# 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; Hamers et al., 2011a). Antiretroviral therapy (ART) programmes have been established all over the 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 per replication cycle are produced with a mutation rate in the order of 10-5 per nucleotide per replication 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 scale up of antiretroviral (ARV) drugs there is growing evidences 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., 2009a; 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., 2009a). The most likely reasons for the increase of 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 variant (Jakobsen et al., 2010; Supervie et al., 2010; Wittkop et al., 2011; Yerly et al., 1999). The rising prevalence of drug resistant variants narrows treatment options (Hamers et al., 2012; Hamers et al., 2011a; Hamers et al., 2011b). Thus, ARV treatment, management and surveillance of drug resistant HIV variants (Bennett et al., 2009; Bennett et al., 2008) is essential for prolonging the usage of, and optimizing the outcome from, a particular drugs 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 pretreatment HIV drug resistance test, 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 out for exploring clinically relevant low abundance (< 1%) resistant HIV variants (Dudley et al., 2012; Ji et al., 2012; Tsibris et al., 2009; Wang et al., 2007). 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

### 4.2.1 HTS raw sequence reads

A set of forward and reverse primers that bind at targeted genome regions in the HIV genome and is PCR amplified to produce millions of copies of the region. Each copy of the region is then sequenced using HTS method. In an instance, when multiple samples need to be sequenced, the forward and reverse primers set is first attached to the 3’ end of a unique sample specific sequence, which is now described as Multiplex Identifier (MID) (**Figure 4.1**).

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 quickly and sensitively facilitate the 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 analyzing data from new sequencing platforms.

Thus, Seq2Res has been developed (mostly using Python) in a modularized format keeping each distinct analysis block independent of each other thereby enabling easy insertion of new modules to facilitate 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 analyse such data.

### 4.2.2 Development and processing of data using Seq2Res

One of the primary focuses in the development of Seq2Res was make usage as easy as possible by keeping the required number of input files and parameters to a bare minimum. Users need to input the raw sequence data file, the primers used in resequencing (multiple amplicons can be analysed concurrently) and if, present, the multiplex identifiers (MID) used together with their associated patient identifier.

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.

INSERT SECTION DESCRIBING DATA STRUCTURE HERE…needs to be written much better than what’s above…..describe it properly.

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

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**).

The end-users also have an option to set the following required preprocessing input parameters (**Figure 4.3 Step 1**):

1. Number of nucleotide mismatches between user supplied MID and MID in sequence read, which is described as MID tolerance
2. Number of nucleotide mismatches between user supplied primer and the primer in the sequence read, which is described as primer tolerance
3. Quality control parameters for raw data processing.

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

1. **File format pre-processing**

Raw sequence data can be uploaded in one of the three different sequence file formats to the pipeline: a Standard Flowgram Format (SFF) file generated by Roche/454 HTS platform, a FASTQ file, or a combination of FASTA and QUAL files. Depending on the input file, Seq2Res may need pre-processing conversion of the file (**Figure 4.3 step 2**). 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. **Preprocessing of MID and Primer files**

Seq2Res reads the tab delimited MID sequences and the sample names from the MID file and stores the list of paired sample name and its corresponding MID sequence. Seq2Res also reads the tab-delimited primer file containing the name of amplicons, primers (forward and reverse primers), the start and end nucleotide positions of amplicons (**Figure 4.3 step 3**). Seq2Res stores the amplicon names associating with its primers and start-end positions. From the start and end nucleotide positions, Seq2Res processes to find the start and end codon positions in the standard HIV *pol* reference sequence. On the basis of the start and end codon positions of each amplicon, Seq2Res again processes to find the first and the last DRM codon positions of the amplicons in the HIV *pol* reference sequence. The amplicon size from the first and the last DRM codon positions is called the optimal coverage amplicon. Seq2Res now defines a sequence that has the first DRM position and ends at the last DRM position or beyond as the optimal coverage full-length sequence.

