Computational analysis of transcriptional regulation in metazoans

Morgane Thomas-Chollier

ENS
ÉCOLE NORMALE
SUPÉRIEURE

Computational analysis of transcriptional regulation in metazoans

A dissertation submitted for the degree of **Habilitation à Diriger des Recherches** from Ecole normale supérieure, Paris

Morgane Thomas-Chollier

Defense : 14th June 2016

Jury composition:
external experts:
Stein Aerts, Associate Professor KU Leuven (Belgium)
Dave Ferrier, Senior Lecturer University of St Andrews (Scotland)
Yacine Graba, Research Director CNRS Marseille
jury members:
Patrick Charnay, Professor ENS Paris
Denis Thieffry, Professor ENS Paris
Hélène Touzet, Research Director CNRS Lille
Jacques van Helden, Professor AMU Marseille
Michel Vervoort, Professor Université Paris Diderot

May 2016

Acknowledgements

First, I wish to thank the three external experts Stein Aerts, David Ferrier and Yacine Graba for reviewing this work. I would also like to thank all members of the jury for being present at the defense and coming from far: Yacine Graba, Hélène Touzet, Jacques van Helden, but also the more local members: Patrick Charnay, Denis Thieffry and Michel Vervoort. I am deeply grateful to have all of you in my HDR jury.

A particular thank to Denis Thieffry for his constant support and kindness, and for making the lab a welcoming, happy and fulfilling place to come to everyday. His original point of view on my work, both for research and teaching, is an important source of inspiration and questioning. His capacity to calmly face any administrative trouble is remarkable, and a real asset that I envy.

I am deeply grateful to my former supervisors Jacques van Helden and Martin Vingron, but also Michel Vervoort who all encouraged me in pursuing in academia, believing in my capacities to go further, and "co-opting" me, even though I have a second X chromosome.

Special thanks to Jacques van Helden, who is an inspiring scientist. His willingness to trust me with RSAT (even when he disagrees!) is very important.

I would like to thank Sebastiaan Meising, who has become my main experimental collaborator these recent years. His inclination to create a fair collaborative work with bioinformaticians is exemplar. It has been a real pleasure to have tutored Stefanie and Jonas together.

At IBENS, I am thankful to all members of CSB for everyday fun and great work, and the members of Dyogen, the Genomic Platform and the Bioinformatic platform (mostly for fun, but also a bit for work !) and Brigitte and Abdul for their endless help with administrative tasks. Thank you to Stéphane Le Crom for insightful/funny discussions regarding professorship. Special thanks to the members of the Informatic Platform, who are the quiet angels keeping computers working.

My work at ENS also consists in teaching, and I am very thankful to the pedagogic team of the Biology department for their welcome and guidance during these four years. Particular thanks to Pierre, Andrea and Patrick for their support.

Un grand merci à ma Maman, pour tout. Je croyais en avoir fini lors de la thèse de doctorat. Maintenant, c'est sûr, c'est bon, j'ai terminé mes études. Merci bien sûr à ma famille.

Last, but not least, I cannot thank you Benjamin enough for what you bring to my personal and scientific life. We're a team. Raising Andreas along your side is my everyday miracle. Now, it is your turn to write your HDR !

Like my PhD thesis, this document is dedicated in loving memory of *Titou*, because family is not just people with related DNA. I think you would have liked that.

English Summary

This HDR thesis presents my work on transcriptional regulation in metazoans (animals). As a computational biologist, my research activities cover both the development of new bioinformatics tools, and contributions to a better understanding of biological questions. The first part focuses on transcription factors, with a study of the evolution of Hox and ParaHox gene families across metazoans, for which I developed HoxPred, a bioinformatics tool to automatically classify these genes into their groups of homology. Transcription factors regulate their target genes by binding to short cis-regulatory elements in DNA. The second part of this thesis introduces the prediction of these cis-regulatory elements in genomic sequences, and my contributions to the development of user-friendly computational tools (RSAT software suite and TRAP). The third part covers the detection of these cis-regulatory elements using high-throughput sequencing experiments such as ChIP-seq or ChIP-exo. The bioinformatics developments include reusable pipelines to process these datasets, and novel motif analysis tools adapted to these large datasets (RSAT *peak-motifs* and ExoProfiler). As all these approaches are generic, I naturally apply them to diverse biological questions, in close collaboration with experimental groups. In particular, this third part presents the studies uncovering new DNA sequences that are driving or preventing the binding of the glucocorticoid receptor. Finally, my research perspectives are introduced, especially regarding further developments within the RSAT suite enabling cross-species conservation analyses, and new collaborations with experimental teams, notably to tackle the epigenomic remodelling during osteoporosis.

Résumé en français

Cette thèse d'HDR présente mes travaux concernant la régulation transcriptionelle chez les métazoaires (animaux). En tant que biologiste computationelle, mes activités de recherche portent sur le développement de nouveaux outils bioinformatiques, et contribuent à une meilleure compréhension de questions biologiques. La première partie concerne les facteurs de transcriptions, avec une étude de l'évolution des familles de gènes Hox et ParaHox chez les métazoaires. Pour cela, j'ai développé HoxPred, un outil bioinformatique qui classe automatiquement ces gènes dans leur groupe d'homologie. Les facteurs de transcription régulent leurs gènes cibles en se fixant à l'ADN sur des petites régions cis-régulatrices. La seconde partie de cette thèse introduit la prédiction de ces éléments cis-régulateurs au sein de séquences génomiques, et présente mes contributions au développement d'outils accessibles aux non-spécialistes (la suite RSAT et TRAP). La troisième partie couvre la détection de ces éléments cis-régulateurs grâce aux expériences basées sur le séquençage à haut débit comme le ChIP-seq ou le ChIP-exo. Les développements bioinformatiques incluent des pipelines réutilisables pour analyser ces jeux de données, ainsi que de nouveaux outils d'analyse de motifs adaptés à ces grands jeux de données (RSAT *peak-motifs* et ExoProfiler). Comme ces approches sont génériques, je les applique naturellement à des questions biologiques diverses, en étroite collaboration avec des groupes expérimentaux. En particulier, cette troisième partie présente les études qui ont permis de mettre en évidence de nouvelles séquences d'ADN qui favorisent ou empêchent la fixation du récepteur aux glucocorticoides. Enfin, mes perspectives de recherche sont présentées, plus particulièrement concernant les nouveaux développements au sein de la suite RSAT pour permettre des analyses basées sur la conservation inter-espèces, mais aussi de nouvelles collaborations avec des équipes expérimentales, notamment pour étudier le remodelage épigénomique au cours de l'ostéoporose.

Contents

List of abbreviations

aa	amino acid
ChIP	Chromatin Immunoprecipitation
ChIP-seq	Chromatin Immunoprecipitation followed by sequencing
ChIP-exo	Chromatin Immunoprecipitation followed by exonuclease
CRER	Cis-Regulatory Element enriched Region
CRM	Cis-Regulatory Module
DAMID	DNA Adenine Methyltransferase IDentification
ENCODE	ENCyclopedia Of DNA Elements
ENS	Ecole normal supérieure
FRiP	Fraction of Reads in Peaks
GR	Glucocorticoid Receptor
GRN	Gene Regulatory Network
IBENS	Institut de Biologie de lcole normal supérieure
IDR	Irreproducible Discovery Rate
IP	Immunoprecipitation
NRF	Non Redundant Fraction
NRS	Negative Regulatory Sequence
NSC	Normalised Strand Coefficient
PCR	Polymerase Chain Reaction
PG	Paralogous Group
PSSM	Position-specific scoring matrix
RSAT	Regulatory Sequence Analysis Tools
RSC	Relative Strand Correlation
TF	Transcription factor
TFBS	Transcription Factor Binding Site
TRAP	TRanscription factor Affinity Prediction

List of organism names

Scientific name	**phylum**	**common name**
Amphimedon queenslandica	Porifera	sponge
Drosophila melanogaster	Arthropods	fruitfly
Caenorhabditis elegans	Nematodes	roundworm
Mnemiopsis leidyi	Ctenophore	sea walnut
Nematostella vectensis	Cnidaria	starlet sea anemone
Paracentrotus lividus	Ambulacraria	rock sea urchin
Pleurobrachia bachei	Ctenophore	sea gooseberry
Strongylocentrotus purpuratus	Ambulacraria	purple sea urchin
Symsagittifera roscoffensis	Xenacoelomorpha	mint-sauce worm
Trichoplax adherens	Placozoa	-
Xenoturbella bocki	Xenacoelomorpha	-

Foreword

Being a bioinformatician in 2016 is both thrilling and frustrating.

Thrilling, as in less than a decade (barely since my PhD), we have been propelled into the ”Big Data” era of Biology [Stephens et al., 2015]. Improvements in sequencing technologies have led to an explosion of Genomics data. These billions of Terabytes (”Zettabytes”) of sequence data are raising challenges for computer scientists : data compression and storage, accessibility and distribution, development of more efficient algorithms to process these large datasets. The challenge for bioinformaticians is to keep up with these perpetual new developments to obtain biological insights from all these datasets, bridging the gap between computer scientists and experimental biologists. In just a few years, the global demand in bioinformatics skills has exploded, with several job advertisements posted every single day, solely in France[1] !

Today, it is obvious that there are not enough bioinformaticians. It has become ordinary to be approached by experimental biologists desperate to find ”someone to analyse their data”. That is when the frustration comes in, as bioinformaticians are too often considered as a mere service provider, contacted once the raw data are already produced to apply routine pipelines, regardless of the fact that most projects require customised analyses [Chang, 2015]. Frustration also comes from the lack of consensual definition of ’bioinformatician’ [Smith, 2015]. Within the spectrum of bioinformaticians, I came to consider myself as a computational (or *dry*) biologist, motivated by biological questions and using a computer as my bench. In this new Big Data era, collaboration between *wet* and *dry* biologists is becoming the new standard. Bioinformaticians should be involved early in the experimental design, and fair co-authorship on the publications should be customary. Evaluation criteria should be adapted for bioinformatician careers [Chang, 2015], acknowledging that working with multiple collaborators on very diverse biological questions is actually a sign of success rather than dispersion. The evaluation criteria need to be broadened to not only include the production of scientific software, but also recognize the maintenance of these software for the community [Singh Chawla, 2016]. Last, the frustration also comes when reading high-impact journal articles that have questionable and often unreproducible bioinformatic data analyses. During the peer-reviewing process, editors should enforce policies to ask reviewers if the manuscript should be sent to a bioinformatics specialist, similar to the policies often in place for statistics.

[1] source: www.sfbi.fr

Training in bioinformatics has become crucial in recent years. On the one hand, by providing courses and training material [Lewitter, 2006] dedicated to researchers, to alleviate the current bottleneck of sequence data analysis. It is also important to provide user-friendly computer tools to experimentalists, who have the biological expertise to analyse their data, but often lack bioinformatics skills. On the other hand, it is necessary to engage the undergraduate biology students into interdisciplinary work and computational biology, so that the next generation of biologists and clinicians will have essential bioinformatics skills [Brazas et al., 2014].

Even if this dissertation focuses on my research work, teaching takes a huge part of my activity and motivation to be associate professor. I am gladly contributing to the above-mentioned teaching aspects by (i) my engagement in the AVIESAN/IFB school of bioinformatics for researchers, as well as in various trainings for biologists (Belgium, France, Singapore), (ii) developing usable bioinformatics tools (mainly RSAT) and training users via published protocols and workshops, (iii) as vice-president of the French Society of Bioinformatics (SFBI), co-organising the first national meeting dedicated to the teaching of bioinformatics at the undergraduate level, and (iv) at ENS, teaching computational biology to all biology students, and introduce them to the current challenges of the Big Data era.

It is within this framework of transition to this Big Data era that my research contributions are situated. This dissertation tackles diverse biological questions such as the evolutionary analysis of the Hox genes family, and the study of transcriptional regulation, using biological sequence analysis approaches.

References

Brazas, M. D., Lewitter, F., Schneider, M. V., van Gelder, C. W. G., and Palagi, P. M. (2014). A Quick Guide to Genomics and Bioinformatics Training for Clinical and Public Audiences. *PLoS computational biology*, 10(4):e1003510.

Chang, J. (2015). Core services: Reward bioinformaticians. *Nature*, 520(7546):151–152.

Lewitter, F. (2006). Welcome to plos computational biology "education". *PLoS computational biology*.

Singh Chawla, D. (2016). The unsung heroes of scientific software. *Nature*, 529(7584):115–116.

Smith, D. R. (2015). Broadening the definition of a bioinformatician. *Frontiers in genetics*, 6:258.

Stephens, Z. D., Lee, S. Y., Faghri, F., Campbell, R. H., Zhai, C., Efron, M. J., Iyer, R., Schatz, M. C., Sinha, S., and Robinson, G. E. (2015). Big Data: Astronomical or Genomical? *PLoS Biology*, 13(7):e1002195.

Overview

After cellular biology studies, I specialised in bioinformatics during my Masters and moved to Belgium for my PhD. I have then worked for four years as a postdoctoral fellow, including three years in Germany thanks to an *Alexander von Humboldt* fellowship. In 2012, I was appointed associate professor at Ecole normale suprieure Paris, and affiliated for research to the Computational Systems Biology group headed by Denis Thieffry at the Institut de Biologie de l'Ecole normal supérieure (IBENS).

As a computational biologist, I have been involved in projects dealing with various biological questions, as well as diverse bioinformatics approaches. My research interests include development and evolution of metazoans, regulation of transcription and high-throughput functional genomics. In my view, presenting my work in three independent chapters best reflects these three main aspects of my research. Each chapter is organized with a separate introduction to the specific field, my work put in perspective with the state-of-the-art, and a conclusion presenting my current and future projects in this area.

Chapter 1: The immense diversity of animal morphologies and physiologies has always been fascinating for me. I developed this interest in the evolution of animal morphology by studying genes that control embryonic development. The first chapter presents my contributions to the evo-devo field, through the study of the Hox and ParaHox gene families evolution across the animal kingdom.

Chapter 2: Because Hox and ParaHox genes encode transcription factors - proteins that regulate the expression of their target genes by binding on short cis-regulatory elements in DNA - I became acquainted with methods to predict these cis-regulatory elements in genomic sequences, and participated in the development of new computer tools. The second chapter presents my contributions to the regulatory genomics field, through the development of user-friendly tools to study cis-regulatory elements, using binding motifs represented as PSSMs.

Chapter 3: The transition to the Big Data era revolutionised the regulatory genomics field, with the emergence of experimental approaches based on high-throughput sequencing, such as ChIP-seq. Like any technique based on high-throughput sequencing, this approach requires bioinformatics processing and analysis. Many tools for motif analysis could not cope with the resulting very large datasets. The third chapter presents my work on the development of motif analysis tools for high-throughput functional genomic datasets, and on the analysis of ChIP-seq and ChIP-exo datasets targeting the glucocorticoid receptor.

This thesis ends with general concluding remarks and prospects.

Chapter 1
Classification and evolution of Hox proteins

In this chapter, I will first introduce the Hox gene family within the global Homeobox superfamily, as both terms are often sources of confusion. Then, I will present the problem of classifying Hox proteins in homology groups, and the methodological aspects of Hox sequences classification. I will present HoxPred, the tool I started to develop during my PhD and further enhanced during my post-doc, which automatically classifies Hox sequences in their homology groups. I will highlight the contribution of HoxPred to novel insights in the evolutionary history of Hox genes, with a particular emphasis on the most debated questions. I will finally discuss the changes brought to the field during this transition to the Big Data era.

Some sections of this chapter have been published in the following review :

- Thomas-Chollier, M. and Martinez, P. (2016). **Origin of Metazoan Patterning Systems and the Role of ANTP- Class Homeobox Genes.** *eLS*, *John Wiley Sons Ltd, Chichester. http://www.els.net [doi: 10.1002/9780470015902.a0022852.pub2]*.

Related papers:

- Hudry, B., Thomas-Chollier, M., Volovik, Y., Duffraisse, M., Dard, A. e. l., Frank, D., Technau, U., and Merabet, S. (2014). **Molecular insights into the origin of the Hox-TALE patterning system.** *eLife*, 3:e01939.
- Thomas-Chollier, M., Ledent, V., Leyns, L., and Vervoort, M. (2010). **A non-tree-based comprehensive study of metazoan Hox and ParaHox genes prompts new insights into their origin and evolution.** *BMC evolutionary biology*, 10:73.
- Thomas-Chollier, M. and Ledent, V. (2008). **Comparative phylogenomic analyses of teleost fish Hox gene clusters: lessons from the cichlid fish Astatotilapia burtoni: comment.** *BMC Genomics*, 9:35.
- Thomas-Chollier, M., Leyns, L., and Ledent, V. (2007). **HoxPred: automated classification of Hox proteins using combinations of generalised profiles.** *BMC bioinformatics*, 8:247.

1.1 Hox and Homeobox: preventing the confusion

1.1.1 The Homeobox superfamily

The Homeobox gene superfamily encompasses genes bearing a particular 180-nucleotides sequence called homeobox, discovered in the early 80's [McGinnis et al., 1984; Scott and Weiner, 1984]. This homeobox encodes for the homeodomain, a DNA-binding domain of 60 amino acids (aa), which enable the proteins of the Homeobox superfamily to bind on very short DNA stretches and regulate target genes. Additional domains of these proteins can mediate the interaction with other proteins (cofactors), thereby modulating different target genes, and thus contributing to the fine-tuning of gene regulation. The homeobox genes are not restricted to animals, as they are found also in plants and fungi, but not in bacteria or archea. Of note, the homeodomain proteins account for about 15-30% of all transcription factors in animals (for a general review on homeodomains, refer to [Bürglin and Affolter, 2015]).

The homeobox gene superfamily can be subdivided into classes: in animals 11 classes [Holland, 2012] (ANTP, PRD, TALE, POU, CERS, PROS, ZF, LIM, HNF, CUT, and SINE) or 16 classes [Bürglin and Affolter, 2015] have been defined, depending on the degree of refinement that authors impose for their classification. P. Holland acknowledges that classification of homeobox genes based on orthology has limitations, because ancient gene duplications and gene losses are difficult to resolve, and the origin of some particular genes remain unclear, thus hampering their classification. These classes are themselves subdivided into gene families, according to sequence similarity between the homeobox genes and presence of additional sequence domains [Bürglin and Affolter, 2015; Holland, 2012]. The ANTP class is the most studied, as this class alone encompasses a large fraction of homeobox genes (47% of all homeobox genes in the fly *Drosophila melanogaster*).

ANTP genes have only been found in animals (metazoans), it is thus thought that they emerged at the root of all animals. They show remarkable diversity, with 50 gene families [Holland, 2012], that have expanded from a single protoANTP gene through tandem gene duplications. The ANTP class comprises the Hox gene families, as well as ParaHox and HoxL (Hox-linked) gene families like Mnx, Evx, Gbx, Meox, which share strong sequence similarities with the homeobox sequence of Hox genes and were likely clustered, and NKL (NK-linked) gene families [Hui et al., 2011; Ferrier, 2016].

1.1.2 Hox genes: a hundred years story

Hox genes have a hundred years story, and constitute the most famous, and yet somewhat mysterious genes in Biology. The story begins with a monstrous mutant fruitfly *Drosophila melanogaster*

isolated by C. Bridge in 1915, showing an homeotic transformation of the third thoracic segment, thus having four wings instead of two. Deciphering the molecular basis of this ”bithorax” mutant was of huge interest to understand the genetic control of developmental mechanisms. In the 1950's, E.B. Lewis conducted an extensive genetic analysis of the bithorax mutant and uncovered the bithorax complex (BX-C) of genes [Lewis, 1978]. W.J. Gehring had isolated another monstrous fruitfly bearing legs instead of antennas on its head, a phenotype resulting from a mutation in the gene Antennapedia (Antp). The Antp gene was found to be a member of a gene complex similar to BX-C, named the Antennapedia complex (ANT-C) [Kaufman et al., 1980]. The ANT-C and BX-C complexes were cloned and sequenced in the 1980's, revealing the 180-nucleotides sequence called ”homeobox”, common to these genes [McGinnis et al., 1984; Scott and Weiner, 1984]. It appeared that many more genes share this sequence and thus belong to the homeobox gene superfamily. D. Duboule uncovered that the mouse Hox genes are clustered and arranged in the same order as in the fruitfly [Duboule and Dollé, 1989]. The organisation of Hox genes in clusters was thus considered as a rule for all animals, and since the beginning of the 1990's, Hox genes from a wide range of animal species were being sequenced to better understand the evolution of the Hox clusters. It then became obvious that some organisms do not show this clustered organisation (such as the nematode *Caenorhabditis elegans*), or only partially (such as the fruitfly *Drosophila melanogaster*). The commonly accepted explanation is that they have lost this particular organisation during evolutionary time through chromosomal rearrangements and gene losses.

Hox genes have been defined in various manners (reviewed in [Ball et al., 2007]), depending on the inclusion of their organisation in clusters and spatial colinearity. I will consider hereafter a Hox gene as an ANTP class homeobox gene orthologous to one of the Hox group of vertebrates and drosophila [Miller and Ball, 2008]. The Hox groups refer to the groups of homology in which Hox genes can be classified. These groups, called paralogous groups, will be thoroughly described in section 1.2.

1.1.3 The sister family of ParaHox genes

ParaHox genes are members of the Gsx, Pdx/Xlox and Cdx homeobox genes families. Similarly to the Hox genes, they are organised into a gene cluster in some animals. It is thought that the Hox and ParaHox gene clusters originated by duplication of a single ancestral 'protoHox' cluster of 2-4 genes [Brooke et al., 1998]. The timing of this duplication, and the exact gene content of this protoHox cluster nevertheless remains elusive (see [Quiquand et al., 2009; Thomas-Chollier et al., 2010] for various alternative scenarios, and [Ferrier, 2015] for a recent review including mechanistic aspects).

1.2 Classification of Hox proteins

1.2.1 Hox homology groups

A long history of tandem duplications of Hox genes have generated the Hox clusters found in living organisms. Some duplications have occurred a long time ago in a putative ancestor, while some duplications appear to be more recent, and are thus restricted to a given taxonomic group. Besides, individual Hox genes have also been lost in various species. The exact evolutionary history of Hox genes is thus difficult to decipher, and that is why classification of individual Hox genes into homologous families is necessary. This classification is intrinsically linked to the organisation of Hox genes into clusters.

Hox genes in the fruitfly *Drosophila melanogaster* are organised into a split cluster, while mammals have four Hox clusters, located on different chromosomes (Fig. 1.1A). In the mouse, 39 members of the Hox gene family have been found, organised on the HoxA, HoxB, HoxC and HoxD

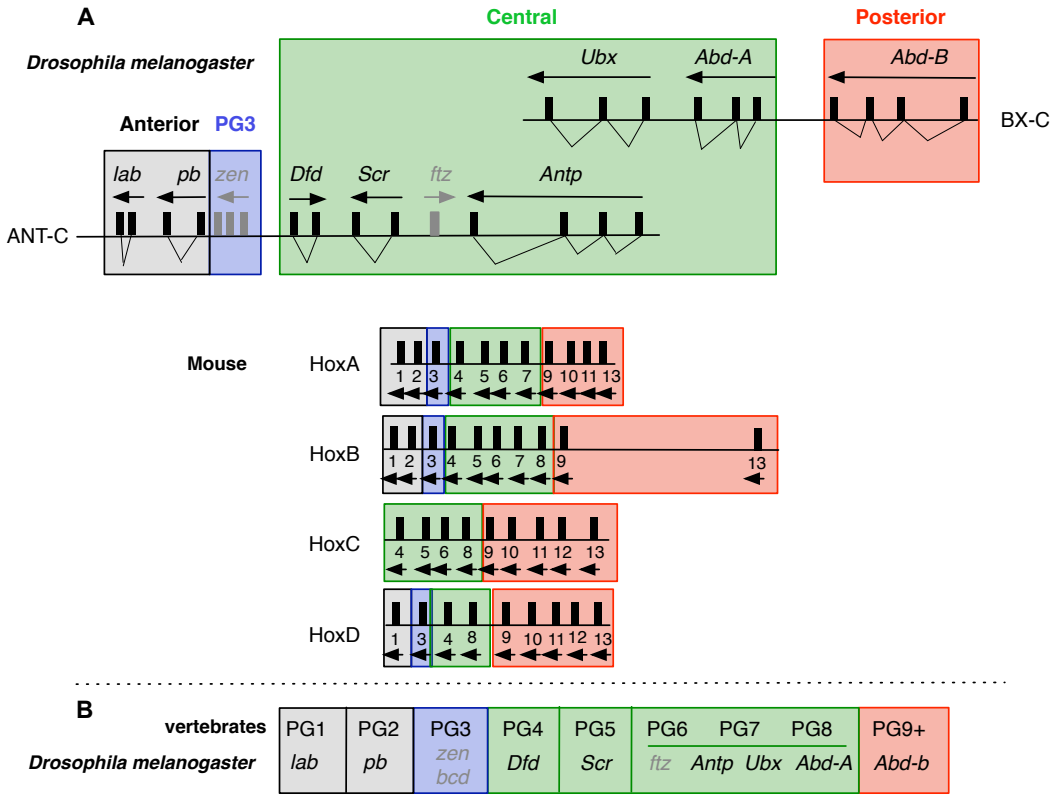

Figure 1.1: Hox gene homology groups. A. The four broad groups of classification (Anterior, PG3, Central and Posterior) mapped onto the Hox clusters of mouse and *Drosophila melanogaster*. The broad groups of homology are depicted with colored boxes. Shaded genes are derived from Hox genes, but are not true Hox genes. Mouse Hox genes are numbered according to their paralogous group; mouse clusters do not contain genes from the PG14. The representation of Hox genes clusters includes structural details, relative distances and a clear separation of the *Drosophila melanogaster* ANT-C and BX-C complexes as advocated in [Duboule, 2007], to highlight structural differences. **B.** Correspondence between the vertebrate PGs and protostome homology groups, represented by *Drosophila melanogaster* gene names. There is no clear direct relationships between the individual genes of PG6-PG8 genes and *ftz*, *Antp*, *Ubx* and *abd-A*.

clusters. It is thought that these clusters result from genome duplications during early vertebrate evolution, which have quadruplicated the ancestral vertebrate cluster that comprised 14 genes. These duplications were followed by mutation events, which have led to different losses of Hox genes within each cluster (reviewed in [Pascual-Anaya et al., 2013]).

Hox genes can be classified in homology groups, based on sequence similarity, as well as their position in the cluster. In vertebrates, Hox genes fall into one of the 14 known Paralogous Groups (PGs) [Scott, 1993; Ferrier, 2004]. For example, the mouse HoxA1, HoxB1 and HoxD1 genes (there is no HoxC1 gene in the mouse) belong to the paralogous group 1 (PG1). By homology, the *Drosophila melanogaster* labial gene can be classified into PG1 as well (Fig. 1.1B). When no clear homology to these vertebrate groups may be found, Hox genes can be classified into broader classes (Anterior, PG 3, Central and Posterior) [Finnerty and Martindale, 1998] (Fig. 1.1A).

1.2.2 Classification methods

My work has been focused on the methodological aspects of Hox sequence classification. I will first comment on the nature of available Hox sequences, highlighting how it may affect classification and our understanding of Hox data. I will then introduce the classification methods commonly used for Hox sequences, putting a particular emphasis on the strengths and weaknesses of each method.

The nature of Hox sequences

When this project started, Hox genes were commonly detected by Polymerase Chain Reaction (PCR) survey. It consists in using degenerated fragments of a homeodomain sequence as a probe to search for Hox genes in another organism. This low-cost technique has brought insights into the Hox content of many organisms, but it has intrisic weaknesses that are worth mentioning. First, only a very small fragment of the protein was usually sequenced (often restricted to the 60 aa of the homeodomain, sometimes only the most central 25 aa of the homeodomain), which hampered the assignment to an homology group. Second, because less-conserved Hox genes were not detected by PCR survey, this method did not ensure that the complete Hox content of an organism was revealed. Third, PCR fragments did not provide information on the organisation of the cluster. This is why efforts were made to sequence larger genomic fragments encompassing the complete Hox cluster (e.g. [Cameron et al., 2006; Hoegg et al., 2007]), later replaced by complete genome sequences, offering the most comprehensive view of the Hox clusters, especially for species with disintegrated clusters like the sea anemone *Nematostella vectensis* [Ryan et al., 2007].

Phylogenetic trees

Phylogenetic tree reconstruction is widely used to classify new Hox sequences in their homology groups. The underlying principle is to compile a collection of reference sequences, in addition to the sequences to classify. By analysing how the new sequences group with the reference sequences on the tree, it is possible to decipher the relationships between these sequences and classify them. This technique nevertheless requires manual work, and classification is highly dependent on the phylogenetic reconstruction method, leading to conflicting results [Ryan et al., 2007]. This well-known problem is a direct consequence of the short size and very weak phylogenetic signal of the sequences that can be aligned (usually restricted to the conserved homeodomain) [Kourakis and Martindale, 2000].

