

**THÈSE DE DOCTORAT  
DE SORBONNE UNIVERSITÉ**

Spécialité: Bioinformatique  
École doctorale n. 515: Complexité du vivant

réalisée sous la direction de Rayan Chikhi

**Sequence Bioinformatics  
Institut Pasteur/CNRS – USR3756**

présentée par

**Luc Bassel**

**From sequences to knowledge, improving and  
learning from sequence alignments.**

**Soutenue le 2022-07-07**

devant le jury composé de:

TOPOLINO Alfredo	Professeur	Univ. Genève	Rapporteur
SE-YENG Fang	Professeur	Univ. Shanghai	Rapporteur
CASTAFIORE Bianca	Cantatrice	Scala di Milano	Examinateuse
LAMPION Serafon	Assureur		Invité
CHIKHI Rayan	PhD	Institut Pasteur	Directeur de Thèse



# Acknowledgments

Here will go my acknowledgments Sed commodo posuere pede. Mauris ut est. Ut quis purus. Sed ac odio. Sed vehicula hendrerit sem. Duis non odio. Morbi ut dui. Sed accumsan risus eget odio. In hac habitasse platea dictumst. Pellentesque non elit. Fusce sed justo eu urna porta tincidunt. Mauris felis odio, sollicitudin sed, volutpat a, ornare ac, erat. Morbi quis dolor. Donec pellentesque, erat ac sagittis semper, nunc dui lobortis purus, quis congue purus metus ultricies tellus. Proin et quam. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Praesent sapien turpis, fermentum vel, eleifend faucibus, vehicula eu, lacus.

Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Donec odio elit, dictum in, hendrerit sit amet, egestas sed, leo. Praesent feugiat sapien aliquet odio. Integer vitae justo. Aliquam vestibulum fringilla lorem. Sed neque lectus, consectetur at, consectetur sed, eleifend ac, lectus. Nulla facilisi. Pellentesque eget lectus. Proin eu metus. Sed porttitor. In hac habitasse platea dictumst. Suspendisse eu lectus. Ut mi mi, lacinia sit amet, placerat et, mollis vitae, dui. Sed ante tellus, tristique ut, iaculis eu, malesuada ac, dui. Mauris nibh leo, facilisis non, adipiscing quis, ultrices a, dui.

Morbi luctus, wisi viverra faucibus pretium, nibh est placerat odio, nec commodo wisi enim eget quam. Quisque libero justo, consectetur a, feugiat vitae, porttitor eu, libero. Suspendisse sed mauris vitae elit sollicitudin malesuada. Maecenas ultricies eros sit amet ante. Ut venenatis velit. Maecenas sed mi eget dui varius euismod. Phasellus aliquet volutpat odio. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Pellentesque sit amet pede ac sem eleifend consectetur. Nullam elementum, urna vel imperdiet sodales, elit ipsum pharetra ligula, ac pretium ante justo a nulla. Curabitur tristique arcu eu metus. Vestibulum lectus. Proin mauris. Proin eu nunc eu urna hendrerit faucibus. Aliquam auctor, pede consequat laoreet varius, eros tellus scelerisque quam, pellentesque hendrerit ipsum dolor sed augue. Nulla nec lacus.



# **Glossary**

This is the glossary



# Contents

<b>Acknowledgments</b>	<b>i</b>
<b>Glossary</b>	<b>iii</b>
<b>General Introduction</b>	<b>9</b>
Organization of this manuscript . . . . .	9
Research output . . . . .	9
<b>1. What is Sequence data ?</b>	<b>11</b>
1.1. Biological sequences, a primer . . . . .	11
1.2. Obtaining sequence data . . . . .	15
1.3. Sequencing errors, how to account for them ? . . . . .	21
1.4. The special case of homopolymers . . . . .	23
1.5. Conclusion . . . . .	26
References for chapter 1 . . . . .	26
<b>2. Aligning sequence data</b>	<b>41</b>
2.1. What is an alignment ? . . . . .	41
2.2. How do we speed up pairwise alignment ? . . . . .	45
2.3. Multiple sequence alignment . . . . .	49
2.4. The specificities of read-mapping . . . . .	51
References for chapter 2 . . . . .	52
<b>3. Contribution 1: Improving read alignment by exploring a sequence transformation space</b>	<b>69</b>
Abstract . . . . .	70
3.1. Introduction . . . . .	70
3.2. Methods . . . . .	71
3.3. Datasets and Pipelines . . . . .	78
3.4. Results . . . . .	79
3.5. Discussion . . . . .	85
3.6. Limitations of this study . . . . .	85
3.7. Code availability . . . . .	86
Supplementary information . . . . .	86
References for chapter 3 . . . . .	86

<b>4. Learning from alignments</b>	<b>89</b>
4.1. Alignments are a rich source of information . . . . .	89
4.2. Preprocessing the alignment for machine learning . . . . .	89
4.3. How to learn from ALNs . . . . .	91
<b>5. Viruses, HIV and drug resistance</b>	<b>93</b>
5.1. What are viruses ? . . . . .	93
5.2. Getting to know HIV . . . . .	95
5.3. Drug resistance in HIV . . . . .	101
References for chapter 5 . . . . .	103
<b>6. Contribution 2: Inferring mutation roles from sequence alignments using machine learning</b>	<b>125</b>
Abstract . . . . .	126
Author summary . . . . .	126
6.1. Introduction . . . . .	127
6.2. Materials and methods . . . . .	129
6.3. Results . . . . .	135
6.4. Discussion and perspectives . . . . .	145
Acknowledgments . . . . .	146
Supporting Information . . . . .	148
References for chapter 6 . . . . .	148
<b>7. Learning alignments, an interesting perspective</b>	<b>153</b>
7.1. Learning pairwise alignment . . . . .	153
7.2. What else could we learn ? . . . . .	153
<b>Global conclusion</b>	<b>155</b>
HPC part . . . . .	155
HIV part . . . . .	155
Final words . . . . .	156
<b>A. Supporting Information for “Mapping-friendly sequence reductions: going beyond homopolymer compression”</b>	<b>157</b>
A.1. “TandemTools” dataset generation . . . . .	157
A.2. MSR performance comparison . . . . .	159
A.3. Origin of incorrectly mapped reads of high mapping quality on whole human genome . . . . .	160
A.4. Analyzing read origin on whole human genome . . . . .	161
A.5. Performance of MSRs on the Drosophila genome . . . . .	166
References for appendix A . . . . .	166
<b>B. Supporting Information for “HIV and DRMs”</b>	<b>167</b>
B.1. Detailed list of HIV-1 protein structures used for figure generation. . . . .	167

<b>C. Supporting Information for “Using Machine Learning and Big Data to Explore the Drug Resistance Landscape in HIV”</b>	<b>169</b>
C.1. S1 Appendix (Technical appendix) . . . . .	169
C.2. S1 Fig. . . . .	173
C.3. S2 Fig. . . . .	174
C.4. S3 Fig. . . . .	175
C.5. S1 Table. . . . .	176
C.6. S2 Appendix. (Fisher exact tests) . . . . .	178
C.7. S1 Data. . . . .	178
C.8. S2 Data. . . . .	179
References for Appendix C . . . . .	179
<b>Global References</b>	<b>181</b>



# List of Figures

1.1. Double-helix structure of DNA . . . . .	12
1.2. Effect of frameshift mutations . . . . .	15
1.3. Caption . . . . .	17
1.4. Homopolymer length as fraction of the genome . . . . .	24
1.5. Distribution of homopolymer lengths per base in the human genome (for homopolymers of length $\geq 4$ ) . . . . .	24
1.6. Homopolymer compression can help resolve ambiguities due to sequencing errors . . . . .	26
2.1. Example global alignment with the Needleman-Wunsch algorithm . . . . .	43
2.2. 2 heuristic methods to speed up alignment: . . . . .	46
2.3. Example of kmers sharing a minimizer . . . . .	48
2.4. Overview of the progressive alignment process. . . . .	50
3.1. Representing and counting Streaming sequence reductions. . . . .	73
3.2. SSR equivalence classes for a fixed partition of the inputs. . . . .	76
3.3. Illustration of how a respective mapq threshold is chosen for each of our evaluated MSRs. . . . .	80
3.4. Graph representations of our highlighted MSRs: $MSR_E$ , $MSR_F$ , and $MSR_P$ . . . . .	81
3.5. Performance of our 58 selected mapping-friendly sequence reductions across genomes on reads simulated by <code>nanosim</code> . . . . .	82
5.1. Main steps of HIV-1 replication cycle . . . . .	97
5.3. 3D structure of HIV-1 Reverse-transcriptase . . . . .	100
5.2. Structure and main components of a mature HIV-1 virion. . . . .	123
6.1. Classifier Performance on UK and African datasets. . . . .	136
6.2. Discrimination between sequences having at least one RAM, and those having none on sequences with training features corresponding to known RAMs removed. . . . .	138
6.3. Relative risk of the new mutations with regards to known RAMs on the UK dataset . . . . .	141
6.4. Structure of HIV-1 RT with highlighted important sites. . . . .	143
A.1. Histogram of the original simulated positions for the incorrectly mapped reads using <code>minimap2</code> at high mapping qualities across the whole human genome, for several transformation methods. . . . .	160

A.2. Origin of correctly and incorrectly mapped raw reads . . . . .	161
A.3. Origin of correctly (teal) and incorrectly (red) mapped reads, transformed with HPC . . . . .	162
A.4. Origin of correctly (teal) and incorrectly (red) mapped reads, transformed with MSR <sub>E</sub> . . . . .	163
A.5. Origin of correctly (teal) and incorrectly (red) mapped reads, transformed with MSR <sub>P</sub> . . . . .	164
A.6. Origin of correctly (teal) and incorrectly (red) mapped reads, transformed with MSR <sub>F</sub> . . . . .	165
A.7. Results of the <code>paftools mapeval</code> evaluation on reads simulated and mapped to whole <i>Drosophila melanogaster</i> and <i>Escherichia coli</i> (Genbank ID U00096.2) genomes. . . . .	166
C.1. Relative risks of the new mutations with regards to known RAMs on the African dataset . . . . .	173
C.2. Closeup structural view of the entrance of the NNIBP of HIV-1 RT . . . . .	174
C.3. Closeup structural view of the active site of HIV-1 RT. . . . .	175

# List of Tables

1.1. Comparison of sequencing technology characteristics. . . . .	21
3.1. Performance of MSRs, HPC, and raw mappings across different mappers and reference sequences. . . . .	83
6.1. Summary of the UK and African datasets. . . . .	131
6.2. All training and testing datasets used during this study. . . . .	134
6.3. Analysis of new potential RAMs. . . . .	140
A.1. Comparing performance of MSRs on the whole human genome, whole <i>Drosophila melanogaster</i> genome, repeated regions of the whole human genome and synthetic centromeric sequence. . . . .	159
C.1. Detailed view of the characteristics of new potential RAMs . . . . .	177



# General Introduction

- Explain shortly that 2 quite different subjects linked by alignment and sequence data.

## Organization of this manuscript

- Organisation of the manuscript

## Research output

### Journal publications

- **Bassel, Luc**, Paul Medvedev and Rayan Chikhi. 2022. “**Mapping-friendly sequence reductions: going beyond homopolymer compression**”. *iScience* DOI goes here (*Adapted as Chapter 3*)
- **Bassel, Luc**<sup>1</sup>, Anna Zhukova<sup>1</sup>, Christian J Villabona-Arenas, Katherine E Atkins, Stéphane Hué, and Olivier Gascuel. 2021. “**Drug Resistance Mutations in HIV: New Bioinformatics Approaches and Challenges.**” *Current Opinion in Virology* 51 (December): 56–64.  
[10.1016/j.coviro.2021.09.009](https://doi.org/10.1016/j.coviro.2021.09.009) (*Used as the basis for Section 5.3.4*)
- **Bassel, Luc**, Anna Tostevin, Christian Julian Villabona-Arenas, MartinePeeters, Stéphane Hué, and Olivier Gascuel. 2021. “**Using Machine Learning and Big Data to Explore the Drug Resistance Landscape in HIV.**” *PLOS Computational Biology* 17 (8): e1008873.  
[10.1371/journal.pcbi.1008873](https://doi.org/10.1371/journal.pcbi.1008873). (*Adapted as Chapter 6*)
- Zhukova, Anna, **Luc Bassel**, Frédéric Lemoine, Marie Morel, JakubVoznica, and Olivier Gascuel. 2021. “**Origin, Evolution and Global Spread of SARS-CoV-2.**” *Comptes Rendus. Biologies* 344 (1): 57–75.  
[10.5802/crbiol.29](https://doi.org/10.5802/crbiol.29).

---

<sup>1</sup>Co-first authors: Luc Bassel and Anna Zhukova

## CHAPTER 0

- Lemoine, Frédéric, Luc Bassel, Jakub Voznica, and Olivier Gascuel.2020. “**COVID-Align: accurate online alignment of hCoV-19 genomes using a profile HMM**” *Bioinformatics*, 37 (12): 1761-1762.  
[10.1093/bioinformatics/btaa871](https://doi.org/10.1093/bioinformatics/btaa871).

### Presentations and posters

- “**Mapping-friendly sequence reductions: going beyond homopolymer compression**” proceedings talk, **RECOMB-SEQ 2022**. San Diego, USA (*May 21<sup>st</sup> 2022*)
- “**Can we improve analyses be transforming DNA?**” Joint RECOMB-SEQ RECOMB-CCB scientific **communication session<sup>2</sup>**. San Diego, USA (*May 21<sup>st</sup> 2022*).
- “**Machine learning approaches to reveal resistance mutations in HIV**” Poster at **MCEB 2019**. Porquerolles, France (*May 29<sup>th</sup> 2019*)

---

<sup>2</sup>2nd place prize awarded

# 1. What is Sequence data ?

## 1.1. Biological sequences, a primer

To fully understand the work that was done during this thesis, as well as the choices that were made, some basic knowledge of molecular biology and genetics is needed. If you are already familiar with biological sequences, feel free to skip ahead to section 1.2.

### 1.1.1. What is DNA ?

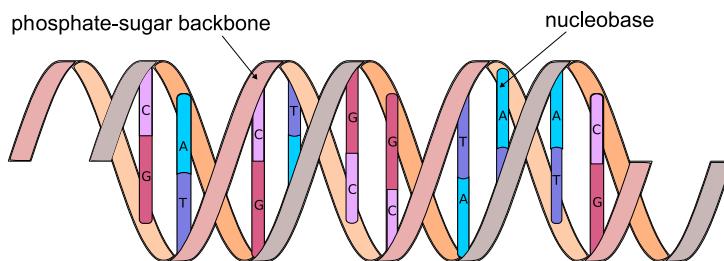
DesoxyriboNucleic Acid (DNA) is one of the most important molecules there is, without it complex life as we know it is impossible. It contains all the genetic information of a given organism, that is to say all the information necessary for the organism to: 1) function as a living being and 2) make a perfect copy of itself. This is the case for the overwhelming majority of living organisms on planet earth, from elephants to potatoes, to micro-organisms like bacteria.

DNA is a polymer, composed of monomeric units called nucleotides. Each nucleotide is composed of Ribose (a five carbon sugar) on which are attached a phosphate group as well as one of four nucleobases: Adenine (A), Cytosine (C), Guanine (G) or Thymine (T). These four types of nucleotide<sup>1</sup> monomers link up with one-another, through phosphate-sugar bonds, creating a single strand of DNA. The ordered sequence of these four types of nucleotides in strand encodes all the genetic information necessary for the organism to function. Nucleotides in a strand form strong complementary bonds with nucleotides from another strand, A with T and C with G. These bonds allows two strands of DNA to form the double-helix structure of DNA [1] shown in Figure 1.1. The specificity of nucleotide bonds ensure that the two strands of the double helix are complementary and that the information contained in one strand can be recovered from the other. This ensures a certain structural stability to the DNA molecule and a way to recover the important information that could be lost due to a damaged strand.

The amount of DNA necessary to encode the information varies greatly from organism to organism: 5400 base pairs (5.4kBp) for the  $\varphi X174$  phage [2], 4.9MBp for *Escherichia coli* [3], 3.1GBp for *Homo sapiens* [4] all the way up to almost 150GBp for *Paris japonica*,

---

<sup>1</sup>I will prefer the term “nucleotide” instead of “nucleobase” from now on. Chemically speaking they do not refer to the exact same molecule, but for our purposes you can consider these terms interchangeable.



**Figure 1.1.: Double-helix structure of DNA**

Each strand of DNA has a phosphate-sugar backbone on which are attached nucleobases. The two strands are linked by complementary bonds between the nucleobases of different strands (A bonding with T and C bonding with G).

a Japanese mountain flowering plant [5]. While very small genome size tend to occur in smaller, simpler organisms, genome size does not correlate with organism complexity [6].

### 1.1.2. From Information to action

#### 1.1.2.1. Proteins, their structure and their role

The double stranded DNA molecules present in the cells of a living organism contain information only; in order for the organism to live this information must be read and translated into actions. Most of the actions necessary for “life” are taken by large molecules called proteins, they have a very wide range of functions from catalyzing reactions in the cell to giving it its structure [7].

Proteins are macromolecules, that are made up of one or several chains of amino acids. These chains then link together and fold up in a specific three dimensional structure, giving the protein the shape it needs to fulfill its goal. This structure is determined by the sequence of amino acids, and a given protein can be identified by this amino acid sequence [7].

This sequence is directly dependent on the information contained in the DNA. First the DNA is transcribed in a similar, but single stranded, molecule called RNA (Ribonucleic Acid) which encodes the same sequence. This RNA molecule is then translated into a protein by the following process [8]:

1. Nucleotides in the RNA sequence are read in groups of three called codons.
2. These codons are read sequentially along the RNA molecule.
3. Each codon corresponds to an amino acid, according to the genetic code.
4. The sequence of codons in RNA (*and by extension DNA*) determines the sequence of amino acids.

## 1.1. BIOLOGICAL SEQUENCES, A PRIMER

5. The translation process is stopped when a specific type of codon (*a “Stop” codon*) is read.

With four types of nucleotides and codons grouping three nucleotides there are  $4^3 = 64$  possible codons. However, as stated above, proteins are only made up of 20 different amino acids, meaning that several different codons correspond to the same amino acid. This gives the translation process a certain robustness to errors that can occur when the DNA is copied to create a new cell, or when it is transformed into RNA prior to protein translation.

The portion of DNA that is read to create the protein is said to be “coding”, and is called a gene. There are several thousands of genes in the human genome [9] resulting in proteins executing thousands of different functions in a cell. In human beings, coding DNA represents only 1% to 2% of the total genome [10, 11]. The large majority of the DNA in a human being is not translated into proteins, a portion of it has a regulatory role, controlling transcription and translation, but the role remains unknown for the rest of the human genome [12, 13].

### 1.1.2.2. Making mistakes

Going from DNA sequence to protein is quite a complicated process involving several steps, it is therefore possible for a mistake to happen. There are several mechanisms to avoid mistakes and alteration of the genetic information: the complementary nature of the two strands of DNA, the redundant nature of the genetic code as well as error correction mechanisms in the molecules (*called “polymerases”*) that read and write DNA and RNA. However, despite all that, some errors in the nucleic acid (DNA and RNA) or protein sequences still make it through, these are called mutations.

**1.1.2.2.1. Where can mistakes happen ?** There are several sources of error that can alter the genetic information [14]:

- **DNA replication:** When a cell divides, or when an organism reproduces, the DNA molecule must be copied in order to preserve and transmit genetic information. This process has a very low rate of errors, with as low as one error for every billion to every hundred billion of replicated base pairs [15]. This is due to the fact that the DNA polymerase (the protein responsible for copying DNA molecules), has a relatively low error rate to start with, but mostly to the error correcting mechanisms that are present in certain cells and bacteria [16].
- **RNA transcription:** Since errors in RNA transcripts are less important than in replicated DNA, RNA polymerases have a much higher error rate than their DNA counterparts. This error rate has been estimated to be between four errors for each million [17] to two errors for each hundred thousand [18] transcribed bases.

## CHAPTER 1

- **Protein translation:** The process of translating RNA to a protein is done by proteins called ribosomes. This is a very error prone process with a mistranslation rate estimated to be of the order of one error for every 10,000 codons translated [19]
- **Other mutagenic events:** Many external events and factors have been shown to provoke mutations in exposed DNA such as Ionizing radiation [20], UV rays [21], Toxins [22], heat Stress [23], cold stress [24] or oxidative stress [25].

**1.1.2.2.2. What kind of errors are possible?** In biological sequences (nucleic acids and proteins), mutations can result from one of three error modes:

- **Substitutions**, where the original base units (nucleotide or amino acid) is mistakenly replaced by another one, for instance inserting an A instead of a G during RNA transcription.
- **Insertions**, where a new base unit not present in the original sequence is added to the newly synthesized biological sequence.
- **deletions**, where a base unit from the original sequence is skipped and not taken into account when synthesizing the new sequence.

While these three types of errors occur both in nucleic acids and proteins there are some things to consider about the consequences of nucleic acid mutations on protein synthesis. Due to the redundant nature of the genetic code mentioned in Section 1.1.2.1, some mutations in the nucleic acid sequence will result in the same protein sequence and therefore not have altered protein activity. Some mutations however will result in a substitution at the amino acid level which could potentially lead to a physicochemically altered or even non-functional protein. Finally, insertion and deletion errors (collectively called indels) can have big consequences on resulting proteins. Inserting or deleting nucleotides in multiples of three will result in the insertion/deletion of amino acids in the resulting protein, any other length of indel will result in what is called a frameshift mutation [26]. These mutations causes changes in the codons, potentially resulting in a completely different amino-acid sequence, including premature stop codon apparition as shown in Figure 1.2.

**1.1.2.2.3. What effect can mutations have ?** As we stated above, some mutations in DNA may have no repercussions, some others can lead to non-functional proteins. In some cases mutations can be associated with a trait in the mutated individual. For example a single mutation in a gene linked with coagulation can lead to pathological Leiden thrombophilia [27], a single amino acid deletion in the CFTR protein leads to (*the very deadly*) cystic fibrosis [28], and many mutations have been linked to complex diseases like type 2 diabetes [29, 30]. All mutational effects are not necessarily bad for the organism though, and mutations are essential for bacteria [31] or viruses like HIV [32] to develop resistance to treatment (*more on that in Chapters 5 and 6*).

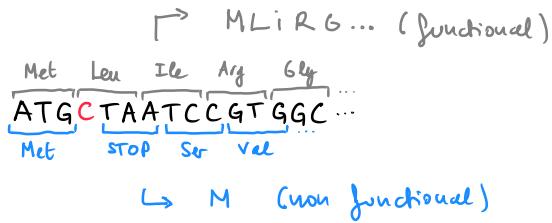


Figure 1.2.: Effect of frameshift mutations

The deletion of a single C (highlighted in red) in the original DNA sequence leads to a change in the codons read during translation. The original codons (shown in grey, with corresponding amino acids, above the sequence) translate to the functional protein. The new codons caused by the deletion (shown in blue, with corresponding amino acids, below the sequence), induce a premature STOP codon leading to a non-functional protein.

While some mutations, their mechanisms and consequences have been thoroughly studied, in many cases mutations are simply linked to a trait. Since it is easier to show correlation than causation, and that the former does not necessarily imply the latter, it is important to further study mutations of notice to understand their potential consequences.

## 1.2. Obtaining sequence data

In many fields, especially in computational biology, we need to know what genetic information the studied organism has, that is to say, what is the exact sequence of nucleotides that make up its DNA. The process of figuring out this sequence is, perhaps unsurprisingly, called sequencing. And a sequence that is inferred from this process is called a *sequencing read* or, more commonly, just a *read*.

### 1.2.1. Sanger sequencing, a breakthrough

The first sequencing method was developed in 1977 [33]. Sanger *et al.* devised a simple method to read the sequence of nucleotides that make up a DNA sequence known as “chain termination sequencing” or more commonly “Sanger sequencing” (*represented in Figure 1.3*).

To understand Sanger sequencing, one must first understand how to synthesize DNA. As we stated in Section 1.1.1, DNA is built up from building blocks that we called nucleotides, more specifically deoxynucleotidetriphosphates or dNTPs. these dNTPs are

## CHAPTER 1

made up of a sugar (deoxyribose) a nucleobase (A, T, G or C) and 3 phosphate groups. By successively adding these dNTPs at the end of an existing DNA molecule, we extend it, linking one of the phosphates of the dNTP to an oxygen atom on the last nucleotide of the DNA molecule. Let us now consider a dideoxynucleotidetriphosphate (ddNTP), which is identical to a dNTP except we remove a specific oxygen atom. This ddNTP can be added to the growing molecule of DNA like regular dNTPS, but since it is missing that one oxygen atom no more dNTPs or ddNTPs can be added to the DNA molecule after this one. The elongation is terminated and we call these ddNTPs chain-terminators. This combination of DNA synthesis followed by termination are at the heart of Sanger sequencing.

It is important to note that while dNTPs and ddNTPs refer to nucleotides with any nucleobase, we can refer to specific dNTPs by replacing the “N” with the base of choice. For example, dATP refers to the dNTP that has adenine as a base. Similarly we have dCTP, dGTP and dTTP (as well as ddATP, ddCTP, ddGTP and ddTTP).

1. The first step of Sanger sequencing (and most sequencing methods) is to amplify the DNA molecule we wish to sequence, *i.e.* make plenty of copies of it (usually through a process called PCR). These clones of the sequence are then separated into their two complementary strands one of which will be used as a template for the sequencing steps.
2. The second step is to prepare 4 different sequencing environments (*think of it as 4 test tubes*). In each environment we introduce an equal mix of the 4 dNTPs, that will be used to elongate new DNA molecules from the amplified templates, and a single type of ddNTP. So in the first test tube we will have only ddATP, ddCTP in the second, *et caetera*. In addition, these ddNTP are marked, at first with radioactive isotopes, and later on with dyes. This marking means that we can observe the location of these ddNTP later on.
3. Then an equal portion of the template is introduced in each environment with DNA polymerases (that will add the nucleotides to elongate a sequence that is complementary to the template), and short specific DNA molecules called primers that are necessary for the polymerases to start synthesizing new DNA.
4. During synthesis the chain will be elongated with dNTPs by the polymerase and the reaction will be stopped once a ddNTP is incorporated. At the end of this process we have plenty of fragments of DNA in each test tube, and we know that these fragments end with a specific base in a given environment. For example, in the test tube where we added ddATP, we know that all the fragments end with an A, and that we have all the possible fragments that start at the beginning of the template and end with an A. If the template is AACTA, then the fragments we would get in the ddATP test tube would be A, AA, and AACTA.
5. Then, a sample from each environment is taken and deposited in a gel, each in its own lane. A process called electrophoresis is then used to separate the fragments according to their weight. By applying an electrical current to the gel, the fragments of DNA will migrate away from where they were deposited along their

## 1.2. OBTAINING SEQUENCE DATA

lane in the gel. Lighter, shorted DNA fragments will travel further than heavier ones. We then get clusters of fragments ordered by weight (and therefore by length) called bands. With the marked ddNTP we can reveal these bands in the gel.

6. We know that: 1) bands are ordered by weight; 2) consecutive bands correspond to the addition of a single nucleotide; 3) in a specific lane fragments corresponding to a band end with a specific base. This knowledge is enough to deduce the sequence of the template we sequenced. An example gel is shown in Figure 1.3.

This process allowed Sanger *et al.* to sequence the first genome, of a  $\varphi X174$  bacteriophage, in 1977 [2]. Although revolutionnary, this method was costly, time consuming and labor intensive. Adjustments to this method were made in order to make it faster and less expensive. An important step was to change the way ddNTPs were marked. By using fluorescent markers, each base having a distinct “color”, we can eliminate the need to have 4 different environments and lanes in the gel [34, 35]. This also paved the way for automating sequencing, each fluoresently marked band can be excited with a laser, and the resulting specific wavelength can be recorded by optical systems and the corresponding base automatically deduced [36] (Also see Figure 1.3). Other improvements were made such as using capillary electrophoresis instead of gel electrophoresis.

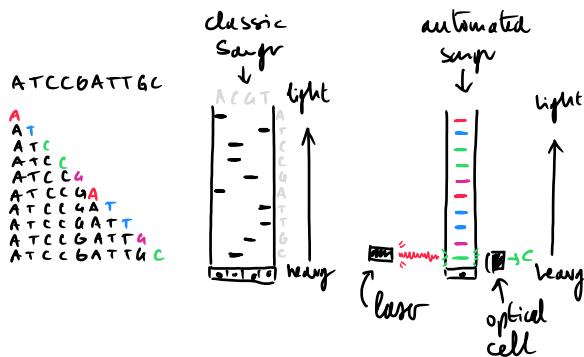


Figure 1.3.: Caption

(WIP)

These gradual improvements to the Sanger sequencing protocol, allowed to sequence longer and more accurate reads, with the latest technologies resulting in reads reaching 1 ,000 base pairs with an accuracy of 99.999% [37]. These improvements also resulted in a lower cost for sequencing, which was greatly decreased from around \$1000 per base-pair [38] to only \$0.5 per kilobase [37]. Finally these technological improvements also increased the throughput of sequencing machines from around 1 kilobase per day [38] to 120 kilobases per hour [39].

Despite these improvements, for ambitious endeavours such as the human genome project, sequencing was a massive undertaking: the first human genome is estimated to

## CHAPTER 1

have cost between 500 million and 1 billion US dollars to sequence [40].

### 1.2.2. Next-generation sequencing

Through these large sequencing projects and the genomics field in general, the richness and usefulness of sequence data was made ever more apparent. This growing need of sequence data ushered in a new era of sequencing with the development of many new sequencing methods designed to have a higher throughput and a lower cost than Sanger sequencing. This second generation of sequencing technologies, also referred to as Next-Generation Sequencing (NGS) or Massively parallel sequencing. While there were different technologies, there are a few common key points [41]:

- As with Sanger sequencing we need to amplify and clone the DNA template, however since these technologies result in short reads than Sanger sequencing, the DNA we want to sequence must first be randomly broken up into small template fragments before being amplified.
- The amplified template fragments are attached to some sort of solid support, resulting in a physical support with billions of template fragments attached to it.
- As in Sanger sequencing, DNA molecules, complementary to the template fragments, are elongated. This happens for billions of fragments at the same time (hence the “massively parallel” epithet).
- The addition of specific nucleotides to a chain are detected in real time, and there is no chain termination. There is no need for the long step of electrophoresis. These detections are simultaneous for all the molecules being elongated at once.

The result of these steps is a very large number of short reads. With data analysis these short reads can be used to deduce longer sequences and eventually the original whole genome sequence through a process called *assembly*.

The main NGS method is called “sequencing by synthesis”, developed by a company: Illumina. It is commonly referred to as *Illumina sequencing*. This method is based on *reversible chain terminators*, developed at the Institut Pasteur in the 90’s [42]. These are marked dNTPs that can be used to elongate DNA molecules, but that have an additional molecular group that makes them terminators by default. However this terminating group can be removed once the NTP is included in a DNA molecule allowing the elongation process to continue. These dNTPs are fluorescently marked and when excited with a laser they emit light with a distinctive color. During Illumina sequencing, these reversible chain terminators are included to millions of fragments at the same time, stopping elongation. At this point all the fragments are excited with a laser and an optical system takes a picture of the emitted colors for all the fragments at once. In this image, a pixel loosely corresponds to a sequenced fragment, and its color to the most recently added dNTP. The terminating groups are then cleaved and the process

## 1.2. OBTAINING SEQUENCE DATA

can start over by incorporating a new batch of reversible terminators. By observing the successive images we can deduce the sequence of added nucleotides for each sequenced fragment and obtain all of our reads.

Another NGS method is called pyrosequencing, commercialized by 454 Life Sciences. Contrary to Illumina sequencing, this method does not use reversible chain terminators. Instead it uses a special enzyme called luciferase that emits light as specific dNTPs are added. This process is repeated for the 4 dNTPs (similarly to Sanger sequencing) and from the light emissions we can deduce the sequence of nucleotides [43].

These technologies yield reads around 150 nucleotides for Illumina and 400nt for pyrosequencing [44], this is much shorter than the 1kB reads obtainable from the latest Sanger sequencing technologies. However the throughputs are much higher [39]: 2.5 to 12.5 Gigabases per hour for Illumina and 30 Megabases per hour for pyrosequencing. Costs are also quite low: \$0.07 and \$10 per Megabase for Illumina and pyrosequencing respectively. The per-base sequencing accuracies are also quite high, up to 99.9% for both Illumina [45] and pyrosequencing [39]. A summary of the key characteristics for various sequencing technologies can be found in Table 1.1. The lower cost and higher throughput has made the Illumina sequencing technology the dominant one, the company estimating that 90% of the world's sequencing data was generated with Illumina machines in 2015 [46].

### 1.2.3. Long read sequencing

Although NGS technologies revolutionized the sequencing world, recent efforts have been made to get longer reads. These third-generation methods generate reads of tens of kilobases and are commonly called *long-read sequencing* method. Long reads have a host of applications [47] for which short NGS reads might not be well suited: *De novo* assembly of large complex genomes, studying complex repetitive regions such as centromeres or telomeres or detection of structural variants. They have recently been used to assemble the first truly complete human genome, including telomeric and centromeric regions [4].

The two available long read technologies are: Single Molecule Real Time sequencing (SMRT), commercialized by Pacific Biosciences (PacBio) and Nanopore sequencing, commercialized by Oxford Nanopore Technologies (ONT). While both technologies are quite different they result in much longer reads than even Sanger sequencing in real time, without the need for chain terminators or separate sequencing reactions, with a high throughput and at a reasonably low cost.

SMRT sequencing was first developed in 2009 [48], before being commercialized and furthered by PacBio. The basic principle is as follows:

1. Fragment and amplify DNA to obtain a very large number of DNA templates.

## CHAPTER 1

2. Link both strands of each DNA template together with known sequences called *bell adapters*. Denature the DNA to create a single stranded, circular DNA molecule.
3. Primers and polymerases are attached to the circular molecule specifically on one of the bell adapters.
4. Add the circular DNA template, primer, polymerases complexes to a SMRT chip. This chip is essentially a large aluminium surface with hundreds of thousands of microscopic wells called *Zero-Mode Waveguides* (ZMWs) only 100nm in diameter [49]. The polymerases are chemically bonded to the bottom of each of these ZMWs so we effectively get a single DNA template and polymerase per well.
5. Fluorescently marked dNTPs are incorporated progressively in each of the wells. When a marked dNTP is incorporated in the newly synthesized DNA strand, light of a specific wavelength is emitted.
6. The size of these ZMWs make the detection of the fluorescence possible with an optical system. Incorporation of dNTPs in each ZMW can be detected simultaneously in a parallel fashion and the resulting sequences deduced.

Nanopore sequencing, though of in the eighties, further developed along the years [50] and first commercialized by ONT in 2014 [51], is completely different from all the sequencing technologies previously mentioned. Where all the other ones are based on synthesizing a complementary DNA strand and detecting specific dNTP incorporation in some way or another, there is no synthesis in nanopore sequencing. The principle relies on feeding a single strand of a DNA template through a small hole in a membrane, a nanopore, at a controlled speed. As the nucleotides go through the nanopore, the electric current is formed between both sides of the membrane. This current can be measured and is specific to the succession of 5 to 6 nucleotides inside the nanopore channel at any given time. By looking at the evolution of the electric current as the DNA strand goes through the nanopore, we can deduce the sequence of nucleotides through a process called *base calling*. Base calling is usually done with machine learning methods, mainly artificial neural networks [52]. In the flow cells used in ONT sequencers, there are hundreds of thousands of nanopores, spread out over a synthetic membrane, allowing for massively parallel sequencing as well. Theoretically, since this method is not based on synthesis, the upper limit for read length is only limited by the length of the template, and in practice ONT sequencing produces the longest reads.

Both technologies yield long reads, the median and highest read lengths being 10 kilobases and 60 kilobases respectively for PacBio sequencing [53]. For nanopore the median read lengths of 10 to 12 kilobases [54, 55] are similar to PacBio, but it can also yield ultra-long reads of 1 up to 2.3 megabases long [56, 57, 58]. The length of the reads and parallel nature of these two technologies allow these sequencers to have truly massive throughputs. PacBio sequencers can sequence between 2 and 11 gigabases per hour and ONT from 12.5 gigabases per hour, up to a staggering 260 gigabases per hour for the latest ONT PromethION machines [55]. The cost of sequencing with these machines, while higher than for Illumina sequencers, remains reasonably affordable at \$0.32 and \$0.13 per megabase for PacBio and ONT respectively [59]. These characteristics are

### 1.3. SEQUENCING ERRORS, HOW TO ACCOUNT FOR THEM ?

technology	read length (nt)	throughput (nt/hour)	cost (\$/Mb)	accuracy
Sanger	1 000	120 $10^3$	\$500	99.999%
Illumina	150	2.5-12.5 $10^9$	\$0.07	99.9%
Pyrosequencing	400	30 $10^6$	\$10	99.9%
PacBio SMRT	10 000 (up to 60 000)	2-11 $10^9$	\$0.32	85-92%
Nanopore	12 000 (up to $2.5 \cdot 10^6$ )	12.5-260 $10^9$	\$0.13	87-98%

Table 1.1.: Comparison of sequencing technology characteristics.

Characteristics for the latest sequencers were used for the Sanger sequencing entry. The length is given in nucleotides, throughputs in sequences nucleotides per hour and cost in US dollars per megabase.

summarized in Table 1.1 along with other sequencing technologies.

The length, throughput and sequencing cost of both these technologies paint a pretty picture, and indeed they have proved useful in many settings, sequencing accuracy however is the main problem with these technologies. The per-base sequencing accuracy has been estimated to be between 85% and 92% for PacBio sequencers and 87% to 98% for ONT machines [60, 55, 61]. This accuracy is much lower than either Sanger sequencing or Illumina reads. Characterizing, correcting and accounting for these errors is widely studied and it will be discussed in more detail in Sections 1.3 and 1.4.

While most of the mentioned technologies can also be adapted and used to sequence RNA instead of DNA [62, 63], directly sequencing proteins remains a challenge. The sequence of amino acids making up a protein is usually deduced from the codons in sequenced DNA or RNA after detection of potentially coding regions called open reading frames (ORFs). Development of methods to directly sequence protein molecules using mass spectrometry was started not very long after Sanger sequencing [64] and improved [65]. New methods are still being developed [66] but protein sequencing is still a challenge.

### 1.3. Sequencing errors, how to account for them ?

Sequencing technologies are not perfect, they make errors as we can see from the various accuracy rates reported in Section 1.2. For technologies based on nucleic acid synthesis (i.e. everything except ONT), since they use polymerases it stands to reason that the same three types of errors, described in Section 1.1.2.2, occur: substitutions, insertions and deletions. For long read technologies though, most of the errors do not come from the polymerase, but from signal processing used to deduce the sequence. Since both technologies execute single molecule sequencing, the signal to noise ratio is low [67, 68] making base calling more complicated.

## CHAPTER 1

This explains the discrepancy in error rates between short and long read sequencing technologies: the former getting as low as  $10^{-4}$  or  $10^{-5}$  after computational processing [69] where the latter are between 10% and 15%. This high error rate long reads is bothersome and many efforts have been made lower this error rate, computationally or technologically.

### 1.3.1. Error correction methods

The long read error-correction literature and toolset is rich and active [70, 71, 72]. There are two main ways to correct errors: 1) hybrid methods where higher-accuracy shorter reads are used to correct errors and 2) non-hybrid methods where only the long-reads are used.

In Non-hybrid methods [70, 73], by finding regions that overlap fairly well between reads and taking the consensus (i.e. the majority nucleotide at each position) of the overlapped regions some errors can be eliminated. In many analyses and sequencing data processing pipelines, the first step is to break up the reads into all possible overlapping subsequences of length  $k$  called k-mers (e.g the 3-mers of the ATTGC are ATT, TTG and TGC). Rare k-mers in the read dataset, i.e. k-mers that appear only a handful of times in all the reads, are likely the result of an error and filtering them out can improve analysis. One or both of these procedures are implemented in several pieces of commonly used software like the wtdbg [74], and canu [75] assemblers or the standalone daccord long-read correcter [76]. In some cases, errors are corrected not on the raw reads but after having assembled the long reads into long continuous sequences (contigs), this process is called polishing. The ntEdit polisher [77] also filters out rare kmers to correct errors. The Arrow [78] and Nanopolish [79] polishers correct the assembly using the raw PacBio and ONT long reads respectively.

Hybrid methods, as their name suggest, makes use of short reads to correct errors in long reads. By finding similar regions between the short and long reads we can use the higher accuracy of short reads to correct the long ones. This is implemented in many pieces of software proovread [80], Jabba [81], PBcR [82] or LORDEC [83]. Short reads can also be used to polish long read assemblies with tools like Pilon [84] or Racon [85]. The first complete human genome was assembled and polished using many different sequencing technologies including PacBio, ONT and Illumina technologies [4].

### 1.3.2. More accurate sequencing methods

While a lot of effort is being put into error correction, another angle of attack to lower the error rate of long reads is to improve the sequencing technology.

In 2019, PacBio introduced HiFi reads, based on a circular consensus (CCS) technique [86]. During SRMT sequencing the 2 strands are linked together by ball adapters to form a circular DNA template (c.f. Section 1.2.3), the central idea of CCS is to sequence this

## 1.4. THE SPECIAL CASE OF HOMOPOLYMERS

molecule multiple time by going over the circle multiple times. In the resulting long sequence the, known bell adapter sequences can be removed, and a consensus sequence can be built from the multiple passes over the same DNA template. This results in long-read accuracies of 99.8% to 99.9% [86, 55]. This works because PacBio sequencing errors are mostly randomly distributed along the sequenced template (more on that in Section 1.4.2), therefore it is unlikely that the same error will appear in multiple passes over the same template portion.

For ONT sequencing, most improvement efforts have been focused on base-callers. These tools were originally based on Hidden Markov Models [87] (HMMs), but gradually they have been shifting over to neural network based deep learning methods [88, 89, 52, 73] with faster inference times and better performance.

A technologically agnostic method using unique molecular identifiers added during the template preparation phase, and consensus sequencing has been shown, in specific contexts, to improve the accuracies of both ONT and PacBio CCS long reads to 99.59% and 99.93% respectively [90].

Finally, new sequencing technologies are being developed, like built in error-correction short-read technologies yielding error-free reads of up to 200 nucleotides [91]. Illumina also recently announced its own high-throughput, high-accuracy long-read sequencing technology in 2022 [92], although details about the performance and technology are scarce.

## 1.4. The special case of homopolymers

Despite improvement in error correction methods and sequencing technologies, certain genetic patterns are particularly difficult to process, homopolymers are one such pattern.

### 1.4.1. Homopolymers and the human genome

*Homopolymers* consist of a stretch of repeated nucleotides (i.e.  $\geq 2$ ) occurring at some point in the genome. For example the sequence AAAA is a length 4 adenine homopolymer.

- What are homopolymers -> repeated stretches of identical nucleotides
- On the CHM13 whole human genome assembly v1.1, 50% of the 3Gb are in homopolymers of size 2 or more, and 10% are in homopolymers of size 4 or more. The longest homopolymer in the CHM13 v1.1 whole human genome assembly is 86. Homopolymers of short/medium lengths make up a significant part of the human genome (Figure 1.4).
- HPs are more often A/T rich than G/C rich in the human genome (Figure 1.5).

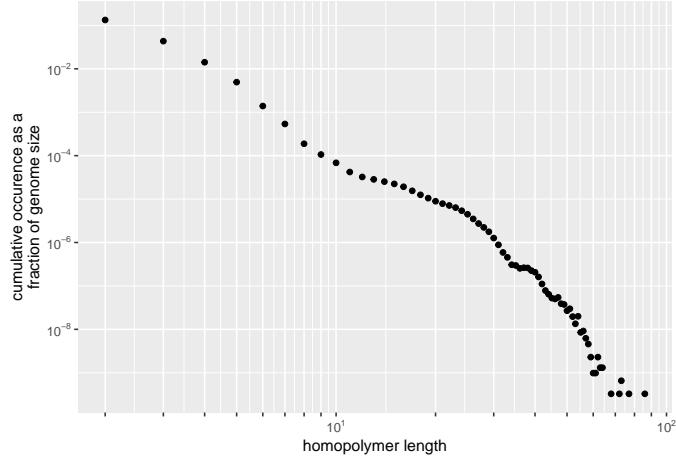


Figure 1.4.: Homopolymer length as fraction of the genome

- More than 1.4 million homopolymers 4-mers and up in the exome (i.e. DNA that is translated into proteins) [93]
- according to [94], in the GRCh38 human genome assembly, more than 1.9 Mb are in homopolymers of length 8 or higher, representing about a thousandth of the genome, still a lot of bases to make mistakes on! Longest HP is 90.

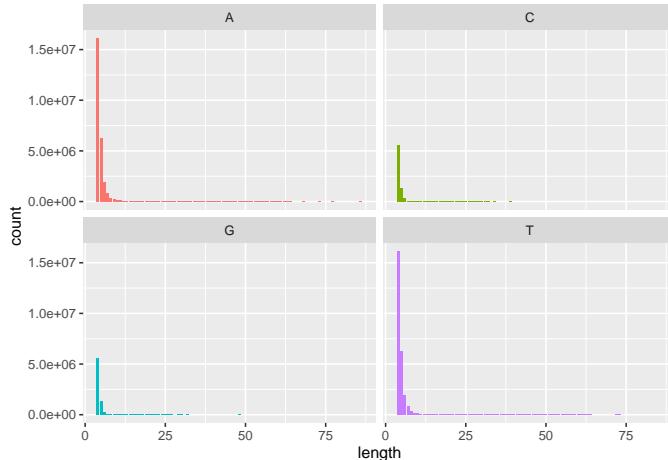


Figure 1.5.: Distribution of homopolymer lengths per base in the human genome (for homopolymers of length  $\geq 4$ .)

#### 1.4.2. Homopolymers and long reads

Mainly indels, in certain regions of the genomes, particularly homopolymers. For ONT [95, 96], for PacBIO and ONT [97, 67]. ONT has more homopolymer errors than PacBIO

[55]. For short reads and PacBio the error rate usually goes up as the homopolymer length grows, however ONT's error rate in HPs does not depend on the length (even though it's the highest it's flat) [98]. Most HiFi errors are still indels in homopolymers [86]

### 1.4.3. Accounting for homopolymers

#### 1.4.3.1. Specific error correction

- In some cases homopolymers are taken into special consideration during polishing [99, 79], and computational approaches are developed to deal with homopolymers [100, 101]
- Specific technologies:
  - R.10 Nanopore: [102, 73]
  - Solid state nanopores: [103, 104]
  - BaseCallers which deal with homopolymers, like can be seen in Fig 3 of [52] (scrappie raw 1.4.1 or guppy)
- Avoid using HPs in barcode sequences [105, 106].
- A little different but HPs have to be taken into account when developing DNA based data storage [107]

#### 1.4.3.2. HPC Trick

homopolymers are still tricky regions even with long high accuracy reads -> so HPC

- HPC takes repeated runs of a single nucleotide and compresses them to a single occurrence
- Can help resolve ambiguities in some cases c.f. Figure 1.6.
- Used in many software tools:
  - assembly: HiCanu [108], MDBG [109], pyrosequencing assembly [110], wtdbg2 [74], shasta [111]
  - clustering: [112]
  - error correction: LSC [113] and LSCPlus [114]
  - alignment: minimap2 [115], winnowmap2 [116].
  - Satellite tandem repeat analysis [117]

reconstruct the whole genome.

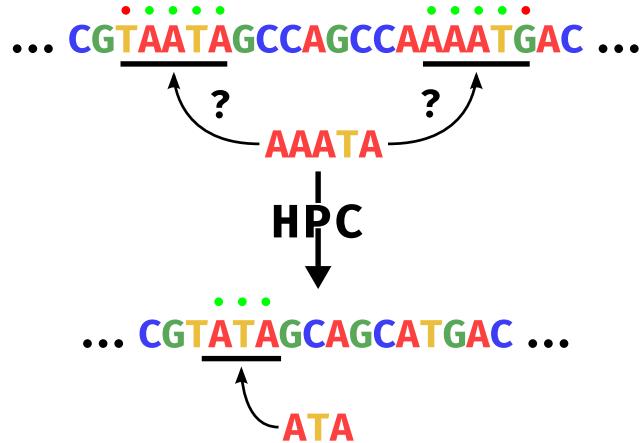


Figure 1.6.: Homopolymer compression can help resolve ambiguities due to sequencing errors

A read with homopolymer related sequencing errors can be homologous to two different regions of the reference genome. After applying HPC, this ambiguity is resolved and the read is homologous to only one region.

## 1.5. Conclusion

- **Sequencing is an active field:** new methods, both in short reads (ultima [118] ) and long reads (illumina infinity [92] ), as well as improving existing technologies, PacBio (by acquiring Omnicome [119]) developing a new accurate short read technology [120]
- **Long reads are good:** Super useful for plenty of very important tasks but...
- **Errors are bad:** particularly in HPs since they are an important part of the genome and a major source of error in long reads
- **Computational view:** sequences Essentially a text file. Just a long succession of letters, so we can apply String/text algorithmics to it.

## References for chapter 1

- [1] J. D. Watson and F. H. C. Crick. “The Structure of Dna”. In: *Cold Spring Harbor Symposia on Quantitative Biology* 18 (Jan. 1, 1953). tex.ids= watson-STRUCTUREDNA1953 PMID: 13168976 publisher: Cold Spring Harbor Laboratory Press, pp. 123–131. doi: [10.1101 / SQB.1953.018.01.020](https://doi.org/10.1101/SQB.1953.018.01.020). URL: <http://symposium.cshlp.org/content/18/123> (cit. on p. 11).

## REFERENCES FOR CHAPTER 1

- [2] F. Sanger, G. M. Air, et al. “Nucleotide Sequence of Bacteriophage X174 DNA”. In: *Nature* 265.5596 (5596 Feb. 1977), pp. 687–695. ISSN: 1476-4687. DOI: [10.1038/265687a0](https://doi.org/10.1038/265687a0). URL: <https://www.nature.com/articles/265687a0> (visited on 05/17/2022) (cit. on pp. 11, 17).
- [3] Colin T. Archer, Jihyun F. Kim, et al. “The genome sequence of *E. coli* W (ATCC 9637): comparative genome analysis and an improved genome-scale reconstruction of *E. coli*”. In: *BMC Genomics* 12.1 (Jan. 6, 2011), p. 9. DOI: [10.1186/1471-2164-12-9](https://doi.org/10.1186/1471-2164-12-9). URL: <https://doi.org/10.1186/1471-2164-12-9> (cit. on p. 11).
- [4] Sergey Nurk, Sergey Koren, et al. “The complete sequence of a human genome”. In: *Science* 376.6588 (Apr. 2022). Publisher: American Association for the Advancement of Science, pp. 44–53. DOI: [10.1126/science.abj6987](https://doi.org/10.1126/science.abj6987). URL: <https://www.science.org/doi/10.1126/science.abj6987> (cit. on pp. 11, 19, 22, 71, 78).
- [5] Jaume Pellicer, Michael F. Fay, and Ilia J. Leitch. “The largest eukaryotic genome of them all?” In: *Botanical Journal of the Linnean Society* 164.1 (Sept. 1, 2010), pp. 10–15. DOI: [10.1111/j.1095-8339.2010.01072.x](https://doi.org/10.1111/j.1095-8339.2010.01072.x). URL: <https://doi.org/10.1111/j.1095-8339.2010.01072.x> (cit. on p. 12).
- [6] H. C. Macgregor. “C-Value Paradox”. In: ed. by Sydney Brenner and Jefferey H. Miller. DOI: [10.1006/rwgn.2001.0301](https://doi.org/10.1006/rwgn.2001.0301). New York: Academic Press, Jan. 1, 2001, pp. 249–250. DOI: [10.1006/rwgn.2001.0301](https://doi.org/10.1006/rwgn.2001.0301). URL: <https://www.sciencedirect.com/science/article/pii/B0122270800003013> (cit. on p. 12).
- [7] Bruce Alberts, Alexander Johnson, et al. *Molecular Biology of the Cell. 4th edition*. Garland Science, 2002. URL: <https://www.ncbi.nlm.nih.gov/books/NBK26916/> (cit. on p. 12).
- [8] F. H. C. Crick, Leslie Barnett, et al. “General Nature of the Genetic Code for Proteins”. In: *Nature* 192.4809 (Dec. 1961). Number: 4809 Publisher: Nature Publishing Group, pp. 1227–1232. DOI: [10.1038/1921227a0](https://doi.org/10.1038/1921227a0). URL: <https://www.nature.com/articles/1921227a0> (cit. on p. 12).
- [9] International Human Genome Sequencing Consortium. “Finishing the euchromatic sequence of the human genome”. In: *Nature* 431.7011 (Oct. 2004). Number: 7011 Publisher: Nature Publishing Group, pp. 931–945. DOI: [10.1038/nature03001](https://doi.org/10.1038/nature03001). URL: <https://www.nature.com/articles/nature03001> (cit. on p. 13).
- [10] Ran Elkon and Reuven Agami. “Characterization of noncoding regulatory DNA in the human genome”. In: *Nature Biotechnology* 35.8 (Aug. 2017). Number: 8 Publisher: Nature Publishing Group, pp. 732–746. DOI: [10.1038/nbt.3863](https://doi.org/10.1038/nbt.3863). URL: <https://www.nature.com/articles/nbt.3863> (cit. on p. 13).
- [11] Gilbert S. Omenn. “Reflections on the HUPO Human Proteome Project, the Flagship Project of the Human Proteome Organization, at 10 Years”. In: *Molecular & Cellular Proteomics : MCP* 20 (Feb. 26, 2021). PMID: 33640492 PMCID: PMC8058560, p. 100062. DOI: [10.1016/j.mcpro.2021.100062](https://doi.org/10.1016/j.mcpro.2021.100062). URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8058560/> (cit. on p. 13).

## CHAPTER 1

- [12] Svetlana A. Shabalina and Nikolay A. Spiridonov. “The mammalian transcriptome and the function of non-coding DNA sequences”. In: *Genome Biology* 5.4 (Mar. 25, 2004), p. 105. DOI: [10.1186/gb-2004-5-4-105](https://doi.org/10.1186/gb-2004-5-4-105). URL: <https://doi.org/10.1186/gb-2004-5-4-105> (cit. on p. 13).
- [13] ENCODE Project Consortium. “An Integrated Encyclopedia of DNA Elements in the Human Genome”. In: *Nature* 489.7414 (Sept. 6, 2012). PMID: 22955616 PMCID: PMC3439153, pp. 57–74. DOI: [10.1038/nature11247](https://doi.org/10.1038/nature11247). URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3439153/> (cit. on p. 13).
- [14] Nimrat Chatterjee and Graham C. Walker. “Mechanisms of DNA damage, repair, and mutagenesis: DNA Damage and Repair”. In: *Environmental and Molecular Mutagenesis* 58.5 (June 2017), pp. 235–263. DOI: [10.1002/em.22087](https://doi.org/10.1002/em.22087). URL: <https://onlinelibrary.wiley.com/doi/10.1002/em.22087> (cit. on p. 13).
- [15] Iwona J. Fijalkowska, Roel M. Schaaper, and Piotr Jonczyk. “DNA replication fidelity in Escherichia coli: a multi-DNA polymerase affair”. In: *FEMS microbiology reviews* 36.6 (Nov. 2012). PMID: 22404288 PMCID: PMC3391330, pp. 1105–1121. DOI: [10.1111/j.1574-6976.2012.00338.x](https://doi.org/10.1111/j.1574-6976.2012.00338.x). URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3391330/> (cit. on p. 13).
- [16] Leslie Pray. “DNA replication and causes of mutation”. In: *Nature education* 1.1 (2008), p. 214 (cit. on p. 13).
- [17] Jean-François Gout, W. Kelley Thomas, et al. “Large-scale detection of in vivo transcription errors”. In: *Proceedings of the National Academy of Sciences* 110.46 (Nov. 12, 2013). Publisher: Proceedings of the National Academy of Sciences, pp. 18584–18589. DOI: [10.1073/pnas.1309843110](https://doi.org/10.1073/pnas.1309843110). URL: <https://www.pnas.org/doi/full/10.1073/pnas.1309843110> (cit. on p. 13).
- [18] Jean-Francois Gout, Weiyi Li, et al. “The landscape of transcription errors in eukaryotic cells”. In: *Science Advances* 3.10 (Oct. 20, 2017). PMID: 29062891 PMCID: PMC5650487, e1701484. DOI: [10.1126/sciadv.1701484](https://doi.org/10.1126/sciadv.1701484). URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5650487/> (cit. on p. 13).
- [19] Dmitri Shcherbakov, Youjin Teo, et al. “Ribosomal mistranslation leads to silencing of the unfolded protein response and increased mitochondrial biogenesis”. In: *Communications Biology* 2.1 (Oct. 17, 2019). Number: 1 Publisher: Nature Publishing Group, pp. 1–16. DOI: [10.1038/s42003-019-0626-9](https://doi.org/10.1038/s42003-019-0626-9). URL: <https://www.nature.com/articles/s42003-019-0626-9> (cit. on p. 14).
- [20] Omar Desouky, Nan Ding, and Guangming Zhou. “Targeted and non-targeted effects of ionizing radiation”. In: *Journal of Radiation Research and Applied Sciences* 8.2 (Apr. 1, 2015), pp. 247–254. DOI: [10.1016/j.jrras.2015.03.003](https://doi.org/10.1016/j.jrras.2015.03.003). URL: <https://www.sciencedirect.com/science/article/pii/S1687850715000333> (cit. on p. 14).

## REFERENCES FOR CHAPTER 1

- [21] Jürgen Kiefer. "Effects of Ultraviolet Radiation on DNA". In: ed. by Günter Obe and Vijayalaxmi. DOI: 10.1007/978-3-540-71414-9\_3. Berlin, Heidelberg: Springer, 2007, pp. 39–53. doi: 10.1007 / 978 - 3 - 540 - 71414 - 9 \_ 3. URL: [https://doi.org/10.1007/978-3-540-71414-9\\_3](https://doi.org/10.1007/978-3-540-71414-9_3) (cit. on p. 14).
- [22] J. W. Bennett and M. Klich. "Mycotoxins". In: *Clinical Microbiology Reviews* 16.3 (July 2003). PMID: 12857779 PMCID: PMC164220, pp. 497–516. doi: 10.1128 / CMR.16.3.497-516.2003. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC164220/> (cit. on p. 14).
- [23] O.L. Kantidze, A.K. Velichko, et al. "Heat Stress-Induced DNA Damage". In: *Acta Naturae* 8.2 (2016). PMID: 27437141 PMCID: PMC4947990, pp. 75–78. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4947990/> (cit. on p. 14).
- [24] C. D. Gregory and A. E. Milner. "Regulation of cell survival in Burkitt lymphoma: implications from studies of apoptosis following cold-shock treatment". In: *International Journal of Cancer* 57.3 (May 1, 1994). PMID: 8169005, pp. 419–426. doi: 10.1002/ijc.2910570321 (cit. on p. 14).
- [25] Anat Gafter-Gvili, Boris Zingerman, et al. "Oxidative Stress-Induced DNA Damage and Repair in Human Peripheral Blood Mononuclear Cells: Protective Role of Hemoglobin". In: *PLoS ONE* 8.7 (July 9, 2013). PMID: 23874593 PMCID: PMC3706398, e68341. doi: 10.1371/journal.pone.0068341. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3706398/> (cit. on p. 14).
- [26] J. R. Roth. "Frameshift mutations". In: *Annual Review of Genetics* 8 (1974). tex.ids= rothFRAMESHIFTMUTATIONS1974 PMID: 4613260, pp. 319–346. doi: 10.1146/annurev.ge.08.120174.001535 (cit. on p. 14).
- [27] Jody Lynn Kujovich. "Factor V Leiden thrombophilia". In: *Genetics in Medicine* 13.1 (Jan. 1, 2011), pp. 1–16. doi: 10.1097/GIM.0b013e3181faa0f2. URL: <https://www.sciencedirect.com/science/article/pii/S1098360021040430> (cit. on p. 14).
- [28] Garry R. Cutting. "Cystic fibrosis genetics: from molecular understanding to clinical application". In: *Nature Reviews Genetics* 16.1 (Jan. 2015). Number: 1 Publisher: Nature Publishing Group, pp. 45–56. doi: 10.1038 / nrg3849. URL: <https://www.nature.com/articles/nrg3849> (cit. on p. 14).
- [29] Christian Fuchsberger, Jason Flannick, et al. "The genetic architecture of type 2 diabetes". In: *Nature* 536.7614 (Aug. 2016). Number: 7614 Publisher: Nature Publishing Group, pp. 41–47. doi: 10.1038 / nature18642. URL: <https://www.nature.com/articles/nature18642> (cit. on p. 14).
- [30] Andrew P Morris, Benjamin F Voight, et al. "Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes". In: *Nature genetics* 44.9 (Sept. 2012). PMID: 22885922 PMCID: PMC3442244, pp. 981–990. doi: 10 . 1038 / ng . 2383. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3442244/> (cit. on p. 14).

## CHAPTER 1

- [31] N. Woodford and M. J. Ellington. “The emergence of antibiotic resistance by mutation”. In: *Clinical Microbiology and Infection* 13.1 (Jan. 1, 2007), pp. 5–18. DOI: [10.1111/j.1469-0691.2006.01492.x](https://doi.org/10.1111/j.1469-0691.2006.01492.x). URL: <https://www.sciencedirect.com/science/article/pii/S1198743X14615500> (cit. on p. 14).
- [32] Soo-Yon Rhee, Matthew J. Gonzales, et al. “Human immunodeficiency virus reverse transcriptase and protease sequence database”. In: *Nucleic Acids Research* 31.1 (Jan. 1, 2003), pp. 298–303. DOI: [10.1093/nar/gkg100](https://doi.org/10.1093/nar/gkg100). URL: <https://academic.oup.com/nar/article/31/1/298/2401450> (cit. on pp. 14, 103).
- [33] F. Sanger, S. Nicklen, and A. R. Coulson. “DNA Sequencing with Chain-Terminating Inhibitors”. In: *Proceedings of the National Academy of Sciences* 74.12 (Dec. 1977), pp. 5463–5467. DOI: [10.1073/pnas.74.12.5463](https://doi.org/10.1073/pnas.74.12.5463). URL: <https://www.pnas.org/doi/abs/10.1073/pnas.74.12.5463> (visited on 05/17/2022) (cit. on p. 15).
- [34] Lloyd M. Smith, Steven Fung, et al. “The Synthesis of Oligonucleotides Containing an Aliphatic Amino Group at the 5’ Terminus: Synthesis of Fluorescent DNA Primers for Use in DNA Sequence Analysis”. In: *Nucleic Acids Research* 13.7 (Apr. 11, 1985), pp. 2399–2412. ISSN: 0305-1048. DOI: [10.1093/nar/13.7.2399](https://doi.org/10.1093/nar/13.7.2399). URL: <https://doi.org/10.1093/nar/13.7.2399> (visited on 05/17/2022) (cit. on p. 17).
- [35] Lloyd M. Smith, Jane Z. Sanders, et al. “Fluorescence Detection in Automated DNA Sequence Analysis”. In: *Nature* 321.6071 (6071 June 1986), pp. 674–679. ISSN: 1476-4687. DOI: [10.1038/321674a0](https://doi.org/10.1038/321674a0). URL: <https://www.nature.com/articles/321674a0> (visited on 05/17/2022) (cit. on p. 17).
- [36] Wilhelm Ansorge, Brian Sproat, et al. “Automated DNA sequencing: ultrasensitive detection of fluorescent bands during electrophoresis”. In: *Nucleic Acids Research* 15.11 (June 11, 1987), pp. 4593–4602. DOI: [10.1093/nar/15.11.4593](https://doi.org/10.1093/nar/15.11.4593). URL: <https://doi.org/10.1093/nar/15.11.4593> (cit. on p. 17).
- [37] Jay Shendure and Hanlee Ji. “Next-generation DNA sequencing”. In: *Nature Biotechnology* 26.10 (Oct. 2008). Number: 10 Publisher: Nature Publishing Group, pp. 1135–1145. DOI: [10.1038/nbt1486](https://doi.org/10.1038/nbt1486). URL: <https://www.nature.com/articles/nbt1486> (cit. on p. 17).
- [38] Francis S. Collins, Michael Morgan, and Aristides Patrinos. “The Human Genome Project: Lessons from Large-Scale Biology”. In: *Science* 300.5617 (Apr. 11, 2003). Publisher: American Association for the Advancement of Science, pp. 286–290. DOI: [10.1126/science.1084564](https://doi.org/10.1126/science.1084564). URL: <https://www.science.org/doi/10.1126/science.1084564> (cit. on p. 17).
- [39] Lin Liu, Yinhui Li, et al. “Comparison of Next-Generation Sequencing Systems”. In: *Journal of Biomedicine and Biotechnology* 2012 (July 5, 2012), e251364. ISSN: 2314-6133. DOI: [10.1155/2012/251364](https://doi.org/10.1155/2012/251364). URL: <https://www.hindawi.com/journals/bmri/2012/251364/> (visited on 05/16/2022) (cit. on pp. 17, 19).

- [40] *The Cost of Sequencing a Human Genome*. URL: <https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost> (cit. on p. 18).
- [41] Michael L. Metzker. “Sequencing Technologies — the next Generation”. In: *Nature Reviews Genetics* 11.1 (1 Jan. 2010), pp. 31–46. ISSN: 1471-0064. DOI: [10.1038/nrg2626](https://doi.org/10.1038/nrg2626). URL: <https://www.nature.com/articles/nrg2626> (visited on 05/16/2022) (cit. on p. 18).
- [42] Bruno Canard and Robert S. Sarfati. “DNA polymerase fluorescent substrates with reversible 3'-tags”. In: *Gene* 148.1 (Oct. 11, 1994), pp. 1–6. DOI: [10.1016/0378-1119\(94\)90226-7](https://doi.org/10.1016/0378-1119(94)90226-7). URL: <https://www.sciencedirect.com/science/article/pii/0378111994902267> (cit. on p. 18).
- [43] P. Nyren, B. Pettersson, and M. Uhlen. “Solid Phase DNA Minisequencing by an Enzymatic Luminometric Inorganic Pyrophosphate Detection Assay”. In: *Analytical Biochemistry* 208.1 (Jan. 1, 1993), pp. 171–175. DOI: [10.1006/abio.1993.1024](https://doi.org/10.1006/abio.1993.1024). URL: <https://www.sciencedirect.com/science/article/pii/S0003269783710249> (cit. on p. 19).
- [44] Elaine R. Mardis. “A decade’s perspective on DNA sequencing technology”. In: *Nature* 470.7333 (Feb. 2011). Number: 7333 Publisher: Nature Publishing Group, pp. 198–203. DOI: [10.1038/nature09796](https://doi.org/10.1038/nature09796). URL: <https://www.nature.com/articles/nature09796> (cit. on p. 19).
- [45] Nicholas Stoler and Anton Nekrutenko. “Sequencing error profiles of Illumina sequencing instruments”. In: *NAR Genomics and Bioinformatics* 3.1 (Mar. 1, 2021), lqab019. DOI: [10.1093/nargab/lqab019](https://doi.org/10.1093/nargab/lqab019). URL: <https://doi.org/10.1093/nargab/lqab019> (cit. on p. 19).
- [46] *Sequencing Technology / Sequencing by synthesis*. URL: <https://www.illumina.com/science/technology/next-generation-sequencing/sequencing-technology.html> (cit. on p. 19).
- [47] Martin O Pollard, Deepti Gurdasani, et al. “Long reads: their purpose and place”. In: *Human Molecular Genetics* 27.R2 (Aug. 1, 2018), R234–R241. DOI: [10.1093/hmg/ddy177](https://doi.org/10.1093/hmg/ddy177). URL: <https://doi.org/10.1093/hmg/ddy177> (cit. on p. 19).
- [48] John Eid, Adrian Fehr, et al. “Real-Time DNA Sequencing from Single Polymerase Molecules”. In: *Science* 323.5910 (Jan. 2, 2009). Publisher: American Association for the Advancement of Science, pp. 133–138. DOI: [10.1126/science.1162986](https://doi.org/10.1126/science.1162986). URL: <https://www.science.org/doi/10.1126/science.1162986> (cit. on p. 19).
- [49] M. J. Levene, J. Korlach, et al. “Zero-Mode Waveguides for Single-Molecule Analysis at High Concentrations”. In: *Science* 299.5607 (Jan. 31, 2003). Publisher: American Association for the Advancement of Science, pp. 682–686. DOI: [10.1126/science.1079700](https://doi.org/10.1126/science.1079700). URL: <https://www.science.org/doi/10.1126/science.1079700> (cit. on p. 20).

## CHAPTER 1

- [50] James Clarke, Hai-Chen Wu, et al. “Continuous base identification for single-molecule nanopore DNA sequencing”. In: *Nature Nanotechnology* 4.4 (Apr. 2009). Number: 4 Publisher: Nature Publishing Group, pp. 265–270. DOI: [10.1038/nano.2009.12](https://doi.org/10.1038/nano.2009.12). URL: <https://www.nature.com/articles/nano.2009.12> (cit. on p. 20).
- [51] David Deamer, Mark Akeson, and Daniel Branton. “Three decades of nanopore sequencing”. In: *Nature Biotechnology* 34.5 (May 2016). Number: 5 Publisher: Nature Publishing Group, pp. 518–524. DOI: [10.1038/nbt.3423](https://doi.org/10.1038/nbt.3423). URL: <https://www.nature.com/articles/nbt.3423> (cit. on p. 20).
- [52] Ryan R. Wick, Louise M. Judd, and Kathryn E. Holt. “Performance of neural network basecalling tools for Oxford Nanopore sequencing”. In: *Genome Biology* 20.1 (June 24, 2019), p. 129. DOI: [10.1186/s13059-019-1727-y](https://doi.org/10.1186/s13059-019-1727-y). URL: <https://doi.org/10.1186/s13059-019-1727-y> (cit. on pp. 20, 23, 25).
- [53] Anthony Rhoads and Kin Fai Au. “PacBio Sequencing and Its Applications”. In: *Genomics, Proteomics & Bioinformatics*. SI: Metagenomics of Marine Environments 13.5 (Oct. 1, 2015), pp. 278–289. DOI: [10.1016/j.gpb.2015.08.002](https://doi.org/10.1016/j.gpb.2015.08.002). URL: <https://www.sciencedirect.com/science/article/pii/S1672022915001345> (cit. on p. 20).
- [54] Camilla L.C. Ip, Matthew Loose, et al. “MinION Analysis and Reference Consortium: Phase 1 data release and analysis”. In: *F1000Research* 4 (Oct. 15, 2015). PMID: 26834992 PMCID: PMC4722697, p. 1075. DOI: [10.12688/f1000research.7201.1](https://doi.org/10.12688/f1000research.7201.1). URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4722697/> (cit. on p. 20).
- [55] Glennis A. Logsdon, Mitchell R. Vollger, and Evan E. Eichler. “Long-Read Human Genome Sequencing and Its Applications”. In: *Nature Reviews Genetics* 21.10 (10 Oct. 2020), pp. 597–614. ISSN: 1471-0064. DOI: [10.1038/s41576-020-0236-x](https://doi.org/10.1038/s41576-020-0236-x). URL: <https://www.nature.com/articles/s41576-020-0236-x> (visited on 05/16/2022) (cit. on pp. 20, 21, 23, 25).
- [56] Miten Jain, Sergey Koren, et al. “Nanopore sequencing and assembly of a human genome with ultra-long reads”. In: *Nature Biotechnology* 36.4 (Apr. 2018). Number: 4 Publisher: Nature Publishing Group, pp. 338–345. DOI: [10.1038/nbt.4060](https://doi.org/10.1038/nbt.4060). URL: <https://www.nature.com/articles/nbt.4060> (cit. on p. 20).
- [57] *Thar she blows! Ultra long read method for nanopore sequencing · Loman Labs.* URL: <http://lab.loman.net/2017/03/09/ultrareads-for-nanopore/> (cit. on p. 20).
- [58] Alexander Payne, Nadine Holmes, et al. “BulkVis: a graphical viewer for Oxford nanopore bulk FAST5 files”. In: *Bioinformatics* 35.13 (July 1, 2019), pp. 2193–2198. DOI: [10.1093/bioinformatics/bty841](https://doi.org/10.1093/bioinformatics/bty841). URL: <https://doi.org/10.1093/bioinformatics/bty841> (cit. on p. 20).
- [59] Valentine Murigneux, Subash Kumar Rai, et al. “Comparison of long-read methods for sequencing and assembly of a plant genome”. In: *GigaScience* 9.12 (Nov. 30, 2020), giaa146. DOI: [10.1093/gigascience/giaa146](https://doi.org/10.1093/gigascience/giaa146). URL: <https://doi.org/10.1093/gigascience/giaa146> (cit. on p. 20).

- [60] Mark J. P. Chaisson, John Huddleston, et al. “Resolving the complexity of the human genome using single-molecule sequencing”. In: *Nature* 517.7536 (Jan. 2015). Number: 7536 Publisher: Nature Publishing Group, pp. 608–611. DOI: [10.1038/nature13907](https://doi.org/10.1038/nature13907). URL: <https://www.nature.com/articles/nature13907> (cit. on p. 21).
- [61] Miten Jain, Hugh E. Olsen, et al. “The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community”. In: *Genome Biology* 17.1 (Nov. 25, 2016), p. 239. DOI: [10.1186/s13059-016-1103-0](https://doi.org/10.1186/s13059-016-1103-0). URL: <https://doi.org/10.1186/s13059-016-1103-0> (cit. on p. 21).
- [62] Mingye Hong, Shuang Tao, et al. “RNA sequencing: new technologies and applications in cancer research”. In: *Journal of Hematology & Oncology* 13.1 (Dec. 4, 2020), p. 166. DOI: [10.1186/s13045-020-01005-x](https://doi.org/10.1186/s13045-020-01005-x). URL: <https://doi.org/10.1186/s13045-020-01005-x> (cit. on p. 21).
- [63] Fatih Ozsolak and Patrice M. Milos. “RNA sequencing: advances, challenges and opportunities”. In: *Nature Reviews Genetics* 12.2 (Feb. 2011). Number: 2 Publisher: Nature Publishing Group, pp. 87–98. DOI: [10.1038/nrg2934](https://doi.org/10.1038/nrg2934). URL: <https://www.nature.com/articles/nrg2934> (cit. on p. 21).
- [64] D F Hunt, J R Yates, et al. “Protein sequencing by tandem mass spectrometry.” In: *Proceedings of the National Academy of Sciences* 83.17 (Sept. 1986). Publisher: Proceedings of the National Academy of Sciences, pp. 6233–6237. DOI: [10.1073/pnas.83.17.6233](https://doi.org/10.1073/pnas.83.17.6233). URL: <https://www.pnas.org/doi/abs/10.1073/pnas.83.17.6233> (cit. on p. 21).
- [65] Bryan John Smith. *Protein Sequencing Protocols*. Springer Science & Business Media, 2002 (cit. on p. 21).
- [66] Laura Restrepo-Pérez, Chirlmin Joo, and Cees Dekker. “Paving the way to single-molecule protein sequencing”. In: *Nature Nanotechnology* 13.9 (Sept. 2018). Number: 9 Publisher: Nature Publishing Group, pp. 786–796. DOI: [10.1038/s41565-018-0236-6](https://doi.org/10.1038/s41565-018-0236-6). URL: <https://www.nature.com/articles/s41565-018-0236-6> (cit. on p. 21).
- [67] Jason L Weirather, Mariateresa de Cesare, et al. “Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis”. In: *F1000Research* 6 (June 19, 2017). PMID: 28868132 PMCID: PMC5553090, p. 100. DOI: [10.12688/f1000research.10571.2](https://doi.org/10.12688/f1000research.10571.2). URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5553090/> (cit. on pp. 21, 24).
- [68] Yunhao Wang, Yue Zhao, et al. “Nanopore sequencing technology, bioinformatics and applications”. In: *Nature Biotechnology* 39.11 (Nov. 2021). Number: 11 Publisher: Nature Publishing Group, pp. 1348–1365. DOI: [10.1038/s41587-021-01108-x](https://doi.org/10.1038/s41587-021-01108-x). URL: <https://www.nature.com/articles/s41587-021-01108-x> (cit. on p. 21).
- [69] Xiaotu Ma, Ying Shao, et al. “Analysis of error profiles in deep next-generation sequencing data”. In: *Genome Biology* 20.1 (Mar. 14, 2019), p. 50. DOI: [10.1186/s13059-019-1659-6](https://doi.org/10.1186/s13059-019-1659-6). URL: <https://doi.org/10.1186/s13059-019-1659-6> (cit. on p. 22).

## CHAPTER 1

- [70] Leandro Lima, Camille Marchet, et al. “Comparative assessment of long-read error correction software applied to Nanopore RNA-sequencing data”. In: *Briefings in Bioinformatics* 21.4 (July 15, 2020), pp. 1164–1181. DOI: [10.1093/bib/bbz058](https://doi.org/10.1093/bib/bbz058). URL: <https://doi.org/10.1093/bib/bbz058> (cit. on p. 22).
- [71] Shuhua Fu, Anqi Wang, and Kin Fai Au. “A comparative evaluation of hybrid error correction methods for error-prone long reads”. In: *Genome Biology* 20.1 (Feb. 4, 2019), p. 26. DOI: [10.1186/s13059-018-1605-z](https://doi.org/10.1186/s13059-018-1605-z). URL: <https://doi.org/10.1186/s13059-018-1605-z> (cit. on p. 22).
- [72] Haowen Zhang, Chirag Jain, and Srinivas Aluru. “A comprehensive evaluation of long read error correction methods”. In: *BMC Genomics* 21.6 (Dec. 21, 2020), p. 889. DOI: [10.1186/s12864-020-07227-0](https://doi.org/10.1186/s12864-020-07227-0). URL: <https://doi.org/10.1186/s12864-020-07227-0> (cit. on p. 22).
- [73] Shanika L. Amarasinghe, Shian Su, et al. “Opportunities and Challenges in Long-Read Sequencing Data Analysis”. In: *Genome Biology* 21.1 (Feb. 7, 2020), p. 30. ISSN: 1474-760X. DOI: [10.1186/s13059-020-1935-5](https://doi.org/10.1186/s13059-020-1935-5). URL: <https://doi.org/10.1186/s13059-020-1935-5> (visited on 05/16/2022) (cit. on pp. 22, 23, 25).
- [74] Jue Ruan and Heng Li. “Fast and accurate long-read assembly with wtDBG2”. In: *Nature Methods* 17.2 (Feb. 2020). Number: 2 Publisher: Nature Publishing Group, pp. 155–158. DOI: [10.1038/s41592-019-0669-3](https://doi.org/10.1038/s41592-019-0669-3). URL: <https://www.nature.com/articles/s41592-019-0669-3> (cit. on pp. 22, 25).
- [75] Sergey Koren, Brian P. Walenz, et al. “Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation”. In: *Genome Research* 27.5 (Jan. 5, 2017). Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab PMID: 28298431, pp. 722–736. DOI: [10.1101/gr.215087.116](https://doi.org/10.1101/gr.215087.116). URL: <https://genome.cshlp.org/content/27/5/722> (cit. on p. 22).
- [76] German Tischler and Eugene W. Myers. *Non Hybrid Long Read Consensus Using Local De Bruijn Graph Assembly*. Tech. rep. DOI: [10.1101/106252](https://doi.org/10.1101/106252) Section: New Results Type: article. Feb. 6, 2017, p. 106252. DOI: [10.1101/106252](https://doi.org/10.1101/106252). URL: <https://www.biorxiv.org/content/10.1101/106252v1> (cit. on p. 22).
- [77] René L Warren, Lauren Coombe, et al. “ntEdit: scalable genome sequence polishing”. In: *Bioinformatics* 35.21 (Nov. 1, 2019), pp. 4430–4432. DOI: [10.1093/bioinformatics/btz400](https://doi.org/10.1093/bioinformatics/btz400). URL: <https://doi.org/10.1093/bioinformatics/btz400> (cit. on p. 22).
- [78] N Lance Hepler, M Brown, et al. “An improved circular consensus algorithm with an application to detect HIV-1 Drug-Resistance associated mutations (DRAMs)”. In: 2016 (cit. on p. 22).

- [79] Jared T. Simpson, Rachael E. Workman, et al. “Detecting DNA cytosine methylation using nanopore sequencing”. In: *Nature Methods* 14.4 (Apr. 2017). Number: 4 Publisher: Nature Publishing Group, pp. 407–410. DOI: [10.1038/nmeth.4184](https://doi.org/10.1038/nmeth.4184). URL: <https://www.nature.com/articles/nmeth.4184> (cit. on pp. 22, 25).
- [80] Thomas Hackl, Rainer Hedrich, et al. “proovread : large-scale high-accuracy PacBio correction through iterative short read consensus”. In: *Bioinformatics* 30.21 (Nov. 1, 2014), pp. 3004–3011. DOI: [10.1093/bioinformatics/btu392](https://doi.org/10.1093/bioinformatics/btu392). URL: <https://doi.org/10.1093/bioinformatics/btu392> (cit. on pp. 22, 42).
- [81] Giles Miclotte, Mahdi Heydari, et al. “Jabba: hybrid error correction for long sequencing reads”. In: *Algorithms for Molecular Biology* 11.1 (May 3, 2016), p. 10. DOI: [10.1186/s13015-016-0075-7](https://doi.org/10.1186/s13015-016-0075-7). URL: <https://doi.org/10.1186/s13015-016-0075-7> (cit. on p. 22).
- [82] Sergey Koren, Michael C. Schatz, et al. “Hybrid error correction and de novo assembly of single-molecule sequencing reads”. In: *Nature Biotechnology* 30.7 (July 2012). Number: 7 Publisher: Nature Publishing Group, pp. 693–700. DOI: [10.1038/nbt.2280](https://doi.org/10.1038/nbt.2280). URL: <https://www.nature.com/articles/nbt.2280> (cit. on pp. 22, 42).
- [83] Leena Salmela and Eric Rivals. “LoRDEC: accurate and efficient long read error correction”. In: *Bioinformatics* 30.24 (Dec. 15, 2014), pp. 3506–3514. DOI: [10.1093/bioinformatics/btu538](https://doi.org/10.1093/bioinformatics/btu538). URL: <https://doi.org/10.1093/bioinformatics/btu538> (cit. on p. 22).
- [84] Bruce J. Walker, Thomas Abeel, et al. “Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement”. In: *PLOS ONE* 9.11 (Nov. 19, 2014). Publisher: Public Library of Science, e112963. DOI: [10.1371/journal.pone.0112963](https://doi.org/10.1371/journal.pone.0112963). URL: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0112963> (cit. on p. 22).
- [85] Robert Vaser, Ivan Sović, et al. “Fast and accurate de novo genome assembly from long uncorrected reads”. In: *Genome Research* 27.5 (May 2017). PMID: 28100585 PMCID: PMC5411768, pp. 737–746. DOI: [10.1101/gr.214270.116](https://doi.org/10.1101/gr.214270.116) (cit. on p. 22).
- [86] Aaron M. Wenger, Paul Peluso, et al. “Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome”. In: *Nature Biotechnology* 37.10 (Oct. 2019), pp. 1155–1162. DOI: [10.1038/s41587-019-0217-9](https://doi.org/10.1038/s41587-019-0217-9). URL: [http://www.nature.com/articles/s41587-019-0217-9](https://www.nature.com/articles/s41587-019-0217-9) (cit. on pp. 22, 23, 25).
- [87] Winston Timp, Jeffrey Comer, and Aleksei Aksimentiev. “DNA Base-Calling from a Nanopore Using a Viterbi Algorithm”. In: *Biophysical Journal* 102.10 (May 16, 2012), pp. L37–L39. DOI: [10.1016/j.bpj.2012.04.009](https://doi.org/10.1016/j.bpj.2012.04.009). URL: <https://www.sciencedirect.com/science/article/pii/S0006349512004511> (cit. on p. 23).
- [88] Peter Peresíni, Vladimír Boža, et al. “Nanopore base calling on the edge”. In: *Bioinformatics* 37.24 (Dec. 15, 2021), pp. 4661–4667. DOI: [10.1093/bioinformatics/btab528](https://doi.org/10.1093/bioinformatics/btab528). URL: <https://doi.org/10.1093/bioinformatics/btab528> (cit. on p. 23).

