Title: Development of novel approaches for HIV drug resistance detection using nanopore sequencing technology.

Background

HIV-antiretroviral-therapy (HART) is currently a widely used clinical option for managing HIV-1 infected people and is effective for reducing the viral burden to very low levels that are undetectable to conventional HIV-viral-load (HVL) tests[1]. HIV viruses within an HIV-infected person may differ in the drugs they are susceptible to and these viruses continue evolving to evade the drugs. The major cause(s) of HART treatment failure in people undergoing therapy is the emergence of viruses with HIV-drug-resistance (HIVDR) mutations[2]. The triggers of HIVDR are mainly the polymerase enzyme driven HIV replication cycle that is inherently error prone and characterized by nucleotide substitutions and in the presence of ART pressure exacerbates the problem into selection of mutations that lead to HIVDR. HIVDR may also be a consequence of the antigenic shift and drift or transmitted from one HIV infected person to a new host, referred to as pre-treatment HIVDR. Monitoring of HIVDR is imperative for informing public health policy since its population levels can be harnessed to guide decisions about potent regimens that may be included in HIV combinational antiviral therapies. Further, to meet the "third 95" goal of the 2030 UNAIDS 95-95-95 goals that aims to have 95% of HIV-infected people that are receiving HART to maintain viral suppression, strengthens the need for efficient and expeditious and cost effective HIVDR monitoring structures and methods.

The determination of mutations that confer HIVDR requires genotypic characterization of the circulating viruses in the host and currently it is most commonly done using first-generation Sanger sequencing methods and or second-generation sequencing technologies[3]. Sanger sequencing methods have a low depth and derive consensus nucleotide sequences from populations that are greater than 20% of the HIV sequences represented in the multiple viruses present per nucleotide position. Despite having low error rates of approximately 0.1%, shortcomings inherent with the Sanger sequencing methods are the inability to evaluate minority variants that appear in less than 20% of the sequences, inability for real-time data analysis since the entire coverage of the target amplicon needs to be fully attained before analysis can commence and it cannot permit the determination of whether resistance mutations are from the same or different viruses.

In 2014 a novel sequencing technology the MinION, that relies on a nanoscale protein pore, the nanopore that is embedded in a graphene membrane was released. When in an electrically charged solution the membrane can detect DNA or RNA passing through the pores by changing the ionic current during the translocation and the ionic change is characteristic to a particular nucleotide[4]. In this way the nucleotide sequence can be derived using computational algorithms in real-time. Nanopore sequencing produces long reads that can be over 100 X 10^3 base pairs and is fast and able to achieve over 250 bases per second. The data can be analyzed in real-time since each read represents a single molecule and can have coverage of an entire target amplicon. Notwithstanding its advantages, it has an inherent high error rate relative to Sanger sequencing resulting from a high signal-to-noise ratio that affects the base-calling algorithm's accuracy[5]. HIVDR monitoring using the nanopore methods requires careful evaluation to establish its reliability for clinical management of patients. Considering that it's error rates may have a huge implication for patient's quality of life if a wrong base-call results in a non-synonymous outcome that does not code for a drug resistant mutation emphasizes the need for developing optimized approaches for nanopore HIVDR monitoring.

Objectives

1. Compare the HIVDR scoring between Sanger and Nanopore sequencing method of matched samples to evaluate the concordance of results and.
2. Develop a method to PCR amplify low plasma viral load samples, genotype these with the nanopore method and derive HIVDR mutations.
3. Identify improvements to the base-calling algorithms that may obviate high error rates.
4. Provide an expeditious analysis pipeline for analysis of HIVDR of Nanopore generated sequences.
5. Establish if the various base-calling errors of the Nanopore sequencing are occurring in a stochastic or non-stochastic manner.

Methodology

The project will begin with evaluating in-house databases for sequences derived from Sanger methods and their HIVDR profile determined using the Stanford HIV drug resistance database. Archived plasma samples for these will be processed using conventional PCR for HIV-1 polymerase, reverse transcriptase and integrase amplicons for sequencing using the nanopore method to establish baseline differences between the methods in identifying HIVDR. Pilot runs for the Nanopore sequencing will be done for 2 – 5 hours to establish the optimum depth that will be reasonable for HIVDR testing. The sequences will be demultiplexed and base-called using Guppy software, quality control to remove short reads and low error rate reads. Reads will be aligned to relevant HIV consensus reference genomes and collapsed into multiple consensuses using Megahit and USEARCH software at a similarity threshold of 75% to 95% to establish the optimal value. The positive predictive values will then be determined after establishing the contrast between the HIVDR identification of the matched Sanger and Nanopore sequences for sequences occurring at a threshold of over 20%. The Sanger sequencing method has been widely used for HIVDR determination and suffices to be considered the gold-standard. Therefore, understanding the performance of nanopore sequencing against Sanger sequencing is useful information since this will tell us the basic target we need to aim for to improve nanopore sequencing for HIVDR. Key to note though is that the base-calling accuracy of nanopore sequencing has been greatly improved by novel flow cell technologies coupled with optimized translocation speeds. Nevertheless, the residual base-calling error requires to be evaluated against the gold-standard to direct the optimization for HIVDR clinical diagnostics.

Investigation of the nanopore sequencing method will be done using spiked plasma samples to optimize the genotyping method for low HVL samples. One of the problems that affects the success of PCR amplification is low amount of genomic template. Usually patients with low HVL, typically below 1000 copies/mL present PCR amplification problems. Well characterized plasma samples with known HVL will be serially diluted to mimic low HVL samples and then PCR amplified to establish their sensitivity. The optimal viral load dilution will be used to spike plasma samples for nanopore sequencing. Samples will be normalized to have a known similar concentration of spiked plasma. Spiking improves the PCR success rate of low viral load samples and these can also be genotyped. Analysis pipelines to remove the spike sample' sequences from the patient samples will be developed.

