Pacific Bioscience’s Single Molecule Read Time technology is a new commercial technology available in the market ([www.pacificbiosciences.com](http://www.pacificbiosciences.com)). It is considered the third generation sequencing technology. The technology does not involve PCR amplification of the template DNA; instead the base sequencing is done on single molecule of a DNA, enabling the detection of variation at molecule level. Besides this advantage, it provides the read length of about 10,000 bases (Eid et al., 2009; McCarthy, 2010).

DNA sequencing takes place in the zero mode waveguide (ZMW) (Levene et al., 2003). ZMW is a nano-size chamber that is 7 nanometer in diameter and 10 nanometer in depth (McCarthy, 2010). A DNA template and polymerase complex is immobilized at the base of a ZMW and different color flurophore labeled nucleotides are added into a ZMW chamber. During base incorporation at sequencing step, nucleotide fluorescence is detected with the light that illuminates the ZMW chamber, followed by cleavage of the flurophore.

1.12 HIV-1 Drug resistance Genotyping in the era of high throughput sequencing (HTS)

As conventional Sanger-based genotyping is unable to characterize the HIV viral quasispecies at less than 20% prevalence, a true HIV diversity cannot be ascertained (Schuurman et al., 2002). An alternative genotyping method is required that has the ability to sequence the HIV population to “deeper” level and characterize the overall spectrum of true viral diversity in the viral quasispecies.

Different approaches like sensitive real time PCR drug resistant test (Johnson et al., 2008), allele-specific RT-PCR (Palmer et al., 2006) and single genome sequencing (Palmer et al., 2005) were developed but were highly expensive and difficult to implement as a HIV drug resistant diagnostic tool.

High throughput sequencing technologies are able to generate millions of sequence reads (reviewed in (Metzker, 2005, 2009)) and, therefore, have the capability to sequence low abundance variants in the viral quasispecies. Thus, HTS technologies hold great potential to be a more sensitive, cost-effective alternative to Sanger sequencing for HIV drug resistance genotyping.

HIV drug resistance genotyping requires the correct identification of drug resistant mutation in HIV sequence data. The drug resistant mutations to conventional ARVs occur across the HIV protease and reverse transcriptase genes. Because the pattern of drug resistant mutations occurring in the HIV genes determines the resistance level (Hamers et al., 2012; Hoffmann et al., 2007), the entirety of each of the genes must be sequenced to cover all of the relevant DRMs with long reads preferable to ensure co-occurring DRMs are correctly identified.

Therefore, the next generation sequencing technology that produces the longest sequence read length must be the choice for HIV genotyping. Among the current next generation HTS technologies, Roche/454 has the longest read length (up to 600 bases) (Loman et al., 2012) and has been used widely for HIV genotypic drug resistance test (Astrovskaya et al., 2011; Bansode et al., 2013; Beerenwinkel and Zagordi, 2011; Bordoni et al., 2008; Delobel et al., 2011; Dudley et al., 2012; Eriksson et al., 2008; Fischer et al., 2010; Hedskog et al., 2010; Lataillade et al., 2010; Le et al., 2009). We describe Roche/454 technology as the HTS from here.

An advantage of HTS is that it allows for the sequencing of multiple samples at the same time, with the addition of a specific tag sequence for each sample (Hamady et al., 2008), which enables low cost drug resistant genotyping per sample (Dudley et al., 2012). This means that low cost large scale HIV drug resistance genotyping might be possible in low and middle-income countries, for example, sub-Saharan Africa, where the prevalence of HIV infection is high.

Although HTS is considered as an alternative to conventional method to characterize the HIV diversity in the viral population, it is required that HTS has the sensitivity to genotype the HIV genes correctly. However, there are inherent sequencing errors in the sequence data produced by HTS. The errors might be generated before actual sequencing of the viral genes at PCR amplification or during actual sequencing in HTS.

