Evolution and functional relevance of non-coding elements in vertebrate genomes

Anamaria Necsulea

Université Claude Bernard - Lyon 1
Laboratoire de Biométrie et Biologie Évolutive

Évolution et fonctionnalité des éléments non-codants dans les génomes des vertébrés

Evolution and functional relevance of non-coding elements in vertebrate genomes

Anamaria NECSULEA

Habilitation à diriger les recherches

Soutenue publiquement le 31/01/2024

Jury:

Sarah DJEBALI	INSERM, IRSD, Toulouse	Rapportrice
Nicolas GALTIER	CNRS, ISEM, Montpellier	Examinateur
Vincent LACROIX	UCBL, LBBE, Lyon	Examinateur
Hugues ROEST CROLLIUS	CNRS, IBENS, Paris	Rapporteur
Claire ROUGEULLE	CNRS, Epigénétique et destin cellulaire, Paris	Examinatrice
Marie SÉMON	ENS, LBMC, Lyon	Rapportrice

Contents

Remerciements

Il me semble que l'écriture d'un chapitre de remerciements est toujours difficile. Dans mon cas, cela est dû au grand nombre de personnes que j'aimerais remercier comme il se doit, mais aussi à la peur de ne pas trouver les mots justes pour exprimer toute ma reconnaissance. Tout au long de ma carrière de recherche, j'ai eu la chance d'être entourée par des personnes extraordinaires, qui m'ont fourni d'excellents modèles à suivre et qui m'ont soutenue lors des étapes difficiles. Ici je fais le choix de les présenter dans un ordre chronologique approximatif, sachant que lorsque des ambiguïtés temporelles se sont présentées, toutes les dates pertinentes ont été retenues (oui, une description précise du matériel et méthodes est importante même dans le chapitre de remerciements).

Je remercie donc les enseignant.e.s et les chercheur.e.s qui m'ont fait découvrir la bioinformatique, lors de mes études à l'INSA de Lyon. Parmi les nombreux noms qui me viennent à l'esprit, je dois citer Guillaume Beslon, Hubert Charles, Laurent Duret, Christian Gautier, Carole Knibbe, Jean Lobry et Marie-France Sagot. J'ai découvert la bioinformatique lorsque nous étions jeunes toutes les deux. C'est grâce à ces enseignant.e.s et chercheur.e.s que j'ai acquis une grande affection pour ce domaine de recherche, que je garde encore aujourd'hui.

Je remercie les collègues qui m'ont entourée et encadrée de près ou de loin lors de ma thèse au LBBE. A commencer bien sûr par Jean Lobry, mon directeur de thèse, qui m'a accordé une confiance demesurée et qui m'a soutenue lors de mes nombreux écarts par rapport au chemin établi. Je remercie également Laurent Duret, avec qui j'ai eu la chance de travailler lors de ma dernière année de thèse et pendant un peu de temps après. Depuis ce moment, Laurent n'a cessé de m'impressionner par son immense culture, sa rigueur scientifique, sa brillance et sa modestie, le tout couronné par un don pour les jeux de mots que j'essaye vainement d'imiter (tout comme le reste de ses qualités). C'est également lors de ma thèse que j'ai rencontré Bastien Boussau, qui est rapidement devenu un modèle pour moi, en plus d'un collaborateur et ami, et à qui je dois énormément. Travailler avec Bastien reste toujours très stimulant et je me réjouis de pouvoir le rejoindre dans notre future équipe de recherche.

Je remercie les collègues et collaborateur.rice.s suisses, que j'ai rencontré.e.s lors de mon séjour post-doctoral à Lausanne : Henrik Kaessmann et Denis Duboule, qui m'ont accueillie dans leurs laboratoires respectifs et qui m'on fait découvrir de nouvelles façons de faire de la recherche ; Fabrice Darbellay et Rita Amândio, qui m'ont appris de nouvelles approches pour l'étude des longs ARNs non-codants ; Markus Heim, qui m'a permis de faire de la recherche un peu plus appliquée que d'habitude. Je remercie Magali Soumillon, une super-héroïne qui peut faire à la fois de la bioinformatique de haut niveau, des expériences complexes en biologie moléculaire et du pilotage automobile extrême. Je remercie Maria Warnefors d'avoir rendu le coin "Ana-Maria" du bureau très agréable à vivre ; Diego Cortéz et Katja Guschanski pour tous les fous rires ; Philippe Julien, Iris Finci, Francesco Carelli, Margarida Cardoso-Moreira,

Zhongyi Wang et Angélica Liechti pour les bons moments passés ensemble.

Je remercie tou.te.s les étudiant.e.s avec qui j'ai eu l'opportunité de travailler ces dernières années : Iris Finci et Adem Bilican, les premier.e.s doctorant.e.s que j'ai eu la chance d'encadrer avec Henrik Kaessmann ; Oliver Selmoni, Athimed El-Taher, Florian Bénitière, Hoedric Huguet, Timothée Kastylevsky, Thomas Lahaie, Basile Sugranes, Estelle Champion, Hugo Seytier, Victor Lefebvre, Adam Boussif et Florian Blanchard, que j'ai (co-)encadré pour leurs stages de Licence ou Master à Lausanne et à Lyon. Je remercie tout particulièrement Alexandre Laverré, le premier doctorant que j'ai eu la chance d'encadrer au LBBE. Ce fut un vrai plaisir de travailler avec Alex, qui est toujours souriant, motivé et curieux.

Je remercie tous les collègues des pôles administratif, informatique et biotechnologique, sans lesquels notre laboratoire ne pourrait pas fonctionner. En particulier, je remercie Stéphane Delmotte, Lionel Humblot, Simon Penel, Bruno Spataro et Nathalie Arbasetti, qui ont passé une bonne partie de leur temps à résoudre mes problèmes informatiques et administratifs, sans perdre leur patience et leur bonne humeur.

Je remercie les collègues qui m'ont donné l'opportunité de participer aux enseignements à l'Université Claude Bernard - Lyon 1, en particulier Vincent Lacroix, qui est en grande partie moteur de la réussite du Master Bioinformatique.

Je remercie les membres du jury d'avoir pris le temps de lire ce manuscrit : Sarah Djebali, Nicolas Galtier, Vincent Lacroix, Hugues Roest Crollius, Claire Rougeulle et Marie Sémon. Ce fut un honneur de bénéficier de leur expertise dans les domaines de l'évolution moléculaire, de la génomique comparative et de la génomique fonctionnelle.

Je remercie les amis que j'ai connus grâce au laboratoire et à qui je dois énormément en dehors du laboratoire: Laurent Duret et Céline Blasco, qui sont à eux deux entièrement responsables de ma culture cinématographique et de ma forme physique, Simon Penel, avec qui je partage une fascination inexplicable pour Horror Express et Dementia 13, Bastien Boussau et Mathilde Paris, les amis de 40 ans (et ce n'est pas fini). Enfin, je remercie Philippe Veber, mon plus cher collègue de bureau.

Chapter 1

Introduction

In this manuscript, I will present a summary of the research that I conducted after obtaining my PhD in 2008, as well as the research projects that I am planning to pursue during the coming years. When I was a PhD student, the publication of the human genome sequence was still fresh [1], but several hundreds of genome sequences were already available for prokaryotes, as well as for a few eukaryotes. Several approaches that could quantify gene at the genome-wide level were already common practice, for example microarrays [2] or serial analyses of gene expression (SAGE) [3]. It was already possible to evaluate the positions at which proteins bind along the genome using the ChIP-chip approach, which combined chromatin immunoprecipitation with whole-genome DNA microarrays [4]. This approach opened the way for the study of gene expression regulatory mechanisms, at the genome-wide level. It was also already possible to investigate the three-dimensional conformation of the chromatin, thanks to the chromosome conformation capture (3C) approach [5]. All of these techniques, along with many others, were thus already setting the stage for the functional genomics revolution that started a few years afterwards, with the increasing availability and affordability of next generation sequencing (NGS) techniques. Coupled with NGS, these techniques became truly applicable genome-wide, their sensitivity to detect unfrequent events increased considerably, and their output became digital and thus easier to model statistically [6].

During my PhD, I studied the evolutionary processes that drive local strand asymmetries in nucleotide composition, related to the genomic organization of DNA replication and transcription units. In practice, my research projects involved analyzing nucleotide composition characteristics, such as the GC-content and the GC-skew, as well as synonymous codon frequencies. Because my work was limited to analyzing simple features of genome sequences, I was somewhat envious of the more complex data types that were rapidly accumulating at the time thanks to innovations in molecular biology technologies. At my PhD defense, one of the members of the jury asked an unusual question: if a fairy (specialized in molecular biology and evolution matters) would grant me any wish, what would I ask for? Although - presumably - it would have been in the fairy's powers to give me the answer to any question I was interested in, I decided I would ask for unlimited data, rather than for direct answers to biological questions. One of the best feelings when doing research is understanding something, getting the answer to a question that one is interested in. For me, the process of obtaining the answers, designing and performing the actual analyses that bring together all the pieces of the puzzle, is part of the enjoyment. I would not want to get the answers on a silver platter, without getting a chance to play with the data. Although fairies do not exist, I consider my-

self lucky to be a researcher during a period marked by outstanding technological innovations in molecular biology and genetics, that give us access to previously unimaginable types and quantities of data.

As a post-doctoral researcher, my wish to study more complex molecular data was granted. My host laboratory generated an extensive collection of transcriptome sequencing (RNA-seq) data, for several amniote species and major organs. We used this data collection to analyze the evolution of protein-coding gene expression patterns [7]. We also used the RNA-seq data to identify long non-coding RNAs (abbreviated lncRNAs and simply defined as long transcripts that do not encode proteins) in each of the species and to compare their repertoires, their sequences and their expression patterns across species [8]. Studying lncRNAs made me face one of the most important challenges in the "big data" era in molecular biology: making sense of the data. Around the same time, the ENCODE consortium had published its first genome-wide analyses of biochemical activities or characteristics, including transcription, transcription factor binding and histone modifications [9]. The ENCODE consortium concluded that the 80% of the human genome had (biochemical) functions. This estimate was in striking contrast with the proportion of the human genome that is thought to be subject to purifying selection, estimated at ∼5.5% through comparisons among mammalian species [10] or at ∼9% through a comparison among human genomes [11]. The conflict between the notions of biochemical activity and selected biological function became overt [12].

More than 10 years afterwards, the debate is unfortunately not over, and confusions between biochemical activities and biological functions are still very frequent. This is particularly true when studying lncRNAs. Such transcripts are now readily detected with sensitive RNA sequencing approaches. It is not yet clear how many lncRNA *loci* are present in the human genome. Depending on the annotation database, the number of human lncRNA *loci* varies between ∼18,000 (in RefSeq) and ∼95,000 (in NONCODE or LncBook), as of late 2022 [13]. This number is either comparable with or much higher than the estimated number of protein-coding genes in the human genome, which revolves around ∼20,000 [13]. However, while protein-coding genes have been studied extensively since the beginnings of molecular biology and genetics, much less is known about lncRNAs. At present, biological functions and functionality remain strictly hypothetical for the great majority of the tens of thousands of lncRNAs that were identified recently with transcriptome sequencing data. In this context, studying the evolution of lncRNAs is essential. Evidently, the degree of evolutionary conservation of lncRNAs (or of other genomic elements), does not provide a perfect assessment of their functionality. Purifying selection can be difficult to detect, especially for elements that have recently acquired their biological functions. Conversely, evidence of purifying selection at a non-coding genomic locus often does not suffice to pinpoint the exact characteristic of the locus that is under selection (see also chapter 2). Nevertheless, even if (evidence of) evolutionary conservation is not perfectly synonymous with (evidence of) functionality, identifying those lncRNAs that are subject to purifying selection can help decide what lncRNAs warrant a detailed investigation. Given the large number of lncRNAs, which cannot all be thoroughly investigated experimentally, this type of filter is urgently needed. With this motivation, one of my main objectives when studying the evolution of lncRNAs and other non-coding DNA elements is to understand their functionality. I will present some of my research on lncRNA evolution and functionality in chapter 2.

While the main focus of my first post-doctoral fellowship was lncRNA evolution, quickly afterwards I became more interested in the evolution of gene expression regulation mech-

anisms, in particular of chromatin contacts involving gene promoters and distal regulatory mechanisms. I was exposed to this exciting topic during my stay in the Laboratory of Developmental Genomics, at the EPFL, Lausanne, whose main research objective was to understand the regulation of the *HoxD* genes in the context of mouse development. My colleagues had been among the first to use chromosome conformation capture techniques to identify the regulatory elements that control *HoxD* expression in different contexts [14]. They often investigated the chromatin contacts between *HoxD* genes and *cis*-acting regulatory elements with the 4C-seq technique, a derivative of the original chromosome conformation capture approach, which identifies interactions involving a pre-defined target genomic region [15]. A similar approach was more recently proposed to identify chromatin interactions between a pre-defined set of target genomic regions (such as gene promoters) and the entire genome [16]. Thanks to this type of approach, regulatory interactions between gene promoters and *cis*-regulatory elements can be detected with high sensitivity, at the genome-wide level. However, as is the case with all high-sensitivity functional genomics methods, the question of the biological relevance of the detected chromatin interactions also arises. Not all chromatin interactions detected with this type of approach are expected to be biologically relevant, and not all interactions between promoters and other genomic segments are expected to have regulatory roles. Here again, studying the evolution of chromatin interaction landscapes is one possible approach towards better understanding their functional relevance. During my first years as a CNRS researcher, I initiated a comparative analysis of promoter-centered chromatin contacts between human and mouse, which became the research topic of Alexandre Laverré's PhD. I will present the main results of Alexandre's work in chapter 4.

In chapter 5, I will present a recently started research project, in which I aim to investigate the relationship between genome rearrangements and the evolution of *cis*-regulatory landscapes. In particular, this project aims to determine whether genome rearrangements that drastically alter *cis*-regulatory landscapes might in some cases lead to gene pseudogenization. Although I have not divided this manuscript into "past research activities" and "research projects", the chapters are presented in approximate chronological order of my research interests. The results presented in chapter 2 are nearly all finalized and published, while the project presented in chapter 5 is very preliminary.

Chapter 2

Evolution and functionality of long non-coding RNAs

In this chapter, I will briefly discuss the functionality of long non-coding RNAs and how it can be assessed with evolutionary approaches. Over the past few years, several reviews have discussed the evidence for lncRNA functionality [17, 18], the appropriate methodology towards testing lncRNA functions [19, 20], and the biological functions that have been attributed to lncRNAs so far [21]. The topic is very much under debate, both in the field of evolutionary biology and in the field of molecular biology and genetics.

2.1 Are long non-coding RNAs functional or transcriptional noise?

Before discussing lncRNA functionality, we need to set a clear definition of "function", which is a difficult question. An intuitive definition of function for a genomic element could be given through the notion of usefulness or even necessity for the organism that carries it. This means that the presence of the genomic element is beneficial for the organism and that alterations or losses of the element would be deleterious for the organism. In this case, the evolution of the genomic element should be governed by natural selection, in particular purifying selection against deleterious mutations, but also potentially positive selection in favor of new, advantageous mutations (if the element can be further optimized, or if a new optimum emerges following a change in conditions). The function of the element would then be directly determined by natural selection; not all features of the element need to be under selection. This definition corresponds to the notion of *selected function* [12]. Throughout the text, I use the term "functionality" to indicate the presence of a biological function for a genomic element, even if the exact function is unknown.

As others have unambiguously stated before [12], evolutionary biologists cannot agree with the affirmation made by the original ENCODE project [9], according to which biochemical activity implies biological function. If we applied this definition, all lncRNAs detected with RNA sequencing would be functional, simply because they are transcribed. An indication that this cannot be the case, at least for the human genome, is the fact that erroneous transcripts (for example mRNA transcripts containing premature stop codons, which will be eventually be degraded by the nonsense-mediated decay machinery [22]) can be detected

in high-throughput transcriptome sequencing data [23]. Cellular machineries, including the RNA polymerases and the spliceosome, are not error-free. The error rate is believed to vary among *loci*, depending on the fitness cost of errors. For example, highly expressed genes, for which errors are more costly, have lower alternative splicing rates than weakly expressed genes [24]. Error rates are also expected to vary among species, depending on efficiency of natural selection or the effective population size N_e. For example, fewer splicing variants are detected in species with large N_e such as *Drosophila*, than in species with small N_e such as human [25]. This observation is in agreement with the "drift barrier" hypothesis proposed by Michael Lynch [26]. In this context, it is noteworthy that fewer lncRNAs were detected for species with large N_e than for species with small N_e [17, 27–29]. If lncRNA production is not too costly in terms of energy ressources (and given that lncRNAs are generally very weakly expressed [17] and presumably not translated, energy costs should be limited), then they should persist in species with small N_e even if they are not beneficial (or even if they are slightly deleterious) for the organism.

The question in the title is clearly overly simplistic. Some lncRNAs are not only functional but also essential for the organisms that carry them. However, it is unlikely that the numerous lncRNAs that are currently detected with high-throughput sequencing approaches are all functional (at least in species with small N_e, such as human and other vertebrates). The answer depends on the specific lncRNA that one is interested in. If we agree that lncRNAs cannot all be functional, and cannot all be transcriptional noise, then the question of how to quantify and distinguish the two classes remains open. Investigating the evolutionary characteristics of lncRNAs can provide a useful - though by no means perfect - approach towards understanding which lncRNAs are more likely to be functional. I will present below a brief summary of what is known about the evolution of lncRNAs, including my past contributions to this topic.

2.2 Evolution of long non-coding RNAs

2.2.1 Evolution of lncRNA sequences

The first evolutionary analyses of lncRNA sequences in mammals were performed before the NGS revolution, using the transcript repertoires obtained with full length cDNA sequencing by the FANTOM consortium [30]. It was shown that the sequences of $\sim$ 3,000 mouse "macro-RNAs" were subject to detectable purifying selection pressures, in particular on promoter regions and splice signals [30, 31]. The density of functional regions within mouse macro-RNAs was estimated to be around 5% [31]. A few years later, analyses of an extensive collection of transcriptome sequencing data showed that three quarters of the human genome is transcribed [32]. Analyses of lncRNAs defined with RNA-seq and other types of transcriptome assays by the GENCODE project [33] showed that their exonic sequences were much less conserved than protein-coding gene exons, but significantly more conserved than ancestral repeats [28]. These studies evaluated long-term evolutionary sequence conservation, using estimations of selective constraint derived from comparisons of mammalian genome alignments. The weak selective constraint observed on lncRNA sequences could thus in principle be explained by the presence of recently acquired, lineage-specific functions. However, analyses of within-species single nucleotide polymorphism data also found little evidence of selective constraint on human lncRNA sequences, contrary to *Drosophila* lncRNAs [29].

2.2.2 Evolution of lncRNA repertoires and expression patterns

The first evolutionary analysis of lncRNA repertoires and transcription patterns was performed on two closely related mouse subspecies and on rat [34]. In this study, only 60% of lncRNA loci showed conserved transcription among rodents, compared to 90% for protein-coding genes. Moreover, only those lncRNAs with evidence of transcriptional conservation had sequences that were more conserved during evolution than neighboring neutrally-evolving sequences [34]. The rapid evolution of lncRNA repertoires was later confirmed at broader evolutionary scales [8, 35, 36]. During my post-doctoral work, I annotated lncRNAs in 10 vertebrate species, using RNA-seq data from 8 organs [8]. I showed that only about 10% of human lncRNAs were conserved in mouse, meaning that transcribed homologous sequences could be identified. More than 10,000 lncRNAs could be detected in each species. Most of these lncRNAs were predicted to be lineage-specific, and only 400 lncRNAs were predicted to have originated in the ancestor of all amniote species included in the study. Thus, the sequences and the expression of long non-coding RNAs evolved much more rapidly than those of protein-coding genes.

The expression patterns of lncRNAs also differed considerably from those of protein-coding genes. LncRNAs had narrower expression patterns than protein-coding genes and were predominantly expressed in the testes [8]. We analyzed the transcriptomes in isolated cell populations during mouse spermatogenesis and found that lncRNAs were specifically expressed in the germline, in spermatocytes (meiotic cells) and early-stage spermatids (post-meiotic, haploid cells). The widespread transcription observed in these cell types was accompanied by an overall open chromatin state. We hypothesized that the extensive transcription of lncRNAs and other non-coding regions (including pseudogenes and transposable elements) was favored by the extensive chromatin remodeling that is known to take place during the late stages of spermatogenesis [37].

2.2.3 Conserved lncRNAs in embryonic development

Interestingly, we found that those lncRNAs that were highly conserved showed evidence for a potential involvement in developmental processes, although they were mainly detected using RNA-seq data from adult organs (with the exception of the placenta, we had sampled only adult organs). Specifically, the promoters of the highly conserved lncRNAs were enriched in conserved binding sites for developmental transcription factors, including homeobox transcription factors [8]. This motivated me to continue my work on lncRNA evolution by adding a temporal dimension in the comparative analysis. I joined the Laboratory of Developmental Genomics at the EPFL and collaborated with Fabrice Darbellay, a PhD student in the laboratory, to generate a collection of transcriptome sequencing data for mouse and rat, for four organs (brain, kidney, liver and testis) and five developmental stages (two embryonic stages, newborns, young and aged adult individuals). For deeper evolutionary comparisons, we also generated transcriptome sequencing data for chicken, for the two embryonic stages.

This analysis confirmed that overall, lncRNAs evolve much more rapidly than protein-coding genes, even for lncRNAs that were detected in embryonic organs [38]. However, there was indeed an association between expression in early development and evolutionary conservation. Long non-coding RNAs that were conserved between rodents and chicken were twice more likely to be predominantly expressed in embryonic stages, compared to rodent-specific or species-specific lncRNAs [38].

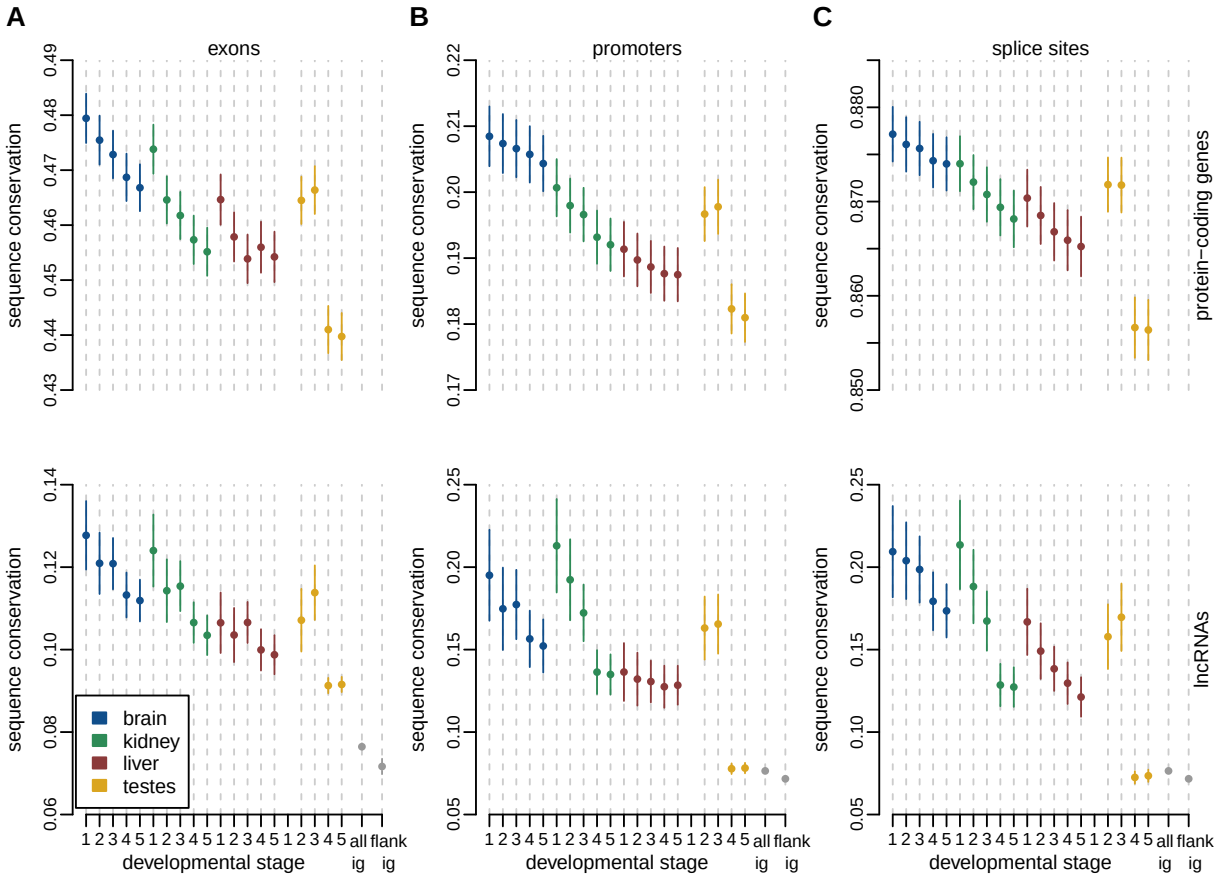

Figure 2.1: Long-term sequence conservation (measured with the PhastCons score for placental mammals) for lncRNAs and protein-coding gene regions. Genes are divided into subsets that are expressed above noise levels (TPM>1) in combinations of organs and developmental stages. Developmental stages are numbered from 1 to 5 and include two embryonic stages, newborns, young and aged adult individuals. Top: protein-coding genes. Bottom: lncRNAs. Three categories of regions were analyzed: exons, promoters and splice sites. Gray dots represent all intergenic regions and flanking intergenic regions of lncRNAs. Figure from Darbellay & Necsulea [38].

