Single molecule sequencing of M13 virus genome without amplification

Luyang Zhao[1], Liwei Deng[1], Gailing Li[1], Huan Jin[1], Jinsen Cai[1], Huan Shang[1], Yan Li[1], Haomin Wu[1], Weibin Xu[1], Lidong Zeng[1], Renli Zhang[2], Huan Zhao[3], Ping Wu[1], Zhiliang Zhou[1], Jiao Zheng[1], Pierre Ezanno[1], Qin Yan[1], Michael W. Deem[4], Jiankui He[5]*

[1]Direct Genomics Co., Ltd., Shenzhen, Guangdong, 518055, China

[2]Reproductive Medical Center of Guangdong General Hospital & Guangdong Academy of Medical Sciences, 106 Zhongshan Er Road, Guangzhou 510080, China.

[3]Shenzhen Armed Police Hospital Reproductive Center, 8 Jinhu Road, Luohu District, Shenzhen 518029, China.

[4]Departments of Bioengineering and Physics & Astronomy, Rice University, Houston, TX, 77005, USA

[5]Department of Biology, South University of Science and Technology of China, Shenzhen, Guangdong, 518058, China

*corresponding: hejk@sustc.edu.cn

ABSTRACT

Third generation sequencing is a direct measurement of DNA/RNA sequences at the single molecule level without amplification. In this study, we report sequencing of the genome of the M13 virus by a new single molecule sequencing platform. Our platform detects single molecule fluorescence by the total internal reflection microscope technique, with sequencing-by-synthesis chemistry. We sequenced the genome of M13 to a depth of 316x and 100% coverage. The consensus sequence accuracy is 100%. We demonstrated that single molecule sequencing has no significant GC bias.

INTRODUCTION

Since the first two publications of the human genome sequences [1, 2], scientists around the world have embarked on a quest for next-generation sequencing (NGS) technologies. The resulting progress has revolutionized fields ranging from academic research to clinic diagnosis [3, 4]. Applications in the field of precision medicine (see review [5] for precise definition) include cancer diagnosis [6, 7] and inherited disease diagnosis [8, 9]. Progress in NGS technologies has brought, ethical questions as well [10, 11]. Promising applications include detection of pathogenic organisms [12, 13] and forensic sciences [14, 15]. At present, the cost of the sample library preparation process for NGS is still a significant part of the total cost of genome sequencing. Simple operation, cost effective sample preparation, generation of high throughput data, and more sensitive instruments are key requirements of the sequencing market in the future.

Since the early stages of DNA sequencing, single molecule (SM) sequencing is a key technological development. SM sequencing was first experimented in late 80s [16] and is now seen as the next step in the evolution of NGS [17]. Different SM sequencing technologies have rapidly developed over the past decade, with progress on read length, sequencing time, and data throughput. Principles of these technologies exhibit notable differences. Three companies and their own technologies are now well known: (i) the first true single molecule sequencing (tSMS) combining with sequencing-by-synthesis (SBS) [18] technology from Helicos Biosciences [19, 20] which is the technology we are improving; (ii) single molecule real time (SMRT) sequencing technology from Pacific Biosciences which provides super long read length (longer than 10k bases [21, 22]), but relatively low throughput; and (iii) Oxford Nanopore platform based on the direct electrical detection of single DNA molecule through α-hemolysin nanopores on which surface exonuclease enzyme molecules are attached [23], which provides long read (6k bases [21]) but limited accuracy and low throughput.

Despite the advantages of NGS platforms, the preparation of DNA libraries generally requires a preliminary step based on PCR amplification. This process introduces bias and ultimately can result in wrong interpretation of raw data sets [24, 25]. The Illumina sequencing platform, with which

most current sequencing is performed, produces data sets showing uneven coverage and serious defects in GC-poor or GC-rich regions. Low coverage regions could be interpreted as sequencing errors by most current assemblers [26], and high coverage regions could be interpreted as repetitive sequences [27, 28]. Much effort has gone into improving protocols of library preparation to reduce or fully suppress GC bias [29, 30].

