École Doctorale SEVAB

Sciences Ecologiques, Vétérinaires, Agronomiques & Bioingénéries

Exploration de génomes par analyse de données omiques

Sylvain Foissac, INRAE Toulouse

Manuscrit destiné à l'obtention de

l'**Habilitation à Diriger des Recherches**

dans l'espoir d'une soutenance le 25 juin 2024, devant le jury composé de :

Christine Gaspin	Directrice de Recherche, INRAE Toulouse
Vincent Lacroix	Maître de conférences, Université de Lyon
Claire Lemaitre	Chargée de Recherche, INRIA Rennes
Frédérique Pitel	Directrice de Recherche, INRAE Toulouse
Marie-France Sagot	Directrice de Recherche, INRIA Lyon
Patricia Thébault	Professeuse des Universités, Université de Bordeaux

Table des matières

1 Cursus et informations diverses **3**

1.1 Parcours scientifique . . . 3

a. Contexte . . . 3

b. Formation universitaire . . . 4

c. L'épopée ENCODE . . . 4

d. Détour et retour, *in extremis* . . . 6

e. Épilogue . . . 6

1.2 Activités diverses . . . 7

a. Encadrement et formation par la recherche . . . 7

Co-encadrement de Masters . . . 7

Co-encadrement de thèses . . . 8

Participations à des comités et jury de thèse et de recrutement . . . 9

b. Responsabilités collectives, animation scientifique . . . 9

1.3 *Curriculum Vitae* . . . 10

2 Contexte scientifique et rappels **14**

2.1 La vie, la cellule, l'ADN . . . 14

2.2 Gène et expression génique . . . 16

a. La transcription . . . 16

b. La maturation . . . 17

c. La traduction . . . 17

d. Modèle général de l'expression génique . . . 18

3 Activités de recherche **20**

3.1 L'annotation génomique . . . 20

a. Contexte, objectif et enjeux . . . 20

b. Annotation de gènes codants dans un génome de plante . . . 22

Le logiciel EuGène . . . 23

Extension du modèle pour intégrer l'épissage alternatif 24
Extension du modèle pour intégrer les gènes chevauchants 26
Conclusion 26
c. Annotation v2.0 : vers une nouvelle vision du gène 27
Effacement des frontières géniques 27
Prévalence des mécanismes alternatifs et du non codant 28
Annotation de nouveaux petits ARNs 29
d. Annotation des génomes animaux 33
Conclusion : définir le gène et son expression ? 34
3.2 La génomique 3D : de la structure à la fonction 38
a. La génomique 3D et la technologie Hi-C 38
La génomique spatiale 38
Le protocole Hi-C 39
b. Annotation de structures 3D par analyse de données Hi-C 40
Des paires de lectures aux matrices Hi-C 40
Les structures identifiables 41
Application aux génomes animaux 44
c. Vers l'analyse différentielle de données Hi-C 45
Projet Treediff 45
Projet HiCDOC 46
3.3 Perspectives 46
a. Vers l'annotation fonctionnelle, translationnelle et régulationnelle 46
b. Méthodes d'analyse de données de génomique 3D 48
3.4 Productions scientifiques 48
a. Articles publiés dans des revues internationales à comité de lecture . . . 49
b. Chapitres d'ouvrage 52
c. Communications dans des conférences 52
d. Jeux de données 53
e. Logiciels 54

Remerciements **56**

Bibliographie **65**

– Chapitre 1 –

Cursus et informations diverses

1.1 Parcours scientifique

Préambule

Cette partie a pour objectif de présenter mon parcours professionnel, son contexte et ses principales étapes. Ou comment, après des années de vagabondage international à enchaîner contrats précaires, rencontres et situations improbables, j'ai fini par revenir au pays [a], auprès des miens. Avec en fil conducteur de ma trajectoire scientifique l'articulation de deux disciplines principales : **la biologie moléculaire** et **la bioinformatique**.

a. ou *Mon Païs*, comme chantait Claude Nougaro.

a. Contexte

Je suis né en 1977, **l'année du premier séquençage de génome complet** par la technique qui fut pendant plus de trente ans la référence pour séquencer l'ADN (Sanger et al. 1977). J'ai donc environ l'âge du génie génétique et de l'ADN recombinant, également apparus dans les années 70 avec les techniques et outils moléculaires permettant la manipulation de fragments génomiques : première bactérie (Mandel et al. 1970) puis souris (Jaenisch et al. 1974) transgéniques, premières utilisations d'enzymes de restriction pour la cartographie d'ADN (Danna et al. 1971) ou la production d'insuline (Villa-Komaroff et al. 1978). Cette décennie marque un tournant de la biologie moléculaire.

b. Formation universitaire

C'est lors de mes études universitaires à l'université Paul Sabatier de Toulouse III (Fig. 1.1, étape 1) que **je découvre la complexité des bases moléculaires du vivant** et la fascination que cela suscite en moi. Initialement dans une filière de biologie cellulaire, je spécialise mon cursus dans **la génétique et la biologie moléculaire** pour en obtenir une Maîtrise (Master 1) en 2000. J'apprends ainsi les détails et les limites du dogme de la biologie moléculaire et du modèle d'expression génique, sans me douter que j'allais plus tard participer aux travaux remettant en question ces mêmes enseignements.

Durant ces études, je découvre également une nouvelle discipline alors émergeante et très prometteuse : **la bioinformatique**. Je postule à différentes formations dites de 3^{eme} cycle et choisis le DEA (Master 2) de Paris Jussieu/Diderot intitulé "Analyse des Génomes et Modélisation Moléculaire", où j'apprends notamment les bases de différentes matières dont l'algorithmique, la programmation, l'apprentissage machine et l'optimisation (Fig. 1.1, étape 2). Fort de ces connaissances et face à la grisaille du climat parisien, je décide de chercher un stage de DEA plus au sud. C'est ainsi que je commence à travailler à l'INRA de Toulouse Auzeville, à Castanet-Tolosan, sous la direction d'un jeune directeur de recherche du nom de Thomas Schiex. Celui-ci me propose de poursuivre avec **une thèse de bioinformatique sur le thème de la détection de gènes** (Fig. 1.1, étape 3).

> Incidemment, ma thèse commence donc l'année de la fameuse double publication de la première version du génome humain séquencée et annotée. Alors en tant que spectateur, j'assiste ainsi à l'arrivée au coude à coude de cette course de géants, avec d'un côté le consortium international publiant dans *Nature* (The International Human Genome Sequencing Consortium 2001) et de l'autre le redoutable C. Venter et son entreprise privée *Celera Genomics* publiant dans *Science* (Venter et al. 2001). Face à l'ampleur de ce qui est accompli, j'ai conscience de vivre à distance un épisode majeur dans l'histoire de la biologie, depuis mon bureau de simple thésard, à Castanet-Tolosan.

c. L'épopée ENCODE

En fin de thèse, je suis attiré par les travaux du groupe de Roderic Guigó au *Centre de Regulació Genòmica* (CRG) de Barcelone, groupe qui par ailleurs a fait partie de la fameuse aventure du génome humain en participant à son annotation (Venter et al. 2001). La ville elle-même m'attirant également pour plusieurs raisons [1], je tente de prendre contact par courriel. Contre toute attente, je reçois une réponse, puis une proposition de rendez-vous téléphonique,

1. sa proximité, son climat et l'inévitable souvenir alors récent du film "L'auberge Espagnole".

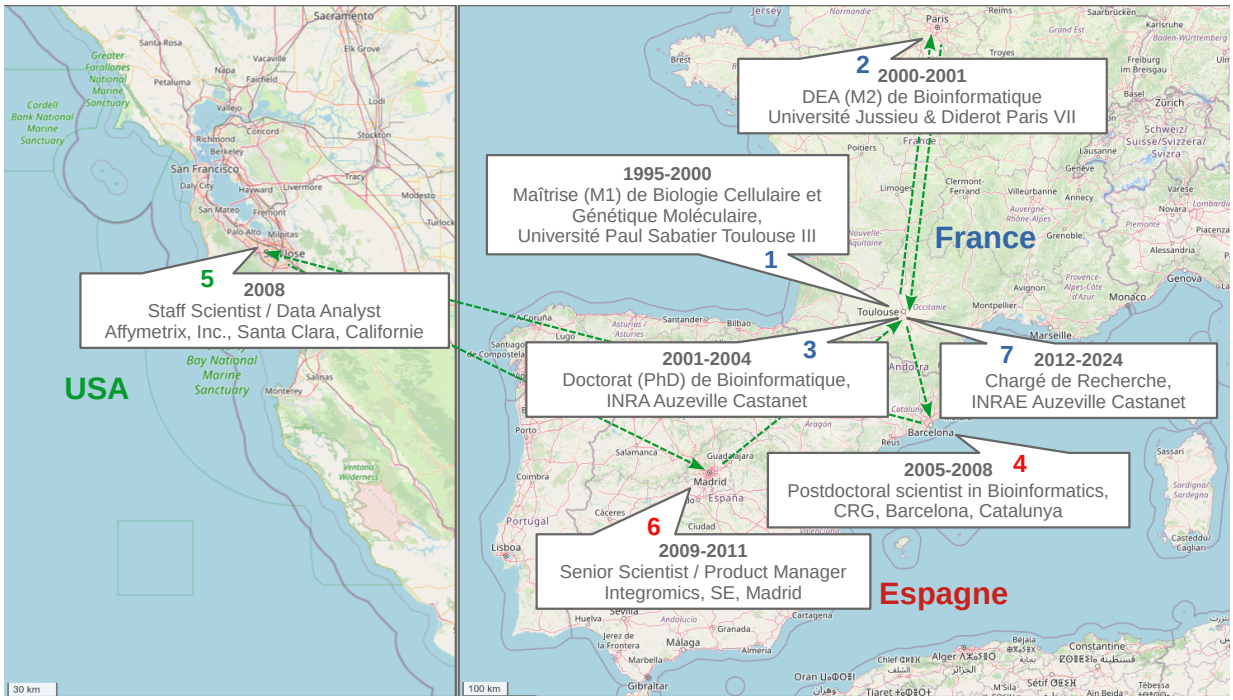

FIGURE 1.1 – Étapes principales de mon parcours universitaire et professionnel.

puis de séminaire sur place, puis de contrat postdoctoral.

Initialement prévu pour un an, **mon séjour postdoctoral à Barcelone se prolonge sur trois ans** pour plusieurs raisons [2] (Fig. 1.1, étape 4). J'apprends l'analyse de données massives dites omiques, issues de génomique, transcriptomique, etc. J'ai la chance, entre autres, de participer au projet ENCODE sur l'annotation du génome humain, et de collaborer avec les autres membres du consortium international. Je fais ainsi connaissance avec le groupe de Tom Gingeras, de l'entreprise américaine Affymetrix, alors référence mondiale dans la biotechnologie des puces à ADN. Je réussis à me faire inviter pour un séjour d'un mois sur place, près de San Francisco en Californie, à côté d'universités prestigieuses dont je voyais passer les noms évocateurs en faisant la biblio de ma thèse : Stanford, Berkeley, UCSC… pas tout à fait le même exotisme que Castanet-Tolosan.

À ma grande surprise, Tom me propose un poste de type CDI [3] pour un salaire brut de 85, 000 dollars annuels, énorme pour moi même en comptant les prélèvements, les impôts et le coût de la vie relativement élevé [4]. Acceptant la proposition, **je pars donc travailler dans une**

2. toujours le climat bien sûr, la qualité de l'environnement scientifique, et d'autres raisons que je ne peux détailler ici.

3. à cela près que le terme "Durée Indéfinie" est à prendre au pied de la lettre vu les conditions contractuelles locales qui permettent à l'employeur comme à l'employé de mettre unilatéralement fin au contrat du jour au lendemain.

4. surtout pour ce qui est loyer, santé, fromage et vin.

multinationale de la mythique *Silicon Valley*, côtoyant ainsi dans le train [5] les employés de Google, Apple et eBay (Fig. 1.1, étape 5). Mais le rêve américain est de courte durée, car moins d'un an après mon arrivée le laboratoire déménage au tout aussi prestigieux mais moins ensoleillé *Cold Spring Harbor Laboratory* dans la région de New York. Quelque peu refroidi par le nom du laboratoire, et malgré les conditions avantageuses de l'offre qui m'est proposée pour m'installer sur la côte est, je choisis de revenir sur le vieux continent, plus riche pour moi en expériences personnelles, passées et prévues.

d. Détour et retour, *in extremis*

Je décroche alors un poste de *Senior Scientist* à Integromics, une entreprise *spin-off* de l'université de Madrid qui développe et commercialise des logiciels d'analyse de données pour des laboratoires publics et privés (Fig. 1.1, étape 6). Initialement membre du groupe R&D, je lance le développement d'un nouveau produit pour l'analyse de données issues de séquençage à haut débit, technologie prenant progressivement le pas sur les puces à ADN. Je passe *Product Manager* sur une suite logicielle dont le portefeuille client approche le million d'euros de chiffre d'affaire annuel. Supervisant l'ensemble de la ligne de production, de la conception à la vente en passant par le développement, le test et le marketing, j'apprends beaucoup dans des domaines aussi variés que nouveaux. Cependant, je me perds un peu dans des activités qui m'éloignent dangereusement de la science [6].

Dans ce contexte, il m'est de plus en plus difficile de poursuivre mes travaux de recherche, notamment en collaboration avec des anciens collègues (CRG, Affymetrix, consortium ENCODE). Je candidate alors pour des postes de chercheur en France, dont celui qui m'est attribué à l'INRA de Toulouse, dans un laboratoire voisin de celui de ma thèse (Fig. 1.1, étape 7). C'est ainsi qu'après presque huit ans à l'étranger, dont la moité dans le privé, **j'intègre en 2012 la fonction publique en tant que Chargé de Recherche**, poste que j'ai le plaisir d'occuper à l'heure où je rédige ces lignes.

e. Épilogue

Incidemment, l'année de ma prise de fonction sont publiés plusieurs articles du projet ENCODE, dont le principal paraît dans un numéro spécial de la prestigieuse revue *Nature*. Malgré l'absence d'instituts français dans le consortium, j'ai cette fois la petite fierté de voir mon nom dans la liste des signataires. Bien entendu, il y figure en caractères minuscules en toute fin

5. ainsi que sur les terrains de sport du campus de Stanford où je jouais au *touch rugby*.

6. Exemple de réalisations improbables : une étude de marché pour convaincre les actionnaires de financer un projet, des communiqués de presse, une vidéo de démonstration de logiciel, du démarchage téléphonique pour vendre ou faire renouveler des licences, et, paroxysme du *packaging* virtuel en ligne, l'image d'un faux emballage pour un objet qui n'existe pas.

d'article, enfoui dans une liste de centaines de noms, à la mesure de ma contribution toute relative. Mais il y figure néammoins.

Curieusement, si de tels éclats de prestige glanés pendant mon périple ont suffisamment auréolé mon dossier pour m'octroyer un poste, **je suis finalement revenu à Castanet-Tolosan, mon point de départ**. Par rapport à ma position de spectateur distant d'il y a vingt ans, la différence est que je suis passé de simple thésard à... simple chercheur. Mais surtout, j'ai davantage conscience de la chance que cela représente. Comme le montrent régulièrement mes collègues, on peut faire de la science de qualité chez nous aussi.

Même à Castanet-Tolosan.

1.2 Activités diverses

Préambule

Cette partie présente quelques-unes de mes activités professionnelles ne correspondant pas directement à de la production scientifique, essentiellement de l'animation scientifique et/ou implication dans divers collectifs.

a. Encadrement et formation par la recherche

Co-encadrement de Masters

Année	Étudiant-e	Co-encadré-e avec	Lieu	Formation
2015	Marjorie Mersch	Frédérique Pitel	GenPhySE, INRAE	M2 Univ. Toulouse
2017	Matthew Smart	Sarah Djebali, Kylie Munyard	GenPhySE, INRAE	M1 Univ. Curtin, Australie
2017	Raphaël Taris	Céline Noirot, Frédérique Pitel	MIAT, INRAE	M2 Univ. Toulouse
2018	Camille Mestre	Thomas Faraut, Sarah Djebali	GenPhySE, INRAE	M2 Univ. Toulouse
2019	Cyril Kurylo	Matthias Zytnicki	GenPhySE, INRAE	M2 Univ. Toulouse
2019	Marie Jeremy	Frédérique Pitel	GenPhySE, INRAE	M2 Univ. Toulouse
2022	Tess Azevedo	Sarah Djebali, Cervin Guyomar	GenPhySE, INRAE	M1 INSA Toulouse
2023	Gwendaelle Cardenas	Nathalie Vialaneix	MIAT, INRAE	M2 Univ. Rennes
2024	Victor Lefebvre	Anamaria Necsulea, Sarah Djebali	LBBE, CNRS	M2 Univ. Lyon

Co-encadrement de thèses

Maria Marti Marimon

date	2016-2018
titre	3D genome conformation and gene expression in fetal pig muscle at late gestation
co-encadrée avec	Martine Yerle, INRAE
réalisée à	GenPhySE, INRAE de Toulouse
soutenue le	9/11/2018
discipline	biologie cellulaire/moléculaire et bioinformatique
lien	https://theses.fr/2018INPT0099
note	première déclaration “officielle” de co-encadrement de thèse

Marjorie Mersch

date	2016-2018
titre	Analyse de la méthylation de l’ADN par séquençage haut-débit chez la Poule
co-encadrée avec	Frédérique Pitel, INRAE
réalisée à	GenPhySE, INRAE de Toulouse
soutenue le	30/10/2018
discipline	bioinformatique et biologie moléculaire
lien	https://theses.fr/2018INPT0107
note	co-encadrement non affiché pour raisons administratives mais contribution substantielle cependant

Nathanaël Randriamihamison

date	2018-2021
titre	Classification Ascendante Hiérarchique sous Contrainte de Contiguïté pour l’Analyse de données Hi-C
co-encadrée avec	Nathalie Vialaneix, INRAE, Marie Chavent, INRIA, Pierre Neuvial, CNRS
réalisée à	MIAT, INRAE de Toulouse
soutenue le	27/10/2021
discipline	mathématiques et statistique appliquées
lien	https://www.theses.fr/2021TOU30108
note	contribution superficielle, du moins sur les aspects scientifiques

Élise Jorge

date	2023-2025
titre	Analyse comparative de données de génomique 3D
co-encadrée avec	Nathalie Vialaneix, INRAE, et Pierre Neuvial, CNRS
réalisée à	GenPhySE, INRAE de Toulouse
soutenue le	en cours
discipline	statistique appliquée, bioinformatique
note	thèse en cours pour laquelle je suis déclaré encadrant principal (avec dérogation HDR)

Participations à des comités et jury de thèse et de recrutement

- 2016 : jury de thèse de Sergio Espeso Gil, Universitat Pompeu Fabra, Catalunya, Barcelona, en tant qu'examinateur.
- 2018-2019 : comité de thèse d'Alexandre Heurteau, CNRS, Université Toulouse.
- 2019 : jury de thèse de Leandro Lima, Université Lyon 1, en tant qu'examinateur (https://www.theses.fr/2019LYSE1055).
- 2019-2020 : membre de la commission de recrutement CRCN-TH (2019) puis du jury professionnel (2020) pour l'unité GABI, INRAE Jouy-en-Josas, en tant que rapporteur.
- 2021-2022 : comité de thèse de Fabien Degalez, INRAE, Agrocampus Ouest, Rennes.
- 2021-2022 : comité de thèse de Chloé Cerutti, INRAE, Université Toulouse.
- 2022-2023 : comité de thèse de Samira Ghazali, INSERM, Université Toulouse.
- 2024-2025 : comité de thèse de Sébastien Cabanac, INRAE, Université Toulouse.

b. Responsabilités collectives, animation scientifique

- Représentant de l'équipe Dynagen puis des CR au Conseil de Service de l'unité GenPhySE entre 2016 et 2020.
- Représentant du personnel en Commission Administrative Paritaire (CAP), élu sur deux mandatures de 2016 à 2022.
- Représentant du personnel en Commission Consultative Paritaire des Contractuels (CCPC), mandaté depuis 2023.
- Membre de la Commission Scientifique Spécialisée (CSS) "GVA - Génétique Végétale et Animale" chargée de l'évaluation des CR et DR INRAE, nommé pour la mandature 2020-2024.

- Depuis décembre 2018 : animateur du groupe de travail transdisciplinaire **CHROCOGEN : Chromatin Conformation and Gene Expression** sur la génomique 3D et les liens entre structure et fonction de la chromatine. Organisation de séminaires et discussions scientifiques avec des présentations le plus souvent sous la forme de "journal club", ou de partages informels de résultats sur des projets en cours. Initialement petit groupe inter-unités du centre INRAE de Toulouse réunissant collègues *wet* et *dry*, le développement de la thématique de la génomique 3D sur la région a permis d'inclure des membres de divers laboratoires du campus universitaire voisin. La période COVID, le passage progressif vers le distanciel, les mobilités et le "réseautage" ont fait élargir le périmètre géographique du groupe, qui comprend aujourd'hui une soixantaine d'inscrit-es de différentes affiliations (INRAE, INSERM, CNRS, CEA et Universités) sur plusieurs ville (Toulouse, Paris, Bordeaux, Lyon). En cinq ans d'existence, toujours sur la base du volontariat de ses membres, le réseau a proposé une quarantaine de présentations et quelques événements ponctuels [7].
  Informations, programme et inscription : https://groupes.renater.fr/sympa/info/chrocogen. Financement sur la période 09/2023 - 02/2025 par le métaprogramme INRAE Digit-BIO via le projet ChrocoNET : https://digitbio.hub.inrae.fr/rubriques-verticales2/nos-actions/consortia/consortium-chroconet-2023-2025
- Depuis mai 2024 : **responsable du Pôle Scientifique "Génomique Structurale et Fonctionnelle"** de l'unité GenPhySE, en charge de la prospective et de l'animation scientifique des trois équipes de recherche faisant partie dudit pôle.

1.3 *Curriculum Vitae*

Sylvain Foissac, Research Scientist

INRAE GenPhySE, ch. de Borde Rouge, 31326 Castanet-Tolosan, France

http://genoweb.toulouse.inra.fr/~sfoissac

sylvain.foissac@inrae.fr

Research statement

I am interested in **the fundamental mechanisms of life at the molecular scale**, particularly at the genome level. My goal is to understand how the genome is organized, functions, and how its linear and three-dimensional structures regulate gene expression.

To achieve this, **I design, develop, and utilize computational methods and tools to**

7. dont par exemple un atelier de type "mini-symposium" sur la génomique 3D lors de l'édition 2024 de la conférence nationale de bioinformatique JOBIM à Toulouse.

analyze genomics, transcriptomics, and other -omics data, mostly obtained from high-throughput sequencing experiments. My expertise lies at the interface between various fields, combining molecular biology, bioinformatics, genomics, biostatistics, and computational science. My research focuses on **genome annotation, 3D genomics, and comparative data analysis**.

Summary

- About **20 years of experience in computational biology and omics data analysis**.
- Scientific expertise in structural and functional genomics.
- Publication record of **31 peer-reviewed articles** and 3 book chapters.

Experience

- Nov 2012 - present : Research Scientist ("CRCN") at **INRAE** in Toulouse, France.
- Dec 2010 - Oct 2012 : Product Manager at **Integromics, S.L.** in Madrid, Spain.
- Dec 2008 - Dec 2010 : Senior Research Scientist at **Integromics, S.L.** in Madrid, Spain.
- Apr 2008 - Nov 2008 : Data Analyst at **Affymetrix, Inc.** in Santa Clara, California, USA.
- Apr 2005 - Mar 2008 : PostDoctoral researcher at the **CRG** in Barcelona, Catalunya, Spain.

Education

- 2001-2004 : PhD thesis in Bioinformatics. Toulouse 3 University, France.
- 2000-2001 : Master degree in Bioinformatics. Paris VI & VII University, France.
- 1995-2000 : Bachelor degree in Cellular and Molecular Genetics. Toulouse 3 University, France.

Publications

– Degalez et al (2024). Enriched atlas of lncRNA and protein-coding genes for the GRCg7b chicken assembly and its functional annotation across 47 tissues. **Scientific Reports**. 10.1038/s41598-024-56705-y.

– Neuvial et al (2024). A two-sample tree-based test for hierarchically organized genomic signals. **Journal of the Royal Statistical Society Series C : Applied Statistics**. 10.1093/jrsssc/qlae011.

– Dufour et al (2024). Cell specification and functional interactions in the pig blastocyst inferred from single-cell transcriptomics and uterine fluids proteomics. **Genomics**. 10.1016/j.ygeno.2023.110780.

– Kurylo et al (2023). TAGADA : a scalable pipeline to improve genome annotations with RNA-seq data. **NAR Genomics And Bioinformatics**. 10.1093/nargab/lqad089.

– Hoellinger et al (2023). Enhancer/gene relationships : need for more reliable genome-wide reference sets. **Frontiers in Bioinformatics**. 10.1093/nar/gkx920.

– Jehl et al (2021). RNA-Seq Data for Reliable SNP Detection and Genotype Calling : Interest for Coding Variant Characterization and Cis-Regulation Analysis by Allele-Specific Expression in Livestock Species. **Frontiers in Genetics**. 10.3389/fgene.2021.655707.

– Marti-Marimon et al (2021). Major Reorganization of Chromosome Conformation During Muscle Development in Pig. **Frontiers in Genetics**. 10.3389/fgene.2021.748239.

– Jehl et al (2020). An integrative atlas of chicken long non-coding genes and their annotations across 25 tissues. **Scientific Reports**. 10.1038/s41598-020-77586-x.

– Giuffra et al (2019). Functional Annotation of Animal Genomes (FAANG) : Current Achievements and Roadmap. **Annual Review of Animal Biosciences**. 10.1146/annurev-animal-020518-114913.

– Foissac et al (2019). Multi-species annotation of transcriptome and chromatin structure in domesticated animals. **BMC Biology**. 10.1186/s12915-019-0726-5.

– Fève et al (2017). Identification of a t(3 ;4)(p1.3 ;q1.5) translocation breakpoint in pigs using somatic cell hybrid mapping and high-resolution mate-pair sequencing. **PLoS ONE**. 10.1371/journal.pone.0187617.

– David et al (2017). Genome-wide epigenetic studies in chicken : A review. **Epigenomes**. 10.3390/epigenomes1030020.

– Muret et al (2017). Long noncoding RNA repertoire in chicken liver and adipose tissue. **Genetics Selection Evolution**. 10.1186/s12711-016-0275-0.

– Andersson et al (2015). Coordinated international action to accelerate genome-to-phenome with FAANG, the Functional Annotation of Animal Genomes project. **Genome Biology**. 10.1186/s13059-015-0622-4.

