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 2022-05-18

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

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.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

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.

Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vitae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

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.

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

Goal of this Thesis

Thesis organization

TO COME

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.

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, creating the phosphate-sugar backbone of single strand of DNA. The ordered sequence of these four types of nucleotides in a single strand encodes all the genetic information necessary to life. Nucleotides can form strong complementary bonds with each other, A with T and C with G. These bonds allows two strands of DNA to link up with each other, forming the double-helix structure of DNA as 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.

1.1.2. What are proteins ?

- Proteins are the action molecules of living organisms
- They are obtained from DNA by translation where DNA -> RNA -> Protein
- Mutations in the DNA sequence can induce mutations in proteins:
  - illnesses

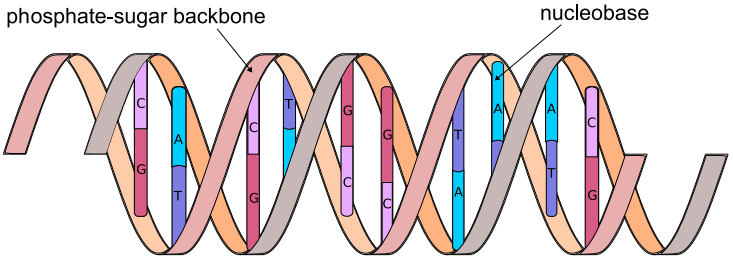

Figure 1.1.: **Double-helix structure of DNA**

– acquired traits

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 (Sanger, Nicklen, and Coulson 1977). Sanger *et al.* devised a simple method to read the sequence of nucleotides that make up a DNA sequence.

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 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 (Sanger, Air, et al. 1977). 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 (Smith, Fung, et al. 1985; Smith, Sanders, et al. 1986). The fluorescence also allowed for automated reading ***(base-calling ?)*** with optical systems.

I need to speak of performance / throughput of these methods here.

1.2.2. Next-generation sequencing

Developed to lower cost and more throughput

- Massively parallel
- Long reads sequencing

Quick summary on PacBio and ONT

As a conclusion, mention the work done on protein sequencing, but we usually get the protein sequence from the DNA sequence that is translated from codons.

1.3. Sequencing errors, how to account for them ?

1.3.1. What kind of errors happen ?

- Substitutions
- Insertions
- deletions

1.3.2. Long read errors

Mainly indels, in certain regions of the genomes, particularily homopolymers.

1.3.3. HPC

- HPC takes repeated runs of a single nucleotide and compresses them to a single occurence
- Empirically is has been shown to improve mapping and other applications

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.

2.1.1. How to align ?

Either decide to align globally or locally. Usually done with dynamic programming by assigning a score to mismatches, indels, … and then choosing the alignment with minimal score.

short example of local alignment

gap open, gap extend, etc…

2.1.2. Substitution models

mismatch depends on models, BLOSUM, PAML, …

2.1.3. Why align ?

- pairwise alignment to compare two similar sequences
- mapping to get an idea of where a short sequence comes from on a larger reference. (sequencing reads)

2.2. How do we speed up local alignment ?

2.2.1. Indexed … BLAST,…

short explanation of BLAST or suffix-arrays, …

2.2.2. Seeded, minimap2, ...

How minimap works, rough idea

2.3. MSA

When we need to compare a lot of individuals together we can do MSA.

NP-hard problem so we need heuristics or tricks

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

2.3.0.1. Progressive

guide tree, clustering of sequences then refine alignment. Good heuristic but with larger datasets, becomes harder

2.3.0.2. HMMs / index alignment / pairwise

Example of COVID where homology is high so we can get away with using HMMS / pairwise to root sequence. *(point to appendinx with covid align ?)*

3. Mapping-friendly sequence reductions: going beyond homopolymer compression

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

originally published in iScience in XXX
doi:10.1371/journal.pcbi.1008873

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) (Dohm et al. 2020). Many algorithmic problems, such as alignment, become trivial if not for sequencing errors (Gusfield 1997). 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 (Li 2018).

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 Miller et al. (2008). It is used by a wide range of error-correction algorithms, e.g. for 454 data (Bragg et al. 2012), PacBio data (Au et al. 2012), and Oxford Nanopore data (Sahlin and Medvedev 2021). HPC is used in alignment, e.g. by the widely used minimap2 aligner (Li 2018). HPC is also used in long-read assembly, e.g. HiCanu (Nurk, Walenz, et al. 2020), SMARTdenovo (Liu et al. 2021), or mdBG (Ekim, Berger, and Chikhi 2021). HPC is also used for clustering transcriptome reads according to gene family of origin (Sahlin and Medvedev 2020). 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 (Nurk, Koren, et al. 2021). 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.

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

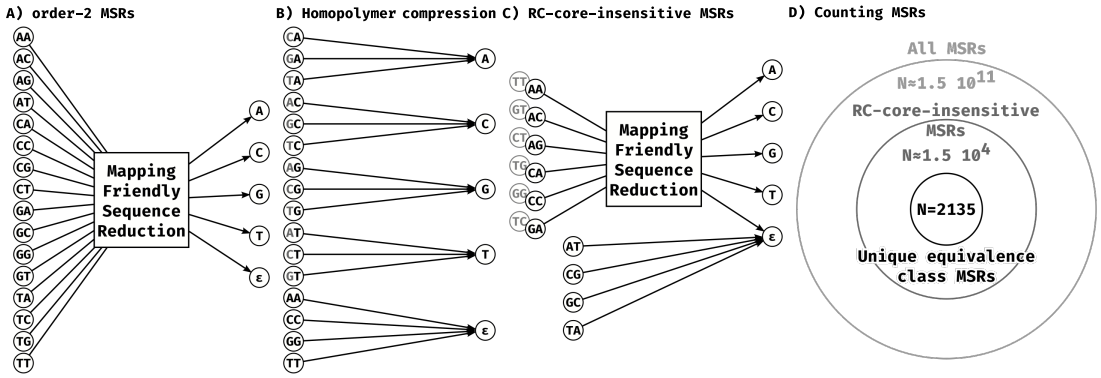

Figure 3.1.: **Representing and counting Streaming sequence reductions.** **A**: General representation of an order-2 Streaming sequence reduction as a mapping of 16 input dinucleotides, to the 4 nucleotide outputs and the empty character ε. **B**: Homopolymer compression is an order-2 SSR. All dinucleotides 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.

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 IsComp$(t(i), t(j))$ = 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.

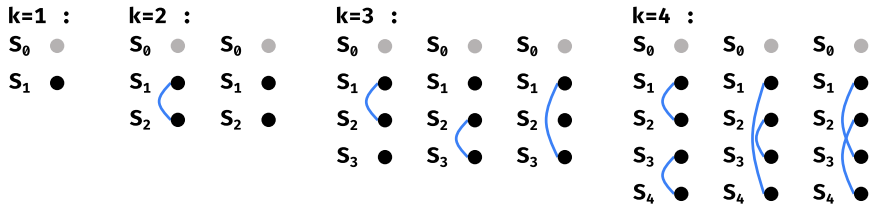

Figure 3.2.: **SSR equivalence classes for a fixed partition of the inputs.** S_0 is always assigned ε, so it is represented by a gray node. A blue link between S_i and an S_j denotes that IsComp$(t(i), t(j))$ = true. The equivalence classes are determined by the Watson-Crick complementary relationships 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 $S_0, \ldots, S_k$, with $S_1, \ldots, S_k$ 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 $\left\{ {n \atop k} \right\}$ (Graham, Knuth, and Patashnik 1994, 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 (Nurk, Koren, et al. 2021). 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) (Adams et al. 2000).

3. **Synthetic centromeric sequence:** This sequence was obtained from the `TandemTools` mapper test data (Mikheenko et al. 2020). 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` (Yang et al. 2017) 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`(Li 2018) 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` (Li 2018). 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

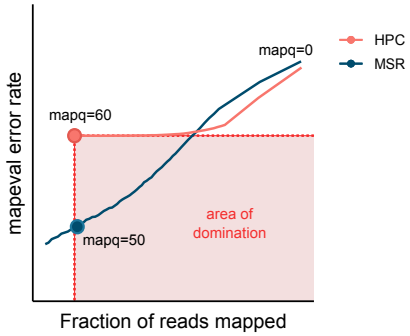

Figure 3.3.: **Illustration of how a respective mapq threshold is chosen for each of our evaluated MSRs.**
The orange dot shows the error rate and fraction of reads mapped for HPC at mapq threshold 60. Anything below and to the right of this point is strictly better than HPC 60, i.e. it has both a lower 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.

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.

- **MSR$_F$**: The MSR with the highest fraction of mapped reads at a mapq threshold of 0.
- **MSR$_E$**: The MSR with the lowest error rate at its respective mapq threshold.
- **MSR$_P$**: 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$_F$, MSR$_E$, MSR$_P$. An intriguing property is that they output predominantly As and Ts, with MSR$_P$ assigning 2 input pairs to the G/C output whereas MSR$_E$ and MSR$_F$ assign only one. Additionally, MSR$_E$ and MSR$_P$ 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.

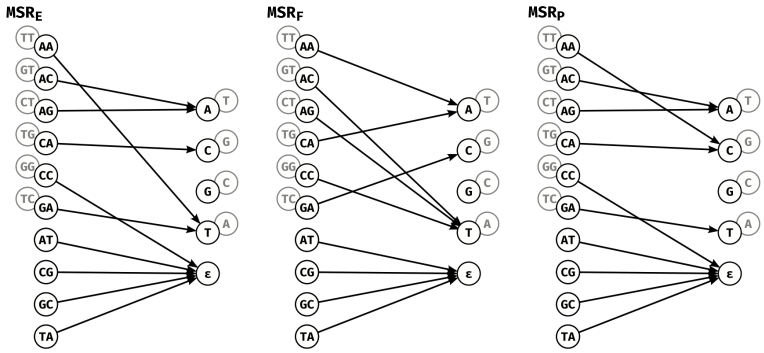

Figure 3.4.: **Graph representations of our highlighted MSRs: MSR$_E$, MSR$_F$, and MSR$_P$.**
MSR$_E$ has the lowest error rate of among MSRs at the highest mapq threshold for which it performs better than HPC at mapq 60, MSR$_F$ has the highest fraction of reads mapped at mapq 60 and MSR$_P$ has the 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.

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.

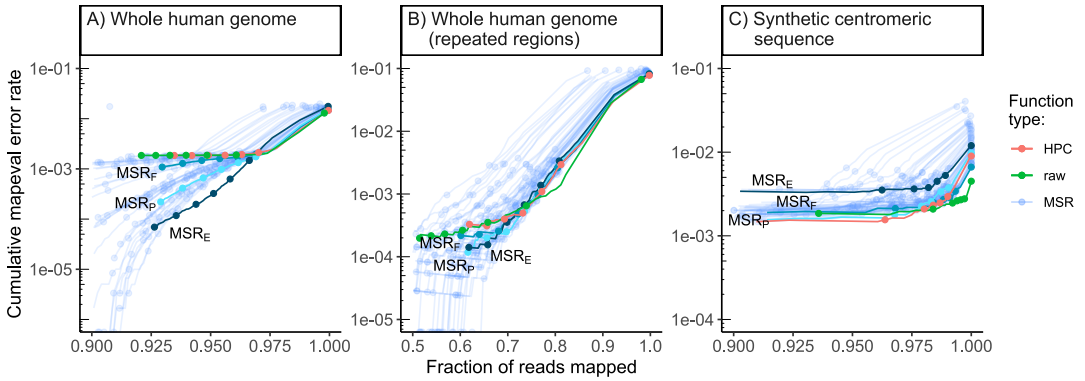

Figure 3.5.: **Performance of our 58 selected mapping-friendly sequence reductions across genomes on reads simulated by `nanosim`**
Panel **A)** shows the whole human genome assembly, **B)** the subset of mapped reads from panel B that originate from repetitive regions, and **C)** the "TandemTools" synthetic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

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≥ 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) (Jain et al. 2020). The same options as `minimap2` were used, and k-mers were counted using `meryl` (Rhie et al. 2020), considering the top 0.02% as repetitive (as suggested by the `winnowmap2` usage guide).

