De novo analysis of splicing from RNA-seq data: models, algorithms and applications

Vincent Lacroix

Habilitation à Diriger des Recherches
Université Claude Bernard Lyon 1

De novo analysis of splicing from RNA-seq data: models, algorithms and applications

Vincent Lacroix

5 July 2023

Jury:

Hélène Touzet	Directrice de Recherche CNRS, Lille	Rapporteure
Eduardo Eyras	EMBL Australia Group Leader & Professor ANU	Rapporteur
Daniel Gautheret	Professeur Université Paris-Saclaix	Rapporteur
Sophie Schbath	Directrice de Recherche INRAE, Paris-Saclaix	Examinatrice
Céline Brochier-Armanet	Professeure Université Claude Bernard Lyon 1	Présidente

Contents

Acknowlegdments

I would first like to thank Hélène Touzet, Daniel Gautheret and Eduardo Eyras for accepting to act as reviewers for my work, and Sophie Schbath and Céline Brochier-Armanet for accepting to be in my jury. I am honoured to benefit from the expertise of all of you and I am looking forward to the discussion we will have on the day of the defence.

I would like to express my deep gratitude to my mentors, Marie-France and Roderic. Both of you are internationally recognised researchers, passionate for what they do. I feel really lucky to be able to learn from you. I hope I will be up to the task for transmitting the values that you gave me. Optimism, hard work, never quitting, no boundaries, nice atmosphere, having fun, generosity.

I would like to thank my PhD students.

Paulo, you were my first PhD student. I had no idea how to advise a PhD student, but you made it sound easy. You were always positive about the ideas we tried. Always in a good mood, making jokes, working hard when necessary. Working with you was a pleasure. I do not feel I have been advising you, you knew what to do. I am very lucky to have met you.

Gustavo, you joined the group at the very beginning of KisSplice. The prototype was a mess, you rewrote everything from scratch and worked very hard to get it working on real data. The few months before submitting the first KisSplice paper were really intense. I was continuously giving you feedback on counter-examples we had not taken into account in our model. This was tough, but I think we ended up with a very solid work. After the publication, we could rest a bit, but knowing that there was a giant biconnected component that our algorithm could not explore was giving us headaches. With your will, you managed to solve this before the end of you PhD. Congratulations. Really.

Alice, you joined as an engineer hired by INRIA to help develop KisSplice. During the year you spent on this topic before moving to metabolism, you developped kissplice2refgenome, which would enable us to apply KisSplice to pathologies. Thank you for opening these possibilities. I remember you were here at the very first meeting where we talked with Patrick Edery about the Taybi-Linder syndrome. We did not know at the time that this would lead us so far.

Hélène, you were my first co-supervision outside of the team, with Cristina Vieira. This was really great for me to be able to learn on transposable elements, you played also a central role in finishing the SNP paper, and mentoring Mathilde Boutigny to code kissplice2reftranscriptome. Thank you.

Clara, you were officially fully supervised by Didier Auboeuf, but I feel have played the role of a co-supervisor. I think you are very talented, working in medical applications for a bioinformatician is a challenging choice, because people may think you only apply existing methods, when it is so much more. Your role in applying KisSplice to the IAV dataset was instrumental. Thank you. The team you formed with Audric on TALS was also terrific.

Audric, you are so optimistic about science, it helps me, really. Always trying to understand more what is in the data. Finding novel signals, generalising this at the scale of a very large dataset. Your scientific curiosity has no end. You are also very solid. You made a wonderful link with Sylvie Mazoyer and Patrick Edery, who cosupervised your PhD.

Leandro, thank you for existing. Everyone wants to work with you, you are so nice. After several students trying, you were the only one to beat Gustavo's code and propose the current version of KisSplice. You are so modest that you will say it is a very naive and bad algorithm. I disagree. It enabled us to work with instances that we could not finish with the original algorithm. This opened many new applications.

Camille, you were one of the 5 students benchmarking our newly created master

in bioinformatics. I managed to attract you in the group for a PhD during which you initiated our move towards Nanopore data. I am impressed by your solidity. During the PhD, you traversed Covid, and you became a mother. Many students would have quit. You stood up. Bravo.

I want to thank the postdocs and engineers who made essential contributions to the work presented here: Camille Marchet, Janice Kielbassa, Lilia Brinza, Emilie Chautard, Xavi Castells, Eric Cumunel I want to thank all the master students who participated in the life of the lab and contributed to the scientific output presented here; Tiffany Delhomme, Ardi Tampuu, Erwan Scaon, Maxence Morgat, Patrick Van Tran, Sylvain Bonnet, Mathilde Boutigny, Hermes Paraqindes, Victor Deguise, Sylvie Nguyen, Pascal Oberbach, Adrien Raimbault, Pierre Gerenton, Sasha Darmon.

I would like to thank my collaborators who taught me what I know now. One of the reason why I chose this line of work is to be able to keep on learning all my life. Thank you very much to all scientists who accept to share what they know with others. In particular, I want to thank Leen Stougie and Alberto Marchetti-Spaccamela for initiating me to graph algorithms, to Fabien Jourdan for graph drawing, to Franck Picard, Sophie Schbath and Stéphane Robin for graph statistics, to Anne Morgat and Alain Viari for metabolic networks. During my postdoc, the atmosphere in Roderic's lab was simply great. I want to thank Sarah Djebali, Julien Lagarde, Sylvain Foissac, Micha Sammeth, Thomas Derrien, Paolo Ribeca, Christoforos Nikolaou, Hagen Tilgner, who taught me all I know about genomic/transcritomic data analysis. I really knew nothing when I arrived. I also want to thank Juan Valcarcel for introducing me to splicing. Anne Bergeron, who was visiting the lab, for enabling me to publish a RNA-seq paper in 2008.

I want to thank Pierre Peterlongo and Rayan Chikhi for being so enthusiastic and efficient when we started to use de Bruijn graphs for analying SNPs and alternative splicing in raw reads.

I want to thank Didier Auboeuf and Cyril Bourgeois for enabling me to validate experimentally the predictions of KisSplice. I want to thank Patrick Edery, Sylvie Mazoyer, Marion Delous, Rémy Bordonné, Alicia Besson, Audrey Putoux, Anne-Louise Leutenegger who initiated me to the minor spliceosome and the use of patient data. I want to thank Laurent Jacob for initiating me to machine learning and exporting de Bruijn Graphs to bacterial GWAS. I want to thank Nadia Naffakh and Vincent Navratil for initiating me to viral genomics. I want to thank Jean-Marc Aury for initiating me to Nanopore data. I want to thank Rita Rebollo, Cristina Vieira and Marie Fablet for initiating me to transposable elements.

I want to thank Laurent Duret, with whom I had discussions on splicing that deeply changed my way of seeing it. Many splice variants are indeed noise.

I want to thank all the members of the Baobab team, for making it such a great place to work. Thank you Marie-France for creating such a wonderful working environment, international, open-minded, free. Thank you to all the members of the LBBE, I feel honoured to be part of this lab, where I have the feeling that many ideas emerge and then spread. Thank you to Christian Gautier for hosting me in his lab when I was a M2 student. You have been a model for me. Thank you to Dominique Mouchiroud, Manolo Gouy, Fabrice Vavre and Emmanuel Desouhant for renewing their trust over the years.

I want to thank all the colleagues involved in the Bioinformatics Master: Arnaud Mary, Céline Brochier-Armanet, Guillaume Launay, Fabien Duchateau, Philippe Veber, Anamaria Necsulea, Laurent Guéguen, Marc Bailly-Béchet, Carole Knibbe, Sabine Peres, Emmanuel Bettler, Gilbert Deléage, Sophie Ayciriex, Annabelle Haudry. I also want to thank Christelle Lopes, Anne-Béatrice Dufour and Jean Lobry for the Licence BISM I also want to thank Nicolas Lechopier, Lucie Dalibert and Eric Tannier for the course on Ethics in Bioinformatics. I want to thank the group of colleagues who set up the course "Climate and Transitions": Bastien Boussau, Gilles Escarguel,

Anne-Laure Fougères, Ivan Gentil, Chloé Maréchal, Vincent Perrier, Philippe Poncharal, Yann Voituron. I feel privileged to be part of this group, trying to help with my own means. I really hope that this course will provoke deep changes at university and in the rest of the world.

And I want to thank the three sunshines of my life. You give me strength and happiness. Every day.

Chapter 1

Introduction

Writing this document is the occasion for me to summarise and try and take some distance with what I did as a scientist in the last 18 years. I chose a narrative chronological style, trying to tell the story of my trajectory, from my discovery of research during the PhD, followed by my experience as a postdoc, to the definition of my own research lines since I was recruited as an assistant professor. In this document, I will not enter the details of each work presented. I will rather try to present the question addressed and an intuition of the solution proposed to tackle the problem. When relevant, I will also explain the genesis of the project, as I find it interesting to know how new questions arise, often through discussions with colleagues from other disciplines.

During these 18 years, I was not only a scientist. I was also a teacher. This is a part of my work that I enjoy, because the impact on society is more direct. I also find it rewarding and resting. After a day of intense teaching, I am tired but my mind is free. If students fail at their exams, I share only a portion of the responsability. If they succeed, I am happy for them. This document also contains my trajectory as a teacher, from teaching courses that are well defined, to choosing what courses should be taught in the curriculum of a bioinformatician.

Last but not least, in the last few years, I have been strongly affected by the ecological crisis. Before 2018, I had diffuse knowledge that there was a problem but I had not understood its magnitude. This is modifying in depth many aspects of my life, including the way I want to do research. For now, the best action I have found is to participate in an inderdisciplinary group who set up a course for undergraduate students so that they understand the scientific bases of what is the ecological crisis. This led to a MOOC (https://foad.univ-lyon1.fr/course/view.php?id=13), mandatory for all L1 students in Science in Lyon 1, but freely accessible to any interested citizen. In spring 2023 it has also been used in Lyon 2, and other courses from social sciences will be added next year. Concerning my research, I have modified the way I do research (travelling, computing, feeding). I now want to modify the objects of my research. This is an ongoing process.

Chapter 2

My first steps in bioinformatics

2.1 Master

For me bioinformatics is a scientific domain at the intersection of biology, computer science and mathematics. I really like those three domains, and I feel privileged to be able to carry on working in an interdisciplinary field where I do not have to choose one discipline at the expense of another. As a matter of fact, I would have liked to carry on studying history, philosophy, languages, but this was not possible, and when I was 18, I had to choose science, at the expense of the rest. I entered INSA (Institut National des Sciences Appliquées) and the two first years were quite general, with a lot of mathematics, physics and some computer science. This gave me good methodological bases. At the end of the first two years, I chose to enter a newly created department called Bioinformatics & Modelling. The main reason for my choice was that there was a good amount of mathematics (I did not want to go too fast to applications), and room for interdisciplinarity. I learned a lot in this department, at the contact of great teachers who gave a lot of energy and enthusiasm: Jean-Michel Fayard, Guillaume Beslon, Hubert Charles, Christian Gautier, Sandrine Charles, Daniel Chessel, Jacques Estève, Alain Pavé. I also met Marie-France Sagot, who taught a class of algorithmics. It is probably the class where I understood the least what was taught. But this intrigued me and I figured there was a lot to learn from her. During the 5th year at INSA, we were free to do what we wanted (which is really great when I think about it), and I decided to join a DEA (Diplôme d'Études Approfondies) at University. I needed to find a lab where to do my internship. I talked to Marie-France, saying I did not know anything about algorithms, but that it sounded interesting. She said: "you can always catch up from books things others learned on the benches of university". I spent part of my summer reading Gusfield's "Algorithms on Strings, Trees and Sequences" [Gusfield, 1997] and in September, I started working with Marie-France. She put me to work on finding motifs in metabolic networks (she had previously worked on finding motifs in sequences [Sagot, 1998, Marsan and Sagot, 2000], and I later understood the relationship between the two topics). This year of DEA was really central for me. I discovered research and university. At INSA, I had always had the feeling of more or less mastering what I was doing. Of course, there were disciplines that I found harder than others, but I was given exercises where there was a solution. With a reasonable amount of work, I could always get the answer. At university, the feeling of mastering had completely disappeared. For the first time in my life, I realised that the amount of things I knew was extremely limited and the amount of things I did not know was much much larger. I was given problems that did not have any solution. And people

around me were thinking that this was normal. This was a bit scary, but I liked it.

Thanks to Marie-France, my beginnings in research were immediately international, attending conferences and being able to actually meet the people who were writing the papers I had read. She was also very available, giving me advice, freedom and support. When I joined Marie-France's group, there were already strong connections to Brazil, with brazilian PhD students spending part of their PhD in France. I travelled there twice myself (University of Sao Paulo, where I could present my work and initiate a collaboration with Cristina Gomes Fernandes, and Recife, to attend a bioinformatics conference organised by Marie-France). I also travelled to Paris, Lisbon, Bertinoro, Mallorca to attend conferences, and in some occasions present my work. Every time, there was a great atmosphere for meeting, discussing. At Wabi, we met Alberto Marchetti-Spaccamela and Leen Stougie, who are now long standing collaborators. I went to Rome and Amterdam several times during my PhD to work with them on various topics, including elementary modes.

2.2 PhD work

Scientifically, the core of my PhD work was algorithmic. I started from a real example, brought to me by Anne Morgat and Alain Viari, it is reproduced in Figure 2.1.

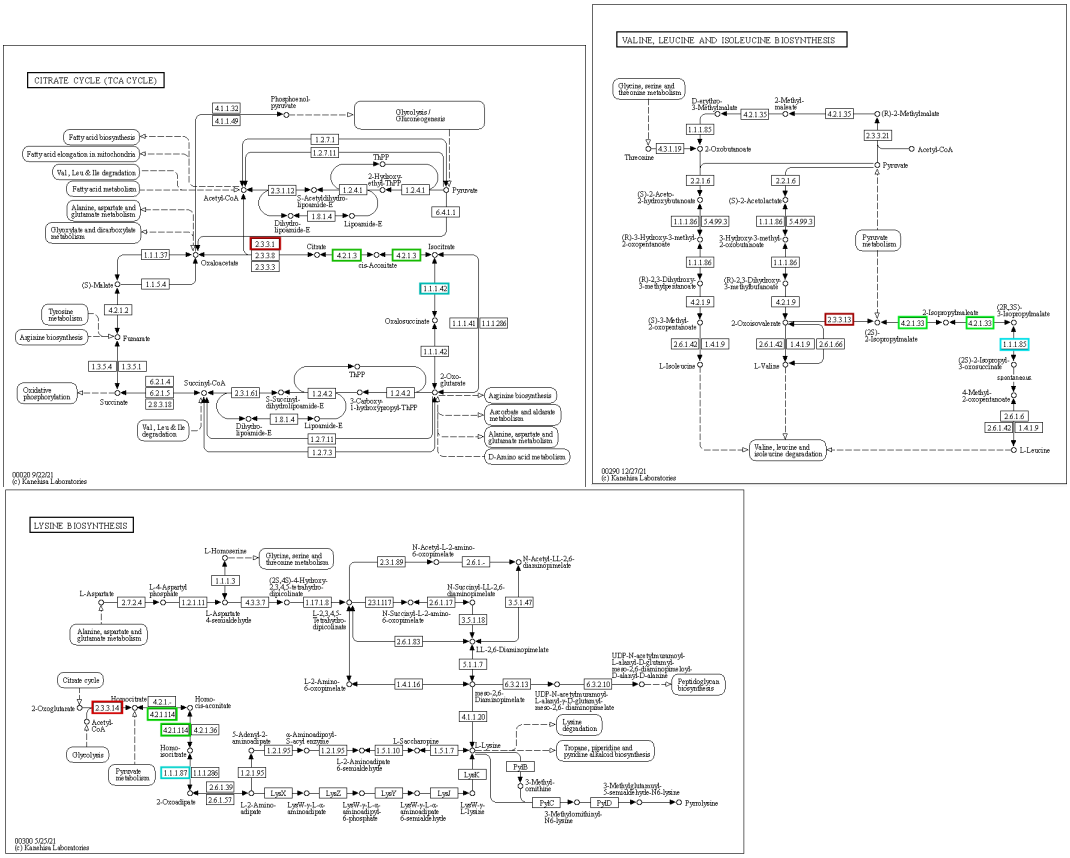

Figure 2.1: The motif 2.3.3, 4.2.1, 4.2.1, 1.1.1 is repeated in three distinct metabolic pathways. Numbers stand for enzyme commission numbers (EC). Each number corresponds to a type of reaction. In each case, the initial substrate is transformed into the final product using an acyltransferase (2.3.3), an hydro-lyase (4.2.1) twice, and an oxydoreductase (1.1.1).

The observation, made by Anne Morgat was that the biosynthetic pathway of leucine, lysine and part of the TCA cycle share the same mechanistic steps. Such pattern is intriguing because it may have a single evolutionary origin and have emerged through duplication and subfunctionalisation of the ancestral enzymes. It may also correspond to convergent evolution.

We set up for finding other repeated motifs in metabolic networks. Was it an isolated example ? Could we find others ? This was the basis for my PhD work. At the time, there was a lot of interest for network motifs, as introduced by Uri Alon [Milo et al., 2002]. Those had been successfully applied to regulatory networks [Shen-Orr et al., 2002] where the topology of a subnetwork can be used to predict the systems dynamics. In our case however, the topology alone could not describe the reaction mechanisms. We needed to introduce node labels in the definition. We decided to define a graph motif as a multiset of node labels. An occurrence was a connected subgraph whose set of node labels matched the motif. This definition captured the example brought to me by Anne Morgat, but it was more general. In particular, it enabled to find occurrences where the order to the reaction mechanisms had been inverted, such as the example presented in Figure 2.2.

