Exome Sequencing FAQ

2011-05-13 (Version 1)

Contents

1. What is exome sequencing technology?..2

2. What is included in an entire project of exome sequencing? How long does an exome sequencing project take?..2

3. What is BGI's experience in the field of exome sequencing?..2

4. Does BGI have any exome-related publications? 3

5. What research applications are appropriate for exome sequencing?................................3

6. What is the time limit for DNA extraction of collected samples to ensure high-quality DNA?..3

7. What are the sample requirements for exome sequencing? What is the minimum amount required?..3

8. What are the possible reasons for samples failing BGI's QC even after several deliveries? What precautions should be taken when shipping the samples?.....................3

9. What criteria are used in DNA sample quality control at BGI?.......................................4

10. What is the standard for determining the qualified DNA sample? 4

11. What are the procedures used for exome sequencing?..4

12. What are the normal requirements for the effective mean depth for each sample? 6

13. What are the procedures and methods used for exome sequencing analysis? 7

14. What kind of data and reports are generated upon completion of the project?.................9

15. What are the sizes of datasets associated with different sequencing depths? 9

16. How should samples be selected for genetic research of complex diseases?....................9

1. What is exome sequencing technology?

   Exome sequencing refers to a genome analysis method utilizing sequence capture technology to selectively capture exomes within the coding regions of the human genome, followed by target fragment enrichment, and high-throughput sequencing. In contrast to the traditional method of PCR, exome sequencing has dramatically increased the research efficiency of exome regions within the human genome and significantly reduced the cost of research. This technology is not only used in the research of complex diseases such as diabetes, obesity, cancer, and monogenic diseases, but is also considered an effective method for rapid screening of Mendelian heritable diseases.

2. What is included in an entire project of exome sequencing? How long does an exome sequencing project take?

   The following procedures are performed assuming appropriate samples are provided by our customers:

   - sample QC to ensure certain requirements are achieved,
   - exome library construction,
   - sequencing using the Illumina HisSeq® 2000,
   - bioinformatics analysis.

   The turn-around times for these projects are as follows:

   - sample number over 100: library construction and sequencing within 40 work days, bioinformatics analysis in 25 work days,
   - sample number less than 100: entire project completion within 4 months, but may be longer if there are an extremely large number of samples.

3. What is BGI's experience in the field of exome sequencing?

   Since the second half of 2010, most of the exome projects at BGI used the Agilent SureSelect platform. Since 2011 the service based on NimbleGen SeqCap EZ has been available.

   With the maturation of the technology and steady increase in data quality, more than 6,000 exome samples have been sequenced, and more than 600 exome libraries can be constructed weekly. BGI participates in a series of international and domestic exome projects on major diseases, serves as one of the vital members of the international cancer genome project, and collaborates with numerous well-known institutes, hospitals, universities, and pharmaceutical companies world-wide.

4. Does BGI have any exome-related publications?

   See the following references for BGI related publications:

   - Sequencing of fifty human exomes reveals adaptation to high altitude. (Science) DOI: 10.1126/science.1190371,

   - Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. (Nature Genetics) DOI:10.1038/ng.680,

   - TGM6 identified as a novel causative gene of spinocerebellar ataxias using exome sequencing. (Brain) DOI:10.1093/brain/awq323.

5. What research applications are appropriate for exome sequencing?

   Because the coding region sequence accounts for most of the design of the probes, exome sequencing is primarily applicable to investigations of complex diseases caused by potential variations within coding regions. It is particularly appropriate for research of diseases associated with rare mutations.

6. What is the time limit for DNA extraction of collected samples to ensure high-quality DNA?

   Pathological or normal samples collected during clinic operations should be preserved in liquid nitrogen immediately after incision if they cannot be treated immediately upon collection from operating room. Typically, extraction work within three months provides qualified DNA samples, if proper methods are applied.

7. What are the sample requirements for exome sequencing? What is the minimum amount required?

   - sample purity should be OD260/280 =1.8~2.0,

   - sample concentration should be as high as possible and no less than 50ng/µl,

   - sample quantity should be no less than 6ug for each sample.