– Rubio-Peña et al (2015). Modeling of autosomal-dominant retinitis pigmentosa in Caenorhabditis elegans uncovers a nexus between global impaired functioning of certain splicing factors and cell type-specific apoptosis. **RNA**. 10.1261/rna.053397.115.

– Bonnet et al (2013). An overview of gene expression dynamics during early ovarian folliculogenesis : Specificity of follicular compartments and bi-directional dialog. **BMC Genomics**. 10.1186/1471-2164-14-904.

– Djebali et al (2012). Landscape of transcription in human cells. **Nature**. 10.17615/r8aq-6h68.

– Djebali et al (2012). Evidence for transcript networks composed of chimeric rnas in human cells. **PLoS ONE**. 10.1371/journal.pone.0028213.

– The ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. **Nature**. 10.1038/nature11247.

– Ozsolak et al (2010). Comprehensive polyadenylation site maps in yeast and human reveal pervasive alternative polyadenylation. **Cell**. 10.1016/j.cell.2010.11.020.

– Kapranov et al (2010). New class of gene-termini-associated human RNAs suggests a novel

RNA copying mechanism. **Nature**. 10.1038/nature09190.

– Fejes-Toth et al (2009). Post-transcriptional processing generates a diversity of 5'-modified long and short RNAs. **Nature**. 10.1038/nature07759.

– Djebali et al (2008). Efficient targeted transcript discovery via array-based normalization of RACE libraries. **Nature Methods**. 10.1038/nmeth.1216.

– Piqué et al (2008). A Combinatorial Code for CPE-Mediated Translational Control. **Cell**. 10.1016/j.cell.2007.12.038.

– Sammeth et al (2008). A general definition and nomenclature for alternative splicing events. **PLoS Computational Biology**. 10.1371/journal.pcbi.1000147.

– Foissac et al (2008). Genome Annotation in Plants and Fungi : EuGene as a Model Platform. **Current Bioinformatics**. 10.2174/157489308784340702.

– Foissac et al (2007). ASTALAVISTA : dynamic and flexible analysis of alternative splicing events in custom gene datasets. **Nucleic acids research**. 10.1093/nar/gkm311.

– The ENCODE Project Consortium (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. **Nature**. 10.1038/nature0587.

– Denoeud et al (2007). Prominent use of distal 5' transcription start sites and discovery of a large number of additional exons in ENCODE regions. **Genome Research**. 10.1101/gr.5660607.

– Foissac et al (2005). Integrating alternative splicing detection into gene prediction. **BMC Bioinformatics**. 10.1186/1471-2105-6-25.

– Foissac et al (2003). EUGÈNE'HOM : A generic similarity-based gene finder using multiple homologous sequences. **Nucleic Acids Research**. 10.1093/nar/gkg586.

– Chapitre 2 –

Contexte scientifique et rappels

Préambule

Ce chapitre a pour objectif de rappeler quelques bases de biologie moléculaire potentiellement utiles pour la suite. Pour des raisons de narration, les connaissances présentées ici datent du temps où je les ai apprises, il y a donc plus de vingt ans. De plus, le manque de temps, d'espace et de compétence [a] m'a imposé plusieurs raccourcis et approximations [b]. De nature optimiste, j'en fais donc appel à l'indulgence du lectorat. Dans tous les cas, pour toute information complémentaire sur le contenu de cette partie, le plus simple reste probablement de demander à un téléphone, qui demandera à son tour à une intelligence artificielle. Reste sinon le meilleur moyen de s'informer en biologie, que je recommande souvent aux plus jeunes : d'abord Wikipedia, puis les résumés sur PubMed, et enfin les articles scientifiques eux-mêmes, sachant que solliciter l'expertise d'un-e collègue s'insère avantageusement à n'importe quelle étape de ce pipeline [c].

a. la mienne bien sûr, mais aussi celle des modèles de langage génératifs, encore trop insatisfaisants à mon goût malheureusement.

b. sans compter mes incorrigibles tentatives d'humour, beaucoup plus rares dans le chapitre suivant cependant.

c. idéalement agrémenté de café, thé ou bière.

2.1 La vie, la cellule, l'ADN

Tout organisme vivant est constitué d'une ou plusieurs cellules. La cellule, unité de base de la vie, contient l'ensemble de l'information génétique appelé génome qui caractérise l'organisme et son espèce. La présence d'un noyau, compartiment renfermant le génome, distingue les organismes dits eucaryotes, à cellule(s) nucléée(s), des procaryotes, qui en sont

dépourvus. Malgré leur importance et leur supériorité à plusieurs niveaux (on peut citer par exemple leur nombre, leur ubiquité, la longévité de leur suprématie sur la planète et leur capacité à en coloniser tous les habitats), les procaryotes ne seront tout simplement pas abordés dans la suite de ce manuscrit, sous prétexte qu'ils sont incapables de le lire.

Le support moléculaire de l'information génétique est l'acide désoxyribonucléique ADN (ou *DNA* [1]). L'ADN génomique se présente sous la forme d'une hélice composée de deux brins (Fig 2.1). Chaque brin résulte d'un enchaînement linéaire et orienté d'élements appelés nucléotides, composés chacun de trois parties : un acide phosphorique, un sucre et une base azotée ("base"). **On distingue pour l'ADN quatre types de nucléotides représentés par les lettres A, C, G et T**, qui ne diffèrent que par leur base, respectivement une adénine, une cytosine, une guanine et une thymine. La longueur d'un brin étant fonction du nombre de nucléotides et donc de bases dans la chaîne, l'unité de longueur de l'ADN est la paire de base (*base pair : bp*) ou simplement base (b).

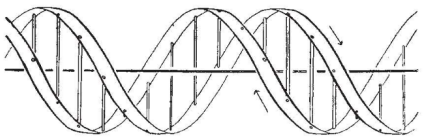

FIGURE 2.1 – Structure schématique de l'ADN. Les deux bandes en hélice représentent les chaînes nucléotidiques, les traits verticaux les appariements entre bases, la ligne horizontale l'axe virtuel de l'hélice et les flèches l'orientation antiparallèle. Source : Watson et al. 1953b, d'après notamment les travaux de R. Franklin.

Dans un brin d'ADN, chaque nucléotide est lié au précédent par son atome de carbone appelé 5' et au suivant par son carbone 3' (selon la numérotation conventionnelle des carbones composant le squelette du sucre). Le premier nucléotide de la chaîne présente donc un carbone 5' libre d'une telle liaison et le dernier un carbone 3', **ce qui définit pour tout brin une extrémité dite 5', une autre 3', et un sens 5' → 3'**. Les deux brins de l'ADN génomiques sont antiparallèles car en orientations 5' → 3' opposées (Fig 2.1). L'appariement entre bases complémentaires permet l'hybridation entre brins, formant des paires de bases impliquant majoritairement soit une adénine et une thymine (paire A-T), soit une guanine et une cytosine (paire C-G).

Par conséquent, on peut représenter l'information génétique contenue dans l'ADN par une succession de lettres A, C, G et T identifiant les bases successives d'un des brins. Par convention, une séquence d'ADN représente les bases de 5' vers 3' dans le sens de lecture, la séquence de l'autre brin pouvant se déduire par complémentarité (et inversion). Par exemple, la séquence

1. Malgré mes efforts, ce manuscrit contient un mélange de termes anglophones et francophones. J'en appelle encore une fois à l'indulgence du lectorat.

reverse-complémentaire de `GATTACA` est `TGTAATC`.

2.2 Gène et expression génique

La fonction principale du génome est de permettre le stockage, la transmission et l'expression de son information génétique. Dans ce contexte, certaines régions du génome jouent un rôle particulièrement important dans le vivant : **les gènes**. Longtemps considéré comme l'unité fonctionnelle de l'information héréditaire, le gène est un concept aussi fondamental que difficile à définir de façon précise et consensuelle. Généralement, on peut concevoir un gène comme la partie d'une molécule d'ADN génomique nécessaire à la synthèse dans la cellule d'un produit biologique fonctionnel, généralement une protéine ou un ARN (Fig. 2.2). Quel qu'il soit, ce produit est une séquence d'éléments dont la nature est déterminée par celle de la région d'ADN d'origine.

> **Le processus de production d'une protéine ou d'un ARN à partir du gène correspondant s'appelle l'expression génique**. Il s'agit du mécanisme moléculaire à la base du fonctionnement de tout organisme vivant, et plus généralement de la vie sous toutes ses formes connues à ce jour.

Comme évoqué plus haut, il ne sera fait mention par la suite que de l'expression génique eucaryote, qui concerne à peu près tout ce qui n'est pas bactérien[2]. Le processus peut se décomposer en trois étapes principales (Fig. 2.2).

a. La transcription

La transcription produit une copie de l'information génétique contenue sur une partie de l'ADN génomique par la synthèse d'une molécule d'acide ribonucléique ARN. Pendant la transcription, un complexe moléculaire comprenant une protéine de polymérisation d'ARN (*RNA polymerase*) se fixe sur l'ADN au niveau d'une région située en début du gène : le promoteur. Puis le complexe se déplace en suivant un des brins d'ADN, construisant un brin d'ARN en assemblant des nucléotides l'un après l'autres (polymérisation). Les nucléotides de l'ARN diffèrent de ceux de l'ADN par un atome d'oxygène dans leur sucre, et par la présence de la base uracile (U) à la place de la thymine (T) de l'ADN. L'élongation de l'ARN s'effectue en produisant une séquence identique à celle d'un des brins d'ADN (aux substitutions T $\rightarrow$ U près), qui définit le sens de transcription. En fin de gène, la polymérase d'ARN se décroche de l'ADN, terminant la transcription. La molécule d'ARN produite, appelée transcrit primaire,

2. incluant donc la plupart de la vague biomasse globalement croissante qui constitue tant bien que mal l'auteur de ces lignes.

précurseur d'ARN messager, pré-messager ou encore pré-ARNm (*pre-mRNA*) passe ensuite par une étape de maturation.

b. La maturation

La maturation est une étape de modification du pré-ARNm qui produit un ARN messager mature (ARNm ou *mRNA*). En particulier, les étapes de **coiffage (*capping*) et de polyadénlyation** protègent l'ARN d'une dégradation précoce en ajoutant respectivement une coiffe (*cap*) à l'extrémité 5' et une queue polyA à l'extrémité 3'. Une autre étape importante de la maturation est **l'épissage**, qui consiste à exciser certaines parties internes de l'ARN pré-messager (les introns) pour rabouter entre elles les autres parties par conséquent conservées (les exons). Les jonctions exon-intron, appelées sites d'épissage, se caractérisent par une séquence de base plus ou moins variable selon la position dans le site. Par exemple, les introns commencent généralement par `GT` en 5' et se terminent par `AG` en 3'. La maturation produit donc un transcrit mature (ARNm ou *mRNA*) dont la séquence résulte de la concaténation des exons du transcrit primaire. Il est toutefois important de noter que ces étapes de coiffage, polyadénylation et épissage ne s'appliquent pas forcément à tous les transcrits.

> **L'épissage alternatif** est une variation dans le modèle général de l'expression génique, qui permet de produire différentes séquences d'ARNm à partir d'un même transcrit primaire. En fonction des conditions biologique et en partie du hasard [a], une variabilité dans l'utilisation des sites d'épissage peut provoquer une différence d'introns excisés entre ARNm. Il en résulte que plusieurs structures exon-intron coexistent au sein d'un même gène. Ce mécanisme peut se combiner avec d'autres, comme une polyadénylation alternative par exemple.
>
> *a.* une composante stochastique intervient dans la plupart des mécanismes moléculaires.

c. La traduction

La traduction consiste à produire une protéine à partir de l'information portée par l'ARN messager. Comme pour la maturation, il s'agit d'une étape facultative, de nombreux ARNs assurant leur fonction sans coder pour une protéine. Le complexe moléculaire réalisant la traduction, **le ribosome**, est lui-même constitué d'un ensemble de protéines et d'ARNs non-codants. Toute protéine est en premier lieu caractérisée par une séquence d'acides aminés, elle-même déterminée par une partie dite codante de la séquence de l'ARNm. La séquence codante :

- est parcourue par le ribosome dans le sens 5' $\rightarrow$ 3'.

- est composée de triplets de bases (codons), le code génétique permettant au ribosome d'associer à chacun d'eux un des vingt acides aminés existants.
- commence par un codon start `ATG` et se termine par un des trois codons stop `TAA`, `TAG` ou `TGA`.
- est flanquée par des régions non traduites appelées 5' et 3' UTR, respectivement en début et fin d'ARNm.

Les codons stop, les seuls à ne pas avoir d'acide aminé associé [3], provoquent le décrochement du ribosome et la libération de la protéine ainsi constituée.

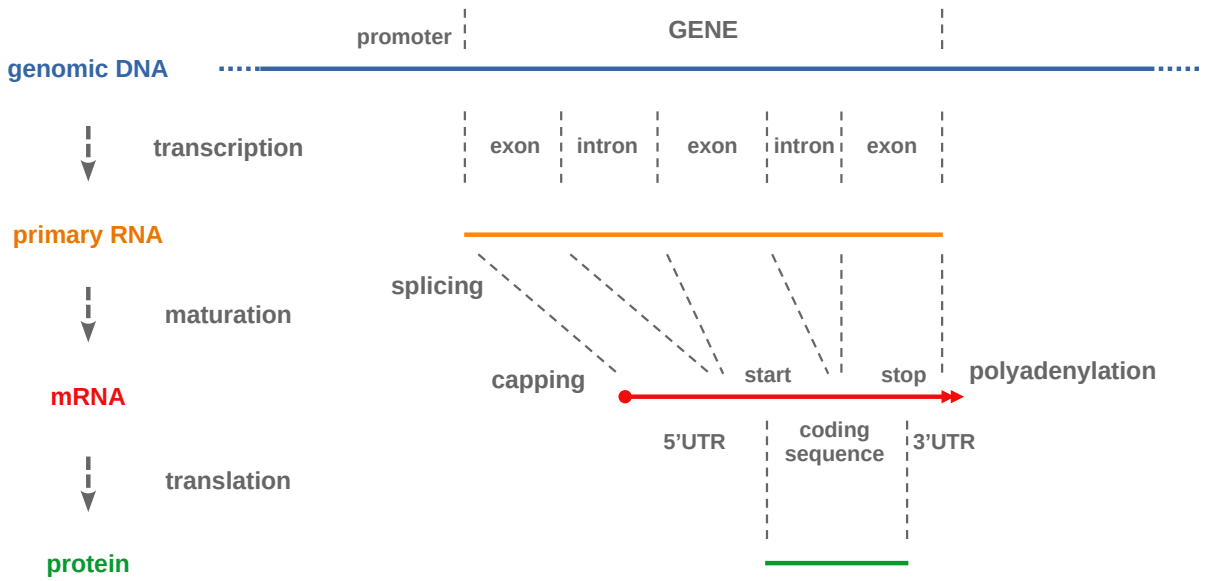

FIGURE 2.2 – Représentation schématique du modèle général de l'expression génique eucaryote. Vision simpliste enseignée en fin du siècle dernier, encore d'actualité pour beaucoup de scientifiques. Échelle et proportions non valides.

d. Modèle général de l'expression génique

De ces notions basiques découle **un modèle général de l'expression génique eucaryote**, qui repose sur le dogme fondamental de la biologie moléculaire, et qui sert de base à la vision que l'on a généralement du génome (Fig. 2.2) : on considère souvent les gènes comme

3. sauf de rares exceptions comme la sélénocystéine et la pyrrolysine, pouvant parfois s'associer aux codons UGA et UAG, respectivement.

des îlots de transcription, séparés entre eux par de grands espaces intergéniques (concept de l'"ADN poubelle"), servant à produire des protéines qui assurent les fonctions biologiques de l'organisme. Il s'agit bien sûr d'une vision historique et simplifiée, et déjà pendant mes études de nombreuses variations et exceptions m'avaient été présentées : cas des ARN non codants (ncRNAs), des petits ARNs, de l'épissage alternatif... mais globalement, ce modèle général reste encore tributaire de la vision la plus généralement répandue, même chez les biologistes.

– Chapitre 3 –

Activités de recherche

Préambule

Ce chapitre, normalement le plus sérieux du manuscrit, a pour but de présenter mes activités de recherche, leur contexte, leur objectif, leur motivation et leur trajectoire. La première section, relativement dense par endroits, traite le thème de l'annotation. La suivante, plus succinte, celui de la génomique 3D. La fin est une liste de mes productions et d'autres activités.

3.1 L'annotation génomique

a. Contexte, objectif et enjeux

Qu'il s'agisse de continents, planètes ou génomes, il n'est d'exploration sans cartographie. Au séquençage génomique, qui identifie la nature et l'ordre des bases d'ADN contenues dans les chromosomes, succède ainsi l'annotation, chargée de renseigner leurs parties potentiellement fonctionnelles dans la cellule. À l'instar d'une annotation d'article ou de livre, qui surligne, souligne ou entoure certaines parties du texte, l'annotation d'un génome appose de l'information sur une séquence, en premier lieu la position des gènes. Les travaux de nombreuses disciplines reposent sur la cartographie génomique produite, faisant de l'annotation une étape cruciale dans la compréhension du vivant.

Concrètement, la nature linéaire de la séquence caractérisant un espace à une dimension, les régions d'intérêt se localisent donc avec des intervalles délimités par deux positions génomiques (début et fin) pour un chromosome spécifique. L'information de brin peut également être nécessaire pour certains processus moléculaires orientés, comme la transcription qui ne produit

pas le même transcrit selon le brin d'ADN utilisé comme matrice. Ainsi, un fichier d'annotation peut décrire divers éléments fonctionnels (gènes, transcrits, exons, sites de fixation de facteurs de transcription, etc.) dans un format standard, typiquement de type texte tabulé (GTF, BED, etc.). L'information la plus attendue d'une annotation est typiquement la position et le nom de tous les gènes du génome, avec la structure exon-intron des transcrits associés.

Or, produire une annotation n'est pas une mince affaire. Il s'agit d'intégrer plusieurs types d'information, par exemple la séquence génomique elle-même, la séquence de produits d'expression génique comme des transcrits ou des protéines, et des connaissances déjà établies sur d'autres génomes ou le même. Comme pour le séquençage, l'annotation demande d'importantes ressources humaines et informatiques, ce qui explique que les mêmes groupes réalisaient souvent les deux opérations. La production d'un nouveau génome de référence s'accompagnait d'une annotation conjointe, et la publication associée décrivait l'ensemble des résultats, produits par de grands consortiums (The International Human Genome Sequencing Consortium 2001 ; The Mouse Genome Sequencing Consortium 2002). Depuis, des groupes se sont dédiés à l'annotation, en se spécialisant sur une espèce ou un groupe d'espèces (Gramates et al. 2022 ; Denoyelle et al. 2021 ; The ENCODE Project Consortium 2012 ; Yue et al. 2014) ou de façon plus générique avec par exemple les annotations ENSEMBL de l'EBI (Martin et al. 2022), les annotations RefSeq du NCBI (Farrell et al. 2021) ou celles de l'UCSC (Nassar et al. 2022). Comme l'annotation a longtemps relevé davantage de l'artisanat que de l'industrie, la qualité des méthodes automatiques restant bien inférieure à celle des annotations plus "manuelles" (Searle et al. 2010), chaque groupe a développé sa propre expertise. Conséquemment, les méthodes diffèrent. Il en résulte une qualité d'annotation très variable en termes de fiabilité et complétion en fonction du groupe d'origine, de l'espèce, voire d'une version à une autre. Cette variabilité, bien que documentée depuis les premières versions du génome humain par exemple (Hogenesch et al. 2001), reste un problème critique pour la communauté car elle impacte la quasi-totalité des études génomiques réalisées dans le monde (Zhao et al. 2015 ; Chisanga et al. 2022). De plus, contrairement à ce que l'on pourrait croire, les spectaculaires progrès technologiques réalisés dans le domaine du séquençage n'ont pas forcément facilité la tâche de l'annotation. En effet, l'augmentation du débit de production de données a multiplié non seulement le nombre de séquences génomiques à annoter mais aussi celui des séquences transcriptomiques à intégrer.

Au cours de ma carrière, j'ai abordé différents aspects de l'annotation génomique, en progressant sur deux axes principaux : (1) l'évolution des technologies et (2) ma curiosité scientifique.

1. L'évolution constante des biotechnologies demande un renouvellement régulier des méthodes d'analyse, dans le domaine de l'annotation et de la bioinformatique en général. Bien que souvent mise en avant, il ne s'agit pas que de l'augmentation exponentielle du volume des données produites, mais aussi de l'évolution du type de ces données. Par

exemple, la transition entre les puces à ADN *microarrays* et le séquençage à haut débit comme technologie de référence pour la transcriptomique a modifié la nature du signal produit. Or, les méthodes permettant d'analyser un signal continu, comme l'intensité lumineuse issue d'hybridation dans les puces, diffèrent de celles permettant d'analyser un signal discret, comme le nombre de lectures séquencées et alignées. Autre exemple dans le domaine du séquençage : les méthodes n'ont pas seulement été affectées par l'augmentation du nombre de séquences à traiter, mais aussi par l'évolution de leur taille, que ce soit pour des applications comme l'assemblage ou l'alignement.

2. Deuxième moteur de mon parcours, la curiosité scientifique m'a toujours poussé à explorer au-delà des connaissances établies, quitte à les remettre parfois en question. Prolongeant les études universitaires, la recherche permet de continuer à apprendre tout le long de la carrière, ce qui, dans nos domaines à évolution rapide, relève autant de la nécessité impérieuse que du plaisir. En outre, elle permet de contribuer à certaines avancées, ce qui est toujours gratifiant. De façon générale, que l'on participe ou non aux travaux y aboutissant, la découverte d'un nouveau mécanisme moléculaire, d'une technologie révolutionnaire, d'un algorithme brillant ou d'un modèle renversant le consensus établi procure toujours une stimulation intellectuelle d'une rare intensité. Cette soif de découverte et de questionnement des bases moléculaires du vivant a toujours été pour moi une source de motivation majeure, nourrissant continuellement mon enthousiasme parfois excessif.

Ce chapitre vise à montrer comment l'articulation de ces deux axes m'a permis de participer à des découvertes conduisant à remettre en question la vision alors établie de la structure et du fonctionnement du génome eucaryote.

b. Annotation de gènes codants dans un génome de plante

Fiche Synthétique

- **Lieu** : INRA Auzeville (thèse)
- **Génome** : plantes, surtout *Arabidopsis thaliana* (arabette) et *Oryza sativa* (riz)
- **Cible** : gènes et transcrits codants
- **Données traitées** : génomes et transcriptomes (ESTs de séquençage Sanger)
- **Contribution principale** : méthode et logiciel
- **Publications co-signées** : Foissac et al. 2003 ; Foissac 2004 ; Foissac et al. 2005 ; Foissac et al. 2008

Le logiciel EuGène

Mes premiers travaux concernent l'annotation de gènes codant pour des protéines, cible relativement simple mais prioritaire. L'objectif de la détection de gènes est d'identifier la position et la composition exon-intron de tous les gènes présents dans une séquence génomique donnée, en précisant donc la position de début et de fin de chaque exon. Concrètement, lors de ma thèse à l'INRA de Toulouse, j'ai contribué au développement de nouvelles fonctions dans le logiciel EuGène, initialement utilisé pour annoter le génome de la plante modèle *Arabidopsis thaliana* (arabette). En bref, EuGène fonctionne ainsi :

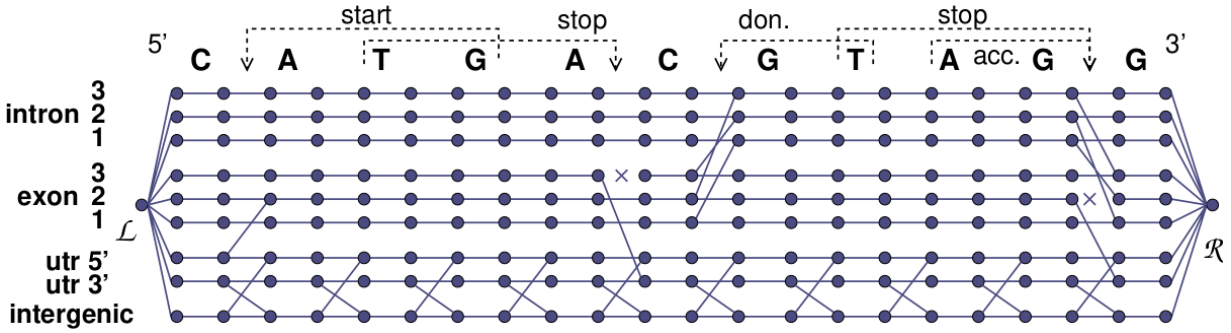

FIGURE 3.1 – Le graphe orienté acyclique modélisant les structures de gène possibles dans EuGène (Foissac et al. 2005). Les pistes horizontales représentent les annotations (états) possibles. Chaque nucléotide de la séquence (haut) peut être dans un état (gauche) soit intergénique, 5' ou 3'UTR, exon (seules les parties codantes sont considérées, avec trois phases possibles) ou intron (trois phases aussi selon la position du site donneur d'épissage dans l'exon précédant). Les signaux détectés dans la séquence permettent les changements de piste (arcs obliques). Pour simplifier la figure, seules les pistes correspondant au brin sens sont représentées.

1. Un graphe orienté acyclique est d'abord construit pour représenter l'ensemble des prédictions possibles (Fig. 3.1). La taille du graphe est proportionnelle à celle de la séquence en longueur, avec des "pistes" horizontales qui représentent, pour chaque position génomique, les différentes annotations possibles pour le nucléotide correspondant (ou "état" : région intergénique, exon, intron, etc.). Les pistes sont formées par des arcs horizontaux, modélisant une continuité d'état entre deux positions successives, mais des arcs obliques permettent des transitions d'un état à l'autre, par exemple pour passer d'un état exonique à un état intronique au niveau d'un site d'épissage. Le principe est que (1) tout chemin traversant le graphe définit une prédiction de structure génique (et donc une annotation) pour la séquence, et (2) il existe un chemin pour toute annotation possible. Par exemple, un chemin qui ne passe que par la piste intergénique indique une absence totale de gène.

2. Afin d'identifier le meilleur chemin et donc la prédiction la plus probable, des poids sont assignés aux arcs du graphe en fonction de diverse sources d'information qui dépendent de modèles préalablement estimés sur des séquences dont on connaît la nature. Par exemple, des modèles capturant la composition en divers mots (*k-mers*) d'exons et introns déjà annotés permettent de pondérer les arcs d'état, et des logiciels externes de détection de signaux biologiques permettent de pondérer les arcs de transition. L'objectif est de pouvoir attribuer un score global à toute prédiction possible en additionnant le score de tous les arcs composant le chemin correspondant.
3. Enfin, un algorithme de type plus court chemin, variante de l'algorithme de Bellman/Viterbi, identifie le chemin de score maximum (ou de moindre coût, ce qui est équivalent). Grâce à la programmation dynamique, l'algorithme bénéficie d'une complexité linéaire en temps et en espace par rapport à la taille de la séquence.

