Novel Application for Accurate Monkey's Exome Sequencing

– Empowering research for your drug development

1. Introduction

Due to the close proximity in evolution, non-human primates are often used as models for studying human diseases and drug development. They have contributed to pre-clinical studies, including the discovery of vaccines, drug development, and behavioral research[1]. Cynomolgus macaque (*Macaca fascicularis*) is the most widely used non-human primate among these models[2].

It is important to know the genetic information of experimental materials during the pre-clinical research process. Before the availability of BGI's monkey exome capturing array (MECA), most researchers used CNV microarrays or whole genome sequencing to obtain detailed information from monkeys. The CNV microarray restricts the ability to detect data that are specific to annotations and contents. Moreover, the important, interpretable part of genome, the exome, represents only 1% of the genome. Monkey exome sequencing using MECA can sequence more samples and is more precise at a lower cost than whole genome sequencing.

BGI recently completed the whole genome assembly and gene annotation of rhesus macaque (*Macaca mulatta*) and cynomolgus macaque (*Macaca fascicularis*),[3,4] with 150 G of sequencing data at an average depth of about 50X. Based on previous research and our understanding of genome sequence information, BGI has released the first monkey exome sequencing platform using next-generation sequencing technology and MECA. MECA is a proprietary exome capture array designed by BGI for capturing the entire monkey exome. The combination of this revolutionary array and BGI's high-throughput sequencing technology not only simplifies the workflow of exome sequencing experiments but also improves cost-effectiveness and turnaround time.

Here we report the results of monkey exome sequencing using MECA to capture the whole exome combined with Illumina HiSeq™ 2000 sequencing.

2. Methodology

2.1 Experimental design

Sequencing of the Chinese rhesus macaque (CR) and Cynomolgus macaque (CE) was conducted using MECA and human exome array to capture all exons. They were then sequenced on Illumina HiSeq™ 2000 in parallel. To test the capacity of MECA, we pooled two samples into one library in a single array. Output data were mapped to the reference sequence, Indian rhesus macaque. Sequencing quality was measured based on data mapped to target region, coverage of target region, and distribution of per-base sequencing depth.

