COLD SPRING HARBOR LABORATORY

WATSON SCHOOL OF BIOLOGICAL SCIENCES

PHD THESIS

Promoter Evolution in *Drosophila*: Non-Coding Transcription & Transposon-Driven Innovation

---

Author:

Philippe J BATUT

Supervisor:

Pr. Thomas R GINGERAS

Contents

Chapter 1

Introduction

1.1 Overview

The diversity of forms and lifestyles among living organisms, and the processes by which it has been generated, have long been a source of fascination and puzzlement. Since Darwin and Wallace, we have come to view this diversity as the result of evolutionary processes, driven by gradual change under the directing forces of natural selection. However, the precise mechanisms underlying evolutionary change have proven elusive. In particular, explaining the diversity of animal forms in terms of the evolution of developmental processes has long been a daunting question for embryologists. With the advent of molecular biology in the 20^{th} century and a deeper comprehension of the genetic mechanisms that specify development, there has been a renewal of interest in these questions, and this novel perspective has led to a better understanding of developmental evolution. A certain number of fundamental principles have emerged regarding animal development and the processes by which it evolves.

Development in metazoans is orchestrated by the execution of complex gene regulatory programs encoded in the sequence of the genome. The expression of thousands of loci, in a precise temporal sequence and in well-defined spatial patterns, establishes and progressively refines the patterning of embryonic structures. Beyond this organizing function, gene expression also specifies the progressive differentiation of cells throughout the embryo into increasingly specialized cell types, ultimately imparting upon them the particular phenotypic traits that underlie their biological functions. This staggeringly complex decision-making process is integrated into large gene regulatory networks. The highly hierarchical structure of these networks reflects the causal relationships between series of developmental events, as well as their temporal order of occurrence. Indeed, the processes of developmental control are best described as cascades of decisions that progressively specify the fate of individual cells, and thus restrict the sets of choices available to them at subsequent stages. Decrypting the structure and function

of developmental gene regulatory networks is therefore crucial to our understanding of development, yet at this point both the nature of the genes involved and the mechanisms that regulate their activity remain insufficiently characterized.

Although our knowledge of the genes that control development has expanded dramatically since the first pioneering insights into developmental genetics, recent studies of transcription in higher eukaryotes have revealed tremendous complexity. The transcriptomes of various organisms have been found to comprise an unexpected diversity of protein-coding and non-coding transcripts, and is it now apparent that transcription pervades much of genomic space. Classical protein-coding loci often generate multiple transcript isoforms through alternative splicing and other processes, and also give rise to overlapping transcripts that do not seem to encode polypeptides. Genomic intervals between protein-coding loci, long thought to consist of largely inert "junk DNA", have been shown to produce large numbers of non-coding transcripts. Although there has been much debate regarding the potential functions and biological relevance of pervasive non-coding transcription, in recent years a large body of work has provided overwhelming evidence that much of it does have biological functions. Whether and how such transcripts participate in the control of development is currently the subject of intense investigation.

Transcriptional regulation is thought to play a crucial role in the developmental control of gene activity in metazoans. Genetic and biochemical studies over the last decades have illuminated the molecular processes through which such regulation can be achieved. Transcription from individual promoters is regulated primarily by the binding of sequence-specific transcription factors at *cis*-regulatory elements such as enhancers or insulators, and functional interactions between multiple factors underlie the complex computations performed by these regulatory elements. Although the characterization of well-defined model systems has yielded many significant insights, our understanding of the mechanisms underlying transcriptional control remains limited. On a more global level, establishing how these mechanisms are integrated across genes into complex coordinated networks is a central problem in systems biology. Large-scale studies of developmental gene expression, along with the functional dissection of large numbers of individual regulatory elements, will be necessary to achieve a global understanding of developmental processes.

In accordance with its critical role in controlling development, changes in transcriptional regulation have long been postulated to be important drivers of developmental evolution. The unique properties of regulatory changes, as opposed to changes in protein-coding sequences for instance, also seem

to make them attractive candidates from a theoretical standpoint. And indeed, detailed studies of the evolution of developmental and morphological phenotypes have often traced the genetic causes back to regulatory mutations. More recently, genome-wide surveys of regulatory sequences have also revealed fast evolutionary dynamics in diverse lineages. Together, these observations suggest that regulatory variation should be of particular relevance to the evolution of development and organismal phenotypes. However, the precise molecular mechanisms underlying regulatory variation, their effects on gene expression, and the types of genes that are most often affected by evolutionarily relevant mutations are currently largely uncharacterized.

Among the mechanisms that participate in the remodeling of regulatory interactions, the activity of transposable elements is of particular interest. Since their discovery, it has been hypothesized that transposons may play a role in gene regulation, and that they could be potent drivers of regulatory change. The ability of transposons to influence gene expression was actually the main reason they were identified in the first place. Their capacity to physically move throughout their host genomes, and thus perhaps distribute identical regulatory elements to multiple loci, made them appear early on as potentially important contributors to regulatory evolution. Although there has been controversy in the decades following their discovery over the extent and adaptive value of transposon domestication, it is now becoming clear that they have played an important role in the evolution of regulatory networks, at least in certain lineages. The extent of this phenomenon in various organisms is still being determined, however, and its relevance to the regulation of developmental gene expression has not been investigated in depth.

The primary goal of the work presented here is to further our understanding of developmental transcriptomes and the processes by which they evolve. In particular, the influence of mutational mechanisms that recycle pre-existing genomic elements, such as the co-option of transposons as regulatory modules, constitutes an important point of interest. We chose to focus on transcriptional promoters, a class of regulatory elements that play a central role in integrating regulatory inputs and determining levels of transcription. The direct measurement of transcription start site (TSS) usage provides a powerful tool to map promoters on a genome-wide scale and to quantify their contribution to transcriptomes. Importantly, such experiments yield an assessment of the impact of promoters on gene expression that is difficult to obtain for other types of regulatory elements. However, techniques allowing these measurements currently have strong limitations. Therefore we developed RAMPAGE (RNA

Annotation and Mapping of Promoters for the Analysis of Gene Expression), a high-fidelity TSS usage profiling technique based on massively parallel sequencing of 5'-complete complementary DNAs. Using this technique, we characterized the landscape of promoter activity throughout the life cycle of *Drosophila melanogaster*. This analysis revealed widespread non-coding transcription, as well as over 1,300 transposon-derived promoters with developmentally regulated expression. It also allowed us to explore the hypothesis that transposons can distribute promoters with stereotyped developmental expression patterns throughout the *D. melanogaster* genome.

In order to investigate patterns of developmental gene expression divergence between species, we also profiled promoter activity throughout embryonic development in 5 *Drosophila* species. This approach allowed us to address questions regarding the evolutionary rates of promoter gain and loss and the pace of quantitative expression divergence. It also provided a unique opportunity to study the selective forces that shape transcriptome evolution, as well as the influence that these forces exert on particular genes or categories of genes. In addition, the comparative analysis of functional data across species is a powerful strategy to identify biologically relevant aspects of genome function through evolutionary conservation. Such strategies are particularly valuable when attempting to assess the conservation of aspects of genome function that cannot at this point be predicted directly from genome sequence, such as transcriptional output. This allowed us to directly tackle the contentious question of non-coding transcription functionality in a developmental setting, by directly measuring the intensity of purifying selection on this type of transcription.

This introduction will be organized into three main sections. Section 1.2 will provide an overview of the current understanding of transcriptome complexity and transcriptional regulation. I will describe how genome-wide surveys of transcription have transformed our understanding of genome function, and feature the main tenets of the controversy regarding the biological importance of non-coding transcription. I will also present the evidence supporting the widespread functionality of this transcription and describe the known molecular functions of non-coding transcripts, drawing from studies in various organisms. The last part of this section will introduce the fundamental aspects of transcriptional regulation in eukaryotes. Section 1.3 will focus on regulatory evolution, and its relevance to developmental and morphological changes in metazoans. With a certain emphasis on examples from *Drosophila*, I will illustrate salient aspects and principles of regulatory evolution. I will also introduce the molecular mechanisms that underlie this phenomenon, and in particular the relevance of transposable

elements. Finally, Section 1.4 will present functional genomics approaches to the study of evolutionary processes. I will provide a brief overview of modern techniques for transcriptome analysis, and describe the current understanding of transcriptome evolution in *Drosophila*.

1.2 Eukaryotic transcriptomes: Complexity & Regulation

Genome-wide surveys of transcription in recent years have revealed a picture far more complex than anticipated. Protein-coding genes generate a tremendous diversity of transcripts through alternative promoter and cleavage/polyadenylation site usage, as well as alternative splicing. In addition, eukaryotic cells produce large numbers of non-coding transcripts, both intergenic and overlapping protein-coding loci. Although the biological relevance of these transcripts has been a matter of intense debate, the genetic and biochemical characterization of a growing number of examples is illuminating the breadth of molecular roles they perform. This work has been gradually leading to the realization that virtually every biological process in higher eukaryotes requires the involvement of some form of non-coding transcription. The expression of these sophisticated transcriptional landscapes is a finely regulated process, modulated by environmental and developmental cues.

The first part of this section will provide an overview of the current understanding of transcriptome complexity, drawing from insights in various systems. I will also briefly describe the controversy surrounding the significance of pervasive transcription and introduce the arguments most central to it. The second part will focus more specifically on non-coding transcription and the current knowledge regarding its prevalence, molecular activities, and biological relevance. A detailed discussion of the known molecular functions of non-coding transcription is aimed at making the case for their diversity and importance, and provides a justification for our interest in this process. In addition, I will introduce recent genome-wide evidence of purifying selection on non-coding RNA (ncRNA) loci, supporting the hypothesis that many ncRNAs perform significant biological roles. Finally, the third part will introduce the fundamental concepts of transcriptional regulation, including the underlying molecular mechanisms, the genomic organization of regulatory elements, and the architecture of gene regulatory networks.

1.2.1 The emerging complexity of transcriptional landscapes

Genome-wide studies of transcription

With the completion of large-scale genome sequencing projects and the advent of new technologies for genome-wide functional studies, our understanding of the extent of transcriptional activity in eukaryotic genomes has changed drastically. The human genome, in particular, has been the focus of intense collaborative efforts geared towards understanding the organization of its functional elements. The Encyclopedia of DNA Elements (ENCODE) project, which started soon after the human genome was sequenced, has been a major contributor to our new understanding of transcriptional landscapes. Early studies using tiling microarray technology revealed the existence of a multitude of new transcripts, which either overlap protein-coding loci or lie in intergenic space (Kapranov et al. (2007)). Despite some controversy regarding the specificity of the techniques used and the biological relevance of the transcripts identified, these observations fundamentally altered our vision of transcriptional activity. The view has emerged that transcription is not simply a limited process that operates on well-separated topological units, but a pervasive phenomenon that affects most of the genome, following intricate, often overlapping patterns at many loci (reviewed in Kapranov et al. (2007)). In addition, such work prompted the realization that many transcripts have little or no protein-coding potential, suggesting that non-coding transcription and non-coding transcripts may play a much larger role in human biology than previously anticipated.

More recently, the use of high-throughput shotgun complementary DNA (cDNA) sequencing, a technique known as RNA-seq, has confirmed the essence of these early conclusions, and provided enhanced maps of human transcriptomes. According to the latest data, it is estimated that up to 75% of mappable sequences in the human genome are transcribed in at least some cell types or physiological conditions (Djebali et al. (2012)). A large portion of the transcribed genome is attributable to introns, and some evidence of their post-transcriptional processing into shorter products suggests that many introns are not merely spacers, but do harbor functional elements. There are classical examples of such processing of introns into stable and functional products, such as small nucleolar RNAs (snoRNAs), which are involved in guiding chemical modifications to other RNAs, and microRNAs (miRNAs), which are involved in port-transcriptional gene regulation. Introns also encode complex arrays of sequences that specify the fate of primary transcripts – most notably, they contain diverse sequence elements involved

in the regulation of splicing. In addition, we are now beginning to truly appreciate the prevalence of non-coding transcription. Thousands of short ncRNAs, including the majority of miRNAs, are generated through the processing of dedicated, independent primary transcripts. Strikingly, the number of known long non-coding RNAs (lncRNAs) in humans has grown in a few years from a few classical examples, such as *Xist* or *H19*, to well over 13,000 (Djebali et al. (2012), Guttman et al. (2009), see also GENCODE v19 annotations). Figure 1.1 provides an illustration of the complexity of transcription patterns and the abundance of non-coding transcripts at the well-characterized human *HoxA* locus. Notably, several of these ncRNAs now have established functions in the regulation of neighboring Hox genes (Wang et al. (2011), Bertani et al. (2011), Maamar et al. (2013), Zhao et al. (2013)). In addition to these well-defined non-coding transcriptional units, there is also ample evidence for seemingly more diffuse transcription patterns producing non-coding transcripts from many canonical Pol II promoters and transcriptional enhancers (Kapranov et al. (2007), Fejes-Toth et al. (2009), Taft et al. (2009), Kim et al. (2010), Schlesinger et al. (2013)). Overall, while we are still in the early days of its functional characterization, it has become obvious that non-coding transcription is a major phenomenon with acute relevance to gene regulation, development and disease. This topic will be discussed in more detail below.

Transcription is not puzzling solely in its breadth: a whole other level of complexity arises from the staggering diversity of transcript isoforms encoded by individual loci. Alternative splicing in particular, but also the use of alternative promoters and alternative cleavage and polyadenylation sites, generate multitudes of products by simply rearranging in various combinations the set of functional elements available at any given locus. The discovery of splicing, and alternative splicing, heralded the promise of a new level of gene expression regulation at the post-transcriptional level. It was not immediately clear, however, whether alternative splicing was an exotic phenomenon restricted to a small number of exceptional genes, or whether on the contrary it would prove to be a general regulatory step across the genome. It seems fair to say that the outcome has probably exceeded anyone's expectations: it is now estimated that at least 95% of multi-exonic transcripts are alternatively spliced (Wang et al. (2008), reviewed in Nilsen and Graveley (2010)). Protein-coding genes in particular make extensive use of this process, which is now firmly established as a major contributor to proteome diversity (Nilsen and Graveley (2010)). Although the case of the human genome is of particular value because of the sheer scale of functional characterization efforts, it is becoming clear that, at least to an extent, the principles enumerated above do apply to other eukaryotes as well. In mammals, thorough surveys of

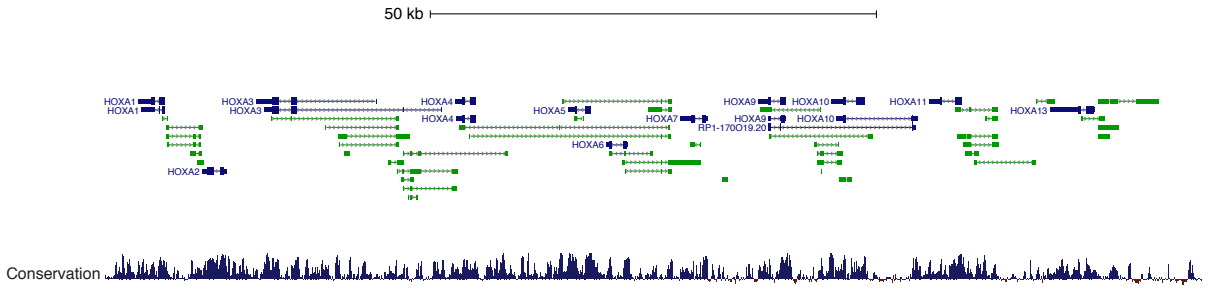

Figure 1.1. Organization of transcription at the human *HoxA* locus.
The human *HoxA* locus encodes the well-known protein-coding *HoxA* genes (blue), as well as a large number of non-protein-coding transcription units (green). Only the highest-quality manually curated annotations generated by GENCODE (version 17) are included in this figure. Notably, several lncRNAs encoded by the locus are known to have important roles in the regulation of neighboring protein-coding genes (see main text). Figure generated by the UCSC Genome Browser.

the mouse transcriptome have revealed a complexity that essentially mirrors that observed in humans: a majority of the mouse genome is transcribed, alternative splicing is omnipresent, and non-coding transcription is extremely prevalent. Among invertebrates, *D. melanogaster* and *C. elegans* have been the most extensively studied on a genome-wide scale. In both cases, the genomes are much smaller and the level of complexity is lower than in mammals, but similar phenomena are nonetheless observed on a scale that is far from negligible. Approximately 75% of the *D. melanogaster* genome can be transcribed, and alternative splicing has been observed at about 60% of multi-exonic genes (Graveley et al. (2010)). A recent survey identified 1,119 putative long intergenic non-coding RNA (lincRNA) loci, suggesting that non-coding transcription may be more prevalent than anticipated in fly (Young et al. (2012)). In *C. elegans*, transcriptional landscapes are not quite as complex, but about 60% of genes display alternative splicing (Gerstein et al. (2010)), and there are at least 170 lincRNA loci identified to date (Nam and Bartel (2012)).

Interestingly, even in organisms that were once thought to have simple and stereotyped transcription patterns, recent insights have started to challenge the accepted wisdom. In the yeast *S. cerevisiae*, a host of efforts in the past few years have identified a number of condition-specific non-coding stable unannotated transcripts (SUTs), at least 1,500 transcripts antisense to protein-coding genes, as well as a plethora of unstable transcripts only detectable after impairment of the RNA degradation machinery: cryptic unstable transcripts (CUTs) and Xrn1-sensitive unstable transcripts (XUTs) (van Dijk et al. (2011), Jacquier (2009)). In addition, novel techniques allowing high-resolution mapping of the boundaries of individual transcripts have revealed extensive isoform diversity at most loci. Even prokaryotes seem to have their fair share of functional non-coding RNAs, such as the CRISPR-derived RNAs involved in the repression of bacteriophages (Marraffini and Sontheimer (2011)). There have also been reports of widespread antisense transcription at protein-coding loci in bacteria, and such sense-antisense transcript pairs may be processed into short RNAs by an RNAseIII enzyme (Lasa et al. (2011)). The function of this antisense transcription is unknown. These observations in such a variety of organisms underscore how general widespread non-coding transcription is, and clearly warrant further exploration of its regulation and of its potential biological functions.

Functional transcription or transcriptional noise?

The wealth of transcriptome data accumulated over the past decade has certainly contributed to a new understanding of genome organization, but it has also stirred significant controversy regarding the reality as well as the biological relevance of this long-hidden transcriptional "dark matter" (van Bakel et al. (2010), Ponting and Belgard (2010), Clark et al. (2011), Kapranov and St Laurent (2012)). Widespread transcription in *S. cerevisiae* should, in some authors' opinion, be dismissed as "transcriptional noise" – merely the result of non-specific initiation by RNA Pol II due to the limited fidelity of the process (Struhl (2007)). This "junk transcription" would simply be an unavoidable side effect of genic transcription, with no purpose and no function, and the resulting transcripts devoid of any biological relevance. Although this is a rather extreme position, it remains to be determined to what extent non-specific transcriptional activity may contribute to the overall transcriptional output of eukaryotic genomes.

In mammals, some experts have also criticized the evidence supporting widespread non-coding transcription. The specificity of low signal detection on tiling arrays has been questioned, igniting debate over the true extent of the phenomenon. The use of high-throughput sequencing technology has largely resolved these issues, but there remains some controversy regarding the detection of very low signals (van Bakel et al. (2010)). It has been suggested that the extent of transcriptional activity may have been significantly overestimated, and that most transcripts branded by others as distinct entities actually constitute extensions of protein-coding transcripts, such as novel exons or extended untranslated regions (UTRs) (van Bakel et al. (2010)). This position, however, has since been sharply challenged (Clark et al. (2011)). As technology progresses, however, these technical issues are being resolved. Consequently, a consensus in the field has crystallized around the idea that the majority of mammalian genomes are transcribed in complex patterns, and that non-coding transcription clearly plays essential roles in some biological processes.

Beyond the question of their mere existence, the question of the biological significance of these troves of uncharacterized and atypical transcripts has now come to the forefront. In contrast to protein-coding genes, for which we can make use of our clear understanding of the genetic code, no rules are known that could guide the prediction of function from the primary sequence of non-coding transcripts. This lack of interpretive rules has also made it difficult to determine whether the loci encoding ncRNAs are under any sort of selective constraint. Although there is now accumulating evidence, from both com-

parative genomics and population genetics, that large numbers of ncRNAs are indeed under purifying selection (Guttman et al. (2009), Derrien et al. (2012), Young et al. (2012), Haerty and Ponting (2013)), the exact breadth and strength of selection remains unclear. This topic will be discussed in more detail in the next subsection. The low abundance of many transcripts is another source of skepticism towards ncRNAs (van Bakel et al. (2010)). Indeed, a significant proportion seems to have steady-state abundances of only a few copies per cell, and many even seem to be present in less than one copy per cell on average. It is being debated whether such rare molecules could possibly make significant contributions to any biological process. There are, however, some examples of this: the human $ncRNA_{CCND1}$, despite a steady-state abundance of 2-4 copies per cell, does play a clear role in the regulation of the *CCND1* gene (Wang et al. (2008)).

Ultimately, only thorough experimental testing of individual cases will provide a definitive answer to the open question of ncRNA functionality. There are, however, a fast-growing number of examples clearly showing that ncRNAs can perform a variety of molecular functions. It is also becoming increasingly clear that they participate in extremely diverse biological processes, and that their involvement in some of those processes displays extremely deep evolutionary conservation. The increasing use of high-throughput reverse genetics and phenotyping approaches is now making it feasible to rigorously ask these questions on a global scale, and the next few years will likely bring exciting new developments.

1.2.2 Non-coding transcription

Functional long non-coding RNAs: Early insights

Functional non-coding RNAs constitute the core machinery of many fundamental biological processes: ribosomal RNAs are the main scaffold and catalytic component of ribosomes, and small nucleolar RNAs are key actors in their biogenesis; transfer RNAs implement the genetic code; small nuclear RNAs constitute some catalytic component of the spliceosome; the TERC component of telomerase has a central role in chromosome biology. In spite of this, recognition of the diversity of non-coding transcripts and their molecular functions has only begun to emerge in recent years.

The mouse H19 transcript was one of the first molecules thought to represent an emerging class of long ncRNAs. It was initially described as a large liver-specific transcript, and it was soon recognized that it had little protein-coding potential and no conserved open reading frame (ORF) (Brannan et al.

(1990)). Despite being localized to the cytoplasm, it also did not seem to template any protein production in mouse cells (Brannan et al. (1990)). It has since been established that the *H19* locus is conserved and imprinted across therian mammals, in which it regulates placental growth (Smits et al. (2008)). It is also known to play a role in tumor suppression. Although H19 is believed to act in part as a precursor for the miR-675 microRNA, patterns of sequence conservation across the locus suggest additional functions for the full primary transcript as well (Keniry et al. (2012)).

The Xist RNA was another early example, initially identified as the product of a gene exclusively expressed from the human inactive X chromosome (Brown et al. (1991)). It has since been established that the Xist transcript is a non-coding RNA essential for X chromosome inactivation (Penny et al. (1996)). Although the details of the process remain the subject of active research, it has been shown that the Xist RNA acts at least in part by directly binding to and recruiting in *cis* the Polycomb repressive complex 2 (PRC2) (Zhao et al. (2008)). Details of Xist function will be elaborated below.

Although these early examples were long perceived as puzzling examples of an atypical biological function of RNA, it has now become clear that non-coding RNA genes are present in the tens of thousands in many eukaryotic genomes. There is mounting evidence that many of these loci have been or are currently under selective pressure in various organisms, suggesting that functionality might the rule rather than the exception. Interestingly, it seems that the mechanism of action of the Xist RNA, in particular, might have provided an early example of a much more general phenomenon. Indeed, there have been numerous reports of physical interactions between non-coding transcripts and transcriptional co-activators or co-repressors, and it has been proposed that such interactions might be at the heart of lncRNA biology.

The role of *Xist* in random X chromosome inactivation

The mammalian Xist offers one of the most famous and best-characterized examples of functional lncRNA to date. As such, it is arguably the only one for which so many functional aspects have been formally tested and established. Therefore it is a useful case to consider, both as a clear illustration of lncRNA biology and as a compelling argument for further focus on non-coding transcription.

In mammals, dosage compensation for the expression of X-linked genes is achieved by random inactivation of one of the two X chromosomes in females (Augui et al. (2011)). The study of X-autosome translocations in the early 1980s lead to the identification of a locus required for X inactivation, which

was termed the X inactivation center (*Xic*) (Rastan (1983)). Further work identified a gene within the human *Xic* with a very unique expression pattern: it appeared to be expressed exclusively from the inactive X chromosome, where the vast majority of genes are tightly repressed (Brown et al. (1991)). That exceptional gene was named X-inactive specific transcripts (*Xist*), and it was proposed that it might play a role in X inactivation. The finding that a focal deletion of the 5' region of *Xist* abolished X inactivation demonstrated this to indeed be the case (Penny et al. (1996)). Bewilderment ensued when it appeared that the ~15 kb Xist transcript had no conserved ORF and localized almost exclusively to the nucleus, which led to the suggestion that the active product of the gene might be a functional non-coding RNA (Brockdorff et al. (1992)).

In support of this hypothesis, it was subsequently shown that the expression of an ectopic Xist cDNA, even from an autosome, is sufficient to initiate silencing in *cis* (Wutz and Jaenisch (2000)). The mechanisms underlying the silencing process are still unclear, but the Xist RNA appears to coat the entire inactive X chromosome (Clemson et al. (1996), Chaumeil et al. (2006)). Recent data shows that, starting from the locus it is expressed from, the RNA spreads across the chromosome by hopping to loci in close proximity in three-dimensional space (Engreitz et al. (2013)). How this spreading is restricted to a single X chromosome is unknown. This coating by Xist seems to create a somewhat distinct nuclear compartment, into which genes get recruited as they are progressively silenced (Chaumeil et al. (2006)). Focused *Xist* mutations showed that different domains of the RNA are required for chromosome coating and silencing activity (Wutz et al. (2002)). Repeat A, a region genetically required for silencing, was shown to physically interact with PRC2 (Zhao et al. (2008)), a complex responsible for catalyzing the transcriptionally repressive trimethylation of H3K27, which is thought to be a key step in the silencing process (Plath et al. (2003)). Intriguingly, association with PRC2 has been observed for other lncRNAs in eukaryotes (Zhao et al. (2010), Guttman et al. (2011)). There have also been reports of Xist physically interacting with PRC1, however it is unclear whether this is a direct interaction (Zhao et al. (2008)). Repeat C, in turn, is required for proper Xist RNA localization to the inactive X (Wutz et al. (2002)), and it was shown that it acts by binding to the YY1 protein (Jeon and Lee (2011)). YY1 binds to both the *Xist* locus and the Xist transcript, and is required for the nucleation of Xist RNA binding to the inactive X chromosome.

Although these observations converge towards a potential mechanism for Xist-induced chromosome silencing, the details remain unclear, and the reality is likely to be complex. For instance,

knockdown of PRC2 components (Eed and Ezh2) does not seem to abrogate gene silencing, possibly because of functional redundancy with another pathway (Zhao et al. (2008)). The evolutionary origins of the gene long remained murky, but recent research has shed some light on this topic. In a somewhat perverse twist of fate, there is evidence that the *Xist* gene evolved in part by pseudogenization of an ancestral protein-coding gene (Duret et al. (2006), Elisaphenko et al. (2008)). Interestingly, other parts of the gene seem to be derived from fragments of several transposable elements (Elisaphenko et al. (2008)).

Molecular functions of non-coding transcription

During the last few years, our knowledge of the molecular functions of non-coding transcription has exploded. Following the discovery of RNA interference (RNAi, Fire et al. (1998)), a rich body of research has illuminated the molecular functions of siRNAs and miRNAs, and characterized the machinery that mediates them (Ghildiyal and Zamore (2009), Malone et al. (2009), Krol et al. (2010), Castel and Martienssen (2013)). More recently, many detailed studies of individual lncRNAs have made clear that these transcripts perform an impressive diversity of molecular tasks (Ponting et al. (2009), Wilusz et al. (2009), Nagano and Fraser (2011), Guttman and Rinn (2012), Ulitsky and Bartel (2013)). One common theme that seems to be emerging, however, is that many lncRNAs seem to have roles in transcriptional regulation. A number of examples suggest that physical interactions between lncRNAs and transcriptional co-activators or co-repressors play mechanistic roles in this process.

Generally, one could propose three categories of functional non-coding transcription. The first one regroups cases in which the act of transcription itself, rather than the RNA product of the process, fulfills a function. The second one encompasses RNAi in a broad sense – that is, all Argonaute-mediated functions of classical small RNAs: siRNAs, miRNAs and Piwi-interacting RNAs (piRNAs). The third one, for lack of a better classification, includes all Argonaute-independent processes by which a non-coding RNA molecule can influence a biological process.

There are a few cases in which it is solidly established that non-coding transcription performs a clear functional role, independently of the RNA molecule that is synthesized in the process. At the *S. cerevisiae GAL* locus, the *GAL10* gene was shown to have an antisense transcriptional unit, named *GAL10-ncRNA* (Houseley et al. (2008)). When glucose is present in the growth medium, transcription of *GAL10-ncRNA* promotes increased H3K36 trimethylation and reduced H3 acetylation across the *GAL10*

locus. This in turn represses initiation from the *GAL10* promoter. Such processes are also at play in metazoans, for instance at the *Ultrabithorax* (*Ubx*) locus in *D. melanogaster* (Petruk et al. (2006)). In this case, the *bithoraxoid* (*bxd*) transcriptional unit lies upstream of *Ubx* in the same orientation, and its transcription represses that of the *Ubx* gene by transcriptional interference, possibly by interference with transcription factor binding at the *Ubx* promoter. This *bxd*-mediated repression has an important role in embryo patterning, as a focal deletion of the upstream *bxd* promoter causes homeotic defects.

Since the initial description of RNAi in *C. elegans*, intensive research on the topic has identified several related small RNA-based processes operating in eukaryotes, and characterized the molecular machinery that mediates them. Common to all these pathways is the use of short RNA molecules (~20-30 nucleotides) to provide specificity to the gene silencing activities of effector Argonaute (AGO) proteins. Small RNAs form intermolecular complexes with AGOs, and guide them via their complementarity to target transcripts. Three main classes of small RNAs that associate with different AGOs have been described. miRNAs, processed by the enzymes Dicer and Drosha from a stem-loop structure in longer precursors, serve as guides for the targeting of long transcripts (mRNAs and others) for post-transcriptional gene silencing (PTGS) (Krol et al. (2010)). This effect is mediated by the destabilization of target transcripts as well as translational repression. siRNAs, generated by Dicer from long double-stranded precursors, guide PTGS as well as transcriptional gene silencing (TGS) in a variety of organisms (Ghildiyal and Zamore (2009), Castel and Martienssen (2013)). Finally, piRNAs associate with a specific class of AGOs (Piwi proteins) and are mostly involved in the silencing of transposable elements by TGS and PTGS, mainly but not exclusively in the germline (Malone et al. (2009)).

The AGO-independent activities of non-coding RNAs have been the focus of intense research in recent years, and appear to be extremely diverse. Several well-characterized lncRNAs have been shown to associate with proteins that write, read or erase chromatin modifications, as well as other cofactors (Ponting et al. (2009), Wilusz et al. (2009), Nagano and Fraser (2011), Guttman and Rinn (2012), Rinn and Chang (2012), Ulitsky and Bartel (2013)). Most notably, lncRNAs such as HOTAIR (Rinn et al. (2007)) and Xist (Zhao et al. (2008)) in mammals and COLDAIR (Heo and Sung (2011)) in plants have been shown to interact directly with the PRC2 complex, which they recruit to target loci either in *cis* (COLDAIR) or in *trans* (HOTAIR, Xist). A recent study estimated that up to 30% of lncRNAs in human embryonic stem cells interact, directly or not, with at least one of 12 chromatin regulators assayed (Guttman et al. (2011)). Conversely, other ncRNAs have been found to displace transcriptional

cofactors from their target loci by competitive binding. For instance, the Gas5 RNA acts as a decoy by associating with the DNA-binding domain of the human glucocorticoid receptor, thus preventing its interaction with genomic binding sites (Kino et al. (2010)). There is also evidence that many ncRNAs associate *in vivo* with the catalytic domain of the mammalian maintenance DNA methyltransferase DNMT1, and this seems to block DNA methylation at the transcription sites of these ncRNAs (Di Ruscio et al. (2013)). By a similar mechanism, lncRNAs transcribed from a *PTEN* pseudogene act as miRNA sponges, derepressing the *PTEN* gene by competitive binding to miRNAs (Poliseno et al. (2010)). There are also instances of allosteric regulation of protein cofactors: in response to DNA damage, ncRNAs transcribed from the human *CCND1* locus bind to the TLS protein, facilitating its inhibition of the p300 co-activator and thus promoting silencing of the *CCND1* gene (Wang et al. (2008)). A role in the establishment of nuclear architecture has been demonstrated for the Neat1 lncRNA, which is required for the assembly of nuclear bodies called paraspeckles (Mao et al. (2011)). Some lncRNAs also appear to carry out functions in the cytoplasm. For instance, the human NRON RNA inhibits the activity of the NFAT transcription factor by preventing its import to the nucleus (Willingham et al. (2005)).

The discovery of large numbers of ncRNAs in eukaryotes, together with the characterization of the diversity of their molecular functions, have contributed to bringing them to the forefront of genomics. In vindication of the need for such focus, there is by now ample evidence that non-coding transcription is involved in a very broad range of biological processes, from vernalization in plants (Heo and Sung (2011)) to embryo patterning in fly (Petruk et al. (2006)) and the p53 response in mammals (Huarte et al. (2010)).

Evolutionary conservation & Population genetics

Despite recent advances in our knowledge of the biochemical activities of non-coding transcription, both the prevalence and the physiological relevance of these activities remain a matter of debate. Although definitive answers must await further experimental investigation, much can already be learned from genetics. Recent lines of evidence from comparative genomics, population genetics and the genetics of human disease seem to suggest that a sizeable proportion of non-coding transcription in various organisms might be physiologically relevant and under selective pressure. It has already become clear, however, that non-coding genes differ drastically from their protein-coding counterparts in their rates of evolutionary gain and loss, the pace of their sequence divergence, and the fitness effects of individual

mutations.

The evolutionary conservation of sequence (or other genomic features) beyond neutral-regime expectations provides strong evidence of selective pressure, implying functionality. With the availability of large numbers of complete genomes, the practical use of this concept in comparative genomics has become a standard. Some well-characterized ncRNA loci are known to have orthologs in other species that fulfill similar functions. For instance, *Xist* orthologs are involved in X chromosome inactivation throughout the therian lineage (Duret et al. (2006)). Two zebrafish lincRNA genes, *cyrano* and *megamind*, were recently found to have short patches of deep sequence conservation, and to be required for normal embryonic development. In a striking experiment, it was shown that the mouse and human orthologs of these lincRNAs were able to rescue the developmental defects induced by knockdown of the endogenous transcripts (Ulitsky et al. (2011)). These examples, however, underscore the difficulty of assessing orthology, let alone the conservation of function, for non-coding genes: beyond short stretches of conserved sequence (<300bp), these zebrafish loci had absolutely no sequence similarity to their mammalian counterparts. In fact, a substantial number of potentially orthologous lncRNA genes have been identified at syntenic positions between species, yet primary sequence conservation is undetectable (Ulitsky et al. (2011)). One explanation for the lack of obvious conservation of sequence features is our ignorance of the rules of the game. For protein-coding sequence, we have a rather detailed understanding of the meaning of different types of mutations (e.g., synonymous or non-synonymous), and this informs our interpretation of sequence variation between individuals or across species. The lack of interpretive rules (e.g., the relationship between primary sequence and RNA secondary structure) makes the task much more difficult for non-coding sequence.

In aggregate, it was shown in multiple independent studies that mammalian lincRNAs have levels of sequence conservation that are intermediate between those of protein-coding genes and (supposedly) neutrally-evolving ancestral repeats, although the effect of selection is weak (Guttman et al. (2009), Derrien et al. (2012), Young et al. (2012)). lincRNA exons are substantially less conserved than protein-coding exons, whereas the promoters of both classes of genes seem to be under almost the same level of constraint (Guttman et al. (2009), Derrien et al. (2012)). Analyses of interspecies conservation of *D. melanogaster* lincRNA exons have led to similar conclusions (Young et al. (2012)). A recent experimental survey of the liver transcriptome in 3 species of rodents concluded that the rates of gain and loss of lincRNA genes are higher than those of their protein-coding counterparts (Kutter et al. (2012)).

Although this study focused on a relatively small number of loci (150-600 per species), it clearly establishes that the two classes of genes have substantially different evolutionary turnover rates, at least in this tissue. Further supporting this hypothesis, a genome-wide computational study of human lncRNAs estimated that up to a third may have arisen in the primate lineage (Derrien et al. (2012)).

Analyses of interspecies conservation, although powerful, only detect selective pressures that are sustained over the divergence time of the species included in the comparison. They are by design not capable of detecting selection acting in a single species or a small subclade. Given the observation stated above that many lncRNA genes seem evolutionarily recent, there is a dire need to assess the magnitude of ongoing selective pressures. Population genetics provides a powerful means to measure the effects of very recent, if not current selection. Indeed, negative selection is expected to suppress the population frequency of deleterious variants, and therefore low levels of polymorphism over a locus or a class of loci are a clear hallmark of its effects. With the recent completion of large population sequencing projects in human and *Drosophila*, such approaches are now becoming practical on a genome-wide scale. One such survey found that lincRNA exons in *D. melanogaster* have a clear excess of low-frequency variants, although the suppression of polymorphism was still weaker than on protein-coding exons (Haerty and Ponting (2013)). Accordingly, a significant proportion of polymorphisms at these loci were estimated to have weakly to moderately deleterious fitness effects, although variants with strongly deleterious effects were far more prevalent in protein-coding genes. Intriguingly, the same analysis conducted on human lincRNAs failed to detect any effect of recent negative selection, and most variants at those loci seem to evolve under a neutral or near-neutral regime (Haerty and Ponting (2013)). While it cannot be ruled out that these results reflect profound differences in the biology of lncRNAs in human and fly, it is more likely that they reflect differences in the effectiveness of natural selection in the two species. *Drosophila* has been estimated to have an effective population size 3 orders of magnitude larger than human, theoretically allowing selection on variants with fitness effects up to 3 orders of magnitude smaller. Variants with similar fitness effects in human would segregate under a regime dominated by genetic drift, much as completely neutral variants would. Overall, it can be concluded that *Drosophila* lncRNAs as a class are under significant selective pressure, implying widespread functionality. Selection has little if any influence on human lncRNAs, allowing no conclusion to be drawn as to their potential functionality (Haerty and Ponting (2013)).

The association of individual polymorphisms with deleterious phenotypes provides very strong

evidence for function in select cases. Although it was not recognized as such at the time, one of the first non-coding transcription mutants in *Drosophila* was described by Edward Lewis: the *pbx*^{1} mutation at the *Bithorax* locus, which leads to homeotic transformations (Lewis (1978)), was later found to be a deletion of the upstream *bxd* promoter already mentioned earlier (Petruk et al. (2006)). In human, it is now clear that many miRNAs have roles in the pathophysiology of various cancers (Farazi et al. (2013)). Variants of the MIAT lncRNA gene have been found to be associated with myocardial infarction (Ishii et al. (2006)). Polymorphisms of the ANRIL lncRNA at the *INK4/ARF* locus are associated with higher risk for several cancers, coronary disease and type 2 diabetes (Pasmant et al. (2011)). The ANRIL RNA is involved in *INK4a* silencing through direct recruitment of PRC1. HOTAIR overexpression is a strong predictor of breast cancer metastasis and death, and experimental modulation of HOTAIR expression affects the invasiveness of epithelial cancer cells in a mouse xenograft model (Gupta et al. (2010)). These effects are thought to be mediated through ectopic recruitment of PRC2 by HOTAIR to metastasis-suppressor genes. It has also been proposed that increased expression of a transcript antisense to the *BACE1* gene (BACE1-AS) in Alzheimer's disease stabilizes the BACE1 mRNA, leading to higher levels of the BACE1 enzyme and in turn increased synthesis of amyloid-beta (Faghihi et al. (2008)).

1.2.3 Regulatory logic & Mechanisms of transcriptional regulation

Fundamental principles of transcriptional regulation

Transcriptional regulation is the ensemble of all molecular processes that participate in specifying levels of transcriptional activity at loci throughout the genome. Research on this topic focuses on two principal aspects: the mechanistic basis for changes in transcriptional activity in response to developmental or environmental cues, and the processes responsible for the maintenance of transcriptional activity and cellular state.

Seminal work in the early 1960s on bacterial genetic systems, most importantly the Lac operon and the Lambda phage in *E. coli*, established that responses to environmental cues in these contexts are genetically encoded (Jacob and Monod (1961)). It was also determined that the expression of a new phenotype involves the transcription of genes into an unstable RNA intermediate (the messenger) (Jacob and Monod (1961), Brenner et al. (1961)), and subsequent translation into proteins. In these cases, the main regulatory step was found to be the decision of whether or not to transcribe the gene.

Genetic dissection of both systems revealed that expression of a particular protein-coding sequence (the structural gene) is under the control of a distinct sequence module in *cis* (the operator), which responds to the cellular concentration of a soluble factor (the regulator) that is encoded by another structural gene. Despite the diversity and complexity of molecular mechanisms uncovered since then, the binding of sequence-specific *trans* regulators to *cis*-regulatory modules remains the conceptual cornerstone of the modern understanding of transcriptional regulation.

In addition to this, work on the Lambda phage also revealed two other fundamental principles of gene regulation. The first one is the notion of cooperativity. It was observed that direct physical interactions between molecules of lambda repressor binding to neighboring sites on the DNA greatly enhance repressor occupancy above a critical threshold concentration. This feature enables gene expression to respond in a switch-like fashion to varying concentrations of repressor (Johnson et al. (1979)). Direct or indirect cooperative interactions, either between molecules of the same transcription factor or between different factors, have since been found to underlie the workings of many known eukaryotic *cis*-regulatory sequences (Small et al. (1992), Lebrecht et al. (2005), Spitz and Furlong (2012)). The second essential principle uncovered was positive feedback in transcriptional regulation circuits, a feature that enables cellular memory. Indeed, the lambda so-called repressor also has the ability to activate the transcription of its own gene (Ptashne et al. (1976)). This feedback loop imparts bistability to the system and ensures that, once the repressor is expressed above a threshold concentration, it will keep being expressed until external signals actively reverse that decision (Ptashne (2011)). Positive feedback is an extremely important feature of virtually all known gene regulatory networks, and is thought to be key to cellular memory in many systems (Davidson (2010)).

The model of gene regulation by diffusible sequence-specific regulators and short *cis*-regulatory DNA sequences offers features that make it very amenable to the assembly of complex gene regulatory networks, and these advantages were recognized early on. Britten and Davidson argued that the repetition of *cis*-regulatory sequences throughout genomes must be the basis for the implementation of "batteries" of genes that are capable of responding in an integrated fashion to a common signal (Britten and Davidson (1969)). Although the fine details are complex, this intuition has been abundantly validated in the decades that followed their initial claims. It is now clear that concerted transcriptional regulation by transcription factors targeting stereotyped sequences at large sets of target loci is central to cellular differentiation and metazoan development (Davidson (2010)). Research on *Drosophila* in

the 1980s established that the classical homeotic genes, which establish segmental identity during early embryonic development (Lewis (1978)), all encode sequence-specific transcription factors (Hoey and Levine (1988), Desplan et al. (1988), Levine and Hoey (1988)). These genes are expressed in different domains of the early embryo that are defined by the local abundance of maternally deposited factors. In turn, their products bind distinct sets of genomic sites to activate complex developmental programs in the appropriate embryonic segments. Other classical developmental regulators (Nusslein-Volhard and Wieschaus (1980)) were subsequently shown to also be transcription factors, and there is strong evidence that the early patterning of *Drosophila* embryos is specified almost exclusively by cascades of transcriptional regulation (Segal et al. (2008)).

Much research has focused on the elucidation of the basic molecular mechanisms that underlie the function of transcription factors. Further genetic and biochemical characterization of the Lac and Lambda systems in *E. coli*, as well as the GAL system in yeast, shed light on this question. Transcriptional activators were shown to physically interact with the basal transcription machinery, and it was found that the DNA-binding and *trans*-activating activities of the yeast Gal4p activator are fulfilled by distinct domains of the protein (Brent and Ptashne (1985)). These observations led to a model of activation by recruitment, in which the function of activators is simply to recruit RNA polymerase to a given sequence, thereby increasing its effective concentration in the vicinity of the promoter (Ptashne (2005)). This early model focused explicitly on the recruitment of the basal transcription machinery to the promoter as the main rate-limiting step in the process of gene induction. It was supported by "activator bypass" experiments, in which the activation domain of Gal4p was deleted, and its DNA-binding domain was fused directly to a component of RNA polymerase II (Barberis et al. (1995)). This manipulation led to robust activation of the target gene, reinforcing the notion of RNA polymerase recruitment and pre-initiation complex (PIC) formation as the major rate-limiting step.

This general model has since undergone significant elaboration, and some major revisions. The view of activators as simple bridging factors between DNA and the basal transcription machinery has been largely vindicated, but in eukaryotes this machinery appears to be exceedingly complex. Far from consisting solely of a Pol II holoenzyme, it includes a great diversity of large macromolecular complexes that are all required for proper gene induction. These are involved, for instance, in the ATP-dependent remodeling of nucleosomes (e.g., SWI/SNF) (Struhl (1999)), the covalent modification of histones (SAGA) (Kouzarides (2007)), or the recruitment of other macromolecular complexes (Media-

tor) (Malik and Roeder (2010)). Different genes have different requirements for each of these cofactors, and in a twist that blurs the line between sequence-specific and general transcription factors (GTFs), it has recently been argued that some classical GTFs can have highly cell type-specific and gene-specific functions (Goodrich and Tjian (2010)). In stark contrast to prokaryotes, eukaryotic genomes are packaged into chromatin, which has two critical implications. The first one is that, in the absence of specific signals, nucleosomes impart global transcriptional repression on most of the genome. Genome accessibility is now recognized as a key determinant of transcriptional activity (Bell et al. (2011)). The second implication of chromatin packaging is the opportunity for the transcriptional machinery to use histones as a platform for signaling between various regulators and effectors. A multitude of chromatin marks, including DNA methylation and dozens of histone modifications, have been shown to play key roles in many aspects of genome function, in particular transcription (Kouzarides (2007)).

Importantly, the past decade has also seen a shift away from the concept of PIC formation as the main rate-limiting step in gene induction, with the growing recognition that post-initiation transcriptional pausing is a widespread and finely regulated process (Adelman and Lis (2012)). Pausing is the phenomenon by which a polymerase that has already cleared its promoter and transitioned into elongation can stop transcribing, while remaining stably associated with the gene and retaining the ability to resume transcription upon release. Dedicated factors, such as the negative elongation factor (NELF), get recruited to some genes and actively promote pausing. Other factors, such as the DRB sensitivity-inducing factor (DSIF) and positive transcription elongation factor b (P-TEFb), can subsequently override this pausing in response to specific signals. This phenomenon, once thought to be an exotic feature of a few genes, has been shown to be very common in metazoans, and to play some crucial physiological roles (reviewed in Adelman and Lis (2012)). For instance, pausing has an important developmental function in the mesoderm of early *Drosophila* embryos, where pause release allows rapid, robust and synchronous transcriptional induction across populations of cells (Boettiger and Levine (2009)). This synchrony was shown to substantially increase the phenotypic robustness of tissue morphogenesis (Lagha et al. (2012)).

In higher eukaryotes, the genomic organization of regulatory sequence elements appears to be very intricate. Many *cis*-regulatory modules (CRMs) are capable of binding multiple transcription factors, often both activators and repressors. The interplay between these factors allows CRMs to perform complex computations, such as the integration of signals from distinct inputs (Spitz and Furlong

(2012)). Whereas in bacteria and yeast most regulatory sequences lie in the immediate vicinity of their target promoters, the regulatory modules of multicellular organisms have a very distributed and non-colinear organization. Some regulatory sequences can be located megabases away from their target promoters, and often have no influence on the activity of intervening genes (Amano et al. (2009), Spitz and Furlong (2012)). The regulation of individual promoters can rely on the integration of inputs from multiple CRMs and other regulatory elements, such as enhancers, silencers, insulators or locus control regions (LCRs). Many genes also have multiple alternative promoters, which are often under the control of regulatory sequences with distinct specificities (Carninci et al. (2005), Lenhard et al. (2012)). Although the maintenance of gene expression levels and cellular states largely relies on transcription factor-mediated positive feedback (Alon (2007), Davidson (2010)), there is abundant evidence that chromatin modification-mediated feedbacks also play roles. For instance, DNA methylation and the trimethylation of H3K27 by the PRC2 complex are thought to both repress transcription and be stably inherited through mitosis (Wigler et al. (1981), Hansen and Helin (2009)).

Organization and function of RNA Polymerase II core promoters

Core promoters are genomic elements that bind and position general transcription factors and RNA polymerase, thus defining transcription start sites. They form the platform for the assembly of transcriptional preinitiation complexes (PICs), which are the ensemble of GTFs and RNA polymerase required for the onset of transcriptional initiation. Promoters for RNA polymerase II, which transcribes all protein-coding genes and most lncRNAs in eukaryotes, have varied architectures and sequence motif compositions (Figure 1.2). They are much more diverse and less stereotyped than their prokaryotic counterparts, and their functional classification, as well as the molecular basis of their function, are the focus of ongoing research. In addition to their diversity in individual genomes, they also display systematic differences between distantly related species.

Genetic and biochemical studies of model promoters have shed light on their mechanisms of action. They are usually constituted of a set of degenerate sequence motifs mediating low-affinity interactions with GTFs and the basal transcription machinery. The precise composition of this set is variable, however, and no single sequence element seems absolutely necessary for promoter function. A canonical TATA box, probably the best-characterized promoter element in eukaryotes, is present at a subset of promoters, and is involved in binding the TATA-binding protein (TBP), a component of the transcrip-

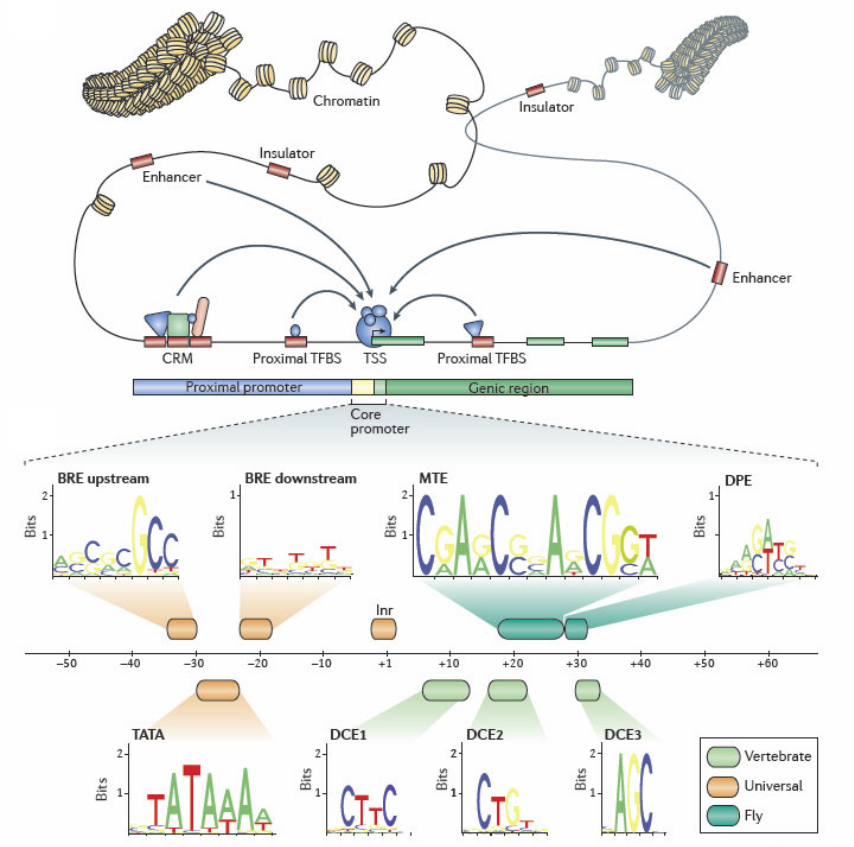

Figure 1.2. Eukaryotic RNA Polymerase II core promoters
RNA Pol II promoters play a prominent role in the regulation of mRNAs and lncRNAs, as well as other transcripts. They constitute the platform on which the inputs from all regulatory elements (proximal *cis*-regulatory modules (CRMs), distal enhancers, silencers) are integrated to determine the transcriptional output of the gene. A certain number of sequence motifs are often found at core promoter, but not all of them. Some motifs are specific to vertebrates or invertebrates. Figure reproduced from Lenhard et al. (2012).

tion factor IID complex (TFIID). The TFIIB recognition element (BRE) specifically binds another GTF, TFIIB. Two other well-characterized promoter motifs, the Initiator (INR) and the downstream promoter element (DPE), are thought to mediate interactions with the TBP-associated factor (TAF) subunits of TFIID. It is believed that the combination of several such motifs constitutes a docking site for the general transcription machinery, and provides directionality to the promoter by correctly orienting the PIC. The arrangement of binding sites defines the position of the transcription start site, which tends to be located either within an INR element or at a more degenerate pyrimidine-purine dinucleotide. Although some promoters have well-defined single initiation sites (”sharp” promoters), most allow initiation at a larger set of positions (”broad” promoters), occasionally spanning up to about 300 bp of genomic sequence (reviewed in Lenhard et al. (2012)).

Large-scale studies of transcription initiation in *Drosophila* and mammals have suggested a tentative functional classification of metazoan promoters into three categories (Carninci et al. (2005), Hoskins et al. (2011), Lenhard et al. (2012)). Tissue-specific promoters expressed in the adult organism tend to have precisely positioned TATA boxes and INR motifs, and generally display sharp initiation patterns. Ubiquitously expressed housekeeping genes include either a DNA recognition element (DRE) or other motifs that are currently poorly characterized, and have broad initiation patterns. Finally, developmentally regulated promoters generally have an INR motif, and sometimes a downstream promoter element (DPE) as well. Unlike ubiquitously expressed TSSs, they tend to have rather sharp initiation patterns.

Recent work has challenged some of these classical views, and may lead to a new understanding of promoter function. The idea that GTFs are simply recruited to core promoters by sequence-specific activators, but do not themselves play a regulatory role, has come under scrutiny after some classical GTFs were shown to have cell type and gene-specific functions. These studies also identified cell type-specific ”non-prototypical” core promoter recognition factors, such as dedicated TAFs and TBP-related factors (TRFs) (Goodrich and Tjian (2010)). High-throughput surveys of transcriptomes or of transcriptional activity have uncovered extremely widespread bidirectional initiation at *Drosophila* and mammalian promoters (Seila et al. (2008), Core et al. (2008), Kwak et al. (2013), Sigova et al. (2013)), calling into question common views on the role of promoters in orienting Pol II. New data suggests that a further degree of directionality is enforced by transcription termination signals and transcript degradation (Almada et al. (2013), Ntini et al. (2013)), and possibly by chromatin looping (Tan-Wong

et al. (2012)). High-resolution mapping of initiation complexes throughout the human genome has cast doubt on the notion of differential sequence motif composition between classes of promoters. Indeed, it was shown that the vast majority of promoters do in fact share common motifs including TATA, BRE and INR, but those motifs are substantially more degenerate than previously recognized (Venters and Pugh (2013)). There remains the possibility, however, that near-consensus and weak motifs may not be functionally equivalent.

Molecular mechanisms of enhancer function

Transcriptional enhancers are classically defined as *cis*-regulatory DNA sequences capable of stimulating transcription from basal promoters, regardless of genomic distance and orientation relative to the target promoter. They were initially discovered in the SV40 virus genome, where a 72 bp minimal sequence was found to enhance transcription from a reporter gene by over two orders of magnitude (Banerji et al. (1981), Moreau et al. (1981)). Although it was speculated that such sequence modules might play roles in gene regulation, and sequences with similar activities were quickly discovered in other viral genomes, it was unclear how relevant this phenomenon would be to endogenous regulation in eukaryotic genomes. The discovery of the first eukaryotic enhancer in an intron of the mouse immunoglobulin heavy chain gene legitimately ushered long-range transcriptional enhancers into the field of gene regulation (Banerji et al. (1983)). Importantly, whereas viral enhancers had strong activity in diverse cell types, the immunoglobulin gene enhancer displayed exquisite cell type specificity.

Pioneering work in *Drosophila* in the early 1990s provided a detailed characterization of a developmental enhancer, the regulatory module responsible for the expression of the *even-skipped* (*eve*) gene in the second of the 7 stripes of its spatial expression pattern in early embryos. *eve* stripe 2 expression depends genetically on several genes of the "gap" class (*bicoid*, *hunchback*, *giant*, *Krüppel*) and *sloppy-paired 1*, which are early regulators of embryo segmentation (Stanojevic et al. (1991)). A 480 bp sequence located ~1 kb upstream of the eve TSS was shown to recapitulate stripe 2 expression, and to encode binding sites for the products of all of these genes (Figure 1.3) (Stanojevic et al. (1989), Small et al. (1991), Small et al. (1992),). Mutations of putative binding sites abolished the binding of cognate DNA-binding proteins *in vitro*, and drastically altered the spatial expression pattern *in vivo*. The *eve* stripe 2 enhancer includes binding sites for both activators and repressors, both being essential for proper expression (Small et al. (1991)). Cooperativity between transcription factors, the precise extent

of which remains a matter of debate (Ilsley et al. (2013)), is thought to be important for the establishment of a sharp stripe of expression from crude gradients of activators and repressors (Small et al. (1992)) (Figure 1.3).

Cooperativity has since been implicated in sharpening spatial expression patterns for other developmental genes as well. Direct physical interactions between bicoid molecules bound to neighboring genomic sites are essential for the establishment of precise expression domains for hunchback, giant and Krüppel, and mutants defective in cooperative DNA binding die during embryogenesis with head and thorax defects (Lebrecht et al. (2005)). Cooperativity is not, however, a universal feature of eukaryotic enhancers. Whereas switch-like, digital regulatory behavior has great advantages in some contexts, in others a more linear, analog response is preferable. For instance, non-cooperative binding of human NF-κb to clusters of binding sites allows a graded transcriptional response to varying extracellular concentrations of the TNF-α cytokine (Giorgetti et al. (2010)).

Detailed biochemical and structural studies on model regulatory modules, such as the virus-inducible interferon beta (IFN-β) enhancer in human, form the basis of our current understanding of the molecular mechanisms that underlie their function. The IFN-β enhancer, located ~50 bp upstream of the TSS of the IFN-β gene, is constitutively nucleosome-free. The TSS, on the other hand, is masked by a strongly positioned nucleosome under non-inducing conditions, and this arrangement strongly represses transcription. Virus infection triggers the activation of three sets of transcription factors: NF-κb, interferon regulatory facors (IRFs) and ATF-2/c-Jun heterodimers (Thanos and Maniatis (1995), Agalioti et al. (2000)). These TFs, together with the architectural protein HMG I(Y), bind cooperatively to the IFN-β enhancer, forming a dense deoxyribonucleoprotein complex dubbed "enhanceosome" (Panne et al. (2007)) (Figure 1.4). The TFs together recruit the GCN5 histone acetyltransferase complex, followed by CBP in a complex with RNA polymerase II. Acetylation of the TSS nucleosome by GCN5 and CBP facilitates the recruitment of the SWI/SNF chromatin remodeling complex. Remodeling of the TSS nucleosome allows the binding of TFIID to the TATA box, which causes the docking of PolII at the core promoter and triggers transcription initiation (Thanos and Maniatis (1995), Agalioti et al. (2000)).

