complexes than TADs in bilateria. The Hi-C data would also facilitate the comparison of GRBs and the 3D structure of the *Nematostella* genome, thereby addressing the questions on the origin of GRBs and TADs. This analysis would provide great insights into the origin and evolution of both GRBs and TADs in primitive metazoan species and provide a model for the evolution of long-range gene regulation in metazoa.

References

Abnizova, I et al. (2007). “Statistical Information Characterization of Conserved Non-Coding Elements in Vertebrates”. In: *Journal of Bioninformatics and Computational Biology* 5.2B, pp. 533–547.

Acemel, Rafael D., Ignacio Maeso, and José Luis Gómez-Skarmeta (2017). “Topologically associated domains: a successful scaffold for the evolution of gene regulation in animals”. In: *Wiley Interdisciplinary Reviews: Developmental Biology* 6.3, pp. 1–19.

Ahituv, Nadav et al. (2007). “Deletion of ultraconserved elements yields viable mice”. In: *PLoS Biology* 5.9, pp. 1906–1911.

Akalin, Altuna et al. (2009). “Transcriptional features of genomic regulatory blocks”. In: *Genome Biology* 10.4, R38.

Akalin, Altuna et al. (2015). “Genomation: A toolkit to summarize, annotate and visualize genomic intervals”. In: *Bioinformatics* 31.7, pp. 1127–1129.

Altschul, Stephen F. et al. (1990). “Basic Local Alignment Search Tool”. In: *Journal of Molecular Biology* 215.3, pp. 403–410.

Amores, Angel et al. (2011). “Genome evolution and meiotic maps by massively parallel DNA sequencing: Spotted gar, an outgroup for the teleost genome duplication”. In: *Genetics* 188.4, pp. 799–808.

Andersson, Robin et al. (2014). “An atlas of active enhancers across human cell types and tissues”. In: *Nature* 507.7493, pp. 455–461.

Arnold, Cosmas D et al. (2013). “Genome-Wide Quantitative Enhancer Activity Maps Identified by STARR-seq”. In: *Science* 339.6123, pp. 1074–1077.

Arnosti, David N. and Meghana M. Kulkarni (2005). “Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards?” In: *Journal of Cellular Biochemistry* 94.5, pp. 890–898.

Asthana, Saurabh et al. (2007). “Widely distributed noncoding purifying selection in the human genome.” In: *Proceedings of the National Academy of Sciences* 104.30, pp. 12410–12415.

Babarinde, Isaac Adeyemi and Naruya Saitou (2013). “Heterogeneous tempo and mode of conserved noncoding sequence evolution among four mammalian orders”. In: *Genome Biology and Evolution* 5.12, pp. 2330–2343.

Babarinde, Isaac Adeyemi and Naruya Saitou (2016). “Genomic Locations of Conserved Noncoding Sequences and Their Proximal Protein-Coding Genes in Mammalian Expression Dynamics”. In: *Molecular biology and evolution* 33.7, pp. 1807–1817.

Balanda, Kevin P and H L Macgillivray (1988). “Kurtosis : A Critical Review”. In: *The American Statistician* 42.2, pp. 111–119.

Barski, Artem et al. (2007). “High-Resolution Profiling of Histone Methylations in the Human Genome”. In: *Cell* 129.4, pp. 823–837.

Barthel, Robert et al. (2003). “Regulation of tumor necrosis factor alpha gene expression by mycobacteria involves the assembly of a unique enhanceosome dependent on the coactivator proteins CBP/p300.” In: *Molecular and cellular biology* 23.2, pp. 526–533.

Bejerano, Gill et al. (2004). “Ultraconserved elements in the human genome”. In: *Science* 304.5675, pp. 1321–1325.

Bejerano, Gill et al. (2006). “A distal enhancer and an ultraconserved exon are derived from a novel retroposon”. In: *Nature* 441.1, pp. 87–90.

Berlivet, Soizik et al. (2013). “Clustering of Tissue-Specific Sub-TADs Accompanies the Regulation of HoxA Genes in Developing Limbs”. In: *PLoS Genetics* 9.12, e1004018.

Berna, Luisa and Fernando Alvarez-Valin (2014). “Evolutionary genomics of fast evolving tunicates”. In: *Genome Biology and Evolution* 6.7, pp. 1724–1738.

Bernstein, Bradley E. et al. (2005). “Genomic maps and comparative analysis of histone modifications in human and mouse”. In: *Cell* 120.2, pp. 169–181.

Bernstein, Bradley E. et al. (2006). “A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells”. In: *Cell* 125.2, pp. 315–326.

Bhatia, Shipra et al. (2014). “A survey of ancient conserved non-coding elements in the PAX6 locus reveals a landscape of interdigitated cis-regulatory archipelagos”. In: *Developmental Biology* 387.2, pp. 214–228.

Bird, Christine P. et al. (2007). “Fast-evolving noncoding sequences in the human genome”. In: *Genome Biology* 8.6, R118.

Birnbaum, Ramon Y. et al. (2012). “Coding exons function as tissue-specific enhancers of nearby genes”. In: *Genome Research* 22.6, pp. 1059–1068.

Birney, Ewan et al. (2007). “Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project”. In: *Nature* 447.7146, pp. 799–816.

Blanchette, Mathieu et al. (2004). “Aligning multiple genomic sequences with the threaded blockset aligner”. In: *Genome Research* 14.4, pp. 708–715.

Bonn, Stefan et al. (2012). “Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development”. In: *Nature Genetics* 44.2, pp. 148–156.

Bossen, Claudia et al. (2015). “The chromatin remodeler Brg1 activates enhancer repertoires to establish B cell identity and modulate cell growth”. In: *Nature Immunology* 16.7, pp. 775–784.

Boyd, J. Lomax et al. (2015). “Human-chimpanzee differences in a FZD8 enhancer alter cell-cycle dynamics in the developing neocortex”. In: *Current Biology* 25.6, pp. 772–779.

Bradley, Robert K. et al. (2010). “Binding site turnover produces pervasive quantitative changes in transcription factor binding between closely related drosophila species”. In: *PLoS Biology* 8.3, e1000343.

Carnell, Ammie N. and Jay I. Goodman (2003). “The long (LINEs) and the short (SINEs) of it: Altered methylation as a precursor to toxicity”. In: *Toxicological Sciences* 75.2, pp. 229–235.

Carninci, Piero et al. (2006). “Genome-wide analysis of mammalian promoter architecture and evolution”. In: *Nature Genetics* 38.6, pp. 626–635.

Carroll, Sean B. (2005). “Evolution at two levels: On genes and form”. In: *PLoS Biology* 3.7, pp. 1159–1166.

Carroll, Sean B. (2008). “Evo-Devo and an Expanding Evolutionary Synthesis: A Genetic Theory of Morphological Evolution”. In: *Cell* 134.1, pp. 25–36.

Chapman, Jarrod A. et al. (2010). “The dynamic genome of Hydra”. In: *Nature* 464.7288, pp. 592–596.

Cheng, Jen Hao et al. (2015). “Genome-wide analysis of enhancer RNA in gene regulation across 12 mouse tissues”. In: *Scientific Reports* 5, pp. 1–9.

Chiang, Charleston W K et al. (2008). “Ultraconserved elements: Analyses of dosage sensitivity, motifs and boundaries”. In: *Genetics* 180.4, pp. 2277–2293.

Chuong, Edward B., Nels C. Elde, and Cédric Feschotte (2017). “Regulatory activities of transposable elements: From conflicts to benefits”. In: *Nature Reviews Genetics* 18.2, pp. 71–86.

Clapier, Cedric R. et al. (2017). “Mechanisms of action and regulation of ATP-dependent chromatin-remodelling complexes”. In: *Nature Reviews Molecular Cell Biology* 18.7, pp. 407–422.

Clarke, Shoa L. et al. (2012). “Human Developmental Enhancers Conserved between Deuterostomes and Protostomes”. In: *PLoS Genetics* 8.8.