2.2 Workflow

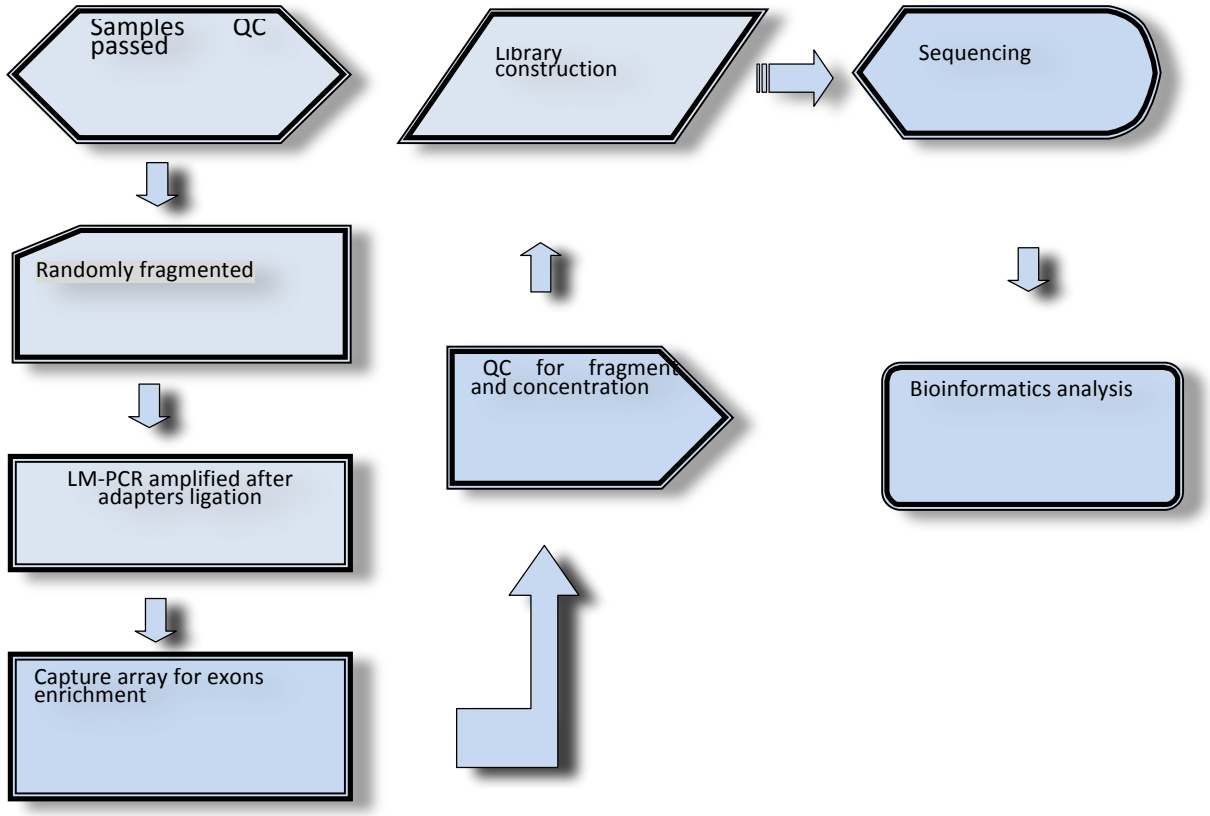

Figure 1 The workflow of monkey exome sequencing

2.3 Exons extraction and enrichment from the whole genome

MECA was designed by BGI and produced by NimbleGen Roche. Its capacity is about 50M, covering all exons of the Chinese rhesus macaque and Cynomolgus macaque. The library was prepared using Illumina Paired-End Genomic DNA Sample Prep Kit. Sample and exome libraries were hybridized after amplification using LM-PCR. Then, DNA segments that were not complements of probes were washed off and captured DNA was amplified by LM-PCR again. We performed qPCR as a QC step to measure the enrichment efficiency and used Aglient 2100 to QC the library.

2.4 Results:

⌘ **Output data and average depth**

We sequenced three types of libraries from each Chinese rhesus macaque (CR) and Cynomolgus macaque (CE) sample, including one sample with one MECA capture (labeled CR_Single and CE_Single), two pooled samples with one MECA capture (labeled CR_Pooling and CE_Pooling) and one sample with one human array capture (labeled CR_human array and CE_human array). The output data of the CR samples were all approximately 1.3G. The output data of the CE samples were 1.5G at approximately 40X sequencing depth. The data were mapped to the Indian rhesus macaque as a reference. We found that the mapped data captured by MECA were greater than that captured using the human exome array, indicating that MECA has higher specificity.

Table 1. Output data of CR and CE captured by MECA and human exome capture array

Output Data	CE Pooling	CE Single	CE_human exome array	CR Single	CR Pooling	CR_human array
Data mapped to target region (Mb)	1,407.59	1,380.94	1,180.97	1,507.45	1,529.93	1,247.36
Mean depth of target region	45.09	44.23	37.83	47.08	47.79	38.96

⌘ **The distribution of per-base sequencing depth**

As shown in Figure 2, the distribution of per-base sequencing depth from the data of single and pooled samples captured by MECA approximately follows a Poisson distribution, which indicates that the target region nucleotide is evenly captured. However, the distribution of per-base sequencing depth from the data of samples captured using the human exome array did not follow a Poisson distribution, which indicates a part of the target region may be missed. These results are shown in Figure 3.

