Statistical methods for analysing high throughput sequencing data

Hugues Richard

Habilitation à diriger des recherches
présentée et soutenue publiquement le 14 février 2022 par
Hugues RICHARD

Statistical methods for analysing high throughput sequencing data

From sequences to genome-wide summary statistics

Methods and Research development
Robert Koch Institute
Germany

Lab. Computional and Quantitative Biology
Sorbonne Université
France

COMPOSITION DU JURY

Mme	Bonizzoni Paola	Rapportrice	Professeure
Mr	Gautheret Daniel	Rapporteur	Professeur
Mr	Rivals Eric	Rapporteur	DR
Mr	Gesteira Costa Filho Ivan	Examinateur	Professeur
Mme	Thomas-Chollier Morgane	Examinatrice	MCF
Mr	Zucker Jean-Daniel	Examinateur	DR

This page left intentionally left blank

v1.0 - 2021-12-10 - original version
v.1.1 - 2022 - 03 - correction after defense

Contents

Acknowledgments

I would like to thank Paola Bonizzoni, Daniel Gautheret, and Eric Rivals for accepting to be reviewer for this manuscript. Thanks to Morgane Thomas Chollier, Ivan Gesteira Costa Filho and Jean-Daniel Zucker for participating in the jury. I hope you will enjoy reading this text as much as I had writing it.

I had the opportunity and the chance to interact and work with many gifted people. In several cases, it was possible because of a good work environment, an efficient recruiting strategy and a positive lab culture. I have heard many horror stories about other labs and I am therefore thankful to Bernard, Martin, Alessandra, Bernhard and Stephan for creating a positive work culture to promote exciting science. In particular, I would like to thank Alessandra for the opportunities I had at the LCQB. Freshly finishing my postdoc, she offered me multiple options, letting me act in complete freedom as a true act of mentoring. It was also overwhelming to experience and participate in the growth of the laboratory during its first decade. It felt like the perfect environment for me, and helped me transit between computational/statistical science and biology.

In any collaboration, there is a cognitive space created by the interaction between people, and in this space, this is were magic happens and each one can bring its own perspective and expertise. I would like to thank the numerous person I had the chance to work with: thanks for your ideas, for the conversations, and especially for being patient with me when time was pressing: Elodie, Pierre, Anish, Marcel, Marc, David; and Adel, Alessandra, Alexandre, Angela, Annalisa, Bogdan, Constance, Gilles, Iris, Laurent, Juliana, Martin, Martin, Mathilde, Matt, Rayan, Sabrina, Stefan, Stephan. Each new subject always starts with sparkles of curiosity and genuine interest, proceeds through the ocean of all possible questions until it finally reaches destination in the form of an article.

Finally, all of the last years would not have been possible without the support of Miriam. Even though she was jokingly commenting on "my research", she greatly helped me shape the document and carefully looked over the numerous errors. We started as a team of two and are now proceeding as a lab of four with the 2 funniest and sweetest interns one could think of. Finally thanks to my family and especially my mom for steadily and silently orienting me towards academia, I love it. I am sure she would be happy to learn that I am defending my habilitation and that she would profess a loud “c’est pas trop tôt !” of relief.

Introduction

The discovery of the structure of the DNA molecule (Franklin & Gosling, 1953; Watson & Crick, 1953; Wilkins, Stokes, & Wilson, 1953) and the formulation of the central dogma of molecular biology (F. Crick, 1970; F. H. Crick, 1958) marked the beginning of a new era in Biology. It laid the foundation for Biology to become a quantitative science, in which large amounts of molecular data would be integrated into a detailed description of living systems. At that point, we are in reach of understanding how information in a cell is stored, accessed, read and processed in an unprecedented and comprehensive manner (National Research Council, 2005). This transformation can be attributed to a steady technological improvement –through miniaturization and parallelisation, of the devices that measure and probe the molecular content of the cell (Metzker, 2010).

The central dogma (Cobb, 2017; F. Crick, 1970) states what characterises the flow of genetic information: it is stored in genomes, composed of DNA, then transcribed into RNA, which can have enzymatic/catalytic activity on its own, or be translated into protein. Proteins are the main biochemical protagonists of the cells, they can form protein interaction complexes, and bind to DNA or RNA as well (see Fig. 1 below). High Throughput DNA Sequencing (HTS) provides means of collecting molecular data that informs on each step of this flow: It informs on the regulatory mechanisms by providing readouts of the raw genetic material; It describes the functioning of the cell at a particular timepoint and scale by sampling the population of RNAs; It highlights the interactions of proteins to DNA or RNA by sequencing bound fragments; finally, it hints at chromosomal architecture through DNA-DNA interactions sites –see magnifying glasses in Figure 1. As a result, each of those assays yields a snapshot of the cell at the molecular level.

Cells are complex systems, depending on the condition the same genotype can give rise to different phenotypes or tissue types. Even though all those dimensions of the cells (species, tissues, temporal state, disease) cannot be assessed and compared at once, high throughput assays broaden our measuring capacity. For more than two decades, as the prices of DNA sequencing plummeted (Metzker, 2010), we have been taking more and more of those molecular pictures of biological systems. Nowadays, the genotype of a dozen individuals can be obtained in just a few days and for less than a thousand euros[1], whereas the first human genome took more than a decade and 10 billions dollars to be completed (Lander et al., 2001; Venter et al., 2001). For this reason, it is not surprising that pathogens detection and epidemiological surveillance can now be made at a molecular level (Oude Munnink et al., 2021) and on-the-fly (Gardy, Loman, & Rambaut, 2015; Quick et al., 2016).

The data in its rawest form consists of short readouts of DNA sequences in large quantities (millions to billions of reads). It cannot be used *per se* to describe the biological system. For best results, we have to integrate our current understanding of molecular biology while considering

[1] As part of a black Friday offer, I got my genome sequenced last year at 30X coverage for less that 200 euros

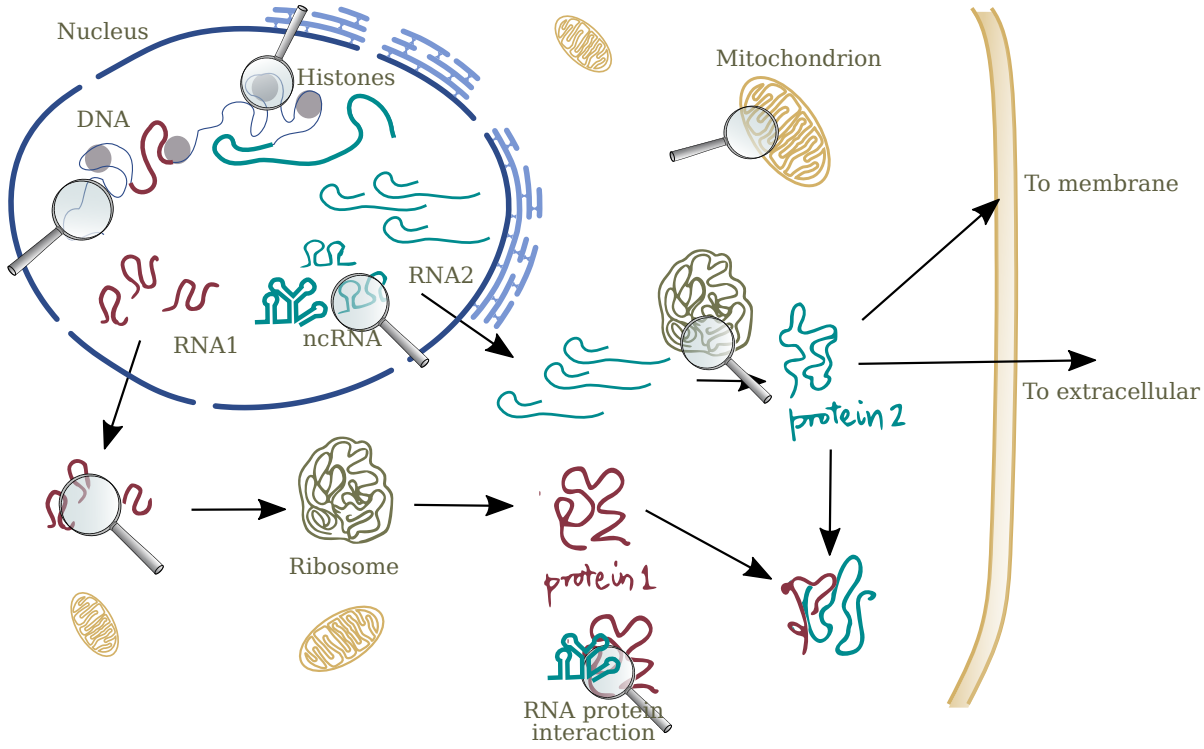


Figure 1: Schematic representation of the different layers of molecular information together with the flow of genomic information DNA→RNA→protein→interaction. The DNA molecule is simplified as a string with no consideration for its spatial organization. Nowadays high throughput assays can measure those systems at various levels, in this case represented by magnifying lenses. Thanks to Anish MS Shrestha for coming up with the idea of this figure.

the limitations of the experimental assay in order to *reduce* the raw sequences to a set of relevant compiled measures (or summary statistics). Compiling those summary statistics raises a number of methodological questions, which requires collaboration between different disciplines: molecular biology obviously, as well as statistical modelling, computer science and mathematics or physics. The scale of data, involving large quantities (Terabytes of sequences), spread among billions of positions, implies that algorithmic complexity and efficiency are an important part of the design. In the past years, I have been mainly working on proposing efficient statistical methods to reduce the raw measurements into a set of relevant summary statistics.

In the spirit of an habilitation thesis, the purpose of this document is twofold. First, I wanted to take the opportunity to give a general view on a set of common methodological problems emerging from the analysis of raw data (chapter 1) and highlight my different contributions in this context. Second, I wanted to construct a more personal opinion on the effective promises and limitations of omics data in the life science (chapter 2).

Within the first chapter, after a brief summary of the different sequencing platforms, I will detail the experimental steps for the preparation of sequencing libraries (the so called protocol). The specifics of library preparation are important, as it determines how we should analyse the sequences afterward. Protocols have multiple variations, depending mainly on the type of assay (e.g. sequencing of genome, transcriptome or protein-DNA interaction) and on the underlying questions (*e.g.* genotyping, functional analysis, mutant/WT comparison), but overall the principal steps

remain the same. I thus decided to present all types of assays side by side, in order to give the reader the possibility to summarise any experiment according to a set of core properties.

Due to those shared attributes, there is a common denominator to all computational workflows analysing the sequencing results: sequence comparison, feature counting/sequence census, data visualisation, summary statistics, determination of significance and functional analysis. The data produced at this stage can then potentially lead to the validation of a biological hypothesis and/or the generation of a new one. Side by side, I will describe the steps of any computational analysis of HTS data (section 1.3).

Like most data intensive applications, the analysis of HTS data requires to a good balance between three aspects: modelling, genericity, and algorithmic complexity. In other words, the model should account for the specificities of the sequencing protocol, while keeping a formulation general enough to be transferred among variations, and enables an algorithmic formulation which can scale up with data size. I will present a selection of general methodological problems associated to HTS data analysis, especially for sequence comparison and summary statistics (section 1.4). They will provide a rationale for the probabilistic models detailed afterward (section 1.5). My contributions will be used as illustrations of models used or developed along this text.

I will then shift from methodological questions to an exemplary case of biological data analysis by presenting a work in collaboration with the Diatom Genomics team at LCQB (section 1.6). This work was more directed toward quantitative biology and concerned the analysis of the sequenced small RNAs fragments from the diatom *Phaeodactylum tricornutum* in various conditions. Small RNAs were completely uncharacterized in diatoms and we ultimately wanted to determine their characteristics as well as in which regulatory processes they participate. We managed to categorise virtually all the sequences present in the sample and validate their presence experimentally. But there was no significant effect between conditions, and our conclusions remained mainly observational.

This personal experience was an eye opener about the limitations of HTS based analysis. Genome-wide assays surely accelerate the pace of new discoveries, but they have shifted some of the research focus in biology from hypothesis-driven to resource based summaries and diagrams (Stern, 2019). Descriptive science (generation of new hypothesis by measurement of a system) and proper validation of a hypothesis through a carefully though experimental setup are common antagonists when thinking of the proper way of doing science and as expected, this raised debate in the decade after the publication of the draft human genome (Golub, 2010; Weinberg, 2010). In the second chapter, I will start by discussing those questions and continue afterward by commenting on the epistemology of big data and data intensive science in biology.

References

Cobb, M. (2017). 60 years ago, Francis Crick changed the logic of biology. *PLOS Biology*, *15*(9), 1–8. doi:10.1371/journal.pbio.2003243

Crick, F. (1970). Central dogma of molecular biology. *Nature*, *227*(5258), 561–563.

Crick, F. H. (1958). On protein synthesis. *Symp Soc Exp Biol*, *12*, 138–163.

Franklin, R. E., & Gosling, R. G. (1953, April). Molecular configuration in sodium thymonucleate. *Nature*, *171*(4356), 740–741.

Gardy, J., Loman, N. J., & Rambaut, A. (2015). Real-time digital pathogen surveillance — the time is now. *Genome Biology*, *16*(1), 155. doi:10.1186/s13059-015-0726-x

Golub, T. (2010). Counterpoint: Data first. *Nature*, *464*, 679.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., ... International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*, 860–921.

Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nat Rev Genet*, *11*(1), 31–46. doi:10.1038/nrg2626

National Research Council. (2005). *Mathematics and 21st Century Biology*. doi:10.17226/11315

Oude Munnink, B. B., Worp, N., Nieuwenhuijse, D. F., Sikkema, R. S., Haagmans, B., Fouchier, R. A. M., & Koopmans, M. (2021). The next phase of SARS-CoV-2 surveillance: Real-time molecular epidemiology. *Nature Medicine*, *27*(9), 1518–1524. doi:10.1038/s41591-021-01472-w

Quick, J., Loman, N. J., Duraffour, S., Simpson, J. T., Severi, E., Cowley, L., ... Carroll, M. W. (2016). Real-time, portable genome sequencing for Ebola surveillance. *Nature*, *530*(7589), 228–232. doi:10.1038/nature16996

Stern, C. D. (2019). The 'Omics revolution: How an obsession with compiling lists is threatening the ancient art of experimental design. *Bioessays*, *41*(12), e1900168. doi:10.1002/bies.201900168

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., ... Zhu, X. (2001). The sequence of the human genome. *Science*, *291*(5507), 1304–1351. doi:10.1126/science.1058040. eprint: http://science.sciencemag.org/content/291/5507/1304.full.pdf

Watson, J. D., & Crick, F. H. C. (1953). Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature*, *171*, 737–738.

Weinberg, R. (2010). Point: Hypotheses first. *Nature*, *464*, 678 EP -.

Wilkins, M. H. F., Stokes, A. R., & Wilson, H. R. (1953, April). Molecular structure of deoxypentose nucleic acids. *Nature*, *171*(4356), 738–740.

Chapter 1

High Throughput Sequencing Assays: an Overview for the Analyst

1.1 Sequencing Platforms

Ever since the determination of the first full genomic sequence of an organism almost three decades ago with *H. hinfluenza* (Fleischmann et al., 1995), and the landmark publication of the human genome at a "finished grade" ten years later (International Human Genome Sequencing Consortium, 2004), there has been a significant shift. Automated sequencing by chain termination (e.g. Sanger sequencing (Sanger, Nicklen, & Coulson, 1977)), is replaced by a set of technologies whose throughput is increasing exponentially fast, putting sequencing assays/experiments within reach of a very large audience (small research labs to citizen scientists). In theory, anyone is now in reach of generating sequences that could disrupt medicine, health, and ancestry (Chiu & Miller, 2019; Esplin, Oei, & Snyder, 2014), but also to impact previously unexpected areas of research (Miller, 2010). These technologies, commonly termed High Throughput Sequencing (HTS) technologies, all rely on miniaturisation and parallelisation of clonal template generation, which allow to sequence millions to billions of DNA fragments at the same time. The first generation of HTS consists in a variety of techniques that mainly fall into two broad categories: sequencing by ligation (SBL) and sequencing by synthesis (SBS). Briefly, SBS is a DNA-polymerase dependent method for sequencing (used by the Solid and Roche sequencers and is now phased out), and SBL involves a probe sequence bound to a fluorophore (mainly Illumina). I refer to the following reviews for more details on those technologies (Glenn, 2011; Metzker, 2010).

Each of the sequencing platform has its advantages and drawbacks, and offer various tradeoffs depending on throughput, cost and read lengths. These need to be taken into account depending on the primary biological question that is to be addressed. A large majority of the sequences produced nowadays are sequenced on one of the Illumina sequencing platforms, in part due to its reduced cost per bp (Kodama, Shumway, & Leinonen, 2012). Illumina proposes a large range of sequencing platforms, and has on average the most cost effective technology [1]. This puts the expected price of a human genome sequenced at 30X coverage (the norm) between 900$ and 8000$.

Last generation of HTS technologies are based on single molecules sequencing and allow to sequence longer reads (up to a few hundred thousands of bp). The two most common are SMRT

[1] as of December 2018 the cost was between 10$ and 60$ per Gbp, depending on the machine. More info are compiled on this online spreadsheet

by Pacific Bioscience (Eid et al., 2009), or Minion by Oxford Nanopore Technology ("The long view on sequencing", 2018; van Dijk, Jaszczyszyn, Naquin, & Thermes, 2018). Other companies, such as 10x genomics, use microfluidics to barcode sequences from the same fragment, leading to synthetic long reads after attribution. Longer reads are crucial for some application such as whole genome sequencing, because genomes are inherently straddled with repeats (Berlin et al., 2015; Jain et al., 2018). It also makes it possible to sequence full transcript mRNA isoforms directly (Workman et al., 2018). Nevertheless, the throughput of the long read technologies is still much lower than Illumina sequencers.

Figure 1.1 summarises the relation between read length and throughput for most of the sequencing platforms (Nederbragt, 2016). A large proportion of the platforms operate in the region of [1-10 Gbp throughput] and [100-300bp read length] range.

Even though each platform developed its own proprietary technology to carry out the nucleotidic readouts, the steps preceding sequencing –*e.g.* the preparation of the raw sequence templates, remains nearly the same. Apart from the specifics of the sequences targeted for enrichment, most of the protocols share a common backbone for library preparation: starting from the biological material (a population of cells in a particular condition or from a given tissue type), cells are lysed, and then DNA or RNA molecules are purified, extracted, sheared into fragments, and possibly enriched. Then adapters are ligated and the fragments are sequenced.

In the next two sections, we will compile the steps of a typical workflow of library preparation, mentioning the different types of assays in parallel. A sequencing experiment consists of two parts: the first step is the preparation of a library of sequences which will be loaded onto the sequencer, the second step is the computational analysis of the resulting sequences.

1.2 Canonical Steps of Library Preparation

Three of the most common assays are genome sequencing, transcriptome sequencing and sequencing of DNA/RNA bound to proteins (Plocik & Graveley, 2013). We will limit our presentation in the following to "bulk" sequence analysis (on a population of cells). A large number of single cell assays have been developed in the last decade (Stegle, Teichmann, & Marioni, 2015), but this is beyond the scope of this presentation.

1.2.1 Library Preparation

Each type of sequencing assay is designed to enrich for a particular category of sequences (see Figure 1):

A. **Genomes**: The term Genomes refers either to the whole genome sequence (WGS) of a clonal population of cells, a targeted sequencing of specific regions (*e.g.* the set of exons: the exome), the genome of a single cell, or an heterogeneous population of cells (e.g. tumoral samples, or bacterial communities – metagenomes).

B. **Transcriptomes**: the population of messenger RNAs, mature or all. They can be selected in a particular size range (short RNAs (Ghildiyal & Zamore, 2009)), and/or according to the post-transcriptional modification of mature RNAs (poly-A tail (Sultan et al., 2008), capped RNA (Carninci et al., 2006), all messenger RNAs), or based on their subcellular localisation (cytosol, nucleus, ribosome associated).

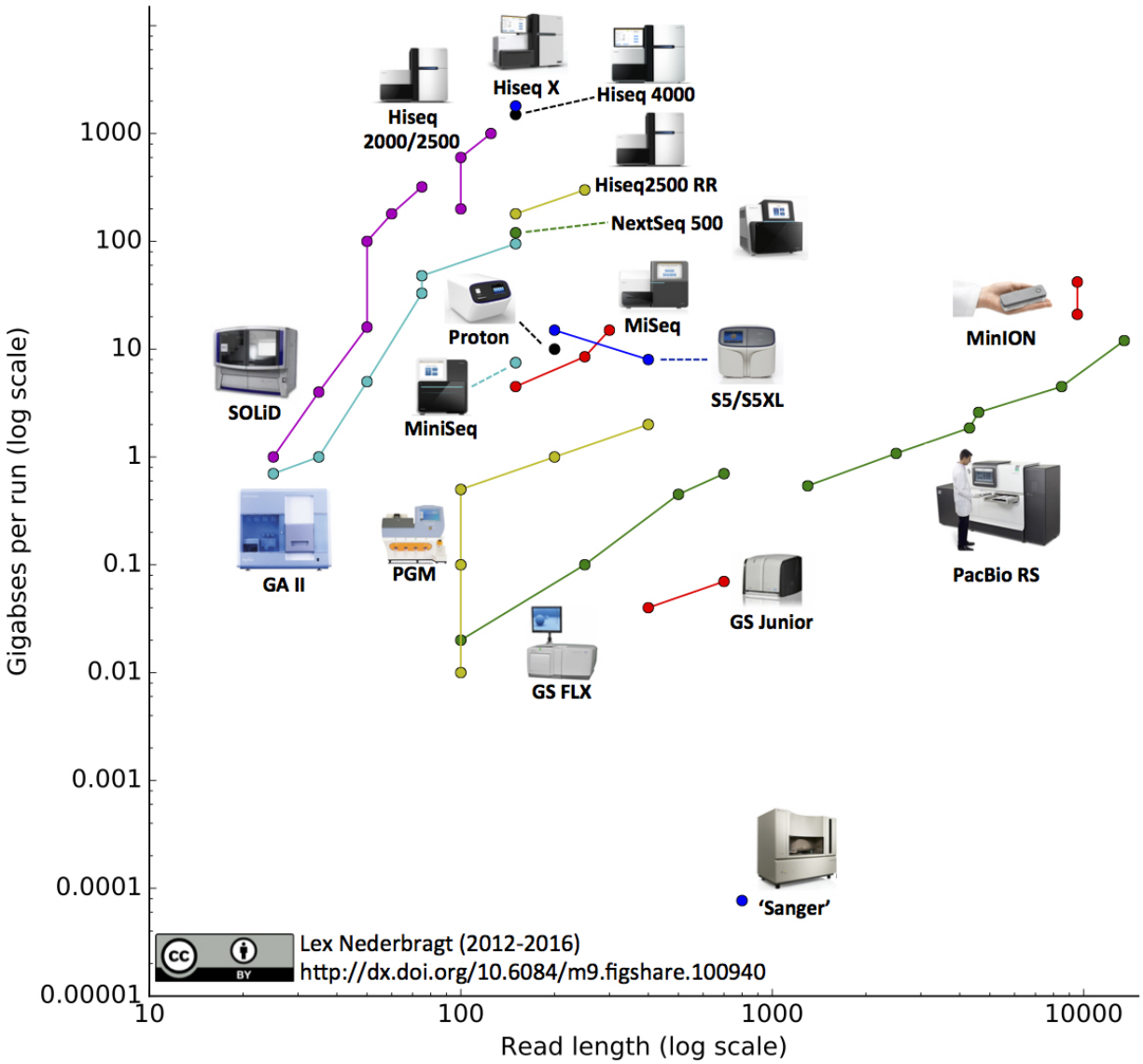


Figure 1.1: from (Nederbragt, 2016), comparison of throughput and read length for most platforms on a log-log scale. The value reported for throughput is meant for one run on the platform and does not take into account the time needed to complete it (which can range from a few hours to several days). Some technologies, like GS FLX and GS Junior from Roche or Solid have been discontinued and are therefore not mentioned in the main text.

C. Sequences involved in to **protein-DNA or protein-RNA interaction**. That can be fragments of DNA binding to a given protein (for instance transcription factors or enhancers) or to histone having a given modification (epigenome), or an RNA Binding Protein (RBP).

In the following, when a given assay uses a specific protocol step, I will reference it with the corresponding letter: **A**, **B** or **C** (see figure 1.2)

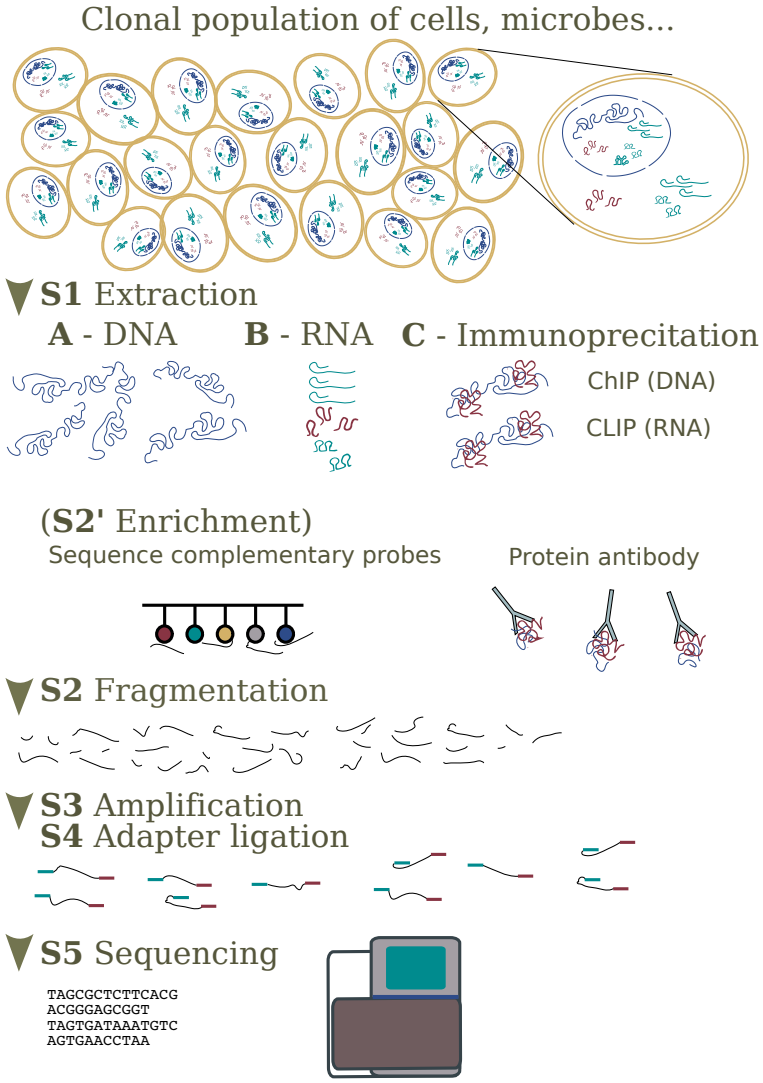


Figure 1.2: A summary of the canonical steps of library preparation. Three different sequencing assays are presented side by side in order to highlight their common step (**A**- genome isolates (Metzker, 2010), **B**-transcriptome sequencing (Ozsolak & Milos, 2010), **C**-protein bound fragments (Furey, 2012; Lee & Ule, 2018)). Details about each step of the protocol can be found in the text.

Here are the main protocol steps before getting sequence reads from the sequencer:

S1 - Sequence preparation and extraction/purification

A. After breaking cell membranes, genomic sequences are precipitated (using for instance ethanol) and purified from other cellular content.

B. To obtain RNA sequences RNA purification is performed: RNA degradation is stopped and the molecules are separated followed by the removal of possible DNA-contaminant with a DNase.

Ribosomal RNAs can make up more than 80% of the RNA content of the cell. In order to avoid using an unwanted amount of the sequencing capacity on it, they thus either have to be removed using ribosome specific hybridisation probes[2], or the mature mRNAs must be enriched afterwards (using poly-dT pulling, or capping). In eukaryotic cells, RNA purification can target only the cytoplasmic fraction (mature mRNA) or other fractions (nuclear, whole cell, mitochondria bound).

The extraction of RNAs is directly followed by a step of reverse transcription into cDNA using either random primers or poly-dT primers, having each their positional preference (poly-dT tend to enrich the 3'-end, while random primer have some preferred motifs see section 1.2.3). When the assay targets the population of small-RNAs, a size selection step is also performed.

C. Sequences bound to a protein are crosslinked, *i.e.* contact and proximity points are transformed into strong covalent bonds by chemical means for DNA (formaldehyde) (Furey, 2012) or by radiation for RNA (UV-light, to specifically crosslink direct protein-RNA interaction) (Lee & Ule, 2018). In the context of DNA-binding, transcription factors or enhancers are commonly assayed. Let's note also that specific antibodies are available for multiple histone marks, making it possible to assess also the state of the epigenome, by means of the histone modification imprinted along the chromosomes (Furey, 2012).
In the case of RNA bound protein, a step of reverse transcription is performed (as in B). In this case, the reverse transcriptase is expected to be interrupted by the peptides bound at the cross link site. The 5' end of the fragments are thus enriched in the exact cross linking site.

S2 - Fragmentation Sequences are then sheared into smaller fragments, usually by sonication or chemical means with a restriction enzyme. For genome and transcriptome sequencing, this step is performed to ensure that the fragments size will be compatible with the capacity of the sequencing platform (see Figure1.1). For immunoprecipitation based assays, the fragmentation step will enrich fragment on the protein-bound regions. In some cases (**B** and **C**), some variations are common:

B. Fragmentation can precede the first reverse transcription (mentioned in S1) and be performed directly on the population of RNAs. Obviously, no fragmentation is done for small non coding RNAs.

C. For protein-DNA interaction, sequence fragmentation with ChIP-Seq can be done by sonication, resulting in approx. 200bp fragments around the binding site, or with restriction enzymes such as MNase (Barski et al., 2007) or exonuclease (Rhee & Pugh, 2011), with fragments more precisely delineating the binding region. For protein-RNA interaction, an RNase is used in order to obtain an optimal RNA fragment size distribution (typically between 30 and 200bp). The most used CLIP techniques (HITS-CLIP, eCLIP, iCLIP, CLIP stands for Cross-Linking ImmunoPrecipitation (Lee & Ule, 2018)) perform this fragmentation before immunoprecipitation (step 2').

The fragments are then selected in a given size range (sequences are deposed on a gel and cut out after electrophoresis). In the case of CLIP techniques, the specificity of the antibody (step 2') can be assessed by running an SDS PAGE[3] gel and verifying that the band has a size compatible with

[2]such as the Ribominus kit from Thermo Fisher for instance

[3]SDS page is an electrophoresis method that allows protein separation by mass (from wikipedia https://en.wikipedia.org/wiki/SDS-PAGE).

the target protein, when bound to DNA. Additional bands indicate contamination with background proteins.

For genome sequences, longer fragments have more potential for resolving repeats induced ambiguities, thus leading to more accurate reconstructions of the whole genomic sequence. Mate-pair libraries, by ligating fragment ends and then shearing sequences around this ligation points, provide an alternative to obtain sequences coming from the end of longer fragments (a few Kbp long).

S2' - Enrichment of Sequences This step can occur at different points within the protocol, but is commonly performed after fragmentation. The goal is to enrich the sample in sequences with some properties such as genomic localisation, or by means of their sequence characteristics (**A, B**). For protein binding assays (**C**), an antibody for the protein is used to filter by immunoprecipitation (IP) of the fragments. The specificity of the antibody used is crucial, as it determines the amount of resulting *bona fide* sequencing material and the quality of the library.

A,**B**. A set of sequence probes pulls down fragments coming from predefined regions. This allows for instance to reduce a WGS library to the set of sequences covering the exons (exome sequencing). Note in some protocols the pull down is done after the whole library preparation (e.g. just before **S5**). An alternative technique for enrichment is to perform PCRs in solution.

C. The target protein or the targeted histone modification is immunoprecipitated with an antibody specific of the protein. The specificity of the antibody is of prime importance, as DNA contamination and unspecific enrichment or indirect binding can create cross hybridisation and unwanted readout (Meyer & Liu, 2014). Various experimental techniques perform in parallel a control experiment –called *Input*- to give an indication about non specific binding, in order to mitigate such biases. This experiment can consist in performing directly a fragmentation on the DNA, or using the non specific antibody Immunoglobulin G (IgG) for ChIP, or to generate a *size-matched* input cut of a SDS page gel without IP (eCLIP protocol) (Lee & Ule, 2018). Owing to their broader distribution, those Input controls need to be sequenced at greater depth. We detail a little more this bias below (see section 1.2.3).

S3,S4 - Amplification of the material (optional) and ligation of sequence adapters Some assays need a large number of cells (as much as ≈ 10 million for ChIP-Seq) to start with. Depending on the amount of starting material, some variations on PCR can be recommended to optimize the amplification (Furey, 2012). Optimization of the fragmentation step are also possible. After amplification, sequence adapters are added to the end of the fragments in order to fix them on the flow cell.

S5 - Sequencing Reads, usually shorter than the fragments, are sequenced (usually 50 bp to 300 bp long reads for Illumina). Note that reads can be sequenced either on one side of the fragment (single end) or on both sides (paired-end) after doing template switching. The resulting data is a set of sequences obtained in form of a text file (two files for paired-end) containing the sequences, and for each base pair, the quality of the readout (the probability to contain an error).

1.2.2 Other Assays

Many variations of the protocols mentioned above have been developped[4]. For instance it allows to concentrate the sequencing power on other specific categories of molecules. Let us mention a few other popular protocols for the sake of completeness.

Chromatin free regions: DNase-footprinting (Hesselberth et al., 2009) or ATAC-Seq (Buenrostro, Giresi, Zaba, Chang, & Greenleaf, 2013) experiments use respectively a DNase or a transposase which are preferentially active in unbound DNA regions. The resulting sequence fragments are thus enriched in open chromatin regions, regions that have gene-regulatory functions, such as promoters, enhancers, silencers, insulators, and locus control regions. This allows to simultaneously identify all types of regulatory regions in a genome-wide manner (Gusmao, Allhoff, Zenke, & Costa, 2016).

