

A brief utilization report on the Illumina HiSeq 2000 sequencer

Lin Liu*, Ni Hu, Bo Wang, Minfeng Chen, Juan Wang, Zhijian Tian, Yimin He and Danni Lin

Beijing Genomics Institute (BGI), 4th floor, Building #11, Beishan Industrial Zone, Yantian District, Shenzhen 518083, Guangdong Province, China

(Received 13 August 2011)

As the first and largest user of next-generation Illumina HiSeq 2000 (HiSeq) sequencers in the world, Beijing Genomics Institute (BGI) has extensive experience utilizing HiSeq, covering almost all sequencing areas, including human, animal, plant and microbe, and almost all sequencing types, including *de novo* sequencing, resequencing, metagenome sequencing, RNA-Seq, Digital Gene Express. BGI has managed to convert the full sequencing power from 20 Genome Analyzer II (GA) to, currently, 100 HiSeq 2000 sequencers, with the useful supplementation of 15 SOLiD™ 4 systems from Applied Biosystems. A brief evaluation and comparison of the two Illumina sequencing systems is provided in this manuscript, including library construction, quality control, multiplexing and automation, as well as troubleshooting guidelines and a few demo cases utilizing HiSeq 2000 at BGI.

Keywords: next-generation sequencing; genome; sequencer; HiSeq; troubleshooting; quality control

Introduction

Genomic sequencing has enabled us to find out the most fundamental information of life. Genomics is a cutting-edge and rapidly advancing area of research, with scientists fiercely competing to develop a better understanding of sequencing methods and bioinformatics. Numerous species have been sequenced which has facilitated seed incubation, personal genome and medicine, GWAS population disease study, evolution, as well as finding new species and target drugs. Well-known genome projects include the 1000 Genomes Project (2007), 1000 Plant and Animal Reference Genomes Project (www.genomics.cn/en/research.php?type=show&id=497) and the Earth Microbiome Project (Gilbert et al., 2010).

As early as in 1970s, two methods were developed independently by two groups – Maxam and Gilbert, and Sanger and Coulson (Feldman 2001) – with both teams receiving the 1980 Nobel Prize. The traditional Sanger–Coulson capillary sequencing technology, also known as dideoxy chain-termination technology, was widely applied in sequencing cloning vectors, PCR products and even shot-gun sequencing for large genomes such as human. It uses a single DNA chain as a template, with the addition of dNTP and fluorescent ddNTP in PCR amplification, to get DNA products of various lengths, followed by electrophoresis separation by size and induced fluorescence detection of the bases. Applied Biosystems introduced the first automatic sequencing machine (named AB370) onto market in 1986 (Figure 1). This automatic and integrated machine

adopted capillary electrophoresis instead of agarose gel electrophoresis, making the sequencing results more accuracy and robust. It could detect 96 bases at a time, 500,000 bases in one day, and the read length could reach 600 bases. Nevertheless, it had clear disadvantages, such as the size of the machine, long analysis time and low yield.

Due to the growth in applications and the urgent need of powerful sequencing technologies, next generation sequencing (NGS), also referred to as high-throughput sequencing (HTS), became available in the 2000s. The main advantages of NGS compared with traditional sequencing technologies are high throughput, high speed, less requirements in the amount of input samples and preparation before sequencing, and a more affordable price. All NGS systems share some additional features: (1) they require library fragments to be amplified to form multiple copies on a solid phase (glass slide or microbead); (2) sequencing reactions take place in repeating steps and signals are collected and analyzed automatically; (3) they have both paired end and mate pair sequencing. This development triggered the widespread use and application of sequencing and analysis in areas such as personal genome, personal medicine, RNA expression analysis, methylation, etc. Next generation sequencers are typically represented by three companies: Illumina, Applied Biosystem (SOLiD) and Roche454. Currently, the HiSeq 2000, developed by Illumina, can reach 600 G/run with improved truseq-v3 reagents, and promises to produce

*Corresponding author. Email: liulin@genomics.org.cn

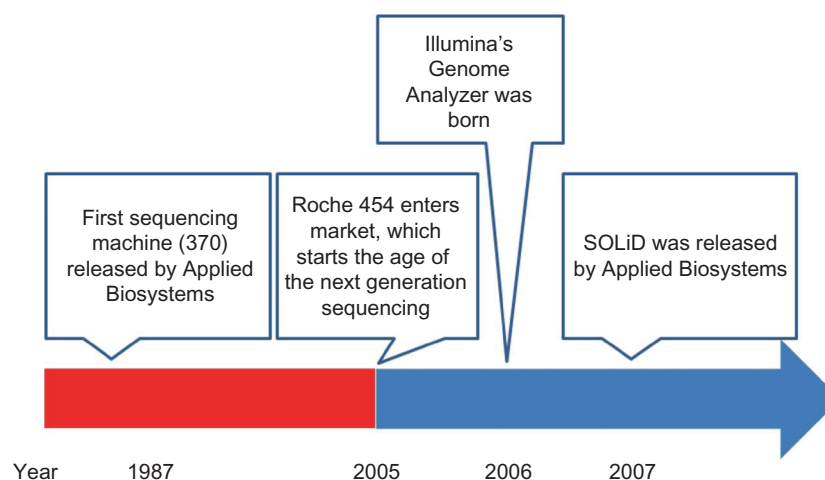


Figure 1. The developmental process of sequencing systems. First sequencing machine (370) was released by Applied Biosystems in 1986 and it led to the era of first generation sequencing. In 2005, Roche 454 entered market and it started the era of next generation sequencing. In 2006 Illumina's genome analyzer was born and SOLiD was released by Applied Biosystems in the following year.

a 1-T/run yield by the end of 2011. Read length and the single read error rate, however, is not comparable to the traditional Sanger–Coulson method, with NGS generally capable of reaching a 100–250-bp read length and an error rate less than 1%, which may cause problems in *de novo* assembly. Software has been developed to overcome this problem, with SOAP (Li et al., 2008; Wang et al., 2008), developed by BGI, is one of the most widely used tool. Many software packages are shared by the scientific community and can be found free on the internet at the National Center for Biotechnology Information (www.ncbi.nlm.nih.gov), Broad Institute (www.broadinstitute.org) or Wellcome Trust Sanger Institute (www.sanger.ac.uk, <http://www.genomesonline.org/>).

Scientists occasionally refer to new, emerging real-time single-molecular sequencing technology as NGS, represented by Pacific Biosciences and Helicos BioSciences (Thompson and Steinmann, 2010; Pacific Biosciences 2010); otherwise, it can be considered third-generation. The key difference between NGS and single-molecular sequencing technology is that, in NGS, DNA molecules have to be amplified before sequencing, while the latter enables us to sequence a single DNA molecule; thus, avoid the time, error and cost associated with sample preparation and amplification. Single Molecule Real Time (SMRT™) DNA sequencing from Pacific Biosciences is used as an example to illustrate how single molecule sequencing works. It will be detailed later in this report along with its advantages.

NGS technology was first developed by Roche in 2005, based on the principle of pyrosequencing used in the 454 sequencing system. The principle is that DNA fragments, which are ligated to the 454-specific adapters, are amplified with emulsion PCR using amplification beads

before sequencing. The beads, containing amplified fragments, are then loaded into the PicoTiterPlate device for sequencing. Polymerase incorporates the nucleotides complementary to the template strand of DNA. When each nucleotide is added, a fluorescent signal is generated and the strength of the signal, recorded by CCD camera, enables analysis of the base sequence. The latest development from Roche is the GS FLX Titanium series, which extends the read length from 400 to 1000 (Shendure and Ji, 2008; La Scola et al., 2008).

Illumina released their first NGS instrument, the Solexa GA, in 2006, recording a yield of 20 G/run. Through improvements in polymerase in cluster amplification, and upgraded hardware and software, the yield was improved to 50 G/run in 2009. The latest 150PE GAIIX series can attain 85 G/run. The HiSeq 2000 has been available since 2010 and, at the launch, the number of cycles per run was aimed to reach 209 with a throughput of 200 G/run (Holt and Jones, 2008; MacLean et al., 2009).

The principle of sequencing is as follows: DNA fragments are loaded onto a glass flow-cell (FC) containing specific adapters for DNA fragments to attach to and are then amplified using bridge PCR in the Cluster Station (CS). The flow-cell, containing the amplified DNA fragments (clusters), sequences the DNA using a robust four-color DNA sequencing-by-synthesis technology. With the help of polymerase, fluorescence-labeled nucleotides are incorporated onto the strand. The 3'OH bonds of the nucleotides are inactivated to avoid incorporation of more than one base at one cycle. Images are taken to identify the corresponding base, after the addition of each nucleotide. The fluorescent is then removed and the 3'-end is unblocked to enable the next incorporation (Holt and Jones, 2008).

Table 1. Comparison of NGS platforms.

	454 Sequencing	Illumina	SOLiD™ TM
Latest System	GS FLX Titanium series	HiSeq 2000	SOLiD™ 5500 Systems
Sequencing chemistry	Pyrosequencing	Sequence by synthesis	Sequencing by ligation
Amplification approach	Emulsion PCR	Bridge amplification	Emulsion PCR
Read length	1000 bp	209 bp	150 bp
Error rate	~ 1%	> 1%	~0.01%
Run Time per run	~ 10 h	13 days	~ 7 days
Throughput per run	400 – 600 Mb	~ 600 Gb	~ 300 Gb

SOLiD™ was released by Applied Biosystems in 2007. Like 454 pyrosequencing, DNA fragments are amplified with emulsion PCR before sequencing. The beads are then immobilized onto a flow-cell for sequencing. Universal sequencing primers are annealed to the fragments and then fluorescently labeled di-base probes are ligated to the fragments with the help of DNA ligase. Another round of sequencing begins with the universal primer attaching to the $n-1$, $n-2$, $n-3$, . . . position. This allows for proofreading of the sequencing data. Thus, the accuracy of SOLiD™ Systems can reach 99.94%, while the newly developing SOLiD™ 5500 and SOLiD™ 5500xl Systems, which became available in the late 2010, can reach 99.99% accuracy (Margulies et al., 2005; Mardis, 2008). Analysis software was also upgraded to BioScope from COLOUR for convenience and more powerful pipeline analysis. The second generation sequencing technologies are summarized in Table 1.

SMRT™ DNA sequencing is based on the observation of each single DNA polymerase molecule working in a continuous, progressive manner. This technology contains three components: the SMRT cell, the phospholinked nucleotides and the detection platform. The SMRT cell contains thousands of zero-mode waveguides (ZMWs), which act as nanophotonic visualization chambers. Each ZMW allows only a single DNA polymerase molecule to attach to the bottom surface permanently. The phospholinked nucleotides, each type labeled with a different colored fluorophore, are added before the start of the reaction. Because the size of the ZMW is very small and the DNA polymerase incorporates multiple bases within a second, the detection platform has been specifically designed for this technology.

During the reaction, the DNA polymerase is incorporated with the complementary nucleotides, and the engaged fluorophore emits fluorescent light whose color corresponds to the base identity. Once incorporation is completed, the polymerase cleaves the bond holding the fluorophore and the nucleotide, leading to the dye diffusing out of the detection volume. After incorporation, the signal immediately returns to the baseline and the process is repeated.

This new technology has several advantages compared to the next generation sequencing. Read length of the

SMRT™ DNA sequencing is more than double that of the next generation sequencing. The sample needed for running SMRT™ DNA sequencing is ~1 ng of DNA, which suggests that no PCR amplification is needed. This also lowers the risk of having systematic amplification bias. Single molecule, real-time analysis allows the system to observe the structural and cell-type variations not accessible with next generation sequencing. It also has the ability to observe and capture kinetic information, which means it can measure and access the dynamics and timing of enzymatic incorporation. For example, direct detection of DNA methylation (without bisulfite conversion) can be done by monitoring the arrival and duration of the resulting fluorescence pulses, thus yielding information about polymerase kinetics. The power of this technology can be further extended by changing the chemistry without modifying the machinery, such as detecting mRNA translation by the ribosomal complex immobilized to the ZMW and adding a different fluorophore-labeled tRNA (Gupta, 2008; Flusberg et al., 2010).

Comparison of GA and HiSeq

At the beginning of 2010, Illumina launched a new sequencer, the HiSeq 2000. Although sequencing chemistry and methodology is the same as in the GA, it has notable improvements in yield, reagent packing, software and operations, but the quality parameters are not directly comparable due to significant changes in hardware structure and function. For example, optic bomb and light routine changes in the HiSeq improved the initial intensity level qualification to 2500 in read1, which is relatively similar to the sequencing quality of 1500 in the GA, and there are two more boosts of intensity in the 100PE run in the HiSeq. Over a 1-year application experience, BGI has achieved an average 270 G/run yield at the end of 2010, with a sequencing quality success rate and sequencer use saturation rate of >85%. Sequencing quality pass is defined as an error rate of <2% in the 100PE, a GC bias within 1.5% and a yield within 30% variation of expectations. Various aspects of the GA and HiSeq are compared below.

Yield

Compared to the GA (Table 2), the HiSeq has an available surface area in the flow-cell (FC) that is more than five times in the GA, including the dual surface. At its launch, the recommended cluster density per mm square was far less than the GA. Over the years, problems with reagents (NaOH, flow-cell, LFN, etc.) and software defects (bugs, focus problems, etc.) associated with quality problems of Illumina products have delayed the process and speed of cluster density improvement. It took BGI a year to improve the yield from 200 to 270 G, with an improvement in cluster density per unit area of at least 20% over the GA, without introducing samples or data bias. Currently, with the improved FC surface area and upgraded software, which allows recognition of more clusters, BGI has achieved an average 600 G/run yield in human resequencing samples. Within the “road map” declared by Illumina, and with the aid of truseq-v3 cluster generation and SBS reagents, which allow 150PE prolonged sequencing, by the end of 2011, the HiSeq yield is promised to reach 1 T/run, as human genome sequencing costs continue to drop from \$10K to \$1000 (Figure 2).