Une limitation majeure de la méthode, retrouvée dans tous les autres prédicteurs de gènes du même type, résidait dans son incapacité à identifier plus d'une structure optimale par séquence. Par conséquent, on ne pouvait détecter au mieux qu'un seul transcrit par gène.

Extension du modèle pour intégrer l'épissage alternatif

Déjà bien documenté chez l'homme et la souris au début des années 2000, l'épissage alternatif était encore considéré comme un événement relativement peu fréquent chez les plantes (Foissac 2004 ; Foissac et al. 2005). Comme souvent, l'arrivée massive de données "omiques" (ici de transcriptomique) a nécessité une double évolution, à la fois en bioinformatique au niveau des méthodes d'analyse, et en biologie au niveau du modèle général d'expression génique.

> Dans le cadre de ma thèse, j'ai donc choisi de m'attaquer à ce problème en développant une méthode de prédiction de gènes qui prenne en compte la possibilité d'avoir plusieurs transcrits par gène. Ceci impliquait le développement d'une nouvelle méthode d'intégration de données, la modification du modèle existant, la conception d'un algorithme efficace de résolution et l'implémentation logicielle. J'ai réalisé cela en partant du logiciel EuGène, que j'ai dû reprendre en profondeur pour mener à bien ces travaux.

Les étapes sont les suivantes :

Les données disponibles pour détecter la présence de transcrits alternatifs (ou "variants") consistaient en un grand nombre de séquences partielles de transcrits appelées ESTs pour *Expressed Sequence Tags*, similaires dans l'esprit à ce qu'on aurait de nos jours avec du séquençage RNA-seq - à bien moindre quantité et qualité bien sûr. Ces ESTs peuvent être alignés individuellement sur le génome, fournissant ainsi une information partielle sur la structure exon-intron du transcrit correspondant. Deux ESTs venant du même variant donnent des alignements com-

patibles. Pour trouver des variants différents on peut donc comparer les alignements deux à deux pour lister les incompatibilités.

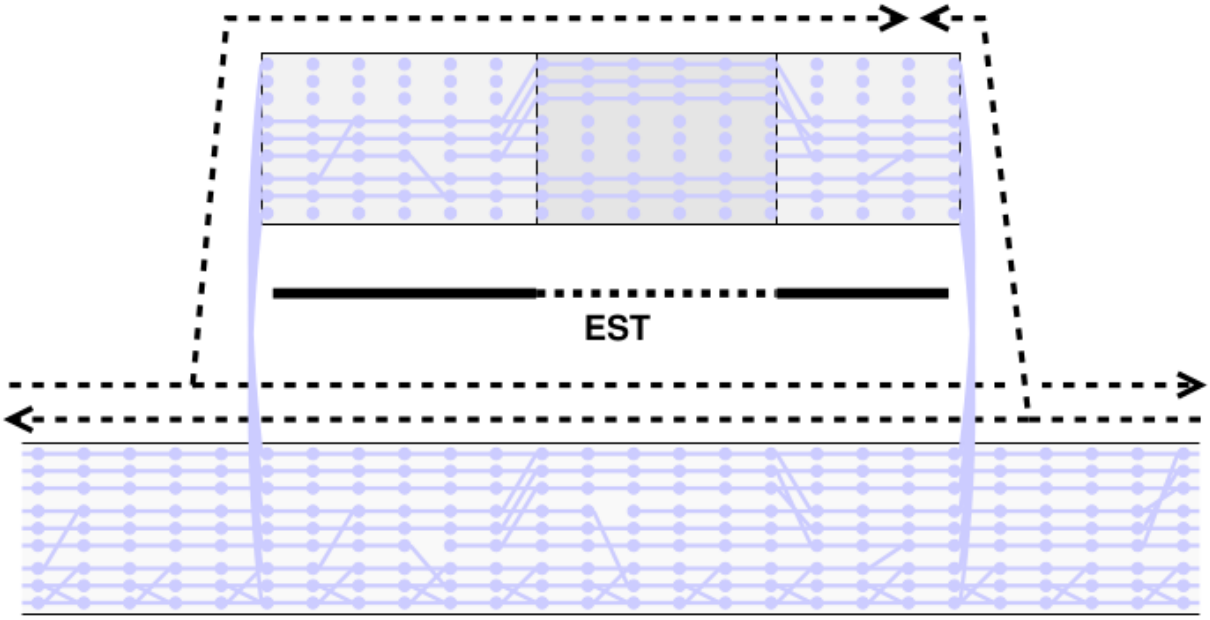

FIGURE 3.2 – Prise en compte de l'épissage alternatif dans EUGÈNE. Le graphe principal (bas) représente les possibles prédictions de structure génique pour la séquence génomique entière. Pour prendre en compte l'information apportée par l'EST aligné (milieu), une déviation locale du graphe (haut) est branchée de part et d'autre de l'alignement. Les arcs de cette déviation parallèle sont modifiés pour forcer toute prédiction qui la traverse à respecter la structure de l'alignement (ici, les pistes exoniques sont retirées le long de l'intron). Le graphe principal est parcouru une fois dans les deux sens et pour chaque EST une jonction est réalisée en reliant les chemins optimaux venant des deux extrémités du graphe principal aux points de branchement de la déviation, produisant la prédiction optimale en accord avec l'EST (Foissac et al. 2005).

La modification du modèle consiste, pour chaque EST, à construire une extension du graphe principal en dupliquant la portion de graphe correspondant à la région de l'alignement, formant ainsi une déviation parallèle locale raccordée par ses extrémités au graphe principal (Fig. 3.2). L'information spécifique de chaque EST est ensuite injectée localement dans l'extension correspondante, ce qui évite de modifier le graphe principal. Le principe est que tout chemin passant dans une extension produit une solution compatible avec l'EST correspondant.

L'algorithme modifié consiste à d'abord parcourir le graphe principal en stockant les résultats intermédiaires du plus court chemin à chaque position [1], puis à parcourir chaque extension sé-

1. plus précisément, deux passes sont nécessaires, une dans le sens "forward" et une dans le sens "reverse", en stockant à chaque position le résultat du plus court chemin provenant du début et de la fin du graphe respectivement

parément. Pour chacune d'entre elles, le résultat du plus court chemin la traversant est connecté aux résultats intermédiaires du graphe principal aux positions des extrémités de l'extension, produisant ainsi l'ensemble des plus courts chemins passant par chaque extension (Fig. 3.2). L'idée sous-jacente est que si C_1 est le plus court chemin entre a et b et C_2 le plus court chemin entre b et c, alors le plus court chemin entre a et c est la jonction $C_1 + C_2$.

En termes de complexité algorithmique, la méthode permet de rester linéaire en temps et en espace mémoire par rapport à la quantité de données en entrée (génome + transcriptome). Plus de détails sont disponibles dans la publication dédiée (Foissac et al. 2005).

Extension du modèle pour intégrer les gènes chevauchants

Une autre extension du modèle général d'expression génique concerne la définition même du concept de gène, en particulier l'idée que les gènes peuvent être considérés comme des entités individuelles contenues dans des intervalles génomiques bornés et distincts.

J'ai commencé me pencher sur cette question pendant ma thèse. En effet, une des améliorations demandées par des utilisateurs du logiciel EuGène concernait l'intégration d'ESTs particuliers correspondant à des extrémités de cDNA dits "pleine taille". L'objectif était de délimiter les unités de transcription et donc les gènes. Pour prendre en compte cette information, il suffisait de forcer les chemins dans le graphe à "entrer" et "sortir" d'un gène aux positions de ces ESTs. Or, sur certaines données, le logiciel retournait une erreur du fait d'une impossibilité de satisfaire les contraintes imposées. La raison était que certains de ces ESTs délimitaient parfois des intervalles chevauchants, l'espace séparant deux gènes adjacents pouvant se réduire à néant, voire atteindre des valeurs négatives en fonction des transcrits considérés. Les limites de certains gènes semblaient donc variables. Pour prendre en compte la possibilité de chevauchement et éviter les erreurs du logiciel, j'ai dû modifier le modèle existant pour ajouter des pistes d'états combinant UTR 5' et 3' à la même position, permettant de passer d'un gène à un autre sans devoir repasser par une piste intergénique intermédiaire.

La version résultante du logiciel EuGène a permis à ses utilisateurs de réaliser une nouvelle annotation du génome d'*A. thaliana* lors du projet CATMA (Crowe 2003).

Conclusion

> Ces travaux ont montré que même sur un génome relativement simple et compact, la prise en compte de données abondantes produites par de nouvelles technologies peut demander de substantielles innovations méthodologiques, voire conceptuelles. En particulier, le modèle général de l'expression génique présentait des limites qu'il fallait dépasser.

c. Annotation v2.0 : vers une nouvelle vision du gène

Effacement des frontières géniques

Fiche Synthétique

- **Lieu** : CRG (Barcelone, Catalogne) et Affymetrix (Santa Clara, USA)
- **Génome** : *Homo sapiens* (humain)
- **Cible** : gènes et transcrits codants et non codants
- **Données traitées** : transcriptomes par puces *tiling arrays*
- **Contribution principale** : analyses de données
- **Publications co-signées** : The ENCODE Project Consortium 2007 ; Denoeud et al. 2007 ; Djebali et al. 2008 ; Djebali et al. 2012b

Au sortir de ma thèse, je souhaitais donc questionner à la lumière de nouvelles données la vision du génome sur laquelle se basait mon enseignement universitaire, qui présentait des gènes bien distincts les uns des autres. Il fallait donc choisir une espèce disposant de données riches. Or, l'espèce la qui bénéficie de la plus grande attention de la communauté scientifique n'est autre que celle-là même de ses membres : *Homo sapiens*[2]. J'ai donc poursuivi mes travaux d'annotation en travaillant sur le génome humain, principalement dans le contexte du consortium international ENCODE, à partir de données de transcriptomique.

Plus précisément, nous analysions principalement des données produites par technologie de puces à ADN, en particulier des *tiling arrays* d'Affymetrix. Chacune de ces puces présentait à sa surface un ensemble de sondes oligonucléotidiques dont les séquences (chevauchantes, d'où le nom des puces évoquant le concept de tuilage) représentaient la plupart des parties non répétées du génome humain. L'originalité de cette technologie est que, contrairement aux puces traditionnelles qui ne pouvaient généralement mesurer que l'expression de transcrits déjà connus, les *tiling arrays* d'Affymetrix permettaient de scanner l'ensemble des séquences génomiques et donc ouvrait une approche exploratoire novatrice. Pour identifier les régions représentées dans le transcriptome d'un échantillon donné, il fallait extraire l'ARN, produire de l'ADNc par transcription reverse[3], et hybrider sur une puce *tiling array*. Le signal que l'on analysait correspondait à une série de positions génomiques avec pour chacune une valeur d'intensité d'hybridation qui reflétait la quantité d'ARN provenant de la région génomique correspondante[4].

2. Le fait que 100% des personnes qui demandent ou attribuent des financements de recherche appartiennent à cette espèce explique certainement pourquoi c'est la plus fournie en données omiques.
3. cette étape explique pourquoi on parle de puces à ADN même pour du transcriptome.
4. contrairement à ce qu'on entend parfois, les expériences de transcriptomique ne donnent généralement

Une succession de sondes avec un signal positif définissait ce que nous appelions un *transfrag*, pour fragment transcrit. Or, bien qu'une partie de ces transfrags correspondait bien aux exons de transcrits connus, la plupart (63%) révélait un signal de transcription dans des régions encore non annotées, augmentant considérablement la portion à considérer comme transcrite du génome (The ENCODE Project Consortium 2007).

Pour explorer plus précisément les frontières des gènes, une technique consistait à réaliser une expérience dite de RACE (*Rapid Amplification of cDNA Ends*) avant hybridation sur *tiling array*. Ceci impliquait une amplification préalable des cDNA par RT-PCR à partir d'amorces ciblant des régions candidates, généralement des exons de gènes connus ou des nouveaux *transfrags*. Le signal obtenu par ces expériences combinant RACE et *tiling array* produisait ce que nous appelions *RACEfrags*[5]. Ces travaux (et d'autres qui ont suivi) révélaient que globalement, plus on cherchait à définir la frontière d'un gène, plus celle-ci semblait s'éloigner, parfois jusqu'à des centaines de Kb, retombant soit dans l'exon d'un gène connu, soit dans une région intronique ou intergénique, brouillant ainsi les frontières entre gènes (The ENCODE Project Consortium 2007 ; Denoeud et al. 2007 ; Djebali et al. 2008 ; Djebali et al. 2012b).

Prévalence des mécanismes alternatifs et du non codant

Fiche Synthétique

- **Lieu** : CRG (Barcelone, Catalogne), Affymetrix (Santa Clara, USA), Integromics (Madrid, Espagne)
- **Génome** : *Homo sapiens* (humain), *Saccharomyces cerevisae* (levure), *Xenopus laevis* (batracien) et autres.
- **Cible** : gènes et transcrits codants et non codants
- **Données traitées** : annotations et transcriptomes
- **Contribution principale** : méthode et analyses de données
- **Publications co-signées** : Foissac et al. 2007 ; Sammeth et al. 2008 ; Foissac et al. 2015 ; Ozsolak et al. 2010 ; Piqué et al. 2008

La multiplicité des nouveaux transcrits découverts soulevait la question de leur biogénèse. La plupart des transcrits contenant plusieurs exons, l'épissage alternatif se révélait un producteur

aucune mesure directe du niveau de transcription des gènes mais mesurent des quantités de transcrits, résultant de la différence entre production et dégradation.

5. du moins officiellement, car il nous arrivait parfois d'utiliser le terme *CRAPfrags* en interne, en particulier dans des moments de stress et de fatigue, peu avant l'heure du *confcall* hebdomadaire où nous devions présenter les derniers résultats d'analyse aux collaborateurs américains du consortium.

primordial de complexité transcriptomique. En collaboration étroite avec un autre postdoc enthousiaste de l'équipe, j'ai poursuivi mes investigations sur l'épissage alternatif, ce qui nous a permis d'aboutir à une nouvelle nomenclature permettant une classification exhaustive des événements d'épissage alternatifs (Foissac et al. 2007 ; Sammeth et al. 2008 ; Foissac et al. 2015).

Autre conséquence de cette "explosion" transcriptomique : la proportion toujours croissante d'ARNs séquencés qui ne codent (probablement) pas pour des protéines. Les gènes non-codant pour des protéines, autrefois considérés comme une exception, ont bénéficié d'une attention croissante de la part de la communauté et comptent aujourd'hui pour la majorité des gènes annotés chez l'humain (Nurk et al. 2022). Une partie de mes travaux s'est inscrite dans cette dynamique (The ENCODE Project Consortium 2007 ; Djebali et al. 2012b ; The ENCODE Project Consortium 2012 ; Djebali et al. 2012a).

Enfin, parmi les processus moléculaires à l'origine de la diversité transcriptomique figure également une des étapes de la maturation de l'ARN en charge de définir l'extrémité terminale en 3'. À cette position, régulièrement –et incorrectement [6]– appelée "terminaison de la transcription", s'opère le clivage de l'ARN en cours de transcription, suivi de la polyadénylation de l'extrémité ainsi générée (l'ajout de la queue polyA faisant partie de l'étape de maturation des transcrits primaires). Des variations de cette position de clivage et polyadénylation, potentiellement combinées avec l'épissage alternatif, contribuent substantiellement à la richesse du transcriptome (Ozsolak et al. 2010 ; Sammeth et al. 2008). Ces variations peuvent avoir de nombreuses conséquences fonctionnelles car les UTR 3' hébergent divers signaux impliqués dans des processus majeurs affectant les ARNm, comme la stabilité, l'export cytoplasmique et la traduction (Ozsolak et al. 2010 ; Piqué et al. 2008).

Annotation de nouveaux petits ARNs

Fiche Synthétique

- **Lieu** : Affymetrix (Santa Clara, USA)
- **Génome** : *Homo sapiens* (humain)
- **Cible** : gènes et transcrits courts et longs, codants et non codants
- **Données traitées** : transcriptomes par RNA-seq
- **Contribution principale** : analyses de données
- **Publications co-signées** : Fejes-Toth et al. 2009

La dernière exploration transcriptomique révisant mes connaissances universitaires fut ap-

6. l'ARN Polymérase poursuit en effet la transcription au-delà du site de clivage.

portée par l'arrivée d'une technologie révolutionnaire : le séquençage à haut débit, dit de nouvelle génération. C'est lors de mon séjour en Californie en 2008 que j'ai manipulé pour la première fois ce type de données, paradoxalement produites par l'entreprise Solexa, compétiteur direct de mon employeur d'alors Affymetrix [7]. Le projet émanait d'une collaboration entre notre groupe, dirigé par Tom Gingeras à Affymetrix, et celui de Greg Hannon, alors au Cold Spring Harbor Laboratory.

> **L'objectif de cette étude** était d'utiliser la nouvelle technologie de séquençage pour caractériser la partie du transcriptome représentée par des "petits" ARNs, généralement définis comme faisant moins de 200nt de long. Suite à un heureux concours de circonstances, j'eus la chance de réaliser la quasi-totalité des analyses bioinformatiques du projet. De nombreux types de petits ARNs ayant déjà été identifiés, il s'agissait d'en découvrir de nouveaux, en adoptant donc une approche exploratoire (Fejes-Toth et al. 2009).

La première étape consistait à retirer des alignements de lectures tout ce qui chevauchait les petits ARNs déjà répertoriés dans les bases de données existantes, afin d'écarter ce qui était susceptible d'être connu. L'analyse des lectures restantes a montré que la plupart provenait de régions géniques du même sens, en majorité d'exons de gènes codants qui produisent des ARN messagers de taille supérieure à 200nt. Une minorité seulement de ces petits ARNs étaient associés aux promoteurs de gènes connus, les autres se répartissant dans des exons internes. Cette accumulation de petits ARNs au sein d'exons de transcrits plus longs suggérait une production par découpage de ces derniers. Or, l'analyse fortuite de CAGE-tags, séquences chargées de capturer l'extrémité 5' des ARNs matures, a révélé un profil de couverture similaire, à l'intérieur d'exons annotés (Fig. 3.3). Les CAGE-tags ciblant le *cap*, une structure particulière des ARNs les protégeant d'une dégradation précoce en extrémité 5', ces résultats suggéraient un ajout de *cap* à une extrémité issue d'un clivage, ce qui non seulement n'avait jamais été documenté, mais en plus impliquait l'intervention d'une protéine encore non identifiée dans le génome humain.

Les analyses que j'ai réalisées dans le cadre de cette étude, comprenant l'estimation de la prévalence du phénomène à l'échelle du génome et l'analyse d'alignements de CAGE-tags sur des jonctions exon-exon, ont permis de mettre en évidence [8] un nouveau mécanisme moléculaire capable de couper des longs ARNs puis de protéger l'extrémité 5' de certains fragments par *recapping* (Fejes-Toth et al. 2009). Le modèle original résultant, qui produit des petits ARNs assez stables dans les cellules pour être détectable par séquençage, est présenté en Fig. 3.4.

7. Solexa fut achetée par Illumina en 2007, et Affymetrix, reine des puces qui n'a pas résisté à la déferlante du séquençage, par Thermo Fisher en 2016.

8. après avoir convaincu les collaborateurs du CSHL, ce qui fut encore plus difficile que de convaincre les reviewers de Nature par la suite.

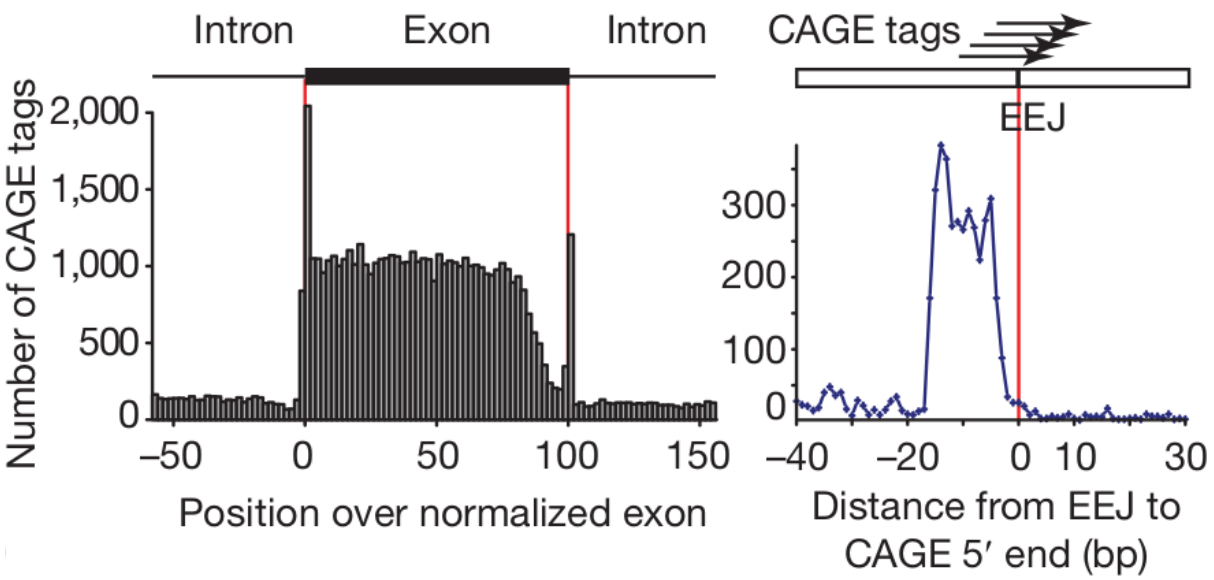

FIGURE 3.3 – Distribution des extrémités 5' de CAGE-tags alignés sur des régions contenant des exons internes de gènes codants (gauche) et des jonctions exon-exon (droite). En ordonnée est indiqué le nombre de lectures obtenues par capture de *cap* CAGE, séquencées et alignées à chaque position relative par rapport aux extrémités d'exons internes. La position relative permet de cumuler les profils de couverture d'exons de différentes longueurs. L'accumulation spécifique de lectures dans les exons suggère une création d'extrémité 5' par clivage plutôt que par initiation de transcription. De plus, le défaut de couverture en fin d'exon s'explique par la difficulté d'aligner sur le génome des lectures hybrides chevauchant deux exons. L'alignement de droite, réalisé à partir des lectures non alignées sur le génome puis réalignées sur des jonctions exon-exon reconstituées à partir d'annotations de référence, montre une accumulation d'alignements commençant précisément en amont des jonctions, confirmant la production d'ARN cellulaires clivés puis recappés en 5' (Fejes-Toth et al. 2009).

Épilogue

En guise de dénouement heureux à cette aventure américaine, il s'avère que peu après la parution de notre étude, un groupe de biochimistes de l'université de l'Ohio publiait un article annonçant la découverte chez l'homme d'un complexe protéique capable de régénérer une coiffe à partir d'une extrémité clivée d'ARNm, confirmant ainsi l'existence du mécanisme de *recapping* que nous avions proposé pour expliquer nos résultats (Otsuka et al. 2009).

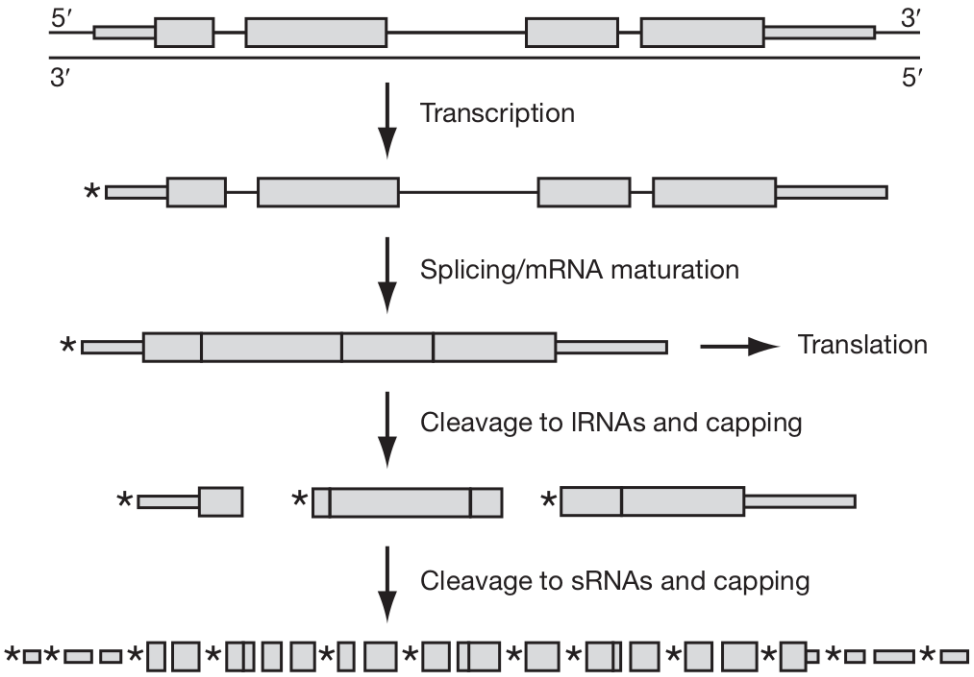

FIGURE 3.4 – Schéma du nouveau processus moléculaire proposé pour expliquer les résultats obtenus par séquençage de petits ARNs et de CAGE-tags. Divers ARNs courts et longs sont produits par clivage d'ARNm, certains étant par la suite protégés en 5' par "recapping" (Fejes-Toth et al. 2009).

d. Annotation des génomes animaux

Fiche Synthétique

- **Lieu** : INRAE Toulouse
- **Génome** : *Bos taurus* (vache), *Gallus gallus* (poule), *Sus scrofa* (porc), *Capra hircus* (chèvre)
- **Cible** : gènes et transcrits, éléments régulateurs
- **Données traitées** : RNA-seq, ATAC-seq, Hi-C, WGBS
- **Contribution principale** : analyses de données, montage et gestion de projet, management, conseil
- **Publications co-signées** : Andersson et al. 2015 ; Muret et al. 2017 ; Foissac et al. 2019a ; Giuffra et al. 2019 ; Jehl et al. 2020 ; Jehl et al. 2021 ; Kurylo et al. 2023a ; Degalez et al. 2024

Depuis ma prise de fonction à l'INRAE, j'ai mis à disposition mes compétences acquises à l'étranger en contribuant à l'annotation génomique d'animaux à intérêt agronomique, poursuivant ainsi mes travaux de recherche dans ce domaine.