Advantages of the sample preparation for SM sequencing, combined with massively parallel short reads covalently captured onto an engineered surface, ideally fit the requirements of clinical diagnosis using DNA sequencing. Advantages of SM sequencing include (i) a simple and time−saving sample preparation consisting briefly of DNA shearing followed by poly-A tailing and 3' end blocking steps, (ii) absence of base substitution introduced by the limited fidelity of DNA polymerases routinely used in PCR to amplify genomic DNA in samples, and (iii) the possibility of sequencing RNA molecules as well as DNA in order to investigate transcriptomic aspects of gene expression.

Our approach is devised to provide simple operation and high-throughput, unbiased data. Recently, we have demonstrated a direct targeted sequencing of cancer related gene mutations at the SM level [31]. In this paper, we describe the performance of our new GenoCare platform for SM sequencing without preliminary PCR amplification. The vector M13mp18 whose sequence is derived from the genome of the bacteriophage M13 was sequenced to the depth of 316x with 100% coverage. More importantly, no significant GC bias was observed.

EXPERIMENTAL CONSIDERATIONS

SAMPLE PREPARATION

M13: M13mp18 cloning vector was purchased from NEB, Beijing, China, and used as received. The sequence of the M13mp18 cloning vector is derived from the M13 phage [32] and contains 7249 bp. In this study, we used this cloning vector as DNA raw material to re-sequence, analyze, and compare with the reference sequence.

Oligonucleotide Primers: 5’ amine functionalized Poly-T oligonucleotides were purchased from Sangon and used as received.

M13 genomic DNA preparation process was illustrated in **Figure 1**.

1. DNA fragmentation

The M13mp18 cloning vector (from NEB, ref. N4018S) was used as raw DNA material to be sequenced by our platform. This cloning vector was first randomly fragmented into dsDNA fragments of about 200 bp using NEBNext® dsDNA Fragmentase® (from NEB, ref M0348S). Then, DNA fragments were purified using Agencourt AMPure XP beads (from Beckman, ref. A63881). The concentration of DNA was assessed by UV absorption using a Nanodrop 2000 device.

2. Poly-A tailing and blocking

Multiple incorporations of 50-100 dATP at the 3' end of ssDNA fragments from the cloning vector resulted in a poly-A tail. This reaction completed within 20 minutes. In a second step, poly-A tailed 3' ends were blocked by incorporating the Cyanine 3 dideoxy ATP (Cy3-ddATP from PERKINELMER, ref. NEL586001EA). The blocking reaction completed within 30 minutes using the enzyme Terminal Transferase (from NEB, ref. M0315) such that the incorporation of reversible terminators at the 3' end of the template strands was prevented.

SURFACES AND TEMPLATE CAPTURE

Surface Chemistry: Sequencing surfaces were prepared on 110×74 mm epoxy-coated glass coverslips (SCHOTT, Jena, Germany). Poly-T oligonucleotides were covalently bond to surface.

Flow Cells: The above functionalized glass coverslip was assembled with a 1.0 mm thick glass slide by a pressure sensitive adhesive to form a flow cell. The flow cell has 16 channels, determined by the adhesive shape. For the M13 sequencing in this experiment, ~0.5% of one channel was imaged.

Template Capture (Hybridization): The surface of the flow-cell was chemically modified by anchoring poly-T ssDNA strands at their 5' end, in order to capture poly-A tailed strands from the library once they were injected inside the flow-cell at 55 °C. Then non-hybridized templates were washed away by 150 mM HEPES, 1X SSC and 0.1% SDS, followed by 150 mM HEPES and 150 mM NaCl.

SEQUENCING REACTIONS

The GenoCare Platform: All the sequencing reactions were implemented on the GenoCare platform The GenoCare is an automated single molecule sequencer with three major components: fluorescence imaging system, microfluidic system, and the stage to control the movement of sample. The imaging system is based on total internal reflection fluorescence (TIRF) microscopy [31].