## CHAPTER 1

- [89] Vladimír Boža, Broňa Brejová, and Tomáš Vinař. “DeepNano: Deep recurrent neural networks for base calling in MinION nanopore reads”. In: *PLOS ONE* 12.6 (June 5, 2017). Publisher: Public Library of Science, e0178751. doi: [10.1371/journal.pone.0178751](https://doi.org/10.1371/journal.pone.0178751). URL: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0178751> (cit. on p. 23).
- [90] Søren M. Karst, Ryan M. Ziels, et al. “High-accuracy long-read amplicon sequences using unique molecular identifiers with Nanopore or PacBio sequencing”. In: *Nature Methods* 18.2 (Feb. 2021). Number: 2 Publisher: Nature Publishing Group, pp. 165–169. doi: [10.1038/s41592-020-01041-y](https://doi.org/10.1038/s41592-020-01041-y). URL: <https://www.nature.com/articles/s41592-020-01041-y>. (cit. on p. 23).
- [91] Zitian Chen, Wenxiong Zhou, et al. “Highly accurate fluorogenic DNA sequencing with information theory-based error correction”. In: *Nature Biotechnology* 35.12 (Dec. 2017). Number: 12 Publisher: Nature Publishing Group, pp. 1170–1178. doi: [10.1038/nbt.3982](https://doi.org/10.1038/nbt.3982). URL: <https://www.nature.com/articles/nbt.3982> (cit. on p. 23).
- [92] *High Performance Long Read Assay Enables Contiguous Data up to 10Kb on Existing Illumina Platforms*. URL: [https://www.illumina.com/content/illumina-marketing/amr/en\\_US/science/genomics-research/articles/infinity-high-performance-long-read-assay.html](https://www.illumina.com/content/illumina-marketing/amr/en_US/science/genomics-research/articles/infinity-high-performance-long-read-assay.html) (cit. on pp. 23, 26).
- [93] Gergely Ivády, László Madar, et al. “Analytical parameters and validation of homopolymer detection in a pyrosequencing-based next generation sequencing system”. In: *BMC Genomics* 19.1 (Feb. 21, 2018), p. 158. doi: [10.1186/s12864-018-4544-x](https://doi.org/10.1186/s12864-018-4544-x). URL: <https://doi.org/10.1186/s12864-018-4544-x> (cit. on p. 24).
- [94] A. Sina Booeshaghi and Lior Pachter. “Pseudoalignment facilitates assignment of error-prone Ultima Genomics reads”. In: (). doi: [10.1101/2022.06.04.494845](https://doi.org/10.1101/2022.06.04.494845) (cit. on p. 24).
- [95] Clara Delahaye and Jacques Nicolas. “Sequencing DNA with nanopores: Troubles and biases”. In: *PLOS ONE* 16.10 (Oct. 1, 2021). Publisher: Public Library of Science, e0257521. doi: [10.1371/journal.pone.0257521](https://doi.org/10.1371/journal.pone.0257521). URL: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0257521> (cit. on p. 24).
- [96] Sara Goodwin, James Gurtowski, et al. “Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome”. In: *Genome Research* 25.11 (Jan. 11, 2015). Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab PMID: 26447147, pp. 1750–1756. doi: [10.1101/gr.191395.115](https://doi.org/10.1101/gr.191395.115). URL: <https://genome.cshlp.org/content/25/11/1750> (cit. on p. 24).
- [97] Juliane C Dohm, Philipp Peters, et al. “Benchmarking of Long-Read Correction Methods”. In: *NAR Genomics and Bioinformatics* 2.2 (June 1, 2020). ISSN: 2631-9268. doi: [10.1093/nargab/lqaa037](https://doi.org/10.1093/nargab/lqaa037) (cit. on pp. 24, 70).

- [98] Jonathan Foox, Scott W. Tighe, et al. “Performance assessment of DNA sequencing platforms in the ABRF Next-Generation Sequencing Study”. In: *Nature Biotechnology* 39.9 (Sept. 2021). Number: 9 Publisher: Nature Publishing Group, pp. 1129–1140. DOI: [10.1038/s41587-021-01049-5](https://doi.org/10.1038/s41587-021-01049-5). URL: <https://www.nature.com/articles/s41587-021-01049-5> (cit. on p. 25).
- [99] Yao-Ting Huang, Po-Yu Liu, and Pei-Wen Shih. “Homopolish: a method for the removal of systematic errors in nanopore sequencing by homologous polishing”. In: *Genome Biology* 22.1 (Mar. 31, 2021), p. 95. DOI: [10.1186/s13059-021-02282-6](https://doi.org/10.1186/s13059-021-02282-6). URL: <https://doi.org/10.1186/s13059-021-02282-6> (cit. on p. 25).
- [100] Franka J. Rang, Wigard P. Kloosterman, and Jeroen de Ridder. “From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy”. In: *Genome Biology* 19.1 (July 13, 2018), p. 90. DOI: [10.1186/s13059-018-1462-9](https://doi.org/10.1186/s13059-018-1462-9). URL: <https://doi.org/10.1186/s13059-018-1462-9> (cit. on p. 25).
- [101] Peter Sarkozy, Ákos Jobbág, and Peter Antal. “Calling Homopolymer Stretches from Raw Nanopore Reads by Analyzing k-mer Dwell Times”. In: ed. by Hannu Eskola, Outi Väisänen, et al. IFMBE Proceedings. Singapore: Springer, 2018, pp. 241–244. DOI: [10.1007/978-981-10-5122-7\\_61](https://doi.org/10.1007/978-981-10-5122-7_61) (cit. on p. 25).
- [102] *R10.3: the newest nanopore for high accuracy nanopore sequencing – now available in store*. Section: News. URL: <http://nanoporetech.com/about-us/news/r103-newest-nanopore-high-accuracy-nanopore-sequencing-now-available-store> (cit. on p. 25).
- [103] Lei Zhou, Kun Li, et al. “Detection of DNA homopolymer with graphene nanopore”. In: *Journal of Vacuum Science & Technology B* 37.6 (Nov. 2019). Publisher: American Vacuum Society, p. 061809. DOI: [10.1116/1.5116295](https://doi.org/10.1116/1.5116295). URL: <https://avs.scitation.org/doi/full/10.1116/1.5116295> (cit. on p. 25).
- [104] Yusuke Goto, Itaru Yanagi, et al. “Identification of four single-stranded DNA homopolymers with a solid-state nanopore in alkaline CsCl solution”. In: *Nanoscale* 10.44 (2018). Publisher: Royal Society of Chemistry, pp. 20844–20850. DOI: [10.1039/C8NR04238A](https://doi.org/10.1039/C8NR04238A). URL: <https://pubs.rsc.org/en/content/articlelanding/2018/nr/c8nr04238a> (cit. on p. 25).
- [105] John A. Hawkins, Stephen K. Jones, et al. “Indel-correcting DNA barcodes for high-throughput sequencing”. In: *Proceedings of the National Academy of Sciences* 115.27 (July 3, 2018). Publisher: Proceedings of the National Academy of Sciences, E6217–E6226. DOI: [10.1073/pnas.1802640115](https://doi.org/10.1073/pnas.1802640115). URL: <https://www.pnas.org/doi/full/10.1073/pnas.1802640115> (cit. on p. 25).
- [106] Amrita Srivathsan, Bilgenur Baloğlu, et al. “A MinION™-based pipeline for fast and cost-effective DNA barcoding”. In: *Molecular Ecology Resources* 18.5 (2018). \_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/1755-0998.12890>, pp. 1035–1049. DOI: [10.1111/1755-0998.12890](https://doi.org/10.1111/1755-0998.12890). URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/1755-0998.12890> (cit. on p. 25).

## CHAPTER 1

- [107] Yixin Wang, Md. Noor-A-Rahim, et al. “Construction of Bio-Constrained Code for DNA Data Storage”. In: *IEEE Communications Letters* 23.6 (June 2019). Conference Name: IEEE Communications Letters, pp. 963–966. DOI: [10.1109/LCOMM.2019.2912572](https://doi.org/10.1109/LCOMM.2019.2912572) (cit. on p. 25).
- [108] Sergey Nurk, Brian P. Walenz, et al. “HiCanu: Accurate Assembly of Segmental Duplications, Satellites, and Allelic Variants from High-Fidelity Long Reads”. In: *Genome Research* 30.9 (Jan. 9, 2020), pp. 1291–1305. ISSN: 1088-9051, 1549-5469. DOI: [10.1101/gr.263566.120](https://doi.org/10.1101/gr.263566.120). pmid: [32801147](#) (cit. on pp. 25, 71).
- [109] Barış Ekim, Bonnie Berger, and Rayan Chikhi. “Minimizer-Space de Bruijn Graphs: Whole-Genome Assembly of Long Reads in Minutes on a Personal Computer”. In: *Cell Systems* 12.10 (Oct. 20, 2021), 958–968.e6. ISSN: 2405-4712. DOI: [10.1016/j.cels.2021.08.009](https://doi.org/10.1016/j.cels.2021.08.009) (cit. on pp. 25, 71).
- [110] Jason R. Miller, Arthur L. Delcher, et al. “Aggressive Assembly of Pyrosequencing Reads with Mates”. In: *Bioinformatics* 24.24 (Dec. 15, 2008), pp. 2818–2824. ISSN: 1367-4803. DOI: [10.1093/bioinformatics/btn548](https://doi.org/10.1093/bioinformatics/btn548) (cit. on pp. 25, 71).
- [111] Kishwar Shafin, Trevor Pesout, et al. “Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes”. In: *Nature Biotechnology* 38.9 (Sept. 2020). Number: 9 Publisher: Nature Publishing Group, pp. 1044–1053. DOI: [10.1038/s41587-020-0503-6](https://doi.org/10.1038/s41587-020-0503-6). URL: <https://www.nature.com/articles/s41587-020-0503-6> (cit. on p. 25).
- [112] Kristoffer Sahlin and Paul Medvedev. “De Novo Clustering of Long-Read Transcriptome Data Using a Greedy, Quality Value-Based Algorithm”. In: *Journal of Computational Biology* 27.4 (Apr. 1, 2020), pp. 472–484. DOI: [10.1089/cmb.2019.0299](https://doi.org/10.1089/cmb.2019.0299) (cit. on pp. 25, 71).
- [113] Kin Fai Au, Jason G. Underwood, et al. “Improving PacBio Long Read Accuracy by Short Read Alignment”. In: *PLOS ONE* 7.10 (Oct. 4, 2012), e46679. ISSN: 1932-6203. DOI: [10.1371/journal.pone.0046679](https://doi.org/10.1371/journal.pone.0046679) (cit. on pp. 25, 71).
- [114] Ruifeng Hu, Guibo Sun, and Xiaobo Sun. “LSCplus: a fast solution for improving long read accuracy by short read alignment”. In: *BMC Bioinformatics* 17.1 (Nov. 9, 2016), p. 451. DOI: [10.1186/s12859-016-1316-y](https://doi.org/10.1186/s12859-016-1316-y). URL: <https://doi.org/10.1186/s12859-016-1316-y> (cit. on p. 25).
- [115] Heng Li. “Minimap2: Pairwise Alignment for Nucleotide Sequences”. In: *Bioinformatics* 34.18 (Sept. 15, 2018), pp. 3094–3100. ISSN: 1367-4803. DOI: [10.1093/bioinformatics/bty191](https://doi.org/10.1093/bioinformatics/bty191) (cit. on pp. 25, 48, 51, 71, 78, 79, 159).
- [116] Chirag Jain, Arang Rhie, et al. “Weighted Minimizer Sampling Improves Long Read Mapping”. In: *Bioinformatics* 36 (Supplement\_1 July 1, 2020), pp. i111–i118. ISSN: 1367-4803. DOI: [10.1093/bioinformatics/btaa435](https://doi.org/10.1093/bioinformatics/btaa435) (cit. on pp. 25, 52, 81, 159).

## REFERENCES FOR CHAPTER 1

- [117] Christophe Van Neste, Filip Van Nieuwerburgh, et al. “Forensic STR analysis using massive parallel sequencing”. In: *Forensic Science International: Genetics*. Analysis and biostatistical interpretation of complex and low template DNA samples 6.6 (Dec. 1, 2012), pp. 810–818. DOI: [10.1016/j.fsigen.2012.03.004](https://doi.org/10.1016/j.fsigen.2012.03.004). URL: <https://www.sciencedirect.com/science/article/pii/S1872497312000798> (cit. on p. 25).
- [118] Gilad Almogy, Mark Pratt, et al. “Cost-efficient whole genome-sequencing using novel mostly natural sequencing-by-synthesis chemistry and open fluidics platform”. In: (). DOI: [10.1101/2022.05.29.493900](https://doi.org/10.1101/2022.05.29.493900) (cit. on p. 26).
- [119] *Short-read sequencing by binding*. URL: <https://www.pacb.com/technology/sequencing-by-binding/> (cit. on p. 26).
- [120] Arif E. Cetin, Pinar Iyidogan, et al. “Plasmonic Sensor Could Enable Label-Free DNA Sequencing”. In: *ACS Sensors* 3.3 (Mar. 23, 2018). Publisher: American Chemical Society, pp. 561–568. DOI: [10.1021/acssensors.7b00957](https://doi.org/10.1021/acssensors.7b00957). URL: <https://doi.org/10.1021/acssensors.7b00957> (cit. on p. 26).



## 2. Aligning sequence data

### 2.1. What is an alignment ?

We want to compare individuals, species, whatever. To do this we need to compare what is comparable. Alignment to the rescue.

- We want to find similarities in 2 sequences and put these similarities together, so we can compare sequences.
- When we align 2 sequences together we talk about **pairwise** alignment as opposed to **multiple** alignment where we align more than 2 sequences all together. We will first focus on pairwise alignment.

2 ways of going about for pairwise alignment [121]:

- **Global alignment:** we look at the entirety of the 2 sequences and take all that information into account when aligning
- **Local alignment:** we look at the sequences portion by portion, trying to find the best place where they match up.

#### 2.1.1. Why align ?

- hamming distance is an easy method [122]
  - Although it is easy to compute -> sequences must have same length
  - When dealing with DNA/proteins we have to be able to deal with insertions/deletions and hamming cannot do that...
- This is very similar to other well known problems in computer science: the string-edit problem [123] and the Levenshtein distance [124]
- Alignments are used in many cases so that:
  - we can examine similarities/differences between them (i.e. comparative genomics ) [125]
  - we can infer (usually with multiple alignment):

## CHAPTER 2

- \* evolutionary relationships (phylogenetics), and most methods to reconstruct phylogenetic trees take as input a multiple alignment: [126, 127, 128, 129, 130]
- \* protein:
  - structure [131, 132]
  - function [133, 134]
- we can correct sequencing errors [80, 82, 135].
- Structural variant detection [136, 137]

### 2.1.2. How to align two sequences ?

The seminal method for pairwise alignment was the Needleman-Wünsch algorithm [138] based on a dynamic programming method. A decade later, the Smith-Waterman algorithm [139] was developed with similar ideas to perform local alignment. Both are still used today for pairwise alignment.

Dynamic programming is a method to solve complex problems by breaking it into smaller sub-problems and solving each one optimally and separately [140, 141], it is widely used when we wish to have a precise alignment between 2 sequences.

#### 2.1.2.1. Global alignment

- We do global alignment when we expect two sequences to be related and of similar genes/proteins, etc...
- short presentation of NW algo (c.f. Figure 2.1):
  - The score of an alignment can be defined recursively -> dynamic programming
  - Fill out DP matrix:
  - Traceback to find optimal alignment: start in the bottom right and follow the arrows up to the upper left. Diagonal: Match/mismatch, vertical arrow: gap in “columns” sequence, horizontal arrow: gap in “rows” sequence.
- This algorithm although very precise, has a time complexity of  $O(nm)$  where  $n$  and  $m$  are the lengths of the sequences to align [121]. Some methods have been proposed to speed up [142], however the complexity is still  $O(nm / \log(n))$ . Lower bounds have been studied and there is not much optimization we can do if we want to keep an optimal exact alignment [143, 144]. If we want to do better we have to rely on heuristics.

## 2.1. WHAT IS AN ALIGNMENT ?

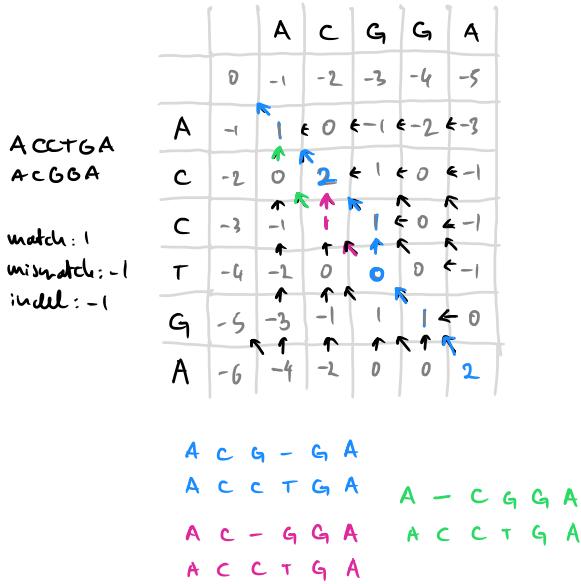


Figure 2.1.: **Example global alignment with the Needleman-Wunsch algorithm**  
The match score is 1, the mismatch and indel scores are both set to -1. This results in 3 equally optimal global alignments.

- Another issue is space complexity since we need to store the matrix, we get  $O(nm)$  as well for space, so if we wish to align 2 human genomes we would need to store  $\approx 10^{19}$  matrix cells, which would amount to 10 Exabytes of storage if we use 8bit integers (meaning it would take an entire data center to store that).
  - However in practice we can do much better than that, and construct an optimal alignment in linear space complexity  $O(n + m)$  [145] meaning we would only need a couple gigabytes to store the matrix for 2 human genomes.
  - Above idea resulting in Myers-miller algorithm [146], implemented in the EMBOSS stretcher tool [147]

### 2.1.2.2. Local alignment

- We want to align a subsequence to a subsequence.
- Similar ideas to NW, except we start the traceback at the maximally valued cell and stop when we hit a score of 0. i.e.
- In the example in Figure 2.1 there are two possible starting cells with a score of 2. This will yield 2 optimal local alignments: AC to AC and GA to GA.

## CHAPTER 2

- In terms of complexity is the same as NW in quadratic in time and space [121] but can be taken down to linear space with the same approach as NW.
- Optimizations were made and resulted in the Huang and Miller algorithm [148] which is implemented in the EMBOSS Lalign tool [147], and the Waterman Eggert algorithm [149].

Both methods are implemented in many different software tools and are used when performing pairwise alignments of short sequences [150, 151, 147] with version implemented for specific CPU instruction sets [152] or GPUs [153] to speed up alignment.

### 2.1.3. Substitution models / scoring

When scoring an alignment we can use different scoring models/substitution matrices [154]:

- A lot of work has been done on protein scoring matrices
  - Log-Odds models, based on the fact that mutations are not equiprobable, and some mutations will be much more common ( $I \leftrightarrow L$  in proteins)
  - PAM (Point Accepted Mutations) [155], gotten from studying closely related protein sequences, estimate the probability of one amino acid changing to another one over time. With more data refinements have been made for PAM-like matrices [156].
  - BLOSUM [157], similar idea to PAM except they were constructed on several “blocks”. A block is a segment of a protein that is very well conserved within a family and computed the probabilities on these blocks.
  - Some matrices estimated with ML instead of log-odds methods [158, 159]
  - Model specific matrices:
    - \* Transmembrane matrices [160, 161]
    - \* Disordered regions in proteins: [162]
    - \* context-specific matrices [163]
    - \* Specific organisms like *P. falciparum* with pfSSM [164] or HIV [165]
    - \* specific to global alignment [166]
- Some on DNA alignment
  - You can derive a matrix with methods similar to PAM [167]
  - You can do codon substitution matrices [168], or combine codon matrices with AA matrices [169]

## 2.2. HOW DO WE SPEED UP PAIRWISE ALIGNMENT ?

### 2.1.4. Dealing with gaps

- biologically longer deletions are more likely than plenty of shorter ones
  - Here a short example of 2 alignments with different gap strategies
- Semi global alignment (i.e. gaps on the ends of the sequence are free...)
- Affine gap penalties, proposed by Gotoh [170]
- Non affine gap penalties [171, 172, 173]

## 2.2. How do we speed up pairwise alignment ?

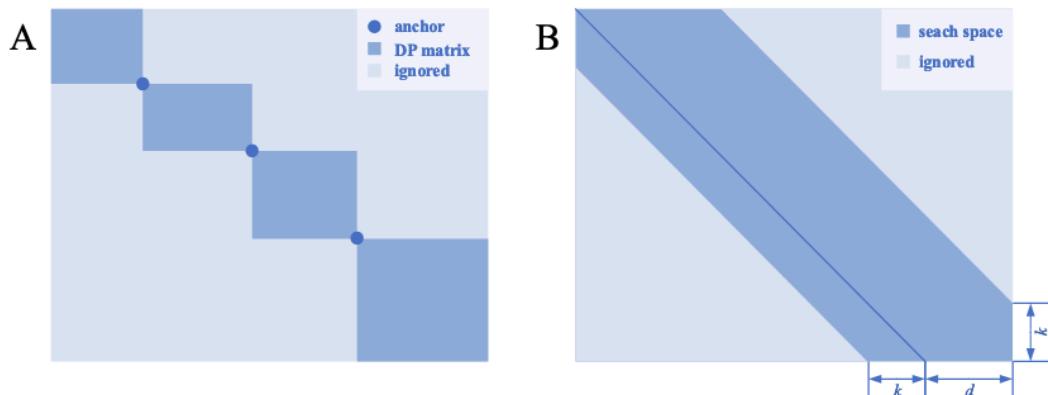
Although NW and SW give us optimal alignments [174] when dealing with large sequences they become impractical due to time and space complexity. We need heuristics, review [175], .

### 2.2.1. Change the method

- Bounded DP [176, 177], we can make an assumption about the relationship between the sequences, there will probably not be many gaps, therefore the scores we will used will be concentrated around the diagonal of the DP matrix c.f. 2.2
- New Algorithms like the wavefront algorithm [178]
- HMMs:
  - PairHMMs can be used to align 2 sequences, in some cases it is mathematically equivalent to NW [179].
  - Software: HHsearch [180], HMMer [181], MCALIGN2 [182] is used to efficiently search for alignments in large databases
- You can also align with Fast Fourier Transforms (FFT) as in MAFFT [183] by quickly computing correlations with FFT to find homologous regions and then using these to align.

### 2.2.2. Seed and extend with data structures

- General idea:
  - Seed + extend for local alignment
  - Find anchors for global alignment
  - In both cases: divide and conquer approach: as you can restrict the DP matrix by blocks defined by the anchors/seeds c.f. 2.2
  - Many ways of finding seeds [184] and to index them [185]
  - 2 use cases:
    - \* global alignment: align 2 large sequences that would overload computing resources.
    - \* local alignment:
      - to a database: ie. find hits in many sequences.
      - to a reference: ie find best hit in one sequence.



**Figure 1.** Two heuristic algorithms for pairwise sequence alignment. (A) A dynamic programming matrix, which is separated by several anchors, which is certain to be in the optimal path. (B) A shape-based bounded dynamic programming matrix in which the light blue block is calculation-free because these states are thought to be less likely to be in the optimal path.

#### Figure 2.2.: 2 heuristic methods to speed up alignment:

divide and conquer and bounded dynamic programming. Adapted from [175] ([Original figure here as a placeholder, I will adapt it](#))

- Building indices/databases (useful when you want to try aligning a query sequence to a bunch of possible targets, i.e. in order to search for homology, this does approximate local alignment):

## 2.2. HOW DO WE SPEED UP PAIRWISE ALIGNMENT ?

### 2.2.2.1. Hash tables

- BLAST [186]
  - We break up our sequences into overlapping “words” i.e. all possible short sub-sequences.
  - Construct a database of words with positions in the reference sequence(s)
  - Get list of words for your query sequence and generate possible “hits”, i.e. words that align with your query words with a score higher than a threshold.
  - Scan the database for these generated words, if you find one it is a hit and these 2 words are a candidate position for a local alignment.
  - extend the local alignment with SW in both directions from the hit to get a local alignment, stop extending if the alignment score gets too low.
  - Plenty of variants: BLASTP for proteins, BLASTN or MEGABLAST for DNA, BLASTX for comparing DNA query to a protein database, PSI-BLAST that iteratively refines the alignment by building multiple alignments from the statistically significant alignments, learns a specific scoring matrix from this and starts over, other people have refined the heuristic for even faster seed prediction as in UBLAST [187].
  - It is important to note that it is a heuristic method, it has a decreased sensitivity to the SW algorithm which is optimal [188], it is however much faster and can allow users to search for similarities among millions of sequences at once
  - Particularly useful, one of the most cited papers in the world: quickly search for homologs of unknown sequences, ... available as a web service hosted by NCBI (<https://blast.ncbi.nlm.nih.gov/Blast.cgi>).
- FASTA [189], derived from FASTP [190] which could not deal with gaps, preceded BLAST.
  - Similarly to BLAST, break query up into overlapping words
  - Scan the library for exact-matching words to build up regions with high similarity (save 10 best regions)
  - Score regions with a substitution matrix
  - High scoring regions are combined to build an approximate alignment
  - Highest scoring ungapped alignment is realigned with banded SW.

- Both methods are very Fast, taking only a couple seconds to find approximate local alignments for 100 sequences [191] in a database of over 80 million sequences [192], with HMMs being quite slow. It is much faster to use these than to try to do the same task with SW or NW [193].
- Both BLAST and FASTA compute a statistical p-value for a pairwise alignment and an expected number of random hits when scanning the database (E-value) [194, 195]
- Minimap2 [115] minimizer (Figure 2.3) hash table
- Diamond [196, 197] which indexes both the reference and query at the same time.



Figure 2.3.: Example of kmers sharing a minimizer

#### 2.2.2.2. Suffix trees / suffix arrays

- Used in many pattern matching problems [123].
- Software: AVID uses suffix trees to find anchors [198], MUMmer finds largest identical subsequences between 2 sequences to anchor alignments using suffix trees [199]

#### 2.2.2.3. FM Index

- FM index [200]:
- In some cases when using suffix trees take too much memory we can use an FM index which is based on the burrows wheeler transform [201].
- used for in exact and approximate string matching [121]

## 2.3. MULTIPLE SEQUENCE ALIGNMENT

- Software: BWT-SW uses an FM-index to speed up local alignment [202], Bowtie 2 uses an FM index to find seeds [203], BWA and BWA-SW use a similar idea [204, 205], BWA-MEM [206] and CUSHAW [207] also uses FM indices to find exact matches to seed a local alignment .

### 2.3. Multiple sequence alignment

When we need to compare a lot of individuals together we can do MSA, essential task in many bioinformatics analyses [208].

NP-hard [209, 210] problem if you do it with DP so we need heuristics or tricks

Even if we align all sequences pairwise we need to then combine all gaps and stuff -> complicated.

#### 2.3.1. Progressive alignment

guide tree, clustering of sequences then refine alignment. Good heuristic but with larger datasets, becomes harder. [211] Main MSA method.

1. Compute pairwise distance matrix for sequence set:
  - either by doing  $N(N-1)/2$  pairwise alignments
  - Or alignment free methods to speed things up [212, 213]
2. Build guide tree from distances (neighbor joining, UPGMA, ...)
3. Align sequences one by one according to the tree, from the leafs (i.e sequences) to the root (full MSA).

Problems -> keeps gaps and if bad alignment at first steps error propagates (“once a gap, always a gap” [211]).

In order to curtail this problem we have iterative refinement [208]:

1. create MSA (e.g. progressive)
2. divide MSA into 2 groups + remove columns with only gaps
3. realign with profile alignment
4. Redo 2+3 until no improvement is made, according to some scoring function (Weighted sum-of-pair [214], or others like log-odds, correlation [215] or consistency [216])

Some of the most used MSA software uses these methods of progressive/iterative refinement:

- CLUSTAL-W [217] and CLUSTAL-X [218]

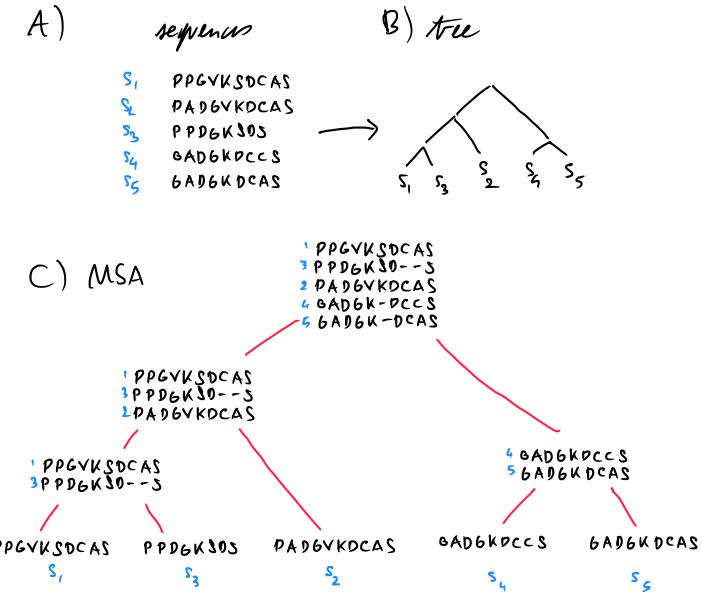


Figure 2.4.: Overview of the progressive alignment process.

A) sequences to align B) guide tree constructed from distances C) Alignment steps along the guide tree and resulting MSA. Adapted from [121]

- T-Coffee [219]
- MUSCLE [220, 221]
- MAFFT [183]
- ProbCons [222]

Different strengths/weaknesses so many reviews and benchmarks to make a choice [223, 224, 225, 226, 227]

### 2.3.2. HMMs in profile-profile or seq-profile alignments

HMM method similar performance to clustal-w [228], HHMer [181], MSAProbs [229] and COVID-align [230]

Example of COVID where homology is high so we can get away with aligning pairwise to ancestral-sequence NextClade/NextAlign [231]

### 2.3.3. Other methods (short)

- Simulated annealing to speed up DP [232, 233, 234]
- Genetic algorithm, review [235], can speed things up [236]
  - SAGA [237], VGDA [238], GAPAM [239]
- recently: regressive method [240] root to leaf allows to align a very large number of sequences: 1.4 million!

## 2.4. The specificities of read-mapping

Huge review of mapping in [241]

### 2.4.1. What is read-mapping ?

Task: locally/semi-globally align independent sequencing reads to a reference genome.

Plenty of aligners both for short reads (reviews [242, 243, 244]) and long-reads ([115]. Active field and there are benchmarks [245] and reviews [241, 246].

- Mapping quality:
  - Intro and definition, quite a loosely defined term.
    - \* Different definitions/implementations ?
  - Mapping quality from tandem simulation [247]
- Technical aspects to speed up, since we usually have a couple refs and many query reads, the index strategy described earlier is particularly well suited:
  - hash tables: [115]
  - FM index / BWT: BWT-SW [202], Bowtie 2 [203], BWA, BWA-SW [204, 205], BWA-MEM [206] and CUSHAW [207].
- Divide and conquer as well: [248]
- These heuristics also become more and more needed as sequencing technologies yield longer and longer reads.
- Also hardware acceleration for short [249] and long [250, 251, 252] reads.

### 2.4.2. Challenges of read-mapping

- low homology / sequencing errors... make mapping and other tasks hard [123]
  - Particularly with long reads since sequencing error rate is higher: specific long read mappers that take this into account:
    - \* PB: BLASR [253], lordFAST [254]
    - \* ONT: GraphMap [255]
    - \* Both: NGMLR [256], MashMap [257], DuploMap [258]
- Repetitive regions (centromeres, telomeres) make it hard to map [241]. Specific tools have been developed for this:
  - winnowmap [116], winnowmap2 [259]
  - tandemtools [260] for tandem repeats.
- homopolymers can also be a problem, so HPC is implemented in mappers \*(c.f. 1.4.3.2)\*

## References for chapter 2

- [80] Thomas Hackl, Rainer Hedrich, et al. “proovread : large-scale high-accuracy PacBio correction through iterative short read consensus”. In: *Bioinformatics* 30.21 (Nov. 1, 2014), pp. 3004–3011. DOI: [10.1093/bioinformatics/btu392](https://doi.org/10.1093/bioinformatics/btu392). URL: <https://doi.org/10.1093/bioinformatics/btu392> (cit. on pp. 22, 42).
- [82] Sergey Koren, Michael C. Schatz, et al. “Hybrid error correction and de novo assembly of single-molecule sequencing reads”. In: *Nature Biotechnology* 30.7 (July 2012). Number: 7 Publisher: Nature Publishing Group, pp. 693–700. DOI: [10.1038/nbt.2280](https://doi.org/10.1038/nbt.2280). URL: <https://www.nature.com/articles/nbt.2280> (cit. on pp. 22, 42).
- [115] Heng Li. “Minimap2: Pairwise Alignment for Nucleotide Sequences”. In: *Bioinformatics* 34.18 (Sept. 15, 2018), pp. 3094–3100. ISSN: 1367-4803. DOI: [10.1093/bioinformatics/bty191](https://doi.org/10.1093/bioinformatics/bty191) (cit. on pp. 25, 48, 51, 71, 78, 79, 159).
- [116] Chirag Jain, Arang Rhie, et al. “Weighted Minimizer Sampling Improves Long Read Mapping”. In: *Bioinformatics* 36 (Supplement\_1 July 1, 2020), pp. i111–i118. ISSN: 1367-4803. DOI: [10.1093/bioinformatics/btaa435](https://doi.org/10.1093/bioinformatics/btaa435) (cit. on pp. 25, 52, 81, 159).
- [121] Wing-Kin Sung. *Algorithms in Bioinformatics: A Practical Introduction*. DOI: [10.1201/9781420070347](https://doi.org/10.1201/9781420070347). New York: Chapman and Hall/CRC, Oct. 10, 2011. DOI: [10.1201/9781420070347](https://doi.org/10.1201/9781420070347) (cit. on pp. 41, 42, 44, 48, 50).
- [122] Richard Wesley Hamming. *Coding and Information Theory*. tex.ids= hamming1980coding googlebooksid: ed5QAAAAMAAJ lccn: 79015159. Prentice-Hall, 1980 (cit. on p. 41).

- [123] Dan Gusfield. *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*. DOI: 10.1017/CBO9780511574931. Cambridge: Cambridge University Press, 1997. doi: 10.1017/CBO9780511574931. URL: <https://www.cambridge.org/core/books/algorithms-on-strings-trees-and-sequences/F0B095049C7E6EF5356F0A26686C20D3> (cit. on pp. 41, 48, 52, 70).
- [124] V. I. Levenshtein. “Binary Codes Capable of Correcting Deletions, Insertions and Reversals”. In: *Soviet Physics Doklady* 10 (Feb. 1, 1966). ADS Bibcode: 1966SPhD...10..707L, p. 707. URL: <https://ui.adsabs.harvard.edu/abs/1966SPhD..10..707L> (cit. on p. 41).
- [125] Ross C. Hardison. “Comparative Genomics”. In: *PLOS Biology* 1.2 (Nov. 17, 2003). Publisher: Public Library of Science, e58. doi: 10.1371/journal.pbio.0000058. URL: <https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.0000058> (cit. on p. 41).
- [126] Joseph Felsenstein. “Evolutionary trees from DNA sequences: A maximum likelihood approach”. In: *Journal of Molecular Evolution* 17.6 (Nov. 1, 1981), pp. 368–376. doi: 10.1007/BF01734359. URL: <https://doi.org/10.1007/BF01734359> (cit. on p. 42).
- [127] Sudhir Kumar, Koichiro Tamura, and Masatoshi Nei. “MEGA: Molecular Evolutionary Genetics Analysis software for microcomputers”. In: *Bioinformatics* 10.2 (Apr. 2, 1994), pp. 189–191. doi: 10.1093/bioinformatics/10.2.189. URL: <https://doi.org/10.1093/bioinformatics/10.2.189> (cit. on p. 42).
- [128] Alexey M Kozlov, Diego Darriba, et al. “RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference”. In: *Bioinformatics* 35.21 (Nov. 1, 2019), pp. 4453–4455. doi: 10.1093/bioinformatics/btz305. URL: <https://doi.org/10.1093/bioinformatics/btz305> (cit. on p. 42).
- [129] Stéphane Guindon, Jean-François Dufayard, et al. “New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0”. In: *Systematic Biology* 59.3 (May 1, 2010), pp. 307–321. doi: 10.1093/sysbio/syq010. URL: <https://doi.org/10.1093/sysbio/syq010> (cit. on p. 42).
- [130] Morgan N. Price, Paramvir S. Dehal, and Adam P. Arkin. “FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments”. In: *PLOS ONE* 5.3 (Mar. 10, 2010), e9490. doi: 10.1371/journal.pone.0009490. URL: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0009490> (cit. on p. 42).
- [131] John Jumper, Richard Evans, et al. “Highly accurate protein structure prediction with AlphaFold”. In: *Nature* 596.7873 (Aug. 2021). Number: 7873 Publisher: Nature Publishing Group, pp. 583–589. doi: 10.1038/s41586-021-03819-2. URL: <https://www.nature.com/articles/s41586-021-03819-2> (cit. on p. 42).

## CHAPTER 2

- [132] Kevin Karplus, Christian Barrett, et al. “Predicting protein structure using only sequence information”. In: *Proteins: Structure, Function, and Bioinformatics* 37.S3 (1999), pp. 121–125. DOI: [10.1002/\(SICI\)1097-0134\(1999\)37:3+<121::AID-PROT16>3.0.CO;2-Q](https://doi.org/10.1002/(SICI)1097-0134(1999)37:3+<121::AID-PROT16>3.0.CO;2-Q). URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291097-0134%281999%2937%3A3%20%3C121%3A%3AAID-PROT16%3E3.0.CO%3B2-Q> (cit. on p. 42).
- [133] James D Watson, Roman A Laskowski, and Janet M Thornton. “Predicting protein function from sequence and structural data”. In: *Current Opinion in Structural Biology*. Sequences and topology/Nucleic acids 15.3 (June 1, 2005), pp. 275–284. DOI: [10.1016/j.sbi.2005.04.003](https://doi.org/10.1016/j.sbi.2005.04.003). URL: <https://www.sciencedirect.com/science/article/pii/S0959440X05000825> (cit. on p. 42).
- [134] David Lee, Oliver Redfern, and Christine Orengo. “Predicting protein function from sequence and structure”. In: *Nature Reviews Molecular Cell Biology* 8.12 (Dec. 2007), pp. 995–1005. DOI: [10.1038/nrm2281](https://doi.org/10.1038/nrm2281). URL: <https://www.nature.com/articles/nrm2281> (cit. on p. 42).
- [135] Leena Salmela and Jan Schröder. “Correcting errors in short reads by multiple alignments”. In: *Bioinformatics* 27.11 (June 1, 2011), pp. 1455–1461. DOI: [10.1093/bioinformatics/btr170](https://doi.org/10.1093/bioinformatics/btr170). URL: <https://doi.org/10.1093/bioinformatics/btr170> (cit. on p. 42).
- [136] Paul Medvedev, Monica Stanciu, and Michael Brudno. “Computational methods for discovering structural variation with next-generation sequencing”. In: *Nature Methods* 6.11 (Nov. 2009). Number: 11 Publisher: Nature Publishing Group, S13–S20. DOI: [10.1038/nmeth.1374](https://doi.org/10.1038/nmeth.1374). URL: <https://www.nature.com/articles/nmeth.1374> (cit. on p. 42).
- [137] Medhat Mahmoud, Nastassia Gobet, et al. “Structural variant calling: the long and the short of it”. In: *Genome Biology* 20.1 (Nov. 20, 2019), p. 246. DOI: [10.1186/s13059-019-1828-7](https://doi.org/10.1186/s13059-019-1828-7). URL: <https://doi.org/10.1186/s13059-019-1828-7> (cit. on p. 42).
- [138] Saul B. Needleman and Christian D. Wunsch. “A general method applicable to the search for similarities in the amino acid sequence of two proteins”. In: *Journal of Molecular Biology* 48.3 (Mar. 28, 1970), pp. 443–453. DOI: [10.1016/0022-2836\(70\)90057-4](https://doi.org/10.1016/0022-2836(70)90057-4). URL: [http://www.sciencedirect.com/science/article/pii/0022283670900574](https://www.sciencedirect.com/science/article/pii/0022283670900574) (cit. on p. 42).
- [139] T. F. Smith and M. S. Waterman. “Identification of common molecular subsequences”. In: *Journal of Molecular Biology* 147.1 (Mar. 25, 1981), pp. 195–197. DOI: [10.1016/0022-2836\(81\)90087-5](https://doi.org/10.1016/0022-2836(81)90087-5). URL: <https://www.sciencedirect.com/science/article/pii/0022283681900875> (cit. on p. 42).
- [140] Stephen P. Bradley, Arnoldo C. Hax, and Thomas L. Magnanti. *Applied Mathematical Programming*. Google-Books-ID: MSWdWv3Gn5cC. Addison-Wesley Publishing Company, 1977 (cit. on p. 42).

## REFERENCES FOR CHAPTER 2

- [141] Richard Bellman. “The theory of dynamic programming”. In: *Bulletin of the American Mathematical Society* 60.6 (1954), pp. 503–515. DOI: [10.1090/S0002-9904-1954-09848-8](https://doi.org/10.1090/S0002-9904-1954-09848-8). URL: <https://www.ams.org/bull/1954-60-06/S0002-9904-1954-09848-8/> (cit. on p. 42).
- [142] William J. Masek and Michael S. Paterson. “A faster algorithm computing string edit distances”. In: *Journal of Computer and System Sciences* 20.1 (Feb. 1, 1980), pp. 18–31. DOI: [10.1016/0022-0000\(80\)90002-1](https://doi.org/10.1016/0022-0000(80)90002-1). URL: <https://www.sciencedirect.com/science/article/pii/0022000080900021> (cit. on p. 42).
- [143] Nguyen Xuan Vinh, Julien Epps, and James Bailey. “Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance”. en. In: *Journal of Machine Learning Research* 11 (2010), p. 18 (cit. on pp. 42, 134, 171).
- [144] J. D. Ullman, A. V. Aho, and D. S. Hirschberg. “Bounds on the Complexity of the Longest Common Subsequence Problem”. In: *Journal of the ACM* 23.1 (Jan. 1, 1976), 1–12. DOI: [10.1145/321921.321922](https://doi.org/10.1145/321921.321922). URL: <https://doi.org/10.1145/321921.321922> (cit. on p. 42).
- [145] D. S. Hirschberg. “A linear space algorithm for computing maximal common subsequences”. In: *Communications of the ACM* 18.6 (June 1, 1975), 341–343. DOI: [10.1145/360825.360861](https://doi.org/10.1145/360825.360861). URL: <https://doi.org/10.1145/360825.360861> (cit. on p. 43).
- [146] Eugene W. Myers and Webb Miller. “Optimal alignments in linear space”. In: *Bioinformatics* 4.1 (Mar. 1, 1988), pp. 11–17. DOI: [10.1093/bioinformatics/4.1.11](https://doi.org/10.1093/bioinformatics/4.1.11). URL: <https://doi.org/10.1093/bioinformatics/4.1.11> (cit. on p. 43).
- [147] Peter Rice, Ian Longden, and Alan Bleasby. “EMBOSS: the European molecular biology open software suite”. In: *Trends in genetics* 16.6 (2000). tex.ids=riceEMBOSSEuropeanMolecular publisher: Elsevier current trends, 276–277 (cit. on pp. 43, 44).
- [148] Xiaoqiu Huang and Webb Miller. “A time-efficient, linear-space local similarity algorithm”. In: *Advances in Applied Mathematics* 12.3 (Sept. 1, 1991), pp. 337–357. DOI: [10.1016/0196-8858\(91\)90017-D](https://doi.org/10.1016/0196-8858(91)90017-D). URL: <https://www.sciencedirect.com/science/article/pii/019688589190017D> (cit. on p. 44).
- [149] Michael S. Waterman and Mark Eggert. “A new algorithm for best subsequence alignments with application to tRNA-rRNA comparisons”. In: *Journal of Molecular Biology* 197.4 (Oct. 20, 1987), pp. 723–728. DOI: [10.1016/0022-2836\(87\)90478-5](https://doi.org/10.1016/0022-2836(87)90478-5). URL: <https://www.sciencedirect.com/science/article/pii/0022283687904785> (cit. on p. 44).
- [150] Jason E. Stajich, David Block, et al. “The Bioperl Toolkit: Perl Modules for the Life Sciences”. In: *Genome Research* 12.10 (Jan. 10, 2002). Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab PMID: 12368254, pp. 1611–1618.

## CHAPTER 2

- DOI: [10.1101/gr.361602](https://doi.org/10.1101/gr.361602). URL: <https://genome.cshlp.org/content/12/10/1611> (cit. on p. 44).
- [151] Robert C. Gentleman, Vincent J. Carey, et al. “Bioconductor: open software development for computational biology and bioinformatics”. In: *Genome Biology* 5.10 (Sept. 15, 2004), R80. DOI: [10.1186/gb-2004-5-10-r80](https://doi.org/10.1186/gb-2004-5-10-r80). URL: <https://doi.org/10.1186/gb-2004-5-10-r80> (cit. on p. 44).
  - [152] Jeff Daily. “Parasail: SIMD C library for global, semi-global, and local pairwise sequence alignments”. In: *BMC Bioinformatics* 17.1 (Feb. 10, 2016), p. 81. DOI: [10.1186/s12859-016-0930-z](https://doi.org/10.1186/s12859-016-0930-z). URL: <https://doi.org/10.1186/s12859-016-0930-z> (cit. on p. 44).
  - [153] W. Frohmberg, M. Kierzynka, et al. “G-PAS 2.0 – an improved version of protein alignment tool with an efficient backtracking routine on multiple GPUs”. In: *Bulletin of the Polish Academy of Sciences: Technical Sciences* 60.3 (Dec. 1, 2012), pp. 491–494. DOI: [10.2478/v10175-012-0062-1](https://doi.org/10.2478/v10175-012-0062-1). URL: <http://journals.pan.pl/dlibra/publication/96876.edition/83624/content> (cit. on p. 44).
  - [154] Stephen F Altschul. “Substitution Matrices”. In: John Wiley & Sons, Ltd, 2013. DOI: [10.1002/9780470015902.a0005265.pub3](https://doi.org/10.1002/9780470015902.a0005265.pub3). URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470015902.a0005265.pub3> (cit. on p. 44).
  - [155] M.O. Dayhoff, R.M. Schwartz, and B.C. Orcutt. “A Model of Evolutionary Change in Proteins”. In: *A Model of Evolutionary Change in Proteins* (1978), pp. 345–352 (cit. on p. 44).
  - [156] T. Müller and M. Vingron. “Modeling amino acid replacement”. In: *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology* 7.6 (2000). PMID: 11382360, pp. 761–776. DOI: [10.1089/10665270050514918](https://doi.org/10.1089/10665270050514918) (cit. on p. 44).
  - [157] S. Henikoff and J. G. Henikoff. “Amino acid substitution matrices from protein blocks”. In: *Proceedings of the National Academy of Sciences* 89.22 (Nov. 15, 1992). PMID: 1438297, pp. 10915–10919. DOI: [10.1073/pnas.89.22.10915](https://doi.org/10.1073/pnas.89.22.10915). URL: <https://www.pnas.org/content/89/22/10915> (cit. on p. 44).
  - [158] S. Whelan and N. Goldman. “A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach”. In: *Molecular Biology and Evolution* 18.5 (May 2001). PMID: 11319253, pp. 691–699. DOI: [10.1093/oxfordjournals.molbev.a003851](https://doi.org/10.1093/oxfordjournals.molbev.a003851) (cit. on p. 44).
  - [159] Si Quang Le and Olivier Gascuel. “An Improved General Amino Acid Replacement Matrix”. In: *Molecular Biology and Evolution* 25.7 (July 1, 2008), pp. 1307–1320. DOI: [10.1093/molbev/msn067](https://doi.org/10.1093/molbev/msn067). URL: <https://doi.org/10.1093/molbev/msn067> (cit. on p. 44).

## REFERENCES FOR CHAPTER 2

- [160] Tobias Müller, Sven Rahmann, and Marc Rehmsmeier. “Non-symmetric score matrices and the detection of homologous transmembrane proteins”. In: *Bioinformatics* 17.suppl\_1 (June 1, 2001), S182–S189. doi: [10.1093/bioinformatics/17.suppl\\_1.S182](https://doi.org/10.1093/bioinformatics/17.suppl_1.S182). URL: [https://doi.org/10.1093/bioinformatics/17.suppl\\_1.S182](https://doi.org/10.1093/bioinformatics/17.suppl_1.S182) (cit. on p. 44).
- [161] P. C. Ng, J. G. Henikoff, and S. Henikoff. “PHAT: a transmembrane-specific substitution matrix. Predicted hydrophobic and transmembrane”. In: *Bioinformatics (Oxford, England)* 16.9 (Sept. 2000). PMID: 11108698, pp. 760–766. doi: [10.1093/bioinformatics/16.9.760](https://doi.org/10.1093/bioinformatics/16.9.760) (cit. on p. 44).
- [162] Rakesh Trivedi and Hampapathalu Adimurthy Nagarajaram. “Amino acid substitution scoring matrices specific to intrinsically disordered regions in proteins”. In: *Scientific Reports* 9.1 (Nov. 8, 2019). Number: 1 Publisher: Nature Publishing Group, p. 16380. doi: [10.1038/s41598-019-52532-8](https://doi.org/10.1038/s41598-019-52532-8). URL: <https://www.nature.com/articles/s41598-019-52532-8> (cit. on p. 44).
- [163] Nalin C. W. Goonesekere and Byungkook Lee. “Context-specific amino acid substitution matrices and their use in the detection of protein homologs”. In: *Proteins: Structure, Function, and Bioinformatics* 71.2 (2008), pp. 910–919. doi: [10.1002/prot.21775](https://doi.org/10.1002/prot.21775). URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.21775> (cit. on p. 44).
- [164] Umadevi Paila, Rohini Kondam, and Akash Ranjan. “Genome bias influences amino acid choices: analysis of amino acid substitution and re-compilation of substitution matrices exclusive to an AT-biased genome”. In: *Nucleic Acids Research* 36.21 (Dec. 2008). PMID: 18948281 PMCID: PMC2588515, pp. 6664–6675. doi: [10.1093/nar/gkn635](https://doi.org/10.1093/nar/gkn635) (cit. on p. 44).
- [165] David C. Nickle, Laura Heath, et al. “HIV-Specific Probabilistic Models of Protein Evolution”. In: *PLoS ONE* 2.6 (June 6, 2007). PMID: 17551583 PMCID: PMC1876811, e503. doi: [10.1371/journal.pone.0000503](https://doi.org/10.1371/journal.pone.0000503). URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1876811/> (cit. on p. 44).
- [166] Mihaela E. Sardiu, Gelio Alves, and Yi-Kuo Yu. “Score statistics of global sequence alignment from the energy distribution of a modified directed polymer and directed percolation problem”. In: *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics* 72.6 Pt 1 (Dec. 2005). PMID: 16485984, p. 061917. doi: [10.1103/PhysRevE.72.061917](https://doi.org/10.1103/PhysRevE.72.061917) (cit. on p. 44).
- [167] F. Chiaromonte, V. B. Yap, and W. Miller. “Scoring pairwise genomic sequence alignments”. In: DOI: 10.1142/9789812799623\_0012. WORLD SCIENTIFIC, Dec. 2001, pp. 115–126. doi: [10.1142/9789812799623\\_0012](https://doi.org/10.1142/9789812799623_0012). URL: [https://www.worldscientific.com/doi/abs/10.1142/9789812799623\\_0012](https://www.worldscientific.com/doi/abs/10.1142/9789812799623_0012) (cit. on p. 44).
- [168] Adrian Schneider, Gina M. Cannarozzi, and Gaston H. Gonnet. “Empirical codon substitution matrix”. In: *BMC bioinformatics* 6 (June 1, 2005). PMID: 15927081 PMCID: PMC1173088, p. 134. doi: [10.1186/1471-2105-6-134](https://doi.org/10.1186/1471-2105-6-134) (cit. on p. 44).

## CHAPTER 2

- [169] Adi Doron-Faigenboim and Tal Pupko. “A Combined Empirical and Mechanistic Codon Model”. In: *Molecular Biology and Evolution* 24.2 (Feb. 1, 2007), pp. 388–397. DOI: [10.1093/molbev/msl175](https://doi.org/10.1093/molbev/msl175). URL: <https://doi.org/10.1093/molbev/msl175> (cit. on p. 44).
- [170] Osamu Gotoh. “An improved algorithm for matching biological sequences”. In: *Journal of Molecular Biology* 162.3 (Dec. 15, 1982), pp. 705–708. DOI: [10.1016/0022-2836\(82\)90398-9](https://doi.org/10.1016/0022-2836(82)90398-9). URL: <https://www.sciencedirect.com/science/article/pii/0022283682903989> (cit. on p. 45).
- [171] Steven A. Benner, Mark A. Cohen, and Gaston H. Gonnet. “Empirical and Structural Models for Insertions and Deletions in the Divergent Evolution of Proteins”. In: *Journal of Molecular Biology* 229.4 (Feb. 20, 1993), pp. 1065–1082. DOI: [10.1006/jmbi.1993.1105](https://doi.org/10.1006/jmbi.1993.1105). URL: <https://www.sciencedirect.com/science/article/pii/S0022283683711058> (cit. on p. 45).
- [172] Xun Gu and Wen-Hsiung Li. “The size distribution of insertions and deletions in human and rodent pseudogenes suggests the logarithmic gap penalty for sequence alignment”. In: *Journal of Molecular Evolution* 40.4 (Apr. 1, 1995), pp. 464–473. DOI: [10.1007/BF00164032](https://doi.org/10.1007/BF00164032). URL: <https://doi.org/10.1007/BF00164032> (cit. on p. 45).
- [173] Michael S. Waterman. “Efficient sequence alignment algorithms”. In: *Journal of Theoretical Biology* 108.3 (June 7, 1984), pp. 333–337. DOI: [10.1016/S0022-5193\(84\)80037-5](https://doi.org/10.1016/S0022-5193(84)80037-5). URL: <https://www.sciencedirect.com/science/article/pii/S0022519384800375> (cit. on p. 45).
- [174] William R. Pearson and Webb Miller. “[27] Dynamic programming algorithms for biological sequence comparison”. In: vol. 210. Numerical Computer Methods. DOI: [10.1016/0076-6879\(92\)10029-D](https://doi.org/10.1016/0076-6879(92)10029-D). Academic Press, Jan. 1, 1992, pp. 575–601. DOI: [10.1016/0076-6879\(92\)10029-D](https://doi.org/10.1016/0076-6879(92)10029-D). URL: <https://www.sciencedirect.com/science/article/pii/007668799210029D> (cit. on p. 45).
- [175] Jiannan Chao, Furong Tang, and Lei Xu. “Developments in Algorithms for Sequence Alignment: A Review”. In: *Biomolecules* 12.4 (Apr. 6, 2022), p. 546. DOI: [10.3390/biom12040546](https://doi.org/10.3390/biom12040546). URL: <https://www.mdpi.com/2218-273X/12/4/546> (cit. on pp. 45, 46).
- [176] John L. Spouge. “Speeding up Dynamic Programming Algorithms for Finding Optimal Lattice Paths”. In: *SIAM Journal on Applied Mathematics* 49.5 (Oct. 1989). tex.ids= spougeSpeedingDynamicProgramming1989 publisher: Society for Industrial and Applied Mathematics, pp. 1552–1566. DOI: [10.1137/0149094](https://doi.org/10.1137/0149094). URL: <https://pubs.siam.org/doi/abs/10.1137/0149094> (cit. on p. 45).
- [177] James W. Fickett. “Fast optimal alignment”. In: *Nucleic Acids Research* 12.1Part1 (Jan. 11, 1984), pp. 175–179. DOI: [10.1093/nar/12.1Part1.175](https://doi.org/10.1093/nar/12.1Part1.175). URL: <https://doi.org/10.1093/nar/12.1Part1.175> (cit. on p. 45).

## REFERENCES FOR CHAPTER 2

- [178] Santiago Marco-Sola, Juan Carlos Moure, et al. “Fast gap-affine pairwise alignment using the wavefront algorithm”. In: *Bioinformatics* btaa777 (Sept. 11, 2020). DOI: [10.1093/bioinformatics/btaa777](https://doi.org/10.1093/bioinformatics/btaa777). URL: <https://doi.org/10.1093/bioinformatics/btaa777> (cit. on p. 45).
- [179] Richard Durbin, Sean R. Eddy, et al. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. DOI: 10.1017/CBO9780511790492. Cambridge: Cambridge University Press, 1998. DOI: [10.1017/CBO9780511790492](https://doi.org/10.1017/CBO9780511790492). URL: <https://www.cambridge.org/core/books/biological-sequence-analysis/921BB7B78B745198829EF96BC7E0F29D> (cit. on p. 45).
- [180] Johannes Söding. “Protein homology detection by HMM-HMM comparison”. In: *Bioinformatics (Oxford, England)* 21.7 (Apr. 1, 2005). PMID: 15531603, pp. 951–960. DOI: [10.1093/bti125](https://doi.org/10.1093/bti125) (cit. on p. 45).
- [181] Robert D. Finn, Jody Clements, and Sean R. Eddy. “HMMER web server: interactive sequence similarity searching”. In: *Nucleic Acids Research* 39. Web Server issue (July 1, 2011). PMID: 21593126 PMCID: PMC3125773, W29–W37. DOI: [10.1093/nar/gkr367](https://doi.org/10.1093/nar/gkr367). URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3125773/> (cit. on pp. 45, 50).
- [182] Jun Wang, Peter D. Keightley, and Toby Johnson. “MCALIGN2: Faster, accurate global pairwise alignment of non-coding DNA sequences based on explicit models of indel evolution”. In: *BMC Bioinformatics* 7.1 (June 8, 2006), p. 292. DOI: [10.1186/1471-2105-7-292](https://doi.org/10.1186/1471-2105-7-292). URL: <https://doi.org/10.1186/1471-2105-7-292> (cit. on p. 45).
- [183] Kazutaka Katoh, Kazuharu Misawa, et al. “MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform”. In: *Nucleic Acids Research* 30.14 (July 15, 2002), pp. 3059–3066. DOI: [10.1093/nar/gkf436](https://doi.org/10.1093/nar/gkf436). URL: <https://doi.org/10.1093/nar/gkf436> (cit. on pp. 45, 50).
- [184] Yanni Sun and Jeremy Buhler. “Choosing the best heuristic for seeded alignment of DNA sequences”. In: *BMC Bioinformatics* 7.1 (Mar. 13, 2006), p. 133. DOI: [10.1186/1471-2105-7-133](https://doi.org/10.1186/1471-2105-7-133). URL: <https://doi.org/10.1186/1471-2105-7-133> (cit. on p. 46).
- [185] Heng Li and Nils Homer. “A Survey of Sequence Alignment Algorithms for Next-Generation Sequencing”. In: *Briefings in Bioinformatics* 11.5 (Sept. 1, 2010), pp. 473–483. ISSN: 1467-5463. DOI: [10.1093/bib/bbp015](https://doi.org/10.1093/bib/bbp015). URL: <https://doi.org/10.1093/bib/bbp015> (visited on 05/16/2022) (cit. on p. 46).
- [186] S. F. Altschul, W. Gish, et al. “Basic local alignment search tool”. In: *Journal of Molecular Biology* 215.3 (Oct. 5, 1990). PMID: 2231712, pp. 403–410. DOI: [10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2) (cit. on p. 47).
- [187] Robert C. Edgar. “Search and clustering orders of magnitude faster than BLAST”. In: *Bioinformatics* 26.19 (Oct. 1, 2010), pp. 2460–2461. DOI: [10.1093/bioinformatics/btq461](https://doi.org/10.1093/bioinformatics/btq461). URL: <https://doi.org/10.1093/bioinformatics/btq461> (cit. on p. 47).

## CHAPTER 2

- [188] Eugene G. Shpaer, Max Robinson, et al. “Sensitivity and Selectivity in Protein Similarity Searches: A Comparison of Smith–Waterman in Hardware to BLAST and FASTA”. In: *Genomics* 38.2 (Dec. 1, 1996). tex.ids= shpaerSensitivitySelectivityProtein1996a, pp. 179–191. doi: [10.1006/geno.1996.0614](https://doi.org/10.1006/geno.1996.0614). URL: <https://www.sciencedirect.com/science/article/pii/S088875439690614X> (cit. on p. 47).
- [189] W. R. Pearson and D. J. Lipman. “Improved tools for biological sequence comparison”. In: *Proceedings of the National Academy of Sciences of the United States of America* 85.8 (Apr. 1988). PMID: 3162770 PMCID: PMC280013, pp. 2444–2448. doi: [10.1073/pnas.85.8.2444](https://doi.org/10.1073/pnas.85.8.2444) (cit. on p. 47).
- [190] D. J. Lipman and W. R. Pearson. “Rapid and sensitive protein similarity searches”. In: *Science (New York, N.Y.)* 227.4693 (Mar. 22, 1985). PMID: 2983426, pp. 1435–1441. doi: [10.1126/science.2983426](https://doi.org/10.1126/science.2983426) (cit. on p. 47).
- [191] Ganapathi Varma Saripella, Erik L. L. Sonnhammer, and Kristoffer Forslund. “Benchmarking the next generation of homology inference tools”. In: *Bioinformatics* 32.17 (Sept. 9, 2016). Publisher: Oxford University Press PMID: 27256311, p. 2636. doi: [10.1093/bioinformatics/btw305](https://doi.org/10.1093/bioinformatics/btw305). URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5013910/> (cit. on p. 48).
- [192] Robert D. Finn, Penelope Coggill, et al. “The Pfam protein families database: towards a more sustainable future”. In: *Nucleic Acids Research* 44.Database issue (Jan. 1, 2016). Publisher: Oxford University Press PMID: 26673716, p. D279. doi: [10.1093/nar/gkv1344](https://doi.org/10.1093/nar/gkv1344). URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4702930/> (cit. on p. 48).
- [193] Nadia Essoussi and Sondes Fayech. “A comparison of four pair-wise sequence alignment methods”. In: *Bioinformation* 2.4 (Dec. 28, 2007). PMID: 21670797 PMCID: PMC2255065, pp. 166–168. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2255065/> (cit. on p. 48).
- [194] S Karlin and S F Altschul. “Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes.” In: *Proceedings of the National Academy of Sciences* 87.6 (Mar. 1990). tex.ids= karlinMethodsAssessingStatistical1990 PMCID: PMC53667 PMID: 2315319, pp. 2264–2268. doi: [10.1073/pnas.87.6.2264](https://doi.org/10.1073/pnas.87.6.2264). URL: <https://pnas.org/doi/full/10.1073/pnas.87.6.2264> (cit. on p. 48).
- [195] Richard Mott. “Alignment: Statistical Significance”. In: \_eprint: <https://onlinelibrary.wiley.com/doi/10.1038/npg.els.0005264>. John Wiley & Sons, Ltd, 2005. doi: [10.1038/npg.els.0005264](https://doi.org/10.1038/npg.els.0005264). URL: <https://onlinelibrary.wiley.com/doi/abs/10.1038/npg.els.0005264> (cit. on p. 48).
- [196] Benjamin Buchfink, Chao Xie, and Daniel H. Huson. “Fast and sensitive protein alignment using DIAMOND”. In: *Nature Methods* 12.1 (Jan. 2015). PMID: 25402007, pp. 59–60. doi: [10.1038/nmeth.3176](https://doi.org/10.1038/nmeth.3176) (cit. on p. 48).

- [197] Benjamin Buchfink, Klaus Reuter, and Hajk-Georg Drost. “Sensitive protein alignments at tree-of-life scale using DIAMOND”. In: *Nature Methods* 18.4 (Apr. 2021). Number: 4 Publisher: Nature Publishing Group, pp. 366–368. doi: [10.1038/s41592-021-01101-x](https://doi.org/10.1038/s41592-021-01101-x). URL: <https://www.nature.com/articles/s41592-021-01101-x> (cit. on p. 48).
- [198] Nick Bray, Inna Dubchak, and Lior Pachter. “AVID: A Global Alignment Program”. In: *Genome Research* 13.1 (Jan. 1, 2003). PMID: 12529311 PMCID: PMC430967, pp. 97–102. doi: [10.1101/gr.789803](https://doi.org/10.1101/gr.789803). URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC430967/> (cit. on p. 48).
- [199] Arthur L. Delcher, Adam Phillippy, et al. “Fast algorithms for large-scale genome alignment and comparison”. In: *Nucleic Acids Research* 30.11 (June 1, 2002). PMID: 12034836 PMCID: PMC117189, pp. 2478–2483. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC117189/> (cit. on p. 48).
- [200] P. Ferragina and G. Manzini. “Proceedings 41st Annual Symposium on Foundations of Computer Science”. In: tex.ids= ferraginaOpportunisticDataStructures2000a ISSN: 0272-5428. Nov. 2000, pp. 390–398. doi: [10.1109/SFCS.2000.892127](https://doi.org/10.1109/SFCS.2000.892127) (cit. on p. 48).
- [201] Michael Burrows and David Wheeler. *A Block-Sorting Lossless Data Compression Algorithm*. Tech. rep. 1994 (cit. on p. 48).
- [202] T. W. Lam, W. K. Sung, et al. “Compressed indexing and local alignment of DNA”. In: *Bioinformatics* 24.6 (Mar. 15, 2008), pp. 791–797. doi: [10.1093/bioinformatics/btn032](https://doi.org/10.1093/bioinformatics/btn032). URL: <https://doi.org/10.1093/bioinformatics/btn032> (cit. on pp. 49, 51).
- [203] Ben Langmead and Steven L. Salzberg. “Fast Gapped-Read Alignment with Bowtie 2”. In: *Nature Methods* 9.4 (4 Apr. 2012), pp. 357–359. ISSN: 1548-7105. doi: [10.1038/nmeth.1923](https://doi.org/10.1038/nmeth.1923) (cit. on pp. 49, 51).
- [204] Heng Li and Richard Durbin. “Fast and accurate short read alignment with Burrows–Wheeler transform”. In: *Bioinformatics* 25.14 (July 15, 2009), pp. 1754–1760. doi: [10.1093/bioinformatics/btp324](https://doi.org/10.1093/bioinformatics/btp324). URL: <https://doi.org/10.1093/bioinformatics/btp324> (cit. on pp. 49, 51).
- [205] Heng Li and Richard Durbin. “Fast and accurate long-read alignment with Burrows–Wheeler transform”. In: *Bioinformatics* 26.5 (Mar. 1, 2010), pp. 589–595. doi: [10.1093/bioinformatics/btp698](https://doi.org/10.1093/bioinformatics/btp698). URL: <https://doi.org/10.1093/bioinformatics/btp698> (cit. on pp. 49, 51).
- [206] Heng Li. *Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM*. May 26, 2013. arXiv: [1303.3997 \[q-bio\]](https://arxiv.org/abs/1303.3997) (cit. on pp. 49, 51).
- [207] Yongchao Liu and Bertil Schmidt. “Long read alignment based on maximal exact match seeds”. In: *Bioinformatics* 28.18 (Sept. 15, 2012), pp. i318–i324. doi: [10.1093/bioinformatics/bts414](https://doi.org/10.1093/bioinformatics/bts414). URL: <https://doi.org/10.1093/bioinformatics/bts414> (cit. on pp. 49, 51).

## CHAPTER 2

- [208] David J Russell, ed. *Multiple Sequence Alignment Methods*. Vol. 1079. Methods in Molecular Biology. DOI: 10.1007/978-1-62703-646-7. Totowa, NJ: Humana Press, 2014. DOI: [10.1007/978-1-62703-646-7](https://doi.org/10.1007/978-1-62703-646-7). URL: <http://link.springer.com/10.1007/978-1-62703-646-7> (cit. on p. 49).
- [209] Lusheng Wang and Tao Jiang. “On the Complexity of Multiple Sequence Alignment”. In: *Journal of Computational Biology* 1.4 (Jan. 1994). Publisher: Mary Ann Liebert, Inc., publishers, pp. 337–348. DOI: [10.1089/cmb.1994.1.337](https://doi.org/10.1089/cmb.1994.1.337). URL: <https://www.liebertpub.com/doi/abs/10.1089/cmb.1994.1.337> (cit. on p. 49).
- [210] Winfried Just. “Computational Complexity of Multiple Sequence Alignment with SP-Score”. In: *Journal of Computational Biology* 8.6 (Nov. 2001). Publisher: Mary Ann Liebert, Inc., publishers, pp. 615–623. DOI: [10.1089/106652701753307511](https://doi.org/10.1089/106652701753307511). URL: <https://www.liebertpub.com/doi/abs/10.1089/106652701753307511> (cit. on p. 49).
- [211] Da-Fei Feng and Russell F. Doolittle. “Progressive sequence alignment as a pre-requisite to correct phylogenetic trees”. In: *Journal of Molecular Evolution* 25.4 (Aug. 1, 1987), pp. 351–360. DOI: [10.1007/BF02603120](https://doi.org/10.1007/BF02603120). URL: <https://doi.org/10.1007/BF02603120> (cit. on p. 49).
- [212] David T. Jones, William R. Taylor, and Janet M. Thornton. “The rapid generation of mutation data matrices from protein sequences”. In: *Bioinformatics* 8.3 (June 1, 1992), pp. 275–282. DOI: [10.1093/bioinformatics/8.3.275](https://doi.org/10.1093/bioinformatics/8.3.275). URL: <https://doi.org/10.1093/bioinformatics/8.3.275> (cit. on p. 49).
- [213] B E Blaisdell. “A measure of the similarity of sets of sequences not requiring sequence alignment.” In: *Proceedings of the National Academy of Sciences* 83.14 (July 1986). Publisher: Proceedings of the National Academy of Sciences, pp. 5155–5159. DOI: [10.1073/pnas.83.14.5155](https://doi.org/10.1073/pnas.83.14.5155). URL: <https://www.pnas.org/doi/abs/10.1073/pnas.83.14.5155> (cit. on p. 49).
- [214] Stephen F. Altschul, Raymond J. Carroll, and David J. Lipman. “Weights for data related by a tree”. In: *Journal of Molecular Biology* 207.4 (June 20, 1989), pp. 647–653. DOI: [10.1016/0022-2836\(89\)90234-9](https://doi.org/10.1016/0022-2836(89)90234-9). URL: <https://www.sciencedirect.com/science/article/pii/0022283689902349> (cit. on p. 49).
- [215] Robert C. Edgar and Kimmen Sjölander. “A comparison of scoring functions for protein sequence profile alignment”. In: *Bioinformatics* 20.8 (May 22, 2004), pp. 1301–1308. DOI: [10.1093/bioinformatics/bth090](https://doi.org/10.1093/bioinformatics/bth090). URL: <https://doi.org/10.1093/bioinformatics/bth090> (cit. on p. 49).
- [216] C Notredame, L Holm, and D G Higgins. “COFFEE: an objective function for multiple sequence alignments.” In: *Bioinformatics* 14.5 (June 1, 1998), pp. 407–422. DOI: [10.1093/bioinformatics/14.5.407](https://doi.org/10.1093/bioinformatics/14.5.407). URL: <https://doi.org/10.1093/bioinformatics/14.5.407> (cit. on p. 49).

- [217] Julie D. Thompson, Desmond G. Higgins, and Toby J. Gibson. “CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice”. In: *Nucleic Acids Research* 22.22 (Nov. 11, 1994), pp. 4673–4680. DOI: [10.1093/nar/22.22.4673](https://doi.org/10.1093/nar/22.22.4673). URL: <https://academic.oup.com/nar/article/22/22/4673/2400290> (cit. on p. 49).
- [218] Julie D. Thompson, Toby J. Gibson, et al. “The CLUSTAL\_X Windows Interface: Flexible Strategies for Multiple Sequence Alignment Aided by Quality Analysis Tools”. In: *Nucleic Acids Research* 25.24 (Dec. 1, 1997), pp. 4876–4882. DOI: [10.1093/nar/25.24.4876](https://doi.org/10.1093/nar/25.24.4876). URL: <https://doi.org/10.1093/nar/25.24.4876> (cit. on p. 49).
- [219] Cédric Notredame, Desmond G Higgins, and Jaap Heringa. “T-coffee: a novel method for fast and accurate multiple sequence alignment” Edited by J. Thornton”. In: *Journal of Molecular Biology* 302.1 (Sept. 8, 2000), pp. 205–217. DOI: [10.1006/jmbi.2000.4042](https://doi.org/10.1006/jmbi.2000.4042). URL: <http://www.sciencedirect.com/science/article/pii/S0022283600940427> (cit. on p. 50).
- [220] Robert C. Edgar. “MUSCLE: a multiple sequence alignment method with reduced time and space complexity”. In: *BMC Bioinformatics* 5.1 (Aug. 19, 2004), p. 113. DOI: [10.1186/1471-2105-5-113](https://doi.org/10.1186/1471-2105-5-113). URL: <https://doi.org/10.1186/1471-2105-5-113> (cit. on p. 50).
- [221] Robert C. Edgar. “MUSCLE: multiple sequence alignment with high accuracy and high throughput”. In: *Nucleic Acids Research* 32.5 (Mar. 1, 2004), pp. 1792–1797. DOI: [10.1093/nar/gkh340](https://doi.org/10.1093/nar/gkh340). URL: <https://doi.org/10.1093/nar/gkh340> (cit. on p. 50).
- [222] Chuong B. Do, Mahathi S.P. Mahabhashyam, et al. “ProbCons: Probabilistic consistency-based multiple sequence alignment”. In: *Genome Research* 15.2 (Feb. 2005). PMID: 15687296 PMCID: PMC546535, pp. 330–340. DOI: [10.1101/gr.2821705](https://doi.org/10.1101/gr.2821705). URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC546535/> (cit. on p. 50).
- [223] Cédric Notredame. “Recent Evolutions of Multiple Sequence Alignment Algorithms”. In: *PLOS Computational Biology* 3.8 (Aug. 31, 2007). Publisher: Public Library of Science, e123. DOI: [10.1371/journal.pcbi.0030123](https://doi.org/10.1371/journal.pcbi.0030123). URL: <https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.0030123> (cit. on p. 50).
- [224] Cédric Notredame. “Recent Progress in Multiple Sequence Alignment: A Survey”. In: *Pharmacogenomics* 3.1 (Jan. 2002), pp. 131–144. ISSN: 1462-2416. DOI: [10.1517/14622416.3.1.131](https://doi.org/10.1517/14622416.3.1.131). URL: <https://www.futuremedicine.com/doi/abs/10.1517/14622416.3.1.131> (visited on 05/16/2022) (cit. on p. 50).

## CHAPTER 2

- [225] Robert C Edgar and Serafim Batzoglou. “Multiple Sequence Alignment”. In: *Current Opinion in Structural Biology*. Nucleic Acids/Sequences and Topology 16.3 (June 1, 2006), pp. 368–373. ISSN: 0959-440X. DOI: [10.1016/j.sbi.2006.04.004](https://doi.org/10.1016/j.sbi.2006.04.004). URL: <https://www.sciencedirect.com/science/article/pii/S0959440X06000704> (visited on 05/16/2022) (cit. on p. 50).
- [226] Fabiano Sviatopolk-Mirsky Pais, Patrícia de Cássia Ruy, et al. “Assessing the efficiency of multiple sequence alignment programs”. In: *Algorithms for Molecular Biology* 9.1 (Mar. 6, 2014), p. 4. DOI: [10.1186/1748-7188-9-4](https://doi.org/10.1186/1748-7188-9-4). URL: <https://doi.org/10.1186/1748-7188-9-4> (cit. on p. 50).
- [227] J. D. Thompson, F. Plewniak, and O. Poch. “BAliBASE: a benchmark alignment database for the evaluation of multiple alignment programs.” In: *Bioinformatics* 15.1 (Jan. 1, 1999), pp. 87–88. DOI: [10.1093/bioinformatics/15.1.87](https://doi.org/10.1093/bioinformatics/15.1.87). URL: <https://academic.oup.com/bioinformatics/article/15/1/87/218377> (cit. on p. 50).
- [228] Sean R Eddy. “Multiple Alignment Using Hidden Markov Models”. In: (), p. 7 (cit. on p. 50).
- [229] Yongchao Liu, Bertil Schmidt, and Douglas L. Maskell. “MSAProbs: multiple sequence alignment based on pair hidden Markov models and partition function posterior probabilities”. In: *Bioinformatics* 26.16 (Aug. 15, 2010), pp. 1958–1964. DOI: [10.1093/bioinformatics/btq338](https://doi.org/10.1093/bioinformatics/btq338). URL: <https://doi.org/10.1093/bioinformatics/btq338> (cit. on p. 50).
- [230] Frédéric Lemoine, Luc Bassel, et al. “COVID-Align: Accurate online alignment of hCoV-19 genomes using a profile HMM”. In: *Bioinformatics* btaa871 (Oct. 12, 2020). tex.ids= lemoineCOVIDAlignAccurateOnline2020. DOI: [10.1093/bioinformatics/btaa871](https://doi.org/10.1093/bioinformatics/btaa871). URL: <https://doi.org/10.1093/bioinformatics/btaa871> (cit. on p. 50).
- [231] Ivan Aksamentov, Cornelius Roemer, et al. “Nextclade: clade assignment, mutation calling and quality control for viral genomes”. In: *Journal of Open Source Software* 6.67 (Nov. 30, 2021), p. 3773. DOI: [10.21105/joss.03773](https://doi.org/10.21105/joss.03773). URL: <https://joss.theoj.org/papers/10.21105/joss.03773> (cit. on p. 50).
- [232] Jin Kim, Sakti Pramanik, and Moon Jung Chung. “Multiple sequence alignment using simulated annealing”. In: *Bioinformatics* 10.4 (July 1, 1994), pp. 419–426. DOI: [10.1093/bioinformatics/10.4.419](https://doi.org/10.1093/bioinformatics/10.4.419). URL: <https://doi.org/10.1093/bioinformatics/10.4.419> (cit. on p. 51).
- [233] Masato Ishikawa, Tomoyuki Toya, et al. “Multiple sequence alignment by parallel simulated annealing”. In: *Bioinformatics* 9.3 (June 1, 1993), pp. 267–273. DOI: [10.1093/bioinformatics/9.3.267](https://doi.org/10.1093/bioinformatics/9.3.267). URL: <https://doi.org/10.1093/bioinformatics/9.3.267> (cit. on p. 51).
- [234] Hongwei Huo and Vojislav Stojkovic. “Third International Conference on Natural Computation (ICNC 2007)”. In: vol. 2. ISSN: 2157-9563. Aug. 2007, pp. 270–274. DOI: [10.1109/ICNC.2007.139](https://doi.org/10.1109/ICNC.2007.139) (cit. on p. 51).

- [235] Biswanath Chowdhury and Gautam Garai. “A review on multiple sequence alignment from the perspective of genetic algorithm”. In: *Genomics* 109.5 (Oct. 1, 2017), pp. 419–431. DOI: [10.1016/j.ygeno.2017.06.007](https://doi.org/10.1016/j.ygeno.2017.06.007). URL: <http://www.sciencedirect.com/science/article/pii/S0888754317300551> (cit. on p. 51).
- [236] Ching Zhang and Andrew K.C. Wong. “A genetic algorithm for multiple molecular sequence alignment”. In: *Bioinformatics* 13.6 (Dec. 1, 1997), pp. 565–581. DOI: [10.1093/bioinformatics/13.6.565](https://doi.org/10.1093/bioinformatics/13.6.565). URL: <https://doi.org/10.1093/bioinformatics/13.6.565> (cit. on p. 51).
- [237] C Notredame and D G Higgins. “SAGA: sequence alignment by genetic algorithm.” In: *Nucleic Acids Research* 24.8 (Apr. 15, 1996). PMID: 8628686 PMCID: PMC145823, pp. 1515–1524. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC145823/> (cit. on p. 51).
- [238] Farhana Naznin, Ruhul Sarker, and Daryl Essam. “Vertical decomposition with Genetic Algorithm for Multiple Sequence Alignment”. In: *BMC Bioinformatics* 12.1 (Aug. 25, 2011), p. 353. DOI: [10.1186/1471-2105-12-353](https://doi.org/10.1186/1471-2105-12-353). URL: <https://doi.org/10.1186/1471-2105-12-353> (cit. on p. 51).
- [239] Farhana Naznin, Ruhul Sarker, and Daryl Essam. “Progressive Alignment Method Using Genetic Algorithm for Multiple Sequence Alignment”. In: *IEEE Transactions on Evolutionary Computation* 16.5 (Oct. 2012). Conference Name: IEEE Transactions on Evolutionary Computation, pp. 615–631. DOI: [10.1109/TEVC.2011.2162849](https://doi.org/10.1109/TEVC.2011.2162849) (cit. on p. 51).
- [240] Edgar Garriga, Paolo Di Tommaso, et al. “Large multiple sequence alignments with a root-to-leaf regressive method”. In: *Nature Biotechnology* 37.12 (Dec. 2019). Number: 12 Publisher: Nature Publishing Group, pp. 1466–1470. DOI: [10.1038/s41587-019-0333-6](https://doi.org/10.1038/s41587-019-0333-6). URL: <https://www.nature.com/articles/s41587-019-0333-6> (cit. on p. 51).
- [241] Mohammed Alser, Jeremy Rotman, et al. “Technology dictates algorithms: recent developments in read alignment”. In: *Genome Biology* 22.1 (Aug. 26, 2021), p. 249. DOI: [10.1186/s13059-021-02443-7](https://doi.org/10.1186/s13059-021-02443-7). URL: <https://doi.org/10.1186/s13059-021-02443-7> (cit. on pp. 51, 52).
- [242] Sophie Schbath, Véronique Martin, et al. “Mapping Reads on a Genomic Sequence: An Algorithmic Overview and a Practical Comparative Analysis”. In: *Journal of Computational Biology* 19.6 (June 2012). PMID: 22506536 PMCID: PMC3375638, pp. 796–813. DOI: [10.1089/cmb.2012.0022](https://doi.org/10.1089/cmb.2012.0022). URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3375638/> (cit. on p. 51).
- [243] Ayat Hatem, Doruk Bozdağ, et al. “Benchmarking short sequence mapping tools”. In: *BMC Bioinformatics* 14.1 (June 7, 2013), p. 184. DOI: [10.1186/1471-2105-14-184](https://doi.org/10.1186/1471-2105-14-184). URL: <https://doi.org/10.1186/1471-2105-14-184> (cit. on p. 51).