Chromosomal conformation ("C"-technologies): Chromosome organisation in eukaryotic cells can be studied by sequencing pairs of genomic loci in physical interaction. The technique combines protein crosslinking (like ChIP) with proximity ligation of DNA. Briefly cells nuclei are isolated, and then formaldehyde is used to crosslink the chromatin proteins with DNA (Schmitt, Hu, & Ren, 2016). Cross-linked DNA is then fragmented using restriction enzymes. The ends of the fragments are religated in conditions that favour juxtaposed DNA fragments. Multiple variations of this protocol exist (3C, Hi-C, etc.).

DNA/RNA modifications: The most common DNA chemical modification is 5-methylcytosine (5mC). 5mC methylation for instance can be assessed using bisulfite treatment, a biochemical process which deaminates every unmethylated cytosine residues turning them into uracil while keeping the 5mC residues unaffected. A subsequent PCR finishes the transformation by turning the uracils (*e.g.* the unmethylated bases) into thymines.[5] Stretches of positions with A to T mutation thus correspond to non methylated regions and the other way round.

1.2.3 Biases and Artifacts

Preparing a sequencing library involves multiple experimental steps, each having its own effect on the population of fragments. The protocol can induce biases in favouring or avoiding some sequence fragments that are not related to the aim of the assay. Let us review the biases that have been identified so far at each stage. As technology improves, some of those effects can be mitigated, but most of the points listed below remain important sources of noise that should be accounted for when analysing the sequence data.

Fragmentation technique: While fragmentation by sonication is generally reported to be uniform, restriction enzymes notoriously preferentially cut the DNA at particular sites (Chung et al., 2010; Meyer & Liu, 2014), therefore affecting fragment ends placement for ChIP or CLIP protocols. Furthermore, even if the fragmentation is theoretically uniform on DNA *in vitro*, the chromatin

[4] as of May 2018, Illumina website lists more than 40 variations of the protocols for genome sequencing, and above 50 for RNA content characterisation. The DNA or RNA protein interactions protocols exists in at least 30 variations as well. See https://www.illumina.com/techniques/sequencing/ngs-library-prep/library-prep-methods.html

[5] Use of those types of assays could be decreasing, as single molecule sequencers can theoretically measure the various modifications of a base directly.

structure of the genome will have an influence on the location of sequence cuts *in vivo*. Certain cuts have more chance to occur in the open chromatin regions (Meyer & Liu, 2014). Likewise, the RNA secondary structure, consisting of loops and stem regions, can have an influence on fragmentation or priming efficiency and thus influence the relative abundances of different reads along the sequence (J. Li, Jiang, & Wong, 2010; Zheng, Chung, & Zhao, 2011).

Fragment size selection: Selecting a fragment size obviously impacts the physical coverage of the sequences in the sample, and limits the size of the events which can be observed within one fragment. Additionally, when the sample consists in a collection of relatively short sequences, as it is the case for mRNAs, size selection will create a bias next to the transcript ends (Griebel et al., 2012).

PCR biases: At each PCR cycle, the differences in composition and length within the population of DNA fragments result in uneven amplification efficiency. A common way to summarise this effect is to consider GC-bias, because GC composition plays a direct role in annealing temperature (Benjamini & Speed, 2012). Apart from the GC content of the fragment, the genomic sequence surrounding the 5' end of the reads may affect the uniformity of read distribution (K. D. Hansen, Brenner, & Dudoit, 2010; J. Li et al., 2010). This bias may be due to PCR as well. In the case of RNA sequencing, it could also be mediated by the formation of structures limiting primer binding during the reverse transcription step (J. Li et al., 2010). Note that the PCR amplification effects can be mitigated by adding short (6-10nt) random sequence barcodes (so called Unique Molecular Identifier (UMI)) prior to the amplification (S3) step. Each original fragments can be deduplicated using the fact that the combination of the UMI and read sequence is unique.

Probe based enrichment: Enrichment based methods can be used to select specific populations of RNAs, such as polyA RNAs with poly-dT pulling. This can have an impact on the fragments that are retrieved, as there can be breaks near to transcripts 3'-end which will result in an increase of fragment coming from there. Enrichment can also target specific regions, like for exome sequencing. In this case the enrichment probes are targeting a set of sequences designed according to a reference. Probes annealing depends on its GC-content, with a natural impact on the relative sequences representativity.

Antibody based enrichment: This bias is related to the efficiency and the specificity of the antibody used to perform immunoprecipitation. As mentioned previously, non specific enrichment, as well as cross enrichment of non direct binding can occur. These effects can be reproducible across conditions and were reported as Hyper-ChIPpable regions for ChIP (Teytelman, Thurtle, Rine, & van Oudenaarden, 2013) or background regions (Friedersdorf & Keene, 2014; Reyes-Herrera, Speck-Hernandez, Sierra, & Herrera, 2015) and cross link motifs (CL-motifs) (Haberman et al., 2017) in the case of CLIP. Theses biases can be mitigated by designing different controls, or input experiments as I mentioned previously.

DNA contamination: When purifying the population of RNA (RNA-Seq and CLIP-Seq), great care has to be applied, as some DNA from the sample can still be present. We and others quantified this noise for the first RNA-Seq experiments, which were doing selection using poly-dT pulling

(Sultan et al., 2008; E. T. Wang et al., 2008a). The DNA contamination noise was observed to be relatively low, at expression levels 100 folds lower on average than expressed regions. Different variants of the protocol or improper use of DNAse can however have dramatic impact on the background expression levels observed.

Sequencing errors: Illumina platform is mainly reported to have substitution errors (Dohm, Lottaz, Borodina, & Himmelbauer, 2008). There can also be biases from the base caller or induced by amplification, which will favour certain types of errors next to specific sequence content (Allhoff et al., 2013; Saad et al., 2018).

In the last decade, the expectation from those assays would follow a typical "Gartner hype cycle" phenomenon[6]: after the high expectation brought by the new technology settles down, thorough studies report alarming results and low reproducibility. The systematic effects that were discovered guide in turn the development of improved experimental protocols. It also leads consortia responsible of large datasets production to publish guidelines and sets of best practices, such as the one proposed by ENCODE for ChIP-seq (Landt et al., 2012) or by GEUVADIS for RNA-Seq ('t Hoen et al., 2013). Nowadays, we can consider that the assays presented before have attained a good level of maturity. In any case, any new protocol appearing could in theory be analysed for systematic biases. Due to the randomness inherent to the preparation of fragments and the sample size (millions of reads), a thorough statistical analysis can identify systematic library preparation effects. I will detail in section 1.5 how simple model assumptions enable to account for those biases. We should however note that it is in practice a daunting task to design those models.

1.3 Analysis of the Sequence Reads

The collection of reads produced by the sequencer, basically a text file, cannot be interpreted by simple examination. The raw sequences need to be compared, combined, aggregated and counted to build numerical and visual representations. Constructing those representations (I call them summary statistics[7]), relies on internal rules derived from our knowledge in molecular biology. Their final form is determined by the questions at the origin of the experiment. A summary statistics can for instance be the genotype of an individual after having done DNA sequencing, a table summarising the abundance levels of all transcripts in an RNA-Seq assay, or even the gene model annotation. To obtain those summary statistics, every sequencing data analysis pipeline has to follows a few common steps: quality control, sequence comparison, feature counting/sequence census, data visualisation, generation of summary statistics. After the summary statistics are produced an additional step of functional analysis and biological interpretation is usually done. Each step in this pipeline motivates the methodological questions and the corresponding computational methods that I will present in the following sections.

Let's have first a bird's eye view of the typical steps of a traditional data analysis (see Figure 1.3A):

1. **Quality control** (read/library properties)
   First of all, the set of sequences is analysed to confirm that the quality of the library is suf-

[6] https://en.wikipedia.org/wiki/Hype_cycle
[7] they are not necessarily limited to statistics, and can also consist in processed sequence data, such as assembly.

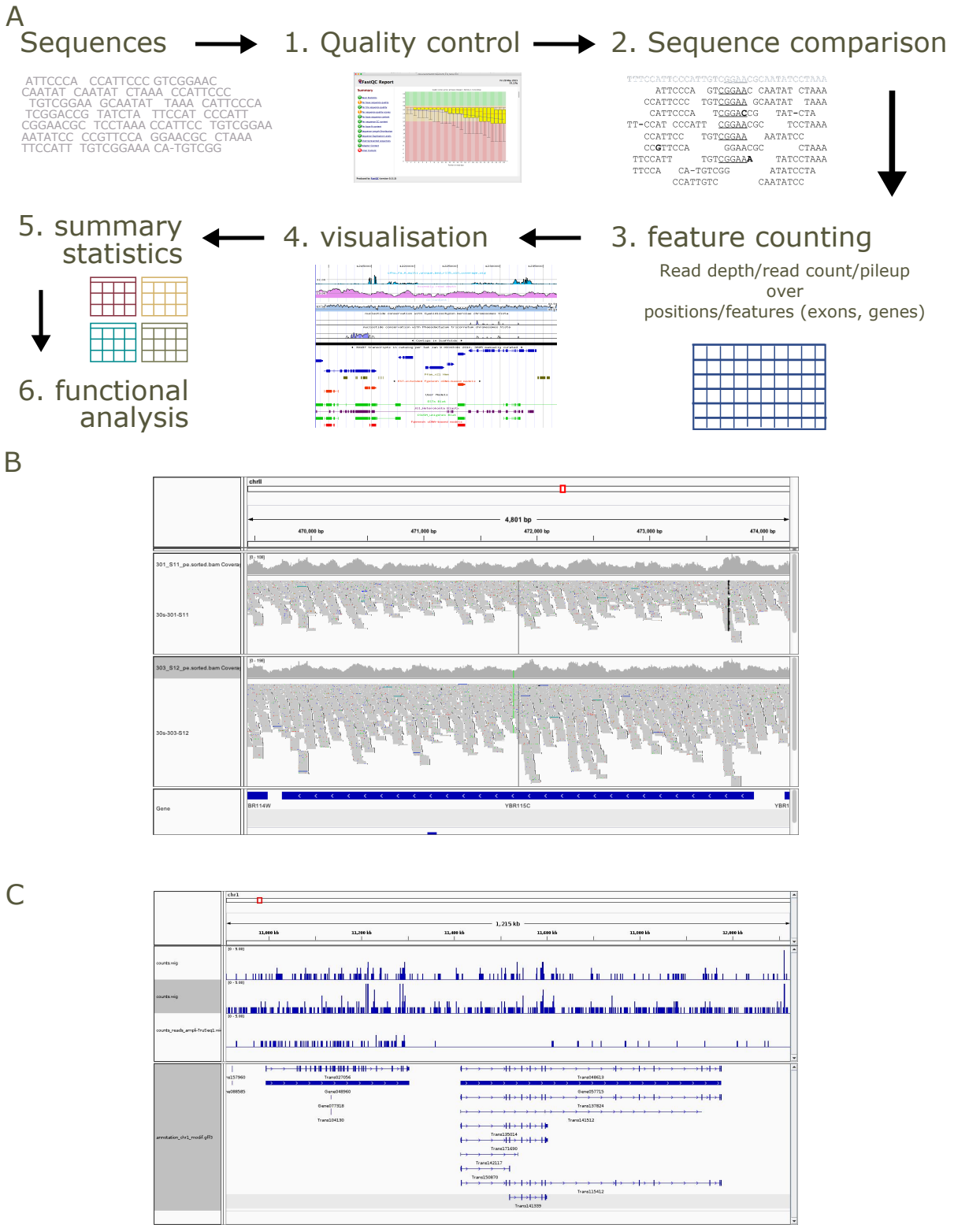


Figure 1.3: A - Typical steps of a sequence data analysis pipeline. B - Example of a visualisation obtain with the IGV genome browser/ The aligned reads are represented from two genome sequencing experiment, together with the computed read depth. Genomic coordinates are reported on the top. C - Genome browser view of of three RNA-seq experiments. The read depth from the experiment is reported on the top, with the genes and and transcript annotation below.

ficient and that the sequencing run went as expected. Quality plot will include summary information from the PHRED quality scores of the reads, as well as enrichment statistics on the reads compositional properties. This can highlight possible biases related to adapter sequences enrichment within the reads, or unexpected variations in sequence composition within the population of reads (for instance PCR or base caller artifacts) (Andrews, n.d.; Ewels, Magnusson, Lundin, & Kaller, 2016).

2. **Sequence comparison** - Alignments, Assembly, Error Correction.
   Then, sequences have to be *compared*, *e.g.* either by pairwise comparisons (sequence assembly), or by aligning them to a set of reference sequences (sequence alignment). This step is crucial, it gives a first structure on the set of reads by arranging them with respect to each other. The data structure describing the results of the comparisons can range from a tiling over the reference genome (sequence alignment), to sequence graphs summarising pairwise overlaps between the reads (Overlap Layout Consensus, de Bruijn graph). The technique to derive sequence alignments is based on finding stretches of sequences that are more similar to each other than it would be expected by chance, given the genomic content. An implicit probabilistic modelling of sequence content, accounting for possible biases (sequencing errors, genetics of the species...) is implicitly specified to find the most probable alignment. Since the amount of sequences is very large, we need to align them as fast as possible. The computational cost for exact alignment -quadratic in sequence length- is prohibitive given the scale of the data. Thus efficient indexing structures and approximations are usually used to reduce the search space (H. Li & Homer, 2010). As a note, sequence comparison can also be used to automatically correct sequencing errors *before* alignment or assembly (I mention this problem below in section 1.5.5).

3. **Feature Counting** - Extraction of raw statistics (read counts, coverage profiles)
   The first basic information extracted from sequence alignment is the list of reads and their starting position. Practical summaries can be computed, first at a bp precision, such as *read depth* (the number of reads overlapping each position) or *read count* (the number of reads 5' ends starts). Often more general raw statistics are computed and quantify the amount of reads sharing the same feature, over a range of properties. The first obvious statistic aggregates and counts reads according to their localisation (let it be an exon, a gene, an upstream region, or simply equally spaced bins). Those raw counts inform on the amount of evidence retrieved per location. They are obviously proportional to sequencing power (the more we sequence, the more evidence we will get), but they are also estimates on the amount of molecules in the sample. Obtaining abundance estimate is one of the main aim in sequencing experiments (clearly for RNA-Seq assays, see Figure 1.3C). Another statistic, mostly used in a genome sequencing framework, merges all reads per position in a *pile up* of each of the nucleotides aligned. This is a simple summary from a sequence alignement (information about consecutive positions of a read are lost) but it can be used as an input by genotyping tools.

To facilitate exchange, traceability, and reproducibility, a few data formats that report information about sequence alignment and raw statistics have been developped (W J Kent, Zweig, Barber, Hinrichs, & Karolchik, 2010; H. Li et al., 2009). These standardised formats usually serve as an entry point for the analyses that follow: data visualisation, extraction of summary statistics or functional analysis. Quite often these raw statistics will also be a matrix containing count values for each feature as rows, and the different conditions or individuals as columns. As most classical data

analysis techniques operate on numerical matrices, exploratory data analysis usually can take place at this point.

4. **Visualisation** (genome browser)
   Owing to the sheer amount of data and the recent advances in interactive environment for data visualisation and analysis, data visualisation has become a staple of NGS data analysis. Traditional summary plots are produced on the fly by interactive genome browser (W James Kent et al., 2002; J. T. Robinson et al., 2011) and display the set of all reads alignments on a genomic interval, jointly with raw statistics and sequence annotation. The visualisation can go down to the level of single reads (see Figure 1.3B/C) up to chromosome wide view. This allows data analyst and experimentalist to inspect visually the various properties of the dataset and guides the formulation of hypothesis based on the data. Other representations make use of graph layouts and can be used to represent sequence assembly data (Wick, Schultz, Zobel, & Holt, 2015), summarise genome rearrangements (Krzywinski et al., 2009), or visualise transcripts isoforms (Rogers, Thomas, Reddy, & Ben-Hur, 2012).

5. Inference of **Summary Statistics** (list of variants, genome sequence, species abundance, expression level of exons/genes/isoforms, haplotyping)
   The specifics of the experiment's design defines the summary statistics which are required. For genome sequencing, the ultimate goal is simply to uncover the underlying genotype. When there is heterogeneity in the population (tumoral samples or metagenomes for instance), it will be the set of somatic variants. When a reference sequence is available, it breaks down into sub-questions such as haplotyping or variants annotation. For RNA-Seq data, the first goal is to detect and quantify the expression level (transcripts abundance) for each annotated transcript, usually starting from counts on regions and the reads spanning exon-exon junctions. For protein with specific DNA/RNA binding (Transcription Factor, RNA Binding Proteins), a set of intervals, detected by accumulation of reads at certain locations, describe the protein binding landscape. In the case of histone modifications, binding signal can be more spread and the summary will consist in profiles of regions characteristic of a certain type of binding. In all cases, the estimates will have to account for the different sources of variability (sampling variability, experimental variability, and biological variability).

6. **Functional analysis**
   After summary statistics are produced, an additional series of bioinformatics analysis are needed to enable the user to infer more information about the functions of the different biological entities. Quite often, this functional analysis starts from a set of summary statistics reported in a traditional tabular format –running over features and conditions. They can for instance be analysed using conventional data analysis and statistics, or serve for heterogeneous data integration and joint analysis across different assays (ENCODE Project, 2012; Kundaje et al., 2015). However, functional analysis has a broad meaning, as a large variety of analysis scenarios can exist from this point. I will describe only some case studies more closely related to my personal experience.

   With genomic experiments, beyond making genotypes, together with metadata available as a server (for instance for upcoming evolutionary studies), phenotype/genotype relationship is

first investigated using statistical techniques. This concerns for instance using Genome Wide Association Studies (GWAS), QTL mapping, rare variant analysis, positive selection, or in depth functional annotation of protein domains.

For transcriptomics studies, the first question usually is to extend the summary statistics to a complete characterisation of the population of transcripts (Alamancos, Agirre, & Eyras, 2014; Beretta, Bonizzoni, Vedova, Pirola, & Rizzi, 2014; Conesa et al., 2016; Steijger et al., 2013). Importantly, sequence reads provide an global picture of the transcribed fraction, and allow to catalogue and improve the annotation of all species of transcripts (Consortium et al., 2014; de Rie et al., 2017; Djebali et al., 2012; Tran, Souiai, Romero-Barrios, Crespi, & Gautheret, 2016), including protein encoding mRNAs as well as non-coding regulatory, structural or catalytic RNAs. New populations of RNAs can be singled-out through the analysis of read depth profiles and their localisation, combined with read sequence properties (Morillon & Gautheret, 2019). This had a dramatic impact in enlarging the bestiary of known non coding RNAs such as miRNA, but also circRNAs (T. B. Hansen et al., 2013; Memczak et al., 2013) or lincRNA (Hon et al., 2017). I was involved in a project for the in depth characterisation of the population of small RNAs in diatoms (Rogato et al., 2014), that I will detail later in section 1.6. An associated question is to detect and quantify changes in gene and transcripts abundance between conditions (Steijger et al., 2013), while accounting for variations due to technical as well as biological effects.

For ChIP-Seq and CLIP-Seq data, one of the main question is to estimate the sequence motif(s) associated with the binding regions, a classical problem in the analysis of biological sequences (Santana-Garcia et al., 2019; Tompa et al., 2005). In the case of ChIP/CLIP, it is coupled with sequencing data from enriched regions, in order to summarise the protein binding characteristics (Thomas-Chollier, Herrmann, et al., 2012; Zambelli, Pesole, & Pavesi, 2013)

This functional analysis step is also about **data integration**, where the information from the different assays is combined to provide a more global view. For instance, ChIP-Seq is often combined with RNA-Seq assays on the same experimental type, in order to link predicted bound regions to the expression values of the downstream genes (Thormann et al., 2018). Similarly variation in gene expression data are probed for differences in the genome content when combining DNA and RNA assays on a population of multiple individuals (eQTL analysis) (Heinig et al., 2010; Pickrell et al., 2010) This step of data integration is also very important when studying the determinants of the epigenome "code" (Ernst et al., 2011; Karlić, Chung, Lasserre, Vlahoviček, & Vingron, 2010; Mammana & Chung, 2015).

I chose to describe a generic pipeline in order to motivate some of the main methodological questions emerging for the computational analysis of sequence reads and that I will detail now.

1.4 Methodological Problems Emerging from Sequence Data

A series of fundamental methodological questions emerge during the analysis of high throughput sequencing data. I will present specifically the basic problems behind the tasks of sequence comparison, of summary statistics computation, and of functional analysis. Feature counting or data visualisation, which present their own algorithmic and statistical challenges, are not mentioned here.

This selection of general methodological problems is influenced from my own experience and is naturally not comprehensive. It still covers a variety of use cases. I will show afterwards (section 1.5) how to tackle most of those problems within the statistical framework of hierarchical models, using some of my contributions as an illustration.

1.4.1 Sequence Comparison

In the context of HTS, when a *reference* sequence is available, the alignment problem can be stated as:

> Given a set of *query* sequence reads, align them to their most likely originating position onto a set of *reference* sequences.

Solving this question was first framed as an optimisation problem. Given an alignment scoring function, usually evaluated as the sum of the contributions from each pair of aligned nucleotides, find the alignment with the highest score. Positions not aligning between the two sequences (creating *gaps*) are given penalties depending on their length (see below 1.5.3 for a formal definition).
Pairwise alignment was first used for detecting homologies between related sequences, using global (Needleman & Wunsch, 1970) and then local alignment (Gotoh, 1982; Smith & Waterman, 1981). Specific cases, such as spliced alignment necessitate specific score functions (van Nimwegen, Paul, Sheridan, & Zavolan, 2006). Traditional pairwise alignment can be solved exactly using dynamic programming with a cost in $\mathcal{O}(\ell n)$, where ℓ, n are the length of the query and the reference. For HTS reads the problem is slightly different, as a very large number of reads need to be aligned to the reference in a limited amount of time (H. Li & Homer, 2010). An added complication is the size of the reference (10^6 to 10^{10} bp) and its highly repetitive content (Treangen & Salzberg, 2012).

This triggered the development of efficient indexing techniques to efficiently elicit good candidate seeds (David Weese & Siragusa, 2017). Those seeds are are selected and extended using powerful heuristics (to control the tradeoff between the sensitive start from all candidates and the complexity caused by repetitive regions). The index is constructed from the reference genome (but it needs to be compressed to fit in memory) (H. Li & Durbin, 2009), or the reads (when they are expected to be overlapping a lot) (D. Weese, Emde, Rausch, Döring, & Reinert, 2009), or both (for alignment on multi-organisms reference databases, e.g. (Buchfink, Xie, & Huson, 2014; Siragusa, Weese, & Reinert, 2013). Data structure considered for indexing are hash tables, or suffix trees/suffix arrays and Burrows-Wheeler transform/FM-index. Suffix trees can summarise the set of all subsequences of the reference genome and allow to compute maximal overlapping regions once this data structure is built (A. M. S. Shrestha, Frith, & Horton, 2014).

There are multiple variations to the classical scoring scheme (Durbin, Eddy, Krogh, & Mitchison, 1998). One consists of allowing reads to have gaps of arbitrary length in their alignment, by extending the scoring scheme with a constant gap penalty. This type of alignment is practical to find reads identifying structural variants between the query and the reference, or when performing cDNA to genome alignments. However, if the cost for exact alignment with gap affine penalty is in $\mathcal{O}(\ell \cdot n)$, the cost for split alignment is in $\mathcal{O}(\ell \cdot n^2)$, which requires again appropriate data structures to balance the accuracy/speed tradeoff.

Repeats, combined with sequencing errors, create uncertainties for read placement. As a result, several alignments can possess almost the same score although result in placing the read at completely different locations. We thus always need to quantify the confidence of each reported alignments, given

the scoring scheme and the reference sequence. Fortunately, as we will see in next section (1.5) the alignment problem can be methodologically reduced into an estimation problem with Hidden Markov models (Durbin et al., 1998).

Finally, let's mention genome assembly, which is an important methodological problem based on sequence comparison, in fact one of the oldest problem in bioinformatics related to the sequencing revolution. There is no reference sequence available, so one possible strategy is to construct a graph summarising all pairwise similarities between reads with an overlap graph (called Overlap Layout Consensus Graph) (Nagarajan & Pop, 2013). In order to find the overlaps between the reads, it is necessary to compute all read pairwise alignments. This is not feasible with current read quantity, so associated data structure that can index the reads have been developed to perform the overlap in place (Simpson & Durbin, 2012). Within Laurent David PhD project (co-supervised with Alessandra Carbone), he developed a method based on Overlap Layout Consensus Graph for the targeted assembly of genes from metagenomics samples (David, Vicedomini, Richard, & Carbone, 2020). Presenting assembly would be a detour from the main methodological presentation on alignment, so I will not provide additional details.

1.4.2 Summary Statistics in genomics

A set of variant calls is the first summary needed from a genomic assay. Let's consider to simplify that we sequenced a query genome from a clonal population of cell. A *call* is the information describing a location from the query which is different from the reference. If a combination of multiple genomes would be considered, an estimate of the proportions of each of these could be also reported. According to the number of basepairs involved in the variation of sequence, two general types of variants are usually defined:

- Single Nucleotide Variants (SNV) or short insertions/deletions (indels). They are annotated using the set of aligned sequences around the variant location. The most common methods use Bayesian model, integrating local estimate of basepair errors, mapping errors, together with global estimates of sampling variability and polymorphism, and calls the genotypes which are the most likely *a posteriori* (Garrison & Marth, 2012)). Most current versions perform haplotyping as well in the case of polyploid genomes (Rimmer et al., 2014). More straightforward tools call SNV and indels by applying filters on read coverage using the reads pileup file (Danecek et al., 2011). The filters are set up to control the global false discovery rate. Recently, leveraging on the wealth of human ground-truth data available, Machine (Deep) Learning methods have been applied to the problem of SNV and short indels calling with unprecedented success (Poplin et al., 2018).

- Structural Variants (SVs) are defined as differences involving more than a few dozen bp (commonly above 50bp). SVs can be detected according to different strategies:

  - Direct evidence from the split-alignment of multiple reads. In this case, most of the methodological problems come from the alignment step. With collaborators at the University of Tokyo (co-contributor: A. MS Shrestha, helped by M. Frith and K. Asai), we designed a new framework that aligns a group of reads identifying a SV (A. M. Shrestha, Asai, Frith, & Richard, 2018). I will provide more detail in section 1.5.3.

- Mate-Pair or Paired-end data. When a fragment surrounds an SV, the alignment of the sequences at its ends can be used as an indirect evidence. Depending on the kind of variant, it will break some of the properties expected from the read pair (deletions lead to longer insert size, duplication to inverse read pair orientation). Alexandre Gillet developed such a kind of strategy during his PhD thesis (supervisors G. Fischer and I. Lafontaine) and I contributed the statistical analysis for detecting unexpected pairs (Gillet-Markowska, Richard, Fischer, & Lafontaine, 2015).
- From Read Depth (RD) data. For clonal populations, the distribution of read depth revolves around a mean coverage value with a variability resulting from fragments sampling and other steps of library preparation. Significant changes in the observed read depth coverage are an indirect indication of a deletion (drop in RD) or change in the Copy Number Variant (CNV, fold change in RD) (Ye, Schulz, Long, Apweiler, & Ning, 2009).
- From local assembly of the non aligned reads followed by a split alignment of the resulting contigs.

These different lines of evidence can also be combined to produce consolidated calls. Various tools follow this strategy now (Cameron et al., 2017; English et al., 2015; Layer, Chiang, Quinlan, & Hall, 2014; Rausch et al., 2012; Rimmer et al., 2014).

One could wonder why a set of variant calls is preferred to a complete assembly of the genome from the reads. One simple reason is that most of the existing data structures were developed with only one reference genome in mind. Furthermore, in the case of ploidy or mixture of genomes, there is an identifiability issue: given a set of variants they can be phased in multiple equivalent ways. There are multiple reasons for that. First, owing to limited read length combined with genomic repetitions and ploidy, most genome cannot be accurately assembled. Likewise mixture of genomes will present specific challenges [8]. Second, data structures that allow to treat collections of genomes (*e.g.* pangenomes) comprehensively have their own implementation challenges and emerged recently (Computational Pan-Genomics, 2018). Most of those considerations are now limited to the problem of haplotype construction, which consists of disentangling variants on the same chromosomal copy for a polyploid organism, or determining quasispecies for a population of viruses.

1.4.3 Summary Statistics and Functional Analysis in Transcriptomics

The central aim of RNA sequencing is to report the set of transcripts present, estimate their abundance, and update the gene models. As seen previously (section 1.2.1), transcriptomic assays are quantitative: the number of reads originating from a genomic region is related to the quantity of transcripts issued from that region. Given a model of reads sampling, we can obtain a first summary statistics for genes and transcripts by estimating their abundance using existing gene annotation, (see below in section 1.5.4 and (Pachter, 2011; Richard et al., 2010)).

RNA-seq precisely collects sequences from the expressed transcripts as well. Thus, functional analysis deals with refining the list of expressed RNAs, and understanding their possible disregulation between conditions. This can be summarized as the following methodological problems: annotation of transcripts (sequences and boundaries), differential expression, and discovery of new populations of RNAs and of their properties.

[8] this problem is methodologically similar to transcript quantification, which is presented in the next section

Quantification

Transcript quantification means to estimate the number of transcribed molecules for each transcript expressed in the cell population (Mäder, Nicolas, Richard, Bessières, & Aymerich, 2011; Ozsolak & Milos, 2010), given the set of read sequences. The usual problem is stated as: given sequence alignments, a reference annotation of exons boundaries, and an annotation of transcript isoforms, estimates the amount of reads within each isoform, normalized by sequencing depth and isoform length. However to facilitate the analysis, statistics are often combined at various levels of granularity. The most practical would be to pool expression level per gene (Mortazavi, Williams, McCue, Schaeffer, & Wold, 2008; Sultan et al., 2008), per exon (Anders & Huber, 2010; M. D. Robinson, McCarthy, & Smyth, 2010), per transcript isoform (Richard et al., 2010; Trapnell et al., 2012), or at the level of individual alternative splicing events (Katz, Wang, Airoldi, & Burge, 2010; Shen et al., 2014; Tran et al., 2016; E. T. Wang et al., 2008b). Global gene expression level can be decomposed respectively as the expression level of each exon, each transcript or each alternative splicing event. Note that this methodological question, which we could call source estimation, appears quite often in other contexts. Some examples are for instance the proportion estimation from mixture of genomes that I mentioned previously, or the quantification of metagenomes as a mixture of source samples.

In all cases, the reads are *de facto* sampled randomly according to the abundance of the molecule (the number of molecules in the library), and various factors influence the results. Thus inference made for obtaining abundance statistics needs to account for this randomness (reads are sampled according to their abundance/the number of molecules in the library), and possible biases coming from the different steps during library preparation (biological and experimental variability).

In a more general form, annotation is unknown or only partially known, and the set of isoforms that are compatible with sequence evidence is enumerated as quantification proceeds (see annotation below).

Transcripts Annotation

Complete transcript annotation has proven to be a complex task (Steijger et al., 2013), and can be divided into subproblems.

Annotation of transcript boundaries It first means describing the list of all transcription start site (TSS) and transcription end site (TES), as well as the structure of exons. TSS and TES can be detected by searching for significant changes in expression level (Mirauta, Nicolas, & Richard, 2014; Tran et al., 2016; Trapnell et al., 2010) (see section 1.5.4) or by analysing specific assays enriching reads around the transcript ends(Kanamori-Katayama et al., 2011). The exon boundaries can be also identified using split-alignment of reads that overlap exon-exon junctions, usually taking into account sequence signals next to acceptor/donor sites (Dobin et al., 2013; Iwata & Gotoh, 2012; Jean, Kahles, Sreedharan, De Bona, & Ratsch, 2010; Philippe, Salson, Commes, & Rivals, 2013; K. Wang et al., 2010). "Calling" the junctions implies then to be able to correct some artifacts that can impair proper annotation, such as junction coming from transcriptional noise (inaccurate cutting by the splicing machinery) or ambiguous boundaries (micro-homologies next to the splice donor and acceptor sites) (van Nimwegen et al., 2006). Alignment can also be done directly on the splice graph (Denti et al., 2018).

Transcript Isoform reconstruction Given the set of existing transcript boundaries, multiple transcript isoforms can overlap on the same genomic region. One following task is then to reconstruct the set of isoforms that best explain aligned sequences and boundaries (Bernard, Jacob, Mairal, & Vert, 2014; B. Li & Dewey, 2011; Richard et al., 2010; Trapnell et al., 2010) given existing databases (Koscielny et al., 2009) or *de novo* (Haas et al., 2013; Robertson et al., 2010; Schulz, Zerbino, Vingron, & Birney, 2012). Transcript isoform reconstruction is a complex task that has strong identification issues, as the limited length of the reads will imply that the number of possible isoforms will increase exponentially with the number of exons and the enumeration of all isoforms is not possible. Usually this task is combined with the estimation of transcripts' abundance, by adding a regularisation constraint on the lowly expressed transcripts. In this case it can be rewritten under a convex optimisation framework where the set of isoforms does not need to be enumerated explicitly (Bernard et al., 2014).

Differential Analysis

When multiple conditions or cellular states are assayed, the most pressing question is to detect transcripts whose concentrations changed significantly according to the conditions. In other words, the question is to find RNAs whose abundance changed, after having disentangled biological effect from experimental variability. After having aggregated the read evidence per feature (gene, exons, isoforms) and normalised the expression signal between conditions –accounting for sequencing depth, this problem can be efficiently tackled within a linear model framework. One particularity here is that the data is essentially digital and thus values are accounted for with count models (Anders & Huber, 2010; Love, Huber, & Anders, 2014; M. D. Robinson et al., 2010). I worked on methods for differential expression in two contexts: differential exons usage between conditions (Richard et al., 2010), *de-novo* detection of regions that are differentially expressed between conditions (Mirauta, Nicolas, & Richard, 2013).