Cooper, Gregory M. et al. (2005). “Distribution and intensity of constraint in mammalian genomic sequence”. In: *Genome Research* 15.7, pp. 901–913.

Core, Leighton J. et al. (2014). “Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers”. In: *Nature Genetics* 46.12, pp. 1311–1320.

Crane, Emily et al. (2015). “Condensin-Driven Remodeling of X-Chromosome Topology during Dosage Compensation”. In: *Nature* 523.7559, pp. 240–244.

Creyghton, M. P. et al. (2010). “Histone H3K27ac separates active from poised enhancers and predicts developmental state”. In: *Proceedings of the National Academy of Sciences* 107.50, pp. 21931–21936.

Danks, Gemma et al. (2013). “OikoBase: A genomics and developmental transcriptomics resource for the urochordate Oikopleura dioica”. In: *Nucleic Acids Research* 41.D1, pp. 845–853.

Davies, Kalina T.J., Georgia Tsagkogeorga, and Stephen J. Rossiter (2014). “Divergent evolutionary rates in vertebrate and mammalian specific conserved non-coding elements (CNEs) in echolocating mammals”. In: *BMC Evolutionary Biology* 14.1, p. 261.

De Silva, Dilrini R., Richard Nichols, and Greg Elgar (2014). “Purifying selection in deeply conserved human enhancers is more consistent than in coding sequences”. In: *PLoS ONE* 9.7, pp. 1–10.

DeCarlo, Lawrence T. (1997). “On the Meaning and Use of Kurtosis”. In: *Psychological Methods* 2.3, pp. 292–307.

Degner, Jacob F. et al. (2012). “DNase-I sensitivity QTLs are a major determinant of human expression variation”. In: *Nature* 482.7385, pp. 390–394.

Deininger, Prescott L. et al. (2003). “Mobile elements and mammalian genome evolution”. In: *Current Opinion in Genetics and Development* 13.6, pp. 651–658.

Denoeud, France et al. (2010). “Plasticity of animal genome architecture unmasked by rapid evolution of a pelagic tunicate.” In: *Science* 330.6009, pp. 1381–1385.

Derti, Adnan et al. (2006). “Mammalian ultraconserved elements are strongly depleted among segmental duplications and copy number variants”. In: *Nature Genetics* 38.10, pp. 1216–1220.

Dickel, Diane E. et al. (2018). “Ultraconserved Enhancers Are Required for Normal Development”. In: *Cell* 172.3, 491–499.e15.

Dimitrieva, Slavica and Philipp Bucher (2013). “UCNEbase - A database of ultraconserved non-coding elements and genomic regulatory blocks”. In: *Nucleic Acids Research* 41.Database issue, pp. D101–9.

Dixon, Jesse R., David U. Gorkin, and Bing Ren (2016). “Chromatin Domains: The Unit of Chromosome Organization”. In: *Molecular Cell* 62.5, pp. 668–680.

Dixon, Jesse R. et al. (2012). “Topological domains in mammalian genomes identified by analysis of chromatin interactions”. In: *Nature* 485.7398, pp. 376–380.

Dixon, Jesse R. et al. (2015). “Chromatin architecture reorganization during stem cell differentiation”. In: *Nature* 518.7539, pp. 331–336.

Doglio, Laura et al. (2013). “Parallel Evolution of Chordate Cis-Regulatory Code for Development”. In: *PLoS Genetics* 9.11, e1003904.

Dong, Xianjun, David Fredman, and Boris Lenhard (2009). “Synorth: exploring the evolution of synteny and long-range regulatory interactions in vertebrate genomes”. In: *Genome Biology* 10.8, R86.

Dong, Xianjun et al. (2009). “Exonic remnants of whole-genome duplication reveal cis-regulatory function of coding exons”. In: *Nucleic Acids Research* 38.4, pp. 1071–1085.

Dong, Xinran et al. (2016). “Genome-Wide Identification of Regulatory Sequences Undergoing Accelerated Evolution in the Human Genome”. In: *Molecular Biology and Evolution* 33.10, pp. 2565–2575.

Dowen, Jill M. et al. (2014). “Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes”. In: *Cell* 159.2, pp. 374–387.

Dubchak, Inna et al. (2000). “Active conservation of noncoding sequences revealed by three-way species comparisons”. In: *Genome Research* 10.9, pp. 1304–1306.

Eichenlaub, Michael P. and Laurence Ettwiller (2011). “De novo genesis of enhancers in vertebrates”. In: *PLoS Biology* 9.11.

Emera, Deena et al. (2016). “Origin and evolution of developmental enhancers in the mammalian neocortex”. In: *Proceedings of the National Academy of Sciences* 113.19, E2617–E2626.

Engström, Pär G. et al. (2007). “Genomic regulatory blocks underlie extensive microsynteny conservation in insects”. In: *Genome Research* 17.12, pp. 1898–1908.

Erceg, Jelena et al. (2014). “Subtle Changes in Motif Positioning Cause Tissue-Specific Effects on Robustness of an Enhancer’s Activity”. In: *PLoS Genetics* 10.1.

Ernst, Jason and Manolis Kellis (2012). “ChromHMM: Automating chromatin-state discovery and characterization”. In: *Nature Methods* 9.3, pp. 215–216.

Falcon, S. and R. Gentleman (2007). “Using GOstats to test gene lists for GO term association”. In: *Bioinformatics* 23.2, pp. 257–258.

Feschotte, Cédric (2008). “Transposable elements and the evolution of regulatory networks”. In: *Nature Reviews Genetics* 9.5, pp. 397–405.

Feschotte, Cédric and Ellen J. Pritham (2007). “DNA Transposons and the Evolution of Eukaryotic Genomes”. In: *Annual Review of Genetics* 41.1, pp. 331–368.

Fierst, Janna L. et al. (2015). “Reproductive Mode and the Evolution of Genome Size and Structure in Caenorhabditis Nematodes”. In: *PLoS Genetics* 11.6, pp. 1–25.

Fisher, Shannon et al. (2006). “Conservation of RET regulatory function from human to zebrafish without sequence similarity”. In: *Science* 312.5771, pp. 276–279.

Forrest, Alistair R.R. et al. (2014). “A promoter-level mammalian expression atlas”. In: *Nature* 507.7493, pp. 462–470.

Gerstein, M B et al. (2010). “Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project”. In: *Science* 330.6012, pp. 1775–1787.

Girard, Lisa and Michael Freeling (1999). “Regulatory changes as a consequence of transposon insertion”. In: *Developmental Genetics* 25.4, pp. 291–296.

Gittelman, Rachel M. et al. (2015). “Comprehensive identification and analysis of human accelerated regulatory DNA”. In: *Genome Research* 25.9, pp. 1245–1255.

Glazov, Evgeny A. et al. (2005). “Ultraconserved elements in insect genomes: A highly conserved intronic sequence implicated in the control of homothorax mRNA splicing”. In: *Genome Research* 15.6, pp. 800–808.

Gómez-Marín, Carlos et al. (2015). “Evolutionary comparison reveals that diverging CTCF sites are signatures of ancestral topological associating domains borders”. In: *Proceedings of the National Academy of Sciences* 112.24, pp. 7542–7547.

Goode, Debbie K. et al. (2005). “Highly conserved regulatory elements around the SHH gene may contribute to the maintenance of conserved synteny across human chromosome 7q36.3”. In: *Genomics* 86.2, pp. 172–181.

Guo, Ya et al. (2015). “CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function”. In: *Cell* 162.4, pp. 900–910.

Haberle, Vanja and Alexander Stark (2018). “Eukaryotic core promoters and the functional basis of transcription initiation”. In: *Nature Reviews Molecular Cell Biology*.

Haberle, Vanja et al. (2014). “Two independent transcription initiation codes overlap on vertebrate core promoters”. In: *Nature* 507.7492, pp. 381–385.

Haberle, Vanja et al. (2015). “CAGEr: Precise TSS data retrieval and high-resolution promoterome mining for integrative analyses”. In: *Nucleic Acids Research* 43.8, e51.