## CHAPTER 2

- [244] Stefan Canzar and Steven L. Salzberg. “Short Read Mapping: An Algorithmic Tour”. In: *Proceedings of the IEEE* 105.3 (Mar. 2017). Conference Name: Proceedings of the IEEE, pp. 436–458. DOI: [10.1109/JPROC.2015.2455551](https://doi.org/10.1109/JPROC.2015.2455551) (cit. on p. 51).
- [245] Karel Břinda, Valentina Boeva, and Gregory Kucherov. “RNF: a general framework to evaluate NGS read mappers”. In: *Bioinformatics* 32.1 (Jan. 1, 2016), pp. 136–139. DOI: [10.1093/bioinformatics/btv524](https://doi.org/10.1093/bioinformatics/btv524). URL: <https://doi.org/10.1093/bioinformatics/btv524> (cit. on p. 51).
- [246] Matthew Ruffalo, Thomas LaFramboise, and Mehmet Koyutürk. “Comparative analysis of algorithms for next-generation sequencing read alignment”. In: *Bioinformatics (Oxford, England)* 27.20 (Oct. 15, 2011). PMID: 21856737, pp. 2790–2796. DOI: [10.1093/bioinformatics/btr477](https://doi.org/10.1093/bioinformatics/btr477) (cit. on p. 51).
- [247] Ben Langmead. “A tandem simulation framework for predicting mapping quality”. In: *Genome Biology* 18.1 (Aug. 10, 2017), p. 152. DOI: [10.1186/s13059-017-1290-3](https://doi.org/10.1186/s13059-017-1290-3). URL: <https://doi.org/10.1186/s13059-017-1290-3> (cit. on p. 51).
- [248] Hsin-Nan Lin and Wen-Lian Hsu. “Kart: a divide-and-conquer algorithm for NGS read alignment”. In: *Bioinformatics (Oxford, England)* 33.15 (Aug. 1, 2017). PMID: 28379292 PMCID: PMC5860120, pp. 2281–2287. DOI: [10.1093/bioinformatics/btx189](https://doi.org/10.1093/bioinformatics/btx189) (cit. on p. 51).
- [249] Corey B. Olson, Maria Kim, et al. “2012 IEEE 20th International Symposium on Field-Programmable Custom Computing Machines”. In: tex.ids= olsonHardwareAccelerationShort2012a. Apr. 2012, pp. 161–168. DOI: [10.1109/FCCM.2012.621036](https://doi.org/10.1109/FCCM.2012.621036) (cit. on p. 51).
- [250] Peng Chen, Chao Wang, et al. “Accelerating the Next Generation Long Read Mapping with the FPGA-Based System”. In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 11.5 (Sept. 2014). Conference Name: IEEE/ACM Transactions on Computational Biology and Bioinformatics, pp. 840–852. DOI: [10.1109/TCBB.2014.2326876](https://doi.org/10.1109/TCBB.2014.2326876) (cit. on p. 51).
- [251] Hajime Suzuki and Masahiro Kasahara. “Introducing difference recurrence relations for faster semi-global alignment of long sequences”. In: *BMC Bioinformatics* 19.1 (Feb. 19, 2018), p. 45. DOI: [10.1186/s12859-018-2014-8](https://doi.org/10.1186/s12859-018-2014-8). URL: <https://doi.org/10.1186/s12859-018-2014-8> (cit. on p. 51).
- [252] Alberto Zeni, Giulia Guidi, et al. “2020 IEEE International Parallel and Distributed Processing Symposium (IPDPS)”. In: ISSN: 1530-2075. May 2020, pp. 462–471. DOI: [10.1109/IPDPS47924.2020.00055](https://doi.org/10.1109/IPDPS47924.2020.00055) (cit. on p. 51).
- [253] Mark J. Chaisson and Glenn Tesler. “Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory”. In: *BMC Bioinformatics* 13.1 (Sept. 19, 2012), p. 238. DOI: [10.1186/1471-2105-13-238](https://doi.org/10.1186/1471-2105-13-238). URL: <https://doi.org/10.1186/1471-2105-13-238> (cit. on p. 52).

## REFERENCES FOR CHAPTER 2

- [254] Ehsan Haghshenas, S. Cenk Sahinalp, and Faraz Hach. “lordFAST: sensitive and Fast Alignment Search Tool for LOng noisy Read sequencing Data”. In: *Bioinformatics (Oxford, England)* 35.1 (Jan. 1, 2019). PMID: 30561550 PMCID: PMC6298053, pp. 20–27. doi: [10.1093/bioinformatics/bty544](https://doi.org/10.1093/bioinformatics/bty544) (cit. on p. 52).
- [255] Ivan Sović, Mile Šikić, et al. “Fast and sensitive mapping of nanopore sequencing reads with GraphMap”. In: *Nature Communications* 7.1 (Apr. 15, 2016). Number: 1 Publisher: Nature Publishing Group, p. 11307. doi: [10.1038/ncomms11307](https://doi.org/10.1038/ncomms11307). URL: <https://www.nature.com/articles/ncomms11307> (cit. on p. 52).
- [256] Fritz J. Sedlazeck, Philipp Rescheneder, et al. “Accurate detection of complex structural variations using single-molecule sequencing”. In: *Nature Methods* 15.6 (June 2018). Number: 6 Publisher: Nature Publishing Group, pp. 461–468. doi: [10.1038/s41592-018-0001-7](https://doi.org/10.1038/s41592-018-0001-7). URL: <https://www.nature.com/articles/s41592-018-0001-7> (cit. on p. 52).
- [257] Chirag Jain, Alexander Dilthey, et al. “A Fast Approximate Algorithm for Mapping Long Reads to Large Reference Databases”. In: *Journal of Computational Biology* 25.7 (July 1, 2018). PMID: 29708767 PMCID: PMC6067103, pp. 766–779. doi: [10.1089/cmb.2018.0036](https://doi.org/10.1089/cmb.2018.0036). URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6067103/> (cit. on p. 52).
- [258] Timofey Prodanov and Vikas Bansal. “Sensitive Alignment Using Paralogous Sequence Variants Improves Long-Read Mapping and Variant Calling in Segmental Duplications”. In: *Nucleic Acids Research* 48.19 (Nov. 4, 2020), e114. ISSN: 0305-1048. doi: [10.1093/nar/gkaa829](https://doi.org/10.1093/nar/gkaa829) (cit. on pp. 52, 85).
- [259] Chirag Jain, Arang Rhee, et al. “Long-read mapping to repetitive reference sequences using Winnowmap2”. In: *Nature Methods* 19.6 (June 2022). Number: 6 Publisher: Nature Publishing Group, pp. 705–710. doi: [10.1038/s41592-022-01457-8](https://doi.org/10.1038/s41592-022-01457-8). URL: <https://www.nature.com/articles/s41592-022-01457-8> (cit. on p. 52).
- [260] Alla Mikheenko, Andrey V Bzikadze, et al. “TandemTools: Mapping Long Reads and Assessing/Improving Assembly Quality in Extra-Long Tandem Repeats”. In: *Bioinformatics* 36 (Supplement\_1 July 1, 2020), pp. i75–i83. ISSN: 1367-4803. doi: [10.1093/bioinformatics/btaa440](https://doi.org/10.1093/bioinformatics/btaa440) (cit. on pp. 52, 78).



### 3. Contribution 1: Improving read alignment by exploring a sequence transformation space

*Intro to this chapter within the context of the thesis to go here!*

Recall that:

- Homopolymers are a problem (c.f. 1.4.2)
- Mapping is hard (c.f. 2.4)
- HPC has been used successfully used to improve mapping (c.f. 1.4.3.2)

This chapter was written as an article titled:

**“Mapping-friendly sequence reductions: going beyond homopolymer compression”**

it was published in **DATE HERE**, in the iScience proceedings of the RECOMB-SEQ 2022 conference ([doi:10.1371/journal.pcbi.1008873](https://doi.org/10.1371/journal.pcbi.1008873)).

The author list, complete with affiliations is given below:

Luc Bassel<sup>1,2\*</sup>, Paul Medvedev<sup>3,4,5</sup>, Rayan Chikhi<sup>1</sup>

**1** Sequence Bioinformatics, Department of Computational Biology, Institut Pasteur, Paris, France

**2** Sorbonne Université, Collège doctoral, Paris, France

**3** Department of Computer Science and Engineering, Pennsylvania State University, University Park, Pennsylvania, United States of America

**4** Department of Biochemistry and Molecular Biology, Pennsylvania State University, University Park, Pennsylvania, United States of America

**5** Center for Computational Biology and Bioinformatics, Pennsylvania State University, University Park, Pennsylvania, United States of America

## Abstract

Sequencing errors continue to pose algorithmic challenges to methods working with sequencing data. One of the simplest and most prevalent techniques for ameliorating the detrimental effects of homopolymer expansion/contraction errors present in long read data is homopolymer compression. It collapses runs of repeated nucleotides, with the intuitive goal of removing some of the sequencing errors and often improving mapping sensitivity. Though our intuitive understanding justifies why homopolymer compression works, it in no way implies that it is the best transformation that can be done. In this paper, we explore if there are transformations that can be applied in the same pre-processing manner as homopolymer compression that would achieve better alignment sensitivity. We introduce a more general framework than homopolymer compression, called mapping-friendly sequence reductions. We transform the reference and the reads using these reductions and then apply an alignment algorithm. We demonstrate that some mapping-friendly sequence reductions lead to improved mapping accuracy, outperforming homopolymer compression.

### 3.1. Introduction

Sequencing errors continue to pose algorithmic challenges to methods working with read data. In short-read technologies, these tend to be substitution errors, but in long reads, these tend to be short insertions and deletions; most common are expansions or contractions of homopolymers (i.e. reporting 3 As instead of 4) [97]. Many algorithmic problems, such as alignment, become trivial if not for sequencing errors [123]. Error correction can often decrease the error rate but does not eliminate all errors. Most tools therefore incorporate the uncertainty caused by errors into their underlying algorithms. The higher the error rate, the more detrimental its effect on algorithm speed, memory, and accuracy. While the sequencing error rate of any given technology tends to decrease over time, new technologies entering the market typically have high error rates (e.g. Oxford Nanopore Technologies). Finding better ways to cope with sequencing error therefore remains a top priority in bioinformatics.

One of the simplest and most prevalent techniques for ameliorating the detrimental effects of homopolymer expansion/contraction errors is *homopolymer compression* (abbreviated HPC). HPC simply transforms runs of the same nucleotide within a sequence into a single occurrence of that nucleotide. For example, HPC applied to the sequence AAAGGGTTA yields the sequence AGTA. To use HPC in an alignment algorithm, one first compresses the reads and the reference, then aligns each compressed read to the compressed reference, and finally reports all alignment locations, converted into the coordinate system of the uncompressed reference. HPC effectively removes homopolymer expansion/contraction errors from the downstream algorithm. Though there is a trade-off with specificity of the alignment (e.g. some of the compressed alignments may

## 3.2. METHODS

not correspond to true alignments) the improvement in mapping sensitivity usually outweighs it [115].

The first use of HPC that we are aware of was in 2008 as a pre-processing step for 454 pyrosequencing data in the Celera assembler [110]. It is used by a wide range of error-correction algorithms, e.g. for 454 data [261], PacBio data [113], and Oxford Nanopore data [262]. HPC is used in alignment, e.g. by the widely used minimap2 aligner [115]. HPC is also used in long-read assembly, e.g. HiCanu [108], SMARTdenovo [263], or mdBG [109]. HPC is also used for clustering transcriptome reads according to gene family of origin [112]. Overall, HPC has been widely used, with demonstrated benefits.

Though our intuitive understanding justifies why HPC works, it in no way implies that it is the best transformation that can be done. Are there transformations that can be applied in the same pre-processing way as HPC that would achieve better alignment sensitivity? In this work, we define a more general notion which we call *mapping-friendly sequence reductions*. In order to efficiently explore the performance of all reductions, we identify two heuristics to reduce the search space of reductions. We then identify a number of mapping-friendly sequence reductions which are likely to yield better mapping performance than HPC. We evaluate them using two mappers (`minimap2` and `winnowmap2`) on three simulated datasets (whole human genome, human centromere, and whole *Drosophila* genome). We show that some of these functions provide vastly superior performance in terms of correctly placing high mapping quality reads, compared to either HPC or using raw reads. For example, one function increased the mapping accuracy of `minimap2` by an order of magnitude over the entire human genome, keeping an identical fraction of reads mapped.

We also evaluate whether HPC sensitivity gains continue to outweigh the specificity cost with the advent of telomere-to-telomere assemblies [4]. These contain many more low-complexity and/or repeated regions such as centromeres and telomeres. HPC may increase mapping ambiguity in these regions by removing small, distinguishing, differences between repeat instances. Indeed, we find that neither HPC nor our mapping-friendly sequence reductions perform better than mapping raw reads on centromeres, hinting at the importance of preserving all sequence information in repeated regions.

## 3.2. Methods

### 3.2.1. Streaming sequence reductions

We wish to extend the notion of homopolymer compression to a more general function while maintaining its simplicity. What makes HPC simple is that it can be done in a streaming fashion over the sequence while maintaining only a local context. The algorithm can be viewed simply as scanning a string from left to right and, at each new character, outputting that character if and only if it is different from the previous

## CHAPTER 3

character. In order to prepare for generalizing this algorithm, let us define a function  $g^{\text{HPC}} : \Sigma^2 \rightarrow \Sigma \cup \{\varepsilon\}$  where  $\Sigma$  is the DNA alphabet,  $\varepsilon$  is the empty character, and

$$g^{\text{HPC}}(x_1 \cdot x_2) = \begin{cases} x_2 & \text{if } x_1 \neq x_2 \\ \varepsilon & \text{if } x_1 = x_2 \end{cases}$$

Now, we can view HPC as sliding a window of size 2 over the sequence and at each new window, applying  $g^{\text{HPC}}$  to the window and concatenating the output to the growing compressed string. Formally, let  $x$  be a string, which we index starting from 1. Then, the HPC transformation is defined as

$$f(x) = x[1, \ell - 1] \cdot g(x[1, \ell]) \cdot g(x[2, \ell + 1]) \cdots g(x[|x| - \ell + 1, |x|]) \quad (3.1)$$

where  $\ell = 2$  and  $g = g^{\text{HPC}}$ . In other words,  $f$  is the concatenation of the first  $\ell - 1$  characters of  $x$  and the sequence of outputs of  $g$  applied to a sliding window of length  $\ell$  over  $x$ . The core of the transformation is given by  $g$  and the size of the context  $\ell$ , and  $f$  is simply the wrapper for  $g$  so that the transformation can be applied to arbitrary length strings.

With this view in mind, we can generalize HPC while keeping its simplicity by 1) considering different functions  $g$  that can be plugged into Equation (3.1) increasing the context that  $g$  uses (i.e. setting  $\ell > 2$ ). Formally, for a given alphabet  $\Sigma$  and a context size  $\ell$ , a function  $T$  mapping strings to strings is said to be an *order- $\ell$*  Streaming sequence reduction (abbreviated *SSR*) if there exists some  $g : \Sigma^\ell \rightarrow \Sigma \cup \{\varepsilon\}$  such that  $T = f$ .

Figure 3.1A shows how an SSR can be visualized as a directed graph. Observe that an order- $\ell$  SSR is defined by a mapping between  $|\Sigma|^\ell$  inputs and  $|\Sigma| + 1$  outputs. For example, for  $\ell = 2$ , there are  $n = 16$  inputs and  $k = 5$  outputs. Figure 3.1B visualizes HPC in this way.

Since we aim to use SSRs in the context of sequencing data, we need to place additional restrictions on how they handle reverse complements. For example, given two strings  $x$  (e.g. a read) and  $y$  (e.g. a substring of the reference), a mapper might check if  $x = RC(y)$ . When strings are pre-processed using an SSR  $f$ , it will end up checking if  $f(x) = RC(f(y))$ . However,  $x = RC(y)$  only implies that  $f(x) = f(RC(y))$ . In order to have it also imply that  $f(x) = RC(f(y))$ , we need  $f$  to be commutative with  $RC$ , i.e. applying SSR then  $RC$  needs to be equivalent to applying  $RC$  then SSR. We say that  $f$  is *RC-insensitive* if for all  $x$ ,  $f(RC(x)) = RC(f(x))$ . Observe that HPC is RC-insensitive.

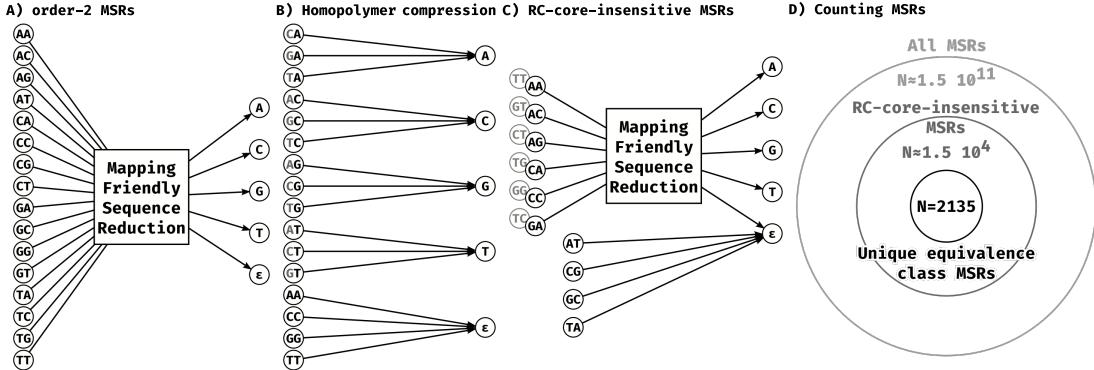


Figure 3.1.: Representing and counting Streaming sequence reductions.

**A:** General representation of an order-2 Streaming sequence reduction as a mapping of 16 input dinucleotides, to the 4 nucleotide outputs and the empty character  $\varepsilon$ . **B:** Homopolymer compression is an order-2 SSR. All dinucleotides except those that contain the same nucleotide twice map to the second nucleotide of the pair. The 4 dinucleotides that are the two same nucleotides map to the empty character  $\varepsilon$ . **C:** Our RC-core-insensitive order-2 SSRs are mappings of the 6 representative dinucleotide inputs to the 4 nucleotide outputs and the empty character  $\varepsilon$ . The 4 dinucleotides that are their own reverse complement are always mapped to  $\varepsilon$ . The remaining 6 dinucleotides are mapped to the complement of the mapped output of the reverse complement dinucleotide input. For example, if AA is mapped to C, then TT (the reverse complement of AA) will be mapped to G (the complement of C). **D:** Number of possible SSR mappings under the different restrictions presented in the main text. All mappings from 16 dinucleotide inputs to 5 outputs (as in panel A) are represented by the outermost circle. All RC-core-insensitive mappings (as in panel C) are represented by the medium circle. All RC-core-insensitive mappings with only one representative of each equivalence class are represented by the innermost circle.

### 3.2.2. Restricting the space of Streaming sequence reductions

To discover SSRs that improve mapping performance, our strategy is to put them all to the test by evaluating the results of an actual mapping software over a simulated test dataset reduced by each SSR. However, even with only 16 inputs and 5 outputs, the number of possible  $g$  mappings for order-2 SSRs is  $5^{16} \approx 1.5 \cdot 10^{11}$ , which is prohibitive to enumerate. In this section, we describe two ideas for reducing the space of SSRs that we will test. In subsection 3.2.2.1, we show how the restriction to RC-insensitive mappings can be used to reduce the search space. In subsection 3.2.2.2, we exploit the natural symmetry that arises due to Watson-Crick complements to further restrict the search space.

These restrictions reduce the number of order-2 SSRs to only , making it feasible to test

all of them. Figure 3.1D shows an overview of our restriction process.

### 3.2.2.1. Reverse complement-core-insensitive Streaming sequence reductions

Consider an SSR defined by a function  $g$ , as in Equation (3.1). Throughout this paper we will consider SSRs that have a related but weaker property than RC-insensitive. We say that an SSR is *RC-core-insensitive* if the function  $g$  that defines it has the property that for every  $\ell$ -mer  $x$  and its reverse complement  $y$ , we have that either  $g(x)$  is the reverse complement of  $g(y)$  or  $g(x) = g(y) = \varepsilon$ . We will restrict our SSR search space to RC-core-insensitive reductions in order to reduce the number of SSRs we will need to test.

Let us consider what this means for the case of  $\ell = 2$ , which will be the focal point of our experimental analysis. There are 16  $\ell$ -mers(i.e. dinucleotides) in total. Four of them are their own reverse complement: AT, TA, GC, CG. The RC-core-insensitive restriction forces  $g$  to map each of these to  $\varepsilon$ , since a single nucleotide output cannot be its own reverse complement. This leaves 12  $\ell$ -mers, which can be broken down into 6 pairs of reverse complements. For each pair, we can order them in lexicographical order and write them as  $(AA, TT)$ ,  $(AC, GT)$ ,  $(AG, CT)$ ,  $(CA, TG)$ ,  $(CC, GG)$ , and  $(GA, TC)$ . Defining  $g$  can then be done by assigning an output nucleotide to the first  $\ell$ -mer in each of these pairs (Figure 3.1C). For example, we can define an SSR by assigning  $g(AA) = C$ ,  $g(AC) = C$ ,  $g(AG) = A$ ,  $g(CA) = A$ ,  $g(CC) = T$ , and  $g(GA) = G$ . As an example, let us apply the corresponding SSR to an example read  $r$ :

$$\begin{array}{ll} r = \text{TAAGTTGA} & f(RC(r)) = \text{TCACCTG} \\ f(r) = \text{TCAGGTG} & RC(f(r)) = \text{CACCTGA} \\ RC(r) = \text{TCAACTTA} & \end{array}$$

Observe that the first  $\ell - 1$  nucleotides of  $r$  (shown in red) are copied as-is, since we do not apply  $g$  on them (as per Equation (3.1)). As we see in this example, this implies that  $f(RC(r))$  is not necessarily equal to  $RC(f(r))$ ; thus an RC-core-insensitive SSR is not necessarily an RC-insensitive SSR. However, an RC-core-insensitive SSR has the property that for all strings  $r$ , we have  $f(RC(r))[\ell, |r|] = RC(f(r))[1, |r| - \ell + 1]$ . In other words, if we drop the  $\ell - 1$  prefix of  $f(RC(r))$  and the  $\ell - 1$  suffix of  $RC(f(r))$ , then the two strings are equal. Though we no longer have the strict RC-insensitive property, this new property suffices for the purpose of mapping long reads. Since the length of the read sequences will be much greater than  $\ell$  (in our results we will only use  $\ell = 2$ ), having a mismatch in the first or last nucleotide will be practically inconsequential.

It is important to note though that there may be other RC-insensitive functions not generated by this construction. For instance, HPC cannot be derived using this method

## 3.2. METHODS

(as it does not map the di-nucleotides AT,TA,GC and CG to  $\varepsilon$ ), and yet it is RC-insensitive.

We can count the number of RC-core-insensitive SSRs. Let us define  $i(\ell)$  the number of input assignments necessary to fully determine the RC-core-insensitive SSR; one can think of this as the degrees-of-freedom in choosing  $g$ . As we showed, for  $\ell = 2$ , we have  $i(\ell) = 6$ . The number of RC-core-insensitive SSRs is then  $5^{i(\ell)}$ . Therefore, for  $\ell = 2$ , instead of  $5^{16}$  possible mappings we have at most  $5^6 \approx 1.5 \cdot 10^4$  RC-core-insensitive mappings (Figure 3.1D). For an odd  $\ell > 2$ , there are no  $\ell$ -mers that are their own reverse complements, hence  $i(\ell) = 4^\ell/2$ . If  $\ell$  is even then there are  $4^{\ell/2}$  inputs that are their own reverse complements (i.e. we take all possible sequences of length  $\ell/2$  and reconstruct the other half with reverse complements). Thus,  $i(\ell) = (4^\ell - 4^{\ell/2})/2$ .

### 3.2.2.2. Equivalence classes of SSRs

When performing preliminary tests, we noticed that swapping  $A \leftrightarrow T$  and/or  $C \leftrightarrow G$ , as well as swapping the whole  $A/T$  pair with the  $C/G$  pair in the SSR outputs did not affect the performance. In other words, we could exchange the letters of the output in a way that preserves the Watson-Crick complementary relation. Intuitively, this can be due to the symmetry induced by reverse complements in nucleic acid strands, though we do not have a more rigorous explanation for this effect. In this section, we will formalize this observation by defining the notion of SSR equivalence. This will reduce the space of SSRs that we will need to consider by allowing us to evaluate only one SSR from each equivalence class.

Consider an RC-core-insensitive SSR defined by a function  $g$ , as in Equation (3.1). An  $\ell$ -mer is canonical if it is not lexicographically larger than its reverse complement. Let  $I$  be the set of all  $\ell$ -mers that are canonical and are not reverse complements of each other. Such an SSR's *dimension*  $k$  is the number of distinct nucleotides that can be output by  $g$  on inputs from  $I$  (not counting  $\varepsilon$ ). The dimension can range from 1 to 4. Next, observe that  $g$  maps all elements of  $I$  to one of  $k+1$  values (i.e.  $\Sigma \cup \varepsilon$ ). The output of  $g$  on  $\ell$ -mers not in  $I$  is determined by its output on  $\ell$ -mers in  $I$ , since we assume the SSR is RC-core-insensitive. We can therefore view it as a partition of  $I$  into  $k+1$  sets  $S_0, \dots, S_k$ , and then having a function  $t$  that is an injection from  $\{1, \dots, k\}$  to  $\Sigma$  that assigns an output letter to each partition. Further, we permanently assign the output letter for  $S_0$  to be  $\varepsilon$ . Note that while  $S_0$  could be empty,  $S_1, \dots, S_k$  cannot be empty by definition of dimension. For example, the SSR used in Section 3.2.2.1 has dimension four and corresponds to the partition  $S_0 = \{\}, S_1 = \{AG, CA\}, S_2 = \{CC\}, S_3 = \{AA, AC\}, S_4 = \{GA\}$ , and to the injection  $t(1) = A, t(2) = T, t(3) = C$ , and  $t(4) = G$ .

Let  $\text{IsCOMP}(x, y)$  be a function that returns true if two nucleotides  $x, y \in \Sigma \cup \{\varepsilon\}$  are Watson-Crick complements, and false otherwise. Consider two SSRs of dimension  $k$  defined by  $S_0, \dots, S_k, t$  and  $S'_0, \dots, S'_k, t'$ , respectively. We say that they are equivalent iff all the following conditions are met:

## CHAPTER 3

- $S_0 = S'_0$ ,
- there exists a permutation  $\pi$  of  $\{1, \dots, k\}$  such that for all  $1 \leq i \leq k$ , we have  $S_i = S'_{\pi(i)}$ ,
- for all  $1 \leq i < j \leq k$ , we have  $\text{IsCOMP}(t(i), t(j)) = \text{IsCOMP}(t'(\pi(i)), t'(\pi(j)))$ .

One can verify that this definition is indeed an equivalence relation, i.e. it is reflexive, symmetric, and transitive. Therefore, we can partition the set of all SSRs into equivalence classes based on this equivalence relation. One caveat is that a single SSR defined by a function  $g$  may correspond to multiple SSRs of the form  $S_0, \dots, S_k, t$ . However, these multiple SSRs are equivalent, hence the resulting equivalence classes are not affected. Furthermore, we can assume that there is some rule to pick one representative SSR for its equivalence class; the rule itself does not matter in our case.

Figure 3.1 shows the equivalence classes for  $\ell = 2$ , for a fixed partition. An equivalence class can be defined by which pair of classes  $S_i$  and  $S_j$  have complementary outputs under  $t$  and  $t'$ . Let us define  $o(k)$  as the number of equivalence classes for a given partition and a given  $k$ . Then Figure 3.1 shows that  $o(1) = 1$ ,  $o(2) = 2$  and  $o(3) = o(4) = 3$ . There are thus only 9 equivalence classes for a given partition.

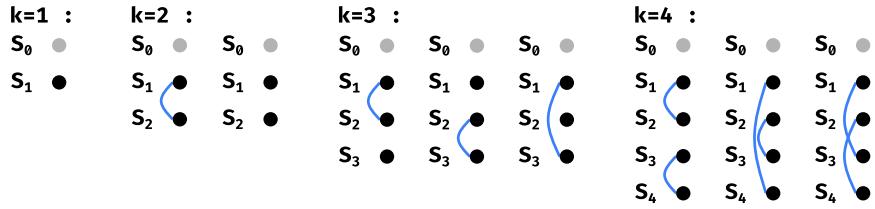


Figure 3.2.: SSR equivalence classes for a fixed partition of the inputs.

$S_0$  is always assigned  $\varepsilon$ , so it is represented by a gray node. A blue link between  $S_i$  and  $S_j$  denotes that  $\text{IsCOMP}(t(i), t(j)) = \text{true}$ . The equivalence classes are determined by the Watson-Crick complementary relationships between the rest of the parts, i.e. by all the possible ways to draw the blue links.

### 3.2.2.3. Counting the number of restricted SSRs

In this section, we derive a formula for the number of restricted MSRs, i.e. MSRs that are RC-core-insensitive and that are representative for their equivalence class. Consider the class of RC-core-insensitive MSRs with dimension  $k$ . In subsection 3.2.2.1, we derived that the degrees-of-freedom in assigning  $\ell$ -mers to an output is  $i(\ell) = 4^\ell / 2$  if  $\ell$  is odd and  $i(\ell) = (4^\ell - 4^{\ell/2}) / 2$  if  $\ell$  is even. Let  $C(\ell, k)$  be the number of ways that  $i(\ell)$   $\ell$ -mers can be partitioned into  $k + 1$  sets  $S_0, \dots, S_k$ , with  $S_1, \dots, S_k$  required to be non-empty. Then, in subsection 3.2.2.2, we have derived  $o(k)$ , the number of MSR equivalence classes for each such partition. The number of restricted MSRs can then be written as

### 3.2. METHODS

$$N(\ell) = \sum_{k=1}^4 C(\ell, k) \cdot o(k) \quad (3.2)$$

To derive the formula for  $C(\ell, k)$ , we first recall that the number of ways to partition  $n$  elements into  $k$  non-empty sets is known as the Stirling number of the second kind and is denoted by  $\{n\}_k$  [264, p.265]. It can be computed using the formula

$$\{n\}_k = \frac{1}{k!} \sum_{i=0}^k (-1)^i \binom{k}{i} (k-i)^n$$

Let  $j$  be the number of the  $i(\ell)$   $\ell$ -mers that are assigned to  $S_0$ . Note this does not include the  $\ell$ -mers that are self-complementary that are forced to be in  $S_0$ . Let  $C(\ell, k, j)$  be the number of ways that  $i(\ell)$   $\ell$ -mers can be partitioned into  $k+1$  sets  $S_0, \dots, S_k$ , such that  $j$  of the  $\ell$ -mers go into  $|S_0|$  and  $S_1, \dots, S_k$  to are non-empty. We need to consider several cases depending on the value of  $j$ :

- In the case that  $j = 0$ , we are partitioning the  $i(\ell)$  inputs among non-empty sets  $S_1, \dots, S_k$ . Then  $C(\ell, k, j) = \{i(\ell)\}_k$ .
- In the case that  $1 \leq j \leq i(\ell) - k$ , there are  $\binom{i(\ell)}{j}$  ways to choose which  $j$   $\ell$ -mers are in  $S_0$ , and  $\{i(\ell)-j\}_k$  ways to partition the remaining  $\ell$ -mers into  $S_1, \dots, S_k$ . Hence,  $C(\ell, k, j) = \binom{i(\ell)}{j} \{i(\ell)-j\}_k$ .
- In the case that  $j > i(\ell) - k$ , it is impossible to partition the remaining  $k$  (or fewer)  $\ell$ -mers into  $S_1, \dots, S_k$  such that the sets are non-empty. Recall that as we assume the dimension is  $k$ , each set must contain at least one element. Hence,  $C(\ell, k, j) = 0$ .

Putting this together into Equation (3.2), we get

$$N(\ell) = \sum_{k=1}^4 o(k) \left( \{i(\ell)\}_k + \sum_{j=1}^{i(\ell)-k} \binom{i(\ell)}{j} \{i(\ell)-j\}_k \right)$$

For  $\ell = 2$ , we have  $N(2) = 2,135$  restricted MSRs, which is several orders of magnitude smaller than the initial  $5^{16}$  possible MSRs and allows us to test the performance of all of them. for order-3 MSRs we get  $N(3) = 2.9 \cdot 10^{21}$  which much smaller than the full search space of  $5^{4^3} \approx 5.4 \cdot 10^{44}$ , for order-4 MSRs we get a similar reduction in search space with  $N(4) = 9.4 \cdot 10^{84}$  as opposed to the full search space of  $5^{4^4} \approx 8.6 \cdot 10^{178}$ . For these higher order MSRs, although the restricted search space is much smaller than the full naive one, it is still too large to exhaustively search.

### 3.3. Datasets and Pipelines

#### 3.3.1. Datasets

The following three reference sequences were used for evaluation:

1. **Whole human genome:** This reference sequence is a whole genome assembly of the CHM13hTERT human cell line by the Telomere-to-Telomere consortium [@ 4]. We used the 1.1 assembly release (Genbank Assembly ID [GCA\\_009914755.3](#)).
2. **Whole *Drosophila* genome:** This reference sequence is a whole genome assembly of a *Drosophila melanogaster*, release 6.35 (Genbank Assembly ID [GCA\\_000001215.4](#)) [265].
3. **Synthetic centromeric sequence:** This sequence was obtained from the TandemTools mapper test data [260]. It is a simulated centromeric sequence that is inherently difficult to map reads to. Appendix A.1 describes how it was constructed.

#### 3.3.2. Simulation pipeline

Given a reference sequence, simulated reads were obtained using `nanosim` [266] with the `human_NA12878_DNA_FAB49712_guppy_flipflop` pre-trained model, mimicking sequencing with an Oxford Nanopore instrument. The number of simulated reads was chosen to obtain a theoretical coverage of whole genomes around 1.5x, this resulted in simulating  $\approx 6.6 \cdot 10^5$  reads for the whole human genome and  $\approx 2.6 \cdot 10^4$  reads for the whole *Drosophila* genome. Since the centromeric sequence is very short, we aimed for a theoretical coverage of 100x which resulted in  $\approx 1.3 \cdot 10^4$  simulated reads.

For each evaluated SSR, the reads as well as the reference sequence were reduced by applying the SSR to them. The reduced reads were then mapped to the reduced reference using `minimap2`[115] with the `map-ont` preset and the `-c` flag to generate precise alignments. Although HPC is an option in `minimap2` we do not use it and we evaluate HPC as any of the other SSRs by transforming the reference and reads prior to mapping. The starting coordinates of the reduced reads on the reduced reference were updated to reflect deletions incurred by the reduction process. The mapping results with translated coordinates were filtered to keep only the primary alignments. This process was done for each of our 2135 SSRs as well as with HPC and the original untransformed reads (denoted as *raw*).

### 3.3.3. Evaluation pipeline

We use two metrics to evaluate the quality of a mapping of a simulated read set. The first is the *fraction of reads mapped*, i.e. that have at least one alignment. The second is the *error rate*, which is the fraction of mapped reads that have an incorrect location as determined by `paftools mapeval` [115]. This tool considers a read as correctly mapped if the intersection between its true interval of origin, and the interval where it has been mapped to, is at least 10% of the union of both intervals.

Furthermore, we measure the error rate as a function of a given *mapping quality threshold*. Mapping quality (abbreviated mapq) is a metric reported by the aligner that indicates its confidence in read placement; the highest value (60) indicates that the mapping location is likely correct and unique with high probability, and a low value (e.g. 0) indicates that the read has multiple equally likely candidate mappings and that the reported location cannot be trusted. The error rate at a mapq threshold  $t$  is then defined as the error rate of reads whose mapping quality is  $t$  or above. For example, the error rate at  $t = 0$  is the error rate of the whole read set, while the error rate at  $t = 60$  is the error rate of only the most confident read mappings. Observe that the error rate decreases as  $t$  increases.

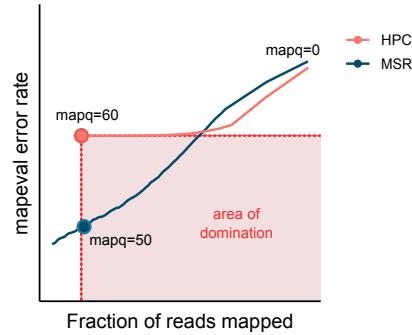
## 3.4. Results

### 3.4.1. Selection of mapping-friendly sequence reductions

We selected a set of “promising” SSRs starting from all of the SSRs enumerated in Section 3.2.2, that we call *mapping-friendly sequence reductions* (abbreviated *MSR*). The selection was performed by considering an independent read set of lower (0.5x) coverage, simulated from the whole human genome reference. This dataset is separate from the ones used for evaluation. Note that overfitting MSRs to a particular genome is acceptable in applications where a custom MSR can be used for each genome. Yet in this work, the same set of selected MSRs will be used across all genomes.

For each evaluated SSR, we selected, if it exists, the highest mapq threshold for which the mapped read fraction is higher and the error rate is lower than HPC at mapq 60 (0.93 and  $2.1 \cdot 10^{-3}$  respectively). Figure 3.3 illustrates the idea. Then we identified the 20 SSRs that have the highest fraction of reads mapped at their respective thresholds. Similarly we identified the 20 SSRs with the lowest error rate. Finally we select the 20 SSRs that have the highest percentage of thresholds “better” than HPC at mapq 60; i.e. the number of mapq thresholds for which the SSR has both a higher fraction of reads mapped and lower error rate than HPC at a mapq threshold of 60, divided by the total number of thresholds (=60).

The union of these 3 sets of 20 SSRs resulted in a set of 58 “promising” MSRs. Furthermore, we will highlight three MSRs that are “best in their category”, i.e.



**Figure 3.3.: Illustration of how a respective mapq threshold is chosen for each of our evaluated MSRs.**

The orange dot shows the error rate and fraction of reads mapped for HPC at mapq threshold 60. Anything below and to the right of this point is strictly better than HPC 60, i.e. it has both a lower error rate and higher fraction of reads mapped. If an evaluated MSR does not pass through this region, then it is discarded from further consideration. In the figure, the blue MSR does pass through this region, indicating that it is better than HPC 60. We identify the leftmost point (marked as a blue dot) and use the mapq threshold at that point as the respective threshold.

- **MSR<sub>F</sub>:** The MSR with the highest fraction of mapped reads at a mapq threshold of 0.
- **MSR<sub>E</sub>:** The MSR with the lowest error rate at its respective mapq threshold.
- **MSR<sub>P</sub>:** The MSR with the highest percentage of mapq thresholds for which it is “better” than HPC at mapq 60.

Figure 3.4 shows the actual functions  $\text{MSR}_F$ ,  $\text{MSR}_E$ ,  $\text{MSR}_P$ . An intriguing property is that they output predominantly As and Ts, with  $\text{MSR}_P$  assigning 2 input pairs to the G/C output whereas  $\text{MSR}_E$  and  $\text{MSR}_F$  assign only one. Additionally,  $\text{MSR}_E$  and  $\text{MSR}_P$  both assign the {CC,GG} input pair to the deletion output  $\varepsilon$  removing any information corresponding to repetitions of either G or C from the reduced sequence. Overall this means the reduced sequences are much more AT-rich than their raw counterparts, but somehow information pertinent to mapping is retained.

### 3.4.2. Mapping-friendly sequence reductions lead to lower mapping errors on whole genomes

Across the entire human genome, at high mapping quality thresholds (above 50), our selected 58 MSRs generally have lower mapping error rate than HPC and raw Figure 3.5A and Table 3.1. This is not surprising, as we selected those MSRs partly on the criteria of outperforming HPC at mapq 60; however, it does demonstrate that we did not overfit to the simulated reads used to select the MSRs.

### 3.4. RESULTS

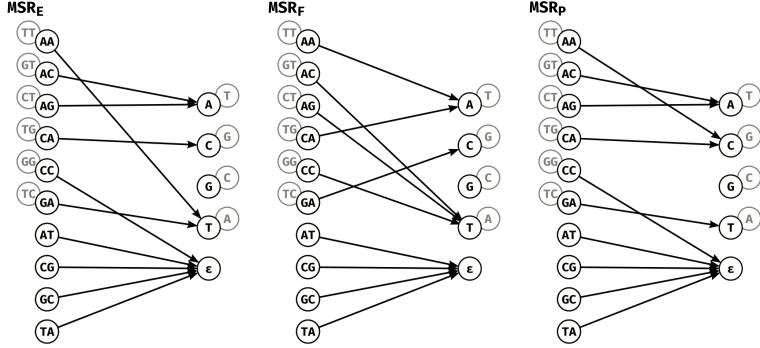


Figure 3.4.: **Graph representations of our highlighted MSRs:  $\text{MSR}_E$ ,  $\text{MSR}_F$ , and  $\text{MSR}_P$ .**

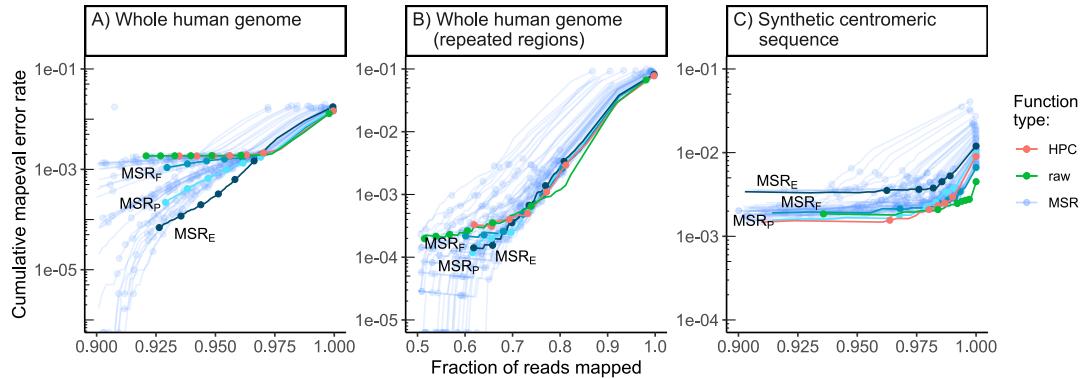
$\text{MSR}_E$  has the lowest error rate of among MSRs at the highest mapq threshold for which it performs better than HPC at mapq 60,  $\text{MSR}_F$  has the highest fraction of reads mapped at mapq 60 and  $\text{MSR}_P$  has the highest percentage of mapq thresholds for which it outperforms HPC at mapq 60. The grayed out nodes represent the reverse complement of input dinucleotides and outputs, as in Figure 3.1C. For example for  $\text{MSR}_E$ , AA is mapped to T, so TT is mapped to A.

Mapping quality is only an indication from the aligner to estimate whether a read mapping is correct, and according to Figure 3.5A the mapping error rate of most MSRs is low even for mapping qualities lower than 60. Therefore, we choose to compare MSR-mapped reads with lower mapping qualities against raw or HPC-mapped reads with the highest (60) mapping quality (which is the mapping quality thresholds most practitioners would use by default).

Table 3.1 shows that the three selected MSRs outperform both HPC and raw in terms of mapping error rate, with similar fractions of mapped reads overall. For example on the human genome, at  $\text{mapq} \geq 50$ ,  $\text{MSR}_F$ ,  $\text{MSR}_P$  and  $\text{MSR}_E$  all map more reads than either HPC or raw at  $\text{mapq}=60$ , and  $\text{MSR}_P$  and  $\text{MSR}_E$  also have error rates an order of magnitude lower than either HPC or raw.

To evaluate the robustness of MSRs E, F and P we investigated the impact of mapping to a different organism or using another mapper. To this effect we repeated the evaluation pipeline in these different settings:

- Using the *Drosophila melanogaster* whole genome assembly as reference and mapping with `minimap2`
- Using the whole human genome assembly as reference but mapping with `winnnowmap2`(version 2.02) [116]. The same options as `minimap2` were used, and k-mers were counted using `meryl` [267], considering the top 0.02% as repetitive (as suggested by the `winnnowmap2` usage guide).



**Figure 3.5.: Performance of our 58 selected mapping-friendly sequence reductions across genomes on reads simulated by nanosim**

Panel **A**) shows the whole human genome assembly, **B**) the subset of mapped reads from panel B that originate from repetitive regions, and **C**) the “TandemTools” synthetic centromeric reference sequence. We highlighted the best-performing mapping-friendly sequence reductions as MSR E, F and P, respectively in terms of cumulative mapeval error rate, fraction of reads mapped, and percentage of better thresholds than HPC. Each point on a line represents, from left to right, the mapping quality thresholds 60, 50, 40, 30, 20, 10 and 0. For the first point of each line, only reads of mapping quality 60 are considered, and the y value represents the rate of these reads that are not correctly mapped, the x value represents the fraction of simulated reads that are mapped at this threshold. The next point is computed for all reads of mapping quality  $\geq 50$ , etc. The rightmost point on any curve represents the mapping error rate and the fraction of mapped reads for all primary alignments. The x-axes are clipped for lower mapped read fractions to better differentiate HPC, raw and MSRs E, F and P.

## 3.4. RESULTS

MSRs E, F and P behave very similarly in both of these contexts compared to HPC/raw: by selecting mapped reads with  $\text{mapq} \geq 50$  for the three MSRs we obtain a similar fraction of mapped reads with much lower error rates (Table 3.1). A noticeable exception is the `winnowmap2` experiment, where a larger fraction of raw reads are mapped than any other MSR and even HPC. A more detailed results table can be found in Table A.1, and a graph of MSR performance on the whole Drosophila genome in Figure A.7. As Figure A.7 shows, we also evaluated these MSRs on a whole *Escherichia coli* (Genbank ID U00096.2) genome, where we observed similar results, albeit the best MSRs do not seem to be one of our three candidates. This might mean that specific MSRs are more suited to particular types of genomes.

Whole human genome minimap2			Whole human genome winnowmap2			Whole Drosophila genome minimap2		
mapq	fraction	error	fraction	error	fraction	error		
HPC	60	0.935 +0%	1.85e-03 + 0%	0.894 +0%	1.43e-03 + 0%	0.957 +0%	2.27e-03 + 0%	
raw	60	0.921 -1%	1.86e-03 + 0%	<b>0.932 +4%</b>	1.75e-03 +23%	0.958 +0%	2.27e-03 - 0%	
MSR <sub>F</sub>	50	<b>0.938 +0%</b>	1.29e-03 -30%	0.886 -1%	3.82e-04 -73%	<b>0.960 +0%</b>	1.37e-03 - 39%	
MSR <sub>E</sub>	50	0.936 +0%	<b>1.17e-04 -94%</b>	0.820 -8%	<b>8.93e-05 -94%</b>	0.954 -0%	<b>0.00 -100%</b>	
MSR <sub>P</sub>	50	<b>0.938 +0%</b>	4.15e-04 -78%	0.845 -6%	1.14e-04 -92%	0.957 +0%	8.11e-04 - 64%	

Table 3.1.: **Performance of MSRs, HPC, and raw mappings across different mappers and reference sequences.**

For each reference sequence and mapper pair, we report the fraction of reads mapped (“fraction” columns), the `paftools mapeval` mapping error rate (“error” columns). The percentage differences are computed w.r.t to the respective HPC value. For HPC and the raw these metrics were obtained for alignments of mapping quality of 60. For MSRs E, F and P these metrics were obtained for alignments of mapping quality  $\geq 50$ .

### 3.4.3. Mapping-friendly sequence reductions increase mapping quality on repeated regions of the human genome

To evaluate the performance of our MSRs specifically on repeats, we extracted the simulated reads for which an overlap with repeated region of the whole human genome was greater than 50% of the read length. We then evaluated the MSRs on these reads only. Repeated regions were obtained from <https://t2t.gi.ucsc.edu/chm13/hub/t2t-chm13-v1.1/rmsk/rmsk.bigBed>.

We obtained similar results as on the whole human genome, with MSRs E, F and P performing better than HPC at mapq 50 (Figure 3.5B). At a mapq threshold of 50, the error rate is 53%, 31%, and 39% lower than HPC at mapq 60 for MSRs E, F and P respectively, while the fraction of mapped reads remains slightly higher. At mapq=60, raw has a error rate 40% lower than HPC but it the mapped fraction is also 17% lower.

### 3.4.4. Raw mapping improves upon HPC on centromeric regions

On the “TandemTools” centromeric reference, HPC consistently maps a smaller fraction of reads than raw, across all mapping quality thresholds (Figure 3.5C). Additionally, the error rate for raw is often inferior to that of HPC. The same is true for our selected MSRs: most of them have comparable performance to HPC, but none of them outperform raw mapping (Figure 3.5C).

We conjecture this is due to the highly repetitive nature of centromeres. HPC likely removes small unique repetitions in the reads and the reference that might allow mappers to better match reads to a particular occurrence a centromeric pattern. Mapping raw reads on the other hand preserves all bases in the read and better differentiates repeats. Therefore it seems inadvisable to use HPC when mapping reads to highly repetitive regions of a genome, such as a centromere.

### 3.4.5. Positions of incorrectly mapped reads across the entire human genome

To study how MSRs E, F, and P improve over HPC and raw mapping in terms of error rate on the human genome, we selected all the primary alignments that `paftools mapeval` reported as incorrectly mapped. For HPC and raw, only alignments of mapping quality equal to 60 were considered. To report a comparable fraction of aligned reads, we selected alignments of mapping quality  $\geq 50$  for MSRs. We then reported the origin of those incorrectly mapped reads on whole human genome reference, shown per-chromosome in Appendix A.1.

We observe that erroneously mapped reads are not only those from centromeres, and instead originate from many other genomic regions. MSRs E and P have a markedly lower number of these incorrect mappings than either HPC or raw, with 1118 incorrect mappings for raw mappings and 1130 for HPC as opposed to 549, 970 and 361 for MSRs E, F and P respectively. This stays true even for difficult regions of the genome such as chromosome X, where raw and HPC have 70 incorrect mappings as opposed MSRs E and P that have 39, and 27 errors respectively.

We also investigated where all simulated reads were mapped on the whole human genome assembly, for raw, HPC and MSRs E,F and P in Figures A.2 to A.6. The correctly mapped reads are, as expected, evenly distributed along each chromosome. The incorrectly mapped reads are however unevenly distributed. For most chromosomes there is a sharp peak in the distribution of incorrectly mapped reads, located at the position of the centromere. For the acrocentric chromosomes, there is a second peak corresponding to the “stalk” satellite region, with an enrichment of incorrectly mapped reads. This is expected since both centromeres and “stalks” are repetitive regions which are a challenge for mapping. For chromosomes 1, 9 and 16 however the majority of incorrectly mapped reads originate in repeated regions just after the centromere.

### 3.5. Discussion

We have introduced the concept of mapping-friendly sequence reduction and shown that it improves the accuracy of the popular mapping tool `minimap2` on simulated Oxford Nanopore long reads.

We focused on reads with high mapping quality (50-60), as it is a common practice to disregard reads with low mapping quality [258, 268, 269]. However across all mapped reads ( $\text{mapq} \geq 0$ ), HPC and our MSRs have lower mapping accuracies than raw reads, consistent with the recommendation made in `minimap2` to not apply HPC to ONT data. Despite this, we newly show the benefit of using HPC (and our MSRs) with `minimap2` on ONT data when focusing on high mapping quality reads. Furthermore MSRs provide a higher fraction of high-mapq reads compared to both raw and HPC, as shown on the human and *Drosophila* genomes.

A natural future direction is to also test whether our MSRs perform well on mapping Pacific Biosciences long reads. Furthermore, it is important to highlight that our sampling of MSRs is incomplete. This is of course due to only looking at functions having  $\ell = 2$ , but also to the operational definition of RC-core-insensitive functions, and finally to taking representatives of equivalence classes. An interesting future direction would consist in exploring other families of MSRs, especially those that would include HPC and/or close variations of it.

Additionally, our analyses suggests to not perform HPC on centromeres and other repeated regions, hinting at applying sequence transformations to only some parts of the genomes. We leave this direction for future work.

### 3.6. Limitations of this study

Our proposed MSRs improve upon HPC at mapq 60, both in terms of fraction of reads mapped and error rate, on whole human and *Drosophila melanogaster* genomes. We chose these sequences because they were from organisms that we deemed different enough, however it would be interesting to verify if our proposed MSRs are still advantageous on very different organisms, e.g. more bacterial or viral genomes. This would allow us to assess the generalizability of our proposed MSRs.

We made the choice of using simulated data to be able to compute a mapping error rate. Some metrics, such as fraction of reads mapped might still be informative with regards to the mapping performance benefits of MSRs, even on real data. Evaluating the MSRs on real data might be more challenging but would offer insight into real-world usage of such pre-processing transformations.

Finally, the restrictions we imposed to define RC-core-insensitive MSRs though intuitively understandable are somewhat arbitrary, so exploring a larger search space might

## CHAPTER 3

be beneficial. Alternatively for higher order MSRs, even with our restrictions, the search spaces remain much too large to be explored exhaustively. To mitigate this problem, either further restrictions need to be found, or an alternative, optimization-based exploration method should be implemented.

### 3.7. Code availability

The scripts and pipelines used to obtain the results, as well as do the analysis and figures are available in an online repository at [github.com/lucbassel/MSR\\_discovery](https://github.com/lucbassel/MSR_discovery)

### Supplementary information

Supporting Information can be found in Appendix A

### References for chapter 3

- [4] Sergey Nurk, Sergey Koren, et al. “The complete sequence of a human genome”. In: *Science* 376.6588 (Apr. 2022). Publisher: American Association for the Advancement of Science, pp. 44–53. DOI: [10.1126/science.abj6987](https://doi.org/10.1126/science.abj6987). URL: <https://www.science.org/doi/10.1126/science.abj6987> (cit. on pp. 11, 19, 22, 71, 78).
- [97] Juliane C Dohm, Philipp Peters, et al. “Benchmarking of Long-Read Correction Methods”. In: *NAR Genomics and Bioinformatics* 2.2 (June 1, 2020). ISSN: 2631-9268. DOI: [10.1093/nargab/lqaa037](https://doi.org/10.1093/nargab/lqaa037) (cit. on pp. 24, 70).
- [108] Sergey Nurk, Brian P. Walenz, et al. “HiCanu: Accurate Assembly of Segmental Duplications, Satellites, and Allelic Variants from High-Fidelity Long Reads”. In: *Genome Research* 30.9 (Jan. 9, 2020), pp. 1291–1305. ISSN: 1088-9051, 1549-5469. DOI: [10.1101/gr.263566.120](https://doi.org/10.1101/gr.263566.120). pmid: [32801147](https://pubmed.ncbi.nlm.nih.gov/32801147/) (cit. on pp. 25, 71).
- [109] Barış Ekim, Bonnie Berger, and Rayan Chikhi. “Minimizer-Space de Bruijn Graphs: Whole-Genome Assembly of Long Reads in Minutes on a Personal Computer”. In: *Cell Systems* 12.10 (Oct. 20, 2021), 958–968.e6. ISSN: 2405-4712. DOI: [10.1016/j.cels.2021.08.009](https://doi.org/10.1016/j.cels.2021.08.009) (cit. on pp. 25, 71).
- [110] Jason R. Miller, Arthur L. Delcher, et al. “Aggressive Assembly of Pyrosequencing Reads with Mates”. In: *Bioinformatics* 24.24 (Dec. 15, 2008), pp. 2818–2824. ISSN: 1367-4803. DOI: [10.1093/bioinformatics/btn548](https://doi.org/10.1093/bioinformatics/btn548) (cit. on pp. 25, 71).
- [112] Kristoffer Sahlin and Paul Medvedev. “De Novo Clustering of Long-Read Transcriptome Data Using a Greedy, Quality Value-Based Algorithm”. In: *Journal of Computational Biology* 27.4 (Apr. 1, 2020), pp. 472–484. DOI: [10.1089/cmb.2019.0299](https://doi.org/10.1089/cmb.2019.0299) (cit. on pp. 25, 71).

## REFERENCES FOR CHAPTER 3

- [113] Kin Fai Au, Jason G. Underwood, et al. “Improving PacBio Long Read Accuracy by Short Read Alignment”. In: *PLOS ONE* 7.10 (Oct. 4, 2012), e46679. ISSN: 1932-6203. DOI: [10.1371/journal.pone.0046679](https://doi.org/10.1371/journal.pone.0046679) (cit. on pp. 25, 71).
- [115] Heng Li. “Minimap2: Pairwise Alignment for Nucleotide Sequences”. In: *Bioinformatics* 34.18 (Sept. 15, 2018), pp. 3094–3100. ISSN: 1367-4803. DOI: [10.1093/bioinformatics/bty191](https://doi.org/10.1093/bioinformatics/bty191) (cit. on pp. 25, 48, 51, 71, 78, 79, 159).
- [116] Chirag Jain, Arang Rhie, et al. “Weighted Minimizer Sampling Improves Long Read Mapping”. In: *Bioinformatics* 36 (Supplement\_1 July 1, 2020), pp. i111–i118. ISSN: 1367-4803. DOI: [10.1093/bioinformatics/btaa435](https://doi.org/10.1093/bioinformatics/btaa435) (cit. on pp. 25, 52, 81, 159).
- [123] Dan Gusfield. *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*. DOI: 10.1017/CBO9780511574931. Cambridge: Cambridge University Press, 1997. DOI: [10.1017/CBO9780511574931](https://doi.org/10.1017/CBO9780511574931). URL: <https://www.cambridge.org/core/books/algorithms-on-strings-trees-and-sequences/F0B095049C7E6EF5356F0A26686C20D3> (cit. on pp. 41, 48, 52, 70).
- [258] Timofey Prodanov and Vikas Bansal. “Sensitive Alignment Using Paralogous Sequence Variants Improves Long-Read Mapping and Variant Calling in Segmental Duplications”. In: *Nucleic Acids Research* 48.19 (Nov. 4, 2020), e114. ISSN: 0305-1048. DOI: [10.1093/nar/gkaa829](https://doi.org/10.1093/nar/gkaa829) (cit. on pp. 52, 85).
- [260] Alla Mikheenko, Andrey V Bzikadze, et al. “TandemTools: Mapping Long Reads and Assessing/Improving Assembly Quality in Extra-Long Tandem Repeats”. In: *Bioinformatics* 36 (Supplement\_1 July 1, 2020), pp. i75–i83. ISSN: 1367-4803. DOI: [10.1093/bioinformatics/btaa440](https://doi.org/10.1093/bioinformatics/btaa440) (cit. on pp. 52, 78).
- [261] Lauren Bragg, Glenn Stone, et al. “Fast, Accurate Error-Correction of Amplicon Pyrosequences Using Acacia”. In: *Nature Methods* 9.5 (5 May 2012), pp. 425–426. ISSN: 1548-7105. DOI: [10.1038/nmeth.1990](https://doi.org/10.1038/nmeth.1990) (cit. on p. 71).
- [262] Kristoffer Sahlin and Paul Medvedev. “Error Correction Enables Use of Oxford Nanopore Technology for Reference-Free Transcriptome Analysis”. In: *Nature Communications* 12.1 (1 Jan. 4, 2021), p. 2. ISSN: 2041-1723. DOI: [10.1038/s41467-020-20340-8](https://doi.org/10.1038/s41467-020-20340-8) (cit. on p. 71).
- [263] Hailin Liu, Shigang Wu, et al. “SMARTdenovo: A de Novo Assembler Using Long Noisy Reads”. In: *Gigabyte* 2021 (Mar. 8, 2021), pp. 1–9. DOI: [10.46471/gigabyte.15](https://doi.org/10.46471/gigabyte.15) (cit. on p. 71).
- [264] Ronald L. Graham, Donald Ervin Knuth, and Oren Patashnik. *Concrete Mathematics: A Foundation for Computer Science*. 2nd ed. Reading, Mass: Addison-Wesley, 1994. 657 pp. ISBN: 978-0-201-55802-9 (cit. on p. 77).
- [265] M. D. Adams, S. E. Celniker, et al. “The Genome Sequence of Drosophila Melanogaster”. In: *Science (New York, N.Y.)* 287.5461 (Mar. 24, 2000), pp. 2185–2195. ISSN: 0036-8075. DOI: [10.1126/science.287.5461.2185](https://doi.org/10.1126/science.287.5461.2185). pmid: [10731132](https://pubmed.ncbi.nlm.nih.gov/10731132/) (cit. on p. 78).

## CHAPTER 3

- [266] Chen Yang, Justin Chu, et al. “NanoSim: Nanopore Sequence Read Simulator Based on Statistical Characterization”. In: *GigaScience* 6.4 (Apr. 1, 2017). ISSN: 2047-217X. DOI: [10.1093/gigascience/gix010](https://doi.org/10.1093/gigascience/gix010) (cit. on pp. 78, 157).
- [267] Arang Rhie, Brian P. Walenz, et al. “Merqury: Reference-Free Quality, Completeness, and Phasing Assessment for Genome Assemblies”. In: *Genome Biology* 21.1 (Sept. 14, 2020), p. 245. ISSN: 1474-760X. DOI: [10.1186/s13059-020-02134-9](https://doi.org/10.1186/s13059-020-02134-9) (cit. on p. 81).
- [268] Heng Li. *New Strategies to Improve Minimap2 Alignment Accuracy*. Aug. 7, 2021. arXiv: [2108.03515 \[q-bio\]](https://arxiv.org/abs/2108.03515) (cit. on p. 85).
- [269] Heng Li, Jonathan M. Bloom, et al. “A Synthetic-Diploid Benchmark for Accurate Variant-Calling Evaluation”. In: *Nature Methods* 15.8 (8 Aug. 2018), pp. 595–597. ISSN: 1548-7105. DOI: [10.1038/s41592-018-0054-7](https://doi.org/10.1038/s41592-018-0054-7) (cit. on p. 85).

# 4. Learning from alignments

## 4.1. Alignments are a rich source of information

### 4.1.1. Pairwise alns

we can compare sequences and say if an organism, or in the case of mapping get an idea of where on the genome we are sequencing

### 4.1.2. MSA

Here we have richer

#### 4.1.2.1. Clustering

- Phylogenetic trees
- Evolutionary inference
- Protein structure prediction

#### 4.1.2.2. ML

- Alphafold
- Predict location / function
- predict characteristics i.e. resistance, ...

## 4.2. Preprocessing the alignment for machine learning

In order to do some learning we need to have the data in digestible form

### 4.2.1. Embedding the alignment

We need a way to represent, the position and the character in a sequence

## CHAPTER 4

### **4.2.1.1. Physico-chemical embeddings**

AAIndex, or other embeddings, we add extra info, but we also make a string choice when deciding what features to add

### **4.2.1.2. Generalistic categorical embeddings**

One-Hot, etc..., easily interpretable...

### **4.2.1.3. Learned embeddings**

language models, transformers, etc...

Powerful but hard to interpret what the model actually learns. i.e. “black box”

## **4.2.2. Choosing a learning target**

Of course once we have the data in digestible form we need an objective, a goal and once again a multitude

### **4.2.2.1. Regression**

Either resistance level, IC50, ...

### **4.2.2.2. Classification**

Resistant or not, compartments in the cell, ...

### **4.2.2.3. Task-based...**

end-to-end training like aligning sequences, this is harder because it requires developing a custom differentiable scoring function based on the task.

## 4.3. How to learn from ALNs

### 4.3.1. Tests and statistical learning

- correlation
- Fisher
- Multiple testing ?

### 4.3.2. Taking interactions into account

- Regressions w/ regularization
- RF
- ...

### 4.3.3. Deep Learning

- Steiner et al...
- plenty of other refs (DRMs + ML section from our minireview in Current opinions in virology 2021)



# 5. Viruses, HIV and drug resistance

## 5.1. What are viruses ?

Viruses occupy a strange place in the tree of life, with many debating if they are actually alive or not. André Lwoff gave what is probably the most fitting definition: “*viruses are viruses*” [270]. Despite this ambiguity, viruses share some common characteristics which allow us to define them as intracellular parasites [271]:

1. Viruses have some type of genetic information, contained in either DNA or RNA
2. This genetic information is protected by some form of envelope
3. They use the cellular machinery of host cells to make copies of themselves.

While we all know that viruses can be pathogenic and dangerous (the recent example of SARS-CoV2 springs to mind), that is not necessarily the case and some viruses like GBV-C [272] and certain strains of H5N1 *Influenza* [273] are non pathogenic and essentially harmless.

Viruses have been discovered for all three domains of life: Eukaryota, Bacteria and Archaea. In Eukaryota many viruses have been discovered for animals (both vertebrate [274] and invertebrate [275]), plants [276], protozoa [277], chromista [278] and even fungi [279]. Bacterial viruses known as phages have been known to exist since the beginning of the 20th century [280, 281]. These bacteriophages are being considered as a therapeutic alternative to antibiotics [282, 283] which could help with multi-drug-resistant bacterial pathogens. Archaea are also known to have their own viral infections [284, 285].

Strangely even viruses of viruses seem to exist such as the plant satellite virus [286, 287] or hepatitis delta virus [288, 289]. These “*viroids*” do not infect viral hosts *per se* but they cannot replicate on their own. Replication must happen during co-infection with a larger virus. More recently, true viruses of viruses called virophages have been discovered. These virophages like sputnik [290] or zamilon [291] specifically infect giant viruses.

There is a huge diversity of viruses affecting all types of life, and new viruses are being discovered all the time [292]. This diversity hints at a rich and long evolutionary history. When and where viruses originated is still under study [293, 294] and we might never know how they emerged, it is however believed that they may have played an important role in the emergence of eukaryotic cells [295]. This co-evolution between virus and host cell shows a strong link between the two organisms and some parts of the human genome

## CHAPTER 5

are likely of ancient viral origin [296, 297]. It has been estimated that 1% to 8% of the human genome are endogenous retroviral sequences [298, 299].

The rich diversity of viruses is reflected in the variety of genetic information support, replication strategy, physical and genomic size, as well as shape. The differences in genetic information support and replication strategy form the basis of the Baltimore virus classification system [300], still used today [301] to classify virus lineages.

As stated above all viruses have some genetic information, this information is stored either as DNA or as RNA, which is the molecule of choice for 70% of human pathogenic viruses [302] (*HIV and SARS-CoV 2 are RNA viruses*).

For DNA viruses, the molecule can be double-stranded as for *Herpesvirus* [303, 304], single-stranded like in the case of *Papillomavirus* [305] or even circular in the case of the Hepatitis B virus [306]. This molecular diversity is also present in RNA viruses where the RNA molecule can be double-stranded like for *Rotavirus* [307], or single-stranded. Furthermore, for single-stranded RNA viruses the strand can either be positive (*i.e. can be directly translated into a protein*) like Hepatitis C virus [308] or *Poliovirus* [309, 310]; conversely there are negative-strand RNA viruses, for which the complementary strand of RNA must be synthesized before translation into a protein, such as the Influenza or Measles viruses [311].

This diversity in genetic information support implies a necessary diversity in replication strategy. The main replication strategies are as follows [312]:

- The RNA molecule is directly copied as RNA. This is the strategy followed by single-stranded RNA coronaviruses [313], Dengue viruses [314] or Hepatitis C virus [315].
- The DNA molecule is directly replicated as DNA. this can happen for both single-stranded DNA viruses like *Papillomavirus* [316] and double-stranded DNA viruses like Herpes simplex virus [317].
- The DNA molecule is replicated by going through an RNA intermediary like Hepatitis B virus [318].
- The RNA molecule is replicated by going through a DNA intermediary. This strategy is used by retroviruses that integrate this viral DNA intermediary into the host DNA, like HIV-1 (see Section 5.2.2).

Finally, the genetic diversity of viruses is reflected in their physical characteristics: viruses come in all shapes and sizes. Physical size range from 17nm for plant satellite viruses [319] to the giant 400nm *Mimivirus* [320]. Genomic size is also quite variable, there is a stark contrast between the 860 bp *Circovirus SFBeef* and the 2.5 Mbp *Pandoravirus salinus* genomes [321]. Viruses come in a variety of shapes [322]: icosahedral for HIV, helical for the tobacco mosaic virus or a distinctive head-tail shape for bacteriophages.

Although there are a large number of viruses, and many of them are of great importance for human health, we will now focus on one virus of particular importance: Human Immunodeficiency Virus otherwise known as HIV.

## 5.2. Getting to know HIV

### 5.2.1. Quick Presentation of HIV

HIV is a single-stranded RNA retrovirus that is responsible for the Acquired Immune Deficiency Syndrome (AIDS) pandemic that has been around for the last couple decades. This virus is transmitted through sexual contact or through blood. Sexual activity is the largest transmission factor followed by intravenous drug use [323, 324].

The HIV infects cells of the host immune system, specifically CD4 T-cell lymphocytes and destroys them due to its replication process [325]. CD4 T-cells are an essential part of the immune system response helping fight against infection in humans. An HIV infection typically starts with an asymptomatic phase that can last years, followed by a growth in viral replication leading to a decrease in CD4 cells which progresses into AIDS [326]. During AIDS, when the CD4 cell count is low enough opportunistic diseases such as pneumonia or tuberculosis [327] can easily infect the host, leading to death when the immune system is weak enough.

The HIV/AIDS is one of the deadliest pandemics in history, estimated to have led to the death of 36 million people [328]. In 2010 [329] approximately 33 million people were infected with HIV, 2.6 million of which were due to new infections, and 1.8 million died of AIDS. Most of the new infections happened in economically developing regions of the world, 70% of them coming from sub-Saharan Africa [329]. As of 2020, these numbers have decreased with “only” 1.5 million new infections and 680,000 AIDS deaths, which is encouraging from a public health perspective.

The HIV-1 virus was discovered simultaneously in 1983 by Françoise Barré-Sinoussi, Luc Montagnier [330] and Robert Gallo [331]. There exists a second HIV-2 virus discovered shortly after HIV-1 [332], it is however less transmissible than HIV-1 which is largely responsible for the global HIV/AIDS pandemic [333]. In Africa in 2006, HIV-1 infections were rising where HIV-2 were declining [334].

While both viruses are of zoonotic origin, from transmissions of Simian Immunodeficiency Virus (SIV) from primates to humans. HIV-1 most likely originates from an SIV present in chimpanzees [335, 336, 337], and HIV-2 from an SIV present in Sooty mangabeys [338, 339, 340].

Several independent such transmissions have resulted in 4 lineages of HIV-1 labeled groups M, N, O and P [341] (similarly HIV-2 is split into groups A to H also resulting from independent zoonotic transmissions). While Groups N and P have been identified in

## CHAPTER 5

only a handful of individuals in Cameroon , and group O is estimated to a few thousand cases in western Africa, the majority of the pandemic is due to viruses from group M.

The most recent common ancestor, *i.e. the putative virus that founded group M*, is estimated to have originated in what is now the Democratic Republic of Congo [342, 343, 344] at some point between 1910 and 1931 [345, 342, 346].

Group M is further subdivided into 9 subtypes each with distinct genetic characteristics, labeled A to K [347, 337]. Like in many viruses [348], when 2 genetically different strains of HIV co-infect a single host there is a risk genetic recombination leading to a new strain [349]. During recombination, a new genome is formed from parts of the original genomes. This can lead to new strains that can spread and form lineages of their own. HIV strains resulting from recombination are called Circulating Recombinant Forms (CRFs). There are currently 118 identified HIV-1 CRFs in the Los Alamos National Laboratory HIV sequence database [350] (1 for HIV-2). There also exist many unique recombinant forms (URFs). Recombination can be particularly bothersome, complicating evolutionary analyses [351], or facilitating drug resistance and hindering vaccine development [352].

While subtype C represented almost half of global infections from 2004 to 2007, subtype B is the majority subtype in richer countries of North America and Western Europe [353] where sequencing efforts are more common. This accounts for an over-representation of subtype B sequences in public databases such as the Los Alamos sequence database where 54% of sequences are of the B subtype and only 15% are C [354].

### 5.2.2. The replication cycle of HIV

The viruses replication cycle and its immune-cell host specificity are what makes it particularly dangerous. This replication cycle can broadly be categorized into 9 separate steps [355, 356] shown in Figure 5.1.

1. An HIV virion binds itself to the CD4 host cell through membrane proteins.
2. The virion envelope and host cell membrane fuse together allowing the viral genetic material and proteins to enter the host cell.
3. The viral RNA is reverse-transcribed into viral DNA
4. The viral DNA is integrated into the host cell genome
5. The integrated viral DNA is transcribed by the host cell machinery into multiple copies of viral RNA
6. The viral RNA is translated into immature viral polyproteins
7. The viral polyproteins are cleaved to form individual viral proteins.
8. The newly synthesized viral RNA and viral proteins gather around the host-cell membrane which starts budding to create a new virion

## 5.2. GETTING TO KNOW HIV

9. Once the budding is complete, the virion is released from the host cell and matures before being able to infect other CD4 cells and replicate again.

The successive infection of CD4 cells by HIV virions leads to cellular death due to inflammatory response and/or activation of apoptosis [357, 358]. The gradual depletion of CD4 cells in the infected individual's body lead to the suppression of the immune system, eventually leading to AIDS.

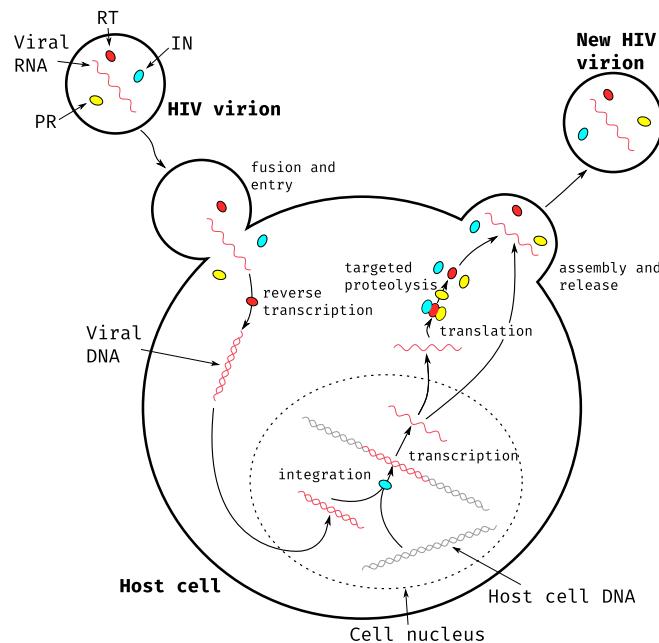


Figure 5.1.: Main steps of HIV-1 replication cycle

The HIV virion contains viral RNA and three essential proteins: Reverse Transcriptase (RT) represented in red, Integrase (IN) represented in cyan and Protease (PR) represented in yellow.

### 5.2.3. Genetics of HIV

The replication cycle described in Section 5.2.2 is made possible by the 15 proteins of HIV. These proteins are coded by 9 separate genes [359]. An overview of the HIV proteins, their structure and localization within the viral particle can be seen in Figure 5.2.

The HIV genome is made up of three main genes each coding for polyproteins and six genes coding for proteins with regulatory and accessory roles. The three polyproteins correspond to long chains of amino acids which are subsequently cleaved at specific positions to result in separate viral proteins.

## CHAPTER 5

The *gag* (“group-specific-antigen”) gene codes for the Gag polyprotein which, once cleaved, results in four proteins with mainly structural roles:

- The Matrix protein (MA or p17) lines the internal surface if the virion membrane maintaining the shape and structural integrity of the virion.
- The Capsid protein (CA or p24) forms an inner core (the capsid) inside the virion around the viral RNA. It helps protect the viral genetic information.
- The Nucleocapsid protein (NC or p7) binds with the viral RNA inside the capsid, stabilizing the molecule and further protecting the genetic information.
- The p6 protein is a small, largely unstructured protein [360] that is suspected of playing a role in virion budding and release from the host cell at the end of the replication cycle [361, 362].

The *pol* (“polymerase”) gene coding for the Pol polyprotein. After cleaving this results in the three essential viral enzymes at the heart of the replication cycle:

- The Protease (PR) is responsible for cleaving the Gag, Pol and Env polyproteins to get the individual viral proteins. Without it the individual viral proteins cannot come into being and therefore cannot function, stopping viral replication.
- The Reverse Transcriptase (RT or p51/p66) is responsible for synthesizing viral DNA from the viral RNA template contained in the virion. This is the first step in hijacking the cellular machinery for replication. Without viral DNA, HIV replication is impossible.
- The Integrase (IN) is responsible for integrating the viral DNA produced by RT in to the host cell DNA. Once the viral DNA is inside the host genome it can be transcribed and then translated (as described in Section 1.1) to produce new copies of the viral RNA and proteins. Without this integration step the viral genetic information cannot be expressed and the replication cycle is stopped.

The *env* (“envelope”) gene code for the Env, the third and last polyprotein. The two resulting proteins coat the membrane of the virion and are responsible for binding with the CD4 host cells.

- The Surface protein (SU or gp120) binds to receptors on the surface of CD4 cells and allow the virion to attach itself to the host cell [363]. It also enables membrane fusion, the essential first step in the viral replication cycle [364].
- The Transmembrane protein (TM or gp41) anchors SU into the virion membrane.

The 6 remaining genes all code for single proteins. Two of these have essential regulatory roles and the remaining four have accessory roles.

The *tat* (“trans-activator of transcription”) gene codes for Tat, the first essential regulatory protein. Tat activates and promotes transcription leading to more numerous and

longer copies of the viral RNA [365]. The *rev* (for “regulator of virion”) gene codes for Rev, the second essential regulatory protein. Rev helps transcribed viral RNA exist the nucleus of the host cell in order to be translated to viral proteins or be packages in new, budding virions [366].

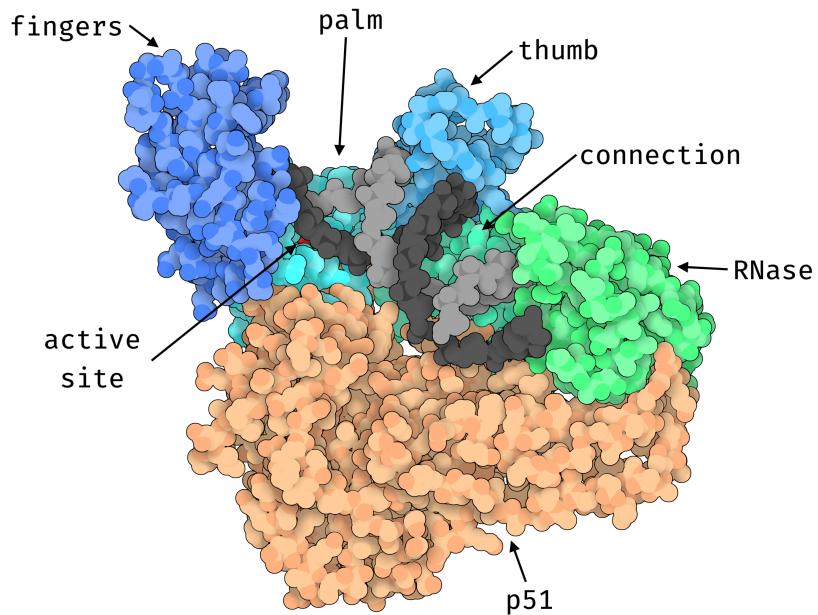
The remaining four accessory genes are as follows: *nef* (“negative regulatory factor”) code for the Nef protein which prevents the production of the CD4 cellular defense proteins increasing infectivity [367]; *vif* (“viral infectivity factor”) codes for the Vif protein which also increases viral infectivity [368]; *vpu* (“viral protein U”) codes for Vpu which likely helps during release of new virions [369, 368] as well as preventing production of CD4 in the host cell, it is not believed to be present in the mature virion as it binds to host cellular membranes [370]; *vpr* (“viral protein R”) likely helps viral DNA enter the host cell nucleus and prevents the natural host cell reproduction cycle [371].

The existence of a 10th HIV-1 gene was suggested in 1988 [372], overlapping the *env* gene and coding for proteins on the other strand of viral DNA than the other genes. This putative gene was named *asp* (“antisense protein”) and Asp transcripts were isolated during an HIV-1 infection in 2002 [373]. The function of this protein is still unknown but it has been shown to have a strong evolutionary correlation with HIV-1 group M responsible for the pandemic [374]. Although this Asp protein is still a source of debate, it is under active research [375].

#### 5.2.4. Main protein mechanisms in replication

All 3 are cleaved and matured from the pol polyprotein

- RT: [377], c.f. Figure 5.3
  - 2 subunits: p66 and p51 (i.e. same chain but 1 is cleaved). p66=functional, p51=structural mainly (except for NNIBP).
  - Fold to form a canal -> fingers/palm/thumb with the polymerase active site in the palm
  - RNA strand and DNA go through the canal and are synthesized in the palm
  - RNase domain at the other end of p66 to cleave the RNA strand off of the DNA once it's RTed.
  - Once the RNA has been RTed the RT does another pass to synthesize the complementary DNA strand, this time the RNase is not active.



**Figure 5.3.: 3D structure of HIV-1 Reverse-transcriptase**

the different domains of the p66 subunit are labeled and shown in different shades of blue and green. The structural p51 subunit is shown in orange. The RNA template is shown in dark gray and the newly synthesized DNA strand in light gray. The polymerase active site is shown in red although mostly hidden by the RNA template. The 3D visualization was produced with Illustrate [378] using the 2hmi PDB structure.

- IN: It pre-processes the ends of the viral DNA, cleaves the host DNA and binds the viral DNA to the host DNA (i.e. strand transfer). [379, 380, 381]. Multimer -> chains link up. Each chain has:
  - Core catalytic domain: Binds with viral DNA and performs the integration
  - C-terminal domain: Binds with any DNA, adding stability
  - N-terminal domain: Binds with  $Zn^{2+}$  and helps multimerisation (i.e integrase stability)
- PR: cleaves gag and pol polyproteins to produce individual structural and regulatory proteins. Symmetric dimer structure [382, 383], 2 identical chains.