   The required DNA quantity for a single library construction is 3ug. Also, BGI needs a reserve backup DNA amount should there be a library construction failure. The sample surplus will be properly stored and returned to the collaborators at the end of project upon request.

8. What are the possible reasons for samples failing BGI's QC even after several deliveries? What precautions should be taken when shipping the samples?

   Please extract DNA from fresh blood or tissue samples or samples preserved in liquid nitrogen as much as possible. We recommend that DNA samples be precipitated by 75% alcohol, so that they can be transported at room temperature. Tissue samples must be frozen with liquid nitrogen and shipped with dry ice within 72 hours. If ice bags are used for shipment, please ensure the samples can be received within 24 hours.

Note the following precautions:

- Record the sample name, sample collection date, and sample type on the surface of the collection tube or aluminum packaging foil using an oil-based marker, and avoid contact with ethanol-based organic solvents.

- When completing the *Sample Information Sheet,*, the collaborator should record in detail, the treatment condition, storage requirements, storage time, etc. Additionally, fill in the detailed experimental design (including the patterns of test and control groups, if the samples need pooling for library construction, and the pooling method) or authorize BGI to record it. This information helps our technical staff determine the most appropriate strategy to carry out the experiment.

- Assess the DNA sample quality by verifying the integrity of electrophoresis band. UV and the Agilent 2100 Bioanalyzer can be used to test sample concentration and purity. Also provide this information to BGI so that our quality check is evidence-based and the testing results can fully reflect the actual sample condition. On the wall of the EP tube, clearly indicate the sample name, purity, volume, and the extraction date. This information should match the information on the information sheet.

9. What criteria are used in DNA sample quality control at BGI?

   BGI uses the Qubit® fluorometer and agarose gel electrophoresis as quality checks. The testing results that are provided in the QC report include concentration, volume, total mass, and test conclusion of sample purity of RNA or protein contamination.

10. What is the standard for determining the qualified DNA sample?

    For DNA samples, the critical criteria are as follows:

    - the main belt of DNA is clear,

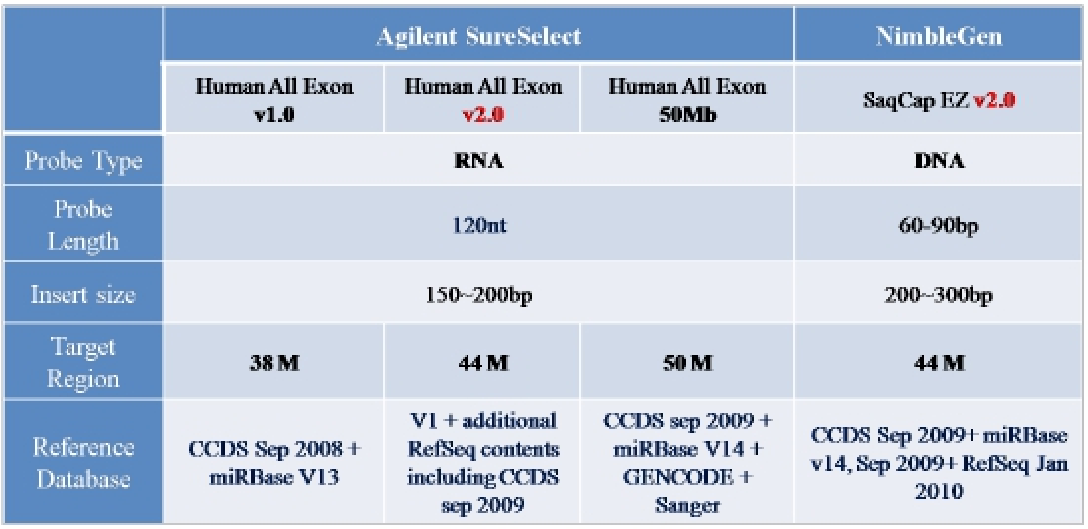
    - the total mass is greater than 3ug.

