SORBONNE UNIVERSITÉ

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 Blassel

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

Soutenue le 2022-06-16

devant le jury composé de:

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

Aknowledgments

Here will go my aknowledgments 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, consectetuer at, consectetuer 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, consectetuer 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 consectetuer. 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

List of Figures

List of Tables

General Introduction

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

Organization of this manuscript

- Organisation of the manuscript

Publications produced

- **Blassel, Luc**, Paul Medvedev and Rayan Chikhi. 2022. **"Mapping-friendly sequence reductions: going beyond homopolymer compression"**. *iScience* DOI goes here *(Included in Chapter 3)*
- **Blassel, Luc**[1], Anna Zhukova[1], 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
- **Blassel, 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. *(Included in Chapter 6)*
- Zhukova, Anna, **Luc Blassel**, 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.
- Lemoine, Frédéric, **Luc Blassel**, 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.

[1]co-first authors: Luc Blassel and Anna Zhukova

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 biology and more particularly genetics are needed. If you are already familiar with biological sequences, feel free to skip ahead to section 1.2.

1.1.1. What is DNA ?

Desoxyribo**N**ucleic **A**cid (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) of Thymine (T). These 4 types of nucleotide 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,* 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].

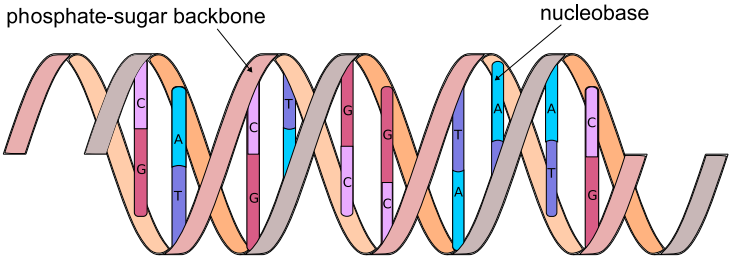

Figure 1.1.: **Double-helix structure of DNA**
Each strand of DNA has a phosphate-sugar backbone on which are attached nucleobases. The two strands are linked by complementary bonds between the nucleobases of different strands (A bonding with T and C bonding with G).

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 contains only information; in order for the organism to live, this information must be read and transformed 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 it's 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 3 dimensional structure, giving the protein the shape it needs to fulfill it's 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 transcripted in a similar, but single stranded, molecule called RNA 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 3 called a codon.
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.
5. The translation process is stopped when a specific type of codon is read.

With 4 types of nucleotides and codons grouping 3 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 a portion 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 2 strands of DNA, the redundant nature of the genetic code as well as error correction mechanisms in the molecules that read and write DNA and RNA *(polymerases)*. However, despite all that, some errors still make it through.

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 transmit genetic information. This process has a very low rate of errors, with as low as 1 error for every billion to every hundred billions of base pairs replicated [15]. This is due to the fact that the DNA polymerase (the protein that is responsible for copying DNA molecules), has a relatively error rate to start with, but mostly due to the error correcting mechanisms that are present in certain cells and bacteria [16].
- **RNA transcription:** error rate between 4 errors for each million [17] to 2 errors for each hundred thousand [18] base pairs transcribed.
- **Other mutagenic events:** Ionizing radiation [19], UV rays [20], Toxins [21], heat Stress [22], cold stress [23], oxidative stress [24].

1.1.2.2.2. What kind of errors are possible?

- substitution
- insertion
- deletion

1.1.2.2.3. What effect can mutations have ? SNPs, many individual mutations have been linked to certain traits (however, a trait is not usually caused only by a single mutation, plus hard to prove causality)

- Diseases:
  - Pathologic coagulation linked to a single mutation in DNA coding for a coagulation protein [25].
  - Cystic Fibrosis, a deletion of a single amino acid in the protein leads to fatal disease [26]
  - Plenty of mutation in specific genes have been linked with type 2 diabetes [27, 28]
- Added traits, like drug resistance:
  - Dozens of mutations proven to confer HIV resistance to commonly used drugs [29]
  - A large amount of mutations have been shown to confer antibacterial resistance in a range of bacteria [30]

1.2. Obtaining sequence data

In order to study living organisms we need to be able to obtain their genetic information, i.e figure out a way to get the sequence of nucleobases that make up their DNA.

1.2.1. Sanger sequencing, a breakthrough

The first true sequencing method was developed in 1977 [31]. Sanger *et al.* devised a simple method to read the sequence of nucleotides that make up a DNA sequence *(also represented in Figure 1.2)*.

1. Clone sequence / amplify
2. Prepare 4 different sequencing environments with a majority of dNTP (ie regular nucleotides) and in each a single type of ddNTP (a terminator). ddNTP are marked
3. In each test tube add DNA polymerase, primers and denatures DNA fragments you want to sequence
4. Sequence is replicated until incorporation of ddNTP stopping reaction
5. Separate replicated fragments by gel electrophoresis (i.e shorter fragments go further), 1 ddNTP type in each lane
6. With marked you can see which nucleotide is present at a given position

This allowed Sanger *et al.* to sequence the first whole $\varphi X174$ bacteriophage genome [2]. This method, although revolutionnary was costly and time consuming.

The marking of primers and ddNTP with fluorescence allowed to do the polymerization in a single test tube and use a single lane for electrophoresis [32, 33]. The fluorescence also allowed for automated reading with optical systems. Plus switching to capillary electrophoresis allowed for automation and speeding up. Shotgun sequencing also speed up stuff at the reconstruct the whole genome.

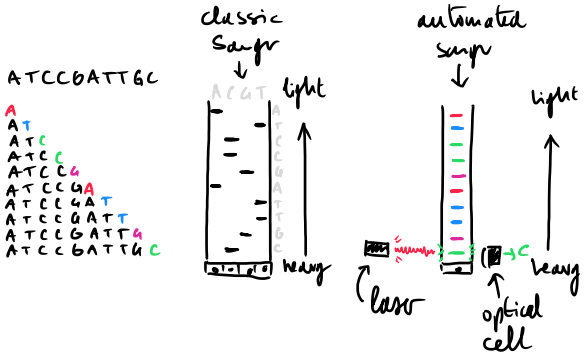

Figure 1.2.: **Caption**
(WIP)

With the latest Sanger sequencing technologies, sequenced sequences can reach 1kBp with an accuracy of 99.999% [34]. Although very time consuming and expensive at it's dawn, the sequencing of the first human genome costing between 500 million and 1 billion dollars [35]. Over time and with technological advancments, the cost of Sanger sequencing was greatly reduced from \$1000 per Bp [36] to \$0.5 per kBp [34], and the throughput increased from 1 kBp per day [36] to up to 120 kBp per hour [37].

1.2.2. Next-generation sequencing

Developed to lower cost and more throughput

- Massively parallel *(no need to detail technology, just numbers i.e. throughput and cost) (i.e. 2nd gen)* [38]
  - Key points:
    - Break up DNA to sequence in smaller fragments
    - Amplify fragments (i.e. clone them)

    - Attach fragments to solid support so we can thousands to billions of sequencing reactions at once
    - detect the polymerization as it happens, speeding up the process.
  - Illumina:
    - 150nt [39]
    - 98% accuracy, $0.07 per MBp, 2.5 to 12.5 Gbp per hour [37]
    - Detect polymerization by adding marked terminating NTP, imaging and cleaving terminating group to allow further polymerization.
  - pyrosequencing:
    - 400nt [39]
    - 99.9% accuracy, $10 per MBp, 30Mbp per hour [37]
    - Detect polymerization with a reaction that emits light during the incorporation of a dNTP to the synthesized DNA, measure the light emission to determine the base that was added.

- Long reads sequencing (i.e. 3rd gen):
  - Introduce need for longer reads:
    - Better for analyzing complex and repetitive parts of the genome, and were used for assembly **[CITE]**, structural variant detection **[CITE]**, etc…
    - Real time sequencing, i.e.
    - High throughput and
  - PacBIO SMRT [40]:
    - Tech summary:
      - Link 2 strands of DNA to sequence with a hairpin sequence to create circular DNA
      - Capture inserted with polymerase fixed to bottom of a well (up to 8million wells in 1 sequencing cell)
      - Florescently marked dNTP are added and excited with laser and light emission recorded by camera
      - Stop when you reach the hairpin
    - Most reads around 10kb in length with reads up to 60kb [41]
    - 85% to 92% accuracy [42, 43]
    - $0.32 per Mbp [44]
    - 2Gbp to 11Gbp per hour [43]

- NanoPore [45]:
  - Tech summary:
    - 100k nanopores spread out on synthetic membrane
    - Protein splits the 2 strands and pushes 1 of the brands through a nanopore at a controlled rate
    - As DNA goes through the pore it disrupts the electrical current between 2 sides of the membrane
    - Disruptions are characteristic of the 5/6 bases in the nanopore channel -> deduce sequence from current disruptions (i.e. base calling)
  - Median read length around 10kb - 12kb [46, 43].
  - you can get ultra long reads, of almost 1Mb [47, 48] and even up to 2.3 Mb [49].
  - 87% to 98% accuracy [50, 43]
  - $0.13 per Mbp [44]
  - 12.5Gbp to 260Gbp per hour !! [43]

As a conclusion, quickly mention the work done on protein sequencing:

- has been going on for a long time [51, 52] with mass spectrometry
- But still ongoing with new methods [53]

but we usually deduce the protein sequence from the DNA sequence that is translated from codons after detecting ORFs.

1.3. Sequencing errors, how to account for them ?

Just like mutations we can get substitutions and indels.

general context of sequencing errors, i.e. short reads have high accuracy (After computational correction you can get 10^{-4} to 10^{-5} error rate [54]) , long reads high error rate as stated before, 10% to 15%.

1.3.1. Error correction methods

- In Non-Hybrid error correcting methods only the long reads are used, by overlapping them and taking the consensus of the overlap it is possible to eliminate some errors, as well as filtering very rare subsequences likely due to a sequencing error [55, 56]. These methods are applied in several pieces of software like wtdbg2 [57], canu [58], or daccord [59].
- In Hybrid methods, shorter more accurate reads are used to correct long-read errors, it has been used in software like proovread [60], Jabba [61], PBcR [62] or LoRDEC [63]
- In some cases it can be useful to correct errors after having assembled the raw reads in a process called polishing. One can use non-hybrid methods by filtering out rare subsequences like in ntEdits , or correcting with raw PacBio or ONT data as is done in Arrow [64] or Nanopolish [65] respectively. Conversly, hybrid methods using short reads also exist to correct assembled reads like Pilon [66] or Racon [67].
- Error correction bibliography is rich and many people are working on this [55, 68, 69]

1.3.2. More accurate sequencing methods

- CCS / HiFi stuff -> 99.8% to 99.9% accuracy on long reads [70, 43] because errors are randomly distributed. Most errors are still indels in homopolymers [70]
- Using unique molecular identifiers and consensus reads, you can reduce error rate to 99.59% and 99.93% accuracies for ONT and PacBio CCS [71].
- Improve base calling models for ONT by switching from HMMs to deep learning, [72, 73, 74, 56]
- Built in error correction [75], results in shorter reads with 99.82% accuracy (200-250nt reads though). Error free up to 200bp.
- To come Illumina Infinity [76], high throughput long reads.

1.4. The special case of homopolymers

1.4.1. Homopolymers and the genome

- 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 length or more, and 10% are in homopolymers fof 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 1.3).
- More than 1.4 million homopolymers 4-mers and up in the exome (i.e. DNA that is translated into proteins) [77]
- according to [78], 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.

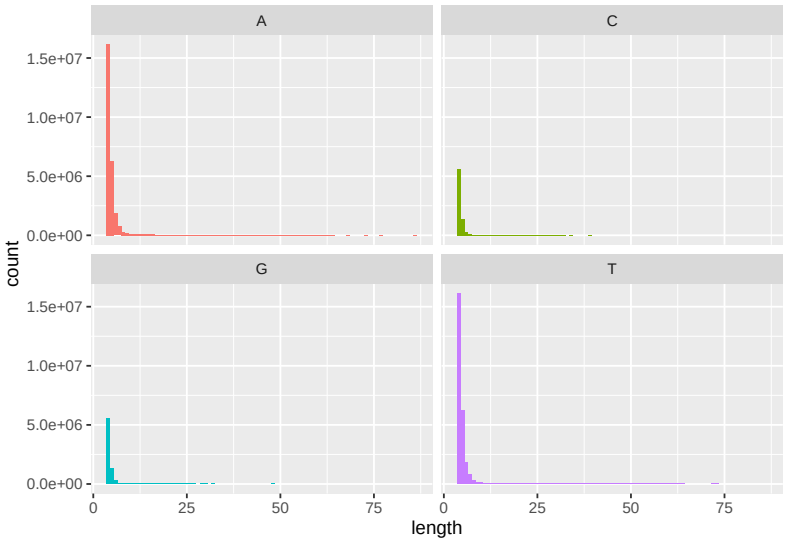

Figure 1.3.: Distribution of homopolymer lengths per base in the human genome (for homopolyers of length ≥ 4.)

1.4.2. Homopolymers and long reads

Mainly indels, in certain regions of the genomes, particularly homopolymers. For ONT [79, 80], for PacBIO and ONT [81, 82]. ONT has more homopolymer errors than PacBIO [43]. 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) [83].

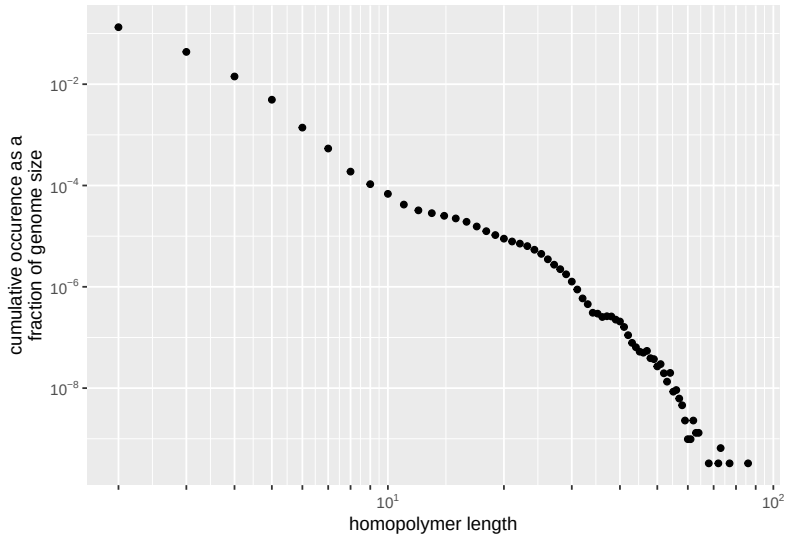

Figure 1.4.: Homopolymer length as fraction of the genome

1.4.3. Accounting for homopolymers

1.4.3.1. Specific error correction

- In some cases homopolymers are taken into special consideration during polishing [84, 65], and computational approaches are developed to deal with homopolymers [85, 86]
- Specific technologies:
  - R.10 Nanopore: [87, 56]
  - Solid state nanopores: [88, 89]
  - BaseCallers which deal with homopolymers, like can be seen in Fig 3 of [74] (scrappie raw 1.4.1 or guppy)
- Avoid using HPs in barcode sequences [90, 91].
- A little different but HPs have to be taken into account when developing DNA based data storage [92]

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 1.5.

- Used in many software tools for long reads [93, 94, 95, 57, 96, 97, 98] and even pyrosequencing reads [99]

reconstruct the whole genome.

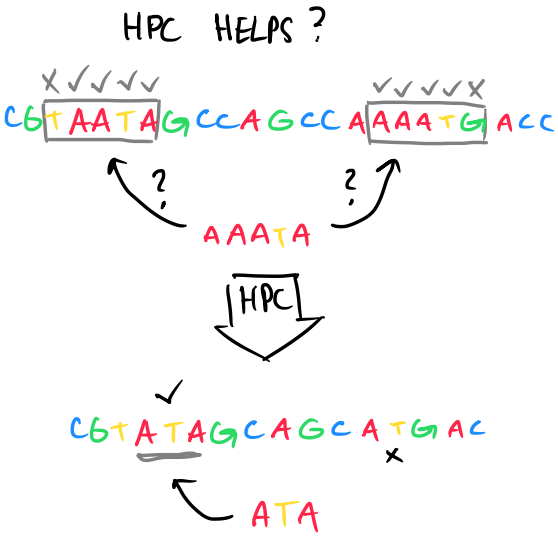

Figure 1.5.: **Caption**
(WIP)

1.5. Conclusion

- **Sequencing is an active field:** new methods, both in short reads (ultima) and long reads (illumina infinity), as well as improving existing technologies
- **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= watsonSTRUCTUREDNA1953 PMID: 13168976 publisher: Cold Spring Harbor Laboratory Press, pp. 123–131. DOI: 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. URL: https://www.nature.com/articles/265687a0 (visited on 05/17/2022) (cit. on pp. 11, 15).