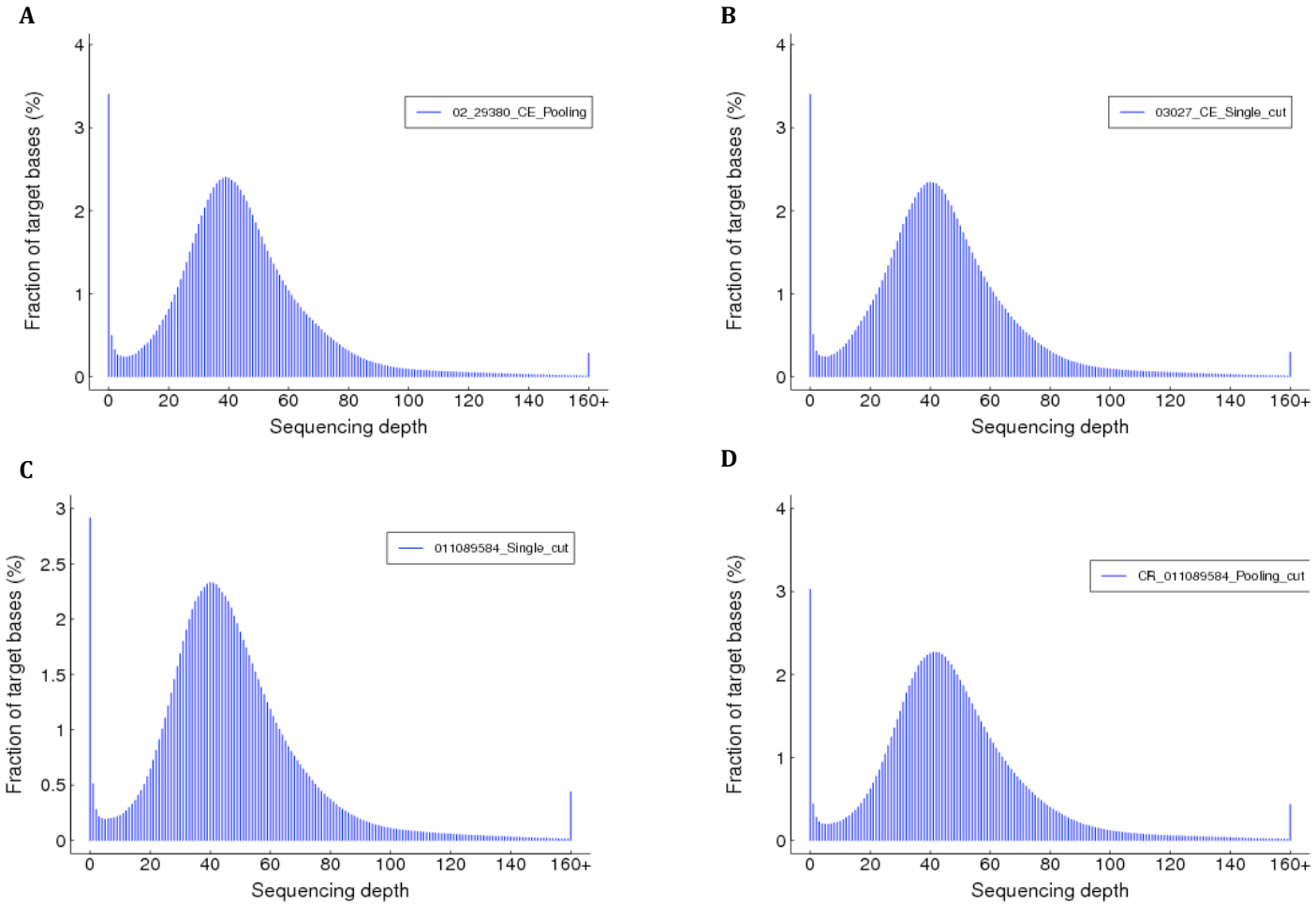

Figure 2. The distribution of per-base sequencing depth of CE (A & B) and CR (C &D) captured using MECA

