Published in final edited form as:

Nat Biotechnol. 2013 July ; 31(7): 592–594. doi:10.1038/nbt.2616.

Read and assembly metrics do not affect clinical utility of rapid whole genome sequencing platforms for mapping outbreaks

SR Harris¹, ME Török^{2,3,4}, EJP Cartwright^{2,3,4}, M Quail¹, SJ Peacock^{1,2,3,4}, and J Parkhill¹

¹Wellcome Trust Sanger Institute, Hinxton, United Kingdom

²University of Cambridge, Cambridge, United Kingdom

³Cambridge University Hospitals NHS Foundation Trust, Cambridge, United Kingdom

⁴Cambridge Health Protection Agency Microbiology Laboratory, Cambridge, United Kingdom

In their paper “Performance comparison of bench-top high-throughput sequencing platforms”¹, Loman and colleagues provide a detailed comparison of the metrics associated with three different platforms in the assembly of a single genome. Information was given on read-level metrics such as length, accuracy and alignment, and on assembly-level metrics such as contig N50 and gap number. The results were discussed in the context of the utility of whole-genome sequencing for public health microbiology.

We believe, however, that one of the primary uses for sequencing in clinical microbiology (at least initially) will be in the detection of pathogen transmission events and the investigation of outbreaks of infectious diseases^{2–7}. In this context, what is important is less the *de novo* assembly and analysis of single genomes, but more the identification of discriminatory single nucleotide polymorphisms (SNPs) in the shared core genome of multiple isolates, and their use to rule patients into, or out of, an outbreak. Therefore, we compared the utility of the two most cost-effective rapid platforms (Illumina MiSeq and Ion Torrent PGM) for resolving relationships amongst strains of methicillin-resistant *Staphylococcus aureus* (MRSA) isolated during a recently reported nosocomial outbreak⁴.

In brief, an outbreak investigation was initiated when the infection control team at the Cambridge University Hospitals NHS Foundation Trust noted that three infants on the Special Care Baby Unit were MRSA-positive on screening swabs. A retrospective look-back over six months identified a total of 17 MRSA-positive patients in three distinct clusters. We extracted and sequenced DNA from 24 MRSA isolates – 15 from patients suspected to be involved in the outbreak (corresponding to P1 to P15 in Figure 1A and 1B of the paper)⁴ and nine from control patients elsewhere in the hospital. The isolates were sequenced on HiSeq, MiSeq and Ion Torrent platforms with read-lengths of 76, 150 and ~150bp, to an average coverage of 204.81-, 77.74- and 71.98-fold, respectively. We employed a simple processing pipeline for each data set, using the read-mapping and SNP calling parameters described in the methods section. All three platforms clearly discriminated outbreak from non-outbreak MRSA strains (with an average of 13,154 SNP differences between outbreak and non-outbreak isolates called by MiSeq and 13,297 differences called by the PGM). All platforms also clearly discriminated amongst the near-identical outbreak strains, identifying a total of 23 SNPs amongst the 15 strains. The initial MiSeq data called one extra SNP in one

outbreak strain, which was subsequently identified as a read mapping error. However, this did not affect the identified relationships amongst the strains. In every case, the HiSeq and MiSeq results were identical. The larger number of SNPs seen in the Ion Torrent data for the outbreak to non-outbreak comparison supports previous data which showed that Ion Torrent had a higher false-positive SNP rate when comparing more distantly-related organisms⁸.

In conclusion, we found that despite the different read metrics and error profiles identified previously, both rapid-turnaround platforms (MiSeq and Ion Torrent) produced comparable clinically actionable data when applied to a real-world outbreak, in terms of discriminating out-break from non-outbreak strains, and identifying the small number of discriminatory SNPs within the outbreak. The error profiles of individual reads and the differences in quality of assembly do not have any affect on the utility of the output when sufficient read-coverage is obtained in mapping-based approaches. Decisions on which platform to use for clinical microbiology should therefore be based on other considerations, such as ease-of-use, complete pipeline turnaround time (including DNA and library preparation, not just machine run time)⁹, and per-base and per-run costs. Although this comparison has concentrated on read-mapping and SNP calling in outbreak situations, small indels can contain useful phylogenetic information, and *de-novo* assembled genomes clearly have other uses in clinical microbiology, including generation of accurate reference sequences, and identification of novel genes in outbreak strains. For these uses, the performance of the different platforms does vary, and comparative analyses such as that produced by Loman *et al.*¹ should be taken into account when choosing an appropriate platform.

