nature biotechnology

Performance comparison of benchtop high-throughput sequencing platforms

Nicholas J Loman[1], Raju V Misra[2], Timothy J Dallman[2], Chrystala Constantinidou[1], Saheer E Gharbia[2], John Wain[2,3] & Mark J Pallen[1]

Three benchtop high-throughput sequencing instruments are now available. The 454 GS Junior (Roche), MiSeq (Illumina) and Ion Torrent PGM (Life Technologies) are laser-printer sized and offer modest set-up and running costs. Each instrument can generate data required for a draft bacterial genome sequence in days, making them attractive for identifying and characterizing pathogens in the clinical setting. We compared the performance of these instruments by sequencing an isolate of *Escherichia coli* O104:H4, which caused an outbreak of food poisoning in Germany in 2011. The MiSeq had the highest throughput per run (1.6 Gb/run, 60 Mb/h) and lowest error rates. The 454 GS Junior generated the longest reads (up to 600 bases) and most contiguous assemblies but had the lowest throughput (70 Mb/run, 9 Mb/h). Run in 100-bp mode, the Ion Torrent PGM had the highest throughput (80–100 Mb/h). Unlike the MiSeq, the Ion Torrent PGM and 454 GS Junior both produced homopolymer-associated indel errors (1.5 and 0.38 errors per 100 bases, respectively).

Over the past decade and a half, genome sequencing has transformed almost every corner of the biomedical sciences, including the study of bacterial pathogens[1]. In the last five years, high-throughput (or 'next-generation') sequencing technologies have delivered a step change in our ability to sequence genomes, whether human or bacterial[2,3]. Since arriving in the marketplace, these technologies have undergone sustained technical improvement, which, twinned with lively competition between alternative platforms, has placed genome sequencing in a state of permanent revolution.

Although high-throughput sequencing has seen extensive use in bacteriology, for example, in the genomic epidemiology of bacterial pathogens[4], until recently sequencing platforms were tailored chiefly toward large-scale applications, focused on the race to the '$1,000 human genome', with footprints, workflows, reagent costs and run times poorly matched to the needs of small laboratories studying small genomes. However, three different benchtop high-throughput sequencing instruments are currently available, all capable of sequencing bacterial genomes in a matter of days (**Table 1**).

[1]Centre for Systems Biology, University of Birmingham, Birmingham, UK. [2]Health Protection Agency, London, UK. [3]School of Medicine, University of East Anglia, Norwich, UK. Correspondence should be addressed to M.J.P. (m.pallen@bham.ac.uk) or J.W. (j.wain@uea.ac.uk).

Received 19 December 2011; accepted 30 March 2012; published online 22 April 2012; corrected online 23 April 2012 (details online); doi:10.1038/nbt.2198

The 454 GS Junior from Roche was released in early 2010 and is a smaller, lower-throughput version of the 454 GS FLX machine, exploiting similar emulsion PCR and pyrosequencing approaches, but with lower set-up and running costs. The Ion Torrent Personal Genome Machine (PGM) was launched in early 2011 (ref. 5). Like the 454 GS Junior, this technology exploits emulsion PCR. It also incorporates a sequencing-by-synthesis approach, but uses native dNTP chemistry and relies on a modified silicon chip to detect hydrogen ions released during base incorporation by DNA polymerase (making it the first 'post-light' sequencing instrument). The Illumina MiSeq was announced in January 2011 and began to ship to customers in the fourth quarter of 2011. The MiSeq is based on the existing Solexa sequencing-by-synthesis chemistry[6] but has dramatically reduced run times compared to the Illumina HiSeq (fastest run 4 h versus 1.5 d for 36-cycle sequencing or 16 h versus 8.5 d for 200-cycle sequencing), made possible by a smaller flow cell, reduced imaging time and faster microfluidics.

We wished to compare the performance of these three sequencing platforms by analyzing data with commonly used assembly and analysis pipelines. We therefore benchmarked these platforms by using them to sequence the genome of an isolate from the recent outbreak of food-borne illness caused by Shiga-toxin-producing *E. coli* O104:H4, which struck Germany between May and July 2011. This outbreak was responsible for >4,000 infections and more than 40 deaths[7]. Previous whole-genome sequencing efforts applied to isolates from the outbreak yielded novel diagnostic reagents and provided important clues as to the nature, origins and evolution of the outbreak strain[8–12]. These efforts also demonstrated the utility of an 'open-source' approach to outbreak genomics that included rapid sequencing, a liberal approach to data release and use of crowdsourcing[10]. Although all infections during the outbreak were acquired in Germany, travelers took their infections back to other countries in North America and Europe, including the United Kingdom[7]. Here, we have focused on a single *E. coli* isolate of serotype O104 from the United Kingdom, which was epidemiologically linked to the German outbreak.

RESULTS

Creation of reference assembly

To permit comparisons of benchtop sequencing data, we generated a reference assembly for *E. coli* O104:H4 strain 280 (UK Health Protection Agency's materials identifier H112160280) using established high-throughput sequencing platforms. This isolate was recovered from a female traveler returning from Germany who had

Table 1 Price comparison of benchtop instruments and sequencing runs

Platform	List price	Approximate cost per run	Minimum throughput (read length)	Run time	Cost/Mb	Mb/h
454 GS Junior	$108,000	$1,100	35 Mb (400 bases)	8 h	$31	4.4
Ion Torrent PGM						
(314 chip)	$80,490[a,b]	$225[c]	10 Mb (100 bases)	3 h	$22.5	3.3
(316 chip)		$425	100 Mb[d] (100 bases)	3 h	$4.25	33.3
(318 chip)		$625	1,000 Mb (100 bases)	3 h	$0.63	333.3
MiSeq	$125,000	$750	1,500 Mb (2 × 150 bases)	27 h	$0.5	55.5

Note pricing may vary between countries and/or sales territories. Instrument prices do not include service contracts. Sample prices do not include the cost of generating the initial fragmented genomic DNA library with adaptors (an additional cost of between $75–200 depending on method used). Cost per megabase assumes one sample and one sample sequencing kit per run. Unless stated, pricing information is from the online supplement of ref. 3.
[a]Ion Torrent PGM pricing from Invitrogen US territory website (http://www.invitrogen.com/, accessed 21 February 2012). [b]Price includes Ion Torrent PGM, server, OneTouch and OneTouch ES sample automation systems. [c]Ion Torrent PGM prices include chip and sample preparation kit. [d]Configuration used in this study.

developed hemolytic uremic syndrome and thrombotic thrombocytopenic purpura. The isolate was confirmed as typical of the outbreak strain (ST678, *stx-2* positive and intimin negative)[12].

