# Chapter 4

**Development of Seq2Res: A computational Drug Resistance Testing Pipeline for the Analysis and Management of HIV-1 Roche/454 Sequence Data**

## 4.1 Introduction

UNAIDS estimates that approximately 34 million individuals - including 2.6 million newly infected - were living with HIV and as many as 1.6 million died of HIV-related illnesses in the year 2011 alone (UNAIDS, 2012). The Sub-Saharan African region has the highest prevalence of HIV infections (Asamoah-Odei et al., 2004). Antiretroviral therapy (ART) programmes have been established over the entire region (Blower et al., 2005; Herbst et al., 2009; Nattrass, 2006; Stringer et al., 2006) with the aim of suppressing viral replication, resulting in a lower viral load (Autran et al., 1997; Li et al., 1998; Mocroft et al., 2010) and thereby extending the life expectancy of HIV positive individuals (Fang et al., 2007; Harrison et al., 2010; Mills et al., 2011). By the end of 2011, 8 million people from low and middle-income countries were receiving the life saving drugs (20 times more than in 2003) (UNAIDS, 2012; WHO, 2011).

In the order of 1010 new viruses are produced per replication cycle with a mutation rate in the order of 10-5 per nucleotide per cycle (Ho et al., 1995; Mansky, 1996a, b, 1998; Mansky and Temin, 1995). The high mutation rate is due to the error prone reverse transcriptase enzyme that transcribes viral RNA to cDNA during the replication (Bebenek et al., 1989; Berkhout et al., 2001; Preston et al., 1988; Roberts et al., 1988). The accumulation of random mutations in HIV leads to development of drug resistance (Johnson et al., 2008).

With the scale up of antiretroviral (ARV) drugs there is growing evidence suggesting that drug resistant HIV can exist as minor variants in individuals undergoing treatment (Aghokeng et al., 2011; Dudley et al., 2012; Gupta et al., 2012; Lataillade et al., 2010; Li et al., 2011; Little et al., 2002; Simen et al., 2009; Yebra et al., 2011). This observation is also supported by several research studies on population-level surveillance of HIV drug resistant variants (Dudley et al., 2012; Hamers et al., 2012; Hamers et al., 2011a; Hamers et al., 2011b; Lataillade et al., 2010; Li et al., 2011; Little et al., 2002; Simen et al., 2009). The most likely reasons for the increase in drug resistant viral variants in the general HIV-infected population is poor adherence during therapy (Bangsberg et al., 2003; Golin et al., 2002; Low-Beer et al., 2000) and an increased rate of transmission of resistant viral variants (Hamers et al., 2011a; Jakobsen et al., 2010. The rising prevalence of drug resistant variants narrows treatment options [Hamers, 2012 #1232; Supervie et al., 2010; Wittkop et al., 2011; Yerly et al., 1999). Thus, ARV treatment, management and surveillance of drug resistant HIV variants (Bennett et al., 2009) is essential for prolonging the usage of, and optimizing the outcome from, a particular drug cocktail (Adetunji et al., 2012). Therefore, the World Health Organization (WHO) guideline ([http://www.**who**.int/**hiv**/drugresistance/](http://www.who.int/hiv/drugresistance/)) recommends that a pretreatment HIV drug resistance test is carried out, but this has only been possible in resource-rich countries (Aberg et al., 2009; Clumeck et al., 2008; Gazzard et al., 2008; Kaplan et al., 2009; Masur and Kaplan, 2009; Perfect et al., 2010) .

Both individual and population-level screening of resistant HIV variants must be done routinely for tracking of resistant viruses and to prevent the exhaustion of ART treatment options. However, due to the cost factor, regular pretreatment resistance testing in low and middle-income countries, was not possible until the advent of high throughput sequencing (HTS). HTS is a robust, reliable and affordable way to explore clinically relevant low abundance (< 1%) resistant HIV variants (Dudley et al., 2012; Ji et al., 2012; Tsibris et al., 2009; Wang et al., 2007). However, the huge volume of sequence data from HTS is a challenge for data analysis and management.

This chapter describes the development and testing of a computational tool designed to facilitate low cost HIV drug resistance test using HTS technologies.

## 4.2 Methods and Materials

While HTS approaches provide an exciting prospect for enabling high-throughput, low-cost HIV drug resistance genotyping, the sheer volume of data generated by such sequencing platforms means that the bioinformatics requirements for the management, analysis and interpretation of the data is immense. The use of HTS for routine HIV drug resistance genotyping requires a bioinformatics platform that can facilitate fast and sensitive analysis of data by individuals, such as clinicians and wet-lab researchers, with little, or no, bioinformatics expertise.

Further, the rapidly evolving field of high-throughput sequencing means that any bioinformatics platform must be robust and easy to adapt to analyze data from new sequencing platforms.

Thus, Seq2Res has been developed (mostly using the Python high level programming language) in a modularized format, keeping each distinct analysis block independent of the other, thereby enabling easy insertion of new modules to allow the analysis of new data-types. Given that the vast majority of HIV resistance data has been generated using the Roche/454 platform, Seq2Res has been initially developed to analyze such data.

### 4.2.1 Structure of HTS raw sequence reads

The sequence reads obtained directly from the sequencing instrument, before preprocessing are the raw sequence reads. All raw sequence reads in a file are in the 5’ to 3’ end orientation as nucleotides are added to the 3’ end during sequencing. A raw HTS sequence read consists of key sequence, an MID (Multiplex Identifier) sequence to label a sample, a primer sequence to identify a specific amplicon region and the actual amplicon region sequence from 5’ to the 3’ end (**Figure 4.1**).

The key sequence consists of four nucleotides that are used by Roche/454 high throughput sequencing platforms to calibrate the measurement of optical emission to count the number of nucleotides added to the growing strand during sequencing. The key sequence is removed from raw sequence reads before analysis.

The key sequence is followed by a sequence to identify a sample. In the event that multiple samples are sequenced concurrently, each sample is tagged with a unique identifier sequence called MID. Roche/454 has approximately 150 standard MID sequences. The MID sequences are usually ten nucleotides long, however, MIDs can be designed using any unique custom sequence. The MID sequence is removed from further analysis after separating the sequences in MID, and thus sample, categories.

The MID sequence is followed by a short sequence to identify the genomic amplicon region called a primer. A genomic region can be sequenced in both directions and therefore the primer can be either in a forward or reverse orientation. Primer sequences are designed to bind specifically at the 5’ region of its associated genomic region’s forward (forward primer) and reverse (reverse primer) strands.

The primer sequence is followed by a genomic amplicon sequence. This is the sequence that is of interest and is analyzed.

### 4.2.2 Seq2Res: Required Data

One of the primary focuses in the development of Seq2Res was to make usage as easy as possible by keeping the required number of files and parameters to a minimum. Users are therefore only required to provide the raw sequence data file, the primer information (multiple amplicons can be analysed concurrently) and, if present, the MID used together with their associated patient identifier.

### 4.2.2.1: Raw sequence reads file

The raw data file can be in a number of formats:

1. The standard flowgram format (sff) file.
2. FASTQ format
3. FASTA format together with the associated QUAL file.

### 4.2.2.2: Primer file containing amplicon primers

The primer file is a five column, tab-delimited file containing the name of each amplicon in the first column, the forward and reverse primer sequences in the second and third column, and the amplicon start and end positions relative to the HXB2 HIV *pol* reference sequence in the fourth and fifth columns, respectively (**Figure 4.2 A**).

### 4.2.2.3: MID file containing sample identifiers

The MID file contains two tab-delimited columns of data of which the first column contains the Roche/454 standard MID name (or the actual nucleotide sequence of the MID used) with the second column containing the sample name associated with that MID (**Figure 4.2 B**).

### 4.2.2.4: Threshold Prevalence cutoff

The threshold prevalence cutoff is the minimum percentage of the viral sequence reads in a sample that are predicted resistant, intermediately resistance or sensitive to a drug, which is required to call a viral sample as resistant or intermediate or susceptible respectively to the antiretroviral drugs. The percentage of the viral sequences is checked in the order that is predicted resistant, intermediately resistant and sensitive to a drug. If the percentage of sequence reads predicted as resistant to a drug is greater or equal to the cutoff, the viral population in the sample is called as resistant to the drug. If it is less than the cutoff, then the percentage of sequence reads predicted intermediate resistance to a drug is checked. If this is greater or equal to the cutoff value, then the viral population in a sample is called as intermediate resistant. If the percentage of intermediate resistant sequences is also less than cutoff value, the viral population in a sample is called as susceptible to the drug. The default prevalence cutoff is 15%.