Hare, Emily E. et al. (2008). “Sepsid even-skipped enhancers are functionally conserved in Drosophila despite lack of sequence conservation”. In: *PLoS Genetics* 4.6, e1000106.

Harmston, N., A. Baresic, and B. Lenhard (2013). “The mystery of extreme non-coding conservation”. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 368.1632, pp. 20130021–20130021.

Harmston, Nathan et al. (2017). “Topologically associating domains are ancient features that coincide with Metazoan clusters of extreme noncoding conservation”. In: *Nature Communications* 8.1, p. 441.

Harris, R. S. (2007). “Improved pairwise alignment of genomic DNA”. PhD thesis. The Pennsylvania State University.

He, Qiye et al. (2011). “High conservation of transcription factor binding and evidence for combinatorial regulation across six Drosophila species”. In: *Nature Genetics* 43.5, pp. 414–421.

Heintzman, Nathaniel D. et al. (2007). “Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome”. In: *Nature Genetics* 39.3, pp. 311–318.

Hiller, Michael, Bruce T. Schaar, and Gill Bejerano (2012). “Hundreds of conserved non-coding genomic regions are independently lost in mammals”. In: *Nucleic Acids Research* 40.22, pp. 11463–11476.

Hoegg, Simone et al. (2004). “Phylogenetic timing of the fish-specific genome duplication correlates with the diversification of teleost fish”. In: *Journal of Molecular Evolution* 59.2, pp. 190–203.

Hoffman, Michael M et al. (2012). “Unsupervised pattern discovery in human chromatin structure through genomic segmentation”. In: *Nature methods* 9.5, pp. 473–476.

Hraba-Renevey, Suzanne and Michel Kress (1988). “Expression of a mouse replacement histone H3.3 gene with a highly conserved 3’ noncoding region during SV40- and polyoma-induced Go to S-phase transition”. In: *Nucleic acids research* 17.7, pp. 2449–2461.

Hsieh, C.-L. et al. (2014). “Enhancer RNAs participate in androgen receptor-driven looping that selectively enhances gene activation”. In: *Proceedings of the National Academy of Sciences* 111.20, pp. 7319–7324.

Hubisz, Melissa J. and Katherine S. Pollard (2014). “Exploring the genesis and functions of Human Accelerated Regions sheds light on their role in human evolution”. In: *Current Opinion in Genetics and Development* 29, pp. 15–21.

Indjeian, Vahan B. et al. (2016). “Evolving New Skeletal Traits by cis-Regulatory Changes in Bone Morphogenetic Proteins”. In: *Cell* 164.1-2, pp. 45–56.

Infante, Carlos R. et al. (2015). “Shared Enhancer Activity in the Limbs and Phallus and Functional Divergence of a Limb-Genital cis-Regulatory Element in Snakes”. In: *Developmental Cell* 35.1, pp. 107–119.

Irvine, Steven Q. (2013). “Study of Cis-regulatory Elements in the Ascidian Ciona intestinalis”. In: *Current Genomics* 14.1, pp. 56–67.

Iwafuchi-Doi, Makiko et al. (2016). “The Pioneer Transcription Factor FoxA Maintains an Accessible Nucleosome Configuration at Enhancers for Tissue-Specific Gene Activation”. In: *Molecular Cell* 62.1, pp. 79–91.

Jacob, François and Jacques Monod (1961). “Genetic regulatory mechanisms in the synthesis of proteins”. In: *Journal of Molecular Biology* 3.3, pp. 318–356.

Jiang, Peng and Mona Singh (2014). “CCAT: Combinatorial Code Analysis Tool for transcriptional regulation”. In: *Nucleic Acids Research* 42.5, pp. 2833–2847.

Johnson, K. D. et al. (2003). “Highly Restricted Localization of RNA Polymerase II within a Locus Control Region of a Tissue-Specific Chromatin Domain”. In: *Molecular and Cellular Biology* 23.18, pp. 6484–6493.

Johnson, Rory et al. (2006). “Identification of the REST regulon reveals extensive transposable element-mediated binding site duplication”. In: *Nucleic Acids Research* 34.14, pp. 3862–3877.

Junion, Guillaume et al. (2012). “A transcription factor collective defines cardiac cell fate and reflects lineage history”. In: *Cell* 148.3, pp. 473–486.

Kajimoto, Y and P Rotwein (1991). “Structure of the chicken insulin-like growth factor I gene reveals conserved promoter elements.” In: *The Journal of biological chemistry* 266.15, pp. 9724–9731.

Kamal, M., X. Xie, and E. S. Lander (2006). “A large family of ancient repeat elements in the human genome is under strong selection”. In: *Proceedings of the National Academy of Sciences* 103.8, pp. 2740–2745.

Kasowski, Maya et al. (2010). “Variation in transcription factor binding among humans”. In: *Science* 328.5975, pp. 232–235.

Kent, W. J. et al. (2002). “The Human Genome Browser at UCSC”. In: *Genome Research* 12.6, pp. 996–1006.

Kent, W James (2002). “BLAT — The BLAST -Like Alignment Tool”. In: *Genome Research* 12, pp. 656–664.

Kiełbasa, Szymon M. et al. (2011). “Adaptive seeds tame genomic sequence comparison”. In: *Genome Research* 21.3, pp. 487–493.

Kikuta, Hiroshi et al. (2007a). “Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates”. In: *Genome Research* 17.5, pp. 545–555.

Kikuta, Hiroshi et al. (2007b). “Retroviral enhancer detection insertions in zebrafish combined with comparative genomics reveal genomic regulatory blocks - A fundamental feature of vertebrate genomes”. In: *Genome Biology* 8.Suppl. 1, S4.

Kim, Tae Hoon et al. (2005). “A high-resolution map of active promoters in the human genome”. In: *Nature* 436.7052, pp. 876–880.

Kim, Tae Kyung et al. (2010). “Widespread transcription at neuronal activity-regulated enhancers”. In: *Nature* 465.7295, pp. 182–187.

Kimble, Judith and David Hirsh (1979). “The postembryonic cell lineages of the hermaphrodite and male gonads in Caenorhabditis elegans”. In: *Developmental Biology* 70.2, pp. 396–417.

Kimura-Yoshida, C. (2004). “Characterization of the pufferfish Otx2 cis-regulators reveals evolutionarily conserved genetic mechanisms for vertebrate head specification”. In: *Development* 131.1, pp. 57–71.

King, Nicole et al. (2008). “The genome of the choanoflagellate Monosiga brevicollis and the origin of metazoans”. In: *Nature* 451.7180, pp. 783–788.

Kulkarni, Meghana M. and David N. Arnosti (2003). “Information display by transcriptional enhancers”. In: *Development* 130.26, pp. 6569–6575.

Kunarso, Galih et al. (2010). “Transposable elements have rewired the core regulatory network of human embryonic stem cells”. In: *Nature Genetics* 42.7, pp. 631–634.

Kundaje, Anshul et al. (2015). “Integrative analysis of 111 reference human epigenomes”. In: *Nature* 518.7539, pp. 317–330.

Kvon, Evgeny Z et al. (2016). “Progressive Loss of Function in a Limb Enhancer during Snake Evolution Article Progressive Loss of Function in a Limb Enhancer during Snake Evolution”. In: *Cell* 167, pp. 633–642.

Lai, Fan et al. (2013). “Activating RNAs associate with Mediator to enhance chromatin architecture and transcription”. In: *Nature* 494.7438, pp. 497–501.

Lampe, Xavier et al. (2008). “An ultraconserved Hox-Pbx responsive element resides in the coding sequence of Hoxa2 and is active in rhombomere 4”. In: *Nucleic Acids Research* 36.10, pp. 3214–3225.

Langmead, Ben et al. (2009). “Ultrafast and memory-efficient alignment of short DNA sequences to the human genome”. In: *Genome Biology* 10.3, R25.