1.4.4 Summary Statistics and Functional Analysis for ChIP/CLIP-seq

ChIP-Seq or CLIP-Seq reads inform on the binding of a target protein to DNA or RNA, respectively. After fragmentation, read pair ends are expected to accumulate around the bound target protein. Thus, the summary statistic will combine read locations to deduce the set genomic intervals bound to the target protein. It is obtained by detecting so called *peaks*, regions where the read depth of the bound fragments is higher than expected from the background. Background read depth can be estimated using the input experiment (see 1.2.1) or using read depth genome wide distribution when no input was produced. This analysis is done by integrating two lines of evidence: first defining candidate peaks where the reads accumulate using longitudinal clustering methods, and then decide for each cluster if its read depth is significantly higher than expected (e.g. given the depth on the input region). Ideally, the number of reads binding to a location is proportional to the binding affinity of the protein to the region and this value can be reported. Peaks detection can also integrate a step of *peak classification*, where the profile of read placement is used to to separate different proteins (for instance broad for histones, and narrow for Transcription Factor Binding Sites). In the case of truncation based CLIP experiments, 5' ends of fragments are enriched in cross link locations. Thus read starts are additionally used as diagnostic events to infer cross link sites.

Once the peaks have been identified, it is usually followed by functional analysis, where the

question is to find the set of sequence motifs that the best explain the peaks location (Thomas-Chollier, Darbo, et al., 2012; Thomas-Chollier, Herrmann, et al., 2012). It is meant to detail *a posteriori* the sequence determinant of binding for the protein. The problem of motifs detection is almost as old as sequence alignment (Durbin et al., 1998).

1.5 Probabilistic Models for Sequencing Assays

As I detailed previously, each step of library preparation comes with its own artifacts and biases (section 1.2.3). Furthermore, randomness is at the heart of all sequencing protocols: random sampling by sequencing. Probabilistic models provide a principled way to integrate those sources of uncertainties while adjusting to the type of biases present in each library.

A natural class of models in this context are hierarchical models, where the observations (the sequences, or their abundance) can be explained by the value of one or multiple latent (hidden) variables (*e.g.* the similarity between sequences, the original amount of an RNA molecule). Conditional probabilities are specified to link the variables of the model, whose general structure is summarised with a dependency graph (Figure 1.4 left). By specifying the graph structure and the conditional distributions we explicit the relationship between latent variables and observations. When the structure of the dependency graph has good properties (for instance is a Directed Acyclic Graph (Bishop, 2006)) we can infer, given the observation, the more likely values of the latent variables.

State-space models (SSM), are a subclass of models that account for the dependencies between neighbouring position in a sequence of observations. They make the assumption that the distribution of the latent variable at a position, depends only on the values at the previous positions (see Figure 1.4 right). As genomic data is indeed sequential, SSM are a very common modelling choice in computational biology. SSM also provide a good cost/reward balance by allowing to devise versatile models which are still tractable for exact solutions or give fast answers with good approximations.

In the following, I will first provide a general introduction to hierarchical and state-space models, by recalling the main statistical questions related to inference and estimation, and I will illustrate the use of these models on two problems I worked on: sequence alignment and estimation from read counts. In the latter case, an important part of the problem consists in modelling the observations correctly (*i.e.* not under/overestimating the different factors of dispersion). I will provide in section 1.5.5 a compilation of the common models that have been proposed to account for variability in sequence counts and detail some model we developed.

1.5.1 State-space Models and Hidden Markov Models

We consider sequences of size T, with observations $(y_t)_{t=1,\dots,T}$ (or $y_{1:T}$) and a corresponding process on hidden states $(x_t)_{t=1,\dots,T}$ (or $x_{1:T}$). State space models are aiming at the reconstruction of the sequence $x_{1:T}$, which is not directly available, from the observed measurements $y_{1:T}$. To do so, it uses two components: a probabilistic model that link observations to the latent space states, and a description of the latent space dynamics. We will use in the following SSM with emission densities: $\pi(y_t \mid x_t) =: e(y_t; x_t)$ and a Markov model on the latent variable x_t with values in Ω_x, $\pi(x_t \mid x_{t-1}) =: k(x_t \mid x_{t-1})$. The model on the latent variable k is also refered to as the transition kernel. When Ω_x is a discrete space, we usually call the SSM a Hidden Markov Model (HMM). HMM are one of the most common probabilistic model in computational biology (Durbin et al.,

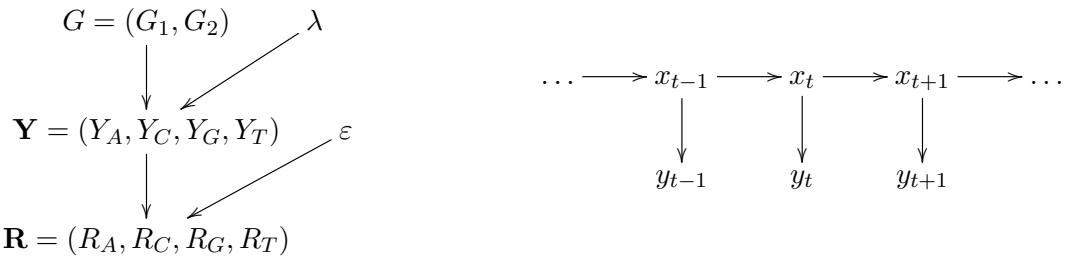


Figure 1.4: Examples of probability diagram for two hierarchical models. Left: a model explaining the letters piling up at one position in a diploid genome (see 1.3 and 1.4.2), given the genotype $\mathbf{G}$, the expected coverage λ and the error rate μ. Fragments are first sampled according to a Poisson distribution for instance: $\mathbf{Y} \mid G = (A, G) \sim (\mathcal{P}(\lambda/2), 0, \mathcal{P}(\lambda/2), 0)$ (more on the Poisson distribution in section 1.5.5). The observations are the read counts $\mathbf{R}$. They can be subject to possible error rates according to a binomial distribution. For instance $R_A \mid \mathbf{Y} = \mathbf{y} \sim \mathcal{B}(y_A, (1-\varepsilon)) + \mathcal{B}(y_C + y_G + y_T, \varepsilon/3)$. Right: Example of a traditional state space model, where the observations are the $(y_t)_{t=1,\ldots,T}$ and the hidden variables are the $(x_t)_{t=1,\ldots,T}$. The x_t are distributed according to a Markov chain. This model can be used for instance in the estimation of expression levels from read counts (see 1.5.4). Note that in both models, most information about the model come from the edges that are absent from the graph as they are the one implying conditional independency relationships.

1998). This model can be extended by allowing dependency of the observations y_t on more hidden states ($\pi(y_t \mid x_{t-k:t})$), or on previous observations ($\pi(y_t \mid x_{t-k:t}, y_{t-\ell:t-1})$).

In the context of sequence data analysis, we are interested in reconstructing the path with highest posterior probability $\arg\max_{x_{1:T}} \pi(x_{1:T} \mid y_{1:T})$ and the marginal posterior probabilities of hidden states $\pi(x_t \mid y_{1:T})$ as well as characterising differences between experiments with estimated parameters. We design such models to be able to get simple and meaningful statistics where model complexity (capturing the sources of variability in the data) and its computational tractability (for parameter estimation) are well balanced.

A naive approach to evaluate those quantities would require to enumerate all states combinations (of size $|\Omega_x|^T$ for discrete spaces). However, by using the recurrences embedded in the model structure, those quantities can be obtained efficiently through an iterative sequence of updates (Rabiner & Juang, 1986):

- $\pi(x_t \mid y_{1:t-1})$ - *prediction.* The distribution of x_t conditioned by the sequence of previous observations –$y_{1:t-1}$, can be obtained by integrating over all possible x_{t-1} values (markovian property on the x_t).

$$\pi(x_t \mid y_{1:t-1}) \propto \int \pi(x_{t-1} \mid y_{1:t-1}) \cdot k(x_t \mid x_{t-1}) \, \mathrm{d}x_{t-1}$$

- $\pi(x_t \mid y_{1:t})$ - *filtering.* The distribution of x_t conditioned by the sequence of observations up to the current one $y_{1:t}$, is deduced from the prediction density updated with the likelihood of the observed data at t,

$$\pi(x_t \mid y_{1:t}) \propto \pi(x_t \mid y_{1:t-1}) \cdot e(y_t \mid x_t)$$

- $\pi(x_t \mid y_{1:T})$ - smoothing. The posterior distribution of x_t accounting for the complete sequence of observations $y_{1:T}$ can be written in terms of filtering and prediction, after integrating over

all possible upcoming state values x_{t+1}.

$$\pi(x_t \mid y_{1:T}) \propto \pi(x_t \mid y_{1:t}) \cdot \int \frac{\pi(x_{t+1} \mid y_{1:T}) \cdot k(x_{t+1} \mid x_t)}{\pi(x_{t+1} \mid y_{1:t})} \, \mathrm{d}x_{t+1}$$

The likelihood of the observed sequence can also be obtained by carrying out similar recurrences on $\pi(y_t \mid y_{t-1})$.

When Ω_x is discrete, reconstructing the path of hidden variables allocation with the highest posterior probability $\arg\max_{x_{1:T}} \pi(x_{1:T} \mid y_{1:T})$ can also be obtained with similar recurrences (Rabiner & Juang, 1986).

$$\begin{aligned}
\phi_t(j) &:= \log \max_{x_1,\ldots,x_t} \pi(x_{1:t-1}, x_t = j \mid y_{1:t}) \\
\phi_t(j) &= \max_{i \in \Omega_x} [\phi_{t-1}(i) + \log k(j \mid i)] + \log e(y_t \mid j)
\end{aligned}$$

The quantity $\max_i \phi_T(i)$ returns the highest log probability for the sequence and a corresponding path can then be reconstructed by backtracking.

I am exploiting this class of models at two different steps of the sequence analysis pipeline (section 1.3), sequence comparison and production of summary statistics. The first step uses raw sequences as input, and the questions are then related to comparing a set of reads to a reference sequence (such as a genome or a transcriptome). In this case we consider ordered pairs of observations for $(y) = (r, q)$ where $r = r_1 \ldots r_T$ and $q = q_1 \ldots q_S$ are both sequences of letters taken from the DNA alphabet $\Sigma = \{\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t}\}$. r is the reference sequence and q the read sequence (the query) that we want to align. In the other case, we are working with summary statistics consisting of aggregated counts (per position or per feature) and want to consider directly abundances of certain molecules or coverage of regions in the genome. In this case, the observations y_t are integer count data.

I will first present a straightforward hierarchical model that estimates transcript isoforms proportions, given the read counts within exonic regions (Richard et al., 2010). The next subsection deals with the general formulation in the case of sequence alignment, and introduces one of my contribution for joint split alignment of reads (A. M. Shrestha et al., 2018) . The section after introduces two transition kernels we developed for transcriptional profiling (Mirauta et al., 2014) and the analysis of protein RNA binding (Krakau, Richard, & Marsico, 2017).

1.5.2 Isoform Abundance Estimation

We look as the problem of estimating transcript abundance from RNA sequencing data. I will present a simple version of the model that I proposed at the dawn of the technology. This problem has since received a good amount of methodological development, jointly with consortium efforts towards proper benchmarking of the methods (Steijger et al., 2013).

Contribution: Estimating transcript proportions from exonic read counts We consider a gene with n exons, and the observations are the number of reads aligning within each exon $\mathbf{y} = (y_1, \ldots, y_n)$.

The absolute transcript proportion is not directly accessible so we aim to reconstruct the expected number of reads contributing to each of the k isoforms $\mathbf{x} = (x_1, \ldots, x_k)$. We hypothesise that we know the exon composition of each isoform, which is encoded using a binary matrix $I_{e,j}$ such

that $I_{e,j} = 1$ if exon e is part of isoform j. The expected read count in an exon y_e is the sum of all transcripts contributions (p_e is a normalisation factor related to the length of each exon : $p_e = \frac{l_e}{\sum_{i=1}^{n} l_i}$):

$$y_e = \sum_{j=1}^{k} \frac{p_e}{\sum_i p_i \cdot I_{i,j}} \cdot I_{e,j} \cdot x_k$$

For a transcript j we can specify a hierarchical model for the distribution of counts of isoform j in exon e.
First, the count for each isoform follows a Poisson distribution

$$X_j \sim \mathcal{P}(\lambda_j) \qquad \text{with } \lambda_j = \lambda \cdot \frac{1}{\sum_i p_i \cdot I_{i,j}} \cdot q_j$$

and the conditional distribution for Y_e^j given x_j reads in the transcripts is a multinomial

$$(Y_1^j, \ldots, Y_n^j) \mid X_j = x_j \sim \mathcal{M}\left(\left(\frac{p_e}{\sum_i p_i \cdot I_{i,j}} \cdot I_{e,j}\right)_{e=1,\ldots,n}, x_j\right)$$

λ is a normalisation factor depending on the depth of sequencing of the library and on the relative proportion of the isoform. This model is fully parameterised and the estimation of the parameters can be solved directly with an EM algorithm.

I designed the model and implemented estimation and inference with this model as part of a set of tools we developed for the analysis of alternative splicing in RNA-Seq data. The methodological parts were done with Marcel Schulz and with Marc Sultan, we performed experimental validations using a large panel of quantitative RT-PCR experiments. This work took place during my postdoc at the Max Planck Institute for Molecular Genetics in Berlin (Richard et al., 2010).

State of the art now: The problem of transcript quantification evolved a lot since to account for most systematics biases. The first extension of the model was to consider the contribution of each fragment to the transcript abundance as a hidden variable (B. Li & Dewey, 2011). This way, it is possible to jointly estimate longitudinal effects of fragments placements, fragment length, as well as sequence composition effects when looking for transcript abundance (introduced in section 1.5.5). In (Pachter, 2011), a general review of hierarchical models that can be considered for isoform abundance quantification was detailed. However, those model formulations are linear in the number of read alignments for each iteration of the EM algorithm, and cannot scale up as sequencing throughput increases. The problem thus shifted to exploiting an approximate factorisation of the likelihood function (Nicolae, Mangul, Măndoiu, & Zelikovsky, 2011; Zakeri, Srivastava, Almodaresi, & Patro, 2017) together with *quasi-matching* of the reads to be able to perform reasonable quantification in a short amount of time (Bray, Pimentel, Melsted, & Pachter, 2016; Patro, Duggal, Love, Irizarry, & Kingsford, 2017). As I mentioned previously, the same methodological question was also framed in other contexts, such as metagenomics profiling or mixture of samples estimation, and can be treated within this formal framework.

1.5.3 Sequence Alignment

In the case of pairwise alignment, the question can be stated as, starting from two sequences of letters, r (a reference) and q (a query), find the best way to align one to the other. There are broadly two type of alignment: global (end to end) and local (best scoring). In practice, when globally aligning two sequences, we rewrite them as $\tilde{r}$ and $\tilde{q}$ by inserting gap characters `"-"` such that $|\tilde{q}| = |\tilde{r}|$, as represented below on an example.

$$\begin{array}{ll} r = \texttt{tatcgtacgggagcaaatgt} & \tilde{r} = \texttt{tatcgtacgggagcaaatgt} \\ q = \texttt{tatcgtgcgcatgt} & \tilde{q} = \texttt{tatcgtgc----gc--atgt} \end{array}$$

As we already mentioned in 1.4.1, the traditional formulation for finding an alignment between two sequences involves a score function that rewards matches, and penalises mismatches and non aligned segments. A corresponding probabilistic model uses a HMM that is defined with a sequence of observations as a pair of sequences $(r_{1:T}, q_{1:S})$ with values in $\{\texttt{a}, \texttt{c}, \texttt{g}, \texttt{t}\}$, and a hidden process $\mathbf{x}$ that informs on the state of the two emitted letters (are they aligned or not). This process takes values in $\mathcal{A} = \{\texttt{M}, \texttt{I}, \texttt{D}, \texttt{B}, \texttt{E}\}$ and emits pairs of letters (columns of the alignment), given the type of aligned bases with the following emission probabilities:

$$\begin{aligned} e(r_t, q_s; \texttt{M}) &= \sigma(r_t, q_s) && \text{(a Match of letters } r_t \text{ and } q_s\text{).} \\ e(-, q_s; \texttt{D}) &= \phi(q_s) && \text{(a Deletion in front of } q_s\text{)} \\ e(r_t, -; \texttt{I}) &= \psi(r_t) && \text{(an Insertion of } r_t\text{)} \end{aligned}$$

States `B` and `E` are not emitting observations (`"silent"`) and anchor the alignment at the beginning and the end (see Figure 1.5).

The corresponding transition kernel k considers different types of successions of alignments, according to how we consider the alignment of the query to the reference should be. The two most common cases are global and local alignments. For instance, in the case of aligning reads to genome, the query is much shorter and a local alignment kernel, where match states are surrounded by insertions and deletions around the query, is most appropriate (Durbin et al., 1998).

As an example, let's consider the simple case of a gapless alignment and write down its likelihood. The c first letters of R and the d letters of Q are unaligned, the next e letters of R and Q are aligned, and the final f letters of R and g letters of Q are unaligned ($c+e+f = T$ and $d+e+g = S$) . The probability of the alignment is (Frith, 2019):

$$\begin{aligned} \pi(r_{1:T}, q_{1:S}, \mathbf{x}) = &\left(\prod_{t=1}^{c} k(\texttt{D},\texttt{D})\psi_{r_t}\right)(1 - k(\texttt{D},\texttt{D})) \cdot \left(\prod_{s=1}^{d} k(\texttt{I},\texttt{I})\phi_{q_s}\right)(1 - k(\texttt{I},\texttt{I})) \\ &\left(\prod_{t=1}^{e} k(\texttt{M},\texttt{M})\sigma(r_{c+t}, q_{d+t})\right)(1 - k(\texttt{M},\texttt{M})) \\ &\left(\prod_{t=1}^{f} k(\texttt{D},\texttt{D})\psi_{r_{c+e+t}}\right)(1 - k(\texttt{D},\texttt{D})) \cdot \left(\prod_{s=1}^{g} k(\texttt{I},\texttt{I})\phi_{q_{d+e+s}}\right)(1 - k(\texttt{I},\texttt{I})). \end{aligned}$$

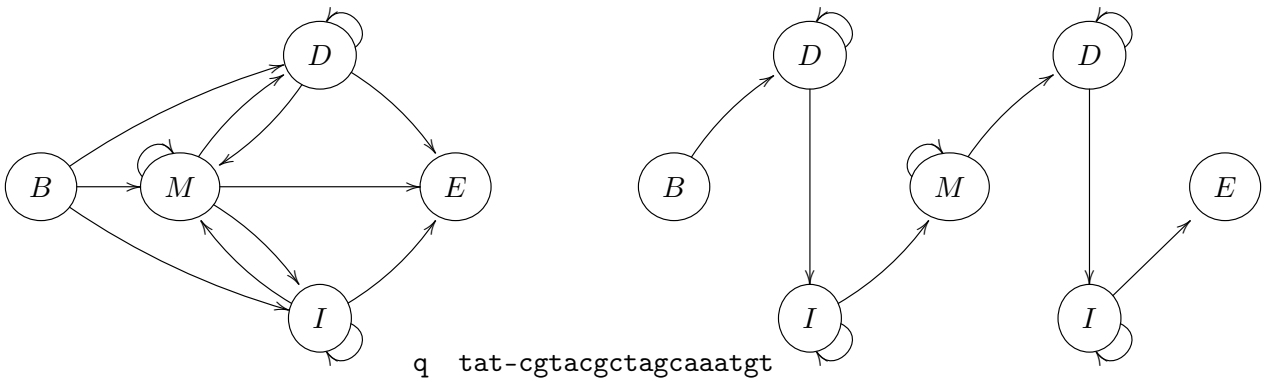


Figure 1.5: Top: examples of transition kernels used for pairwise global alignment (left) and ungapped local alignment (right). Bottom: A global alignment of the sequences q and r with the corresponding sequence of states $\mathbf{x}$ underneath. Note that an other alignment of r with the same score and probability exists.

We can factor this probability with a constant K defined as[9]:

$$K = \left(\prod_{t=1}^{T} k(\mathtt{D},\mathtt{D})\phi_{r_t}\right)\left(\prod_{s=1}^{S} k(\mathtt{I},\mathtt{I})\phi_{q_s}\right)(1-k(\mathtt{D},\mathtt{D}))^2(1-k(\mathtt{I},\mathtt{I}))^2(1-k(\mathtt{M},\mathtt{M}))$$

and taking the logarithm of the normalized term, we are left with maximizing:

$$\ln\left(\frac{1}{C}\,\pi(r_{1:T}, q_{1:S}, \mathbf{x})\right) = \sum_{t=1}^{e} \ln\left(\frac{k(\mathtt{M},\mathtt{M})}{k(\mathtt{D},\mathtt{D})k(\mathtt{I},\mathtt{I})}\frac{\sigma(r_{c+t}, q_{d+t})}{\phi_{r_{c+t}}\psi_{s_{d+t}}}\right)$$

With this formulation finding the most probable alignment path is equivalent to finding the alignment of maximal local score, where the score can be defined as the sum of the contributions from the individual matching bases, reweighted according to their expected background occurrence distributions, ϕ and ψ. We could thus define a substitution score matrix S in the following way:

$$S(a,b) = \lambda \cdot \ln\left(\frac{k(\mathtt{M},\mathtt{M})}{k(\mathtt{D},\mathtt{D})k(\mathtt{I},\mathtt{I})}\frac{\sigma(a,b)}{\phi_a\psi_b}\right) \quad a,b \in \Sigma^2$$

This highlights the link between score based and model based alignments (λ is a rescaling constant). The computation carry out in a similar way for alignments with gaps affine costs (Frith, 2019).

Probabilistic treatment of the alignment problem is extremely important due to the large amount of repeats that are present in most genomes and the necessity to identify regions in the target sequence that will be difficult to align or sequences that cannot be placed reliably. In the case of the alignments of sequencing reads this quantity is referred to as the mapping quality, e.g. the PHRED scaled probability that the read is misaligned. This was introduced for the first time for HTS reads with the MAQ aligner (H. Li, Ruan, & Durbin, 2008) and it is now a fixture of most aligners.

[9]Note that is very similar to considering a likelihood ratio with a model where r and q are not aligned.

Contribution: a principled framework for joint split alignment With collaborators at the University of Tokyo (A. MS Shrestha, M. Frith, K. Asai), we designed a new framework that aligns a group of reads identifying a SV. I designed the first draft of the method during a research stay at the CBRC in Tokyo and the core implementation and evaluation was then done with Anish MS Shrestha. We are considering here *split-alignments*, alignments where two different portions of a read align to disjoint genomic locations on the reference. We proposed a new principled framework that exploits the probabilistic set-up of the pair-HMM, in order to score a set of sequences that are identifying conjointly a structural variant. The need for a probabilistic treatment is apparent when one considers the ambiguities stemming from repeated regions. Indeed, even under ideal conditions of no sequencing errors and very high coverage, we estimated that 40% of deletions cannot be identified with certainty by pairwise alignments of 100bp reads(A. M. Shrestha et al., 2018).

To solve this problem, we first reformulate the joint split-alignment problem as a profile-to-sequence alignment problem. We consider that given the reference genome $\mathcal{G}$ and a set $\mathcal{R}$ of reads originating from a region $\mathcal{H}$ in the assessed genome and that contains an SV. Our aim is to find the split-alignment of $\mathcal{H}$ to $\mathcal{G}$ which determines the position of the SV. Note as well that the sequence of $\mathcal{H}$ in the SV locus is not observed. We are scoring a split alignment A according to two types of scores: (1) $S_{\text{genome}}(A; \mathcal{H}, \mathcal{G})$ accounts for the evolutionary divergence between $\mathcal{G}$ and $\mathcal{H}$; and (2) $S_{\text{sequencer}}$ which weights the agreement of each read to $\mathcal{H}$. We also consider that a profile matrix C can be constructed from the multiple sequence alignment of the reads to $\mathcal{H}$. We can now score $\mathcal{H}$ and an alignment A of the matrix C to $\mathcal{G}$ as:

$$\begin{aligned} S(\mathcal{H}, A; \mathcal{G}, \mathcal{R}) =& S_{\text{genome}}(A; \mathcal{H}, \mathcal{G}) \\ &+ \sum_{i=1}^{\ell} \sum_{x \in \mathcal{D}} C_{xi} \times S_{\text{sequencer}}(\mathcal{H}_i, x), \end{aligned}$$

The parameters used in the scoring schemes are shown in Figure 1.6a. S_{genome} consists of a substitution matrix, affine gap penalties for small indels, and a constant penalty for large splits, and $S_{\text{sequencer}}$ consists of a substitution matrix and linear gap penalties. The values in both scoring schemes can be adjusted to reflect the relative importance of the reference genome and of the read sequences. We show an application of our scoring model on a toy example in Figure 1.6b. Note that the choice of linear gap penalties allows us to express the alignment score equivalently as the score of C-to-$\mathcal{G}$ alignment, with each column of C treated independently.

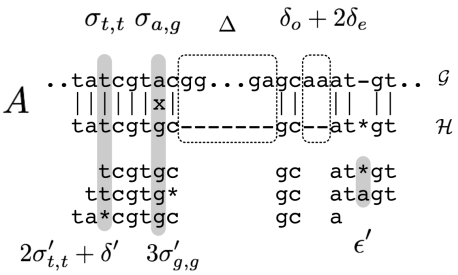
This extended model reconstruct the implicitly sequenced genome around breakpoints by doing profile to sequence alignment. I order to be able to compute those profile-to-sequence split alignments in a reasonable running time, we developed an approximate solution to the problem using as seed alignments pairwise local alignments of the reads to the reference. Our method follows multiple steps, as depicted in Figure 1.7.

In developing the method, we made extensive use of the underlying probabilistic model to be able to balance the running time constraints with the accuracy of the results. This proved to be efficient. Our method, despite being implemented in a scripting langage, has in practice running time on par with more optimised tools to which we compared us: Seghemel, Deli, Lumpy and Splazers and Platypus (detailed in (A. M. Shrestha et al., 2018)). There were already different split alignment strategies proposed previously, but none until now was considering the problem of joint alignments nor giving it a probabilistic treatment.

a

	$\mathcal{H}$-to-$\mathcal{G}$ (S_{genome})	Sequencer ($S_{\text{sequencer}}$)
substitution	σ	σ'
deletion open	δ_o	-
deletion extend	δ_e	δ'
insertion open	ϵ_o	-
insertion extend	ϵ_e	ϵ'
split	Δ	-

b



c

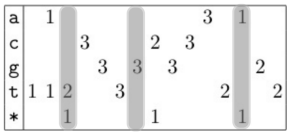


Figure 1.6: Scoring a joint split-alignment, from (A. M. Shrestha, Asai, Frith, & Richard, 2018). (a) Parameters of the scoring scheme. (b) Toy example with an alignment A of three reads identifying two deletions (dashed boxes). Computation of the score is indicated on some columns of the alignment – highlighted in gray, with contributions from S_{genome} and $S_{\text{sequencer}}$ at the top and at the bottom, respectively. (c) The profile matrix for the example (a 15bp count profile). The choice of linear gap penalties in $S_{\text{sequencer}}$ allows us to express the joint split-alignment as the alignment of C to $\mathcal{G}$, with columns of C treated independently. The gray columns match the ones in (b).

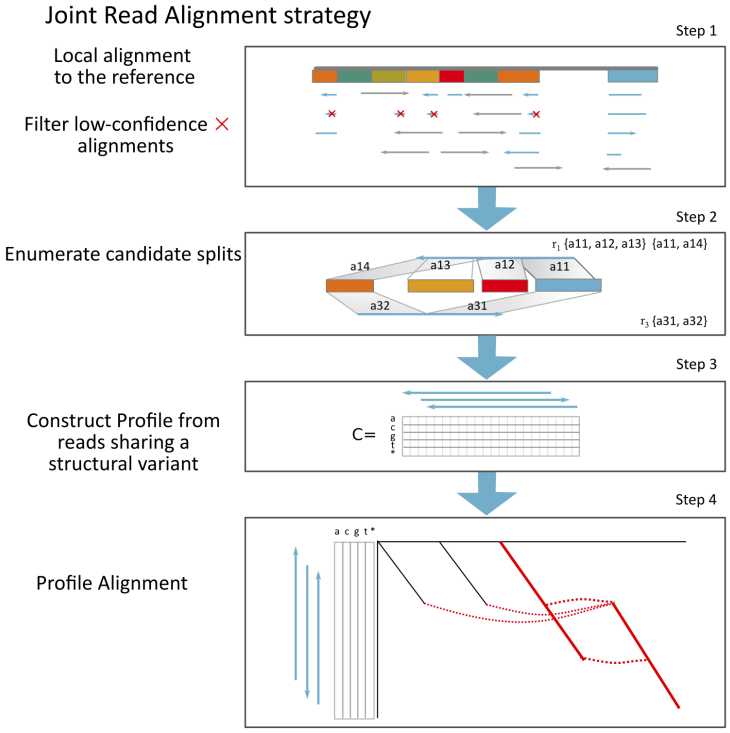


Figure 1.7: The main steps of our Joint Read Alignment strategy. First, we perform local pairwise alignment of each read to $\mathcal{G}$ (Step 1). Next, we extract reads that are likely to have originated from SV sites by enumerating candidate SVs that can be inferred from their local alignments (Step 2). We group these reads according to the candidate variant site they point to. Each group along with $\mathcal{G}$ forms an instance of the joint split-alignment problem described previously. We solve each instance by constructing a profile matrix of the reads (Step 3), and finding its maximum-scoring alignment to $\mathcal{G}$ (Step 4). Finally we compute a confidence value for the joint split-alignments (Step 5).

We demonstrated the advantages of our method, over other split-aligners, by applying it to the problem of identifying medium and large deletions ($\geq$ 20bp) from typical human genome resequencing datasets.

State of the art now: The specific problem of split alignment has not received much attention lately. This can be attributed to the fact that the attention of the community moved towards long reads and the renewed need for efficient and accurate alignment techniques. The construction of sequence profiles on the other end experienced a renewed interest, with many publications aiming at constructing the sequence profile in form of a Partial Order Graph for error correction (polishing) in long reads (Gao et al., 2020; Vaser, Sović, Nagarajan, & Šikić, 2017).

1.5.4 HMM Models using Count Observations

HMM are a natural choice for segmenting longitudinal data into a set of coherent groups. They provide at the same time the group delineation and estimates on the properties of each group. HMM are one of the most successful models in bioinformatics, they are used since the dawn of sequencing for annotation (Burge & Karlin, 1997; Nicolas et al., 2002). With the HTS revolution, those methods started to be used for clustering the signal coming from the depth of aligned reads along the genomic sequence.

First contribution: Parseq

After I integrated the LCQB, I wanted to consider the problem of transcript abundance estimation in a more general framework. Starting from a work of Pierre Nicolas (researcher, INRAE) on tiling array data (Nicolas et al., 2009), we supervised Bogdan Mirauta's PhD. The subject was to develop a method to reconstruct the expression levels at each genomic location, given the RNA-Seq read counts.

Without considering prior annotation, one way is to use the fact that neighbouring bases likely come from the same transcript and thus that their expression level should marginally change. The observations y_t are here the counts from the 5' end of reads. A Hidden Markov Model can integrate this information about neighbouring positions, by reconstructing the unobserved expression levels x_t given the counts y_t. We use a transition kernel that can detail the different events occurring along the sequence with an impact on the expression level. A position can be either expressed ($\mathbf{I}_{\{x_t>0\}}$) or non-expressed ($\mathbf{I}_{\{x_t=0\}}$). Within expressed regions there can be changes in x_t, either due to overlapping transcripts, or to other effects (summarised with the function g below). The transition kernel can then simply write as:

$$k(x_t, x_{t-1}) = \mathbf{I}_{\{x_{t-1}=0\}}[(1-\eta)\delta_0(x_t) + \eta f(x_t)] + \mathbf{I}_{\{x_{t-1}>0\}}[(1-\beta_0)g(x_t; x_{t-1}) + \beta_0\delta_0(x_t)]$$

η and β_0 give the probabilities of stopping and starting transcription, while f and g are generic densities for expressed positions. Following a the preliminary work of Pierre Nicolas on tiling array data (Nicolas et al., 2009), the g density is as well a mixture, with three components: (1) unchanged transcription level, or for changes that differ by their amplitudes and are referred as (2) shifts (large amplitude) and (3) drifts (small amplitude). Figure 1.8 represent the different possible movements on the transitions kernel along a sequence.

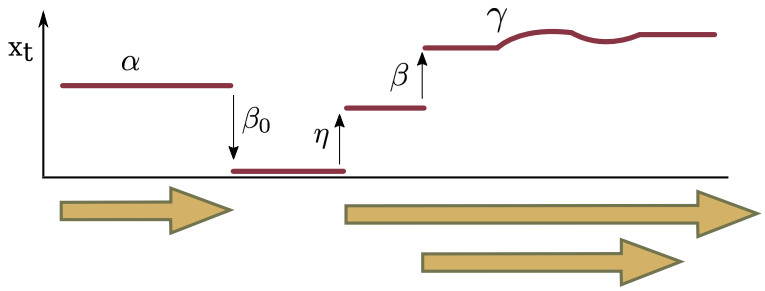


Figure 1.8: Illustration of the different changes in the expression level x_t that are taken into account with the model. Genes are represented with orange arrows below. The distribution $g(x_t; x_{t-1})$ can be decomposed in a term of change of expression (shifts, probability β) and a term for proportional changes (drifts, probability γ).

Applying such model gives a rich and detailed description of the transcriptional landscape along the genome: breakpoints in expression level, indicative of starts and end of transcription, or the probability that a given region is expressed. Integrating the information about x_t estimates additionally enables annotating transcribed regions in the sequence as well and transcription starts and end (see Figure 1.9) The main challenge in this context was twofold. First the hidden process is in a continuous state space. Estimation and inference is not amenable by the classical methods presented previously (section 1.5.1) and need to be substituted by particle filtering algorithms build on sequential Monte Carlo principles (Doucet & Johansen, 2009). The sequence length, ranging in millions of bp increases the difficulty as well. Then, the sequence counts within transcribed regions exhibited much more variability that what could be expected with traditional models. This triggered the development a new distribution for read counts with extra variance which I describe in section 1.5.5.

As an estimate on expression level is derived with confidence interval for each position, this information can also be used in a multiple conditions setup to detect differentially expressed regions, and we proposed a prototype for this question in (Mirauta et al., 2013).

