A Software Tool to Process HIV Ultra-deep Sequencing Nucleotide Data

Chien-Chi Lo[1], Brian Gaschen[1], Charles Calef[1], Thomas Leitner[1] and Bette Korber[12]

***Short Abstract* — Massive sequence reads generated by 454 Life Sciences pyrosequencing technology enable a variety of applications on genome research. A method for processing the huge amount of datasets for subsequent analysis is necessary. We have developed a software tool for this purpose. Given input with ultra-deep sequencing sequences and a reference sequence in FASTA format, the program will output with compressed, identity-tallied and codon-aligned sequence alignment. This tool has been successfully applied on a recent study to reveal dynamic HIV-1 escape and large population shifts during drug treatment.**

***Keywords*— 454; Pyrosequencing;Ultra-deep sequencing; Alignment**

I. Introduction

Ultra deep sequencing reads produced by the 454 Life science pyrosequencers provides approaches for detecting rare HIV-1 variants and estimating the viral population within an infected individual [1-2]. For the subsequent analysis and data interpretation, we need to align deep sequencing datasets which is up to 100,000 reads and for protein level analysis, we need to make the nucleotide alignment codon-aligned. These huge number of reads make the traditional multiple sequence alignment very inefficient [3]. The frequent insertion and deletion in single base of 454 reads also make it hard to align well [4-6]. In addition, the codon/position information of the respect reference sequence needs to be taken into consideration for the alignment. We developed a program which present a general method for processing sequences obtained from ultra deep sequencing methodologies where alignment of one variable sequence amplicon re-sequenced 1,000s to 100,000s of times requires alignment, clean up, re-alignment and initial pre-processing to exclude problematic sequences.

II. Methods

Our filtering and alignment process began by compressing all identical sequences into sets such that a single representative sequence was included; each unique sequence was named to indicate how many times it was found in the sample. These sequences were rank-ordered from most to least common and sequentially numbered to give every sequence a unique ID based on its frequency. This compression step was repeated several times through the filtering process.

The reverse complement sequences were generated from the entire set, aligned each sequence in a pair-wise alignment to reference sequence [7], and each sequence has a user-set minimal similarity to reference sequence. Thus, only the forward direction version of each sequence was retained, and sequences that were not a good match in either direction were cut. Identical sequences were compressed again. Then program eliminated sequences that were too short to span our minimum region of interest by user-defined length.

Next, the pair-wise alignments of each sequence relative to the reference strain were created by the alignment tool [7], and made an in-frame multiple sequence pairwise DNA alignment by sequentially combining the pair-wise alignments, keeping the alignment in-frame for the multiple sequence alignment using reference sequence coding information. This part of program combined the strategy we used in the GeneCutter and SynchAlign tool at the Los Alamos HIV database [8].

Then, user can define the interesting region based on the reference sequencing to trim out. Finally, program segregated any sequence containing a frame-shift (single or double base gaps or insertions), or that had simple compensatory changes withing the trimmed region, or that did not fully span the region of interest.

III. Results and Conclusion

We have developed and used the program to analyses 950,000 HIV Env V3 sequences from four patients at three time points that spanned the emergence of CCR-5 antagonist drug resistance [9]. The process steps retained 80-95% of the original 454 reads for subsequent analysis. While the tool was initially designed for HIV, the program provided should be applicable to any sequenced organism.

Acknowledgements: This work was funded by Los Alamos direct research LDRD funding and NIH Center for HIV/AIDS Vaccine Immunology u01 A104785.

E-mail: chienchi@lanl.gov
[1]Los Alamos National Laboratories, Los Alamos, NM, USA.
[2]Santa Fe Institute, Santa Fe, NM, USA;

References

[1] Mitsuya Y. et. al. (2008) Minority human immunodeficiency virus type 1 variants in antiretroviral-naive persons with reverse transcriptase codon 215 revertant mutations. *J. Virol*. **82**, 10747-55
[2] Eriksson N. et. al. (2008) Viral population estimation using pyrosequencing. *PloS Comput. Biol*. **4**, e1000074
[3] Thompson J.D, Higgins D.G., Gibson T.J. (1994) Clustal w: improving the sensitivity of progressive multiple alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*. **22**, 4673–4690
[4] Ronaghi M., Uhlen M., Nyren P. (1998) A Sequencing Method Based on Real-Time Pyrophosphate. *Science* **281**, 363-364
[5] Huse S.M. et. al. (2007) Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol*. **8**, R143
[6] Quinlan A.R. et. al. (2008) Pyrobayes: an improved base caller for SNP discovery in pyrosequences. *Nat. Methods* **5,** 179-81