[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. 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. URL: https://www.science.org/doi/10.1126/science.abj6987 (cit. on pp. 11, 59, 66).

[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. URL: https://doi.org/10.1111/j.1095-8339.2010.01072.x (cit. on p. 11).

[6] H. C. Macgregor. "C-Value Paradox". In: ed. by Sydney Brenner and Jefferey H. Miller. DOI: 10.1006/rwgn.2001.0301. New York: Academic Press, Jan. 1, 2001, pp. 249–250. DOI: 10.1006/rwgn.2001.0301. URL: https://www.sciencedirect.com/science/article/pii/B0122270800003013 (cit. on p. 11).

[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. 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. 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. 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. 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. 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. 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. 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. 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. 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. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5650487/ (cit. on p. 13).

[19] 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. URL: https://www.sciencedirect.com/science/article/pii/S1687850715000333 (cit. on p. 13).

[20] 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 (cit. on p. 13).

[21] 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. 13).

[22] 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. 13).

[23] 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. 13).

[24] 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. 13).

[25] 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).

[26] 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).

[27] 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).

[28] 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).

[29] 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. URL: https://academic.oup.com/nar/article/31/1/298/2401450 (cit. on p. 14).

[30] 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. URL: https://www.sciencedirect.com/science/article/pii/S1198743X14615500 (cit. on p. 14).

[31] 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. URL: https://www.pnas.org/doi/abs/10.1073/pnas.74.12.5463 (visited on 05/17/2022) (cit. on p. 14).

[32] 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. URL: https://doi.org/10.1093/nar/13.7.2399 (visited on 05/17/2022) (cit. on p. 15).

[33] 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. URL: https://www.nature.com/articles/321674a0 (visited on 05/17/2022) (cit. on p. 15).

[34] 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. URL: https://www.nature.com/articles/nbt1486 (cit. on p. 15).

[35] *The Cost of Sequencing a Human Genome*. URL: https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost (cit. on p. 15).

[36] 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. URL: https://www.science.org/doi/10.1126/science.1084564 (cit. on p. 15).

[37] Lin Liu, Yinhu 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. URL: https://www.hindawi.com/journals/bmri/2012/251364/ (visited on 05/16/2022) (cit. on pp. 15, 16).

[38] 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. URL: https://www.nature.com/articles/nrg2626 (visited on 05/16/2022) (cit. on p. 15).

[39] 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. URL: https://www.nature.com/articles/nature09796 (cit. on p. 16).

[40] 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. URL: https://www.science.org/doi/10.1126/science.1162986 (cit. on p. 16).

[41] 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. URL: https://www.sciencedirect.com/science/article/pii/S1672022915001345 (cit. on p. 16).

[42] 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. URL: https://www.nature.com/articles/nature13907 (cit. on p. 16).

[43] 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. URL: https://www.nature.com/articles/s41576-020-0236-x (visited on 05/16/2022) (cit. on pp. 16–19).

[44] 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. URL: https://doi.org/10.1093/gigascience/giaa146 (cit. on pp. 16, 17).

[45] 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. URL: https://www.nature.com/articles/nnano.2009.12 (cit. on p. 17).

[46] 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. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4722697/ (cit. on p. 17).

[47] 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. URL: https://www.nature.com/articles/nbt.4060 (cit. on p. 17).

[48] *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. 17).

[49] 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. URL: https://doi.org/10.1093/bioinformatics/bty841 (cit. on p. 17).

[50] 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. URL: https://doi.org/10.1186/s13059-016-1103-0 (cit. on p. 17).

[51] 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. URL: https://www.pnas.org/doi/abs/10.1073/pnas.83.17.6233 (cit. on p. 17).

[52] Bryan John Smith. *Protein Sequencing Protocols*. Springer Science & Business Media, 2002 (cit. on p. 17).

[53] 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. URL: https://www.nature.com/articles/s41565-018-0236-6 (cit. on p. 17).

[54] 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. URL: https://doi.org/10.1186/s13059-019-1659-6 (cit. on p. 17).

[55] 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. URL: https://doi.org/10.1093/bib/bbz058 (cit. on p. 18).

[56] 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. URL: https://doi.org/10.1186/s13059-020-1935-5 (visited on 05/16/2022) (cit. on pp. 18, 20).

[57] 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. URL: https://www.nature.com/articles/s41592-019-0669-3 (cit. on pp. 18, 21).

[58] 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. URL: https://genome.cshlp.org/content/27/5/722 (cit. on p. 18).

[59] 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. URL: https://www.biorxiv.org/content/10.1101/106252v1 (cit. on p. 18).

[60] 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. URL: https://doi.org/10.1093/bioinformatics/btu392 (cit. on pp. 18, 34).

[61] 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. URL: https://doi.org/10.1186/s13015-016-0075-7 (cit. on p. 18).

[62] 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. URL: https://www.nature.com/articles/nbt.2280 (cit. on pp. 18, 34).

[63] 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. URL: https://doi.org/10.1093/bioinformatics/btu538 (cit. on p. 18).

[64] 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. 18).

[65] 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. URL: https://www.nature.com/articles/nmeth.4184 (cit. on pp. 18, 20).

[66] 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. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0112963 (cit. on p. 18).

[67] 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 (cit. on p. 18).

[68] 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. URL: https://doi.org/10.1186/s13059-018-1605-z (cit. on p. 18).

[69] 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. URL: https://doi.org/10.1186/s12864-020-07227-0 (cit. on p. 18).

[70] 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. URL: http://www.nature.com/articles/s41587-019-0217-9 (cit. on p. 18).

[71] 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. URL: https://www.nature.com/articles/s41592-020-01041-y. (cit. on p. 18).

[72] 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. URL: https://doi.org/10.1093/bioinformatics/btab528 (cit. on p. 18).

[73] 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. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0178751 (cit. on p. 18).

[74] 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. URL: https://doi.org/10.1186/s13059-019-1727-y (cit. on pp. 18, 20).

[75] 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. URL: https://www.nature.com/articles/nbt.3982 (cit. on p. 18).

[76] *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 (cit. on p. 18).

[77] 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. URL: https://doi.org/10.1186/s12864-018-4544-x (cit. on p. 19).

[78] A. Sina Booeshaghi and Lior Pachter. “Pseudoalignment facilitates assignment of error-prone Ultima Genomics reads”. In: (). DOI: 10.1101/2022.06.04.494845 (cit. on p. 19).

[79] 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. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0257521 (cit. on p. 19).

[80] 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. URL: https://genome.cshlp.org/content/25/11/1750 (cit. on p. 19).

[81] 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 (cit. on pp. 19, 58).

[82] 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. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5553090/ (cit. on p. 19).

[83] 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. URL: https://www.nature.com/articles/s41587-021-01049-5 (cit. on p. 19).

[84] 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. URL: https://doi.org/10.1186/s13059-021-02282-6 (cit. on p. 20).

[85] 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. URL: https://doi.org/10.1186/s13059-018-1462-9 (cit. on p. 20).

[86] Peter Sarkozy, Ákos Jobbágy, 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 (cit. on p. 20).

[87] *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. 20).

[88] 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. URL: https://avs.scitation.org/doi/full/10.1116/1.5116295 (cit. on p. 20).

[89] 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. URL: https://pubs.rsc.org/en/content/articlelanding/2018/nr/c8nr04238a (cit. on p. 20).

[90] 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. URL: https://www.pnas.org/doi/full/10.1073/pnas.1802640115 (cit. on p. 20).

[91] 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. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/1755-0998.12890 (cit. on p. 20).

[92] 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 (cit. on p. 20).

[93] 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 (cit. on pp. 21, 59).

[94] 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. pmid: 32801147 (cit. on pp. 21, 59).

[95] 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 (cit. on pp. 21, 39, 59, 66, 67, 117).

[96] 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. URL: https://www.nature.com/articles/s41587-020-0503-6 (cit. on p. 21).

[97] 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 (cit. on pp. 21, 59).

[98] 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 (cit. on pp. 21, 59).

[99] 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 (cit. on pp. 21, 59).

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 [100]:

- **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 [101]
  - 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 [102] and the Levenshtein distance [103]
- Alignments are used in many cases so that:
  - we can examine similarities/differences between them (i.e. comparative genomics) [104]
  - we can infer (usually with multiple alignment):

    - evolutionary relationships (phylogenetics), and most methods to reconstruct phylogenetic trees take as input a multiple alignment: [105, 106, 107, 108, 109]
    - protein:
        - structure [110, 111]
        - function [112, 113]
  - we can correct sequencing errors [60, 62, 114].
  - Structural variant detection [115, 116]

2.1.2. How to align two sequences ?

The seminal method for pairwise alignment was the Needleman-Wünsch algorithm [117] based on a dynamic programming method. A decade later, the Smith-Waterman algorithm [118] 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 [119, 120], 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:
  - The score of an alignment can be defined recursively -> dynamic programming
  - Fill out DP matrix
  - Traceback to find optimal alignment
- Example… (can be at the same time as the algo presentation)
- This algorithm although very precise, has a time complexity of $O(nm)$ where n and m are the lengths of the sequences to align [100]. Some methods have been proposed to speed up [121], 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 [122, 123]. If we want to do better we have to rely on heuristics.
- 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 $O(n + m)$ [124] meaning we would only need a couple gigabytes to store the matrix for 2 human genomes.
  - Above idea resulting in Myers-miller algorithm [125], implemented in the EMBOSS stretcher tool [126]

2.1.2.2. Local alignment

- Similar ideas to NW
- Basic example of NW
- In terms of complexity is the same as NW in quadratic in time and space [100] 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 [127] which is implemented in the EMBOSS Lalign tool [126], and the Waterman Eggert algorithm [128].

Both methods are implemented in many different software tools and are used when performing pairwise alignments of short sequences [129, 130, 126] with version implemented for specific CPU instruction sets [131] or GPUs [132] to speed up alignment.

2.1.3. Substitution models / scoring

When scoring an alignment we can use different scoring models/substitution matrices [133]:

- 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) [134], 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 [135].
  - BLOSUM [136], 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 instad of log-odds methods [137, 138]
  - Model specific matrices:
    * Transmembrane matrices [139, 140]

    - Disordered regions in proteins: [141]
    - context-specific matrices [142]
    - Specific organisms like *P. falciparum* with pfSSM [143] or HIV [144]
    - specific to global alignment [145]

- Some on DNA alignment
  - You can derive a matrix with methods similar to PAM [146]
  - You can do codon substitution matrices [147], or combine codon matrices with AA matrices [148]

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 [149]
- Non affine gap penalties [150, 151, 152]

2.2. How do we speed up pairwise alignment ?

Although NW and SW give us optimal alignments [153] when dealing with large sequences they become impractical due to time and space complexity. We need heuristics, review [154], .

2.2.1. Change the method

- Bounded DP [155, 156], 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. 2.1
- HMMs: ***(I Need to find out if they actually are faster for pairwise, and how fast)***
  - PairHMMs can be used to align 2 sequences, in some cases it is mathematically equivalent to NW [157].
  - Software: HHsearch [158], HMMer [159], MCALIGN2 [160] is used to efficiently search for alignments in large databases

- You can also align with Fast Fourier Transforms (FFT) as in MAFFT [161] 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. 2.1
  - Many ways of finding seeds [162] and to index them [163]
  - 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.
- 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):

2.2.2.1. Hash tables

- BLAST [164]
  - 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.

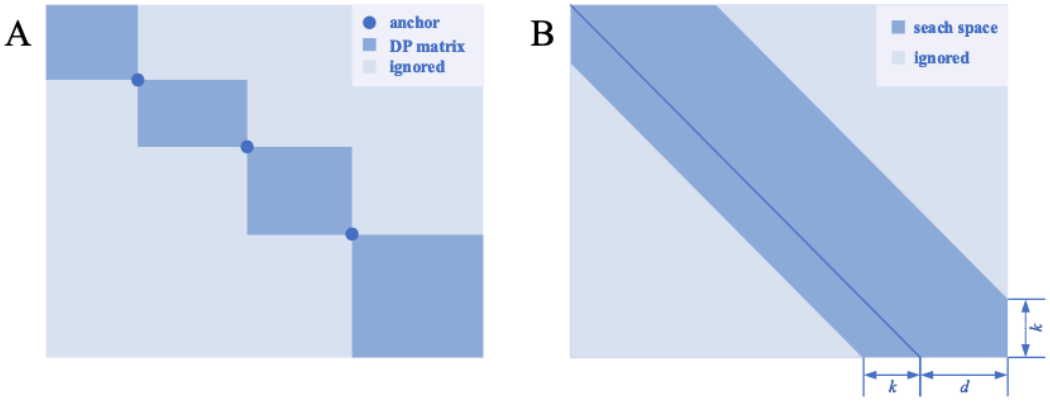

Figure 1. Two heuristic algorithms for pairwise sequence alignment. (**A**) A dynamic programming matrix, which is separated by several anchors, which is certain to be in the optimal path. (**B**) A shape-based bounded dynamic programming matrix in which the light blue block is calculation-free because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

Figure 2.1.: **2 heuristic methods to speed up alignment:** divide and conquer and bounded dynamic programming. Adapted from [154] **(Original figure here as a placeholder, I will adapt it)**

  - 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 [165].
  - It is important to note that it is a heuristic method, it has a decreased sensitivity to the SW algorithm which is optimal [166], 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 [167], derived from FASTP [168] 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 [169] in a database of over 80 million sequences [170], with HMMs being quite slow. It is much faster to use these than to try to do the same task with SW or NW [171].
- 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) [172, 173]
- Minimap2 [95] minimizer hash table
- Diamond [174, 175] which indexes both the reference and query at the same time.

2.2.2.2. Suffix trees / suffix arrays

- Used in many pattern matching problems [102].
- Software: AVID uses suffix trees to find anchors [176], MUMmer finds largest identical subsequences between 2 sequences to anchor alignments using suffix trees [177]

2.2.2.3. FM Index

- FM index [178]:
- 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 [179].
- used for in exact and approximate string matching [100]
- Software: BWT-SW uses an FM-index to speed up local alignment [180], Bowtie 2 uses an FM index to find seeds [181], BWA and BWA-SW use a similar idea [182, 183], BWA-MEM [184] and CUSHAW [185] 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 [186].

NP-hard [187, 188] 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. [189] 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 [190, 191]
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" [189]).

In order to curtail this problem we have iterative refinement [186]:

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 [192], or others like log-odds, correlation [193] or consistency [194])

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

- CLUSTAL-W [195] and CLUSTAL-X [196]
- T-Coffee [197]
- MUSCLE [198, 199]
- MAFFT [161]
- ProbCons [200]

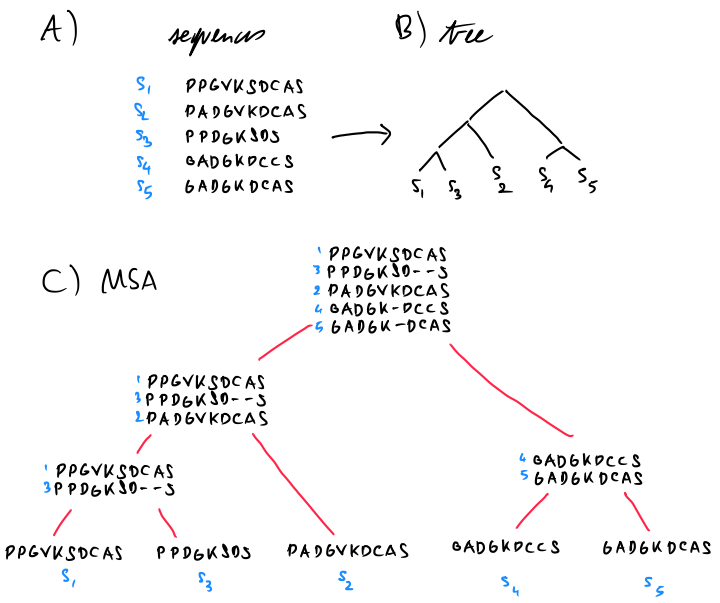

Figure 2.2.: **Overview of the progressive alignment process.**
A) sequences to align B) guide tree constructed from distances C) Alignment steps along the guide tree and resulting MSA. Adapted from [100]

Different strenghts/weaknesses so many reviews and benchmarks to make a choice [201, 202, 203, 204, 205]

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

HMM method similar performance to clustal-w [206], HHMer [159] and COVID-align COVID-align [207]

Example of COVID where homology is high so we can get away with aligning pairwise to ancestral-sequence NextClade/NextAlign [208]

2.3.3. Other optimization methods (short)

- Simulated annealing to speed up DP [209, 210]
- Genetic algorithm, review [211]
  - SAGA [212]
- recently: regressive method [213] root to leaf allows to align a very large number of sequences: 1.4 million!