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.2, 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 (Prodanov and Bansal 2020; Li 2021; Li et al. 2018). 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

Adams, M. D. et al. (Mar. 24, 2000). "The Genome Sequence of Drosophila Melanogaster". In: *Science (New York, N.Y.)* 287.5461, pp. 2185–2195. ISSN: 0036-8075. DOI: 10.1126/science.287.5461.2185. pmid: 10731132 (cit. on p. 25).

Au, Kin Fai et al. (Oct. 4, 2012). "Improving PacBio Long Read Accuracy by Short Read Alignment". In: *PLOS ONE* 7.10, e46679. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0046679 (cit. on p. 18).

Bragg, Lauren et al. (May 2012). "Fast, Accurate Error-Correction of Amplicon Pyrosequences Using Acacia". In: *Nature Methods* 9.5 (5), pp. 425–426. ISSN: 1548-7105. DOI: 10.1038/nmeth.1990 (cit. on p. 18).

Dohm, Juliane C et al. (June 1, 2020). "Benchmarking of Long-Read Correction Methods". In: *NAR Genomics and Bioinformatics* 2.2. ISSN: 2631-9268. DOI: 10.1093/nargab/lqaa037 (cit. on p. 18).

Ekim, Barış, Bonnie Berger, and Rayan Chikhi (Oct. 20, 2021). "Minimizer-Space de Bruijn Graphs: Whole-Genome Assembly of Long Reads in Minutes on a Personal Computer". In: *Cell Systems* 12.10, 958–968.e6. ISSN: 2405-4712. DOI: 10.1016/j.cels.2021.08.009 (cit. on p. 18).

Graham, Ronald L., Donald Ervin Knuth, and Oren Patashnik (1994). *Concrete Mathematics: A Foundation for Computer Science*. 2nd ed. Reading, Mass: Addison-Wesley. 657 pp. ISBN: 978-0-201-55802-9 (cit. on p. 24).

Gusfield, Dan (1997). "Algorithms on stings, trees, and sequences: Computer science and computational biology". In: *Acm Sigact News* 28.4, pp. 41–60 (cit. on p. 18).

Jain, Chirag et al. (July 1, 2020). "Weighted Minimizer Sampling Improves Long Read Mapping". In: *Bioinformatics* 36 (Supplement_1), pp. i111–i118. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btaa435 (cit. on pp. 30, 73).

Li, Heng (Sept. 15, 2018). "Minimap2: Pairwise Alignment for Nucleotide Sequences". In: *Bioinformatics* 34.18, pp. 3094–3100. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bty191 (cit. on pp. 18, 26, 73).

– (Aug. 7, 2021). *New Strategies to Improve Minimap2 Alignment Accuracy*. arXiv: 2108.03515 [q-bio] (cit. on p. 32).

Li, Heng et al. (Aug. 2018). "A Synthetic-Diploid Benchmark for Accurate Variant-Calling Evaluation". In: *Nature Methods* 15.8 (8), pp. 595–597. ISSN: 1548-7105. DOI: 10.1038/s41592-018-0054-7 (cit. on p. 32).

Liu, Hailin et al. (Mar. 8, 2021). "SMARTdenovo: A de Novo Assembler Using Long Noisy Reads". In: *Gigabyte* 2021, pp. 1–9. DOI: 10.46471/gigabyte.15 (cit. on p. 18).

Mikheenko, Alla et al. (July 1, 2020). "TandemTools: Mapping Long Reads and Assessing/Improving Assembly Quality in Extra-Long Tandem Repeats". In: *Bioinformatics* 36 (Supplement_1), pp. i75–i83. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btaa440 (cit. on p. 26).

Miller, Jason R. et al. (Dec. 15, 2008). "Aggressive Assembly of Pyrosequencing Reads with Mates". In: *Bioinformatics* 24.24, pp. 2818–2824. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btn548 (cit. on p. 18).

Nurk, Sergey, Sergey Koren, et al. (2021). "The Complete Sequence of a Human Genome". In: *bioRxiv : the preprint server for biology*. DOI: 10.1101/2021.05.26.445798 (cit. on pp. 19, 25).

Nurk, Sergey, Brian P. Walenz, et al. (Jan. 9, 2020). "HiCanu: Accurate Assembly of Segmental Duplications, Satellites, and Allelic Variants from High-Fidelity Long Reads". In: *Genome Research* 30.9, pp. 1291–1305. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.263566.120. pmid: 32801147 (cit. on p. 18).

Prodanov, Timofey and Vikas Bansal (Nov. 4, 2020). "Sensitive Alignment Using Paralogous Sequence Variants Improves Long-Read Mapping and Variant Calling in Seg-

mental Duplications". In: *Nucleic Acids Research* 48.19, e114. ISSN: 0305-1048. DOI: 10.1093/nar/gkaa829 (cit. on p. 32).

Rhie, Arang et al. (Sept. 14, 2020). "Merqury: Reference-Free Quality, Completeness, and Phasing Assessment for Genome Assemblies". In: *Genome Biology* 21.1, p. 245. ISSN: 1474-760X. DOI: 10.1186/s13059-020-02134-9 (cit. on p. 30).

Sahlin, Kristoffer and Paul Medvedev (Apr. 1, 2020). "De Novo Clustering of Long-Read Transcriptome Data Using a Greedy, Quality Value-Based Algorithm". In: *Journal of Computational Biology* 27.4, pp. 472–484. DOI: 10.1089/cmb.2019.0299 (cit. on p. 18).

– (Jan. 4, 2021). "Error Correction Enables Use of Oxford Nanopore Technology for Reference-Free Transcriptome Analysis". In: *Nature Communications* 12.1 (1), p. 2. ISSN: 2041-1723. DOI: 10.1038/s41467-020-20340-8 (cit. on p. 18).

Yang, Chen et al. (Apr. 1, 2017). "NanoSim: Nanopore Sequence Read Simulator Based on Statistical Characterization". In: *GigaScience* 6.4. ISSN: 2047-217X. DOI: 10.1093/gigascience/gix010 (cit. on pp. 26, 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, compartiments 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...
- others

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. Presentation of HIV

- pandemic
- history

5.2.2. Replication cycle of HIV

- proteins (+ computational representation as a string of letters)
- full 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. Using Machine Learning and Big Data to Explore the Drug Resistance Landscape in HIV

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

originally published in PLoS Computational Biology in August 2021
doi:10.1371/journal.pcbi.1008873

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 (Lepri et al. 2000; Verhofstede et al. 2007). Furthermore, drug-resistant HIV strains can be transmitted to treatment-naive individuals and further spread throughout the population over time (Hué et al. 2009; Mourad et al. 2015; Zhukova et al. 2017). 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 (Bennett et al. 2009; Hammond et al. 1998; Wensing et al. 2016). 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 (Dudoit and Laan 2007; Villabona-Arenas et al. 2016). 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 (Maddison and FitzJohn 2015)) 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 (Sham and Purcell 2014), 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 (Lengauer and Sing 2006) or other indicators from phenotypic resistance assays such as PhenoSense (J. Zhang et al. 2005). Many methods have been used to predict a resistance level: Support Vector Machines (SVMs) (Niko Beerenwinkel et al. 2003), k-Nearest Neighbors (KNN) and Random Forests (RFs) (Shen et al. 2016), and more recently Artificial Neural Networks (ANNs) (Yu, Weber, and Harrison 2014; Sheik Amamuddy, Bishop, and Tastan Bishop 2017). 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 (N. Beerenwinkel et al. 2001; Araya and Hazelhurst 2009), ensemble classifier chains (Riemenschneider et al. 2016; Heider et al. 2013) and also ANNs (Drăghici and Potter 2003). Most recently Steiner *et al.* (Steiner, Gibson, and Crandall 2020) 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.* (Steiner, Gibson, and Crandall 2020) 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 ($n \approx 4,000$) (Villabona-Arenas et al. 2016). 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 (Villabona-Arenas et al. 2016). 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 (Villabona-Arenas et al. 2016), 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 (Villabona-Arenas et al. 2016). 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 (Mooney et al. 2018). 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 (Tibshirani 1996; Brier 1950; Gascuel et al. 1998), 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

		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.

value to amino acids, we transform a categorical variable into an ordinal variable. Any ordering of amino acids is hard to 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 (Villabona-Arenas et al. 2016). 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 (Goeman and Solari 2014). 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 (Rennie et al. 2003); the higher (≈ 1.0) and the lower (≈ 0) conditional probabilities correspond to the most important features.

2. Logistic regression (LR) with L1 regularization (LASSO) (Tibshirani 1996) 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 (Breiman 2001).

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 (Alvarez Melis and Jaakkola 2018; Hastie, Tibshirani, and Friedman 2009; Q. Zhang, Y. N. Wu, and Zhu 2018). 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 (Gascuel et al. 1998).
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 (Villabona-Arenas et al. 2016). 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

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

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).

6.2.3. Measuring classifier performance

To compare the performance of our classifiers we used balanced accuracy (Brodersen et al. 2010), 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 (Nguyen Xuan Vinh, Julien Epps, and Bailey 2010). 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 (Brier 1950) 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

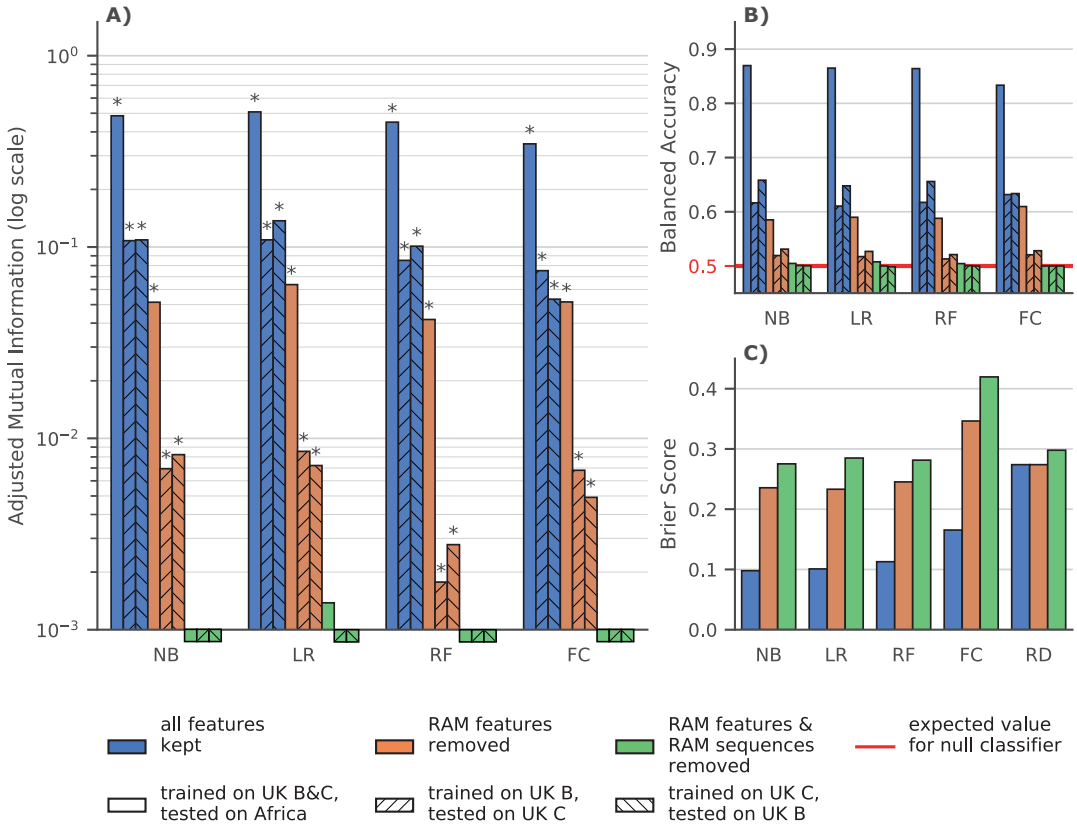

Figure 6.1.: **Classifier Performance on UK and African datasets.**
NB: naive Bayes, **LR**: Logistic Regression with Lasso regularization, **RF**: Random Forest, **FC**: Fisher Classifier, **RD**: Agnostic random probabilistic classifier (this classifier predicts, as the probability of a sample belonging 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 (Villabona-Arenas et al. 2016)) 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