HiSeq operation

The simplified operation of the HiSeq also benefited from signal capture by scanning (Table 3): a 100PE single flow-cell run takes 8 days in HiSeq (13 days for a dual flow-cell

run) instead of 10 days in GA. In GA, area imaging uses the TIRF scanning method and only one CCD for imaging. HiSeq, on the other hand, uses epi-fluorescence and TDI (time delay integration) line-scanning technology with four CCDs for imaging. This saves imaging time and takes ~12.5 min/surface for imaging. Furthermore, HiSeq can perform dual FC alternation of imaging and chemistry, so that it is more efficient (User’s guide available from <https://icom.illumina.com/download/summary/>). It is simple to load the flow cells in the HiSeq, compared to the need for prism and oil in GA. The flow-cells are held by vacuum and a LED switch ensures the correct connection.

HiSeq reagent

Compared with GA reagents, the CGK kits for cluster generation in the cBot are pre-configured and plug-and-play reagents, which are packed in a 96-plate. Only one manifold is used in cBot instead of multiple, and the HiSeq has both 50- and 200-cycles SBS kits for convenient operation of various sequencing lengths. Other improvements include barcoding, color-coded packing and a customer-friendly reagent quality control (QC) and order-tracking website under Illumina, which will be released soon. With same cluster density in GA and HiSeq, we have observed that the sample load amount was less in HiSeq than GA.

Table 2. Comparison of the HiSeq and GA.

	HiSeq (at 300G)	GA IIx (at 50G)	GA IIx (at >85G)
Available surface area (mm ²)	2880 (1.7 mm FC)	510	510
Tile (mm ²)	5.7	0.531	0.531
Tile number/lane	32	120	120
FC surfaces	Dual	One	One
FC number/run	Dual	One	One
Gb per run	300	50	95
Gb per day	37.5	5	7
PF cluster density (K/mm ²)	600	490	620
Read length	2×100	2×100	2×150

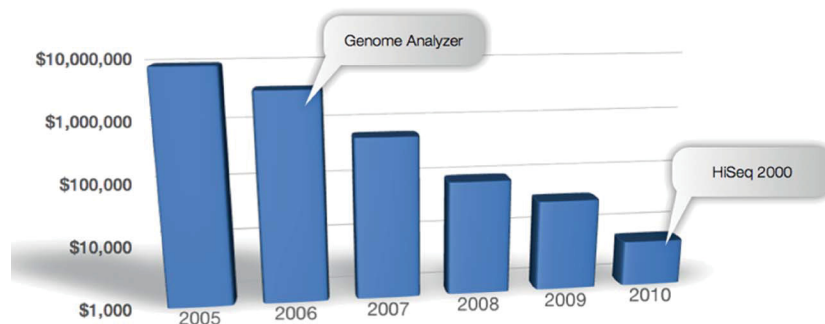


Figure 2. The cost of human resequencing using GA and HiSeq. Price per 30× human genome is shown. In 2010, with HiSeq 2000, the cost per human genome dropped below \$10K. It is expected that by the end of 2011, when HiSeq is upgraded, to yield 1T/run and the cost per human genome could reach \$1000.

Table 3. Advantages of line imaging.

	Area imaging	Line imaging
Stages & filter movement delays	–	+
Data transfer delays	–	+
Practical data acquisition limit	+	+
Data quality/background rejection	–	+

Although the basic reagent chemistry and manufacturing pipeline of GA and HiSeq has not changed greatly, reagent QC in the manufacturing and delivery process has proven problematic throughout 2010 and some problems took up to six months to be resolved, particularly NaOH acidification, LFN purity, the FC dark area, debris contamination and the FC gradient.

NaOH acidification could hamper denaturation of the DNA duplex during hybridization steps in cBot and, therefore, decrease cluster density, which appeared as density fluctuations in random lanes, even in one FC. LFN purity problems affect its removal and, thus, prephasing could appear abnormally high, with consequential effects on the error rate and quality score. Debris in the reagents could potentially block tubing in the cBot and HiSeq machinery, so it is strictly prohibited and requires checking all small packing to locate debris in advance (the cBot data sheet is available from: http://www.illumina.com/documents/products/datasheets/datasheet_cbot.pdf and the HiSeq2000 specification sheet is available from http://www.illumina.com/documents/products/datasheets/datasheet_hiseq2000.pdf).

Reagent QC is described in detail later in this article.

HiSeq software

The HiSeq Control Sequencing Software (HCS) provides wizard-driven step-by-step instructions for the user. In a previous version of the HCS1.0 series, the focus was on debugging to ensure a smooth run. Meanwhile, recipes have been constantly modified to fix intensity problems, reagent usage, etc. The new dynamic focus feature introduced in the HCS1.1.37 has the benefit of fast reaction time, as it partially uses focus position of previous cycles, but requires a T density exceeding 50 K/mm². Without warning, it has caused out-of-focus problems in random and index cycles of almost all samples, so we have had to downgrade the HiSeq to the former version HCS1.0.69 to ensure consistent quality. In the HCS1.1.37.8, focus method has reverted to ATGC four bases. Dark-area quality problems with the flow-cells have worsened the focus situation. At the beginning of read2, if the former cycles are blank with indexes, then the initial quality was always bad. In HCS1.3, a template algorithm has been upgraded to use CHASTITY (highest color corrected intensity over sum of highest and second highest

intensity) instead of signal-to-noise as the merge criteria for clusters, thus improving the PF% (pass filtering rate), especially at higher raw densities, and sequencing quality, and duplication rate. It extended the template generation time to ~4 h per flow-cell. Its real time analysis (RTA) enables monitoring the running status and, in RTA1.10.28, the color matrix uses the first four cycles at the beginning of read1 and read2. Also, in-lane control alignment could be generated at cycle52. The RTA run status of all HiSeq can be checked remotely in Sequencing Analysis Viewer (SAV), and this facilitated keeping track of multiple HiSeq run status in real time, which is especially important for a large genome center such as BGI. In applying Illumina software, bugs and defects are regularly found, analyzed and reported to Illumina with suggestions for solutions for those still not fixed. Data QC in BGI is described in detail later in this article.

Optics

In contrast to GA, there are intensity boosts in HiSeq at the 76th cycle in read1 and the 150th cycle in read2, leading to a higher intensity and improved quality score. This does not necessarily lead to better sequencing quality as reflected by the overall error rate. There are two laser lamps (red and green) compared to a single lamp in GA. The excited laser beam passes four semitransparent filters for A, T and C, and one full reflection mirror for G, sequentially, and is imaged by four CCDs individually for A, T, C and G. The GA has only one CCD. In HiSeq, base G has the lowest initial intensity of all the four bases (3000–4000), while, in GA, base G is the highest (800–1000).

GA uses one camera to focus and take pictures, and identifies clusters by the signal-to-noise ratio. HiSeq initially applied a fifth, small camera to focus but, since the advent of the HCS1.1.37, this camera is no longer present in HiSeq (see the HCS1.1.37.19 release notes available from <https://icom.illumina.com/Download/Summary/cvYwj8n2QEuKFMqCUyfyEQ>).

Quality comparison

The error rate comparison between GA and HiSeq, using human samples, is listed in Table 4. The sequencing quality review is described in detail later in this article.

Library construction improvements

DNA

Throughout DNA preparation from starting genome DNA to fragmentation, size selection, amplification and library QC, BGI has managed all types of DNA samples efficiently and built-up its own library workflow coupled with ISO9001 and 5S quality management systems, with a success rate for generating quality libraries of over 95%

Table 4. Error rate between GA and HiSeq.

	Human Resequencing		PF	Error rate	Read length
	Cluster density	Cluster density			
GA	28,3817/tile	534K/mm ²	86%	1.23; 1.52	2 × 100
HiSeq	309,0000/tile	542K/mm ²	84%	1.16; 1.92	2 × 100

and a routine daily output of about 100 libraries. Exon, fosmid, BAC libraries are handled automatically to save man-power and costs, as described later in this article.

Initial genome DNA quality is important because it directly affects the quality of the library. Therefore, choosing an effective detection method for the initial DNA material is necessary. Nanodrop test results are not accurate owing to polysaccharides or other interferences. Gel densitometry and fluorescence analysis to estimate the concentration of DNA is more suitable. Thus, we used gel electrophoresis analysis software or a Qubit fluorometer HS/BR dsDNA assay kit to determine the concentration of DNA.

The less genome DNA used for library construction, the higher duplication rate. We usually use 3–5 µg DNA input for construction of short insert (insert size less than 800 bp) libraries. Initial DNA could be as little as 1 µg, but the success rate for quality library construction may drop from >95% to 85–90%.

In DNA fragmentation, adaptive focused acoustics (AFA) technology is very good, due to its efficient, fast, batch operation and low-cost properties. It also fragments DNA of less than 200 bp, and very little DNA is lost during the fragmentation process (Gilbert et al., 2010). We now fragment DNA samples using AFA technology in a 24- or 96-well format from Covaris. We also have Covaris S2 and Bioruptor for fragmenting different DNAs. Bioruptor is low cost and can fragment 12 samples at one time.

In size selection and electrophoresis, we usually excise a 2-mm wide gel slice of the desired size. Libraries that contain an acceptably narrow size distribution of paired end reads are necessary. Low voltage and a prolonged electrophoresis time to 2 h can help improve the resolution of the DNA fragments. During electrophoresis, the temperature of the buffer increases, causing gel deformation (especially for the low range agarose) or impact DNA separation. We now use a pre-cooling electrophoresis buffer and control room temperature during the electrophoresis process. Meanwhile, back-up gel slices are saved adjacent to the selected size, in an attempt to save sample and time in reconstructing the library, when the selected piece does not have good sequencing data. In gel extraction, we also routinely slice gel at room temperature instead of heating to 50°C in QG buffer, as high temperatures cause GC content bias during gel extraction.

In library amplification, it is essential to choose the appropriate set of conditions for each PCR. With too much template, non-specific PCR products will be generated, but too little template will cause higher PCR duplicates. Otherwise, a greater numbers of PCR cycles could lead to greater GC content bias in the resulting sequences. Usually, 10 cycles are used to avoid duplication and GC bias.

In PCR products purification, a clean library is the one that is free of adapter and primer dimers or single stranded DNA. We routinely purify PCR products by gel extraction.

The PCR reaction will increase duplicates and cause GC bias. This phenomenon is bad for genome assembly, especially for high or low G+C content samples. It influences the distribution of read coverage across the targeted sequencing regions. Therefore, we construct amplification-free (PCR-free) libraries for the high or low G+C content samples. The amount of input DNA required is 10 µg. In addition, we have designed more than 20 types of index PCR-free adapters so as not to waste data. For input DNA as a single size PCR product, 1 µg is needed for library construction without purification.

Regarding library insert size selection, there are no specific requirements for re-sequencing. For some specific samples (samples with short repeats), however, the 500-bp library is better than the 200-bp library because longer reads are better for mapping. For *de novo* assembly, the insert gradient is extremely important for scaffold construction. Longer fragments are conducive to scaffold assembly; shorter fragments increase the utilization of longer fragments and assembly is improved (Quail et al., 2008; Kozarewa et al., 2009).

In DNA library QC, two systems are used – Agilent 2100 for insert size detection and qPCR for concentration detection. Qualified libraries should have an insert size that varies within 10% of expectation (Panaro et al., 2000; also see the TruSeq DNA sample preparation guide (15005180A) available from <https://icom.illumina.com/download/summary/vclodijXk0uzMzq-jWcozQ>).

Large insert DNA (2–5K)

A mate pair library of 2–5K insert requires 15–20 µg of genomic DNA, which ensures enough size-selected DNA for circularization, whose ligation efficiency is between 2 and 10%, and to avoid high duplication and bias. If input

is less than 15 µg, the success rate of library construction could drop to <80%, with possibly higher duplication, and diversity of the library could be affected. The success rate of constructing a 10K library is stable at ~90% and that of 20K is 75%. For a 20K library, input DNA of 30–50 µg is required. Before PCR, the DNA is only 10–50 ng, so a PCR-free library cannot be constructed.

There are two QC steps in mate pair library preparation. First, check the size distribution and quantify the purified DNA using an Agilent Bioanalyzer 2100 (20K cannot use in this method; thus the success rate is even lower). Construction would not be continued if obvious smearing of the DNA distribution is seen. Secondly, quantify the circularized DNA using a Qubit Quantitation Platform. To assess the circularization efficiency, molecule numbers before and after circularization steps should be known. BGI takes the relatively easy step of summing the fragmented and purified molecules after circularization divided by the molecules before circularization.

To avoid overloading of samples of large libraries (5K and larger), suitable electrophoresis equipment and a gel comb should be used. Generally, we recommend aligning multiple wells and extending the electrophoresis time (14–16 h), because it appears to decrease the chance that short fragments become entangled with longer fragments in the gel, and ensures separation and size selection.

A hydroshear device is used to shear DNA at the beginning, because it provides tighter distribution of sheared DNA fragments. The tighter the initial shearing event, the greater the fraction of the sample in the desired size range. Nebulization and sonication are too strong to fragment the genome evenly. Some genomic DNA is intact, while others are sheared into pieces that are too small.

Generally, we slice two or three gel pieces and select one for circularization, according to the results of Agilent 2100 (Chen et al., 2009; see also, the Mate Pair library sequencing sample preparation guide available from <https://icom.illumina.com/download/summary/LAJkz5z08kOWWwwsoZFjmg>).

sRNA

Compared to the Illumina method and 1.5V sRNA protocols, BGI's improvements are represented as follows: about 3–5-fold fewer adapters are used, there are more index choices, the introduction of separation of 18–30 nt fragments on the first day before adapter ligation to narrow the size of choice.

PCR cycles are usually between 11 and 15, and 18–20 for serum samples, based on the gel picture and input amount. Extra care is taken on the first day in separating sRNA and an appropriate size is selected. If the size is too small, sequencing output data will not be ideal.

