Comparative genomics and the emergence of evolutionary innovations in Vertebrates

Camille Berthelot

Habilitation à diriger des recherches

université PARIS-SACLAY

Comparative genomics and the emergence of evolutionary innovations in Vertebrates

Habilitation à diriger des recherches
de l'Université Paris-Saclay

présentée et soutenue à Paris, le 26/09/2022, par

Camille BERTHELOT

Composition du jury

Daniel J. MACQUEEN Professor, University of Edinburgh	Rapporteur
Judith ZAUGG Group leader, EMBL Heidelberg	Rapportrice
Gilles FISCHER DR CNRS, Sorbonne Université	Rapporteur
Odile LECOMPTE Professeur, Université de Strasbourg	Examinatrice
Olivier LESPINET Professeur, Université de Paris Saclay	Examinateur
Marie SEMON MCU HDR, ENS Lyon	Examinatrice

English summary

Genome evolution is the source of much of the transmissible phenotypic variability we observe within and between species. Over the past ten years, the advancements of sequencing technologies have made it possible to explore and compare the genomes of multiple species to understand how evolution acts on vertebrate genomes, but also how the genomes of these diverse species relate to our own. Evidence shows that vertebrate genomes evolve at multiple scales, from base pair substitutions to large-scale rearrangements of chromosomal structure, and these modifications can all result in functional changes in gene products and expression programs. During my scientific career, I have studied vertebrate genome evolution across multiple clades, timescales and levels of resolution in order to better understand how genomic changes can result in evolutionary novelty. This habilitation thesis summarizes my work on the evolution of genome organisation in paleopolyploid fishes and vertebrates, and on the evolution of gene expression in mammals and fishes. Since September 2021, I lead the Comparative Functional Genomics group at Institut Pasteur. In the final section of the manuscript, I discuss how my lab is heading forward to uncover the functional mechanisms of evolutionary innovations in primates and other mammalian groups.

Keywords – Evolutionary genomics, comparative genomics, phylogenomics, ancestral genomes, comparative transcriptomics, evolution of gene regulation

Résumé en français

L'évolution du génome est la source de la majorité de la variabilité phénotypique transmissible observée entre individus et entre espèces. Au cours des dernières années, l'avancée des technologies de séquençage a rendu possible l'exploration et la comparaison des génomes de nombreuses espèces pour éclairer les mécanismes par lesquels l'évolution façonne les génomes de vertébrés, mais aussi à quel degré ces génomes sont similaires au nôtre. L'observation montre que les génomes de vertébrés évoluent à de multiples échelles, de la substitution de base aux réarrangements à large échelle de la structure des chromosomes, et toutes ces changements sont susceptibles de modifier fonctionnellement les gènes ou leurs programmes d'expression. Au cours de ma carrière scientifique, j'ai étudié l'évolution des génomes de vertébrés à travers différents clades, échelles de temps et niveaux de résolution pour mieux comprendre comment ces changements génomiques permettent l'apparition d'innovations évolutives. Cette these d'habilitation résume mes travaux sur l'évolution de l'organisation du génome chez les poisons paleopolyploïdes et chez les vertébrés, ainsi que sur l'évolution de l'expression des gènes chez les mammifères et les poissons. Depuis septembre 2021, je dirige le groupe Génomique Fonctionnelle Comparative à l'Institut Pasteur. Dans la dernière section du manuscrit, je discute des directions futures de mon laboratoire pour illuminer les mécanismes fonctionnels d'innovations évolutives chez les primates et d'autres groupes de mammifères.

Mots-clés – Génomique évolutive, génomique comparative, phylogénomique, génomes ancestraux, transcriptomique comparative, évolution de la régulation génique

Table of contents

Curriculum Vitae

Personal information

Name	Camille Berthelot
Date of birth	21 April 1985 (37 yo)
Work address	Institut Pasteur – CNRS UMR 3525 – INSERM UA12 Comparative Functional Genomics group Bâtiment François Jacob 26-02-08A 25-28 rue du Docteur Roux 75724 Paris Cedex 15 France
E-mail	camille.berthelot@pasteur.fr
Phone	+33 1 86 46 79 46
Website	https://research.pasteur.fr/en/member/camille-berthelot/
ORCID	0000-0001-5054-2690

Research positions

From 2021	**Junior group leader** Institut Pasteur, Paris, France – CNRS UMR 3525 – INSERM UA12
From 2016	**Tenured Research Scientist (CRCN)** National Institute for Health and Medical Research (INSERM, France) Institut de Biologie de l'Ecole normale supérieure (IBENS), Paris, France – CNRS UMR 8197 – INSERM U1024
2012 – 2015	**Postdoctoral fellow** European Molecular Biology Laboratory – European Bioinformatics Institute (EMBL-EBI), Cambridge, UK – *PI: Paul Flicek*
2009 – 2012	**PhD student** Institut de Biologie de l'Ecole normale supérieure (IBENS), Paris, France – CNRS UMR 8197 – INSERM U1024. *PI : Hugues Roest Crollius*
2007 – 2008	**Master student** Institut Jacques Monod, Paris, France. *PI* : *Eva-Maria Geigl*

Education

2009 – 2012	**PhD in Comparative Genomics and Bioinformatics** *Thesis*: Evolutionary mechanisms of gene order perturbation in vertebrate genomes. *Supervisor*: Hugues Roest Crollius, Institut de Biologie de l'Ecole Normale Supérieure, Paris, France *Jury members*: Daniel Gautheret, Nadia El-Mabrouk, Herve Isambert, Aoife McLysaght, Eric Tannier. *Defense:* 28/09/2012 https://www.theses.fr/2012PA112192
2006 – 2007	**Biology & Geology teaching degree** ("Agregation") – National teaching exam, rank 17th
2004 – 2008	**Bachelor and Master** in Molecular Biology Ecole Normale Supérieure de Lyon, France National entry exam, rank 30th.

Grants

As Principal Investigator

2021 – 2026 **Institut Pasteur G5 Grant** – Junior group leadership
2020 – 2024 **European Research Council (ERC**) – Starting Grant EVOMENS
2016 – 2017 **INSERM Starting Grant**

As Collaborator (as part of the DYOGEN lab)

2019 – 2022 **INSERM Cross-cutting Program** – Genomic Variability
PI: Emmanuelle Génin (INSERM UMR 1087, Brest, France)
WP leaders : Hugues Roest Crollius, Anne-Louise Leutenegger
2017 – 2021 **Agence Nationale de la Recherche (ANR)** – Collaborative Project GenoFish
PI : Yann Guiguen (INRAe Rennes, France)

As Supervisor

2021 – 2022 **EndoFrance** – EndoMens pilot project
Fondation pour la Recherche sur l'Endométriose – EndoMens pilot project
Pilot research grants to Axelle Brulport (postdoc)

Major collaborations

From 2021 Jérémy Terrien and Aude Anzeraey (MNHN Paris)
From 2020 Steven Knafo (APHP – Kremlin Bicetre Hospital) and Julien Bouvier (I2BC, Paris Saclay)
From 2020 Ludivine Doridot (Institut Cochin, Paris) and Angela Goncalves (DKFZ Heidelberg, Germany)
From 2019 Lyne Fellmann (Simian Laboratory, U. Strasbourg) and Pau Molina Vila (Primatology station, CNRS UAR 846, Rousset sur Arc)
From 2019 Pascal Rihet, Benoît Ballester and Aitor Gonzalez (TAGC, Marseille)
2019 – 2021 Ingo Braasch (U. Michigan, USA)
Outcome: *Thompson et al. Nature Genetics 2021*
2016 – 2018 Melody Clark (British Antarctic Survey, Cambridge UK)
Outcome: *Berthelot et al. GBE 2019*
From 2012 Duncan Odom (CRUK, Cambridge, UK – now DKFZ Heidelberg, Germany) and Diego Villar (CRUK, Cambridge, UK – now QMUL London, UK)
Outcome: *Parey et al., in progress*
Berthelot, Villar et al. Nature Eco Evo 2018
Villar, Berthelot et al. Cell 2015
From 2010 Yann Guiguen (INRAe Rennes, France)
Outcome: *Parey et al., submitted*
Parey et al., submitted
Parey et al. MBE 2020
Pasquier et al. JEZB 2018
Berthelot et al. Nature Comm 2014
2009 – 2013 Kerstin Howe and Derek Stemple (Sanger Institute, Cambridge UK)
Outcome: *Howe et al. Nature 2013*

Scientific responsibilities

2021	**Conference co-chair**, GRC for Ecological and Evolutionary Genomics (delayed to 2023 due to COVID-19)
2018 – 2019	**Scientific committee member**, JOBIM (national French conference for bioinformatics)
2018 – 2019	**Scientific committee member**, Young Researchers in Life Sciences
2017	**Conference vice-chair**, GRC for Ecological and Evolutionary Genomics

Memberships

Member of the French Society for Bioinformatics (SFBI)
Member of the Society for Molecular Biology and Evolution (SMBE)
Member of the GDR BioSimia (French network for biomedical research on non-human primates)
Member of the GDR BIM (French network for bioinformatics and mathematical models for biology)

Review activities

Invited peer review for *Cell* (3), *Nature* (4), *Science* (1), *Nature Genetics* (2), *Nature Communications* (3), *Nature Ecology and Evolution* (4), *AJHG* (1), *eLife* (1), *Cell Reports* (1), *Molecular Biology and Evolution* (3), *Genome Biology and Evolution* (2) and others.

Grant peer review - BBSRC UK (2021).

Juries and committees

Recruitment

2021	Jury member for the recruitment of two tenured research scientists, INRAe
	Jury member for PhD fellowships, Sorbonne Université, Paris
2020	Jury member for the recruitment of an assistant professor, Sorbonne Université, Paris
	Jury member for the recruitment of an assistant professor, U. Aix-Marseille
2016	Jury member for the recruitment of two assistant professors, U. Pierre et Marie Curie, Paris

PhD juries

2022	Examinator for Alexandre Laverré, LBBE, Lyon
2019	Defense opponent for Gareth Gillard, CIGENE, Oslo, Norway
	Examinator for Martin Silvert, Institut Pasteur, Paris
	Defense opponent for Srinidhi Varadharajan, CIGENE, Oslo, Norway
	Examinator for Victoire Baillet, IBENS, Paris
2018	Thesis reviewer for Céline Le Béguec, U. Rennes 1

PhD committees

From 2021	Scientific expert for Adrien Dufour, INRAe Toulouse
	Mentor for Claire Lansonneur, IBENS, Paris
From 2020	Scientific expert for Salomé Nashed, Sorbonne Université, Paris
2019 – 2021	Doctoral school liaison for Anna Ferraioli, Sorbonne Université, Paris
	Mentor for Federica Mantica, CRG Barcelona, Spain

Scientific expert for Alexandre Laverré, LBBE, Lyon

Supervision and Mentoring

Postdoctoral scientists

From 2020	Axelle Brulport
	Malgorzata Gazda

PhD students

2018 – 2021	Elise Parey (75% - Sorbonne Université)

Master students

From 2021	Bruno Raquillet (30% - M2 Sorbonne Université)
	Eulalie Liorzou (50% - M2 ENS Paris)
2017	Baptiste Ameline (100% - M2 Université de Nantes – now a postdoc at the University Hospital of Basel, Switzerland)
2015	Céline Le Béguec (100% - M2 Université de Rennes – now a lecturer at IUT Saint-Brieuc)

Undergraduate students

2014	Alissa Williams (100% - Wofford College, USA – now a postdoc at Vanderbilt University, USA)
2012	Judith Abecassis (100% - L3 ENS Paris – now a postdoc at Mines Paris Tech)

EMBL Alumni Mentorship program

From 2021	Ioannis Sarropoulos, PhD student, University of Heidelberg, Germany

Teaching

From 2021	**Formal course and hands-on training** - "Machine Learning for Genetic Variant Interpretation" for Masters/PhD students (5hrs/year), PSL, Paris
2021	**Guest lecture**, Human Population Genomics and Genetic Epidemiology course for Masters/PhD students, Institut Pasteur, Paris
2018	**Guest lecture**, DESC Cytogenetics, Institut Imagine, Paris
From 2017	**Formal course** – "Karyotype evolution in vertebrates" (3hrs/year), MD curriculum, Research track, Université de Paris
2017	**Guest lecture**, “Computational analysis of cis-regulatory sequences” course for Master students, ENS Paris
2017	**Hands-on bioinformatics projects**, L3 students, ENS Paris
2012 – 2015	**Pratical course** - "ChIP-seq data analysis" for 1rst-year Ph.D students (18 hrs/year), EMBL, Heidelberg, Germany
2008 – 2011	**Pratical course** – "Molecular biology 101" for 2nd-year Bachelor students (64 hrs/year), Université Paris Diderot, France, as a teaching assistant (“monitrice”)

Outreach

2021	**Podcast interview** by Lucas Denis, Université de Bourgogne Franche-Comté
2020	**Introductory lecture** - “Genomic medicine, a challenge at the interface of biology and computer science” for L3 students, ENS Paris
2015	**Pint of Science**, Cambridge, UK (general public lecture)

2013	**The Female of the Species**, EMBL-EBI, UK (outreach for female early-career researchers)

Representation

2018 – 2021	Elected representative for research staff, IBENS Council
2012 – 2014	Elected Postdoc representative, EMBL-EBI, UK
2010 – 2012	Elected Student representative, IBENS Council

Oral communications

Conferences

2022	Conference Jacques Monod "Origins of Metazoans", Roscoff, France – **invited**
2021	Groupement de Recherche BioSimia, Paris, France – **invited**
	Groupement de Recherche BIM, Lyon, France – **invited**
	AQuatic Models for Human Disease (AQMHD), online – **invited**
	ISMB/ECCB, online – **invited**
	Alignments and Phylogeny (AlPhy), online – **invited**
2019	International Conference on Integrated Salmonid Biology (ICISB), Edinburgh, UK – **invited**
	French Clinical Genetics Society, Lille, France – **invited**
2018	European Society for Human Genomics, Milan, Italy – **invited**
2017	Gordon Research Conference on Ecological and Evolutionary Genomics, Biddeford, USA – **invited**
	The Biology of Genomes, Cold Spring Harbor, USA
2016	JOBIM, Lyon, France – **highlight presentation**
2015	The Biology of Genomes, Cold Spring Harbor, USA
	CASIM V, Cambridge, UK – **invited**
2014	Livestock Genomics – Genome Informatics Satellite Meeting, Cambridge, UK – **invited**
	Society for Molecular Biology and Evolution, San Juan, Puerto Rico
2012	Groupe de Travail en Génomique Comparative, Lille, France – **invited**
	Genome Informatics, Cambridge, UK
	Otto Warburg Summer School for Evolutionary Genomics
2011	JOBIM, Rennes, France – **best talk award**
	Alignments and Phylogeny (AlPhy), Marseille, France

Invited seminars (selected)

2021	Centre for Genomic Regulation, Barcelona, Spain
	Institut de Génomique Fonctionnelle de Lyon, France
	Université de Lille, France
2020	Museum Koenig, Leibniz, Germany
	National Autonomous University of Mexico
2019	University of Oslo, Norway
	FAANG workshop, EU COST actions, Ljubljana, Slovenia
	Laboratoire de Biométrie et Biologie Evolutive, Lyon, France
	Institut Curie, Paris, France
	Hôpital de la Pitié Salpétrière, Paris, France
2018	Centre of Integrative Genomics, Oslo, Norway
2017	Université Pierre et Marie Curie, Paris, France

Publications

Summary	
Number of publications:	15
as first/co-first author:	6
as last/co-last author:	4
H-index:	11
Total number of citations:	5,747 *(source: Google Scholar)*

* joint first authors

joint corresponding or senior authors

1. E. Parey, A. Louis, J. Montfort, O. Bouchez, C. Roques, C. Iampietro, J. Lluch, A. Castinel, C. Donnadieu, T. Desvignes, C. Floi Bucao, E. Jouanno, M. Wen, S. Mejri, R. Dirks, H. Jansen, C. Henkel, WJ. Chen, M. Zahm, C. Cabau, C. Klopp, A. W. Thompson, M. Robinson-Rechavi, I. Braasch, G. Lecointre, J. Bobe, J. H. Postlethwait, **C. Berthelot**#, H. Roest Crollius#, Y. Guiguen#. Genome structures resolve the early diversification of teleost fishes. *BioRxiv* (2022). *Submitted, under review.*

2. E. Parey, A. Louis, J. Montfort, Y. Guiguen, H. Roest Crollius#, **C. Berthelot**#, A high resolution comparative atlas across 74 fish genomes illuminates teleost evolution after whole genome duplication. *BioRxiv* (2022). *Submitted, under review.*

3. M. Muffato, A. Louis, N. Thi Thuy Nguyen, J. Lucas, **C. Berthelot**#, H. Roest Crollius#, Reconstruction of hundreds of reference ancestral genomes across the eukaryotic kingdom. *BioRxiv* (2022). *Submitted, under review.*

4. L. Moyon, **C. Berthelot**, A. Louis, N. T. T. Nguyen, H. Roest Crollius, Classification of non-coding variants with high pathogenic impact. *PLoS Genetics*. **18**, e1010191 (2022).

5. A. W. Thompson, M. B. Hawkins, E. Parey, D. J. Wcisel, T. Ota, K. Kawasaki, E. Funk, M. Losilla, O. E. Fitch, Q. Pan, R. Feron, A. Louis, J. Montfort, M. Milhes, B. L. Racicot, K. L. Childs, Q. Fontenot, A. Ferrara, S. R. David, A. R. McCune, A. Dornburg, J. A. Yoder, Y. Guiguen, H. Roest Crollius, **C. Berthelot**, M. P. Harris, I. Braasch, The bowfin genome illuminates the developmental evolution of ray-finned fishes. *Nature Genetics*, 1–12 (2021).

6. E. Parey, A. Louis, C. Cabau, Y. Guiguen, H. Roest Crollius#, **C. Berthelot**#, Synteny-Guided Resolution of Gene Trees Clarifies the Functional Impact of Whole-Genome Duplications. *Molecular Biology and Evolution*. **37**, 3324–3337 (2020).

7. **C. Berthelot**, J. Clarke, T. Desvignes, H. William Detrich, P. Flicek, L. S. Peck, M. Peters, J. H. Postlethwait, M. S. Clark, Adaptation of Proteins to the Cold in Antarctic Fish: A Role for Methionine? *Genome Biol Evol*. **11**, 220–231 (2019).

8. C. Sacerdot, A. Louis, C. Bon, **C. Berthelot**, H. R. Crollius, Chromosome evolution at the origin of the ancestral vertebrate genome. *Genome Biology*. **19**, 166 (2018).

9. **C. Berthelot***, D. Villar*, J. E. Horvath, D. T. Odom, P. Flicek, Complexity and conservation of regulatory landscapes underlie evolutionary resilience of mammalian gene expression. *Nature Ecology & Evolution*. **2**, 152–163 (2018).

10. J. Pasquier, I. Braasch, P. Batzel, C. Cabau, J. Montfort, T. Nguyen, E. Jouanno, C. **Berthelot, C**. Klopp, L. Journot, J. H. Postlethwait, Y. Guiguen, J. Bobe, Evolution of gene expression after whole-genome duplication: New insights from the spotted gar genome. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*. **328**, 709–721 (2017).

11. **C. Berthelot**, M. Muffato, J. Abecassis, H. Roest Crollius, The 3D Organization of Chromatin Explains Evolutionary Fragile Genomic Regions. *Cell Reports*. **10**, 1913–1924 (2015).

12. D. Villar*, **C. Berthelot***, S. Aldridge, T. F. Rayner, M. Lukk, M. Pignatelli, T. J. Park, R. Deaville, J. T. Erichsen, A. J. Jasinska, J. M. A. Turner, M. F. Bertelsen, E. P. Murchison, P. Flicek, D. T. Odom, Enhancer Evolution across 20 Mammalian Species. *Cell*. **160**, 554–566 (2015).

13. **C. Berthelot**, F. Brunet, D. Chalopin, A. Juanchich, M. Bernard, B. Noël, P. Bento, C. Da Silva, K. Labadie, A. Alberti, J.-M. Aury, A. Louis, P. Dehais, P. Bardou, J. Montfort, C. Klopp, C. Cabau, C. Gaspin, G. H. Thorgaard, M. Boussaha, E. Quillet, R. Guyomard, D. Galiana, J. Bobe, J.-N. Volff, C. Genêt, P. Wincker, O. Jaillon, H. R. Crollius, Y. Guiguen, The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. *Nat Commun*. **5** (2014), doi:10.1038/ncomms4657.

