

CHARACTERIZING NEXT GENERATION SEQUENCING ERROR AND THE CONSEQUENCES FOR THE STUDY OF INTRA-PATIENT VIRAL DIVERSITY

John Archer¹, Greg Baillie², Simon J. Watson², Paul Kellam², Andrew Rambaut³ and David L. Robertson¹

¹Faculty of Life Sciences, University of Manchester, Manchester; ²Wellcome Trust Sanger Institute, Cambridge; ³Institute of Evolutionary Biology, University of Edinburgh, Edinburgh
Contact: john.archer@manchester.ac.uk

Introduction

Second generation sequencing platforms have provided an unprecedented insight into pathogen variation especially in relation to the detection of minority variants [1,2,3,4]. However there is uncertainty surrounding their ability to accurately and consistently detect low frequency variants [5,6].

Previously we presented software for the mapping and alignment of viral read data generated on the 454 Life Sciences platform [1]. Here we extend the underlying framework, by usage of a novel read storage system (Fig. 1) as well as the incorporation of multithreading, so that it is applicable to data generated on the Illumina platform.

We apply the framework to the development of a pipeline for the comparison of variation present within reads, derived from 12 H1N1 infected individuals, generated on both the 454 Life Sciences and Illumina platforms simultaneously. Data from the H1N1 genome as particularly useful for characterizing indel error as within this genome such biologically generated traits are uncommon.

Results

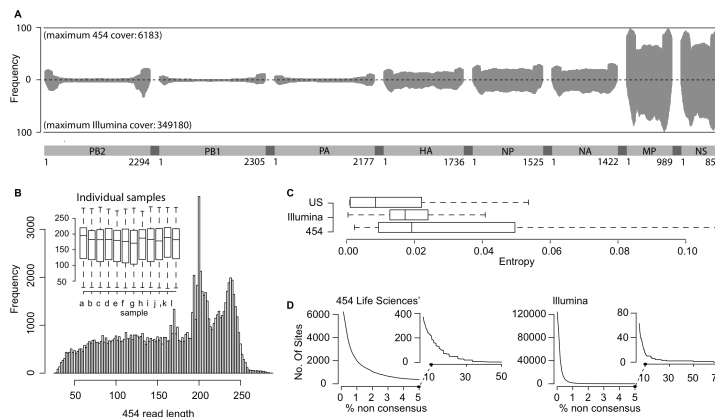


Figure 3 – Mapping and data characterization. (A) Normalized read coverage following sample pooling. Coverage obtained from the 454 platform is displayed above the dotted line while the coverage obtained on the Illumina platform is mirrored below. (B) Read length variation across the 454 data. The inset box and whisker plot shows lengths within each sample. (C) Entropy within data obtained from both platforms as well as that present within the US H1N1 data. (D) Number of variable sites containing non-consensus nucleotides.

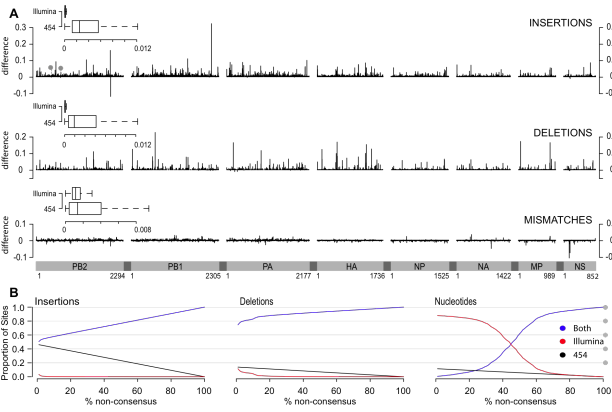
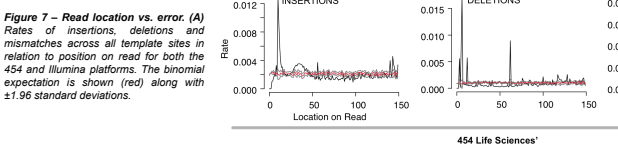


Figure 5 – Cross platform differences. (A) Subtraction of the expected per-site ratios obtained between non-consensus and consensus states for insertions, deletions and mismatches on the Illumina platform from those observed on the 454 platform. The inset box and whisker plots represent the absolute values categorized according to their original polarity. (B) The proportion of sites displaying variation that are in agreement between platforms (blue); threshold levels are depicted on the x-axis. Red and black indicate variable sites that are present on within data from single platform only (see key).



Methods

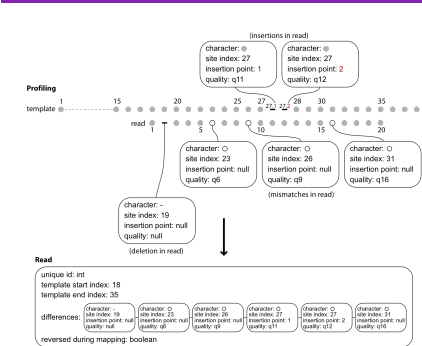


Figure 1 – Read Storage. Profiling of the read in relation to the template (top) and subsequent data structure used to store the read within the assembly.