### 5.3. DRUG RESISTANCE IN HIV

- Ionization of an Aspartate residue in each active site produces acid-base reaction that cleaves precursor [384]. (active site)
- Flaps are flexible parts of PR that can open and close [385], probably controls access of substrates to active site [386]

## 5.3. Drug resistance in HIV

When on ART, virus evolves under selective pressure and develops resistance -> treatment failure.

### 5.3.1. What is ART ?

ART is very important in reducing the viral load and therefore reducing transmission [387], however it does not “cure” HIV, it controls it drastically reducing mortality and morbidity [388].

Mostly target the RT and PR proteins

History [389, 390]:

- azidothymidine (AZT), an RTI, was shown to be effective against HIV in 1987 [391], only 3 years later resistant strains of the virus were circulating [392], because it was a monotherapy.
- Between 1988 and 1995 four more RTIs were approved by the FDA and a combination therapy was shown to be effective [393]
- In 2008 there were 30 antiretroviral drugs have approved by the FDA [394], of which 12 were RTI, 10 PIs and 5 combination drugs, often available as a single pill.
- In 2019 there are 11 licensed single pill regimens in Europe [395].

Other drugs:

- entry inhibitors: have been shown to be effective 2 have been approved in the USA [396]
- integrase inhibitors (INI): recent focus on development, 5 approved drugs since 2007 [397]
- Pre-exposure prophylaxis (PrEP), putting uninfected but at risk populations on ART before any known exposure. Has been shown effectively lower risk of infection in several studies [398, 399, 400]. When adherence is maintained studies have shown a infection risk between 44% and 100% [401].  
In 2022 there are 1 PrEP regimen approved for use in Europe [402].

## CHAPTER 5

- Action of ART drugs:
  - RTI:
    - \* NRTI: nucleoside analogue RTI work by binding with the active site of RT blocking DNA polymerization -> chain terminating inhibitors
    - \* NNRTI: non-nucleoside analogue RTI, bind to other a hydrophobic pocket behind the active site the RT, changing its flexibility and rendering it inactive [403]
  - PI: bind to active site of protease to hinder action [404]
  - Entry inhibitors are being found [405, 406]. enfuvirtide binds to gp41 and prevents fusion [407]
  - INI: prevent strand transfer where viral DNA is first linked to host DNA [408]

However Drug resistance has emerged for most treatment types even PrEP [409, 410] but seem to be rare and due to pre-treatment but unknown HIV infection [411]

### 5.3.2. How can HIV evade ART ?

review [412, 413]

- NRTI [414, 415]:
  - Prevent incorporation of NRTI, like M184V [416]
  - Remove incorporated NRTI: TAMs [417, 418]
- NNRTI: mutations limiting access to the binding pocket like Y181C or changing the shape of it like K103N [419]. dangerous because a single mutation produces viruses that are both fit and highly resistant [412] + cross-resistance
- PI: make active site bigger reducing affinity to inhibitors but also to natural substrate [420], also change flap conformation to deny access to active site to PI [421].
- INI: reduce affinity to inhibitors [422]

### 5.3.3. Consequences on global health

- Transmission of drug resistant viruses, particularly between treatment-naive individuals which is the main transmission mode in the UK [423] and Switzerland [424], creating long lasting drug resistance reservoirs in naive population.
- Dangerous because there can be therapeutic failure of first line regimen. [425]
- Fitness is lower for some DRMs [426, 427] which means there is evolutionary pressure for these mutations to disappear in the absence of treatment

- the reversion to WT can last a long time with a wide range of median loss times from 1 to 13 years [428], depending on the type of mutation (NRTI > NNRTI).

### 5.3.4. Finding DRMs

(This sections build upon a review we did [429])

- Finding and categorizing DRMs is an important task.
- Sequence databases useful in discovery:
  - UK-CHIC [430] and the uk HIV drug resistance database (<https://www.hivrdb.org.uk/>)
  - Swiss cohort study (<https://www.shcs.ch/>)
  - PANGEA [431]
  - Stanford HIV Drug resistance database (<https://hivdb.stanford.edu/>) with sequences [32]
- Databases of mutations are kept [432] with regularly update lists of important DRMs [433, 434] and tools to help practitioners detect and interpret resistance [435].
- stat tests
  - multiple testing
  - phylogenetic correlation
- assays
- novel approaches
  - deep learning
  - ...

## References for chapter 5

- [32] Soo-Yon Rhee, Matthew J. Gonzales, et al. “Human immunodeficiency virus reverse transcriptase and protease sequence database”. In: *Nucleic Acids Research* 31.1 (Jan. 1, 2003), pp. 298–303. DOI: [10.1093/nar/gkg100](https://doi.org/10.1093/nar/gkg100). URL: <https://academic.oup.com/nar/article/31/1/298/2401450> (cit. on pp. 14, 103).
- [270] A. Lwoff. “The concept of virus”. In: *Journal of General Microbiology* 17.2 (Oct. 1957). tex.ids= lwoffConceptVirus1957a PMID: 13481308, pp. 239–253. DOI: [10.1099/00221287-17-2-239](https://doi.org/10.1099/00221287-17-2-239) (cit. on p. 93).
- [271] Philip D Minor. “Viruses”. In: \_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/9780470015902.a0000441.pub3>. John Wiley & Sons, Ltd, 2014. DOI: [10.1002/9780470015902.a0000441.pub3](https://doi.org/10.1002/9780470015902.a0000441.pub3). URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470015902.a0000441.pub3> (cit. on p. 93).

## CHAPTER 5

- [272] Jack T. Stapleton, Steven Foug, et al. “The GB viruses: a review and proposed classification of GBV-A, GBV-C (HGV), and GBV-D in genus Pegivirus within the family Flaviviridae”. In: *The Journal of General Virology* 92.Pt 2 (Feb. 2011). PMID: 21084497 PMCID: PMC3081076, pp. 233–246. doi: [10.1099/vir.0.027490-0](https://doi.org/10.1099/vir.0.027490-0). URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3081076/> (cit. on p. 93).
- [273] Naoki Yamamoto, Yoshihiro Sakoda, et al. “Characterization of a non-pathogenic H5N1 influenza virus isolated from a migratory duck flying from Siberia in Hokkaido, Japan, in October 2009”. In: *Virology Journal* 8.1 (Feb. 11, 2011), p. 65. doi: [10.1186/1743-422X-8-65](https://doi.org/10.1186/1743-422X-8-65). URL: <https://doi.org/10.1186/1743-422X-8-65> (cit. on p. 93).
- [274] Mang Shi, Xian-Dan Lin, et al. “The evolutionary history of vertebrate RNA viruses”. In: *Nature* 556.7700 (Apr. 2018). Number: 7700 Publisher: Nature Publishing Group, pp. 197–202. doi: [10.1038/s41586-018-0012-7](https://doi.org/10.1038/s41586-018-0012-7). URL: <https://www.nature.com/articles/s41586-018-0012-7> (cit. on p. 93).
- [275] Jean R. Adams and Jean R. Bonami, eds. *Atlas of Invertebrate Viruses*. DOI: 10.1201/9781315149929. Boca Raton: CRC Press, Aug. 31, 2017. doi: [10.1201/9781315149929](https://doi.org/10.1201/9781315149929) (cit. on p. 93).
- [276] Pierre Lefevre, Darren P. Martin, et al. “Evolution and ecology of plant viruses”. In: *Nature Reviews Microbiology* 17.10 (Oct. 2019). Number: 10 Publisher: Nature Publishing Group, pp. 632–644. doi: [10.1038/s41579-019-0232-3](https://doi.org/10.1038/s41579-019-0232-3). URL: <https://www.nature.com/articles/s41579-019-0232-3> (cit. on p. 93).
- [277] A. L. Wang and C. C. Wang. “Viruses of parasitic protozoa”. In: *Parasitology Today* 7.4 (Jan. 1, 1991), pp. 76–80. doi: [10.1016/0169-4758\(91\)90198-W](https://doi.org/10.1016/0169-4758(91)90198-W). URL: <https://www.sciencedirect.com/science/article/pii/016947589190198W> (cit. on p. 93).
- [278] Gustavo Fermin, Sudeshna Mazumdar-Leighton, and Paula Tenant. “Viruses of prokaryotes, protozoa, fungi, and chromista”. In: Publisher: Academic Press. 2018, p. 217 (cit. on p. 93).
- [279] Suvi Sutela, Anna Poimala, and Eeva J Vainio. “Viruses of fungi and oomycetes in the soil environment”. In: *FEMS Microbiology Ecology* 95.9 (Sept. 1, 2019), fiz119. doi: [10.1093/femsec/fiz119](https://doi.org/10.1093/femsec/fiz119). URL: <https://doi.org/10.1093/femsec/fiz119> (cit. on p. 93).
- [280] F. W. Twort. “AN INVESTIGATION ON THE NATURE OF ULTRAMICROSCOPIC VIRUSES.” In: *The Lancet* 186.4814 (Dec. 4, 1915). Publisher: Elsevier, pp. 1241–1243. doi: [10.1016/S0140-6736\(01\)20383-3](https://doi.org/10.1016/S0140-6736(01)20383-3). URL: [https://www.thelancet.com/journals/lancet/article/PIIS0140-6736\(01\)20383-3/fulltext](https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(01)20383-3/fulltext) (cit. on p. 93).

## REFERENCES FOR CHAPTER 5

- [281] M. Delbrock. “Bacterial Viruses or Bacteriophages”. In: *Biological Reviews* 21.1 (1946). \_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1469-185X.1946.tb00451.x>. pp. 30–40. DOI: [10.1111/j.1469-185X.1946.tb00451.x](https://doi.org/10.1111/j.1469-185X.1946.tb00451.x). URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-185X.1946.tb00451.x> (cit. on p. 93).
- [282] Jason R Clark and John B March. “Bacterial viruses as human vaccines?” In: *Expert Review of Vaccines* 3.4 (Aug. 1, 2004). Publisher: Taylor & Francis \_eprint: <https://doi.org/10.1586/14760584.3.4.463> PMID: 15270651, pp. 463–476. DOI: [10.1586/14760584.3.4.463](https://doi.org/10.1586/14760584.3.4.463). URL: <https://doi.org/10.1586/14760584.3.4.463> (cit. on p. 93).
- [283] H E van Kan-Davelaar, J C M van Hest, et al. “Using viruses as nanomedicines”. In: *British Journal of Pharmacology* 171.17 (2014). \_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/bph.12662> pp. 4001–4009. DOI: [10.1111/bph.12662](https://doi.org/10.1111/bph.12662). URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/bph.12662> (cit. on p. 93).
- [284] David Prangishvili, Tamara Basta, et al. “Viruses of the Archaea”. In: \_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/9780470015902.a0000774.pub3> DOI: [10.1002/9780470015902.a0000774.pub3](https://doi.org/10.1002/9780470015902.a0000774.pub3). John Wiley & Sons, Ltd, 2016, pp. 1–9. DOI: [10.1002/9780470015902.a0000774.pub3](https://doi.org/10.1002/9780470015902.a0000774.pub3). URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470015902.a0000774.pub3> (cit. on p. 93).
- [285] David Prangishvili, Patrick Forterre, and Roger A. Garrett. “Viruses of the Archaea: a unifying view”. In: *Nature Reviews Microbiology* 4.11 (Nov. 2006). Number: 11 Publisher: Nature Publishing Group, pp. 837–848. DOI: [10.1038/nrmicro1527](https://doi.org/10.1038/nrmicro1527). URL: <https://www.nature.com/articles/nrmicro1527> (cit. on p. 93).
- [286] R.I.B. Francki. “Plant virus satellites”. In: *ANNUAL REVIEW OF MICROBIOLOGY RESEARCH* (). tex.ids= franckiPlantVirusSatellites, franckiPlantVirus-Satellitesa tex.recid: 10193 place: 1985. URL: <http://worldveg.tind.io/record/10193> (cit. on p. 93).
- [287] Ping Xu and Marilyn J Roossinck. “Plant Virus Satellites”. In: \_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/9780470015902.a0000771.pub2> DOI: [10.1002/9780470015902.a0000771.pub2](https://doi.org/10.1002/9780470015902.a0000771.pub2). John Wiley & Sons, Ltd, 2011. DOI: [10.1002/9780470015902.a0000771.pub2](https://doi.org/10.1002/9780470015902.a0000771.pub2). URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470015902.a0000771.pub2> (cit. on p. 93).
- [288] M. M. Lai. “The molecular biology of hepatitis delta virus”. In: *Annual review of biochemistry* 64 (Jan. 1, 1995). PMID: 7574482, pp. 259–286. DOI: [10.1146/annurev.bi.64.070195.001355](https://doi.org/10.1146/annurev.bi.64.070195.001355). URL: <https://doi.org/10.1146/annurev.bi.64.070195.001355> (cit. on p. 93).

## CHAPTER 5

- [289] Sarah A Hughes, Heiner Wedemeyer, and Phillip M Harrison. “Hepatitis delta virus”. In: *The Lancet* 378.9785 (July 2, 2011), pp. 73–85. DOI: [10.1016/S0140-6736\(10\)61931-9](https://doi.org/10.1016/S0140-6736(10)61931-9). URL: <https://www.sciencedirect.com/science/article/pii/S0140673610619319> (cit. on p. 93).
- [290] Christelle Desnues, Mickaël Boyer, and Didier Raoult. “Chapter 3 - Sputnik, a Virophage Infecting the Viral Domain of Life”. In: ed. by Małgorzata Łobocka and Wacław T. Szybalski. Vol. 82. Bacteriophages, Part A. DOI: 10.1016/B978-0-12-394621-8.00013-3. Academic Press, Jan. 1, 2012, pp. 63–89. DOI: [10.1016/B978-0-12-394621-8.00013-3](https://doi.org/10.1016/B978-0-12-394621-8.00013-3). URL: <https://www.sciencedirect.com/science/article/pii/B9780123946218000133> (cit. on p. 93).
- [291] Morgan Gaia, Samia Benamar, et al. “Zamilon, a Novel Virophage with Mimiviridae Host Specificity”. In: *PLOS ONE* 9.4 (Apr. 18, 2014). Publisher: Public Library of Science, e94923. DOI: [10.1371/journal.pone.0094923](https://doi.org/10.1371/journal.pone.0094923). URL: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0094923> (cit. on p. 93).
- [292] Robert C. Edgar, Jeff Taylor, et al. “Petabase-scale sequence alignment catalyses viral discovery”. In: *Nature* 602.7895 (Feb. 2022). Number: 7895 Publisher: Nature Publishing Group, pp. 142–147. DOI: [10.1038/s41586-021-04332-2](https://doi.org/10.1038/s41586-021-04332-2). URL: <https://www.nature.com/articles/s41586-021-04332-2> (cit. on p. 93).
- [293] Arshan Nasir, Ethan Romero-Severson, and Jean-Michel Claverie. “Investigating the Concept and Origin of Viruses”. In: *Trends in Microbiology* 28.12 (Dec. 1, 2020), pp. 959–967. DOI: [10.1016/j.tim.2020.08.003](https://doi.org/10.1016/j.tim.2020.08.003). URL: <https://www.sciencedirect.com/science/article/pii/S0966842X20302304> (cit. on p. 93).
- [294] Patrick Forterre and David Prangishvili. “The origin of viruses”. In: *Research in Microbiology*. Special issue on the origin of life and microbiology 160.7 (Sept. 1, 2009), pp. 466–472. DOI: [10.1016/j.resmic.2009.07.008](https://doi.org/10.1016/j.resmic.2009.07.008). URL: <https://www.sciencedirect.com/science/article/pii/S0923250809001065> (cit. on p. 93).
- [295] Patrick Forterre. “The origin of viruses and their possible roles in major evolutionary transitions”. In: *Virus Research*. Comparative Genomics and Evolution of Complex Viruses 117.1 (Apr. 1, 2006), pp. 5–16. DOI: [10.1016/j.virusres.2006.01.010](https://doi.org/10.1016/j.virusres.2006.01.010). URL: <https://www.sciencedirect.com/science/article/pii/S0168170206000293> (cit. on p. 93).
- [296] JD Boeke and JP Stoye. “Retrotransposons, Endogenous Retroviruses, and the Evolution of Retroelement”. In: (1997). PMID: 21433351. URL: <https://www.ncbi.nlm.nih.gov/books/NBK19468> (cit. on p. 94).
- [297] Shohei Kojima, Kohei Yoshikawa, et al. “Virus-like insertions with sequence signatures similar to those of endogenous nonretroviral RNA viruses in the human genome”. In: *Proceedings of the National Academy of Sciences* 118.5 (Feb. 2, 2021). Publisher: Proceedings of the National Academy of Sciences, e2010758118. DOI: [10.1073/pnas.2010758118](https://doi.org/10.1073/pnas.2010758118). URL: <https://www.pnas.org/doi/10.1073/pnas.2010758118> (cit. on p. 94).

- [298] R Löwer, J Löwer, and R Kurth. “The viruses in all of us: characteristics and biological significance of human endogenous retrovirus sequences.” In: *Proceedings of the National Academy of Sciences* 93.11 (1996). tex.eprint: <https://www.pnas.org/doi/pdf/10.1073/pnas.93.11.5177>, pp. 5177–5184. DOI: [10.1073/pnas.93.11.5177](https://doi.org/10.1073/pnas.93.11.5177). URL: <https://www.pnas.org/doi/abs/10.1073/pnas.93.11.5177> (cit. on p. 94).
- [299] David J. Griffiths. “Endogenous retroviruses in the human genome sequence”. In: *Genome Biology* 2.6 (June 5, 2001), reviews1017.1. DOI: [10.1186/gb-2001-2-6-reviews1017](https://doi.org/10.1186/gb-2001-2-6-reviews1017). URL: <https://doi.org/10.1186/gb-2001-2-6-reviews1017> (cit. on p. 94).
- [300] D. Baltimore. “Expression of animal virus genomes”. In: *Bacteriological Reviews* 35.3 (Sept. 1971). PMID: 4329869 PMCID: PMC378387, pp. 235–241. DOI: [10.1128/br.35.3.235-241.1971](https://doi.org/10.1128/br.35.3.235-241.1971) (cit. on p. 94).
- [301] Eugene V. Koonin, Mart Krupovic, and Vadim I. Agol. “The Baltimore Classification of Viruses 50 Years Later: How Does It Stand in the Light of Virus Evolution?” In: *Microbiology and Molecular Biology Reviews* 85.3 (Aug. 18, 2021). Publisher: American Society for Microbiology, e00053–21. DOI: [10.1128/MMBR.00053-21](https://doi.org/10.1128/MMBR.00053-21). URL: <https://journals.asm.org/doi/full/10.1128/MMBR.00053-21> (cit. on p. 94).
- [302] Esteban Domingo and Celia Perales. “RNA Virus Genomes”. In: \_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/9780470015902.a0001488.pub3> DOI: [10.1002/9780470015902.a0001488.pub3](https://doi.org/10.1002/9780470015902.a0001488.pub3). John Wiley & Sons, Ltd, 2018, pp. 1–12. DOI: [10.1002/9780470015902.a0001488.pub3](https://doi.org/10.1002/9780470015902.a0001488.pub3). URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470015902.a0001488.pub3> (cit. on p. 94).
- [303] Duncan J. McGeoch, Frazer J. Rixon, and Andrew J. Davison. “Topics in herpesvirus genomics and evolution”. In: *Virus Research*. Comparative Genomics and Evolution of Complex Viruses 117.1 (Apr. 1, 2006), pp. 90–104. DOI: [10.1016/j.virusres.2006.01.002](https://doi.org/10.1016/j.virusres.2006.01.002). URL: <https://www.sciencedirect.com/science/article/pii/S0168170206000219> (cit. on p. 94).
- [304] Paul Boehmer and Amitabh Nimonkar. “Herpes Virus Replication”. In: *IUBMB Life* 55.1 (2003). \_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1080/1521654031000070645>, pp. 13–22. DOI: [10.1080/1521654031000070645](https://doi.org/10.1080/1521654031000070645). URL: <https://onlinelibrary.wiley.com/doi/abs/10.1080/1521654031000070645> (cit. on p. 94).
- [305] Mathijs H Brentjens, Kimberly A Yeung-Yue, et al. “Human papillomavirus: a review”. In: *Dermatologic Clinics* 20.2 (Apr. 1, 2002), pp. 315–331. DOI: [10.1016/S0733-8635\(01\)00028-6](https://doi.org/10.1016/S0733-8635(01)00028-6). URL: <https://www.sciencedirect.com/science/article/pii/S0733863501000286> (cit. on p. 94).
- [306] Alan Kay and Fabien Zoulim. “Hepatitis B virus genetic variability and evolution”. In: *Virus Research*. Replicative and Evolutionary Aspects of Hepatitis Viruses 127.2 (Aug. 1, 2007), pp. 164–176. DOI: [10.1016/j.virusres.2007.02.021](https://doi.org/10.1016/j.virusres.2007.02.021). URL:

## CHAPTER 5

- <https://www.sciencedirect.com/science/article/pii/S0168170207000615> (cit. on p. 94).
- [307] U D Parashar, J S Bresee, et al. “Rotavirus.” In: *Emerging Infectious Diseases* 4.4 (1998). PMID: 9866732 PMCID: PMC2640254, pp. 561–570. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2640254/> (cit. on p. 94).
- [308] Peter Simmonds. “Variability of hepatitis C virus”. In: *Hepatology* 21.2 (Feb. 1, 1995), pp. 570–583. DOI: [10.1016/0270-9139\(95\)90121-3](https://doi.org/10.1016/0270-9139(95)90121-3). URL: <https://www.sciencedirect.com/science/article/pii/0270913995901213> (cit. on p. 94).
- [309] Eckard Wimmer, Christopher U. T. Hellen, and Xuemei Cao. “Genetics of poliovirus”. In: *Annual Review of Genetics* 27 (Jan. 1, 1993). Publisher: Annual Reviews, Inc., pp. 353–437. URL: <https://go.gale.com/ps/i.do?p=AONE&sw=w&issn=00664197&v=2.1&it=r&id=GALE%7CA15539705&sid=googleScholar&linkaccess=abs> (cit. on p. 94).
- [310] Vincent R. Racaniello. “One hundred years of poliovirus pathogenesis”. In: *Virology*. Virology 50th Anniversary Issue 344.1 (Jan. 5, 2006), pp. 9–16. DOI: [10.1016/j.virol.2005.09.015](https://doi.org/10.1016/j.virol.2005.09.015). URL: <https://www.sciencedirect.com/science/article/pii/S0042682205005830> (cit. on p. 94).
- [311] P Palese, H Zheng, et al. “Negative-strand RNA viruses: genetic engineering and applications.” In: *Proceedings of the National Academy of Sciences* 93.21 (Oct. 15, 1996). Publisher: Proceedings of the National Academy of Sciences, pp. 11354–11358. DOI: [10.1073/pnas.93.21.11354](https://doi.org/10.1073/pnas.93.21.11354). URL: <https://www.pnas.org/doi/abs/10.1073/pnas.93.21.11354> (cit. on p. 94).
- [312] Esteban Domingo and Celia Perales. “Virus Evolution”. In: [eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9780470015902.a0000436.pub3](https://onlinelibrary.wiley.com/doi/pdf/10.1002/9780470015902.a0000436.pub3) DOI: [10.1002/9780470015902.a0000436.pub3](https://doi.org/10.1002/9780470015902.a0000436.pub3). John Wiley & Sons, Ltd, 2014. DOI: [10.1002/9780470015902.a0000436.pub3](https://doi.org/10.1002/9780470015902.a0000436.pub3). URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470015902.a0000436.pub3> (cit. on p. 94).
- [313] Philip V'kovski, Annika Kratzel, et al. “Coronavirus biology and replication: implications for SARS-CoV-2”. In: *Nature Reviews Microbiology* 19.3 (Mar. 2021). Number: 3 Publisher: Nature Publishing Group, pp. 155–170. DOI: [10.1038/s41579-020-00468-6](https://doi.org/10.1038/s41579-020-00468-6). URL: <https://www.nature.com/articles/s41579-020-00468-6> (cit. on p. 94).
- [314] Anne Tuiskunen Bäck and Åke Lundkvist. “Dengue viruses – an overview”. In: *Infection Ecology & Epidemiology* 3 (Aug. 30, 2013). PMID: 24003364 PMCID: PMC3759171, 10.3402/iee.v3i0.19839. DOI: [10.3402/iee.v3i0.19839](https://doi.org/10.3402/iee.v3i0.19839). URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3759171/> (cit. on p. 94).
- [315] Lynn B. Dustin, Barbara Bartolini, et al. “Hepatitis C virus: life cycle in cells, infection and host response, and analysis of molecular markers influencing the outcome of infection and response to therapy”. In: *Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases* 22.10 (Oct. 2016). PMID: 27592089 PMCID: PMC5627509,

## REFERENCES FOR CHAPTER 5

- pp. 826–832. DOI: [10.1016/j.cmi.2016.08.025](https://doi.org/10.1016/j.cmi.2016.08.025). URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5627509/> (cit. on p. 94).
- [316] Meelis Kadaja, Toomas Silla, et al. “Papillomavirus DNA replication — From initiation to genomic instability”. In: *Virology*. Small Viruses, Big Discoveries: The Interwoven Story of the Small DNA Tumor Viruses 384.2 (Feb. 20, 2009), pp. 360–368. DOI: [10.1016/j.virol.2008.11.032](https://doi.org/10.1016/j.virol.2008.11.032). URL: <https://www.sciencedirect.com/science/article/pii/S0042682208007691> (cit. on p. 94).
- [317] Sandra K. Weller and Donald M. Coen. “Herpes Simplex Viruses: Mechanisms of DNA Replication”. In: *Cold Spring Harbor Perspectives in Biology* 4.9 (Sept. 2012). PMID: 22952399 PMCID: PMC3428768, a013011. DOI: [10.1101/cshperspect.a013011](https://doi.org/10.1101/cshperspect.a013011). URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3428768/> (cit. on p. 94).
- [318] Juergen Beck and Michael Nassal. “Hepatitis B virus replication”. In: *World Journal of Gastroenterology : WJG* 13.1 (Jan. 7, 2007). PMID: 17206754 PMCID: PMC4065876, pp. 48–64. DOI: [10.3748/wjg.v13.i1.48](https://doi.org/10.3748/wjg.v13.i1.48). URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4065876/> (cit. on p. 94).
- [319] Jesse D. Pyle and Karen-Beth G. Scholthof. “Chapter 58 - Biology and Pathogenesis of Satellite Viruses”. In: ed. by Ahmed Hadidi, Ricardo Flores, et al. DOI: [10.1016/B978-0-12-801498-1.00058-9](https://doi.org/10.1016/B978-0-12-801498-1.00058-9). Boston: Academic Press, Jan. 1, 2017, pp. 627–636. DOI: [10.1016/B978-0-12-801498-1.00058-9](https://doi.org/10.1016/B978-0-12-801498-1.00058-9). URL: <https://www.sciencedirect.com/science/article/pii/B9780128014981000589> (cit. on p. 94).
- [320] Didier Raoult, Stéphane Audic, et al. “The 1.2-megabase genome sequence of Mimivirus”. In: *Science (New York, N.Y.)* 306.5700 (Nov. 19, 2004). PMID: 15486256, pp. 1344–1350. DOI: [10.1126/science.1101485](https://doi.org/10.1126/science.1101485) (cit. on p. 94).
- [321] José A. Campillo-Balderas, Antonio Lazcano, and Arturo Becerra. “Viral Genome Size Distribution Does not Correlate with the Antiquity of the Host Lineages”. In: *Frontiers in Ecology and Evolution* 3 (2015). URL: [https://www.frontiersin.org/article/10.3389/fevo.2015.00143](https://doi.org/10.3389/fevo.2015.00143) (cit. on p. 94).
- [322] Alan J Cann. “Virus Structure”. In: [\\_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9780470015902.a0000439.pub2](https://doi.org/10.1002/9780470015902.a0000439.pub2). John Wiley & Sons, Ltd, 2015, pp. 1–9. DOI: [10.1002 / 9780470015902 . a0000439 . pub2](https://doi.org/10.1002/9780470015902.a0000439.pub2). URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470015902.a0000439.pub2> (cit. on p. 94).
- [323] Florian Hladik and M. Juliana McElrath. “Setting the stage: host invasion by HIV”. In: *Nature Reviews Immunology* 8.6 (June 2008). Number: 6 Publisher: Nature Publishing Group, pp. 447–457. DOI: [10.1038/nri2302](https://doi.org/10.1038/nri2302). URL: <https://www.nature.com/articles/nri2302> (cit. on p. 95).

## CHAPTER 5

- [324] George M. Shaw and Eric Hunter. "HIV Transmission". In: *Cold Spring Harbor Perspectives in Medicine* 2.11 (Jan. 11, 2012). Publisher: Cold Spring Harbor Laboratory Press PMID: 23043157, a006965. DOI: [10.1101/cshperspect.a006965](https://doi.org/10.1101/cshperspect.a006965). URL: <http://perspectivesinmedicine.cshlp.org/content/2/11/a006965> (cit. on p. 95).
- [325] Robin A. Weiss. "How Does HIV Cause AIDS?" In: *Science* 260.5112 (May 28, 1993). Publisher: American Association for the Advancement of Science, pp. 1273–1279. DOI: [10.1126/science.8493571](https://doi.org/10.1126/science.8493571). URL: <https://www.science.org/doi/10.1126/science.8493571> (cit. on p. 95).
- [326] Anne Melhuish and Penny Lewthwaite. "Natural history of HIV and AIDS". In: *Medicine* 46.6 (June 1, 2018), pp. 356–361. DOI: [10.1016/j.mpmed.2018.03.010](https://doi.org/10.1016/j.mpmed.2018.03.010). URL: <https://www.sciencedirect.com/science/article/pii/S1357303918300744> (cit. on p. 95).
- [327] John F. Murray, Charles P. Felton, et al. "Pulmonary complications of the acquired immunodeficiency syndrome". In: *New England Journal of Medicine* 310.25 (1984). tex.eprint: <https://doi.org/10.1056/NEJM198406213102529>, pp. 1682–1688. DOI: [10.1056/NEJM198406213102529](https://doi.org/10.1056/NEJM198406213102529). URL: <https://doi.org/10.1056/NEJM198406213102529> (cit. on p. 95).
- [328] Shrikanth Sampath, Anwar Khedr, et al. "Pandemics Throughout the History". In: *Cureus* 13.9 (Sept. 20, 2021). Publisher: Cureus. DOI: [10.7759/cureus.18136](https://doi.org/10.7759/cureus.18136). URL: <https://www.cureus.com/articles/69273-pandemics-throughout-the-history> (cit. on p. 95).
- [329] World Health Organization. *Global report: UNAIDS report on the global AIDS epidemic 2010*. OCLC: 780905192. Geneve: World Health Organization, 2010. URL: [https://www.unaids.org/globalreport/Global\\_report.htm](https://www.unaids.org/globalreport/Global_report.htm) (cit. on p. 95).
- [330] F. Barré-Sinoussi, J. C. Chermann, et al. "Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS)". In: *Science (New York, N.Y.)* 220.4599 (May 20, 1983). PMID: 6189183, pp. 868–871. DOI: [10.1126/science.6189183](https://doi.org/10.1126/science.6189183) (cit. on p. 95).
- [331] R. C. Gallo, P. S. Sarin, et al. "Isolation of human T-cell leukemia virus in acquired immune deficiency syndrome (AIDS)". In: *Science (New York, N.Y.)* 220.4599 (May 20, 1983). PMID: 6601823, pp. 865–867. DOI: [10.1126/science.6601823](https://doi.org/10.1126/science.6601823) (cit. on p. 95).
- [332] François Clavel, Denise Guétard, et al. "Isolation of a New Human Retrovirus from West African Patients with AIDS". In: *Science* 233.4761 (July 18, 1986). Publisher: American Association for the Advancement of Science, pp. 343–346. DOI: [10.1126/science.2425430](https://doi.org/10.1126/science.2425430). URL: <https://www.science.org/doi/abs/10.1126/science.2425430> (cit. on p. 95).

## REFERENCES FOR CHAPTER 5

- [333] Peter B. Gilbert, Ian W. McKeague, et al. “Comparison of HIV-1 and HIV-2 infectivity from a prospective cohort study in Senegal”. In: *Statistics in Medicine* 22.4 (Feb. 28, 2003), pp. 573–593. DOI: [10.1002/sim.1342](https://doi.org/10.1002/sim.1342). URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.1342> (cit. on p. 95).
- [334] Maarten F. Schim van der Loeff, Akum Aveika Awasana, et al. “Sixteen years of HIV surveillance in a West African research clinic reveals divergent epidemic trends of HIV-1 and HIV-2”. In: *International Journal of Epidemiology* 35.5 (Oct. 2006). PMID: 16543363, pp. 1322–1328. DOI: [10.1093/ije/dyl037](https://doi.org/10.1093/ije/dyl037) (cit. on p. 95).
- [335] Feng Gao, Elizabeth Bailes, et al. “Origin of HIV-1 in the chimpanzee Pan troglodytes troglodytes”. In: *Nature* 397.6718 (Feb. 1999). Number: 6718 Publisher: Nature Publishing Group, pp. 436–441. DOI: [10.1038/17130](https://doi.org/10.1038/17130). URL: <https://www.nature.com/articles/17130> (cit. on p. 95).
- [336] Donald J. Hamel, Jean-Louis Sankalé, et al. “Twenty years of prospective molecular epidemiology in Senegal: changes in HIV diversity”. In: *AIDS research and human retroviruses* 23.10 (Oct. 2007). PMID: 17961103, pp. 1189–1196. DOI: [10.1089/aid.2007.0037](https://doi.org/10.1089/aid.2007.0037) (cit. on p. 95).
- [337] Paul M. Sharp and Beatrice H. Hahn. “Origins of HIV and the AIDS Pandemic”. In: *Cold Spring Harbor Perspectives in Medicine*: 1.1 (Sept. 2011). PMID: 22229120 PMCID: PMC3234451, a006841. DOI: [10.1101/cshperspect.a006841](https://doi.org/10.1101/cshperspect.a006841). URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3234451/> (cit. on pp. 95, 96).
- [338] Vanessa M. Hirsch, Robert A. Olmsted, et al. “An African primate lentivirus (SIVsm) closely related to HIV-2”. In: *Nature* 339.6223 (June 1989). Number: 6223 Publisher: Nature Publishing Group, pp. 389–392. DOI: [10.1038/339389a0](https://doi.org/10.1038/339389a0). URL: <https://www.nature.com/articles/339389a0> (cit. on p. 95).
- [339] Feng Gao, Ling Yue, et al. “Human infection by genetically diverse SIVSM-related HIV-2 in West Africa”. In: *Nature* 358.6386 (Aug. 1992). Number: 6386 Publisher: Nature Publishing Group, pp. 495–499. DOI: [10.1038/358495a0](https://doi.org/10.1038/358495a0). URL: <https://www.nature.com/articles/358495a0> (cit. on p. 95).
- [340] Z. Chen, P. Telfier, et al. “Genetic characterization of new West African simian immunodeficiency virus SIVsm: geographic clustering of household-derived SIV strains with human immunodeficiency virus type 2 subtypes and genetically diverse viruses from a single feral sooty mangabey troop”. In: *Journal of Virology* 70.6 (June 1996). PMID: 8648696 PMCID: PMC190237, pp. 3617–3627. DOI: [10.1128/JVI.70.6.3617-3627.1996](https://doi.org/10.1128/JVI.70.6.3617-3627.1996) (cit. on p. 95).
- [341] Joris Hemelaar. “The origin and diversity of the HIV-1 pandemic”. In: *Trends in Molecular Medicine* 18.3 (Mar. 1, 2012), pp. 182–192. DOI: [10.1016/j.molmed.2011.12.001](https://doi.org/10.1016/j.molmed.2011.12.001). URL: <https://www.sciencedirect.com/science/article/pii/S1471491411002103> (cit. on p. 95).

## CHAPTER 5

- [342] Michael Worobey, Marlea Gemmel, et al. “Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960”. In: *Nature* 455.7213 (Oct. 2008). Number: 7213 Publisher: Nature Publishing Group, pp. 661–664. DOI: [10.1038/nature07390](https://doi.org/10.1038/nature07390). URL: <https://www.nature.com/articles/nature07390> (cit. on p. 96).
- [343] N. Vidal, M. Peeters, et al. “Unprecedented degree of human immunodeficiency virus type 1 (HIV-1) group M genetic diversity in the Democratic Republic of Congo suggests that the HIV-1 pandemic originated in Central Africa”. In: *Journal of Virology* 74.22 (Nov. 2000). PMID: 11044094 PMCID: PMC110924, pp. 10498–10507. DOI: [10.1128/jvi.74.22.10498-10507.2000](https://doi.org/10.1128/jvi.74.22.10498-10507.2000) (cit. on p. 96).
- [344] Nuno R. Faria, Andrew Rambaut, et al. “The early spread and epidemic ignition of HIV-1 in human populations”. In: *Science (New York, N.Y.)* 346.6205 (Oct. 3, 2014). PMID: 25278604 PMCID: PMC4254776, pp. 56–61. DOI: [10.1126/science.1256739](https://doi.org/10.1126/science.1256739). URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4254776/> (cit. on p. 96).
- [345] B. Korber, M. Muldoon, et al. “Timing the ancestor of the HIV-1 pandemic strains”. In: *Science (New York, N.Y.)* 288.5472 (June 9, 2000). PMID: 10846155, pp. 1789–1796. DOI: [10.1126/science.288.5472.1789](https://doi.org/10.1126/science.288.5472.1789) (cit. on p. 96).
- [346] Andrew Rambaut, David Posada, et al. “The causes and consequences of HIV evolution”. In: *Nature Reviews Genetics* 5.1 (Jan. 2004). Number: 1 Publisher: Nature Publishing Group, pp. 52–61. DOI: [10.1038/nrg1246](https://doi.org/10.1038/nrg1246). URL: <https://www.nature.com/articles/nrg1246> (cit. on p. 96).
- [347] Francine E. McCutchan. “Global epidemiology of HIV”. In: *Journal of Medical Virology* 78.S1 (2006). \_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/jmv.20599>, S7–S12. DOI: [10.1002/jmv.20599](https://doi.org/10.1002/jmv.20599). URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/jmv.20599> (cit. on p. 96).
- [348] Marcos Pérez-Losada, Miguel Arenas, et al. “Recombination in viruses: Mechanisms, methods of study, and evolutionary consequences”. In: *Infection, Genetics and Evolution* 30 (Mar. 2015). PMID: 25541518 PMCID: PMC7106159, pp. 296–307. DOI: [10.1016/j.meegid.2014.12.022](https://doi.org/10.1016/j.meegid.2014.12.022). URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7106159/> (cit. on p. 96).
- [349] D. L. Robertson, B. H. Hahn, and P. M. Sharp. “Recombination in AIDS viruses”. In: *Journal of Molecular Evolution* 40.3 (Mar. 1995). tex.ids=robertsonRecombinationAIDSViruses1995a PMID: 7723052, pp. 249–259. DOI: [10.1007/BF00163230](https://doi.org/10.1007/BF00163230) (cit. on p. 96).
- [350] *HIV Circulating Recombinant Forms (CRFs)*. URL: <https://www.hiv.lanl.gov/content/sequence/HIV/CRFs/CRFs.html> (cit. on p. 96).
- [351] David Posada, Keith A. Crandall, and Edward C. Holmes. “Recombination in evolutionary genomics”. In: *Annual Review of Genetics* 36 (2002). PMID: 12429687, pp. 75–97. DOI: [10.1146/annurev.genet.36.040202.111115](https://doi.org/10.1146/annurev.genet.36.040202.111115) (cit. on p. 96).

## REFERENCES FOR CHAPTER 5

- [352] Barbara S. Taylor, Magdalena E. Sobieszczyk, et al. “The Challenge of HIV-1 Subtype Diversity”. In: *New England Journal of Medicine* 358.15 (Apr. 10, 2008). Publisher: Massachusetts Medical Society \_eprint: <https://doi.org/10.1056/NEJMra0706737> PMID: 18403767, pp. 1590–1602. DOI: [10.1056/NEJMra0706737](https://doi.org/10.1056/NEJMra0706737). URL: <https://doi.org/10.1056/NEJMra0706737> (cit. on p. 96).
- [353] Joris Hemelaar, Eleanor Gouws, et al. “Global trends in molecular epidemiology of HIV-1 during 2000–2007”. In: *AIDS (London, England)* 25.5 (Mar. 13, 2011). PMID: 21297424 PMCID: PMC3755761, pp. 679–689. DOI: [10.1097/QAD.0b013e328342ff93](https://doi.org/10.1097/QAD.0b013e328342ff93). URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3755761/> (cit. on p. 96).
- [354] *Distribution of all HIV-1 sequences: WORLD*. URL: <https://www.hiv.lanl.gov/components/sequence/HIV/geo/geo.comp> (cit. on p. 96).
- [355] Eric O. Freed. “HIV-1 Replication”. In: *Somatic Cell and Molecular Genetics* 26.1 (Nov. 1, 2001), pp. 13–33. DOI: [10.1023/A:1021070512287](https://doi.org/10.1023/A:1021070512287). URL: <https://doi.org/10.1023/A:1021070512287> (cit. on p. 96).
- [356] Monique R. Ferguson, Daniel R. Rojo, et al. “HIV-1 replication cycle”. In: *Clinics in Laboratory Medicine* 22.3 (Sept. 2002). PMID: 12244589, pp. 611–635. DOI: [10.1016/s0272-2712\(02\)00015-x](https://doi.org/10.1016/s0272-2712(02)00015-x) (cit. on p. 96).
- [357] M. L. Gougeon, A. G. Laurent-Crawford, et al. “Direct and indirect mechanisms mediating apoptosis during HIV infection: contribution to in vivo CD4 T cell depletion”. In: *Seminars in Immunology* 5.3 (June 1, 1993), pp. 187–194. DOI: [10.1006/smim.1993.1022](https://doi.org/10.1006/smim.1993.1022). URL: <https://www.sciencedirect.com/science/article/pii/S1044532383710225> (cit. on p. 97).
- [358] K. K. Vidya Vijayan, Krithika Priyadarshini Karthigeyan, et al. “Pathophysiology of CD4+ T-Cell Depletion in HIV-1 and HIV-2 Infections”. In: *Frontiers in Immunology* 8 (May 23, 2017). PMID: 28588579 PMCID: PMC5440548, p. 580. DOI: [10.3389/fimmu.2017.00580](https://doi.org/10.3389/fimmu.2017.00580). URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5440548/> (cit. on p. 97).
- [359] A. D. Frankel and J. A. Young. “HIV-1: fifteen proteins and an RNA”. In: *Annual Review of Biochemistry* 67 (1998). PMID: 9759480, pp. 1–25. DOI: [10.1146/annurev.biochem.67.1.1](https://doi.org/10.1146/annurev.biochem.67.1.1) (cit. on p. 97).
- [360] Torgils Fossen, Victor Wray, et al. “Solution Structure of the Human Immunodeficiency Virus Type 1 p6 Protein \*”. In: *Journal of Biological Chemistry* 280.52 (Dec. 30, 2005). Publisher: Elsevier PMID: 16234236, pp. 42515–42527. DOI: [10.1074/jbc.M507375200](https://doi.org/10.1074/jbc.M507375200). URL: [https://www.jbc.org/article/S0021-9258\(19\)47854-X/abstract](https://www.jbc.org/article/S0021-9258(19)47854-X/abstract) (cit. on p. 98).
- [361] H G Göttlinger, T Dorfman, et al. “Effect of mutations affecting the p6 gag protein on human immunodeficiency virus particle release.” In: *Proceedings of the National Academy of Sciences* 88.8 (Apr. 15, 1991). Publisher: Proceedings of the National Academy of Sciences, pp. 3195–3199. DOI: [10.1073/pnas.88.8.3195](https://doi.org/10.1073/pnas.88.8.3195). URL: <https://www.pnas.org/doi/abs/10.1073/pnas.88.8.3195> (cit. on p. 98).

## CHAPTER 5

- [362] M Huang, J M Orenstein, et al. “p6Gag is required for particle production from full-length human immunodeficiency virus type 1 molecular clones expressing protease”. In: *Journal of Virology* 69.11 (Nov. 1995). Publisher: American Society for Microbiology, pp. 6810–6818. DOI: [10.1128/jvi.69.11.6810-6818.1995](https://doi.org/10.1128/jvi.69.11.6810-6818.1995). URL: <https://journals.asm.org/doi/abs/10.1128/jvi.69.11.6810-6818.1995> (cit. on p. 98).
- [363] S Bour, R Geleziunas, and M A Wainberg. “The human immunodeficiency virus type 1 (HIV-1) CD4 receptor and its central role in promotion of HIV-1 infection”. In: *Microbiological Reviews* 59.1 (Mar. 1995). Publisher: American Society for Microbiology, pp. 63–93. DOI: [10.1128/mr.59.1.63-93.1995](https://doi.org/10.1128/mr.59.1.63-93.1995). URL: <https://journals.asm.org/doi/abs/10.1128/mr.59.1.63-93.1995> (cit. on p. 98).
- [364] L. D. Hernandez, L. R. Hoffman, et al. “Virus-cell and cell-cell fusion”. In: *Annual Review of Cell and Developmental Biology* 12 (1996). tex.ids= hernandezVIR-USCELLCELLCELLFUSION1996 PMID: 8970739, pp. 627–661. DOI: [10.1146/annurev.cellbio.12.1.627](https://doi.org/10.1146/annurev.cellbio.12.1.627) (cit. on p. 98).
- [365] Ka Jones and Bm Peterlin. “Control of Rna Initiation and Elongation at the Hiv-1 Promoter”. In: *Annual Review of Biochemistry* 63 (1994). Place: Palo Alto Publisher: Annual Reviews WOS:A1994NV05900021, pp. 717–743. DOI: [10.1146/annurev.bi.63.070194.003441](https://doi.org/10.1146/annurev.bi.63.070194.003441). URL: <https://www.annualreviews.org/doi/10.1146/annurev.bi.63.070194.003441> (cit. on p. 99).
- [366] Thomas J. Hope. “Viral RNA export”. In: *Chemistry & Biology* 4.5 (May 1, 1997), pp. 335–344. DOI: [10.1016/S1074-5521\(97\)90124-1](https://doi.org/10.1016/S1074-5521(97)90124-1). URL: <https://www.sciencedirect.com/science/article/pii/S1074552197901241> (cit. on p. 99).
- [367] A. Mangasarian and D. Trono. “The multifaceted role of HIV Nef”. In: *Research in Virology* 148.1 (Jan. 1, 1997), pp. 30–33. DOI: [10.1016/S0923-2516\(97\)81909-7](https://doi.org/10.1016/S0923-2516(97)81909-7). URL: <https://www.sciencedirect.com/science/article/pii/S0923251697819097> (cit. on p. 99).
- [368] é. A. Cohen, R. A. Subramanian, and H. G. Göttlinger. “Role of Auxiliary Proteins in Retroviral Morphogenesis”. In: ed. by Hans-Georg Kräusslich. Current Topics in Microbiology and Immunology. DOI: [10.1007/978-3-642-80145-7\\_7](https://doi.org/10.1007/978-3-642-80145-7_7). Berlin, Heidelberg: Springer, 1996, pp. 219–235. DOI: [10.1007/978-3-642-80145-7\\_7](https://doi.org/10.1007/978-3-642-80145-7_7). URL: [https://doi.org/10.1007/978-3-642-80145-7\\_7](https://doi.org/10.1007/978-3-642-80145-7_7) (cit. on p. 99).
- [369] Robert A Lamb and Lawrence H Pinto. “Do Vpu and Vpr of Human Immunodeficiency Virus Type 1 and NB of Influenza B Virus Have Ion Channel Activities in the Viral Life Cycles?” In: *Virology* 229.1 (Mar. 3, 1997), pp. 1–11. DOI: [10.1006/viro.1997.8451](https://doi.org/10.1006/viro.1997.8451). URL: <https://www.sciencedirect.com/science/article/pii/S0042682297984518> (cit. on p. 99).
- [370] Nabab Khan and Jonathan D. Geiger. “Role of Viral Protein U (Vpu) in HIV-1 Infection and Pathogenesis”. In: *Viruses* 13.8 (July 27, 2021). PMID: 34452331 PMCID: PMC8402909, p. 1466. DOI: [10.3390/v13081466](https://doi.org/10.3390/v13081466). URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8402909/> (cit. on p. 99).

## REFERENCES FOR CHAPTER 5

- [371] Michael Emerman. “HIV-1, Vpr and the cell cycle”. In: *Current Biology* 6.9 (Sept. 1, 1996), pp. 1096–1103. DOI: [10.1016/S0960-9822\(02\)00676-0](https://doi.org/10.1016/S0960-9822(02)00676-0). URL: <https://www.sciencedirect.com/science/article/pii/S0960982202006760> (cit. on p. 99).
- [372] R. H. Miller. “Human immunodeficiency virus may encode a novel protein on the genomic DNA plus strand”. In: *Science (New York, N.Y.)* 239.4846 (Mar. 18, 1988). PMID: 3347840, pp. 1420–1422. DOI: [10.1126/science.3347840](https://doi.org/10.1126/science.3347840) (cit. on p. 99).
- [373] Sylvie Briquet and Catherine Vaquero. “Immunolocalization Studies of an Antisense Protein in HIV-1-Infected Cells and Viral Particles”. In: *Virology* 292.2 (Jan. 20, 2002), pp. 177–184. DOI: [10.1006/viro.2001.1224](https://doi.org/10.1006/viro.2001.1224). URL: <https://www.sciencedirect.com/science/article/pii/S0042682201912243> (cit. on p. 99).
- [374] Elodie Cassan, Anne-Muriel Arigon-Chifolleau, et al. “Concomitant emergence of the antisense protein gene of HIV-1 and of the pandemic”. In: *Proceedings of the National Academy of Sciences* 113.41 (Oct. 11, 2016). Publisher: National Academy of Sciences Section: Biological Sciences PMID: 27681623, pp. 11537–11542. DOI: [10.1073/pnas.1605739113](https://doi.org/10.1073/pnas.1605739113). URL: <https://www.pnas.org/content/113/41/11537> (cit. on p. 99).
- [375] Juliette Savoret, Nathalie Chazal, et al. “A Pilot Study of the Humoral Response Against the AntiSense Protein (ASP) in HIV-1-Infected Patients”. In: *Frontiers in Microbiology* 11 (2020). URL: [https://www.frontiersin.org/article/10.3389/fmicb.2020.00020](https://doi.org/10.3389/fmicb.2020.00020) (cit. on p. 99).
- [376] Christine Zardecki, Shuchismita Dutta, et al. “PDB-101: Educational resources supporting molecular explorations through biology and medicine”. In: *Protein Science* 31.1 (2022). \_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/pro.4200>, pp. 129–140. DOI: [10.1002/pro.4200](https://doi.org/10.1002/pro.4200). URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/pro.4200> (cit. on p. 123).
- [377] Stefan G. Sarafianos, Bruno Marchand, et al. “Structure and Function of HIV-1 Reverse Transcriptase: Molecular Mechanisms of Polymerization and Inhibition”. In: *Journal of molecular biology* 385.3 (Jan. 2009), pp. 693–713. ISSN: 0022-2836. DOI: [10.1016/j.jmb.2008.10.071](https://doi.org/10.1016/j.jmb.2008.10.071) (cit. on pp. 99, 142).
- [378] David S. Goodsell, Ludovic Autin, and Arthur J. Olson. “Illustrate: Software for Biomolecular Illustration”. In: *Structure (London, England: 1993)* 27.11 (Nov. 5, 2019). PMID: 31519398 PMCID: PMC6834899, 1716–1720.e1. DOI: [10.1016/j.str.2019.08.011](https://doi.org/10.1016/j.str.2019.08.011) (cit. on p. 100).
- [379] Thang K. Chiu and David R. Davies. “Structure and Function of HIV-1 Integrase”. In: *Current Topics in Medicinal Chemistry* 4.9 (May 1, 2004). tex.ids= chiuStructureFunctionHIV1, pp. 965–977. DOI: [10.2174/1568026043388547](https://doi.org/10.2174/1568026043388547) (cit. on p. 100).

## CHAPTER 5

- [380] Dominic Esposito and Robert Craigie. “HIV Integrase Structure and Function”. In: ed. by Karl Rlaramorosch, Frederick A. Murphy, and Aaron J. Shawn. Vol. 52. tex.ids= espositoHIVIntegraseStructure1999a DOI: 10.1016/S0065-3527(08)60304-8. Academic Press, Jan. 1, 1999, pp. 319–333. URL: <https://www.sciencedirect.com/science/article/pii/S0065352708603048> (cit. on p. 100).
- [381] Olivier Delelis, Kevin Carayon, et al. “Integrase and integration: biochemical activities of HIV-1 integrase”. In: *Retrovirology* 5 (Dec. 17, 2008). PMID: 19091057 PMCID: PMC2615046, p. 114. DOI: 10.1186/1742-4690-5-114. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2615046/> (cit. on p. 100).
- [382] Laurence H. Pearl and William R. Taylor. “A structural model for the retroviral proteases”. In: *Nature* 329.6137 (Sept. 1987). Number: 6137 Publisher: Nature Publishing Group, pp. 351–354. DOI: 10.1038/329351a0. URL: <https://www.nature.com/articles/329351a0> (cit. on p. 100).
- [383] Sergei Gulnik, John W. Erickson, and Dong Xie. “HIV protease: Enzyme function and drug resistance”. In: vol. 58. DOI: 10.1016/S0083-6729(00)58026-1. Academic Press, Jan. 1, 2000, pp. 213–256. DOI: 10.1016/S0083-6729(00)58026-1. URL: <https://www.sciencedirect.com/science/article/pii/S0083672900580261> (cit. on p. 100).
- [384] Abelardo M. Silva, Raúl E. Cachau, et al. “Inhibition and catalytic mechanism of HIV-1 aspartic protease”. In: *Journal of Molecular Biology* 255.2 (Jan. 1, 1996), pp. 321–340. DOI: 10.1006/jmbi.1996.0026. URL: <https://www.sciencedirect.com/science/article/pii/S0022283696900261> (cit. on p. 101).
- [385] Viktor Hornak, Asim Okur, et al. “HIV-1 protease flaps spontaneously open and reclose in molecular dynamics simulations”. In: *Proceedings of the National Academy of Sciences* 103.4 (Jan. 24, 2006). Publisher: Proceedings of the National Academy of Sciences, pp. 915–920. DOI: 10.1073/pnas.0508452103. URL: <https://www.pnas.org/doi/10.1073/pnas.0508452103> (cit. on p. 101).
- [386] Yuqi Yu, Jinan Wang, et al. “Structural insights into HIV-1 protease flap opening processes and key intermediates”. In: *RSC Advances* 7.71 (Sept. 15, 2017). Publisher: The Royal Society of Chemistry, pp. 45121–45128. DOI: 10.1039/C7RA09691G. URL: <https://pubs.rsc.org/en/content/articlelanding/2017/ra/c7ra09691g> (cit. on p. 101).
- [387] Robert W. Eisinger, Carl W. Dieffenbach, and Anthony S. Fauci. “HIV Viral Load and Transmissibility of HIV Infection: Undetectable Equals Untransmittable”. In: *JAMA* 321.5 (Feb. 5, 2019), pp. 451–452. DOI: 10.1001/jama.2018.21167. URL: <https://doi.org/10.1001/jama.2018.21167> (cit. on p. 101).
- [388] Frank J. Palella, Kathleen M. Delaney, et al. “Declining Morbidity and Mortality among Patients with Advanced Human Immunodeficiency Virus Infection”. In: *New England Journal of Medicine* 338.13 (Mar. 26, 1998). Publisher: Massachusetts Medical Society \_eprint: <https://doi.org/10.1056/NEJM199803263381301>

## REFERENCES FOR CHAPTER 5

- PMID: 9516219, pp. 853–860. DOI: [10.1056/NEJM199803263381301](https://doi.org/10.1056/NEJM199803263381301). URL: <https://doi.org/10.1056/NEJM199803263381301> (cit. on p. 101).
- [389] Andrew R. Zolopa. “The evolution of HIV treatment guidelines: Current state-of-the-art of ART”. In: *Antiviral Research*. Twenty-five Years of Antiretroviral Drug Development: Progress and Prospects 85.1 (Jan. 1, 2010), pp. 241–244. DOI: [10.1016/j.antiviral.2009.10.018](https://doi.org/10.1016/j.antiviral.2009.10.018). URL: <https://www.sciencedirect.com/science/article/pii/S0166354209005051> (cit. on p. 101).
- [390] Steven S. Forsythe, William McGreevey, et al. “Twenty Years Of Antiretroviral Therapy For People Living With HIV: Global Costs, Health Achievements, Economic Benefits”. In: *Health Affairs* 38.7 (July 2019). Publisher: Health Affairs, pp. 1163–1172. DOI: [10.1377/hlthaff.2018.05391](https://doi.org/10.1377/hlthaff.2018.05391). URL: <https://www.healthaffairs.org/doi/10.1377/hlthaff.2018.05391> (cit. on p. 101).
- [391] Margaret A. Fischl, Douglas D. Richman, et al. “The Efficacy of Azidothymidine (AZT) in the Treatment of Patients with AIDS and AIDS-Related Complex”. In: *New England Journal of Medicine* 317.4 (July 23, 1987). Publisher: Massachusetts Medical Society \_eprint: <https://doi.org/10.1056/NEJM198707233170401> PMID: 3299089, pp. 185–191. DOI: [10.1056/NEJM198707233170401](https://doi.org/10.1056/NEJM198707233170401). URL: <https://doi.org/10.1056/NEJM198707233170401> (cit. on p. 101).
- [392] Douglas D. Richman. “Susceptibility to nucleoside analogues of zidovudine-resistant isolates of human immunodeficiency virus”. In: *The American Journal of Medicine*. Dideoxycydine (ddC): A Potent Antiretroviral Agent for Human Immunodeficietion Virus Infection 88.5, Supplement 2 (May 21, 1990), S8–S10. DOI: [10.1016/0002-9343\(90\)90414-9](https://doi.org/10.1016/0002-9343(90)90414-9). URL: <https://www.sciencedirect.com/science/article/pii/0002934390904149> (cit. on p. 101).
- [393] R. M. Gulick, J. W. Mellors, et al. “Treatment with indinavir, zidovudine, and lamivudine in adults with human immunodeficiency virus infection and prior antiretroviral therapy”. In: *The New England Journal of Medicine* 337.11 (Sept. 11, 1997). PMID: 9287228, pp. 734–739. DOI: [10.1056/NEJM199709113371102](https://doi.org/10.1056/NEJM199709113371102) (cit. on p. 101).
- [394] Lucia Palmisano and Stefano Vella. “A brief history of antiretroviral therapy of HIV infection: success and challenges”. In: *Annali dell'Istituto Superiore Di Sanita* 47.1 (2011). PMID: 21430338, pp. 44–48. DOI: [10.4415/ANN\\_11\\_01\\_10](https://doi.org/10.4415/ANN_11_01_10) (cit. on p. 101).
- [395] *One-pill-a-day HIV treatments*. URL: <https://www.aidsmap.com/about-hiv/one-pill-a-day-hiv-treatments> (cit. on p. 101).
- [396] John C. Tilton and Robert W. Doms. “Entry inhibitors in the treatment of HIV-1 infection”. In: *Antiviral Research*. Twenty-five Years of Antiretroviral Drug Development: Progress and Prospects 85.1 (Jan. 1, 2010), pp. 91–100. DOI: [10.1016/j.antiviral.2009.07.022](https://doi.org/10.1016/j.antiviral.2009.07.022). URL: <https://www.sciencedirect.com/science/article/pii/S0166354209003945> (cit. on p. 101).

## CHAPTER 5

- [397] Kimberly K. Scarsi, Joshua P. Havens, et al. “HIV-1 Integrase Inhibitors: A Comparative Review of Efficacy and Safety”. In: *Drugs* 80.16 (Nov. 1, 2020), pp. 1649–1676. DOI: [10.1007/s40265-020-01379-9](https://doi.org/10.1007/s40265-020-01379-9). URL: <https://doi.org/10.1007/s40265-020-01379-9> (cit. on p. 101).
- [398] Robert M. Grant, Javier R. Lama, et al. “Preexposure Chemoprophylaxis for HIV Prevention in Men Who Have Sex with Men”. In: *New England Journal of Medicine* 363.27 (Dec. 30, 2010). Publisher: Massachusetts Medical Society \_eprint: <https://doi.org/10.1056/NEJMoa1011205> PMID: 21091279, pp. 2587–2599. DOI: [10.1056/NEJMoa1011205](https://doi.org/10.1056/NEJMoa1011205). URL: <https://doi.org/10.1056/NEJMoa1011205> (cit. on p. 101).
- [399] Jared M. Baeten, Deborah Donnell, et al. “Antiretroviral prophylaxis for HIV prevention in heterosexual men and women”. In: *The New England Journal of Medicine* 367.5 (Aug. 2, 2012). PMID: 22784037 PMCID: PMC3770474, pp. 399–410. DOI: [10.1056/NEJMoa1108524](https://doi.org/10.1056/NEJMoa1108524) (cit. on p. 101).
- [400] Susan P. Buchbinder and Albert Liu. “Pre-exposure prophylaxis and the promise of combination prevention approaches”. In: *AIDS and behavior* 15 Suppl 1 (Apr. 2011). PMID: 21331801 PMCID: PMC3064892, S72–79. DOI: [10.1007/s10461-011-9894-1](https://doi.org/10.1007/s10461-011-9894-1) (cit. on p. 101).
- [401] James, IV Riddell, K. Rivet Amico, and Kenneth H. Mayer. “HIV Preexposure Prophylaxis: A Review”. In: *JAMA* 319.12 (Mar. 27, 2018), pp. 1261–1268. DOI: [10.1001/jama.2018.1917](https://doi.org/10.1001/jama.2018.1917). URL: <https://doi.org/10.1001/jama.2018.1917> (cit. on p. 101).
- [402] EMA. *Truvada*. Sept. 17, 2018. URL: <https://www.ema.europa.eu/en/medicines/human/EPAR/truvada> (cit. on p. 101).
- [403] R. M. Esnouf, J. Ren, et al. “Unique features in the structure of the complex between HIV-1 reverse transcriptase and the bis(heteroaryl)piperazine (BHAP) U-90152 explain resistance mutations for this nonnucleoside inhibitor”. In: *Proceedings of the National Academy of Sciences of the United States of America* 94.8 (Apr. 15, 1997). PMID: 9108091 PMCID: PMC20554, pp. 3984–3989. DOI: [10.1073/pnas.94.8.3984](https://doi.org/10.1073/pnas.94.8.3984) (cit. on p. 102).
- [404] Noel A. Roberts, Joseph A. Martin, et al. “Rational Design of Peptide-Based HIV Proteinase Inhibitors”. In: *Science* 248.4953 (Apr. 20, 1990). Publisher: American Association for the Advancement of Science, pp. 358–361. DOI: [10.1126/science.2183354](https://doi.org/10.1126/science.2183354). URL: <https://www.science.org/doi/10.1126/science.2183354> (cit. on p. 102).
- [405] José A Esté and Amalio Telenti. “HIV entry inhibitors”. In: *The Lancet* 370.9581 (July 7, 2007), pp. 81–88. DOI: [10.1016/S0140-6736\(07\)61052-6](https://doi.org/10.1016/S0140-6736(07)61052-6). URL: <https://www.sciencedirect.com/science/article/pii/S0140673607610526> (cit. on p. 102).

## REFERENCES FOR CHAPTER 5

- [406] J. Michael Kilby and Joseph J. Eron. "Novel Therapies Based on Mechanisms of HIV-1 Cell Entry". In: *New England Journal of Medicine* 348.22 (May 29, 2003). Publisher: Massachusetts Medical Society, pp. 2228–2238. DOI: [10.1056/NEJMra022812](https://doi.org/10.1056/NEJMra022812). URL: <https://www.nejm.org/doi/10.1056/NEJMra022812> (cit. on p. 102).
- [407] Courtney V Fletcher. "Enfuvirtide, a new drug for HIV infection". In: *The Lancet* 361.9369 (May 10, 2003), pp. 1577–1578. DOI: [10.1016/S0140-6736\(03\)13323-5](https://doi.org/10.1016/S0140-6736(03)13323-5). URL: <https://www.sciencedirect.com/science/article/pii/S0140673603133235> (cit. on p. 102).
- [408] Yves Pommier, Allison A. Johnson, and Christophe Marchand. "Integrase inhibitors to treat HIV/Aids". In: *Nature Reviews Drug Discovery* 4.3 (Mar. 2005). Number: 3 Publisher: Nature Publishing Group, pp. 236–248. DOI: [10.1038/nrd1660](https://doi.org/10.1038/nrd1660). URL: <https://www.nature.com/articles/nrd1660> (cit. on p. 102).
- [409] David C. Knox, Peter L. Anderson, et al. "Multidrug-Resistant HIV-1 Infection despite Preexposure Prophylaxis". In: *New England Journal of Medicine* 376.5 (Feb. 2, 2017). Publisher: Massachusetts Medical Society, pp. 501–502. DOI: [10.1056/NEJMc1611639](https://doi.org/10.1056/NEJMc1611639). URL: <https://www.nejm.org/doi/10.1056/NEJMc1611639> (cit. on p. 102).
- [410] Christopher B. Hurt, Joseph J. Eron, and Myron S. Cohen. "Pre-exposure prophylaxis and antiretroviral resistance: HIV prevention at a cost?" In: *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America* 53.12 (Dec. 2011). PMID: 21976467 PMCID: PMC3246871, pp. 1265–1270. DOI: [10.1093/cid/cir684](https://doi.org/10.1093/cid/cir684) (cit. on p. 102).
- [411] Kevin M. Gibas, Polly van den Berg, et al. "Drug Resistance During HIV Pre-Exposure Prophylaxis". In: *Drugs* 79.6 (Apr. 1, 2019), pp. 609–619. DOI: [10.1007/s40265-019-01108-x](https://doi.org/10.1007/s40265-019-01108-x). URL: <https://doi.org/10.1007/s40265-019-01108-x> (cit. on p. 102).
- [412] Palanee Ammaranond and Sayompoo Sanguansittanan. "Mechanism of HIV antiretroviral drugs progress toward drug resistance". In: *Fundamental & Clinical Pharmacology* 26.1 (2012). \_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1472-8206.2011.01009.x>, pp. 146–161. DOI: [10.1111/j.1472-8206.2011.01009.x](https://doi.org/10.1111/j.1472-8206.2011.01009.x). URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1472-8206.2011.01009.x> (cit. on p. 102).
- [413] François Clavel and Allan J. Hance. "HIV Drug Resistance". In: *New England Journal of Medicine* 350.10 (Mar. 4, 2004). Publisher: Massachusetts Medical Society \_eprint: <https://www.nejm.org/doi/pdf/10.1056/NEJMra025195> PMID: 14999114, pp. 1023–1035. DOI: [10.1056/NEJMra025195](https://doi.org/10.1056/NEJMra025195). URL: <https://www.nejm.org/doi/full/10.1056/NEJMra025195> (cit. on p. 102).

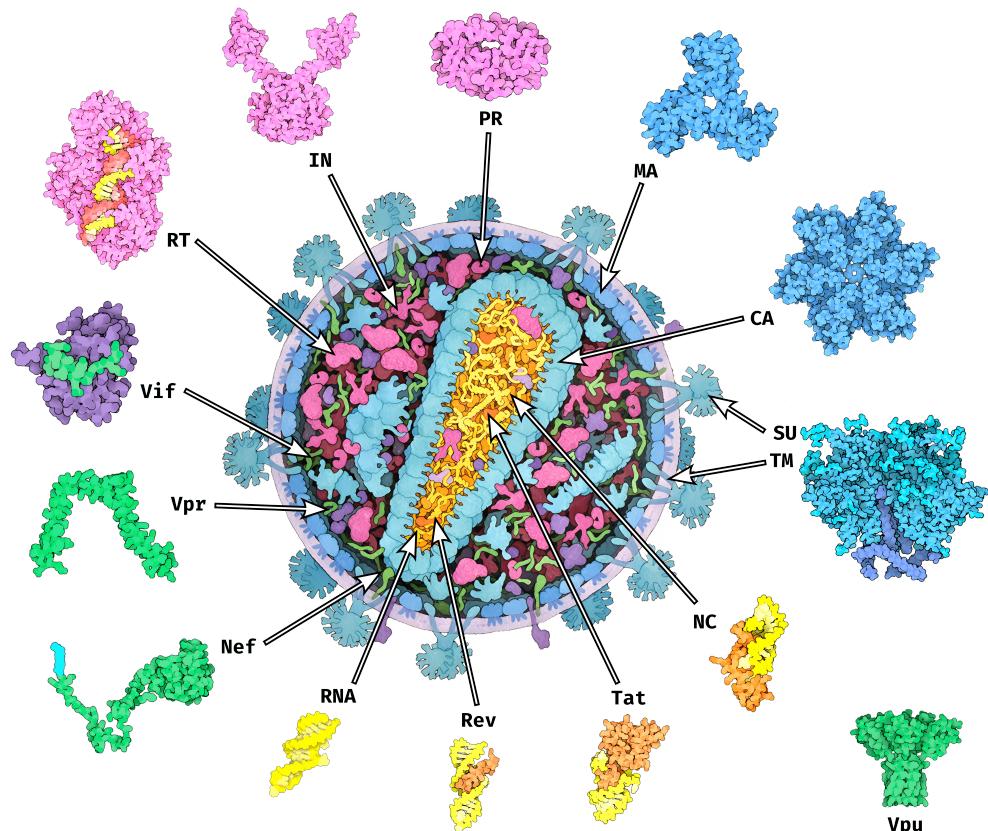
## CHAPTER 5

- [414] Luis Menéndez-Arias. “Mechanisms of resistance to nucleoside analogue inhibitors of HIV-1 reverse transcriptase”. In: *Virus Research*. Retroviral Reverse Transcription 134.1 (June 1, 2008), pp. 124–146. DOI: [10.1016/j.virusres.2007.12.015](https://doi.org/10.1016/j.virusres.2007.12.015). URL: <https://www.sciencedirect.com/science/article/pii/S0168170207004686> (cit. on p. 102).
- [415] N. Sluis-Cremer, D. Arion, and M. A. Parniak\*. “Molecular mechanisms of HIV-1 resistance to nucleoside reverse transcriptase inhibitors (NRTIs)”. In: *Cellular and Molecular Life Sciences CMLS* 57.10 (Sept. 1, 2000), pp. 1408–1422. DOI: [10.1007/PL00000626](https://doi.org/10.1007/PL00000626). URL: <https://doi.org/10.1007/PL00000626> (cit. on p. 102).
- [416] S. G. Sarafianos, K. Das, et al. “Lamivudine (3TC) resistance in HIV-1 reverse transcriptase involves steric hindrance with beta-branched amino acids”. In: *Proceedings of the National Academy of Sciences of the United States of America* 96.18 (Aug. 31, 1999). PMID: 10468556 PMCID: PMC17836, pp. 10027–10032. DOI: [10.1073/pnas.96.18.10027](https://doi.org/10.1073/pnas.96.18.10027) (cit. on p. 102).
- [417] P. R. Meyer, S. E. Matsuura, et al. “A mechanism of AZT resistance: an increase in nucleotide-dependent primer unblocking by mutant HIV-1 reverse transcriptase”. In: *Molecular Cell* 4.1 (July 1999). PMID: 10445025, pp. 35–43. DOI: [10.1016/s1097-2765\(00\)80185-9](https://doi.org/10.1016/s1097-2765(00)80185-9) (cit. on p. 102).
- [418] Paul L. Boyer, Stefan G. Sarafianos, et al. “Selective Excision of AZTMP by Drug-Resistant Human Immunodeficiency Virus Reverse Transcriptase”. In: *Journal of Virology* 75.10 (May 2001). PMID: 11312355 PMCID: PMC114238, pp. 4832–4842. DOI: [10.1128/JVI.75.10.4832-4842.2001](https://doi.org/10.1128/JVI.75.10.4832-4842.2001). URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC114238/> (cit. on p. 102).
- [419] Steven G. Deeks. “Nonnucleoside Reverse Transcriptase Inhibitor Resistance”. In: *JAIDS Journal of Acquired Immune Deficiency Syndromes* 26 (Mar. 1, 2001), S25. URL: [https://journals.lww.com/jaids/Abstract/2001/03011/Nonnucleoside\\_Reverse\\_Transcriptase\\_Inhibitor.4.aspx](https://journals.lww.com/jaids/Abstract/2001/03011/Nonnucleoside_Reverse_Transcriptase_Inhibitor.4.aspx) (cit. on p. 102).
- [420] Annemarie M. J. Wensing, Noortje M. van Maarseveen, and Monique Nijhuis. “Fifteen years of HIV Protease Inhibitors: raising the barrier to resistance”. In: *Antiviral Research*. Twenty-five Years of Antiretroviral Drug Development: Progress and Prospects 85.1 (Jan. 1, 2010), pp. 59–74. DOI: [10.1016/j.antiviral.2009.10.003](https://doi.org/10.1016/j.antiviral.2009.10.003). URL: <https://www.sciencedirect.com/science/article/pii/S0166354209004902> (cit. on p. 102).
- [421] Nese Kurt Yilmaz, Ronald Swanstrom, and Celia A. Schiffer. “Improving Viral Protease Inhibitors to Counter Drug Resistance”. In: *Trends in Microbiology* 24.7 (July 1, 2016), pp. 547–557. DOI: [10.1016/j.tim.2016.03.010](https://doi.org/10.1016/j.tim.2016.03.010). URL: <https://www.sciencedirect.com/science/article/pii/S0966842X16300026> (cit. on p. 102).
- [422] Jose-Luis Blanco, Vici Varghese, et al. “HIV-1 Integrase Inhibitor Resistance and Its Clinical Implications”. In: *The Journal of Infectious Diseases* 203.9 (May 1, 2011), pp. 1204–1214. DOI: [10.1093/infdis/jir025](https://doi.org/10.1093/infdis/jir025). URL: <https://doi.org/10.1093/infdis/jir025> (cit. on p. 102).

- [423] Raphaël Mourad, François Chevennet, et al. “A Phylotype-Based Analysis Highlights the Role of Drug-Naïve HIV-Positive Individuals in the Transmission of Antiretroviral Resistance in the UK”. ENGLISH. In: *Aids* 29.15 (Sept. 2015), pp. 1917–1925. ISSN: 0269-9370. DOI: [10.1097/QAD.0000000000000768](https://doi.org/10.1097/QAD.0000000000000768) (cit. on pp. 102, 127).
- [424] Sara M. Drescher, Viktor von Wyl, et al. “Treatment-Naïve Individuals Are the Major Source of Transmitted HIV-1 Drug Resistance in Men Who Have Sex With Men in the Swiss HIV Cohort Study”. In: *Clinical Infectious Diseases* 58.2 (Jan. 15, 2014), pp. 285–294. DOI: [10.1093/cid/cit694](https://doi.org/10.1093/cid/cit694). URL: <https://doi.org/10.1093/cid/cit694> (cit. on p. 102).
- [425] Ragna S Boerma, T Sonia Boender, et al. “High levels of pre-treatment HIV drug resistance and treatment failure in Nigerian children”. In: *Journal of the International AIDS Society* 19.1 (2016). \_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.7448/IAS.19.1.21140>, p. 21140. DOI: [10.7448/IAS.19.1.21140](https://doi.org/10.7448/IAS.19.1.21140). URL: <https://onlinelibrary.wiley.com/doi/abs/10.7448/IAS.19.1.21140> (cit. on p. 102).
- [426] Denise Kühnert, Roger Kouyos, et al. “Quantifying the fitness cost of HIV-1 drug resistance mutations through phylodynamics”. In: *PLOS Pathogens* 14.2 (Feb. 20, 2018). Publisher: Public Library of Science, e1006895. DOI: [10.1371/journal.ppat.1006895](https://doi.org/10.1371/journal.ppat.1006895). URL: <https://journals.plos.org/plospathogens/article?id=10.1371/journal.ppat.1006895> (cit. on p. 102).
- [427] Thibault Mesplède, Peter K. Quashie, et al. “Viral fitness cost prevents HIV-1 from evading dolutegravir drug pressure”. In: *Retrovirology* 10.1 (Feb. 22, 2013), p. 22. DOI: [10.1186/1742-4690-10-22](https://doi.org/10.1186/1742-4690-10-22). URL: <https://doi.org/10.1186/1742-4690-10-22> (cit. on p. 102).
- [428] Hannah Castro, Deenan Pillay, et al. “Persistence of HIV-1 Transmitted Drug Resistance Mutations”. In: *The Journal of Infectious Diseases* 208.9 (Nov. 1, 2013), pp. 1459–1463. DOI: [10.1093/infdis/jit345](https://doi.org/10.1093/infdis/jit345). URL: <https://academic.oup.com/jid/article/208/9/1459/2192846> (cit. on p. 103).
- [429] Luc Bassel, Anna Zhukova, et al. “Drug resistance mutations in HIV: new bioinformatics approaches and challenges”. In: *Current Opinion in Virology* 51 (Dec. 1, 2021), pp. 56–64. DOI: [10.1016/j.coviro.2021.09.009](https://doi.org/10.1016/j.coviro.2021.09.009). URL: <https://www.sciencedirect.com/science/article/pii/S1879625721001073> (cit. on p. 103).
- [430] UK CHIC Steering Committee. “The creation of a large UK-based multicentre cohort of HIV-infected individuals: The UK Collaborative HIV Cohort (UK CHIC) Study”. In: *HIV Medicine* 5.2 (2004), pp. 115–124. DOI: [10.1111/j.1468-1293.2004.00197.x](https://doi.org/10.1111/j.1468-1293.2004.00197.x). URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1468-1293.2004.00197.x> (cit. on p. 103).
- [431] Lucie Abeler-Dörner, Mary K. Grabowski, et al. “PANGEA-HIV 2: Phylogenetics And Networks for Generalised Epidemics in Africa”. In: *Current Opinion in HIV and AIDS* 14.3 (May 2019), 173–180. DOI: [10.1097/COH.0000000000000542](https://doi.org/10.1097/COH.0000000000000542). URL: <https://doi.org/10.1097/COH.0000000000000542> (cit. on p. 103).

## CHAPTER 5

- [432] Robert W. Shafer. “Rationale and Uses of a Public HIV Drug-Resistance Database”. In: *The Journal of Infectious Diseases* 194.Supplement\_1 (Sept. 1, 2006), S51–S58. DOI: [10.1086/505356](https://doi.org/10.1086/505356). URL: [https://academic.oup.com/jid/article/194/Supplement\\_1/S51/894267](https://academic.oup.com/jid/article/194/Supplement_1/S51/894267) (cit. on p. 103).
- [433] Annemarie M. Wensing, Vincent Calvez, et al. “2019 update of the drug resistance mutations in HIV-1”. In: *Topics in Antiviral Medicine* 27.3 (Sept. 2019). PMID: 31634862 PMCID: PMC6892618, pp. 111–121. URL: <https://pubmed.ncbi.nlm.nih.gov/31634862/> (cit. on p. 103).
- [434] Shauna A Clark, Charles Calef, and John W Mellors. “Mutations in retroviral genes associated with drug resistance”. In: *HIV sequence compendium* (2007). tex.ids= clarkMutationsRetroviralGenes publisher: Citeseer, 58–158 (cit. on p. 103).
- [435] Tommy F. Liu and Robert W. Shafer. “Web Resources for HIV Type 1 Genotypic-Resistance Test Interpretation”. In: *Clinical Infectious Diseases* 42.11 (June 2006), pp. 1608–1618. ISSN: 1058-4838. DOI: [10.1086/503914](https://doi.org/10.1086/503914) (cit. on p. 103).



**Figure 5.2.: Structure and main components of a mature HIV-1 virion.**

Structural proteins MA, CA, SU and TM are represented in Blue, functional enzymes RT, IN and PR in pink, RNA binding proteins Rev, Tat and NC in orange and accessory proteins Vif, Nef, Vpr and Vpu in green. Viral RNA is shown in yellow. The phospholipd membrane of the virion is shown in a light purple color. The p6 protein is not represented as it is largely unstructured. Vpu is not believed to be present in the HIV virion.

Figure adapted from PDB101 [376] ([PDB101.rcsb.org](http://PDB101.rcsb.org), CC By 4.0 License, detailed list of structures used available in Appendix [Appendix B](#)).



## 6. Contribution 2: Inferring mutation roles from sequence alignments using machine learning

*Intro to this chapter within the context of the thesis to go here!*

This chapter was written as an article titled:

**“Using Machine Learning and Big Data to Explore the Drug Resistance Landscape in HIV”**

it was originally published in August 2021, in *PLoS Computational Biology* ([doi:10.1371/journal.pcbi.1008873](https://doi.org/10.1371/journal.pcbi.1008873)).

The author list, complete with affiliations is given below:

Luc Bassel<sup>1,2\*</sup>, Anna Tostevin<sup>3</sup>, Christian Julian Villabona-Arenas<sup>4,5</sup>, Martine Peeters<sup>6</sup>, Stéphane Hué<sup>4,5</sup>, Olivier Gascuel<sup>1,7#</sup> On behalf of the UK HIV Drug Resistance Database<sup>^</sup>

**1** Unité de Bioinformatique Évolutive, Institut Pasteur, Paris, France

**2** Sorbonne Université, Collège doctoral, Paris, France

**3** Institute for Global Health, UCL, London, UK

**4** Department of Infectious Disease Epidemiology, London School of Hygiene and Tropical Medicine, London, UK

**5** Centre for Mathematical Modelling of Infectious Diseases, London School of Hygiene and Tropical Medicine, London, UK

**6** TransVIHMI (Recherches Translationnelles sur VIH et Maladies Infectieuses), Université de Montpellier, Institut de Recherche pour le Développement, INSERM, Montpellier, France

**7** Institut de Systématique, Evolution, Biodiversité (ISYEB), UMR 7205 - Muséum National d’Histoire Naturelle, CNRS, SU, EPHE and UA, Paris, France

# Current address: Institut de Systématique, Evolution, Biodiversité (ISYEB), UMR 7205 - Muséum National d’Histoire Naturelle, CNRS, SU, EPHE and UA, Paris, France

\* luc.bassel@pasteur.fr (LB)

\* olivier.gascuel@mnhn.fr (OG)

<sup>^</sup> Membership list can be found in the acknowledgments section

## Abstract

Drug resistance mutations (DRMs) appear in HIV under treatment pressure. DRMs are commonly transmitted to naive patients. The standard approach to reveal new DRMs is to test for significant frequency differences of mutations between treated and naive patients. However, we then consider each mutation individually and cannot hope to study interactions between several mutations. Here, we aim to leverage the ever-growing quantity of high-quality sequence data and machine learning methods to study such interactions (i.e. epistasis), as well as try to find new DRMs.