14. K. Howe, M. D. Clark, C. F. Torroja, J. Torrance, **C. Berthelot**, M. Muffato, J. E. Collins, S. Humphray, K. McLaren, L. Matthews, S. McLaren, I. Sealy, M. Caccamo, C. Churcher, C. Scott, J. C. Barrett, R. Koch, G.-J. Rauch, S. White, W. Chow, B. Kilian, L. T. Quintais, J. A. Guerra-Assunção, Y. Zhou, Y. Gu, J. Yen, J.-H. Vogel, T. Eyre, S. Redmond, R. Banerjee, J. Chi, B. Fu, E. Langley, S. F. Maguire, G. K. Laird, D. Lloyd, E. Kenyon, S. Donaldson, H. Sehra, J. Almeida-King, J. Loveland, S. Trevanion, M. Jones, M. Quail, D. Willey, A. Hunt, J. Burton, S. Sims, K. McLay, B. Plumb, J. Davis, C. Clee, K. Oliver, R. Clark, C. Riddle, D. Elliott, G. Threadgold, G. Harden, D. Ware, S. Begum, Beverley Mortimore, G. Kerry, P. Heath, B. Phillimore, A. Tracey, N. Corby, M. Dunn, C. Johnson, J. Wood, S. Clark, S. Pelan, G. Griffiths, M. Smith, R. Glithero, P. Howden, N. Barker, Christine Lloyd, C. Stevens, J. Harley, K. Holt, G. Panagiotidis, J. Lovell, H. Beasley, C. Henderson, D. Gordon, K. Auger, D. Wright, J. Collins, C. Raisen, L. Dyer, K. Leung, L. Robertson, K. Ambridge, D. Leongamornlert, S. McGuire, R. Gilderthorp, C. Griffiths, D. Manthravadi, S. Nichol, G. Barker, S. Whitehead, M. Kay, J. Brown, C. Murnane, E. Gray, M. Humphries, N. Sycamore, D. Barker, D. Saunders, J. Wallis, A. Babbage, S. Hammond, M. Mashreghi-Mohammadi, L. Barr, S. Martin, P. Wray, A. Ellington, N. Matthews, M. Ellwood, R. Woodmansey, G. Clark, J. D. Cooper, A. Tromans, D. Grafham, C. Skuce, R. Pandian, R. Andrews, E. Harrison, A. Kimberley, J. Garnett, N. Fosker, R. Hall, P. Garner, D. Kelly, C. Bird, S. Palmer, I. Gehring, A. Berger, C. M. Dooley, Z. Ersan-Ürün, C. Eser, H. Geiger, M. Geisler, L. Karotki, A. Kirn, J. Konantz, M. Konantz, M. Oberländer, S. Rudolph-Geiger, M. Teucke, C. Lanz, G. Raddatz, K. Osoegawa, B. Zhu, A. Rapp, S. Widaa, C. Langford, F. Yang, S. C. Schuster, N. P. Carter, J. Harrow, Z. Ning, J. Herrero, S. M. J. Searle, A. Enright, R. Geisler, R. H. A. Plasterk, C. Lee, M. Westerfield, P. J. de Jong, L. I. Zon, J. H. Postlethwait, C. Nüsslein-Volhard, T. J. P. Hubbard, H. R. Crollius, J. Rogers, D. L. Stemple, The zebrafish reference genome sequence and its relationship to the human genome. *Nature*. **496**, 498–503 (2013).

15. S. Champlot*, **C. Berthelot***, M. Pruvost, E. A. Bennett, T. Grange, E.-M. Geigl, An Efficient Multistrategy DNA Decontamination Procedure of PCR Reagents for Hypersensitive PCR Applications. *PLOS ONE*. **5**, e13042 (2010).

Book chapters

1. E. Parey, H. Roest Crollius, **C. Berthelot**, SCORPiOs, a novel method to reconstruct gene phylogenies in the context of the known WGD event. *Polyploidy: Methods and Protocols*, edited by Y. Van de Peer. Methods in Molecular Biology, Springer Nature (in press).

2. E. Parey, **C. Berthelot**, Les duplications complètes du génome, une source de redondance à l'échelle du génome entier. *Fonctions et évolution des séquences répétées dans les génomes*, edited by G.-F. Richard. ISTE Press (in press).

3. J. Bobe, L. Marandel, S. Panserat, P. Boudinot, **C. Berthelot**, E. Quillet, J.-N. Volff, C. Genêt, O. Jaillon, H. R. Crollius, Y. Guiguen, in *Genomics in Aquaculture*, S. MacKenzie, S. Jentoft, Eds. Academic Press, San Diego (2016) pp. 21–43.

Research projects

1. Evolution of genome organization

1.1. Introduction

During the course of my scientific career, I have devoted a significant part of my research to understanding how vertebrate genomes become reorganized over time, and eventually result in the karyotypes and gene arrangements that we observe today. Genome organization evolves through a combination of mutational processes – which modify chromosomes; and selective processes – which remove those modifications when they negatively impact genome function. Both processes are generally poorly characterized in eukaryotic genomes, and here I introduce a number of key concepts that run through my past research on this topic. These projects have focused on leveraging comparative genomics to extend our knowledge of functional genomic structures, and how evolution disrupts and rewires these structures through karyotypic reorganization.

1.1.1. Mutational processes of karyotype evolution

In eukaryotes, karyotypes are combinations of linear chromosomes which can change in number, content and organization during evolution (Sankoff, 2003). These reorganizations are of two types: some modify the order of DNA sequences in the genome, while some modify the DNA content. Sequences are re-ordered by chromosomal rearrangements, which include inversions and translocations, where the number of chromosomes is not modified; and chromosomal fusions and fissions, which reduce or increase the chromosome set. On the other hand, deletions and duplications reduce or increase the sequence content. In practice, chromosomal rearrangements observed in nature are often a more complex combination of these canonical classes (Kloosterman *et al.*, 2015; Kronenberg *et al.*, 2018; Jiao and Schneeberger, 2020). Additionally, a particular case on which I have worked extensively are whole-genome duplications, where the entire set of chromosomes is duplicated, resulting in polyploid individuals.

How these rearrangements occur in cells remains an area of debate, and several molecular mechanisms have been implicated, including non-homologous recombination (Lupski and Stankiewicz, 2005), DNA breakage repair (Soutoglou *et al.*, 2007), polymerase slippage during replication (Lee, Carvalho and Lupski, 2007), and generally meiotic errors. What is however known is that rearrangements are frequent, with many chromosomal variants existing as polymorphisms in populations, including in humans (Sudmant *et al.*, 2015). In some species, especially in plants and yeasts, individuals with different levels of ploidy can coexist within the population (Soltis, Visger and Soltis, 2014). Moreover, karyotypes vary substantially even between closely related species, which can have very different chromosome numbers or significant chromosomal reorganization (McClintock, 1984; Thybert *et al.*, 2018; Yin *et al.*, 2021). The dynamics through which rearrangements occur in eukaryotic evolution are not well understood, although the rearrangement rate is known to vary substantially between clades (Coghlan *et al.*, 2005).

An outstanding question in the field has been whether those rearrangements are largely neutral, and are distributed in genomes according to mechanical susceptibility to breakage; or whether they are frequently selected against – or for, if they are adaptive.

1.1.2. Selective pressures on karyotype evolution

Chromosomal rearrangements can result in fitness differences between individuals and therefore be subjected to selection: this is prominently the case for deletions and translocations, which can result in incomplete sets of chromosomes. Deletions are indeed a major cause of genetic diseases in humans (Shapira, 1998). However, the selective impact of other types of rearrangements such as inversions is more subtle, especially when they simply reorganize the genome without loss of sequence. Because genes depend on non-coding sequences in their local environment for their correct expression, these rearrangements can potentially modify gene function and therefore be selected negatively or positively (Coghlan *et al.*, 2005; Spielmann, Lupiáñez and Mundlos, 2018). This phenomenon is well documented in development, where several important developmental genes have remained in linkage all through vertebrate evolution, as rearrangements that separate them from regulatory environment are removed by selection (Engström *et al.*, 2007; Kikuta *et al.*, 2007). However, the fraction of the genome that evolves under such regulatory constraints is debated. Other selective pressures apply on rearranged chromosomes – especially recombination suppression – but they mostly affect populational dynamics, rather than long-term evolution, and I will not discuss them here.

Additionally, duplications pose another challenge to selection, as they can modify the molecular equilibrium in the cell. Genes that are involved in protein complexes are typically expressed at levels that preserve the stoichiometric balance between all components in the complex. This equilibrium, termed gene dosage balance, can be disrupted when some but not all genes are duplicated, and therefore expressed at higher levels (Veitia, 2004). Whole-genome duplications are thought to be tolerated in evolution in part because they preserve this gene dosage balance, while providing new genetic material that can be diverted to new functions and help with adaptation (Makino and McLysaght, 2010; Veitia and Birchler, 2021).

1.2. Dynamics of chromosomal rearrangements in mammals

As a PhD student under the supervision of Hugues Roest Crollius at the Institut de Biologie de l'Ecole normale supérieure in Paris, my first project investigated the distribution of chromosomal rearrangements in mammalian genomes. Mammalian chromosomes are largely collinear between species (e.g. syntenic), but are punctually rearranged by chromosomal inversions, translocations, fusions and fissions (Ferguson-Smith and Trifonov, 2007). Extensive evidence supports that chromosomal rearrangement breakpoints are not randomly distributed in mammalian genomes (Pevzner and Tesler, 2003; Sankoff and Trinh, 2005; Lemaitre *et al.*, 2009; Von Grotthuss, Ashburner and Ranz, 2010), but the origin of this distribution has been hotly debated. Two main scenarios have been proposed to explain this observation: some regions may be more prone to breakage, or natural selection may have purged breaks where genomic organization is functionally important (Peng, Pevzner and Tesler, 2006; Becker and Lenhard, 2007).

Using comparative genomics tools developed in the lab, I analyzed the distribution of over 800 chromosomal breakpoints that occurred during genome evolution from an ancestral mammal to five of its extant descendant species (human, mouse, dog, cow and horse). Our results revealed that the local frequency of breakpoints in mammalian genomes correlates strongly with the local proportion of open chromatin, suggesting that active chromatin regions may be more vulnerable to breakage or erroneous repair, leading to chromosomal reorganization. Further, I showed that the rearrangement distribution observed in mammalian genomes can be closely reproduced by simulations of genome evolution, where open chromatin regions physically close in the nucleus are the major promoter of chromosomal rearrangements (**Figure 1**). I also tested alternative scenarios – for example, whether non-homologous recombination of transposable elements in physical contact in the nucleus would result in a similar pattern – and found that simulations based on open chromatin regions were unique in reproducing the observations from real data.

During this work, I also investigated whether gene regulation enforces selective constraints on chromosomal rearrangements. Indeed, I found that the local proportion of conserved non-coding sequence correlates negatively with rearrangement breakpoints, suggesting that rearrangements occurring in genomic regions densely populated by regulatory elements are subjected to negative selection. However, this effect was small and only accounted for 3% of the variance in breakpoint frequency. Together, these results allowed us to conclude that the distribution of rearrangements in mammalian genomes is largely driven by a mechanical propensity to breakage or misrepair in open chromatin regions, with a small effect of selection against reorganization of local gene regulation, possibly restricted to specific genomic regions. While this work received modest attention at the time of publication, it found regained interest in recent years, as functional investigation of DNA breakage in cellular models and cancer have independently confirmed that active genomic regions in physical proximity in the nucleus display higher rates of rearrangements (Zhang *et al.*, 2012).

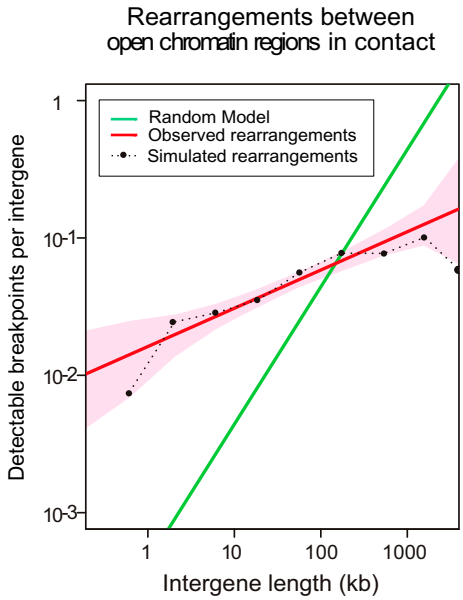

Figure 1. *Evolutionary simulations of chromosomal rearrangements in the human genome. Pairs of non-coding open chromatin regions were randomly selected according to their probability of contacting each other in 3D in the nucleus to simulate rearrangement breakpoints. The distribution of simulated breakpoints between genes (dotted) resembles that observed in real data (red), and is strikingly different from an uniform distribution (green) due to the non-linear distribution of open chromatin.*

Related publication

C. Berthelot, M. Muffato, J. Abecassis, H. Roest Crollius, The 3D Organization of Chromatin Explains Evolutionary Fragile Genomic Regions. *Cell Reports*. **10**, 1913–1924 (2015). *IF: 9.4*

1.3. Genome evolution after whole-genome duplication

1.3.1. Ancestral polyploidy in the zebrafish and rainbow trout genomes

Building on my experience with genome organization evolution, I was involved during my PhD in the sequencing and analysis of the reference genomes for two model species: zebrafish (*Danio rerio*) and rainbow trout (*Oncorhynchus mykiss*). Zebrafish is a major model organism for developmental biology, evolutionary biology and neuroscience (Lieschke and Currie, 2007), while rainbow trout is widely studied for agronomical, ecological and population genetics purposes (Thorgaard *et al.*, 2002). Both species are descended from an ancient whole genome duplication which occurred in the ancestor of teleost fish, around 320 million years ago (Jaillon *et al.*, 2004). Additionally, the rainbow trout genome underwent another round of duplication at the origin of Salmonids (~80 million years ago; Macqueen and Johnston, 2014). For both organisms, my main task was to identify and characterize paralogous genes and chromosomal regions that arose from the genome-wide duplication. Because whole-genome duplications produce new evolvable copies for every gene, they are thought to contribute substantially to adaptation, isolation and phenotypic robustness. Yet, the structural and functional consequences of whole-genome duplications are not well understood. Whole-genome duplications are overall rare in animals, unlike plants (Van de Peer, Mizrachi and Marchal, 2017); paleopolyploid fish represent a key taxon to study the evolutionary phenomenon which gave rise, amongst others, to many of the multigenic gene families present in the human genome today (Dehal and Boore, 2005).

The zebrafish genome project was coordinated by Kerstin Howe and Derek Stemple (Wellcome Trust Sanger Institute, UK), and the analysis was performed in collaboration with the Ensembl team at EMBL-EBI (UK), who produced the genome annotation and comparative resources. In this genome, the whole-genome duplication event is ancient, and the ancestral polyploid structure has been in part obscured by chromosomal rearrangements, small-scale duplications and massive loss of supernumerary genes since the duplication event. We showed that most zebrafish genes have orthologs in human, and that a substantial fraction of zebrafish genes (26%) are still present in two copies inherited from the polyploidization. We also identified anciently paralogous genomic regions in the zebrafish genome, providing the first genome-wide cartography of the whole-genome duplication event in this model organism.

In rainbow trout, where the polyploidization event is more recent, the recently duplicated chromosomal structure is still very apparent (**Figure 2**). We showed that despite 80 million years of evolution, the majority of duplicated genes remain present in two copies in the rainbow trout genome. Indeed, some duplicated genomic regions retain such a high degree of sequence similarity due to local maintenance of meiotic recombination (Allendorf *et al.*, 2015; Lien *et al.*, 2016), that they could not be differentiated in this original genome assembly and resulted in collapsed genomic regions. Using phylogenetic analysis of gene histories and conserved synteny, we reconstructed the evolution of the rainbow trout karyotype through both the teleost and the salmonid whole genome duplications (**Figure 2**). Further, I confirmed that gene families related to transcriptional regulation and development have been

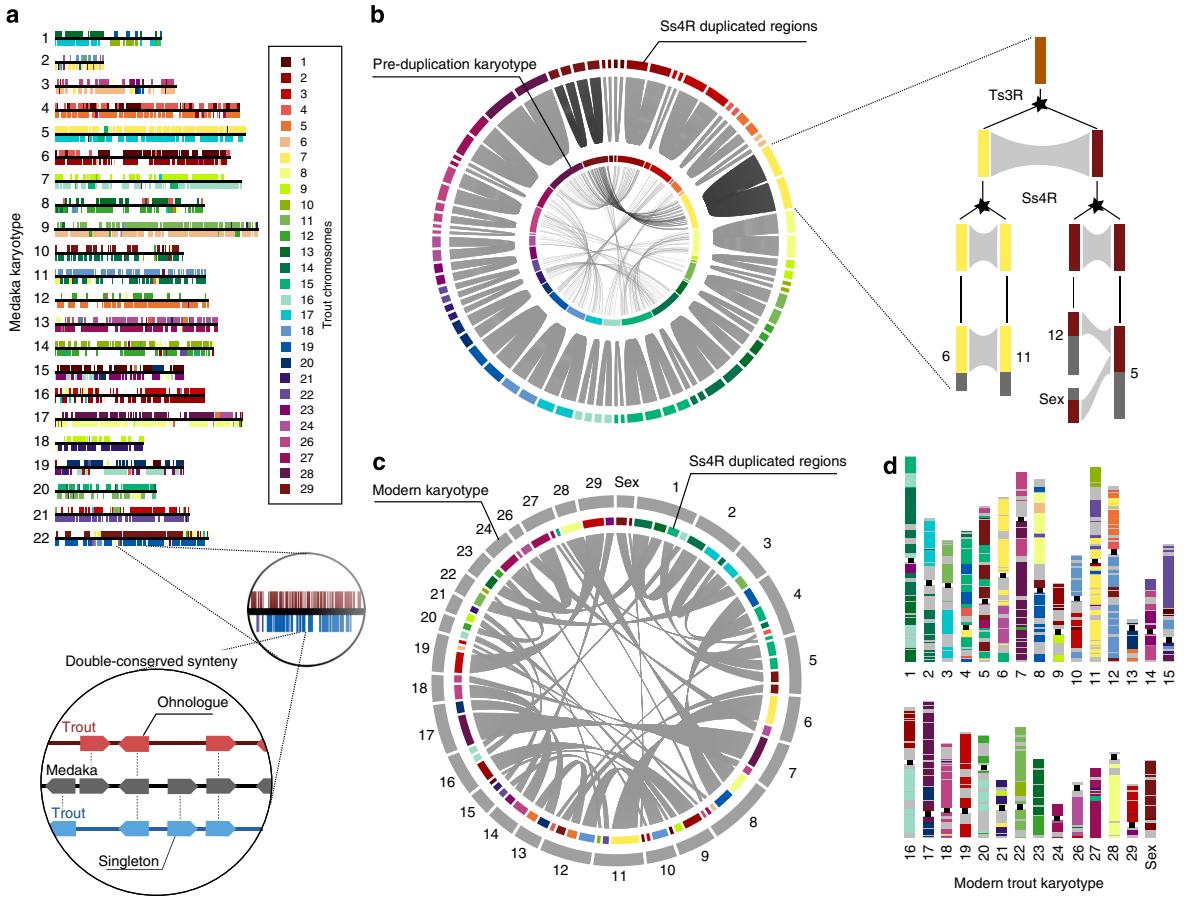

Figure 2. *Evolutionary history of the rainbow trout genome.* ***a.*** *Double-conserved synteny between the trout and medaka chromosomes : because rainbow trout has undergone a whole genome duplication event, each medaka chromosome has two orthologous chromosomes in rainbow trout.* ***b.*** *Imbrication of the teleost (Ts3R) and salmonid (Ss4R) whole-genome duplications in the structure of the rainbow trout genome. Duplicated regions in the rainbow trout genome (outer circle) could be grouped into 31 ancestral chromosomes before the Ss4R duplication event (inner circle). This ancestral karyotype was itself a paleopolyploid from the Ts3R, and the ancestral chromosomes can be grouped in duplicated pairs, represented by grey links.* ***c-d.*** *Distribution of duplicated genomic regions on the rainbow trout karyotype, joined by grey links on the circular representation; colours as in (b).*

preferentially retained over successive rounds of whole-genome duplication, suggesting that these genes may provide raw matter for evolutionary innovations.

Related publications

K. Howe, M. D. Clark, C. F. Torroja, J. Torrance, **C. Berthelot**, M. Muffato, J. E. Collins, S. Humphray, K. McLaren, L. Matthews, S. McLaren, I. Sealy, M. Caccamo, C. Churcher, C. Scott, J. C. Barrett, R. Koch, G.-J. Rauch, S. White, W. Chow, B. Kilian, L. T. Quintais, J. A. Guerra-Assunção, Y. Zhou, Y. Gu, J. Yen, J.-H. Vogel, T. Eyre, S. Redmond, R. Banerjee, J. Chi, B. Fu, E. Langley, S. F. Maguire, G. K. Laird, D. Lloyd, E. Kenyon, S. Donaldson, H. Sehra, J. Almeida-King, J. Loveland, S. Trevanion, M. Jones, M. Quail, D. Willey, A. Hunt, J. Burton, S. Sims, K. McLay, B. Plumb, J. Davis, C. Clee, K. Oliver, R. Clark, C. Riddle, D. Elliott, G. Threadgold, G. Harden, D. Ware, S. Begum, Beverley Mortimore, G. Kerry, P. Heath, B. Phillimore, A. Tracey, N. Corby, M. Dunn, C. Johnson, J. Wood, S. Clark, S. Pelan, G. Griffiths, M. Smith, R. Glithero, P. Howden, N. Barker, Christine Lloyd, C. Stevens, J. Harley, K. Holt, G. Panagiotidis, J. Lovell, H. Beasley, C. Henderson, D. Gordon, K. Auger, D. Wright, J. Collins, C. Raisen, L. Dyer, K. Leung, L. Robertson, K. Ambridge, D. Leongamornlert, S. McGuire, R. Gilderthorp, C. Griffiths, D. Manthravadi, S. Nichol, G. Barker, S. Whitehead, M. Kay, J. Brown, C. Murnane, E. Gray, M. Humphries, N. Sycamore, D. Barker, D. Saunders, J. Wallis, A. Babbage, S. Hammond, M. Mashreghi-Mohammadi, L. Barr, S. Martin, P. Wray, A. Ellington, N. Matthews, M. Ellwood, R. Woodmansey, G. Clark, J. D. Cooper, A. Tromans, D. Grafham, C. Skuce, R. Pandian, R. Andrews, E. Harrison, A. Kimberley, J. Garnett, N. Fosker, R. Hall, P. Garner, D. Kelly, C. Bird, S. Palmer, I. Gehring, A. Berger, C. M. Dooley, Z. Ersan-Ürün, C. Eser, H. Geiger, M. Geisler, L. Karotki, A. Kirn, J. Konantz, M. Konantz, M. Oberländer, S. Rudolph-Geiger, M. Teucke, C. Lanz, G. Raddatz, K. Osoegawa, B. Zhu, A. Rapp, S. Widaa, C. Langford, F. Yang, S. C. Schuster, N. P. Carter, J. Harrow, Z. Ning, J. Herrero, S. M. J. Searle, A. Enright, R. Geisler, R. H. A. Plasterk, C. Lee, M. Westerfield, P. J. de Jong, L. I. Zon, J. H. Postlethwait, C. Nüsslein-Volhard, T. J. P. Hubbard, H. R. Crollius, J. Rogers, D. L. Stemple, The zebrafish reference genome sequence and its relationship to the human genome. *Nature*. **496**, 498–503 (2013). *IF: 38.3*

C. Berthelot, F. Brunet, D. Chalopin, A. Juanchich, M. Bernard, B. Noël, P. Bento, C. Da Silva, K. Labadie, A. Alberti, J.-M. Aury, A. Louis, P. Dehais, P. Bardou, J. Montfort, C. Klopp, C. Cabau, C. Gaspin, G. H. Thorgaard, M. Boussaha, E. Quillet, R. Guyomard, D. Galiana, J. Bobe, J.-N. Volff, C. Genêt, P. Wincker, O. Jaillon, H. R. Crollius, Y. Guiguen, The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. *Nat Commun*. **5** (2014), doi:10.1038/ncomms4657. *IF: 14.9*

1.3.2. Comparative genomics in paleopolyploid fishes

From 2017 to 2021, I co-supervised Elise Parey's PhD work with Hugues Roest Crollius as part of an international collaborative project to further explore the functional consequences of whole-genome duplications in teleost fishes. Teleosts represent about half of all vertebrate species and are remarkably diverse in terms of phenotypic and environmental niche adaptations, which makes them an outstanding model group for evolutionary, ecological and functional genomics (Nelson, Grande and Wilson, 2016). Yet, despite a growing number of sequence reference genomes, large-scale comparative analysis remains challenging in teleosts due to the specifics of their genomic organization. As legacy of their common whole genome duplication 320 million years ago, a large fraction of teleost genomes remain in duplicate paralogous copies (Jaillon *et al.*, 2004; Howe *et al.*, 2013). This ancestral polyploidy confounds the detailed identification of orthologous genomic regions across teleost species, and therefore of their specific evolutionary dynamics, which we set out to investigate in more detail. The aim of the consortium was, broadly speaking, to compare over 75 fish reference genomes with multiple instances of genome duplications in the phylogeny and provide insight into the mechanisms of karyotype, gene and sex

evolution after whole-genome duplication events. This work was funded by ANR, the French National Research Agency, and involved collaborators in France, Switzerland and the USA – amongst which Yann Guiguen (INRAe Rennes), Ingo Braasch (U. Michigan) and John Posthlewait (U. Oregon).