QC includes three quantifications. First, Agilent2100 for size and concentration, then qPCR quantification. After a comparison of the two methods where both qualify, the libraries can be pooled and the mixed library is quantified by qPCR (SmallRNA Sample Prep Guide (1004239 D), available from <https://icom.illumina.com/download/summary/nEIIhtFpu0aN9f-NHUiL2w>).

mRNA

Compared with Illumina's protocol, BGI has optimized a number of steps:

- (1) Reaction time is optimized. At the synthesis of second strand cDNA, the time is reduced from 2.5 h (Illumina) to 2 h.
- (2) Two Gel purifications, employing a 50-bp DNA ladder, are performed to obtain libraries with a better concentrated and precise insert size, compared to the one purification in Illumina's protocol.
- (3) In precipitation of RNA, the total reaction volume is increased from 86 to 392 µl, and the inoculation time is doubled compared to Illumina's protocol.

Compared with Illumina's protocol, BGI optimized the system to reduce reaction time in synthesizing SS cDNA, increased gel purification from once to twice to obtain libraries with a better concentrated and more precise insert size (mRNA-Seq Sample Prep Guide (1004898 D) <https://icom.illumina.com/download/summary/A4zkeV3eLUOO2mziSSR1g>).

Strand-specific transcriptome library preparation

To distinguish the polarity of transcript, Parkhomchuk et al. (Parkhomchuk et al. 2009) developed a simple method to construct a cDNA library to sequence the specific strand of cDNA and ascertain the origin of mRNA. Briefly, compared to Illumina's protocol for preparing samples for sequencing of mRNA, after first strand cDNA synthesis, we use dUTP to replace dTTP for second strand cDNA synthesis. After adapter ligation, the DNA strand containing the 'U' base will be digested with UNG enzyme, so only one DNA strand will be used as a template for the forward PCR step (Mwololo et al., 2010).

Non-coding RNA library preparation

Illumina's protocol is as follows: total RNA is used as starting material to construct the cDNA library, then the cDNA library is treated by DSN enzyme.

BGI's protocol is as follows: first, polyA RNA is isolated from total RNA and sequenced. Second, polyA(-)

RNA is treated by a Ribominus kit from Invitrogen to deplete rRNA, then sequenced and, by combine polyA RNA and polyA(-) plus rRNA(-) RNA sequence information, we can identify some ncRNAs.

Automation

BGI is currently developing four automation platforms: Agilent Bravo, Hamilton Bonadu, Tecan and Apricot Designs. BGI has achieved some progress in hardware maintenance and protocol design. They are being used for multiple tasks, such as reagent setup, library construction of fosmid, BAC, exon, and qPCR. The weekly output of libraries is ~200 plates. Apricot Designs' i-pipette Pro 500 μ l is mainly used in 96-well plate ethanol and TE-based liquid handling. Agilent Bravo workstations are assembled in the Shenzhen and Hong Kong laboratories, and are used in DNA purification, qPCR, RNA-Seq, Sure-select, reagent setup, etc. Four plates of DNA samples can be purified simultaneously with Ampure Beads and Jinmaige Beads (from GenMag Biotechnology) and completed in 38 min. Hamilton STARlet is assembled in Hong Kong and mainly used in DNA Purification, RNA-Seq, Sure-select and index-pooling. Tecan is assembled in the Health Department of Shenzhen, and mainly uses an 8-channel system to transfer samples to 96-well plates or 1.5-ml micro-centrifuge tubes. The 96-channel is also sometimes used.

Library QC system

BGI has applied a fully self-developed qPCR methodology to quantify libraries before loading the sequencers. Only libraries equipped with both adapter and primer could be amplified in real time and amplified later on the flow-cell. Therefore, quantification via qPCR is more accurate than the Agilent 2100 bioanalyzer, which nonspecifically detect all DNA fragments. In addition, qPCR is very sensitive and low concentrated libraries can also be quantified.

There are about 20 standard libraries for use in qPCR quantification. Each has its own quantification curve for flow-cell clusters and each standard is chosen based on type (animal/plant large genome, or bacterial, PCR-free or Chip-Seq) of library, size of insert, GC content, etc. For example, PCR-free libraries require about half the loading amount for the desired cluster density on the flow-cell, compared to normal PCR-amplified libraries. Standard libraries are chosen from those with good sequencing data, where the qPCR amplification curve are very close to the old standard sample and, especially, where the 1-nM point overlapped, to guarantee overall stability of quantification QC (Figure 3). Samples stored more than 3 months should be requantified. Standard samples were aliquoted as 1 nM and stored. To use, dilute four times to form five titration concentrations: 1000, 100, 10, 1 and 0.1 pM. It is a

requirement that at least four points are in the same line (Figure 4). At quantification qPCR, all libraries for examination were generally diluted 100 times, with a reaction volume of 10 μ l.

Normally, qPCR and Agilent 2100 quantifications have a relatively fixed ratio relationship, which could be used for QC operation error and degradation situations, etc. If the ratio is out of range, a requantification is normally carried out. Large insert libraries and methylation libraries are exceptions of this rule, and are still under study.

Quantification and cluster density are summarized and analyzed weekly and minor adjustments made for more accurate cluster predictions. Standard sample r^2 values should be above 0.995 and amplification efficiency should be above 80%.

HiSeq operation troubleshooting

There are several critical steps where quality of a run needs to be checked to ensure final output. At the 1st cycle, flow-cell defects, such as dark-area, is checked. After the 4th cycle, cluster density, signal intensity and FWHM are checked. After the 13th cycle, phasing and prephasing values are checked. After the 25th cycle, %PF, % base and Q30 are checked. The main parameters are cluster density and distribution, cluster phenotype, initial intensity, signal decay, GC content and distribution, %PF, phasing and prephasing, quality value, error rate, % align, etc. From experience, we have assigned a quality pass value to each parameter and, if unqualified, several related factors may be abnormal; then, troubleshooting is carried out for quality assurance improvement actions. Here, we list some general problems (Table 5), elaborating on topics that need further attention to differentiate reagent, software or machinery defects from sample or library features, and urging Illumina R&D to resolve existing problems or to improve the BGI production or QC system.

HiSeq & cBot tips

- (1) Rehybridization is possible if intensity drops too low in HiSeq, which is rarely a problem in GA.
- (2) When not in use, the HiSeq sipper should be soaked in water and postwashed once a day to prevent clogging.
- (3) If the gasket is changed after priming in every new run, as Illumina suggests, there will be a high rate of clogging, so gasket change does not need to be frequent.
- (4) Illumina suggests a NaOH wash after every run in HiSeq. This causes rusting of the pumps; about 40% of our HiSeq in Hong Kong has shown rusting.
- (5) HiSeq will automatically adjust the sides of the lanes, based on clusters. If the clusters are too

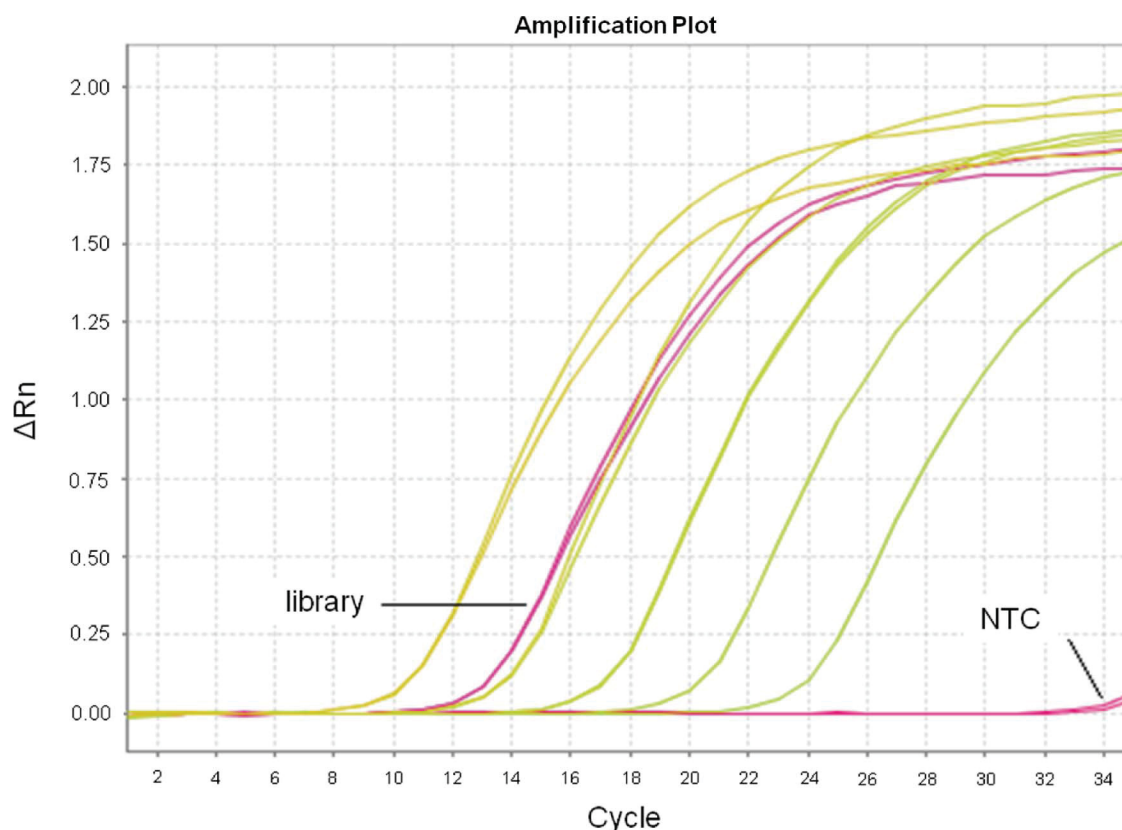


Figure 3. Amplification Plot in qPCR. Twenty different kinds of libraries with good sequencing data were chosen as standards, and the qPCR amplification curves of similar-type test libraries were compared to the standards for quantification. New standards were also compared to former ones, especially among dilutions when the 1-nM point was overlapped to guarantee overall stability of quantification QC.

little, there will be a problem and run can not be initiated. This can be overcome by changing the local disk file `<EnableAutoCenter>true` in "runParameters.xml", from "true" to "false". The software will skip the automatic side adjustment and go on.

Data QC system

With HiSeq, imaging information, including intensity information (*.cif), is converted to a base-calling file (*.bcl) by RTA, when Q30 is generated. The .bcl file is converted to a .qseq file via a Bcl converter. It is different from GA, whose RTA only produces .cif files. GA uses OLB to convert the .cif file to a .qseq file. The process of generating a .qseq file can also calculate regular output parameters, such as raw cluster, PF cluster (%PF), FCI, phasing and prephasing. Pipeline then analyzes the .qseq file to generate a .fastq file for mismatch information by ELAND, if a reference sequence is available (Figure 6). Then, BGI uses self-developed Sequencing Quality Control Software (SQCS) (Figure 5) to further the analysis for QC purposes, including Q20 (percentage of bases with quality value ≥ 20), PF, GC content, base distribution, quality

curve (with checkfastq .cpp and fqcheck_distribute.pl), adapter composition (with filter_adapter.cpp), error rate distribution (with error_rate.pl), insert size and distribution (with insert_size.pl), intensity curve (with signal.pl), index composition analysis (with index_stat.pl), align rate and average error rate, average insert size, etc. (with report.pl). There is a difference between RTA PF and pipeline PF. RTA gives the average PF of all 32 tiles in HiSeq; fast QC check (4 tiles) and full analysis in the pipeline PF is calculated as medium number. This difference is vast with flow-cell quality problems, such as with the dark-area or gradient. We changed the setting of pipeline analysis to use RTA PF value for the fast QC check for accurate early detection of density problems (Table 6), while still keeping full analysis pipeline PF for accurate yield report.

Reagent problems and solutions

Dark area case of FC

Dark-area problem is referred as an extreme, non-even (low-density) seeding of adapters on the surface of flow-cells. When scanned with fluorescent reagent during

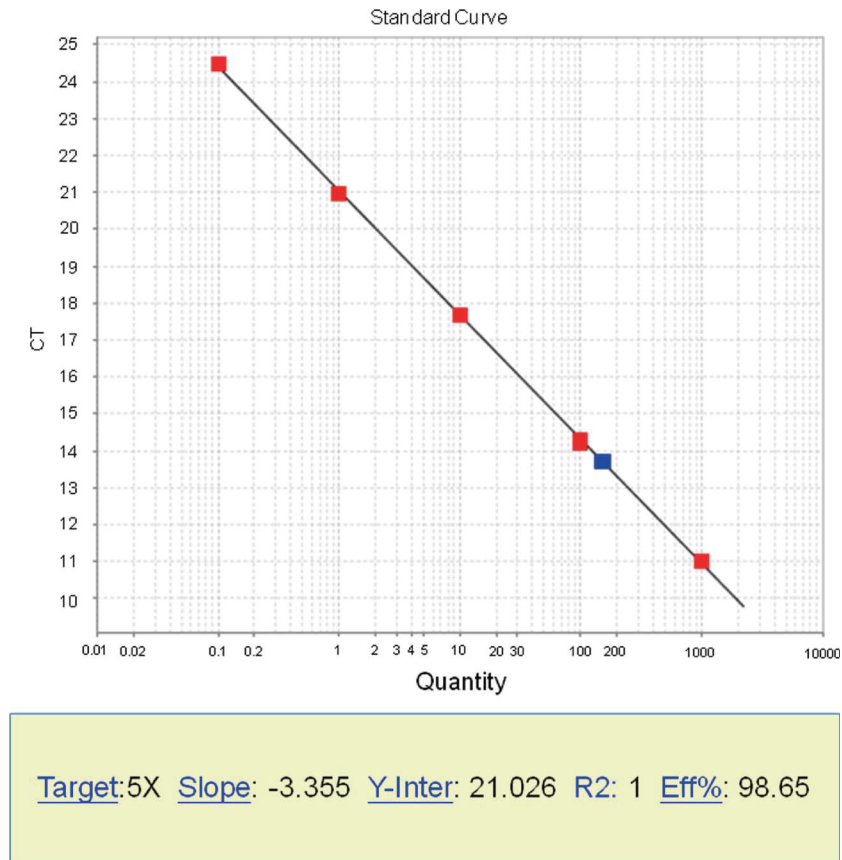


Figure 4. Standard curve in qPCR. Standard libraries were diluted to 1000, 100, 10, 1 and 0.1pM. It is required that at least four points are in the same line.

Table 5. Some problems detected in BGI and troubleshooting guide.