As in the case of the IFN-β gene, enhancer activation is often a stepwise process. So-called pioneer factors sometimes "prime" enhancers without inducing transcription, instead conferring upon them the competence to bind additional regulators (Zaret and Carroll (2011)). These particular factors are often capable of interacting with regulatory sequences that are inaccessible to other factors, for instance

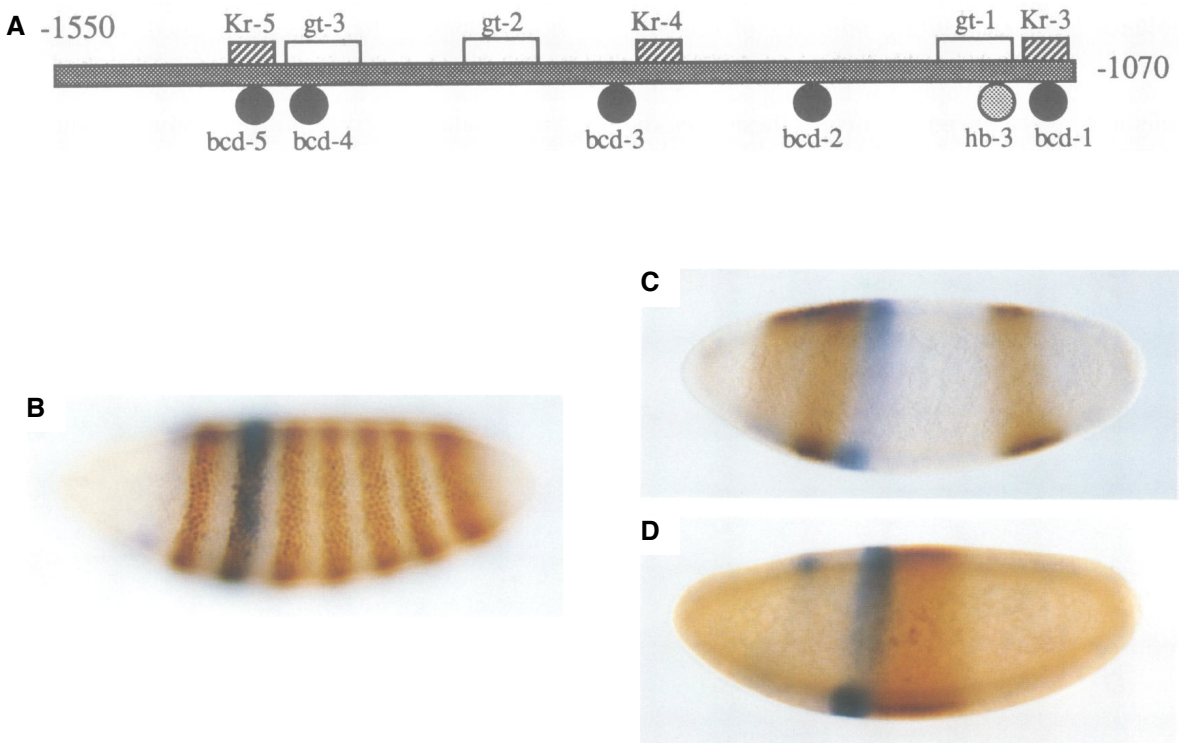

Figure 1.3. Organization and function of the *D. melanogaster eve* stripe 2 enhancer
(A) General organization of TFBSs at the *eve* stripe 2 enhancer. (B) A reporter gene driven by the minimal stripe 2 enhancer (blue) recapitulates the stripe 2 expression pattern of the endogenous *eve* gene (brown). (C, D) The expression domains of the giant (C) and Krüppel (D) repressors (in brown) define the boundaries of the eve expression domain (blue). Figure reproduced from Small et al. (1992).

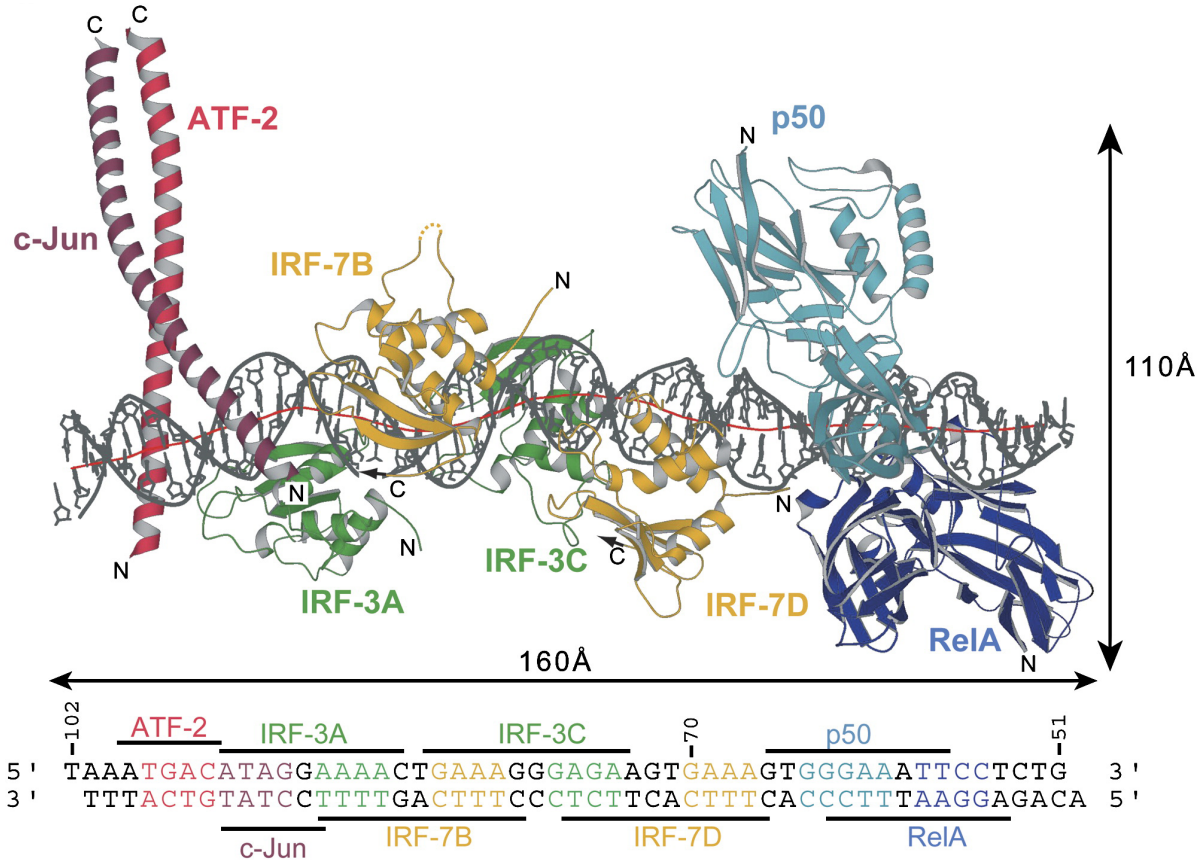

Figure 1.4. Atomic structure of the IFN-β enhanceosome
The crystallographic structure model of the enhanceosome bound by all the sequence-specific transcription factors that are known to regulate its activity shows the overall organization of binding sites, as well as the spatial arrangement of transcription factors. Figure reproduced from Panne et al. (2007).

because of nucleosome occupancy or DNA methylation. Their binding can promote local remodeling of chromatin or demethylation of DNA, thus potentiating the locus for the possible binding of other factors (Almer et al. (1986), Xu et al. (2009), Stadler et al. (2011)). In developmental settings, these successive events can sometimes be well separated temporally. For instance, the pioneer factor FoxD3 binds to the albumin gene enhancer in human embryonic stem cells, preventing DNA methylation at a FoxA1 binding site. This binding site, however, will only become occupied when FoxA1 is expressed upon endoderm differentiation. Even then, FoxA1 itself does not induce transcription, instead acting as a placeholder for the binding of further factors. Upon terminal hepatocyte differentiation, the expression and binding of these additional factors finally leads to transcriptional induction (Xu et al. (2009), Cirillo et al. (2002)).

More recent work has yielded numerous important insights into enhancer function in higher eukaryotes. To explain how some regulatory elements could act on their target genes over large genomic distances, it had long been postulated that distal regulatory elements must physically associate with their target promoters through chromatin looping (Choi and Engel (1988)). Studies of the β-globin locus provided direct evidence for this type of direct interaction (Carter et al. (2002), Tolhuis et al. (2002)), and recent work in embryonic stem cells implicated the Mediator co-activator complex, as well as cohesin, in the establishment of such chromatin loops (Kagey et al. (2010)). Very recent studies in mammalian cells identified large clusters of enhancer-like modules, dubbed "super-enhancers", that display exceptionally high transcription factor occupancy and activity (Whyte et al. (2013)). Many of those drive the expression of key developmental regulators and are directly activated by the products of their target genes, thus implementing strong positive feedback loops that may be important for the epigenetic inheritance of cell fate decisions. Mechanistically, intriguing findings have raised numerous new questions. DNA methylation patterns are highly predictive of enhancer activity (Schlesinger et al. (2013)), prompting debate as to the exact causal relationships between transcription factor binding, DNA methylation and transcriptional induction. There is also very strong evidence that many active enhancers are transcribed (Kim et al. (2010), Schlesinger et al. (2013)). The presence of canonical TATA boxes in some of them, and indications that these motifs may be under purifying selection, suggest that the transcription of enhancers may play a role in their function (Schlesinger et al. (2013)). The advent of high-throughput approaches for the genetic dissection of regulatory modules offers promising avenues for the study of eukaryotic regulatory sequences (Melnikov et al. (2012), Kheradpour et al. (2013)). Characterization

of the Lac operon promoter by such methods has yielded an extremely detailed biophysical model of the promoter, including accurate *ab initio* measurements of the strength of all intermolecular interactions involved (Kinney et al. (2010)). These technological and conceptual advances hold out hope that outstanding questions, such as the role of functional interactions between transcription factors, the grammatical rules of enhancer organization, or the role of cooperativity, may find some answers in the near future.

Genome-wide perspective

Dissecting the genomic architecture of transcriptional regulation in higher eukaryotes has proven daunting due to the large size of genomes, the complexity of regulatory landscapes, and the intricacy of functional interactions between regulatory sequences. In the past decade, however, technological advances and the financing of large functional annotation projects have led to rapid progress in our understanding of regulatory sequence.

Advances in microarray and DNA sequencing technologies have spurred fast-paced innovation in approaches to read out genome function. Chromatin profiling, in particular, has emerged as a powerful tool for the detection of regulatory activity on a genome-wide scale (Heintzman et al. (2009), Ernst et al. (2011)). Indeed, combinatorial profiles of DNA methylation and histone modifications are characteristic of the type and activity status of specific classes of regulatory elements (Schlesinger et al. (2013)). Because of the versatility and relative ease of profiling chromatin landscapes, it has now become standard practice to use such strategies to identify regulatory elements *ab initio*. The latest work by the ENCODE consortium identified almost 400,000 enhancer-like and over 70,000 promoter-like regions throughout the human genome (Bernstein et al. (2012)). Similar work in mouse identified over 234,000 putative enhancers and 53,000 putative promoters (Shen et al. (2012)). Even in the much more compact genome of *D. melanogaster*, chromatin profiling has led to the annotation of over 14,000 putative *cis*-regulatory modules and 7,000 insulators (Negre et al. (2011)). In vivo DNaseI footprinting in human cell lines and tissues revealed the existence of at least 8.4 million putative transcription factor binding sites, many of which display canonical sequence motifs and signs of purifying selection (Neph et al. (2012)). In accordance with this apparent regulatory complexity, genome-wide surveys of transcriptional initiation have identified large numbers of promoters in various genomes, and established that many genes use multiple alternative promoters (Carninci et al. (2005), Carninci et al. (2006), Hoskins et al. (2011)).

Functional interactions between promoters and distal regulatory elements are at this point poorly characterized, and even more poorly understood. The organization of eukaryotic genomes into multiple distinct domains may underlie these interactions, with individual chromosomes largely confined to their own ”chromosome territories” (Cremer and Cremer (2001), Lanctot et al. (2007)). These architectural features are thought to have implications for many aspects of genome function, including gene expression (Lanctot et al. (2007)). The rules underlying this organization, however, are only beginning to be uncovered. Recent efforts to map physical interactions in the human genome have revealed a complex hierarchical organization, with megabase-scale topological domains mostly invariant between cell types and even between species, and submegabase-scale interactions displaying much more plasticity during cell differentiation (Dixon et al. (2012), Phillips-Cremins et al. (2013), Jin et al. (2013)). This profound reorganization of fine-scale architecture is thought to reflect the cell type-specific remodeling of functional interactions between regulatory modules, such as those between enhancers and their target promoters. How these interactions are specified remains to be determined, as putative regulatory elements generally do not interact with the closest transcription start site (Sanyal et al. (2012)). Recent data suggests that CTCF and cohesin tend to anchor broad invariant domains, whereas Mediator and cohesin seem involved in mediating promoter-enhancer interactions (Phillips-Cremins et al. (2013)). There is also genetic evidence that Mediator and cohesin are required for functional interactions between enhancers and promoters in human cells (Kagey et al. (2010)). The precise relationship between physical contact and transcriptional induction is unclear, as some enhancers are already in contact with their target promoters before being fully induced by cognate signals (Jin et al. (2013)).

Although the organizing principles themselves have not yet emerged, it is abundantly clear already that distal regulatory elements and the three-dimensional organization of genomes have acute relevance to development (Attanasio et al. (2013)) and evolution (Cotney et al. (2013)), as well as to human disease (Hindorff et al. (2009), Bernstein et al. (2012), Attanasio et al. (2013)). Indeed, it has been estimated that up to 88% of disease-associated single-nucleotide polymorphisms (SNPs) yielded by various genome-wide association studies (GWAS) lie outside of protein-coding sequence (Hindorff et al. (2009)), and are substantially enriched in putative enhancer and promoter regions (Bernstein et al. (2012)).

Gene regulatory networks in development

Although much of gene regulation can be understood at the level of individual genes or even regulatory modules, some other aspects of regulatory behavior are inherent properties of broader gene regulatory networks (GRNs). The emerging properties of these higher-level systems underlie complex computations and dynamic patterns of expression: for instance binary decisions, memory of regulatory state or, conversely, transient responses (Davidson (2010)). The positive feedback loop of the previously mentioned Lambda phage regulatory network, which provides long-term memory to that system, offers a classical illustration of this principle . In addition to this feedback, mutual inhibitory interactions between the repressor and its antagonist, Cro, provide Lambda with a switch mechanism: whichever of the two regulators first reaches a high enough concentration fully represses its antagonist, thus sealing the fate of the expression state of the cell (reviewed in Ptashne (2011)).

Such stereotyped topological features are known as sub-circuits, or network motifs (Alon (2007)). They are repeatedly found in large numbers of regulatory networks, from bacteria to vertebrates, in which they perform similar logic computations. For instance, several positive feedback loops are thought to form the basis for the maintenance of embryonic stem cell fate in mammals. The master regulators of ES cell fate, Oct4, Sox2 and Nanog, in addition to activating a large set of regulator and effector genes, also directly activate the transcription of the genes that encode them (Boyer et al. (2005), Young (2011)). The versatility of network motifs is a consequence of the fact that their functional attributes depend on their topology, rather than on the particular regulator genes that compose them (Davidson (2010)). A recent theoretical study showed that, of all the possible topologies of 3-node networks, only 2 are capable of mediating transient responses (Ma et al. (2009)). This demonstrates in a telling manner that topology, rather than individual genes, is key to understanding network effects.

It has been proposed that most developmental GRNs may be complex, hierarchical assemblages of a modest number of stereotypical sub-circuit types, each performing one of the finite number of tasks required for development: for instance interpreting initial inputs, stably maintaining a defined regulatory state, or excluding other regulatory states (Oliveri et al. (2008), Davidson (2010)). Dissection of the GRN specifying the skeletogenic mesoderm in sea urchin, one of the best-characterized metazoan developmental regulatory networks to date, revealed such a modular structure: a succession of sub-circuits performing atomic processing tasks, from input signal transduction to the expression of terminal

differentiation genes (Oliveri et al. (2008)). There are also indications that early embryo patterning in *Drosophila* may well be governed by a similarly structured GRN (Davidson (2010), Stathopoulos and Levine (2005), Segal et al. (2008)). It is a cascade of regulatory interactions that establish anterior-posterior and dorsal-ventral patterns by recursively subdividing presumptive embryonic domains. The regulatory states of these presumptive domains ultimately determine the expression of distinct cellular phenotypes. Stereotyped network sub-circuits are also repeated in this GRN, and some perform tasks similar to those they have in other networks.

The ability to model complex gene regulatory networks in a quantitative and predictive manner will ultimately be key to our understanding of the genetic basis of multicellular development. Their characterization, however, is extremely complex and work-intensive, and large-scale network inference in higher eukaryotes remains a challenge (Yosef et al. (2013)). But the ability to generate data at an ever-increasing throughput, as well as new ways to analyze and interpret this data, are offering promising avenues to do so.

1.3 Rewiring circuits: Regulatory evolution in eukaryotes

Regulatory changes are believed to play a prominent role in the evolution of development and morphological phenotypes. Their unique properties set them apart from other types of genetic changes such as protein-coding mutations and gene duplications, and confer them an exceptional potential to generate phenotypic novelty. My interest in regulatory evolution largely stems from the recognition of these unique features, and this section will discuss these properties and their relevance to the evolution of development. I will first provide a brief historical perspective of principles and hypotheses regarding regulatory changes and their impact on animal development. Then I will describe a set of well-characterized examples of morphological evolution that illustrate them. A series of studies of the evolution of wing pigmentation patterns in *Drosophila*, in particular, will provide the basis for a more detailed discussion of the mechanisms by which regulatory innovation can create complex developmental patterns. This detailed discussion of the features that make them unique and worthy of interest is intended as a rationale for my focus on regulatory changes in development, which constitutes a fundamental premise of my project. An assessment of the adaptive value of the contributions of regulatory changes will follow. Finally, the last two subsections will describe the mechanisms of regulatory sequence evolution at the

molecular level, and the particular role of transposable elements in driving regulatory evolution.

Historical perspective & General considerations

As our modern understanding of gene regulation started to emerge in the early days of molecular biology, a number of authors started to recognize that the diversification of higher eukaryotes must have involved something else than the creation of completely new genes, and that regulatory variation could represent a powerful means to generate phenotypic novelty (Britten and Davidson (1969), Britten and Davidson (1971)). In their famous 1975 paper, King and Wilson compared a number of proteins that had been sequenced at the time in both humans and chimpanzees. In the face of striking similarity for all these proteins, they offered this hypothesis: "We suggest that evolutionary changes in anatomy and way of life are more often based on changes in the mechanisms controlling the expression of genes than on sequence changes in proteins" (King and Wilson (1975)). François Jacob similarly argued that the "chemical structures and functions" of diverse organisms are highly similar, and that evolution "is a matter of regulation rather than of structure" (Jacob (1977)).

Despite these early intuitions, the molecular mechanisms by which this diversification could possibly occur were and remain elusive. Early findings regarding the genetic determinism of embryonic development in *Drosophila* yielded some precious insights, however. The characterization of homeotic genes, in particular, seemed to offer reasonable hypotheses to explain some macroevolutionary transitions (Lewis (1978)). Indeed, there appeared to exist a finite number of genes capable of specifying the identity and complex phenotype of each segment in the body, and mutations of these loci were capable of striking reorganizations of the body plan by reassigning segment identities. Mutations near the *Ultrabithorax* locus, for instance, could create flies with a second pair of wings in place of halteres, similar to those of other insects (Lewis (1978)).

In the conceptual framework of hierarchical GRNs that specify developmental processes, this translates into the notion that relatively simple regulatory changes at key nodes in GRNs could have considerable influence on organismal phenotypes (Britten and Davidson (1971), Stern and Orgogozo (2009), Davidson (2010)). One important feature in evolution, recognized since the early days of comparative embryology and still remarkably current, is that history and contingency have a remarkable impact on subsequent changes. Historical constraints define future possibilities, and the layering of novel features onto existing ones gradually adds further constraints (Jacob (1977), Davidson and Erwin

(2006)). In terms of GRNs, the corollary of this is that once created, utilized and built upon, certain regulatory sub-circuits become absolute necessities, with greatly diminished potential for further change. It has been argued that some such sub-circuits highly resilient to evolutionary change, dubbed network "kernels", do indeed exist, and perform comparable fundamental tasks in distantly related organisms (Davidson and Erwin (2006)).

The notion of pleiotropy is also key to many theoretical arguments made in support of regulatory evolution as a driver of developmental evolution (Carroll (2005), Stern and Orgogozo (2009)). Indeed, in metazoans individual regulatory proteins (TFs, signaling molecules, receptors, etc) are often used in multiple different developmental processes, with context-dependent effects. Pleiotropy restricts the evolutionary potential of gene products, as any change beneficial to one process is likely to have other, potentially deleterious effects on other processes in which the molecule engages. Because of the "compartmentation" of regulatory information into multiple genetically separable modules, such as enhancers, *cis*-regulatory mutations can more easily circumvent such pleiotropic effects (Kirschner and Gerhart (1998), Raff and Raff (2000), Carroll (2005)). The same argument is often made to predict the relative scarcity of *trans* regulatory mutations, as those will likely display considerable pleiotropy due to their influence on many target genes (Carroll (2005), Stern and Orgogozo (2009)). And indeed, a number of recent studies have hinted at the conclusion that *cis*-regulatory mutations are more frequently fixed in nature than mutations of *trans* regulators (Wittkopp et al. (2008), McManus et al. (2010)).

Much debate still surrounds the types of mutations most relevant to developmental evolution, their phenotypic effect sizes, and the number of causal mutations that explain natural evolutionary changes (Stern (2000)). Classical micromutationist views, based mostly on population genetics theory and largely uninformed by the actual mechanisms through which genes direct development, hold that large numbers of mutations with minute effects must form the basis of adaptation. Macromutationist views hold that large-effect mutations do appear occasionally in natural populations and can be beneficial enough to be selected for. They are supported in part by the observation of experimentally induced mutations with surprisingly large effects. These questions have only begun to be answered by detailed studies of evolutionarily relevant genetic variation.

Regulatory changes in morphological evolution

Over the past two decades, the convergence of classical evolutionary theory and developmental genetics has produced a plethora of studies aimed at understanding the genetic basis for morphological evolution. For the reasons explained above, many of them have focused, in particular, on the role of regulatory evolution. Here I provide a few examples that illustrate the relevance of regulatory changes to the evolution of development and morphology in various organisms.

As stated earlier, the discovery of homeotic genes in *Drosophila* provided a solid genetic basis for attempting to understand the evolution of body plans in metazoans. The discovery that the emergence of the classical 8-gene *Hox* set predated the radiation of arthropods (Grenier et al. (1997)) forced the realization that the tremendous diversity of their body plans must have been brought about by some modifications of the activity of these very 8 genes (Palopoli and Patel (1996), Popadic et al. (1998), Carroll (1995)). It was originally proposed by Edward Lewis that changes to *Hox* genes must explain the major evolutionary transitions in arthropods, such as those from "millipede-like" ancestors to four-winged insects to six-legged, two-winged insects like *Drosophila* (Lewis (1978)). In keeping with this theory, it has been argued that the transition from a crustacean-like to a hexapod body plan was caused by changes to the expression patterns of *Ultrabithorax* (*Ubx*) and *Abdominal-A* (*Abd-A*), as well as protein-coding mutations of *Ubx* (Ronshaugen et al. (2002)). It has also been proposed that changes in the expression of some *Hox* genes, as well as changes in their sets of target genes, explain major differences between Lepidoptera and Diptera, such as the presence of larval abdominal limbs and two pairs of wings in Lepidoptera (Warren et al. (1994)). Some authors have argued that changes to *Hox* gene expression also underlie the evolution of other characters, such as head structures, in arthropods (Popadic et al. (1998)).

Such hypotheses about macroevolutionary transitions between distantly related taxa are exceedingly difficult to test, however, and largely rely on circumstantial evidence (Stern (2000)). In contrast, more modest changes between closely related species have been characterized in much greater detail. For instance, subtle variations in the regulation of *Ubx* are responsible for the evolution of the patterns of trichomes (hair-like structures) on the legs of adults in the *Drosophila* genus (Stern (1998)). Likewise, regulatory evolution of the *Shavenbaby* (*svb*) gene underlies the variation in larval trichome patterns between closely related *Drosophila* species (Sucena and Stern (2000), Sucena et al. (2003)). High-

resolution genetic mapping of the loci responsible for these phenotypic differences identified three distinct loci closely linked to the *svb* gene that overlap three enhancers involved in its regulation (McGregor et al. (2007)). These observations revealed that the evolution of this trait required multiple small-effect *cis*-regulatory mutations at a single gene, in agreement with classical micromutationist theory. In addition, the fact that regulatory mutations of *svb* have repeatedly been involved in the evolution of this trait in multiple lineages illustrates the idea that some genes are indeed hotspots for the evolution of a character, perhaps due to their key position in particular GRNs (Stern and Orgogozo (2009), Stern (2013), Prud'homme et al. (2006),).

Cis-regulatory mutations also account for the evolution of diverse traits in other organisms. For instance, multiple mutations affecting the regulation of *unpaired-like*, a homolog of a *D. melanogaster* gene regulating cell proliferation and differentiation, are responsible for changes in the size and shape of wings among different species of *Nasonia* wasps (Loehlin and Werren (2012)). Likewise, *cis*-regulatory mutations of the *Pitx1* gene are responsible for the reduction of pelvic fin size in stickleback fish (Shapiro et al. (2004)). Novel enhancer-like regulatory activities in the embryonic limb have evolved in the human lineage, and may underlie certain human-specific morphological traits (Cotney et al. (2013)).

This set of studies, and many others, point to a major role for *cis*-regulatory mutations in the evolution of development and morphology, and have begun to refine our understanding of the genetic and developmental mechanisms that underlie these transitions. This is not to say, however, that protein-coding mutations do not matter to morphological evolution: for instance, the evolution of a head crest in pigeons was recently shown to result from a single non-synonymous mutation in the *EphB2* gene (Shapiro et al. (2013)). The relative contributions of both mechanisms, and the precise influence of small- and large-effect mutations, remain very much a matter of debate.

Evolution of pigmentation patterns in *Drosophila*

The evolution of adult pigmentation patterns in *Drosophila* offers a particularly valuable case study, as it has been characterized more thoroughly than any other morphological phenotype, and it illustrates a number of the principles that I introduced earlier. Below I explore this example through a set of studies published over the last decade.

The dark pigmentation of adult body parts (abdomen, bristles, wings, etc) is due to the local production of dark melanin during pupal development by the product of the *yellow* (*y*) gene (Wittkopp

et al. (2002)). Other genes modulate this process, such as *ebony* (*e*), whose product inhibits melanin production. *y* expression in pupae generally prefigures the adult pigmentation pattern, and the evolution of *y* regulation has repeatedly been found to play a major role in the evolution of adult pigmentation (Carroll (2005)). In addition to its role in the formation of multiple independent pigmentation patterns, *y* is also expressed in some neurons and is thought to play an important role in male courtship behavior (Radovic et al. (2002), Drapeau et al. (2003), Drapeau et al. (2006)). This accumulation of multiple independently selectable functions makes *y* a stereotypical example of a gene in which protein-coding mutations would have extremely pleiotropic effects. As a consequence, selective pressure on each trait would likely dramatically reduce the evolutionary plasticity of all other traits. Instead, pigmentation pattern evolution has repeatedly evaded pleiotropy-related constraints by exploiting *cis*-regulatory mutations at the *y* locus to independently modify expression patterns in various parts of the body (Wittkopp et al. (2002), Carroll (2005), Gompel et al. (2005)). Through the separation of regulatory functions specifying different phenotypes into physically distinct modules, the *y* locus has been ideally shaped by selection for this genetic uncoupling of phenotypes. In the few cases in which the genetic causes of evolutionary change have been finely mapped, there appear to be several causal mutations involved, in agreement with the classical micromutationist view that evolutionary change is brought about by the progressive accumulation of small-effect mutations (Gompel et al. (2005)).

A study by Wittkopp et al. explored the genetic causes of the divergence in adult body pigmentation patterns (Wittkopp et al. (2002)). *yellow* expression in pupae differs between distantly related *Drosophila* species, and it is correlated with the distribution of melanin on the adult body. Transgenes from *D. subobscura* and *D. virilis* in *D. melanogaster* drive species-specific expression patterns, which points to the implication of *cis*-regulatory mutations. However, *D. melanogaster* transgenes in *D. virilis* do not fully recapitulate the original expression pattern, suggesting that there are differences in *trans* regulators as well. In addition, ectopic *y* expression was found to specify ectopic pigmentation in some genetic backgrounds only. An unidentified modifier mutation on another chromosome (not *ebony*) was found to be responsible for this effect, implying that epistatic interactions also modulate pigmentation patterns.

Work by Gompel et al. focused on the evolution of a wing spot in *D. biarmipes* (Gompel et al. (2005)). Wing pigmentation patterns are very diverse throughout the *Drosophila* genus, and are thought to have varied functions in camouflage, mimicry, thermoregulation and mate selection. The presence

of a dark pigmentation spot on the anterior part of the wing tips of males is a derived character in *D. biarmipes*, a close relative of *D. melanogaster*. Both species also display homogeneous light pigmentation throughout the wing. *yellow* expression in pupae prefigures the adult pigmentation pattern, and the two species have homologous wing enhancers upstream of *y*. In *D. melanogaster*, the *D. biarmipes* enhancer drives transgene expression in a pattern reminiscent of the wing spot, indicating that mutations within the enhancer account to a first approximation for the evolution of the new pigmentation pattern. Further analysis revealed that the *D. biarmipes* element has acquired binding sites for the Engrailed repressor, as well as for an unidentified activator. The *engrailed* gene has a deeply conserved role in wing development, and is expressed in the posterior half of pupal wings. Ectopic *y* expression in *D. melanogaster* wings is not sufficient to cause ectopic pigmentation, suggesting that other loci are involved. Incidentally, the male wing spot in *D. biarmipes* is associated with a localized down-regulation of the melanin-inhibitory *ebony* gene in pupal wings, suggesting that changes affecting *ebony* expression have played a role. Overall, this study showed that the evolution of a novel pigmentation pattern involved multiple *cis*-regulatory mutations at the *y* locus, as well as additional changes at other loci. The *y* locus mutations, by creating new regulatory inputs, co-opted the deeply conserved expression patterns of developmental regulators such as engrailed. The evolution of elaborate wing patterns in *D. guttifera* appears to have followed a similar strategy, through the co-option of a preexisting wingless expression pattern. These observations suggest a straightforward evolutionary path for the emergence of complex features, through the layering of new regulatory interactions onto existing ones.

Very recent work by Arnoult et al. (Figure 1.5) has built upon this study, and led to a detailed understanding of the genetic mechanisms underlying the emergence and subsequent diversification of a wing pigmentation trait in the *D. melanogaster* group (Arnoult et al. (2013)). The male-specific wing spot of *D. biarmipes* has homologs in several other species that display variation in their shapes and intensities. *yellow* expression in pupal wings foreshadows the adult pigmentation patterns, and it is driven by homologous enhancers in all species. The transcription factor Distal-less (Dll), a well-defined regulator of wing patterning, has several binding sites in the *D. biarmipes* enhancer, and is both necessary and sufficient to trigger its activation. Dll was also found to somehow repress the expression of *ebony*. Interestingly, ectopic Dll expression is sufficient to direct ectopic pigmentation, but only in spotted species. Together, these observations suggest that the initial emergence of the wing spot involved the evolution of regulatory interactions between Dll and at least two different pigmentation genes. Intriguingly, the

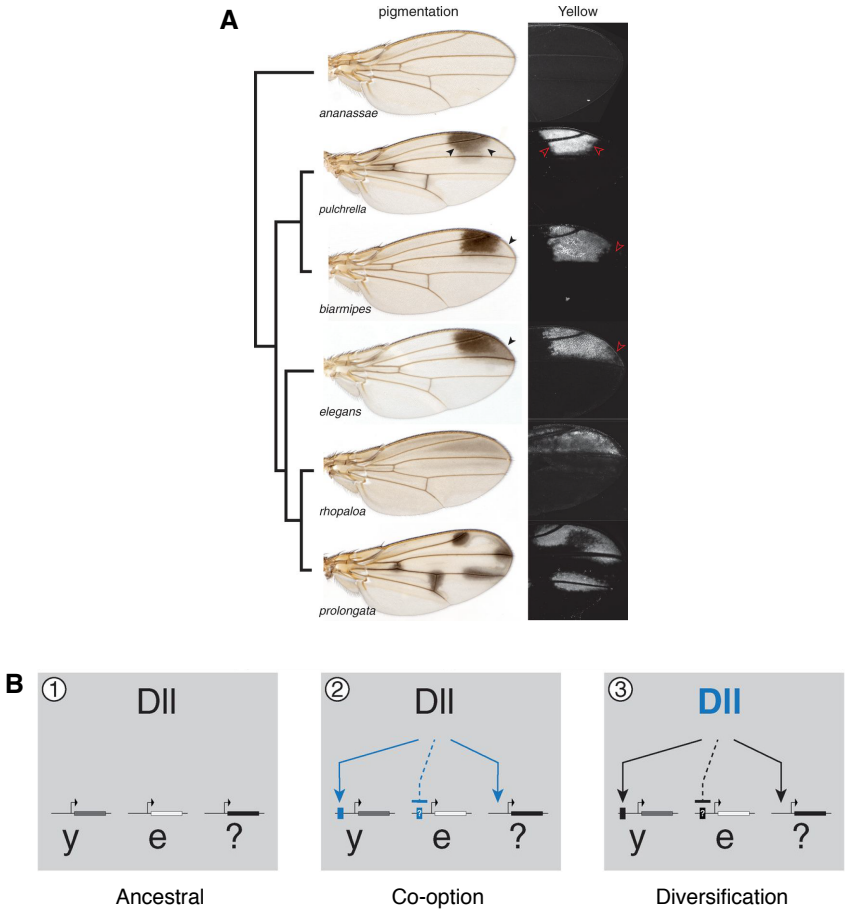

Figure 1.5. Evolution of wing pigmentation spots in *Drosophila*
(A) Phylogenetic tree (left) and wing pigmentations patterns (middle) of several *Drosophila* species. The expression pattern of the *yellow* gene in developing wings (right) prefigures the adult pigmentation pattern. (B) Model for the evolution of the wing pigmentation spot. The co-option of regulatory interactions by pigmentation genes is thought to underlie the initial emergence of the spot. Subsequent changes in the expression of an upstream regulator of these genes were then responsible for the diversification of wing spot shapes. Figure reproduced from Arnoult et al. (2013).

D. biarmipes enhancer recapitulates the exact spot expression pattern as a transgene in *D. biarmipes*, but not in *D. melanogaster*, suggesting subtle differences in *trans* regulators between the two species. Strikingly, at late developmental stages, after wing patterning is complete, Dll expression patterns diverge between the different spotted species, and in all cases accurately prefigure the adult pigmentation patterns. This body of work indicates that the wing spot originally emerged through the recruitment of multiple regulatory inputs (Dll, engrailed, etc) by multiple pigmentation genes, and that the subsequent diversification of spot characteristics involved modifications of the expression of the main inducer, Dll. It illustrates how Dll has become a recurrent target for the diversification of spot patterns, and suggests a model in which different evolutionary events might arise from changes at different hierarchical levels of developmental gene regulatory networks. Possibly, the emergence of morphological novelties requires changes to the regulation of effector genes, whereas their subsequent diversification involves the spatial redistribution of upstream master regulators.

These studies have together contributed a great deal to our understanding of the role of *cis*-regulatory changes in morphological evolution. They seem to confirm a number of theoretical predictions regarding the avoidance of pleiotropic effects, the co-option of existing regulatory information, the number and effect sizes of causal mutations, and the existence of "hotspot" genes repeatedly targeted in multiple lineages for the evolution of the same trait. Interestingly, the evolution of color patterns in mouse is caused by *cis*-regulatory mutations at the *Agouti* locus (Manceau et al. (2011)), raising the possibility that some of these principles may be generalizable.

Relevance to adaptive evolution

Although it is by now well established that *cis*-regulatory mutations contribute to phenotypic evolution, it is not entirely clear yet to what extent this evolution is adaptive (Fay and Wittkopp (2008)). In recent years, comparative genomics and population genetics have provided powerful tools to start investigating this question. In vertebrates and humans in particular, the vast amounts of population variation data that have been generated have been harnessed for this purpose, and analyses point to a very prevalent role for regulatory variation in adaptation on various time scales. Recently generated population data for *Drosophila* should provide similar insights in the near future and allow comparisons with vertebrates.

Freshwater adaptation in sticklebacks constitutes a powerful model system, and it provides a useful illustration of the possible contributions of regulatory variation to recent evolution. Marine

sticklebacks have colonized freshwater habitats formed since the last ice age in many locations around the world, and have progressively adapted to this new environment. The repeated evolution of similar phenotypic traits in similar environments demonstrates unambiguously that these changes have been driven by natural selection. Genome sequencing of multiple wild marine and freshwater isolates from diverse geographical locations allowed the detection of recurrently selected variants across the genome. This analysis revealed that, although protein-coding variants did play a role, it is predominantly *cis*-regulatory variants that have driven adaptation (Jones et al. (2012)).

Many recent studies have investigated the prevalence of heritable variation affecting gene expression, which is the first necessary condition for evolution. In diverse organisms, there is growing evidence that regulatory variation indeed abounds in natural populations. In *Drosophila*, transcriptome profiling in inbred lines recently derived from wild isolates has revealed extremely widespread heritable variation in transcript levels (Ayroles et al. (2009)). In human, genome-wide mapping of genetic variants affecting transcript levels, known as expression quantitative trait loci (eQTLs), has identified thousands of common variants segregating in natural populations (Lappalainen et al. (2013)). This study also identified hundreds of additional QTLs affecting alternative splicing, and another one revealed the existence of abundant genetic variation affecting protein levels (Wu et al. (2013)). It has also been shown that polymorphisms affect chromatin accessibility at thousands of loci across the human genome, and that these variants often overlap predicted transcription factor binding sites (Degner et al. (2012)). Similarly in mouse, there is abundant standing genetic variation affecting the function of individual enhancers (Heinz et al. (2013)), and enhancers have been shown to regulate developmental phenotypes such as craniofacial morphology (Attanasio et al. (2013)). In conclusion, it is very clear that regulatory variation is abundant in natural populations of various organisms, and that this variation has the potential to produce selectable phenotypic effects.

In addition, there is widespread evidence of fast regulatory divergence between species. Two recent studies have focused on the divergence of tissue-specific gene expression and splice isoform abundance across multiple vertebrates (Merkin et al. (2012), Barbosa-Morais et al. (2012)). Both differ very substantially between species, and alternative splicing was found to diverge particularly fast. However, although a relative enrichment for changes in alternative splicing affecting phosphorylation sites and protein-protein interaction domains suggests a contribution of positive selection, it is unclear to what extent transcriptome evolution is adaptive. A number of authors have even argued that it may

be primarily neutral, driven by genetic drift and buffered by compensation at other levels of regulation (Khaitovich et al. (2004), Khaitovich et al. (2005), Fay and Wittkopp (2008), Staubach et al. (2010), Khan et al. (2013), McManus et al. (2013)).

On the other hand, a large body if recent work has provided evidence that regulatory variation contributes substantially to lineage-specific adaptation. Comparative genomics approaches have revealed widespread adaptive acquisition of new regulatory sequences throughout the vertebrate radiation. Regulatory innovation seems to have occurred in three successive waves that preferentially affected different functional categories of genes (Lowe et al. (2011)). Another study identified almost a thousand conserved non-coding sequences that display markedly accelerated evolution in the human lineage, suggesting that they are under positive selection for human-specific traits (Prabhakar et al. (2006)). Estimates of selective forces acting on human transcription factor binding sites point to an important contribution of regulatory mutations to adaptive evolution since the divergence from chimpanzee (Arbiza et al. (2013)). The analysis of population genetics data has revealed reduced polymorphism, implying purifying selection, on recently acquired regulatory functions in human (Ward and Kellis (2012)). Human-specific constraint in non-conserved regulatory regions was detected, notably, near genes with roles in nerve growth and color vision, consistent with purifying selection for novel functions.

Interestingly, at least one recent study has attributed a significant role for regulatory variation in population-specific adaptation within the human species. Using population data, Fraser argued that regulatory mutations drive local adaptations in different human populations throughout the world. Strikingly, he found that regulatory mutations were over ten times more likely than protein sequence changes to underlie local adaptation (Fraser (2013)).

Molecular mechanisms of regulatory innovation

Efforts to characterize the evolutionary dynamics of regulatory sequences and of their functional output have revealed a complex picture. There is strong evidence of widespread purifying selection on individual transcription factor binding sites (TFBSs) in *Drosophila* (Clark et al. (2007)) and in mammals (Arbiza et al. (2013)), yet comparative studies of TF binding have demonstrated fast-paced turnover of TFBSs, particularly in mammals (Moses et al. (2006), Odom et al. (2007), Stefflova et al. (2013)). Our mechanistic understanding of regulatory sequences is still limited, and it is difficult to predict how rearrangements of regulatory sequences affect their ultimate function.

The *Drosophila eve* stripe 2 enhancer, described earlier, provides one of the best-characterized examples of regulatory sequence evolution to date. Precise spatiotemporal expression of *eve* in early embryos is crucial, and accordingly the function of the stripe 2 enhancer is tightly conserved across several species as distantly related as *D. melanogaster* and *D. pseudoobscura* (Ludwig et al. (1998), Ludwig et al. (2000), Ludwig et al. (2005)) or even sepsids (Hare et al. (2008)). Surprisingly, however, the sequence of the enhancer appears to have diverged substantially between species (Ludwig et al. (1998), Ludwig et al. (2000), Ludwig et al. (2005), Ho et al. (2009)). An early study revealed that, of the 17 *D. melanogaster* TFBSs identified at the time, only 3 are perfectly conserved across 4 *Drosophila* species considered (Ludwig et al. (1998)). Compared to the *D. melanogaster* element, the *D. pseudoobscura* enhancer was found to lack a Bicoid binding site, but to have an additional Krüppel binding site, and several other TFBSs are weakened by substitutions. Importantly, the missing Bicoid binding site is known to be essential for enhancer function in *D. melanogaster* (Ludwig et al. (1998)). To explain this high degree of functional conservation despite substantial sequence divergence, it has been proposed that TFBS mutations with small effects on *eve* expression can reach fixation by drift at an appreciable rate, and that stabilizing selection maintains enhancer function in the long run by favoring mutations that offset these weakly deleterious effects. (Ludwig et al. (2000), Ludwig et al. (2005)) According to this model, homologous regulatory modules are expected to differ by many functionally compensatory mutations.

Stefflova, Thybert et al. used comparative profiling of TF binding in the liver of multiple rodents to estimate the degree of TFBS divergence between closely related species, and to investigate the determinants of conservation and change (Stefflova et al. (2013)). In agreement with previous work, they observed fast turnover of individual TFBSs. By measuring binding for several TFs that are known to often co-bind the same *cis*-regulatory modules, they were able to demonstrate that cooperativity between factors is a major determinant of TFBS conservation. Binding sites in co-bound regions are under stronger purifying selection than isolated sites, and the loss of a binding site often causes the loss of binding of other TFs in the same module.

A genome-wide study of binding sites for 78 human TFs provided insights into the selective forces that shape regulatory evolution (Arbiza et al. (2013)). About a third of all nucleotides in these TFBSs were found to evolve under a non-neutral regime. The information content of each position in a TFBS motif and the overall affinity of individual motifs have a strong influence on the magnitude of

negative selection. Positive selection was also found to be prevalent, showing that variation in these TFBS has contributed significantly to adaptation in the human lineage.

The rate of TFBS gain and loss is much lower in *Drosophila* than in mammals, although it is still consequential (Moses et al. (2006), Odom et al. (2007), Bradley et al. (2010), He et al. (2011), Stefflova et al. (2013)). It has often been proposed that differences in genome architecture can account for these different dynamics, as most fly genomes are considerably smaller, and more densely covered by constrained functional elements, than their mammalian counterparts. Some authors have also argued that population genetics probably also partly explains this discrepancy, as effective population sizes in *Drosophila* are thought to be dramatically larger than in mammals, which greatly potentiates the influence of natural selection (Stefflova et al. (2013)).

Although the mutational mechanisms driving the gain, loss and modification of regulatory elements are diverse, a large body of work now supports the long-standing hypothesis that transposable elements play an important role in the evolution of regulatory sequence. Through their ability to distribute stereotyped regulatory modules throughout their host genomes, they are believed to contribute greatly to the assembly of complex gene regulatory networks. The next section explores this hypothesis and the evidence supporting it.

Transposons as regulatory elements & Vectors of genetic innovation

Transposons are genetic elements capable of physically moving, or transposing, to other loci within the genome that hosts them. They are broadly classified according to their molecular mechanisms of transposition (Kazazian (2004)). One class transposes through direct excision from the locus of origin and integration at a new one, and is referred to as DNA transposons. Other elements are first copied into an RNA intermediate, before being reverse-transcribed into DNA and integrated at a new locus, and are termed retrotransposons. These fall into two classes: those with long terminal repeats (LTR) and those without, respectively known as LTR and non-LTR retrotransposons. The latter include two superfamilies of well-known elements, called long and short interspersed elements (LINEs and SINEs). All three main classes of transposons – DNA, LTR and non-LTR – include two related types of elements. Some encode the enzymatic machinery required for their own transposition, and are therefore called autonomous elements. Others, thought to have evolved from autonomous transposons by deletion of internal sequences, do not encode this machinery and require the expression of related autonomous

elements to mediate their transposition. Those are known as non-autonomous elements. Transposons have been discovered in both prokaryotes and eukaryotes, and constitute a very significant portion of the genomes of higher eukaryotes. For instance, they account for about 50% of all bases in the human genome (Lander et al. (2001)) and over 80% of the maize genome (Schnable et al. (2009)).

When Barbara McClintock discovered transposons in maize, she described them as mobile "controlling elements" affecting the expression of individual genes, as made evident by the modification of specific phenotypes (McClintock (1956)). The mechanisms underlying this control were unknown, as the molecular underpinnings of gene regulation were only beginning to be investigated, and the broader relevance of the phenomenon was unclear at the time. Yet, she noted early on that "controlling elements appear to reflect the presence in the nucleus of highly integrated systems operating to control gene action", hinting at the possibility of a more general regulatory role. These findings resonated with the ideas of Britten and Davidson, and their theoretical considerations regarding the potential roles of repeated sequences in the organization of the gene regulatory networks orchestrating multicellular development. They underscored the outstanding potential of mobile genetic elements for driving the evolution of such networks through the "saltatory replication" of regulatory elements (Britten and Davidson (1969), Britten and Davidson (1971)).

Subsequently, as more was discovered about their properties and their impact on host genomes, transposons came to be viewed primarily as genomic parasites, surviving and thriving solely because of their ability to replicate faster than they are eliminated. This "selfish DNA" hypothesis elegantly relies on the sufficiency of selectively near-neutral proliferation as its sole explanatory concept, and gained considerable traction (Orgel and Crick (1980), Charlesworth et al. (1994)). The low population frequency of many mobile element insertions in *Drosophila* appeared consistent with this theory (Charlesworth et al. (1992), Charlesworth et al. (1994)). It was also supported by the discovery of deleterious effects associated with transposons, and the scarcity of known beneficial effects. The phenomenon of hybrid dysgenesis, by which crosses between different strains of *Drosophila* can have dramatically reduced fertility, is caused by the unleashing of transposons present in only one of the two strains (Pelisson (1981), Bingham et al. (1982), Malone et al. (2009)). Other work in fly found that ectopic recombination between different insertions of an element is a significant source of selective pressure limiting transposon proliferation (Charlesworth et al. (1992), Charlesworth et al. (1994)). This balance between rapid expansion and weak negative selection can potentially explain the maintenance of mobile elements around

a steady-state copy number.

These views also appeared consistent with the more recent discovery of dedicated molecular mechanisms that mediate the silencing of mobile elements. In the germline as well as in somatic tissues, RNA interference-related pathways repress the expression of transposons at the transcriptional and post-transcriptional levels (Malone et al. (2009)). These mechanisms have been characterized in depth in *Drosophila*. In *D. melanogaster* ovaries, the piRNA pathway uses short RNAs to guide the transcriptional and post-transcriptional silencing of transposons by Argonautes of the Piwi clade (Malone et al. (2009), Malone et al. (2009), Rozhkov et al. (2013)). Chromosomal loci that generate primary piRNAs, which are responsible for initiating this process, are composed primarily of transposon fragments. For this reason, they are thought to constitute a catalog of elements present in the genome. It is believed that the insertion of new types of transposons into such clusters underlies their silencing. Through this organization, the piRNA pathway is able to provide a form of adaptive immunity against transposons (Malone et al. (2009)). In somatic tissues, transposons are post-transcriptionally silenced by Ago2 and endogenous siRNAs (Ghildiyal et al. (2008), Czech et al. (2008), Ghildiyal and Zamore (2009)). Despite these repression mechanisms, transposable elements have long been known to be expressed in certain somatic tissues, and to encode *cis*-regulatory sequences that determine their expression specificity (Ding and Lipshitz (1994), Bronner et al. (1995), Kerber et al. (1996), Graveley et al. (2010)). Their patterns of expression are thought to be the result of a balance between intrinsic regulatory elements and silencing by the host (Ghildiyal and Zamore (2009)).

More recent data has come to support at least partly the early views of transposons as occasionally beneficial, almost symbiotic genetic elements. Work in plants has confirmed the role of transposons in gene regulation, as well as other aspects of chromosome biology, and characterized some of the mechanisms by which they act (Lippman et al. (2004)). The study of an ongoing burst of transposition in rice revealed a near-complete avoidance of integration into exons, as well as weak effects of new insertions on gene expression under normal conditions, thus explaining how genomes can withstand even robust expansions of some elements. Strikingly, new insertions had the ability to render neighboring genes stress-responsive, illustrating how transposons can rewire regulatory circuits in potentially beneficial ways (Naito et al. (2009)). In mammals, early comparative genomics work identified deeply conserved regulatory modules derived from transposable elements, and suggested that they might be more prevalent than many believed at the time (Bejerano et al. (2006), Silva et al. (2003), Lowe et al.

(2007)). Large-scale surveys of transcription initiation sites in mouse and human have revealed the existence of tens of thousands of active retrotransposon-derived promoters, many of which are likely to drive the expression of protein-coding genes (Faulkner et al. (2009), Djebali et al. (2012)). A broad survey of DNaseI-hypersensitive sites in human samples identified hundreds of thousands of putative regulatory elements in transposable elements (Thurman et al. (2012)). LTR retrotransposons have globally reshaped the transcriptional network of the *p53* tumor-suppressor gene in primates, and account for about a third of human-specific p53 binding sites (Wang et al. (2007)). The genome-wide binding profiles of OCT4, NANOG and CTCF have diverged very substantially between human and mouse embryonic stem cells, and transposable elements have contributed up to 25% of all new binding sites in each lineage (Kunarso et al. (2010)). Large-scale GRN remodeling by a Eutherian-specific transposable element, MER20, is thought to have contributed to the evolution of pregnancy by distributing hundreds of transcriptional enhancers, insulators and repressors throughout the genome (Lynch et al. (2011)). Multiple waves of retrotransposon expansion have diversified the repertoire of CTCF binding sites in several mammalian lineages, and the new sites often function as transcriptional insulators (Schmidt et al. (2012)). It was also recently shown that transposons have had a major influence on the diversification and regulation of vertebrate long non-coding RNAs (Kapusta et al. (2013)).

Transposons are now viewed as major contributors to the evolution of gene regulation at the transcriptional and post-transcriptional levels, in an echo to the visionary hypotheses of McClintock, Davidson and Britten. A recent survey of selective constraint across 29 mammalian genomes identified over 280,000 conserved non-exonic elements that have been co-opted from transposons (Lindblad-Toh et al. (2011)). Many of those are thought to have roles in gene regulation. Furthermore, it was estimated that approximately 19% of Eutherian-specific conserved elements are derived from transposons, a number that underscores how truly central they are to the evolutionary process. It remains unclear at this point, however, what their exact contributions are to developmental gene expression in various organisms, and in particular in *Drosophila*. Transposons are less abundant in some fly species, such as *D. melanogaster*, than they are in most mammals (Clark et al. (2007)). Their population dynamics are also different, and once again it is possible that genome architecture and population genetics may account for quantitative differences between flies and mammals. There is ample variation between different species, however, and more recent population frequency data supports the idea that transposon insertions can indeed be adaptive in wild *D. melanogaster* populations (Petrov et al. (2011)). Further

efforts in both functional genomics and population genetics will be required to address these issues in the future.

1.4 A functional genomics approach to the study of evolution

Technological improvements over the past decade have allowed genome-wide comparisons of transcriptional output and regulatory mechanisms between species. Various forms of cDNA sequencing have been used to establish a census of RNA molecules present in the transcriptomes of different organisms, and chromatin immunoprecipitation methods coupled to sequencing (ChIP-seq) have explored the divergence of protein-DNA interactions involved in transcriptional regulation. In this section, I will provide a brief overview of current methods for transcriptome analysis and of the biological insights they have yielded. The first part describes the techniques currently available for genome-wide surveys of transcriptomes, and highlights particular improvements that would benefit comparative transcriptome analysis. The second part describes the current state of transcriptome evolution studies in *Drosophila*, and discusses unresolved questions that warrant further investigation.

High-throughput techniques for the survey of transcriptomes

The democratization of high-throughput sequencing has led to the development of a large number of techniques for transcriptome analysis based on complementary DNA (cDNA) sequencing. Given the current impracticality of carrying out full-length cDNA sequencing on a large scale, different methods put an emphasis on accurately recovering different features of a transcriptome.

The most generic, versatile and widely used technique is shotgun cDNA sequencing, commonly known as RNA-seq (Mortazavi et al. (2008)). This method is conceptually a direct successor to expressed sequence tag (EST) sequencing, as it simply consists of the analysis of random short fragments of cDNA molecules. It has the advantage of providing coverage for the entire length of the template RNA, with the exception of the very ends, which are not captured by most common protocols (Levin et al. (2010), Steijger et al. (2013)). This renders the precise detection of transcription start sites (TSSs) impossible. In addition, the overlap of transcribed sequences for a gene that possesses more than a single promoter – the rule rather than the exception – makes it extremely difficult to deconvolve the contributions of individual promoters. A recent assessment of computational methods for RNA-seq data

analysis revealed overall poor performance for the identification of promoters and the quantification of transcript abundance. Therefore, RNA-seq is currently best suited for the analysis of RNA abundance at the level of whole genes, or for the analysis of local splicing patterns (Steijger et al. (2013)).

For studies that focus specifically on transcriptional regulation, however, it is crucial to be able to accurately decipher the activity levels of individual promoters, which are the functional units on which such regulation acts. This need motivated the development of approaches based on 5'-complete cDNA sequencing, which can accurately pinpoint TSSs. The most widely known of those is CAGE (Cap Analysis of Gene Expression), a technique based on the biotinylation of the 5' cap of Pol II transcripts and the affinity purification of 5'-complete cDNAs associated with them (Kodzius et al. (2006), Valen et al. (2009)). CAGE has made numerous valuable contributions, but its specificity for TSSs is still limited (see Chapter 2), it is a very cumbersome protocol, and it only yields very short sequence tags (~27 bases). This last point is a major limitation for the analysis of TSSs in repeated sequences and, crucially, it does not allow the attribution of individual promoters to annotated genes or transcripts. This makes the interpretation of CAGE data rather ambiguous. Other protocols have been developed to address these issues, and allow paired-end sequencing of longer cDNA fragments (Ni et al. (2010), Plessy et al. (2010)). This feature allows the attribution of promoters to annotations; however, it comes at the cost of a sharp drop in specificity for TSSs (see Chapter 2).

Other techniques capture 3'-complete cDNAs, and are specifically geared towards the analysis of cleavage and polyadenylation sites (Jan et al. (2011), Hoque et al. (2013)). Additionally, new computational methods for transcriptome reconstruction rely on the integrative analysis of multiple data types – for instance, RNA-seq, CAGE and 3' end sequencing (Li et al. (2011)). Although the identification of transcript ends should be improved by such methods, the accuracy of reconstruction will still be limited by the nature of RNA-seq data. Ultimately, only methods based on full-length cDNA sequencing can be optimally accurate for the detection of TSSs and the characterization of the transcripts they generate. Such approaches, based on Sanger sequencing or more recently on third-generation platforms (Sharon et al. (2013) and Appendix 2), are cumbersome and their throughput is still too low for most applications.

The limitations of current techniques for TSS identification and promoter activity profiling prompted us to develop a new method that would address several of these shortcomings. This method, RAMPAGE (RNA Annotation and Mapping of Promoters for the Analysis of Gene Expression), de-

livers substantial improvements in terms of TSS specificity, cDNA insert length, sequence read length, transcript quantification and protocol streamlining. The method and its validation are described in depth in Chapter 2.

Drosophila as a model system for transcriptome evolution

D. melanogaster, with its compact and well-annotated genome, short life cycle, small size and well-characterized development, is ideally suited as a model system for genome-wide studies of developmental gene expression in a complex metazoan. In addition, the availability of inbred laboratory strains and reference genomes for multiple species has made it a favorite for transcriptome evolution studies. Over the last decade, a number of surveys have explored the variation of gene expression within and between species, and the selective forces that shape regulatory evolution.

In agreement with previous analyses of small gene sets, early genome-wide comparative studies found abundant regulatory variation both within (Rifkin et al. (2003), Ayroles et al. (2009)) and between (Rifkin et al. (2003)) species. For instance, an analysis of gene expression divergence among several strains of *D. melanogaster* and other *Drosophila* species revealed that over a quarter of all genes display significant expression differences between at least two strains or species (Rifkin et al. (2003)). In support of theoretical considerations regarding mutations with pleiotropic effects, it was observed that genes encoding transcription factors have evolutionarily stable expression specificities, whereas the expression of their target genes is substantially less constrained (Rifkin et al. (2003)). Similarly, regulatory changes in *cis* are somewhat more prevalent than those in *trans*, suggesting that negative selection is potentiated by the pleiotropy of *trans* mutations (Wittkopp et al. (2004), Wittkopp et al. (2008), McManus et al. (2010)).

Importantly, the comparison of gene expression variance both within and between species allowed the analysis of the selective forces that shape regulatory divergence at the level of individual genes (Rifkin et al. (2003)). Purifying selection emerged as the main evolutionary force shaping regulatory evolution, with most genes showing little expression variation both within and between species. Many other genes, however, display high expression divergence between species but low polymorphism within *D. melanogaster*, suggesting widespread lineage-specific selection on regulatory mutations. These observations suggest that positive selection is the main driver of regulatory change, and that the contribution of genetic drift has been comparatively minor within the *melanogaster* subgroup. This is in contrast

to other studies claiming a dominant role for drift (Khaitovich et al. (2005)), and large-scale identification of causal genetic variants and of the selective forces acting on them will be necessary to rigorously address this question.

Mutation accumulation assays, which artificially minimize the effects of selection in order to reveal the full spectrum of possible mutations, demonstrated great potential for rapid regulatory divergence in *Drosophila* (Rifkin et al. (2005)). The genetic architecture of gene regulation appears to be such that mutations with an influence on gene expression are frequent, and their effect sizes are often moderate to large. This is in contrast with the much lower rate of gene expression divergence observed between species, which confirms that the expression of most genes is under purifying selection. It cannot be ruled out, however, that a proportion of such mutations may normally be selected out for reasons that are independent of their effects on gene expression.