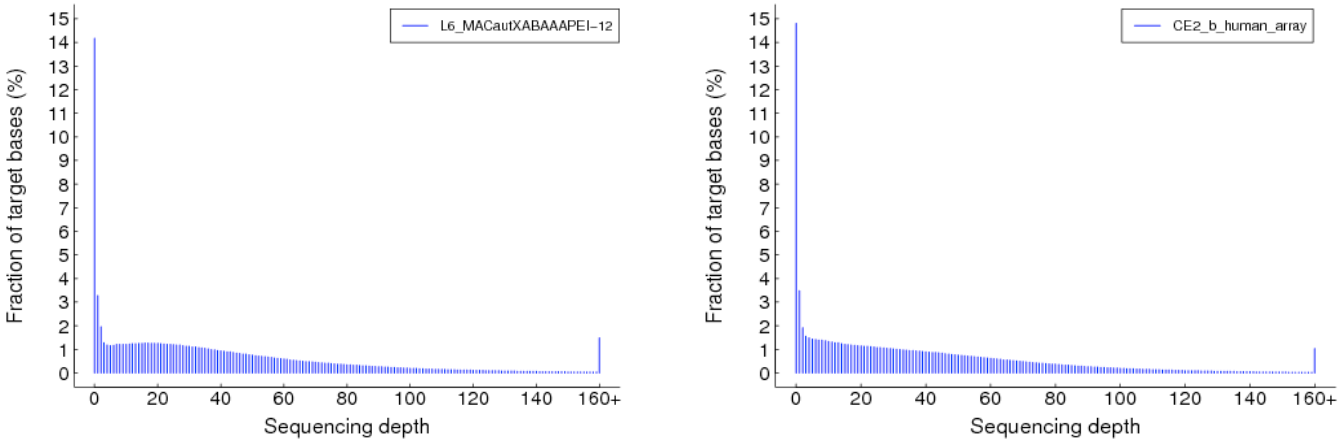

Figure 3. The distribution of per-base sequencing depth of CR (left) and CE (right) captured using human exome array

The X-axis represents the sequencing depth; the Y-axis represents the percentage of total target regions at a given sequencing depth.

⌘ **Coverage of target region**

As illustrated in Figure 4 and Figure 5, there is no significant difference between pooling or single library constructions. However, the coverage when using the human exome array capture on monkey's samples decreased dramatically. At the same time, MECA results were nearly 90% at 20X, which indicates that the performance of monkey exome sequencing is as good as mature human whole exome sequencing.