- Dans le contexte du consortium international FAANG (*Functional Annotation of ANimal Genomes*), qui se voudrait le pendant d'ENCODE pour les animaux, j'ai co-monté et co-dirigé le projet pilote français FR-AgENCODE qui visait à améliorer l'annotation génomique de quatre espèces : la vache *Bos taurus*, la poule *Gallus gallus*, le porc *Sus scrofa* et la chèvre *Capra hircus*. Combinant des données d'expression génique (RNA-seq), d'accessibilité de la chromatine (ATAC-seq) et de génomique 3D (Hi-C) sur une centaine d'échantillons, ce projet a apporté des connaissances nouvelles sur ces génomes, confirmant le caractère générique des avancées précédemment observées sur le génome humain (Foissac et al. 2019a). Le succès de ce projet pilote a confirmé le positionnement de l'INRAE au coeur de la coordination internationnale FAANG (Andersson et al. 2015 ; Foissac et al. 2019a ; Giuffra et al. 2019).
- En parallèle, par le biais de ces travaux et du réseau de collaborations développé lors de cette initiative, j'ai (plus ou moins directement) contribué à poursuivre la caractérisation du transcriptome non codant de ces espèces, en particulier chez la poule (Muret et al. 2017 ; Foissac et al. 2019a ; Jehl et al. 2020 ; Jehl et al. 2021 ; Degalez et al. 2024).
- Enfin, dans le cadre notamment du projet EU H2020 GENE-SWitCH, une suite du projet pilote FR-AgENCODE, j'ai co-supervisé le développement d'un pipeline d'analyse de

données RNA-seq permettant d'enrichir une annotation existante avec de nouveaux gènes et transcrits (Kurylo et al. 2023a). Outre son implémentation sous la forme d'un pipeline portable et reproductible poétiquement nommé TAGADA (*Transcript And Gene Annotation, Deconvolution and Analysis*), l'originalité de l'approche par rapport aux outils existants réside dans sa capacité à intégrer des résultats provenant de nombreux échantillons dans une annotation globale et robuste, à quantifier l'expression des gènes et des transcrits résultants (en plus de ceux de l'annotation existante), et à détecter les transcrits non-codants. En outre, la flexibilité du pipeline nous a permis d'intégrer des données de séquençage RNA-seq par longues et courtes lectures pour réaliser une annotation jusqu'alors inédite, combinant les avantages des technologies d'Illumina et de PacBio (Kurylo et al. 2023a). Plus largement, le projet a produit divers types de données en quantité qui sont encore en cours d'analyse : expression génique (RNA-seq), accessibilité de la chromatine (ATAC-seq), méthylation de l'ADN (WGBS), marques épigénétiques (ChIP-seq) et interactions chromatiniennes (capture Hi-C), le tout dans divers tissus au cours du développement embryonnaire chez le porc et la poule.

Conclusion : définir le gène et son expression ?

Encore valide et acceptable il y a deux ou trois décennies, le modèle général de l'expression génique eucaryote est clairement dépassé aujourd'hui [9], au point de remettre en question la définition du gène.

> En effet, la notion même de gène semble manifestement difficile –voire impossible– à capturer, comme en atteste la récurrence de publications posant frontalement la question de sa définition depuis le milieu du siècle dernier sans pour autant aboutir à un quelconque consensus (Stadler 1954 ; Demerec 1955 ; Portin 1993 ; Snyder et al. 2003 ; Burian 2004 ; Kapranov et al. 2005 ; Pearson 2006 ; Carninci 2006 ; Pennisi 2007 ; Fox Keller et al. 2007 ; Gerstein et al. 2007 ; Gingeras 2007 ; Pesole 2008 ; Stadler et al. 2009 ; Portin et al. 2017). Parmi les découvertes disruptives motivant un tel questionnement, on peut en citer trois qui compromettent particulièrement le schéma de l'expression génique.

1. **Les mécanismes alternatifs et multiples variantes** des étapes de l'expression génique remettent en question la pertinence du modèle général et du concept historique "un gène ⇒ un ARNm ⇒ une protéine ⇒ une fonction" (Portin et al. 2017).
   - Alors que les variations de début et fin de transcription détruisent la bijection gène-transcrit, l'épissage alternatif fait de même avec la bijection transcrit-protéine. La combinatoire résultant de l'utilisation conjointe de promoteurs, sites d'épissage et

9. en espérant que ce ne soit pas le cas de l'auteur de ces lignes.

sites de polyadénylation alternatifs augmente d'autant plus le degré de ramification du modèle en multipliant les connections entre gène, transcrit, protéine et fonction. De plus, l'évanescence des frontières de gènes, qui fluctuent et s'estompent à mesure des études transcriptomiques réalisées, brouille la notion d'îlots géniques distincts (Wright et al. 2021), et dessine un continuum transcriptomique dans lequel s'inscrivent les ramifications précédemment mentionnées. Il en résulte un réseau transcriptomique global et intriqué, où s'enchevêtrent des transcrits de tailles, orientations, fonctions et voies de production diverses (Fig. 3.5).

- D'autres mécanismes ouvrent des voies alternatives dans le modèle, fragilisant sa structure : le *backsplicing* par exemple, qui génère des ARN circulaires apparemment non codants (Nemeth et al. 2023), le *trans-splicing* qui permet la fusion de transcrits primaires issus de loci génomiques distants pour former un transcrit mature potentiellement codant (Robertson et al. 2007), les gènes à sélénocystéine qui enfreignent le code génétique en ne s'arrêtant pas au stop (Mariotti et al. 2012 ; Belew et al. 2014), le décalage de phase ou *programmed ribosomal frameshifting* (Belew et al. 2014), le *recursive splicing* (Sibley et al. 2015), la copie d'ARN par *RNA-Dependent RNA polymerase* (Kapranov et al. 2010a), l'*RNA editing*, l'épitranscriptomique, ... la littérature regorge de ces mécanismes plus fascinants les uns que les autres.
- Enfin, n'oublions pas que les molécules d'ARN les plus nombreuses de la cellule sont des ARN ribosomaux, qui, bien que servant à la synthèse protéique, ne codent pas eux-même pour des protéines. De fait, les gènes de ces ARNr produisent de longs transcrits primaires polycistroniques dont les parties destinées à devenir des ARNr fonctionnels sont séparées par des régions destinées à la dégradation. Or, le processus de maturation de ces transcrits ne fait pas intervenir d'épissage mais des combinaisons complexes de clivage par réaction endo- et exo-nucléiques (Henras et al. 2014), à l'encontre du modèle prétendu "général" de l'expression génique dont il reste souvent tenu à l'écart.

2. **La transcription généralisée ou *pervasive transcription*** participe également de l'obsolescence du modèle classique de l'expression génique. Il est désormais globalement admis que la plupart sinon la quasi totalité du génome peut se retrouver transcrit en ARN par un phénomène de transcription généralisée non limité aux gènes codant pour des protéines. Ce phénomène, abondamment documenté et confirmé par différentes technologies (Kapranov et al. 2002 ; Bertone et al. 2004 ; Burian 2004 ; Carninci et al. 2005 ; Kapranov et al. 2007a ; Kapranov et al. 2007b ; Dinger et al. 2009 ; Jacquier 2009 ; The ENCODE Project Consortium 2012 ; Jensen et al. 2013 ; Wade et al. 2014 ; Palazzo et al. 2020 ; Goszczynski et al. 2021), a donné naissance au concept de *dark matter* du gé-

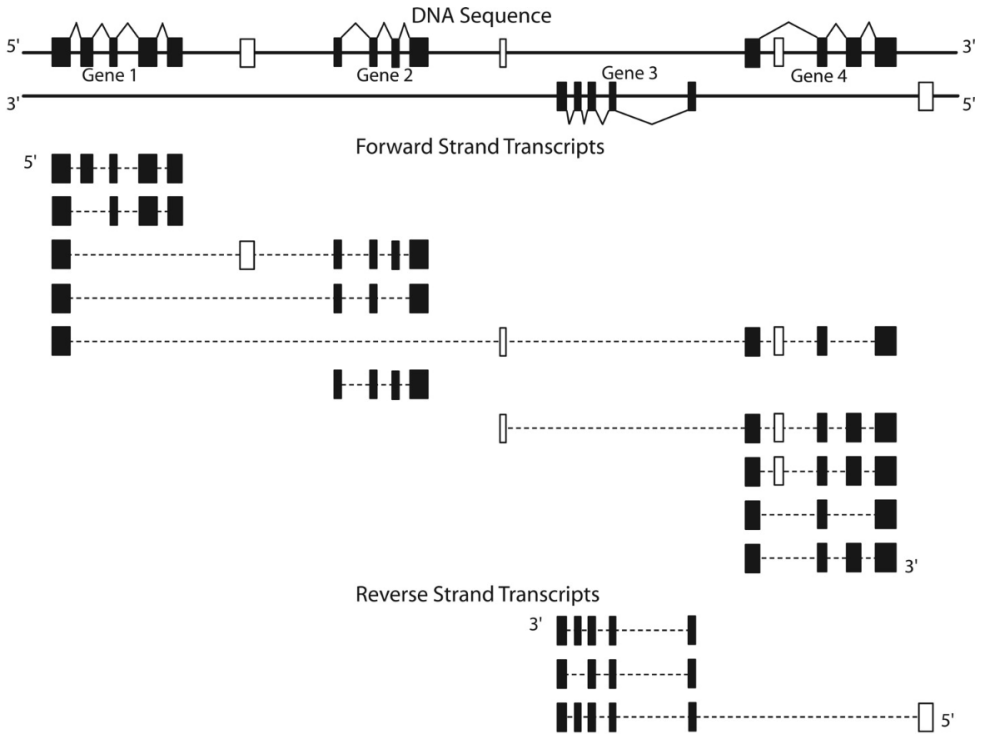

FIGURE 3.5 – Représentation du type de réseau transcriptomique couramment obtenu à partir de données de séquençage profond d'ARN (bas) par rapport aux annotations typiquement utilisées le millénaire dernier (haut). Source : Gerstein et al. 2007.

nome (Johnson et al. 2005 ; Wang et al. 2021) et, conséquemment, à une vive controverse scientifique quant à l'éventuelle prévalence, nouveauté et fonction de cette matière noire. Les débats sur le sujet (Bakel et al. 2010 ; Robinson 2010 ; Lee Phillips 2010 ; Clark et al. 2011 ; Bakel et al. 2011), rejoignant la polémique suscitée par la communication autour des résultats du projet ENCODE et de critiques parfois particulièrement virulentes (Graur et al. 2013), s'articulent autour de la proportion de cette transcription généralisée qui n'est guère plus que du bruit de transcription, ou *transcriptional noise* (Raj et al. 2006 ; Blake et al. 2006 ; Ponjavic et al. 2007 ; Struhl 2007 ; Losick et al. 2008 ; Xu et al. 2009 ; Pertea et al. 2018).

Indépendamment des aspects fonctionnels, il ressort clairement aujourd'hui que l'existence de la transcription généralisée est indéniable. Tout d'abord, en termes quantitatifs, la masse d'ARN nucléaire provenant de régions ne codant pas pour des protéines dépasse celle des ARN codants, même en écartant les ARN ribosomaux et mitochondriaux (Kapranov et al. 2010b). Par ailleurs, en termes de nombre, au niveau génomique, on trouve davantage de gènes non-codants que de gènes codants annotés chez les animaux (Kurylo et al. 2023a), et jusqu'à deux fois plus chez l'homme par exemple (Nurk et al. 2022). En-

fin, qu'ils soient courts, longs ou circulaires, une littérature abondante détaille la diversité des types d'ARN non codants fonctionnels déjà identifiés et des rôles qu'ils jouent dans la cellule, ne permettant plus de les ignorer en bloc sous prétexte d'une prétendue absence de conservation ou de fonction (Mattick et al. 2023 ; Nemeth et al. 2023).

Finalement, au-delà du débat réducteur "*dark matter vs. junk DNA*" ou "*pervasive transcription vs. transcriptional noise*", la question subsiste d'identifier la partie fonctionnelle et informative du transcriptome, si possible à l'aide de méthodes et outils récents (Wang et al. 2021). Reste donc à distinguer le signal du bruit.

3. La dernière révolution en date du modèle d'expression génique vient de la *pervasive translation*, ou traduction généralisée en micropeptides. Des techniques comme le ribosome *profiling*, ou Ribo-seq, permettent de détecter les parties d'ARN qui sont engagées dans les ribosomes, indiquant les régions du transcriptome en cours de traduction dans la cellule (Ingolia et al. 2009 ; Ingolia et al. 2011 ; Lee et al. 2012 ; Schott et al. 2021). Ces technologies ont révélé l'étendue jusqu'alors insoupçonnée de la traduction généralisée de petites phase de lectures *small Open Reading Frames* (sORFs) localisées dans les ARN "non codants" et dans les parties "non traduites" des ARNm. Ainsi, on se retrouve par exemple avec des transcrits bicistroniques, qui ont une phase codante principale et une sORF en 3' ou 5'UTR (cas des *upstream ORFs* ou uORFs).

La traduction de ces sORFs produit des micropeptides, d'un ordre de longueur de la dizaine à la centaine de codons environ. Plusieurs études montrent à quel point leur abondance est considérable, et documentent de multiples exemples de fonctions biologiques qui leur sont attribuées (Frith et al. 2006 ; Andrews et al. 2014 ; Heesch et al. 2019 ; Chen et al. 2020 ; Wei et al. 2020 ; Patraquim et al. 2022 ; Barczak et al. 2023).

Pour résumer, on peut dire que les découvertes récentes des nombreuses ramifications, variantes et sophistications du processus d'expression génique enrichissent considérablement le modèle historique (Fig. 3.6).

Personnellement, il se trouve que j'ai passé plus de vingt ans à chercher des gènes. J'en ai trouvé et annoté des centaines de milliers, dans divers génomes d'animaux et de végétaux, en mobilisant des connaissances en constant renouvellement, à l'aide de multiples technologies en front de science.

Aujourd'hui, je dois avouer que je ne sais toujours pas ce que le mot gène veut dire exactement.

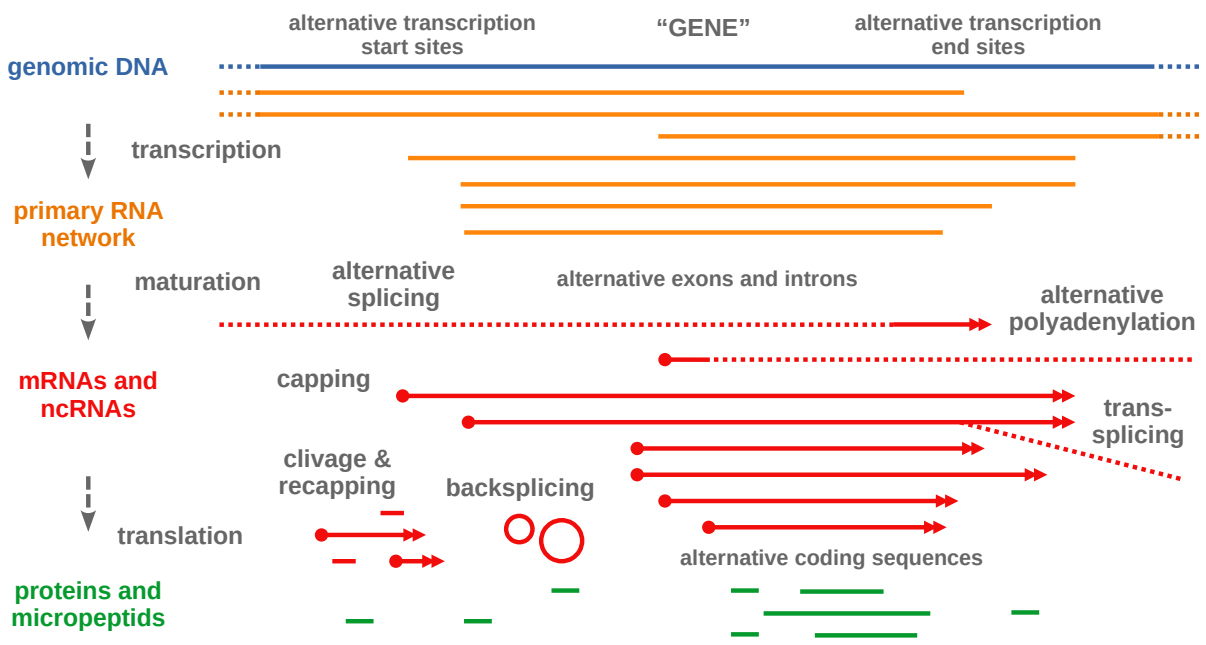

FIGURE 3.6 – Modèle actualisé de l'expression génique eucaryote prenant compte de certains mécanismes alternatifs. Échelle et proportions non valides.

3.2 La génomique 3D : de la structure à la fonction

PRÉAMBULE

Les derniers progrés technologiques qui ont le plus influencé l'orientation de mes recherches concernent l'étude de la conformation tridimensionnelle de la chromatine. En effet, l'arrivée récente de données massives de génomique 3D a apporté une nouvelle dimension à la compréhension du génome et de son fonctionnement, repoussant encore les limites du modèle jusqu'alors principalement linéaire de l'expression génique. Cette partie présente ce nouveau domaine de recherche que j'ai choisi d'investir.

a. La génomique 3D et la technologie Hi-C

La génomique spatiale

La longueur cumulée de l'ADN présent dans une seule cellule animale atteint l'ordre du mètre, quand le noyau qui le contient fait moins d'un centième de millimètre. Or, les contraintes s'exerçant sur la conformation spatiale de la chromatine relèvent autant de la structure que de

la fonction du génome. En effet, les processus moléculaires fondamentaux du vivant, comme la réplication et la transcription, dépendent pour leur bon fonctionnement de modifications coordonnées de la structure de la chromatine. Pas d'expression génique possible sans décompaction de la chromatine sur un promoteur par exemple, ou rapprochement physique d'un régulateur. La structure 3D du génome joue ainsi un rôle crucial dans le fonctionnement de la cellule (Lupiáñez et al. 2015 ; Oudelaar et al. 2020).

Les évolutions technologiques ont fortement influencé l'étude de cette structure 3D (Davies et al. 2017). En permettant de visualiser dans les cellules la position de sondes lumineuses conçues pour s'hybrider sur des régions génomiques d'intérêt, la microscopie fut pendant des décennies la technologie de référence pour explorer la structure spatiale du génome. Avec l'hybridation *in situ* en fluorescence (*FISH*) par exemple, on peut estimer la distance physique entre deux régions génomiques. La nécessité de cibler spécifiquement chaque région d'intérêt et le relatif manque de précision de la mesure optique limitent cependant la résolution et le débit de cette approche. Avec la capacité d'estimer des centaines de milliers de distances entre positions génomiques qui couvrent potentiellement l'ensemble du génome, la technologie Hi-C (*High-throughput Chromatin Conformation Capture*) est devenue depuis son introduction au début des années 2000 une nouvelle référence pour la génomique 3D à haut-débit et haute résolution (Belmont 2014 ; Bonev et al. 2016 ; Davies et al. 2017 ; Kempfer et al. 2019).

Le protocole Hi-C

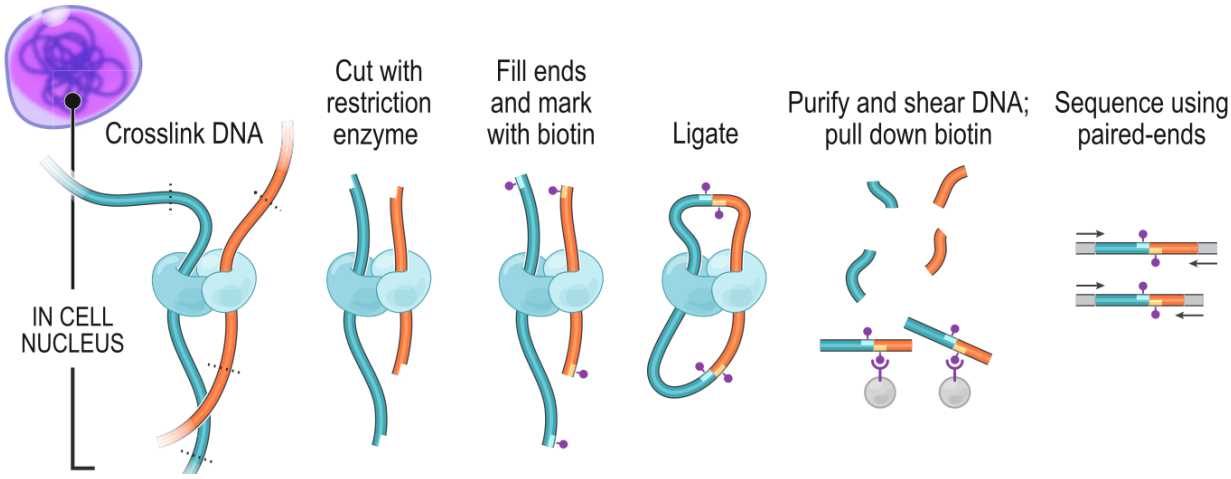

FIGURE 3.7 – Schéma du protocole expérimental de la technologie Hi-C, version *in situ* (Rao et al. 2014). Les différentes étapes sont détaillées dans le corps du texte.

En bref, le protocole Hi-C comprend les étapes suivantes (Fig. 3.7) :

1. Stabilisation de la structure chromatinienne présente dans un échantillon de cellules.

2. Fragmentation de l'ADN, souvent par enzyme(s) de restriction, afin de créer des extrémités libres réparties sur l'ensemble du génome.
3. Marquage des ces extrémités à la biotine pour permettre une récupération ultérieure.
4. Ajout de ligase pour permettre à des fragments d'ADN spatialement proches les uns des autres de se lier entre eux par leurs extrémités libres (*proximity-dependant ligation*, voir plus bas).
5. Récupération et séquençage dit *Paired-End* des deux extrémités de chaque fragment hybride, produisant des paires de lectures permettant d'identifier les séquences précédemment liées.

> **Le principe fondamental de la technologie Hi-C, appelé *proximity-dependant ligation*, repose sur le fait que la probabilité de générer une liaison entre deux fragments est inversement corrélée à la distance spatiale qui les sépare**. Autrement dit pour schématiser, plus deux régions génomiques sont proches, plus elles risquent de se retrouver associées dans un même fragment hybride, dont le séquençage des extrémités permet d'identifier l'association. En réalisant l'expérience sur un grand nombre de cellules, typiquement des centaines de milliers, on peut ainsi utiliser le nombre de paires de lectures associant deux régions génomiques comme une estimation (ou *proxy*) de la distance spatiale entre ces régions dans le noyau des cellules.

Note : contrairement à ce qu'une interprétation approximative de la méthode laisse trop souvent croire, l'Hi-C ne donne strictement parlant aucune mesure directe sur d'éventuels contacts ou interactions physiques réelles entre régions génomiques *in vivo*. En effet, ne sont comptées que des liaisons artificiellement générées au cours de l'expérience lors de l'étape de *proximity-dependant ligation*.

b. Annotation de structures 3D par analyse de données Hi-C

Des paires de lectures aux matrices Hi-C

Le pipeline d'analyse classique de données Hi-C part des lectures produites qu'il faut d'abord "nettoyer" et aligner sur une séquence génomique de référence. Ensuite, une matrice d'interactions est constituée pour comptabiliser le nombre d'associations observées entre régions génomiques. Pour cela, on utilise le génome pour les deux axes de la matrice, en le segmentant en intervalles (ou *bins*) afin de former une grille, dont la largeur de maille définit la résolution de l'analyse : plus les bins sont petits, plus la granularité est fine et la résolution élevée. Chaque élément de la matrice contient le nombre de paires de lectures associant les deux intervalles génomiques correspondants, que l'on peut représenter par une carte d'interactions comme en

Fig. 3.8. A partir des mêmes données de séquençage on peut utiliser plusieurs tailles de bins pour produire des matrices et donc des cartes de différentes résolutions, sachant que les résolutions fines demandent beaucoup de lectures et donc de grandes profondeurs de séquençage pour être informatives.

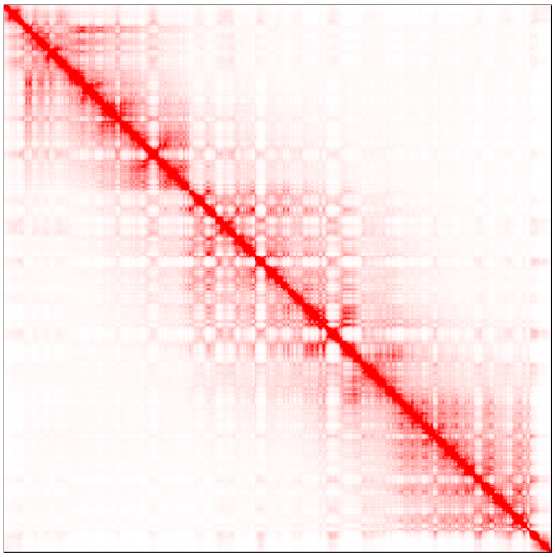

FIGURE 3.8 – Carte Hi-C représentant les interactions entre régions de 500 Kb du chromosome 1 dans un échantillon de foie porcin (Foissac et al. 2019a).

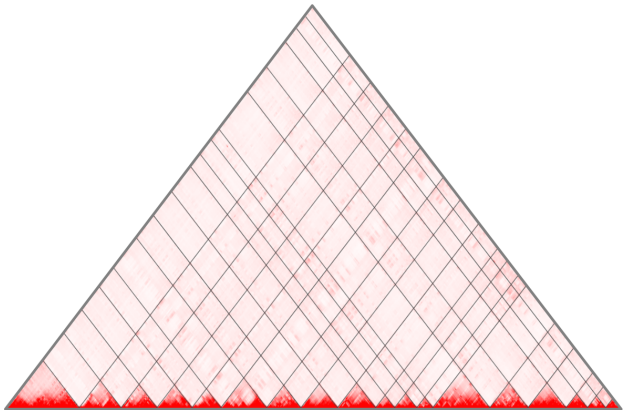

FIGURE 3.9 – Carte Hi-C tout génome à 500 Kb de résolution, réalisée à partir d'échantillons de muscle porcin à 90 jours de développement, avec les chromosome alignés en bas de gauche à droite (de 1 à 18 autosomes). Les territoires chromosomiques contribuent à la prévalence des interactions intra-chromosomes (Marti-Marimon et al. 2021).

Quelques propriétés générales des matrices Hi-C (Fig. 3.8) :

- la matrice est symétrique, car les interactions ne sont pas orientées. Il est d'ailleurs courant de n'en représenter qu'une moitié, sous la forme d'un triangle isocèle dont la base est la diagonale de la matrice (Fig. 3.9).
- la plupart du signal est concentré le long de la diagonale, car les régions adjacentes le long du génome sont forcément proches en 3D : la distance linéaire borne la distance spatiale.
- globalement, l'intensité du signal diminue en s'éloignant de la diagonale, faisant de la distance génomique le déterminant principal du nombre d'interactions.
- les matrices sont souvent creuses (*sparse*), avec une majorité de comptages nuls ou correspondant à du bruit de fond, et ce d'autant plus que la résolution est fine.