It has also been found that individual lineages progressively accumulate compensatory mutations in *cis* and *trans* that affect the expression of individual genes in opposite ways (Landry et al. (2005), McManus et al. (2010)). This is reminiscent of the patterns of sequence divergence at the *eve* stripe 2 enhancer, and suggests a similar combination of small-effect mutations and stabilizing selection. Interestingly, the existence of such *cis-trans* compensatory pairs was found to underlie the severe dysregulation of certain genes in interspecific hybrids (Landry et al. (2005), Meiklejohn et al. (2013)). It is thought that this phenomenon could in certain cases form the basis for reproductive isolation between populations. The high prevalence of genetic incompatibilities segregating in *D. melanogaster* populations raises the possibility that regulatory mutations such as those may perhaps play a role in speciation (Corbett-Detig et al. (2013)).

Importantly, the developmental context of gene expression shapes natural variation. For instance, there are stronger evolutionary constraints on phases of upregulation of individual genes than on phases of downregulation. A recent study of embryonic gene expression between several *Drosophila* species revealed a pattern of divergence consistent with the classical hourglass model (Kalinka et al. (2010)). The degree of interspecific divergence varies between developmental stages and it is minimized at mid-embryogenesis, at the phylotypic stage – the stage at which, based on morphological criteria, the embryos of different species in the clade are most similar.

Beyond these phenomenological studies, the focus has shifted more recently to the identification of molecular mechanisms underlying regulatory divergence. Comparative studies of transcription

factor binding have revealed a rate of divergence much slower than that observed in mammals (Moses et al. (2006), He et al. (2011)). This is likely to be attributable in part to the much smaller effective population sizes of mammalian species, which limits the efficacy of selection. These observations suggest a dominant role for purifying selection in shaping TFBS evolution in *Drosophila*. There is, however, still a dynamic turnover of binding sites. The adaptive value of lineage-specific sites, and the functional relationship of TFBS turnover to the evolution of gene expression, remained to be determined.

We are still in the early days of comparative functional studies, and there remain many open questions. Although the evolution of transcription factor binding sites is beginning to be investigated, we still know little about the types of mutations and the selective forces that underlie their evolution. And importantly, we know nothing about the evolutionary dynamics of other types of regulatory elements. Crucially, comparative surveys of transcription have so far focused exclusively on protein-coding genes. With the growing recognition that a large fraction of developmental transcriptomes consists of non-coding RNAs, and given the difficulties in assessing their functional conservation from sequence evolution patterns, there is a dire need of investigating their conservation directly with functional assays.

1.5 Final remarks

Much evidence has accumulated over the years that regulatory changes play a major role in the evolution of development and morphology. Detailed studies of individual phenotypes have identified regulatory mutations as the cause of numerous developmental changes, in a variety of metazoans. These examples have started illuminating the molecular and developmental mechanisms through which regulatory variation affects organismal phenotypes. Importantly, many of these changes have been shown to be adaptive and to be driven by natural selection.

Little is known, however, about the mutations that drive regulatory evolution, the genes they are most likely to effectively act on, or their precise effects on gene expression and organism fitness. Genome-wide comparative studies of regulatory mechanisms at the molecular level hold the potential to reveal key aspects of the processes through which genetic mutations drive changes in gene expression and phenotypes. The particular types of mutations that drive regulatory evolution are diverse, and they include exotic mechanisms, such as the transposition of mobile genetic elements, which can have immediate strong effects and seamlessly implement similar complex changes at multiple loci.

In recent years, comparative functional work has been undertaken to analyze the evolution of gene expression patterns, as well as the evolutionary dynamics of individual regulatory elements such as transcription factor binding sites. Such a focus on well-defined genetic elements is crucial, as it allows a far more detailed and thorough assessment of the selective forces at play. Indeed, genomic sequences are the level at which we have the most powerful tools to dissect the architecture of evolutionary transitions, for both theoretical and technical reasons. They are also the most fundamental level at which, ultimately, we need to be able to understand them.

More diverse types of genomic elements need to be investigated, however, for a clearer and more complete understanding to emerge. Promoters are of particular interest, as they constitute the platform for the integration of inputs from various regulatory elements. Importantly, it is also possible to directly measure the effects of promoter mutations on gene expression, by monitoring either transcriptional activity or, as a proxy, steady-state transcript abundance. By jointly analyzing patterns of sequence variation and transcription, we can gain insights into the relationship between the genotype and the intermediate phenotype of transcriptional output.

The genes involved in orchestrating development and driving its evolution also constitute the focal point of a growing area of research, as their functional diversity seems far greater than anticipated. Despite great progress over the last three decades, we still have a very fragmentary view of the genetic basis of developmental control, and of the architecture of gene regulatory networks. With the discovery of large numbers of non-coding transcriptional units and the characterization of the molecular functions of many non-coding RNAs, it has also become clear that such atypical genes need to be integrated to our understanding of many biological processes. Their potential involvement in various aspects of development, as well as the molecular mechanisms through which they might act, will need to be carefully studied. Functional non-coding transcription units were hidden gems of the *Drosophila Bithorax* locus and mammalian *Hox* clusters, and others are likely hiding in plain sight. Their influence – if any – on developmental evolution warrants rigorous assessment as well, in particular because their evolutionary dynamics seem very different from those of protein-coding genes.

The work I am presenting here focuses on the evolution of transcriptional promoters in *Drosophila*. Specifically, it addresses the question of the influence of transposable elements on transcriptome evolution, and investigates some aspects of long-standing hypotheses regarding transposon co-option. Through a direct analysis of the functional conservation of promoter activity, it also explores the relevance of

non-coding transcription to embryonic development.

My results will be divided into three Chapters. A careful analysis of promoter activity in a developmental context required the development of adequate experimental and analytical tools, and Chapter 2 focuses on the development of RAMPAGE, a high-performance method for TSS identification and promoter activity profiling. Both experimental and computational aspects of the method will be described, along with analyses validating the approach. Chapter 3 will be dedicated to the characterization of genome-wide promoter activity profiles throughout the life cycle of *D. melanogaster*, and to the analysis of the role of transposons in the control of developmental gene regulation. Transposon-derived promoters were found to drive the expression of hundreds of protein-coding and non-coding transcripts, and there is evidence supporting the hypothesis that transposons can indeed distribute stereotyped regulatory elements throughout *Drosophila* genomes. Finally, Chapter 4 will describe the results of a comparative study of promoter activity throughout the embryonic development of 5 *Drosophila* species. This analysis uncovered dynamic patterns of promoter gain and loss throughout the clade, and confirmed a role for transposons in fostering regulatory innovation in individual lineages. Strikingly, we discovered thousands of promoters driving the expression of putative non-coding transcripts, and there is strong evidence that many of them are under purifying selection, implying that they have conserved functions in embryonic development.

1.6 Acknowledgements

There are many people to thank for making these years in Cold Spring Harbor as exceptional as they have been, both scientifically and personally. I would like to start by expressing my gratitute towards Tom, for always being willing to give advice, and yet giving me as much freedom as he did to pursue my interests. This made my years in the lab both a thrilling time and a great learning experience. I am also grateful to Felix Schlesinger and Alex Dobin for making every work day a little bit more fun with often stimulating, if occasionally erratic, discussions, and for teaching me just about everything I can do with a computer beyong sending email. And to Sudipto, for turning that lonely lab room of ours into a more lively place, and for the good company on these late nights of spinning and pipetting in the colonies. Thanks also to everyone else in the lab, past and present, for making it easy and enjoyable to work there.

I am deeply grateful to Alex Gann, outstanding mentor formally and not, guide and friend, and voice of reason even as he might be aiming for outrage... I had never heard so many bad jokes about the French, but I even almost liked those.

I am forever indebted to the school, for bringing together so many engaging and stimulating instructors, and running such a great program for us. Said instructors also deserve full credit for making these courses as fun and challenging as they were. Many thanks to our dual Deans, Leemor Joshua-Tor and Alex Gann, for their leadership. I also have a special thought for the WSBS staff, for always making life so much easier for us, even in those particular instances when we were not making it much easier for them... Sometimes I fear that the only thing they will remember of me is my very unique perception of time, but then again, I did do that to myself.

I would like to thank my thoroughly helpful and ever-encouraging thesis committee, Greg Hannon, Alex Gann, Josh Dubnau, Zach Lippman and Adrian Krainer, for their advice and support throughout my project. These meetings were both productive and stimulating, and they really made a difference. I am also very grateful to Brenton Graveley for accepting to serve on my thesis jury.

I would also like to take this opportunity to thank the Florence Gould Foundation, for their financial support throughout these years.

On a more personal note, I would like to thank Emily for all her affection, her company and her support, as well as her merciless proofreading. Just for the record, writing a thesis will never be quite

the same as writing a thesis while eating cornbread. Just sayin'.

To my comrades of WSBS vintage 2008, thank you so much for all the good fun and the good spirit of our early days here, when our instructors were teaching us a thing or two about our new home... Among those, the fact that when folks around here talk about the "work hard, play hard" culture at CSHL, they really do mean the first bit. A special thought for Dario Bressan, without whom this thesis might not have materialized without a fiercer struggle.

My thoughts go out to the great friends I have or had here, they know who they are, be they the Brooklyn crew or the others.

Finally, last but certainly not least, infinite thanks to my family, for letting me pursue whatever terrible career path I feel passionate about, and for their unwavering support.

Chapter 2

Promoter activity profiling by paired-end sequencing of 5'-complete complementary DNAs

This chapter was originally published in *Genome Research* under the title:
"High-fidelity promoter profiling reveals widespread alternative promoter usage and transposon-driven developmental gene expression "

Many eukaryotic genes possess multiple alternative promoters with distinct expression specificities. Therefore, comprehensively annotating promoters and deciphering their individual regulatory dynamics is critical for gene expression profiling applications, and for our understanding of regulatory complexity. To achieve this, we have developed RAMPAGE, a novel promoter activity profiling approach that combines extremely specific 5'-complete cDNA sequencing with an integrated data analysis workflow to address the limitations of current techniques. RAMPAGE features a streamlined protocol for fast and easy generation of highly multiplexed sequencing libraries, offers very high transcription start site specificity, generates accurate and reproducible promoter expression measurements, and yields extensive transcript connectivity information through paired-end cDNA sequencing.

2.1 Introduction

In recent years, a large body of work has been uncovering the complexities of transcriptional regulation in eukaryotes. The landscapes of transcription, surveyed with ever-increasing scrutiny, reveal intricate genetic architectures from which originate myriads of protein-coding and non-coding transcripts (Kapranov et al. (2007), Djebali et al. (2012)). The regulatory blueprints that orchestrate the spatiotemporal dynamics of eukaryotic transcriptomes mirror this complexity. Large-scale surveys of chromatin modifications and transcription factor occupancy in diverse organisms have started to shed light on the abundance of *cis*-regulatory modules (Bernstein et al. (2012), Ernst et al. (2011), Negre et al. (2011), Shen et al. (2012)), their relevance to development and disease (Bernstein et al. (2012), Lindblad-Toh et al. (2011)), and the structure of the gene regulatory networks they implement (Marbach et al. (2012), Suzuki et al. (2009)). Additionally, genome-wide studies of transcription start site (TSS) usage have shown that many genes possess alternative promoters, highlighting the importance of their contribution to the diversity of gene expression patterns (Carninci et al. (2006), Suzuki et al. (2009)). TSSs are of particular interest, because in addition to harboring many transcription factor binding sites, the promoters they are embedded in constitute the platforms where the transcriptional machinery integrates the inputs from cognate cis-regulatory elements. They are also worthy of attention from an experimental standpoint, since the quantification of transcripts coming from individual TSSs allows for precise measurements of the final output of these molecular computations.

The explosion of experimental and computational approaches in functional genomics that accompanied the advent of second-generation sequencing has been a major driving force behind our progress in uncovering and understanding this regulatory complexity. For the study of TSS location and activity, however, even state-of-the-art, high-resolution techniques based on 5'-complete cDNA sequencing (Kodzius et al. (2006), Ni et al. (2010), Plessy et al. (2010)) are currently lacking in multiple aspects. Here we address these issues and present RAMPAGE (RNA Annotation and Mapping of Promoters for the Analysis of Gene Expression), a very accurate 5'-complete cDNA sequencing approach that allows for the *ab initio* identification of TSSs at base-pair resolution, the quantification of their expression and the characterization of their transcripts. We engineered our protocol to take full advantage of the paired-end sequencing capabilities of current high-throughput platforms, thus yielding crucial transcript connectivity information. Importantly, this feature allows us to rigorously connect TSSs to

the genes they drive the expression of based on direct cDNA evidence. Our method also provides much higher specificity for TSSs than current approaches, and we developed a streamlined two-day protocol that allows the barcoding and pooling of multiple samples after the very first step, thus greatly facilitating library multiplexing and preparation. For the analysis of this data, we have developed an integrated analysis pipeline that relies on the unique features of the data to maximize TSS specificity, transcript connectivity information recovery, and quantification accuracy. At the core of this pipeline lies a novel peak-calling algorithm for TSS discovery that was specifically tailored to filter out multiple types of noise (i.e., random distortions of the underlying signal by technical factors) associated with 5'-complete cDNA sequencing. Here I give an overview of the method and present an evaluation of its performance; a detailed protocol and additional technical considerations can be found in Appendix 1.

2.2 Results

RAMPAGE: Multiplexed paired-end sequencing of 5'-complete cDNAs

5'-complete cDNA sequencing has proven to be a challenging task, despite significant contributions over the years from several approaches that have relied on diverse strategies. CAGE (Kodzius et al. (2006)) is based on the biotinylation of the 7-methylguanosine cap of RNA Polymerase II (Pol II) transcripts and pulldown of the 5'-complete cDNAs they are hybridized to, a technique known as "cap-trapping" (Carninci et al. (1996)). CAGEscan (Plessy et al. (2010)) and other approaches (Islam et al. (2011)) exploit some unique features of reverse-transcriptase enzymes to add adaptors to the end of 5'-complete first-strand cDNAs during the reverse-transcription step, in a process dubbed "template-switching" (Hirzmann et al. (1993)). PEAT (Ni et al. (2010)) and similar techniques rely on the ligation of an RNA adaptor to the 5' end of capped transcripts ("oligo-capping"), similarly to conventional 5'-RACE.

CAGE relies on a protocol that, although scalable, is cumbersome, and requires input amounts on the order of 50 μg of total RNA. Its main limitation is the impossibility to sequence more than short 5' tags (~27 bases) from the cDNAs, which makes unambiguous read mapping impossible for large parts of eukaryotic genomes, precludes evidence-driven assignment of novel TSSs to gene annotations, and yields no transcript structure information. This has been a major impediment to the analysis of novel TSSs in general, and of repeat-borne TSSs in particular. The specificity of CAGE for TSSs is also

currently limited (please see "Assessment of assay performance" below). CAGEscan does allow paired-end sequencing of cDNA inserts, but with lower TSS specificity (Plessy et al. (2010)). PEAT also allows for paired-end sequencing, although only 20 bp can be sequenced from each end due to the cloning procedure used, but this is again at the expense of specificity. Moreover, adaptor ligation is mediated by T4 RNA ligase 1, an enzyme known to have strong sequence biases (Zhenodarova et al. (1989)), which is detrimental to accurate transcript representation. Finally, this complex protocol requires large quantities of starting material (~150 μg total RNA), which is impractical for most samples.

To address these challenges, we developed RAMPAGE by modifying and combining the two orthogonal 5'-selection approaches of template-switching and cap-trapping (see Appendix 1 and Figure 2.1). In comparison to current approaches, RAMPAGE has the key advantage of yielding long paired reads as opposed to short sequence tags, while also offering greatly improved specificity for TSSs. Library preparation and multiplexing is greatly facilitated by the fact that individual samples are barcoded and pooled after the very first step of the protocol, allowing almost the entire workflow to be carried out on a single library (see Appendix 1). Additionally, all steps from the biological sample to the pooled cDNA products can be carried out in 96-well plates, and our full workflow from RNA to library can be completed in 2 days, making library preparation simple and very scalable. Input material requirements are on the order of 10 to 20-fold lower than for conventional CAGE.

Computational analysis of RAMPAGE data

We designed an integrated computational strategy that makes extensive use of the unique features of the data to enhance the accuracy and quality of our analysis. All analysis steps from raw sequencing data to TSS clusters, expression level estimates and partial transcript models can be performed in a single process for a set of samples. The complete analysis workflow is summarized in the Methods section and Appendix 1.

The cornerstone of this pipeline is a novel peak-calling algorithm for TSS discovery that implements several noise-filtering strategies to greatly improve our ability to discriminate between true TSSs and background signal. As in other high-throughput assays, robustly detectable signal must be distinguished from a background that may have multiple possible origins. Additionally, most eukaryotic promoters do not use a single position as their TSS, but allow transcription initiation at several positions. The shapes of these TSS clusters (TSCs) vary between promoters, from sharp (a few nu-

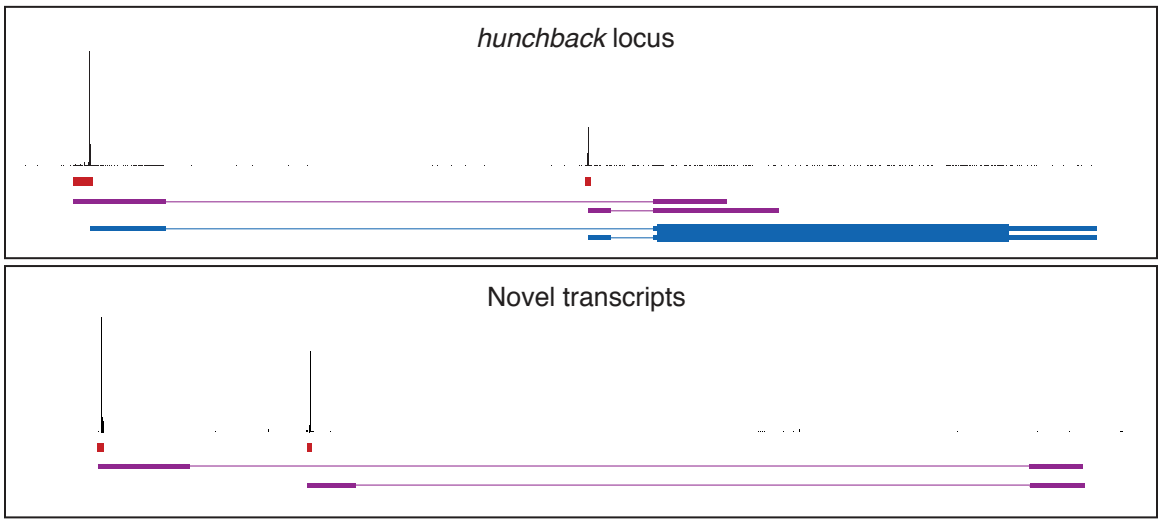

Figure 2.1. Overview of RAMPAGE data
Graphical representation of the data at the *hunchback* gene locus, and at an unannotated locus harboring novel transcripts. For each panel, the top track shows the density of cDNA 5' ends per position on the upper strand, which can be interpreted as a single base-resolution profile of transcription initiation activity. The second track represents the peaks (i.e., TSS clusters) called from that density profile. The third track shows the partial transcript models reconstructed *ab initio* from our sequencing data using Cufflinks. For the upper panel, the fourth track displays Flybase transcript annotations. For the second panel, note that paired-end information allows one to infer a functional link between the two promoters, which appear to be alternative promoters for a common locus.

cleotides) to broad (at least 100 nucleotides) (Carninci et al. (2006)). Therefore, previous analyses of 5'-complete cDNA sequencing data have usually made use of some strategy to group individual TSSs into clusters (Carninci et al. (2006), Ni et al. (2010), Plessy et al. (2010)). Building upon this work, we devised a novel approach to identify TSCs, which we define operationally as regions of statistically significant clustering of RAMPAGE 5' end tags. Critically, our peak-calling algorithm was designed to make extensive use of paired-end information and to correct for several sources of noise inherent to 5'-complete cDNA sequencing.

Firstly, we expect the background distribution of signal per genomic position to be overdispersed due to at least two technical factors: failures of reverse-transcriptase to reach the 5' end of its template are expected to be more likely at specific sites of a given transcript (e.g., strong secondary structures), and PCR duplicates in the libraries can randomly amplify the signal at individual positions. Both effects will lead to the data looking more "peaky" than the actual landscape of transcription initiation is. To attenuate these effects, we make use of an overdispersed distribution (negative binomial) to model background signal, and we remove PCR duplicates from our datasets prior to peak-calling. For our purposes, we define PCR duplicates as read pairs that share similar alignment coordinates (start, end, splice sites) and an identical reverse-transcription primer sequence (which we use as a pseudo-random single-molecule barcode).

Secondly, non-specific signal coming from non-5'-complete cDNAs represents another source of background, which is complex because the amount of non-specific signal depends on transcript abundance. In the absence of an appropriate correction, this will lead to highly expressed transcripts contributing many false-positive TSCs. To limit this effect, other authors have made use of independent RNA-seq data to filter CAGE signal (Hoskins et al. (2011)), but this approach requires the generation of additional datasets for all samples under study. Harnessing paired-end information, we make use of the fact that coverage by downstream sequencing reads (i.e., the 3'-most portion of our cDNAs) can provide us with an estimate of transcript abundance at internal (non-TSS) positions. We model background from incomplete cDNAs as linearly proportional to transcript abundance as measured by downstream read coverage, and show this approach to greatly improve our ability to distinguish between true TSSs and spurious internal signal (Figure 2.2).

These features were incorporated into a sliding window algorithm that scans the genome and assesses the significance of local signal enrichment given the null distribution. Downstream read cov-

erage in the same window is used to correct for local transcript abundance, by subtracting from the raw signal a pseudocount proportional to this coverage. After false discovery rate (FDR) correction by the Benjamini-Hochberg method, enriched windows in close proximity to each other are merged into peaks, and those are subsequently trimmed at the edges down to the first base with signal.

Our data yields rich information about transcript structure and connectivity, which allows us to connect these TSCs to annotated genes based on rigorous cDNA evidence. This is an extremely important feature, since complex transcriptional architectures (Djebali et al. (2012), Kapranov et al. (2007)) make the promoter-transcript relationships at many loci otherwise difficult to decipher. Additionally, we take advantage of the fact that the downstream portions of the inserts are distributed over broad regions of the targets to gain knowledge about medium-range transcript connectivity. In the current implementation, reads from individual TSCs are processed through Cufflinks to produce partial transcript models.

Assessment of assay performance

The combination of template-switching and cap-trapping yields libraries that are highly enriched for 5'-complete cDNAs, as can be judged from the distribution of raw signal over annotated transcripts (Figure 2.2). For individual transcript annotations, we estimate that the median proportion of 5' tags in TSS regions is over 90% (Figure 2.3). Comparisons to similar D. melanogaster data generated by CAGE or PEAT revealed a dramatic improvement in specificity over these previous methods (Figure 2.3). In turn, the peak calls are themselves extremely highly enriched over annotated TSSs (Figure 2.2). Analysis of ChIP-seq histone modification profiles confirmed that the vast majority of peaks display chromatin features characteristic of TSSs (data not shown). The downstream read transcript abundance correction proves to be very effective at filtering out spurious signal in internal regions of transcripts, while having a very limited effect on sensitivity for annotated TSSs (Figure 2.2).

In terms of topological resolution, extensive comparisons on equivalent samples with a large RNA ligase-mediated 5'-RACE (RLM-RACE) dataset (Hoskins et al. (2011)) show very strong agreement between the two techniques (Figure 2.4). This demonstrates that RAMPAGE achieves single-base topological resolution in TSS detection, which has previously not been possible with CAGE (Hoskins et al. (2011)).

Gene expression quantification accuracy was benchmarked against standard shotgun RNA-seq

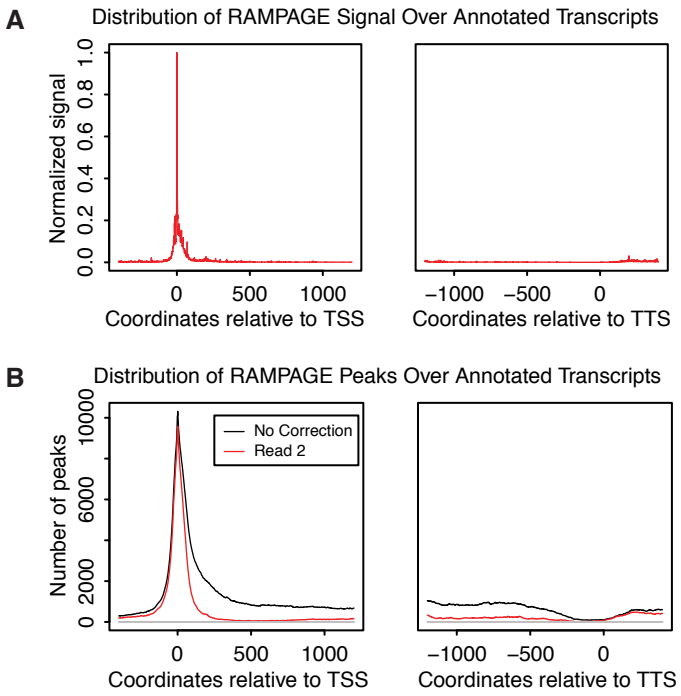

Figure 2.2. Distribution of raw signal and peaks over transcript annotations
(A) Metaprofile of signal density over all Flybase r5.32 transcript annotations. TSS: Transcription start site; TTS: Transcription Termination Site. (B) Metaprofile of peak density over annotated transcripts. (Red curve: downstream read coverage correction; black curve: no correction; all other peakcalling parameters were kept identical)

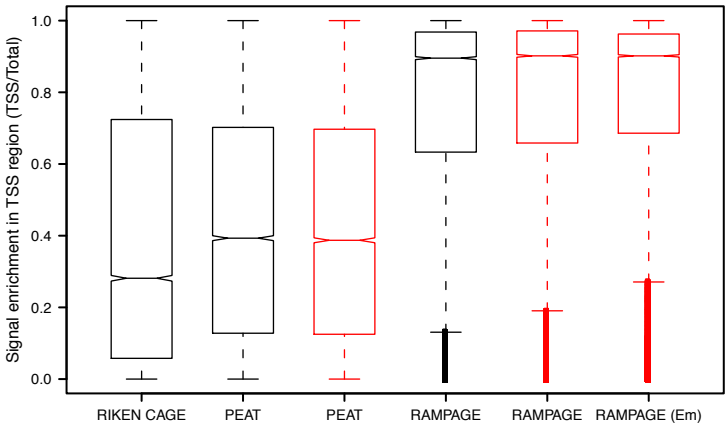

Figure 2.3. Signal enrichment at annotated TSSs
In order to compare 5' enrichment performance, we defined for each annotated transcript our 5' enrichment metric as the number of tags in the TSS region (annotated TSS +/- 150 bp) divided by the total number of tags in the transcript (TSS region or rest of the transcript). For each dataset, this ratio was computed for all Flybase r5.32 transcripts having at least 10 tags overall, and we plotted the distribution of ratios. We compared the RIKEN institute CAGE modENCODE dataset (mixed-stage embryos, Hoskins et al. (2011)), the published PEAT dataset (mixed-stage embryos, Ni et al. (2010)) and our adult female flies dataset. All datasets were brought to the same size by subsampling reads. Black boxplots: for all datasets, we considered only the first 20 templated bases of read 1 and discarded any other data. Reads were mapped using the same parameters. This is intended as a fair comparison of protocol performance, independent of sequencing data type. Red boxplots: the full data (paired-end) was used. RAMPAGE (Em) plot: to control for sample type effects (embryos vs. adults), we ran the same analysis on our embryos data (all 24 full datasets pooled).

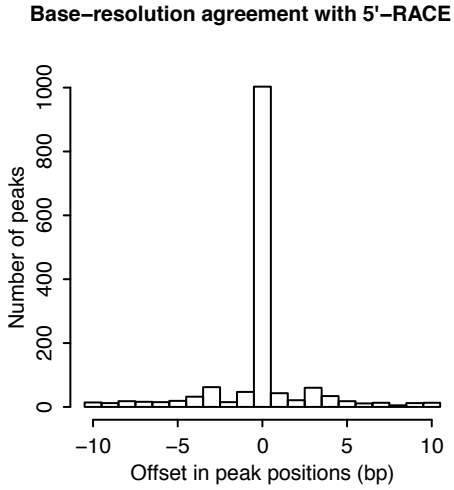

Figure 2.4. Topological agreement between RAMPAGE and 5'-RACE
Histogram of the cross-correlation of TSS cluster positioning by RLM-RACE (Hoskins et al. (2011)) and by our method. For each cluster, we determined the positional offset (in base pairs) that maximizes the cross-correlation between the data from the two methods.

data from adult male and female D. melanogaster that was generated by the modENCODE consortium (Graveley et al. (2010)). This comparison showed good agreement between the techniques for absolute quantification (data not shown), and excellent agreement for relative quantification (Figure 2.5). Expression level estimates are very reproducible, even between full biological replicates (Figure 2.5).

2.3 Discussion

We have developed and validated a method for high-throughput, high-quality discovery of TSSs, the characterization of the transcripts that emanate from them, and the quantification of their expression. We propose this approach, which directly delineates promoter-specific expression and offers a simple workflow and optimized sample multiplexing, as an advantageous alternative to standard RNA-seq for many gene expression profiling applications. Importantly, this library preparation method will also be easily portable to other sequencing platforms with minimal alterations. This is particularly attractive as new technologies yielding greater read lengths are now allowing us to move towards large-scale full-length cDNA sequencing. Preliminary work conducted towards the generation of full-length cDNA libraries using a modified version of the protocol and their sequencing on the Pacific Biosciences single-molecule platform is described in Appendix 2.

2.4 Methods

Fly stocks and sample collections:
Stocks of the *y; cn bw sp* strain were maintained in standard cornmeal medium bottles in a 24ºC incubator. 0- to 12-hour-old flies were sorted by sex and kept in vials with cornmeal medium for 5 days, then snap-frozen.

RNA extraction:
Total RNA was extracted from adult flies using Trizol (Invitrogen) according to the manufacturer's instructions and treated with DNAseI (Roche). For the human K562 cell line, RNA was extracted using Trizol according to the manufacturer's instructions and treated with DNAseI (Roche). We systematically confirmed on a Bioanalyzer (Agilent) that the RNA was of very high quality. 5'monophosphate

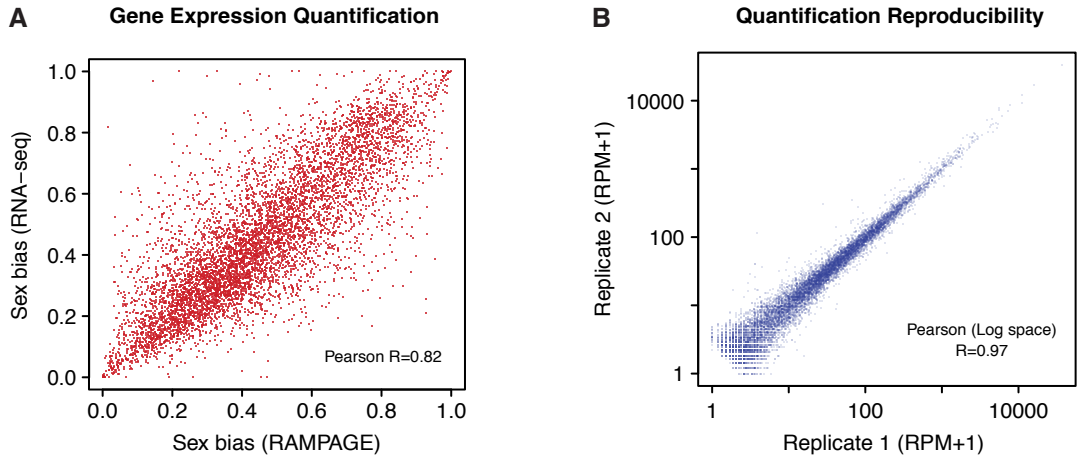

Figure 2.5. Relative transcript quantification with RAMPAGE
(A) Comparison of RAMPAGE and standard RNA-seq (Graveley et al. (2010)) performance for relative quantification of gene expression. We compared the measures of sex bias in the expression of genes obtained by the two methods. (B) Reproducibility of expression level measurements between biological replicates.

species – mainly ribosomes – were depleted by TEX digest (see Appendix 1).

Library preparation & sequencing:

Libraries were prepared from adult male and female total RNA. Reverse-transcription was run in parallel with different barcoded oligos, and the samples were pooled right after reverse-transcription. Libraries from biological replicates were prepared independently. The 5'-complete cDNA selection strategy relies on the combination of two orthogonal enrichment methods: reverse-transcriptase template-switching, and cap-trapping. The template-switching approach is based on the ability of reverse-transcriptase to add linker sequences to the ends of 5'-complete cDNAs – preferentially if they are made from capped transcripts (Fig. S1). Cap-trapping relies on the biotinylation of capped RNA molecules and specific pulldown of their associated 5'-complete cDNAs. The libraries were run on a DNA HS Bioanalyzer chip for quality control, quantified by quantitative PCR, and sequenced on one lane each on an Illumina GAIIx sequencer (2x76bp). See Appendix 1 for detailed experimental procedures.

Sequencing reads alignment:

The sequences corresponding to the library identification barcode and the reverse-transcription primer were trimmed prior to mapping. Trimmed reads were mapped with STAR, with parameters described in Tables S2-3. All uniquely mapping reads were kept. As a rescue strategy for multiply mapping reads, if all alignments for those reads started within an annotated transposon and overlapped the same gene annotation, the alignment starting in the closest transposon insertion was selected. All non-rescued multi-mappers were discarded.

Data analysis pipeline:

PCR duplicates, defined as reads sharing the same alignment coordinates (start, end and splice sites) were removed from the individual datasets. To avoid over-collapsing, we took advantage of the fact that the long random sequence (15-mer) of our reverse-transcription primer often primes with mismatches. We used this sequence as a pseudo-random barcode allowing us to distinguish between true duplicates (same barcode) and independent identical inserts. All collapsed datasets were then combined prior to peak-calling. The density of cDNA 5' ends across the genome was determined from this combined dataset, as well as the density of coverage by second (i.e., downstream) sequencing reads. Peaks were

called by a sliding window algorithm that assesses the significance of local signal enrichment given a null distribution. Downstream read coverage in the same window was used to correct for local transcript abundance, by subtracting from the raw signal a pseudocount proportional to this coverage. After FDR correction, significant windows in close proximity to each other were merged into peaks, and those were trimmed at the edges down to the first base with signal. (Parameters: window width 15 bases, null distribution negative binomial with k=1, background weight 0.8, FDR 0.001%, merging range 150 bases). These peaks were connected to annotated genes based on cDNA structure information. For each peak, if we could find at least 2 inserts having their 5' in the peak and overlapping an annotated exon of a gene, the peak was functionally linked to that gene. If a peak could potentially be linked to several genes, ties were broken by removing all links that were 5-fold weaker than the strongest one. For quantification, the signal for each peak and each timepoint was derived from the uncollapsed datasets, and normalized to dataset size (defined as the total number of reads attributed to any genic TSS). We built partial transcript models by running Cufflinks separately on the set of reads coming from each peak for each given dataset, and collapsing all transcripts for each peak using Cuffmerge. This pipeline was implemented with scripts written in Python, including Scipy and Numpy. BEDtools (Quinlan and Hall 2010) v2.11.2 and Cufflinks (Trapnell et al. 2010) v1.0.3 were used for some analyses, and plotting was done in R. See Appendix 1 for details of the analysis.

Comparison with 5'-RACE:

Our adult flies RAMPAGE data (replicate 2, sexes pooled) was compared to the modENCODE adult flies 5'-RACE dataset (see Supplementary Information). For each RAMPAGE peak that was $\leq$500 bases wide and for which there were at least 5 tags in each dataset (exactly in the peak for RAMPAGE, in the peak +/- 10 bases for RACE), we determined the positional offset that maximizes the cross-correlation between the 2 signals.

Comparison with RNA-seq:

The modENCODE 5-day-old adult flies RNA-seq data (Graveley et al. 2010) was mapped with STAR, and the expression of annotated genes was quantified using Cufflinks. Sex bias = Male expression

(RPM) / (Male expression + Female expression).

Reproducibility between biological replicates:

RAMPAGE libraries were generated for two independent batches of adult *D. melanogaster* females, and sequenced on separate flowcells on Illumina GAIIx sequencers (8.3M and 16.7M million reads). The second dataset was randomly subsampled to match the size of the first one. Both datasets were mapped in parallel with the same parameters, duplicates were collapsed, and the datasets were pooled prior to peak-calling. Expression values for this common set of intervals were derived from each uncollapsed dataset.

Chapter 3

High-fidelity promoter profiling reveals widespread alternative promoter usage and transposon-driven developmental gene expression

This chapter was originally published in *Genome Research* under the title:
"High-fidelity promoter profiling reveals widespread alternative promoter usage and transposon-driven developmental gene expression "

The landscape of promoters and their regulation throughout development are poorly characterized in most eukaryotic genomes. To gain a better understanding of genome organization and expression, we developed RAMPAGE, a method for transcription start site discovery and transcript quantification based on 5'-complete cDNA sequencing. We used this approach in a genome-wide study of promoter activity throughout 36 stages of the life cycle of *Drosophila melanogaster*, and describe here a comprehensive dataset that represents the first available developmental time course of promoter usage. We found that over 40% of developmentally expressed genes have at least 2 promoters, and that alternative promoters generally implement distinct regulatory programs. Transposable elements, long proposed to play a central role in the evolution of their host genomes through their ability to regulate gene expression, contribute at least 1,300 promoters shaping the developmental transcriptome of *D. melanogaster*. Hundreds of these promoters drive the expression of annotated genes, and transposons often impart their own expression specificity upon the genes they regulate. These observations provide support for the theory that transposons may drive regulatory innovation through the distribution of stereotyped cis-regulatory modules throughout their host genomes.

3.1 Introduction

Recent large-scale studies of RNA Polymerase II (Pol II) initiation sites in eukaryotes have identified unexpectedly large numbers of promoters. These observations have raised questions regarding the biological relevance of these promoters, and their regulation throughout development and in response to physiological clues. Even in well-characterized model organisms such as *D. melanogaster*, there are only sparse high-quality annotations of promoters, and no genome-wide studies of their developmental expression profiles have been conducted.

Using a high-throughput, high-fidelity approach, we set out to profile promoter activity genome-wide throughout the life cycle of *D. melanogaster*, so as to have a complete view of the transcriptional landscape and of the diversity of expression patterns in this model organism. This rich dataset reveals that over 40% of all genes are expressed from at least two promoters, underscoring the pervasiveness of this phenomenon in *Drosophila*. Importantly, we found that alternative promoters generally have uncorrelated developmental expression patterns, which reveals that they most often implement independent regulatory programs. These observations suggest that the emergence of alternative promoters has been a major driving force underlying the evolutionary diversification of gene expression programs. Our analyses also uncovered a widespread role for transposons in the developmental regulation of transcription, with ~1,300 transposon-embedded promoters driving developmentally regulated expression of diverse sets of transcripts.

Transposable elements have been shown to influence gene expression in a variety of organisms, including plants (Lippman et al. (2004), McClintock (1956), Naito et al. (2009)), *Drosophila* (Lipatov et al. (2005), Rouget et al. (2010)) and mammals (Bejerano et al. (2006), Nigumann et al. (2002)). This regulatory potential, together with the ability of transposons to disseminate stereotyped sequence modules throughout their host genomes, has led to the proposal that transposon expansion and domestication may be a powerful force underlying the assembly of complex regulatory networks (Britten and Davidson (1969), Feschotte (2008)), in particular by providing promoters for host genes (Faulkner et al. (2009), Nigumann et al. (2002), van de Lagemaat et al. (2003)). Their potential contribution to developmental gene expression, however, is currently only supported by modest evidence in mammals (Cohen et al. (2009), Macfarlan et al. (2012), Peaston et al. (2004)) and has been reported to be extremely rare in *Drosophila* (Lipatov et al. (2005)). Furthermore, it is unclear whether transposons actually distribute

promoters with stereotyped regulatory logics through a copy-and-paste mechanism.

We found that transposons from diverse classes have been co-opted to drive the expression of hundreds of annotated genes. Many of these transposons appear to have conferred their intrinsic regulatory specificity to the genes they drive, which demonstrates that they do distribute pre-programmed regulatory modules to multiple loci. A case study of *roo* element LTRs uncovered the existence of a core promoter and of a complex set of transcription factor binding sites that underlie these intrinsic regulatory properties.

3.2 Results

TSS discovery and expression profiling throughout the D. melanogaster *life cycle*

This methodological approach was used to study promoter activity dynamics throughout the life cycle of *D. melanogaster* (24 embryonic stages, 5 larval, 5 pupal, 2 adult). We sampled embryonic development, a period of fast transitions, at high temporal resolution (1 hour). All sequencing data were mapped to the genome with our spliced read aligner, STAR (Dobin et al. (2012)). Stringent peak-calling identified 31,080 high-confidence TSCs (versus 12,454 in the most recent global study (Hoskins et al. (2011))), 76% of which could be unambiguously assigned to 12,706 annotated genes based on cDNA structure (see Methods). The remaining 7,421 TSCs drive novel transcripts, which we partially characterize. Of the genic TSCs, as many as 39.6% are unannotated in Flybase r5.32. Our results are consistent with the known structure and expression dynamics of well-characterized developmental regulators, including the differential expression of alternative promoters, and represent to our knowledge the first genome-wide developmental time course of promoter activity (Figure 3.1).

The use of alternative promoters is very common in *D. melanogaster*, with over 40% of developmentally expressed genes having at least 2 promoters (Figure 3.2). In contrast, Flybase annotations only attribute alternative promoters to 14.8% of genes (see Methods). The discovery of so many promoters with relatively shallow sequencing of complex samples and a stringent analysis indicates that alternative promoter usage is an extremely frequent phenomenon, even in a relatively simple metazoan genome. Importantly, alternative promoters tend to drive expression in uncorrelated patterns (Figure 3.2 & Appendix 3 Figure 5.19). This shows that they generally implement distinct regulatory programs, as suggested previously (Carninci et al. (2006), Rach et al. (2009)). Further analysis of 1,295 genes that

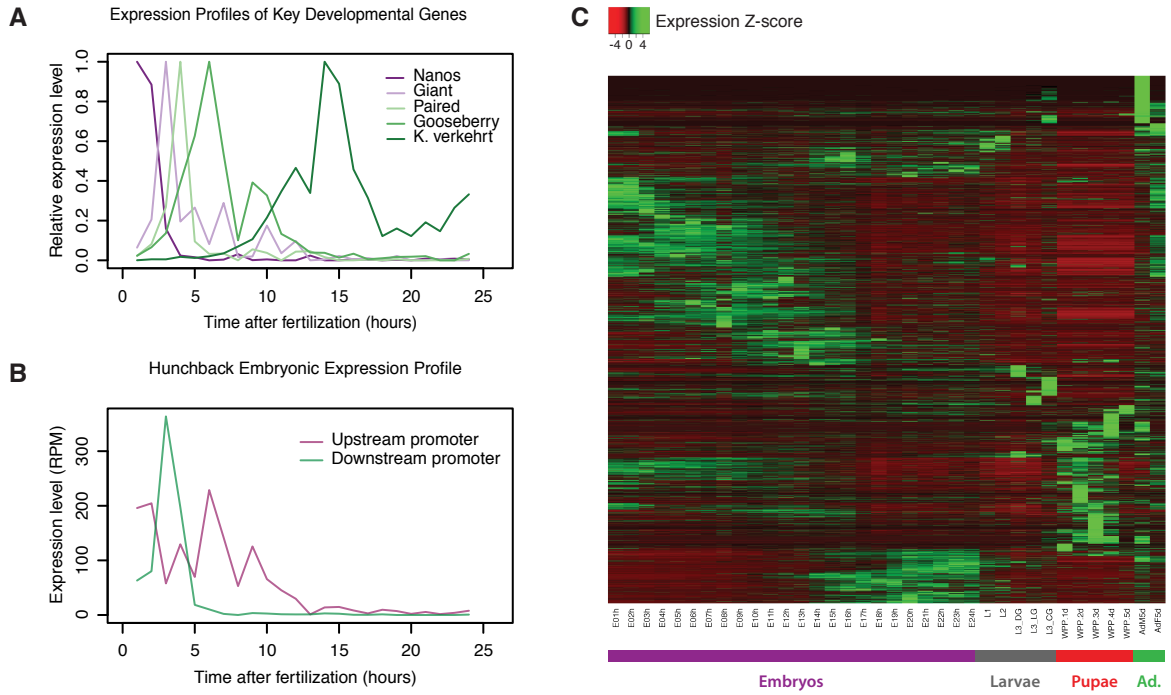

Figure 3.1. Genome-wide promoter activity dynamics
(A) Expression profiles of well-characterized key developmental genes during embryonic development. Note the sharpness of the profiles afforded by the high temporal resolution of the timecourse. K. verkhert: *Krotzkopf verkhert*. (B) Differential expression of alternative promoters (*hunchback* locus). Our data fully recapitulates the expression pattern for *hb* that has been characterized in previous work. The hb mRNAs transcribed from the upstream (maternal) promoter are predominant immediately after egg laying, and decay rapidly as the downstream promoter starts being expressed, displaying maximal expression 2-3 hours after egg laying. The upstream promoter is active again with a second peak at 5-6 hours. (C) Heatmap representing the z-score normalized expression profiles for the 24,264 promoters we could attribute to annotated genes based on cDNA structure.

undergo clear developmental transitions between alternative promoters revealed that these transitions occur in a great diversity of temporal patterns, throughout the entire life cycle (Figure 3.2).

The analysis of our high-resolution data shows that many genes undergo very fast transitions during embryonic development, their expression changes often spanning a large fraction of their dynamic range (median 60.8%) within a single hour (Figure 3.2 & Appendix 3 Figure 5.20). Some of these abrupt regulatory transitions can sometimes be of a very large magnitude on an absolute scale (Figure 3.2). Functional annotation analysis of the fastest-changing genes revealed significant enrichment for categories related to transcription factor activity, tissue morphogenesis, and cell-cell contacts (data not shown).

Role of transposons in developmental gene regulation

We set out to investigate the role of transposons in the developmental regulation of transcription. For certain time points, up to 1.6% of the transcriptome was the product of transcription initiating in TEs (Figure 3.3). Prompted by previous reports of developmental expression of transposons (Ding and Lipshitz (1994), Mozer and Benzer (1994), Parkhurst and Corces (1987)), we established expression profiles for individual subfamilies (Methods). Virtually all transposon subfamilies display clear developmental regulation (Figure 3.3), in diverse patterns. This is consistent with the view that transposons have intrinsic properties governing their own expression, as shown previously for individual cases (Bronner et al. (1995), Naito et al. (2009), Udomkit et al. (1996)). With regards to regulatory innovation, this makes transposons particularly interesting as a versatile toolkit of mobile regulatory modules with diverse properties.

To search for instances of transposons providing promoters for host genes, we mined our data for transposon-contained TSCs that drive the expression of annotated exons (Figure 3.3). We thus found 182 high-confidence TSCs derived from multiple classes of TEs (Figure 3.3) that drive the expression of 152 annotated genes. RNA ligase-mediated 5'-RACE on selected candidates validated our findings (Appendix 3 Figure 5.21). Figure (Figure 3.3) illustrates one such case, where a solo LTR from a *297* element provides an unannotated alternative promoter for the *TM4SF* gene. Their temporal patterns of expression are diverse, with subpopulations being active at any developmental stage sampled (Figure 3.3). Very importantly, the expression profiles of these transposon-derived TSCs are generally uncorrelated with the profiles of alternative promoters of the same gene (data not shown), which suggests that

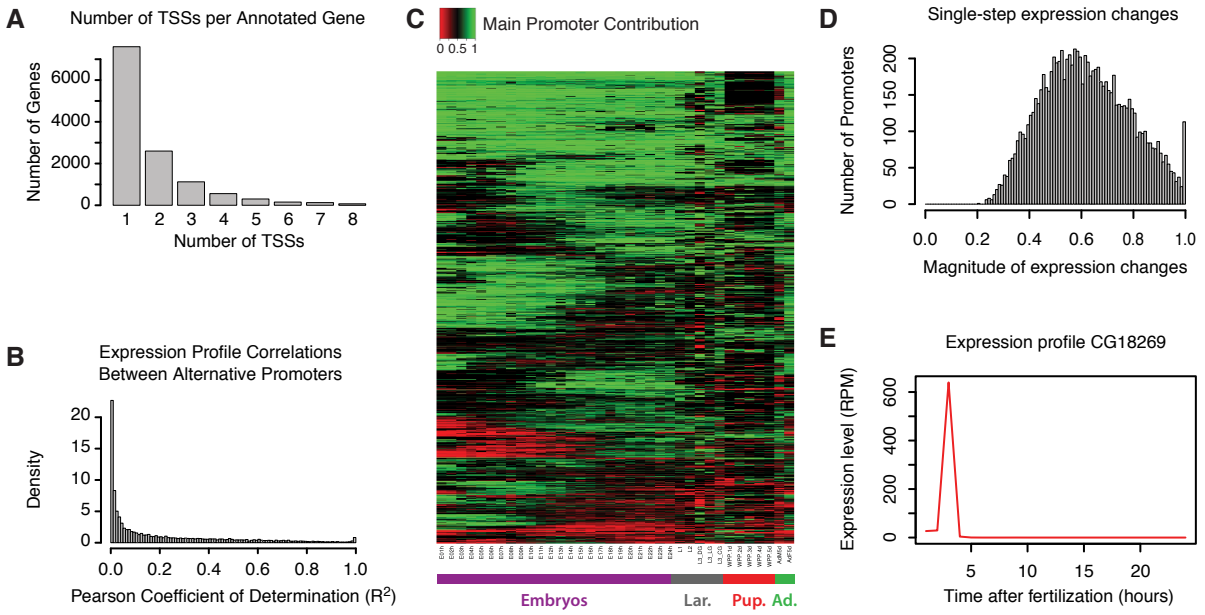

Figure 3.2. Widespread alternative promoter usage

(A) Number of TSSs detected per annotated gene. Over 40% of all expressed genes have at least 2 alternative TSSs. (A small number of genes are excluded from the graph (>10 TSSs) but these are probably affected by technical artifacts.) (B) Distribution of pairwise Pearson's coefficients of determination (R^2) between the full expression profiles (36 timepoints) of alternative promoters. This gives a measure of the similarity between the expression profiles of alternative promoters. Only TSCs with a maximum expression level at least 10 RPM were included. Note the overall absence of correlation (median coefficient: 0.108). (C) Temporal dynamics of developmental transitions between alternative promoters. The heatmap represents the fraction of total expression contributed by the main promoter at each timepoint, for 1,295 genes that display pronounced transitions between promoters (see Methods). Note the diversity in the timing of promoter transitions. (D) Maximal fraction of the dynamic range of the profile of a given TSS spanned in a single hour during embryonic development (24 timepoints, 0-24h). Median: 60.8%. Only genes whose expression range spans at least an order of magnitude and whose maximum expression level exceeds 10 RPM were considered in this analysis. (E) Example of a gene with fast transitions kinetics of high absolute magnitude.

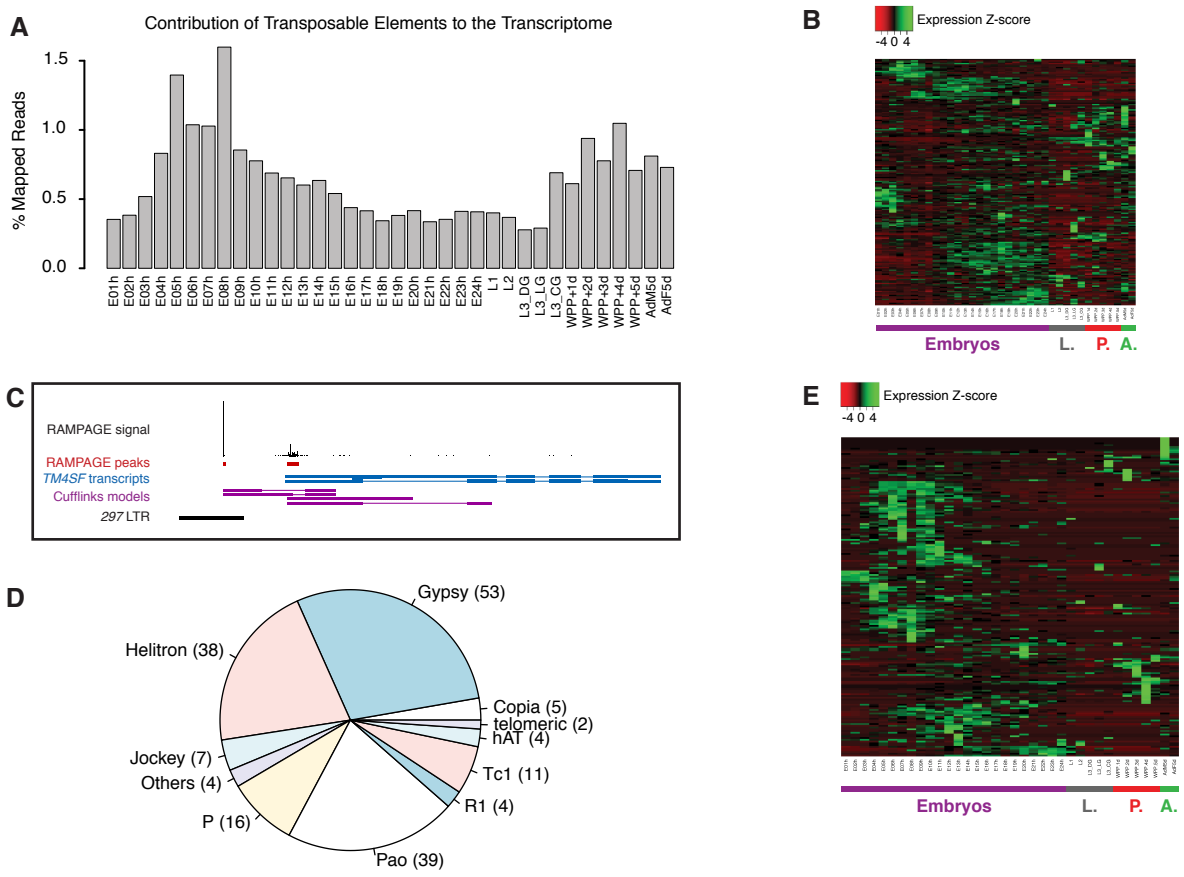

Figure 3.3. Transposon expression and co-option
. (A) Contribution of transcription initiating within transposable elements to the developmental transcriptome. For each time point, we report the proportion of all mapped reads (aligned uniquely or to multiple locations) for which the 5' end lies in an annotated transposon. (B) Z-score-normalized expression profiles for all annotated classes of transposable elements. Note the developmental regulation of virtually all classes, as well as the disparity of patterns across classes. (C) A *297* LTR provides a strong alternative promoter for the *TM4SF* gene. (D) Subfamilies of transposable elements providing TSSs for annotated genes. The number of TSCs for each subfamily is reported in brackets (total 182). (E) Z-score-normalized expression profiles for all transposon-derived genic TSCs. The diversity of expression profiles underscores the versatility of transposons as regulatory modules.

the emergence of the transposon TSCs did constitute genuine regulatory innovation. All major classes of *D. melanogaster* transposons are represented (LTR, LINE, DNA, Helitrons; see Figure 3.3), although LTR retrotransposons alone – predominantly those of the *gypsy* and *pao* families – account for a little over half of all instances. Not only full-length LTR retrotransposon insertions, but also solo LTRs and other fragments, are found to provide genic TSCs.

Importantly, these 182 TSCs represent a very stringently selected set, which may lead us to underestimate the pervasiveness of the phenomenon. To obtain a more accurate estimate, we optimized our peak-calling strategy to increase sensitivity for weaker TSCs, such as those active only in rare cell types. Retaining a still stringent threshold of at least 3 tags in a single time point (see Methods), we thus discovered an additional 333 transposon-borne TSCs driving the expression of annotated genes, bringing the total number to 515. We expect that deeper sequencing and targeted examination of rare cell types will lead to dramatic revisions of this initial estimate.

Furthermore, our initial high-confidence set includes 779 transposon-borne TSCs driving the expression of novel transcripts. To provide further evidence of the biological relevance of transposon-driven developmental transcription, we sought to better characterize these non-genic transcripts. From our data, we could reconstruct Cufflinks partial transcript models for 509 of the aforementioned 779 non-genic transposon-derived promoters (total 598 transcripts). Out of 598 transcript models, 209 are clearly spliced, showing that these transcripts often undergo post-transcriptional processing. Out of 598, at least 198 transposon-driven transcript models (from 161 promoters) contain at least 30% non-transposon sequences, which demonstrates that TE-derived promoters often drive the expression of neighboring non-repeat regions. This is bound to be an underestimate, since our transcript models are usually partial. We hypothesize that the creation of promoters by transposons could be a very powerful evolutionary mechanism for the creation of novel non-coding RNA genes. Strikingly, 112 of the 598 non-genic transcripts are antisense to Flybase-annotated mRNA transcripts. Another 61 overlap annotated transcripts on the same genomic strand. The abundance of such gene-overlapping transcripts points to a potentially important role of transposon-driven non-coding transcription in the regulation of gene expression.

Transposons distribute promoters with pre-programmed regulatory logics

We next investigated whether the transposons that contribute TSCs to host genes have similar expression profiles to the transposon class they belong to. This would imply that transposons contribute functional modules with predetermined and stereotyped regulatory logics to host genes. We show that the 182 high-confidence transposon-derived genic TSCs overall have a clear tendency to share the expression profiles of their class of origin (Figure 3.4). This trend becomes even clearer when focusing on TSCs derived from specific classes of elements. In particular, the 18 TSCs derived from the LTRs of *roo* elements are expressed in temporal patterns that display compelling similarity to each other and to the overall class pattern (Figure 3.4). This observation also holds true for other classes of elements (Figure 3.4). *roo*-driven expression was clearly detectable in profiles established by standard RNA-seq, indicating that these elements drive the expression of full-length genic transcripts (Appendix 3 Figure 5.22).

This is quite a striking result, since the detection of such broadly correlated patterns is only possible if a large fraction of gene-driving insertions possess the same specificity. As the analysis of certain transposon sequences has shown, however, a large number of diverse transcription factor binding site (TFBS) motifs can often be found throughout the length of the sequence (see for example Lynch et al. (2011)). Thus, different fragments derived from the same original element may confer vastly different expression specificities, or even carry out other molecular functions. For instance, different human MER20 insertions can bear, in the same cell type, chromatin profiles that are characteristic of either transcriptional enhancers, repressors, or insulators (Lynch et al. 2011). Therefore, even a strict interpretation of the copy-and-paste model does not necessarily imply simple and systematic expression profile correlations between fragments belonging to the same TE family. We conclude that our observations strongly support the hypothesis that transposons often impart their own regulatory properties upon the genes they drive the expression of.

