Whole Genome Re-Sequencing FAQ

2011-04-28 (Version 1)

Contents

1. What is whole genome re-sequencing? 2
2. What are the strengths of whole genome re-sequencing? 2
3. What are the sample requirements for human genome re-sequencing? 2
4. What human reference genome is used? Which version is currently used? 3
5. What workflow is used with this technology? 3
6. What are read, pair end, and consensus? 3
7. What is difference between SNV and SNP? What method is used to validate the accuracy of a called SNP? 3
8. What are the effects of repeat sequence regions in the human genome on re-sequencing? 3
9. What are the differences in whole genome re-sequencing between somatic chromosomes and sex chromosomes? 4
10. There are several mature technologies in studying complex diseases at BGI. In what situations does BGI recommend using WGS? 4
11. There are several mature technologies in studying complex diseases at BGI. In what situations does BGI recommend using WGS? 4

1. What is whole genome re-sequencing?

   Whole-genome re-sequencing is the sequencing of an individual in comparison to a known reference genome. By comparing sequences among individuals or populations, the genomic variations can be analyzed and provide a comprehensive analysis of gene sequence and structure variations. BGI applies the combinational approach of short-reads, paired-end, and short insert-sizes in re-sequencing. Whole genome re-sequencing is a powerful tool for scientific research and industry application as it can be used to detect disease-related mutations at the whole genome level.

2. What are the strengths of whole genome re-sequencing?

   Strengths of whole genome re-sequencing include:

   - The technology workflow is mature.
   - Large structural and copy number variations in the genome can be tested.
   - Novel and rare hereditary variations can be found.
   - More comprehensive and reliable information on genetic variation is derived.
   - Small data deviation, high uniformity, and an accurate genomic information of the samples are inherent in the approach.

3. What are the sample requirements for human genome re-sequencing?

   Sample requirements include the following:

   - DNA samples must be free from degradation, protein, and RNA contamination.
   - Sample DNA quantity should be:
     - for single library construction: ≥6ug: DNA
     - for single sample sequencing: ≥10ug: DNA
     - for multiple library construction (N times): 3× (N+1） ug
   - Concentration: ≥50 ng/ul
   - Purity: OD260/280 Ratio: 1.8-2.0

4. What human reference genome is used? Which version is currently used?

   Generally hg18 or hg19 is used as the reference for human genome re-sequencing.

   - For HG18, we are using NCBI build 36.3 with a genome size of 3,107,677,273 and effective size of 2,881,421,700 (non-N bases).
   - For HG19, we are using GRC 37(March 3, 2009) with a genome size of 3,137,161,264 and an effective size of 2,897,310,462 (non-N bases).

5. What workflow is used with this technology?

   The workflow is illustrated in Figure 1.

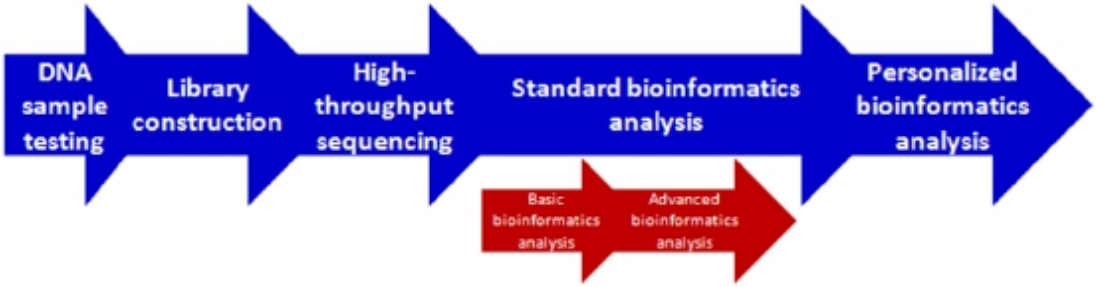

Fig. 1 Whole genome re-sequencing workflow

6. What are read, pair end, and consensus?

   - Reads are the smallest units of base sequencing.
   - Pair ends, or paired-end tags, also known as PET, refer to the short sequences at the 5' and 3' ends of the DNA fragment of interest, which can be a piece of genomic DNA or cDNA.
   - Consensus is a single sequence that represents, at each subsequent position, the variation found within corresponding columns of a multiple sequence alignment.

7. What is difference between SNV and SNP?

   SNP is single nucleotide polymorphism and SNV is single nucleotide variation. If the changes of a single nucleotide in tumor cells have not been validated in a population, it is called a SNV. Otherwise, it is called a SNP.

8. What method is used to validate the accuracy of a called SNP?

   In the Yan Huang Project, Sanger sequencing and array is used to validate the accuracy of a SNP. Sanger sequencing is acknowledged as a preferred standard in sequencing.

9. What are the effects of repeat sequence regions in the human genome on re-sequencing?

   In theory, a sequence of a gene comes from one region of the genome. If the sequence is a repeat sequence in the reference genome, 5 different regions can be aligned. When we choose a region randomly, there is an 80% chance (4 out of 5) of selecting incorrectly. If the tested sequence has a repeated sequence, then the effects of choosing the wrong regions will be low. That means that a real repeat sequence does not lead to a high error rate in re-sequencing.

10. What are the differences in whole genome re-sequencing between somatic chromosomes and sex chromosomes?

    If the sample is female (XX), there is no difference in sequencing depth between somatic and sex chromosomes. If the sample is male, because the sex chromosome is XY, the sequencing depth is half of the somatic chromosome. However, as the homology between X and Y chromosomes is high, there can be large differences between X and Y chromosome.

11. There are several mature technologies in studying complex diseases at BGI. In what situations does BGI recommend using WGS?

    We recommend WGS in the following situations:

    - If the researcher has sufficient funds and little background knowledge for specific disease, we recommend using WGS plus genotyping.

    - We also recommend WGS as a tool to identify the virus insert fragment in a host genome for infectious diseases.