Fill & Lock: Since the hybridization of poly-T primer with poly-A tailed template may not be perfect, a step to fill the remaining dATP on the template with dTTP before the real sequencing process starts is necessary. After hybridization, the temperature of the flow-cell was lowered to 37 °C. The unpaired adenine nucleotides of poly-A tailed template strand were paired by multiple incorporations of natural thymine nucleotides at the 3' end of primer strands. A mixture of dATP, dCTP, and dGTP reversible terminators were added to block further incorporation so that the template was locked in place and ready for sequencing.

Nucleotide Addition: Reversible terminators were adopted in the sequencing-by-synthesis approach. They are modified nucleotides, which are composed of nucleotide triphosphates, a fluorophore (Atto647N), disulfide linker, and an inhibitor group. The design of the inhibitor effectively blocks the incorporation of next nucleotide before the cleavage of the disulfide bond of the previous reversible terminator.

The DNA extension was carried out at 37℃ in Tris buffer containing polymerase, one of the four nucleotides and other salts. The components of this system are available with the use instructions from Direct Genomics.

RESULTS AND DISCUSSION

Sequencing Process. Our sequencing-by-synthesis (SBS) scheme is shown in **Figure 1**. Sample preparation is simple and fast, especially without amplification. M13 genomic DNA was sheared into fragments of ~200bp, a length of 50~100nt poly-A was added to the end, and blocked by ddATP-Cy3. Sequencing surfaces were chemically modified and covalently bound with poly-T, which can be hybridized with target DNA. Once annealed, residual dATP were filled with natural nucleotides, and locked with one reversible terminator. The sequencing of target DNA was then ready to commence.

The single molecule SBS process has been described elsewhere [31] Each cycle includes terminator incorporation, imaging, cleavage of fluorophore, and capping of residual bonds. The GenoCare platform adopts the total internal fluorescence microscopy (TIRF) for the observation of single molecules. The integration time was 200 ms to guarantee a good signal-to-noise ratio and reduce the photobleaching of dyes. We used 0.5% of one flow-cell channel to resequence the M13 virus

genome as a demonstration of the GenoCare performance. We sequenced 80 cycles (40 quads of CTAG), and the images were analyzed (**Supporting Information**) to reach 100% coverage and an average depth of 316x for each base. The instrument run time was 9 hours, and sample preparation took 3 hours.

Genome Coverage. About 100,000 reads were uniquely aligned to the reference genome, which accounted for 25.4% of the total reads. Reads were filtered by several criteria: 1) reads that were less than 13 bases after alignment were discarded, 2) reads that included a sequence exactly matching the terminator addition order and indicative of non-specific adsorption were also discarded, and 3) reads that could be mapped to multiple locations on the reference genome were excluded. From these data, we can determine the error rate. As is shown in **Table 1**, the dominant error was deletion (1.65%), followed by insertion (0.78%) and substitution (0.69%). Most of the reads aligned perfectly to the reference without any error, and the most error allowed per sequence by our algorithm was 3 (**Figure S1**). The average coverage depth for each base is 316x, and the minimum coverage is 14x (**Figure 2a**). When the average coverage depth reaches 10x, genome coverage rate climbs to 100% (**Figure 2b**). The Integrative Genomics Viewer (IGV) gives a clear picture of mapping against the known M13 genome reference (**Figure S2**).

Read Length. The read length for this M13 sequencing run is shown in **Figure 3**. With the sequences less than 13 bases discarded due to poor unique mapping rate, this average read length is 22 bases (**Table 1**), given that only 80 base incorporation cycles were conducted. For the 7.2 kb M13 genome, a read length of 22 bases gives more than adequate specificity for alignment. Before alignment, the length distribution shows a peak at around 25 bases, while many reads were lost due to errors and non-uniqueness of mapping during alignment, causing a decrease of throughput and lowering of the average read length.