Identifying paralogy in anciently duplicated genomes

A key difficulty identified early in the project was to reliably identify orthologous genes (descended from the same gene copy) and paralogous genes (descended from different duplicates) across multiple species in a clade with such complex genomic histories. Indeed, genome duplications result in non-parsimonious gene phylogenies which are difficult to reconstruct accurately based on gene sequence information only (Zwaenepoel and Van de Peer, 2019). To address this problem, we reasoned that the residual structure of duplicated chromosomes could be leveraged as a complement to sequence evolution to infer gene histories. In paleopolyploid genomes, genes are embedded in syntenic duplicated regions, meaning that phylogenetic signal from entire gene neighborhoods can be utilized to resolve the history of individual genes when sequence alone is inconclusive (Byrne and Wolfe, 2005; Catchen, Conery and Postlethwait, 2009). We developed a method named SCORPiOs (Synteny-guided CORrection of Paralogies and Orthologies), which constructs orthology groups between multiple species based on gene neighborhoods instead of sequence, and uses this syntenic information to guide phylogenetic gene tree inference (**Figure 3**). We showed that our model substantially improves gene phylogenies in the presence of whole-genome duplication events, and reveals how gene duplicates likely contributed to evolutionary innovations in the vascular system, pigmentation repertoire and retina in fish species.

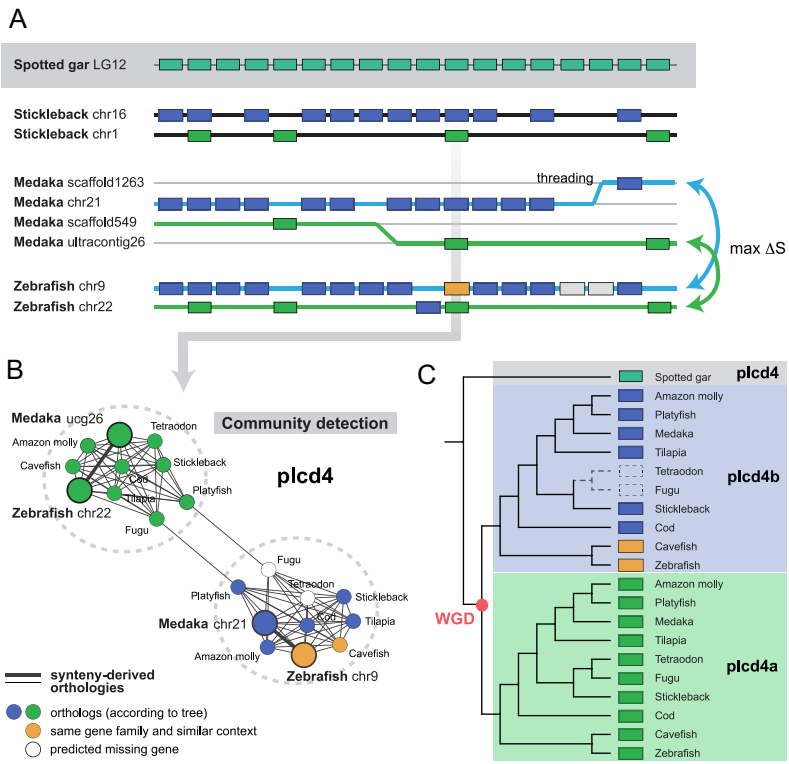

Figure 3. *Overview of SCORPiOs.* ***A.*** *The method uses an unduplicated outgroup genome (here, spotted gar) as a proxy for the local ancestral gene order. SCORPiOs reconstructs the most likely orthologous duplicated regions (max ΔS) between pairs of paleotetraploid genomes, based on shared orthologies and gene retention patterns.* ***B.*** *For each gene family, SCORPiOs constructs a graph where nodes represent genes, and edges join genes that belong to orthologous duplicated regions based on (A). This graph is then separated using a community detection algorithm to obtain orthogroups based on syntenic information.* ***C.*** *The syntenic orthogroups in (B) are implemented as a constraint when inferring the phylogenetic tree, resulting in a synteny- and sequence-consistent gene history.*

Related publication

E. Parey, A. Louis, C. Cabau, Y. Guiguen, H. Roest Crollius[#], **C. Berthelot**[#], Synteny-Guided Resolution of Gene Trees Clarifies the Functional Impact of Whole-Genome Duplications. *Molecular Biology and Evolution*. **37**, 3324–3337 (2020). *IF: 16.2*

Software

SCORPiOs: a synteny-guided gene tree correction pipeline for clades that have undergone a whole-genome duplication event. https://github.com/DyogenIBENS/SCORPIOS

Book chapters

E. Parey, H. Roest Crollius, **C. Berthelot**, SCORPiOs, a novel method to reconstruct gene phylogenies in the context of the known WGD event. *Polyploidy: Methods and Protocols*, edited by Y. Van de Peer. Methods in Molecular Biology, Springer Nature (in press).

E. Parey, **C. Berthelot**, Les duplications complètes du génome, une source de redondance à l'échelle du génome entier. *Fonctions et évolution des sequences répétées dans les génomes*, edited by G.-F. Richard. ISTE Press (in press).

Following up on this work, we set out to reconstruct the evolutionary history of teleost karyotypes through and after the whole genome duplication event (WGD). To this end, we built on a published reconstruction of the ancestral karyotype of teleosts (Nakatani and McLysaght, 2017). This karyotype had been estimated to contain 13 chromosomes, and the broad genomic locations descended from each of these 13 chromosomes had previously been delineated across the genomes of four widely-studied reference teleost species (zebrafish, medaka, stickleback and tetraodon). However, this ancestral resource corresponds to the genomic organization *before* the tetraploidization event: the distribution of each pair of duplicated ancestral chromosomes in extant teleost genomes was not described, hindering the identification of orthologous and paralogous genomic regions across species. To address this issue, we combined the gene phylogeny methodology developed in SCORPiOs with the available pre-WGD karyotype reconstruction and established the first high resolution comparative atlas of paleopolyploid regions across 74 teleost fish genomes (**Figure 4**). We used the synteny-corrected inferences of paralogs and orthologs across all 74 species produced by SCORPiOs, mapped them to the ancestral chromosomes identified in Nakatani and McLysaght, and leveraged the homology relationships across species to identify (i) paralogous chromosomal regions within genomes and (ii) the orthology relationships of these regions between species.

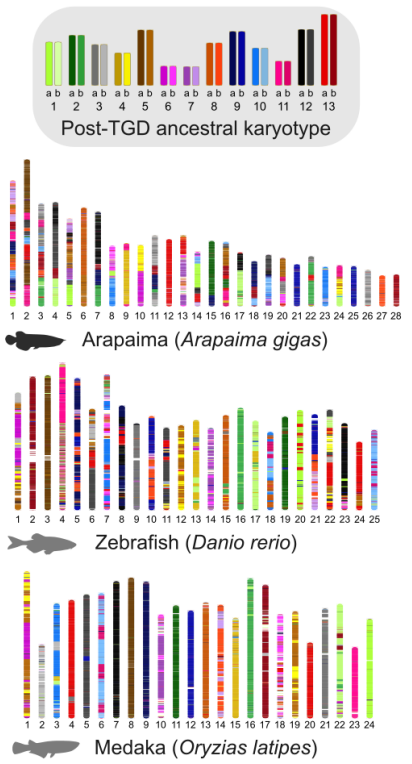

Figure 4. *The ancestral teleost genome after duplication, and the inferred location of genomic regions descended from each ancestral chromosome in the genomes of three representative teleost fishes: the arapaima (belonging to the Osteoglossiformes order), zebrafish (Cypriniformes order), and medaka (Beloniformes order).*

Genome-wide mechanisms of rediploidization in teleost fishes

This comparative homology atlas across teleosts represented a fantastic resource to study the tetraploidization and rediploidization mechanisms that affected the ancestor of teleosts. Indeed, the mechanisms through which the ancestral teleost initially became polyploid has remained controversial. Broadly, polyploidization can occur either (i) through the doubling of a single parental genome (autopolyploidy), typically due to errors during meiosis; or (ii) through the hybridization of two different parental genomes, followed by doubling (allopolyploidy; Stebbins, 1947; Mason and Wendel, 2020). The mode of polyploidization has important consequences on the structural and functional evolution of descendant genomes. Indeed, polyploids undergo a complex process named rediploidization and essentially return to a diploid organization through a combination of chromosomal rearrangements, local deletions, and the pseudogenization of a

substantial fraction of their redundant, duplicated genes (Kellis, Birren and Lander, 2004; Garsmeur *et al.*, 2014). As autopolyploids contain two highly related subgenomes, they typically experience a balanced rediploidization across duplicated chromosomes and can maintain prolonged recombination between duplicated chromosomes, which behave as tetrads during meiosis (tetrasomic inheritance; Allendorf *et al.*, 2015; Robertson *et al.*, 2017). Allopolyploids, however, can contain substantially differentiated initial subgenomes, depending on the time of divergence between their parental species before hybridization. This initial dissimilarity frequently results in “subgenome dominance”, where one subgenome is over-expressed, under stronger selection, and retains a larger fraction of genes than the other during the rediploidization process (Garsmeur *et al.*, 2014; Session *et al.*, 2016; Cheng *et al.*, 2018). Therefore, understanding the mechanisms through which paleopolyploid genomes became duplicated is fundamental to investigate how polyploidization can lead to species diversification, subpartition of gene functions, phenotypic evolution and adaptative innovation, as they condition the evolvability of the subgenomes.

Previous to our work, conflicting evidence had been put forward regarding the initial duplication mechanisms of the ancestral teleost genome. Phylogenetic analysis of homology relationships between the *Hox* gene clusters across fish species had suggested that at least some genomic regions still maintained recombination by the time the Osteoglossiform and Clupeocephala clades diverged 267 million years ago, so about 60 million years after the WGD event (Martin and Holland, 2014). This behavior was suggestive of an autotetraploid origin, although some allotetraploids from closely related parents can exhibit localized meiotic recombination between subgenomes (Li *et al.*, 2021). Conversely, an analysis of 5,589 gene loci in the genomes of 8 teleosts had revealed an unbalanced retention of WGD paralogs on duplicated chromosomes, suggesting an allotetraploid origin (Conant, 2020). To address this open question, we performed the first genome-wide analysis of rediploidization patterns

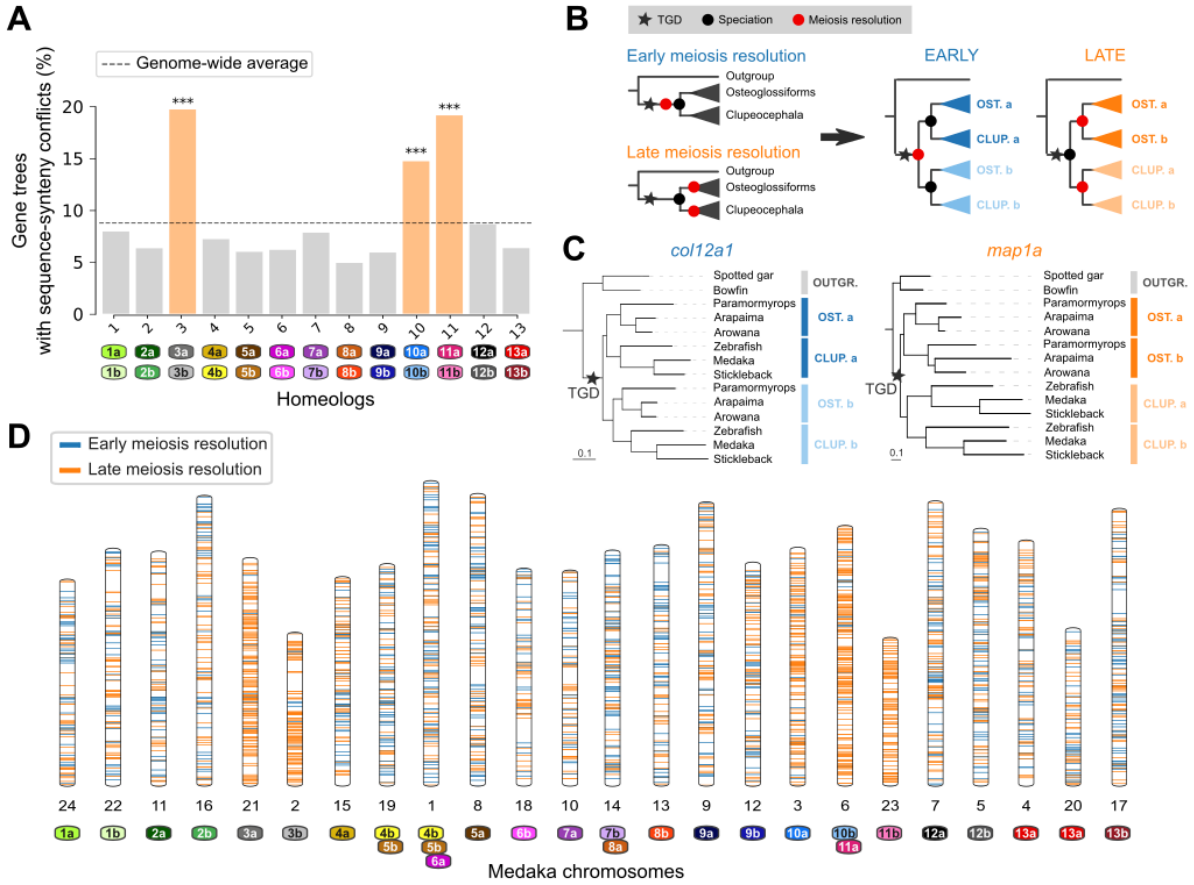

Figure 5. *Evidence for prolonged meiotic recombination in the ancestral tetraploid teleost genome.* ***A.*** *Genes on ancestral chromosome pairs 3, 10 and 11 exhibit a high rate of sequence vs. synteny conflicts in the placement of their duplication nodes.* ***B.*** *Late meiosis resolution after speciation (left) results in specific gene tree topologies (right), which group paralogs by lineage instead of by canonical orthology.* ***C.*** *Examples of genes with early (col12a1) and late (map1a) meiosis resolution.* ***D.*** *Distribution of regions descended from delayed meiosis resolution in the medaka genome.*

from the teleost whole-genome duplication. We identified three pairs of duplicate ancestral chromosomes with strong evidence of delayed rediploidization and prolonged meiotic recombination, which perdured after the Osteoglossiform/Clupeocephala divergence (**Figure 5**). This phenomenon was detected based on gene homology relationships: when duplicated chromosomes recombine during meiosis, paralogous genes behave more like alleles and keep exchanging sequences. Paralog sequences only start diverging once meiotic recombination has ceased, frequently after speciation and in a lineage-specific manner (Robertson *et al.*, 2017; Gundappa *et al.*, 2021). This characteristic pattern is detectable from sequence vs. synteny conflicts: synteny suggests that the duplication is ancestral, as both paralogs are embedded in large duplicated chromosomes dating back from the WGD, but sequence divergence suggests that the duplications are lineage-specific and more recent. We designed an extension to SCORPiOs that searches for these discordant tree topologies and explicitly tests whether they support delayed meiosis resolution (SCORPiOs LOReLEi). This work provided the first evidence that meiotic recombination was maintained for at least 60 million years across entire duplicated chromosomes in the teleost ancestor, suggesting that the initial subgenomes were highly similar and providing support for an autotetraploid origin to teleosts.

Autopolyploidy is classically associated with a balanced rediploidization of the subgenomes, where a substantial fraction of paralog genes return to a single-copy state by loss of one copy, but without favoring one duplicated chromosome over the other (Garsmeur *et al.*, 2014). In teleost fish, however, we found evidence that rediploidization has been unequal, with at least 5 pairs of ancestral chromosomes out of 13 displaying a bias for gene retention on one chromosome. Of these 5 pairs of chromosomes, four also displayed marked differences in selective pressure on gene sequences, as measured by the ratio of non-synonymous to synonymous substitutions in genes (dN/dS; overlap was significant, $p = 0.03$, Fisher's exact test). These differences were however not always consistent with the direction of the retention bias – some chromosomes exhibited higher gene retention but lower selection while the contrary was true in other chromosome pairs. They were also not exclusive of an history of delayed meiosis resolution, as at least two chromosome pairs experienced both delayed and biased rediploidization, suggesting that initial sequence similarity is ultimately irrelevant to biases in gene retention. While these results could not explain why a probable autotetraploid experienced such biased rediploidization, our findings are important because they suggest that biased gene retention is not the hallmark of allopolyploids, and is insufficient to conclude about the mode of polyploidization, which had been the general consensus so far.

Related publication

E. Parey, A. Louis, J. Montfort, Y. Guiguen, H. Roest Crollius[#], **C. Berthelot**[#]. A high-resolution comparative atlas across 74 fish genomes illuminates teleost evolution after whole-genome duplication. *bioRxiv* (2022). *Submitted, under review.*

1.3.3. Phylogenomics using genome structures

In parallel to our investigations of the karyotypic and functional evolution of paleopolyploid fishes, Elise Parey, Hugues Roest Crollius and I also contributed to investigations of the phylogeny of fish species as part of the same international consortium. Ray-finned fishes represent a monophyletic group including teleosts (96% of species), amiiformes (bowfins) and lepisosteiformes (gars) and diverged from other vertebrates ~385 million years ago (Near *et al.*, 2012). The topologies of these old speciations has been a historical area of controversy between paleontologists, systematists and molecular evolutionists, with disagreements on the sister group to teleosts and the early taxonomic relationships within teleosts (reviewed in Dornburg and Near, 2021). These relationships are important to interpret the early evolutionary history of fishes and their fossil record, but also to ground analyses of the teleost whole-genome duplication with an appropriate species phylogeny and outgroups.

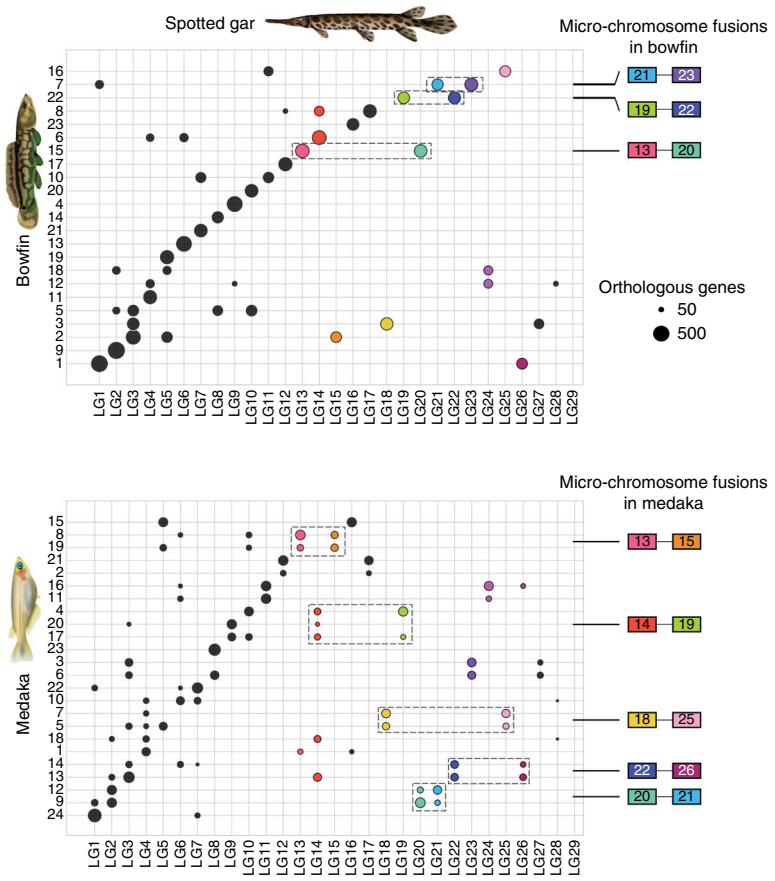

Figure 6. *Ancestral vertebrate micro-chromosomes (highlighted in color) are conserved in spotted gar, but fused into larger chromosomes in bowfin and medaka. However, the fusions are independent and correspond to different arrangements of the ancestral micro-chromosomes.*

While systematics are not at the core of my scientific interests, we became intrigued by the idea that comparisons of genome organizations could potentially be useful to resolve long-standing phylogenetic questions, where phylogenomic analyses based on sequence alignments have given ambiguous results. This is especially relevant for old taxonomic relationships, as genomic rearrangements are more rare than nucleotide substitutions and therefore less affected by saturation, reversions and convergence at those time scales. As chromosome-scale reference genome assemblies are becoming the norm, allowing whole-genome comparisons of genome structure, the question seemed particularly timely.

