Error profiles for next generation sequencing technologies

Melanie Schirmer[1], Linda D'Amore[2], Neil Hall[1], Christopher Quince[2]✉
[1]University of Glasgow, United Kingdom
[2]University of Liverpool, United Kingdom

Motivation and Objectives

Next generation sequencing has revolutionized genome research and marked the start of a new era. These new technologies present us with unprecedented amounts of data - but with this sequencing data come errors that are not only platform specific but also depend on the library preparation method and the type of sequencing (i.e. amplicon or metagenome). Illumina's sequencing platforms are currently among the most utilized platforms as they are able to generate millions of reads at relatively low cost – but Illumina error profiles are still poorly understood. A better knowledge of the error profiles is essential for sequence analysis and vital in order to draw valid conclusions. It has been reported that the major source of errors for Illumina are substitution-type miscalls (Archer *et al.*, 2012). We developed a program that enables us to infer error profiles based on sequencing data from mock communities. This allows us to study and compare different errors and biases introduced by different sequencing machines, different library preparation methods as well as different types of sequencing. Here, we present the metagenome error profiles for a mock community that was sequenced on the Genome Analyzer (GA) II for the standard Illumina library preparation method (TruSeq). Being able to infer error profiles for individual sequencing runs has the potential to greatly improve our ability to correct errors and thus enhance further sequencing analysis.

Methods

For our error profiles we used a diverse mock community that was constructed by combining even amounts of purified genomic DNAs (Shakya *et al.,* 2013). The mock community consists of 49 bacterial genomes and 10 archaeal genomes covering most phyla and the community also contains closely related species and strain pairs. We sequenced a sample of the mock community on the GA II. The libraries for the sample were prepared with the standard Illumina library preparation method (TruSeq) with a starting amount of 500ng of DNA. This yielded about 6 million forward and reverse reads of 101bp.

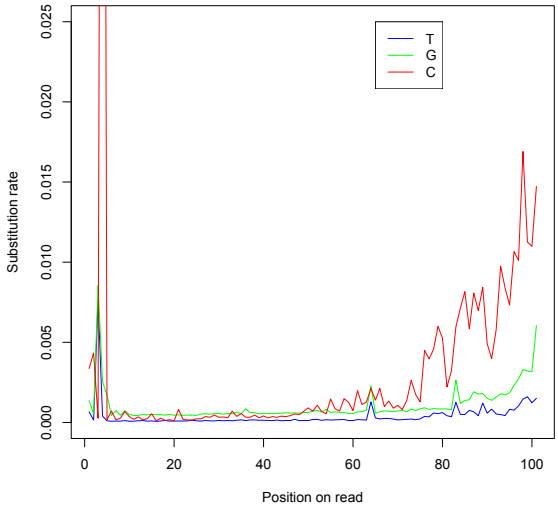
First, we aligned the reads with BWA (Li and Durbin, 2009) against the 59 reference genomes. Then we converted the files to SAM format and generated the MD tag with samtools (Li *et al.*, 2009). Based on the resulting files our program infers position and nucleotide specific substitution rates. Whenever a substitution is encountered, we identify the reference nucleotide based on the MD tag and the substituting nucleotide on the read is determined based on the extended CIGAR string. The output of our program consists of four 4x101 matrices (one for each possible "original" nucleotide) for the set of forward and reverse reads, respectively, in which we store the number of observed substitution types for each position of the read. We then normalize these matrices as follows: We count the number of occurrences of, for example, A on the read for each position, add the number of detected substitutions from A to T, G and C, respectively, and subtract the number of substitutions from T, G and C, respectively, to A at this position. This accounts for errors and reflects the true number of A's. In addition, our program computes the overall insertion and deletion rate.

To verify our algorithm we extended our read simulation program (Schirmer *et al.*, 2012) to generate reads based on error profiles of the above described format. We simulated one million forward and reverse reads based on the error profiles inferred from the GA II run. The error profiles, inferred from the simulated reads, concurred with the original error profiles used to simulate the reads and thus validates the algorithm.

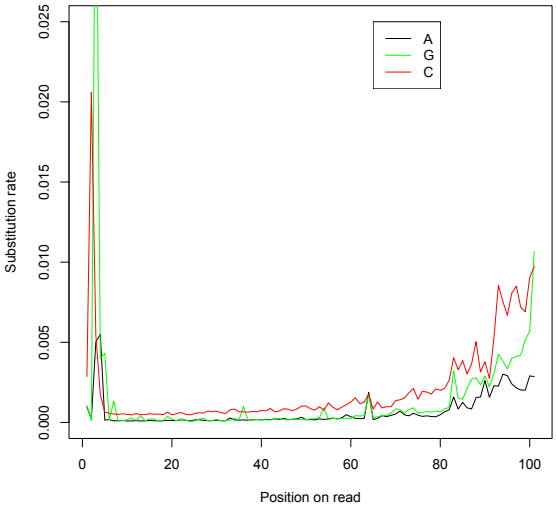
Results and Discussion

The GA II error profiles show a strong increase in the number of substitutions towards the end of the read. The average substitution rate for the forward reads is $\approx$0.004, where several spikes were observed across the first 10bp as well as an increase in substitutions starting from the middle of the read towards the end of the read. For