In order to identify *cis*-elements in the transposons that could explain these regulatory properties, we focused on *roo* LTR TSCs, the largest group with clear class-specific expression patterns. The analysis of multiple sequence alignments revealed little divergence among all these insertions and relative to the class consensus (Figure 3.5). The consensus LTR sequence was found to have matches to 6 TFBS motifs (Nub, Tin, Vnd, Btd and Br_z4, q-value < 0.05 for each instance; Bap, q-value=0.075; see Figure 3.5). With the exception of *br*, all the genes encoding transcription factors predicted to bind

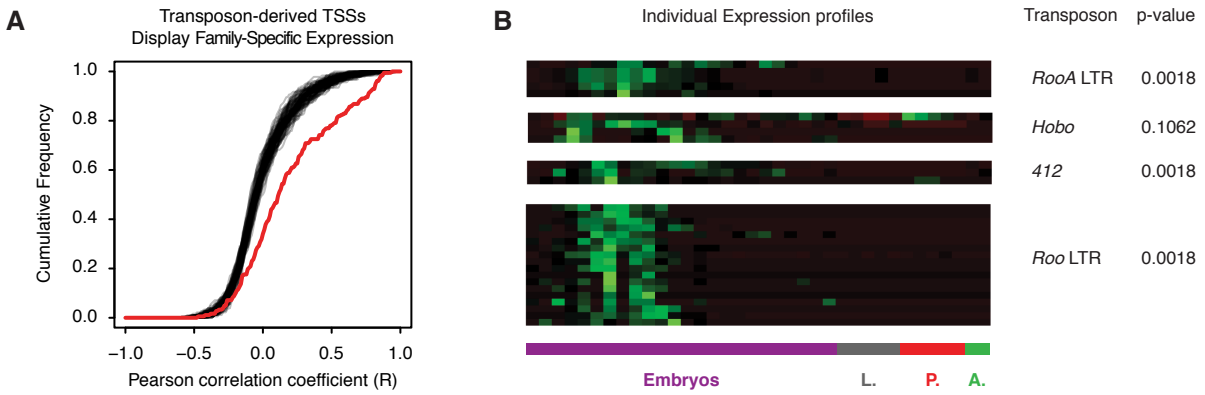

Figure 3.4. Transposons impart their own expression specificity upon genes
(A) Cumulative distribution of pairwise Pearson correlation coefficients (R) between individual transposon-derived TSCs and the class of TEs they are derived from (red curve). This measures the similarity between the expression profile of a given gene-driving insertion and the overall profile of the class it belongs to. The black curves show 100 simulations in which the TSS-transposon class pairs were randomized. Permutation test (10,000 randomizations) p=0.0001. (B) Z-score-normalized expression profiles for individual subfamilies of transposons. Bonferroni-corrected P-values from permutation tests quantify the significance of the similarity between each group of TSCs and its cognate class profile. NB: 0.0018 is the limit of the power of the statistical tests.

these motifs have expression profiles consistent with that of the *roo* LTRs (Figure 3.5). The analysis of endogenous truncated LTR copies is consistent with a role for these sequences in transcriptional regulation (Appendix 3 Figure 5.23). Embryonic expression of *roo* transposons has previously been shown to require the mesoderm-determining genes *twist* and *snail* (Bronner et al. 1995). It is also known that the *tin* and *vnd* genes are direct targets of the TWI transcription factor (Lee et al. 1997; Mellerick and Nirenberg 1995; Yin et al. 1997), and that *bap* is a direct target of TIN (Zaffran et al. 2001). Additionally, we show that the TSS is at the same position in all of the LTRs of interest (Appendix 3 Figure 5.24), and that it overlaps a canonical core promoter Initiator (INR) sequence (Figure 3.5). Overall, this analysis shows that *roo* LTRs possess a proper Pol II core promoter and *cis*-regulatory elements that can explain their expression specificity.

Population genetics of transposon-derived genic TSCs

To explore the evolutionary implications of our observations, we used existing data (Petrov et al. (2011)) on the population frequencies of many transposons, including 56 of the TSC-bearing insertions we identified (see Methods). Of those insertions, 45 are estimated to be rare or very rare variants in the wild North-American (NA) populations studied. Notably, 42 of these rare variants were absent from the ancestral African (AF) populations the NA ones split from 10,000-16,000 years ago – a number which again underscores the power of this mutational mechanism to continuously create standing variation for regulatory networks. Additionally, we found that 11 variants (20% of total) are either common (4) or fixed (7) in NA populations, showing clearly that transposon-derived variants can make significant contributions to population gene pools.

3.3 Discussion

We measured promoter activity throughout the life cycle of *D. melanogaster*, thus providing a high-quality reference dataset for the community. Importantly, this dataset offers particularly high temporal resolution (1 hour) for the period of embryonic development. We observed a very widespread use of alternative promoters as a means to implement differential regulation in a developmental context. Our results also show that transposons contribute large numbers of developmentally expressed TSSs, and support a long-hypothesized mechanism through which transposons distribute pre-assembled *cis* modules throughout the genome. These modules appear to affect the developmental regulation of hundreds

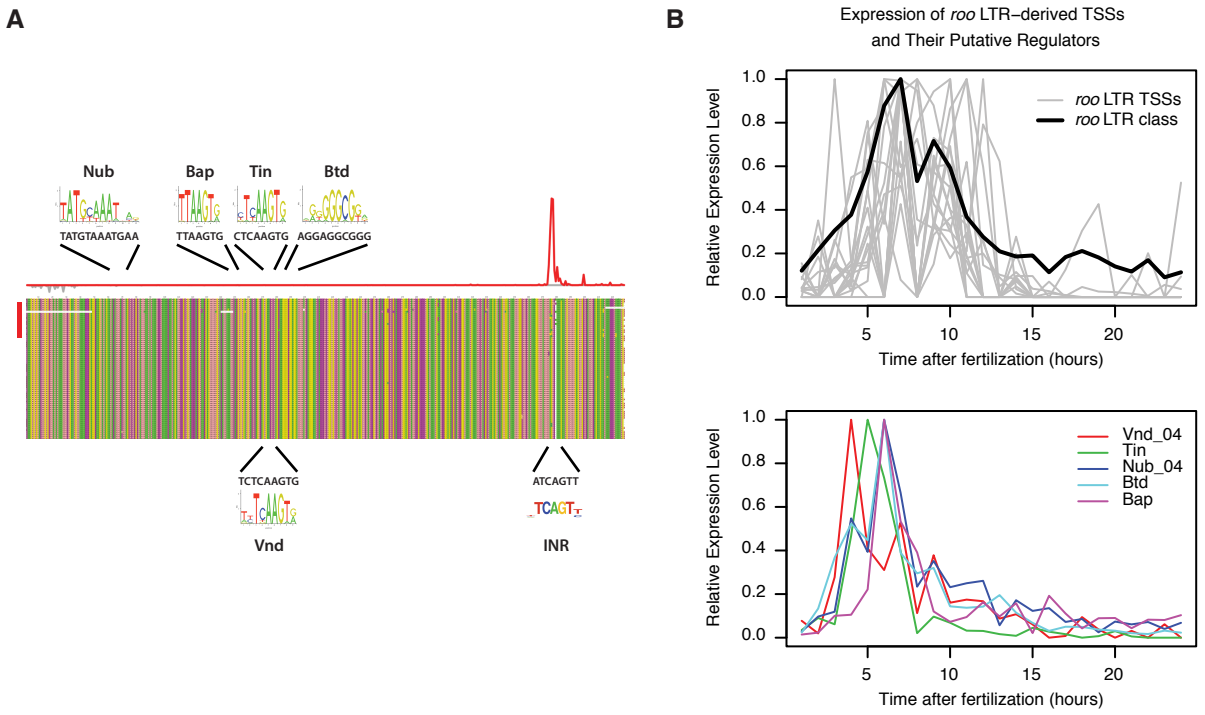

Figure 3.5. Core promoters and *cis*-regulatory elements in *roo* transposons
(A) Multiple alignment of the sequences of the 18 LTRs providing TSCs for host genes (red bar on the left) to the *roo* consensus (upper sequence) and to a set of full-length LTRs with high similarity to the class consensus. The histogram above shows the density of tags on the upper (red) and lower (grey) strands. The positions of various sequence motifs are depicted, along with the logo of the known motif and the actual consensus sequence of the LTR. The TFBSs for NUB and BAP and the Initiator sequence (INR) are on the upper strand; the TFBSs for TIN, VND and BTD are on the lower strand. (B) Expression profiles of the genes encoding putative regulators of *roo* LTRs. *nub* and *vnd* have more than one TSS, and only the one with the expression profile most consistent with *roo* LTRs is shown.

of genes and non-coding transcripts. We expect that further study of more complex genomes with higher transposon contents, such as mammalian or plant genomes, will uncover even greater numbers of such instances. Additionally, our study focused very specifically on transposons providing promoters, but these elements have been shown to have the potential to also contribute transcription factor binding sites, enhancers, silencers, insulators or microRNA target sites (Bourque (2009), Bourque et al. (2008), Lindblad-Toh et al. (2011), Lynch et al. (2011)). Overall, our observations underscore the potential of transposons as a powerful and versatile creative force in regulatory innovation.

3.4 Methods

Fly stocks & sample collections:

Stocks of the *y; cn bw sp* strain were maintained in standard cornmeal medium bottles in a 24ºC incubator. Embryo collections were performed in population cages (Flystuff, #59-116). 2- to 7-day-old flies were left to acclimatize to the cage for at least 48h and regularly fed with grape juice-agar plates (Flystuff, #47-102) generously loaded with yeast paste. After two 2-hour pre-lays, embryos were collected in 1-hour windows and aged appropriately (24 timepoints, 0-24h). Embryos were washed with deionized water, dechorionated for 90 sec with 50% bleach, rinsed abundantly with water, and snap-frozen in liquid nitrogen. Larvae and pupae were collected as described previously (Graveley et al. (2010)). For L1 and L2 stages, 2-hour embryo collections were aged for 42 or 66 hours, larvae were briefly rinsed with deionized water and snap-frozen. For L3 stages, embryos were transferred to bottles containing cornmeal medium supplemented with 0.05% bromophenol blue, and wandering L3 larvae were staged based on gut staining (dark, light or clear gut) and snap-frozen. For pupae, 2-hour embryo collections were transferred to standard cornmeal medium bottles, the positions of new white prepupae on the walls of the bottle were marked, and pupae were collected and snap-frozen at the desired age. For adults, 0- to 12-hour-old flies were sexed and kept in vials with cornmeal medium for 5 days, then snap-frozen.

RNA extraction:

Total RNA was extracted from adult flies using Trizol (Invitrogen) according to the manufacturer's instructions and treated with DNAseI (Roche). Extraction from embryos, larvae and pupae was performed using a Beadbeater (Biospec, Cat. #607) and the RNAdvance Tissue kit (Agencourt #A32649) accord-

ing to the manufacturer's instructions, including DNAseI treatment. For the human K562 cell line, RNA was extracted using Trizol according to the manufacturer's instructions and treated with DNAseI (Roche). We systematically checked on a Bioanalyzer (Agilent) that the RNA was of very high quality. 5'monophosphate species – mainly ribosomes – were depleted by TEX digest (Supplementary Methods).

Library preparation & sequencing:

Three multiplexed libraries were prepared: one for embryos (24 barcoded samples), one for larvae and pupae (10 samples), and one for adults (2 samples). The reverse-transcription was run in parallel for all samples destined to the same library, and the samples were pooled right after reverse-transcription. The libraries were run on a DNA HS Bioanalyzer chip for quality control, quantified by quantitative PCR, and sequenced on one lane each on an Illumina GAIIx (adults, 2x76bp) or HiSeq (embryos, larvae and pupae, 2x101bp). See Appendix 1 for a detailed protocol.

Sequencing reads alignment:

The sequences corresponding to the library identification barcode and the reverse-transcription primer were trimmed prior to mapping. Trimmed reads were mapped with STAR, with parameters described in Tables S2-3. All uniquely mapping reads were kept. As a rescue strategy for multiply mapping reads, if all alignments for those reads started within an annotated transposon and overlapped the same gene annotation, the alignment starting in the closest transposon insertion was selected. All non-rescued multi-mappers were discarded.

Data analysis pipeline:

Data was analyzed as described in Chapter I and Appendix 1. Parameters for peak-calling: window width 15 bases, null distribution negative binomial with k=4, background weight 0.5, FDR 0.01, merging range 150 bases.

Alternative promoters in Flybase:

The number of distinct TSSs was counted for all Flybase r5.32 mRNA and ncRNA transcript annotations for which we could detect expression in our dataset. Since our peakcalling algorithm merges

windows closer than 150bp, we also merged together annotated TSSs within 150bp of each other, for the fairness of the comparison.

Identification of weaker peaks:
Weaker peaks were identified by calling peaks from the individual (non combined) collapsed datasets, to increase sensitivity for briefly expressed peaks. We also used slightly less stringent parameters (window width 10 bases, null distribution negative binomial with k=5, no downstream read background correction, FDR 0.05, merging range 150 bases), and retained all peaks supported by at least 3 independent tags. To filter out contributions from background signal in the body of transcripts, we discarded any peaks that overlapped annotated exons. We then combined the peaks from all datasets and merged any peaks closer than 50bp using BEDtools (mergeBed). Peaks were attributed to genes based on evidence from at least one cDNA.

Genome annotations: Transcript annotations were obtained from FlyBase (release 5.32). Analyses performed involved all transcripts annotated as "mRNA" or "ncRNA". Transposable element RepeatMasker annotations were downloaded from the UCSC Genome Browser. We corrected the annotation of the DNAREP1_DM element to "Helitron", based on analysis by Kapitonov and Jurka (Kapitonov and Jurka 2007).

Correlation of expression profiles between alternative TSSs:
All genic TSSs having a maximum expression level of at least 5 reads per million (RPM) were considered. We computed Pearson's coefficient of determination (R^2) for all possible pairs of alternative promoters.

Developmental transitions between alternative promoters. For all genes with maximum expression at least 10 RPM for at least 5 consecutive timepoints that had at least 2 alternative promoters, we computed the fraction of the total gene expression at each individual timepoint that was contributed by the main TSS (defined as the one that contributes the largest proportion of the total expression over the whole time series). This metric is represented as a heatmap for 1,295 genes that underwent clear transitions between alternative promoters (difference at least 0.5 between the maximum and minimum of the main

promoter fraction). (Note: a default value of 0.5 (black) was attributed to all timepoints were total gene expression <10RPM.)

Analysis of fast-regulation genes:

All genes with maximum expression levels at least 10RPM during embryonic development were considered for this analysis (full set). The fastest-changing genes were defined as those that overall undergo at least 10-fold expression level variations and display single-step variations of at least 85% of their full dynamic range. Functional category enrichment in the fast gene set relative to the full set was assessed using the DAVID database tools (Huang et al. 2009).

TSSs in transposons:

BEDtools (intersectBed) was used to search for TSSs ovelapping transposons, and we retained all TSSs that overlapped a transposon over at least 50% of their length.

Transposon subfamily profiles:

Transposon subfamily profiles were established by considering all alignments (from uniquely or multiply mapping reads) starting within any insertion of the class, weighed by the inverse of the number of alignments for the read. These profiles were normalized to the total number of transposon-derived reads in each dataset.

Expression profile comparisons between TE-derived TSCs and transposon classes:

The expression profiles of transposon-overlapping TSCs were paired to their cognate transposon class profile, and Pearson's correlation coefficient was computed for every such pair. The statistical significance of the overall similarity between profiles was assessed by a permutation test (following the recommendations of Phipson & Smyth (Phipson and Smyth (2010))) in which the TSC profiles were paired to random transposon class profiles (or, alternatively, to random genic TSC profiles). The same strategy was applied to transposons coming from individual classes. In that case, we conducted the permutation tests on all classes for which there were at least 3 TSCs by pairing the individual TSCs to

random transposons, and the p-values were adjusted for multiple testing by Bonferroni correction.

roo LTR sequence analysis:

We retrieved the sequences of the 18 *roo* LTR insertions bearing genic TSCs, of all other annotated insertions with length at least 420bp and RepeatMasker alignment score at least 4000 (chrUextra excluded, 50 insertions). Multiple sequence alignments were generated using MUSCLE (default parameters) on the EMBL website, and visualized using Jalview. Consensus transposon sequences were downloaded from Flybase. The LTR sequence we used corresponds to the first 429bp of the *roo* consensus (see Flybase). We used FIMO (Grant et al. 2011) to search for matches to TFBS motifs from the Jaspar Core Insecta database (Bryne et al. 2008), using default parameters and a 4th-order Markov background model derived from the whole genome. A custom script was used to search for matches to previously characterized core promoter motifs (FitzGerald et al. (2006)) (TATA, INR, INR1, DPE, DPE1).

Population genetics data analysis:

We used the genotyping data for Flybase-annotated transposon insertions from Petrov et al. (Petrov et al. (2011)). Each transposon-contained TSC was attributed to a Flybase transposon annotation if it fully overlapped one of them (108 insertions). Allele frequency data were available for 56 insertions.

Chapter 4

Promoter evolution patterns reveal deep conservation of non-coding transcription in *Drosophila* embryos

Regulatory changes are thought to have contributed significantly to the evolution of metazoans, and yet little is known about the evolutionary dynamics of the genomic elements that regulate gene expression. To experimentally probe the functional conservation of transcriptional promoters, we generated genome-wide profiles of promoter activity throughout embryonic development in five *Drosophila* species. We found that promoter gain and loss have been very active processes throughout the ~25 million years since the last common ancestor of these species. Transposable elements, as well as changes in the directionality of ancestral promoters, contribute substantially to this phenomenon. The principles of core promoter organization, despite being under strong purifying selection, were found to change over broader timescales, as exemplified by the increased usage of Initiator and DPE1 motifs in the *melanogaster* subgroup. Developmental gene expression profiles displayed evidence of strong purifying selection, although diversity in the behavior of individual genes was observed. Determinants of the degree of expression divergence include gene function, as well as systems-level constraints on developmental stages. However, the intensity of selection on expression specificity is independent of selection on protein-coding sequences. Importantly, we discovered over 3,600 novel putative non-coding RNA promoters, many of which are functionally preserved between highly diverged species. We present evidence of purifying selection acting on these promoters, both at the level of sequence and at the level of expression specificity. This suggests that non-coding transcription in *Drosophila* is not only prevalent, but also fulfills essential functions during embryonic development.

4.1 Introduction

There has been growing interest for decades in the idea that regulatory innovation may be paramount to the evolution of multicellular organisms, particularly in the case of morphological evolution (Carroll (2008), Stern and Orgogozo (2009), Davidson and Erwin (2006)). Theoretical considerations regarding modularity, evolvability and pleiotropy, which have recently garnered some experimental support, make such views both appealing and intuitive (Carroll (2008), Stern and Orgogozo (2009), Davidson and Erwin (2006)). Indeed, studies in a variety of organisms have uncovered a great deal of evolutionary plasticity in gene expression control, both at the transcriptional and post-transcriptional levels (Rifkin et al. (2003), McManus et al. (2010), Merkin et al. (2012), Barbosa-Morais et al. (2012)). The extent to which regulatory evolution is a target of selection is unknown, however, and the prevalence of neutral changes remains a matter of controversy (Khaitovich et al. (2004), Khaitovich et al. (2005)).

Recent advances in sequencing technologies, along with the availability of a growing number of reference genomes, have made it readily feasible to collect functional data on a genome-wide scale for multiple eukaryotic species. A number of studies have implemented this strategy over the past few years, in organisms ranging from yeast to fly and mammals (Tsankov et al. (2011), McManus et al. (2010), Merkin et al. (2012)). The rewards of such approaches are two-fold: they illuminate the inner workings of the evolutionary process, and they provide a measure of the selective forces that operate on various aspects of genome function. The second point is of great importance to the effort to decode and interpret genetic information. Indeed, comparative analysis of genome sequence alone is seldom capable of predicting the evolution of molecular function, let alone of organismal phenotypes. Only the direct measurement of functional characteristics can provide information regarding genotype-phenotype relationships in this context. In particular, such data-driven approaches are ideally suited to investigate the evolution of poorly understood genomic elements, such as regulatory regions or long non-coding RNA (lncRNA) loci.

A few seminal studies have focused on the particular case of transcription factor binding sites (TFBSs). Work in mammals has revealed strikingly fast evolutionary dynamics for TFBSs, and a diverse set of selective constraints acting on different loci and different timescales (Odom et al. (2007), Stefflova et al. (2013), Steijger et al. (2013)). Other data has suggested that the pace of change might be far slower in fly, possibly due to a more prevalent role for purifying selection in more compact genomes (Moses

et al. (2006), He et al. (2011)). In neither case, however, could the impact of TFBS gain and loss on gene expression be assessed, and it is very unclear what selective forces – if any – underlie these dynamics. Beyond TFBSs, very little is known about other types of genomic elements involved in the regulation of gene expression.

We chose to focus on promoters, which play a central role in the control gene expression – they are the nexus of transcriptional regulation, on which almost all regulatory inputs converge to be integrated and produce a specific response (Lenhard et al. (2012)). Importantly, whereas the influence of other types of regulatory elements on gene expression is generally difficult to predict, one can directly measure the effects of promoter changes on the genes they control. This provides us with a unique opportunity to connect patterns of sequence variation to gene expression patterns and gene function. To achieve this, we used a novel transcriptome profiling approach that allows the expression of individual promoters to be accurately monitored in order to profile promoter expression throughout embryonic development in 5 *Drosophila* species.

Our analyses revealed that, despite widespread purifying selection, promoter gain and loss have been very prevalent throughout the evolutionary history of the clade. Interestingly, we identified genetic recycling mechanisms as a source of regulatory novelty: both transposable elements and changes in the directionality of ancestral promoters regions contribute significantly to the evolution of transcriptional regulation. The grammatical rules of promoter design, though largely conserved, do change perceptibly over time. We show here that reliance on Initiator and DPE1 core promoter motifs, for instance, has increased throughout the evolution of the *melanogaster* subgroup. The specificity of expression of individual promoters was found to be under significant purifying selection overall, and its evolution is shaped by a complex combination of forces. Systems-level constraints related to gene function and developmental stages emerged as important contributors to patterns of regulatory divergence.

We discovered 4,050 promoters driving the expression of putative long non-coding RNAs – a much larger number than anticipated. Of those, 1,047 are functionally preserved between *D. melanogaster* and *D. pseudoobscura*, suggesting that they have been under selective pressure for over 25 million years. Indeed, we found evidence of strong purifying selection acting on both the primary sequence and the specificity of expression of these promoters. These observations suggest the existence of a trove of long non-coding RNA genes playing deeply conserved roles in a variety of crucial developmental processes.

4.2 Results

Multispecies Promoter Expression Profiling Throughout Embryonic Development

We have previously developed a novel transcript profiling method, RAMPAGE (RNA Analysis for Mapping of Promoters and Analysis of Gene Expression), in order to study the expression of individual promoters. This approach, which is based on high-throughput sequencing of 5'-complete complementary DNA (cDNA) molecules, allows the identification of transcription start sites (TSSs) with single-base resolution, and the direct measurement of the contribution of individual TSSs to steady-state transcriptomes. Although other 5'-complete cDNA sequencing methods exist (Kodzius et al. (2006), Ni et al. (2010), Plessy et al. (2010)), RAMPAGE offers far greater specificity for TSSs, while also providing valuable transcript connectivity information through paired-end sequencing of medium-length cDNA inserts (see Chapter 2). Since individual eukaryotic promoters often allow productive transcription initiation from multiple neighboring positions (Carninci et al. (2006), Lenhard et al. (2012)), we use a dedicated peak-calling algorithm to group neighboring TSSs into TSS clusters (TSCs) corresponding to individual promoters (see Methods). For brevity, from here on we will refer to the contribution of individual TSCs to the steady-state transcriptome as "promoter expression".

We harnessed this approach to generate developmental transcriptome profiles at both high temporal resolution (1 hour) and high sequence coverage (124-170 million read pairs per species) for 5 *Drosophila* species spanning over 25 million years of evolution: *D. melanogaster*, *D. simulans*, *D. erecta*, *D. ananassae* and *D. pseudoobscura*. We focused on embryonic development, a crucial period during which the body plan is established and the primordia for all larval organs are generated. As illustrated in a graphical representation of the resulting TSS expression landscape at an individual locus (Figure 4.1), one can directly compare the activity of single promoters across species and infer patterns of evolution from this data. Importantly, paired-end sequencing of cDNAs provides an evidence-based method to attribute novel TSS clusters to existing gene annotations, and also provides useful information about overall transcript structure (Figure 4.1).

For each species, peak-calling on the full time series data identified between 22k and 27k high-confidence TSCs (see Methods). The distribution of raw RAMPAGE signal (Appendix 4 Figure 5.25) and of TSCs (Appendix 4 Figure 5.26) over Flybase-annotated loci confirms the very high specificity of our assay for true transcription start sites. The analysis of biological replicates for the *D. melanogaster*

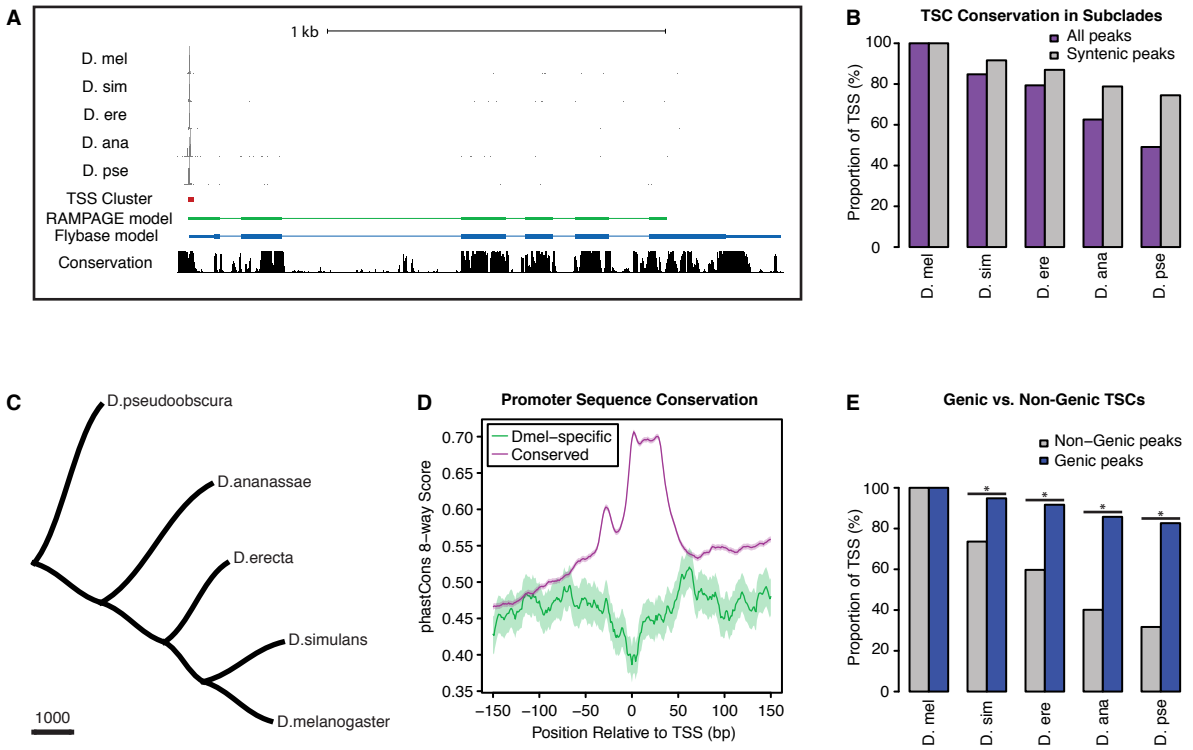

Figure 4.1. Genome-wide patterns of promoter gain and loss

(A) Distribution of RAMPAGE signal at the *NLaz* locus in 5 *Drosophila* species. The top 5 tracks are bar plots of RAMPAGE signal density (number of reads) over the *D. melanogaster NLaz* locus. For non-*melanogaster* species, sequencing reads were simply mapped to the appropriate genomes and their coordinates translated to orthologous positions based on whole-genome alignments. The red box represents the TSC identified from the *D. melanogaster* data. A partial transcript model (green) was generated *de novo* by running Cufflinks on RAMPAGE reads – note the agreement with existing transcript annotations (blue). The last track represents sequence conservation scores (phastCons). (B) Proportion of *D. melanogaster* TSCs reproducibly discovered in biological duplicates (first pair of bars) and functionally conserved in all species of subclades of increasing sizes. Subclades include all descendants of a common ancestor, and are are designated by the species that is most distantly related to *D. melanogaster*. (C) The species phylogeny can be accurately reconstructed directly from patterns of TSC gain and loss. The presence/absence of each TSC was treated as a discrete character and the unrooted tree reconstructed using the Phylip software package. (D) Average profiles of sequence conservation (phastCons scores) over the TSCs functionally conserved between all 5 species and those specific to *D. melanogaster*. Note the strong excess of sequence conservation over functionally conserved TSCs. (E) TSCs driving the expression of Flybase-annotated genes display a far higher degree of functional conservation than "orphan" TSCs ($p<0.01$ for all pairwise comparisons; chi-square test with Bonferroni correction).

time course confirmed that the detection of individual TSCs is extremely reproducible (Figure 4.1). Based on direct cDNA evidence, we could attribute 82% of *D. melanogaster* TSCs to Flybase-annotated genes. The remaining 3,693 TSCs, for which no cDNA overlaps an annotated exon, appear to drive the expression of independent transcripts. Given the quality of Flybase coding sequence annotations, it is highly likely that these "orphan" TSCs drive the expression of unannotated, non-protein-coding transcripts.

Promoter birth and death are pervasive despite strong purifying selection

We first analyzed the overall patterns of TSC gain and loss throughout the clade under study from a *D. melanogaster*-centric perspective. As a global measure of functional conservation, we assessed the proportion of *D. melanogaster* TSCs that were found to be active in all species of progressively larger subclades (see Methods). We observed a clear trend of sharply declining TSC conservation with increasing evolutionary distance. Only 49% of all peaks were found to be functionally conserved in all 5 species (Figure 4.1, purple bars). Given the incompleteness and the poor quality of some genome assemblies, this could be an underestimate of the true conservation rate. Therefore, we also assessed the conservation of only those TSCs for which syntenic alignments could be found in all species of a particular subclade (Figure 4.1, grey bars). About 75% of such peaks were fully conserved. Since some peaks must lack satisfactory syntenic alignments owing to genuine large insertions or deletions, this number is likely to be an overestimate, and we expect the actual conservation level to lie somewhere between the two boundaries proposed here. Improvements to genome assemblies should settle this issue in the future. Analyzing TSC conservation between species pairs or from a *D. simulans*-centric perspective showed similar trends (Appendix 4 Figure 5.27)).

Identical analysis of biological replicates of the *D. melanogaster* time course showed the false positive rate for gain/loss event detection to be under 0.1% (Figure 4.1). Although TSCs with lower expression levels tended to be less conserved, general trends were shared between TSCs of all expression levels (Appendix 4 Figure 5.28). Low overall expression in whole embryos may reflect weak transcription in individual cells and/or a restriction of expression to specific cell types. Furthermore, although a gain/loss of expression during embryogenesis may reflect a change in expression specificity rather than a complete gain/loss of function, the vast majority (91.4%) of *D. pseudoobscura* TSCs that were inferred to be lost in *D. melanogaster* based on embryo data were never found to be expressed at any other stage

of the life cycle. Therefore, we believe that our strategy accurately and robustly detects true promoter gain and loss events. Consistent with this, we could reconstruct the known species phylogeny by simply treating the presence or absence of individual TSCs as discrete characters in a standard parsimony framework (Figure 4.1).

Importantly, functional conservation is reflected in the conservation of promoter sequence: promoters found to be active in all species display a much higher degree of sequence conservation than *D. melanogaster*-specific promoters, which appear no more constrained than surrounding regions (Figure 4.1). This suggests that a large number of promoters have been under intense and sustained purifying selection since the last common ancestor of all 5 species. To confirm this finding, we compared the evolutionary rates of gain and loss between the TSCs that were attributed to annotated genes – which are strongly expected to play some constrained functional role – and those that were not. Although a large number of these "orphan" TSCs are far from evolving neutrally (see below), we found a stark contrast in the degree of functional conservation of the two classes (Figure 4.1). This discrepancy in conservation rates reveals a particularly high degree of constraint on the promoters of annotated genes, and clearly identifies purifying selection as a major force shaping their evolutionary dynamics.

While purifying selection does emerge as a key player, a sizeable proportion of TSCs (25-50%) are not shared between all species, underscoring the inherently fluid nature of the regulatory landscape in *Drosophila*. Comparisons to published data on the evolution of Twist binding sites revealed that overall, promoters evolve as rapidly as Twist TFBSs do, and possibly even faster (Appendix 4 Figure 5.29). Although a rigorous comparison between such disparate data types is difficult, this appears to show that promoters and TFBSs do not evolve on wildly different timescales. Given the relative complexity of core promoters, this suggests that either weak negative selection or positive selection on a subpopulation of loci underlies these rather wide-ranging changes in the transcriptional landscape.

Interestingly, we found mutational mechanisms that recycle existing genetic elements for the generation of novel promoters to be prevalent. We and others have previously shown that transposable elements contribute large numbers of functional promoters in various eukaryotic organisms, including *D. melanogaster* (see Chapter 3). This observation is confirmed here in all 4 other species. The prevalence of transposon-derived TSCs seems to scale linearly with the overall transposon content of genomes (Appendix 4 Figure 5.30). In all species, many of these transposon-derived promoters drive the expression of annotated genes (Appendix 4 Figure 5.30). We also found that ~20% of all TSC gain and

loss events seem to arise from shifts in the directionality of conserved promoters: in these cases, TSC gain and loss stems from ancestrally unidirectional promoters becoming bidirectional, and vice versa. This is particularly interesting in light of recent studies pointing at widespread bidirectional initiation at eukaryotic promoters, with the directionality of productive transcription being enforced at the level of elongation and/or RNA stability. We speculate that such directionality-enforcing mechanisms may present particular evolvability, and therefore contribute significantly to regulatory innovation.

Core promoter syntax, albeit strongly constrained, does evolve perceptibly

The syntactic rules that underlie the design of RNA Polymerase II core promoters in eukaryotes are unclear, although new data has suggested some unifying principles (Lenhard et al. (2012), Venters and Pugh (2013)). Accordingly, it is largely unknown to what degree core promoter elements are free to diverge over time, and to what extent they differ between clades (Lenhard et al. (2012)). A number of core promoter sequence motifs identified in *Drosophila* appear not to be present in human, or at least to have diverged beyond recognition, showing that change is certainly possible over very long timescales. It remains to be determined, however, whether consensus sequence motifs can change over shorter timescales.

We searched the promoter regions of all 5 species for 15 core promoter sequence motifs identified in *D. melanogaster* by previous studies (FitzGerald et al. (2006)). When comparing the distribution of motifs over promoters that were found to be functionally conserved between pairs of species, we found that overall their sequence composition was strikingly similar, even between the most distant species (Figure 4.2). This observation held true both for sequence elements that are precisely positioned relative to the main TSS, and for elements that display much more flexible positioning.

Although these observations suggest strong selective constraints on promoter design principles, they do not directly demonstrate the effects of purifying selection at individual promoters. Therefore, we analyzed base pair-level measures of selective constraint (phyloP scores) at classes of promoters that displayed varying motif compositions (Figure 4.2). Compared to the average of all promoter regions, those that contained Initiator (INR) and TATA box motifs displayed a clear excess of sequence conservation precisely in the region where these motifs lie. This unambiguously demonstrates that individual core promoter elements are under selective pressure, reinforcing the notion that promoter design principles are mostly invariant over the timescales under study.

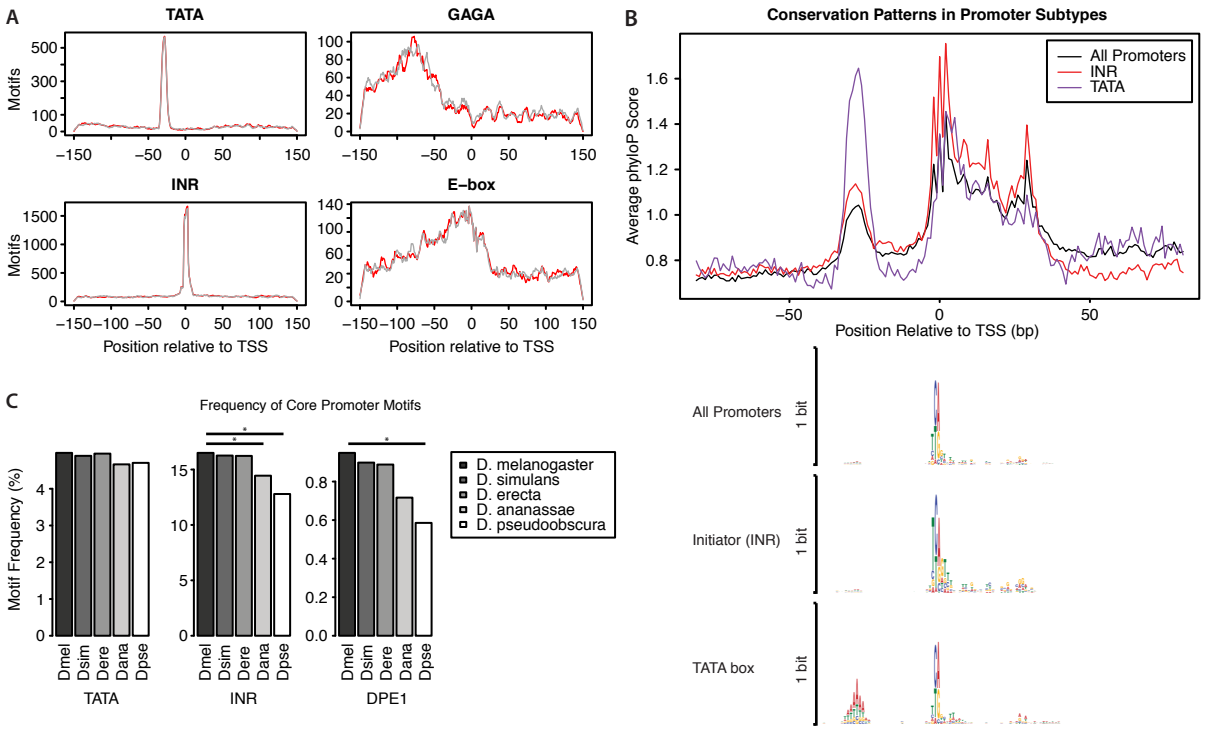

Figure 4.2. Core promoter motifs and evolution of promoter syntax
(A) Core promoter sequence motif usage is highly similar between orthologous promoter regions that are functionally conserved between pairs of species (here, *D. melanogaster* and *D. erecta*). (B) Individual motif instances are under purifying selection. Base-wise conservation scores (phyloP) are increased at specific positions in classes of promoters that have matches to known sequence motifs. For instance, note the excess of sequence conservation at TATA box promoters relative to all promoters, specifically over the TATA box motif. The sequence motif logos were derived from all promoter regions of a given class. TSCs overlapping protein-coding sequence on either strand were excluded from this analysis. (C) The INR and DPE1 motifs are less frequently used in *D. pseudoobscura* and *D. ananassae* than in the 3 other species. For each species, the bar plots represent the proportion of TSCs with matches to sequence motifs. Only *D. melanogaster* TSCs found to be functionally conserved in the target species were included in this analysis. Pairwise comparisons to *D. melanogaster*: chi-square test with Benjamini-Hochberg correction for multiple testing (stars above graph denote p<0.01).

Despite this overall conservation of promoter organization, we were able to confidently identify some trends of change in core promoter elements (Figure 4.2). Strikingly, the INR motif – one of the most prevalent and well-characterized – was found to be significantly less frequent in the 2 species most distant from *D. melanogaster*, namely *D. ananassae* and *D. pseudoobscura*. A similar trend was observed for the DPE1 motif, but not at all for the TATA box motif. Controls suggested that these effects were not likely to be attributable to whole-genome alignment errors at large evolutionary distances (Appendix 4 Figure 5.31). It is important to stress that these comparisons were carried out strictly on a set of orthologous promoter regions that are functionally conserved across all species. Therefore, our observations clearly suggest that the *melanogaster* subgroup (*melanogaster*, *simulans* and *erecta*) has evolved a broader reliance on INR and DPE1 elements relative to its ancestors.

Evolution of developmental gene expression and systems-level constraints

Precise spatiotemporal expression patterns are at the heart of embryo segmentation and morphogenesis, and there is evidence for strong conservation in the expression of well-characterized developmental regulators. On a more global scale, however, it is unclear to what extent the expression of promoters and genes varies between species, although previous studies have pointed to extensive evolutionary plasticity (Rifkin et al. (2003), McManus et al. (2010)). It is also unknown whether this plasticity is contingent upon gene function, and for protein-coding genes, how it relates to the plasticity of coding sequences.

The comparison of temporal expression profiles across species requires correcting for differences in developmental timing so that orthologous time points may be compared. Building upon previously described methods, we chose to define developmental timing based on global genome expression, and we aligned the time series to one another to maximize overall similarities between transcriptomes. We used an implementation of time-warping algorithms that was specifically developed for the treatment of data from different species. This process consisted primarily of linear transformations, and the analysis of well-characterized genes confirmed the high quality of the alignments (Appendix 4 Figure 5.32).

As measurement accuracy is a function of expression level, we used *D. melanogaster* biological replicates to identify an expression range within which we could expect reproducible estimates of expression profiles, and only considered promoters with a maximum expression level of at least 25 RPM

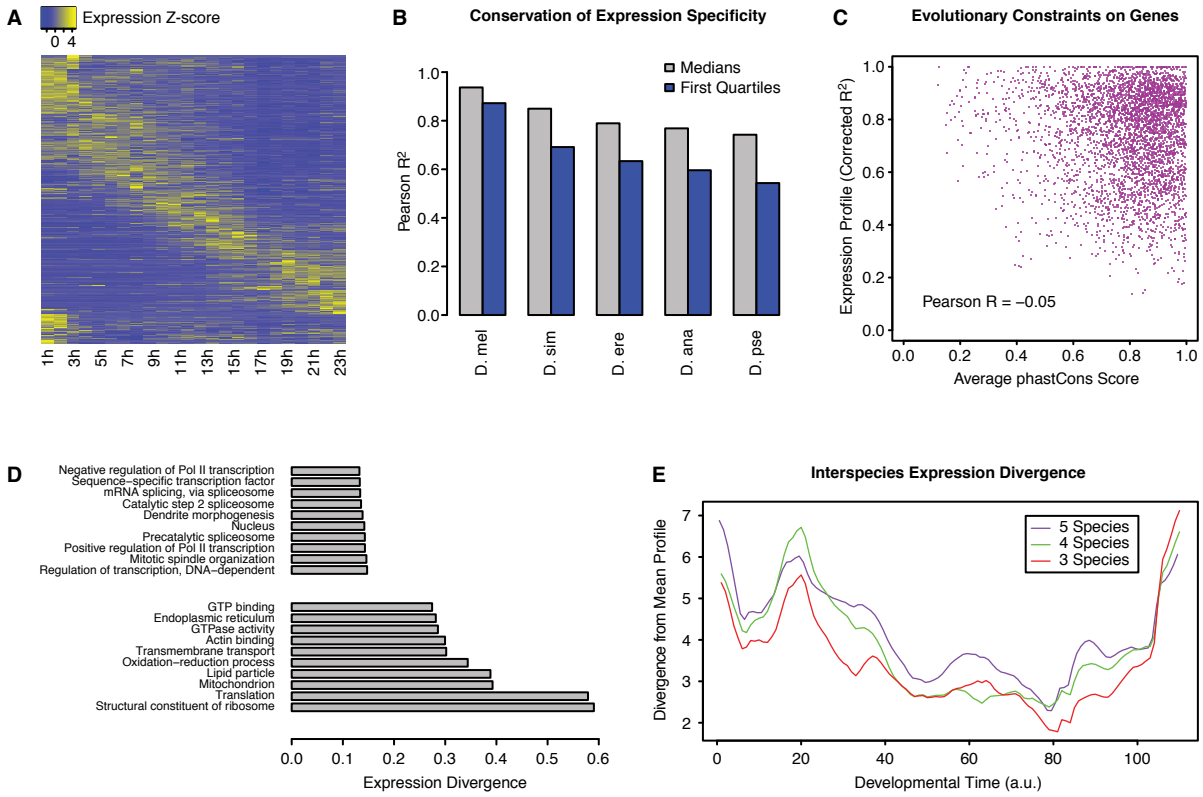

Figure 4.3. Selective pressures on developmental patterns of gene expression
(A) Developmental expression profiles of individual *D. melanogaster* promoters. Only promoters with a maximum expression level at least 25 RPM (8,668 promoters), for which reproducibility between biological replicates is very high, are included. (B) Conservation of the temporal expression profiles of individual promoters. For each subclade, we computed the average correlation coefficient between all pairs of species for each individual gene. The graph shows the median and first quartile over all genes with orthologs across the subclade. (C) Selective pressures on protein-coding sequences and temporal expression profiles are independent. For each protein-coding gene with orthologs in all 5 species and maximum expression at least 25 RPM in *D. melanogaster* (3,914 genes), we are plotting the average profile correlation between all pairs of species (corrected, see Methods) versus the average phastCons score over the coding sequence. (D) The evolutionary divergence of expression specificity varies widely between Gene Ontology (GO) categories. For each gene with orthologs in all 5 species and maximum expression at least 25 RPM, we computed a measure of overall divergence across the clade (see Methods). The barplot shows the average divergence per GO categories, for the 20 categories with the lowest (top) and greatest (bottom) divergence. (E) Selective constraints on gene expression specificity vary between developmental stages. We performed principal component analysis (PCA) on all genes with orthologs in all 5 species and maximum expression at least 25 RPM in *D. melanogaster*, and considered gene expression in the space of the first 3 components. The plot represents the average Euclidian distance of each species to the average profile across species, for each time point (4 species: excludes *D. pseudoobscura*; 3 species: excludes *D. pseudoobscura* and *D. ananassae*).

in *D. melanogaster* for subsequent analyses (Figure 4.3). For those promoters, the median Pearson R^2 between replicates was 0.95 (Figure 4.3 and Appendix 4 Figure 5.33). In general, for all genes for which clear orthologs could be identified, expression profiles were tightly conserved between species: for the full clade (5 species), the median R^2 was 0.75 (Figure 4.3 and Appendix 4 Figure 5.33). Extensive variation could be observed, however. For instance, the *hunchback* gene, whose precise expression is crucial for proper segmentation, displayed extreme conservation in the expression of both of its promoters (Appendix 4 Figure 5.33). On the other hand, the promoter of the *RpL19* ribosomal protein gene showed large differences in its expression between species (Appendix 4 Figure 5.33).

As a first inquiry into the sources of such gene-to-gene variation, we focused on the relationship between selective constraints on expression specificity and protein-coding sequence. Strikingly, we found the two to be entirely uncorrelated: at the level of individual genes, the two types of constraints seem to act absolutely independently of each other (Figure 4.3). We then investigated whether variable selective pressure on expression specificity might find its source in gene function. Grouping genes according to their Gene Ontology (GO) annotation terms showed this to be the case: the degree of expression pattern conservation differs widely between GO categories (Figure 4.3). Functions related to the regulation of transcription and splicing dominated the top of the conservation scale, in accordance with the known molecular function of many master regulators of early development. Categories related to the core translational machinery and cytoskeletal structures were prevalent at the bottom of the list. We also found clear evidence for differential selective pressure across developmental stages, in accordance with other studies. Indeed, we found the degree of interspecies divergence in gene expression to follow the ”hourglass” pattern previously described (Kalinka et al. (2010), Figure 4.3). The very minimum of our divergence metric was reached only after what is considered the insect phylotypic stage, but that stage had near-minimum divergence. This slight discrepancy with previous work might be explained by the species considered or the genes included in each study. Overall, our findings point to a complex ensemble of interwoven selective pressures – some of them acting on complex systems-level properties – shaping the evolution of developmental gene expression.

Deep conservation of over a thousand long non-coding RNA promoters

Studying promoter expression in a phylogenetic framework provides a unique opportunity to address the question of long non-coding RNA (lncRNA) conservation and functionality. Indeed, the conserva-

tion of features beyond neutral-rate expectations provides the ultimate proof of unambiguous, selectable biological function. Such approaches, however, have been complicated by the fact that lncRNA transcript sequences are under rather loose constraint overall. Therefore, our ability to pinpoint TSSs with single-base accuracy gives us unprecedented leverage to detect otherwise elusive sequence conservation patterns. Furthermore, beyond promoter sequence conservation, we are also in a position to assess selective constraint on the specificity of expression of these loci.

We found 3,693 embryonic TSCs in *D. melanogaster* that could not be functionally linked to any Flybase-annotated gene, and therefore represent putative lncRNA promoters. We also identified TSCs for 357 Flybase-annotated lncRNAs. Of these 4,050 TSCs overall, 2,435 could be aligned to all other genomes, and 1,047 were functionally shared between *D. melanogaster* and *D. pseudoobscura*. The similarity of their expression patterns with those of particular sets of protein-coding genes suggests a broad diversity of potential developmental functions (Appendix 4 Figure 5.34).

These deeply conserved elements constitute a putative core set of *Drosophila* lncRNA promoters. We found strong evidence of their conservation at the sequence level: indeed, the pattern and intensity of sequence constraint around these TSCs is comparable to that observed at all functionally conserved TSCs taken together (Figure 4.4). In terms of expression pattern conservation, we found lncRNA promoters to be under a degree of constraint well beyond that of many functional categories of protein-coding genes (Figure 4.4). Both observations taken together argue strongly for sustained selective pressure on over a thousand putative lncRNA promoters for at least 25 million years.

Furthermore, this is a stringently selected set, and many TSCs were excluded simply because of the poor quality of genome assemblies. To place a more reasonable lower bound on the true number of conserved lncRNA promoters, we focused on those shared between the 3 species of the *melanogaster* subgroup. These 1,836 promoters display a high degree of sequence conservation within the subgroup (Appendix 4 Figure 5.35), suggesting strong lineage-specific selective constraints. Importantly, we only considered lncRNAs expressed during a very short developmental period. We previously reported 7,421 putative lncRNA promoters in an analysis of the whole life cycle (see Chapter 3), and also detected expression of only 205 of 1,119 recently identified lncRNA (Young et al. (2012)). This suggests that we are only beginning to scratch the surface of lncRNA biology in *Drosophila*, and that many more loci may be under selective constraint.

There is obviously a need to fully characterize the transcripts generated from these promoters,

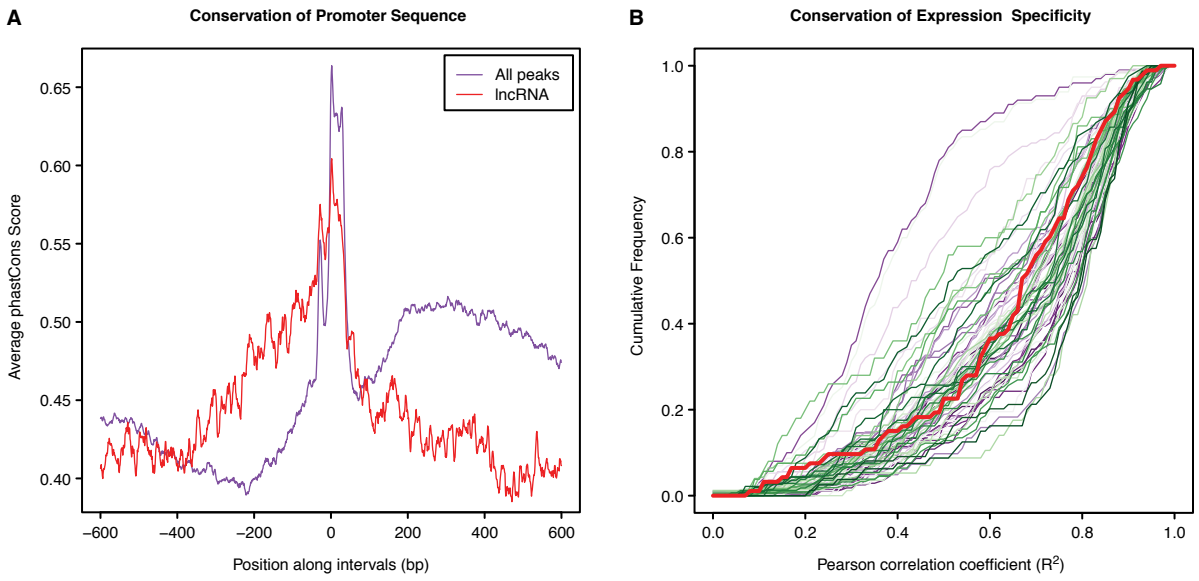

Figure 4.4. Strong purifying selection on long non-coding RNA promoters
(A) The sequences of functionally conserved lncRNA promoters are under comparable selective pressure to those of protein-coding genes. (B) The developmental expression profiles of functionally conserved lncRNA promoters are far more constrained than those of many categories of protein-coding genes.

and to rigorously assess both their independence from annotated protein-coding genes and their own protein-coding potential. We are currently making progress on this issue, using two different strategies. First, we are starting to analyze full-length transcript annotations recently generated from multiple short-read data types by the modENCODE consortium (personal communication). Second, we are working towards generating full-length transcript sequences by using a modified version of our protocol adapted for sequencing on the Pacific Biosciences platform (see Appendix 2).

We are planning to characterize the spatial expression patterns of a few of the most highly conserved examples by RNA fluorescent *in situ* hybridization (FISH). We will also attempt to knock down their expression *in vivo* using standard short hairpin RNA (shRNA) tools from the Transgenic RNAi Project (TRiP) at Harvard Medical School.

4.3 Discussion

Our analyses provide, to our knowledge, the first genome-wide overview of promoter evolution in *Drosophila* and its relationship to developmental expression patterns. Although we found individual promoters to be under significant purifying selection, promoter birth and death are very active processes, and have been prevalent throughout the history of the clade. Overall, the design principles of Pol II core promoters are under considerable selective pressure, and individual instances of canonical sequence motifs display hallmarks of strong purifying selection. In spite of this, we found evidence of some plasticity on longer time scales: indeed, our findings show that the prevalence of INR and DPE1 motifs in Pol II promoters is higher in the *melanogaster* subgroup. This particular pattern is consistent with the possibility that these motifs were less prevalent ancestrally, and that increased motif usage is a derived state.

Developmental expression patterns, as expected, were found to be under intense selective pressure. The factors that modulate evolutionary constraints appear diverse and complex. We found no relationship between the selective forces acting on the protein-coding sequences of genes and on their expression specificity. On the other hand, the particular functions that genes perform appear to be an important factor, as different functional categories display different degrees of divergence. Systems-level constraints, such as those operating on developmental stages, also appear to play an important role.

Our data revealed the existence of thousands of novel promoters in *D. melanogaster*, many of

which drive the expression of unannotated, most likely non-coding transcripts. Strikingly, we found that over one thousand of those are expressed in the embryos of *Drosophila* species over 25 million years apart. We found strong evidence of purifying selection at these promoters, at the levels of primary sequence and expression specificity. For instance, the upstream promoter of the *bithoraxoid* (*bxd*) non-coding transcriptional unit, which plays a critical role in the regulation of *Ultrabithorax* (*Ubx*) expression, has one of the most tightly conserved expression profiles that we have observed. These elements represent a putative core set of embryonic lncRNA promoters in *Drosophila*, and the diversity of their expression patterns suggests they may have extremely varied developmental roles. Importantly, their molecular functions are entirely unknown in the vast majority of cases, and may prove diverse as well.

Arguably, the most important topic that we have been unable to address here is the nature of the selective forces that drive changes in expression patterns. The relative contributions of neutral and adaptive changes to transcriptome evolution are difficult to infer from such a sparse sampling of species in the clade. However, extending the analysis presented here to more species would provide precise estimates of evolutionary rates of expression divergence at a fine phylogenetic scale. The comparison of those rates between lineages would yield significant insights regarding the selective forces at play. For instance, a constant rate of change throughout all branches of the tree would be consistent with neutral drift. Conversely, sudden shifts from a stable ancestral expression profile to a different, yet equally stable derived profile would be diagnostic of adaptive changes.

Such insights would be valuable, but obviously this approach still remains largely phenomenological. A true understanding of selective forces cannot emerge without a true understanding of the molecular mechanisms underlying regulatory evolution. Identifying those with accuracy on a broad genomic scale will require approaches that can effectively explore the space of possible genetic variation at regulatory sequences, and the impact of this variation on transcriptional output. The recent development of massively parallel reporter assays (MPRA) may provide one route for this type of large-scale exploration of sequence-function relationships.

4.4 Methods

Fly stocks:

All *Drosophila* strains were obtained from the *Drosophila* Species Stock Center at UC San Diego, CA (https://stockcenter.ucsd.edu/info/welcome.php). For each species considered we worked with the reference genome strain. Stock numbers: *D. melanogaster* #14021-0231.36, *D. simulans* #14021-0251.195, *D. erecta* #14021-0224.01, *D. ananassae* #14024-0371.13, *D. pseudoobscura* #14011-0121.94. Stocks were maintained on standard cornmeal medium. Embryo collections in consecutive one-hour intervals were conducted as described in Chapter 2.

RNA Extraction and Library preparation:

Sample homogenization, RNA extraction and DNaseI treatment were carried out as described in Chapter 3. The quality of every sample was assessed on a Bioanalyzer RNA Nano chip. RAMPAGE libraries were prepared as described in Chapter 2. For every time series, each sample bas labeled with a different sequence barcode during reverse-transcription, and all samples for the series were then pooled and processed together, as described in Appendix 1. Quality control and library quantification were carried out on a Bioanalyzer DNA High Sensitivity chip.

Genome references and annotations:

All reference sequences and annotations were obtained from Flybase (http://flybase.org). *D. melanogaster* release 5.49, *D. simulans* r1.4, *D. erecta* r1.3, *D. ananassae* r1.3, *D. pseudoobscura* r2.9.

Primary data processing:

Data for each time series was processed independently using the pipeline described in Appendix 1. Reads were mapped to the appropriate reference genomes using STAR. Peaks were called on the pooled data from whole time series, using parameters optimized to yield good TSS specificity with respect to annotations and comparable numbers of peaks for all species. All peaks overlapping Flybase-annotated rDNA repeats were filtered out. *D. melanogaster* replicate 1: window size (w) = 15nt, negative binomial dispersion parameter (k) = 50, FDR=15%, read 2 background weight (b) = 0.5, merging distance (d) = 150nt (24,831 peaks). *D. melanogaster* replicate 2: w=15, k=10, FDR=10%, b=0.5, d=150 (24,093

peaks). *D. simulans*: w=15, k=50, FDR=10%, b=0.5, d=150 (25,133 peaks). *D. erecta*: w=15, k=50, FDR=30%, b=0.5, d=150 (22,463 peaks). *D. ananassae*: w=15, k=20, FDR=5%, b=0.5, d=150 (26,867 peaks). *D. pseudoobscura* w=15, k=20, FDR=10%, b=0.5, d=150 (25,839 peaks).

TSC conservation:

Functional conservation was assessed for all peaks with at least 15 RAMPAGE tags that did not map to heterochromatic regions or chr4 in *D. melanogaster*, or orthologous regions in other species. We translated the genomic coordinates of each peak in each species to coordinates in the multiple sequence alignment of all genomes (15-way MultiZ alignment from UCSC, http://hgdownload.soe.ucsc.edu/goldenPath/dm3/multiz15way). To be considered for analysis, each peak was required to have a unique syntenic alignment in all other species considered, defined as follows: both ends of an 800-bp window centered on the middle of the peak had to map to the same strand of the same chromosome or scaffold, 50% of bases had to be aligned (i.e., not in assembly gaps), and 25% of bases had to be aligned to bases (as opposed to alignment gaps). Raw 5' signal for each genome was also translated to multiple alignment coordinates. For each peak from each species, functional conservation was assessed by counting the number of RAMPAGE tags in each species. A peak was considered absent in a target species if it had at least a 100-fold lower signal than in the reference species. Peaks with <100 tags in the reference species were considered absent if they had no detectable signal in a target species.