We first investigated this line of thought while collaborating with Ingo Braasch (U. Michigan) on the reference genome of the bowfin (*Amia calva*). The bowfin is the only surviving species of a once speciose clade according to the fossil record, the Halecomorphi, and the position of this clade relative to the teleosts and the gars was debated. Molecular phylogenies consensually place the bowfin as the sister group of gars (Near *et al.*, 2012; Braasch *et al.*, 2016; Bi *et al.*, 2021), but bowfin and teleosts share a number of derived morphological and karyotypic characters (Sallan, 2014). Specifically, neither of them have micro-chromosomes, unlike the gar, chicken, and likely the vertebrate ancestor. To resolve this question, we investigated the chromosomal structure of the bowfin and compared it to that of spotted gar and medaka, a representative teleost fish. We showed that while both bowfin and teleosts experienced fusions of micro-chromosomes into larger chromosomes, these fusions were entirely independent and did not constitute a shared derived character (**Figure 6**). Additionally, we implemented a phylogenomics strategy to infer a species phylogeny from comparisons of genome organization, by calculating the fraction of shared gene adjacencies between pairs of genomes and building a distance-based tree. This analysis consolidated the molecular phylogenetic evidence based on sequence alignments, and further cemented that bowfins are the sister group of gars, together forming the Holostean clade.

Our collaborators in the GenoFish project then turned their attention to the early diversification of teleost fishes, which was the subject of much deeper controversy (Dornburg and Near, 2021). The three oldest clades in the teleost phylogeny are the Elopomorphs (eels, tarpons), the Osteoglossiforms (bonytongues, arowanas) and the Clupeocephala (all other teleosts). Very few shared derived characters have been identified to ascertain the relationships between these taxonomic groups, and different molecular phylogenetic studies had been unable to reach a consensus (Takezaki, 2021). To tackle this question, the GenoFish consortium sequenced high-quality, chromosome-scale reference genomes for seven elopomorph species, as the genomic resources for this clade in particular were poor and fragmented. Using this novel data in combination with reference genomes from 18 other teleosts

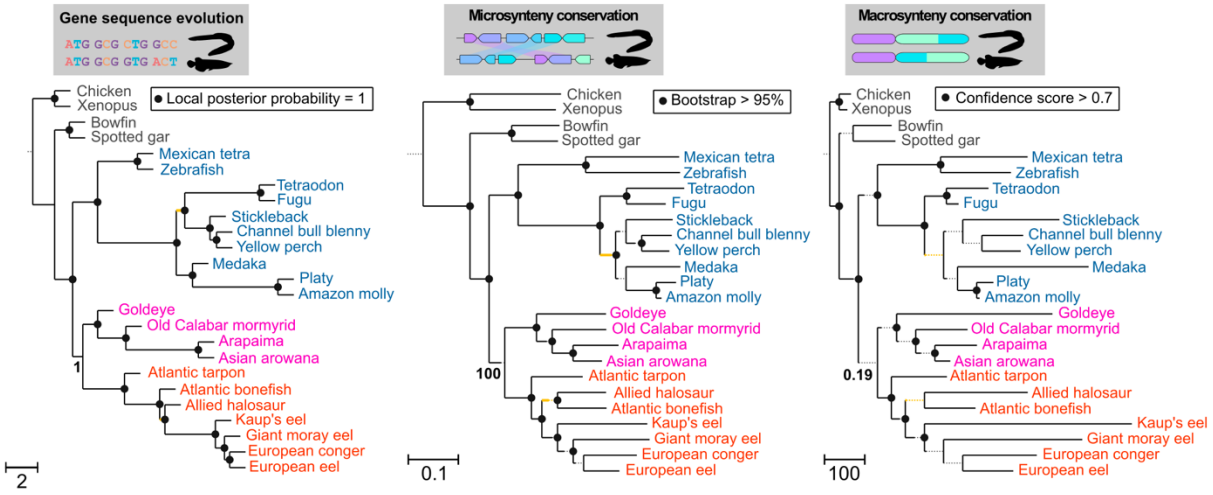

Figure 7. *Phylogenetic trees of teleost species based on sequence evolution, microsynteny conservation and macrosyntenic similarity. Elopomorph species are in orange; Osteoglossiforms in pink; Clupeocephala in blue; outgroups on grey. All three methodologies place Elopomorphs and Osteoglossiforms as a monophyletic group, sister to Clupeocephala.*

and vertebrate outgroups, we explored the taxonomic relationships of these three ancient teleost clades with phylogenomics evidence collected at three levels of granularity: (i) a consensus sequence tree, based on alignments from 955 strict 1-1 orthologs; (ii) a microsynteny conservation tree, based on conserved gene adjacencies as previously done for the bowfin genome; and (iii) a macrosynteny similarity tree, based on shared chromosomal rearrangements between species. All three analyses converged into a single evolutionary scenario placing Elopomorphs and Osteoglossiforms as a monophyletic group, sister to the Clupeocephala (**Figure 7**). This was a surprising finding, as this hypothesis had been evoked but never formally retained because of the lack of morphological characters supporting this phylogeny based on modern species and in the fossil record. However, we discovered a shared, derived cytogenetic character in support of the clade, as Elopomorphs and Osteoglossiforms share a fusion of two ancestral chromosomes, which is unlikely to have occurred twice independently. Together, these results provided strong evidence for the monophyly of Elopomorphs and Osteoglossiforms, a clade that we named Eloposteoglossocephala, and resolved 50 years of debate about the early diversification of teleost fishes.

Related publications

A.W. Thompson, M.B. Hawkins, E. Parey, D.J. Wcisel, T. Ota, K. Kawasaki, E. Funk, M. Losilla, O.E. Fitch, Q. Pan, R. Feron, A. Louis, J. Montfort, M. Milhes, B.L. Racicot, K.L. Childs, Q. Fontenot, A. Ferrara, S.R. David, A.R. McCune, A. Dornburg, J.A. Yoder, Y. Guiguen, H. Roest Crollius, **C. Berthelot**, M.P. Harris, I. Braasch. The bowfin genome illuminates the developmental evolution of ray-finned fishes. *Nature Genetics* 1–12 (2021). *IF: 38.3*

E. Parey, A. Louis, J. Montfort, O. Bouchez, C. Roques, C. Iampietro, J. Lluch, A. Castinel, C. Donnadieu, T. Desvignes, C. Floi Bucao, E. Jouanno, M. Wen, S. Mejri, R. Dirks, H. Jansen, C. Henkel, WJ. Chen, M. Zahm, C. Cabau, C. Klopp, A. W. Thompson, M. Robinson-Rechavi, I. Braasch, G. Lecointre, J. Bobe, J. H. Postlethwait, **C. Berthelot**[#], H. Roest Crollius[#], Y. Guiguen[#]. Genome structures resolve the early diversification of teleost fishes. *BioRxiv* (2022). *Submitted, under review.*

1.4. Ancestral genomes through the eukaryotic kingdom

Upon returning to the Institut de Biologie de l'ENS as an INSERM researcher in 2016, I also picked up again my interest in chromosomal rearrangements and ancestral genomes, and became involved in a

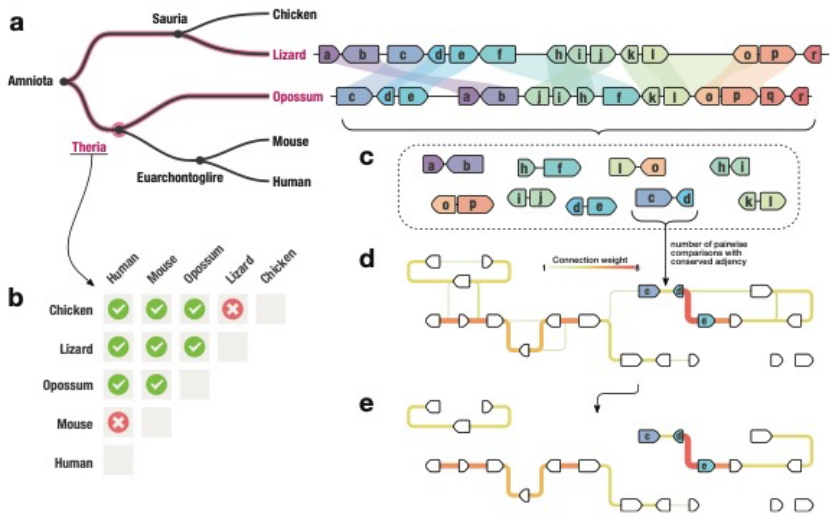

Figure 8. *Principles of AGORA.* ***a.*** *Conserved gene arrangements are identified between genome pairs.* ***b.*** *Comparisons are informative to reconstruct an ancestor if this ancestor lies on the evolutionary path between those genomes (green ticks).* ***c-d.*** *Conserved adjacencies observed in informative pairwise comparisons (c) are collected in a graph structure (d) where nodes are genes and links are weighted conserved adjacencies.* ***e.*** *This graph is linearized by traversing the links of maximal weight, providing the contiguous and parsimonious ancestral gene order.*

long-standing project to reconstruct the genome organization of ancestral species using comparative genomics methods. This work was initiated by Matthieu Muffato and Hugues Roest Crollius back in 2008 when Matthieu was a PhD student in the lab, but several major advances were implemented during the COVID-19 pandemic of 2020-2021, as we revived the project during lockdowns. This project ambitions to provide ancestral reference time-points to study the evolutionary dynamics of chromosomes and genes across different taxa. Much in the way that ancestral sequence reconstruction is fundamental to study the mechanisms of gene evolution, reconstructing the detailed karyotypic structure and organization of long-lost ancestors of extant species would open the door to investigating whether chromosomal rearrangements are favored or excluded by selection in certain genomic regions or are involved in the acquisition of specific phenotypic innovations (Coghlan *et al.*, 2005; Farré *et al.*, 2016; Kim *et al.*, 2017; Rhie *et al.*, 2021).

The methodology developed by Matthieu Muffato is named AGORA (Algorithms for Gene Order Reconstruction in Ancestors) and relies on the parsimonious assumption that gene arrangements shared between the genomes of two species are ancestral, and therefore must have existed in every ancestor since their divergence (**Figure 8**). These gene arrangements can be direct, contiguous gene-to-gene adjacencies, but can also be more distant linkage between marker genes on chromosomes. By comparing gene order across species, AGORA builds weighted gene arrangement graphs, where nodes represent ancestral genes and edges represent orthologous adjacencies observed in two or more extant genomes. These graphs can then be linearized to extract the putative ancestral gene order. While similar strategies have been implemented in other methods over the years (Ma *et al.*, 2006; Jones *et al.*, 2012; Duchemin *et al.*, 2017; Kim *et al.*, 2017), AGORA stands out in two major ways. First, AGORA is able to integrate gene arrangement information at multiple resolution scales, producing “contigs” of adjacent ancestral genes but also “scaffolding” them all the way to chromosome-scale ancestral genome assemblies, given sufficient input information. Second, AGORA can scale to very large datasets, processing hundreds of vertebrate-sized genomes in a few hours with relatively modest computational requirements and reconstructing the genomes of their ancestors at every phylogenetic node.

Using this methodology, we reconstructed the genomes of a total of 624 ancestors of vertebrates, plants and fungi, all of which were made available through the Genomicus comparative genomics database managed by Hugues Roest Crollius’s lab (https://www.genomicus.bio.ens.psl.eu/). We extensively benchmarked the resulting ancestral genomes against curated reconstructions from key ancestral species, and showed that our workflow performs as well as, and often outperforms, much more computationally intensive state-of-the-art methods. The structures of ancestral genomes reconstructed by AGORA very much resemble extant genomes, as they contain similar gene counts, similar sets of reference single-copy genes, and are assembled into chromosome-sized blocks for the majority of ancestors younger than 100 million years (**Figure 9**). We further showed that these reference ancestral genomes can be used to trace the history of chromosomal rearrangements at high resolution across

entire clades. We expect that these ancestral genome reconstruction will enable new lines of investigation into the evolutionary dynamics of karyotypes and their involvement in the acquisition of evolutionarily selected traits.

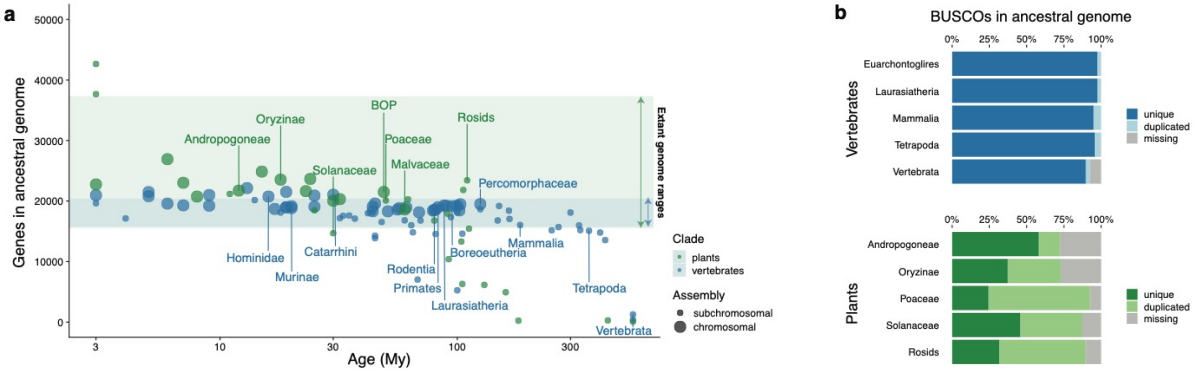

Figure 9. *Completion of ancestral genomes reconstructed by AGORA.* ***a.*** *Gene content and assembly continuity of 77 vertebrate and 33 plant ancestral genomes reconstructed by AGORA. The ranges of gene contents of extant vertebrate and plant genomes are highlighted as blue and green shading, respectively.* ***b.*** *Representation of Benchmark Universal Single Copy Orthologs (BUSCOs) in AGORA's ancestral genomes. Plant genomes, which have undergone rounds of whole-genome duplications, frequently contain a large fraction of duplicated genes.*

Related publication

M. Muffato, A. Louis, N. Thi Thuy Nguyen, J. Lucas, **C. Berthelot**[#], H. Roest Crollius[#], Reconstruction of hundreds of reference ancestral genomes across the eukaryotic kingdom. *BioRxiv* (2022). *Submitted, under review.*

2. Evolution of gene expression and regulation

2.1. Introduction

Since my PhD, and in parallel to my work on the evolution of karyotypes, I have also been interested in the functional aspects of genome evolution and, in particular, how changes in gene expression participate to species divergence and to the acquisition of novel phenotypes. Understanding how vertebrate genomes are regulated to sustain the diversity of cell types and responses necessary to the organism remains a key challenge, and here I present a brief overview of the vertebrate functional genomics concepts behind this part of my work.

2.1.1. Mechanisms of gene regulation in vertebrates

As complex multicellular organisms, vertebrates deploy a single genome into a multitude of organs, cell types and functions. This deployment is operated through the mobilization of regulatory elements, non-coding DNA elements that activate or repress gene transcription and control the spatiotemporal expression of genes during development and beyond (Heinz *et al.*, 2015). Regulatory elements are able to bind proteins known as transcription factors, which can in turn recruit the transcriptional machinery (Shlyueva, Stampfel and Stark, 2014). Transcription factors are tissue- and cell-type-specific to a degree, and in combination, they are thought to underlie the programmed cellular fate decisions and cellular responses to the environment of vertebrate organisms.

Gene expression is controlled by two major types of cis-regulatory elements: (1) promoters, which lie upstream of the transcription start site (TSS) and act as “switches”, turning gene expression on and off; and (2) enhancers, which can be located hundreds of kilobases away from their targets and control the tissue-specific activation of genes (Andersson *et al.*, 2014; Shlyueva, Stampfel and Stark, 2014; Andersson and Sandelin, 2020). Most genes have one or a few promoters, depending on their number of TSSs; however, genes can be regulated by a very large number of enhancers, especially those with complex spatiotemporal patterns of expression. Conversely, a single enhancer can target several genes, while most promoters activate only one gene. These regulatory elements are thought to operate in a concerted manner: enhancers bound by their activating transcription factors contact promoters via DNA looping in the nucleus, which promotes the recruitment of the RNA polymerase machinery and eventual transcriptional activation.

Promoters and enhancers can be characterized biochemically based on their epigenomic features: they correspond to regions of open chromatin (sensitive to DNAses and transposases), carry specific histone modifications and, in the case of enhancers, produce short, bi-directional transcripts called eRNAs (Shlyueva, Stampfel and Stark, 2014; Heinz *et al.*, 2015; Li, Notani and Rosenfeld, 2016; Andersson and Sandelin, 2020). Harnessing these biochemical properties has identified almost a million such regulatory elements in the human genome active across 190 tissues, with likely more to be discovered in unexplored cell types, developmental stages, or active only under specific environmental circumstances (ENCODE Project Consortium, 2012; Roadmap Epigenomics Consortium *et al.*, 2015). However, in most cases it remains unclear which genes are targeted by these regulatory elements, or what fraction of these non-coding elements recruit the biochemical markings of promoters and enhancers without fulfilling any actual regulatory functions (Gasperini, Tome and Shendure, 2020).

2.1.2. Evolutionary dynamics of gene regulation in vertebrates

Vertebrates have mostly conserved body plans and organismal functions, but also display considerable lineage-specific adaptations such as the loss of limbs in snakes (Kvon *et al.*, 2016), or the acquisition of lineage-specific organs such as the placenta in placental mammals (Lynch *et al.*, 2011) and an additional cardiac chamber in teleost fishes (Moriyama *et al.*, 2016). Along with protein evolution, modifications of

gene expression regulation are hypothesized to play a major role in the adoption of these new phenotypic traits.

Gene expression is generally conserved in vertebrates, with the same organs having more similar transcriptomic profiles between species than they do with other organs in the same species (Brawand *et al.*, 2011; Breschi, Gingeras and Guigó, 2017; Cardoso-Moreira *et al.*, 2019). These tissue-specific gene expression programs are controlled by tissue-specific sets of transcription factors that are largely conserved across species (Villar, Flicek and Odom, 2014). The evolutionary conservation of gene expression is however not perfect, and transcriptional program conservation erodes with evolutionary distance. This divergence is expected to be a combination of neutral drift, negative selection that maintains tissue-specific expression programs, and in some cases, positive selection where new functions are gained through regulatory changes (Romero, Ruvinsky and Gilad, 2012). Additionally, gene duplication is a major contributor to evolutionary novelty through changes in gene expression, by providing new gene copies that can acquire divergent or new expression patterns, a process called subfunctionalization (partition of ancestral gene functions and/or expression territories between sister duplicates) or neofunctionalization (acquisition of a novel function and/or expression territory; Conant and Wolfe, 2008).

Unlike gene expression, however, regulatory elements have been shown to diverge relatively rapidly in vertebrates. Transcription factor binding sites experience fast turnover: while the motifs recognized for protein-DNA binding remain essentially the same across species, their locations in the genome change rapidly: for most TFs, conservation of binding sites between human and mouse fall in the 5-30% range (Odom *et al.*, 2007; Schmidt *et al.*, 2010). This turnover is not limited to binding site modifications within otherwise conserved regulatory regions. Indeed, enhancers especially exhibit fast evolutionary divergence in vertebrates, and many enhancers detected using biochemical evidence are not active in other related species, even when an orthologous sequence exists (Degner *et al.*, 2012; Cotney *et al.*, 2013; Villar, Flicek and Odom, 2014; Zhou *et al.*, 2014; Reilly *et al.*, 2015). This is in line with the low sequence conservation of most regulatory elements detected in the human genome, which experience much lower negative selection than coding sequences (ENCODE Project Consortium, 2012; Reilly and Noonan, 2016).

2.2. Evolution of gene expression after whole-genome duplication

My first foray into gene expression evolution was during my work on the rainbow trout genome, towards the end of my PhD. In addition to genome organization evolution, I also studied how gene expression changes after whole-genome duplication and may lead to functional innovations. With our collaborators at INRA, we produced transcriptomic data across 15 tissues in trout, and identified four major patterns of gene expression evolution after whole-genome duplication (**Figure 10**), where (1) both copies retained the ancestral pattern (green); (2) both copies retained the ancestral pattern but with a disequilibrium in expression levels in favor of one copy (orange); (3) one copy retained the ancestral pattern while one diverged in expression (purple); and (4) one copy retained the ancestral pattern while the other was largely repressed, and possibly on the path for pseudogenization (red). These four patterns were associated with differences in sequence divergence and in gene functions, with genes involved in environment perception and response enriched in categories of high differentiation. These results suggested that new gene copies produced by whole-genome duplication may become appropriated by new processes and result in evolutionary innovations via changes in their transcription patterns.

Related publication

C. Berthelot, F. Brunet, D. Chalopin, A. Juanchich, M. Bernard, B. Noël, P. Bento, C. Da Silva, K. Labadie, A. Alberti, J.-M. Aury, A. Louis, P. Dehais, P. Bardou, J. Montfort, C. Klopp, C. Cabau, C.