GC Bias. In accordance with the predicted effect of a PCR-free sample preparation, no obvious GC bias was observed under a window of 100 bases in which the GC content fluctuates in the range 22-69% (**Figure 4a**). The y-axis is an average of coverage depth in all 100-base windows with the same GC percentage. The distribution of base frequency in the reference as function of the CG content shows an almost identical shape to the depth distribution calculated from the sequencing result (**Figure 4b and Supporting Information**) and the R^2 (goodness of fit) of those two curves is 0.9946, which indicates that no coverage bias is observed in this experiment.

CONCLUSION

In this paper, we demonstrated the new GenoCare platform for single molecule sequencing. GenoCare is an automated desktop type of sequencer for dedicated use in the clinic. Compared to traditional next generation sequencing, sample preparation of single molecule sequencing is simpler and faster. Most importantly it does not require the use of PCR amplification, which effectively limits the GC bias. For example, on the Illumina system, a major NGS platform, it has been reported that GC bias leads to an uneven coverage or even no coverage of reads across the genome.

The cloning vector M13mp18 was sequenced on this new platform. A total of 80 cycles was run. Overall sequencing took 12 hours including sample preparation, instrument run time and data analysis. Eventually an average of 316x coverage depth and 22 bases read length were achieved. The consensus accuracy reached 100% once each base was sequenced at least 10 times. There was no apparent GC bias observed in this experiment, demonstrating the advantage of single molecule sequencing.

REFERENCES

[1] E. S. Lander *et al.*, “Initial sequencing and analysis of the human genome,” *Nature*, vol. 409, no. 6822, pp. 860–921, Feb. 2001.

[2] J. C. Venter *et al.*, “The sequence of the human genome,” *Science*, vol. 291, no. 5507, pp. 1304–1351, Feb. 2001.

[3] K. V. Voelkerding, S. A. Dames, and J. D. Durtschi, “Next-generation sequencing: from basic research to diagnostics,” *Clin. Chem.*, vol. 55, no. 4, pp. 641–658, Apr. 2009.

[4] J. Zhang, R. Chiodini, A. Badr, and G. Zhang, “The impact of next-generation sequencing on genomics,” *J. Genet. Genomics Yi Chuan Xue Bao*, vol. 38, no. 3, pp. 95–109, Mar. 2011.

[5] D. Roden and R. Tyndale, “Genomic Medicine, Precision Medicine, Personalized Medicine: What’s in a Name?,” *Clin. Pharmacol. Ther.*, vol. 94, no. 2, pp. 169–172, Aug. 2013.

[6] L. Dong *et al.*, “Clinical Next Generation Sequencing for Precision Medicine in Cancer,” *Curr. Genomics*, vol. 16, no. 4, pp. 253–263, Aug. 2015.

[7] Y. Xue and W. R. Wilcox, “Changing paradigm of cancer therapy: precision medicine by next-generation sequencing,” *Cancer Biol. Med.*, vol. 13, no. 1, pp. 12–18, Mar. 2016.

[8] W. Zhang, H. Cui, and L.-J. C. Wong, “Application of next generation sequencing to molecular diagnosis of inherited diseases,” *Top. Curr. Chem.*, vol. 336, pp. 19–45, 2014.

[9] H. Daoud *et al.*, “Next-generation sequencing for diagnosis of rare diseases in the neonatal intensive care unit,” *CMAJ Can. Med. Assoc. J.*, vol. 188, no. 11, pp. E254–E260, Aug. 2016.

[10] A. J. Clarke, “Managing the ethical challenges of next-generation sequencing in genomic medicine,” *Br. Med. Bull.*, vol. 111, no. 1, pp. 17–30, Sep. 2014.

[11] K. Sénécal, K. Thys, D. F. Vears, K. Van Assche, B. M. Knoppers, and P. Borry, “Legal approaches regarding health-care decisions involving minors: implications for next-generation sequencing,” *Eur. J. Hum. Genet.*, vol. 24, no. 11, pp. 1559–1564, Nov. 2016.