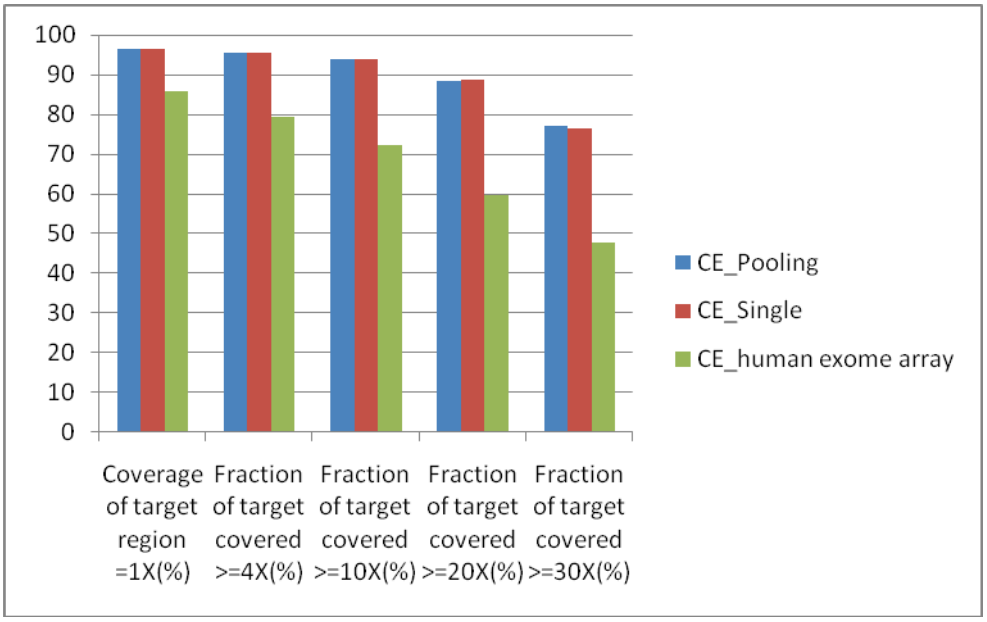

Figure 4. Coverage of target region statistic for CE using both monkey exome array and human exome array

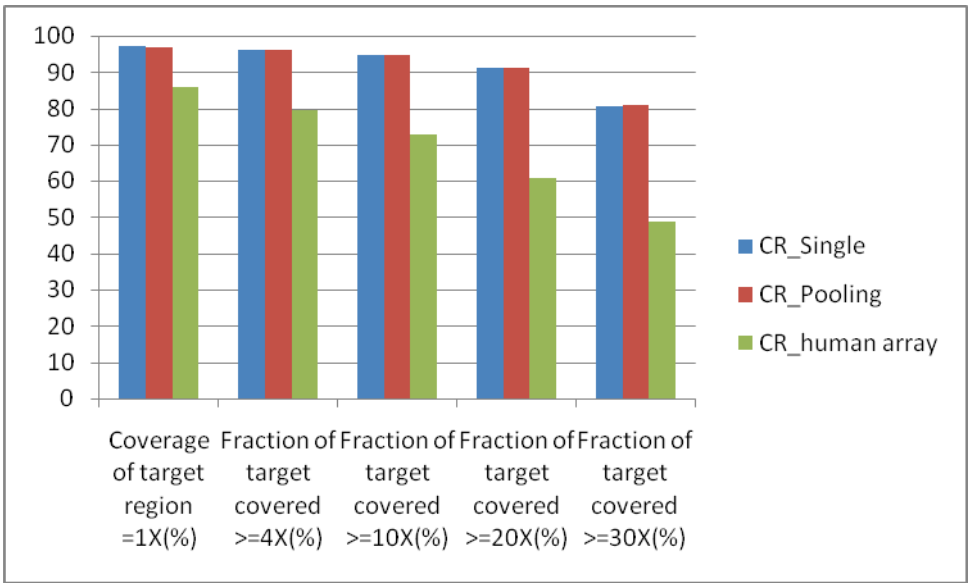

Figure 5.Coverage of target region statistic for CR using both monkey exome array and human exome array

Coverage of target region is the number of base sites covered by no less than one read in target region/the size of target region, which is an important indicator for sequencing quality. Fraction of target region covered>=4x means the number of base sites with a sequencing depth of no less than 4 in the target region/the size of target region, which is also the threshold for SNP calling. Fraction of target region covered>=10x: the number of base sites with a sequencing depth of no less than 10 in the target region/the size of target region. Fraction of target region covered>=20x: the number of base sites with a sequencing depth of no less than 20 in the target region/the size of target region.

⌘ **Capture specificity**

To assess the quality of probe design in the capturing array, capture specificity was measured using the ratio of the effective sequence on target (Mb)/the total effective yield (Mb). A comparison of the results shows that MECA has higher capture specificity than commercial human exome array on monkey samples as shown in Figures 6 and 7.