### 4.2.3: Advanced parameters

While the three required files are sufficient to run analysis in Seq2Res, advance end-users have an option to set the following required preprocessing parameters:

4.2.3.1: Quality-trimming parameters

These parameters are supplied to the trimming tool. QTrim iteratively trims out poor quality nucleotides from the ends of a sequence, based on the mean of quality scores across the sequence. The quality trimming parameters that end-users can set are:

1. Mean quality: This is a minimum mean quality score across a sequence read. Every sequence read must satisfy this criterion for further analysis to take place. If a sequence does not satisfy the mean quality criteria, it is removed. The default mean quality score is 20.
2. Minimum read length: This is the minimum length, counted in base pairs, required for a sequence read in order for further analysis to take place. If a sequence does not meet the minimum read length cutoff, it is removed. The default minimum read length is 50 base pairs.
3. Mode of trimming: Users can set two modes of trimming, which are a) trimming from 3’ end, or b) trimming from 5’ and 3’ ends.

4.2.3.2: Demultiplex parameters

The demultiplex parameters refer to the primer and MID tolerance, as well as the key sequence length.

1. Primer tolerance refers to the number of nucleotide mismatches between the user-supplied primer and the primer in a sequence read. For an amplicon of a sequence to be identified, a primer has to appear in the sequence. But due to sequencing errors, the primers may not appear exactly as supplied in the primer file. The primer tolerance allows the amplicon of the sequence to be identified although there are some errors, less or equal to, the primer tolerance. The default primer tolerance is 3.
2. MID tolerance refers to the number of nucleotide mismatches between the user supplied MID and the MID in sequence read. Similar to the primer tolerance, some errors in an MID sequence can be tolerated. The default MID tolerance is 2.
3. Key sequence length: This is usually a nucleotide sequence of 4 unique bases at the 5’ end of a raw sequence read. The key sequence is not a part of sequence reads and is removed. The default value is 4. If no key sequence is present in sequence reads, a value of zero must be supplied.

### 4.2.3 Development and processing of data using Seq2Res

The steps undertaken by each module of Seq2Res are described below. The output of each module serves as the input for the next and, thus, makes for easy swapping/replacement of modules in future versions (**Figure 4.3**).

1. **Pre-processing of submitted files**

Depending on the input file, Seq2Res may need to first pre-process the file. Since the subsequent steps in the Seq2Res pipeline work only with FASTQ files, the other formats are converted to FASTQ format during the pre-processing. An SFF file is converted to a FASTQ file with a tool called sff\_extractor (<http://bioinf.comav.upv.es/sff_extract/index.htm>) while a FASTA file with a paired quality scores file is merged to a FASTQ file.

1. **Processing optimal full length positions**

The start and end nucleotide positions that defines the full length of the amplicons are supplied in the primer file. Seq2Res processes these supplied positions to associate the start and end codon positions with the standard HIV *pol* reference sequence. Because Seq2Res only counts the drug resistant codon positions in the amplicons, the start and end codon positions for full length are processed to find the first and last drug resistant codon positions in the amplicons (**Figure 4.4**). The start and end codon positions for the amplicons are redefined by the first and last drug resistant codon positions in the amplicons respectively and this new start and end codon positions are defined as the optimal full-length positions. Seq2Res considers all the amplicon sequence reads covering optimal full-length positions for downstream analysis. This is a critical step as there may be a large number of sequence reads that are not full length according to the amplicon borders but satisfy the optimal full-length criteria.

1. **Demultiplexing**

Seq2Res utilizes Biopython ([www.biopython.org](http://www.biopython.org)) (Cock et al., 2009) packages to read individual sequence reads from the FASTQ file and removes any key sequence, if present, from the sequence reads. For the first step of the demultiplexing Seq2Res searches for the MID in every sequence read which is located at the 5’ region of a sequence read. A subsequence, of length that is equal to MID length, is obtained from 5’ end of the sequence read. The subsequence is then pair-wise global aligned with every MIDs in the input list of MIDs. The MID with the number of mismatches less or equal to the MID tolerance in the alignment (default MID tolerance 2) identifies the sample of the sequence. The MID is added at the sequence identifier and the MID subsequence is deleted from the sequence read. Thus, every sequence read is either tagged with the details of the MID in the sequence identifier or is discarded from subsequent analysis.

Next, Seq2Res searches for the primer in each sequence read in a similar way to identify the MID. A subsequence equal to a primer length is obtained from 5’ end of the sequence read. The subsequence is then pair-wise globally aligned with every primer in the input list of forward and reverse primers. The primer, either forward or reverse, with the number of mismatches less or equal to user threshold of primer tolerance (the default primer tolerance is 3) is selected and added in the sequence ID along with the strand forward (+) or strand reverse (-).

If the sequence reads are generated using Primer ID technology [Jabara, 2011 #1188], Seq2Res searches by primer, MID and Primer ID using the PIDA algorithm. At this stage the PIDA algorithm also generates the consensus sequences representing each primer ID and it is only these consensus sequences that are passed through for subsequent analysis.

1. **Advanced Sequence Reads Quality Control**

The non-discarded sequences are quality trimmed in the next step. During sequencing, Roche/454 HTS calls a base with a certain probability (Margulies et al., 2005), and this score for each base is saved in the quality file. The scores range from minimum of zero to a maximum of 40. A score of 10 would mean that the probability of a wrong base call is 1 in 10, 20 means 1 in a 100, and 40 means 1 in a 10000 (Cock et al., 2010). In general, the quality scores of a base decrease as the sequence length increases. A lower quality score indicates that a base has a lower probability of being correct and including these poor quality bases would compromise the data analysis. Seq2Res uses QTrim (<https://hiv.sanbi.ac.za/tools/qtrim>) – an in-house designed quality-trimming tool for removing these lower quality bases. Default parameter settings in QTrim are: mean quality score of 20, minimum read length of 50 nucleotides and trimming from 3’ end of a sequence read. Seq2Res allows flexibility to control these default QTrim parameter settings.

1. **Binning of sequence reads by amplicon and MID**

Once sequence reads have been quality trimmed the information contained in the sequence IDs is used to bin groups of sequences on the basis of sample/MID, and then by amplicon. All subsequent analysis is performed on each ‘bin’ of sequences concurrently using a high-performance computing cluster.

1. **Reference Mapping**

The high quality sequence reads in each bin are mapped to the full HXB2 *pol* gene reference sequence. This mapping step is one of the most critical steps of the entire pipeline as it is here that we are able to correct both PCR- and sequencing-induced errors in sequence reads. RAMICS (Rapid Amplicon Mapping In Codon Space) is a tool that has been developed by Imogen Wright (a PhD student in the research group of Prof Simon Travers, manuscript currently under review in Nucleic Acids Research) that maps sequence reads to a reference sequence using hidden Markov models in “codon-space” and is capable of identifying, and accounting for, PCR- and sequencing-induced errors in sequence reads during the mapping process. RAMICS is particularly powerful at homopolymer regions, accurately identifying the over- or under-call of nucleotides. Further, RAMICS is able to identify whether each sequence read is full length and maps to the entire amplicon.

In Seq2Res, RAMICS maps amplicon sequence reads to a subsequence of the reference sequence that correspond to amplicon specific optimal full-length sequence. For example, if an amplicon optimal start and end codon positions are 66 and 100 respectively, which correspond to the HIV *pol* reference sequence, RAMICS copies the nucleotide sequence from codon 66 to 100 of the reference sequence and maps the amplicon sequence reads to it.

RAMICS provides a novel ‘fasta-like’ (fastm) output file (**Figure 4.5**) that, for each read, details what amino acid is present at each position in the read relative to the HXB2 reference sequence. Any single or double nucleotide insertions resulting from PCR or sequencing error are, thus, removed from subsequent analysis. Any full codon sized insertions or deletions are identified and documented in the fastm output file.

1. **Codon position translation**

For every sequence read written to the fastm file, Seq2Res translates each codon position that was produced with respect to the reference sequence (the HXB2 full *pol sequence*) into codon positions relative to the gene(s) that the amplicon covers (protease, reverse transcriptase or integrase). For example, the codon positions 57 and 156 in HIV *pol* sequence are translated to the protease codon position 1 and reverse transcriptase codon position 1 respectively.

For each read, the amino acid present at all positions evaluated as part of the Stanford HIV resistance testing algorithm (Liu and Shafer, 2006) are extracted and saved to a file in the format required for submission to the Stanford algorithm (**Figure 4.6**).

1. **Resistance Testing**

For resistance testing we use a locally installed version of the Stanford HIVdb resistance-testing algorithm (Sierra – kindly provided by Tommy Liu and Robert Schafer) (Liu and Shafer, 2006). For each sequence read contained in the submitted file, the Sierra algorithm produces the drug resistance result in a XML file detailing their resistance level to all drugs– Resistant (R), Intermediate (I) or Susceptible (S) as well as the relevant drug resistant mutations (DRMs) present.