Phylogeny reconstruction:

The peaks from all species were merged and collapsed in multiple alignment space to generate a non-redundant set of all peaks in the clade. The conservation of these peaks was assessed as described above. The phylogenetic tree was inferred by treating the presence/absence of each peak as a 2-state discrete character, sequentially using the MIX and PARS program of the PHYLIP suite according to the recommendations of the software documentation (http://evolution.genetics.washington.edu/phylip.html).

Sequence conservation:

Per-base conservation scores were computed by running the phastCons and phyloP programs of the PHAST suite v1.1 on the MultiZ alignment according to the recommendations of the software documentation (http://compgen.bscb.cornell.edu/phast). Depending on the subclade of interest, some species

were excluded from the alignment for certain analyses. Pre-computed phastCons scores for the full 15-way alignment were downloaded from UCSC (http://hgdownload.soe.ucsc.edu/goldenPath/dm3/-phastCons15way).

Core promoter motifs:

For analyses of motif composition, we only considered *D. melanogaster* TSCs that were functionally conserved across all 5 species. We used pairwise chained alignments downloaded from UCSC (http://hgdownload.soe.ucsc.edu/downloads.html#fruitfly) to align the most heavily used position of each TSC (i.e., the main TSS) to all other genomes. Peaks for which the maximum position could not be aligned to all genomes were excluded from the analysis. A custom script was used to search for matches to previously characterized core promoter motifs (FitzGerald et al. 2006) within a 301-bp window centered on the main TSS. Consensus sequences for sets of peaks with matches to individual motifs were computed using MEME v4.9.0 (http://meme.nbcr.net/meme).

Time series alignment:

Z-score transformed gene expression time series from all species were registered to one another using the GTEM suite according to the recommendations of the software documentation (http://flydev.berkeley-.edu/cgi-bin/GTEM/index.htm). One-to-one orthology calls from Flybase (2012 release 2) were used to match gene expression profiles between species. We pre-processed pairs of datasets (*D. melanogaster* and another species) to compensate for differences in annotation quality and peak-calling between species. We identified orthologs of TSCs that had detectable expression (at least 10 tags) but initially failed to be called in one species. In addition, when a functionally conserved TSC had been attributed to an annotated gene in one species but not the other, we corrected this discrepancy by attributing it to the gene in both species. For the *D. ananassae* dataset, the 8^{th} time point failed to yield acceptable data, and was excluded from the analysis. All time series were upsampled 5-fold and smoothed with a 2-hour window size using RZ-Smooth v4.1. Optimal global alignment paths between *D. melanogaster* and the other datasets were computed with T-Warp v3.2 with Pearson distance matrices (3-hour window). M-Align v2.8 was used to align each series to the *D. melanogaster* reference and smooth the final aligned series (1-hour window). The expression profiles of individual TSCs were registered to one another with M-Align, using the optimal alignment path computed for gene expression profiles. Prior to alignment,

we used the UCSC liftOver tool (http://hgdownload.cse.ucsc.edu/admin/exe/linux.x86_64/liftOver) to identify *D. melanogaster* TSCs that aligned well (at least 50% of bases aligned) to all other genomes. The temporal expression profiles of those orthologous genomic positions only were aligned.

Expression profile & Coding sequence conservation:

We measured the conservation of individual expression profiles (TSCs or genes) across a clade as the average Pearson R^2 for all pairwise comparisons of species within the clade. For each *D. melanogaster* protein-coding gene, we considered all genomic positions annotated as coding sequence and computed the average 15-way phastCons score over all these positions.

Global interspecies expression divergence:

In order to estimate the overall divergence between datasets, we first performed principal component analysis (PCA) jointly on all 5 time series using R. Only genes with one-to-one orthologs in all 5 species were included. For further analysis, we only considered gene expression in the space of the first 3 components. To account for small inaccuracies, the PCA-transformed time series were aligned to each other a second time using GTEM (see above), with very minor consequences. We defined the consensus developmental path across the 5 species as the average expression path in the space of the first 3 components. We defined the clade-wide divergence at any time point as the average Euclidean distance of all species paths to the consensus path.

Other software:

Custom analysis scripts were written in Python 2.7 (http://www.python.org). R was used for plotting (http://www.r-project.org).

Chapter 5

Discussion

In the continuity of recent studies of regulatory evolution, the purpose of the work presented here was to explore transcriptome complexity and evolutionary divergence in a developmental context. We focused on promoters, which play a central role in gene regulation, but the evolutionary dynamics of which have not been characterized in depth.

In order to identify promoters and quantify their transcriptional output we developed RAMPAGE, a TSS detection method based on massively parallel sequencing of 5'-complete cDNAs. Through the development of both experimental and computational procedures, this approach features substantial improvements relative to existing techniques. These include increased specificity for TSSs, the ability to sequence medium-sized cDNAs and thus partially characterize transcript structure, and significant streamlining of the library preparation protocol. In contrast to shotgun RNA-seq, this method directly measures the contribution of individual promoters to the transcriptome, and is thus ideally suited for the characterization of the expression specificity of these unitary regulatory elements. As such, we hope it will in the future be useful for other expression profiling studies in which the primary focus is on transcriptional regulation. We have recently started profiling promoter activity in human tissues as part of the ENCODE consortium, and we plan to progressively scale up these efforts. In order to better annotate transcripts based on high-throughput data, we have also been working with collaborators in Peter Bickel's group at UC Berkeley on the joint analysis of RAMPAGE and RNA-seq data for *ab initio* transcript modeling.

Current high-throughput sequencing platforms only offer limited read lengths, but as those are improved and third-generation sequencers become available, we anticipate that it will soon become practical to undertake full-length cDNA sequencing on a large scale (Sharon et al. (2013)). We modified our protocol to prepare full-length cDNA libraries suitable for sequencing on the only long-read platform

available at this point, which is commercialized by Pacific Biosciences (see Appendix 2). Although improvements still need to be made, preliminary results are very promising: most cDNAs sequenced are indeed full-length, and accurately match established transcript annotations. We will be working on optimizing our library preparation protocol in the near future. Additionally, an ongoing collaboration with the Schatz lab at CSHL aims at improving analytical methods to compensate for the current shortcomings of the PacBio platform. Importantly, the data analysis strategy that I have developed for RAMPAGE would greatly benefit from longer reads.

To gain a better understanding of transcriptome complexity and regulation in a developmental setting, I profiled promoter activity throughout the life cycle of *D. melanogaster*. This work revealed a significant influence of transposable elements on gene regulation: indeed, over 1,300 transposon-derived promoters were found to drive the expression of protein-coding and non-coding transcripts. The observed expression of transposons in family-specific patterns is in agreement with previous studies (Graveley et al. (2010)), but the formal demonstration that they themselves bear active promoters and that they have been co-opted to regulate the expression of numerous host protein-coding genes represents, to the best of our knowledge, a notable advance. Importantly, promoter expression profiling in other *Drosophila* species has led to qualitatively similar conclusions, although the prevalence of the phenomenon depends on the transposon content of individual genomes.

The scope of transposon domestication in *Drosophila* is probably narrower than it seems to be in mammals, and the population dynamics of repeated sequences are much different between the two clades. Therefore we look forward to comparing our current work with similar surveys in mammals, and in particular in human. Recent comparative genomics analyses have identified literally hundreds of thousands of transposon-derived conserved elements in the human genome (Lindblad-Toh et al. (2011)), and there is now a pressing need to make progress on their functional annotation. Furthermore, a number of studies have suggested that transposon domestication is still currently an active process in mammals – therefore it is likely that many more elements that are not conserved between species have driven lineage-specific innovation (Lindblad-Toh et al. (2011), Schmidt et al. (2012), Bourque et al. (2008)). We expect that further functional work in human and other mammals will yield important new insights into the peculiar relationship between transposons and their host genomes.

In order to better understand the evolution of developmental transcriptomes, we compared promoter expression profiles at high temporal resolution throughout embryonic development in 5 *Drosophila*

species. This analysis revealed a dynamic remodeling of genome-wide promoter landscapes throughout the evolution of the clade, suggesting that promoter gain and loss play a substantial role in the evolution of gene regulation. Curiously, the syntactic rules of promoter design do not seem to be set in stone, and subtle changes in the overall composition of core promoters can be observed at moderate evolutionary distances. Despite significant plasticity in the genomic organization of regulatory elements, the developmental expression dynamics of individual genes appear to be under substantial selective pressure. However, this general trend does not by any means apply uniformly to the whole genome. Selective pressures on expression specificity are shaped by gene function, as well as system-level constraints on developmental stages.

Arguably, the most significant finding of this comparative study is the prevalence of non-coding transcription during embryonic development, and the deep evolutionary conservation of promoters driving that transcription. Although more controls are needed to assess the independence of these transcripts from protein-coding loci, as well as their own protein-coding potential, we identified over a thousand promoters driving the expression of putative long non-coding RNAs that have been conserved for over 25 million years. Those promoters display hallmarks of strong purifying selection, both at the level of genomic sequence and at the level of expression specificity. If these results can be confirmed, they would point to a role of non-coding transcription in crucial stages of development that is much more far-reaching than had been anticipated. We are looking forward to carrying out further characterization of the transcripts identified and, if possible, we are hoping to conduct loss-of-function experiments to test the implication of individual candidates in the control of embryonic development. We believe that the formal demonstration of a role of non-coding transcription in such a fundamental process would represent a significant step forward.

Beyond this question, we are also planning on investigating whether the emergence of individual non-coding transcription units might have played an adaptive role in the evolution of embryonic development in *Drosophila*. Using the data we have generated, it should be possible to identify promoters that represent evolutionary innovations in the *melanogaster* subgroup (that is, promoters not present in either *D. pseudoobscura* or *D. ananassae*), and that have come under purifying selection in that subclade. Using the same methods we have used so far, we can assess the intensity of recent negative selection on the sequence and expression patterns of these *melanogaster* subgroup-specific promoters. The identification of recently evolved functional promoters would provide a strong vindication of the

hypothesis that non-coding transcription plays a role in lineage-specific adaptation. Again, functional assays to investigate the potential role of recently evolved lncRNAs in embryonic development would provide invaluable insights, and we are hoping to be able to carry out such experiments in the near future.

Ultimately, we speculate that it should be possible to transition from such candidate-driven approaches to more general loss-of-function screens. Although the construction of transgenic short hairpin RNA (shRNA) libraries targeting all lncRNAs identified in this study would represent a significant amount of work, it does not seem unreasonable for a well-established laboratory to undertake such an endeavor. Along with the implementation of medium-throughput assays to monitor development in *D. melanogaster*, such an approach would permit a general assessment of the precise role of non-coding transcription in embryonic development in *Drosophila*. We look forward to seeing whether such experiments actually become a reality, as they would provide a great opportunity to test the hypotheses we have put forward here.

Overall, we hope that this work will have contributed somewhat to our understanding of regulatory evolution, and to the investigation of the roles of non-coding transcription in metazoan development. Further work in *Drosophila*, as well as parallel efforts in mammals, will put our hypotheses to the test and assess their generality across diverse organisms.

Appendix 1: Detailed RAMPAGE Protocol

This protocol was originally published in Current Protocols in Molecular Biology under the title: ”RAMPAGE: Promoter Activity Profiling by Paired-End Sequencing of 5’-Complete cDNAs”

RAMPAGE: Promoter Activity Profiling by Paired-End Sequencing of 5′-Complete cDNAs

Philippe Batut[1] and Thomas R. Gingeras[1]

[1]Watson School of Biological Sciences, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York

ABSTRACT

RNA annotation and mapping of promoters for analysis of gene expression (RAMPAGE) is a method that harnesses highly specific sequencing of 5′-complete complementary DNAs to identify transcription start sites (TSSs) genome-wide. Although TSS mapping has historically relied on detection of 5′-complete cDNAs, current genome-wide approaches typically have limited specificity and provide only scarce information regarding transcript structure. RAMPAGE allows for highly stringent selection of 5′-complete molecules, thus allowing base-resolution TSS identification with a high signal-to-noise ratio. Paired-end sequencing of medium-length cDNAs yields transcript structure information that is essential to interpreting the relationship of TSSs to annotated genes and transcripts. As opposed to standard RNA-seq, RAMPAGE explicitly yields accurate and highly reproducible expression level estimates for individual promoters. Moreover, this approach offers a streamlined 2- to 3-day protocol that is optimized for extensive sample multiplexing, and is therefore adapted for large-scale projects. This method has been applied successfully to human and *Drosophila* samples, and in principle should be applicable to any eukaryotic system. *Curr. Protoc. Mol. Biol.* 104:25B.11.1-25B.11.16.

Keywords: transcription start site • promoter • RAMPAGE • high-throughput sequencing • expression profiling

INTRODUCTION

This unit presents a protocol for RNA Annotation and Mapping of Promoters for Analysis of Gene Expression (RAMPAGE), a method for genome-wide identification of transcription start sites (TSSs) and quantification of promoter activity (Batut et al., 2013). RAMPAGE is based on synthesis of 5′-complete cDNAs from eukaryotic total RNA samples, and their sequencing on Illumina high-throughput platforms.

Previous methods for high-throughput sequencing of 5′-complete cDNAs have failed to achieve high specificity for TSS identification, and often provide only scarce sequence information in the form of 20- to 30-base "tags" (Ni et al., 2010; Valen et al., 2009). This makes their alignment to reference genomes problematic (especially for studying repeat sequences), and yields no information regarding transcript structure. This is a major pitfall, as transcript connectivity is essential to revealing the nature of products transcribed from individual promoters. Transcript connectivity is also key to understanding relationships between functionally related elements, such as alternative promoters. The RAMPAGE approach achieves greatly increased TSS specificity through the combination of two orthogonal enrichment strategies: template switching (Hirzmann et al., 1993) and cap trapping (Carninci et al., 1996). Template switching makes use of unique properties of certain reverse-transcriptase enzymes to add adaptor sequences to the end of

Current Protocols in Molecular Biology 25B.11.1-25B.11.16, October 2013
Published online October 2013 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/0471142727.mb25b11s104

Figure 5.1

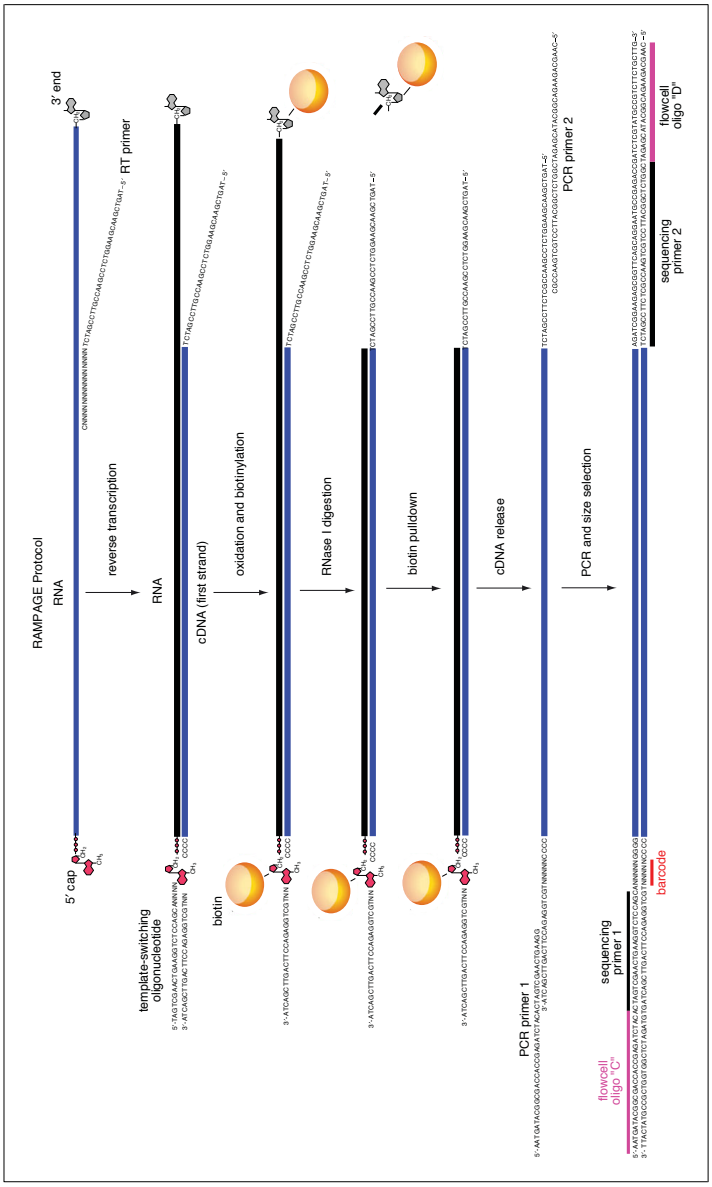

Figure 25B.11.1 Preparation of RAMPAGE library. Ribosome-depleted RNA is reverse-transcribed with random primers bearing an Illumina adaptor sequence overhang. Under the conditions used, the reverse transcriptase will often add a few non-templated Cs when it reaches the 5′ end of the template, especially if the template is capped. A template-switching oligo (TSO), which has three riboguanosines at its 3′ end, can hybridize to the terminal Cs, prompting the enzyme to switch templates and add the TSO sequence to the end of the newly synthesized cDNA. Since the TSO bears the other Illumina adaptor sequence, resulting 5′-complete cDNAs are amplifiable, whereas non-5′-complete molecules are not. The next steps implement the cap-trapping strategy, in which riboses with free 2′- and 3′-hydroxyl groups are oxidized and biotinylated, and single-stranded portions of RNA are digested by RNase I. This leaves biotin groups at only the 5′ ends of capped transcripts hybridized to 5′-complete cDNAs, which can then be recovered on streptavidin-coated beads. After PCR amplification and size selection, the cDNAs selected by these two orthogonal strategies can be directly sequenced on Illumina platforms.

Figure 5.2

5′-complete cDNAs, while cap trapping is based on biotinylation and pulldown of capped RNA molecules and their associated 5′-complete cDNAs. The method is illustrated in Figure 25B.11.1 and described in detail in Basic Protocol 1. Analysis is described in Basic Protocol 2.

PREPARATION OF 5′-COMPLETE cDNAs FOR PAIRED-END SEQUENCING

BASIC PROTOCOL 1

The template-switching step of this protocol makes use of a set of 40 six-base barcodes designed to have GC contents between 20% and 80% and a minimum Hamming distance (i.e., number of differing positions) of three between any two barcodes in the set. The latter requirement ensures that barcodes with one sequencing error can still be unambiguously identified, thus maximizing the proportion of barcodes recovered while minimizing the risk of barcode misassignment. The sequences of the 40 barcoded oligos are listed in Table 25B.11.1. Note that, in order for template-switching to occur, the last three residues of all TSOs must be riboguanosines (rG; Zhu et al., 2001).

The addition of barcodes early in the workflow allows for very efficient multiplexing by allowing most of the procedure to be performed on large pools of samples. This streamlined protocol permits completion of the full procedure in 2 to 3 days. The resulting libraries are suitable for paired-end sequencing on Illumina platforms (GA_{II}, HiSeq, MiSeq). The length of sequences is limited only by the capabilities of the platform.

NOTE: All synthetic oligonucleotides were synthesized on a 100-μmol scale (IDT) and purified by standard desalting, unless otherwise specified. It is best to order TSOs in batches of 250 nmol.

Materials

DNaseI-treated total RNA
Terminator (TEX) enzyme with buffer A (Epicentre, cat. no. TER51020)
Molecular-biology grade water (Sigma-Aldrich, cat. no. 95284-100ML)
Agencourt RNAClean XP kit (Beckman Coulter, cat. no. A63987)
70% (v/v) ethanol, freshly prepared
Reverse transcription (RT) primer:
  400 μM rampage_RT:
  5′-TAGTCGAACGAAGGTCTCCGAACCGCTCTTCCGATCT$(N)_{15}$
Template-switching oligonucleotides (TSOs, Table 25B.11.1):
  4 mM rampage_TS_ :
  5′- TAGTCGAACTGAAGGTCTCCAGCA***NNNNNN***rGrGrG
SuperScript III reverse transcriptase (Invitrogen, 200 U/μl, cat. no. 18080-085), with first-strand buffer and 100 mM DTT
10 mM dNTP mix (Invitrogen, cat. no. 18427-013)
Sorbitol/trehalose solution (see recipe)
5 M betaine (Sigma-Aldrich, cat. no. B0300-1VL)
qPCR primers:
  10 μM CAGEscan-erF:
  5′-AATGATACGGCGACCACCGAGATCTACACTAGTCGAACTGAAGG
  10 μM CAGEscan-erR:
  5′-CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTG AACCGCTCTTCCGATCT
Power SYBR Green premix (Applied Biosystems, cat. no. 4367659)
Sodium periodate ($NaIO_4$, ≥ 99.8%, Sigma-Aldrich, cat. no. 311448-5G)
1 M sodium acetate (NaOAc), pH 4.5: prepare from commercial 3 M NaOAc, pH 5.5 (Ambion, cat. no. AM9740)
40% (v/v) glycerol (Sigma-Aldrich, cat. no. G5516-100ML)

Figure 5.3

1 M Tris-Cl, pH 7.0 and 8.5: prepare from commercial pH 7.4 stock (Sigma-Aldrich, cat. no. T2194–100ML) by adjusting pH with HCl or NaOH
Biotin hydrazide, long arm (Vector Laboratories, cat. no. SP-1100)
1 M sodium citrate, pH 6.0 (Sigma-Aldrich, cat. no. S1804-500G)
0.5 M EDTA, pH 8.0 (Ambion, cat. no. AM9260G)
5 to 10 U/μl RNase I (Promega, cat. no. M4261)
10 mg/ml MPG streptavidin beads (PureBiotech, cat. no. MSTR0502)
E. coli tRNA, DNA and protein free (see Support Protocol)
Wash buffers 1 to 4 (see recipes)
10 M NaOH (Sigma-Aldrich, cat. no. 72068-100ML)
Agencourt AMPure XP kit (Beckman Coulter, cat. no. A63881)
Ex Taq Hot Start (HS) polymerase with buffer and 2.5 mM dNTP mix (Clontech, cat. no. RR006A)
Sequencing primers:
rampage_r1 (custom primer):
5′- TAGTCGAACTGAAGGTCTCCAGCA
SBS8 (standard Illumina primer):
5′- CGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT

200-μl high-recovery PCR tubes (Axygen, cat. no. PCR-02-L-C)
Thermal cycler (BioRad)
4 C cold block
Quantitative PCR system (e.g., Applied Biosystems 7300 real-time PCR system)
Magnet for bead separation
Bioanalyzer (Agilent)
Bioanalyzer RNA Nano chip kit (Agilent)
Optical 96-well, 200-μl qPCR plates with covers (Applied Biosystems, cat. nos. N801-0560 and 4311971)
1.7-ml high-recovery microcentrifuge tubes (Axygen, cat. no. MCT-175-L-C)
Bioanalyzer DNA High-Sensitivity chip kit (Agilent)

Additional reagents and equipment for sequencing on an Illumina platform

Degrade 5′-monophosphate RNAs (terminator digest)

1. For each TSO, place 5 μg DNaseI-treated total RNA in a 10-μl volume in a high-recovery 200-μl PCR tube (or 96-well plate if processing many samples).

   Each of the 40 TSOs is processed separately in steps 1–15 (40 TEX digests, 40 RT reactions). These are pooled into a single library at step 16. When pooling n libraries (n × 5 μg starting material), it is advised to not exceed a total of 50 μg starting material.

2. Denature 5 min at 65 C in a thermal cycler and place immediately in a cold block cooled to 4 C on ice for 2 min (or in an ice-water bath).

   Proper denaturation is important for degradation efficiency, as secondary structures can protect transcripts from digestion.

   This protocol has been successfully performed using as little as 2 to 3 μg input per sample when pooling multiple libraries after the reverse-transcription step. The quality of the RNA should be checked at that point by running on a Bioanalyzer RNA Nano chip. RNA integrity is absolutely crucial to generation of high-quality libraries.

3. Prepare a 20-μl digestion mix by adding:

   2 μl TEX buffer A
   3 μl 1 U/μl TEX
   5 μl H_2O.

Figure 5.4

Table 25B.11.1 Template-Switching Oligonucleotides (TSOs)[a]

Name	Sequence (5′ to 3′)
rampage_TS_01	TAGTCGAACTGAAGGTCTCCAGCA**AGGTAA**rGrGrG
rampage_TS_02	TAGTCGAACTGAAGGTCTCCAGCA**CACTAC**rGrGrG
rampage_TS_03	TAGTCGAACTGAAGGTCTCCAGCA**TTGGTC**rGrGrG
rampage_TS_04	TAGTCGAACTGAAGGTCTCCAGCA**GTGTCA**rGrGrG
rampage_TS_05	TAGTCGAACTGAAGGTCTCCAGCA**GCCGAA**rGrGrG
rampage_TS_06	TAGTCGAACTGAAGGTCTCCAGCA**TGATCA**rGrGrG
rampage_TS_07	TAGTCGAACTGAAGGTCTCCAGCA**CTGTAT**rGrGrG
rampage_TS_08	TAGTCGAACTGAAGGTCTCCAGCA**CGACTG**rGrGrG
rampage_TS_09	TAGTCGAACTGAAGGTCTCCAGCA**TTCCAG**rGrGrG
rampage_TS_10	TAGTCGAACTGAAGGTCTCCAGCAA**CTCTT**rGrGrG
rampage_TS_11	TAGTCGAACTGAAGGTCTCCAGCA**GGATAC**rGrGrG
rampage_TS_12	TAGTCGAACTGAAGGTCTCCAGCA**TTAACG**rGrGrG
rampage_TS_13	TAGTCGAACTGAAGGTCTCCAGCA**GAGTGC**rGrGrG
rampage_TS_14	TAGTCGAACTGAAGGTCTCCAGCAA**AGGAC**rGrGrG
rampage_TS_15	TAGTCGAACTGAAGGTCTCCAGCA**CGCGTT**rGrGrG
rampage_TS_16	TAGTCGAACTGAAGGTCTCCAGCAA**TGCGT**rGrGrG
rampage_TS_17	TAGTCGAACTGAAGGTCTCCAGCAA**TAAGC**rGrGrG
rampage_TS_18	TAGTCGAACTGAAGGTCTCCAGCAA**TCTCC**rGrGrG
rampage_TS_19	TAGTCGAACTGAAGGTCTCCAGCA**TAACTC**rGrGrG
rampage_TS_20	TAGTCGAACTGAAGGTCTCCAGCAA**TAGAG**rGrGrG
rampage_TS_21	TAGTCGAACTGAAGGTCTCCAGCAA**GCCTA**rGrGrG
rampage_TS_22	TAGTCGAACTGAAGGTCTCCAGCA**TGTAGT**rGrGrG
rampage_TS_23	TAGTCGAACTGAAGGTCTCCAGCAA**AAACGG**rGrGrG
rampage_TS_24	TAGTCGAACTGAAGGTCTCCAGCA**CCTACG**rGrGrG
rampage_TS_25	TAGTCGAACTGAAGGTCTCCAGCAA**CTAGA**rGrGrG
rampage_TS_26	TAGTCGAACTGAAGGTCTCCAGCA**CCCTCT**rGrGrG
rampage_TS_27	TAGTCGAACTGAAGGTCTCCAGCA**GGTATA**rGrGrG
rampage_TS_28	TAGTCGAACTGAAGGTCTCCAGCA**GATCCC**rGrGrG
rampage_TS_29	TAGTCGAACTGAAGGTCTCCAGCA**CAATGT**rGrGrG
rampage_TS_30	TAGTCGAACTGAAGGTCTCCAGCA**GCGTTG**rGrGrG
rampage_TS_31	TAGTCGAACTGAAGGTCTCCAGCAA**ACTGA**rGrGrG
rampage_TS_32	TAGTCGAACTGAAGGTCTCCAGCA**CCAATA**rGrGrG
rampage_TS_33	TAGTCGAACTGAAGGTCTCCAGCA**GCGACT**rGrGrG
rampage_TS_34	TAGTCGAACTGAAGGTCTCCAGCA**GGGGAT**rGrGrG
rampage_TS_35	TAGTCGAACTGAAGGTCTCCAGCA**TCTTCC**rGrGrG
rampage_TS_36	TAGTCGAACTGAAGGTCTCCAGCA**GAACAT**rGrGrG
rampage_TS_37	TAGTCGAACTGAAGGTCTCCAGCA**TCGAAG**rGrGrG
rampage_TS_38	TAGTCGAACTGAAGGTCTCCAGCA**TGCTGC**rGrGrG
rampage_TS_39	TAGTCGAACTGAAGGTCTCCAGCA**CTGCTA**rGrGrG
rampage_TS_40	TAGTCGAACTGAAGGTCTCCAGCA**GACGTG**rGrGrG

[a]Library identification barcodes are indicated in bold, and barcode numbers (01– 40) are included in TSO names.

Figure 5.5

4. Incubate 90 min at 30 C in a thermal cycler.
5. Purify digested RNA using an RNAClean XP kit as follows:
   a. Add 30 µl RNAClean XP bead suspension and mix thoroughly by vortexing or pipetting.
   b. Precipitate 5 min at room temperature.
   c. Place on a magnet for 3 min and carefully remove the supernatant.
   d. Wash twice with 100 µl freshly prepared 70% ethanol.
   e. Air dry for 2 min (without completely drying out the beads).
   f. Elute by adding 7.5 µl H_2O, resuspending beads well by pipetting, and incubating 3 min at room temperature.
   g. Place on magnet until beads are well separated (3 to 5 min) and recover supernatant.
6. Assess extent of ribosomal RNA degradation by running the samples on a Bioanalyzer RNA Nano chip according to manufacturer's instructions.

Reverse-transcribe RNAs

7. Mix reverse-transcription (RT) primers with sample as follows:

   7.5 µl TEX-treated RNA
   1 µl 400 µM rampage_RT
   1 µl 4 mM rampage_TS_ .

8. Denature 10 min at 65 C and immediately place in an ice-cold metal block for 2 min.

   Proper denaturation is important for reverse transcription efficiency, as secondary structures can diminish processivity of the enzyme.

9. Add RT reaction mix (28.4 µl/reaction):

   7.5 µl first-strand buffer
   1.9 µl 10 mM dNTP mix
   7.5 µl sorbitol/trehalose solution
   1.9 µl 100 mM DTT
   5.6 µl 5 M betaine
   4 µl 200 U/µl SuperScript III RT.

10. Incubate in a thermal cycler as follows:

    10 sec at 4 C
    1 min at 22 C
    30 min at 42 C
    15 min at 75 C
    Hold at 4 C.

11. Perform RNAClean XP cleanup as in step 5, using the following volumes:

    65 µl bead suspension
    150 µl 70% ethanol wash
    40 µl H_2O to elute.

    Recovered samples can be stored up to 2 months at – 20 C. However, RNA integrity is still crucial, so all necessary precautions should be taken to prevent degradation.

Figure 5.6

Quantify and pool library (facultative)

As library-specific sequence barcodes are added during RT, it is possible to pool samples at this stage. Accurate quantification of individual libraries is important to ensure equal representation of all libraries in the pool. For sensitivity and accuracy, we favor the following quantitative PCR (qPCR)-based assay to perform this quantification.

12. Prepare 10-fold serial dilutions from 10^{-1} to 10^{-3} for each sample. Prepare a longer series of dilutions (10^{0} to 10^{-5}) for one of the libraries (chosen randomly) to generate a standard curve.

    The standard curve is built from an arbitrarily chosen sample by plotting Ct as a function of the logarithm of its concentration, with the undiluted sample corresponding to 1 unit. The best fit is determined by linear regression, and this is used to compute the concentration of the other samples (relative to this one) given their measured Ct values.

13. Distribute 2 µl of each dilution to duplicate wells of an optical 96-well qPCR plate.

14. Add qPCR mix (18 µl/reaction):

    0.8 µl 10 µM CAGEscan-erF primer
    0.8 µl 10 µM CAGEscan-erR primer
    10 µl Power SYBR Green mix
    6.4 µl H_2O.

15. Run the following qPCR program:

1 cycle:	2 min at 95 C
	10 sec at 55 C
	2 min at 68 C
39 cycles:	15 sec at 95 C
	10 sec at 65 C
	2 min at 68 C.

16. Pool all libraries in equimolar amounts based on qPCR quantification, aiming for a total that corresponds to 30 to 50 µg starting material (total, non-TEX treated RNA).

17. Reduce the total sample volume by RNAClean XP precipitation as in step 11, using a bead-to-sample volume ratio of 1.8:1.

Oxidize 5′-cap

18. Prepare 250 mM $NaIO_4$ solution by dissolving 26.7 mg $NaIO_4$ in 500 µl H_2O.

    Periodate is used to oxidize ribose residues that bear free 2′- and 3′-hydroxyl groups. All riboses in 3′-terminal nucleotides and 5′-cap structures are affected. Ribose residues whose 2′- and 3′-hydroxyls have been oxidized to aldehydes can be biotinylated by reaction with biotin hydrazide.

    This solution is light-sensitive. It should always be prepared fresh and kept covered in aluminum foil on ice.

19. Add 2 µl of 1 M NaOAc, pH 4.5, to the RNA/cDNA solution.

    The pH of this solution is critical.

20. Add 2 µl of 250 mM $NaIO_4$, mix well, and incubate 45 min on ice in the dark (or in foil).

21. Stop reaction by adding 2 µl of 40% glycerol and mixing well by pipetting.

22. Add 14 µl of 1 M Tris-Cl, pH 8.5, and mix well by pipetting.

Figure 5.7

23. Perform RNAClean XP cleanup using the following volumes:

    105 µl bead suspension
    200 µl 70% ethanol
    40 µl H_2O to elute.

Biotinylate 5′-cap

24. Prepare 15 mM biotin solution by dissolving 4.2 mg biotin hydrazide in 750 µl H_2O. Cover with aluminum foil and vortex 20 to 30 min at room temperature.

    Vortexing this long is necessary because biotin does not dissolve well in water. This solution should always be prepared fresh and kept on ice, covered in foil.

25. Add 4 µl of 1 M sodium citrate, pH 6.0, to the oxidized sample, then add 13.5 µl of 15 mM biotin solution and mix well by pipetting.

26. Incubate 14 to 15 hr in the dark (covered with foil) at room temperature.

    No cleanup is required after this step.

Digest with RNaseI

27. Prepare RNaseI mix (per reaction):

    6 µl 1 M Tris-Cl, pH 8.5
    1 µl 0.5 M EDTA, pH 8.0
    5 µl 10 U/µl RNaseI.

28. Add 12 µl RNaseI mix to the biotinylated sample, mix well by pipetting, and incubate 30 min at 37 C.

    When pooling many libraries, the incubation time can be extended to 60 min.

29. Incubate 5 min at 65 C and immediately place on ice for 2 min.

30. Perform RNAClean XP cleanup using the following volumes:

    125 µl bead suspension
    200 µl 70% ethanol
    40 µl H_2O to elute.

Perform streptavidin pulldown (cap trapping)

31. During RNaseI digest and cleanup, prepare magnetic streptavidin beads as follows:

    a. Resuspend beads by vortexing vigorously and transfer 100 µl to a 1.7-ml tube.
    b. Add 1.5 µl of 20 µg/µl *E. coli* tRNA, mix well, and incubate 30 min at room temperature, vortexing every 3 min to resuspend the beads.
    c. Place on magnetic stand for 3 min and remove supernatant.
    d. Add 50 µl wash buffer 1, resuspend well by pipetting, separate on magnetic stand, and remove supernatant. Repeat once.
    e. Resuspend beads in 80 µl wash buffer 1.

    The tRNA must be DNase-treated (see Support Protocol) prior to use.

32. Add 80 µl washed bead suspension to RNaseI-treated sample and incubate 30 min at room temperature, mixing by gentle vortexing every 3 min.

33. Place on magnetic stand for 5 min, then remove and discard supernatant.

34. Add 150 µl wash buffer 1, resuspend by pipetting, place on magnetic stand for 3 min, then remove and discard supernatant.

Figure 5.8

35. Using the same procedures, wash:

    Once with wash buffer 2
    Twice with wash buffer 3
    Twice with wash buffer 4.

    Make sure the supernatant is completely removed after the final wash.

36. Elute sample from beads:

    a. Add 65 µl of 50 mM NaOH and mix well by pipetting.
    b. Incubate 10 min at room temperature, vortexing gently every 2 to 3 min.
    c. Place on magnetic stand for 3 min.
    d. Transfer supernatant to a tube on ice containing 12 µl of 1 M Tris-Cl, pH 7.0.

37. Perform AMPure XP cleanup as described for RNAClean XP (step 5) using the following volumes:

    130 µl bead suspension
    200 µl 70% ethanol wash
    73 µl H_2O to elute.

Amplify by PCR

38. Prepare PCR mix as follows (100 µl/reaction):

    73 µl template
    10 µl *Ex Taq* buffer
    8 µl 2.5 mM dNTP mix
    4 µl 10 µM rampage_F primer
    4 µl 10 µM rampage_R primer
    1 µl 5 U/µl *Ex Taq* HS.

39. Amplify product using the following program:

1 cycle:	75 sec at 95 C
	10 sec at 55 C
	2 min at 68 C
16 cycles:	15 sec at 95 C
	10 sec at 65 C
	2 min at 68 C
1 cycles:	5 min at 68 C.

    It is essential to recover exactly 100 µl from this reaction, as exact volumes and ratios determine the size selection range and efficiency of recovery.

Perform size selection by differential precipitation

40. Perform a first AMPure XP cleanup to precipitate and remove large inserts:

    a. Add 52 µl bead suspension to 100 µl PCR product (0.52:1 ratio).
    b. Precipitate 5 min, place on magnet 3 min.
    c. Transfer supernatant to a new tube and discard the beads.

41. For the second cleanup, prepare a bead-enriched AMPure XP suspension by transferring 80 µl AMPure XP suspension to new tube, placing on the magnet for 3 min, and discarding 60 µl supernatant. Resuspend beads well in the remaining 20 µl.

Figure 5.9

42. Perform second cleanup to precipitate and recover medium inserts, discarding short ones.

    a. Add 18 µl enriched bead suspension to the supernatant from the first precipitation and mix well.
    b. Precipitate 5 min, separate on magnet 3 min.
    c. Remove and discard supernatant.
    d. Wash three times with 300 µl of 70% ethanol.
    e. Air dry until no ethanol remains (without overdrying the beads).
    f. Add 20 µl H_2O and incubate 5 min at room temperature to elute.
    g. Recover supernatant and discard beads.

Quantify and perform quality control

43. Run the final library on a Bioanalyzer High-Sensitivity DNA chip according to manufacturer's instructions for quality control and preliminary quantification. Run undiluted samples as well as 10^{-1} and 10^{-2} dilutions to make sure at least one measurement will fall within the dynamic range of the assay.

    The expected size range is 300 to 1000 bp.

44. Adjust the concentration of the library to 10 nM.

45. Sequence on an Illumina platform (GA_{II}, HiSeq, MiSeq) using the following conditions:

    Paired-end run
    Read length as desired
    Loading concentration as recommended by platform manufacturer
    Sequencing primers:
    Read 1: rampage_r1 (custom primer)
    Read 2: SBS8 (standard Illumina primer).

SUPPORT PROTOCOL

PREPARATION OF tRNA STOCK SOLUTION

tRNA is used to saturate nonspecific RNA interactions with streptavidin-coated beads. The tRNA must be carefully treated with DNase and protease and then purified prior to use.

Materials

E. coli tRNA (type XX, Sigma-Aldrich, cat. no. R1753-500UN)
RQ1 RNase-free DNase with buffer (Promega, cat. no. M6101)
0.5 M EDTA, pH 8.0 (Ambion, cat. no. AM9260G)
10% SDS (Sigma-Aldrich, cat. no. G05030-500ML-F)
Proteinase K (New England Biolabs, cat. no. P8102S)
Agencourt RNAClean XP kit (Beckman Coulter, cat. no. A63987)
70% (v/v) ethanol

1.5-ml microcentrifuge tube
Magnet for bead separation

Perform DNase and protease digestion

1. Dissolve 30 mg tRNA in 400 µl water in a 1.5-ml microcentrifuge tube.

2. Add the following, then incubate 2 hr at 37 C:

    45 µl RQ1 DNase buffer
    30 µl of 1 U/µl RQ1 DNase.

Figure 5.10

3. Add the following, then incubate 30 min at 45 C:

   10 µl 0.5 M EDTA, pH 8.0
   10 µl 10% SDS
   10 µl 10 mg/ml proteinase K.

Purify tRNA

4. Add 900 µl RNAClean XP bead suspension, mix well, and allow to precipitate 5 min at room temperature.
5. Place on a magnet for 5 to 10 min (until solution is clear).
6. Remove and discard the supernatant.
7. Wash three times with 1.8 ml of 70% ethanol. Remove ethanol.
8. Microcentrifuge for several seconds to bring all contents to the bottom of the tube.
9. Place back on magnet for 1 min, then remove any residual ethanol.
10. Air dry for 3 min.
11. Add 1.5 ml water, mix well by pipetting, and incubate 5 min at room temperature to elute tRNA.
12. Place on magnet for 5 to 10 min (until solution is clear).
13. Recover supernatant and store in small aliquots (e.g., 100 µl) up to 1 year at – 20 C.

ANALYSIS OF SEQUENCE DATA FOLLOWING RAMPAGE

BASIC PROTOCOL 2

Based on our experience with RAMPAGE, we designed an integrated data processing workflow that makes extensive use of the unique features of the data to enhance the accuracy and quality of analysis. Basic processing has proven, in our experience, to be an important contributor to the quality of the output of RAMPAGE assays. The following description covers all analysis steps, from raw sequencing data to TSS clusters, expression level estimates, and partial transcript models. For further explanation of the data analysis, see Background Information.

1. *Align cDNAs to the reference genome*. We use STAR software (Dobin et al., 2012) for its speed and accuracy; however, any short-read alignment program capable of spliced alignment of paired-end Illumina data would, in principle, be suitable. The library identification barcode (first 6 bases of read 1) as well as the RT primer sequence (first 15 bases of read 2) must be trimmed off prior to mapping. Since a few G's are added at the very 5′ end during cloning, it is important that the alignment program be able to automatically trim off these non-genomic bases during mapping. The STAR algorithm and others have this capability.
2. *Filter uniquely mapping reads*. The aim of this step is to exclude from the analysis any reads for which the locus of origin cannot be unambiguously determined due to limited sequence information or genomic repeats.
3. *Collapse PCR duplicates*. PCR duplicates must be removed to improve both the specificity of peak-calling and the accuracy of transcript quantification. Collapsing is performed based on full alignment coordinates (start, end, splice sites). To avoid over-collapsing, we use the sequence of the RT primer as a pseudo-random single-molecule barcode. Indeed, since this long oligo often primes RT with mismatches, cDNAs derived from distinct RNA molecules often have different RT primer sequences.

Figure 5.11

4. *Determine density of cDNA 5′ ends at all genomic positions*. For each cDNA sequence, record the genomic position to which the 5′-most base of the cDNA aligns. For the whole dataset, this is best represented as an intensity ("wiggle") file.

5. *Determine coverage by downstream reads at all genomic positions*. Extract the alignment coordinates of all downstream reads from the full alignments. Then, for each genomic position, record the number of downstream reads that cover it. For the whole dataset, this is also best represented as an intensity ("wiggle") file.

6. *Perform peak-calling using data from steps 4 and 5*. We use a sliding window algorithm that, for each position in the genome, assesses the statistical enrichment of 5′ signal within a window surrounding that position. The background distribution used to test significance is a negative binomial (the dispersion parameter can be optimized for each dataset). The coverage by downstream reads in the same window is used to subtract a pseudo-count from the 5′ signal, and thus render significance harder to achieve at highly transcribed exonic positions (see Background Information for further explanation). Neighboring significant windows are fused into peaks, which are then trimmed at the edges down to the first base with signal.

7. *Attribute individual TSS clusters (TSCs) to annotations using data from 3 and 6*. TSCs can be attributed to annotated genes if reads that initiate within them also overlap annotated exons. We usually require two independent (collapsed) cDNAs to support that association. In case of ties (one TSC linked to more than one annotation), all associations supported by 5-fold fewer reads (or less) than the strongest association are removed. Among other things, this filters out spurious associations to downstream genes due to run-off transcription from the appropriate (upstream) gene.

8. *Quantify 5′ end signal over individual TSCs using data from 4 and 6*. Count the number of 5′ tags covering each TSC.

9. *Normalize expression values for sequencing depth*. For this purpose, we consider the "total transcriptome" of interest to be the ensemble of all transcripts in the sample that initiate within any of the TSCs called as statistically significant. Therefore, we normalize the expression value of each TSC to the sum of the expression values of all TSCs. This normalized measure of expression is usually reported in reads per million (rpm).

10. *Reconstruct partial transcript models*. For each TSC, we extract all cDNAs that have their 5′ end within its boundaries (on the same strand), and convert the alignments to BAM format. Each of these bundles of reads is then run through Cufflinks (Trapnell et al., 2010) to generate transcript models.

REAGENTS AND SOLUTIONS

Use molecular-biology-grade or RNase-free water in all recipes and protocol steps. For common stock solutions, see ***APPENDIX 2****; for suppliers, see* ***APPENDIX 4****.*

Sorbitol/trehalose solution, 3.3 M/0.66 M

D(−)-Sorbitol (Wako Pure Chemicals, cat. no. 194-03752)
D(+)-Trehalose dihydrate (Sigma-Aldrich, cat. no. T9531-25G)

Place 2 ml RNase-free water in a 50-ml conical tube. Weigh 8.02 g trehalose directly into the tube. Add 3 ml RNase-free water and mix. Weigh 17.8 g sorbitol directly into the tube. Add another 5.5 ml water and mix. Finally, add water to give a total volume of 30 ml and mix well. Transfer to a 100-ml RNase-free glass bottle and autoclave at 121 C for 30 min. Store 1.5-ml aliquots up to 6 months at room temperature protected from light with aluminum foil.

Figure 5.12

This solution must be prepared precisely as described. It is important that water be added progressively, as it can be difficult to prepare such concentrated solutions accurately.

Wash buffer 1

45 ml 5 M NaCl (Ambion, cat. no. AM9760G, final 4.5 M)
5 ml 0.5 M EDTA, pH 8.0 (final 50 mM)
Store up to several months at room temperature

Wash buffer 2

3 ml 5 M NaCl (Ambion, cat. no. AM9760G, final 0.3 M)
0.1 ml 0.5 M EDTA, pH 8.0 (final 1 mM)
46.9 ml H_2O
Store up to several months at room temperature

Wash buffer 3

0.1 ml 0.5 M EDTA, pH 8.0 (final 1 mM)
2 ml 10% SDS (Sigma-Aldrich, cat. no. G05030-500ML-F, final 0.4%)
25 ml 1 M NaOAc, pH 6.1 (final 0.5 M)
1 ml 1 M Tris·Cl, pH 8.5 (*APPENDIX 2A*; final 20 mM)
21.9 ml H_2O
Store up to several months at room temperature

Wash buffer 4

0.1 ml 0.5 M EDTA, pH 8.0 (final 1 mM)
25 ml 1 M NaOAc, pH 6.1 (final 0.5 M)
0.5 ml 1 M Tris·Cl, pH 8.5 (*APPENDIX 2A*; final 10 mM)
24.4 ml H_2O
Store up to several months at room temperature

COMMENTARY

Background Information

Advantages of RAMPAGE

The detection and mapping of 5′-complete cDNAs has long been the method of choice to identify transcription start sites at high resolution, traditionally by primer extension assays or by cap trapping and Sanger sequencing (Carninci et al., 1996). The advent of high-throughput sequencing platforms has created the opportunity for new methods that could perform the same task on a genome-wide scale. Additionally, massively parallel sequencing allows for transcript quantification through the counting of cDNA fragments, as is done for instance in standard (shotgun) RNA-seq (Wang et al., 2009). Compared to shotgun RNA-seq, however, the 5′-complete cDNA sequencing approach has the critical advantage of explicitly preserving TSS-specific information, thus faithfully delineating the expression profiles of individual promoters.

Other methods for 5′-complete cDNA sequencing have been developed previously (Ni et al., 2010; Plessy et al., 2010; Valen et al., 2009), but provide only limited specificity for TSS detection (Batut et al., 2013). Most of these allow for sequencing of only short sequence tags, which renders mapping to reference genomes very problematic. Moreover, this dearth of sequence information prevents evidence-based assignment of novel promoters to annotated genes. Few protocols allow for paired-end sequencing of medium-sized cDNA fragments (Ni et al., 2010; Plessy et al., 2010), and those protocols offer the poorest TSS specificity (Batut et al., 2013). Additionally, all these protocols are cumbersome and often require large amounts of input material, which makes their application to rare samples and their parallelization problematic (Batut et al., 2013).

With RAMPAGE, one can achieve highly specific 5′-complete cDNA preparation that allows for paired-end sequencing to the full capability of current Illumina platforms. Input material requirements are on the order of 2 to 5 μg of total RNA, which is easily manageable for most samples. Additionally, sample multiplexing greatly improves the throughput

Figure 5.13

of the library preparation process through the addition of sequence barcodes very early in the protocol, which allows almost the entire process to be carried out on large pools of samples in a single tube.

Finally, it is important to stress the portability of this protocol. By substituting the proper adaptor sequences, it should be readily feasible to adapt the method for sequencing on other platforms. This is a notable advantage, as new technologies offering significantly greater read lengths are beginning to emerge.

Recommendations for RAMPAGE data analysis

The most distinctive part of the data analysis pipeline described here is a novel peak-calling algorithm for TSS cluster finding that implements several noise-filtering strategies to improve the ability to discriminate between true TSSs and background signal. As with all biochemical assays, biologically relevant signal must be distinguished from background signals, which may have multiple origins. Moreover, the vast majority of eukaryotic promoters do not display transcription initiation at a single position, but instead allow initiation at multiple sites. The precise length and shape of TSS clusters vary between promoters, from sharp (one or a few nucleotides) to broad (≥ 100 nucleotides) (Carninci et al., 2006). Therefore, previous analyses of 5′-complete cDNA sequencing data have usually made some attempt at grouping individual TSSs into functionally meaningful local clusters (Carninci et al., 2006; Ni et al., 2010; Plessy et al., 2010). Elaborating on this existing work, we devised a novel approach to identify TSCs, which we define operationally as regions of statistically significant clustering of cDNA 5′ ends. Critically, this peak-calling algorithm was designed to make extensive use of paired-end information and to correct for several sources of noise inherent to 5′-complete cDNA sequencing.

Firstly, the null (background) distribution of signal per genomic position is expected to be overdispersed due to at least two technical factors. Failures of reverse transcriptase to processively reach the very 5′ end of its template will be more likely at specific sites of a given transcript (e.g., strong secondary structures), and PCR duplicates generated during the library preparation process can randomly distort the signal at individual positions. Both effects make the raw data seem more "peaky" than the actual landscape of transcription initiation. To attenuate these effects, we make use of an overdispersed statistical distribution (negative binomial) to model background signal, and we remove PCR duplicates from datasets prior to peak-calling. For these purposes, we define PCR duplicates as read pairs that share similar alignment coordinates (start, end, splice sites) and an identical reverse-transcription primer sequence (which we use as a pseudo-random single-molecule barcode).

Secondly, non-5′-complete cDNAs represent another source of background, which will manifest itself mostly over exons. This type of background is complex, because the amount of nonspecific signal depends on transcript abundance. In the absence of an appropriate correction, these artifacts will yield many false-positive TSCs over highly expressed transcripts. Taking advantage of paired-end sequence information, we make use of the fact that coverage by downstream sequencing reads (i.e., the 3′-most portion of our cDNAs) can provide an estimate of transcript abundance at internal (non-TSS) positions. We model background from incomplete cDNAs as linearly proportional to transcript abundance as measured by downstream read coverage. However imperfect this approach might seem, it greatly improves our ability to distinguish between true TSSs and spurious internal signal (Batut et al., 2013).

These specially designed features were incorporated into a sliding window algorithm that scans the whole genome and assesses the significance of local signal enrichment given the null distribution. Downstream read coverage in the same window is used to correct for local transcript abundance by subtracting from the raw 5′ end signal a pseudocount proportional to this coverage. After false discovery rate (FDR) correction using the method of Benjamini and Hochberg (1995), enriched windows in close proximity to each other are merged into peaks, which are subsequently trimmed at the edges down to the first base with signal.

This assay provides extensive information about transcript structure and connectivity, which allows one to connect TSCs to annotated genes based on rigorous experimental determination of cDNA structure. This is a crucial point, since the complex transcriptional architecture of eukaryotic genomes (Kapranov et al., 2007; Djebali et al., 2012) makes promoter-transcript relationships at many loci otherwise difficult to decipher. Additionally, this method takes advantage of the fact that

Figure 5.14

the downstream portions of cDNA inserts are distributed over broad regions of the targets to gain knowledge about medium-range transcript connectivity. Using the current workflow, reads from individual TSCs are processed through Cufflinks to produce partial transcript models.

Critical Parameters and Troubleshooting

Low TSS specificity

Sample quality is crucial to the success of the assay. RNA degradation leads to a higher contribution of non-5′-complete cDNAs in the libraries, thus reducing TSS specificity. RNA quality should be checked by running the samples on Bioanalyzer RNA Nano chips before starting library preparation. Only the highest quality samples should be used.

Poor TSS specificity may also be due to RNA degradation during library preparation, which may be linked to RNase contamination. It is critical that all tubes, pipets, reagents, and work surfaces be RNase-free, and that gloves be worn at all times.

Finally, low TSS specificity could also be caused by incomplete RNase I digestion, which would lead to the capture of non-5′-complete cDNAs through the pulldown of remaining biotin groups at the 3′ end of transcripts. In this case, increasing the RNase I digestion time should be considered.

High ribosomal RNA content

An unusually high rRNA content in the final libraries (>10% to 15% of reads) typically indicates a low-efficiency Terminator (TEX) digest. Analysis of the samples on a Bioanalyzer RNA Nano (or Pico) chip after TEX digest and cleanup should show a near-complete disappearance of rRNA peaks. If rRNA peaks are still prominent, it may be necessary to increase the incubation time for TEX digestion. Proper denaturation of the samples prior to TEX digestion is also critical, as the processivity of this enzyme is affected by secondary structures. Samples should be denatured for a full 5 min at 65 C and then cooled to 4 C within seconds, as renaturation before the addition of enzyme will have adverse effects.

Shift in distribution of library insert size

When preparing libraries from human or *Drosophila* RNA, final product sizes are typically distributed broadly between 300 and 1,000 bp. Samples from other species may yield different insert sizes. A likely cause of strongly skewed distributions in human or *Drosophila* samples is failure of the size selection procedure, which could result from inaccurate pipetting (suspension-to-sample ratios are critical) or from contamination or alteration of the AMPure XP suspension. Alternatively, abnormally short insert sizes could result from RNA degradation before or during library preparation (see Low TSS specificity, above).

Low pass-filtering rate for sequencing lane

We have recently observed low pass-filter rates despite acceptable cluster densities, and this seems to be related to new Illumina analysis pipeline algorithms or parameters. This is likely due to the stretch of three or four G's after the library ID barcode (bases 7-8 of read 1). For these few cycles, all clusters across the flow cell incorporate the same base, and this compromises cluster calling and/or base calling. We have significantly improved the results by spiking in 10% to 15% phiX and using a separate control lane on the flowcell (either phiX control or anything with roughly homogeneous base composition, e.g., genomic DNA, exome).

Anticipated Results

When starting from 5 μg total RNA samples from human or *Drosophila*, one should usually expect a final concentration of 10 to 40 nM (total volume 20 μl), with a broad size distribution spanning 300 to 1,000 bp. In *Drosophila* samples, 70% to 80% of reads should map uniquely to the dm3 reference genome. When compared to Flybase r5.32 annotations, over 90% of uniquely mapped reads (median across annotations) that fall within an annotated transcript are within ± 150 bp of the annotated TSS.

Time Considerations

For few samples that are not pooled, the procedure typically requires 2 full days. All steps through the setup of the biotinylation reaction can be performed on day 1, the biotinylation reaction itself can be incubated overnight, and all other steps can easily be performed on day 2. If processing and pooling many samples, the procedure is more comfortably split over 3 days: all steps through reverse-transcription on day 1, steps from qPCR quantification to setup of the biotinylation reaction on day 2, and all remaining steps on day 3. It is also possible, for convenience, to store the samples at – 20 C after any RNAClean XP or AMPure XP cleanup step.

Figure 5.15

Literature Cited

Batut, P., Dobin, A., Plessy, C., Carninci, P., and Gingeras, T.R. 2013. High-fidelity promoter profiling reveals widespread alternative promoter usage and transposon-driven developmental gene expression. *Genome Res*. 23:169-180.

Benjamini, Y. and Hochberg, Y. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Stat. Methodol*. 57:289-300.

Carninci, P., Kvam, C., Kitamura, A., Ohsumi, T., Okazaki, Y., Itoh, M., Kamiya, M., Shibata, K., Sasaki, N., Izawa, M., Muramatsu, M., Hayashizaki, Y., and Schneider, C. 1996. High-efficiency full-length cDNA cloning by biotinylated CAP trapper. *Genomics* 37:327-336.

Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., Semple, C.A.M., Taylor, M.S., Engstrom, P.G., Frith, M.C., Forrest, A.R.R., Alkema, W.B., Tan, S.L., Plessy, C., Kodzius, R., Ravasi, T., Kasukawa, T., Fukuda, S., Kanamori-Katayama, M., Kitazume, Y., Kawaji, H., Kai, C., Nakamura, M., Konno, H., Nakano, K., Mottagui-Tabar, S., Arner, P., Chesi, A., Gustincich, S., Persichetti, F., Suzuki, H., Grimmond, S.M., Wells, C.A., Orlando, V., Wahlestedt, C., Liu, E.T., Harbers, M., Kawai, J., Bajic, V.B., Hume, D.A., and Hayashizaki, Y. 2006. Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genet*. 38:626-635.

Djebali, S., Davis, C.A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., Xue, C., Marinov, G.K., Khatun, J., Williams, B.A., Zaleski, C., Rozowsky, J., Roder, M., Kokocinski, F., Abdelhamid, R.F., Alioto, T., Antoshechkin, I., Baer, M.T., Bar, N.S., Batut, P., Bell, K., Bell, I., Chakrabortty, S., Chen, X., Chrast, J., Curado, J., Derrien, T., Drenkow, J., Dumais, E., Dumais, J., Duttagupta, R., Falconnet, E., Fastuca, M., Fejes-Toth, K., Ferreira, P., Foissac, S., Fullwood, M.J., Gao, H., Gonzalez, D., Gordon, A., Gunawardena, H., Howald, C., Jha, S., Johnson, R., Kapranov, P., King, B., Kingswood, C., Luo, O.J., Park, E., Persaud, K., Preall, J.B., Ribeca, P., Risk, B., Robyr, D., Sammeth, M., Schaffer, L., See, L.H., Shahab, A., Skancke, J., Suzuki, A.M., Takahashi, H., Tilgner, H., Trout, D., Walters, N., Wang, H., Wrobel, J., Yu, Y., Ruan, X., Hayashizaki, Y., Harrow, J., Gerstein, M., Hubbard, T., Reymond, A., Antonarakis, S.E., Hannon, G., Giddings, M.C., Ruan, Y., Wold, B., Carninci, P., Guigo, R., and Gingeras, T.R. 2012. Landscape of transcription in human cells. *Nature* 489:101-108.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. 2012. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* 29:15-21.

Hirzmann, J., Luo, D., Hahnen, J., and Hobom, G. 1993. Determination of messenger-RNA 5′-ends by reverse transcription of the cap structure. *Nucleic Acids Res*. 21:3597-3598.

Kapranov, P., Willingham, A.T., and Gingeras, T.R. 2007. Genome-wide transcription and the implications for genomic organization. *Nat. Rev. Genet*. 8:413-423.