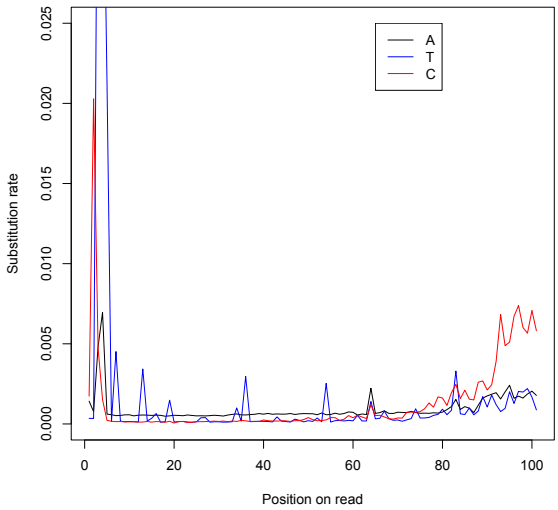
the reverse reads the average substitution rate is ≈0.012. The start of the reverse reads also shows several spikes in the error profile but less compared to the forward reads. Though smaller spikes were observed across the whole read length. The substitution rate starts to increase after the first third of the read and is overall significantly higher towards the end of the read compared to the forward reads. We observed the highest substitution rates for A and the lowest substitution rates for G for both forward and reverse reads (disregarding the first 10bp). Subsequently we examined the frequencies of the nucleotides for each position across the reads to test for possible



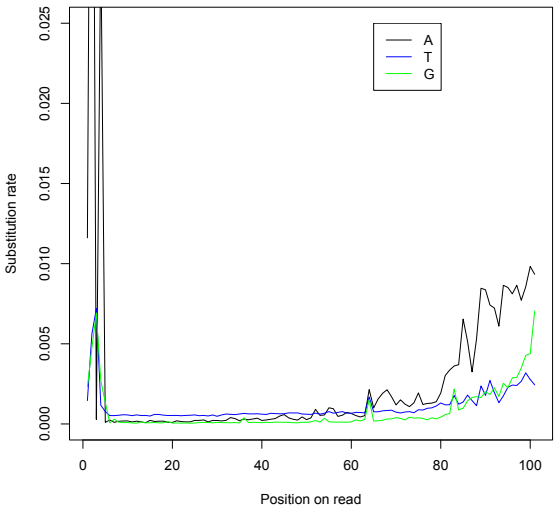
(a) R1 reads: original nucleotide A



(b) R1 reads: original nucleotide T



(c) R1 reads: original nucleotide G



(d) R1 reads: original nucleotide C

Figure 1: Error profile for forward reads: The x-axis indicates the position on the read and the y-axis the substitution rate (# of observed substitution/# of occurrences of the "original" nucleotide). Each subfigure represents one of the four possible original nucleotides for which different types of substitutions are indicated by different colors.

artifacts, as these could explain the spikes in the error profile at the read-start. For metagenomic data sets we expect a uniform frequency distribution across the reads for all nucleotides. Here, we identified fluctuations within the first 10bp that sufficiently account for the increased error rates across these positions. Separating the error profiles according to the different substitution types presented further insights. Figure 1 shows that - independent of the original nucleotide – a substitution with C is the most common error towards the end of the read. If the original nucleotide is a C a substitution with A is the most common error. Inferring error profiles for different sequencing machines, library preparation methods and sequencing types has great potential for error correction. It also enables us to infer error profiles for individual sequencing runs by including a mock community (e.g. instead of PhiX). We will extend our research to different sequencers, more library preparation methods and different types of sequencing to identify differences and similarities in the error profiles and as a possible guideline for experimental design.

Acknowledgements

This research is part of a project funded by the Technology Strategy Board. M.S. is supported by Unilever R&D Port Sunlight, Bebington, United Kingdom.

References

Archer J, Baillie G, Watson SJ, Kellam P, Rambaut A et al. (2012) Analysis of high-depth sequence data for studying viral diversity: a comparison of next generation sequencing platforms using Segminator II. *BMC Bioinformatics* **13**(1), 47.

Li H and Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics* **25**, 1754-60.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J et al. (2009). The sequence alignment/map format and samtools. *Bioinformatics*,**25**(16), 2078–2079

Schirmer M, Sloan WT and Quince C (2012) Benchmarking of viral haplotype reconstruction programmes: An overview of the capacities and limitations of currently available programmes. *Brief. Bioinfo.* (online - ahead of print]). doi: 10.1093/bib/bbs081

Shakya M, Quince C, Campbell J, Yang ZK, Schadt C *et al.* (2013) Comparative meta-genomic and rRNA microbial diversity characterization using archaeal and bacterial synthetic communities. *Environmental microbiology*. doi:10.1111/1462-2920.12086