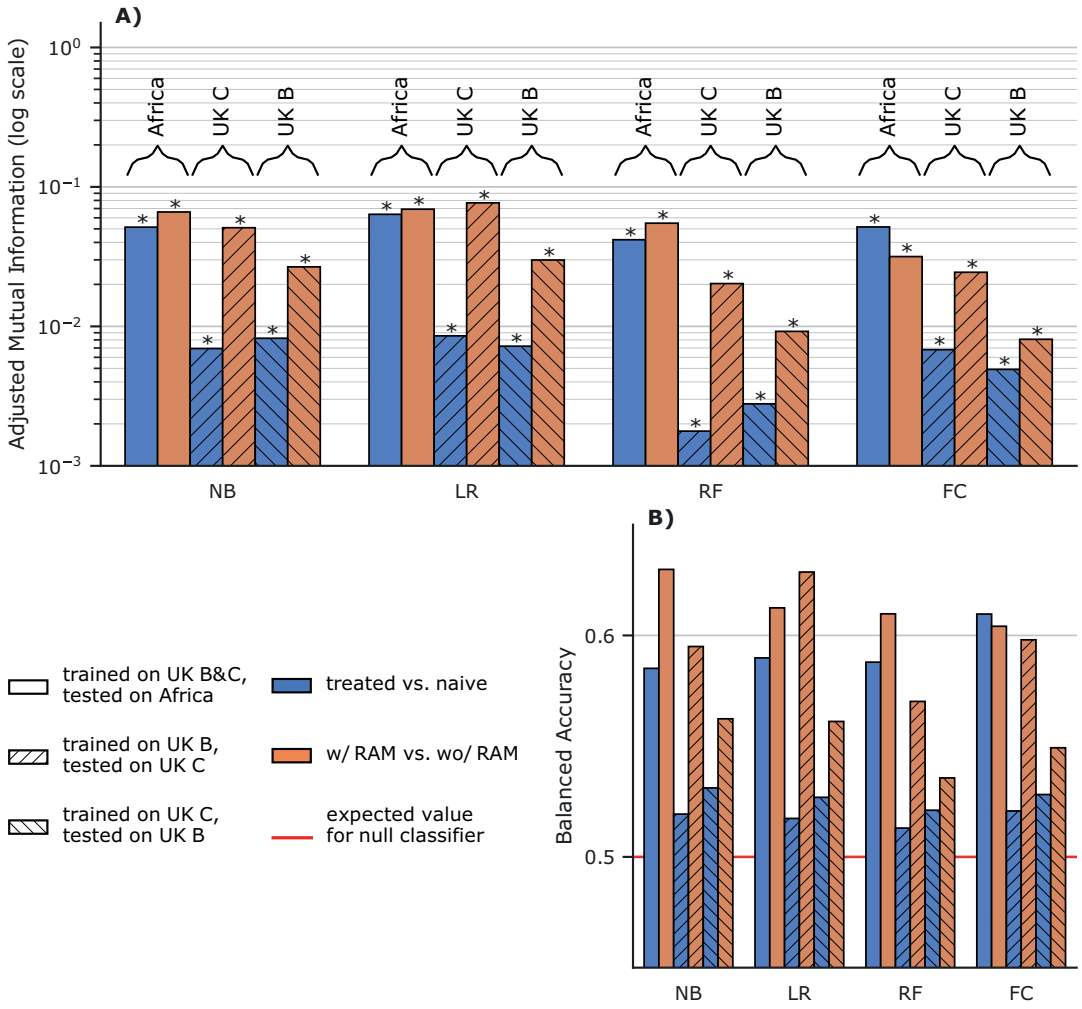

Figure 6.2.: **Discrimination between sequences having at least one RAM, and those having none on sequences with training features corresponding to known RAMs removed.**
NB: naive Bayes, **LR**: Logistic Regression with Lasso regularization, **RF**: Random Forest, **FC**: Fisher Classifier. **A)** Adjusted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

each classification task, a ranking of mutations potentially associated with resistance 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.

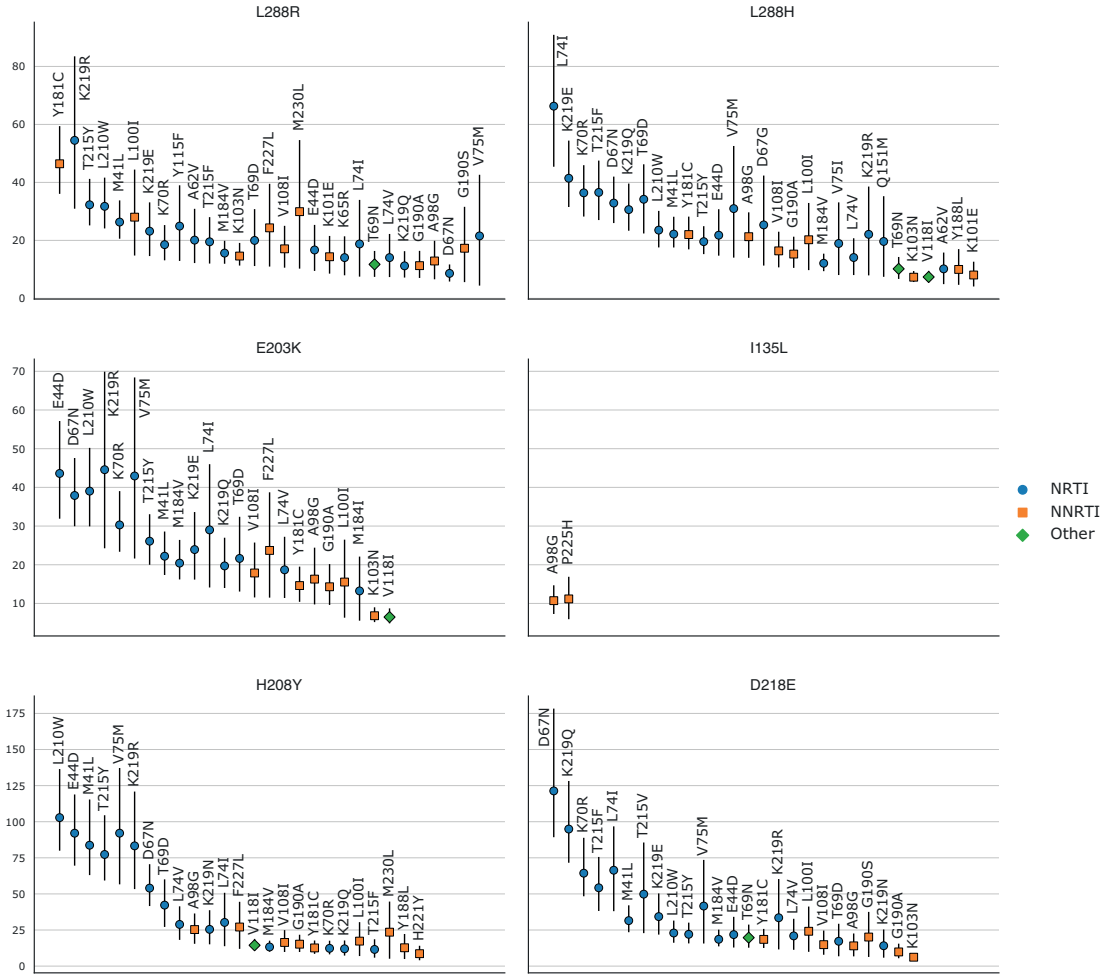

Figure 6.3.: **Relative risk of the new mutations with regards to known RAMs on the UK dataset**
(i.e. the prevalence of the new mutation in sequences with a given known RAM divided by the prevalence of the new mutation in sequences without this RAM). RRs were only computed for mutations (new and RAMs) 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 (Schrödinger, LLC 2015). 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 (Sarafianos et al. 2009). 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

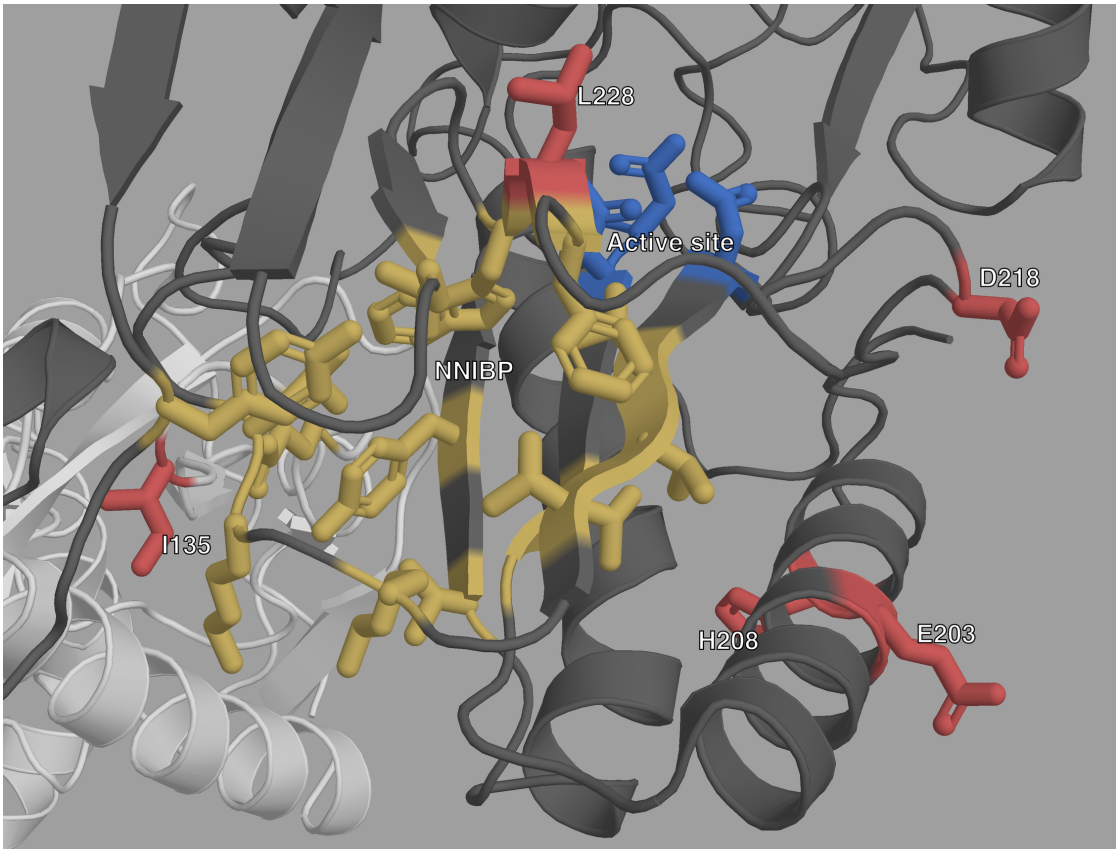

Figure 6.4.: **Structure of HIV-1 RT with highlighted important sites.**
The p66 subunit is colored dark gray and the p51 subunit white. The active site is highlighted in blue, and the NNIBP is highlighted in yellow. The sites of new mutations are colored in red.

very low BLOSUM62 score, of -3 and -2 respectively (Table 6.3). Arginine (R) and 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 (Rhee et al. 2007) and were suggested to be associated with reduced susceptibility to didanosine (De Luca et al. 2007; Marcelin et al. 2006). I135L has been observed in sequences with reduced susceptibility to NNRTIs (Brown et al. 2000). H208Y has been associated with NNRTI and NRTI resistance (Clark et al. 2006) and it has been suggested that it has an accessory role in NRTI resistance (Nebbia et al. 2007). E203K, D218E, L228RH and H208Y have all been mentioned in (Saracino et al. 2006) 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 (T. T. Wu et al. 2009). 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).