Gaspin, G. H. Thorgaard, M. Boussaha, E. Quillet, R. Guyomard, D. Galiana, J. Bobe, J.-N. Volff, C. Genêt, P. Wincker, O. Jaillon, H. R. Crollius, Y. Guiguen, The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. *Nat Commun*. **5** (2014), doi:10.1038/ncomms4657. *IF: 14.9*

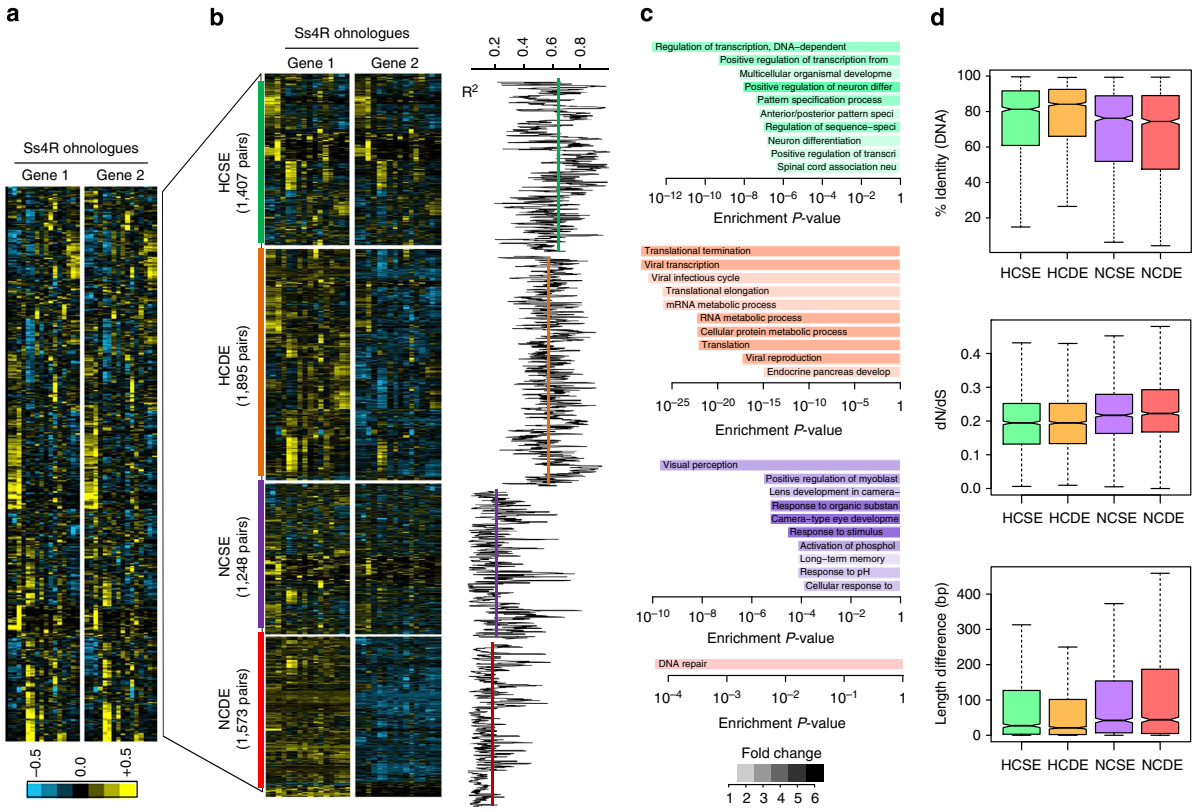

Figure 10. *Expression of duplicated genes in rainbow trout.* ***a.*** *Expression levels in 15 tissues, normalized and centered independently for each gene.* ***b.*** *Expression levels in 15 tissues, normalized and centered by pair of duplicated genes, delineating four groups of genes with either high or low correlation (HC/NC groups) and either similar or different average expression levels (SE/DE groups).* ***c.*** *Functional enrichments in each of the four categories of duplicated genes.* ***d.*** *Sequence comparisons between duplicated genes shows that gene pairs with correlated expression (NC groups) tend to have higher sequence conservation and to be under stronger selective pressure, showing that functional divergence is associated with sequence divergence.*

2.3. Evolution of transcriptional regulation in mammalian genomes

2.3.1. Evolution of enhancers and promoters in mammals

In 2012, I started my postdoc at EMBL-EBI in Cambridge in Paul Flicek's lab, a specialist of the evolution of vertebrate genomes and their functional mobilization across tissues. Paul was involved in a fruitful long-term collaboration with Duncan Odom's lab at CRUK Cambridge to investigate the evolution of gene expression and regulation in mammals, and I became the computational lead on a large comparative genomics effort between both labs to analyze the genome-wide conservation of regulatory elements using liver as a model tissue. In collaboration with Diego Villar, a postdoc in the Odom lab, we profiled the liver epigenomic landscapes in twenty mammalian species spanning over 180 million years of evolution, using a combination of histone mark specific chromatin immunoprecipitation assays (ChIP-seq). In each species, we profiled the genomic regions enriched in H3K4me3 and H3K27ac histone modifications, which in combination are highly enriched at active promoters, while active enhancers typically carry H3K27ac but not H3K4me3. These assays allowed us to delineate the main active gene

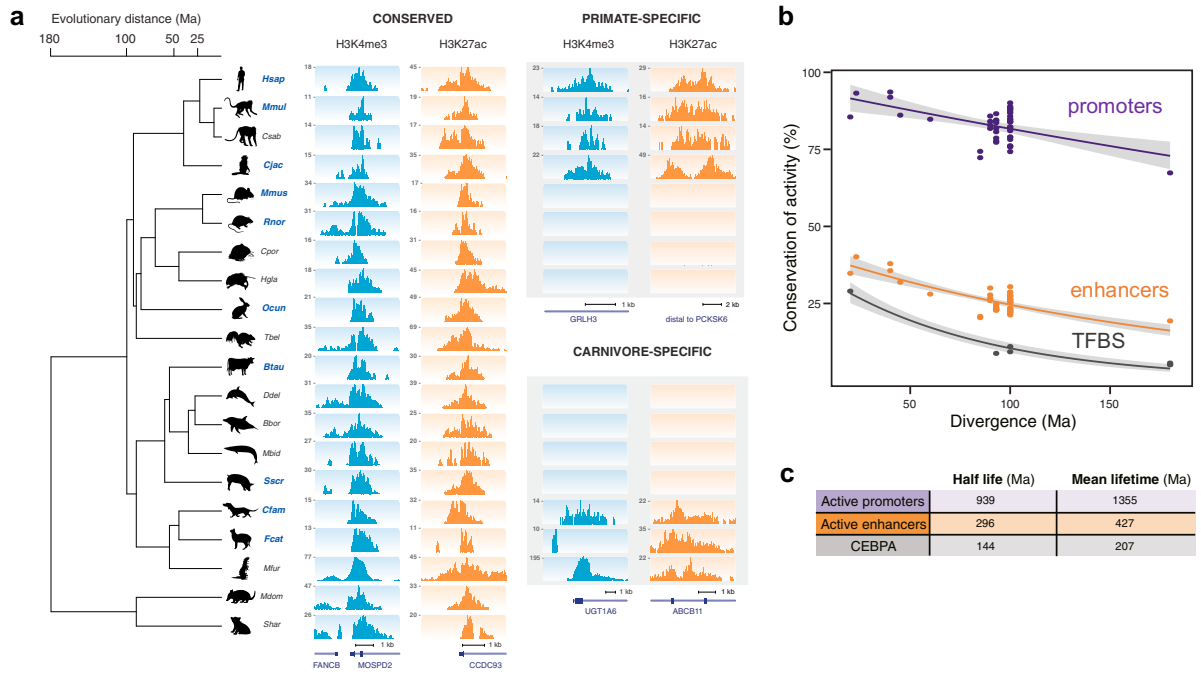

Figure 12. *Evolutionary conservation of liver regulatory elements across 20 mammalian species.* ***a.*** *Profiles of H3K4me3 and H3K27ac histone modification enrichment in example genomic regions, where activity is either conserved across all study species, specific to primates or specific to carnivores.* ***b.*** *Conservation of regulatory activity over evolutionary time, showing that promoters are largely conserved while enhancers diverge rapidly, although not as fast as individual transcription factor biding sites (CEBPA used as a representative example). Lines represent an exponential decay fit, greyed areas represent the 95% confidence interval of the fit.* ***c.*** *Half-lives and mean lifetimes of regulatory elements and transcription factor binding sites (CEBPA used as a representative example), calculated from the exponential decay fits in* ***b****.*

regulatory elements in each species, which we further compared across species using whole-genome alignments (**Figure 11a**). At the time when these experiments took place, only a handful of studies had compared active regulatory landscapes between key mammals, typically human and mouse (Degner *et al.*, 2012; Shibata *et al.*, 2012; Xiao *et al.*, 2012; Cotney *et al.*, 2013): while it was becoming apparent that regulatory elements were significantly less conserved than anticipated and observed in lineages with smaller genomes such as flies (Arnold *et al.*, 2014), the evolutionary dynamics and breadth of conservation of these elements was poorly understood.

Our results demonstrated that the evolutionary dynamics of promoters and enhancers differ dramatically: while promoters are typically conserved across species, with half-lives similar to genes, enhancers are evolutionarily labile and turn over quickly in mammals (**Figure 11b-c**). We showed that this evolutionary plasticity is not due to drastic differences in sequence evolution or content between promoters and enhancers, as most enhancers had detectable sequence orthologs across comparable species subsets to promoters; but these orthologous sequences were typically not active as enhancers in other species. We

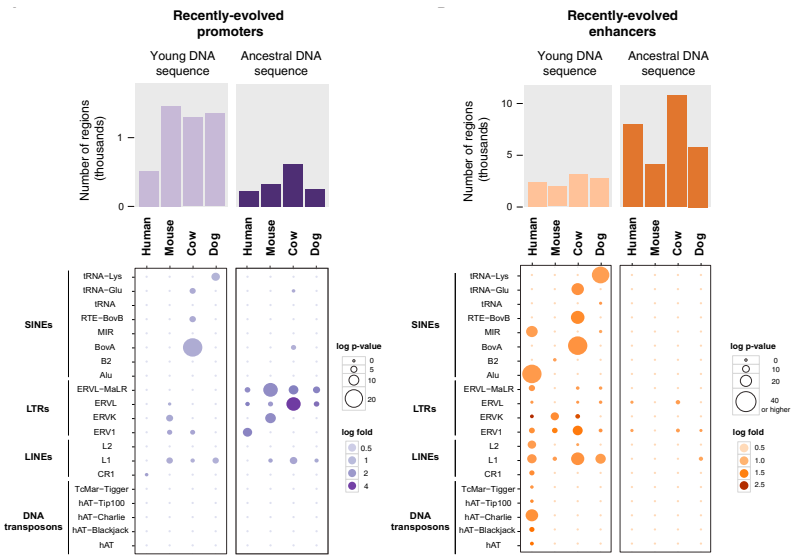

Figure 11. *Sequence age and transposable elements enrichments in liver promoters and enhancers recently evolved in human, mouse, cow and dog.*

further identified a set of liver promoters and enhancers whose activity is highly conserved in mammals, and showed that this activity conservation associates with enhanced evolutionary constraint at the sequence level compared to other regulatory elements of the same type.

Finally, we investigated recently-evolved regulatory elements that are either lineage- or species-specific in our dataset: these represented a significant fraction (~30%) of all detected enhancers. These recently-evolved regulatory elements have different evolutionary origins (**Figure 12**). We showed that new promoters typically correspond to recently acquired sequences as well as recycled LTR-type transposons, especially endoretroviruses (ERVs) which are known to produce long non-coding RNAs in mammals (Fort *et al.*, 2014; Hoeppner *et al.*, 2018). New enhancers, however, typically lie in older DNA sequences and presumably arise through mutations and evolutionary tinkering (Yokoyama, Zhang and Ma, 2014; Kratochwil and Meyer, 2015). About 25% of new enhancers come from more recently acquired sequences, and these are enriched in young transposable elements, which have been shown to be important contributors to regulatory novelty in mammalian genomes (Kapusta *et al.*, 2013; Trizzino *et al.*, 2017; Fueyo *et al.*, 2022). This work was considered a landmark contribution to our collective understanding of the evolutionary dynamics of regulatory landscapes in mammals, and was highlighted in cover articles in *Nature Genetics* and *Current Biology*.

Related publication

D. Villar*, **C. Berthelot***, S. Aldridge, T. F. Rayner, M. Lukk, M. Pignatelli, T. J. Park, R. Deaville, J. T. Erichsen, A. J. Jasinska, J. M. A. Turner, M. F. Bertelsen, E. P. Murchison, P. Flicek, D. T. Odom, Enhancer Evolution across 20 Mammalian Species. *Cell*. **160**, 554–566 (2015). *IF: 41.6*

2.3.2. Resilience of gene expression to regulatory change

The results from our 2015 publication in *Cell* raised considerable questions (and some degree of skepticism) as they were in apparent contradiction with the high conservation of gene expression levels observed in mammalian tissues, which are thought to be controlled by tissue-specific enhancers. To address these questions, we further pursued this line of work in a second, related project that investigated how the evolution of a gene's regulatory landscape – the collection of active regulatory elements around it – correlates with modifications of the transcription levels of this gene. We profiled the liver transcriptome in 15 out of the 20 mammalian species initially included in the study for which high-quality RNA samples could be generated, and quantified gene expression to complement the previously generated regulatory landscapes in the liver of these species. We showed that mammalian

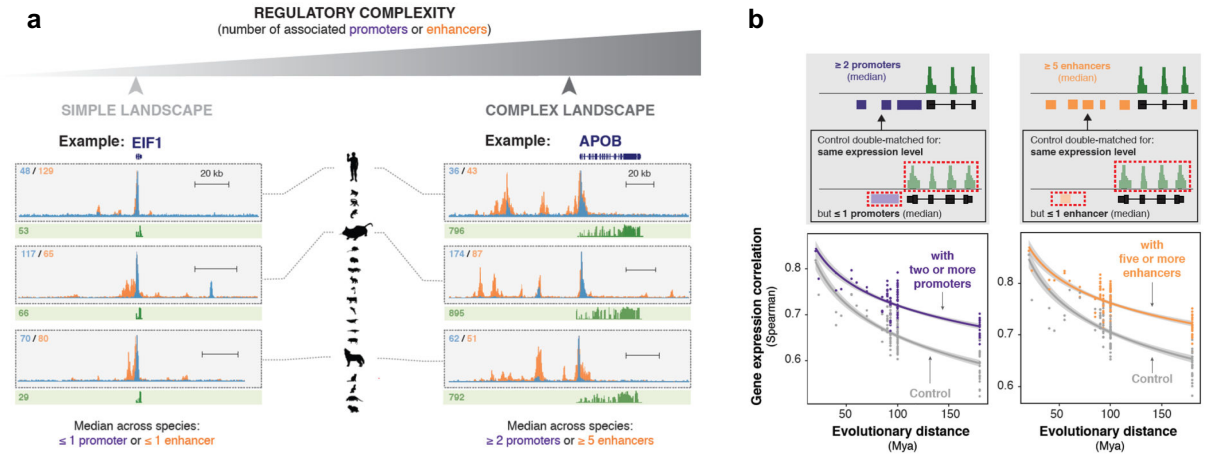

***Figure 13**. Complexity of regulatory landscapes underlies gene expression conservation. **a.** Examples of genes with low (EIF1) and high (APOB) regulatory landscape complexity in liver, and their surrounding histone modification landscapes in three representative mammalian species (human, mouse and dog). **b.** Genes with complex regulatory landscapes retain more conserved gene expression over evolutionary time, even when controlling for the effects of expression level.*

genes retain their regulatory complexity through evolution, although their individual regulatory elements can experience a high degree of turnover: genes embedded in complex landscapes are surrounded by large numbers of active regulatory elements in most species, while genes with minimal regulatory landscapes rarely acquire large numbers of new regulatory elements (**Figure 13a**). Moreover, we discovered that genes with complex regulatory landscapes exhibit higher and more tightly conserved expression in mammals (**Figure 13b**). The conservation of individual regulatory elements, while significantly associated with conserved gene expression, was not a dominant effect, suggesting that the evolutionary resilience of gene expression in mammals is largely the result of selective pressure distributed over many regulatory elements that can replace and compensate one another. Interestingly, we showed that recently-acquired regulatory elements significantly contribute to this ongoing turnover.

Our observations have since been consistently reproduced across other tissues, cell types and species, and have contributed to reframe gene expression regulation from a tightly regulated process to a complex, collective effort with many weak actors complementing and buffering each other's effects (Hill, Vande Zande and Wittkopp, 2020). This work further highlighted the need for methodological developments to rigorously test whether the activity of specific regulatory elements and gene expression levels evolve under neutral or selective pressures, and answer hypothesis-driven questions beyond the description of genome-wide evolutionary dynamics.

Related publication

C. Berthelot*, D. Villar*, J. E. Horvath, D. T. Odom, P. Flicek, Complexity and conservation of regulatory landscapes underlie evolutionary resilience of mammalian gene expression. *Nature Ecology & Evolution*. **2**, 152–163 (2018). *IF: 15.4*

2.4. Functional mutations in human regulatory elements

Upon returning to the Institut de Biologie de l'ENS as a faculty member, I became involved in a project exploring the functional potential of genetic variants occurring in non-coding regulatory elements in human. This project is in collaboration with Hugues Roest Crollius and with a nationwide consortium coordinated by INSERM (Transversal Program for Genetic Variability), which aims to analyze and interpret genetic variation in an extensively phenotyped sample of the French population (Zins, Goldberg, and CONSTANCES team, 2015). During this project, I collaborated with and mentored Lambert Moyon, a PhD student under Hugues's supervision, who developed a machine-learning model which scores non-coding genetic variants based on their evolutionary conservation, their epigenomic features (such as histone modifications or transcription factor binding sites in representative tissues), their sequence context and their tridimensional contacts with genes of interest, in order to identify non-coding variants with a probable phenotypic impact. The software uses a random forest algorithm to classify genetic variants based on a predicted pathogenicity score, and was trained by comparing well-identified disease-causing non-coding variants to tailored lists of benign control variants that alleviate a

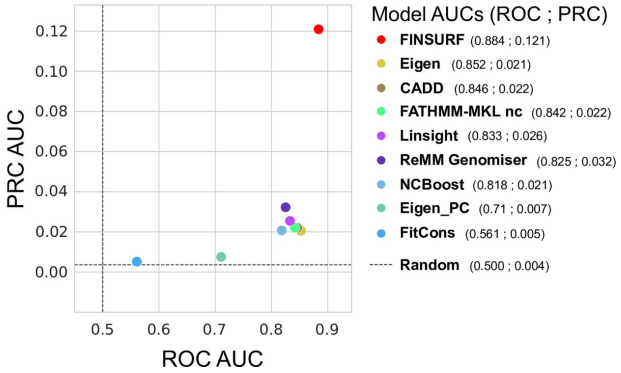

Figure 14. *Performances of FINSURF compared to other state-of-the-art algorithms to separate pathogenic non-coding genetic variants from control polymorphisms. All algorithms were applied to the same set of test variants (which are fully independent of the set used to train FINSURF). ROC: Receiver Operating Characteristic curve; PRC: Precision Recall Curve; AUC: Area Under the Curve. Higher ROC and PRC AUC indicate that the model is more discriminative and more sensitive.*

number of known positional biases linked to gene density and other genomic heterogeneities in the human genome. This software, named FINSURF, outperforms state-of-the-art methods (**Figure 14**) and provides visualization tools to interpret the molecular mechanisms of action of candidate variants. We are currently expanding this model to (i) select tissue-specific descriptors to identify candidate genetic variants tailored to phenotypes or diseases of interest, and (ii) predict whether variants result in over- or under-expression of target genes, a project carried out by Franklin Delehelle, a postdoc in the Roest Crollius lab.

Related publication

L. Moyon, **C. Berthelot**, A. Louis, N. T. T. Nguyen, H. Roest Crollius, Classification of non-coding variants with high pathogenic impact. *PLoS Genetics*. **18**, e1010191 (2022).

Software

FINSURF: Functional Interpretation of Non-coding Sequences Using Random Forests. https://github.com/DyogenIBENS/FINSURF

Web server

FINSURF online: https://www.finsurf.bio.ens.psl.eu/

3. Current and future research directions

In 2019, I obtained an ERC Starting Grant to start my independent lab, and was recruited shortly after as a group leader at the Institut Pasteur in Paris in an international call for tenure-track junior groups. My lab effectively started in September 2021, after a long year of false starts, lab closures and distanced work due to the COVID-19 pandemic. I recruited two postdocs in 2020 to kick off the ERC project, and although our activities have been severely curtailed by the restrictions in place to contain the pandemic, their input, help and enthusiasm has been invaluable in setting up a functional lab in a minimal amount of time as we moved to our new premises. Since joining Institut Pasteur, we have hosted two M2 students, a tech student, and I have recently recruited an experienced research technician to become our lab manager and facilitate the daily running of operations in the lab. I have also recruited a PhD student, the first who will be entirely under my supervision, and who will join us in October 2022 after doing her M2 internship with us earlier this year. Altogether, these past months have been a challenging – but rewarding – experience, and it has been a humbling joy to witness the young scientists under my care come together and gel as a team, as I myself learn the ropes of effective supervision, lab management and administrative wrangling. In this part of the manuscript, I describe the scientific projects that are either ongoing or under development in the lab.

3.1. The evolution of menstruation

I became interested in the evolution of reproductive traits, and especially menstruation, towards the end of my postdoc. Menstruation is one of these evolutionary traits that “do not make sense”: what can possibly be the adaptive advantage conferred by losing blood and tissue every month, which seems wasteful and inefficient and is associated with disorders and diseases in a large fraction of the population? In this section, I give an overview of the main tenets and goals of the central project of my lab, which has received financial support from ERC and Institut Pasteur.