1. **Demultiplexing**

Seq2Res utilizes Biopython ([www.biopython.org](http://www.biopython.org)) (Cock et al., 2009) packages to read individual sequences from the FASTQ file. For the first step of the demultiplexing Seq2Res searches for the MID in every sequence read which is located in the first X nucleotides at the 5’ end of a sequence read (x is the MID length). A subsequence of length X is extracted 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 details of the MID with the number of mismatches less or equal to the MID tolerance in the alignment (default MID tolerance 2) are 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, we search for the primer in each sequence read in a similar manner as was done for identifying the MID. Again, a subsequence of the length equal to a primer length is obtained from 5’ end of the sequence read. The subsequence is then pair-wise global 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 or primer tolerance (Default primer tolerance 3) is selected and tagged at the sequence ID along with the strand forward (+) or strand reverse (-).

If the sequence reads are generated using Primer ID technology, 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, 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. 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. Seq2Res allows a flexibility to control all the features of QTrim.

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

One sequence reads have been quality trimmed the information contained in their sequence Ids is used to bin them together 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.

RAMICS outputs a novel ‘fasta-like’ (fastm) file (**Figure 4.4**) 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 output 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 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 (Table 4.1) (**REF**) are extracted and saved to a file in the format required for submission to the Stanford algorithm (**Figure 4.5**).

1. **Resistance Testing**

For resistance testing we use 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 outputs the result in a XML file detailing their drug 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 extracts the resistance level for every DRMs in the sequences from the XML file, and tabulates it in a tab-delimited file. Each row in the tabulated file contains user and sequence information in the first ten columns. 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 to a specific drug (R, I or S) appears at odd columns and the pertinent DRMs to that resistance call follow in each of the even columns (**Table 4.2**). For example, a resistance call “R” to Abacavir (ABD) may appear at column 11 and DRMs K65R, K70R, V75I, F77L from a sequence that are resistant (R) to the first drug Abacavir (ABC) may appear in column 12. If DRMs that result in a resistance call to a particular drug are not present in the sequence, a dash appears in the columns for resistance level and DRMs for that drug. Thus, this step generates an easily searchable table from which all other results and conclusions are generated.

From the table, the prevalence of sequence reads that are highly resistant (R), intermediate resistant (I) or susceptible (S) to a particular antiretroviral drug are calculated. The prevalence of resistant and intermediate sequences for an antiretroviral drug are compared with the user defined cutoff prevalence and the antiretroviral drug is classified as resistant or intermediate resistance or susceptible to a drug using the following conditions in order:

1. If the prevalence of resistant sequences to a drug is greater than user defined cutoff, the drug is classified as Resistant.
2. If the prevalence of resistant sequences and intermediate resistant sequences to a drug is collectively greater or equal to user-defined cutoff or the prevalence of intermediate resistance to a drug is greater than user defined cutoff, the drug is classified as intermediate resistant.
3. If not both of the above, then the drug is classified as susceptible.

The antiretroviral drugs and their classification as resistant, intermediate or susceptible are tabulated as high throughput drug resistant report. Each classified drug is color-coded in the report as – Red for highly resistant, Orange for intermediate resistant and green for susceptible. The frequency of drug resistant mutations (DRMs) is also calculated from RAMICS output FASTM file and generates an “ultra deep prevalence plot” showing the prevalence of each DRM and a horizontal red line representing the cutoff prevalence.

1. **Consensus like report generation**

In a consensus like analysis, all pertinent DRMs with a prevalence in the entire dataset of greater than, or equal to, the user-defined prevalence cutoff are written to a single Sierra input file representing a “consensus-like” sequence. This sequence is analysed using the Stanford algorithm and the results are written to resistance table in the same manner as the “ultra-deep” analysis.

### 4.2.3 Evaluating the sensitivity of Seq2Res.

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](http://hivdb.stanford.edu/DR/asi/releaseNotes/index.html" \l "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 analyse these data using Seq2Res, the sequences in both the test data were fragmented into three ‘amplicons’ per sequence with an overlap of 100 nucleotides between the adjacent subsequences. The first subsequence covers the entire HIV protease gene and so it is called as fragment PR. The second subsequence covers the first half of HIV reverse transcriptase gene, so it is called RT1 fragment and the third subsequence covers the second half of HIV reverse transcriptase gene, so it is called RT2 fragment.