Second contribution: PureCLIP

During her PhD with Analisa Marsico (main supervisor, Max-Planck institute for Molecular Genetics) and myself, Sabrina Krakau worked on the analysis of CLIP-Seq data. When we assay CLIP-seq data, we aim at reconstructing both the bound regions and the exact positions where there was a crosslink between the protein and the RNA (1.4.4). The raw observations are the reads aligned to the genome, from which surrogate statistics are constructed: read start counts k_t and fragment density c_t (see Figure1.10a and b). Starting from those two tracks we thus want infer two complimentary information:

- which regions are enriched in bound sequences (process $S_t^{(1)}$, described by two states, enriched or non enriched). The enrichment state influences directly fragment density.
- On which positions had crosslink taken place (process $S_t^{(2)}$ with two states, crosslinked or not crosslinked). In those regions, the amount of reads' 5' end is higher than expected, given the fragment density.

We thus start with observations that are the read start counts ($k_t \in \mathbb{N}$) and the pulled down fragment density ($c_t \in \mathbb{R}^+$) (Figure 1.10b). Note that conceptually, c_t is very similar to the expression level x_t

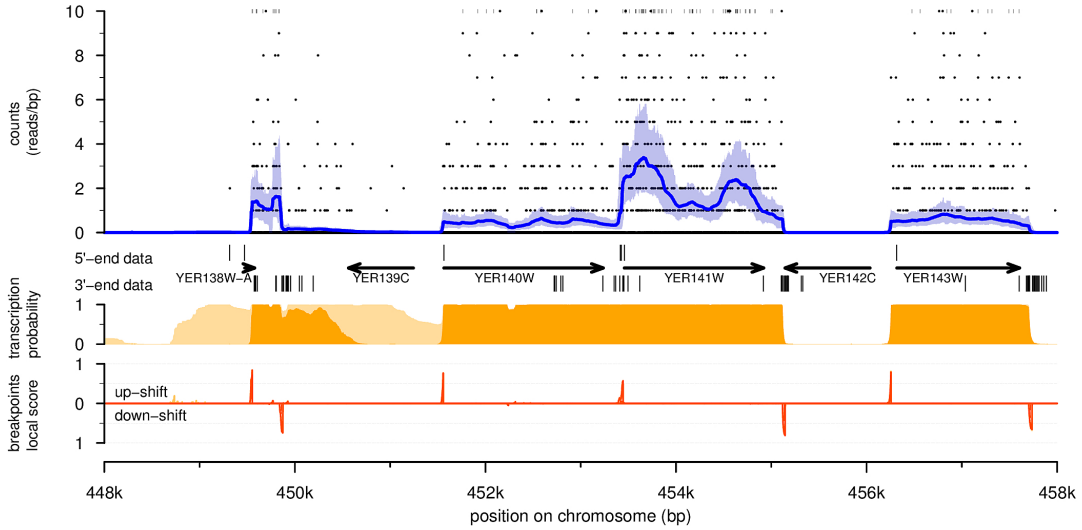


Figure 1.9: Transcriptional landscape reconstruction with Parseq, from (Mirauta, Nicolas, & Richard, 2014). Example of results on a 10 kbp region of the first strand of *S. cerevisiae* chromosome V (dataset SRR121907). From top to bottom: read counts (dots) and the estimated expression profile (blue line) with its 95% credibility interval (light blue area); annotated CDSs (arrows) complemented with specific data sets of 5'-ends and 3'-ends (brown); probability of transcription with a cut-off on expression level set to 0^+ (light orange) or 0.1 reads/bp (orange); Local score in high scoring segments for the detection of breakpoints associated with up-shifts and down-shifts (red). This example illustrates the detection of overlapping transcription units (up-shifts before YER140W and YER141W) and incomplete termination sites (down-shift after YER138W-A).

estimated with Parseq. However, to avoid overcomplicating the model this part was simplified and the values for c_t where estimated using kernel density estimation of the read counts. Indeed, the end goal of PureCLIP is to provide a binary annotation on the enriched/non enriched states and, as we will see below, to integrate other covariates for the detection of crosslink sites so that fragment density could be estimated with this method without significant loss.

We consider thus a two dimensional hidden process $\mathbf{x}_t = (S_t^{(1)}, S_t^{(2)})$ with a total of 4 possible states and which tracks the joint state of enrichment and crosslinking with a Markov model of order 1.

$$C_t \mid S_t^{(1)} = i \sim \Gamma_{\mathrm{LT}(\tau)}(\mu_i, \delta_i)$$
$$K_t \mid C_t = c_t, S_t^{(2)} = j \sim \mathcal{B}_{\mathrm{LT}(0)}(\hat{n}_t, p_j)$$

Where $i \in \{0, 1\}$ (resp $j \in \{0, 1\}$) is the subscript for enriched and non enriched (resp. cross-linked and not cross-linked). $\hat{n}_t$ is the expected number of fragment starts and is deduced from the fragment density c_t using a simple linear regression. $\Gamma_{\mathrm{LT}(\tau)}$ and $\mathcal{B}_{\mathrm{LT}(0)}$ are the truncated Gamma and Binomial distribution.

The HMM thus infers regions where the fragments are enriched (*e.g.* fragment density is higher than expected), and within those positions the crosslink sites where the reads start accumulate. By HMM standards, the model constructed is relatively simple, but its strength comes from the fact that we can integrate additional covariates, such as the presence of motifs or information on background binding (as described previously in section 1.2.3).

In the two models that were presented, the dispersion expected from read counts is an important

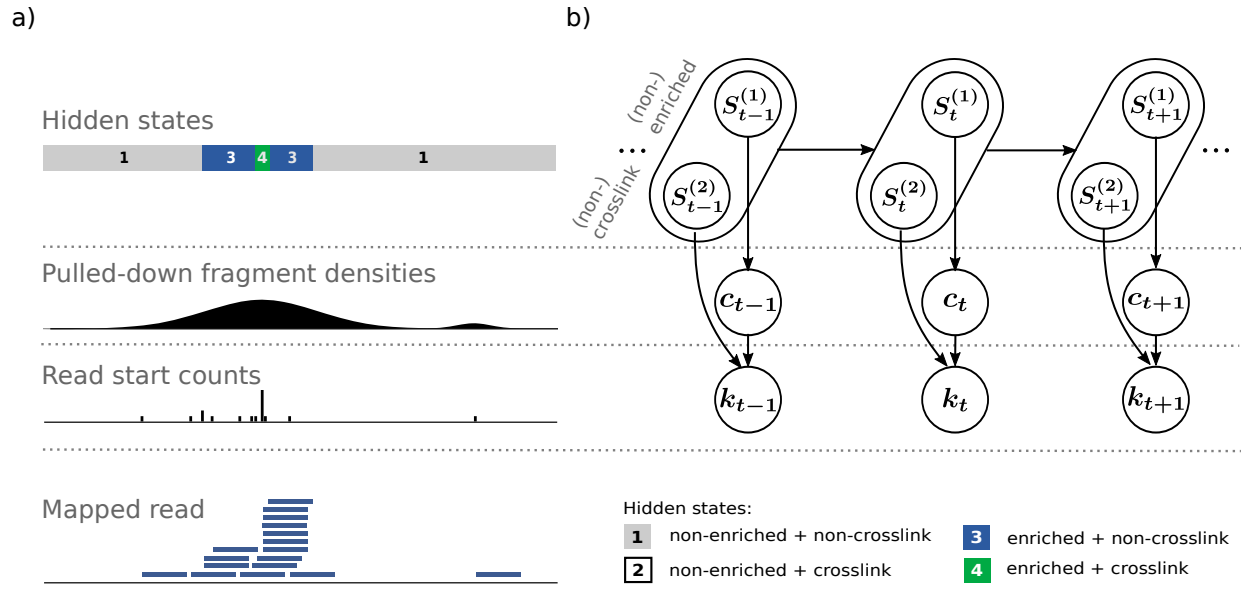


Figure 1.10: Summary of the modeling framework for analyzing RNA-protein interaction landscape. **a)** From (Krakau, Richard, & Marsico, 2017). Starting from the mapped reads (bottom), two signals that serve as observations in the HMM are derived for all nucleotide positions: individual read start counts and pulled-down fragment densities, obtained from smoothed read start counts. The model aims to reconstruct the most likely sequence of hidden states (top) from these signals. **b)** A graphical representation of the corresponding HMM.

factor in being able to make inference about the unobserved variables. Multiple distributions of counts can be proposed to account for overdispersion, and we provide a detailed presentation in the following section. Those models are also instrumental in estimating the terms of biological variance when testing for differential expression between conditions (I will not detail further on this part).

1.5.5 Count Models for Sequence Abundance

Most of the models presented in the last section relate to sequence counts for the observations. They thus need to specify an explicit emission distribution that will correspond to the count statistics (read depth, or read counts). These models are also at the core of other functional analysis problems, such as differential expression.

Simplest count model

To first explicitly account for the randomness coming from the sampling-by-sequencing process, we can start by hypothesizing that sampling across positions is uniform. Then the count Y_t obtained from the reads starting at any location t simply follows a binomial distribution $\mathcal{B}(n, p_t)$ (n being the total number of reads sampled):

$$\pi(y_t) = \binom{n}{y_t} p_t^{y_t} (1 - p_t)^{n - y_t}$$

p_t being the frequency at which we expect this position to be sampled (its proportion over all possible positions under an ideal scenario). We just mentioned in previous section a use of the binomial distribution in its truncated version for the detection of crosslink sites.

Often, the events considered are relatively rare (*e.g.* $p \ll 10^{-3}$) and n is at least 10^5 to 10^6, and the binomial can be safely approximated by a Poisson distribution $\mathcal{P}(\lambda)$ with rate $\lambda = n \cdot p_t$.

$$\pi(y_t) = e^{-\lambda} \cdot \frac{\lambda^{y_t}}{y_t!}$$

Poisson assumption alleviates the shortcomings of considering a Gaussian approximation for y_t, which has sometimes been proposed as an easy choice. Gaussian approximation is not justified by the digital type of the data, and it will fail to model low counts accurately (Audic & Claverie, 1997; Cai et al., 2004). However, the hypothesis of uniform sampling, which is at the heart of the Poisson distribution is too strict in most cases, as it implies that the variance of the count is equal to its rate λ. Multiples factors, such as biological variability, protocol related biases, or non uniform longitudinal sampling induce an implicit heterogeneity on the rate, such that the counts usually exhibit extra-variability (Anders & Huber, 2010).

Finite mixture models

Heterogeneity can be easily be integrated by means of mixture models. They can consist in finite mixture or in the continuous Gamma mixture model, also called Negative Binomial.

One example of finite mixture models arise naturally when considering the counts of k-mers Y, *e.g.* the substrings of the reads of a given length k. Under an ideal scenario where each k-mer is unique in the genome, and due to uniform sampling, the k-mer would ideally follow a Poisson distribution with expected count $\lambda = \frac{n \cdot p}{\ell - k + 1}$. However, sequencing errors are common and decrease the observed count of *bona-fide* k-mers, while creating artefactual ones. If we hypothesise a uniform error rate of ε at each base (thus a probability of $\varepsilon/3$ for each substitution), a k-mer with exactly i errors is distributed according to a Poisson distribution with an expectation $\mu_i = \lambda \cdot (\varepsilon/3)^i (1-\varepsilon)^{k-i}$ (we neglect cases where an other k-mer, at a short hamming distance, would contribute to the read depth). The distribution of Y is obtained by summing over the possible number of errors, which results in a mixture of Poisson distributions with rates μ_i, where the proportions are given by a binomial distribution $\mathcal{B}(\varepsilon, k)$. The mixture writes simply as:

$$\pi(y) = \sum_{i=0}^{k} \mathcal{B}(i; \varepsilon, k) \cdot \mathcal{P}(y; \mu_i)$$

Note that repeats in the genome, which we did not account for here, can also be considered as having an expected count that will increase in proportion of the number of copies that are present (*e.g.* $2\mu_i$ for 2 copies, etc.).

Contribution: detection of sequencing errors I utilised these mixture models in the problem of detecting reads possessing sequencing errors. Together with colleagues at the Free University in Berlin (David Weese, Manuel Holtgrewe) and at Carnegie Melon University in Pittsburgh (Marcel Schulz) we developed Fiona, an automatic tool for the correction of sequencing errors in HTS libraries (Schulz et al., 2014). Briefly, a suffix array data structure is constructed from the reads and cutoffs are tabulated from the genome properties to decide which k-mers will be identified as potential errors. The reads possessing errors are then processed sequentially for error correction (see Figure 1.11)

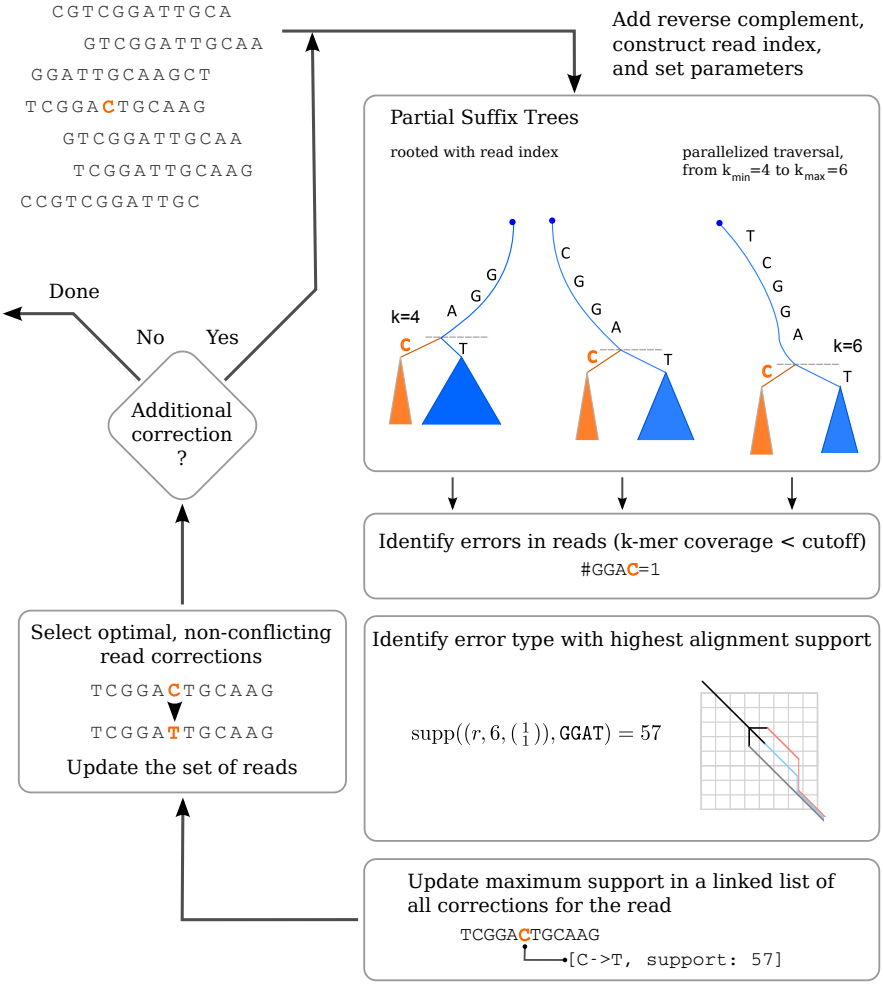


Figure 1.11: The Fiona strategy illustrated on a toy example from (Schulz et al., 2014). A set of partial suffix arrays are built from the set of reads and their reverse complement. The trees are traversed in parallel to detect and correct errors. Potential errors in the reads are identified as nodes in the tree according to their coverage (*e.g.* the substring `GGAC`, covered by only one read). The correction with the highest support is chosen to correct the read at that position. Due to the parallel traversal of the tree, all possible corrections on a read are recorded in a linked list, which reports the positions of corrections as well as their current maximal support. After traversal, the reads are updated by applying all non conflicting corrections in order of decreasing support. Once all reads have been corrected, the algorithm repeats the procedure until the number of corrections have been achieved.

The strategy implemented in Fiona combines the detection of candidate sequencing errors with a thoughtful implementation based on suffix arrays and banded alignment. I only mention the link to k-mer abundance distributions. Statistical error detection consists in determining a k-mer abundance value under which a string is deemed erroneous. This strategy was first proposed in the correction module of the EULER assembler, where least abundant substrings (usually singletons) were selected. More general methods, using alignment on the paths of a de Bruijn graph, were also developed (Salmela & Rivals, 2014). In a classical general modelling framework (David Weese, Schulz, & Richard, 2017), we aim to classify a k-mer has having either no-error ($z = 0$) or at least one error ($z > 0$). With Fiona, we used a log-odds ratio to classify k-mers:

$$\log \frac{P(Y_k = c \mid z > 0)}{P(Y_k = c \mid z = 0)} + w$$

The constant w impacts the proportion of erroneous reads detected. We can match the setup of a naive Bayes classifier and set it according to the prior probability of each category: $w = \log(1 - (1 - \varepsilon)^k)/(1 - \varepsilon)^k$. With Fiona, we first used the previous mixture model to get the explicit distribution of $Y_k \mid z$. Then, under a model of uniform read placement, the only parameters needed to determine the threshold value are the genome length n and the average error rate ε.

Contribution: Species abundance estimates using NPMLE Finite mixtures are also commonly used in ecology for answering general questions about species diversity. In a traditional ecology experiment, the species are sampled at a location and their abundance x_j –the number of times species j is observed, is recorded. However, as time and resource are limited, an exhaustive characterization of the biotope is not possible. The statistical questions are thus to estimate the species diversity. How many new species do we expect to discover if we sample twice as much? three times as much? What is the total number of species in the biotope? When did a species got extinct (D. L. Roberts & Solow, 2003)?

When I started my postdoc at the Max Planck for Molecular Genetics under Martin Vingron supervision, he asked me to look at the species diversity problem and adapt it to the analysis of HTS libraries. For whole genome sequencing (clonal population), the problem is relatively easy as one can fix upfront the amount of sequencing needed to reach a given coverage. In the case of transcriptomes or metagenomes, we are in the framework described for ecology: the species are replaced by the expressed genes or the microbial species, and we do not know how many are present in total –*e.g.* how complex the library is (Daley & Smith, 2013). When the first RNA-Seq experiments were made at the institute, one of the question was then to know whether the sequencing capacity was sufficient to detect most of the expressed genes (Sultan et al., 2008).

We consider that mapped reads are sampled independently and with replacement from the whole population of transcripts. The count x_j of gene j follows a Poisson distribution of parameter λ. The distribution of gene abundance can be summarised by describing the distribution over genes frequencies (the number of genes with count k):

$$y_k = \sum_{\text{genes } j} \mathbf{I}_{\{x_j = k\}}$$

S is the total number of different genes expressed obtained after sequencing ($S = \sum_{k>0} y_k$).

We see that the y_k are realisations of a finite mixture of Poisson distributions, whose rates depend

on the frequencies of all genes. We need to estimate the components of the mixture in order to predict y_0, the number of expressed gene that are still not sequenced. However, this problem can be impossible to solve, as there may be an arbitrary number of undetected species whose detection rate is lower than the sampling effort. The total number of species that can exist has to be constrained in some way, using regularization techniques such as the penalized likelihood (J.-P. Z. Wang & Lindsay, 2005).

Estimation of gene diversity is the same as estimating y_0 conditionally on S. Let us denote as f the marginal distribution over read counts, which can be written as:

$$f(c, Q) = \pi(X_j = k) = \int e^{-\lambda} \frac{\lambda^k}{k!} dQ(\lambda)$$

Where Q denotes the distribution of Poisson rates amongst the set of genes. We thus estimate the distribution over read counts $\hat{Q}$ using a Non Parametric Maximum Likelihood (NPMLE) approach (J.-P. Z. Wang & Lindsay, 2005). An estimator to y^0 can then be written as

$$y^0 = S \cdot \frac{f(0, \hat{Q})}{1 - f(0, \hat{Q})}$$

In the context of the RNA-Seq experiment, we could show that the extrapolation on single libraries was in line with the results obtained by pooling, and that the increase in sequencing power made possible by RNA-seq was able to recover most of the expressed genes (see Figure 1.12)

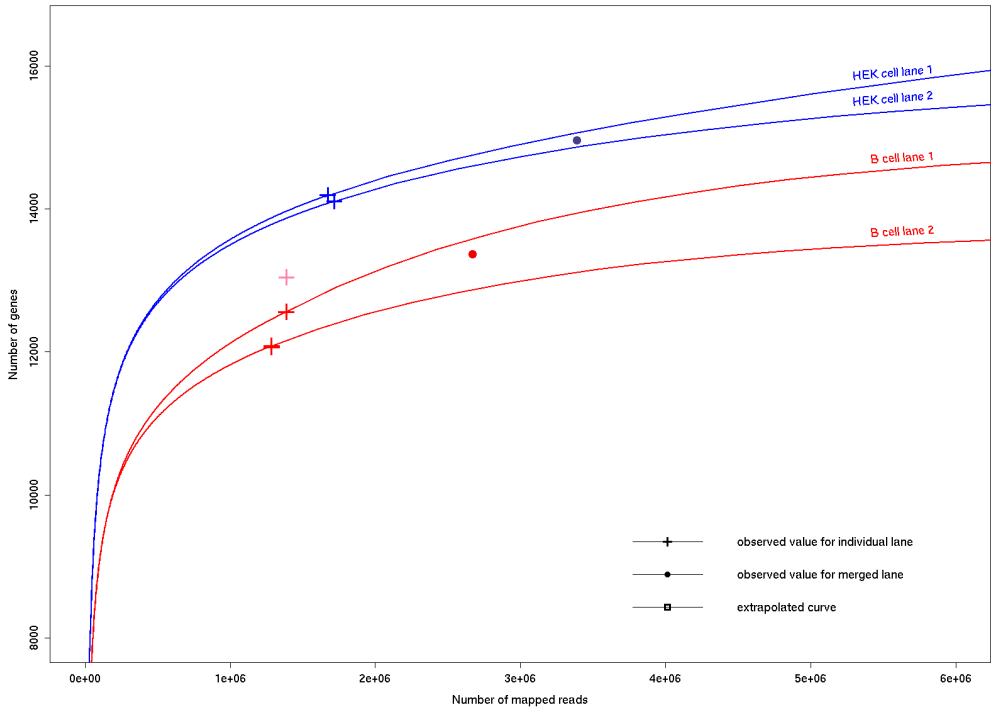


Figure 1.12: **Dynamic range of RNA-Seq**, from (Sultan et al., 2008). Rank abundance curves (RACs) showing the total number of mapped reads (x-axis) versus the total number of identified genes (y-axis). Data points show the observed values on individual (crosses) and merged (dots) experiments. Curves were extrapolated as follows: for values smaller or equal than the sample size, the number of expected genes was obtained by sub-sampling. For values greater than the sample size, the number of expected new genes was computed from the statistical analysis by Poisson mixtures.

The Gamma-Poisson mixture model

A more general extension of the Poisson distribution to account for overdispersion is the negative binomial model, that can be defined as a continuous mixture of Poisson distributions, where rates are distributed according to a Gamma distribution.

$$\lambda \sim \Gamma(a, 1/\kappa)$$
$$Y \sim \mathcal{P}(n \cdot \lambda)$$

resulting in a general distribution for the Negative Binomial as:

$$\mathbb{P}(Y = y) = \pi(y\ ;\ \kappa, n) = \frac{\Gamma(a+y)}{k!\,\Gamma(a)}\, p^y\, (1-p)^a$$

where p is a function of κ and n:

$$\frac{n}{\kappa} = \frac{1-p}{p}$$

and the mean and variance can be expressed as:

$$m = a \cdot \frac{\kappa}{n}$$
$$\sigma^2 = m + \frac{1}{a} \cdot m^2.$$

We note the model $\mathcal{NB}(m, \kappa)$ for a mean m and an overdispersion κ.
In practice this model establishes a quadratic relationship between the mean counts and their variance, while using only one more free parameter than the Poisson distribution. It remains quite flexible on the types of distribution for the rates. It has thus been a common choice for count distribution when extra variance is expected and that a finite mixture is not a realistic option. It was shown to provide a reasonable adjustment when looking at the distribution of read depth in genome sequencing (Bentley et al., 2008). It is also commonly used to model the variability in count between biological replicates for instance for detection differential expression in RNA-Seq experiments (Anders & Huber, 2010; Love et al., 2014; M. D. Robinson et al., 2010).

In the case of the PureCLIP Tool, we used a truncated version of the gamma function for the parameters describing the expression level e_t:

$$k_{\text{emission}}(e_t \mid s_t^{(1)}) \sim \text{trunc}\Gamma(\alpha_{s_t^{(1)}}, \beta_{s_t^{(1)}})$$

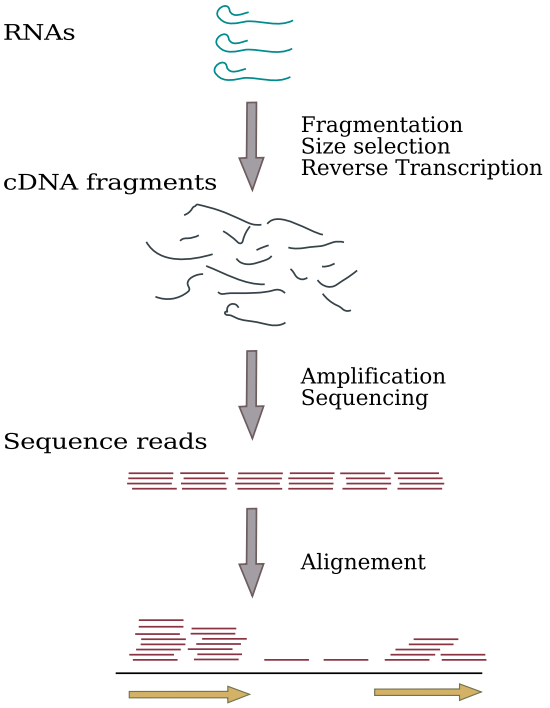
the other observation, the count of reads 5' end, is then conditionally drawn according to a binomial distribution

$$k_{\text{crosslink}}(c_t \mid e_t, s_t^{(1)}, s_t^{(2)}) \sim \mathcal{B}(\hat{n}(e_t), p_{s_t^{(2)}})$$

Note that in this model, the observations are two dimensional, they combine a coverage term e_t, and a crosslinking term c_t. The coverage e_t is estimated using a smoothing window over the read depth.

Contribution: A hierarchical model with extra-variance

For the analysis of longitudinal RNA-Seq data, at the core of the Parseq tool, the Negative Binomial failed to account for the relationship between mean and variance, as well as the proportion of regions with no counts. We thus proposed a more general hierarchical model that mimics mechanistically the different steps of the library preparation (Mirauta et al., 2014).



f_t: number of initial cDNA fragments
x_t: expression level
$f_t \sim \mathcal{P}(\frac{x_t}{a})$

a_t: amplification of mean a
$a_t \sim \Gamma(a, \kappa)$

y_t counts of 5' ends at position t
$y_t \sim \mathcal{P}(f_t \cdot a_t)$
$\mathbb{E}(y_t \mid x_t) = x_t$

Figure 1.13: left: canonical steps of the RNA-Seq protocol, including: i) RNA fragmentation and Reverse Transcription; ii) cDNA amplification and iii) final read sampling by sequencing. On the right, alongside the protocol steps, variables that model quantities are represented.

We aim to explicitly account for three steps of the protocols (see also 1.2.1): (i) initial molecule sampling and fragmentation, (ii) amplification, and (iii) final sampling by sequencing. To do so, we introduce auxiliary variables corresponding to the number of fragments sampled by the experiment f_t, and to an amplification term a_t. Given an expression level x_t for a region (which would be $n \cdot p_t$ in the uniform case), we have:

$$f_t \mid x_t \sim \mathcal{P}(\frac{x_t}{a})$$

The amplification is hypothesized to take place with global rate

$$a_t \sim \Gamma(a, \kappa)$$

the number of reads aligned at t are then sampled according to a Poisson distribution at a rate given by the number of amplified fragments

$$y_t \mid f_t, a_t \sim \mathcal{P}(f_t \cdot a_t)$$

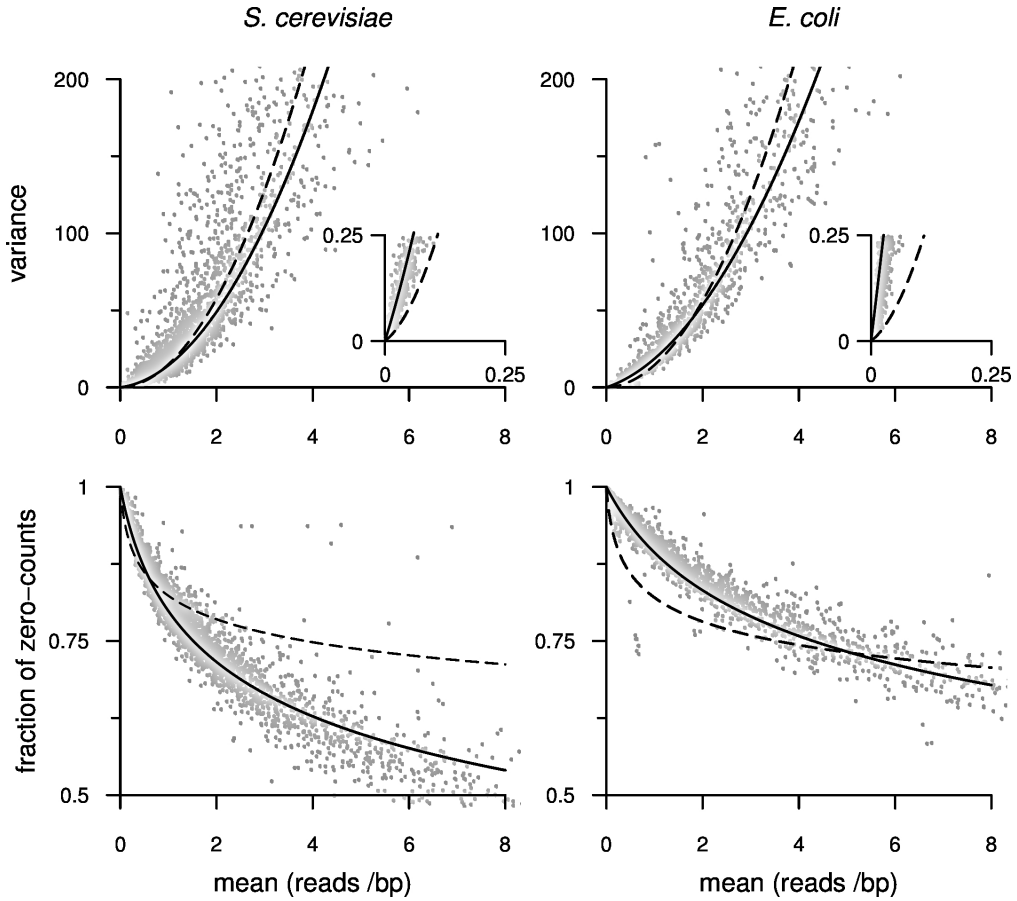


Figure 1.14: Distribution of read counts inside regions of homogeneous expression, from (Mirauta, Nicolas, & Richard, 2014). *Saccharomyces cerevisiae* data-set SRR121907 (left); *Escherichia coli* data-set SRR794838 (right). Each long open-reading frame (ORF, region without in-frame stop codon) identified on the genome is represented by a dot. Dashed lines show the fit of the negative binomial model with overdispersion parameter estimated via variance (reads^2/bp^2) versus mean (reads/bp) regression; plain lines show the fit with the Parseq model.

we can write the distribution of y_t given x_t as a mixture of negative binomials, whose constributions are poisson distributed:

$$y_t \mid x_t \sim \sum_{f_t=0}^{\infty} \mathcal{P}(f_t; \frac{x_t}{a}) \cdot \mathcal{NB}(y_t; f_t \cdot a, \kappa)$$

In practice we also add a term accounting for local bias s_t and that occurs prior to fragment selection (*e.g.* $f_t \sim \mathcal{P}(x_t/a \cdot s_t)$)

When controlling the adjustment of this model on preliminary RNA-Seq data we see that the negative binomial fails to account for extra variability when the read counts are low (Figure 1.14, insets). Furthermore the fraction of positions with no read starts is as well better captured by the mechanistic model.

Use of covariates to explain read count

In addition to the estimation of a broad library-wide variance term, it is possible to inspect specific factors that are responsible for the skew in read counts. Indeed, various steps of library preparation induce biases (section 1.2.3), that can be caused by sequence content or other biological properties.

By making some assumption on the shape of the data –within a single isoform gene, the transcript abundance is expected to be constant- one can tabulate the observed read counts with other sources of information, such as sequence content, or position in the gene. Then, linear models are used to explain the part of the variability of the counts within the gene bodies that can be explained by those covariates. Multiples methods have been developped, for instance (J. Li et al., 2010) proposed a linear model to estimate the effect of position specific sequence content around the position where the sequence is mapping:

$$\log y_{gt} = \log y_g + \sum_{k \in -\ell,\ldots,\ell} \sum_{\sigma \in \{A,C,G\}} \xi_\sigma^k \mathbf{I}_{\{\sigma_{t+k}=\sigma\}}$$

Here, the observed read count at position t, y_{gt} can deviate from the gene baseline y_g, according to the nucleotide composition ℓ bp around the 5' end of the read. The variables ξ_σ^k weight the contribution of the nucleotide σ k bp apart from the 5' of the read (with ξ_T^k set to 0). Those covariates can act on the values of the parameters through various link function. This was proposed for correction of abundance based on the GC-content of the fragment (Benjamini & Speed, 2012) or also in the case of RNA-Seq data for accounting for different sequence contexts (J. Li et al., 2010; Love et al., 2014; A. Roberts, Trapnell, Donaghey, Rinn, & Pachter, 2011).

Contribution: integration of input data and motifs for CLIP data For the PureCLIP model, Sabrina Krakau integrated two types of informations as covariates in the model: the signal from the input experiment and non specific binding motifs (called CL motifs, see section 1.2.3). The relationships in both cases are learned using generalized linear models (GLM) that link either the expected coverage c_t to the observed value of the input with a Gamma-GLM, or the probability of cross linking $p_{j,t}$ to the occurence of specific sequence motifs.