We used the Roche 454 GS FLX+ system to generate very long fragment reads (modal read length, 812 bases; maximum read length, 1,170 bases). Additionally, Roche 454 GS FLX was used to sequence an 8-kb insert paired-end library using Titanium chemistry. The reads were assembled into contigs, which were scaffolded to produce a draft reference assembly. Mean coverage depth for the assembly was 32-fold.

The use of abundant long reads and long-insert, paired-end information resulted in a very high-quality draft genome assembly consisting of three scaffolds. Of the bases in the assembly, 99.42% are Q64 bases (representing accuracy of one miscall around every 2.5 M bases), 99.54% are Q40 (one miscall every 10,000 bases) or higher. Bases with a quality score <40 were masked with a lower-case letter and excluded from further analysis. The largest scaffold corresponded to the chromosome (5,340,015 bp), the two smaller scaffolds corresponded to two large plasmids (pESBL and pAA). The 1.5-kb plasmid sequence was present in a single contig. Although each scaffold represented a single circular replicon, 153 gaps remained within the scaffolds. These gaps represent repetitive regions longer than the mean read length and shorter than the paired-end insert library, which cannot be resolved by this sequencing strategy.

Characteristics of reads from benchtop sequencers

Genome depth, evenness of coverage, read length and read quality are the four major factors that determine the ability to reconstruct genome sequences from sequence data. There were large differences in the number, predicted quality and length of reads obtained from the three platforms (**Table 2** and **Fig. 1**). The 454 GS Junior produced the longest reads, with a mean length of 522 bases, but had the lowest throughput of the three instruments (70–71 megabases). Ion Torrent PGM runs generated over four times the throughput of 454 GS Junior but generated the shortest reads (mean 121 bases). The MiSeq run produced the greatest throughput (1.6 gigabases) with reads slightly longer than those by Ion Torrent PGM, permitting the multiplexing of seven *E. coli* strains on a single run when targeting 40-fold coverage of each genome. MiSeq reads were paired-end, that is, fragments were sequenced in both directions. Across the reference chromosome, coverage was generally even for all technologies. However, in the MiSeq data we saw a peak associated with the Shiga toxin–producing phage. A similar, but smaller, peak was detectable in the Ion Torrent PGM data (**Supplementary Fig. 1**). These peaks may be explained by the occurrence of phage lysis in the cultures used to prepare the DNA for sequencing. Differences in relative coverage levels were also seen in the pESBL and pAA plasmids between instruments, possibly due to the use of different DNA shearing techniques in library preparation.

Because each manufacturer uses a unique software implementation to generate base-quality score predictions, direct comparison of these scores between platforms is difficult. We recalibrated quality scores for each instrument by first aligning reads to the reference genome. By observing the counts of matched and mismatched bases in each aligned read, a new quality score can be calculated, called alignment quality. We used a scoring system, previously published[13], which takes into account substitutions, insertions and deletions. Mismatches resulting in deletions are assigned randomly to the position of one of the adjacent bases in the read. Alignment quality scores measured in this way generally had good agreement with predicted scores, with the Ion Torrent PGM generally underestimating quality scores and the other instruments slightly overestimating them (**Supplementary Fig. 2**). The MiSeq produced the highest quality reads, owing to a low substitution error rate (0.1 substitutions per 100 bases) and the near absence of indel errors compared to the other platforms. The Ion Torrent PGM showed a steadily decreasing accuracy across the read to the 100th base. However, soft clipping of low-quality read ends by the BWA alignment software serves to make the accuracy appear to increase after this point as mismatches are not counted in soft-clipped (unaligned) parts of the read[14].

Comparison of the frequency of indels through alignment to the reference demonstrated that Ion Torrent PGM reads had 1.5 indels per 100 bases (1.72 indels per read). The 454 GS Junior had

Table 2 Run and alignment metrics for benchtop sequencers

Platform (run)	Number of reads	Total bases	Modal read length in bases	Mean read length in bases (s.d.)	Alignment coverage		
					Chromosome	Large plasmids	Reads aligned (%)
454 GS Junior (1)	135,992	70,999,968	518	522 (46)	11.50	5.66	99
454 GS Junior (2)	137,528	71,710,564	516	521 (47)	11.54	5.39	99
Ion Torrent PGM (1)	2,483,868	303,579,279	123	122 (11)	46.60	53.33	90
Ion Torrent (2)	2,154,577	260,017,346	123	120 (16)	39.33	43.80	89
MiSeq (1)	11,708,156	1,652,529,000	150	141 (22)	–	–	–
MiSeq (1) demultiplexed strain 280	1,766,516	250,356,566	150	141 (21)	22.11	625.46	99

Metrics for each sequencing run are shown as well as results of alignment against the reference sequence. Depth of coverage for the chromosome and two large plasmids (pESBL and pAA) are shown with the percentage of reads that align. For the MiSeq run, the sequence metrics are shown for the entire run as well as the results of de-multiplexing *E. coli* O104:H4 strain 280. Alignment statistics for the entire run are not shown as two strains sequenced were of *E. coli* isolates unrelated to the outbreak strain.