A woman will, on average, menstruate 450 times during her lifetime. Menstruation corresponds to the shedding of the uterine lining (endometrium) when fecundation has not occurred. This physiological process affects half of the human species and is associated with severe gynaecological conditions (Dunselman *et al.*, 2014). Yet, the molecular pathways responsible for menstruation remain relatively understudied despite their tremendous importance for human health and reproduction (Evans *et al.*, 2016). This tissue has garnered renewed interest in recent years due to its high disease prevalence rates and its potential as a source of easily accessible stem cells (Evans *et al.*, 2016): functional studies have highlighted important changes in gene expression along the human hormonal cycle, but the menstrual time point was often excluded (Krjutškov *et al.*, 2016; Lucas *et al.*, 2018; Wang *et al.*, 2018).

Interestingly, menstruation is a recent evolutionary innovation in primates, and has appeared convergently at least three times in mammals: menstruation has also been observed in the spiny mouse, the elephant shrew, and several bat species (Emera, Romero and Wagner, 2012; Bellofiore *et al.*, 2017). Expression of this trait varies substantially even amongst related primates: humans, chimpanzees and Old World monkeys such as macaques menstruate, while lemurs do not, and the expression of the trait in New World monkeys such as marmosets, saimiris and capuchins is variable and does not strictly follow the phylogenetic tree (Strassmann, 1996). Non-menstruating primates experience a similar hormonal cycle with an early phase governed by estrogen followed by a late phase under progesterone control, but their endometrium is

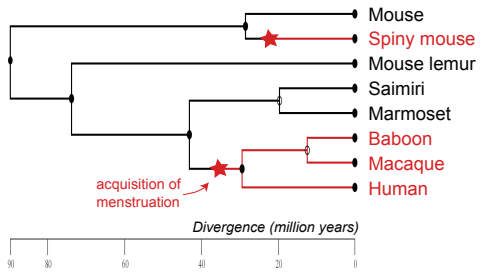

***Figure 15**. Mammalian species selected to study the molecular emergence of menstruation. The selected species cover two convergent apparitions of menstruation in rodents and primates.*

reabsorbed by the uterine wall at the end of the cycle, as in most other mammals.

While some phenotypic differences between closely related species are due to protein sequence changes, most of them result from regulatory changes modifying the spatial and temporal patterns of gene expression (Romero, Ruvinsky and Gilad, 2012). Comparisons of gene expression and regulation between species have made critical contributions to our understanding of evolutionary innovations across a wide variety of traits (Reilly and Noonan, 2016), such as the genetic networks involved in the complexity of the human brain (Reilly *et al.*, 2015; Emera *et al.*, 2016) and the features of the human face(Prescott *et al.*, no date). Comparative evolutionary studies of the female reproductive tract have so far exclusively focused on post-fecundation mechanisms, especially placentation (Lynch *et al.*, 2011; Chuong *et al.*, 2013; Griffith *et al.*, 2017). While other primate- and human-specific traits have commended attention in brain (Reilly *et al.*, 2015), limb (Cotney *et al.*, 2013) or testis (Soumillon *et al.*, 2013), menstruation has been largely ignored. As menstruation has been gained and/or lost several times in mammals, a comparative study has considerable leverage to identify the mechanisms involved in its evolution. Indeed, menstruating species should be functionally more similar compared to non-menstruating, more closely-related species specifically for pathways involved in this trait. To decipher the molecular genetics of menstruation, we are performing an integrated profiling of gene expression and active regulatory landscapes in uterine lining samples from six primate species and two rodents, four of which menstruate and four of which do not (**Figure 15**). This analysis will allow us to investigate outstanding questions regarding the functional underpinnings of menstruation, a key physiological process in human reproduction and a major evolutionary innovation in the primate reproductive tract.

3.1.1. Characterizing the late-cycle uterine endometrium in primates and rodents

Tissue collection. No publicly accessible collection exists for primate endometrial tissue, and functional genomics data from this tissue in human and model animals is scarce and inconsistent. To address this, we are collecting healthy endometrial tissue from human donors 8 days prior and 2 days into menses, in collaboration with Prof. Geoffroy Canlorbe in the gynaecology service at Hôpital Pitié-Salpétrière (Paris). We are working with three major primate research facilities with extensive experience in reproductive biology (CNRS Primatology Station in Rousset-sur-Arc; Simian Laboratory at Strasbourg University; National History Museum lemur research facility in Brunoy) to track the hormonal cycles and collect endometrial tissue at matched time points in five primate species spanning the three main primate families (mouse lemurs, saimiris, marmosets, macaques and baboons). We are also collaborating with the mouse facility at Institut Pasteur to collect comparable samples in genetically diverse mice strains, and with Dr Kathy Millen's lab at Seattle Children's Hospital to collect matched spiny mice samples, which is the only known menstruating rodent and represents a convergent evolution of the trait compared to primates.

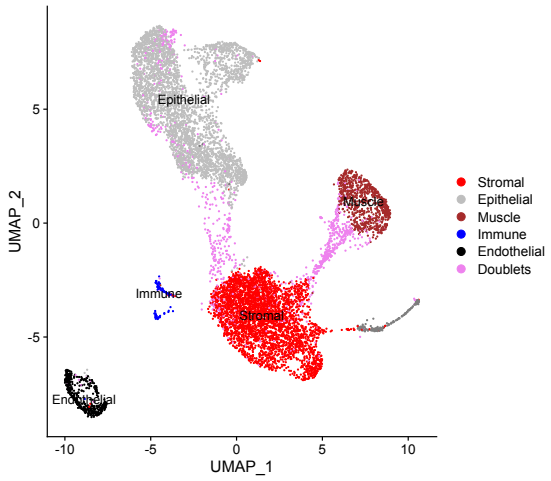

***Figure 16**. Main cell populations in the macaque uterine endometrium, based on their single-nuclei transcriptomic profiles projected in a 2D space using UMAP (Uniform Manifold Approximation and Projection). All expected cell types are represented.*

Single-nuclei transcriptomics and epigenomics. Our main objective is to unveil the functional changes in transcriptional regulation and gene expression that result in menstruation instead of uterine lining resorption. Cellular heterogeneity across samples can majorly confound comparative functional genomics analyses done on bulk tissue, and we are therefore carrying out functional analyses of the endometrium at the cell type level. We are profiling the transcriptomes and open chromatin regions in the collected tissue samples using combined single-nuclei RNA-seq and ATAC-

seq, which gives us access to the different cell populations present in the endometrial lining (**Figure 16**). Our first priority is to produce an integrated functional landscape of gene expression and regulation in stromal and glandular cells, the two main cell types in the endometrium, before and during menses, in all six primates and in both rodent species.

3.1.2. Obtaining cellular models of the uterine endometrium

A second key aspect of our current work is to develop tractable lab models to perform comparative analyses of the uterine endometrium across species without relying on fresh tissue samples from live animals, which are difficult to obtain. Over the past few years, endometrial organoids have emerged as a powerful tool to study endometrial function and disease (Boretto *et al.*, 2017; Turco *et al.*, 2017). Organoids are self-forming 3D cell assemblies grown in culture from primary tissue samples, which reproduce the cellular organization of their tissue of origin better than classical 2D tissue culture. Endometrial organoids have been successfully obtained from a variety of endometrium cell sources in both human and mouse, including from menstrual fluid in humans (Cindrova-Davies *et al.*, 2021). These organoids are either composed of epithelial cells or a combination of epithelial and stromal cells, and exhibit the key cellular functions and hormone responses of uterine endometrium.

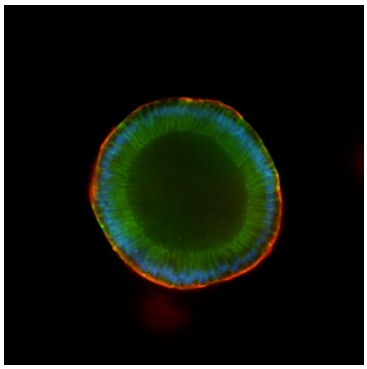

Figure 17*. Endometrial epithelial organoid obtained from macaque tissue observed by confocal microscopy. Green: E-Cadherin (epithelial cell membranes); blue: DAPI (nuclei); red: Laminin (basal lamina).*

Since arriving at Institut Pasteur, we have been adapting these techniques to produce, maintain and hormonally stimulate endometrial organoids from different species (**Figure 17**). Ultimately, our goal is to set up a collection of organoids from a variety of primate and non-primate species, including species that are not necessarily accessible for experimentation. To this aim, we have started coordinating with several zoos in France to retrieve uterine samples from animals that either died of natural causes or have to be euthanized for humane reasons. Our first objective is to stimulate organoids from multiple menstruating and non-menstruating species with similar estrogen and progesterone regimens in controlled culture environments, and profile the transcriptomes of the organoids. We expect that this experiment will reveal which transcriptomic changes separating menstruating and non-menstruating are inherent to their cellular programmes, and which result from differences in their hormonal and cellular environments.

3.2. Human genetic variation and menstrual diseases

As we started digging into the evolution of menstruation, it became clear to me that the functional underpinnings of the female reproductive tract are a largely neglected area of research, despite their obvious importance for both evolutionary genetics and medical research. We recently developed two projects that investigate human variation in uterine phenotypes, which complement our mammalian evolutionary project presented above.

3.2.1. Illuminating the mechanisms of endometriosis using menstrual fluid

Endometriosis is a chronic gynecologic disease that occurs when endometrial tissue, which normally lines the inside of the uterus, grows outside of the uterine cavity (Laux-Biehlmann, d'Hooghe and Zollner, 2015; Saunders and Horne, 2021). Endometriosis is thought to develop when endometrial fragments contained in menstrual fluid flow back into the abdomen through the Fallopian tubes during periods and pathologically attach in the body cavity. This backflow phenomenon has been observed in 90% of

women of reproductive age, but only 10% will develop endometriosis. Endometriosis symptoms are enormously diverse, and diagnosis requires examination by expert specialists. Time to diagnosis is typically long and difficult, on average 7 years in France, and invasive surgery remains the state-of-the-art methodology for definitive diagnosis. Today, a high-stakes challenge in patient care is the development of new methods to diagnose endometriosis early, quickly, non-invasively and with high confidence (Chapron *et al.*, 2019; Hudson, 2022).

The objective of this project is to characterize the endometrial cellular populations contained in menstrual fluid in endometriosis patients and control donors, in order to identify potential biomarkers for non-invasive endometriosis diagnosis. We are collecting menstrual flow from healthy donors and endometriosis patients to identify the different cell types coming from the uterus endometrial lining present in the menstrual fluid. Our objective is to compare both groups of donors using a combination of single-cell transcriptomics and bulk transcriptomics on sorted cell populations, in order to detect changes in cell proportions, cellular viability and gene expression that differentiate endometriosis patients from healthy controls. To further characterize these changes, we plan to cultivate donor endometrial cells as organoids and assess the functionality of the cells (growth rate, viability, clonality, somatic mutation load) to identify functional differences between patients and controls. Characterizing menstrual fluid as a biological tissue in health and disease will further our understanding of disease mechanisms involved in endometriosis. We expect that this project will help identify biomarkers to detect endometriosis early, easily and non-invasively in order to reduce diagnostic delays and improve patient care.

This project is performed in tight partnership with Ludivine Doridot (Institut Cochin, Paris), an expert in reproductive biology and functional genomics, and in collaboration with Angela Goncalves (DKFZ, Heidelberg), who specializes in somatic tissue evolution and cancer processes. In my lab, this project has been largely driven by Axelle Brulport (experimental postdoc), who took leadership on the project and obtained two pilot grants from the EndoFrance patient association and the Fondation pour la Recherche sur l'Endométriose to pursue the proof of concept.

3.2.2. Evolutionary signatures of uterine functions on the human genome

Genome evolution can provide critical insight into gene functions and disease mechanisms by documenting how genetic modifications behave through natural selection (Benton *et al.*, 2021). This evolutionary approach is particularly relevant to study the uterus, as genetic variations affecting reproductive functions impact fertility and therefore evolutionary fitness. Interestingly, the uterus is a variable organ in humans, with both uterine life-history traits and disease prevalence rates displaying variation within and between human populations with different genetic ancestries (Bougie *et al.*, 2019; Giuliani, As-Sanie and Marsh, 2020). In this project, we hypothesize that genetic variations modifying uterine functions have been, and remain, crucial contributors to human evolution and disease, and we propose to leverage recent evolutionary signals in the human genome to illuminate the genetic mechanisms of uterine functions.

To investigate how uterine functions have shaped the recent evolution of the human genome, we are relying on public datasets correlating common polymorphisms to uterine traits and diseases in hundreds of thousands European, Asian and African individuals from genome-wide association studies (GWAS; datasets from the UK Biobank, FinnGen, and EBI GWAS databases; Bycroft *et al.*, 2018; Buniello *et al.*, 2019; Kurki *et al.*, 2022). We have integrated 272 GWAS studies on variable uterus phenotypes and curated a list of 877 common genetic variants statistically associated with these phenotypes of interest. These genetic variants allowed us to delineate a map of 386 regions of interest in the human genome that strongly associate with female reproductive functions (**Figure 18**). We are in the process of validating which of those genomic regions are active in the human uterus and link them to specific cell types using gene expression and regulation data from public sources and produced by the lab for the menstruation evolution project.

Our objective is then to investigate whether the coding and non-coding elements of these regions have been subjected to natural selection – negative and positive – in the recent human past. We plan to explore patterns of losses of genetic diversity between human populations over the past 100,000 years (Laval *et al.*, 2021), as well as the dynamics of these genomic regions during the evolution of the human lineage since its divergence from the chimpanzee (6 Mya). Eventually, we want to integrate those results with our study of endometrial tissue evolution across primates and rodents, with the expectation that functional genomic regions that have contributed to the emergence of menstruation during mammalian evolution remain potential hotspots of adaptation in the human genome, as they impact key reproductive functions and fitness.

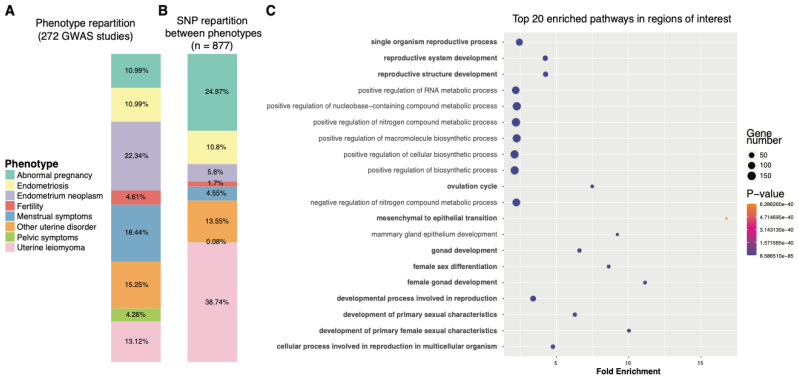

Figure 18. *Identification of genomic regions involved in female reproductive functions using flag SNPs from a meta-analysis of genome-wide association studies.* ***A.*** *Repartition of the uterine-related phenotypes amongst GWAS studies.* ***B.*** *Repartition of SNPs across phenotypes.* ***C.*** *Top 20 biological pathways enriched around the flag SNPs, computed with GREAT v.4.*

3.3. Methodological advances for comparative functional genomics

Complementary to our experimental work, my lab includes computational biologists and bioinformaticians, and one of our areas of research is how we can improve upon the state of the art in terms of methods for comparisons of functional and genomic information across species. Almost all of the methods to compare gene expression or regulation currently at our disposal were developed to be used between conditions within a single species. These methods suffer from serious flaws when applied to comparisons between species (Romero, Ruvinsky and Gilad, 2012; Dunn, Luo and Wu, 2013; Dunn *et al.*, 2018). Firstly, they do not account for technical factors that may differ between different reference genomes, such as quality of the genome annotation, ability to uniquely map sequencing reads, etc. Secondly, these methods do not consider that differences between species will reflect a combination of functionally relevant modifications and neutral evolutionary divergence: as such, most functional genomics methods are confounded when applied to comparisons across species. In the same way that raw comparisons of sequence similarity are not necessarily informative about conservation of function, functional genomics methods require adaptations to decipher neutral from functional change, and this has been a fascinating problem that we are trying to address to different ways as required for specific biological problems.

3.3.1. Phylogenetic models for the evolution of regulatory elements

Phylogenetic models are foundational in evolutionary studies by providing statistical models describing how a trait – a binary phenotype, a quantitative trait or a molecular trait, such as a sequence – has changed during the evolution of a group of species (Liò and Goldman, 1998). This statistical model can then be extended to test evolutionary hypotheses: for example, whether the trait evolves under selection or neutrality, has evolved significantly faster in a specific branch or clade, or presents different selective optima in different parts of the evolutionary tree. Phylogenetic models are widely applied to sequence analysis, where the frequencies and patterns of sequence substitutions either at the DNA or the protein

levels are used to infer the relationships between species, as well as the evolutionary dynamics of specific genes and non-coding sequences (Yang, 1997; Whelan, Liò and Goldman, 2001). In quantitative trait evolution, phylogenetic models are frequently applied to infer the existence of stabilizing or directional selection on measurable phenotypes, and to test whether separate traits co-evolve and may be functionally related (Manceau, Lambert and Morlon, 2017). In both fields, the statistical methods underlying the models are typically parametric, and fit the measured trait differences between species or individuals to an expected statistical distribution to estimate how the trait likely evolved along the species tree given the observed data.

I got interested in extending phylogenetic models to transcriptomics and regulatory genomics data during my postdoc, as I became aware that the methodologies used in the field at the time – and to a large extent, still used today – were inappropriate. Most comparative transcriptomics and regulomics studies rely on methods developed to detect differential gene expression or differential transcription factor binding between conditions, and typically perform pairwise comparisons between species. Neither of these methods account for the relatedness between species, embedded into the phylogenetic structure of the tree (Dunn, Luo and Wu, 2013; Dunn *et al.*, 2018). As a result, most studies perform a large number of non-independent comparisons, resulting in high false-discovery rates; and they cannot conclude on whether the detected changes correspond to selective changes or neutral drift. Ideally, comparative functional genomics should aim to model changes of gene expression or regulatory activity along the species tree to identify evolutionary branches where the magnitude of change is greater than expected, which could be evidence of selection (Price *et al.*, 2022).

Such an approach has previously been developed by Rori Rohlfs and Rasmus Nielsen for gene expression data, named EVE (Expression Variance and Evolution; Rohlfs, Harrigan and Nielsen, 2014; Rohlfs and Nielsen, 2015). This method models the variance of gene expression measured within species and between species to detect differences in gene expression between clades, and is described as a phylogenetic ANOVA test. The underlying statistical model assumes that gene expression evolves as an Ornstein-Uhlenbeck process, so that gene expression tends to drift neutrally as a Brownian process (variance increases proportionally to time), but is also constrained by selection which drives at back towards an optimal value. EVE models gene expression variance within and across species as a function of time using maximum likelihood to estimate the selective optimum and strength of selection from the existing data. As a corollary, the model can also detect phylogenetic changes in selective optima between groups of species, which is its most used feature.

We are adapting this approach for regulatory functional genomics data in order to detect changes in gene regulatory region usage between species (**Figure 19**). Our objective is to propose phylogeny-aware strategies to identify sets of regulatory elements that were recruited on branches of interest in the species tree, and that are enriched in regions under selection in that branch (for other strategies, see Yang *et al.*, 2018; Dukler, Huang and Siepel, 2020). Compared to gene expression data, regulatory data presents a much lower sequence conservation between species, a higher evolutionary turnover

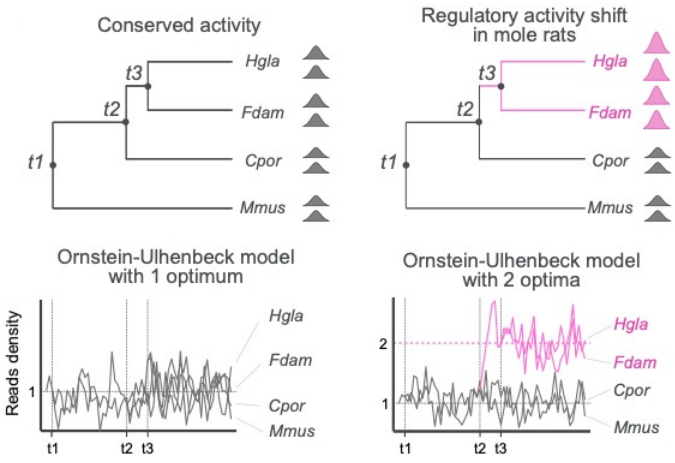

Figure 19. *Schematic representation of phylogenetic analysis of regulatory activity in a group of rodents: the naked mole rat (Heterocephalus glaber; Hgla), the Damaraland mole rat (Fukomys damarensis; Fdam), the guinea pig (Cavia porcellus; Cpor), and the mouse (Mus musculus; Mmus). We detect regulatory regions where the activity signal, measured from histone mark ChIP-seq, has changed significantly in the mole rat clade (between time points t2 and t3). The maximum likelihood model for these regions contain two selective optima (right), one for mole rats and one for the outgroups. The lower panels represent evolutionary simulations based on the parameters of the maximum likelihood models above.*

represent interesting lines of inquiry from both mathematical and evolutionary standpoints, which we are exploring.

We are applying these methods to our project on the evolution of menstruation in mammals, where we are generating ATAC-seq data from the uterine lining of different mammalian species. We are also using this approach in a separate project in collaboration with Diego Villar (Queen Mary University, London) which investigates the evolution of gene regulation in liver and heart in mole rats. Mole rats are rodents that evolved striking adaptations to their subterranean ecological niche, including somatosensory regression, hypoxia tolerance, circadian clock modifications and metabolic adaptations (Kim *et al.*, 2011). The molecular bases of several of these adaptations are well-described in heart and liver cells, and we study how these adaptations are encoded genetically through modifications of gene expression regulation compared to non-adapted outgroup species such as guinea pig or mouse (**Figure 19**). Our preliminary results show that the regulation of several pathways involved in hypoxia resistance, metabolic processing of sugars and immune processes is modified in mole rats, in line with phenotypic observations.