Hox signatures

The concept of Hox signatures, also known as 'characteristic residues' or 'diagnostic residues', has been pioneered by [Sharkey et al., 1997], and further extended by [Telford, 2000]. The underlying idea is that, at some positions, some amino acids that are exclusively found in a specific homologous group (e.g. the position pointed by an arrow have a Methionine residue exclusively in the Ubx sequences, Fig.1.2). These positions thus contain 'diagnostic residues' for a given homologous group.

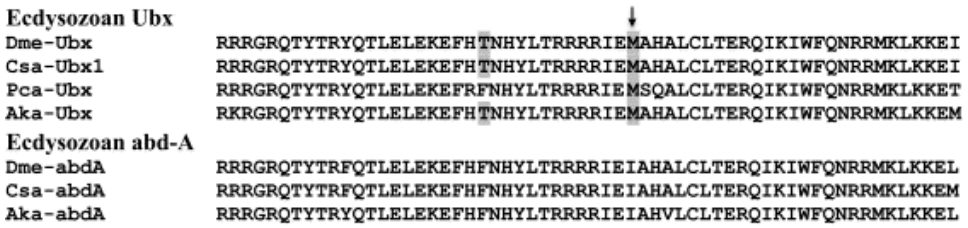

Figure 1.2: Examples of Hox genes signatures. Alignment of protostomes Hox sequences from [Balavoine et al., 2002]. The arrow indicates a diagnostic residue for Ubx sequences. Refer to Fig. 1.5 to visualize the position of ecdysozoans in the animal species tree. Species abbreviations are: Dme *Drosophila melanogaster*, Csa *Cupiennius salei*, Pca *Priapulus caudatus*, Aka *Acanthokara kaputensis*, Hro *Helobdella robusta*, Pvu *Patella vulgata*, Hme *Hirudo medicinalis*, Lan *Lingula anatina*, Pni *Polycelis nigra*.

Hox signatures bring significant information for classification of Hox sequences. Unfortunately, this manual approach is laborious, and some signatures are difficult to define when based on a combination of positions [Sarkar et al., 2002]. As more sequences are available, signatures are also susceptible to change. Two projects had addressed the question of Hox signatures with bioinformatics approaches. The first one [Sarkar et al., 2002] aimed at discovering the signatures and then using them as classification rules. We showed that it lacked accuracy [Thomas-Chollier et al., 2007]. The second one [Ogishima and Tanaka, 2007] aimed at discovering signatures - without classification purposes - in regions outside the homeodomain, thereby preventing its use

for most Hox sequences. There was thus no automated classification method for Hox proteins, which was problematic considering the ever-growing amount of sequences to analyse. It is in this context that I developed HoxPred (detailed below in Section 1.2.3).

Sequence similarity scores

Another automated method is to classify Hox sequences based on the highest similarity score to an annotated sequence (e.g. using BLAST-based approaches). While such approaches are fine for a first rough estimation, they fail to distinguish between highly similar homeodomain sequences. Recently, an approach based on all-against-all pairwise sequence similarity, followed by a clustering and specific visualisation of this pairwise sequence similarity (CLANS) has been used to provide a large-scale classification of Hox sequences [Hueber et al., 2010]. This approach is able to use regions flanking the homeodomain and full-length sequences. As the authors haven't compared their classification to our preceding work, I will discuss their results in Section 1.3.

1.2.3 HoxPred: a motif-based approach to classify Hox sequences

HoxPred is a Hox-dedicated computer program designed to classify Hox protein sequences, without phylogenetic reconstructions. The requirements were as follows:

- scale with the increasing amount of sequence data to classify (or re-classify)
- target the predominant source of data, namely the homeodomain region
- be able to discriminate among these highly-conserved sequences
- process many sequences in a small amount of time
- be accessible through both user-friendly and programmatic interfaces

The method was described and evaluated in [Thomas-Chollier et al., 2007] ; Figure 1.3 illustrates the general approach. The underlying principle is an extension of the Hox signatures. However, instead of attempting to explicitly discover the few key positions that would define a given homology group, the homeodomain is considered in its entirety as a motif, and described as a generalised profile (Fig.1.3A). Optimal combinations of such profiles allow the classification of sequences, through a supervised classification approach (Fig. 1.3B) in which discriminant functions are trained to assign sequences to predefined homology groups (Fig.1.3C). This technique thus differs from pattern searching techniques where a sequence either matches or not a given pattern that describes qualitatively a motif. The discriminant functions of HoxPred moreover allows the use of the information of multiple profiles, which increases the accuracy of the predictions [Thomas-Chollier et al., 2007].

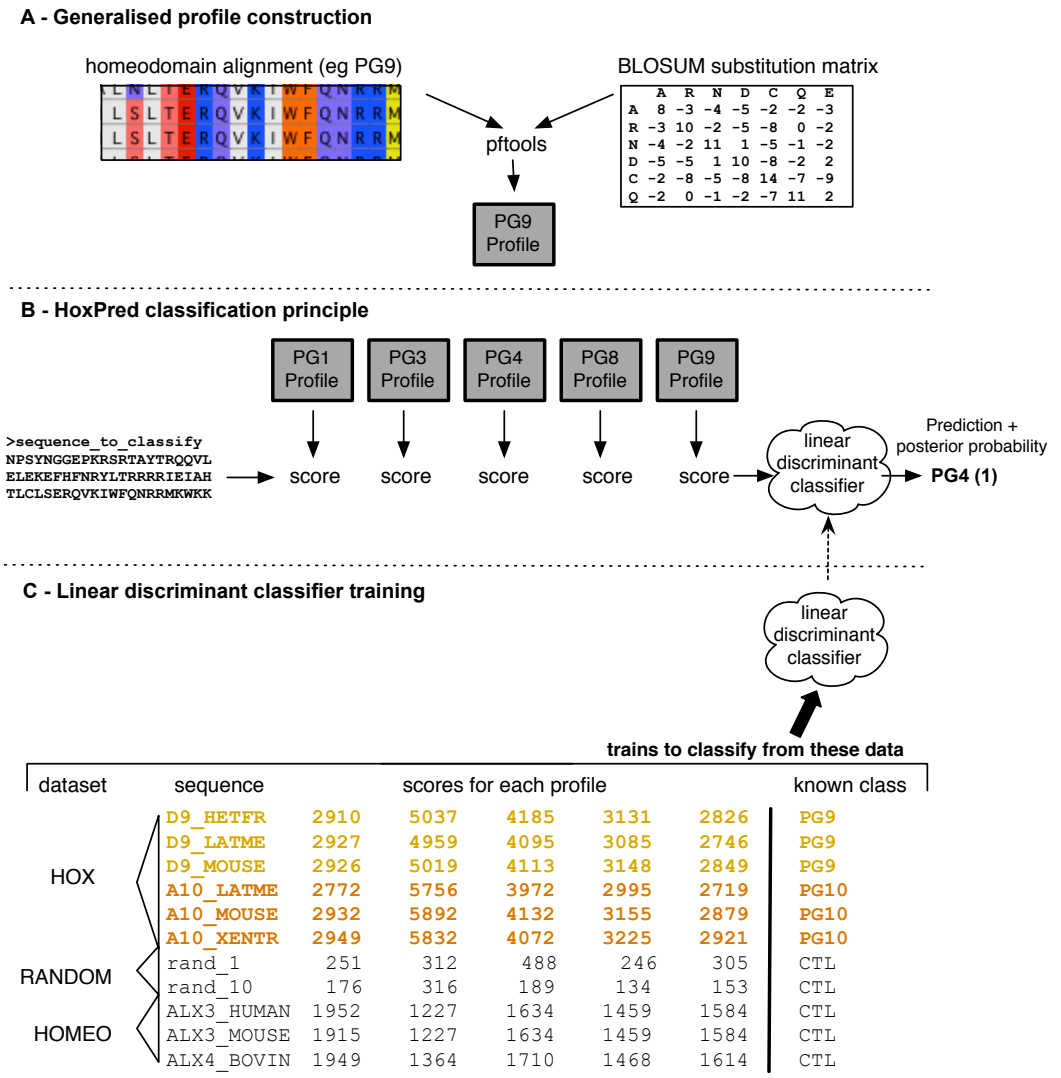

Figure 1.3: HoxPred classification approach. from [Thomas-Chollier et al., 2010]. **A.** Generalised profile construction. A multiple alignment is built from a set of non-redundant homeodomain sequences that belong to a given homology group (PG9 for this illustration). This alignment then serves to generate the corresponding generalised profile. This profile is a scoring matrix that allows to assign a score to a sequence, based on its similarity with the profile. Contrary to more simple pattern search technique, a profile can provide scores for residues that were not originally found at a given position of the motif. These scores are residue-specific, and extrapolated by using a substitution matrix when building the profile. **B.** HoxPred classification principle. The sequence to classify is scored by an optimal combination of profiles. The resulting vector of scores then serves as input to a discriminant function that has been previously trained to classify such a vector of scores into a specific class (eg PG4). **C.** Linear discriminant classifier training. The training phase aims at generating the discriminant function. The training dataset comprises sequences for which the class is known. They can be HOX, RANDOM or HOMEO (non-hox homeobox) sequences. All sequences are scored by the profiles, so that each sequence is represented by a vector of scores. The classifier is then trained to classify such vector of scores into their associated class (specified on the right). CTL is the control class.

Originally designed for vertebrate Hox sequence classification, it has proven successful in clarifying the evolutionary history of the *HoxC1a* genes in teleost fish [Thomas-Chollier and Ledent, 2008]. HoxPred was later extended to study the Hox and ParaHox sequences at the scale of metazoans [Thomas-Chollier et al., 2010]. This enabled the first large-scale study of Hox genes, simultaneously investigating 310 metazoan species accounting for more than 10,000 homeodomain

genes. This study addressed several fundamental and unsolved questions regarding the origin and evolution of the Hox and ParaHox genes, as detailed below.

1.3 New insights into the evolution of Hox genes in metazoans

In this section, I will highlight the contribution of HoxPred to novel insights in the evolutionary history of Hox genes. This requires an outline of the framework of the animal phylogeny. Figure 1.5 depicts the (simplified) evolutionary relationships between organisms of the animal kingdom (Metazoa). I will present the views obtained with non-tree based classifications (HoxPred and CLANS) on the most debated questions.

1.3.1 The uncertain origin of deuterostome Posterior genes

Posterior Hox genes of deuterostomes such as cephalochordates (amphioxus), urochordates (sea squirt), and ambulacraria (echinoderms and hemichordates), can not be confidently related to specific vertebrate PG using phylogenetic analyses (reviewed in [Pascual-Anaya et al., 2013]). It has been therefore proposed that the blurred relationships between Hox Posterior genes would be explained by an accelerated evolution rate of these genes, a process called 'Deuterostome Posterior Flexibility' [Ferrier et al., 2000]. An alternative hypothesis suggests multiple independent duplications to shape the posterior portion of the Hox clusters [Ferrier et al., 2000]. HoxPred classifications enabled to propose a global model for Posterior genes evolution in bilaterians (Fig.1.4)[Thomas-Chollier et al., 2010].

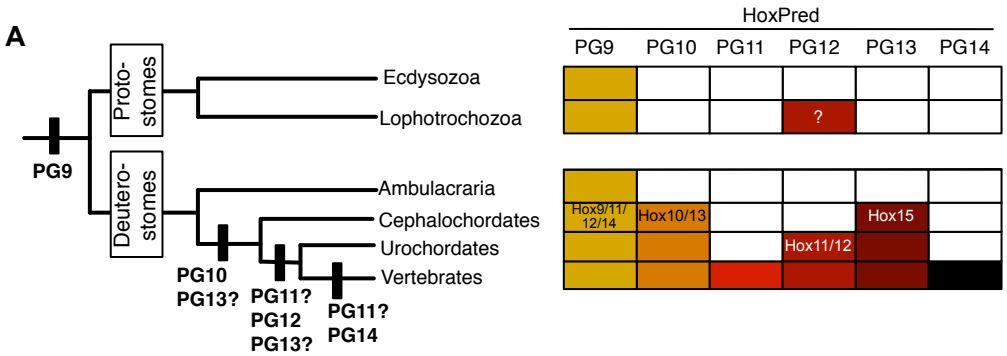

Figure 1.4: Models for the evolution of Posterior Hox genes in bilaterians. from [Thomas-Chollier et al., 2010]. The predicted PGs for each phylogenetic group are indicated with colors in the tables. Inside these tables, the names of the genes are indicated when HoxPred predictions differ from their current annotation. The possible emergence of individual PGs are indicated on the schematic tree with vertical bars (only the PG content is considered, not the actual number of genes belonging to each PG, i.e. lineage-specific duplication and losses of individual genes are not indicated). Given that both protostomes and deuterostomes have PG9 predictions, it seems that a *Hox9* gene was already present in Urbilateria. PG10 would have emerged in deuterostomes, in the lineage leading to chordates. After the divergence of cephalochordates, the lineage leading to urochordates and vertebrates would have acquired PG12. PG14 appeared in vertebrates. With respect to PG11, this group could have emerged either before or after the split between urochordates and vertebrates. Considering that both *Ciona intestinalis* and *Oikopleura dioica* have disintegrated clusters and likely miss PGs, we cannot exclude a possible loss of PG11 in urochordates. The emergence of PG13 is uncertain due to the prediction of the amphioxus *Hox15* gene as PG13. It could either be early in the chordate lineage, or in the last common ancestor of urochordates and vertebrates.

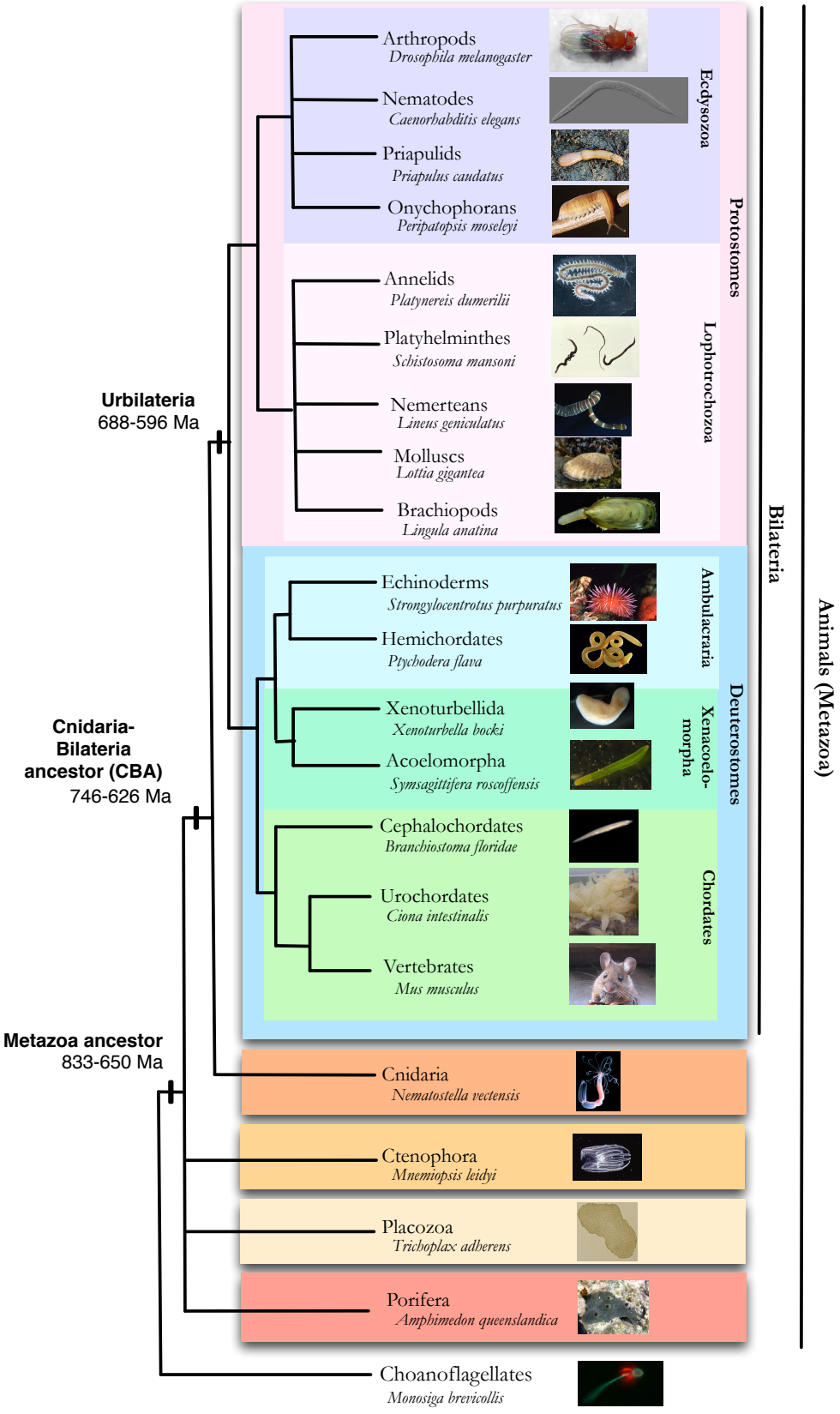

Figure 1.5: Simplified consensual animal phylogeny. The tree depicts the evolutionary relationships between living animals. Choanoflagellates are added as outgroup. Species names written in italics are given as examples for each phylum. Putative ancestors are indicated. Divergence time estimations (Million years ago (Ma)) are from [dos Reis et al., 2015]. Position of basal metazoans and Xenacoelomorpha are controversial, see main text.

Our analysis of HoxPred assignments favors the hypothesis of multiple independent duplications over the 'Deuterostome Posterior Flexibility' hypothesis alone. In particular, HoxPred assigned all amphioxus Posterior genes to PG9 and PG10, with the exception of *Hox15*, predicted as PG13, suggesting that the amphioxus *Hox11-14* genes would have arisen from duplications of *Hox9*- and *Hox10*-like genes, independent of those which produced the vertebrate PG11 to PG14 Posterior Hox genes. Results with CLANS [Hueber et al., 2010] are coherent for *Hox9-12*, classified as PG9/10, but *Hox13* and *Hox14* appear more similar to PG11-13. *Hox15* is considered to be specific to amphioxus, rather than classified as PG13. These results thus point to a mixture of 'Deuterostome Posterior Flexibility' and independent duplications. A recent review, integrating phylogenetic tree based, HoxPred and CLANS results, concludes on an ancestral chordate with three Hox groups : PG9/10, PG11/12 and PG13/14 [Pascual-Anaya et al., 2013]. On a side note, this review supports our suggestion to rename some Hox genes [Thomas-Chollier et al., 2010].

1.3.2 The bilaterian Central genes enigma

Phylogenetic approaches fail to decipher the relationships between the three very similar homeodomain sequences of the Central groups PG6-PG8, and it is well-known that the position of the gene in the cluster is not a decisive criterion, because of inversions, duplications or gene loss [Balavoine et al., 2002]. In particular, the evolutionary relationships between the protostome Central genes, but also between the deuterostome Central genes remain unclear. Based on HoxPred results, we proposed a possible evolutionary scenario with PG6 and PG7 already present in the bilaterian ancestor [Thomas-Chollier et al., 2010] (Fig.1.6). Sequences outside the homeodomain (sometimes called 'para-peptide') can be important to classify the central Hox genes [Balavoine et al., 2002]. Interestingly, the CLANS approach performed better when adding flanking regions than using the homeodomain sequence only [Hueber et al., 2010], prompting Hueber *et al.* to perform a more detailed analysis of the Central gene classification [Hueber et al., 2013].

In line with HoxPred results, they found that (i) protostome *Antp* sequences cluster with vertebrate *Hox7* sequences, (ii) ambulacrarian and amphioxus Central sequences would derive from independent duplications of an ancestral PG7 gene, and (iii) PG8 is restricted to vertebrates. Hueber *et al.* consequently also conclude that PG7 would be ancestral to all bilaterians. However, their result do not support PG6 as ancestral, but rather specific to vertebrates. As of today, many full-length Hox sequences are available ; extending HoxPred with regions larger to the homeodomain would probably improve the classification of Central genes.

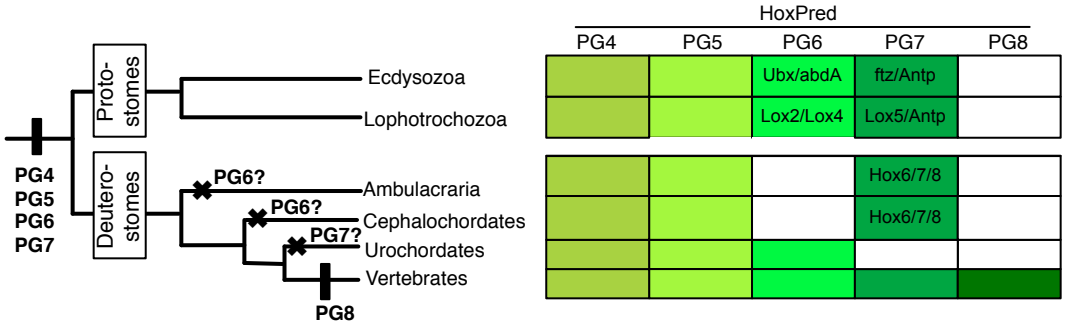

Figure 1.6: Models for the evolution of Central Hox genes in bilaterians. from [Thomas-Chollier et al., 2010]. The predicted PGs for each phylogenetic group are indicated with colors in the tables. Inside these tables, the names of the genes are indicated when HoxPred predictions differ from their current annotation. The possible emergence of individual PGs are indicated on the schematic tree with vertical bars (only the PG content is considered, not the actual number of genes belonging to each PG, i.e. lineage-specific duplication and losses of individual genes are not indicated). Four Central PGs were present in Urbilateria (PG4, PG5, PG6 and PG7). PG6 and PG7 would have been independently lost within deuterostomes. PG8 emerged in vertebrates.

1.3.3 The Cnidarian Hox genes controversy

The exact Hox content of the sea anemone *Nematostella vectensis* gave rise to a controversy, mostly because the classification of these cnidarian genes relative to bilaterian Hox homologous groups was highly dependent on the phylogenetic reconstruction method (reviewed in [Moreno and Martinez, 2010]). Although some studies have challenged the notion that cnidarians have true Hox genes [Quiquand et al., 2009], experts globally agree on the presence of seven Hox genes dispersed in the *N. vectensis* genome, including members of the anterior group. Two non-anterior genes are particularly difficult to classify, namely *anthox1* and *anthox1a*, initially classified as Central/Posterior [Ferrier and Holland, 2001; Ryan et al., 2006], and then as Posterior [Ryan et al., 2007; Quiquand et al., 2009], cnidarian-specific [Chourrout et al., 2006] , cnidarian-specific posterior subgroups [Chiori et al., 2009] and even non-Hox [Kamm et al., 2006]. HoxPred classification as Central [Thomas-Chollier et al., 2010] is thus in agreement with the non-anterior classification.

Two independent approaches support the presence of *bona fide* Hox genes in this organism. First, synteny analyses between *N. vectensis* and bilaterian genomes uncovered Hox and ParaHox loci [Hui et al., 2008]. Second, functional analyses on two Hox proteins in *N. vectensis* have revealed that they form complexes with Pbx, a major Hox cofactor in bilaterians [Hudry et al., 2014]. In addition, the complex formed by Pbx with the non-anterior *anthox1a* protein binds on the same DNA sequences that are bound by Central Hox proteins of bilaterians [Hudry et al., 2014], which provide functional evidence supporting the HoxPred classification in Central class for this gene.

Studies in other cnidarian species fail to provide a clear-cut view for the non-anterior Hox-like genes. Three cnidarian-specific classes (CnoxA, CnoxB and CnoxC) have been defined [Chiori et al., 2009; Reddy et al., 2015], which could derive from a Central or Posterior ancestral Hox gene

present in the cnidarian-bilaterian common ancestor. Altogether, a consensus view is emerging: the various cnidarians studies show unexpected diverse repertoires of Hox-like genes, some of which have arisen from lineage-specific, and perhaps cnidarian-specific, duplication events. It may be thus necessary to add these cnidarian-specific classes to HoxPred. Functional analyses, as performed in *N. vectensis*, may be the key to correctly classify these genes not only based on sequence similarities, but also based on their function and interaction modes with co-factors.

1.3.4 Evolutionary relationships between Hox and ParaHox

The ParaHox cluster of genes has long been supposed to be the sister cluster of the Hox cluster, with the *Gsx*, *Xlox* and *Cdx* genes corresponding to the Anterior, PG3 and Posterior groups, respectively [Brooke et al., 1998]. Using HoxPred, we revisited how the ParaHox genes can be related to the Hox genes [Thomas-Chollier et al., 2010]. Our results grouping *Gsx* and *Xlox* to PG3 do not support the traditional grouping of *Gsx* with PG1, but are consistent with a phylogenetic analysis that regroups *Gsx* and *Xlox* into a PG2/PG3 group [Quiquand et al., 2009]. *Cdx* genes were consistently predicted in the Central group, rather than in the Posterior group. Interestingly, CLANS has been very recently used to revisit this question as well [Hueber and Frickey, 2016]. In global agreement with our results, this study confirmed the relationships between *Cdx* and the Central group, *Xlox* and PG3, and *Gsx* and PG2/PG3.

1.3.5 The scarce but increasing knowledge on basal metazoans

The metazoan basal groups (close to the root of the animal phylogenetic tree) are the phyla Porifera (sponges), Ctenophora (comb jellies) and Placozoa (trichoplax) (Fig. 1.5). Their phylogenetic relationships have been, and are still, under debate due to methodological considerations, from the selection of taxa and characters, to the use of different phylogenetic algorithms [Whelan et al., 2015]. Figure 1.7 schematizes the major alternative positions that have been given to these groups over time.

Uncovering the homeobox gene repertoire of early-branching metazoans has been a matter of many converging research efforts initiated in the early 90's. It was already clear at that time that deciphering the origin and evolution of this gene family could bring insights into the evolution of developmental processes, and consequently the emergence and diversity of morphological novelties.

In ctenophores, the complete genomes of *Mnemiopsis leidyi* [Ryan et al., 2010], *Pleurobrachia bachei* and ten ctenophore transcriptomes [Moroz et al., 2014] corroborate the previously-reported absence of Hox and ParaHox genes in this phylum. Of note, this last study mentions a *Cdx* gene in *M. leidyi*, but our reanalysis of the datasets rather clusters this gene with the ANTP HoxL subclass.

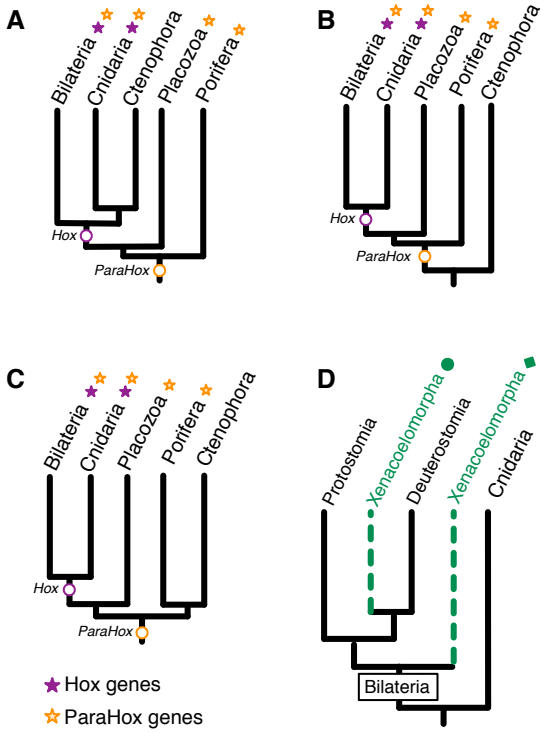

Figure 1.7: Phylogenetic relationships among the early branching metazoan phyla. from [Thomas-Chollier and Martinez, 2016]. The alternate topologies are based on the studies of (a) Porifera sister to the remaining metazoans e.g. [Philippe et al., 2009], (b) Ctenophora as sister to the remaining metazoans e.g. [Moroz et al., 2014], (c) Ctenophora + Porifera as sister groups [Ryan et al., 2013], and (d) Xenacoelomorpha (dotted green branches) as sister to the remaining bilaterians (diamond; [Hejnol et al., 2009]) or as members of Deuterostomia (circle; [Philippe et al., 2011]). The presence of Hox and ParaHox genes is indicated by purple and orange stars, respectively. The deduced parsimonious emergence of these gene families in the frame of each tree topology is indicated with circles.