2.2.4 Conservation of lncRNA promoters and splice signals

In this study, we also re-evaluated the extent of long-term sequence conservation on different regions of lncRNAs (Figure 2.1). We confirmed that promoters and splice sites were the most conserved regions of lncRNAs, almost twice more conserved than their exonic regions [38]. This result confirmed previous observations made for fruitfly and human lncRNAs [29]. The high levels of sequence conservation observed for protein-coding gene splice sites is not surprising given that these regions are crucial for correct transcript processing. The fact that lncRNA splice sites also show increased sequence conservation levels compared to flanking intergenic regions (Figure 2.1) suggests that correct transcript processing is also important at least for some lncRNAs. The fact that lncRNA promoters generally have higher levels of sequence conservation than exons could be explained by the fact that a large proportion of lncRNA promoters are also thought to act as regulatory elements for other genes [39], thus accumulating another (putative) biological function (see also section 2.3 below).

Notably, the sequence conservation pattern described above depended on the organ and developmental stage where the long non-coding RNAs were expressed. The excess of sequence conservation on lncRNA splice sites and promoters compared to exons was only observed for those lncRNAs that were expressed in somatic organs or in early developmental stages. For those lncRNAs that were expressed in adult testes (*i.e.*, the great majority of all the lncRNAs that we could study), exonic sequence conservation levels were generally higher than the conservation levels of their promoters and splice sites, which were similar to the neighboring intergenic regions (Figure 2.1).

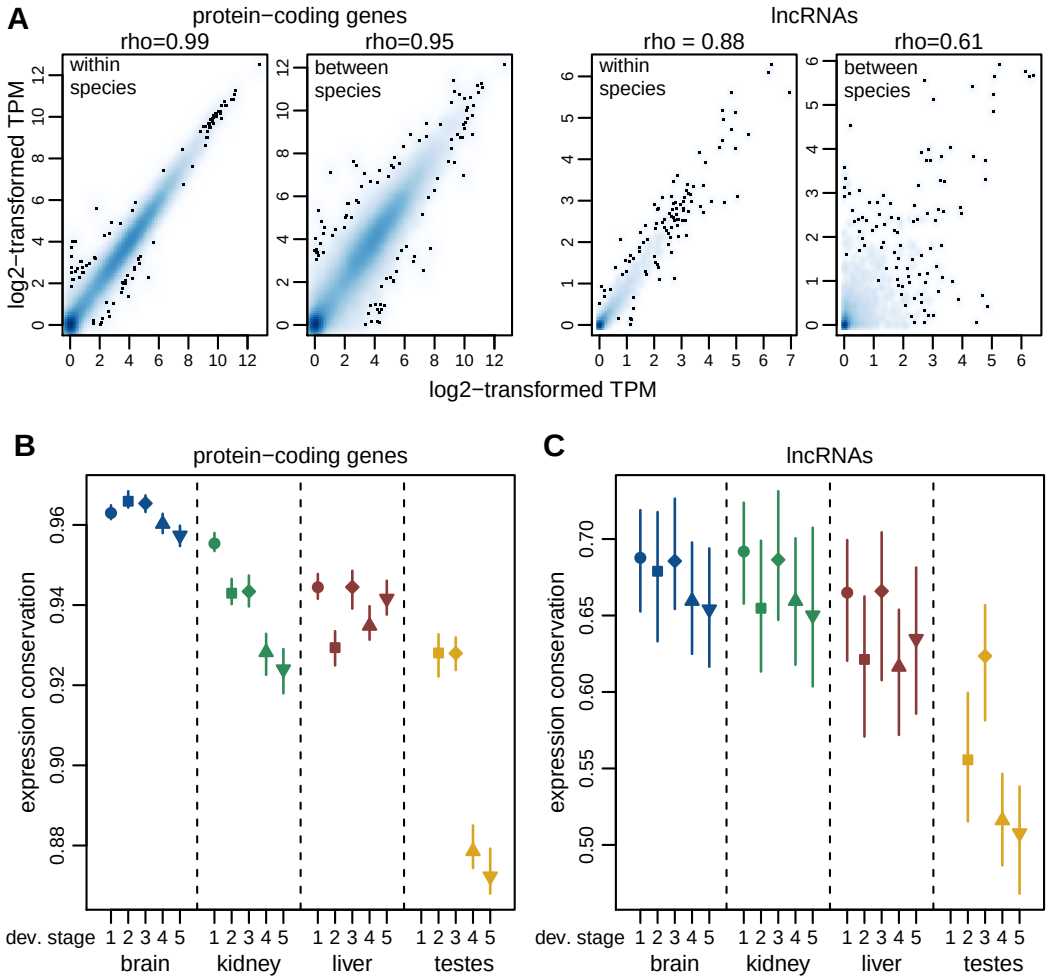

Figure 2.2: Expression conservation between mouse and rat, for lncRNAs and protein-coding genes, in various organs and developmental stages. Panel A: scatter plots of expression levels (log2-transformed TPM values) between individuals from a same species and between species, for protein-coding genes and lncRNAs. Panel B: expression conservation index for protein-coding genes, for the 4 organs and 5 developmental stages. Panel C: same as B, for lncRNAs. Vertical segments represent 95% confidence intervals constructed with 100 bootstrap replicates. Figure from Darbellay & Necsulea [38].

2.2.5 Rapid evolution of lncRNA expression levels

The slight excess of sequence conservation on lncRNA promoters might also indicate that the biological function carried by some of these loci resides in the act of transcription, rather than in the RNA molecule itself, as was previously shown for *e.g.* the *Airn* lncRNA in mouse [40]. To test to what extent lncRNA expression levels might be under selective constraint, we devised an expression conservation index. This index is inspired by the classical McDonald-Kreitman test, which contrasts between-species divergence and within-species variation to test for natural selection [41]. We simply computed the expression level correlation among all orthologous lncRNAs (or protein-coding genes) for mouse and rat, and divided it by the average expression level correlation between pairs of individuals from the same species, for a given organ/developmental stage combination. This ratio can be seen as an indication of the degree of expression conservation between species, which takes into consideration the biological and technical variability that can be observed among individuals from the same species. It measures expression conservation at the transcriptome-wide level, not on gene-by-gene basis. For both protein-coding genes and lncRNAs, this analysis shows that expression levels are more constrained during embryonic development than in adult organisms, as expected (Figure 2.2). Likewise, for both categories of genes, the greatest extent of expression conservation is observed in the brain and in the embryonic stages of the other somatic organs, while the lowest expression conservation levels are found for the adult testes (Figure 2.2). As observed for sequence conservation analyses, lncRNA expression levels are much less conserved than protein-coding gene expression levels, in all organs and developmental stages (Figure 2.2).

2.2.6 Limited evidence for lncRNA functionality from evolutionary studies

This study and similar comparative transcriptomics analyses across species, organs and developmental stages [42] have provided several important insights into the global patterns of lncRNA evolution in mammals and other vertebrates. While the great majority of mammalian lncRNAs are predominantly expressed in the adult testes [8, 37, 38, 42], it seems likely that the minority of lncRNAs that are expressed in somatic organs and in earlier developmental stages are enriched in functionally relevant transcripts. These lncRNAs are more conserved than testes-specific lncRNAs (though still much less conserved than protein-coding genes), both in terms of sequences and in terms of expression levels. These transcripts might thus be prioritized in the search for functional lncRNAs. Interestingly, for transcripts expressed during embryonic development or in somatic organs, the strongest signals for selective constraint do not come from lncRNA exonic sequences, but from their promoters and from their splice sites. Combined with evidence stemming from *in vitro* perturbations of lncRNAs [43], this finding suggests that for some of these lncRNAs the biological function may not reside in the non-coding RNA molecule that is produced by the locus, but may be achieved by other functional elements embedded in the locus, or may reside in the process of transcription and splicing rather than in its product [40, 43]. Pinpointing the selected biological function of a lncRNA-producing genomic locus (even after ascertaining that a function is likely to exist) is a difficult task. In section 2.3 below, I will provide a brief overview of the challenges of lncRNA functional investigations.

2.3 The challenges of investigating lncRNA functions

Studying the evolution of lncRNA *loci* is not sufficient to prove or disprove their functionality and to understand their functions. Detailed *in vitro* and *in vivo* investigations of lncRNA functions are needed. However, these investigations are challenging due to the multitude of biological functions that could be achieved by a single lncRNA *locus*, which all need to be thoroughly tested in different ways [19]. It is important to state that lncRNA *loci* can have selected biological functions that are independent of the actual RNA molecule produced by the *loci*. This fact can seem counter-intuitive at first, especially given the model set by protein-coding genes, where mRNAs are essential carriers of information if not direct cellular actors. Yet, there are several situations where functional genomic regions can produce RNA molecules as dispensable by-products.

2.3.1 Associations between lncRNAs and enhancers

It is now well established that the promoters of some mammalian lncRNAs have dual roles as regulatory elements (in particular expression enhancers) for neighboring genes. More generally, it is increasingly acknowledged that gene promoters and expression enhancers do not represent two distinct classes of expression regulatory elements, but have many overlapping characteristics: similar chromatin structure, capability of activating transcription at a neighboring locus, capability of initiating transcription at the locus itself [44]. Expression enhancers, which are typically predicted based on their histone modification signatures, often generate bidirectional transcripts [45]. These transcripts are short and unstable but could be identified with RNA-seq techniques that specifically target the 5' end of the transcripts [46]. While these unstable enhancer-associated RNA molecules (or eRNAs) can not be detected with classical RNA-seq approaches, there is evidence that stable long non-coding RNAs can also be produced by some enhancer elements. Notably, it was shown that about half of mouse lncRNAs detected with RNA-seq are transcribed from genomic elements that would be typically classified as expression enhancers based on their chromatin modification patterns [39]. It then becomes important to ask whether the transcripts produced at the enhancer *locus* are functionally relevant, or just a by-product of the transient fixation of transcription factors and of the RNA polymerase. The presence of eRNAs was reported to stabilize the promoter-enhancer loops that mediate expression activation, in the context of estrogen-dependent transcriptional regulation [47]. However, there is increasing evidence that this is not always the case, at least for enhancer-associated lncRNAs.

The importance of investigating the additional roles of lncRNA *loci* as RNA-independent regulatory elements is well illustrated by *lincRNA_p21* example [48]. This lncRNA was originally reported to function as a gene expression repressor, potentially controlling the expression of hundreds of genes as part of the canonical p53 transcriptional response (p53 is a tumor-supressor protein and a key factor in cellular stress responses [49]). This biological function was proposed following observations that *lincRNA_p21* was regulated by the p53 protein, and that *in vitro* knockdown of *lincRNA_p21* resulted in differential expression for hundreds of genes, many of which were also differentially expressed upon knockdown of the p53 protein itself [48]. The lncRNA was thus presented as an important regulator of gene expression in *trans*. This model was later refuted through *in vivo* studies in mouse, which showed that the subcellular localization of the lncRNA was incompatible with *trans*-regulation of gene expression [50]. It was proposed instead that the RNA molecule acts as a *cis*-regulator of the

neighboring *p21* gene [50]. Further analyses of genetic deletions in the *lincRNA_p21 locus* in the mouse model later showed that the *cis*-regulatory functions were independent of the presence of the RNA, and that they were carried out by multiple enhancer DNA elements embedded in the *locus* [51]. Thus, at least in this context, the *lincRNA_p21* RNA molecule was functionally irrelevant.

Their dual roles as gene expression enhancers can help explain why lncRNA promoters tend to be more conserved than their exonic regions. In contrast, the excess of sequence conservation on lncRNA splice signals compared to exonic regions seems to be in favor of RNA-dependent functions - indeed, if the sequence of the lncRNA is not important, why would its splicing pattern be under selection? However, there is now evidence that splicing at lncRNA *loci* may contribute to the regulation of neighboring genes, and that this is independent of the RNA molecule that is produced by the *lncRNA* locus [43]. In agreement with this finding, it was reported that enhancers that produce spliced lncRNAs have increased activity compared to other enhancers [52]. The mechanisms that could explain how lncRNA processing may affect the regulation of the neighboring genes are not yet perfectly understood, but could involve cotranscriptional processes that alter histone methylation at the *locus*, as reported for the *COOLAIR* lncRNA in *Arabidopsis* [53].

2.3.2 Other RNA-independent functions at lncRNA *loci*

The presence of enhancer elements is not the only biological confounding factor that can mislead initial assessments of lncRNA functions. An interesting example is the *Airn* lncRNA, which is transcribed in antisense of the parentally imprinted gene *Igf2r* [54]. Transcription of this lncRNA, which is itself parentally imprinted, is required to repress the neighboring genes, in *cis* [55]. It was later shown that the *Airn* RNA molecule is not required to silence the *Igf2r* gene. Analyses of mutant genotypes with transcription termination signals inserted at various positions downstream of the *Airn* promoter showed that transcriptional overlap between *Airn* and the promoter of the *Igf2r* gene was required for the silencing of the latter, but that the *Airn* RNA molecule was likely not involved in the gene repression process [40]. For this *locus*, the proposed model for gene expression regulation involves transcriptional interference between the two sense-antisense promoters [40].

2.3.3 The case of *Hotair*: one lncRNA can hide another

The challenges of investigating lncRNA functions are well illustrated by the case of *Hotair*. This lncRNA was originally discovered in human, along with other non-coding RNAs embedded at *Hox* loci [56]. In animals, *Hox* genes are crucial transcription factors, which control the development of the main body axis but also the development of appendages such as the limbs, the genitalia *etc.* [57]. In vertebrates, *Hox* genes are organized into 4 clusters named A, B, C and D, which derive from the two rounds of whole-genome duplications that occurred in the vertebrate ancestor. During development, the series of *Hox* genes within each cluster are transcriptionally activated in a precise temporal order and in specific spatial domains. Their expression patterns are controlled by complex regulatory mechanisms, which are still extensively studied today.

The discovery of lncRNAs produced from the immediate vicinity of *Hox* genes created a great deal of interest in their potential involvement in *Hox* gene regulation [56]. In particular, it was reported that the *Hotair* lncRNA, which is produced from the intergenic region between

Hoxc11 and *Hoxc12* in the *HoxC* cluster, regulates the expression of *Hoxd* genes in *trans* [56]. Specifically, it was proposed that Hotair is required to target the Polycomb represssive complex PRC2 to the *HoxD* locus, thus contributing to gene silencing [56].

The original report of the regulatory function of *Hotair* was based on *in vitro* experiments in human fibroblasts. It was later shown that a homologous lncRNA can be recovered in mouse, but that the mouse and human lncRNAs do not have conserved exon-intron structures [58]. Moreover, the deletion of the entire mouse *HoxC* cluster, thus including the *Hotair locus*, had no detectable effect on the expression of *HoxD* genes [58]. These results were interpreted as potential functional differences between the human and mouse *Hotair* lncRNAs, but the contradictory results could also be explained by differences in experimental design (*in vitro* and *in vivo* experiments, knock-down of the *Hotair* RNA *vs.* knock-out of the entire *HoxC* cluster). To disentangle these hypotheses, the group of researchers that discovered *Hotair* genetically engineered a mouse model carrying a targeted deletion of the *Hotair locus*. They reported that the mutant mice carried several malformations, including wrist effects and homeotic transformations in the axial skeleton [59]. They also reported that the deletion of *Hotair* resulted in *HoxD* gene derepression in mouse fibroblasts, consistently with the original reports in human [59].

The claims regarding the roles of *Hotair* in *HoxD* gene regulation were revisited by Rita Amândio, a PhD student in the Laboratory of Developmental Genomics led by Denis Duboule. I closely collaborated with Rita on this project, thus having the opportunity to discover firsthand the intricacies of this case study. First, Rita re-analyzed the phenotypes of the mice carrying the targeted deletion described above [59]. She was not able to confirm the skeletal malformations that were reported originally, and could only report a mild morphological difference for one caudal vertebra [60]. Rita then generated RNA sequencing data for several tissues that were relevant for the proposed *Hotair* functions (the forelimbs and the hindlimbs, the external genitalia and three trunk segments), for wild type mice and for the mice carrying the *Hotair* deletion. We analyzed this transcriptome data, first aiming to determine whether *HoxD* genes were differentially expressed between the two mouse genotypes. This was not the case, contrary to what was reported originally. However, we uncovered significant differences in the expression levels of the *Hoxc11* and *Hoxc12* genes, which are the immediate neighbors of *Hotair* [60]. This result was again in striking contradiction with the original study [59]. To better understand the sources of these discrepancies, we performed a detailed examination of the transcriptional profile at the *Hotair* locus, in wild type and mutant mice (Figure 2.3).

With this analysis, we were able to make several unexpected observations. First, we saw that another small transcript could be detected between the *Hoxc12* and *Hoxc11* genes, on the same strand (Figure 2.3). This transcript appears to initiate from a CpG island promoter, just like the *Hoxc12* and *Hoxc11* genes. We named this transcript *AHotair*, for Antisense of Hotair (sadly, my proposal of naming it *Hoxc11¾* was not seriously considered). Interestingly, in wild type mice this transcript appears to terminate close to the *Hotair* termination site on the opposite strand (Figure 2.3). In the mutant mice, the termination sites of *Hotair* and *AHotair* are both deleted, but the *AHotair* CpG island promoter remains and now generates a longer transcript that continues until *Hoxc11*. This transcript could perhaps continue beyond this boundary, creating an alternative isoform of *Hoxc11*. This might explain why higher expression levels were found for *Hoxc11* in mutant mice in the RNA-seq data. With short-read RNA-seq we were not able to determine the full-length isoforms of these transcripts, and further work is needed to better characterize them.

On the *Hotair* strand this time, we observed that the initiation sites of the lncRNA were not all affected by the deletion. A more distal initiation site is still present and is able to generate a transcript that extends until the termination site of *Hoxc12* on the opposite strand (Figure 2.3). We named this transcript *Ghostair*, for Ghost of Hotair, because it lingers after the announced death of the *Hotair* lncRNA. The termination site of *Ghostair* is very close to the *Hoxc12* termination site. Because *Hoxc12* is more weakly expressed in mutant compared to wild type mice, it is tempting to speculate that this particular transcription localization might play a role in its downregulation, but we have no evidence to support this claim.

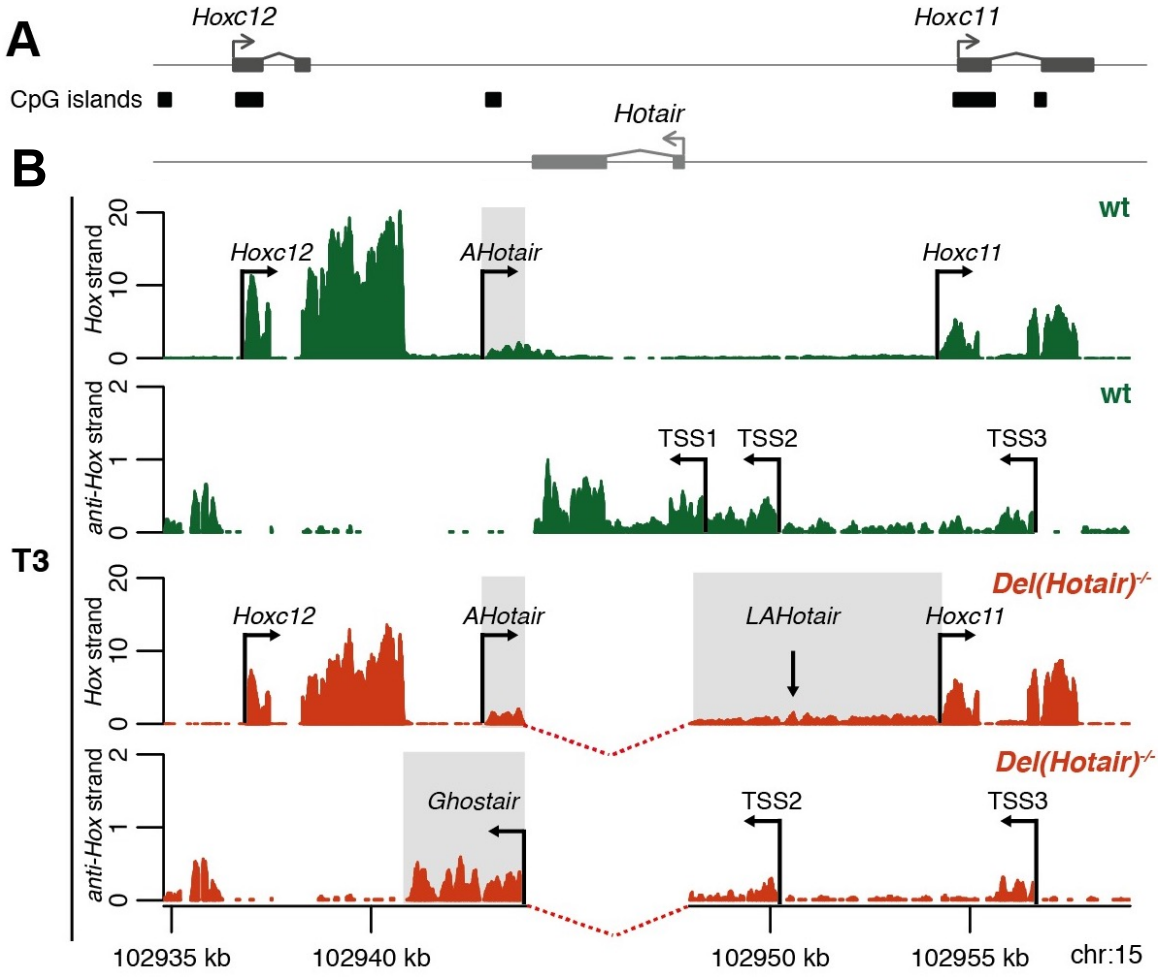

Figure 2.3: Transcription profile around the *Hotair* locus. A. Localization of *Hotair* on the antisense strand compared to *Hoxc* genes. B. RNA-seq coverage along the genomic region containing *Hotair*, in the posterior trunk of wild type mice (green) and in mice carrying a deletion of the *Hotair* locus (orange). Figure adapted from Amândio *et al.* [60].

Our analyses of the *Hotair* mutant mice showed that the transcriptional profile around this *locus* is much more complex than originally imagined. This specific deletion of the *Hotair* gene does not only result in the absence of the main *Hotair* isoform, but also in the origination of two new transcripts, *AHotair* and *Ghostair*. Any phenotypic changes or gene expression modifications between the two mouse genotypes might thus be due not only to the absence of *Hotair*, but also to the presence of these new RNA molecules, at the very least. We cannot exclude that the deletion of this genomic region affected other functional DNA elements that

may co-localize with the *Hotair* region of origin. To unambiguously assign the proposed biological functions to the Hotair RNA molecule, additional experiments are needed, including a "rescue" experiment where *Hotair* is re-introduced in the genetic background of the mutant mice (a reintroduction would be compatible with its proposed function as a *trans*-acting expression regulator). Such experiments are difficult to perform, especially *in vivo* in the mouse model, which might help explain their conspicuous paucity in the existing lncRNA literature. We also did not attempt to rescue the *Hotair* phenotypes in our revisit of the original claims, but in our defense that was mainly because we could not confirm any of the reported phenotypes and thus had nothing to rescue.

Beyond its implications for the biological functions of *Hotair*, this case study is a good illustration of the challenges of investigating lncRNA functions. A guideline for the investigation of lncRNA functions *in vivo* was already proposed [19], but the advice it gave is not consistently followed. I have the feeling that much of the current lncRNA literature consists of initial over-optimistic claims of lncRNA functions, followed by more careful reassessments and walk-backs of the original claims, as was the case for *lincRNA_p21*. This dynamics seems to be particularly prevalent for claims related to the roles of lncRNAs in human diseases. I will discuss more about this in section 2.4 below.