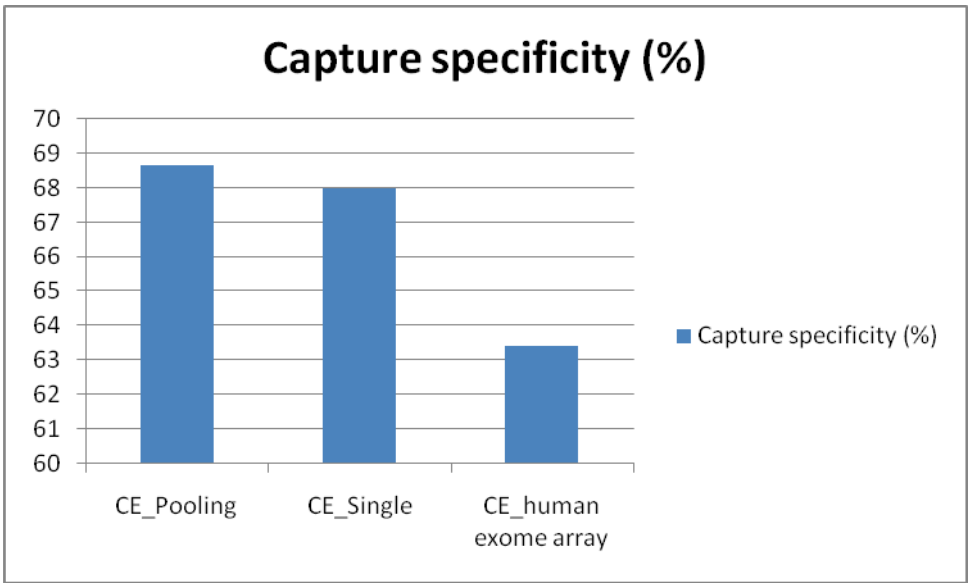

Figure 6. Capture specificity statistic for CE using both monkey exome array and human exome array

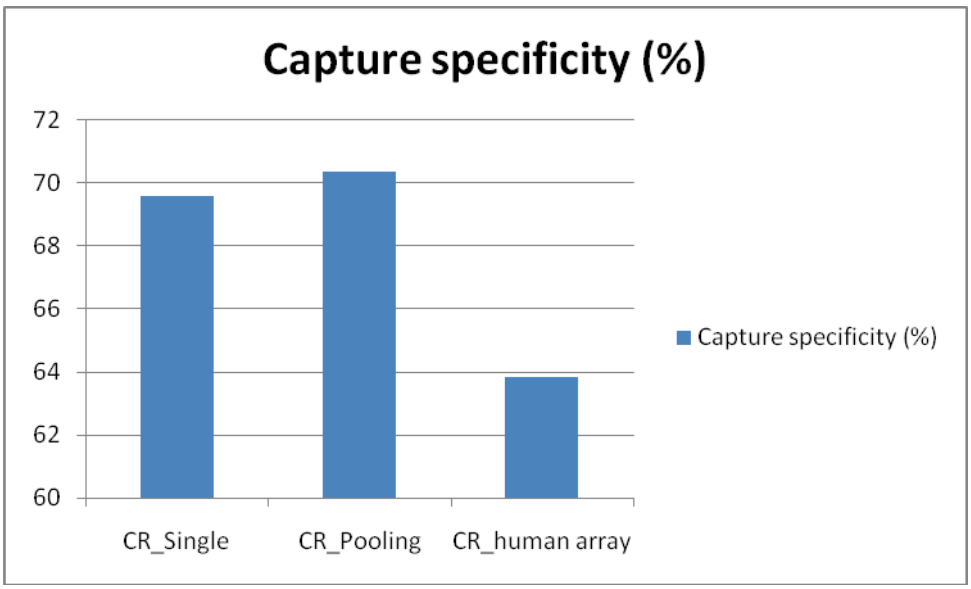

Figure 7.Capture specificity statistic for CR using both monkey exome array and human exome arra

3. Conclusions

Monkey exome sequencing methodology is an efficient research strategy to selectively sequence the coding regions of the monkey genome. Based on the comparison of the results of single sample with one MECA, two samples with one MECA, and single sample with one human exome array, we conclude that:

1. The two-pooled samples with one MECA condition show similar performance as single sample;

2. MECA is more suitable for monkey exome research than the human exome array;

3. MECA is currently the best technique for monkey exome sequencing without compromising results.

4. References

[1] Vanessa Schmitt, Julia Fischer. Representational format determines numerical competence in monkeys, ***Nature Communication***, 2011. 2:257

[2] Steelandt S, Dufour V, Broihanne M-H, *et al*. Can Monkeys Make Investments Based on Maximized Pay-off? ***PLoS ONE***. 2011. 6(3):e17801

[3] Fang X, Zhang Y, Zhang R, *et al*. Genome sequence and global sequence variation map with 5.5 million SNPs in Chinese rhesus macaque, ***Genome Biology***. 2011, 12(7):R63.