2.4. The specifities of read-mapping

Huge review of mapping in [214]

- Mapping is effectively the task of finding a bunch of independent local alignments between query sequences and a target reference sequence.
- context of short read (query) on large reference sequence (even with long reads they are usually much smaller than the reference genome)
- Problems:
  - Repetitive regions (centromeres, telomeres) make it hard [214]
  - low homology / sequencing errors… make mapping and other tasks hard [102]
- short-reads mapping
  - benchmarks/review [215, 216]
  - Algorithms in short read mapping [217] (basically same thing as speeding up section)
  - Hardware accelerated [218]
- Long read mapping:

  - software: winnowmap [219], winnowmap2 [220] and tandemtools [221] for repetitive regions.

- Mapping quality:
  - Intro and definition, quite a loosely defined term.
  - Mapping quality from tandem simulation [222]

- Benchmarking is of interest of field [223]

References for chapter 2

[60] 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. URL: https://doi.org/10.1093/bioinformatics/btu392 (cit. on pp. 18, 34).

[62] 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. URL: https://www.nature.com/articles/nbt.2280 (cit. on pp. 18, 34).

[95] 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 (cit. on pp. 21, 39, 59, 66, 67, 117).

[100] Wing-Kin Sung. *Algorithms in Bioinformatics: A Practical Introduction*. DOI: 10.1201/9781420070347. New York: Chapman and Hall/CRC, Oct. 10, 2011. DOI: 10.1201/9781420070347 (cit. on pp. 33–35, 39, 41).

[101] Richard Wesley Hamming. *Coding and Information Theory*. tex.ids= hamming1980coding googlebooksid: ed5QAAAAMAAJ lccn: 79015159. Prentice-Hall, 1980 (cit. on p. 33).

[102] 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. 33, 39, 42, 58).

[103] 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. 33).

[104] 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. 33).

[105] 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. 34).

[106] 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. 34).

[107] 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. 34).

[108] 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. 34).

[109] 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. 34).

[110] 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. 34).

[111] 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. 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. 34).

[112] 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. URL: https://www.sciencedirect.com/science/article/pii/S0959440X05000825 (cit. on p. 34).

[113] 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. URL: https://www.nature.com/articles/nrm2281 (cit. on p. 34).

[114] 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. URL: https://doi.org/10.1093/bioinformatics/btr170 (cit. on p. 34).

[115] 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. URL: https://www.nature.com/articles/nmeth.1374 (cit. on p. 34).

[116] 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. URL: https://doi.org/10.1186/s13059-019-1828-7 (cit. on p. 34).

[117] 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. URL: http://www.sciencedirect.com/science/article/pii/0022283670900574 (cit. on p. 34).

[118] 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. URL: https://www.sciencedirect.com/science/article/pii/0022283681900875 (cit. on p. 34).

[119] 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. 34).

[120] 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. URL: https://www.ams.org/bull/1954-60-06/S0002-9904-1954-09848-8/ (cit. on p. 34).

[121] 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. URL: https://www.sciencedirect.com/science/article/pii/0022000080900021 (cit. on p. 34).

[122] 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. 34, 92, 127).

[123] 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. URL: https://doi.org/10.1145/321921.321922 (cit. on p. 34).

[124] 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. URL: https://doi.org/10.1145/360825.360861 (cit. on p. 35).

[125] 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. URL: https://doi.org/10.1093/bioinformatics/4.1.11 (cit. on p. 35).

[126] 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 p. 35).

[127] 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. URL: https://www.sciencedirect.com/science/article/pii/019688589190017D (cit. on p. 35).

[128] 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. URL: https://www.sciencedirect.com/science/article/pii/0022283687904785 (cit. on p. 35).

[129] 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. URL: https://genome.cshlp.org/content/12/10/1611 (cit. on p. 35).

[130] 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. URL: https://doi.org/10.1186/gb-2004-5-10-r80 (cit. on p. 35).

[131] 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. URL: https://doi.org/10.1186/s12859-016-0930-z (cit. on p. 35).

[132] 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. URL: http://journals.pan.pl/dlibra/publication/96876/edition/83624/content (cit. on p. 35).

[133] Stephen F Altschul. "Substitution Matrices". In: John Wiley & Sons, Ltd, 2013. DOI: 10.1002/9780470015902.a0005265.pub3. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470015902.a0005265.pub3 (cit. on p. 35).

[134] 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. 35).

[135] 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 (cit. on p. 35).

[136] 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. URL: https://www.pnas.org/content/89/22/10915 (cit. on p. 35).

[137] 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 (cit. on p. 35).

[138] 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. URL: https://doi.org/10.1093/molbev/msn067 (cit. on p. 35).

[139] 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. URL: https://doi.org/10.1093/bioinformatics/17.suppl_1.S182 (cit. on p. 35).

[140] 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 (cit. on p. 35).

[141] 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. URL: https://www.nature.com/articles/s41598-019-52532-8 (cit. on p. 36).

[142] 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. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.21775 (cit. on p. 36).

[143] 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 (cit. on p. 36).

[144] 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. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1876811/ (cit. on p. 36).

[145] 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 (cit. on p. 36).

[146] 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. URL: https://www.worldscientific.com/doi/abs/10.1142/9789812799623_0012 (cit. on p. 36).

[147] 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 (cit. on p. 36).

[148] 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. URL: https://doi.org/10.1093/molbev/msl175 (cit. on p. 36).

[149] 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. URL: https://www.sciencedirect.com/science/article/pii/0022283682903989 (cit. on p. 36).

[150] 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. URL: https://www.sciencedirect.com/science/article/pii/S0022283683711058 (cit. on p. 36).

[151] 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. URL: https://doi.org/10.1007/BF00164032 (cit. on p. 36).

[152] 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. URL: https://www.sciencedirect.com/science/article/pii/S0022519384800375 (cit. on p. 36).

[153] 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. Academic Press, Jan. 1, 1992, pp. 575–601. DOI: 10.1016/0076-6879(92)10029-D. URL: https://www.sciencedirect.com/science/article/pii/007668799210029D (cit. on p. 36).

[154] 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. URL: https://www.mdpi.com/2218-273X/12/4/546 (cit. on pp. 36, 38).

[155] 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. URL: https://epubs.siam.org/doi/abs/10.1137/0149094 (cit. on p. 36).

[156] 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. URL: https://doi.org/10.1093/nar/12.1Part1.175 (cit. on p. 36).

[157] 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. 36).

[158] 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. 36).

[159] 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. 36, 42).

[160] 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. 36).

[161] 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. 37, 40).

[162] 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. 37).

[163] 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/bbq015. URL: https://doi.org/10.1093/bib/bbq015 (visited on 05/16/2022) (cit. on p. 37).

[164] 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. 37).

[165] 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. 38).

[166] 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. URL: https://www.sciencedirect.com/science/article/pii/S088875439690614X (cit. on p. 38).

[167] 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 (cit. on p. 38).

[168] 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 (cit. on p. 38).

[169] 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. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5013910/ (cit. on p. 39).

[170] 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. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4702930/ (cit. on p. 39).

[171] 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. 39).

[172] 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. URL: https://pnas.org/doi/full/10.1073/pnas.87.6.2264 (cit. on p. 39).

[173] Richard Mott. "Alignment: Statistical Significance". In: _eprint: https://onlinelibrary.wiley.com/doi/pdf/10 DOI: 10.1038/npg.els.0005264. John Wiley & Sons, Ltd, 2005. DOI: 10.1038/npg.els.0005264. URL: https://onlinelibrary.wiley.com/doi/abs/10.1038/npg.els.0005264 (cit. on p. 39).

[174] 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 (cit. on p. 39).

[175] 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. URL: https://www.nature.com/articles/s41592-021-01101-x (cit. on p. 39).

[176] 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. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC430967/ (cit. on p. 39).

[177] 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. 39).

[178] 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 (cit. on p. 39).

[179] Michael Burrows and David Wheeler. *A Block-Sorting Lossless Data Compression Algorithm*. Tech. rep. 1994 (cit. on p. 39).

[180] 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. URL: https://doi.org/10.1093/bioinformatics/btn032 (cit. on p. 39).

[181] 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 (cit. on p. 39).

[182] 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. URL: https://doi.org/10.1093/bioinformatics/btp324 (cit. on p. 39).

[183] 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. URL: https://doi.org/10.1093/bioinformatics/btp698 (cit. on p. 39).

[184] Heng Li. *Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM*. May 26, 2013. arXiv: 1303.3997 [q-bio] (cit. on p. 39).

[185] 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. URL: https://doi.org/10.1093/bioinformatics/bts414 (cit. on p. 39).

[186] 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. URL: http://link.springer.com/10.1007/978-1-62703-646-7 (cit. on p. 40).

[187] 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. URL: https://www.liebertpub.com/doi/abs/10.1089/cmb.1994.1.337 (cit. on p. 40).

[188] 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. URL: https://www.liebertpub.com/doi/abs/10.1089/106652701753307511 (cit. on p. 40).

[189] Da-Fei Feng and Russell F. Doolittle. "Progressive sequence alignment as a prerequisitetto correct phylogenetic trees". In: *Journal of Molecular Evolution* 25.4 (Aug. 1, 1987), pp. 351–360. DOI: 10.1007/BF02603120. URL: https://doi.org/10.1007/BF02603120 (cit. on p. 40).

[190] 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. URL: https://doi.org/10.1093/bioinformatics/8.3.275 (cit. on p. 40).

[191] 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. URL: https://www.pnas.org/doi/abs/10.1073/pnas.83.14.5155 (cit. on p. 40).

[192] 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. URL: https://www.sciencedirect.com/science/article/pii/0022283689902349 (cit. on p. 40).

[193] 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. URL: https://doi.org/10.1093/bioinformatics/bth090 (cit. on p. 40).

[194] 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. URL: https://doi.org/10.1093/bioinformatics/14.5.407 (cit. on p. 40).

[195] 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. URL: https://academic.oup.com/nar/article/22/22/4673/2400290 (cit. on p. 40).

[196] 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. URL: https://doi.org/10.1093/nar/25.24.4876 (cit. on p. 40).

[197] Cédric Notredame, Desmond G Higgins, and Jaap Heringa. "T-coffee: a novel method for fast and accurate multiple sequence alignment11Edited by J. Thornton". In: *Journal of Molecular Biology* 302.1 (Sept. 8, 2000), pp. 205–217. DOI: 10.1006/jmbi.2000.4042. URL: http://www.sciencedirect.com/science/article/pii/S0022283600940427 (cit. on p. 40).

[198] 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. URL: https://doi.org/10.1186/1471-2105-5-113 (cit. on p. 40).

[199] 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. URL: https://doi.org/10.1093/nar/gkh340 (cit. on p. 40).

[200] 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. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC546535/ (cit. on p. 40).

[201] 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. URL: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.0030123 (cit. on p. 42).

[202] 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. URL: https://www.futuremedicine.com/doi/abs/10.1517/14622416.3.1.131 (visited on 05/16/2022) (cit. on p. 42).

[203] 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. URL: https://www.sciencedirect.com/science/article/pii/S0959440X06000704 (visited on 05/16/2022) (cit. on p. 42).

[204] 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. URL: https://doi.org/10.1186/1748-7188-9-4 (cit. on p. 42).

[205] 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. URL: https://academic.oup.com/bioinformatics/article/15/1/87/218377 (cit. on p. 42).

[206] Sean R Eddy. “Multiple Alignment Using Hidden Markov Models”. In: (), p. 7 (cit. on p. 42).

[207] Frédéric Lemoine, Luc Blassel, 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. URL: https://doi.org/10.1093/bioinformatics/btaa871 (cit. on p. 42).

[208] 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. URL: https://joss.theoj.org/papers/10.21105/joss.03773 (cit. on p. 42).

[209] 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. URL: https://doi.org/10.1093/bioinformatics/10.4.419 (cit. on p. 42).

[210] 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. URL: https://doi.org/10.1093/bioinformatics/9.3.267 (cit. on p. 42).

[211] 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. URL: http://www.sciencedirect.com/science/article/pii/S0888754317300551 (cit. on p. 42).

[212] 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. 42).

[213] 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. URL: https://www.nature.com/articles/s41587-019-0333-6 (cit. on p. 42).

[214] 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. URL: https://doi.org/10.1186/s13059-021-02443-7 (cit. on p. 42).

[215] 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. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3375638/ (cit. on p. 42).

[216] 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. URL: https://doi.org/10.1186/1471-2105-14-184 (cit. on p. 42).

[217] 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 (cit. on p. 42).

[218] 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.2012.36 (cit. on p. 42).

[219] 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 (cit. on pp. 43, 69, 117).

[220] 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. URL: https://www.nature.com/articles/s41592-022-01457-8 (cit. on p. 43).

[221] 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 (cit. on pp. 43, 66).

[222] 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. URL: https://doi.org/10.1186/s13059-017-1290-3 (cit. on p. 43).

[223] 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. URL: https://doi.org/10.1093/bioinformatics/btv524 (cit. on p. 43).

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*).
The author list, complete with affiliations is given below:

Luc Blassel[1,2*], Paul Medvedev[3,4,5], Rayan Chikhi[1]

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 preprocessing 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) [81]. Many algorithmic problems, such as alignment, become trivial if not for sequencing errors [102]. 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 AAAGGTTA 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

not correspond to true alignments) the improvement in mapping sensitivity usually outweighs it [95].

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 [99]. It is used by a wide range of error-correction algorithms, e.g. for 454 data [224], PacBio data [93], and Oxford Nanopore data [225]. HPC is used in alignment, e.g. by the widely used minimap2 aligner [95]. HPC is also used in long-read assembly, e.g. HiCanu [94], SMARTdenovo [226], or mdBG [98]. HPC is also used for clustering transcriptome reads according to gene family of origin [97]. 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

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

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

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 ℓ over x. The core of the transformation is given by g and the size of the context ℓ, 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 Σ and a context size ℓ, a function T mapping strings to strings is said to be an *order-ℓ* Streaming sequence reduction (abbreviated *SSR*) if there exists some $g : \Sigma^\ell \to \Sigma \cup \{\varepsilon\}$ such that $T = f$.

Figure 3.1A shows how an SSR can be visualized as a directed graph. Observe that an order-ℓ 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.

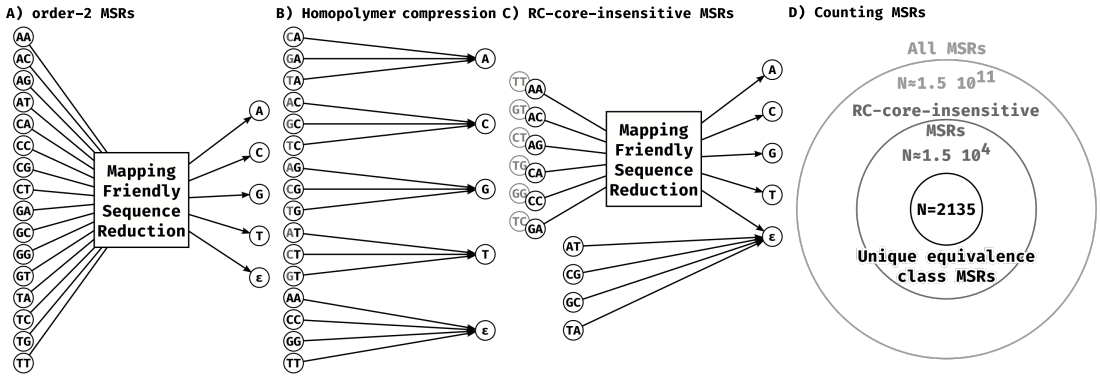

Figure 3.1.: **Representing and counting Streaming sequence reductions.** **A**: General representation of an order-2 Streaming sequence reduction as a mapping of 16 input dinucleotides, to the 4 nucleotide outputs and the empty character ε. **B**: Homopolymer compression is an order-2 SSR. All dinucleotides 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 ε. **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 ε. The 4 dinucleotides that are their own reverse complement are always mapped to ε. 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 ℓ-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 ℓ-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 ε, since a single nucleotide output cannot be its own reverse complement. This leaves 12 ℓ-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 ℓ-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:

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

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 ℓ (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

(as it does not map the di-nucleotides AT,TA,GC and CG to ε), 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 ℓ-mers that are their own reverse complements, hence $i(\ell) = 4^\ell/2$. If ℓ 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 ℓ-mer is canonical if it is the not lexicographically larger than its reverse complement. Let I be the set of all ℓ-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 ε). 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 ℓ-mers not in I is determined by its output on ℓ-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, ..., S_k, and then having a function t that is an injection from $\{1, \dots, k\}$ to Σ that assigns an output letter to each partition. Further, we permanently assign the output letter for S_0 to be ε. 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 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 $S'_0, , S'_k, t'$, respectively. We say that they are equivalent iff all the following conditions are met:

- $S_0 = S'_0$,
- there exists a permutation π 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 $\textsc{IsComp}(t(i), t(j)) = \textsc{IsComp}(t'(\pi(i)), t'(\pi(j)))$.