Problem	Cause
No cluster or fewer cluster; fewer cluster and lower GC	FC defect; NaOH reagent with incorrect pH and concentration; fluidics are blocked during cluster generation
Read N in certain cycle or in index cycle; lower PF, FWHM > 4 for special sample such as small RNA, DGE, etc	The dynamic focus method of HCS1.1.37 depends on the T cluster density, which is the minimum number (50K/mm ²) of “T” clusters per mm ² required
Intensity graph is abnormal (bridge-like); G intensity is very low	SMR not mixed very well
High phasing	ICR not mixed well
High prephasing	LFN with bad quality; valve problem; thermal station problem
Water pours out from reagent compartment	The condensation pump may not keep up depending on environmental conditions
GT/AC intensity are zero	Green/Red laser problem
Out of focus in the first cycle of R2 with non index library and index library in the same FC	HCS 1.1.37 and HCS 1.3.1 can not support this kind of running
Index recipe lost when FCA is PE run and FCB is PEI run	HCS 1.3.1 can not support this kind of running

sequencing, such areas usually appear as black (dark) areas without clusters. The phenotype of dark-area problems typically appear as a straight-line on the edge, straight-line in the middle, wavy-edge of a lane, or total black-out of a lane. It not only caused decreased yield, but also the PF rate and sequencing quality is affected and dropped quickly. The reason is associated with focusing quality, especially

the dark area around the middle of a lane. Based on experience of failure rate and cost, BGI decided to stop runs when dark-area exceeds one third of the whole surface of the flow-cell, if not all are located in the middle of lanes. A typical dark-area FC is shown in Figure 7. With dynamic focus for clusters, the effect on quality is worsened compared to GA.

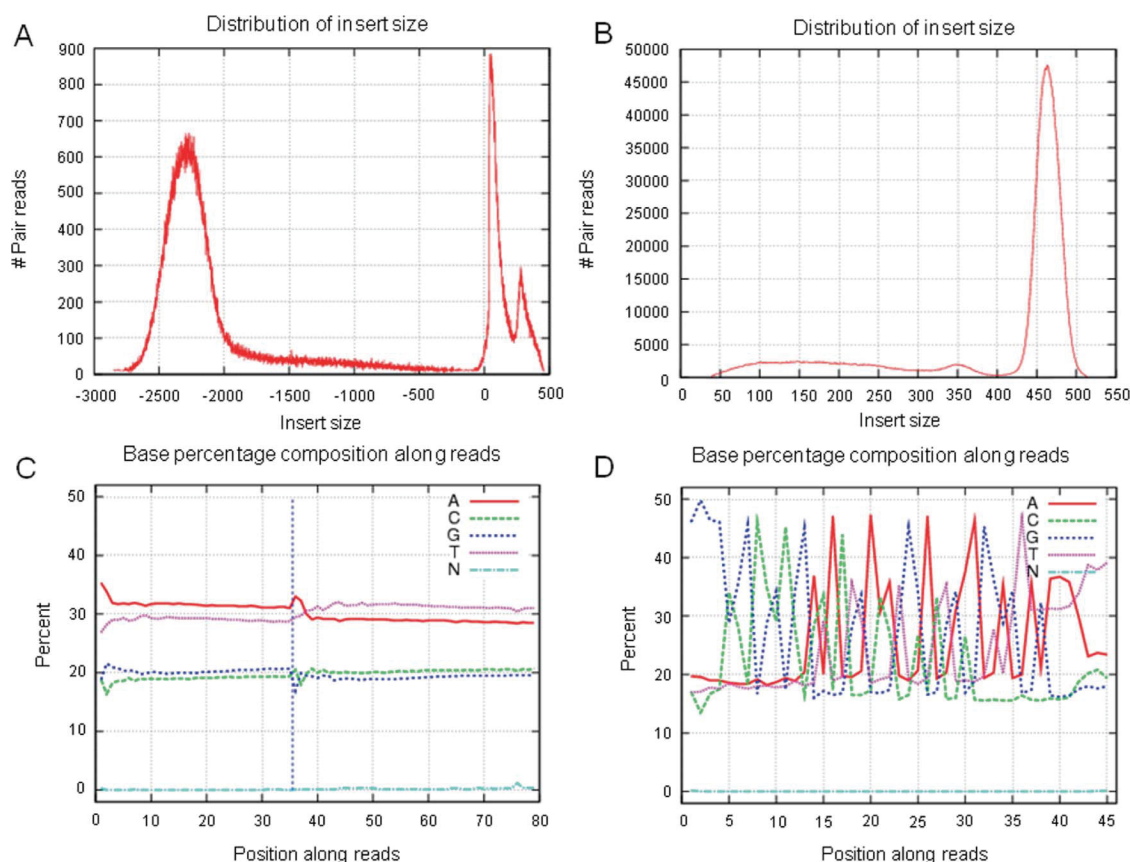


Figure 5. SQCS output. (A) Mate-pair library with abnormal small insert contamination may be due to incomplete digestion of linear DNA, insufficient biotinylated tag purification or nicked DNA genome. This could lead to assembly problems. (B) A 500-bp library with smear smaller contaminants. This may be caused by bad resolution of gel or gel contamination. (C) Abnormal base composition. Reasons may be complex genome or expression, such as Chinese shrimp, library preparation (inrandom fragmentation or ligation, adapter contamination, etc), or sequencing. (D) Abnormal base fluctuation due to adapter contamination.

Gradient across the FC

There is a cluster density gradient across the FC in HiSeq due to a manufacturing defect. A typical FC gradient is shown in Table 7. The gradient not only affects the PF and, thus, yield of sequencing, but also overall quality (which is basically related to the focusing algorithm). Usually, smaller libraries have a bigger gradient. A 200-bp DGE library cluster variation could be up to 150%. To guarantee sequencing quality and yield, cluster expectation decreases by up to 30% and cost is increased.

Reagent QC

BGI always checks the reagent condition and performs reagent quality tests and analyses when reagent shipments arrive. Checking includes packing impairment, amount and consumption of dry ice, and regular logistics. Reagents with different lot numbers are stored separately and tested accordingly. In our experience, different lots of reagents usually differ in quality and phenotype. Standard human

libraries with a verified quality are used as in-lane controls; quality-stable project samples are used on good-condition HiSeq for reagent QC. In this step, the analysis parameters mainly include cluster density, PF, phasing/pre-phasing, intensity conditions, quality value, error rate, etc. Preliminary test results are verified and released to NGS platform and related departments in BGI within 2 days after fast QC analysis of test runs. Thereafter, a total of 10% of each lot of reagents will be selected randomly for further reagent QC tests. Analysis parameters include cluster density, PF, phasing/pre-phasing, intensity conditions, error rate, quality value, GC content, GC distribution, etc. We have also made changes in RTA reports of cluster density in the fast QC check for early detection of FC problems to ensure progress within projects (as detailed in the “Data QC system” section of this paper).

Reagent problem are sometimes shown as GC bias. For example, in 2009, the GA v2 cluster generation reagent with Taq polymerase for first strand synthesis resulted in a lower GC content. HiSeq truseq-v3, which is available this

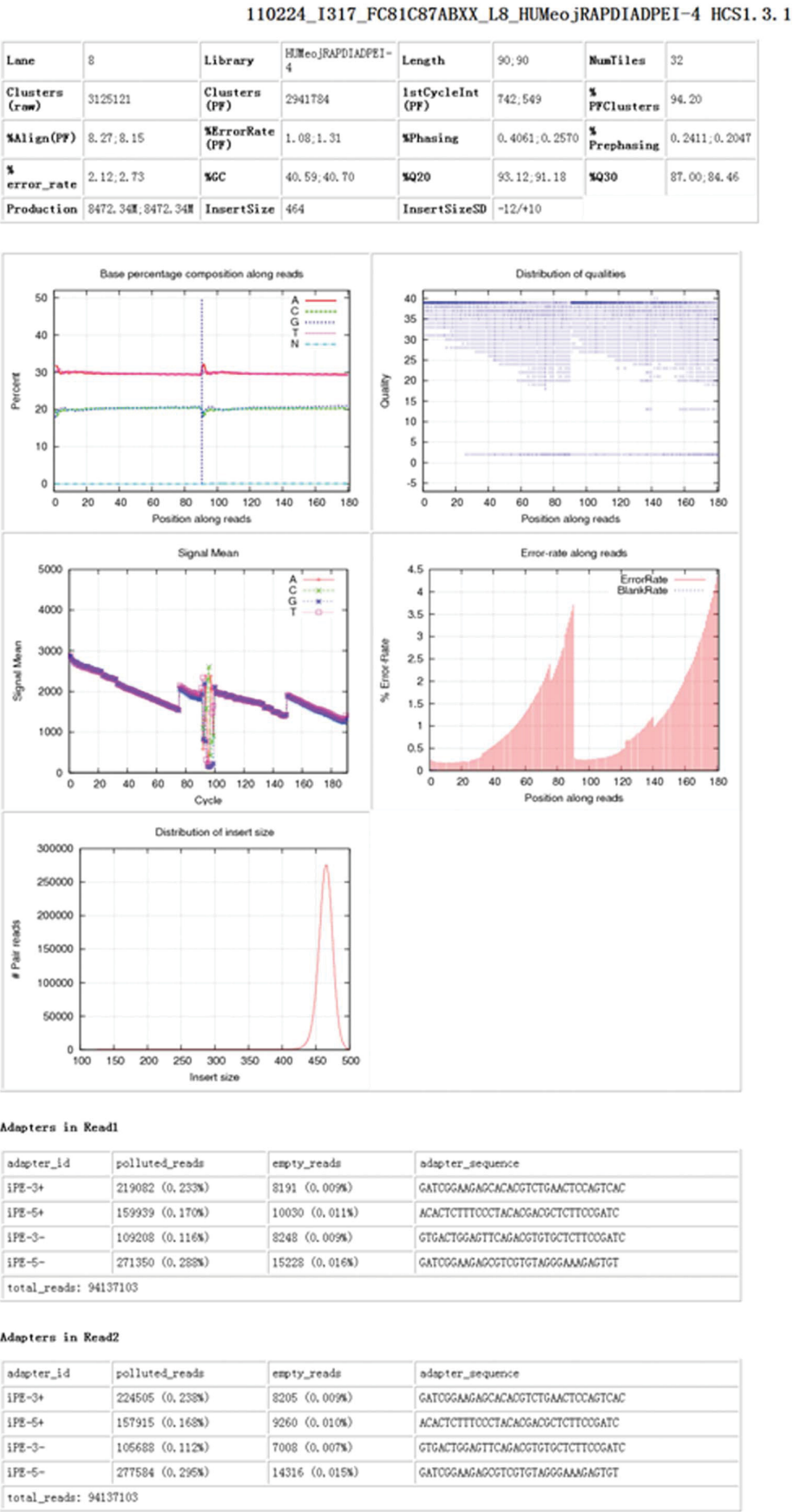


Figure 6. BGI's BMS system QC webpage.

Table 6. Fast QC check cluster and PF comparison between RTA and unmodified pipeline.

With FC problem				Without FC problem			
RTA		Pipeline		RTA		Pipeline	
Raw cluster*	PF%	Raw cluster*	PF%	Raw cluster*	PF%	Raw cluster*	PF%
2.30%	0.01%	45.24%	−8.88%	2.29%	0.02%	0.33%	−0.31%
2.32%	0.02%	−18.56%	−3.66%	2.29%	0.01%	−4.55%	−0.40%
2.50%	−0.07%	36.22%	−22.87%	2.34%	0.00%	−0.02%	−0.09%
2.30%	−0.05%	11.63%	−14.48%	2.30%	0.00%	−4.03%	−0.17%
2.29%	−0.02%	1.08%	−0.85%	2.30%	0.05%	−0.35%	−0.12%
2.48%	0.01%	−2.81%	−0.43%	2.36%	0.03%	−1.01%	−0.92%
2.41%	−0.03%	7.52%	3.01%	2.35%	−0.05%	−2.81%	0.25%
2.39%	−0.05%	−5.68%	−0.12%	2.31%	−0.03%	−5.18%	−0.51%

Note: 1, FC problem refers to dark-area problem. 2, * In fast QC check, this value is calculated as RTA/pipeline, reporting raw clusters divided by the raw cluster numbers from the final complete pipeline analysis, which is considered the most accurate reflection of overall clusters for a run.

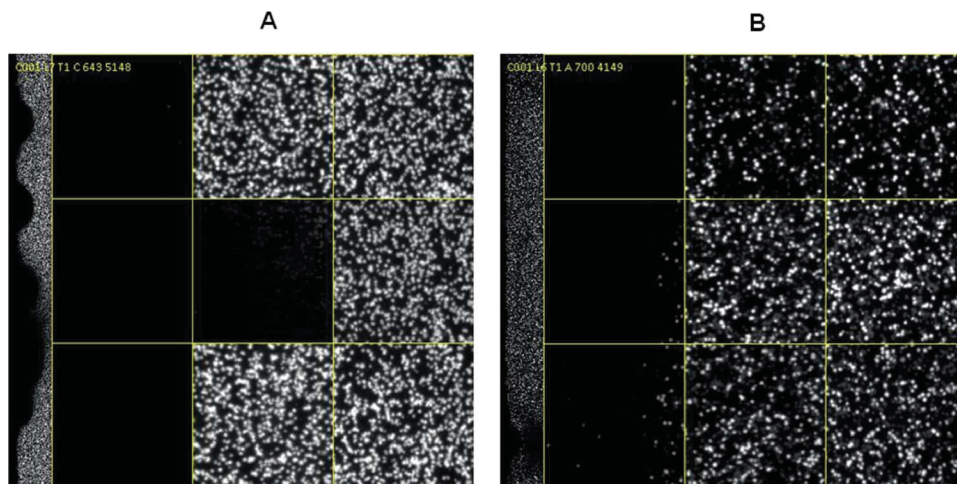


Figure 7. Dark-area FC. (A) Phenotype as wavy-edge dark area up to 50% of the lane. (B) Phenotype as straight-line and irregular dark area on one end of the lane. This could also lead to sequencing quality problems as low PF, bad focus, low Q20 and high error rate, etc.