A PCR amplification step to generate millions of template DNA precedes the actual HTS sequencing step. The PCR amplification step has embedded errors such as DNA recombination (Kanagawa, 2003; Meyerhans et al., 1990; Yang et al., 1996), DNA synthesis errors (Hughes and Totten, 2003; Mansky and Temin, 1995) and DNA re-sampling errors (Liu et al., 1996). These errors add artificial variation in the HIV-1 population, confounding the real ones.

developed

Furthermore, the well-known error in HTS is the homopolymer error, which is generation of insertion or deletion of a base due to miscalculation of fluorescence light detection corresponding to the number of bases incorporated in the region of single repetitive base (Bordoni et al., 2008; Gilles et al., 2011; Kunin et al., 2009; Loman et al., 2012; Luo et al., 2012; Margulies et al., 2005; Wang et al., 2007). Homopolymer error needs to be accounted and identified correctly as there are homopolymer regions in HIV protease and reverse transcriptase genes and these regions have drug resistance mutations (**Figure 1.15**).

A quality score is attached to all the bases in sequence reads, which indicates a probability that the base call is incorrect (Cock et al., 2010; Schmieder and Edwards, 2011). Quality scores of the bases can be used as a possible way to learn that the sequenced base is an error both there are billions of bases in a sequence data file and manual intervention for error profiling is not possible. Nonetheless, poor quality bases might compromise the downstream data analysis (Schmieder and Edwards, 2011). Quality trimming tools are used to trim out poor quality bases to ensure only high quality bases are analyzed to generate high quality result. However, PCR related base error could still have high quality scores and therefore, quality trimming of sequence data only is not enough remove the base errors.

A technique to correctly identify the sequencing error irrespective of base quality score is multiple sequence alignment (Bebenek et al., 1993; Edgar, 2004; Katoh et al., 2009) of HTS sequences with a standard reference. As amino acids in genes are coded by codon in the genes, the mapping at “codon-space” (Delport et al., 2010) can accurately identify the sequencing errors.

A new technology has been introduced that enables tagging of every viral sequence with a specific sequence (called PrimerID) before the PCR and sequencing step (Jabara et al., 2011). The technology enables tracking of every sequence originating from a template viral sequence, which can then be used to generate a consensus sequence that represents the original viral sequence (Jabara et al., 2011). In theory, the sequencing errors are rectified while generating consensus sequences.

HTS technology associated with PrimerID technology (Jabara et al., 2011) might be widely used in future to correct errors in homopolymer region besides other errors. Drug resistance test with accurate genotyping of HIV quasispecies would give accurate drug susceptibility for the antiretroviral drugs.

1.12 Thesis Outline

Chapter 2: The chapter introduces and describes, in detail, a novel algorithm QTrim for quality trimming of Roche/454 ultra-deep high throughput sequence data. QTrim is optimized for both poor and high quality data. HIV-1 resistance test requires high quality genotypic data and QTrim is designed to quality control the data for the test. QTrim is compared to other widely used quality trimming tools and the comparative analysis result is also presented. The entire chapter is produced as a paper for publication entitled “QTrim: A novel tool for the quality trimming of sequence reads generated using the Roche/454 sequencing platform” (Shrestha, RK and Travers, Simon; being reviewed).

Chapter 3: The chapter introduces Primer ID technique recently developed to correct ultra-deep sequencing errors. The chapter also discusses a comparative analysis of sequence data with and without Primer ID approach and the impact of the technology on downstream analysis. The Primer ID approach sequence data is obtained from a study conducted to observe HIV virus response on the vaccine. The study was conducted in the University of Cape Town.

Chapter 4: The chapter introduces and describes the computational pipeline called Seq2Res that facilitates low cost HIV drug resistance test. The chapter describes a workflow of the pipeline, the requirements of the pipeline, HIV drug resistance output files and plots that summaries overall analysis. The test data is obtained from a study called CIPRA-SA, which is a trial of antiretroviral drug monitoring strategy in resource poor setting. The chapter is included in paper publication.

Chapter 5: This chapter discusses on comparison of clinical data and sequence data analyzed using Seq2Res pipeline. Excitingly, the sequence data analyzed using Seq2Res agrees to clinical data.

Astrovskaya, I, Tork, B, Mangul, S, Westbrooks, K, Mandoiu, I, Balfe, P, Zelikovsky, A (2011) Inferring viral quasispecies spectra from 454 pyrosequencing reads. *BMC Bioinformatics* **12 Suppl 6**: S1.

Bansode, V, McCormack, GP, Crampin, AC, Ngwira, B, Shrestha, RK, French, N, Glynn, JR, Travers, SA (2013) Characterizing the emergence and persistence of drug resistant mutations in HIV-1 subtype C infections using 454 ultra deep pyrosequencing. *BMC Infect Dis* **13**: 52.