11. What are the procedures used for exome sequencing?

    Presently, BGI uses two platforms for exome capture, Agilent SureSelect and NimbleGen SeqCap EZ. For the latter platform BGI only provides 44M version 2.0 capture service.

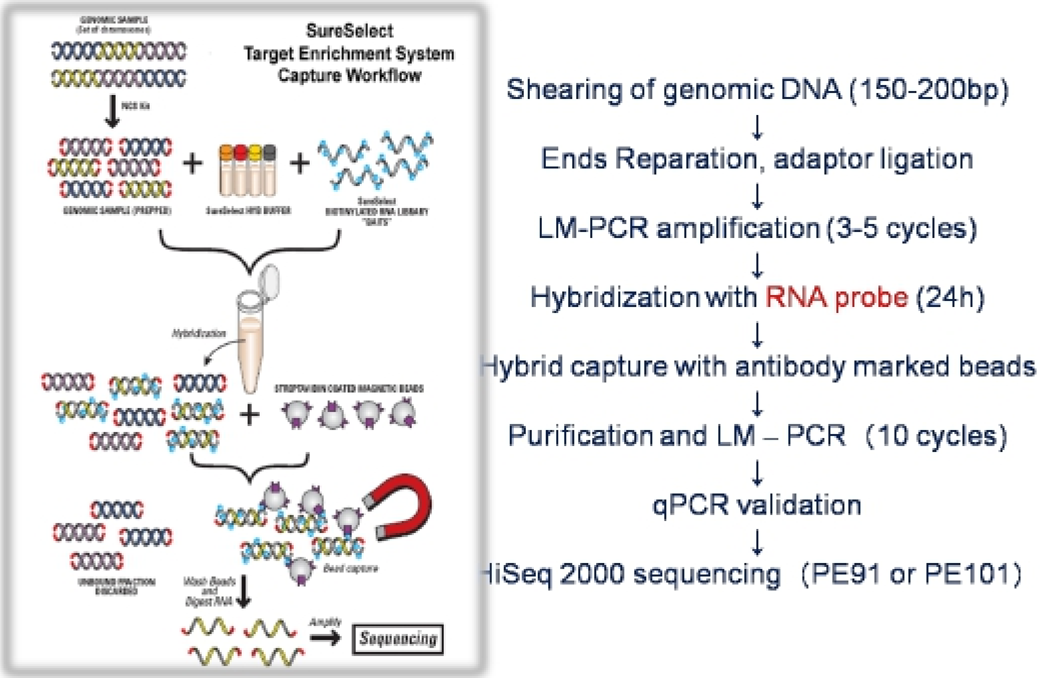
    The average DNA fragment size of SureSelect platform is 150bp, whereas it is 200-300bp for NimbleGen SeqCap EZ platform. Both platforms are relatively mature and as a strategic partner of NimbleGen and Agilent, BGI strongly influences the chip design.

Table 1: Information about exome capture platforms in BGI



	Agilent SureSelect			NimbleGen
	Human All Exon v1.0	Human All Exon v2.0	Human All Exon 50Mb	SaqCap EZ v2.0
Probe Type	RNA			DNA
Probe Length	120nt			60-90bp
Insert size	150~200bp			200~300bp
Target Region	38 M	44 M	50 M	44 M
Reference Database	CCDS Sep 2008 + miRBase V13	V1 + additional RefSeq contents including CCDS sep 2009	CCDS sep 2009 + miRBase V14 + GENCODE + Sanger	CCDS Sep 2009+ miRBase v14, Sep 2009+ RefSeq Jan 2010

All of Agilent capture methods use the same procedure for exome library construction and are as shown in figure 1. Note that there are three different types of kits.