Supporting Information

Supporting Information can be found in the appendix B

References for chapter 6

Alvarez Melis, David and Tommi Jaakkola (2018). "Towards Robust Interpretability with Self-Explaining Neural Networks". In: *Advances in Neural Information Processing Systems 31*. Ed. by S. Bengio et al. Curran Associates, Inc., pp. 7775–7784 (cit. on p. 51).

Araya, Seare Tesfamichael and Scott Hazelhurst (Jan. 2009). "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, pp. 62–72. ISSN: 0035-919X. DOI: 10.1080/00359190909519238 (cit. on p. 46).

Beerenwinkel, N. et al. (Nov. 2001). "Geno2pheno: Interpreting Genotypic HIV Drug Resistance Tests". In: *IEEE Intelligent Systems* 16.6, pp. 35–41. ISSN: 1941-1294. DOI: 10.1109/5254.972080 (cit. on p. 46).

Beerenwinkel, Niko et al. (July 2003). "Geno2pheno: Estimating Phenotypic Drug Resistance from HIV-1 Genotypes". In: *Nucleic Acids Research* 31.13, pp. 3850–3855. ISSN: 0305-1048. DOI: 10.1093/nar/gkg575 (cit. on p. 46).

Bennett, Diane E. et al. (Mar. 2009). "Drug Resistance Mutations for Surveillance of Transmitted HIV-1 Drug-Resistance: 2009 Update". en. In: *PLOS ONE* 4.3, e4724. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0004724 (cit. on p. 45).

Breiman, Leo (Oct. 2001). "Random Forests". en. In: *Machine Learning* 45.1, pp. 5–32. ISSN: 1573-0565. DOI: 10.1023/A:1010933404324 (cit. on p. 51).

Brier, Glenn W. (Jan. 1950). "Verification of Forecasts Expressed in Terms of Probability". en. In: *Monthly Weather Review* 78.1, pp. 1–3. ISSN: 0027-0644 (cit. on pp. 48, 52).

Brodersen, Kay Henning et al. (Aug. 2010). "The Balanced Accuracy and Its Posterior Distribution". In: *2010 20th International Conference on Pattern Recognition*, pp. 3121–3124. DOI: 10.1109/ICPR.2010.764 (cit. on pp. 52, 84).

Brown, Andrew J. Leigh et al. (Nov. 2000). "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, pp. 10269–10273. ISSN: 0022-538X (cit. on p. 63).

Clark, Shauna A. et al. (Apr. 2006). "Reverse Transcriptase Mutations 118I, 208Y, and 215Y Cause HIV-1 Hypersusceptibility to Non-Nucleoside Reverse Transcriptase Inhibitors". en-US. In: *AIDS* 20.7, pp. 981–984. ISSN: 0269-9370. DOI: 10.1097/01.aids.0000222069.14878.44 (cit. on p. 63).

De Luca, Andrea et al. (Dec. 2007). "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, pp. 1645–1653. ISSN: 0022-1899. DOI: 10.1086/522231 (cit. on p. 63).

Drăghici, Sorin and R. Brian Potter (Jan. 2003). "Predicting HIV Drug Resistance with Neural Networks". In: *Bioinformatics* 19.1, pp. 98–107. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/19.1.98 (cit. on p. 46).

Dudoit, Sandrine and Mark J. van der Laan (Dec. 2007). *Multiple Testing Procedures with Applications to Genomics*. en. Springer Science & Business Media. ISBN: 978-0-387-49317-6 (cit. on p. 45).

Gascuel, Olivier et al. (Aug. 1998). "Twelve Numerical, Symbolic and Hybrid Supervised Classification Methods". In: *International Journal of Pattern Recognition and Artificial Intelligence* 12.05, pp. 517–571. ISSN: 0218-0014. DOI: 10.1142/S0218001498000336 (cit. on pp. 48, 51).

Goeman, Jelle J. and Aldo Solari (2014). "Multiple Hypothesis Testing in Genomics". en. In: *Statistics in Medicine* 33.11, pp. 1946–1978. ISSN: 1097-0258. DOI: 10.1002/sim.6082 (cit. on p. 50).

Hammond, Jennifer et al. (Dec. 1998). "Mutations in Retroviral Genes Associated with Drug Resistance". en. In: *Human retroviruses and AIDS*, pp. 11136–11179 (cit. on p. 45).

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (Aug. 2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. en. Springer Science & Business Media. ISBN: 978-0-387-84858-7 (cit. on p. 51).

Heider, Dominik et al. (Aug. 2013). "Multilabel Classification for Exploiting Cross-Resistance Information in HIV-1 Drug Resistance Prediction". In: *Bioinformatics* 29.16, pp. 1946–1952. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btt331 (cit. on p. 46).

Hué, Stéphane et al. (Mar. 2009). "Demonstration of Sustained Drug-Resistant Human Immunodeficiency Virus Type 1 Lineages Circulating among Treatment-Naïve Individuals". en. In: *Journal of Virology* 83.6, pp. 2645–2654. ISSN: 0022-538X, 1098-5514. DOI: 10.1128/JVI.01556-08 (cit. on p. 45).

Lengauer, Thomas and Tobias Sing (Oct. 2006). "Bioinformatics-Assisted Anti-HIV Therapy". en. In: *Nature Reviews Microbiology* 4.10, pp. 790–797. ISSN: 1740-1534. DOI: 10.1038/nrmicro1477 (cit. on p. 46).

Lepri, Alessandro Cozzi et al. (Mar. 2000). "Resistance Profiles in Patients with Viral Rebound on Potent Antiretroviral Therapy". en. In: *The Journal of Infectious Diseases* 181.3, pp. 1143–1147. ISSN: 0022-1899. DOI: 10.1086/315301 (cit. on p. 45).

Maddison, Wayne P. and Richard G. FitzJohn (Jan. 2015). "The Unsolved Challenge to Phylogenetic Correlation Tests for Categorical Characters". en. In: *Systematic Biology* 64.1, pp. 127–136. ISSN: 1063-5157. DOI: 10.1093/sysbio/syu070 (cit. on p. 45).

Marcelin, Anne-Genevieve et al. (2006). "Impact of HIV-1 Reverse Transcriptase Polymorphism at Codons 211 and 228 on Virological Response to Didanosine". en. In: *Antiviral Therapy*, p. 8 (cit. on p. 63).

Mooney, Alyssa C. et al. (July 2018). "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, pp. 2368–2379. ISSN: 1573-3254. DOI: 10.1007/s10461-018-2155-9 (cit. on p. 48).

Mourad, Raphaël et al. (Sept. 2015). "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, pp. 1917–1925. ISSN: 0269-9370. DOI: 10.1097/QAD.0000000000000768 (cit. on p. 45).

Nebbia, G. et al. (May 2007). "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, pp. 1013–1016. ISSN: 0305-7453. DOI: 10.1093/jac/dkm067 (cit. on p. 63).

Rennie, Jason D et al. (2003). "Tackling the Poor Assumptions of Naive Bayes Text Classifiers". en. In: *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pp. 616–623 (cit. on p. 50).

Rhee, Soo-Yon et al. (May 2007). "HIV-1 Subtype B Protease and Reverse Transcriptase Amino Acid Covariation". en. In: *PLOS Computational Biology* 3.5, e87. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.0030087 (cit. on p. 63).

Riemenschneider, Mona et al. (Feb. 2016). "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, p. 10. ISSN: 1756-0381. DOI: 10.1186/s13040-016-0089-1 (cit. on p. 46).

Saracino, A. et al. (2006). "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, pp. 9–17. ISSN: 1096-9071. DOI: 10.1002/jmv.20500 (cit. on p. 63).

Sarafianos, Stefan G. et al. (Jan. 2009). "Structure and Function of HIV-1 Reverse Transcriptase: Molecular Mechanisms of Polymerization and Inhibition". In: *Journal of molecular biology* 385.3, pp. 693–713. ISSN: 0022-2836. DOI: 10.1016/j.jmb.2008.10.071 (cit. on p. 60).

Schrödinger, LLC (Nov. 2015). "The PyMOL Molecular Graphics System, Version 1.8" (cit. on p. 60).

Sham, Pak C. and Shaun M. Purcell (May 2014). "Statistical Power and Significance Testing in Large-Scale Genetic Studies". en. In: *Nature Reviews Genetics* 15.5, pp. 335–346. ISSN: 1471-0064. DOI: 10.1038/nrg3706 (cit. on p. 46).

Sheik Amamuddy, Olivier, Nigel T. Bishop, and Özlem Tastan Bishop (Aug. 2017). "Improving Fold Resistance Prediction of HIV-1 against Protease and Reverse Transcriptase Inhibitors Using Artificial Neural Networks". In: *BMC Bioinformatics* 18.1, p. 369. ISSN: 1471-2105. DOI: 10.1186/s12859-017-1782-x (cit. on p. 46).

Shen, ChenHsiang et al. (Aug. 2016). "Automated Prediction of HIV Drug Resistance from Genotype Data". In: *BMC Bioinformatics* 17.8, p. 278. ISSN: 1471-2105. DOI: 10.1186/s12859-016-1114-6 (cit. on p. 46).

Steiner, Margaret C., Keylie M. Gibson, and Keith A. Crandall (May 2020). "Drug Resistance Prediction Using Deep Learning Techniques on HIV-1 Sequence Data". en. In: *Viruses* 12.5, p. 560. DOI: 10.3390/v12050560 (cit. on p. 46).

Tibshirani, Robert (1996). "Regression Shrinkage and Selection Via the Lasso". en. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1, pp. 267–288. ISSN: 2517-6161. DOI: 10.1111/j.2517-6161.1996.tb02080.x (cit. on pp. 48, 50).

Verhofstede, Chris et al. (2007). "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, pp. 1254–1260. ISSN: 1096-9071. DOI: 10.1002/jmv.20950 (cit. on p. 45).

Villabona-Arenas, Christian Julian et al. (Nov. 2016). "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, p. 2577. ISSN: 0269-9370. DOI: 10.1097/QAD.0000000000001233 (cit. on pp. 45–48, 50, 51, 55).

Vinh, Nguyen Xuan, Julien Epps, and James Bailey (2010). "Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance". en. In: *Journal of Machine Learning Research* 11, p. 18 (cit. on pp. 52, 85).

Wensing, A. M. et al. (Dec. 2016). "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, pp. 132, 132–133. ISSN: 2161-5861 (cit. on p. 45).

Wu, Tong Tong et al. (Mar. 2009). "Genome-Wide Association Analysis by Lasso Penalized Logistic Regression". en. In: *Bioinformatics* 25.6, pp. 714–721. ISSN: 1460-2059, 1367-4803. DOI: 10.1093/bioinformatics/btp041 (cit. on p. 64).

Yu, Xiaxia, Irene T. Weber, and Robert W. Harrison (July 2014). "Prediction of HIV Drug Resistance from Genotype with Encoded Three-Dimensional Protein Structure". In: *BMC Genomics* 15.5, S1. ISSN: 1471-2164. DOI: 10.1186/1471-2164-15-S5-S1 (cit. on p. 46).

Zhang, Jie et al. (Apr. 2005). "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, pp. 439–444. ISSN: 1525-4135. DOI: 10.1097/01.qai.0000147526.64863.53 (cit. on p. 46).

Zhang, Quanshi, Ying Nian Wu, and Song-Chun Zhu (2018). "Interpretable Convolutional Neural Networks". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8827–8836 (cit. on p. 51).

Zhukova, Anna et al. (Dec. 2017). "The Role of Phylogenetics as a Tool to Predict the Spread of Resistance". en. In: *The Journal of Infectious Diseases* 216.suppl_9, S820–S823. ISSN: 0022-1899. DOI: 10.1093/infdis/jix411 (cit. on p. 45).

7. Learning alignments, an interesting perspective

7.1. Learning pairwise alignment

7.1.1. Transformers and deep embedding

7.1.2. DEDAL

7.1.3. predictind an alignment

7.2. What else could we learn ?

7.2.1. Learn to predict seeds or starting positions

7.2.2. Learn pre-processing functions

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` (Yang et al. 2017) and assembled using a centromere-tailored pipeline (Bzikadze and Pevzner 2020). A 10kbp deletion was then added to this assembly. The resulting sequence is the one we refer to as the "Centromeric sequence".