Table 7. A typical FC with gradient cluster densities.

Tile	Raw Clusters /mm ²	Tile	Raw clusters /mm ²	Tile	Raw clusters /mm ²	Tile	Raw clusters /mm ²
1	376,902	21	374,945	41	387,298	61	379,394
2	412,696	22	402,663	42	419,527	62	408,727
3	453,912	23	441,351	43	454,679	63	442,497
4	500,519	24	482,639	44	499,203	64	480,924
5	549,637	25	530,444	45	546,842	65	528,081
6	601,133	26	582,299	46	594,390	66	577,710
7	651,514	27	631,731	47	642,518	67	625,534
8	693,286	28	669,136	48	698,479	68	680,951

year, could provide a solution for generating less biased sequences and better representation of GC-rich regions, due to new cluster amplification polymerase and SBS polymerase. BGI considers a GC content above 65% as a high GC genome. By analyzing GC distribution of human re-sequencing, coverage at the high GC region is obviously improved.

Duplicates

To improve the scaffold length of assembling and at the same time avoid difficulties in circulizing the long DNA fragment, we have tried a 800-bp insert library without circulization. We got higher rates of duplication reads for the 800-bp insert libraries during conversion from GA to HiSeq.

Duplication is related to genome size, repeat region of the genome, sample nature (expression level), insert size of libraries, sequencing depth (yield), amplification method, and the algorithm for image analysis and base-calling. It not only affects useful data output, but also the quality index for sample preparation and the sequencing system. However, for the 800-bp insert library, we have found that one of the reasons for high duplicates is related to the insert size. In general, the cluster diameter will be larger if we sequence longer libraries. However, in the process of image analysis, a large facula from a large cluster may be split into several smaller areas and base-called to several reads with the same base sequence and with very close position coordinates, which is known as “base-calling duplication”. This kind of duplication can be detected by comparing both the base sequence and the position coordinates. We calculate the average rate of base-calling duplication for each insert-size sequenced on GAIIX and HiSeq (Table 8). Base-calling duplication increases with the insert-size of library, and the data of HiSeq has more base-calling duplication than that of GAIIX. Additionally, the quality and accurate rate of 800bp-insert libraries are usually worse than short libraries because the quality of the clusters.

However, duplication of 800-bp insert library decreased significantly with the use of the HCS 1.3 in HiSeq, due to changes in the template merge algorithm with the chastity value of the base calls (highest color-corrected intensity over the sum of the highest and second highest) instead of signal-to-background ratio. Here, we select a human re-sequencing 800-bp insert library to compare duplication in different HiSeq software versions (Table 9).

Table 8. Rate of base-calling duplication for different insert-sizes.

Insert-size of library (bp)	% Basecalling duplication	
	GAIIX	HiSeq 2000
200	0.1	0.3
500	0.3	3.9
800	4.5	13.3

Table 9. The duplicates in different HiSeq software versions.

HCS version	1.3	1.1.37
Cluster(raw)/mm ²	304,916	473,335
Cluster(PF)/mm ²	290,947	410,264
%PF	95.42	86.68
Total reads	1,669,681	2,307,737
Duplicate reads	54,242 (3.25%)	727,328 (31.52%)
Base-calling duplicate reads	10,846 (0.65%)	684,476 (29.66%)

HiSeq yield/PF/quality study

To explore the maximum yield of HiSeq, we sequenced nine lanes with different cluster densities using a standard human library in HCS 1.3. The library's length is 450 bp and total sequencing length is 91PE, 90 bp available in read1 and 90 bp in read2. We aligned the reads to the reference (NCBI HG18) by SOAP2.20 (<http://soap.genomics.org.cn/>) and calculated the mapped rate of the reads and the mismatch rate of the bases. The data is shown in Table 10 and Figure 8. If we ignore the data quality, the maximum yield is 22.7 G/lane at 988K clusters/mm². Three phenomena indicate a decrease in quality when the cluster density increases, especially when it is higher than 786K/mm². The %PF (pass filtering rate of Illumina base-calling) decreases, the %Q20 (percentage of bases with quality value ≥ 20) drops and the unmapped rate and mismatch rate increase as a result of alignment. These are not obvious when the density raises from 319 to 596K/mm², but become serious when the density reaches 786K/mm².

Additionally, the GC content of reads decreases when the density increases, which is also observed in GA, due to complex factors in the bridge amplification process on the flow-cell and the image analysis algorithm. The GC bias at high density has been modified by reagent and software updates but is still difficult to resolve completely. The problem of GC bias becomes more significant in the depth distribution versus GC content. We split the genome to 1-Kbp windows and calculate the sequence depth and GC content in each window (Figure 9). The shape of the depth distribution changes slightly and the GC bias is not a problem at low and medium densities (Lane A–C). It is acceptable at high densities (Lane D–E) but, at extreme densities (Lane F–I), the GC bias is a major problem.

For the human DNA library, the best raw cluster density is between 500 to 600K/mm², considering the yield, the quality and the GC bias. At this density, the yield from one 101PE run with two flow-cells on the HiSeq can reach 250–300G bases. Correspondingly, the yield from one 101PE run of GA is about 40–50G bases at the optimum density. In normal data from 101PE runs on HiSeq, the error rate from human libraries is 1–2% and the Q20 is 80–95%, which is no different from GA.

GC bias has recently been improved through reagent upgrades from truseq v2 to truseq v3 by adopting new polymerase and buffers, along with FC enlargement (scanning area changed from 2949 to 4424 mm²), and software algorithm upgrades (cluster density changed from 488 to 825K/mm² for a standard human library) to achieve output data of 200–600G. The *Mycobacterium tuberculosis* genome was sequenced by both Truseq v2 and Truseq v3 reagents. It was found that, with Truseq v3 reagent, GC bias in the high GC region was reduced compared to Truseq v2 reagent (Figure 10). The reported GC content of

Table 10. HiSeq data with different cluster densities.

Lane	Raw cluster density (K/mm ²)	% PF clusters	Lane yield (G bases)	% Q20 bases		% Mapped reads		Pair-mapped	% Mismatch bases		% GC	
				Read 1	Read 2	Read 1	Read 2		Read 1	Read 2	Read 1	Read 2
A	319	90.6	9.3	95.2	94.0	95.1	94.3	88.8	0.61	0.69	40.83	40.83
B	389	90.2	11.4	94.9	93.3	95.1	94.1	87.8	0.59	0.60	40.82	40.83
C	440	89.3	12.7	94.6	92.8	95.0	93.9	87.6	0.61	0.63	40.77	40.79
D	521	88.0	14.8	93.9	92.1	94.9	93.7	87.2	0.64	0.68	40.69	40.72
E	596	86.0	16.6	93.6	86.0	95.5	93.2	85.3	0.58	0.94	40.33	40.49
F	786	82.3	21.0	90.1	85.1	94.6	93.4	86.1	0.80	1.39	40.04	40.08
G	938	73.3	22.2	86.4	81.4	93.7	88.0	80.4	0.92	1.48	39.48	39.43
H	988	71.0	22.7	86.7	79.6	93.3	83.7	76.3	0.92	1.62	39.42	39.25
I	1020	65.6	21.7	85.0	76.5	92.5	72.7	66.2	1.09	1.70	39.13	38.68

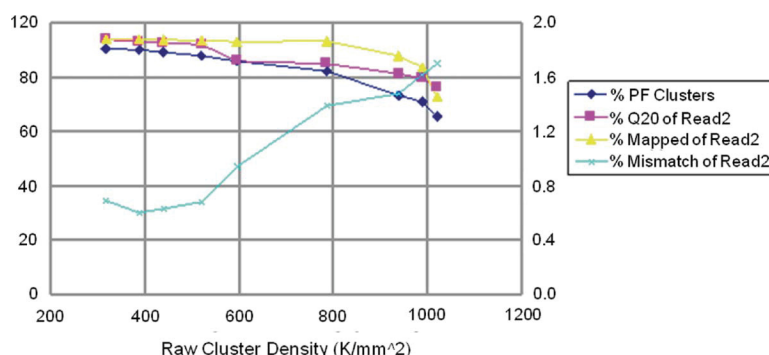


Figure 8. HiSeq quality data at different cluster densities. With increase in cluster density, the PF, Q20 and map rate decreased and mismatch rate increased. Quality has to be considered at the PF turnover point and before reaching the maximum yield and cluster density. In this figure, a density of around 700K/mm² is critical for maximum yield.

Mycobacterium tuberculosis is ~60%. A sliding windows of 500 bp was used in analysis of the GC distribution. GC distributes quite evenly using Truseq v3 reagent, while GC depth varies significantly in the range 55–70% with Truseq v2. Another significant improvement associated with the new scanning chemicals and SBS enzyme in truseq v3 versus truseq v2 is that signal decay is reduced and sequencing accuracy improved (Table 1)

In-lane control

We have applied an in-lane control library to evaluate the sequence quality and quantification of the sample library. Before the advent of indexed libraries and multiplexing sequencing, the control library (Phix or Human) occupied one lane for each flow-cell in calculating the error rate and evaluating the quality of the run. In fact, only 1% of the data from one lane is sufficient to achieve the purpose of control. Presently, we use in-lane control with an indexed library. We mix the indexed control library and the indexed library from another sample using a fixed content of control library versus 1% of the mixed library. Normally, the cluster number of the control will be stable; that of the sample will fluctuate due to the variation in quantification. If there

is a problem with the flow-cell, the cluster numbers of the control and sample will be inconsistent, but the effect of the flow-cell on the control and sample are not equal. Using in-lane control, we can distinguish the problem of flow-cell or the quantification of the sample (Table 12).

Relationship of input sample amount and cluster density in HiSeq with different software versions

The 800-bp insert libraries have relatively lower cluster densities than regular 500 bp libraries as regards difficulties in quantification and cluster amplification. Duplication is another potential factor affecting yield (detailed in the “Duplication” section of this paper). When HiSeq control software was upgraded from HCS 1.1.37.8 to HCS1.3, cluster density versus sample input amount was studied. Four 800-bp insert libraries were tested in both software versions at the same and different concentrations. The results indicated that, with the same input amount of sample, HCS1.3 generated a relatively lower cluster density and, to achieve a specific cluster density, HCS1.3 required more input amount of sample (Table 13). This is also true of the 200-bp insert and 500-bp insert libraries, but the concentration difference is lower compared to a 800-bp insert

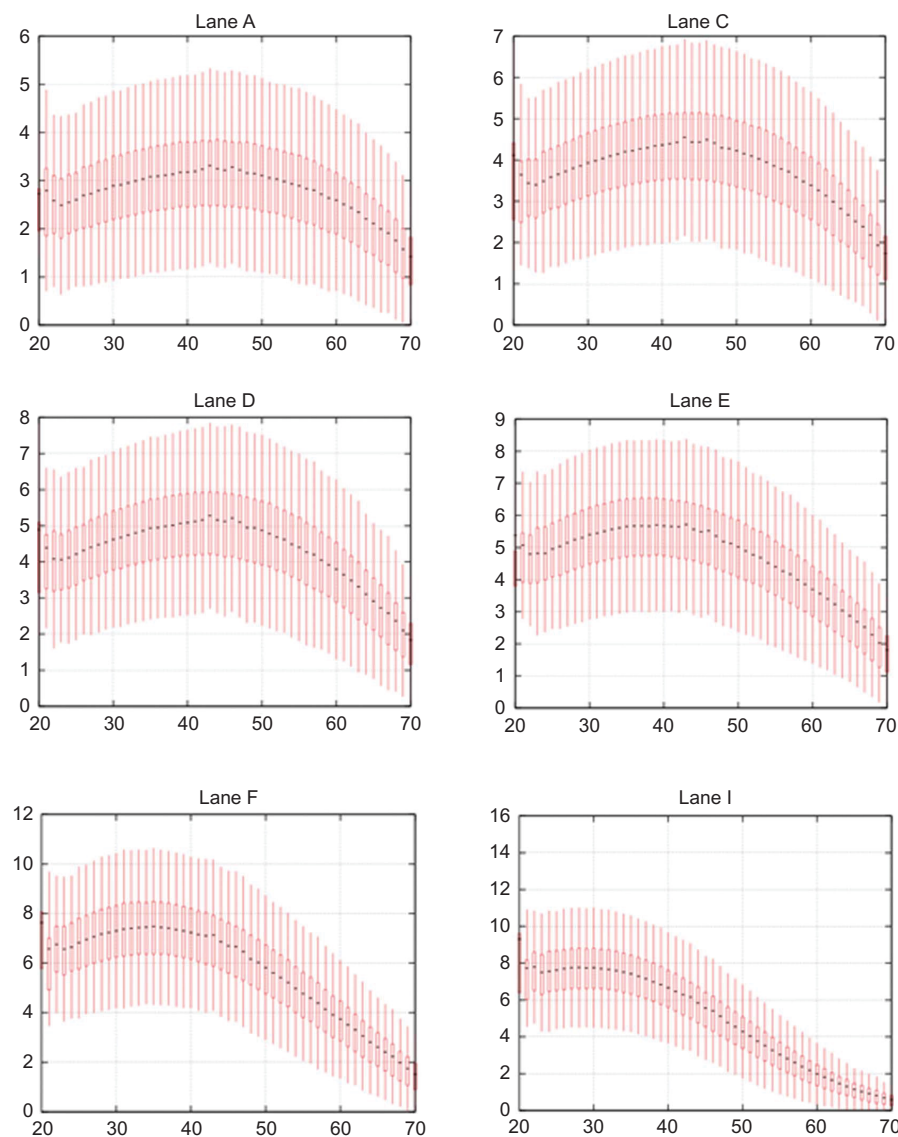


Figure 9. Depth distribution versus GC content corresponding to Table 10. The genome was split into 1-kbp windows. Depth of sequences and GC was shown in each window. With an increase in cluster density, the GC content became lower and the GC bias becomes a problem. At high density in Lane F and I, GC distribution was abnormal.

library. This is in accordance with the new template merge algorithm feature of HCS1.3. Nevertheless, the question remains as to whether it will cause data bias. By analyzing GC distribution with the same amount of data, the difference is negligible, but further studies on data correlation are needed.