[12] R. R. Miller, V. Montoya, J. L. Gardy, D. M. Patrick, and P. Tang, “Metagenomics for pathogen detection in public health,” *Genome Med.*, vol. 5, no. 9, p. 81, Sep. 2013.

[13] F. Thorburn, S. Bennett, S. Modha, D. Murdoch, R. Gunson, and P. R. Murcia, “The use of next generation sequencing in the diagnosis and typing of respiratory infections,” *J. Clin. Virol.*, vol. 69, pp. 96–100, Aug. 2015.

[14] S. M. Aly and D. M. Sabri, “Next generation sequencing (NGS): a golden tool in forensic toolkit,” *Arch. Med. Sadowej Kryminol.*, vol. 65, no. 4, pp. 260–271, 2015.

[15] C. Børsting and N. Morling, “Next generation sequencing and its applications in forensic genetics,” *Forensic Sci. Int. Genet.*, vol. 18, pp. 78–89, Sep. 2015.

[16] J. H. Jett *et al.*, “High-speed DNA sequencing: an approach based upon fluorescence detection of single molecules,” *J. Biomol. Struct. Dyn.*, vol. 7, no. 2, pp. 301–309, Oct. 1989.

[17] L. T. Sam *et al.*, “A Comparison of Single Molecule and Amplification Based Sequencing of Cancer Transcriptomes,” *PLOS ONE*, vol. 6, no. 3, p. e17305, Mar.

2011.
[18] I. Braslavsky, B. Hebert, E. Kartalov, and S. R. Quake, “Sequence information can be obtained from single DNA molecules,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 100, no. 7, pp. 3960–3964, Apr. 2003.
[19] J. F. Thompson and K. E. Steinmann, “Single molecule sequencing with a HeliScope genetic analysis system,” *Curr. Protoc. Mol. Biol. Ed. Frederick M Ausubel Al*, vol. Chapter 7, p. Unit7.10, Oct. 2010.
[20] P. Milos, “Helicos BioSciences,” *Pharmacogenomics*, vol. 9, no. 4, pp. 477–480, Apr. 2008.
[21] J. A. Reuter, D. Spacek, and M. P. Snyder, “High-Throughput Sequencing Technologies,” *Mol. Cell*, vol. 58, no. 4, pp. 586–597, May 2015.
[22] H. P. J. Buermans and J. T. den Dunnen, “Next generation sequencing technology: Advances and applications,” *Biochim. Biophys. Acta*, vol. 1842, no. 10, pp. 1932–1941, Oct. 2014.
[23] H. Lu, F. Giordano, and Z. Ning, “Oxford Nanopore MinION Sequencing and Genome Assembly,” *Genomics Proteomics Bioinformatics*, vol. 14, no. 5, pp. 265–279, Oct. 2016.
[24] E. L. van Dijk, Y. Jaszczyszyn, and C. Thermes, “Library preparation methods for next-generation sequencing: tone down the bias,” *Exp. Cell Res.*, vol. 322, no. 1, pp. 12–20, Mar. 2014.
[25] Y.-C. Chen, T. Liu, C.-H. Yu, T.-Y. Chiang, and C.-C. Hwang, “Effects of GC Bias in Next-Generation-Sequencing Data on De Novo Genome Assembly,” *PLOS ONE*, vol. 8, no. 4, p. e62856, Apr. 2013.
[26] H. Chitsaz *et al.*, “Efficient de novo assembly of single-cell bacterial genomes from short-read data sets,” *Nat. Biotechnol.*, vol. 29, no. 10, pp. 915–921, Sep. 2011.
[27] J. C. Dohm, C. Lottaz, T. Borodina, and H. Himmelbauer, “Substantial biases in ultra-short read data sets from high-throughput DNA sequencing,” *Nucleic Acids Res.*, vol. 36, no. 16, p. e105, Sep. 2008.
[28] D. Aird *et al.*, “Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries,” *Genome Biol.*, vol. 12, no. 2, p. R18, 2011.
[29] S. O. Oyola *et al.*, “Optimizing Illumina next-generation sequencing library preparation for extremely AT-biased genomes,” *BMC Genomics*, vol. 13, p. 1, Jan. 2012.
[30] I. Kozarewa, Z. Ning, M. A. Quail, M. J. Sanders, M. Berriman, and D. J. Turner, “Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes,” *Nat. Methods*, vol. 6, no. 4, pp. 291–295, Apr. 2009.
[31] Y. Gao *et al.*, “Single molecule targeted sequencing for cancer gene mutation detection,” *Sci. Rep.*, vol. 6, p. 26110, May 2016.
[32] C. Yanisch-Perron, J. Vieira, and J. Messing, “Improved M13 phage cloning vectors and host strains: nucleotide sequences of the M13mp18 and pUC19 vectors,” *Gene*, vol. 33, no. 1, pp. 103–119, 1985.

