Quail et al. BMC Genomics 2012, 13:341
http://www.biomedcentral.com/1471-2164/13/341

BMC Genomics

RESEARCH ARTICLE Open Access

A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers

Michael A Quail*, Miriam Smith, Paul Coupland, Thomas D Otto, Simon R Harris, Thomas R Connor, Anna Bertoni, Harold P Swerdlow and Yong Gu

Abstract

Background: Next generation sequencing (NGS) technology has revolutionized genomic and genetic research. The pace of change in this area is rapid with three major new sequencing platforms having been released in 2011: Ion Torrent's PGM, Pacific Biosciences' RS and the Illumina MiSeq. Here we compare the results obtained with those platforms to the performance of the Illumina HiSeq, the current market leader. In order to compare these platforms, and get sufficient coverage depth to allow meaningful analysis, we have sequenced a set of 4 microbial genomes with mean GC content ranging from 19.3 to 67.7%. Together, these represent a comprehensive range of genome content. Here we report our analysis of that sequence data in terms of coverage distribution, bias, GC distribution, variant detection and accuracy.

Results: Sequence generated by Ion Torrent, MiSeq and Pacific Biosciences technologies displays near perfect coverage behaviour on GC-rich, neutral and moderately AT-rich genomes, but a profound bias was observed upon sequencing the extremely AT-rich genome of *Plasmodium falciparum* on the PGM, resulting in no coverage for approximately 30% of the genome. We analysed the ability to call variants from each platform and found that we could call slightly more variants from Ion Torrent data compared to MiSeq data, but at the expense of a higher false positive rate. Variant calling from Pacific Biosciences data was possible but higher coverage depth was required. Context specific errors were observed in both PGM and MiSeq data, but not in that from the Pacific Biosciences platform.

Conclusions: All three fast turnaround sequencers evaluated here were able to generate usable sequence. However there are key differences between the quality of that data and the applications it will support.

Keywords: Next-generation sequencing, Ion torrent, Illumina, Pacific biosciences, MiSeq, PGM, SMRT, Bias, Genome coverage, GC-rich, AT-rich

Background

Sequencing technology is evolving rapidly and during the course of 2011 several new sequencing platforms were released. Of note were the Ion Torrent Personal Genome Machine (PGM) and the Pacific Biosciences (PacBio) RS that are based on revolutionary new technologies.

The Ion Torrent PGM "harnesses the power of semiconductor technology" detecting the protons released as nucleotides are incorporated during synthesis [1]. DNA fragments with specific adapter sequences are linked to and then clonally amplified by emulsion PCR on the surface of 3-micron diameter beads, known as Ion Sphere Particles. The templated beads are loaded into proton-sensing wells that are fabricated on a silicon wafer and sequencing is primed from a specific location in the adapter sequence. As sequencing proceeds, each of the four bases is introduced sequentially. If bases of that type are incorporated, protons are released and a signal is detected proportional to the number of bases incorporated.

PacBio have developed a process enabling single molecule real time (SMRT) sequencing [2]. Here, DNA polymerase molecules, bound to a DNA template, are

* Correspondence: mq1@sanger.ac.uk
Wellcome Trust Sanger Institute, Hinxton, UK

attached to the bottom of 50 nm-wide wells termed zero-mode waveguides (ZMWs). Each polymerase is allowed to carry out second strand DNA synthesis in the presence of γ-phosphate fluorescently labeled nucleotides. The width of the ZMW is such that light cannot propagate through the waveguide, but energy can penetrate a short distance and excite the fluorophores attached to those nucleotides that are in the vicinity of the polymerase at the bottom of the well. As each base is incorporated, a distinctive pulse of fluorescence is detected in real time.