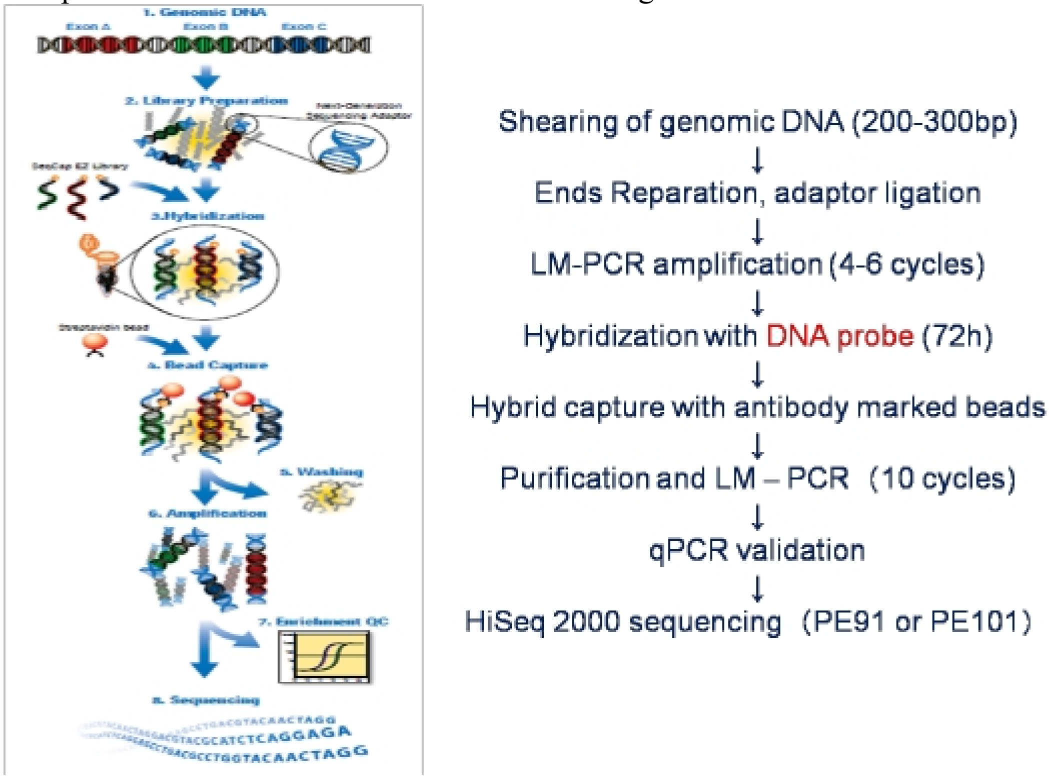


GENOMIC SAMPLE (Set of Chromosomes)
SureSelect Target Enrichment System Capture Workflow
NGS Kit
GENOMIC SAMPLE (PREPPED)
SureSelect HYB BUFFER
SureSelect BIOTINYLATED RNA LIBRARY "BAITS"
Hybridization
STREPTAVIDIN COATED MAGNETIC BEADS
Bead capture
UNBOUND FRACTION DISCARDED
Wash Beads and Digest RNA
Amplify
Sequencing

Shearing of genomic DNA (150-200bp)
↓
Ends Reparation, adaptor ligation
↓
LM-PCR amplification (3-5 cycles)
↓
Hybridization with RNA probe (24h)
↓
Hybrid capture with antibody marked beads
↓
Purification and LM – PCR (10 cycles)
↓
qPCR validation
↓
HiSeq 2000 sequencing (PE91 or PE101)

Figure 1: Agilent SureSelect exome capture experimental pipeline.

The procedure for NimbleGen EZ are shown in figure 2.



Shearing of genomic DNA (200-300bp)
↓
Ends Reparation, adaptor ligation
↓
LM-PCR amplification (4-6 cycles)
↓
Hybridization with DNA probe (72h)
↓
Hybrid capture with antibody marked beads
↓
Purification and LM – PCR （10 cycles)
↓
qPCR validation
↓
HiSeq 2000 sequencing （PE91 or PE101）

Figure 2: NimbleGen SeqCap EZ exome capture experimental pipeline.