A.2. MSR performance comparison

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`(Li 2018) and `winnowmap2`(Jain et al. 2020). 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%	**0.00 –100%**	0.954 –1%	**0.00 –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.

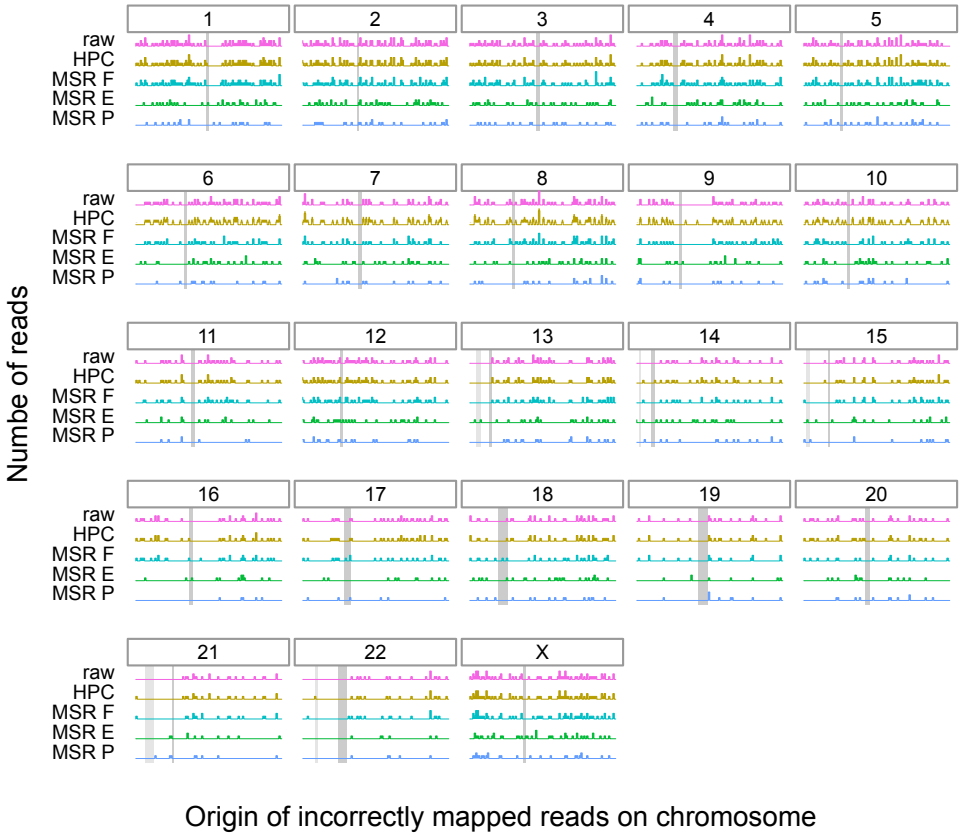

Figure A.1.: **Histogram of the original simulated positions for the incorrectly mapped reads using `minimap2` at high mapping qualities across the whole human genome, for several transformation methods.**
For a given chromosome, each row represents the number of simulated reads starting at that particular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

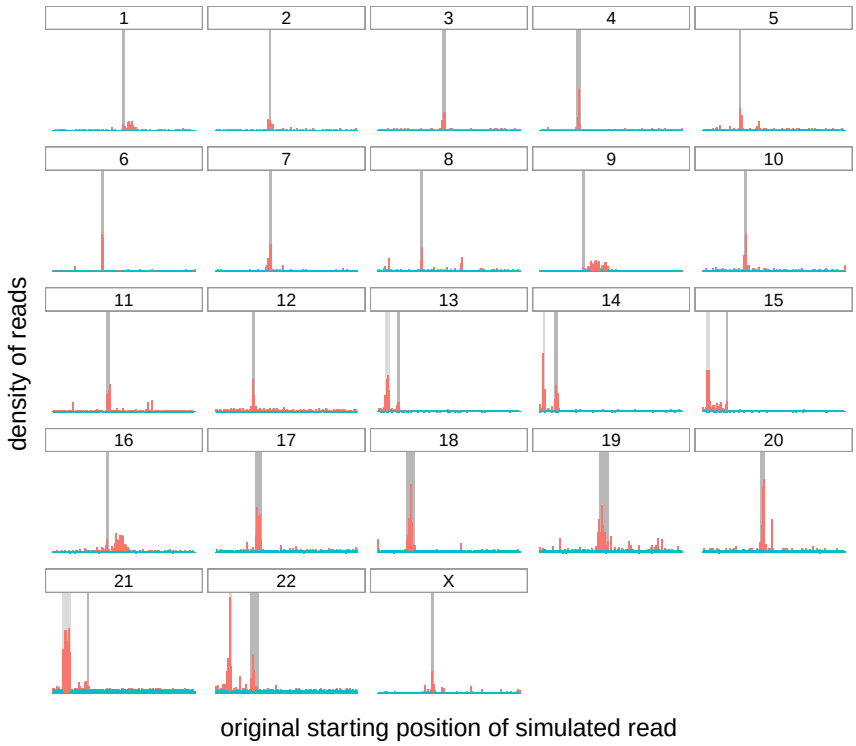

Figure A.2.: **Origin of correctly and incorrectly mapped raw reads**
Distribution of the origin of correctly and incorrectly mapped simulated reads (in teal and red respectively) on the different chromosomes of the whole human genome. The dark grey rectangle for each chromosome represents the centromere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

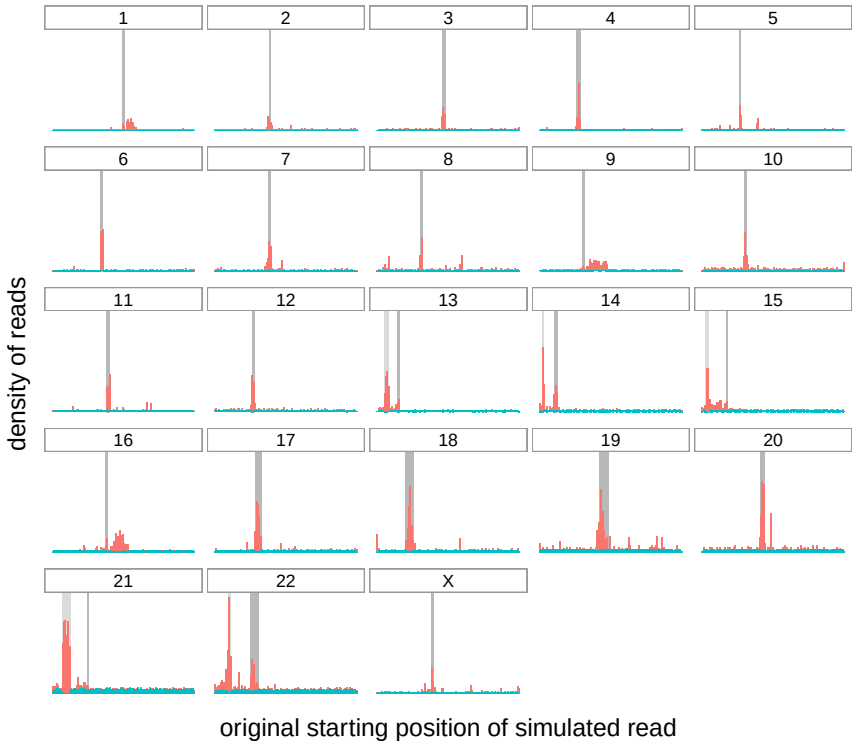

Figure A.3.: **Origin of correctly (teal) and incorrectly (red) mapped reads, transformed with HPC**
Distribution of the origin of correctly and incorrectly mapped simulated reads (in teal and red respectively) on the different chromosomes of the whole human genome. The dark grey rectangle for each chromosome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

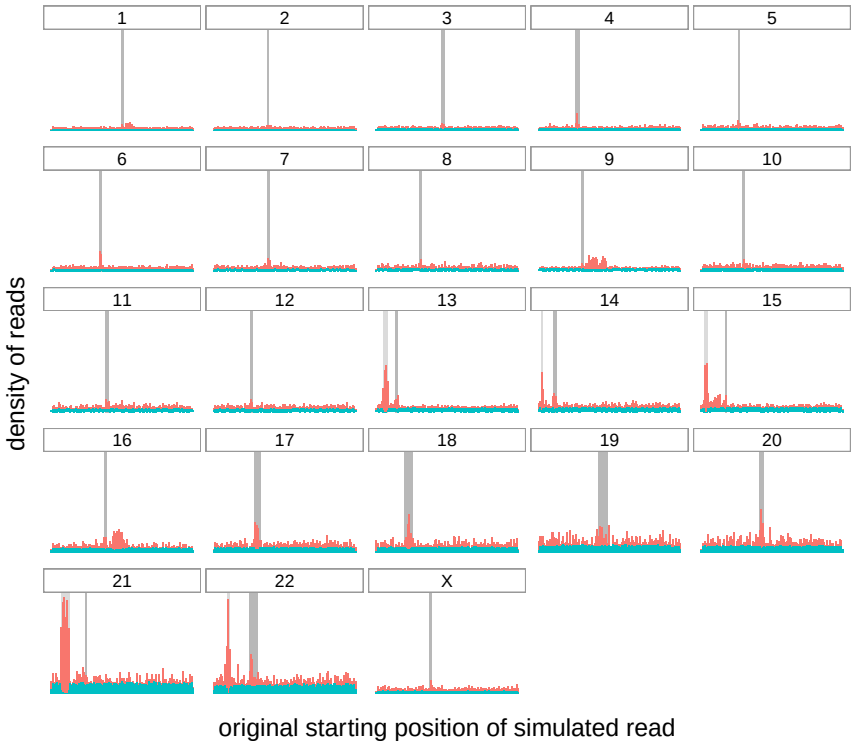

Figure A.4.: **Origin of correctly (teal) and incorrectly (red) mapped reads, transformed with $\mathbf{MSR_E}$**
Distribution of the origin of correctly and incorrectly mapped simulated reads (in teal and red respectively) on the different chromosomes of the whole human genome. The dark grey rectangle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

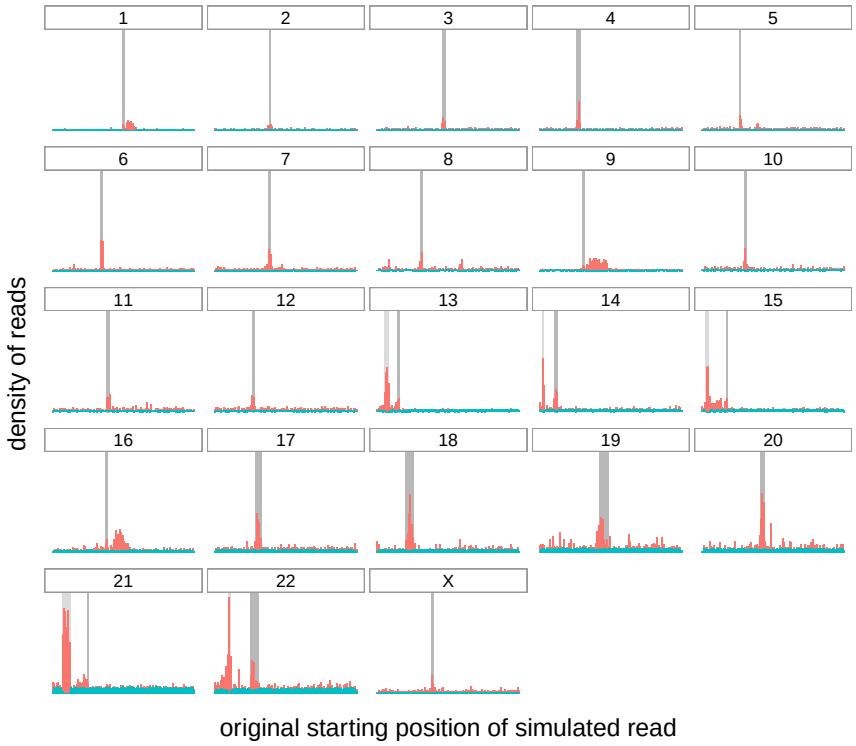

Figure A.5.: **Origin of correctly (teal) and incorrectly (red) mapped reads, transformed with MSR_P**
Distribution of the origin of correctly and incorrectly mapped simulated reads (in teal and red respectively) on the different chromosomes of the whole human genome. The dark grey rectangle for each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