At each position t, the expected mean $\mu_{i,t}$ is supposed to be multiplicatively proportional to the background signal b_t (the shape parameter of each Gamma distribution is supposed to be constant):

$$\ln(\mu_{i,t}) = \alpha_{i,0} + \alpha_{i,1} b_t.$$

This way, we can make use of the input data to account for regions highly transcribed or bound by highly abundant background proteins.

The probability of cross-linking is adjusted using a (zero-truncated) binomial logistic regression on a set of learned motifs:

$$\ln \frac{p_{j,t}}{1 - p_{j,t}} = \beta_{j,0} + \beta_{j,m_t} x_{m_t,t},$$

where m_t is the best scoring motif observed at position t. The set of motifs are learned calling crosslink site on the input data using the basic version of PureCLIP (without covariates). Motifs are then detected around the detected position. The resulting set of motifs provides a picture of the fragments that are commonly crosslinked to RNA background proteins.

1.6 Analysing the small RNAs populations in *P. tricornutum*

This project, where I put into practice each and every steps of the sequence analysis workflow by myself, provides a good last section for this chapter. Although I was working on sequence analysis for

a while, my main concern before this project was the development of methods and their validation, and were taking place as part of a large working group. Until then I never had performed all the analysis by myself.

The project was initiated by the Diatom Genomics group of Angela Falciatore at LCQB (group now at IBPC) and we tag teamed with Alessandra Rogato (postdoc in Angela's lab, then researcher at Stazione Dohm, Naples). I would do the data analysis and Alessandra would design and perform experimental validations (verifying that the elements identified were present, as well as functional characterisations). Until now, I presented a subdiscipline that we could call *bioinformatics-as-a-method*: the aim is to provide a new general solution to a particular analysis problem. Here, no new methods had to be developed, but I rather carefully selected, validated and reiterated classical bioinformatics and data analysis in order to propose new biological hypotheses: *bioinformatics-as-a-result*, or quantitative biology (to mention the lab's name). It is also different from the development of an analysis workflow (*bioinformatics-as-a-process*), as the new object were uncharacterized in this species and each analysis was truly exploratory.

The goal was to follow up on a first discovery by the group about silencing in *Phaeodactylum tricornutum*, a marine diatom. Diatoms are unicellular eukaryotes and present a valuable model system to address questions about the evolution and diversification of gene regulatory mechanisms in eukaryotes. The Diatom genomics group had already shown that the expression of anti-sense or inverted repeat sequences of selected target genes can trigger efficient gene silencing in *P. tricornutum* and they described as well the presence of genes encoding a predicted Dicer-like protein, an Argonaute-like protein and a potential RNA-dependent RNA polymerase (RdRP) (De Riso et al., 2009), which are all proteins involved in the RNA silencing pathways.

To gain insight into the population of small RNAs in diatoms, library of short RNA fragments were sequenced in different light and iron conditions. We then assessed which types of small RNAs populations could be present by first looking at size distribution (see Figure 1.15 top) and the small RNA localisation. We observed 3 characteristic lengths that corresponded to:

1. a very abundant RNA population of 24-25 nt in length originating from intergenic regions and, in particular, from a unique 80 bp region on chromosome 2 (Figure 1.15A, "Others" inset),
2. tRNA fragments mainly enriched in the 19-20 nt fraction and, to a minor extent, in the 30-33 nt fraction (Figure 1.15A, "Known ncRNAs" inset),
3. repeat and coding regions enriched in 25-30 nt fragments (Figure 1.15B).

We then analysed the aligned reads and observed two types of behaviours. On the one hand, some reads match very specific genome locations and form, after alignment, characteristic piles of thousands of copies of the same sequence, accumulated on a single strand (panel A). On the other hand, there are regions, of a few thousand bases in length, that are covered by overlapping reads, accumulated on both strands and, at times, forming several piles distributed with a periodic pattern (panel B).

For the first distribution type, after a strict filtering step eliminating possible artifacts, we characterised 50 candidate regions that appeared to be specific to sRNAs lying in tRNAs and in intergenic regions. These are organised in three groups. Thirty regions correspond to already annotated non-coding RNA structures, including a highly abundant candidate (representing 0.4% of all aligned reads) that overlaps the U2 snRNA gene located on chromosome 5, and the 5S ribosomal RNA on chromosome 3. The majority of the candidate regions overlap with 28 different tRNA sequences (corresponding to 22 different codons), suggesting that they may represent tRNA-related small RNAs

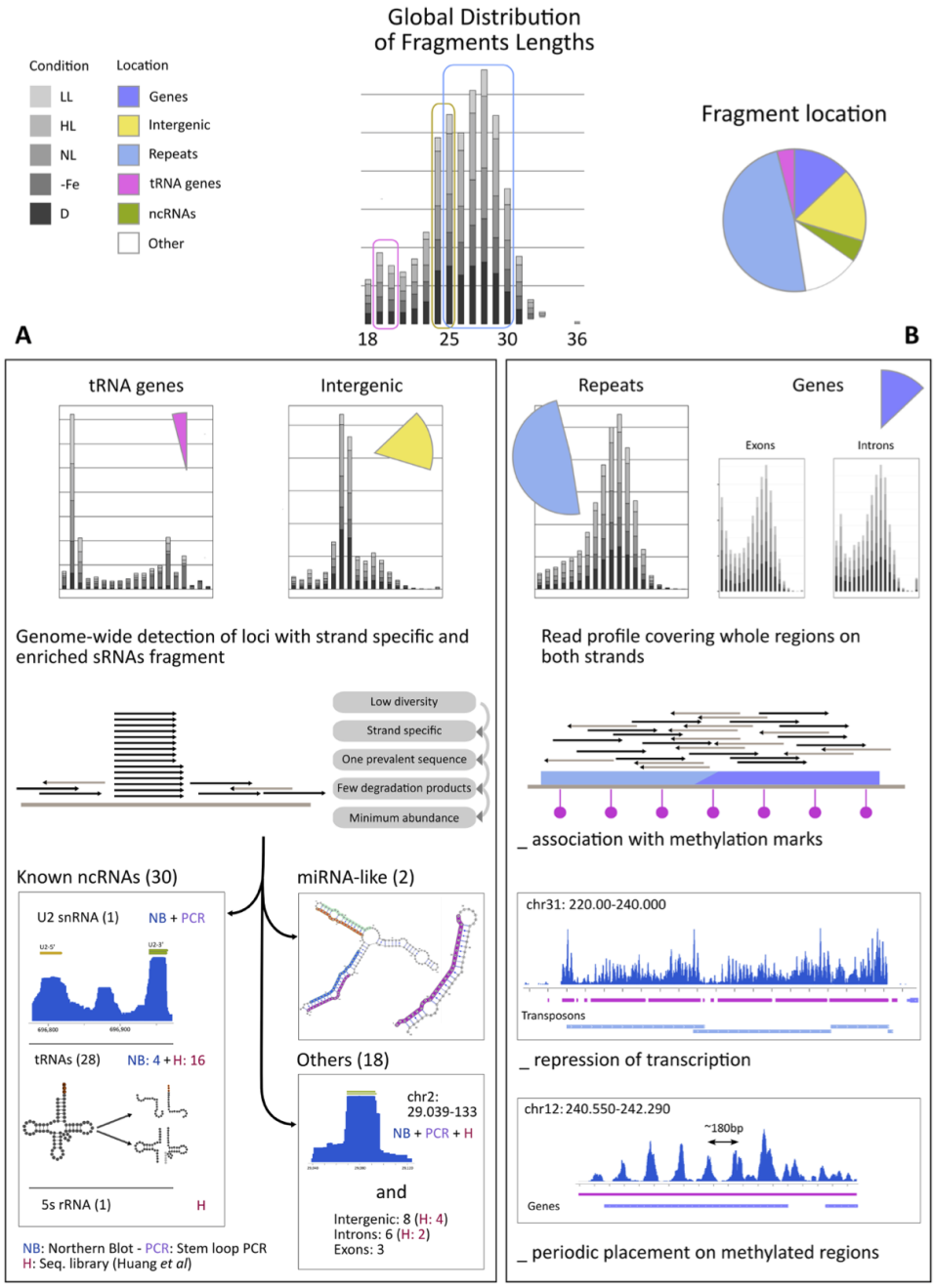


Figure 1.15: Workflow of the small RNAs analysis in *P. tricornutum*, from (Rogato et al., 2014). Top. Fragment lengths distribution of reads (histogram, center) is reported in a grey color scale distinguishing the five experimental conditions (LL, HL, NL, -Fe, D). The distribution of fragment location is also reported (pie chart, right) with a color scale indicating genes, intergenic regions, repeat regions, tRNA genes, ncRNAs and other loci. We distinguish two workflows described in boxes A and B, characterized by different local loci distributions of reads along the genome. (A) Sequence specific distribution of fragment lengths that is systematically observed for tRNA genes and intergenic regions. Reads were filtered in five steps, described in the 5 grey boxes. We obtained three main groups of results, indicated by squared boxes (number of predicted sRNAs is reported in parenthesis). The number of predicted sRNAs that were experimentally validated is also indicated, together with the experimental technique (NB, Northern Blot; PCR, Stem Loop PCR; H, sequencing data from (Huang, He, & Wang, 2011)). (B) Distribution of fragment lengths that covers loci with overlapping reads and accumulated on both strands. This distribution pattern has been observed to either Tranposable Elements (TEs) or coding genes, associated to methylation. Examples of the periodic placement of sRNAs on three Codi LTR-retrotransposons on chromosome 31 and on a protein coding gene on chromosome 12 are reported. Color palette for TEs and genes is the same as above, and Highly Methylated regions are represented in purple

(tRFs). The remaining 20 candidates do not overlap with regions related to non coding RNAs. Two of them resemble miRNA-like molecules, supported by a stable precursor structure, and 18 of them have no particular associated structure. These predictions were well supported across different libraries (28 regions detected in at least two libraries) and on an independent sequencing experiment published by Huang and coworkers (Huang, He, & Wang, 2011) that we integrated in our pipeline (24 regions are also detected). We were able to experimentaly validate the presence of a few of those candidates regions by Northern Blot or Stem Loop PCR (see annotationss in Figure 1.15, panel A).

The second kind of sRNAs distribution appeared to be specific to transposable elements and to coding genes (Figure 1.15B). Three major observations were made: 1. reads overlap over relatively long regions that are typically methylated, 2. the sRNAs accumulation correlates with repression of transcription, and 3. the sRNA profile displays periodic patterns at a distance varying within the 180-200 nt interval.

The characterization of the sRNA transcriptomes revealed a small ncRNA landscape in diatoms that is much more complex than anticipated. We identified and characterized different functional categories of small RNAs with different sizes, suggesting the presence of distinct biosynthetic pathways. Based on the sequence data available, my analysis highlighted that 93.5% of sequence specific reads can be explained by their accumulation in well classified loci covering 8% of the whole genome.

Even though this first analysis provided a very comprehensive characterisation of the *P. tricornutum* small RNAs, we did not validated or showed a functional effect. It is unfortunately often the case with such kind of analyses (Hul et al., 2020). I recently had the opportunity to participate in a follow up project analysing the small RNAs and the transcriptional response in different Dicer Knockout lines. Generating the KO lines was a work done by Emilya Gripioti during her PhD thesis, under the supervision Frederic Verret (Hellenic Center for Marine Research, Heraklion) and Kriton Katanlidis (Institute of Molecular Biology and Biotechnology, Heraklion). The main results from this analysis were submitted as an article and are now in revision.

References

't Hoen, P. A. C., Friedlander, M. R., Almlof, J., Sammeth, M., Pulyakhina, I., Anvar, S. Y., ... Lappalainen, T. (2013). Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories. *Nat Biotechnol*, *31*(11), 1015–1022. doi:10.1038/nbt.2702

Alamancos, G. P., Agirre, E., & Eyras, E. (2014). Methods to study splicing from high-throughput RNA sequencing data. *Methods Mol Biol*, *1126*, 357–397. doi:10.1007/978-1-62703-980-2{_}26

Allhoff, M., Schonhuth, A., Martin, M., Costa, I. G., Rahmann, S., & Marschall, T. (2013). Discovering motifs that induce sequencing errors. *BMC Bioinformatics*, *14 Suppl 5*, S1. doi:10.1186/1471-2105-14-S5-S1

Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, *11*(10), R106. doi:10.1186/gb-2010-11-10-r106

Andrews, S. (n.d.). FastQC A Quality Control tool for High Throughput Sequence Data. Retrieved 2012, from http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Audic, S., & Claverie, J. M. (1997). The significance of digital gene expression profiles. *Genome Res*, *7*(10), 986–995.

Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D. E., Wang, Z., ... Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. *Cell*, *129*(4), 823–837. doi:10.1016/j.cell.2007.05.009

Benjamini, Y., & Speed, T. P. (2012). Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res*, *40*(10), e72. doi:10.1093/nar/gks001

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., ... Smith, A. J. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, *456*(7218), 53–59. doi:10.1038/nature07517

Beretta, S., Bonizzoni, P., Vedova, G. D., Pirola, Y., & Rizzi, R. (2014). Modeling alternative splicing variants from rna-seq data with isoform graphs. *Journal of Computational Biology*, *21*(1), 16–40. PMID: 24200390. doi:10.1089/cmb.2013.0112. eprint: https://doi.org/10.1089/cmb.2013.0112

Berlin, K., Koren, S., Chin, C.-S., Drake, J. P., Landolin, J. M., & Phillippy, A. M. (2015). Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nat Biotechnol*, *33*(6), 623–630. doi:10.1038/nbt.3238

Bernard, E., Jacob, L., Mairal, J., & Vert, J.-P. (2014). Efficient RNA isoform identification and quantification from RNA-Seq data with network flows. *Bioinformatics*, *30*(17), 2447–2455. doi:10.1093/bioinformatics/btu317

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature biotechnology*, *34*(5), 525–527.

Buchfink, B., Xie, C., & Huson, D. H. (2014). Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, *12*, 59 EP -.

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., & Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods*, *10*(12), 1213–1218. doi:10.1038/nmeth.2688

Burge, C., & Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology*, *268*(1), 78–94. doi:https://doi.org/10.1006/jmbi.1997.0951

Cai, L., Huang, H., Blackshaw, S., Liu, J. S., Cepko, C., & Wong, W. H. (2004). Clustering analysis of SAGE data using a Poisson approach. *Genome Biol*, *5*(7), R51. doi:10.1186/gb-2004-5-7-r51

Cameron, D. L., Schröder, J., Penington, J. S., Do, H., Molania, R., Dobrovic, A., ... Papenfuss, A. T. (2017). GRIDSS: sensitive and specific genomic rearrangement detection using positional de Bruijn graph assembly. *Genome Research*. doi:10.1101/gr.222109.117. eprint: http://genome.cshlp.org/content/early/2017/11/02/gr.222109.117.full.pdf+html

Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., ... Hayashizaki, Y. (2006). Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet*, *38*(6), 626–635. doi:10.1038/ng1789

Chiu, C. Y., & Miller, S. A. (2019). Clinical metagenomics. *Nat Rev Genet*, *20*(6), 341–355. doi:10.1038/s41576-019-0113-7

Chung, H.-R., Dunkel, I., Heise, F., Linke, C., Krobitsch, S., Ehrenhofer-Murray, A. E., ... Vingron, M. (2010). The effect of micrococcal nuclease digestion on nucleosome positioning data. *PLoS One*, *5*(12), e15754. doi:10.1371/journal.pone.0015754

Computational Pan-Genomics, C. (2018). Computational pan-genomics: Status, promises and challenges. *Brief Bioinform*, *19*(1), 118–135. doi:10.1093/bib/bbw089

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., ... Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology*, *17*(1), 13. doi:10.1186/s13059-016-0881-8

Consortium, T. F., the RIKEN PMI, (DGT), C., Forrest, A. R. R., Kawaji, H., Rehli, M., ... Hayashizaki, Y. (2014). A promoter-level mammalian expression atlas. *Nature*, *507*, 462 EP -.

Daley, T., & Smith, A. D. (2013). Predicting the molecular complexity of sequencing libraries. *Nat Methods*, *10*(4), 325–327. doi:10.1038/nmeth.2375

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., ... Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158. doi:10.1093/bioinformatics/btr330

David, L., Vicedomini, R., Richard, H., & Carbone, A. (2020). Targeted domain assembly for fast functional profiling of metagenomic datasets with S3A. *Bioinformatics*, *36*(13), 3975–3981. doi:10.1093/bioinformatics/btaa272. eprint: https://academic.oup.com/bioinformatics/article-pdf/36/13/3975/33459041/btaa272.pdf

De Riso, V., Raniello, R., Maumus, F., Rogato, A., Bowler, C., & Falciatore, A. (2009). Gene silencing in the marine diatom Phaeodactylum tricornutum. *Nucleic Acids Research*, *37*(14), e96–e96. doi:10.1093/nar/gkp448. eprint: https://academic.oup.com/nar/article-pdf/37/14/e96/16752973/gkp448.pdf

de Rie, D., Abugessaisa, I., Alam, T., Arner, E., Arner, P., Ashoor, H., ... de Hoon, M. J. L. (2017). An integrated expression atlas of miRNAs and their promoters in human and mouse. *Nature Biotechnology*, *35*, 872 EP -.

Denti, L., Rizzi, R., Beretta, S., Vedova, G. D., Previtali, M., & Bonizzoni, P. (2018). Asgal: Aligning rna-seq data to a splicing graph to detect novel alternative splicing events. *BMC Bioinformatics*, *19*(1), 444. doi:10.1186/s12859-018-2436-3

Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., ... Gingeras, T. R. (2012). Landscape of transcription in human cells. *Nature*, *489*(7414), 101–108. doi:10.1038/nature11233

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., ... Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-Seq aligner. *Bioinformatics*, *29*(1), 15–21. doi:10.1093/bioinformatics/bts635

Dohm, J. C., Lottaz, C., Borodina, T., & Himmelbauer, H. (2008). Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res*, *36*(16), e105. doi:10.1093/nar/gkn425

Doucet, A., & Johansen, A. M. (2009). A tutorial on particle filtering and smoothing: Fifteen years later. *The Oxford Handbook of Nonlinear Filtering*, 656–704.

Durbin, R., Eddy, S. R., Krogh, A., & Mitchison, G. (1998). *Biological sequence analysis: Probabilistic models of proteins and nucleic acids*. doi:10.1017/CBO9780511790492

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., ... Turner, S. (2009). Real-time DNA sequencing from single polymerase molecules. *Science*, *323*(5910), 133–138. doi:10.1126/science.1162986

ENCODE Project, C. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, *489*(7414), 57–74. doi:10.1038/nature11247

English, A. C., Salerno, W. J., Hampton, O. A., Gonzaga-Jauregui, C., Ambreth, S., Ritter, D. I., ... Gibbs, R. A. (2015). Assessing structural variation in a personal genome—towards a human reference diploid genome. *BMC Genomics*, *16*(1), 286. doi:10.1186/s12864-015-1479-3

Ernst, J., Kheradpour, P., Mikkelsen, T. S., Shoresh, N., Ward, L. D., & Epstein, C. B. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, *473*. doi:10.1038/nature09906

Esplin, E. D., Oei, L., & Snyder, M. P. (2014). Personalized sequencing and the future of medicine: Discovery, diagnosis and defeat of disease. *Pharmacogenomics*, *15*(14), 1771–1790. doi:10.2217/pgs.14.117

Ewels, P., Magnusson, M., Lundin, S., & Kaller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, *32*(19), 3047–3048. doi:10.1093/bioinformatics/btw354

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., ... Merrick, J. M. (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae rd. *Science*, *269*(5223), 496–512.

Friedersdorf, M. B., & Keene, J. D. (2014). Advancing the functional utility of PAR-CLIP by quantifying background binding to mRNAs and lncRNAs. *Genome Biol*, *15*(1), R2. doi:10.1186/gb-2014-15-1-r2

Frith, M. C. (2019). How sequence alignment scores correspond to probability models. *Bioinformatics*, *36*(2), 408–415. doi:10.1093/bioinformatics/btz576. eprint: https://academic.oup.com/bioinformatics/article-pdf/36/2/408/31962819/btz576.pdf

Furey, T. S. (2012). ChIP-seq and beyond: New and improved methodologies to detect and characterize protein-DNA interactions. *Nat Rev Genet*, *13*(12), 840–852. doi:10.1038/nrg3306

Gao, Y., Liu, Y., Ma, Y., Liu, B., Wang, Y., & Xing, Y. (2020). abPOA: an SIMD-based C library for fast partial order alignment using adaptive band. *Bioinformatics*, *37*(15), 2209–2211. doi:10.1093/bioinformatics/btaa963. eprint: https://academic.oup.com/bioinformatics/article-pdf/37/15/2209/40080650/btaa963.pdf

Garrison, E., & Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. *arXiv e-prints*, arXiv:1207.3907. arXiv: 1207.3907 [q-bio.GN]

Ghildiyal, M., & Zamore, P. D. (2009). Small silencing RNAs: An expanding universe. *Nat Rev Genet*, *10*(2), 94–108. doi:10.1038/nrg2504

Gillet-Markowska, A., Richard, H., Fischer, G., & Lafontaine, I. (2015). Ulysses: accurate detection of low-frequency structural variations in large insert-size sequencing libraries. *Bioinformatics*, *31*(6), 801–808. doi:10.1093/bioinformatics/btu730. eprint: http://bioinformatics.oxfordjournals.org/content/31/6/801.full.pdf+html

Glenn, T. C. (2011). Field guide to next-generation dna sequencers. *Molecular Ecology Resources*, *11*(5), 759–769. doi:10.1111/j.1755-0998.2011.03024.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1755-0998.2011.03024.x

Gotoh, O. (1982). An improved algorithm for matching biological sequences. *Journal of Molecular Biology*, *162*(3), 705–708. doi:https://doi.org/10.1016/0022-2836(82)90398-9

Griebel, T., Zacher, B., Ribeca, P., Raineri, E., Lacroix, V., Guigo, R., & Sammeth, M. (2012). Modelling and simulating generic RNA-Seq experiments with the flux simulator. *Nucleic Acids Res*, *40*(20), 10073–10083. doi:10.1093/nar/gks666

Gusmao, E. G., Allhoff, M., Zenke, M., & Costa, I. G. (2016). Analysis of computational footprinting methods for DNase sequencing experiments. *Nature Methods*, *13*, 303 EP -.

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., ... Regev, A. (2013). De novo transcript sequence reconstruction from rna-seq using the trinity platform for reference generation and analysis. *Nature Protocols*, *8*(8), 1494–1512. doi:10.1038/nprot.2013.084

Haberman, N., Huppertz, I., Attig, J., Konig, J., Wang, Z., Hauer, C., ... Ule, J. (2017). Insights into the design and interpretation of iCLIP experiments. *Genome Biol*, *18*(1), 7. doi:10.1186/s13059-016-1130-x

Hansen, K. D., Brenner, S. E., & Dudoit, S. (2010). Biases in illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Res*, *38*(12), e131. doi:10.1093/nar/gkq224

Hansen, T. B., Jensen, T. I., Clausen, B. H., Bramsen, J. B., Finsen, B., Damgaard, C. K., & Kjems, J. (2013). Natural rna circles function as efficient microRNA sponges. *Nature*, *495*, 384 EP -.

Heinig, M., Petretto, E., Wallace, C., Bottolo, L., Rotival, M., Lu, H., ... Cook, S. A. (2010). A trans-acting locus regulates an anti-viral expression network and type 1 diabetes risk. *Nature*, *467*, 460 EP -.

Hesselberth, J. R., Chen, X., Zhang, Z., Sabo, P. J., Sandstrom, R., Reynolds, A. P., ... Stamatoyannopoulos, J. A. (2009). Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nat Methods*, *6*(4), 283–289. doi:10.1038/nmeth.1313

Hon, C.-C., Ramilowski, J. A., Harshbarger, J., Bertin, N., Rackham, O. J. L., Gough, J., ... Forrest, A. R. R. (2017). An atlas of human long non-coding RNAs with accurate 5' ends. *Nature*, *543*, 199 EP -.

Huang, A., He, L., & Wang, G. (2011). Identification and characterization of microRNAs from Phaeodactylum tricornutum by high-throughput sequencing and bioinformatics analysis. *BMC Genomics*, *12*(1), 337. doi:10.1186/1471-2164-12-337

Hul, M. V., Roy, T. L., Prifti, E., Dao, M. C., Paquot, A., Zucker, J.-D., ... Cani, P. D. (2020). From correlation to causality: The case of subdoligranulum. *Gut Microbes*, *12*(1), 1849998. PMID: 33323004. doi:10.1080/19490976.2020.1849998. eprint: https://doi.org/10.1080/19490976.2020.1849998

International Human Genome Sequencing Consortium. (2004). Finishing the euchromatic sequence of the human genome. *Nature*, *431*(7011), 931–45. doi:10.1038/nature03001

Iwata, H., & Gotoh, O. (2012). Benchmarking spliced alignment programs including Spaln2, an extended version of Spaln that incorporates additional species-specific features. *Nucleic Acids Res*, *40*(20), e161. doi:10.1093/nar/gks708

Jain, M., Koren, S., Miga, K. H., Quick, J., Rand, A. C., Sasani, T. A., ... Loose, M. (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology*, *36*, 338 EP -.

Jean, G., Kahles, A., Sreedharan, V. T., De Bona, F., & Ratsch, G. (2010). RNA-Seq read alignments with PALMapper. *Curr Protoc Bioinformatics*, *Chapter 11*, Unit 11.6. doi:10.1002/0471250953.bi1106s32

Kanamori-Katayama, M., Itoh, M., Kawaji, H., Lassmann, T., Katayama, S., Kojima, M., ... Hayashizaki, Y. (2011). Unamplified cap analysis of gene expression on a single-molecule sequencer. *Genome Research*, *21*(7), 1150–1159. doi:10.1101/gr.115469.110. eprint: http://genome.cshlp.org/content/21/7/1150.full.pdf+html

Karlić, R., Chung, H.-R., Lasserre, J., Vlahoviček, K., & Vingron, M. (2010). Histone modification levels are predictive for gene expression. *Proceedings of the National Academy of Sciences*, *107*(7), 2926–2931. doi:10.1073/pnas.0909344107. eprint: https://www.pnas.org/content/107/7/2926.full.pdf

Katz, Y., Wang, E. T., Airoldi, E. M., & Burge, C. B. (2010). Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature Methods*, *7*, 1009 EP -.

Kent, W. J. [W J], Zweig, A. S., Barber, G., Hinrichs, A. S., & Karolchik, D. (2010). BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics*, *26*(17), 2204–2207. doi:10.1093/bioinformatics/btq351

Kent, W. J. [W James], Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., & Haussler, D. (2002). The human genome browser at UCSC. *Genome Res*, *12*(6), 996–1006. doi:10.1101/gr.229102

Kodama, Y., Shumway, M., & Leinonen, R. (2012). The Sequence Read Archive: explosive growth of sequencing data. *Nucleic Acids Res*, *40*(Database issue), D54–6. doi:10.1093/nar/gkr854

Koscielny, G., Texier, V. L., Gopalakrishnan, C., Kumanduri, V., Riethoven, J.-J., Nardone, F., ... Gautheret, D. (2009). Astd: The alternative splicing and transcript diversity database. *Genomics*, *93*(3), 213–220. doi:https://doi.org/10.1016/j.ygeno.2008.11.003

Krakau, S., Richard, H., & Marsico, A. (2017). PureCLIP: Capturing target-specific protein-RNA interaction footprints from single-nucleotide CLIP-seq data. *Genome Biology*, *18*(1), 240. doi:doi:10.1186/s13059-017-1364-2

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., ... Marra, M. A. (2009). Circos: an information aesthetic for comparative genomics. *Genome Res*, *19*(9), 1639–1645. doi:10.1101/gr.092759.109

Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., ... Roadmap Epigenomics Consortium. (2015). Integrative analysis of 111 reference human epigenomes. *Nature*, *518*(7539), 317–330. doi:10.1038/nature14248

Landt, S. G., Marinov, G. K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., ... Snyder, M. (2012). ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res*, *22*(9), 1813–1831. doi:10.1101/gr.136184.111

Layer, R. M., Chiang, C., Quinlan, A. R., & Hall, I. M. (2014). LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol*, *15*(6), R84. doi:10.1186/gb-2014-15-6-r84

Lee, F. C. Y., & Ule, J. (2018). Advances in CLIP Technologies for Studies of Protein-RNA Interactions. *Mol Cell*, *69*(3), 354–369. doi:10.1016/j.molcel.2018.01.005

Li, B., & Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, *12*(1), 323. doi:10.1186/1471-2105-12-323

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. doi:10.1093/bioinformatics/btp324. eprint: https://academic.oup.com/bioinformatics/article-pdf/25/14/1754/605544/btp324.pdf

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., ... Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. doi:10.1093/bioinformatics/btp352

Li, H., & Homer, N. (2010). A survey of sequence alignment algorithms for next-generation sequencing. *Brief Bioinform*, *11*(5), 473–483. doi:10.1093/bib/bbq015

Li, H., Ruan, J., & Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res*, *18*(11), 1851–1858. doi:10.1101/gr.078212.108

Li, J., Jiang, H., & Wong, W. H. (2010). Modeling non-uniformity in short-read rates in RNA-Seq data. *Genome Biol*, *11*(5), R50. doi:10.1186/gb-2010-11-5-r50

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*, *15*(12), 550. doi:10.1186/s13059-014-0550-8

Mäder, U., Nicolas, P., Richard, H., Bessières, P., & Aymerich, S. (2011). Comprehensive identification and quantification of microbial transcriptomes by genome-wide unbiased methods. *Current Opinion in Biotechnology*, *22*(1), 32–41. **16 citations** (ISI base of knowledge). doi:10.1016/j.copbio.2010.10.003

Mammana, A., & Chung, H.-R. (2015). Chromatin segmentation based on a probabilistic model for read counts explains a large portion of the epigenome. *Genome Biology*, *16*(1), 151. doi:10.1186/s13059-015-0708-z

Memczak, S., Jens, M., Elefsinioti, A., Torti, F., Krueger, J., Rybak, A., ... Rajewsky, N. (2013). Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature*, *495*, 333 EP -.

Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nat Rev Genet*, *11*(1), 31–46. doi:10.1038/nrg2626

Meyer, C. A., & Liu, X. S. (2014). Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. *Nat Rev Genet*, *15*(11), 709–721. doi:10.1038/nrg3788

Miller, G. (2010). Forensics. familial DNA testing scores a win in serial killer case. *Science*, *329*(5989), 262. doi:10.1126/science.329.5989.262

Mirauta, B., Nicolas, P., & Richard, H. (2013). Pardiff: Inference of Differential Expression at Base-Pair Level from RNA-Seq Experiments. In A. Petrosino, L. Maddalena, & P. Pala (Eds.), *Iciap international workshops, naples, italy, september 9-13, 2013. proceedings* (Vol. 8158, pp. 418–427). Lecture Notes in Computer Science. doi:10.1007/978-3-642-41190-8_45

Mirauta, B., Nicolas, P., & Richard, H. (2014). Parseq: reconstruction of microbial transcription landscape from RNA-Seq read counts using state-space models. *Bioinformatics*, *30*(10), 1409–1416. doi:10.1093/bioinformatics/btu042. eprint: http://bioinformatics.oxfordjournals.org/content/30/10/1409.full.pdf+html

Morillon, A., & Gautheret, D. (2019). Bridging the gap between reference and real transcriptomes. *Genome Biology*, *20*(1), 112. doi:10.1186/s13059-019-1710-7

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods*, *5*. doi:10.1038/nmeth.1226

Nagarajan, N., & Pop, M. (2013). Sequence assembly demystified. *Nature Reviews Genetics*, *14*(3), 157–167. doi:10.1038/nrg3367

Nederbragt, L. (2016). Developments in NGS. doi:10.6084/m9.figshare.100940.v9

Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*, *48*(3), 443–453.

Nicolae, M., Mangul, S., Măndoiu, I. I., & Zelikovsky, A. (2011). Estimation of alternative splicing isoform frequencies from RNA-Seq data. *Algorithms for molecular biology*, *6*(1), 1–13.

Nicolas, P., Bize, L., Muri, F., Hoebeke, M., Rodolphe, F., Ehrlich, S. D., ... Bessières, P. (2002). Mining Bacillus subtilis chromosome heterogeneities using hidden Markov models. *Nucleic*

Acids Research, *30*(6), 1418–1426. doi:10.1093/nar/30.6.1418. eprint: https://academic.oup.com/nar/article-pdf/30/6/1418/9901258/301418.pdf

Nicolas, P., Leduc, A., Robin, S., Rasmussen, S., Jarmer, H., & Bessières, P. (2009). Transcriptional landscape estimation from tiling array data using a model of signal shift and drift. *Bioinformatics (Oxford, England)*, *25*(18), 2341–2347. doi:10.1093/bioinformatics/btp395

Ozsolak, F., & Milos, P. M. (2010). RNA sequencing: Advances, challenges and opportunities. *Nature Reviews Genetics*, *12*, 87 EP -.

Pachter, L. (2011). Models for transcript quantification from RNA-Seq. *arXiv e-prints*, arXiv:1104.3889. arXiv: 1104.3889 [q-bio.GN]

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature methods*, *14*(4), 417–419.

Philippe, N., Salson, M., Commes, T., & Rivals, E. (2013). Crac: An integrated approach to the analysis of rna-seq reads. *Genome Biology*, *14*(3), R30. doi:10.1186/gb-2013-14-3-r30

Pickrell, J. K., Marioni, J. C., Pai, A. A., Degner, J. F., Engelhardt, B. E., Nkadori, E., ... Pritchard, J. K. (2010). Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*, *464*, 768 EP -.