In the enigmatic phylum Placozoa, *Trichoplax adhaerens* includes one ParaHox-related gene, *Trox-2*, classified as a ParaHox Gsx (including with HoxPred [Thomas-Chollier et al., 2010]). *Trichoplax adhaerens* might be a secondarily simplified organism that would have lost Hox/ParaHox genes; this *Gsx* gene would thus be the remnant of a wider set of Hox/ParaHox genes present in the ancestors [Monteiro et al., 2006]. This scenario is substantiated by the proposed basal position of Ctenophores or Porifera (rather than this phylum) in the species tree (Fig.1.7), and the recent unravelling of both Hox and ParaHox 'ghost loci' in *T. adhaerens*, using genomic synteny and Monte Carlo-based simulations [Mendivil Ramos et al., 2012]. These 'ghost loci' are defined independently of phylogenetic reconstructions, and provide evidence that the genomic region in which Hox and ParaHox genes are located in other animals is also present in this placozan genome.

Regarding Porifera, analyses of the complete genome sequence of the demosponge *Amphimedon queenslandica* concluded on the absence of Hox/ParaHox genes in this phylum [Larroux et al., 2007]. As Hox and ParaHox "ghost loci" were also predicted in *A. queenslandica* [Mendivil Ramos et al., 2012], the long-thought apparent absence of Hox/ParaHox genes in sponges might simply result from a small sampling effect. Indeed, a *Cdx*-like gene has recently been uncovered in two calcareous sponges, and 'ghost loci' for the Hox genes have been predicted, supporting the view that the absence of Hox/ParaHox genes is the result of a lineage-specific loss in some sponges like *A. queenslandica* [Fortunato et al., 2014]. This crucial finding revolutionised our view on the

emergence of the Hox/ParaHox genes, suggesting that these genes arose directly within the first metazoans (in the hypothesis of a very basal position of Porifera as in Figure 1.7A,C) or within the early branches (in the hypothesis of a less basal position of Porifera as in Figure 1.7B). Of note, this *Cdx*-like gene is very divergent, and classified as 'control' by HoxPred. Altogether, the emerging view points to sponges having a more complex gene repertoire than previously thought, with distinct patterns of gene family losses. In summary, the Hox/ParaHox genes predate the sponges and have not been found (yet ?) in ctenophores.

1.3.6 Towards a definite position of Xenacoelomorpha as deuterostomes ?

Xenacoelomorpha is a clade of worm-like marine animals (Fig.1.5), whose position within the bilaterian phylogenetic tree still remains enigmatic and debated, currently placed within the Deuterostomia or at the base of Bilateria (Fig.1.7D) (reviewed in [Nakano, 2015; Haszprunar, 2015]). This group comprises xenoturbellids, acoels and nemertodermatids.

Based on PCR surveys, acoels and nemertodermatids have a small number of Hox genes: one anterior, one central and one posterior (possibly duplicated in certain species by lineage-specific events) [Cook et al., 2004; Jimenez-Guri et al., 2006; Moreno et al., 2009] (Fig.1.8). In addition, a PG3 member is hypothetically present in acoels [Baguñà and Riutort, 2004], but no recent publication has confirmed this. This small Hox gene content notably resembles the one of *Xenoturbella bocki* [Fritzsch et al., 2008], although Fritzsch *et al.* concluded that there was no particular sequence similarity between Hox genes of these two groups.

At the end of my PhD, I re-analysed these sequences (Thomas-Chollier, 2008, PhD thesis), incorporating a larger sampling of bilaterian Hox genes. This revealed that some reported Hox genes were contaminations from mollusc DNA. More interestingly, and in support of the grouping of the Xenacoelomorpha species, this new analysis revealed that *Xenoturbella* and acoelomorph Hox sequences are more similar among themselves, than to any other phylogenetic group.

Unravelling the entire Hox content of a Xenacoelomorpha species is necessary to verify and complete these findings. To this end, I am involved in a consortium that has now produced genomes and some transcriptomes for five Acoelomorpha species and *Xenoturbella bocki*. This consortium includes among others Max Telford, Albert Poutska, Hervé Philippe and Pedro Martinez, with whom I recently wrote a review [Thomas-Chollier and Martinez, 2016]. My contribution is to analyze the Hox/ParaHox genes (and extend to the homeobox superfamily). This unpublished and ongoing work confirms various contaminations in the previous studies (Fig.1.8). In the context of a basal position in bilaterians (Fig.1.9A), my results suggest that PG2 (and possibly PG3) emerged after the divergence between Xenacoelomorpha and other bilaterians. Alternatively, in the context of a position closer to Deuterostomes (Fig.1.9B), these results suggest that PG2 and PG3 would

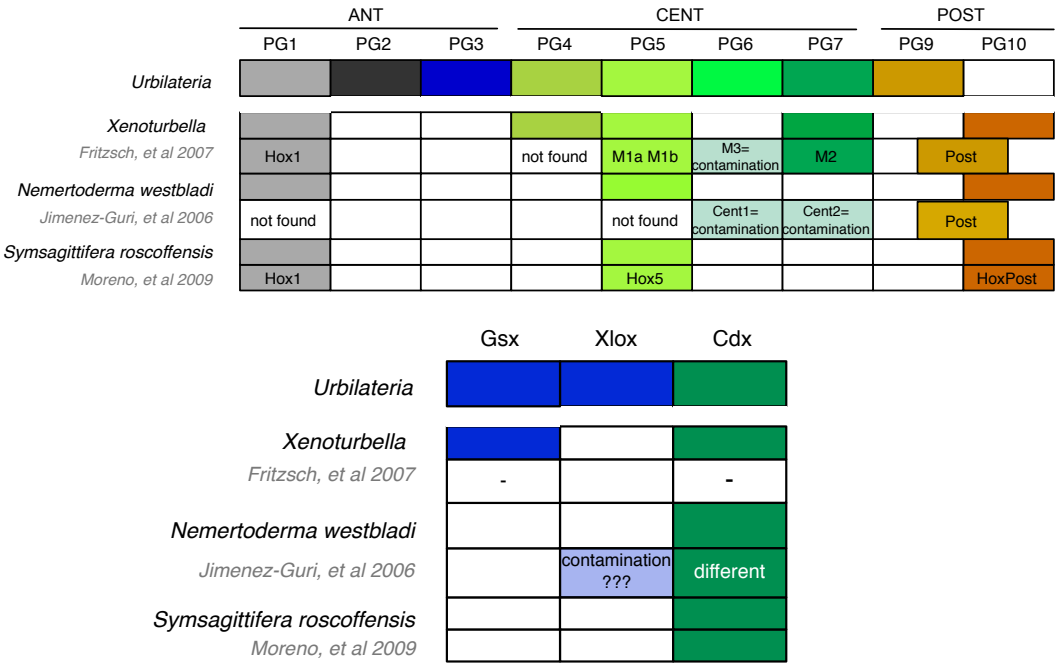

Figure 1.8: Hox and ParaHox genes in Xenacoelomorpha. unpublished. The predicted PGs for each species are indicated with colors in the tables. Inside these tables, the names of the genes from former studies are indicated. Rows starting with just the name of the species correspond to the genes I found from the complete genomes. The Urbilateria repertoire is from [Thomas-Chollier et al., 2010], following the hypothesis of xenaceolomorphs being secondarily-simplified deuterostomes.

have been lost at the base of this group. In both cases, *Xlox* would have been lost at the base of this group. The HoxPred classification of some Posterior genes in PG10 would support a position within Deuterostome, according to the model from Figure1.4. In addition, it has been shown that the acoel *Symsagittifera roscoffensis* Hox4/5 sequence has a suite of 6 residues downstream the homeodomain that is only found in Deuterostomes [Deutsch, 2008].

In Xenacoelomorpha, the small number of Hox genes is alternatively interpreted as a derived state [Deutsch, 2008] or as evidence for a basal position in bilaterians [Haszprunar, 2015] (Fig.1.7D). As mentioned above, the phylogenetic position of Acoelomorpha is an ongoing debate, in which the interpretation of this reduced Hox number is a key argument. The Hox and ParaHox repertoire is more complex in *Xenoturbella bocki* than in the studied acoels. It contains PG4 and PG7 genes (consistent with the above-mentioned hypothesis of PG7 being ancestral in bilaterians) and a *Gsx* gene (Fig.1.8). These findings thus invalidate the hypothesis that the small number of Hox genes in acoels represents the ancestral bilaterian repertoire, as at least PG4, PG7, *Gsx* and *Xlox* would have been lost in these organisms (Fig.1.9). The team supporting the basal bilaterian position has very recently published a phylogenomic study with 15 xenacoelomorphs transcriptomes [Cannon et al., 2016]. Interestingly, the argument about Hox genes is not mentioned at all in this study. Finally, the Hox genes are not organized into a cluster in the acoel *S. roscoffensis*

[Moreno et al., 2009]. A reduced number of Hox genes associated with a disintegrated cluster organisation would thus support acoels (and possibly all xenacoelomorphs) being secondarily-simplified organisms, regardless of their definite position within bilaterians.

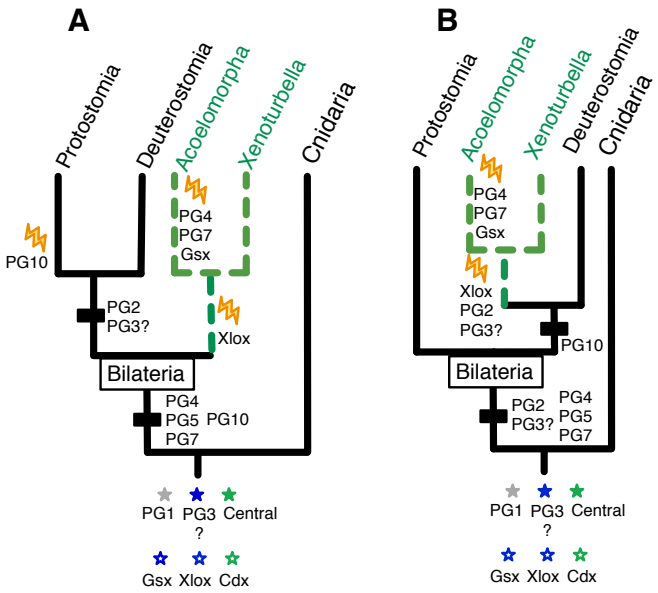

Figure 1.9: Interpretation of the Hox/ParaHox repertoires in the alternative positions of Xenacoelomorpha. unpublished. The putative Hox/ParaHox content of the cnidarian-bilaterian ancestor is indicated at the bottom. The deduced parsimonious emergence of PGs in the frame of each tree topology is indicated with black bars. The lightening indicates gene loss. **A**. Xenacoelomorpha (dotted green branches) as sister to the remaining bilaterians or **B** as members of Deuterostomia.

1.4 Conclusion : the rise and fall of Hox genes

This concluding remark is inspired by D. Duboule, who wrote ”The rise and fall of Hox gene clusters” [Duboule, 2007], when it became clear that well-ordered clusters were not the rule for Hox genes. In my view, the transition to the Big Data era has brought to an end the systematic quest of Hox genes across metazoans. We have shifted from a time when the only sequences available for many animals were the small PCR fragments of Hox genes, to a wealth of complete genomes and transcriptomes, for which detailed analyses of Hox contents have not been reported, even for basal metazoans (e.g. [Nichols et al., 2012; Riesgo et al., 2015]). This actually provides substantial public material for future computational studies on the evolution of Hox genes, especially in groups where small taxon sampling has shown limits to conclusions made for the whole group (e.g. sponges not having Hox/ParaHox genes). To automatically detect and classify Hox genes from full genome sequences, HoxPred remains a state-of-the-art approach, still used by several research groups.

Why have the Hox genes lost their primacy ? First, sponges and trichoplax have ParaHox -but no Hox- genes, which lead to the hypothesis that the ParaHox genes have more evolutionary constraints than Hox genes [Quiquand et al., 2009]. Second, the plasticity of Hox repertoires, prone

to gene duplications or massive gene losses as in sponges, has somewhat limited the inferences between the Hox content and the complexity of an animal. Whole genome analyses moreover provided evidence for a large diversity and a more complex gene repertoire than previously thought in the early-branching metazoans [Ferrier, 2015], thus shifting the focus to other genes (e.g. the Hox cofactors *Pbx* and *Meis* [Merabet and Galliot, 2015]) or other important biological functions in evo-devo studies (e.g. the neural system). Yet, the Big Data era has enabled tremendous discoveries related to Hox genes (e.g. the temporal dynamics of Hox clusters in vertebrates [Noordermeer et al., 2014]). Very little is still known concerning the roles of Hox/ParaHox genes in early divergent taxa, but what we have learned over the last few years suggests that this area of research has a very promising (though challenging) future ahead.

References

Baguñà, J. and Riutort, M. (2004). The dawn of bilaterian animals: the case of acoelomorph flatworms. *BioEssays : news and reviews in molecular, cellular and developmental biology*, 26(10):1046–1057.

Bailey, T., Krajewski, P., Ladunga, I., Lefebvre, C., Li, Q., Liu, T., Madrigal, P., Taslim, C., and Zhang, J. (2013). Practical guidelines for the comprehensive analysis of ChIP-seq data. *PLoS computational biology*, 9(11):e1003326.

Balavoine, G., de Rosa, R., and Adoutte, A. (2002). Hox clusters and bilaterian phylogeny. *Molecular phylogenetics and evolution*, 24(3):366–373.

Ball, E. E., de Jong, D. M., Schierwater, B., Shinzato, C., Hayward, D. C., and Miller, D. J. (2007). Implications of cnidarian gene expression patterns for the origins of bilaterality is the glass half full or half empty? *Integrative and Comparative Biology*, 47(5):701–711.

Brooke, N. M., Garcia-Fernàndez, J., and Holland, P. W. (1998). The ParaHox gene cluster is an evolutionary sister of the Hox gene cluster. *Nature*, 392(6679):920–922.

Bürglin, T. R. and Affolter, M. (2015). Homeodomain proteins: an update. *Chromosoma*.

Cameron, R. A., Rowen, L., Nesbitt, R., Bloom, S., Rast, J. P., Berney, K., Arenas-Mena, C., Martinez, P., Lucas, S., Richardson, P. M., Davidson, E. H., Peterson, K. J., and Hood, L. (2006). Unusual gene order and organization of the sea urchin hox cluster. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, 306(1):45–58.

Cannon, J. T., Vellutini, B. C., Smith, J., Ronquist, F., Jondelius, U., and Hejnol, A. (2016). Xenacoelomorpha is the sister group to Nephrozoa. *Nature*, 530(7588):89–93.

Chiori, R., Jager, M., Denker, E., Wincker, P., Da Silva, C., Le Guyader, H., Manuel, M. e. l., and Qu e innec, E. (2009). Are Hox genes ancestrally involved in axial patterning? Evidence from the hydrozoan Clytia hemisphaerica (Cnidaria). *PLoS ONE*, 4(1):e4231.

Chourrout, D., Delsuc, F., Chourrout, P., Edvardsen, R. B., Rentzsch, F., Renfer, E., Jensen, M. F., Zhu, B., de Jong, P., Steele, R. E., and Technau, U. (2006). Minimal ProtoHox cluster inferred from bilaterian and cnidarian Hox complements. *Nature*, 442(7103):684–687.

Cook, C. E., Jimenez-Guri, E., Akam, M., and Salo, E. (2004). The Hox gene complement of acoel flatworms, a basal bilaterian clade. *Evolution & Development*, 6(3):154–163.

Deutsch, J. S. (2008). Do acoels climb up the "Scale of Beings"? *Evolution & Development*, 10(2):135–140.

dos Reis, M., Thawornwattana, Y., Angelis, K., Telford, M. J., Donoghue, P. C. J., and Yang, Z. (2015). Uncertainty in the Timing of Origin of Animals and the Limits of Precision in Molecular Timescales. *Current biology*, 25(22):2939–2950.

Duboule, D. (2007). The rise and fall of Hox gene clusters. *Development (Cambridge, England)*, 134(14):2549–2560.

Duboule, D. and Dollé, P. (1989). The structural and functional organization of the murine HOX gene family resembles that of Drosophila homeotic genes. *The EMBO Journal*, 8(5):1497–1505.

Ferrier, D. (2016). Evolution of homeobox gene clusters in animals: the Giga-cluster and primary versus secondary clustering. *Frontiers in Ecology and Evolution*.

Ferrier, D. E., Minguillón, C., Holland, P. W., and Garcia-Fernàndez, J. (2000). The amphioxus Hox cluster: deuterostome posterior flexibility and Hox14. *Evolution & Development*, 2(5):284–293.

Ferrier, D. E. K. (2004). Hox genes: Did the vertebrate ancestor have a Hox14? *Current biology*, 14(5):R210–1.

Ferrier, D. E. K. (2015). The origin of the Hox/ParaHox genes, the Ghost Locus hypothesis and the complexity of the first animal. *Briefings in functional genomics*.

Ferrier, D. E. K. and Holland, P. W. H. (2001). Ancient origin of the Hox gene cluster. *Nature reviews Genetics*, 2(1):33–38.

Finnerty, J. R. and Martindale, M. Q. (1998). The evolution of the Hox cluster: insights from outgroups. *Current opinion in genetics & development*, 8(6):681–687.

Fortunato, S. A. V., Adamski, M., Ramos, O. M., Leininger, S., Liu, J., Ferrier, D. E. K., and Adamska, M. (2014). Calcisponges have a ParaHox gene and dynamic expression of dispersed NK homeobox genes. *Nature*, 514(7524):620–623.

Fritzsch, G., Böhme, M. U., Thorndyke, M., Nakano, H., Israelsson, O., Stach, T., Schlegel, M., Hankeln, T., and Stadler, P. F. (2008). PCR survey of Xenoturbella bocki Hox genes. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, 310(3):278–284.

Haszprunar, G. (2015). Review of data for a morphological look on Xenacoelomorpha (Bilateria incertae sedis). *Organisms Diversity & Evolution*.

Hejnol, A., Obst, M., Stamatakis, A., Ott, M., Rouse, G. W., Edgecombe, G. D., Martinez, P., Baguna, J., Bailly, X., Jondelius, U., Wiens, M., Muller, W. E. G., Seaver, E., Wheeler, W. C., Martindale, M. Q., Giribet, G., and Dunn, C. W. (2009). Assessing the root of bilaterian animals with scalable phylogenomic methods. *Proceedings of the Royal Society B: Biological Sciences*, 276(1677):4261–4270.

Hoegg, S., Boore, J. L., Kuehl, J. V., and Meyer, A. (2007). Comparative phylogenomic analyses of teleost fish Hox gene clusters: lessons from the cichlid fish Astatotilapia burtoni. *BMC Genomics*, 8:317.

Holland, P. W. H. (2012). Evolution of homeobox genes. *Wiley Interdisciplinary Reviews: Developmental Biology*, 2(1):31–45.

Hudry, B., Thomas-Chollier, M., Volovik, Y., Duffraisse, M., Dard, A. e. l., Frank, D., Technau, U., and Merabet, S. (2014). Molecular insights into the origin of the Hox-TALE patterning system. *eLife*, 3:e01939.

Hueber, S. and Frickey, T. (2016). Solving Classification Problems for Large Sets of Protein Sequences with the Example of Hox and ParaHox Proteins. *Journal of Developmental Biology*, 4(1):8.

Hueber, S. D., Rauch, J., Djordjevic, M. A., Gunter, H., Weiller, G. F., and Frickey, T. (2013). Analysis of central Hox protein types across bilaterian clades: on the diversification of central Hox proteins from an Antennapedia/Hox7-like protein. *Developmental Biology*, 383(2):175–185.

Hueber, S. D., Weiller, G. F., Djordjevic, M. A., and Frickey, T. (2010). Improving Hox protein classification across the major model organisms. *PLoS ONE*, 5(5):e10820.

Hui, J. H. L., Holland, P. W. H., and Ferrier, D. E. K. (2008). Do cnidarians have a ParaHox cluster? Analysis of synteny around a Nematostella homeobox gene cluster. *Evolution & Development*, 10(6):725–730.

Hui, J. H. L., McDougall, C., Monteiro, A. S., Holland, P. W. H., Arendt, D., Balavoine, G., and Ferrier, D. E. K. (2011). Extensive Chordate and Annelid Macrosynteny Reveals Ancestral Homeobox Gene Organization. *Molecular biology and evolution*, 29(1):157–165.

Jimenez-Guri, E., Paps, J., Garcia-Fernàndez, J., and Salo, E. (2006). Hox and ParaHox genes in Nemertodermatida, a basal bilaterian clade. *The International journal of developmental biology*, 50(8):675–679.

Kamm, K., Schierwater, B., Jakob, W., Dellaporta, S. L., and Miller, D. J. (2006). Axial patterning and diversification in the cnidaria predate the Hox system. *Current biology*, 16(9):920–926.

Kaufman, T. C., Lewis, R., and Wakimoto, B. (1980). Cytogenetic Analysis of Chromosome 3 in DROSOPHILA MELANOGASTER: The Homoeotic Gene Complex in Polytene Chromosome Interval 84a-B. *Genetics*, 94(1):115–133.

Kourakis, M. J. and Martindale, M. Q. (2000). Combined-method phylogenetic analysis of Hox and ParaHox genes of the metazoa. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, 288(2):175–191.

Larroux, C., Fahey, B., Degnan, S. M., Adamski, M., Rokhsar, D. S., and Degnan, B. M. (2007). The NK homeobox gene cluster predates the origin of Hox genes. *Current biology*, 17(8):706–710.

Lewis, E. B. (1978). A gene complex controlling segmentation in Drosophila. *Nature*, 276(5688):565–570.

McGinnis, W., Levine, M. S., Hafen, E., Kuroiwa, A., and Gehring, W. J. (1984). A conserved DNA sequence in homoeotic genes of the Drosophila Antennapedia and bithorax complexes. *Nature*, 308(5958):428–433.

Mendivil Ramos, O., Barker, D., and Ferrier, D. E. K. (2012). Ghost loci imply Hox and ParaHox existence in the last common ancestor of animals. *Current biology*, 22(20):1951–1956.

Merabet, S. and Galliot, B. (2015). The TALE face of Hox proteins in animal evolution. *Frontiers in genetics*, 6:267.

Miller, D. J. and Ball, E. E. (2008). Cryptic complexity captured: the Nematostella genome reveals its secrets. *Trends in genetics : TIG*, 24(1):1–4.

Monteiro, A. S., Schierwater, B., Dellaporta, S. L., and Holland, P. W. H. (2006). A low diversity of ANTP class homeobox genes in Placozoa. *Evolution & Development*, 8(2):174–182.

Moreno, E. and Martinez, P. (2010). Origin of Bilaterian Hox Patterning System. *eLS*.

Moreno, E., Nadal, M., Baguñà, J., and Martinez, P. (2009). Tracking the origins of the bilaterian Hox patterning system: insights from the acoel flatworm Symsagittifera roscoffensis. *Evolution & Development*, 11(5):574–581.

Moroz, L. L., Kocot, K. M., Citarella, M. R., Dosung, S., Norekian, T. P., Povolotskaya, I. S., Grigorenko, A. P., Dailey, C., Berezikov, E., Buckley, K. M., Ptitsyn, A., Reshetov, D., Mukherjee, K., Moroz, T. P.,

Bobkova, Y., Yu, F., Kapitonov, V. V., Jurka, J., Bobkov, Y. V., Swore, J. J., Girardo, D. O., Fodor, A., Gusev, F., Sanford, R., Bruders, R., Kittler, E., Mills, C. E., Rast, J. P., Derelle, R., Solovyev, V. V., Kondrashov, F. A., Swalla, B. J., Sweedler, J. V., Rogaev, E. I., Halanych, K. M., and Kohn, A. B. (2014). The ctenophore genome and the evolutionary origins of neural systems. *Nature*, 510(7503):109–114.

Nakano, H. (2015). What is Xenoturbella? *Zoological letters*, 1:22.

Nichols, S. A., Roberts, B. W., Richter, D. J., Fairclough, S. R., and King, N. (2012). Origin of metazoan cadherin diversity and the antiquity of the classical cadherin/β-catenin complex. *Proceedings of the National Academy of Sciences*, 109(32):13046–13051.

Noordermeer, D., Leleu, M., Schorderet, P., Joye, E., Chabaud, F., Duboule, D., and Krumlauf, R. (2014). Temporal dynamics and developmental memory of 3D chromatin architecture at Hox gene loci. *eLife*, 3.

Ogishima, S. and Tanaka, H. (2007). Missing link in the evolution of Hox clusters. *Gene*, 387(1-2):21–30.

Pascual-Anaya, J., D'Aniello, S., Kuratani, S., and Garcia-Fernàndez, J. (2013). Evolution of Hox gene clusters in deuterostomes. *BMC Developmental Biology*, 13:26.

Philippe, H., Brinkmann, H., Copley, R. R., Moroz, L. L., Nakano, H., Poustka, A. J., Wallberg, A., Peterson, K. J., and Telford, M. J. (2011). Acoelomorph flatworms are deuterostomes related to Xenoturbella. *Nature*, 470(7333):255.

Philippe, H., Derelle, R., Lopez, P., Pick, K., Borchiellini, C., Boury-Esnault, N., Vacelet, J., Renard, E., Houliston, E., Quéinnec, E., Da Silva, C., Wincker, P., Le Guyader, H., Leys, S., Jackson, D. J., Schreiber, F., Erpenbeck, D., Morgenstern, B., Wörheide, G., and Manuel, M. (2009). Phylogenomics revives traditional views on deep animal relationships. *Current biology*, 19(8):706–712.

Quiquand, M., Yanze, N., Schmich, J. u. r., Schmid, V., Galliot, B., and Piraino, S. (2009). More constraint on ParaHox than Hox gene families in early metazoan evolution. *Developmental Biology*, 328(2):173–187.

Reddy, P. C., Unni, M. K., Gungi, A., Agarwal, P., and Galande, S. (2015). Evolution of Hox-like genes in Cnidaria: Study of Hydra Hox repertoire reveals tailor-made Hox-code for Cnidarians. *Mechanisms of development*, 138 Pt 2:87–96.

Riesgo, A., Farrar, N., Windsor, P. J., Giribet, G., and Leys, S. P. (2015). The Analysis of Eight Transcriptomes from All Poriferan Classes Reveals Surprising Genetic Complexity in Sponges. *Molecular biology and evolution*, 31(5):1102–1120.

Ryan, J. F., Burton, P. M., Mazza, M. E., Kwong, G. K., Mullikin, J. C., and Finnerty, J. R. (2006). The cnidarian-bilaterian ancestor possessed at least 56 homeoboxes: evidence from the starlet sea anemone, Nematostella vectensis. *Genome Biology*, 7(7):R64.

Ryan, J. F., Mazza, M. E., Pang, K., Matus, D. Q., Baxevanis, A. D., Martindale, M. Q., and Finnerty, J. R. (2007). Pre-bilaterian origins of the Hox cluster and the Hox code: evidence from the sea anemone, Nematostella vectensis. *PLoS ONE*, 2(1):e153.

Ryan, J. F., Pang, K., Program, N. C. S., Mullikin, J. C., Martindale, M. Q., and Baxevanis, A. D. (2010). The homeodomain complement of the ctenophore Mnemiopsis leidyi suggests that Ctenophora and Porifera diverged prior to the ParaHoxozoa. *EvoDevo*, 1(1):9.

Ryan, J. F., Pang, K., Schnitzler, C. E., Nguyen, A.-D., Moreland, R. T., Simmons, D. K., Koch, B. J., Francis, W. R., Havlak, P., Smith, S. A., Putnam, N. H., Haddock, S. H. D., Dunn, C. W., Wolfsberg, T. G., Mullikin, J. C., Martindale, M. Q., and Baxevanis, A. D. (2013). The genome of the ctenophore Mnemiopsis leidyi and its implications for cell type evolution. *Science (New York, NY)*, 342(6164):1242592–1242592.

Sarkar, I. N., Thornton, J. W., Planet, P. J., Figurski, D. H., Schierwater, B., and DeSalle, R. (2002). An automated phylogenetic key for classifying homeoboxes. *Molecular phylogenetics and evolution*, 24(3):388–399.

Scott, M. P. (1993). A rational nomenclature for vertebrate homeobox (HOX) genes. *Nucleic Acids Research*, 21(8):1687–1688.

Scott, M. P. and Weiner, A. J. (1984). Structural relationships among genes that control development: sequence homology between the Antennapedia, Ultrabithorax, and fushi tarazu loci of Drosophila. *Proceedings of the National Academy of Sciences of the United States of America*, 81(13):4115–4119.

Sharkey, M., Graba, Y., and Scott, M. P. (1997). Hox genes in evolution: protein surfaces and paralog groups. *Trends in genetics : TIG*, 13(4):145–151.

Telford, M. J. (2000). Turning Hox "signatures" into synapomorphies. *Evolution & Development*, 2(6):360–364.