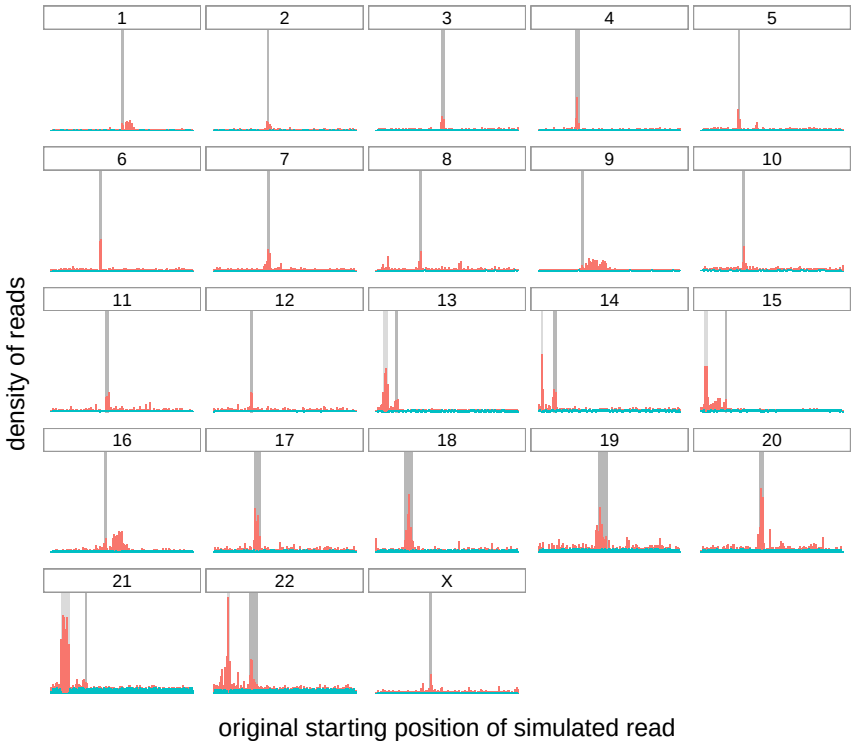

Figure A.6.: **Origin of correctly (teal) and incorrectly (red) mapped reads, transformed with $\mathbf{MSR_F}$**
Distribution of the origin of correctly and incorrectly mapped simulated reads (in teal and red respectively) on the different chromosomes of the whole human genome. The dark grey rectangle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

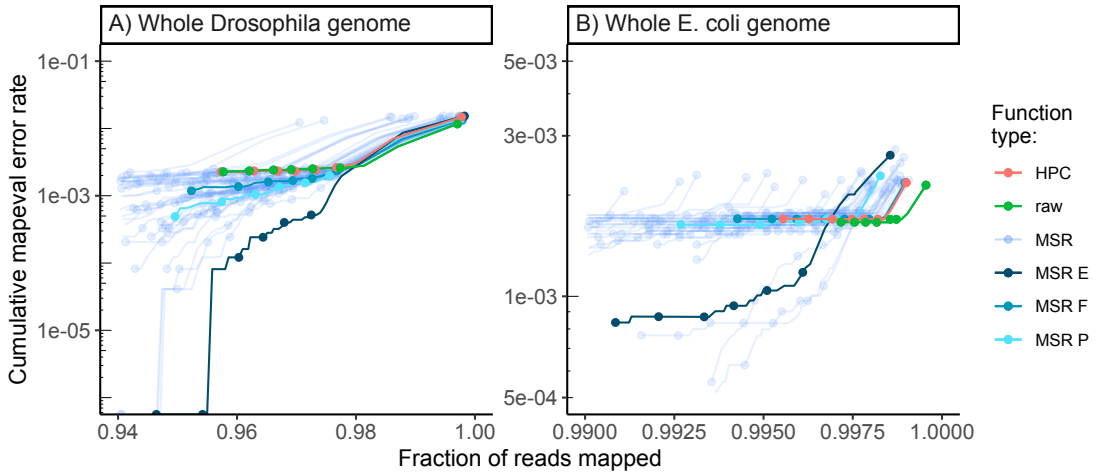

Figure A.7.: Results of the `paftools mapeval` evaluation on reads simulated and mapped to whole *Drosophila melanogaster* and *Escherichia coli* (Genbank ID U00096.2) genomes. MSRs E, F and P are shown in different shades of blue to differentiate them from other MSRs. Reads were simulated with `nanosim`,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

References for appendix A

Bzikadze, Andrey V. and Pavel A. Pevzner (Nov. 2020). "Automated Assembly of Centromeres from Ultra-Long Error-Prone Reads". In: *Nature Biotechnology* 38.11 (11), pp. 1309–1316. ISSN: 1546-1696. DOI: 10.1038/s41587-020-0582-4 (cit. on p. 73).

Jain, Chirag et al. (July 1, 2020). "Weighted Minimizer Sampling Improves Long Read Mapping". In: *Bioinformatics* 36 (Supplement_1), pp. i111–i118. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btaa435 (cit. on pp. 30, 73).

Li, Heng (Sept. 15, 2018). "Minimap2: Pairwise Alignment for Nucleotide Sequences". In: *Bioinformatics* 34.18, pp. 3094–3100. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bty191 (cit. on pp. 18, 26, 73).

Yang, Chen et al. (Apr. 1, 2017). "NanoSim: Nanopore Sequence Read Simulator Based on Statistical Characterization". In: *GigaScience* 6.4. ISSN: 2047-217X. DOI: 10.1093/gigascience/gix010 (cit. on pp. 26, 73).

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 (Castresana 2000) 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 (McGinnis et al. 2018).

B.1.2. Classifiers

We used classifier implementations from the scikit-learn python library (Pedregosa et al. 2011), `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 (Virtanen et al. 2020), and corrected p-values for multiple testing with the statsmodels python library (Seabold and Perktold 2010) 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 (Brodersen et al. 2010).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 (Nguyen Xuan Vinh, Julien Epps, and Bailey 2010). 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 (N. X. Vinh and J. Epps 2009).

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

MI was used to compute the G statistic, which follows the chi-square distribution under the null hypothesis (Harremoes 2014). 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 (Pedregosa et al. 2011) 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 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.

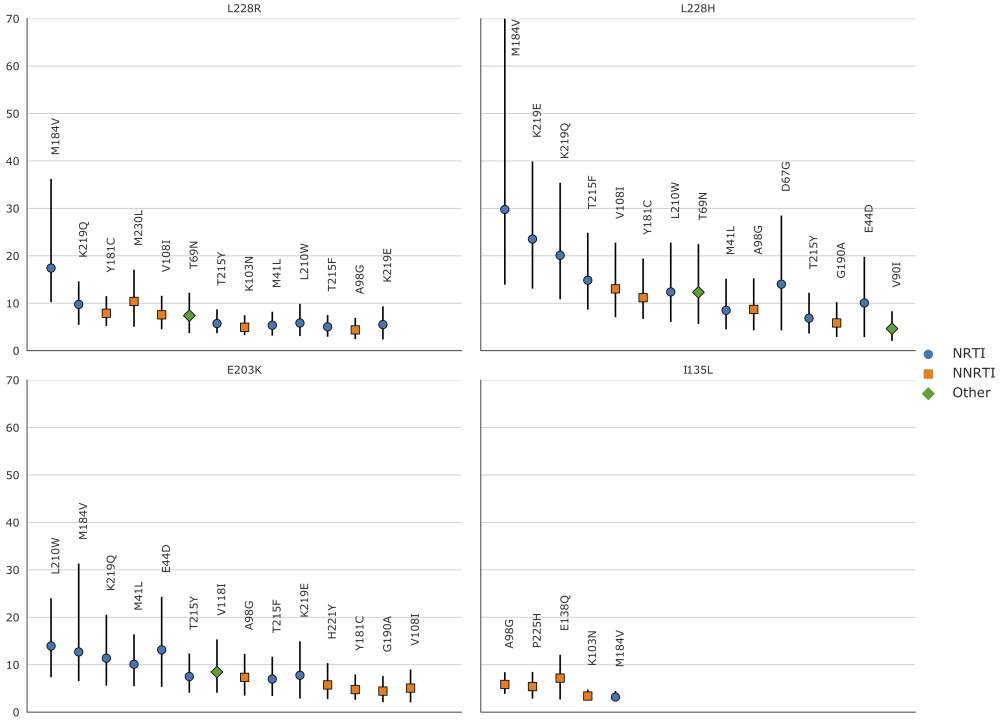

Figure B.1.: **Relative risks of the new mutations with regards to known RAMs on the African dataset**
(i.e. the prevalence of the new mutation in sequences with a given RAM divided by the prevalence of the new mutation in sequences without the RAM). RRs were only computed for mutations (new and RAMs)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

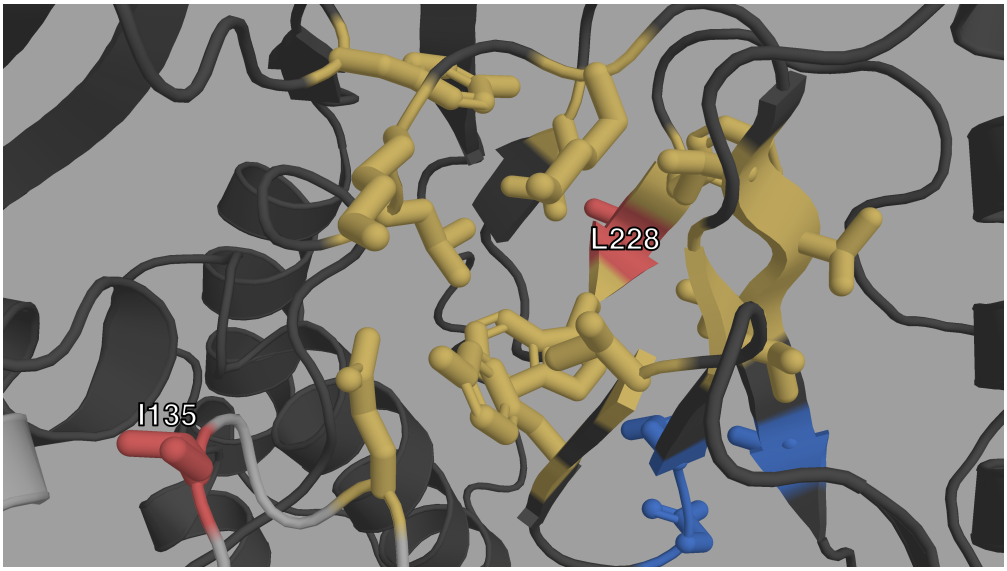

Figure B.2.: **Closeup structural view of the entrance of the NNIBP of HIV-1 RT**
The p66 subunit is colored in dark gray, the p51 subunit in light gray. The NNIBP is highlighted in yellow. The active site is colored in blue. We can see the physical proximity of I135 (red) to the entrance of the NNIBP.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

B.4. S3 Fig.