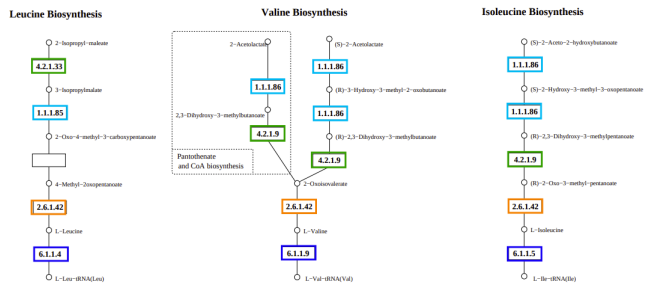

Figure 2.2: The motif 1.1.1, 1.1.1, 4.2.1, 2.6.1, 6.1.1 is repeated in three metabolic pathways: leucine biosynthesis, isoleucine biosynthesis, valine biosynthesis. There is also an occurrence of the motif that is split among two metabolic pathways: "Valine biosyntesis" and "Panthothenate and Coa biosynthesis".

As a matter of fact, in this definition, not only the order could be changed, but the topology itself could be completely changed. A path could be considered similar to a star. During my PhD, I realised that this definition was too general when applying it to metabolism, and I considered alternative definitions where the topology is fixed. My original definition has however been used in other contexts since, such as in the context of protein interaction networks [Bruckner et al., 2010], where nodes are proteins and edges are detected protein interactions. Since the detection of protein interactions is difficult and leads to many false positives and false negatives, it makes sense to use a motif definition where the topoloy is not fixed.

It was also used in the context of social networks, where a team needs to be built based on existing social relationships [Sikora, 2011]. For instance, a team of two software developers, a secretary and a commercial need to work together, and the assumption is that they will work better together if they know each other. Each person is a node in the graph, edges denote their social relationships, node labels denote their job. A connected subgraph induced by the node labels gives a valid team.

Overall, the contribution of my PhD was mostly methodological. It had a limited impact on metabolism but the combinatorial problem I had introduced attracted quite a lot of interest from the graph community, also generating applications to other fields.

For my postdoc, I really wanted to have a deeper impact on biology. I decided to go to Roderic Guigo's group in the Center for Genomic Regulation (CRG), Barcelona, a place where I could get closer to data, and how it was produced. I stayed 14 months

in Roderic's group. It was short, but very intense.

Chapter 3

Postdoc at the Center for Genomic Regulation

3.1 ENCODE & Chimeric RNAs

The first months, I worked on ENCODE, a large project which was funding Roderic's group. When I joined, they had just published the analysis of 1% of the human genome [Birney et al., 2007]. They were by then looking into scaling the analysis to chr21 and 22. They had large amounts of data, Roderic's group was responsible for the bioinformatics analyses. Race-arrays were used. A primer was designed in chr21 or chr22. RACE products (i.e. RNA containing the primer) were analysed. One of the most surprising results when I joined was that many RACE products were extending beyond the annotated transcription termini. In many cases, the product could even be traced to another chromosome. Tom Gingeras, the PI responsible for the project, was really excited about the results, and we spent a large amount of time trying to explain how this data could have been generated. One of the mechanisms that we were thinking of was trans-splicing, i.e. two pre-mRNAs from two distinct loci, are joined together at the splicing step. This mechanism had been very rarely described, hence the excitement. We however first had to rule out other explanations. Using stringent filters, we convinced ourselves that technical artifacts of the RACE-array experiment could be controlled. The next step was to rule out the possibility that these chimeric RNAs could have been generated by genome rearrangements. This turned out to be the most likely explanation, because we were working on Hela cell lines, which are notoriously rearranged. The appealing explanation of trans-splicing fade away. It could be that some chimeras were generated by this mechanism, but fishing them in the crowd of all chimeras was not possible. The whole process took a lot of energy. The paper was submitted to Nature, Genome Research, reviewed twice in each journal but rejected every time, mostly because we were trying to convince reviewers that there was a lot of trans-splicing in the data. As a young postdoc led by an influential PI, I really thought that these chimeras were generated by trans-splicing. I did not know much about this mechanism, and for me it seemed just as likely as any other mechanism. My boss was enthusiastic about it, so I figured it was probably true. I was underestimating the bias introduced by the motivation to make a discovery. This was really an instructive experience. Along the process, reviewers played a very important part, pushing us to ground our claims. The paper was finally published in Plos One [Djebali et al., 2012b].

My interest in chimeras continued afterwards, when the group of Alfonso Valencia contacted us to try and see if any of the chimeras we were detecting, regardless of the mechanism which had generated them, could be supported by proteomics

data. We found some of them, hence renewing the interest of the community in this [Frenkel-Morgenstern et al., 2012].

3.2 The beginning of RNA-seq

In parallel to this work on ENCODE, I worked on what became my own postdoc project. We were in 2008, and this was the beginning of Solexa sequencing (later known as Illumina). Anne Bergeron was visiting Roderic's lab at the time, and we started to work on the following question. Is it possible to reconstruct all the alternative transcripts of a gene based on short RNA-seq reads ? In our first formulation, we considered that all combinations of annotated exons could potentially form a valid transcript. We also considered an optimistic setting where all bases were correctly sequenced, with the number of reads obtained from each transcript being proportional to its abundance. Figure 3.1 summarises the problem we were trying to solve.

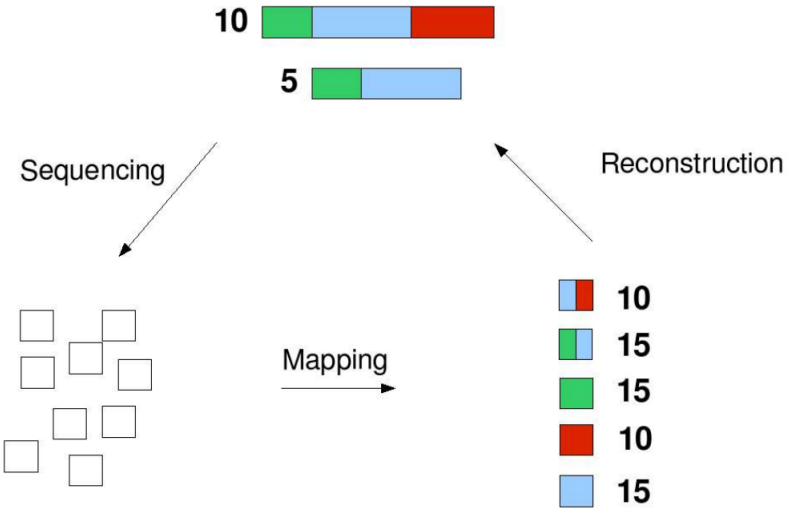

Figure 3.1: **Transcriptome Reconstruction** A transcriptome is a set of alternative transcripts, together with their abundance. Here a gene with 3 exons and two alternative transcripts is represented. This transcriptome is sequenced through short reads. Short reads are mapped back to a reference genome, reads mapping to each exon and exon junction are separated and counted. The problem of transcriptome reconstruction is to identify the transcriptome using only the counts.

We wrote down the equations and we soon realised that, even with a gene with three exons, we were not able to tell apart one transcriptome from another. We were missing information. The problem was not identifiable. An example is given in Figure 3.2.

More generally, we were able to clarify that this non-uniqueness of the solution was caused by sub-structures that we called interchangeable sets, i.e. sets of transcripts which share the same signature in terms of exon and exon junction counts (Figure 3.3).

We next showed that, even with paired-end reads, the problem was still present, even with 5 exons. On the positive side, we showed that, if we restrict the problem to the quantification of transcripts that are already fully annotated (i.e. we do not allow for the discovery of new transcripts), then in practice, the vast majority of transcripts annotated in Gencode are identifiable.

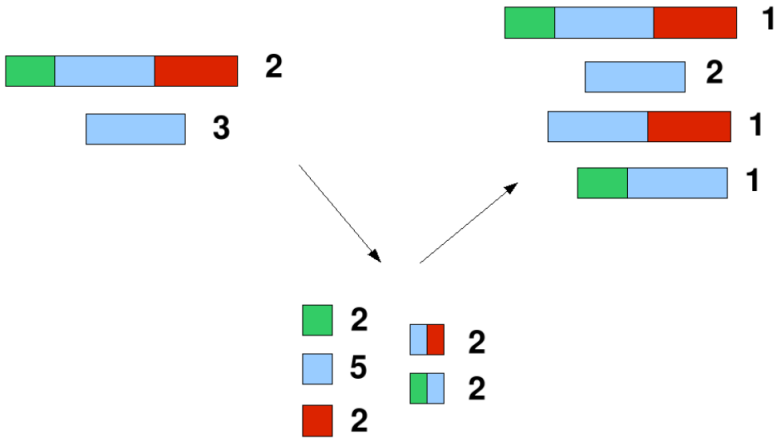

Figure 3.2: **Non unicity of transcriptome reconstruction** Suppose we are trying to sequence the transcriptome located on the upper left of the figure. Each colour corresponds to an exon. The transcriptome contains two distinct transcripts, one with 3 exons (green, blue, red) and one with one exon (blue). The first transcript is present in two copies, the second transcript in three copies. Sequencing this transcriptome with short reads enables to quantify exon and exon junctions. These counts are compatible with the initial transcriptome, but also with an alternative transcriptome containing a different set of transcripts (ex: transcript blue-red was not present in the initial transcriptome).

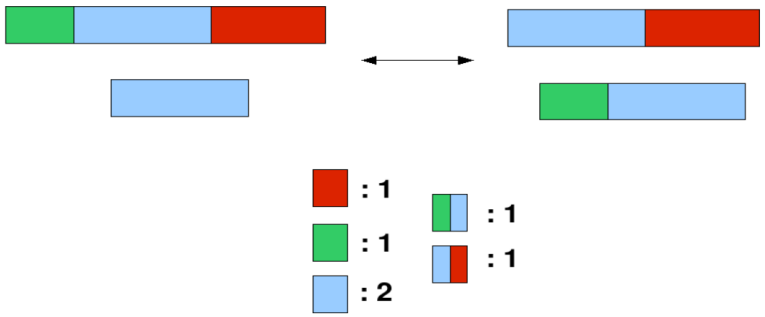

Figure 3.3: **Interchangeable Sets** Sets of transcripts that have exactly the same exon and exon junction counts.

To our knowledge, this issue of non-uniqueness of a solution is still scarcely adressed in alternative splicing analysis (and in many other problems in bioinformatics). It is often assumed that users would not know what to do if several solutions are given. Dedicated tools are required to help navigate the space of solutions, but developing them takes time and is often not considered a priority. We ourselves only reported the issue of non-uniqueness [Lacroix et al., 2008] but did not develop a software to handle it. Very recently, the group of Carl Kingsford proposed to explicitely model non-uniqueness of the solution and propose a confidence range of expression for each transcript [Zheng et al., 2022].

3.3 My splicing initiation

The rest of my post-doc was dedicated to real data analysis. I had the chance to publish quickly my work on the non-uniqueness of transcriptome reconstruction, which enabled me to be less pressured for the remaining of my work there. I worked on analysing data generated through a collaboration with the group of Chris Smith. The link was through Juan Varcarcel. The group of Chris Smith was interested in a splicing factor called PTB [Spellman et al., 2007]. They already had microarray data to study the impact on splicing of knocking down this splicing factor. Through the collaboration with us, they wanted to know if they could reproduce their findings with RNA-seq data. We started to analyse the data, and were able to indeed reproduce part of the results, but there were discordances between the micro-array data and the RNA-seq data. The original purpose of their work was not to perform a comparison of both technologies. As a matter of fact, some of the authors of the collaboration were employees of Affymetrix (who were commercialising the microarrays) and had a limited interest in claiming that RNA-seq could replace microarrays. Besides, at this time, we were blinded by chimeric RNAs and we were also using Chris Smith's dataset to try and find chimeras. This caused us to lose time, not focusing on the main interest of this dataset, which was to characterise the targets of PTB and understand more finely the mechanism of splicing regulation. In the end, the RNA-seq data was not included in the paper, which was published based solely on microarray data [Llorian et al., 2010], but the expertise I had gained was to be used in many other contexts.

During my postdoc, I was also involved in two other side projects. The first one was the attempt to develop an RNA-seq read mapper. At the time I started analysing RNA-seq data, there was no mapper dedicated to this task. In particular, there was no mapper which could identify reads spanning exon junctions, especially for reads as short as 36nt. Together with Paolo Ribeca, who joined Roderic's group as a postdoc when I was there, we started to work on a method that could systematically try all splits for each read (for instance 18-18, 17-19, 16-20, etc), map separately each side of the split to the reference genome, and then, once all reads are split-mapped, summarise the results by reporting exon junctions which were supported by at least two distinct splits (for instance 2 reads with 17-19 and 1 read with 15-21). Our idea was that if a junction is supported by 5 reads which have exactly the same split, this could correspond to a mapping artifact. Unfortunately, although the idea was original, we did not have enough time to develop it, and soon enough, several RNA-seq mappers were published, among which the widely used TopHat [Trapnell et al., 2009], for which Paolo and I acted as reviewers. Paolo continued to work on mapping software and published a general purpose mapper [Marco-Sola et al., 2012] afterwards. The second side project I was involved in was done in collaboration with Micha Sammeth, who was already a postdoc in Roderic's group when I joined. Micha was already an expert in the bioinformatics analysis of alternative splicing [Foissac and Sammeth, 2007] and he was really interested in the potential of RNA-seq data to identify and quantify alternative transcripts of the same gene. We were thinking that read coverage could be used as a proxy for the level of inclusion of an exon in a transcript, i.e. an exon which is often included in a transcript would have a higher read coverage. Inspecting real data, we soon realised that the read coverage of an exon was governed by many other factors, one of which was its position within the transcript. And this depended on the precise RNA-seq protocol that was used. Exons located at the 3'end of the transcript tended to be more covered when using a protocol where cDNA was synthesized using polydT primers. Exons located at the 5'end of the transcript tended to be more covered when using a protocol where cDNA was synthesized uisng random hexamers. Such biases were still present, but to a lesser extent when fragmentation

of transcripts was performed prior to cDNA synthesis. The most challenging part of this project was certainly that, at this time, the RNA-seq protocol was still changing regularly. At each change, we evaluated the novel bias and tried to understand which step had generated it. This led us to propose to the community a software that could realistically simulate RNA-seq reads [Griebel et al., 2012], taking into account explictely each step of the RNA-seq protocol. The final RNA-seq protocol adopted worldwide performs fragmentation prior to cDNA synthesis, and uses random hexamers to initiate first strand cDNA synthesis. The FluxSimulator is still widely used. This work really helped me to understand that it is essential to know how the data is produced, in order not to mis-interpret results during data analysis.

My intention when I had joined Roderic's group was to learn how data was produced and analysed. My expectations were largely fulfilled, and I could clearly have stayed much longer and continued to learn more on genomics and data analysis. I was however interested in settling down in France and have kids. A position of assistant professor was opening in Lyon. I applied and got hired.

Chapter 4

Models and algorithms for local de novo assembly of RNA-seq data

4.1 The beginnings of KisSplice

When I finished my postdoc in Barcelona and came back to Lyon, I brought with me my expertise in RNA-seq data analysis, and I figured I could apply it to the more general case of transcriptome assembly, where no reference genome is available. The idea was to make the most of this technique, which, in principle, did not require any reference genome, and therefore could be applied to any organism, even those for which no genomic resources were available.

The use of de Bruijn graphs (DBG) for assembling genomes traces back to the Pevzner, Tang and Waterman paper [Pevzner et al., 2001] (Figure 4.1. This was then an alternative to the traditional overlap layout consensus (OLC) approach. In the OLC approach, reads correspond to nodes, and overlaps to edges between nodes. In contrast, in the DBG approach, reads are first split into k-mers, i.e. words of length k. Then each k-mer is a node and edges are overlaps of exactly k-1 nt between k-mers. Importantly, a k-mer present in several reads is represented as a single node.

The popularisation of short reads clearly gave a boost to the use of de Bruijn graphs with the use of Velvet [Zerbino and Birney, 2008] for genome assembly and later Oases [Schulz et al., 2012] for transcriptome assembly.

At this time in the lab, we had many informal meetings where we gathered in a room with 5 to 10 people and exchanged ideas. I remember one of these meetings where Alain Viari, Marie-France Sagot, Henri Soldano, Nadia Pisanti, Pierre Peterlongo and I were there. We were discussing about many topics. I was just arriving from Barcelona, and had spent some time trying to understand Velvet [Zerbino and Birney, 2008] and the use of de Bruijn Graphs. I brought this in the list of topics we could discuss. We were wondering how assemblers dealt with SNPs and sequencing errors. We were making drawings and at some point it appeared quite clear that a SNP generates a special pattern in the graph, a bubble (Figure 4.2). A bubble is a vertex-disjoint pair of paths with the same source and target. For a SNP, each path of the bubble has a length of exactly k k-mers.