Leal, Francisca and Martin J. Cohn (2016). “Loss and Re-emergence of Legs in Snakes by Modular Evolution of Sonic hedgehog and HOXD Enhancers”. In: *Current Biology* 26.21, pp. 2966–2973.

Lee, Alison P. et al. (2011). “Ancient vertebrate conserved noncoding elements have been evolving rapidly in teleost fishes”. In: *Molecular Biology and Evolution* 28.3, pp. 1205–1215.

Lee, Tong Ihn et al. (2006). “Control of Developmental Regulators by Polycomb in Human Embryonic Stem Cells”. In: *Cell* 125.2, pp. 301–313.

Lelli, Katherine M., Matthew Slattery, and Richard S. Mann (2012). “Disentangling the many layers of eukaryotic transcriptional regulation.” In: *Annual Review of Genetics* 43, pp. 43–68.

Lemaire, C, R Heilig, and J L Mandel (1988). “The chicken dystrophin cDNA: striking conservation of the C-terminal coding and 3’ untranslated regions between man and chicken.” In: *The EMBO journal* 7.13, pp. 4157–62.

Lenhard, Boris, Albin Sandelin, and Piero Carninci (2012). “Metazoan promoters: Emerging characteristics and insights into transcriptional regulation”. In: *Nature Reviews Genetics* 13.4, pp. 233–245.

Levy, Samuel, Sridhar Hannenhalli, and Christopher Workman (2001). “Enrichment of regulatory signals in conserved non-coding genomic sequence”. In: *Bioinformatics* 17.10, pp. 871–877.

Li, Wenbo et al. (2013). “Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation”. In: *Nature* 498.7455, pp. 516–520.

Lieberman-Aiden, Erez et al. (2009). “Comprehensive mapping of long-range interactions reveals folding principles of the human genome”. In: *Science* 326.5950, pp. 289–293.

Lin, Yin C. et al. (2012). “Global changes in the nuclear positioning of genes and intra-and interdomain genomic interactions that orchestrate B cell fate”. In: *Nature Immunology* 13.12, pp. 1196–1204.

Loots, Gabriela G. et al. (2002). “rVista for comparative sequence-based discovery of functional transcription factor binding sites”. In: *Genome Research* 12.5, pp. 832–839.

Lowe, C. B., G. Bejerano, and D. Haussler (2007). “Thousands of human mobile element fragments undergo strong purifying selection near developmental genes”. In: *Proceedings of the National Academy of Sciences* 104.19, pp. 8005–8010.

Lowe, Craig B. et al. (2011). “Three periods of regulatory innovation during vertebrate evolution”. In: *Science* 333.6045, pp. 1019–1024.

Lu, Xu et al. (2008). “The effect of H3K79 dimethylation and H4K20 trimethylation on nucleosome and chromatin structure”. In: *Nature Structural and Molecular Biology* 15.10, pp. 1122–1124.

Manning, Gerard et al. (2008). “The protist, Monosiga brevicollis, has a tyrosine kinase signaling network more elaborate and diverse than found in any known metazoan.” In: *Proceedings of the National Academy of Sciences of the United States of America* 105.28, pp. 9674–9679.

Marcovitz, Amir, Robin Jia, and Gill Bejerano (2016). “"Reverse Genomics" Predicts Function of Human Conserved Noncoding Elements”. In: *Molecular Biology and Evolution* 33.5, pp. 1358–1369.

Margueron, Raphael, Patrick Trojer, and Danny Reinberg (2005). “The key to development: Interpreting the histone code?” In: *Current Opinion in Genetics and Development* 15.2, pp. 163–176.

McCole, Ruth B. et al. (2014). “Abnormal Dosage of Ultraconserved Elements Is Highly Disfavored in Healthy Cells but Not Cancer Cells”. In: *PLoS Genetics* 10.10, e1004646.

McEwen, Gayle K. et al. (2006). “Ancient duplicated conserved noncoding elements in vertebrates: A genomic and functional analysis”. In: *Genome Research* 16.4, pp. 451–465.

McLean, Cory Y. et al. (2011). “Human-specific loss of regulatory DNA and the evolution of human-specific traits”. In: *Nature* 471.7337, pp. 216–219.

McManus, C. Joel et al. (2010). “Regulatory divergence in Drosophila revealed by mRNA-seq”. In: *Genome Research* 20.6, pp. 816–825.

Meader, Stephen, Chris P. Ponting, and Gerton Lunter (2010). “Massive turnover of functional sequence in human and other mammalian genomes”. In: *Genome Research* 20.10, pp. 1335–1343.

Menoret, Delphine et al. (2013). “Genome-wide analyses of Shavenbaby target genes reveals distinct features of enhancer organization”. In: *Genome Biology* 14.8, R86.

Mikkelsen, Tarjei S. et al. (2007). “Genome-wide maps of chromatin state in pluripotent and lineage-committed cells”. In: *Nature* 448.7153, pp. 553–560.

Mousavi, Kambiz et al. (2013). “ERNAs Promote Transcription by Establishing Chromatin Accessibility at Defined Genomic Loci”. In: *Molecular Cell* 51.5, pp. 606–617.

Nagy, László G. et al. (2014). “Latent homology and convergent regulatory evolution underlies the repeated emergence of yeasts”. In: *Nature Communications* 5.

Narendra, Varun et al. (2015). “CTCF establishes discrete functional chromatin domains at the Hox clusters during differentiation”. In: *Science* 347.6225, pp. 1017–1021.

Navratilova, Pavla et al. (2009). “Systematic human/zebrafish comparative identification of cis-regulatory activity around vertebrate developmental transcription factor genes”. In: *Developmental Biology* 327.2, pp. 526–540.

Navratilova, Pavla et al. (2017). “Sex-specific chromatin landscapes in an ultra-compact chordate genome”. In: *Epigenetics and Chromatin* 10.1, pp. 1–18.

Nichols, S. A. et al. (2012). “Origin of metazoan cadherin diversity and the antiquity of the classical cadherin/β-catenin complex”. In: *Proceedings of the National Academy of Sciences* 109.32, pp. 13046–13051.

Nora, Elphège P. et al. (2012). “Spatial partitioning of the regulatory landscape of the X-inactivation centre”. In: *Nature* 485.7398, pp. 381–385.

Osterwalder, Marco et al. (2018). “Enhancer redundancy provides phenotypic robustness in mammalian development”. In: *Nature* 554.7691, pp. 239–243.

Panne, Daniel (2008). “The enhanceosome”. In: *Current Opinion in Structural Biology* 18.2, pp. 236–242.

Panne, Daniel, Tom Maniatis, and Stephen C. Harrison (2004). “Crystal structure of ATF-2/c-Jun and IRF-3 bound to the interferon-β enhancer”. In: *EMBO Journal* 23.22, pp. 4384–4393.

Panne, Daniel, Tom Maniatis, and Stephen C. Harrison (2007). “An Atomic Model of the Interferon-β Enhanceosome”. In: *Cell* 129.6, pp. 1111–1123.

Paps, Jordi and Peter W.H. Holland (2018). “Reconstruction of the ancestral metazoan genome reveals an increase in genomic novelty”. In: *Nature Communications* 9.1, pp. 1–8.

Paris, Mathilde et al. (2013). “Extensive Divergence of Transcription Factor Binding in Drosophila Embryos with Highly Conserved Gene Expression”. In: *PLoS Genetics* 9.9, e100374.

Parker, Hugo J. et al. (2014). “A reporter assay in lamprey embryos reveals both functional conservation and elaboration of vertebrate enhancers”. In: *PLoS ONE* 9.1, e85492.

Patwardhan, Rupali P. et al. (2012). “Massively parallel functional dissection of mammalian enhancers in vivo”. In: *Nature Biotechnology* 30.3, pp. 265–270.

Pearson, Karl P (1905). “"Das Fehlergesetz und Seine Verallgemeinerungen Durch Fechner und Pearson." A Rejoinder”. In: *Biometrika* 4.1, pp. 169–212.

Pennacchio, Len A. et al. (2006). “In vivo enhancer analysis of human conserved non-coding sequences”. In: *Nature* 444.7118, pp. 499–502.