2.4 LncRNA relevance for human health

The discovery that the human genome is pervasively transcribed into non-coding RNAs [32] brought new hope for the study of many human diseases, for which the molecular mechanisms were not yet perfectly understood or for which molecular drug targets could not yet be identified. Soon after the realization that the human genome encodes tens of thousands of lncRNAs, these transcripts became an important research focus in biomedicine. One of the many diseases for which lncRNA research became highly prevalent during the past decade is hepatocellular carcinoma (HCC). HCC is a major cause of cancer-related death world-wide. Because HCC is generally detected late, surgery to remove tumors is not possible for the majority of patients, and available chemotherapies have only a limited effect on patient survival [61]. The search for new disease biomarkers and for new molecular targets for chemotherapy is thus understandably intense for HCC. I became interested in the relevance of lncRNAs for HCC following a collaboration with Markus Heim at the Department of Biomedicine of the University of Basel, during which we investigated the transcriptional response to hepatitis C infection and interferon treatment in the human liver [62]. Together with other collaborators, Markus's research group set out to investigate the genomic, transcriptomic and proteomic differences between HCC tumors and healthy liver tissues [63]. They generated transcriptome sequencing data for more than a hundred HCC patients, for tumor and adjacent liver tissues, as well as for healthy livers. This data enabled us to analyze the expression patterns of lncRNAs in HCC, and to reevaluate the existing claims regarding lncRNA relevance for this disease [64].

In this project, we started out by doing a literature search, to get a better idea of how many lncRNAs were previously reported to be important for HCC. We queried PubMed with the keyword "hepatocellular carcinoma" and retrieved the titles and abstracts of the corresponding PubMed records. We then searched the abstracts to identify gene names, by matching words with a list of human gene names obtained from the Ensembl database. We were thus

able to identify articles that specifically mentioned lncRNA names in their abstracts; note that we did not systematically checked the nature of the association between these lncRNAs and HCC reported in each article. We found that the number of HCC-related articles that cite lncRNAs increased dramatically between 2009 and 2019, with almost 7% of all HCC-related publications mentioning lncRNAs in their abstracts (Figure 2.4). However, the overwhelming majority of lncRNAs were cited in just one HCC-related article (Figure 2.4B). This is understandable because the research field is still very new, but it nevertheless highlights the need to re-assess and reproduce lncRNA-related claims in HCC.

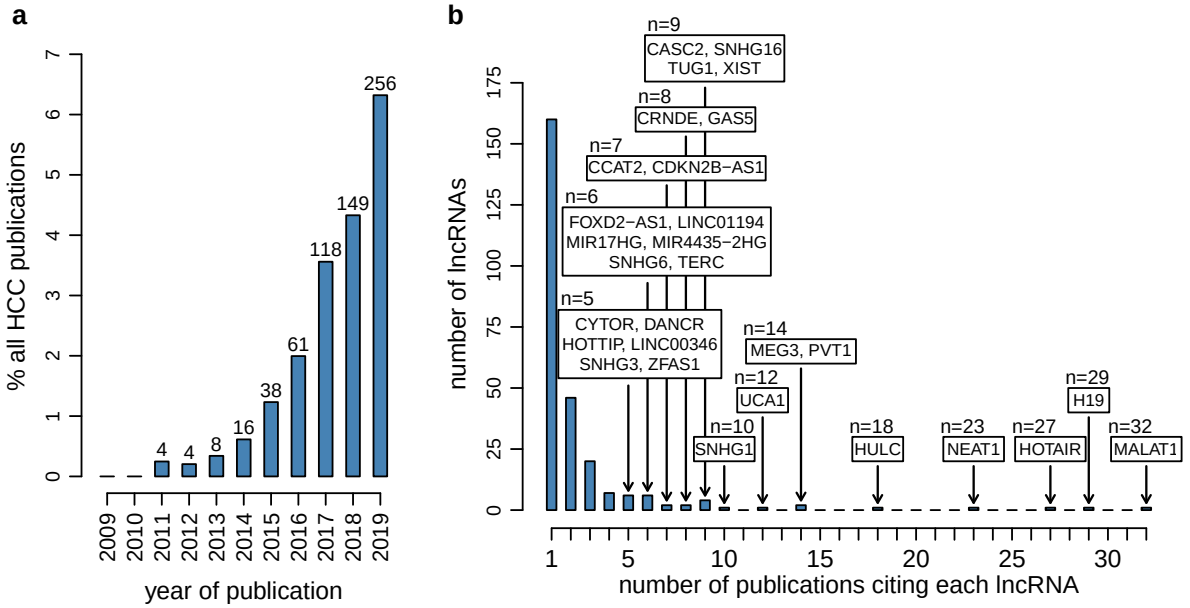

Figure 2.4: Dramatic increase in the number of HCC-related scientific publications that cite lncRNAs. A. Number of HCC-articles citing lncRNAs published each year between 2009 and 2019. B. Histogram of the number of HCC-related publications that cite each lncRNA. Figure from Necsulea *et al.* [64].

We then analyzed into more depth the lncRNAs that were reported to be associated with HCC by at least 5 publications. At the top of the list (Figure 2.4B) we found several lncRNAs that have been extensively studied in many biological contexts, not just in HCC. For example, *NEAT1* and *MALAT1* (also initially named *NEAT2*) were discovered in a screen for nuclear-enriched transcripts [65], more than 15 years ago. *H19*, the first lncRNA ever described [66], is a parentally-imprinted transcript that contributes to the control of placenta development and embryonic growth [67]. Also at the top of the list is *HOTAIR*, the human homologue of the mouse *Hotair* lncRNA, which we discussed in depth in the subsection 2.3.3 above. The most cited lncRNAs also include *XIST*, which is the well known regulator of X-chromosome inactivation in placental mammals [68]. The fact that the most frequently cited lncRNAs in association with HCC were originally described in other biological contexts is not surprising, given that these lncRNAs were reported to have important roles in gene expression regulation and are thus likely to attract attention when attempting to understand the molecular mechanisms driving tumorigenesis.

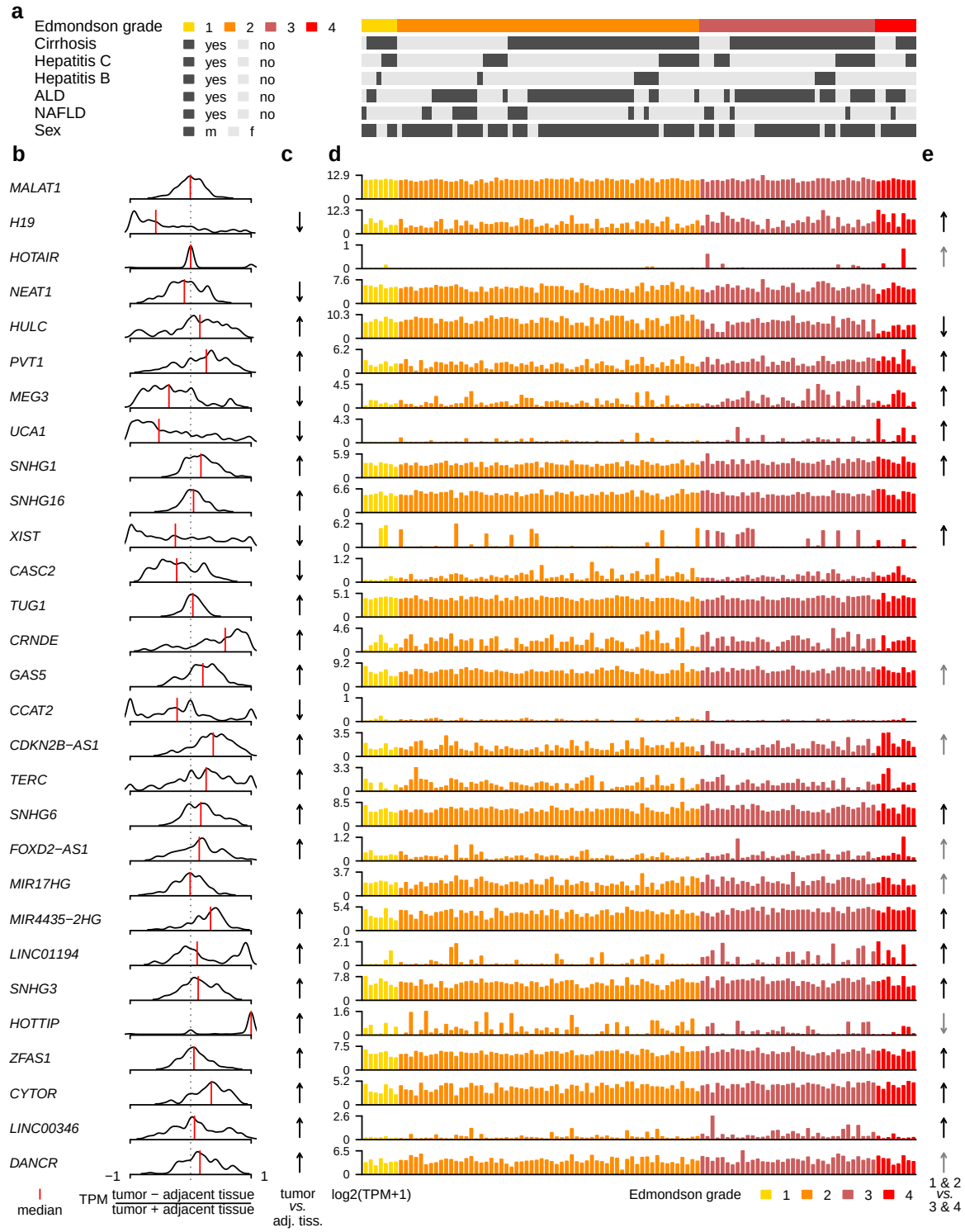

Figure 2.5: Expression patterns of HCC-associated lncRNAs. A. Distribution of sample characteristics for the transcriptome data included in our analysis. The Edmondson-Steiner grade represents the degree of tumor differentiation (grades 1 to 4 from low tumor differentiation to high tumor differentiation, corresponding to early and late cancer stages). B. Expression patterns of 29 HCC-associated lncRNAs (density plots of differences between tumors and adjacent samples across patients). C. Arrows indicate whether significant expression differences were found between tumors and adjacent tissues. D. Barplot of the expression patterns in the tumor samples, colored by tumor grade. E. Arrows indicate whether significant expression differences were found between low-grade and high-grade tumors. Figure from Necsulea *et al.* [64].

We then attempted to test whether the expression patterns of these frequently HCC-associated lncRNAs were coherent with their previously proposed biological roles. However, in many cases it was difficult to understand what these roles might be, because of conflicting claims in the literature. For example, *MALAT1* (short for Metastasis Associated Lung Adenocarcinoma Transcript 1), was initially reported to promote metastasis in breast cancer [69], but was more recently proposed to inhibit metastasis in the same cancer type [70]. In HCC, *MALAT1* is generally proposed to promote cancer progression (see *e.g.* Li *et al.* [71]). In our transcriptome data, *MALAT1* was not differentially expressed between tumors and adjacent tissue samples, nor was it more highly expressed in advanced stage tumors (Figure 2.5), as one might have expected from a tumorigenesis-promoting gene. Likewise, *H19* was proposed to act as an oncogene [72] and as a tumor suppressor [73]. In our data, *H19* is expressed at lower levels in tumors than in adjacent samples, but at higher levels in late-stage tumors compared to early stage tumors (Figure 2.5). This complex expression dynamics is not consistent with a simple role for *H19* in HCC. We also observed that *HOTAIR* is not differentially expressed between tumors and adjacent samples, and is barely detectable in all but two RNA-seq samples (Figure 2.5). Thus, our data does not support an important role for *HOTAIR* in HCC, despite the numerous articles claiming so.

Evidently, analyses of lncRNA expression levels in tumors, adjacent tissues and healthy livers are not sufficient to claim that these transcripts can act as tumor suppressors or oncogenes. However, numerous recent publications in this research field rely on little more than comparative transcriptomics analyses to do so. Experimental validations are often performed *in vitro* using immortalized cell lines, which may not recapitulate well the situations encountered *in vivo*. Clearly, *in vivo* experiments are difficult to perfom when investigating human diseases, in the absence of an animal model that could recapitulate their molecular underpinnings. The development of better-suited model systems, such as biopsy-derived organoids [74], may help understand the molecular mechanisms driving HCC, including the roles of lncRNAs therein. Moreover, as the lncRNA research field gains in maturity, publications claiming major biological roles for lncRNAs (whether related to HCC or in other contexts) without proper validation will hopefully become less frequent.

At present, it is difficult to navigate in the recent lncRNA literature without being overwhelmed by the sheer number of articles claiming biological functions for these transcripts, which often do not stand up to scrutiny. The increase in the number of predatory journals in biomedicine and biology in general has greatly contributed to this problem [75]. Because of this, I have found it difficult to keep working on this topic, and I am gradually shifting towards other research subjects. In section 2.5 I will try to explain why I found working on lncRNA challenging, and why I am nevertheless still very curious about lncRNAs and following this research topic from a distance.

2.5 Future research around and lncRNAs

During my post-doctoral research at the University of Lausanne, I was mainly interested in the evolution of lncRNAs although I also participated in other projects. I was keen to follow my initial assessment of the evolution of lncRNAs with an evo-devo approach, and I joined the Laboratory of Developmental Genomics with this objective. The research environment in this laboratory was very different from the ones I had experienced in Lyon or previously in Lausanne. For the first time, I was part of a research group where bioinformatics

approaches were very much secondary, and where "evolution" meant observable phenotypic changes rather than mutation biases, gene conversion, recombination, changes in expression levels *etc.* My 3½ years in this research group were an amazing opportunity to learn new things. I was definitely greatly influenced by the philosophy of the laboratory. I moved away from my original interests in lncRNAs not only because I discovered other research questions, but also because I started seeing more and more the bad part of the lncRNA literature. I was part of a laboratory where biological hypotheses were tested with exquisite genetic manipulations, where controls were carefully designed and where considerable time and financial resources were invested towards doing high-quality genetics research. Partly because of this exposure, I became acutely aware of the lack of rigour that characterizes some of the recent lncRNA research. The *Hotair* case study was eye-opening, especially since I had been enthusiastic upon reading the first report of this lncRNA and other presumed lncRNA regulators of *Hox* gene expression in *trans*. My tendency to be skeptical about lncRNA functionality claims was thus further reinforced during my stay in the Laboratory of Developmental Genomics.

While the literature claiming crucial lncRNA functions in human diseases is plagued by questionable research published in predatory journals, the field of lncRNA evolution is much healthier. I could have thus decided to pursue my original research objective. However, I decided against continuing in this direction because the competition in this field was of excellent quality, starting with my post-doctoral advisor, Henrik Kaessmann. I was able to finalize my first proposed research project [38], but I realized I could not bring any major contributions to this topic given the high-quality work that was rapidly accumulating in this area [42, 76].

Despite the decision to change research directions, I am still interested in some aspects of lncRNA biology, and more generally in the functionality of various transcriptome features. In particular, I am collaborating with Carina Mugal to study the potential implications of lncRNAs in the establishment of reproductive barriers during speciation in flycatchers. We started this collaboration by co-supervising Hugo Seytier for his M2 internship in 2023. Hugo identified several lncRNAs with divergent expression patterns between two flycatcher species and their F1 hybrids, which may represent good candidates for mechanisms involved in reproductive isolation. In a more distantly related project, I am collaborating with Laurent Duret to study the relationship between transcriptome complexity and effective population sizes in eukaryotes. We have co-supervised Florian Bénitière for his M2 internship in 2018. Florian is now finalizing his PhD, co-supervised by Laurent Duret and Tristan Lefébure, and more distantly co-advised by myself. Florian's first research paper showed that the prevalence of alternative splicing is negatively correlated with effective population sizes, supporting the hypothesis that many of the low-frequency alternative splicing variants that are detected with high-throughput RNA-seq data likely represent biological noise [25].

One of the aspects that initially attracted me towards lncRNA research is their involvement of lncRNAs in X-chromosome inactivation, which is still today one of their major, unambiguously proven functional roles. I am still fascinated by this topic, in particular by the convergent recruitment of non-homologous lncRNAs in the X-inactivation process in placental and marsupial mammals [77], and by the lineage-specific emergence of other lncRNAs involved in the control of gene expression on the X chromosome [78]. The evolutionary origin of the *Xist* lncRNA, which is the key player of X-chromosome inactivation in placental mammals, is also very exciting. Duret *et al.* [79] showed that *Xist* appeared in the ancestor of placental mammals following the pseudogenization of a protein-coding gene. In a recent project, I have attempted to explore into more detail the evolutionary events that may have

led to the emergence of *Xist*. I will briefly describe this project in chapter 5.

Chapter 3

Functionality of long-range chromatin interactions

Since 2018, my research has shifted away from long non-coding RNAs and towards the evolution of *cis*-regulatory mechanisms of gene expression. I have been particularly interested in the regulatory relationships that take place at long genomic distances between gene promoters and distal *cis*-regulatory elements. This was the main focus of Alexandre Laverré's PhD thesis, whom I co-supervised with Eric Tannier. In this chapter, I will present an introduction to this topic. I will focus on the functional significance of long-range chromatin interactions in vertebrates, and to a lesser extent on their evolution, which will be discussed in depth in chapter 4. This chapter largely corresponds to the first draft of a review article, which I am currently preparing following an invitation from Genome Biology and Evolution. This article will soon be available as a preprint on bioRxiv. Note that the figures that will accompany this article have yet to be prepared. To illustrate the PCHi-C technique, on which we based many of the analyses described in chapter 4, I include a figure from Schoenfelder *et al.* [16], for which I have yet to obtain usage permission.

3.1 Complex *cis*-regulatory landscapes in vertebrates

Variations in gene expression are central to the biology of complex multicellular organisms. Within species, gene expression levels vary among cell types, developmental stages, physiological states, in response to external stimuli, etc. Gene expression changes are also thought to play an important role in establishing phenotypic differences between species [80]. This expression versatility is enabled by intricate regulatory mechanisms, which involve interactions between *trans*-acting factors (*e.g.*, proteins or non-coding RNAs) and DNA sequences in *cis* (*i.e.*, found on the same chromosome as the target gene), either in the immediate vicinity of the gene (proximal elements) or further away (distal elements). *Cis*-regulatory DNA elements are typically further divided into enhancers and silencers depending on their activating or repressing roles on gene expression, although this seemingly clear division is an oversimplification of biological reality [81]. Both types of elements were discovered more than three decades ago [82, 83], but knowledge about their biochemical characteristics and about their modes of action has rapidly accumulated in recent years, thanks to technological innovations.

Expression enhancers are commonly predicted based on their chromatin modification sig-

natures, which include methylation and acetylation of histone tails [84]. Chromatin immunoprecipitation followed by sequencing (ChIP-seq) experiments that target specific histone marks [85] or the proteins that deposit them [86] have become standard approaches to map enhancer elements genome-wide. Techniques that aim to identify open chromatin regions, such as DNAse hypersensitivity [87] or transposase accessibility assays [88], are also typically used for enhancer prediction. Furthermore, following the discovery that enhancers are frequently transcribed into (often short-lived, unstable) non-coding RNAs [45, 89], techniques that can capture nascent transcripts have also been used to predict the positions of these regulatory elements [46]. Importantly, the elements detected with these techniques should be considered as candidate or predicted enhancers, as their ability to activate expression is typically not tested. Other approaches, such as STARR-seq [90], can directly test the ability of putative enhancer sequences to drive gene expression *in vitro*, in constructs where the enhancer is inserted close to a reporter gene and a minimal promoter. However, these methods also cannot offer a guarantee of the activity of the predicted element *in vivo*, which likely depends on additional factors including chromatin accessibility.

In contrast with enhancers, gene expression silencers have proven more difficult to predict at the genome-wide level, mainly because silencer-specific chromatin signatures are lacking [91]. Recently, silencers were predicted genome-wide based on subtractive chromatin status analyses, which identified open chromatin regions that did not have the histone modification profiles typical of enhancer or insulator elements [92]. Open chromatin regions that carry the repressive histone modification H3K27me3 were also proposed as putative silencer elements [93, 94]. The silencer activity of candidate elements defined with such approaches could then be tested *in vitro* with massively parallel reporter assays [95, 96]. Similar to enhancer elements, silencers are thought to control gene expression in a position-independent manner, through chromatin contacts that bring them into physical proximity with gene promoters [91, 94]. Interestingly, in *Drosophila* almost all elements identified as silencers in a given cellular context were found to act as enhancers in different contexts [97], indicating that regulatory elements with dual roles as activators or repressors are more common than originally thought. *Cis*-regulatory elements with dual activator/repressor roles also exist in vertebrates [98], although it is not yet clear whether they are the exception or the rule.

Comparisons of genome-wide chromatin modification maps across cell types and tissues showed that putative enhancers are often cell type or tissue-specific, much more so than gene promoters [99, 100]. These elements may thus contribute to the modularity of gene expression regulation, by activating gene expression in specific spatio-temporal contexts [101]. Individual enhancer elements may carry binding sites for multiple transcription factors, which further complicates the combinatorial control of gene expression [102]. Moreover, multiple elements with similar spatiotemporal patterns of activity may contribute to the regulation of the same gene, potentially contributing to the robustness of gene expression control [103].

The complexity of *cis*-regulatory mechanisms of gene expression in vertebrates is further underlined by the fact that many regulatory elements are situated far from their target genes. The first long-range relationships between *cis*-regulatory elements and gene promoters were discovered through genetic manipulation approaches or through functional dissections of loci involved in human genetic diseases [104–106]. These pioneering studies showed that enhancers can be situated several hundreds of kilobases (kb) away from their target genes, and that they did not necessarily control the expression of the closest neighboring genes. It was later shown with chromosome conformation capture (3C) techniques [5] that these

regulatory elements and the target gene promoters are brought into physical proximity in the nucleus [14, 107, 108]. These long-range chromatin interactions (also called chromatin contacts or chromatin loops) are now commonly detected at the genome-wide level, thanks to the development of 3C-derived techniques. Below, we will summarize what is currently known about the prevalence, the functional relevance and the evolutionary implications of these interactions.

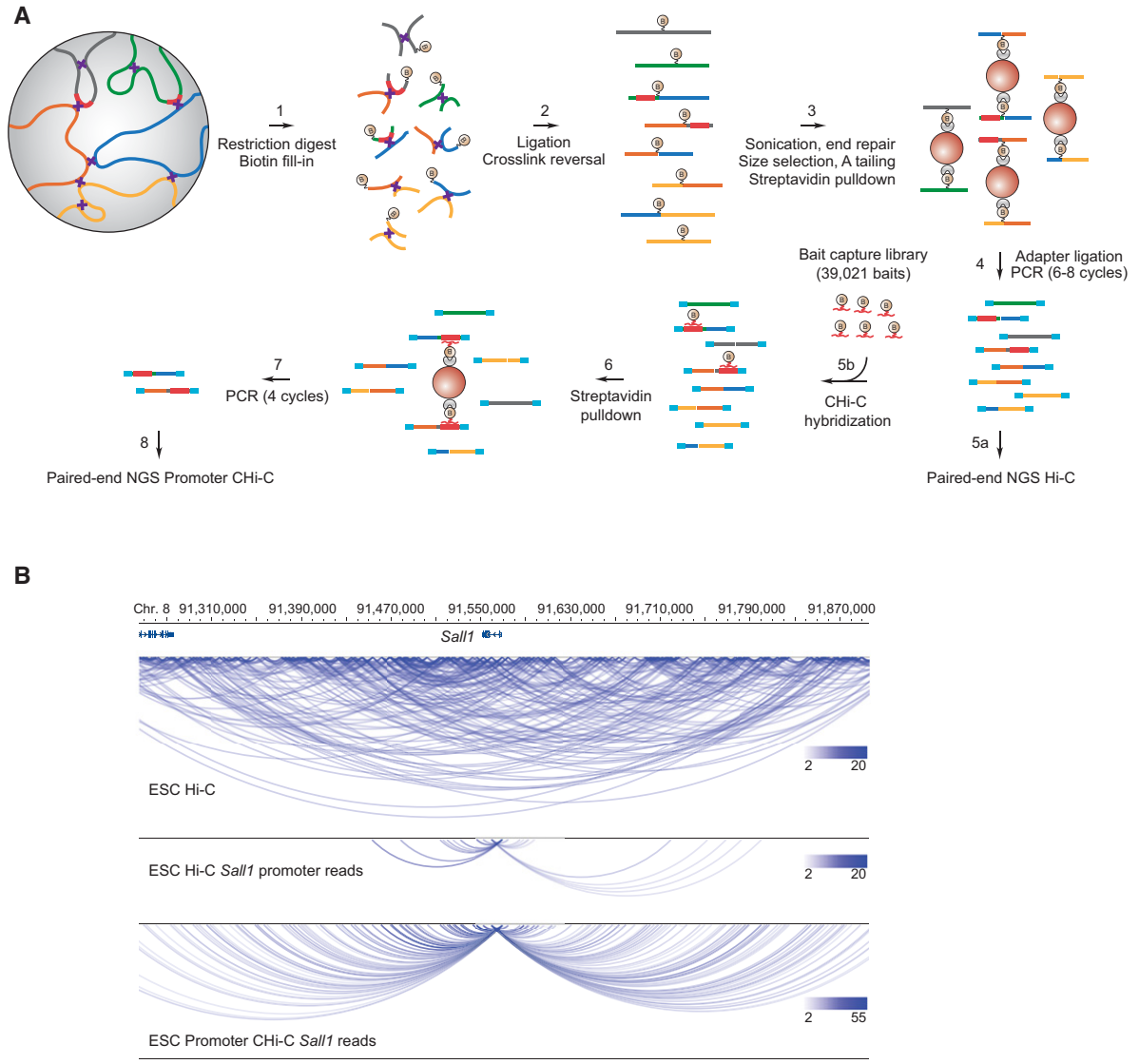

Figure 3.1: Illustration of the PCHi-C technique. A. Illustration of the PCHi-C protocol, which includes chromatin cross-linking, proximity ligation and capture of regions of interest using RNA baits. B. Illustration of the chromatin contacts that can be detected with the Hi-C approach (two panels at the top) and with the PCHi-C approach (bottom panel). Figure from Schoenfelder *et al.* [16].