Index adapter and sample mix check-point

With the increase in HiSeq yield, samples are pooled in a more complex manner in one lane for sequencing, with single or combined indices. Currently, BGI uses over 200 index sequences.

However, double-indexed libraries and some others, such as PCR products, have unbalanced bases in the initial

cycles during sequencing. This leads to lower PF values (calculated over the first 25 cycles) and affects both yield and quality. We usually decrease the cluster density, use more matched indices or pool more libraries to ensure quality (Tables 14 and 15 and Figure 11).

To prevent sample mix-up, we combine PCR and mass spectrometry to check SNP as the ID of human samples (see Appendix 1: BGI sample ID PCR primers).

Completed and ongoing projects at BGI

BGI has completed a number of plant genome projects. In 2010, using the Illumina GA II platform, BGI re-sequenced a total of 17 wild and 14 cultivated soybean

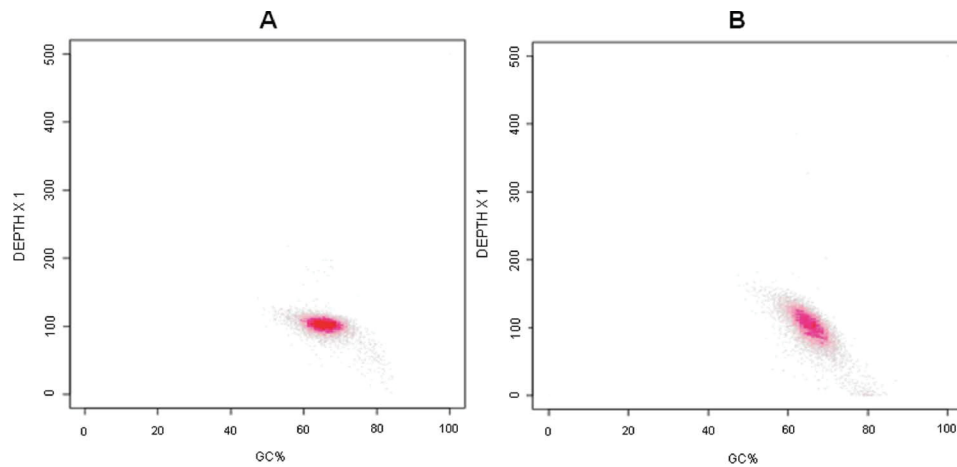


Figure 10. Correlation of GC content and genome base depth in *Mycobacterium tuberculosis*. The horizontal axis is GC content and the vertical axis is depth of gene base. (A) Truseq v3 reagent. (B) Truseq v2 reagent. Each data point corresponds to the number of reads recorded for a 500-bp window.

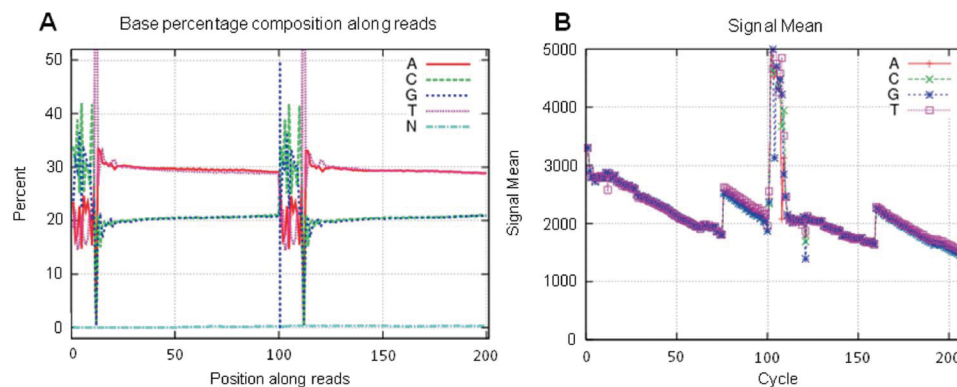


Figure 11. Improvement in sequencing quality by pooling more index libraries. Index library A shown above corresponds to Table 15

Table 11. Error rate between truseq v3 and truseq v2 (Error rate in A and B was calculated by ELAND).

Reagent	Read Length (bp)	Raw Clusters / tile	%PF	% Error rate	%GC	%Q20
(A) The same library (human sample) was sequenced by Truseq v2 kit and Truseq v3 kit						
v2 kit	100;100	4029347	90.5	1.18;1.51	40.92;41.09	92.36;88.63
v3 kit	100;100	4341583	89.7	0.69;1.20	41.71;41.74	95.95;89.38
(B) A human sample was re-sequenced by Truseq v3.						
	150; 150			1.64;1.77	43.28;43.32	88.75;89.00

Table 12. Distinguishing the problem with in-lane control.

Total clusters/mm ²	Control clusters/mm ²	% content of control	Comment
449,403	4045	0.9	Normal
482,630	4054	0.8	Normal
427,840	3936	0.9	Normal
312,557	4001	1.3	The sample's content is less than expected due to the variation in quantification
198,840	2983	1.5	The flowcell's problem affect the cluster number of both control and sample, but not equally

Table 13. Relationship between concentration and HCS version.

Library ID	HCS version	Raw clusters/mm ²	%PF clusters	Concentration (pM)
L1	HCS 1.3	305,627	94.97	10.2
	HCS 1.1.37.8	494,285	85.56	10.2
	HCS 1.3	420,748	94.23	14.6
	HCS 1.1.37.8	462,178	86.93	10.5
L2	HCS 1.3	283,338	95.74	10.5
	HCS 1.1.37.8	454,908	86.68	10.5
	HCS 1.3	418,738	94.07	16.5
	HCS 1.1.37.8	423,747	86.66	10.8
L3	HCS 1.3	304,918	95.42	10.5
	HCS 1.1.37.8	452,968	86.44	10.5
	HCS 1.3	376,128	94.89	15.2
	HCS 1.1.37.8	398,178	88.98	10.8
L4	HCS 1.3	305,088	95.38	10.5
	HCS 1.1.37.8	430,658	87.05	10.5
	HCS 1.3	377,702	95.09	15
	HCS 1.1.37.8	417,313	87.7	10.8

Table 14. Improvement in multiplex sequencing quality by lowering cluster density.

Library	Raw clusters /mm ²	PF clusters /mm ²	%PF clusters	%GC	%Q20
LA	820,385	504,048	61.78	53.01	70.54
	288,623	271,915	94.6	53.39	92.05
LB	681,785	495,652	72.8	53.04	84
	373,292	328,353	88.8	53.42	90.57

Table 15. Improvement in sequencing quality by pooling more index libraries.

	Raw clusters /mm ²	PF cluster /mm ²	PF%	%Q20	Read Length
Index library A (>100 libraries pooled)	503,617	437,221	86.87	91.43	100×2

genomes to an average of approximately $\times 5$ depth and $>90\%$ coverage (Lam et al. 2010). A total of 901.75 million (M) paired-end reads of 45- or 76-bp read length (180 Gb of sequence) was generated. All sequence reads were aligned against the reference genome using SOAP2 with parameters that included sequence similarity, pair-end relationships and sequence quality. We compared the patterns of genetic variation between wild and cultivated soybeans and identified higher allelic diversity in wild soybeans. The data provided a valuable resource for the analysis of wild soybeans and to facilitate future breeding and quantitative trait analysis.

In a similar case, in 2009, BGI reported the draft genome sequence of *Cucumis sativus* var. *sativus* L. (Huang et al. 2009), using a novel combination of traditional Sanger and next-generation Illumina GA sequencing technologies. We generated a total of 26.5 billion high-quality base pairs, or 72.2-fold genome coverage, of which the Sanger reads provided 3.9-fold coverage and the Illumina GA reads provided 68.3-fold coverage. The GA reads ranged in length from 42 to 53 bp at that time.

BGI is now running a huge genome project called “1000 Animal and Plant Genome Project”, most of which are being sequenced using the new Illumina platform, HiSeq.

Like other genome centers, most of the sequencing power of BGI is focused on human genomes. Together with the Sanger Institute and NIH, BGI initiated the 1000 Genome Project (The 1000 Genomes Project Consortium 2010), which aims to provide detailed characterization of human genome sequence variations as a foundation for investigating the relationship between genotype and phenotype. In the pilot phase of this project, the goal is to develop and compare different strategies for genome-wide sequencing with high-throughput platforms. There are three projects: low-coverage whole-genome sequencing of 179 individuals from four populations, high-coverage sequencing of two mother–father–child trios, and exon-targeted sequencing of 697 individuals from seven populations. A total of 4.9 terabases of DNA sequence was generated in nine sequencing centres using three sequencing technologies: the illumina GA (I and II), AB SOLiD™

System (1.0 and 2.0) and the 454 GS FLX. At BGI, using Illumina GA, 231 Gb of sequence was generated in the low-coverage project and 76 Gb of sequence in the Trio project. These methods and public data have supported the next phase of or human genetic research. The Illumina HiSeq platform is running for the next phase of the 1000 Genome Project.

BGI not only sequence large genomes but is also involved in studies on small genomes, such as microbes. BGI is part of the Metagenomics of the Human Intestinal Tract Program (metaHIT). To obtain a broad overview of human gut microbial genes and understand the impact of gut microbes on human health and well-being, it is crucial to assess their genetic potential. BGI used Illumina Genome Analyser (GA) technology to carry out deep sequencing of total DNA from faecal samples of 124 European adults (Qin et al. 2010). A total of 576.7 Gb of sequence was generated, almost 200 times more than in all previous studies, assembled into contigs with a predicted 3.3 million unique open reading frames (ORFs). This gene catalogue contains virtually all prevalent gut microbial genes in our cohort, provides a broad view of the functions important for bacterial life in the gut and indicates that many bacterial species are shared by different individuals. The results have also shown that short-read metagenomic sequencing can be used for global characterization of the genetic potential of ecologically complex environments. The metaHIT program has moved to the second phase with the number of people sampled increased to 300 and all samples being sequenced using the HiSeq.

Acknowledgements

We thank Dr Li-jun Ma for critical reading the manuscript, the tech group in the sequencing platform of BGI for valuable comments and Jingxiang Li for general support of this work.

References

- Chen Y, Gelfond JAL, McManus LM, Shireman PK. 2009. Reproducibility of quantitative RT-PCR array in miRNA expression profiling and comparison with microarray analysis. *BMC Genom.* 10:407.
- Feldman B. 2001. The Nobel Prize: A History of Genius, Controversy, and Prestige, p252, 379. Arcade Publishing
- Flusberg BA, Webster DR, Lee JH, Travers KJ, Olivares EC, Clark TA, Korlach J, Turner SW. 2010. Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat Methods.* 7:461–465.
- Gilbert JA, Meyer F, Jansson J, Gordon J, Pace N, Tiedje J, Ley R, Fierer N, Field D, Kyrpides N, et al. 2010. The earth microbiome project, report of the 1st EMP meeting on sample selection and acquisition, Argonne National Laboratory, October 6th 2010. *Stand Genomic Sci.* 3(3):249–253.
- Gupta PK. 2008. Single-molecule DNA sequencing technologies for future genomics research. *Trends Biotechnol.* 26:602–611.
- Holt RA, Jones SJM. 2008. The new paradigm of flow cell sequencing. *Genome Res.* 18:839–846.
- Huang S, Li R, Zhang Z, Li L, Gu X, Fan W, Lucas WJ, Wang X, Xie B, Ni P, et al. 2009. The genome of the cucumber, *Cucumis sativus* L. *Nat Genet.* 41:1275–1281.
- Kozarewa I, Ning Z, Quail MA, Sanders MJ, Berriman M, Turner DJ. 2009. Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nat Methods.* 6:291–295.
- Lam H-M, Xu X, Liu X, Chen W, Yang G, Wong F-L, Li M-W, He W, Qin N, Wang B, et al. 2010. Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. *Nat Genet.* 42:1053–1059.
- La Scola R, Elkarkouri K, Li W, Wahab T, Fournous G, Rolain J-M, Biswas S, Drancourt M, Robert C, Audic S, Lofdahl S, Raoult D. 2008. Rapid comparative genomic analysis for clinical microbiology: the *Francisella tularensis* paradigm. *Genome Res.* 18:742–750.
- Li R, Li Y, Kristiansen K, Wang J. 2008. SOAP: short oligonucleotide alignment program. *Bioinformatics* 24:713–714.
- MacLean D, Jones JDG, Studholme DJ. 2009. Application of 'next-generation' sequencing technologies to microbial genetics. *Nat Rev Microbiol.* 7:287–296.
- Mardis ER. 2008. The impact of next-generation sequencing technology on genetics. *Trends Genet.* 24:133–141.
- Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen Y-J, Chen Z, et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437: 376–380.
- Mwololo JK, Munyua JK, Muturi PW, Munyiri SW. 2010. An overview of advances in bioinformatics and its application in functional genomics. *J Anim Plant Sci.* 6:645–652.
- Pacific Biosciences. 2010. Single molecule real time (SMRTM) DNA sequencing. Technology background. Available from: http://www.pacificbiosciences.com/assets/files/pacbio_technology_background.pdf
- Panaro NJ, Yuen PK, Sakazume T, Fortina P, Kricka LJ, Wilding P. 2000. Evaluation of DNA fragment sizing and quantification by the agilent 2100. Bioanalyzer. *Clin Chem.* 46:1851–1853.
- Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T, et al. 2010. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464:59–65.
- Quail MA, Kozarewa I, Smith F, Scally A, Stephens PJ, Durbin R, Swerdlow H, Turner DJ. 2008. A large genome center's improvements to the Illumina sequencing system. *Nat Methods.* 5:1005–1010.
- Shendure J, Ji H. 2008. Next-generation DNA sequencing. *Nat Biotechnol.* 26:1135–1145.
- The 1000 Genomes Project. 2007. Report of a workshop to plan a 10'' deep catalog of human genetic variation [cited 2010 Jan 30]. Available from: <http://www.1000genomes.org/docs/1000Genomes-MeetingReport.pdf>
- The 1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. *Nature* 467:1061–1073.
- Thompson IF, Steinmann KE. 2010. Single molecule sequencing with a HeliScope genetic analysis system. *Curr Protoc Mol Biol.* DOI: 10.1002/0471142727.mb0710s92.
- Wang J, Wang W, Li R, Li Y, Tian G, Goodman L, Fan W, Zhang J, Li J, Zhang J, Guo Y, Feng B, et al. 2008. The diploid genome sequence of an Asian individual. *Nature* 456:60–65.