We did not immediately realise that this idea could be exploited to search for such patterns. The meeting ended, and most of the people from the meeting went to do something else. Pierre and I continued discussing about it during the following weeks, and we said, ok, why not try and devise an algorithm which explicitely searches for these patterns in the raw data. There were already papers published on SNP finding

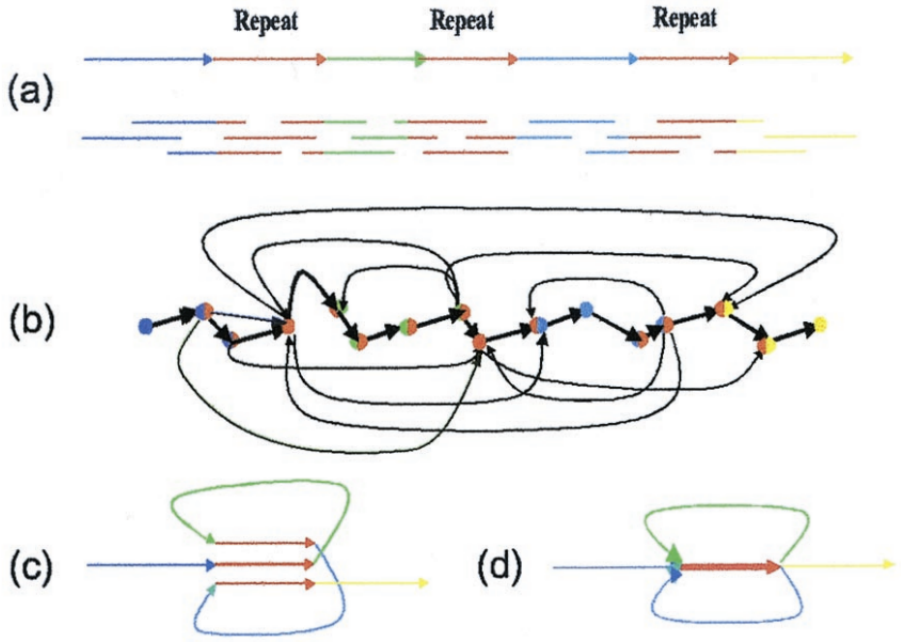

Figure 4.1: **Comparison of the DeBruijn Graph approach and the Overlap Consensus Layout approach**. Figure from [Pevzner et al., 2001] (a) DNA sequence with a triple repeat R; (b) the layout graph; (c) construction of the de Bruijn graph by gluing repeats; (d) de Bruijn graph

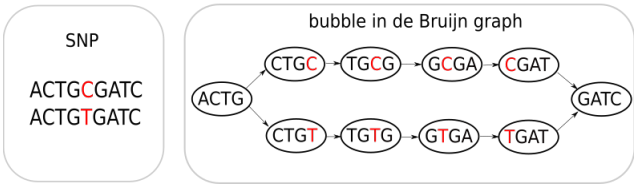

Figure 4.2: A SNP in a genome generates a bubble in the associated de Bruijn graph (here k = 4)

but the vast majority was relying on mapping reads to a reference genome. Initiatives for finding SNPs without a reference genome were either performing a first step of de novo assembly of the short reads themselves with external software and then mapping the same short reads to the assembled reference [Ratan et al., 2010], or using k-mer frequency spectra to identify from the raw reads, those which could be associated to a SNP [Cannon et al., 2010] but the output was k-mers, not SNPs. The originality of our approach was that we were directly searching for SNPs in the raw data.

Two main difficulties had to be dealt with: sequencing errors and inexact repeats. Sequencing errors can in principle generate bubbles in the de Bruijn graph generated from the sequencing reads. An easy solution to discard them is to notice that one of the path of the bubble (the one corresponding to the error) will be supported by very few reads (much fewer than the main path). In practice, in the implementation, an easy solution to deal with this is to discard k-mers seen less than c times in the data. Since bubbles associated with sequencing errors are mostly composed of rare k-mers, this trick removes them for the graph.

Inexact repeats can also generate bubbles in the de Bruijn graph. Discarding these bubbles in more involved because in this case, each path of the bubble will be supported by the same number of reads. Our idea was to notice that in practice most bubbles associated with repeats were branching (Figure 4.3, i.e., at least one node in a path has outdegree or indegree more than 2. The reason why those bubbles are branching is that genomic repeats often correspond to families with more than two copies. This is the case of transposable elements. This is also the case of microsatellites. Such families of repeats do not generate a single bubble but a set of interconnected bubbles. Each bubble in the set is branching, a feature we can use during bubble enumeration. The case where a repeat is present in only two copies and therefore generates a non-branching bubble clearly happens in practice. However, in order to correspond to a bubble, the copies of the repeats need to have diverged at a site flanked by k constant nucleotides. This restricts to low divergence repeats. Our heuristic to restrict to non-branching bubbles was quite efficient in practice. In terms of recall, our main issue was that we could not detect SNPs located less than k nt apart. We clarified this by calling them isolated SNPs.

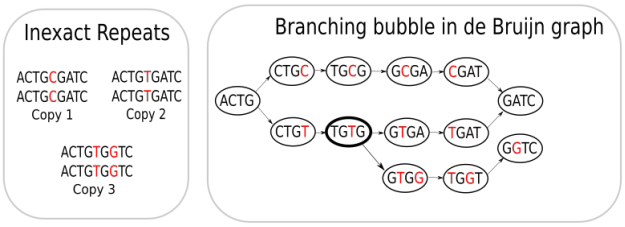

Figure 4.3: A repeat family with three copies generates intertwined bubbles. Each bubble is branching. Node TGTG is branching.

Our first paper on SNPs dates back to 2010 [Peterlongo et al., 2010]. In this paper, we presented the bubble model, a first implementation and a benchmark on simulated data and on Lenski's experimental evolution data [Barrick et al., 2009]. The software was further developped by Pierre. It became DiscoSNP [Uricaru et al., 2015], a software with a very low memory footprint compared to widely used SNP calling softwares like Cortex [Iqbal et al., 2012]. On my side, I started to focus on transcriptomes, where SNPs are also relevant but other patterns emerge.

One feature that was original with RNA-seq data and you cannot find with genomics is alternative splicing. There were several de novo transcriptome assemblers that were published or about to be published at the time [Grabherr et al., 2011, Robertson et al., 2010, Schulz et al., 2012]. In the Trans-Abyss paper, from the group of Inanc Birol [Robertson et al., 2010], one observation attracted my attention. They were claiming that an alternative splice junction would correspond in the DBG to a set of exactly $k-1$ consecutive k-mers flanked by two branching k-mers.

I naturally wanted to clarify what kind of bubble was generated by such an AS event. We worked on the model and found that the shorter path length indeed often contains $k-1$ k-mers, but not always. In fact, every time one of the flanking exons shares a nt with the skipped exon, the shorter path length is smaller.

The model is therefore the following. An alternative splicing event corresponds to a pair of vertex-disjoint paths between s and t, with the shorter path length containing at most $k-1$ nodes.

Ideally, all the bubbles satisfying these constraints would correspond to AS events, and all AS events would correspond to bubbles satisfying these constraints. This is not the case. Some AS events are not captured by the model. This is the case of mutually exclusive exons. This is also the case of alternative transcription start/end,

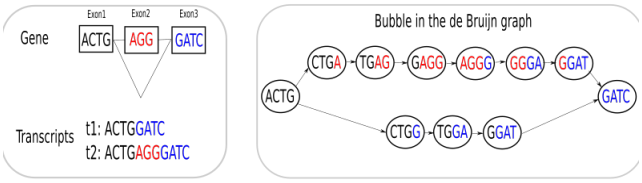

Figure 4.4: A gene with three exons can produce two alternative transcripts depending if exon 2 is included (transcript 1) or excluded (transcript 2). Such alternative splicing event corresponds to a bubble in the corresponding de Bruijn graph. The shorter path of the bubble contains k-1 nodes.

although this is different biologically in the sense that this does not correspond to a splice site choice, but to alternative promoter of polyadenylation, which are distinct biological processes.

We saw earlier that sequencing errors could generate bubbles that could be mistaken for SNPs. This is not true for AS events, because of the constraint that the shorter path contains at most $k-1$ nodes.

Again, the main issue comes from repeats, which can generate many types of bubbles, including the ones satisfying the shorter path length constraint. Our first ideas to rule out false positives was to exploit the idea that, in the case of repeat-induced bubbles, the upper path and the lower path of the bubbles would have very similar sequences. In practice, we computed all bubbles and then discarded all the ones where the edit distance between the upper path and the lower path was below a threshold (3 in practice). This was a pragmatic choice based on the analysis of real examples which gave good initial results.

The first version of KisSplice [Sacomoto et al., 2012] was benchmarked against Trinity [Grabherr et al., 2011] using a dataset of 71M reads from human and brain liver tissues from the Illumina Body Map project (ERP000546).KisSplice found 3497 AS events while Trinity found 1123 out of which 553 were common. This is expected that KisSplice is more sensitive than Trinity because it solves an easier task and can afford to be more exhaustive. What surprised us more was that some events were found by Trinity and not KisSplice. Detailed analysis of a sample of these cases revealed that many of them were true AS events.

4.2 The quest for hidden bubbles

Knowing that we were missing bubbles kept us awake some time. The reason why this happened was simple. In a perfect world where every gene is highly expressed and no k-mer is shared among genes, the DBG associated to a transcriptome contains one connected component per gene. This bijection between genes and connected components is not true for mainly two reasons. First, many genes are poorly expressed, hence they are covered by reads which are not tiling. A single gene may therefore be split in several components. Second, genes may share k-mers and every time they do, they are merged in the same component. This happens for instance in the case of (recent) paralogous genes, which diverged from the same ancestral gene after a duplication event. This also happens when the gene contains a repeated element. Repeated elements are very frequent in eukaryotic genomes. They were thought to be much less present in eukaryotic transcriptomes, because repeats were not supposed to be expressed. Some repeats, like transposable elements, can be expressed when they are active. Old copies of transposable elements may also be expressed just because they overlap a gene, either in the exon, but also in the introns. Introns are not supposed to be present in

RNA-seq data, because the protocol captures polyA+ RNA, which is supposed to be fully processed. In practice, there remains always some pre-mRNA in the mix and some introns are sequenced as well [Tilgner et al., 2012]. In human transcriptomes, the most abundant transposable element is Alu. There is more than 1M copies of this element in the genome [Batzer and Deininger, 2002], many of which lie in introns. The presence of Alu is the main reason why there is a giant biconnected component in human transcriptomes. Discarding the full biconnected component created by Alu is not a good strategy because we may miss important splicing events located in genes with Alu in the introns.

In all the datasets we analysed so far, whatever the species considered (mouse, drosophila, dog, zebrafish, mosquito) there was always a giant component in the transcriptome.

Because we wanted to have a general method which is agnostic to the set of transposable elements that may be present in the transcriptome, we tried to characterise what could be the form of the subgraph induced by repeats in a transcriptome. We focused on high copy number low divergence repeats, because they are the ones that cause trouble when traversing a DBG.

After several attempts to solve the problem directly, we realised it was difficult. One of the formulation indeed leads to a NP-hard problem. This does not mean there is no room for proposing exact algorithms, but we decided to go for another choice. If it is not possible to identify and remove repeats from the DBG, maybe it is possible to avoid to traverse them. A simple idea that we exploit is the number of branching nodes in each path of the bubbles. Bounding to 5 branching nodes gives already very good results, and enables to fish back the bubbles that we knew we were missing. There is still room for improvement in refining this criterion, but a pragmatic approach of 5 branches gives good results. These results on the difficulty of directly identifying repeats in the DBG, and the strategy to avoid them instead of traversing them was published in AMB in 2017 [Lima et al., 2017]. There is clearly room for improvements in explicitely modelling repeats in transcriptome assembly, a topic we will come back to in the perspectives.

4.3 Improving the bubble enumeration algorithm

Along the years, we have made many attempts to improve the bubble enumeration algorithm, both from a theoretical and a practical point of view. One of the difficulty with enumeration problems is that the size of the output can be exponential. Therefore, we cannot hope to have a polynomial algorithm for enumerating all bubbles. We can however try to bound the time spent between finding one bubble and the next. This is what we did in [Sacomoto et al., 2013], where we propose a polynomial delay algorithm (O(n(m+nlogn) for general graphs, where n is the number of nodes and m is the number of arcs). In the particular case of de Bruijn Graphs, where the degree is bounded, we could find a delay of (O(n(m+log α)), where α is the number of nodes in the longest path of the bubble).

From a practical point of view, outputting an exponential number of bubbles is however of limited interest, because end users will not analyse them all. Several possibilities arise. One is to output a basis of bubbles, which is a set of polynomial size, for which we can derive the full set of bubbles using simple operations. This was done in [Acuña et al., 2017]. This line of research is interesting because it enables to implicitly manipulate the full list of bubbles without discarding any. In its current version, the generator set that is used is however not easy to interpret. Many of the bubbles in this set correspond to repeats or combination of repeats and splicing events, and do not have a direct biological interpretation. The second direction we followed is to abandon the objective of enumerating all bubbles and instead focus on a subset:

those which do not overlap high-copy number low divergence repeats, those repeats being the main cause for the combinatorial explosion. In practice, we achieve this by restricting to bubbles with at most 5 branches. This additional constraint enables us to use simple algorithms, which do not necessarily have a good theoretical behaviour, but work well in practice. The algorithm that is currently used in KisSplice was proposed and implemented by Leandro Lima. It is very simple. It starts from a source node s, enumerates all paths that start from this source and have at most 5 branches. Then combines these paths together to try and find bubbles, then removes s and starts from another node. At some point in the computation, the memory could fail because we are storing a potentially exponential number of paths. In practice, this does not happen, because DBG are sparse, and restricting to paths with at most 5 branches prevents us from entering subgraphs with an exponential number of paths. More recent trials show that enumerating bubbles with at most 30 branches still works well. The issue becomes to process the output efficiently. The number of bubbles obtained is much larger. And many are false positives induced by repeats. To summarise, there is a thin frontier between bubbles associated to repeats that we do not want to enumerate, and bubbles associated to AS events that we do want to enumerate. More work on modelling is still required to undertand better where this frontier lies. More thoughts on this are given in the perspective section. Long reads may clearly help.

Chapter 5

Development of KisSplice along with applications to various contexts where splicing is altered

5.1 First application of KisSplice to real data: Myotonic dystrophy

The story of the development of KisSplice is intimately linked to the collaborative work it enabled us to initiate. One of the first collaboration was with the group of Nicolas Charlet Berguerand. It really gave a boost to the project. Nicolas had collected data from patients with Myotonic Dystrophy. He had aquired RNA-seq data but had no skills to analyse it. Didier Auboeuf put me in contact with Nicolas, and we started to analyse his data. At this time, KisSplice was just a prototype and it was the opportunity for us to try and scale to a large dataset. He had a total of 1 billion reads. This was also the opportunity to clarify which events we were missing and how to fish them back. Together with Vincent Navratil, from the PRABI, we ran KisSplice on his dataset (it took one week) and found a large list of genes which were mis-spliced. We gave this list to Nicolas, who said, nice, but where is SCN5A ? It should be in the list. We went back to the data, and were able to flag, in the large graph, which nodes corresponded to SCN5A, then pull from the large graph the subgraph induced by these nodes. The reason why we had missed it was because at this time, our enumeration algorithm stopped after a timeout. It was correctly enumerating bubbles located in most biconnected components of the graph, but was failing on the largest biconnected component, where SCN5A was located. The possibility to inspect the subraph induced by SCN5A confirmed that our idea of restricting to bubbles with less than 5 branches would be efficient here.

What was really nice with this collaboration was that it enabled us to improve our method because Nicolas already knew what he was looking for. This was one of those rare examples where you have access to a false negative of your method. The project was however also frustrating in a way because our method could also find many other candidates that we found as mis-spliced, some of them with a much larger difference between patients and controls than with SCN5A,

However, the transcriptome analysis was just one aspect of the work of Nicolas, and much work had been done already on this gene when we entered the collaboration. It was too late to propose other candidates. The published paper [Freyermuth et al., 2016]

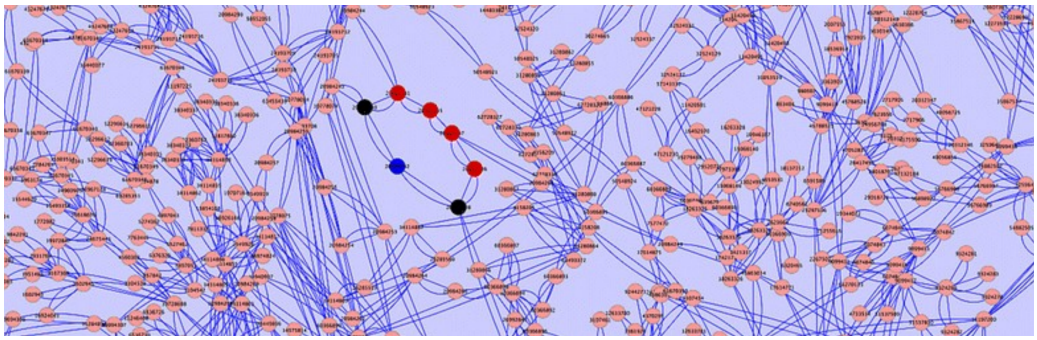

Figure 5.1: A splicing event in SCN5A, flanked by repeats. Red nodes correspond to the inclusion of the exon, blue nodes to the exclusion, black nodes to the flanking exons. Non branching nodes are compacted. The bubble corresponding the AS event is surrounded by complex regions of the graph which correspond to high copy number low divergence repeats.

was therefore centered on SCN5A, and the full list of candidates was published as supplementary material.

This collaboration was really a good opportunity for us to boost KisSplice. This first work would open the door to other collaborations on spliceosomopathies.

5.2 Comparison to a mapping-first approach and experimental validations

The connection with Nicolas Charlet Berguerand had been through Didier Auboeuf, whom I had met through Gilles Thomas, who had recently arrived in Lyon and was setting up his bioinformatics platform for cancer genomics. Didier's expertise in splicing was clearly an ideal match for my starting expertise in the bioinformatics of splicing. We therefore started to work together. Didier had already developed his own bioinformatics software (FastDB [de la Grange et al., 2005]) and he was interested in developing a pipeline for analysing RNA-seq data. His initial idea was to map the reads to the human genome, and analyse reads mapping to exon junctions to assess which exons are skipped and which are included. My approach based on de novo transcriptome assembly was complementary, and we decided to compare the two approaches, to assess which method performed best. The originality of KisSplice is that it is a de novo assembler. It therefore does not require a reference genome. Many other methods, including MISO [Katz et al., 2010], Cufflinks [Trapnell et al., 2010], use a reference genome, and seem therefore more adequate to work with human. There are however several issues when trying to use a reference genome as a first step in the process. First, reads may be poorly mapped if they stem from unannotated exon junctions. Assembling reads into unitigs and mapping unitigs may facilitate this task. Second, reads are short and may map to multiple locations whenever they stem from a repeat which is larger than the read. Again, assembling reads into unitigs prior to mapping may facilitate the mapping, especially if not all copies of the repeat are expressed. In order to explore those two ideas, we focused on a dataset available through the ENCODE project [Djebali et al., 2012a], SK-N-SH cell lines treated by retinoic acid, and searched for exon skipping events using two methods. The first is a mapping-first approach, named FarLine, developed in the group of Didier Auboeuf. The second is KisSplice. The first result we obtain was that the overlap between the two approaches was very low (Figure 5.2).