Les structures identifiables

Quatre principaux types de structure chromatinienne se distinguent dans les données Hi-C, à différents niveaux d'échelle (Fig. 3.10) (Bonev et al. 2016 ; Mota-Gómez et al. 2019 ; Oudelaar

et al. 2020).

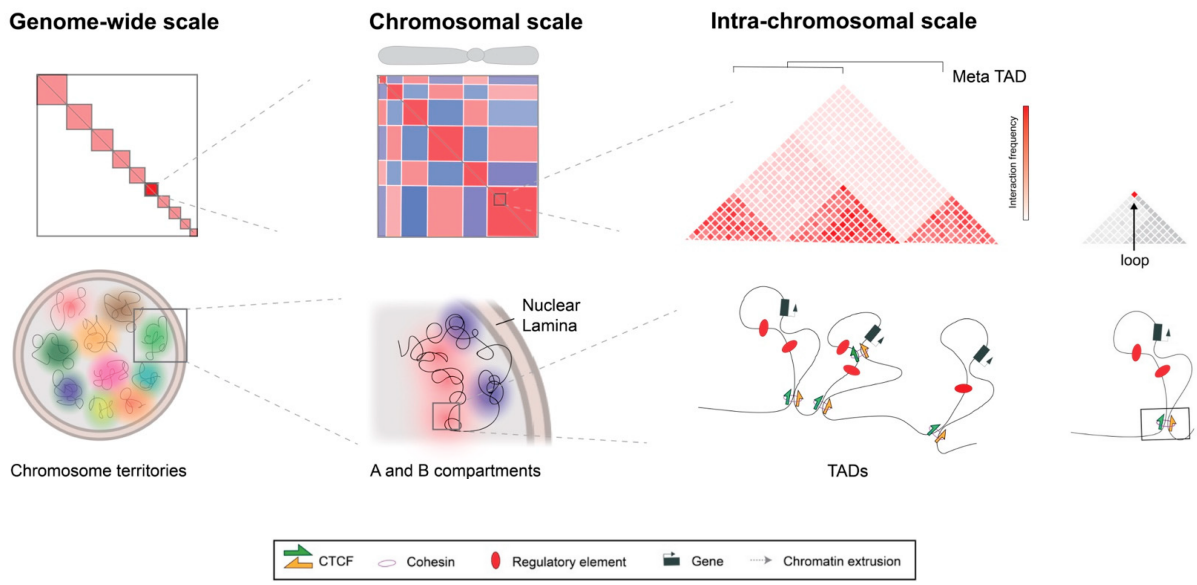

FIGURE 3.10 – Structures typiques de la chromatine à différents niveaux d'échelle. De gauche à droite : territoires chromosomiques, compartiments chromatiniens, domaines topologiques (TADs) et boucles de chromatine. Source : Mota-Gómez et al. 2019.

- **Les territoires chromosomiques**. Dans les cartes Hi-C représentant le génome entier, les chromosomes délimitent typiquement des blocs visibles d'interactions plus fréquentes en leur sein (Fig. 3.9). La proportion d'interactions intra-chromosomiques varie en fonction de l'espèce, du tissu et des facteurs expérimentaux.
- **Les compartiments chromatiniens**. Deux principaux types complémentaires de profil d'interactions semblent émerger par position génomique, formant parfois des bandes régulières alternant des régions de fortes et faibles valeurs. L'alternance et le contraste entre ces bandes forment des motifs de damier, particulièrement visibles si l'on retire l'effet de la distance génomique sur les valeurs par une normalisation dite *observed/expected* qui prend en compte l'éloignement par rapport à la diagonale (Fig. 3.10 et 3.12). Ces deux types de profil segmentent le génome de façon bipartite en délimitant des compartiments dont la taille atteint l'ordre du Mb. Ces compartiments traduisent une séparation de phase physique de la chromatine qui peut se trouver dans deux état principaux : compartiment "A" pour la chromatine plutôt ouverte, accessible et active en transcription, et compartiment "B" pour la chromatine plus compacte et moins transcrite globalement

(Fig. 3.10).

- **Les TADs (*Topologically Associating Domains*)**. Les TADs sont des régions du génome au sein desquelles les interactions sont plus fréquentes que vers l'extérieur. Avec des tailles variant entre le Kb et le Mb, ils forment des carrés visibles le long de la diagonale de cartes Hi-C entières, et des triangles à la base de cartes triangulaires (Fig. 3.10 et 3.11). Délimités par des frontières riches en sites de fixation de la protéine CTCF, ils hébergent d'importantes relations de régulation d'expression génique entre activateurs et promoteurs.
- **Les boucles de chromatine**. Certains mécanismes nécessitent la proximité physique de deux positions génomiques distantes linéairement. Par exemple, la transcription d'un gène peut être déclenchée par le rapprochement entre son promoteur et un site de régulation sur lequel est fixé un facteur de transcription (Fig. 3.10). Un tel rapprochement de régions précises, appelé boucle, peut se détecter sur des cartes à résolution fine (idéalement de l'ordre du Kb par bin) par une tache locale en dehors de la diagonale (Fig. 3.10 et 3.13).

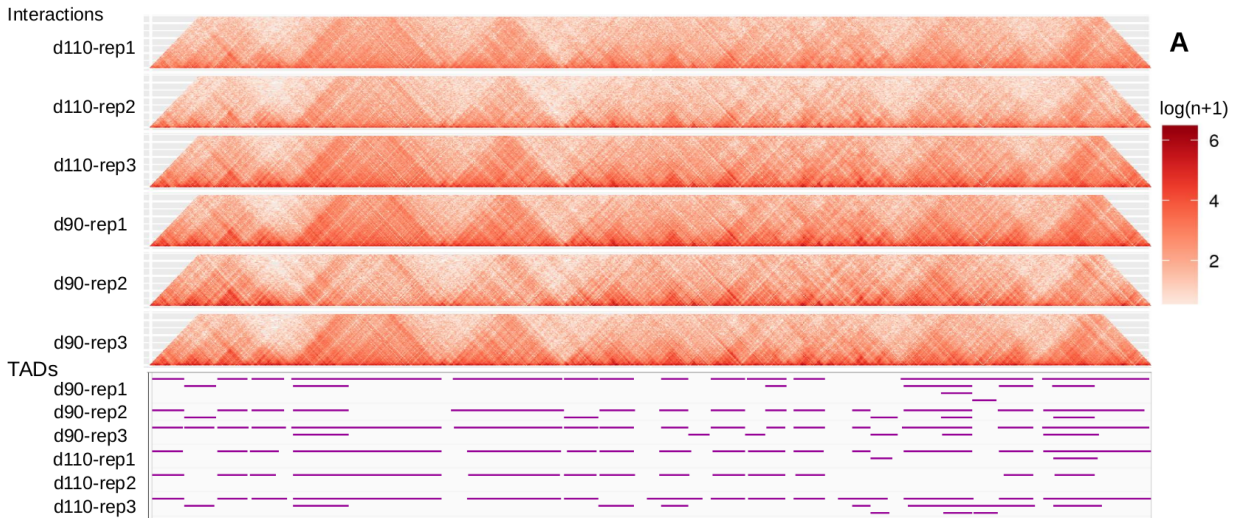

FIGURE 3.11 – Annotation de domaines topologiques TADs prédits sur des données Hi-C obtenues à partir d'échantillons de muscle porcin à deux stades de développement (90 jours et 110 jours post-fertilisation), avec trois répliques biologiques par stade. Haut : cartes Hi-C, chromosome 1, région (165-200 Mb). Bas : prédictions de TADs pour chaque échantillon. Les TADs (bas) correspondent généralement aux triangles (haut). Source : Marti-Marimon et al. 2021.

Pour chacun de ces types de structure (sauf les territoires chromosomaux qui sont évidents) des méthodes et outils logiciels ont été développés pour permettre leur détection à partir de données Hi-C, avec plus ou moins de succès (Dali et al. 2017 ; Zufferey et al. 2018 ; Marti-Marimon et al. 2021 ; Liu et al. 2023).

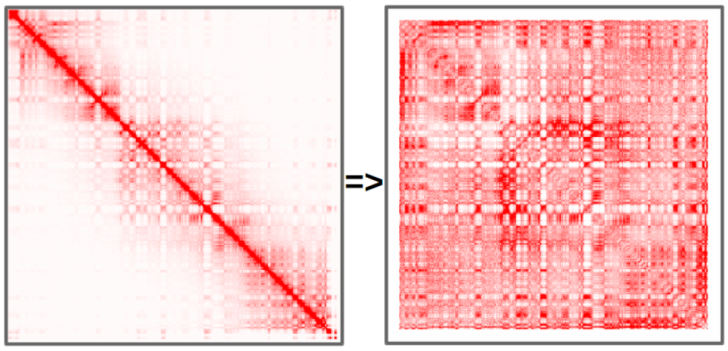

FIGURE 3.12 – Compartimentation visible de la chromatine dans les données Hi-C. Gauche : carte Hi-C non normalisée de foie porcin, chromosome 1, résolution 500 Kb. Droite : matrice normalisée par la distance linéaire pour atténuer l'effet de la diagonale (*observed/expected*). L'alternance de régions complémentaires indique les transitions entre compartiments "A" et "B". Source : Foissac et al. 2019a.

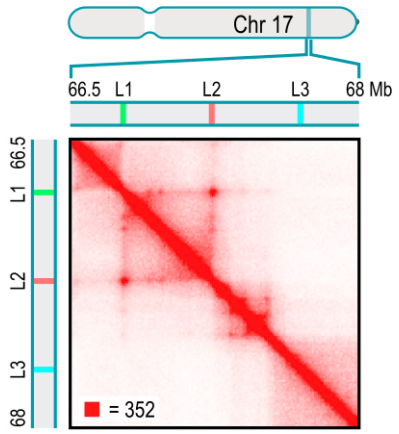

FIGURE 3.13 – Boucle de chromatine visible sur des données Hi-C humaines (lignée cellulaire GM12878, résolution 5 Kb), la tache foncée indiquant un rapprochement entre les positions L1 et L2. Source : Rao et al. 2014.

Application aux génomes animaux

La problématique de la structure chromatinienne ayant été abordée lors du projet ENCODE humain, c'est assez naturellement que, suite à mon intégration à l'INRAE, j'ai souhaité apporter cette nouvelle dimension à l'annotation des génomes animaux. En particulier, lors du montage du projet pilote FR-AgENCODE en 2014 pour l'annotation de quatre espèces (porc, poule, vache et chèvre), nous avons intégré un volet de caractérisation structurale par Hi-C, bien que le nombre de laboratoires réalisant ce type d'expérience dans le monde se comptait sur les doigts de la main à l'époque. C'est donc dans le cadre de ce projet qu'ont été réalisées les premières expériences Hi-C à l'INRAE, expériences qui se font désormais en routine à la plateforme de séquençage. Ces travaux ont permis la première caractérisation de la structure 3D du génome de ces espèces dans le cadre du consortium international FAANG (Foissac et al. 2019a).

Puis, dans le cadre de la première thèse que j'ai co-encadrée à l'INRAE (2016-2019), nous avons entrepris d'apporter une dimension "tout-génome" à la caractérisation d'interactions 3D par des expériences Hi-C sur des échantillons de muscle foetal porcin. Ces travaux ont apporté de nouvelles connaissances sur le rôle de la conformation chromatinienne pendant la maturation, étape du développement embryonnaire porcin qui joue un rôle critique dans la mortalité néonatale des porcelets (Marti-Marimon et al. 2021). Cette étude a également permis de lancer une dynamique d'animation scientifique autour du thème de la génomique 3D, notamment par

le biais d'un réseau de collaborations interdisciplinaire encore actif à ce jour [10].

c. Vers l'analyse différentielle de données Hi-C

Afin d'identifier les liens entre structure et fonction du génome, une approche classique consiste à comparer les structures présentes dans des échantillons cellulaires venant de groupes qui diffèrent par un facteur d'intérêt : tissu d'origine, état de santé, condition expérimentale, etc. L'objectif est d'associer à ce facteur une différence significative de conformation de la chromatine, comme on le fait dans le cas des analyses différentielles d'expression génique par exemple. La particularité ici est la même que pour les méthodes de détection : elles diffèrent en fonction du type de structure considéré. Autre distinction majeure : alors que de nombreuses méthodes et outils existent pour l'analyse différentielle de données RNA-seq par exemple, ce n'est pas le cas pour l'Hi-C, méthode plus récente et surtout plus onéreuse. En effet, du fait de la nature matricielle des interactions, le nombre de lectures à séquencer pour atteindre une résolution donnée augmente globalement avec le carré de celle-ci. Il en résulte un nombre généralement très faible de répliques expérimentales réalisées par condition, la plupart des études se limitant à des duplicats, souvent fusionnés d'ailleurs pour maximiser la résolution de l'analyse. Dès lors, tant le besoin que la possibilité de méthodes statistiquement pertinentes ont longtemps manqué. Pour autant, la réduction constante du coût de séquençage fait évoluer la situation, ouvrant la voie au développement et à l'utilisation de méthodes adaptées.

Ainsi, par le biais de collaborations interdisciplinaires, je me suis investi ces dernières années dans le développement de méthodes différentielles à deux niveaux d'échelle essentiellement : au niveau des TADs et des compartiments de la chromatine.

Projet Treediff

L'objectif du projet est d'identifier des différences de structure 3D du génome entre deux groupes d'échantillons cellulaires à partir de matrices Hi-C les caractérisant. En bref, la première version de la méthode réalise un clustering hiérarchique des éléments de la diagonale pour transformer chaque matrice en un arbre. Un test statistique faisant intervenir les distances entre paires de feuilles –distances cophénétiques– et une agrégation de p-valeurs permet ensuite de statuer sur l'égalité entre groupes d'un sous-arbre donné.

La méthode, abordée lors d'une précédente thèse que j'ai co-encadrée (Randriamihamison 2021), a fait l'objet d'un stage M2, d'un package R [11], d'une publication (Neuvial et al. 2024), et d'une nouvelle thèse en cours pour laquelle je suis déclaré comme encadrant principal [12].

10. Assez joliment nommé Chrocogen
11. `TreeDiff` : https://cran.r-project.org/web/packages/treediff/index.html
12. afin de m'obliger à enfin passer l'HDR.

Projet HiCDOC

Le projet HiCDOC vise à détecter les compartiments A et B de la chromatine dans des échantillons pour lesquels on dispose de matrices Hi-C. Il s'agit en particulier de développer une méthode d'analyse comparative qui permette de détecter des différences significatives de compartimentation entre groupes de matrices avec répliques.

Contrairement à la plupart des autres outils qui utilisent la méthode classique de la PCA (*Principal Component Analysis*) pour détecter les compartiments, comme initialement proposé par Lieberman-Aiden et al. 2009, HiCDOC adopte une approche originale basée sur du clustering des régions génomiques par k-means contraint, prenant en compte les répliques de chaque condition. La méthode fait l'objet d'un package R Bioconductor [13] et d'un article en préparation.

3.3 Perspectives

a. Vers l'annotation fonctionnelle, translationnelle et régulationnelle

La plupart de l'annotation réalisée dans le cadre de mes activités de recherche a consisté à l'analyse de données omiques pour réaliser une cartographie d'éléments génomiques d'intérêt. Il s'agit cependant d'un processus principalement unidirectionnel, incomplet, statique et descriptif.

- Unidirectionnel car si les connaissances apportées par les découvertes successives de nouveaux mécanismes moléculaires servent régulièrement à alimenter les processus d'annotation par la création de nouveaux pipelines ou par l'amélioration d'existants, l'annotation en soi n'apporte que trop rarement de nouvelles connaissances biologiques.
- Incomplet car même en comptant sur la coordination internationale et une réduction constante du coût de séquençage, on peut difficilement imaginer que l'annotation des animaux d'élevage bénéficie pour chaque génome d'un investissement comparable à celui qui est fait pour l'homme. Il n'y aura pas l'équivalent d'un projet ENCODE pour chaque espèce.
- Statique et descriptif car si la multiplication des couches d'annotation (omiques mesurées, cellules ciblées, conditions expérimentales...) peut permettre la construction de réseaux de gènes ou régulateurs, seules des corrélations peuvent en émerger, sans apporter de lien de causalité.

13. https://bioconductor.org/packages/release/bioc/html/HiCDOC.html

Dans la continuité de mes travaux passés et en cours, une voie de recherche que je souhaite explorer pour dépasser ces limitations consiste à utiliser les annotations produites pour développer une approche intégrative multi-omiques et multi-espèces visant deux objectifs :

1. Inférer des annotations afin de compléter d'inévitables données manquantes sur certains échantillons, types cellulaires et espèces qui ne font pas l'objet d'expériences omiques. En effet, le transfert d'annotation entre tissus, espèces et types de données omiques est un enjeu majeur pour l'avenir de la génomique animale. Le développement d'approches de type *machine learning* représente une piste prioritaire pour la prédiction d'annotations. Les modèles d'intelligence artificielle évoluant, la possibilité additionnelle de récupérer des informations biologiques pertinentes capturées pendant la phase d'apprentissage ouvre des perspectives intéressantes de valoriser les résultats en complétant le prédictif par de l'explicatif.
2. Utiliser les annotations pour produire de la connaissance sur les mécanismes moléculaires biologiques à l'oeuvre dans les cellules. L'objectif est de modéliser les réseaux de régulation d'expression génique afin de proposer des hypothèses de relations régulatrices et des régions candidates à tester de façon expérimentale dans des systèmes cellulaires rapporteurs. Les avancées technologiques tels que l'édition de génome et la culture d'organoïdes offrent des possibilités considérables pour passer d'une position d'observation et de description de l'organisation du génome à une position plus active où l'on teste expérimentalement son fonctionnement, révélant des relations de causalité dans les mécanismes à l'oeuvre. L'annotation informe sur l'anatomie du génome, la génomique fonctionnelle sur sa physiologie.

Ainsi, à partir des annotations produites par la communauté, il serait possible de les compléter, d'en tirer de l'information biologique sur les éléments explicatifs, de proposer des relations fonctionnelles candidates à tester et, idéalement, de permettre d'intégrer les résultats expérimentaux pour améliorer les modèles et alimenter des échanges entre modélisation et expérimentation.

En termes de moyens, le contexte local est particulièrement favorable au développement de ces approches. En effet, la restructuration de notre unité m'a justement permis de monter une nouvelle équipe sur la base du projet combinant la multi-omiques multi-espèces pour améliorer la connaissance des mécanismes moléculaires responsables de la régulation de l'expression génique animale. Après avoir ainsi créé l'équipe REGLISS (***RE**gulatory **G**enomics for* **LI***vestock **S**pecie**S***) [14], j'ai pris la responsabilité du nouveau Pôle de Génomique Structurale et Fonctionnelle, qui a la charge de la prospective scientifique des trois équipes qui le composent.

14. Le nom initialement proposé de **GE**nétique **M**o**LE**culaire et **C**ellulaire de l'**EX**pression (GEMLECEX) fut abandonné pour des raisons de phonétique.

Ce rôle d'animateur scientifique encourage les échanges entre collègues d'expertises et de compétences diverses, ce qui résonne particulièrement avec mon enthousiasme et mon goût pour la transdisciplinarité.

b. Méthodes d'analyse de données de génomique 3D

Dans le domaine de la génomique 3D les thèmes de recherche à développer ne manquent pas. Ma priorité se porte sur la poursuite des approches comparatives pour la détection de différences significatives de structure 3D entre groupes d'échantillons. L'objectif est en premier lieu de finaliser les travaux sur les différents niveaux d'échelle abordés (projets TreeDiff et HiCDOC en particulier).

Dans un second temps, il s'agira simplement de profiter des progrès réalisés sur les aspects méthodologiques pour les appliquer aux animaux d'intérêt agronomiques, en profitant des données produites dans le contexte du consortium international FAANG (*Functional Annotation of ANimal Genomes*). Dans ce domaine également l'approche multi-espèces représente un potentiel intéressant pour apporter une dimension évolutive à l'exploration des mécanismes moléculaires responsables de la relation entre structure et fonction des génomes. Les approches de génomique fonctionnelle mentionnées ci-dessus ouvrent des pistes de recherche pour aborder les aspects de causalité à l'œuvre dans ces liens structure-fonction.

En termes de moyens, ce thème bénéficie d'une dynamique tout à fait réjouissante. Le sympathique réseau Chrocogen que j'ai la chance d'animer offre un cadre propice d'échange, de discussion et de collaboration entre scientifiques d'expertises et affiliations très diverses. J'encourage d'ailleurs toute personne intéressée à visiter notre page Chrocogen pour consulter le programme et s'inscrire à la chrocoliste de diffusion pour s'informer des prochains chrocotalks.

3.4 Productions scientifiques

Fiche Synthétique

- **Total publication :** 31 *peer-reviewed* articles, 3 chapitres de livres.
- **Authorship :** 7 articles signés en (co-)premier auteur, 3 en dernier ou *corresponding* auteur (les 3 dans les cinq dernières années).
- **Domaine :** essentiellement biologie (moléculaire/cellulaire) et bioinformatique.
- **Journaux les plus fréquents :** 4 articles dans *Nature*, 2 dans *Cell*, *Nucleic Acids Research*, *Scientific Reports*, *Frontiers in Genetics*, *PLoS ONE*, 1 dans *Nature Methods*,

Genome Research, *PLoS Computational Biology*, *BMC Biology*, etc.

- **Liens et accès pdf :** http://genoweb.toulouse.inra.fr/~sfoissac/website/publications.html.

a. Articles publiés dans des revues internationales à comité de lecture

Andersson L, Archibald AL, Bottema CD, Brauning R, Burgess SC, Burt DW, Casas E, Cheng HH, Clarke L, Couldrey C et al. (2015). "Coordinated international action to accelerate genome-to-phenome with FAANG, the Functional Annotation of Animal Genomes project". In : *Genome Biology* 16.1, p. 1-6. ISSN : 1474760X. DOI : 10.1186/s13059-015-0622-4.

Bonnet A, Cabau C, Bouchez O, Sarry J, Marsaud N, Foissac S, Woloszyn F, Mulsant P et Mandon-Pepin B (2013). "An overview of gene expression dynamics during early ovarian folliculogenesis : Specificity of follicular compartments and bi-directional dialog". In : *BMC Genomics* 14.1, p. 1-19. ISSN : 14712164. DOI : 10.1186/1471-2164-14-904.

David SA, Mersch M, Foissac S, Collin A, Pitel F et Coustham V (2017). "Genome-wide epigenetic studies in chicken : A review". In : *Epigenomes* 1.3, p. 20. ISSN : 20754655. DOI : 10.3390/epigenomes1030020.

Degalez F, Charles M, Foissac S, Zhou H, Guan D, Fang L, Klopp C, Allain C, Lagoutte L, Lecerf F et al. (2024). "Enriched atlas of lncRNA and protein-coding genes for the GRCg7b chicken assembly and its functional annotation across 47 tissues". In : *Scientific Reports* 14.1, p. 6588. DOI : 10.1038/s41598-024-56705-y.

Denoeud F, Kapranov P, Ucla C, Frankish A, Castelo R, Drenkow J, Lagarde J, Alioto T, Manzano C, Chrast J et al. (2007). "Prominent use of distal 5' transcription start sites and discovery of a large number of additional exons in ENCODE regions". In : *Genome Research* 17.6, p. 746-759. ISSN : 10889051. DOI : 10.1101/gr.5660607.

Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F et al. (2012a). "Landscape of transcription in human cells". In : *Nature* 489.7414, p. 101-108. ISSN : 1476-4687. DOI : 10.1038/nature11233.

Djebali S, Kapranov P, Foissac S, Lagarde J, Reymond A, Ucla C, Wyss C, Drenkow J, Dumais E, Murray RR et al. (2008). "Efficient targeted transcript discovery via array-based normalization of RACE libraries". In : *Nature Methods* 5.7, p. 629-635. ISSN : 15487091. DOI : 10.1038/nmeth.1216.

Djebali S, Lagarde J, Kapranov P, Lacroix V, Borel C, Mudge JM, Howald C, Foissac S, Ucla C, Chrast J et al. (2012b). "Evidence for transcript networks composed of chimeric rnas in human cells". In : *PLoS ONE* 7.1, e28213. ISSN : 19326203. DOI : 10.1371/journal.pone.0028213.

Dufour A, Kurylo C, Stöckl JB, Laloë D, Bailly Y, Manceau P, Martins F, Turhan AG, Ferchaud S, Pain B et al. (2024). "Cell specification and functional interactions in the pig blastocyst inferred from single-cell transcriptomics and uterine fluids proteomics". In : *Genomics* 116.2, p. 110780. ISSN : 0888-7543. DOI : 10.1016/j.ygeno.2023.110780.

Fejes-Toth K, Sotirova V, Sachidanandam R, Assaf G, Hannon GJ, Kapranov P, Foissac S, Willingham AT, Duttagupta R, Dumais E et al. (2009). "Post-transcriptional processing generates a diversity of 5'-modified long and short RNAs". In : *Nature* 457.7232, p. 1028-1032. ISSN : 00280836. DOI : 10.1038/nature07759.

Fève K, Foissac S, Pinton A, Mompart F, Esquerré D, Faraut T, Yerle M et Riquet J (2017). "Identification of a t(3 ;4)(p1.3 ;q1.5) translocation breakpoint in pigs using somatic cell hybrid mapping and high-resolution mate-pair sequencing". In : *PLoS ONE* 12.11, p. 1-17. ISSN : 19326203. DOI : 10.1371/journal.pone.0187617.

Foissac S, Bardou P, Moisan A, Cros MJ et Schiex T (2003). "EUGÈNE'HOM : A generic similarity-based gene finder using multiple homologous sequences". In : *Nucleic Acids Research* 31.13, p. 3742-3745. ISSN : 03051048. DOI : 10.1093/nar/gkg586.

Foissac S, Djebali S, Munyard K, Vialaneix N, Rau A, Muret K, Esquerré D, Zytnicki M, Derrien T, Bardou P et al. (2019a). "Multi-species annotation of transcriptome and chromatin structure in domesticated animals". In : *BMC Biology* 17.1, p. 1-25. ISSN : 17417007. DOI : 10.1186/s12915-019-0726-5.

Foissac S, Gouzy J, Rombauts S, Mathe C, Amselem J, Sterck L, Peer Y de, Rouze P et Schiex T (2008). "Genome Annotation in Plants and Fungi : EuGene as a Model Platform". In : *Current Bioinformatics* 3.2, p. 87-97. ISSN : 15748936. DOI : 10.2174/157489308784340702.