In recent years, the sequencing industry has been dominated by Illumina, who have adopted a sequencing-by-synthesis approach [3], utilizing fluorescently labeled reversible-terminator nucleotides, on clonally amplified DNA templates immobilized to an acrylamide coating on the surface of a glass flowcell. The Illumina Genome Analyzer and more recently the HiSeq 2000 have set the standard for high throughput massively parallel sequencing, but in 2011 Illumina released a lower throughput fast-turnaround instrument, the MiSeq, aimed at smaller laboratories and the clinical diagnostic market.

Here we evaluate the output of these new sequencing platforms and compare them with the data obtained from the Illumina HiSeq and GAIIx platforms. Table 1 gives a summary of the technical specifications of each of these instruments.

Results

Sequence generation

Platform specific libraries were constructed for a set of microbial genomes *Bordetella pertussis* (67.7% GC, with some regions in excess of 90% GC content), *Salmonella* Pullorum (52% GC), *Staphylococcus aureus* (33% GC) and *Plasmodium falciparum* (19.3% GC, with some regions close to 0% GC content). We routinely use these to test new sequencing technologies, as together their sequences represent the range of genomic landscapes that one might encounter.

PCR-free [4] Illumina libraries were uniquely barcoded, pooled and run on a MiSeq flowcell with paired 150 base reads plus a 6-base index read and also on a single lane of an Illumina HiSeq with paired 75 base reads plus an 8-base index read (Additional file 1: Table S1). Illumina libraries prepared with amplification using Kapa HiFi polymerase [5] were run on a single lane of an Illumina GA IIx with paired 76 base reads plus an 8-base index read and on a MiSeq flowcell with paired 150 base reads plus a 6-base index read. PCR-free libraries represent an improvement over the standard Illumina library preparation method as they result in more even sequence coverage [4] and are included here alongside libraries prepared with PCR in order to enable comparison to PacBio which has an amplification free workflow.

Ion Torrent libraries were each run on a single 316 chip for a 65 cycles generating mean read lengths of 112–124 bases (Additional file 1: Table S2). Standard PacBio libraries, with an average of 2 kb inserts, were run individually over multiple SMRT cells, each using C1 chemistry, and providing ≥20x sequence coverage data for each genome (Additional file 1: Table S3).

The datasets generated were mapped to the corresponding reference genome as described in Methods. For a fair comparison, all sequence datasets were randomly down-sampled (normalized) to contain reads representing a 15x average genome coverage.

Table 1 Technical specifications of Next Generation Sequencing platforms utilised in this study

Platform	Illumina MiSeq	Ion Torrent PGM	PacBio RS	Illumina GAIIx	Illumina HiSeq 2000
Instrument Cost*	$128 K	$80 K**	$695 K	$256 K	$654 K
Sequence yield per run	1.5-2Gb	20-50 Mb on 314 chip, 100-200 Mb on 316 chip, 1Gb on 318 chip	100 Mb	30Gb	600Gb
Sequencing cost per Gb*	$502	$1000 (318 chip)	$2000	$148	$41
Run Time	27 hours***	2 hours	2 hours	10 days	11 days
Reported Accuracy	Mostly > Q30	Mostly Q20	<Q10	Mostly > Q30	Mostly > Q30
Observed Raw Error Rate	0.80 %	1.71 %	12.86 %	0.76 %	0.26 %
Read length	up to 150 bases	~200 bases	Average 1500 bases**** (C1 chemistry)	up to 150 bases	up to 150 bases
Paired reads	Yes	Yes	No	Yes	Yes
Insert size	up to 700 bases	up to 250 bases	up to 10 kb	up to 700 bases	up to 700 bases
Typical DNA requirements	50-1000 ng	100-1000 ng	~1 μg	50-1000 ng	50-1000 ng

* All cost calculations are based on list price quotations obtained from the manufacturer and assume expected sequence yield stated.
** System price including PGM, server, OneTouch and OneTouch ES.
*** Includes two hours of cluster generation.
**** Mean mapped read length includes adapter and reverse strand sequences. Subread lengths, i.e. the individual stretches of sequence originating from the sequenced fragment, are significantly shorter.