[4] Yan G, Zhang G, Fang X,, *et al*. Genome sequencing and comparison of two nonhuman primate animal models, the cynomolgus and Chinese rhesus macaques, ***Nature Biotechnology***. 2011. 29(11):1019-23

Appendix

➲ Sample requirement:

1. Sample condition: DNA samples without degradation and RNA contamination;
2. Sample quantity (single): ≥ 6 μg (for two library construction in the event of a failure);
3. Sample concentration: ≥ 50 ng/μL;
4. Sample purity: $OD_{260/280}$ =1.8~2.0.

➲ Turnaround time:

Turnaround times are based on target region sequencing due to the fact that extra arrays are not prepared, and we have to wait until we receive the capture array from the supplier.

➲ FAQ:

1. What is the advantage of exome sequencing technology?

Answer: Regardless of the depth of research on the genome of monkeys, the crucial part of the genome is the exome which accounts for only 1% of the genome. Exome sequencing analysis is more efficient, reliable, and accurate for revealing common, rare, and novel variants located in the exon regions.

2. How much of the monkey genome can the array cover?

Answer: The array can cover 50M of the monkey genome. However, we can add other regions that may be of interest to clients, in which case the capacity can be enlarged.

3. How can I determine a suitable sequencing coverage depth, and how much data will be generated?

Answer: We recommended sequencing at 50X minimum which represents 90% of the detectable bases within the exome could be captured. The data output is determined by sequencing depth and targeted regions.

4. What key factors affect the analysis of exome sequencing?

Answer: The main factors that affect the results of exome sequencing include the quality and quantity of DNA samples, the capture specificity of hybridization, and the quality of the sequencing. BGI's experience shows that we provide stable performance for exome capture and sequencing.

Contact Us

North and South America

BGI Americas Corporation

One Broadway, 3rd Floor Cambridge, MA 02142
U.S.A. Tel: +1-617-500-2741 Fax: +1-617-500-2742
Email: info@bgiamericas.com

Europe

BGI Europe

Copenhagen Bio Science Park Ole Maaloes Vej 3, 2200
Copenhagen, Denmark Tel: +45-7026-0806
Email: bgieurope@genomics.cn

China

BGI Shenzhen (HQ)

Beishan Industrial Zone Yantian District, Shenzhen,
518083, China Tel: 400-706-6615 (within
China) +86-755-25273620
Email: tech@bgitechsolutions.com

Hong Kong

BGI Hong Kong

Dai Fu Street, Tai Po Industrial Estate, Tai Po, New
Territories, Hong Kong
Enquiry Email: bgihk.info@genomics.cn Tel:
+852-3610-3510 Fax: +852-2636-5406
Sample receiving Email: bgihk.sample@genomics.cn
Contact: Derry Yuen Ka Leung Tel: +852-3610-3511

Japan

BGI Japan

Kobe KIMEC Center BLDG. 8F
1-5-2 Minatojima-minamimachi, chuo-ku, Kobe City,
Hyogo-pref. 650-0047 JAPAN Tel: 078-599-6108 Fax:
078-599-6109 Email: bgijapan@genomics.org.cn

Asia Pacific / Oceania

BGI Asia Pacific

Main Building 2nd Floor, Beishan Industrial Zone Yantian
District, Shenzhen, 518083, China
Tel: +(86)25273120 Fax: +(86)25011756
Email: p_info_asiapacific@genomics.cn