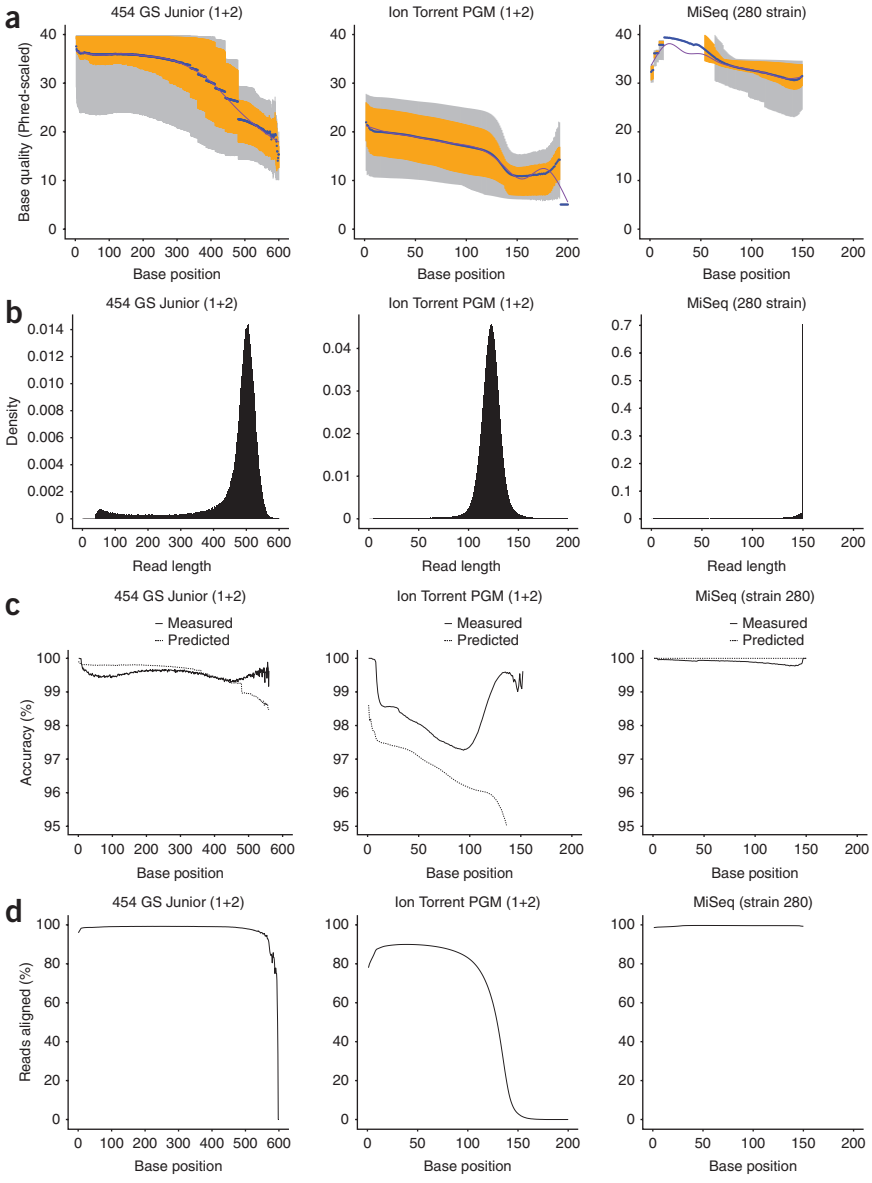

Figure 1 Evaluation of read length and quality from benchtop sequencers. (**a**) Box plots generated by the *qrqc* software package showing the predicted per-base quality score for combined sequencing runs for each benchtop instrument at each read position created by the *qrqc* package. Gray shaded bands indicate the 10% and 90% quantiles, orange shaded bands indicate the lower and upper quartiles, the blue dot is the median. A purple smooth curve is fit through the distributions[16]. Quality scores are given as Phred-scaled quality values where $Q = -10 \log_{10} P$ (P is the probability of the base call being correct). (**b**) Histograms showing read lengths produced by each instrument. (**c**) Comparison of the predicted and measured accuracy for each benchtop sequencer. Predicted accuracy is determined by multiplying the number of alignments of bases of each quality score by the probability of an incorrect base call ($10^{-Q/10}$). The sum of these values is divided by the number of aligned bases to give a measurement of accuracy. (**d**) The percentage of reads aligned at each read position.

0.38 indels per 100 bases (1.74 indels per read). In contrast, indels were detected very infrequently in MiSeq data with <0.001 indels per 100 bases. These results were confirmed by alignment to two other reference genomes sequenced with other sequencing technologies (**Supplementary Tables 1–3**). As with 454 sequencing, the major source of indels in Ion Torrent PGM data are runs of identical bases (homopolymers). Comparison of homopolymer accuracy between Ion Torrent PGM and 454 GS Junior demonstrated that Ion Torrent PGM was less accurate when calling homopolymers of any length (**Supplementary Fig. 3**). The dominant source of error was deletions, with accuracy rates as low as 60% for homopolymers 6 bases or longer.

Comparison of *de novo* assemblies

The use of high-throughput sequencing for the discovery of differences in gene content and arrangement relies on the generation of accurate *de novo* assemblies. We compared draft, *de novo* assemblies from benchtop instruments using a variety of metrics. Assembly metrics such as total assembly size and N50 (a statistic for describing the distribution of contig lengths in an assembly) (ref. 15) give a guide to assembly completeness or fragmentation but not accuracy. An ideal assembly produces a single accurate contig for each replicon, but this is rarely possible owing to the presence of long repeat sequences. When comparing assemblies produced by benchtop *de novo* sequencers, we saw two major groupings of assembly quality. Heavily fragmented assemblies were obtained with Ion Torrent PGM data (single runs or combined), 454 GS Junior (single runs) and MiSeq (**Fig. 2** and **Supplementary Table 4**). Less fragmented assemblies were obtained when reads from two 454 GS Junior runs were combined to increase depth of coverage (98 contigs versus 150 contigs using the assembler program MIRA) and when paired-end information was used to scaffold contigs generated from the MiSeq data (200 scaffolds versus 311 contigs using CLC Assembly Cell). Scaffolds produced by assembly of Illumina MiSeq paired-end data gave output containing runs of ambiguous 'N' bases between 1 and 352 bases in length (81 such runs in CLC Assembly Cell output, 153 runs in Velvet output).