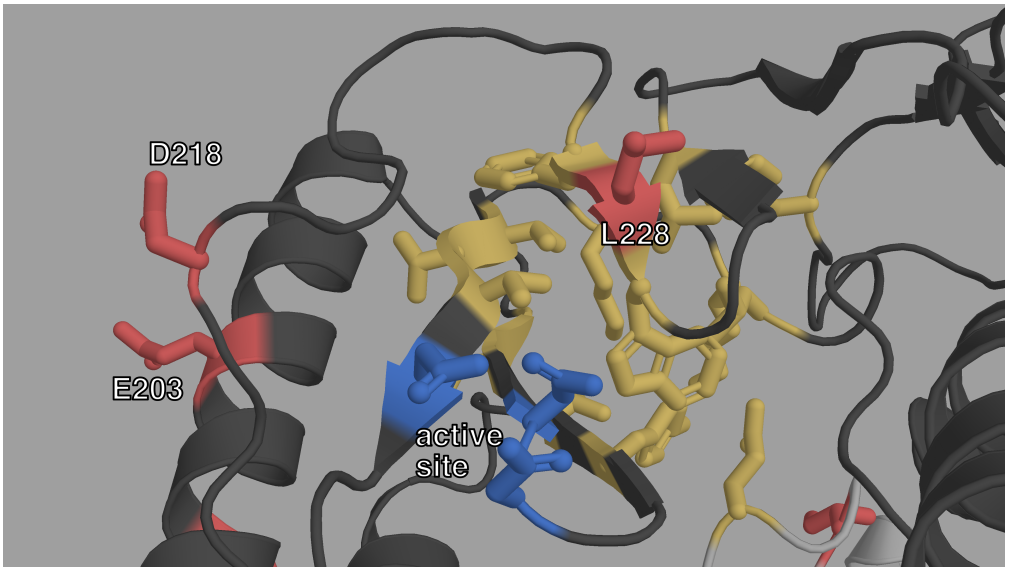

Figure B.3.: **Closeup structural view of the active site of HIV-1 RT.** The p66 subunit is colored in dark gray, the p51 subunit in light gray. The active site is highlighted in blue. The NNIBP is colored in yellow. L228, E203 and D218 (red) are also very close on either side of the active site.

B.5. S1 Table.

Detailed table of "new mutation" characteristics.

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.

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/).

References for Appendix B

Brodersen, Kay Henning et al. (Aug. 2010). "The Balanced Accuracy and Its Posterior Distribution". In: *2010 20th International Conference on Pattern Recognition*, pp. 3121–3124. DOI: 10.1109/ICPR.2010.764 (cit. on pp. 52, 84).

Castresana, J. (Apr. 2000). "Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis". en. In: *Molecular Biology and Evolution* 17.4, pp. 540–552. ISSN: 0737-4038. DOI: 10.1093/oxfordjournals.molbev.a026334 (cit. on p. 84).

Harremoes, Peter (June 2014). "Mutual Information of Contingency Tables and Related Inequalities". In: *2014 IEEE International Symposium on Information Theory*. Honolulu, HI, USA: IEEE, pp. 2474–2478. ISBN: 978-1-4799-5186-4. DOI: 10.1109/ISIT.2014.6875279 (cit. on p. 85).

McGinnis, Will et al. (Jan. 2018). *Scikit-Learn-Contrib/Categorical-Encoding: Release For Zenodo*. Zenodo. DOI: 10.5281/ZENODO.1157110 (cit. on p. 84).

Pedregosa, Fabian et al. (2011). "Scikit-Learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12.Oct, pp. 2825–2830. ISSN: ISSN 1533-7928 (cit. on pp. 84, 85).

Seabold, Skipper and Josef Perktold (2010). "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, pp. 92–96. DOI: 10.25080/Majora-92bf1922-011 (cit. on p. 84).

Vinh, N. X. and J. Epps (June 2009). "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*, pp. 84–91. DOI: 10.1109/BIBE.2009.19 (cit. on p. 85).

Vinh, Nguyen Xuan, Julien Epps, and James Bailey (2010). “Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance”. en. In: *Journal of Machine Learning Research* 11, p. 18 (cit. on pp. 52, 85).

Virtanen, Pauli et al. (Mar. 2020). “SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python”. en. In: *Nature Methods* 17.3, pp. 261–272. ISSN: 1548-7105. DOI: 10.1038/s41592-019-0686-2 (cit. on p. 84).

Global References

Adams, M. D. et al. (Mar. 24, 2000). "The Genome Sequence of Drosophila Melanogaster". In: *Science (New York, N.Y.)* 287.5461, pp. 2185–2195. ISSN: 0036-8075. DOI: 10.1126/science.287.5461.2185. pmid: 10731132 (cit. on p. 25).

Alvarez Melis, David and Tommi Jaakkola (2018). "Towards Robust Interpretability with Self-Explaining Neural Networks". In: *Advances in Neural Information Processing Systems 31*. Ed. by S. Bengio et al. Curran Associates, Inc., pp. 7775–7784 (cit. on p. 51).

Araya, Seare Tesfamichael and Scott Hazelhurst (Jan. 2009). "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, pp. 62–72. ISSN: 0035-919X. DOI: 10.1080/00359190909519238 (cit. on p. 46).

Au, Kin Fai et al. (Oct. 4, 2012). "Improving PacBio Long Read Accuracy by Short Read Alignment". In: *PLOS ONE* 7.10, e46679. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0046679 (cit. on p. 18).

Beerenwinkel, N. et al. (Nov. 2001). "Geno2pheno: Interpreting Genotypic HIV Drug Resistance Tests". In: *IEEE Intelligent Systems* 16.6, pp. 35–41. ISSN: 1941-1294. DOI: 10.1109/5254.972080 (cit. on p. 46).

Beerenwinkel, Niko et al. (July 2003). "Geno2pheno: Estimating Phenotypic Drug Resistance from HIV-1 Genotypes". In: *Nucleic Acids Research* 31.13, pp. 3850–3855. ISSN: 0305-1048. DOI: 10.1093/nar/gkg575 (cit. on p. 46).

Bennett, Diane E. et al. (Mar. 2009). "Drug Resistance Mutations for Surveillance of Transmitted HIV-1 Drug-Resistance: 2009 Update". en. In: *PLOS ONE* 4.3, e4724. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0004724 (cit. on p. 45).

Bragg, Lauren et al. (May 2012). "Fast, Accurate Error-Correction of Amplicon Pyrosequences Using Acacia". In: *Nature Methods* 9.5 (5), pp. 425–426. ISSN: 1548-7105. DOI: 10.1038/nmeth.1990 (cit. on p. 18).

Breiman, Leo (Oct. 2001). "Random Forests". en. In: *Machine Learning* 45.1, pp. 5–32. ISSN: 1573-0565. DOI: 10.1023/A:1010933404324 (cit. on p. 51).

Brier, Glenn W. (Jan. 1950). "Verification of Forecasts Expressed in Terms of Probability". en. In: *Monthly Weather Review* 78.1, pp. 1–3. ISSN: 0027-0644 (cit. on pp. 48, 52).

Brodersen, Kay Henning et al. (Aug. 2010). "The Balanced Accuracy and Its Posterior Distribution". In: *2010 20th International Conference on Pattern Recognition*, pp. 3121–3124. DOI: 10.1109/ICPR.2010.764 (cit. on pp. 52, 84).

Brown, Andrew J. Leigh et al. (Nov. 2000). "Reduced Susceptibility of Human Immunodeficiency Virus Type 1 (HIV-1) from Patients with Primary HIV Infection to Nonnu-

cleoside Reverse Transcriptase Inhibitors Is Associated with Variation at Novel Amino Acid Sites". In: *Journal of Virology* 74.22, pp. 10269–10273. ISSN: 0022-538X (cit. on p. 63).

Bzikadze, Andrey V. and Pavel A. Pevzner (Nov. 2020). "Automated Assembly of Centromeres from Ultra-Long Error-Prone Reads". In: *Nature Biotechnology* 38.11 (11), pp. 1309–1316. ISSN: 1546-1696. DOI: 10.1038/s41587-020-0582-4 (cit. on p. 73).

Castresana, J. (Apr. 2000). "Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis". en. In: *Molecular Biology and Evolution* 17.4, pp. 540–552. ISSN: 0737-4038. DOI: 10.1093/oxfordjournals.molbev.a026334 (cit. on p. 84).

Clark, Shauna A. et al. (Apr. 2006). "Reverse Transcriptase Mutations 118I, 208Y, and 215Y Cause HIV-1 Hypersusceptibility to Non-Nucleoside Reverse Transcriptase Inhibitors". en-US. In: *AIDS* 20.7, pp. 981–984. ISSN: 0269-9370. DOI: 10.1097/01.aids.0000222069.14878.44 (cit. on p. 63).

De Luca, Andrea et al. (Dec. 2007). "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, pp. 1645–1653. ISSN: 0022-1899. DOI: 10.1086/522231 (cit. on p. 63).

Dohm, Juliane C et al. (June 1, 2020). "Benchmarking of Long-Read Correction Methods". In: *NAR Genomics and Bioinformatics* 2.2. ISSN: 2631-9268. DOI: 10.1093/nargab/lqaa037 (cit. on p. 18).

Drăghici, Sorin and R. Brian Potter (Jan. 2003). "Predicting HIV Drug Resistance with Neural Networks". In: *Bioinformatics* 19.1, pp. 98–107. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/19.1.98 (cit. on p. 46).

Dudoit, Sandrine and Mark J. van der Laan (Dec. 2007). *Multiple Testing Procedures with Applications to Genomics*. en. Springer Science & Business Media. ISBN: 978-0-387-49317-6 (cit. on p. 45).

Ekim, Barış, Bonnie Berger, and Rayan Chikhi (Oct. 20, 2021). "Minimizer-Space de Bruijn Graphs: Whole-Genome Assembly of Long Reads in Minutes on a Personal Computer". In: *Cell Systems* 12.10, 958–968.e6. ISSN: 2405-4712. DOI: 10.1016/j.cels.2021.08.009 (cit. on p. 18).

Gascuel, Olivier et al. (Aug. 1998). "Twelve Numerical, Symbolic and Hybrid Supervised Classification Methods". In: *International Journal of Pattern Recognition and Artificial Intelligence* 12.05, pp. 517–571. ISSN: 0218-0014. DOI: 10.1142/S0218001498000336 (cit. on pp. 48, 51).

Goeman, Jelle J. and Aldo Solari (2014). "Multiple Hypothesis Testing in Genomics". en. In: *Statistics in Medicine* 33.11, pp. 1946–1978. ISSN: 1097-0258. DOI: 10.1002/sim.6082 (cit. on p. 50).

Graham, Ronald L., Donald Ervin Knuth, and Oren Patashnik (1994). *Concrete Mathematics: A Foundation for Computer Science*. 2nd ed. Reading, Mass: Addison-Wesley. 657 pp. ISBN: 978-0-201-55802-9 (cit. on p. 24).

Gusfield, Dan (1997). "Algorithms on stings, trees, and sequences: Computer science and computational biology". In: *Acm Sigact News* 28.4, pp. 41–60 (cit. on p. 18).

Hammond, Jennifer et al. (Dec. 1998). "Mutations in Retroviral Genes Associated with Drug Resistance". en. In: *Human retroviruses and AIDS*, pp. 11136–11179 (cit. on p. 45).

Harremoes, Peter (June 2014). "Mutual Information of Contingency Tables and Related Inequalities". In: *2014 IEEE International Symposium on Information Theory*. Honolulu, HI, USA: IEEE, pp. 2474–2478. ISBN: 978-1-4799-5186-4. DOI: 10.1109/ISIT.2014.6875279 (cit. on p. 85).

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (Aug. 2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. en. Springer Science & Business Media. ISBN: 978-0-387-84858-7 (cit. on p. 51).

Heider, Dominik et al. (Aug. 2013). "Multilabel Classification for Exploiting Cross-Resistance Information in HIV-1 Drug Resistance Prediction". In: *Bioinformatics* 29.16, pp. 1946–1952. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btt331 (cit. on p. 46).

Hué, Stéphane et al. (Mar. 2009). "Demonstration of Sustained Drug-Resistant Human Immunodeficiency Virus Type 1 Lineages Circulating among Treatment-Naïve Individuals". en. In: *Journal of Virology* 83.6, pp. 2645–2654. ISSN: 0022-538X, 1098-5514. DOI: 10.1128/JVI.01556-08 (cit. on p. 45).

Jain, Chirag et al. (July 1, 2020). "Weighted Minimizer Sampling Improves Long Read Mapping". In: *Bioinformatics* 36 (Supplement_1), pp. i111–i118. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btaa435 (cit. on pp. 30, 73).