1. **Resistance report presentation**

Seq2Res processes a XML file to obtain sequence name, DRMs and resistance level of the DRMs in each sequence. The information obtained per sequence is stored in a row in a tabulated tab-delimited file. The first ten columns of each row in the file contain user and sequence information. The order of the information in the first ten columns are USERID, JOBID, input file name, sample name, the MID of the sample, a sequence read ID, number of the sequences with that sequence read ID, the amplicon name of the sequence, strand orientation (forward or reverse) of the sequence and the length of the sequence. From column 11, a resistance level (R, I or S) to a specific drug appears at odd columns and the pertinent DRMs to that resistance call follow in each of the even columns. For example, a resistance call “R” to Abacavir (ABC) may appear at column 11 and DRMs K65R from a sequence that are resistant (R) to the first drug ABC may appear in column 12. Similarly, the resistance level and DRMs for the drug didanosine (d4T) appear in column 13 and 14 respectively. This is followed for a defined order of antiretroviral drugs (**Table 4.1**). If a sequence read does not cover DRM codon positions associated with a drug, a “-“ appears in the columns for that drug. For example, if a sequence covers only protease gene region and not reverse transcriptase gene, the table lists out resistance levels and the associated DRMs for PI drugs and “-“ for reverse transcriptase inhibitor drugs. Thus, this step generates an easily searchable table from which all other results and conclusions are generated.

From the table, the number of sequence reads that are predicted to have high resistance (R), intermediate resistance (I) or susceptibility (S) by local Sierra to a particular antiretroviral drug is calculated. The algorithm classifies an antiretroviral drug as either resistant, intermediate or susceptible to the viral population in a sample using the following conditions:

1. If an amplicon covers one gene, the percentage of sequence reads predicted resistant, intermediately resistant and sensitive are calculated. If the prevalence of resistant sequence reads to a drug is greater or equal to the prevalence cutoff, the viral population in the sample is called as resistant to the drug. If the viral sample is not resistant and the prevalence of predicted intermediate sequences reads is greater or equal to the prevalence cutoff, the viral population in a sample is reported as intermediately resistant to the drug. If the viral sample is neither resistant nor intermediate resistant, it is called as susceptible to the associated drug.
2. If more than one amplicon covers one gene, for example X and Y amplicons for the RT gene, the following conditions are applied:
   1. If the percentage resistance for either X or Y is above the prevalence cutoff, the sample is considered resistant to the associated drugs. The reported prevalence is equal to the prevalence of the amplicon with the highest number of resistant calls.
   2. If the percentage resistance for both X and Y are less than the cutoff, the percentage intermediate resistance for either amplicon above the cutoff is reported. As with the resistance prevalence, the reported intermediate resistance is equal to the prevalence of the amplicon with the highest number of intermediate calls.
   3. If the percentage resistance and intermediate resistance for X and Y are both less than the cutoff, the sample is called as susceptible to the associated drug.

For example, if the prevalence cutoff is 15% and two amplicons RT1 and RT2 cover the gene RT. Then

1. If the percentage resistance found for RT1 is 4% and the percentage resistance found for RT2 is 60%, the sample is called as resistant to the associated drug.
2. If for the drug ABC, the percentage resistance found for RT1 amplicon is 16% and the percentage resistance found for RT2 amplicon is 20%, the sample is called as resistant to the associated drug from amplicon RT2.
3. If the percentage intermediate resistance found for of RT1 is 2% and the percentage intermediate resistance found for RT2 is 20%, the sample is called as intermediate resistant to the associated drug.
4. If the percentage intermediate resistance found for RT1 is 20% and the percentage intermediate resistance found for RT2 is 25%, the sample is called as intermediate resistant to the associated drug from amplicon RT2.
5. If the percentage of sensitive for RT1 is 90% and the percentage of sensitive for RT2 is 99%, the drug is called as susceptible from amplicon RT2.

The antiretroviral drug classes (NNRTI, NRTI, PI, and IN), the drugs, the predicted resistance levels of the sample to each drug, as well as the drug resistant mutations associated with each drug are shown in the drug resistance report.

### 4.2.3 Quality Analysis in Seq2Res

Seq2Res inherits all the features of the QTrim quality-trimming tool. All the trimming results are available in Seq2Res. The demultiplexed sequence reads from every sample in a single file is trimmed and generates analytical graphical plots of the untrimmed and trimmed data for direct comparison. The plots show the comparison of trimmed and untrimmed data on mean quality of sequence reads against number of sequence reads (**Figure 4.7**), sequence read length verses number of sequence reads (**Figure 4.8**) and the trend of quality scores across sequence reads (**Figure 4.9**).

**4.2.4 Graphical analysis of DRM prevalence**

From the RAMICS produced fastm files for every sample, the percentage mutations at known standard drug resistant codon positions (**Table 4.2**) are calculated. The mutations that confer viral resistance to drugs are then grouped by the drug class – PI, NRTI, NNRTI and IN. A bar plot is generated for each drug class, showing the prevalence of each DRM and a red horizontal line that cuts through the plot representing the user defined prevalence cutoff for quick observation of DRMs with prevalence below or above or on the red line (**Figure 4.10**).

### 4.2.5 Evaluating the sensitivity of Seq2Res.

### 4.2.5.1: Confirmation of the accuracy of the locally installed version of Sierra

The first step of evaluating the sensitivity of Seq2Res was to ensure that the locally installed version of the Stanford resistance testing algorithm (Sierra) was fully functioning and sensitive on sequences of lengths that are comparable to 454 sequence reads.

To achieve this, we acquired two datasets of sample data from the Stanford Database website (<http://hivdb.stanford.edu/DR/asi/releaseNotes/index.html#sampledata>). Both datasets consisted of sequences, generated using population-based Sanger-style sequencing, with an associated genotyping result. The first dataset consisted of 2055 sequences (Dataset A) while the second consisted of 5838 sequences (Dataset B)

In order to analyze these data using Seq2Res, the sequences in both the test data were fragmented into three ‘amplicons’ per sequence with some overlap between the adjacent subsequences. The fragments corresponded to HXB2 positions 55 to 159 (PR), 50 to 298 (RT1) and 290 to 399 (RT2).

The resulting amplicons were individually mapped to the HXB2 *pol* sequence using RAMICS and then submitted to the locally installed version of Sierra. The resulting resistance calls for each linked set of amplicons were then compared with the known genotypic call for each ‘parental’ sequence.

### 4.2.6 Test Data for simulation

While the above approach is appropriate for ascertaining if the drug resistance calls on consensus sequences with a known genotype are correct it does not fully evaluate a pipeline developed to undertake resistance genotyping on high throughput sequencing data. Thus, we undertook a comprehensive analysis of simulated HTS sequence data to fully test the performance of Seq2Res.

Sequences covering the entire *pol* gene from five individuals were selected from dataset B (above). The selection criteria required that each sequence had to have at least the K65R, K103N mutations in reverse transcriptase (in order to evaluate the capacity of Seq2Res to call the correct DRM at homopolymer regions) as well as at least one other reverse transcriptase mutations and one or more protease DRMs. These were saved as resistant sequences.

A ‘susceptible’ sequence was generated for each resistant sequence by reverting each of the DRMs back to wild type. Thus, the final dataset that was used to simulate the HTS data contained ten sequences in total – five resistant and five susceptible sequences. The resistance profile of each of these ‘seed’ sequences was evaluated using the Stanford HIV resistance algorithm.

### 4.2.7 Simulation of high throughput sequencing amplicons

Each of the ‘seed’ sequences was then fragmented into three subsequences representing an individual amplicon. These fragments corresponded to HXB2 positions 169 – 469 (PR), 466 – 812 (RT1) and 672 – 1017 (RT2)

In order to simulate the fragmented PR, RT1 and RT2, we chose a next generation sequence simulator called ART (Huang et al., 2012) as the features of tool include functions to produce sequencing-like errors. ART can emulate the Roche/454 sequencing process to simulate high throughput sequence data and include Roche/454 specific error models in homopolymer, substitution and insertion-deletion errors, as well as flexible base quality profiles. These are all customizable parameters and ART allows user-supplied sequence read profiles. A real Roche/454 sequence dataset was provided to ART to generate sequence read profiles. The profile was then used to simulate 20,000 Roche/454 sequence reads for each PR, RT1 and RT2 fragment (amplicon sequences) in FASTQ format.

### 4.2.7.1 Generation of different known prevalence of DRM data

For each amplicon from each patient the simulated sensitive and resistant sequences were randomly sampled to generate HTS datasets containing 10,000 sequence reads with a known mix of resistant and sensitive sequences. Seven datasets were generated for each amplicon each patient representing 0.1%, 1%, 5%, 10%, 15%, 20% and 50% prevalence of resistant sequences in the dataset (**Table 4.3**). The final simulated dataset comprised of 105 fastq files corresponding to 7 prevalence levels x 3 amplicon x 5 patients. Each of these fastq files was submitted to Seq2Res and the observed results compared with the expected results.