Ni, T., Corcoran, D.L., Rach, E.A., Song, S., Spana, E.P., Gao, Y., Ohler, U., and Zhu, J. 2010. A paired-end sequencing strategy to map the complex landscape of transcription initiation. *Nat. Methods*. 7:521-527.

Plessy, C., Bertin, N., Takahashi, H., Simone, R., Salimullah, M., Lassmann, T., Vitezic, M., Severin, J., Olivarius, S., Lazarevic, D., Hornig, N., Orlando, V., Bell, I., Gao, H., Dumais, J., Kapranov, P., Wang, H., Davis, C.A., Gingeras, T.R., Kawai, J., Daub, C.O., Hayashizaki, Y., Gustincich, S., and Carninci, P. 2010. Linking promoters to functional transcripts in small samples with nanoCAGE and CAGEscan. *Nat. Methods*. 7:528-534.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol*. 28:511-515.

Valen, E., Pascarella, G., Chalk, A., Maeda, N., Kojima, M., Kawazu, C., Murata, M., Nishiyori, H., Lazarevic, D., Motti, D., Marstrand, T.T., Tang, M.H.E., Zhao, X., Krogh, A., Winther, O., Arakawa, T., Kawai, J., Wells, C., Daub, C., Harbers, M., Hayashizaki, Y., Gustincich, S., Sandelin, A., and Carninci, P. 2009. Genome-wide detection and analysis of hippocampus core promoters using DeepCAGE. *Genome Res*. 19:255-265.

Wang, Z., Gerstein, M., and Snyder, M. 2009. RNA-Seq: A revolutionary tool for transcriptomics. *Nat. Rev. Genet*. 10:57-63.

Zhu, Y.Y., Machleder, E.M., Chenchik, A., Li, R. and Siebert, P.D. 2001. Reverse transcriptase template switching: A SMART approach for full-length cDNA library construction. *Biotechniques* 30:892-897.

Figure 5.16

Appendix 2: Full-length cDNA sequencing on the Pacific Biosciences platform with a modified RAMPAGE protocol

This Appendix features work done in collaboration with two other people in the Gingeras group, whom I gratefully acknowledge. Lei-Hoon See participated in the library preparation process, and Alexander Dobin analyzed the data.

Introduction

Until now, genome-wide transcriptome surveys have only been feasible using microarrays or high-throughput short read sequencing platforms. These technologies, unfortunately, only provide limited information regarding transcript connectivity – that is, the exact combination of sequences included in each individual RNA molecule. However it is becoming clear that alternative splicing, together with alternative promoter and termination site usage, hold out the potential for the synthesis of vast numbers of possible transcripts at individual loci, through the combinatorial use of available sequence modules. Only a subset of these possibilities are likely to be realized, though, and identifying those has become a central issue in transcriptomics. Computational methods have been developed to tackle this issue by attempting to recover connectivity information from short-read data, generally with mixed results (Steijger et al., Nat Methods 2013). Recently, the emergence of novel medium- to high-throughput sequencing platforms offering substantially increased read lengths has made it possible to envision solving this problem experimentally with full-length cDNA sequencing. The only platform available at this point, commercialized by Pacific Biosciences (PacBio), currently offers median read lengths of about 5kb, making it possible to start testing such approaches. Its main limitation appears to be the quality of the sequences produced, with a raw per-base error rate of approximately 15%. This poor quality can be offset by the possibility to generate higher-quality ”circular consensus” (CCS) reads by repeatedly sequencing the same molecule, but as the overall number of bases that can be sequenced in a single run

is limited, this comes at the cost of sequencing only shorter inserts.

We and others have started developing strategies to generate full-length cDNA libraries, sequence them on the Pacific Biosciences platform, and analyze this new type of data. A first medium-scale study was even published recently (Sharon et al. (2013)). However, current library preparation protocols are crude and have limited specificity for full-length molecules, having generally been built upon commercially available kits for classical cloning and Sanger sequencing. They also often have input material requirements that are impractical for many samples. Analytical methods are somewhat inadequate as well, as they generally make use of tools designed for the analysis of high-quality Sanger sequences.

The RAMPAGE protocol offers several features that make it both amenable to and potentially powerful for full-length cDNA preparation. First, this protocol offers very high TSS detection specificity in its current form (partial, 5'-complete cDNA molecules). Second, it may be adapted to generate full-length molecules with very few modifications – perhaps as little as changing the reverse-transcription primer. Third, high-quality 5'-complete libraries can currently be obtained from limited amounts of material (5μg total RNA), raising the possibility that a modified protocol would not require much more. Here I present very encouraging preliminary results obtained for a recent pilot library.

Results

To generate full-length cDNAs, we modified the RAMPAGE protocol by simply using an oligo-dT reverse-transcription primer and slightly shorter PCR primers (see Methods below). We prepared a full-length cDNA library from 10μg of adult female *D. melanogaster* total RNA, and sequenced it on a single SMRT cell. This run yielded 64,629 raw reads, including 31,505 reads with circular consensus. Only the latter were considered for further analysis. The raw consensus read length distribution has a median of 808 bases (Figure 5.17A), which is somewhat shorter than expected. We attribute this in part to the library preparation protocol, as the PCR amplification step likely disfavors longer inserts. The exclusive use of circular consensus reads is also a strong limitation in this regard. We are currently working on addressing these issues (see Discussion).

We mapped the reads to the *D. melanogaster* reference genome using STAR, an aligner developed in-house that is optimized for the mapping of spliced cDNA reads and can withstand moderately

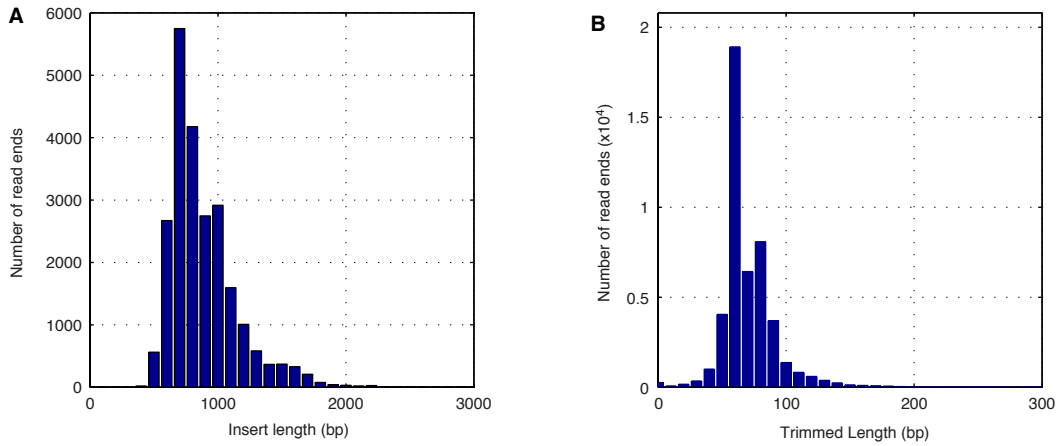

Figure 5.17. PacBio sequencing data quality
Characteristics and mapping of RAMPAGE PacBio reads. (A) Raw read length distribution for all circular consensus reads. Adaptor sequences are included (111 bases total). (B) Number of bases trimmed off each read end during mapping.

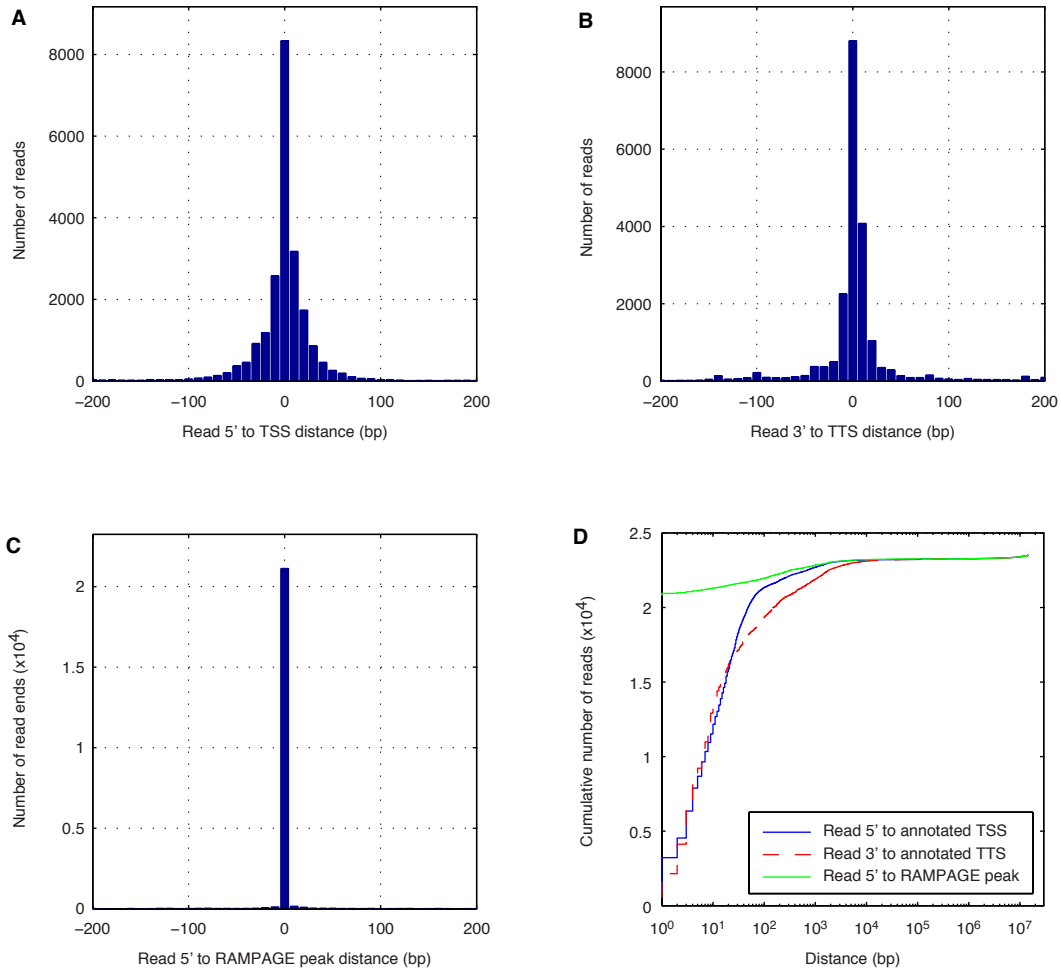

Figure 5.18. Most cDNAs are full-length molecules
(A) Histogram of genomic distances between the 5' end of read alignments and the closest Flybase-annotated transcription start site (TSS). (B) Histogram of genomic distances between the 3' end of read alignments and the closest Flybase-annotated transcription termination site (TTS). (C) Histogram of genomic distances between the 5' end of read alignments and the closest Illumina-based RAMPAGE peak. (D) Cumulative representation of the same 3 distributions.

high sequencing error rates (Dobin et al. (2012)). Approximately 75% of the reads could be mapped to a unique genomic location. Adaptor sequences were not removed prior to mapping, but approximately 60-70 bases were automatically trimmed off each read end (Figure 5.17B), which corresponds roughly to the average length of adaptors. This shows that the cDNA inserts themselves were generally mapped from end to end, with only minor sequence losses.

We assessed the general accuracy of the data with regards to the 5' and 3' ends of transcripts by comparing read alignment coordinates to high-quality Flybase transcript annotations. This indeed revealed very high accuracy at both ends (Figure 5.18A,B,D). To account for annotation insufficiencies, we compared the PacBio reads to TSSs identified for the same sample type using the standard, Illumina-based version of RAMPAGE. The agreement between the two datasets is striking, which confirms that the adaptation to PacBio was successful (Figure 5.18C-D).

Discussion

These preliminary results show that RAMPAGE holds great potential for adaptation to full-length cDNA sequencing. Tests of the protocol on the PacBio platform were promising, judging by the high accuracy of transcript 5' and 3' end mapping. Importantly, the amount of material used here is also an order of magnitude lower than that required by other published protocols.

Hurdles do remain, though. Most importantly, only short- to medium-length sequences were obtained here. We believe this is attributable in part to the PCR amplification step, and we are planning on trying to mitigate this problem by using semi-suppressive PCR. Limiting the analysis to circular consensus reads is also a major impediment to improving read lengths, and we are considering methods to make use of non-consensus reads. The laboratory of Michael Schatz at CSHL has developed methods to correct long low-quality PacBio reads using short high-quality Illumina reads. These methods were initially geared towards genomic DNA sequencing data, and we are now collaborating with the Schatz group to adapt them to cDNA data. This increase in read quality should also improve mapping accuracy.

Concurrently, we are also optimizing methods to precisely identify adaptor sequences at both ends of low-quality cDNA sequences. As the 5' and 3' adaptors are different, this will allow the assignment of individual reads to their genomic strand of origin. It might also slightly improve the precision of the mapping of cDNA ends. Given the low accuracy of the reads, this recognition is somewhat

challenging, and we are considering implementing the Smith-Waterman alignment algorithm for this task.

The results presented here are very preliminary, but they do suggest that with some improvements the RAMPAGE protocol will be suitable for high-accuracy full-length cDNA sequencing from moderate amounts of input material. The optimization of both library cloning protocols and data analysis methods should improve performance in the near future. With additional increases in the accuracy and throughput of third-generation platforms, full-length cDNA sequencing may ultimately replace current transcriptome analysis methods. We note that the data analysis strategy implemented for RAMPAGE (see Chapter 2 & Appendix 1) would only work better with full-length cDNA data.

Methods

Sample collection and RNA extraction: *D. melanogaster* stocks (strain *y; cn b sp*) were maintained on standard cornmeal medium. Approximately 10 5-day-old female adults were collected, and total RNA was extracted using the Agencourt RNAdvance Tissue kit (see Chapter 1).

Library preparation: Libraries were prepared from 10μg of DNase-treated, TEX-digested total RNA, according to the standard RAMPAGE protocol (see Appendix 1) with the following primers. Reverse-transcription primer: 5'- TCCTGCTGAACCGCTCTTCCGATCT$(T)_{20}$VN; Template-switching primer: 5'- TAGTCGAACTGAAGGTCTCCAGCA$(N)_5$rGrGrG; Forward PCR primer: 5'- ACCACCGAGATCTACACTAGTCGAACTGAAGGTCTCCAGC; Reverse PCR primer: 5'- GATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT. PCR amplification was conducted as follows: 95ºC for 75s; 55ºC for 10min; 68ºC for 6min; 15 cycles of (95ºC for 15s; 65ºC for 10s; 68ºC for 3min); 68ºC for 5min. Library quality control and quantification was performed on a Bioanalyzer High-Sensitivity DNA chip. PacBio library preparation was conducted using a commercial PacBio kit according to the manufacturer's instructions. The library was loaded by magnetic loading.

Data analysis: PacBio CCS reads were mapped to Flybase *D. melanogaster* genome and annotations version 5.49 with STAR 2.3.1u compiled for long reads, using the following parameters: –genomeDir Flybase_Dmel5.49 –outFilterMultimapScoreRange 20 –outFilterScoreMinOverLread 0 –outFilterMatchNminOverLread 0.66 –outFilterMismatchNmax 1000 –winAnchorMultimapNmax 200 –seedSearchLmax 30 –seedSearchStartLmax 12 –seedPerReadNmax 100000 –seedPerWindowNmax

100 –alignIntronMax 50000 –alignTranscriptsPerReadNmax 100000 –alignTranscriptsPerWindowNmax 10000 For the TSS/TTS specificity plots, the distance from 5'/3' ends of each read to the nearest TSS/TTS was calculated. For the comparison to RAMPAGE, the distance from the ends of a read to the nearest boundary of all RAMPAGE peaks was calculated, assuming 0 distance for the ends falling inside RAMPAGE peaks. The Illumina RAMPAGE data was the that described in Chapter 2.

Appendix 3: Supplementary figures for Chapter 3

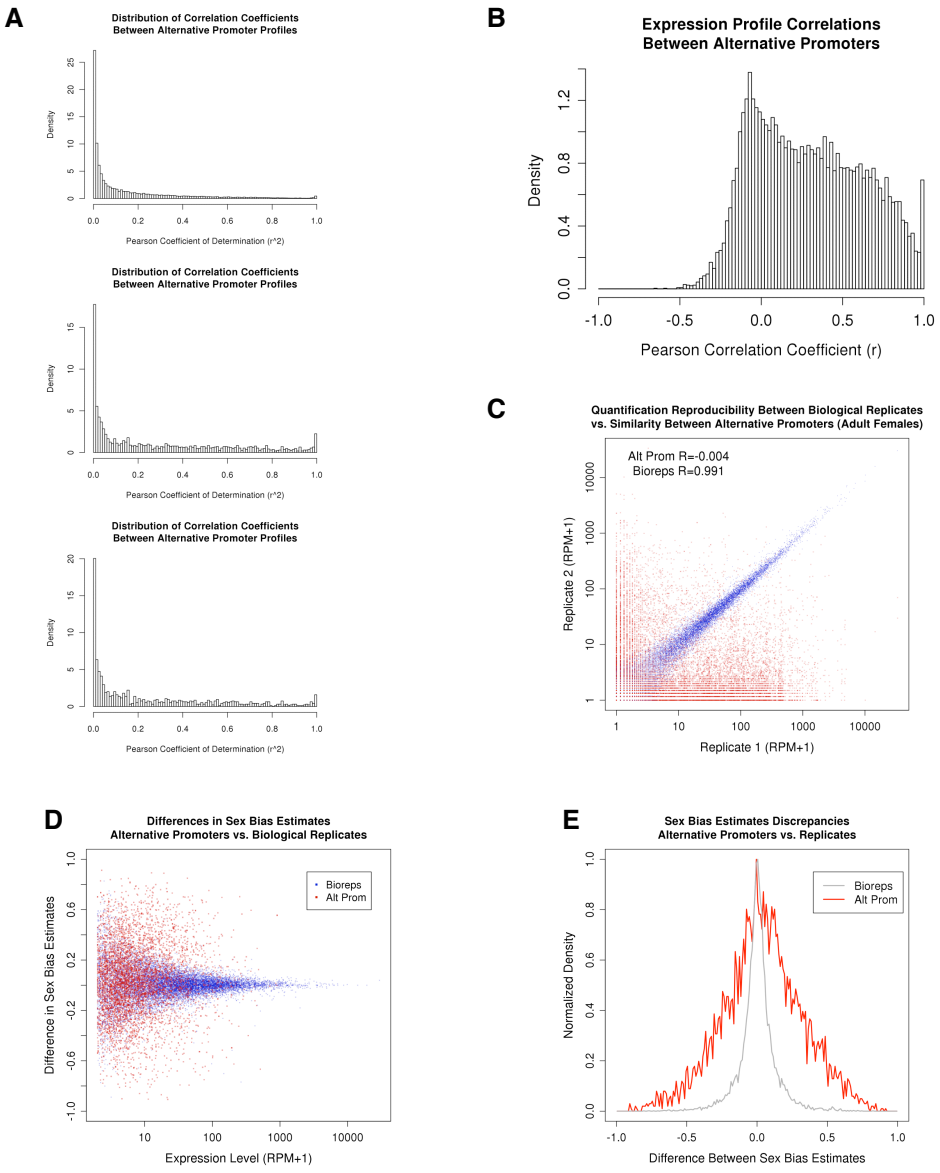

Figure 5.19. Correlations between alternative promoters and Quantification accuracy
(A) Distribution of correlation coefficients (R^2) between alternative promoters considering all genes with maximum expression level at least 1 RPM (top), at least 50 RPM (middle), or maximum expression level between 50 and 400 RPM (bottom). (B) Distribution of correlation coefficients (R) for all genes with expression level at least 5 RPM. (C) Differential expression of alternative promoters (red) compared to reproducibility of promoter expression estimates between biological replicates (blue). (D) Difference in sex bias estimates between alternative promoters (red) or between replicates for the same promoter (blue), as a function of maximum expression level. (E) Histogram of the same data as D, for all genes with maximum expression level at least 1 RPM.

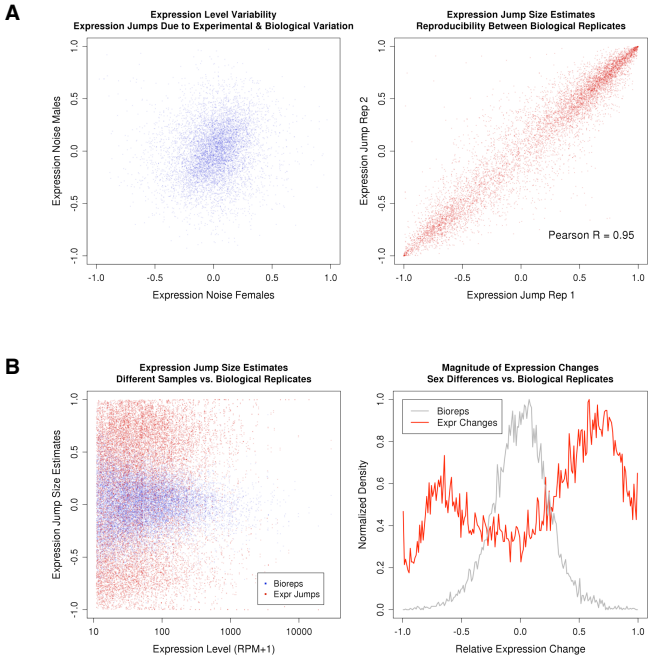

Figure 5.20. Expression level transitions and Quantification accuracy
(A) Lack of correlation of cross-replicate variance between samples (left) and reproducibility of cross-sample expression level transitions between biological replicates (right). (B) Left panel: Expression level transitions observed between samples (red) compared to the reproducibility of expression level transition estimates between replicates (blue), as a function of maximum expression level. Right panel: Same data as in left panel plotted as a histogram, for all promoters with maximum expression level at least RPM.

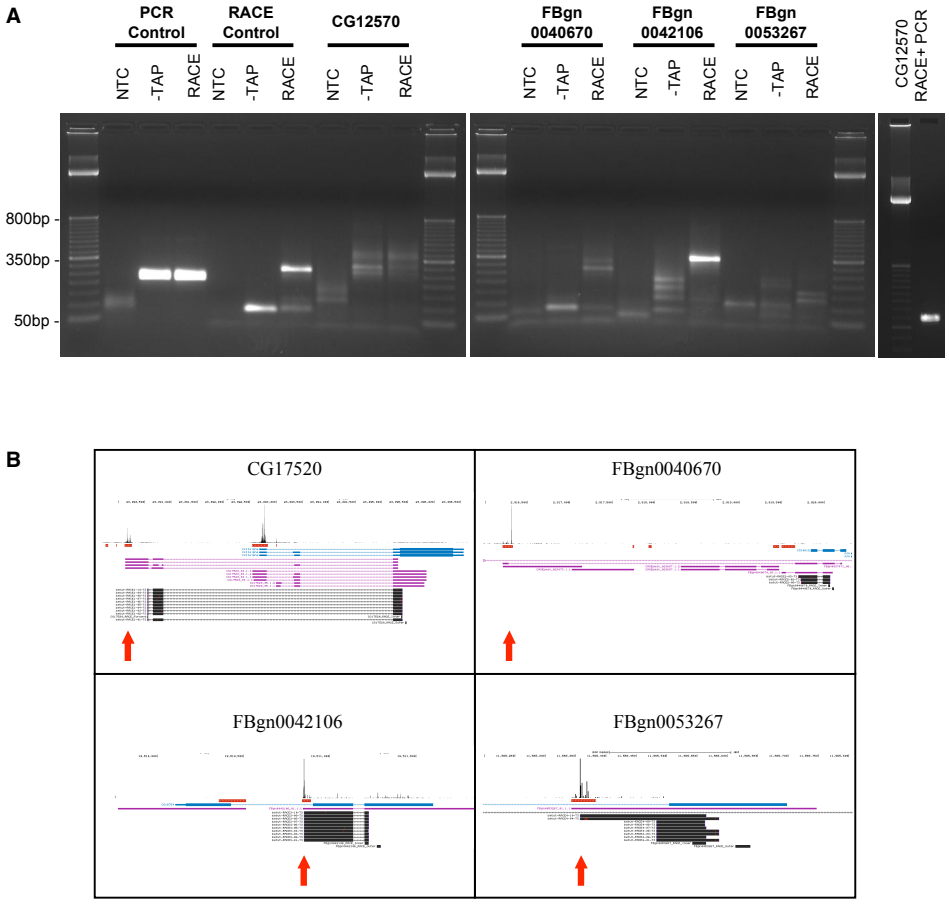

Figure 5.21. Validation of transposon-derived TSCs by 5'-RACE
(A) Agarose gel electrophoresis results showing the 5'-RACE products (NTC: No-template control, -TAP: Control without tobacco acid phosphatase treatment, RACE: 5–RACE). (B) Genome browser screenshots showing, from top to bottom: RAMPAGE tag 5' ends on the positive strand, RAMPAGE peaks (red), Flybase transcript annotations (blue), Cufflinks models from RAMPAGE data (purple), RACE products and primers (black). The target promoters are indicated with red arrows. (C) FBgn0040670 results. Upper panel: Genome browser screenshot showing collapsed RAMPAGE reads mapping to the transposon-derived TSS. Multiple independent reads clearly support transcription of the gene from this TSS. Lower panel: The 3 RACE products (out of 10) that did map to the locus have their 5' ends at different positions, which neither annotations nor independent DNAse-seq data suggest are bona fide TSSs. In contrast, the RAMPAGE-predicted peak is supported by strong DNAse-seq signal. Tracks from top to bottom: RAMPAGE 5' ends (+ strand), RAMPAGE peaks, Flybase transcript annotations, Cufflinks models, 5'-RACE products, RAMPAGE 5' ends (- strand), DNAse-seq data for 4 stages of embryogenesis (from BDTNP).

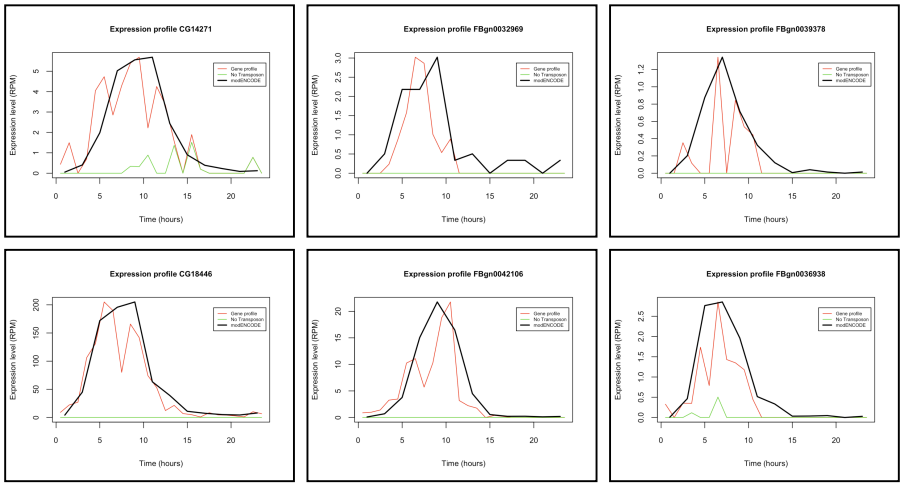

Figure 5.22. *roo*-driven expression is detectable in standard RNA-seq profiles
Comparison of the temporal expression profiles obtained by RAMPAGE or by standard shotgun RNA-seq (modENCODE consortium) for genes with *roo* LTR-derived promoters. For each panel, the RAMPAGE profiles including or not the contribution from the *roo* promoter are shown sider by side with the RNA-seq profile. (All the genes having such a promoter and for which the profiles with and without transposon contribution are clearly distinguishable are show). In all cases, the contribution from the transposon is very clearly reflected in the RNA-seq profile.

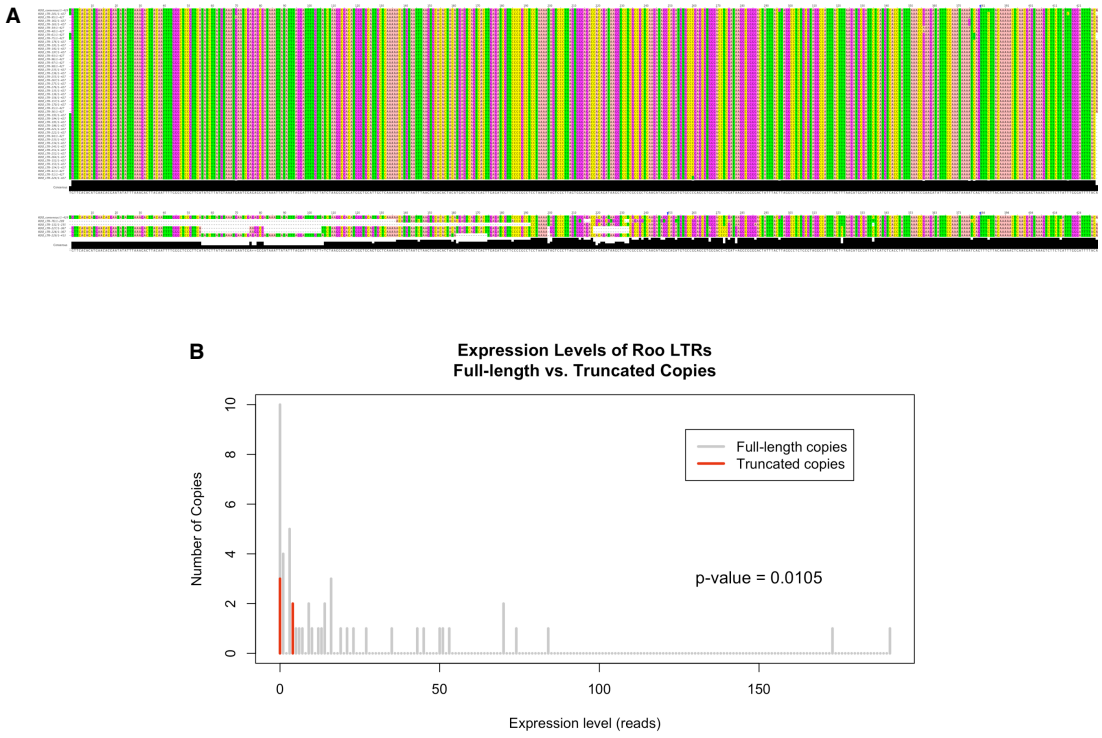

Figure 5.23. Expression of full-length and truncated copies of *roo* LTRs
To test whether the predicted TFBSs are required for the proper expression of *roo* LTRs, we compared the expression patterns of either full-length or truncated individual copies. To reduce the chances of mismapping reads, we excluded from this analysis all copies present on scaffolds U and Uextra or in the heterochomatic portions of any chromosome. Full-length copies are defined as those at least 420 bp long and having a RepeatMasker alignment score greater than 4,000. All copies with length at least 350 bp and score at least 2,500 were aligned to each other using MUSCLE and manually scanned for copies in which at least one of the TFBSs was deleted. For each copy, we report the total number of collapsed uniquely mapping reads. Copies with TFBS deletions display markedly reduced expression compared to full-length ones (p-value = 0.0105, one-tailed permutation test).

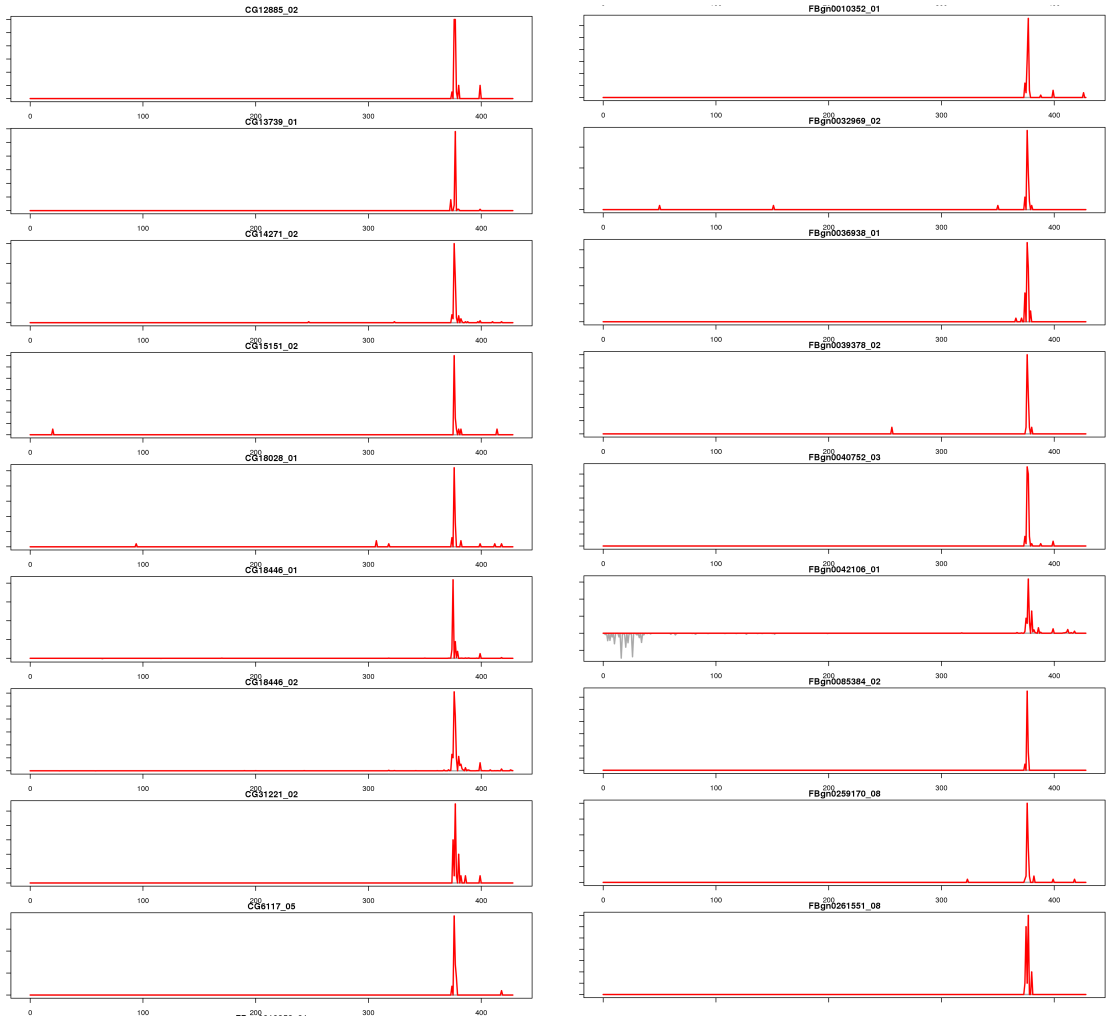

Figure 5.24. Raw RAMPAGE signal over 18 *roo* LTRs driving the expression of protein-coding genes is plotted as a function of the position along the LTR sequence. The coordinates are in the space of the multiple sequence alignment (see Methods).

Appendix 4: Supplementary figures for Chapter 4

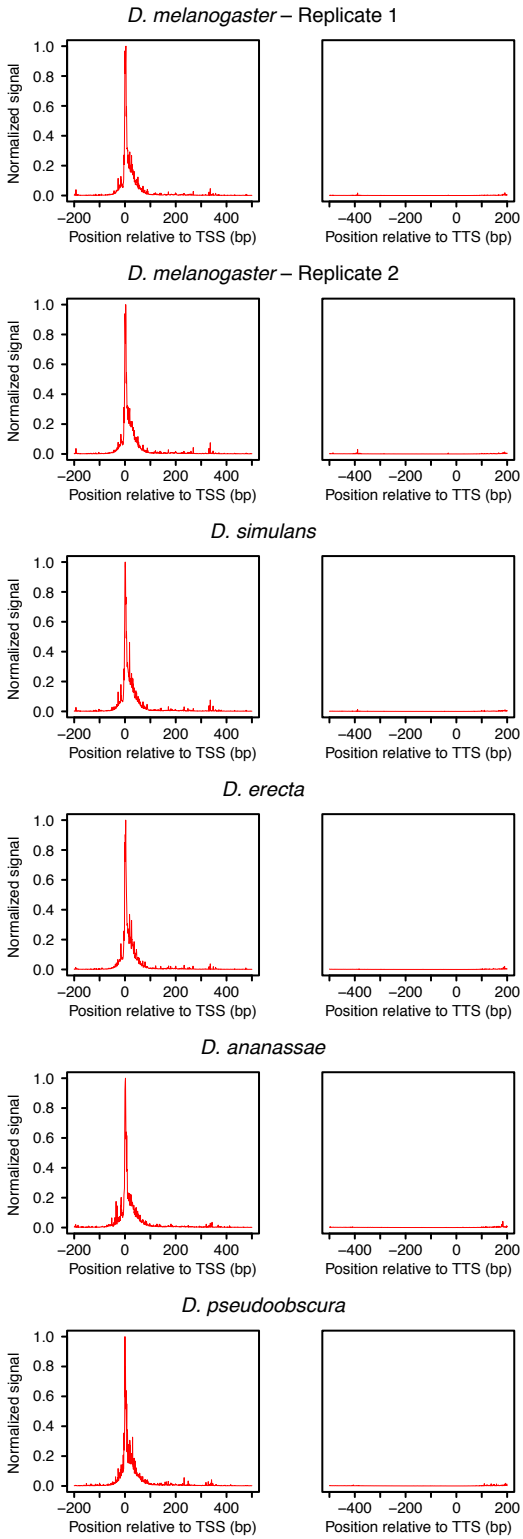

Figure 5.25. Distribution or raw RAMPAGE signal over transcript annotations
For each species, RAMPAGE reads were mapped to the appropriate genome. The raw 5' signal was then converted to orthologous *D. melanogaster* coordinates using chained pairwise alignments from UCSC. Metaprofiles were constructed by summing signal intensity over Flybase r5.49 mRNA annotations.

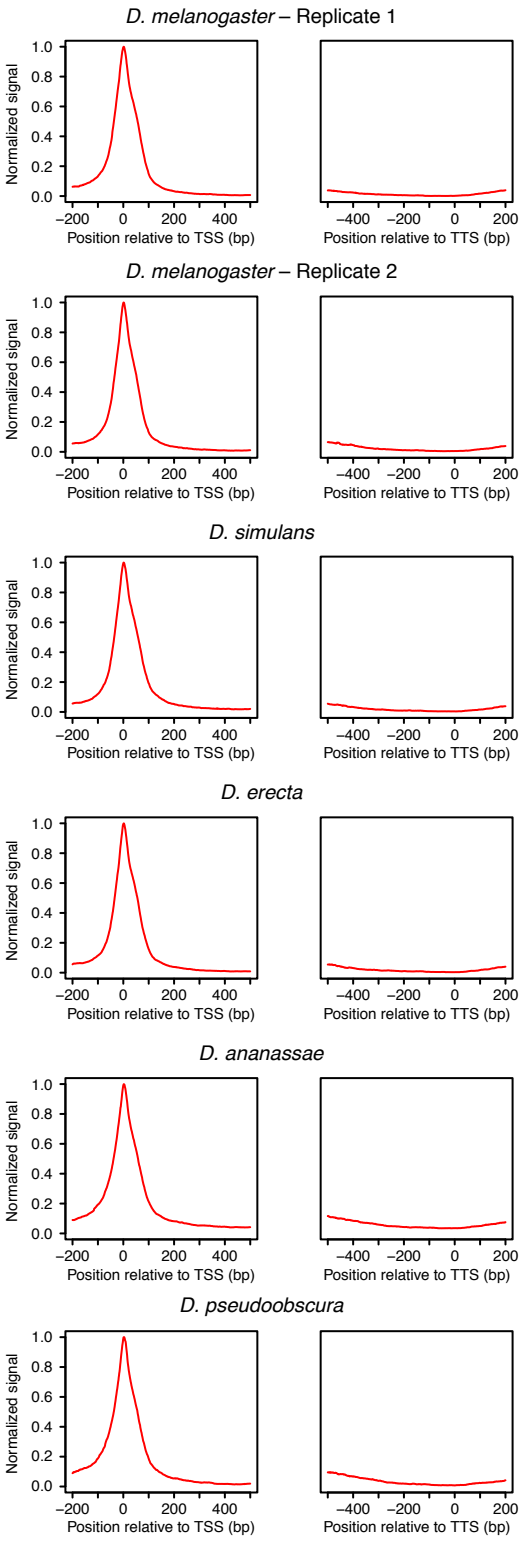

Figure 5.26. Distribution or RAMPAGE peaks over transcript annotations
For each species, RAMPAGE reads were mapped to the appropriate genome and peaks called as described in Chapter 4. The peak coordinates were then converted to orthologous *D. melanogaster* coordinates using chained pairwise alignments and the liftOver tool from UCSC. Metaprofiles were constructed by summing signal intensity over Flybase r5.49 mRNA annotations.

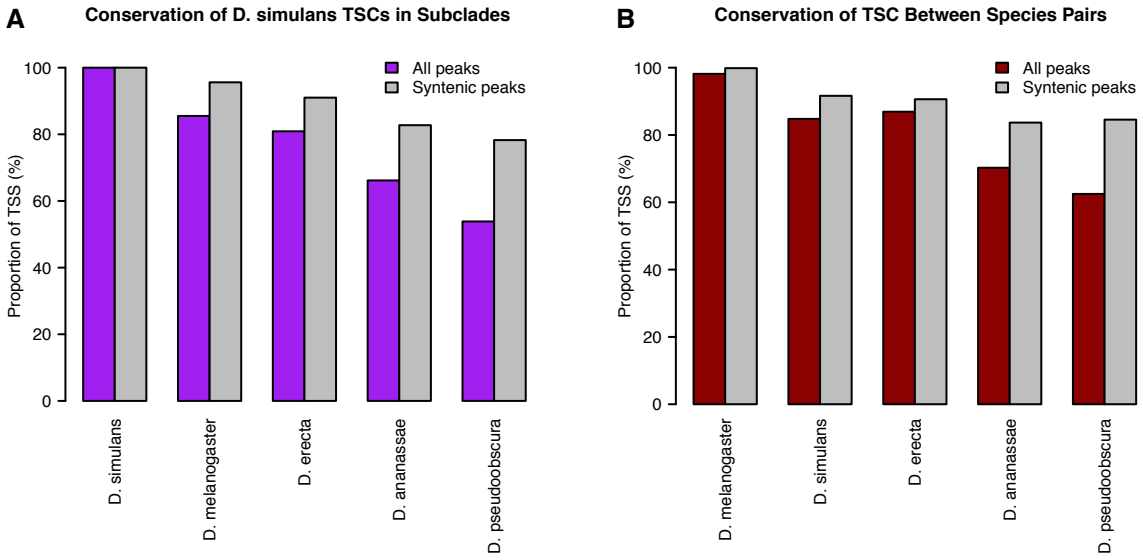

Figure 5.27. Alternative analyses of TSC conservation
TSC conservation was quantified as described in Chapter 4. (A) *D. simulans*-centric analysis. (B) Quantification of TSC conservation between species pairs, as opposed to subclades.

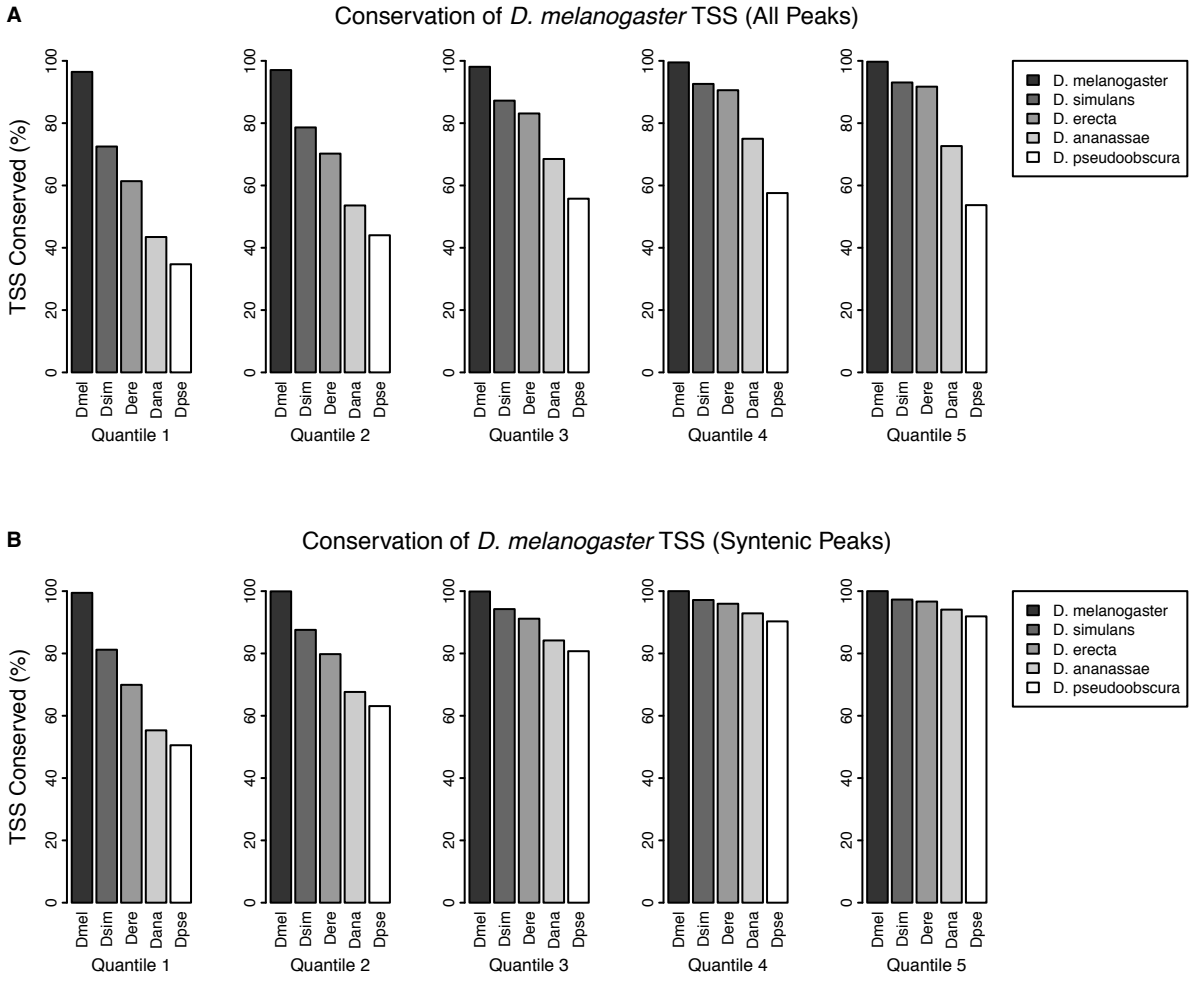

Figure 5.28. TSC conservation by expression quantiles
D. melanogaster TSCs were categorized into 5 expression quantiles based on total raw signal for the full time series. Functional conservation was assessed as described in Chapter 4. (A) Conservation of all TSCs. (B) Conservation of TSCs with syntenic alignments in all 5 species.

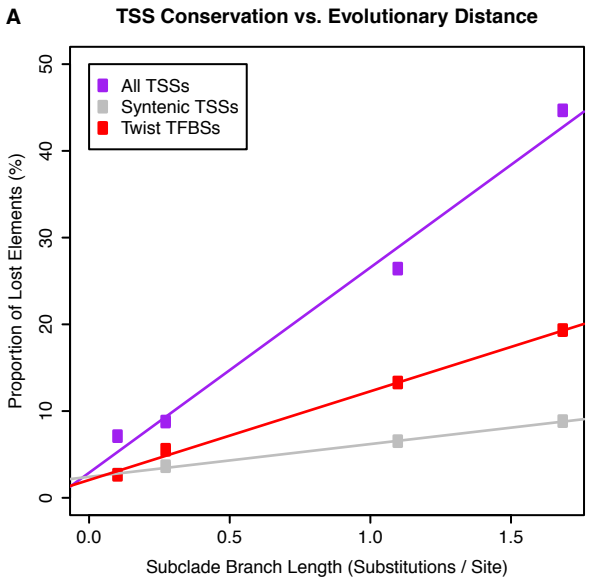

Figure 5.29. Evolutionary rates of gain and loss for TSCs and Twist TFBSs See Chapter 4 for methods. Twist TFBS data from He et al. (2011).

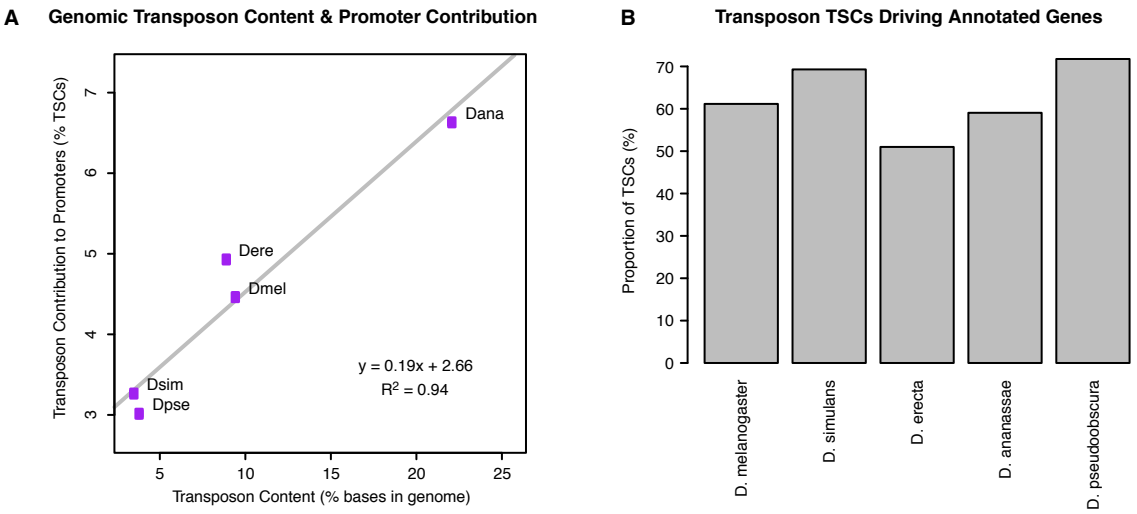

Figure 5.30. Transposons contribute many genic TSCs
Analysis based on RepeatRunner transposons annotations by the *Drosophila* 12 genomes consortium. (A) Proportion of genic TSCs that overlap transposon annotations, plotted as a function of total genome coverage by transposon annotations. (B) Proportion of transposon-derived TSCs that drive the expression of annotated genes in each species.

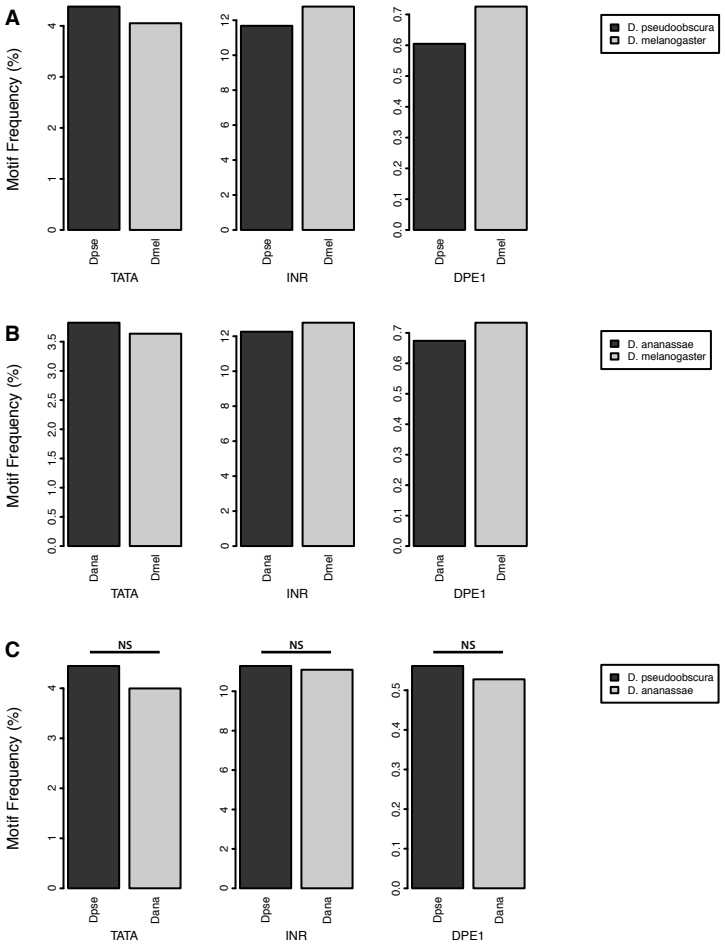

Figure 5.31. Core promoter syntax evolution: Control analyses
Analyses aimed at assessing the impact of whole-genome alignment errors on relative frequencies of core promoter usage. If the *D. melanogaster*-centric analysis detected fewer INR and DPE1 motifs in *D. pseudoobscura* and *D. ananassae* because of alignment errors, the trends would be expected to be reversed when performing comparisons in the other direction. This was not observed, which supports the reality of the trends reported in Chapter 4. (A) Comparison of *D. pseudoobscura* TSCs to their *D. melanogaster* orthologs. (B) Comparison of *D. ananassae* TSCs to their *D. melanogaster* orthologs. (C) Direct comparison between *D. pseudoobscura* and *D. ananassae*.

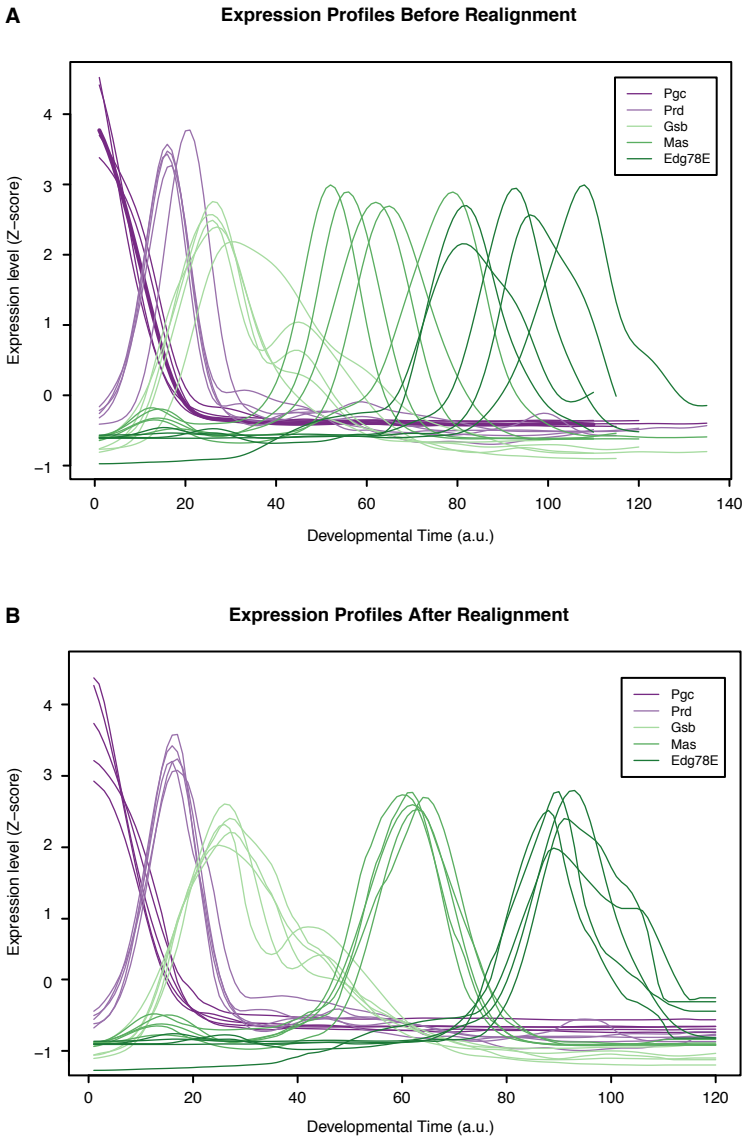

Figure 5.32. Time series alignment by time-warping of gene expression profiles
Global gene expression profiles from all species were aligned to the *D. melanogaster* time series as described in Chapter 4. This figure shows the expression profiles for well-characterized developmental regulators before (A) and after (B) alignment. The time scale corresponds to the absolute *D. melanogaster* developmental time (24 hours divided into 120 units by upsampling).

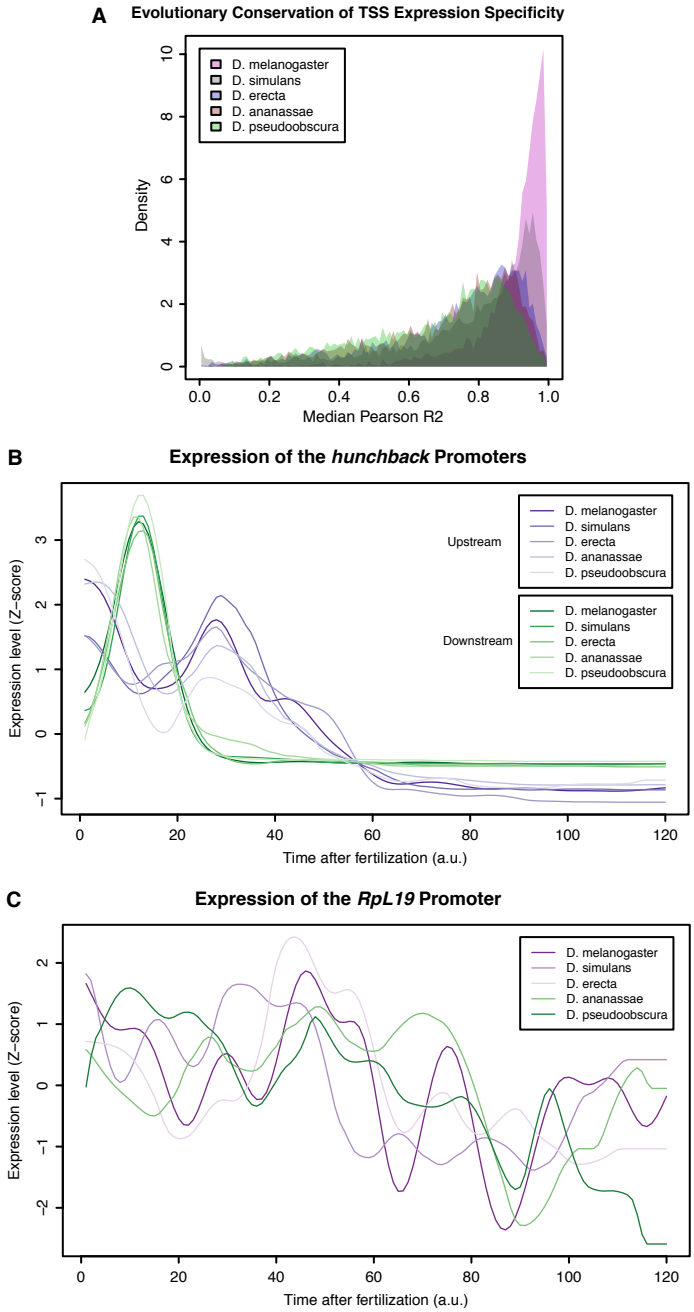

Figure 5.33. Evolutionary conservation of gene expression profiles
(A) Distribution of average correlation coefficients for all orthologous genes between pairs of species. (B) Aligned expression profiles for the 2 promoters of the *hunchback* gene. (C) Aligned expression profiles for the *RpL19* gene.