Perry, Michael W. et al. (2010). “Shadow enhancers foster robustness of drosophila gastrulation”. In: *Current Biology* 20.17, pp. 1562–1567.

Phillips-Cremins, Jennifer E. et al. (2013). “Architectural protein subclasses shape 3D organization of genomes during lineage commitment”. In: *Cell* 153.6, pp. 1281–1295.

Plessy, Charles et al. (2005). “Enhancer sequence conservation between vertebrates is favoured in developmental regulator genes”. In: *Trends in Genetics* 21.4, pp. 207–210.

Pokholok, Dmitry K. et al. (2005). “Genome-wide map of nucleosome acetylation and methylation in yeast”. In: *Cell* 122.4, pp. 517–527.

Policarpi, Cristina et al. (2017). “Enhancer SINEs Link Pol III to Pol II Transcription in Neurons”. In: *Cell Reports* 21.10, pp. 2879–2894.

Pollard, Katherine S. et al. (2006). “Forces shaping the fastest evolving regions in the human genome”. In: *PLoS Genetics* 2.10, pp. 1599–1611.

Polychronopoulos, Dimitris et al. (2017). “Conserved non-coding elements: Developmental gene regulation meets genome organization”. In: *Nucleic Acids Research* 45.22, pp. 12611–12624.

Pope, Benjamin D. et al. (2014). “Topologically associating domains are stable units of replication-timing regulation”. In: *Nature* 515.7527, pp. 402–405.

Prabhakar, Shyam et al. (2008). “Human-specific gain of function in a developmental enhancer”. In: *Science* 321.5894, pp. 1346–1350.

Rada-Iglesias, Alvaro et al. (2011). “A unique chromatin signature uncovers early developmental enhancers in humans”. In: *Nature* 470.7333, pp. 279–285.

Rao, Suhas S.P. et al. (2014). “A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping”. In: *Cell* 159.7, pp. 1665–1680.

Rebeiz, M. et al. (2011). “Evolutionary origin of a novel gene expression pattern through co-option of the latent activities of existing regulatory sequences”. In: *Proceedings of the National Academy of Sciences* 108.25, pp. 10036–10043.

Reddy, Timothy E. et al. (2012). “Effects of sequence variation on differential allelic transcription factor occupancy and gene expression”. In: *Genome Research* 22.5, pp. 860–869.

Rhee, Ina et al. (2002). “DNMT1 and DNMT3b cooperate to silence genes in human cancer cells”. In: *Nature* 416.6880, pp. 552–556.

Ritter, Deborah I. et al. (2012). “Transcriptional enhancers in protein-coding exons of vertebrate developmental genes”. In: *PLoS ONE* 7.5, e35202.

Rivera, Chloe M. and Bing Ren (2013). “Mapping human epigenomes”. In: *Cell* 155.1, pp. 39–55.

Ross, Gj (2015). “Parametric and nonparametric sequential change detection in R: The cpm package”. In: *Journal of Statistical Software* 66.3, pp. 1–20.

Rouault, J. P. et al. (1993). “Sequence analysis reveals that the BTG1 anti-proliferative gene is conserved throughout evolution in its coding and 3’ non-coding regions”. In: *Gene* 129.2, pp. 303–306.

Roy, Sushmita et al. (2010). “Identification of functional elements and regulatory circuits by Drosophila modENCODE”. In: *Science* 330.6012, pp. 1787–1797.

Ruppert, David (1987). “What is kurtosis? An influence function approach”. In: *American Statistician* 41.1, pp. 1–5.

Ryba, Tyrone et al. (2010). “Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types”. In: *Genome Research* 20.6, pp. 761–770.

Ryu, Taewoo, Loqmane Seridi, and Timothy Ravasi (2012). “The evolution of ultraconserved elements with different phylogenetic origins”. In: *BMC Evolutionary Biology* 12.1, p. 236.

Sagai, Tomoko et al. (2004). “Phylogenetic conservation of a limb-specific, cis-acting regulator of Sonic hedgehog (Shh)”. In: *Mammalian Genome* 15.1, pp. 23–34.

Sandelin, Albin et al. (2004). “Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes”. In: *BMC Genomics* 5.1, p. 99.

Sanges, Remo et al. (2013). “Highly conserved elements discovered in vertebrates are present in non-syntenic loci of tunicates, act as enhancers and can be transcribed during development”. In: *Nucleic Acids Research* 41.6, pp. 3600–3618.

Sasaki, T. et al. (2008). “Possible involvement of SINEs in mammalian-specific brain formation”. In: *Proceedings of the National Academy of Sciences* 105.11, pp. 4220–4225.

Scally, Aylwyn et al. (2012). “Insights into hominid evolution from the gorilla genome sequence”. In: *Nature* 483.7388, pp. 169–175.

Schmidt, Dominic et al. (2010). “Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding.” In: *Science* 328.5981, pp. 1036–40.

Schmitt, Anthony D. et al. (2016). “A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome”. In: *Cell Reports* 17.8, pp. 2042–2059.

Schwaiger, Michaela et al. (2014). “Evolutionary conservation of the eumetazoam gene regulatory landscape - Supplemental Figures”. In: *Genome Research* 24, pp. 639–650.

Schwartz, Scott et al. (2003). “Human-mouse alignments with BLASTZ.” In: *Genome research* 13.1, pp. 103–107.

Sebé-Pedrós, Arnau et al. (2010). “Ancient origin of the integrin-mediated adhesion and signaling machinery”. In: *Proceedings of the National Academy of Sciences* 107, pp. 10142–10147.

Sebé-Pedrós, Arnau et al. (2011). “Unexpected repertoire of metazoan transcription factors in the unicellular holozoan capsaspora owczarzaki”. In: *Molecular Biology and Evolution* 28.3, pp. 1241–1254.

Sebé-Pedrós, Arnau et al. (2012). “Premetazoan Origin of the Hippo Signaling Pathway”. In: *Cell Reports* 1.1, pp. 13–20.

Sebé-Pedrós, Arnau et al. (2013a). “Early evolution of the T-box transcription factor family”. In: *Proceedings of the National Academy of Sciences* 110.40, pp. 16050–16055.

Sebé-Pedrós, Arnau et al. (2013b). “Regulated aggregative multicellularity in a close unicellular relative of metazoa”. In: *eLife* 2013.2, pp. 1–20.

Sebé-Pedrós, Arnau et al. (2016). “The Dynamic Regulatory Genome of Capsaspora and the Origin of Animal Multicellularity”. In: *Cell* 165.5, pp. 1224–1237.

Sebé-Pedrós, Arnau et al. (2018). “Cnidarian Cell Type Diversity and Regulation Revealed by Whole-Organism Single-Cell RNA-Seq”. In: *Cell* 173.6, 1520–1534.e20.

Sexton, Tom et al. (2012). “Three-dimensional folding and functional organization principles of the Drosophila genome”. In: *Cell* 148.3, pp. 458–472.

Shibata, Yoichiro et al. (2012). “Extensive evolutionary changes in regulatory element activity during human origins are associated with altered gene expression and positive selection”. In: *PLoS Genetics* 8.6, e1002789.

Shinzato, Chuya et al. (2011). “Using the Acropora digitifera genome to understand coral responses to environmental change”. In: *Nature* 476.7360, pp. 320–323.

Shiraki, T. et al. (2003). “Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage”. In: *Proceedings of the National Academy of Sciences* 100.26, pp. 15776–15781.

Shogren-Knaak, Michael et al. (2006). “Histone H4-K16 acetylationcontrols chromatin structure and protein interactions”. In: *Science* 311.18, pp. 844–847.

Siepel, Adam et al. (2005). “Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes”. In: *Genome Research* 15.8, pp. 1034–1050.

Singer, Tatjana et al. (2010). “LINE-1 retrotransposons: Mediators of somatic variation in neuronal genomes?” In: *Trends in Neurosciences* 33.8, pp. 345–354.