3.3.2. Comparative genomics in the single cell era

Finally, we are also interested in the development of methodologies to identify and compare orthologous cell types between species from single cell sequencing data. Intuitively, orthologous cell types largely perform the same functions in different organisms, and should therefore present highly correlated gene expression programs. Although this intuition holds generally true, matching cell types between species based on their transcriptomic profiles measured by single-cell sequencing remains challenging for both biological and technical reasons (Geirsdottir *et al.*, 2019; Shafer, 2019). Gene repertoires and gene expression diverge over time, which can obscure the correlations of expression especially when looking at specific sub-populations within larger cell types. Additionally, gene expression data as measured by single-cell transcriptomics remains sparse: with a sequencing depth of 10,000 to 20,000 reads per cell, only a stochastic fraction of the transcriptome of each cell is captured (Stegle, Teichmann and Marioni, 2015). This statistical heterogeneity can be partially offset by pooling similar cells to obtain a more realistic, averaged view of the transcriptomic profile of the cell type. However, this pooling is usually based on cell clustering relying on these same under-sampled transcriptomes and may be biased in different ways across species. As a result, matching fine-grained cell populations across species is non-trivial and often contends with technical uncertainty and error in how cell populations have been defined in each species separately. Another approach consists in integrating single-cell transcriptomes of both species in a common reference space, and then clustering into subtypes containing orthologous cells from both species (Liu *et al.*, 2020). Unfortunately, this methodology frequently yields poor results, where biological signal is largely obscured by technical noise and data distortion due to the integration procedure.

Identifying orthologous cell types and potentially functional changes in their transcriptomic programs is central to several of our research projects. As such, we are benchmarking different strategies to integrate and correlate gene expression data from single cells across two or more species. These methods will be a cornerstone of data analysis for our project on the evolution of menstruation. However, we have also started evaluating these methods on another project investigating the orthology of spinal cord neural cell types in human and mouse, in collaboration with Steven Knafo (Hôpital Kremlin-Bicêtre) and Julien Bouvier (Institut de Neurosciences, Paris Saclay). Spinal cord neurons are responsible for transmitting sensory-motor influxes between the limbs and the brain, and for feedback loops controlling reflex limb motion. Spinal cord contains an array of neuronal populations with excitatory and inhibitory functions, motoneurons, interneurons, and more refined functional subcategories characterized by their axonal projections and localization in the spine (Osseward *et al.*, 2021). My collaborators are interested in developing models and treatments for spinal cord injuries, where the communication between brain and limb is severed. Studying these processes is mostly done in mouse models; however, the homology of neural cell types in the human and mouse spinal cords are not characterized, limiting the applicability

to translational medicine. We have generated single-cell transcriptomes from human and mouse spinal cords in order to identify categories of neurons that purportedly fulfill the same function, and are currently exploring spatial transcriptomics and optogenetics cell tracing to complement this transcriptomic data. Our objective is to obtain an atlas of orthologous spinal neural cell types in human and mouse, based on gene expression, spatial localization and axonal projection, all of which contribute to neuronal function.

References

Allendorf, F.W. *et al.* (2015) 'Effects of Crossovers Between Homeologs on Inheritance and Population Genomics in Polyploid-Derived Salmonid Fishes', *Journal of Heredity*, 106(3), pp. 217–227. Available at: https://doi.org/10.1093/jhered/esv015.

Andersson, R. *et al.* (2014) 'An atlas of active enhancers across human cell types and tissues', *Nature*, 507(7493), pp. 455–461. Available at: https://doi.org/10.1038/nature12787.

Andersson, R. and Sandelin, A. (2020) 'Determinants of enhancer and promoter activities of regulatory elements', *Nature Reviews Genetics*, 21(2), pp. 71–87. Available at: https://doi.org/10.1038/s41576-019-0173-8.

Arnold, C.D. *et al.* (2014) 'Quantitative genome-wide enhancer activity maps for five Drosophila species show functional enhancer conservation and turnover during cis-regulatory evolution', *Nature Genetics*, 46(7), pp. 685–692. Available at: https://doi.org/10.1038/ng.3009.

Becker, T.S. and Lenhard, B. (2007) 'The random versus fragile breakage models of chromosome evolution: a matter of resolution', *Molecular Genetics and Genomics*, 278(5), pp. 487–491.

Bellofiore, N. *et al.* (2017) 'First evidence of a menstruating rodent: the spiny mouse (Acomys cahirinus)', *American Journal of Obstetrics and Gynecology*, 216(1), p. 40.e1-40.e11. Available at: https://doi.org/10.1016/j.ajog.2016.07.041.

Benton, M.L. *et al.* (2021) 'The influence of evolutionary history on human health and disease', *Nature Reviews Genetics*, 22(5), pp. 269–283. Available at: https://doi.org/10.1038/s41576-020-00305-9.

Bi, X. *et al.* (2021) 'Tracing the genetic footprints of vertebrate landing in non-teleost ray-finned fishes', *Cell* [Preprint]. Available at: https://doi.org/10.1016/j.cell.2021.01.046.

Boretto, M. *et al.* (2017) 'Development of organoids from mouse and human endometrium showing endometrial epithelium physiology and long-term expandability', *Development*, 144(10), pp. 1775–1786. Available at: https://doi.org/10.1242/dev.148478.

Bougie, O. *et al.* (2019) 'Influence of race/ethnicity on prevalence and presentation of endometriosis: a systematic review and meta-analysis', *BJOG: an international journal of obstetrics and gynaecology*, 126(9), pp. 1104–1115. Available at: https://doi.org/10.1111/1471-0528.15692.

Braasch, I. *et al.* (2016) 'The spotted gar genome illuminates vertebrate evolution and facilitates human-teleost comparisons', *Nature Genetics*, 48(4), pp. 427–437. Available at: https://doi.org/10.1038/ng.3526.

Brawand, D. *et al.* (2011) 'The evolution of gene expression levels in mammalian organs', *Nature*, 478(7369), pp. 343–348. Available at: https://doi.org/10.1038/nature10532.

Breschi, A., Gingeras, T.R. and Guigó, R. (2017) 'Comparative transcriptomics in human and mouse', *Nature Reviews Genetics*, 18(7), pp. 425–440. Available at: https://doi.org/10.1038/nrg.2017.19.

Buniello, A. *et al.* (2019) 'The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019', *Nucleic Acids Research*, 47(D1), pp. D1005–D1012. Available at: https://doi.org/10.1093/nar/gky1120.

Bycroft, C. *et al.* (2018) 'The UK Biobank resource with deep phenotyping and genomic data.', *Nature*, 562(7726), pp. 203–209. Available at: https://doi.org/10.1038/s41586-018-0579-z.

Byrne, K.P. and Wolfe, K.H. (2005) 'The Yeast Gene Order Browser: Combining curated homology and syntenic context reveals gene fate in polyploid species', *Genome Research*, 15(10), pp. 1456–1461. Available at: https://doi.org/10.1101/gr.3672305.

Cardoso-Moreira, M. *et al.* (2019) 'Gene expression across mammalian organ development', *Nature*, 571(7766), pp. 505–509. Available at: https://doi.org/10.1038/s41586-019-1338-5.

Catchen, J.M., Conery, J.S. and Postlethwait, J.H. (2009) 'Automated identification of conserved synteny after whole-genome duplication', *Genome Research*, 19(8), pp. 1497–1505. Available at: https://doi.org/10.1101/gr.090480.108.

Chapron, C. *et al.* (2019) 'Rethinking mechanisms, diagnosis and management of endometriosis', *Nature Reviews Endocrinology*, 15(11), pp. 666–682. Available at: https://doi.org/10.1038/s41574-019-

0245-z.

Cheng, F. *et al.* (2018) 'Gene retention, fractionation and subgenome differences in polyploid plants', *Nature Plants*, 4(5), pp. 258–268. Available at: https://doi.org/10.1038/s41477-018-0136-7.

Chuong, E.B. *et al.* (2013) 'Endogenous retroviruses function as species-specific enhancer elements in the placenta', *Nature Genetics*, 45(3), pp. 325–329. Available at: https://doi.org/10.1038/ng.2553.

Cindrova-Davies, T. *et al.* (2021) 'Menstrual flow as a non-invasive source of endometrial organoids', *Communications Biology*, 4(1), pp. 1–8. Available at: https://doi.org/10.1038/s42003-021-02194-y.

Coghlan, A. *et al.* (2005) 'Chromosome evolution in eukaryotes: a multi-kingdom perspective', *Trends in Genetics*, 21(12), pp. 673–682. Available at: https://doi.org/10.1016/j.tig.2005.09.009.

Conant, G.C. (2020) 'The lasting after-effects of an ancient polyploidy on the genomes of teleosts', *PLOS ONE*, 15(4), p. e0231356. Available at: https://doi.org/10.1371/journal.pone.0231356.

Conant, G.C. and Wolfe, K.H. (2008) 'Turning a hobby into a job: How duplicated genes find new functions', *Nature Reviews Genetics*, 9(12), pp. 938–950. Available at: https://doi.org/10.1038/nrg2482.

Cotney, J. *et al.* (2013) 'The Evolution of Lineage-Specific Regulatory Activities in the Human Embryonic Limb', *Cell*, 154(1), pp. 185–196. Available at: https://doi.org/10.1016/j.cell.2013.05.056.

Degner, J.F. *et al.* (2012) 'DNase I sensitivity QTLs are a major determinant of human expression variation', *Nature*, 482(7385), pp. 390–394. Available at: https://doi.org/10.1038/nature10808.

Dehal, P. and Boore, J.L. (2005) 'Two Rounds of Whole Genome Duplication in the Ancestral Vertebrate', *PLoS Biol*, 3(10), p. e314. Available at: https://doi.org/10.1371/journal.pbio.0030314.

Dornburg, A. and Near, T.J. (2021) 'The Emerging Phylogenetic Perspective on the Evolution of Actinopterygian Fishes', *Annual Review of Ecology, Evolution, and Systematics*, 52(1), pp. 427–452. Available at: https://doi.org/10.1146/annurev-ecolsys-122120-122554.

Duchemin, W. *et al.* (2017) 'DeCoSTAR: Reconstructing the Ancestral Organization of Genes or Genomes Using Reconciled Phylogenies', *Genome Biology and Evolution*, 9(5), pp. 1312–1319. Available at: https://doi.org/10.1093/gbe/evx069.

Dukler, N., Huang, Y.-F. and Siepel, A. (2020) 'Phylogenetic Modeling of Regulatory Element Turnover Based on Epigenomic Data', *Molecular Biology and Evolution*, 37(7), pp. 2137–2152. Available at: https://doi.org/10.1093/molbev/msaa073.

Dunn, C.W. *et al.* (2018) 'Pairwise comparisons across species are problematic when analyzing functional genomic data', *Proceedings of the National Academy of Sciences*, 115(3), pp. E409–E417. Available at: https://doi.org/10.1073/pnas.1707515115.

Dunn, C.W., Luo, X. and Wu, Z. (2013) 'Phylogenetic Analysis of Gene Expression', *Integrative and Comparative Biology*, 53(5), pp. 847–856. Available at: https://doi.org/10.1093/icb/ict068.

Dunselman, G. a. J. *et al.* (2014) 'ESHRE guideline: management of women with endometriosis', *Human Reproduction*, 29(3), pp. 400–412. Available at: https://doi.org/10.1093/humrep/det457.

Emera, D. *et al.* (2016) 'Origin and evolution of developmental enhancers in the mammalian neocortex', *Proceedings of the National Academy of Sciences*, 113(19), pp. E2617–E2626. Available at: https://doi.org/10.1073/pnas.1603718113.

Emera, D., Romero, R. and Wagner, G. (2012) 'The evolution of menstruation: A new model for genetic assimilation', *BioEssays*, 34(1), pp. 26–35. Available at: https://doi.org/10.1002/bies.201100099.

ENCODE Project Consortium (2012) 'An integrated encyclopedia of DNA elements in the human genome', *Nature*, 489(7414), pp. 57–74. Available at: https://doi.org/10.1038/nature11247.

Engström, P.G. *et al.* (2007) 'Genomic regulatory blocks underlie extensive microsynteny conservation in insects', *Genome research*, 17(12), pp. 1898–1908.

Evans, J. *et al.* (2016) 'Fertile ground: human endometrial programming and lessons in health and disease', *Nature Reviews Endocrinology*, 12(11), pp. 654–667. Available at: https://doi.org/10.1038/nrendo.2016.116.

Farré, M. *et al.* (2016) 'Novel Insights into Chromosome Evolution in Birds, Archosaurs, and Reptiles', *Genome Biology and Evolution*, 8(8), pp. 2442–2451. Available at: https://doi.org/10.1093/gbe/evw166.

Ferguson-Smith, M.A. and Trifonov, V. (2007) 'Mammalian karyotype evolution', *Nature Reviews Genetics*, 8(12), pp. 950–962. Available at: https://doi.org/10.1038/nrg2199.

Fort, A. *et al.* (2014) 'Deep transcriptome profiling of mammalian stem cells supports a regulatory role for retrotransposons in pluripotency maintenance', *Nature Genetics*, 46(6), pp. 558–566. Available at: https://doi.org/10.1038/ng.2965.

Fueyo, R. *et al.* (2022) 'Roles of transposable elements in the regulation of mammalian transcription', *Nature Reviews Molecular Cell Biology*, pp. 1–17. Available at: https://doi.org/10.1038/s41580-022-00457-y.

Garsmeur, O. *et al.* (2014) 'Two Evolutionarily Distinct Classes of Paleopolyploidy', *Molecular Biology and Evolution*, 31(2), pp. 448–454. Available at: https://doi.org/10.1093/molbev/mst230.

Gasperini, M., Tome, J.M. and Shendure, J. (2020) 'Towards a comprehensive catalogue of validated and target-linked human enhancers', *Nature Reviews Genetics*, pp. 1–19. Available at: https://doi.org/10.1038/s41576-019-0209-0.

Geirsdottir, L. *et al.* (2019) 'Cross-Species Single-Cell Analysis Reveals Divergence of the Primate Microglia Program', *Cell*, 179(7), pp. 1609-1622.e16. Available at: https://doi.org/10.1016/j.cell.2019.11.010.

Giuliani, E., As-Sanie, S. and Marsh, E.E. (2020) 'Epidemiology and management of uterine fibroids', *International Journal of Gynaecology and Obstetrics: The Official Organ of the International Federation of Gynaecology and Obstetrics*, 149(1), pp. 3–9. Available at: https://doi.org/10.1002/ijgo.13102.

Griffith, O.W. *et al.* (2017) 'Embryo implantation evolved from an ancestral inflammatory attachment reaction', *Proceedings of the National Academy of Sciences*, p. 201701129. Available at: https://doi.org/10.1073/pnas.1701129114.

Gundappa, M.K. *et al.* (2021) 'Genome-wide reconstruction of rediploidization following autopolyploidization across one hundred million years of salmonid evolution', *Molecular Biology and Evolution*, p. msab310. Available at: https://doi.org/10.1093/molbev/msab310.

Heinz, S. *et al.* (2015) 'The selection and function of cell type-specific enhancers', *Nature Reviews Molecular Cell Biology*, 16(3), pp. 144–154. Available at: https://doi.org/10.1038/nrm3949.

Hill, M.S., Vande Zande, P. and Wittkopp, P.J. (2020) 'Molecular and evolutionary processes generating variation in gene expression', *Nature Reviews Genetics*, pp. 1–13. Available at: https://doi.org/10.1038/s41576-020-00304-w.

Hoeppner, M.P. *et al.* (2018) 'An Evaluation of Function of Multicopy Noncoding RNAs in Mammals Using ENCODE/FANTOM Data and Comparative Genomics', *Molecular Biology and Evolution*, 35(6), pp. 1451–1462. Available at: https://doi.org/10.1093/molbev/msy046.

Hudson, N. (2022) 'The missed disease? Endometriosis as an example of "undone science"', *Reproductive Biomedicine & Society Online*, 14, pp. 20–27. Available at: https://doi.org/10.1016/j.rbms.2021.07.003.

Jaillon, O. *et al.* (2004) 'Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype', *Nature*, 431(7011), pp. 946–957. Available at: https://doi.org/10.1038/nature03025.

Jiao, W.-B. and Schneeberger, K. (2020) 'Chromosome-level assemblies of multiple Arabidopsis genomes reveal hotspots of rearrangements with altered evolutionary dynamics', *Nature Communications*, 11(1), pp. 1–10. Available at: https://doi.org/10.1038/s41467-020-14779-y.

Jones, B.R. *et al.* (2012) 'ANGES: reconstructing ANcestral GEnomeS maps', *Bioinformatics*, 28(18), pp. 2388–2390. Available at: https://doi.org/10.1093/bioinformatics/bts457.

Kapusta, A. *et al.* (2013) 'Transposable Elements Are Major Contributors to the Origin, Diversification, and Regulation of Vertebrate Long Noncoding RNAs', *PLOS Genetics*, 9(4), p. e1003470. Available at: https://doi.org/10.1371/journal.pgen.1003470.

Kellis, M., Birren, B.W. and Lander, E.S. (2004) 'Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae', *Nature*, 428(6983), pp. 617–624. Available at: https://doi.org/10.1038/nature02424.

Kikuta, H. *et al.* (2007) 'Genomic regulatory blocks encompass multiple neighboring genes and maintain

conserved synteny in vertebrates', *Genome research*, 17(5), pp. 545–555.

Kim, E.B. *et al.* (2011) 'Genome sequencing reveals insights into physiology and longevity of the naked mole rat', *Nature*, 479(7372), pp. 223–227. Available at: https://doi.org/10.1038/nature10533.

Kim, J. *et al.* (2017) 'Reconstruction and evolutionary history of eutherian chromosomes', *Proceedings of the National Academy of Sciences*, 114(27), pp. E5379–E5388. Available at: https://doi.org/10.1073/pnas.1702012114.

Kloosterman, W.P. *et al.* (2015) 'Characteristics of de novo structural changes in the human genome', *Genome Research*, 25(6), pp. 792–801. Available at: https://doi.org/10.1101/gr.185041.114.

Kratochwil, C.F. and Meyer, A. (2015) 'Evolution: Tinkering within Gene Regulatory Landscapes', *Current Biology*, 25(7), pp. R285–R288. Available at: https://doi.org/10.1016/j.cub.2015.02.051.

Krjutškov, K. *et al.* (2016) 'Single-cell transcriptome analysis of endometrial tissue', *Human Reproduction*, 31(4), pp. 844–853. Available at: https://doi.org/10.1093/humrep/dew008.

Kronenberg, Z.N. *et al.* (2018) 'High-resolution comparative analysis of great ape genomes', *Science*, 360(6393), p. eaar6343. Available at: https://doi.org/10.1126/science.aar6343.

Kurki, M.I. *et al.* (2022) 'FinnGen: Unique genetic insights from combining isolated population and national health register data'. medRxiv, p. 2022.03.03.22271360. Available at: https://doi.org/10.1101/2022.03.03.22271360.

Kvon, E.Z. *et al.* (2016) 'Progressive Loss of Function in a Limb Enhancer during Snake Evolution', *Cell*, 167(3), pp. 633-642.e11. Available at: https://doi.org/10.1016/j.cell.2016.09.028.

Laux-Biehlmann, A., d'Hooghe, T. and Zollner, T.M. (2015) 'Menstruation pulls the trigger for inflammation and pain in endometriosis', *Trends in Pharmacological Sciences*, 36(5), pp. 270–276. Available at: https://doi.org/10.1016/j.tips.2015.03.004.

Laval, G. *et al.* (2021) 'Sporadic occurrence of recent selective sweeps from standing variation in humans as revealed by an approximate Bayesian computation approach', *Genetics*, 219(4), p. iyab161. Available at: https://doi.org/10.1093/genetics/iyab161.

Lee, J.A., Carvalho, C.M.B. and Lupski, J.R. (2007) 'A DNA Replication Mechanism for Generating Nonrecurrent Rearrangements Associated with Genomic Disorders', *Cell*, 131(7), pp. 1235–1247. Available at: https://doi.org/10.1016/j.cell.2007.11.037.

Lemaitre, C. *et al.* (2009) 'Analysis of fine-scale mammalian evolutionary breakpoints provides new insight into their relation to genome organisation', *BMC genomics*, 10(1), p. 335.

Li, J.-T. *et al.* (2021) 'Parallel subgenome structure and divergent expression evolution of allo-tetraploid common carp and goldfish', *Nature Genetics*, 53(10), pp. 1–11. Available at: https://doi.org/10.1038/s41588-021-00933-9.

Li, W., Notani, D. and Rosenfeld, M.G. (2016) 'Enhancers as non-coding RNA transcription units: recent insights and future perspectives', *Nature Reviews Genetics*, 17(4), pp. 207–223. Available at: https://doi.org/10.1038/nrg.2016.4.

Lien, S. *et al.* (2016) 'The Atlantic salmon genome provides insights into rediploidization', *Nature*, 533(7602), pp. 200–205. Available at: https://doi.org/10.1038/nature17164.

Lieschke, G.J. and Currie, P.D. (2007) 'Animal models of human disease: zebrafish swim into view', *Nature Reviews. Genetics*, 8(5), pp. 353–367. Available at: https://doi.org/10.1038/nrg2091.

Liò, P. and Goldman, N. (1998) 'Models of Molecular Evolution and Phylogeny', *Genome Research*, 8(12), pp. 1233–1244. Available at: https://doi.org/10.1101/gr.8.12.1233.

Liu, J. *et al.* (2020) 'Jointly defining cell types from multiple single-cell datasets using LIGER', *Nature Protocols*, 15(11), pp. 3632–3662. Available at: https://doi.org/10.1038/s41596-020-0391-8.

Lucas, E.S. *et al.* (2018) 'Reconstruction of the Decidual Pathways in Human Endometrial Cells Using Single-Cell RNA-Seq', *bioRxiv*, p. 368829. Available at: https://doi.org/10.1101/368829.

Lupski, J.R. and Stankiewicz, P. (2005) 'Genomic Disorders: Molecular Mechanisms for Rearrangements and Conveyed Phenotypes', *PLoS Genet*, 1(6), p. e49. Available at: https://doi.org/10.1371/journal.pgen.0010049.