The number of contigs that can be mapped unambiguously to the reference gives a measure of reference genome coverage. Differences in reference genome coverage were seen when comparing assemblies from each platform (**Supplementary Table 4**). None of the assemblies generated aligned unambiguously to 100% of the reference. Contigs obtained from the 454 GS Junior data aligned to the largest proportion of the reference, with 3.72% of the reference unmapped, compared to 4.6% for Ion Torrent PGM and 3.95% for MiSeq. The MIRA assembler produced the assemblies with the highest coverage of the reference genome for each data type.

The Ion Torrent PGM assemblies had large numbers of gaps (**Fig. 3**), compared to assemblies obtained from 454 GS Junior and MiSeq data. Increasing sequence coverage by combining assemblies from the two Ion Torrent PGM runs reduced the number of gaps in

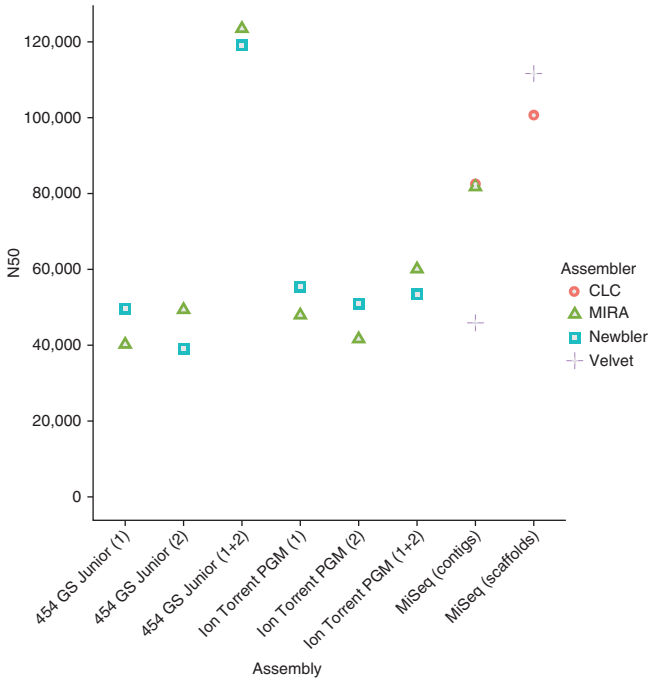

Figure 2 N50 contig sizes from assemblies generated from sequence data for each sequencing platform. A selection of popular genome assemblers have been used. The N50 contig size is calculated using the total genome length of the *E. coli* strain 280 reference sequence, rather than the sum total of contig lengths.

the assembly. With the MIRA assembly, the combined Ion Torrent PGM run showed 38% fewer gaps than Ion Torrent PGM run number two alone. However, many miscalls in long homopolymeric tracts remained, so that in assemblies produced from combining both Ion Torrent PGM data sets, large numbers of contigs were disrupted either by contig breaks or apparent frameshifts. Of the 2,017 gaps seen in the combined Ion Torrent PGM assembly produced by the Newbler assembler (1,811 gaps for MIRA), around a third to a quarter were due to gaps associated with ends of contig or unmapped sequence, the rest being associated with homopolymeric tracts. Although the likelihood of a gap increases with the length of the homopolymer, the number of very short homopolymers (2–3 residues) resulting in assembly gaps was significantly higher for this platform than for the 454 GS Junior. Manual inspection of assembly alignments revealed that many of the indels associated with short homopolymeric tracts demonstrated strand bias, with the correct call predominantly associated with either forward or reverse reads and the erroneous sequences associated with the opposite strand (**Supplementary Fig. 4**). Although problems with homopolymers are known to result from flow cell–based chemistries, it is unclear why this strand bias should occur with Ion Torrent technology. However, scrutiny of other public data sets from this instrument (http://mira-assembler.sourceforge.net/docs/DefinitiveGuideToMIRA.html) suggests it is a pervasive problem.

Benchtop assemblies and public health microbiology

A key test for a genome-sequencing technology is whether it can deliver trustworthy new insights into the biology of the organism under scrutiny. We therefore evaluated how *de novo* assemblies generated from data from each platform performed in reporting features of biological interest in the outbreak strain. For some features, all analyses did well—for example, all documented the presence and accurate full-length sequence of the genes encoding the Shiga toxin type-2 subunits. However, at the other extreme, in some instances, all instruments did badly—for instance, in all assemblies the two larger plasmids were broken into multiple contigs, which could not be readily assigned to chromosome or plasmid without alignment to the reference genome.

We used 31 protein sequences linked to pathogen biology as queries in translated BLAST searches of the assemblies obtained from the benchtop sequencing platforms (**Table 3**). No assembly contained a full set of full-length sequences. The best MiSeq assembly captured 29/31 full-length sequences, the best 454 GS Junior assembly found 26 and the best Ion Torrent PGM assembly found 23. Perhaps the most challenging targets in the survey were the four serine protease autotransporters encoded in the genome of the outbreak strain. None of the platforms managed to recover all four genes as full-length fragments. This is because these genes contain multiple domains and some domains exist as multiple copies in the genome, which are assembled into repeat consensus contigs that cannot be unambiguously placed in the genome. Notably, the choice of assembler affected the ability to detect certain genes; for example, CLC Assembly Cell and MIRA run with Illumina MiSeq were able to reconstruct all four of the aggregative adhesion fimbrial genes tested, whereas Velvet could reconstruct only two.