We trained classifiers to discriminate between Reverse Transcriptase Inhibitor (RTI)-experienced and RTI-naive samples on a large HIV-1 reverse transcriptase (RT) sequence dataset from the UK ( $n \approx 55,000$ ), using all observed mutations as binary representation features. To assess the robustness of our findings, our classifiers were evaluated on independent data sets, both from the UK and Africa. Important representation features for each classifier were then extracted as potential DRMs. To find novel DRMs, we repeated this process by removing either features or samples associated to known DRMs. When keeping all known resistance signal, we detected sufficiently prevalent known DRMs, thus validating the approach. When removing features corresponding to known DRMs, our classifiers retained some prediction accuracy, and six new mutations significantly associated with resistance were identified. These six mutations have a low genetic barrier, are correlated to known DRMs, and are spatially close to either the RT active site or the regulatory binding pocket. When removing both known DRM features and sequences containing at least one known DRM, our classifiers lose all prediction accuracy. These results likely indicate that all mutations directly conferring resistance have been found, and that our newly discovered DRMs are accessory or compensatory mutations. Moreover, apart from the accessory nature of the relationships we found, we did not find any significant signal of further, more subtle epistasis combining several mutations which individually do not seem to confer any resistance.

## Author summary

Almost all drugs to treat HIV target the Reverse Transcriptase (RT) and Drug resistance mutations (DRMs) appear in HIV under treatment pressure. Resistant strains can be transmitted and limit treatment options at the population level. Classically, multiple statistical testing is used to find DRMs, by comparing virus sequences of treated and naive populations. However, with this method, each mutation is considered individually and we cannot hope to reveal any interaction (epistasis) between them. Here, we used machine learning to discover new DRMs and study potential epistasis effects. We applied this approach to a very large UK dataset comprising  $\approx 55,000$  RT sequences. Results robustness was checked on different UK and African datasets.

Six new mutations associated to resistance were found. All six have a low genetic barrier and show high correlations with known DRMs. Moreover, all these mutations are close

to either the active site or the regulatory binding pocket of RT. Thus, they are good candidates for further wet experiments to establish their role in drug resistance. Importantly, our results indicate that epistasis seems to be limited to the classical scheme where primary DRMs confer resistance and associated mutations modulate the strength of the resistance and/or compensate for the fitness cost induced by DRMs.

## 6.1. Introduction

Drug resistance mutations (DRMs) arise in Human Immunodeficiency Virus-1 (HIV-1) due to antiretroviral treatment pressure, leading to viral rebound and treatment failure [436, 437]. Furthermore, drug-resistant HIV strains can be transmitted to treatment-naïve individuals and further spread throughout the population over time [438, 423, 439]. These transmitted resistant variants limit baseline treatment options and have clinical and public health implications worldwide. Almost all drugs to treat HIV target the reverse transcriptase (RT), encoded by the *pol* gene. Lists of DRMs are regularly compiled and updated by experts in the field, based on genotype analyses and phenotypic resistance tests or clinical outcome in patients on ART [440, 441, 442]. However, with the development of new antiretroviral drugs that target RT but also other regions of the *pol* gene like protease or integrase, and the use of anti-retrovirals in high risk populations by pre-exposure prophylaxis (PREP), it is important to further our understanding of HIV polymorphisms and notably the interactions between mutations and epistatic effects.

Among known DRMs, some mutations, such as M184V, directly confer resistance to antiretrovirals, more precisely the commonly used NRTI, 3TC (lamivudine) and FTC (emtricitabine), and are called primary or major drug resistance mutations, while some mutations like E40F have an accessory role and increases drug resistance when appearing alongside primary DRMs. Moreover, some mutations like S68G seem to have a compensatory role, but are not known to confer any resistance nor modulate resistance induced by primary DRMs. All of these mutations might have different functions in the virus, but they are all known to be associated with drug resistance phenomena. Therefore, during the rest of this article we will refer to all of these known mutations as resistance associated mutations (RAMs), rather than DRMs which is too specific, and our goal will be to search for new RAMs and study the interactions between known RAMs and the new ones.

Classically, new RAMs have been found using statistical testing and large multiple sequence alignments (MSA) of the studied protein [443, 444]. Tests are performed for mutations of interest on a given MSA to check if they are associated with the treatment status and outcome of the individual the viral sequences were sampled from. The test significance is corrected for multiple testing as all mutations associated to every MSA position is virtually a resistance mutation and tested. After this preliminary statistical search, the selected mutations are scrutinized to remove the effects of phylogenetic correlation (i.e. typically counting two sequences which are identical or closely related due to

## CHAPTER 6

transmission rather than independent acquisition twice [445]) and check that the same mutation occurred several times in different subtypes and populations being treated with the same drug. Then, these mutations can be further experimentally tested in vitro or in vivo to validate phenotypic resistance. This method has worked well, but by design it is not ideal for studying the effect of several mutations at once, since if we have to test all couples or triplets of mutations, we quickly lose statistical power when correcting for multiple testing [446], due to the large number of tests to perform. Moreover, phylogenetic correlation is again a critical issue with such an approach.

Machine learning has been extensively used to predict resistance to antiretrovirals from sequence data. There are two main approaches to predicting resistance from sequence data. Regression, where machine learning models are trained to predict the value of a drug resistance indicator, typically  $IC_{50}$  fold change in response to a given drug [447] or other indicators from phenotypic resistance assays such as PhenoSense [448]. Many methods have been used to predict a resistance level: Support Vector Machines (SVMs) [449], k-Nearest Neighbors (KNN) and Random Forests (RFs) [450], and more recently Artificial Neural Networks (ANNs) [451, 452]. Alternatively, this task has also been approached as a classification problem. Given a certain threshold on a phenotypic resistance measure, sequences are given a label of "resistant" or "susceptible" to a certain drug. Machine learning classifiers are then trained to predict that label. For this task, SVMs and decision trees have been used [453, 454], ensemble classifier chains [455, 456] and also ANNs [457]. Most recently Steiner *et al.* [458] have used Deep Learning Architectures to predict resistance status (i.e. classification) from sequence data. Since phenotypic assays are more complicated and costly to perform than simple genotyping, there is a limited number of sequences paired with a resistance level. This is the main limitation of these studies since machine learning methods typically benefit from a large amount of training data. This is especially true for deep neural networks which can need hundreds of thousands of training samples for certain tasks and architectures. However, despite this limitation, approaches proposed in these studies seem to have fairly good predictive accuracy. It is important to note that all of these studies aim to predict if a given sequence is resistant or not to a given drug, they do not aim to find new potential RAMs. Although Steiner *et al.* [458] have checked that known DRM positions are captured by their models and found several positions potentially associated to resistance, it is not the main goal of their method.

It is accepted in machine learning that there is a trade-off between model accuracy and model interpretability. In these previous studies the goal was to make the most accurate predictions possible, using complex models such as SVMs and ANNs, therefore sacrificing interpretability. Here, we have a different approach, using simpler models that might be less accurate but whose predictions we can understand and interpret. We train these models to discriminate RTI-naive from RTI-experienced sequences. Without the need for phenotypic data, we are able to use much larger HIV-1 RT sequence datasets from the UK ( $n \approx 55,000$ ) (<http://www.hivrdb.org.uk/>) and Africa ( $n \approx 4,000$ ) [444]. By using interpretable models, we can extract mutations that are important for determining if a sequence is treated or not and potentially find new mutations potentially associated to resistance. Furthermore, we aim to detect associations between mutations and their

## 6.2. MATERIALS AND METHODS

effect on antiretroviral resistance in order to study potential underlying epistasis. The African and UK datasets are very different both from genetic and treatment history standpoints, therefore training classifiers on the UK dataset and testing them on the African one, should guarantee the robustness of our findings and greatly alleviate phylogenetic correlation effects. In the following sections, we first describe the data then the methods used. Our results include the assessment of the performance of our classifiers even when trained on data devoid of any known resistance-associated signal; as well as a description of the main features (prevalence and correlation to known mutations, genetic barrier and structural analysis) of six potentially resistance associated mutations, newly discovered thanks to our approach. These results and perspectives are discussed in the concluding section.

## 6.2. Materials and methods

### 6.2.1. Data

In this study, we used all the drug resistance mutations that appeared in the Stanford HIV Drug resistance database, both for NRTI (Nucleoside Reverse Transcriptase Inhibitors; <https://hivdb.stanford.edu/dr-summary/comments/NRTI/>) and NNRTI (Non Nucleoside RTI; <https://hivdb.stanford.edu/dr-summary/comments/NNRTI/>) as known RAMs. To discover new RAMs, assess their statistical significance and study potential epistatic effects, we used two datasets of HIV-1 RT sequences. A large one ( $n = 55,539$ ) from the UK HIV Drug Resistance Database (<http://www.hivrdb.org.uk/>) and a smaller ( $n = 3,990$ ) one from 10 different western, eastern and central African countries [444]. In the UK dataset, sequences from RTI-naive individuals formed the majority class with 41,921 sequences (75%). In the African dataset, both classes were more balanced with 2,316 RTI-naive sequences (58%). In the UK dataset, RTI-naive sequences had at least one known RAM in 25% of cases, most likely due to transmissions to naive patients or undisclosed treatment history, against 48% in RTI-experienced sequences, thus making the discrimination between the RTI-experienced and RTI-naive sequences particularly difficult. In the African dataset this distribution was more contrasted, with only 14% of RTI-naive sequences having at least one known RAM, versus 83% of RTI-experienced sequences. The African dataset was also much more genetically diverse with 24 different subtypes and CRFs compared to the 2 subtypes (B and C) that we retained for this study from the UK cohort. The majority of the sequences from the African dataset were samples from Cameroon (27%), Democratic Republic of Congo (17%), Burundi (15%), Burkina Faso (13%) and Togo (11%).

It is important to note that RTI-experienced sequences in both of these datasets can be considered as resistant to treatment. Since the viral load was sufficiently high to allow for sequencing of the virus, we can consider that the ART has failed. However, in some cases this resistance might be caused by non adherence to ART, rather than by the presence of RAMs, therefore adding some noise to the relationship between treatment

## CHAPTER 6

status and resistance.

In addition to differences in size, balance between RTI-naive and experienced classes, and the genetic difference between the UK and African datasets, there are also significant differences resulting from differing treatment strategies. In the UK and other higher income countries, the treatment is often tailored to the individual with genotype testing, which result in specific treatment as well as thorough follow-ups and high treatment adherence. In the African countries of the dataset that we used, the treatment is ZDV/ d4T (NRTI) + 3TC (NRTI) + NVP/EFV (NNRTI) in most cases [444], and this treatment is generalized to the affected population, with poorer follow-up and adherence than in the UK. This discrepancy could lead to different mutations arising in both datasets, however since the treatment strategy is a combination of both NRTI and NNRTI drug classes, as in many countries, similar RAMs arise [444]. Furthermore, there is potentially more uncertainty in the African dataset than in the UK. For example some individuals may have unofficially taken antiretroviral drugs, but still identify themselves as RTI-naive, or report having some form of ART while not having been treated for HIV [459]. All of this explains the high prevalence of multiple resistance in the African data set: the median number of RAMs in sequences containing at least one RAM is 3 in the African sequences, while it is 1 in UK sequences (Table 6.1). Thus, we can say that African sequences are highly resistant, with possibly different mutations and epistatic effects, compared to their UK counterparts.

All these differences between the two datasets helped us to assess the generalizability of our method and the robustness of the results. That is to say, if signal extracted from the UK dataset was still relevant on such a different dataset as the African one, we could be fairly reassured in regard to the biological and epidemiological relevance of the observed signal.

Sequences in both African and UK datasets were already aligned. In order to avoid overly gappy regions of our alignment we selected only positions 41 to 235 of RT for our analysis. We used the Sierra web service (<https://hivdb.stanford.edu/page/webservice/>) to get amino acid positions relative to the reference HXB2 HIV genome. This allowed us to determine all the amino acids present at each reference position in both datasets, among which we distinguished the “reference amino acids” for each position, corresponding to the B and C subtype reference sequences obtained from the Los Alamos sequence database (<http://www.hiv.lanl.gov/>). All the other, non-reference amino acids are named “mutations” in the following, and the set of mutations was explored to reveal new potential RAMs.

To train our supervised classification methods [460, 461, 462], the sequence data needed to be encoded to numerical vectors. A common and intuitive way to do so is to create a single feature in the dataset for each position of the sequence to encode. Each amino acid is then assigned an integer value, and an amino acid sequence is represented by a succession of integers corresponding to each amino acid. There is, however, one drawback with this method: by assigning an integer value to amino acids, we transform a categorical variable into an ordinal variable. Any ordering of amino acids is hard to

## 6.2. MATERIALS AND METHODS

		UK	Africa
size		55539	3990
RTI naive	with known RAMs	11429 (21%)	318 (8%)
	without known RAMs	30492 (55%)	1998 (50%)
RTI experienced	with known RAMs	6633 (12%)	1388 (35%)
	without known RAMs	6985 (13%)	286 (7%)
sequences with $\geq 2$ known RAMs		8034 (14%)	1308 (33%)
max known RAM number		13	17
Median known RAM number		1	3
number of subtypes / CRFs		2	24
subtypes / CRFs	A	0 (0%)	472 (12%)
	B	37806 (68%)	64 (2%)
	C	17733 (32%)	702 (18%)
	CRF02 AG	0 (0%)	1477 (37%)

Table 6.1.: **Summary of the UK and African datasets.**

Percentages are computed with regards to the size of the considered dataset (e.g. 21% of the sequences of the UK dataset are RTI-naive and have at least one known RAM). The median number of RAMs was computed only on sequences that had at least one known RAM.

justify and might introduce bias. To avoid this, we represented each sequence by a binary vector using one-hot encoding. For each position in the sequence to be encoded, amino acids corresponding to mutations are mapped to a binary vector denoting its presence or absence in the sequence. For example, at site 184, amino acids M, G, I, L, T and V are present in the UK dataset. After encoding we will have 5 binary features corresponding to the M184G, M184I, M184L, M184T and M184V mutations. We did not encode the reference amino acid M, but only the mutated amino acids. With this method each mutation in the dataset ( $n = 1,318$ ) corresponds to a single feature. Some of these features corresponded to known RAMs (e.g., M184I and M184V) and are named (known) RAM features in the following ( $n = 121$ ). This encoding allows the classifiers to consider specific mutations and potentially link them to resistance.

### 6.2.2. Classifier training

In order to find new potential RAMs, we first followed the conventional multiple testing approach [444]. We first used Fisher exact tests to identify which of these mutations were significantly associated with anti-retroviral treatment. All the resulting p-values were then corrected for multiple testing using the Bonferroni correction [463]. Those for which the corrected p-value was  $\leq 0.05$  were then considered as significantly associated with treatment and potentially implicated in resistance.

This method was complemented by our parallel, machine learning based approach. In order to extract potential RAMs, we trained several classifiers to discriminate between RTI-experienced and RTI-naive sequences represented by the binary vectors described above. This classification task does not need any phenotypic resistance measure, allowing us to use much larger and more readily available datasets than other machine learning based approaches previously mentioned. Once the classifiers were trained, we extracted the most important representation features, which corresponded to potentially resistance-associated mutations (PRAM in short). To this aim we chose three interpretable supervised learning classification methods so as to be able to extract those features:

1. Multinomial naive Bayes (NB), which estimates conditional probabilities of being in the RTI-experienced class given a set of representation features [464]; the higher ( $\approx 1.0$ ) and the lower ( $\approx 0$ ) conditional probabilities correspond to the most important features.
2. Logistic regression (LR) with L1 regularization (LASSO) [460] which assigns weights to each of the features, whose sign denotes the importance to one of the 2 classes, and whose absolute value denotes the weight of this importance.
3. Random Forest (RF) , which has feature importance measures based on the Gini impurity in the decision trees [465].

## 6.2. MATERIALS AND METHODS

Interpretability was the main driver behind our classification method choice, with the conditional probabilities of NB, the weight or LR and the importance values of RF, we can easily extract which mutations are driving the discrimination of RT sequences. This is why we did not choose to use ANNs which could have led to an increase in accuracy at the cost of interpretability [466, 467, 468]. Moreover, these three classification methods have the potential to detect epistatic effects. With RF, the discrimination is based on the combination of a few features (i.e. mutations), while with LR the features are weighted positively or negatively, thus making it possible to detect cumulative effects resulting from a large number of mutations, which individually have no discrimination power. Naive Bayes is a very simple approach, generally fairly accurate, and in between the two others in terms of explanatory power [462].

In order to be able to compare all these approaches in a common framework, we devised a very simple classifier out of the results of the Fisher exact tests. This "Fisher classifier" (FC) predicts a sequence as RTI-experienced if it has at least one of the mutations significantly associated to treatment. In this way, we were able to compute metrics for all classification methods and compare their performance.

It is important to note that in all of these approaches we chose to discriminate RTI-naive from RTI-experienced sequences, regardless of the type of RTI received. One of the reasons is that we did not have detailed enough treatment history for sequences in the UK and African datasets. Moreover, even without segmenting by treatment type, the size of the training set and the power of our classification methods were both high enough to be able to detect all kinds of resistance associated mutations. We shall see (Result section) that we were able to determine the likely treatment involved by further examining the important extracted features and comparing them to known RAMs. Furthermore, since the treatment strategies are so different between the UK and African sequences, training on sequences having received different treatments should increase the robustness of our classifiers and the relevance of the mutations selected as potentially associated to resistance.

To avoid phylogenetic confounding factors (e.g. transmitted mutations within a specific country or region), and avoid finding mutations potentially specific to a given subtype, we split the training and testing sets by HIV-1 M subtype. This resulted in training a set of classifiers on all subtype B sequences of the UK dataset and testing them on subtype C sequences from the UK dataset, training another set of classifiers on the subtype C sequences of the UK dataset and testing on the subtype B sequences from the UK dataset, as well as training a final set of classifiers on the whole UK dataset, but testing it on the smaller African dataset with a completely different phylogenetic makeup and treatment context [444]. Furthermore, in order to identify novel RAMs and study the behavior of the classifiers, we repeated this training scheme on both datasets, each time removing resistance-associated signal incrementally: first by removing all representation features corresponding to known RAMs from the dataset, and second by removing all sequences that had at least one known RAM. This resulted in each type of classifier being trained and tested 9 times, on radically different sets to ensure the interpretability and robustness of the results (see Table 6.2).

Signal removal level	Trained on		Tested on	
None	UK, subtype B	(37806)	UK, subtype C	(17733)
	UK, subtype C	(17733)	UK, subtype B	(37806)
	UK, subtypes B & C	(55539)	Africa, all subtypes	(3990)
Known RAM features removed	UK, subtype B	(37806)	UK, subtype C	(17733)
	UK, subtype C	(17733)	UK, subtype B	(37806)
	UK, subtypes B & C	(55539)	Africa, all subtypes	(3990)
Known RAM features & sequences with $\geq 1$ known RAM removed	UK, subtype B	(24422)	UK, subtype C	(13055)
	UK, subtype C	(13055)	UK, subtype B	(24422)
	UK, subtypes B & C	(37477)	Africa, all subtypes	(2284)

Table 6.2.: **All training and testing datasets used during this study.**

The number of sequences in each dataset is shown in parentheses

### 6.2.3. Measuring classifier performance

To compare the performance of our classifiers we used balanced accuracy [469], which is the average of accuracies (i.e. percentages of well-classified sequences) computed separately on each class of the test set. This score takes into account, and corrects for, the imbalance between RTI-naive and RTI-experienced samples, which would lead to a classifier always predicting a sequence as RTI-naive getting a classical accuracy score of up to 77% (i.e. the frequency of naive sequences in the UK dataset). We also computed the adjusted mutual information (AMI) between predicted and true sequence labels, which is a normalized version of MI allowing comparison of performance on differently sized test sets [143]. Additionally, mutual information (MI) was used to compute p-values and assess the significance of the classifiers' predictive power. The probabilistic performance of the classifiers was evaluated using an adapted Brier score [461] more suited to binary classification, which is the mean squared difference between the actual class (coded by 1 and 0 for the RTI-experienced and RTI-naive samples respectively) and the predicted probability of being RTI-experienced. This approach refines the standard accuracy measure by rewarding methods that well approximate the true status of the sample (eg. predicting a probability of 0.9 while the true status is 1); conversely, binary methods (predicting 0 or 1, but no probabilities) will be penalized if they are often wrong. The Brier approach thus assigns better scores to methods that recognize their ignorance than to methods producing random predictions.

## 6.3. Results

### 6.3.1. Classifier performance & interpretation

As can be seen in Fig 6.1A and 6.1B, when all RAM features and sequences were kept in the training and testing sets, classifiers had good prediction accuracy, with the machine learning classifiers slightly outperforming the “Fisher” classifier. When removing RAM features from the training and testing sets, the classifiers retained a significant prediction accuracy, especially with the African data set and its multiple RAMs that are observed in a large number of sequences (but removed in this experiment). In this configuration the ML classifiers had a similar performance to the “Fisher” classifier, except for the random forest that is slightly less accurate, likely due to overfitting. Also, when removing sequences that had known RAMs, every classifier lost all prediction accuracy, and none could distinguish RTI-naive from RTI-experienced sequences. Regarding the Brier score, we see the advantage of the machine learning classifiers over the “Fisher” classifier, which is worse than random predictions when known RAMs are removed. The ability of machine learning classifiers to quantify the resistance status should be an asset for many applications.

The fact that classifiers retained prediction accuracy after removing known RAM corresponding features suggests that there was some residual, unknown resistance-associated signal in the data. The fact that this same power was non-existent when removing the known RAM-containing sequences from the training and testing sets, indicates that this residual signal was contained in these already mutated sequences. This suggests that the mutations that are found in the RAM removed experiment (see list below) are most likely accessory mutations that accompany known RAMs. This also suggests that all primary DRMs (i.e., that directly confer antiretroviral resistance) have been identified, which is reassuring from a public health perspective.

The performance discrepancy between the UK and African test sets can be explained by several factors. Firstly, African sequences that have known RAMs are more likely to have multiple RAMs, and thus more (known and unknown) resistance-associated features than their UK counterparts (c.f. Table 6.1). This means that resistant African sequences are easier to detect even when removing known RAMs. Secondly, RTI-naive sequences in the UK test sets are more likely to have known RAMs than their African counterparts (c.f. Table 6.1) and therefore more companion mutations. This means that the RTI-naive sequences in the UK test set are more likely to be misclassified as RTI-experienced than in the African test set.

### 6.3.2. Additional classification results

The fact that, when looking at classifiers trained without known RAMs , “Fisher” classifiers perform as well as the machine learning ones, leads us to believe that there is little interaction between mutations that would explain resistance better than taking

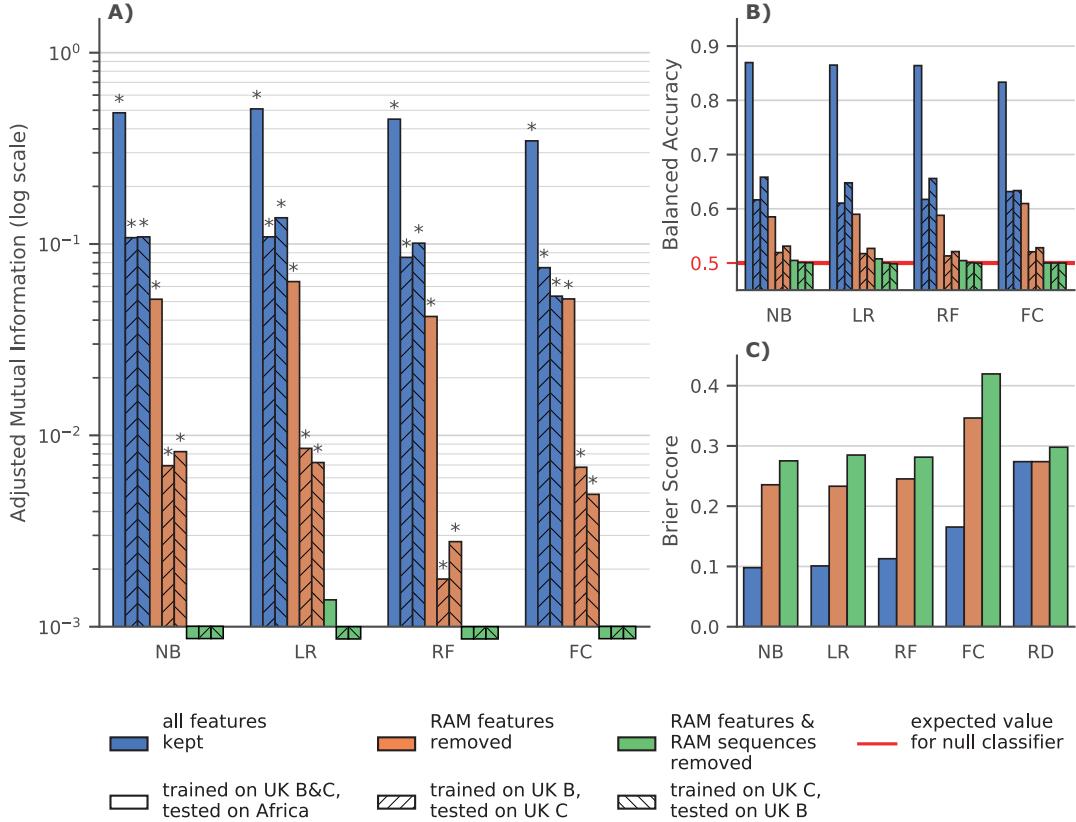


Figure 6.1.: **Classifier Performance on UK and African datasets.**

**NB:** naive Bayes, **LR:** Logistic Regression with Lasso regularization, **RF:** Random Forest, **FC:** Fisher Classifier, **RD:** Agnostic random probabilistic classifier (this classifier predicts, as the probability of a sample belonging to a class, the frequency of that class in the training data). **A)** Adjusted mutual information (higher is better) between ground truth and predictions by classifiers trained on dataset with all features (blue), without features corresponding to known RAMs (orange) and without RAM features and without sequences that have at least 1 known RAM (green). Hatching indicates the training set on which a classifier was trained and the testing set on which the performance was measured. The expected value for a null classifier is 0, and 1 for a perfect classifier and a \* denotes that the p-value derived from mutual information is  $\leq 0.05$ . For example when trained with all features all the classifiers have a significative MI. Conversely when removing RAM features and RAM sequences none of the classifiers have a significative MI and only LR trained on the entirety of the UK dataset has an  $AMI > 10^{-3}$  **B)** Balanced Accuracy score, i.e. average of accuracies per-class (higher is better) for the same classifiers as in a). The red line at  $y = 0.5$  is the expected balanced accuracy for a null classifier that only predicts the majority class as well as a random uniform (i.e. 50/50) classifier. **C)** Brier score, which is the mean squared difference between the sample's experience to RTI and the predicted probability of being RTI experienced (lower is better), for the same classifiers as in **A)** and **B)**.

## 6.3. RESULTS

each mutation separately. It is therefore likely that the kind of epistatic phenomena we were looking for, combining several mutations that do not induce any resistance when taken separately, do not come into play here. We are in a classical scheme where primary DRMs confer resistance and associated mutations reinforce the strength of the resistance and/or compensate for the fitness cost induced by primary DRMs.

It is important to remember that in the previous section we were trying (as usual, e.g. see [444]) to find novel mutations associated with resistance by discriminating RTI-naive from RTI-experienced sequences, both with the statistical tests and the classifiers. However, this is intrinsically biased and noisy. Indeed, a RTI-naive sequence is not necessarily susceptible to RTIs as a resistant strain could have been transmitted to the individual. Conversely, an RTI-experienced sequence may not be resistant to treatment, due to poor ART adherence for example. We must therefore keep in mind that the noisy nature of the relationship between resistance and treatment status is partly responsible for the lower performance of classifiers trained on the UK sequences with reduced signal. Moreover, as all the additional resistance signal we detected is associated to the sequences having at least one known RAM (see above), we performed another analysis trying to discriminate between the sequences having at least one known RAM and those having none. The goal was to check that the mutations we discovered by discriminating RTI-experienced from RTI-naive samples, are truly accessory and compensatory mutations. As can be seen in Fig 6.2A and 6.2B, the classifiers trained to discriminate sequences that have at least one known RAM from those that have none, on datasets from which all features corresponding to known RAMs were removed, perform much better than classifiers trained to discriminate RTI-experienced from RTI-naive sequences. This increase in performance is especially visible for classifiers tested on UK sequences (more difficult to classify than the African ones, see above), with an AMI often almost one order of magnitude higher for the known-RAM presence/absence classification task. This further reinforces our belief that all there is a fairly strong residual resistance-signal in sequences that contain known RAMs, due to new accessory and compensatory mutations identified by our classifiers and Fisher tests. As a side note, Logistic regression (LR) consistently outperforms other classifiers, a tendency already observed in Fig 6.1.

### 6.3.3. Identifying new mutations from classifiers

We assessed the importance of each mutation in the learned internal model of all the classifiers, in the setting where all known RAMs have been removed from the training dataset. For the Fisher classifier, we used one minus the p-value of the exact Fisher test as the importance value, therefore the more significantly associated mutations have the higher importance value and were ranked first. For a given classification task, we ranked each mutation according to the appropriate importance value for each classifier (see above), trained on the B or C subtypes, with the highest importance value having a rank of 0. We then computed the average rank for each mutation and each classification task (RTI-naive/RTI-experienced and RAM present/RAM absent). This gave us, for each classification task, a ranking of mutations potentially associated with resistance

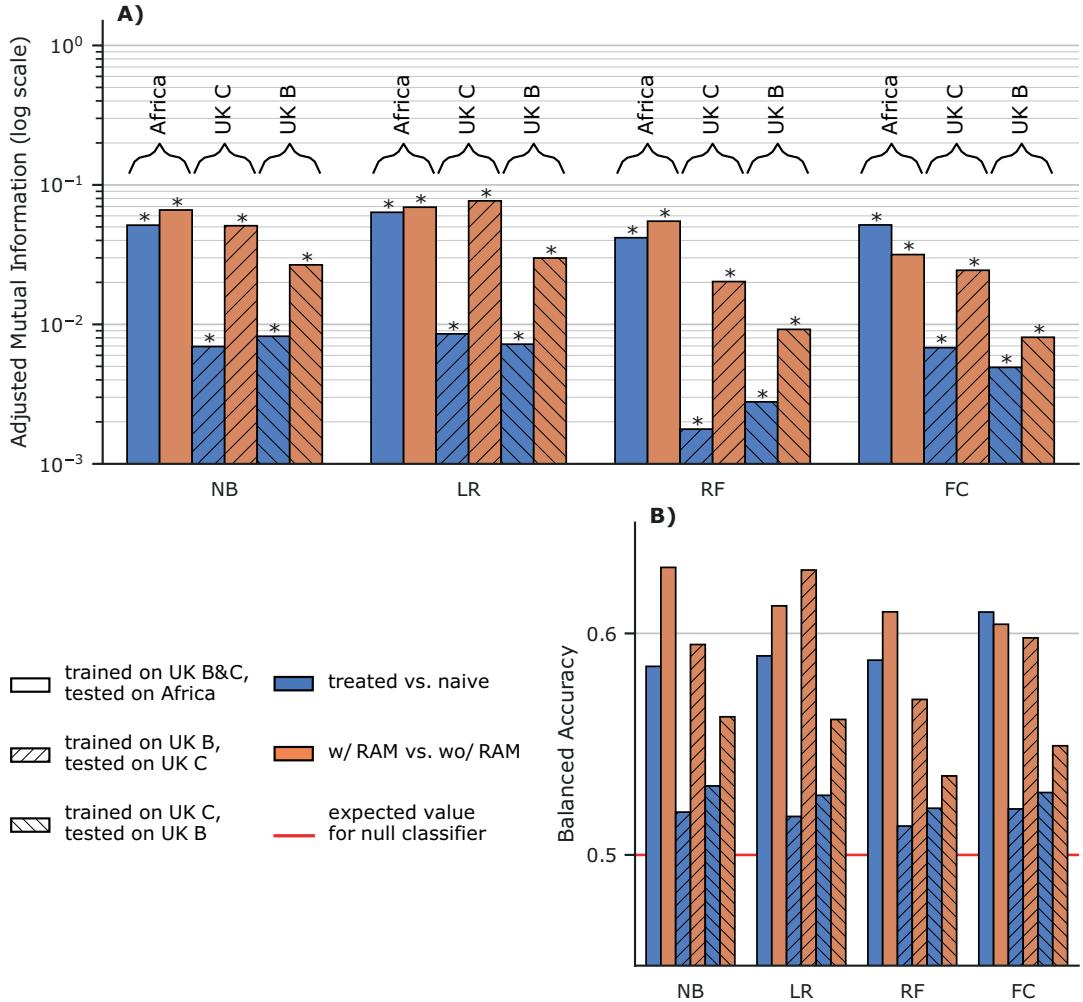


Figure 6.2.: Discrimination between sequences having at least one RAM, and those having none on sequences with training features corresponding to known RAMs removed.

**NB:** naive Bayes, **LR:** Logistic Regression with Lasso regularization, **RF:** Random Forest, **FC:** Fisher Classifier. **A)** Adjusted mutual information (higher is better) for classifiers trained without features corresponding to known RAMs. The classifiers are either trained to discriminate RTI-naive from RTI-experienced sequences (blue), or sequences with at least one known RAM from sequences that have none (orange). Hatching and braced annotations indicate the training and testing sets resulting in a given performance measure. **B)** Balanced accuracy, i.e. average of accuracies per-class for the same classifiers as in **A)** (higher is better). The red line at  $y = 0.5$  is the expected value for a classifier only predicting the majority class as well as a random uniform (50/50) classifier.

## 6.3. RESULTS

that took into account the importance given to this new mutation by each classifier trained on this task. Mutations that were in the 10 most important mutations for both of the classification tasks were considered of interest. Based on these criteria we selected the following potentially resistance-associated mutations (w.r.t. the HXB2 reference genome): L228R, L228H, E203K, D218E, I135L and H208Y. These mutations are referred to as “new mutations” in the rest of this study.

To check the epistatic nature of these selected mutations we computed the relative risk  $RR(new, X)$  between a new mutation and a binary character  $X$ .  $RR(new, X)$  was computed from the contingency table between  $new$  and  $X$  as follows:

	X present	X absent	$RR(new, X) = \frac{A}{A+C} \div \frac{B}{B+D}$
new present	A	B	
new absent	C	D	

The RR gives us a measure for how over-represented each of our new mutations is in sequences that have the  $X$  character compared to those that don’t.

To get a general idea of this over-representation, for each new mutation we computed  $RR(new, treatment)$  comparing the prevalence of the new mutation in RTI-experienced and RTI-naive sequences. We also computed  $RR(new, withRAM)$  comparing the prevalence the new mutation in sequences having at least one known RAM and sequences that have none. Both of these RRs are shown in Table 6.3 for each new mutation.

We then computed  $RR(new, RAM)$  for each known RAM present in more than 0.1% of UK sequences and the new mutations. In Fig 6.3 we see the RRs for which the lower bound of the 95% confidence interval, computed on 1000 bootstrap samples from the UK dataset, was greater than 4.

### 6.3.4. Detailed analysis of potentially resistance-associated mutations

As can be seen in Table 6.3, all of these new mutations except for I135L, are highly over-represented in RTI-experienced sequences and sequences that already have known RAMs, with lower bounds on the 95% RR CI always greater than 5, and often exceeding 10. When looking at the RRs computed for individual RAMs on the UK dataset (Fig 6.3), this impression is confirmed with very high over-representation of these new mutations potentially associated with resistance in sequences that have a given known RAM, with 95% RR lower CI bounds sometimes greater than 80 (H208Y/L210W and D218E/D67N), and most of the time greater than 10. with the noticeable exception of I135L where only 2 known RAMs give RRs with lower CI bounds greater than 4. The RRs computed on the African dataset (C.1) tell a similar story albeit with smaller RR values due to a smaller number of occurrences of both new mutations and known RAMs.

	codon distance		UK		$RR(new, X)$		p-value
	min	avg	B62	count	<i>treatment</i>	<i>any RAM</i>	
<b>L228R</b>	1	1.16	-2	227 (0.4%)	18.1 [12.9;27.3]	115.7 [55.1;507.3]	$3.4 \cdot 10^{-31}$
<b>E203K</b>	1	1.31	1	256 (0.5%)	11 [8.2;15.1]	20.1 [13.7;32.1]	$1.1 \cdot 10^{-14}$
<b>D218E</b>	1	1	2	168 (0.3%)	13.1 [9.0;19.6]	27 [16.3;57.0]	$3.3 \cdot 10^{-10}$
<b>L228H</b>	1	1.12	-3	287 (0.5%)	6.4 [5.1;8.4]	9.2 [6.9;12.6]	$4.4 \cdot 10^{-16}$
<b>I135L</b>	1	1.16	2	540 (1.0%)	1.8 [1.5;2.1]	2.4 [2.0;2.8]	$5.9 \cdot 10^{-08}$
<b>H208Y</b>	1	1.10	2	205 (0.4%)	8.8 [6.5;12.5]	14.9 [9.9;23.6]	$1.2 \cdot 10^{-05}$
<b>RAMs</b>	1 [1;2]	1.35 [1;2.44]	0 [-2;3]	58 (0.1%) [2;1842]	8.3 [0.6; $\infty$ ]	26.4 [1.4; $\infty$ ]	$3.1 \cdot 10^{-2}$ $[2.3 \cdot 10^{-58};1]$

Table 6.3.: Analysis of new potential RAMs.

**Codon distance:** For each new mutation we computed the minimum number of nucleotide mutations to go from the wild amino acid codons to those of the mutated amino acid, as well as the average codon distance between both amino acids, weighted by the prevalence of each wild and mutated codon at the given position in the UK dataset. **B62:** BLOSUM62 similarity values (e.g. D218E = 2, reflecting that E and D are both negatively charged and highly similar). **Count:** We looked at the number of occurrences of each new potential RAM in the UK dataset and the corresponding prevalence in parentheses. **Relative risks:** We computed  $RR(new, treatment)$  (e.g. L228R is 18.1 times more prevalent in RTI-experienced sequences compared to RTI-naive sequences in the UK dataset). We also computed  $RR(new, any RAM)$  (e.g. L228R is 115.7 times more prevalent in sequences that have at least one known RAM than in sequences that have none in the UK dataset). The 95% confidence intervals shown under each RR were computed with 1000 bootstrap samples of size  $n = 55,000$  drawn with replacement from the whole UK dataset. **p-values:** Fisher exact tests were done on the African dataset (to avoid confounding effects due to phylogenetic correlation) to see if each of these new mutations were more prevalent in RTI-experienced sequences. The same metrics were computed for all known RAMs, the median values are shown in the last two lines of this table, as well as the 5<sup>th</sup> and 95<sup>th</sup> percentiles which are shown underneath.  $RR(RAM, any RAM)$  values were computed for any RAM except itself to avoid always having infinite ratios.

### 6.3. RESULTS

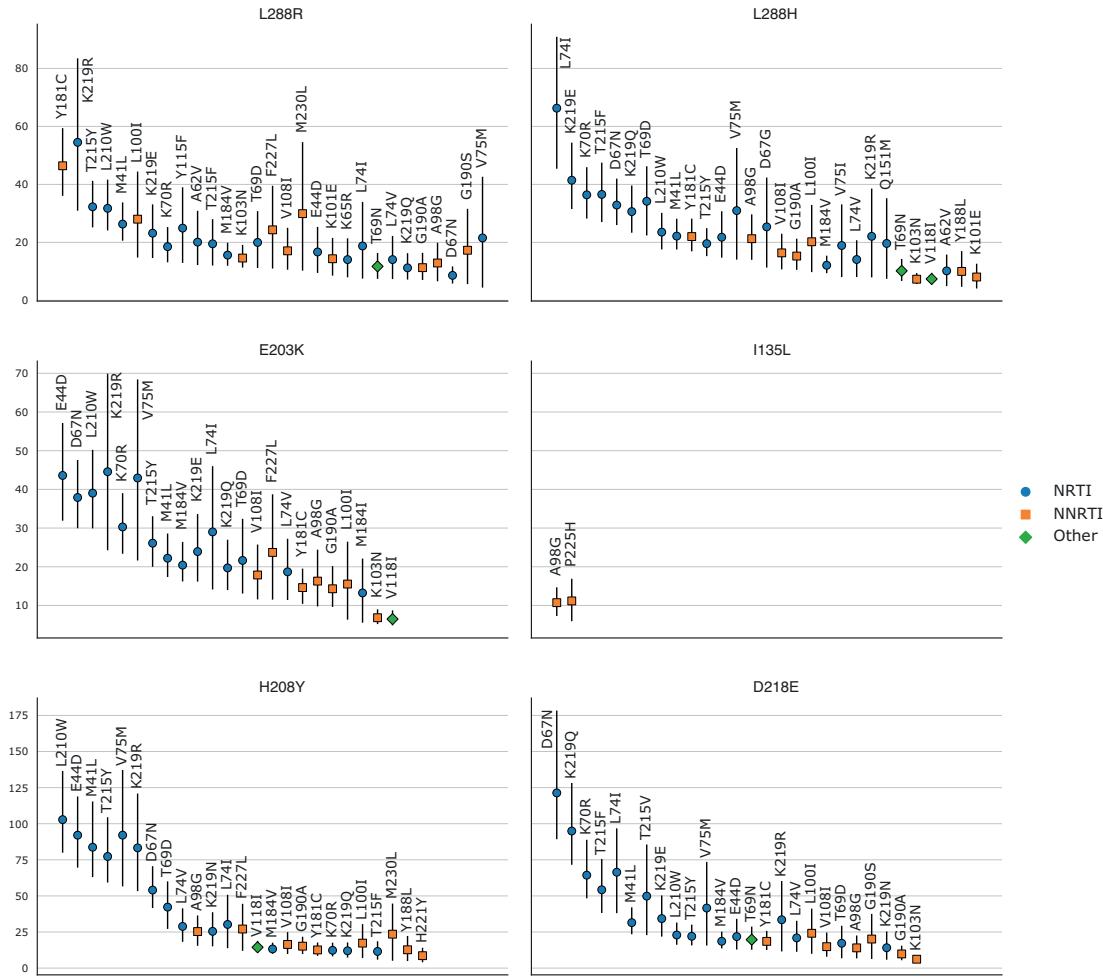


Figure 6.3.: **Relative risk of the new mutations with regards to known RAMs on the UK dataset**

(i.e. the prevalence of the new mutation in sequences with a given known RAM divided by the prevalence of the new mutation in sequences without this RAM). RRs were only computed for mutations (new and RAMs) that appeared in at least 0.1% (=55) sequences. 95% confidence intervals, represented by vertical bars, were computed with 1000 bootstrap samples of UK sequences. Only RRs with a lower CI boundary greater than 4 are shown. The shape and color of the point represents the type of RAM as defined by Stanford's HIVDB. Blue circle: NRTI, orange square: NNRTI, green diamond: Other. RR values are shown from left to right, by order of decreasing values on the lower bound of the 95% CI.

## CHAPTER 6

The genetic barrier to resistance for each of these new mutations is quite low, with a minimum of 1 base change for each of them (Table 6.3). We also computed the average codon distance (i.e. number of different bases), weighted by the prevalence of wild and mutated codons at the given positions in the UK (Table 6.3) and Africa (Table C.5) datasets, and in each case the average codon distance was always close to 1. In other words, at the amino acid level these mutations are expected to be relatively frequent. However, their frequencies are much higher in treated/with-RAM sequences than in naive/without-RAM ones (Table 6.3). Moreover, if we look at the BLOSUM62 scores (Table 6.3), some of these mutations induce some substantial changes in physicochemical properties, most notably at site 228, which reinforces again the likelihood that these mutations are associated with resistance. These metrics were also computed for all known RAMs (Table 6.3). For all these metrics, and the 6 new potential RAMs, values are contained between the 5<sup>th</sup> and 95<sup>th</sup> percentiles computed on known RAMs, except for the BLOSUM score of L228H that corresponds to a drastic physicochemical change. To gain more insight on these new mutations we also observed their spatial location on the 3-D HIV-1 RT structure using PyMol [470]. HIV-1 RT is a heterodimer with two subunits translated from the same sequence with different lengths and 3-D structures. The smaller p51 subunit (440 AAs) has a mainly structural role, while the larger p66 (560 AAs) subunit has the active site at positions 110, 185 and 186. The p66 subunit also has a regulatory pocket behind the active site: the non-nucleoside inhibitor binding pocket (NNIBP) formed of several sites of the p66 subunit as well as site 138 of the p51 subunit. Nucleoside RT Inhibitors (NRTI) are nucleotide analogs and bind in the active site, blocking reverse transcription. Non-Nucleoside RT Inhibitors (NNRTI) bind in the NNIBP, changing the protein conformation and blocking reverse transcription. More details on the structure and function of HIV-1 RT can be found in [377]. A general view of where the new mutations are situated with regards to the other important sites of HIV-1 RT is shown in Fig 6.4, and is detailed below.

### 6.3.4.1. L228R / L228H

L228R is the most important of these new mutations according to the feature importance ranking done above. This is reflected in the very high over-representation in RTI-experienced sequences and sequences with known RAMs shown in Table 6.3. When looking at the detailed RRs shown in Fig 6.3, we observe that L228R presents high RR values with mainly NRTI RAMs, but also with NNRTI RAMs such as Y181C and L100I, and this is even more so for RRs computed on the African dataset (C.1). L228H is very similar in all regards to L228R, however its highest RRs are exclusively with NRTI RAMs.

Site 228 of the p66 subunit is located very close to the active site of RT, where NRTIs operate (Figs 6.4 and C.3) which could explain the role that L228R and L228H seem to have in NRTI resistance. However, site 228 of the p66 subunit is also between sites 227 and 229 which are both part of the NNIBP. Furthermore, both L228H and L228R have very low BLOSUM62 score, of -3 and -2 respectively (Table 6.3). Arginine (R) and

### 6.3. RESULTS

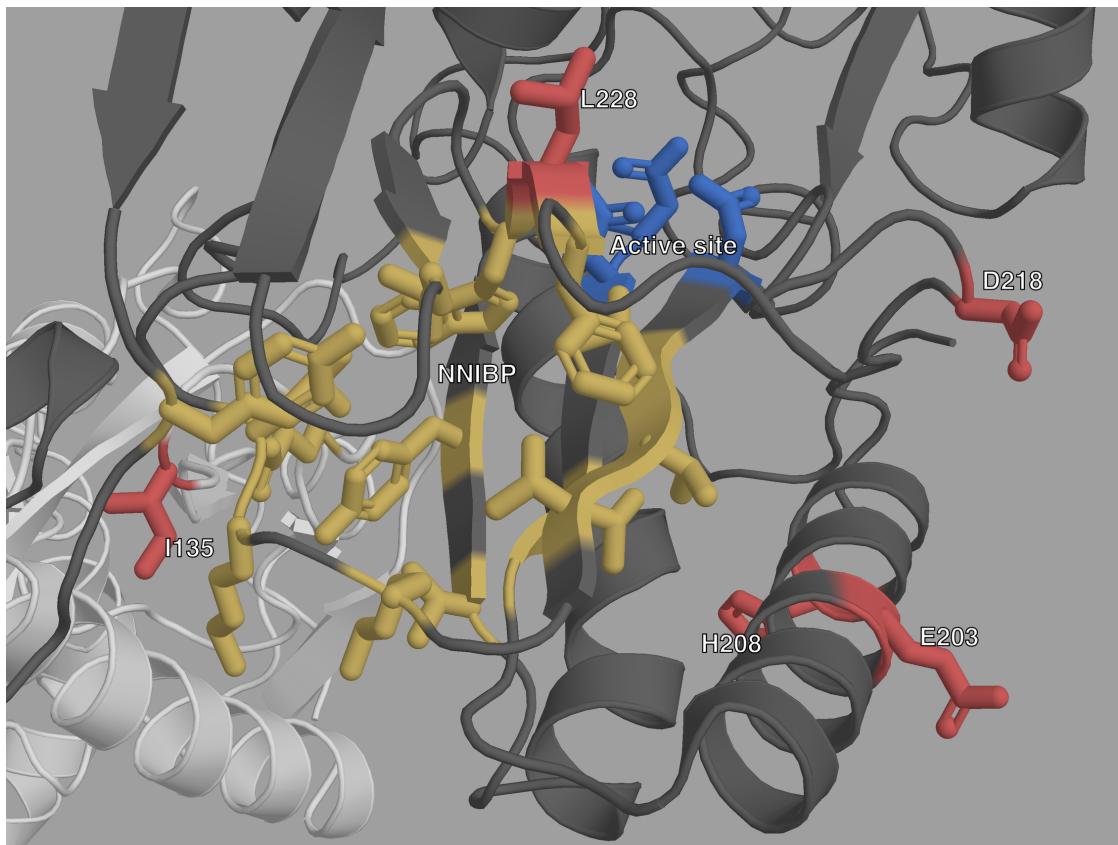


Figure 6.4.: Structure of HIV-1 RT with highlighted important sites.

The p66 subunit is colored dark gray and the p51 subunit white. The active site is highlighted in blue, and the NNIBP is highlighted in yellow. The sites of new mutations are colored in red.

## CHAPTER 6

Histidine (H) are both less hydrophobic than Leucine (L), and have positively charged side-chains. This important change in physicochemical properties could explain the role they both seem to have in NRTI resistance. However, while both Arginine and Histidine are larger than Leucine, Arginine is also fairly larger than Histidine, which is aromatic. This difference between both residues might explain the association L228R seems to have with NNRTI resistance that L228H does not have.

### 6.3.4.2. E203K / H208Y

Both E203K and H208Y are highly over-represented in RTI-experienced sequences and sequences with known RAMs. They both have high RR values for NRTI RAMs. Furthermore the most highly valued RAM RRs in Fig 6.3, are very similar for E203K and H208Y. Structurally they are close to each other on an alpha helix which is close to the active site.

Both E203K and H208Y have positive, albeit not maximal, BLOSUM62 scores, meaning they are fairly common substitutions. However, these mutations induce some change in physicochemical properties with Tyrosine (Y) being less polar than Histidine (H), and the change from Glutamic Acid (E) to Lysine (K) corresponding to a change from a negatively charged side chain to a positively charged one.

All this, combined with their structural proximity and the shared high RR values for single RAMs, suggests a similar role in NRTI resistance.

### 6.3.4.3. I135L

In Table 6.3 and Fig 6.3, we observe that I135L has the lowest RR values of all the new mutations, with CI bounds lower than 2 in Table 6.3's general RRs. However, it is the most prevalent of the new mutations. If we look at the detailed RRs of Fig 6.3, we see that I135L is significantly over-represented in sequences with NNRTI RAMs, specifically A98G and P225H. Structurally this makes sense: On the p66 subunit, site 135 is on the outside, far from both the active site and the NNIBP. However, site 135 on the p51 subunit is located very close to the NNIBP (Figs 6.3 and C.2).

The BLOSUM62 score for this substitution is quite high (Table 6.3), which is expected since both residues are very similar to one another, differing only by the positioning of one methyl group. However, Leucine (L) is less hydrophobic than Isoleucine (I), despite they are still both classified as hydrophobic residues (Table C.5).

The proximity between site 135 and the pocket in which NNRTI RAMs bind, as well as the high RR values for these NNRTI RAMs leads us to believe that I135L could play a subtle accessory role in NNRTI resistance, either by enhancing the effect of some NNRTI RAMs (typically, A98G and P225H), or by compensating for loss of fitness.

### 6.3.4.4. D218E

D218E is also highly over-represented in both RTI-experienced sequences and sequences with known RAMs. It has infinite RR values in the African dataset (Table 6.3), because it is quite rare in this dataset, and all of its 25 occurrences are in sequences that have at least one known RAM and are RTI-experienced. In fact, from the UK dataset we can see that D218E has some of the highest RR values for individual RAMs (along with H208Y). The majority of these very high RR values occur for NRTI RAMs. Site 218 on the p66 subunit is quite close to the RT active site, which could explain the role D218E seems to have in NRTI resistance. Aspartic acid (D) and Glutamic acid (E) are very similar amino acids, both acidic with negatively charged side-chains, as reflected in their fairly high BLOSUM62 score, the main difference between both being molecular weight, with E being slightly larger than D.

## 6.4. Discussion and perspectives

Our method has allowed us to identify six mutations that might play a role in drug resistance in HIV. These mutations are significantly over-represented in RTI-experienced sequences, as well as sequences exhibiting at least one other known RAM. The fact that models trained on the UK are still performant on such a different dataset as the African one strongly suggests that the learned classifier models have acquired generalized knowledge on resistance. For all of these new mutations their spatial positioning on HIV-1 RT is consistent with our conclusions, as all were either close to the active site or the regulatory binding pocket.

Some of the mutations we have identified as potentially associated with resistance have been mentioned in previous studies. L228R/H have been observed before [471] and were suggested to be associated with reduced susceptibility to didanosine [472, 473]. I135L has been observed in sequences with reduced susceptibility to NNRTIs [474]. H208Y has been associated with NNRTI and NRTI resistance [475] and it has been suggested that it has an accessory role in NRTI resistance [476]. E203K, D218E, L228RH and H208Y have all been mentioned in [477] as probably linked to phenotypic resistance to NRTI and NNRTI.

However, none of these mutations has been experimentally confirmed as conferring or helping with drug resistance to the best of our knowledge. The fact that we find them again with a big data analysis of highly different sequences and involved statistical selection procedure combining multiple testing and machine learning, and that we have very high significance, clearly indicates their potential role in resistance. Therefore, we believe they are sufficiently linked to drug resistance that they garner a closer inspection either in-vitro or in-vivo to determine the mechanisms that could allow them to play a role in resistance.

With our machine classifiers we seem to have found some RAMs of an accessory nature,

## CHAPTER 6

over-represented in sequences already containing known RAMs. This is a form of epistasis, where the interaction between the main RAM and the accessory RAM is important. However, we did not manage to find subtler forms of epistasis, in our dataset, where two mutations separately have no effect on resistance but have an effect together. This is partly indicated by the fact that there is a limited performance gap between the Fisher exact tests and more sophisticated classifiers, that are able to reveal significant association of mutations, while each individual mutation has low prediction power. However, one advantage of machine learning classifiers, is that they are probabilistic, meaning that they can give more nuanced insights into the nature or resistance level of a given sequence than the classical binary presence/absence of RAMs approach. In this regard logistic regression appears as a method of choice, showing similar or better performance than other classifiers, and an easy interpretation that is facilitated by the lasso regularization which performs a simple feature selection and retains the most important ones. Similar results were already observed on other sequence analysis tasks [478]. In order to investigate the second form of epistasis further we tested each pair of mutations in the UK dataset ( $n = 867,903$ ) with Fisher exact tests to see if they were linked to treatment status. In order to mitigate the effects of phylogenetic correlation which are sure to have an effect in this type of setting, we tested the pairs that were significantly associated to treatment ( $n = 1,309$ ) again on the African dataset. We also compared these results to the Fisher exact tests executed for each single mutation. We did not find any pair of mutations that was significantly associated, to treatment where neither member were significantly associated individually. Moreover, we only found 3 significantly associated pairs of mutations that did not include at least one known RAM, and they all included one of our newly found potential RAM: L228R + I142V, L228R + F214L and L228H + F214L (see appendix C.6 for details).

With therapeutic strategies targeting multiple proteins that are now used, there might be some epistatic effects with other regions of the HIV genome that are targeted by some of the drugs. These potential effects however, lie outside the scope of this study.

Because of the lack of detailed treatment history metadata, we did not distinguish mutations arising from NRTIs or NNRTIs. We believe that a large amount of high quality sequence data, along with a sufficiently detailed log of treatments and drugs the sequences were exposed to, could allow us to use our machine-learning approach to find mutations related to specific drugs and thus furthering our knowledge of HIV drug resistance, giving clinicians more tools to manage and help infected patients.

## Acknowledgments

We thank Anna Zhukova, Frédéric Lemoine and Marie Morel for their help and suggestions.

We also thank the UK HIV Drug Resistance Database and the UK Collaborative HIV Cohort:

#### 6.4. DISCUSSION AND PERSPECTIVES

**Steering committee:** David Asboe, Anton Pozniak (Chelsea & Westminster Hospital, London); Patricia Cane (Public Health England, Porton Down); David Chadwick (South Tees Hospitals NHS Trust, Middlesbrough); Duncan Churchill (Brighton and Sussex University Hospitals NHS Trust); Simon Collins (HIV i-Base, London); Valerie Delpech (National Infection Service, Public Health England); Samuel Douthwaite (Guy's and St. Thomas' NHS Foundation Trust, London); David Dunn, Kholoud Porter, Anna Tostevin, Oliver Stirrup (Institute for Global Health, UCL); Christophe Fraser (University of Oxford); Anna Maria Geretti (Institute of Infection and Global Health, University of Liverpool); Rory Gunson (Gartnavel General Hospital, Glasgow); Antony Hale (Leeds Teaching Hospitals NHS Trust); Stéphane Hué (London School of Hygiene and Tropical Medicine); Michael Kidd (Public Health England, Birmingham Heartlands Hospital); Linda Lazarus (Expert Advisory Group on AIDS Secretariat, Public Health England); Andrew Leigh-Brown (University of Edinburgh); Tamyo Mbisa (National Infection Service, Public Health England); Nicola Mackie (Imperial NHS Trust, London); Chloe Orkin (Barts Health NHS Trust, London); Eleni Nastouli, Deenan Pillay, Andrew Phillips, Caroline Sabin (University College London, London); Kate Templeton (Royal Infirmary of Edinburgh); Peter Tilston (Manchester Royal Infirmary); Erik Volz (Imperial College London, London); Ian Williams (Mortimer Market Centre, London); Hongyi Zhang (Addenbrooke's Hospital, Cambridge).

**Coordinating Center:** Institute for Global Health, UCL (David Dunn, Keith Fairbrother, Anna Tostevin, Oliver Stirrup)

**Centers contributing data:** Clinical Microbiology and Public Health Laboratory, Addenbrooke's Hospital, Cambridge (Justine Dawkins); Guy's and St Thomas' NHS Foundation Trust, London (Emma Cunningham, Jane Mullen); PHE – Public Health Laboratory, Birmingham Heartlands Hospital, Birmingham (Michael Kidd); Antiviral Unit, National Infection Service, Public Health England, London (Tamyo Mbisa); Imperial College Health NHS Trust, London (Alison Cox); King's College Hospital, London (Richard Tandy); Medical Microbiology Laboratory, Leeds Teaching Hospitals NHS Trust (Tracy Fawcett); Specialist Virology Centre, Liverpool (Elaine O'Toole); Department of Clinical Virology, Manchester Royal Infirmary, Manchester (Peter Tilston); Department of Virology, Royal Free Hospital, London (Clare Booth, Ana Garcia-Diaz); Edinburgh Specialist Virology Centre, Royal Infirmary of Edinburgh (Lynne Renwick); Department of Infection & Tropical Medicine, Royal Victoria Infirmary, Newcastle (Matthias L Schmid, Brendan Payne); South Tees Hospitals NHS Trust, Middlesbrough (David Chadwick); Department of Virology, Barts Health NHS Trust, London (Mark Hopkins); Molecular Diagnostic Unit, Imperial College, London (Simon Dustan); University College London Hospitals (Stuart Kirk); West of Scotland Specialist Virology Laboratory, Gartnavel, Glasgow (Rory Gunson, Amanda Bradley-Stewart).

## Supporting Information

Supporting Information can be found in the appendix C

### References for chapter 6

- [143] Nguyen Xuan Vinh, Julien Epps, and James Bailey. “Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance”. en. In: *Journal of Machine Learning Research* 11 (2010), p. 18 (cit. on pp. 42, 134, 171).
- [377] Stefan G. Sarafianos, Bruno Marchand, et al. “Structure and Function of HIV-1 Reverse Transcriptase: Molecular Mechanisms of Polymerization and Inhibition”. In: *Journal of molecular biology* 385.3 (Jan. 2009), pp. 693–713. ISSN: 0022-2836. DOI: [10.1016/j.jmb.2008.10.071](https://doi.org/10.1016/j.jmb.2008.10.071) (cit. on pp. 99, 142).
- [423] Raphaël Mourad, François Chevennet, et al. “A Phylotype-Based Analysis Highlights the Role of Drug-Naive HIV-Positive Individuals in the Transmission of Antiretroviral Resistance in the UK”. ENGLISH. In: *Aids* 29.15 (Sept. 2015), pp. 1917–1925. ISSN: 0269-9370. DOI: [10.1097/QAD.0000000000000768](https://doi.org/10.1097/QAD.0000000000000768) (cit. on pp. 102, 127).
- [436] Alessandro Cozzi Lepri, Caroline A. Sabin, et al. “Resistance Profiles in Patients with Viral Rebound on Potent Antiretroviral Therapy”. en. In: *The Journal of Infectious Diseases* 181.3 (Mar. 2000), pp. 1143–1147. ISSN: 0022-1899. DOI: [10.1086/315301](https://doi.org/10.1086/315301) (cit. on p. 127).
- [437] Chris Verhofstede, Filip Van Wanzele, et al. “Detection of Drug Resistance Mutations as a Predictor of Subsequent Virological Failure in Patients with HIV-1 Viral Rebounds of Less than 1,000 RNA Copies/ML”. en. In: *Journal of Medical Virology* 79.9 (2007), pp. 1254–1260. ISSN: 1096-9071. DOI: [10.1002/jmv.20950](https://doi.org/10.1002/jmv.20950) (cit. on p. 127).
- [438] Stéphane Hué, Robert J. Gifford, et al. “Demonstration of Sustained Drug-Resistant Human Immunodeficiency Virus Type 1 Lineages Circulating among Treatment-Naïve Individuals”. en. In: *Journal of Virology* 83.6 (Mar. 2009), pp. 2645–2654. ISSN: 0022-538X, 1098-5514. DOI: [10.1128/JVI.01556-08](https://doi.org/10.1128/JVI.01556-08) (cit. on p. 127).
- [439] Anna Zhukova, Teresa Cutino-Moguel, et al. “The Role of Phylogenetics as a Tool to Predict the Spread of Resistance”. en. In: *The Journal of Infectious Diseases* 216.suppl\_9 (Dec. 2017), S820–S823. ISSN: 0022-1899. DOI: [10.1093/infdis/jix411](https://doi.org/10.1093/infdis/jix411) (cit. on p. 127).
- [440] Diane E. Bennett, Ricardo J. Camacho, et al. “Drug Resistance Mutations for Surveillance of Transmitted HIV-1 Drug-Resistance: 2009 Update”. en. In: *PLOS ONE* 4.3 (Mar. 2009), e4724. ISSN: 1932-6203. DOI: [10.1371/journal.pone.0004724](https://doi.org/10.1371/journal.pone.0004724) (cit. on p. 127).

## REFERENCES FOR CHAPTER 6

- [441] Jennifer Hammond, Charles Calef, et al. “Mutations in Retroviral Genes Associated with Drug Resistance”. en. In: *Human retroviruses and AIDS* (Dec. 1998), pp. 11136–11179 (cit. on p. 127).
- [442] A. M. Wensing, V. Calvez, et al. “2017 Update of the Drug Resistance Mutations in HIV-1., 2017 Update of the Drug Resistance Mutations in HIV-1”. eng. In: *Topics in antiviral medicine, Topics in Antiviral Medicine* 24, 24.4, 4 (Dec. 2016), pp. 132, 132–133. ISSN: 2161-5861 (cit. on p. 127).
- [443] Sandrine Dudoit and Mark J. van der Laan. *Multiple Testing Procedures with Applications to Genomics*. en. Springer Science & Business Media, Dec. 2007. ISBN: 978-0-387-49317-6 (cit. on p. 127).
- [444] Christian Julian Villabona-Arenas, Nicole Vidal, et al. “In-Depth Analysis of HIV-1 Drug Resistance Mutations in HIV-Infected Individuals Failing First-Line Regimens in West and Central Africa”. en-US. In: *AIDS* 30.17 (Nov. 2016), p. 2577. ISSN: 0269-9370. DOI: [10.1097/QAD.0000000000001233](https://doi.org/10.1097/QAD.0000000000001233) (cit. on pp. 127–130, 132, 133, 137).
- [445] Wayne P. Maddison and Richard G. FitzJohn. “The Unsolved Challenge to Phylogenetic Correlation Tests for Categorical Characters”. en. In: *Systematic Biology* 64.1 (Jan. 2015), pp. 127–136. ISSN: 1063-5157. DOI: [10.1093/sysbio/syu070](https://doi.org/10.1093/sysbio/syu070) (cit. on p. 128).
- [446] Pak C. Sham and Shaun M. Purcell. “Statistical Power and Significance Testing in Large-Scale Genetic Studies”. en. In: *Nature Reviews Genetics* 15.5 (May 2014), pp. 335–346. ISSN: 1471-0064. DOI: [10.1038/nrg3706](https://doi.org/10.1038/nrg3706) (cit. on p. 128).
- [447] Thomas Lengauer and Tobias Sing. “Bioinformatics-Assisted Anti-HIV Therapy”. en. In: *Nature Reviews Microbiology* 4.10 (Oct. 2006), pp. 790–797. ISSN: 1740-1534. DOI: [10.1038/nrmicro1477](https://doi.org/10.1038/nrmicro1477) (cit. on p. 128).
- [448] Jie Zhang, Soo-Yon Rhee, et al. “Comparison of the Precision and Sensitivity of the Antivirogram and PhenoSense HIV Drug Susceptibility Assays”. en-US. In: *JAIDS Journal of Acquired Immune Deficiency Syndromes* 38.4 (Apr. 2005), pp. 439–444. ISSN: 1525-4135. DOI: [10.1097/01.qai.0000147526.64863.53](https://doi.org/10.1097/01.qai.0000147526.64863.53) (cit. on p. 128).
- [449] Niko Beerenwinkel, Martin Däumer, et al. “Geno2pheno: Estimating Phenotypic Drug Resistance from HIV-1 Genotypes”. In: *Nucleic Acids Research* 31.13 (July 2003), pp. 3850–3855. ISSN: 0305-1048. DOI: [10.1093/nar/gkg575](https://doi.org/10.1093/nar/gkg575) (cit. on p. 128).
- [450] ChenHsiang Shen, Xiaxia Yu, et al. “Automated Prediction of HIV Drug Resistance from Genotype Data”. In: *BMC Bioinformatics* 17.8 (Aug. 2016), p. 278. ISSN: 1471-2105. DOI: [10.1186/s12859-016-1114-6](https://doi.org/10.1186/s12859-016-1114-6) (cit. on p. 128).
- [451] Xiaxia Yu, Irene T. Weber, and Robert W. Harrison. “Prediction of HIV Drug Resistance from Genotype with Encoded Three-Dimensional Protein Structure”. In: *BMC Genomics* 15.5 (July 2014), S1. ISSN: 1471-2164. DOI: [10.1186/1471-2164-15-S5-S1](https://doi.org/10.1186/1471-2164-15-S5-S1) (cit. on p. 128).

## CHAPTER 6

- [452] Olivier Sheik Amamuddy, Nigel T. Bishop, and Özlem Tastan Bishop. “Improving Fold Resistance Prediction of HIV-1 against Protease and Reverse Transcriptase Inhibitors Using Artificial Neural Networks”. In: *BMC Bioinformatics* 18.1 (Aug. 2017), p. 369. ISSN: 1471-2105. DOI: [10.1186/s12859-017-1782-x](https://doi.org/10.1186/s12859-017-1782-x) (cit. on p. 128).
- [453] N. Beerenwinkel, T. Lengauer, et al. “Geno2pheno: Interpreting Genotypic HIV Drug Resistance Tests”. In: *IEEE Intelligent Systems* 16.6 (Nov. 2001), pp. 35–41. ISSN: 1941-1294. DOI: [10.1109/5254.972080](https://doi.org/10.1109/5254.972080) (cit. on p. 128).
- [454] Seare Tesfamichael Araya and Scott Hazelhurst. “Support Vector Machine Prediction of HIV-1 Drug Resistance Using the Viral Nucleotide Patterns”. In: *Transactions of the Royal Society of South Africa* 64.1 (Jan. 2009), pp. 62–72. ISSN: 0035-919X. DOI: [10.1080/00359190909519238](https://doi.org/10.1080/00359190909519238) (cit. on p. 128).
- [455] Mona Riemenschneider, Robin Senge, et al. “Exploiting HIV-1 Protease and Reverse Transcriptase Cross-Resistance Information for Improved Drug Resistance Prediction by Means of Multi-Label Classification”. In: *BioData Mining* 9.1 (Feb. 2016), p. 10. ISSN: 1756-0381. DOI: [10.1186/s13040-016-0089-1](https://doi.org/10.1186/s13040-016-0089-1) (cit. on p. 128).
- [456] Dominik Heider, Robin Senge, et al. “Multilabel Classification for Exploiting Cross-Resistance Information in HIV-1 Drug Resistance Prediction”. In: *Bioinformatics* 29.16 (Aug. 2013), pp. 1946–1952. ISSN: 1367-4803. DOI: [10.1093/bioinformatics/btt331](https://doi.org/10.1093/bioinformatics/btt331) (cit. on p. 128).
- [457] Sorin Drăghici and R. Brian Potter. “Predicting HIV Drug Resistance with Neural Networks”. In: *Bioinformatics* 19.1 (Jan. 2003), pp. 98–107. ISSN: 1367-4803. DOI: [10.1093/bioinformatics/19.1.98](https://doi.org/10.1093/bioinformatics/19.1.98) (cit. on p. 128).
- [458] Margaret C. Steiner, Keylie M. Gibson, and Keith A. Crandall. “Drug Resistance Prediction Using Deep Learning Techniques on HIV-1 Sequence Data”. en. In: *Viruses* 12.5 (May 2020), p. 560. DOI: [10.3390/v12050560](https://doi.org/10.3390/v12050560) (cit. on p. 128).
- [459] Alyssa C. Mooney, Chadwick K. Campbell, et al. “Beyond Social Desirability Bias: Investigating Inconsistencies in Self-Reported HIV Testing and Treatment Behaviors Among HIV-Positive Adults in North West Province, South Africa”. en. In: *AIDS and Behavior* 22.7 (July 2018), pp. 2368–2379. ISSN: 1573-3254. DOI: [10.1007/s10461-018-2155-9](https://doi.org/10.1007/s10461-018-2155-9) (cit. on p. 130).
- [460] Robert Tibshirani. “Regression Shrinkage and Selection Via the Lasso”. en. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1 (1996), pp. 267–288. ISSN: 2517-6161. DOI: [10.1111/j.2517-6161.1996.tb02080.x](https://doi.org/10.1111/j.2517-6161.1996.tb02080.x) (cit. on pp. 130, 132).
- [461] Glenn W. Brier. “Verification of Forecasts Expressed in Terms of Probability”. en. In: *Monthly Weather Review* 78.1 (Jan. 1950), pp. 1–3. ISSN: 0027-0644 (cit. on pp. 130, 134).
- [462] Olivier Gascuel, Bernadette Bouchon-Meunier, et al. “Twelve Numerical, Symbolic and Hybrid Supervised Classification Methods”. In: *International Journal of Pattern Recognition and Artificial Intelligence* 12.05 (Aug. 1998), pp. 517–571. ISSN: 0218-0014. DOI: [10.1142/S0218001498000336](https://doi.org/10.1142/S0218001498000336) (cit. on pp. 130, 133).

## REFERENCES FOR CHAPTER 6

- [463] Jelle J. Goeman and Aldo Solari. “Multiple Hypothesis Testing in Genomics”. en. In: *Statistics in Medicine* 33.11 (2014), pp. 1946–1978. ISSN: 1097-0258. DOI: [10.1002/sim.6082](https://doi.org/10.1002/sim.6082) (cit. on p. 132).
- [464] Jason D Rennie, Lawrence Shih, et al. “Tackling the Poor Assumptions of Naive Bayes Text Classifiers”. en. In: *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*. 2003, pp. 616–623 (cit. on p. 132).
- [465] Leo Breiman. “Random Forests”. en. In: *Machine Learning* 45.1 (Oct. 2001), pp. 5–32. ISSN: 1573-0565. DOI: [10.1023/A:1010933404324](https://doi.org/10.1023/A:1010933404324) (cit. on p. 132).
- [466] David Alvarez Melis and Tommi Jaakkola. “Towards Robust Interpretability with Self-Explaining Neural Networks”. In: *Advances in Neural Information Processing Systems 31*. Ed. by S. Bengio, H. Wallach, et al. Curran Associates, Inc., 2018, pp. 7775–7784 (cit. on p. 133).
- [467] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. en. Springer Science & Business Media, Aug. 2009. ISBN: 978-0-387-84858-7 (cit. on p. 133).
- [468] Quanshi Zhang, Ying Nian Wu, and Song-Chun Zhu. “Interpretable Convolutional Neural Networks”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 8827–8836 (cit. on p. 133).
- [469] Kay Henning Brodersen, Cheng Soon Ong, et al. “The Balanced Accuracy and Its Posterior Distribution”. In: *2010 20th International Conference on Pattern Recognition*. Aug. 2010, pp. 3121–3124. DOI: [10.1109/ICPR.2010.764](https://doi.org/10.1109/ICPR.2010.764) (cit. on pp. 134, 170).
- [470] Schrödinger, LLC. “The PyMOL Molecular Graphics System, Version 1.8”. Nov. 2015 (cit. on p. 142).
- [471] Soo-Yon Rhee, Tommy F. Liu, et al. “HIV-1 Subtype B Protease and Reverse Transcriptase Amino Acid Covariation”. en. In: *PLOS Computational Biology* 3.5 (May 2007), e87. ISSN: 1553-7358. DOI: [10.1371/journal.pcbi.0030087](https://doi.org/10.1371/journal.pcbi.0030087) (cit. on p. 145).
- [472] Andrea De Luca, Simona Di Giambenedetto, et al. “Improved Interpretation of Genotypic Changes in the HIV-1 Reverse Transcriptase Coding Region That Determine the Virological Response to Didanosine”. en. In: *The Journal of Infectious Diseases* 196.11 (Dec. 2007), pp. 1645–1653. ISSN: 0022-1899. DOI: [10.1086/522231](https://doi.org/10.1086/522231) (cit. on p. 145).
- [473] Anne-Genevieve Marcelin, Philippe Flandre, et al. “Impact of HIV-1 Reverse Transcriptase Polymorphism at Codons 211 and 228 on Virological Response to Didanosine”. en. In: *Antiviral Therapy* (2006), p. 8 (cit. on p. 145).

## CHAPTER 6

- [474] Andrew J. Leigh Brown, Heather M. Precious, et al. “Reduced Susceptibility of Human Immunodeficiency Virus Type 1 (HIV-1) from Patients with Primary HIV Infection to Nonnucleoside Reverse Transcriptase Inhibitors Is Associated with Variation at Novel Amino Acid Sites”. In: *Journal of Virology* 74.22 (Nov. 2000), pp. 10269–10273. ISSN: 0022-538X (cit. on p. 145).
- [475] Shauna A. Clark, Nancy S. Shulman, et al. “Reverse Transcriptase Mutations 118I, 208Y, and 215Y Cause HIV-1 Hypersusceptibility to Non-Nucleoside Reverse Transcriptase Inhibitors”. en-US. In: *AIDS* 20.7 (Apr. 2006), pp. 981–984. ISSN: 0269-9370. DOI: [10.1097/01.aids.0000222069.14878.44](https://doi.org/10.1097/01.aids.0000222069.14878.44) (cit. on p. 145).
- [476] G. Nebbia, Caroline A. Sabin, et al. “Emergence of the H208Y Mutation in the Reverse Transcriptase (RT) of HIV-1 in Association with Nucleoside RT Inhibitor Therapy”. en. In: *Journal of Antimicrobial Chemotherapy* 59.5 (May 2007), pp. 1013–1016. ISSN: 0305-7453. DOI: [10.1093/jac/dkm067](https://doi.org/10.1093/jac/dkm067) (cit. on p. 145).
- [477] A. Saracino, L. Monno, et al. “Impact of Unreported HIV-1 Reverse Transcriptase Mutations on Phenotypic Resistance to Nucleoside and Non-Nucleoside Inhibitors”. en. In: *Journal of Medical Virology* 78.1 (2006), pp. 9–17. ISSN: 1096-9071. DOI: [10.1002/jmv.20500](https://doi.org/10.1002/jmv.20500) (cit. on p. 145).
- [478] Tong Tong Wu, Yi Fang Chen, et al. “Genome-Wide Association Analysis by Lasso Penalized Logistic Regression”. en. In: *Bioinformatics* 25.6 (Mar. 2009), pp. 714–721. ISSN: 1460-2059, 1367-4803. DOI: [10.1093/bioinformatics/btp041](https://doi.org/10.1093/bioinformatics/btp041) (cit. on p. 146).

# **7. Learning alignments, an interesting perspective**

## **7.1. Learning pairwise alignment**

### **7.1.1. DEDAL**

- reference to transformer embedding
- Predict substitution matrix
- Reference other similar works
- drawback: only on proteins

### **7.1.2. predicting an alignment**

- Transformer models can also predict tokens -> predict “CIGAR string” or a aligned sequence.
- Challenges:
  - Longer sequences in DNA
  - Size difference in the case of mapping
  - Less information in a single nucleotide token than in proteins....

## **7.2. What else could we learn ?**

### **7.2.1. Learn to predict seeds or starting positions**

- DeepMinimizer
- predict start position given a pair of sequences

## CHAPTER 7

### 7.2.2. Learn pre-processing functions

i.e. either connections in MSR graph or sequence 2 sequence models

# Global conclusion

## HPC part

- We have developed a method to improve mapping by pre-processing biological sequences
  - In terms of error rate and fraction of mapped reads
  - Although transformations selected on whole human genome generalises to *D. melanogaster* and *E. coli* genomes.
- Where to go next ?
  - explore a larger space of transformations:
    - \* develop more efficient ways to explore search space
    - \* new search space reduction methods
  - Explore different applications: assembly, clustering, ...
  - Explore different types of transformations, i.e. less constraints given by us, ML/Seq2Seq,...

## HIV part

- We have used machine learning in order to find new drug resistance mutations in HIV
  - We have showed a link to drug resistance for 6 mutations of the RT-pol protein, currently not classified as DRMS, but they have been identified as potentially linked to resistance previously
  - These mutations seem to be accessory mutations and do not
  - There seems to be no complex epistasis phenomena
- Where to go next:
  - Confirm experimentally / *in vivo* the role these mutations have in resistance

## APPENDIX

- Explore more sensitive methods, lots of work on interpretable DL -> restricted by dataset size...
- Explore metadata rich data, e.g. per treatment,
- Explore other organisms for which we have data -> HCV ?

## Final words

Alignments are fundamental, improvements in the creation and analysis steps are crucial and likely to help other people gain insight in biological processes. *etc...*

## **A. Supporting Information for “Mapping-friendly sequence reductions: going beyond homopolymer compression”**

### **A.1. “TandemTools” dataset generation**

This dataset was obtained by taking a human X chromosome HOR sequence, concatenating it 500 times with added mutations in order to obtain an approximately 1 Mbp long sequence. Then 1200 reads were simulated from the sequence using `nanosim` [266] and assembled using a centromere-tailored pipeline [479]. A 10kbp deletion was then added to this assembly. The resulting sequence is the one we refer to as the “Centromeric sequence”.

## APPENDIX A

## A.2. MSR performance comparison

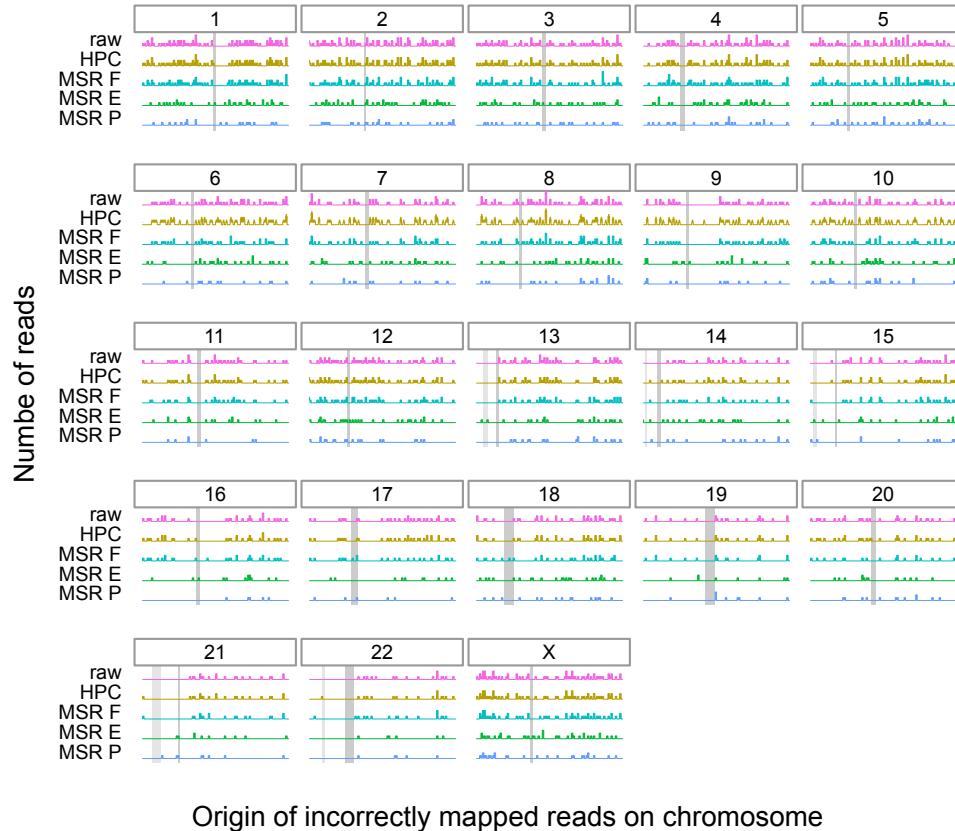
Table A.1.: Comparing performance of MSRs on the whole human genome, whole *Drosophila melanogaster* genome, repeated regions of the whole human genome and synthetic centromeric sequence.