Small, S, D N Arnosti, and M Levine (1993). “Spacing ensures autonomous expression of different stripe enhancers in the even-skipped promoter.” In: *Development* 119.3, pp. 762–772.

Smit, A.F.A., R. Hubley, and P. Green. *RepeatMasker Open-4.0.* `http://www.repeatmasker.org`. Accessed: 2018-02-22.

Smith, Robin P. et al. (2013). “Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model”. In: *Nature Genetics* 45.9, pp. 1021–1028.

Spieler, Derek et al. (2014). “Restless Legs Syndrome-Associated intronic common variant in Meis1 alters enhancer function in the developing telencephalon”. In: *Genome Research* 24.4, pp. 592–603.

Stark, Alexander et al. (2007). “Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures”. In: *Nature* 450.7167, pp. 219–232.

Stephen, Stuart et al. (2008). “Large-scale appearance of ultraconserved elements in tetrapod genomes and slowdown of the molecular clock”. In: *Molecular Biology and Evolution* 25.2, pp. 402–408.

Stone, Jonathon R. and Gregory A. Wray (2001). “Rapid evolution of cis-regulatory sequences via local point mutations”. In: *Molecular Biology and Evolution* 18.9, pp. 1764–1770.

Su, Ming et al. (2014). “Evolution of Alu Elements toward Enhancers”. In: *Cell Reports* 7.2, pp. 376–385.

Suga, Hiroshi et al. (2012). “Genomic survey of premetazoans shows deep conservation of cytoplasmic tyrosine kinases and multiple radiations of receptor tyrosine kinases”. In: *Science Signaling* 5.222, pp. 26–28.

Suga, Hiroshi et al. (2013). “The Capsaspora genome reveals a complex unicellular prehistory of animals”. In: *Nature Communications* 4, pp. 1–9.

Suga, Hiroshi et al. (2014). “Earliest holozoan expansion of phosphotyrosine signaling”. In: *Molecular Biology and Evolution* 31.3, pp. 517–528.

Suganuma, Tamaki and Jerry L. Workman (2011). “Signals and Combinatorial Functions of Histone Modifications”. In: *Annual Review of Biochemistry* 80.1, pp. 473–499.

Sun, Hong, Geir Skogerbø, and Runsheng Chen (2006). “Conserved distances between vertebrate highly conserved elements”. In: *Human Molecular Genetics* 15.19, pp. 2911–2922.

Sur, Inderpreet and Jussi Taipale (2016). “Accelerated evolution of conserved noncoding sequences in humans”. In: *Nat Reviews Cancer* 314.5800, p. 786.

Tan, G. (2018). "Computational Genomics of Regulatory Elements and Regulatory Territories". PhD thesis.

Tan, Ge (2015). *CNEr: CNE Detection and Visualization.*

Thanos, Dimitris and Tom Maniatis (1995). "Virus induction of human IFNβ gene expression requires the assembly of an enhanceosome". In: *Cell* 83.7, pp. 1091–1100.

The C. elegans Sequencing Consortium (1998). "Genome Sequence of the Nematode C. elegans : A Platform for Investigating Biology". In: *Science* 282.5396, pp. 2012–2018.

Thornburg, Bartley G., Valer Gotea, and Wojciech Makałowski (2006). "Transposable elements as a significant source of transcription regulating signals". In: *Gene.* Vol. 365, pp. 104–110.

Vavouri, Tanya et al. (2007). "Parallel evolution of conserved non-coding elements that target a common set of developmental regulatory genes from worms to humans". In: *Genome Biology* 8.2, R15.

Vettese-Dadey, M et al. (1996). "Acetylation of histone H4 plays a primary role in enhancing transcription factor binding to nucleosomal DNA in vitro." In: *The EMBO journal* 15.10, pp. 2508–2518.

Villar, Diego, Paul Flicek, and Duncan T. Odom (2014). "Evolution of transcription factor binding in metazoans-mechanisms and functional implications". In: *Nature Reviews Genetics* 15.4, pp. 221–233.

Viturawong, Tar et al. (2013). "A DNA-Centric Protein Interaction Map of Ultraconserved Elements Reveals Contribution of Transcription Factor Binding Hubs to Conservation". In: *Cell Reports* 5.2, pp. 531–545.

Voolstra, Christian R. et al. (2017). "Comparative analysis of the genomes of Stylophora pistillata and Acropora digitifera provides evidence for extensive differences between species of corals". In: *Scientific Reports* 7.1, p. 17583.

Walter, Klaudia et al. (2005). “Striking nucleotide frequency pattern at the borders of highly conserved vertebrate non-coding sequences”. In: *Trends in Genetics* 21.8, pp. 436–440.

Wang, Dong et al. (2011). “Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA”. In: *Nature* 474.7351, pp. 390–397.

Wang, Jianli et al. (2009). “Large number of ultraconserved elements were already present in the jawed vertebrate ancestor”. In: *Molecular Biology and Evolution* 26.3, pp. 487–490.

Wang, Ting et al. (2007). “Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53.” In: *Proceedings of the National Academy of Sciences* 104.47, pp. 18613–8.

Wang, Yuan Liang et al. (2014). “TRF2, but not TBP, mediates the transcription of ribosomal protein genes”. In: *Genes and Development* 28.14, pp. 1550–1555.

Warnefors, Maria et al. (2016). “Combinatorial Gene Regulatory Functions Underlie Ultraconserved Elements in Drosophila”. In: *Molecular Biology and Evolution* 33.9, pp. 2294–2306.

White, J (1988). “The anatomy”. In: *The Nematode Caenorhabditis elegans*, pp. 81–122.

Wicker, Thomas et al. (2005). “The repetitive landscape of the chicken genome”. In: *Genome Research* 15.1, pp. 126–136.

Wilson, Michael D et al. (2008). “Species-specific transcription in mice carrying human chromosome 21”. In: *Science* 322.5900, pp. 434–438.

Wittkopp, Patricia J., Belinda K. Haerum, and Andrew G. Clark (2008). “Regulatory changes underlying expression differences within and between Drosophila species”. In: *Nature Genetics* 40.3, pp. 346–350.

Wittkopp, Patricia J. and Gizem Kalay (2012). “Cis-regulatory elements: Molecular mechanisms and evolutionary processes underlying divergence”. In: *Nature Reviews Genetics* 13.1, pp. 59–69.

Woolfe, Adam et al. (2005). “Highly conserved non-coding sequences are associated with vertebrate development”. In: *PLoS Biology* 3.1, e7.

Wray, Gregory A. (2007). “The evolutionary significance of cis-regulatory mutations”. In: *Nature Reviews Genetics* 8.3, pp. 206–216.

Yaffe, D et al. (1985). “Highly conserved sequences in the 3’ untranslated region of mRNAs coding for homologous proteins in distantly related species.” In: *Nucleic acids research* 13.10, pp. 3723–37.

Yan, Liying et al. (2016). “Epigenomic Landscape of Human Fetal Brain, Heart, and Liver”. In: *Journal of Biological Chemistry* 291.9, pp. 4386–4398.

Young, Robert S. et al. (2017). “Bidirectional transcription initiation marks accessible chromatin and is not specific to enhancers”. In: *Genome Biology* 18.1, pp. 1–11.

Yu, Miao and Bing Ren (2017). “The Three-Dimensional Organization of Mammalian Genomes”. In: *Annual Review of Cell and Developmental Biology* 33.1, pp. 265–289.

Zabidi, Muhammad A. et al. (2015). “Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation”. In: *Nature* 518.7540, pp. 556–559.

Zerbino, Daniel R. et al. (2018). “Ensembl 2018”. In: *Nucleic Acids Research* 46.D1, pp. D754–D761.

Zheng, Wei et al. (2011). “Regulatory Variation Within and Between Species”. In: *Annual Review of Genomics and Human Genetics* 12.1, pp. 327–346.