12. What are the normal requirements for the effective mean depth for each sample?

The definition of effective mean depth (xX) is that the total number of bases mapped to the target region is equivalent to x times of the whole target region length. The size of target region is determined by the method used for library construction.

90% of the target region will be covered when the sequencing depth has reached 6X. Based on BGI’s project experience, the coverage reaches a plateau when the sequencing depth is above 20X as shown in figure 3.

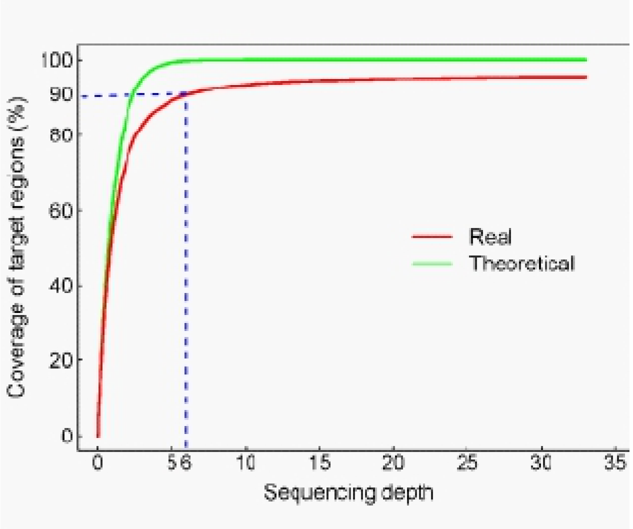


Figure 3: Correlation between sequencing depth and target region coverage.

We recommend increasing the average sequencing depth to 30X or above. Although this increase contributes minimally to the improvement of target region coverage, it improves the likelihood of identifying potential rare variants associated with disease and improves the accuracy of detecting SNPs. The collaborator can select an effective mean depth based on budget and the purpose of the research.

13. What are the procedures and methods used for exome sequencing analysis?

After sequencing, raw data are generated in fastq format. The pipeline of bioinformatics analysis is shown in figure 4.