Results using `minimap2` [115] and `winnowmap2` [116]. The number of simulated reads for each reference sequence is given in parentheses and called  $n$ . Results are reported for mapq thresholds of 60, 50 and 0. The best performance for each category is highlighted in bold. The percentage difference are computed w.r.t HPC at each given threshold.

mapping friendly sequence reduction	mapq=60		mapq $\geq$ 50		any mapq	
	fraction	error	fraction	error	fraction	error
<b>Whole Drosophila melanogaster genome - minimap2 (n = 25 764)</b>						
HPC	0.957 +0%	2.27e-03 +0%	<b>0.963 +0%</b>	2.34e-03 +0%	<b>0.998 +0%</b>	1.48e-02 +0%
raw	<b>0.958 +0%</b>	2.27e-03 -0%	0.962 -0%	2.34e-03 +0%	0.997 -0%	<b>1.17e-02 -21%</b>
MSR <sub>F</sub>	0.952 -1%	1.18e-03 -48%	0.960 -0%	1.37e-03 -41%	<b>0.998 +0%</b>	1.36e-02 -8%
MSR <sub>E</sub>	0.946 -1%	<b>0 -100%</b>	0.954 -1%	<b>0 -100%</b>	<b>0.998 +0%</b>	1.53e-02 +3%
MSR <sub>P</sub>	0.950 -1%	4.90e-04 -78%	0.957 -1%	8.11e-04 -65%	<b>0.998 -0%</b>	1.39e-02 -6%
<b>Whole Drosophila melanogaster genome - winnowmap2 (n = 25 764)</b>						
HPC	0.923 +0%	1.51e-03 +0%	0.930 +0%	1.59e-03 +0%	0.989 +0%	1.50e-02 +0%
raw	<b>0.949 +3%</b>	1.92e-03 +27%	<b>0.954 +3%</b>	1.99e-03 +26%	<b>0.995 +1%</b>	<b>1.33e-02 -12%</b>
MSR <sub>F</sub>	0.918 -1%	<b>1.27e-03 -16%</b>	0.925 -0%	<b>1.30e-03 -18%</b>	0.987 -0%	1.37e-02 -9%
MSR <sub>P</sub>	0.905 -2%	1.33e-03 -12%	0.912 -2%	1.53e-03 -3%	0.983 -1%	1.40e-02 -7%
MSR <sub>E</sub>	0.905 -2%	1.42e-03 -6%	0.912 -2%	1.49e-03 -6%	0.983 -1%	1.44e-02 -4%
<b>Synthetic centromeric sequence - minimap2 (n = 12 673)</b>						
HPC	0.870 +0%	<b>1.36e-03 + 0%</b>	0.964 +0%	<b>1.56e-03 + 0%</b>	<b>1.000 +0%</b>	9.00e-03 + 0%
raw	<b>0.936 +8%</b>	1.86e-03 +36%	<b>0.984 +2%</b>	2.09e-03 +34%	<b>1.000 +0%</b>	<b>4.50e-03 -50%</b>
MSR <sub>E</sub>	0.885 +2%	3.39e-03 +149%	0.962 -0%	3.53e-03 +127%	<b>1.000 +0%</b>	1.20e-02 +33%
MSR <sub>F</sub>	0.850 -2%	2.04e-03 +50%	0.968 +0%	2.12e-03 +36%	<b>1.000 +0%</b>	6.63e-03 -26%
MSR <sub>P</sub>	0.898 +3%	1.58e-03 +16%	0.968 +0%	1.79e-03 +15%	<b>1.000 +0%</b>	9.78e-03 + 9%
<b>Synthetic centromeric sequence - winnowmap2 (n = 12 673)</b>						
HPC	0.775 + 0%	<b>1.32e-03 + 0%</b>	<b>0.822 +0%</b>	<b>1.82e-03 + 0%</b>	0.997 +0%	8.37e-02 + 0%
raw	<b>0.850 +10%</b>	2.04e-03 +54%	0.890 +8%	1.95e-03 +7%	<b>0.999 +0%</b>	<b>4.60e-02 -45%</b>
MSR <sub>E</sub>	0.795 + 2%	2.28e-03 +73%	0.846 +3%	2.52e-03 +38%	0.997 -0%	6.96e-02 -17%
MSR <sub>F</sub>	0.820 + 6%	1.83e-03 +38%	0.867 +6%	2.27e-03 +25%	0.997 -0%	5.97e-02 -29%
MSR <sub>P</sub>	0.780 + 1%	1.62e-03 +22%	0.829 +1%	2.09e-03 +15%	0.997 -0%	8.65e-02 + 3%
<b>Whole human genome - minimap2 (n = 655 594)</b>						
HPC	<b>0.935 +0%</b>	1.85e-03 + 0%	<b>0.942 +0%</b>	1.85e-03 + 0%	<b>1.000 +0%</b>	1.46e-02 + 0%
raw	0.921 -1%	1.86e-03 + 0%	0.927 -2%	1.86e-03 + 1%	0.998 -0%	<b>1.29e-02 -11%</b>
MSR <sub>E</sub>	0.926 -1%	<b>6.92e-05 -96%</b>	0.936 -1%	<b>1.17e-04 -94%</b>	0.999 -0%	1.76e-02 +20%
MSR <sub>P</sub>	0.929 -1%	2.20e-04 -88%	0.938 -0%	4.15e-04 -78%	0.999 -0%	1.55e-02 + 6%
MSR <sub>F</sub>	0.930 -1%	1.09e-03 -41%	0.938 -0%	1.29e-03 -30%	<b>1.000 -0%</b>	1.51e-02 + 4%
<b>Whole human genome - winnowmap2 (n = 655 594)</b>						
HPC	0.894 + 0%	1.43e-03 + 0%	0.902 +0%	1.49e-03 + 0%	0.988 +0%	1.92e-02 + 0%
raw	<b>0.932 + 4%</b>	1.75e-03 +23%	<b>0.937 +4%</b>	1.79e-03 +20%	<b>0.994 +1%</b>	<b>1.43e-02 -26%</b>
MSR <sub>F</sub>	0.874 - 2%	2.81e-04 -80%	0.886 -2%	3.82e-04 -74%	0.984 -0%	1.94e-02 + 1%
MSR <sub>E</sub>	0.795 -11%	<b>6.33e-05 -96%</b>	0.820 -0%	<b>8.93e-05 -94%</b>	0.971 -2%	2.08e-02 + 9%
MSR <sub>P</sub>	0.826 - 8%	8.68e-05 -94%	0.845 -6%	1.14e-04 -92%	0.975 -1%	2.11e-02 +10%
<b>Whole Human genome (repeated regions) - minimap2 (n = 68 811)</b>						
HPC	<b>0.619 + 0%</b>	3.29e-04 + 0%	0.656 + 0%	3.10e-04 + 0%	<b>0.998 +0%</b>	7.79e-02 + 0%
raw	0.514 -17%	1.98e-04 -40%	0.539 -18%	2.16e-04 -30%	0.981 -2%	<b>6.69e-02 -14%</b>
MSR <sub>F</sub>	0.601 - 3%	2.18e-04 -34%	0.640 - 2%	2.27e-04 -27%	<b>0.998 -0%</b>	8.15e-02 + 5%
MSR <sub>E</sub>	0.618 - 0%	1.41e-04 -57%	<b>0.658 + 0%</b>	<b>1.55e-04 -50%</b>	0.997 -0%	8.23e-02 + 6%
MSR <sub>P</sub>	0.616 - 1%	<b>1.18e-04 -64%</b>	0.656 + 0%	1.99e-04 -36%	0.997 -0%	8.31e-02 + 7%
<b>Whole Human genome (repeated regions) - winnowmap2 (n = 68 811)</b>						
HPC	0.525 + 0%	1.24e-03 + 0%	0.557 + 0%	1.49e-03 + 0%	0.950 +0%	1.19e-01 + 0%
raw	<b>0.648 +23%</b>	1.26e-03 + 1%	<b>0.672 +21%</b>	1.49e-03 + 0%	<b>0.968 +2%</b>	<b>8.09e-02 -32%</b>
MSR <sub>F</sub>	0.482 - 8%	<b>1.63e-03 +31%</b>	0.516 - 7%	1.83e-03 +23%	0.940 -1%	1.21e-01 + 2%
MSR <sub>E</sub>	0.366 -30%	6.35e-04 -49%	0.405 -27%	<b>9.32e-04 -37%</b>	0.911 -4%	1.38e-01 +17%
MSR <sub>P</sub>	0.415 -21%	9.45e-04 -24%	0.451 -19%	1.16e-03 -22%	0.920 -3%	1.39e-01 +17%

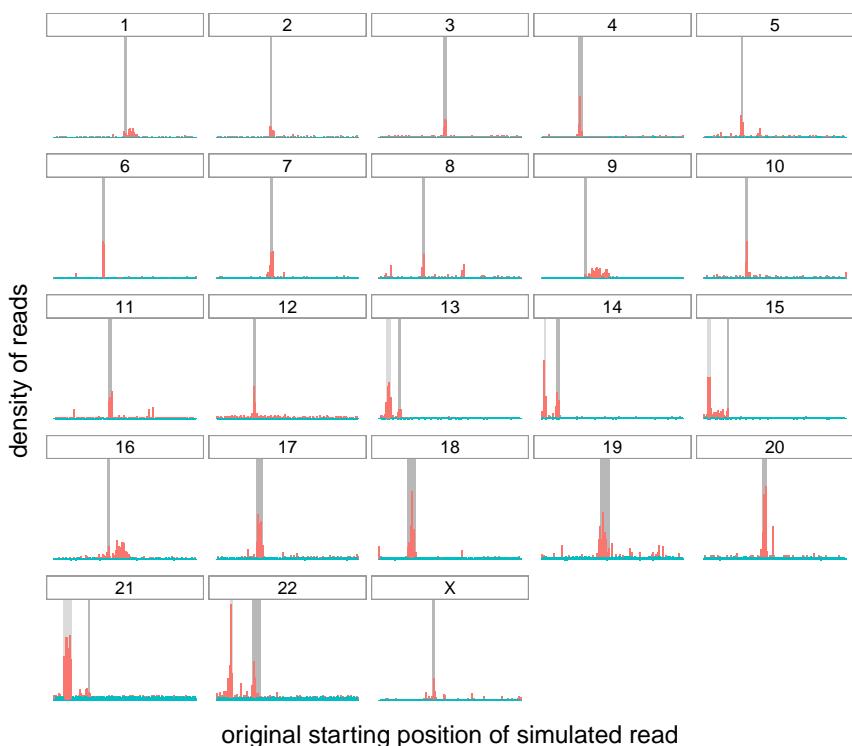
## APPENDIX A

### A.3. Origin of incorrectly mapped reads of high mapping quality on whole human genome.



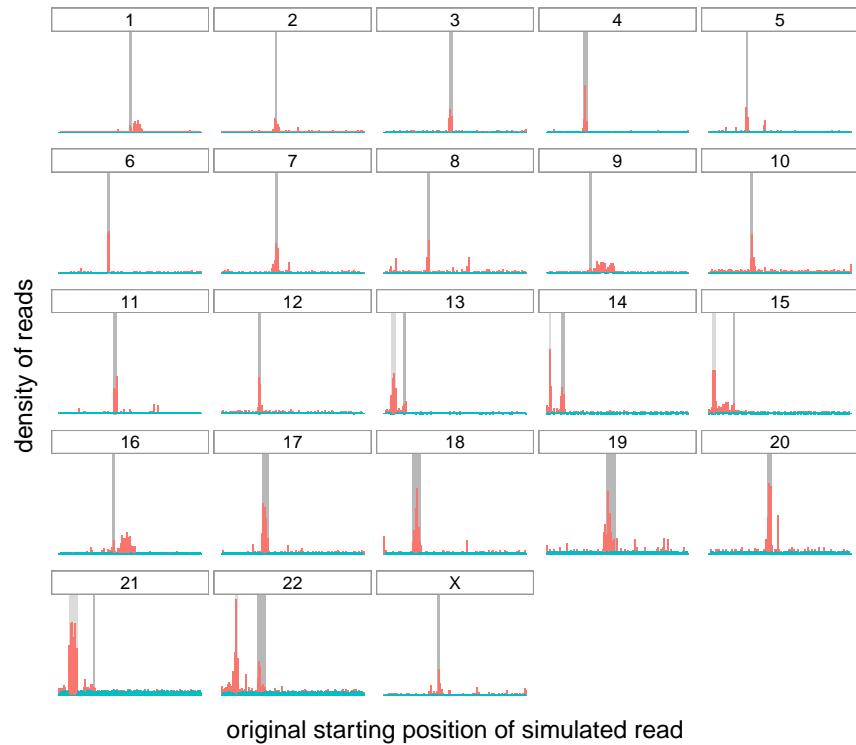
**Figure A.1.: Histogram of the original simulated positions for the incorrectly mapped reads using `minimap2` at high mapping qualities across the whole human genome, for several transformation methods.**

For a given chromosome, each row represents the number of simulated reads starting at that particular region. The dark gray rectangle represents the position of the centromere for that chromosome, obtained from annotations provided by the T2T consortium (<http://t2t.gi.ucsc.edu/chm13/hub/t2t-chm13-v1.1/>). Similarly for chromosomes 13, 14, 15, 21 and 22, a lighter gray rectangle represents the position of the “stalk” satellites also containing repetitive regions. For HPC and raw reads only alignments of mapping quality 60 were considered. To provide a fair comparison, alignments with mapping qualities  $\geq 50$  were considered for MSRs E, F and P.

**A.4. Analyzing read origin on whole human genome****Figure A.2.: Origin of correctly and incorrectly mapped raw reads**

Distribution of the origin of correctly and incorrectly mapped simulated reads (in teal and red respectively) on the different chromosomes of the whole human genome. The dark grey rectangle for each chromosome represents the centromere of that chromosome. The lighter gray rectangle on chromosomes 13, 14, 15, 21 and 22 correspond to satellites denoted as “stalk”, another repetitive region.

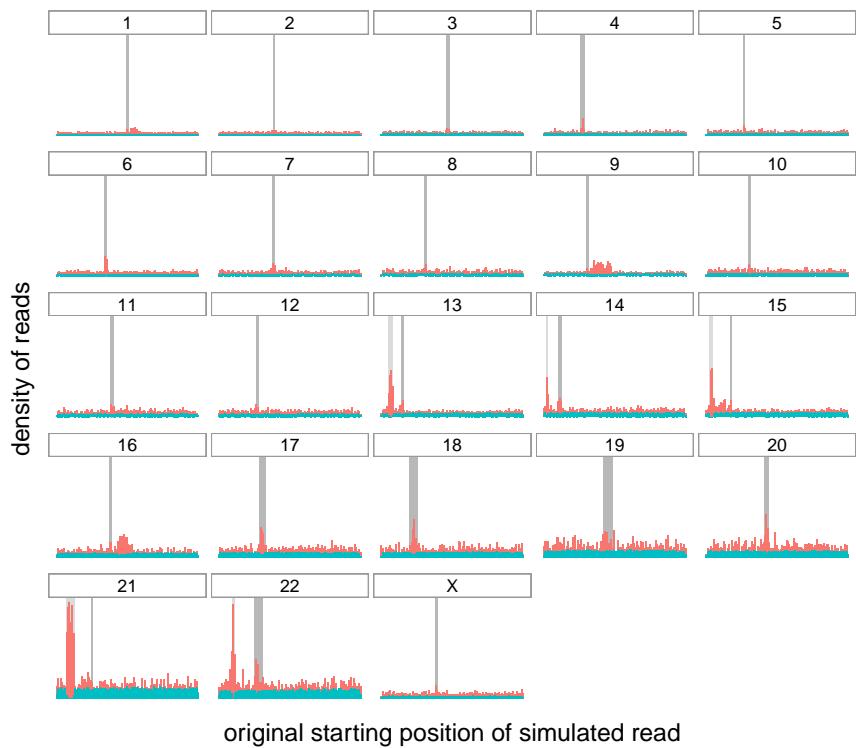
## APPENDIX A



**Figure A.3.: Origin of correctly (teal) and incorrectly (red) mapped reads, transformed with HPC**

Distribution of the origin of correctly and incorrectly mapped simulated reads (in teal and red respectively) on the different chromosomes of the whole human genome. The dark grey rectangle for each chromosome represents the centromere of that chromosome. The lighter gray rectangle on chromosomes 13, 14, 15, 21 and 22 correspond to satellites denoted as “stalk”, another repetitive region.

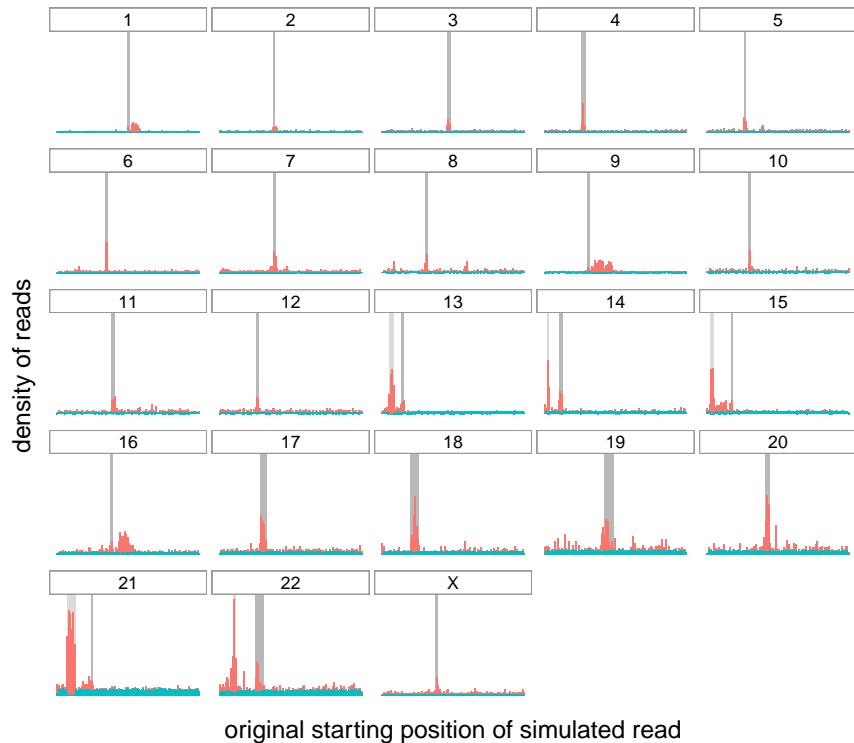
#### A.4. ANALYZING READ ORIGIN ON WHOLE HUMAN GENOME



**Figure A.4.: Origin of correctly (teal) and incorrectly (red) mapped reads, transformed with  $\text{MSR}_E$**

Distribution of the origin of correctly and incorrectly mapped simulated reads (in teal and red respectively) on the different chromosomes of the whole human genome. The dark grey rectangle for each chromosome represents the centromere of that chromosome. The lighter gray rectangle on chromosomes 13, 14, 15, 21 and 22 correspond to satellites denoted as “stalk”, another repetitive region.

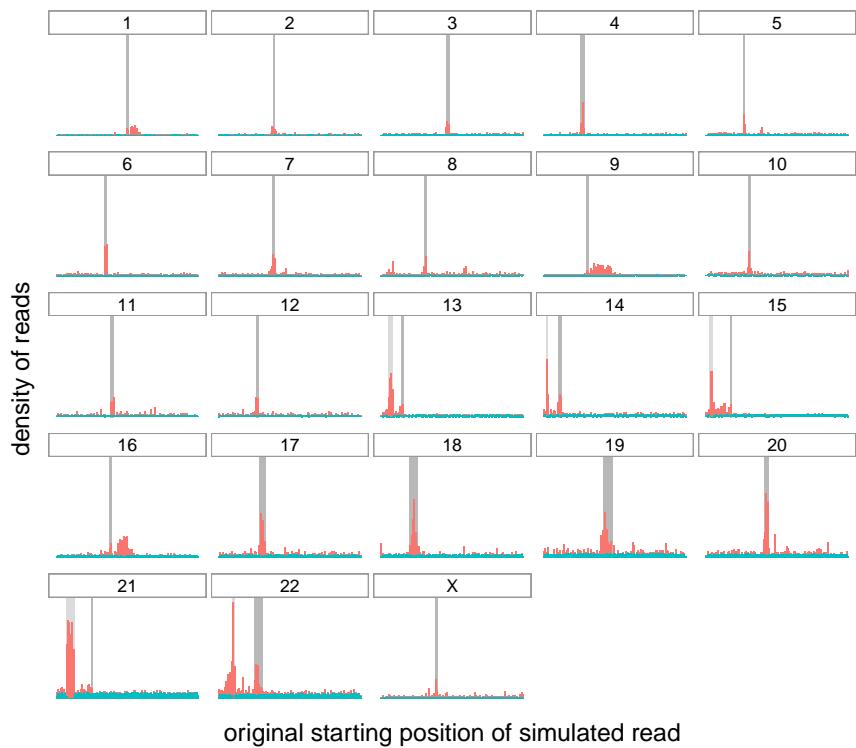
## APPENDIX A



**Figure A.5.: Origin of correctly (teal) and incorrectly (red) mapped reads, transformed with  $\text{MSR}_P$**

Distribution of the origin of correctly and incorrectly mapped simulated reads (in teal and red respectively) on the different chromosomes of the whole human genome. The dark grey rectangle for each chromosome represents the centromere of that chromosome. The lighter gray rectangle on chromosomes 13, 14, 15, 21 and 22 correspond to satellites denoted as “stalk”, another repetitive region.

#### A.4. ANALYZING READ ORIGIN ON WHOLE HUMAN GENOME



**Figure A.6.: Origin of correctly (teal) and incorrectly (red) mapped reads, transformed with  $\text{MSR}_F$**

Distribution of the origin of correctly and incorrectly mapped simulated reads (in teal and red respectively) on the different chromosomes of the whole human genome. The dark grey rectangle for each chromosome represents the centromere of that chromosome. The lighter gray rectangle on chromosomes 13, 14, 15, 21 and 22 correspond to satellites denoted as “stalk”, another repetitive region.

## A.5. Performance of MSRs on the *Drosophila* genome

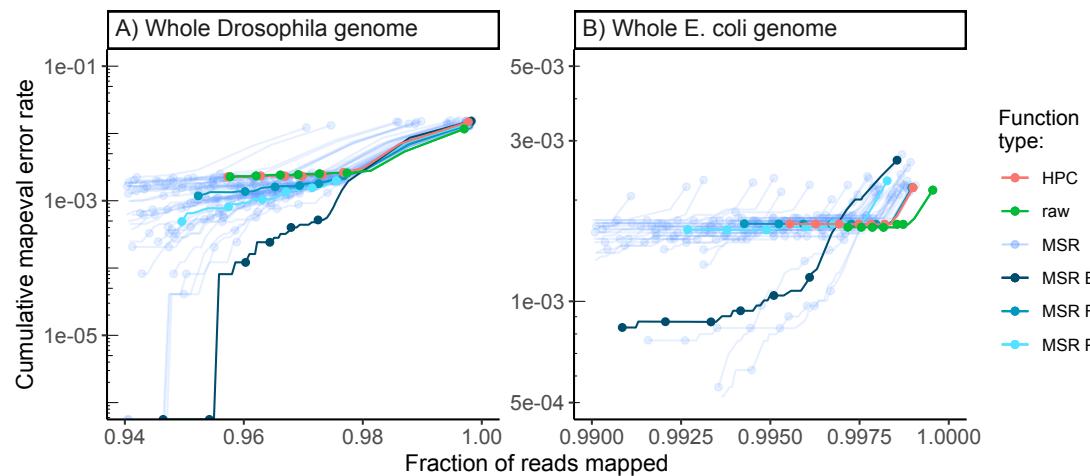


Figure A.7.: Results of the `paftools mapeval` evaluation on reads simulated and mapped to whole *Drosophila melanogaster* and *Escherichia coli* (Genbank ID U00096.2) genomes.

MSRs E, F and P are shown in different shades of blue to differentiate them from other MSRs. Reads were simulated with `nanosim`, and mapped with `minimap2`.

## References for appendix A

- [115] Heng Li. “Minimap2: Pairwise Alignment for Nucleotide Sequences”. In: *Bioinformatics* 34.18 (Sept. 15, 2018), pp. 3094–3100. ISSN: 1367-4803. DOI: [10.1093/bioinformatics/bty191](https://doi.org/10.1093/bioinformatics/bty191) (cit. on pp. 25, 48, 51, 71, 78, 79, 159).
- [116] Chirag Jain, Arang Rhie, et al. “Weighted Minimizer Sampling Improves Long Read Mapping”. In: *Bioinformatics* 36 (Supplement\_1 July 1, 2020), pp. i111–i118. ISSN: 1367-4803. DOI: [10.1093/bioinformatics/btaa435](https://doi.org/10.1093/bioinformatics/btaa435) (cit. on pp. 25, 52, 81, 159).
- [266] Chen Yang, Justin Chu, et al. “NanoSim: Nanopore Sequence Read Simulator Based on Statistical Characterization”. In: *GigaScience* 6.4 (Apr. 1, 2017). ISSN: 2047-217X. DOI: [10.1093/gigascience/gix010](https://doi.org/10.1093/gigascience/gix010) (cit. on pp. 78, 157).
- [479] Andrey V. Bzikadze and Pavel A. Pevzner. “Automated Assembly of Centromeres from Ultra-Long Error-Prone Reads”. In: *Nature Biotechnology* 38.11 (11 Nov. 2020), pp. 1309–1316. ISSN: 1546-1696. DOI: [10.1038/s41587-020-0582-4](https://doi.org/10.1038/s41587-020-0582-4) (cit. on p. 157).

## B. Supporting Information for “HIV and DRMs”

### B.1. Detailed list of HIV-1 protein structures used for figure generation.

The images for HIV-1 structures used in Figure 5.2 were obtained from: <https://cdn.rcsb.org/pdb101/learn/resource/biology-of-hiv/>. They are licensed under a Creative Commons By 4.0 license which allows reuse and adaptation for non commercial use.

PDB structure IDs:

- **SU** and **TM**: [4nco](#)
- **MA**: [1hiw](#)
- **CA**: [3h47](#)
- **NC**: [1a1t](#)
- **RT**: [1hys](#) (for Figure 5.2) and [2hmi](#) (for Figure 5.3)
- **IN**: [1ex4](#)
- **PR**: [1hpv](#)
- **Vpu**: [1pi7](#) and [1vpu](#)
- **Vif**: [3dcg](#)
- **Vpr**: [1esx](#)
- **Nef**: [1avv](#) and [1qa5](#)
- **Rev**: [1etf](#)
- **Tat**: [1biv](#) and [1jfw](#)



## **C. Supporting Information for “Using Machine Learning and Big Data to Explore the Drug Resistance Landscape in HIV”**

### **C.1. S1 Appendix (Technical appendix).**

#### **C.1.1. Data**

##### **C.1.1.1. Data Availability**

The policy of the UK HIV Drug Resistance Database is to make DNA sequences available to any bona fide researcher who submits a scientifically robust proposal, provided data exchange complies with Information Governance and Data Security Policies in all the relevant countries. This includes replication of findings from published studies, although the researcher would be encouraged to work with the main author of the published paper to understand the nuances of the data. Enquiries should be addressed to [iph.hivrd@ucl.ac.uk](mailto:iph.hivrd@ucl.ac.uk) in the first instance. More information on the UK dataset is also available on the UK CHIC homepage: [www.ukchic.org.uk](http://www.ukchic.org.uk). Amino acid sequences are made available along with a metadata file.

The West and central African dataset is available as supplementary information along with a metadata file containing HIV subtype, treatment information and known RAM presence/absence for each sequence.

Predictions made for each sequence of both datasets, by all of the trained classifiers are made available as part of the supplementary data as well as synthetic results from which the figures of the paper were drawn. The importance values for each mutation and each trained classifier are also made available.

All the data and metadata files made available are hosted in the online repository linked to this project at the following URL:

[github.com/lucblassel/HIV-DRM-machine-learning/tree/main/data](https://github.com/lucblassel/HIV-DRM-machine-learning/tree/main/data)

## APPENDIX C

### C.1.1.2. Data Preprocessing

For both the African and UK datasets, the sequences were truncated to keep sites 41 to 235 of the RT protein sequence before encoding. This truncation was needed to avoid the perturbation to classifier training due to long gappy regions at the beginning and end of the UK RT alignment caused by shorter sequences. These positions were determined with the Gblocks software [480] with default parameters, except for the Maximum number of sequences for a flanking position, set to 50,000, and the Allowed gap positions, which was set to “All”. The encoding was done with the `OneHotEncoder` from the category-encoders python module [481].

### C.1.2. Classifiers

We used classifier implementations from the scikit-learn python library [482], `RandomForestClassifier` for the random forest classifier, `MultinomialNB` for Naïve Bayes and `LogisticRegressionCV` for logistic regression.

`RandomForestClassifier` was used with default parameters except:

- "n\_jobs"=4
- "n\_estimators"=5000

`LogisticRegressionCV` was used with the following parameters:

- "n\_jobs"=4
- "cv"=10
- "Cs"=100
- "penalty"='l1'
- "multi\_class"='multinomial'
- "solver"='saga'
- "scoring"='balanced\_accuracy'

`MultinomialNB` was used with default parameters.

For the Fisher exact tests, we used the implementation from the `scipy` python library [483], and corrected p-values for multiple testing with the `statsmodels` python library [484] using the "Bonferroni" method.

### C.1.3. Scoring

To evaluate classifier performance several measures were used. We computed balanced accuracy instead of classical accuracy, because it can be overly optimistic, especially when assessing a highly biased classifier on an unbalanced test set [469]. The balanced accuracy is computed using the following formula, where  $TP$  and  $TN$  are the number of true positives and true negatives respectively, and  $FP$  and  $FN$  are the number of

### C.1. S1 APPENDIX (TECHNICAL APPENDIX).

false positives and false negatives respectively:

$$\text{balanced accuracy} = \frac{1}{2} \left( \frac{TP}{TP+FP} + \frac{TN}{TN+FN} \right)$$

We also computed adjusted mutual information (AMI). We chose it over mutual information (MI) because it has an upper bound of 1 for a perfect classifier and is not dependent on the size of the test set, allowing us to compare the performance for differently sized test sets [143]. The adjusted mutual information of variables  $U$  and  $V$  is defined by the following formula, where  $MI(U, V)$  is the mutual information between variables  $U$  and  $V$ ,  $H(X)$  is the entropy of the variable  $X$  ( $= U$  or  $V$ ) and  $E\{MI(U, V)\}$  is the expected MI, as explained in [485].

$$AMI(U, V) = \frac{MI(U, V) - E\{MI(U, V)\}}{\frac{1}{2}[H(U) + H(V)] - E\{MI(U, V)\}}$$

MI was used to compute the  $G$  statistic, which follows the chi-square distribution under the null hypothesis [486]. This was used to compute p-values for each of our classifiers and assess the significance of their performance.  $G$  is defined by equation below, where  $N$  is the number of samples.

$$G = 2 \cdot N \cdot MI(U, V)$$

Finally, to check the probabilistic predictive power of the classifiers we also computed the Brier score which is the mean squared difference between the ground truth and the predicted probability of being of the positive class for every sequence in the test set (therefore lower is better for this metric). The Brier score is defined in equation below, where  $p_t$  is the predicted probability of being of the positive class for sample  $t$  and  $o_t$  is the actual class (0 or 1, 1=positive class) of sample  $t$ :

$$\text{Brier score} = \frac{1}{N} \sum_{t=1}^N (p_t - o_t)^2$$

We used the following implementations from the scikit-learn python library [482] with default options:

- `balanced_accuracy_score`
- `mutual_info_score`
- `adjusted_mutual_info_score`
- `brier_score_loss`

We used the relative risk to observe the relationship between one of our new mutations and a binary character  $X$  such as treatment status or presence/absence of a known

## APPENDIX C

RAM.

$$\begin{aligned} RR(new, X) &= \frac{\text{prevalence}(new \text{ mutation} \mid X = 1)}{\text{prevalence}(new \text{ mutation} \mid X = 0)} \\ &= \frac{|(new = 1) \cap (X = 1)|}{|(X = 1)|} \div \frac{|(new = 1) \cap (X = 0)|}{|(X = 0)|} \end{aligned}$$

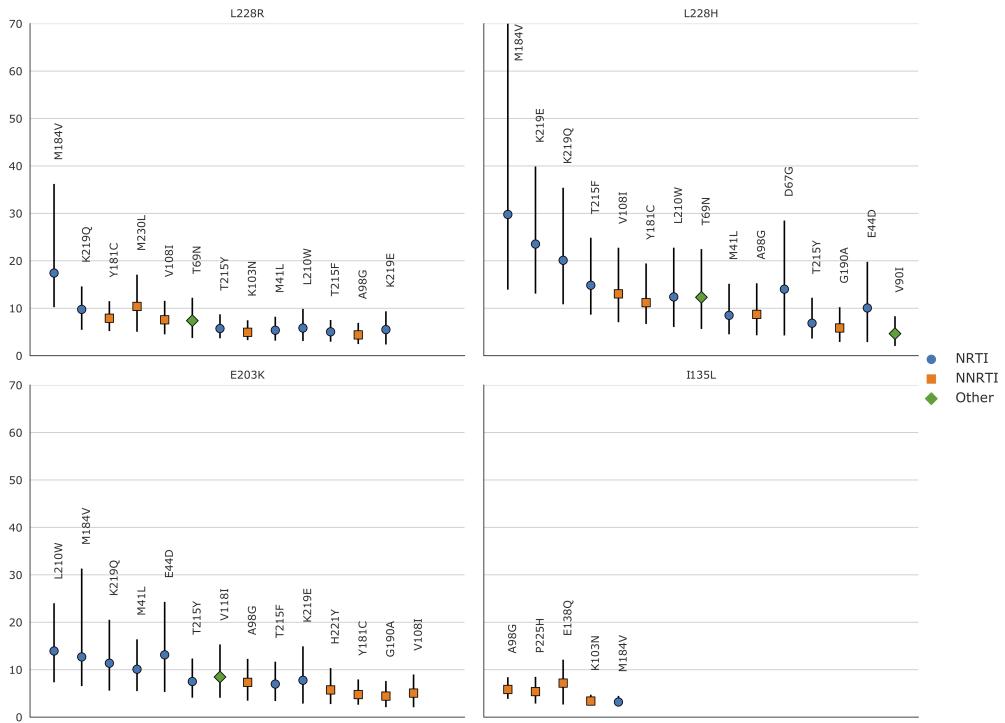
**C.2. S1 Fig.**

Figure C.1.: Relative risks of the new mutations with regards to known RAMs on the African dataset

(i.e. the prevalence of the new mutation in sequences with a given RAM divided by the prevalence of the new mutation in sequences without the RAM). RRs were only computed for mutations (new and RAMs) that appeared in at least 30 sequences, which is why RRs were not computed for H208Y and D218E. 95% confidence intervals, represented by vertical bars, were computed with 1000 bootstrap samples of the African sequences. Only RRs with a lower CI boundary greater than 2 are shown. The shape and color of the point represents the type of RAM as defined by Stanford's HIVDB. Blue circle: NRTI, orange square: NNRTI, green diamond: Other. For the RR of L228H with regards to M184V, the upper CI bound is infinite. The new RAMs have high RR values for known RAMs similar to those obtained on the UK dataset. We also arrive at similar conclusions, I135L being associated with NNRTIs, E203K and L228H to NRTI and L228R to both. RR values are shown from left to right, by order of decreasing values on the lower bound of the 95% CI.

## APPENDIX C

### C.3. S2 Fig.

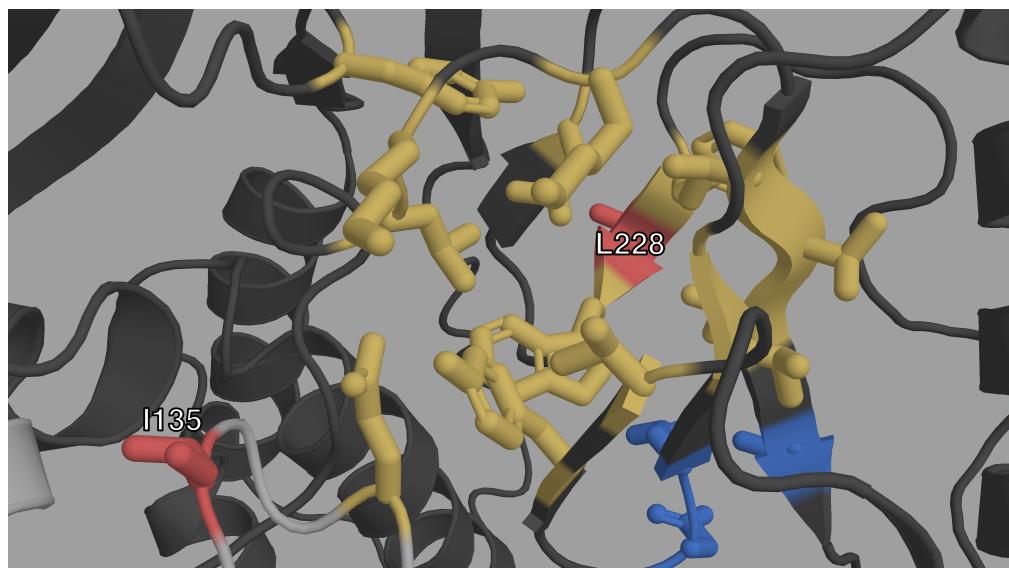


Figure C.2.: Closeup structural view of the entrance of the NNIBP of HIV-1 RT

The p66 subunit is colored in dark gray, the p51 subunit in light gray. The NNIBP is highlighted in yellow. The active site is colored in blue. We can see the physical proximity of I135 (red) to the entrance of the NNIBP. We can also see how L228 (red) is between 2 AAs of the NNIBP.

**C.4. S3 Fig.**

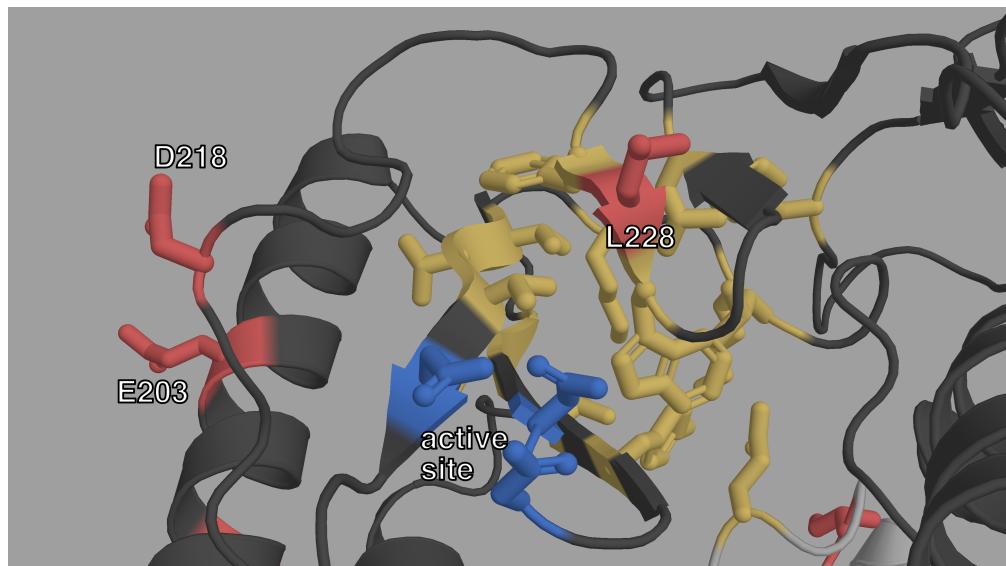


Figure C.3.: Closeup structural view of the active site of HIV-1 RT.

The p66 subunit is colored in dark gray, the p51 subunit in light gray. The active site is highlighted in blue. The NNIBP is colored in yellow. L228, E203 and D218 (red) are also very close on either side of the active site.

APPENDIX C

**C.5. S1 Table.**

	rank	codon distance				UK		Africa				p-value	B62	Dayhoff category shift	Change in			
		T/N	W/W	min	UK	Africa	count	ratio $\rho(new, treatment)$	$\rho(new, with RAM)$	count	ratio $\rho(new, treatment)$	$\rho(new, with RAM)$			net charge	polarity	hydrophobicity index	molecular weight
L228R	0	0	1	1.16	1.21	227 (0.4%)	18.1 [12.9;27.3]	115.7 [55.1;507.3]	98 (2.5%)	32.5 [15.4;147.1]	42.4 [17.8; $\infty$ ]	2.0E-30	-2	e → d	1	5.6	-0.93	43.03
E203K	1	1	1	1.31	1.33	256 (0.5%)	11.0 [8.2;15.1]	20.1 [13.7;32.1]	56 (1.4%)	14.1 [6.7;71.9]	17.4 [8.2;83.7]	6.4E-14	1	c → d	2	-1	0.68	-0.94
D218E	2	3	1	1	1	168 (0.3%)	13.1 [9.0;19.6]	27.0 [16.3;57.0]	25 (0.6%)	$\infty$ [ $\infty$ ; $\infty$ ]	$\infty$ [ $\infty$ ; $\infty$ ]	2.0E-09	2	c → c	0	-0.7	0.01	14.03
L228H	3	4	1	1.12	1.17	287 (0.5%)	6.4 [5.1;8.4]	9.2 [6.9;12.6]	53 (1.3%)	23.1 [9.4; $\infty$ ]	34.1 [12.0; $\infty$ ]	2.7E-15	-3	e → d	0	5.5	-0.92	23.99
I135L	4	6	1	1.16	1.13	540 (1.0%)	1.8 [1.5;2.1]	2.4 [2.0;2.8]	134 (3.4%)	2.6 [1.8;3.8]	2.4 [1.7;3.4]	2.6E-07	2	e → e	0	-0.3	-0.69	0
H208Y	8	9	1	1.10	1.12	205 (0.4%)	8.8 [6.5;12.5]	14.9 [9.9;23.6]	13 (0.3%)	$\infty$ [ $\infty$ ; $\infty$ ]	$\infty$ [ $\infty$ ; $\infty$ ]	7.3E-05	2	d → f	0	-4.2	1.27	26.03

Table C.1.: Detailed view of the characteristics of new potential RAMs

**Rank:** For each new mutation we computed the aggregate feature importance ranks for the RTI-naive / RTI-experienced and known RAM present / known RAM absent classification tasks. **Codon distance:** We computed the minimum number of nucleotide mutations to go from the wild amino acid codons to those of the mutated amino acid, as well as the average codon distance between both amino acids, weighted by the prevalence of each wild and mutated codon in the UK and the African datasets. **Count (both UK and Africa):** We looked at the number of apparitions of each new potential RAM in the UK and African datasets and the corresponding prevalence in parentheses. **Ratio (both UK and Africa):** We computed the prevalence ratio  $\rho(new, treatment)$  (e.g. L228R is 18.1 times more prevalent in RTI-experienced sequences compared to RTI-naive sequences in the UK dataset). We also computed the prevalence ratio  $\rho(new, anyRAM)$  (e.g. L228R is 115.7 times more prevalent in sequences that have at least one known RAM than in sequences that have none in the UK dataset). The 95% confidence intervals shown under each ratio were computed with 1000 bootstrap samples of size  $n = 55,000$  drawn with replacement from the whole UK dataset (The same procedure was done on the African dataset with size  $n = 3990$ ). **p-values:** Fisher exact tests were done on the African dataset to see if each of these new mutations were more prevalent in RTI-experienced sequences; p-value were corrected with the Bonferroni method for the six simultaneous tests. **B62:** BLOSUM62 similarity values (e.g. D218E = 2, reflecting that E and D are both negatively charged and highly similar). **Dayhoff category shift:** The change in Dayhoff amino acid category is written thusly: “starting category → ending category”. These categories are as follows: *a*: Sulfur polymerization. *b*: Small, *c*: Acid and amide, *d*: Basic, *e*: Hydrophobic and *f*: aromatic. **Physico-chemical change:** Change in physicochemical properties was obtained by subtracting the property value of the wild-type amino acid from the mutated amino acid. All values were obtained from the AAindex database [487]

## C.6. S2 Appendix. (Fisher exact tests)

**Fisher exact tests on pairs of mutations.** A detailed explanation of the procedure followed to test pairs of mutations for association with treatment. Detailed numerical results are also given.

In order to study epistasis further we conducted Fisher exact tests between every pair of mutations in the UK dataset ( $n = 867,903$ ) and the treatment status, corrected the p-values with the Bonferroni method with an overall risk level  $\alpha = 0.05$ . Out of these tests, 1,309 pairs were significantly associated with treatment status. 424 out of 1,309 these pairs were two known RAMs, 806 of these pairs contained one known RAM and only 79 tests had pairs involving no known RAM at all. Furthermore out of these 1,309 significantly associated pairs, 829 contained two mutations that were significantly associated to treatment when testing mutations one by one. In 478 pairs, one of the two mutations is associated to treatment on its own, and the remaining 2 pairs, none of the mutations were significantly associated with treatment on their own. These 2 pairs were K103R + V179D and T165I + K173Q. The first pair, is a pair of known RAMs and this interaction is characterized in the HIVDb database (<https://hivdb.stanford.edu/dr-summary/comments/NNRTI/>). The second pair is made up of new mutations, and the corrected p-value is 0.02. In the Standford HIVDB, T165I has been associated to a reduction in EFV susceptibility.

Out of the 1,309 pairs significantly associated to treatment, 151 contained at least one of our 6 new potential RAMs, in 6 cases the pair was made up of 2 of them.

In the UK dataset, phylogenetic correlation is likely very impactful with regards to these tests. Indeed, the sequences are far from being independent. In order to alleviate this effect we decided to test the sigficative pairs again on the African dataset, and once more correct with the Bonferroni procedure.

Out of the 1,309 tests 294 have significative p-values after correction. Out of these 221 pairs were composed of 2 mutations individually significatively associated with treatment. The remaining 73 pairs had one mutation significantly associated with treatment.

Out of the 221 significative tests, 156 pairs were composed of 2 known RAMS while 135 had one known RAM in the pair. The remaining 3 pairs that do not contain a known RAM all contained either L228R or L228H which are both part of our 6 potential RAMS.

## C.7. S1 Data.

**Archive of figure generating data.** A zip archive containing the processed data used to generate each panel of the main figures.

<https://doi.org/10.1371/journal.pcbi.1008873.s007> (ZIP)

## C.8. S2 Data.

**List of known DRMs.** A .csv file containing all the known RAMs used in this project as well as the corresponding feature name in the encoded datasets. Obtained from ([hivdb.stanford.edu/dr-summary/comments/NRTI/](http://hivdb.stanford.edu/dr-summary/comments/NRTI/)) and ([hivdb.stanford.edu/dr-summary/comments/NNRTI/](http://hivdb.stanford.edu/dr-summary/comments/NNRTI/)).

<https://doi.org/10.1371/journal.pcbi.1008873.s008> (CSV)

## References for Appendix C

- [443] Nguyen Xuan Vinh, Julien Epps, and James Bailey. “Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance”. en. In: *Journal of Machine Learning Research* 11 (2010), p. 18 (cit. on pp. 42, 134, 171).
- [469] Kay Henning Brodersen, Cheng Soon Ong, et al. “The Balanced Accuracy and Its Posterior Distribution”. In: *2010 20th International Conference on Pattern Recognition*. Aug. 2010, pp. 3121–3124. doi: [10.1109/ICPR.2010.764](https://doi.org/10.1109/ICPR.2010.764) (cit. on pp. 134, 170).
- [480] J. Castresana. “Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis”. en. In: *Molecular Biology and Evolution* 17.4 (Apr. 2000), pp. 540–552. ISSN: 0737-4038. doi: [10.1093/oxfordjournals.molbev.a026334](https://doi.org/10.1093/oxfordjournals.molbev.a026334) (cit. on p. 170).
- [481] Will McGinnis, Hbghhy, et al. *Scikit-Learn-Contrib/Categorical-Encoding: Release For Zenodo*. Zenodo. Jan. 2018. doi: [10.5281/ZENODO.1157110](https://doi.org/10.5281/ZENODO.1157110) (cit. on p. 170).
- [482] Fabian Pedregosa, Gaël Varoquaux, et al. “Scikit-Learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12.Oct (2011), pp. 2825–2830. ISSN: ISSN 1533-7928 (cit. on pp. 170, 171).
- [483] Pauli Virtanen, Ralf Gommers, et al. “SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python”. en. In: *Nature Methods* 17.3 (Mar. 2020), pp. 261–272. ISSN: 1548-7105. doi: [10.1038/s41592-019-0686-2](https://doi.org/10.1038/s41592-019-0686-2) (cit. on p. 170).
- [484] Skipper Seabold and Josef Perktold. “Statsmodels: Econometric and Statistical Modeling with Python”. en. In: *Proceedings of the 9th Python in Science Conference*. Ed. by Stéfan van der Walt and Jarrod Millman. 2010, pp. 92–96. doi: [10.25080/Majora-92bf1922-011](https://doi.org/10.25080/Majora-92bf1922-011) (cit. on p. 170).
- [485] N. X. Vinh and J. Epps. “A Novel Approach for Automatic Number of Clusters Detection in Microarray Data Based on Consensus Clustering”. In: *2009 Ninth IEEE International Conference on Bioinformatics and BioEngineering*. June 2009, pp. 84–91. doi: [10.1109/BIBE.2009.19](https://doi.org/10.1109/BIBE.2009.19) (cit. on p. 171).

## APPENDIX C

- [486] Peter Harremoes. “Mutual Information of Contingency Tables and Related Inequalities”. In: *2014 IEEE International Symposium on Information Theory*. Honolulu, HI, USA: IEEE, June 2014, pp. 2474–2478. ISBN: 978-1-4799-5186-4. DOI: [10.1109/ISIT.2014.6875279](https://doi.org/10.1109/ISIT.2014.6875279) (cit. on p. 171).
- [487] Shuichi Kawashima, Piotr Pokarowski, et al. “AAindex: amino acid index database, progress report 2008”. In: *Nucleic Acids Research* 36.suppl\_1 (Jan. 1, 2008), pp. D202–D205. DOI: [10.1093/nar/gkm998](https://doi.org/10.1093/nar/gkm998). URL: [https://academic.oup.com/nar/article/36/suppl\\_1/D202/2508449](https://academic.oup.com/nar/article/36/suppl_1/D202/2508449) (cit. on p. 177).

# Global References

- [1] J. D. Watson and F. H. C. Crick. “The Structure of Dna”. In: *Cold Spring Harbor Symposia on Quantitative Biology* 18 (Jan. 1, 1953). tex.ids= watson-STRUCTUREDNA1953 PMID: 13168976 publisher: Cold Spring Harbor Laboratory Press, pp. 123–131. DOI: [10.1101/SQB.1953.018.01.020](https://doi.org/10.1101/SQB.1953.018.01.020). URL: <http://symposium.cshlp.org/content/18/123> (cit. on p. 11).
- [2] F. Sanger, G. M. Air, et al. “Nucleotide Sequence of Bacteriophage X174 DNA”. In: *Nature* 265.5596 (5596 Feb. 1977), pp. 687–695. ISSN: 1476-4687. DOI: [10.1038/265687a0](https://doi.org/10.1038/265687a0). URL: <https://www.nature.com/articles/265687a0> (visited on 05/17/2022) (cit. on pp. 11, 17).
- [3] Colin T. Archer, Jihyun F. Kim, et al. “The genome sequence of E. coli W (ATCC 9637): comparative genome analysis and an improved genome-scale reconstruction of E. coli”. In: *BMC Genomics* 12.1 (Jan. 6, 2011), p. 9. DOI: [10.1186/1471-2164-12-9](https://doi.org/10.1186/1471-2164-12-9). URL: <https://doi.org/10.1186/1471-2164-12-9> (cit. on p. 11).
- [4] Sergey Nurk, Sergey Koren, et al. “The complete sequence of a human genome”. In: *Science* 376.6588 (Apr. 2022). Publisher: American Association for the Advancement of Science, pp. 44–53. DOI: [10.1126/science.abj6987](https://doi.org/10.1126/science.abj6987). URL: <https://www.science.org/doi/10.1126/science.abj6987> (cit. on pp. 11, 19, 22, 71, 78).
- [5] Jaume Pellicer, Michael F. Fay, and Ilia J. Leitch. “The largest eukaryotic genome of them all?” In: *Botanical Journal of the Linnean Society* 164.1 (Sept. 1, 2010), pp. 10–15. DOI: [10.1111/j.1095-8339.2010.01072.x](https://doi.org/10.1111/j.1095-8339.2010.01072.x). URL: <https://doi.org/10.1111/j.1095-8339.2010.01072.x> (cit. on p. 12).
- [6] H. C. Macgregor. “C-Value Paradox”. In: ed. by Sydney Brenner and Jefferey H. Miller. DOI: [10.1006/rwgn.2001.0301](https://doi.org/10.1006/rwgn.2001.0301). New York: Academic Press, Jan. 1, 2001, pp. 249–250. DOI: [10.1006/rwgn.2001.0301](https://doi.org/10.1006/rwgn.2001.0301). URL: <https://www.sciencedirect.com/science/article/pii/B0122270800003013> (cit. on p. 12).
- [7] Bruce Alberts, Alexander Johnson, et al. *Molecular Biology of the Cell. 4th edition.* Garland Science, 2002. URL: <https://www.ncbi.nlm.nih.gov/books/NBK26916/> (cit. on p. 12).
- [8] F. H. C. Crick, Leslie Barnett, et al. “General Nature of the Genetic Code for Proteins”. In: *Nature* 192.4809 (Dec. 1961). Number: 4809 Publisher: Nature Publishing Group, pp. 1227–1232. DOI: [10.1038/1921227a0](https://doi.org/10.1038/1921227a0). URL: <https://www.nature.com/articles/1921227a0> (cit. on p. 12).

## REFERENCES

- [9] International Human Genome Sequencing Consortium. “Finishing the euchromatic sequence of the human genome”. In: *Nature* 431.7011 (Oct. 2004). Number: 7011 Publisher: Nature Publishing Group, pp. 931–945. DOI: [10.1038/nature03001](https://doi.org/10.1038/nature03001). URL: <https://www.nature.com/articles/nature03001> (cit. on p. 13).
- [10] Ran Elkon and Reuven Agami. “Characterization of noncoding regulatory DNA in the human genome”. In: *Nature Biotechnology* 35.8 (Aug. 2017). Number: 8 Publisher: Nature Publishing Group, pp. 732–746. DOI: [10.1038/nbt.3863](https://doi.org/10.1038/nbt.3863). URL: <https://www.nature.com/articles/nbt.3863> (cit. on p. 13).
- [11] Gilbert S. Omenn. “Reflections on the HUPO Human Proteome Project, the Flagship Project of the Human Proteome Organization, at 10 Years”. In: *Molecular & Cellular Proteomics : MCP* 20 (Feb. 26, 2021). PMID: 33640492 PMCID: PMC8058560, p. 100062. DOI: [10.1016/j.mcpro.2021.100062](https://doi.org/10.1016/j.mcpro.2021.100062). URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8058560/> (cit. on p. 13).
- [12] Svetlana A. Shabalina and Nikolay A. Spiridonov. “The mammalian transcriptome and the function of non-coding DNA sequences”. In: *Genome Biology* 5.4 (Mar. 25, 2004), p. 105. DOI: [10.1186/gb-2004-5-4-105](https://doi.org/10.1186/gb-2004-5-4-105). URL: <https://doi.org/10.1186/gb-2004-5-4-105> (cit. on p. 13).
- [13] ENCODE Project Consortium. “An Integrated Encyclopedia of DNA Elements in the Human Genome”. In: *Nature* 489.7414 (Sept. 6, 2012). PMID: 22955616 PMCID: PMC3439153, pp. 57–74. DOI: [10.1038/nature11247](https://doi.org/10.1038/nature11247). URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3439153/> (cit. on p. 13).
- [14] Nimrat Chatterjee and Graham C. Walker. “Mechanisms of DNA damage, repair, and mutagenesis: DNA Damage and Repair”. In: *Environmental and Molecular Mutagenesis* 58.5 (June 2017), pp. 235–263. DOI: [10.1002/em.22087](https://doi.org/10.1002/em.22087). URL: <https://onlinelibrary.wiley.com/doi/10.1002/em.22087> (cit. on p. 13).
- [15] Iwona J. Fijalkowska, Roel M. Schaaper, and Piotr Jonczyk. “DNA replication fidelity in Escherichia coli: a multi-DNA polymerase affair”. In: *FEMS microbiology reviews* 36.6 (Nov. 2012). PMID: 22404288 PMCID: PMC3391330, pp. 1105–1121. DOI: [10.1111/j.1574-6976.2012.00338.x](https://doi.org/10.1111/j.1574-6976.2012.00338.x). URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3391330/> (cit. on p. 13).
- [16] Leslie Pray. “DNA replication and causes of mutation”. In: *Nature education* 1.1 (2008), p. 214 (cit. on p. 13).
- [17] Jean-François Gout, W. Kelley Thomas, et al. “Large-scale detection of in vivo transcription errors”. In: *Proceedings of the National Academy of Sciences* 110.46 (Nov. 12, 2013). Publisher: Proceedings of the National Academy of Sciences, pp. 18584–18589. DOI: [10.1073/pnas.1309843110](https://doi.org/10.1073/pnas.1309843110). URL: <https://www.pnas.org/doi/full/10.1073/pnas.1309843110> (cit. on p. 13).
- [18] Jean-François Gout, Weiyi Li, et al. “The landscape of transcription errors in eukaryotic cells”. In: *Science Advances* 3.10 (Oct. 20, 2017). PMID: 29062891 PMCID: PMC5650487, e1701484. DOI: [10.1126/sciadv.1701484](https://doi.org/10.1126/sciadv.1701484). URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5650487/> (cit. on p. 13).

- [19] Dmitri Shcherbakov, Youjin Teo, et al. “Ribosomal mistranslation leads to silencing of the unfolded protein response and increased mitochondrial biogenesis”. In: *Communications Biology* 2.1 (Oct. 17, 2019). Number: 1 Publisher: Nature Publishing Group, pp. 1–16. doi: [10.1038/s42003-019-0626-9](https://doi.org/10.1038/s42003-019-0626-9). URL: <https://www.nature.com/articles/s42003-019-0626-9> (cit. on p. 14).
- [20] Omar Desouky, Nan Ding, and Guangming Zhou. “Targeted and non-targeted effects of ionizing radiation”. In: *Journal of Radiation Research and Applied Sciences* 8.2 (Apr. 1, 2015), pp. 247–254. doi: [10.1016/j.jrras.2015.03.003](https://doi.org/10.1016/j.jrras.2015.03.003). URL: <https://www.sciencedirect.com/science/article/pii/S1687850715000333> (cit. on p. 14).
- [21] Jürgen Kiefer. “Effects of Ultraviolet Radiation on DNA”. In: ed. by Günter Obe and Vijayalaxmi. DOI: [10.1007/978-3-540-71414-9\\_3](https://doi.org/10.1007/978-3-540-71414-9_3). Berlin, Heidelberg: Springer, 2007, pp. 39–53. doi: [10.1007/978-3-540-71414-9\\_3](https://doi.org/10.1007/978-3-540-71414-9_3). URL: [https://doi.org/10.1007/978-3-540-71414-9\\_3](https://doi.org/10.1007/978-3-540-71414-9_3) (cit. on p. 14).
- [22] J. W. Bennett and M. Klich. “Mycotoxins”. In: *Clinical Microbiology Reviews* 16.3 (July 2003). PMID: 12857779 PMCID: PMC164220, pp. 497–516. doi: [10.1128/CMR.16.3.497-516.2003](https://doi.org/10.1128/CMR.16.3.497-516.2003). URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC164220/> (cit. on p. 14).
- [23] O.L. Kantidze, A.K. Velichko, et al. “Heat Stress-Induced DNA Damage”. In: *Acta Naturae* 8.2 (2016). PMID: 27437141 PMCID: PMC4947990, pp. 75–78. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4947990/> (cit. on p. 14).
- [24] C. D. Gregory and A. E. Milner. “Regulation of cell survival in Burkitt lymphoma: implications from studies of apoptosis following cold-shock treatment”. In: *International Journal of Cancer* 57.3 (May 1, 1994). PMID: 8169005, pp. 419–426. doi: [10.1002/ijc.2910570321](https://doi.org/10.1002/ijc.2910570321) (cit. on p. 14).
- [25] Anat Gafter-Gvili, Boris Zingerman, et al. “Oxidative Stress-Induced DNA Damage and Repair in Human Peripheral Blood Mononuclear Cells: Protective Role of Hemoglobin”. In: *PLoS ONE* 8.7 (July 9, 2013). PMID: 23874593 PMCID: PMC3706398, e68341. doi: [10.1371/journal.pone.0068341](https://doi.org/10.1371/journal.pone.0068341). URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3706398/> (cit. on p. 14).
- [26] J. R. Roth. “Frameshift mutations”. In: *Annual Review of Genetics* 8 (1974). tex.ids= rothFRAMESHIFTMUTATIONS1974 PMID: 4613260, pp. 319–346. doi: [10.1146/annurev.ge.08.120174.001535](https://doi.org/10.1146/annurev.ge.08.120174.001535) (cit. on p. 14).
- [27] Jody Lynn Kujovich. “Factor V Leiden thrombophilia”. In: *Genetics in Medicine* 13.1 (Jan. 1, 2011), pp. 1–16. doi: [10.1097/GIM.0b013e3181faa0f2](https://doi.org/10.1097/GIM.0b013e3181faa0f2). URL: <https://www.sciencedirect.com/science/article/pii/S1098360021040430> (cit. on p. 14).
- [28] Garry R. Cutting. “Cystic fibrosis genetics: from molecular understanding to clinical application”. In: *Nature Reviews Genetics* 16.1 (Jan. 2015). Number: 1 Publisher: Nature Publishing Group, pp. 45–56. doi: [10.1038/nrg3849](https://doi.org/10.1038/nrg3849). URL: <https://www.nature.com/articles/nrg3849> (cit. on p. 14).

## REFERENCES

- [29] Christian Fuchsberger, Jason Flannick, et al. “The genetic architecture of type 2 diabetes”. In: *Nature* 536.7614 (Aug. 2016). Number: 7614 Publisher: Nature Publishing Group, pp. 41–47. DOI: [10.1038/nature18642](https://doi.org/10.1038/nature18642). URL: <https://www.nature.com/articles/nature18642> (cit. on p. 14).
- [30] Andrew P Morris, Benjamin F Voight, et al. “Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes”. In: *Nature genetics* 44.9 (Sept. 2012). PMID: 22885922 PMCID: PMC3442244, pp. 981–990. DOI: [10.1038/ng.2383](https://doi.org/10.1038/ng.2383). URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3442244/> (cit. on p. 14).
- [31] N. Woodford and M. J. Ellington. “The emergence of antibiotic resistance by mutation”. In: *Clinical Microbiology and Infection* 13.1 (Jan. 1, 2007), pp. 5–18. DOI: [10.1111/j.1469-0691.2006.01492.x](https://doi.org/10.1111/j.1469-0691.2006.01492.x). URL: <https://www.sciencedirect.com/science/article/pii/S1198743X14615500> (cit. on p. 14).
- [32] Soo-Yon Rhee, Matthew J. Gonzales, et al. “Human immunodeficiency virus reverse transcriptase and protease sequence database”. In: *Nucleic Acids Research* 31.1 (Jan. 1, 2003), pp. 298–303. DOI: [10.1093/nar/gkg100](https://doi.org/10.1093/nar/gkg100). URL: <https://academic.oup.com/nar/article/31/1/298/2401450> (cit. on pp. 14, 103).
- [33] F. Sanger, S. Nicklen, and A. R. Coulson. “DNA Sequencing with Chain-Terminating Inhibitors”. In: *Proceedings of the National Academy of Sciences* 74.12 (Dec. 1977), pp. 5463–5467. DOI: [10.1073/pnas.74.12.5463](https://doi.org/10.1073/pnas.74.12.5463). URL: <https://www.pnas.org/doi/abs/10.1073/pnas.74.12.5463> (visited on 05/17/2022) (cit. on p. 15).
- [34] Lloyd M. Smith, Steven Fung, et al. “The Synthesis of Oligonucleotides Containing an Aliphatic Amino Group at the 5 Terminus: Synthesis of Fluorescent DNA Primers for Use in DNA Sequence Analysis”. In: *Nucleic Acids Research* 13.7 (Apr. 11, 1985), pp. 2399–2412. ISSN: 0305-1048. DOI: [10.1093/nar/13.7.2399](https://doi.org/10.1093/nar/13.7.2399). URL: <https://doi.org/10.1093/nar/13.7.2399> (visited on 05/17/2022) (cit. on p. 17).
- [35] Lloyd M. Smith, Jane Z. Sanders, et al. “Fluorescence Detection in Automated DNA Sequence Analysis”. In: *Nature* 321.6071 (6071 June 1986), pp. 674–679. ISSN: 1476-4687. DOI: [10.1038/321674a0](https://doi.org/10.1038/321674a0). URL: <https://www.nature.com/articles/321674a0> (visited on 05/17/2022) (cit. on p. 17).
- [36] Wilhelm Ansorge, Brian Sproat, et al. “Automated DNA sequencing: ultrasensitive detection of fluorescent bands during electrophoresis”. In: *Nucleic Acids Research* 15.11 (June 11, 1987), pp. 4593–4602. DOI: [10.1093/nar/15.11.4593](https://doi.org/10.1093/nar/15.11.4593). URL: <https://doi.org/10.1093/nar/15.11.4593> (cit. on p. 17).
- [37] Jay Shendure and Hanlee Ji. “Next-generation DNA sequencing”. In: *Nature Biotechnology* 26.10 (Oct. 2008). Number: 10 Publisher: Nature Publishing Group, pp. 1135–1145. DOI: [10.1038/nbt1486](https://doi.org/10.1038/nbt1486). URL: <https://www.nature.com/articles/nbt1486> (cit. on p. 17).

- [38] Francis S. Collins, Michael Morgan, and Aristides Patrinos. “The Human Genome Project: Lessons from Large-Scale Biology”. In: *Science* 300.5617 (Apr. 11, 2003). Publisher: American Association for the Advancement of Science, pp. 286–290. DOI: [10.1126/science.1084564](https://doi.org/10.1126/science.1084564). URL: <https://www.science.org/doi/10.1126/science.1084564> (cit. on p. 17).
- [39] Lin Liu, Yinhui Li, et al. “Comparison of Next-Generation Sequencing Systems”. In: *Journal of Biomedicine and Biotechnology* 2012 (July 5, 2012), e251364. ISSN: 2314-6133. DOI: [10.1155/2012/251364](https://doi.org/10.1155/2012/251364). URL: <https://www.hindawi.com/journals/bmri/2012/251364/> (visited on 05/16/2022) (cit. on pp. 17, 19).
- [40] *The Cost of Sequencing a Human Genome*. URL: <https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost> (cit. on p. 18).
- [41] Michael L. Metzker. “Sequencing Technologies — the next Generation”. In: *Nature Reviews Genetics* 11.1 (1 Jan. 2010), pp. 31–46. ISSN: 1471-0064. DOI: [10.1038/nrg2626](https://doi.org/10.1038/nrg2626). URL: <https://www.nature.com/articles/nrg2626> (visited on 05/16/2022) (cit. on p. 18).
- [42] Bruno Canard and Robert S. Sarfati. “DNA polymerase fluorescent substrates with reversible 3'-tags”. In: *Gene* 148.1 (Oct. 11, 1994), pp. 1–6. DOI: [10.1016/0378-1119\(94\)90226-7](https://doi.org/10.1016/0378-1119(94)90226-7). URL: <https://www.sciencedirect.com/science/article/pii/0378111994902267> (cit. on p. 18).
- [43] P. Nyren, B. Pettersson, and M. Uhlen. “Solid Phase DNA Minisequencing by an Enzymatic Luminometric Inorganic Pyrophosphate Detection Assay”. In: *Analytical Biochemistry* 208.1 (Jan. 1, 1993), pp. 171–175. DOI: [10.1006/abio.1993.1024](https://doi.org/10.1006/abio.1993.1024). URL: <https://www.sciencedirect.com/science/article/pii/S0003269783710249> (cit. on p. 19).
- [44] Elaine R. Mardis. “A decade’s perspective on DNA sequencing technology”. In: *Nature* 470.7333 (Feb. 2011). Number: 7333 Publisher: Nature Publishing Group, pp. 198–203. DOI: [10.1038/nature09796](https://doi.org/10.1038/nature09796). URL: <https://www.nature.com/articles/nature09796> (cit. on p. 19).
- [45] Nicholas Stoler and Anton Nekrutenko. “Sequencing error profiles of Illumina sequencing instruments”. In: *NAR Genomics and Bioinformatics* 3.1 (Mar. 1, 2021), lqab019. DOI: [10.1093/nargab/lqab019](https://doi.org/10.1093/nargab/lqab019). URL: <https://doi.org/10.1093/nargab/lqab019> (cit. on p. 19).
- [46] *Sequencing Technology / Sequencing by synthesis*. URL: <https://www.illumina.com/science/technology/next-generation-sequencing/sequencing-technology.html> (cit. on p. 19).
- [47] Martin O Pollard, Deepti Gurdasani, et al. “Long reads: their purpose and place”. In: *Human Molecular Genetics* 27.R2 (Aug. 1, 2018), R234–R241. DOI: [10.1093/hmg/ddy177](https://doi.org/10.1093/hmg/ddy177). URL: <https://doi.org/10.1093/hmg/ddy177> (cit. on p. 19).

## REFERENCES

- [48] John Eid, Adrian Fehr, et al. “Real-Time DNA Sequencing from Single Polymerase Molecules”. In: *Science* 323.5910 (Jan. 2, 2009). Publisher: American Association for the Advancement of Science, pp. 133–138. DOI: [10.1126/science.1162986](https://doi.org/10.1126/science.1162986). URL: <https://www.science.org/doi/10.1126/science.1162986> (cit. on p. 19).
- [49] M. J. Levine, J. Korlach, et al. “Zero-Mode Waveguides for Single-Molecule Analysis at High Concentrations”. In: *Science* 299.5607 (Jan. 31, 2003). Publisher: American Association for the Advancement of Science, pp. 682–686. DOI: [10.1126/science.1079700](https://doi.org/10.1126/science.1079700). URL: <https://www.science.org/doi/10.1126/science.1079700> (cit. on p. 20).
- [50] James Clarke, Hai-Chen Wu, et al. “Continuous base identification for single-molecule nanopore DNA sequencing”. In: *Nature Nanotechnology* 4.4 (Apr. 2009). Number: 4 Publisher: Nature Publishing Group, pp. 265–270. DOI: [10.1038/nnano.2009.12](https://doi.org/10.1038/nnano.2009.12). URL: <https://www.nature.com/articles/nnano.2009.12> (cit. on p. 20).
- [51] David Deamer, Mark Akeson, and Daniel Branton. “Three decades of nanopore sequencing”. In: *Nature Biotechnology* 34.5 (May 2016). Number: 5 Publisher: Nature Publishing Group, pp. 518–524. DOI: [10.1038/nbt.3423](https://doi.org/10.1038/nbt.3423). URL: <https://www.nature.com/articles/nbt.3423> (cit. on p. 20).
- [52] Ryan R. Wick, Louise M. Judd, and Kathryn E. Holt. “Performance of neural network basecalling tools for Oxford Nanopore sequencing”. In: *Genome Biology* 20.1 (June 24, 2019), p. 129. DOI: [10.1186/s13059-019-1727-y](https://doi.org/10.1186/s13059-019-1727-y). URL: <https://doi.org/10.1186/s13059-019-1727-y> (cit. on pp. 20, 23, 25).
- [53] Anthony Rhoads and Kin Fai Au. “PacBio Sequencing and Its Applications”. In: *Genomics, Proteomics & Bioinformatics*. SI: Metagenomics of Marine Environments 13.5 (Oct. 1, 2015), pp. 278–289. DOI: [10.1016/j.gpb.2015.08.002](https://doi.org/10.1016/j.gpb.2015.08.002). URL: <https://www.sciencedirect.com/science/article/pii/S1672022915001345> (cit. on p. 20).
- [54] Camilla L.C. Ip, Matthew Loose, et al. “MinION Analysis and Reference Consortium: Phase 1 data release and analysis”. In: *F1000Research* 4 (Oct. 15, 2015). PMID: 26834992 PMCID: PMC4722697, p. 1075. DOI: [10.12688/f1000research.7201.1](https://doi.org/10.12688/f1000research.7201.1). URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4722697/> (cit. on p. 20).
- [55] Glennis A. Logsdon, Mitchell R. Vollger, and Evan E. Eichler. “Long-Read Human Genome Sequencing and Its Applications”. In: *Nature Reviews Genetics* 21.10 (10 Oct. 2020), pp. 597–614. ISSN: 1471-0064. DOI: [10.1038/s41576-020-0236-x](https://doi.org/10.1038/s41576-020-0236-x). URL: <https://www.nature.com/articles/s41576-020-0236-x> (visited on 05/16/2022) (cit. on pp. 20, 21, 23, 25).
- [56] Miten Jain, Sergey Koren, et al. “Nanopore sequencing and assembly of a human genome with ultra-long reads”. In: *Nature Biotechnology* 36.4 (Apr. 2018). Number: 4 Publisher: Nature Publishing Group, pp. 338–345. DOI: [10.1038/nbt.4060](https://doi.org/10.1038/nbt.4060). URL: <https://www.nature.com/articles/nbt.4060> (cit. on p. 20).

- [57] *Thar she blows! Ultra long read method for nanopore sequencing* · Loman Labs. URL: <http://lab.loman.net/2017/03/09/ultrareads-for-nanopore/> (cit. on p. 20).
- [58] Alexander Payne, Nadine Holmes, et al. “BulkVis: a graphical viewer for Oxford nanopore bulk FAST5 files”. In: *Bioinformatics* 35.13 (July 1, 2019), pp. 2193–2198. DOI: [10.1093/bioinformatics/bty841](https://doi.org/10.1093/bioinformatics/bty841). URL: <https://doi.org/10.1093/bioinformatics/bty841> (cit. on p. 20).
- [59] Valentine Murigneux, Subash Kumar Rai, et al. “Comparison of long-read methods for sequencing and assembly of a plant genome”. In: *GigaScience* 9.12 (Nov. 30, 2020), giaa146. DOI: [10.1093/gigascience/giaa146](https://doi.org/10.1093/gigascience/giaa146). URL: <https://doi.org/10.1093/gigascience/giaa146> (cit. on p. 20).
- [60] Mark J. P. Chaisson, John Huddleston, et al. “Resolving the complexity of the human genome using single-molecule sequencing”. In: *Nature* 517.7536 (Jan. 2015). Number: 7536 Publisher: Nature Publishing Group, pp. 608–611. DOI: [10.1038/nature13907](https://doi.org/10.1038/nature13907). URL: <https://www.nature.com/articles/nature13907> (cit. on p. 21).
- [61] Miten Jain, Hugh E. Olsen, et al. “The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community”. In: *Genome Biology* 17.1 (Nov. 25, 2016), p. 239. DOI: [10.1186/s13059-016-1103-0](https://doi.org/10.1186/s13059-016-1103-0). URL: <https://doi.org/10.1186/s13059-016-1103-0> (cit. on p. 21).
- [62] Mingye Hong, Shuang Tao, et al. “RNA sequencing: new technologies and applications in cancer research”. In: *Journal of Hematology & Oncology* 13.1 (Dec. 4, 2020), p. 166. DOI: [10.1186/s13045-020-01005-x](https://doi.org/10.1186/s13045-020-01005-x). URL: <https://doi.org/10.1186/s13045-020-01005-x> (cit. on p. 21).
- [63] Fatih Ozsolak and Patrice M. Milos. “RNA sequencing: advances, challenges and opportunities”. In: *Nature Reviews Genetics* 12.2 (Feb. 2011). Number: 2 Publisher: Nature Publishing Group, pp. 87–98. DOI: [10.1038/nrg2934](https://doi.org/10.1038/nrg2934). URL: <https://www.nature.com/articles/nrg2934> (cit. on p. 21).
- [64] D F Hunt, J R Yates, et al. “Protein sequencing by tandem mass spectrometry.” In: *Proceedings of the National Academy of Sciences* 83.17 (Sept. 1986). Publisher: Proceedings of the National Academy of Sciences, pp. 6233–6237. DOI: [10.1073/pnas.83.17.6233](https://doi.org/10.1073/pnas.83.17.6233). URL: <https://www.pnas.org/doi/abs/10.1073/pnas.83.17.6233> (cit. on p. 21).
- [65] Bryan John Smith. *Protein Sequencing Protocols*. Springer Science & Business Media, 2002 (cit. on p. 21).
- [66] Laura Restrepo-Pérez, Chirlmin Joo, and Cees Dekker. “Paving the way to single-molecule protein sequencing”. In: *Nature Nanotechnology* 13.9 (Sept. 2018). Number: 9 Publisher: Nature Publishing Group, pp. 786–796. DOI: [10.1038/s41565-018-0236-6](https://doi.org/10.1038/s41565-018-0236-6). URL: <https://www.nature.com/articles/s41565-018-0236-6> (cit. on p. 21).

## REFERENCES

- [67] Jason L Weirather, Mariateresa de Cesare, et al. “Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis”. In: *F1000Research* 6 (June 19, 2017). PMID: 28868132 PMCID: PMC5553090, p. 100. DOI: [10.12688/f1000research.10571.2](https://doi.org/10.12688/f1000research.10571.2). URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5553090/> (cit. on pp. 21, 24).
- [68] Yunhao Wang, Yue Zhao, et al. “Nanopore sequencing technology, bioinformatics and applications”. In: *Nature Biotechnology* 39.11 (Nov. 2021). Number: 11 Publisher: Nature Publishing Group, pp. 1348–1365. DOI: [10.1038/s41587-021-01108-x](https://doi.org/10.1038/s41587-021-01108-x). URL: <https://www.nature.com/articles/s41587-021-01108-x> (cit. on p. 21).
- [69] Xiaotu Ma, Ying Shao, et al. “Analysis of error profiles in deep next-generation sequencing data”. In: *Genome Biology* 20.1 (Mar. 14, 2019), p. 50. DOI: [10.1186/s13059-019-1659-6](https://doi.org/10.1186/s13059-019-1659-6). URL: <https://doi.org/10.1186/s13059-019-1659-6> (cit. on p. 22).
- [70] Leandro Lima, Camille Marchet, et al. “Comparative assessment of long-read error correction software applied to Nanopore RNA-sequencing data”. In: *Briefings in Bioinformatics* 21.4 (July 15, 2020), pp. 1164–1181. DOI: [10.1093/bib/bbz058](https://doi.org/10.1093/bib/bbz058). URL: <https://doi.org/10.1093/bib/bbz058> (cit. on p. 22).
- [71] Shuhua Fu, Anqi Wang, and Kin Fai Au. “A comparative evaluation of hybrid error correction methods for error-prone long reads”. In: *Genome Biology* 20.1 (Feb. 4, 2019), p. 26. DOI: [10.1186/s13059-018-1605-z](https://doi.org/10.1186/s13059-018-1605-z). URL: <https://doi.org/10.1186/s13059-018-1605-z> (cit. on p. 22).
- [72] Haowen Zhang, Chirag Jain, and Srinivas Aluru. “A comprehensive evaluation of long read error correction methods”. In: *BMC Genomics* 21.6 (Dec. 21, 2020), p. 889. DOI: [10.1186/s12864-020-07227-0](https://doi.org/10.1186/s12864-020-07227-0). URL: <https://doi.org/10.1186/s12864-020-07227-0> (cit. on p. 22).
- [73] Shanika L. Amarasinghe, Shian Su, et al. “Opportunities and Challenges in Long-Read Sequencing Data Analysis”. In: *Genome Biology* 21.1 (Feb. 7, 2020), p. 30. ISSN: 1474-760X. DOI: [10.1186/s13059-020-1935-5](https://doi.org/10.1186/s13059-020-1935-5). URL: <https://doi.org/10.1186/s13059-020-1935-5> (visited on 05/16/2022) (cit. on pp. 22, 23, 25).
- [74] Jue Ruan and Heng Li. “Fast and accurate long-read assembly with wtdbg2”. In: *Nature Methods* 17.2 (Feb. 2020). Number: 2 Publisher: Nature Publishing Group, pp. 155–158. DOI: [10.1038/s41592-019-0669-3](https://doi.org/10.1038/s41592-019-0669-3). URL: <https://www.nature.com/articles/s41592-019-0669-3> (cit. on pp. 22, 25).
- [75] Sergey Koren, Brian P. Walenz, et al. “Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation”. In: *Genome Research* 27.5 (Jan. 5, 2017). Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab PMID: 28298431, pp. 722–736. DOI: [10.1101/gr.215087.116](https://doi.org/10.1101/gr.215087.116). URL: <https://genome.cshlp.org/content/27/5/722> (cit. on p. 22).

- [76] German Tischler and Eugene W. Myers. *Non Hybrid Long Read Consensus Using Local De Bruijn Graph Assembly*. Tech. rep. DOI: 10.1101/106252 Section: New Results Type: article. Feb. 6, 2017, p. 106252. doi: [10.1101/106252](https://doi.org/10.1101/106252). URL: <https://www.biorxiv.org/content/10.1101/106252v1> (cit. on p. 22).
- [77] René L Warren, Lauren Coombe, et al. “ntEdit: scalable genome sequence polishing”. In: *Bioinformatics* 35.21 (Nov. 1, 2019), pp. 4430–4432. doi: [10.1093/bioinformatics/btz400](https://doi.org/10.1093/bioinformatics/btz400). URL: <https://doi.org/10.1093/bioinformatics/btz400> (cit. on p. 22).
- [78] N Lance Hepler, M Brown, et al. “An improved circular consensus algorithm with an application to detect HIV-1 Drug-Resistance associated mutations (DRAMs)”. In: 2016 (cit. on p. 22).
- [79] Jared T. Simpson, Rachael E. Workman, et al. “Detecting DNA cytosine methylation using nanopore sequencing”. In: *Nature Methods* 14.4 (Apr. 2017). Number: 4 Publisher: Nature Publishing Group, pp. 407–410. doi: [10.1038/nmeth.4184](https://doi.org/10.1038/nmeth.4184). URL: <https://www.nature.com/articles/nmeth.4184> (cit. on pp. 22, 25).
- [80] Thomas Hackl, Rainer Hedrich, et al. “proovread : large-scale high-accuracy PacBio correction through iterative short read consensus”. In: *Bioinformatics* 30.21 (Nov. 1, 2014), pp. 3004–3011. doi: [10.1093/bioinformatics/btu392](https://doi.org/10.1093/bioinformatics/btu392). URL: <https://doi.org/10.1093/bioinformatics/btu392> (cit. on pp. 22, 42).
- [81] Giles Miclotte, Mahdi Heydari, et al. “Jabba: hybrid error correction for long sequencing reads”. In: *Algorithms for Molecular Biology* 11.1 (May 3, 2016), p. 10. doi: [10.1186/s13015-016-0075-7](https://doi.org/10.1186/s13015-016-0075-7). URL: <https://doi.org/10.1186/s13015-016-0075-7> (cit. on p. 22).
- [82] Sergey Koren, Michael C. Schatz, et al. “Hybrid error correction and de novo assembly of single-molecule sequencing reads”. In: *Nature Biotechnology* 30.7 (July 2012). Number: 7 Publisher: Nature Publishing Group, pp. 693–700. doi: [10.1038/nbt.2280](https://doi.org/10.1038/nbt.2280). URL: <https://www.nature.com/articles/nbt.2280> (cit. on pp. 22, 42).
- [83] Leena Salmela and Eric Rivals. “LoRDEC: accurate and efficient long read error correction”. In: *Bioinformatics* 30.24 (Dec. 15, 2014), pp. 3506–3514. doi: [10.1093/bioinformatics/btu538](https://doi.org/10.1093/bioinformatics/btu538). URL: <https://doi.org/10.1093/bioinformatics/btu538> (cit. on p. 22).
- [84] Bruce J. Walker, Thomas Abeel, et al. “Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement”. In: *PLOS ONE* 9.11 (Nov. 19, 2014). Publisher: Public Library of Science, e112963. doi: [10.1371/journal.pone.0112963](https://doi.org/10.1371/journal.pone.0112963). URL: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0112963> (cit. on p. 22).
- [85] Robert Vaser, Ivan Sović, et al. “Fast and accurate de novo genome assembly from long uncorrected reads”. In: *Genome Research* 27.5 (May 2017). PMID: 28100585 PMCID: PMC5411768, pp. 737–746. doi: [10.1101/gr.214270.116](https://doi.org/10.1101/gr.214270.116) (cit. on p. 22).