Workflow

All the platforms have library preparation protocols that involve fragmenting genomic DNA and attaching specific adapter sequences. Typically this takes somewhere between 4 and 8 hours for one sample. In addition, the Ion Torrent template preparation has a two hour emulsion PCR and a template bead enrichment step.

In the battle to become the platform with the fastest turnaround time, all the manufacturers are seeking to streamline library preparation protocols. Life Technologies have developed the Ion Xpress Fragment Library Kit that has an enzymatic "Fragmentase" formulation for shearing starting DNA, thereby avoiding the labour of physical shearing and potentially enabling complete library automation. We tested this kit on our four genomes alongside the standard library kit with physical shearing and found both to give equal genomic representation (see Additional file 2: Figure S1 for results obtained with *P. falciparum*). Illumina purchased Epicentre in order to package the Nextera technology with the MiSeq. Nextera uses a transposon to shear genomic DNA and simultaneously introduce adapter sequences [6]. The Nextera method can produce sequencing ready DNA in around 90 minutes and gave us remarkably even genome representation (Additional file 2: Figures S2 and Additional file 2: Figure S3) with *B. pertussis* and *S. aureus*, but produced a very biased sequence dataset from the extremely AT-rich *P. falciparum* genome.

Genome coverage and GC bias

To analyse the uniformity of coverage across the genome we tabulated the depth of coverage seen at each position of the genome. We utilized the coverage plots described by Lam et al., [7] that depict; the percentage of the genome that is covered at a given read depth, and genome coverage at different read depths respectively, for each dataset (Figure 1) alongside the ideal theoretical coverage that would be predicted based on Poisson behaviour.

In the context of the GC-rich genome of *B. pertussis*, most platforms gave similar uniformity of sequence coverage, with the Ion Torrent data giving slightly more uneven coverage. In the *S. aureus* genome the PGM performed better. The PGM gave very biased coverage when sequencing the extremely AT-rich *P. falciparum* genome (Figure 1). This affect was also evident when we plotted coverage depth against GC content (Additional file 2: Figure S4). Whilst the PacBio platform gave a sequence dataset with quite even coverage on GC and extremely AT-rich contexts, it did demonstrate slight but noticeable unevenness of coverage and bias towards GC-rich sequences with the *S. aureus* genome. With the GC-neutral *S.* Pullorum genome all platforms gave equal coverage with unbiased GC representation (data not shown).

The most dramatic observation from our results was the severe bias seen when sequencing the extremely AT-rich genome of *P. falciparum* on the PGM. The result of this was deeper than expected coverage of the GC-rich *var* and subtelomeric regions and poor coverage within introns and AT-rich exonic segments (Figure 2), with approximately 30% of the genome having no sequence coverage whatsoever. This bias was observed with libraries prepared using both enzymatic and physical shearing (Additional file 2: Figure S1).

In a recent study to investigate the optimal enzyme for next generation library preparation [5], we found that the enzyme used for fragment amplification during next generation library preparation can have a significant influence on bias. We found the enzyme Kapa HiFi amplifies fragments with the least bias, giving even coverage, close to that obtained without amplification. Since the PGM has two amplification steps, one during library preparation and the other emulsion PCR (emPCR) for template amplification, we reasoned that this might be the cause of the observed bias. Substituting the supplied Platinum Taq enzyme with Kapa HiFi for the nick translation and amplification step during library preparation profoundly reduced the observed bias (Figure 3). We were unable to further improve this by use of Kapa HiFi for the emPCR (results not shown).

Of the four genomes sequenced, the *P. falciparum* genome is the largest and most complex and contains a significant quantity of repetitive sequences. We used *P. falciparum* to analyse the effect of read length versus mappability. As the PacBio pipeline doesn't generate a mapping quality value and to ensure a fair comparison, we remapped the reads of all technologies using the k-mer based mapper, SMALT [9], and then analysed coverage across the *P. falciparum* genome (Additional file 3: Table S4). This data confirms the poor performance of Ion Torrent on the *P. falciparum* genome, as only 65% of the genome is covered with high quality (>Q20) reads compared to ~98-99% for the other platforms. Whilst the mean mapped readlength of the PacBio reads with this genome was 1336 bases, average subread length (the length of sequence covering the genome) is significantly less (645 bases). The short average subread length is due to preferential loading of short fragment constructs in the library and the effect of lag time (non-imaged bases) after sequencing initiation, the latter resulting in sequences near the beginning of library constructs not being reported.

As the median length of the PacBio subreads for this data set are just 600 bases, we compared their coverage with an equivalent amount of *in silico* filtered reads of >620 bases. This led to a very small decrease in the percentage of bases covered. Using paired reads on the

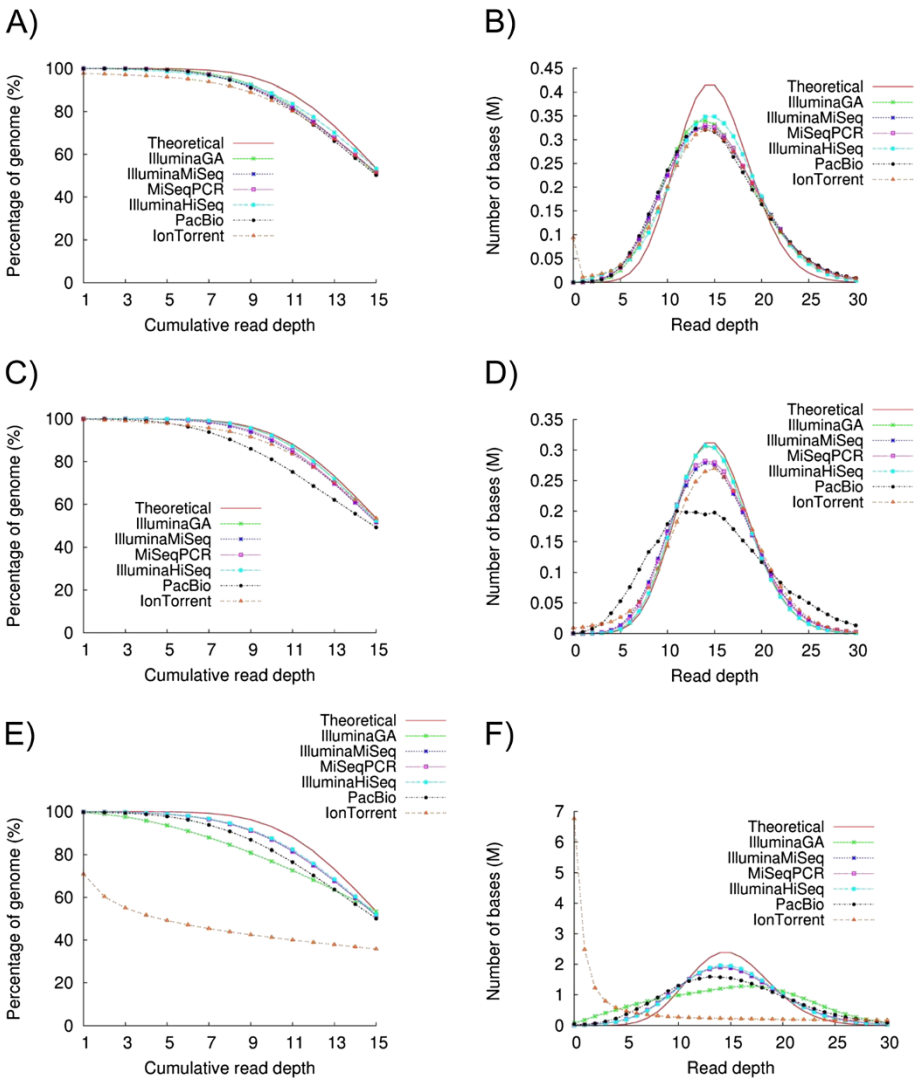

Figure 1 Genome coverage plots for 15x depth randomly downsampled sequence coverage from the sequencing platforms tested. **A**) The percentage of the *B. pertussis* genome covered at different read depths; **B**) The number of bases covered at different depths for *B. pertussis;* **C**) The percentage of the *S. aureus* genome covered at different read depths; **D**) The number of bases covered at different depths for *S. aureus;* **E**) The percentage of the *P. falciparum* genome covered at different read depths; and **F**) The number of bases covered at different depths for *P. falciparum*.

Illumina MiSeq, however, gave a strong positive effect, with 1.1% more coverage being observed from paired-end reads compared to single-end reads.

Error rates

We observed error rates of below 0.4% for the Illumina platforms, 1.78% for Ion Torrent and 13% for PacBio sequencing (Table 1). The number of error-free reads, without a single mismatch or indel, was 76.45%, 15.92% and 0% for, MiSeq, Ion Torrent and PacBio, respectively. The error heatmap in Figure 2A shows that the PacBio errors are distributed evenly over the chromosome. We manually inspected the regions where Ion Torrent and Illumina generated more errors. Illumina produced errors after long (> 20-base) homopolymer tracts [10] (Figure 4A).

Also evident in the MiSeq data, were strand errors due to the GGC motif [11]. Following the finding that the motif GGC generates strand-specific errors, we analyzed this phenomenon in the MiSeq data for *P. falciparum* (Additional file 4: Table S5). We observed that the error is mostly generated by GC-rich motifs, principally GGCGGG. We found no evidence for an error if the triplet after the GGC is AT-rich. Other MiSeq datasets also showed this artifact (data not shown). In addition to this being a strand-specific issue, it appears that this is a read-specific phenomenon. Whilst there is a quality drop in the first read following these GC-rich motifs, there is

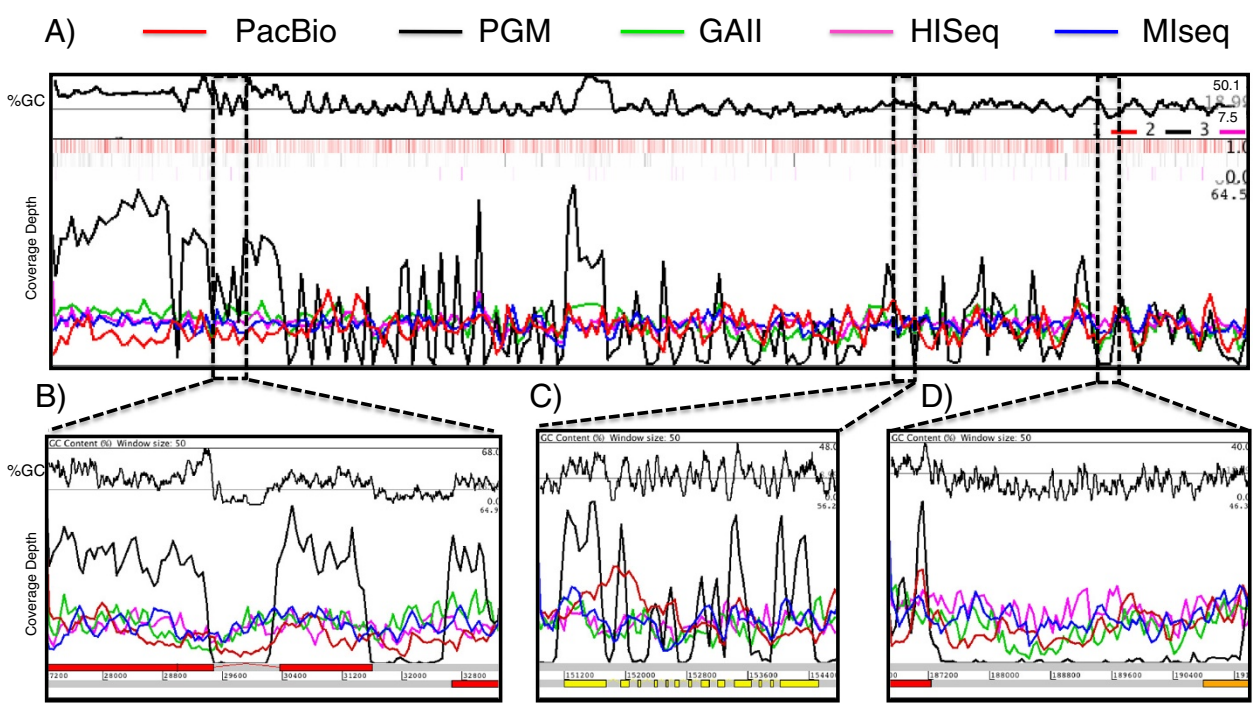

Figure 2 Artemis genome browser [8] **screenshots illustrating the variation in sequence coverage of a selected region of *P. falciparum* chromosome 11, with 15x depth of randomly normalized sequence from the platforms tested.** In each window, the top graph shows the percentage GC content at each position, with the numbers on the right denoting the minimum, average and maximum values. The middle graph in each window is a coverage plot for the dataset from each instrument; the colour code is shown above graph a). Each of the middle graphs shows the depth of reads mapped at each position, and below that in B-D are the coordinates of the selected region in the genome with gene models on the (+) strand above and (–) strand below. **A**) View of the first 200 kb of chromosome 11. Graphs are smoothed with window size of 1000. A heatmap of the errors, normalized by the amount of mapping reads is included just below the GC content graph (PacBio top line, PGM middle and MiSeq bottom). **B**) Coverage over region of extreme GC content, ranging from 70% to 0%. **C**) Coverage over the gene PF3D7_1103500. **D**) Example of intergenic region between genes PF3D7_1104200 and PF3D7_1104300. The window size of B, C and D is 50 bp.

a striking loss of quality in read 2, where the reads have nearly half the mean quality value compared to the read 1 reads for GC-rich triplets that follow the GGC motif. We could observe this low quality in read 2 in all our analysed Illumina lanes. For AT-rich motifs the ratio is nearly 1 (1.03).

Ion Torrent didn't generate reads at all for long (> 14-base) homopolymer tracts, and could not predict the correct number of bases in homopolymers >8 bases long. Very few errors were observed following short homopolymer stretches in the MiSeq data (Figure 4B). Additionally, we observed strand-specific errors in the PGM data but were unable to associate these with any obvious motif (Figure 4C).

SNP calling

In order to determine whether or not the higher error rates observed with the PGM and PacBio affected their ability to call SNPs, we aligned the reads from the *S. aureus* genome, for which all platforms gave good sequence representation, against the reference genome of the closely related strain USA300_FPR3757 [12], and compared the SNPs called against those obtained by aligning the reference sequences of the two genomes (Figure 5 and Additional file 5: Table S6). In order to create a fair comparison we initially used the same randomly normalized 15x datasets used in our analysis of genome coverage, which according to the literature [3] is sufficient to accurately call heterozygous variants but found that that was insufficient for the PacBio datasets where a 190x coverage was used.

Overall the rate of SNP calling was slightly higher for the Ion Torrent data than for Illumina data (chi square p value 3.15E-08), with approximately 82% of SNPs being correctly called for the PGM and 68-76% of the SNPs being detected from the Illumina data (Figure 5A). Conversely, the rate of false SNP calls was higher with Ion Torrent data than for Illumina data (Figure 5B). SNP calling from PacBio data proved more problematic, as existing tools are optimized for short-read data and not for high error-rate long-read data. We were reliant on SNPs called by the SMRT portal pipeline for this analysis. Our results showed that SNP detection from PacBio data was not as accurate as that from the other

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS
Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS
Sync your system to PACER to automate legal marketing.