Foissac S et Sammeth M (2007). "ASTALAVISTA : dynamic and flexible analysis of alternative splicing events in custom gene datasets". In : *Nucleic acids research* 35.suppl_2, W297-W299. DOI : 10.1093/nar/gkm311.

Foissac S et Schiex T (2005). "Integrating alternative splicing detection into gene prediction". In : *BMC Bioinformatics* 6.1, p. 1-10. ISSN : 14712105. DOI : 10.1186/1471-2105-6-25.

Giuffra E, Tuggle CK et the FAANG Consortium (2019). "Functional Annotation of Animal Genomes (FAANG) : Current Achievements and Roadmap". In : *Annual Review of Animal Biosciences* 7, p. 65-88. ISSN : 21658110. DOI : 10.1146/annurev-animal-020518-114913.

Hoellinger T, Mestre C, Aschard H, Le Goff W, Foissac S, Faraut T et Djebali S (2023). "Enhancer/gene relationships : need for more reliable genome-wide reference sets". In : *Frontiers in Bioinformatics* 3, p. 1092853. DOI : 10.1093/nar/gkx920.

Jehl F, Degalez F, Bernard M, Lecerf F, Lagoutte L, Désert C, Coulée M, Bouchez O, Leroux S, Abasht B et al. (2021). "RNA-Seq Data for Reliable SNP Detection and Genotype Calling : Interest for Coding Variant Characterization and Cis-Regulation Analysis by Allele-Specific

Expression in Livestock Species". In : *Frontiers in Genetics* 12, p. 1104. ISSN : 16648021. DOI : 10.3389/fgene.2021.655707.

Jehl F, Muret K, Bernard M, Boutin M, Lagoutte L, Désert C, Dehais P, Esquerré D, Acloque H, Giuffra E et al. (2020). "An integrative atlas of chicken long non-coding genes and their annotations across 25 tissues". In : *Scientific reports* 10.1, p. 20457. DOI : 10.1038/s41598-020-77586-x.

Kapranov P, Ozsolak F, Kim SW, Foissac S, Lipson D, Hart C, Roels S, Borel C, Antonarakis SE, Monaghan AP et al. (2010a). "New class of gene-termini-associated human RNAs suggests a novel RNA copying mechanism". In : *Nature* 466.7306, p. 642-646. ISSN : 14764687. DOI : 10.1038/nature09190.

Kurylo C, Guyomar C, Foissac S et Djebali S (2023a). "TAGADA : a scalable pipeline to improve genome annotations with RNA-seq data". In : *NAR Genomics And Bioinformatics* 5.4, lqad089. ISSN : 2631-9268. DOI : 10.1093/nargab/lqad089.

Marti-Marimon M, Vialaneix N, Lahbib-Mansais Y, Zytnicki M, Camut S, Robelin D, Yerle-Bouissou M et Foissac S (2021). "Major Reorganization of Chromosome Conformation During Muscle Development in Pig". In : *Frontiers in Genetics* 12, p. 1895. ISSN : 16648021. DOI : 10.3389/fgene.2021.748239.

Muret K, Klopp C, Wucher V, Esquerré D, Legeai F, Lecerf F, Désert C, Boutin M, Jehl F, Acloque H et al. (2017). "Long noncoding RNA repertoire in chicken liver and adipose tissue". In : *Genetics Selection Evolution* 49.1, p. 1-17. ISSN : 12979686. DOI : 10.1186/s12711-016-0275-0.

Neuvial P, Randriamihamison N, Chavent M, Foissac S et Vialaneix N (2024). "A two-sample tree-based test for hierarchically organized genomic signals". In : *Journal of the Royal Statistical Society Series C : Applied Statistics*, qlae011. DOI : 10.1093/jrsssc/qlae011.

Ozsolak F, Kapranov P, Foissac S, Kim SW, Fishilevich E, Monaghan AP, John B et Milos PM (2010). "Comprehensive polyadenylation site maps in yeast and human reveal pervasive alternative polyadenylation". In : *Cell* 143.6, p. 1018-1029. ISSN : 00928674. DOI : 10.1016/j.cell.2010.11.020.

Piqué M, López JM, Foissac S, Guigó R et Méndez R (2008). "A Combinatorial Code for CPE-Mediated Translational Control". In : *Cell* 132.3, p. 434-448. ISSN : 00928674. DOI : 10.1016/j.cell.2007.12.038.

Rubio-Peña K, Fontrodona L, Aristizábal-Corrales D, Torres S, Cornes E, García-Rodríguez FJ, Serrat X, González-Knowles D, Foissac S, Porta-De-La-Riva M et al. (2015). "Modeling of autosomal-dominant retinitis pigmentosa in Caenorhabditis elegans uncovers a nexus between global impaired functioning of certain splicing factors and cell type-specific apoptosis". In : *RNA* 21.12, p. 2119-2131. ISSN : 14699001. DOI : 10.1261/rna.053397.115.

Sammeth M, Foissac S et Guigó R (2008). "A general definition and nomenclature for alternative splicing events". In : *PLoS Computational Biology* 4.8, e1000147. ISSN : 1553734X. DOI : 10.1371/journal.pcbi.1000147.

The ENCODE Project Consortium (2007). "Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project". In : *Nature* 447.7146, p. 799-816. ISSN : 0028-0836. DOI : 10.1038/nature0587.

— (2012). "An integrated encyclopedia of DNA elements in the human genome". In : *Nature* 489.7414, p. 57-74. ISSN : 14764687. DOI : 10.1038/nature11247.

b. Chapitres d'ouvrage

Djebali S, Wucher V, Foissac S, Hitte C, Corre E et Derrien T (2017). "Bioinformatics pipeline for transcriptome sequencing analysis". In : *Methods in Molecular Biology*. T. 1468. Springer Protocols, p. 201-219. DOI : 10.1007/978-1-4939-4035-6_14.

Foissac S et Sammeth M (2015). "Analysis of alternative splicing events in custom gene datasets by AStalavista". In : *Methods in Molecular Biology*. T. 1269. Springer Protocols, p. 379-392. DOI : 10.1007/978-1-4939-2291-8_24.

Neuvial P, Foissac S et Vialaneix N (2023). "Comprendre l'organisation spatiale de l'ADN à l'aide de la statistique". In : *L'Interdisciplinarité*. T. 1. CNRS Editions, p. 172-179.

c. Communications dans des conférences

Quelques exemples de communications orales et/ou poster.

Acloque H, Harrison P, Lakhal W, Martin F, Archibald A, Beinat M, Davey M, Djebali S, Foissac S, Guizard S et al. (2022). "Extensive functional genomics information from early developmental time points for pig and chicken (Talk)". In : *Proceedings of 12th World Congress on Genetics Applied to Livestock Production (WCGALP)*. Wageningen Academic Publishers, p. 2281-2284. DOI : 10.3920/978-90-8686-940-4_550.

Degalez F et al. (2023). "A lncRNA gene-enriched atlas for GRCg7b chicken genome and its functional annotation across 47 tissues (Poster)". In : *ISAG*. Cape Town, South Africa.

Djebali S, Foissac S, Vialaneix N, Munyard K, Rau A, Faraut T, Lagarrigue S, Acloque H et Giuffra E (2019). "Chromatin accessibility conservation across four livestock species (Talk)". In : *Inproceedings of the International Society for Animal Genetics (ISAG 2019)* (7-12 juill. 2019). Llieda, Spain.

Foissac S (2019). "Functional annotation of livestock genomes : chromatin structure and regulation of gene expression (Invited talk)". In : *Journal of Animal Science (Proceedings of the ASAS/ASDS Midwest Joint Meeting)*. T. 97. Suppl. 2. Omaha, NE, USA, p. 15-16. DOI : 10.1093/jas/skz122.028.

— (2022). "TAGADA : Transcripts and Genes Assembly, Deconvolution, Analysis ; a Nextflow pipeline to improve genome annotations with RNA-seq data (Invited talk)". In : PRBB-CRG. Barcelona, Catalunya.

Foissac S, Djebali S, Vialaneix N, Zytnicki M, Rau A, Lagarrigue S, Acloque H et Giuffra E (2019b). "Multi-level conservation of chromosome conformation across livestock species reveals evolutionary links between genome structure and function (Talk)". In : *Inproceedings of the International Society for Animal Genetics (ISAG 2019)* (7-12 juill. 2019). Llieda, Spain.

Kurylo C, Maigné E, Foissac S et Zytnicki M (2023b). "Prediction and differential analysis of chromatin compartments from Hi-C data (Poster)". In : *ISMB/ECCB*. Lyon, France.

Lahbib-Mansais Y, Marti-Marimon M, Vialaneix N, Foissac S, Bouissou-Matet M et Liaubet L (2019). "Organisation nucléaire et expression génique lors du développement chez le porc". In : *Séminaire du réseau EpiPHASE* (26-27 juin 2019). Castanet-Tolosan, France.

Randriamihamison N, Chavent M, Foissac S, Vialaneix N et Neuvial P (2020). "Classification ascendante hiérarchique sous contrainte de contigüité pour l'analyse différentielle de données Hi-C". In : *Journées de Statistique de la SFdS (volume exceptionnel)*.

d. Jeux de données

Je travaille le plus possible dans un esprit de science ouverte, dans le cadre de projets qui produisent des données de type FAIR. Pour les données omiques, le principe est de rendre les séquences publiques au plus tôt après séquençage, avec comme seule restriction d'usage de ne pas publier avant le groupe producteur des données.

Par exemple, les données produites dans le cadre du projet GENE-SWitCH sont déposées dans des registres publics structurés, riches en métadonnées, accessibles sur le portail du *FAANG Data Center* à l'EMBL/EBI :

- Séquences : https://data.faang.org/dataset?standard=FAANG&project=GENE-SWitCH.
- Résultats d'analyse : https://data.faang.org/analysis?standard=FAANG&project=GENE-SWitCH.

Pour plus d'information sur la politique FAIR de données du consortium FAANG : https://www.faang.org/data-share-principle et https://data.faang.org/home.

J'utilise également des ressources nationales au besoin, comme par exemple le repository INRAE Omics Dataverse, hébergé par recherche.data.gouv.fr :

- Companion dataverse pour "Major reorganization of chromosome conformation during muscle development in pig" : https://doi.org/10.15454/DOMEHB
- Companion dataverse pour "TAGADA : a scalable pipeline to improve genome annotations with RNA-seq data"" : https://doi.org/10.57745/3UGLXW

Ces dépôts "dataverse" diffusent en libre accès (FAIR) des informations liées à un projet et/ou une publication : données de type *Supplementary Files* ou ne pouvant être hébergées par l'éditeur, lien vers toutes les données relatives au projet (séquences, métadata, résultats), preprint éventuel de l'article, scripts non disponibles ailleurs, etc.

Eventuellement, d'autres informations peuvent être mises à disposition par d'autres canaux, comme par exemple par le site fragencode https://www.fragencode.org/.

e. Logiciels

En terme de développement logiciel, mes contributions ont considérablement évolué au cours de ma carrière. Pendant ma thèse il y a vingt ans, j'ai beaucoup contribué au code du logiciel EuGène (http://eugene.toulouse.inra.fr/). Depuis, mes contributions relèvent de deux formes : soit le modelage artisanal de scripts "à l'ancienne" parfois monolignes et souvent peu montrables pour les analyses exploratoires (bash, awk, perl), soit en accompagnant ou supervisant le développement de logiciels codés par des personnes plus compétentes que moi. Mis à part mon passage dans le privé, où j'étais devenu responsable de plusieurs produits logiciels commerciaux, je ne me suis impliqué que dans des productions *open-source*. En particulier, depuis mon recrutement à l'INRAE, j'ai contribué au développements de plusieurs logiciels d'analyse de données omiques :

- TAGADA : un logiciel d'analyse de données RNA-seq pour l'annotation génomique sous forme d'un pipeline open source encapsulé (Kurylo et al. 2023a). https://github.com/FAANG/analysis-TAGADA#readme
- HiCDOC : un package Bioconductor (R) pour l'analyse de données de génomique 3D de type Hi-C visant à identifier des différences significatives de compartimentation chromatinienne entre conditions biologiques d'intérêt. https://bioconductor.org/packages/release/bioc/html/HiCDOC.html
- Treediff : un package R du CRAN pour l'analyse de données de génomique 3D de type Hi-C visant à détecter des différences significatives entre groupes de dendrogrammes

associés à des conditions biologiques d'intérêt. https://cran.r-project.org/web/packages/treediff/index.html

Remerciements

Merci !

- Merci Christine, Claire, Marie-France, Patricia, Pitou et Vincent pour avoir accepté de faire partie de mon jury d'HDR. Pour les trois premières, merci d'avoir accepté de rapporter mon manuscrit, et merci de l'avoir fait d'une si belle façon.
- Merci aux personnes que j'ai déjà remerciées dans la (déjà) longue partie dédiée de ma thèse, partie qui est probablement la plus lue, réussie et intéressante du manuscrit – avec peut-être un peu l'état de l'art. Merci encore Thomas de m'avoir lancé dans le monde de la recherche et de m'en avoir appris autant, concernant les aspects scientifiques, pratiques et éthiques. Merci aussi de m'avoir aidé au-delà de la thèse.
- Merci aux collègues de la fabuleuse époque de Barcelone.

  Merci tout d'abord Roderic de m'avoir recruté en post-doc et d'avoir toujours renouvelé mon contrat, surtout au terme d'une première année particulièrement difficile sur le plan personnel. Merci de m'avoir donné une seconde chance. Merci aussi pour ton *management*, ta tolérance, ton engagement, ta vision de la science. Merci de nous avoir fourni un environnement confortable, abrité des conflits des hautes sphères, propice aux échanges, à la solidarité, à l'émulation, aux collaborations (en tous genres), à la stimulation intellectuelle et au travail de qualité. Merci pour tout ce que j'ai appris à ton contact. Merci aussi de nous avoir couverts lors du fameux incident diplomatique des *kisses* malencontreusement transmis par erreur à l'éditrice américaine d'un prestigieux journal de génomique, épisode qui aurait pu sérieusement compromettre la suite de ma carrière scientifique. Merci enfin – et bravo – pour ce que tu as accompli avec le groupe auquel j'ai eu la chance d'appartenir, car c'est exceptionnel. Je ne connais pas d'autre *group leader* ayant réussi à instaurer et maintenir une telle relation avec ses post-docs, PhD students, et autres *alumni* sur une période couvrant près de deux décennies pour qu'une quarantaine d'entre elles/eux s'organisent et se retrouvent depuis plusieurs pays et continents pour une fête surprise à l'occasion de ses 60 ans. *Moltes gràcies, i visca Catalunya* !

  France, merci beaucoup pour ton aide si précieuse – voire cruciale – à mon arrivée, ta franchise, ton entrain, ton humour, ta patience à mon égard et les longues discussions

tard le soir au labo. Merci de m'avoir ramassé. La meilleure fausse couverture de Nature affichera toujours ton visage.

Julien, merci pour tout, pro et perso. Merci d'avoir assuré aussi bien et longtemps sur ENCODE, pour les analyses, le support, les échanges, les scripts, l'inspection des données, les contacts. Merci pour tes nombreuses contributions plus ou moins volontaires à la page *Geek Of The Week* sur le wiki du labo. Merci de m'avoir souvent aidé à garder le cap. Merci pour les fous rires. Merci pour ton attitude, merci d'être et de rester un amour. Gros bisou tio, datte d'autocorrect et à très bientôt.

Hey Micha ! Thanks a lot for everything, all the projects and the parties, the fun, the jokes, the nights in the lab spent writing, coding, testing, decoding ("woof !"), the fun, the laughs, the little smileys still hiding in our posters and articles, the fun, the quotes I still use (*"No risk, no fun !"*, *"I see colors !"*, *"Sant Hilari, sant hilari"*...), the astalavista server goodies, the fun, the Miss Splicing poster, the amazing energy in everything, the C/Princesa parties, the flyers, the cervezabiramigos, the kisses to all the authors, the fun, the best new year party ever ending up as Batman & Robin at the police station, the fun... Did I mention the fun ? So much work and coffee, so many beers and crazy stories !... Asta, compañero !

Christoforos, thanks for all the great moments in your company. Not much work together actually, which is not bad *per se* you would agree. Looking forward to our next meeting/call/gathering... even if it is about science. Thanks for the inspiration, the knowledge, for your generosity, your humility and your panache !

Thanks to the locals/old ones from the IMIM era for the assistance and the friendly spirit : Robert, Mar, Nuria, Eduardo, "gff2ps" Pep, Enrique, Oscar, Montse and the others. Thanks Jan-Jaap for teaching me so many diverse things (how to escape from vi, to curse in Dutch, to shoot a cat with insuline...), merci Arnaud pour les brèves, gracias Luis B. por el regalito que aun tengo conmigo. Thanks to the old ones from the CRG era for sharing the workspace and good moments : "U12-splicing" Tyler, "who broke the cluster" Hagen, "Benfica" Francisco, "so where is Charles" Charles, "intern" Sonja, "oh my god" Ania, "crapfrags" Vincent, and the other "Old farts from Guigo's lab". Thanks Romina for the precious assistance to all of us. Thanks to the newer/younger ones, to the nextflow team (Björn, Jose, Cédric, etc), the PRBB people (Filipe & Laura, Raik, Silvia, Sarah, Megan, Mireya), all the collaborators (including Juan and the splice girls), the many I forgot the names. Gracias Maria por esta primera colaboracion y por todo lo que te debo, incluso lo poco que pudo aprender de español o de motivos CPE en 3'UTR, y por los preciosos momentos contigo. Espero ver a tu sonrisa otra vez algun dia. Thanks to Colin and Rory for the GAE musketeers hiking trips with Julien and me, although there is less

and less hiking and more and more food, drinks, taxi, "cancer pizza", cava police, C* & B*, n*scape, sh*ms and other * since that marvellous Aiguestortes' *carros de foc* – and in particular since that 30km endless first day in Menorca. Thanks to all the friends from that time for the good moments, including Pauline, Eleanna, Valentina, Marta, Fabio and Laia.

- Côté Californien, merci aux collègues d'Affymetrix. Thanks Tom for hiring me, for trusting me, and for sharing your vision and your passion. Phil, thanks a lot for all the collaborations, the discussions about the analyses, the results. Thanks for bringing me along the Helicos data analysis, post-Affy. That was so intense ! I will always remember the night we got that email from Fatih starting by "Paper is accepted. Congratulations to all". We have done quite some work together, over many years and workplaces, including Barcelona, Santa Clara, New York, Boston, Madrid... maybe Toulouse and Xiamen one day ! Thanks to all the nice ENCODE collaborators around the world, at the UCSC, EBI or RIKEN. And thanks a lot Lila for welcoming me in the Bay Area, for assisting me to get all set up, for bearing with my awkward humour and for sharing so much !
- Vient la saga Integromics... Muchas gracias a tod@s los compis de Tres Cantos y de Granada ! Primero a los jefes : Jose-Maria, Alberto, y el Gran Pajaro Marco, jefe de los pajaros ! Ah, los pajaros... nuestro traductor universal y catador de leche Quino, siempre listo para sacar un chiste, Henrique que se lo veia y se lo sabia todo sobre tod@s, pero que nunca podia pedir los postres con fresa que queria por miedo de una cancion (pobrecito !), el pajarito Roberto que un dia consigo aprenderle a Quino un poco de Chino para hablar con una camarera – la unica idioma que no conocia ! Pajaros, nunca me olvidare el falso email que hicimos a la cliente de Ramon "Madame Lederc" con la maravillosa traduccion castellano/frances de Quino. Nunca tampoco me olvidare la comida esta – al Encuentro era ? –, cuando el viejo nos dejo para "un" chupito la botella de aguardiente de su pueblo que acabamos para volver totalmente borrachos en la ofi al mediodia. Como se dice : *"Chupito, chupito, ..."*). Hilmar, – y no "Gilmar" –, gracias por tu ayuda, tu mente, tu personalidad, por salvarnos tantas veces, tambien por ser tan buena persona y por conseguir esta mezcla unica de acento aleman y andaluz. Patricia, gracias perchita por la complicidad, la energia y la locura compartida. Gracias a Eduardo, Tatiana, Patricio, Isabel, Ramon, Mariana, Imad y los demas de la ofi por el animo, la solidaridad y el espirito positivo incluso en los momentos dificiles. Gracias tambien a los de Granada, estos campeones, especialmente Juan, Ale, Miguel, Jose y Manu. Thanks to Jonathan, Gordy and the US team, to the UK team and to the Bulgarian team.
- Enfin, merci aux collègues de l'INRA, version "Science & Impact" puis INRAE.

  En premier lieu merci Thomas pour m'avoir fait (re-)venir, remis en piste vers la recherche

et lancé dans mon nouvel environnement. Merci pour ton humour (si si, c'est du premier degré !), pour ton indulgence envers – entre autres – mon démarrage laborieux et pour ton aide régulière. Merci aux membres actuels et passés de l'équipe DYNAGEN pour l'accueil et les échanges. Merci en particulier David pour la camaraderie, pour les bons plans de sortie (en falaise, en ville ou ailleurs), les bons moments hors boulot et pour le soutien dans l'adversité.

Sarah, merci pour le beau morceau de carrière partagé. Plus de quinze ans de collaboration, de la Catalogne à l'Occitanie ! Tant de bioinfo, de pipelines, de données, de formatage, de parsing, de production et analyse de résultats, de rédaction d'articles, de remplissage de "XX (XX%)" dans des drafts, de *conference calls* puis visios. Quinze ans de discussions, d'idées, d'annotations, de gènes, de transcrits, d'exons, de codes, de scripts, de graphes, de tableaux, de nombres, de pourcentages, de p-valeurs, de catégories, classes, slides... sans compter les combinaisons : par exemple, les discussions en visio sur des scripts destinés à formater des résultats de pipeline d'analyse de données pour faire des slides avec des tableaux contenant des nombres et des pourcentages de catégories de gènes et transcrits. Ah ça, c'est sûr qu'il faut s'accrocher parfois, mais le pire c'est qu'on se comprend ! Je nous souhaite encore beaucoup de collaborations aussi productives qu'agréables. Merci Sarah.

Merci ensuite Hervé, pour ton esprit d'initiative, ta compétence et ces échanges sympas et stimulants. Merci d'avoir été assez fou pour lancer l'Hi-C à l'INRA avec moi – sans toi je n'en serais pas à la 3D ! –, de m'avoir donné le répondant qu'il me fallait côté "wet lab" et la confiance d'oser des projets (parfois trop ?) ambitieux scientifiquement. On a bien assuré dans l'ensemble, et l'Xtra-seq restera sûrement mon plus bel échec ;-) Merci chef !

Un énorme merci à Nathalie, toutes versions inclues : NV^2, NV, Lisbeth Salander, Tuxette, "*woman with a mission*" – pardon pour ces mauvais guillemets –, Nic et les autres. Nath, merci d'avoir réussi à m'apprendre des notions de stat (oui sans -s, car on dit LA Statistique... j'aurai au moins retenu ça !). Cet exploit en dit long sur ta pédagogie et ta patience, dont il te faut d'ailleurs si souvent faire preuve pour supporter mon penchant pour l'économie d'effort et la procrastination, mes excès d'enthousiasme, mon bavardage débordant, mes innombrables questions – même en rando ! – et ma tendance à râler régulièrement : que ce soit sur Linux (plus beaucoup), LaTeX (pas mal), et surtout sur R (*#RCgalR* !). Plus sérieusement, merci pour ton aide cruciale et déterminante dans quasiment tout ce que j'ai entrepris depuis plus de 5 ans, que ce soit la fin de ma conversion au 100% linux, mon dossier CR1 ou tous les projets au bout desquels j'ignore comment j'aurais pu arriver sans toi (FR-AgENCODE et pig3Dgenome par exemple). Merci donc par exemple pour les analyses de données, les conseils experts en analyse de données, les

explications de méthodes, les rédactions/relectures (articles, dossiers, formulaires, etc), l'assistance technique, les co-encadrements, le sauvetage en procédures administratives insurmontables pour moi, la gestion de deadlines, le suivi des démarches, les relances d'emails, les chrocotalks, la contribution à l'essor de la dynamique Hi-C (aspects méthodos et autres), l'aide sur cette HDR même (les encouragements sans pression, les astuces en LaTeX et la gestion intérimaire du mini-symposium JOBIM entre autres), les innombrables coups de main en R donc (les scripts que je pirate comme un cochon avec/dans du code bash ou awk, les installations et mises à jour de packages/librairies, la maintenance du Renv fragencode, etc), en sysadmin, en git, etc. Et tout ça en plus de tout le reste que tu gères en parallèle ! Comment c'est possible ? MystR. Merci aussi pour ta confiance, ton intégrité, ta générosité, ta sensibilité, ton efficacité, et tout le reste. Merci de m'avoir sauvé : je te dois une de mes vies. J'espère rester longtemps "ton gentil bioinformaticien". Et ttccc.

Toute ma gratitude à monsieur Matthias, camarade de plus de 20 ans sur qui le temps – entre autres – ne semble avoir prise. Principalement camarade de jeu dans une ancienne vie (sans revenir sur la période de thèse un peu barrée avec flag/touch-rugby ou diplomacy, on peut mentionner l'époque postdoc avec la transmission d'appart à Barcelone et surtout l'épique voyage en ascenseur avec vue panoramique du prestigieux hall d'hôtel à Sitges – j'espère un jour retrouver cette photo mytique, qui valait bien que je passasse la fin de nuit sur un banc de gare !) et plus largement aujourd'hui, ta discrétion et ton humilité ne sauraient ombrager la valeur des nombreuses contributions dont je te suis gré, de fragencode aux projets Hi-C. Merci Matthias pour ton aide et ta patience, peut-être pourrai-je un jour te rendre la pareille. Et merci de m'avoir expliqué, ce jour où l'on s'est croisés dans le couloir juste avant mon passage pour l'audition du concours CR2, que je n'avais pas à présenter de projet de recherche : peut-être n'aurais-je pas eu le poste sinon.

Côté FR-AgENCODE, merci encore à Sarah, Nath, Matthias, Hervé et Thomas. Merci Elisabetta de m'avoir donné cette occasion d'investir cette thématique et d'accéder à FAANG, malgré le prix à payer. Merci Sandy pour les échanges sympas, le boulot que tu abats et les autres collaborations sur les lncRNAs avec tes *boys*. Andrea, Thomas D., merci pour votre expertise et votre bonne humeur, en visio ou IRL. Thanks Kylie for your visit and your kindness, it was sweet having you around. Sans pouvoir les nommer toustes, merci aux collègues de fragencode qui ont aidé à faire avancer les choses et surmonter les difficultés. Merci aux collègues qui se sont attribué la conception et l'implémentation du "*FRAGENCODE project*" (que j'ai pourtant co-monté et co-piloté) dans l'article valorisant son volet ressources tissulaires sans m'avoir ni cité ni même informé, pour ce que cette ommission révèle d'une part, et, par contraste, pour la belle mise en valeur

des qualités professionnelles et humaines des autres collègues avec qui j'ai la chance de travailler.