3.2 Identification and prevalence of long-range chromatin interactions

Long-range chromatin interactions between gene promoters and distal regulatory elements can now be detected at a genome-wide level with several types of chromatin conformation assays. Among the many derivatives of the original chromosome conformation capture (3C) approach, Hi-C is probably the most widely used today [109]. However, the complexity of the libraries prepared with this technique, which can capture chromatin interactions between all possible pairs of genomic fragments, is too high to enable detection of specific contacts at high resolution. Other related approaches are better suited for the investigation of regulatory interactions between promoters and distal elements. For example, an application of the ChIA-PET technique, which combines immunoprecipitation of a protein of interest with proximity ligation of DNA fragments found in the same chromatin complex, was able to reveal several thousands of long-distance promoter-enhancer interactions in human cells [110]. A similar approach that combines chromatin immunoprecipitation with Hi-C was likewise able to reveal long-range associations between promoters and enhancers [111]. The promoter-capture Hi-C (PCHi-C) approach, a derivative of Hi-C that targets a set of predefined gene promoters, was specifically designed for the identification of putative regulatory interactions [16, 112]. Moreover, the Micro-C and capture Micro-C approaches, which differ from standard Hi-C approaches in that genomic fragments are obtained by micrococcal nuclease digestion rather than by treatment with restriction enzymes, also show great potential for the fine-scale detection of promoter-enhancer interactions [113, 114]. These techniques each have specific advantages and disadvantages for the detection of regulatory interactions. For example, ChIA-PET and HiChIP target interactions mediated by a specific protein or protein complex, and can thus specifically identify chromatin contacts associated with transcription factors or with the RNA polymerase, thus enriching for genuine regulatory interactions. On the contrary, the PCHi-C approach does not rely on the presence of a predefined protein and can thus detect chromatin interactions that pre-date transcriptional activation [115].

These fine-scale investigations of putative regulatory interactions in human and mouse showed that about two thirds of chromatin interactions occurred between an element and its nearest gene promoters, while the remainder bypassed at least one active or inactive promoter [16, 110, 112]. PCHi-C analyses showed that the frequency of interactions between pairs of genomic elements decreases rapidly with increasing genomic distances between the two elements, as expected based on previous knowledge obtained from Hi-C data [16, 112]. Nevertheless, the median distance between pairs of interacting promoters and enhancers was generally above 100 kilobases (kb), although this was dependent on the analyzed cell type [16, 112]. Depending on the technique and on the cell type that was analyzed, several thousands of chromatin interactions could be detected between regions separated by more than 500 kb [16, 110, 112]. A non-negligible proportion of ultra-long-range interactions, spanning tens of megabases (Mb), was also revealed recently [116].

Importantly, long-range interactions between gene promoters and putative regulatory elements are not restricted to actively expressed genes. Some of the chromatin interactions observed for inactive genes could be explained through regulatory contacts with silencer elements [91, 94]. Indeed, genomic regions carrying the repressive chromatin mark H3K27me3 are known to form chromatin contact clusters, which span long genomic distances in *cis* and include *trans*-interactions [117, 118]. However, not all chromatin interactions involving in-

active genes have repressive roles. It is increasingly acknowledged that chromatin contacts between gene promoters and activating regulatory elements can precede gene expression and that the pre-formed interactions may contribute to the efficiency of transcriptional activation [119]. Interestingly, PCHi-C data analyses consistently showed that transcriptionally active promoters tended to interact with more distant genomic elements, compared to inactive promoters [16, 112].

The first reports of long-range regulatory relationships between promoters and distal *cis*-acting elements were all related to developmental transcription factors, such as *DACH*, *SOX9* and *SHH* [104, 105, 120]. Given that genes with functions in embryonic development and in transcriptional regulation are significantly enriched in the vicinity of large intergenic regions, which may harbor distant regulatory elements [121], an over-representation of these biological functions is perhaps expected. However, genome-wide analyses of promoter-enhancer contacts did not reveal any particular functional enrichment among the genes that are generally involved in long-range interactions [16, 112, 122, 123]. Nevertheless, developmental genes play a key role in a specific class of long-range chromatin interactions, namely those mediated by the Polycomb repressive complex, which carry H3K27me3 histone marks [16, 117, 124]. Perhaps consistently, the prevalence of long-range interactions was also reported to vary significantly among cell types in mouse, with higher frequencies observed for embryonic stem cells, where the Polycomb repressive complex is known to play an important role in gene repression [123].

3.3 Long-range regulatory interactions and topologically associating domains

Long-range regulatory interactions need to be discussed in conjunction with topologically associating domains (TADs), which are large genomic regions in which chromatin interactions occur preferentially, to the exclusion of neighboring regions [125]. Because TADs can be uncovered with genome-wide Hi-C approaches, which are easier to implement than the capture-based techniques needed to identify fine-scale long-range chromatin interactions, they have been extensively investigated in the past decade. Their biological and physical characteristics have been discussed in numerous in-depth reviews [126–129]. Here, we will only discuss TADs insofar as they pertain to the establishment and the evolution of long-range regulatory interactions.

TADs were originally presented as a central feature of the organization of *cis*-regulatory interactions in mammalian genomes [125]. Indeed, TADs tend to overlap with clusters of co-regulated promoters and enhancers [100]. A commonly accepted view is that TADs function as facilitators of chromatin interactions between promoters and enhancers, which also act to inhibit regulatory interference from regulatory elements outside of TADs, to prevent aberrant gene expression [115]. However, additional regulatory mechanisms must be at work within individual TADs to confer expression specificity to the genes embedded therein, which do not all show correlated expression patterns [115]. Moreover, regulatory interactions between promoters and enhancers can cross TAD boundaries. The great majority of long-range chromatin interactions detected with PCHi-C occur within TADs, but 6 to 10% were found to bridge TAD boundaries in mouse cells [16]. In human cells, about one third of PCHi-C long range interactions were found to occur across TAD boundaries [122]. Additional data

from genetic manipulations introducing artificial TAD boundaries in mouse cells showed that strong chromatin interactions between promoters and enhancers, in particular those with high regulatory activity, could bypass these boundaries [130, 131]. Thus, the coherence between long-range chromatin contacts and TADs is far from perfect. This may be in part due to the technical difficulty of identifying these genome architecture features [132].

Viewing TADs as a higher-order control structure for regulatory chromatin interactions has led to the proposal that they may have had a major influence on the evolution of gene regulation [133]. In particular, the fact that TAD boundaries inhibit regulatory interactions (at least to some extent), was proposed to play a role in the integration of newly-evolved regulatory elements in gene expression control networks. According to this hypothesis, restricting newly-evolved regulatory elements to a limited number of putative target genes within the same TAD may have reduced their potential deleterious effects and facilitated their maintenance through evolutionary time [133]. Conversely, the evolution of new TAD structures through genomic rearrangements was proposed to have contributed to functional innovations in gene expression regulation [134]. Moreover, it was proposed that TADs may represent a fertile environment for the evolution of new regulatory relationships, potentially contributing to the emergence of pleiotropy for key developmental transcription factors [135]. These hypotheses are attractive but need to be thoroughly tested through detailed comparative analyses of TADs and regulatory chromatin interactions. The evolution of TADs has already begun to be explored. However, such evolutionary studies are still at an embryonic stage, as illustrated by the difficulty of reaching a consensus on whether TADs are conserved or not [136]. Comparative studies of comparable chromosome conformation data are still scarce [136], but are starting to become available [137]. Further such comparisons are needed to explore the complex question of TAD evolution.

While the impact of evolutionary changes of TADs on gene expression regulation remains for now speculative, genetic manipulations performed on model organisms have provided valuable data on the effects of TAD perturbations. These can have a wide range of consequences on gene expression and on phenotypes in mammals, ranging from mild changes to highly deleterious effects. It was shown that alterations of TAD configurations, in particular of TAD boundaries, are responsible for some human limb malformations [138] and for the activation of oncogenes in human cancer cells [139]. Likewise, the formation of new TADs was proposed to explain the deleterious effects of segmental duplications around the *Sox9* locus in mouse [140]. In other contexts, changes in TAD structures had no detectable phenotypic effects and only mild effects on gene expression [131, 141]. These contrasting results show that it is impossible to propose a unified model for the implications of TADs in gene expression regulation. In this context, we propose that individual chromatin interactions may be a better suited unit for the study of gene expression regulation rather than complex TADs. Thanks to the improvement of high-resolution chromatin conformation capture technologies, it is gradually becoming possible to disentangle the effects of TADs and of individual regulatory chromatin contacts on the control of gene expression.

3.4 Are chromatin interactions required for gene expression regulation?

It is important to note that the causal relationship between chromatin interactions and gene expression regulation is not as clear as it may seem at first. The commonly accepted model is that chromatin loops are needed to bring distal enhancers in physical proximity with gene promoters, in order to activate gene expression. This model relies on substantial evidence for chromatin contacts between promoters and regulatory elements, and is also supported by genetic manipulations experiments, which successfully activated gene expression by forcing chromatin loops with distant enhancers [142–144]. However, there is also evidence that physical proximity between promoters and enhancers is not always required for gene activation. For example, it was shown that distal enhancers could activate *Shh* expression in mouse, during the transition from embryonic stem cells to neural progenitors, without an increase in physical proximity [145]. Contrary to what was expected, live imaging experiments showed that the physical distance between enhancers and the target gene promoter increased following enhancer activation [145]. These results suggested that gene regulation may be mediated by additional chromatin conformation changes, in addition to the now well-studied chromatin loops [145]. Other indications that chromatin interactions may not be needed for gene regulation came from genetic manipulations that abrogated chromatin interactions between the *Sox2* gene and its control region in mouse embryonic stem cells, with no observable effect on transcription [146]. Interestingly, live-cell imaging had previously been unable to reveal a physical proximity in the nucleus between *Sox2* and the same control region during transcriptional activation [147], although a long-range interaction was consistently detected with chromatin conformation capture data [146].

Additional mechanistic models of gene expression regulation were proposed to explain why discrete chromatin interactions between promoters and distal enhancers may not always be necessary to activate gene expression. For example, a transcription factor (TF) activity gradient model was proposed to explain how distant enhancers might act to activate gene expression without directly contacting the promoter [148]. According to this model, the binding of a TF to a distant enhancer would result in a post-translational modification (*e.g.* acetylation) of the TF protein. The modified TF would then diffuse towards the gene promoter, where it would activate transcription [148]. Another model posits that transcriptional activation might be achieved through spatial clustering of multiple activator elements and of RNA polymerases, rather than by the formation of discrete loops between promoters and enhancers [149]. This "hub" model for transcriptional activation is supported by evidence from imaging experiments, which revealed spatial clusters of enhancers and recruited regulatory factors in the nucleus [150].

In addition to these biological considerations, technical aspects must also be taken into account before attributing causal regulatory roles to chromatin contacts detected with genome-wide technologies such as PCHi-C, HIChIP, ChIA-PET etc. These techniques are characterized by high sensitivity for the detection of discrete chromatin interactions between gene promoters and other genomic regions. However, there is no reason to believe that the predicted chromatin interactions all have regulatory roles, or even that they are all biologically relevant. Computational methods aiming to predict chromatin interactions from this type of data rely on models of the probability of chromatin contact between two genomic regions depending on the linear genomic distance between the two regions and other factors [151, 152].

Predicted chromatin contacts are those deemed to be more frequent than expected, given the values taken by the variables considered in the model. False positives may arise from factors that are not accounted for in the models, or from inaccurate estimations of the confounding factor values. For example, strong interactions may be predicted between regions that are closer in the genome of the sample under scrutiny than in the reference genome used for the analysis, due to structural variants or to genome assembly errors. Even excluding false positives, chromatin interactions detected between promoters and other genomic regions are not all claimed to have regulatory roles. The genomic regions contacted by promoters are enriched in histone modifications that are typical of regulatory elements, but do not all overlap with predicted regulatory elements [16, 112, 153]. Importantly, Hi-C-derived approaches are not able to pinpoint associations between gene promoters and individual regulatory elements, as interactions are detected between restriction fragments, which can be much longer than a typical enhancer. Overall, promoter-centered chromatin contacts must be considered to be enriched in, but not synonymous with, regulatory interactions.

3.5 Long-range regulatory interactions constrain genome evolution

The existence of long-range regulatory interactions raises important questions with respect to their evolutionary robustness and to the constraints they might impose on large-scale genome evolution. The distance between two genomic regions found on the same chromosome is positively correlated with the probability of observing a genomic rearrangement (a large-scale inversion, translocation, duplication or deletion) within the interval. Thus, pairs of gene promoters and regulatory elements are more likely to be directly affected by rearrangements if they are separated by large genomic distances. As a consequence of genomic rearrangements, the chromatin contact between the two elements may be abrogated. Indeed, chromatin interactions are rarely detected in *trans*, or in *cis* but at distances above 10 Mb [16, 112]. Thus, selective pressures acting to maintain long-range regulatory interactions may eliminate some genomic rearrangements. Interestingly, the presence of large genomic regions that were maintained in conserved synteny for long evolutionary periods was used to predict long-range regulatory relationships, much before the deployment of dedicated chromatin conformation capture techniques [154–156]. A similar principle was used to predict regulatory elements and their target genes on the mammalian X chromosome, many of which were successfully validated with transgenics experiments in zebrafish [157]. The properties of the regulatory relationships between enhancer elements and target genes predicted with these approaches are consistent with analyses based on chromatin conformation data. For example, the number of enhancers predicted to be associated with each target gene, based on synteny conservation, is positively correlated with gene expression breadth [158].

Another interesting finding with respect to the impact of long-range regulatory interactions on genome evolution is that the recombination rate is reduced between interacting elements, leading to the presence of "recombination valleys" in regulatory domains [159]. This suggested that combinations of alleles in gene promoters and distal regulatory elements may be favored by selection and thus need to be transmitted together, although mechanistic explanations involving DNA methylation were also proposed [159]. However, it was later shown that regions involved in long-range chromatin interactions are not in linkage disequilibrium, despite the previously observed reduction in recombination rate [160].

3.6 Evolution of chromatin interactions between promoters and regulatory elements

To date, very few studies have directly addressed the evolution of chromatin interactions between promoters and regulatory elements. However, the evolutionary conservation of individual interactions was often tested when investigating the phenotypic effects of the disruption of chromatin architecture. One of the first studies reporting a strong phenotypic effect of structural variants in TADs was based on the observation of genomic rearrangements in patients with limb malformations, which were then recapitulated through genetic manipulations in the mouse model [138]. The chromatin interaction profile was found to be similar in patient-derived fibroblasts and in mouse cells, for the key limb developmental genes that were under investigation [138]. Similar chromatin interactions between genetically engineered mouse strains and patients suffering from genetic limb malformations were also found at other developmental genes [161]. At larger evolutionary distances, conserved regulatory interactions were found for HoxD genes between mouse and chicken [162]. However, differences in chromatin architecture affecting the regulation of the *Shh* gene were also found between mouse and human [163].

These examples, which pertain to a very narrow class of genes responsible for correct embryonic development, cannot be used to make a general statement regarding the evolutionary conservation of long-range regulatory interactions at the genome-wide level. However, at present, genome-wide comparative studies of chromatin contacts are still scarce. We have previously performed a comparative analysis of putative regulatory interactions, based on human and mouse PCHi-C data [153]. This analysis was based on heterogeneous collections of chromatin conformation data, derived from different cell types for human and mouse, and may thus have under-estimated the true extent to which regulatory chromatin interactions are conserved during evolution. Nevertheless, this study was able to show that chromatin interactions between promoters and predicted enhancers were significantly more conserved than expected by chance. Specifically, on average, about 12% of promoter-enhancer contacts were conserved between human and mouse, while only 1% of conservation was expected based on simulated data [153]. Interestingly, the excess of contact conservation compared to simulations was stronger for promoter-enhancer contacts separated by large genomic distances, suggesting that long-range interactions may be under stronger purifying selection. This study also confirmed that genomic rearrangements that alter chromatin interactions are likely counter-selected, as the homologous sequences of promoter-enhancer pairs detected with PCHi-C data in human or mouse were found to be in conserved synteny in a wide range of vertebrate species [153].

3.7 Interplay between the evolution of regulatory chromatin interactions and the evolution of gene expression

The question of the contribution of long-range chromatin interactions to gene expression regulation is also important from an evolutionary perspective. If they have a substantial contribution to gene expression regulation, evolutionary changes in long-range chromatin in-

teractions should lead to evolutionary changes in gene expression. This hypothesis was tested indirectly in a comparative analysis of PCHi-C interactions between human and mouse [153]. This study showed that the extent of evolutionary conservation of regulatory landscapes was positively associated with the evolutionary conservation of gene expression patterns. This result is encouraging and consistent with the commonly accepted role of chromatin interactions in the control of gene expression. We note however that the association between regulatory evolution and gene expression evolution was mild, consistent with previous work which failed to uncover a strong relationship between the two evolutionary parameters [164]. This analysis suffered from several limitations, which may explain its lack of power for this particular question. In particular, due to limited data availability, chromatin interactions and gene expression patterns were not evaluated using the same tissues and cell types. Furthermore, as only two species could be analyzed, the evolutionary tempo of chromatin interactions could not be assessed.

3.8 Future steps towards understanding the evolution of regulatory chromatin interactions

At present, genome-wide evolutionary analyses of interactions between gene promoters and distal regulatory elements are strikingly scarce. With the exception of the study cited above [153], we are not aware of any other comprehensive comparison across species. Note that comparative studies of TADs are more widespread, and have been reviewed elsewhere [134, 136]. Nevertheless, as the associations between TADs and long-range regulatory interactions are far from perfect, the latter also need to be analyzed in depth from an evolutionary perspective. Evolutionary analyses are needed for many reasons, not least of which is the potential to better understand the functional relevance of chromatin interactions. As with all other functional genomics approaches which are now commonly used to investigate the biochemical activities and structural features of the genome, the techniques designed to identify chromatin interactions likely capture many spurious, artefactual interactions. These may be false positives of the approach, or simply biological noise, which is a reality of the clearly suboptimal vertebrate genomes [12]. Evolutionary studies have the potential to filter out the noise from the biologically relevant interactions, much like they are needed to better predict functional long non-coding RNAs [165] or functional *cis*-regulatory elements [166].

As chromatin conformation capture approaches (such as PCHi-C, HiChIP, ChIA-PET) become more accessible, both technically and financially, they will likely set the stage for in depth evolutionary analyses, performed with comparable data. Ideally, such comparisons should combine multiple homologous tissues or cell types, across a wide range of species. Importantly, transcriptome sequencing data obtained from the same biological samples should provide a solid basis for joint evolutionary analyses of *cis*-regulatory interactions and of gene expression. Comparing long-range chromatin interactions along a phylogenetic tree (rather than just between two species) opens further possibilities for evolutionary analyses, in addition to better evaluating the rate of regulatory evolution at broader time scales. With multiple species, it becomes possible to infer ancestral states and to identify changes that occurred specifically in a given lineage. These changes could then be associated with gene expression changes, or to lineage-specific genome rearrangements. Thus, a broader evolutionary perspective could be obtained on the evolution and on the functional impact of long-range *cis*-regulatory interactions in vertebrates.

Chapter 4

Evolution of long-range chromatin interactions

In this chapter, I will present a comparative analysis between human and mouse, which allowed us to obtain the first insights into the evolution of promoter-enhancer chromatin interactions at a genome-wide scale. This was the main focus of Alexandre Laverré's PhD thesis. As this work is now published [153], I will only briefly discuss the main results and the main challenges that we encountered in the process.

4.1 Comparative analyses in the absence of perfectly comparable data

When we started working on this project in 2018, several technologies specifically designed to identify chromatin interactions with putative regulatory roles were already established (see also chapter 3 above). In particular, the promoter capture Hi-C (PCHi-C) technique had been used in a number of publications to describe chromatin interactions between gene promoters and other genomic regions, in human and mouse. I thus believed that the time was ripe to perform a comparative analysis of regulatory chromatin interactions. Alexandre combined all the publicly available PCHi-C data for human and mouse, taking care to include only data generated with similar experimental protocols [16, 112] and to re-process it with a common computational pipeline. However, we quickly realized that the available samples were derived from very different cell types in the two species. They also greatly differed in terms of sequencing depth and potentially even capture efficiency. We were not able to account for all the technical factors that could affect our results, although we attempted to perform downsampling analyses to control for differential detection power among samples. In particular, it was impossible to do the analysis using only comparable biological samples. Only three analyzed cell types were similar between two species (embryonic stem cells, B cells and adipocytes), but even these cell types were not perfectly comparable, as they were not obtained with identical protocols. We are thus aware that we likely under-estimated the true extent of evolutionary conservation of promoter-centered chromatin contacts. However, given that the majority of detected interactions were shared across several cell types in each species, we reasoned that the data contains a non-negligible proportion of "constitutive" contacts, and that the comparison between species is thus justified. We also confirmed previous

observations that promoter-enhancer contacts can occur in the absence of (or prior to) gene expression activation [119], which could explain the presence of these constitutive interactions.

At present, there are many more publicly available PCHi-C samples for human and mouse (more than 300 results for each of the two species in the SRA database, as of november 2023). However, publicly available PCHi-C data exists for only two other species, namely chicken and pig. The existing data is clearly not sufficient to perform evolutionary analyses at broader scales. Moreover, the issue of comparability between biological samples is still very much relevant, as these data come from a wide variety of tissues, cell types and experimental settings. I have recently initiated a collaborative project (MetaEvoChroCo, a collaboration between researchers at LBBE and LEHNA in Lyon, and INRAe and INSERM in Toulouse), whose main aim is to perform a comprehensive comparative analyses of regulatory landscapes, gene expression patterns and metabolic phenotypes in birds. In this project, we propose to generate comparable PCHi-C and RNA-seq data for multiple species, across tissues that are relevant for energy metabolism. Hopefully, if the project is funded, this data will enable us to perform an unbiased comparison of regulatory landscapes at broader evolutionary scales.

A key aspect of our study was that we generated simulated PCHi-C data, which recapitulate the genomic distribution of the regions contacted by gene promoters. This simulated data enabled us to evaluate what is expected by chance, for all our analyses. This is important because the distribution of chromatin contacts is far from uniform along the genome. In particular, gene promoters contact more often their neighboring genomic regions, and the probability of contact decreases rapidly with the genomic distance between the two regions. Our simulated data reproduces this behavior (Figure 4.1), which enables us to easily control for the effect of the distance to the contacting gene promoter.