Plocik, A. M., & Graveley, B. R. (2013). New insights from existing sequence data: Generating breakthroughs without a pipette. *Mol Cell*, *49*(4), 605–617. doi:10.1016/j.molcel.2013.01.031

Poplin, R., Chang, P.-C., Alexander, D., Schwartz, S., Colthurst, T., Ku, A., ... DePristo, M. A. (2018). A universal SNP and small-indel variant caller using deep neural networks. *Nature Biotechnology*, *36*(10), 983–987. doi:10.1038/nbt.4235

Rabiner, L., & Juang, B. (1986). An introduction to hidden Markov models. *IEEE ASSP Magazine*, *3*(1), 4–16. doi:10.1109/MASSP.1986.1165342

Rausch, T., Zichner, T., Schlattl, A., Stütz, A. M., Benes, V., & Korbel, J. O. (2012). DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, *28*(18), i333.

Reyes-Herrera, P. H., Speck-Hernandez, C. A., Sierra, C. A., & Herrera, S. (2015). BackCLIP: a tool to identify common background presence in par-clip datasets. *Bioinformatics*, *31*(22), 3703–3705. doi:10.1093/bioinformatics/btv442

Rhee, H. S., & Pugh, B. F. (2011). Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell*, *147*(6), 1408–1419. doi:10.1016/j.cell.2011.11.013

Richard, H., Schulz, M. H., Sultan, M., Nurnberger, A., Schrinner, S., Balzereit, D., ... Yaspo, M.-L. (2010). Prediction of alternative isoforms from exon expression levels in RNA-Seq experiments. *Nucleic Acids Research*. **60 citations** (ISI base of knowledge). doi:10.1093/nar/gkq041. eprint: http://nar.oxfordjournals.org/content/early/2010/02/11/nar.gkq041.full.pdf+html

Rimmer, A., Phan, H., Mathieson, I., Iqbal, Z., Twigg, S. R. F., Wilkie, A. O. M., ... Consortium, W. (2014). Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nature Genetics*, *46*(8), 912–918. doi:10.1038/ng.3036

Roberts, A., Trapnell, C., Donaghey, J., Rinn, J. L., & Pachter, L. (2011). Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biology*, *12*(3), R22. doi:10.1186/gb-2011-12-3-r22

Roberts, D. L., & Solow, A. R. (2003). When did the dodo become extinct? *Nature*, *426*(6964), 245–245. doi:10.1038/426245a

Robertson, G., Schein, J., Chiu, R., Corbett, R., Field, M., Jackman, S. D., ... Birol, I. (2010). De novo assembly and analysis of rna-seq data. *Nature Methods*, *7*(11), 909–912. doi:10.1038/nmeth.1517

Robinson, J. T., Thorvaldsdottir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative genomics viewer. *Nat Biotechnol*, *29*(1), 24–26. doi:10.1038/nbt.1754

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*. doi:10.1093/bioinformatics/btp616

Rogato, A., Richard, H., Voss, B., Sarazin, A., Navarro, S. C., Navarro, L., ... Falciatore, A. (2014). The diversity of small non coding RNA populations in the diatom Phaeodactylum Tricornotum. *BMC Genomics*, *15*(1), 698. doi:10.1186/1471-2164-15-698

Rogers, M. F., Thomas, J., Reddy, A. S., & Ben-Hur, A. (2012). SpliceGrapher: detecting patterns of alternative splicing from rna-seq data in the context of gene models and EST data. *Genome Biology*, *13*(1), R4. doi:10.1186/gb-2012-13-1-r4

Saad, C., Noe, L., Richard, H., Leclerc, J., Buisine, M.-P., Touzet, H., & Figeac, M. (2018). DiNAMO: highly sensitive DNA motif discovery in high-throughput sequencing data. *BMC Bioinformatics*, *19*(1), 223. doi:10.1186/s12859-018-2215-1

Salmela, L., & Rivals, E. (2014). LoRDEC: accurate and efficient long read error correction. *Bioinformatics*, *30*(24), 3506–3514. doi:10.1093/bioinformatics/btu538. eprint: https://academic.oup.com/bioinformatics/article-pdf/30/24/3506/17144171/btu538.pdf

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, *74*(12), 5463–5467.

Santana-Garcia, W., Rocha-Acevedo, M., Ramirez-Navarro, L., Mbouamboua, Y., Thieffry, D., Thomas-Chollier, M., ... Medina-Rivera, A. (2019). RSAT variation-tools: An accessible and flexible framework to predict the impact of regulatory variants on transcription factor binding. *Computational and Structural Biotechnology Journal*. doi:10.1016/j.csbj.2019.09.009

Schmitt, A. D., Hu, M., & Ren, B. (2016). Genome-wide mapping and analysis of chromosome architecture. *Nat Rev Mol Cell Biol*, *17*(12), 743–755. doi:10.1038/nrm.2016.104

Schulz, M. H., Weese, D., Holtgrewe, M., Dimitrova, V., Niu, S., Reinert, K., & Richard, H. (2014). Fiona: a parallel and automatic strategy for read error correction. *Bioinformatics*, *30*(17), i356–i363. doi:10.1093/bioinformatics/btu440. eprint: http://bioinformatics.oxfordjournals.org/content/30/17/i356.full.pdf+html

Schulz, M. H., Zerbino, D. R., Vingron, M., & Birney, E. (2012). Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics*, *28*(8), 1086–1092. doi:10.1093/bioinformatics/bts094. eprint: https://academic.oup.com/bioinformatics/article-pdf/28/8/1086/18531388/bts094.pdf

Shen, S., Park, J. W., Lu, Z.-x., Lin, L., Henry, M. D., Wu, Y. N., ... Xing, Y. (2014). rMATS: robust and flexible detection of differential alternative splicing from replicate rna-seq data. *Proceedings of the National Academy of Sciences*, *111*(51), E5593–E5601. doi:10.1073/pnas.1419161111. eprint: https://www.pnas.org/content/111/51/E5593.full.pdf

Shrestha, A. M. S., Frith, M. C., & Horton, P. (2014). A bioinformatician's guide to the forefront of suffix array construction algorithms. *Brief Bioinform*, *15*(2), 138–154. doi:10.1093/bib/bbt081

Shrestha, A. M., Asai, K., Frith, M., & Richard, H. (2018). Jointly aligning a group of DNA reads improves accuracy of identifying large deletions. *Nucleic Acids Research*, *46*(3), e18. doi:doi:10.1093/nar/gkx1175

Simpson, J. T., & Durbin, R. (2012). Efficient de novo assembly of large genomes using compressed data structures. *Genome Research*, *22*(3), 549–556. doi:10.1101/gr.126953.111. eprint: http://genome.cshlp.org/content/22/3/549.full.pdf+html

Siragusa, E., Weese, D., & Reinert, K. (2013). Fast and accurate read mapping with approximate seeds and multiple backtracking. *Nucleic Acids Research*, *41*(7), e78–e78. doi:10.1093/nar/gkt005. eprint: http://oup.prod.sis.lan/nar/article-pdf/41/7/e78/25341856/gkt005.pdf

Smith, T., & Waterman, M. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*, *147*(1), 195–197. doi:https://doi.org/10.1016/0022-2836(81)90087-5

Stegle, O., Teichmann, S. A., & Marioni, J. C. (2015). Computational and analytical challenges in single-cell transcriptomics. *Nat Rev Genet*, *16*(3), 133–145. doi:10.1038/nrg3833

Steijger, T., Abril, J. F., Engström, P. G., Kokocinski, F., Consortium, T. R., Akerman, M., ... Zhang, M. Q. (2013). Assessment of transcript reconstruction methods for RNA-seq. *Nature Methods*, *10*, 1177 EP -.

Sultan, M., Schulz, M. H., Richard, H., Magen, A., Klingenhoff, A., Scherf, M., ... Yaspo, M.-L. (2008). A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science*, *321*(5891), 956–960. doi:10.1126/science.1160342

Teytelman, L., Thurtle, D. M., Rine, J., & van Oudenaarden, A. (2013). Highly expressed loci are vulnerable to misleading chip localization of multiple unrelated proteins. *Proc Natl Acad Sci U S A*, *110*(46), 18602–18607. doi:10.1073/pnas.1316064110

The long view on sequencing. (2018). *Nature Biotechnology*, *36*, 287 EP -.

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

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

Thormann, V., Rothkegel, M. C., Schopflin, R., Glaser, L. V., Djuric, P., Li, N., ... Meijsing, S. H. (2018). Genomic dissection of enhancers uncovers principles of combinatorial regulation and cell type-specific wiring of enhancer-promoter contacts. *Nucleic Acids Res*, *46*(6), 2868–2882. doi:10.1093/nar/gky051

Tompa, M., Li, N., Bailey, T. L., Church, G. M., De Moor, B., Eskin, E., ... Zhu, Z. (2005). Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol*, *23*(1), 137–144. doi:10.1038/nbt1053

Tran, V. D. T., Souiai, O., Romero-Barrios, N., Crespi, M., & Gautheret, D. (2016). Detection of generic differential rna processing events from rna-seq data. *RNA Biology*, *13*(1), 59–67. PMID: 26849165. doi:10.1080/15476286.2015.1118604. eprint: https://doi.org/10.1080/15476286.2015.1118604

Trapnell, C., Hendrickson, D. G., Sauvageau, M., Goff, L., Rinn, J. L., & Pachter, L. (2012). Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature Biotechnology*, *31*, 46 EP -.

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., ... Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, *28*, 511 EP -.

Treangen, T. J., & Salzberg, S. L. (2012). Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics*, *13*(2), 36–46.

van Dijk, E. L., Jaszczyszyn, Y., Naquin, D., & Thermes, C. (2018). The Third revolution in sequencing technology. *Trends in Genetics*, *34*(9), 666–681. doi:10.1016/j.tig.2018.05.008

van Nimwegen, E., Paul, N., Sheridan, R., & Zavolan, M. (2006). SPA: a probabilistic algorithm for spliced alignment. *PLoS Genet*, *2*(4), e24. doi:10.1371/journal.pgen.0020024

Vaser, R., Sović, I., Nagarajan, N., & Šikić, M. (2017). Fast and accurate de novo genome assembly from long uncorrected reads. *Genome research*, *27*(5), 737–746.

Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., ... Burge, C. B. (2008a). Alternative isoform regulation in human tissue transcriptomes. *Nature*, *456*(7221), 470–476. doi:10.1038/nature07509

Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., ... Burge, C. B. (2008b). Alternative isoform regulation in human tissue transcriptomes. *Nature*, *456*(7221), 470–476. doi:10.1038/nature07509

Wang, K., Singh, D., Zeng, Z., Coleman, S. J., Huang, Y., Savich, G. L., ... Liu, J. (2010). MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res*, *38*(18), e178. doi:10.1093/nar/gkq622

Wang, J.-P. Z., & Lindsay, B. G. (2005). A penalized nonparametric maximum likelihood approach to species richness estimation. *Journal of the American Statistical Association*, *100*(471), 942–959. doi:10.1198/016214504000002005. eprint: https://doi.org/10.1198/016214504000002005

Weese, D. [D.], Emde, A.-K., Rausch, T., Döring, A., & Reinert, K. (2009). RazerS - fast read mapping with sensitivity control. *Genome Research*, *19*(9), 1646–1654.

Weese, D. [David], Schulz, M. H., & Richard, H. (2017). DNA-Seq error correction based on substring indices. In M. Elloumi (Ed.), *Algorithms for next-generation sequencing data: Techniques, approaches, and applications* (pp. 147–166). doi:doi:10.1007/978-3-319-59826-0_7

Weese, D. [David], & Siragusa, E. (2017). Full-text indexes for high-throughput sequencing. In *Algorithms for next-generation sequencing data: Techniques, approaches, and applications (ed: Elloumi, mourad)* (pp. 41–75). doi:10.1007/978-3-319-59826-0_2

Wick, R. R., Schultz, M. B., Zobel, J., & Holt, K. E. (2015). Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics*, *31*(20), 3350–3352. doi:10.1093/bioinformatics/btv383

Workman, R. E., Tang, A., Tang, P. S., Jain, M., Tyson, J. R., Zuzarte, P. C., ... Timp, W. (2018). Nanopore native RNA sequencing of a human poly(A) transcriptome. *bioRxiv*. doi:10.1101/459529. eprint: https://www.biorxiv.org/content/early/2018/11/09/459529.full.pdf

Ye, K., Schulz, M. H., Long, Q., Apweiler, R., & Ning, Z. (2009). Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*, *25*(21), 2865–2871. doi:10.1093/bioinformatics/btp394. eprint: https://academic.oup.com/bioinformatics/article-pdf/25/21/2865/6059071/btp394.pdf

Zakeri, M., Srivastava, A., Almodaresi, F., & Patro, R. (2017). Improved data-driven likelihood factorizations for transcript abundance estimation. *Bioinformatics*, *33*(14), i142–i151. doi:10.

1093/bioinformatics/btx262. eprint: https://academic.oup.com/bioinformatics/article-pdf/33/14/i142/25157301/btx262.pdf

Zambelli, F., Pesole, G., & Pavesi, G. (2013). Motif discovery and transcription factor binding sites before and after the next-generation sequencing era. *Brief Bioinform*, *14*(2), 225–237. doi:10.1093/bib/bbs016

Zheng, W., Chung, L. M., & Zhao, H. (2011). Bias detection and correction in RNA-sequencing data. *BMC Bioinformatics*, *12*, 290. doi:10.1186/1471-2105-12-290

Chapter 2

Perspectives on Omics and Big Data in Biology

2.1 Promises and limitations of Omics data

2.1.1 Context

Signal processing has been my main scientific interest in the last decade. Raw sequence data are computationally processed into a few meaningful statistical digests. This strategy of producing precise molecular descriptions is usually considered the *de facto* path towards scientific discoveries, owing to the large amount of sequencing data generated. Accordingly, when we analysed the population of small RNAs in diatoms (section 1.6), we were more ambitious. We extended our goal beyond a set of summary tables and aimed to uncover the mechanisms of non coding RNA mediated silencing in diatoms. Final objective was a biological question: How are small RNAs produced and what is their potential regulatory role? We presumed to generate a load of biological hypotheses from the data, and test those hypothesis experimentally afterward.

Alas, I realised soon enough that, bewitched by the promises of High Throughput Sequencing, my expectations were exaggerated. We could detail or catalog the population of short sequences to unprecedented depths, but without additional experiments the biology of those RNAs would remain elusive.[1] This experience was quite surprising. I started my PhD a few years after the sequencing of the human genome, and I always implicitly assumed that by describing cells with more and more data –sequencing or whatsoever, we would be left mainly with statistical and modelling issues. Indeed, the stream of technological revolutions taking place in the Life Sciences was turning upside down our way of doing research. A common belief at the turn of the century was: we just go on and mine the data, and the biological discoveries will unfold (Anderson, 2008; Kell & Oliver, 2004). Obviously, thing are a bit more complicated. More data does not necessarily mean more biological knowledge (Allen, 2001a), and it does not always translate into a mechanistic understanding or an explanation (Allen, 2001b). The stark contrast, between the merely descriptive aspect of sequencing assays, and the promises from the postgenomic era led me to a critical examination of the prospects and limitations of the use of omics data for scientific discovery.

The technological revolution brought first by HTS disrupted all disciplines in molecular biology, moving from a narrow view of a few elements to genome-wide measures of the molecular activity.

[1] We indeed performed additional experiments, but they were mainly validation of HTS based observations.

Global discoveries were made possible (Brown & Botstein, 1999) and disrupted many common assumptions in molecular biology, starting for instance with the surprising low number of genes in multicellular organisms (Lander et al., 2001), followed by the multiple layers of transcriptional regulations in eukaryotic organism, including miRNAs (Bartel, 2009), the unexpected prevalence of alternative splicing (Wang et al., 2008) or the existence of pervasive tranlation (Ingolia, Lareau, & Weissman, 2011). At the same time, the modern promise of omics took form as new kind of biology, where large-scale data production would be expected to bring new scientific knowledge (Golub, 2010). Those results were not produced by "simply looking at the data", but by directing it towards a particular scientific question (Allen, 2001b). A debate already started around the danger of large scale projects to simply produce catalogue without a specific question in mind (R. Weinberg, 2010). Woese (Woese, 2004) advocated the imperious need for a working plan: "without a guiding vision, there is no road ahead; ... [life] science becomes an engineering discipline, concerned with temporal practical problems".

There is a gap in perception on the use of omics and its gain to biology. How can we accommodate the aim of cataloguing the molecular basis of living organisms and "let the data speak for itself" (Anderson, 2008), with the traditional scientific method, which is rooted into the formulation of alternative hypothesis and their experimental test? What is the impact on our scientific practice? And what will define a good scientific practice in our Data-Rich world (Elliott, Cheruvelil, Montgomery, & Soranno, 2016)?

More generally, what is the role of "Big Data" in biology? How does it interface with computationally intensive analyses? Can automated methods (*e.g.* Machine Learning) transform data into knowledge, and scientific discovery be carried out directly by machines (Anderson, 2008; Nickles, 2018)? The large quantities of data produced, and the methods proposed for their processing, are not explicit. Will it make biology not intelligible for human? (Anonymous, 2000; Nickles, 2018)

Understanding possibilities and limits has important implications: it provides a panorama of the scientific practice nowadays, and helps me to understand which stand a bioinformatician should take if he wants to move aside from signal processing and engineering. It also has a direct impact on clarifying how different practices or projects are evaluated by funding agencies, and how a grant application is judged.

Those questions will be considered from different angles, first recalling how it started with the reductionistic/holistic debate. High throughput assays, by enabling a *holistic* view of the cellular processes, were historically cast in the 2000s as a remedy to the limitations of *reductionistic* approaches, which *de facto* relied on the hypothetico-deductive scientific method. I will detail how reductionism and holism[2] are defined in the life science, in order to clarify how this dichotomy was established, and wether or not they should be seen as opposite practices. Additionally, Omics methods are mainly descriptive, which is at odds with hypothesis-driven science and this raise additional questions on the various dimensions of scientific practice.

More recent accounts of the scientific method in biology show that, rather than entrenching reductionist methods and holistic schools, current practice combines different approaches. They alternates between them to narrow our knowledge gap: investigation/discovery, experimental tests of hypothesis, computational modelling and methods/technology development. I will summarise some of the descriptions proposed in the last decade by philosophers of science. Following (Callebaut, 2012), we will see that *Scientific perspectivism*, an empiricist thesis proposed by Giere (Giere,

[2]sometimes referred to as System biology as well

2006), which shares themes with van Fraassen (van Fraassen, 2008), is a compelling framework to understand current practices making sense of omics data.

Finally, omics data is an emanation/product of a much larger revolution impacting all areas of science: Big Data Driven Research. Big data applied to biology is not new, it already existed three centuries ago during Linnaeus taxonomical effort (Müller-Wille & Charmantier, 2012), and is more general than only the generation of omics data, englobing as well Database creation and curation. We can still ask how omics data creation interfaces with resource building? What is the interest of generating those resources? Again, shouldn't the scientific question be put first (R. Weinberg, 2010)? Will Big Data in Biology, combined with the formidable computing power and the recent revolution we observe in the areas of machine learning, allow us to create automatically new knowledge? What are the short and mid term challenges?

2.1.2 Reductionistic and Holistic science

Let us recall the dichotomy between holism and reductionism in molecular biology. I will concentrate the discussion on omics data, our main theme. Each omics experiment provides a global and unbiased molecular view on a biological system –it literally is a digital molecular footprint: counting all transcribed molecules, characterising the whole genome.
For this reason, already in its infancy (first complete chromosomes sequenced, microarray analyses), omics data was hailed as a new mode of exploration which would describe the entire system down from its molecular parts, so called *holistic* approach.

> "Exploration means looking around, observing, describing and mapping undiscovered territory, not testing theories or models. The goal is to discover things we neither knew nor expected, and to see relationships and connections among the elements, whether previously suspected or not. It follows that this process is not driven by hypothesis and should be as model-independent as possible. We should use the unprecedented experimental opportunities that the genome sequences provide to take a fresh, comprehensive and openminded look at every question in biology. If we succeed, we can expect that many of the new models that emerge will defy conventional wisdom" (Brown & Botstein, 1999)

During exploratory analysis, the scientist starts by detecting patterns or regularities from the molecular observations, and then proceeds upwards by proposing a biological phenomena to explain them: inductive reasoning at its best. To be more general, holistic science recognises that all molecular constituent are interconnected and aims to study biological systems in their global complexity (whole organisms or populations of organisms). This practice was also called systems biology[3]: "There is now a golden opportunity for system-level analysis to be grounded in molecular-level understanding, resulting in a continuous spectrum of knowledge" (Kitano, 2002).

With the massive accumulation of molecular data started in the beginning of the century, Systems Biology rose to prominence and was hailed as the long needed "move away from reductionism" (Regenmortel, 2004). The advent of a new kind of modelling in biology was taking place (Mazzocchi, 2008). What do we really understand by reductionism? Put loosely, reductionism describes a system by dividing it into a set of core parts to study them separately. Nevertheless, a precise definition is often lacking. As Richard Dawkins once wrote: "Reductionism is one of those things, like sin,

[3]Systems biology englobe slightly more than that, for instance bottom up modelling.

that is only mentioned by people who are against it" (Dawkins, 1986). In fact, reductionism can have different acceptations and therefore be seen from different points of view. Either ontological, epistemological or methodological (Brigandt & Love, 2017) (see figure 2.1). We need to know exactly which type is relevant when we criticise reductionism in the life sciences.

Ontological reductionism is an assertion made about the entities in the world. To the question: Is there anything more than physical objects, properties, events, etc?, The answer of ontological reductionism is no. For instance there are no vital forces or souls. This is not of particular interest for us in the debate.

The two remaining types of reductionism are often confused and it is important to clarify in which aspect they differ.

Epistemological reductionism follows the idea that "the knowledge about one scientific domain can be reduced to another body of scientific knowledge" (Brigandt & Love, 2017). This claim -of epistemic reduction between disciplines, is a very general claim, as when Crick famously said: "the ultimate aim of the modern movement in biology is to explain all biology in terms of physics and chemistry" (Crick, 1966). One could also qualify this opinion as a reductionism of "fundamentalists" (Woese, 2004). It has strongly shaped research in the life science, especially in molecular biology from the 1960s on, when research moved on from the structure of the DNA, to the search for the "code" and culminating with the pursuit the Human Genome Project (Richardson & Stevens, 2015). Woese already noted that this vision of biology remained stucked in an influence from pre-XXth century physics, where all phenomena would derive from theories and models, and all explanations would be based on chemistry and physics (Mazzocchi, 2008; Woese, 2004). This epistemological reductionism creates a hierarchy designating some disciplines to be more "fundamental" (physics, chemistry), and others resulting in mere consequence of the "laws of the universe" (S. Weinberg, 1987) [4]. It also lends the idea of the existence of an ultimate level of truth where knowledge would "bottom out" [5]. The nefarious implications of this principle had already been criticised by many when they rejected epistemic reductionism (Mayr, 1988; Woese, 2004). Indeed, even if there have been successes, such as the epistemic reduction of thermodynamics to statistical physics, this is an other challenge across disciplines. For instance, how can particle physics explain biological phenomenons such as self replication or evolution? At a molecular level, Anfinsen's hypothesis postulates that the global free energy minimum of a protein sequence determines its folding. From this hypothesis it should be possible to *reduce* the problem of protein folding to solving a set of physical equations. Recent results from protein structure prediction have clearly shown that this goal is unattainable (Jumper et al., 2021). A general argument is that Physics and Biology are *epistemologically discontinuous*. Even modern physics with concepts such as entanglement or the incertitude principle does not agree anymore with epistemological reduction between discipline[6]. This type of reductionism was referred to by some of its critics, but if "exploring the epistemic relationships between different disciplines might be grist in the mill for a philosopher of science" (Fang & Casadevall, 2011), it does not reflect the scientific practice in the life science. We can put it aside for our purpose, but let's emphasise that epistemic reduction was often referred to in the critics of reductionism in the Life Sciences and sometimes mixed up with experimental reduction presented below (Regenmortel, 2004).

[4] starting around those arguments, "the unreasonable effectiveness of mathematics" also pops up in the conversation
[5] I will treat this point more in detail in the next section
[6] Woese wrote "in a metaphysical sense Molecular Biology was outdated from the onset!" (Woese, 2004)

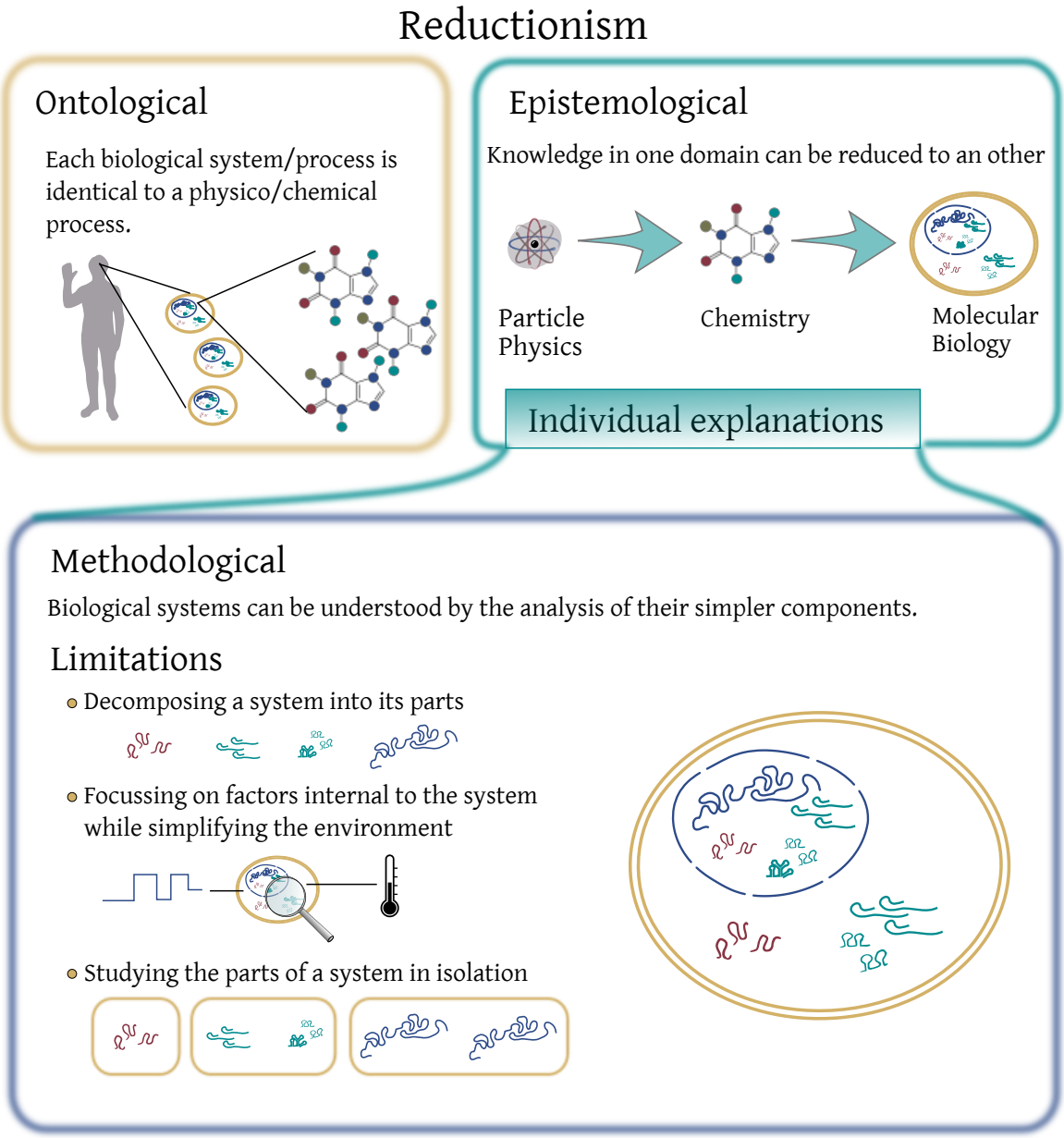


Figure 2.1: Illustration of the different kinds of reductionisms and their limitations. Reductionism can make different claim about the relation between scientific domains. Top left: Ontological reductionism implies that there is no vitalism or no soul. Top right: Epistemological reductionism (also called fundamentalist reductionism). Bottom: Methodological reductionism is a specific case of epistemological reductionism applied to individual explanation (it is also called experimental reductionism). We are mainly interested in understanding the limitations of this later type as described in (Kaiser, 2011))

Explanatory reductionism is a specific type of epistemic reduction, when applied to individual explanation. This technique aims to transform a higher-level explanation (for example a cellular one) into a lower level one (for instance a molecular one) of the same phenomenon (Kaiser, 2011). Explanatory reduction relates the *explanandum* (the phenomenon to be explained) to the *explanans* (the explanatory relevant factors). More specifically the relation of reduction holds between a representation (or description or model) of the phenomenon (or behaviour of a system) to be explained and the representation of the causal factors referred to in the explanation. The most relevant and used type of reductionism is a subtype of explanatory reductionism.

Methodological (or experimental) reductionism claims that biological phenomena and systems can be understood by the analysis of their smaller components. This strategy traces back to Descartes and his clockwork model ("divide each difficulty into as many parts as possible and as might be necessary for its adequate solution" (Descartes, 1637/1850, p. 61) . This is one of the most common technique used in biology to prove scientific facts. What are the characteristics of this type of reductionism in the life science? Let us take a a simple example from molecular biology with cell lines. Cell lines are applied as a model systems to decipher the sequence determinant mechanisms of gene expression. By using a combination of ChIP-Seq, RNA-Seq and mutagenesis experiments one can link transcription to the presence of sequence motifs and the binding of a transcription factor in the promoter region (section 1.1). Gene expression is reduced to the combined effect of sequence content and transcription factor binding. We started from a broad definition of reductionism, but this example illustrates that we are interested in considering the limitations of **methodological reductionism**. Now, what are the main limitations that are mentioned by its practitioners? Marie Kaiser (Kaiser, 2011) identifies three key ingredients that are criticised for the reductive explanations in the life science (see Figure 2.1 bottom).

The first limitation is the decomposition of a system into its parts. This characteristic is common to all reductive methods: the system is decomposed into components and those components are used to explain it as a whole. In the case of molecular biology, the components are objects such as genes and mutations, or transcript molecules.[7] For omics data, the objects considered "bottom out" at a molecular level. This first point is what people usually think about for reductive explanation, but there are two other important characteristics commonly observed.

The second limitation is focussing on factor internal to the system while simplifying the environment. An example given by Kaiser is the case of protein folding. One describes the fold (internal factor) while fixing a set of "background conditions" (temperature, pH-value, salt concentration) necessary for the folding to occur. The external factors are considered as a fixed input.

The third limitation is studying the parts of a system *in isolation*. Note that this is conceptually different from the first limitation (decomposition of the system into parts). Also the parts are never studied in complete isolation (it is rarely experimentally possible). One can understand that the parts are not investigated *in situ* –in the context of the system they are part of, but detached from it (e.g., *in vitro*). In the context of molecular biology, one example would be the study of Knock-Out lines, in which genes are inactivated sequentially.

We can now address the limitations of the reductionist explanations when applied to biologi-

[7]Note that the lower level objects do not have to be molecular parts to be components used in a reductive explanation.

cal systems. Living cells are *complex* systems (Mazzocchi, 2008): cellular activity is the result of combining multiple layers of regulation and molecular interactions, tuned by evolution over multiple generations. To understand regulation, it is important to take into account multiple elements, ranging over up to 6 orders of magnitude (from the single gene or transcript, to the global organisation of a group of cells). Sequential perturbation of the system components (studying the parts in isolation), and monitoring of its impact on the system (decomposition), will rapidly face a combinatorial wall. Also, because we simplify the environment when studying cell lines, we are not always capable of generalizing molecular results to other tissues or complete organisms.

> "To summarize, it can be said that the more complex the organisation of a system is, and the more its parts are integrated and interdependent on each other, the more limited are the insights into the system one can achieve by utilising the two reductive methods: decomposition and investigation of parts in isolation." (Kaiser, 2011).

In the past, the reductionist technique depicted showed a few limitations when studying cellular systems. A common problem was related to the definition of genes, originally described as elementary functional building blocks (*parts*). This kind of simplification fails to describe how the same genotype can give rise to different cell types. In particular, mechanisms such as pleiotropy (multiple fonctions for one gene) or epistasis (higher level interactions between genetic elements) are not amenable to such reductionist view. Although not necessarily reductionist in itself, it also propelled simplifying popular views about genes a single explanatory variables of traits or placeholder for essences (Heine, 2017) A second limitation is the existence of emergent properties: properties of a system can emerge at a higher level than predicted from the properties of its parts. For instance the virulence of a microbe cannot be attributed solely to the microbe or to the host, but is the result of the interaction between them and with the environment (Casadevall, Fang, & Pirofski, 2011)

Omics and the holistic/reductionist debate As we understand the different limitations of a reductionist approach applied in the life science. But what does omics data add to the search for a holistic explanation? Is it by itself sufficient to "move away from reductionism"?

First there is a temporal effect. What we see as holistic is not definitive, it is also motivated by shifts in technology. What we perceived as a limitation of reductionism can change in a decade as the technology improves. The omen from (Brown & Botstein, 1999) reported previously was motivated, 20 years ago, by the advances of microarray technology, which was later represented as ripe with limitations by RNA-Seq. RNA-Seq is now being replaced by single cells and spatial transcriptomics. So in a few decades, we moved from the characterisation of a few genes, to the whole populations of expressed transcripts, to their intercellular resolution, and we are now aiming at an even more global picture. Is that being holistic?

Let us consider a typical omics experiment. For instance an RNA-Seq assay can be performed to explain the shifts in the transcriptional landscape in specific cell types or among cells with different genetic backgrounds. Such a setup necessarily operates on a limited number of conditions (it "simplifies the environment"), usually considering cell lines of fixing the cell type(s) ("studying the parts of the system in isolation"). So, although those assays were at some point in the past associated with holistic science, we understand that they present at least two of the characteristics of experimental reductionism highlighted by Kaiser. One counter argument usually brought forward here is that the system should be more integrated, that we should aim at a "more complete picture",

by integrating other information, such as DNA conformation or epigenetic marks, to be considered sufficiently holistic. But adding experimental assays, such as protein expression or genotype, or the epigenome, does not change the fact we are still following a reductionist agenda by simplifying the environment and studying the parts in isolation. Omics-Holism, the new biology for the XXIst century, may have propelled systems approach, even though it provides reductionist explanations.