### 4.2.8 Computational Resources

Seq2Res is currently housed at the South African National Bioinformatics Institute (SANBI). The computing infrastructure used to run Seq2Res consists of two Blade Servers (PowerEdge M610x) each comprising 12 core processors (Intel(R) Xeon(R) CPU X5675 @ 3.07GHz), 32-gigabyte memory and a graphical processing unit (GPUs) (NVIDIA®Tesla™ M2090).

An application programming interface (API) has been developed to run Seq2Res over the internet. A web front end that uses the API has been developed for biologists and clinicians to upload the HIV genotypic data and other required files and parameters for the viral drug resistance testing of associating patients. The API can also be executed from command line executable scripts. The API was developed by Baruch Lubinsky.

### 4.3 Results

### 4.3.1 Seq2Res running time

The running time of Seq2Res depends on several factors, including the Internet connection at SANBI in order to download the uploaded sequence and parameter files, the number of raw sequence reads, amplicons and samples in the uploaded sequence file, and the number of jobs running on the available servers. Therefore, it is challenging to provide an exact running time for a Seq2Res process.

Nonetheless, we tested Seq2Res at the current Internet speed at SANBI (2 megabytes/second) with no other processes running on the servers, using 119 simulated files. Each file contained one sample, one amplicon and 10,000 sequence reads and we allowed a two-minute time interval between each submission using the API with command line scripting. After the completion of each job, and email is sent to the user. The start time was noted as the time of job submission and the time of job completion was noted as the time recorded in the email of each process. We observed that the average time for each job execution was three minutes. However, the execution time increased with an increase in samples and amplicons in the input raw sequence file. We also executed Seq2Res locally in the SANBI computational infrastructure on Roche/454 Junior and FLX plates, bypassing the internet connection. The raw sequence files from the Junior plates had 48 samples, with each sample containing three amplicons. The FLX plate consists of eight sections and a raw sequence file from each section had eight samples, with each of the samples containing three amplicons. The execution time for a file from Roche/454 Junior plate required on average of 30 minutes and a file from a section of Roche/454 FLX plate required on average of 15 minutes.

### 4.3.2. Quality Trim analysis of simulated data

The quality trimming report for the 105 simulated datasets showed that there was no sequence reads discarded in any dataset as a result of poor quality. Because the same simulator ART using the same quality profile produced all the simulated datasets, we analyzed the quality of one of the 105 simulated datasets. The mean quality of the sequence reads in the dataset did not seem to change before trimming and after trimming. The read mean quality score was observed to be between 30 and 32 in both untrimmed and trimmed state (**Figure 4.11**). The median quality score at every 10th base position from all sequence reads was observed to be above 30 across the sequence reads (**Figure 4.12**). Similar quality scores in sequence reads were observed in the other 104 simulated datasets.

### 4.3.3 Comparison of Mutation and Resistance Level Calls using the Sierra web service and Seq2Res

For every sequence in dataset A and dataset B, the DRMs reported using the Sierra web service (Liu and Shafer, 2006) was compared to those reported for the same sequence analysed by the Seq2Res. The one to one comparison of DRMs in every sequence reported by Sierra web service and RAMICS as part of Seq2Res, showed that except for a single drug resistant mutation in reverse transcriptase codon position 236 in a sequence from dataset A, there was a 100% congruency in DRMs calls. In the sequence from dataset A, Sierra web service reported amino acid Leucine (236L) while RAMICS reported wild type amino acid Proline (236P) at the discordant codon position.

### 4.3.4 Drug Resistant Mutations in the selected test sequences for simulation

The ten ‘seed’ sequences used for simulation were analyzed using web sierra. The five ‘resistant’ sequences were confirmed to have multiple DRMs, many of which convey resistance to various drugs DRMs (**Table 4.4**). The five ‘sensitive’ sequences, on the other hand, were confirmed to contain no DRMs (**Table 4.4**).

The Sierra web service reports DRMs resistance levels in five categories including high-level resistance, intermediate resistance, low-level resistance, potential low-level resistance and susceptible. The five categories are converted to a three class resistance level, which inlcludes Resistance (R), Intermediate (I) and Susceptible (S) classifications, according to the Stanford HIVDB algorithm (**Table 4.5**). The five resistant sequences with DRMs were observed to be highly resistant to most of the antiretroviral drugs while the five sequences without any DRMs were observed to be sensitive to the antiretroviral drugs (**Table 4.6**).

### 4.3.5 Optimal codon positions of the amplicons in the simulated datasets

The nucleotide start and end positions that were defined corresponding to the HXB2 *pol* reference sequence, for the amplicon PR were 169 and 469, for RT1 were 466 and 812 and for RT2 were 672 and 1017. The difference in the nucleotide start and end positions provides the full-length of the amplicons. The codon positions of the nucleotide start and end positions corresponding to the reference sequence for PR were 57 and 157, for RT1 were 156 and 271 and for RT2 were 224 and 339. Seq2Res processed these start and end codon positions of the full-length amplicons to get the first and last DRM codon positions in the amplicons. These first and last DRM codon positions in the amplicons are the optimal full-length codon positions and the sequence in between the positions covers all DRMs of interest. The optimal full-length start and end codon positions obtained for PR was 66 and 149, for RT1 was 195 and 270 and for RT2 was 224 and 336 (**Table 4.7**). In further downstream processing, the amplicon sequence reads that extend from optimal start to end codon positions are considered although they are not necessarily the full-length amplicon.

### 4.3.6 Prevalence of known drug resistant mutations

Each simulated dataset had a defined proportion of resistant and sensitive sequences and therefore the prevalence of DRMs in the dataset is known prior to the analyses. The analyses of simulated data using Seq2Res showed that, in all cases, the observed prevalences of the DRMs were very almost identical to that of the expected prevalence (**Figure 4.13**).

We also investigated the observed prevalence of DRMs in simulated amplicon datasets from all samples analyzed by Seq2Res. The mean observed prevalence of DRMs from three simulated amplicon datasets, PR, RT1 and RT2, in a sample was calculated. The mean observed prevalence from the three amplicons at all expected prevalence levels (0.1%, 1%, 5%, 10%, 15%, 20% and 50%) in the samples was marginally greater or lower than the expected prevalence (**Table 4.8**). The mean of the observed prevalence across the samples also showed that the observed prevalence was again marginally greater or lower than the expected prevalence (**Table 4.8**).

### 4.3.7 Resistance calls to antiretroviral drugs

All the simulated datasets (105 datasets), each containing one amplicon, were executed in Seq2Res at the default prevalence cutoff of 15%. The simulated datasets had sequence reads with a known prevalence of highly resistant DRMs. The prevalence of sequence reads that were predicted as drug resistant/intermediate to an associated antiretroviral drug using local Sierra algorithms were calculated using an inbuilt algorithm. Based on the calculated prevalence of resistant sequence reads, the algorithm also called an antiretroviral drug either resistant or intermediate or sensitive to the sample.

As expected, we observed that at the default prevalence cutoff of 15%, all the antiretroviral drugs were called as sensitive in the amplicons PR, RT1 and RT2 simulated datasets of all samples with expected DRMs prevalence below 15% (0.1%, 1%, 5% and 10%) (**Table 4.9**),

In the PR amplicon simulated datasets with 15% expected prevalence of DRMs from all samples, all PI drugs were called as sensitive in the samples 56252, 63377, and 4134, whereas the samples 21354 and 2368 have intermediate drug resistance call (**Table 4.10**). In the amplicons RT1 and RT2 simulated datasets with 15% expected prevalence of DRMs from all samples, almost all NRTI drugs were called as resistant or intermediate resistant whereas NNRTI drugs EFV and NVP were called as resistant and ETR and RPV were called as sensitive (**Table 4.10**).

Again, as expected, in the amplicons PR, RT1 and RT2 simulated datasets with expected prevalence of 20% and 50%, most of the drugs were called as resistant or intermediate resistance (**Table 4.11**).

### 4.3.8 Seq2Res web Application Programming Interface (API) and web outputs

The Seq2Res web interface (available at <http://hiv.sanbi.ac.za/tools/#/seq2res>) (**Figure 4.14**) has been created to enable easy execution of the pipeline for HIV drug resistance genotyping by people with little or no bioinformatics experience.

A click on “Submit job” takes users to the Seq2Res job submission page (**Figure 4.15**). Users can provide a job name and upload a raw sequence file, primer file containing the forward and reverse primer details and MID file containing sample specific tags sequences. While the required parameters are kept to the bare minimum on the initial website to avoid confusion, users can also set a number of other parameters for the analysis in the advanced options. Users are informed by email about the completion of their job.