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.

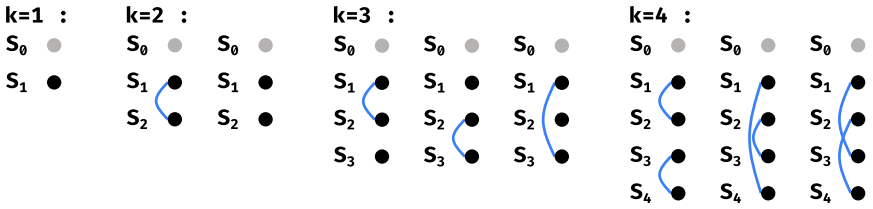

Figure 3.2.: **SSR equivalence classes for a fixed partition of the inputs.** S_0 is always assigned ε, so it is represented by a gray node. A blue link between S_i and an S_j denotes that $\textsc{IsComp}(t(i), t(j)) = \text{true}$. The equivalence classes are determined by the Watson-Crick complementary 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 ℓ-mers to an output is $i(\ell) = 4^\ell/2$ if ℓ is odd and $i(\ell) = (4^\ell - 4^{\ell/2})/2$ if ℓ is even. Let $C(\ell, k)$ be the number of ways that $i(\ell)$ ℓ-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

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

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 $\left\{ {n \atop k} \right\}$ [227, p.265]. It can be computed using the formula

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

Let j be the number of the $i(\ell)$ ℓ-mers that are assigned to S_0. Note this does not include the ℓ-mers that are self-complementary that are forced to be in S_0. Let $C(\ell, k, j)$ be the number of ways that $i(\ell)$ ℓ-mers can be partitioned into $k+1$ sets $S_0, \ldots, S_k$, such that j of the ℓ-mers go into $|S_0|$ and $S_1, \ldots, S_k$ 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 $S_1, \ldots, S_k$. Then $C(\ell, k, j) = \left\{ {i(\ell) \atop k} \right\}$.
- In the case that $1 \leq j \leq i(\ell) - k$, there are $\binom{i(\ell)}{j}$ ways to choose which j ℓ-mers are in S_0, and $\left\{ {i(\ell)-j \atop k} \right\}$ ways to partition the remaining ℓ-mers into $S_1, \ldots, S_k$. Hence, $C(\ell, k, j) = \binom{i(\ell)}{j} \left\{ {i(\ell)-j \atop k} \right\}$.
- In the case that $j > i(\ell) - k$, it is impossible to partition the remaining k (or fewer) ℓ-mers into $S_1, \ldots, S_k$ 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

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

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) [228].

3. **Synthetic centromeric sequence:** This sequence was obtained from the `TandemTools` mapper test data [221]. 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` [229] 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`[95] 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` [95]. 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 higest 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.

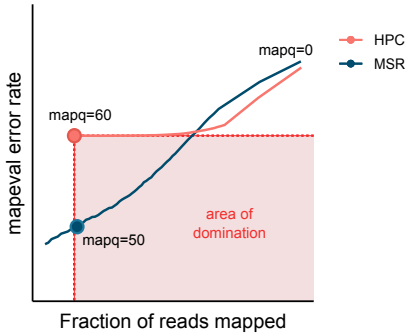

Figure 3.3.: **Illustration of how a respective mapq threshold is chosen for each of our evaluated MSRs.**
The orange dot shows the error rate and fraction of reads mapped for HPC at mapq threshold 60. Anything below and to the right of this point is strictly better than HPC 60, i.e. it has both a lower 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**$_{\mathrm{F}}$: The MSR with the highest fraction of mapped reads at a mapq threshold of 0.
- **MSR**$_{\mathrm{E}}$: The MSR with the lowest error rate at its respective mapq threshold.
- **MSR**$_{\mathrm{P}}$: 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 MSR$_{\mathrm{F}}$, MSR$_{\mathrm{E}}$, MSR$_{\mathrm{P}}$. An intriguing property is that they output predominantly As and Ts, with MSR$_{\mathrm{P}}$ assigning 2 input pairs to the G/C output whereas MSR$_{\mathrm{E}}$ and MSR$_{\mathrm{F}}$ assign only one. Additionally, MSR$_{\mathrm{E}}$ and MSR$_{\mathrm{P}}$ both assign the {CC,GG} input pair to the deletion output ε 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.

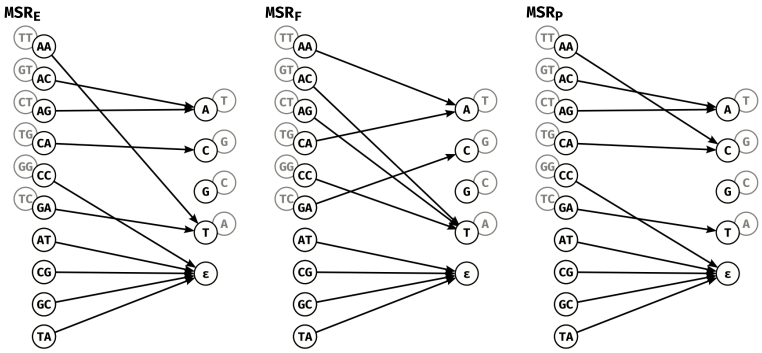

Figure 3.4.: **Graph representations of our highlighted MSRs: MSR_E, MSR_F, and MSR_P.**
MSR_E has the lowest error rate of among MSRs at the highest mapq threshold for which it performs better than HPC at mapq 60, MSR_F has the highest fraction of reads mapped at mapq 60 and MSR_P has the highest percentage 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 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 mapq$\geq$ 50, MSR_F, MSR_P and MSR_E all map more reads than either HPC or raw at mapq=60, and MSR_P and 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 `winnowmap2`(version 2.02) [219]. The same options as `minimap2` were used, and k-mers were counted using `meryl` [230], considering the top 0.02% as repetitive (as suggested by the `winnowmap2` usage guide).

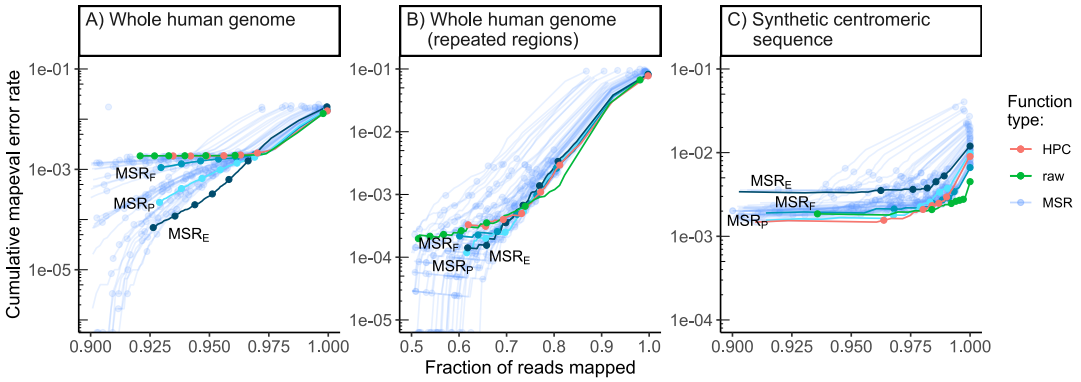

Figure 3.5.: **Performance of our 58 selected mapping-friendly sequence reductions across genomes on reads simulated by `nanosim`**
Panel **A)** shows the whole human genome assembly, **B)** the subset of mapped reads from panel B that originate from repetitive regions, and **C)** the "TandemTools" synthetic 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 ≥ 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.

MSRs E, F and P behave very similarly in both of these contexts compared to HPC/raw: by selecting mapped reads with mapq≥ 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%	**0.932 +4%**	1.75e-03 +23%	0.958 +0%	2.27e-03 − 0%
MSR_F	50	**0.938 +0%**	1.29e-03 −30%	0.886 −1%	3.82e-04 −73%	**0.960 +0%**	1.37e-03 − 39%
MSR_E	50	0.936 +0%	**1.17e-04 −94%**	0.820 −8%	**8.93e-05 −94%**	0.954 −0%	**0.00 −100%**
MSR_P	50	**0.938 +0%**	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 ≥ 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 ≥ 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 [231, 232, 233]. However across all mapped reads (mapq ≥ 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

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/lucblassel/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. URL: https://www.science.org/doi/10.1126/science.abj6987 (cit. on pp. 11, 59, 66).

[81] 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 (cit. on pp. 19, 58).

[93] 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 (cit. on pp. 21, 59).

[94] 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. pmid: 32801147 (cit. on pp. 21, 59).

[95] 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 (cit. on pp. 21, 39, 59, 66, 67, 117).

[97] 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 (cit. on pp. 21, 59).

[98] 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 (cit. on pp. 21, 59).

[99] 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 (cit. on pp. 21, 59).

[102] 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. 33, 39, 42, 58).

[219] 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 (cit. on pp. 43, 69, 117).

[221] 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 (cit. on pp. 43, 66).

[224] 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 (cit. on p. 59).

[225] 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 (cit. on p. 59).

[226] 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 (cit. on p. 59).

[227] 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. 65).

[228] 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. pmid: 10731132 (cit. on p. 66).

[229] 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 (cit. on pp. 66, 115).

[230] 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 (cit. on p. 69).

[231] 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 (cit. on p. 73).

[232] Heng Li. *New Strategies to Improve Minimap2 Alignment Accuracy*. Aug. 7, 2021. arXiv: 2108.03515 [q-bio] (cit. on p. 73).

[233] 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 (cit. on p. 73).

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

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 one 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. HIV and DRMs

5.1. What are viruses ?

small presentation / definition of viruses

DNA / RNA viruses

5.2. What is HIV ?

5.2.1. Quick Presentation of HIV

- pandemic
- history

5.2.2. Replication cycle of HIV

- proteins (+ computational representation as a string of letters)
- schematic cycle

5.3. Drug resistance in HIV

When on ART, virus evolves under selectuve pressure and develops resitance -> treatment failure.

5.3.1. How does ART work

target the proteins, RT, PR, IN (small history of ART)

5.3.2. different types of resistance

- NRTI
- NNRTI
- Entry inhibitors
- PI
- INSTI

5.3.3. Consequences on global health

Transmitted DRMS can be very serious , … however fitness cost, …

5.3.4. Finding DRMS

- Consortiums / HIVDB, UK-CHIC, …
- stat tests
  - multiple testing
  - phylogenetic correlation
- assays
- novel approaches
  - deep learning
  - …

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*).
The author list, complete with affiliations is given below:

Luc Blassel[1,2*], Anna Tostevin[3], Christian Julian Villabona-Arenas[4,5], Martine Peeters[6], Stéphane Hué[4,5], Olivier Gascuel[1,7#] On behalf of the UK HIV Drug Resistance Database[^]

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.blassel@pasteur.fr (LB)
* olivier.gascuel@mnhn.fr (OG)

^ 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 [234, 235]. Furthermore, drug-resistant HIV strains can be transmitted to treatment-naive individuals and further spread throughout the population over time [236, 237, 238]. 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 [239, 240, 241]. However, with the developement 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 [242, 243]. 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

transmission rather than independent acquisition twice [244]) 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 [245], 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 [246] or other indicators from phenotypic resistance assays such as PhenoSense [247]. Many methods have been used to predict a resistance level: Support Vector Machines (SVMs) [248], k-Nearest Neighbors (KNN) and Random Forests (RFs) [249], and more recently Artificial Neural Networks (ANNs) [250, 251]. 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 [252, 253], ensemble classifier chains [254, 255] and also ANNs [256]. Most recently Steiner *et al.* [257] 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.* [257] 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 ($n \approx 4,000$) [243]. 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

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 [243]. 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

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 [243], 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 [243]. 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 [258]. 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 [259, 260, 261], 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

		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 ≥ 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 [243]. 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 [262]. Those for which the corrected p-value was ≤ 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 [263]; the higher (≈ 1.0) and the lower (≈ 0) conditional probabilities correspond to the most important features.

2. Logistic regression (LR) with L1 regularization (LASSO) [259] 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 [264].

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 [265, 266, 267]. 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 [261].
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 [243]. 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 [268], 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 [122]. 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 [260] 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); conversly, 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 sore, 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

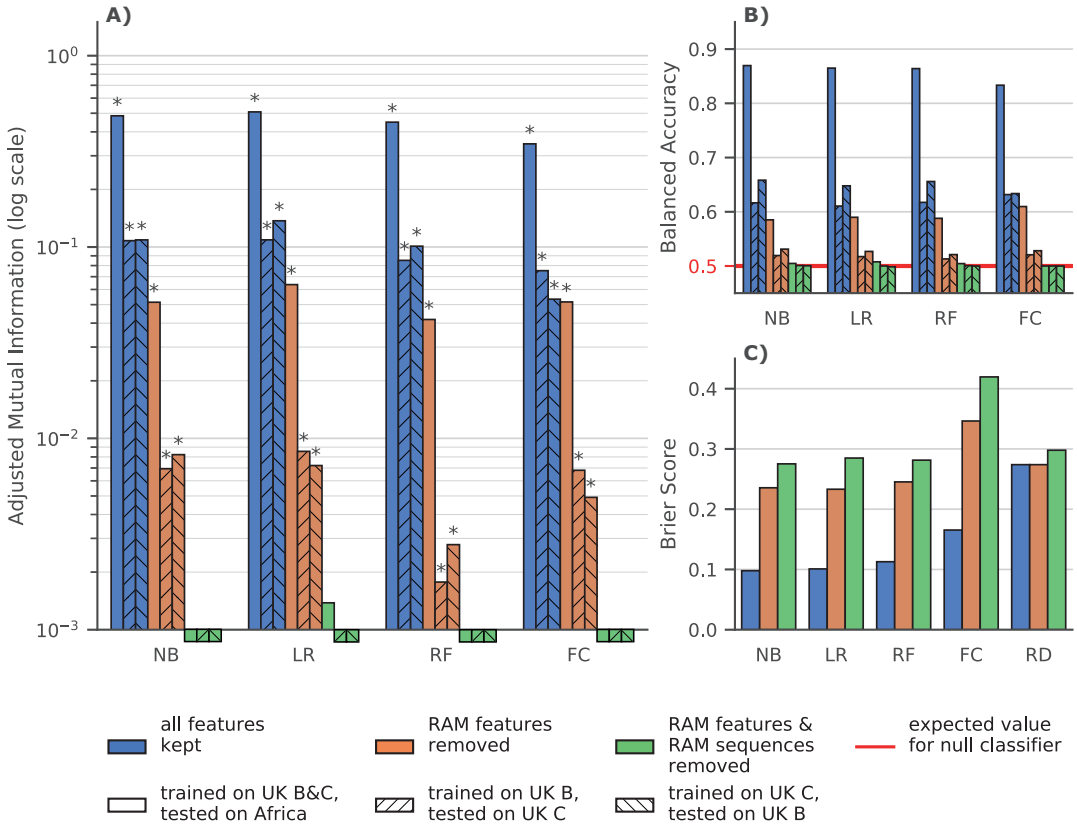

Figure 6.1.: **Classifier Performance on UK and African datasets.**
NB: naive Bayes, **LR**: Logistic Regression with Lasso regularization, **RF**: Random Forest, **FC**: Fisher Classifier, **RD**: Agnostic random probabilistic classifier (this classifier predicts, as the probability of a sample belonging 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 ≤ 0.05. For example when trained with all features all the classifiers have a significative MI. Conversly 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)**.