Nanopore sequencing technology can produce long reads and each sequence is from a single DNA or RNA molecule. This therefore allows the identification of the viruses that a particular HIVDR mutation(s) occur. Alignment of the collapsed sequences will be visualized in a reading frame to select those with particular HIVDR mutations and analyzed using phylogenetic and structure modelling methods to identify any association of HIVDR mutations to particular viruses.

To further understand the factors that may affect the emergence and or persistence of HIVDR, PCR amplification of patient samples will be done using random primers that can PCR amplify the viral species in the samples. The amplicons will be Nanopore sequenced and a metagenomics analysis pipeline done to identify the viruses present. This will mainly aim at identifying if any association is present between the HIVDR genotypes and magnitude to other viral pathogens in a patient. Nanopore base-calling algorithms will be assessed to identify any improvements that may reduce the error rates. Further sequence analysis to evaluate if the base-calling errors are stochastic or occur in particular motifs will be done through aligning the sequences to multiple references in a reading frame for quick visualization of differences, insertions and or deletions that may be base-calling artifacts. This will provide important information about how the base-calling algorithms may be improved for higher accuracy.

To expedite the interpretation of the Nanopore sequencing data reporting, computational scripts to parse the data to query HIVDR databases and generates reports is key. This should put into consideration the practicability of accomplishing tasks in field settings without internet, computer servers or advanced laboratories. Python libraries including SierraPy for creating a localized mirror of the Stanford HIVDR database, and data manipulation libraries like PANDAS and JSON will handy. Shell scripts will be mainly used as parsers to call the various scripts and integrate the computation analysis and reporting tasks. Containerization will be done to virtualize the various computational resources so that they can run on standalone machines and also to reduce the incompatibilities brought about by software upgrades.

My Interest in the Project and Suitability

In my early research career I accomplished [specific skills deleted]. I gained a wealth of experience in a vast scope of contemporary scientific technologies and methods, usually generating a deluge of data from the tasks that I accomplished. I began to appreciate the need for computational methods to analyze the huge data generated that included mainly genotypic data and immune response data. This was the impetus that fueled my true passion which was always computation applications for data mining. Consequently, I had a paradigm shift into computational studies aiming to acquire the basics of computational algorithms and apply these to biological data analysis and or problems. However, throughout my studies at [institution deleted] where I pursued [programme of study deleted], I was always struggling to get to speed with the rather very intensive programme. I felt this did not give me the opportunity to delve deep into big data analyses, though it was the most important real introduction into what I always wanted to do. The [deleted] opportunity revealed to me the limitless investigation that I can apply to medical research and biology data using computational approaches and most importantly how to transform this data into useful practical application. This since increased my rigor to take personal initiatives to get better at skill sets that are directly useful in this regard particularly scripting to develop in-house programs so that I can apply practical computational skills to simplify research analyses and interpretation. This PhD will give me the opportunity to have most of my time occupied in developing the skills set that will make me a seasoned bioinformatician. I am suitable for this programme since I have been doing [deleted] and my experience in [deleted] is vast. I also possess an understanding of computational skills more than the average user, I am comfortable with [skills deleted] since my current role involves [skills and tasks deleted]. Training and capacity building is one of my key ideologies and I have actively trained colleagues and interns both formally and informally. [Statement example deleted.] The PhD will involve a great deal of learning and developing novel methods that will be added capacity to [statement example] since I plan to train and pass on the expertise.

Importance of the Project Implementation

Next generation sequencing (NGS) is currently in the spotlight as a technique that may be replacing the 1st generation HIVDR Sanger sequencing method because of the increasing concern about the implications of low diversity HIVDR. The Nanopore sequencing method is a promising NGS option as it is relatively cheap since it has a minimal capital expenditure for it to be set up compared to Sanger. Nanopore can also be used in field settings where there is very limited infrastructure and produces real time data. The method however has not been fine-tuned for clinical diagnostic purpose therefore this project is imperative to fill this gap. This project once optimized and has collected enough evidence of satisfactory data to support clinical diagnostic use will eminently replace the Sanger method that has inherently expensive maintenance and running costs.

References

1. Cohen MS, Smith MK, Muessig KE, Hallett TB, Powers KA, Kashuba AD: **Antiretroviral treatment of HIV-1 prevents transmission of HIV-1: where do we go from here?** *Lancet* 2013, **382**(9903):1515-1524.
2. Lima ENC, Piqueira JRC, Camargo M, Galinskas J, Sucupira MC, Diaz RS: **Impact of antiretroviral resistance and virological failure on HIV-1 informational entropy**. *J Antimicrob Chemother* 2018, **73**(4):1054-1059.
3. Stranneheim H, Lundeberg J: **Stepping stones in DNA sequencing**. *Biotechnol J* 2012, **7**(9):1063-1073.
4. Jain M, Olsen HE, Paten B, Akeson M: **The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community**. *Genome Biol* 2016, **17**(1):239.
5. Ji H, Sandstrom P, Paredes R, Harrigan PR, Brumme CJ, Avila Rios S, Noguera-Julian M, Parkin N, Kantor R: **Are We Ready for NGS HIV Drug Resistance Testing? The Second "Winnipeg Consensus" Symposium**. *Viruses* 2020, **12**(6).

TOTAL WORD COUNT = 2089
(Excludes headers, footers and project title)