Integration of whole-genome sequencing into existing practice in a public health laboratory requires backwards compatibility

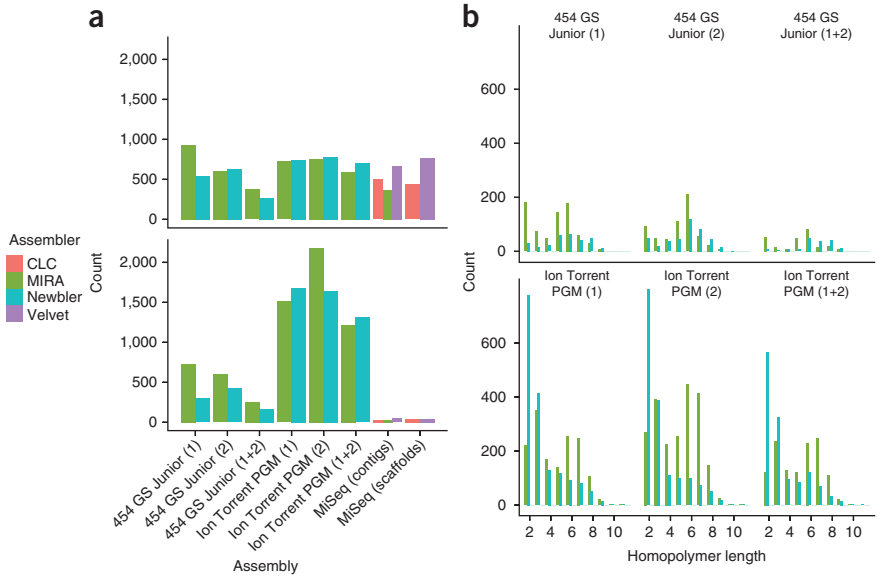

Figure 3 An analysis of gaps when aligning draft *de novo* assemblies to the reference genome. (**a**, top panel) The number of gaps that are not associated with homopolymeric tracts, for example, contig breaks, misassemblies or missing sequence. (bottom panel) The number of gaps that are associated with homopolymeric tracts for each draft assembly. (**b**) The length of erroneously called homopolymeric tracts for each 454 GS Junior and Ion Torrent PGM assembly.

Table 3 Full-length identical matches of clinically important proteins against draft assemblies

Assembly	Assembler	Adhesins	Antibiotic resistance	Serotype antigens	Microcins	Shiga toxin A/B	SPATEs	Tellerium resistance	MLST genes	Percentage found
280 Reference	Reference	4/4	6/6	2/2	1/1	2/2	4/4	12/12	7/7	100.00
C236-11	Reference	4/4	6/6	2/2	1/1	2/2	3/4	12/12	7/7	97.37
MiSeq (contigs)	CLC	4/4	6/6	2/2	1/1	2/2	2/4	12/12	7/7	94.74
MiSeq (scaffolds)	CLC	4/4	6/6	2/2	1/1	2/2	2/4	12/12	7/7	94.74
MiSeq (contigs)	MIRA	4/4	6/6	2/2	1/1	2/2	1/4	12/12	7/7	92.11
MiSeq (contigs)	Velvet	2/4	6/6	2/2	1/1	2/2	3/4	12/12	7/7	92.11
MiSeq (scaffolds)	Velvet	2/4	6/6	2/2	1/1	2/2	3/4	12/12	7/7	92.11
454 Junior (1+2)	Newbler	3/4	6/6	0/2	0/1	2/2	3/4	12/12	7/7	86.84
454 Junior (1)	Newbler	1/4	5/6	0/2	0/1	2/2	2/4	12/12	7/7	76.32
454 Junior (1+2)	MIRA	2/4	6/6	0/2	1/1	2/2	0/4	12/12	6/7	76.32
454 Junior (2)	Newbler	2/4	6/6	0/2	0/1	2/2	3/4	10/12	6/7	76.32
Ion Torrent (1+2)	MIRA	2/4	5/6	1/2	1/1	2/2	1/4	9/12	7/7	73.68
Ion Torrent (1+2)	Newbler	1/4	6/6	2/2	1/1	1/2	3/4	8/12	5/7	71.05
Ion Torrent (1)	Newbler	1/4	6/6	2/2	0/1	2/2	1/4	8/12	6/7	68.42
Ion Torrent (2)	Newbler	1/4	5/6	0/2	0/1	2/2	3/4	8/12	6/7	65.79
454 Junior (1)	MIRA	0/4	5/6	0/2	1/1	1/2	2/4	9/12	6/7	63.16
Ion Torrent (1)	MIRA	1/4	5/6	0/2	1/1	2/2	1/4	5/12	7/7	57.89
Ion Torrent (2)	MIRA	1/4	4/6	0/2	0/1	1/2	1/4	8/12	6/7	55.26
454 Junior (2)	MIRA	1/4	1/6	0/2	0/1	0/2	0/4	10/12	6/7	47.37

Protein coding sequences were searched against draft assemblies for each benchtop instrument using translated BLAST (tblastn, part of the BLAST 2.2.22 package). The results show the number of matches that are identical to the sequence in the reference assembly. For MLST sequences, the nucleotide sequences and nucleotide BLAST (blastn) was used. A summary of BLAST results can be found in **Supplementary Table 5**. SPATEs, serine protease autotransporters.

with existing typing methods. We therefore attempted to generate multi-locus sequence typing (MLST) profiles from each assembly. An accurate MLST profile was generated for the outbreak strain by all assemblies using MiSeq data. However, some 454 GS Junior and Ion Torrent PGM assemblies generated indel errors in at least one housekeeping gene.