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 [243]) 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

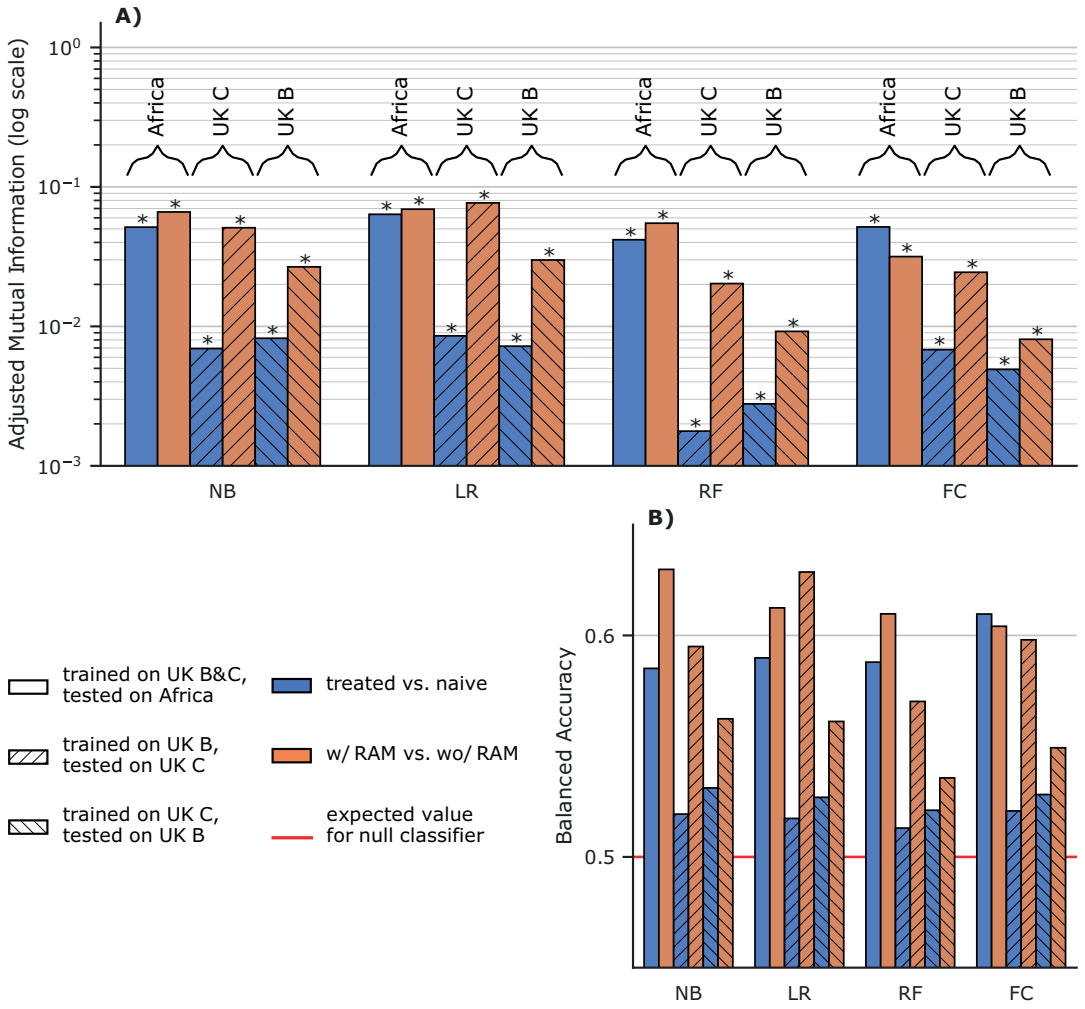

Figure 6.2.: **Discrimination between sequences having at least one RAM, and those having none on sequences with training features corresponding to known RAMs removed.**
NB: naive Bayes, **LR**: Logistic Regression with Lasso regularization, **RF**: Random Forest, **FC**: Fisher Classifier. **A)** Adjusted 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.

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
new present	A	B
new absent	C	D

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

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 (B.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			
					$RR(new, X)$		
	min	avg	B62	count	*treatment*	*any RAM*	p-value
L228R	1	1.16	-2	227 (0.4%)	18.1 [12.9;27.3]	115.7 [55.1;507.3]	$3.4 \cdot 10^{-31}$
E203K	1	1.31	1	256 (0.5%)	11 [8.2;15.1]	20.1 [13.7;32.1]	$1.1 \cdot 10^{-14}$
D218E	1	1	2	168 (0.3%)	13.1 [9.0;19.6]	27 [16.3;57.0]	$3.3 \cdot 10^{-10}$
L228H	1	1.12	-3	287 (0.5%)	6.4 [5.1;8.4]	9.2 [6.9;12.6]	$4.4 \cdot 10^{-16}$
I135L	1	1.16	2	540 (1.0%)	1.8 [1.5;2.1]	2.4 [2.0;2.8]	$5.9 \cdot 10^{-08}$
H208Y	1	1.10	2	205 (0.4%)	8.8 [6.5;12.5]	14.9 [9.9;23.6]	$1.2 \cdot 10^{-05}$
RAMs	1 [1;2]	1.35 [1;2.44]	0 [-2;3]	58 (0.1%) [2;1842]	8.3 [0.6;∞]	26.4 [1.4;∞]	$3.1 \cdot 10^{-2}$ [$2.3 \cdot 10^{-58}$;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^{th} and 95^{th} percentiles which are shown underneath. $RR(RAM, any\ RAM)$ values were computed for any RAM except itself to avoid always having infinite ratios.

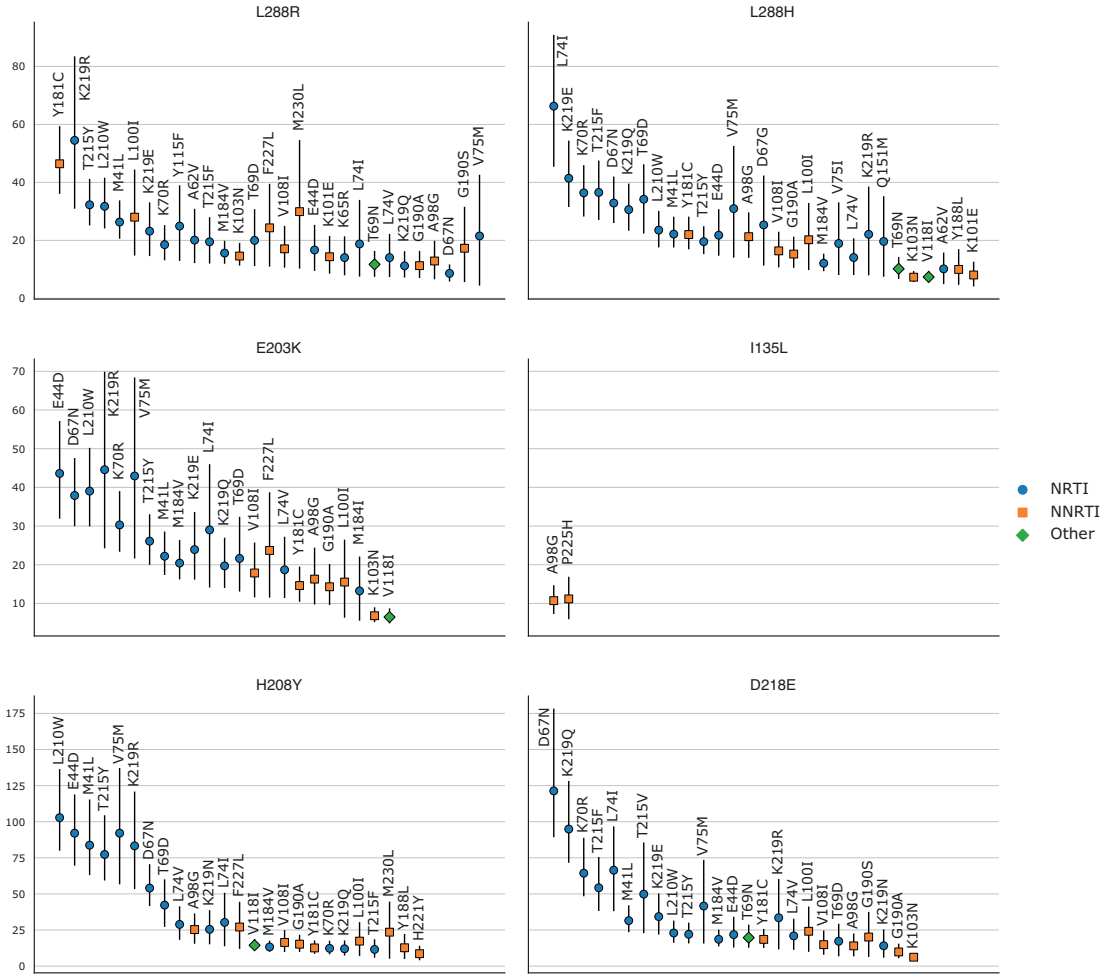

Figure 6.3.: **Relative risk of the new mutations with regards to known RAMs on the UK dataset**
(i.e. the prevalence of the new mutation in sequences with a given known RAM divided by the prevalence of the new mutation in sequences without this RAM). RRs were only computed for mutations (new and RAMs) 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.

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 B.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^{th} and 95^{th} 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 [269]. 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 [270]. 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 h for RRs computed on the African dataset (B.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 B.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

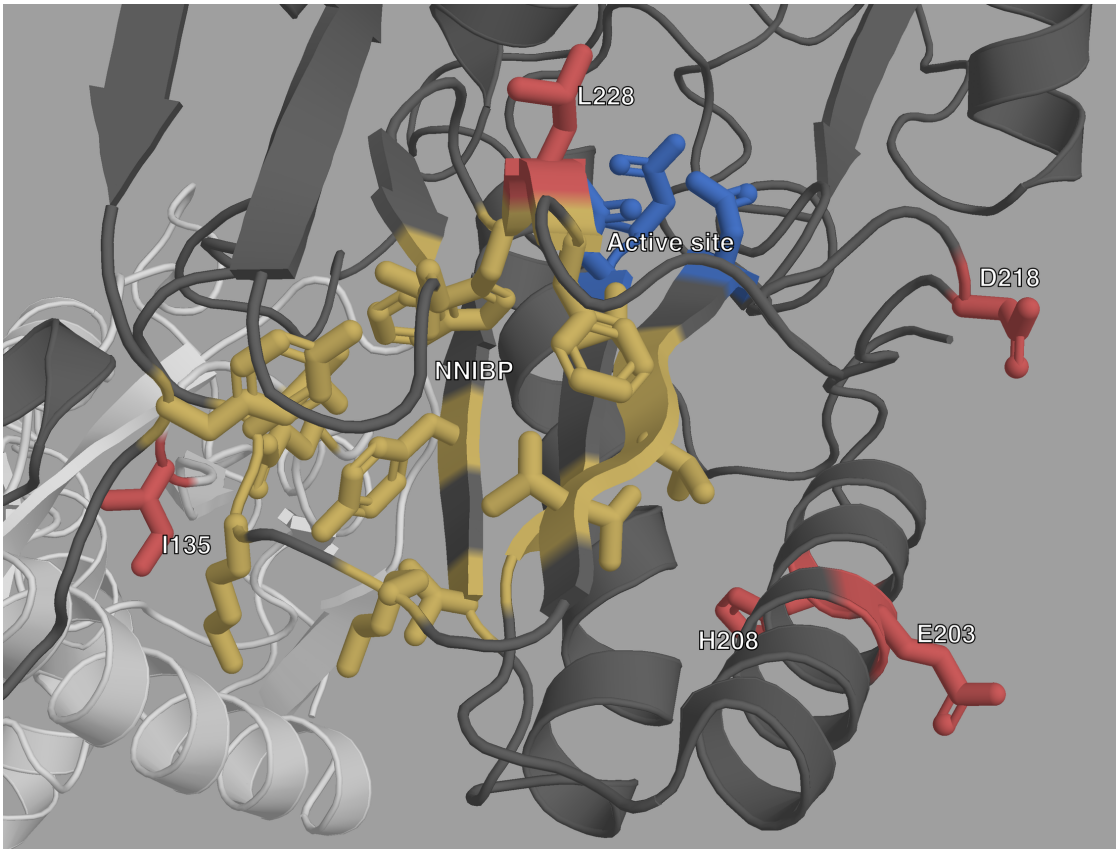

Figure 6.4.: **Structure of HIV-1 RT with highlighted important sites.** The p66 subunit is colored dark gray and the p51 subunit white. The active site is highlighted in blue, and the NNIBP is highlighted in yellow. The sites of new mutations are colored in red.

Histidine (H) are both less hydrophobic that 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 B.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 B.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 [271] and were suggested to be associated with reduced susceptibility to didanosine [272, 273]. I135L has been observed in sequences with reduced susceptibility to NNRTIs [274]. H208Y has been associated with NNRTI and NRTI resistance [275] and it has been suggested that it has an accessory role in NRTI resistance [276]. E203K, D218E, L228RH and H208Y have all been mentioned in [277] 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,

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 [278]. 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 B.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:

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).

CHAPTER 6

Supporting Information

Supporting Information can be found in the appendix B

References for chapter 6

[122] 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. 34, 92, 127).

[234] 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 (cit. on p. 85).

[235] Chris Verhofstede, Filip Van Wanzeele, 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 (cit. on p. 85).

[236] 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 (cit. on p. 85).

[237] 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 (cit. on p. 85).

[238] 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 (cit. on p. 85).

[239] 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 (cit. on p. 85).

[240] 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. 85).

[241] 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. 85).

[242] 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. 85).

[243] 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 (cit. on pp. 85–88, 90, 91, 95).

[244] 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 (cit. on p. 86).

[245] 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 (cit. on p. 86).

[246] 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 (cit. on p. 86).

[247] 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 (cit. on p. 86).

[248] 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 (cit. on p. 86).

[249] 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 (cit. on p. 86).

[250] 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 (cit. on p. 86).

[251] 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 (cit. on p. 86).

[252] 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 (cit. on p. 86).

[253] 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 (cit. on p. 86).

[254] 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 (cit. on p. 86).

[255] 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 (cit. on p. 86).

[256] 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 (cit. on p. 86).

[257] 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 (cit. on p. 86).

[258] 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 (cit. on p. 88).

[259] 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 (cit. on pp. 88, 90).

[260] 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. 88, 92).

[261] 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 (cit. on pp. 88, 91).

[262] 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 (cit. on p. 90).

[263] 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. 90).

[264] Leo Breiman. “Random Forests”. en. In: *Machine Learning* 45.1 (Oct. 2001), pp. 5–32. ISSN: 1573-0565. DOI: 10.1023/A:1010933404324 (cit. on p. 90).

[265] 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. 91).

[266] 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. 91).

[267] 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. 91).

[268] 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 (cit. on pp. 92, 126).

[269] Schrödinger, LLC. “The PyMOL Molecular Graphics System, Version 1.8”. Nov. 2015 (cit. on p. 100).

[270] 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 (cit. on p. 100).

[271] 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 (cit. on p. 103).

[272] 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 (cit. on p. 103).

[273] 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. 103).

[274] 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. 103).

[275] 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 (cit. on p. 103).

[276] 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 (cit. on p. 103).

[277] 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 (cit. on p. 103).

[278] 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 (cit. on p. 104).

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 an 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

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 mehods
  - 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

- 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` [229] and assembled using a centromere-tailored pipeline [279]. 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` [95] and `winnowmap2` [219]. 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.