We just understood that, although omics data analysis is reductionist at its core, its ability to exhaustively describe certain parts of a system have made it seem more holistic. But being exhaustive is not enough: "feckless [..] systems biology may merely describe phenomena without providing explanation or mechanistic insight" (Fang & Casadevall, 2011). This is manifest with omics assays that are observational by design. Multiple conditions could be considered, but they only work for one hypothesis at a time. This leads us to an other important axis of debate on the generation of scientific knowledge: can we be satisfied with biology limited to being a descriptive science or should we stick to the scientific method of testing hypothesis? We need to speak about another dimension of the scientific practice on the place of omics assays, being mainly descriptive, and their ability to generate scientific knowledge (see Figure 2.2).

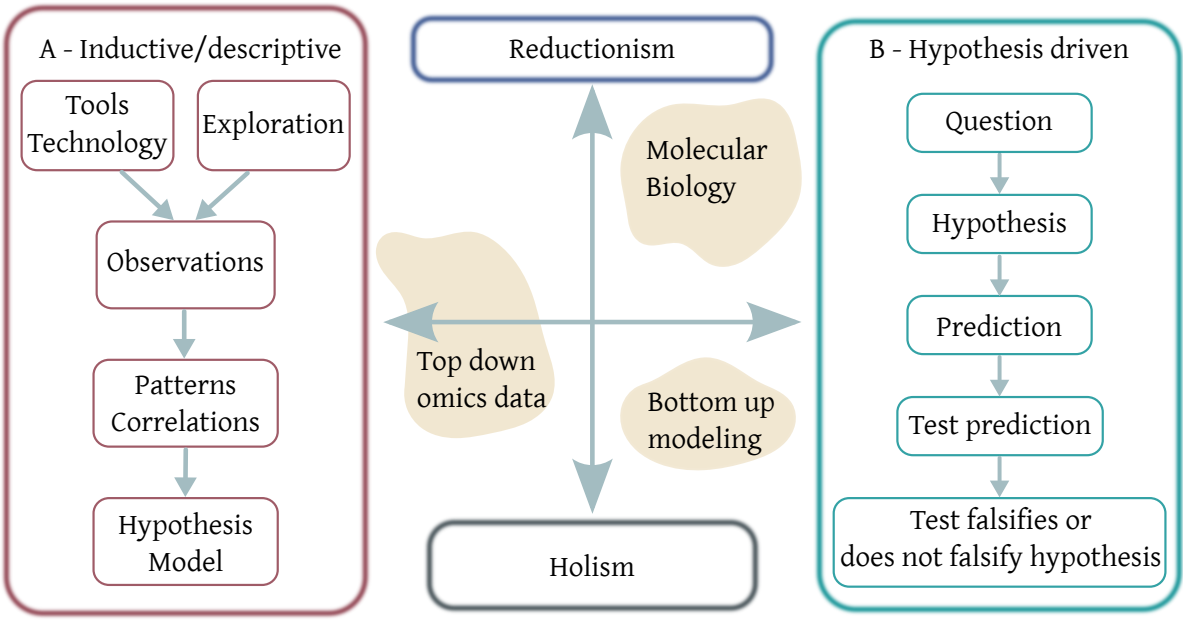


Figure 2.2: The various dimensions of research inquiry . Different types of scientific experiments are represented by light orange shapes on a 2D plane separating the type of approach (Reductionism/Holism) on the vertical axis and the scientific method (A - Inductive / B - Hypothesis driven) on the horizontal axis.

Descriptive and Hypothetico-deductive science As we mentioned earlier, the strong point of omics assays is that they allow "exploration", and "looking around". Using molecular biology basic principles and statistical models, we apply experimental reductionism to bring a rich description of the system. We can then search for recurrent patterns (e.g. groups of genes co-regulated, segregating mutations), and propose novel scientific hypotheses (figure 2.2A, left panel). It is not enough. We collect precise observations, but we are in practice as clueless as a detective examining a crime scene. The experiment only hints at clues, it is *circumstantial evidence* but not a direct proof.

In all scientific inquiries, accumulating empirical data is not sufficient *per se* to make scientific claims. Even if I observe the same pattern multiple times, it can always be disproved by additional

evidence –one has to be careful with hasty generalisations. In the case of omics assays, publication pressure, combined with high sequencing costs, can provide a good ground for such tendency, and to strong claims are made that do not necessarily lead to biological knowledge. For instance, under-appreciated technological errors can rapidly get branded as biological knowledge [8]. Another problem can be the over-interpretation of simple facts. Most famously the ENCODE project reported that 80% of the human genome was assigned a biochemical "function" (ENCODE Project, 2012), meaning that on those regions at least some biochemical reaction was occurring. It caused an uproar from scientists which argued that those claims where inflated, likely to attract media coverage (Doolittle, 2013; Graur et al., 2013). It is an archetypical example of the problem of revamping observational data as a scientific claim, without a model to support it.

The canonical way of proving hypothesis is to use the hypothetico-deductive (HD) methods (it is coined as the scientific method in scientific textbooks). It works by proposing one or multiple hypothesis to be tested by mean of an experiment. The prediction of the hypothesis leads to a clear interpretation of the implication of different outcomes on the theory (figure 2.2B, right panel). This implies that the scientific hypothesis needs to be clearly specified and tested to prove scientific facts. Popper further clarified what can be done. Indeed, no theory can ever be proved directly (Popper, 1959), and a scientific theory has to be able to make predictions which can be disproved in some way (*falsifiable*). As Popper writes, the hypothesis under scrutiny cannot be verified but has to be falsifiable by the test in order to prove anything. Back to our crime scene analogy, this corresponds to the strategy advocated by the fictional detective Sherlock Holmes: formulate rival hypothesis as to the identity of the suspect and eliminate them until one hypothesis remains: "When you have eliminated the impossible, whatever remains, however improbable, must be the truth" (Doyle, 1890).

Even though HD is considered the norm, we cannot disregard the gain of descriptive science to various disciplines. Some important scientific disciplines, such as paleontology, astronomy or the study of evolution are essentially descriptive (Casadevall & Fang, 2008). In all of those cases, we cannot make interventions on the system. For instance, paleontology is mainly limited to the observation of fossils. Laws of physics can also be considered descriptive, although we use them everyday to make predictions. We remark that descriptive studies are usually a first attempt to consider novel questions.

Conversely, one critic of HD methods is that they are too narrow-minded, they can only provide *post hoc* confirmation. For instance, during RNA-Seq analysis, the typical summary statistics consist in the list of differentially expressed genes between a normal or wild type condition and one or multiple mutants. Those genes are then annotated, and functional enrichment is computed. We then usually only *confirm* that the regulatory changes observed conform to the behaviour we expect from our model. It was even advocated that the method of *strong inference* –the process of sequential inductive inference- can provide a much more efficient technique for scientific thinking (Platt, 1964). So, although we need a HD technique to prove facts, it will not generate hypothesis. Runs of explanatory analysis are necessary to formulate hypothesis and lay out a research agenda.

To summarize, rather than an antagonism between holism and reductionism or descriptive and HD, the scientific practice nowadays is often an *iterative process* that integrates those aspects of scientific research.

[8] See for instance (Poon et al., 2016) where the authors reported a very high within host *influenza* diversity but the thorough reanalysis from (Xue & Bloom, 2019) showed that this was only due to an experimental artifact related to pooling of the samples.

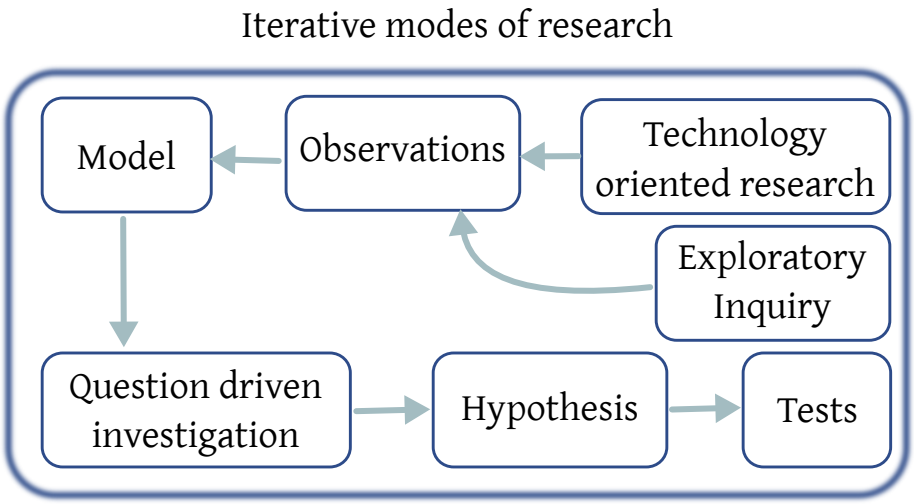


Figure 2.3: A sketch of the main steps of iterative research inquiry that characterise current scientific methods.

2.1.3 What is the scientific method anyway?

When describing modern scientific practice, authors give different accounts, by emphasizing on a back and forth (Kell & Oliver, 2004), Iteration (Beard & Kushmerick, 2009; Elliott et al., 2016; O'Malley, Elliott, & Burian, 2010) or even refer to a new dimension of the practice, where model based hypothesis can be specified in order to test a theory (Voit, 2019). The very interesting account of the different phases in miRNA research from (O'Malley et al., 2010) shows that researchers have to alternate between different modes of investigation in order to leverage the tools at their disposal (Casadevall & Fang, 2008; Fang & Casadevall, 2011). (Elliott et al., 2016) also describe an open-ended, non-linear process:

> "Scientist attempt to answer research questions with observations, field studies, or integrated databases (Sabina Leonelli, 2014), they engage in exploratory inquiry or modelling exercises to detect patterns in available data (...), or they create new tools, techniques and methods (Beard & Kushmerick, 2009; O'Malley et al., 2010) – all of which enable them to test hypothesis, answer questions or gather additional data more effectively."

Some other authors show how current practice moves from hypothesis generation, translates them into a model, and makes predictions that are amenable to a HsD method (Beard & Kushmerick, 2009). What is important is that we can extract different ingredients, common to all practices (Callebaut, 2012) (see Figure 2.3): exploratory inquiry, technology oriented research, and question driven investigation. Presented from this perspective, this alternance, this back and forth, seems completely reasonable. Although we started this chapter guided by single methods and theories, we appreciate now that scientific questions in the Life Sciences are complex and are not amenable to understanding with a single strategy in mind.

Why did this whole debate emerges in the first place? In my opinion, this stems from an implicit stance in Systems Biology which says that measuring everything will give us a *complete* understanding of the biological system. This idea is commonly shared among scientists so far. A

complete understanding will eventually lead us to a core set of true rules that describe the universe. In the end of the XXth century, Objectivists championed the idea that "there are truth out there to be discovered, truths that once discovered will form a permanent part of human knowledge" (S. Weinberg, 2001). Objectivism follows a long line of debates in philosophy of science about realism and the notion of truth. It bears some familiarity with epistemological reductionism mentioned earlier (if indeed some absolute truths exist, then they could be reduced into scientific disciplines (S. Weinberg, 1987)). Although the debate was running for a few centuries among philosophers of science, objectivism is often implicitly assumed by science practitioners. The question of whether the theoretical claims of science possess an absolute truth value is not obvious and was already addressed to by various philosophical schools [9]. Among them, criticism came from constructivists, mainly scientists coming from Humanities and social sciences, who argued that scientific knowledge is a sort of "social construct". There is no truth but a consensus that scientists attain regarding what they *say* to have found. This lead in the nineties to the *Science war* between the two extreme positions.

Figure 2.4: Blind monks examining an elephant, (ca 1814-1818) from Katsushika Hokusai (1760-1849)

Those ideas can be illustrated with an old parable. It is the one of the "Blind monks and the Elephant" (See figure 2.4 for a reproduction from the Edo period). This story originated in the ancient Indian subcontinent and a famous poem by John Godfrey Saxe in the 19th century summarises it well:

It was six men of Indostan
To learning much inclined,

[9]Pragmatism is for instance a movement which has been criticising realism starting in the late nineteenth century

Who went to see the Elephant
(Though all of them were blind),
That each by observation
Might satisfy his mind

The *First* approached the Elephant,
And happening to fall
Against his broad and sturdy side,
At once began to bawl:
"God bless me!—but the Elephant
Is very like a wall!"

And subsequently each blind man has a different interpretation on what the elephant is, a wall, a snake, a spear, a tree, a fan or a rope. The poem concludes "And so these men of Indostan, disputed loud and long [...] Though each was partly right, they were all in the wrong!". The critic from Saxe was aimed towards religious scholars and their interpretation of the divine, but the parable can sustain many interesting interpretations. It can illustrate how scientists cope with empirical evidences: each instrument provides a different measurement (one could say a different *perspective*), and we obviously need to conciliate those to obtain *a representation* of the elephant. In other words, empirical science has to work by integrating information coming from different perspectives to understand a phenomenon.

If we now want to apply Weinberg's objectivist stance, we should add that (1) there is indeed *an elephant in the room* and (2) that the monks can ultimately reach a *truthful* understanding of it. On the contrary, constructivists would argue that, no matter the number of monks, or how long they study their part of the Elephant, they will always come up with a social construction explaining what the animal truly is. Even though they can come to an agreement, the representation they advance needs not to correspond to a scientific truth. They would add that given that "Reality seems capable of sustaining more than one account given of it, depending upon the goals of those who engage with it" (Shapin, 1982, p.194), the ultimate knowledge claims are *contigent* to the context, they are shaped by multiple additional factors such as research process, or societal implications at play. In the end, the constructivists mainly lost this Science wars, which does not mean that we should be satisfied with objectivism.

Scientific perspectivism was proposed "to develop an understanding of scientific claims that mediates between the strong objectivism of most scientist [...] and the constructivism found largely among historians and sociologists of science" (Giere, 2006, p. 3). It substitutes the question of *model* and *representation* in lieu and place of *theories* and *truth*. It first invalidates the question of the existence of the Elephant by moving the center of the debate on our empirical approach to the world, mediated by instruments measurements. Giere uses the example of Color vision. Color is perceived as the combination of certain wavelengths that are then transcoded in the retinal system by cone receptors. It cannot be explained from a complete objectivist point of view. Think for instance of the color of a red car. It is not an inherent property, it can change with the ambient lightning or the position of the viewer. One cannot say that it is completely subjective either, as a color still corresponds to a certain wavelength. Rather, Color vision can be well described by describing the asymmetrical interaction between the viewer and the object being observed, it is perspectival. Giere then moves on by elaborating on scientific instruments, showing that they, as

well "interact with only restricted aspects of the world" (p. 59) and that they "process input from the environment in ways peculiar to their own makeup" (p. 59). Likewise, scientific theorising can be seen as fundamentally perspectival. Giere compares scientific models to maps, they are designed to represent certain aspects of the structure of the world. Like a map, this correspondance only makes sense when we try to resolve it empirically, *e.g.* when we use the features represented on the map to find our way in town. Models do not carry an inherent *truth value*. Thus, the claims we can make about the world are always of a conditional form: "According to this highly confirmed theory (or reliable instrument) the world seems to be such and such" (p. 5-6). We understand also that if our monks would have had more of a perspectivist take, they would have not gotten into a fight about the status of the Elephant.

Likely, Van Fraassen (Van Fraassen, 1980), although he has a slightly diverging view on scientific theories, develops the concept of perspective in representations when he speaks about the theory of images (the *Bildtheorie*). Perspective is associated with painting and is the way one projects a 3D scene on a 2D painting. It relates to the content of the representation, and how the elements represented relate to it. Spatial perspective involves occlusion (revealing what things are like from one angle is incompatible with simultaneously revealing other parameters), grain (texture-fading), angle (limited range of depictions) and marginal distortions. Different perspectives can imply the account for some properties, but using one obliges not accounting for the other (one can see the interesting analogy with apparently contradicting experiments). That means that such perspectives cannot be simply combined in a third one. Again, those examples based on images, highlight that scientific work does not have to aim for an *absolute* explanation of reality. "Science aims to give us theories which are empirically adequate; and acceptance of a theory involves as belief only that it is empirically adequate. This is the statement of the anti-realist position I advocate; I shall call it constructive empiricism"(Van Fraassen, 1980, p. 10), repeated in (van Fraassen, 2008, p. 317).

Current scientific practice, which involves iterating over various approaches and techniques, can be well accounted for by perspectivism. This view moves away from the "flat-earth" take of system biology (one big exhaustive network) and accounts for the complex, multilevel, multiscale characteristics of biology (Callebaut, 2012). We understand that more data does not necessarily brings more truth. By interacting with omics data, we may recognise that, according to our current knowledge in molecular biology and our statistical analysis, "the world seems to be such and such". It helps put into context the promises and the limitations of the scientific practice with omics data, which is an empirical verification to validate model and hypothesis to fit with our measurements, based on a proper integrative agenda. It trivialises cases where different experiments produces observations that may not seem compatible, in fact according to our limited set of current perspectives. Further, the notion of *complexity* mentioned earlier is encompassed in our interactions with the biological system, and is part of the perspective too (it is only as complex as our instrument and knowledge allows us to describe it at this point in time).

2.1.4 An intermediary conclusion... and a perspective

Which of the questions asked in the beginning did we answer so far? We saw that the omics data, besides being an incredible technological revolution and operating at different scale, essentially brings the scientific methods and techniques we already knew. Further, even though omics experiments have the potential to provide new hypothesis and theories, those should not be overhyped until they can really be put in practice and their significance is confirmed after modelling and model

testing. We also showed that current feeling about Systems biology and the promises of omics data is overestimated because of the implicit objectivist stance of most scientists. Scientific perspectivism is a theory which allows to replace current promises and limitations with an empiricist take, where each experiment and line of evidence are considered together, with data generation and the underlying model. It is no surprise that current accounts of the vivid field of Data Science convey such a perspectivist take (Bourne, 2021).

Looking at recent developments, various ways of using omics data are showing integrated research agendas that are typically perspectivist. A first interesting technique is the use of randomisation (in the sequence space) as a core experimental principle of the experiment. Thus, it is alleviating some of the limitations related to descriptive assays. Among multiple examples, let us cite two recent ones that proposed to uncover the logic of gene regulation by generating millions of random promoter sequences and monitoring the resulting gene expression (de Boer et al., 2020), or probing mRNA binding energy by monitoring changes in the protein bound fraction (Smyth et al., 2015)). Although the assays are still heavily reductionist (the environment is simplified), it is possible, for instance for the first assay, to build, directly from the data, interpretable models that accurately predict the expression driven by the promoter sequences: empirical-omics at its best.
Another typical example is coined as multi omics data integration. It combines genome-scale metabolomic models with omics measurements and constraints on enzymatic reactions derived from the bibliography to generate and test hypotheses regarding typical reaction constants (Ebrahim et al., 2016; King et al., 2015). These methods can be sometimes advertised as solving the "Big Data to Knowledge" challenge. They are mainly providing a framework to integrate modelling with inductive descriptive science (lower left corner in figure 2.2).

The last example calls for the more generic question of what Big Data in biology means? Would it be different from omics data we already mentioned until now? Which type of knowledge is the processing of Big Data creating? Does it justify the creation of large data ressources, rather than pursuing a clear hypothesis driven agenda?

2.2 Is Big Data driven biology intelligible?

Until now, we considered the questions related to the accumulation, processing, and storage of raw molecular data (e.g. omics data), but Big Data in Biology can be considered in a broader epistemological framework. In fact, the very concept of Big Data emerged when digital data could be accumulated in large quantities (a huge `Volume`) and at unprecedented pace (a high `Velocity`). This pile of digital data can then be mined and interrogated using computational methods and modelling. The nature of omics is one of the premisses of Big Data in Biology [10]. However, the interpretation of what makes a large `Volume` or a high `Velocity` is often contested: those attributes cannot be defined on an absolute scale. They evolve constantly with technological advances: the news of James Watson's genome "sequenced at high speed" in 2008 (it took 4 months), sounds like a long time in view of the 24h needed for Illumina's Novaseq to sequence a handful of human genomes.

It is not only about having a *lot* of data *rapidly*, the integration of various sources of data (*e.g.* a

[10] If the term Big Data was coined in the nineties, let us remind that the accumulation of large quantities of data was a common feat in science for a long time. Naturalists in the XIXth century already had cope with the storage and organisation of large collections of data (Strasser, 2012).

diverse `Variety`) is often mentioned as an important additional ingredient (Sabina Leonelli, 2020). It is more about combining everything in a coherent framework and interrogating the resulting ressources in order to make new claims outside of their original scope. This is clear in biological research, where building databases/repository allows to access, store and analyse the large amount of data generated benefits a lot to a diverse array of research communities. The creation of model organisms databases, building on genome sequencing project, provides a good example. Typically, the data would be decontextualised –taken out of their native experimental and naming production territory, and then recontextualised –transported into a proper ontological framework together with metadata information (Sabina Leonelli, 2017). The first produced data consisted mainly in genomic sequences, associated to a significant curatorial work. The second wave of consortium efforts was engaged to broaden our view according to other dimensions of molecular biology. Those projects, financed in the last two decades, produced large `Volume` of raw data to deepen our understanding of model organisms through high throughput functional screens in normal and disease cells (TAGC, ENCODE, FANTOM projects), to increase our knowledge about populations (1000 or 1001 genome project), to fill in the tree of life (10k genome project, European Reference Genome Atlas), or to interrogate environmental and ecological data (Tara ocean, Metasub project). Data from those projects concerns primarily genomic and functional sequencing but annotated images or other molecular data can also be integrated. Model organism databases pose a challenging curatorial work: molecular data need to be linked and captured with an accurate representation of the complexity and the dynamics of an always evolving organism (Sabina Leonelli, 2017). We can retain the 3 following "`V`-words" to be associated with big-data: `Volume`, `Velocity`, and `Variety` [11].

It has to be stated that Big data is not only about generating a large amount of data and processing it into summary statistics! Data production is the result of a deliberate preparation, it goes beyond the choice of an instrument and the corresponding techniques needed to process the raw data. (Big) Data is more than the resulting text files and database schemas. Leonelli pointed out that data is curated and packaged before entering a database and used a metaphor of "data journeys" (Sabina Leonelli, 2017). Thus, processing and summaries are a first step in the journey that data is taking, in order to be further used for new purposes and by other research communities. To ensure a safe trip –one which will not deform/damage the information carried over by the data- this first step of data production, analysis and summary has to be circumscribed within a commonly agreed set of norms and standards. It reminds previous large scale data efforts of the XIX$^{\text{th}}$ century, which developed a common system of measures (the *mètre étalon* followed by the establishment of metrology institutes across Europe[12]). Such enterprises marked the transition of data to be considered a reusable asset, a commodity in digital form (Sabina Leonelli, 2019).
It is thus important to define standards in experimental techniques for data production and reproducible schemes for data processing. Likely, standardisation efforts emerged for high throughput data in biology, such as MAQC and SEQC for microarrays and sequencing data (Su et al., 2014), or the ChIP-Seq best practices within the ENCODE project (see chapter 1). It is still a challenging enterprise with "biological data on modern organisms[, they] are heterogeneous both in their content and in their format; [they] are curated and re-purposed to address the needs of highly disparate and

[11] other properties such as `Veracity`, `Value`, `Volatility` and `Validity` can also be considered, see (Sabina Leonelli, 2020), (Kitchin, 2014; Laney, 2001) or (Kitchin & McArdle, 2016) for a more in-depth definition

[12] the first *Physikalische-Technische-Reichsanstalt* was founded in Berlin in 1887, followed by Paris and London.

fragmented epistemic communities" (Sabina Leonelli, 2014). The other important development is about data naming: construct a common set of terms (e.g. an ontology).

Is it sufficient to produce those catalogues and maps, even though they are operating at an unprecedented scale? Critics emerged with the sequencing of the human genome: such a project is mainly descriptive and lacks a proper hypothesis to be tested. People argued that the amount of funding required is too high in comparison to the biological insight that would be obtained with a set of well designed HD experiments (R. Weinberg, 2010). The cartography efforts engulf a large amount of funding and is not necessarily paying off in terms of our understanding. Accumulation of omics experiments does not necessarily bring more than a large and very accurate list of molecular components (Stern, 2019). What is the point, when no proper question was asked? Indeed large scale projects bring a low amount of mechanistic insights. The debate around ENCODE come again to mind. This was also shown by Lopez-Rubio and Ratti when they perform a meta-analysis of the collection of articles from the TAGC project (López-Rubio & Ratti, 2021). Although justified to some extent, the properties of large scale data projects postulated by its critics may be inaccurate and too limiting. For instance a common perception is that the scope of such projects is limited to the creation of data warehouses and the establishment of data production standards. However, compiling experiments in a model organism database is a much more general task than simply building up data repositories. One could say that there are simply no "raw" data in model organism databases (Sabina Leonelli, 2017). Extending the scope to large scale projects generating omics data (e.g. ENCODE or TAGC), we could argue that the work involved goes beyond something purely descriptive. Each RNA-Seq experiment tests for instance thousands of –possibly weak- hypothesis at the same time. The data represented for visual exploration on the webservers already compiles results from those experiments, allowing the user to get his own opinion, but within the testing/data analysis framework of the consortium. This data packaging lays the ground for the next iterations of hypothesis and modelling from people interrogating the database with particular questions in mind, and the data is reused outside of its creation ground.

The epithets given to such projects –too descriptive, concerned with the production of lists, a mere fishing expedition- may also be too short sighted. The focus of the critics is more directed towards the publications accompanying the original data release and does not consider the lasting impact of the collection/taxonomic effort on Science. Let us remind that the first long lists of animals and plants compiled by taxonomists in the modern period paved the way for the unveiling of the theory of evolution and natural selection. The first technological advances in molecular biology led us to the validation of synthetic evolution. We are now occupied with a greater task of quantifying the organisation and the action of living systems at the molecular level and are therefore coping with higher levels of complexity. Such projects highlighted the limits and the advantages of our current biological models.

So it is not anymore really a question of knowing if it is "good" or "bad", it is already an important part of research and has motivated the construction of large data infrastructures such as the European Bioinformatics Institute. It is however important to keep in mind that one cannot generate a lot of data just for the sake of it. Those efforts should be done with different goals in mind to realise large-scale project data as an asset. Again there must be a research agenda. Also the FAIR principles for data sharing come to mind.

The question seems rather to be able to understand the relation that can be drawn between

Big Data and knowledge. How far can Big Data be a mean to the automated generation of scientific knowledge? In a very influential presentation in 2007, Jim Gray claimed that we entered a fourth paradigm, in which we rely more and more heavily on data-intensive work and computational ressources when doing scientific research. The *Data-intensive science* paradigm follows the three first paradigms of experimental evidence, of theoretical science and of modelling/computation, and can be summarised in two core properties (S. Leonelli, 2012): "One is the intuition that induction from existing data is vindicated as a crucial form of scientific inference, which can guide and inform experimental research, and the other is the central role of machines, and thus of automated reasoning, in extracting meaningful patterns from data". Meaningful signal is thus detected by looking for associations between the different dimensions in the data, *e. g.* by relying heavily on correlations to draw conclusions (Mayer-Schönberger & Cukier, 2013). Although traditionally seen as a drawback in classical H-D method, correlation and induction were epitomised as an important feature of intensive data-science, leading to extreme claims such as "correlation supersedes causation" and a prophetized "end of the theory" (Anderson, 2008). Although this model agnostic view carries a simplistic discourse on intensive data science, it raises interesting questions about the epistemic scope of notions such as knowledge or understanding.

Machine Learning (ML) provides an environment for *generalising* predictions on a target variable y (a class, a real value), given a set $\mathbf{x}$ of covariates that are observed. ML operates inductively by starting from a *training set* to learn the machine parameters and evaluate its performance, and then apply it to new data. The basic goal is to either provide answers on a specifically stated problem by predicting a response, or to detect structures and patterns in the data –usually by means of correlations. The application of ML to the Life Sciences has already shown successes, for instance in image classification (CT-scans, tumor grading). Interestingly in the last decade the practice of ML has shifted hands, from the development of optimisation techniques for specific problems, to the generic design of machine architecture by the data practitioners (and the emergence of data science). More recently, the use of ML went a step forward on the problem of predicting the fold of a protein from its amino acid sequence. The deep learning machine AlphaFold2 could provide predictions at an accuracy on par with experimental techniques (Jumper et al., 2021), making an incredible leap forward to the whole protein structure field. It is interesting to note that the problem was depicted as being "solved" in the press [13]. However "Dramatic ability to predict is not the same as explanatory understanding" (Nickles, 2018), and we could argue that the *know-how* to provide a very accurate answer to a problem is not the same as to *know-that* a model can explain most of the phenomenon under scrutiny. From the classical theory-centric view of science, we build knowledge through claims we can make about the world. We publish those claims in books and journals and this add it to the corpus of science. Conversely, machines only give very accurate answers, but usually without justifications. If we are to rely more and more on those systems, "Generative insight would be replaced by black box consequentialism." (Nickles, 2018).

This tension between *knowledge as an ability to predict* against *knowledge as a way to describe the world* adds interesting questions for our everyday scientific practice (Weinberger, 2017): "We thought knowledge was about finding the order hidden in the chaos. We thought it was about simplifying the world. It looks like we were wrong. Knowing the world may require giving up on understanding

[13] for instance science magazine titled on 2020-11-30: 'The game has changed'. AI triumphs at solving protein structures

it". Indeed expert AI models are mainly black-boxes, and it is simply not possible to comprehend the model built derive principle by "looking at" the model (for instance the parameters of the fifteen AF2 models necessitate around 3.5GB of disk space). Ironically, the more we "know-how" to tackle complex problem with ML, the less we understand about those systems, or how they function as black boxes. This issue was identified by the ML communities, and the concept of eXplainable AI (systems which can describe features responsible for a prediction) became more and more prominent in the last years and has seen multiple developments. Indeed, data-models developed have to be accountable for their choice, in order to be used into the decision process in health or policy, and to debunk possible hidden biases that could arise from the data. But XAI analysis are *postmortem*, they are performed given a set of datapoints.

This polarity between knowing and understanding also brings back consideration on the validity of epistemological reductionism (see 2.1.2). Indeed, one exemplary reductionistic idea was Anfinsen's dogma/the thermodynamics hypothesis (the structure of a protein is determined by the physical properties of its sequence alone). It is clear now that, although protein structure can be accurately predicted from the amino acid sequence, physical models are not the solution, and only ML managed to crush the problem (at least its first step). This shift in perspective has even been more pregnant in the domain of Natural Language Processing, where translations based on syntactic models showed their intrinsic limitations as soon as simple, word counts based system started leveraging terascale corpus data (Nickles, 2018). In fact, there was a significant misunderstanding hidden in the western science approach to knowledge. We thought that because models, *e.g.* simple claims about the world were a beautiful and practical tool to describe it, knowledge should also be like that. However, the models we build are just a representation of the world, and there is no reason for reality to be simple nor beautiful. There may not even be any truth that is attainable (remember the elephant in section 2.1.3). It does not have to worry us if we just want to obtain solutions to each instance. Although it is alien to our understanding or cognition (Nickles, 2018; Weinberger, 2017) this "know-how" also counts as knowledge. But if we want to understand a problem from our limited human perspective, a wrong simple model will still be more useful.

Even though the system constructed is not intelligible, it does not mean it is completely theory free. Indeed, data-intensive science is in a sense *agnostic to science*, and the main problem is automated curve fitting. However, data accumulation is not theory free, but done with a set of implicit questions in mind (even if this concerns a lot of Volume and at a high scale). As demonstrated previously, it needs to be repurposed to travel to other applications. The use of omics data, being usually an application of experimental reductionism, is indeed loaded with theories about gene regulations. Wolfgang Pietsch goes further in the analysis, and uses classificatory trees and non parametric regression as examples to argue further that in those cases ML techniques are theory-laden in an external sense (for the factors outside the phenomenon) but that the internal connection between variables are largely theory-free (Pietsch, 2015). Although we do not think that this separation is always so clear cut for ML (we presented multiple hierarchical models previously where the variables where connected in a principled way).

Likewise, owing to the descriptive nature of Big Data, causal relationships are also commonly thought to be unattainable. As written previously, Inductive reasoning is used recurrently. This is based on the idea that data is present in such `Volume` and `Variety`, it can exhaustively covers all the cases of interest ($n = all$). Roughly sketched, we see we can apply eliminative induction in

use to deduce causal relationships (as when promoter activity is summarised using a mutagenesis experiment, see 2.1.4). Although a very simple type of causality is inferred (the type of influence of a variable on another is not specified, etc.), Pietsch gave a very detailed account, which demonstrated that as soon as some properties of data intensive science are fulfilled, eliminative induction can be used (Pietsch, 2016). AI model can extract some form of causal relationship, and are to some extent theory-laden, but those same models are "flat" and do not describe a hierarchy, a more diagrammatic understanding of the elements. In this case, simple intelligible human models are on the contrary more practical. So AI can find links, even prove them, but those links are rough indication, and they will never be "novel" (Ratti, 2020).

What is the progress we can envision, the effect of ML and data intensive science, if models are ultra accurate but not intelligible? Are we abandon the classical goal of science, making claims about the world? We can turn back to the perspectivist take on representations. We are already using books, diagrams, or computer systems as cognitive extension devices to help and guide our interpretation of phenomenon. In particular, advanced data visualisation techniques have proved very useful. Beyond the success of AlphaFold2, end-to-end learning, where ML is trained to predict a complex output, is an efficient way of returning representations. Is it foreseeable that instead of providing intelligible models, we could teach machine to construct intelligible representations? Attaining this goal, at least from a perspectivist point of view, could be enough: according to this highly accurate and almost exhaustively trained machine, we can provide a representation of the world as such and such.