Thomas-Chollier, M. and Ledent, V. (2008). Comparative phylogenomic analyses of teleost fish Hox gene clusters: lessons from the cichlid fish Astatotilapia burtoni: comment. *BMC Genomics*, 9:35.

Thomas-Chollier, M., Ledent, V., Leyns, L., and Vervoort, M. (2010). A non-tree-based comprehensive study of metazoan Hox and ParaHox genes prompts new insights into their origin and evolution. *BMC evolutionary biology*, 10:73.

Thomas-Chollier, M., Leyns, L., and Ledent, V. (2007). HoxPred: automated classification of Hox proteins using combinations of generalised profiles. *BMC bioinformatics*, 8:247.

Thomas-Chollier, M. and Martinez, P. (2016). Origin of Metazoan Patterning Systems and the Role of ANTPClass Homeobox Genes. *eLS*.

Whelan, N. V., Kocot, K. M., and Halanych, K. M. (2015). Employing Phylogenomics to Resolve the Relationships among Cnidarians, Ctenophores, Sponges, Placozoans, and Bilaterians. *Integrative and Comparative Biology*, 55(6):1084–1095.

Chapter 2
Computational prediction of cis-regulatory elements

This chapter will be devoted to the 'traditional' approaches to predict cis-regulatory elements within genomes, before the advent of the ChIP-seq technique and other high-throughput epigenetic methods that will be covered in the next chapter. Interestingly, these 'traditional' approaches are still heavily used these days, to analyse in more detail the epigenomic datasets, or to predict cis-regulatory elements in genomes for which no epigenomic datasets are available. This is the case for many non-model organisms that have been sequenced, but that do not have established protocols for epigenomics. I will first briefly introduce the cis-regulatory elements and the notion of motifs. I will then present RSAT and my contributions to this suite of tools dedicated to the analysis of cis-regulatory elements. Next, I will present a complementary tool named TRAP. Finally, I will conclude on my current and future projects in this area.

Related papers:

- Medina-Rivera A*, Defrance M*, Sand O*, Herrmann C, Castro-Mondragon J, Delerce J, Jaeger S, Blanchet C, Vincens P, Caron C, Staines DM, Contreras-Moreira B, Artufel M, Charbonnier hamvongsa L, Hernandez C, Thieffry D, Thomas-Chollier M#, van Helden J# (2015). **RSAT 2015: Regulatory Sequence Analysis Tools.** *Nucleic Acids Research*, 43(W1):W50-W56.
- Hudry B, Thomas-Chollier, M, Volovik Y, Duffraisse M, Dard A, Frank D, Technau U, and Merabet S (2014). **Molecular insights into the origin of the Hox-TALE patterning system.** *eLife*, 3:e01939.
- Thomas-Chollier M, Defrance M, Medina-Rivera A, Sand O, Herrmann C, Thieffry D, van Helden J (2011). **RSAT 2011: regulatory sequence analysis tools.** *Nucleic Acids Research*, 39, W86-91.
- Medina-Rivera A, Abreu-Goodger C, Thomas-Chollier M, Salgado H, Collado-Vides J, van Helden J (2011). **Theoretical and empirical quality assessment of transcription factor-binding motifs.** *Nucleic Acids Research*, 39, 808-824.
- Thomas-Chollier M, Hufton A, Heinig M, O'Keeffe S, El Masri N, Roider HG, Manke T, Vingron M (2011). **Transcription factor binding predictions using TRAP for the analysis of ChIP-seq data and regulatory SNPs.** *Nature Protocols*, 6, 1860-1869.
- Sand O, Thomas-Chollier M, van Helden J (2009). **Retrieve-ensembl-seq: user-friendly and large-scale retrieval of single or multi-genome sequences from Ensembl.** *Bioinformatics*, 25, 2739-2740.
- Thomas-Chollier M*, Sand O*, Turatsinze J-V, Janky R, Defrance M, Vervisch E, Brohe S, van Helden J (2008). **RSAT: regulatory sequence analysis tools.** *Nucleic Acids Research*, 36, W119-27.
- Sand O, Thomas-Chollier M, Vervisch E, van Helden J (2008). **Analyzing multiple data sets by interconnecting RSAT programs via SOAP Web services: an example with ChIP-chip data.** *Nature Protocols*, 3, 1604-1615.
- Turatsinze J-V*, Thomas-Chollier M*, Defrance M, van Helden J (2008). **Using RSAT to scan genome sequences for transcription factor binding sites and cis-regulatory modules.** *Nature Protocols*, 3, 1578-1588.

*=co-first author #=co-corresponding author

2.1 Cis-regulatory elements and DNA binding motifs

2.1.1 Transcriptional regulation and cis-regulatory elements

Transcriptional regulation underlies the fine-tuned expression of genes in their biological context: specific cell type, developmental stage, in response to a particular stimulus... Transcriptional regulation is mediated by specific proteins named transcription factors (TFs), which bind to very short regions of DNA named cis-regulatory elements or transcription factor binding sites (TFBSs). Uncovering these regulatory elements hidden in the genomes is thus critical to understand the regulation of gene expression.

Several TFs can jointly bind to DNA on closely-located TFBSs (forming a cis-regulatory module (CRM)) to cooperatively fine-tune the expression of the target gene, or they can compete with each other for the same TFBS. TFBSs are very short (6-20 bp) and degenerate, *i.e.* a given TF is able to bind to slightly different sites with slightly different binding affinities. In metazoans, these TFBSs can be located upstream or downstream of the target gene, either in proximal or distal locations, or within an intron (and even in coding exons!). Altogether, these characteristics make TFBSs difficult to predict based on genomic sequences alone.

2.1.2 Building and describing a DNA binding motif

Specific bioinformatics approaches have been developed to identify TFBSs in DNA sequences for many years (reviewed in [Wasserman and Sandelin, 2004; GuhaThakurta, 2006; Aerts, 2012]). Many of these approaches are intrinsically based on the notion of DNA binding motifs, which account for TFBSs being degenerate (see [D'haeseleer, 2006] for an introduction to DNA sequence motifs). These motifs encode the binding specificity of TFs, and can be represented synthetically in various ways, termed motif descriptors (reviewed in [Bucher et al., 1996]). These motif descriptors include string-based, matrix-based and sequence logo representations (Fig. 2.1).

To build a motif, the first step is generally to align a set of sequences (*e.g.* experimentally validated TFBSs) (Fig. 2.1A) and choose a motif descriptor. Consensus sequences (Fig. 2.1B,C) are very synthetic motif descriptors but have inherent weaknesses. On the one hand, the strict consensus loses the information relative to the non-predominant letters at a given position (Fig. 2.1B). On the other hand, the degenerate consensus loses the information about the most frequent nucleotide (Fig. 2.1C). Position-specific scoring matrix (PSSM) is more expressive than the consensus sequence, as it keeps the information from all nucleotides (Fig. 2.1D). The matrix should be read as follows: each column represents one position of the motif and each row represents one nucleotide. The PSSM represented in figure 2.1D is a count matrix, because each cell contains the number of times each nucleotide is found at each position of the motif. From this count matrix,

it is possible to derive a frequency matrix (PFM) or various types of weight matrices (PWM). The sequence logo (Fig. 2.1E) is commonly used for a graphical representation of a motif.

PSSMs are widely-supported by analysis tools, and still represent the current standard despite being an approximation of the TF binding specificity [Stormo, 2013]. Yet, more complex motif descriptors have been employed for DNA binding motifs (listed in [Slattery et al., 2014]), for example to take into account dependencies between positions of the motifs (in PSSMs, all columns are independent). In this transition to the Big Data era, the datasets have become sufficiently large to train such complex descriptors. They comprise extension of PSSM to di-nucleotides [Zhao et al., 2012] and hidden markov models (HMMs) that can take into account variable motif lengths [Mathelier and Wasserman, 2013]. Despite the initial enthusiasm for these more complex descriptors[1], the improvement of performance seems rather marginal, except for particular TF families such as Zinc fingers [Zhao et al., 2012; Weirauch et al., 2013]. With the exponential growth of motif databases these last few years [Mathelier et al., 2016], I would tend to say that PSSMs are here to stay.

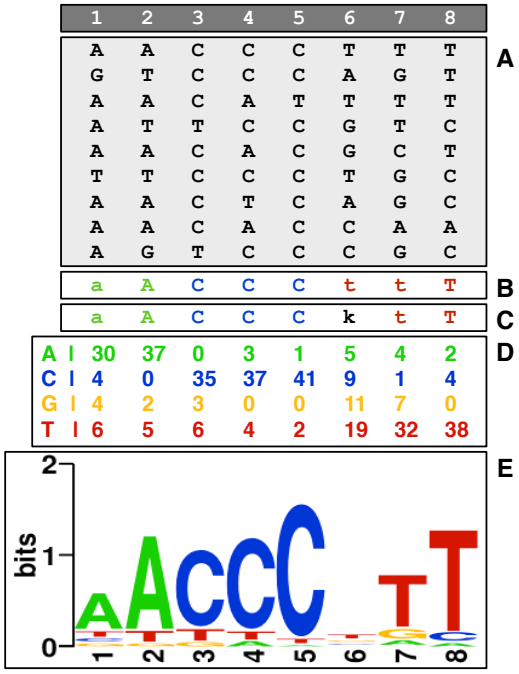

Figure 2.1: Representations of the binding specificity of a Transcription Factor. from [Turatsinze et al., 2008] **A.** Subset of the collection of 44 sites for the TF Krüppel of *Drosophila melanogaster*, taken from ORegAnno database and aligned using the program MEME. B,C,D and E are based on the whole collection of Krüppel sites. **B.** strict consensus of the selected sites. **C.** degenerate consensus using the IUPAC code for ambiguous nucleotides. **D.** position-specific scoring matrix (PSSM) obtained using RSAT *convert-matrix*. Each column of the matrix represents one position of the motif and the numbers indicate the nucleotide absolute frequencies at this position of the aligned sites. **E.** sequence Logo obtained using WebLogo (http://weblogo.berkeley.edu/logo.cgi). Each column represents one position of the motif, and the letters indicate which nucleotides are found at a given position. The total height of each column reflects its information content. The height of each letter is proportional to the frequency of the corresponding nucleotide at the given position.

[1]W. Wasserman opened his talk at an INSERM workshop in Bordeaux in 2011 by the strong statement 'PSSMs are dead!'

2.2 RSAT: Regulatory Sequence Analysis Tools

2.2.1 A well-established suite of tools for regulatory sequence analysis

The Regulatory Sequence Analysis Tools (RSAT)[2] is a software suite integrating a wide variety of programs to analyse cis-regulatory elements in genomic sequences. Its main alternative is the MEME suite [Bailey et al., 2009]. The RSAT suite has been established by Jacques van Helden (currently professor at Aix-Marseille Université, France). Since its initial development in 1998 [van Helden et al., 1998, 2000a], RSAT has provided uninterrupted service and has broadened its applications, following advances in the field of regulatory genomics.

In the earlier days, support was restricted to the yeast genome, and the server was centred on the string-based pattern-discovery algorithms *oligo-analysis* and *dyad-analysis* [van Helden et al., 2000b]. Soon, the server expanded to the building blocks of the actual suite: modular tools that can be chained to enable a complete analysis (sequence retrieval, core analysis, visualisation of the results, random controls), accessible through a Web server allowing usage by non-specialists,

[2] `http://rsat.eu`

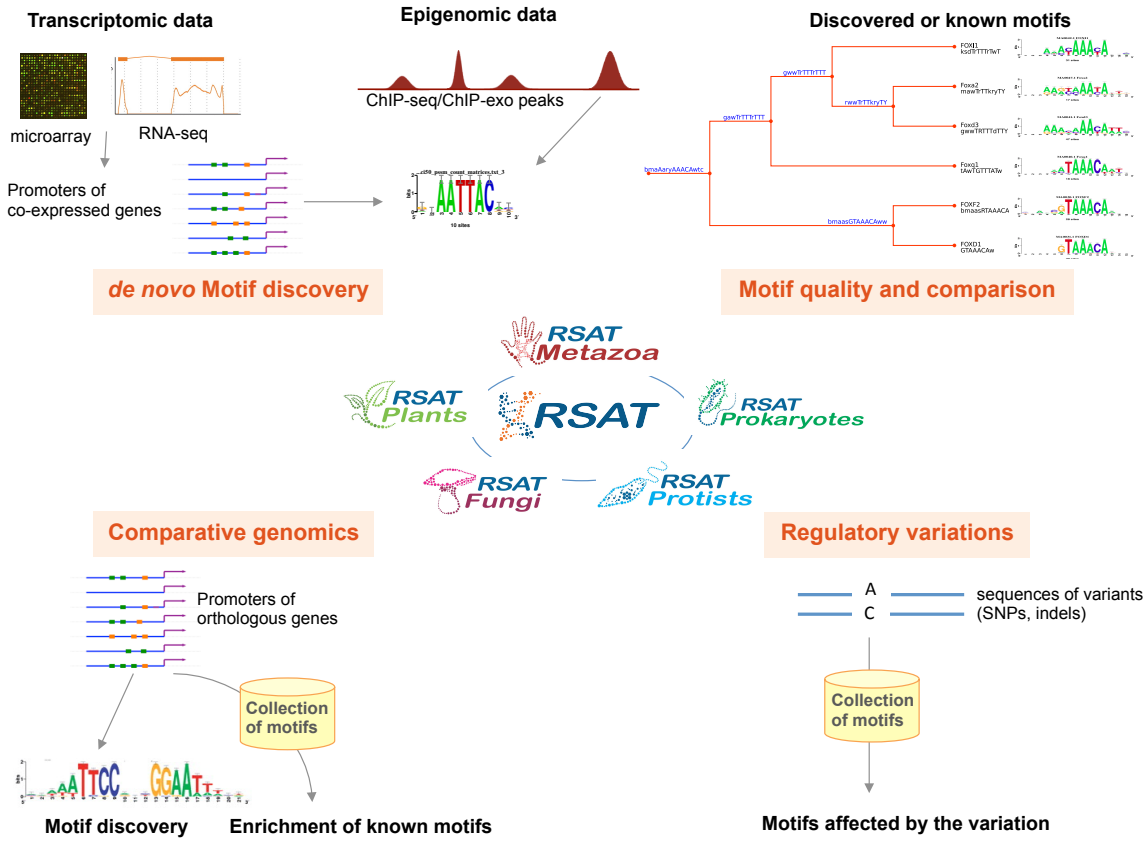

Figure 2.2: Overview of the main applications of RSAT. from [Medina-Rivera et al., 2015]

and supporting a large number of genomes, instead of focusing on a handful of model organisms [van Helden, 2003]. Some tools (like Patser) were developed by other labs, but integrated within RSAT to offer access through a graphical interface.

Over the years, several developers have contributed to RSAT when joining J. van Helden's lab. My involvement started in 2007, with the inclusion of new tools to support PSSMs and an in-depth remodelling of the Web server [Thomas-Chollier et al., 2008; Medina-Rivera et al., 2011]. This remodelling was necessary to accommodate the addition of new tools to the interface, and increase user-friendliness. My personal interest in metazoan genomes prompted the development of tools to include genomes from EnsEMBL [Thomas-Chollier et al., 2008; Sand et al., 2009]. Over the years, many more genomes have been added (amounting to 1794 in 2011 [Thomas-Chollier et al., 2011a], 3314 in 2015 [Medina-Rivera et al., 2015]). Five public Web sites dedicated to specific taxonomic groups are now in place [Medina-Rivera et al., 2015]. Last, several new tools have been developed, in particular to enable the analyses of high-throughput datasets [Thomas-Chollier et al., 2011a, 2012b] and regulatory variations [Medina-Rivera et al., 2015] (Fig. 2.2).

Nowadays, RSAT is a widely-used and established bioinformatics suite ($>$2500 citations, 15000 requests/month on the Web server, invitation for the NAR Web server issue of 2015). In addition to non-specialist users, bioinformatician users have motivated the development of programmatic access [Sand et al., 2008] and virtual machines to facilitate the local installation of the suite [Medina-Rivera et al., 2015]. Training of users is also important for the RSAT team, and I have personally been committed to education through courses, workshops and published protocols [Turatsinze et al., 2008; Sand et al., 2008; Thomas-Chollier et al., 2012a]. Over the years, my role in the RSAT team has broadened and I am now co-maintaining the suite with J. van Helden, and supervising students contributing to RSAT.

I will detail below one of the RSAT developments in which I was primarily involved, related to the 'traditional' approaches to detect cis-regulatory elements with PSSMs.

2.2.2 matrix-scan: a comprehensive PSSM pattern-matching program

Pattern-matching is a commonly-used approach to scan genomic sequences for locating putative cis-regulatory elements resembling a given motif. When the motif is described as a PSSM, the underlying concept is to find DNA segments that are more similar to the PSSM than to the expected background genomic DNA. The PSSM is used to score each segment of the sequence to analyse, and only segments with a score higher than a predefined threshold are considered as a 'hit', *i.e.* a putative binding site. To scan sequences with PSSMs, a variety of 'hit-based' programs have been developed (see references in [Turatsinze et al., 2008; Aerts, 2012]). Globally, these programs differ on the following points: supported background models, calculation of P-values,

efficiency. In addition, a distinct group of programs is dedicated to the detection of cis-regulatory modules (CRMs), *i.e.* clusters of TFBSs predictions (reviewed in [Van Loo and Marynen, 2009; Aerts, 2012]). The underlying hypothesis is that combinations of TFBS predictions are more likely to correspond to binding sites than isolated predictions.

Initially, the pattern-matching tool Patser [Hertz and Stormo, 1999] was accessible through RSAT Web interface. This program was very useful, but limited in terms of background models. With Jean-Valery Turatsinze, we implemented in RSAT the program *matrix-scan* [Turatsinze et al., 2008; Thomas-Chollier et al., 2008]. The key characteristic of this program is the calculation of P-values for background models defined as higher-order Markov chains. These P-values are important to estimate the expected number of false positive predictions. These P-values can also be used as a threshold, rather than the usually-used weight score. This is important since the ranges of weight scores are specific to each PSSM, thus a given weight (*e.g.* 5) could be a stringent threshold for a given PSSM, but a loose one for another PSSM. The *matrix-scan* program combines various features from other programs, and is easily accessible on the Web interface. Due to the increasing size of the datasets to scan, Matthieu Defrance has developed a much faster version of *matrix-scan* [Thomas-Chollier et al., 2011a], which can be used as a standalone program. *matrix-scan* is very popular, and has been independently evaluated [Dabrowski et al., 2015].

matrix-scan now represents one of the core programs of RSAT, and is heavily used by other RSAT programs. It has enabled the development of *matrix-quality* [Medina-Rivera et al., 2011], a tool that compares the distributions of PSSMs weight scores, and that can be used to evaluate the quality of PSSMs on real datasets. Interestingly, this tool can also be used to observe the enrichment of a motif in datasets. For example, I used it to reveal the specific enrichment of Hox/Pbx motif in endodermal promoters (and not in ectodermal promoters) of the sea anemone *N. vectensis* genome [Hudry et al., 2014]. The main advantage of this approach, compared to other motif enrichment programs, is that we compare the complete score distributions, without the need to apply a threshold on the weight score.

To predict putative CRMs, we have implemented the search for Cis-Regulatory Enriched Regions (CRERs) [Turatsinze et al., 2008], which correspond to regions that have a higher number of TFBS predictions than expected by chance. These CRERs may contain TFBS predictions for various transcription factors, as *matrix-scan* supports the scanning with multiple matrices as input. Initially embedded within *matrix-scan*, the detection of CRERs has been re-designed as an independent program (*crer-scan*) to increase its computing efficiency [Medina-Rivera et al., 2015].

Although the introduction of P-values and CRERs aims at reducing the number of false predictions, this remains a well-known issue in pattern matching approaches. Such overabundant false predictions led to the 'futility theorem', stating that most predictions will not have a functional role

[Wasserman and Sandelin, 2004]. Interestingly, these last months have somehow challenged this assertion, revealing the biological importance of low-affinity binding sites that deviates from the consensus motif. It has been shown that the binding specificity of a Hox protein (and its cofactors) is mediated by a cluster of low-affinity binding sites, which is evolutionary conserved [Crocker et al., 2015]. Another study supports the view that the specificity of an enhancer relies on a combination of imperfect matches to the consensus binding sites, in a sub-optimised order [Farley et al., 2015]. In our own work (see next chapter), we have found that the glucocorticoid receptor is recognising binding sites that largely deviate from the consensus sequence *in vivo*. Altogether, the emerging view that low-affinity binding sites are widespread and critical for gene regulation is opening new exciting perspectives in the field (reviewed in [Merabet and Lohmann, 2015; Crocker et al., 2016]). This has prompted the question whether "we need to reconsider the stringent criteria generally used in computational analyses and predictions of genome-wide binding data for identifying cis-regulatory sequences" [Merabet and Lohmann, 2015]. Indeed, this new paradigm will undoubtedly influence the bioinformatics methods to detect cis-regulatory elements.

2.3 TRAP: TRanscription factor Affinity Prediction

2.3.1 Energy-based models of TF-DNA binding affinity

Binding affinity denotes the strength of the TF-DNA interaction [Furey, 2012], which leads to the notion of high- and low-affinity binding sites. The specificity of a given TF denotes its capacity to distinguish between different sequences (this takes into account the differences in binding affinity for all possible binding sites [Stormo and Zhao, 2010]). In section 2.1.2, I introduced the PSSMs in which the elements of the matrix are nucleotide counts (or frequencies). These PSSMs aim to model the binding specificity of TFs in a simple probabilistic framework. A distinct biophysical energy-based framework has also been proposed, in which TF-DNA interactions are considered in terms of binding energies (reviewed in [Stormo, 2013], see also [Slattery et al., 2014]). In this framework, the elements of the PSSMs are energy contributions of each base, taken independently at each position, and when summed, determine the total binding free energy of any sequence.

During my postdoc in Martin Vingron's lab (MPIMG, Berlin, Germany), I used the program TRAP [Roider et al., 2007; Manke et al., 2008] previously developed in this group, which supports such energy matrices to calculate the total affinity of a TF for a sequence. I will present below this program with its advantages and limits, then my contributions to expand its usage by the community, and finally, I will briefly present recently-developed approaches within the energy-based framework.

2.3.2 TRAP: predictions of transcription factor affinities with an energy model

The TRanscription factor Affinity Prediction (TRAP) method calculates the affinity of transcription factors for DNA sequences, on the basis of a biophysical model of the binding energies between a TF and DNA [Roider et al., 2007; Manke et al., 2008]. In contrast with the 'hit-based' methods like RSAT *matrix-scan*, a sequence segment is viewed as a continuous fragment for which the total affinity for a given TF can be calculated, rather than as a binding site or not a binding site. This circumvents the main limitation of hit-based methods: selecting an optimal threshold to separate the predicted TFBSs from the background. TRAP does not require a threshold value, as the program sums the affinity of each segment for a given TF over the total length of a sequence (Fig. 2.3).

The main advantage of TRAP is that all positions in the sequence contribute to the overall affinity, including low-affinity sites. As mentioned above, detecting low-affinity binding sites is difficult with hit-based methods, as it requires lowering the threshold, and consequently results in further increasing the number of false predictions. Moreover, TRAP takes advantage of the PSSMs of hit-based approach (such as the collections provided by the JASPAR database [Mathelier et al.,

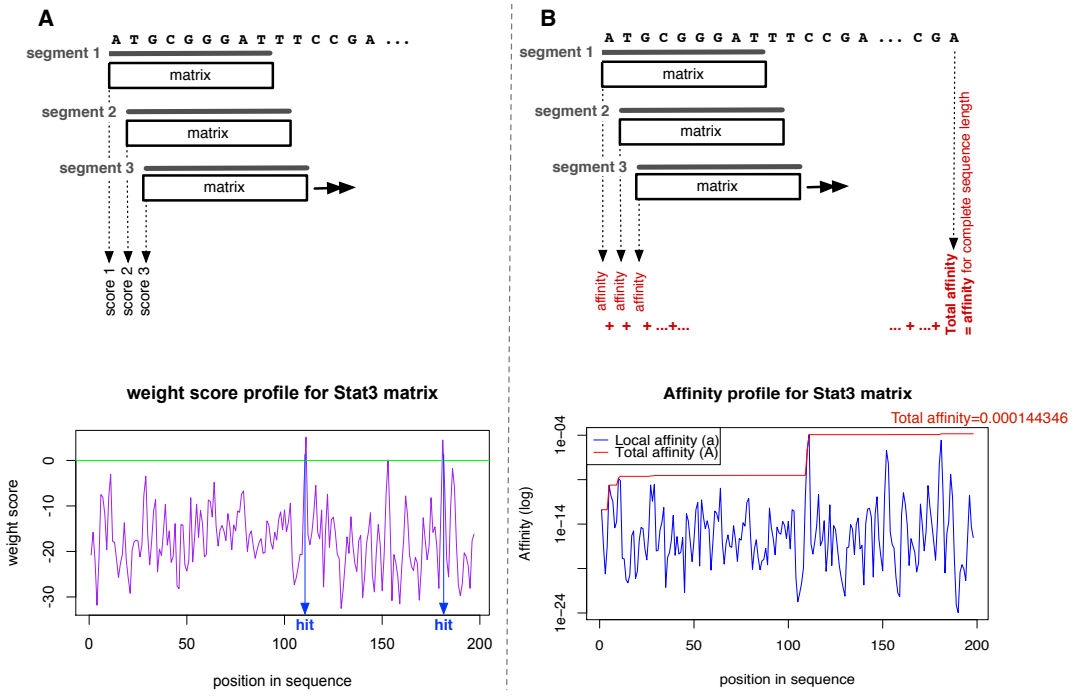

Figure 2.3: General principle of hit-based and affinity-based approaches. from [Thomas-Chollier et al., 2011b]. **A.** Hit-based method. The matrix is aligned to the sequence, and a score is given to each aligned segment. Only scores reaching the predefined threshold (indicated as a green line for an arbitrary value of 0 on the plot) are retained, and the sequence positions corresponding to these scores are reported as TFBS hits (indicated by blue arrows on the plot). B. Affinity-based method as implemented in TRAP. The energy matrix is aligned to the sequence, and an affinity value is calculated for each aligned segment. These affinity values are summed over the entire sequence, to obtain the total affinity value for this sequence. The cumulative TRAP score is shown in red.

2016]) by first converting them into position-specific energy matrices. TRAP does not return the position of the putative binding sites, but hit-based methods can be used to further refine the results by identifying the precise position of a putative TFBS. This is why TRAP and hit-based approaches are complementary rather than competitive approaches.

The main limitation of TRAP is related to the P-value calculation [Manke et al., 2008], which is necessary to normalise the affinity values across multiple matrices, and ask questions such as which TFs have the highest relative binding affinity for a given sequence. To calculate the P-value, the distribution of affinities in background sequences must be pre-calculated, which is achieved by parametrizing each matrix individually. The limitation relies in this step of parametrisation, as it does not work for all matrices [Manke et al., 2008], it needs to be defined for each size of sequences to be treated, and it is computationally demanding (for parametrising a motif database such as JASPAR, a computer cluster is necessary). This means that only a few pre-calculated backgrounds are available for the users, and that users will not easily train matrices for personal background sequences.

My involvement in TRAP was principally to deliver the method as a usable tool for the community. The experience I gained with RSAT allowed me to refactor the program into modular sub-tools, to offer a Web-based access designed for non-specialist users[3], and to provide a protocol guiding users for best usage of the program [Thomas-Chollier et al., 2011b]. A R version of TRAP has been concomitantly developed by M. Heinig[4].

2.3.3 A bright future for energy models ?

In recent years, approaches based on energy models have been highlighted, and new methodologies have been proposed to build the energy matrices directly from high-throughput datasets [Stormo, 2013; Slattery et al., 2014]. In particular, they have been shown to perform well on an independent evaluation of various methodologies [Weirauch et al., 2013]. In connection with the view that low-affinity binding sites are biologically important for gene regulation, the energy models are better suited than hit-based methods [Crocker et al., 2016]. Advances in machine learning approaches may nevertheless shadow energy models in the near future, as more complex models obtained with deep learning on TF-DNA binding experiment datasets has systematically outperformed all previous methods [Alipanahi et al., 2015; Park and Kellis, 2015].

[3] `http://trap.molgen.mpg.de`
[4] `https://github.com/matthuska/tRap`

2.4 Current projects and perspective

In this transition to the Big Data era, I am convinced of the crucial importance of providing user-friendly tools to experimentalists, who have the biological expertise to analyse their data, but often lack adequate bioinformatics skills. I will thus pursue my long-standing collaboration with J. van Helden, ensuring the maintenance and new developments of RSAT. We will continue our efforts regarding accessibility of the tools (particularly to biologists having non-published genomes to analyse) and training of users. In this respect, we are both partners in the recently-accepted European COST action 'Gene Regulation Ensemble Effort for the Knowledge Commons'[5].