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

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

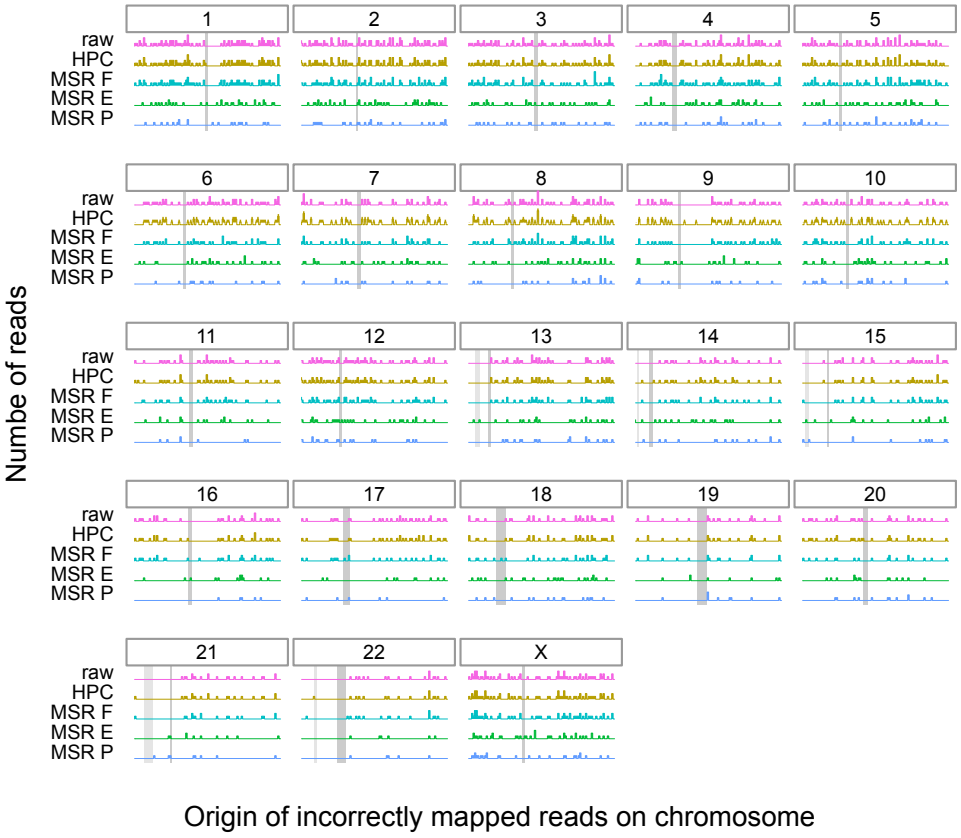

Figure A.1.: **Histogram of the original simulated positions for the incorrectly mapped reads using `minimap2` at high mapping qualities across the whole human genome, for several transformation methods.**
For a given chromosome, each row represents the number of simulated reads starting at that particular 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 ≥ 50 were considered for MSRs E, F and P.

A.4. Analyzing read origin on whole human genome

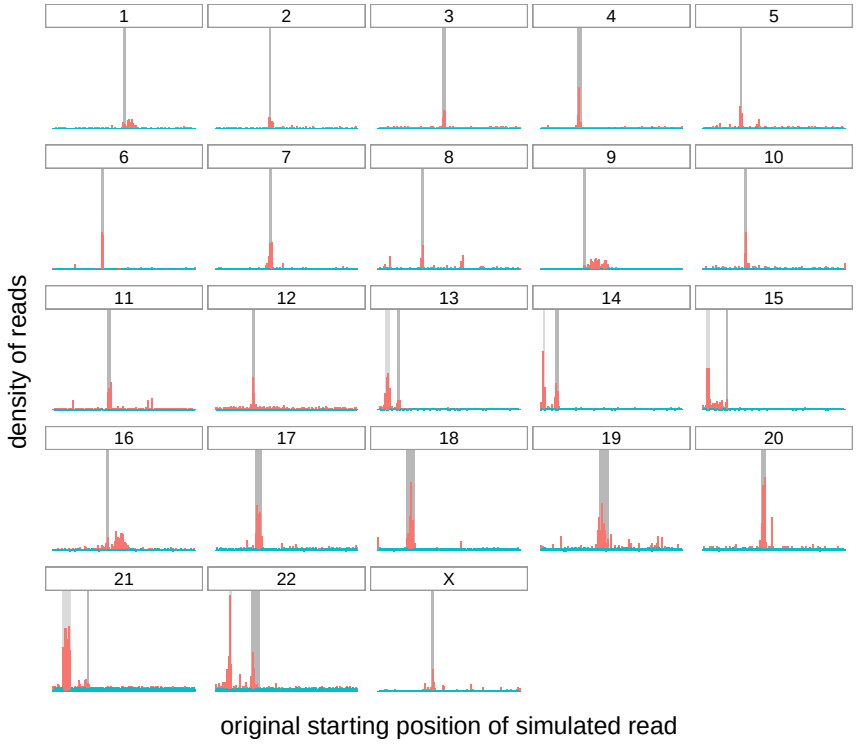

Figure A.2.: **Origin of correctly and incorrectly mapped raw reads**
Distribution of the origin of correctly and incorrectly mapped simulated reads (in teal and red respectively) on the different chromosomes of the whole human genome. The dark grey rectangle for each chromosome represents the centromere of that chromosome. The lighter gray rectangle on chromosomes 13, 14, 15, 21 and 22 correspond to satellites denoted as “stalk”, another repetitive region.

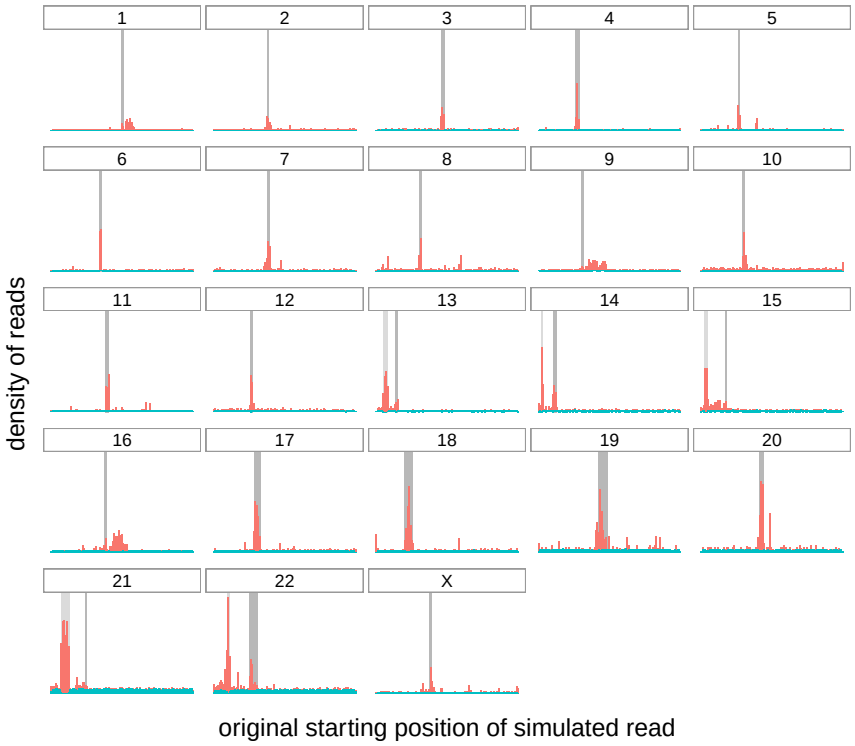

Figure A.3.: **Origin of correctly (teal) and incorrectly (red) mapped reads, transformed with HPC**
Distribution of the origin of correctly and incorrectly mapped simulated reads (in teal and red respectively) on the different chromosomes of the whole human genome. The dark grey rectangle for each chromosome 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.

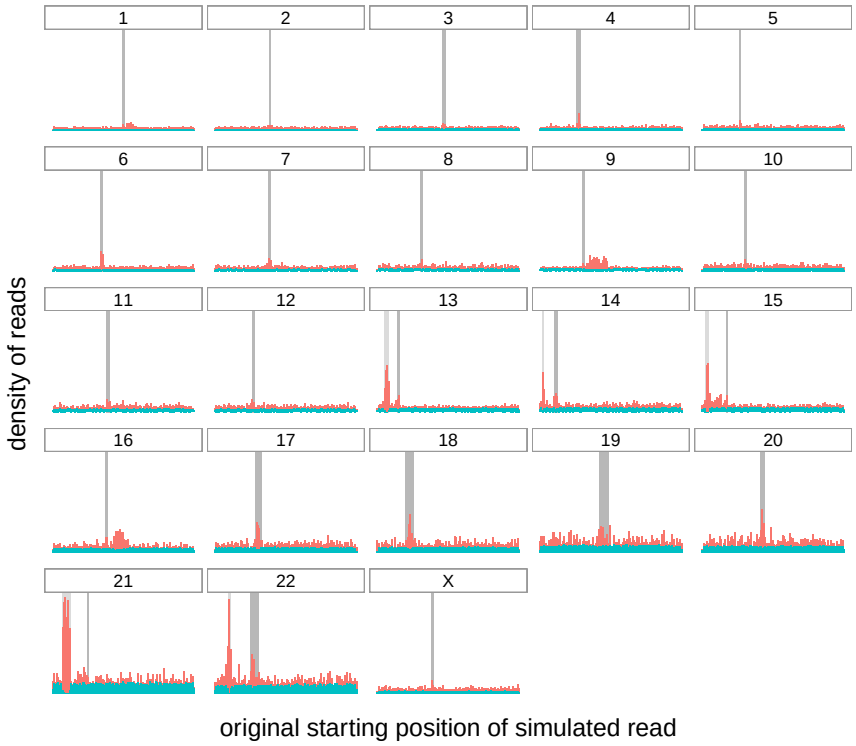

Figure A.4.: **Origin of correctly (teal) and incorrectly (red) mapped reads, transformed with $\mathbf{MSR_E}$**
Distribution of the origin of correctly and incorrectly mapped simulated reads (in teal and red respectively) on the different chromosomes of the whole human genome. The dark grey rectangle 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.

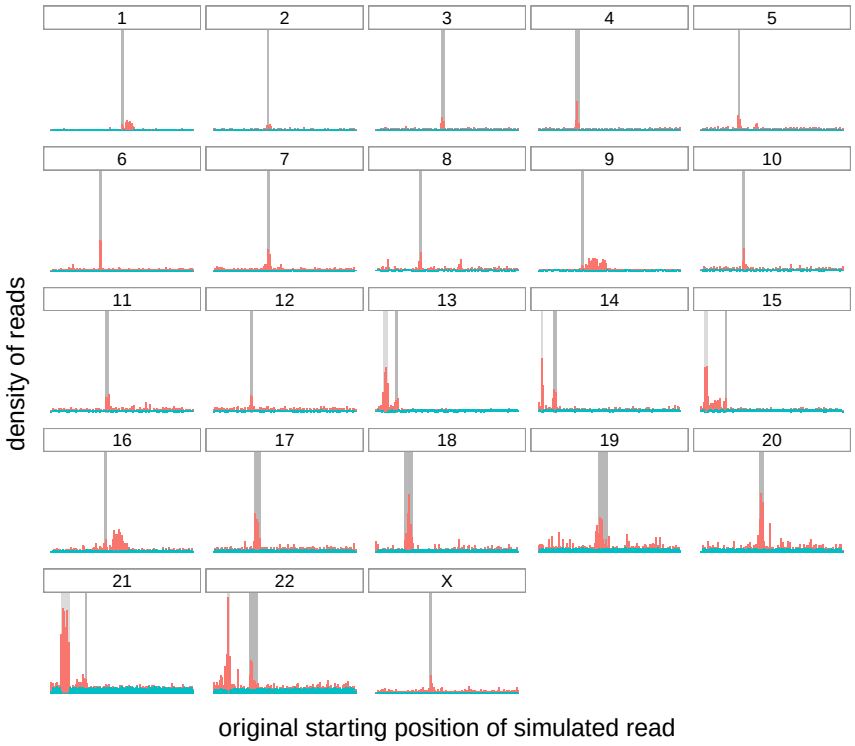

Figure A.5.: **Origin of correctly (teal) and incorrectly (red) mapped reads, transformed with $\mathbf{MSR_P}$**
Distribution of the origin of correctly and incorrectly mapped simulated reads (in teal and red respectively) on the different chromosomes of the whole human genome. The dark grey rectangle 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.

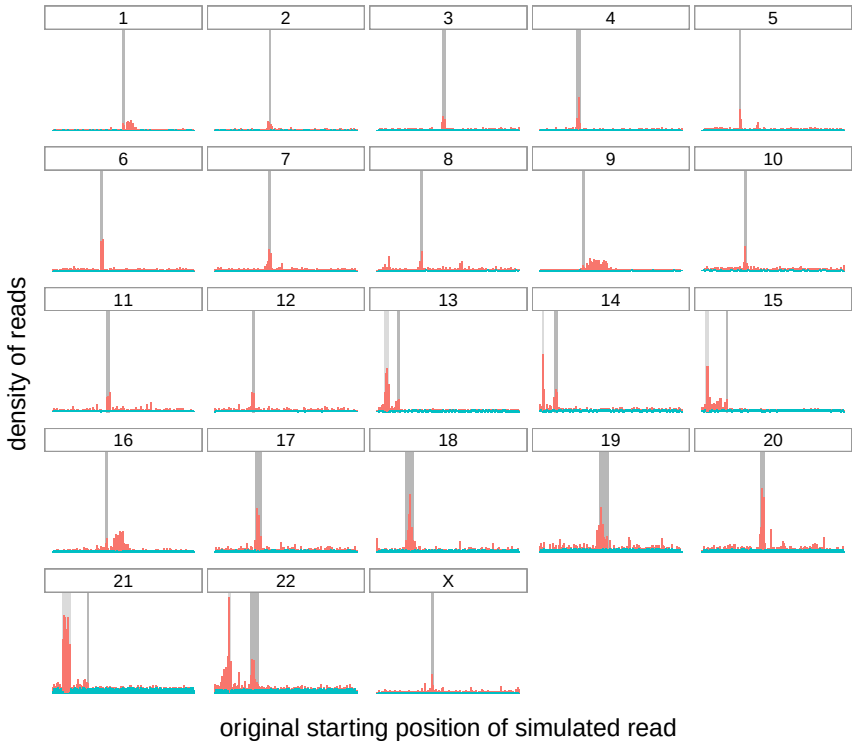

Figure A.6.: **Origin of correctly (teal) and incorrectly (red) mapped reads, transformed with $\mathbf{MSR_F}$**
Distribution of the origin of correctly and incorrectly mapped simulated reads (in teal and red respectively) on the different chromosomes of the whole human genome. The dark grey rectangle 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

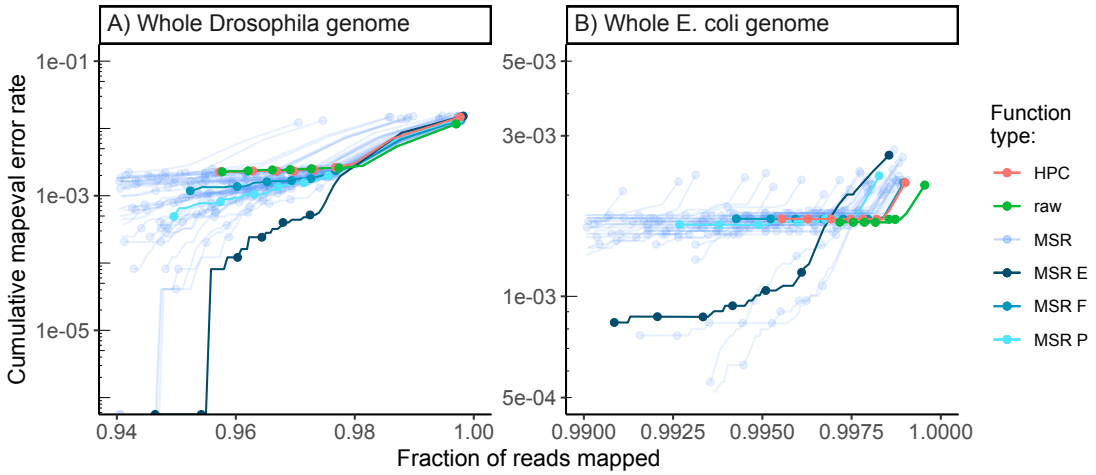

Figure A.7.: **Results of the `paftools mapeval` evaluation on reads simulated and mapped to whole *Drosophila melanogaster* and *Escherichia coli* (Genbank ID U00096.2) genomes.**
MSRs E, F and P are shown in different shades of blue to differentiate them from other MSRs. Reads were simulated with `nanosim`, and mapped with `minimap2`.

References for appendix A

[95] 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 (cit. on pp. 21, 39, 59, 66, 67, 117).

[219] 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 (cit. on pp. 43, 69, 117).

[229] 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 (cit. on pp. 66, 115).

[279] 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 (cit. on p. 115).

B. Supporting Information for "Using Machine Learning and Big Data to Explore the Drug Resistance Landscape in HIV"

B.1. S1 Appendix (Technical appendix).

