**ABSTRACT**

In countries with a high burden of HIV, drug resistance testing is not routinely performed in a clinical setting due to the cost. The advent of Roche/454 pyrosequencing holds great promise in the development of a high-throughput, robust, reliable and affordable HIV drug resistance test. Further, the depth of coverage obtained using Roche/454 massive parallel pyrosequencing means that the presence of clinically relevant low abundance drug resistant viral variants within an individual’s viral population can also be explored.

The sheer volume of data generated by such an approach means that there is a need for powerful, sensitive and user-friendly bioinformatics applications for management and analysis of the data. We have developed a bioinformatics pipeline (Seq2Res) that takes sequence data directly from the Rohce/Roche/454 sequencing platform and outputs drug resistance information for each sample.

As part of this pipeline we have developed a novel approach (QTrim) for quality trimming of the Roche/454 sequence reads. Roche/454 sequence data has a quality score associated with every nucleotide and these need to be accounted for prior to downstream analysis of the data, for accurate drug resistance results. We compared the performance of QTrim with widely used algorithms on both good and poor quality Roche/454 sequence data. We evaluated methods based on mean read length, total number of reads and the percentage of poor quality bases in the resulting output. We find that our approach performs marginally better than the next best method for good quality data and significantly outperforms all methods when poor quality data is analyzed.

Sequence reads often contain sequencing errors and PCR associated errors. These errors can be corrected generating a consensus sequence from the sequence reads derived from a viral template sequence. Sequence reads deriving from the same viral template sequence can be identified incorporating a Primer ID to the cDNA primer during reverse transcription process. Seq2Res is capable of processing sample specific sequence data produced using Primer ID, generate consensus sequences and generate drug resistance report for the sample.

We tested the sensitivity of Seq2Res drug resistant mutation detection using the two real consensus sequence datasets downloaded from Stanford Database. Drug resistant mutations in the real sequence datasets were obtained from Web Sierra HIV drug resistance interpretation algorithm and Seq2Res. The mutations from both the Web Sierra and Seq2Res were compared. We found that Seq2Res identified all the drug resistant mutations that were reported by the Web Sierra. In addition, Seq2Res was also able to identify a drug resistant mutation that was reported by Web Sierra incorrectly.

We also tested the ability of Seq2Res to report the prevalence of the drug resistant mutations correctly using five simulated datasets each containing known prevalence of drug resistant mutations at known codon positions.

Finally, we use Seq2Res for drug resistance prediction from the real biological datasets. The datasets were generated as part of the CIPRA-SA (Comprehensive International Program for Research in AIDS in South Africa). 831 HIV infected individuals were enrolled in the study that included single dose nevirapine (sdNVP) experienced patients through prevention from mother to child (PMTCT) therapy and antiretroviral naïve patients. 562 of 831 patients were followed; the rest were not included for death, toxicity, withdrawal of consent or lost to follow up. The date of NVP exposure for the drug-experienced patients was known. Baseline samples were obtained before antiretroviral treatment (ART) initiation. Virologic failure (VF) was defined as decline of viral load less than 1.5 log10 from baseline to 12 weeks of treatment or two consecutive samples from a patient taken four weeks apart have viral load greater than 1000 RNA copies/ml. VF to ART was exhibited in 79 patients including 15 patients who failed second line ART. 71% of these patients received the drug combination D4T-3TC-EFV, 20% received D4T-3TC-NVP, 8% received D4T-3TC-LPV/r and 1% received D4T-3TC-NLF. 471 samples were genotyped using Roche/454 and 630 samples were genotyped using Roche/454 FLX. Consensus sequence for 349 samples from conventional genotyping was also available. Genotypic data from FLX and Junior were input in to Seq2Res pipeline for resistance prediction in the samples.

Roche/454 FLX and Roche/454 Junior genotypic data for baseline samples showed that the number of samples that were predicted resistant increased when the prevalence of resistance was decreased to 1% cutoff. The number of samples predicted resistant was higher in PMTCT group than in no-PMTCT group at the prevalence cutoff 15% and below. It showed that NVP resistance was more likely to develop in NVP exposed patients in PMTCT group than in the drug naïve group. The time (in days) of NVP exposure before ART initiation was calculated for the individuals in both PMTCT group and no-PMTCT group. The result showed that the prediction of NVP resistance significantly correlates (p < 0.05) with time since NVP exposure. The samples from patients that receive sdNVP within ~6 month before ART initiation were predicted resistant while the samples from patients that receive sdNPV after ~2 years were predicted sensitive. This showed that inclusion of NVP in the ART cocktail for the patients who received NPV before 6 months or less time could compromise the therapy.

Roche/454 FLX and Roche/454 Junior genotypic data from first line virologic failure (VF1) samples were studied for resistance. Seq2Res reported that up to 100% of the VF1 samples from PMTCT and ~65% of the VF1 samples from no-PMTCT were resistant.

Keywords: Roche/454 pyrosequencing, HIV Antiretroviral Drug Resistance Testing, antiretroviral therapy, Low Abundance Viral Variants, QTrim, Quality Trimming, Primer ID, Consensus sequence, Drug resistant mutations, prevalence, single dose Nevirapine, genotype, virologic failure, PMTCT