In particular, there are many exon skipping events that are found only by the

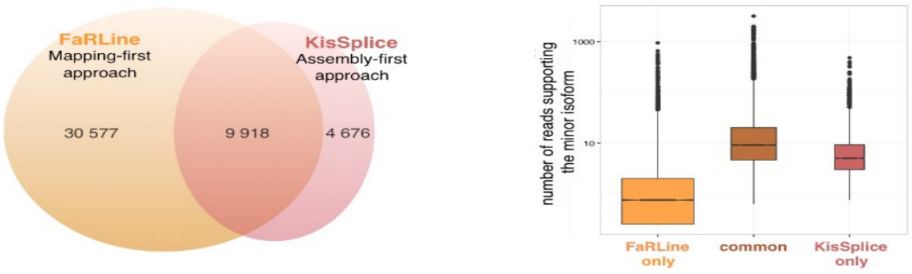

Figure 5.2: Out of a total of 45171 exon skipping events, 22% are found by both approaches, 68% are found only by mapping and 10% are found only by assembly. Most mapping-specific events correspond to rare variants.

mapping-first approach. A closer look at these mapping-specific events reveals that the vast majority correspond to rare variants, supported by less than 5 reads. The question of the biological relevance of these rare splice variants is clearly an open question in the field. If the purpose is to catalog them exhaustively, then the mapping approach should clearly be preferred. The reason why they are not found by the assembly-first approach is simply because they correspond to disconnected components in the DBG and do not correspond to bubbles. This being said, if the purpose is to be exhaustive, then mapping-first is not sufficient either, because there are variants which are not rare, and are found only by assembly. If we choose to disregard rare variants (less than 5 reads and relative abundance less than 10%, Figure 5.3) out of a total of 6444 events, 68% are found by both approaches, 22% are found only by assembly and 10% are found only by mapping.

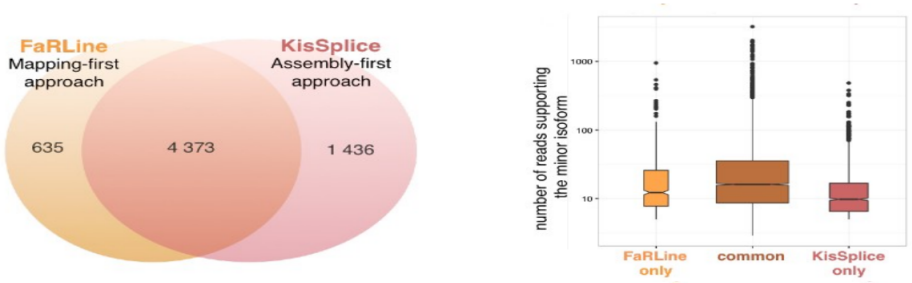

Figure 5.3: After filtering out rare variants (i.e. less than 5 reads, or relative abundance less than 10%), out of a total of 6444 exon skipping events, 4373 are found by both approaches, 635 are found only by mapping, 1436 are found only by assembly

A systematic analysis of events found by one method and not the other reveals that that assembly-first approaches are stonger than mapping-first approaches to identify novel exon skipping events, and AS events located in recent paralogs. Conversely, mapping-first approaches are stronger than assembly-first approaches for finding rare transcripts (i.e. assembly requires more coverage), and to identify annotated skipped exons which correspond to exonised Alus. This is due to the fact that we restrict the enumeration of bubbles to at most 5 branches.

One strong message of our work was that it is possible to go beyond the simple intersection strategy when two methods do not have the same predictions. It is possible to explain why some method fails to find one type of instance. Taking the union of both methods is actually the strategy we recommend.

This collaboration with the group of Didier was also the occasion to extensively confront KisSplice's prediction to experimental validation (Figure 5.4). This was a tough work, long and tedious, but necessary to convince biologists that our predictions could be trusted.

Reassuringly, we could validate experimentally the vast majority of the events we tested (41 out of 48). Out of the events we could not validate, we noticed an enrichment in rare variants and complex events.

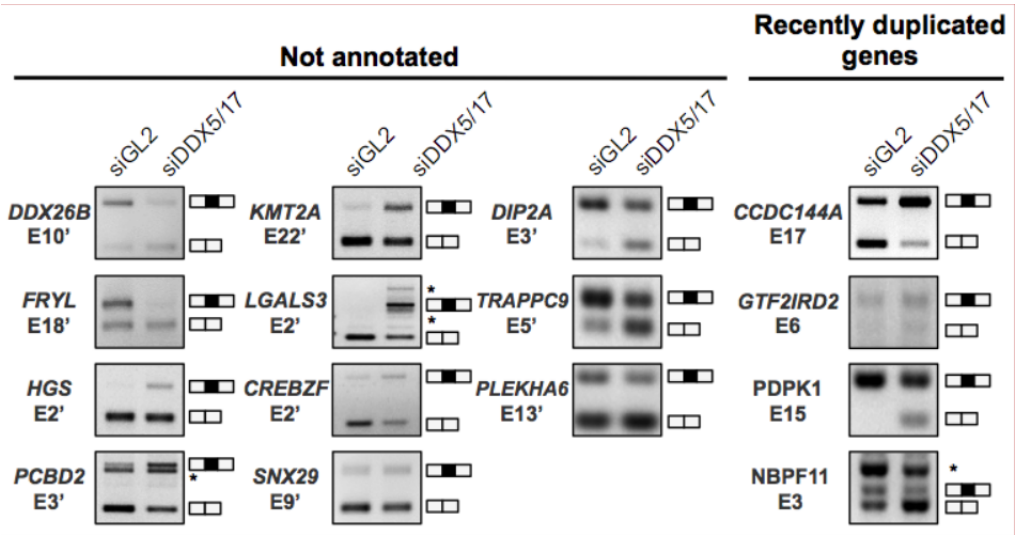

Figure 5.4: Experimental validations by RT-PCR of exon skipping events found only by KisSplice and not FarLine

For rare variants, we decided to dig a bit further and stratified cases to validate in bins of abundance. We found that variants covered by less than 5 reads were much harder to validate. Among the variants which were covered by more reads, but could not be validated, the vast majority had a minor isoform which had a low relative coverage (less than 10%). Overall, this is not surprising that these rare events are hard to validate. Indeed, when using RT-PCR, it suffices that the first rounds of PCR are unsuccessful with the minor isoform for it to be outpaced by the major isoform. This issue is amplified by the fact that the sample used for validation is not the same as the sample used for prediction. Hence, the abundance of the minor isoform may be a bit less in the sample used for validation compared to the sample used for prediction.

For complex events, we did not go much further at the time. In some cases, like LGALS3 or NBPF11, we see additional bands on the gel which correspond to the existence of additional variants. In most cases, KisSplice also predicts these additional bands, but since KisSplice is pairwise, the predictions will be redundant (variant 1 Vs variant 2, variant 1 Vs variant 3, variant 2 Vs variant 3). The possibility to identify and quantify complex events directly (and not pairwise) was later explored in the PhD of Camille Sessegolo [Sessegolo, 2021]

5.3 Taybi-Linder Syndrome and minor spliceosome

One of the positive consequence of having developed bioinformatics software to analyse splicing was that it enabled me to establish collaborations with biologists interested in splicing, and lacking the bioinformatics knowledge. I was contacted in 2014 by Patrick Edery and Anne-Louise Leutenegger, who needed help to analyse RNA-seq data from patients suffering from the Taybi-linder syndrome, a rare spliceosomopathy

caused by a mutation in RNU4ATAC, the gene encoding U4atac, a small nuclear RNA which is an essential component of the minor spliceosome [Edery et al., 2011]. They had recently characterised the mutation and showed, using RT-qPCR, that some minor introns were poorly spliced in patients. I knew nothing about the minor spliceosome. I was intrigued, and I decided to enter this collaboration. This was the beginning of a 6-year collaboration. During this collaboration, I realised that my understanding of splicing was very limited, and that I needed to understand better at least some of the mechanistic details.

I knew that splicing was carried out by a complex cellular machinery called the spliceosome, but I knew very little about it. It happens to be one of the most complex molecular machinery known to date, conserved from yeast to human and is composed of small nuclear RNAs and numerous proteins [Matera and Wang, 2014]. It is estimated that every human cell contains ∼100,000 spliceosomes, which are responsible for removing over 200,000 different intron sequences [Chen and Moore, 2015]. Spliceosome assembly is a complex process which needs to be restarted for each intron. Human cells contain two types of spliceosome: the major spliceosome responsible for removing 99.5% of introns and the minor spliceosome, which removes the remaining 0.5%. They differ in their composition in snRNAs. In the major spliceosome, U1 and U2 are involved in intron recognition, whereas it is U11 and U12 in the minor spliceosome. U4, U5 and U6 are involved in splicing itself for the major spliceosome whereas it is U4atac, U5 and U6atac for the minor spliceosome [Turunen et al., 2013]. Essential components of the minor spliceosome were primarily identified in animals and plants, and then in protists and fungi, indicating an early origin of the minor spliceosome [Russell et al., 2006].

Introns recognised by the minor spliceosome are different. The donor site and the acceptor site are not systematically GT/AG. They can be AT/AC (in 1/3 of the cases), hence the U4atac naming. The branch site is also very conserved. Figure 5.5 shows the two.

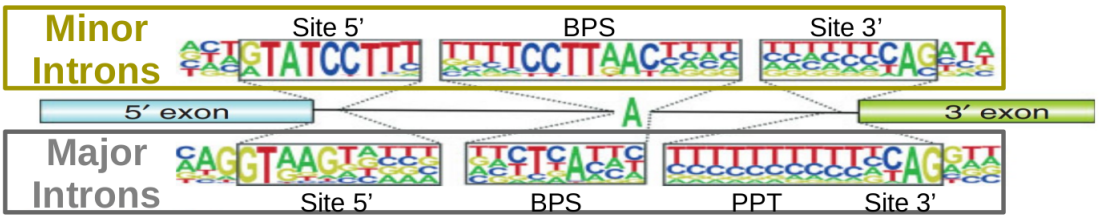

Figure 5.5: Consensus sequence of the human U12- and U2- type introns [Turunen et al., 2013]. BPS: Branch Point Site. PPT: PolyPyrimidine Tract.

Working on this pathology was instructive also because I learned how it was to work with patient data. Not easy at all. The data is precious. Especially for a rare disease. It took us 5 years to build a cohort of 8 patients. This may seem ridiculous for bioinformaticians working on diseases like cancer, but in the case of TALS, with <1000 cases known worldwide, this is really a different world. A consequence of the sparsity of the data was that the tissue sampled for each patient was not always consistent. For most patients, we had fibroblasts (derived from skin samples), and for one patient, we also had a lymphoblastoid cell line (derived from a blood sample). Although in all tissues, we clearly see that minor introns are more poorly spliced in patients compared to controls, the effect is much clearer in lymphoblasts than in fibroblasts. To date, I do no think that there is a clearcut explanation for this. Other spliceosomopathies like retinisis pigmentosa are very tissue-specific, and affect primarily the retina, although the mutated gene is a ubiquitously expressed splicing factor. One hypothesis is related

to the endogenous high rodopsin (a light-sensitive receptor protein) turnover within a limited time interval, which creates a strong demand for a robust splicing machinery [Bujakowska et al., 2009]. In the case of TALS, all minor introns are mis-spliced, hence all tissues should be affected. The main reason for sampling skin and blood is that this is easy to get from patients. Fibroblasts are also considered as relevant tissues, because skin is indeed altered in patients.

One of the main difficulties we had was that we tried to compare to data obtained from patients suffering from another pathology, the Roifman syndrome, related to a mutation in the same gene: U4atac [Merico et al., 2015]. This pathology has overlapping clinical signs, but is much milder. In particular, patients may live much longer (over 30 Vs less than 10 for TALS). The expectation when inspecting molecular data would therefore be that the transcriptome is much more perturbed in TALS than in Roifman. And this is not at all what we saw when we started analysing the data. Minor introns are indeed mis-spliced in TALS fibroblats, but the effect is very weak for each intron. In contrast, in Roifman mononuclear cells, mis-splicing is much more pronounced. Only when we started to analyse our lymphoblast TALS data did we understand that the tissue was a confounding factor. It may seem an obvious point now that we have the explanation, but it took us time, because since we had only one sample from lymphoblasts, we had originally chosen to present only the fibroblast cohort, leaving the lymphoblast TALS aside, not analysed.

Clearly, skin and blood are not the most relevant tissue for this disease. Brain would be the most informative tissue, but this is not accessible for patients. The collaboration evolved towards zebrafish, where most clinical signs of the pathology could be recapitulated. I was less involved in this part. [Khatri et al., 2023]

The data we analysed from fibroblats and lymphoblastoid cell lines however enabled us to make valuable contributions to the fundamental understanding of the minor spliceosome [Cologne et al., 2019]. First, we were able to reclassify some introns as U12 instead of U2 (red dots in Figure 5.6). Second, we could clarify that some introns may be spliced both by the minor and the major spliceosome, with a switch of splice site. Finally, in contrast with previous reports, we clarified that U12 introns are not especially more poorly spliced than U2 introns in physiological conditions. The efficiency of splicing primarily depends on the expression level of the gene. The more highly expressed, the better spliced [Saudemont et al., 2017]. While there is an evolutionary explanation to this, this does not explain all the data, and there is likely room for a mechanistic explanation. If the gene is highly expressed, then spliceosome assembly may be facilitated because components of the splicing machinery are easier to recruit.

To summarise this part, I would say that I learned to work with patient data. I learned what was the minor spliceosome. Among the questions that remain open and for which I would be interested in finding an answer are:

1. What is the basis for tissue specificity of splicing defaults ?
2. Why is there a minor spliceosome at all ?
3. Among the 800 genes that are mis-spliced in patients, which ones are dispensable ?

Among the future directions, the aquisition of proteomics data for lymphoblastoid cell lines might be a good lead towards mechanistic explanations. We already have evidence that the protein levels of INTS7 and INTS10 are decreased in patients [Almentina Ramos Shidi et al., 2023]. Among all U12 genes, clarifying which ones are affected at the protein level would certainly help.

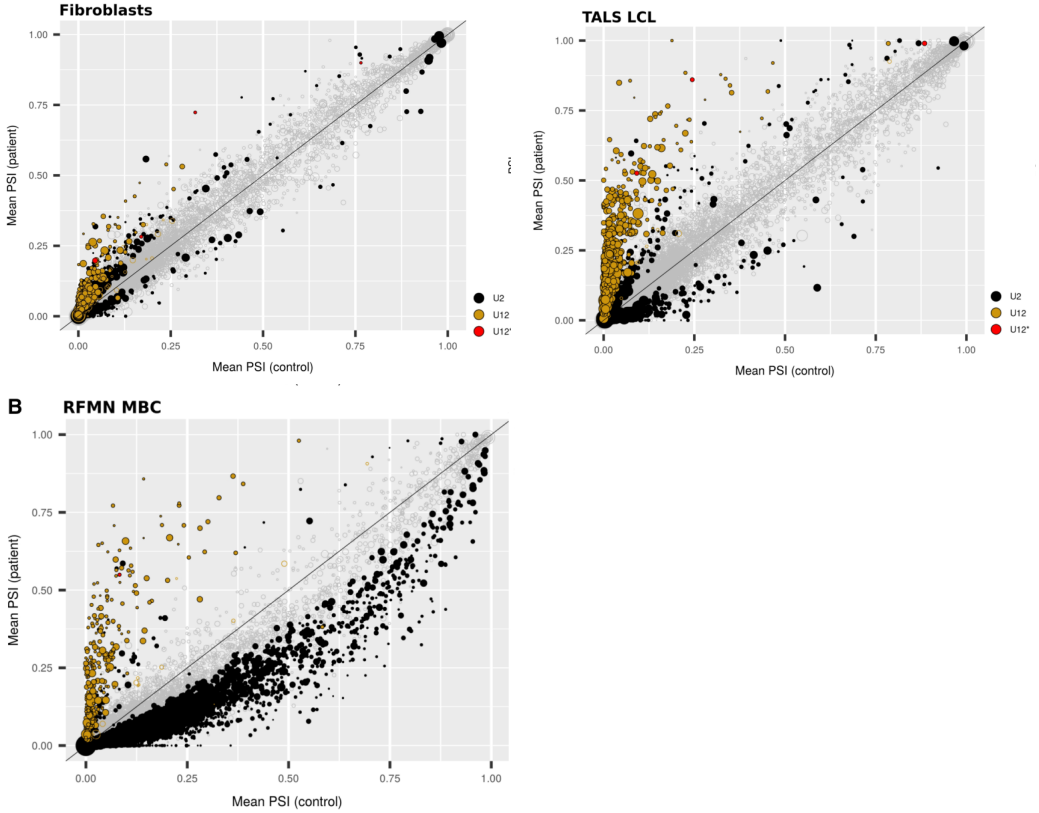

Figure 5.6: Comparison of U2- and U12-type IR levels in TALS patient and control cells. Analysis of the (A) fibroblast data sets or (B) lymphoblast cell lines and (C) Roifman mononuclear blood cells data sets. Plots of the mean U2- and U12-type IR levels expressed with the Percent Spliced In (PSI) metric and obtained for the patients' versus the controls' data sets (PSI-plots). Each circle represents an intron: the color indicates its type (U12⋆ means U2-type intron proposed to be reclassified as U12-type in this study), the size indicates the amount of the corresponding transcript, and the filing status indicates the sig- nificance of the IR level (filled circle: FDR $\leq 5\%$; unfilled circle: FDR $> 5\%$). The intron position respective to the line indicates whether the intron is more retained in patients (above the line) or controls (below the line). The further a point is from this line, the greater the intron dPSI.