DISCUSSION

In our evaluation, all three benchtop sequencing platforms generated useful draft genome sequences of the German *E. coli* outbreak strain. All could be judged suitable for bacterial genome sequencing, in producing assemblies that mapped to 95% or more of the reference genome and recovered the vast majority of coding sequences. As expected, no instrument could generate completely accurate one-contig-per-replicon assemblies that might equate to a finished genome. Thus, for each technology there is a trade-off between advantages and disadvantages. In our survey, the MiSeq generated the highest throughput per run and lowest error rate of the instruments, without significant indel errors and the lowest rate of substitution errors (although accuracy does drop off toward the ends of reads). However, the MiSeq delivered shorter read lengths than the 454 GS Junior, probably a significant factor in the lower quality assemblies produced from MiSeq data. Even with paired-end sequencing, the single scaffold assemblies from the MiSeq are interrupted by unfillable gaps, representing difficult-to-resolve repeats. The MiSeq was the longest-running instrument, with paired-end, 150-base sequencing on a pre-release instrument taking >27 h (60 Mb/h).

The 454 GS Junior delivered the longest read length but the lowest throughput (8 Mb/h during a 9-h run) and suffered from errors in homopolymeric tracts, even when assembled at high coverage. Each Ion Torrent PGM run produced the shortest reads and the worst performance with homopolymers. However, it delivered the fastest throughput (80–100 Mb/h) and shortest run time (~3 h). This platform has also shown the greatest improvement in performance in recent months. An assembly for the outbreak strain generated in May 2011 from data from the original Ion Torrent 314 chip contained >3,000 contigs[10], whereas, in this study, data from the recently available 316 chip were assembled into <400 contigs.

Speed, set-up, running costs and simplicity of workflow are also important factors when comparing these platforms. The Ion Torrent PGM is the lowest-price instrument. The cost per base of generating sequence data appears to be an order of magnitude higher for the 454 GS Junior than for the other two platforms. The MiSeq workflow has the fewest manual steps as template amplification is done directly on the instrument without manual intervention in contrast to the Ion Torrent PGM and 454 GS Junior, which require preparation of amplified sequence libraries through emulsion PCR and enrichment stages off the instrument. The Ion Torrent PGM is notable for offering three differently priced sequencing-chip reagents, which gives flexibility when designing experiments, as a choice can be made based on the throughput required. Since this study was carried out, a paired-end protocol for the Ion Torrent PGM has been announced, similar to that for the MiSeq, which requires a second sequencing reaction to be done immediately after the first, which also has the effect of doubling the run-time (http://www.iontorrent.com/lib/images/PDFs/pe_appnote_v12b.pdf).

One important conclusion from this evaluation is that saying that one has "sequenced a bacterial genome" means different things on different benchtop sequencing platforms. Potential users of these technologies need to be sensitive to these differences, particularly when comparing or combining data generated on different platforms. It is also important to ask (i) to what extent errors can be corrected by comparison to reference data, (ii) when it is safe to use a mapping approach that makes assumptions about the resemblance of a novel sequence to an existing reference sequence and (iii) how much one should have to rely on human insight rather than automated analyses and pipelines. In this study, we set a tough test by evaluating algorithmically generated *de novo* assemblies. However, during the real-world test case of the German *E. coli* outbreak, even the Ion Torrent platform, using the 314 chip with its low throughput and high error rate, delivered useful insights into the biology and evolution of the outbreak strain—for example, a homopolymer error in an MLST profile was easily corrected by manual comparison to database sequences[9,10]. We are thus confident that benchtop high-throughput sequencing platforms are poised to make a decisive impact on diagnostic and public health microbiology in the near future.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturebiotechnology/.

Accession codes. 454 sequences have been deposited into the Short Read Archive under study number SRA048574, with run accessions SRR388806 (454 GS Junior run 1), SRR388807 (454 GS Junior run 2), SRR388808 (454 FLX+) and SRR388809 (454 Titanium 8 kb paired-end). Ion Torrent PGM sequences have been deposited under study number SRA048511, with accessions SRR389193 (Ion Torrent PGM run 1), SRR389194 (Ion Torrent PGM run 2). The multiplexed MiSeq reads have been deposited under study number SRA048664. Assembly files and analysis scripts have been uploaded to a public Github repository (https://github.com/nickloman/benchtop-sequencing-comparison).

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

We gratefully acknowledge the blogging community for helpful discussion in the comments section of our blog (http://pathogenomics.bham.ac.uk/blog/), and in particular to B. Chevreux, J. Johnson, K. Robison and L. Nederbragt. We are grateful to C. Hercus at Novocraft for help with the Novoalign software and to A. Darling for help with Mauve Assembly Metrics. We thank Roche Diagnostics, UK, for 454 GS FLX+ and 454 FLX paired-end sequencing, technical support and helpful discussion. We thank Life Technologies for early access to 316 chips and instrument fluidics upgrade. We thank G. Smith and Illumina UK for early access to the MiSeq platform and public release of *E. coli* outbreak-strain data. We thank the three anonymous reviewers for their many helpful suggestions for improving the manuscript. The xBASE facility and N.J.L. are funded by BBSRC grant BBE0111791.

AUTHOR CONTRIBUTIONS

N.J.L., J.W., S.E.G. and M.J.P. conceived the experiments; J.W. and S.G. supplied the strains; N.J.L., R.V.M. and T.J.D. carried out the bioinformatics analysis; C.C. performed the Ion Torrent sequencing; and S.E.G. and R.V.M. performed the 454 GS Junior sequencing. N.J.L. and M.J.P. wrote the manuscript. All authors commented on the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/naturebiotechnology/.