Bebenek, K, Abbotts, J, Wilson, SH, Kunkel, TA (1993) Error-prone polymerization by HIV-1 reverse transcriptase. Contribution of template-primer misalignment, miscoding, and termination probability to mutational hot spots. *J Biol Chem* **268**: 10324-10334.

Beerenwinkel, N, Zagordi, O (2011) Ultra-deep sequencing for the analysis of viral populations. *Curr Opin Virol* **1**: 413-418.

Bordoni, R, Bonnal, R, Rizzi, E, Carrera, P, Benedetti, S, Cremonesi, L, Stenirri, S, Colombo, A, Montrasio, C, Bonalumi, S, Albertini, A, Bernardi, LR, Ferrari, M, De Bellis, G (2008) Evaluation of human gene variant detection in amplicon pools by the GS-FLX parallel Pyrosequencer. *BMC Genomics* **9**: 464.

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.

Delobel, P, Saliou, A, Nicot, F, Dubois, M, Trancart, S, Tangre, P, Aboulker, JP, Taburet, AM, Molina, JM, Massip, P, Marchou, B, Izopet, J (2011) Minor HIV-1 variants with the K103N resistance mutation during intermittent efavirenz-containing antiretroviral therapy and virological failure. *PLoS One* **6**: e21655.

Delport, W, Scheffler, K, Botha, G, Gravenor, MB, Muse, SV, Kosakovsky Pond, SL (2010) CodonTest: modeling amino acid substitution preferences in coding sequences. *PLoS Comput Biol* **6**.

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

Eid, J, Fehr, A, Gray, J, Luong, K, Lyle, J, Otto, G, Peluso, P, Rank, D, Baybayan, P, Bettman, B, Bibillo, A, Bjornson, K, Chaudhuri, B, Christians, F, Cicero, R, Clark, S, Dalal, R, Dewinter, A, Dixon, J, Foquet, M, Gaertner, A, Hardenbol, P, Heiner, C, Hester, K, Holden, D, Kearns, G, Kong, X, Kuse, R, Lacroix, Y, Lin, S, Lundquist, P, Ma, C, Marks, P, Maxham, M, Murphy, D, Park, I, Pham, T, Phillips, M, Roy, J, Sebra, R, Shen, G, Sorenson, J, Tomaney, A, Travers, K, Trulson, M, Vieceli, J, Wegener, J, Wu, D, Yang, A, Zaccarin, D, Zhao, P, Zhong, F, Korlach, J, Turner, S (2009) Real-time DNA sequencing from single polymerase molecules. *Science* **323**: 133-138.

Eriksson, N, Pachter, L, Mitsuya, Y, Rhee, SY, Wang, C, Gharizadeh, B, Ronaghi, M, Shafer, RW, Beerenwinkel, N (2008) Viral population estimation using pyrosequencing. *PLoS Comput Biol* **4**: e1000074.

Fischer, W, Ganusov, VV, Giorgi, EE, Hraber, PT, Keele, BF, Leitner, T, Han, CS, Gleasner, CD, Green, L, Lo, CC, Nag, A, Wallstrom, TC, Wang, S, McMichael, AJ, Haynes, BF, Hahn, BH, Perelson, AS, Borrow, P, Shaw, GM, Bhattacharya, T, Korber, BT (2010) Transmission of single HIV-1 genomes and dynamics of early immune escape revealed by ultra-deep sequencing. *PLoS One* **5**: e12303.

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.

Hamady, M, Walker, JJ, Harris, JK, Gold, NJ, Knight, R (2008) Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. *Nat Methods* **5**: 235-237.

Hamers, RL, Sigaloff, KC, Wensing, AM, Wallis, CL, Kityo, C, Siwale, M, Mandaliya, K, Ive, P, Botes, ME, Wellington, M, Osibogun, A, Stevens, WS, Rinke de Wit, TF, Schuurman, R (2012) Patterns of HIV-1 drug resistance after first-line antiretroviral therapy (ART) failure in 6 sub-Saharan African countries: implications for second-line ART strategies. *Clin Infect Dis* **54**: 1660-1669.