5.4 The FluHit project: host-splicing alteration upon viral infection

I was contacted in 2017 by Nadia Naffakh, who had recently identified the RED-SMU1 complex as an essential actor in Influenza A Virus (IAV) infection [Fournier et al., 2014]. These two splicing factors were recruited by the virus to obtain the correct amount of protein NS1 and NEP, NEP being a spliced version of NS1, located on segment 8 of the genome (Figure 5.7). The intuition of Nadia was that, if RED-SMU1 were recruited by the virus, then they would not be available for their natural targets and this would cause mis-splicing of the host genes. More generally, the question was to assess what was the impact of IAV on the splicing of its host, and to what extent this impact was mediated by its interaction with RED/SMU1.

In order to answer this question, we proposed an RNA-seq experiment following the design in Figure 5.8.

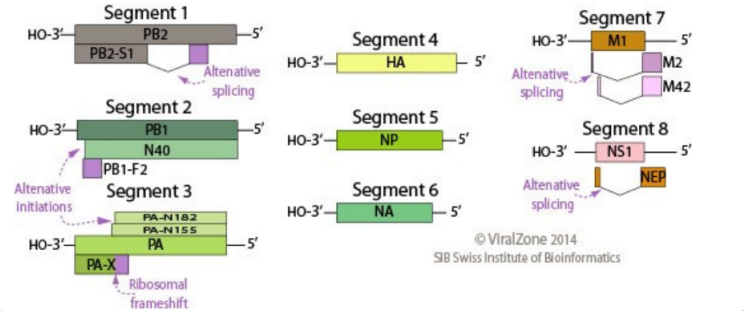

Figure 5.7: Genome of Influenza A Virus. Segment 8 yields two proteins. The relative abundance of NEP is controlled by RED/SMU1.

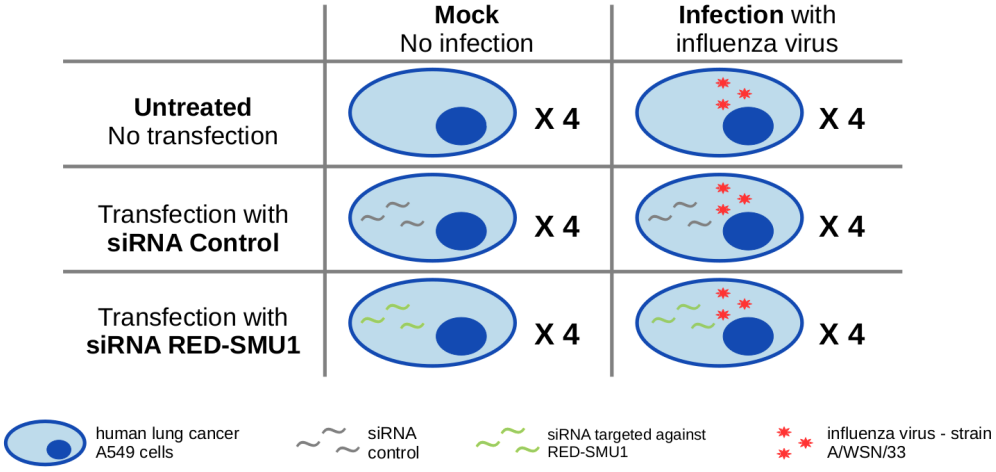

Figure 5.8: Experimental Design of the FluHit project.

The participation to the design of the study was really a nice experience. The number of replicates was confortable. The sequencing depth was not easy to design. Indeed, for infected cells, up to 50% of RNAs are from the virus, not the host. In order to have enough depth to study host splicing, we had to sequence deeper in infected cells (dual RNA-seq) Vs non infected cells. The choice of the strain of the virus, the cell line (A549), the multiplicity of infection (MOI: 5PFU/cell), the time after infection (6h) was chosen by Nadia.

When we started the study, it was already well documented that IAV induced a host transcriptional shutoff achieved by reduction of cellular mRNA levels [Bercovich-Kinori et al., 2016]. We therefore first characterised the impact of the virus infection in terms of gene expression (using Htseq-count and DESeq2). The splicing analysis was novel. We chose to use KisSplice, because it had the benefit of finding novel splice sites, which do not occur in physiological conditions. In the version of KisSplice that we initially used at the beginning of the project, we did not take into account that the RNA-seq data could be stranded. This happened to be crucial for this project. Indeed, Clara Benoit-Pilven was analysing the initial results of the project. Since we had a lot of intron retentions predicted, she was manually inspecting some cases to make sure that KisSplice's predictions were correct. She fell on Prickle1, where we an intron located in the 3'UTR was predicted to be spliced in infected cells (Figure 5.9). The intron was not annotated before, the splice sites were GT-AG and the junction was supported by many reads. The intron seemed trustable. The strange point was that Prickle1 was on the minus strand, and this intron was on the plus strand. Zooming out, Clara realised that these

junction reads were actually stemming from a gene located 10kb upstream on the plus strand (PPHLN1). In non-infected cells, there were no reads mapping downstream the annotated transprition termination of the gene, but in infected cells, reads were mapping several kb dowstream, suggesting that transcription termination was altered in infected cells. We therefore tried to see if this result could be generalised, and it happened to be the case ! We found many other examples (Figure 5.10). Unluckily for us, this discovery was being made by other groups at the same time, and was published before we could finish our work [Zhao et al., 2018, Heinz et al., 2018, Bauer et al., 2018]. We therefore decided to focus our work on splicing, which was our initial focus. There was one paper which reported splicing defaults in IAV infected cell-lines but the analysis was limited, possibly due to the shallow depth of the sequencing experiment, and the bioinformatics analysis was restricted to annotated exons [Fabozzi et al., 2018].

Figure 5.9: Upper track: non infected cells, lower track: cells infected by IAV. Reads map downtream the annotated transcription terminus of gene PPHNL1 in infected cells. This may cause a mis-interpretation of the splicing pattern in PRICKLE1, the downstream gene.

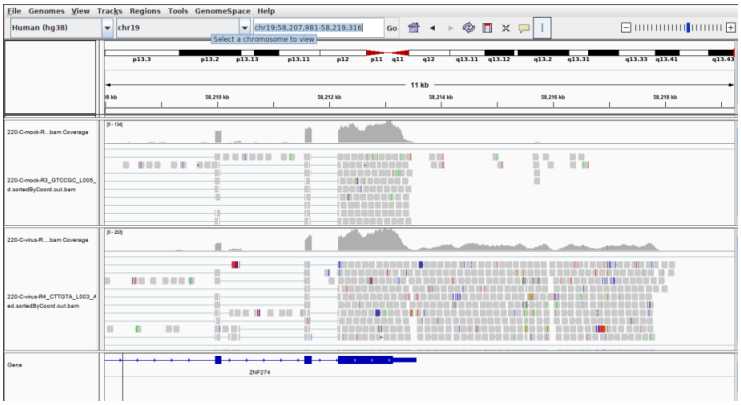

Figure 5.10: Upper track: non infected cells, lower track: cells infected by IAV. Reads map downtream the annotated transcription terminus of ZNF274 in infected cells.

Overall, when we recapitulate the three types of signals (gene expression, splicing and transcription termination), we clearly see that infected cells are very different from non-infected cells (Figure 5.11). Importantly, the genes involved in each type of perturbation are not the same, which suggests that the perturbations in splicing are

not a mere consequence of the perturbation of transcription (Figure 5.12).

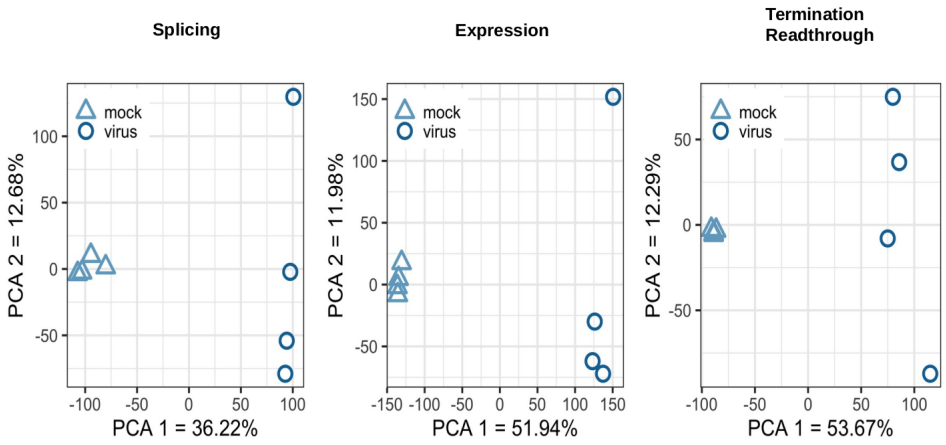

Figure 5.11: Principal Component Analysis of the dataset based on Expression, Splicing and Termination Readthrough. In each case, the first axis of the PCA explains more than 30% of the variance and clearly separates well infected Vs non infected.

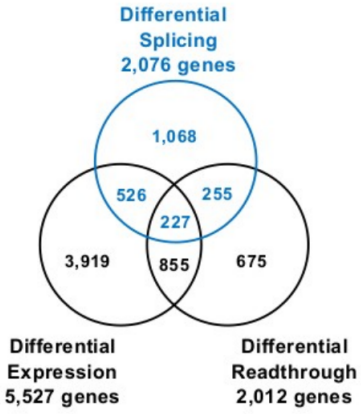

Figure 5.12: Genes whose initation of transcription, splicing, and termination of transcription are altered by the virus are not the same, suggesting that these three perturbations are independent.

Focusing more specifically on splicing, we can add that approximately 25% of AS events were not annotated before (Figure 5.13). These events would have been overseen if we had used a method which relied solely on annotations.

Finally, we also noticed that introns were better spliced in infected Vs non-infected cells (Figure 5.14). This result is counter-intuitive, because we would expect that, since splicing is altered, introns would be more poorly spliced. A possible explanation for this could be that, since transcription is slowed down in infected cells, the mRNAs we sample could seem overall better spliced simply because they have been transcribed a long time ago, hence splicing if fully terminated, whereas in non-infected cells, the pre-mRNAs that are being transcribed are being spliced when we sample. Consistent with this hypothesis, we observe a slightly positive correlation between decreased expression and increased splicing (R^2=5%, p=$7 * 10^{-11}$). This explanation is however most

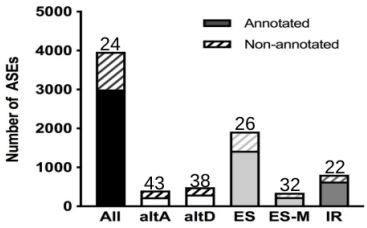

Figure 5.13: For each category of splicing events, number of annotated and novel splice sites.

probably not the only one, because the incompletion of splicing does not concern all introns of a gene, but specific introns. Other mechanisms explaining the "over-splicing" phenomenon remain to be determined.

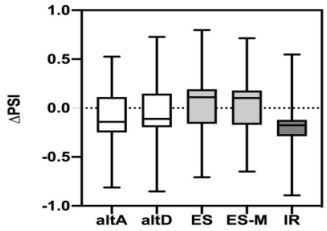

Figure 5.14: For each catergory of splicing event, distribution of deltaPSI. Introns are overall better spliced in infected cells.

Among the genes which exhibited altered splicing in infected cells, we noticed an enrichment in genes involved in the regulation of transcription by RNA-polII (GO:0006357). This could point to so-far unexplored mechanisms for viral-induced host shut-off.

Taking advantage of the full design of our study, we also explored to what extent knocking down RED could recapitulate the splicing alterations seen in infected cells. We did find a set of 678 genes whose splicing was altered both upon RED depletion and IAV infection. Out of those, the majority (491) corresponded to consistent changes (inclusion level of the exon increases, or decreases in both datasets). Altough these numbers are clearly higher than expected under the hypothesis of full independance between IAV infection and RED depletion, there remains a large number of genes whose splicing alterations cannot be explained by RED sequestration by IAV. Other mechanisms seem involved, some which may be direct (through the HNRP K splicing factor [Thompson et al., 2020]) or indirect (splicing factors which are differentially expressed and themselves induce altered splicing).

Chapter 6

Perspectives

6.1 Data perspectives

Developing methods in a field where technologies evolve fast is challenging because methods may quickly be outdated. Since 2008, the read length of Illumina sequencing has kept up increasing. When I started in the field, we had single reads of 36nt, the common practice now is to have paired-end reads of 150nt. In most applications, this read length is largely sufficient and I am convinced many things can still be done in transcriptomics with the data that has been produced in the last 10 years.

Producing more data with a more recent technology seems relevant only for a subset of applications, such as the study of repeats, but the claim I heard in several occasions that long reads will ultimately replace short reads seems ungrounded to me. In the history of technologies, there is no replacement, but addition of a new technology in the repertoire of existing ones.

I must say I am a bit worried by the pace at which data is being produced in the field and I am not sure what I should do as a bioinformatician. One avenue of research is clearly to facilitate data re-use, instead of always producing new datasets that remain superficially analysed.

The temptation for a biologist to aquire his own data is however still very present. Presenting results along with a new dataset obtained with a recent technology can be easier to publish than presenting similar results based on the analysis of an older dataset from someone else. I think it should be the opposite. If novel results can be derived from an existing dataset, then they should be highly valued as they have been obtained with a limited carbon footprint.

There is a need for explicit policies that encourage data re-use and possibly others that restrict data production.

If we do not put any limit, the tendency will be to accumulate more data, the consequence being to buy more computing clusters to be able to process this amount of data. Even though it may give work for many years to the young bioinformatians I train in our master, I do not want to encourage this dynamic. We need to get organised to decide, given a biological question, how much new data needs to be produced. If the amount of new data is too large, then maybe this question can remain unanswered a bit longer.

At my own scale, from a more local perspective, my intention is to continue to provide the community with methods that can use short reads. As much as possible, these methods should be run on a local computer, without the need for a cluster. In this way, I would not encourage biologists to buy new sequencing machines, produce new data, or buy new computer clusters.

Alongside, I want to develop specific tools for dealing with repeats. These methods will first use long reads. I plan to transfer the knowledge I will get from the long read

analysis to the short reads field, so that users who cannot afford the aquisition of long reads can benefit from these methods.

6.2 Methods perspectives

The algorithm that is currently used by default in KisSplice to enumerate bubbles is quite simple. For every pair of vertices s and t, we compute all s,t paths with at most five branches, and combine them in a pairwse manner, so as to form bubbles with a shorter path length of at most $k-1$ k-mers. The number of paths between two nodes can in principle be exponential and this algorithm therefore has a very bad theoretical behaviour. In practice, de Bruijn Graphs are sparse, and the fact that we restrict to paths with a maximum number of branches is key to the success of this algorithm. In terms of running time, KisSplice takes a few hours to run on datasets as large as 1G reads (Table 6.1). The limiting step is not the enumeration of bubbles, but the quantification step. Although the algorithm for quantifying could be improved, we currently consider that the running time is acceptable and we do not plan to invest in this in the short term.

Species	# reads	Genome size	# AS events	Time (mn)
Zebrafish	130M	1.4G	32K	90
Rhodnius prolixus	250M	4.8G	78K	213
Aedes aegypti	415M	1.7G	32K	106
Canis familiaris	960M	2.8G	69K	246

Table 6.1: Running time for KisSplice

The main point we would like to improve is our ability to retrieve bubbles which correspond to true AS events, but have more than 5 branches.

We tried the basic approach of simply allowing to enumerate bubbles with at most 10, 20, 30 branches. We saw that the number of bubbles increases very fast, and most of them are artifactual bubbles induced by repeats. Within this large output, there is however a subset which corresponds to true AS events. Some of them are complex splicing events, where there is a combination of exon skipping events and alternative donor/acceptor sites which together yield bubbles with possibly more than 5 branches. Some correspond to a combination of AS events and genetic polymorphism. For instance, if the skipped exon contains SNPs, each of them can generate up to two branches in the path. Finally, some correspond to a combination of AS events and repeats. For instance if the skipped exon itself corresponds to a repeated element. This happens for instance in human for exonised Alus. This last category is particularly hard to deal with and is most likely underannotated in current genomes. Having a method that can correctly identify these cases of exonisation of transposable elements could be very valuable.

Our first lead towards obtaining an algorithm for finding only relevant high-branching bubbles is to first flag nodes associated to repeats in the DBG, and then search for bubbles with a shorter path containing only regular nodes and a longer path containing repeat nodes. The flagging of repeat nodes can be done using a simple metric that we introduce and corresponds to a generalisation of the notion of degree in DBG. For each node in the non-compacted DBG, β corresponds to the number of nodes that can be reached from this node with a path of at most 20 (this parameter can be adjusted). Nodes located within a repeat will have many direct neighbors. Nodes located outside of a repeat will have few neighbours and hence a low value of β. Once β is computed for each node, we choose a threshold, say 100, and label each node with $\beta > 100$ as a repeat node. Then each connected component of the subgraph induced by the repeat nodes is a distinct repeat family. The choice of the thresholds is indeed delicate, but

is chosen so as to have connected components of reasonable size. We can also check, using repeat annotations obtained through RepeatMasker, than the repeat families we detect indeed correspond to true repeat families. Initial results using data from a dog transcirptome show that some repeat families are split into several components, while some repeat families are joined in the same component. We also see many connected components with high β which do not correspond to any annotated repeat family. There is clearly something to dig deeper in this direction. However, our initial goal is not to identify repeats precisely. An approximate identification may be sufficient. Once this is done, we can consider all pairs of regular nodes (with $\beta < 50$) and try to find all bubbles with a lower path containing only regular nodes and an upper path passing through a repeat component. Importantly, given one entry point and one outgoing point of a repeat component, we do not consider all the paths that connect them. We only say that there is at least one path, and we output the shortest one. In this way, we break the combinatorial explosion which is inherent to this. From a biological point of view, the precise sequence of the inclusion isoform that we output may be wrong, but the structure should be correct. The hope, with this new method, is to be able to catalog exhaustively alternative splicing events which include repeated regions.

During his M2 intership in spring 2022, Sasha Darmon implemented the identification of the repeat components. He should start a PhD in October 2023 and will work on this algorithm.