B.1.1. Data

B.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.hivrdb@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. 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

B.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 [280] 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 [281].

B.1.2. Classifiers

We used classifier implementations from the scikit-learn python library [282], `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 [283], and corrected p-values for multiple testing with the statsmodels python library [284] using the `"Bonferroni"` method.

B.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 [268].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

false positives and false negatives respectively:

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

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 [122]. 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 [285].

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

MI was used to compute the G statistic, which follows the chi-square distribution under the null hypothesis [286]. 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:

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

We used the following implementations from the scikit-learn python library [282] 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 B

RAM.

$$RR(new, X) = \frac{prevalence\,(new\ mutation \mid X = 1)}{prevalence\,(new\ mutation \mid X = 0)}$$

$$= \frac{|(new = 1) \cap (X = 1)|}{|(X = 1)|} \div \frac{|(new = 1) \cap (X = 0)|}{|(X = 0)|}$$

B.2. S1 Fig.

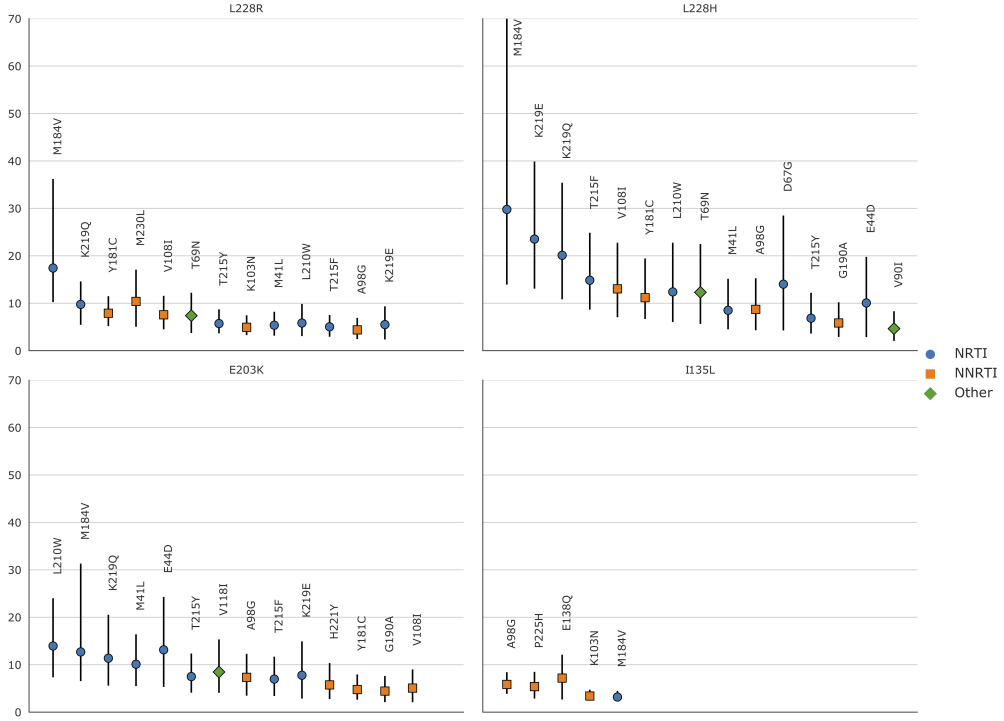

Figure B.1.: **Relative risks of the new mutations with regards to known RAMs on the African dataset**
(i.e. the prevalence of the new mutation in sequences with a given RAM divided by the prevalence of the new mutation in sequences without the RAM). RRs were only computed for mutations (new and RAMs) 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.

B.3. S2 Fig.

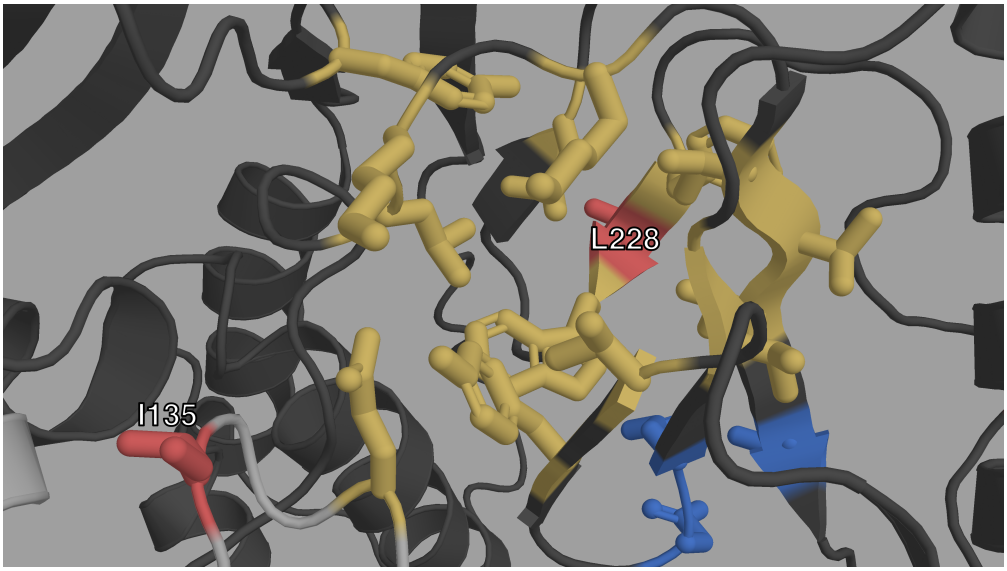

Figure B.2.: **Closeup structural view of the entrance of the NNIBP of HIV-1 RT**
The p66 subunit is colored in dark gray, the p51 subunit in light gray. The NNIBP is highlighted in yellow. The active site is colored in blue. We can see the physical proximity of I135 (red) to the entrance of the NNIBP. We can also see how L228 (red) is between 2 AAs of the NNIBP.

B.4. S3 Fig.

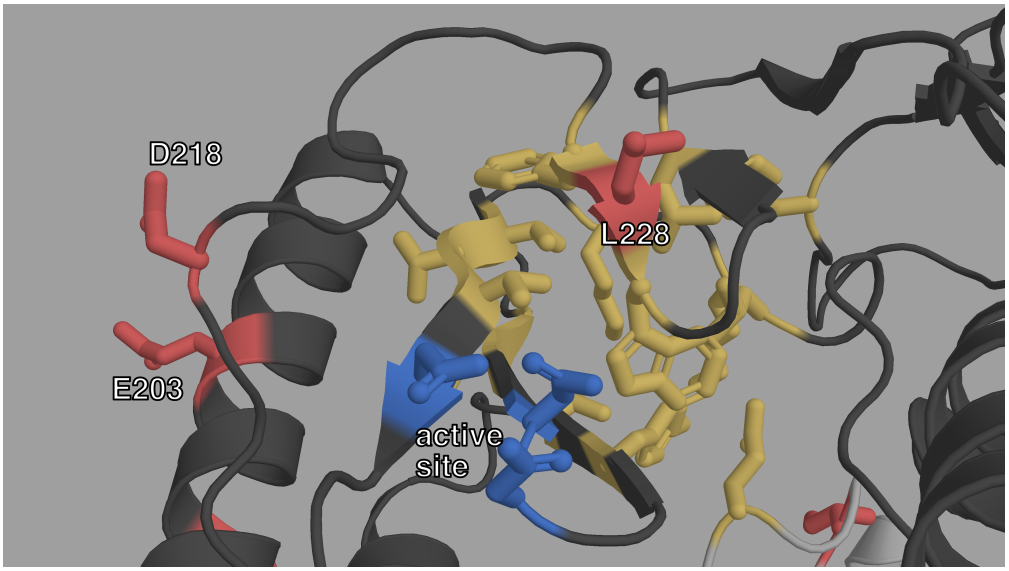

Figure B.3.: **Closeup structural view of the active site of HIV-1 RT.** The p66 subunit is colored in dark gray, the p51 subunit in light gray. The active site is highlighted in blue. The NNIBP is colored in yellow. L228, E203 and D218 (red) are also very close on either side of the active site.

B.5. S1 Table.

	rank		codon distance			UK			Africa			p-value	B62	Dayhoff category shift	Change in net charge	polarity	hydrophobicity index	molecular weight
	T/N	W/W	min	UK	Africa	count	ratio $\rho(new, treatment)$	$\rho(new, with\ RAM)$	count	ratio $\rho(new, treatment)$	$\rho(new, with\ RAM)$							
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;∞]	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%)	∞ [∞;∞]	∞ [∞;∞]	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;∞]	34.1 [12.0;∞]	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%)	∞ [∞;∞]	∞ [∞;∞]	7.3E-05	2	d → f	0	-4.2	1.27	26.03

Table B.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 wil-type amino acid from the mutated amino acid. All values were obtained from the AAindex database [287]

B.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 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.

B.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)

B.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/) and (hivdb.stanford.edu/dr-summary/comments/NNRTI/).

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

References for Appendix B

[122] 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. 34, 92, 127).

[268] 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 (cit. on pp. 92, 126).

[280] 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 (cit. on p. 126).

[281] Will McGinnis, Hbghhy, et al. *Scikit-Learn-Contrib/Categorical-Encoding: Release For Zenodo*. Zenodo. Jan. 2018. DOI: 10.5281/ZENODO.1157110 (cit. on p. 126).

[282] 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. 126, 127).

[283] 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 (cit. on p. 126).

[284] 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 (cit. on p. 126).

[285] 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 (cit. on p. 127).

[286] 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 (cit. on p. 127).

[287] 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. URL: https://academic.oup.com/nar/article/36/suppl_1/D202/2508449 (cit. on p. 133).

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= watsonSTRUCTUREDNA1953 PMID: 13168976 publisher: Cold Spring Harbor Laboratory Press, pp. 123–131. DOI: 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. URL: https://www.nature.com/articles/265687a0 (visited on 05/17/2022) (cit. on pp. 11, 15).

[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. 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. URL: https://www.science.org/doi/10.1126/science.abj6987 (cit. on pp. 11, 59, 66).

[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. URL: https://doi.org/10.1111/j.1095-8339.2010.01072.x (cit. on p. 11).

[6] H. C. Macgregor. "C-Value Paradox". In: ed. by Sydney Brenner and Jefferey H. Miller. DOI: 10.1006/rwgn.2001.0301. New York: Academic Press, Jan. 1, 2001, pp. 249–250. DOI: 10.1006/rwgn.2001.0301. URL: https://www.sciencedirect.com/science/article/pii/B0122270800003013 (cit. on p. 11).

[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. 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. 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. 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. 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. 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. 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. 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. 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. 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. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5650487/ (cit. on p. 13).

[19] 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. URL: https://www.sciencedirect.com/science/article/pii/S1687850715000333 (cit. on p. 13).

[20] 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 (cit. on p. 13).

[21] 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. 13).

[22] 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. 13).

[23] 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. 13).

[24] 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. 13).

[25] 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).

[26] 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).

[27] 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).

[28] 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).

[29] 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. URL: https://academic.oup.com/nar/article/31/1/298/2401450 (cit. on p. 14).

[30] 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. URL: https://www.sciencedirect.com/science/article/pii/S1198743X14615500 (cit. on p. 14).

[31] 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. URL: https://www.pnas.org/doi/abs/10.1073/pnas.74.12.5463 (visited on 05/17/2022) (cit. on p. 14).

[32] 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. URL: https://doi.org/10.1093/nar/13.7.2399 (visited on 05/17/2022) (cit. on p. 15).

[33] 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. URL: https://www.nature.com/articles/321674a0 (visited on 05/17/2022) (cit. on p. 15).

[34] 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. URL: https://www.nature.com/articles/nbt1486 (cit. on p. 15).

[35] *The Cost of Sequencing a Human Genome*. URL: https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost (cit. on p. 15).

[36] 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. URL: https://www.science.org/doi/10.1126/science.1084564 (cit. on p. 15).

[37] Lin Liu, Yinhu 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. URL: https://www.hindawi.com/journals/bmri/2012/251364/ (visited on 05/16/2022) (cit. on pp. 15, 16).

[38] 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. URL: https://www.nature.com/articles/nrg2626 (visited on 05/16/2022) (cit. on p. 15).

[39] 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. URL: https://www.nature.com/articles/nature09796 (cit. on p. 16).

[40] 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. URL: https://www.science.org/doi/10.1126/science.1162986 (cit. on p. 16).

[41] 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. URL: https://www.sciencedirect.com/science/article/pii/S1672022915001345 (cit. on p. 16).

[42] 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. URL: https://www.nature.com/articles/nature13907 (cit. on p. 16).

[43] 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. URL: https://www.nature.com/articles/s41576-020-0236-x (visited on 05/16/2022) (cit. on pp. 16–19).

[44] 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. URL: https://doi.org/10.1093/gigascience/giaa146 (cit. on pp. 16, 17).

[45] 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. URL: https://www.nature.com/articles/nnano.2009.12 (cit. on p. 17).

[46] 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. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4722697/ (cit. on p. 17).

[47] 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. URL: https://www.nature.com/articles/nbt.4060 (cit. on p. 17).

[48] *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. 17).

[49] 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. URL: https://doi.org/10.1093/bioinformatics/bty841 (cit. on p. 17).

[50] 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. URL: https://doi.org/10.1186/s13059-016-1103-0 (cit. on p. 17).

[51] 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. URL: https://www.pnas.org/doi/abs/10.1073/pnas.83.17.6233 (cit. on p. 17).

[52] Bryan John Smith. *Protein Sequencing Protocols*. Springer Science & Business Media, 2002 (cit. on p. 17).

[53] 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. URL: https://www.nature.com/articles/s41565-018-0236-6 (cit. on p. 17).

[54] 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. URL: https://doi.org/10.1186/s13059-019-1659-6 (cit. on p. 17).

[55] 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. URL: https://doi.org/10.1093/bib/bbz058 (cit. on p. 18).

[56] 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. URL: https://doi.org/10.1186/s13059-020-1935-5 (visited on 05/16/2022) (cit. on pp. 18, 20).

[57] 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. URL: https://www.nature.com/articles/s41592-019-0669-3 (cit. on pp. 18, 21).

[58] 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. URL: https://genome.cshlp.org/content/27/5/722 (cit. on p. 18).

[59] 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. URL: https://www.biorxiv.org/content/10.1101/106252v1 (cit. on p. 18).

[60] 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. URL: https://doi.org/10.1093/bioinformatics/btu392 (cit. on pp. 18, 34).

[61] 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. URL: https://doi.org/10.1186/s13015-016-0075-7 (cit. on p. 18).

[62] 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. URL: https://www.nature.com/articles/nbt.2280 (cit. on pp. 18, 34).

[63] 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. URL: https://doi.org/10.1093/bioinformatics/btu538 (cit. on p. 18).

[64] 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. 18).

[65] 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. URL: https://www.nature.com/articles/nmeth.4184 (cit. on pp. 18, 20).

[66] 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. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0112963 (cit. on p. 18).

[67] 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 (cit. on p. 18).

[68] 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. URL: https://doi.org/10.1186/s13059-018-1605-z (cit. on p. 18).

[69] 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. URL: https://doi.org/10.1186/s12864-020-07227-0 (cit. on p. 18).

[70] 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. URL: http://www.nature.com/articles/s41587-019-0217-9 (cit. on p. 18).

[71] 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. URL: https://www.nature.com/articles/s41592-020-01041-y. (cit. on p. 18).

[72] 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. URL: https://doi.org/10.1093/bioinformatics/btab528 (cit. on p. 18).

[73] 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. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0178751 (cit. on p. 18).

[74] 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. URL: https://doi.org/10.1186/s13059-019-1727-y (cit. on pp. 18, 20).

[75] 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. URL: https://www.nature.com/articles/nbt.3982 (cit. on p. 18).

[76] *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 (cit. on p. 18).

[77] 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. URL: https://doi.org/10.1186/s12864-018-4544-x (cit. on p. 19).

[78] A. Sina Booeshaghi and Lior Pachter. “Pseudoalignment facilitates assignment of error-prone Ultima Genomics reads”. In: (). DOI: 10.1101/2022.06.04.494845 (cit. on p. 19).

[79] 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. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0257521 (cit. on p. 19).

[80] 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. URL: https://genome.cshlp.org/content/25/11/1750 (cit. on p. 19).

[81] 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 (cit. on pp. 19, 58).

[82] 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. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5553090/ (cit. on p. 19).