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.

### 4.2.4 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’ sequences was generated for each resistant sequence by reverting each of the DRMs back to wildtype. 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.

### 4.2.5 Simulation of high throughput sequencing amplicons

Each of the ‘seed’ sequences was then fragmented into three subsequences representing an individual amplicon. These amplicons corresponded to HXB2 positions X-Y (PR), M-N (RT1) and J-K (RT2)

ART is a next generation sequence data simulator tool (Huang et al., 2012). ART can create a sequence profile from Roche/454 real sequence data and use the profile to generate Roche/454 sequence data with real sequencing errors from a single sequence. The fragment PR, RT1 and RT2 FASTA sequence file was submitted to the ART tool to simulate 20,000 Roche/454 sequence reads for each fragment, now called amplicon sequences, in FASTQ format.

### 4.2.5.2Generation of different known prevalence of DRM data

For each 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 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.

### 4.2.6 Computational Infrastructures

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).

Seq2Res is available to be run online via an API. A web front end has been developed which uses the API make Seq2Res easily available to clinicians and researchers. The web front end allows user to upload the HIV genotypic data for the viral drug resistance test and then download the drug resistance report. The API can also be accessed from other interfaces such as command line scripts. The API and the web front end have been developed by Baruch Lubinsky.

### 4.3 Results

### 4.3.1 Comparison of Mutation call and Resistance Level Call for the Mutations from Web Sierra and Seq2Res

For the submitted “large data set” and “very large data sets”, web Sierra reported mutated amino acids at the drug resistant codon positions and the resistance level – R, I or S - of those drug resistant mutated amino acids. Seq2Res mapped the sequences to HIV *pol* reference sequence and reported the amino acids at all the codon positions, which are then translated to codon position in genomic region and then submitted to local sierra, which reported the resistance level of amino acids at drug resistant codon positions.

The comparison of DRMs reported by web Sierra and Seq2Res mapping tool – RAMICS showed that except for a single drug resistant mutation in reverse transcriptase codon position 236 in a sequence from “large data set”, there was a 100% congruency in mutation calls. The web Sierra reported amino acid Leucine (236L) while RAMICS reported wild type amino acid Proline (236P) at the discordant codon position.

The alignment of the reference sequence and the sequence containing the discordant codon position was obtained from RAMICS sequence alignment output. We manually reviewed the alignment in an alignment viewer tool. 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 4.6 1A**). The alignment at the codon positions after the codon position 234 showed wide type (**Figure 4.6 1B**). On the other hand, bases on the alignment reported by web Sierra, amino acid at codon positions 234 was proline (P) while 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 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 web Sierra. Therefore, we concluded that web Sierra alignment had no deletion at codon position 234 (**Figure 4.6 2C**). This resulted the rest of the sequences to go out of reading frame, reporting different amino acids than the wild type at all codon positions (**Figure 4.6 2D)**.

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

The ten ‘seed’ sequences used for simulation we are also analysed using web sierra. The five ‘resistant’ sequences were confirmed to 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**)..

Web Sierra reports drug resistivity to DRMs in five resistance levels - high-level resistance, intermediate resistance, low-level resistance, potential low-level resistance and susceptible – that are converted to three resistance level – Resistance (R), Intermediate (I) and Susceptible (S) - according to 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 five sequences without any DRMs were observed to be sensitive to the antiretroviral drugs (**Table 4.6**).

### 4.3.2.1. Seq2Res analysis of simulated data

**Initial demultiplexing**

All 105 simulated datasets, each containing 10,000 sequence reads and varying proportions of resistant and sensitive sequence reads were analyzed independently of each other using Seq2Res. In Seq2Res, initial demultiplex step involved identifying sequence reads that had no primer and MID tag sequence and discarding them from subsequent analysis.