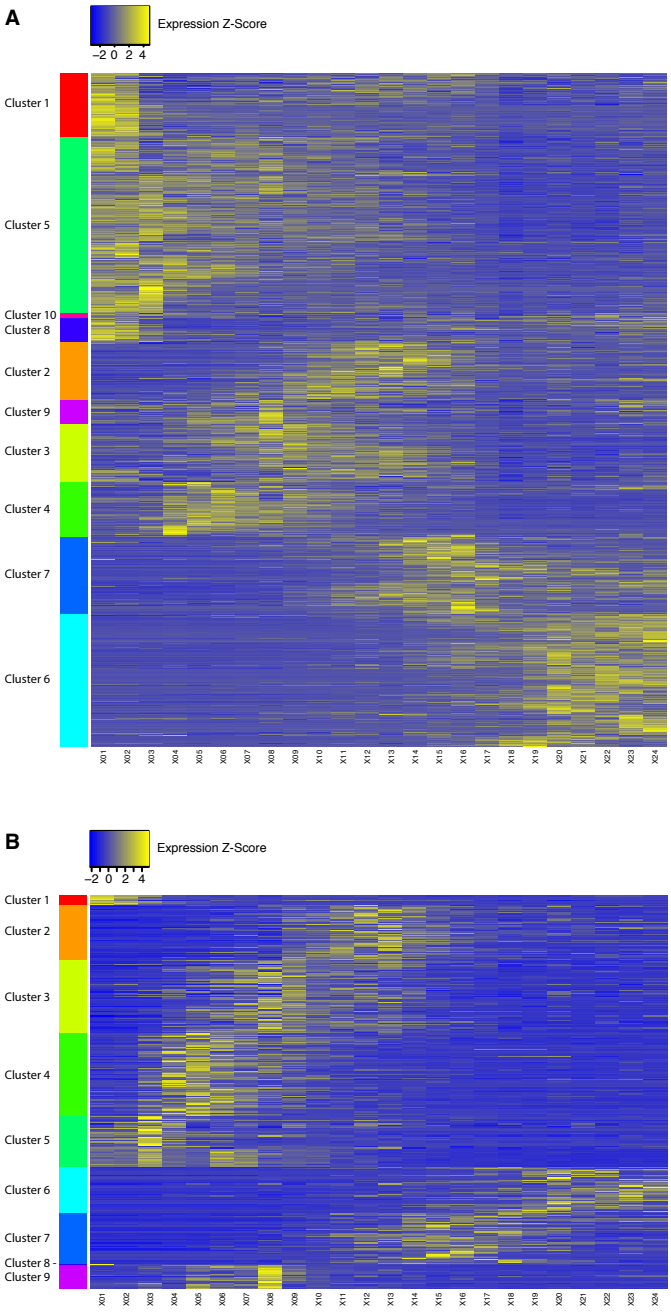

Figure 5.34. Clustering of *D. melanogaster* developmental expression profiles
In order to analyze coexpression between putative lncRNA and protein-coging gene promoters, we first grouped protein-coding gene promoter expression profiles by k-means clustering (10 clusters) to define reference coexpression sets (A). We then clustered both lncRNA and protein-coging gene promoter profiles together, and extracted the lncRNA promoter profiless corresponding to each expression cluster previously defined (B).

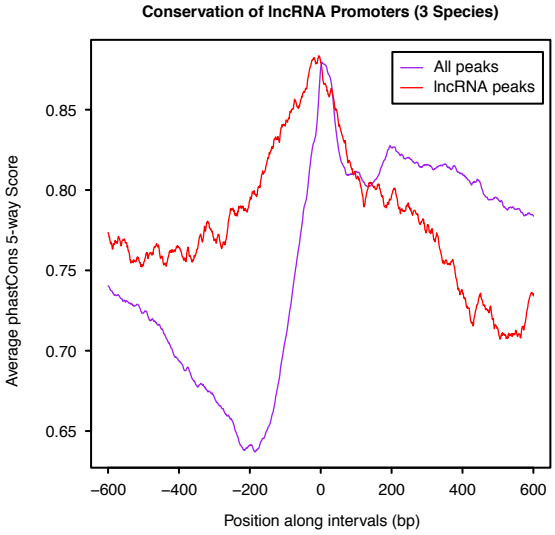

Figure 5.35. Sequence conservation of *melanogaster* subgroup lncRNA TSCs
In order to assess whether lncRNA TSCs that are shared at least within the *melanogaster* subgroup are under purifying selection, we plotted *melanogaster* subgroup-specific phastCons scores over TSCs. These phastCons scores were computed by including exclusively the genomes of *melanogaster* subgroup species in the input multiple sequence alignment.

Bibliography

Adelman K, Lis JT, 2012. Promoter-proximal pausing of rna polymerase ii: emerging roles in metazoans. *Nat Rev Genet*, **13**: 720–31. ISSN 1471-0064 (Electronic) 1471-0056 (Linking). doi: 10.1038/nrg3293. 10.

Agalioti T, Lomvardas S, Parekh B, Yie J, Maniatis T, Thanos D, 2000. Ordered recruitment of chromatin modifying and general transcription factors to the ifn-beta promoter. *Cell*, **103**: 667–78. ISSN 0092-8674 (Print) 0092-8674 (Linking). 4.

Almada AE, Wu X, Kriz AJ, Burge CB, Sharp PA, 2013. Promoter directionality is controlled by u1 snrnp and polyadenylation signals. *Nature*, **499**: 360–3. ISSN 1476-4687 (Electronic) 0028-0836 (Linking). doi:10.1038/nature12349. 7458.

Almer A, Rudolph H, Hinnen A, Horz W, 1986. Removal of positioned nucleosomes from the yeast pho5 promoter upon pho5 induction releases additional upstream activating dna elements. *EMBO J*, **5**: 2689–96. ISSN 0261-4189 (Print) 0261-4189 (Linking). 10.

Alon U, 2007. Network motifs: theory and experimental approaches. *Nature Reviews Genetics*, **8**: 450–461. ISSN 1471-0056. doi:10.1038/Nrg2102. 6.

Amano T, Sagai T, Tanabe H, Mizushina Y, Nakazawa H, Shiroishi T, 2009. Chromosomal dynamics at the shh locus: limb bud-specific differential regulation of competence and active transcription. *Dev Cell*, **16**: 47–57. ISSN 1878-1551 (Electronic) 1534-5807 (Linking). doi:10.1016/j.devcel.2008.11.011. 1.

Arbiza L, Gronau I, Aksoy BA, Hubisz MJ, Gulko B, Keinan A, Siepel A, 2013. Genome-wide inference of natural selection on human transcription factor binding sites. *Nat Genet*, **45**: 723–9. ISSN 1546-1718 (Electronic) 1061-4036 (Linking). doi:10.1038/ng.2658. 7.

Arnoult L, Su KF, Manoel D, Minervino C, Magrina J, Gompel N, Prud'homme B, 2013. Emergence and diversification of fly pigmentation through evolution of a gene regulatory module. *Science*, **339**: 1423–6. ISSN 1095-9203 (Electronic) 0036-8075 (Linking). doi:10.1126/science.1233749. 6126.

Attanasio C, Nord AS, Zhu Y, Blow MJ, Li Z, Liberton DK, Morrison H, Plajzer-Frick I, Holt A, Hosseini R, Phouanenavong S, Akiyama JA, Shoukry M, Afzal V, Rubin EM, FitzPatrick DR, Ren B, Hallgrimsson B, Pennacchio LA, Visel A, 2013. Fine tuning of craniofacial morphology by distant-acting enhancers. *Science*, **342**: 1241006. ISSN 1095-9203 (Electronic) 0036-8075 (Linking). doi: 10.1126/science.1241006. 6157.

Augui S, Nora EP, Heard E, 2011. Regulation of x-chromosome inactivation by the x-inactivation centre. *Nat Rev Genet*, **12**: 429–42. ISSN 1471-0064 (Electronic) 1471-0056 (Linking). doi:10.1038/nrg2987. 6.

Ayroles JF, Carbone MA, Stone EA, Jordan KW, Lyman RF, Magwire MM, Rollmann SM, Duncan LH, Lawrence F, Anholt RR, Mackay TF, 2009. Systems genetics of complex traits in drosophila melanogaster. *Nat Genet*, **41**: 299–307. ISSN 1546-1718 (Electronic). doi:10.1038/ng.332. 3.

Banerji J, Olson L, Schaffner W, 1983. A lymphocyte-specific cellular enhancer is located downstream of the joining region in immunoglobulin heavy chain genes. *Cell*, **33**: 729–40. ISSN 0092-8674 (Print) 0092-8674 (Linking). 3.

Banerji J, Rusconi S, Schaffner W, 1981. Expression of a beta-globin gene is enhanced by remote sv40 dna sequences. *Cell*, **27**: 299–308. ISSN 0092-8674 (Print) 0092-8674 (Linking). 2 Pt 1.

Barberis A, Pearlberg J, Simkovich N, Farrell S, Reinagel P, Bamdad C, Sigal G, Ptashne M, 1995. Contact with a component of the polymerase ii holoenzyme suffices for gene activation. *Cell*, **81**: 359–68. ISSN 0092-8674 (Print) 0092-8674 (Linking). 3.

Barbosa-Morais NL, Irimia M, Pan Q, Xiong HY, Gueroussov S, Lee LJ, Slobodeniuc V, Kutter C, Watt S, Colak R, Kim T, Misquitta-Ali CM, Wilson MD, Kim PM, Odom DT, Frey BJ, Blencowe BJ, 2012. The evolutionary landscape of alternative splicing in vertebrate species. *Science*, **338**: 1587–93. ISSN 1095-9203 (Electronic) 0036-8075 (Linking). doi:10.1126/science.1230612. 6114.

Bejerano G, Lowe CB, Ahituv N, King B, Siepel A, Salama SR, Rubin EM, Kent WJ, Haussler D, 2006. A distal enhancer and an ultraconserved exon are derived from a novel retroposon. *Nature*, **441**: 87–90. ISSN 0028-0836. doi:10.1038/Nature04696. 7089.

Bell O, Tiwari VK, Thoma NH, Schubeler D, 2011. Determinants and dynamics of genome accessibility. *Nat Rev Genet*, **12**: 554–64. ISSN 1471-0064 (Electronic) 1471-0056 (Linking). doi:10.1038/nrg3017. 8.

Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M, 2012. An integrated encyclopedia of dna elements in the human genome. *Nature*, **489**: 57–74. ISSN 1476-4687 (Electronic) 0028-0836 (Linking). doi:10.1038/nature11247. 7414.

Bertani S, Sauer S, Bolotin E, Sauer F, 2011. The noncoding rna mistral activates hoxa6 and hoxa7 expression and stem cell differentiation by recruiting mll1 to chromatin. *Mol Cell*, **43**: 1040–6. ISSN 1097-4164 (Electronic) 1097-2765 (Linking). doi:10.1016/j.molcel.2011.08.019. 6.

Bingham PM, Kidwell MG, Rubin GM, 1982. The molecular basis of p-m hybrid dysgenesis: the role of the p element, a p-strain-specific transposon family. *Cell*, **29**: 995–1004. ISSN 0092-8674 (Print) 0092-8674 (Linking). 3.

Boettiger AN, Levine M, 2009. Synchronous and stochastic patterns of gene activation in the drosophila embryo. *Science*, **325**: 471–473. ISSN 0036-8075. doi:10.1126/Science.1173976. 5939.

Bourque G, 2009. Transposable elements in gene regulation and in the evolution of vertebrate genomes. *Current Opinion in Genetics & Development*, **19**: 607–612. ISSN 0959-437X. doi:10.1016/J.Gde.2009.10.013. 6.

Bourque G, Leong B, Vega VB, Chen X, Lee YL, Srinivasan KG, Chew JL, Ruan Y, Wei CL, Ng HH, Liu ET, 2008. Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome Research*, **18**: 1752–1762. ISSN 1088-9051. doi:10.1101/Gr.080663.108. 11.

Boyer LA, Lee TI, Cole MF, Johnstone SE, Levine SS, Zucker JP, Guenther MG, Kumar RM, Murray HL, Jenner RG, Gifford DK, Melton DA, Jaenisch R, Young RA, 2005. Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell*, **122**: 947–56. ISSN 0092-8674 (Print) 0092-8674 (Linking). doi:10.1016/j.cell.2005.08.020. 6.

Bradley RK, Li XY, Trapnell C, Davidson S, Pachter L, Chu HC, Tonkin LA, Biggin MD, Eisen MB, 2010. Binding site turnover produces pervasive quantitative changes in transcription factor binding between closely related drosophila species. *PLoS Biol*, **8**: e1000343. ISSN 1545-7885 (Electronic) 1544-9173 (Linking). doi:10.1371/journal.pbio.1000343. 3.

Brannan CI, Dees EC, Ingram RS, Tilghman SM, 1990. The product of the h19 gene may function as an rna. *Mol Cell Biol*, **10**: 28–36. ISSN 0270-7306 (Print) 0270-7306 (Linking). 1.

Brenner S, Jacob F, Meselson M, 1961. An unstable intermediate carrying information from genes to ribosomes for protein synthesis. *Nature*, **190**: 576–581. ISSN 0028-0836 (Print) 0028-0836 (Linking).

Brent R, Ptashne M, 1985. A eukaryotic transcriptional activator bearing the dna specificity of a prokaryotic repressor. *Cell*, **43**: 729–36. ISSN 0092-8674 (Print) 0092-8674 (Linking). 3 Pt 2.

Britten RJ, Davidson EH, 1969. Gene regulation for higher cells - a theory. *Science*, **165**: 349–&. ISSN 0036-8075. 3891.

Britten RJ, Davidson EH, 1971. Repetitive and non-repetitive dna sequences and a speculation on origins of evolutionary novelty. *Quarterly Review of Biology*, **46**: 111–&. ISSN 0033-5770. 2.

Brockdorff N, Ashworth A, Kay GF, McCabe VM, Norris DP, Cooper PJ, Swift S, Rastan S, 1992. The product of the mouse xist gene is a 15 kb inactive x-specific transcript containing no conserved orf and located in the nucleus. *Cell*, **71**: 515–26. ISSN 0092-8674 (Print) 0092-8674 (Linking). 3.

Bronner G, Taubert H, Jackle H, 1995. Mesoderm-specific b104 expression in the drosophila embryo is mediated by internal cis-acting elements of the transposon. *Chromosoma*, **103**: 669–75. ISSN 0009-5915 (Print) 0009-5915 (Linking). 10.

Brown CJ, Ballabio A, Rupert JL, Lafreniere RG, Grompe M, Tonlorenzi R, Willard HF, 1991. A gene from the region of the human x inactivation centre is expressed exclusively from the inactive x chromosome. *Nature*, **349**: 38–44. ISSN 0028-0836 (Print) 0028-0836 (Linking). doi:10.1038/349038a0. 6304.

Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C, Kodzius R, Shimokawa K, Bajic VB, Brenner SE, Batalov S, Forrest AR, Zavolan M, Davis MJ, Wilming LG, Aidinis V, Allen JE, Ambesi-Impiombato A, Apweiler R, Aturaliya RN, Bailey TL, Bansal M, Baxter L, Beisel KW, Bersano T, Bono H, Chalk AM, Chiu KP, Choudhary V, Christoffels A, Clutterbuck DR, Crowe ML, Dalla E, Dalrymple BP, de Bono B, Della Gatta G, di Bernardo D, Down T, Engstrom P, Fagiolini M, Faulkner G, Fletcher CF, Fukushima T, Furuno M, Futaki S, Gariboldi M, Georgii-Hemming P, Gingeras TR, Gojobori T, Green RE, Gustincich S, Harbers M, Hayashi Y, Hensch TK, Hirokawa N, Hill D, Huminiecki L, Iacono M, Ikeo K, Iwama A, Ishikawa T, Jakt M, Kanapin A, Katoh M, Kawasawa Y, Kelso J, Kitamura H, Kitano H, Kollias G, Krishnan SP, Kruger A, Kummerfeld SK, Kurochkin IV, Lareau LF, Lazarevic D, Lipovich L, Liu J, Liuni S, McWilliam S, Madan Babu M, Madera M, Marchionni L, Matsuda H, Matsuzawa S, Miki H, Mignone F, Miyake S, Morris K, Mottagui-Tabar S, Mulder N, Nakano N, Nakauchi H, Ng P, Nilsson

R, Nishiguchi S, Nishikawa S, Nori F, Ohara O, Okazaki Y, Orlando V, Pang KC, Pavan WJ, Pavesi G, Pesole G, Petrovsky N, Piazza S, Reed J, Reid JF, Ring BZ, Ringwald M, Rost B, Ruan Y, Salzberg SL, Sandelin A, Schneider C, Schonbach C, Sekiguchi K, Semple CA, Seno S, Sessa L, Sheng Y, Shibata Y, Shimada H, Shimada K, Silva D, Sinclair B, Sperling S, Stupka E, Sugiura K, Sultana R, Takenaka Y, Taki K, Tammoja K, Tan SL, Tang S, Taylor MS, Tegner J, Teichmann SA, Ueda HR, van Nimwegen E, Verardo R, Wei CL, Yagi K, Yamanishi H, Zabarovsky E, Zhu S, Zimmer A, Hide W, Bult C, Grimmond SM, Teasdale RD, Liu ET, Brusic V, Quackenbush J, Wahlestedt C, Mattick JS, Hume DA, Kai C, Sasaki D, Tomaru Y, Fukuda S, Kanamori-Katayama M, Suzuki M, Aoki J, Arakawa T, Iida J, Imamura K, Itoh M, Kato T, Kawaji H, Kawagashira N, Kawashima T, Kojima M, Kondo S, Konno H, Nakano K, Ninomiya N, Nishio T, Okada M, Plessy C, Shibata K, Shiraki T, Suzuki S, Tagami M, Waki K, Watahiki A, Okamura-Oho Y, Suzuki H, Kawai J, Hayashizaki Y, 2005. The transcriptional landscape of the mammalian genome. *Science*, **309**: 1559–63. ISSN 1095-9203 (Electronic). doi:10.1126/science.1112014. 5740.

Carninci P, Kvam C, Kitamura A, Ohsumi T, Okazaki Y, Itoh M, Kamiya M, Shibata K, Sasaki N, Izawa M, Muramatsu M, Hayashizaki Y, Schneider C, 1996. High-efficiency full-length cdna cloning by biotinylated cap trapper. *Genomics*, **37**: 327–336. ISSN 0888-7543. 3.

Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple CAM, Taylor MS, Engstrom PG, Frith MC, Forrest ARR, Alkema WB, Tan SL, Plessy C, Kodzius R, Ravasi T, Kasukawa T, Fukuda S, Kanamori-Katayama M, Kitazume Y, Kawaji H, Kai C, Nakamura M, Konno H, Nakano K, Mottagui-Tabar S, Arner P, Chesi A, Gustincich S, Persichetti F, Suzuki H, Grimmond SM, Wells CA, Orlando V, Wahlestedt C, Liu ET, Harbers M, Kawai J, Bajic VB, Hume DA, Hayashizaki Y, 2006. Genome-wide analysis of mammalian promoter architecture and evolution. *Nature Genetics*, **38**: 626–635. ISSN 1061-4036. doi:10.1038/Ng1789. 6.

Carroll SB, 1995. Homeotic genes and the evolution of arthropods and chordates. *Nature*, **376**: 479–85. ISSN 0028-0836 (Print) 0028-0836 (Linking). doi:10.1038/376479a0. 6540.

Carroll SB, 2005. Evolution at two levels: On genes and form. *Plos Biology*, **3**: 1159–1166. ISSN 1544-9173. doi:10.1371/journal.pbio.0030245. 7.

Carroll SB, 2008. Evo-devo and an expanding evolutionary synthesis: A genetic theory of morphological evolution. *Cell*, **134**: 25–36. ISSN 0092-8674. doi:10.1016/J.Cell.2008.06.030. 1.

Carter D, Chakalova L, Osborne CS, Dai YF, Fraser P, 2002. Long-range chromatin regulatory interactions in vivo. *Nat Genet*, **32**: 623–6. ISSN 1061-4036 (Print) 1061-4036 (Linking). doi: 10.1038/ng1051. 4.

Castel SE, Martienssen RA, 2013. Rna interference in the nucleus: roles for small rnas in transcription, epigenetics and beyond. *Nat Rev Genet*, **14**: 100–12. ISSN 1471-0064 (Electronic) 1471-0056 (Linking). doi:10.1038/nrg3355. 2.

Charlesworth B, Lapid A, Canada D, 1992. The distribution of transposable elements within and between chromosomes in a population of drosophila melanogaster. ii. inferences on the nature of selection against elements. *Genet Res*, **60**: 115–30. ISSN 0016-6723 (Linking). 2.

Charlesworth B, Sniegowski P, Stephan W, 1994. The evolutionary dynamics of repetitive dna in eukaryotes. *Nature*, **371**: 215–220. ISSN 0028-0836. 6494.

Chaumeil J, Le Baccon P, Wutz A, Heard E, 2006. A novel role for xist rna in the formation of a repressive nuclear compartment into which genes are recruited when silenced. *Genes Dev*, **20**: 2223–37. ISSN 0890-9369 (Print) 0890-9369 (Linking). doi:10.1101/gad.380906. 16.

Choi OR, Engel JD, 1988. Developmental regulation of beta-globin gene switching. *Cell*, **55**: 17–26. ISSN 0092-8674 (Print) 0092-8674 (Linking). 1.

Cirillo LA, Lin FR, Cuesta I, Friedman D, Jarnik M, Zaret KS, 2002. Opening of compacted chromatin by early developmental transcription factors hnf3 (foxa) and gata-4. *Mol Cell*, **9**: 279–89. ISSN 1097-2765 (Print) 1097-2765 (Linking). 2.

Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, Markow TA, Kaufman TC, Kellis M, Gelbart W, Iyer VN, Pollard DA, Sackton TB, Larracuente AM, Singh ND, Abad JP, Abt DN, Adryan B, Aguade M, Akashi H, Anderson WW, Aquadro CF, Ardell DH, Arguello R, Artieri CG, Barbash DA, Barker D, Barsanti P, Batterham P, Batzoglou S, Begun D, Bhutkar A, Blanco E, Bosak SA, Bradley RK, Brand AD, Brent MR, Brooks AN, Brown RH, Butlin RK, Caggese C, Calvi BR, Bernardo de Carvalho A, Caspi A, Castrezana S, Celniker SE, Chang JL, Chapple C, Chatterji S, Chinwalla A, Civetta A, Clifton SW, Comeron JM, Costello JC, Coyne JA, Daub J, David RG, Delcher AL, Delehaunty K, Do CB, Ebling H, Edwards K, Eickbush T, Evans JD, Filipski A, Findeiss S, Freyhult E, Fulton L, Fulton R, Garcia AC, Gardiner A, Garfield DA, Garvin BE, Gibson G, Gilbert D, Gnerre S, Godfrey J, Good R, Gotea V, Gravely B, Greenberg AJ, Griffiths-Jones S, Gross S, Guigo R, Gustafson EA, Haerty W, Hahn MW, Halligan DL, Halpern AL, Halter GM, Han MV, Heger A, Hillier L, Hinrichs AS, Holmes I, Hoskins RA, Hubisz MJ, Hultmark D, Huntley MA, Jaffe DB, Jagadeeshan S, Jeck WR, Johnson J, Jones CD, Jordan WC, Karpen GH, Kataoka E, Keightley PD, Kheradpour P, Kirkness EF, Koerich LB, Kristiansen K, Kudrna D, Kulathinal RJ, Kumar S, Kwok R, Lander E, Langley CH, Lapoint R, Lazzaro BP, Lee SJ, Levesque L, Li R, Lin CF, Lin MF, Lindblad-Toh K, Llopart A, Long M, Low L, Lozovsky E, Lu J, Luo M, Machado CA, Makalowski W, Marzo M, Matsuda M, Matzkin L, McAllister B, McBride CS, McKernan B, McKernan K, Mendez-Lago M, Minx P, Mollenhauer MU, Montooth K, Mount SM, Mu X, Myers E, Negre B, Newfeld S, Nielsen R, Noor MA, O'Grady P, Pachter L, Papaceit M, Parisi MJ, Parisi M, Parts L, Pedersen JS, Pesole G, Phillippy AM, Ponting CP, Pop M, Porcelli D, Powell JR, Prohaska S, Pruitt K, Puig M, Quesneville H, Ram KR, Rand D, Rasmussen MD, Reed LK, Reenan R, Reily A, Remington KA, Rieger TT, Ritchie MG, Robin C, Rogers YH, Rohde C, Rozas J, Rubenfield MJ, Ruiz A, Russo S, Salzberg SL, Sanchez-Gracia A, Saranga DJ, Sato H, Schaeffer SW, Schatz MC, Schlenke T, Schwartz R, Segarra C, Singh RS, Sirot L, Sirota M, Sisneros NB, Smith CD, Smith TF, Spieth J, Stage DE, Stark A, Stephan W, Strausberg RL, Strempel S, Sturgill D, Sutton G, Sutton GG, Tao W, Teichmann S, Tobari YN, Tomimura Y, Tsolas JM, Valente VL, Venter E, Venter JC, Vicario S, Vieira FG, Vilella AJ, Villasante A, Walenz B, Wang J, Wasserman M, Watts T, Wilson D, Wilson RK, Wing RA, Wolfner MF, Wong A, Wong GK, Wu CI, Wu G, Yamamoto D, Yang HP, Yang SP, Yorke JA, Yoshida K, Zdobnov E, Zhang P, Zhang Y, Zimin AV, Baldwin J, Abdouelleil A, Abdulkadir J, Abebe A, Abera B, Abreu J, Acer SC, Aftuck L, Alexander A, An P, Anderson E, Anderson S, Arachi H, Azer M, Bachantsang P, Barry A, Bayul T, Berlin A, Bessette D, Bloom T, Blye J, Boguslavskiy L, Bonnet C, Boukhgalter B, Bourzgui I, Brown A, Cahill P, Channer S, Cheshatsang Y, Chuda L, Citroen M, Collymore A, Cooke P, Costello M, D'Aco K, Daza R, De Haan G, DeGray S, DeMaso C, Dhargay N, Dooley K, Dooley E, Doricent M, Dorje P, Dorjee K, Dupes A, Elong R, Falk J, Farina A, Faro S, Ferguson D, Fisher S, Foley CD, Franke A, Friedrich D, Gadbois L, Gearin G, Gearin CR, Giannoukos G, Goode T, Graham J, Grandbois E, Grewal S, Gyaltsen K, Hafez N, Hagos B, Hall J, Henson C, Hollinger A, Honan T, Huard MD, Hughes L, Hurhula B, Husby ME, Kamat A, Kanga B, Kashin S, Khazanovich D, Kisner P, Lance K, Lara M, Lee W, Lennon N, Letendre F, LeVine

R, Lipovsky A, Liu X, Liu J, Liu S, Lokyitsang T, Lokyitsang Y, Lubonja R, Lui A, MacDonald P, Magnisalis V, Maru K, Matthews C, McCusker W, McDonough S, Mehta T, Meldrim J, Meneus L, Mihai O, Mihalev A, Mihova T, Mittelman R, Mlenga V, Montmayeur A, Mulrain L, Navidi A, Naylor J, Negash T, Nguyen T, Nguyen N, Nicol R, Norbu C, Norbu N, Novod N, O'Neill B, Osman S, Markiewicz E, Oyono OL, Patti C, Phunkhang P, Pierre F, Priest M, Raghuraman S, Rege F, Reyes R, Rise C, Rogov P, Ross K, Ryan E, Settipalli S, Shea T, Sherpa N, Shi L, Shih D, Sparrow T, Spaulding J, Stalker J, Stange-Thomann N, Stavropoulos S, Stone C, Strader C, Tesfaye S, Thomson T, Thoulutsang Y, Thoulutsang D, Topham K, Topping I, Tsamla T, Vassiliev H, Vo A, Wangchuk T, Wangdi T, Weiand M, Wilkinson J, Wilson A, Yadav S, Young G, Yu Q, Zembek L, Zhong D, Zimmer A, Zwirko Z, Alvarez P, Brockman W, Butler J, Chin C, Grabherr M, Kleber M, Mauceli E, MacCallum I, 2007. Evolution of genes and genomes on the drosophila phylogeny. *Nature*, **450**: 203–18. ISSN 1476-4687 (Electronic). doi:10.1038/nature06341. 7167.

Clark MB, Amaral PP, Schlesinger FJ, Dinger ME, Taft RJ, Rinn JL, Ponting CP, Stadler PF, Morris KV, Morillon A, Rozowsky JS, Gerstein MB, Wahlestedt C, Hayashizaki Y, Carninci P, Gingeras TR, Mattick JS, 2011. The reality of pervasive transcription. *PLoS Biol*, **9**: e1000625; discussion e1001102. ISSN 1545-7885 (Electronic) 1544-9173 (Linking). doi:10.1371/journal.pbio.1000625. 7.

Clemson CM, McNeil JA, Willard HF, Lawrence JB, 1996. Xist rna paints the inactive x chromosome at interphase: evidence for a novel rna involved in nuclear/chromosome structure. *J Cell Biol*, **132**: 259–75. ISSN 0021-9525 (Print) 0021-9525 (Linking). 3.

Cohen CJ, Lock WM, Mager DL, 2009. Endogenous retroviral ltrs as promoters for human genes: A critical assessment. *Gene*, **448**: 105–114. ISSN 0378-1119. doi:10.1016/J.Gene.2009.06.020. 2.

Corbett-Detig RB, Zhou J, Clark AG, Hartl DL, Ayroles JF, 2013. Genetic incompatibilities are widespread within species. *Nature*, **504**: 135–7. ISSN 1476-4687 (Electronic) 0028-0836 (Linking). doi:10.1038/nature12678. 7478.

Core LJ, Waterfall JJ, Lis JT, 2008. Nascent rna sequencing reveals widespread pausing and divergent initiation at human promoters. *Science*, **322**: 1845–1848. ISSN 0036-8075. doi:10.1126/Science.1162228. 5909.

Cotney J, Leng J, Yin J, Reilly SK, DeMare LE, Emera D, Ayoub AE, Rakic P, Noonan JP, 2013. The evolution of lineage-specific regulatory activities in the human embryonic limb. *Cell*, **154**: 185–96. ISSN 1097-4172 (Electronic) 0092-8674 (Linking). doi:10.1016/j.cell.2013.05.056. 1.

Cremer T, Cremer C, 2001. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nat Rev Genet*, **2**: 292–301. ISSN 1471-0056 (Print) 1471-0056 (Linking). doi:10.1038/35066075. 4.

Czech B, Malone CD, Zhou R, Stark A, Schlingeheyde C, Dus M, Perrimon N, Kellis M, Wohlschlegel JA, Sachidanandam R, Hannon GJ, Brennecke J, 2008. An endogenous small interfering rna pathway in drosophila. *Nature*, **453**: 798–U7. ISSN 0028-0836. doi:10.1038/Nature07007. 7196.

Davidson EH, 2010. Emerging properties of animal gene regulatory networks. *Nature*, **468**: 911–20. ISSN 1476-4687 (Electronic) 0028-0836 (Linking). doi:10.1038/nature09645. 7326.

Davidson EH, Erwin DH, 2006. Gene regulatory networks and the evolution of animal body plans. *Science*, **311**: 796–800. ISSN 1095-9203 (Electronic) 0036-8075 (Linking). doi:10.1126/science.1113832. 5762.

Degner JF, Pai AA, Pique-Regi R, Veyrieras JB, Gaffney DJ, Pickrell JK, De Leon S, Michelini K, Lewellen N, Crawford GE, Stephens M, Gilad Y, Pritchard JK, 2012. Dnase i sensitivity qtls are a major determinant of human expression variation. *Nature*, **482**: 390–4. ISSN 1476-4687 (Electronic) 0028-0836 (Linking). doi:10.1038/nature10808. 7385.

Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel A, Knowles DG, Lagarde J, Veeravalli L, Ruan X, Ruan Y, Lassmann T, Carninci P, Brown JB, Lipovich L, Gonzalez JM, Thomas M, Davis CA, Shiekhattar R, Gingeras TR, Hubbard TJ, Notredame C, Harrow J, Guigo R, 2012. The gencode v7 catalog of human long noncoding rnas: analysis of their gene structure, evolution, and expression. *Genome Res*, **22**: 1775–89. ISSN 1549-5469 (Electronic) 1088-9051 (Linking). doi:10.1101/gr.132159.111. 9.

Desplan C, Theis J, O'Farrell PH, 1988. The sequence specificity of homeodomain-dna interaction. *Cell*, **54**: 1081–90. ISSN 0092-8674 (Print) 0092-8674 (Linking). 7.

Di Ruscio A, Ebralidze AK, Benoukraf T, Amabile G, Goff LA, Terragni J, Figueroa ME, De Figueiredo Pontes LL, Alberich-Jorda M, Zhang P, Wu M, D'Alo F, Melnick A, Leone G, Ebralidze KK, Pradhan S, Rinn JL, Tenen DG, 2013. Dnmt1-interacting rnas block gene-specific dna methylation. *Nature*. ISSN 1476-4687 (Electronic) 0028-0836 (Linking). doi:10.1038/nature12598.

Ding D, Lipshitz HD, 1994. Spatially regulated expression of retrovirus-like transposons during drosophila melanogaster embryogenesis. *Genetical Research*, **64**: 167–181. ISSN 0016-6723. 3.

Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B, 2012. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, **485**: 376–80. ISSN 1476-4687 (Electronic) 0028-0836 (Linking). doi:10.1038/nature11082. 7398.

Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, Xue C, Marinov GK, Khatun J, Williams BA, Zaleski C, Rozowsky J, Roder M, Kokocinski F, Abdelhamid RF, Alioto T, Antoshechkin I, Baer MT, Bar NS, Batut P, Bell K, Bell I, Chakrabortty S, Chen X, Chrast J, Curado J, Derrien T, Drenkow J, Dumais E, Dumais J, Duttagupta R, Falconnet E, Fastuca M, Fejes-Toth K, Ferreira P, Foissac S, Fullwood MJ, Gao H, Gonzalez D, Gordon A, Gunawardena H, Howald C, Jha S, Johnson R, Kapranov P, King B, Kingswood C, Luo OJ, Park E, Persaud K, Preall JB, Ribeca P, Risk B, Robyr D, Sammeth M, Schaffer L, See LH, Shahab A, Skancke J, Suzuki AM, Takahashi H, Tilgner H, Trout D, Walters N, Wang H, Wrobel J, Yu Y, Ruan X, Hayashizaki Y, Harrow J, Gerstein M, Hubbard T, Reymond A, Antonarakis SE, Hannon G, Giddings MC, Ruan Y, Wold B, Carninci P, Guigo R, Gingeras TR, 2012. Landscape of transcription in human cells. *Nature*, **489**: 101–8. ISSN 1476-4687 (Electronic) 0028-0836 (Linking). doi:10.1038/nature11233. 7414.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR, 2012. Star: ultrafast universal rna-seq aligner. *Bioinformatics*, **29**: 15–21. ISSN 1367-4811 (Electronic) 1367-4803 (Linking). doi:10.1093/bioinformatics/bts635. 1.

Drapeau MD, Cyran SA, Viering MM, Geyer PK, Long AD, 2006. A cis-regulatory sequence within the yellow locus of drosophila melanogaster required for normal male mating success. *Genetics*, **172**: 1009–30. ISSN 0016-6731 (Print) 0016-6731 (Linking). doi:10.1534/genetics.105.045666. 2.

Drapeau MD, Radovic A, Wittkopp PJ, Long AD, 2003. A gene necessary for normal male courtship, yellow, acts downstream of fruitless in the drosophila melanogaster larval brain. *J Neurobiol*, **55**: 53–72. ISSN 0022-3034 (Print) 0022-3034 (Linking). doi:10.1002/neu.10196. 1.

Duret L, Chureau C, Samain S, Weissenbach J, Avner P, 2006. The xist rna gene evolved in eutherians by pseudogenization of a protein-coding gene. *Science*, **312**: 1653–1655. ISSN 0036-8075. doi: 10.1126/Science.1126316. 5780.

Elisaphenko EA, Kolesnikov NN, Shevchenko AI, Rogozin IB, Nesterova TB, Brockdorff N, Zakian SM, 2008. A dual origin of the xist gene from a protein-coding gene and a set of transposable elements. *PLoS One*, **3**: e2521. ISSN 1932-6203 (Electronic) 1932-6203 (Linking). doi:10.1371/journal.pone.0002521. 6.

Engreitz JM, Pandya-Jones A, McDonel P, Shishkin A, Sirokman K, Surka C, Kadri S, Xing J, Goren A, Lander ES, Plath K, Guttman M, 2013. The xist lncrna exploits three-dimensional genome architecture to spread across the x chromosome. *Science*. ISSN 1095-9203 (Electronic) 0036-8075 (Linking). doi:10.1126/science.1237973.

Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M, Ku M, Durham T, Kellis M, Bernstein BE, 2011. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, **473**: 43–9. ISSN 1476-4687 (Electronic) 0028-0836 (Linking). doi:10.1038/nature09906. 7345.

Faghihi MA, Modarresi F, Khalil AM, Wood DE, Sahagan BG, Morgan TE, Finch CE, St Laurent r G, Kenny PJ, Wahlestedt C, 2008. Expression of a noncoding rna is elevated in alzheimer's disease and drives rapid feed-forward regulation of beta-secretase. *Nat Med*, **14**: 723–30. ISSN 1546-170X (Electronic) 1078-8956 (Linking). doi:10.1038/nm1784. 7.

Farazi TA, Hoell JI, Morozov P, Tuschl T, 2013. Micrornas in human cancer. *Adv Exp Med Biol*, **774**: 1–20. ISSN 0065-2598 (Print) 0065-2598 (Linking). doi:10.1007/978-94-007-5590-1\ensuremath{_}1.

Faulkner GJ, Kimura Y, Daub CO, Wani S, Plessy C, Irvine KM, Schroder K, Cloonan N, Steptoe AL, Lassmann T, Waki K, Hornig N, Arakawa T, Takahashi H, Kawai J, Forrest ARR, Suzuki H, Hayashizaki Y, Hume DA, Orlando V, Grimmond SM, Carninci P, 2009. The regulated retrotransposon transcriptome of mammalian cells. *Nature Genetics*, **41**: 563–571. ISSN 1061-4036. doi: 10.1038/Ng.368. 5.

Fay JC, Wittkopp PJ, 2008. Evaluating the role of natural selection in the evolution of gene regulation. *Heredity (Edinb)*, **100**: 191–9. ISSN 1365-2540 (Electronic) 0018-067X (Linking). doi:10.1038/sj.hdy.6801000. 2.

Fejes-Toth K, Sotirova V, Sachidanandam R, Assaf G, Hannon GJ, Kapranov P, Foissac S, Willingham AT, Duttagupta R, Dumais E, Gingeras TR, Lab ACSH, 2009. Post-transcriptional processing generates a diversity of 5'-modified long and short rnas. *Nature*, **457**: 1028–1032. ISSN 0028-0836. doi:10.1038/Nature07759. 7232.

Feschotte C, 2008. Transposable elements and the evolution of regulatory networks. *Nature Reviews Genetics*, **9**: 397–405. ISSN 1471-0056. doi:10.1038/Nrg2337. 5.

Fire A, Xu S, Montgomery MK, Kostas SA, Driver SE, Mello CC, 1998. Potent and specific genetic interference by double-stranded rna in caenorhabditis elegans. *Nature*, **391**: 806–11. ISSN 0028-0836 (Print) 0028-0836 (Linking). doi:10.1038/35888. 6669.

FitzGerald PC, Sturgill D, Shyakhtenko A, Oliver B, Vinson C, 2006. Comparative genomics of drosophila and human core promoters. *Genome Biol*, **7**: R53. ISSN 1465-6914 (Electronic) 1465-6906 (Linking). doi:10.1186/gb-2006-7-7-r53. 7.

Fraser HB, 2013. Gene expression drives local adaptation in humans. *Genome Res*. ISSN 1549-5469 (Electronic) 1088-9051 (Linking). doi:10.1101/gr.152710.112.

Gerstein MB, Lu ZJ, Van Nostrand EL, Cheng C, Arshinoff BI, Liu T, Yip KY, Robilotto R, Rechtsteiner A, Ikegami K, Alves P, Chateigner A, Perry M, Morris M, Auerbach RK, Feng X, Leng J, Vielle A, Niu W, Rhrissorrakrai K, Agarwal A, Alexander RP, Barber G, Brdlik CM, Brennan J, Brouillet JJ, Carr A, Cheung MS, Clawson H, Contrino S, Dannenberg LO, Dernburg AF, Desai A, Dick L, Dose AC, Du J, Egelhofer T, Ercan S, Euskirchen G, Ewing B, Feingold EA, Gassmann R, Good PJ, Green P, Gullier F, Gutwein M, Guyer MS, Habegger L, Han T, Henikoff JG, Henz SR, Hinrichs A, Holster H, Hyman T, Iniguez AL, Janette J, Jensen M, Kato M, Kent WJ, Kephart E, Khivansara V, Khurana E, Kim JK, Kolasinska-Zwierz P, Lai EC, Latorre I, Leahey A, Lewis S, Lloyd P, Lochovsky L, Lowdon RF, Lubling Y, Lyne R, MacCoss M, Mackowiak SD, Mangone M, McKay S, Mecenas D, Merrihew G, Miller r D M, Muroyama A, Murray JI, Ooi SL, Pham H, Phippen T, Preston EA, Rajewsky N, Ratsch G, Rosenbaum H, Rozowsky J, Rutherford K, Ruzanov P, Sarov M, Sasidharan R, Sboner A, Scheid P, Segal E, Shin H, Shou C, Slack FJ, Slightam C, Smith R, Spencer WC, Stinson EO, Taing S, Takasaki T, Vafeados D, Voronina K, Wang G, Washington NL, Whittle CM, Wu B, Yan KK, Zeller G, Zha Z, Zhong M, Zhou X, Ahringer J, Strome S, Gunsalus KC, Micklem G, Liu XS, Reinke V, Kim SK, Hillier LW, Henikoff S, Piano F, Snyder M, Stein L, Lieb JD, Waterston RH, 2010. Integrative analysis of the caenorhabditis elegans genome by the modencode project. *Science*, **330**: 1775–87. ISSN 1095-9203 (Electronic) 0036-8075 (Linking). doi:10.1126/science.1196914. 6012.

Ghildiyal M, Seitz H, Horwich MD, Li CJ, Du TT, Lee S, Xu J, Kittler ELW, Zapp ML, Weng ZP, Zamore PD, 2008. Endogenous sirnas derived from transposons and mrnas in drosophila somatic cells. *Science*, **320**: 1077–1081. ISSN 0036-8075. doi:10.1126/Science.1157396. 5879.

Ghildiyal M, Zamore PD, 2009. Small silencing rnas: an expanding universe. *Nature Reviews Genetics*, **10**: 94–108. ISSN 1471-0056. doi:10.1038/Nrg2504. 2.

Giorgetti L, Siggers T, Tiana G, Caprara G, Notarbartolo S, Corona T, Pasparakis M, Milani P, Bulyk ML, Natoli G, 2010. Noncooperative interactions between transcription factors and clustered dna binding sites enable graded transcriptional responses to environmental inputs. *Mol Cell*, **37**: 418–28. ISSN 1097-4164 (Electronic) 1097-2765 (Linking). doi:10.1016/j.molcel.2010.01.016. 3.

Gompel N, Prud'homme B, Wittkopp PJ, Kassner VA, Carroll SB, 2005. Chance caught on the wing: cis-regulatory evolution and the origin of pigment patterns in drosophila. *Nature*, **433**: 481–487. ISSN 0028-0836. doi:10.1038/Nature03235. 7025.

Goodrich JA, Tjian R, 2010. Unexpected roles for core promoter recognition factors in cell-type-specific transcription and gene regulation. *Nature Reviews Genetics*, **11**: 549–558. ISSN 1471-0056. doi: 10.1038/Nrg2847. 8.

Graveley BR, Brooks AN, Carlson JW, Duff MO, Landolin JM, Yang L, Artieri CG, van Baren MJ, Boley N, Booth BW, Brown JB, Cherbas L, Davis CA, Dobin A, Li R, Lin W, Malone JH, Mattiuzzo NR, Miller D, Sturgill D, Tuch BB, Zaleski C, Zhang D, Blanchette M, Dudoit S, Eads B, Green RE, Hammonds A, Jiang L, Kapranov P, Langton L, Perrimon N, Sandler JE, Wan KH, Willingham A, Zhang Y, Zou Y, Andrews J, Bickel PJ, Brenner SE, Brent MR, Cherbas P, Gingeras TR, Hoskins RA, Kaufman TC, Oliver B, Celniker SE, 2010. The developmental transcriptome of drosophila melanogaster. *Nature*. ISSN 1476-4687 (Electronic) 0028-0836 (Linking). doi:10.1038/nature09715.

Grenier JK, Garber TL, Warren R, Whitington PM, Carroll S, 1997. Evolution of the entire arthropod hox gene set predated the origin and radiation of the onychophoran/arthropod clade. *Curr Biol*, **7**: 547–53. ISSN 0960-9822 (Print) 0960-9822 (Linking). 8.

Gupta RA, Shah N, Wang KC, Kim J, Horlings HM, Wong DJ, Tsai MC, Hung T, Argani P, Rinn JL, Wang Y, Brzoska P, Kong B, Li R, West RB, van de Vijver MJ, Sukumar S, Chang HY, 2010. Long non-coding rna hotair reprograms chromatin state to promote cancer metastasis. *Nature*, **464**: 1071–6. ISSN 1476-4687 (Electronic) 0028-0836 (Linking). doi:10.1038/nature08975. 7291.

Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, Huarte M, Zuk O, Carey BW, Cassady JP, Cabili MN, Jaenisch R, Mikkelsen TS, Jacks T, Hacohen N, Bernstein BE, Kellis M, Regev A, Rinn JL, Lander ES, 2009. Chromatin signature reveals over a thousand highly conserved large non-coding rnas in mammals. *Nature*, **458**: 223–227. ISSN 0028-0836. doi:10.1038/Nature07672. 7235.

Guttman M, Donaghey J, Carey BW, Garber M, Grenier JK, Munson G, Young G, Lucas AB, Ach R, Bruhn L, Yang X, Amit I, Meissner A, Regev A, Rinn JL, Root DE, Lander ES, 2011. lincrnas act in the circuitry controlling pluripotency and differentiation. *Nature*, **477**: 295–300. ISSN 1476-4687 (Electronic) 0028-0836 (Linking). doi:10.1038/nature10398. 7364.

Guttman M, Rinn JL, 2012. Modular regulatory principles of large non-coding rnas. *Nature*, **482**: 339–46. ISSN 1476-4687 (Electronic) 0028-0836 (Linking). doi:10.1038/nature10887. 7385.

Haerty W, Ponting CP, 2013. Mutations within lncrnas are effectively selected against in fruitfly but not in human. *Genome Biol*, **14**: R49. ISSN 1465-6914 (Electronic) 1465-6906 (Linking). doi: 10.1186/gb-2013-14-5-r49. 5.

Hansen KH, Helin K, 2009. Epigenetic inheritance through self-recruitment of the polycomb repressive complex 2. *Epigenetics*, **4**: 133–138. ISSN 1559-2294. 3.

Hare EE, Peterson BK, Iyer VN, Meier R, Eisen MB, 2008. Sepsid even-skipped enhancers are functionally conserved in drosophila despite lack of sequence conservation. *PLoS Genet*, **4**: e1000106. ISSN 1553-7404 (Electronic) 1553-7390 (Linking). doi:10.1371/journal.pgen.1000106. 6.

He Q, Bardet AF, Patton B, Purvis J, Johnston J, Paulson A, Gogol M, Stark A, Zeitlinger J, 2011. High conservation of transcription factor binding and evidence for combinatorial regulation across six drosophila species. *Nat Genet*, **43**: 414–20. ISSN 1546-1718 (Electronic) 1061-4036 (Linking). doi:10.1038/ng.808. 5.

Heintzman ND, Hon GC, Hawkins RD, Kheradpour P, Stark A, Harp LF, Ye Z, Lee LK, Stuart RK, Ching CW, Ching KA, Antosiewicz-Bourget JE, Liu H, Zhang XM, Green RD, Lobanenkov VV, Stewart R, Thomson JA, Crawford GE, Kellis M, Ren B, 2009. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, **459**: 108–112. ISSN 0028-0836. doi:10.1038/Nature07829. 7243.

Heinz S, Romanoski CE, Benner C, Allison KA, Kaikkonen MU, Orozco LD, Glass CK, 2013. Effect of natural genetic variation on enhancer selection and function. *Nature*, **503**: 487–92. ISSN 1476-4687 (Electronic) 0028-0836 (Linking). doi:10.1038/nature12615. 7477.

Heo JB, Sung S, 2011. Vernalization-mediated epigenetic silencing by a long intronic noncoding rna. *Science*, **331**: 76–9. ISSN 1095-9203 (Electronic) 0036-8075 (Linking). doi:10.1126/science. 1197349. 6013.

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA, 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A*, **106**: 9362–7. ISSN 1091-6490 (Electronic) 0027-8424 (Linking). doi: 10.1073/pnas.0903103106. 23.

Hirzmann J, Luo D, Hahnen J, Hobom G, 1993. Determination of messenger-rna 5'-ends by reverse transcription of the cap structure. *Nucleic Acids Research*, **21**: 3597–3598. ISSN 0305-1048. 15.

Ho MCW, Johnsen H, Goetz SE, Schiller BJ, Bae E, Tran DA, Shur AS, Allen JM, Rau C, Bender W, Fisher WW, Celniker SE, Drewell RA, 2009. Functional evolution of cis-regulatory modules at a homeotic gene in drosophila. *Plos Genetics*, **5**: –. ISSN 1553-7390. doi:10.1371/Journal.Pgen. 1000709. 11.

Hoey T, Levine M, 1988. Divergent homeo box proteins recognize similar dna sequences in drosophila. *Nature*, **332**: 858–61. ISSN 0028-0836 (Print) 0028-0836 (Linking). doi:10.1038/332858a0. 6167.

Hoque M, Ji Z, Zheng D, Luo W, Li W, You B, Park JY, Yehia G, Tian B, 2013. Analysis of alternative cleavage and polyadenylation by 3' region extraction and deep sequencing. *Nat Methods*, **10**: 133–9. ISSN 1548-7105 (Electronic) 1548-7091 (Linking). doi:10.1038/nmeth.2288. 2.

Hoskins RA, Landolin JM, Brown JB, Sandler JE, Takahashi H, Lassmann T, Yu C, Booth BW, Zhang D, Wan KH, Yang L, Boley N, Andrews J, Kaufman TC, Graveley BR, Bickel PJ, Carninci P, Carlson JW, Celniker SE, 2011. Genome-wide analysis of promoter architecture in drosophila melanogaster. *Genome Res*. ISSN 1549-5469 (Electronic) 1088-9051 (Linking). doi:10.1101/gr.112466.110.

Houseley J, Rubbi L, Grunstein M, Tollervey D, Vogelauer M, 2008. A ncrna modulates histone modification and mrna induction in the yeast gal gene cluster. *Mol Cell*, **32**: 685–95. ISSN 1097-4164 (Electronic) 1097-2765 (Linking). doi:10.1016/j.molcel.2008.09.027. 5.

Huarte M, Guttman M, Feldser D, Garber M, Koziol MJ, Kenzelmann-Broz D, Khalil AM, Zuk O, Amit I, Rabani M, Attardi LD, Regev A, Lander ES, Jacks T, Rinn JL, 2010. A large intergenic noncoding rna induced by p53 mediates global gene repression in the p53 response. *Cell*, **142**: 409–419. ISSN 0092-8674. doi:10.1016/J.Cell.2010.06.040. 3.

Ilsley GR, Fisher J, Apweiler R, Depace AH, Luscombe NM, 2013. Cellular resolution models for even skipped regulation in the entire drosophila embryo. *Elife*, **2**: e00522. ISSN 2050-084X (Electronic). doi:10.7554/eLife.00522.

Ishii N, Ozaki K, Sato H, Mizuno H, Saito S, Takahashi A, Miyamoto Y, Ikegawa S, Kamatani N, Hori M, Nakamura Y, Tanaka T, 2006. Identification of a novel non-coding rna, miat, that confers risk of myocardial infarction. *J Hum Genet*, **51**: 1087–99. ISSN 1434-5161 (Print) 1434-5161 (Linking). doi:10.1007/s10038-006-0070-9. 12.

Islam S, Kjallquist U, Moliner A, Zajac P, Fan JB, Lonnerberg P, Linnarsson S, 2011. Characterization of the single-cell transcriptional landscape by highly multiplex rna-seq. *Genome Res*, **21**: 1160–7. ISSN 1549-5469 (Electronic) 1088-9051 (Linking). doi:10.1101/gr.110882.110. 7.

Jacob F, 1977. Evolution and tinkering. *Science*, **196**: 1161–6. ISSN 0036-8075 (Print) 0036-8075 (Linking). 4295.

Jacob F, Monod J, 1961. Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol*, **3**: 318–56. ISSN 0022-2836 (Print) 0022-2836 (Linking).

Jacquier A, 2009. The complex eukaryotic transcriptome: unexpected pervasive transcription and novel small rnas. *Nat Rev Genet*, **10**: 833–44. ISSN 1471-0064 (Electronic) 1471-0056 (Linking). doi: 10.1038/nrg2683. 12.

Jan CH, Friedman RC, Ruby JG, Bartel DP, 2011. Formation, regulation and evolution of caenorhabditis elegans 3'utrs. *Nature*, **469**: 97–101. ISSN 1476-4687 (Electronic) 0028-0836 (Linking). doi: 10.1038/nature09616. 7328.

Jeon Y, Lee JT, 2011. Yy1 tethers xist rna to the inactive x nucleation center. *Cell*, **146**: 119–33. ISSN 1097-4172 (Electronic) 0092-8674 (Linking). doi:10.1016/j.cell.2011.06.026. 1.

Jin F, Li Y, Dixon JR, Selvaraj S, Ye Z, Lee AY, Yen CA, Schmitt AD, Espinoza CA, Ren B, 2013. A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature*, **503**: 290–4. ISSN 1476-4687 (Electronic) 0028-0836 (Linking). doi:10.1038/nature12644. 7475.

Johnson AD, Meyer BJ, Ptashne M, 1979. Interactions between dna-bound repressors govern regulation by the lambda phage repressor. *Proc Natl Acad Sci U S A*, **76**: 5061–5. ISSN 0027-8424 (Print) 0027-8424 (Linking). 10.

Jones FC, Grabherr MG, Chan YF, Russell P, Mauceli E, Johnson J, Swofford R, Pirun M, Zody MC, White S, Birney E, Searle S, Schmutz J, Grimwood J, Dickson MC, Myers RM, Miller CT, Summers BR, Knecht AK, Brady SD, Zhang H, Pollen AA, Howes T, Amemiya C, Baldwin J, Bloom T, Jaffe DB, Nicol R, Wilkinson J, Lander ES, Di Palma F, Lindblad-Toh K, Kingsley DM, 2012. The genomic basis of adaptive evolution in threespine sticklebacks. *Nature*, **484**: 55–61. ISSN 1476-4687 (Electronic) 0028-0836 (Linking). doi:10.1038/nature10944. 7392.

Kagey MH, Newman JJ, Bilodeau S, Zhan Y, Orlando DA, van Berkum NL, Ebmeier CC, Goossens J, Rahl PB, Levine SS, Taatjes DJ, Dekker J, Young RA, 2010. Mediator and cohesin connect gene expression and chromatin architecture. *Nature*, **467**: 430–435. ISSN 0028-0836. doi:10.1038/ Nature09380. 7314.

Kalinka AT, Varga KM, Gerrard DT, Preibisch S, Corcoran DL, Jarrells J, Ohler U, Bergman CM, Tomancak P, 2010. Gene expression divergence recapitulates the developmental hourglass model. *Nature*, **468**: 811–U102. ISSN 0028-0836. doi:10.1038/Nature09634. 7325.

Kapranov P, St Laurent G, 2012. Dark matter rna: Existence, function, and controversy. *Front Genet*, **3**: 60. ISSN 1664-8021 (Electronic) 1664-8021 (Linking). doi:10.3389/fgene.2012.00060.

Kapranov P, Willingham AT, Gingeras TR, 2007. Genome-wide transcription and the implications for genomic organization. *Nat Rev Genet*, **8**: 413–23. ISSN 1471-0056 (Print). doi:10.1038/nrg2083. 6.

Kapusta A, Kronenberg Z, Lynch VJ, Zhuo X, Ramsay L, Bourque G, Yandell M, Feschotte C, 2013. Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding rnas. *PLoS Genet*, **9**: e1003470. ISSN 1553-7404 (Electronic) 1553-7390 (Linking). doi:10.1371/journal.pgen.1003470. 4.

Kazazian HH, 2004. Mobile elements: Drivers of genome evolution. *Science*, **303**: 1626–1632. ISSN 0036-8075. 5664.

Keniry A, Oxley D, Monnier P, Kyba M, Dandolo L, Smits G, Reik W, 2012. The h19 lincrna is a developmental reservoir of mir-675 that suppresses growth and igf1r. *Nat Cell Biol*, **14**: 659–65. ISSN 1476-4679 (Electronic) 1465-7392 (Linking). doi:10.1038/ncb2521. 7.

Kerber B, Fellert S, Taubert H, Hoch M, 1996. Germ line and embryonic expression of fex, a member of the drosophila f-element retrotransposon family, is mediated by an internal cis-regulatory control region. *Mol Cell Biol*, **16**: 2998–3007. ISSN 0270-7306 (Print) 0270-7306 (Linking). 6.

Khaitovich P, Hellmann I, Enard W, Nowick K, Leinweber M, Franz H, Weiss G, Lachmann M, Paabo S, 2005. Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. *Science*, **309**: 1850–4. ISSN 1095-9203 (Electronic) 0036-8075 (Linking). doi:10.1126/science.1108296. 5742.

Khaitovich P, Weiss G, Lachmann M, Hellmann I, Enard W, Muetzel B, Wirkner U, Ansorge W, Paabo S, 2004. A neutral model of transcriptome evolution. *Plos Biology*, **2**: 682–689. ISSN 1544-9173. doi:10.1371/journal.pbio.0020132. 5.

Khan Z, Ford MJ, Cusanovich DA, Mitrano A, Pritchard JK, Gilad Y, 2013. Primate transcript and protein expression levels evolve under compensatory selection pressures. *Science*, **342**: 1100–4. ISSN 1095-9203 (Electronic) 0036-8075 (Linking). doi:10.1126/science.1242379. 6162.

Kheradpour P, Ernst J, Melnikov A, Rogov P, Wang L, Zhang X, Alston J, Mikkelsen TS, Kellis M, 2013. Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome Res*, **23**: 800–11. ISSN 1549-5469 (Electronic) 1088-9051 (Linking). doi:10.1101/gr.144899.112. 5.

Kim TK, Hemberg M, Gray JM, Costa AM, Bear DM, Wu J, Harmin DA, Laptewicz M, Barbara-Haley K, Kuersten S, Markenscoff-Papadimitriou E, Kuhl D, Bito H, Worley PF, Kreiman G, Greenberg ME, 2010. Widespread transcription at neuronal activity-regulated enhancers. *Nature*, **465**: 182–7. ISSN 1476-4687 (Electronic) 0028-0836 (Linking). doi:10.1038/nature09033. 7295.

King MC, Wilson AC, 1975. Evolution at 2 levels in humans and chimpanzees. *Science*, **188**: 107–116. ISSN 0036-8075. 4184.

Kinney JB, Murugan A, Callan J C G, Cox EC, 2010. Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proc Natl Acad Sci U S A*, **107**: 9158–63. ISSN 1091-6490 (Electronic) 0027-8424 (Linking). doi:10.1073/pnas.1004290107. 20.