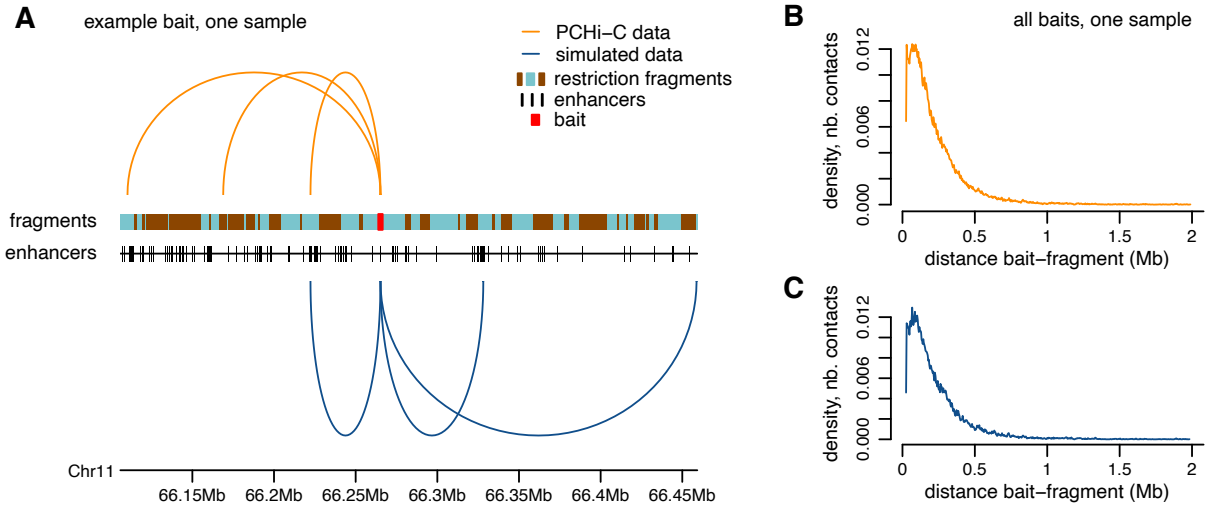

Figure 4.1: Overview of the simulated PCHi-C data that we generated as a control for our analyses. A. Example of real and simulated PCHi-C contacts for a gene promoter (baited genomic fragment). B. Distribution of the genomic distance between pairs of restriction fragments found in chromatin contact. Figure from Laverré *et al.* [153].

4.2 Promoter-centered interactions are enriched in regulatory relationships

One of the first aspects that we wanted to clarify when we started our analyses of PCHi-C data was to what extent the promoter-centered chromatin interactions detected with this technique could be considered to have regulatory roles. The publications describing the first applications of these techniques reported that the genomic regions contacted by gene promoters were enriched in histone marks that are characteristic of enhancers, for actively expressed genes, and that are typical of silenced regions, for inactive genes [16, 112]. However, a direct estimation of the overlap with predicted enhancers was not provided. At the time, genome-wide predictions of putative silencers were not yet available.

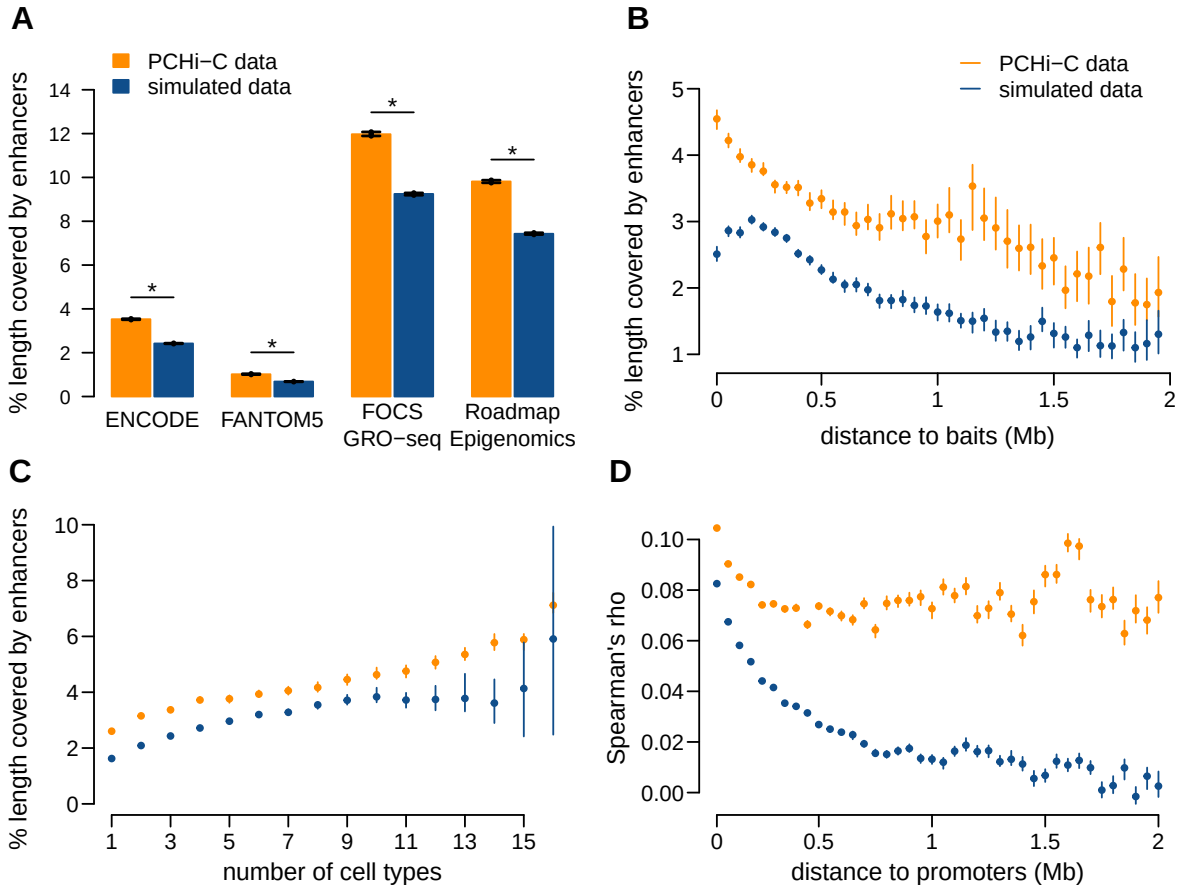

Figure 4.2: Enrichment of regulatory elements among the genomic regions contacted by gene promoters. A. Percentage of the total restriction fragment length covered by predicted enhancers, for restriction fragments contacted by gene promoters. B. Percentage of the length covered by predicted enhancers, as a function of the distance between contacted restriction fragments and baited gene promoters. C. Percentage of the length covered by predicted enhancers, as a function of the number of cell types in which interactions were scored. D. Correlations between promoter activity and enhancer activity, for pairs of promoters and enhancers in chromatin contact, as a function of the genomic distance that separates them. Figure from Laverré *et al.* [153].

We intersected the genomic regions contacted by gene promoters with the coordinates of predicted enhancers, provided by large consortia such as ENCODE, FANTOM etc. We found that promoters were often in contact with enhancers (for example, 36% of genomic regions contacted by promoters overlap with ENCODE enhancers for human PCHi-C data). However, the results were not overwhelming, especially considering that the regions "contacted" by promoters in simulated PCHi-C data also displayed high overlap fractions with predicted enhancers (27% for ENCODE). This is perhaps not surprising, considering that the ENCODE consortium has provided several hundreds of thousands of enhancer predictions, which are scattered throughout the human and mouse genomes. These predicted enhancers are likely not all biologically relevant [166].

We also wanted to evaluate to what extent the pairs of promoters and enhancers that are predicted by PCHi-C data represent genuine regulatory relationships. To do this, we evaluated the correlation between the activity of promoters and the activity of enhancers in each pair, estimated with CAGE (cap analysis of gene expression) by the FANTOM consortium [46], across several hundreds of samples. Perhaps surprisingly at first, the overall correlations were weak (Figure 4.2D). This could be explained by many factors, including the noise inherent to CAGE experiments, especially for low abundance enhancer-associated transcripts. The fact that genes tend to be regulated by multiple enhancers could also explain why individual activity correlations are low, when single enhancers are considered in conjunction with each promoter. However, encouragingly, we found that the activity correlations were higher for PCHi-C data than for simulated data, and that the contrast between the two was higher at large genomic distances (Figure 4.2D).

Thus, the interactions between promoters and enhancers detected with PCHi-C data can be considered to be enriched in genuine regulatory relationships. However, I would like to stress that they are by no means confirmed regulatory relationships. For a more stringent assessment of regulatory relationships, one could combine PCHi-C data with activity correlations between promoters and enhancers. We did not do this in Alexandre's work, but this control could be implemented in the future.

4.3 Genomic sequences contacted by promoters are evolutionarily conserved

We next verified the extent of sequence conservation for the genomic regions contacted by gene promoters. We note that in PCHi-C data interactions are detected between pairs of restriction fragments, obtained after chromatin digestion with one or several restriction enzymes (see Figure 4.1 and chapter 3). Restriction fragments have variable sizes, up to several tens of kilobases, and thus often include more than one enhancer. We thus performed the sequence conservation analysis at two levels, focusing on contacted restriction fragments, or focusing on the enhancers embedded in contacted restriction fragments (Figure 4.3). We measured the extent of sequence conservation through the percentage of aligned sequences between human and mouse, after masking exonic sequences. We found that restriction fragments contacted by gene promoters were significantly more conserved than expected based on our simulations. This was in part due to a much lower proportion of repetitive sequences in the real PCHi-C data (perhaps explained by a bias towards mappable regions in the sequencing analysis pipeline), but the effect remained visible when we analyzed separately repetitive

and non-repetitive sequence (Figure 4.3). However, surprisingly, we observed no sequence conservation difference between PCHi-C and simulated data when we specifically analyzed enhancer sequences (Figure 4.3). This was somewhat disappointing: one of my first intuitions was that intersecting on one hand enhancer predictions based on histone marks, and on the other hand chromatin contacts with gene promoters, would give us a set of elements enriched in functionally relevant enhancers. This does not seem to be the case, although again this question should be revisited with additional, more stringent filters on PCHi-C data (*e.g.*, selecting contacts detected in multiple samples). Thus, the restriction fragments contacted by gene promoters are more conserved than expected by chance not because they overlap with better conserved enhancers, but perhaps because they overlap with more enhancers.

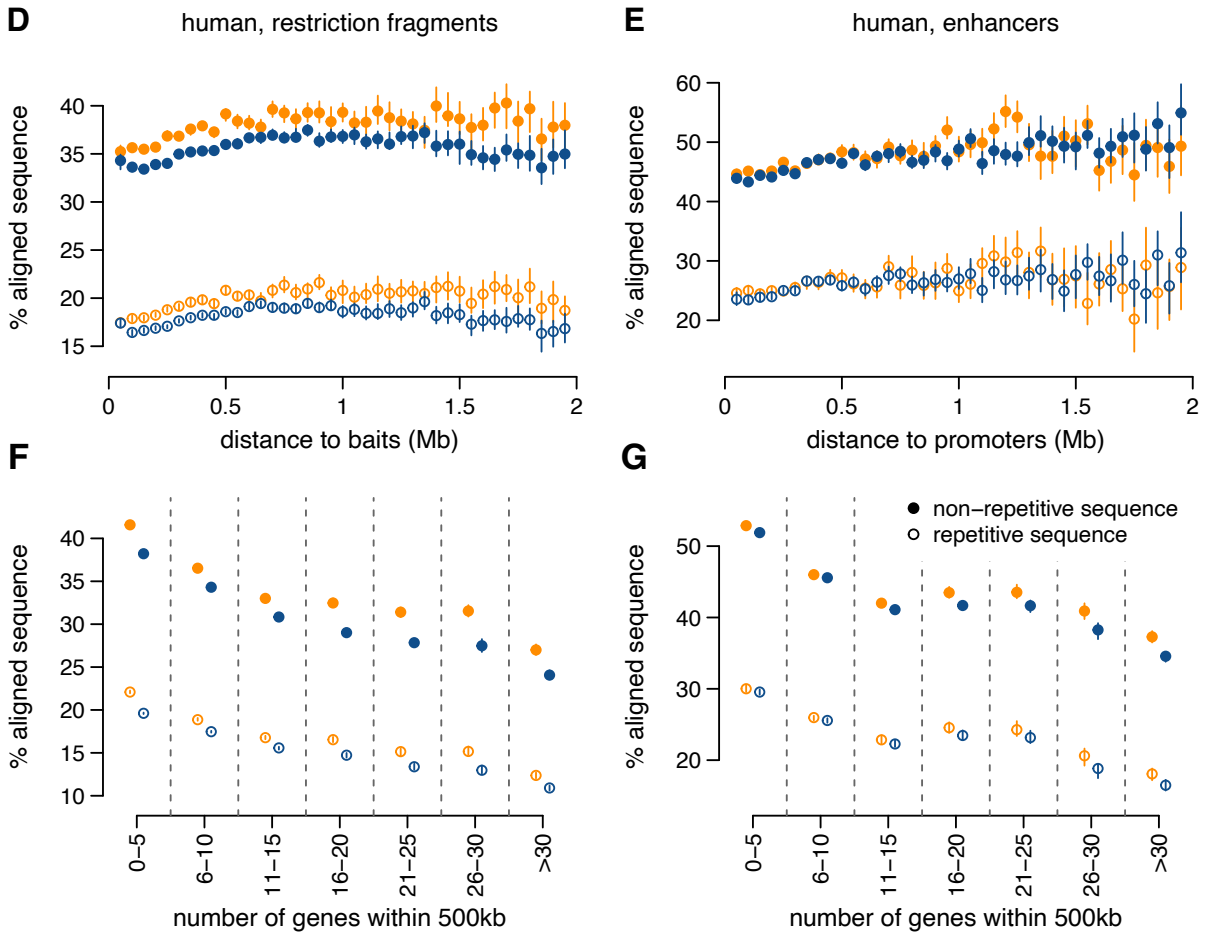

Figure 4.3: Sequence conservation of genomic regions and of enhancers contacted by gene promoters. Left: sequence conservation for restriction fragments contacted by gene promoters. Right: sequence conservation for predicted enhancers (ENCODE dataset) found in the fragments contacted by gene promoters. Solid dots: non-repetitive sequences; empty dots: repetitive sequences. Figure from Laverré *et al.* [153].

We also uncovered a strong negative association between sequence conservation (for both contacted restriction fragments and contacted enhancers) and gene density in the neighboring regions (Figure 4.3). Specifically, enhancers found in gene-poor regions are more conserved during evolution. This result is consistent with previous reports of an enrichment of gene deserts (large intergenic regions, up to several megabases long) around highly conserved developmental genes and transcription factors [121].

4.4 Promoter-centered chromatin contacts are evolutionarily conserved

After these initial exploratory analyses, we arrived at one of our main questions: are chromatin contacts between promoters and other genomic regions conserved during evolution? We focused on chromatin contacts for which both genomic regions were alignable between human and mouse, without ambiguity. We could thus predict putative orthologous contacts, and we said that a chromatin contact detected in one species was conserved in the other species if the orthologous contact was detected in its corresponding PCHi-C data.

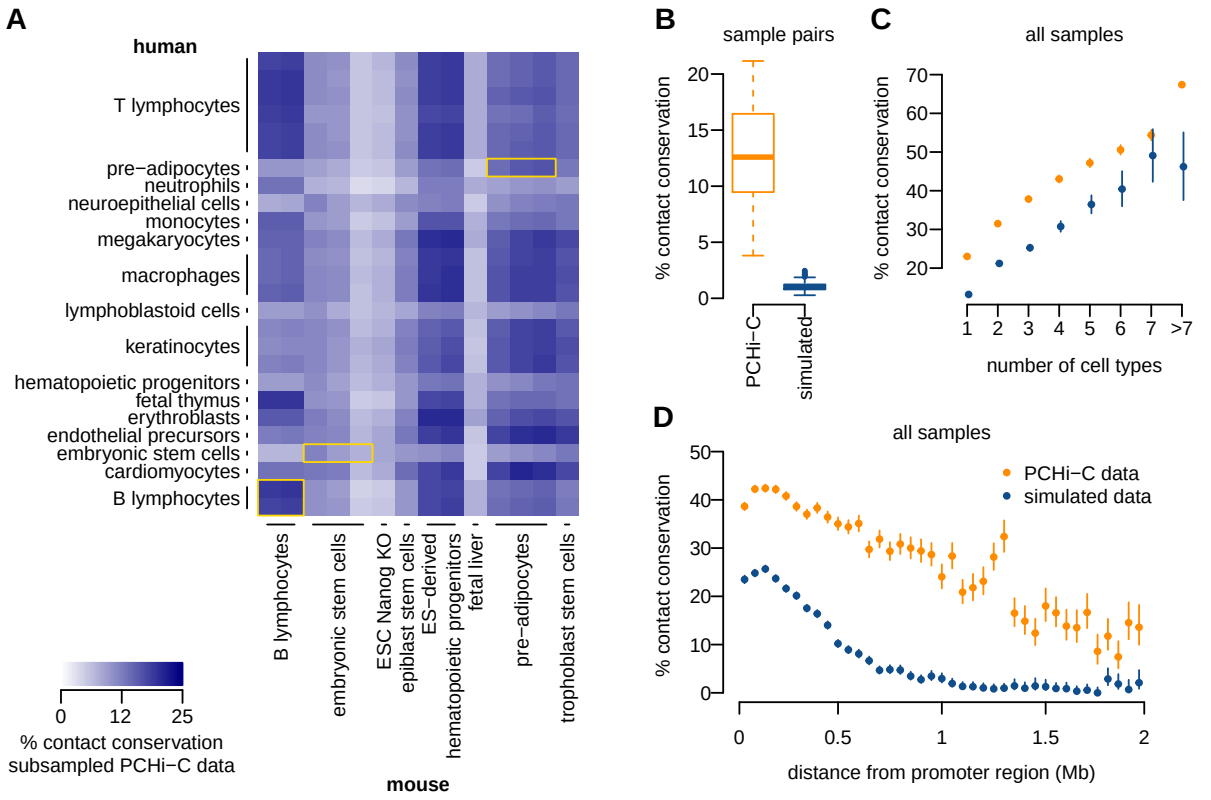

Figure 4.4: Conservation of promoter-centered chromatin contacts between human and mouse. A. Heatmap representation of the percentage of contact conservation between pairs of human and mouse samples. B. Boxplot representing the distribution of the percentages of contact conservation between pairs of samples, for real and simulated PCHi-C data. C. Relationship between the percentage of contact conservation and the number of cell types in which contacts were detected. D. Dependency between the percentage of contact conservation and the genomic distance separating the two contacting regions. All samples are combined for the analyses presented in C and D. Figure from Laverré *et al.* [153].

For this analysis, the excess of conservation in the real PCHi-C data compared to the simulated PCHi-C data was striking (Figure 4.4). This was particularly true for interactions detected in multiple cell types, as expected biologically (housekeeping regulatory relationships may be under stronger functional constraint) and technically (given the disparity in the cell type collections analyzed for human and mouse). The excess of conservation compared to simulated data was particularly strong for interactions between genomic regions separated

by large distances on the linear genome (Figure 4.4D). We initially thought that the peaks of conservation found at distances just below of 1.5 Mb were artefactual, and we did our best to remove them with additional controls. After further manual inspection, it turned out that these highly conserved contacts were very much real and involved *Hox* genes, for which important long-range regulatory contacts were previously reported [14].

4.5 Regulatory evolution and gene expression evolution

Finally, in this project we wanted to re-evaluate the relationship between regulatory evolution and gene expression evolution. This question had been previously addressed, but (to our knowledge) never by defining regulatory relationships with high resolution chromatin contact data. We evaluated regulatory landscape conservation at multiple levels: by computing the average sequence conservation of contacted enhancers for each gene; by determining whether the contacted enhancers were in conserved synteny between human and mouse; and by directly estimating the fraction of conserved contacts between the two species, as described above. To evaluate gene expression evolution, we compared relative expression profiles, obtained from a publicly available transcriptome collection which encompasses comparable organs and developmental stages [76]. Specifically, for each gene, we obtained a relative expression profile by dividing the expression levels (evaluated as transcript per million, or TPM, values) by the maximum observed value among samples. With this procedure, we obtained expression profiles that are more comparable between species than TPMs, which are affected by various species-specific factors (gene annotation differences for example). Our measure of expression conservation can thus capture changes in tissue specificity or developmental stage specificity for example, but we do not claim to detect quantitative differences in expression levels between species.

Our analyses confirmed several previously reported results, such as the positive correlation between the number of contacted enhancers, on one hand, and the gene expression level and the gene expression breadth, on the other hand (Figure 4.5) [164]. However, we were not able to confirm a previously reported association between the number of enhancers attributed to each gene and the extent of expression conservation [164]. This discrepancy could be explained by the fact that we used a different measure of gene expression conservation, but perhaps also by the way in which enhancers are attributed to genes. In previous regulatory evolution studies, enhancers were assumed to control the closest neighboring genes, within a given maximal genomic distance [164]. With PCHi-C data, there is evidence that a considerable proportion (up to one third) of inferred promoter-enhancer relationships bypass the closest promoter [122]. We also discuss this possible source of discrepancy below (section 4.6).

Overall, our analyses uncovered significant, positive associations between the extent of gene expression conservation and the extent of regulatory landscape conservation. These results are intuitively expected, but previous joint evolutionary analyses of gene expression and regulatory mechanisms [164] had only uncovered very mild associations between the two factors. Multiple explanations could be proposed to explain the discrepancies with previous work, including the types of data that were analyzed and the methodology used to evaluate gene expression conservation [153]. Irrespective of these considerations, we interpret these results as a positive motivation to continue using chromatin conformation capture data for the prediction of regulatory relationships.

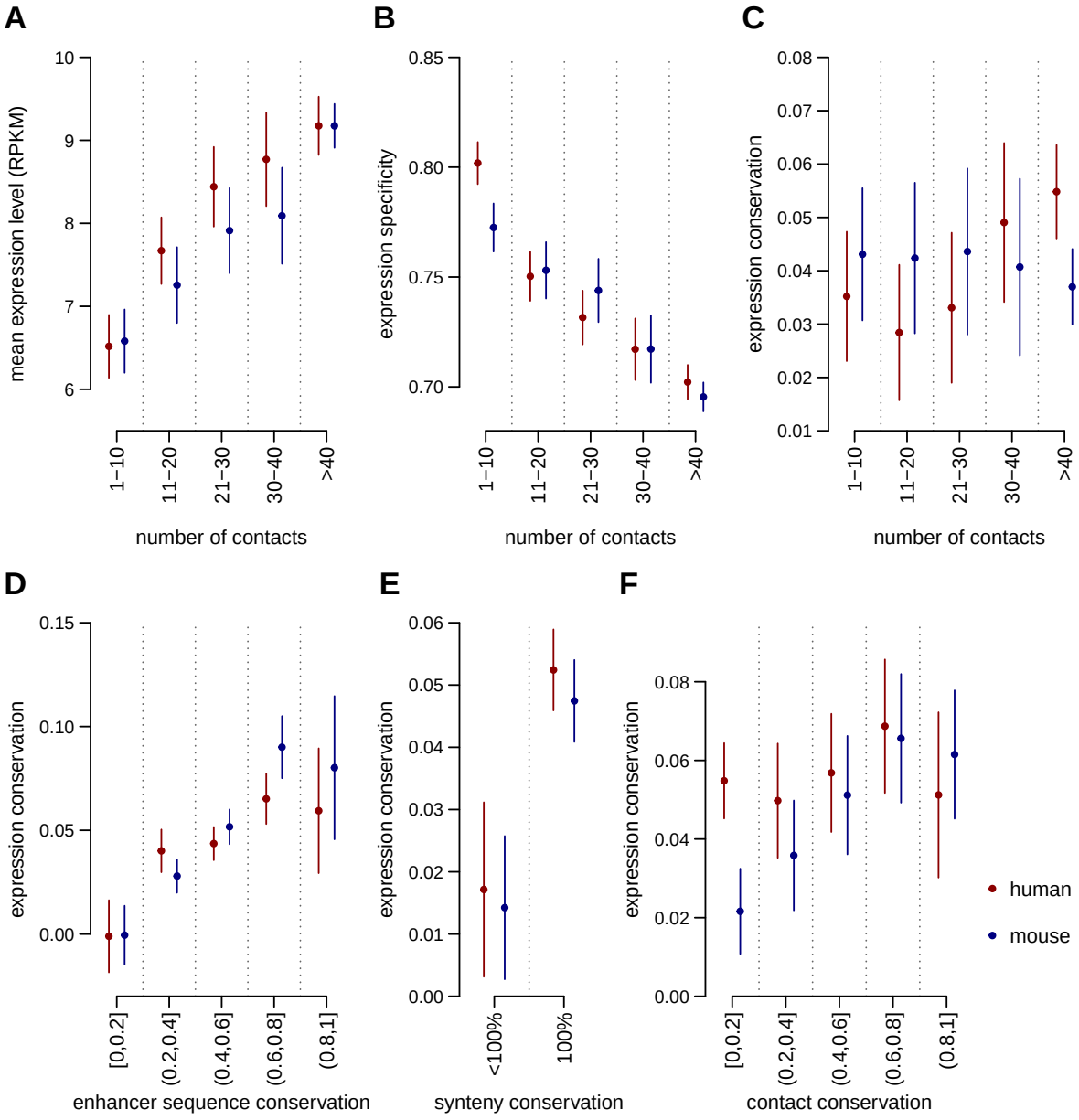

Figure 4.5: Correlation between regulatory evolution and expression evolution. A. Highly expressed genes contact more enhancers than weakly-expressed genes. B. Broadly-expressed genes contact more enhancers than narrowly-expressed genes. C. Expression conservation is not correlated with the number of enhancers contacted by genes. D. Expression conservation is positively correlated with the sequence conservation of contacted enhancers. E. Expression conservation is positively correlated with synteny conservation between promoters and contacted enhancers. F. Expression conservation is positively correlated with promoter-enhancer contact conservation in mouse.