The simulated datasets had no MID tag sequence and thus, demultiplexing the sequence reads with MID tag sequence was skipped. The information “NOMID” was added to the sequence ID. In the next step, sequence reads were demultiplexed using primer information. There was no any sequence reads discarded in this step and that was because sequence reads in the input file were already selected from a larger simulated dataset generated by high throughput simulating tool ART (Huang et al., 2012) using primer before mixing the resistant and sensitive sequence reads. Thus, 100% of sequence reads were passed through subsequent analysis.

**Quality trimming of simulated datasets**

All the sequence reads in every simulated dataset were quality trimmed using QTrim at mean quality score of 20. Interestingly, not a single sequence read was discarded from all 105 simulated datasets. This showed that high quality simulated sequence reads were simulated from ART (Huang et al., 2012) simulator.

In the subsequent analysis, sequence reads were binned together by amplicon and not by MID as no MID information was provided. There was only one amplicon in each dataset; therefore, all the sequences were binned together resulting in only one dataset in all 105 datasets. The dataset generated by binning was subsequently analyzed.

**Reference mapping of binned dataset and optimal full-length evaluation**

Every sequence reads in the dataset generated by binning were mapped to a standard HIV *pol* reference sequence using a mapping tool – RAMICS. For every mapped sequence reads, RAMICS added information either “Y” for optimal full-length if a sequence read was extended from the first to the last DRM positions of its amplicon or “N” if it was not.

HIV drug resistance testing required that a sequence read covered all DRM position of its amplicon. Therefore, sequence reads that were not optimal full-length were discarded from subsequent analysis. The number of sequence reads that were discarded as non optimal full-length was observed to be correlated for a specific amplicon across the five samples and the seven predefined prevalence levels. The average number of sequence reads discarded in amplicon PR was 180, in amplicon RT1 was 2535 and for RT2 was 396 across the five samples and seven predefined prevalence levels (**Table 4.7**). Surprisingly, the number of sequence reads for sample 56252 in amplicon RT2 was none at all seven predefined prevalence levels while other samples had lost an average of 500 sequence reads for the same amplicon across the seven predefined prevalence levels. Thus, for RT2 amplicon an average number of sequence reads discarded as non optimal full-length reduced to 396. The resulting dataset of optimal full-length sequence reads were passed for subsequent downstream analysis.

**Codon translation and Resistance testing**

RAMICS output FASTA format like FASTM file that reported a sequence ID followed by a list of comma separated codon positions with respect to the reference sequence and amino acids at those positions of every full-length sequence read. The codon positions were translated to the codon positions with respect to the gene and only the codon positions that were in the list of standard drug resistant mutation codon positions for the gene were selected. Subsequently, these selected codon positions of the sequences were saved in a file and analyzed in the subsequent downstream processes.

The file containing the translated and selected drug resistant mutation codon positions with amino acids was passed through multiple genotypic resistance interpretation algorithms in local Sierra (Liu and Shafer, 2006) for HIV drug resistance testing on every codon position of a sequence. Sierra algorithm produced an output XML file containing a resistance level –R, I or S - for codon position in every sequence read.

**Prevalence of known drug resistant mutations**

For every sequence read, the drug resistant mutations of the sequences and their resistance levels were extracted and tabulated in tab-delimited file such that for defined sequence of antiretroviral drugs, the table listed out the resistance level R, I or S and the DRMs of a sequence for each drug. The tab-delimited file was obtained for all 105 input simulated dataset.

Sequence reads in each datasets were derived from a resistant and a sensitive sequence reads by simulation, and therefore, the dataset derived from a sample’s resistant and sensitive sequence must have the DRMs from the resistant sequence as reported by web Sierra (Liu and Shafer, 2006).

Of the 105 datasets, every simulated dataset was analyzed for the prevalence of DRMs that it contained from tab-delimited file. From a tabulated file for every input simulated dataset, out of the total sequences that went through drug resistance testing in local Sierra, the prevalence of sequence reads containing each known mutations for the sample was calculated (**Figure 4.7**).

Each simulated dataset had defined proportion of resistant and sensitive sequences and therefore, the prevalence of known mutations in the dataset would be expected to be similar to the prevalence of resistant sequences in the mixture. Analysis of simulated data using Seq2Res showed that, in all cases, the observed prevalences of the known mutations very similar to the expected prevalence (**Figure 4.7**).