Figures and Captions

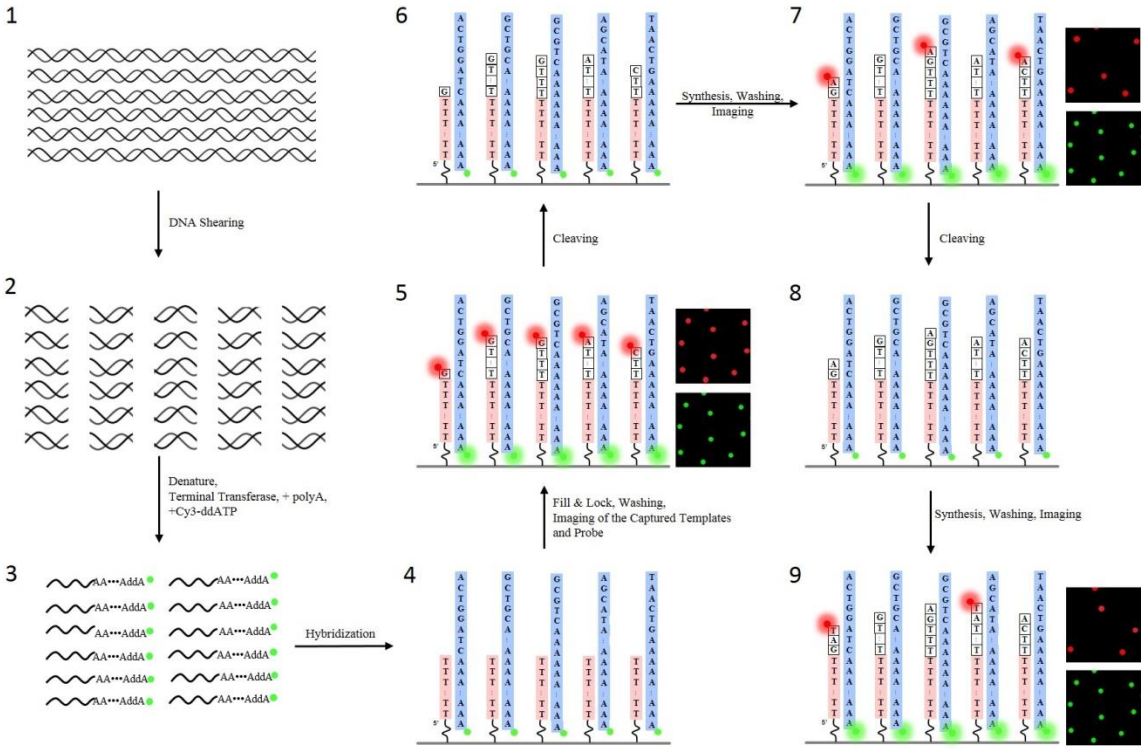

Figure 1. Sample preparation and sequencing process for single molecule sequencing of biological samples.

