Quantitative deep sequencing and HIV-1 population dynamics

John Archer1, Andrew Rambaut2 and David L. Robertson1

Faculty of Life Sciences, University of Manchester, Manchester; 1Institute of Evolutionary Biology, University of Edinburgh, Edinburgh

Contact: John.archer@manchester.ac.uk

Abstract

The 454 Life Sciences’ GS-FLX pyrosequencing platform has parallelized the determination of low frequency variants within the viral population. Studies using such data have come to depend on the development of bespoke mapping pipelines. As a consequence of platform and viral specific error models that takes insertions, deletions and dependent insertions into account the frequency of the individual V3 variants (Fig. 4, panel A and B – inset trees). In support of this, we observe that the CXCR4-using viral population that emerge on drug do so from pre-existing low frequency variants present prior to therapy (Figs. 3 and 4). This is particularly evident for patient 18 where a single low frequency (0.04%) CXCR4 - using V3 variant presented at the first time point constitutes the majority of the population in weeks 2 and 16 (Fig. 4, panel A).

Introduction

The 454 Life Sciences’ GS-FLX pyrosequencing platform has parallelized the determination of nucleotide order within genetic material resulting in the ability to generate extremely large data sets [2]. For viral populations the vast majority of short sequence segments produced, termed reads, in conjunction with intrinsic error rates associated with the sequencing platform [3] pose challenging computational problems (Fig. 1). However, these data have the potential to provide a previously unprecedented level of insight into pathogen variation. This is particularly important in the detection of minority variants, for example, those associated with drug resistance [1,4]. Here we apply our framework to the detection of HIV-1 resistance to the CCR5 antagonist vicriviroc, present in the form of CXCR4- using variants, within temporally sampled data. Software is available at: http://www.bioinf.manchester.ac.uk/segminator/

Results

Table: Sample data with read and variant information

<table>
<thead>
<tr>
<th>Sample</th>
<th>Variant</th>
<th>Read</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample A</td>
<td>12 (V3)</td>
<td>500</td>
<td>0.05</td>
</tr>
<tr>
<td>Sample B</td>
<td>10 (V3)</td>
<td>300</td>
<td>0.03</td>
</tr>
</tbody>
</table>

We provide further evidence that the V3 region of HIV-1 harbours a low frequency CCR5-comprising minority population. Here we have demonstrated that the use of a data specific template lowers the level of sequence variation between the template and reads thus reducing data loss during the mapping phase (Fig. 3).

Discussion

Here we have demonstrated that the use of a data specific template lowers the level of sequence variation between the template and reads thus reducing data loss during the mapping phase (Fig. 3).

Platform based insertions were corrected for by comparison to the consensus template reducing the number of reads that are discarded due to frame shift error. We propose the development of platform and viral specific error models that takes insertions, deletions and mismatches into account.

In support of Tsibris et al. [1], we observe that the CXCR4-using viral population that emerge on drug do so from pre-existing low frequency variants present prior to therapy (Figs. 3 and 4). This is particularly evident for patient 18 where a single low frequency (0.04%) CXCR4 - using V3 variant presented at the first time point constitutes the majority of the population in weeks 2 and 16 (Fig. 4, panel A).

The observed extent of the divergence in the phylogenetic trees is misleading without taking into account the frequency of the individual V3 variants (Fig. 4, panel A and B – inset trees). For example at week 0 (patient 18) a single CCR5 – using variant makes up 70.12% of the population. The next two most frequent variants account for 17.54 and 6.52%. The remaining 5.82% of the population at this time point is what constitutes the observed “bushiness” of the phylogeny. It will be important to consider how much if this low frequency variation is acceptable in the context of therapy. For example is it clusters of low frequency variants that important (Fig 4, panel A and B – week 8) or do individual variants need to be considered?

References


Figure 1: Speed versus variation

Figure 2: Mapping, extraction and phenotype determination - (A) Sample id’s and corresponding number of reads obtained in [1]. (B) Number of V3 spanning reads extracted. (C) Reads from each data set were aligned using the charge rule [5] and PSSM [6]. Identical reads were removed, and the number of V3 sequences extracted for co-receptor usage determination relative to the original template were removed.

Figure 3: Phylogenetic analysis - (A) Evidence for the emergence of a majority CCR5-using viral population from low frequency variants for patient 18. Colors (see key) indicate sampling time. The scale bar represents substitutions per site. The inset tree indicates the frequency of each variant. The red circle at week 16 indicated a re-emergence of CCR5-using virus due to failure to conform to treatment.

Figure 4: Discussion - (A) Patient 18inset tree indicates the frequency of each variant. The red circle at week 16 indicated a re-emergence of CCR5-using virus due to failure to conform to treatment. (B) Chosen as the best data set for patient 18. The red circle at week 16 indicated a re-emergence of CCR5-using virus due to failure to conform to treatment.