Analysis of all the observed prevalences across the samples from the same expected prevalence level showed that the mean of the observed prevalence are equal to the corresponding expected prevalence and the standard deviation is less than 1 except at the expected prevalence of 50% (**Table 4.8**). The high standard deviation at expected prevalence of 50% was due to higher observed prevalence (55.42%) for DRM K103N in amplicon RT2 of sample 2368 (**Figure 4.7 E**) and lower observed prevalence (46.75%) for DRM K65R in amplicon RT1 of sample 56252 (**Figure 4.7 A**) than expected prevalence of 50%.

Further analysis on the dataset of sample 2368’s amplicon RT2 at 50% expected prevalence showed that there were 4440 sequences with DRM K103N and 5057 sequences without DRM K103N in the final analysis result. As sequences with DRM were lower than sequences without DRM, the observed prevalence was lower than expected prevalence.

Similarly, in dataset of sample 56252’s amplicon RT1 at 50% expected prevalence showed that there were 4881 sequences with DRM K65R and 3780 sequences without DRM K65R in the final analysis. Therefore, the observed prevalence was higher than the expected prevalence.

**Effect of simulation error in the observed prevalence**

All 105 simulated datasets were prepared with a fixed proportion of resistant and sensitive simulated sequence, thus, the number of simulated resistant sequences in each simulated dataset was known. The final number of sequences for prevalence analysis and the initial number of sequences with the same DRM had to be same with their difference zero. However, in few amplicons, the number of simulated sequences with a DRM in the final prevalence analysis was observed to be greater than the number of simulated sequence with the same mutation mixed in fixed proportion while creating the data. Those extra numbers of DRMs were created by simulation errors (**Table 4.9**).

Sequencing error was observed to effect at the low expected prevalence level. Few sequences with simulation error had increased the observed prevalence up to twice the expected prevalence (**Table 4.9**).

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

The Seq2Res web interface (available at <http://hiv.sanbi.ac.za/tools/#/seq2res>) 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 4.9**). 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 4.10**). 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 4.11**). 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 4.12**).

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 (consensus-like or ultra-deep depending upon the end-users selection upon job submission) using the DRMs with prevalence greater or equal to defined cutoff prevalence (**Figure 4.13**). The susceptibility of drugs for the patient is shown by color code – Red: Resistant, Orange: Intermediate resistant, Green: Susceptible (**Figure 4.13**).

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 4.14**).

Back at the bottom of the result page obtained after clicking a job name, there is a link “Further Page”, which takes user to the advanced result page that contains drug resistant report generated using high throughput sequences. A click on the link “Ultra deep drug resistance report” downloads high throughput drug resistance reports of all patients in one PDF file. Each patient specific high throughput drug resistance report is a table that displays antiretroviral drug classes – NRTIs, NNRTIs, PIs or INs in first column, all clinical antiretroviral drugs in second column, the number of sequences analyzed in third column, the percentage of sequences that are resistant to corresponding drug in fourth column, the percentage of sequences that are intermediate resistant to the corresponding drug in fifth column and the DRMs in the sixth column (**Figure 4.15**). The susceptibility of drugs for the patient is shown by color code is same as in “consensus like” drug resistant report.

Below the link for high throughput drug resistant report, there is a link to download the “consensus like” DRMs each patient.

### 4.4 Discussion and Conclusion

HIV drug resistance test is essential to characterize the viral population (Baba et al., 2005; Simen et al., 2009a) and to treat HIV infected individuals with correct combination of antiretroviral drugs to suppress the viral replication for longer time period and increase their life expectancy (Harrison et al., 2010; van Sighem et al., 2010). Roche/454 high throughput sequencing technology has shown great potential to genotype even the minor HIV variants that are in most cases drug resistant variants (Lataillade et al., 2010; Simen et al., 2007; Simen et al., 2009b; Varghese et al., 2009). But Roche/454 sequencing technology produces up to a million sequences ([www.454.com](http://www.454.com)) and manual analysis at any processing step is not feasible. Seq2Res computational pipeline is designed to analyze this huge amount of data for HIV drug resistance test and produce sample specific drug resistance report in both HTS method and Consensus like method; and also show the prevalence of DRMs.