Figure 2 – The data analysis pipeline. The preprocessing of the template sequence prior to read mapping is outlined (center). The fragments titled “k-mers” are all the unique words within the template sequence. These are stored along with their corresponding locations. On the left hand side all k-mers of equal length, extracted from the read, are shown. The plot indicates the frequency of k-mer matches across the template sequence for a single read. Dashed boxes indicate processing events that take place within the framework.

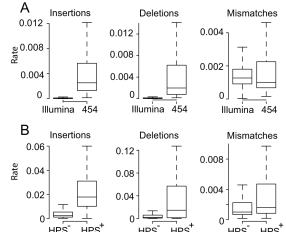
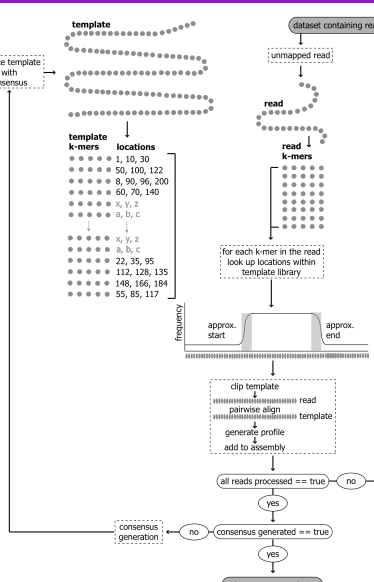


Figure 4 – Cross platform rate variation. (A) Ratios between non-consensus and consensus states for insertions, deletions and mismatches respectively as observed on each individual platform. (B) Ratios between non-consensus and consensus states for insertions, deletions and mismatches observed on the 454 platform within hps+ and hps- categories.

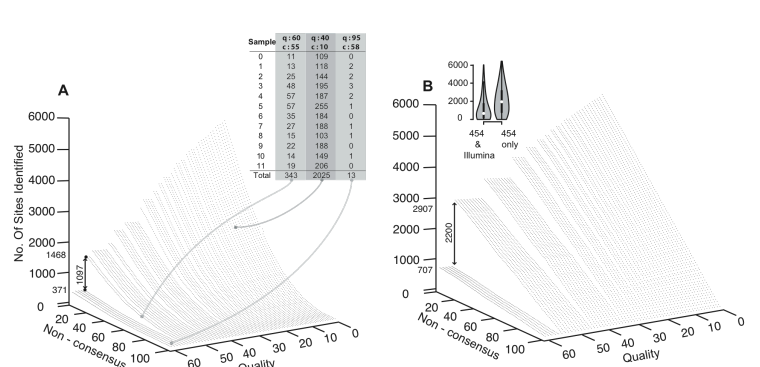


Figure 6 – Mismatches confirmed on both platforms. (A) The number of variable sites, within each sample that have been verified by both platforms at differing quality and non-consensus values. The inset are examples from three points on the topology. The arrow indicates the drop in variable sites identified between the 55th and 56th quality score percentiles. (B) Same as (A) but only data from the 454 Life Sciences platform is used. The inset plot depicts the differences between the numbers of variable sites cross validated on both platforms to those confirmed on the 454 platform.

Discussion

That the median entropies obtained from 12 H1N1 samples, on both platforms, are significantly higher than that of the median entropy obtained from all of the H1N1 sequences sampled within the US (Fig.3D) highlights the importance of quantifying platform introduced variation. In the data from the 454 Life Sciences platform that the ratio of insertions and deletions between hps+ and hps- regions (Fig. 4B) is similar to that previously observed within a genome with very a different evolutionary strategy [6] highlights the consistent nature of this error.

When the topology of the relationship between the percent of non-consensus nucleotides at a site, the quality of these nucleotides and the number of variable sites identified within each sample is explored, the only outstanding feature lies between the 55th and 56th quality score percentile (Fig. 6A). Here the decrease in the number of cross platform validated sites is a strong indication that the quality score of the base is the most reliable factor for identifying platform error. Without cross platform validation the number of variable sites identified within each sample is consistently higher (Fig. 6B). This will make identifying low level variation within viral populations difficult using a single platform.

Acknowledgments

We would like to thank BBSRC for funding.



References

- Archer J, et al. The evolutionary analysis of emerging low frequency HIV-1 CXCR4 using variants through time—an ultra-deep approach. PLoS Comput Biol 8: e1001022.
- Palmer S, et al. (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.
- Rozera G, et al. (2005) Massively parallel pyrosequencing highlights minority variants in the HIV-1 env quasipieces deriving from lymphomonocytic sub-populations. Retrovirology 6: 15.
- Medlin D, et al. (2008) Microbiology in the post-genomic era. Nat Rev Microbiol 6: 419-430.
- Brockman W, et al. (2008) Quality scores and SNP detection in sequencing-by-synthesis systems. Genome Res 18: 763-770.
- Wang C, et al. (2007) Characterization of mutation spectra with ultra-deep pyrosequencing: application to HIV-1 drug resistance. Genome Res 17: 1195-1201.