4.6 Defining target genes for regulatory elements with PCHi-C data

As mentioned before, with our data, we were unable to uncover a previously reported association between the number of enhancers predicted to regulate each gene and the extent of gene expression conservation [153, 164]. One of the possible reasons for this discrepancy is the way that enhancers were attributed to gene. In the absence of chromatin conformation data, enhancers are traditionally assumed to regulate the closest neighboring genes, within a certain maximum genomic distance. This definition of "regulatory domains" is at the basis of GREAT [167], a widely-used method for the inference of functional associations for predicted regulatory elements. However, analyses of PCHi-C data largely invalidate this definition of regulatory domains. As a follow-up on Alexandre Laverré's work, we exploited the PCHi-C data collection for human and mouse to propose a new tool for the inference of functional associations for regulatory elements. We named this tool GOntact, as a contraction between Gene Ontology and contacts. A preprint is already available [168].We are currently working to implement this tool as a webserver, which will be available at http://gontact.univ-lyon1.fr.

Chapter 5

Genome rearrangements and the evolution of regulatory landscapes

5.1 Biased distributions of phenotypic effects for observed rearrangements

The presence of long-range regulatory interactions between promoters and enhancers raises the question of their robustness to genomic rearrangements, and of the functional constraints that they may impose on large-scale genome evolution. The presence of conserved synteny at large evolutionary scales between genes and conserved non-coding sequences was interpreted as evidence for conservation of long-range regulatory relationships, even before such interactions could be predicted with chromatin conformation data [154–157]. This reasoning is confirmed by numerous articles, which report deleterious phenotypic effects for genomic rearrangements that alter chromatin conformations [138, 140, 161, 169]. However, it is important to note that these considerations may not be applicable at the genome-wide level, as they are based on analyses of highly conserved genes, including key developmental transcription factors. Most likely, the fitness effects of genomic rearrangements that perturb regulatory landscapes vary depending on the functional importance of the genes that are affected, and on the robustness of their regulatory mechanisms. In our evolutionary comparison of chromatin contacts, we also observed that promoter-enhancer relationships were conserved in synteny significantly more often than expected [153]. This observed average effect may also be due to the presence of strong constraints on a small subset of regulatory interactions.

Interestingly, analyses of *Drosophila* laboratory strains with highly rearranged genomes did not reveal any detectable effect on gene expression, beyond cases where gene bodies were directly touched by the rearrangements [170]. A biased representation of possible fitness effects of genome rearrangements may also explain this result: the laboratory strains under investigation were all viable, which indicates that genome rearrangements with highly deleterious effects were excluded from this pool of genotypes. This biased phenotypic representation is the opposite of the one observed for case studies of genomic rearrangements involved in human diseases, where highly deleterious rearrangements are likely favored [138].

While discussing these observations during Alexandre Laverré's PhD project, we realized

that our analysis was also biased in terms of phenotypic effects of genome rearrangements, although we based it on all analyzable genes. Indeed, these analyzable genes are those that are orthologous between human and mouse. We were not able to analyze cases where a gene was lost in one of the two lineages, nor did we analyze lineage-specific gene duplications. By excluding gene losses, we have perhaps discarded a subset of genomic rearrangements with extreme effects on regulatory landscapes. We can imagine a scenario where a large-scale inversion (for example) separates a gene from its most important distant regulatory elements. If, following this drastic change in its regulatory landscape, the gene is no longer expressed in the tissues where it performs its function, the gene may become pseudogenized following a relaxation of purifying selection. Indeed, sequence-altering mutations on a gene that is already not correctly expressed are unlikely to have any further deleterious effects on the organism. The gene could thus become pseudogenized and lost with evolutionary time. Evidently, this scenario only applies to dispensable genes, otherwise the original rearrangement would have highly deleterious phenotypic consequences.

I decided to follow up on this idea, and proposed Master internships to investigate the association between genomic rearrangements and gene losses. This topic was addressed by Victor Lefebvre during his M1 internship and by Thomas Lahaie during his M2 internship, in 2023. They each addressed a different aspect of this research question, as briefly described below.

5.2 Genomic rearrangements around the *Xist/Lnx3* locus

One of my motivations for pursuing this idea was a peculiar situation that I observed around *Xist*, the long non-coding RNA responsible for X chromosome inactivation in placental mammals (it does seem that lncRNAs follow me, even though I tried to move away from them). In 2006, Laurent Duret and collaborators showed that *Xist* originated in placental mammals through the pseudogenization of a protein-coding gene, named *Lnx3* [79]. One of the circumstances that helped him uncover this exciting origination story was the remarkable conservation of synteny around *Xist* (Figure 5.1).

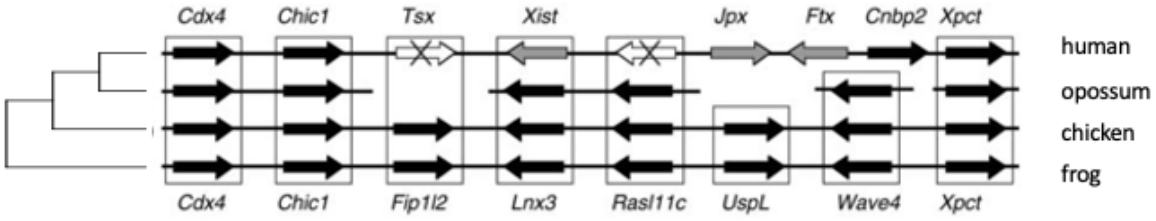

Figure 5.1: Genomic organization around the *Xist/Lnx3* locus, in human, opossum, chicken and frog. Figure adapted from Duret *et al.* [79].

The *Xist/Lnx3* locus is flanked by the highly conserved *Cdx4*, *Chic1*, *Xpct* genes, which are maintained in synteny across tetrapod species (Figure 5.1). Among these genes, *Cdx4* is a developmental transcription factor involved in the patterning of the anterior/posterior body axis. I came across *Cdx4* in an unrelated project, in which I collaborated with Isabel Guerreiro (a post-doc in Denis Duboule's laboratory in Geneva) with the aim of studying the genetic basis of axial elongation in snakes. In this project, Isabel and I discovered that *Cdx4* is in fact

part of a very ancient tandem duplication, which likely originated in the ancestor of tetrapods (and which is completely unrelated with the whole-genome duplications that occurred in the vertebrate ancestor, which generated a different set of *Cdx* paralogues). Strikingly, we found that different lineages lost different copies of the tandem duplicates: placental mammals and snakes both have just a single copy of *Cdx4*, but they kept different paralogous copies. In other species, such as birds and marsupial mammals, the two copies are preserved. The functional implications of this curious evolutionary event are still under investigation (Guerreiro and Necsulea, manuscript in preparation). However, this finding motivated me to look beyond the boundaries of Laurent's figure from 2006, which stopped at *Cdx4* (Figure 5.1 and Figure 5.2).

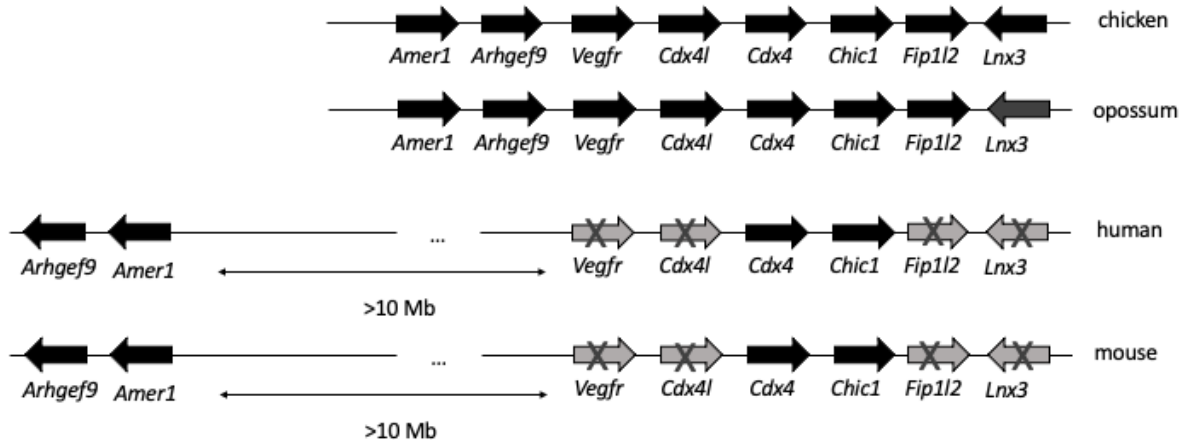

Figure 5.2: A broader view of the genomic organization around the *Xist/Lnx3* locus, in human, mouse, opossum and chicken. Figure adapted from Victor Lefebvre's internship report.

If we look upstream of *Cdx4*, we notice that in addition to *Cdx4l* (the ancient paralogue of *Cdx4*), placental mammals also lost *Vegfr* (Figure 5.2). Moreover, the two next closest genes (*Amer1* and *Arhgef9*) were relocated at a distance of more than 10 Mb from *Cdx4*, and are now in the opposite transcriptional orientation (Figure 5.2). We can be confident of the direction of the evolutionary change because chicken and opossum both show the same genomic organization. This suggests that (at least) one large inversion occured in the immediate vicinity of *Cdx4* in placental mammals. This inversion may have directly perturbed the sequences of *Cdx4l* and/or *Vegfr*, for which we cannot find traces in placental mammal genomes.

Thus, at least one large genomic rearrangement occurred in close proximity to the *Xist/Lnx3 locus* in the ancestor of placental mammals, that is, approximately (on a geological scale) at the same time as the pseudogenization of *Lnx3*, *Fip1l2*, *Cdx4l* and *Vegfr*. The sequences of the last two genes may have been directly affected the genomic rearrangement. However, the gene bodies of *Lnx3* and *Fip1l2* were clearly not part of the synteny breakpoint. For these genes, we cannot exclude the hypothesis that the genomic rearrangement affected their regulatory chromatin interactions.

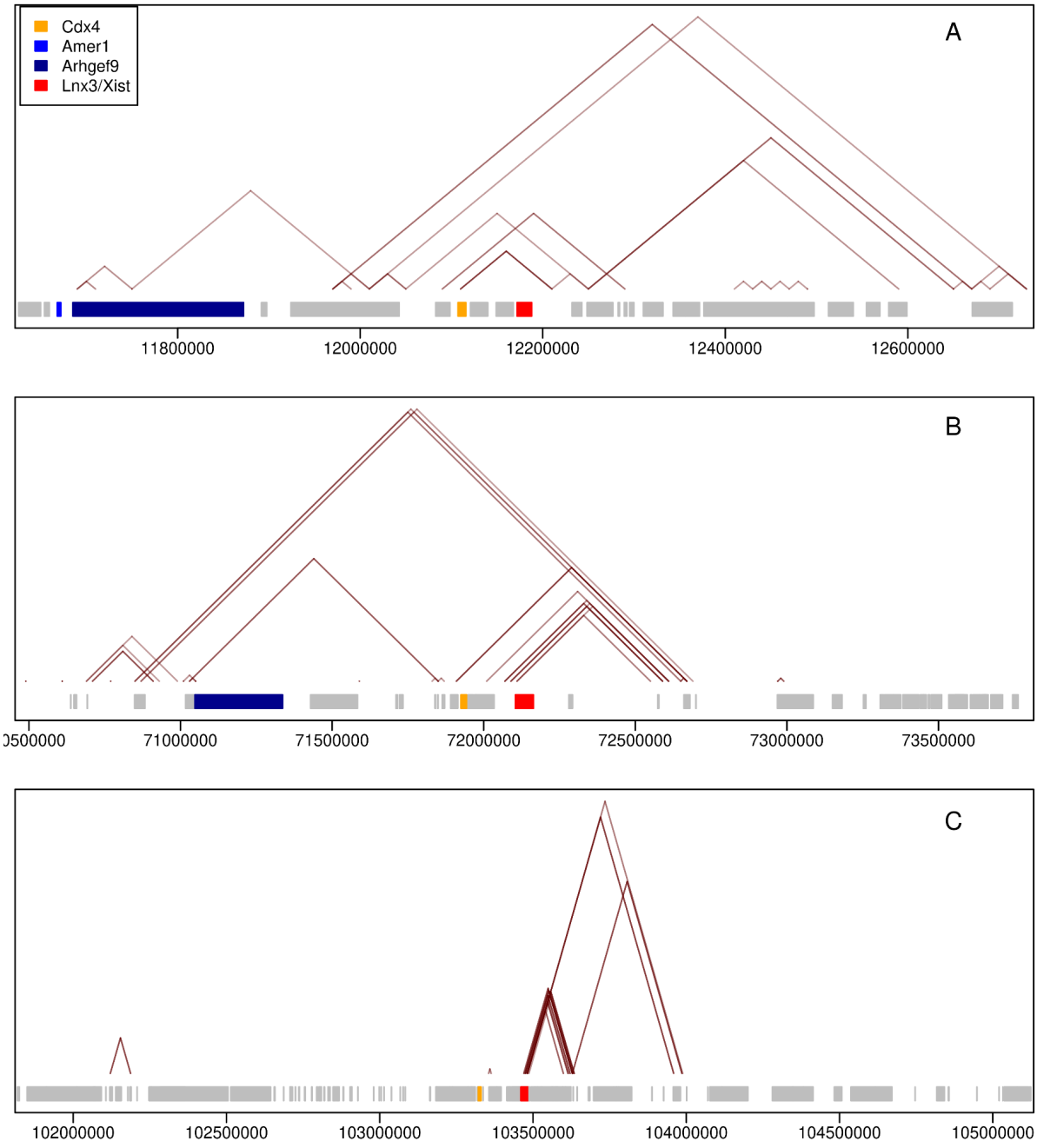

Figure 5.3: Chromatin interactions detected with FitHic2 using Hi-C data for chicken (A), opossum (B), and detected using PCHi-C data for mouse (C). The *Cdx4*, *Lnx3/Xist*, *Amer1*, *Arhgef9* genes are displayed in color (the latter two genes are not visible in mouse). Figure adapted from Victor Lefebvre's internship report.

5.3 Evolutionary change in chromatin interactions around the *Xist/Lnx3* locus

During his M1 internship, Victor Lefebvre analyzed the distribution of chromatin interactions around the *Xist/Lnx3*, in opossum and chicken. He used Hi-C data from muscle, which was publicly available for both species. He detected discrete chromatin interactions with FitHic2 [171] and compared these interactions with those observed in PCHi-C data in mouse [153]. Although the Hi-C data is underpowered for the detection of discrete chromatin interactions compared to PCHi-C data, Victor was still able to uncover some interesting differences between species in the structure of chromatin contacts (Figure 5.3).

In agreement with previous reports, according to which *Xist* is found close to a TAD boundary in placental mammals [125], PCHi-C data show that the *Xist* promoter preferentially contacts elements upstream to itself (on the right-hand side of the image in Figure 5.3). However, in chicken *Lnx3* does not appear to be situated on the boundary of a TAD. We can observe genomic interactions that start upstream of *Lnx3* (right-hand side of the image) and that end in the region upstream of *Cdx4* (left-hand side of the image) in both chicken and opossum. These interactions could have been affected by the genomic rearrangements; in any case, they are not observed in mouse.

This analysis does not allow us to conclude on the effect of this large-scale inversion on the regulatory landscapes of *Lnx3/Xist*, not least because the Hi-C data does not offer enough power to detect fine-scale interactions. We plan to explore this aspect further, using PCHi-C data for the chicken (data kindly provided ahead of publication by H. Acloque, S. Foissac and S. Djebali, INRAe and INSERM). Victor will continue this project as part of his M2 internship.

5.4 Gene losses are enriched close to synteny breakpoints

A causal relationship between the neighboring rearrangement and the loss of *Lnx3* can only remain speculative, even if we had abundant chromatin conformation data. To further test the original hypothesis, we turned to genome-wide analyses, as always in computational evolutionary biology. As part of his Master 2 internship in 2023, Thomas Lahaie investigated the potential association between gene losses and synteny breakpoints. He focused on gene losses that occurred in the ancestor of placental mammals, which include *Lnx3*. Through comparative analyes of gene families, Thomas identified approximately 120 protein-coding genes which were lost in the ancestor of placental mammals, but which were kept as a single copy in the chicken genome. We call these genes "placental-lost".

We first asked whether these placental-lost genes were spatially clustered in the chicken genome. Indeed, assuming that genomic rearrangements had drastically perturbed the regulatory landscapes of the lost genes, a genomic clustering is expected, because neighboring genes often shared regulatory interactions. To do this, for each placental-lost gene, we computed its distance to the nearest other placental-lost gene. If no other placental-lost gene was found on the same chromosome, we set the distance to 1e9 (a value close to the chicken genome size), to denote the absence of a close neighbor. We then compared the distribution of these distances between placental-lost neighbors to the random expectation, obtained by randomly re-shuffling the labels ("placental-lost" or "other") of genes. Our results confirm this

intuition: placental-lost genes are found more closely together than expected in the chicken genome (figure 5.4). We note that this genomic clustering could also be explained by alternative hypotheses, such as co-localization of functionally enriched genes, which might be concomittantly lost following a phenotypic change.

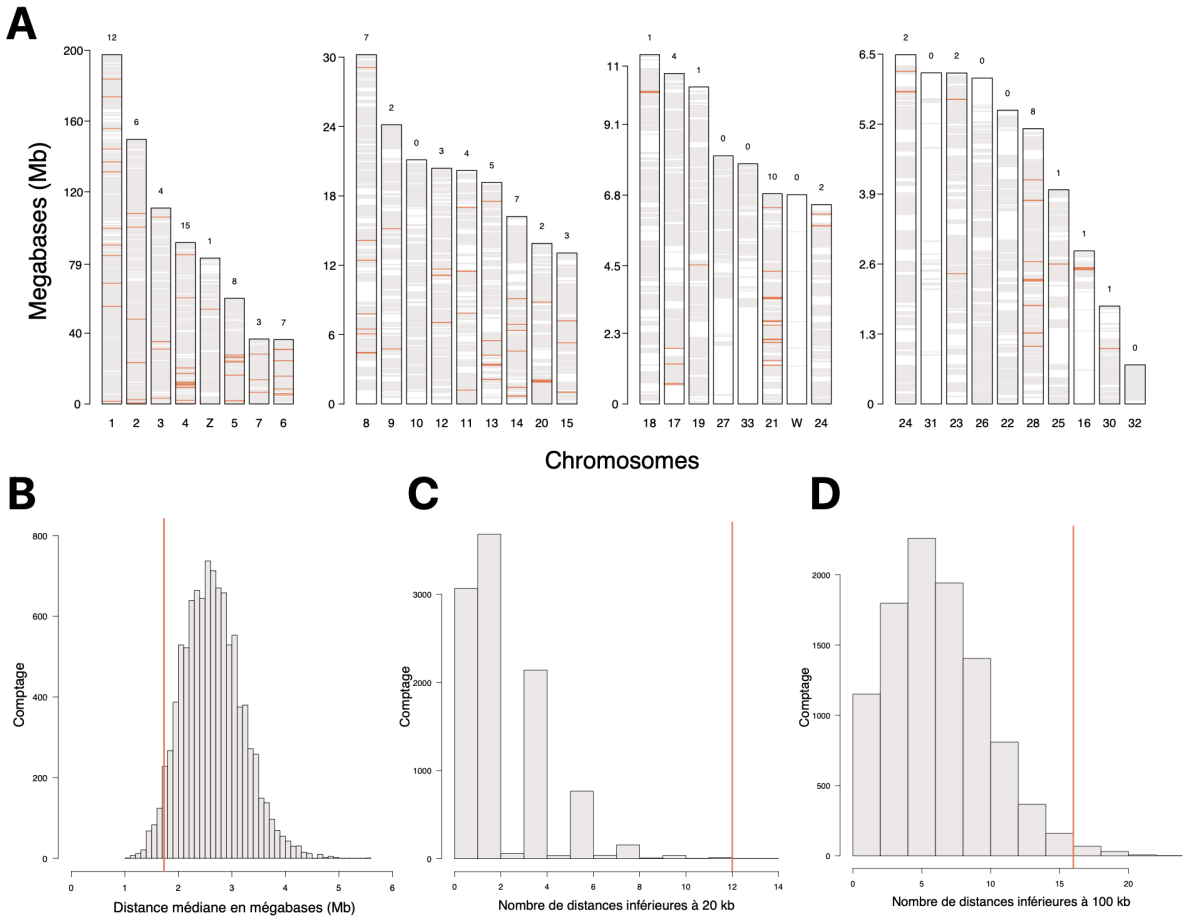

Figure 5.4: Genomic clustering of placental-lost genes in the chicken genome. A. Position of placental-lost genes on the chicken chromosomes. B. Distribution of the median distance between neighboring pairs of placental-lost genes. C. Number of placental-lost gene neighbors separated by at most 20 kb. D. Number of placental-lost gene neighbors separated by at most 100 kb. B,C,D. The gray histogram represents the random expectation obtained through simulations. The red vertical bar represents the observed value.

To consolidate this result, Thomas identified synteny breakpoints by comparing the positions of orthologous genes between chicken, opossum, mouse and human. He again focused on synteny breakpoints that were found between chicken and placental mammals, but not between chicken and opossum. We called these synteny breakpoints "placental-breakpoints". We observed that placental-lost genes were found significantly closer than expected by chance to genomic rearrangement breakpoints (figure 5.5). Interestingly, overlaps (distances of 0 kb) between placental-lost genes and placental-breakpoints were not more frequent than expected. However, short distances (excluding overlaps but below 50 kb) were over-represented in the observed data compared to the randomizations (figure 5.5) This observation conforts

our original hypothesis. Indeed, we do not aim to uncover rearrangements which directly affected gene sequences, but rearrangements which perturbed regulatory landscapes while leaving gene bodies intact.

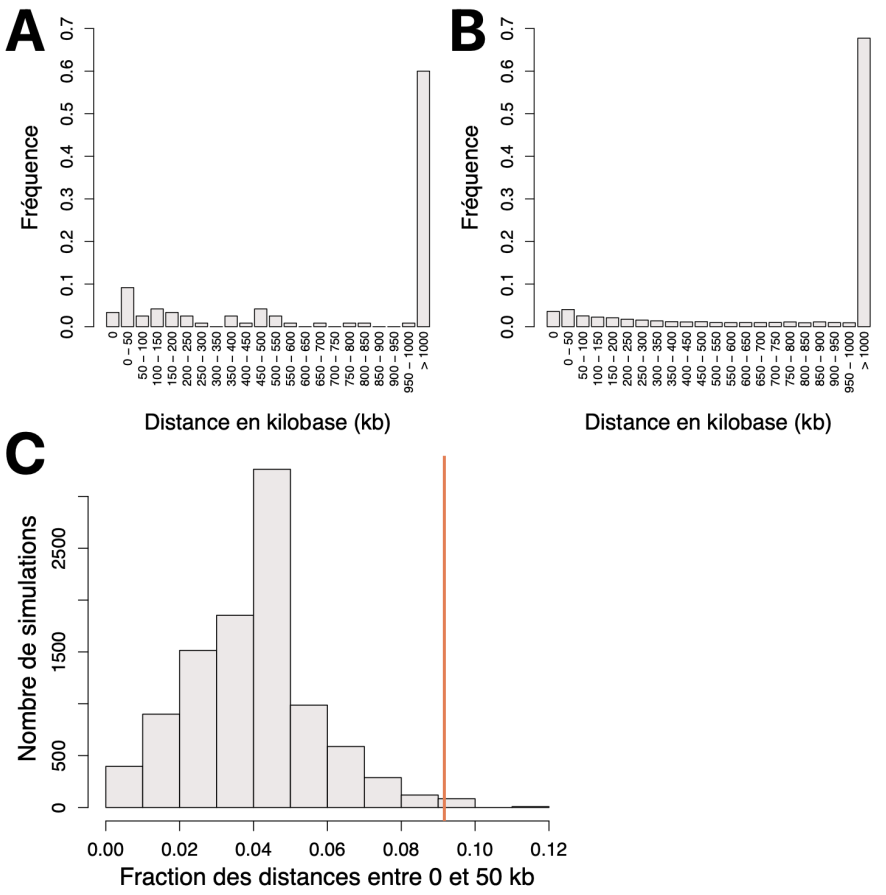

Figure 5.5: Proximity between placental-lost genes and placental-breakpoints, in the chicken genome. A. Distribution of the distance between placental-lost genes and placental-breakpoints. B. Randomly expected distribution of the distance between placental-lost genes and placental-breakpoints, obtained through permutations. C. Fraction of distances between placental-lost genes and placental-breakpoints that are below 50 kb.