Clicking the “My Jobs” button at Seq2Res homepage takes users to a page containing the list of all the jobs that the user has submitted (**Figure 4.16**). Job details like the name of job, the date of job submission and the status of the job processing - “complete” or “pending” or “error” while processing.

A click on a job from the list of jobs initially displays two plots that show the overall sequence analysis results of all samples in the input file for that job. The first plot in the result page shows the number of sequences discarded in each step of Seq2Res processing and the number of sequences that are passed in downstream analysis for making final result (**Figure 4.17**). The second plot shows the number of sequences in forward and reverse strands per amplicon per sample on which HIV drug resistance genotyping was performed (**Figure 4.18**).

A list of individual patient results in the form of links are listed down the results page, below the above mentioned two plots. A click on a green “show” button at the left of each patient and MID name takes users to a drug resistance result page of the selected patient. A table at the top of the sample specific result page shows the drug resistant report of the selected patient (**Figure 4.19**). The susceptibility of drugs for the patient is shown by color code – Red: Resistant, Orange: Intermediate resistant, Green: Susceptible (**Figure 4.19**).

Below the table in a patient specific result page, there are four bar plots – one plot for each drug class – NRTI, NNRTI, PI and Integrase Inhibitors (INs). Each bar plot shows the prevalence of drug class specific DRMs, the DRMs and the coverage (number of reads) at each DRM codon position (**Figure 4.20**).

### 4.4 Discussion and Conclusion

HIV drug resistance testing is essential to characterize the viral population (Baba et al., 2005; Simen et al., 2009) and to treat HIV infected individuals with the correct combination of antiretroviral drugs to suppress the viral replication for longer time periods and thereby increase life expectancy (Harrison et al., 2010; van Sighem et al., 2010). The Roche/454 HTS technology has shown great potential to genotype even the minor HIV variants that are clinically relevant (Lataillade et al., 2010; Simen et al., 2007; Simen et al., 2009; Varghese et al., 2009). But HTS technology produces up to a million sequences ([www.454.com](http://www.454.com)) and manual analysis at any processing step is not feasible. Our Seq2Res computational pipeline is designed to analyze this huge amount of data for HIV drug resistance testing and produce sample specific drug resistance reports..

HIV drug resistance is strongly associated with the pattern of DRMs (Bennett et al., 2009; Johnson et al., 2009; Johnson et al., 2010; Zhang et al., 2010). Therefore, it is required to sequence the entirety of HIV genes associated with drug resistancs, to reveal the DRMs patterns. Among the current high throughput sequencing technologies, Roche/454 has the ability to generate sequence reads of up to 1000 nucleotides length ([www.454.com](http://www.454.com)). However, the reverse transcriptase gene covers over a thousand bases and therefore needs to be sequenced in two parts, where each amplicon’s “start” and “end” positions are marked by the positions of first base position of the forward primer and last base position of reverse primer. The full-length amplicon sequences that are defined by user provided nucleotide “start” and “end” positions were analyzed to consider all the DRMs that the amplicon covered. However, a large number of sequences were discarded even though these sequences covered all the DRMs of the amplicon. The DRM codons were usually located at a position away from the 5’ and 3’ ends in the amplicon sequence reads. The sequences were discarded as non full-length because they did not cover those non-DRM codon positions at the ends. Therefore, Seq2Res finds the first and last DRM codon positions in the amplicon and considers the sequences covering those positions as optimal full-length. This saved large number of sequences from being discarded as non full-length.

We tested the sensitivity of the RAMICS mapping tool and the accuracy of the locally installed version Sierra, using two datasets A and B. The sequence reads in each dataset were fragmented into 3 subsequences with some overlaps, generating three amplicon datasets from each dataset. The sequence reads in each dataset were mapped to the reference sequence with RAMICS and the mutation calls at DRM codon positions produced by RAMICS, were submitted to local Sierra.

The comparison of mutation calls and resistance calls obtained from RAMICS and local Sierra to that from the Sierra web service showed that there was only one mutation call that was in different, the ultimate resistance calls for that sequence read were the same.

In order to ascertain why the mutation call did not correspond, we manually reviewed the RAMICS alignment. RAMICS showed that there was a deletion of nucleotide ‘T’ in the codon position 234, which result to unknown amino acid at that position (**Figure 4.21 1A**). The alignment showed that at codon positions 235 and onward, there were wild type amino acids in the sequence read (**Figure 4.21 1B**).

The alignment of the sequence with the discordant mutation call could not be obtained from the Sierra web service as we did not know which reference mapping tool was used. However, based on the amino acids that Sierra web sierra reported at codon position 234, 235 and 236, we were able to identify the Sierra web service alignment. It indicated that the amino acid at codon positions 234 was proline (P) while the wild type was leucine (L), at codon position 235 was isoleucine while wild type was histamine (H) and at codon position 236 was leucine while wild type was proline (P). Based on this codon positions information, we edited the RAMICS alignment and removed the deletion manually at codon position 234 from RAMICS alignment. The removal of the deletion showed exactly the same amino acids at codon positions 234, 235 and 236 as reported by Sierra web service. Therefore, we concluded that the Sierra web service alignment of the sequence read had no deletion at codon position 234 (**Figure 4.21 2C**). This resulted in the mutation P236L in the sequence reads (**Figure 4.21 2D.** This showed that Seq2Res mapping tool, RAMICS, had better alignment and called the mutations accurately as the mapping was done at ‘codon-space”. While most of the mapping tools including muscle (Edgar, 2004a, b), Clustalw (Larkin et al., 2007), MAFFT (Katoh et al., 2009; Katoh et al., 2002), Mosaik (Lee et al., 2013), T-coffee (Notredame et al., 2000), Kalign (Lassmann and Sonnhammer, 2005) maps sequences to a reference sequence at nucleotide space, RAMICS maps at “codon-space” and thus is able to identify PCR sequencing error and genuine indels.

Since the mutation calls and resistance calls by RAMICS and local Sierra were accurate, we evaluated the complete Seq2Res pipeline using the simulated datasets. The simulated datasets were generated using the simulator tool ART (Huang et al., 2012). The simulator was provided with real high throughput sequence profile. The number of research publications (Gilles et al., 2011; Huse et al., 2007; Kunin et al., 2009; Margulies et al., 2005; Moore et al., 2006; Vera et al., 2008; Wang et al., 2007) showed that Roche/454 high throughput pyrosequencing has high insertion and deletion error rate at the homopolymer region. As we are using the simulated datasets to evaluate that Seq2Res works, we undertook the homopolymer region analysis to know the similarity level of the simulated datasets to the real high throughput sequence data.

Out of the DRMs present in the “seed” sequence, we analyzed the homopolymer errors only in the DRMs G48V, I54V, K65R and K103N present within or adjacent to the homopolymer region in HIV reference (**Table 4.12**) in the simulated datasets. The homopolymer errors in the simulated sequences derived from sensitive “seed” sequence and resistant “seed” sequence were analyzed independently.

In the simulated datasets from sample 2368, we observed up to four nucleotide insertions and three nucleotide deletions (**Table 4.13**). The highly occurring error was one nucleotide insertion (+1) or one nucleotide deletion (-1). We also observed that the highest percentage of simulated sequence reads in 2368’s sample dataset without any indels was 95.96%, which would mean that the remaining 4.04% of the sequence reads have homopolymer errors (**Table 4.13**). On the other hand, a dataset of the sample 2368 had 70.07% of sequence reads without any homopolymer indels and the remaining 29.93% had no homopolymer indels. Therefore, there was inconsistent in the homopolymer errors between the simulated datasets from the same simulator tool. We observed similar amount of indels in other sample’s simulated datasets as well.

We compared the simulated homopolymer errors with the published reports on homopolymer error rates in real datasets. Gilles et al (Gilles et al., 2011) reported that they observed **5.97% ±1.33** homopolymer related error of the total 45,895 erroneous bases in the 454 GS-FLX Titanium bases. Huse et.al (Huse et al., 2007) reported that **39%** of the errors were related to homopolymer out of 159,981 total errors. Alike the inconsistency in the homopolymer errors in between the simulated datasets, the published papers also showed that there was inconsistency about the homopolymer errors in the real high throughput pyrosequencing datasets.

The simulated sequence reads derived from resistant “seed” sequence and sensitive “seed” sequence were commixed together at fixed proportions to generate datasets with varying known prevalence (0.1%, 1%, 5%, 10%, 15%, 20% and 50%) of simulated sequence reads containing DRMs. The “seed” sequences, from which the simulated datasets were prepared, were tested to Sierra web service (Liu and Shafer, 2006) to obtain mutation calls and resistance calls.