## REFERENCES

- [86] Aaron M. Wenger, Paul Peluso, et al. “Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome”. In: *Nature Biotechnology* 37.10 (Oct. 2019), pp. 1155–1162. DOI: [10.1038/s41587-019-0217-9](https://doi.org/10.1038/s41587-019-0217-9). URL: <http://www.nature.com/articles/s41587-019-0217-9> (cit. on pp. 22, 23, 25).
- [87] Winston Timp, Jeffrey Comer, and Aleksei Aksimentiev. “DNA Base-Calling from a Nanopore Using a Viterbi Algorithm”. In: *Biophysical Journal* 102.10 (May 16, 2012), pp. L37–L39. DOI: [10.1016/j.bpj.2012.04.009](https://doi.org/10.1016/j.bpj.2012.04.009). URL: <https://www.sciencedirect.com/science/article/pii/S0006349512004511> (cit. on p. 23).
- [88] Peter Perešíni, Vladimír Boža, et al. “Nanopore base calling on the edge”. In: *Bioinformatics* 37.24 (Dec. 15, 2021), pp. 4661–4667. DOI: [10.1093/bioinformatics/btab528](https://doi.org/10.1093/bioinformatics/btab528). URL: <https://doi.org/10.1093/bioinformatics/btab528> (cit. on p. 23).
- [89] Vladimír Boža, Broňa Brejová, and Tomáš Vinař. “DeepNano: Deep recurrent neural networks for base calling in MinION nanopore reads”. In: *PLOS ONE* 12.6 (June 5, 2017). Publisher: Public Library of Science, e0178751. DOI: [10.1371/journal.pone.0178751](https://doi.org/10.1371/journal.pone.0178751). URL: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0178751> (cit. on p. 23).
- [90] Søren M. Karst, Ryan M. Ziels, et al. “High-accuracy long-read amplicon sequences using unique molecular identifiers with Nanopore or PacBio sequencing”. In: *Nature Methods* 18.2 (Feb. 2021). Number: 2 Publisher: Nature Publishing Group, pp. 165–169. DOI: [10.1038/s41592-020-01041-y](https://doi.org/10.1038/s41592-020-01041-y). URL: <https://www.nature.com/articles/s41592-020-01041-y>. (cit. on p. 23).
- [91] Zitian Chen, Wenxiong Zhou, et al. “Highly accurate fluorogenic DNA sequencing with information theory-based error correction”. In: *Nature Biotechnology* 35.12 (Dec. 2017). Number: 12 Publisher: Nature Publishing Group, pp. 1170–1178. DOI: [10.1038/nbt.3982](https://doi.org/10.1038/nbt.3982). URL: <https://www.nature.com/articles/nbt.3982> (cit. on p. 23).
- [92] *High Performance Long Read Assay Enables Contiguous Data up to 10Kb on Existing Illumina Platforms*. URL: [https://www.illumina.com/content/illumina-marketing/amr/en\\_US/science/genomics-research/articles/infinity-high-performance-long-read-assay.html](https://www.illumina.com/content/illumina-marketing/amr/en_US/science/genomics-research/articles/infinity-high-performance-long-read-assay.html) (cit. on pp. 23, 26).
- [93] Gergely Ivády, László Madar, et al. “Analytical parameters and validation of homopolymer detection in a pyrosequencing-based next generation sequencing system”. In: *BMC Genomics* 19.1 (Feb. 21, 2018), p. 158. DOI: [10.1186/s12864-018-4544-x](https://doi.org/10.1186/s12864-018-4544-x). URL: <https://doi.org/10.1186/s12864-018-4544-x> (cit. on p. 24).
- [94] A. Sina Booeshaghi and Lior Pachter. “Pseudoalignment facilitates assignment of error-prone Ultima Genomics reads”. In: (). DOI: [10.1101/2022.06.04.494845](https://doi.org/10.1101/2022.06.04.494845) (cit. on p. 24).

- [95] Clara Delahaye and Jacques Nicolas. “Sequencing DNA with nanopores: Troubles and biases”. In: *PLOS ONE* 16.10 (Oct. 1, 2021). Publisher: Public Library of Science, e0257521. DOI: [10.1371/journal.pone.0257521](https://doi.org/10.1371/journal.pone.0257521). URL: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0257521> (cit. on p. 24).
- [96] Sara Goodwin, James Gurtowski, et al. “Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome”. In: *Genome Research* 25.11 (Jan. 11, 2015). Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab PMID: 26447147, pp. 1750–1756. DOI: [10.1101/gr.191395.115](https://doi.org/10.1101/gr.191395.115). URL: <https://genome.cshlp.org/content/25/11/1750> (cit. on p. 24).
- [97] Juliane C Dohm, Philipp Peters, et al. “Benchmarking of Long-Read Correction Methods”. In: *NAR Genomics and Bioinformatics* 2.2 (June 1, 2020). ISSN: 2631-9268. DOI: [10.1093/nargab/lqaa037](https://doi.org/10.1093/nargab/lqaa037) (cit. on pp. 24, 70).
- [98] Jonathan Foox, Scott W. Tighe, et al. “Performance assessment of DNA sequencing platforms in the ABRF Next-Generation Sequencing Study”. In: *Nature Biotechnology* 39.9 (Sept. 2021). Number: 9 Publisher: Nature Publishing Group, pp. 1129–1140. DOI: [10.1038/s41587-021-01049-5](https://doi.org/10.1038/s41587-021-01049-5). URL: <https://www.nature.com/articles/s41587-021-01049-5> (cit. on p. 25).
- [99] Yao-Ting Huang, Po-Yu Liu, and Pei-Wen Shih. “Homopolish: a method for the removal of systematic errors in nanopore sequencing by homologous polishing”. In: *Genome Biology* 22.1 (Mar. 31, 2021), p. 95. DOI: [10.1186/s13059-021-02282-6](https://doi.org/10.1186/s13059-021-02282-6). URL: <https://doi.org/10.1186/s13059-021-02282-6> (cit. on p. 25).
- [100] Franka J. Rang, Wigard P. Kloosterman, and Jeroen de Ridder. “From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy”. In: *Genome Biology* 19.1 (July 13, 2018), p. 90. DOI: [10.1186/s13059-018-1462-9](https://doi.org/10.1186/s13059-018-1462-9). URL: <https://doi.org/10.1186/s13059-018-1462-9> (cit. on p. 25).
- [101] Peter Sarkozy, Ákos Jobbág, and Peter Antal. “Calling Homopolymer Stretches from Raw Nanopore Reads by Analyzing k-mer Dwell Times”. In: ed. by Hannu Eskola, Outi Väisänen, et al. IFMBE Proceedings. Singapore: Springer, 2018, pp. 241–244. DOI: [10.1007/978-981-10-5122-7\\_61](https://doi.org/10.1007/978-981-10-5122-7_61) (cit. on p. 25).
- [102] *R10.3: the newest nanopore for high accuracy nanopore sequencing – now available in store*. Section: News. URL: <http://nanoporetech.com/about-us/news/r103-newest-nanopore-high-accuracy-nanopore-sequencing-now-available-store> (cit. on p. 25).
- [103] Lei Zhou, Kun Li, et al. “Detection of DNA homopolymer with graphene nanopore”. In: *Journal of Vacuum Science & Technology B* 37.6 (Nov. 2019). Publisher: American Vacuum Society, p. 061809. DOI: [10.1116/1.5116295](https://doi.org/10.1116/1.5116295). URL: <https://avs.scitation.org/doi/full/10.1116/1.5116295> (cit. on p. 25).

## REFERENCES

- [104] Yusuke Goto, Itaru Yanagi, et al. “Identification of four single-stranded DNA homopolymers with a solid-state nanopore in alkaline CsCl solution”. In: *Nanoscale* 10.44 (2018). Publisher: Royal Society of Chemistry, pp. 20844–20850. DOI: [10.1039/C8NR04238A](https://doi.org/10.1039/C8NR04238A). URL: <https://pubs.rsc.org/en/content/articlelanding/2018/nr/c8nr04238a> (cit. on p. 25).
- [105] John A. Hawkins, Stephen K. Jones, et al. “Indel-correcting DNA barcodes for high-throughput sequencing”. In: *Proceedings of the National Academy of Sciences* 115.27 (July 3, 2018). Publisher: Proceedings of the National Academy of Sciences, E6217–E6226. DOI: [10.1073/pnas.1802640115](https://doi.org/10.1073/pnas.1802640115). URL: <https://www.pnas.org/doi/full/10.1073/pnas.1802640115> (cit. on p. 25).
- [106] Amrita Srivathsan, Bilgenur Baloğlu, et al. “A MinION™-based pipeline for fast and cost-effective DNA barcoding”. In: *Molecular Ecology Resources* 18.5 (2018). \_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/1755-0998.12890>, pp. 1035–1049. DOI: [10.1111/1755-0998.12890](https://doi.org/10.1111/1755-0998.12890). URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/1755-0998.12890> (cit. on p. 25).
- [107] Yixin Wang, Md. Noor-A-Rahim, et al. “Construction of Bio-Constrained Code for DNA Data Storage”. In: *IEEE Communications Letters* 23.6 (June 2019). Conference Name: IEEE Communications Letters, pp. 963–966. DOI: [10.1109/LCOMM.2019.2912572](https://doi.org/10.1109/LCOMM.2019.2912572) (cit. on p. 25).
- [108] Sergey Nurk, Brian P. Walenz, et al. “HiCanu: Accurate Assembly of Segmental Duplications, Satellites, and Allelic Variants from High-Fidelity Long Reads”. In: *Genome Research* 30.9 (Jan. 9, 2020), pp. 1291–1305. ISSN: 1088-9051, 1549-5469. DOI: [10.1101/gr.263566.120](https://doi.org/10.1101/gr.263566.120). pmid: [32801147](https://pubmed.ncbi.nlm.nih.gov/32801147/) (cit. on pp. 25, 71).
- [109] Barış Ekim, Bonnie Berger, and Rayan Chikhi. “Minimizer-Space de Bruijn Graphs: Whole-Genome Assembly of Long Reads in Minutes on a Personal Computer”. In: *Cell Systems* 12.10 (Oct. 20, 2021), 958–968.e6. ISSN: 2405-4712. DOI: [10.1016/j.cels.2021.08.009](https://doi.org/10.1016/j.cels.2021.08.009) (cit. on pp. 25, 71).
- [110] Jason R. Miller, Arthur L. Delcher, et al. “Aggressive Assembly of Pyrosequencing Reads with Mates”. In: *Bioinformatics* 24.24 (Dec. 15, 2008), pp. 2818–2824. ISSN: 1367-4803. DOI: [10.1093/bioinformatics/btn548](https://doi.org/10.1093/bioinformatics/btn548) (cit. on pp. 25, 71).
- [111] Kishwar Shafin, Trevor Pesout, et al. “Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes”. In: *Nature Biotechnology* 38.9 (Sept. 2020). Number: 9 Publisher: Nature Publishing Group, pp. 1044–1053. DOI: [10.1038/s41587-020-0503-6](https://doi.org/10.1038/s41587-020-0503-6). URL: <https://www.nature.com/articles/s41587-020-0503-6> (cit. on p. 25).
- [112] Kristoffer Sahlin and Paul Medvedev. “De Novo Clustering of Long-Read Transcriptome Data Using a Greedy, Quality Value-Based Algorithm”. In: *Journal of Computational Biology* 27.4 (Apr. 1, 2020), pp. 472–484. DOI: [10.1089/cmb.2019.0299](https://doi.org/10.1089/cmb.2019.0299) (cit. on pp. 25, 71).

- [113] Kin Fai Au, Jason G. Underwood, et al. “Improving PacBio Long Read Accuracy by Short Read Alignment”. In: *PLOS ONE* 7.10 (Oct. 4, 2012), e46679. ISSN: 1932-6203. DOI: [10.1371/journal.pone.0046679](https://doi.org/10.1371/journal.pone.0046679) (cit. on pp. 25, 71).
- [114] RuiFeng Hu, Guibo Sun, and Xiaobo Sun. “LSCplus: a fast solution for improving long read accuracy by short read alignment”. In: *BMC Bioinformatics* 17.1 (Nov. 9, 2016), p. 451. DOI: [10.1186/s12859-016-1316-y](https://doi.org/10.1186/s12859-016-1316-y). URL: <https://doi.org/10.1186/s12859-016-1316-y> (cit. on p. 25).
- [115] Heng Li. “Minimap2: Pairwise Alignment for Nucleotide Sequences”. In: *Bioinformatics* 34.18 (Sept. 15, 2018), pp. 3094–3100. ISSN: 1367-4803. DOI: [10.1093/bioinformatics/bty191](https://doi.org/10.1093/bioinformatics/bty191) (cit. on pp. 25, 48, 51, 71, 78, 79, 159).
- [116] Chirag Jain, Arang Rhie, et al. “Weighted Minimizer Sampling Improves Long Read Mapping”. In: *Bioinformatics* 36 (Supplement\_1 July 1, 2020), pp. i111–i118. ISSN: 1367-4803. DOI: [10.1093/bioinformatics/btaa435](https://doi.org/10.1093/bioinformatics/btaa435) (cit. on pp. 25, 52, 81, 159).
- [117] Christophe Van Neste, Filip Van Nieuwerburgh, et al. “Forensic STR analysis using massive parallel sequencing”. In: *Forensic Science International: Genetics. Analysis and biostatistical interpretation of complex and low template DNA samples* 6.6 (Dec. 1, 2012), pp. 810–818. DOI: [10.1016/j.fsigen.2012.03.004](https://doi.org/10.1016/j.fsigen.2012.03.004). URL: <https://www.sciencedirect.com/science/article/pii/S1872497312000798> (cit. on p. 25).
- [118] Gilad Almogy, Mark Pratt, et al. “Cost-efficient whole genome-sequencing using novel mostly natural sequencing-by-synthesis chemistry and open fluidics platform”. In: (). DOI: [10.1101/2022.05.29.493900](https://doi.org/10.1101/2022.05.29.493900) (cit. on p. 26).
- [119] *Short-read sequencing by binding*. URL: <https://www.pacb.com/technology/sequencing-by-binding/> (cit. on p. 26).
- [120] Arif E. Cetin, Pinar Iyidogan, et al. “Plasmonic Sensor Could Enable Label-Free DNA Sequencing”. In: *ACS Sensors* 3.3 (Mar. 23, 2018). Publisher: American Chemical Society, pp. 561–568. DOI: [10.1021/acssensors.7b00957](https://doi.org/10.1021/acssensors.7b00957). URL: <https://doi.org/10.1021/acssensors.7b00957> (cit. on p. 26).
- [121] Wing-Kin Sung. *Algorithms in Bioinformatics: A Practical Introduction*. DOI: [10.1201/9781420070347](https://doi.org/10.1201/9781420070347). New York: Chapman and Hall/CRC, Oct. 10, 2011. DOI: [10.1201/9781420070347](https://doi.org/10.1201/9781420070347) (cit. on pp. 41, 42, 44, 48, 50).
- [122] Richard Wesley Hamming. *Coding and Information Theory*. tex.ids= hamming1980coding googlebooksid: ed5QAAAAMAAJ lccn: 79015159. Prentice-Hall, 1980 (cit. on p. 41).
- [123] Dan Gusfield. *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*. DOI: [10.1017/CBO9780511574931](https://doi.org/10.1017/CBO9780511574931). Cambridge: Cambridge University Press, 1997. DOI: [10.1017/CBO9780511574931](https://doi.org/10.1017/CBO9780511574931). URL: <https://www.cambridge.org/core/books/algorithms-on-strings-trees-and-sequences/F0B095049C7E6EF5356F0A26686C20D3> (cit. on pp. 41, 48, 52, 70).

## REFERENCES

- [124] V. I. Levenshtein. “Binary Codes Capable of Correcting Deletions, Insertions and Reversals”. In: *Soviet Physics Doklady* 10 (Feb. 1, 1966). ADS Bibcode: 1966SPhD...10..707L, p. 707. URL: <https://ui.adsabs.harvard.edu/abs/1966SPhD..10..707L> (cit. on p. 41).
- [125] Ross C. Hardison. “Comparative Genomics”. In: *PLOS Biology* 1.2 (Nov. 17, 2003). Publisher: Public Library of Science, e58. DOI: [10.1371/journal.pbio.0000058](https://doi.org/10.1371/journal.pbio.0000058). URL: <https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.0000058> (cit. on p. 41).
- [126] Joseph Felsenstein. “Evolutionary trees from DNA sequences: A maximum likelihood approach”. In: *Journal of Molecular Evolution* 17.6 (Nov. 1, 1981), pp. 368–376. DOI: [10.1007/BF01734359](https://doi.org/10.1007/BF01734359). URL: <https://doi.org/10.1007/BF01734359> (cit. on p. 42).
- [127] Sudhir Kumar, Koichiro Tamura, and Masatoshi Nei. “MEGA: Molecular Evolutionary Genetics Analysis software for microcomputers”. In: *Bioinformatics* 10.2 (Apr. 2, 1994), pp. 189–191. DOI: [10.1093/bioinformatics/10.2.189](https://doi.org/10.1093/bioinformatics/10.2.189). URL: <https://doi.org/10.1093/bioinformatics/10.2.189> (cit. on p. 42).
- [128] Alexey M Kozlov, Diego Darriba, et al. “RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference”. In: *Bioinformatics* 35.21 (Nov. 1, 2019), pp. 4453–4455. DOI: [10.1093/bioinformatics/btz305](https://doi.org/10.1093/bioinformatics/btz305). URL: <https://doi.org/10.1093/bioinformatics/btz305> (cit. on p. 42).
- [129] Stéphane Guindon, Jean-François Dufayard, et al. “New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0”. In: *Systematic Biology* 59.3 (May 1, 2010), pp. 307–321. DOI: [10.1093/sysbio/syq010](https://doi.org/10.1093/sysbio/syq010). URL: <https://doi.org/10.1093/sysbio/syq010> (cit. on p. 42).
- [130] Morgan N. Price, Paramvir S. Dehal, and Adam P. Arkin. “FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments”. In: *PLOS ONE* 5.3 (Mar. 10, 2010), e9490. DOI: [10.1371/journal.pone.0009490](https://doi.org/10.1371/journal.pone.0009490). URL: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0009490> (cit. on p. 42).
- [131] John Jumper, Richard Evans, et al. “Highly accurate protein structure prediction with AlphaFold”. In: *Nature* 596.7873 (Aug. 2021). Number: 7873 Publisher: Nature Publishing Group, pp. 583–589. DOI: [10.1038/s41586-021-03819-2](https://doi.org/10.1038/s41586-021-03819-2). URL: <https://www.nature.com/articles/s41586-021-03819-2> (cit. on p. 42).
- [132] Kevin Karplus, Christian Barrett, et al. “Predicting protein structure using only sequence information”. In: *Proteins: Structure, Function, and Bioinformatics* 37.S3 (1999), pp. 121–125. DOI: [10.1002/\(SICI\)1097-0134\(1999\)37:3+<121::AID-PROT16>3.0.CO;2-Q](https://doi.org/10.1002/(SICI)1097-0134(1999)37:3+<121::AID-PROT16>3.0.CO;2-Q). URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291097-0134%281999%2937%3A3%20%3C121%3A%3AAID-PROT16%3E3.0.CO%3B2-Q> (cit. on p. 42).

- [133] James D Watson, Roman A Laskowski, and Janet M Thornton. “Predicting protein function from sequence and structural data”. In: *Current Opinion in Structural Biology*. Sequences and topology/Nucleic acids 15.3 (June 1, 2005), pp. 275–284. DOI: [10.1016/j.sbi.2005.04.003](https://doi.org/10.1016/j.sbi.2005.04.003). URL: <https://www.sciencedirect.com/science/article/pii/S0959440X05000825> (cit. on p. 42).
- [134] David Lee, Oliver Redfern, and Christine Orengo. “Predicting protein function from sequence and structure”. In: *Nature Reviews Molecular Cell Biology* 8.12 (Dec. 2007), pp. 995–1005. DOI: [10.1038/nrm2281](https://doi.org/10.1038/nrm2281). URL: <https://www.nature.com/articles/nrm2281> (cit. on p. 42).
- [135] Leena Salmela and Jan Schröder. “Correcting errors in short reads by multiple alignments”. In: *Bioinformatics* 27.11 (June 1, 2011), pp. 1455–1461. DOI: [10.1093/bioinformatics/btr170](https://doi.org/10.1093/bioinformatics/btr170). URL: <https://doi.org/10.1093/bioinformatics/btr170> (cit. on p. 42).
- [136] Paul Medvedev, Monica Stanciu, and Michael Brudno. “Computational methods for discovering structural variation with next-generation sequencing”. In: *Nature Methods* 6.11 (Nov. 2009). Number: 11 Publisher: Nature Publishing Group, S13–S20. DOI: [10.1038/nmeth.1374](https://doi.org/10.1038/nmeth.1374). URL: <https://www.nature.com/articles/nmeth.1374> (cit. on p. 42).
- [137] Medhat Mahmoud, Nastassia Gobet, et al. “Structural variant calling: the long and the short of it”. In: *Genome Biology* 20.1 (Nov. 20, 2019), p. 246. DOI: [10.1186/s13059-019-1828-7](https://doi.org/10.1186/s13059-019-1828-7). URL: <https://doi.org/10.1186/s13059-019-1828-7> (cit. on p. 42).
- [138] Saul B. Needleman and Christian D. Wunsch. “A general method applicable to the search for similarities in the amino acid sequence of two proteins”. In: *Journal of Molecular Biology* 48.3 (Mar. 28, 1970), pp. 443–453. DOI: [10.1016/0022-2836\(70\)90057-4](https://doi.org/10.1016/0022-2836(70)90057-4). URL: [http://www.sciencedirect.com/science/article/pii/0022283670900574](https://www.sciencedirect.com/science/article/pii/0022283670900574) (cit. on p. 42).
- [139] T. F. Smith and M. S. Waterman. “Identification of common molecular subsequences”. In: *Journal of Molecular Biology* 147.1 (Mar. 25, 1981), pp. 195–197. DOI: [10.1016/0022-2836\(81\)90087-5](https://doi.org/10.1016/0022-2836(81)90087-5). URL: <https://www.sciencedirect.com/science/article/pii/0022283681900875> (cit. on p. 42).
- [140] Stephen P. Bradley, Arnoldo C. Hax, and Thomas L. Magnanti. *Applied Mathematical Programming*. Google-Books-ID: MSWdWv3Gn5cC. Addison-Wesley Publishing Company, 1977 (cit. on p. 42).
- [141] Richard Bellman. “The theory of dynamic programming”. In: *Bulletin of the American Mathematical Society* 60.6 (1954), pp. 503–515. DOI: [10.1090/S0002-9904-1954-09848-8](https://doi.org/10.1090/S0002-9904-1954-09848-8). URL: <https://www.ams.org/bull/1954-60-06/S0002-9904-1954-09848-8/> (cit. on p. 42).

## REFERENCES

- [142] William J. Masek and Michael S. Paterson. “A faster algorithm computing string edit distances”. In: *Journal of Computer and System Sciences* 20.1 (Feb. 1, 1980), pp. 18–31. DOI: [10.1016/0022-0000\(80\)90002-1](https://doi.org/10.1016/0022-0000(80)90002-1). URL: <https://www.sciencedirect.com/science/article/pii/0022000080900021> (cit. on p. 42).
- [143] Nguyen Xuan Vinh, Julien Epps, and James Bailey. “Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance”. en. In: *Journal of Machine Learning Research* 11 (2010), p. 18 (cit. on pp. 42, 134, 171).
- [144] J. D. Ullman, A. V. Aho, and D. S. Hirschberg. “Bounds on the Complexity of the Longest Common Subsequence Problem”. In: *Journal of the ACM* 23.1 (Jan. 1, 1976), 1–12. DOI: [10.1145/321921.321922](https://doi.org/10.1145/321921.321922). URL: <https://doi.org/10.1145/321921.321922> (cit. on p. 42).
- [145] D. S. Hirschberg. “A linear space algorithm for computing maximal common subsequences”. In: *Communications of the ACM* 18.6 (June 1, 1975), 341–343. DOI: [10.1145/360825.360861](https://doi.org/10.1145/360825.360861). URL: <https://doi.org/10.1145/360825.360861> (cit. on p. 43).
- [146] Eugene W. Myers and Webb Miller. “Optimal alignments in linear space”. In: *Bioinformatics* 4.1 (Mar. 1, 1988), pp. 11–17. DOI: [10.1093/bioinformatics/4.1.11](https://doi.org/10.1093/bioinformatics/4.1.11). URL: <https://doi.org/10.1093/bioinformatics/4.1.11> (cit. on p. 43).
- [147] Peter Rice, Ian Longden, and Alan Bleasby. “EMBOSS: the European molecular biology open software suite”. In: *Trends in genetics* 16.6 (2000). tex.ids=riceEMBOSSEuropeanMolecular publisher: Elsevier current trends, 276–277 (cit. on pp. 43, 44).
- [148] Xiaoqiu Huang and Webb Miller. “A time-efficient, linear-space local similarity algorithm”. In: *Advances in Applied Mathematics* 12.3 (Sept. 1, 1991), pp. 337–357. DOI: [10.1016/0196-8858\(91\)90017-D](https://doi.org/10.1016/0196-8858(91)90017-D). URL: <https://www.sciencedirect.com/science/article/pii/019688589190017D> (cit. on p. 44).
- [149] Michael S. Waterman and Mark Eggert. “A new algorithm for best subsequence alignments with application to tRNA-rRNA comparisons”. In: *Journal of Molecular Biology* 197.4 (Oct. 20, 1987), pp. 723–728. DOI: [10.1016/0022-2836\(87\)90478-5](https://doi.org/10.1016/0022-2836(87)90478-5). URL: <https://www.sciencedirect.com/science/article/pii/0022283687904785> (cit. on p. 44).
- [150] Jason E. Stajich, David Block, et al. “The Bioperl Toolkit: Perl Modules for the Life Sciences”. In: *Genome Research* 12.10 (Jan. 10, 2002). Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab PMID: 12368254, pp. 1611–1618. DOI: [10.1101/gr.361602](https://doi.org/10.1101/gr.361602). URL: <https://genome.cshlp.org/content/12/10/1611> (cit. on p. 44).

- [151] Robert C. Gentleman, Vincent J. Carey, et al. “Bioconductor: open software development for computational biology and bioinformatics”. In: *Genome Biology* 5.10 (Sept. 15, 2004), R80. DOI: [10.1186/gb-2004-5-10-r80](https://doi.org/10.1186/gb-2004-5-10-r80). URL: <https://doi.org/10.1186/gb-2004-5-10-r80> (cit. on p. 44).
- [152] Jeff Daily. “Parasail: SIMD C library for global, semi-global, and local pairwise sequence alignments”. In: *BMC Bioinformatics* 17.1 (Feb. 10, 2016), p. 81. DOI: [10.1186/s12859-016-0930-z](https://doi.org/10.1186/s12859-016-0930-z). URL: <https://doi.org/10.1186/s12859-016-0930-z> (cit. on p. 44).
- [153] W. Frohmberg, M. Kierzynka, et al. “G-PAS 2.0 – an improved version of protein alignment tool with an efficient backtracking routine on multiple GPUs”. In: *Bulletin of the Polish Academy of Sciences: Technical Sciences* 60.3 (Dec. 1, 2012), pp. 491–494. DOI: [10.2478/v10175-012-0062-1](https://doi.org/10.2478/v10175-012-0062-1). URL: <http://journals.pan.pl/dlibra/publication/96876/edition/83624/content> (cit. on p. 44).
- [154] Stephen F Altschul. “Substitution Matrices”. In: John Wiley & Sons, Ltd, 2013. DOI: [10.1002/9780470015902.a0005265.pub3](https://doi.org/10.1002/9780470015902.a0005265.pub3). URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470015902.a0005265.pub3> (cit. on p. 44).
- [155] M.O. Dayhoff, R.M. Schwartz, and B.C. Orcutt. “A Model of Evolutionary Change in Proteins”. In: *A Model of Evolutionary Change in Proteins* (1978), pp. 345–352 (cit. on p. 44).
- [156] T. Müller and M. Vingron. “Modeling amino acid replacement”. In: *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology* 7.6 (2000). PMID: 11382360, pp. 761–776. DOI: [10.1089/10665270050514918](https://doi.org/10.1089/10665270050514918) (cit. on p. 44).
- [157] S. Henikoff and J. G. Henikoff. “Amino acid substitution matrices from protein blocks”. In: *Proceedings of the National Academy of Sciences* 89.22 (Nov. 15, 1992). PMID: 1438297, pp. 10915–10919. DOI: [10.1073/pnas.89.22.10915](https://doi.org/10.1073/pnas.89.22.10915). URL: <https://www.pnas.org/content/89/22/10915> (cit. on p. 44).
- [158] S. Whelan and N. Goldman. “A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach”. In: *Molecular Biology and Evolution* 18.5 (May 2001). PMID: 11319253, pp. 691–699. DOI: [10.1093/oxfordjournals.molbev.a003851](https://doi.org/10.1093/oxfordjournals.molbev.a003851) (cit. on p. 44).
- [159] Si Quang Le and Olivier Gascuel. “An Improved General Amino Acid Replacement Matrix”. In: *Molecular Biology and Evolution* 25.7 (July 1, 2008), pp. 1307–1320. DOI: [10.1093/molbev/msn067](https://doi.org/10.1093/molbev/msn067). URL: <https://doi.org/10.1093/molbev/msn067> (cit. on p. 44).
- [160] Tobias Müller, Sven Rahmann, and Marc Rehmsmeier. “Non-symmetric score matrices and the detection of homologous transmembrane proteins”. In: *Bioinformatics* 17.suppl\_1 (June 1, 2001), S182–S189. DOI: [10.1093/bioinformatics/17.suppl\\_1.S182](https://doi.org/10.1093/bioinformatics/17.suppl_1.S182). URL: [https://doi.org/10.1093/bioinformatics/17.suppl\\_1.S182](https://doi.org/10.1093/bioinformatics/17.suppl_1.S182) (cit. on p. 44).

## REFERENCES

- [161] P. C. Ng, J. G. Henikoff, and S. Henikoff. “PHAT: a transmembrane-specific substitution matrix. Predicted hydrophobic and transmembrane”. In: *Bioinformatics (Oxford, England)* 16.9 (Sept. 2000). PMID: 11108698, pp. 760–766. DOI: [10.1093/bioinformatics/16.9.760](https://doi.org/10.1093/bioinformatics/16.9.760) (cit. on p. 44).
- [162] Rakesh Trivedi and Hampapathalu Adimurthy Nagarajaram. “Amino acid substitution scoring matrices specific to intrinsically disordered regions in proteins”. In: *Scientific Reports* 9.1 (Nov. 8, 2019). Number: 1 Publisher: Nature Publishing Group, p. 16380. DOI: [10.1038/s41598-019-52532-8](https://doi.org/10.1038/s41598-019-52532-8). URL: <https://www.nature.com/articles/s41598-019-52532-8> (cit. on p. 44).
- [163] Nalin C. W. Goonesekere and Byungkook Lee. “Context-specific amino acid substitution matrices and their use in the detection of protein homologs”. In: *Proteins: Structure, Function, and Bioinformatics* 71.2 (2008), pp. 910–919. DOI: [10.1002/prot.21775](https://doi.org/10.1002/prot.21775). URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.21775> (cit. on p. 44).
- [164] Umadevi Paila, Rohini Kondam, and Akash Ranjan. “Genome bias influences amino acid choices: analysis of amino acid substitution and re-compilation of substitution matrices exclusive to an AT-biased genome”. In: *Nucleic Acids Research* 36.21 (Dec. 2008). PMID: 18948281 PMCID: PMC2588515, pp. 6664–6675. DOI: [10.1093/nar/gkn635](https://doi.org/10.1093/nar/gkn635) (cit. on p. 44).
- [165] David C. Nickle, Laura Heath, et al. “HIV-Specific Probabilistic Models of Protein Evolution”. In: *PLoS ONE* 2.6 (June 6, 2007). PMID: 17551583 PMCID: PMC1876811, e503. DOI: [10.1371/journal.pone.0000503](https://doi.org/10.1371/journal.pone.0000503). URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1876811/> (cit. on p. 44).
- [166] Mihaela E. Sardiu, Gelio Alves, and Yi-Kuo Yu. “Score statistics of global sequence alignment from the energy distribution of a modified directed polymer and directed percolation problem”. In: *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics* 72.6 Pt 1 (Dec. 2005). PMID: 16485984, p. 061917. DOI: [10.1103/PhysRevE.72.061917](https://doi.org/10.1103/PhysRevE.72.061917) (cit. on p. 44).
- [167] F. Chiaromonte, V. B. Yap, and W. Miller. “Scoring pairwise genomic sequence alignments”. In: DOI: [10.1142/9789812799623\\_0012](https://doi.org/10.1142/9789812799623_0012). WORLD SCIENTIFIC, Dec. 2001, pp. 115–126. DOI: [10.1142/9789812799623\\_0012](https://doi.org/10.1142/9789812799623_0012). URL: [https://www.worldscientific.com/doi/abs/10.1142/9789812799623\\_0012](https://www.worldscientific.com/doi/abs/10.1142/9789812799623_0012) (cit. on p. 44).
- [168] Adrian Schneider, Gina M. Cannarozzi, and Gaston H. Gonnet. “Empirical codon substitution matrix”. In: *BMC bioinformatics* 6 (June 1, 2005). PMID: 15927081 PMCID: PMC1173088, p. 134. DOI: [10.1186/1471-2105-6-134](https://doi.org/10.1186/1471-2105-6-134) (cit. on p. 44).
- [169] Adi Doron-Faigenboim and Tal Pupko. “A Combined Empirical and Mechanistic Codon Model”. In: *Molecular Biology and Evolution* 24.2 (Feb. 1, 2007), pp. 388–397. DOI: [10.1093/molbev/msl175](https://doi.org/10.1093/molbev/msl175). URL: <https://doi.org/10.1093/molbev/msl175> (cit. on p. 44).

- [170] Osamu Gotoh. “An improved algorithm for matching biological sequences”. In: *Journal of Molecular Biology* 162.3 (Dec. 15, 1982), pp. 705–708. DOI: [10.1016/0022-2836\(82\)90398-9](https://doi.org/10.1016/0022-2836(82)90398-9). URL: <https://www.sciencedirect.com/science/article/pii/0022283682903989> (cit. on p. 45).
- [171] Steven A. Benner, Mark A. Cohen, and Gaston H. Gonnet. “Empirical and Structural Models for Insertions and Deletions in the Divergent Evolution of Proteins”. In: *Journal of Molecular Biology* 229.4 (Feb. 20, 1993), pp. 1065–1082. DOI: [10.1006/jmbi.1993.1105](https://doi.org/10.1006/jmbi.1993.1105). URL: <https://www.sciencedirect.com/science/article/pii/S0022283683711058> (cit. on p. 45).
- [172] Xun Gu and Wen-Hsiung Li. “The size distribution of insertions and deletions in human and rodent pseudogenes suggests the logarithmic gap penalty for sequence alignment”. In: *Journal of Molecular Evolution* 40.4 (Apr. 1, 1995), pp. 464–473. DOI: [10.1007/BF00164032](https://doi.org/10.1007/BF00164032). URL: <https://doi.org/10.1007/BF00164032> (cit. on p. 45).
- [173] Michael S. Waterman. “Efficient sequence alignment algorithms”. In: *Journal of Theoretical Biology* 108.3 (June 7, 1984), pp. 333–337. DOI: [10.1016/S0022-5193\(84\)80037-5](https://doi.org/10.1016/S0022-5193(84)80037-5). URL: <https://www.sciencedirect.com/science/article/pii/S0022519384800375> (cit. on p. 45).
- [174] William R. Pearson and Webb Miller. “[27] Dynamic programming algorithms for biological sequence comparison”. In: vol. 210. Numerical Computer Methods. DOI: [10.1016/0076-6879\(92\)10029-D](https://doi.org/10.1016/0076-6879(92)10029-D). Academic Press, Jan. 1, 1992, pp. 575–601. DOI: [10.1016/0076-6879\(92\)10029-D](https://doi.org/10.1016/0076-6879(92)10029-D). URL: <https://www.sciencedirect.com/science/article/pii/007668799210029D> (cit. on p. 45).
- [175] Jiannan Chao, Furong Tang, and Lei Xu. “Developments in Algorithms for Sequence Alignment: A Review”. In: *Biomolecules* 12.4 (Apr. 6, 2022), p. 546. DOI: [10.3390/biom12040546](https://doi.org/10.3390/biom12040546). URL: <https://www.mdpi.com/2218-273X/12/4/546> (cit. on pp. 45, 46).
- [176] John L. Spouge. “Speeding up Dynamic Programming Algorithms for Finding Optimal Lattice Paths”. In: *SIAM Journal on Applied Mathematics* 49.5 (Oct. 1989). tex.ids= spougeSpeedingDynamicProgramming1989 publisher: Society for Industrial and Applied Mathematics, pp. 1552–1566. DOI: [10.1137/0149094](https://doi.org/10.1137/0149094). URL: <https://pubs.siam.org/doi/abs/10.1137/0149094> (cit. on p. 45).
- [177] James W. Fickett. “Fast optimal alignment”. In: *Nucleic Acids Research* 12.1Part1 (Jan. 11, 1984), pp. 175–179. DOI: [10.1093/nar/12.1Part1.175](https://doi.org/10.1093/nar/12.1Part1.175). URL: <https://doi.org/10.1093/nar/12.1Part1.175> (cit. on p. 45).
- [178] Santiago Marco-Sola, Juan Carlos Moure, et al. “Fast gap-affine pairwise alignment using the wavefront algorithm”. In: *Bioinformatics* btaa777 (Sept. 11, 2020). DOI: [10.1093/bioinformatics/btaa777](https://doi.org/10.1093/bioinformatics/btaa777). URL: <https://doi.org/10.1093/bioinformatics/btaa777> (cit. on p. 45).

## REFERENCES

- [179] Richard Durbin, Sean R. Eddy, et al. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. DOI: 10.1017/CBO9780511790492. Cambridge: Cambridge University Press, 1998. doi: 10.1017/CBO9780511790492. URL: <https://www.cambridge.org/core/books/biological-sequence-analysis/921BB7B78B745198829EF96BC7E0F29D> (cit. on p. 45).
- [180] Johannes Söding. “Protein homology detection by HMM-HMM comparison”. In: *Bioinformatics (Oxford, England)* 21.7 (Apr. 1, 2005). PMID: 15531603, pp. 951–960. doi: 10.1093/bioinformatics/bti125 (cit. on p. 45).
- [181] Robert D. Finn, Jody Clements, and Sean R. Eddy. “HMMER web server: interactive sequence similarity searching”. In: *Nucleic Acids Research* 39.Web Server issue (July 1, 2011). PMID: 21593126 PMCID: PMC3125773, W29–W37. doi: 10.1093/nar/gkr367. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3125773/> (cit. on pp. 45, 50).
- [182] Jun Wang, Peter D. Keightley, and Toby Johnson. “MCALIGN2: Faster, accurate global pairwise alignment of non-coding DNA sequences based on explicit models of indel evolution”. In: *BMC Bioinformatics* 7.1 (June 8, 2006), p. 292. doi: 10.1186/1471-2105-7-292. URL: <https://doi.org/10.1186/1471-2105-7-292> (cit. on p. 45).
- [183] Kazutaka Katoh, Kazuharu Misawa, et al. “MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform”. In: *Nucleic Acids Research* 30.14 (July 15, 2002), pp. 3059–3066. doi: 10.1093/nar/gkf436. URL: <https://doi.org/10.1093/nar/gkf436> (cit. on pp. 45, 50).
- [184] Yanni Sun and Jeremy Buhler. “Choosing the best heuristic for seeded alignment of DNA sequences”. In: *BMC Bioinformatics* 7.1 (Mar. 13, 2006), p. 133. doi: 10.1186/1471-2105-7-133. URL: <https://doi.org/10.1186/1471-2105-7-133> (cit. on p. 46).
- [185] Heng Li and Nils Homer. “A Survey of Sequence Alignment Algorithms for Next-Generation Sequencing”. In: *Briefings in Bioinformatics* 11.5 (Sept. 1, 2010), pp. 473–483. ISSN: 1467-5463. doi: 10.1093/bib/bbp015. URL: <https://doi.org/10.1093/bib/bbp015> (visited on 05/16/2022) (cit. on p. 46).
- [186] S. F. Altschul, W. Gish, et al. “Basic local alignment search tool”. In: *Journal of Molecular Biology* 215.3 (Oct. 5, 1990). PMID: 2231712, pp. 403–410. doi: 10.1016/S0022-2836(05)80360-2 (cit. on p. 47).
- [187] Robert C. Edgar. “Search and clustering orders of magnitude faster than BLAST”. In: *Bioinformatics* 26.19 (Oct. 1, 2010), pp. 2460–2461. doi: 10.1093/bioinformatics/btq461. URL: <https://doi.org/10.1093/bioinformatics/btq461> (cit. on p. 47).

- [188] Eugene G. Shpaer, Max Robinson, et al. “Sensitivity and Selectivity in Protein Similarity Searches: A Comparison of Smith–Waterman in Hardware to BLAST and FASTA”. In: *Genomics* 38.2 (Dec. 1, 1996). tex.ids= shpaerSensitivitySelectivityProtein1996a, pp. 179–191. doi: [10.1006/geno.1996.0614](https://doi.org/10.1006/geno.1996.0614). URL: <https://www.sciencedirect.com/science/article/pii/S088875439690614X> (cit. on p. 47).
- [189] W. R. Pearson and D. J. Lipman. “Improved tools for biological sequence comparison”. In: *Proceedings of the National Academy of Sciences of the United States of America* 85.8 (Apr. 1988). PMID: 3162770 PMCID: PMC280013, pp. 2444–2448. doi: [10.1073/pnas.85.8.2444](https://doi.org/10.1073/pnas.85.8.2444) (cit. on p. 47).
- [190] D. J. Lipman and W. R. Pearson. “Rapid and sensitive protein similarity searches”. In: *Science (New York, N. Y.)* 227.4693 (Mar. 22, 1985). PMID: 2983426, pp. 1435–1441. doi: [10.1126/science.2983426](https://doi.org/10.1126/science.2983426) (cit. on p. 47).
- [191] Ganapathi Varma Saripella, Erik L. L. Sonnhammer, and Kristoffer Forslund. “Benchmarking the next generation of homology inference tools”. In: *Bioinformatics* 32.17 (Sept. 9, 2016). Publisher: Oxford University Press PMID: 27256311, p. 2636. doi: [10.1093/bioinformatics/btw305](https://doi.org/10.1093/bioinformatics/btw305). URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5013910/> (cit. on p. 48).
- [192] Robert D. Finn, Penelope Coggill, et al. “The Pfam protein families database: towards a more sustainable future”. In: *Nucleic Acids Research* 44.Database issue (Jan. 1, 2016). Publisher: Oxford University Press PMID: 26673716, p. D279. doi: [10.1093/nar/gkv1344](https://doi.org/10.1093/nar/gkv1344). URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4702930/> (cit. on p. 48).
- [193] Nadia Essoussi and Sondes Fayech. “A comparison of four pair-wise sequence alignment methods”. In: *Bioinformation* 2.4 (Dec. 28, 2007). PMID: 21670797 PMCID: PMC2255065, pp. 166–168. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2255065/> (cit. on p. 48).
- [194] S Karlin and S F Altschul. “Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes.” In: *Proceedings of the National Academy of Sciences* 87.6 (Mar. 1990). tex.ids= karlinMethodsAssessingStatistical1990 PMCID: PMC53667 PMID: 2315319, pp. 2264–2268. doi: [10.1073/pnas.87.6.2264](https://doi.org/10.1073/pnas.87.6.2264). URL: <https://pnas.org/doi/full/10.1073/pnas.87.6.2264> (cit. on p. 48).
- [195] Richard Mott. “Alignment: Statistical Significance”. In: \_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1038/npg.els.0005264>. John Wiley & Sons, Ltd, 2005. doi: [10.1038/npg.els.0005264](https://doi.org/10.1038/npg.els.0005264). URL: <https://onlinelibrary.wiley.com/doi/abs/10.1038/npg.els.0005264> (cit. on p. 48).
- [196] Benjamin Buchfink, Chao Xie, and Daniel H. Huson. “Fast and sensitive protein alignment using DIAMOND”. In: *Nature Methods* 12.1 (Jan. 2015). PMID: 25402007, pp. 59–60. doi: [10.1038/nmeth.3176](https://doi.org/10.1038/nmeth.3176) (cit. on p. 48).

## REFERENCES

- [197] Benjamin Buchfink, Klaus Reuter, and Hajk-Georg Drost. “Sensitive protein alignments at tree-of-life scale using DIAMOND”. In: *Nature Methods* 18.4 (Apr. 2021). Number: 4 Publisher: Nature Publishing Group, pp. 366–368. DOI: [10.1038/s41592-021-01101-x](https://doi.org/10.1038/s41592-021-01101-x). URL: <https://www.nature.com/articles/s41592-021-01101-x> (cit. on p. 48).
- [198] Nick Bray, Inna Dubchak, and Lior Pachter. “AVID: A Global Alignment Program”. In: *Genome Research* 13.1 (Jan. 1, 2003). PMID: 12529311 PMCID: PMC430967, pp. 97–102. DOI: [10.1101/gr.789803](https://doi.org/10.1101/gr.789803). URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC430967/> (cit. on p. 48).
- [199] Arthur L. Delcher, Adam Phillippy, et al. “Fast algorithms for large-scale genome alignment and comparison”. In: *Nucleic Acids Research* 30.11 (June 1, 2002). PMID: 12034836 PMCID: PMC117189, pp. 2478–2483. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC117189/> (cit. on p. 48).
- [200] P. Ferragina and G. Manzini. “Proceedings 41st Annual Symposium on Foundations of Computer Science”. In: tex.ids= ferraginaOpportunisticDataStructures2000a ISSN: 0272-5428. Nov. 2000, pp. 390–398. DOI: [10.1109/SFCS.2000.892127](https://doi.org/10.1109/SFCS.2000.892127) (cit. on p. 48).
- [201] Michael Burrows and David Wheeler. *A Block-Sorting Lossless Data Compression Algorithm*. Tech. rep. 1994 (cit. on p. 48).
- [202] T. W. Lam, W. K. Sung, et al. “Compressed indexing and local alignment of DNA”. In: *Bioinformatics* 24.6 (Mar. 15, 2008), pp. 791–797. DOI: [10.1093/bioinformatics/btn032](https://doi.org/10.1093/bioinformatics/btn032). URL: <https://doi.org/10.1093/bioinformatics/btn032> (cit. on pp. 49, 51).
- [203] Ben Langmead and Steven L. Salzberg. “Fast Gapped-Read Alignment with Bowtie 2”. In: *Nature Methods* 9.4 (4 Apr. 2012), pp. 357–359. ISSN: 1548-7105. DOI: [10.1038/nmeth.1923](https://doi.org/10.1038/nmeth.1923) (cit. on pp. 49, 51).
- [204] Heng Li and Richard Durbin. “Fast and accurate short read alignment with Burrows–Wheeler transform”. In: *Bioinformatics* 25.14 (July 15, 2009), pp. 1754–1760. DOI: [10.1093/bioinformatics/btp324](https://doi.org/10.1093/bioinformatics/btp324). URL: <https://doi.org/10.1093/bioinformatics/btp324> (cit. on pp. 49, 51).
- [205] Heng Li and Richard Durbin. “Fast and accurate long-read alignment with Burrows–Wheeler transform”. In: *Bioinformatics* 26.5 (Mar. 1, 2010), pp. 589–595. DOI: [10.1093/bioinformatics/btp698](https://doi.org/10.1093/bioinformatics/btp698). URL: <https://doi.org/10.1093/bioinformatics/btp698> (cit. on pp. 49, 51).
- [206] Heng Li. *Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM*. May 26, 2013. arXiv: [1303.3997 \[q-bio\]](https://arxiv.org/abs/1303.3997) (cit. on pp. 49, 51).
- [207] Yongchao Liu and Bertil Schmidt. “Long read alignment based on maximal exact match seeds”. In: *Bioinformatics* 28.18 (Sept. 15, 2012), pp. i318–i324. DOI: [10.1093/bioinformatics/bts414](https://doi.org/10.1093/bioinformatics/bts414). URL: <https://doi.org/10.1093/bioinformatics/bts414> (cit. on pp. 49, 51).

- [208] David J Russell, ed. *Multiple Sequence Alignment Methods*. Vol. 1079. Methods in Molecular Biology. DOI: 10.1007/978-1-62703-646-7. Totowa, NJ: Humana Press, 2014. doi: [10.1007/978-1-62703-646-7](https://doi.org/10.1007/978-1-62703-646-7). URL: <http://link.springer.com/10.1007/978-1-62703-646-7> (cit. on p. 49).
- [209] Lusheng Wang and Tao Jiang. “On the Complexity of Multiple Sequence Alignment”. In: *Journal of Computational Biology* 1.4 (Jan. 1994). Publisher: Mary Ann Liebert, Inc., publishers, pp. 337–348. doi: [10.1089/cmb.1994.1.337](https://doi.org/10.1089/cmb.1994.1.337). URL: <https://www.liebertpub.com/doi/abs/10.1089/cmb.1994.1.337> (cit. on p. 49).
- [210] Winfried Just. “Computational Complexity of Multiple Sequence Alignment with SP-Score”. In: *Journal of Computational Biology* 8.6 (Nov. 2001). Publisher: Mary Ann Liebert, Inc., publishers, pp. 615–623. doi: [10.1089/106652701753307511](https://doi.org/10.1089/106652701753307511). URL: <https://www.liebertpub.com/doi/abs/10.1089/106652701753307511> (cit. on p. 49).
- [211] Da-Fei Feng and Russell F. Doolittle. “Progressive sequence alignment as a pre-requisite to correct phylogenetic trees”. In: *Journal of Molecular Evolution* 25.4 (Aug. 1, 1987), pp. 351–360. doi: [10.1007/BF02603120](https://doi.org/10.1007/BF02603120). URL: <https://doi.org/10.1007/BF02603120> (cit. on p. 49).
- [212] David T. Jones, William R. Taylor, and Janet M. Thornton. “The rapid generation of mutation data matrices from protein sequences”. In: *Bioinformatics* 8.3 (June 1, 1992), pp. 275–282. doi: [10.1093/bioinformatics/8.3.275](https://doi.org/10.1093/bioinformatics/8.3.275). URL: <https://doi.org/10.1093/bioinformatics/8.3.275> (cit. on p. 49).
- [213] B E Blaisdell. “A measure of the similarity of sets of sequences not requiring sequence alignment.” In: *Proceedings of the National Academy of Sciences* 83.14 (July 1986). Publisher: Proceedings of the National Academy of Sciences, pp. 5155–5159. doi: [10.1073/pnas.83.14.5155](https://doi.org/10.1073/pnas.83.14.5155). URL: <https://www.pnas.org/doi/abs/10.1073/pnas.83.14.5155> (cit. on p. 49).
- [214] Stephen F. Altschul, Raymond J. Carroll, and David J. Lipman. “Weights for data related by a tree”. In: *Journal of Molecular Biology* 207.4 (June 20, 1989), pp. 647–653. doi: [10.1016/0022-2836\(89\)90234-9](https://doi.org/10.1016/0022-2836(89)90234-9). URL: <https://www.sciencedirect.com/science/article/pii/0022283689902349> (cit. on p. 49).
- [215] Robert C. Edgar and Kimmen Sjölander. “A comparison of scoring functions for protein sequence profile alignment”. In: *Bioinformatics* 20.8 (May 22, 2004), pp. 1301–1308. doi: [10.1093/bioinformatics/bth090](https://doi.org/10.1093/bioinformatics/bth090). URL: <https://doi.org/10.1093/bioinformatics/bth090> (cit. on p. 49).
- [216] C Notredame, L Holm, and D G Higgins. “COFFEE: an objective function for multiple sequence alignments.” In: *Bioinformatics* 14.5 (June 1, 1998), pp. 407–422. doi: [10.1093/bioinformatics/14.5.407](https://doi.org/10.1093/bioinformatics/14.5.407). URL: <https://doi.org/10.1093/bioinformatics/14.5.407> (cit. on p. 49).

## REFERENCES

- [217] Julie D. Thompson, Desmond G. Higgins, and Toby J. Gibson. “CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice”. In: *Nucleic Acids Research* 22.22 (Nov. 11, 1994), pp. 4673–4680. DOI: [10.1093/nar/22.22.4673](https://doi.org/10.1093/nar/22.22.4673). URL: <https://academic.oup.com/nar/article/22/22/4673/2400290> (cit. on p. 49).
- [218] Julie D. Thompson, Toby J. Gibson, et al. “The CLUSTAL\_X Windows Interface: Flexible Strategies for Multiple Sequence Alignment Aided by Quality Analysis Tools”. In: *Nucleic Acids Research* 25.24 (Dec. 1, 1997), pp. 4876–4882. DOI: [10.1093/nar/25.24.4876](https://doi.org/10.1093/nar/25.24.4876). URL: <https://doi.org/10.1093/nar/25.24.4876> (cit. on p. 49).
- [219] Cédric Notredame, Desmond G Higgins, and Jaap Heringa. “T-coffee: a novel method for fast and accurate multiple sequence alignment” Edited by J. Thornton”. In: *Journal of Molecular Biology* 302.1 (Sept. 8, 2000), pp. 205–217. DOI: [10.1006/jmbi.2000.4042](https://doi.org/10.1006/jmbi.2000.4042). URL: <http://www.sciencedirect.com/science/article/pii/S0022283600940427> (cit. on p. 50).
- [220] Robert C. Edgar. “MUSCLE: a multiple sequence alignment method with reduced time and space complexity”. In: *BMC Bioinformatics* 5.1 (Aug. 19, 2004), p. 113. DOI: [10.1186/1471-2105-5-113](https://doi.org/10.1186/1471-2105-5-113). URL: <https://doi.org/10.1186/1471-2105-5-113> (cit. on p. 50).
- [221] Robert C. Edgar. “MUSCLE: multiple sequence alignment with high accuracy and high throughput”. In: *Nucleic Acids Research* 32.5 (Mar. 1, 2004), pp. 1792–1797. DOI: [10.1093/nar/gkh340](https://doi.org/10.1093/nar/gkh340). URL: <https://doi.org/10.1093/nar/gkh340> (cit. on p. 50).
- [222] Chuong B. Do, Mahathi S.P. Mahabhashyam, et al. “ProbCons: Probabilistic consistency-based multiple sequence alignment”. In: *Genome Research* 15.2 (Feb. 2005). PMID: 15687296 PMCID: PMC546535, pp. 330–340. DOI: [10.1101/gr.2821705](https://doi.org/10.1101/gr.2821705). URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC546535/> (cit. on p. 50).
- [223] Cédric Notredame. “Recent Evolutions of Multiple Sequence Alignment Algorithms”. In: *PLOS Computational Biology* 3.8 (Aug. 31, 2007). Publisher: Public Library of Science, e123. DOI: [10.1371/journal.pcbi.0030123](https://doi.org/10.1371/journal.pcbi.0030123). URL: <https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.0030123> (cit. on p. 50).
- [224] Cédric Notredame. “Recent Progress in Multiple Sequence Alignment: A Survey”. In: *Pharmacogenomics* 3.1 (Jan. 2002), pp. 131–144. ISSN: 1462-2416. DOI: [10.1517/14622416.3.1.131](https://doi.org/10.1517/14622416.3.1.131). URL: <https://www.futuremedicine.com/doi/abs/10.1517/14622416.3.1.131> (visited on 05/16/2022) (cit. on p. 50).

- [225] Robert C Edgar and Serafim Batzoglou. “Multiple Sequence Alignment”. In: *Current Opinion in Structural Biology*. Nucleic Acids/Sequences and Topology 16.3 (June 1, 2006), pp. 368–373. ISSN: 0959-440X. DOI: [10.1016/j.sbi.2006.04.004](https://doi.org/10.1016/j.sbi.2006.04.004). URL: <https://www.sciencedirect.com/science/article/pii/S0959440X06000704> (visited on 05/16/2022) (cit. on p. 50).
- [226] Fabiano Sviatopolk-Mirsky Pais, Patrícia de Cássia Ruy, et al. “Assessing the efficiency of multiple sequence alignment programs”. In: *Algorithms for Molecular Biology* 9.1 (Mar. 6, 2014), p. 4. DOI: [10.1186/1748-7188-9-4](https://doi.org/10.1186/1748-7188-9-4). URL: <https://doi.org/10.1186/1748-7188-9-4> (cit. on p. 50).
- [227] J. D. Thompson, F. Plewniak, and O. Poch. “BAliBASE: a benchmark alignment database for the evaluation of multiple alignment programs.” In: *Bioinformatics* 15.1 (Jan. 1, 1999), pp. 87–88. DOI: [10.1093/bioinformatics/15.1.87](https://doi.org/10.1093/bioinformatics/15.1.87). URL: <https://academic.oup.com/bioinformatics/article/15/1/87/218377> (cit. on p. 50).
- [228] Sean R Eddy. “Multiple Alignment Using Hidden Markov Models”. In: (), p. 7 (cit. on p. 50).
- [229] Yongchao Liu, Bertil Schmidt, and Douglas L. Maskell. “MSAProbs: multiple sequence alignment based on pair hidden Markov models and partition function posterior probabilities”. In: *Bioinformatics* 26.16 (Aug. 15, 2010), pp. 1958–1964. DOI: [10.1093/bioinformatics/btq338](https://doi.org/10.1093/bioinformatics/btq338). URL: <https://doi.org/10.1093/bioinformatics/btq338> (cit. on p. 50).
- [230] Frédéric Lemoine, Luc Bassel, et al. “COVID-Align: Accurate online alignment of hCoV-19 genomes using a profile HMM”. In: *Bioinformatics* btaa871 (Oct. 12, 2020). tex.ids= lemoineCOVIDAlignAccurateOnline2020. DOI: [10.1093/bioinformatics/btaa871](https://doi.org/10.1093/bioinformatics/btaa871). URL: <https://doi.org/10.1093/bioinformatics/btaa871> (cit. on p. 50).
- [231] Ivan Aksamentov, Cornelius Roemer, et al. “Nextclade: clade assignment, mutation calling and quality control for viral genomes”. In: *Journal of Open Source Software* 6.67 (Nov. 30, 2021), p. 3773. DOI: [10.21105/joss.03773](https://doi.org/10.21105/joss.03773). URL: <https://joss.theoj.org/papers/10.21105/joss.03773> (cit. on p. 50).
- [232] Jin Kim, Sakti Pramanik, and Moon Jung Chung. “Multiple sequence alignment using simulated annealing”. In: *Bioinformatics* 10.4 (July 1, 1994), pp. 419–426. DOI: [10.1093/bioinformatics/10.4.419](https://doi.org/10.1093/bioinformatics/10.4.419). URL: <https://doi.org/10.1093/bioinformatics/10.4.419> (cit. on p. 51).
- [233] Masato Ishikawa, Tomoyuki Toya, et al. “Multiple sequence alignment by parallel simulated annealing”. In: *Bioinformatics* 9.3 (June 1, 1993), pp. 267–273. DOI: [10.1093/bioinformatics/9.3.267](https://doi.org/10.1093/bioinformatics/9.3.267). URL: <https://doi.org/10.1093/bioinformatics/9.3.267> (cit. on p. 51).
- [234] Hongwei Huo and Vojislav Stojkovic. “Third International Conference on Natural Computation (ICNC 2007)”. In: vol. 2. ISSN: 2157-9563. Aug. 2007, pp. 270–274. DOI: [10.1109/ICNC.2007.139](https://doi.org/10.1109/ICNC.2007.139) (cit. on p. 51).

## REFERENCES

- [235] Biswanath Chowdhury and Gautam Garai. “A review on multiple sequence alignment from the perspective of genetic algorithm”. In: *Genomics* 109.5 (Oct. 1, 2017), pp. 419–431. DOI: [10.1016/j.ygeno.2017.06.007](https://doi.org/10.1016/j.ygeno.2017.06.007). URL: <http://www.sciencedirect.com/science/article/pii/S0888754317300551> (cit. on p. 51).
- [236] Ching Zhang and Andrew K.C. Wong. “A genetic algorithm for multiple molecular sequence alignment”. In: *Bioinformatics* 13.6 (Dec. 1, 1997), pp. 565–581. DOI: [10.1093/bioinformatics/13.6.565](https://doi.org/10.1093/bioinformatics/13.6.565). URL: <https://doi.org/10.1093/bioinformatics/13.6.565> (cit. on p. 51).
- [237] C Notredame and D G Higgins. “SAGA: sequence alignment by genetic algorithm.” In: *Nucleic Acids Research* 24.8 (Apr. 15, 1996). PMID: 8628686 PMCID: PMC145823, pp. 1515–1524. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC145823/> (cit. on p. 51).
- [238] Farhana Naznin, Ruhul Sarker, and Daryl Essam. “Vertical decomposition with Genetic Algorithm for Multiple Sequence Alignment”. In: *BMC Bioinformatics* 12.1 (Aug. 25, 2011), p. 353. DOI: [10.1186/1471-2105-12-353](https://doi.org/10.1186/1471-2105-12-353). URL: <https://doi.org/10.1186/1471-2105-12-353> (cit. on p. 51).
- [239] Farhana Naznin, Ruhul Sarker, and Daryl Essam. “Progressive Alignment Method Using Genetic Algorithm for Multiple Sequence Alignment”. In: *IEEE Transactions on Evolutionary Computation* 16.5 (Oct. 2012). Conference Name: IEEE Transactions on Evolutionary Computation, pp. 615–631. DOI: [10.1109/TEVC.2011.2162849](https://doi.org/10.1109/TEVC.2011.2162849) (cit. on p. 51).
- [240] Edgar Garriga, Paolo Di Tommaso, et al. “Large multiple sequence alignments with a root-to-leaf regressive method”. In: *Nature Biotechnology* 37.12 (Dec. 2019). Number: 12 Publisher: Nature Publishing Group, pp. 1466–1470. DOI: [10.1038/s41587-019-0333-6](https://doi.org/10.1038/s41587-019-0333-6). URL: <https://www.nature.com/articles/s41587-019-0333-6> (cit. on p. 51).
- [241] Mohammed Alser, Jeremy Rotman, et al. “Technology dictates algorithms: recent developments in read alignment”. In: *Genome Biology* 22.1 (Aug. 26, 2021), p. 249. DOI: [10.1186/s13059-021-02443-7](https://doi.org/10.1186/s13059-021-02443-7). URL: <https://doi.org/10.1186/s13059-021-02443-7> (cit. on pp. 51, 52).
- [242] Sophie Schbath, Véronique Martin, et al. “Mapping Reads on a Genomic Sequence: An Algorithmic Overview and a Practical Comparative Analysis”. In: *Journal of Computational Biology* 19.6 (June 2012). PMID: 22506536 PMCID: PMC3375638, pp. 796–813. DOI: [10.1089/cmb.2012.0022](https://doi.org/10.1089/cmb.2012.0022). URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3375638/> (cit. on p. 51).
- [243] Ayat Hatem, Doruk Bozdağ, et al. “Benchmarking short sequence mapping tools”. In: *BMC Bioinformatics* 14.1 (June 7, 2013), p. 184. DOI: [10.1186/1471-2105-14-184](https://doi.org/10.1186/1471-2105-14-184). URL: <https://doi.org/10.1186/1471-2105-14-184> (cit. on p. 51).

- [244] Stefan Canzar and Steven L. Salzberg. “Short Read Mapping: An Algorithmic Tour”. In: *Proceedings of the IEEE* 105.3 (Mar. 2017). Conference Name: Proceedings of the IEEE, pp. 436–458. DOI: [10.1109/JPROC.2015.2455551](https://doi.org/10.1109/JPROC.2015.2455551) (cit. on p. 51).
- [245] Karel Břinda, Valentina Boeva, and Gregory Kucherov. “RNF: a general framework to evaluate NGS read mappers”. In: *Bioinformatics* 32.1 (Jan. 1, 2016), pp. 136–139. DOI: [10.1093/bioinformatics/btv524](https://doi.org/10.1093/bioinformatics/btv524). URL: <https://doi.org/10.1093/bioinformatics/btv524> (cit. on p. 51).
- [246] Matthew Ruffalo, Thomas LaFramboise, and Mehmet Koyutürk. “Comparative analysis of algorithms for next-generation sequencing read alignment”. In: *Bioinformatics (Oxford, England)* 27.20 (Oct. 15, 2011). PMID: 21856737, pp. 2790–2796. DOI: [10.1093/bioinformatics/btr477](https://doi.org/10.1093/bioinformatics/btr477) (cit. on p. 51).
- [247] Ben Langmead. “A tandem simulation framework for predicting mapping quality”. In: *Genome Biology* 18.1 (Aug. 10, 2017), p. 152. DOI: [10.1186/s13059-017-1290-3](https://doi.org/10.1186/s13059-017-1290-3). URL: <https://doi.org/10.1186/s13059-017-1290-3> (cit. on p. 51).
- [248] Hsin-Nan Lin and Wen-Lian Hsu. “Kart: a divide-and-conquer algorithm for NGS read alignment”. In: *Bioinformatics (Oxford, England)* 33.15 (Aug. 1, 2017). PMID: 28379292 PMCID: PMC5860120, pp. 2281–2287. DOI: [10.1093/bioinformatics/btx189](https://doi.org/10.1093/bioinformatics/btx189) (cit. on p. 51).
- [249] Corey B. Olson, Maria Kim, et al. “2012 IEEE 20th International Symposium on Field-Programmable Custom Computing Machines”. In: tex.ids= olsonHardwareAccelerationShort2012a. Apr. 2012, pp. 161–168. DOI: [10.1109/FCCM.2012.621036](https://doi.org/10.1109/FCCM.2012.621036) (cit. on p. 51).
- [250] Peng Chen, Chao Wang, et al. “Accelerating the Next Generation Long Read Mapping with the FPGA-Based System”. In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 11.5 (Sept. 2014). Conference Name: IEEE/ACM Transactions on Computational Biology and Bioinformatics, pp. 840–852. DOI: [10.1109/TCBB.2014.2326876](https://doi.org/10.1109/TCBB.2014.2326876) (cit. on p. 51).
- [251] Hajime Suzuki and Masahiro Kasahara. “Introducing difference recurrence relations for faster semi-global alignment of long sequences”. In: *BMC Bioinformatics* 19.1 (Feb. 19, 2018), p. 45. DOI: [10.1186/s12859-018-2014-8](https://doi.org/10.1186/s12859-018-2014-8). URL: <https://doi.org/10.1186/s12859-018-2014-8> (cit. on p. 51).
- [252] Alberto Zeni, Giulia Guidi, et al. “2020 IEEE International Parallel and Distributed Processing Symposium (IPDPS)”. In: ISSN: 1530-2075. May 2020, pp. 462–471. DOI: [10.1109/IPDPS47924.2020.00055](https://doi.org/10.1109/IPDPS47924.2020.00055) (cit. on p. 51).
- [253] Mark J. Chaisson and Glenn Tesler. “Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory”. In: *BMC Bioinformatics* 13.1 (Sept. 19, 2012), p. 238. DOI: [10.1186/1471-2105-13-238](https://doi.org/10.1186/1471-2105-13-238). URL: <https://doi.org/10.1186/1471-2105-13-238> (cit. on p. 52).

## REFERENCES

- [254] Ehsan Haghshenas, S. Cenk Sahinalp, and Faraz Hach. “lordFAST: sensitive and Fast Alignment Search Tool for LOng noisy Read sequencing Data”. In: *Bioinformatics (Oxford, England)* 35.1 (Jan. 1, 2019). PMID: 30561550 PMCID: PMC6298053, pp. 20–27. DOI: [10.1093/bioinformatics/bty544](https://doi.org/10.1093/bioinformatics/bty544) (cit. on p. 52).
- [255] Ivan Sović, Mile Šikić, et al. “Fast and sensitive mapping of nanopore sequencing reads with GraphMap”. In: *Nature Communications* 7.1 (Apr. 15, 2016). Number: 1 Publisher: Nature Publishing Group, p. 11307. DOI: [10.1038/ncomms11307](https://doi.org/10.1038/ncomms11307). URL: <https://www.nature.com/articles/ncomms11307> (cit. on p. 52).
- [256] Fritz J. Sedlazeck, Philipp Rescheneder, et al. “Accurate detection of complex structural variations using single-molecule sequencing”. In: *Nature Methods* 15.6 (June 2018). Number: 6 Publisher: Nature Publishing Group, pp. 461–468. DOI: [10.1038/s41592-018-0001-7](https://doi.org/10.1038/s41592-018-0001-7). URL: <https://www.nature.com/articles/s41592-018-0001-7> (cit. on p. 52).
- [257] Chirag Jain, Alexander Dilthey, et al. “A Fast Approximate Algorithm for Mapping Long Reads to Large Reference Databases”. In: *Journal of Computational Biology* 25.7 (July 1, 2018). PMID: 29708767 PMCID: PMC6067103, pp. 766–779. DOI: [10.1089/cmb.2018.0036](https://doi.org/10.1089/cmb.2018.0036). URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6067103/> (cit. on p. 52).
- [258] Timofey Prodanov and Vikas Bansal. “Sensitive Alignment Using Paralogous Sequence Variants Improves Long-Read Mapping and Variant Calling in Segmental Duplications”. In: *Nucleic Acids Research* 48.19 (Nov. 4, 2020), e114. ISSN: 0305-1048. DOI: [10.1093/nar/gkaa829](https://doi.org/10.1093/nar/gkaa829) (cit. on pp. 52, 85).
- [259] Chirag Jain, Arang Rhie, et al. “Long-read mapping to repetitive reference sequences using Winnowmap2”. In: *Nature Methods* 19.6 (June 2022). Number: 6 Publisher: Nature Publishing Group, pp. 705–710. DOI: [10.1038/s41592-022-01457-8](https://doi.org/10.1038/s41592-022-01457-8). URL: <https://www.nature.com/articles/s41592-022-01457-8> (cit. on p. 52).
- [260] Alla Mikheenko, Andrey V Bzikadze, et al. “TandemTools: Mapping Long Reads and Assessing/Improving Assembly Quality in Extra-Long Tandem Repeats”. In: *Bioinformatics* 36 (Supplement\_1 July 1, 2020), pp. i75–i83. ISSN: 1367-4803. DOI: [10.1093/bioinformatics/btaa440](https://doi.org/10.1093/bioinformatics/btaa440) (cit. on pp. 52, 78).
- [261] Lauren Bragg, Glenn Stone, et al. “Fast, Accurate Error-Correction of Amplicon Pyrosequences Using Acacia”. In: *Nature Methods* 9.5 (5 May 2012), pp. 425–426. ISSN: 1548-7105. DOI: [10.1038/nmeth.1990](https://doi.org/10.1038/nmeth.1990) (cit. on p. 71).
- [262] Kristoffer Sahlin and Paul Medvedev. “Error Correction Enables Use of Oxford Nanopore Technology for Reference-Free Transcriptome Analysis”. In: *Nature Communications* 12.1 (1 Jan. 4, 2021), p. 2. ISSN: 2041-1723. DOI: [10.1038/s41467-020-20340-8](https://doi.org/10.1038/s41467-020-20340-8) (cit. on p. 71).
- [263] Hailin Liu, Shigang Wu, et al. “SMARTdenovo: A de Novo Assembler Using Long Noisy Reads”. In: *Gigabyte* 2021 (Mar. 8, 2021), pp. 1–9. DOI: [10.46471/gigabyte.15](https://doi.org/10.46471/gigabyte.15) (cit. on p. 71).

- [264] Ronald L. Graham, Donald Ervin Knuth, and Oren Patashnik. *Concrete Mathematics: A Foundation for Computer Science*. 2nd ed. Reading, Mass: Addison-Wesley, 1994. 657 pp. ISBN: 978-0-201-55802-9 (cit. on p. 77).
- [265] M. D. Adams, S. E. Celniker, et al. “The Genome Sequence of *Drosophila Melanogaster*”. In: *Science (New York, N.Y.)* 287.5461 (Mar. 24, 2000), pp. 2185–2195. ISSN: 0036-8075. DOI: [10.1126/science.287.5461.2185](https://doi.org/10.1126/science.287.5461.2185). pmid: 10731132 (cit. on p. 78).
- [266] Chen Yang, Justin Chu, et al. “NanoSim: Nanopore Sequence Read Simulator Based on Statistical Characterization”. In: *GigaScience* 6.4 (Apr. 1, 2017). ISSN: 2047-217X. DOI: [10.1093/gigascience/gix010](https://doi.org/10.1093/gigascience/gix010) (cit. on pp. 78, 157).
- [267] Arang Rhie, Brian P. Walenz, et al. “Merqury: Reference-Free Quality, Completeness, and Phasing Assessment for Genome Assemblies”. In: *Genome Biology* 21.1 (Sept. 14, 2020), p. 245. ISSN: 1474-760X. DOI: [10.1186/s13059-020-02134-9](https://doi.org/10.1186/s13059-020-02134-9) (cit. on p. 81).
- [268] Heng Li. *New Strategies to Improve Minimap2 Alignment Accuracy*. Aug. 7, 2021. arXiv: [2108.03515 \[q-bio\]](https://arxiv.org/abs/2108.03515) (cit. on p. 85).
- [269] Heng Li, Jonathan M. Bloom, et al. “A Synthetic-Diploid Benchmark for Accurate Variant-Calling Evaluation”. In: *Nature Methods* 15.8 (8 Aug. 2018), pp. 595–597. ISSN: 1548-7105. DOI: [10.1038/s41592-018-0054-7](https://doi.org/10.1038/s41592-018-0054-7) (cit. on p. 85).
- [270] A. Lwoff. “The concept of virus”. In: *Journal of General Microbiology* 17.2 (Oct. 1957). tex.ids= lwoffConceptVirus1957a PMID: 13481308, pp. 239–253. DOI: [10.1099/00221287-17-2-239](https://doi.org/10.1099/00221287-17-2-239) (cit. on p. 93).
- [271] Philip D Minor. “Viruses”. In: \_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/9780470015902.a0000441.pub3>. John Wiley & Sons, Ltd, 2014. DOI: [10.1002/9780470015902.a0000441.pub3](https://doi.org/10.1002/9780470015902.a0000441.pub3). URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470015902.a0000441.pub3> (cit. on p. 93).
- [272] Jack T. Stapleton, Steven Foug, et al. “The GB viruses: a review and proposed classification of GBV-A, GBV-C (HGV), and GBV-D in genus Pegivirus within the family Flaviviridae”. In: *The Journal of General Virology* 92.Pt 2 (Feb. 2011). PMID: 21084497 PMCID: PMC3081076, pp. 233–246. DOI: [10.1099/vir.0.027490-0](https://doi.org/10.1099/vir.0.027490-0). URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3081076/> (cit. on p. 93).
- [273] Naoki Yamamoto, Yoshihiro Sakoda, et al. “Characterization of a non-pathogenic H5N1 influenza virus isolated from a migratory duck flying from Siberia in Hokkaido, Japan, in October 2009”. In: *Virology Journal* 8.1 (Feb. 11, 2011), p. 65. DOI: [10.1186/1743-422X-8-65](https://doi.org/10.1186/1743-422X-8-65). URL: <https://doi.org/10.1186/1743-422X-8-65> (cit. on p. 93).
- [274] Mang Shi, Xian-Dan Lin, et al. “The evolutionary history of vertebrate RNA viruses”. In: *Nature* 556.7700 (Apr. 2018). Number: 7700 Publisher: Nature Publishing Group, pp. 197–202. DOI: [10.1038/s41586-018-0012-7](https://doi.org/10.1038/s41586-018-0012-7). URL: <https://www.nature.com/articles/s41586-018-0012-7> (cit. on p. 93).

## REFERENCES

- [275] Jean R. Adams and Jean R. Bonami, eds. *Atlas of Invertebrate Viruses*. DOI: 10.1201/9781315149929. Boca Raton: CRC Press, Aug. 31, 2017. DOI: 10.1201/9781315149929 (cit. on p. 93).
- [276] Pierre Lefevre, Darren P. Martin, et al. “Evolution and ecology of plant viruses”. In: *Nature Reviews Microbiology* 17.10 (Oct. 2019). Number: 10 Publisher: Nature Publishing Group, pp. 632–644. DOI: 10.1038/s41579-019-0232-3. URL: <https://www.nature.com/articles/s41579-019-0232-3> (cit. on p. 93).
- [277] A. L. Wang and C. C. Wang. “Viruses of parasitic protozoa”. In: *Parasitology Today* 7.4 (Jan. 1, 1991), pp. 76–80. DOI: 10.1016/0169-4758(91)90198-W. URL: <https://www.sciencedirect.com/science/article/pii/016947589190198W> (cit. on p. 93).
- [278] Gustavo Fermin, Sudeshna Mazumdar-Leighton, and Paula Tenant. “Viruses of prokaryotes, protozoa, fungi, and chromista”. In: Publisher: Academic Press. 2018, p. 217 (cit. on p. 93).
- [279] Suvi Sutela, Anna Poimala, and Eeva J Vainio. “Viruses of fungi and oomycetes in the soil environment”. In: *FEMS Microbiology Ecology* 95.9 (Sept. 1, 2019), fiz119. DOI: 10.1093/femsec/fiz119. URL: <https://doi.org/10.1093/femsec/fiz119> (cit. on p. 93).
- [280] F. W. Twort. “AN INVESTIGATION ON THE NATURE OF ULTRAMICROSCOPIC VIRUSES.” In: *The Lancet* 186.4814 (Dec. 4, 1915). Publisher: Elsevier, pp. 1241–1243. DOI: 10.1016/S0140-6736(01)20383-3. URL: [https://www.thelancet.com/journals/lancet/article/PIIS0140-6736\(01\)20383-3/fulltext](https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(01)20383-3/fulltext) (cit. on p. 93).
- [281] M. Delbrück. “Bacterial Viruses or Bacteriophages”. In: *Biological Reviews* 21.1 (1946). \_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1469-185X.1946.tb00451.x>, pp. 30–40. DOI: 10.1111/j.1469-185X.1946.tb00451.x. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-185X.1946.tb00451.x> (cit. on p. 93).
- [282] Jason R Clark and John B March. “Bacterial viruses as human vaccines?” In: *Expert Review of Vaccines* 3.4 (Aug. 1, 2004). Publisher: Taylor & Francis \_eprint: <https://doi.org/10.1586/14760584.3.4.463> PMID: 15270651, pp. 463–476. DOI: 10.1586/14760584.3.4.463. URL: <https://doi.org/10.1586/14760584.3.4.463> (cit. on p. 93).
- [283] H E van Kan-Davelaar, J C M van Hest, et al. “Using viruses as nanomedicines”. In: *British Journal of Pharmacology* 171.17 (2014). \_eprint: <https://onlinelibrary.wiley.com/doi/abs/10.1111/bph.12662> (cit. on p. 93).

- [284] David Prangishvili, Tamara Basta, et al. “Viruses of the Archaea”. In: \_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/9780470015902.a0000774.pub3> DOI: 10.1002/9780470015902.a0000774.pub3. John Wiley & Sons, Ltd, 2016, pp. 1–9. DOI: [10.1002 / 9780470015902 . a0000774 . pub3](https://doi.org/10.1002/9780470015902.a0000774.pub3). URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470015902.a0000774.pub3> (cit. on p. 93).
- [285] David Prangishvili, Patrick Forterre, and Roger A. Garrett. “Viruses of the Archaea: a unifying view”. In: *Nature Reviews Microbiology* 4.11 (Nov. 2006). Number: 11 Publisher: Nature Publishing Group, pp. 837–848. DOI: [10.1038 / nrmicro1527](https://doi.org/10.1038/nrmicro1527). URL: <https://www.nature.com/articles/nrmicro1527> (cit. on p. 93).
- [286] R.I.B. Francki. “Plant virus satellites”. In: *ANNUAL REVIEW OF MICROBIOLOGY RESEARCH* (). tex.ids=franckiPlantVirusSatellites, franckiPlantVirus-Satellitesa tex.recid: 10193 place: 1985. URL: <http://worldveg.tind.io/record/10193> (cit. on p. 93).
- [287] Ping Xu and Marilyn J Roossinck. “Plant Virus Satellites”. In: \_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/9780470015902.a0000771.pub2> DOI: 10.1002/9780470015902.a0000771.pub2. John Wiley & Sons, Ltd, 2011. DOI: [10.1002 / 9780470015902 . a0000771 . pub2](https://doi.org/10.1002/9780470015902.a0000771.pub2). URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470015902.a0000771.pub2> (cit. on p. 93).
- [288] M. M. Lai. “The molecular biology of hepatitis delta virus”. In: *Annual review of biochemistry* 64 (Jan. 1, 1995). PMID: 7574482, pp. 259–286. DOI: [10.1146 / annurev.bi.64.070195.001355](https://doi.org/10.1146/annurev.bi.64.070195.001355). URL: <https://doi.org/10.1146/annurev.bi.64.070195.001355> (cit. on p. 93).
- [289] Sarah A Hughes, Heiner Wedemeyer, and Phillip M Harrison. “Hepatitis delta virus”. In: *The Lancet* 378.9785 (July 2, 2011), pp. 73–85. DOI: [10.1016 / S0140-6736\(10\)61931-9](https://doi.org/10.1016/S0140-6736(10)61931-9). URL: <https://www.sciencedirect.com/science/article/pii/S0140673610619319> (cit. on p. 93).
- [290] Christelle Desnues, Mickaël Boyer, and Didier Raoult. “Chapter 3 - Sputnik, a Virophage Infecting the Viral Domain of Life”. In: ed. by Małgorzata Łobocka and Wacław T. Szybalski. Vol. 82. Bacteriophages, Part A. DOI: 10.1016/B978-0-12-394621-8.00013-3. Academic Press, Jan. 1, 2012, pp. 63–89. DOI: [10.1016 / B978-0-12-394621-8.00013-3](https://doi.org/10.1016/B978-0-12-394621-8.00013-3). URL: <https://www.sciencedirect.com/science/article/pii/B9780123946218000133> (cit. on p. 93).
- [291] Morgan Gaia, Samia Benamar, et al. “Zamilon, a Novel Virophage with Mimiviridae Host Specificity”. In: *PLOS ONE* 9.4 (Apr. 18, 2014). Publisher: Public Library of Science, e94923. DOI: [10.1371 / journal . pone . 0094923](https://doi.org/10.1371/journal.pone.0094923). URL: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0094923> (cit. on p. 93).