2.4.1 matrix-clustering: reducing motif redundancy using a dynamic visualisation of clusters

We are currently finalising *matrix-clustering*, a program to cluster PSSMs based on their similarity, with a dynamic visualisation allowing to browse the motif trees and collapse/expand each branch to reduce the redundancy in a semi-automated way. For this work, I have co-supervised Jaime Mondragon, PhD student in J. van Helden's lab.

I have contributed to the original idea, the design of a report displaying interpretable results, and to the applications of this program to tackle the important problem of motif redundancy, occurring within and across motif collections (JASPAR, TRANSFAC). One goal is to pave the way towards a non-redundant public motif collection, which will reduce computing time when performing motif analyses and will constitute a valuable resource for the community. Another goal is to facilitate the integration of results obtained with multiple motif discovery programs, as illustrated in figure 2.4. This example reveals the strength of this program in correctly aligning the motifs, and revealing groups that are very difficult to detect by eye. Of note, this example is based on the same data we analysed previously [Thomas-Chollier et al., 2012b], yet, we were unable before to pinpoint by eye the Oct/Ocr motif among all motif variants. Interestingly, this and the MORE motifs are only reported by RSAT *peak-motifs* (see next chapter for a description of this tool) and not by the concurrent programs MEME-ChIP and HOMER.

This program is unique and constitutes a methodological breakthrough that can concretely tackle the current avalanche of PSSMs (redundant) collections, for example by clustering the meta-database footprintDB [Sebastian and Contreras-Moreira, 2014] or resolving automatically the internal redundancy of the 1000 motifs in JASPAR, currently achieved by hand [Mathelier et al., 2016].

[5] http://www.cost.eu/COST_Actions/ca/CA15205

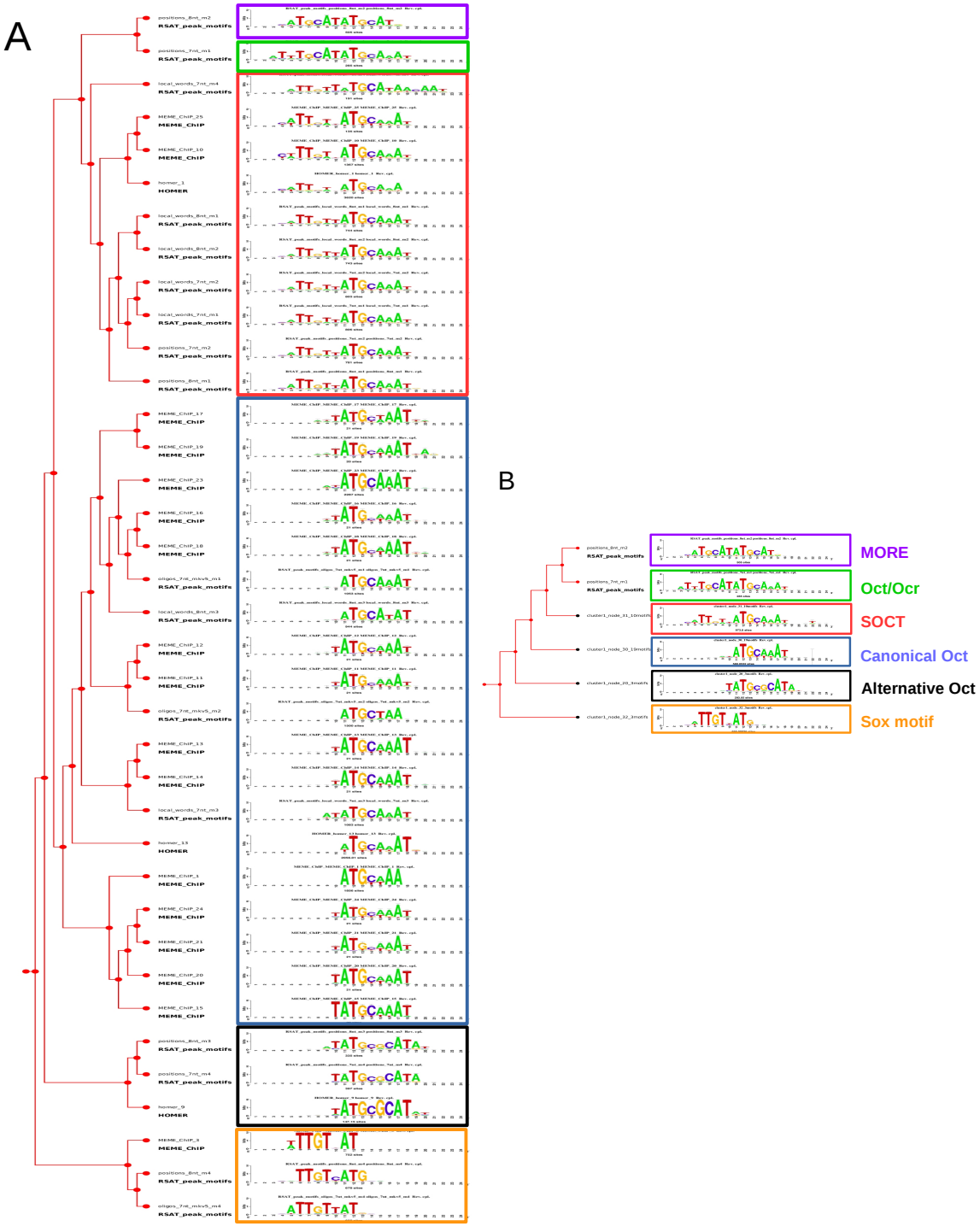

Figure 2.4: Example of PSSM clustering with *matrix-clustering*. from manuscript in preparation. **A.** The hierarchical tree represents one cluster containing 37 motifs, discovered with three distinct programs (RSAT peak-motifs, MEME-ChIP and Homer), from the same TF ChIP-seq dataset of Oct4 [Chen et al., 2008]. Different Oct (and Sox2) variants can be observed (each variant is selected with a different colour), some of which found with multiple programs, and reduced in a semi-automated way to **B.** 6 main motifs, annotated from the literature.

2.4.2 Supporting sequence conservation in RSAT

Until now, RSAT does not natively support cross-species sequence conservation information. However, it is common for users to have a few dozen candidate genes to analyse, by extracting their upstream regions, focusing on cross-species conserved regions (with the hypothesis that conserved regions are more likely to contain biologically functional elements) and scanning them with a collection of PSSMs to predict binding sites and regulatory modules. This generic approach is not yet easily accessible within RSAT to experimental biologists, especially those working with non-model organisms. In the context of the ANR Echinodal project (2014-2018, coordinator: Thierry Lepage), I will supervise the implementation of such a workflow in RSAT, which will be tested on the conserved sequences between the sea urchin genomes of *Paracentrotus lividus* and *Strongylocentrotus purpuratus*.

This project will benefit from a new collaboration with the team of H. Roest Crollius (IBENS, Paris) to extract the conserved non-coding regions from the Genomicus database [Louis et al., 2015]. Together, we have just obtained a grant from the 'Institut Français de Bioinformatique' (RSA-Ticus, 2016-2018) to better connect Genomicus and RSAT. This will allow users to easily analyse large-scale functional genomics datasets, and prioritise candidates for experimental validation, also for non-model organisms.

References

Aerts, S. (2012). Computational strategies for the genome-wide identification of cis-regulatory elements and transcriptional targets. *Current topics in developmental biology*, 98:121–145.

Alipanahi, B., Delong, A., Weirauch, M. T., and Frey, B. J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology*, 33(8):831–838.

Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., Ren, J., Li, W. W., and Noble, W. S. (2009). MEME Suite: tools for motif discovery and searching. *Nucleic Acids Research*, 37(Web Server issue):W202–8.

Bucher, P., Karplus, K., Moeri, N., and Hofmann, K. (1996). A flexible motif search technique based on generalized profiles. *Computers & chemistry*, 20(1):3–23.

Chen, X., Xu, H., Yuan, P., Fang, F., Huss, M., Vega, V. B., Wong, E., Orlov, Y. L., Zhang, W., Jiang, J., Loh, Y.-H., Yeo, H. C., Yeo, Z. X., Narang, V., Govindarajan, K. R., Leong, B., Shahab, A., Ruan, Y., Bourque, G., Sung, W.-K., Clarke, N. D., Wei, C.-L., and Ng, H.-H. (2008). Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*, 133(6):1106–1117.

Crocker, J., Abe, N., Rinaldi, L., Mcgregor, A. P., Frankel, N., Wang, S., Alsawadi, A., Valenti, P., Plaza, S., Payre, F., Mann, R. S., and Stern, D. L. (2015). Low affinity binding site clusters confer hox specificity and regulatory robustness. *Cell*, 160(1-2):191–203.

Crocker, J., Noon, E., and Stern, D. L. (2016). The Soft Touch: Low-Affinity Transcription Factor Binding Sites in Development and Evolution. *Current topics in developmental biology*.

Dabrowski, M., Dojer, N., Krystkowiak, I., Kaminska, B., and Wilczyński, B. (2015). Optimally choosing PWM motif databases and sequence scanning approaches based on ChIP-seq data. *BMC bioinformatics*, 16:140.

D'haeseleer, P. (2006). What are DNA sequence motifs? *Nature Biotechnology*, 24(4):423–425.

Farley, E. K., Olson, K. M., Zhang, W., Brandt, A. J., Rokhsar, D. S., and Levine, M. S. (2015). Suboptimization of developmental enhancers. *Science (New York, NY)*, 350(6258):325–328.

GuhaThakurta, D. (2006). Computational identification of transcriptional regulatory elements in DNA sequence. *Nucleic Acids Research*, 34(12):3585–3598.

Hertz, G. Z. and Stormo, G. D. (1999). Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics (Oxford, England)*, 15(7-8):563–577.

Hudry, B., Thomas-Chollier, M., Volovik, Y., Duffraisse, M., Dard, A. e. l., Frank, D., Technau, U., and Merabet, S. (2014). Molecular insights into the origin of the Hox-TALE patterning system. *eLife*, 3:e01939.

Louis, A., Nguyen, N. T. T., Muffato, M., and Roest Crollius, H. (2015). Genomicus update 2015: KaryoView and MatrixView provide a genome-wide perspective to multispecies comparative genomics. *Nucleic Acids Research*, 43(Database issue):D682–9.

Manke, T., Roider, H. G., and Vingron, M. (2008). Statistical modeling of transcription factor binding affinities predicts regulatory interactions. *PLoS computational biology*, 4(3):e1000039.

Mathelier, A., Fornes, O., Arenillas, D. J., Chen, C.-y., Denay, G., Lee, J., Shi, W., Shyr, C., Tan, G., Worsley-Hunt, R., Zhang, A. W., Parcy, F., Lenhard, B., Sandelin, A., and Wasserman, W. W. (2016). JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Research*, 44(D1):D110–5.

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

Medina-Rivera, A., Abreu-Goodger, C., Thomas-Chollier, M., Salgado, H., Collado-Vides, J., and van Helden, J. (2011). Theoretical and empirical quality assessment of transcription factor-binding motifs. *Nucleic Acids Research*, 39(3):808–824.

Medina-Rivera, A., Defrance, M., Sand, O., Herrmann, C., Castro-Mondragon, J. A., Delerce, J., Jaeger, S., Blanchet, C., Vincens, P., Caron, C., Staines, D. M., Contreras-Moreira, B., Artufel, M., Charbonnier-Khamvongsa, L., Hernandez, C., Thieffry, D., Thomas-Chollier, M., and van Helden, J. (2015). RSAT 2015: Regulatory Sequence Analysis Tools. *Nucleic Acids Research*, 43(W1):W50–6.

Merabet, S. and Lohmann, I. (2015). Toward a new twist in Hox and TALE DNA-binding specificity. *Developmental cell*, 32(3):259–261.

Park, Y. and Kellis, M. (2015). Deep learning for regulatory genomics. *Nature Biotechnology*, 33(8):825–826.

Roider, H. G., Kanhere, A., Manke, T., and Vingron, M. (2007). Predicting transcription factor affinities to DNA from a biophysical model. *Bioinformatics (Oxford, England)*, 23(2):134–141.

Sand, O., Thomas-Chollier, M., and van Helden, J. (2009). Retrieve-ensembl-seq: user-friendly and large-scale retrieval of single or multi-genome sequences from Ensembl. *Bioinformatics (Oxford, England)*, 25(20):2739–2740.

Sand, O., Thomas-Chollier, M., Vervisch, E., and van Helden, J. (2008). Analyzing multiple data sets by interconnecting RSAT programs via SOAP Web services: an example with ChIP-chip data. *Nature Protocols*, 3(10):1604–1615.

Sebastian, A. and Contreras-Moreira, B. (2014). footprintDB: a database of transcription factors with annotated cis elements and binding interfaces. *Bioinformatics (Oxford, England)*, 30(2):258–265.

Slattery, M., Zhou, T., Yang, L., Dantas Machado, A. C., Gordân, R., and Rohs, R. (2014). Absence of a simple code: how transcription factors read the genome. *Trends in biochemical sciences*, 39(9):381–399.

Stormo, G. D. (2013). Modeling the specificity of protein-DNA interactions. *Quantitative biology*, 1(2):115–130.

Thomas-Chollier, M., Darbo, E., Herrmann, C., Defrance, M., Thieffry, D., and van Helden, J. (2012a). A complete workflow for the analysis of full-size ChIP-seq (and similar) data sets using peak-motifs. *Nature Protocols*, 7(8):1551–1568.

Thomas-Chollier, M., Defrance, M., Medina-Rivera, A., Sand, O., Herrmann, C., Thieffry, D., and van Helden, J. (2011a). RSAT 2011: regulatory sequence analysis tools. *Nucleic Acids Research*, 39(Web Server issue):W86–91.

Thomas-Chollier, M., Herrmann, C., Defrance, M., Sand, O., Thieffry, D., and van Helden, J. (2012b). RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets. *Nucleic Acids Research*, 40(4):e31.

Thomas-Chollier, M., Hufton, A., Heinig, M., O'Keeffe, S., Masri, N. E., Roider, H. G., Manke, T., and Vingron, M. (2011b). Transcription factor binding predictions using TRAP for the analysis of ChIP-seq data and regulatory SNPs. *Nature Protocols*, 6(12):1860–1869.

Thomas-Chollier, M., Sand, O., Turatsinze, J.-V., Janky, R., Defrance, M., Vervisch, E., Brohée, S., and van Helden, J. (2008). RSAT: regulatory sequence analysis tools. *Nucleic Acids Research*, 36(Web Server issue):W119–27.

Turatsinze, J.-V., Thomas-Chollier, M., Defrance, M., and van Helden, J. (2008). Using RSAT to scan genome sequences for transcription factor binding sites and cis-regulatory modules. *Nature Protocols*, 3(10):1578–1588.

van Helden, J. (2003). Regulatory sequence analysis tools. *Nucleic Acids Research*, 31(13):3593–3596.

van Helden, J., André, B., and Collado-Vides, J. (1998). Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *Journal of molecular biology*, 281(5):827–842.

van Helden, J., André, B., and Collado-Vides, J. (2000a). A web site for the computational analysis of yeast regulatory sequences. *Yeast (Chichester, England)*, 16(2):177–187.

van Helden, J., Rios, A. F., and Collado-Vides, J. (2000b). Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. *Nucleic Acids Research*, 28(8):1808–1818.

Van Loo, P. and Marynen, P. (2009). Computational methods for the detection of cis-regulatory modules. *Briefings in Bioinformatics*, 10(5):509–524.

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

Weirauch, M. T., Cote, A., Norel, R., Annala, M., Zhao, Y., Riley, T. R., Saez-Rodriguez, J., Cokelaer, T., Vedenko, A., Talukder, S., DREAM5 Consortium, Bussemaker, H. J., Morris, Q. D., Bulyk, M. L., Stolovitzky, G., and Hughes, T. R. (2013). Evaluation of methods for modeling transcription factor sequence specificity. *Nature Biotechnology*, 31(2):126–134.

Zhao, Y., Ruan, S., Pandey, M., and Stormo, G. D. (2012). Improved models for transcription factor binding site identification using nonindependent interactions. *Genetics*, 191(3):781–790.

Chapter 3
The Big Data era of cis-regulatory element detection

The advent of the ChIP-seq experiment has been a turning point in the regulatory genomics field. In this chapter, I will briefly describe this technique, and present *peak-motifs*, the motif discovery tool dedicated to ChIP-seq in RSAT. I will then present the biological insights we obtained on the glucocorticoid receptor (GR) using ChIP-seq. ChIP-seq resolution has been enhanced by the ChIP-exo technique. I will present ExoProfiler, the program we developed to analyse ChIP-exo datasets, and summarise the additional information we obtained on the binding of GR using ChIP-exo. Next, I will highlight the collective know-how gained over the years in producing high-quality ChIP-seq datasets, and introduce recent techniques that may become the future standard approaches. Finally, I will conclude on my current and future projects.

Related papers:

- Telorac J, Prykhozhij SV, Schoene S, Meierhofer D, Sauer S, Thomas-Chollier M#, Meijsing SH# (2016). **Identification and characterization of DNA sequences that prevent glucocorticoid receptor binding to nearby response elements.** *Nucleic Acids Research*, in press.
- Starick S*, Ibn-Salem J*, Jurk M*, Hernandez C, Love MI, Chung H, Vingron M, Thomas-Chollier M#, Meijsing SH# (2015). **ChIP-exo signal associated with DNA-binding motifs provide insights into the genomic binding of the glucocorticoid receptor and cooperating transcription factors.** *Genome Research*, 25(6):825-35.
- Thomas-Chollier M*, Watson L* , Cooper S, Pufall MA, Liu JS, Borzym K, Vingron M, Yamamoto K.R , Meijsing SH (2013). **A naturally occuring insertion of a single amino acid rewires transcriptional regulation by glucocorticoid receptor isoforms.** *Proceedings of the National Academy of Sciences of the United States of America*,110(44):17826-31.
- Thomas-Chollier M, Herrmann C, Defrance M, Sand O, Thieffry D, van Helden J (2012). **RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets.** *Nucleic Acids Research* 40, e31.
- Thomas-Chollier M, Darbo E, Herrmann C, Defrance M, Thieffry D, van Helden J (2012). **A complete workflow for the analysis of full-size ChIP-seq (and similar) data sets using peak-motifs.** *Nature Protocols* 7, 1551568.

*=co-first author #=co-corresponding author

3.1 The ChIP-seq revolution

3.1.1 ChIP-seq: a high-throughput approach to detect DNA binding regions

The Chromatin Immuno-Precipitation (ChIP) has been widely used for years to study *in vivo* protein-DNA interactions, for example to detect DNA bound to a given TF or to histones bearing particular chemical modifications on their tails. The technique itself is not recent, but the methods to analyse the bound DNA have been refined over the years, first limiting the analyses to individual loci, now widened to genome-scale thanks to crucial advances in sequencing technology (Fig. 3.1). The so-called 'ChIP-seq' approach, developed in 2007, has been rapidly adopted to become the current standard, as it combined many advantages towards other approaches [Mardis, 2007], although presenting some limitations (reviewed in [Park, 2009; Liu et al., 2010]). A major driving force in the wide adoption and improvement of ChIP-seq has been its use in the ENCODE project[1]. As of today, more than 2700 ChIP-seq experiments have been produced solely by ENCODE. This large experience allowed ENCODE to release guidelines for the community [Landt et al., 2012]. To increase the production of datasets, ChIP-seq assays can be automatised with robots [Aldridge et al., 2013].

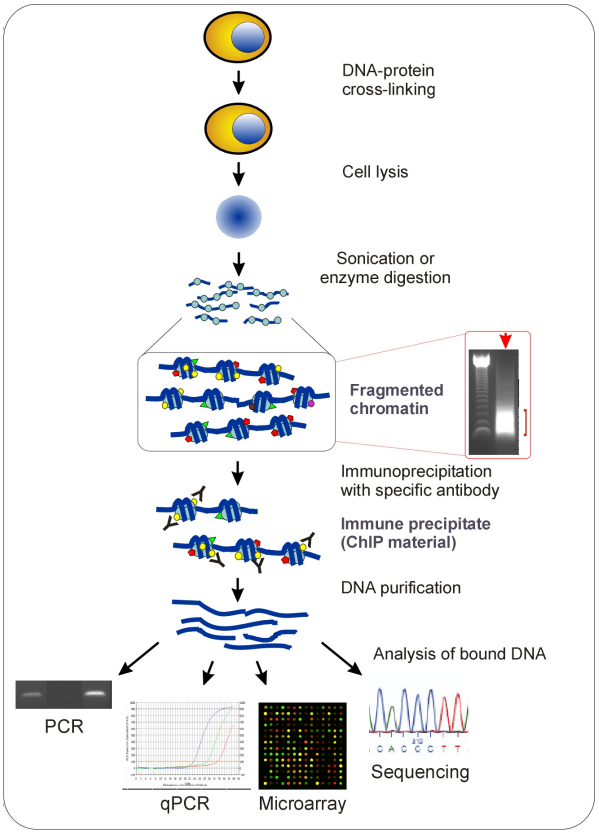

Figure 3.1: Overview of a ChIP experiment ending with various DNA detection techniques. from [Collas and Dahl, 2008]. A cross-linking reagent (formaldehyde is routinely used) is applied on the cells to covalently link proteins and DNA. DNA is then fragmented, and only fragments of the desired length are selected. A specific antibody is then used to retrieve the protein of interest, still bound to DNA. DNA and proteins are then dissociated, so that DNA can be assayed to identify the regions bound to the protein. Several assays are used: PCR-based for small-scale analyses, microarray (ChIP-on-chip) for larger analyses, and later sequencing (ChIP-seq) for full-genome analyses, thanks to the advent of the high-throughput sequencing techniques.

[1] https://www.encodeproject.org

Like any dataset produced with high-throughput sequencing, ChIP-seq requires a computational processing of the raw data, to ultimately obtain the binding profile of the studied protein. This processing consists in multiple steps, performed by distinct programs [Landt et al., 2012; Bailey et al., 2013; Nakato and Shirahige, 2016], which may be superseded by newer tools, since novel bioinformatics methods are still being developed at fast pace. In an ideal ChIP-seq experiment targeting a TF, all bound regions (called 'peaks') would represent direct binding sites and point to the exact TFBS. However, a real dataset contains non-specific regions, or regions corresponding to indirect binding. In addition, the resolution of the peaks (typically 200-400bp) is not precise enough to pinpoint the TFBSs (6-10bp). In practice, to locate the TFBSs and identify the peaks corresponding to direct binding of the TF, a step of motif discovery in all peaks is generally performed to infer the motif(s), followed by a pattern-matching step (*e.g.* with *matrix-scan*) to scan the peaks with the found motifs.

3.1.2 RSAT peak-motifs: motif discovery in full-size datasets

Many programs for motif discovery have been implemented over the years [Tompa et al., 2005]. However, an important bottleneck for most of these tools is that the underlying algorithms were originally developed to discover binding motifs from a small set of co-regulated promoters, and can hardly treat the thousands of peaks produced by ChIP-seq experiments (reviewed in [Zambelli et al., 2013]). This limitation is typically circumvented by restricting motif discovery to a few hundreds peak regions and by truncating the peaks to a maximal width (e.g. 100 bp) to further reduce the total size of the sequence set. However, given the power of the genome-wide experimental approach, one would like to be able to analyze the full dataset. Some alternative algorithms support the analysis of large-scale data sets but are only available via a Unix shell interface, and are thus of poor usability for life-science researchers. Interestingly, ChIP-seq datasets are inherently different from the previous promoter-based datasets when considering motif discovery approaches. Indeed, ChIP-seq datasets are expected to contain more TFBSs under the summit of the peaks, thus enabling approaches based on positional-bias to discover motifs.

I have contributed to the development of a comprehensive pipeline within RSAT called *peak-motifs* [Thomas-Chollier et al., 2012b], motivated by the pressing need for a statistically reliable, time-efficient and user-friendly framework to analyze full datasets of ChIP-seq peaks or similar data (CLIP-seq, DNAse I,...) (Fig. 3.2). This motif discovery approach was significantly faster than other available alternatives, thereby allowing processing of full ChIP-seq datasets, even from the web server. At that time, *peak-motifs* was the only tool that performed a complete motif analysis, in addition to offering a user-friendly web interface without any restriction on sequence size or number of peaks (refer to [Tran and Huang, 2014] for a recent survey of tools and [Boeva, 2016]

for a review). As of today, *peak-motifs* has become a reference, to which newer programs are compared (e.g. [Ding et al., 2014]) and it has been extensively reviewed recently [Lihu and Holban, 2015]. We have furthermore issued a protocol to guide users in the proper usage of the program, and biological interpretation of the results [Thomas-Chollier et al., 2012a].

This program keeps being improved over the years. In particular, the combined use of multiple motif-discovery algorithms in *peak-motifs* generates redundant motifs. To interpret and synthesise the results, a time-consuming step of grouping similar motifs by hand is often necessary. The development of *matrix-clustering* (presented in section 2.4.1) resolves this limitation, by enabling the automatic clustering of all motifs reported by *peak-motifs*.

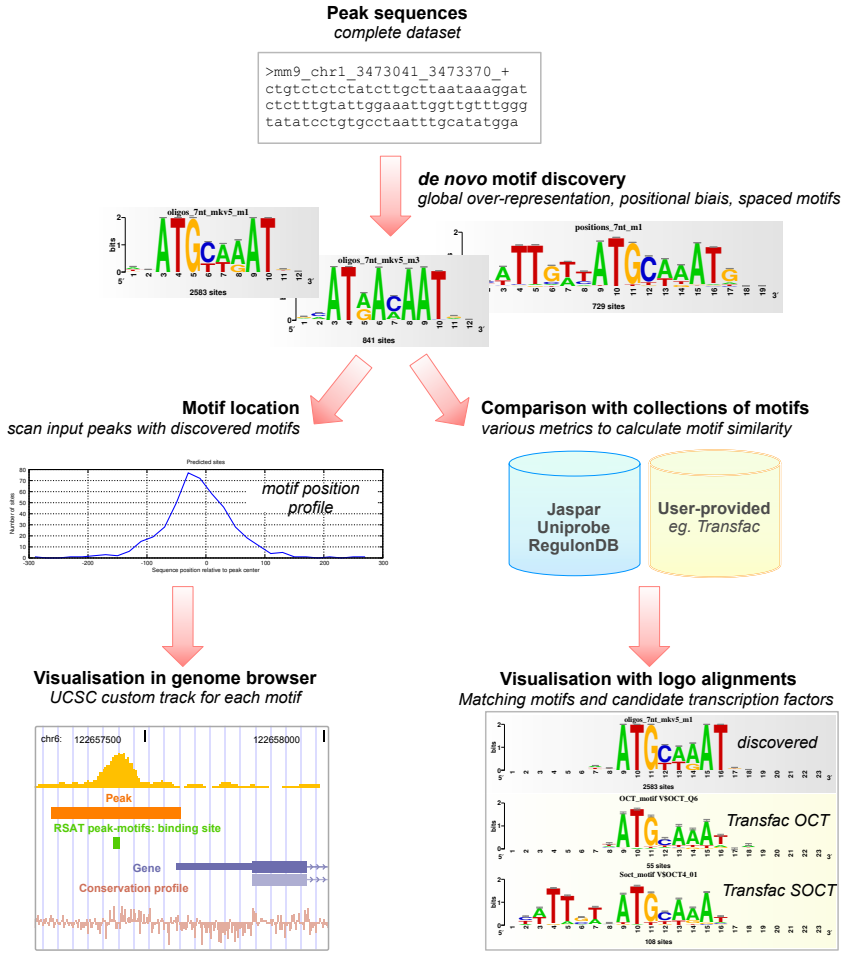

Figure 3.2: Schematic flow chart of the peak-motifs pipeline. from [Thomas-Chollier et al., 2012b]. For sake of clarity, only the main analysis steps are depicted. The pipeline takes as input a set of peak sequences, and runs several *de novo* motif discovery algorithms based on different detection criteria: over-representation, differential representation (test versus control), global position bias or local over-representation along the centred peaks. Transcription factors are predicted by matching discovered motifs against several public motif databases and/or against user-uploaded motif collections. Peak sequences are scanned with the discovered motifs to predict precise binding positions. These positions are then automatically exported as an annotation track for UCSC genome browser, thus enabling a flexible visualisation in their genomic context.

3.1.3 New insights in the binding of glucocorticoid receptor to DNA from ChIP-seq datasets

During my postdoc in Martin Vingron's lab (2009-2012, MPIMG, Berlin, Germany), I gained expertise in ChIP-seq data analysis and started to collaborate with the experimental group of Sebastiaan Meijsing (MPIMG, Berlin, Germany). Over the years, I performed (and later supervised) most bioinformatics analyses of the data produced by this team, and we are currently still developing joint projects. Together with S. Meijsing, we have co-supervised a master student (Jonas Ibn-Salem) and we are co-tutoring a PhD student (Stefanie Shoene).