These results are clearly very preliminary, but are encouraging enough to motivate further analyses. Victor Lefebvre will continue some of these research directions as part of his M2 internship in 2024. He will also analyze the divergence of regulatory chromatin interactions following another type of genome rearrangement, namely gene duplications. Indeed, at present very little is known about the evolution of regulatory landscapes following gene duplications. We will thus be able to explore other aspects of the relationship between *cis*-regulatory landscape evolution and genome rearrangements.

Chapter 6

Conclusion

I had originally planned to write my "habilitation à diriger les recherches" manuscript in 2019. I unfortunately missed that self-imposed deadline and then had to postpone it for several years due to some difficulties during the Covid period. Although it was a difficult exercise, writing this manuscript was an opportunity to revise my priorities, among the numerous research projects that I started but had not been able to finish. It was also very motivating. Writing about them allowed me to see some old projects in a new light, and to some extent to reconsider my move away from some research directions.

Although I have tried my best to present most of the projects I participate in under a common title, I think I have not misled anyone into believing that I carry out a unified research program. I am perhaps too easily distracted, or attracted, by new and exciting ideas. I think this is also part of why I am happy to be a researcher during a period marked by so many technological innovations. I consider myself very privileged to be able to do this type of work. With this sense of privilege also comes a feeling of guilt, of not doing more easily applicable research that might perhaps have at least a minuscule contribution to solving some of the problems of our society. Although I am mainly interested in evolutionary biologiy, I am thus always eager to collaborate with medical researchers. I am also keen to contribute to the training of the next generation of researchers, by teaching in various Master programs and supervising students during their internships.

Chapter 7

Curriculum vitae

Laboratoire de Biométrie et Biologie Évolutive — UMR CNRS 5558
Université Claude Bernard - Lyon 1 — (+33) 07 84 59 69 20
Lyon, France — anamaria.necsulea@univ-lyon1.fr

Date and place of birth: August 19th 1981, Bucharest, Romania. Romanian nationality.

Education

PhD in Bioinformatics
University Claude Bernard Lyon 1, Lyon, France — *October 2005 - June 2008*

M.S. in Ecology, Evolution and Biometry
University Claude Bernard Lyon 1, Lyon, France — *October 2004 - July 2005*

Engineering degree in Bioinformatics and Mathematical Modeling
National Institute of Applied Science (INSA), Lyon, France — *September 2000 - July 2005*

Research positions

Researcher (Chargée de recherche CNRS)
Biometry & Evolutionary Biology Laboratory, University Lyon 1, France — *October 2016 - present*

Researcher
Laboratory of Developmental Genomics, EPFL, Switzerland — *November 2012 - September 2016*

Post-doctoral researcher
Center for Integrative Genomics, University of Lausanne, Switzerland — *April 2009 - October 2012*

Post-doctoral researcher
Biometry & Evolutionary Biology Laboratory, University Lyon 1, France — *July 2008 - March 2009*

Visiting scientist
Department of Biology and Biochemistry, University of Bath, United Kingdom *February 2009*

Doctoral researcher
Biometry & Evolutionary Biology Laboratory, University Lyon 1, France *October 2005 - June 2008*

Grants and fellowships

March 2023, Starting grant for collaborative projects, Fédération de Recherche Bio-Environnement Santé, Université de Lyon (with Sandrine Moja, Université Jean Monnet Saint Étienne, and Christine Oger, Pôle Rhône-Alpes de Bioinformatique).

January 2018 - July 2022, ANR "Jeunes Chercheuses et Jeunes Chercheurs" grant "LncEvoSys".

March 2017, Starting grant for collaborative projects, Fédération de Recherche Bio-Environnement Santé, Université de Lyon (with Florence Hommais, University Claude Bernard - Lyon 1).

January 2017 - December 2017 Impulsion grant, University of Lyon.

November 2012 - October 2015 Ambizione fellowship, Swiss National Science Foundation.

September 2012 (declined) SystemsX Transition Post-Doc fellowship.

August 2009 - July 2012 FEBS long-term post-doctoral fellowship.

February 2009 International Conflict Research Institute visiting fellowship, University of Bath.

October 2005 - September 2008 PhD scholarship, French Ministry for Higher Education and Research.

September 2002 - September 2005 Eiffel excellence scholarship, French Ministry for Foreign Affairs.

Awards

Young researcher excellence prize, Faculty of Biology and Medicine, University of Lausanne (2013).

Swiss Institute of Bioinformatics Young Bioinformatician Award (2013).

Post-doctoral researcher travel award, Society for Molecular Biology and Evolution (2010).

National mathematics competitions, Romania (1994-2000).

Teaching

Methods for the analysis of transcriptomics data, Bioinformatics Master program, University Claude-Bernard Lyon 1, 2019-2023.

Methods for the analysis of genomic and transcriptomics data, Biodiversity, Ecology and Evolution Master program, University Claude-Bernard Lyon 1, 2023.

Introduction to bioinformatics for the analysis of next-generation sequencing data, CNRS Formation (2018-2023).

Molecular evolution course, 2nd year Biology Bachelor, University of Lausanne, 2015-2016.

Supervised PhD students

Alexandre Laverré (co-supervised with Eric Tannier), University of Lyon, 2018-2022.

Iris Finci (co-supervised with Henrik Kaessmann), University of Lausanne, 2010-2014.

Adem Bilican (co-supervised with Henrik Kaessmann), University of Lausanne, 2010-2014.

Academic service and scientific animation

Organization of popular science animations in the framework of the "Fête de la Science" with the CNRS and UCBL, Lyon (2017-2023).

Reviewer for Molecular Biology and Evolution, Genome Biology and Evolution, BMC Genomics, PLoS journals, Nature, Nature Communications, Scientific Reports, Trends in Genetics, *etc.*

Reviewer for the Swiss National Science Foundation (2019-2023).

Symposium organisation on the topic “Non-coding RNAs in development and evolution”, Annual Meeting of the Society for Molecular Biology and Evolution, Chicago, July 2013.

Symposium organisation on the topic “Adaptive and non-adaptive evolution of gene expression and regulation”, Annual Meeting of the Society for Molecular Biology and Evolution, Vienna, July 2015.

Undergraduate student mentoring at the Annual Meeting of the Society for Molecular Biology and Evolution, 2012 (Dublin) and 2013 (Chicago).

Publications

* equal contribution

1. Teoli J, Fablet M, Bardel C, Necşulea A, Labalme A, Lejeune H, Lemaitre JF, Gueyffier F, Sanlaville S, Vieira C, Marais GAB, Plotton I. Transposable element expression with varia-

tion in sex chromosome number supports a toxic Y effect on human longevity. *submitted*. https://doi.org/10.1101/2023.08.03.550779.

2. Bénitière F, Necşulea A, Duret L. Random genetic drift sets an upper limit on mRNA splicing accuracy in metazoans. *submitted*. https://doi.org/10.1101/2022.12.09.519597. Recommended by PCI Evol Biol.

3. Laverré A, Tannier E, Veber P, Necşulea A. GOntact: using chromatin contacts to infer Gene Ontology enrichments for *cis*-regulatory elements. *submitted*. https://doi.org/10.1101/2022.06.13.495495.

4. Necşulea A, Veber P, Boldanova T, Ng CKY, Wieland S, Heim MH. LncRNA analyses reveal increased levels of non-coding centromeric transcripts in hepatocellular carcinoma. *submitted*. https://doi.org/10.1101/2021.03.03.433778.

5. Laverré A, Tannier E, Necşulea A (2022) Long-range promoter–enhancer contacts are conserved during evolution and contribute to gene expression robustness. **Genome Research**, 32, 280-296.

6. Darbellay F, Necşulea A (2020). Comparative transcriptomics analyses across species, organs, and developmental stages reveal functionally constrained lncRNAs. **Molecular Biology and Evolution**, 37, 240-259.

7. Necşulea A (2020). Phylogenomics and genome annotation. Book chapter in: Phylogenetics in the Genomic Era (open access book). https://hal.science/hal-02535669/document.

8. Saudemont B, Popa A, Parmley JL, Rocher V, Blugeon C, Necşulea A, Meyer E, Duret L (2017) The fitness cost of mis-splicing is the main determinant of alternative splicing patterns. **Genome Biology**, 18, 1-15.

9. Boldanova T, Suslov A, Heim MH*, Necşulea A* (2017) Transcriptional response to hepatitis C virus infection and interferon alpha treatment in the human liver. **EMBO Molecular Medicine**, 9, 816-834.

10. Schep R, Necşulea A, Rodriguez-Carballo E, Guerreiro I, Andrey G, Nguyen Huynh TA, Marcet V, Zákány J, Duboule D, Beccari L (2016) Control of *Hoxd* gene transcription in the mammary bud by hijacking a pre-existing regulatory landscape. **Proceedings of the National Academy of Sciences USA**, 113, E7720-E7729.

11. Amândio AR, Necşulea A, Joye E, Duboule D (2016) *Hotair* is dispensable for mouse development. **PLoS Genetics**, 12, e1006232.

12. Beccari L, Yakushiji-Kaminatsui N, Woltering JM, Necşulea A, Lonfat N, Rodriguez-Carballo E, Mascrez B, Yamamoto S, Kuroiwa A, Duboule D (2016) HOX13 proteins control the regulatory switch between TADs at the *HoxD* locus. **Genes and Development**, 30, 1172-1186.

13. Wiberg RA, Halligan DL, Ness RW, Necşulea A, Kaessmann H, Keightley PD (2015) Assessing recent selection and functionality at long non-coding RNA loci in the mouse genome. **Genome Biology and Evolution**, 7, 2432-2444.

14. Necşulea A, Kaessmann H (2014) Evolutionary dynamics of coding and non-coding transcriptomes. **Nature Reviews Genetics**, 15, 734-748.

15. Necşulea A, Soumillon M, Warnefors M, Liechti A, Daish T, Baker J, Grützner F, Kaessmann H (2014) The evolution of lncRNA repertoires and expression patterns in tetrapods. **Nature**, 505, 635-640.

16. Soumillon M, Necşulea A, Weier M, Brawand D, Zhang X, Gu H, Barthès P, Kokkinaki M, Nef S, Gnirke A, Dym M, de Massy B, Mikkelsen TS, Kaessmann H (2013) Cellular source and mechanisms of high transcriptome complexity in the mammalian testis. **Cell Reports**, 3, 2179-2190.

17. Julien P, Brawand D, Soumillon M, Necşulea A, Liechti A, Schütz F, Daish T, Grützner F, Kaessmann H (2012) Mechanisms and evolutionary patterns of mammalian and avian dosage compensation. **PLoS Biology**, 10, e1001328.

18. Brawand D*, Soumillon M*, Necşulea A*, Julien P, Csárdi G, Harrigan P, Weier M, Liechti A, Aximu-Petri A, Kircher M, Albert FW, Zeller U, Khaitovich P, Grützner F, Bergmann S, Nielsen R, Pääbo S, Kaessmann H (2011) The evolution of gene expression levels in mammalian organs. **Nature**, 478, 343-348.

19. Necşulea A, Popa A, Cooper DN, Stenson PD, Mouchiroud D, Gautier C, Duret L (2011) Meiotic recombination favors the spreading of deleterious mutations in human populations. **Human Mutation**, 32, 198-206.

20. Necşulea A, Sémon M, Duret L, Hurst LD (2009) Monoallelic expression and tissue specificity are associated with high crossover rates. **Trends in Genetics**, 25, 519-522.

21. Necşulea A, Guillet C, Cadoret JC, Prioleau MN, Duret L (2009) The relationship between DNA replication and human genome organization. **Molecular Biology and Evolution**, 26, 729-741.

22. Barabote RD, Xie G, Leu DH, Normand P, Necşulea A, Daubin V, Medigue C, Adney WS, Xu XC, Lapidus A, Detter C, Pujic P, Bruce DR, Challacombe JF, Brettin TS, Richardson PM, Berry AM (2009) Complete genome of the cellulolytic thermophile *Acidothermus cellulolyticus* 11B provides insights into its ecophysiological and evolutionary adaptations. **Genome Research**, 19, 1033-1043.

23. Boussau B, Blanquart S, Necşulea A, Lartillot N, Gouy M (2008) Parallel adaptations to high temperatures in the Archean eon. **Nature**, 456, 942-945.

24. Necşulea A, Lobry JR (2007) A new method for assessing the effect of replication on DNA base composition asymmetry. **Molecular Biology and Evolution**, 24, 2169-2179.

25. Necşulea A, Lobry JR (2006) Revisiting the directional mutation pressure theory: The analysis of a particular genomic structure in *Leishmania major*. **Gene**, 385, 28-40.

26. Lobry JR, Necşulea A (2006) Synonymous codon usage and its potential link with optimal growth temperature in prokaryotes. **Gene**, 385, 128-136.

References

1. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921. doi:10.1038/35057062 (2001).
2. Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270,** 467–470. doi:10.1126/science.270.5235.467 (1995).
3. Velculescu, V. E., Zhang, L., Vogelstein, B. & Kinzler, K. W. Serial analysis of gene expression. *Science* **270,** 484–487. doi:10.1126/science.270.5235.484 (1995).
4. Blat, Y. & Kleckner, N. Cohesins bind to preferential sites along yeast chromosome III, with differential regulation along arms versus the centric region. *Cell* **98,** 249–259. doi:10.1016/s0092-8674(00)81019-3 (1999).
5. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *Science* **295,** 1306–1311. doi:10.1126/science.1067799 (2002).
6. Metzker, M. L. Sequencing technologies - the next generation. *Nature Reviews Genetics* **11,** 31–46. doi:10.1038/nrg2626 (2010).
7. Brawand, D. *et al.* The evolution of gene expression levels in mammalian organs. *Nature* **478,** 343–348. doi:10.1038/nature10532 (2011).
8. Necsulea, A. *et al.* The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature* **505,** 635–640 (2014).
9. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74. doi:10.1038/nature11247 (2012).
10. Lindblad-Toh, K. *et al.* A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478,** 476–82. doi:10.1038/nature10530 (2011).
11. Ward, L. D. & Kellis, M. Evidence of abundant purifying selection in humans for recently acquired regulatory functions. *Science* **337,** 1675–1678. doi:10.1126/science.1225057 (2012).
12. Graur, D. *et al.* On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE. *Genome Biology and Evolution* **5,** 578–590. doi:10.1093/gbe/evt028 (2013).
13. Amaral, P. *et al.* The status of the human gene catalogue. *Nature* **622,** 41–47. doi:10.1038/s41586-023-06490-x (2023).

14. Montavon, T. *et al.* A regulatory archipelago controls *Hox* genes transcription in digits. *Cell* **147,** 1132–1145. doi:10.1016/j.cell.2011.10.023 (2011).

15. Gheldof, N., Leleu, M., Noordermeer, D., Rougemont, J. & Reymond, A. Detecting long-range chromatin interactions using the chromosome conformation capture sequencing (4C-seq) method. *Methods in Molecular Biology (Clifton, N.J.)* **786,** 211–225. doi:10.1007/978-1-61779-292-2_13 (2012).

16. Schoenfelder, S. *et al.* The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. *Genome Research* **25,** 582–597. doi:10.1101/gr.185272.114 (2015).

17. Ponting, C. P. & Haerty, W. Genome-wide analysis of human long noncoding RNAs: a provocative review. *Annual Review of Genomics and Human Genetics* **23,** 153–172. doi:10.1146/annurev-genom-112921-123710 (2022).

18. Mattick, J. S. *et al.* Long non-coding RNAs: definitions, functions, challenges and recommendations. *Nature Reviews. Molecular Cell Biology* **24,** 430–447. doi:10.1038/s41580-022-00566-8 (2023).

19. Bassett, A. R. *et al.* Considerations when investigating lncRNA function in vivo. *eLife* **3,** e03058. doi:10.7554/eLife.03058 (2014).

20. Palazzo, A. F. & Lee, E. S. Non-coding RNA: what is functional and what is junk? *Frontiers in Genetics* **6,** 2. doi:10.3389/fgene.2015.00002 (2015).

21. Gil, N. & Ulitsky, I. Regulation of gene expression by *cis*-acting long non-coding RNAs. *Nature Reviews. Genetics* **21,** 102–117. doi:10.1038/s41576-019-0184-5 (2020).

22. Lykke-Andersen, S. & Jensen, T. H. Nonsense-mediated mRNA decay: an intricate machinery that shapes transcriptomes. *Nature Reviews. Molecular Cell Biology* **16,** 665–677. doi:10.1038/nrm4063 (2015).

23. Frankish, A. *et al.* GENCODE: reference annotation for the human and mouse genomes in 2023. *Nucleic Acids Research* **51,** D942–D949. doi:10.1093/nar/gkac1071 (2023).

24. Saudemont, B. *et al.* The fitness cost of mis-splicing is the main determinant of alternative splicing patterns. *Genome Biology* **18,** 208. doi:10.1186/s13059-017-1344-6 (2017).

25. Bénitière, F., Necsulea, A. & Duret, L. *Random genetic drift sets an upper limit on mRNA splicing accuracy in metazoans* Pages: 2022.12.09.519597 Section: New Results. 2023. doi:10.1101/2022.12.09.519597. https://www.biorxiv.org/content/10.1101/2022.12.09.519597v5 (2023).

26. Lynch, M. The frailty of adaptive hypotheses for the origins of organismal complexity. *Proceedings of the National Academy of Sciences of the United States of America* **104 Suppl 1,** 8597–8604. doi:10.1073/pnas.0702207104 (2007).

27. Young, R. S. *et al.* Identification and properties of 1,119 candidate lincRNA loci in the *Drosophila melanogaster* genome. *Genome Biology and Evolution* **4,** 427–442. doi:10.1093/gbe/evs020 (2012).

28. Derrien, T. *et al.* The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Research* **22,** 1775–1789. doi:10.1101/gr.132159.111 (2012).

29. Haerty, W. & Ponting, C. P. Mutations within lncRNAs are effectively selected against in fruitfly but not in human. *Genome Biology* **14,** R49. doi:10.1186/gb-2013-14-5-r49 (2013).

30. Carninci, P. *et al.* The transcriptional landscape of the mammalian genome. *Science* **309,** 1559–1563. doi:10.1126/science.1112014 (2005).

31. Ponjavic, J., Ponting, C. P. & Lunter, G. Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome Res* **17,** 556–65. doi:10.1101/gr.6036807 (2007).

32. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489,** 101–108. doi:10.1038/nature11233 (2012).

33. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Research* **22,** 1760–1774. doi:10.1101/gr.135350.111 (2012).

34. Kutter, C. *et al.* Rapid turnover of long noncoding RNAs and the evolution of gene expression. *PLoS genetics* **8,** e1002841. doi:10.1371/journal.pgen.1002841 (2012).

35. Washietl, S., Kellis, M. & Garber, M. Evolutionary dynamics and tissue specificity of human long noncoding RNAs in six mammals. *Genome Res* **24,** 616–28. http://www.ncbi.nlm.nih.gov/pubmed/24429298 (2014).

36. Hezroni, H. *et al.* Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species. *Cell Reports* **11,** 1110–1122. doi:10.1016/j.celrep.2015.04.023 (2015).

37. Soumillon, M. *et al.* Cellular source and mechanisms of high transcriptome complexity in the mammalian testis. *Cell Reports* **3,** 2179–2190. doi:10.1016/j.celrep.2013.05.031 (2013).

38. Darbellay, F. & Necsulea, A. Comparative transcriptomics analyses across species, organs, and developmental stages reveal functionally constrained lncRNAs. *Molecular Biology and Evolution* **37,** 240–259. doi:10.1093/molbev/msz212 (2020).

39. Marques, A. C. *et al.* Chromatin signatures at transcriptional start sites separate two equally populated yet distinct classes of intergenic long noncoding RNAs. *Genome Biology* **14,** R131. doi:10.1186/gb-2013-14-11-r131 (2013).

40. Latos, P. A. *et al. Airn* transcriptional overlap, but not its lncRNA products, induces imprinted *Igf2r* silencing. *Science* **338,** 1469–1472. doi:10.1126/science.1228110 (2012).

41. McDonald, J. H. & Kreitman, M. Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* **351,** 652–654. doi:10.1038/351652a0 (1991).

42. Sarropoulos, I., Marin, R., Cardoso-Moreira, M. & Kaessmann, H. Developmental dynamics of lncRNAs across mammalian organs and species. *Nature* **571,** 510–514. doi:10.1038/s41586-019-1341-x (2019).

43. Engreitz, J. M. *et al.* Local regulation of gene expression by lncRNA promoters, transcription and splicing. *Nature* **539,** 452–455. doi:10.1038/nature20149 (2016).

44. Andersson, R. & Sandelin, A. Determinants of enhancer and promoter activities of regulatory elements. *Nature Reviews Genetics* **21,** 71–87. doi:10.1038/s41576-019-0173-8 (2020).

45. Kim, T.-K. *et al.* Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465,** 182–187. doi:10.1038/nature09033 (2010).

46. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507,** 455–461. doi:10.1038/nature12787 (2014).

47. Li, W. *et al.* Functional importance of eRNAs for estrogen-dependent transcriptional activation events. *Nature* **498,** 516–520. doi:10.1038/nature12210 (2013).

48. Huarte, M. *et al.* A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. *Cell* **142,** 409–419. doi:10.1016/j.cell.2010.06.040 (2010).

49. Vazquez, A., Bond, E. E., Levine, A. J. & Bond, G. L. The genetics of the p53 pathway, apoptosis and cancer therapy. *Nature Reviews. Drug Discovery* **7,** 979–987. doi:10.1038/nrd2656 (2008).

50. Dimitrova, N. *et al.* LincRNA-p21 activates p21 in *cis* to promote polycomb target gene expression and to enforce the G1/S checkpoint. *Molecular Cell* **54.** Publisher: Elsevier, 777–790. doi:10.1016/j.molcel.2014.04.025 (2014).

51. Groff, A. F. *et al. In vivo* characterization of *Linc-p21* reveals functional cis-regulatory DNA elements. *Cell Reports* **16,** 2178–2186. doi:10.1016/j.celrep.2016.07.050 (2016).

52. Gil, N. & Ulitsky, I. Production of spliced long noncoding RNAs specifies regions with increased enhancer activity. *Cell Systems* **7,** 537–547.e3. doi:10.1016/j.cels.2018.10.009 (2018).

53. Marquardt, S. *et al.* Functional consequences of splicing of the antisense transcript COOLAIR on FLC transcription. *Molecular Cell* **54.** Publisher: Elsevier, 156–165. doi:10.1016/j.molcel.2014.03.026. https://www.cell.com/molecular-cell/abstract/S1097-2765(14)00259-7 (2023) (2014).

54. Lyle, R. *et al.* The imprinted antisense RNA at the Igf2r locus overlaps but does not imprint Mas1. *Nature Genetics* **25,** 19–21. doi:10.1038/75546 (2000).

55. Sleutels, F., Zwart, R. & Barlow, D. P. The non-coding Air RNA is required for silencing autosomal imprinted genes. *Nature* **415,** 810–813. doi:10.1038/415810a (2002).

56. Rinn, J. L. *et al.* Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* **129,** 1311–1323. doi:10.1016/j.cell.2007.05.022 (2007).

57. Mallo, M., Wellik, D. M. & Deschamps, J. Hox genes and regional patterning of the vertebrate body plan. *Developmental Biology* **344,** 7–15. doi:10.1016/j.ydbio.2010.04.024 (2010).

58. Schorderet, P. & Duboule, D. Structural and functional differences in the long non-coding RNA hotair in mouse and human. *PLoS genetics* **7,** e1002071. doi:10.1371/journal.pgen.1002071 (2011).

59. Li, L. *et al.* Targeted disruption of Hotair leads to homeotic transformation and gene derepression. *Cell Reports* **5,** 3–12. doi:10.1016/j.celrep.2013.09.003 (2013).

60. Amândio, A. R., Necsulea, A., Joye, E., Mascrez, B. & Duboule, D. Hotair is dispensible for mouse development. *PLoS genetics* **12.** tex.ids: amandio_hotair_2016-1, e1006232. doi:10.1371/journal.pgen.1006232 (2016).

61. Finn, R. S. *et al.* Atezolizumab plus bevacizumab in unresectable hepatocellular carcinoma. *The New England Journal of Medicine* **382,** 1894–1905. doi:10.1056/NEJMoa1915745 (2020).

62. Boldanova, T., Suslov, A., Heim, M. H. & Necsulea, A. Transcriptional response to hepatitis C virus infection and interferon-alpha treatment in the human liver. *EMBO molecular medicine* **9,** 816–834. doi:10.15252/emmm.201607006 (2017).