En poursuivant dans l'annotation, merci Cervin pour ta venue salvatrice dans GENESWitCH, cruciale pour la thématique, et fort agréable de surcroit. Un vrai plaisir de bosser et échanger avec toi. Dans le même style, merci Cyril pour ta contribution sur TAGADA. Thanks to the FAANG collaborators : to Alan for taking over WP2 (and for bearing with my humour) to the WUR folks for the great collaborations across time (Ole, Jani, Martijn, Martien, Richard etc), and to the others from Roslin/EBI/etc (Sebastien, Fergal, Chris, Alex, Peter, Dan, etc).

Merci à l'équipe "Cyto", avec Martine pour le co-encadrement, Yvette pour la collaboration sur pig3Dgenome. Maria, merci pour ton boulot, ton entrain, tes idées et ta sympathie, qui ont rendu très positive cette première expérience d'encadrement de thèse. Je suis content de notre papier Hi-C (que j'ai eu la bonne surprise de voir cité par un co-reviewer anonyme comme exemple à suivre dans la review d'un autre papier !), sans lequel je n'aurais considéré l'HDR. Je me réjouis à l'idée de prolonger la collaboration avec Alain, Thomas, Anne-Laure – décidément, quelle chance on a avec les nouvelles recrues depuis quelques années ! – et les autres de la nouvelle équipe.

Toujours sur la thématique Hi-C, merci encore Nath, Matthias, Cyril et Sarah, et merci Pierre, renfort déterminant pour la dynamique et le groupe. Pierre, merci pour ta contribution en termes méthodos et humains. Merci de partager mon goût pour les mauvais jeux de mots et autres blagues pathétiques plus ou moins geek – en plus, le fait que Nath les anticipe de plus en plus n'implique pas qu'elle les apprécie de moins en moins. Merci Elise M. pour ton boulot magistral sur HiCDOC. Pour "treediff", merci Marie et les divers stagiaires/doctorant-es, en particulier Elise J. que je remercie pour son énergie et ses nombreuses qualités, tout en lui souhaitant une thèse merveilleuse ! Enfin, merci aux chrocollègues de la chrocoliste pour les chrocotalks, avec en particulier (toujours et encore) Nath, Matthias, Sarah et Pierre pour la plupart du journal club, mais aussi Marion, Raf, Vincent, Solène, Sophie, Thibaut, Anne-Laure, Anouk (merci pour le stage !), Laurence, Alain, Nicolas, Thomas, Annie, Elodie, Guillaume, Hervé, Pascal et les autres pour les chrocoparticipations.

Pitou ! Un gros merci à toi pour tout ! Merci pour tout ce que tu fais résonner en moi, avec ton attitude positive, ton enthousiasme ("C'est génial !") et ton esprit scientifique. Merci d'être source d'inspiration, altruiste et modeste avec ça. Merci pour les "clic-clic", ton indulgence pendant le confinement, les contributions avec Sophie et les autres aux traditionnelles chansons du LGC, et merci d'avoir bien géré cette expérience hallucinante de co-encadrement finalement riche en anecdotes !

Sur le thème de l'épigénétique, merci aussi Céline et Gaëlle pour le renfort sur les analyses de méthylation. Sur le thème du "*single-cell*", merci Pierre, Nath et les autres membres de Ddisc, et Hervé, Cyril, Adrien, Jérôme pour plus4pigs. Jérôme, merci pour les échanges agréables sur 4 ans malgré le tout distanciel.

Revenons à GenPhySE : en premier lieu merci aux membres de la nouvelle équipe REGLISS d'avoir accepté de faire partie de l'aventure et pour m'avoir soutenu lors du montage. Merci donc Agnes, Annie, Carine, Cervin, Guillaume, Laure, Laurence, Mireille, Sabrina, et les plus jeunes. Merci Guillaume d'avoir accepté d'en prendre la direction, et de le faire si bien.

Merci aux collègues qui m'ont filé des coups de main. Parmi les plus sollicité-es toujours pas cité-es, mention spéciale aux gestionnaires budget/mission/rh (Manuela, Valérie, Catherine, Florence, Nancy, Evelyne), au support genotoul (Marie-Stéphane, Didier), aux mousquetaires sigenae (Cédric, Philippe, Patrice), aux "plagistes" de la désormais voisine plateforme de séquençage (Diane, Jérôme, Cécile, Claire). Merci aux autres qui m'ont aidé.

Merci à toutes celles et ceux qui aident aussi à rendre le lieu de travail agréable et convivial, par leurs sourires, rires, blagues, retours sympas après les miennes, et par la culture de l'esprit Bisounours du LGC. Pour la bonne humeur et les discussions (parfois emballées et/ou mémorables) au café ou à table, merci – en vrac et en partie – à Carine, Cervin, Thomas, Pierre, Nath "McFly", Christelle, David, Bertrand, Sonia, Alain, Kamila, Tibo, Anne-Laure, Pitou, Sophie, Katia, Gwen, Stéphane, Laurence, Laure, Manu, Lisa, Elisa, Elise, Julien, Julie, Juliette, Florence, Guillaume, Maguy, Florent, Nathalie, Valérie, Cécile, Yann, Eddie, les jeunes (stagiaires, doctorant-es) trop nombreux-ses pour faire la liste, les gens que j'ai oubliés et les autres fréquentant moins le patio. Merci aussi aux voisins MIAT ou LIPME pour les interactions sympas quand je les croise (Régis, Thomas, Céline, Jérôme, Ludo). Merci aux inconditionnels de l'extérieur pour la compagnie pendant les repas dehors.

Merci aux collègues contribuant aux fonctions collectives de l'unité, de l'animation scientifique à la qualité, aux membres du CS, du CODIR, de ZT, aux IP et aux AP. Merci celles et ceux qui s'occupent du café, des ordis, qui prennent soin du matériel et des espaces partagés. Au niveau du centre, merci les ADASsien-nes (Jérôme, Maguy, Christelle et les autres, du bureau ou responsables d'activités, avec une mention spéciale à Alice, Sophie et Tiphaine pour la gestion de l'escalade), les AUC-COOP, l'atelier vélo (Gérald, Céline et les autres). Merci aux camarades de la section locale (David, Patinet, Didier, Matthias, Guillaume, Michel, MS pour n'en citer qu'une partie), aux RP de la CAPCR (en particulier Sandra, Martine, Rémi, Anne-Marie et Elodie), à certain-es RA (dont Akiko

et Geneviève) et à son ancien président (Laurent). Merci aux collègues engagé-es et/ou impliqué-es dans les instances (CHSCT/F3SCT, CSS, etc), et à celles et eux qui luttent pour préserver une fonction publique de qualité malgré la constante détérioration de la situation depuis des années. On lâche rien !

- Merci à mes ami-es. Mention spéciale aux plus anciens, forcément plus méritants, avec en particulier mes cousins "pas-vraiment-cousins" Lalaina et Guillaume et leur famille (Val, Marlène, Djai), "saladu" JB ("et ouais con !") et mes proches de Barcelone et de cœur Julien, Christoforos et Pauline.

  Merci aux grimpeuses et aux grimpeurs pour les bons moments partagés, soit principalement en salle (Alice, Alain, Sophie, Tiphaine, Tristan, Caleb, Victor, Tristan, Greg et les autres) soit principalement ailleurs : en falaise, en terrasse, au bar, au dernier p'tit pichet, au resto, au barbecue, à la coloc Cazeneuve, en bivouac, autour d'un feu, derrière une guitare, devant une scène, sur la scène, derrière un micro, dans une cave à rock, en camping plus ou moins sauvage, en guinguette plus ou moins éphémère, en camion plus ou moins tagué, sur le canal du midi, en accident de péniche, en route vers Brest à 5 dans une clio, à Tournefeuille, à Muret, vers Montsaunes, à Pessoulens, rue Pargaminière, place St Pierre, aux Carmes, à St Michel, chez moi, en retard, à la bonne franquette, à la belle étoile, en Bretagne, à Lyon, à Naples, en Ariège, à la Clape, dans le Lot – "le plus beau département du monde ! ! !" –, en rivière, en kayak, en brunch du samedimanche, devant du rugby, en réveillon, en déménagement, en crémaillère, en Evasion, en raclette, en fondue, en tristitude, dans le Béarn, dans le Gers, en sauvetage du Bastion à Lectoure, sous une table, dedans après la fermeture, à Paris par hasard (plus en gare que su'l'pont des Arts), au bord ou dans un lac, en grosse flemme, en crise d'ail, sur une plage, en rando dans les Pyrénées, entre Bor et Bar, en pleine nuit (une sirène), au bout de la Terre, au pays des merveilles, en folklore québécois, en moule, en tête, en dévers, sur la même corde, sur la même paroi, sous la même tente, dans le même lit, en cachette, en rêve, en vrac, en vadrouille, en fous rires, en larmes. Pour tout ça, merci Adrien, David, Clarisse, Anne, Céline, Yoann, Léa, Vana, Marie, Marion, Maud, Thomas, Anne-Laure, David et les autres.

  Merci enfin au "nouveau" quartier (Nath, Marie, Van, Tophe, Alexis, Carine, Sylvaine, Alain, Alex, Noslen, Pierre, Matthias, Jordane) pour les afterworks, cinés, quiz, jeux, soirées. Merci plus loin, que ce soit Pessac (Patricia, Thierry), Berlin (Barbara, Hilmar) ou du côté de Layrisse (Barou, Michel et Dominique, Colette, Giselle, Eléa, etc).

- Merci enfin à ma famille.

  Merci à ma famille du Québec pour l'accueil ("un p'tit café avant de partir ?"), et les bons moments, en particulier à Lise, Marcel, Gisou, Angèle, Muguette, Pascal, Katia, Karine.

Merci à mon frérot d'amour ("Youhou frérot !") et à mes parents chéris pour le bonheur, le soutien, le respect, l'amour. TRZ & Mo+, je n'arriverai jamais à exprimer à quel point je vous suis reconnaissant pour tout ce que j'ai vécu et que je vis encore de positif grâce à vous. Sachez que je suis toujours conscient de la chance incroyable et exceptionnelle de vous avoir. La science a beau tenir une place importante dans ma vie, je sais que la magie existe : je la retrouve à chaque passage aux Bourdettes.

Pour terminer, merci à toi, lectrice, lecteur, pour ton temps et ton attention. J'espère que tu as passé un agréable moment.

Du bonheur, la bise et à bientôt !

Bibliographie

Acloque H, Harrison P, Lakhal W, Martin F, Archibald A, Beinat M, Davey M, Djebali S, Foissac S, Guizard S et al. (2022). "Extensive functional genomics information from early developmental time points for pig and chicken (Talk)". In : *Proceedings of 12th World Congress on Genetics Applied to Livestock Production (WCGALP)*. Wageningen Academic Publishers, p. 2281-2284. DOI : 10.3920/978-90-8686-940-4_550.

Andersson L, Archibald AL, Bottema CD, Brauning R, Burgess SC, Burt DW, Casas E, Cheng HH, Clarke L, Couldrey C et al. (2015). "Coordinated international action to accelerate genome-to-phenome with FAANG, the Functional Annotation of Animal Genomes project". In : *Genome Biology* 16.1, p. 1-6. ISSN : 1474760X. DOI : 10.1186/s13059-015-0622-4.

Andrews SJ et Rothnagel JA (2014). "Emerging evidence for functional peptides encoded by short open reading frames". In : *Nature Reviews Genetics* 15.3, p. 193-204. ISSN : 1471-0064. DOI : 10.1038/nrg3520.

Bakel H van, Nislow C, Blencowe BJ et Hughes TR (2010). "Most "Dark Matter" Transcripts Are Associated With Known Genes". In : *PLoS Biology* 8.5, e1000371. ISSN : 1545-7885. DOI : 10.1371/journal.pbio.1000371.

— (2011). "Response to "The Reality of Pervasive Transcription"". In : *PLoS Biology* 9.7, e1001102. ISSN : 1545-7885. DOI : 10.1371/journal.pbio.1001102.

Barczak W, Carr SM, Liu G, Munro S, Nicastri A, Lee LN, Hutchings C, Ternette N, Klenerman P, Kanapin A et al. (2023). "Long non-coding RNA-derived peptides are immunogenic and drive a potent anti-tumour response". In : *Nature Communications* 14.1. ISSN : 2041-1723. DOI : 10.1038/s41467-023-36826-0.

Belew AT, Meskauskas A, Musalgaonkar S, Advani VM, Sulima SO, Kasprzak WK, Shapiro BA et Dinman JD (2014). "Ribosomal frameshifting in the CCR5 mRNA is regulated by miRNAs and the NMD pathway". In : *Nature* 512.7514, p. 265-269. ISSN : 1476-4687. DOI : 10.1038/nature13429.

Belmont AS (2014). "Large-scale chromatin organization : the good, the surprising, and the still perplexing". In : *Current Opinion in Cell Biology* 26, p. 69-78. ISSN : 0955-0674. DOI : 10.1016/j.ceb.2013.10.002.

Bertone P, Stolc V, Royce TE, Rozowsky JS, Urban AE, Zhu X, Rinn JL, Tongprasit W, Samanta M, Weissman S et al. (2004). "Global Identification of Human Transcribed Sequences with Genome Tiling Arrays". In : *Science* 306.5705, p. 2242-2246. ISSN : 1095-9203. DOI : 10.1126/science.1103388.

Blake WJ, Balázsi G, Kohanski MA, Isaacs FJ, Murphy KF, Kuang Y, Cantor CR, Walt DR et Collins JJ (2006). "Phenotypic Consequences of Promoter-Mediated Transcriptional Noise". In : *Molecular Cell* 24.6, p. 853-865. ISSN : 1097-2765. DOI : 10.1016/j.molcel.2006.11.003.

Bonev B et Cavalli G (2016). "Organization and function of the 3D genome". In : *Nature Reviews Genetics* 17.11, p. 661-678. ISSN : 1471-0064. DOI : 10.1038/nrg.2016.112.

Bonnet A, Cabau C, Bouchez O, Sarry J, Marsaud N, Foissac S, Woloszyn F, Mulsant P et Mandon-Pepin B (2013). "An overview of gene expression dynamics during early ovarian folliculogenesis : Specificity of follicular compartments and bi-directional dialog". In : *BMC Genomics* 14.1, p. 1-19. ISSN : 14712164. DOI : 10.1186/1471-2164-14-904.

BRENNER S, JACOB F et MESELSON M (1961). "An Unstable Intermediate Carrying Information from Genes to Ribosomes for Protein Synthesis". In : *Nature* 190.4776, p. 576-581. ISSN : 1476-4687. DOI : 10.1038/190576a0.

Burian R (2004). "Molecular epigenesis, molecular pleiotropy, and molecular gene definitions". In : *History & Philosophy of the Life Sciences* 26.1, p. 59-80. ISSN : 0391-9714. DOI : 10.1080/03919710412331341641.

Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C et al. (2005). "The Transcriptional Landscape of the Mammalian Genome". In : *Science* 309.5740, p. 1559-1563. ISSN : 1095-9203. DOI : 10.1126/science.1112014.

Carninci P (2006). "Tagging mammalian transcription complexity". In : *Trends in Genetics* 22.9, p. 501-510. ISSN : 0168-9525. DOI : 10.1016/j.tig.2006.07.003.

Chen J, Brunner AD, Cogan JZ, Nuñez JK, Fields AP, Adamson B, Itzhak DN, Li JY, Mann M, Leonetti MD et al. (2020). "Pervasive functional translation of noncanonical human open reading frames". In : *Science* 367.6482, p. 1140-1146. ISSN : 1095-9203. DOI : 10.1126/science.aay0262.

Chisanga D, Liao Y et Shi W (2022). "Impact of gene annotation choice on the quantification of RNA-seq data". In : *BMC Bioinformatics* 23.1. ISSN : 1471-2105. DOI : 10.1186/s12859-022-04644-8.

Clark MB, Amaral PP, Schlesinger FJ, Dinger ME, Taft RJ, Rinn JL, Ponting CP, Stadler PF, Morris KV, Morillon A et al. (2011). "The Reality of Pervasive Transcription". In : *PLoS Biology* 9.7, e1000625. ISSN : 1545-7885. DOI : 10.1371/journal.pbio.1000625.

Crowe ML (2003). "CATMA : a complete Arabidopsis GST database". In : *Nucleic Acids Research* 31.1, p. 156-158. ISSN : 1362-4962. DOI : 10.1093/nar/gkg071.

Dali R et Blanchette M (2017). "A critical assessment of topologically associating domain prediction tools". In : *Nucleic Acids Research* 45.6, p. 2994-3005. ISSN : 1362-4962. DOI : 10.1093/nar/gkx145.

Danna K et Nathans D (1971). "Specific cleavage of simian virus 40 DNA by restriction endonuclease of Hemophilus influenzae". In : *Proceedings of the National Academy of Sciences* 68.12, p. 2913-2917. ISSN : 1091-6490. DOI : 10.1073/pnas.68.12.2913.

Darwin C (1859). *On the origin of species by means of natural selection, or, The preservation of favoured races in the struggle for life.* https ://www.biodiversitylibrary.org/bibliography/82303. London John Murray, Albemarle Street 1859, p. 556.

David SA, Mersch M, Foissac S, Collin A, Pitel F et Coustham V (2017). "Genome-wide epigenetic studies in chicken : A review". In : *Epigenomes* 1.3, p. 20. ISSN : 20754655. DOI : 10.3390/epigenomes1030020.

Davies JOJ, Oudelaar AM, Higgs DR et Hughes JR (2017). "How best to identify chromosomal interactions : a comparison of approaches". In : *Nature Methods* 14.2, p. 125-134. ISSN : 1548-7105. DOI : 10.1038/nmeth.4146.

De Vos J et al. (2023). "DNA methylation dynamics regulating embryonic development in pig (Poster)". In : *ISAG*. Cape Town, South Africa.

Degalez F et al. (2023). "A lncRNA gene-enriched atlas for GRCg7b chicken genome and its functional annotation across 47 tissues (Poster)". In : *ISAG*. Cape Town, South Africa.

Degalez F, Charles M, Foissac S, Zhou H, Guan D, Fang L, Klopp C, Allain C, Lagoutte L, Lecerf F et al. (2024). "Enriched atlas of lncRNA and protein-coding genes for the GRCg7b chicken assembly and its functional annotation across 47 tissues". In : *Scientific Reports* 14.1, p. 6588. DOI : 10.1038/s41598-024-56705-y.

Demerec M (1955). "What is a Gene ?-Twenty Years Later". In : *The American Naturalist* 89.844, p. 5-20. ISSN : 1537-5323. DOI : 10.1086/281856.

Denoeud F, Kapranov P, Ucla C, Frankish A, Castelo R, Drenkow J, Lagarde J, Alioto T, Manzano C, Chrast J et al. (2007). "Prominent use of distal 5' transcription start sites and discovery of a large number of additional exons in ENCODE regions". In : *Genome Research* 17.6, p. 746-759. ISSN : 10889051. DOI : 10.1101/gr.5660607.

Denoyelle L, Talouarn E, Bardou P, Colli L, Alberti A, Danchin C, Del Corvo M, Engelen S, Orvain C, Palhière I et al. (2021). "VarGoats project : a dataset of 1159 whole-genome sequences to dissect Capra hircus global diversity". In : *Genetics Selection Evolution* 53.1. ISSN : 1297-9686. DOI : 10.1186/s12711-021-00659-6.

Dinger ME, Amaral PP, Mercer TR et Mattick JS (2009). "Pervasive transcription of the eukaryotic genome : functional indices and conceptual implications". In : *Briefings in Functional Genomics and Proteomics* 8.6, p. 407-423. ISSN : 1477-4062. DOI : 10.1093/bfgp/elp038.

Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F et al. (2012a). "Landscape of transcription in human cells". In : *Nature* 489.7414, p. 101-108. ISSN : 1476-4687. DOI : 10.1038/nature11233.

Djebali S, Foissac S, Vialaneix N, Munyard K, Rau A, Faraut T, Lagarrigue S, Acloque H et Giuffra E (2019). "Chromatin accessibility conservation across four livestock species (Talk)". In : *Inproceedings of the International Society for Animal Genetics (ISAG 2019)* (7-12 juill. 2019). Llieda, Spain.

Djebali S, Kapranov P, Foissac S, Lagarde J, Reymond A, Ucla C, Wyss C, Drenkow J, Dumais E, Murray RR et al. (2008). "Efficient targeted transcript discovery via array-based normalization of RACE libraries". In : *Nature Methods* 5.7, p. 629-635. ISSN : 15487091. DOI : 10.1038/nmeth.1216.

Djebali S, Lagarde J, Kapranov P, Lacroix V, Borel C, Mudge JM, Howald C, Foissac S, Ucla C, Chrast J et al. (2012b). "Evidence for transcript networks composed of chimeric rnas in human cells". In : *PLoS ONE* 7.1, e28213. ISSN : 19326203. DOI : 10.1371/journal.pone.0028213.

Djebali S, Wucher V, Foissac S, Hitte C, Corre E et Derrien T (2017). “Bioinformatics pipeline for transcriptome sequencing analysis”. In : *Methods in Molecular Biology*. T. 1468. Springer Protocols, p. 201-219. DOI : 10.1007/978-1-4939-4035-6_14.

Dufour A, Kurylo C, Stöckl JB, Laloë D, Bailly Y, Manceau P, Martins F, Turhan AG, Ferchaud S, Pain B et al. (2024). “Cell specification and functional interactions in the pig blastocyst inferred from single-cell transcriptomics and uterine fluids proteomics”. In : *Genomics* 116.2, p. 110780. ISSN : 0888-7543. DOI : 10.1016/j.ygeno.2023.110780.

Farrell CM, Goldfarb T, Rangwala SH, Astashyn A, Ermolaeva OD, Hem V, Katz KS, Kodali VK, Ludwig F, Wallin CL et al. (2021). “RefSeq Functional Elements as experimentally assayed nongenic reference standards and functional interactions in human and mouse”. In : *Genome Research* 32.1, p. 175-188. ISSN : 1549-5469. DOI : 10.1101/gr.275819.121.

Fejes-Toth K, Sotirova V, Sachidanandam R, Assaf G, Hannon GJ, Kapranov P, Foissac S, Willingham AT, Duttagupta R, Dumais E et al. (2009). “Post-transcriptional processing generates a diversity of 5’-modified long and short RNAs”. In : *Nature* 457.7232, p. 1028-1032. ISSN : 00280836. DOI : 10.1038/nature07759.

Fève K, Foissac S, Pinton A, Mompart F, Esquerré D, Faraut T, Yerle M et Riquet J (2017). “Identification of a t(3 ;4)(p1.3 ;q1.5) translocation breakpoint in pigs using somatic cell hybrid mapping and high-resolution mate-pair sequencing”. In : *PLoS ONE* 12.11, p. 1-17. ISSN : 19326203. DOI : 10.1371/journal.pone.0187617.

Foissac S (2004). “Localisation de gènes et variants par intégration d’informations”. Thèse de doct. Toulouse, France : Université Paul Sabatier Toulouse III.

— (2019). “Functional annotation of livestock genomes : chromatin structure and regulation of gene expression (Invited talk)”. In : *Journal of Animal Science (Proceedings of the ASAS/ASDS Midwest Joint Meeting)*. T. 97. Suppl. 2. Omaha, NE, USA, p. 15-16. DOI : 10.1093/jas/skz122.028.

— (2022). “TAGADA : Transcripts and Genes Assembly, Deconvolution, Analysis ; a Nextflow pipeline to improve genome annotations with RNA-seq data (Invited talk)”. In : PRBB-CRG. Barcelona, Catalunya.

Foissac S, Bardou P, Moisan A, Cros MJ et Schiex T (2003). “EUGÈNE’HOM : A generic similarity-based gene finder using multiple homologous sequences”. In : *Nucleic Acids Research* 31.13, p. 3742-3745. ISSN : 03051048. DOI : 10.1093/nar/gkg586.

Foissac S, Djebali S, Munyard K, Vialaneix N, Rau A, Muret K, Esquerré D, Zytnicki M, Derrien T, Bardou P et al. (2019a). “Multi-species annotation of transcriptome and chromatin structure in domesticated animals”. In : *BMC Biology* 17.1, p. 1-25. ISSN : 17417007. DOI : 10.1186/s12915-019-0726-5.

Foissac S, Djebali S, Vialaneix N, Zytnicki M, Rau A, Lagarrigue S, Acloque H et Giuffra E (2019b). “Multi-level conservation of chromosome conformation across livestock species reveals evolutionary links between genome structure and function (Talk)”. In : *Inproceedings of the International Society for Animal Genetics (ISAG 2019)* (7-12 juill. 2019). Llieda, Spain.

Foissac S, Gouzy J, Rombauts S, Mathe C, Amselem J, Sterck L, Peer Y de, Rouze P et Schiex T (2008). "Genome Annotation in Plants and Fungi : EuGene as a Model Platform". In : *Current Bioinformatics* 3.2, p. 87-97. ISSN : 15748936. DOI : 10.2174/157489308784340702.

Foissac S et Sammeth M (2007). "ASTALAVISTA : dynamic and flexible analysis of alternative splicing events in custom gene datasets". In : *Nucleic acids research* 35.suppl_2, W297-W299. DOI : 10.1093/nar/gkm311.

— (2015). "Analysis of alternative splicing events in custom gene datasets by AStalavista". In : *Methods in Molecular Biology*. T. 1269. Springer Protocols, p. 379-392. DOI : 10.1007/978-1-4939-2291-8_24.

Foissac S et Schiex T (2005). "Integrating alternative splicing detection into gene prediction". In : *BMC Bioinformatics* 6.1, p. 1-10. ISSN : 14712105. DOI : 10.1186/1471-2105-6-25.

Fox Keller E et Harel D (2007). "Beyond the Gene". In : *PLoS ONE* 2.11, e1231. ISSN : 1932-6203. DOI : 10.1371/journal.pone.0001231.

Frith MC, Forrest AR, Nourbakhsh E, Pang KC, Kai C, Kawai J, Carninci P, Hayashizaki Y, Bailey TL et Grimmond SM (2006). "The Abundance of Short Proteins in the Mammalian Proteome". In : *PLoS Genetics* 2.4, e52. ISSN : 1553-7404. DOI : 10.1371/journal.pgen.0020052.