Table 1. M13 genome sequencing statistics

Cycles Sequenced	Average read length	Coverage			Total Reads	Mapped Reads	Unique Mapped Reads	Unique Mapped Ratio	Sub. Rate	Del. Rate	Ins. Rate
		Average	Max.	Min.							
80	22 bases	316x	717x	18x	409491	104802	103990	25.4%	0.69%	1.65%	0.78%

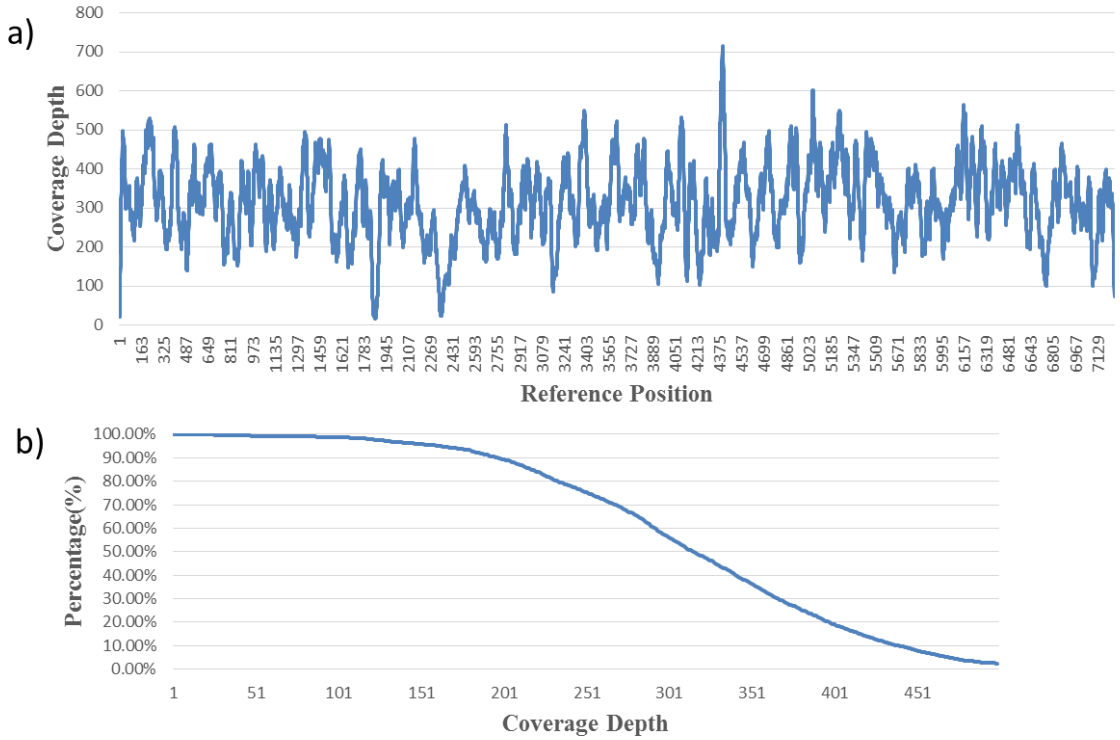

Figure 2. (a) Coverage depth for each base on M13 reference. The average coverage depth is 316x±96x. (b) Coverage rate as a function of coverage depth. 100% coverage was achieved when average coverage depth reached 10X.