Lengauer, Thomas and Tobias Sing (Oct. 2006). "Bioinformatics-Assisted Anti-HIV Therapy". en. In: *Nature Reviews Microbiology* 4.10, pp. 790–797. ISSN: 1740-1534. DOI: 10.1038/nrmicro1477 (cit. on p. 46).

Lepri, Alessandro Cozzi et al. (Mar. 2000). "Resistance Profiles in Patients with Viral Rebound on Potent Antiretroviral Therapy". en. In: *The Journal of Infectious Diseases* 181.3, pp. 1143–1147. ISSN: 0022-1899. DOI: 10.1086/315301 (cit. on p. 45).

Li, Heng (Sept. 15, 2018). "Minimap2: Pairwise Alignment for Nucleotide Sequences". In: *Bioinformatics* 34.18, pp. 3094–3100. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bty191 (cit. on pp. 18, 26, 73).

– (Aug. 7, 2021). *New Strategies to Improve Minimap2 Alignment Accuracy*. arXiv: 2108.03515 [q-bio] (cit. on p. 32).

Li, Heng et al. (Aug. 2018). "A Synthetic-Diploid Benchmark for Accurate Variant-Calling Evaluation". In: *Nature Methods* 15.8 (8), pp. 595–597. ISSN: 1548-7105. DOI: 10.1038/s41592-018-0054-7 (cit. on p. 32).

Liu, Hailin et al. (Mar. 8, 2021). "SMARTdenovo: A de Novo Assembler Using Long Noisy Reads". In: *Gigabyte* 2021, pp. 1–9. DOI: 10.46471/gigabyte.15 (cit. on p. 18).

Maddison, Wayne P. and Richard G. FitzJohn (Jan. 2015). "The Unsolved Challenge to Phylogenetic Correlation Tests for Categorical Characters". en. In: *Systematic Biology* 64.1, pp. 127–136. ISSN: 1063-5157. DOI: 10.1093/sysbio/syu070 (cit. on p. 45).

Marcelin, Anne-Genevieve et al. (2006). "Impact of HIV-1 Reverse Transcriptase Polymorphism at Codons 211 and 228 on Virological Response to Didanosine". en. In: *Antiviral Therapy*, p. 8 (cit. on p. 63).

McGinnis, Will et al. (Jan. 2018). *Scikit-Learn-Contrib/Categorical-Encoding: Release For Zenodo*. Zenodo. DOI: 10.5281/ZENODO.1157110 (cit. on p. 84).

Mikheenko, Alla et al. (July 1, 2020). "TandemTools: Mapping Long Reads and Assessing/Improving Assembly Quality in Extra-Long Tandem Repeats". In: *Bioinformatics* 36 (Supplement_1), pp. i75–i83. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btaa440 (cit. on p. 26).

Miller, Jason R. et al. (Dec. 15, 2008). "Aggressive Assembly of Pyrosequencing Reads with Mates". In: *Bioinformatics* 24.24, pp. 2818–2824. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btn548 (cit. on p. 18).

Mooney, Alyssa C. et al. (July 2018). "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, pp. 2368–2379. ISSN: 1573-3254. DOI: 10.1007/s10461-018-2155-9 (cit. on p. 48).

Mourad, Raphaël et al. (Sept. 2015). "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, pp. 1917–1925. ISSN: 0269-9370. DOI: 10.1097/QAD.0000000000000768 (cit. on p. 45).

Nebbia, G. et al. (May 2007). "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, pp. 1013–1016. ISSN: 0305-7453. DOI: 10.1093/jac/dkm067 (cit. on p. 63).

Nurk, Sergey, Sergey Koren, et al. (2021). "The Complete Sequence of a Human Genome". In: *bioRxiv : the preprint server for biology*. DOI: 10.1101/2021.05.26.445798 (cit. on pp. 19, 25).

Nurk, Sergey, Brian P. Walenz, et al. (Jan. 9, 2020). "HiCanu: Accurate Assembly of Segmental Duplications, Satellites, and Allelic Variants from High-Fidelity Long Reads". In: *Genome Research* 30.9, pp. 1291–1305. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.263566.120. pmid: 32801147 (cit. on p. 18).

Pedregosa, Fabian et al. (2011). "Scikit-Learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12.Oct, pp. 2825–2830. ISSN: ISSN 1533-7928 (cit. on pp. 84, 85).

Prodanov, Timofey and Vikas Bansal (Nov. 4, 2020). "Sensitive Alignment Using Paralogous Sequence Variants Improves Long-Read Mapping and Variant Calling in Segmental Duplications". In: *Nucleic Acids Research* 48.19, e114. ISSN: 0305-1048. DOI: 10.1093/nar/gkaa829 (cit. on p. 32).

Rennie, Jason D et al. (2003). "Tackling the Poor Assumptions of Naive Bayes Text Classifiers". en. In: *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pp. 616–623 (cit. on p. 50).

Rhee, Soo-Yon et al. (May 2007). "HIV-1 Subtype B Protease and Reverse Transcriptase Amino Acid Covariation". en. In: *PLOS Computational Biology* 3.5, e87. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.0030087 (cit. on p. 63).

Rhie, Arang et al. (Sept. 14, 2020). "Merqury: Reference-Free Quality, Completeness, and Phasing Assessment for Genome Assemblies". In: *Genome Biology* 21.1, p. 245. ISSN: 1474-760X. DOI: 10.1186/s13059-020-02134-9 (cit. on p. 30).

Riemenschneider, Mona et al. (Feb. 2016). "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, p. 10. ISSN: 1756-0381. DOI: 10.1186/s13040-016-0089-1 (cit. on p. 46).

Sahlin, Kristoffer and Paul Medvedev (Apr. 1, 2020). "De Novo Clustering of Long-Read Transcriptome Data Using a Greedy, Quality Value-Based Algorithm". In: *Journal of Computational Biology* 27.4, pp. 472–484. DOI: 10.1089/cmb.2019.0299 (cit. on p. 18).

– (Jan. 4, 2021). "Error Correction Enables Use of Oxford Nanopore Technology for Reference-Free Transcriptome Analysis". In: *Nature Communications* 12.1 (1), p. 2. ISSN: 2041-1723. DOI: 10.1038/s41467-020-20340-8 (cit. on p. 18).

Sanger, F., G. M. Air, et al. (Feb. 1977). "Nucleotide Sequence of Bacteriophage X174 DNA". In: *Nature* 265.5596 (5596), pp. 687–695. ISSN: 1476-4687. DOI: 10.1038/265687a0. URL: https://www.nature.com/articles/265687a0 (visited on 05/17/2022) (cit. on p. 12).

Sanger, F., S. Nicklen, and A. R. Coulson (Dec. 1977). "DNA Sequencing with Chain-Terminating Inhibitors". In: *Proceedings of the National Academy of Sciences* 74.12, 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. 12).

Saracino, A. et al. (2006). "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, pp. 9–17. ISSN: 1096-9071. DOI: 10.1002/jmv.20500 (cit. on p. 63).

Sarafianos, Stefan G. et al. (Jan. 2009). "Structure and Function of HIV-1 Reverse Transcriptase: Molecular Mechanisms of Polymerization and Inhibition". In: *Journal of molecular biology* 385.3, pp. 693–713. ISSN: 0022-2836. DOI: 10.1016/j.jmb.2008.10.071 (cit. on p. 60).

Schrödinger, LLC (Nov. 2015). "The PyMOL Molecular Graphics System, Version 1.8" (cit. on p. 60).

Seabold, Skipper and Josef Perktold (2010). "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, pp. 92–96. DOI: 10.25080/Majora-92bf1922-011 (cit. on p. 84).

Sham, Pak C. and Shaun M. Purcell (May 2014). "Statistical Power and Significance Testing in Large-Scale Genetic Studies". en. In: *Nature Reviews Genetics* 15.5, pp. 335–346. ISSN: 1471-0064. DOI: 10.1038/nrg3706 (cit. on p. 46).

Sheik Amamuddy, Olivier, Nigel T. Bishop, and Özlem Tastan Bishop (Aug. 2017). "Improving Fold Resistance Prediction of HIV-1 against Protease and Reverse Transcriptase Inhibitors Using Artificial Neural Networks". In: *BMC Bioinformatics* 18.1, p. 369. ISSN: 1471-2105. DOI: 10.1186/s12859-017-1782-x (cit. on p. 46).

Shen, ChenHsiang et al. (Aug. 2016). "Automated Prediction of HIV Drug Resistance from Genotype Data". In: *BMC Bioinformatics* 17.8, p. 278. ISSN: 1471-2105. DOI: 10.1186/s12859-016-1114-6 (cit. on p. 46).

Smith, Lloyd M., Steven Fung, et al. (Apr. 11, 1985). "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, 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. 12).

Smith, Lloyd M., Jane Z. Sanders, et al. (June 1986). "Fluorescence Detection in Automated DNA Sequence Analysis". In: *Nature* 321.6071 (6071), 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. 12).

Steiner, Margaret C., Keylie M. Gibson, and Keith A. Crandall (May 2020). "Drug Resistance Prediction Using Deep Learning Techniques on HIV-1 Sequence Data". en. In: *Viruses* 12.5, p. 560. DOI: 10.3390/v12050560 (cit. on p. 46).

Tibshirani, Robert (1996). "Regression Shrinkage and Selection Via the Lasso". en. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1, pp. 267–288. ISSN: 2517-6161. DOI: 10.1111/j.2517-6161.1996.tb02080.x (cit. on pp. 48, 50).

Verhofstede, Chris et al. (2007). "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, pp. 1254–1260. ISSN: 1096-9071. DOI: 10.1002/jmv.20950 (cit. on p. 45).

Villabona-Arenas, Christian Julian et al. (Nov. 2016). "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, p. 2577. ISSN: 0269-9370. DOI: 10.1097/QAD.0000000000001233 (cit. on pp. 45–48, 50, 51, 55).

Vinh, N. X. and J. Epps (June 2009). "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*, pp. 84–91. DOI: 10.1109/BIBE.2009.19 (cit. on p. 85).

Vinh, Nguyen Xuan, Julien Epps, and James Bailey (2010). "Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance". en. In: *Journal of Machine Learning Research* 11, p. 18 (cit. on pp. 52, 85).

Virtanen, Pauli et al. (Mar. 2020). "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python". en. In: *Nature Methods* 17.3, pp. 261–272. ISSN: 1548-7105. DOI: 10.1038/s41592-019-0686-2 (cit. on p. 84).

Wensing, A. M. et al. (Dec. 2016). "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, pp. 132, 132–133. ISSN: 2161-5861 (cit. on p. 45).

Wu, Tong Tong et al. (Mar. 2009). "Genome-Wide Association Analysis by Lasso Penalized Logistic Regression". en. In: *Bioinformatics* 25.6, pp. 714–721. ISSN: 1460-2059, 1367-4803. DOI: 10.1093/bioinformatics/btp041 (cit. on p. 64).

Yang, Chen et al. (Apr. 1, 2017). "NanoSim: Nanopore Sequence Read Simulator Based on Statistical Characterization". In: *GigaScience* 6.4. ISSN: 2047-217X. DOI: 10.1093/gigascience/gix010 (cit. on pp. 26, 73).

Yu, Xiaxia, Irene T. Weber, and Robert W. Harrison (July 2014). "Prediction of HIV Drug Resistance from Genotype with Encoded Three-Dimensional Protein Structure".

In: *BMC Genomics* 15.5, S1. ISSN: 1471-2164. DOI: 10.1186/1471-2164-15-S5-S1 (cit. on p. 46).

Zhang, Jie et al. (Apr. 2005). "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, pp. 439–444. ISSN: 1525-4135. DOI: 10.1097/01.qai.0000147526.64863.53 (cit. on p. 46).

Zhang, Quanshi, Ying Nian Wu, and Song-Chun Zhu (2018). "Interpretable Convolutional Neural Networks". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8827–8836 (cit. on p. 51).

Zhukova, Anna et al. (Dec. 2017). "The Role of Phylogenetics as a Tool to Predict the Spread of Resistance". en. In: *The Journal of Infectious Diseases* 216.suppl_9, S820–S823. ISSN: 0022-1899. DOI: 10.1093/infdis/jix411 (cit. on p. 45).