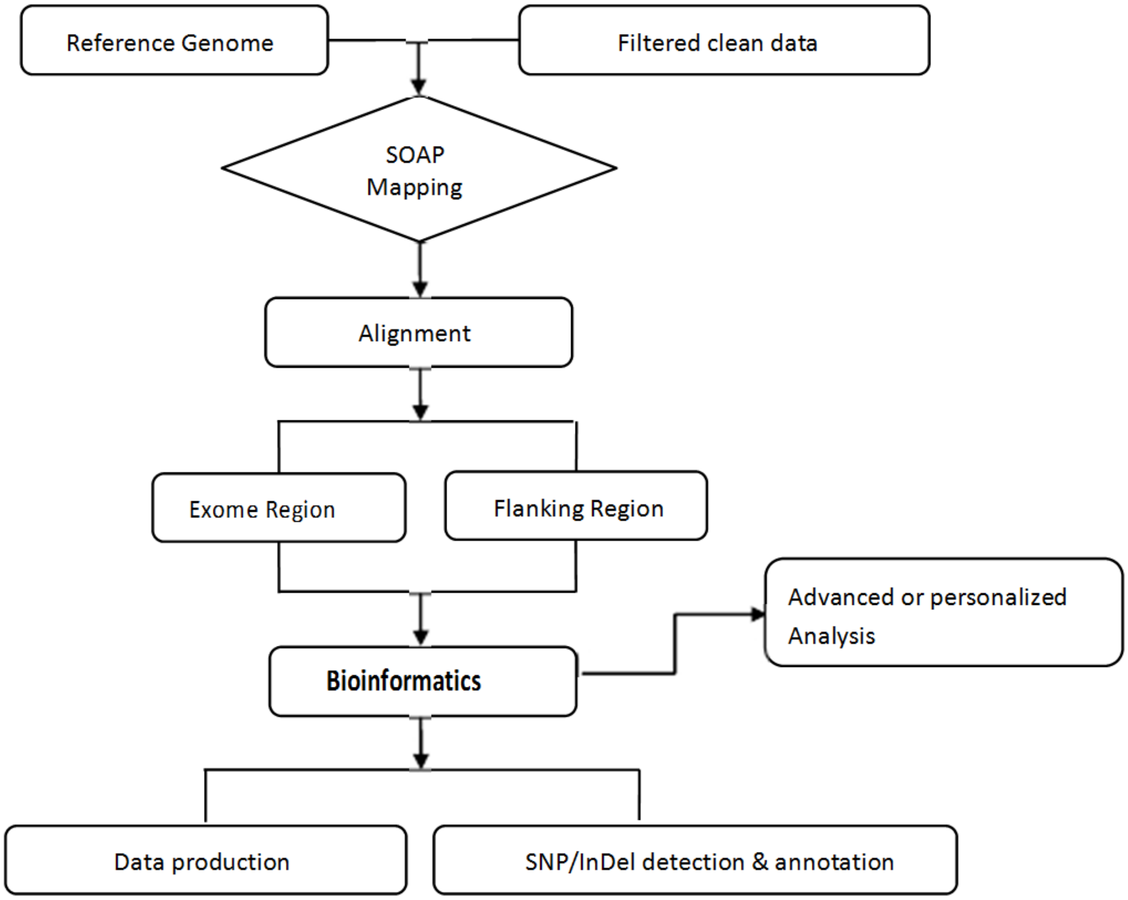


Figure 4: Pipeline for bioinformatics analysis.

Human reference genome build 36 and 37 (i.e Hg18 and Hg19) are both available at BGI. The latter version is set as the default if the reference genome is not specified by the customer.

The adaptor is removed and high-quality reads (filtered clean data) are mapped to the reference sequence, which is implemented using SOAPaligner (software fully developed by BGI). Only the data mapped to reference genome are used for subsequent analysis. For more information on SOAP and the analysis parameters, refer to http://soap.genomics.org.cn/soapaligner.html.

The statistics generated include the amount of data, effective mean depth, coverage, the distribution of per-base sequencing depth, and the cumulative depth distribution in target regions.

SNP refers to the variation of a single nucleotide in contrast to the reference genome. All of the consensus genotypes are identified by SOAPsnp, (see http://soap.genomics.org.cn/soapsnp.html) and then recorded in a CNS file. Subsequently, all loci different from reference genotype are identified from the CNS file. Finally, high confidence SNPs can be identified by applying certain filter conditions. The database used is dbSNP.

14. What kind of data and reports are generated upon completion of the project?

    The project report summarizes the project result.

    The original sequence file generated from sequencing is saved in FASTQ format, and the VCF format is also available upon request.

    Variation identification and annotation are included in the file of bioinformatics analysis results. More details and parameter configurations are described in *Exome Capture Sequencing Project Report*.

    For small datasets, data are uploaded to an FTP server for the collaborators to download. Portable hard drives are also available for larger datasets. A standard Unix or GNU/Linux system is required for customers to perform subsequent analysis. Generally, the sequencing data is saved by BGI for one month following project completion. We can store data for a longer time for an additional fee.

15. What are the sizes of datasets associated with different sequencing depths?

    After image processing, base calling, and assembly, the original sequencing dataset sizes are: effective mean depth of 30X is ~2 to 3G, effective mean depth of 50X is ~4 to 6 G, and effective mean depth of 100X is ~5 to 8G.

16. How should samples be selected for genetic research of complex diseases?

    The optimal criterion is to select disease cases with higher heritability as the research target. The samples of intermediate phenotype for research of complex diseases should have the following features:

    association with the research disease,

    high heritability,

    high dominance,

    early heredity expression among the offspring of the patient and with the same phenotype,

    causality present between pathogenesis and symptom, such as quantitative trait, early onset, and extreme phenotypes, to improve the association between disease and candidate loci determined after filtration.