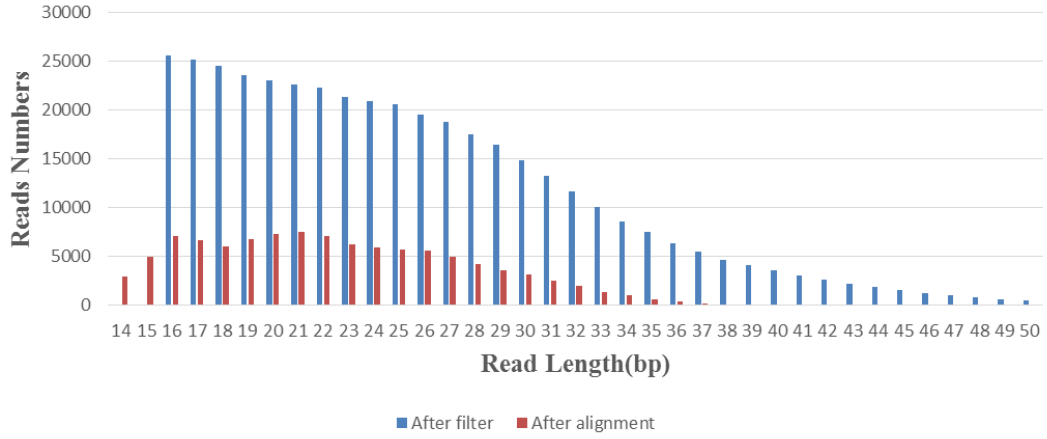

Figure 3. Read length distribution after length and repeat filters (blue bars) and after alignment (red bars).

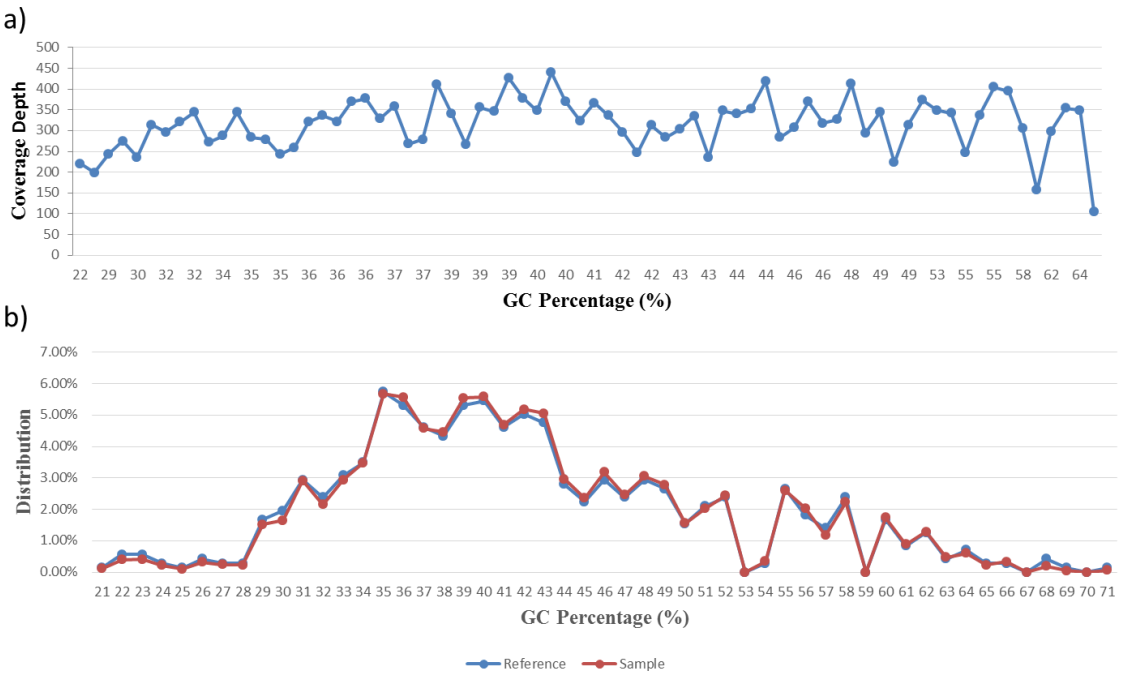

Figure 4. (a) Average depth distribution of all 100-base windows as a function of GC content. From GC content 22% to 69%, the average depth of each window on genome fluctuates in a small range. (b) GC patterns of reference genome (blue curve) and aligned reads (red curve).