The simulated datasets were analyzed using Seq2Res for the mutation calls and the prevalence of the DRMs. Firstly, we tested Seq2Res mutation calls in simulated datasets and found that it was similar to the mutation calls obtained from Sierra web service for the “seed” sequences. Secondly, we tested the resistance calls to antiretroviral drugs based on the percentage of resistant sequences predicted by local Sierra at **default prevalence cutoff 15%.**

In the simulated amplicon datasets with known prevalence of resistant sequences less than 15% (0.1%, 1%, 5% and 10%), the antiretroviral drugs associated with the mutations in the sequence reads were called as sensitive. If the prevalence cutoff was decreased to below 0.1%, it would call the antiretroviral drugs associated with the mutations in the sequence reads as resistant in all the simulated datasets.

In the simulated amplicon datasets with 15% prevalence of resistant sequences, we expected the DRMs associated antiretroviral drugs were called as resistant or intermediate resistant. However, most PI drugs were called as sensitive and few as intermediate resistant. This was because the observed prevalence of DRMs in the amplicons was marginally lower than the expected prevalence. The reason for lower observed prevalence was further investigated. Each simulated amplicon dataset with expected DRMs prevalence of 15% initially had 1500 sequence reads with DRMs out of 10000 sequence reads before analysis. We observed that numbers of sequence reads with DRMs were discarded as non optimal full-length (**Table 4.14**) and this lowered the observed prevalence marginally than the expected prevalence.

In the simulated amplicon datasets with DRMs prevalence above 15% (20% and 50%), all the antiretroviral drugs associated to the DRMs in the sequence reads, as expected, were called as resistant/intermediate resistant.

We also investigated the observed prevalence of DRMs obtained for each amplicon in a sample. The mean observed prevalence of DRMs across the amplicons from a sample at similar expected prevalence showed that the observed prevalence was marginally greater or lower than the expected prevalence. The mean across samples for the same expected prevalence also showed that the observed prevalence was marginally greater or lower than the expected prevalence.

Furthermore, we investigated the reason for the observed prevalence to be marginally greater or lower but not equal to the expected prevalence. We found that there could be two reasons that deviated the observed prevalence from the expected prevalence.

The first reason, as mention above, was that the number of simulated sequence reads both with and without DRMs was discarded as non-optimal full length (**Table 4.15**). The proportion of the discarded simulated sequences with and without DRMs was not same as the proportion that they were in the dataset before analysis. For example, if the dataset before analysis had 1000 sequences with DRMs and 9000 sequences without DRMs, the proportion was 1:10. If 100 sequences with DRMs and 50 sequences without DRMs were discarded as non-optimal full length, their proportion of discarded sequences was 2:1. If the number of sequences with DRMs discarded were higher than those without DRMs in the discarded proportion, the observed prevalence was lower than the expected prevalence. Similarly, if the number of sequences without DRMs discarded were higher than those with DRMs in the discarded proportion, the observed prevalence was higher than the expected prevalence.

The second reason was the increase in the number of sequences with DRMs due to simulation errors at DRM codon positions that generated the DRMs. The type of error that contributed to the generation of DRMs would be substitution as Seq2Res mapping tool - RAMICS would have correction the insertion and deletion errors. This increase in the sequences with DRMs due to simulation errors increased the observed prevalence and was higher than the expected prevalence. This large effect of the simulation errors at DRM codon positions was seen only at low expected prevalence level 0.1% where the observed prevalence was higher.

Thus, we evaluated Seq2Res pipeline for resistance genotyping on high throughput simulated sequence data and our observation showed that the pipeline was able to call correct DRMs and resistance call on DRMs associated antiretroviral drugs.

Bibliography

Aberg, JA, Kaplan, JE, Libman, H, Emmanuel, P, Anderson, JR, Stone, VE, Oleske, JM, Currier, JS, Gallant, JE (2009) Primary care guidelines for the management of persons infected with human immunodeficiency virus: 2009 update by the HIV medicine Association of the Infectious Diseases Society of America. *Clin Infect Dis* **49**: 651-681.

Adetunji, AA, Achenbach, C, Feinglass, J, Darin, KM, Scarsi, KK, Ekong, E, Taiwo, BO, Adewole, IF, Murphy, R (2012) Optimizing Treatment Switch for Virologic Failure during First-Line Antiretroviral Therapy in Resource-Limited Settings. *J Int Assoc Provid AIDS Care* **12**: 236-240.

Aghokeng, AF, Kouanfack, C, Laurent, C, Ebong, E, Atem-Tambe, A, Butel, C, Montavon, C, Mpoudi-Ngole, E, Delaporte, E, Peeters, M (2011) Scale-up of antiretroviral treatment in sub-Saharan Africa is accompanied by increasing HIV-1 drug resistance mutations in drug-naive patients. *AIDS* **25**: 2183-2188.

Asamoah-Odei, E, Garcia Calleja, JM, Boerma, JT (2004) HIV prevalence and trends in sub-Saharan Africa: no decline and large subregional differences. *Lancet* **364**: 35-40.

Autran, B, Carcelain, G, Li, TS, Blanc, C, Mathez, D, Tubiana, R, Katlama, C, Debre, P, Leibowitch, J (1997) Positive effects of combined antiretroviral therapy on CD4+ T cell homeostasis and function in advanced HIV disease. *Science* **277**: 112-116.

Baba, M, Takashima, K, Miyake, H, Kanzaki, N, Teshima, K, Wang, X, Shiraishi, M, Iizawa, Y (2005) TAK-652 Inhibits CCR5-Mediated Human Immunodeficiency Virus Type 1 Infection In Vitro and Has Favorable Pharmacokinetics in Humans. *Antimicrobial Agents and Chemotherapy* **49**: 4584-4591.

Bangsberg, DR, Charlebois, ED, Grant, RM, Holodniy, M, Deeks, SG, Perry, S, Conroy, KN, Clark, R, Guzman, D, Zolopa, A (2003) High levels of adherence do not prevent accumulation of HIV drug resistance mutations. *Aids* **17**: 1925.

Bebenek, K, Abbotts, J, Roberts, JD, Wilson, SH, Kunkel, TA (1989) Specificity and mechanism of error-prone replication by human immunodeficiency virus-1 reverse transcriptase. *J Biol Chem* **264**: 16948-16956.

Bennett, DE, Camacho, RJ, Otelea, D, Kuritzkes, DR, Fleury, H, Kiuchi, M, Heneine, W, Kantor, R, Jordan, MR, Schapiro, JM, Vandamme, AM, Sandstrom, P, Boucher, CA, van de Vijver, D, Rhee, SY, Liu, TF, Pillay, D, Shafer, RW (2009) Drug resistance mutations for surveillance of transmitted HIV-1 drug-resistance: 2009 update. *PLoS One* **4**: e4724.

Berkhout, B, Das, AT, Beerens, N (2001) HIV-1 RNA editing, hypermutation, and error-prone reverse transcription. *Science* **292**: 7.

Blower, S, Bodine, E, Kahn, J, McFarland, W (2005) The antiretroviral rollout and drug-resistant HIV in Africa: insights from empirical data and theoretical models. *AIDS* **19**: 1-14.

Clumeck, N, Pozniak, A, Raffi, F (2008) European AIDS Clinical Society (EACS) guidelines for the clinical management and treatment of HIV-infected adults. *HIV Med* **9**: 65-71.

Cock, PJ, Antao, T, Chang, JT, Chapman, BA, Cox, CJ, Dalke, A, Friedberg, I, Hamelryck, T, Kauff, F, Wilczynski, B (2009) Biopython:freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**: 1422-1423.

Cock, PJA, Fields, CJ, Goto, N, Heuer, ML, Rice, PM (2010) The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research* **38**: 1767-1771.

Dudley, DM, Chin, EN, Bimber, BN, Sanabani, SS, Tarosso, LF, Costa, PR, Sauer, MM, Kallas, EG, O'Connor, DH (2012) Low-cost ultra-wide genotyping using Roche/454 pyrosequencing for surveillance of HIV drug resistance. *PLoS One* **7**: e36494.

Edgar, RC (2004a) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**: 113.

Edgar, RC (2004b) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**: 1792-1797.

Fang, CT, Chang, YY, Hsu, HM, Twu, SJ, Chen, KT, Lin, CC, Huang, LY, Chen, MY, Hwang, JS, Wang, JD, Chuang, CY (2007) Life expectancy of patients with newly-diagnosed HIV infection in the era of highly active antiretroviral therapy. *QJM* **100**: 97-105.

Gazzard, B, Clumeck, N, d'Arminio Monforte, A, Lundgren, JD (2008) Indicator disease-guided testing for HIV--the next step for Europe? *HIV Med* **9 Suppl 2**: 34-40.