63. Ng, C. K. Y. *et al.* Integrative proteogenomic characterization of hepatocellular carcinoma across etiologies and stages. *Nature Communications* **13,** 2436. doi:10.1038/s41467-022-29960-8 (2022).

64. Necsulea, A. *et al. LncRNA analyses reveal increased levels of non-coding centromeric transcripts in hepatocellular carcinoma* Pages: 2021.03.03.433778 Section: New Results. 2021. doi:10.1101/2021.03.03.433778. https://www.biorxiv.org/content/10.1101/2021.03.03.433778v1 (2022).

65. Hutchinson, J. N. *et al.* A screen for nuclear transcripts identifies two linked noncoding RNAs associated with SC35 splicing domains. *BMC genomics* **8,** 39. doi:10.1186/1471-2164-8-39 (2007).

66. Brannan, C. I., Dees, E. C., Ingram, R. S. & Tilghman, S. M. The product of the H19 gene may function as an RNA. *Molecular and Cellular Biology* **10,** 28–36. doi:10.1128/mcb.10.1.28-36.1990 (1990).

67. Chang, S. *et al.* Dysregulated H19/Igf2 expression disrupts cardiac-placental axis during development of Silver-Russell syndrome-like mouse models. *eLife* **11,** e78754. doi:10.7554/eLife.78754 (2022).

68. Brown, C. J. *et al.* Localization of the X inactivation centre on the human X chromosome in Xq13. *Nature* **349,** 82–84. doi:10.1038/349082a0 (1991).

69. Arun, G. *et al.* Differentiation of mammary tumors and reduction in metastasis upon Malat1 lncRNA loss. *Genes & Development* **30,** 34–51. doi:10.1101/gad.270959.115 (2016).

70. Kim, J. *et al.* Long noncoding RNA MALAT1 suppresses breast cancer metastasis. *Nature Genetics* **50,** 1705–1715. doi:10.1038/s41588-018-0252-3 (2018).

71. Li, G.-Z. *et al.* MALAT1/ mir-1-3p mediated BRF2 expression promotes HCC progression via inhibiting the LKB1/AMPK signaling pathway. *Cancer Cell International* **23,** 188. doi:10.1186/s12935-023-03034-1 (2023).

72. Matouk, I. J. *et al.* The H19 non-coding RNA is essential for human tumor growth. *PloS One* **2,** e845. doi:10.1371/journal.pone.0000845 (2007).

73. Yoshimizu, T. *et al.* The H19 locus acts in vivo as a tumor suppressor. *Proceedings of the National Academy of Sciences of the United States of America* **105,** 12417–12422. doi:10.1073/pnas.0801540105 (2008).

74. Nuciforo, S. *et al.* Organoid models of human liver cancers derived from tumor needle biopsies. *Cell Reports* **24,** 1363–1376. doi:10.1016/j.celrep.2018.07.001. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6088153/ (2023) (2018).

75. Sharma, H. & Verma, S. Predatory journals: The rise of worthless biomedical science. *Journal of Postgraduate Medicine* **64,** 226–231. doi:10.4103/jpgm.JPGM_347_18. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6198688/ (2023) (2018).

76. Cardoso-Moreira, M. *et al.* Gene expression across mammalian organ development. *Nature* **571,** 505–509. doi:10.1038/s41586-019-1338-5 (2019).

77. Grant, J. *et al.* Rsx is a metatherian RNA with Xist-like properties in X-chromosome inactivation. *Nature* **487,** 254–258. doi:10.1038/nature11171 (2012).

78. Vallot, C. *et al.* XACT, a long noncoding transcript coating the active X chromosome in human pluripotent cells. *Nature Genetics* **45,** 239–241. doi:10.1038/ng.2530 (2013).

79. Duret, L., Chureau, C., Samain, S., Weissenbach, J. & Avner, P. The Xist RNA gene evolved in eutherians by pseudogenization of a protein-coding gene. *Science* **312,** 1653–5. doi:10.1126/science.1126316 (2006).

80. King, M. C. & Wilson, A. C. Evolution at two levels in humans and chimpanzees. *Science* **188,** 107–116 (1975).

81. Segert, J. A., Gisselbrecht, S. S. & Bulyk, M. L. Transcriptional silencers: driving gene expression with the brakes on. *Trends Genet* **37,** 514–527. doi:10.1016/j.tig.2021.02.002 (2021).

82. Banerji, J., Rusconi, S. & Schaffner, W. Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell* **27,** 299–308 (1981).

83. Brand, A. H., Breeden, L., Abraham, J., Sternglanz, R. & Nasmyth, K. Characterization of a "silencer" in yeast: a DNA sequence with properties opposite to those of a transcriptional enhancer. *Cell* **41,** 41–48. doi:10.1016/0092-8674(85)90059-5 (1985).

84. Calo, E. & Wysocka, J. Modification of enhancer chromatin: what, how, and why? *Molecular Cell* **49,** 825–837. doi:10.1016/j.molcel.2013.01.038 (2013).

85. Aday, A. W., Zhu, L. J., Lakshmanan, A., Wang, J. & Lawson, N. D. Identification of cis regulatory features in the embryonic zebrafish genome through large-scale profiling of H3K4me1 and H3K4me3 binding sites. *Developmental Biology* **357,** 450–462. doi:10.1016/j.ydbio.2011.03.007 (2011).

86. Blow, M. J. *et al.* ChIP-Seq identification of weakly conserved heart enhancers. *Nature Genetics* **42,** 806–810. doi:10.1038/ng.650 (2010).

87. Xi, H. *et al.* Identification and characterization of cell type-specific and ubiquitous chromatin regulatory structures in the human genome. *PLoS genetics* **3,** e136. doi:10.1371/journal.pgen.0030136 (2007).

88. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods* **10,** 1213–1218. doi:10.1038/nmeth.2688. (2022) (2013).

89. De Santa, F. *et al.* A large fraction of extragenic RNA pol II transcription sites overlap enhancers. *PLoS biology* **8,** e1000384. doi:10.1371/journal.pbio.1000384 (2010).

90. Arnold, C. D. *et al.* Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339,** 1074–1077. doi:10.1126/science.1232542 (2013).

91. Pang, B., van Weerd, J. H., Hamoen, F. L. & Snyder, M. P. Identification of non-coding silencer elements and their regulation of gene expression. *Nature Reviews Molecular Cell Biology* **24.** Number: 6 Publisher: Nature Publishing Group, 383–395. doi:10.1038/s41580-022-00549-9. (2023) (2023).

92. Doni Jayavelu, N., Jajodia, A., Mishra, A. & Hawkins, R. D. Candidate silencer elements for the human and mouse genomes. *Nature Communications* **11,** 1061. doi:10.1038/s41467-020-14853-5 (2020).

93. Huang, D., Petrykowska, H. M., Miller, B. F., Elnitski, L. & Ovcharenko, I. Identification of human silencers by correlating cross-tissue epigenetic profiles and gene expression. *Genome Research* **29,** 657–667. doi:10.1101/gr.247007.118 (2019).

94. Cai, Y. *et al.* H3K27me3-rich genomic regions can function as silencers to repress gene expression via chromatin interactions. *Nature Communications* **12.** Number: 1 Publisher: Nature Publishing Group, 719. doi:10.1038/s41467-021-20940-y. (2023) (2021).

95. Pang, B. & Snyder, M. P. Systematic identification of silencers in human cells. *Nature Genetics* **52,** 254–263. doi:10.1038/s41588-020-0578-5 (2020).

96. Hansen, T. J. & Hodges, E. ATAC-STARR-seq reveals transcription factor-bound activators and silencers across the chromatin accessible human genome. *Genome Research* **32,** 1529–1541. doi:10.1101/gr.276766.122 (2022).

97. Gisselbrecht, S. S. *et al.* Transcriptional Silencers in Drosophila Serve a Dual Role as Transcriptional Enhancers in Alternate CellularContexts. *Molecular Cell* **77,** 324–337.e8. doi:10.1016/j.molcel.2019.10.004 (2020).

98. Bessis, A., Champtiaux, N., Chatelin, L. & Changeux, J. P. The neuron-restrictive silencer element: a dual enhancer/silencer crucial for patterned expression of a nicotinic receptor gene in the brain. *Proceedings of the National Academy of Sciences of the United States of America* **94,** 5906–5911. doi:10.1073/pnas.94.11.5906 (1997).

99. Heintzman, N. D. *et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459,** 108–112. doi:10.1038/nature07829 (2009).

100. Shen, Y. *et al.* A map of the cis-regulatory sequences in the mouse genome. *Nature* **488,** 116–120. doi:10.1038/nature11243 (2012).

101. Kim, S. & Wysocka, J. Deciphering the multi-scale, quantitative cis-regulatory code. *Molecular Cell* **83,** 373–392. doi:10.1016/j.molcel.2022.12.032 (2023).

102. Reményi, A., Schöler, H. R. & Wilmanns, M. Combinatorial control of gene expression. *Nature Structural & Molecular Biology* **11,** 812–815. doi:10.1038/nsmb820 (2004).

103. Osterwalder, M. *et al.* Enhancer redundancy provides phenotypic robustness in mammalian development. *Nature* **554,** 239–243. doi:10.1038/nature25461 (2018).

104. Pfeifer, D. *et al.* Campomelic dysplasia translocation breakpoints are scattered over 1 Mb proximal to *SOX9*: evidence for an extended control region. *American Journal of Human Genetics* **65,** 111–124. doi:10.1086/302455 (1999).

105. Lettice, L. A. *et al.* A long-range *Shh* enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Human Molecular Genetics* **12,** 1725–1735. doi:10.1093/hmg/ddg180 (2003).

106. Spitz, F., Gonzalez, F. & Duboule, D. A global control region defines a chromosomal regulatory landscape containing the HoxD cluster. *Cell* **113,** 405–417. doi:10.1016/s0092-8674(03)00310-6 (2003).

107. Tolhuis, B., Palstra, R. J., Splinter, E., Grosveld, F. & de Laat, W. Looping and interaction between hypersensitive sites in the active beta-globin locus. *Molecular Cell* **10,** 1453–1465. doi:10.1016/s1097-2765(02)00781-5 (2002).

108. Splinter, E. *et al.* CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus. *Genes & Development* **20,** 2349–2354. doi:10.1101/gad.399506 (2006).

109. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326,** 289–293. doi:10.1126/science.1181369 (2009).

110. Li, G. *et al.* Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* **148,** 84–98. doi:10.1016/j.cell.2011.12.014 (2012).

111. Mumbach, M. R. *et al.* HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nature Methods* **13.** Number: 11 Publisher: Nature Publishing Group, 919–922. doi:10.1038/nmeth.3999. https://www.nature.com/articles/nmeth.3999 (2023) (2016).

112. Mifsud, B. *et al.* Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nature Genetics* **47,** 598–606. doi:10.1038/ng.3286 (2015).

113. Lee, B. H., Wu, Z. & Rhie, S. K. Characterizing chromatin interactions of regulatory elements and nucleosome positions, using Hi-C, Micro-C, and promoter capture Micro-C. *Epigenetics & Chromatin* **15,** 41. doi:10.1186/s13072-022-00473-4 (2022).

114. Goel, V. Y., Huseyin, M. K. & Hansen, A. S. Region Capture Micro-C reveals coalescence of enhancers and promoters into nested microcompartments. *Nature Genetics* **55,** 1048–1056. doi:10.1038/s41588-023-01391-1 (2023).

115. Schoenfelder, S. & Fraser, P. Long-range enhancer-promoter contacts in gene expression control. *Nature Reviews. Genetics* **20,** 437–455. doi:10.1038/s41576-019-0128-0 (2019).

116. Friman, E. T., Flyamer, I. M., Marenduzzo, D., Boyle, S. & Bickmore, W. A. Ultra-long-range interactions between active regulatory elements. *Genome Research* **33,** 1269–1283. doi:10.1101/gr.277567.122 (2023).

117. Vieux-Rochas, M., Fabre, P. J., Leleu, M., Duboule, D. & Noordermeer, D. Clustering of mammalian Hox genes with other H3K27me3 targets within an active nuclear domain. *Proceedings of the National Academy of Sciences* **112.** Publisher: Proceedings of the National Academy of Sciences, 4672–4677. doi:10.1073/pnas.1504783112. (2023) (2015).

118. Schoenfelder, S. *et al.* Polycomb repressive complex PRC1 spatially constrains the mouse embryonic stem cell genome. *Nature Genetics* **47,** 1179–1186. doi:10.1038/ng.3393 (2015).

119. Paliou, C. *et al.* Preformed chromatin topology assists transcriptional robustness of Shh during limb development. *Proceedings of the National Academy of Sciences* **116.** Publisher: Proceedings of the National Academy of Sciences, 12390–12399. doi:10.1073/pnas.1900672116. (2023) (2019).

120. Nobrega, M. A., Ovcharenko, I., Afzal, V. & Rubin, E. M. Scanning human gene deserts for long-range enhancers. *Science* **302,** 413. doi:10.1126/science.1088328 (2003).

121. Ovcharenko, I. *et al.* Evolution and functional classification of vertebrate gene deserts. *Genome Research* **15.** Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab, 137–145. doi:10.1101/gr.3015505. https://genome.cshlp.org/content/15/1/137 (2023) (2005).

122. Javierre, B. M. *et al.* Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell* **167.** Publisher: Elsevier, 1369–1384.e19. doi:10.1016/j.cell.2016.09.037. https://www.cell.com/cell/abstract/S0092-8674(16)31322-8 (2023) (2016).

123. Novo, C. L. *et al.* Long-Range Enhancer Interactions Are Prevalent in Mouse Embryonic Stem Cells and Are Reorganized upon Pluripotent State Transition. *Cell Reports* **22,** 2615–2627. doi:10.1016/j.celrep.2018.02.040 (2018).

124. Kraft, K. *et al.* Polycomb-mediated genome architecture enables long-range spreading of H3K27 methylation. *Proceedings of the National Academy of Sciences* **119,** e2201883119. doi:10.1073/pnas.2201883119. (2023) (2022).

125. Nora, E. P. *et al.* Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485,** 381–385. doi:10.1038/nature11049 (2012).

126. Tanay, A. & Cavalli, G. Chromosomal domains: epigenetic contexts and functional implications of genomic compartmentalization. *Current Opinion in Genetics & Development* **23,** 197–203. doi:10.1016/j.gde.2012.12.009 (2013).

127. Szabo, Q., Bantignies, F. & Cavalli, G. Principles of genome folding into topologically associating domains. *Science Advances* **5,** eaaw1668. doi:10.1126/sciadv.aaw1668 (2019).

128. Sikorska, N. & Sexton, T. Defining functionally relevant spatial chromatin domains: it is a TAD complicated. *Journal of Molecular Biology* **432,** 653–664. doi:10.1016/j.jmb.2019.12.006 (2020).

129. Da Costa-Nunes, J. A. & Noordermeer, D. TADs: Dynamic structures to create stable regulatory functions. *Current Opinion in Structural Biology* **81,** 102622. doi:10.1016/j.sbi.2023.102622 (2023).

130. Chakraborty, S. *et al.* Enhancer–promoter interactions can bypass CTCF-mediated boundaries and contribute to phenotypic robustness. *Nature Genetics* **55,** 280–290. doi:10.1038/s41588-022-01295-6. (2023) (2023).

131. Rodríguez-Carballo, E. *et al.* Chromatin topology and the timing of enhancer function at the HoxD locus. *Proceedings of the National Academy of Sciences of the United States of America* **117,** 31231–31241. doi:10.1073/pnas.2015083117 (2020).

132. De Wit, E. TADs as the caller calls them. *Journal of Molecular Biology. Perspectives on Chromosome Folding* **432,** 638–642. doi:10.1016/j.jmb.2019.09.026. https://www.sciencedirect.com/science/article/pii/S0022283619305923 (2023) (2020).

133. Acemel, R. D., Maeso, I. & Gómez-Skarmeta, J. L. Topologically associated domains: a successful scaffold for the evolution of gene regulation in animals. *Wiley Interdisciplinary Reviews. Developmental Biology* **6.** doi:10.1002/wdev.265 (2017).

134. Acemel, R. D. & Lupiáñez, D. G. Evolution of 3D chromatin organization at different scales. *Current Opinion in Genetics & Development* **78,** 102019. doi:10.1016/j.gde.2022.102019 (2023).

135. Darbellay, F. & Duboule, D. Topological domains, metagenes, and the emergence of pleiotropic regulations at Hox loci. *Current Topics in Developmental Biology* **116,** 299–314. doi:10.1016/bs.ctdb.2015.11.022 (2016).

136. Eres, I. E. & Gilad, Y. A TAD skeptic: is 3D genome topology conserved? *Trends in genetics: TIG* **37,** 216–223. doi:10.1016/j.tig.2020.10.009 (2021).

137. Foissac, S. *et al.* Multi-species annotation of transcriptome and chromatin structure in domesticated animals. *BMC biology* **17,** 108. doi:10.1186/s12915-019-0726-5 (2019).

138. Lupiáñez, D. G. *et al.* Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* **161,** 1012–1025. doi:10.1016/j.cell.2015.04.004 (2015).

139. Hnisz, D. *et al.* Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science* **351,** 1454–1458. doi:10.1126/science.aad9024 (2016).

140. Franke, M. *et al.* Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature* **538,** 265–269. doi:10.1038/nature19800 (2016).

141. Despang, A. *et al.* Functional dissection of the *Sox9*-Kcnj2 locus identifies nonessential and instructive roles of TAD architecture. *Nature Genetics* **51,** 1263–1271. doi:10.1038/s41588-019-0466-z (2019).

142. Deng, W. *et al.* Reactivation of developmentally silenced globin genes by forced chromatin looping. *Cell* **158,** 849–860. doi:10.1016/j.cell.2014.05.050 (2014).

143. Bartman, C. R., Hsu, S. C., Hsiung, C. C.-S., Raj, A. & Blobel, G. A. Enhancer regulation of transcriptional bursting parameters revealed by forced chromatin looping. *Molecular Cell* **62,** 237–247. doi:10.1016/j.molcel.2016.03.007 (2016).

144. Morgan, S. L. *et al.* Manipulation of nuclear architecture through CRISPR-mediated chromosomal looping. *Nature Communications* **8,** 15993. doi:10.1038/ncomms15993 (2017).

145. Benabdallah, N. S. *et al.* Decreased enhancer-promoter proximity accompanying enhancer activation. *Molecular Cell* **76,** 473–484.e7. doi:10.1016/j.molcel.2019.07.038 (2019).

146. Taylor, T. *et al.* Transcriptional regulation and chromatin architecture maintenance are decoupled functions at the Sox2 locus. *Genes & Development* **36,** 699–717. doi:10.1101/gad.349489.122. (2023) (2022).

147. Alexander, J. M. *et al.* Live-cell imaging reveals enhancer-dependent Sox2 transcription in the absence of enhancer proximity. *eLife* **8,** e41769. doi:10.7554/eLife.41769 (2019).

148. Karr, J. P., Ferrie, J. J., Tjian, R. & Darzacq, X. The transcription factor activity gradient (TAG) model: contemplating a contact-independent mechanism for enhancer–promoter communication. *Genes & Development* **36.** Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab, 7–16. doi:10.1101/gad.349160.121. http://genesdev.cshlp.org/content/36/1-2/7 (2023) (2022).

149. Lim, B. & Levine, M. S. Enhancer-promoter communication: hubs or loops? *Current Opinion in Genetics & Development. Genome Architecture and Expression* **67,** 5–9. doi:10.1016/j.gde.2020.10.001. (2023) (2021).

150. Li, J. *et al.* Single-gene imaging links genome topology, promoter–enhancer communication and transcription control. *Nature Structural & Molecular Biology* **27,** 1032–1040. doi:10.1038/s41594-020-0493-6. https://www.nature.com/articles/s41594-020-0493-6 (2023) (2020).

151. Mifsud, B. *et al.* GOTHiC, a probabilistic model to resolve complex biases and to identify real interactions in Hi-C data. *PloS One* **12,** e0174744. doi:10.1371/journal.pone.0174744 (2017).

152. Cairns, J. *et al.* CHiCAGO: robust detection of DNA looping interactions in Capture Hi-C data. *Genome Biology* **17,** 127. doi:10.1186/s13059-016-0992-2. https://doi.org/10.1186/s13059-016-0992-2 (2020) (2016).

153. Laverré, A., Tannier, E. & Necsulea, A. Long-range promoter–enhancer contacts are conserved during evolution and contribute to gene expression robustness. *Genome Research* **32,** 280–296 (2022).

154. Mongin, E., Dewar, K. & Blanchette, M. Long-range regulation is a major driving force in maintaining genome integrity. *BMC evolutionary biology* **9,** 203. doi:10.1186/1471-2148-9-203 (2009).

155. Mongin, E., Dewar, K. & Blanchette, M. Mapping association between long-range cis-regulatory regions and their target genes using synteny. *Journal of Computational Biology* **18,** 1115–1130. doi:10.1089/cmb.2011.0088. (2023) (2011).

156. Irimia, M. *et al.* Extensive conservation of ancient microsynteny across metazoans due to cis-regulatory constraints. *Genome Research* **22,** 2356–2367. doi:10.1101/gr.139725.112 (2012).

157. Naville, M. *et al.* Long-range evolutionary constraints reveal cis-regulatory interactions on the human X chromosome. *Nature Communications* **6,** 6904. doi:10.1038/ncomms7904 (2015).

158. Clément, Y., Torbey, P., Gilardi-Hebenstreit, P. & Roest Crollius, H. Enhancer-gene maps in the human and zebrafish genomes using evolutionary linkage conservation. *Nucleic Acids Research* **48,** 2357–2371. doi:10.1093/nar/gkz1199 (2020).

159. Liu, Y., Sarkar, A., Kheradpour, P., Ernst, J. & Kellis, M. Evidence of reduced recombination rate in human regulatory domains. *Genome Biology* **18,** 193. doi:10.1186/s13059-017-1308-x (2017).

160. Whalen, S. & Pollard, K. S. Most chromatin interactions are not in linkage disequilibrium. *Genome Research* **29,** 334–343. doi:10.1101/gr.238022.118. http://genome.cshlp.org/content/29/3/334 (2020) (2019).

161. Kragesteen, B. K. *et al.* Dynamic 3D chromatin architecture contributes to enhancer specificity and limb morphogenesis. *Nature Genetics* **50,** 1463–1473. doi:10.1038/s41588-018-0221-x (2018).

162. Yakushiji-Kaminatsui, N. *et al.* Similarities and differences in the regulation of HoxD genes during chick and mouse limb development. *PLoS biology* **16,** e3000004. doi:10.1371/journal.pbio.3000004 (2018).

163. Ushiki, A. *et al.* Deletion of CTCF sites in the SHH locus alters enhancer-promoter interactions and leads to acheiropodia. *Nature Communications* **12,** 2282. doi:10.1038/s41467-021-22470-z (2021).

164. Berthelot, C., Villar, D., Horvath, J. E., Odom, D. T. & Flicek, P. Complexity and conservation of regulatory landscapes underlie evolutionary resilience of mammalian gene expression. *Nature Ecology & Evolution* **2,** 152–163. doi:10.1038/s41559-017-0377-2 (2018).

165. Haerty, W. & Ponting, C. P. No gene in the genome makes sense except in the light of evolution. *Annual Review of Genomics and Human Genetics* **15,** 71–92. doi:10.1146/annurev-genom-090413-025621 (2014).

166. Giudicelli, F. & Roest Crollius, H. On the importance of evolutionary constraint for regulatory sequence identification. *Briefings in Functional Genomics,* elab015. doi:10.1093/bfgp/elab015 (2021).

167. McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nature Biotechnology* **28,** 495–501. doi:10.1038/nbt.1630 (2010).

168. Laverré, A., Tannier, E., Veber, P. & Necsulea, A. *GOntact: using chromatin contacts to infer Gene Ontology enrichments for cis-regulatory elements* Pages: 2022.06.13.495495 Section: New Results. 2022. doi:10.1101/2022.06.13.495495. https://www.biorxiv.org/content/10.1101/2022.06.13.495495v1 (2023).

169. Lupiáñez, D. G., Spielmann, M. & Mundlos, S. Breaking TADs: how alterations of chromatin domains result in disease. *Trends in genetics: TIG* **32,** 225–237. doi:10.1016/j.tig.2016.01.003 (2016).

170. Ghavi-Helm, Y. *et al.* Highly rearranged chromosomes reveal uncoupling between genome topology and gene expression. *Nature Genetics* **51,** 1272–1282. doi:10.1038/s41588-019-0462-3 (2019).

171. Kaul, A., Bhattacharyya, S. & Ay, F. Identifying statistically significant chromatin contacts from Hi-C data with FitHiC2. *Nature Protocols* **15,** 991–1012. doi:10.1038/s41596-019-0273-0 (2020).