6.3 Applications perspectives

I have applied KisSplice to various datasets so far. In many cases, I was guided by the taste for novelty in terms of collaborations. Splicing is an entry point in many fields, in particular for pathologies: cancer, rare diseases, infectiology. In the future, I would like to take a step back from applications to human. There are several motivations for that. First, there is already plenty of talented researchers which work on splicing in human, and I feel that it is a shame to spend too much energy competing with other groups, when there is plenty of other things to be done elsewhere. Second, the climate crisis made me realise that being too human centered is most likely not a solution but an issue. We really need to understand better other species, and learn how to share our shelter with them. My transition away from human applications is for now quite smooth because the current projects I am starting all correspond to vectors of human pathogens. My research should therefore benefit to the community of researchers working on the vector for itself, but also to the community of the pathogen.

The first case is the study of *Rhodnius prolixus*, which is the main vector of the Chagas parasite. I started this collaboration in 2022 with the group of Rafael Dias Mesquita, thanks to the connection through Ariel Silbner. Ariel was visiting the lab because he was collaborating with Marie-France on metabolism. We started to discuss about splicing. He was interested and put me in contact with Rafael, who is working on the annotation of *Rhodnius prolixus*. With the help of Victor Deguise, a M1 student, I ran KisSplice on datasets provided by Rafael, we started by two tissues: fat body and ovaries. I was surprised to find a very large number of exons which were differentially included. In order to dig deeper in the list of candidates, two items are envisionned. First translate into proteins the predicted skipped exons to check if they correspond to known protein domains. Second, run KisSplice on more tissues and output, for each skipped exon, the level of inclusion of the exon in each tissue. Such information could ultimately be added to the genome browser that is being developed by Rafael for the community of biologists working with *Rhodnius prolixus*. His first objective is to annotate one transcript per gene. If we also add a track clarifying which exons can be skipped and in which tissues, this could be very valuable. Clearly, a difficulty of

this project is that there are many genes for which the function is unknown. Knowing which exons can be skipped is a notion that will probably be useful only for the subset of genes for which the function is known.

The second application I would be interested in developping in the coming years is with mosquitoes through a collaboration with the group of Jean-Philippe David at LECA in Grenoble. Jean-Philippe has been working on insecticide resistance using RNA-seq data from anopheles, aedes for many years. During the short time I spent in Grenoble, I visited his lab and we started discussing about the possibility to apply KisSplice to his datasets. The first dataset I tried was *Aedes aegypti*. One population that is sensitive to deltamethrin, one that is resistant. Running KisSplice enabled to identify 50 genes which are differentially spliced. In this initial list, only one corresponded to the list of genes known to be involved in insecticide resistance. Following up on this candidate could be interesting, but at the time Jean-Philippe had found a genome amplification (using WGS data) which contained 3 genes known to be involved in insecticide resistance. This mechanism of gene amplification was more promising and we decided not to follow up for now the splicing. During this process, Jean-Philippe made several suggestions which enabled to improve the interface of KisSplice. We can now more easily select candidates with a volcano plot, we also added information on the conservation of each gene. The idea is that if the differential splicing concerns a conserved gene, then it may be more central. Prioritising candidates is a question in itself. I think we learn a lot on this topic from data exploration combined to discussion with experts. One of the motivating examples I learned from this initial data exploration is the case of bubbles where the lower path mapped to a genomic location, and the upper path mapped to many other genomic locations. This is not a true AS event, but since we find it associated to the phenotype, it is interesting to explore. It turned out to be a novel insertion of a transposable element. This transposable element already has many copies elsewhere in the reference genome. The individual we are analysing additionally contains a novel insertion. This polymorphism of TE insertion is common in mosquitoes, not all individuals have the same insertions. What is interesting here is that the allele frequency of this insertion is much higher in the population that is resistant to deltamethrin compared to the population that is sensitive. We found this initial example by chance. We have seen earlier that KisSplice is not designed to deal correctly with repeats (only bubbles with at most 5 branches are enumerated). Such insertions of TE could therefore be much more numerous and correspond to a set of interesting candidates to explain the phenotype of resistance. A systematic scan of all the cases reported with the current version of KisSplice (which is not optimal for dealing with repeats) reveals that all of them are located in 5'UTR of genes. Clearly more work is required in this direction. This should be possible with the starting of the M1 intership of Pascal Oberbach this spring, working this time on an *Aedes albopictus* dataset, again constrasting two populations, resistant and sensitive to deltamethrin. The lessons learned from the analysis of the *Aedes aegypti* dataset should be useful for the analysis of this dataset.

Finally, beyond the use of KisSplice, I am becoming more and more interested in transposable elements for themselves. I recently worked on the identification of expressed transposable elements using Nanopore data from ovaries and testes of drosophila melanogaster. Using long reads, we are able to have reads which correspond to full-length expressed copies. In 99.9% of the cases, we are able to assign each read to a unique genomic location, therefore clarifying which insertion of a transposable element is being expressed. One of the difficulties in analysing the data comes from the fact that, even though reads can be assigned to a unique genomic location, there are many cases where the read overlaps both a gene and a transposable element. In many cases, it is the gene that is expressed, not the transposable element. We can see this because the (long) read spans the full gene and is longer than the TE. The

transposable element is not active anymore. It corresponds to an old insertion, which happens to fall within the gene. In other cases, it is indeed the transposable element that is expressed, not the gene. Among the TE that are expressed by themselves (not as passengers of their host gene), only a fraction is able to transpose. Indeed, many of the expressed copies are degenerate and contain deletions. Interestingly, we see that some of these copies contain introns that are well spliced.

The world that is opening is very complex and intriguing. I had entered the field of transposable elements because they were causing trouble to enumerate AS events. Let's see what the future is like.

One idea that I like is that some transposable elements like Copia have a retroviral origin. Those viruses lost they ability to infect new hosts. The very idea that a host can domesticate its invadors and internalise them in its genome is quite appealing.

Chapter 7

Supervised PhD students

Since 2010, I was involved in the supervision of 6 PhD students. In all cases, I was co-supervising with a colleague. This was a great way for me to learn how to do this. Clearly, I am still learning because each student has his personality, and the relationship between a supervisor and his PhD student is complex and unique. I hope my students enjoyed their PhDs, that it helped them growing up and that what they learned will be useful for them.

Few of my students stayed in academia. My understanding is that it is essentially because the conditions offered in academia (number of years before obtaining a permanent position, salary) are less attractive than in a company, especially when you have good programming skills. None of my students ever applied for a position of assistant professor. On the one hand, I can understand, because the job is not easy (192 hours of teaching is too much). On the other hand, I feel that I have some responsability in this and if I managed to do my job with less stress, my future students may consider applying for such positions. This means slowing down and declining more of the propositions I receive.

Student	Supervision	Dates	Papers	Job
Paulo Milreu	MF Sagot (40%) V Lacroix (40%) C Gautier (20%)	2010-2013	4	Software Engineer at TecSinapse (Sao Paulo, Brazil)
Gustavo Sacomoto	MF Sagot (40%) V Lacroix (40%) P Crescenzi (20%)	2011-2014	5	Software Engineer at Google (San Francisco, USA)
Hélène Lopez-Maestre	C Vieira (50%) V Lacroix (50%)	2013-2016	3	Bioinformatics Engineer Institut Pasteur (Paris)
Leandro Ishi Soares de Lima	MF Sagot (40%) V Lacroix (40%) G Italiano (20%)	2015-2019	3	Software Engineer at EMBL-EBI (Cambridge, England)
Audric Cologne	P Edery (50%) V Lacroix (50%)	2016-2019	4	Bioinformatics Engineer at BIOASTER (Lyon)
Camille Sessegolo	V Lacroix (50%) A Mary (50%)	2017-2021	1	Bioinformatics Engineer at Argaly (Chambéry)

Table 7.1: PhD students, co-supervisors, dates, number of co-authored papers, job obtained after PhD

Chapter 8

Teaching

8.1 First steps

My first experience as a teacher was when I was 16 or 17. I gave math classes to a teenager who was in 4e or 3e. I had to take the bus for one hour to go to the neighborhood where the class was taught. It was located in Lyon 8e, near the Boulevard des Etats-Unis, in a place where I had never been before, where it seemed easy to disconnect from school to hang out with friends in the streets. What I remember most from this experience was the hope that I was bringing to this home when I arrived. The father was telling to his son, look, it is possible to get good grades. The son was really hard working, interested in learning and eager to please his father. It was very rewarding to teach him.

I later had another experience of teaching in a more fancy suburb, where I clearly was paid for no results at all. The kid was not interested in what I was teaching. His father was busy with his career. I had the feeling of being quite useless.

As a student at INSA, I had several opportunities to explain to others. The first year at INSA is competitive and 15% of the students don't make it to the second year. During the first year, I was not living on the campus, but I stayed after the class to work with others as a group of 5 to 10 students. Those were great experiences. Years later, two friends told me I had played a part in their passing their exams. One of them had a STI (Sciences et Techniques de l'Industrie) baccalauréat before entering INSA. He had simply not been exposed to maths enough before and had a lot to do to cover everything in a short time. But he was motivated. My being there simply catalysed his energy. The math teacher we had, Guy Athanaze, saw this energy in the group, and recruited us the year later to accompany a new process at INSA: promotion diversité. One part of this plan was to go to high schools in disfavored neighborhoods and explain to the students that they should try to apply to INSA. Indeed, the first reason why there are no students from these neighborhoods at INSA was that they simply don't apply. The high school had a plan to help them prepare the exam. My role in this was, once some of them had been admitted at INSA, to accompany their first months, with a regular meeting, once a week, to see if they were getting organised, answer their questions, give them a motivation of what it is like to work for becoming an engineer. The idea was that, at home, they did not have any brother, uncle or cousin that had gone through such studies. We were supposed to compensate for this. The role was more like a mentor, not a teacher. My friend Romain was much better than me at this role, partly because he was older, and mostly because he had a chaotic "parcours", he had quit school for 5 years to do ski competition, and went back to school afterwards. Simply thanks to his will, he decided what life he wanted. He was very inspiring for young students. I think I was too normal. Still, the idea of "mentorat" stayed in part of my head. In many cases, I feel that students need

methods, not content, and they need to have a model to which they can identify.

8.2 Monitorat

During my PhD, I naturally applied for a teaching position. It was called "monitorat" and corresponded to 64 hours of teaching per year. I started by teaching maths to first year biology students. I remember I was very surprised by the public. As a student, I really enjoyed calculating integrals, and deriving functions, just for the pleasure. Here, if the modelling was not linked to biology, students would not get involved. For me, what they were asking me to teach them was much harder because they needed to have a real biological problem, which could be formalised in terms of mathematics, then indeed there could be some math calculation to derive some function, and then in the end, the result should be interpreted biologically. I was mostly using exercises that had been gathered by colleagues along the years, but I also had to invent some exercises for the exams. I therefore discovered how hard it was to use real data to create a valid exam. Very often, I had to invent data that fitted my needs. This was less satisfying but much faster.

The next course I taught was bioinformatics for second year students. At the time, it was an optional course (became mandatory in 2016), followed by 100 students. There was a total of 4 classes of 3 hours. The first 2 were dedicated to sequence analysis. Starting from a DNA sequence, students had to predict which protein it encoded using gene finding programs such as Genscan and Augustus, then align the predicted protein against Swissprot using Blast, to decide if the protein was already known, and/or it had any known homologs. This course is often the first contact of students with bioinformatics. Some like it a lot and ask many questions. Others feel overwhelmed and have a hard time figuring out what is expected from them. Knowing which amount of guidance is required for each student is really one of the difficulties I have when teaching bioinformatics. I used to spend a lot of time on students who do not ask questions, making sure they had not missed the bus. I now target more the average audience, and try to give some advanced questions to the subpopulation of students who want more.

8.3 Assistant Professor

When I was recruited "Maitre de Conférences", I started from doing the same courses as the ones I had done during my monitorat. This helped a lot because I had fewer things to prepare. The novelty was that I was now responsible for the bioinformatics course in second year (15 hours, 100 students), which meant recruiting teachers, reserving rooms, preparing exams. Together with Marc Bailly-Béchet, we decided to focus the course on PAH, an enzyme responsible for the degradation of phenylalanine. Mutations in this gene have been associated to phenyketonuria, a disease which can be detected early and compensated by a regime with low amounts of phenylalaline in the meals. Indeed, this mutation is not detrimental if mutated individuals are not exposed to phenylalaline. This happens to be a nice example of genotype-environment interaction, mentionned in the Genetics course in 2nd year.

I was also given the opportunity to teach a course of 3rd year students, who were wishing to get specialised in bioinformatics, statistics and modelling. The group was limited to say 15 students, but the students were really interested. In this class, they learned how to program with Python, with at the end, an implementation of the Needleman & Wunsh algorithm. They also learned how to use standard bioinformatics software (orf finding, blast to find homologs, multiple alignment, phylogenetic tree reconstrution) throught the Annotathon platform (annotathon.org). Finally, I could insert a practical class where the students could manipulate real Illumina data. This

was possible thanks to the support of the IT in LBBE who gave me access to a machine where I could create accounts for students, and install the software I needed. This was in 2010. At this point, very few students in 3rd year could manipulate real sequencing data. I was quite lucky to be able to teach something that was close to my research. I feel that the students were also interested, some of them were really taking advantage of the freedom I let them on the data, the first version of the subject was indeed very free. We were discussing together the results they were obtaining. The skill to analyse RNA-seq data was also directly useful for them to find internships.

After 10 years, I am still teaching this class of 3rd year students. The work is now much more guided, and the number of students has doubled (now 30 students). This year the course also opened to students from Licence Biodiversité, some of which have not been exposed at all to programming before.

At the master level, when I was recruited, I started by teaching at INSA, because the bioinformatics curriculum was not yet developed at university. At INSA, I was responsible for various courses in 4th year: genomics, transcriptomics. The courses were interesting and I had opportunities to attract students interested in research. Students had a good level in maths. Compared to university, they were a little less autonomous though. Along the years, it became more difficult to maintain this link between INSA and university because a new constraint was introduced: out of the two internships that the students from INSA had to do, one of them at least (and possibly both) had to be in industry. As an ex-student from INSA, I think this is really a shame to add obstacles to students who want to go to academia. The connection between so-called "grandes écoles" and university is already not very developed. I think that many more students from these "grandes écoles" could have been interested in research, but simply did not have the opportunity to try it.

At University, when I was recruited, there was no master in bioinformatics. Students graduated with a Licence in Bioinformatics usually applied either to a master in health, or in ecology. In each of these masters, there were methodological courses, and I participated in teaching these. Along the years, it appeared very clear that there was a need to increase the visibility of bioinformatics at university and create a master specific to this. This is what we did in 2016. This was a crazy adventure, with a lot of work to get everything prepared for the first students to come. Several courses had to be set from scratch, trying to find the right people to teach the notions thought to be essential. This was a very interesting experience. It involved colleagues from Biology (Céline Brochier-Armanet, Arnaud Mary, Marc Bailly-Béchet), from Biochemistry (Gilbert Deléage, Emmanuel Bettler, Guillaume Launay) and Computer Science (Fabien Duchateau, Carole Knibbe). Marc Bailly Béchet and Carole Knibbe, who were very active, unfortunately had to leave the adventure along the way. Marc was recruited in Nice in 2017 and Carole was recruited at INSA in 2018.

Chapter 9

Science and society

9.1 Research Ethics & Ethics for Bioinformatics

During my postdoc in barcelona, I was really interested in genomics, but I had the feeling that my boss was too optmistic about science. For him, there should be no frontiers on what can be done. We should sequence as many genomes as we could. Increasing knowledge was the most important goal, whatever the applications downstream. I really liked his enthusiasm about science and I still have a lot of respect for people who are optimistic about science, but on the other hand, I do not feel like this and my impression is that there is a large disconnection between science and citizens. My way of fighting against this disconnection was to participate to the organisation of a scientific cafe, downtown Barcelona, where we would invite scientists to present their work to a general audience. I could see that people were really interested. Scientists were also happy to present their work.

Coming back to Lyon, I realised that there was the work of Leo Coutellec who had done a summary of what existed in terms of training towards ethics in Lyon (https://studylibfr.com/doc/3680822/les-paysages-de-l-%C3%A9thique), I attended his conference and discussed with people there. I then joined an interdisciplinary group which was forming to build up a course on research ethics. It gathered people from philosophy, history, biology, from different institutions within University of Lyon (one of them was Nicolas Lechopier, with whom I later worked on ethics in bioinformatics). There was a national incentive to train PhD students to research integrity. We widened this to the question of ethics of research. This led to a course proposed to all doctoral schools. After some talks by colleagues, students were invited to present the ethical questions that were present in their PhD work. I could attend several of the classes and I found it very interesting. Correctly formulating the ethical question is often a great starting point.

Since it had to be proposed to a much wider audience, the course became a MOOC. Sarah Carvalo took it in charge. Not being an expert at all in this field, I did not participate to the continuation of this adventure. My role consisted in introducing my colleague Eric Tannier to the group. He participated in the MOOC, now available here: https://www.fun-mooc.fr/fr/cours/ethique-de-la-recherche/

I thought that the reflexive work done by the students was really interesting and when we launched the master in bioinformatics, I immediately insisted for inserting a course on ethics. This was possible under the name: Data Protection, Bioethics and Law. Nicolas Lechopier joined the adventure from the start, with an introductory course on ethics and epistemology. The very first year, there were only 5 students in M2 and we had opened it to volunteers from the master of ecology. It was really a new way for me to teach. The teacher was not standing in front of the class. Tables were organised in a circle. Nicolas would bring initial content in ethics and epistemology,

but the participation was really encouraged. The status of the knower was changed. Every year, we organise a joint course where both Nicolas and I are present. I submit a real case that happened to me as a young bioinformatician and the questions it raises. Then we discuss it together with the students. Along the years, the real cases have also been brought by colleagues and students. To evaluate this course, we ask the students to write a short report on the ethical and law questions that arise in the context of their M2 internships. They have to hand this report in early April. The date is important. It should not be too early because students will not have really entered their topic yet (interships start in mid-january). It should not be too late, because it will enter in conflict with the time they spend on writing their scientific report (June).