Published online at http://www.nature.com/naturebiotechnology/.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Pallen, M.J., Nelson, K. & Preston, G.M. *Bacterial Pathogenomics* (ASM Press, 2007).
2. Metzker, M. Sequencing technologies—the next generation. *Nat. Rev. Genet.* **11**, 31–46 (2010).
3. Glenn, T. Field guide to next-generation DNA sequencers. *Mol. Ecol. Resour.* **11**, 759–769 (2011).
4. Pallen, M., Loman, N. & Penn, C. High-throughput sequencing and clinical microbiology: progress, opportunities and challenges. *Curr. Opin. Microbiol.* **13**, 625–631 (2010).
5. Rothberg, J. *et al.* An integrated semiconductor device enabling non-optical genome sequencing. *Nature* **475**, 348–352 (2011).
6. Bentley, D. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).
7. Frank, C. *et al.* Epidemic profile of Shiga-toxin-producing *Escherichia coli* O104:H4 outbreak in Germany. *N. Engl. J. Med.* **365**, 1771–1780 (2011).
8. Brzuszkiewicz, E. *et al.* Genome sequence analyses of two isolates from the recent *Escherichia coli* outbreak in Germany reveal the emergence of a new pathotype: entero-aggregative-haemorrhagic *Escherichia coli* (EAHEC). *Arch. Microbiol.* **193**, 883–891 (2011).
9. Mellmann, A. *et al.* Prospective genomic characterization of the German enterohemorrhagic *Escherichia coli* O104:H4 outbreak by rapid next generation sequencing technology. *PLoS ONE* **6**, e22751 (2011).
10. Rohde, H. *et al.* Open-source genomic analysis of Shiga-toxin-producing *E. coli* O104:H4. *N. Engl. J. Med.* **365**, 718–724 (2011).
11. Rasko, D. *et al.* Origins of the *E. coli* strain causing an outbreak of hemolytic-uremic syndrome in Germany. *N. Engl. J. Med.* **365**, 709–717 (2011).
12. Grad, Y. *et al.* Genomic epidemiology of the *Escherichia coli* O104:H4 outbreaks in Europe, 2011. *Proc. Natl. Acad. Sci. USA* **109**, 3065–3070 (2012).
13. Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **8**, 186–194 (1998).
14. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
15. Kingsford, C., Schatz, M. & Pop, M. Assembly complexity of prokaryotic genomes using short reads. *BMC Bioinformatics* **11**, 21 (2010).
16. Buffalo, V. qrqc: Quick Read Quality Control R package version 1.9.1 <http://bioinformatics.ucdavis.edu/> (2012).

ONLINE METHODS

Collection of isolates. *E. coli* strain 280 was grown according to the protocol described[17]. To generate enough DNA for sequencing, the isolate was grown on multiple occasions.

Sequencing workflow. A general, simplified workflow for library preparation, amplification and sequencing is shown in **Supplementary Figure 5** with approximate timings for each stage. These stages comprise library preparation from genomic DNA, amplification and sequencing. Library preparation steps are similar for each instrument, involving extraction and purification of genomic DNA, fragmentation through either enzymatic or physical means, fragment size selection and ligation of sequencing adaptors.

Ion Torrent sequencing. Ion Torrent sequencing was performed at the University of Birmingham according to the Ion Torrent protocol (Life Technologies). Total DNA from *E. coli* O104:H4 280 was isolated. Ten milligrams of this DNA was fragmented with a Bioruptor instrument (Diagenode, Liège, Belgium) using the protocol recommended by Life Technologies. A broad profile of fragment sizes (75–500 bp, peak at 255 bp) were obtained that were end-repaired, ligated with Ion Torrent A and P1 adaptors and size selected using E-Gel EX 2% Gel (Invitrogen, Carlsbad, CA) for 150- to 250-bp fragments. The size-selected fragments were amplified and DNA was purified with Agencourt AMPure XP beads (Beckman Coulter Genomics, High Wycombe, UK). The median fragment size of the final library was 200 bp (assessed by a BioAnalyzer High Sensitivity LabChip, Agilent). Library was diluted to 40 pM and two emulsion PCR reactions were set up at two templates per sphere. Sequencing primer and polymerase were added to the final enriched spheres before loading onto the 316 chip. Two 316 chips were run in total. Base calls were generated using version 1.5 of the Ion Torrent software suite and for further analysis, the resulting flowgram files (assembly) or FASTQ files (alignment) were used.

454 GS Junior sequencing. 454 GS Junior sequencing was carried out on an instrument at the Health Protection Agency, Colindale, UK. *E. coli* O104:H4 280 DNA was prepared following the Roche Rapid Library protocol (Roche, Welwyn Garden City, UK), whereby 5 ng/µl was taken from each sample and libraries prepared. Briefly, samples were subjected to the following key steps: DNA fragmentation by nebulization, fragment end-repair, AMPure XP bead preparation (Amersham International, Buckinghamshire, UK), adaptor ligation, small fragment removal, quality assessment using the Agilent 2100 Bioanalyzer, library quantification and finally preparation of working aliquots at a final concentration of 1×10^7 molecules (500 total). Emulsions PCR, enrichment and 454 GS Junior sequencing were carried out per manufacturer's protocols. The resulting flowgram files were used for downstream analysis.

454 GS FLX+ and 454 GS FLX 8-kb titanium sequencing. 454 GS FLX 8-kb titanium paired-end and 454 FLX+ (long read) library construction and sequencing was performed at Roche Diagnostics (Burgess Hill, UK) according to their standard protocols.