Gilles, A, Meglecz, E, Pech, N, Ferreira, S, Malausa, T, Martin, JF (2011) Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing. *BMC Genomics* **12**: 245.

Golin, CE, Liu, H, Hays, RD, Miller, LG, Beck, CK, Ickovics, J, Kaplan, AH, Wenger, NS (2002) A prospective study of predictors of adherence to combination antiretroviral medication. *J Gen Intern Med* **17**: 756-765.

Gupta, RK, Jordan, MR, Sultan, BJ, Hill, A, Davis, DH, Gregson, J, Sawyer, AW, Hamers, RL, Ndembi, N, Pillay, D, Bertagnolio, S (2012) Global trends in antiretroviral resistance in treatment-naive individuals with HIV after rollout of antiretroviral treatment in resource-limited settings: a global collaborative study and meta-regression analysis. *Lancet* **380**: 1250-1258.

Hamers, RL, Schuurman, R, Sigaloff, KC, Wallis, CL, Kityo, C, Siwale, M, Mandaliya, K, Ive, P, Botes, ME, Wellington, M, Osibogun, A, Wit, FW, van Vugt, M, Stevens, WS, de Wit, TF (2012) Effect of pretreatment HIV-1 drug resistance on immunological, virological, and drug-resistance outcomes of first-line antiretroviral treatment in sub-Saharan Africa: a multicentre cohort study. *Lancet Infect Dis* **12**: 307-317.

Hamers, RL, Wallis, CL, Kityo, C, Siwale, M, Mandaliya, K, Conradie, F, Botes, ME, Wellington, M, Osibogun, A, Sigaloff, KC, Nankya, I, Schuurman, R, Wit, FW, Stevens, WS, van Vugt, M, de Wit, TF (2011a) HIV-1 drug resistance in antiretroviral-naive individuals in sub-Saharan Africa after rollout of antiretroviral therapy: a multicentre observational study. *Lancet Infect Dis* **11**: 750-759.

Hamers, RL, Wensing, AM, Back, NK, Arcilla, MS, Frissen, JP (2011b) Multi-nucleoside reverse transcriptase inhibitor resistant HIV type-1 in a patient from Sierra Leone failing stavudine, lamivudine and nevirapine. *Antivir Ther* **16**: 115-118.

Harrison, KM, Song, R, Zhang, X (2010) Life expectancy after HIV diagnosis based on national HIV surveillance data from 25 states, United States. *J Acquir Immune Defic Syndr* **53**: 124-130.

Herbst, AJ, Cooke, GS, Barnighausen, T, KanyKany, A, Tanser, F, Newell, ML (2009) Adult mortality and antiretroviral treatment roll-out in rural KwaZulu-Natal, South Africa. *Bull World Health Organ* **87**: 754-762.

Ho, DD, Neumann, AU, Perelson, AS, Chen, W, Leonard, JM, Markowitz, M (1995) Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection. *Nature* **373**: 123-126.

Huang, W, Li, L, Myers, JR, Marth, GT (2012) ART: a next-generation sequencing read simulator. *Bioinformatics* **28**: 593-594.

Huse, SM, Huber, JA, Morrison, HG, Sogin, ML, Welch, DM (2007) Accuracy and quality of massively parallel DNA pyrosequencing. *Genome biol* **8**: R143.

Jakobsen, MR, Tolstrup, M, Sogaard, OS, Jorgensen, LB, Gorry, PR, Laursen, A, Ostergaard, L (2010) Transmission of HIV-1 drug-resistant variants: prevalence and effect on treatment outcome. *Clin Infect Dis* **50**: 566-573.

Ji, H, Liang, B, Li, Y, Van Domselaar, G, Graham, M, Tyler, S, Merks, H, Sandstrom, P, Brooks, J (2012) Low abundance drug resistance variants in transmitted HIV drug resistance surveillance specimens identified using tagged pooled pyrosequencing. *J Virol Methods* **187**: 314-320.

Johnson, JA, Li, J-F, Wei, X, Lipscomb, J, Irlbeck, D, Craig, C, Smith, A, Bennett, DE, Monsour, M, Sandstrom, P, Lanier, ER, Heneine, W (2008) Minority HIV-1 Drug Resistance Mutations Are Present in Antiretroviral Treatment–Naïve Populations and Associate with Reduced Treatment Efficacy. *PLoS Med* **5**: e158.

Johnson, VA, Brun-Vézinet, Ft, Clotet, B, Gunthard, HF, Kuritzkes, DR, Pillay, D, Schapiro, JM, Richman, DD (2009) Update of the drug resistance mutations in HIV-1: December 2009. *Top HIV Med* **17**: 138–145.

Johnson, VA, Calvez, V, Günthard, HF, Paredes, R, Pillay, D, Shafer, R, Wensing, AM, Richman, DD (2010) 2011 Update of the Drug Resistance Mutations in HIV-1. *HIV Med* **18**: 156–163.

Kaplan, JE, Benson, C, Holmes, KH, Brooks, JT, Pau, A, Masur, H (2009) Guidelines for prevention and treatment of opportunistic infections in HIV-infected adults and adolescents: recommendations from CDC, the National Institutes of Health, and the HIV Medicine Association of the Infectious Diseases Society of America. *MMWR Recomm Rep* **58**: 1-207; quiz CE201-204.

Katoh, K, Asimenos, G, Toh, H (2009) Multiple alignment of DNA sequences with MAFFT. *Methods Mol Biol* **537**: 39-64.

Katoh, K, Misawa, K, Kuma, K, Miyata, T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* **30**: 3059-3066.

Kunin, V, Engelbrektson, A, Ochman, H, Hugenholtz, P (2009) Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environ Microbiol* **12**: 118-123.

Larkin, MA, Blackshields, G, Brown, NP, Chenna, R, McGettigan, PA, McWilliam, H, Valentin, F, Wallace, IM, Wilm, A, Lopez, R, Thompson, JD, Gibson, TJ, Higgins, DG (2007) Clustal W and Clustal X version 2.0. *Bioinformatics* **23**: 2947-2948.

Lassmann, T, Sonnhammer, EL (2005) Kalign--an accurate and fast multiple sequence alignment algorithm. *BMC Bioinformatics* **6**: 298.

Lataillade, M, Chiarella, J, Yang, R, Schnittman, S, Wirtz, V, Uy, J, Seekins, D, Krystal, M, Mancini, M, McGrath, D, Simen, B, Egholm, M, Kozal, M (2010) Prevalence and clinical significance of HIV drug resistance mutations by ultra-deep sequencing in antiretroviral-naive subjects in the CASTLE study. *PLoS One* **5**: e10952.

Lee, W-P, Stromberg, M, Ward, A, Stewart, C, Garrison, E, Marth, GT (2013) MOSAIK: A hash-based algorithm for accurate next-generation sequencing read mapping. *arXiv preprint arXiv:13091149*.

Li, JZ, Paredes, R, Ribaudo, HJ, Svarovskaia, ES, Metzner, KJ, Kozal, MJ, Hullsiek, KH, Balduin, M, Jakobsen, MR, Geretti, AM, Thiebaut, R, Ostergaard, L, Masquelier, B, Johnson, JA, Miller, MD, Kuritzkes, DR (2011) Low-frequency HIV-1 drug resistance mutations and risk of NNRTI-based antiretroviral treatment failure: a systematic review and pooled analysis. *JAMA* **305**: 1327-1335.

Li, TS, Tubiana, R, Katlama, C, Calvez, V, Ait Mohand, H, Autran, B (1998) Long-lasting recovery in CD4 T-cell function and viral-load reduction after highly active antiretroviral therapy in advanced HIV-1 disease. *Lancet* **351**: 1682-1686.

Little, SJ, Holte, S, Routy, JP, Daar, ES, Markowitz, M, Collier, AC, Koup, RA, Mellors, JW, Connick, E, Conway, B, Kilby, M, Wang, L, Whitcomb, JM, Hellmann, NS, Richman, DD (2002) Antiretroviral-drug resistance among patients recently infected with HIV. *N Engl J Med* **347**: 385-394.

Liu, TF, Shafer, RW (2006) Web resources for HIV type 1 genotypic-resistance test interpretation. *Clin Infect Dis* **42**: 1608-1618.

Low-Beer, S, Yip, B, O'Shaughnessy, MV, Hogg, RS, Montaner, JS (2000) Adherence to triple therapy and viral load response. *J Acquir Immune Defic Syndr* **23**: 360-361.

Mansky, LM (1996a) Forward mutation rate of human immunodeficiency virus type 1 in a T lymphoid cell line. *AIDS Res Hum Retroviruses* **12**: 307-314.

Mansky, LM (1996b) The mutation rate of human immunodeficiency virus type 1 is influenced by the vpr gene. *Virology* **222**: 391-400.