The team of S. Meijsing aims at deciphering the molecular mechanisms underlying the binding of TFs to DNA, and the resulting regulation of target genes. They focus on the glucocorticoid receptor (GR), because this nuclear receptor is inducible in the lab by a molecule (a synthetic glucocorticoid), and has been well-studied at the molecular and physiological levels due to its importance as a drug target. The glucocorticoid steroid hormone is indeed mostly known for its anti-inflammatory action and therapeutic usage (*e.g.* allergies), unfortunately associated with many side-effects such as osteoporosis. The glucocorticoid associates with GR in the cell, which then regulates target genes by binding to DNA cis-regulatory elements. We thus aimed at better determining which sequences are driving the binding of GR to its specific regulatory elements, using ChIP-seq data from various cell lines. I will present below insights gained though two studies: (i) the binding specificity of two GR isoforms is partly explained by a subtle but functional motif variant [Thomas-Chollier et al., 2013], (ii) Negative Regulatory Sequences (NRSs) interfere with genomic GR binding through proteins found at sub-nuclear structures called paraspeckles [Telorac et al., 2016].

GRα and GRγ are two naturally occurring isoforms, which differ by a single arginine insertion located in the DNA binding domain. This insertion does not prevent DNA binding, but alters the transcriptional outcome induced by GR. To determine whether this insertion had an effect on GR DNA occupancy, we performed ChIP-seq of GRα and GRγ, and observed that binding regions were remarkably similar, although a small portion of binding regions were isoform-specific. We thus further examined whether the insertion in the gamma isoform altered the sequence preference of GR. We used *peak-motifs* to identify sequence motifs underlying the three classes of binding sites: nondifferential, α-specific and γ-specific binding sites. Although all three motifs look similar, the GRγ motif diverges from the consensus at two positions in the motif. Experimentally mutating a GRγ motif into a non-differential motif restored the activation by the GRα isoform, suggesting differential binding of GRγ to specific sequence motifs explains in part the differential regulation [Thomas-Chollier et al., 2013].

Motif discovery tools are usually designed to find over-represented motifs, but *peak-motifs* in-

tegrates an algorithm that searches for positionally-biaised motifs, thus not specifically directed towards over-representation. Interestingly, in our GR ChIP-seq datasets, we noticed under-represented motifs and reasoned that these may restrict GR binding and contribute -as negative regulatory signals- to guide GR to the appropriate genomic loci. We tested the activity of such under-represented sequences and found that they can indeed interfere with GR binding to nearby GR binding sites, by mechanisms that appear not to involve changes in chromatin accessibility, but rather implicate proteins associated with a specific sub-nuclear structure called paraspeckles. This study uncovered two Negative Regulatory Sequences (NRSs), but additional under-represented sequences were predicted. Notably, such NRS are also detected in ChIP-seq datasets of other TFs, thus pointing to a potential larger spectrum of TFs negatively regulated by these NRSs [Telorac et al., 2016].

3.2 Increasing the ChIP-seq resolution with ChIP-exo

3.2.1 ChIP-exo : a base pair resolution ChIP-seq

The limited resolution of ChIP-seq prompted the development of enhanced near base-pair resolution ChIP protocols (reviewed in [Zentner and Henikoff, 2014], also see [Furey, 2012]). In particular, the ChIP-exo approach [Rhee and Pugh, 2011] is a variation of the ChIP-seq protocol adding an exonuclease that digests the DNA fragment from the 5' end until it reaches the formaldehyde cross-linked protein. Two barriers (one on each strand) therefore encircle the cross linked protein, considerably increasing the resolution of the resulting signal (Fig. 3.3). Unfortunately, the initial report of the technique [Rhee and Pugh, 2011] was somewhat too optimistic, considering that any ChIP-exo region not overlapping a ChIP-seq region was true signal not captured by ChIP-seq, which we have later shown to be erroneous [Starick et al., 2015]. Both experimental and computational tasks are more complicated than for ChIP-seq [Mahony and Pugh, 2015]. This may explain why datasets produced by other teams were published only two years later (*e.g.* [Serandour et al., 2013]).

3.2.2 ExoProfiler : a motif-based approach to analyze ChIP-exo signal

Only a handful of computational approaches have been specifically designed to study ChIP-exo datasets [Zentner and Henikoff, 2014], all directed towards finding peaks, as traditionally done in ChIP-seq. However, calling peaks on the two barriers represented by the ChIP-exo signal only works in the simple situation where a single TF is binding at a location (which was the case in the original datasets [Rhee and Pugh, 2011]). Considering more complex situations (*e.g.*TF like GR binding with co-factors, see below) inevitably hampers peak-calling. We have thus developed ExoProfiler[2], a computational motif-based approach to study the ChIP-exo signal and define footprints

[2]https://github.com/ComputationalSystemsBiology/ExoProfiler

[Starick et al., 2015]. We find that these ChIP-exo footprints are protein-specific and recognition sequence-specific signatures of genomic TF association (Fig. 3.3 illustrate the GR footprint). Importantly, this approach reveals that ChIP-exo captures information about TFs other than the one directly targeted by the antibody in the ChIP-procedure. Consequently, based on the shape of the footprint, one can discriminate between direct and indirect DNA association. The development of ExoProfiler involved Jonas Ibn-Salem (Master2 student co-supervised with S. Meijsing) and Céline Hernandez (bioinformatics engineer supervised by me), who packaged ExoProfiler as a documented tool freely available to the community.

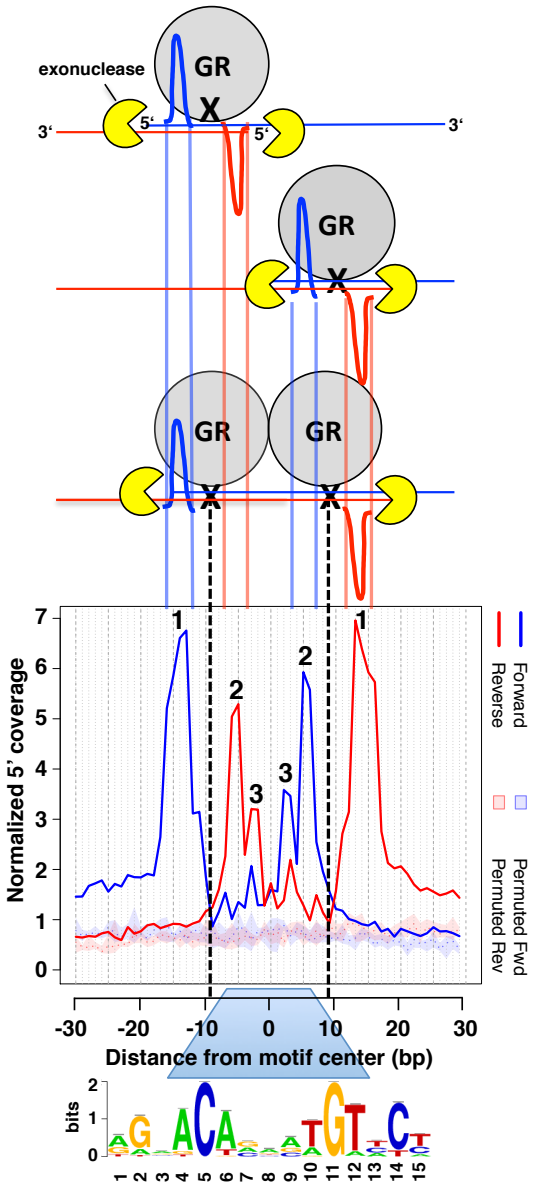

Figure 3.3: Model explaining the footprint profile for GR binding sites. from [Starick et al., 2015]. Inefficiency of cross-linking and monomeric GR binding results in the cross-linking of either one or both GR monomers. The two barriers on each DNA strand (depicted in blue and red) represent the position where the exonuclease stops digesting in proximity of a cross-link. Notably, a population of cells with different cross-link scenarios is analysed, thus explaining the observed footprint profile. Dashed black lines indicate the hypothesised main DNA:GR cross-linking point (in the centre of the peak-pair for each monomer). 1, "outer peaks"; 2 and 3, "inner peaks".

3.2.3 New insights in the binding of glucocorticoid receptor to DNA from ChIP-exo datasets

With S. Meijsing, we used the high resolution of ChIP-exo technology to better understand how GR is recruited to its genomic loci [Starick et al., 2015]. A key finding of our study with ExoProfiler is that GR binds to a broader spectrum of sequences than previously thought, including highly degenerate sequences. Significantly, conventional computational analysis of ChIP-seq peaks, based on sequence overrepresentation, would not identify such degenerate sequences as they are found at similar frequencies at bound and unbound regions. In addition, our study uncovered a TFBS directly recruiting a novel heterodimer of GR and a member of the ETS or TEAD families of TF. ExoProfiler also highlighted indirect binding of GR to DNA via FOX or STAT proteins.

One of the main uncertainty in ChIP-exo data is whether the barriers (stacks of sequence reads) denote exclusively the accumulation of binding signals from multiple cells or could result from PCR artefacts. The recent ChIP-nexus protocol is a simpler variation of ChIP-exo, which moreover takes advantage of barcoding to differentiate true signal from PCR artefacts [He et al., 2015a]. S. Meijsing has already produced several ChIP-nexus datasets targeting GR, which I am currently analyzing.

3.3 Looking back over 8 years of ChIP-seq : quality and biases

The ChIP-seq experiment that yields comprehensive genome-wide binding profile was designed eight years ago. After the initial excitement of this revolutionary approach, it is now time to look back and ask: what is the quality of these profiles? In this section, I will present the guidelines assembled by diverse teams, and highlight the decisive role of the ENCODE project towards producing high-quality datasets. Then, I will introduce various biases, and techniques suggested to correct for them. For a thorough up-to-date review, refer to [Meyer and Liu, 2014].

3.3.1 Producing high-quality datasets

Many experimental approaches are subjected to particular technical biases, and ChIP is no exception. The wide adoption of ChIP-seq by many laboratories, in just a few years, have led to an explosion of datasets, but also of protocols [Arrigoni et al., 2015]. The ENCODE project alone has produced thousands of datasets. Several groups have thus shared their experience to produce high-quality data [Kidder et al., 2011; Chen et al., 2012; Meyer and Liu, 2014]. Because the ENCODE consortium needed to standardize their protocols and computational processing, they have issued some guidelines relevant for ChIP-seq experiments[3] [Landt et al., 2012], discussed in [Nakato and Shirahige, 2016]. These guidelines cover appropriate testing of antibodies, replicates

[3] https://www.encodeproject.org/about/experiment-guidelines/

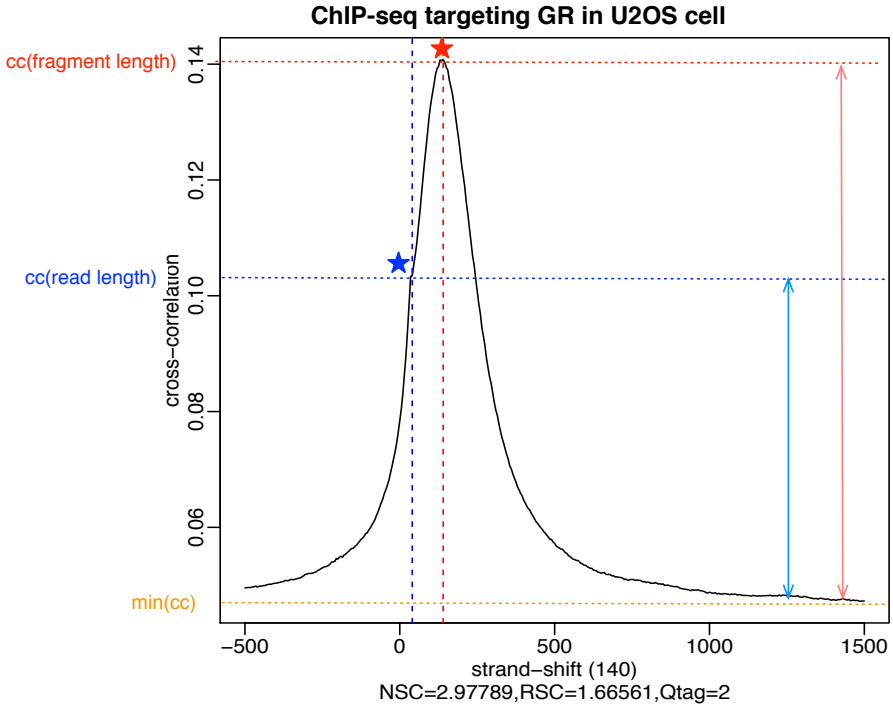

Figure 3.4: Cross-correlation plot for ChIP-seq targeting GR in U2OS cell unpublished. The cross-correlation plot was obtained with phantompeakqualtools (relying on Spp version 1.1) [Marinov et al., 2014] on a dataset produced by S. Meijsing's lab (European Nucleotide Archive accession ERR560463), uniquely mapped reads keeping duplicates. This dataset have a much higher fragment-length peak (red star) than read-length peak (blue star), denoting a high-quality experiment. The NSC and RSC values are accordingly high above the threshold, and the QC (noted Qtag here) is of the highest quality (2).

to assess data variability (two are advised), sequencing depth (minimum 10 million mapped reads in human for TFs, though this might be insufficient for broad histone marks for which up to 60 million reads could be a minimum [Chen et al., 2012]), library complexity (approximately measured by the Non Redundant Fraction (NRF) metric aiming for NRF $\geq$ 0.8 for 10 million reads, which is similar to the 'PCR bottleneck coefficient (PBC)'), and experimental controls to detect and correct biases (input DNA or mock IP 'IgG' specific to each biological context). ENCODE guidelines also include quality metrics to assess the quality of the ChIP-seq datasets :

- Directly on mapped reads: The strand cross-correlation is based on the observation that high-quality experiments produces two densities of reads mapping to the direct and reverse strand, respectively. Two metrics have been defined to assess enrichment in true signal: Normalised Strand Coefficient (NSC) and Relative Strand Correlation (RSC) [Landt et al., 2012]. On a cross-correlation (cc) plot (Fig. 3.4), two peaks are noticeable: fragment-length ('ChIP') peak (red star) and the read-length ('phantom') peak (blue star). A high-quality ChIP-seq dataset have a higher fragment-length peak than read-length peak. The $NSC = cc(fragmentlength)/min(cc)$, NSC $\geq$ 1.05 is recommended.

The $RSC = cc(fragmentlength) - min(cc)/cc(readlength) - min(cc)$, RSC $\geq$ 0.8 is recommended, but QC score have been recently defined to refine this threshold [Marinov et al., 2014], in which RSC $\leq$ 1 has QC=0, 1 $\geq$ RSC $\geq$ 1.5 has QC=1 and RSC $\geq$ 1.5 has QC=2.

- After peak-calling: The Fraction of Reads in Peaks (FRiP) measures the global ChIP enrichment by calculating the fraction of the total mapped reads that fall under a peak. Most of the reads actually correspond to background signal, but in a high-quality ChIP experiment, the FRiP should be at least of 1%. This measure is sensitive to the peak-calling algorithm, and quite dependent on the sequencing depth, so it should only be used as an indicator to perform additional quality checks if the value is under the threshold. The SPOT (Signal Portion of Tags)[4] metrics of ENCODE seems quite similar to FRiP, as it is the percentage of reads that fall in peaks (called 'hotspots' by the authors).

- After peak-calling, with replicates: The Irreproducible Discovery Rate (IDR) is a measure of reproducibility of the detected peaks. The underlying idea is that reproducible peaks should be among the high-ranked peaks and consistent between the two replicates, whereas irreproducible peaks should be among the lower ranks and less consistent. The IDR value is supposed to separate the dataset between reproducible and irreproducible peaks. The peaks above a given IDR threshold (*e.g.* 1%) can be considered reliable. A word of caution however, in that artifactual peaks showing high enrichment (see below) may be reproducible, and thus found among the "reliable" list of peaks.

A retrospective quality assessment of all vertebrate (non-ENCODE) ChIP-seq datasets has also provided important considerations for generating high-quality datasets [Marinov et al., 2014]. Authors used the RSC and NSC values as quality measures and advocate for systematic visual inspection of cross-correlation plots. They stress that the quality metrics proposed by ENCODE are not appropriate for broad peaks, and new metrics should be defined to assess the quality of broad histone marks. Similarly, these metrics should not be directly used to assess datasets where ChIP enrichment is not expected (e.g. knockout TF). Of note, they found that publication in high-impact journals are associated with the largest fraction of low-quality ChIP-seq datasets.

3.3.2 Biases in ChIP experiments

Some groups have investigated the source of biases leading to false enrichment of particular regions [Chen et al., 2012]. Biases in chromatin profiling experiments, including ChIP-seq, have been recently reviewed [Meyer and Liu, 2014], and can be categorised as follows :

[4] http://www.uwencode.org/data/quality/metrics

Chromatin fragmentation: The fragmentation is usually achieved by sonication, which does not cut chromatin homogeneously because of its non-homogenous structure. Indeed, closed chromatin will be more resistant to fragmentation, whereas open chromatin will be easier to cut (Fig. 3.5A). As a consequence, DNA fragments from open chromatin will be more represented, resulting in artificial read enrichment. These sonication-induced biases are specific to each sample, because of its particular chromatin configuration. Some authors advise against using an Input if not sonicated with the ChIP, but admit combining different Input samples if performed in the exact same conditions [Meyer and Liu, 2014].

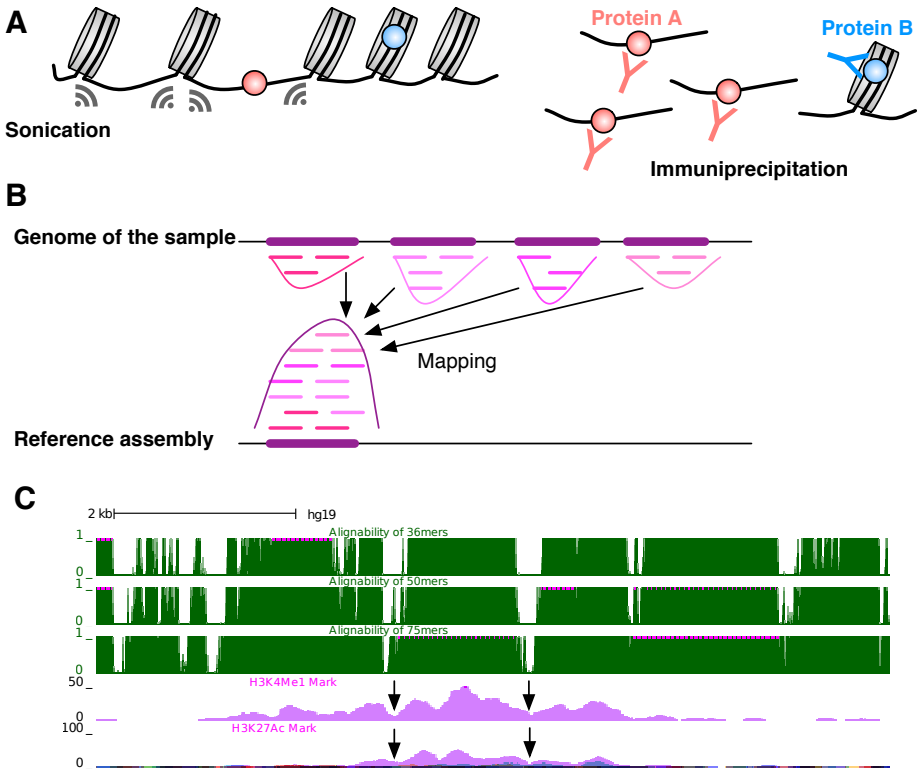

Figure 3.5: unpublished. Panel A was inspired by [Meyer and Liu, 2014] **A.** Chromatin fragmentation bias resulting from a preferential sonication into open-chromatin regions (*e.g.* actively transcribed promoter regions), as tightly packed regions are more difficult to fragment. As a consequence, proteins located in less accessible chromatin regions are more difficult to detect. **B.** Mapping bias resulting from a lower representation of certain regions in the reference assembly compared to the sampled genome (*e.g.* copy number variation). **C.** Low mappability regions results in local decrease of coverage (indicated by arrows). On this illustrative genome browser snapshot, the three green tracks correspond to mappability of 36,50 and 75bp reads as available within the UCSC genome browser; the purple tracks correspond to broad histone marks (H3K4me1 and H3K27ac).

PCR amplification: PCR involves annealing and denaturing DNA fragments at each cycle, which inevitably treat differently fragments of different lengths and with different sequence contents (GC-rich fragments are easier to separate than AT-rich fragments). A well-known issue is the GC bias, consisting in a dependency between the GC content of a region and the number of mapped reads in this region [Benjamini and Speed, 2012]. Study of GC bias in Illumina datasets (this

sequencing platform is used for most ChIP-seq datasets) revealed that this bias results mostly from the GC composition of the full DNA fragment rather than just of the sequenced portion (read), supporting the idea that this bias stems from PCR amplification [Benjamini and Speed, 2012]. It is thus recommended to limit the number of cycles for PCR amplification [Meyer and Liu, 2014].

Mapping: Repetitive elements and differences between the sequenced genome and the reference genome can produce coverage bias in some regions of the genome. For example, cancer cell lines often have extended rearrangements such as duplicated genomic sequences. When mapped on the reference genome, reads originating from different regions additively map to a unique region (Fig. 3.5B). Even in 2016, the human reference genome assembly is still not complete, in particular in centromeric regions where repeat-rich sequences are under-represented [Miga et al., 2015]. Using human data from the 1000 Genome Project, it has been shown that some high ChIP-seq peaks are spurious and correlate with unannotated repeats and high copy number regions [Pickrell et al., 2011].

Another issue stems from local differences of mappability along the genome. Indeed, only uniquely-mapped reads are usually retained during the analysis process. This means that regions of low complexity are predisposed to have a lower coverage of ”uniquely mappable” reads. This problem rather affect broad peaks, such as histone marks, for which the binding profile display an apparent reduced coverage in low-mappability regions (Fig. 3.5C). If mappability is an issue for the experiment (*e.g.* the ChIP-seq targets a TF with TFBSs located in repetitive elements), it is recommended to use longer reads or paired-end reads.

Expression bias/hyper-ChIPable regions: Two influential studies revealed an alarmingly high proportion of artifactual enrichment precisely at highly expressed regions [Teytelman et al., 2013; Park et al., 2013]. These artifactual peaks, termed ’hyper-ChIPable’, are reproducible, appear only for high levels of expression, in any IP dataset (including IgG, and even if the antibody targets a non-DNA binding protein such as the heterologous jellyfish GFP protein) but not particularly in the input control, and cover the whole gene body. Interestingly, this artefact of the ChIP method has been detected in yeast, because the hyper-ChIPable peaks were not corroborating the well-known roles of the studied proteins. Authors of these studies suggest that this phenomenon results from direct or indirect non-specific interactions of the cross-linked proteins with DNA from the open chromatin of highly transcribed regions.

These studies naturally opened a debate. On the post-publication discussion platform Pubpeer[5] and on pubmed[6], Teytelman mentions that for him, cross-linking is not the major problem because lowering the concentration of formaldehyde had minor effects on the detected hyper-ChIPable re-

[5]https://pubpeer.com/publications/591EB69E4EA0D85E6C76D2D9CACC1D
[6]http://www.ncbi.nlm.nih.gov/pubmed/25164749

gions. Yet, various points of the original study [Teytelman et al., 2013] have been cross-examined, in particular the unusually high cross-linking time (1h compared to 10-20min) susceptible to increase the non-specific interactions [Araya et al., 2014]. The second study [Park et al., 2013] nevertheless had a cross-linking time of 15min, which remains in that range. An independent study showed a correlation between increasing cross-linking time and detection of hyper-ChIPable regions, using GFP ChIP-seq (on human cells) [Baranello et al., 2015].

To circumvent this bias, usage of IgG control [Park et al., 2013] or heterologous protein (such as GFP) [Teytelman et al., 2013] in the same conditions as the ChIP (necessary to ensure a similar transcriptome) are advocated to spot the hyper-ChIPable peaks. A cross-linking time of less than 10min would limit this bias [Baranello et al., 2015], in our ChIP-seq targeting GR, S. Meijsing used a cross-linking time of 3min.

TF binding characteristics: Binding of TFs differ in terms of binding affinity, cooperative binding, residence time *in vivo*. Although not mentioned in [Meyer and Liu, 2014], cross-linking is more and more criticised, and its effectiveness varies for different TFs [Gavrilov et al., 2015]. My work on ChIP-exo datasets supports this idea. We noticed that some TFs (such as Fox) cross-link more efficiently, which result in an apparent important signal for these TFs [Starick et al., 2015]. For me, this work at the cross-link resolution was an eye-opener to consider ChIP-seq profiles as the signal of *cross-linked* TFs rather than *bound* TFs.

3.3.3 Which control to use ?

Choosing an appropriate control has been an ongoing debate in recent years [Kidder et al., 2011]. Input DNA aims at controlling for biases due to chromatin fragmentation, as more accessible chromatin [Chen et al., 2012] - or locally more susceptible to shearing because of nucleotide composition [Cheung et al., 2011]- may result in higher background signal. Input DNA is the most commonly used control as (i) it provides a complex library leading to a widely-distributed coverage on the genome [Kidder et al., 2011], and (ii) most peak-calling programs assume that the control is input DNA, and use it to normalise for ChIP enrichment. Input control should be sequenced at higher depth than study samples because of the broader distribution of the reads over the genome [Landt et al., 2012; Chen et al., 2012; Meyer and Liu, 2014]. The IgG (or mock IP) is a non-specific antibody, supposed to provide a better measure of the DNA fragments captured non-specifically by the IP. This control is nevertheless limited by the low complexity of the library, which results from fewer DNA fragments immunoprecipitated by this non-specific antibody [Kidder et al., 2011; Marinov et al., 2014]. Although input has been more widespread, IgG may become more popular, as it can indicate hyper-ChIPable regions [Park et al., 2013]. Importantly, the retrospective study [Marinov et al., 2014] revealed that 20% of control datasets (input and IgG) show artifactual high

enrichment values (comparable to real ChIP datasets) in promoters but also in enhancers, which could be explained by sonication and cross-linking conditions.

Alternative controls are also advocated, like knockout of the targeted TF to account for antibody specificity and non-specific DNA-binding events [Kidder et al., 2011] or antibody against an heterologous protein (such as GFP) [Teytelman et al., 2013]. For ChIP-seq targeting histone modifications, using an antibody against Histone H3 can also serve as control, but it does not provide much difference compared to input [Flensburg et al., 2014]. Adding as control 'spike-in' chromatin from another genome, which should be able to bind the same protein targeted by the antibody, has not been extensively tested yet [Meyer and Liu, 2014]. For inducible TFs, the control can also consist of the uninduced condition. However, such controls are not perfect as induction can alter the chromatin state of the cell [Landt et al., 2012; Meyer and Liu, 2014]. For our studies with GR that is inducible with a hormone, we tested as control the uninduced condition (ethanol vehicle). As GR is not located within the nucleus in absence of the inducing factor, this control was of limited use, as very few sequence reads were finally obtained.

Few teams advocate the use of no control at all, for example by modelling the background from the ChIP dataset, considering that most ChIP reads are background that do not fall under peaks [de Boer et al., 2014]. Having said that, the general view is to rather to take great care of the control (ideally performing several of them), the limitation often being the cost.

3.3.4 Correction with computational approaches

Approaches to correct bias in datasets have been recently reviewed in [Meyer and Liu, 2014]. Of interest, this review also lists some methods to deconvolute ChIP-seq signal when multiple binding sites are located in close proximity, and various normalization techniques proposed to compare ChIP-seq datasets, which is out of the scope of this chapter. I will detail below some approaches to detect and correct for biases, pointing to recently-developed (sometimes unpublished) tools.

Guidelines and quality control: Guidelines for computational analysis of ChIP-seq datasets have been proposed [Bailey et al., 2013; Nakato and Shirahige, 2016]. The current ENCODE phase 3 pipeline specification[7] and implementation[8] can serve as guidance for current analysis pipelines. The quality of a ChIP-seq dataset can first be assessed using the ENCODE metrics (above-mentioned in section 3.1.1: NRF, NSC, RSC, FRiP implemented for example phantompeakqualtools[9] [Marinov et al., 2014] or ChiLin[10] [Meyer and Liu, 2014]. EnCODE-independent tools for quality control of ChIP-seq have been proposed such as CHANCE [Diaz et al., 2012],

[7] https://docs.google.com/document/d/1lG_Rd7fnYgRpSIqrIfuVlAz2dW1VaSQThzk836Db99c/editheading=h.9ecc41kilcvq
[8] https://github.com/kundajelab/TF_chipseq_pipeline
[9] https://code.google.com/p/phantompeakqualtools/
[10] http://cistrome.org/chilin

but haven't been extensively used by the community due to lack of usability and/or updates in this still rapidly changing field. Some newer programs, such as ChIPQC[11] (implemented in R) maybe worth testing.