Methods

Reads were mapped against the chromosome of an EMRSA-15 reference, HO 50960412 (accession number HE681097), using SMALT version 0.6.4 (<http://www.sanger.ac.uk/resources/software/smalt/>). Reads that aligned equally well to two or more locations in the reference sequence were not mapped. For HiSeq and MiSeq data, paired reads were mapped with an expected insert size of between 50 and 1000bp, while for Ion Torrent data reads were mapped as single ended. In all cases, reads containing indels were realigned using the Genome Analysis Toolkit¹⁰ RealignerTargetCreator and IndelRealigner modules. Variants were called using a combination of samtools¹¹, mpileup and bcftools. For HiSeq and Miseq data default options were used, while for Ion Torrent data, settings were chosen as used in the Ion Torrent Variant Caller Plugin: for samtools mpileup, a minimum base quality of 7 was used instead of the default of 13, the coefficient of homopolymer errors was reduced to 50 from the default of 100, the minimum number of reads required for an indel candidate was increased to 4 from a default of one and the phred-scaled gap opening and extension probabilities were reduced to 10 and 17 from the defaults of 40 and 20, respectively. Pseudosequences for each isolate were then created by filtering all bases using the following filters: (1) base must be covered by at least 4 reads, of which at least two must be on each strand; and (2) bases must have a quality score greater than 50 and a mapping quality greater than 30, and the majority base must be present in at least 75% of reads on each strand. Phylogenetic trees of variable sites were reconstructed with RAxML¹², and SNPs reconstructed onto the trees using parsimony. Raw reads for each platform are available

from the European Nucleotide Archive under project accession number ERP000985. Accessions for individual samples are shown in Supplementary Table 1.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

1. Loman NJ, et al. Performance comparison of benchtop high-throughput sequencing platforms. *Nat Biotechnol.* 2012; 30:434–439. doi:nbt.2198 [pii] 10.1038/nbt.2198. [PubMed: 22522955]
2. Eyre DW, et al. A pilot study of rapid benchtop sequencing of *Staphylococcus aureus* and *Clostridium difficile* for outbreak detection and surveillance. *BMJ Open.* 2012; 2 doi:bmjopen-2012-001124 [pii] 10.1136/bmjopen-2012-001124.
3. Gardy JL, et al. Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. *N Engl J Med.* 2011; 364:730–739. doi:10.1056/NEJMoa1003176. [PubMed: 21345102]
4. Harris SR, et al. Whole-genome sequencing for analysis of an outbreak of methicillin-resistant *Staphylococcus aureus*: a descriptive study. *Lancet Infect Dis.* 2012 doi:S1473-3099(12)70268-2 [pii] 10.1016/S1473-3099(12)70268-2. [PubMed: 23158674]
5. Koser CU, et al. Routine use of microbial whole genome sequencing in diagnostic and public health microbiology. *PLoS Pathog.* 2012; 8:e1002824. doi:10.1371/journal.ppat.1002824 PPATHOGENS-D-12-00717 [pii]. [PubMed: 22876174]
6. Koser CU, et al. Rapid whole-genome sequencing for investigation of a neonatal MRSA outbreak. *N Engl J Med.* 2012; 366:2267–2275. doi:10.1056/NEJMoa1109910. [PubMed: 22693998]
7. Snitkin ES, et al. Tracking a hospital outbreak of carbapenem-resistant *Klebsiella pneumoniae* with whole-genome sequencing. *Sci Transl Med.* 2012; 4:148ra116. doi:4/148/148ra116 [pii] 10.1126/scitranslmed.3004129.
8. Quail MA, et al. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics.* 2012; 13:341. doi: 1471-2164-13-341 [pii] 10.1186/1471-2164-13-341. [PubMed: 22827831]
9. Torok ME, et al. Rapid whole-genome sequencing for the investigation of a suspected tuberculosis outbreak. *J Clin Microbiol.* 2012 doi:JCM.02279-12 [pii] 10.1128/JCM.02279-12.
10. McKenna A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 2010; 20:1297–1303. doi:gr.107524.110 [pii] 10.1101/gr.107524.110. [PubMed: 20644199]
11. Li H, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics.* 2009; 25:2078–2079. doi:btp352 [pii] 10.1093/bioinformatics/btp352. [PubMed: 19505943]
12. Stamatakis A. RAXML-VI-HP: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics.* 2006; 22:2688–2690. doi:bt1446 [pii] 10.1093/bioinformatics/bt1446. [PubMed: 16928733]