Mansky, LM (1998) Retrovirus mutation rates and their role in genetic variation. *J Gen Virol* **79 ( Pt 6)**: 1337-1345.

Mansky, LM, Temin, HM (1995) Lower in vivo mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase. *Journal of Virology* **69**: 5087-5094.

Margulies, M, Egholm, M, Altman, WE, Attiya, S, Bader, JS, Bemben, LA, Berka, J, Braverman, MS, Chen, Y-J, Chen, Z, Dewell, SB, Du, L, Fierro, JM, Gomes, XV, Godwin, BC, He, W, Helgesen, S, Ho, CH, Irzyk, GP, Jando, SC, Alenquer, MLI, Jarvie, TP, Jirage, KB, Kim, J-B, Knight, JR, Lanza, JR, Leamon, JH, Lefkowitz, SM, Lei, M, Li, J, Lohman, KL, Lu, H, Makhijani, VB, McDade, KE, McKenna, MP, Myers, EW, Nickerson, E, Nobile, JR, Plant, R, Puc, BP, Ronan, MT, Roth, GT, Sarkis, GJ, Simons, JF, Simpson, JW, Srinivasan, M, Tartaro, KR, Tomasz, A, Vogt, KA, Volkmer, GA, Wang, SH, Wang, Y, Weiner, MP, Yu, P, Begley, RF, Rothberg, JM (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**: 376-380.

Masur, H, Kaplan, JE (2009) New guidelines for the management of HIV-related opportunistic infections. *JAMA* **301**: 2378-2380.

Mills, EJ, Bakanda, C, Birungi, J, Chan, K, Ford, N, Cooper, CL, Nachega, JB, Dybul, M, Hogg, RS (2011) Life expectancy of persons receiving combination antiretroviral therapy in low-income countries: a cohort analysis from Uganda. *Ann Intern Med* **155**: 209-216.

Mocroft, A, Phillips, AN, Ledergerber, B, Smith, C, Bogner, JR, Lacombe, K, Wiercinska-Drapalo, A, Reiss, P, Kirk, O, Lundgren, JD (2010) Estimated average annual rate of change of CD4(+) T-cell counts in patients on combination antiretroviral therapy. *Antivir Ther* **15**: 563-570.

Moore, MJ, Dhingra, A, Soltis, PS, Shaw, R, Farmerie, WG, Folta, KM, Soltis, DE (2006) Rapid and accurate pyrosequencing of angiosperm plastid genomes. *BMC Plant Biol* **6**: 17.

Nattrass, N (2006) South Africa's "rollout" of highly active antiretroviral therapy: a critical assessment. *J Acquir Immune Defic Syndr* **43**: 618-623.

Notredame, C, Higgins, DG, Heringa, J (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol* **302**: 205-217.

Perfect, JR, Dismukes, WE, Dromer, F, Goldman, DL, Graybill, JR, Hamill, RJ, Harrison, TS, Larsen, RA, Lortholary, O, Nguyen, MH, Pappas, PG, Powderly, WG, Singh, N, Sobel, JD, Sorrell, TC (2010) Clinical practice guidelines for the management of cryptococcal disease: 2010 update by the infectious diseases society of america. *Clin Infect Dis* **50**: 291-322.

Preston, BD, Poiesz, BJ, Loeb, LA (1988) Fidelity of HIV-1 reverse transcriptase. *Science* **242**: 1168-1171.

Roberts, JD, Bebenek, K, Kunkel, TA (1988) The accuracy of reverse transcriptase from HIV-1. *Science* **242**: 1171-1173.

Simen, BB, Huppler Hullsiek, K, Novak, RM, MacArthur, RD, Baxter, JD, Huang, C, Lubeski, C, Turenchalk, GS, Braverman, MS, Desany, B (2007) Prevalence of low abundant drug-resistant variants by ultra-deep sequencing in chronically HIV-infected antiretroviral (ARV)-naıve patients and the impact on virological outcomes. In: 16th International HIV Drug Resistance Workshop Barbados.

Simen, BB, Simons, JF, Hullsiek, KH, Novak, RM, MacArthur, RD, Baxter, JD, Huang, C, Lubeski, C, Turenchalk, GS, Braverman, MS, Desany, B, Rothberg, JM, Egholm, M (2009) Low-Abundance Drug-Resistant Viral Variants in Chronically HIV-Infected, Antiretroviral Treatment–Naive Patients Significantly Impact Treatment Outcomes. *Journal of Infectious Diseases* **199**: 693-701.

Stringer, JS, Zulu, I, Levy, J, Stringer, EM, Mwango, A, Chi, BH, Mtonga, V, Reid, S, Cantrell, RA, Bulterys, M, Saag, MS, Marlink, RG, Mwinga, A, Ellerbrock, TV, Sinkala, M (2006) Rapid scale-up of antiretroviral therapy at primary care sites in Zambia: feasibility and early outcomes. *JAMA* **296**: 782-793.

Supervie, V, Garcia-Lerma, JG, Heneine, W, Blower, S (2010) HIV, transmitted drug resistance, and the paradox of preexposure prophylaxis. *Proc Natl Acad Sci U S A* **107**: 12381-12386.

Tsibris, AM, Korber, B, Arnaout, R, Russ, C, Lo, CC, Leitner, T, Gaschen, B, Theiler, J, Paredes, R, Su, Z, Hughes, MD, Gulick, RM, Greaves, W, Coakley, E, Flexner, C, Nusbaum, C, Kuritzkes, DR (2009) Quantitative deep sequencing reveals dynamic HIV-1 escape and large population shifts during CCR5 antagonist therapy in vivo. *PLoS One* **4**: e5683.

UNAIDS (2012) Global Report 2012: UNAIDS Report on the Global AIDS Epidemic. ebookpartnership. com.

van Sighem, AI, Gras, LA, Reiss, P, Brinkman, K, de Wolf, F (2010) Life expectancy of recently diagnosed asymptomatic HIV-infected patients approaches that of uninfected individuals. *AIDS* **24**: 1527-1535.

Varghese, V, Shahriar, R, Rhee, SY, Liu, T, Simen, BB, Egholm, M, Hanczaruk, B, Blake, LA, Gharizadeh, B, Babrzadeh, F, Bachmann, MH, Fessel, WJ, Shafer, RW (2009) Minority variants associated with transmitted and acquired HIV-1 nonnucleoside reverse transcriptase inhibitor resistance: implications for the use of second-generation nonnucleoside reverse transcriptase inhibitors. *J Acquir Immune Defic Syndr* **52**: 309-315.

Vera, JC, Wheat, CW, Fescemyer, HW, Frilander, MJ, Crawford, DL, Hanski, I, Marden, JH (2008) Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. *Mol Ecol* **17**: 1636-1647.

Wang, C, Mitsuya, Y, Gharizadeh, B, Ronaghi, M, Shafer, RW (2007) Characterization of mutation spectra with ultra-deep pyrosequencing: application to HIV-1 drug resistance. *Genome Res* **17**: 1195-1201.

WHO, U (2011) UNAIDS. 2011. Global HIV/AIDS response: Epidemic update and health sector progress towards universal access (Progress Report 2011). *World Health Organization, Geneva, Switzerland*.

Wittkop, L, Gunthard, HF, de Wolf, F, Dunn, D, Cozzi-Lepri, A, de Luca, A, Kucherer, C, Obel, N, von Wyl, V, Masquelier, B, Stephan, C, Torti, C, Antinori, A, Garcia, F, Judd, A, Porter, K, Thiebaut, R, Castro, H, van Sighem, AI, Colin, C, Kjaer, J, Lundgren, JD, Paredes, R, Pozniak, A, Clotet, B, Phillips, A, Pillay, D, Chene, G (2011) Effect of transmitted drug resistance on virological and immunological response to initial combination antiretroviral therapy for HIV (EuroCoord-CHAIN joint project): a European multicohort study. *Lancet Infect Dis* **11**: 363-371.

Yebra, G, de Mulder, M, Perez-Elias, MJ, Perez-Molina, JA, Galan, JC, Llenas-Garcia, J, Moreno, S, Holguin, A (2011) Increase of transmitted drug resistance among HIV-infected sub-Saharan Africans residing in Spain in contrast to the native population. *PLoS One* **6**: e26757.

Yerly, S, Kaiser, L, Race, E, Bru, J-P, Clavel, F, Perrin, L (1999) Transmission of antiretroviral-drug-resistant HIV-1 variants. *The Lancet* **354**: 729.

Zhang, J, Hou, T, Wang, W, Liu, JS (2010) Detecting and understanding combinatorial mutation patterns responsible for HIV drug resistance. *Proc Natl Acad Sci U S A* **107**: 1321-1326.