Appendix 1. Primers for sample SNP check

chr10_49324409
TACCGTCCCCCTTCTGACCCTCACCAGGAAGTGCATGGTTGGAGCAGCATCTGATCCCACCTGGAGTGTGTGGGGTCCCCAAA
GTCCCCACAGTCCCCACAGAGGTATCAGGATCTCTCTCTCCAGCTGGCTCCAGAGCGCCTGGCTGCTCCAGAGATACAGA
CGCTTCTAAGTCCATACAGGCTGGAGACCAGCAGA[C/T]GTGGTACAGAAGTGAGCTCTGCCATTCTTTACTTTGGGGCC
CTGGCACCTTTTGGCCCAACAGTCAAGCTTCTAGGGTCGAAAAAACTCCTCCATTTCCTCTGCAACAGCTGGTTCCTCC
TCTCCGCATCCTCCCGCGCCGTTCCAGAGTTCCGCAGCTTTATTTCCAGCATGATGATTTTTTTCTTTTCTCTGGT
rs1051507 0.490
chr10_120907435
TCTAAGGAGAAAAAAGTACTTAATTAGCAAATATCACAGTTTTATCCAGCAAAGATAATTTAAAAACAAACACATACAACA
AACTCCAAAAGGAATAACACAAAACCCACACTCTCTGGGTCTCATAATACTCACGTAACCTCTGTTTCCATTGATGCTGTTG
AACGCTGCCATGTAGGCACAGAGGAAAACCTGCC[G/A]AGAGGGGCAAATGGATCGCATTTTCGTTTCAAGCATAACAGAAT
TTAGTAATCGTGAAACAGGAAGAATGGCTACCTGAGGTAATAATCACGGACTTGATCCCTTTTCACTGGCGTCATTGACAAAAAT
ACCTTTTAAGAAGACAGTGAACAGAGTTACAAGTGTGCTGTGGAAGGAGAAGTGCAGTAGCAAAGAAGT
rs2275111 0.496
chr12_23587383
ATAATTAGTATGAAGACAAACAGCTACATGAGCTATAGAAATATGGCATATTTCTCAATGGACATAGAATGCTTCATCACCAGA
AGAAGCCTCAATGATGATGTTTATTTTCTGGATAAGCAAATAACACGACCTGTATGGCCTCCAAATCCAGGATCCTTCCACAA
CTGCAGAAAAACAGAACAGTACCTTTATTCATA[C/T]GTCACTTAGTTCTTAATGGTTTTCTTACCCAATATCTTGCTGATG
TTGGAGTTGTGCATGTCAGGAAAGGCTTGAAGGATCTTTCTCCGTTTCATCTTAGCCACACCATGAAGGCATTCAATTGGACG
CTTTATGTGGGGTTCATTGCTACACGCCCTCGGGATTCCCTATAAATTTCTGACTCTGAGACTCCAGCACT
rs7485662 0.515
chr12_39616798
CAGCTTTTCAACAAAACATGATTTACAAATGGAAAAATAATGGCTGAGTTATAGTTACCTCCTGCTGTAATAATACAGGATTTG
CGTAGGTTAAGCCTATGCAAAGACAAGGCCTATTACTGAGTTCAGGATGCATACACTATCTGATGATTTCTGCAGAATTAGGTT
AACTAAATGTAATATTTATACTCTGGAGATTA[T/C]GATGGATCTTGAGCCCTTCATTAAACATTTACAATGTTTCCTACTTTTT
AGCATTCCCTGAGTGGGTAGAACACATCAATGACACAGAGGTGGACATAGGCAGTGATCTCTACTGGCCTTGTTGGGCC
ACAGGAAAGCCCATCCCTACAATCCGATGGTTGAAAAATGGATATGCGGTATGTATGTTCAAGTGCTTTGCT
rs2289522 0.511
chr12_119479715
ACCTACTGCAGCCAAGATAACCCGTTGTGGACACATCTTCTGCTGGGCATGCATCCTGCACTATCTTTCACTGAGTGAGAAGA
CGTGAGTAAATGTCCCATCTGTTACAGTTCCTGTGCATAAGAAGGATCTCAAGAGGTGAGATTGAGACATTTACTCAGTTAGA
TCCCAATCTCTTCACTCCTTGCCCTGCTACATAT[G/A]AGTGTCCATGTTTCAGTGTTGTTGCCACAGAGTCACATCAGTATGT
TGTTGGTGATACATTACGATGCAGCTGATGAAGAGGGAGAAAGGGGTGTTGGTGGCTTTGCCCAAATCCAAATGGATGAAT
GTAGACCATCCCATCTAGGAGGTGAGTTCTTTAAATTTTGGGGACAAATAATCCTGGGTACATTCTTT
rs2292681 0.503
chr13_100062634
ACCACAGTGGTTAATTCTGATGAGAAAGGACACTGGGTTTGGCCCCACACCGCTGTCTTTGAAGCAACTTTTCCAGGAAACA
AACTGGCCAAGCCACATTGGTGCACCACTGCCACTGTTTTGGCCTGATTCTTAATTTAACTGGGCAGAAGTAACACTTACGTC
TTCTGTGAGATGCTGGCACTAAGTCAAGAACTCA[C/T]AGCCTGCGGTTTACGCACCTTGATTTCTGGGATTTCCCGAGC
ACGTTTGCCAACGAGAAGTATGAGAGGTGATCATTAGGTATTAATCCAGTGCCTCTCTTCCAAGTCTTACAGTTGGGCTAA
ATTACCTGGGAGTGAAAAATGGAAAAATAAAGACATGTTAATGTTGTACTCGCATTTTAGCTTCTCTCTATAA
rs1572641 0.485
chr15_75131848
GGAGGCCAGTACCTCAGTCACTTCCACATTACCTACATACACAAAGATTCACTGTGATGGGGATTAAACCATTACAGGGCTGA
CAACAGAGAACCAACTCACTCCGATAGAACTCCTTGGAAGCTGTGCTTACCTGAATAGCTGCAAATGCCAGTGCGGCCCA
GATCACATGCATCATGATTTCTTGTAGCTTCTTAC[T/A]ACTAGAGCCTCACACCCCTGTAACAAACACAGTCATCCTCGTTA
GAAGTGTCTTTAAAGTTAAAGCTGCAGATTACAGGCCACTACCCACTTACAGAAAGGAAAATGACAGAACTGAAAGA
GTGGAAAAAAGATTCCATGATTGGTATAATTACAGACATGTAATTAACAAAGACATGCATTTTATAATAATT
rs11737 0.507
chr15_88027951
AGGTTCTTCACTGTGTCCAGGAGCAAGGGAGGATGCTGCTTACAGAGATCGGAATAAAAGAAGTAGAAAGGATTGGAGCCAT
TCTGTTTCTCTCCTCAGCCACGCTGAACCAAAGAGGATCCTGGGATGCTGATTTCTCTTTCTTTGCCCTGTACATGTAACCTC
GTTTCATCTGCAGGGAGATTTTACATACAGATCCCTGAC[C/A]AGGCTCAGCAGAAGAGAATAGTAGTAGTGGTGAGCAGCCC
TCGTTCCGCATTTCTCTTTGTTGATGCCAGAGGTGAGACCTACGCTCCTCACCCAGAGGATGGTGTACAGATGAAATAAATC
ACACGGTTCAGGTTGGCTAATGTTAAGCATAAGCGAGGTACCAGGTCAGTGGCATGAATGCTCTGCTCAGTTGCTCTG
rs7169981 0.496
chr17_5932068
TCACTTTGCTCAGCACTGAGTGGGAGGGAGGTAGCCTGCAGCACTCATCTCCCGGGTTTCTCTGAGAGTGGCCTTGCCAGG
AAGGCAGAGGTTCCCTGGCTTTTCTAAGGGTGGGCCCATGATTCTGGCCTCTGGTATTAGCCATGCTTCCCTTTCTGCA
GGCACCTACATTGGATGCTTCAGTGACGATGGCCA[C/T]GAGAGGACTCTGAAAGGAGCTGTGTTTTATGACTTGAGAAAGA
TGACTGTCTCCCACTGCCAGGATGCGTGTGCTGAGCGGTGAGTGCTGGGGCCCTGGACTGTTGATTCTAGAGGCAGGAAG
GTCCAGGGACAGTTCCAGAATGAGGAGGCATATCAGCTACCAGTTGGCTGTATAGCATAATTGGAACCTCACTGCTC

rs2302836 0.507

chr17_42723336

CCTGCGCAGGAAACAGAGTCCTAGGGGTGCTCCGGCAGCTTCCAGGTCAAACCTTGGCCTCTGGCACTGCTGGGGTGAGTC
 CAGAGCTACTTGCCAGATTTGGCTTGGGATGGAACCTGGCTCTTGCCATTTCCCGTTTCCTTTCAGTTCAATTTCTTGCTT
 CTTGTGCCCCTTTCTGCTCAGAAAATCCGTTCTAAAGT[A/C]GAGCTGGAAGTGCCTGACCTCCCTGAAGAGTTGTCTCTAT
 CCTTCAATGCCACCTGCCTCAACAATGAGGTATCCCTGGCCTCAAGTCTTGTATGGGACTCAAGATTGGAGACACGGTGAG
 GTGGGCTGGGCAGGGCCTTTGTCTGGAGCATCTGTGGGCACCCAACCCCTTCTTCCTTTGTAGTGACTAAAAAT

rs15908 0.513

chr18_50416306

GTGTGATAATGACAGCAATTGGAGTACTTATAATATGCTTGGTGAATAACAAAGGATCGGCCAATTCCTCCATTGTTATCCAGC
 TATCCACAAATGATGGAGAGTGTGTGACTGTCAAACCTGGAACACCCTCTCTGCTTGTCCACCTACAATGACCACCACTTCA
 ACTGTACCTGCAAGTACAGCCACTGAATCTACA[A/C]CTTCAACAGCTACAGCTGCCACCACTTCCACAGAACCTATAACTGT
 TGCACCTACCGATCATTATAATTTGAACAGCCATAGCCATCACTTCAACTGAAACTTCAACCTCTACCACTTCAACTCA
 GTTTGCAACTATAATAGCTACTCCTATCACTTCAAGTGGAATCATGGCTGCAGTCACTTTAACAGACTCTATAA

rs9947055 0.484

chr18_59321701

GGAAATAAAATAAAAAATAATTTCTTTTCATTTTTGTCTTCAGATTGAAAAACAACCTCAACTCAGAGTCACTGTCACAGTGGAC
 TAATCCCAGCACCATGGCCAATGCCAAGGTCAAACCTCTCCATTCCAAAATTTAAGGTGGAAAAGATGATTGATCCCAAGGCT
 TGTCTGAAAAATCTAGGGCTGAAACATATCTTCAGT[C/C]GAAAGACACATCTGATTTCTCTGGAATGTCAGAGACCAAGGGAGT
 GGCCCTATCAAATGTTATCCACAAAGTGTGCTTAGAAATAACTGAAGATGGTGGGGATTCCATAGAGGTGCCAGGAGCAC
 GGATCCTGCAGCATCAAGGATGAATTGAATGCTGACCATCCCTTTATTACATCATCAGGCACAACAAAACCTCGAAA

rs145556 0.513

chr19_38327601

TTCCCGCTCAGGGGACGGACATGTTGCAAAGTCCTTTCCTCTTTCGGAGCTGAGTGAGAGTCCCAGGACCTTGGAGTTCTCA
 GATTTTGTCTACCAGAGATAAGCTAAGGGCTAATGTTGACCTCGATTTACGACTTCAGCAAACACGGCAGTGACCACTCTCTTC
 TCCTCTCTCTCCCGTTGATTTGAGGATCCGGTGGA[C/T]GGTCTGTGGTTTCGCGATTTTGAGCACAGGCCCAAAGCTCCTGT
 TGTAGAGTGCAGCATCACCGGCGACAGCAGCAGGTGACCTTTCCTCGTCTTACACTGGGAAGAGGGAGGGTGAGCAG
 GCATGCCATAAATAGGTATGCCTGTGACATTTTTGTTTTTAAGTAATAGATATGGGGTCTCACCATGTGGCCT

rs3848596 0.501

chr1_117423809

GAATCATTTGTAGAAAATAGGAGAGTGTAGAGTTCGTGTAGCCAGGCTCGGCATAGCCTCTCATTAGGAATTTATGCTTAGCAT
 TTGGTTATTACCTTTGTCTAGGTCCCTTTGCTACTACACCAGAAGCAGGCATTGGCTTGGTTACTATGGCGAGAAAGTCAGAA
 CCACAAGGAGGAATCTGGGTAAGTGTGGTATT[A/G]TAAGAGTCAGCCTTTATTGAATGCTTAGGCATTGTGCTAAACA
 CGTAGTGTGTTTTCATTTATTCCTCTCAAGAAGTCTATGAGGCAGGTACTATTATTCCTCATTTTACAGATGATGAAAGTGAA
 GCATGGAAGTGTAAAGCAACTTGCTCAAGAATACACAGTTTAGGATCAAAGATGTTGGTTGCCAGCAGTAGAAG

rs2801939 0.519

chr1_153373174

TGAGGGAATCCTGTTGGGCTGAGAGCCAGTACAAATAGTGGAGCCACTCACACTGGTGGCCTTTGCCCTCTCACCTTGCGA
 TGACCCAGAGGTGCGGGTTCTACATAGCATCGGTACAGTGCTGCCCCACGCCTCTTCCCACTTGCCCTGGACTGTGCTGCTCC
 TTCCACTTCTGCTGCTGCAAACCCCGTGAAGGTGT[A/G]TGCCACACCTGGCCTTAAAGAGGGACAGGCTGAAGAGAGGG
 ACAGGCACTCCAAACCTGTCTTGGGGCCACTTTCAGAGCCCCAGCCCTGGGAACCACTCCCACCACAGGCATAAGCTATC
 ACCTAGCAGCCTCAAACCGGTCAGTATTAAGGTTTTCAACCGGAAGGAGGCCAACAGCCCGACAGTGCCATCCCCA