References

Allen, J. F. (2001a). Bioinformatics and discovery: Induction beckons again. *Bioessays*, *23*(1), 104–107. doi:https://doi.org/10.1002/1521-1878(200101)23:1<104::AID-BIES1013>3.0.CO;2-2

Allen, J. F. (2001b). In silico veritas. data-mining and automated discovery: The truth is in there. *EMBO Rep*, *2*(7), 542–544. doi:10.1093/embo-reports/kve139

Anderson, C. (2008). *The end of theory: The data deluge makes the scientific method obsolete.*

Anonymous. (2000). Can biological phenomena be understood by humans? *Nature*, *403*(6768), 345–345. doi:10.1038/35000353

Bartel, D. P. (2009). MicroRNAs: target recognition and regulatory functions. *Cell*, *136*(2), 215–33. doi:10.1016/j.cell.2009.01.002

Beard, D. A., & Kushmerick, M. J. (2009). Strong Inference for Systems Biology. *PLOS Computational Biology*, *5*(8), e1000459–.

Bourne, P. E. (2021). Is "bioinformatics" dead? *PLOS Biology*, *19*(3), 1–3. doi:10.1371/journal.pbio.3001165

Brigandt, I., & Love, A. (2017). Reductionism in biology. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Spring 2017). Metaphysics Research Lab, Stanford University.

Brown, P. O., & Botstein, D. (1999). Exploring the new world of the genome with DNA microarrays. *Nat Genet*, *21*(1 Suppl), 33–37. doi:10.1038/4462

Callebaut, W. (2012). Scientific perspectivism: A philosopher of science's response to the challenge of big data biology. *Stud Hist Philos Biol Biomed Sci*, *43*(1), 69–80. doi:10.1016/j.shpsc.2011.10.007

Casadevall, A., & Fang, F. C. (2008). Descriptive science. *Infect Immun*, *76*(9), 3835–3836. doi:10.1128/IAI.00743-08

Casadevall, A., Fang, F. C., & Pirofski, L.-a. (2011). Microbial virulence as an emergent property: Consequences and opportunities. *PLOS Pathogens*, *7*(7), 1–3. doi:10.1371/journal.ppat.1002136

Crick, F. (1966). *Of molecules and men*. John Danz lectures. University of Washington Press.

Dawkins, R. (1986). *The blind watchmaker*. W.W Norton.

de Boer, C. G., Vaishnav, E. D., Sadeh, R., Abeyta, E. L., Friedman, N., & Regev, A. (2020). Deciphering eukaryotic gene-regulatory logic with 100 million random promoters. *Nature Biotechnology*, *38*(1), 56–65. doi:10.1038/s41587-019-0315-8

Descartes, R. (1637/1850). *Discourse on the method of rightly conducting the reason, and seeking truth in the sciences.* Translated by John Veitch. Edinburgh: Sutherland and Knox.

Doolittle, W. F. (2013). Is junk dna bunk? a critique of ENCODE. *Proceedings of the National Academy of Sciences*, *110*(14), 5294–5300. doi:10.1073/pnas.1221376110. eprint: https://www.pnas.org/content/110/14/5294.full.pdf

Doyle, C. (1890). *The sign of the four*. Penguin.

Ebrahim, A., Brunk, E., Tan, J., O'Brien, E. J., Kim, D., Szubin, R., ... Palsson, B. O. (2016). Multi-omic data integration enables discovery of hidden biological regularities. *Nature Communications*, *7*(1), 13091. doi:10.1038/ncomms13091

Elliott, K. C., Cheruvelil, K. S., Montgomery, G. M., & Soranno, P. A. (2016). Conceptions of good science in our data-rich world. *Bioscience*, *66*(10), 880–889. doi:10.1093/biosci/biw115

ENCODE Project, C. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, *489*(7414), 57–74. doi:10.1038/nature11247

Fang, F. C., & Casadevall, A. (2011). Reductionistic and holistic science. *Infect Immun*, *79*(4), 1401–1404. doi:10.1128/IAI.01343-10

Giere, R. N. (2006). *Scientific perspectivism*. University of Chicago Press.

Golub, T. (2010). Counterpoint: Data first. *Nature*, *464*, 679 EP -.

Graur, D., Zheng, Y., Price, N., Azevedo, R. B., Zufall, R. A., & Elhaik, E. (2013). On the Immortality of Television Sets: Function in the Human Genome According to the Evolution-Free Gospel of ENCODE. *Genome Biology and Evolution*, *5*(3), 578–590. doi:10.1093/gbe/evt028. eprint: https://academic.oup.com/gbe/article-pdf/5/3/578/1567877/evt028.pdf

Heine, S. J. (2017). *DNA is not destiny : The remarkable, completely misunderstood relationship between you and your genes*.

Ingolia, N. T., Lareau, L. F., & Weissman, J. S. (2011). Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell*, *147*(4), 789–802. doi:10.1016/j.cell.2011.10.002

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., ... Hassabis, D. (2021). Highly accurate protein structure prediction with alphafold. *Nature*, *596*(7873), 583–589. doi:10.1038/s41586-021-03819-2

Kaiser, M. I. (2011). The limits of reductionism in the life sciences. *Hist Philos Life Sci*, *33*(4), 453–476.

Kell, D. B., & Oliver, S. G. (2004). Here is the evidence, now what is the hypothesis? the complementary roles of inductive and hypothesis-driven science in the post-genomic era. *Bioessays*, *26*(1), 99–105. doi:10.1002/bies.10385

King, Z. A., Lu, J., Dräger, A., Miller, P., Federowicz, S., Lerman, J. A., ... Lewis, N. E. (2015). BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. *Nucleic Acids Research*, *44*(D1), D515–D522. doi:10.1093/nar/gkv1049. eprint: https://academic.oup.com/nar/article-pdf/44/D1/D515/16661243/gkv1049.pdf

Kitano, H. (2002). Systems biology: A brief overview. *Science*, *295*(5560), 1662–1664. doi:10.1126/science.1069492

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. *Big Data & Society*, *1*(1), 2053951714528481. doi:10.1177/2053951714528481. eprint: https://doi.org/10.1177/2053951714528481

Kitchin, R., & McArdle, G. (2016). What makes Big Data, Big Data? exploring the ontological characteristics of 26 datasets. *Big Data & Society*, *3*(1), 2053951716631130. doi:10.1177/2053951716631130. eprint: https://doi.org/10.1177/2053951716631130

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., ... International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*, 860–921.

Laney, D. (2001). *3D data management: Controlling data volume, velocity, and variety*. META Group.

Leonelli, S. [S.]. (2012). Introduction: Making sense of data-driven research in the biological and biomedical sciences. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *43*(1), 1–3. Data-Driven Research in the Biological and Biomedical Sciences On Nature and Normativity: Normativity, Teleology, and Mechanism in Biological Explanation. doi:https://doi.org/10.1016/j.shpsc.2011.10.001

Leonelli, S. [Sabina]. (2014). What difference does quantity make? on the epistemology of big data in biology. *Big Data Soc*, *1*(1). doi:10.1177/2053951714534395

Leonelli, S. [Sabina]. (2017). Data-centric biology : A philosophical study. Retrieved from https://login.proxy.bib.uottawa.ca/login?url=http://chicago.universitypressscholarship.com/view/10.7208/chicago/9780226416502.001.0001/upso-9780226416335

Leonelli, S. [Sabina]. (2019). The challenges of big data biology. *Elife*, *8*. doi:10.7554/eLife.47381

Leonelli, S. [Sabina]. (2020). Scientific Research and Big Data. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Summer 2020). Metaphysics Research Lab, Stanford University.

López-Rubio, E., & Ratti, E. (2021). Data science and molecular biology: Prediction and mechanistic explanation. *Synthese*, *198*(4), 3131–3156. doi:10.1007/s11229-019-02271-0

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*.

Mayr, E. (1988). The limits of reductionism. *Nature*, *331*(6156), 475–475. doi:10.1038/331475a0

Mazzocchi, F. (2008). Complexity in biology. *EMBO reports*, *9*(1), 10–14. doi:10.1038/sj.embor.7401147

Müller-Wille, S., & Charmantier, I. (2012). Natural history and information overload: The case of linnaeus. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *43*(1), 4–15. Data-Driven Research in the Biological and Biomedical Sciences On Nature and Normativity: Normativity, Teleology, and Mechanism in Biological Explanation. doi:https://doi.org/10.1016/j.shpsc.2011.10.021

Nickles, T. (2018). Alien reasoning: Is a major change in scientific research underway? *Topoi*. doi:10.1007/s11245-018-9557-1

O'Malley, M. A., Elliott, K. C., & Burian, R. M. (2010). From genetic to genomic regulation: Iterativity in microrna research. *Stud Hist Philos Biol Biomed Sci*, *41*(4), 407–417. doi:10.1016/j.shpsc.2010.10.011

Pietsch, W. (2015). Aspects of theory-ladenness in data-intensive science. *Philosophy of Science*, *82*(5), 905–916. doi:10.1086/683328

Pietsch, W. (2016). The causal nature of modeling with big data. *Philosophy & Technology*, *29*(2), 137–171. doi:10.1007/s13347-015-0202-2

Platt, J. R. (1964). Strong inference. *Science*, *146*(3642), 347–353. doi:10.1126/science.146.3642.347. eprint: https://science.sciencemag.org/content/146/3642/347.full.pdf

Poon, L. L. M., Song, T., Rosenfeld, R., Lin, X., Rogers, M. B., Zhou, B., ... Ghedin, E. (2016). Quantifying influenza virus diversity and transmission in humans. *Nature Genetics*, *48*(2), 195–200. doi:10.1038/ng.3479

Popper, K. (1959). *The logic of scientific discovery*. Routledge.

Ratti, E. (2020). What kind of novelties can machine learning possibly generate? the case of genomics. *Studies in History and Philosophy of Science Part A*, *83*, 86–96. doi:https://doi.org/10.1016/j.shpsa.2020.04.001

Regenmortel, M. H. V. V. (2004). Reductionism and complexity in molecular biology. *EMBO reports*, *5*(11), 1016–1020. doi:10.1038/sj.embor.7400284. eprint: https://www.embopress.org/doi/pdf/10.1038/sj.embor.7400284

Richardson, S. S., & Stevens, H. (Eds.). (2015). *Postgenomics: Perspectives on biology after the genome*. Duke University Press.

Shapin, S. (1982). History of science and its sociological reconstructions. *History of Science*, *20*(3), 157–211. doi:10.1177/007327538202000301. eprint: https://doi.org/10.1177/007327538202000301

Smyth, R. P., Despons, L., Huili, G., Bernacchi, S., Hijnen, M., Mak, J., ... Marquet, R. (2015). Mutational interference mapping experiment (MIME) for studying RNA structure and function. *Nature Methods*, *12*(9), 866–872. doi:10.1038/nmeth.3490

Stern, C. D. (2019). The 'Omics revolution: How an obsession with compiling lists is threatening the ancient art of experimental design. *Bioessays*, *41*(12), e1900168. doi:10.1002/bies.201900168

Strasser, B. J. (2012). Data-driven sciences: From wonder cabinets to electronic databases. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *43*(1), 85–87. Data-Driven Research in the Biological and Biomedical Sciences On Nature and Normativity: Normativity, Teleology, and Mechanism in Biological Explanation. doi:https://doi.org/10.1016/j.shpsc.2011.10.009

Su, Z., Łabaj, P., Li, S., Thierry-Mieg, J., Thierry-Mieg, D., Shi, W., ... Consortium, S.-I. (2014). A comprehensive assessment of rna-seq accuracy, reproducibility and information content by the sequencing quality control consortium. *Nature Biotechnology*, *32*(9), 903–914. doi:10.1038/nbt.2957

Van Fraassen, B. C. (1980). *The scientific image* (1st ed.) (O. U. Press., Ed.). Oxford University Press.

van Fraassen, B. C. (2008). *Scientific representation: Paradoxes of perspective*. Oxford University Press.

Voit, E. O. (2019). Perspective: Dimensions of the scientific method. *PLoS Comput Biol*, *15*(9), e1007279. doi:10.1371/journal.pcbi.1007279

Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., ... Burge, C. B. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature*, *456*(7221), 470–476.

Weinberg, R. (2010). Point: Hypotheses first. *Nature*, *464*, 678 EP -.

Weinberg, S. (1987). Newtonianism, reductionism and the art of congressional testimony. *Nature*, *330*(6147), 433–437. doi:10.1038/330433a0

Weinberg, S. (2001). 9. physics and history. In J. A. Labinger & H. Collins (Eds.), (pp. 116–127). doi:doi:10.7208/9780226467245-010

Weinberger, D. (2017). *Our machines now have knowledge we'll never understand.*

Woese, C. R. (2004). A new biology for a new century. *Microbiol Mol Biol Rev*, *68*(2), 173–186. doi:10.1128/MMBR.68.2.173-186.2004

Xue, K. S., & Bloom, J. D. (2019). Reconciling disparate estimates of viral genetic diversity during human influenza infections. *Nature Genetics*, *51*(9), 1298–1301. doi:10.1038/s41588-019-0349-3

Curriculum Vitae

Hugues Richard

44 years old, French, married
RichardH@rki.de
hugues.richard@sorbonne-universite.fr

Robert Koch Institute, Nordufer 20, 13353 Berlin
Work : +49-(0)-30-1875-45039
http://www.lcqb.upmc.fr/hrichard

Working Experience

Since 2019	**Senior Researcher**, Bioinformatics Unit (MF1) Robert Koch Institute, Berlin.
Since 2009	**Associate Professor in Computational Biology and Statistics**, (on leave since 2019) Analytical Genomics team, Lab. of Computational and Quantitative Biology, UMR7238 Computer Science and Engineering Department, Sorbonne Université, Faculty of Science - Pierre & Marie Curie Campus, Paris.
2006-2009	**Post-doctoral Fellowship**, Computational Molecular Biology Dept Max Planck Institute for Molecular Genetic, Berlin.
2001-2006	**Tutor (01-04) and Lecturer (04-06) in Statistics and Bioinformatics** Statistic & Genomes Laboratory, University Evry Val d'Essonne.

Education

2001-2005	**Ph. D., Computational Biology** - *Prediction of the Subcellular Localization of Proteins by their Biological Sequences*, - University Evry Val d'Essonne - France.
2000-2001	**Master's Degree in Mathematics: Analysis and Stochastic Processes** *Marne-la-Vallée University* - France, with recognition.

Awards & Grants

2018-2021	*MASSIV: Modeling Alternative Splicing and its Structural Impact during eVolution*, with E. Laine (ANR JCJC 205k€).
2016-2020	*Dynamic Crosstalks in the Development of Acute Myeloid Leukemias in their Ecosystem*, PI: D. Salort (LCQB) (Plan Cancer, HTE program 110k€ for LCQB, 500k€ total)
2010-2012	**Leave of teaching** (délégation CNRS) at LCQB (UMR 7238)
2012	**Japan Society for the Promotion of Science** (6 months) Computational Biology Research Center, Tokyo, Japan. Hosted by M. Frith.
Fall 2011	**IPAM long research program** (3 months) *Mathematical and Computational Approaches in High-Throughput Genomics*, Institute for Pure and Applied Mathematics, UCLA, USA.
2006-2009	**Max Planck post-doctoral fellowship**.

Analysis of High Throughput Sequencing data

Genomics	Joint Split-Alignement of sequence reads for the annotation of **Structural Variants** (A. M. Shrestha, Asai, Frith, & **Richard**, 2018) Annotation of **low frequency Structural Variants** from Mate-Pair libraries (Gillet-Markowska, **Richard**, Fischer, & Lafontaine, 2015) Fiona, automatic **read error correction** (Schulz* et al., 2014; David Weese, Schulz, & **Richard**, 2017) for genome sequencing experiments **Allogenomics**, prediction of chronic graft lost from exome sequencing data L. Mesnard MD, PhD, Tenon Hospital, Paris
Transcriptomics	Annotation of **Gene model** and structure (Mirauta, Nicolas*, & **Richard***†, 2013, 2014; Warren* et al., 2010), of **Alternative Splicing** Events (**Richard*** et al., 2010; Sultan* et al., 2008), and small RNAs analysis (Rogato* et al., 2014). Analysis of iCLIP/eCLIP Seq data (Krakau, **Richard**†, & Marsico†, 2017). P. Nicolas, INRA, Jouy-en-Josas. **MASSIV** project: assessing the structural impact of alternative splicing during evolution (**Adelphylosofs**; Zea, Laskina, Baudin, **Richard**†, & Laine†, 2021). E. Laine and D. Zea, LCQB, J. Roux, Univ. Basel, Basel.
Metagenomics	**Metasub** Consortium: Metagenomics and Metadesign of Subways and Urban Biomes Christopher E. Mason, Cornell University, New-York.

Native tongue, Fluent, Advanced, Beginner

Research Areas:

- Statistical methods for high-throughput sequencing data:
  - RNA-Seq and transcriptome analysis (Krakau et al., 2017; Mäder, Nicolas, **Richard**, Bessières, & Aymerich, 2011; Mirauta et al., 2013, 2014; **Richard*** et al., 2010; T. Steijger et al., 2013; Sultan* et al., 2008; Warren* et al., 2010),
  - Non Coding RNA and gene model annotation (Rogato* et al., 2014; Warren* et al., 2010)
  - Sequence alignment, detection of gene fusion or genomic variation (Gillet-Markowska et al., 2015; Hao Hu et al., 2010; A. M. Shrestha et al., 2018).
  - Automatic correction of sequencing errors (Schulz* et al., 2014; David Weese et al., 2017).
- Word statistics and Markov chains, Hidden Markov Models (Miele, Bourguignon, Robelin, Nuel, & **Richard**, 2005; **Richard** & Nuel, 2003; Robelin, **Richard**, & Prum, 2003; Robin, Daudin, **Richard**, Sagot, & Schbath, 2002)
- Classification, Protein annotation [31,32]

Software development:

- 9 software tools (Ait-hamlat et al., 2020; Gillet-Markowska et al., 2015; Krakau et al., 2017; Mirauta et al., 2013, 2014; Schulz* et al., 2014; A. M. Shrestha et al., 2018; Warren* et al., 2010; Zea et al., 2021) and 4 web services (**Richard** & Nuel, 2003; Robelin et al., 2003), Ases webserver
- 1 processing pipeline (H. Hu et al., 2015; Hao Hu et al., 2010), 2 librairies (R and C++) (Miele et al., 2005; **Richard*** et al., 2010).

Invited Presentations

[1] Keynote speaker, Polish Bioinformatics Society, 15-17 September 2021.
[2] Keynote speaker, Advanced Genome Science International Symposium, Tokyo University, 10-11 January 2017, http://pags2017.genome-sci.jp/program.html.
[3] Invited lecture, Analysis of Tumoral genome school, Seine-Port, 12-15 May 2014 – *Methods for the analysis of Transcriptome Sequencing Data.*
[4] Invited speaker, RNA-Seq Europe, 2013, Basel, 3-5 December 2013, http://rnaseq-europe.com/speakers – *Beyond gene expression: estimate expression levels and detect transcript boundaries from RNA-seq read counts.*
[5] Invited speaker, 21st International Symposium on Mathematical Programming (ISMP2012), Berlin, 19-24 August 2012 – *Fiona: Automatic correction of sequencing errors in genome sequencing experiments.*
[6] Keynote speaker, Next Generation Sequencing Workshop, organized by IBBE-CNR and Bari University, Bari, 6-8 October 2010, http://mi.caspur.it/workshop_NGS10/ – *Methods for the analysis of RNA-seq data.*
[7] Invited speaker, Workshop on Bioinformatics and High throughput sequencing, organized the ReNaBi network, Paris, 24 March 2010 – *Methods for the analysis of RNA-seq data.*

Conferences with Proceedings (selection)

[8] Mirauta, B., Nicolas, P. **Richard, H**. A Sequential Monte Carlo method for estimating transcriptional landscape at base pair level from RNA-Seq data, Journées Ouvertes en Biologie, Informatique

et Mathématiques (JOBIM), Institut Pasteur, Paris, 28 June- 1st July 2011.
[9] **Richard, H.**, Mucchielli M., Prum B., Képès F., *Hidden Markov Models Hierarchical Classification for Ab-Initio Prediction of Protein Subcellular Localization*, ISMB'05, Detroit, June 2005, PLoS CB poster.
[10] **Richard, H.**, Mucchielli M., Prum B., Képès F. , *Discrimination of the subcellular locations of the yeast proteins by their biological sequence*, JOBIM'04, Montréal, june 2004,n°79.

Communications

[11] Shrestha, AMS, Asai K., Frith M., and **Richard, H.** A new framework for the identification of genomic structural variant using joint alignment of reads, Statistical Methods for Post-Genomic Data, (SMPGD), Lille, 11-12 february 2016.
[12] **Richard, H.**, Weese, D., Holtgrewe, M., Schultz, M. Fiona: A tool for automatic correction of sequencing errors in genome sequencing experiments, 22nd Annual Workshop on Mathematical and Statistical Aspects of Molecular Biology, (MASAMB), Berlin, 10-11 april 2012.
[13] Mirauta, B., Nicolas, P., **Richard, H.** Sequential Monte Carlo - Particle Gibbs inference of Transcriptional Landscape from RNA-Seq Data, Mathematical and Statistical Aspects of Molecular Biology (MASAMB), Berlin, 10-11 april 2012.

Invited Seminars

- Tokyo University 05.26.2017, and Keio University 05.12.2017, Tokyo,
- Ho Chi Minh city University of Technology, Ho Chi Minh City, 07.14.2016,
- CriStaL laboratory, Lille University, Lille, 04.27.2016,
- Lake Arrowhead, Institute for Pure and Applied Mathematics, Los Angeles, 06.11.2014.
- Institute for Cell Biology, UnikKlinik, RWTH, Aachen, 08.04.2014.
- Computational Biology Research Center (CBRC), Tokyo, 02.23.2012.
- Génoscope, Evry, 03.26.2009 – Lab. d'Informatique, de Robotique et de Microélectronique de Montpellier (LIRMM), Montpellier, 01.27.2009 – – Laboratoire Biométrie et Biologie Evolutive (LBBE), Lyon, 01.22.2009. – University College Dublin (UCD), Dublin, 11.21.2008.

Service

- **Referee** for the international scientific journals: Genome Biology, Nucleic Acids Research, Bioinformatics, BMC genomics, BMC bioinformatics, Plos One.
- **Organization of conference**: Statistical Methods for Post-Genomic Data (SMPGD), an international conference which aims at gathering statisticians, computer scientists, and biologists to discuss new statistical methodologies for the analysis of high throughput biological data (http://smpgd2014.sciencesconf.org/ tenth edition, with more than 100 participants)
- **Program Commitee**: ECCB 2016 (La Hague, Netherland), JOBIM2017 (Lille, France).
- **Japan Society for the promotion of Science alumni (JSPS)**: Organization of information meeting for promoting the JSPS program (2016).

Students supervision

Graduate students

- (2017- 2021) Shared PhD supervision of L. David (Ministry of research fellowship), on the functional annotation of large scale metagenomic data.
- (2014- 2018) Shared PhD supervision of S. Krakau, (International Max Planck research School) on methods for predicting RNA-protein interaction from CLIP-Seq data.
- (2010- Dec. 2014) Shared PhD supervision of B. Mirauta (Ministry of research fellowship), on the development of method estimating transcription rate from RNA-Seq data at the basepair level.
- (2007 to 2010) Shared PhD supervision of M. H Schulz, (International Max Planck research School) on a method for the detection and quantification of alternative splicing event on RNA-Seq data.

Master Students

- (March-September 2021), joint with E. Laine: V. Lombard on alternative splicing inspired protein design.
- (March-September 2020), joint with L. Mesnard: P. Delaugère on the genotyping of the MUC1 gene using k-mers.
- (January-July 2020), joint with L. Mesnard: A. Hamza on the analysis of structural variant detection on triplets.
- (February-July 2019), joint with L. Mesnard: M. Sentucq on benchmarking of structural variant detection using
- (February-July 2017), joint with A. Carbone and R. Vicedomini: L. David on targeted assembly of metagenomics data.
- (February-June 2017), joint with L. Mesnard: R. Clerc on the refinement of the allogenomics score for the prediction of renal graft chronic failure.
- (April-September 2014), joint with E. Laine: A. Ait-Amlat on the phylogenetic reconstruction of transcript isoforms evolution
- (April-September 2014), M. Bessoul on the analysis of the periodicity in the small RNA profiles of the diatom *P. tricornutum*.
- (February-August 2013), M. Bessoul, on the development of a method for the inference of methylation levels from bisulfite sequencing experiments.
- (April-September 2010), B. Mirauta which followed with a PhD thesis (see above)
- (Since 2010), I also supervised multiple research projects with 1st year master students. Below is a selection of the subjects: Analysis of bisulfite sequencing data, Correcting sequencing errors by a suffix tree approach, Annotating proteins with variable order markov chains, Evaluating various segmentation methods for the analysis of transcriptome data.

Teaching

I give below the broad themes corresponding for the various courses I taught in.

Bioinformatics:	Next generation sequence data analysis	bachelor & master
	sequence analysis and word statistics.	"
	Microarray/RNA-Seq analysis.	"
	Phylogeny.	"
Statistics:	Multivariate data analysis and Hypothesis testing.	bachelor
	Classification and Pattern Matching.	bachelor & master
	Markovian models	"
Computer Science:	Algorithms.	bachelor
	Programming (C & C++).	"
	Discrete structures	"

List of publications

Abdollahi, N., Albani, A., Anthony, E., Baud, A., Cardon, M., Clerc, R., ... Lopes, A. (2018). Meet-u: Educating through research immersion. *PLOS Computational Biology*, *14*(3), 1–10. doi:10.1371/journal.pcbi.1005992

Ait-hamlat, A., Zea, D. J., Labeeuw, A., Polit, L., **Richard**†, H., & Laine†, E. (2020). Transcripts' evolutionary history and structural dynamics give mechanistic insights into the functional diversity of the jnk family. *Journal of Molecular Biology*, *432*(7), 2121–2140. doi:https://doi.org/10.1016/j.jmb.2020.01.032

Calvignac-Spencer, S., Budt, M., Huska, M., **Richard**, H., Leipold, L., Grabenhenrich, L., ... Hölzer, M. (2021). Rise and fall of sars-cov-2 lineage a.27 in germany. *Viruses*, *13*(8). doi:10.3390/v13081491

Danko*, D., Bezdan*, D., ... (73 authors), **Richard**, H., ... (22 authors), Mason, C. E., & The International MetaSUB consortium. (2021). A global metagenomic map of urban microbiomes and antimicrobial resistance. *Cell*, *184*(13), 3376–3393.e17. doi:https://doi.org/10.1016/j.cell.2021.05.002

David, L., Vicedomini, R., **Richard**†, H., & Carbone†, A. (2020). Targeted domain assembly for fast functional profiling of metagenomic datasets with S3A. *Bioinformatics*, *36*(13), 3975–3981. doi:10.1093/bioinformatics/btaa272. eprint: https://academic.oup.com/bioinformatics/article-pdf/36/13/3975/33459041/btaa272.pdf

Gillet-Markowska, A., **Richard**, H., Fischer, G., & Lafontaine, I. (2015). Ulysses: Accurate detection of low-frequency structural variations in large insert-size sequencing libraries. *Bioinformatics*, *31*(6), 801–808. doi:10.1093/bioinformatics/btu730. eprint: http://bioinformatics.oxfordjournals.org/content/31/6/801.full.pdf+html

Hu, H. [H.], Haas, S. A., Chelly, J., Van Esch, H., Raynaud, M., de Brouwer, A. P. M., ... Kalscheuer, V. M. (2015). X-exome sequencing of 405 unresolved families identifies seven novel intellectual disability genes. *Molecular Psychiatry*, *aop*(current). doi:10.1038/mp.2014.193

Hu, H. [Hao], Wrogemann, K., Kalscheuer, V., Tzschach, A., **Richard**, H., Haas, S. A., ... Chen, W. (2010). Mutation screening in 86 known X-linked mental retardation genes by droplet-based

multiplex PCR and massive parallel sequencing. *The HUGO Journal*. doi:10.1007/s11568-010-9137-y

Krakau, S., **Richard**†, H., & Marsico†, A. (2017). PureCLIP: Capturing target-specific protein-RNA interaction footprints from single-nucleotide CLIP-seq data. *Genome Biology*, *18*(1), 240. doi:doi:10.1186/s13059-017-1364-2

Mäder, U., Nicolas, P., **Richard**, H., Bessières, P., & Aymerich, S. (2011). Comprehensive identification and quantification of microbial transcriptomes by genome-wide unbiased methods. *Current Opinion in Biotechnology*, *22*(1), 32–41. doi:10.1016/j.copbio.2010.10.003

Miele, V., Bourguignon, P.-Y., Robelin, D., Nuel, G., & **Richard**, H. (2005). Seq++: Analyzing biological sequences with a range of markov-related models. *Bioinformatics*, *21*(11), 2783–2784. doi:10.1093/bioinformatics/bti389. eprint: http://bioinformatics.oxfordjournals.org/content/21/11/2783.full.pdf+html

Mirauta, B., Nicolas*, P., & **Richard***†, H. (2013). Pardiff: Inference of differential expression at base-pair level from rna-seq experiments. In A. Petrosino, L. Maddalena, & P. Pala (Eds.), *Iciap international workshops, naples, italy, september 9-13, 2013. proceedings* (Vol. 8158, pp. 418–427). Lecture Notes in Computer Science. doi:10.1007/978-3-642-41190-8_45

Mirauta, B., Nicolas*, P., & **Richard***†, H. (2014). Parseq: Reconstruction of microbial transcription landscape from rna-seq read counts using state-space models. *Bioinformatics*, *30*(10), 1409–1416. doi:10.1093/bioinformatics/btu042. eprint: http://bioinformatics.oxfordjournals.org/content/30/10/1409.full.pdf+html

Richard, H., & Nuel, G. (2003). SPA: Simple web tool to assess statistical significance of dna patterns. *Nucleic Acids Research*, *31*(13), 3679–3681. doi:10.1093/nar/gkg613. eprint: http://nar.oxfordjournals.org/content/31/13/3679.full.pdf+html

Richard*, H., Schulz*, M. H., Sultan*, M., Nurnberger, A., Schrinner, S., Balzereit, D., ... Yaspo, M.-L. (2010). Prediction of alternative isoforms from exon expression levels in rna-seq experiments. *Nucleic Acids Research*. doi:10.1093/nar/gkq041. eprint: http://nar.oxfordjournals.org/content/early/2010/02/11/nar.gkq041.full.pdf+html

Robelin, D., **Richard**, H., & Prum, B. (2003). SIC: A tool to detect short inverted segments in a biological sequence. *Nucleic Acids Research*, *31*(13), 3669–3671. doi:10.1093/nar/gkg596. eprint: http://nar.oxfordjournals.org/content/31/13/3669.full.pdf+html

Robin, S., Daudin, J. J., **Richard**, H., Sagot, M.-F., & Schbath, S. (2002). Occurrence probability of structured motifs in random sequences. *J. of Comp. Biol.* *9*(6), 761–773. doi:10.1089/10665270260518254

Rogato*, A., **Richard***†, H., Voss, B., Sarazin, A., Navarro, S. C., Navarro, L., ... Falciatore, A. (2014). The diversity of small non coding rna populations in the diatom phaeodactylum tricornotum. *BMC Genomics*, *15*(1), 698. doi:10.1186/1471-2164-15-698

Saad, C., Noé, L., **Richard**, H., Leclerc, J., Buisine, M.-P., Touzet, H., & Figeac, M. (2018). DiNAMO: Highly sensitive dna motif discovery in high-throughput sequencing data. *BMC Bioinformatics*, *19*(1), 223. doi:doi:10.1007/978-3-319-59826-0_7

Schulz*, M. H., Weese*, D., Holtgrewe*, M., Dimitrova, V., Niu, S., Reinert, K., & **Richard***†, H. (2014). Fiona: A parallel and automatic strategy for read error correction. *Bioinformatics*, *30*(17), i356–i363. doi:10.1093/bioinformatics/btu440. eprint: http://bioinformatics.oxfordjournals.org/content/30/17/i356.full.pdf+html

Shrestha, A. M., Asai, K., Frith, M., & **Richard**, H. (2018). Jointly aligning a group of dna reads improves accuracy of identifying large deletions. *Nucleic Acids Research*, *46*(3), e18. doi:doi:10.1093/nar/gkx1175

Sommer, A., Fuchs, S., Layer, F., Schaudinn, C., Weber, R. E., **Richard**, H., ... Strommenger, B. (2021). Mutations in the gdpp gene are a clinically relevant mechanism for β-lactam resistance in meticillin-resistant staphylococcus aureus lacking mec determinants. *Microbial Genomics*, *7*(9). doi:https://doi.org/10.1099/mgen.0.000623

Steijger, T. [T.], Abril, J., Engstr, Kokocinski, F., The RGASP Consortium: 58 authors including **Richard** H., Hubbard, T., ... Bertone, P. (2013). Assessment of transcript reconstruction methods for RNA-seq. *Nat. Methods*, *10*(12), 1177–1184. doi:10.1038/nmeth.2714

Sultan*, M., Schulz*, M. H., **Richard***, H., Magen, A., Klingenhoff, A., Scherf, M., ... Yaspo, M.-L. (2008). A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science*, *321*(5891), 956–960. doi:10.1126/science.1160342. eprint: http://www.sciencemag.org/content/321/5891/956.full.pdf

The MetaSUB Consortium: 63 authors including **Richard** H. (2016). Report on the first meeting of the metagenomics and metadesign of the subways and urban biomes (metasub) international consortium. *Microbiome*, *4*(1), 1. doi:10.1186/s40168-016-0168-z

Warren*, W. C., Clayton*, D. F., Ellegren*, H., Arnold*, A. P., ... (65 authors), **Richard**, H., ... Wilson, R. K. (2010). The genome of a songbird. *Nature*, *464*(7289), 757–762. doi:10.1038/nature08819

Weese, D. [David], Schulz, M. H., & **Richard**, H. (2017). Dna-seq error correction based on substring indices. In M. Elloumi (Ed.), *Algorithms for next-generation sequencing data: Techniques, approaches, and applications* (pp. 147–166). doi:doi:10.1007/978-3-319-59826-0_7

Zea, D. J., Laskina, S., Baudin, A., **Richard**†, H., & Laine†, E. (2021). Assessing conservation of alternative splicing with evolutionary splicing graphs. *Genome Research*, *31*(8), 1462–1473. doi:10.1101/gr.274696.120. eprint: http://genome.cshlp.org/content/31/8/1462.full.pdf+html