Along the years, there has been a diversity of teachers in this course, Jos Kafer, Catherine Bourgain, Alain Viari, Laurent Lefevre. The core team of teachers is Lucie Dalibert, Nicolas Lechopier, Yann Bergheaud, Eric Tannier and I. The current name of the course is "Enjeux sociaux, juridiques, éthiques et environnementaux de la bioinformatique". The part on environmental impact of computing is handled by Eric Tannier and Laurent Lefevre.

9.2 Climate and Transitions

In 2020, I received an email from students who were setting up a group of students and teachers to prepare the COP2 which was to be organised in Grenoble in spring. I registered to the first meeting, and I really liked the format. There was a lot of enthusiasm from the students. We had regular meetings (online - covid time) every week during for 10 weeks. It was intense, but defined in time with the final objective of the COP2 in April 2021. There were a lot of ideas that came out of the discussions we had. The representative students attended the COP2 and were able to transmit our proposals. Many universities have signed the COP2 and started implementing changes (https://la-ctes.org/). Some universities even have a full department devoted to the ecological transition. In Lyon, the situation is much less advanced for now.

On the positive side, during the process, Céline Brochier-Armanet, whom I had contacted to know what could be done in terms of teaching the scientific bases of the ecological crisis, had put me in contact with Philippe Poncharal, who was setting up a group of teachers who had expertise in the field. He was interested in having students enroled so that they could state if the course was adapted to their knowledge. Regular meetings started on the setting up of this course. I started by attending these meetings together with the students from the COP2 group. The students could not attend in the long term because it was too demanding and interfered with their studies. I continued and learned enourmously from this group of experts. Vincent Perrier (Energy), Gilles Escarguel (Anthropocene) and Chloé Maréchal (Climate) had already produced content for an optional course they had been teaching for several years. Ivan Gentil, Anne-Laure Fougères, were also extremely pushing towards changing in depth teaching, the group was enlarged to Yann Voituron (Biodiversity) and Bastien Boussau (Feeding). I helped Bastien a little bit for setting up the content. Overall, my contribution to the production of scientific content was limited because I am not an expert. I focused on trying to keep the content accessible to students, making propositions for the quizz sections and the amphi-débats. It was essential for us to have a place where we could really meet the students and discuss all this. We thought that there was a real risk of eco-anxiety and that the students would need to come and talk about it. Given that the course is common to students from maths, physics, biology, etc, it was difficult to find shedules where everyone was available, we were given schedules from 17:30 to 19:00 and we made the terrible mistake to say that the presence at these courses was not mandatory. The results was that there were extremely few

students who attended the amphi-débats. The next year, we completely changed the organisation and we started the year with an amphi in presence where we explain what will happen. We also organised activities that helped the students to participate. The first was inspired from a discussion with a colleague working in Education Sciences, Olivier Morin, who encouraged us to try and orgnanise the students spatially using two contrasting questions. The first was: "in your mind, does the ecological crisis take a central or peripheral place ?" answer from 1 to 10. The second was: "in terms of action, do you think that you can have possibilities to act and make a change" answer from 1 to 10. The answers to these two questions gave them coordinates in a plan with two dimensions. They had to move spatially in the amphitheater. Then they had to discuss with the person standing next to them (this person had given similar answers to the two questions) to understand what had motivated their choice to answer this. Then enlarge to a group of 10 people (there was a total of 150 students), and clarify a series of 2-3 arguments that they had to restitute orally to the full group. I find that this worked really very well, and I was extremely happy that it worked. At this point, there was no content we were trying to teach them. They were expressing themselves freely, and happy to have the microphone. In very few cases, some formulated climato-skeptical arguments, but in most cases the output was informative. Some hopes, few students already aware and active, a big mass of students waiting to know in which direction to head.

Along the semester, there was a total of 4 occasions where we met the students and proposed them activities to provoke reactions. One was centered on the analysis of their individual carbon footprints and was the occasion to discuss about the individual action anyone could take to try and make a change. The main area where students can act is feeding, which represents 1/3 of their carbon footprint (Vs 1/4 of the footprint for the average french inhabitant). The main action is to lower their consumption of meat, especially beef. They also realised that their carbon footprint was lower than the average in France, essentially because they had a lower income. Increasing awareness in their families, friends is therefore expected to have a larger impact. The last session with the students was the occasion to discuss about collective actions that could be taken to make a change. Clearly, individual actions are required but not sufficient at all. A level where students can make proposition is university, where they can ask for actions concerning transportation, feeding, energy consumption. These collective measures have a wider impact, but are more involved because they require to convince a larger number of people, and they require some level of coordination that the students do not always have. Some of the proposals made by students were really inspiring. Encouraging and accompanying those initiatives should really be a priority for the university.

Bibliography

[Acuña et al., 2017] Acuña, V., Grossi, R., Italiano, G. F., Lima, L., Rizzi, R., Sacomoto, G., Sagot, M.-F., and Sinaimeri, B. (2017). On Bubble Generators in Directed Graphs. In Bodlaender, H. L. and Woeginger, G. J., editors, *Graph-Theoretic Concepts in Computer Science*, Lecture Notes in Computer Science, pages 18–31, Cham. Springer International Publishing.

[Almentina Ramos Shidi et al., 2023] Almentina Ramos Shidi, F., Cologne, A., Delous, M., Besson, A., Putoux, A., Leutenegger, A.-L., Lacroix, V., Edery, P., Mazoyer, S., and Bordonné, R. (2023). Mutations in the non-coding RNU4ATAC gene affect the homeostasis and function of the Integrator complex. *Nucleic Acids Research*, 51(2):712–727.

[Barrick et al., 2009] Barrick, J. E., Yu, D. S., Yoon, S. H., Jeong, H., Oh, T. K., Schneider, D., Lenski, R. E., and Kim, J. F. (2009). Genome evolution and adaptation in a long-term experiment with Escherichia coli. *Nature*, 461(7268):1243–1247. Number: 7268 Publisher: Nature Publishing Group.

[Batzer and Deininger, 2002] Batzer, M. A. and Deininger, P. L. (2002). Alu repeats and human genomic diversity. *Nature Reviews Genetics*, 3(5):370–379. Number: 5 Publisher: Nature Publishing Group.

[Bauer et al., 2018] Bauer, D. L. V., Tellier, M., Martínez-Alonso, M., Nojima, T., Proudfoot, N. J., Murphy, S., and Fodor, E. (2018). Influenza Virus Mounts a Two-Pronged Attack on Host RNA Polymerase II Transcription. *Cell Reports*, 23(7):2119–2129.e3.

[Bercovich-Kinori et al., 2016] Bercovich-Kinori, A., Tai, J., Gelbart, I. A., Shitrit, A., Ben-Moshe, S., Drori, Y., Itzkovitz, S., Mandelboim, M., and Stern-Ginossar, N. (2016). A systematic view on influenza induced host shutoff. *eLife*, 5:e18311. Publisher: eLife Sciences Publications, Ltd.

[Birney et al., 2007] Birney, E., Stamatoyannopoulos, J. A., Dutta, A., Guigó, R., Gingeras, T. R., Margulies, E. H., Weng, Z., Snyder, M., Dermitzakis, E. T., Stamatoyannopoulos, J. A., Thurman, R. E., Kuehn, M. S., Taylor, C. M., Neph, S., Koch, C. M., Asthana, S., Malhotra, A., Adzhubei, I., Greenbaum, J. A., Andrews, R. M., Flicek, P., Boyle, P. J., Cao, H., Carter, N. P., Clelland, G. K., Davis, S., Day, N., Dhami, P., Dillon, S. C., Dorschner, M. O., Fiegler, H., Giresi, P. G., Goldy, J., Hawrylycz, M., Haydock, A., Humbert, R., James, K. D., Johnson, B. E., Johnson, E. M., Frum, T. T., Rosenzweig, E. R., Karnani, N., Lee, K., Lefebvre, G. C., Navas, P. A., Neri, F., Parker, S. C. J., Sabo, P. J., Sandstrom, R., Shafer, A., Vetrie, D., Weaver, M., Wilcox, S., Yu, M., Collins, F. S., Dekker, J., Lieb, J. D., Tullius, T. D., Crawford, G. E., Sunyaev, S., Noble, W. S., Dunham, I., Dutta, A., Guigó, R., Denoeud, F., Reymond, A., Kapranov, P., Rozowsky, J., Zheng, D., Castelo, R., Frankish, A., Harrow, J., Ghosh, S., Sandelin, A., Hofacker, I. L., Baertsch, R., Keefe, D., Flicek, P., Dike, S., Cheng, J., Hirsch, H. A., Sekinger, E. A., Lagarde,

J., Abril, J. F., Shahab, A., Flamm, C., Fried, C., Hackermüller, J., Hertel, J., Lindemeyer, M., Missal, K., Tanzer, A., Washietl, S., Korbel, J., Emanuelsson, O., Pedersen, J. S., Holroyd, N., Taylor, R., Swarbreck, D., Matthews, N., Dickson, M. C., Thomas, D. J., Weirauch, M. T., Gilbert, J., Drenkow, J., Bell, I., Zhao, X., Srinivasan, K., Sung, W.-K., Ooi, H. S., Chiu, K. P., Foissac, S., Alioto, T., Brent, M., Pachter, L., Tress, M. L., Valencia, A., Choo, S. W., Choo, C. Y., Ucla, C., Manzano, C., Wyss, C., Cheung, E., Clark, T. G., Brown, J. B., Ganesh, M., Patel, S., Tammana, H., Chrast, J., Henrichsen, C. N., Kai, C., Kawai, J., Nagalakshmi, U., Wu, J., Lian, Z., Lian, J., Newburger, P., Zhang, X., Bickel, P., Mattick, J. S., Carninci, P., Hayashizaki, Y., Weissman, S., Dermitzakis, E. T., Margulies, E. H., Hubbard, T., Myers, R. M., Rogers, J., Stadler, P. F., Lowe, T. M., Wei, C.-L., Ruan, Y., Snyder, M., Birney, E., Struhl, K., Gerstein, M., Antonarakis, S. E., Gingeras, T. R., Brown, J. B., Flicek, P., Fu, Y., Keefe, D., Birney, E., Denoeud, F., Gerstein, M., Green, E. D., Kapranov, P., Karaöz, U., Myers, R. M., Noble, W. S., Reymond, A., Rozowsky, J., Struhl, K., Siepel, A., Stamatoyannopoulos, J. A., Taylor, C. M., Taylor, J., Thurman, R. E., Tullius, T. D., Washietl, S., Zheng, D., Liefer, L. A., Wetterstrand, K. A., Good, P. J., Feingold, E. A., Guyer, M. S., Collins, F. S., Margulies, E. H., Cooper, G. M., Asimenos, G., Thomas, D. J., Dewey, C. N., Siepel, A., Birney, E., Keefe, D., Hou, M., Taylor, J., Nikolaev, S., Montoya-Burgos, J. I., Löytynoja, A., Whelan, S., Pardi, F., Massingham, T., Brown, J. B., Huang, H., Zhang, N. R., Bickel, P., Holmes, I., Mullikin, J. C., Ureta-Vidal, A., Paten, B., Seringhaus, M., Church, D., Rosenbloom, K., Kent, W. J., Stone, E. A., Gerstein, M., Antonarakis, S. E., Batzoglou, S., Goldman, N., Hardison, R. C., Haussler, D., Miller, W., Pachter, L., Green, E. D., Sidow, A., Weng, Z., Trinklein, N. D., Fu, Y., Zhang, Z. D., Karaöz, U., Barrera, L., Stuart, R., Zheng, D., Ghosh, S., Flicek, P., King, D. C., Taylor, J., Ameur, A., Enroth, S., Bieda, M. C., Koch, C. M., Hirsch, H. A., Wei, C.-L., Cheng, J., Kim, J., Bhinge, A. A., Giresi, P. G., Jiang, N., Liu, J., Yao, F., Sung, W.-K., Chiu, K. P., Vega, V. B., Lee, C. W., Ng, P., Shahab, A., Sekinger, E. A., Yang, A., Moqtaderi, Z., Zhu, Z., Xu, X., Squazzo, S., Oberley, M. J., Inman, D., Singer, M. A., Richmond, T. A., Munn, K. J., Rada-Iglesias, A., Wallerman, O., Komorowski, J., Clelland, G. K., Wilcox, S., Dillon, S. C., Andrews, R. M., Fowler, J. C., Couttet, P., James, K. D., Lefebvre, G. C., Bruce, A. W., Dovey, O. M., Ellis, P. D., Dhami, P., Langford, C. F., Carter, N. P., Vetrie, D., Kapranov, P., The ENCODE Project Consortium, Analysis Coordination, Chromatin and Replication, Genes and Transcripts, Integrated Analysis and Manuscript Preparation, Management Group, Multi-species Sequence Analysis, NISC Comparative Sequencing Program*, Baylor College of Medicine Human Genome Sequencing Center*, Washington University Genome Sequencing Center*, Broad Institute*, Children's Hospital Oakland Research Institute*, and Transcriptional Regulatory Elements (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447(7146):799–816. Number: 7146 Publisher: Nature Publishing Group.

[Bruckner et al., 2010] Bruckner, S., Hüffner, F., Karp, R. M., Shamir, R., and Sharan, R. (2010). Topology-free querying of protein interaction networks. *Journal of computational biology*, 17(3):237–252. Publisher: Mary Ann Liebert, Inc. 140 Huguenot Street, 3rd Floor New Rochelle, NY 10801 USA.

[Bujakowska et al., 2009] Bujakowska, K., Maubaret, C., Chakarova, C. F., Tanimoto, N., Beck, S. C., Fahl, E., Humphries, M. M., Kenna, P. F., Makarov, E., Makarova, O., Paquet-Durand, F., Ekström, P. A., van Veen, T., Leveillard, T., Humphries, P., Seeliger, M. W., and Bhattacharya, S. S. (2009). Study of Gene-Targeted Mouse Models of Splicing Factor Gene Prpf31 Implicated in Human Autosomal Domi-

nant Retinitis Pigmentosa (RP). *Investigative Ophthalmology & Visual Science*, 50(12):5927–5933.

[Cannon et al., 2010] Cannon, C. H., Kua, C.-S., Zhang, D., and Harting, J. (2010). Assembly free comparative genomics of short-read sequence data discovers the needles in the haystack. *Molecular Ecology*, 19(s1):147–161. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-294X.2009.04484.x.

[Chen and Moore, 2015] Chen, W. and Moore, M. J. (2015). Spliceosomes. *Current Biology*, 25(5):R181–R183. Publisher: Elsevier.

[Cologne et al., 2019] Cologne, A., Benoit-Pilven, C., Besson, A., Putoux, A., Campan-Fournier, A., Bober, M. B., Die-Smulders, C. E. M. D., Paulussen, A. D. C., Pinson, L., Toutain, A., Roifman, C. M., Leutenegger, A.-L., Mazoyer, S., Edery, P., and Lacroix, V. (2019). New insights into minor splicing—a transcriptomic analysis of cells derived from TALS patients. *RNA*, 25(9):1130–1149. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

[de la Grange et al., 2005] de la Grange, P., Dutertre, M., Martin, N., and Auboeuf, D. (2005). FAST DB: a website resource for the study of the expression regulation of human gene products. *Nucleic Acids Research*, 33(13):4276–4284.

[Djebali et al., 2012a] Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., Xue, C., Marinov, G. K., Khatun, J., Williams, B. A., Zaleski, C., Rozowsky, J., Röder, M., Kokocinski, F., Abdelhamid, R. F., Alioto, T., Antoshechkin, I., Baer, M. T., Bar, N. S., Batut, P., Bell, K., Bell, I., Chakrabortty, S., Chen, X., Chrast, J., Curado, J., Derrien, T., Drenkow, J., Dumais, E., Dumais, J., Duttagupta, R., Falconnet, E., Fastuca, M., Fejes-Toth, K., Ferreira, P., Foissac, S., Fullwood, M. J., Gao, H., Gonzalez, D., Gordon, A., Gunawardena, H., Howald, C., Jha, S., Johnson, R., Kapranov, P., King, B., Kingswood, C., Luo, O. J., Park, E., Persaud, K., Preall, J. B., Ribeca, P., Risk, B., Robyr, D., Sammeth, M., Schaffer, L., See, L.-H., Shahab, A., Skancke, J., Suzuki, A. M., Takahashi, H., Tilgner, H., Trout, D., Walters, N., Wang, H., Wrobel, J., Yu, Y., Ruan, X., Hayashizaki, Y., Harrow, J., Gerstein, M., Hubbard, T., Reymond, A., Antonarakis, S. E., Hannon, G., Giddings, M. C., Ruan, Y., Wold, B., Carninci, P., Guigó, R., and Gingeras, T. R. (2012a). Landscape of transcription in human cells. *Nature*, 489(7414):101–108. Number: 7414 Publisher: Nature Publishing Group.

[Djebali et al., 2012b] Djebali, S., Lagarde, J., Kapranov, P., Lacroix, V., Borel, C., Mudge, J. M., Howald, C., Foissac, S., Ucla, C., Chrast, J., Ribeca, P., Martin, D., Murray, R. R., Yang, X., Ghamsari, L., Lin, C., Bell, I., Dumais, E., Drenkow, J., Tress, M. L., Gelpí, J. L., Orozco, M., Valencia, A., Berkum, N. L. v., Lajoie, B. R., Vidal, M., Stamatoyannopoulos, J., Batut, P., Dobin, A., Harrow, J., Hubbard, T., Dekker, J., Frankish, A., Salehi-Ashtiani, K., Reymond, A., Antonarakis, S. E., Guigó, R., and Gingeras, T. R. (2012b). Evidence for Transcript Networks Composed of Chimeric RNAs in Human Cells. *PLOS ONE*, 7(1):e28213. Publisher: Public Library of Science.