The full-length amplicon sequences that begins from user defined nucleotide “start” position to “end” position are analyzed to consider all the DRMs that the amplicon covers. However, large number of sequences gets discarded although they cover all the DRMs of the amplicon. The DRM codons are usually inwards from the 5’ and 3’ ends in the sequences. The sequences are discarded as not full-length because they don’t cover those non-DRM codon positions at the ends that are not important in the analysis. 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 saves large number of sequences that do not cover non-DRMs codon positions at the ends from being discarded.

Seq2Res DRM calls were compared with the standard tool – web Sierra (Liu and Shafer, 2006) using two sample data. The comparison result was highly congruent; out of over 10683 DRMs in PR and 7867 DRMs in RT genes of sequences in “large data set” and 30765 DRMs in PR and 22569 DRMs in RT genes of sequences in “very large data set”, a single discordant mutation call was observed. However, a close observation in the alignment of the sequence with discordant mutation call and reference sequence showed that Seq2Res mapping tool RAMICS did the alignment correctly in codon space. This showed that Seq2Res was better than web Sierra at DRM call.

Simulated datasets were prepared with known number of simulated sequences containing DRMs. The datasets were analyzed using Seq2Res for the prevalence of the DRMs. Reference mapping of the simulated sequences discarded about 1.8% of PR, 20% of RT1 and 5% of RT2 simulated sequences for all samples at all expected prevalence level (56252’s RT amplicon) as non optimal full-length. Discarding of huge amount of sequences had less effect on the observed prevalence to deviate from the expected prevalence. The observation showed that the observed prevalence of the DRMs in the datasets was very close to the expected prevalence with the mean of observed prevalences from all samples was same as expected prevalence and the standard deviation was close to zero.

Counting the final number of DRMs in the final prevalence analysis and the number of sequences with same DRMs before reference mapping assessed simulation errors in the DRM codon positions. The simulation errors influenced the observed prevalence to be higher than the expected prevalence at the low expected prevalence level.

Roche/454 generates errors at homopolymer region in real sequencing (Gilles et al., 2011). As the simulator tool used the profile of a real sequence data, the homopolymer errors in the simulated data were assessed. The observation showed that for 23% of the sequences, the homopolymer region was sequenced correctly. The chance that simulation generated two insertion errors at homopolymer region was the highest (43.3%) (**Table 4.10).**

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.

Bennett, DE, Myatt, M, Bertagnolio, S, Sutherland, D, Gilks, CF (2008) Recommendations for surveillance of transmitted HIV drug resistance in countries scaling up antiretroviral treatment. *Antivir Ther* **13 Suppl 2**: 25-36.

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.

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.

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.

Jabara, CB, Jones, CD, Roach, J, Anderson, JA, Swanstrom, R (2011) Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID. *Proc Natl Acad Sci U S A* **108**: 20166-20171.

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.

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, Kuma, K, Toh, H, Miyata, T (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res* **33**: 511-518.

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.

Katoh, K, Toh, H (2008) Recent developments in the MAFFT multiple sequence alignment program. *Brief Bioinform* **9**: 286-298.

Katoh, K, Toh, H (2010) Parallelization of the MAFFT multiple sequence alignment program. *Bioinformatics* **26**: 1899-1900.

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.

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.

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.

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

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.

Rhee, SY, Gonzales, MJ, Kantor, R, Betts, BJ, Ravela, J, Shafer, RW (2003) Human immunodeficiency virus reverse transcriptase and protease sequence database. *Nucleic Acids Res* **31**: 298-303.

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

Shafer, RW (2006) Rationale and uses of a public HIV drug-resistance database. *J Infect Dis* **194 Suppl 1**: S51-58.

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 (2009a) 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.

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, Kozal, MJ (2009b) Low-abundance drug-resistant viral variants in chronically HIV-infected, antiretroviral treatment-naive patients significantly impact treatment outcomes. *J Infect Dis* **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.

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.