rs9297 0.511

chr20_49671952

CTCAAAGGGACACGACCCCCAAAAGGTTAAGAGCATGGCATTAGCACAGGGCTAGAAGGAGGTACTCCCAAACCACTACT
 GGAAGAACTCTCCAGCCGGCGCCTGGGTCAATTAATCATCCCATCCAAAGCAACCGTGGTTGATGCTCCCCACCCTAGGTA
 ATTCACATGCTGCCCCCTCCCTTCTTAAGGGGCCACTC[A/G]AGGTTAGGAGGTATAGGACAAATCGAACATCATCCCCA
 CATGGCCCTAACGCAGAGGACACTTTACCAGCCGAAAAACGTGGATGTCTTGGTTTCTGGTCAACAGATGTGCATTCTTCGC
 TGTGCACGTAGCTGTCTCCAGCTTGTCCCCTGTCAGCATCCAAACCTGAAATCATGGCGTGACAGAGGTCTGTTTATT

rs2235862 0.515

chr20_52219626

CCTTTCCTAGGCAGCAATAATGCCTGTTTACAAAAGAGTTGTCAAAAAGAGCCATGTTTTATCCGCAAAATCACCTGCAAAAT
 CAGTGAGCAAGTCTGTGACGACAACTACTTTCAAACGACCATTGTTCAGTTTCGCTGTACAAGTCTTCAACGTGGCCTCTTT
 CATCACAGAGCTCATCTATTCTGCCCCATAAAATC[G/A]GCCAAGACCTTCAAAGAAAACAACCGCAAAAGACACATTTAAAG
 CCGTTGAAGGAAAACAAAACCTTTAAAGCTCACTGAGCTAACTTCAGGTCTTGCTACATCGCATATTCACTGGCCATAATGTTA
 TTCATCTGAAGTCACACTAATTATATATTCTCCCTGGGAACCCACCTTATCCCCTAATGACCTCGTCTCG

rs2296241 0.497

chr22_35237044

GAGGGAAAGACTAAACAGATATATATTTTATTTTCATCTGCTAAATGTCAGAGAGCAAGTTAGACACACATTCCACTAAGCATCA
 AAGGCCCTCGTAGTCTCGGTGGAAGGACAACTCCAGCTCCACATCACTGGTTTAAAGTTTCTTCCTCTGAAAGACACAAAAA
 ATGTAAGAGAAATGAGCCAGGTTGTGTGATATGA[T/C]GTTCACTTTCCTTTCTGAACCTCTGCTGGACTGTGCGTGAGTTTT
 CAACACTTCCATTCCGAAGTTAAGCCTTTCAGATGCTCAGGGCTGGGCAGGTATTTCCCTTCAGTCACCAGCTGGTGGA
 GCAGGAGAGACAACCTTCAGGCTATTTAACCCGGAGCGGACTTTTCCCTCTAGACCACCGGCTTCAATCTGGGA

rs728785 0.518

chr2_141468062

TGCAAAGGAGGCAATCTATGGTCTACTCTTTATGCAAAAGGATGACTGAAAAATTAACCTAGGTTCAATCTGCATCAAAGGTA
TTTTCAGATTGTCACAAACTGCCTTGATAATTCACACACTCAAAGACTCAGGAAAATACAATTTTAGGTGACTTGGAAGAG
AACATACACAGCACTTAACTGAATCCTCATCGCTTCC[G/A]TCTAGGCAGTCATCGTCGCCATCAGATTTCCACCGAGCTTGA
TACAGTGTCTGTTTTGTCAGCGAAACTCTCCAGCTTACATATGTGAGGTAGTGCTTCTCCAGATTAACTAGAGTT
AAAAAAACCCAAGGTATATGAATAATGCTAGCTACAGTCAGTCCATTAATAATTTAAATCTTGTAAGAGTTTTTTGTTTGT

rs13007735 0.485

chr2_204532528

GTCTGTTATAGGAGTCGAGAGCIAAACTTCTCTGGAAATGGGGAAGTGGTGTCTGGGTGTTGAAGCATAAAGATGAGTTTGCA
TGGTGTGTGTGTGAGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGCACGTGTGTGTTTGTGTGTATGAAAGGCAATGGAGAGGGGAAAG
CTTCTTGATAGGGAAGTGGCAGATGGAGAGCGATTTAGA[T/G]AATTTATGCTGAATTTTGTGTACAGATGTGACCCATAATATG
GAACCTCTGGCACCAGCGATGAAGACGTGTGGCCAGTTTCTCTCAACTGAAGTGCAAGATCTCTTATTTCGGGACCACG
GAGAGTCTGACTTAACTACATACATCTTCTGCTGGTGTGTTTGTTCAACTGGAAGAATGACTGTATCAGTCAATG

rs10172036 0.488

chr3_7595789

ATGTGCATCAGTTATGCAGCCC TCTTGACGAAAACAAATCGGATTTATCGCATATTTGAGCAGGGCAAGAAATCAGTAACAGC
TCCCAGATCATAAGCCCAACATCACAACCTGGCAATCACTTCCAGTTAATATCAGTTCAGCTTCTAGGGGTGTTTCATTTGGTT
TGGTGTTGATCCACCAACATCATAGACTA[T/C]GATGAACACAAGACAATGAACCCGTGAGCAAGCCAGAGGGGTTCTC
AAGTGTGACATTACAGATCTCCAATCATTGTCCTTGGGATATAGACTTCTCTCATGGTCACATGTACTGTATGCCATC
AAGACTCGGGGTGTACCCGAGAATTTTAACGAAGCCCAAGCCCATTGGATTCACTATGTACACGACATGTAT

rs1485175 0.493

chr3_187925712

GTGTATGCGAATGCTGATTGCGATCATTAAATGCTGTGTTTTTGTCTGCTCTTTGTAAAGCCGAGGGCCCTGTGGTGACAGCCC
AGTACGACTGCCTCGGCTGTGTGCATCCTATATCAACGCAGAGCCCAGACCTGGAGCCCATTCTGAGACACGGCATTTCAGTA
CTTTAAACAACAACACTCAACATTCTCCCTCTTCA[T/C]GCTTAATGAAGTAAACGGGGCCCAAGACAGGTTTGTCTTTAA
TTCTCTAAGTAGCAGATTACTAAAATTTGTAGATCTTACGTTTAAATGATGATTGATGATTGCTGGTGGTGTTCATGT
GACTAGCAGAGGAAGCGAAGAGCTTTCTCCTTAGATCAAAGTGATTCGGATAGTGCAATATGTAAAGTGAGA

rs1656922	0.489
-----------	-------

chr6 168058593

TCTTTAATAAATTTTAACGATTAATTTGTTTTGTTTACAGATAGTTCTCATTTTTGGTGTGTCATGCTGGCCATACTCTTTAAAAAGT
CTCTTAAGGCCAACTAATGATGTTAAGTGTAAATTAGTTAATAACTATATGTAGTTTAAGATCATGTCACTGTTTCCTCTT
GGAATGAAGTTTGATATTTTAAAGTAATTAAC[G/A]GCATCTACTACCATAGGTACCATTATCTAAATCTGCTGCACATTCAA
GGCTTCTTCTGGATGACATTTTGCTCTCTTGTCTCTACCCTGTCTGGCGAGATGATGTGCTGCACAGCTCACAGGA
GCCATGTCTTGTCTACGACGCTGCAGAGCAATGCCGCCCTGACCATCCAGCTCTTCTCTCAGCTCTTCCA

rs3213589 0.497

chr7 47935452

CCAACACTGCCTCTTGGGTCCCCTAGCTCAGGACCTCCTTCCCAGGCTCCTGATAAGGACACCATGCACCTTACCTTTTGAG
TACTGGTGGTAGGCAGTCACCGCCATTGCCTCAGACATGTTGTGTAGCCTCATTTCAACCCAGAAGTGTCCCCGAAATCCAT
CATCAGACAGTAGAGCCTCTCCAGAAGCCATATGAA[C/A]ACGGAATCCAGATTGGAGGTGCCCGAGCTTCCACTTCCA
GGGAGCAATTAGAACCAGGGTTTCATGAAGTTATCAGAATTCGAGATTAGGATGGCCACAGGAGGATGCTGAGTAGGGGGATA
GAGGATCTCAGAACCAGACTGCGATGCGGTGAAGCTGGGGGTGCGAGGAATGCCAGGCGGAAGGCCGTGGAGGACGTGT

rs2686817 0.519

chr8_59660744

TTCCCTTTTGGAGTCCAATGGCTGTTAACATAAATCTAAATCCTGAGTAACCTACGAGATGCAGATCACCTGGCCAATCAAGA
GGCTCCAGGGACATGTTACGACATGGAAAAACCATGGGTGTTTTTGCCCAAAAAGAGTATGTTCCCTGATGGAAAAGGC
AGGCTTGAGTTTCATTATCTTGAGAACAAAGATCAAG[G/A]CAACTGCAGTAGCTGCTTACCTGTGTGGCCCTGTATTCTC
TCACTGATTTTGTCTCCAGAGGTCCCAACGAGCAGTTCCACAGACTGACTGCCAGATAAAACGGAATTTCCATCCCA
GACAAAGCACCTGCAAGAATGATTAGAAATAGTCTTTTCTTCATCATGAAGGAAGGATATGTTGAACTGGCCAATGTAA

rs12677290 0.496

chr8 143308722

GGATGGTGGATGGGTGGATGGATGGATGGCAGATGGATAGATGGATGGATGGTGGATGGGTAGATGGATGGACAGATG
GACGGTAGACAGGTGGATAAATGGGTGGATGGATAGATGGATGGATGGTGGATGGTTAGATGGATGGTGGATAGGTGAATGA
ATGGATGGTGGATGTGCAGATGAATGGATGGTGAGTAC[G/A]GGTAGGTGGGGTACTCACCACGGGTAGCATCACTTTCTGG
ACAGAGTGGCGATGATGATGATGACAGACAGGCGAGTGCATCCAGCTGATAGGAAGCAGCACTTGATCTTGCTGTCT
CTGGAGCTGGGAGAGAGAGAGAGGTTGTTCCATACCTGGTCTCCAACCTCCCATCTCAGGGGGATTGGGCTTGTCGA

rs11167136 0.490

chr9 36159598

GGTCAACCGCTCTGCAAACAACCATCAATCTGATCCACAGGCCTGAGAAAAGTCTGCTCTCCAGTACCTGCTGCTGATCTGT
TTCAGCCGACAAGAGGCACCATGAAATTGGAATTCACGGAGAAAACTACAATAGCTTCGTGCTGCAGAACCTGAACAGAC
AGAGGAAACGCAAAGAGTACTGGGACATGGCCCTGAG[T/C]GTGGACAACCACGCTCTCTTTGCACATCGCAATGTGCTG
GCTGCTGTCTCCCCATGGTGAGGAGCCTCATCTCCAGCAATGACATGAAGACCGCTGATGAGGCTTTTCATCACCAATTG
ACACCACTTACCTGAGCCCGGTACAGGTGGACAGCTCTGGACTACTTATAGCGGCAAGGTTGGTGATCTCCGAGCAAAA

rs3739609 0.491

chr9_85446251

CTGGTACTTTAGGGCAATTTTATTATTATGAAAAGAGCCACTGACACCAGCTGAAAGAAAATGTTTCTGCCCAGCAGCTGG
 CCCCTCAGCCTTGTGCATGGCCCACTCCTCACACCCTGAGCTCTCGCTGGACCTTAACATTTGTGAACATGATATAACTG
 TTTCTCTACAAACAAACAATACCTGGAGAAGTCACGT[T/A]GTGTCTGGGTTTTTGTTTTTGTTCAGGACCGGA
 TTCCATGGCTCTGTAACCTTGCATGACATTTTACTAAGGTAAGAGACCACCAGGCCTTGGCATAAGCCAAAGCCAGCAGGTA
 AAGAAAACCCTCAGAAAAGACAGGGAAAAAAATACACTCCTGAAATCACAACCTGACTGATGGTAAAGGATTCTAGAAACT

rs2811917 0.519

chr9_106307125

GGTTTGTGTGGACACCTCCCTGGTGCAGTTAATCATGCTGGTGATCAGTGTACTTCTTCTCCCCATGCCAATGCTACTCATTG
 TATCTCTTATGCATTTATCCTCGCCAGTATCCTGAGAATCAGCTCAGTGGAAGGTCGAAGTAAAGCCTTTTCAACGTG
 CACAGCCCACCTGATGGTGGTAGTTTTGTTCTATGGGA[C/T]GGCTCTCTCCATGCACCTGAAGCCCTCCGCTGTAGATTCACA
 GGAAATAGACAAATTTATGGCTTTGGTGTATGCCGGACAAACCCCATGTTGAATCCTATCATCTATAGTCTACGGAACAA
 AGAGGTGAAAGTGGCCTTGAAAAAATTGCTGATTAGAAATCATTTTAATACTGCCTTCATTTCCATCCTCAAATAAC

rs7030820 0.502

chr9_124464328

TTAGCCTTTGGAAACACAGACAGTTACCTGCTAGCAGCCATGGCCATTGACCGCTATGTGGCCATATGTAATCCCTTCCACTA
 CATCACCATTATGAGTCACAGATGCTGTGTCCTGCTTCTGGTTCTCTCCTTCTGCATTCCACATTTTCACTCCCTCCTG
 CACATTCTTCTGACTAATCAGCTCATCTTCTGTGCCTC[A/C]AATGTCATCCATCACTTTTCTGCGATGATCAACCAG
 TGCTAAAATTGTCCTGTTCCCTCCCATTTTGTCAAAGAAATCACAGTAATGACAGAAGGCTTGGCTGTCATAATGACCCCGT
 TTTTCATGCATCATCATCTTTATTTAAGAATCCTCATCACTGTTCTGAAGATTCCCTCAGCTGCTGGAAAGCGTAAAGCATT

rs70156 0.512