Appendix A

Supplementary Figures

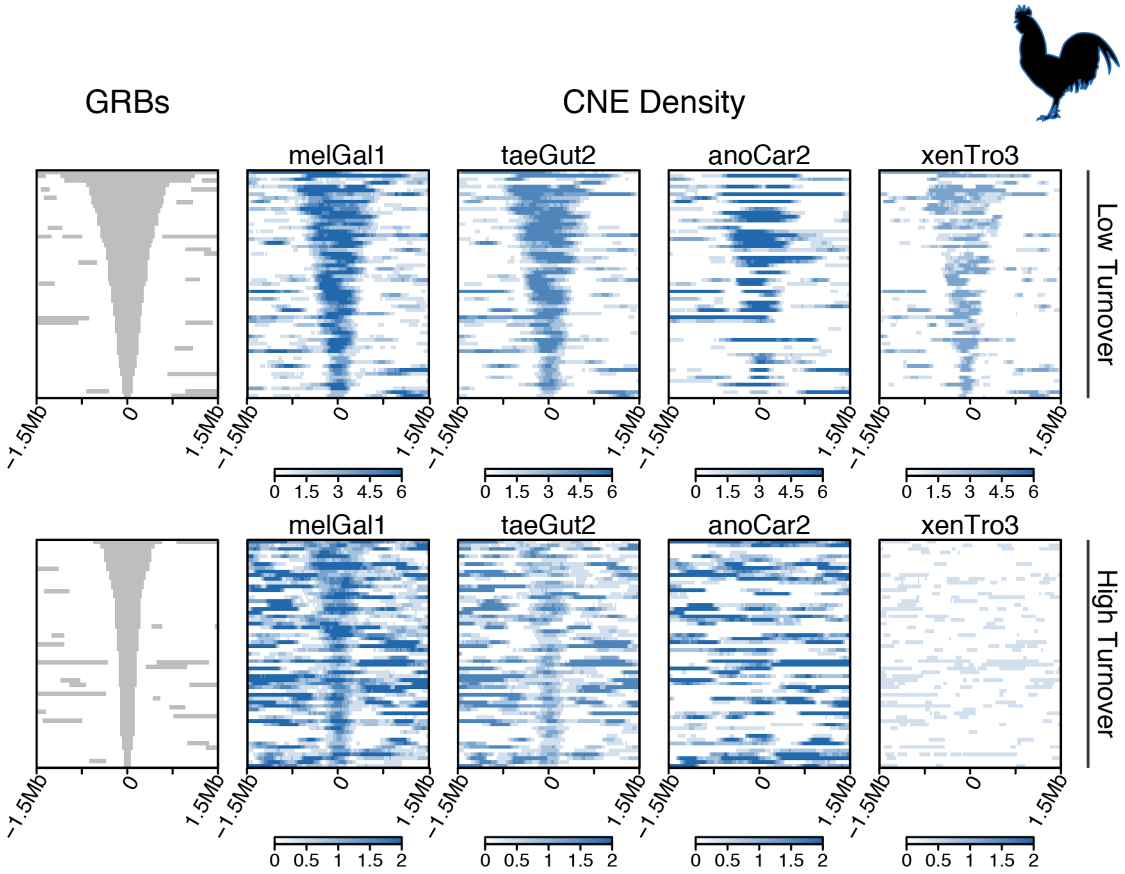

Figure S1: CNE Density in high- and low-turnover galGal4 GRBs. CNE density for each of the species comparisons from galGal4 was plotted in high- and low-turnover GRBs. The grey funnels represent the GRB bounds, while the coloured plots show the CNE density in the same genomic windows. In the ltGRBs, there is visible enrichment of CNEs within the GRBs for all species comparisons, while in the htGRBs, this enrichment is only visible in the most closely related species comparisons and appears to decrease with increasing distance of the comparison.

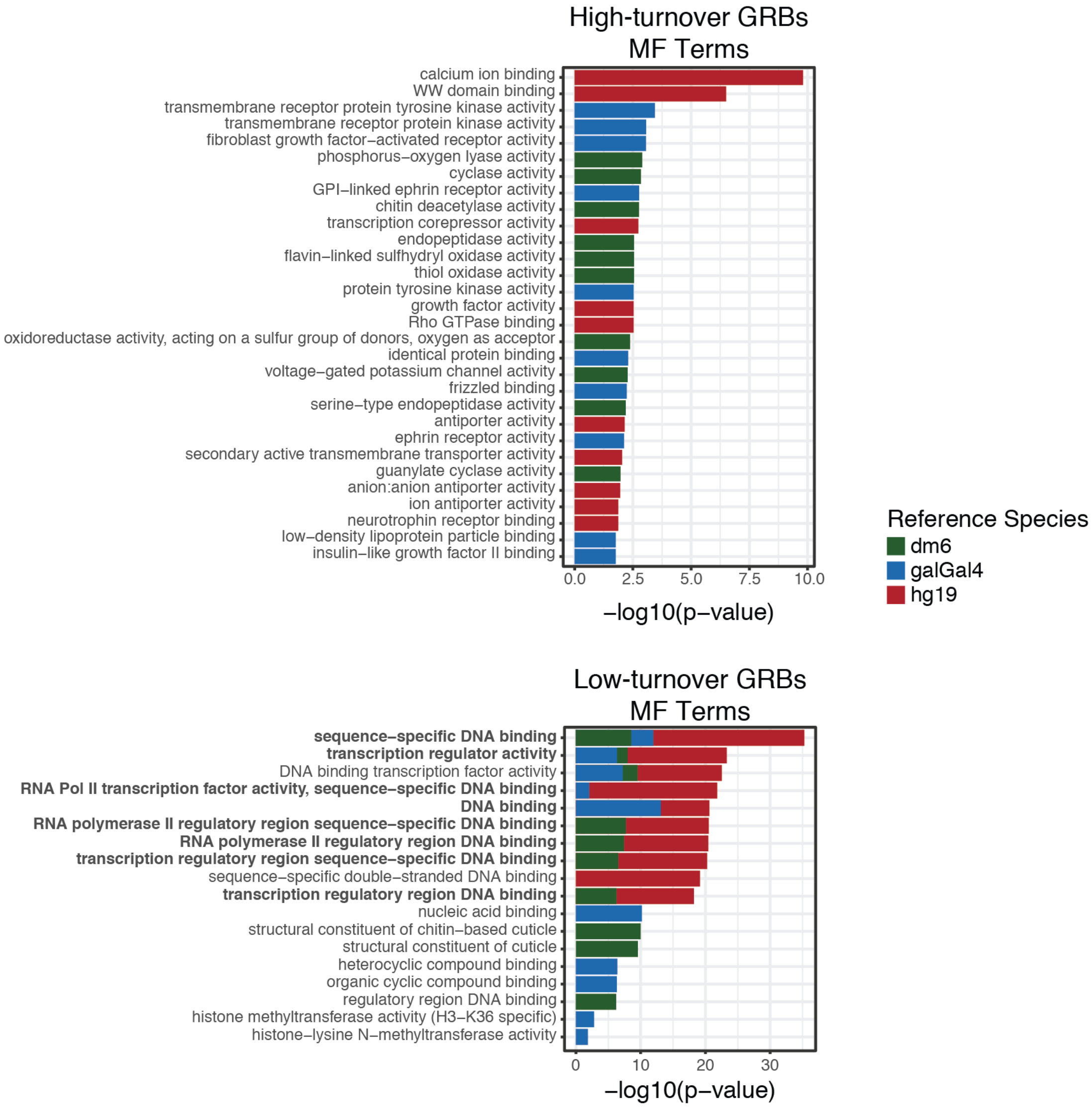

Figure S2: High and Low-turnover GRB target gene MF GO enrichment. Gene ontology (GO) enrichment analysis was performed for the target genes of ht- and ltGRBs. The -log10 p-values for the top 10 enriched MF terms from each reference species are shown here. Terms which occur in more than one reference species are highlighted in bold, and the bars for each reference species' p-value are overlaid. MF GO terms indicate that ht- and ltGRBs are enriched for distinct subsets of developmental genes, and that the enrichment is consistent across the three reference species used.