## REFERENCES

- [292] Robert C. Edgar, Jeff Taylor, et al. “Petabase-scale sequence alignment catalyses viral discovery”. In: *Nature* 602.7895 (Feb. 2022). Number: 7895 Publisher: Nature Publishing Group, pp. 142–147. DOI: [10.1038/s41586-021-04332-2](https://doi.org/10.1038/s41586-021-04332-2). URL: <https://www.nature.com/articles/s41586-021-04332-2> (cit. on p. 93).
- [293] Arshan Nasir, Ethan Romero-Severson, and Jean-Michel Claverie. “Investigating the Concept and Origin of Viruses”. In: *Trends in Microbiology* 28.12 (Dec. 1, 2020), pp. 959–967. DOI: [10.1016/j.tim.2020.08.003](https://doi.org/10.1016/j.tim.2020.08.003). URL: <https://www.sciencedirect.com/science/article/pii/S0966842X20302304> (cit. on p. 93).
- [294] Patrick Forterre and David Prangishvili. “The origin of viruses”. In: *Research in Microbiology*. Special issue on the origin of life and microbiology 160.7 (Sept. 1, 2009), pp. 466–472. DOI: [10.1016/j.resmic.2009.07.008](https://doi.org/10.1016/j.resmic.2009.07.008). URL: <https://www.sciencedirect.com/science/article/pii/S0923250809001065> (cit. on p. 93).
- [295] Patrick Forterre. “The origin of viruses and their possible roles in major evolutionary transitions”. In: *Virus Research*. Comparative Genomics and Evolution of Complex Viruses 117.1 (Apr. 1, 2006), pp. 5–16. DOI: [10.1016/j.virusres.2006.01.010](https://doi.org/10.1016/j.virusres.2006.01.010). URL: <https://www.sciencedirect.com/science/article/pii/S0168170206000293> (cit. on p. 93).
- [296] JD Boeke and JP Stoye. “Retrotransposons, Endogenous Retroviruses, and the Evolution of Retroelement”. In: (1997). PMID: 21433351. URL: <https://www.ncbi.nlm.nih.gov/books/NBK19468> (cit. on p. 94).
- [297] Shohei Kojima, Kohei Yoshikawa, et al. “Virus-like insertions with sequence signatures similar to those of endogenous nonretroviral RNA viruses in the human genome”. In: *Proceedings of the National Academy of Sciences* 118.5 (Feb. 2, 2021). Publisher: Proceedings of the National Academy of Sciences, e2010758118. DOI: [10.1073/pnas.2010758118](https://doi.org/10.1073/pnas.2010758118). URL: <https://www.pnas.org/doi/10.1073/pnas.2010758118> (cit. on p. 94).
- [298] R Löwer, J Löwer, and R Kurth. “The viruses in all of us: characteristics and biological significance of human endogenous retrovirus sequences.” In: *Proceedings of the National Academy of Sciences* 93.11 (1996). tex.eprint: <https://www.pnas.org/doi/pdf/10.1073/pnas.93.11.5177>, pp. 5177–5184. DOI: [10.1073/pnas.93.11.5177](https://doi.org/10.1073/pnas.93.11.5177). URL: <https://www.pnas.org/doi/abs/10.1073/pnas.93.11.5177> (cit. on p. 94).
- [299] David J. Griffiths. “Endogenous retroviruses in the human genome sequence”. In: *Genome Biology* 2.6 (June 5, 2001), reviews1017.1. DOI: [10.1186/gb-2001-2-6-reviews1017](https://doi.org/10.1186/gb-2001-2-6-reviews1017). URL: <https://doi.org/10.1186/gb-2001-2-6-reviews1017> (cit. on p. 94).
- [300] D. Baltimore. “Expression of animal virus genomes”. In: *Bacteriological Reviews* 35.3 (Sept. 1971). PMID: 4329869 PMCID: PMC378387, pp. 235–241. DOI: [10.1128/br.35.3.235-241.1971](https://doi.org/10.1128/br.35.3.235-241.1971) (cit. on p. 94).

- [301] Eugene V. Koonin, Mart Krupovic, and Vadim I. Agol. “The Baltimore Classification of Viruses 50 Years Later: How Does It Stand in the Light of Virus Evolution?” In: *Microbiology and Molecular Biology Reviews* 85.3 (Aug. 18, 2021). Publisher: American Society for Microbiology, e00053–21. doi: [10.1128/MMBR.00053-21](https://doi.org/10.1128/MMBR.00053-21). URL: <https://journals.asm.org/doi/full/10.1128/MMBR.00053-21> (cit. on p. 94).
- [302] Esteban Domingo and Celia Perales. “RNA Virus Genomes”. In: \_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/9780470015902.a0001488.pub3> DOI: [10.1002/9780470015902.a0001488.pub3](https://doi.org/10.1002/9780470015902.a0001488.pub3). John Wiley & Sons, Ltd, 2018, pp. 1–12. doi: [10.1002/9780470015902.a0001488.pub3](https://doi.org/10.1002/9780470015902.a0001488.pub3). URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470015902.a0001488.pub3> (cit. on p. 94).
- [303] Duncan J. McGeoch, Frazer J. Rixon, and Andrew J. Davison. “Topics in herpesvirus genomics and evolution”. In: *Virus Research*. Comparative Genomics and Evolution of Complex Viruses 117.1 (Apr. 1, 2006), pp. 90–104. doi: [10.1016/j.virusres.2006.01.002](https://doi.org/10.1016/j.virusres.2006.01.002). URL: <https://www.sciencedirect.com/science/article/pii/S0168170206000219> (cit. on p. 94).
- [304] Paul Boehmer and Amitabh Nimonkar. “Herpes Virus Replication”. In: *IUBMB Life* 55.1 (2003). \_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1080/1521654031000070645>, pp. 13–22. doi: [10.1080/1521654031000070645](https://doi.org/10.1080/1521654031000070645). URL: <https://onlinelibrary.wiley.com/doi/abs/10.1080/1521654031000070645> (cit. on p. 94).
- [305] Mathijs H Brentjens, Kimberly A Yeung-Yue, et al. “Human papillomavirus: a review”. In: *Dermatologic Clinics* 20.2 (Apr. 1, 2002), pp. 315–331. doi: [10.1016/S0733-8635\(01\)00028-6](https://doi.org/10.1016/S0733-8635(01)00028-6). URL: <https://www.sciencedirect.com/science/article/pii/S0733863501000286> (cit. on p. 94).
- [306] Alan Kay and Fabien Zoulim. “Hepatitis B virus genetic variability and evolution”. In: *Virus Research*. Replicative and Evolutionary Aspects of Hepatitis Viruses 127.2 (Aug. 1, 2007), pp. 164–176. doi: [10.1016/j.virusres.2007.02.021](https://doi.org/10.1016/j.virusres.2007.02.021). URL: <https://www.sciencedirect.com/science/article/pii/S0168170207000615> (cit. on p. 94).
- [307] U D Parashar, J S Bresee, et al. “Rotavirus.” In: *Emerging Infectious Diseases* 4.4 (1998). PMID: 9866732 PMCID: PMC2640254, pp. 561–570. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2640254/> (cit. on p. 94).
- [308] Peter Simmonds. “Variability of hepatitis C virus”. In: *Hepatology* 21.2 (Feb. 1, 1995), pp. 570–583. doi: [10.1016/0270-9139\(95\)90121-3](https://doi.org/10.1016/0270-9139(95)90121-3). URL: <https://www.sciencedirect.com/science/article/pii/0270913995901213> (cit. on p. 94).
- [309] Eckard Wimmer, Christopher U. T. Hellen, and Xuemei Cao. “Genetics of poliovirus”. In: *Annual Review of Genetics* 27 (Jan. 1, 1993). Publisher: Annual Reviews, Inc., pp. 353–437. URL: <https://go.gale.com/ps/i.do?p=AONE&sw=w&issn=00664197&v=2.1&it=r&id=GALE%7CA15539705&sid=googleScholar&linkaccess=abs> (cit. on p. 94).

## REFERENCES

- [310] Vincent R. Racaniello. “One hundred years of poliovirus pathogenesis”. In: *Virology*. Virology 50th Anniversary Issue 344.1 (Jan. 5, 2006), pp. 9–16. DOI: [10.1016/j.virol.2005.09.015](https://doi.org/10.1016/j.virol.2005.09.015). URL: <https://www.sciencedirect.com/science/article/pii/S0042682205005830> (cit. on p. 94).
- [311] P Palese, H Zheng, et al. “Negative-strand RNA viruses: genetic engineering and applications.” In: *Proceedings of the National Academy of Sciences* 93.21 (Oct. 15, 1996). Publisher: Proceedings of the National Academy of Sciences, pp. 11354–11358. DOI: [10.1073/pnas.93.21.11354](https://doi.org/10.1073/pnas.93.21.11354). URL: <https://www.pnas.org/doi/abs/10.1073/pnas.93.21.11354> (cit. on p. 94).
- [312] Esteban Domingo and Celia Perales. “Virus Evolution”. In: [\\_eprint: https://onlinelibrary.wiley.com/doi/10.1002/9780470015902.a0000436.pub3](https://onlinelibrary.wiley.com/doi/10.1002/9780470015902.a0000436.pub3) DOI: [10.1002/9780470015902.a0000436.pub3](https://doi.org/10.1002/9780470015902.a0000436.pub3). John Wiley & Sons, Ltd, 2014. DOI: [10.1002/9780470015902.a0000436.pub3](https://doi.org/10.1002/9780470015902.a0000436.pub3). URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470015902.a0000436.pub3> (cit. on p. 94).
- [313] Philip V'kovski, Annika Kratzel, et al. “Coronavirus biology and replication: implications for SARS-CoV-2”. In: *Nature Reviews Microbiology* 19.3 (Mar. 2021). Number: 3 Publisher: Nature Publishing Group, pp. 155–170. DOI: [10.1038/s41579-020-00468-6](https://doi.org/10.1038/s41579-020-00468-6). URL: <https://www.nature.com/articles/s41579-020-00468-6> (cit. on p. 94).
- [314] Anne Tuiskunen Bäck and Åke Lundkvist. “Dengue viruses – an overview”. In: *Infection Ecology & Epidemiology* 3 (Aug. 30, 2013). PMID: 24003364 PMCID: PMC3759171, 10.3402/iee.v3i0.19839. DOI: [10.3402/iee.v3i0.19839](https://doi.org/10.3402/iee.v3i0.19839). URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3759171/> (cit. on p. 94).
- [315] Lynn B. Dustin, Barbara Bartolini, et al. “Hepatitis C virus: life cycle in cells, infection and host response, and analysis of molecular markers influencing the outcome of infection and response to therapy”. In: *Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases* 22.10 (Oct. 2016). PMID: 27592089 PMCID: PMC5627509, pp. 826–832. DOI: [10.1016/j.cmi.2016.08.025](https://doi.org/10.1016/j.cmi.2016.08.025). URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5627509/> (cit. on p. 94).
- [316] Meelis Kadaja, Toomas Silla, et al. “Papillomavirus DNA replication — From initiation to genomic instability”. In: *Virology*. Small Viruses, Big Discoveries: The Interwoven Story of the Small DNA Tumor Viruses 384.2 (Feb. 20, 2009), pp. 360–368. DOI: [10.1016/j.virol.2008.11.032](https://doi.org/10.1016/j.virol.2008.11.032). URL: <https://www.sciencedirect.com/science/article/pii/S0042682208007691> (cit. on p. 94).
- [317] Sandra K. Weller and Donald M. Coen. “Herpes Simplex Viruses: Mechanisms of DNA Replication”. In: *Cold Spring Harbor Perspectives in Biology* 4.9 (Sept. 2012). PMID: 22952399 PMCID: PMC3428768, a013011. DOI: [10.1101/cshperspect.a013011](https://doi.org/10.1101/cshperspect.a013011). URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3428768/> (cit. on p. 94).

- [318] Juergen Beck and Michael Nassal. "Hepatitis B virus replication". In: *World Journal of Gastroenterology : WJG* 13.1 (Jan. 7, 2007). PMID: 17206754 PMCID: PMC4065876, pp. 48–64. DOI: [10.3748/wjg.v13.i1.48](https://doi.org/10.3748/wjg.v13.i1.48). URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4065876/> (cit. on p. 94).
- [319] Jesse D. Pyle and Karen-Beth G. Scholthof. "Chapter 58 - Biology and Pathogenesis of Satellite Viruses". In: ed. by Ahmed Hadidi, Ricardo Flores, et al. DOI: [10.1016/B978-0-12-801498-1.00058-9](https://doi.org/10.1016/B978-0-12-801498-1.00058-9). Boston: Academic Press, Jan. 1, 2017, pp. 627–636. DOI: [10.1016/B978-0-12-801498-1.00058-9](https://doi.org/10.1016/B978-0-12-801498-1.00058-9). URL: <https://www.sciencedirect.com/science/article/pii/B9780128014981000589> (cit. on p. 94).
- [320] Didier Raoult, Stéphane Audic, et al. "The 1.2-megabase genome sequence of Mimivirus". In: *Science (New York, N.Y.)* 306.5700 (Nov. 19, 2004). PMID: 15486256, pp. 1344–1350. DOI: [10.1126/science.1101485](https://doi.org/10.1126/science.1101485) (cit. on p. 94).
- [321] José A. Campillo-Balderas, Antonio Lazcano, and Arturo Becerra. "Viral Genome Size Distribution Does not Correlate with the Antiquity of the Host Lineages". In: *Frontiers in Ecology and Evolution* 3 (2015). URL: <https://www.frontiersin.org/article/10.3389/fevo.2015.00143> (cit. on p. 94).
- [322] Alan J Cann. "Virus Structure". In: \_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/9780470015902.a0000439.pub2>. John Wiley & Sons, Ltd, 2015, pp. 1–9. DOI: [10.1002/9780470015902.a0000439.pub2](https://doi.org/10.1002/9780470015902.a0000439.pub2). URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470015902.a0000439.pub2> (cit. on p. 94).
- [323] Florian Hladik and M. Juliana McElrath. "Setting the stage: host invasion by HIV". In: *Nature Reviews Immunology* 8.6 (June 2008). Number: 6 Publisher: Nature Publishing Group, pp. 447–457. DOI: [10.1038/nri2302](https://doi.org/10.1038/nri2302). URL: <https://www.nature.com/articles/nri2302> (cit. on p. 95).
- [324] George M. Shaw and Eric Hunter. "HIV Transmission". In: *Cold Spring Harbor Perspectives in Medicine* 2.11 (Jan. 11, 2012). Publisher: Cold Spring Harbor Laboratory Press PMID: 23043157, a006965. DOI: [10.1101/cshperspect.a006965](https://doi.org/10.1101/cshperspect.a006965). URL: <http://perspectivesinmedicine.cshlp.org/content/2/11/a006965> (cit. on p. 95).
- [325] Robin A. Weiss. "How Does HIV Cause AIDS?" In: *Science* 260.5112 (May 28, 1993). Publisher: American Association for the Advancement of Science, pp. 1273–1279. DOI: [10.1126/science.8493571](https://doi.org/10.1126/science.8493571). URL: <https://www.science.org/doi/10.1126/science.8493571> (cit. on p. 95).
- [326] Anne Melhuish and Penny Lewthwaite. "Natural history of HIV and AIDS". In: *Medicine* 46.6 (June 1, 2018), pp. 356–361. DOI: [10.1016/j.mpmed.2018.03.010](https://doi.org/10.1016/j.mpmed.2018.03.010). URL: <https://www.sciencedirect.com/science/article/pii/S1357303918300744> (cit. on p. 95).

## REFERENCES

- [327] John F. Murray, Charles P. Felton, et al. “Pulmonary complications of the acquired immunodeficiency syndrome”. In: *New England Journal of Medicine* 310.25 (1984). tex.eprint: <https://doi.org/10.1056/NEJM198406213102529>. URL: <https://doi.org/10.1056/NEJM198406213102529> (cit. on p. 95).
- [328] Shrikanth Sampath, Anwar Khedr, et al. “Pandemics Throughout the History”. In: *Cureus* 13.9 (Sept. 20, 2021). Publisher: Cureus. DOI: [10.7759/cureus.18136](https://doi.org/10.7759/cureus.18136). URL: <https://www.cureus.com/articles/69273-pandemics-throughout-the-history> (cit. on p. 95).
- [329] World Health Organization. *Global report: UNAIDS report on the global AIDS epidemic 2010*. OCLC: 780905192. Geneve: World Health Organization, 2010. URL: [https://www.unaids.org/globalreport/Global\\_report.htm](https://www.unaids.org/globalreport/Global_report.htm) (cit. on p. 95).
- [330] F. Barré-Sinoussi, J. C. Chermann, et al. “Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS)”. In: *Science (New York, N.Y.)* 220.4599 (May 20, 1983). PMID: 6189183, pp. 868–871. DOI: [10.1126/science.6189183](https://doi.org/10.1126/science.6189183) (cit. on p. 95).
- [331] R. C. Gallo, P. S. Sarin, et al. “Isolation of human T-cell leukemia virus in acquired immune deficiency syndrome (AIDS)”. In: *Science (New York, N.Y.)* 220.4599 (May 20, 1983). PMID: 6601823, pp. 865–867. DOI: [10.1126/science.6601823](https://doi.org/10.1126/science.6601823) (cit. on p. 95).
- [332] François Clavel, Denise Guétard, et al. “Isolation of a New Human Retrovirus from West African Patients with AIDS”. In: *Science* 233.4761 (July 18, 1986). Publisher: American Association for the Advancement of Science, pp. 343–346. DOI: [10.1126/science.2425430](https://doi.org/10.1126/science.2425430). URL: <https://www.science.org/doi/abs/10.1126/science.2425430> (cit. on p. 95).
- [333] Peter B. Gilbert, Ian W. McKeague, et al. “Comparison of HIV-1 and HIV-2 infectivity from a prospective cohort study in Senegal”. In: *Statistics in Medicine* 22.4 (Feb. 28, 2003), pp. 573–593. DOI: [10.1002/sim.1342](https://doi.org/10.1002/sim.1342). URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.1342> (cit. on p. 95).
- [334] Maarten F. Schim van der Loeff, Akum Aveika Awasana, et al. “Sixteen years of HIV surveillance in a West African research clinic reveals divergent epidemic trends of HIV-1 and HIV-2”. In: *International Journal of Epidemiology* 35.5 (Oct. 2006). PMID: 16543363, pp. 1322–1328. DOI: [10.1093/ije/dyl037](https://doi.org/10.1093/ije/dyl037) (cit. on p. 95).
- [335] Feng Gao, Elizabeth Bailes, et al. “Origin of HIV-1 in the chimpanzee Pan troglodytes troglodytes”. In: *Nature* 397.6718 (Feb. 1999). Number: 6718 Publisher: Nature Publishing Group, pp. 436–441. DOI: [10.1038/17130](https://doi.org/10.1038/17130). URL: <https://www.nature.com/articles/17130> (cit. on p. 95).
- [336] Donald J. Hamel, Jean-Louis Sankalé, et al. “Twenty years of prospective molecular epidemiology in Senegal: changes in HIV diversity”. In: *AIDS research and human retroviruses* 23.10 (Oct. 2007). PMID: 17961103, pp. 1189–1196. DOI: [10.1089/aid.2007.0037](https://doi.org/10.1089/aid.2007.0037) (cit. on p. 95).

- [337] Paul M. Sharp and Beatrice H. Hahn. “Origins of HIV and the AIDS Pandemic”. In: *Cold Spring Harbor Perspectives in Medicine*: 1.1 (Sept. 2011). PMID: 22229120 PMCID: PMC3234451, a006841. DOI: [10.1101/cshperspect.a006841](https://doi.org/10.1101/cshperspect.a006841). URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3234451/> (cit. on pp. 95, 96).
- [338] Vanessa M. Hirsch, Robert A. Olmsted, et al. “An African primate lentivirus (SIVsmclosely related to HIV-2)”. In: *Nature* 339.6223 (June 1989). Number: 6223 Publisher: Nature Publishing Group, pp. 389–392. DOI: [10.1038/339389a0](https://doi.org/10.1038/339389a0). URL: <https://www.nature.com/articles/339389a0> (cit. on p. 95).
- [339] Feng Gao, Ling Yue, et al. “Human infection by genetically diverse SIVSM-related HIV-2 in West Africa”. In: *Nature* 358.6386 (Aug. 1992). Number: 6386 Publisher: Nature Publishing Group, pp. 495–499. DOI: [10.1038/358495a0](https://doi.org/10.1038/358495a0). URL: <https://www.nature.com/articles/358495a0> (cit. on p. 95).
- [340] Z. Chen, P. Telfier, et al. “Genetic characterization of new West African simian immunodeficiency virus SIVsm: geographic clustering of household-derived SIV strains with human immunodeficiency virus type 2 subtypes and genetically diverse viruses from a single feral sooty mangabey troop”. In: *Journal of Virology* 70.6 (June 1996). PMID: 8648696 PMCID: PMC190237, pp. 3617–3627. DOI: [10.1128/JVI.70.6.3617-3627.1996](https://doi.org/10.1128/JVI.70.6.3617-3627.1996) (cit. on p. 95).
- [341] Joris Hemelaar. “The origin and diversity of the HIV-1 pandemic”. In: *Trends in Molecular Medicine* 18.3 (Mar. 1, 2012), pp. 182–192. DOI: [10.1016/j.molmed.2011.12.001](https://doi.org/10.1016/j.molmed.2011.12.001). URL: <https://www.sciencedirect.com/science/article/pii/S1471491411002103> (cit. on p. 95).
- [342] Michael Worobey, Marlea Gemmel, et al. “Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960”. In: *Nature* 455.7213 (Oct. 2008). Number: 7213 Publisher: Nature Publishing Group, pp. 661–664. DOI: [10.1038/nature07390](https://doi.org/10.1038/nature07390). URL: <https://www.nature.com/articles/nature07390> (cit. on p. 96).
- [343] N. Vidal, M. Peeters, et al. “Unprecedented degree of human immunodeficiency virus type 1 (HIV-1) group M genetic diversity in the Democratic Republic of Congo suggests that the HIV-1 pandemic originated in Central Africa”. In: *Journal of Virology* 74.22 (Nov. 2000). PMID: 11044094 PMCID: PMC110924, pp. 10498–10507. DOI: [10.1128/jvi.74.22.10498-10507.2000](https://doi.org/10.1128/jvi.74.22.10498-10507.2000) (cit. on p. 96).
- [344] Nuno R. Faria, Andrew Rambaut, et al. “The early spread and epidemic ignition of HIV-1 in human populations”. In: *Science (New York, N.Y.)* 346.6205 (Oct. 3, 2014). PMID: 25278604 PMCID: PMC4254776, pp. 56–61. DOI: [10.1126/science.1256739](https://doi.org/10.1126/science.1256739). URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4254776/> (cit. on p. 96).
- [345] B. Korber, M. Muldoon, et al. “Timing the ancestor of the HIV-1 pandemic strains”. In: *Science (New York, N.Y.)* 288.5472 (June 9, 2000). PMID: 10846155, pp. 1789–1796. DOI: [10.1126/science.288.5472.1789](https://doi.org/10.1126/science.288.5472.1789) (cit. on p. 96).

## REFERENCES

- [346] Andrew Rambaut, David Posada, et al. “The causes and consequences of HIV evolution”. In: *Nature Reviews Genetics* 5.1 (Jan. 2004). Number: 1 Publisher: Nature Publishing Group, pp. 52–61. DOI: [10.1038/nrg1246](https://doi.org/10.1038/nrg1246). URL: <https://www.nature.com/articles/nrg1246> (cit. on p. 96).
- [347] Francine E. McCutchan. “Global epidemiology of HIV”. In: *Journal of Medical Virology* 78.S1 (2006). \_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/jmv.20599>, S7–S12. DOI: [10.1002/jmv.20599](https://doi.org/10.1002/jmv.20599). URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/jmv.20599> (cit. on p. 96).
- [348] Marcos Pérez-Losada, Miguel Arenas, et al. “Recombination in viruses: Mechanisms, methods of study, and evolutionary consequences”. In: *Infection, Genetics and Evolution* 30 (Mar. 2015). PMID: 25541518 PMCID: PMC7106159, pp. 296–307. DOI: [10.1016/j.meegid.2014.12.022](https://doi.org/10.1016/j.meegid.2014.12.022). URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7106159/> (cit. on p. 96).
- [349] D. L. Robertson, B. H. Hahn, and P. M. Sharp. “Recombination in AIDS viruses”. In: *Journal of Molecular Evolution* 40.3 (Mar. 1995). tex.ids=robertsonRecombinationAIDSViruses1995a PMID: 7723052, pp. 249–259. DOI: [10.1007/BF00163230](https://doi.org/10.1007/BF00163230) (cit. on p. 96).
- [350] *HIV Circulating Recombinant Forms (CRFs)*. URL: <https://www.hiv.lanl.gov/content/sequence/HIV/CRFs/CRFs.html> (cit. on p. 96).
- [351] David Posada, Keith A. Crandall, and Edward C. Holmes. “Recombination in evolutionary genomics”. In: *Annual Review of Genetics* 36 (2002). PMID: 12429687, pp. 75–97. DOI: [10.1146/annurev.genet.36.040202.111115](https://doi.org/10.1146/annurev.genet.36.040202.111115) (cit. on p. 96).
- [352] Barbara S. Taylor, Magdalena E. Sobieszczyk, et al. “The Challenge of HIV-1 Subtype Diversity”. In: *New England Journal of Medicine* 358.15 (Apr. 10, 2008). Publisher: Massachusetts Medical Society \_eprint: <https://doi.org/10.1056/NEJMra0706737> PMID: 18403767, pp. 1590–1602. DOI: [10.1056/NEJMra0706737](https://doi.org/10.1056/NEJMra0706737). URL: <https://doi.org/10.1056/NEJMra0706737> (cit. on p. 96).
- [353] Joris Hemelaar, Eleanor Gouws, et al. “Global trends in molecular epidemiology of HIV-1 during 2000–2007”. In: *AIDS (London, England)* 25.5 (Mar. 13, 2011). PMID: 21297424 PMCID: PMC3755761, pp. 679–689. DOI: [10.1097/QAD.0b013e328342ff93](https://doi.org/10.1097/QAD.0b013e328342ff93). URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3755761/> (cit. on p. 96).
- [354] *Distribution of all HIV-1 sequences: WORLD*. URL: <https://www.hiv.lanl.gov/components/sequence/HIV/geo/geo.comp> (cit. on p. 96).
- [355] Eric O. Freed. “HIV-1 Replication”. In: *Somatic Cell and Molecular Genetics* 26.1 (Nov. 1, 2001), pp. 13–33. DOI: [10.1023/A:1021070512287](https://doi.org/10.1023/A:1021070512287). URL: <https://doi.org/10.1023/A:1021070512287> (cit. on p. 96).
- [356] Monique R. Ferguson, Daniel R. Rojo, et al. “HIV-1 replication cycle”. In: *Clinics in Laboratory Medicine* 22.3 (Sept. 2002). PMID: 12244589, pp. 611–635. DOI: [10.1016/s0272-2712\(02\)00015-x](https://doi.org/10.1016/s0272-2712(02)00015-x) (cit. on p. 96).

- [357] M. L. Gougeon, A. G. Laurent-Crawford, et al. “Direct and indirect mechanisms mediating apoptosis during HIV infection: contribution to in vivo CD4 T cell depletion”. In: *Seminars in Immunology* 5.3 (June 1, 1993), pp. 187–194. DOI: [10.1006/smim.1993.1022](https://doi.org/10.1006/smim.1993.1022). URL: <https://www.sciencedirect.com/science/article/pii/S1044532383710225> (cit. on p. 97).
- [358] K. K. Vidya Vijayan, Krithika Priyadarshini Karthigeyan, et al. “Pathophysiology of CD4+ T-Cell Depletion in HIV-1 and HIV-2 Infections”. In: *Frontiers in Immunology* 8 (May 23, 2017). PMID: 28588579 PMCID: PMC5440548, p. 580. DOI: [10.3389/fimmu.2017.00580](https://doi.org/10.3389/fimmu.2017.00580). URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5440548/> (cit. on p. 97).
- [359] A. D. Frankel and J. A. Young. “HIV-1: fifteen proteins and an RNA”. In: *Annual Review of Biochemistry* 67 (1998). PMID: 9759480, pp. 1–25. DOI: [10.1146/annurev.biochem.67.1.1](https://doi.org/10.1146/annurev.biochem.67.1.1) (cit. on p. 97).
- [360] Torgils Fossen, Victor Wray, et al. “Solution Structure of the Human Immunodeficiency Virus Type 1 p6 Protein \*”. In: *Journal of Biological Chemistry* 280.52 (Dec. 30, 2005). Publisher: Elsevier PMID: 16234236, pp. 42515–42527. DOI: [10.1074/jbc.M507375200](https://doi.org/10.1074/jbc.M507375200). URL: [https://www.jbc.org/article/S0021-9258\(19\)47854-X/abstract](https://www.jbc.org/article/S0021-9258(19)47854-X/abstract) (cit. on p. 98).
- [361] H G Göttlinger, T Dorfman, et al. “Effect of mutations affecting the p6 gag protein on human immunodeficiency virus particle release.” In: *Proceedings of the National Academy of Sciences* 88.8 (Apr. 15, 1991). Publisher: Proceedings of the National Academy of Sciences, pp. 3195–3199. DOI: [10.1073/pnas.88.8.3195](https://doi.org/10.1073/pnas.88.8.3195). URL: <https://www.pnas.org/doi/abs/10.1073/pnas.88.8.3195> (cit. on p. 98).
- [362] M Huang, J M Orenstein, et al. “p6Gag is required for particle production from full-length human immunodeficiency virus type 1 molecular clones expressing protease”. In: *Journal of Virology* 69.11 (Nov. 1995). Publisher: American Society for Microbiology, pp. 6810–6818. DOI: [10.1128/jvi.69.11.6810-6818.1995](https://doi.org/10.1128/jvi.69.11.6810-6818.1995). URL: <https://journals.asm.org/doi/abs/10.1128/jvi.69.11.6810-6818.1995> (cit. on p. 98).
- [363] S Bour, R Gelezunas, and M A Wainberg. “The human immunodeficiency virus type 1 (HIV-1) CD4 receptor and its central role in promotion of HIV-1 infection”. In: *Microbiological Reviews* 59.1 (Mar. 1995). Publisher: American Society for Microbiology, pp. 63–93. DOI: [10.1128/mr.59.1.63-93.1995](https://doi.org/10.1128/mr.59.1.63-93.1995). URL: <https://journals.asm.org/doi/abs/10.1128/mr.59.1.63-93.1995> (cit. on p. 98).
- [364] L. D. Hernandez, L. R. Hoffman, et al. “Virus-cell and cell-cell fusion”. In: *Annual Review of Cell and Developmental Biology* 12 (1996). tex.ids= hernandezVIR-USCELLCELLCELLFUSION1996 PMID: 8970739, pp. 627–661. DOI: [10.1146/annurev.cellbio.12.1.627](https://doi.org/10.1146/annurev.cellbio.12.1.627) (cit. on p. 98).

## REFERENCES

- [365] Ka Jones and Bm Peterlin. “Control of Rna Initiation and Elongation at the Hiv-1 Promoter”. In: *Annual Review of Biochemistry* 63 (1994). Place: Palo Alto Publisher: Annual Reviews WOS:A1994NV05900021, pp. 717–743. DOI: [10.1146/annurev.bi.63.070194.003441](https://doi.org/10.1146/annurev.bi.63.070194.003441). URL: <https://www.annualreviews.org/doi/10.1146/annurev.bi.63.070194.003441> (cit. on p. 99).
- [366] Thomas J. Hope. “Viral RNA export”. In: *Chemistry & Biology* 4.5 (May 1, 1997), pp. 335–344. DOI: [10.1016/S1074-5521\(97\)90124-1](https://doi.org/10.1016/S1074-5521(97)90124-1). URL: <https://www.sciencedirect.com/science/article/pii/S1074552197901241> (cit. on p. 99).
- [367] A. Mangasarian and D. Trono. “The multifaceted role of HIV Nef”. In: *Research in Virology* 148.1 (Jan. 1, 1997), pp. 30–33. DOI: [10.1016/S0923-2516\(97\)81909-7](https://doi.org/10.1016/S0923-2516(97)81909-7). URL: <https://www.sciencedirect.com/science/article/pii/S0923251697819097> (cit. on p. 99).
- [368] é. A. Cohen, R. A. Subbramanian, and H. G. Göttlinger. “Role of Auxiliary Proteins in Retroviral Morphogenesis”. In: ed. by Hans-Georg Kräusslich. Current Topics in Microbiology and Immunology. DOI: [10.1007/978-3-642-80145-7\\_7](https://doi.org/10.1007/978-3-642-80145-7_7). Berlin, Heidelberg: Springer, 1996, pp. 219–235. DOI: [10.1007/978-3-642-80145-7\\_7](https://doi.org/10.1007/978-3-642-80145-7_7). URL: [https://doi.org/10.1007/978-3-642-80145-7\\_7](https://doi.org/10.1007/978-3-642-80145-7_7) (cit. on p. 99).
- [369] Robert A Lamb and Lawrence H Pinto. “Do Vpu and Vpr of Human Immunodeficiency Virus Type 1 and NB of Influenza B Virus Have Ion Channel Activities in the Viral Life Cycles?” In: *Virology* 229.1 (Mar. 3, 1997), pp. 1–11. DOI: [10.1006/viro.1997.8451](https://doi.org/10.1006/viro.1997.8451). URL: <https://www.sciencedirect.com/science/article/pii/S0042682297984518> (cit. on p. 99).
- [370] Nabab Khan and Jonathan D. Geiger. “Role of Viral Protein U (Vpu) in HIV-1 Infection and Pathogenesis”. In: *Viruses* 13.8 (July 27, 2021). PMID: 34452331 PMCID: PMC8402909, p. 1466. DOI: [10.3390/v13081466](https://doi.org/10.3390/v13081466). URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8402909/> (cit. on p. 99).
- [371] Michael Emerman. “HIV-1, Vpr and the cell cycle”. In: *Current Biology* 6.9 (Sept. 1, 1996), pp. 1096–1103. DOI: [10.1016/S0960-9822\(02\)00676-0](https://doi.org/10.1016/S0960-9822(02)00676-0). URL: <https://www.sciencedirect.com/science/article/pii/S0960982202006760> (cit. on p. 99).
- [372] R. H. Miller. “Human immunodeficiency virus may encode a novel protein on the genomic DNA plus strand”. In: *Science (New York, N.Y.)* 239.4846 (Mar. 18, 1988). PMID: 3347840, pp. 1420–1422. DOI: [10.1126/science.3347840](https://doi.org/10.1126/science.3347840) (cit. on p. 99).
- [373] Sylvie Briquet and Catherine Vaquero. “Immunolocalization Studies of an Antisense Protein in HIV-1-Infected Cells and Viral Particles”. In: *Virology* 292.2 (Jan. 20, 2002), pp. 177–184. DOI: [10.1006/viro.2001.1224](https://doi.org/10.1006/viro.2001.1224). URL: <https://www.sciencedirect.com/science/article/pii/S0042682201912243> (cit. on p. 99).

- [374] Elodie Cassan, Anne-Muriel Arigon-Chifolleau, et al. “Concomitant emergence of the antisense protein gene of HIV-1 and of the pandemic”. In: *Proceedings of the National Academy of Sciences* 113.41 (Oct. 11, 2016). Publisher: National Academy of Sciences Section: Biological Sciences PMID: 27681623, pp. 11537–11542. DOI: [10.1073/pnas.1605739113](https://doi.org/10.1073/pnas.1605739113). URL: <https://www.pnas.org/content/113/41/11537> (cit. on p. 99).
- [375] Juliette Savoret, Nathalie Chazal, et al. “A Pilot Study of the Humoral Response Against the AntiSense Protein (ASP) in HIV-1-Infected Patients”. In: *Frontiers in Microbiology* 11 (2020). URL: [https://www.frontiersin.org/article/10.3389/fmicb.2020.00020](https://doi.org/10.3389/fmicb.2020.00020) (cit. on p. 99).
- [376] Christine Zardecki, Shuchismita Dutta, et al. “PDB-101: Educational resources supporting molecular explorations through biology and medicine”. In: *Protein Science* 31.1 (2022). \_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/pro.4200>, pp. 129–140. DOI: [10.1002/pro.4200](https://doi.org/10.1002/pro.4200). URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/pro.4200> (cit. on p. 123).
- [377] Stefan G. Sarafianos, Bruno Marchand, et al. “Structure and Function of HIV-1 Reverse Transcriptase: Molecular Mechanisms of Polymerization and Inhibition”. In: *Journal of molecular biology* 385.3 (Jan. 2009), pp. 693–713. ISSN: 0022-2836. DOI: [10.1016/j.jmb.2008.10.071](https://doi.org/10.1016/j.jmb.2008.10.071) (cit. on pp. 99, 142).
- [378] David S. Goodsell, Ludovic Autin, and Arthur J. Olson. “Illustrate: Software for Biomolecular Illustration”. In: *Structure (London, England: 1993)* 27.11 (Nov. 5, 2019). PMID: 31519398 PMCID: PMC6834899, 1716–1720.e1. DOI: [10.1016/j.str.2019.08.011](https://doi.org/10.1016/j.str.2019.08.011) (cit. on p. 100).
- [379] Thang K. Chiu and David R. Davies. “Structure and Function of HIV-1 Integrase”. In: *Current Topics in Medicinal Chemistry* 4.9 (May 1, 2004). tex.ids= chiuStructureFunctionHIV1, pp. 965–977. DOI: [10.2174/1568026043388547](https://doi.org/10.2174/1568026043388547) (cit. on p. 100).
- [380] Dominic Esposito and Robert Craigie. “HIV Integrase Structure and Function”. In: ed. by Karl Rlaramorosch, Frederick A. Murphy, and Aaron J. Shawn. Vol. 52. tex.ids= espositoHIVIntegraseStructure1999a DOI: [10.1016/S0065-3527\(08\)60304-8](https://doi.org/10.1016/S0065-3527(08)60304-8). Academic Press, Jan. 1, 1999, pp. 319–333. URL: <https://www.sciencedirect.com/science/article/pii/S0065352708603048> (cit. on p. 100).
- [381] Olivier Delelis, Kevin Carayon, et al. “Integrase and integration: biochemical activities of HIV-1 integrase”. In: *Retrovirology* 5 (Dec. 17, 2008). PMID: 19091057 PMCID: PMC2615046, p. 114. DOI: [10.1186/1742-4690-5-114](https://doi.org/10.1186/1742-4690-5-114). URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2615046/> (cit. on p. 100).
- [382] Laurence H. Pearl and William R. Taylor. “A structural model for the retroviral proteases”. In: *Nature* 329.6137 (Sept. 1987). Number: 6137 Publisher: Nature Publishing Group, pp. 351–354. DOI: [10.1038/329351a0](https://doi.org/10.1038/329351a0). URL: <https://www.nature.com/articles/329351a0> (cit. on p. 100).

## REFERENCES

- [383] Sergei Gulnik, John W. Erickson, and Dong Xie. “HIV protease: Enzyme function and drug resistance”. In: vol. 58. DOI: 10.1016/S0083-6729(00)58026-1. Academic Press, Jan. 1, 2000, pp. 213–256. DOI: [10.1016/S0083-6729\(00\)58026-1](https://doi.org/10.1016/S0083-6729(00)58026-1). URL: <https://www.sciencedirect.com/science/article/pii/S0083672900580261> (cit. on p. 100).
- [384] Abelardo M. Silva, Raúl E. Cachau, et al. “Inhibition and catalytic mechanism of HIV-1 aspartic protease”. In: *Journal of Molecular Biology* 255.2 (Jan. 1, 1996), pp. 321–340. DOI: [10.1006/jmbi.1996.0026](https://doi.org/10.1006/jmbi.1996.0026). URL: <https://www.sciencedirect.com/science/article/pii/S0022283696900261> (cit. on p. 101).
- [385] Viktor Hornak, Asim Okur, et al. “HIV-1 protease flaps spontaneously open and reclose in molecular dynamics simulations”. In: *Proceedings of the National Academy of Sciences* 103.4 (Jan. 24, 2006). Publisher: Proceedings of the National Academy of Sciences, pp. 915–920. DOI: [10.1073/pnas.0508452103](https://doi.org/10.1073/pnas.0508452103). URL: <https://www.pnas.org/doi/10.1073/pnas.0508452103> (cit. on p. 101).
- [386] Yuqi Yu, Jinan Wang, et al. “Structural insights into HIV-1 protease flap opening processes and key intermediates”. In: *RSC Advances* 7.71 (Sept. 15, 2017). Publisher: The Royal Society of Chemistry, pp. 45121–45128. DOI: [10.1039/C7RA09691G](https://doi.org/10.1039/C7RA09691G). URL: <https://pubs.rsc.org/en/content/articlelanding/2017/ra/c7ra09691g> (cit. on p. 101).
- [387] Robert W. Eisinger, Carl W. Dieffenbach, and Anthony S. Fauci. “HIV Viral Load and Transmissibility of HIV Infection: Undetectable Equals Untransmittable”. In: *JAMA* 321.5 (Feb. 5, 2019), pp. 451–452. DOI: [10.1001/jama.2018.21167](https://doi.org/10.1001/jama.2018.21167). URL: <https://doi.org/10.1001/jama.2018.21167> (cit. on p. 101).
- [388] Frank J. Palella, Kathleen M. Delaney, et al. “Declining Morbidity and Mortality among Patients with Advanced Human Immunodeficiency Virus Infection”. In: *New England Journal of Medicine* 338.13 (Mar. 26, 1998). Publisher: Massachusetts Medical Society \_eprint: <https://doi.org/10.1056/NEJM199803263381301>. PMID: 9516219, pp. 853–860. DOI: [10.1056/NEJM199803263381301](https://doi.org/10.1056/NEJM199803263381301). URL: <https://doi.org/10.1056/NEJM199803263381301> (cit. on p. 101).
- [389] Andrew R. Zolopa. “The evolution of HIV treatment guidelines: Current state-of-the-art of ART”. In: *Antiviral Research*. Twenty-five Years of Antiretroviral Drug Development: Progress and Prospects 85.1 (Jan. 1, 2010), pp. 241–244. DOI: [10.1016/j.antiviral.2009.10.018](https://doi.org/10.1016/j.antiviral.2009.10.018). URL: <https://www.sciencedirect.com/science/article/pii/S0166354209005051> (cit. on p. 101).
- [390] Steven S. Forsythe, William McGreevey, et al. “Twenty Years Of Antiretroviral Therapy For People Living With HIV: Global Costs, Health Achievements, Economic Benefits”. In: *Health Affairs* 38.7 (July 2019). Publisher: Health Affairs, pp. 1163–1172. DOI: [10.1377/hlthaff.2018.05391](https://doi.org/10.1377/hlthaff.2018.05391). URL: <https://doi.org/10.1377/hlthaff.2018.05391> (cit. on p. 101).

## REFERENCES

- [391] Margaret A. Fischl, Douglas D. Richman, et al. “The Efficacy of Azidothymidine (AZT) in the Treatment of Patients with AIDS and AIDS-Related Complex”. In: *New England Journal of Medicine* 317.4 (July 23, 1987). Publisher: Massachusetts Medical Society \_eprint: <https://doi.org/10.1056/NEJM198707233170401> PMID: 3299089, pp. 185–191. DOI: [10.1056/NEJM198707233170401](https://doi.org/10.1056/NEJM198707233170401). URL: <https://doi.org/10.1056/NEJM198707233170401> (cit. on p. 101).
- [392] Douglas D. Richman. “Susceptibility to nucleoside analogues of zidovudine-resistant isolates of human immunodeficiency virus”. In: *The American Journal of Medicine*. Dideoxycydine (ddC): A Potent Antiretroviral Agent for Human Immunodeficietion Virus Infection 88.5, Supplement 2 (May 21, 1990), S8–S10. DOI: [10.1016/0002-9343\(90\)90414-9](https://doi.org/10.1016/0002-9343(90)90414-9). URL: <https://www.sciencedirect.com/science/article/pii/0002934390904149> (cit. on p. 101).
- [393] R. M. Gulick, J. W. Mellors, et al. “Treatment with indinavir, zidovudine, and lamivudine in adults with human immunodeficiency virus infection and prior antiretroviral therapy”. In: *The New England Journal of Medicine* 337.11 (Sept. 11, 1997). PMID: 9287228, pp. 734–739. DOI: [10.1056/NEJM199709113371102](https://doi.org/10.1056/NEJM199709113371102) (cit. on p. 101).
- [394] Lucia Palmisano and Stefano Vella. “A brief history of antiretroviral therapy of HIV infection: success and challenges”. In: *Annali dell'Istituto Superiore Di Sanita* 47.1 (2011). PMID: 21430338, pp. 44–48. DOI: [10.4415/ANN\\_11\\_01\\_10](https://doi.org/10.4415/ANN_11_01_10) (cit. on p. 101).
- [395] *One-pill-a-day HIV treatments*. URL: <https://www.aidsmap.com/about-hiv/one-pill-day-hiv-treatments> (cit. on p. 101).
- [396] John C. Tilton and Robert W. Doms. “Entry inhibitors in the treatment of HIV-1 infection”. In: *Antiviral Research*. Twenty-five Years of Antiretroviral Drug Development: Progress and Prospects 85.1 (Jan. 1, 2010), pp. 91–100. DOI: [10.1016/j.antiviral.2009.07.022](https://doi.org/10.1016/j.antiviral.2009.07.022). URL: <https://www.sciencedirect.com/science/article/pii/S0166354209003945> (cit. on p. 101).
- [397] Kimberly K. Scarsi, Joshua P. Havens, et al. “HIV-1 Integrase Inhibitors: A Comparative Review of Efficacy and Safety”. In: *Drugs* 80.16 (Nov. 1, 2020), pp. 1649–1676. DOI: [10.1007/s40265-020-01379-9](https://doi.org/10.1007/s40265-020-01379-9). URL: <https://doi.org/10.1007/s40265-020-01379-9> (cit. on p. 101).
- [398] Robert M. Grant, Javier R. Lama, et al. “Preexposure Chemoprophylaxis for HIV Prevention in Men Who Have Sex with Men”. In: *New England Journal of Medicine* 363.27 (Dec. 30, 2010). Publisher: Massachusetts Medical Society \_eprint: <https://doi.org/10.1056/NEJMoa1011205> PMID: 21091279, pp. 2587–2599. DOI: [10.1056/NEJMoa1011205](https://doi.org/10.1056/NEJMoa1011205). URL: <https://doi.org/10.1056/NEJMoa1011205> (cit. on p. 101).

## REFERENCES

- [399] Jared M. Baeten, Deborah Donnell, et al. “Antiretroviral prophylaxis for HIV prevention in heterosexual men and women”. In: *The New England Journal of Medicine* 367.5 (Aug. 2, 2012). PMID: 22784037 PMCID: PMC3770474, pp. 399–410. DOI: [10.1056/NEJMoa1108524](https://doi.org/10.1056/NEJMoa1108524) (cit. on p. 101).
- [400] Susan P. Buchbinder and Albert Liu. “Pre-exposure prophylaxis and the promise of combination prevention approaches”. In: *AIDS and behavior* 15 Suppl 1 (Apr. 2011). PMID: 21331801 PMCID: PMC3064892, S72–79. DOI: [10.1007/s10461-011-9894-1](https://doi.org/10.1007/s10461-011-9894-1) (cit. on p. 101).
- [401] James, IV Riddell, K. Rivet Amico, and Kenneth H. Mayer. “HIV Preexposure Prophylaxis: A Review”. In: *JAMA* 319.12 (Mar. 27, 2018), pp. 1261–1268. DOI: [10.1001/jama.2018.1917](https://doi.org/10.1001/jama.2018.1917). URL: <https://doi.org/10.1001/jama.2018.1917> (cit. on p. 101).
- [402] EMA. *Truvada*. Sept. 17, 2018. URL: <https://www.ema.europa.eu/en/medicines/human/EPAR/truvada> (cit. on p. 101).
- [403] R. M. Esnouf, J. Ren, et al. “Unique features in the structure of the complex between HIV-1 reverse transcriptase and the bis(heteroaryl)piperazine (BHAP) U-90152 explain resistance mutations for this nonnucleoside inhibitor”. In: *Proceedings of the National Academy of Sciences of the United States of America* 94.8 (Apr. 15, 1997). PMID: 9108091 PMCID: PMC20554, pp. 3984–3989. DOI: [10.1073/pnas.94.8.3984](https://doi.org/10.1073/pnas.94.8.3984) (cit. on p. 102).
- [404] Noel A. Roberts, Joseph A. Martin, et al. “Rational Design of Peptide-Based HIV Proteinase Inhibitors”. In: *Science* 248.4953 (Apr. 20, 1990). Publisher: American Association for the Advancement of Science, pp. 358–361. DOI: [10.1126/science.2183354](https://doi.org/10.1126/science.2183354). URL: <https://www.science.org/doi/10.1126/science.2183354> (cit. on p. 102).
- [405] José A Esté and Amilio Telenti. “HIV entry inhibitors”. In: *The Lancet* 370.9581 (July 7, 2007), pp. 81–88. DOI: [10.1016/S0140-6736\(07\)61052-6](https://doi.org/10.1016/S0140-6736(07)61052-6). URL: <https://www.sciencedirect.com/science/article/pii/S0140673607610526> (cit. on p. 102).
- [406] J. Michael Kilby and Joseph J. Eron. “Novel Therapies Based on Mechanisms of HIV-1 Cell Entry”. In: *New England Journal of Medicine* 348.22 (May 29, 2003). Publisher: Massachusetts Medical Society, pp. 2228–2238. DOI: [10.1056/NEJMra022812](https://doi.org/10.1056/NEJMra022812). URL: <https://www.nejm.org/doi/10.1056/NEJMra022812> (cit. on p. 102).
- [407] Courtney V Fletcher. “Enfuvirtide, a new drug for HIV infection”. In: *The Lancet* 361.9369 (May 10, 2003), pp. 1577–1578. DOI: [10.1016/S0140-6736\(03\)13323-5](https://doi.org/10.1016/S0140-6736(03)13323-5). URL: <https://www.sciencedirect.com/science/article/pii/S0140673603133235> (cit. on p. 102).
- [408] Yves Pommier, Allison A. Johnson, and Christophe Marchand. “Integrase inhibitors to treat HIV/Aids”. In: *Nature Reviews Drug Discovery* 4.3 (Mar. 2005). Number: 3 Publisher: Nature Publishing Group, pp. 236–248. DOI: [10.1038/nrd1660](https://doi.org/10.1038/nrd1660). URL: <https://www.nature.com/articles/nrd1660> (cit. on p. 102).

- [409] David C. Knox, Peter L. Anderson, et al. “Multidrug-Resistant HIV-1 Infection despite Preexposure Prophylaxis”. In: *New England Journal of Medicine* 376.5 (Feb. 2, 2017). Publisher: Massachusetts Medical Society, pp. 501–502. DOI: [10.1056/NEJMc1611639](https://doi.org/10.1056/NEJMc1611639). URL: <https://www.nejm.org/doi/10.1056/NEJMc1611639> (cit. on p. 102).
- [410] Christopher B. Hurt, Joseph J. Eron, and Myron S. Cohen. “Pre-exposure prophylaxis and antiretroviral resistance: HIV prevention at a cost?” In: *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America* 53.12 (Dec. 2011). PMID: 21976467 PMCID: PMC3246871, pp. 1265–1270. DOI: [10.1093/cid/cir684](https://doi.org/10.1093/cid/cir684) (cit. on p. 102).
- [411] Kevin M. Gibas, Polly van den Berg, et al. “Drug Resistance During HIV Pre-Exposure Prophylaxis”. In: *Drugs* 79.6 (Apr. 1, 2019), pp. 609–619. DOI: [10.1007/s40265-019-01108-x](https://doi.org/10.1007/s40265-019-01108-x). URL: <https://doi.org/10.1007/s40265-019-01108-x> (cit. on p. 102).
- [412] Palanee Ammaranond and Sayompoo Sanguansittanan. “Mechanism of HIV antiretroviral drugs progress toward drug resistance”. In: *Fundamental & Clinical Pharmacology* 26.1 (2012). \_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1472-8206.2011.01009.x>, pp. 146–161. DOI: [10.1111/j.1472-8206.2011.01009.x](https://doi.org/10.1111/j.1472-8206.2011.01009.x). URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1472-8206.2011.01009.x> (cit. on p. 102).
- [413] François Clavel and Allan J. Hance. “HIV Drug Resistance”. In: *New England Journal of Medicine* 350.10 (Mar. 4, 2004). Publisher: Massachusetts Medical Society \_eprint: <https://www.nejm.org/doi/pdf/10.1056/NEJMra025195> PMID: 14999114, pp. 1023–1035. DOI: [10.1056/NEJMra025195](https://doi.org/10.1056/NEJMra025195). URL: <https://www.nejm.org/doi/full/10.1056/NEJMra025195> (cit. on p. 102).
- [414] Luis Menéndez-Arias. “Mechanisms of resistance to nucleoside analogue inhibitors of HIV-1 reverse transcriptase”. In: *Virus Research*. Retroviral Reverse Transcription 134.1 (June 1, 2008), pp. 124–146. DOI: [10.1016/j.virusres.2007.12.015](https://doi.org/10.1016/j.virusres.2007.12.015). URL: <https://www.sciencedirect.com/science/article/pii/S0168170207004686> (cit. on p. 102).
- [415] N. Sluis-Cremer, D. Arion, and M. A. Parniak\*. “Molecular mechanisms of HIV-1 resistance to nucleoside reverse transcriptase inhibitors (NRTIs)”. In: *Cellular and Molecular Life Sciences CMS* 57.10 (Sept. 1, 2000), pp. 1408–1422. DOI: [10.1007/PL00000626](https://doi.org/10.1007/PL00000626). URL: <https://doi.org/10.1007/PL00000626> (cit. on p. 102).
- [416] S. G. Sarafianos, K. Das, et al. “Lamivudine (3TC) resistance in HIV-1 reverse transcriptase involves steric hindrance with beta-branched amino acids”. In: *Proceedings of the National Academy of Sciences of the United States of America* 96.18 (Aug. 31, 1999). PMID: 10468556 PMCID: PMC17836, pp. 10027–10032. DOI: [10.1073/pnas.96.18.10027](https://doi.org/10.1073/pnas.96.18.10027) (cit. on p. 102).

## REFERENCES

- [417] P. R. Meyer, S. E. Matsuura, et al. “A mechanism of AZT resistance: an increase in nucleotide-dependent primer unblocking by mutant HIV-1 reverse transcriptase”. In: *Molecular Cell* 4.1 (July 1999). PMID: 10445025, pp. 35–43. DOI: [10.1016/s1097-2765\(00\)80185-9](https://doi.org/10.1016/s1097-2765(00)80185-9) (cit. on p. 102).
- [418] Paul L. Boyer, Stefan G. Sarafianos, et al. “Selective Excision of AZTMP by Drug-Resistant Human Immunodeficiency Virus Reverse Transcriptase”. In: *Journal of Virology* 75.10 (May 2001). PMID: 11312355 PMCID: PMC114238, pp. 4832–4842. DOI: [10.1128/JVI.75.10.4832-4842.2001](https://doi.org/10.1128/JVI.75.10.4832-4842.2001). URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC114238/> (cit. on p. 102).
- [419] Steven G. Deeks. “Nonnucleoside Reverse Transcriptase Inhibitor Resistance”. In: *JAIDS Journal of Acquired Immune Deficiency Syndromes* 26 (Mar. 1, 2001), S25. URL: [https://journals.lww.com/jaids/Abstract/2001/03011/Nonnucleoside\\_Reverse\\_Transcriptase\\_Inhibitor.4.aspx](https://journals.lww.com/jaids/Abstract/2001/03011/Nonnucleoside_Reverse_Transcriptase_Inhibitor.4.aspx) (cit. on p. 102).
- [420] Annemarie M. J. Wensing, Noortje M. van Maarseveen, and Monique Nijhuis. “Fifteen years of HIV Protease Inhibitors: raising the barrier to resistance”. In: *Antiviral Research*. Twenty-five Years of Antiretroviral Drug Development: Progress and Prospects 85.1 (Jan. 1, 2010), pp. 59–74. DOI: [10.1016/j.antiviral.2009.10.003](https://doi.org/10.1016/j.antiviral.2009.10.003). URL: <https://www.sciencedirect.com/science/article/pii/S0166354209004902> (cit. on p. 102).
- [421] Nese Kurt Yilmaz, Ronald Swanstrom, and Celia A. Schiffer. “Improving Viral Protease Inhibitors to Counter Drug Resistance”. In: *Trends in Microbiology* 24.7 (July 1, 2016), pp. 547–557. DOI: [10.1016/j.tim.2016.03.010](https://doi.org/10.1016/j.tim.2016.03.010). URL: <https://www.sciencedirect.com/science/article/pii/S0966842X16300026> (cit. on p. 102).
- [422] Jose-Luis Blanco, Vici Varghese, et al. “HIV-1 Integrase Inhibitor Resistance and Its Clinical Implications”. In: *The Journal of Infectious Diseases* 203.9 (May 1, 2011), pp. 1204–1214. DOI: [10.1093/infdis/jir025](https://doi.org/10.1093/infdis/jir025). URL: <https://doi.org/10.1093/infdis/jir025> (cit. on p. 102).
- [423] Raphaël Mourad, François Chevennet, et al. “A Phylotype-Based Analysis Highlights the Role of Drug-Naive HIV-Positive Individuals in the Transmission of Antiretroviral Resistance in the UK”. ENGLISH. In: *Aids* 29.15 (Sept. 2015), pp. 1917–1925. ISSN: 0269-9370. DOI: [10.1097/QAD.0000000000000768](https://doi.org/10.1097/QAD.0000000000000768) (cit. on pp. 102, 127).
- [424] Sara M. Drescher, Viktor von Wyl, et al. “Treatment-Naive Individuals Are the Major Source of Transmitted HIV-1 Drug Resistance in Men Who Have Sex With Men in the Swiss HIV Cohort Study”. In: *Clinical Infectious Diseases* 58.2 (Jan. 15, 2014), pp. 285–294. DOI: [10.1093/cid/cit694](https://doi.org/10.1093/cid/cit694). URL: <https://doi.org/10.1093/cid/cit694> (cit. on p. 102).
- [425] Ragna S Boerma, T Sonia Boender, et al. “High levels of pre-treatment HIV drug resistance and treatment failure in Nigerian children”. In: *Journal of the International AIDS Society* 19.1 (2016). \_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.7448/IAS.19.1.7448>

- p. 21140. DOI: [10.7448/IAS.19.1.21140](https://doi.org/10.7448/IAS.19.1.21140). URL: <https://onlinelibrary.wiley.com/doi/abs/10.7448/IAS.19.1.21140> (cit. on p. 102).
- [426] Denise Kühnert, Roger Kouyos, et al. “Quantifying the fitness cost of HIV-1 drug resistance mutations through phylodynamics”. In: *PLOS Pathogens* 14.2 (Feb. 20, 2018). Publisher: Public Library of Science, e1006895. DOI: [10.1371/journal.ppat.1006895](https://doi.org/10.1371/journal.ppat.1006895). URL: <https://journals.plos.org/plospathogens/article?id=10.1371/journal.ppat.1006895> (cit. on p. 102).
- [427] Thibault Mesplède, Peter K. Quashie, et al. “Viral fitness cost prevents HIV-1 from evading dolutegravir drug pressure”. In: *Retrovirology* 10.1 (Feb. 22, 2013), p. 22. DOI: [10.1186/1742-4690-10-22](https://doi.org/10.1186/1742-4690-10-22). URL: <https://doi.org/10.1186/1742-4690-10-22> (cit. on p. 102).
- [428] Hannah Castro, Deenan Pillay, et al. “Persistence of HIV-1 Transmitted Drug Resistance Mutations”. In: *The Journal of Infectious Diseases* 208.9 (Nov. 1, 2013), pp. 1459–1463. DOI: [10.1093/infdis/jit345](https://doi.org/10.1093/infdis/jit345). URL: <https://academic.oup.com/jid/article/208/9/1459/2192846> (cit. on p. 103).
- [429] Luc Bassel, Anna Zhukova, et al. “Drug resistance mutations in HIV: new bioinformatics approaches and challenges”. In: *Current Opinion in Virology* 51 (Dec. 1, 2021), pp. 56–64. DOI: [10.1016/j.coviro.2021.09.009](https://doi.org/10.1016/j.coviro.2021.09.009). URL: <https://www.sciencedirect.com/science/article/pii/S1879625721001073> (cit. on p. 103).
- [430] UK CHIC Steering Committee. “The creation of a large UK-based multicentre cohort of HIV-infected individuals: The UK Collaborative HIV Cohort (UK CHIC) Study”. In: *HIV Medicine* 5.2 (2004), pp. 115–124. DOI: [10.1111/j.1468-1293.2004.00197.x](https://doi.org/10.1111/j.1468-1293.2004.00197.x). URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1468-1293.2004.00197.x> (cit. on p. 103).
- [431] Lucie Abeler-Dörner, Mary K. Grabowski, et al. “PANGEA-HIV 2: Phylogenetics And Networks for Generalised Epidemics in Africa”. In: *Current Opinion in HIV and AIDS* 14.3 (May 2019), 173–180. DOI: [10.1097/COH.0000000000000542](https://doi.org/10.1097/COH.0000000000000542). URL: <https://doi.org/10.1097/COH.0000000000000542> (cit. on p. 103).
- [432] Robert W. Shafer. “Rationale and Uses of a Public HIV Drug-Resistance Database”. In: *The Journal of Infectious Diseases* 194.Supplement\_1 (Sept. 1, 2006), S51–S58. DOI: [10.1086/505356](https://doi.org/10.1086/505356). URL: [https://academic.oup.com/jid/article/194/Supplement\\_1/S51/894267](https://academic.oup.com/jid/article/194/Supplement_1/S51/894267) (cit. on p. 103).
- [433] Annemarie M. Wensing, Vincent Calvez, et al. “2019 update of the drug resistance mutations in HIV-1”. In: *Topics in Antiviral Medicine* 27.3 (Sept. 2019). PMID: 31634862 PMCID: PMC6892618, pp. 111–121. URL: <https://pubmed.ncbi.nlm.nih.gov/31634862/> (cit. on p. 103).
- [434] Shauna A Clark, Charles Calef, and John W Mellors. “Mutations in retroviral genes associated with drug resistance”. In: *HIV sequence compendium* (2007). tex.ids= clarkMutationsRetroviralGenes publisher: Citeseer, 58–158 (cit. on p. 103).

## REFERENCES

- [435] Tommy F. Liu and Robert W. Shafer. “Web Resources for HIV Type 1 Genotypic-Resistance Test Interpretation”. In: *Clinical Infectious Diseases* 42.11 (June 2006), pp. 1608–1618. ISSN: 1058-4838. DOI: [10.1086/503914](https://doi.org/10.1086/503914) (cit. on p. 103).
- [436] Alessandro Cozzi Lepri, Caroline A. Sabin, et al. “Resistance Profiles in Patients with Viral Rebound on Potent Antiretroviral Therapy”. en. In: *The Journal of Infectious Diseases* 181.3 (Mar. 2000), pp. 1143–1147. ISSN: 0022-1899. DOI: [10.1086/315301](https://doi.org/10.1086/315301) (cit. on p. 127).
- [437] Chris Verhofstede, Filip Van Wanzele, et al. “Detection of Drug Resistance Mutations as a Predictor of Subsequent Virological Failure in Patients with HIV-1 Viral Rebounds of Less than 1,000 RNA Copies/Ml”. en. In: *Journal of Medical Virology* 79.9 (2007), pp. 1254–1260. ISSN: 1096-9071. DOI: [10.1002/jmv.20950](https://doi.org/10.1002/jmv.20950) (cit. on p. 127).
- [438] Stéphane Hué, Robert J. Gifford, et al. “Demonstration of Sustained Drug-Resistant Human Immunodeficiency Virus Type 1 Lineages Circulating among Treatment-Naïve Individuals”. en. In: *Journal of Virology* 83.6 (Mar. 2009), pp. 2645–2654. ISSN: 0022-538X, 1098-5514. DOI: [10.1128/JVI.01556-08](https://doi.org/10.1128/JVI.01556-08) (cit. on p. 127).
- [439] Anna Zhukova, Teresa Cutino-Moguel, et al. “The Role of Phylogenetics as a Tool to Predict the Spread of Resistance”. en. In: *The Journal of Infectious Diseases* 216.suppl\_9 (Dec. 2017), S820–S823. ISSN: 0022-1899. DOI: [10.1093/infdis/jix411](https://doi.org/10.1093/infdis/jix411) (cit. on p. 127).
- [440] Diane E. Bennett, Ricardo J. Camacho, et al. “Drug Resistance Mutations for Surveillance of Transmitted HIV-1 Drug-Resistance: 2009 Update”. en. In: *PLOS ONE* 4.3 (Mar. 2009), e4724. ISSN: 1932-6203. DOI: [10.1371/journal.pone.0004724](https://doi.org/10.1371/journal.pone.0004724) (cit. on p. 127).
- [441] Jennifer Hammond, Charles Calef, et al. “Mutations in Retroviral Genes Associated with Drug Resistance”. en. In: *Human retroviruses and AIDS* (Dec. 1998), pp. 11136–11179 (cit. on p. 127).
- [442] A. M. Wensing, V. Calvez, et al. “2017 Update of the Drug Resistance Mutations in HIV-1., 2017 Update of the Drug Resistance Mutations in HIV-1”. eng. In: *Topics in antiviral medicine, Topics in Antiviral Medicine* 24, 24.4, 4 (Dec. 2016), pp. 132, 132–133. ISSN: 2161-5861 (cit. on p. 127).
- [443] Sandrine Dudoit and Mark J. van der Laan. *Multiple Testing Procedures with Applications to Genomics*. en. Springer Science & Business Media, Dec. 2007. ISBN: 978-0-387-49317-6 (cit. on p. 127).
- [444] Christian Julian Villabona-Arenas, Nicole Vidal, et al. “In-Depth Analysis of HIV-1 Drug Resistance Mutations in HIV-Infected Individuals Failing First-Line Regimens in West and Central Africa”. en-US. In: *AIDS* 30.17 (Nov. 2016), p. 2577. ISSN: 0269-9370. DOI: [10.1097/QAD.0000000000001233](https://doi.org/10.1097/QAD.0000000000001233) (cit. on pp. 127–130, 132, 133, 137).

- [445] Wayne P. Maddison and Richard G. FitzJohn. “The Unsolved Challenge to Phylogenetic Correlation Tests for Categorical Characters”. en. In: *Systematic Biology* 64.1 (Jan. 2015), pp. 127–136. ISSN: 1063-5157. DOI: [10.1093/sysbio/syu070](https://doi.org/10.1093/sysbio/syu070) (cit. on p. 128).
- [446] Pak C. Sham and Shaun M. Purcell. “Statistical Power and Significance Testing in Large-Scale Genetic Studies”. en. In: *Nature Reviews Genetics* 15.5 (May 2014), pp. 335–346. ISSN: 1471-0064. DOI: [10.1038/nrg3706](https://doi.org/10.1038/nrg3706) (cit. on p. 128).
- [447] Thomas Lengauer and Tobias Sing. “Bioinformatics-Assisted Anti-HIV Therapy”. en. In: *Nature Reviews Microbiology* 4.10 (Oct. 2006), pp. 790–797. ISSN: 1740-1534. DOI: [10.1038/nrmicro1477](https://doi.org/10.1038/nrmicro1477) (cit. on p. 128).
- [448] Jie Zhang, Soo-Yon Rhee, et al. “Comparison of the Precision and Sensitivity of the Antivirogram and PhenoSense HIV Drug Susceptibility Assays”. en-US. In: *JAIDS Journal of Acquired Immune Deficiency Syndromes* 38.4 (Apr. 2005), pp. 439–444. ISSN: 1525-4135. DOI: [10.1097/01.qai.0000147526.64863.53](https://doi.org/10.1097/01.qai.0000147526.64863.53) (cit. on p. 128).
- [449] Niko Beerenwinkel, Martin Däumer, et al. “Geno2pheno: Estimating Phenotypic Drug Resistance from HIV-1 Genotypes”. In: *Nucleic Acids Research* 31.13 (July 2003), pp. 3850–3855. ISSN: 0305-1048. DOI: [10.1093/nar/gkg575](https://doi.org/10.1093/nar/gkg575) (cit. on p. 128).
- [450] ChenHsiang Shen, Xiaxia Yu, et al. “Automated Prediction of HIV Drug Resistance from Genotype Data”. In: *BMC Bioinformatics* 17.8 (Aug. 2016), p. 278. ISSN: 1471-2105. DOI: [10.1186/s12859-016-1114-6](https://doi.org/10.1186/s12859-016-1114-6) (cit. on p. 128).
- [451] Xiaxia Yu, Irene T. Weber, and Robert W. Harrison. “Prediction of HIV Drug Resistance from Genotype with Encoded Three-Dimensional Protein Structure”. In: *BMC Genomics* 15.5 (July 2014), S1. ISSN: 1471-2164. DOI: [10.1186/1471-2164-15-S5-S1](https://doi.org/10.1186/1471-2164-15-S5-S1) (cit. on p. 128).
- [452] Olivier Sheik Amamuddy, Nigel T. Bishop, and Özlem Tastan Bishop. “Improving Fold Resistance Prediction of HIV-1 against Protease and Reverse Transcriptase Inhibitors Using Artificial Neural Networks”. In: *BMC Bioinformatics* 18.1 (Aug. 2017), p. 369. ISSN: 1471-2105. DOI: [10.1186/s12859-017-1782-x](https://doi.org/10.1186/s12859-017-1782-x) (cit. on p. 128).
- [453] N. Beerenwinkel, T. Lengauer, et al. “Geno2pheno: Interpreting Genotypic HIV Drug Resistance Tests”. In: *IEEE Intelligent Systems* 16.6 (Nov. 2001), pp. 35–41. ISSN: 1941-1294. DOI: [10.1109/5254.972080](https://doi.org/10.1109/5254.972080) (cit. on p. 128).
- [454] Seare Tesfamichael Araya and Scott Hazelhurst. “Support Vector Machine Prediction of HIV-1 Drug Resistance Using the Viral Nucleotide Patterns”. In: *Transactions of the Royal Society of South Africa* 64.1 (Jan. 2009), pp. 62–72. ISSN: 0035-919X. DOI: [10.1080/00359190909519238](https://doi.org/10.1080/00359190909519238) (cit. on p. 128).
- [455] Mona Riemenschneider, Robin Senge, et al. “Exploiting HIV-1 Protease and Reverse Transcriptase Cross-Resistance Information for Improved Drug Resistance Prediction by Means of Multi-Label Classification”. In: *BioData Mining* 9.1 (Feb. 2016), p. 10. ISSN: 1756-0381. DOI: [10.1186/s13040-016-0089-1](https://doi.org/10.1186/s13040-016-0089-1) (cit. on p. 128).

## REFERENCES

- [456] Dominik Heider, Robin Senge, et al. “Multilabel Classification for Exploiting Cross-Resistance Information in HIV-1 Drug Resistance Prediction”. In: *Bioinformatics* 29.16 (Aug. 2013), pp. 1946–1952. ISSN: 1367-4803. DOI: [10.1093/bioinformatics/btt331](https://doi.org/10.1093/bioinformatics/btt331) (cit. on p. 128).
- [457] Sorin Drăghici and R. Brian Potter. “Predicting HIV Drug Resistance with Neural Networks”. In: *Bioinformatics* 19.1 (Jan. 2003), pp. 98–107. ISSN: 1367-4803. DOI: [10.1093/bioinformatics/19.1.98](https://doi.org/10.1093/bioinformatics/19.1.98) (cit. on p. 128).
- [458] Margaret C. Steiner, Keylie M. Gibson, and Keith A. Crandall. “Drug Resistance Prediction Using Deep Learning Techniques on HIV-1 Sequence Data”. en. In: *Viruses* 12.5 (May 2020), p. 560. DOI: [10.3390/v12050560](https://doi.org/10.3390/v12050560) (cit. on p. 128).
- [459] Alyssa C. Mooney, Chadwick K. Campbell, et al. “Beyond Social Desirability Bias: Investigating Inconsistencies in Self-Reported HIV Testing and Treatment Behaviors Among HIV-Positive Adults in North West Province, South Africa”. en. In: *AIDS and Behavior* 22.7 (July 2018), pp. 2368–2379. ISSN: 1573-3254. DOI: [10.1007/s10461-018-2155-9](https://doi.org/10.1007/s10461-018-2155-9) (cit. on p. 130).
- [460] Robert Tibshirani. “Regression Shrinkage and Selection Via the Lasso”. en. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1 (1996), pp. 267–288. ISSN: 2517-6161. DOI: [10.1111/j.2517-6161.1996.tb02080.x](https://doi.org/10.1111/j.2517-6161.1996.tb02080.x) (cit. on pp. 130, 132).
- [461] Glenn W. Brier. “Verification of Forecasts Expressed in Terms of Probability”. en. In: *Monthly Weather Review* 78.1 (Jan. 1950), pp. 1–3. ISSN: 0027-0644 (cit. on pp. 130, 134).
- [462] Olivier Gascuel, Bernadette Bouchon-Meunier, et al. “Twelve Numerical, Symbolic and Hybrid Supervised Classification Methods”. In: *International Journal of Pattern Recognition and Artificial Intelligence* 12.05 (Aug. 1998), pp. 517–571. ISSN: 0218-0014. DOI: [10.1142/S0218001498000336](https://doi.org/10.1142/S0218001498000336) (cit. on pp. 130, 133).
- [463] Jelle J. Goeman and Aldo Solari. “Multiple Hypothesis Testing in Genomics”. en. In: *Statistics in Medicine* 33.11 (2014), pp. 1946–1978. ISSN: 1097-0258. DOI: [10.1002/sim.6082](https://doi.org/10.1002/sim.6082) (cit. on p. 132).
- [464] Jason D Rennie, Lawrence Shih, et al. “Tackling the Poor Assumptions of Naive Bayes Text Classifiers”. en. In: *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*. 2003, pp. 616–623 (cit. on p. 132).
- [465] Leo Breiman. “Random Forests”. en. In: *Machine Learning* 45.1 (Oct. 2001), pp. 5–32. ISSN: 1573-0565. DOI: [10.1023/A:1010933404324](https://doi.org/10.1023/A:1010933404324) (cit. on p. 132).
- [466] David Alvarez Melis and Tommi Jaakkola. “Towards Robust Interpretability with Self-Explaining Neural Networks”. In: *Advances in Neural Information Processing Systems 31*. Ed. by S. Bengio, H. Wallach, et al. Curran Associates, Inc., 2018, pp. 7775–7784 (cit. on p. 133).

- [467] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. en. Springer Science & Business Media, Aug. 2009. ISBN: 978-0-387-84858-7 (cit. on p. 133).
- [468] Quanshi Zhang, Ying Nian Wu, and Song-Chun Zhu. “Interpretable Convolutional Neural Networks”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 8827–8836 (cit. on p. 133).
- [469] Kay Henning Brodersen, Cheng Soon Ong, et al. “The Balanced Accuracy and Its Posterior Distribution”. In: *2010 20th International Conference on Pattern Recognition*. Aug. 2010, pp. 3121–3124. doi: [10.1109/ICPR.2010.764](https://doi.org/10.1109/ICPR.2010.764) (cit. on pp. 134, 170).
- [470] Schrödinger, LLC. “The PyMOL Molecular Graphics System, Version 1.8”. Nov. 2015 (cit. on p. 142).
- [471] Soo-Yon Rhee, Tommy F. Liu, et al. “HIV-1 Subtype B Protease and Reverse Transcriptase Amino Acid Covariation”. en. In: *PLOS Computational Biology* 3.5 (May 2007), e87. ISSN: 1553-7358. doi: [10.1371/journal.pcbi.0030087](https://doi.org/10.1371/journal.pcbi.0030087) (cit. on p. 145).
- [472] Andrea De Luca, Simona Di Giambenedetto, et al. “Improved Interpretation of Genotypic Changes in the HIV-1 Reverse Transcriptase Coding Region That Determine the Virological Response to Didanosine”. en. In: *The Journal of Infectious Diseases* 196.11 (Dec. 2007), pp. 1645–1653. ISSN: 0022-1899. doi: [10.1086/522231](https://doi.org/10.1086/522231) (cit. on p. 145).
- [473] Anne-Genevieve Marcelin, Philippe Flandre, et al. “Impact of HIV-1 Reverse Transcriptase Polymorphism at Codons 211 and 228 on Virological Response to Didanosine”. en. In: *Antiviral Therapy* (2006), p. 8 (cit. on p. 145).
- [474] Andrew J. Leigh Brown, Heather M. Precious, et al. “Reduced Susceptibility of Human Immunodeficiency Virus Type 1 (HIV-1) from Patients with Primary HIV Infection to Nonnucleoside Reverse Transcriptase Inhibitors Is Associated with Variation at Novel Amino Acid Sites”. In: *Journal of Virology* 74.22 (Nov. 2000), pp. 10269–10273. ISSN: 0022-538X (cit. on p. 145).
- [475] Shauna A. Clark, Nancy S. Shulman, et al. “Reverse Transcriptase Mutations 118I, 208Y, and 215Y Cause HIV-1 Hypersusceptibility to Non-Nucleoside Reverse Transcriptase Inhibitors”. en-US. In: *AIDS* 20.7 (Apr. 2006), pp. 981–984. ISSN: 0269-9370. doi: [10.1097/01.aids.0000222069.14878.44](https://doi.org/10.1097/01.aids.0000222069.14878.44) (cit. on p. 145).
- [476] G. Nebbia, Caroline A. Sabin, et al. “Emergence of the H208Y Mutation in the Reverse Transcriptase (RT) of HIV-1 in Association with Nucleoside RT Inhibitor Therapy”. en. In: *Journal of Antimicrobial Chemotherapy* 59.5 (May 2007), pp. 1013–1016. ISSN: 0305-7453. doi: [10.1093/jac/dkm067](https://doi.org/10.1093/jac/dkm067) (cit. on p. 145).

## REFERENCES

- [477] A. Saracino, L. Monno, et al. “Impact of Unreported HIV-1 Reverse Transcriptase Mutations on Phenotypic Resistance to Nucleoside and Non-Nucleoside Inhibitors”. en. In: *Journal of Medical Virology* 78.1 (2006), pp. 9–17. ISSN: 1096-9071. DOI: [10.1002/jmv.20500](https://doi.org/10.1002/jmv.20500) (cit. on p. 145).
- [478] Tong Tong Wu, Yi Fang Chen, et al. “Genome-Wide Association Analysis by Lasso Penalized Logistic Regression”. en. In: *Bioinformatics* 25.6 (Mar. 2009), pp. 714–721. ISSN: 1460-2059, 1367-4803. DOI: [10.1093/bioinformatics/btp041](https://doi.org/10.1093/bioinformatics/btp041) (cit. on p. 146).
- [479] Andrey V. Bzikadze and Pavel A. Pevzner. “Automated Assembly of Centromeres from Ultra-Long Error-Prone Reads”. In: *Nature Biotechnology* 38.11 (11 Nov. 2020), pp. 1309–1316. ISSN: 1546-1696. DOI: [10.1038/s41587-020-0582-4](https://doi.org/10.1038/s41587-020-0582-4) (cit. on p. 157).
- [480] J. Castresana. “Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis”. en. In: *Molecular Biology and Evolution* 17.4 (Apr. 2000), pp. 540–552. ISSN: 0737-4038. DOI: [10.1093/oxfordjournals.molbev.a026334](https://doi.org/10.1093/oxfordjournals.molbev.a026334) (cit. on p. 170).
- [481] Will McGinnis, Hbghhy, et al. *Scikit-Learn-Contrib/Categorical-Encoding: Release For Zenodo*. Zenodo. Jan. 2018. DOI: [10.5281/ZENODO.1157110](https://doi.org/10.5281/ZENODO.1157110) (cit. on p. 170).
- [482] Fabian Pedregosa, Gaël Varoquaux, et al. “Scikit-Learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12.Oct (2011), pp. 2825–2830. ISSN: ISSN 1533-7928 (cit. on pp. 170, 171).
- [483] Pauli Virtanen, Ralf Gommers, et al. “SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python”. en. In: *Nature Methods* 17.3 (Mar. 2020), pp. 261–272. ISSN: 1548-7105. DOI: [10.1038/s41592-019-0686-2](https://doi.org/10.1038/s41592-019-0686-2) (cit. on p. 170).
- [484] Skipper Seabold and Josef Perktold. “Statsmodels: Econometric and Statistical Modeling with Python”. en. In: *Proceedings of the 9th Python in Science Conference*. Ed. by Stéfan van der Walt and Jarrod Millman. 2010, pp. 92–96. DOI: [10.25080/Majora-92bf1922-011](https://doi.org/10.25080/Majora-92bf1922-011) (cit. on p. 170).
- [485] N. X. Vinh and J. Epps. “A Novel Approach for Automatic Number of Clusters Detection in Microarray Data Based on Consensus Clustering”. In: *2009 Ninth IEEE International Conference on Bioinformatics and BioEngineering*. June 2009, pp. 84–91. DOI: [10.1109/BIBE.2009.19](https://doi.org/10.1109/BIBE.2009.19) (cit. on p. 171).
- [486] Peter Harremoes. “Mutual Information of Contingency Tables and Related Inequalities”. In: *2014 IEEE International Symposium on Information Theory*. Honolulu, HI, USA: IEEE, June 2014, pp. 2474–2478. ISBN: 978-1-4799-5186-4. DOI: [10.1109/ISIT.2014.6875279](https://doi.org/10.1109/ISIT.2014.6875279) (cit. on p. 171).
- [487] Shuichi Kawashima, Piotr Pokarowski, et al. “AAindex: amino acid index database, progress report 2008”. In: *Nucleic Acids Research* 36.suppl\_1 (Jan. 1, 2008), pp. D202–D205. DOI: [10.1093/nar/gkm998](https://doi.org/10.1093/nar/gkm998). URL: [https://academic.oup.com/nar/article/36/suppl\\_1/D202/2508449](https://academic.oup.com/nar/article/36/suppl_1/D202/2508449) (cit. on p. 177).



## REFERENCES

### Abstract

This will be the abstract of my PhD. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

### Résumé

Ceci sera le résumé de la thèse. Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetur.