Evaluating the potential of new sequencing technologies for genotyping and variation discovery in human data

Mauricio Carneiro, Carsten Russ, Michael Ross, Patrick Cahill, Stacey Gabriel, Chad Nusbaum, Mark A. DePristo

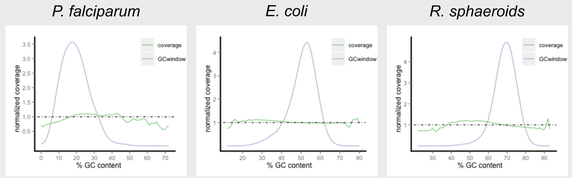
BROAD INSTITUTE

Pacific Biosciences RS

A first look at Pacific Biosciences RS data

Pacific Biosciences technology provides a fundamentally new data type that provides the potential to overcome these limitations by providing significantly longer reads (now averaging >1kb), enabling more unique seeds for reference alignment. In addition, the lack of amplification in the library construction step avoids a common source of base composition bias. With these potential advantages in mind, we here evaluate the utility of the Pacific Biosciences RS platform for human medical resequencing projects by assessing the quality of the raw sequencing data, as well as its use for SNP discovery and genotyping using the Genome Analysis Toolkit (GATK).

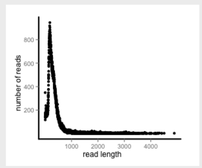
P. falciparum *E. coli* *R. sphaeroids*

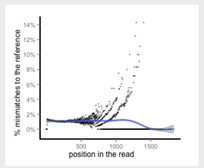


normalized coverage by GC content contrasted with GC windows of the genome shows Pacific Biosciences RS has nearly no bias related

Long reads with insertions as the primary error mode

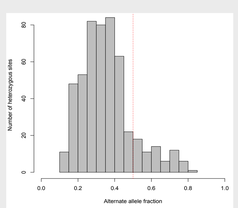
We first defined the basic performance parameters of the data, in terms of read length, base accuracy and error profile, and sequence composition-based representational bias. Using a dataset of 4 runs of data (47,638 reads) from a human sample (see methods), we profiled read length, showing an average of 700 bases over a wide distribution, including 5% of reads >2500 bases. Errors were characterized by comparison to the known reference sequence, showing that the primary error mode was insertions, at 12%, followed by deletions, 2% and apparent mismatch errors at 1%. Further, errors were randomly distributed across the read, rather than increasing in distal positions as is true for other data types.

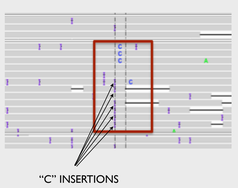




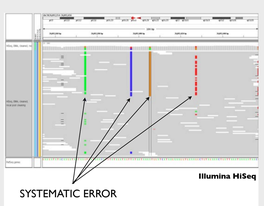
Error rate is high, but error mode is not context-specific

The random nature of Pacific Biosciences error mode empowers the statistical framework of the GATK to distinguish artifacts from real bases. Also due to the high insertion and deletion rate and the length of its reads, Pacific Biosciences data suffers from reference bias (when the aligner prefers to hide the true mismatch inside an insertion for a better scoring alignment). It is important to note however, that reference bias is an artifact of the alignment process, rather than of the data, and can be greatly reduced by locally realigning the reads based on the reference and the data.





"C" INSERTIONS



SYSTEMATIC ERROR

RANDOM ERROR

Comparative Analysis

	sensitivity	specificity	PPV	NPV
PacBio RS	98.1%	100%	100%	68.7%
Ion Torrent	96.2%	100%	100%	54.5%
Illumina MiSeq	98.1%	92.3%	99.6%	70.5%

1000 Genomes Project singleton validation in 8 samples

We sequenced 300 singleton SNPs from the 1000 genomes project public callset that were previously validated with Sanger and Sequenom to evaluate Ion Torrent, Illumina MiSeq and Pacific Biosciences value as a validation tool. All three technologies performed well in the validation assay. A closer look at the first results out of the Pacific Biosciences and Sequenom comparison proved to be extremely informative as the longer reads of the Pacific Bioscience instrument allowed to disambiguate the sites that were wrongly called monomorphic by Sequenom.

	Pacbio called Poly	Pacbio called Mono
sequenom called Poly	218	7
sequenom called Mono	8	12

Visual classification	Result from Pacbio
6 look incredibly good	5 Poly, 1 Reference Bias
1 bad mapping quality	Polymorphic
1 has nearby deletion (unclear)	Reads actually didn't belong at location
50 sites not called by sequenom	27 Polymorphic, 23 lack coverage to disambiguate.

Result Pacbio	#	what went wrong
good sequencing	1	Sequenom was wrong
Alt allele placed on insertion	4	Pacbio Reference Bias
No coverage	1	Reads actually didn't belong at location
Wrong ALT allele called	1	GATK triallelic issue (fixed)

	PacBio RS called Poly	PacBio RS called Mono	Illumina MiSeq called Poly	Illumina MiSeq called Mono
Confirmed de novo SNP	37	1	38	0
Confirmed artifact	1	59	5	55

	Sensitivity	Specificity	PPV	NPV
PacBio RS	97%	98%	97%	98%
Illumina MiSeq	100%	91%	88%	100%

	PacBio RS called Poly	PacBio RS called Mono	Illumina HiSeq called Poly	Illumina HiSeq called Mono
Confirmed de novo SNP	48	0	48	0
Confirmed artifact	5	67	35	37

	Sensitivity	Specificity	PPV	NPV
PacBio RS	100%	93%	91%	100%
Illumina HiSeq	100%	51%	58%	100%

Validating "Hard-To-Validate" sites in NA12878

To evaluate Pacific Biosciences and Illumina MiSeq data for validation/extension, we sequenced 98 variant sites based on Illumina GA1 and GA2 data from the 1000G project. These had been previously validated either as true de novo mutations (38 sites) or false call artifacts (60 sites), using SOLiD, 454 and Sanger sequencing. Performance of data from the two platforms was similar by several metrics. Reference bias obfuscated the one wrong monomorphic call of Pacific Biosciences while the single wrongly called polymorphic site was also miscalled in MiSeq and in our gold standard HiSeq whole genome dataset. From the 5 sites miscalled using data from Illumina MiSeq 2 were in agreement with Pacific Biosciences (one listed above and one not called in Pacific Biosciences due to reference bias), and 3 sites were called polymorphic in error due to noise in the MiSeq data. Pacific Biosciences RS data performed well by all metrics, and at a similar quality to Illumina data, demonstrating that the RS is a powerful tool for follow up validation or extension.

	HiSeq	Sequenom	Pacbio	454
somatic	15	6	12	8
wildtype	0	6	1	0
unknown	0	3	2	7

Pacbio correctly identified a false positive in the original dataset
(unknown in sequenom and 454)

Cancer multi-platform validation experiment

The cancer group at the Broad Institute validated 15 somatic mutations previously identified using Illumina HiSeq with Sequenom, 454 and Pacific Biosciences. In this experiment, Pacific Biosciences was key in identifying the one false positive in the original Illumina HiSeq callset and disambiguating the high false negative rate of Sequenom. The 454 experiment only had coverage in half the sites due to unrelated lab issues. The artifacts that caused Illumina HiSeq data to trigger the false variant in the tumor sample was due to incorrect mapping. The longer reads of Pacific Biosciences were correctly placed elsehere and the artifact variation disappeared.

	Ion Torrent	MiSeq
	TMAP	Broad Pipeline	Broad Pipeline
SNPs	2309	352	368
TiTv	0.65	2.63	2.75
OMNI chip concordance	68%	78%	80%
Indels	2108	524	15

	from the **43** sites in HapMap 3.3	from the **225** sites in our Gold Standard dataset
MiSeq called	43	222
PacBio called	38	197

Visual classification	Visual classification	Visual classification
4 sites not called due to reference bias	16 sites not called due to reference bias	3 sites not called due to low coverage
1 site not called due to low coverage	12 sites not called due to low coverage	

Variant discovery multi and single sample

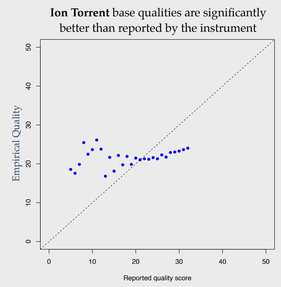
To evaluate Pacific Biosciences and Illumina MiSeq data for variant discovery, we sequenced 177 kb in 61 amplicons from regions across human chromosome 20 of NA12878. These amplicons contained 268 SNPs: 225 that had been validated in our gold standard deep whole genome callset and 43 were present in HapMap. We evaluated the calls made using Ion Torrent and MiSeq data on the 8 samples from the 1000 genomes project assessing the number of calls present in the OMNI chip for those locations and looking at call metrics. The expected number of calls in the region should be around 300 sites and a TiTv >= 2.0. We expect to call approximately 10 real indel events in the region, but do not have a gold standard callset for this region to compare. Results showed that Ion Torrent data clearly benefits from the Broad Institute pipeline due to systematic artifacts generated by the TMap aligner.

Ion Torrent and MiSeq

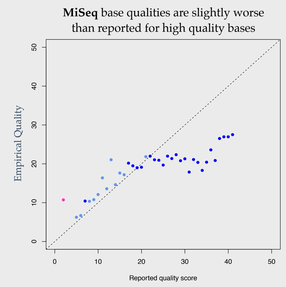
Base qualities and read lengths

Base quality scores are a critical tool for accurate SNP calling, and are used by most analysis algorithms to help distinguish true variation from artifacts. Indeed, the accuracy of the reported base quality scores has a significant impact on the correctness of variation detection. We evaluated the base qualities reported by Ion Torrent and Illumina MiSeq using the GATK base quality score recalibration framework.

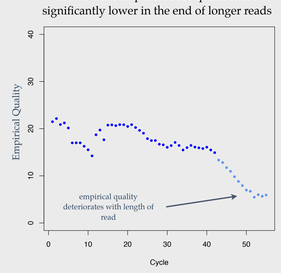
Ion Torrent base qualities are significantly better than reported by the instrument



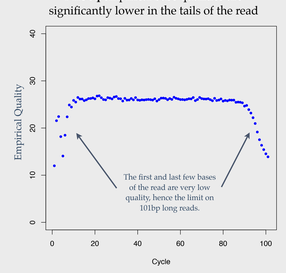
MiSeq base qualities are slightly worse than reported for high quality bases



Ion Torrent empirical base qualities are significantly lower in the end of longer reads

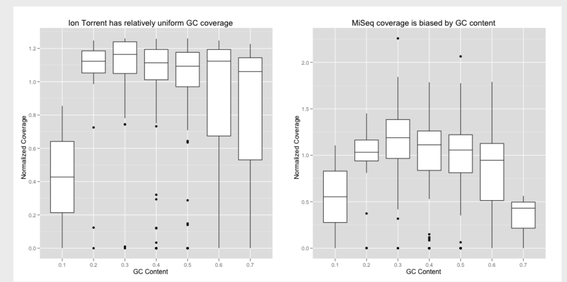


MiSeq empirical base qualities are significantly lower in the tails of the read



Sequence composition-based representational bias

since extreme base composition is a source of representational bias in many sequence data types, we evaluated the performance of Ion Torrent and Illumina MiSeq in a targeted re-sequencing experiment. Ion Torrent showed no GC bias in the normalized coverage plots but Illumina MiSeq displays same deficiency in low and high GC content regions of the human genome as does Illumina's other instrument (HiSeq) with similar chemistry, though great improvement has been made since the earlier versions of the instrument with changes in the chemistry.



Methods

Pacific Biosciences Processing Pipeline

1. We start our processing pipeline with the filtered_subreads.fastq file produced using the Pacific Biosciences SMRT pipeline software.
2. Mapping and Alignment are done using BWA with a heuristic smith waterman algorithm (bwa-sw) and parameters that take advantage of the length of the reads and are more permissive to gap openings and extensions.
3. We add read group information and create a valid BAM file using Picard Tools.
4. We calibrate base qualities using the GATK's Base Quality Score recalibration framework.

Pacific Biosciences has incorporated the pipeline used in this work on the latest version of the SMRT software.

Illumina MiSeq, HiSeq and Ion Torrent Processing Pipeline

HiSeq and MiSeq undergo the standard production pipeline at the Broad Institute. The Production pipeline performs an extra step that has been designed for Illumina data. Indel Realignment is performed before Base Quality Score Recalibration and this tool has a large impact in the quality of the variation (especially indels) identified in these technologies.

Pacific Biosciences would greatly benefit from Indel Realignment, but the current implementation of the tool cannot handle Pacific Biosciences data. We anticipate that future versions of the GATK will address this issue and much improve the reference bias currently plaguing Pacific Biosciences data.

Variant calling for validation and discovery

Variant calling was done using the Genome Analysis Toolkit (GATK) developed at the Broad Institute. For discovery assays, variants were identified using the Unified Genotyper tool of the GATK. For validation assays we used the Genotype And Validate tool of the GATK. The cancer dataset was called using MuTect, a private somatic mutation caller of the GATK. Comparison datasets used for validation were either previously validated using multiple technologies (Sanger, 454, Solid, Complete Genomics, Illumina HiSeq and GA2), or consisted of deep whole genome data of the same sample.