[Edery et al., 2011] Edery, P., Marcaillou, C., Sahbatou, M., Labalme, A., Chastang, J., Touraine, R., Tubacher, E., Senni, F., Bober, M. B., Nampoothiri, S., Jouk, P.-S., Steichen, E., Berland, S., Toutain, A., Wise, C. A., Sanlaville, D., Rousseau, F., Clerget-Darpoux, F., and Leutenegger, A.-L. (2011). Association of TALS Developmental Disorder with Defect in Minor Splicing Component U4atac snRNA.

Science, 332(6026):240–243. Publisher: American Association for the Advancement of Science.

[Fabozzi et al., 2018] Fabozzi, G., Oler, A. J., Liu, P., Chen, Y., Mindaye, S., Dolan, M. A., Kenney, H., Gucek, M., Zhu, J., Rabin, R. L., and Subbarao, K. (2018). Strand-Specific Dual RNA Sequencing of Bronchial Epithelial Cells Infected with Influenza A/H3N2 Viruses Reveals Splicing of Gene Segment 6 and Novel Host-Virus Interactions. *Journal of Virology*, 92(17):e00518–18. Publisher: American Society for Microbiology.

[Foissac and Sammeth, 2007] Foissac, S. and Sammeth, M. (2007). ASTALAVISTA: dynamic and flexible analysis of alternative splicing events in custom gene datasets. *Nucleic Acids Research*, 35(suppl_2):W297–W299.

[Fournier et al., 2014] Fournier, G., Chiang, C., Munier, S., Tomoiu, A., Demeret, C., Vidalain, P.-O., Jacob, Y., and Naffakh, N. (2014). Recruitment of RED-SMU1 Complex by Influenza A Virus RNA Polymerase to Control Viral mRNA Splicing. *PLOS Pathogens*, 10(6):e1004164. Publisher: Public Library of Science.

[Frenkel-Morgenstern et al., 2012] Frenkel-Morgenstern, M., Lacroix, V., Ezkurdia, I., Levin, Y., Gabashvili, A., Prilusky, J., Pozo, A. d., Tress, M., Johnson, R., Guigo, R., and Valencia, A. (2012). Chimeras taking shape: Potential functions of proteins encoded by chimeric RNA transcripts. *Genome Research*, 22(7):1231–1242. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

[Freyermuth et al., 2016] Freyermuth, F., Rau, F., Kokunai, Y., Linke, T., Sellier, C., Nakamori, M., Kino, Y., Arandel, L., Jollet, A., Thibault, C., Philipps, M., Vicaire, S., Jost, B., Udd, B., Day, J. W., Duboc, D., Wahbi, K., Matsumura, T., Fujimura, H., Mochizuki, H., Deryckere, F., Kimura, T., Nukina, N., Ishiura, S., Lacroix, V., Campan-Fournier, A., Navratil, V., Chautard, E., Auboeuf, D., Horie, M., Imoto, K., Lee, K.-Y., Swanson, M. S., de Munain, A. L., Inada, S., Itoh, H., Nakazawa, K., Ashihara, T., Wang, E., Zimmer, T., Furling, D., Takahashi, M. P., and Charlet-Berguerand, N. (2016). Splicing misregulation of SCN5A contributes to cardiac-conduction delay and heart arrhythmia in myotonic dystrophy. *Nature Communications*, 7(1):11067. Number: 1 Publisher: Nature Publishing Group.

[Grabherr et al., 2011] Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., and Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, 29(7):644–652. Number: 7 Publisher: Nature Publishing Group.

[Griebel et al., 2012] Griebel, T., Zacher, B., Ribeca, P., Raineri, E., Lacroix, V., Guigó, R., and Sammeth, M. (2012). Modelling and simulating generic RNA-Seq experiments with the flux simulator. *Nucleic Acids Research*, 40(20):10073–10083.

[Gusfield, 1997] Gusfield, D. (1997). *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*. Cambridge University Press, Cambridge.

[Heinz et al., 2018] Heinz, S., Texari, L., Hayes, M. G. B., Urbanowski, M., Chang, M. W., Givarkes, N., Rialdi, A., White, K. M., Albrecht, R. A., Pache, L., Marazzi, I., García-Sastre, A., Shaw, M. L., and Benner, C. (2018). Transcription Elongation Can Affect Genome 3D Structure. *Cell*, 174(6):1522–1536.e22.

[Iqbal et al., 2012] Iqbal, Z., Caccamo, M., Turner, I., Flicek, P., and McVean, G. (2012). De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nature Genetics*, 44(2):226–232. Number: 2 Publisher: Nature Publishing Group.

[Katz et al., 2010] Katz, Y., Wang, E. T., Airoldi, E. M., and Burge, C. B. (2010). Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature Methods*, 7(12):1009–1015. Number: 12 Publisher: Nature Publishing Group.

[Khatri et al., 2023] Khatri, D., Putoux, A., Cologne, A., Kaltenbach, S., Besson, A., Bertiaux, E., Guguin, J., Fendler, A., Dupont, M. A., Benoit-Pilven, C., Qebibo, L., Ahmed-Elie, S., Audebert-Bellanger, S., Blanc, P., Rambaud, T., Castelle, M., Cornen, G., Grotto, S., Guët, A., Guibaud, L., Michot, C., Odent, S., Ruaud, L., Sacaze, E., Hamel, V., Bordonné, R., Leutenegger, A.-L., Edery, P., Burglen, L., Attié-Bitach, T., Mazoyer, S., and Delous, M. (2023). Deficiency of the minor spliceosome component U4atac snRNA secondarily results in ciliary defects in human and zebrafish. *Proceedings of the National Academy of Sciences*, 120(9):e2102569120. Publisher: Proceedings of the National Academy of Sciences.

[Lacroix et al., 2008] Lacroix, V., Sammeth, M., Guigo, R., and Bergeron, A. (2008). Exact transcriptome reconstruction from short sequence reads. In *Algorithms in Bioinformatics: 8th International Workshop, WABI 2008, Karlsruhe, Germany, September 15-19, 2008. Proceedings 8*, pages 50–63. Springer.

[Lima et al., 2017] Lima, L., Sinaimeri, B., Sacomoto, G., Lopez-Maestre, H., Marchet, C., Miele, V., Sagot, M.-F., and Lacroix, V. (2017). Playing hide and seek with repeats in local and global de novo transcriptome assembly of short RNA-seq reads. *Algorithms for Molecular Biology*, 12(1):2.

[Llorian et al., 2010] Llorian, M., Schwartz, S., Clark, T. A., Hollander, D., Tan, L.-Y., Spellman, R., Gordon, A., Schweitzer, A. C., De La Grange, P., and Ast, G. (2010). Position-dependent alternative splicing activity revealed by global profiling of alternative splicing events regulated by PTB. *Nature structural & molecular biology*, 17(9):1114–1123. Publisher: Nature Publishing Group US New York.

[Marco-Sola et al., 2012] Marco-Sola, S., Sammeth, M., Guigó, R., and Ribeca, P. (2012). The GEM mapper: fast, accurate and versatile alignment by filtration. *Nature Methods*, 9(12):1185–1188. Number: 12 Publisher: Nature Publishing Group.

[Marsan and Sagot, 2000] Marsan, L. and Sagot, M.-F. (2000). Algorithms for Extracting Structured Motifs Using a Suffix Tree with an Application to Promoter and Regulatory Site Consensus Identification. *Journal of Computational Biology*, 7(3-4):345–362. Publisher: Mary Ann Liebert, Inc., publishers.

[Matera and Wang, 2014] Matera, A. G. and Wang, Z. (2014). A day in the life of the spliceosome. *Nature Reviews Molecular Cell Biology*, 15(2):108–121. Number: 2 Publisher: Nature Publishing Group.

[Merico et al., 2015] Merico, D., Roifman, M., Braunschweig, U., Yuen, R. K. C., Alexandrova, R., Bates, A., Reid, B., Nalpathamkalam, T., Wang, Z., Thiruvahindrapuram, B., Gray, P., Kakakios, A., Peake, J., Hogarth, S., Manson, D., Buncic, R., Pereira, S. L., Herbrick, J.-A., Blencowe, B. J., Roifman, C. M., and Scherer, S. W. (2015). Compound heterozygous mutations in the noncoding RNU4ATAC cause Roifman Syndrome by disrupting minor intron splicing. *Nature Communications*, 6(1):8718. Number: 1 Publisher: Nature Publishing Group.

[Milo et al., 2002] Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U. (2002). Network Motifs: Simple Building Blocks of Complex Networks.

Science, 298(5594):824–827. Publisher: American Association for the Advancement of Science.

[Peterlongo et al., 2010] Peterlongo, P., Schnel, N., Pisanti, N., Sagot, M.-F., and Lacroix, V. (2010). Identifying SNPs without a Reference Genome by Comparing Raw Reads. In Chavez, E. and Lonardi, S., editors, *String Processing and Information Retrieval*, Lecture Notes in Computer Science, pages 147–158, Berlin, Heidelberg. Springer.

[Pevzner et al., 2001] Pevzner, P. A., Tang, H., and Waterman, M. S. (2001). An Eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences*, 98(17):9748–9753. Publisher: Proceedings of the National Academy of Sciences.

[Ratan et al., 2010] Ratan, A., Zhang, Y., Hayes, V. M., Schuster, S. C., and Miller, W. (2010). Calling SNPs without a reference sequence. *BMC Bioinformatics*, 11(1):130.

[Robertson et al., 2010] Robertson, G., Schein, J., Chiu, R., Corbett, R., Field, M., Jackman, S. D., Mungall, K., Lee, S., Okada, H. M., Qian, J. Q., Griffith, M., Raymond, A., Thiessen, N., Cezard, T., Butterfield, Y. S., Newsome, R., Chan, S. K., She, R., Varhol, R., Kamoh, B., Prabhu, A.-L., Tam, A., Zhao, Y., Moore, R. A., Hirst, M., Marra, M. A., Jones, S. J. M., Hoodless, P. A., and Birol, I. (2010). De novo assembly and analysis of RNA-seq data. *Nature Methods*, 7(11):909–912. Number: 11 Publisher: Nature Publishing Group.

[Russell et al., 2006] Russell, A. G., Charette, J. M., Spencer, D. F., and Gray, M. W. (2006). An early evolutionary origin for the minor spliceosome. *Nature*, 443(7113):863–866. Number: 7113 Publisher: Nature Publishing Group.

[Sacomoto et al., 2013] Sacomoto, G., Lacroix, V., and Sagot, M.-F. (2013). A polynomial delay algorithm for the enumeration of bubbles with length constraints in directed graphs and its application to the detection of alternative splicing in RNA-seq data. In *Algorithms in Bioinformatics: 13th International Workshop, WABI 2013, Sophia Antipolis, France, September 2-4, 2013. Proceedings 13*, pages 99–111. Springer.

[Sacomoto et al., 2012] Sacomoto, G. A., Kielbassa, J., Chikhi, R., Uricaru, R., Antoniou, P., Sagot, M.-F., Peterlongo, P., and Lacroix, V. (2012). KIS SPLICE: de-novo calling alternative splicing events from RNA-seq data. *BMC Bioinformatics*, 13(6):S5.

[Sagot, 1998] Sagot, M. F. (1998). Spelling approximate repeated or common motifs using a suffix tree. In Lucchesi, C. L. and Moura, A. V., editors, *LATIN'98: Theoretical Informatics*, Lecture Notes in Computer Science, pages 374–390, Berlin, Heidelberg. Springer.

[Saudemont et al., 2017] Saudemont, B., Popa, A., Parmley, J. L., Rocher, V., Blugeon, C., Necsulea, A., Meyer, E., and Duret, L. (2017). The fitness cost of mis-splicing is the main determinant of alternative splicing patterns. *Genome Biology*, 18(1):208.

[Schulz et al., 2012] Schulz, M. H., Zerbino, D. R., Vingron, M., and Birney, E. (2012). Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics*, 28(8):1086–1092.

[Sessegolo, 2021] Sessegolo, C. (2021). *Développement de méthodes bio-informatiques pour l'étude de l'épissage chez les espèces non modèles : épissage complexe et apport des technologies de séquençage de 3eme génération.* These de doctorat, Lyon.

[Shen-Orr et al., 2002] Shen-Orr, S. S., Milo, R., Mangan, S., and Alon, U. (2002). Network motifs in the transcriptional regulation network of Escherichia coli. *Nature Genetics*, 31(1):64–68. Number: 1 Publisher: Nature Publishing Group.

[Sikora, 2011] Sikora, F. (2011). *Aspects algorithmiques de la comparaison d'éléments biologiques*. These de doctorat, Paris Est.

[Spellman et al., 2007] Spellman, R., Llorian, M., and Smith, C. W. (2007). Cross-regulation and functional redundancy between the splicing regulator PTB and its paralogs nPTB and ROD1. *Molecular cell*, 27(3):420–434. Publisher: Elsevier.

[Thompson et al., 2020] Thompson, M. G., Dittmar, M., Mallory, M. J., Bhat, P., Ferretti, M. B., Fontoura, B. M., Cherry, S., and Lynch, K. W. (2020). Viral-induced alternative splicing of host genes promotes influenza replication. *eLife*, 9:e55500. Publisher: eLife Sciences Publications, Ltd.

[Tilgner et al., 2012] Tilgner, H., Knowles, D. G., Johnson, R., Davis, C. A., Chakrabortty, S., Djebali, S., Curado, J., Snyder, M., Gingeras, T. R., and Guigó, R. (2012). Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. *Genome Research*, 22(9):1616–1625. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

[Trapnell et al., 2009] Trapnell, C., Pachter, L., and Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, 25(9):1105–1111.

[Trapnell et al., 2010] Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, 28(5):511–515. Number: 5 Publisher: Nature Publishing Group.

[Turunen et al., 2013] Turunen, J. J., Niemelä, E. H., Verma, B., and Frilander, M. J. (2013). The significant other: splicing by the minor spliceosome. *WIREs RNA*, 4(1):61–76. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/wrna.1141.

[Uricaru et al., 2015] Uricaru, R., Rizk, G., Lacroix, V., Quillery, E., Plantard, O., Chikhi, R., Lemaitre, C., and Peterlongo, P. (2015). Reference-free detection of isolated SNPs. *Nucleic Acids Research*, 43(2):e11.

[Zerbino and Birney, 2008] Zerbino, D. R. and Birney, E. (2008). Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, 18(5):821–829. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

[Zhao et al., 2018] Zhao, N., Sebastiano, V., Moshkina, N., Mena, N., Hultquist, J., Jimenez-Morales, D., Ma, Y., Rialdi, A., Albrecht, R., Fenouil, R., Sánchez-Aparicio, M. T., Ayllon, J., Ravisankar, S., Haddad, B., Ho, J. S. Y., Low, D., Jin, J., Yurchenko, V., Prinjha, R. K., Tarakhovsky, A., Squatrito, M., Pinto, D., Allette, K., Byun, M., Smith, M. L., Sebra, R., Guccione, E., Tumpey, T., Krogan, N., Greenbaum, B., van Bakel, H., García-Sastre, A., and Marazzi, I. (2018). Influenza virus infection causes global RNAPII termination defects. *Nature Structural & Molecular Biology*, 25(9):885–893. Number: 9 Publisher: Nature Publishing Group.

[Zheng et al., 2022] Zheng, H., Ma, C., and Kingsford, C. (2022). Deriving ranges of optimal estimated transcript expression due to nonidentifiability. *Journal of Computational Biology*, 29(2):121–139. Publisher: Mary Ann Liebert, Inc., publishers 140 Huguenot Street, 3rd Floor New ....

Chapter 10

Appendix

Initial version of the introduction, March 2020

The context in which I start writing this document is very particular. This is March 2020, 10. We are in the middle of the covid-19 pandemy. The number of cases and deaths is going up in many countries, including France. Many countries have declared lockdown in order to limit contacts and hence virus transmission. On the one hand, this situation is frightening because, even though we know the number of deaths should decrease, it is still very high (231 in France yesterday). Infected people are not anonymous, some of them are neighbours and relatives. We also cannot exclude that we, ourselves, are contaminated, and show no symptoms. On the other hand, this situation is encouraging because even though everyone is locked down at home, there are many signs of solidarity, one of which is the clapping hands at 8pm everyday to thank all the people who work at hospital. This crisis also reveals that we can very well live our lives by concentrating on fewer essential needs: sleeping, eating, spending time with the family. I bet that the carbon footprint we are having right now is finally down to the acceptable level we were all wishing to obtain to face the ecological crisis that is ahead of us.

I am fundamentally an optimistic person. I think we should take one problem at a time. And target the ones that are both important and achievable. As a scientist, I feel I have been trained so far to be a good bioinformatician. I have been targetting problems related to various fields of biology and computer science. The impact of my work on society has been limited so far, although in the last years, I was involved in clinical applications. I feel my energy would be best used in the coming years to target problems related to the ecological crisis. This includes changing the way I do research, travelling less, reducing my computing habits, analysing smaller but better designed datasets. This includes spreading the word around me that we can change our habits to reduce our carbon footprint and still be happy. I can do this through my teaching. This includes shifting the direct application of my research to problems more directly linked to the ecological crisis. This last item is perharps the least well defined. I feel I need to read a lot to identify problems that I find relevant and for which my expertise can bring something. Trying to enter an area where I wish to contribute but for which I know very little is a little frightening. But after all, as a bioinformatician, I already know the feeling of being an expert in no field at all. And this is probably what I find most appealing in science. Collaborating with experts in their field, and through discussions and exchange, get to know their field better and enrich my understanding of the world. My position of bioinformatician is in a way quite comfortable because many people need advice to analyse their data, hence I can get to know many fields. The main risk for me is clearly to lose myself in too many unrelated projects. I have been able to keep focus during the last 10 years because my efforts have been centered

on the development of a software called KisSplice, which for me is at the center of what I do. It covers a wide spectrum of computational biology as it enables me to work on computational problems (mostly graph algorithms, but also some statistics) and biological applications.