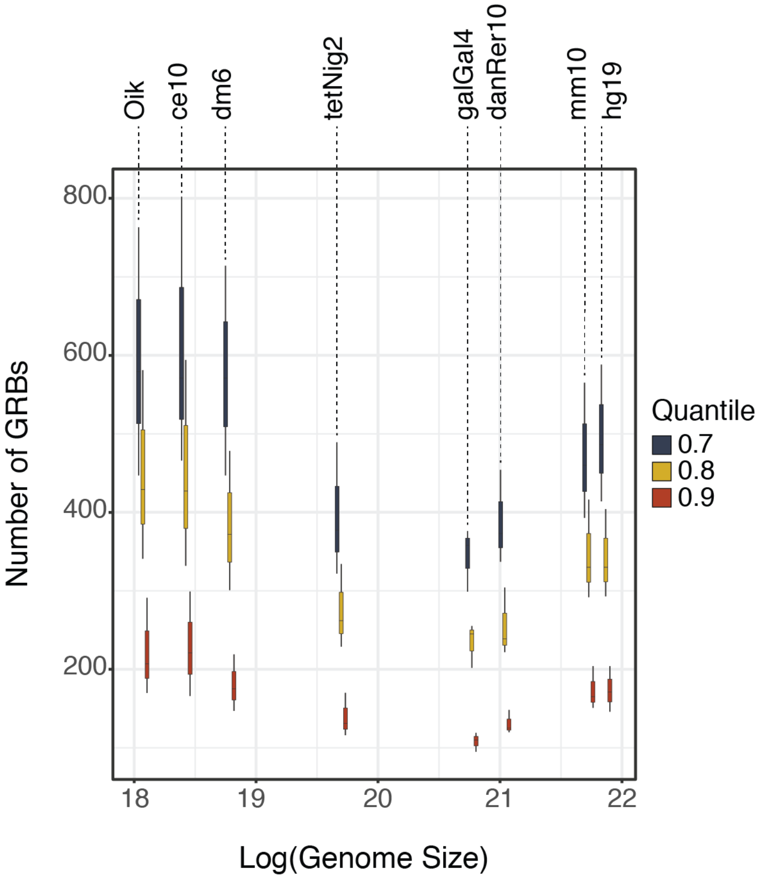

Figure S3: GRB number vs genome size. The number of GRBs predicted from kurtosis-based conservation calculated in three bin sizes was plotted against the log genome size from which they were derived. GRB sets were stratified by the merging quantile used in GRB prediction.

Appendix B

Supplementary Tables

Table S1: Properties of kurtosis-based and CNE density-based GRBs predicted in Chapter 2. The number of GRBs predicted, using either kurtosis- or CNE density-based GRB identification, and the proportion of the genome they cover, are shown for GRBs predicted from human to each of the species listed in the table. Also shown is the width of the intersection between the sets identified using either method, as a proportion of the total width of the GRBs identified using either particular method.

	Dog	**Opossum**	**Chicken**	**Spotted gar**
Number of GRBs (Kurtosis)	559	487	426	400
Number of GRBs (CNE density)	1426	1026	848	403
Proportion of genome covered (Kurtosis)	0.22	0.19	0.13	0.10
Proportion of genome covered (CNE density)	0.32	0.31	0.31	0.14
Intersection width/Total width (Kurtosis)	0.84	0.92	0.95	0.60
Intersection width/Total width (CNE density)	0.58	0.55	0.40	0.43

Table S2: GRB sets identified in ce10 and OikDioicaNorway

Reference species	Conservation measure	Bin size (bp)	Merging quantile	Number of GRBs	Mean width (kb)
ce10	CNE density	-	0.7	634	33.9
ce10	CNE density	-	0.8	518	29.4
ce10	CNE density	-	0.9	334	22.8
ce10	Kurtosis-based conservation	700	0.7	813	33.7
ce10	Kurtosis-based conservation	700	0.8	603	25.5
ce10	Kurtosis-based conservation	700	0.9	301	16.9
ce10	Kurtosis-based conservation	1000	0.7	571	43.6
ce10	Kurtosis-based conservation	1000	0.8	427	30.8
ce10	Kurtosis-based conservation	1000	0.9	221	23
ce10	Kurtosis-based conservation	1300	0.7	467	51.6
ce10	Kurtosis-based conservation	1300	0.8	334	35.1
ce10	Kurtosis-based conservation	1300	0.9	167	24.1
OikDioicaNorway	CNE density	-	0.7	554	22.3
OikDioicaNorway	CNE density	-	0.8	433	18.6
OikDioicaNorway	CNE density	-	0.9	229	16.4
OikDioicaNorway	Kurtosis-based conservation	500	0.7	822	21.3
OikDioicaNorway	Kurtosis-based conservation	500	0.8	634	15.6
OikDioicaNorway	Kurtosis-based conservation	500	0.9	343	10.7
OikDioicaNorway	Kurtosis-based conservation	700	0.7	604	26.7
OikDioicaNorway	Kurtosis-based conservation	700	0.8	455	18.5
OikDioicaNorway	Kurtosis-based conservation	700	0.9	232	12.9
OikDioicaNorway	Kurtosis-based conservation	900	0.7	458	31.9
OikDioicaNorway	Kurtosis-based conservation	900	0.8	358	21.2
OikDioicaNorway	Kurtosis-based conservation	900	0.9	186	14.7

Table S3: Number and mean width of all GRB sets predicted in Chapter 4

Reference species	Bin size (bp)	Merging quantile	Number of GRBs	Mean width (kb)
hg19	20; 30; 40	0.7	588; 486; 414	991.4; 1124.2; 1288.9
hg19	20; 30; 40	0.8	404; 330; 293	851.2; 1003.5; 1108.5
hg19	20; 30; 40	0.9	204; 171; 146	715.9; 873.5; 1006.3
mm10	18; 27; 36	0.7	565; 460; 393	918.9; 1104.5; 1312.9
mm10	18; 27; 36	0.8	416; 330; 292	710.6; 849.9; 1023.5
mm10	18; 27; 36	0.9	204; 165; 151	577.3; 728.2; 832.8
danRer10	9; 14; 18	0.7	454; 373; 337	384.8; 464.2; 530.7
danRer10	9; 14; 18	0.8	304; 239; 222	325.9; 394.5; 437.4
danRer10	9; 14; 18	0.9	148; 125; 120	315.1; 390.4; 422.7
galGal4	7; 10; 14	0.7	376; 358; 299	440.3; 489.3; 552.8
galGal4	7; 10; 14	0.8	255; 245; 202	369.7; 425.5; 482.4
galGal4	7; 10; 14	0.9	119; 110; 95	331.4; 383.4; 446.5
tetNig2	2.4; 3.6; 4.8	0.7	489; 377; 322	100.7; 126.1; 147.7
tetNig2	2.4; 3.6; 4.8	0.8	334; 262; 229	83.1; 103.2; 114
tetNig2	2.4; 3.6; 4.8	0.9	170; 131; 116	79.5; 95.2; 110.7
dm6	1; 1.4; 1.9	0.7	714; 571; 447	52; 66.2; 82.5
dm6	1; 1.4; 1.9	0.8	478; 372; 301	50.3; 66.9; 81.9
dm6	1; 1.4; 1.9	0.9	219; 175; 147	50.6; 64.5; 75.9
ce10	0.7; 1; 1.3	0.7	802; 571; 466	34; 43.6; 51.7
ce10	0.7; 1; 1.3	0.8	594; 427; 332	25.8; 30.8; 35.3
ce10	0.7; 1; 1.3	0.9	299; 221; 166	17; 23; 24.1
OikDioicaNorway	0.5; 0.7; 0.9	0.7	763; 579; 477	22.5; 27.5; 32.5
OikDioicaNorway	0.5; 0.7; 0.9	0.8	581; 429; 341	16.5; 19.2; 21.9
OikDioicaNorway	0.5; 0.7; 0.9	0.9	291; 207; 170	11.5; 13.5; 15.2