Kino T, Hurt DE, Ichijo T, Nader N, Chrousos GP, 2010. Noncoding rna gas5 is a growth arrest- and starvation-associated repressor of the glucocorticoid receptor. *Sci Signal*, **3**: ra8. ISSN 1937-9145 (Electronic). doi:10.1126/scisignal.2000568. 107.

Kirschner M, Gerhart J, 1998. Evolvability. *Proceedings of the National Academy of Sciences of the United States of America*, **95**: 8420–8427. ISSN 0027-8424. 15.

Kodzius R, Kojima M, Nishiyori H, Nakamura M, Fukuda S, Tagami M, Sasaki D, Imamura K, Kai C, Harbers M, Hayashizaki Y, Carninci P, 2006. Cage: cap analysis of gene expression. *Nature Methods*, **3**: 211–222. ISSN 1548-7091. 3.

Kouzarides T, 2007. Chromatin modifications and their function. *Cell*, **128**: 693–705. ISSN 0092-8674 (Print) 0092-8674 (Linking). doi:10.1016/j.cell.2007.02.005. 4.

Krol J, Loedige I, Filipowicz W, 2010. The widespread regulation of microrna biogenesis, function and decay. *Nat Rev Genet*, **11**: 597–610. ISSN 1471-0064 (Electronic) 1471-0056 (Linking). doi:10.1038/nrg2843. 9.

Kunarso G, Chia NY, Jeyakani J, Hwang C, Lu X, Chan YS, Ng HH, Bourque G, 2010. Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat Genet*. ISSN 1546-1718 (Electronic) 1061-4036 (Linking). doi:10.1038/ng.600.

Kutter C, Watt S, Stefflova K, Wilson MD, Goncalves A, Ponting CP, Odom DT, Marques AC, 2012. Rapid turnover of long noncoding rnas and the evolution of gene expression. *PLoS Genet*, **8**: e1002841. ISSN 1553-7404 (Electronic) 1553-7390 (Linking). doi:10.1371/journal.pgen.1002841. 7.

Kwak H, Fuda NJ, Core LJ, Lis JT, 2013. Precise maps of rna polymerase reveal how promoters direct initiation and pausing. *Science*, **339**: 950–3. ISSN 1095-9203 (Electronic) 0036-8075 (Linking). doi:10.1126/science.1229386. 6122.

Lagha M, Bothma JP, Esposito E, Ng S, Stefanik L, Tsui C, Johnston J, Chen K, Gilmour DS, Zeitlinger J, Levine MS, 2012. Paused pol ii coordinates tissue morphogenesis in the drosophila embryo. *Cell*, **153**: 976–87. ISSN 1097-4172 (Electronic) 0092-8674 (Linking). doi:10.1016/j.cell.2013.04.045. 5.

Lanctot C, Cheutin T, Cremer M, Cavalli G, Cremer T, 2007. Dynamic genome architecture in the nuclear space: regulation of gene expression in three dimensions. *Nat Rev Genet*, **8**: 104–15. ISSN 1471-0056 (Print) 1471-0056 (Linking). doi:10.1038/nrg2041. 2.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blocker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan

MJ, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ, 2001. Initial sequencing and analysis of the human genome. *Nature*, **409**: 860–921. ISSN 0028-0836 (Print) 0028-0836 (Linking). doi:10.1038/35057062. 6822.

Landry CR, Wittkopp PJ, Taubes CH, Ranz JM, Clark AG, Hartl DL, 2005. Compensatory cis-trans evolution and the dysregulation of gene expression in interspecific hybrids of drosophila. *Genetics*, **171**: 1813–22. ISSN 0016-6731 (Print) 0016-6731 (Linking). doi:10.1534/genetics.105.047449. 4.

Lappalainen T, Sammeth M, Friedlander MR, t Hoen PA, Monlong J, Rivas MA, Gonzalez-Porta M, Kurbatova N, Griebel T, Ferreira PG, Barann M, Wieland T, Greger L, van Iterson M, Almlof J, Ribeca P, Pulyakhina I, Esser D, Giger T, Tikhonov A, Sultan M, Bertier G, MacArthur DG, Lek M, Lizano E, Buermans HP, Padioleau I, Schwarzmayr T, Karlberg O, Ongen H, Kilpinen H, Beltran S, Gut M, Kahlem K, Amstislavskiy V, Stegle O, Pirinen M, Montgomery SB, Donnelly P, McCarthy MI, Flicek P, Strom TM, Lehrach H, Schreiber S, Sudbrak R, Carracedo A, Antonarakis SE, Hasler R, Syvanen AC, van Ommen GJ, Brazma A, Meitinger T, Rosenstiel P, Guigo R, Gut IG, Estivill X, Dermitzakis ET, Palotie A, Deleuze JF, Gyllensten U, Brunner H, Veltman J, Cambon-Thomsen A, Mangion J, Bentley D, Hamosh A, 2013. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, **501**: 506–11. ISSN 1476-4687 (Electronic) 0028-0836 (Linking). doi:10.1038/nature12531. 7468.

Lasa I, Toledo-Arana A, Dobin A, Villanueva M, de los Mozos IR, Vergara-Irigaray M, Segura V, Fagegaltier D, Penades JR, Valle J, Solano C, Gingeras TR, 2011. Genome-wide antisense transcription drives mrna processing in bacteria. *Proc Natl Acad Sci U S A*, **108**: 20172–7. ISSN 1091-6490 (Electronic) 0027-8424 (Linking). doi:10.1073/pnas.1113521108. 50.

Lebrecht D, Foehr M, Smith E, Lopes FJ, Vanario-Alonso CE, Reinitz J, Burz DS, Hanes SD, 2005. Bicoid cooperative dna binding is critical for embryonic patterning in drosophila. *Proc Natl Acad Sci U S A*, **102**: 13176–81. ISSN 0027-8424 (Print) 0027-8424 (Linking). doi:10.1073/pnas.0506462102. 37.

Lenhard B, Sandelin A, Carninci P, 2012. Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nat Rev Genet*, **13**: 233–45. ISSN 1471-0064 (Electronic) 1471-0056 (Linking). doi:10.1038/nrg3163. 4.

Levin JZ, Yassour M, Adiconis XA, Nusbaum C, Thompson DA, Friedman N, Gnirke A, Regev A, 2010. Comprehensive comparative analysis of strand-specific rna sequencing methods. *Nature Methods*, **7**: 709–U67. ISSN 1548-7091. doi:10.1038/Nmeth.1491. 9.

Levine M, Hoey T, 1988. Homeobox proteins as sequence-specific transcription factors. *Cell*, **55**: 537–40. ISSN 0092-8674 (Print) 0092-8674 (Linking). 4.

Lewis EB, 1978. A gene complex controlling segmentation in drosophila. *Nature*, **276**: 565–70. ISSN 0028-0836 (Print) 0028-0836 (Linking). 5688.

Li JJ, Jiang CR, Brown JB, Huang H, Bickel PJ, 2011. Sparse linear modeling of next-generation mrna sequencing (rna-seq) data for isoform discovery and abundance estimation. *Proc Natl Acad Sci U S A*, **108**: 19867–72. ISSN 1091-6490 (Electronic) 0027-8424 (Linking). doi:10.1073/pnas.1113972108. 50.

Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, Kheradpour P, Ernst J, Jordan G, Mauceli E, Ward LD, Lowe CB, Holloway AK, Clamp M, Gnerre S, Alfoldi J, Beal K, Chang J,

Clawson H, Cuff J, Di Palma F, Fitzgerald S, Flicek P, Guttman M, Hubisz MJ, Jaffe DB, Jungreis I, Kent WJ, Kostka D, Lara M, Martins AL, Massingham T, Moltke I, Raney BJ, Rasmussen MD, Robinson J, Stark A, Vilella AJ, Wen JY, Xie XH, Zody MC, Worley KC, Kovar CL, Muzny DM, Gibbs RA, Warren WC, Mardis ER, Weinstock GM, Wilson RK, Birney E, Margulies EH, Herrero J, Green ED, Haussler D, Siepel A, Goldman N, Pollard KS, Pedersen JS, Lander ES, Kellis M, Inst B, Med BC, Univ W, 2011. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature*, **478**: 476–482. ISSN 0028-0836. doi:10.1038/Nature10530. 7370.

Lipatov M, Lenkov K, Petrov DA, Bergman CM, 2005. Paucity of chimeric gene-transposable element transcripts in the drosophila melanogaster genome. *BMC Biol*, **3**: 24. ISSN 1741-7007 (Electronic). doi:10.1186/1741-7007-3-24.

Lippman Z, Gendrel AV, Black M, Vaughn MW, Dedhia N, McCombie WR, Lavine K, Mittal V, May B, Kasschau KD, Carrington JC, Doerge RW, Colot V, Martienssen R, 2004. Role of transposable elements in heterochromatin and epigenetic control. *Nature*, **430**: 471–476. ISSN 0028-0836. doi: 10.1038/Nature02651. 6998.

Loehlin DW, Werren JH, 2012. Evolution of shape by multiple regulatory changes to a growth gene. *Science*, **335**: 943–7. ISSN 1095-9203 (Electronic) 0036-8075 (Linking). doi:10.1126/science.1215193. 6071.

Lowe CB, Bejerano G, Haussler D, 2007. Thousands of human mobile element fragments undergo strong purifying selection near developmental genes. *Proceedings of the National Academy of Sciences of the United States of America*, **104**: 8005–8010. ISSN 0027-8424. doi:10.1073/Pnas. 0611223104. 19.

Lowe CB, Kellis M, Siepel A, Raney BJ, Clamp M, Salama SR, Kingsley DM, Lindblad-Toh K, Haussler D, 2011. Three periods of regulatory innovation during vertebrate evolution. *Science*, **333**: 1019–24. ISSN 1095-9203 (Electronic) 0036-8075 (Linking). doi:10.1126/science.1202702. 6045.

Ludwig MZ, Bergman C, Patel NH, Kreitman M, 2000. Evidence for stabilizing selection in a eukaryotic enhancer element. *Nature*, **403**: 564–567. ISSN 0028-0836. 6769.

Ludwig MZ, Palsson A, Alekseeva E, Bergman CM, Nathan J, Kreitman M, 2005. Functional evolution of a cis-regulatory module. *PLoS Biol*, **3**: e93. ISSN 1545-7885 (Electronic) 1544-9173 (Linking). doi:10.1371/journal.pbio.0030093. 4.

Ludwig MZ, Patel NH, Kreitman M, 1998. Functional analysis of eve stripe 2 enhancer evolution in drosophila: rules governing conservation and change. *Development*, **125**: 949–958. ISSN 0950-1991. 5.

Lynch VJ, Leclerc RD, May G, Wagner GP, 2011. Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals. *Nature Genetics*, **43**: 1154–U158. ISSN 1061-4036. doi:10.1038/Ng.917. 11.

Ma W, Trusina A, El-Samad H, Lim WA, Tang C, 2009. Defining network topologies that can achieve biochemical adaptation. *Cell*, **138**: 760–73. ISSN 1097-4172 (Electronic) 0092-8674 (Linking). doi:10.1016/j.cell.2009.06.013. 4.

Maamar H, Cabili MN, Rinn J, Raj A, 2013. linc-hoxa1 is a noncoding rna that represses hoxa1 transcription in cis. *Genes Dev*, **27**: 1260–71. ISSN 1549-5477 (Electronic) 0890-9369 (Linking). doi: 10.1101/gad.217018.113. 11.

Macfarlan TS, Gifford WD, Driscoll S, Lettieri K, Rowe HM, Bonanomi D, Firth A, Singer O, Trono D, Pfaff SL, 2012. Embryonic stem cell potency fluctuates with endogenous retrovirus activity. *Nature*. ISSN 1476-4687 (Electronic) 0028-0836 (Linking). doi:10.1038/nature11244.

Malik S, Roeder RG, 2010. The metazoan mediator co-activator complex as an integrative hub for transcriptional regulation. *Nature Reviews Genetics*, **11**: 761–772. ISSN 1471-0056. doi:10.1038/Nrg2901. 11.

Malone CD, Brennecke J, Dus M, Stark A, McCombie WR, Sachidanandam R, Hannon GJ, 2009. Specialized pirna pathways act in germline and somatic tissues of the drosophila ovary. *Cell*, **137**: 522–535. ISSN 0092-8674. doi:10.1016/J.Cell.2009.03.040. 3.

Manceau M, Domingues VS, Mallarino R, Hoekstra HE, 2011. The developmental role of agouti in color pattern evolution. *Science*, **331**: 1062–5. ISSN 1095-9203 (Electronic) 0036-8075 (Linking). doi:10.1126/science.1200684. 6020.

Mao YS, Sunwoo H, Zhang B, Spector DL, 2011. Direct visualization of the co-transcriptional assembly of a nuclear body by noncoding rnas. *Nat Cell Biol*, **13**: 95–101. ISSN 1476-4679 (Electronic) 1465-7392 (Linking). doi:10.1038/ncb2140. 1.

Marbach D, Roy S, Ay F, Meyer PE, Candeias R, Kahveci T, Bristow CA, Kellis M, 2012. Predictive regulatory models in drosophila melanogaster by integrative inference of transcriptional networks. *Genome Res*, **22**: 1334–49. ISSN 1549-5469 (Electronic) 1088-9051 (Linking). doi:10.1101/gr.127191.111. 7.

Marraffini LA, Sontheimer EJ, 2011. Crispr interference: Rna-directed adaptive immunity in bacteria and archaea. *Nat Rev Genet*, **11**: 181–90. ISSN 1471-0064 (Electronic) 1471-0056 (Linking). doi:10.1038/nrg2749. 3.

McClintock B, 1956. Controlling elements and the gene. *Cold Spring Harbor Symposia on Quantitative Biology*, **21**: 197–216. ISSN 0091-7451.

McGregor AP, Orgogozo V, Delon I, Zanet J, Srinivasan DG, Payre F, Stern DL, 2007. Morphological evolution through multiple cis-regulatory mutations at a single gene. *Nature*, **448**: 587–90. ISSN 1476-4687 (Electronic) 0028-0836 (Linking). doi:10.1038/nature05988. 7153.

McManus CJ, Coolon JD, Duff MO, Eipper-Mains J, Graveley BR, Wittkopp PJ, 2010. Regulatory divergence in drosophila revealed by mrna-seq. *Genome Research*, **20**: 816–825. ISSN 1088-9051. doi:10.1101/Gr.102491.109. 6.

McManus J, May G, Spealman P, Shteyman A, 2013. Ribosome profiling reveals post-transcriptional buffering of divergent gene expression in yeast. *Genome Res*. ISSN 1549-5469 (Electronic) 1088-9051 (Linking). doi:10.1101/gr.164996.113.

Meiklejohn CD, Coolon JD, Hartl DL, Wittkopp PJ, 2013. The roles of cis- and trans-regulation in the evolution of regulatory incompatibilities and sexually dimorphic gene expression. *Genome Res*. ISSN 1549-5469 (Electronic) 1088-9051 (Linking). doi:10.1101/gr.156414.113.

Melnikov A, Murugan A, Zhang X, Tesileanu T, Wang L, Rogov P, Feizi S, Gnirke A, Callan J C G, Kinney JB, Kellis M, Lander ES, Mikkelsen TS, 2012. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat Biotechnol*, **30**: 271–7. ISSN 1546-1696 (Electronic) 1087-0156 (Linking). doi:10.1038/nbt.2137. 3.

Merkin J, Russell C, Chen P, Burge CB, 2012. Evolutionary dynamics of gene and isoform regulation in mammalian tissues. *Science*, **338**: 1593–9. ISSN 1095-9203 (Electronic) 0036-8075 (Linking). doi:10.1126/science.1228186. 6114.

Moreau P, Hen R, Wasylyk B, Everett R, Gaub MP, Chambon P, 1981. The sv40 72 base repair repeat has a striking effect on gene expression both in sv40 and other chimeric recombinants. *Nucleic Acids Res*, **9**: 6047–68. ISSN 0305-1048 (Print) 0305-1048 (Linking). 22.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B, 2008. Mapping and quantifying mammalian transcriptomes by rna-seq. *Nat Methods*, **5**: 621–8. ISSN 1548-7105 (Electronic) 1548-7091 (Linking). doi:10.1038/nmeth.1226. 7.

Moses AM, Pollard DA, Nix DA, Iyer VN, Li XY, Biggin MD, Eisen MB, 2006. Large-scale turnover of functional transcription factor binding sites in drosophila. *Plos Computational Biology*, **2**: 1219–1231. ISSN 1553-734X. doi:10.1371/journal.pcbi.0020130. 10.

Mozer BA, Benzer S, 1994. Ingrowth by photoreceptor axons induces transcription of a retrotransposon in the developing drosophila brain. *Development*, **120**: 1049–1058. ISSN 0950-1991. 5.

Nagano T, Fraser P, 2011. No-nonsense functions for long noncoding rnas. *Cell*, **145**: 178–181. ISSN 0092-8674. doi:10.1016/J.Cell.2011.03.014. 2.

Naito K, Zhang F, Tsukiyama T, Saito H, Hancock CN, Richardson AO, Okumoto Y, Tanisaka T, Wessler SR, 2009. Unexpected consequences of a sudden and massive transposon amplification on rice gene expression. *Nature*, **461**: 1130–U232. ISSN 0028-0836. doi:10.1038/Nature08479. 7267.

Nam JW, Bartel DP, 2012. Long noncoding rnas in c. elegans. *Genome Res*, **22**: 2529–40. ISSN 1549-5469 (Electronic) 1088-9051 (Linking). doi:10.1101/gr.140475.112. 12.

Negre N, Brown CD, Ma LJ, Bristow CA, Miller SW, Wagner U, Kheradpour P, Eaton ML, Loriaux P, Sealfon R, Li ZR, Ishii H, Spokony RF, Chen J, Hwang L, Cheng C, Auburn RP, Davis MB, Domanus M, Shah PK, Morrison CA, Zieba J, Suchy S, Senderowicz L, Victorsen A, Bild NA, Grundstad AJ, Hanley D, MacAlpine DM, Mannervik M, Venken K, Bellen H, White R, Gerstein M, Russell S, Grossman RL, Ren B, Posakony JW, Kellis M, White KP, 2011. A cis-regulatory map of the drosophila genome. *Nature*, **471**: 527–531. ISSN 0028-0836. doi:10.1038/Nature09990. 7339.

Neph S, Vierstra J, Stergachis AB, Reynolds AP, Haugen E, Vernot B, Thurman RE, John S, Sandstrom R, Johnson AK, Maurano MT, Humbert R, Rynes E, Wang H, Vong S, Lee K, Bates D, Diegel M, Roach V, Dunn D, Neri J, Schafer A, Hansen RS, Kutyavin T, Giste E, Weaver M, Canfield T, Sabo P, Zhang M, Balasundaram G, Byron R, MacCoss MJ, Akey JM, Bender MA, Groudine M, Kaul R, Stamatoyannopoulos JA, 2012. An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature*, **489**: 83–90. ISSN 1476-4687 (Electronic) 0028-0836 (Linking). doi: 10.1038/nature11212. 7414.

Ni T, Corcoran DL, Rach EA, Song S, Spana EP, Gao Y, Ohler U, Zhu J, 2010. A paired-end sequencing strategy to map the complex landscape of transcription initiation. *Nat Methods*. ISSN 1548-7105 (Electronic) 1548-7091 (Linking). doi:10.1038/nmeth.1464.

Nigumann P, Redik K, Matlik K, Speek M, 2002. Many human genes are transcribed from the antisense promoter of l1 retrotransposon. *Genomics*, **79**: 628–34. ISSN 0888-7543 (Print). doi:10.1006/geno.2002.6758. 5.

Nilsen TW, Graveley BR, 2010. Expansion of the eukaryotic proteome by alternative splicing. *Nature*, **463**: 457–63. ISSN 1476-4687 (Electronic) 0028-0836 (Linking). doi:10.1038/nature08909. 7280.

Ntini E, Jarvelin AI, Bornholdt J, Chen Y, Boyd M, Jorgensen M, Andersson R, Hoof I, Schein A, Andersen PR, Andersen PK, Preker P, Valen E, Zhao X, Pelechano V, Steinmetz LM, Sandelin A, Jensen TH, 2013. Polyadenylation site-induced decay of upstream transcripts enforces promoter directionality. *Nat Struct Mol Biol*. ISSN 1545-9985 (Electronic) 1545-9985 (Linking). doi:10.1038/nsmb.2640.

Nusslein-Volhard C, Wieschaus E, 1980. Mutations affecting segment number and polarity in drosophila. *Nature*, **287**: 795–801. ISSN 0028-0836 (Print) 0028-0836 (Linking). 5785.

Odom DT, Dowell RD, Jacobsen ES, Gordon W, Danford TW, MacIsaac KD, Rolfe PA, Conboy CM, Gifford DK, Fraenkel E, 2007. Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nature Genetics*, **39**: 730–732. ISSN 1061-4036. doi:10.1038/Ng2047. 6.

Oliveri P, Tu Q, Davidson EH, 2008. Global regulatory logic for specification of an embryonic cell lineage. *Proc Natl Acad Sci U S A*, **105**: 5955–62. ISSN 1091-6490 (Electronic) 0027-8424 (Linking). doi:10.1073/pnas.0711220105. 16.

Orgel LE, Crick FHC, 1980. Selfish dna - the ultimate parasite. *Nature*, **284**: 604–607. ISSN 0028-0836. 5757.

Palopoli MF, Patel NH, 1996. Neo-darwinian developmental evolution: can we bridge the gap between pattern and process? *Curr Opin Genet Dev*, **6**: 502–8. ISSN 0959-437X (Print) 0959-437X (Linking). 4.

Panne D, Maniatis T, Harrison SC, 2007. An atomic model of the interferon-beta enhanceosome. *Cell*, **129**: 1111–23. ISSN 0092-8674 (Print) 0092-8674 (Linking). doi:10.1016/j.cell.2007.05.019. 6.

Parkhurst SM, Corces VG, 1987. Developmental expression of drosophila melanogaster retrovirus-like transposable elements. *EMBO J*, **6**: 419–24. ISSN 0261-4189 (Print) 0261-4189 (Linking). 2.

Pasmant E, Sabbagh A, Vidaud M, Bieche I, 2011. Anril, a long, noncoding rna, is an unexpected major hotspot in gwas. *FASEB J*, **25**: 444–8. ISSN 1530-6860 (Electronic) 0892-6638 (Linking). doi:10.1096/fj.10-172452. 2.

Peaston AE, Evsikov AV, Graber JH, de Vries WN, Holbrook AE, Solter D, Knowles BB, 2004. Retrotransposons regulate host genes in mouse oocytes and preimplantation embryos. *Developmental Cell*, **7**: 597–606. ISSN 1534-5807. 4.

Pelisson A, 1981. The i-r system of hybrid dysgenesis in drosophila melanogaster: are i factor insertions responsible for the mutator effect of the i-r interaction? *Mol Gen Genet*, **183**: 123–9. ISSN 0026-8925 (Print) 0026-8925 (Linking). 1.

Penny GD, Kay GF, Sheardown SA, Rastan S, Brockdorff N, 1996. Requirement for xist in x chromosome inactivation. *Nature*, **379**: 131–7. ISSN 0028-0836 (Print) 0028-0836 (Linking). doi:10.1038/379131a0. 6561.

Petrov DA, Fiston-Lavier AS, Lipatov M, Lenkov K, Gonzalez J, 2011. Population genomics of transposable elements in drosophila melanogaster. *Molecular Biology and Evolution*, **28**: 1633–1644. ISSN 0737-4038. doi:10.1093/Molbev/Msq337. 5.

Petruk S, Sedkov Y, Riley KM, Hodgson J, Schweisguth F, Hirose S, Jaynes JB, Brock HW, Mazo A, 2006. Transcription of bxd noncoding rnas promoted by trithorax represses ubx in cis by transcriptional interference. *Cell*, **127**: 1209–21. ISSN 0092-8674 (Print) 0092-8674 (Linking). doi: 10.1016/j.cell.2006.10.039. 6.

Phillips-Cremins JE, Sauria ME, Sanyal A, Gerasimova TI, Lajoie BR, Bell JS, Ong CT, Hookway TA, Guo C, Sun Y, Bland MJ, Wagstaff W, Dalton S, McDevitt TC, Sen R, Dekker J, Taylor J, Corces VG, 2013. Architectural protein subclasses shape 3d organization of genomes during lineage commitment. *Cell*, **153**: 1281–95. ISSN 1097-4172 (Electronic) 0092-8674 (Linking). doi:10.1016/j.cell.2013.04.053. 6.

Phipson B, Smyth GK, 2010. Permutation p-values should never be zero: calculating exact p-values when permutations are randomly drawn. *Stat Appl Genet Mol Biol*, **9**: Article39. ISSN 1544-6115 (Electronic) 1544-6115 (Linking). doi:10.2202/1544-6115.1585. 1.

Plath K, Fang J, Mlynarczyk-Evans SK, Cao R, Worringer KA, Wang H, de la Cruz CC, Otte AP, Panning B, Zhang Y, 2003. Role of histone h3 lysine 27 methylation in x inactivation. *Science*, **300**: 131–5. ISSN 1095-9203 (Electronic) 0036-8075 (Linking). doi:10.1126/science.1084274. 5616.

Plessy C, Bertin N, Takahashi H, Simone R, Salimullah M, Lassmann T, Vitezic M, Severin J, Olivarius S, Lazarevic D, Hornig N, Orlando V, Bell I, Gao H, Dumais J, Kapranov P, Wang H, Davis CA, Gingeras TR, Kawai J, Daub CO, Hayashizaki Y, Gustincich S, Carninci P, 2010. Linking promoters to functional transcripts in small samples with nanocage and cagescan. *Nat Methods*. ISSN 1548-7105 (Electronic) 1548-7091 (Linking). doi:10.1038/nmeth.1470.

Poliseno L, Salmena L, Zhang J, Carver B, Haveman WJ, Pandolfi PP, 2010. A coding-independent function of gene and pseudogene mrnas regulates tumour biology. *Nature*, **465**: 1033–8. ISSN 1476-4687 (Electronic) 0028-0836 (Linking). doi:10.1038/nature09144. 7301.

Ponting CP, Belgard TG, 2010. Transcribed dark matter: meaning or myth? *Hum Mol Genet*. ISSN 1460-2083 (Electronic) 0964-6906 (Linking). doi:10.1093/hmg/ddq362.

Ponting CP, Oliver PL, Reik W, 2009. Evolution and functions of long noncoding rnas. *Cell*, **136**: 629–41. ISSN 1097-4172 (Electronic) 0092-8674 (Linking). doi:10.1016/j.cell.2009.02.006. 4.

Popadic A, Abzhanov A, Rusch D, Kaufman TC, 1998. Understanding the genetic basis of morphological evolution: the role of homeotic genes in the diversification of the arthropod bauplan. *Int J Dev Biol*, **42**: 453–61. ISSN 0214-6282 (Print) 0214-6282 (Linking). 3.

Prabhakar S, Noonan JP, Paabo S, Rubin EM, 2006. Accelerated evolution of conserved noncoding sequences in humans. *Science*, **314**: 786–786. ISSN 0036-8075. doi:10.1126/Science.1130738. 5800.

Prud'homme B, Gompel N, Rokas A, Kassner VA, Williams TM, Yeh SD, True JR, Carroll SB, 2006. Repeated morphological evolution through cis-regulatory changes in a pleiotropic gene. *Nature*, **440**: 1050–3. ISSN 1476-4687 (Electronic) 0028-0836 (Linking). doi:10.1038/nature04597. 7087.

Ptashne M, 2005. Regulation of transcription: from lambda to eukaryotes. *Trends Biochem Sci*, **30**: 275–9. ISSN 0968-0004 (Print) 0968-0004 (Linking). doi:10.1016/j.tibs.2005.04.003. 6.

Ptashne M, 2011. Principles of a switch. *Nat Chem Biol*, **7**: 484–7. ISSN 1552-4469 (Electronic) 1552-4450 (Linking). doi:10.1038/nchembio.611. 8.

Ptashne M, Backman K, Humayun MZ, Jeffrey A, Maurer R, Meyer B, Sauer RT, 1976. Autoregulation and function of a repressor in bacteriophage lambda. *Science*, **194**: 156–61. ISSN 0036-8075 (Print) 0036-8075 (Linking). 4261.

Rach EA, Yuan HY, Majoros WH, Tomancak P, Ohler U, 2009. Motif composition, conservation and condition-specificity of single and alternative transcription start sites in the drosophila genome. *Genome Biology*, **10**: –. ISSN 1474-760X. doi:10.1186/Gb-2009-10-7-R73. 7.

Radovic A, Wittkopp PJ, Long AD, Drapeau MD, 2002. Immunohistochemical colocalization of yellow and male-specific fruitless in drosophila melanogaster neuroblasts. *Biochem Biophys Res Commun*, **293**: 1262–4. ISSN 0006-291X (Print) 0006-291X (Linking). doi:10.1016/S0006-291X(02)00366-2. 4.

Raff EC, Raff RA, 2000. Dissociability, modularity, evolvability. *Evolution & Development*, **2**: 235–237. ISSN 1520-541X. 5.

Rastan S, 1983. Non-random x-chromosome inactivation in mouse x-autosome translocation embryos–location of the inactivation centre. *J Embryol Exp Morphol*, **78**: 1–22. ISSN 0022-0752 (Print) 0022-0752 (Linking).

Rifkin SA, Houle D, Kim J, White KP, 2005. A mutation accumulation assay reveals a broad capacity for rapid evolution of gene expression. *Nature*, **438**: 220–223. ISSN 0028-0836. doi:10.1038/Nature04114. 7065.

Rifkin SA, Kim J, White KP, 2003. Evolution of gene expression in the drosophila melanogaster subgroup. *Nature Genetics*, **33**: 138–144. ISSN 1061-4036. doi:10.1038/Ng1086. 2.

Rinn JL, Chang HY, 2012. Genome regulation by long noncoding rnas. *Annu Rev Biochem*, **81**: 145–66. ISSN 1545-4509 (Electronic) 0066-4154 (Linking). doi:10.1146/annurev-biochem-051410-092902.

Rinn JL, Kertesz M, Wang JK, Squazzo SL, Xu X, Brugmann SA, Goodnough LH, Helms JA, Farnham PJ, Segal E, Chang HY, 2007. Functional demarcation of active and silent chromatin domains in human hox loci by noncoding rnas. *Cell*, **129**: 1311–23. ISSN 0092-8674 (Print) 0092-8674 (Linking). doi:10.1016/j.cell.2007.05.022. 7.

Ronshaugen M, McGinnis N, McGinnis W, 2002. Hox protein mutation and macroevolution of the insect body plan. *Nature*, **415**: 914–7. ISSN 0028-0836 (Print) 0028-0836 (Linking). doi:10.1038/nature716. 6874.

Rouget C, Papin C, Boureux A, Meunier AC, Franco B, Robine N, Lai EC, Pelisson A, Simonelig M, 2010. Maternal mrna deadenylation and decay by the pirna pathway in the early drosophila embryo. *Nature*, **467**: 1128–32. ISSN 1476-4687 (Electronic) 0028-0836 (Linking). doi:10.1038/nature09465. 7319.

Rozhkov NV, Hammell M, Hannon GJ, 2013. Multiple roles for piwi in silencing drosophila transposons. *Genes Dev*, **27**: 400–12. ISSN 1549-5477 (Electronic) 0890-9369 (Linking). doi:10.1101/gad.209767.112. 4.

Sanyal A, Lajoie BR, Jain G, Dekker J, 2012. The long-range interaction landscape of gene promoters. *Nature*, **489**: 109–13. ISSN 1476-4687 (Electronic) 0028-0836 (Linking). doi:10.1038/nature11279. 7414.

Schlesinger F, Smith AD, Gingeras TR, Hannon GJ, Hodges E, 2013. De novo dna demethylation and noncoding transcription define active intergenic regulatory elements. *Genome Res*, **23**: 1601–14. ISSN 1549-5469 (Electronic) 1088-9051 (Linking). doi:10.1101/gr.157271.113. 10.

Schmidt D, Schwalie PC, Wilson MD, Ballester B, Goncalves A, Kutter C, Brown GD, Marshall A, Flicek P, Odom DT, 2012. Waves of retrotransposon expansion remodel genome organization and ctcf binding in multiple mammalian lineages. *Cell*, **148**: 335–48. ISSN 1097-4172 (Electronic) 0092-8674 (Linking). doi:10.1016/j.cell.2011.11.058. 1-2.

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA, Minx P, Reily AD, Courtney L, Kruchowski SS, Tomlinson C, Strong C, Delehaunty K, Fronick C, Courtney B, Rock SM, Belter E, Du F, Kim K, Abbott RM, Cotton M, Levy A, Marchetto P, Ochoa K, Jackson SM, Gillam B, Chen W, Yan L, Higginbotham J, Cardenas M, Waligorski J, Applebaum E, Phelps L, Falcone J, Kanchi K, Thane T, Scimone A, Thane N, Henke J, Wang T, Ruppert J, Shah N, Rotter K, Hodges J, Ingenthron E, Cordes M, Kohlberg S, Sgro J, Delgado B, Mead K, Chinwalla A, Leonard S, Crouse K, Collura K, Kudrna D, Currie J, He R, Angelova A, Rajasekar S, Mueller T, Lomeli R, Scara G, Ko A, Delaney K, Wissotski M, Lopez G, Campos D, Braidotti M, Ashley E, Golser W, Kim H, Lee S, Lin J, Dujmic Z, Kim W, Talag J, Zuccolo A, Fan C, Sebastian A, Kramer M, Spiegel L, Nascimento L, Zutavern T, Miller B, Ambroise C, Muller S, Spooner W, Narechania A, Ren L, Wei S, Kumari S, Faga B, Levy MJ, McMahan L, Van Buren P, Vaughn MW, Ying K, Yeh CT, Emrich SJ, Jia Y, Kalyanaraman A, Hsia AP, Barbazuk WB, Baucom RS, Brutnell TP, Carpita NC, Chaparro C, Chia JM, Deragon JM, Estill JC, Fu Y, Jeddeloh JA, Han Y, Lee H, Li P, Lisch DR, Liu S, Liu Z, Nagel DH, McCann MC, SanMiguel P, Myers AM, Nettleton D, Nguyen J, Penning BW, Ponnala L, Schneider KL, Schwartz DC, Sharma A, Soderlund C, Springer NM, Sun Q, Wang H, Waterman M, Westerman R, Wolfgruber TK, Yang L, Yu Y, Zhang L, Zhou S, Zhu Q, Bennetzen JL, Dawe RK, Jiang J, Jiang N, Presting GG, Wessler SR, Aluru S, Martienssen RA, Clifton SW, McCombie WR, Wing RA, Wilson RK, 2009. The b73 maize genome: complexity, diversity, and dynamics. *Science*, **326**: 1112–5. ISSN 1095-9203 (Electronic) 0036-8075 (Linking). doi:10.1126/science.1178534. 5956.

Segal E, Raveh-Sadka T, Schroeder M, Unnerstall U, Gaul U, 2008. Predicting expression patterns from regulatory sequence in drosophila segmentation. *Nature*, **451**: 535–40. ISSN 1476-4687 (Electronic) 0028-0836 (Linking). doi:10.1038/nature06496. 7178.

Seila AC, Calabrese JM, Levine SS, Yeo GW, Rahl PB, Flynn RA, Young RA, Sharp PA, 2008. Divergent transcription from active promoters. *Science*, **322**: 1849–51. ISSN 1095-9203 (Electronic) 0036-8075 (Linking). doi:10.1126/science.1162253. 5909.

Shapiro MD, Kronenberg Z, Li C, Domyan ET, Pan H, Campbell M, Tan H, Huff CD, Hu H, Vickrey AI, Nielsen SC, Stringham SA, Willerslev E, Gilbert MT, Yandell M, Zhang G, Wang J, 2013. Genomic diversity and evolution of the head crest in the rock pigeon. *Science*, **339**: 1063–7. ISSN 1095-9203 (Electronic) 0036-8075 (Linking). doi:10.1126/science.1230422. 6123.

Shapiro MD, Marks ME, Peichel CL, Blackman BK, Nereng KS, Jonsson B, Schluter D, Kingsley DM, 2004. Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. *Nature*, **428**: 717–23. ISSN 1476-4687 (Electronic) 0028-0836 (Linking). doi:10.1038/nature02415. 6984.

Sharon D, Tilgner H, Grubert F, Snyder M, 2013. A single-molecule long-read survey of the human transcriptome. *Nat Biotechnol*. ISSN 1546-1696 (Electronic) 1087-0156 (Linking). doi:10.1038/nbt.2705.

Shen Y, Yue F, McCleary DF, Ye Z, Edsall L, Kuan S, Wagner U, Dixon J, Lee L, Lobanenkov VV, Ren B, 2012. A map of the cis-regulatory sequences in the mouse genome. *Nature*, **488**: 116–20. ISSN 1476-4687 (Electronic) 0028-0836 (Linking). doi:10.1038/nature11243. 7409.

Sigova AA, Mullen AC, Molinie B, Gupta S, Orlando DA, Guenther MG, Almada AE, Lin C, Sharp PA, Giallourakis CC, Young RA, 2013. Divergent transcription of long noncoding rna/mrna gene pairs in embryonic stem cells. *Proc Natl Acad Sci U S A*, **110**: 2876–81. ISSN 1091-6490 (Electronic) 0027-8424 (Linking). doi:10.1073/pnas.1221904110. 8.

Silva JC, Shabalina SA, Harris DG, Spouge JL, Kondrashovi AS, 2003. Conserved fragments of transposable elements in intergenic regions: evidence for widespread recruitment of mir- and l2-derived sequences within the mouse and human genomes. *Genet Res*, **82**: 1–18. ISSN 0016-6723 (Linking). 1.

Small S, Blair A, Levine M, 1992. Regulation of even-skipped stripe 2 in the drosophila embryo. *EMBO J*, **11**: 4047–57. ISSN 0261-4189 (Print) 0261-4189 (Linking). 11.

Small S, Kraut R, Hoey T, Warrior R, Levine M, 1991. Transcriptional regulation of a pair-rule stripe in drosophila. *Genes Dev*, **5**: 827–39. ISSN 0890-9369 (Print) 0890-9369 (Linking). 5.

Smits G, Mungall AJ, Griffiths-Jones S, Smith P, Beury D, Matthews L, Rogers J, Pask AJ, Shaw G, VandeBerg JL, McCarrey JR, Renfree MB, Reik W, Dunham I, 2008. Conservation of the h19 noncoding rna and h19-igf2 imprinting mechanism in therians. *Nat Genet*, **40**: 971–6. ISSN 1546-1718 (Electronic) 1061-4036 (Linking). doi:10.1038/ng.168. 8.

Spitz F, Furlong EE, 2012. Transcription factors: from enhancer binding to developmental control. *Nat Rev Genet*, **13**: 613–26. ISSN 1471-0064 (Electronic) 1471-0056 (Linking). doi:10.1038/nrg3207. 9.

Stadler MB, Murr R, Burger L, Ivanek R, Lienert F, Scholer A, van Nimwegen E, Wirbelauer C, Oakeley EJ, Gaidatzis D, Tiwari VK, Schubeler D, 2011. Dna-binding factors shape the mouse methylome at distal regulatory regions. *Nature*, **480**: 490–5. ISSN 1476-4687 (Electronic) 0028-0836 (Linking). doi:10.1038/nature10716. 7378.

Stanojevic D, Hoey T, Levine M, 1989. Sequence-specific dna-binding activities of the gap proteins encoded by hunchback and kruppel in drosophila. *Nature*, **341**: 331–5. ISSN 0028-0836 (Print) 0028-0836 (Linking). doi:10.1038/341331a0. 6240.

Stanojevic D, Small S, Levine M, 1991. Regulation of a segmentation stripe by overlapping activators and repressors in the drosophila embryo. *Science*, **254**: 1385–7. ISSN 0036-8075 (Print) 0036-8075 (Linking). 5036.

Stathopoulos A, Levine M, 2005. Genomic regulatory networks and animal development. *Dev Cell*, **9**: 449–62. ISSN 1534-5807 (Print) 1534-5807 (Linking). doi:10.1016/j.devcel.2005.09.005. 4.

Staubach F, Teschke M, Voolstra CR, Wolf JB, Tautz D, 2010. A test of the neutral model of expression change in natural populations of house mouse subspecies. *Evolution*, **64**: 549–60. ISSN 1558-5646 (Electronic) 0014-3820 (Linking). doi:10.1111/j.1558-5646.2009.00818.x. 2.

Stefflova K, Thybert D, Wilson MD, Streeter I, Aleksic J, Karagianni P, Brazma A, Adams DJ, Talianidis I, Marioni JC, Flicek P, Odom DT, 2013. Cooperativity and rapid evolution of cobound transcription factors in closely related mammals. *Cell*, **154**: 530–40. ISSN 1097-4172 (Electronic) 0092-8674 (Linking). doi:10.1016/j.cell.2013.07.007. 3.

Steijger T, Abril JF, Engstrom PG, Kokocinski F, Akerman M, Alioto T, Ambrosini G, Antonarakis SE, Behr J, Bertone P, Bohnert R, Bucher P, Cloonan N, Derrien T, Djebali S, Du J, Dudoit S, Gerstein M, Gingeras TR, Gonzalez D, Grimmond SM, Guigo R, Habegger L, Harrow J, Hubbard TJ, Iseli C, Jean G, Kahles A, Lagarde J, Leng J, Lefebvre G, Lewis S, Mortazavi A, Niermann P, Ratsch G, Reymond A, Ribeca P, Richard H, Rougemont J, Rozowsky J, Sammeth M, Sboner A, Schulz MH, Searle SM, Solorzano ND, Solovyev V, Stanke M, Stevenson BJ, Stockinger H, Valsesia A, Weese D, White S, Wold BJ, Wu J, Wu TD, Zeller G, Zerbino D, Zhang MQ, 2013. Assessment of transcript reconstruction methods for rna-seq. *Nat Methods*, **10**: 1177–84. ISSN 1548-7105 (Electronic) 1548-7091 (Linking). doi:10.1038/nmeth.2714. 12.

Stern DL, 1998. A role of ultrabithorax in morphological differences between drosophila species. *Nature*, **396**: 463–6. ISSN 0028-0836 (Print) 0028-0836 (Linking). doi:10.1038/24863. 6710.

Stern DL, 2000. Evolutionary developmental biology and the problem of variation. *Evolution*, **54**: 1079–91. ISSN 0014-3820 (Print) 0014-3820 (Linking). 4.

Stern DL, 2013. The genetic causes of convergent evolution. *Nat Rev Genet*, **14**: 751–64. ISSN 1471-0064 (Electronic) 1471-0056 (Linking). doi:10.1038/nrg3483. 11.

Stern DL, Orgogozo V, 2009. Is genetic evolution predictable? *Science*, **323**: 746–751. ISSN 0036-8075. doi:10.1126/Science.1158997. 5915.

Struhl K, 1999. Fundamentally different logic of gene regulation in eukaryotes and prokaryotes. *Cell*, **98**: 1–4. ISSN 0092-8674 (Print) 0092-8674 (Linking). doi:10.1016/S0092-8674(00)80599-1. 1.

Struhl K, 2007. Transcriptional noise and the fidelity of initiation by rna polymerase ii. *Nat Struct Mol Biol*, **14**: 103–5. ISSN 1545-9993 (Print) 1545-9985 (Linking). doi:10.1038/nsmb0207-103. 2.

Sucena E, Delon I, Jones I, Payre F, Stern DL, 2003. Regulatory evolution of shavenbaby/ovo underlies multiple cases of morphological parallelism. *Nature*, **424**: 935–938. ISSN 0028-0836. doi:10.1038/Nature01768. 6951.

Sucena E, Stern DL, 2000. Divergence of larval morphology between drosophila sechellia and its sibling species caused by cis-regulatory evolution of ovo/shaven-baby. *Proc Natl Acad Sci U S A*, **97**: 4530–4. ISSN 0027-8424 (Print) 0027-8424 (Linking). 9.

Suzuki H, Forrest ARR, van Nimwegen E, Daub CO, Balwierz PJ, Irvine KM, Lassmann T, Ravasi T, Hasegawa Y, de Hoon MJL, Katayama S, Schroder K, Carninci P, Tomaru Y, Kanamori-Katayama M, Kubosaki A, Akalin A, Ando Y, Arner E, Asada M, Asahara H, Bailey T, Bajic VB, Bauer D, Beckhouse AG, Bertin N, Bjorkegren J, Brombacher F, Bulger E, Chalk AM, Chiba J, Cloonan N, Dawe A, Dostie J, Engstrom PG, Essack M, Faulkner GJ, Fink JL, Fredman D, Fujimori K, Furuno M, Gojobori T, Gough J, Grimmond SM, Gustafsson M, Hashimoto M, Hashimoto T, Hatakeyama M, Heinzel S, Hide W, Hofmann O, Hornquist M, Huminiecki L, Ikeo K, Imamoto N, Inoue S, Inoue Y, Ishihara R, Iwayanagi T, Jacobsen A, Kaur M, Kawaji H, Kerr MC, Kimura R, Kimura S, Kimura Y, Kitano H, Koga H, Kojima T, Kondo S, Konno T, Krogh A, Kruger A, Kumar A, Lenhard B, Lennartsson A, Lindow M, Lizio M, MacPherson C, Maeda N, Maher CA, Maqungo M, Mar J, Matigian NA, Matsuda H, Mattick JS, Meier S, Miyamoto S, Miyamoto-Sato E, Nakabayashi K, Nakachi Y, Nakano M, Nygaard S, Okayama T, Okazaki Y, Okuda-Yabukami H, Orlando V, Otomo J, Pachkov M, Petrovsky N, Plessy C, Quackenbush J, Radovanovic A, Rehli M, Saito R, Sandelin A, Schmeier S, Schonbach C, Schwartz AS, Semple CA, Sera M, Severin J, Shirahige K, Simons C, Laurent GS, Suzuki M, Suzuki T, Sweet MJ, Taft RJ, Takeda S, Takenaka Y, Tan K, Taylor MS, Teasdale RD,

Tegner J, Teichmann S, Valen E, Wahlestedt C, Waki K, Waterhouse A, AWells C, Winther O, Wu L, Yamaguchi K, Yanagawa H, Yasuda J, Zavolan M, Hume DA, Arakawa T, Fukuda S, Imamura K, Kai C, Kaiho A, Kawashima T, Kawazu C, Kitazume Y, Kojima M, Miura H, Murakami K, Murata M, Ninomiya N, Nishiyori H, Noma S, Ogawa C, Sano T, Simon C, Tagami M, Takahashi Y, Kawai J, Hayashizaki Y, Consortium F, Ctr ROS, 2009. The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line. *Nature Genetics*, **41**: 553–562. ISSN 1061-4036. doi:10.1038/Ng.375. 5.

Taft RJ, Glazov EA, Cloonan N, Simons C, Stephen S, Faulkner GJ, Lassmann T, Forrest ARR, Grimmond SM, Schroder K, Irvine K, Arakawa T, Nakamura M, Kubosaki A, Hayashida K, Kawazu C, Murata M, Nishiyori H, Fukuda S, Kawai J, Daub CO, Hume DA, Suzuki H, Orlando V, Carninci P, Hayashizaki Y, Mattick JS, 2009. Tiny rnas associated with transcription start sites in animals. *Nature Genetics*, **41**: 572–578. ISSN 1061-4036. doi:10.1038/Ng.312. 5.

Tan-Wong SM, Zaugg JB, Camblong J, Xu Z, Zhang DW, Mischo HE, Ansari AZ, Luscombe NM, Steinmetz LM, Proudfoot NJ, 2012. Gene loops enhance transcriptional directionality. *Science*, **338**: 671–5. ISSN 1095-9203 (Electronic) 0036-8075 (Linking). doi:10.1126/science.1224350. 6107.

Thanos D, Maniatis T, 1995. Virus induction of human ifn beta gene expression requires the assembly of an enhanceosome. *Cell*, **83**: 1091–100. ISSN 0092-8674 (Print) 0092-8674 (Linking). 7.

Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B, Garg K, John S, Sandstrom R, Bates D, Boatman L, Canfield TK, Diegel M, Dunn D, Ebersol AK, Frum T, Giste E, Johnson AK, Johnson EM, Kutyavin T, Lajoie B, Lee BK, Lee K, London D, Lotakis D, Neph S, Neri F, Nguyen ED, Qu H, Reynolds AP, Roach V, Safi A, Sanchez ME, Sanyal A, Shafer A, Simon JM, Song L, Vong S, Weaver M, Yan Y, Zhang Z, Lenhard B, Tewari M, Dorschner MO, Hansen RS, Navas PA, Stamatoyannopoulos G, Iyer VR, Lieb JD, Sunyaev SR, Akey JM, Sabo PJ, Kaul R, Furey TS, Dekker J, Crawford GE, Stamatoyannopoulos JA, 2012. The accessible chromatin landscape of the human genome. *Nature*, **489**: 75–82. ISSN 1476-4687 (Electronic) 0028-0836 (Linking). doi:10.1038/nature11232. 7414.

Tolhuis B, Palstra RJ, Splinter E, Grosveld F, de Laat W, 2002. Looping and interaction between hypersensitive sites in the active beta-globin locus. *Mol Cell*, **10**: 1453–65. ISSN 1097-2765 (Print) 1097-2765 (Linking). 6.

Tsankov A, Yanagisawa Y, Rhind N, Regev A, Rando OJ, 2011. Evolutionary divergence of intrinsic and trans-regulated nucleosome positioning sequences reveals plastic rules for chromatin organization. *Genome Res*, **21**: 1851–62. ISSN 1549-5469 (Electronic) 1088-9051 (Linking). doi:10.1101/gr.122267.111. 11.

Udomkit A, Forbes S, McLean C, Arkhipova I, Finnegan DJ, 1996. Control of expression of the i factor, a line-like transposable element in drosophila melanogaster. *Embo Journal*, **15**: 3174–3181. ISSN 0261-4189. 12.

Ulitsky I, Bartel DP, 2013. lincrnas: genomics, evolution, and mechanisms. *Cell*, **154**: 26–46. ISSN 1097-4172 (Electronic) 0092-8674 (Linking). doi:10.1016/j.cell.2013.06.020. 1.

Ulitsky I, Shkumatava A, Jan CH, Sive H, Bartel DP, 2011. Conserved function of lincrnas in vertebrate embryonic development despite rapid sequence evolution. *Cell*, **147**: 1537–50. ISSN 1097-4172 (Electronic) 0092-8674 (Linking). doi:10.1016/j.cell.2011.11.055. 7.

Valen E, Pascarella G, Chalk A, Maeda N, Kojima M, Kawazu C, Murata M, Nishiyori H, Lazarevic D, Motti D, Marstrand TT, Tang MHE, Zhao X, Krogh A, Winther O, Arakawa T, Kawai J, Wells C, Daub C, Harbers M, Hayashizaki Y, Gustincich S, Sandelin A, Carninci P, 2009. Genome-wide detection and analysis of hippocampus core promoters using deepcage. *Genome Research*, **19**: 255–265. ISSN 1088-9051. doi:10.1101/Gr.084541.108. 2.

van Bakel H, Nislow C, Blencowe BJ, Hughes TR, 2010. Most "dark matter" transcripts are associated with known genes. *PLoS Biol*, **8**: e1000371. ISSN 1545-7885 (Electronic) 1544-9173 (Linking). doi:10.1371/journal.pbio.1000371. 5.

van de Lagemaat LN, Landry JR, Mager DL, Medstrand P, 2003. Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. *Trends in Genetics*, **19**: 530–536. ISSN 0168-9525. 10.

van Dijk EL, Chen CL, d'Aubenton Carafa Y, Gourvennec S, Kwapisz M, Roche V, Bertrand C, Silvain M, Legoix-Ne P, Loeillet S, Nicolas A, Thermes C, Morillon A, 2011. Xuts are a class of xrn1-sensitive antisense regulatory non-coding rna in yeast. *Nature*, **475**: 114–7. ISSN 1476-4687 (Electronic) 0028-0836 (Linking). doi:10.1038/nature10118. 7354.

Venters BJ, Pugh BF, 2013. Genomic organization of human transcription initiation complexes. *Nature*, **502**: 53–8. ISSN 1476-4687 (Electronic) 0028-0836 (Linking). doi:10.1038/nature12535. 7469.

Wang KC, Yang YW, Liu B, Sanyal A, Corces-Zimmerman R, Chen Y, Lajoie BR, Protacio A, Flynn RA, Gupta RA, Wysocka J, Lei M, Dekker J, Helms JA, Chang HY, 2011. A long noncoding rna maintains active chromatin to coordinate homeotic gene expression. *Nature*. ISSN 1476-4687 (Electronic) 0028-0836 (Linking). doi:10.1038/nature09819.

Wang T, Zeng J, Lowe CB, Sellers RG, Salama SR, Yang M, Burgess SM, Brachmann RK, Haussler D, 2007. Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53. *Proceedings of the National Academy of Sciences of the United States of America*, **104**: 18613–18618. ISSN 0027-8424. doi:10.1073/Pnas.0703637104. 47.

Wang X, Arai S, Song X, Reichart D, Du K, Pascual G, Tempst P, Rosenfeld MG, Glass CK, Kurokawa R, 2008. Induced ncrnas allosterically modify rna-binding proteins in cis to inhibit transcription. *Nature*, **454**: 126–30. ISSN 1476-4687 (Electronic) 0028-0836 (Linking). doi:10.1038/nature06992. 7200.

Ward LD, Kellis M, 2012. Evidence of abundant purifying selection in humans for recently acquired regulatory functions. *Science*. ISSN 1095-9203 (Electronic) 0036-8075 (Linking). doi:10.1126/science.1225057.

Warren RW, Nagy L, Selegue J, Gates J, Carroll S, 1994. Evolution of homeotic gene regulation and function in flies and butterflies. *Nature*, **372**: 458–61. ISSN 0028-0836 (Print) 0028-0836 (Linking). doi:10.1038/372458a0. 6505.

Whyte WA, Orlando DA, Hnisz D, Abraham BJ, Lin CY, Kagey MH, Rahl PB, Lee TI, Young RA, 2013. Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell*, **153**: 307–19. ISSN 1097-4172 (Electronic) 0092-8674 (Linking). doi:10.1016/j.cell.2013.03.035. 2.

Wigler M, Levy D, Perucho M, 1981. The somatic replication of dna methylation. *Cell*, **24**: 33–40. ISSN 0092-8674 (Print) 0092-8674 (Linking). 1.

Willingham AT, Orth AP, Batalov S, Peters EC, Wen BG, Aza-Blanc P, Hogenesch JB, Schultz PG, 2005. A strategy for probing the function of noncoding rnas finds a repressor of nfat. *Science*, **309**: 1570–3. ISSN 1095-9203 (Electronic) 0036-8075 (Linking). doi:10.1126/science.1115901. 5740.

Wilusz JE, Sunwoo H, Spector DL, 2009. Long noncoding rnas: functional surprises from the rna world. *Genes Dev*, **23**: 1494–504. ISSN 1549-5477 (Electronic) 0890-9369 (Linking). doi:10.1101/gad.1800909. 13.

Wittkopp PJ, Haerum BK, Clark AG, 2004. Evolutionary changes in cis and trans gene regulation. *Nature*, **430**: 85–8. ISSN 1476-4687 (Electronic) 0028-0836 (Linking). doi:10.1038/nature02698. 6995.

Wittkopp PJ, Haerum BK, Clark AG, 2008. Regulatory changes underlying expression differences within and between drosophila species. *Nature Genetics*, **40**: 346–350. ISSN 1061-4036. doi:10.1038/Ng.77. 3.

Wittkopp PJ, Vaccaro K, Carroll SB, 2002. Evolution of yellow gene regulation and pigmentation in drosophila. *Curr Biol*, **12**: 1547–56. ISSN 0960-9822 (Print) 0960-9822 (Linking). 18.

Wu L, Candille SI, Choi Y, Xie D, Jiang L, Li-Pook-Than J, Tang H, Snyder M, 2013. Variation and genetic control of protein abundance in humans. *Nature*, **499**: 79–82. ISSN 1476-4687 (Electronic) 0028-0836 (Linking). doi:10.1038/nature12223. 7456.

Wutz A, Jaenisch R, 2000. A shift from reversible to irreversible x inactivation is triggered during es cell differentiation. *Mol Cell*, **5**: 695–705. ISSN 1097-2765 (Print) 1097-2765 (Linking). 4.

Wutz A, Rasmussen TP, Jaenisch R, 2002. Chromosomal silencing and localization are mediated by different domains of xist rna. *Nat Genet*, **30**: 167–74. ISSN 1061-4036 (Print) 1061-4036 (Linking). doi:10.1038/ng820. 2.

Xu N, Papagiannakopoulos T, Pan GJ, Thomson JA, Kosik KS, 2009. Microrna-145 regulates oct4, sox2, and klf4 and represses pluripotency in human embryonic stem cells. *Cell*, **137**: 647–658. ISSN 0092-8674. doi:10.1016/J.Cell.2009.02.038. 4.

Yosef N, Shalek AK, Gaublomme JT, Jin H, Lee Y, Awasthi A, Wu C, Karwacz K, Xiao S, Jorgolli M, Gennert D, Satija R, Shakya A, Lu DY, Trombetta JJ, Pillai MR, Ratcliffe PJ, Coleman ML, Bix M, Tantin D, Park H, Kuchroo VK, Regev A, 2013. Dynamic regulatory network controlling th17 cell differentiation. *Nature*, **496**: 461–8. ISSN 1476-4687 (Electronic) 0028-0836 (Linking). doi:10.1038/nature11981. 7446.

Young RA, 2011. Control of the embryonic stem cell state. *Cell*, **144**: 940–54. ISSN 1097-4172 (Electronic) 0092-8674 (Linking). doi:10.1016/j.cell.2011.01.032. 6.

Young RS, Marques AC, Tibbit C, Haerty W, Bassett AR, Liu JL, Ponting CP, 2012. Identification and properties of 1,119 candidate lincrna loci in the drosophila melanogaster genome. *Genome Biol Evol*, **4**: 427–42. ISSN 1759-6653 (Electronic). doi:10.1093/gbe/evs020. 4.

Zaret KS, Carroll JS, 2011. Pioneer transcription factors: establishing competence for gene expression. *Genes Dev*, **25**: 2227–41. ISSN 1549-5477 (Electronic) 0890-9369 (Linking). doi:10.1101/gad.176826.111. 21.

Zhao H, Zhang X, Frazao JB, Condino-Neto A, Newburger PE, 2013. Hox antisense lincrna hoxa-as2 is an apoptosis repressor in all trans retinoic acid treated nb4 promyelocytic leukemia cells. *J Cell Biochem*, **114**: 2375–83. ISSN 1097-4644 (Electronic) 0730-2312 (Linking). doi:10.1002/jcb.24586. 10.

Zhao J, Ohsumi TK, Kung JT, Ogawa Y, Grau DJ, Sarma K, Song JJ, Kingston RE, Borowsky M, Lee JT, 2010. Genome-wide identification of polycomb-associated rnas by rip-seq. *Molecular Cell*, **40**: 939–953. ISSN 1097-2765. doi:10.1016/J.Molcel.2010.12.011. 6.

Zhao J, Sun BK, Erwin JA, Song JJ, Lee JT, 2008. Polycomb proteins targeted by a short repeat rna to the mouse x chromosome. *Science*, **322**: 750–756. ISSN 0036-8075. doi:10.1126/Science.1163045. 5902.

Zhenodarova SM, Kliagina VP, Maistrenko FG, Pustoshilova NM, Smolianinova OA, 1989. Substrate specificity of t4 rna-ligase. the effect of the nucleotide composition of substrates and the size of phosphate donor on the effectiveness of intermolecular ligation. *Bioorg Khim*, **15**: 478–83. ISSN 0132-3423 (Print) 0132-3423 (Linking). 4.

List of Figures