[83] 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. URL: https://www.nature.com/articles/s41587-021-01049-5 (cit. on p. 19).

[84] 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. URL: https://doi.org/10.1186/s13059-021-02282-6 (cit. on p. 20).

[85] 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. URL: https://doi.org/10.1186/s13059-018-1462-9 (cit. on p. 20).

[86] Peter Sarkozy, Ákos Jobbágy, 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 (cit. on p. 20).

[87] *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. 20).

[88] 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. URL: https://avs.scitation.org/doi/full/10.1116/1.5116295 (cit. on p. 20).

[89] 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. URL: https://pubs.rsc.org/en/content/articlelanding/2018/nr/c8nr04238a (cit. on p. 20).

[90] 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. URL: https://www.pnas.org/doi/full/10.1073/pnas.1802640115 (cit. on p. 20).

[91] 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. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/1755-0998.12890 (cit. on p. 20).

[92] 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 (cit. on p. 20).

[93] 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 (cit. on pp. 21, 59).

[94] 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. pmid: 32801147 (cit. on pp. 21, 59).

[95] 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 (cit. on pp. 21, 39, 59, 66, 67, 117).

[96] 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. URL: https://www.nature.com/articles/s41587-020-0503-6 (cit. on p. 21).

[97] 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 (cit. on pp. 21, 59).

[98] 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 (cit. on pp. 21, 59).

[99] 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 (cit. on pp. 21, 59).

[100] Wing-Kin Sung. *Algorithms in Bioinformatics: A Practical Introduction*. DOI: 10.1201/9781420070347. New York: Chapman and Hall/CRC, Oct. 10, 2011. DOI: 10.1201/9781420070347 (cit. on pp. 33–35, 39, 41).

[101] Richard Wesley Hamming. *Coding and Information Theory*. tex.ids= hamming1980coding googlebooksid: ed5QAAAAMAAJ lccn: 79015159. Prentice-Hall, 1980 (cit. on p. 33).

[102] 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. 33, 39, 42, 58).

[103] 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. 33).

[104] 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. 33).

[105] 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. 34).

[106] 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. 34).

[107] 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. 34).

[108] 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. 34).

[109] 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. 34).

[110] 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. 34).

[111] 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. 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. 34).

[112] 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. URL: https://www.sciencedirect.com/science/article/pii/S0959440X05000825 (cit. on p. 34).

[113] 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. URL: https://www.nature.com/articles/nrm2281 (cit. on p. 34).

[114] 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. URL: https://doi.org/10.1093/bioinformatics/btr170 (cit. on p. 34).

[115] 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. URL: https://www.nature.com/articles/nmeth.1374 (cit. on p. 34).

[116] 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. URL: https://doi.org/10.1186/s13059-019-1828-7 (cit. on p. 34).

[117] 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. URL: http://www.sciencedirect.com/science/article/pii/0022283670900574 (cit. on p. 34).

[118] 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. URL: https://www.sciencedirect.com/science/article/pii/0022283681900875 (cit. on p. 34).

[119] 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. 34).

[120] 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. URL: https://www.ams.org/bull/1954-60-06/S0002-9904-1954-09848-8/ (cit. on p. 34).

[121] 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. URL: https://www.sciencedirect.com/science/article/pii/0022000080900021 (cit. on p. 34).

[122] 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. 34, 92, 127).

[123] 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. URL: https://doi.org/10.1145/321921.321922 (cit. on p. 34).

[124] 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. URL: https://doi.org/10.1145/360825.360861 (cit. on p. 35).

[125] 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. URL: https://doi.org/10.1093/bioinformatics/4.1.11 (cit. on p. 35).

[126] 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 p. 35).

[127] 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. URL: https://www.sciencedirect.com/science/article/pii/019688589190017D (cit. on p. 35).

[128] 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. URL: https://www.sciencedirect.com/science/article/pii/0022283687904785 (cit. on p. 35).

[129] 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. URL: https://genome.cshlp.org/content/12/10/1611 (cit. on p. 35).

[130] 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. URL: https://doi.org/10.1186/gb-2004-5-10-r80 (cit. on p. 35).

[131] 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. URL: https://doi.org/10.1186/s12859-016-0930-z (cit. on p. 35).

[132] 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. URL: http://journals.pan.pl/dlibra/publication/96876/edition/83624/content (cit. on p. 35).

[133] Stephen F Altschul. “Substitution Matrices”. In: John Wiley & Sons, Ltd, 2013. DOI: 10.1002/9780470015902.a0005265.pub3. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470015902.a0005265.pub3 (cit. on p. 35).

[134] 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. 35).

[135] 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 (cit. on p. 35).

[136] 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. URL: https://www.pnas.org/content/89/22/10915 (cit. on p. 35).

[137] 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 (cit. on p. 35).

[138] 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. URL: https://doi.org/10.1093/molbev/msn067 (cit. on p. 35).

[139] 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. URL: https://doi.org/10.1093/bioinformatics/17.suppl_1.S182 (cit. on p. 35).

[140] 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 (cit. on p. 35).

[141] 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. URL: https://www.nature.com/articles/s41598-019-52532-8 (cit. on p. 36).

[142] 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. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.21775 (cit. on p. 36).

[143] 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 (cit. on p. 36).

[144] 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. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1876811/ (cit. on p. 36).

[145] 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 (cit. on p. 36).

[146] 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. URL: https://www.worldscientific.com/doi/abs/10.1142/9789812799623_0012 (cit. on p. 36).

[147] 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 (cit. on p. 36).

[148] 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. URL: https://doi.org/10.1093/molbev/msl175 (cit. on p. 36).

[149] 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. URL: https://www.sciencedirect.com/science/article/pii/0022283682903989 (cit. on p. 36).

[150] 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. URL: https://www.sciencedirect.com/science/article/pii/S0022283683711058 (cit. on p. 36).

[151] 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. URL: https://doi.org/10.1007/BF00164032 (cit. on p. 36).

[152] 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. URL: https://www.sciencedirect.com/science/article/pii/S0022519384800375 (cit. on p. 36).

[153] 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. Academic Press, Jan. 1, 1992, pp. 575–601. DOI: 10.1016/0076-6879(92)10029-D. URL: https://www.sciencedirect.com/science/article/pii/007668799210029D (cit. on p. 36).

[154] 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. URL: https://www.mdpi.com/2218-273X/12/4/546 (cit. on pp. 36, 38).

[155] 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. URL: https://epubs.siam.org/doi/abs/10.1137/0149094 (cit. on p. 36).

[156] 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. URL: https://doi.org/10.1093/nar/12.1Part1.175 (cit. on p. 36).

[157] 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. 36).

[158] 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. 36).

[159] 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. 36, 42).

[160] 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. 36).

[161] 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. 37, 40).

[162] 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. 37).

[163] 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/bbq015. URL: https://doi.org/10.1093/bib/bbq015 (visited on 05/16/2022) (cit. on p. 37).

[164] 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. 37).

[165] 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. 38).

[166] 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. URL: https://www.sciencedirect.com/science/article/pii/S088875439690614X (cit. on p. 38).

[167] 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 (cit. on p. 38).

[168] 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 (cit. on p. 38).

[169] 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. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5013910/ (cit. on p. 39).

[170] 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. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4702930/ (cit. on p. 39).

[171] 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. 39).

[172] 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. URL: https://pnas.org/doi/full/10.1073/pnas.87.6.2264 (cit. on p. 39).

[173] Richard Mott. "Alignment: Statistical Significance". In: _eprint: https://onlinelibrary.wiley.com/doi/pdf/1 DOI: 10.1038/npg.els.0005264. John Wiley & Sons, Ltd, 2005. DOI: 10.1038/npg.els.0005264. URL: https://onlinelibrary.wiley.com/doi/abs/10.1038/npg.els.0005264 (cit. on p. 39).

[174] 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 (cit. on p. 39).

[175] 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. URL: https://www.nature.com/articles/s41592-021-01101-x (cit. on p. 39).

[176] 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. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC430967/ (cit. on p. 39).

[177] 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. 39).

[178] 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 (cit. on p. 39).

[179] Michael Burrows and David Wheeler. *A Block-Sorting Lossless Data Compression Algorithm*. Tech. rep. 1994 (cit. on p. 39).

[180] 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. URL: https://doi.org/10.1093/bioinformatics/btn032 (cit. on p. 39).

[181] 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 (cit. on p. 39).

[182] 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. URL: https://doi.org/10.1093/bioinformatics/btp324 (cit. on p. 39).

[183] 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. URL: https://doi.org/10.1093/bioinformatics/btp698 (cit. on p. 39).

[184] Heng Li. *Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM*. May 26, 2013. arXiv: 1303.3997 [q-bio] (cit. on p. 39).

[185] 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. URL: https://doi.org/10.1093/bioinformatics/bts414 (cit. on p. 39).

[186] 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. URL: http://link.springer.com/10.1007/978-1-62703-646-7 (cit. on p. 40).

[187] 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. URL: https://www.liebertpub.com/doi/abs/10.1089/cmb.1994.1.337 (cit. on p. 40).

[188] 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. URL: https://www.liebertpub.com/doi/abs/10.1089/106652701753307511 (cit. on p. 40).

[189] Da-Fei Feng and Russell F. Doolittle. “Progressive sequence alignment as a prerequisitetto correct phylogenetic trees”. In: *Journal of Molecular Evolution* 25.4 (Aug. 1, 1987), pp. 351–360. DOI: 10.1007/BF02603120. URL: https://doi.org/10.1007/BF02603120 (cit. on p. 40).

[190] 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. URL: https://doi.org/10.1093/bioinformatics/8.3.275 (cit. on p. 40).

[191] 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. URL: https://www.pnas.org/doi/abs/10.1073/pnas.83.14.5155 (cit. on p. 40).

[192] 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. URL: https://www.sciencedirect.com/science/article/pii/0022283689902349 (cit. on p. 40).

[193] 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. URL: https://doi.org/10.1093/bioinformatics/bth090 (cit. on p. 40).

[194] 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. URL: https://doi.org/10.1093/bioinformatics/14.5.407 (cit. on p. 40).

[195] 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. URL: https://academic.oup.com/nar/article/22/22/4673/2400290 (cit. on p. 40).

[196] 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. URL: https://doi.org/10.1093/nar/25.24.4876 (cit. on p. 40).

[197] Cédric Notredame, Desmond G Higgins, and Jaap Heringa. "T-coffee: a novel method for fast and accurate multiple sequence alignment11Edited by J. Thornton". In: *Journal of Molecular Biology* 302.1 (Sept. 8, 2000), pp. 205–217. DOI: 10.1006/jmbi.2000.4042. URL: http://www.sciencedirect.com/science/article/pii/S0022283600940427 (cit. on p. 40).

[198] 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. URL: https://doi.org/10.1186/1471-2105-5-113 (cit. on p. 40).

[199] 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. URL: https://doi.org/10.1093/nar/gkh340 (cit. on p. 40).

[200] 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. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC546535/ (cit. on p. 40).

[201] 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. URL: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.0030123 (cit. on p. 42).

[202] 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. URL: https://www.futuremedicine.com/doi/abs/10.1517/14622416.3.1.131 (visited on 05/16/2022) (cit. on p. 42).

[203] 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. URL: https://www.sciencedirect.com/science/article/pii/S0959440X06000704 (visited on 05/16/2022) (cit. on p. 42).

[204] 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. URL: https://doi.org/10.1186/1748-7188-9-4 (cit. on p. 42).

[205] 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. URL: https://academic.oup.com/bioinformatics/article/15/1/87/218377 (cit. on p. 42).

[206] Sean R Eddy. "Multiple Alignment Using Hidden Markov Models". In: (), p. 7 (cit. on p. 42).

[207] Frédéric Lemoine, Luc Blassel, 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. URL: https://doi.org/10.1093/bioinformatics/btaa871 (cit. on p. 42).

[208] 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. URL: https://joss.theoj.org/papers/10.21105/joss.03773 (cit. on p. 42).

[209] 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. URL: https://doi.org/10.1093/bioinformatics/10.4.419 (cit. on p. 42).

[210] 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. URL: https://doi.org/10.1093/bioinformatics/9.3.267 (cit. on p. 42).

[211] 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. URL: http://www.sciencedirect.com/science/article/pii/S0888754317300551 (cit. on p. 42).

[212] 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. 42).

[213] 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. URL: https://www.nature.com/articles/s41587-019-0333-6 (cit. on p. 42).

[214] 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. URL: https://doi.org/10.1186/s13059-021-02443-7 (cit. on p. 42).

[215] 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. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3375638/ (cit. on p. 42).

[216] 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. URL: https://doi.org/10.1186/1471-2105-14-184 (cit. on p. 42).

[217] 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 (cit. on p. 42).

[218] 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.2012.36 (cit. on p. 42).

[219] 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 (cit. on pp. 43, 69, 117).

[220] 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. URL: https://www.nature.com/articles/s41592-022-01457-8 (cit. on p. 43).

[221] 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 (cit. on pp. 43, 66).

[222] 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. URL: https://doi.org/10.1186/s13059-017-1290-3 (cit. on p. 43).

[223] 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. URL: https://doi.org/10.1093/bioinformatics/btv524 (cit. on p. 43).

[224] 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 (cit. on p. 59).

[225] 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 (cit. on p. 59).

[226] 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 (cit. on p. 59).

[227] 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. 65).

[228] 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. pmid: 10731132 (cit. on p. 66).

[229] 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 (cit. on pp. 66, 115).

[230] 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 (cit. on p. 69).

[231] 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 (cit. on p. 73).

[232] Heng Li. *New Strategies to Improve Minimap2 Alignment Accuracy*. Aug. 7, 2021. arXiv: 2108.03515 [q-bio] (cit. on p. 73).

[233] 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 (cit. on p. 73).

[234] 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 (cit. on p. 85).

[235] Chris Verhofstede, Filip Van Wanzeele, 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 (cit. on p. 85).

[236] 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 (cit. on p. 85).

[237] 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 (cit. on p. 85).

[238] 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 (cit. on p. 85).

[239] 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 (cit. on p. 85).

[240] 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. 85).

[241] 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. 85).

[242] 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. 85).

[243] 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 (cit. on pp. 85–88, 90, 91, 95).

[244] 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 (cit. on p. 86).

[245] 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 (cit. on p. 86).

[246] 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 (cit. on p. 86).

[247] 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 (cit. on p. 86).

[248] 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 (cit. on p. 86).

[249] 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 (cit. on p. 86).

[250] 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 (cit. on p. 86).

[251] 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 (cit. on p. 86).

[252] 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 (cit. on p. 86).

[253] 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 (cit. on p. 86).

[254] 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 (cit. on p. 86).

[255] 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 (cit. on p. 86).

[256] 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 (cit. on p. 86).

[257] 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 (cit. on p. 86).

[258] 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 (cit. on p. 88).

[259] 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 (cit. on pp. 88, 90).

[260] 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. 88, 92).

[261] 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 (cit. on pp. 88, 91).

[262] 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 (cit. on p. 90).

[263] 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. 90).

[264] Leo Breiman. “Random Forests”. en. In: *Machine Learning* 45.1 (Oct. 2001), pp. 5–32. ISSN: 1573-0565. DOI: 10.1023/A:1010933404324 (cit. on p. 90).

[265] 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. 91).

[266] 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. 91).

[267] 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. 91).

[268] 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 (cit. on pp. 92, 126).

[269] Schrödinger, LLC. “The PyMOL Molecular Graphics System, Version 1.8”. Nov. 2015 (cit. on p. 100).

[270] 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 (cit. on p. 100).

[271] 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 (cit. on p. 103).

[272] 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 (cit. on p. 103).

[273] 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. 103).

[274] 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. 103).

[275] 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 (cit. on p. 103).

[276] 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 (cit. on p. 103).

[277] 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 (cit. on p. 103).

[278] 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 (cit. on p. 104).

[279] 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 (cit. on p. 115).

[280] 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 (cit. on p. 126).

[281] Will McGinnis, Hbghhy, et al. *Scikit-Learn-Contrib/Categorical-Encoding: Release For Zenodo*. Zenodo. Jan. 2018. DOI: 10.5281/ZENODO.1157110 (cit. on p. 126).

[282] 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. 126, 127).

[283] 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 (cit. on p. 126).

[284] 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 (cit. on p. 126).

[285] 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 (cit. on p. 127).

[286] 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 (cit. on p. 127).

[287] 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. URL: https://academic.oup.com/nar/article/36/suppl_1/D202/2508449 (cit. on p. 133).

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 consectetuer.