Gerstein MB, Bruce C, Rozowsky JS, Zheng D, Du J, Korbel JO, Emanuelsson O, Zhang ZD, Weissman S et Snyder M (2007). "What is a gene, post-ENCODE? History and updated definition". In : *Genome Research* 17.6, p. 669-681. ISSN : 1088-9051. DOI : 10.1101/gr.6339607.

Gingeras TR (2007). "Origin of phenotypes : Genes and transcripts". In : *Genome Research* 17.6, p. 682-690. ISSN : 1088-9051. DOI : 10.1101/gr.6525007.

Giuffra E, Tuggle CK et the FAANG Consortium (2019). "Functional Annotation of Animal Genomes (FAANG) : Current Achievements and Roadmap". In : *Annual Review of Animal Biosciences* 7, p. 65-88. ISSN : 21658110. DOI : 10.1146/annurev-animal-020518-114913.

Goszczynski DE, Halstead MM, Islas-Trejo AD, Zhou H et Ross PJ (2021). "Transcription initiation mapping in 31 bovine tissues reveals complex promoter activity, pervasive transcription, and tissue-specific promoter usage". In : *Genome Research* 31.4, p. 732-744. ISSN : 1549-5469. DOI : 10.1101/gr.267336.120.

Gramates LS, Agapite J, Attrill H, Calvi BR, Crosby MA, Santos G dos, Goodman JL, Goutte-Gattat D, Jenkins VK, Kaufman T et al. (2022). "FlyBase : a guided tour of highlighted features". In : *Genetics* 220.4. ISSN : 1943-2631. DOI : 10.1093/genetics/iyac035.

Graur D, Zheng Y, Price N, Azevedo RBR, Zufall RA et Elhaik E (2013). "On the Immortality of Television Sets : "Function" in the Human Genome According to the Evolution-Free Gospel of ENCODE". In : *Genome Biology and Evolution* 5.3, p. 578-590. ISSN : 1759-6653. DOI : 10.1093/gbe/evt028.

Heesch S van, Witte F, Schneider-Lunitz V, Schulz JF, Adami E, Faber AB, Kirchner M, Maatz H, Blachut S, Sandmann CL et al. (2019). "The Translational Landscape of the Human Heart". In : *Cell* 178.1, 242-260.e29. ISSN : 0092-8674. DOI : 10.1016/j.cell.2019.05.010.

Henras AK, Plisson-Chastang C, O'Donohue MF, Chakraborty A et Gleizes PE (2014). "An overview of pre-ribosomal RNA processing in eukaryotes". In : *WIREs RNA* 6.2, p. 225-242. ISSN : 1757-7012. DOI : 10.1002/wrna.1269.

Hoellinger T, Mestre C, Aschard H, Le Goff W, Foissac S, Faraut T et Djebali S (2023). "Enhancer/gene relationships : need for more reliable genome-wide reference sets". In : *Frontiers in Bioinformatics* 3, p. 1092853. DOI : 10.1093/nar/gkx920.

Hogenesch JB, Ching KA, Batalov S, Su AI, Walker JR, Zhou Y, Kay SA, Schultz PG et Cooke MP (2001). "A Comparison of the Celera and Ensembl Predicted Gene Sets Reveals Little Overlap in Novel Genes". In : *Cell* 106.4, p. 413-415. ISSN : 0092-8674. DOI : 10.1016/s0092-8674(01)00467-6.

Ingolia NT, Ghaemmaghami S, Newman JRS et Weissman JS (2009). "Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling". In : *Science* 324.5924, p. 218-223. ISSN : 1095-9203. DOI : 10.1126/science.1168978.

Ingolia NT, Lareau LF et Weissman JS (2011). "Ribosome Profiling of Mouse Embryonic Stem Cells Reveals the Complexity and Dynamics of Mammalian Proteomes". In : *Cell* 147.4, p. 789-802. ISSN : 0092-8674. DOI : 10.1016/j.cell.2011.10.002.

Jacob F et Monod J (1961). "Genetic regulatory mechanisms in the synthesis of proteins". In : *Journal of Molecular Biology* 3.3, p. 318-356. ISSN : 0022-2836. DOI : 10.1016/s0022-2836(61)80072-7.

Jacquier A (2009). "The complex eukaryotic transcriptome : unexpected pervasive transcription and novel small RNAs". In : *Nature Reviews Genetics* 10.12, p. 833-844. ISSN : 1471-0064. DOI : 10.1038/nrg2683.

Jaenisch R et Mintz B (1974). "Simian Virus 40 DNA Sequences in DNA of Healthy Adult Mice Derived from Preimplantation Blastocysts Injected with Viral DNA". In : *Proceedings of the National Academy of Sciences* 71.4, p. 1250-1254. ISSN : 1091-6490. DOI : 10.1073/pnas.71.4.1250.

Jehl F, Degalez F, Bernard M, Lecerf F, Lagoutte L, Désert C, Coulée M, Bouchez O, Leroux S, Abasht B et al. (2021). "RNA-Seq Data for Reliable SNP Detection and Genotype Calling : Interest for Coding Variant Characterization and Cis-Regulation Analysis by Allele-Specific Expression in Livestock Species". In : *Frontiers in Genetics* 12, p. 1104. ISSN : 16648021. DOI : 10.3389/fgene.2021.655707.

Jehl F, Muret K, Bernard M, Boutin M, Lagoutte L, Désert C, Dehais P, Esquerré D, Acloque H, Giuffra E et al. (2020). "An integrative atlas of chicken long non-coding genes and their annotations across 25 tissues". In : *Scientific reports* 10.1, p. 20457. DOI : 10.1038/s41598-020-77586-x.

Jensen TH, Jacquier A et Libri D (2013). "Dealing with Pervasive Transcription". In : *Molecular Cell* 52.4, p. 473-484. ISSN : 1097-2765. DOI : 10.1016/j.molcel.2013.10.032.

Johnson JM, Edwards S, Shoemaker D et Schadt EE (2005). "Dark matter in the genome : evidence of widespread transcription detected by microarray tiling experiments". In : *Trends in Genetics* 21.2, p. 93-102. ISSN : 0168-9525. DOI : 10.1016/j.tig.2004.12.009.

Kapranov P, Cawley SE, Drenkow J, Bekiranov S, Strausberg RL, Fodor SPA et Gingeras TR (2002). "Large-Scale Transcriptional Activity in Chromosomes 21 and 22". In : *Science* 296.5569, p. 916-919. ISSN : 1095-9203. DOI : 10.1126/science.1068597.

Kapranov P, Cheng J, Dike S, Nix DA, Duttagupta R, Willingham AT, Stadler PF, Hertel J, Hackermüller J, Hofacker IL et al. (2007a). "RNA Maps Reveal New RNA Classes and a Possible Function for Pervasive Transcription". In : *Science* 316.5830, p. 1484-1488. ISSN : 1095-9203. DOI : 10.1126/science.1138341.

Kapranov P, Drenkow J, Cheng J, Long J, Helt G, Dike S et Gingeras TR (2005). "Examples of the complex architecture of the human transcriptome revealed by RACE and high-density tiling arrays". In : *Genome Research* 15.7, p. 987-997. ISSN : 1088-9051. DOI : 10.1101/gr.3455305.

Kapranov P, Ozsolak F, Kim SW, Foissac S, Lipson D, Hart C, Roels S, Borel C, Antonarakis SE, Monaghan AP et al. (2010a). "New class of gene-termini-associated human RNAs suggests a novel RNA copying mechanism". In : *Nature* 466.7306, p. 642-646. ISSN : 14764687. DOI : 10.1038/nature09190.

Kapranov P, St Laurent G, Raz T, Ozsolak F, Reynolds CP, Sorensen PH, Reaman G, Milos P, Arceci RJ, Thompson JF et al. (2010b). "The majority of total nuclear-encoded non-ribosomal RNA in a human cell is "dark matter" un-annotated RNA". In : *BMC Biology* 8.1. ISSN : 1741-7007. DOI : 10.1186/1741-7007-8-149.

Kapranov P, Willingham AT et Gingeras TR (2007b). "Genome-wide transcription and the implications for genomic organization". In : *Nature Reviews Genetics* 8.6, p. 413-423. ISSN : 1471-0064. DOI : 10.1038/nrg2083.

Kempfer R et Pombo A (2019). "Methods for mapping 3D chromosome architecture". In : *Nature Reviews Genetics* 21.4, p. 207-226. ISSN : 1471-0064. DOI : 10.1038/s41576-019-0195-2.

Kurylo C, Guyomar C, Foissac S et Djebali S (2023a). "TAGADA : a scalable pipeline to improve genome annotations with RNA-seq data". In : *NAR Genomics And Bioinformatics* 5.4, lqad089. ISSN : 2631-9268. DOI : 10.1093/nargab/lqad089.

Kurylo C, Maigné E, Foissac S et Zytnicki M (2023b). "Prediction and differential analysis of chromatin compartments from Hi-C data (Poster)". In : *ISMB/ECCB*. Lyon, France.

Lahbib-Mansais Y, Marti-Marimon M, Vialaneix N, Foissac S, Bouissou-Matet M et Liaubet L (2019). "Organisation nucléaire et expression génique lors du développement chez le porc". In : *Séminaire du réseau EpiPHASE* (26-27 juin 2019). Castanet-Tolosan, France.

Lee S, Liu B, Lee S, Huang SX, Shen B et Qian SB (2012). "Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution". In : *Proceedings of the National Academy of Sciences* 109.37. ISSN : 1091-6490. DOI : 10.1073/pnas.1207846109.

Lee Phillips M (2010). "Existence of RNA "dark matter" in doubt". In : *Nature*. ISSN : 1476-4687. DOI : 10.1038/news.2010.248.

Lieberman-Aiden E, Berkum NL van, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO et al. (2009). "Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome". In : *Science* 326.5950, p. 289-293. ISSN : 1095-9203. DOI : 10.1126/science.1181369.

Liu L, Han K, Sun H, Han L, Gao D, Xi Q, Zhang L et Lin H (2023). “A comprehensive review of bioinformatics tools for chromatin loop calling”. In : *Briefings in Bioinformatics* 24.2. ISSN : 1477-4054. DOI : 10.1093/bib/bbad072.

Losick R et Desplan C (2008). “Stochasticity and Cell Fate”. In : *Science* 320.5872, p. 65-68. ISSN : 1095-9203. DOI : 10.1126/science.1147888.

Lupiáñez DG, Kraft K, Heinrich V, Krawitz P, Brancati F, Klopocki E, Horn D, Kayserili H, Opitz JM, Laxova R et al. (2015). “Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions”. In : *Cell* 161.5, p. 1012-1025. ISSN : 0092-8674. DOI : 10.1016/j.cell.2015.04.004.

Mandel M et Higa A (1970). “Calcium-dependent bacteriophage DNA infection”. In : *Journal of Molecular Biology* 53.1, p. 159-162. ISSN : 0022-2836. DOI : 10.1016/0022-2836(70)90051-3.

Mariotti M, Ridge PG, Zhang Y, Lobanov AV, Pringle TH, Guigo R, Hatfield DL et Gladyshev VN (2012). “Composition and Evolution of the Vertebrate and Mammalian Selenoproteomes”. In : *PLoS ONE* 7.3, e33066. ISSN : 1932-6203. DOI : 10.1371/journal.pone.0033066.

Marti-Marimon M, Vialaneix N, Lahbib-Mansais Y, Zytnicki M, Camut S, Robelin D, Yerle-Bouissou M et Foissac S (2021). “Major Reorganization of Chromosome Conformation During Muscle Development in Pig”. In : *Frontiers in Genetics* 12, p. 1895. ISSN : 16648021. DOI : 10.3389/fgene.2021.748239.

Martin FJ, Amode MR, Aneja A, Austine-Orimoloye O, Azov AG, Barnes I, Becker A, Bennett R, Berry A, Bhai J et al. (2022). “Ensembl 2023”. In : *Nucleic Acids Research* 51.D1, p. D933-D941. ISSN : 1362-4962. DOI : 10.1093/nar/gkac958.

Mattick JS, Amaral PP, Carninci P, Carpenter S, Chang HY, Chen LL, Chen R, Dean C, Dinger ME, Fitzgerald KA et al. (2023). “Long non-coding RNAs : definitions, functions, challenges and recommendations”. In : *Nature Reviews Molecular Cell Biology* 24.6, p. 430-447. ISSN : 1471-0080. DOI : 10.1038/s41580-022-00566-8.

Mendel G (1865). “Versuche uber pflanzen-hybriden”. In : *Vorgelegt in den Sitzungen.*

Mota-Gómez I et Lupiáñez DG (2019). “A (3D-Nuclear) Space Odyssey : Making Sense of Hi-C Maps”. In : *Genes* 10.6, p. 415. ISSN : 2073-4425. DOI : 10.3390/genes10060415.

Muret K, Klopp C, Wucher V, Esquerré D, Legeai F, Lecerf F, Désert C, Boutin M, Jehl F, Acloque H et al. (2017). “Long noncoding RNA repertoire in chicken liver and adipose tissue”. In : *Genetics Selection Evolution* 49.1, p. 1-17. ISSN : 12979686. DOI : 10.1186/s12711-016-0275-0.

Nassar LR, Barber GP, Benet-Pagès A, Casper J, Clawson H, Diekhans M, Fischer C, Gonzalez JN, Hinrichs AS, Lee BT et al. (2022). “The UCSC Genome Browser database : 2023 update”. In : *Nucleic Acids Research* 51.D1, p. D1188-D1195. ISSN : 1362-4962. DOI : 10.1093/nar/gkac1072.

Nemeth K, Bayraktar R, Ferracin M et Calin GA (2023). “Non-coding RNAs in disease : from mechanisms to therapeutics”. In : *Nature Reviews Genetics* 25.3, p. 211-232. ISSN : 1471-0064. DOI : 10.1038/s41576-023-00662-1.

Neuvial P, Foissac S et Vialaneix N (2023). “Comprendre l’organisation spatiale de l’ADN à l’aide de la statistique”. In : *L’Interdisciplinarité*. T. 1. CNRS Editions, p. 172-179.

Neuvial P, Randriamihamison N, Chavent M, Foissac S et Vialaneix N (2024). “A two-sample tree-based test for hierarchically organized genomic signals”. In : *Journal of the Royal Statistical Society Series C : Applied Statistics*, qlae011. DOI : 10.1093/jrsssc/qlae011.

Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, Vollger MR, Altemose N, Uralsky L, Gershman A et al. (2022). “The complete sequence of a human genome”. In : *Science* 376.6588, p. 44-53. ISSN : 1095-9203. DOI : 10.1126/science.abj6987.

Otsuka Y, Kedersha NL et Schoenberg DR (2009). “Identification of a Cytoplasmic Complex That Adds a Cap onto 5’-Monophosphate RNA”. In : *Molecular and Cellular Biology* 29.8, p. 2155-2167. ISSN : 1098-5549. DOI : 10.1128/mcb.01325-08.

Oudelaar AM et Higgs DR (2020). “The relationship between genome structure and function”. In : *Nature Reviews Genetics* 22.3, p. 154-168. ISSN : 1471-0064. DOI : 10.1038/s41576-020-00303-x.

Ozsolak F, Kapranov P, Foissac S, Kim SW, Fishilevich E, Monaghan AP, John B et Milos PM (2010). “Comprehensive polyadenylation site maps in yeast and human reveal pervasive alternative polyadenylation”. In : *Cell* 143.6, p. 1018-1029. ISSN : 00928674. DOI : 10.1016/j.cell.2010.11.020.

Palazzo AF et Koonin EV (2020). “Functional Long Non-coding RNAs Evolve from Junk Transcripts”. In : *Cell* 183.5, p. 1151-1161. ISSN : 0092-8674. DOI : 10.1016/j.cell.2020.09.047.

Patraquim P, Magny EG, Pueyo JI, Platero AI et Couso JP (2022). “Translation and natural selection of micropeptides from long non-canonical RNAs”. In : *Nature Communications* 13.1. ISSN : 2041-1723. DOI : 10.1038/s41467-022-34094-y.

Pearson H (2006). “What is a gene ?” In : *Nature* 441.7092, p. 398-401. ISSN : 1476-4687. DOI : 10.1038/441398a.

Pennisi E (2007). “DNA Study Forces Rethink of What It Means to Be a Gene”. In : *Science* 316.5831, p. 1556-1557. ISSN : 1095-9203. DOI : 10.1126/science.316.5831.1556.

Pertea M, Shumate A, Pertea G, Varabyou A, Breitwieser FP, Chang YC, Madugundu AK, Pandey A et Salzberg SL (2018). “CHESS : a new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise”. In : *Genome Biology* 19.1. ISSN : 1474-760X. DOI : 10.1186/s13059-018-1590-2.

Pesole G (2008). “What is a gene ? An updated operational definition”. In : *Gene* 417.1–2, p. 1-4. ISSN : 0378-1119. DOI : 10.1016/j.gene.2008.03.010.

Piqué M, López JM, Foissac S, Guigó R et Méndez R (2008). “A Combinatorial Code for CPE-Mediated Translational Control”. In : *Cell* 132.3, p. 434-448. ISSN : 00928674. DOI : 10.1016/j.cell.2007.12.038.

Ponjavic J, Ponting CP et Lunter G (2007). “Functionality or transcriptional noise ? Evidence for selection within long noncoding RNAs”. In : *Genome Research* 17.5, p. 556-565. ISSN : 1088-9051. DOI : 10.1101/gr.6036807.

Portin P (1993). “The Concept of the Gene : Short History and Present Status”. In : *The Quarterly Review of Biology* 68.2, p. 173-223. ISSN : 1539-7718. DOI : 10.1086/418039.

Portin P et Wilkins A (2017). “The Evolving Definition of the Term “Gene””. In : *Genetics* 205.4, p. 1353-1364. ISSN : 1943-2631. DOI : 10.1534/genetics.116.196956.

Raj A, Peskin CS, Tranchina D, Vargas DY et Tyagi S (2006). “Stochastic mRNA Synthesis in Mammalian Cells”. In : *PLoS Biology* 4.10, e309. ISSN : 1545-7885. DOI : 10.1371/journal.pbio.0040309.

Randriamihamison N (2021). “Classification Ascendante Hiérarchique sous Contrainte de Contiguïté pour l’Analyse de données Hi-C”. Thèse de doct.

Randriamihamison N, Chavent M, Foissac S, Vialaneix N et Neuvial P (2020). “Classification ascendante hiérarchique sous contrainte de contigüité pour l’analyse différentielle de données Hi-C”. In : *Journées de Statistique de la SFdS (volume exceptionnel)*.

Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES et al. (2014). “A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping”. In : *Cell* 159.7, p. 1665-1680. ISSN : 0092-8674. DOI : 10.1016/j.cell.2014.11.021.

Robertson HM, Navik JA, Walden KKO et Honegger HW (2007). “The Bursicon Gene in Mosquitoes : An Unusual Example of mRNA Trans-splicing”. In : *Genetics* 176.2, p. 1351-1353. ISSN : 1943-2631. DOI : 10.1534/genetics.107.070938.

Robertson M (2010). “The evolution of gene regulation, the RNA universe, and the vexed questions of artefact and noise”. In : *BMC Biology* 8.1. ISSN : 1741-7007. DOI : 10.1186/1741-7007-8-97.

Robinson R (2010). “Dark Matter Transcripts : Sound and Fury, Signifying Nothing ?” In : *PLoS Biology* 8.5, e1000370. ISSN : 1545-7885. DOI : 10.1371/journal.pbio.1000370.

Rubio-Peña K, Fontrodona L, Aristizábal-Corrales D, Torres S, Cornes E, García-Rodríguez FJ, Serrat X, González-Knowles D, Foissac S, Porta-De-La-Riva M et al. (2015). “Modeling of autosomal-dominant retinitis pigmentosa in Caenorhabditis elegans uncovers a nexus between global impaired functioning of certain splicing factors and cell type-specific apoptosis”. In : *RNA* 21.12, p. 2119-2131. ISSN : 14699001. DOI : 10.1261/rna.053397.115.

Sammeth M, Foissac S et Guigó R (2008). “A general definition and nomenclature for alternative splicing events”. In : *PLoS Computational Biology* 4.8, e1000147. ISSN : 1553734X. DOI : 10.1371/journal.pcbi.1000147.

Sanger F, Nicklen S et Coulson AR (1977). “DNA sequencing with chain-terminating inhibitors”. In : *Proceedings of the National Academy of Sciences* 74.12, p. 5463-5467. ISSN : 1091-6490. DOI : 10.1073/pnas.74.12.5463.

Santesmasses D, Mariotti M et Gladyshev VN (2019). “Tolerance to Selenoprotein Loss Differs between Human and Mouse”. In : *Molecular Biology and Evolution* 37.2, p. 341-354. ISSN : 1537-1719. DOI : 10.1093/molbev/msz218.

Schott J, Reitter S, Lindner D, Grosser J, Bruer M, Shenoy A, Geiger T, Mathes A, Dobreva G et Stoecklin G (2021). “Nascent Ribo-Seq measures ribosomal loading time and reveals kinetic impact on ribosome density”. In : *Nature Methods* 18.9, p. 1068-1074. ISSN : 1548-7105. DOI : 10.1038/s41592-021-01250-z.

Searle S, Frankish A, Bignell A, Aken B, Derrien T, Diekhans M, Harte R, Howald C, Kokocinski F, Lin M et al. (2010). “The GENCODE human gene set”. In : *Genome Biology* 11.Suppl 1, P36. ISSN : 1465-6906. DOI : 10.1186/gb-2010-11-s1-p36.

Servant N (2017). “Analysis of chromosome conformation data and application to cancer”. 2017PA066535. Thèse de doct.

Sibley CR, Emmett W, Blazquez L, Faro A, Haberman N, Briese M, Trabzuni D, Ryten M, Weale ME, Hardy J et al. (2015). “Recursive splicing in long vertebrate genes”. In : *Nature* 521.7552, p. 371-375. ISSN : 1476-4687. DOI : 10.1038/nature14466.

Snyder M et Gerstein M (2003). “Defining Genes in the Genomics Era”. In : *Science* 300.5617, p. 258-260. ISSN : 1095-9203. DOI : 10.1126/science.1084354.

Stadler LJ (1954). “The Gene”. In : *Science* 120.3125, p. 811-819. ISSN : 1095-9203. DOI : 10.1126/science.120.3125.811.

Stadler PF, Prohaska SJ, Forst CV et Krakauer DC (2009). “Defining genes : a computational framework”. In : *Theory in Biosciences* 128.3, p. 165-170. ISSN : 1611-7530. DOI : 10.1007/s12064-009-0067-y.

Struhl K (2007). “Transcriptional noise and the fidelity of initiation by RNA polymerase II”. In : *Nature Structural Molecular Biology* 14.2, p. 103-105. ISSN : 1545-9985. DOI : 10.1038/nsmb0207-103.

The ENCODE Project Consortium (2007). “Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project”. In : *Nature* 447.7146, p. 799-816. ISSN : 0028-0836. DOI : 10.1038/nature0587.

— (2012). “An integrated encyclopedia of DNA elements in the human genome”. In : *Nature* 489.7414, p. 57-74. ISSN : 14764687. DOI : 10.1038/nature11247.

The International Human Genome Sequencing Consortium (2001). “Initial sequencing and analysis of the human genome”. In : *Nature* 409.6822, p. 860-921. ISSN : 1476-4687. DOI : 10.1038/35057062.

The Mouse Genome Sequencing Consortium (2002). “Initial sequencing and comparative analysis of the mouse genome”. In : *Nature* 420.6915, p. 520-562. ISSN : 1476-4687. DOI : 10.1038/nature01262.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA et al. (2001). “The Sequence of the Human Genome”. In : *Science* 291.5507, p. 1304-1351. ISSN : 1095-9203. DOI : 10.1126/science.1058040.

Villa-Komaroff L, Efstratiadis A, Broome S, Lomedico P, Tizard R, Naber SP, Chick WL et Gilbert W (1978). “A bacterial clone synthesizing proinsulin.” In : *Proceedings of the National Academy of Sciences* 75.8, p. 3727-3731. ISSN : 1091-6490. DOI : 10.1073/pnas.75.8.3727.

Wade JT et Grainger DC (2014). “Pervasive transcription : illuminating the dark matter of bacterial transcriptomes”. In : *Nature Reviews Microbiology* 12.9, p. 647-653. ISSN : 1740-1534. DOI : 10.1038/nrmicro3316.

Wang MFZ, Mantri M, Chou SP, Scuderi GJ, McKellar DW, Butcher JT, Danko CG et De Vlaminck I (2021). “Uncovering transcriptional dark matter via gene annotation independent single-cell RNA sequencing analysis”. In : *Nature Communications* 12.1. ISSN : 2041-1723. DOI : 10.1038/s41467-021-22496-3.

Watson JD et Crick FHC (1953a). “Genetical Implications of the Structure of Deoxyribonucleic Acid”. In : *Nature* 171.4361, p. 964-967. ISSN : 1476-4687. DOI : 10.1038/171964b0.

Watson JD et Crick FHC (1953b). “Molecular Structure of Nucleic Acids : A Structure for Deoxyribose Nucleic Acid”. In : *Nature* 171.4356, p. 737-738. ISSN : 1476-4687. DOI : 10.1038/171737a0.

Wei LH et Guo JU (2020). “Coding functions of “noncoding” RNAs”. In : *Science* 367.6482, p. 1074-1075. ISSN : 1095-9203. DOI : 10.1126/science.aba6117.

Wright BW, Molloy MP et Jaschke PR (2021). “Overlapping genes in natural and engineered genomes”. In : *Nature Reviews Genetics* 23.3, p. 154-168. ISSN : 1471-0064. DOI : 10.1038/s41576-021-00417-w.

Xu Z, Wei W, Gagneur J, Perocchi F, Clauder-Münster S, Camblong J, Guffanti E, Stutz F, Huber W et Steinmetz LM (2009). “Bidirectional promoters generate pervasive transcription in yeast”. In : *Nature* 457.7232, p. 1033-1037. ISSN : 1476-4687. DOI : 10.1038/nature07728.

Yue F, Cheng Y, Breschi A, Vierstra J, Wu W, Ryba T, Sandstrom R, Ma Z, Davis C, Pope BD et al. (2014). “A comparative encyclopedia of DNA elements in the mouse genome”. In : *Nature* 515.7527, p. 355-364. ISSN : 1476-4687. DOI : 10.1038/nature13992.

Zhao S et Zhang B (2015). “A comprehensive evaluation of ensembl, RefSeq, and UCSC annotations in the context of RNA-seq read mapping and gene quantification”. In : *BMC Genomics* 16.1. ISSN : 1471-2164. DOI : 10.1186/s12864-015-1308-8.

Zufferey M, Tavernari D, Oricchio E et Ciriello G (2018). “Comparison of computational methods for the identification of topologically associating domains”. In : *Genome Biology* 19.1. ISSN : 1474-760X. DOI : 10.1186/s13059-018-1596-9.