Illumina MiSeq sequencing. Illumina MiSeq sequencing was done at Illumina UK, Little Chesterford, UK, on a pre-release, prototype MiSeq instrument. The seven *E. coli* samples were quantified with a Qubit High Sensitivity kit and the total amount of DNA for each sample varied between 523 ng and 954 ng. Samples were sheared with a Covaris S2 instrument followed by end repair, A-tailing and the ligation of TruSeq adaptors containing indexes. Samples were run on a 2% agarose gel (2 samples per gel) and DNA was size selected at 600–700 bp. Ten cycles of PCR were carried out and samples run out on a second 2% agarose gel (two samples per gel). Samples were excised from the gel and quantified with a Qubit high-sensitivity kit. Libraries were diluted to 2 nM in EB plus 0.1% Tween and a pool containing an equimolar concentration of each library was prepared. MiSeq instrument was prepared following routine procedures. Briefly, a standard MiSeq flow cell was inserted into the flow-cell chamber. Next, the DNA sample containing the pool of seven *E. coli* libraries was diluted to 6.2 pmol and pipetted into the sample well on the MiSeq Consumable Cartridge before loading in the chiller section of the MiSeq instrument. A sample sheet was prepared on the MiSeq instrument to provide run details. The run was initiated for 2 × 150 bases of SBS sequencing, including on-board clustering and paired-end preparation, the sequencing of the seven barcode indices and analysis. On the completion of the run, data were base called and demultiplexed on the instrument (provided as Illumina FASTQ files, Phred+64 encoding). FASTQ format files in Illumina 1.5 format were considered for downstream analysis. Although MiSeq produces reads of fixed lengths, tails of these reads may be designated as uncallable as indicated by the read segment quality control indicator, noted by a quality score of two ('B'). In these cases these low-quality tails were trimmed and not used for further analysis.

Bioinformatics

Construction of reference assembly. A high-quality reference sequence for *E. coli* strain 280 was constructed by assembling 454 FLX+ long read data and 454 Titanium paired-end data (8-kb insert) using Newbler 2.6. Newbler was run with parameters -scaffold -tr -cpu 8 -siom 28 -rip. The resulting scaffolds were used for further analysis. Newbler masks certain bases in the assembly regarded as uncertain by assigning it a lower-case nucleotide. These masked bases correspond with bases with a low-quality score. In bacterial genomes these bases are seen predominantly in consensus contigs resulting from long repeat regions, long homopolymeric tracts and contig ends. The resulting assembly was annotated using the automated xBASE annotation pipeline[18], which uses Glimmer for coding sequence prediction[19] and tRNAScan-SE and RNAmmer for stable RNA prediction[20,21].

***De novo* assembly of individual strains.** Assemblies were generated from data generated by each of the benchtop sequencing platforms separately. All data were assembled by MIRA 3.4.0 using default parameters in genome,denovo,accurate mode and the appropriate setting for each instrument type (454,iontor,solexa). Ion Torrent and 454 GS Junior data were additionally assembled with Newbler 2.6 with default parameters. Illumina MiSeq data were additionally assembled using Velvet and CLC Assembly Cell (both de Bruijn graph assemblers). Velvet was run using a *k*-mer value of 55 and exp_cov and cov_cutoff set to auto. The program was run again with -scaffolding off to generate a separate assembly without scaffolds. CLC Assembly Cell version 4.0.6 beta was run with default parameters. *De novo* assemblies were compared for chromosomal coverage and broken genes, among other items using Mauve (mauve_snapshot_2011-08-19) and the Mauve Assembly Metrics package[22]. Assemblies were manually examined using the Tablet viewer[23]. Assembly gaps were inspected using a custom script extract_hp.py, which uses as input gaps reported by Mauve Assembly Metrics. Gaps in the whole-genome alignment that are associated with homopolymeric tracts in the reference (of length two or more) were categorized as homopolymer gaps; other gaps were categorized as assembly gaps. Gaps in the reference sequence were not counted.

Read mapping. For substitution and indel detection, reads from each platform were aligned to the reference assembly using the bwasw module of BWA (version 0.5.9rc1)[24]. The reference genome was indexed with bwa index -a is. The bwasw module was run with default parameters (gap open penalty 5, gap extension penalty 2) using FASTQ files as input. Output BAM files were post-processed using the calmd module of SAMtools, which adds MD tags to each alignment. The MD tag describes the positions of base substitutions. Reads that align to masked bases in the reference genome were excluded from analysis. Read accuracy was determined by a custom Python script (calculate_accuracy.py, available in the Github repository) that uses the pysam module (http://code.google.com/p/pysam/) to read the BAM alignment. The calculate_accuracy script counts mismatches using a published method[14], which counts mismatches resulting from substitutions, insertions and deletions. In the case of deletions, mismatches are assigned to one of the adjacent bases in the read at random. Reads were additionally mapped against *E. coli* strain c236-11 (PacBio and Illumina sequenced) and *E. coli* strain 55989 (Sanger sequenced)[12,25].

For generation of homopolymer accuracy plots, reads for each of the benchtop sequencing platforms were mapped to the reference assembly using Novoalign (version V2.07.13, Novocraft, Malaysia, registered version). Gap penalties were adjusted with parameters as recommended by the documentation -g 20 -x 5. Novoalign was set to align its maximum supported read length of 300 using -n 300. Homopolymeric tract statistics were enabled using the –hpstats option.

17. Chattaway, M., Dallman, T., Okeke, I. & Wain, J. Enteroaggregative *E. coli* O104 from an outbreak of HUS in Germany 2011, could it happen again? *J. Infect. Dev. Ctries.* **5**, 425–436 (2011).
18. Chaudhuri, R. *et al.* xBASE2: a comprehensive resource for comparative bacterial genomics. *Nucleic Acids Res.* **36**, D543–546 (2008).
19. Delcher, A., Bratke, K., Powers, E. & Salzberg, S. Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics* **23**, 673–679 (2007).
20. Lowe, T. & Eddy, S. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955–964 (1997).
21. Lagesen, K. *et al.* RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* **35**, 3100–3108 (2007).
22. Darling, A., Tritt, A., Eisen, J. & Facciotti, M. Mauve assembly metrics. *Bioinformatics* **27**, 2756–2757 (2011).
23. Milne, I. *et al*. Tablet–next generation sequence assembly visualization. *Bioinformatics* **26**, 401–402 (2010).
24. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
25. Touchon, M. *et al.* Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. *PLoS Genet.* **5**, e1000344 (2009).

doi:10.1038/nbt.2198