Lynch, V.J. *et al.* (2011) 'Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals', *Nature Genetics*, 43(11), pp. 1154–1159. Available at: https://doi.org/10.1038/ng.917.

Ma, J. *et al.* (2006) 'Reconstructing contiguous regions of an ancestral genome', *Genome Research*, 16(12), pp. 1557–1565. Available at: https://doi.org/10.1101/gr.5383506.

Macqueen, D.J. and Johnston, I.A. (2014) 'A well-constrained estimate for the timing of the salmonid whole genome duplication reveals major decoupling from species diversification', *Proceedings of the Royal Society B: Biological Sciences*, 281(1778), p. 20132881. Available at: https://doi.org/10.1098/rspb.2013.2881.

Makino, T. and McLysaght, A. (2010) 'Ohnologs in the human genome are dosage balanced and frequently associated with disease', *Proceedings of the National Academy of Sciences*, 107(20), pp. 9270–9274. Available at: https://doi.org/10.1073/pnas.0914697107.

Manceau, M., Lambert, A. and Morlon, H. (2017) 'A Unifying Comparative Phylogenetic Framework Including Traits Coevolving Across Interacting Lineages', *Systematic Biology*, 66(4), pp. 551–568. Available at: https://doi.org/10.1093/sysbio/syw115.

Martin, K.J. and Holland, P.W.H. (2014) 'Enigmatic Orthology Relationships between Hox Clusters of the African Butterfly Fish and Other Teleosts Following Ancient Whole-Genome Duplication', *Molecular Biology and Evolution*, 31(10), pp. 2592–2611. Available at: https://doi.org/10.1093/molbev/msu202.

Mason, A.S. and Wendel, J.F. (2020) 'Homoeologous Exchanges, Segmental Allopolyploidy, and Polyploid Genome Evolution', *Frontiers in Genetics*, 11. Available at: https://doi.org/10.3389/fgene.2020.01014.

McClintock, B. (1984) 'The significance of responses of the genome to challenge', *Science (New York, N.Y.)*, 226(4676), pp. 792–801. Available at: https://doi.org/10.1126/science.15739260.

Moriyama, Y. *et al.* (2016) 'Evolution of the fish heart by sub/neofunctionalization of an *elastin* gene', *Nature Communications*, 7, p. 10397. Available at: https://doi.org/10.1038/ncomms10397.

Nakatani, Y. and McLysaght, A. (2017) 'Genomes as documents of evolutionary history: a probabilistic macrosynteny model for the reconstruction of ancestral genomes', *Bioinformatics*, 33(14), pp. i369–i378. Available at: https://doi.org/10.1093/bioinformatics/btx259.

Near, T.J. *et al.* (2012) 'Resolution of ray-finned fish phylogeny and timing of diversification', *Proceedings of the National Academy of Sciences*, 109(34), pp. 13698–13703. Available at: https://doi.org/10.1073/pnas.1206625109.

Nelson, J.S., Grande, T.C. and Wilson, M.V.H. (2016) *Fishes of the World*. John Wiley & Sons.

Odom, D.T. *et al.* (2007) 'Tissue-specific transcriptional regulation has diverged significantly between human and mouse', *Nature Genetics*, 39(6), pp. 730–732. Available at: https://doi.org/10.1038/ng2047.

Osseward, P.J. *et al.* (2021) 'Conserved genetic signatures parcellate cardinal spinal neuron classes into local and projection subsets', *Science*, 372(6540), pp. 385–393. Available at: https://doi.org/10.1126/science.abe0690.

Peng, Q., Pevzner, P.A. and Tesler, G. (2006) 'The fragile breakage versus random breakage models of chromosome evolution', *PLoS computational biology*, 2(2), p. e14.

Pevzner, P. and Tesler, G. (2003) 'Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution', *Proceedings of the National Academy of Sciences*, 100(13), pp. 7672–7677.

Prescott, S.L. *et al.* (no date) 'Enhancer Divergence and cis-Regulatory Evolution in the Human and Chimp Neural Crest', *Cell* [Preprint]. Available at: https://doi.org/10.1016/j.cell.2015.08.036.

Price, P.D. *et al.* (2022) 'Detecting signatures of selection on gene expression', *Nature Ecology & Evolution*, pp. 1–11. Available at: https://doi.org/10.1038/s41559-022-01761-8.

Reilly, S.K. *et al.* (2015) 'Evolutionary changes in promoter and enhancer activity during human corticogenesis', *Science*, 347(6226), pp. 1155–1159. Available at: https://doi.org/10.1126/science.1260943.

Reilly, S.K. and Noonan, J.P. (2016) 'Evolution of Gene Regulation in Humans', *Annual Review of*

Genomics and Human Genetics, 17(1), pp. 45–67. Available at: https://doi.org/10.1146/annurev-genom-090314-045935.

Rhie, A. *et al.* (2021) 'Towards complete and error-free genome assemblies of all vertebrate species', *Nature*, 592(7856), pp. 737–746. Available at: https://doi.org/10.1038/s41586-021-03451-0.

Roadmap Epigenomics Consortium *et al.* (2015) 'Integrative analysis of 111 reference human epigenomes', *Nature*, 518(7539), pp. 317–330. Available at: https://doi.org/10.1038/nature14248.

Robertson, F.M. *et al.* (2017) 'Lineage-specific rediploidization is a mechanism to explain time-lags between genome duplication and evolutionary diversification', *Genome Biology*, 18, p. 111. Available at: https://doi.org/10.1186/s13059-017-1241-z.

Rohlfs, R.V., Harrigan, P. and Nielsen, R. (2014) 'Modeling Gene Expression Evolution with an Extended Ornstein–Uhlenbeck Process Accounting for Within-Species Variation', *Molecular Biology and Evolution*, 31(1), pp. 201–211. Available at: https://doi.org/10.1093/molbev/mst190.

Rohlfs, R.V. and Nielsen, R. (2015) 'Phylogenetic ANOVA: The Expression Variance and Evolution Model for Quantitative Trait Evolution', *Systematic Biology*, 64(5), pp. 695–708. Available at: https://doi.org/10.1093/sysbio/syv042.

Romero, I.G., Ruvinsky, I. and Gilad, Y. (2012) 'Comparative studies of gene expression and the evolution of gene regulation', *Nature Reviews Genetics*, 13(7), pp. 505–516. Available at: https://doi.org/10.1038/nrg3229.

Sallan, L.C. (2014) 'Major issues in the origins of ray-finned fish (Actinopterygii) biodiversity', *Biological Reviews*, 89(4), pp. 950–971. Available at: https://doi.org/10.1111/brv.12086.

Sankoff, D. (2003) 'Rearrangements and chromosomal evolution', *Current Opinion in Genetics & Development*, 13(6), pp. 583–587. Available at: https://doi.org/10.1016/j.gde.2003.10.006.

Sankoff, D. and Trinh, P. (2005) 'Chromosomal Breakpoint Reuse in Genome Sequence Rearrangement', *Journal of Computational Biology*, 12(6), pp. 812–821. Available at: https://doi.org/10.1089/cmb.2005.12.812.

Saunders, P.T.K. and Horne, A.W. (2021) 'Endometriosis: Etiology, pathobiology, and therapeutic prospects', *Cell*, 184(11), pp. 2807–2824. Available at: https://doi.org/10.1016/j.cell.2021.04.041.

Schmidt, D. *et al.* (2010) 'Five-Vertebrate ChIP-seq Reveals the Evolutionary Dynamics of Transcription Factor Binding', *Science*, 328(5981), pp. 1036–1040. Available at: https://doi.org/10.1126/science.1186176.

Session, A.M. *et al.* (2016) 'Genome evolution in the allotetraploid frog *Xenopus laevis*', *Nature*, 538(7625), p. nature19840. Available at: https://doi.org/10.1038/nature19840.

Shafer, M.E.R. (2019) 'Cross-Species Analysis of Single-Cell Transcriptomic Data', *Frontiers in Cell and Developmental Biology*, 7. Available at: https://www.frontiersin.org/article/10.3389/fcell.2019.00175 (Accessed: 21 January 2022).

Shapira, S.K. (1998) 'An update on chromosome deletion and microdeletion syndromes', *Current opinion in pediatrics*, 10(6), pp. 622–627. Available at: https://doi.org/10.1097/00008480-199810060-00015.

Shibata, Y. *et al.* (2012) 'Extensive Evolutionary Changes in Regulatory Element Activity during Human Origins Are Associated with Altered Gene Expression and Positive Selection', *PLoS Genet*, 8(6), p. e1002789. Available at: https://doi.org/10.1371/journal.pgen.1002789.

Shlyueva, D., Stampfel, G. and Stark, A. (2014) 'Transcriptional enhancers: from properties to genome-wide predictions', *Nature Reviews Genetics*, 15(4), pp. 272–286. Available at: https://doi.org/10.1038/nrg3682.

Soltis, D.E., Visger, C.J. and Soltis, P.S. (2014) 'The polyploidy revolution then…and now: Stebbins revisited', *American Journal of Botany*, 101(7), pp. 1057–1078. Available at: https://doi.org/10.3732/ajb.1400178.

Soumillon, M. *et al.* (2013) 'Cellular Source and Mechanisms of High Transcriptome Complexity in the Mammalian Testis', *Cell Reports*, 3(6), pp. 2179–2190. Available at: https://doi.org/10.1016/j.celrep.2013.05.031.

Soutoglou, E. *et al.* (2007) 'Positional stability of single double-strand breaks in mammalian cells', *Nature Cell Biology*, 9(6), pp. 675–682. Available at: https://doi.org/10.1038/ncb1591.

Spielmann, M., Lupiáñez, D.G. and Mundlos, S. (2018) 'Structural variation in the 3D genome', *Nature Reviews Genetics*, 19(7), pp. 453–467. Available at: https://doi.org/10.1038/s41576-018-0007-0.

Stebbins, G.L. (1947) 'Types of Polyploids: Their Classification and Significance', in M. Demerec (ed.) *Advances in Genetics*. Academic Press, pp. 403–429. Available at: https://doi.org/10.1016/S0065-2660(08)60490-3.

Stegle, O., Teichmann, S.A. and Marioni, J.C. (2015) 'Computational and analytical challenges in single-cell transcriptomics', *Nature Reviews Genetics*, 16(3), pp. 133–145. Available at: https://doi.org/10.1038/nrg3833.

Strassmann, B.I. (1996) 'The Evolution of Endometrial Cycles and Menstruation', *The Quarterly Review of Biology*, 71(2), pp. 181–220.

Sudmant, P.H. *et al.* (2015) 'An integrated map of structural variation in 2,504 human genomes', *Nature*, 526(7571), pp. 75–81. Available at: https://doi.org/10.1038/nature15394.

Takezaki, N. (2021) 'Resolving the Early Divergence Pattern of Teleost Fish Using Genome-Scale Data', *Genome Biology and Evolution*, 13(5), p. evab052. Available at: https://doi.org/10.1093/gbe/evab052.

Thorgaard, G.H. *et al.* (2002) 'Status and opportunities for genomics research with rainbow trout', *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology*, 133(4), pp. 609–646. Available at: https://doi.org/10.1016/S1096-4959(02)00167-7.

Thybert, D. *et al.* (2018) 'Repeat associated mechanisms of genome evolution and function revealed by the Mus caroli and Mus pahari genomes', *Genome Research*, 28(4), pp. 448–459. Available at: https://doi.org/10.1101/gr.234096.117.

Trizzino, M. *et al.* (2017) 'Transposable elements are the primary source of novelty in primate gene regulation', *Genome Research* [Preprint]. Available at: https://doi.org/10.1101/gr.218149.116.

Turco, M.Y. *et al.* (2017) 'Long-term, hormone-responsive organoid cultures of human endometrium in a chemically defined medium.', *Nature cell biology*, 19(5), pp. 568–577. Available at: https://doi.org/10.1038/ncb3516.

Van de Peer, Y., Mizrachi, E. and Marchal, K. (2017) 'The evolutionary significance of polyploidy', *Nature Reviews Genetics*, 18(7), pp. 411–424. Available at: https://doi.org/10.1038/nrg.2017.26.

Veitia, R.A. (2004) 'Gene Dosage Balance in Cellular Pathways: Implications for Dominance and Gene Duplicability', *Genetics*, 168(1), pp. 569–574. Available at: https://doi.org/10.1534/genetics.104.029785.

Veitia, R.A. and Birchler, J.A. (2021) 'Gene-dosage issues: a recurrent theme in whole genome duplication events', *Trends in Genetics* [Preprint]. Available at: https://doi.org/10.1016/j.tig.2021.06.006.

Villar, D., Flicek, P. and Odom, D.T. (2014) 'Evolution of transcription factor binding in metazoans — mechanisms and functional implications', *Nature Reviews Genetics*, advance online publication. Available at: https://doi.org/10.1038/nrg3481.

Von Grotthuss, M., Ashburner, M. and Ranz, J.M. (2010) 'Fragile regions and not functional constraints predominate in shaping gene organization in the genus Drosophila', *Genome research*, 20(8), pp. 1084–1096.

Wang, W. *et al.* (2018) 'Single cell RNAseq provides a molecular and cellular cartography of changes to the human endometrium through the menstrual cycle', *bioRxiv*, p. 350538. Available at: https://doi.org/10.1101/350538.

Whelan, S., Liò, P. and Goldman, N. (2001) 'Molecular phylogenetics: state-of-the-art methods for looking into the past', *Trends in Genetics*, 17(5), pp. 262–272. Available at: https://doi.org/10.1016/S0168-9525(01)02272-7.

Xiao, S. *et al.* (2012) 'Comparative Epigenomic Annotation of Regulatory DNA', *Cell*, 149(6), pp. 1381–1392. Available at: https://doi.org/10.1016/j.cell.2012.04.029.

Yang, Y. *et al.* (2018) 'Continuous-Trait Probabilistic Model for Comparing Multi-species Functional Genomic Data', *Cell Systems*, 7(2), pp. 208-218.e11. Available at: https://doi.org/10.1016/j.cels.2018.05.022.

Yang, Z. (1997) ‘PAML: a program package for phylogenetic analysis by maximum likelihood’, *Computer applications in the biosciences: CABIOS*, 13(5), pp. 555–556.

Yin, Y. *et al.* (2021) ‘Molecular mechanisms and topological consequences of drastic chromosomal rearrangements of muntjac deer’, *Nature Communications*, 12(1), p. 6858. Available at: https://doi.org/10.1038/s41467-021-27091-0.

Yokoyama, K.D., Zhang, Y. and Ma, J. (2014) ‘Tracing the Evolution of Lineage-Specific Transcription Factor Binding Sites in a Birth-Death Framework’, *PLoS Comput Biol*, 10(8), p. e1003771. Available at: https://doi.org/10.1371/journal.pcbi.1003771.

Zhang, Y. *et al.* (2012) ‘Spatial Organization of the Mouse Genome and Its Role in Recurrent Chromosomal Translocations’, *Cell*, 148(5), pp. 908–921. Available at: https://doi.org/10.1016/j.cell.2012.02.002.

Zhou, X. *et al.* (2014) ‘Epigenetic modifications are associated with inter-species gene expression variation in primates’, *Genome Biology*, 15(12), p. 547. Available at: https://doi.org/10.1186/s13059-014-0547-3.

Zins, M., Goldberg, M., and CONSTANCES team (2015) ‘The French CONSTANCES population-based cohort: design, inclusion and follow-up’, *European Journal of Epidemiology*, 30(12), pp. 1317–1328. Available at: https://doi.org/10.1007/s10654-015-0096-4.

Zwaenepoel, A. and Van de Peer, Y. (2019) ‘Inference of Ancient Whole-Genome Duplications and the Evolution of Gene Duplication and Loss Rates’, *Molecular Biology and Evolution*, 36(7), pp. 1384–1404. Available at: https://doi.org/10.1093/molbev/msz088.

Acknowledgements

It has been a pleasure and a privilege to interact with so many fantastic people over the years, who made all of this work possible.

First, I want to thank the members of the jury who agreed to come from – sometimes – far away to review and discuss this body of work: Dan Macqueen, Judith Zaugg, Gilles Fischer, Odile Lecompte, Marie Sémon and Olivier Lespinet. I am deeply grateful to have you in this HDR jury.

I will follow by thanking my scientific mentors, who supported me through thick and thin and helped me grow as a scientist and a person. Hugues Roest Crollius made a gamble on a PhD student with no prior bioinformatics experience, and then welcomed me again in the lab as a permanent scientist, student cheerleader and resident escape-room enthusiast. It has been a pleasure to be in his lab every day, and who knows, maybe someday I'll be back (again). My years in Paul Flicek's lab have been some of my happiest: the freedom, creativity and emulation I discovered there, in collaboration with Duncan Odom and others at EBI and Cambridge, were fundamental to my growth. Paul and Duncan's continued support and trust keep me going when I need a confidence boost, and I am grateful to both of them.

I also want to thank my collaborators, without whom much of this work would not exist: Diego Villar, my partner in crime for all things functional genomics over the past ten years, whose experience remains a touchstone whenever I have an experimental design question in the lab; Yann Guiguen, whose enthusiasm for fish comparative genomics has laid the ground for large parts of my PhD and then junior scientist projects, as well as all the members of Team GenoFish; and my more recent collaborators, who have opened new horizons over the past few years: Ludivine Doridot, Steven Knafo, Angela Goncalves, Lyne Fellmann, Jérémy Terrien and all the members of the CNRS Primatology station. A shout-out as well to the members of the New PI Slack group, whose collective wisdom and willingness to help has made my transition to group leader infinitely easier.

Further thanks to all past and current members of the Dyogen Lab at ENS: Alexandra, Yves, Céline, Lambert, Guillaume, Matthieu, Christine, François and everyone else – a great team, and many laughs around evening games of cards and much needed coffee breaks. I want to dedicate special thanks to Elise Parey, who has been the best possible first PhD student one could wish for, and has thoroughly spoiled me as a supervisor. Many thanks also go to our colleagues at ENS, who helped me through the ups and downs of research life, especially Morgane Thomas Chollier and Laura Cantini.

My biggest thanks go to the members of my team: nothing in the lab would happen without them, and I am most grateful to them for putting their trust and energy into such a junior baby group. Thank you to Axelle, Gosia, Bruno, Eulalie, Anthony and Lucie, whose hard work is already transforming into new and unexpected research directions – it is an immense pleasure to work alongside you all. Special thanks to Marie-Claire, whose administrative magic has been invaluable over the past year to set up and run the group. I also want to thank our new colleagues at Pasteur, especially Eduardo Rocha, director of the Genomes and Genetics department, who has been extremely supportive, as well as Lluis Quintana-Murci, my senior mentor, whose help has been precious.

On a personal note, I want to thank my friends and family, who may not always understand what I do but support me unconditionally nonetheless with their love and advice. Maman, Papa, Valérie, François, Léa, Alex, Bruno and baby Gabrielle: all is well as long as you are here. Plenty of thanks to the "Gang des Lyonnais" and the "Bande du Caousou", our closest bands of merry friends, who are always ready for celebration, commiseration, or a Star Wars GIFs on WhatsApp. You're the best!

Last, but definitely not least, thank you Martin. None of this would be possible without you, and you continue to be there always. I love you.

université
PARIS-SACLAY

Titre : Génomique comparative et émergence d'innovations évolutives chez les Vertébrés

Mots clés : Génomique évolutive, génomique comparative, phylogénomique, génomes ancestraux, transcriptomique comparative, évolution de la régulation génique

Résumé : L'évolution du génome est la source de la majorité de la variabilité phénotypique transmissible observée entre individus et entre espèces. Au cours des dernières années, l'avancée des technologies de séquençage a rendu possible l'exploration et la comparaison des génomes de nombreuses espèces pour éclairer les mécanismes par lesquels l'évolution façonne les génomes de vertébrés, mais aussi à quel degré ces génomes sont similaires au nôtre. L'observation montre que les génomes de vertébrés évoluent à de multiples échelles, de la substitution de base aux réarrangements à large échelle de la structure des chromosomes, et toutes ces changements sont susceptibles de modifier fonctionnellement les gènes ou leurs programmes d'expression. Au cours de ma carrière scientifique, j'ai étudié l'évolution des génomes de vertébrés à travers différents clades, échelles de temps et niveaux de résolution pour mieux comprendre comment ces changements génomiques permettent l'apparition d'innovations évolutives. Cette these d'habilitation résume mes travaux sur l'évolution de l'organisation du génome chez les poisons paleopolyploïdes et chez les vertébrés, ainsi que sur l'évolution de l'expression des gènes chez les mammifères et les poissons. Depuis septembre 2021, je dirige le groupe Génomique Fonctionnelle Comparative à l'Institut Pasteur. Dans la dernière section du manuscrit, je discute des directions futures de mon laboratoire pour illuminer les mécanismes fonctionnels d'innovations évolutives chez les primates et d'autres groupes de mammifères.

Title : Comparative genomics and the emergence of evolutionary innovations in Vertebrates

Keywords : Evolutionary genomics, comparative genomics, phylogenomics, ancestral genomes, comparative transcriptomics, evolution of gene regulation

Abstract : Genome evolution is the source of much of the transmissible phenotypic variability we observe within and between species. Over the past ten years, the advancements of sequencing technologies have made it possible to explore and compare the genomes of multiple species to understand how evolution acts on vertebrate genomes, but also how the genomes of these diverse species relate to our own. Evidence shows that vertebrate genomes evolve at multiple scales, from base pair substitutions to large-scale rearrangements of chromosomal structure, and these modifications can all result in functional changes in gene products and expression programs. During my scientific career, I have studied vertebrate genome evolution across multiple clades, timescales and levels of resolution in order to better understand how genomic changes can result in evolutionary novelty. This habilitation thesis summarizes my work on the evolution of genome organisation in paleopolyploid fishes and vertebrates, and on the evolution of gene expression in mammals and fishes. Since September 2021, I lead the Comparative Functional Genomics group at Institut Pasteur. In the final section of the manuscript, I discuss how my lab is heading forward to uncover the functional mechanisms of evolutionary innovations in primates and other mammalian groups.