Correct for GC bias Prior to peak-calling, samples can be individually corrected for mappability variations and GC bias with programs such as BEADS [Cheung et al., 2011], GCcorrect [Benjamini and Speed, 2012] or BIDCHIPS [Ramachandran et al., 2015]. Authors of these studies suggest to correct each sample independently, as different samples can have different GC biases. This means that 'treatment versus control' normalization should not be performed on raw counts in ChIP-seq, but on corrected counts. Some peak-calling program internally perform such correction, but the popular program MACS does not [Meyer and Liu, 2014].

Remove artifactual peaks Although not mentioned in [Meyer and Liu, 2014], the use of 'blacklists' is a common practice to exclude known dubious genomic regions [Bailey et al., 2013; Nakato and Shirahige, 2016]. The unannotated repeats from [Pickrell et al., 2011] were also provided as a blacklist. Nowadays, the ENCODE DAC blacklist[12] provides a consensus list of regions that show artifactual enrichment of reads, independent of the experiment or the cell line. It is worth mentioning that removing these regions affect the cross-correlation plot and quality measures (RSC,NSC), so it is advised to assess quality before filtering with the blacklist [Carroll et al., 2014]. Because some artifactual signals are cell-type specific (*e.g.* resulting from copy number variations in cancer cell lines), methods have been proposed to identify these specific problematic regions using the input control data. This concept has been integrated within the peak-calling program HMCam [Ashoor et al., 2013], or in a method to produce 'greylists'[13] specific to the studied sample.

A complementary approach to filtering with blacklists, proposed only for the human genome so far, is to modify the reference genomic sequences used for mapping, by adding a set of repeat-rich sequences that are under-represented in the reference assembly (the 'sponge database') [Miga et al., 2015].

In the lab, Céline Hernandez has detected a contamination with cDNA, leading to specific enrichment of reads of the same strand on coding exons. Although not discussed in the literature, this is a problem already encountered by other labs. To systematically detect artifactual peaks stemming from such contamination, our analysis pipeline includes a step of mapping with RNA-seq parameters (to detect exon junctions) and searching for enrichment of unidirectional reads mapping coding exons.

The ENCODE and modENCODE consortia have provided guidelines, quality metrics and blacklists for the studied model organisms (human, mouse, drosophila, roundworm). It is worth noting

[11] http://www.bioconductor.org/packages/release/bioc/html/ChIPQC.html
[12] http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeMapability/wgEncodeDacMapabilityConsensusExcludable.b
[13] http://bioconductor.org/packages/release/bioc/html/GreyListChIP.html

that ChIP-seq experiments performed in non-model organisms benefit from this experience, but the thresholds for quality metrics and the most common artefactual regions need to be defined.

3.4 Novel large-scale experiments for protein-DNA binding

3.4.1 Experimental techniques to study TF-DNA interactions

Many methods have been developed over the years to measure the binding affinity and specificity of a TF to specific sequences, and to obtain genome-wide binding profiles of this TF. The measuring techniques can be divided into *in vitro* and *in vivo* approaches, and small-scale versus large-scale approaches (reviewed in [Stormo and Zhao, 2010; Dey et al., 2012; Levo and Segal, 2014]). Briefly, classical *in vitro* techniques include footprinting assay with DNAseI, Electrophoretic mobility shift assay (EMSA), SELEX, which have been modified into large-scale experiments (HT-SELEX). Protein-binding microarrays (PBM) are also a popular high-throughput *in vitro* technique. While these techniques offer important information on the binding affinity and specificity of a given TF, they do not take into account the local chromatin context, recruitment of co-factors, or the tridimensional structure of the chromatin. Conversely, *in vivo* techniques include ChIP (introduced in section 3.1.1) and DNA adenine methyltransferase identification (DAMID).

As discussed above, ChIP-seq has some limitations: a large number of cells is required, the resolution is about 200bp-500bp and thus does not directly point to the TFBS, and cross-linking introduces some biases. There is thus intense research in developing variations on the ChIP protocol and alternative methods to circumvent these issues (reviewed in [Zentner and Henikoff, 2014; Mahony and Pugh, 2015]). I will highlight below some of these new approaches that may become the new standards in the upcoming years.

3.4.2 Methods for limited number of cells

Standard ChIP-seq requires a large number of cells (10 million), which limit its application to precious samples (e.g. transient developmental cell types, samples requiring to sacrifice many animals). Protocols based on PCR amplification have been developed to reduce the number of cells to 10 000 (Nano-ChIPeq, used for histone modification and LinDA, for both TFs and histone modications). More recently and without additional PCR amplification, ChIPmentation [Schmidt et al., 2015] enables ChIP-seq on 100 000 cells, by using the same hyperactive Tn5 transposase as used for ATAC-seq, a recent technique similar to DNase I sensitivity assay that runs on low input (50 000 cells) [Zentner and Henikoff, 2014]. Another variation succeeds on 10 000-100 000 cells, termed NEXSON, and moreover simplifies the first steps of the ChIP-seq protocol with an efficient extraction of nuclei from formaldehyde-fixed cells [Arrigoni et al., 2015].

3.4.3 Methods for improving resolution

I presented above in section 3.2 the ChIP-exo approach, and its recent variation called ChIP-nexus. High-resolution X-ChIP-seq also provides base-pair resolution [Skene and Henikoff, 2015] by digesting unprotected DNA with MNase until the position of the cross-link, producing barriers similar to ChIP-exo. Computational analysis is supposed to be simpler for high-resolution X-ChIP-seq than ChIP-exo. The ORGANIC protocol (see below) also provides base-pair resolution. Note that DNase-seq and ATAC-seq reveal regions of open chromatin, but their use to detect TFBS at high resolution (digital genomic foot printing) remains controversial (reviewed in [Madrigal, 2015; Sung et al., 2016]).

3.4.4 Methods without cross-links

The cross-linking agent formaldehyde is thought to create artefacts in ChIP-seq data, due to non-specific interactions with DNA (including transient protein-DNA interactions in highly transcribed regions, resulting in hyper-ChIPable regions [Teytelman et al., 2013; Park et al., 2013]) and protein-protein cross-links (discussed in [Zentner and Henikoff, 2014]). Native ChIP (without cross-linking) would thus be more suited, and may become more standard (discussed in [Zentner and Henikoff, 2014]). The ORGANIC protocol combines native ChIP on TFs with high-resolution, similar to high-resolution X-ChIP-seq [Kasinathan et al., 2014]. It is based on low-salt concentration of the buffers, which fixes protein-DNA interactions in a non-covalent manner. A major advantage of this method is the absence of bias from sonication or highly-transcribed regions, removing the need for an input or IgG control. However, it seems that ORGANIC has not been adopted by other labs yet.

DamID is an alternative to ChIP that also does not require cross-linking agents, and does not use antibody, but its main drawback is that it necessitates the TF to be fused to the DAM enzyme. A recent variation of the protocol, Split DamID, is promising to address the problem of detecting binding events of a TF and its co-factor (or dimeric versus monomeric binding), *in vivo*, on native chromatin, for a limited number of cells (10 000) [Hass et al., 2015].

Whichever method will become widely-adopted by the community and replace current ChIP-seq, we can foresee some of its features : (i) limited number of experimental steps that can be standardized and highly reproducible for any types of samples, (ii) cost-effective, (iii) working on a low number of cells, (iv) *in vivo*, (v) on native chromatin, (vi) at base-pair resolution and (vii) with a straightforward computational processing of the data.

3.5 Current projects and perspective

During this transition to the Big Data era, I have been more and more involved in data analyses, mainly from ChIP-seq and ChIP-exo experiments. I grasped the importance and difficulties to find a fulfilling, fair and motivating collaboration with skilled experimentalists. I will thus pursue my fruitful collaboration with S. Meijsing on the glucocorticoid receptor model. I will develop new collaborations as initiated with Pascale Gilardi-Hebenstreit (IBENS, Paris) on the development of the hindbrain in vertebrates, or on haematopoiesis, a central interest of D. Thieffry group. I am already participating to this topic by supervising PhD students for their analyses of genome-scale functional datasets. Samuel Collombet focuses on the reprogramming of B cells into macrophages with Thomas Graf (CRG, Barcelona). Otoniel Rodriguez Jorge directed by Angelica Santana (UAEM, Cuernavaca, Mexico), models T cell activation in neonates. Because more and more datasets must be processed, it is important to have at hand a ChIP-seq processing pipeline that ensures reproducibility of the analyses, traceability and reliability. I am involved in the development of such a pipeline that already facilitate our analyses, and will leave us more time for downstream project-specific analyses.

3.5.1 Towards additional insights in the binding of GR to DNA

In addition to three published studies (cf. sections 3.1.3 and 3.2.3), we are currently finalising two studies conducted by Stefanie Shoene, the PhD student that S. Meijsing and I have been co-tutoring. The first study aims at better understanding which sequence features in TFBSs distinguish GR binding events resulting in the regulation of a gene, from non-regulating binding events. We show that the recognition sequence, specifically of the nucleotides directly flanking the core binding site, differs depending on the strength of GR-dependent activation of nearby genes (Fig. 3.6). Computational and structural studies indicate that these flanking nucleotides change the three-dimensional structure of both the DNA binding site, the tertiary structure of the DNA binding

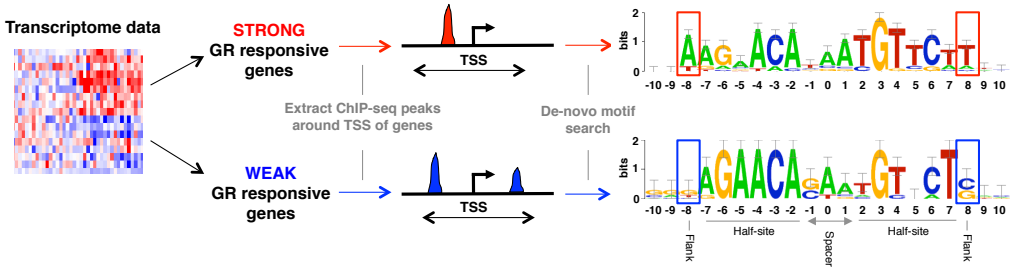

Figure 3.6: Identification of high-activity GBS variants. from manuscript in preparation. Overview of the workflow to identify candidate high-activity GBS variants. Genes were grouped into strong (top 20% highest fold induction) and weak transcriptional responders to dexamethasone treatment. Next, ChIP-seq peaks in a 40kb window centered on the TSS of responder genes were extracted for each group and subjected to *de-novo* motif searches resulting in the depicted motifs. The flank positions (-8 and +8) important for activity of the regulated genes are highlighted by red (A/T) or blue (G/C) boxes.

domain of GR and the quaternary structure of the dimeric complex. The manuscript is currently in revision in the journal Nature Communications. The second study is taking advantage of the STARR-seq experimental approach [Arnold et al., 2013] to decipher the activity of a large quantity of GBS variants. Apart from these studies, we are currently testing the ChIP-nexus protocol to follow-up on our ChIP-exo study.

3.5.2 ChIP-seq processing pipeline using Eoulsan

ChIP-seq data processing requires connecting multiple tools, developed by various teams worldwide, into coherent analyses workflows. These tools need to be replaced by newer versions or concurrent programs very often, due to the rapid developments in the field. Although in-house scripts allow quick development of pipelines, they are not appropriate to ensure reproducibility of the analyses, traceability and reliability. The Genomics Platform at IBENS (headed by Stéphane Le Crom) have developed the Eoulsan framework [Jourdren et al., 2012] to address these problems for RNA-seq pipelines. We have been adapting it to ChIP-seq analyses, for which I have been supervising Céline Hernandez (Bioinformatics engineer), Pierre-Marie Chiaroni (Master2 student) and Cédric Michaut (Master1 student).

3.5.3 ChIP-seq targeting histone modifications

Histone modifications provide information on the state of the chromatin, which can then be interpreted in terms of functional regions. For example, genomic regions bearing the H3K27ac mark are interpreted as enhancer or promoter regions. These histone modifications thus link the epigenomic status with transcriptional regulation. ChIP-seq datasets targeting histone modifications are widespread in the public databases. It is not always possible to obtain a specific antibody against a transcription factor, but there are commercial antibodies against histone modifications. Besides, it is sometimes necessary to first detect the enhancer regions, in order to uncover the transcription factors binding to these regions. This is the case in the ANR-BmBF iBone project (2014-2017, coordinator: Eric Hesse), aiming at studying epigenomic remodelling in osteoporosis. This project will produce ChIP-seq datasets targeting histone modifications, which will allow to detect the enhancer and promoter regions, in which we will perform motif analysis to uncover the potential transcription factors. As the histone modification peaks are very large (often more than 10kb), motif discovery does not perform well on such datasets. Two M2 students (Pierre-Marie Chiaroni and Roberto Tirado Magallanes) have assessed an approach to counter this problem. They have tested tools that perform combinations of multiple histone modification to segment the genome into different states (*e.g.* chromHMM), to find a particular enhancer-like state. This successfully reduces the search space to perform subsequent motif analyses.

Chapter 4
Concluding remarks

Altogether, my work has contributed to the development of bioinformatics tools publicly available to the community (Table 4.1).

Analysis tool	Task	Publication
HoxPred	detection of Hox and ParaHox genes, classification in homology groups	[Thomas-Chollier et al., 2007] [Thomas-Chollier and Ledent, 2008] [Thomas-Chollier et al., 2010]
Eoulsan	processing of ChIP-seq datasets	in preparation
RSAT	motif discovery, pattern-matching, motif comparisons, retrieval of sequences, motif analysis in large-scale datasets	[Turatsinze et al., 2008] [Sand et al., 2008] [Thomas-Chollier et al., 2008] [Sand et al., 2009] [Medina-Rivera et al., 2011] [Thomas-Chollier et al., 2011a] [Thomas-Chollier et al., 2012a] [Thomas-Chollier et al., 2012b] [Medina-Rivera et al., 2015]
TRAP	motif analysis (affinity-based)	[Thomas-Chollier et al., 2011b]
ExoProfiler	motif-based ChIP-exo analysis	[Starick et al., 2015]

Table 4.1: Sequence analysis tools for which I contributed

These tools are complementary and can be jointly applied. For example, I used HoxPred to find the Hox genes in the sea anemone genome, then RSAT to extract the sequences located upstream the genes and to search for Hox binding sites [Hudry et al., 2014]. Some of the tools are intrinsically interconnected: ExoProfiler's motif analysis depends on RSAT Web services, and Eoulsan will soon be directly connected with RSAT *peak-motifs*, to enable a fully automatised ChIP-seq analysis pipeline. Apart from Eoulsan, which is only accessible at the command-line, all other tools were developed with user-friendly interfaces to enable their usage by non-experts.

Regarding Biology, my work has also contributed to a better understanding of (i) the evolution of Hox and ParaHox gene families across metazoans, and of (ii) the transcriptional regulation by the glucocorticoid receptor. Regarding the first point, my contributions consist in proposing models for the evolution of Posterior and Central Hox genes in bilaterians, for the origin of Hox and ParaHox in early metazoans. I have contributed to the detection of new putative Hox genes, and to the improvement of Hox gene annotation. Regarding the second point, my collaboration with S. Meijsing's experimental lab has uncovered new sequences that are driving or preventing the binding of GR (Fig. 4.1). Apart from the classical GR recognition of its consensus sequence (Fig. 4.1A), ChIP-exo experiments have revealed that GR can recognise more degenerate sequences

than anticipated (Fig. 4.1B), and likely forms a heterodimer with ETS or TEAD proteins (Fig. 4.1C). ChIP-seq experiments have allowed us to show that the gamma isoform of GR recognises specific motifs, which are not recognised by the alpha isoform (Fig. 4.1D). ChIP-seq experiments have also unravelled motifs that are under-represented, and that prevent the binding of GR to nearby regulatory elements (Fig. 4.1E).

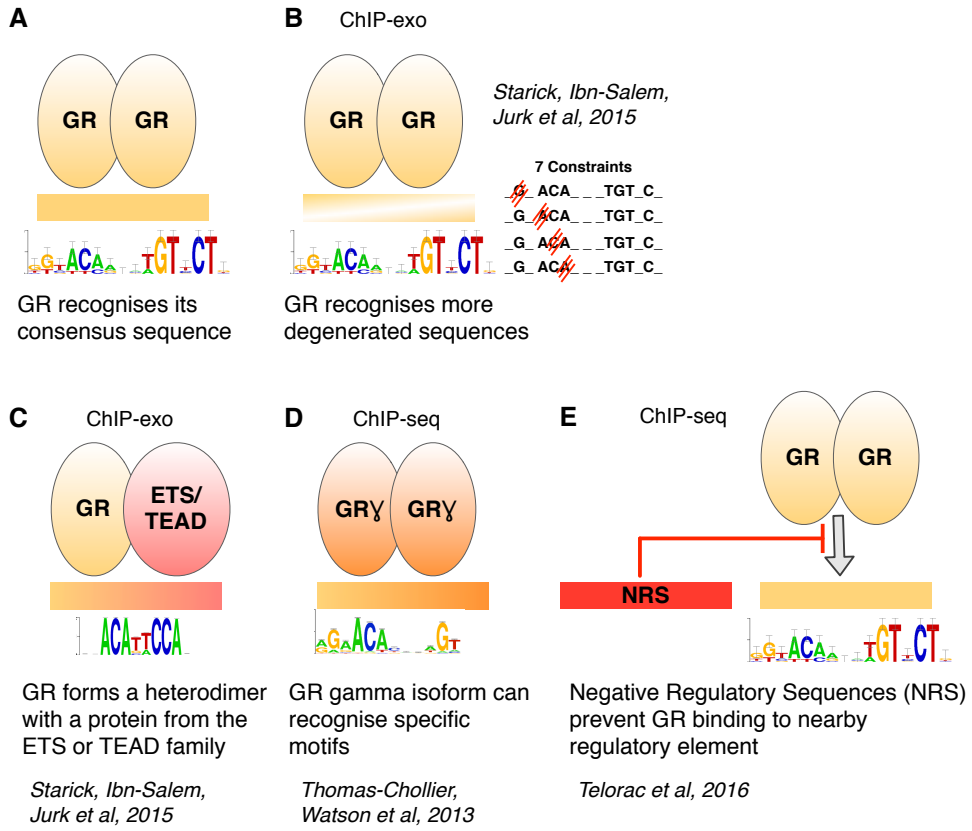

Figure 4.1: Summary of sequences that affect the binding of GR. Panel A depicts the classical recognition of the GR consensus sequence. The other panels summarise the new motifs and likely binding mechanisms uncovered in our work. The experiment (ChIP-seq or ChIP-exo) is indicated, as well as the corresponding publication.

We now have at hand accumulating functional genomics data that provide us with information on which genes are expressed, in response to which regulatory signals, within a particular three dimensional organisation of DNA, and even reaching now the single-cell resolution. Most of these datasets relate to vertebrates, and more particularly human and mouse. Although of huge interest to construct gene regulatory networks (GRN) of TFs with key roles in developmental processes, these datasets are limited when one wants to study the evolution of these GRN across the animal kingdom. To this end, it is already possible to scan genomic sequences to predict regulatory elements in non-model organisms (as I did with basal metazoan genomes to search for Hox motifs [Hudry et al., 2014]), but the results are often deceiving because of the poor quality of assemblies and annotations. Millions of plant and animal genomes are expected in the next ten

years [Stephens et al., 2015], but what will be the quality of these assemblies and annotations? We already see a tendency for decreasing qualities, even for 'key' genomes such as Ctenophore, because producing high-quality and well-annotated assemblies requires a lot of time and effort. Apart from sequence assemblies and annotation, the results of TFBS prediction are hampered by a high number of false predictions. Tentatively, more and more functional genomics datasets (*e.g.* chromatin accessibility with ATAC-seq, or ChIP-seq targeting histone modifications) will be available for non-model organisms, which will reduce the false positives and enable a more systematic study of the evolution of developmental GRN across metazoans. There is currently only one ChIP-seq dataset for a non-model organism (histone modifications in the sea anemone *N. vectensis*) in the Gene Expression Omnibus database. Yet, computational methods for *de novo* assembly of ChIP-seq fragments have already been proposed, to perform motif analysis in ChIP-seq without reference genome assembly [He et al., 2015b].

We have now entered the Big Data era, and more and more datasets will be produced in the near future. I cannot help but wonder : what will we do with these zettabytes of data? Advanced techniques to integrate hundreds of datasets, and extract more information are growing fast (such as "deep learning", actively developed by Google, which released an open source machine learning infrastructure[1]). Yet, increasing our knowledge in Biology should remain the driving force, which could be achieved by exploring already-produced datasets in innovative ways, rather than systematically producing increasing amount of datasets. Using these datasets to define dynamical models for regulatory networks can offer a mechanistic understanding of the biological system, but methods need to be developed to facilitate such integration. Data accessibility poses challenges too, so that raw and processed data continue to be freely available to the community, despite increasing number of datasets (of increasing sizes!). Small labs are already confronted by the problem of disk space cost. To reduce disk usage, a reasonable solution may be to investigate novel systems dedicated to data compression of sequence files.

Ethical and legal aspects of genomic sequences are unfortunately lagging behind, as the technological advances happen at an unprecedented fast pace. What will happen in 2025 when up to 2 billion humans will have their genome sequenced [Stephens et al., 2015], with the technology to edit genomes at hand? I often wonder if we are heading for Gattaca[2].

[1] TensorFlow, november 2015

[2] Gattaca,1997, science fiction film directed by Andrew Niccol, depicting a society where parental genomes are genetically manipulated before *in-vitro* fertilisation, to engender enhanced individuals

References

Aldridge, S., Watt, S., Quail, M. A., Rayner, T., Lukk, M., Bimson, M. F., Gaffney, D., and Odom, D. T. (2013). AHT-ChIP-seq: a completely automated robotic protocol for high-throughput chromatin immunoprecipitation. *Genome Biology*, 14(11):R124.

Araya, C. L., Kawli, T., Kundaje, A., Jiang, L., Wu, B., Vafeados, D., Terrell, R., Weissdepp, P., Gevirtzman, L., Mace, D., Niu, W., Boyle, A. P., Xie, D., Ma, L., Murray, J. I., Reinke, V., Waterston, R. H., and Snyder, M. (2014). Regulatory analysis of the C. elegans genome with spatiotemporal resolution. *Nature*, 512(7515):400–405.

Arnold, C. D., Gerlach, D., Stelzer, C., Boryń, M., Rath, M., and Stark, A. (2013). Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science (New York, NY)*, 339(6123):1074–1077.

Arrigoni, L., Richter, A. S., Betancourt, E., Bruder, K., Diehl, S., Manke, T., and Bönisch, U. (2015). Standardizing chromatin research: a simple and universal method for ChIP-seq. *Nucleic Acids Research*.

Ashoor, H., Hérault, A., Kamoun, A., Radvanyi, F., Bajic, V. B., Barillot, E., and Boeva, V. (2013). HMCan: a method for detecting chromatin modifications in cancer samples using ChIP-seq data. *Bioinformatics (Oxford, England)*, 29(23):2979–2986.

Bailey, T., Krajewski, P., Ladunga, I., Lefebvre, C., Li, Q., Liu, T., Madrigal, P., Taslim, C., and Zhang, J. (2013). Practical guidelines for the comprehensive analysis of ChIP-seq data. *PLoS computational biology*, 9(11):e1003326.

Baranello, L., Kouzine, F., Sanford, S., and Levens, D. (2015). ChIP bias as a function of cross-linking time. *Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology*.

Benjamini, Y. and Speed, T. P. (2012). Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Research*, 40(10):e72.

Boeva, V. (2016). Analysis of Genomic Sequence Motifs for Deciphering Transcription Factor Binding and Transcriptional Regulation in Eukaryotic Cells. *Frontiers in genetics*, 7:24.

Cannon, J. T., Vellutini, B. C., Smith, J., Ronquist, F., Jondelius, U., and Hejnol, A. (2016). Xenacoelomorpha is the sister group to Nephrozoa. *Nature*, 530(7588):89–93.

Carroll, T. S., Liang, Z., Salama, R., Stark, R., and de Santiago, I. (2014). Impact of artifact removal on ChIP quality metrics in ChIP-seq and ChIP-exo data. *Frontiers in genetics*, 5:75.

Chen, Y., Negre, N., Li, Q., Mieczkowska, J. O., Slattery, M., Liu, T., Zhang, Y., Kim, T.-K., He, H. H., Zieba, J., Ruan, Y., Bickel, P. J., Myers, R. M., Wold, B. J., White, K. P., Lieb, J. D., and Liu, X. S. (2012). Systematic evaluation of factors influencing ChIP-seq fidelity. *Nature Methods*, 9(6):609–614.

Cheung, M.-S., Down, T. A., Latorre, I., and Ahringer, J. (2011). Systematic bias in high-throughput sequencing data and its correction by BEADS. *Nucleic Acids Research*, 39(15):e103–e103.

Collas, P. and Dahl, J. A. (2008). Chop it, ChIP it, check it: the current status of chromatin immunoprecipitation. *Frontiers in Bioscience-Landmark*, 13:929–943.

de Boer, B. A., van Duijvenboden, K., van den Boogaard, M., Christoffels, V. M., Barnett, P., and Ruijter, J. M. (2014). OccuPeak: ChIP-Seq peak calling based on internal background modelling. *PLoS ONE*, 9(6):e99844.

Dey, B., Thukral, S., Krishnan, S., Chakrobarty, M., Gupta, S., Manghani, C., and Rani, V. (2012). DNA-protein interactions: methods for detection and analysis. *Molecular and cellular biochemistry*, 365(1-2):279–299.

Diaz, A., Nellore, A., and Song, J. S. (2012). CHANCE: comprehensive software for quality control and validation of ChIP-seq data. *Genome Biology*, 13(10):R98.

Ding, J., Hu, H., and Li, X. (2014). SIOMICS: a novel approach for systematic identification of motifs in ChIP-seq data. *Nucleic Acids Research*, 42(5):e35.

Flensburg, C., Kinkel, S. A., Keniry, A., Blewitt, M. E., and Oshlack, A. (2014). A comparison of control samples for ChIP-seq of histone modifications. *Frontiers in genetics*, 5:329.

Furey, T. S. (2012). ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. *Nature reviews Genetics*, 13(12):840–852.

Gavrilov, A., Razin, S. V., and Cavalli, G. (2015). In vivo formaldehyde cross-linking: it is time for black box analysis. *Briefings in functional genomics*, 14(2):163–165.

Hass, M. R., Liow, H.-h., Chen, X., Sharma, A., Inoue, Y. U., Inoue, T., Reeb, A., Martens, A., Fulbright, M., Raju, S., Stevens, M., Boyle, S., Park, J.-S., Weirauch, M. T., Brent, M. R., and Kopan, R. (2015). SpDamID: Marking DNA Bound by Protein Complexes Identifies Notch-Dimer Responsive Enhancers. *Molecular Cell*, 59(4):685–697.

He, Q., Johnston, J., and Zeitlinger, J. (2015a). ChIP-nexus enables improved detection of in vivo transcription factor binding footprints. *Nature Biotechnology*, 33(4):395–401.

He, X., Cicek, A. E., Wang, Y., Schulz, M. H., Le, H.-S., and Bar-Joseph, Z. (2015b). De novo ChIP-seq analysis. *Genome Biology*, 16:205.

Hudry, B., Thomas-Chollier, M., Volovik, Y., Duffraisse, M., Dard, A. e. l., Frank, D., Technau, U., and Merabet, S. (2014). Molecular insights into the origin of the Hox-TALE patterning system. *eLife*, 3:e01939.

Jourdren, L., Bernard, M., Dillies, M.-A., and Le Crom, S. (2012). Eoulsan: a cloud computing-based framework facilitating high throughput sequencing analyses. *Bioinformatics (Oxford, England)*, 28(11):1542–1543.

Kasinathan, S., Orsi, G. A., Zentner, G. E., Ahmad, K., and Henikoff, S. (2014). High-resolution mapping of transcription factor binding sites on native chromatin. *Nature Methods*, 11(2):203–209.

Kidder, B. L., Hu, G., and Zhao, K. (2011). ChIP-Seq: technical considerations for obtaining high-quality data. *Nature reviews Genetics*, 12(10):918–922.