Hedskog, C, Mild, M, Jernberg, J, Sherwood, E, Bratt, G, Leitner, T, Lundeberg, J, Andersson, B, Albert, J (2010) Dynamics of HIV-1 Quasispecies during Antiviral Treatment Dissected Using Ultra-Deep Pyrosequencing. *PLoS ONE* **5**: e11345.

Hoffmann, C, Minkah, N, Leipzig, J, Wang, G, Arens, MQ, Tebas, P, Bushman, FD (2007) DNA bar coding and pyrosequencing to identify rare HIV drug resistance mutations. *Nucleic Acids Res* **35**: e91.

Hughes, JP, Totten, P (2003) Estimating the accuracy of polymerase chain reaction-based tests using endpoint dilution. *Biometrics* **59**: 505-511.

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.

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.

Kanagawa, T (2003) Bias and artifacts in multitemplate polymerase chain reactions (PCR). *J Biosci Bioeng* **96**: 317-323.

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

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.

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.

Le, T, Chiarella, J, Simen, BB, Hanczaruk, B, Egholm, M, Landry, ML, Dieckhaus, K, Rosen, MI, Kozal, MJ (2009) Low-abundance HIV drug-resistant viral variants in treatment-experienced persons correlate with historical antiretroviral use. *PLoS One* **4**: e6079.

Levene, MJ, Korlach, J, Turner, SW, Foquet, M, Craighead, HG, Webb, WW (2003) Zero-mode waveguides for single-molecule analysis at high concentrations. *Science* **299**: 682-686.

Liu, SL, Rodrigo, AG, Shankarappa, R, Learn, GH, Hsu, L, Davidov, O, Zhao, LP, Mullins, JI (1996) HIV quasispecies and resampling. *Science* **273**: 415-416.

Loman, NJ, Misra, RV, Dallman, TJ, Constantinidou, C, Gharbia, SE, Wain, J, Pallen, MJ (2012) Performance comparison of benchtop high-throughput sequencing platforms. *Nat Biotechnol* **30**: 434-439.

Luo, C, Tsementzi, D, Kyrpides, N, Read, T, Konstantinidis, KT (2012) Direct comparisons of Illumina vs. Roche 454 sequencing technologies on the same microbial community DNA sample. *PLoS One* **7**: e30087.

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.

McCarthy, A (2010) Third generation DNA sequencing: pacific biosciences' single molecule real time technology. *Chem Biol* **17**: 675-676.

Metzker, ML (2005) Emerging technologies in DNA sequencing. *Genome Res* **15**: 1767-1776.

Metzker, ML (2009) Sequencing technologies - the next generation. *Nat Rev Genet* **11**: 31-46.

Meyerhans, A, Vartanian, JP, Wain-Hobson, S (1990) DNA recombination during PCR. *Nucleic Acids Res* **18**: 1687-1691.

Palmer, S, Boltz, V, Martinson, N, Maldarelli, F, Gray, G, McIntyre, J, Mellors, J, Morris, L, Coffin, J (2006) Persistence of nevirapine-resistant HIV-1 in women after single-dose nevirapine therapy for prevention of maternal-to-fetal HIV-1 transmission. *Proc Natl Acad Sci U S A* **103**: 7094-7099.

Palmer, S, Kearney, M, Maldarelli, F, Halvas, EK, Bixby, CJ, Bazmi, H, Rock, D, Falloon, J, Davey, RT, Jr., Dewar, RL, Metcalf, JA, Hammer, S, Mellors, JW, Coffin, JM (2005) Multiple, linked human immunodeficiency virus type 1 drug resistance mutations in treatment-experienced patients are missed by standard genotype analysis. *J Clin Microbiol* **43**: 406-413.

Schmieder, R, Edwards, R (2011) Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27**: 863-864.

Schuurman, R, Brambilla, D, de Groot, T, Huang, D, Land, S, Bremer, J, Benders, I, Boucher, CA (2002) Underestimation of HIV type 1 drug resistance mutations: results from the ENVA-2 genotyping proficiency program. *AIDS Res Hum Retroviruses* **18**: 243-248.

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.

Yang, YL, Wang, G, Dorman, K, Kaplan, AH (1996) Long polymerase chain reaction amplification of heterogeneous HIV type 1 templates produces recombination at a relatively high frequency. *AIDS Res Hum Retroviruses* **12**: 303-306.