Landt, S. G., Marinov, G. K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., Bernstein, B. E., Bickel, P., Brown, J. B., Cayting, P., Chen, Y., DeSalvo, G., Epstein, C., Fisher-Aylor, K. I., Euskirchen, G., Gerstein, M., Gertz, J., Hartemink, A. J., Hoffman, M. M., Iyer, V. R., Jung, Y. L., Karmakar, S., Kellis, M., Kharchenko, P. V., Li, Q., Liu, T., Liu, X. S., Ma, L., Milosavljevic, A., Myers, R. M., Park, P. J., Pazin, M. J., Perry, M. D., Raha, D., Reddy, T. E., Rozowsky, J., Shoresh, N., Sidow, A., Slattery, M., Stamatoyannopoulos, J. A., Tolstorukov, M. Y., White, K. P., Xi, S., Farnham, P. J., Lieb, J. D., Wold, B. J., and Snyder, M. (2012). ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome research*, 22(9):1813–1831.

Levo, M. and Segal, E. (2014). In pursuit of design principles of regulatory sequences. *Nature reviews Genetics*, 15(7):453–468.

Lihu, A. and Holban, Ş. (2015). A review of ensemble methods for de novo motif discovery in ChIP-Seq data. *Briefings in Bioinformatics*, 16(6):964–973.

Liu, E. T., Pott, S., and Huss, M. (2010). Q&A: ChIP-seq technologies and the study of gene regulation. *BMC Biology*, 8:56.

Madrigal, P. (2015). On Accounting for Sequence-Specific Bias in Genome-Wide Chromatin Accessibility Experiments: Recent Advances and Contradictions. *Frontiers in bioengineering and biotechnology*, 3:144.

Mahony, S. and Pugh, B. F. (2015). Protein-DNA binding in high-resolution. *Critical Reviews in Biochemistry and Molecular Biology*, 50(4):269–283.

Mardis, E. R. (2007). ChIP-seq: welcome to the new frontier. *Nature Methods*, 4(8):613–614.

Marinov, G. K., Kundaje, A., Park, P. J., and Wold, B. J. (2014). Large-scale quality analysis of published ChIP-seq data. *G3 (Bethesda, Md.)*, 4(2):209–223.

Meyer, C. A. and Liu, X. S. (2014). Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. *Nature reviews Genetics*, 15(11):709–721.

Miga, K. H., Eisenhart, C., and Kent, W. J. (2015). Utilizing mapping targets of sequences underrepresented in the reference assembly to reduce false positive alignments. *Nucleic Acids Research*, 43(20):e133.

Nakato, R. and Shirahige, K. (2016). Recent advances in ChIP-seq analysis: from quality management to whole-genome annotation. *Briefings in Bioinformatics*.

Park, D., Lee, Y., Bhupindersingh, G., and Iyer, V. R. (2013). Widespread misinterpretable ChIP-seq bias in yeast. *PLoS ONE*, 8(12):e83506.

Park, P. J. (2009). ChIP-seq: advantages and challenges of a maturing technology. *Nature reviews Genetics*, 10(10):669–680.

Pickrell, J. K., Gaffney, D. J., Gilad, Y., and Pritchard, J. K. (2011). False positive peaks in ChIP-seq and other sequencing-based functional assays caused by unannotated high copy number regions. *Bioinformatics (Oxford, England)*, 27(15):2144–2146.

Ramachandran, P., Palidwor, G. A., and Perkins, T. J. (2015). BIDCHIPS: bias decomposition and removal from ChIP-seq data clarifies true binding signal and its functional correlates - Springer. *Epigenetics & chromatin*, 8:33.

Rhee, H. S. and Pugh, B. F. (2011). Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell*, 147(6):1408–1419.

Schmidt, C., Rendeiro, A. F., Sheffield, N. C., and Bock, C. (2015). ChIPmentation: fast, robust, low-input ChIP-seq for histones and transcription factors. *Nature Methods*, 12(10):963–965.

Serandour, A. A., Brown, G., Cohen, J. D., and Carroll, J. S. (2013). Development of an Illumina-based ChIP-exonuclease method provides insight into FoxA1-DNA binding properties . *Bioinformatics (Oxford, England)*, 27(7):1017–1018.

Skene, P. J. and Henikoff, S. (2015). A simple method for generating high-resolution maps of genome-wide protein binding. *eLife*, 4:e09225.

Starick, S. R., Ibn-Salem, J., Jurk, M., Hernandez, C., Love, M. I., Chung, H.-R., Vingron, M., Thomas-Chollier, M., and Meijsing, S. H. (2015). ChIP-exo signal associated with DNA-binding motifs provides insight into the genomic binding of the glucocorticoid receptor and cooperating transcription factors. *Genome research*, 25(6):825–835.

Stephens, Z. D., Lee, S. Y., Faghri, F., Campbell, R. H., Zhai, C., Efron, M. J., Iyer, R., Schatz, M. C., Sinha, S., and Robinson, G. E. (2015). Big Data: Astronomical or Genomical? *PLoS Biology*, 13(7):e1002195.

Stormo, G. D. and Zhao, Y. (2010). Determining the specificity of protein-DNA interactions. *Nature reviews Genetics*, 11(11):751–760.

Sung, M.-H., Baek, S., and Hager, G. L. (2016). Genome-wide footprinting: ready for prime time? *Nature Methods*, 13(3):222–228.

Telorac, J., Prykhozhij, S. V., Schöne, S., Meierhofer, D., Sauer, S., Thomas-Chollier, M., and Meijsing, S. H. (2016). Identification and characterization of DNA sequences that prevent glucocorticoid receptor binding to nearby response elements. *Nucleic Acids Research*.

Teytelman, L., Thurtle, D. M., Rine, J., and van Oudenaarden, A. (2013). Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins. *Proceedings of the National Academy of Sciences*, 110(46):18602–18607.

Thomas-Chollier, M., Darbo, E., Herrmann, C., Defrance, M., Thieffry, D., and van Helden, J. (2012a). A complete workflow for the analysis of full-size ChIP-seq (and similar) data sets using peak-motifs. *Nature Protocols*, 7(8):1551–1568.

Thomas-Chollier, M., Herrmann, C., Defrance, M., Sand, O., Thieffry, D., and van Helden, J. (2012b). RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets. *Nucleic Acids Research*, 40(4):e31.

Thomas-Chollier, M., Watson, L. C., Cooper, S. B., Pufall, M. A., Liu, J. S., Borzym, K., Vingron, M., Yamamoto, K. R., and Meijsing, S. H. (2013). A naturally occuring insertion of a single amino acid rewires transcriptional regulation by glucocorticoid receptor isoforms. *Proceedings of the National Academy of Sciences of the United States of America*, 110(44).

Tompa, M., Li, N., Bailey, T. L., Church, G. M., De Moor, B., Eskin, E., Favorov, A. V., Frith, M. C., Fu, Y., Kent, W. J., Makeev, V. J., Mironov, A. A., Noble, W. S., Pavesi, G., Pesole, G., Régnier, M., Simonis, N., Sinha, S., Thijs, G., van Helden, J., Vandenbogaert, M., Weng, Z., Workman, C., Ye, C., and Zhu, Z. (2005). Assessing computational tools for the discovery of transcription factor binding sites. *Nature Biotechnology*, 23(1):137–144.

Tran, N. T. L. and Huang, C.-H. (2014). A survey of motif finding Web tools for detecting binding site motifs in ChIP-Seq data. *Biology Direct*, 9:4.

Zambelli, F., Pesole, G., and Pavesi, G. (2013). Motif discovery and transcription factor binding sites before and after the next-generation sequencing era. *Briefings in Bioinformatics*, 14(2):225–237.

Zentner, G. E. and Henikoff, S. (2014). High-resolution digital profiling of the epigenome. *Nature reviews Genetics*, 15(12):814–827.

Morgane Thomas-Chollier

mthomas@biologie.ens.fr

Maître de conference ENS
département de Biologie CNU-65

Summary

Born: 06 August 1979 – 36 years
Nationality: French
Training: Master in Biology, PhD in Bioinformatics
Current position: Associate Professor (MCU) at Ecole normale supérieure (ENS), Paris, France
Research interests: transcriptional regulation, high-throughput functional genomics, development and evolution

Education

PhD in Bioinformatics **2004-2008**

Jointly at Vrije Universiteit Brussels (VUB) and Université Libre de Bruxelles (ULB), Brussels, Belgium. Grade: greatest distinction

Master in Bioinformatics and Applied Genomics (M2) **2003**

Ecole Supérieure de Biotechnologie de Strasbourg (European School of the Higher Rhine Universities), Louis Pasteur University, Strasbourg, France. Passed with honours (equivalent to French mention Bien)

Master in Cellular Biology and Physiology (M1) **2002**

"Maîtrise en Biologie cellulaire et Physiologie" at University of Orléans, France. Passed with honours First Class (equivalent to French mention Très Bien)

Research activities

Associate Professor **Since September 2012**

ENS Paris, France. Biology department. Laboratory of Computational Systems Biology (headed by Denis Thieffry).

Postdoctoral Fellow **2009-2012**

Max Planck Institute for Molecular Genetics (MPIMG), Berlin, Germany. Department of Computational Molecular Biology. Supervision : Martin Vingron. **Alexander von Humboldt** foundation and **Max Planck Society** fellowships**.**

Postdoctoral researcher **2008-2009**

ULB, Brussels, Belgium. Bioinformatique des Génomes et des réseaux. Supervision: Jacques van Helden.

PhD thesis **2004-2008**

ULB and VUB, Brussels, Belgium. Supervisors: Luc Leyns (Laboratory of Cell Genetics, VUB), Jacques van Helden (Bioinformatique des Génomes et des réseaux, ULB), and Valérie Ledent (Belgian EMBnet Node). "Evolutionary study of the Hox gene family with matrix-based bioinformatics approaches".

Graduate thesis **2003 (6months)**

CNRS (French National Centre for Scientific Research) in the Microbiology and Genetics Laboratory in Strasbourg. Supervisor: Stéphane Vuillemier. "Study of the putative role of glutathione S-transferases in rhizobacterial genomes".

Awards

2008 Award from the foundation « Alice et David Van Buuren»
2009 Postdoc fellowship from the Alexander von Humboldt foundation

Collaborations

Sebastiaan Meijsing and Albert Poustka, MPIMG Berlin (Germany), Steven Johnsen, Göttingen (Germany), Jacques van Helden, TAGC Marseille (France), Pascale Gilardi-Hebenstreit, Stéphane Le Crom, Hugues Roest-Crollius, IBENS Paris (France), Samir Merabet, IGFL, Lyon (France), Thierry Lepage, IBV, Nice (France), Pedro Martinez, Universitat de Barcelona (Spain), Max Telford, UCL, London (United Kingdom), Angelica Santana, UAEM, Cuernavaca (Mexico).

Publications in peer-reviewed international journals

*= co-first author / #= corresponding author

1. Telorac J, Prykhozhij SV, Schöne S, Meierhofer D, Sauer S, Thomas-Chollier M#, Meijsing SH#. "Identification and characterization of DNA sequences that prevent glucocorticoid receptor binding to nearby response elements". ***Nucleic Acids Research***, in press **2016**
2. Hossan T, Nagarajan S, Baumgart SJ, Xie W, Tirado Magallanes R, Hernandez C, Chiaroni P, Indenbirken D, Spitzner M, Thomas-Chollier M, Grade M, Thieffry D, Grundhoff A, Wegwitz F, Johnsen SA. "The Histone Chaperone SSRP1 is Essential for Wnt Signaling Pathway Activity During Osteoblast Differentiation". ***Stem Cells***, in press **2016**
3. Thomas-Chollier M, Martinez P. "The origin of metazoan patterning systems and the role of ANTP-class homeobox genes". ***eLS***, *John Wiley Sons Ltd, Chichester. http://www.els.net* **2016**
4. Medina-Rivera A*, Defrance M*, Sand O*, Herrmann C, Castro-Mondragon J, Delerce J, Jaeger S, Blanchet C, Vincens P, Caron C, Staines DM, Contreras-Moreira B, Artufel M, Charbonnier-Khamvongsa L, Hernandez C, Thieffry D, Thomas-Chollier M#, van Helden J#. "RSAT 2015: Regulatory Sequence Analysis Tools". ***Nucleic Acids Research***, *43(W1):W50-W56* **2015** (invited submission to the Web Server issue)
5. Starick S*, Ibn-Salem J*, Jurk M*, Hernandez C, Love MI, Chung H, Vingron M, Thomas-Chollier M#, Meijsing SH#. "ChIP-exo signal associated with DNA-binding motifs provide insights into the genomic binding of the glucocorticoid receptor and cooperating transcription factors", ***Genome Research***, *25(6):825-35* **2015**
6. Hudry B, Thomas-Chollier M, Volovik Y , Duffraisse M, Dard A, Dale F, Technau U, Merabet S. "Molecular insights into the origin of the Hox-TALE patterning System", ***eLife***, *3:e01939* **2014**
7. Thomas-Chollier M*, Watson L*, Cooper S, Pufall MA, Liu JS, Borzym K, Vingron M, Yamamoto K.R, Meijsing SH. "A naturally occurring single amino acid insertion rewires transcriptional regulation by Glucocorticoid receptor isoforms", ***Proc. Natl. Acad. Sci. U. S. A.***, *110(44):17826-31* **2013**
8. Thomas-Chollier M#, Darbo E, Herrmann C, Defrance M, Thieffry D, van Helden J#. "From peaks to motifs: a complete workflow for the analysis of full-size ChIP-seq (and similar) datasets", ***Nature Protocols***, *7:1551–1568* **2012**
9. Thomas-Chollier M, Herrman C, Defrance M, Sand O, Thieffry D, van Helden J. "RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets", ***Nucleic Acids Research***, *40(4)* **2012**
10. Thomas-Chollier M#, Hufton A, Heinig M, O'Keeffe S, El Masri N, Roider HG, Manke T, Vingron M. "Transcription factor binding predictions using TRAP for the analysis of ChIP-seq data and regulatory SNPs", ***Nature Protocols***, *6(12):1860-9* **2011**
11. Thomas-Chollier M#, Defrance M, Medina-Rivera A, Sand O, Herrmann C, Thieffry D, van Helden J#. "RSAT 2011: Regulatory Sequence Analysis Tools", ***Nucleic Acids Research***, *39:W86-91* **2011**
12. Medina-Rivera A, Abreu-Goodger C, Thomas-Chollier M, Salgado H, Collado-Vides J, van Helden J. "Theoretical and empirical quality assessment of transcription factor-binding motifs", ***Nucleic Acids Research***, *39(3):808-24* **2011**
13. Thomas-Chollier M#, Ledent V, Leyns L, Vervoort M "A non-tree-based comprehensive study of metazoan Hox and ParaHox genes prompts new insights into their origin and evolution ", ***BMC Evolutionary Biology***, *10:73* **2010** *(highly accessed)*
14. Sand O, Thomas-Chollier M, van Helden J. "Retrieve-ensembl-seq: user-friendly and large-scale retrieval of single or multi-genome sequences from Ensembl.", ***Bioinformatics***, *15;25(20):2739-40* **2009**
15. Thomas-Chollier M*, Sand O*, Turatsinze J-V, Janky R, Defrance M, Vervisch E, van Helden J "RSAT: Regulatory Sequence Analysis Tools", ***Nucleic Acids Research***, *36:W119-W127* **2008**
16. Sand O, Thomas-Chollier M, Vervisch E , van Helden J. "Analyzing multiple datasets by inter-connecting RSAT programs via SOAP Web Services – an example with ChIP-chip data", ***Nature Protocols***, *3:10* **2008**
17. Turatsinze J-V*, Thomas-Chollier M*, Defrance M, van Helden J. "Using RSAT to scan genome sequences for transcription factor binding sites and cis-regulatory modules", ***Nature Protocols***, *3:10* **2008**
18. Thomas-Chollier M#, Ledent V "Comparative phylogenomic analyses of teleost fish Hox gene clusters: lessons from the cichlid fish Astatotilapia burtoni: comment", ***BMC Genomics***, *9:35* **2008**
19. Thomas-Chollier M, Leyns L, Ledent V "HoxPred: automated classification of Hox proteins using combinations of generalised profiles", ***BMC Bioinformatics***, *8:247* **2007**

20. Simionato E, Ledent V, Richards G, Thomas-Chollier M, Kerner P, Coornaert D, Degnan BM, Vervoort M. "Origin and diversification of the basic helix-loop-helix gene family in metazoans: insights from comparative genomics.", ***BMC Evolutionary Biology****, 7:33* **2007** *(highly accessed)*

1025 citations, h-index 12 (Google scholar).
Article n°1 features Stefanie Schoene, the PhD student that I co-supervise with S. Meijsing.
Article n°2 features Roberto Tirado Magallanes and Pierre-Marie Chiaroni, two M2 students that I supervised.
Articles n°2,4,5 features Céline Hernandez, a bioinformatics engineer that I supervised.
Article n°5 features as co-first author Jonas Ibn-Salem, a M2 student that I co-supervised with S. Meijsing.
Article n°6 was mentioned in a perspective article: Ferrier D "Evolutionary developmental biology: The Hox-TALE has been wagging for a long time" **eLife** 3:e02515 2014.

Book Chapter

Leyns L, Piette D and Thomas-Chollier M. (2005). "Evo-devo: paleontologie zonder fossielen?". In : *Evolutie vandaag*, Brussel, VUBPRESS, 125-146 (in Dutch).

Software

HoxPred (http://cege.vub.ac.be/hoxpred): complete development of a classification tool for Hox and ParaHox genes.
RSAT (http://rsat.ulb.ac.be/rsat/): contribution to the development of programs to perform motif scanning, motif discovery in ChIP-seq datasets and motif quality. Design of Web Service access, user interface and analysis workflows.
TRAP (http://trap.molgen.mpg.de): increased the usability of the program, and development of a website.
ExoProfiler (https://github.com/ComputationalSystemsBiology/ExoProfiler): prototyping, supervision of development.

International Conferences

With invitation:

- Bringing Maths to life. October 2015, Naples, Italy.
- 15th Evolutionary Biology Meeting. September 2011, Marseille, France.

With oral presentation:

- *Condition-specific Binding of the Glucocorticoid Receptor*
  Sixth Annual RECOMB/ISCB conference on Regulatory and Systems Genomics. November 2013, Toronto, Canada.
- *RSAT peak-motifs: efficient prediction of transcription binding sites from genome-wide peak sets*
  The next NGS challenge. May 2013, Valencia, Spain.
- *New insights into the origin and evolution of Hox and ParaHox genes*
  3rd Euro Evo Devo Conference (EED). July 2010, Paris, France.

Poster presentations (selection as senior author):

- R Tirado Magallanes, C Hernandez, D Thieffry, M Thomas-Chollier *Evaluation of a probabilistic partitioning approach to systematically refine ChIP-seq peaks location*. [BC]2. June 2015, Basel, Switzerland.
- PM Chiaroni, D Thieffry, M Thomas-Chollier *Prediction of transcription factor motifs and binding sites from multiple histone mark ChIP-seq datasets*. European Conference on Computational Biology (ECCB14). September 2014, Strasbourg, France.

Participation with poster presentations:

- EpiGeneSys: Annual Meeting. November 2014, Barcelona, Spain.
- European Conference of Computational Biology (ECCB/ JOBIM). September 2014, Strasbourg, France.
- Journees Ouvertes de Biologie, Informatique et Mathematiques (JOBIM). July 2013, Toulouse, France.
- SIG regulatory Genomics (ISMB13/ECCB13). July 2013, Berlin, Germany.
- European Conference of Computational Biology (ECCB). September 2010, Ghent, Belgium.
- HOX and TALE homeoproteins in Development and Disease. May2009, Carmona, Spain.
- European Conference of Computational Biology (ECCB). November 2008, Cagliari, Italy.
- 2nd Euro Evo Devo Conference (EED). August 2008, Ghent, Belgium.
- Journées Ouvertes de Biologie, Informatique et Mathématiques (JOBIM). July 2008, Lille, France.
- Benelux Bioinformatics Conference (BBC2007). November 2007, Leuven, Belgium.
- Journées Ouvertes de Biologie, Informatique et Mathématiques (JOBIM). July 2007, Marseille, France.
- Annual meeting of the International Society for Computational Biology (ISMB). June 2005, Detroit, USA.

National conferences

With invitation:

- Journées COMATEGE-SeqBio. November 2015, Orsay, Paris.

With oral presentation:

- *Condition-specific Binding of the Glucocorticoid Receptor*
  Journees Ouvertes de Biologie, Informatique et Mathematiques (JOBIM). July 2015, Clermont-Ferrand, France.

Invited workshops and seminars

Workshops

Abroad:

- Workshop Ecole normale supérieure / National University of Singapore «Joint ENS-CSI Workshop on ChIP-seq data analysis ». March 2014 (3 days), Singapore.
- 1-day Workshop VIB-Bits « Hands-on introduction to ChIP-Seq analysis ». February 2014, May 2015, Leuven, Belgium. Reinvitation for May 2016.

In France:

- Ecole de Bioinformatique AVIESAN « Initiation au traitement des données de génomique obtenues par séquençage à haut débit ». November 2013 (1 week) + October 2014 (1 week), Roscoff. Reinvitation for November 2016.
- NUS-ENS workshop « Novel genome-wide approaches to decipher transcriptional and epigenetic regulation in mammalian cells ». May 2013 (1 day), Paris.
- INSERM workshop « Approches bioinformatique pour décrypter la régulation des génomes ». October 2011 (1 week), Bordeaux.

Seminars

Abroad:

- University Medical Center Hamburg – Eppendorf. October 2015, Hamburg, Germany.
- Centro de Ciencas Genomicas. March 2015, Cuernavaca, Mexico.
- Universidad Autónoma del Estado de Morelos (UAEM). March 2015,Cuernavaca, Mexico.
- Max Planck Institute for Molecular Genetics. November 2014, Berlin, Germany.
- Genome Institute of Singapore. March 2014, Singapore.
- Genetics and Genome Biology - Sick Kids. November 2013, Toronto, Canada.
- Centro de Ciencas Genomicas. March 2010, Cuernavaca, Mexico.

In France:

- UPMC. December 2015, Paris.
- IGFL. December 2015, Lyon.
- CGM. June 2015, Gif-sur-Yvette.
- IGBMC. May 2015, Strasbourg.
- IGBMC. June 2013, Strasbourg.
- Réseau RENABI ChIP-seq, Institut Curie. June 2013, Paris.
- Regional network of bioinformatics engineers. June 2013, Lille.
- Ecole normale supérieure Paris. October 2011, Paris.

Round table

- Invitation by the association "Jeunes Bioinformaticiens de France (JeBiF)" to round tables with master students: « Working abroad ». March 2014, Paris. « Les domaines de la bioinformatique», December 2015, Orsay.

Research supervision

Co-supervision of PhD students: 3

- since 2011: Stefanie Schöene, student at MPIMG Berlin (rate: 50%; S. Meijsing: 50%). In 2015, she obtained the UNESCO-L'Oréal prize « For Women in Science Deutschland ».
- since sept. 2014: Samuel Collombet, student at IBENS (rate: 25%; D. Thieffry: 75%). Participation to his pre-doctoral supervision since December 2012.
- since january. 2016: Céline Hernandez, student at IBENS (rate: 50%; D. Thieffry: 50%).

Supervision of master (M2) students: 5

- 2013: Daniela Garcia (rate: 50%; D. Thieffry: 50%) and Jonas Ibn-Salem (ERASMUS student Frei Universität Berlin, rate: 75%; S. Meijsing: 25%).

- 2014: Pierre-Marie Chiaroni (rate: 90%; D. Thieffry: 10%) and Amhed Vargas Velazquez, (rate: 20%, D. Thieffry: 80%).
- 2015: Roberto Tirado Magallanes (rate 100%).

Supervision of undergraduate students: 4

- 2011: Jan Patrick Pett (L3 student, rate: 100%).
- 2014: Benoit Noël (M1 student, rate: 100%).
- 2016: Cédric Michaut (M1 student, rate: 100%) and Geoffray Brelurut (M1 student, rate: 40%)

Supervision of bioinformatics engineer: 1

- july 2014-december2015: Céline Hernandez (rate: 80%; D. Thieffry: 20%).

Teaching

Ecole normale supérieure (full time, amounting to 192h/year practical teaching "equivalent TD")
L3: Bioinformatics
L3: Biology for non-biologists
M1: Computational Biology
M1+M2: training in maths and informatics
M2: Computational analysis of cis-regulatory elements

Training for researchers (formation permanente)
"Introduction à l'analyse des données de séquençage à haut debit", (3h) UPMC, 2013, 2014.
"Gentle introduction to command-line", (1 day) bioinformatics platform ENS, 2014.
"Using published high-throughput datasets", (1 day) bioinformatics platform ENS, 2014.
"Cours d'Analyse des genomes", (2h) Institut Pasteur Paris. 2013, 2014,2015.

Research responsibilities

Grants obtained

- 2013: ANR-BMBF French-German funding as partner. Title of research project: *Integrative Biology of Osteoanabolic Networks in the Epigenome* (*iBone*). Recruitment of a bioinformatics engineer.
- 2014: ANR (appel à projet générique) as partner. Title of research project: *Characterisation of novel regulators of dorsal-ventral axis formation upstream and downstream of Nodal in the sea urchin and modelling of the gene regulatory network activated by Nodal* (*echiNodal*). Recruitment of a bioinformatics engineer in July 2015.
- 2016: Institut Français de Bioinformatique as partner. Title of research project: *Coupling Genomicus and RSAT to analyze large-scale functional genomics datasets and prioritize candidates for experimental validation (RSATicus)*. Recruitment of a bioinformatics engineer in July 2016.
- 2016: COST as partner. Title: *Gene Regulation Ensemble Effort for the Knowledge* Commons (GREEKO).

Participation to the writing of several grant applications obtained in the name of D. Thieffry (2 ITMO, 1 Merlion)

Bioinformatic software
Co-responsible with Jacques van Helden of the maintenance and scientific direction of the software suite Regulatory Sequence Analysis Tools (RSAT, http://rsat.eu). This bioinformatics suite of tools, which enables analysis of cis-regulatory regions in genome sequences, is very popular (15,000 requests per months) and available through 6 public servers in Europe and Mexico. Last year, we were invited to present the latest developments in the Web Server issue of Nucleic Acids Research. In addition to participation to the maintenance and tool developments, I am also involved in training users.

Organisation of workshops

- Co-organisation with Jacques van Helden of the one-day France Génomique workshop on ChIP-seq. April 2015, Paris, France.
- Co-organisation with Jacques van Helden of the one-day workshop « Analysis of cis-regulatory motifs from high-throughput sequence sets » within the conference JOBIM 2014 / ECCB 2014. September 2014, Strasbourg, France.
- Co-organisation with Sebastiaan Meijsing of the 3 days French-German workshop « Mechanisms of transcriptional and epigenetic regulation ». November 2014, Berlin, Germany. Obtained a financial support from the CNRS-MPG research network in System Biology (GDRE SysBio).

Referee
International journals: *Nucleic Acids Research*, *Bioinformatics*, *PLoS One*, *PLoS Computational Biology*, *BMC Evolutionary Biology*, *Journal of Experimental Zoology part B* and *Mammalian Genomes*.

Teaching responsibilities

- Responsible for 2 teaching modules at M2 level.
- Co-responsible of 3 teaching modules at L3 and M1 levels.
- Participation to establishing of a novel module in L3/M1/M2 entitled "soft skills" (writing reports, design posters, giving oral presentations, understanding the organisation of academic research in France,…).
- Co-responsible of the re-organisation of the website of the Department of Biology of ENS (2013-2015): integration, rationalisation and complete rethinking of all previous research and teaching websites, recruitment and weekly meetings of a web designer, regular presentations to the board of directors.
- Supervision of 4 teaching assistants annually (moniteurs).
- Member of the selection committee for student applications to ENS Biology department in M1 and M2.
- National exam to enter ENS (Concours des ENS) 2014: test of the questions for the Biology written entry exam (BCPST), participation to the correction of this written exam and to admissibility jury.

Administrative responsibilities

- Elected member in the council of the Société Française de BioInformatique (SFBI) since 2014, vice-president since September 2015.
- Elected member of the "conseil d'institut" of IBENS since 2015.
- Organisation committee of the Ecole de Bioinformatique AVIESAN (program, organisation, selection of participants, responsible for the ChIP-seq workshop) since 2014.
- Program committee for the RegGenSIG (ISMB satellite meeting) 2014-2015.
- Responsible for organizing the best presentation/poster awards at JOBIM 2015.
- Participation as jury member for recruiting an associate professor (MCU) in Toulouse, 2015.
- Participation as jury member for recruiting an associate professor in Brussels, 2016.
- PhD committee of Wolfgang Kopp (MPIMG Berlin, Germany) since 2014, Yuvia Perez Rico (UPMC, Paris) and Damien Monet (UPMC, Paris) since 2015, Antonin Thiebaut (UPMC, Paris) since 2016.

Information on career

- 6 months of maternity leave in 2012