

Whole Exome Sequencing and Analysis

Q1. What is whole exome sequencing?

A1. This is an efficient strategy to selectively sequence the coding regions (exons) of the human genome to discover rare or common variants associated with a disorder or phenotype [1]. By focusing sequence production on exons, which represents ~1% of the human genome, many more individuals can be examined at significantly reduced cost and time compared to sequencing their entire genomes. A number of methods are in use, but all generally enrich the DNA sample before sequencing by targeting the exons [2]. Except for a few bases beyond the exon ends, variants in introns, promoters, 3' - and 5' - untranslated regions, or inter-genetic regions generally are not detected. These methods also target only those exons that are well-established and with annotation in a database, such as, consensus CDS, or GENCODE; this subset of exons is roughly 80% of all possible coding sequences found in the UCSC Genome Browser.

Q2. How is whole exome sequencing performed at NISC?

A2. NISC currently employs a solution-based probe hybridization protocol to capture (enrich for) exonic sequences from the DNA sample. The nucleic acid probes are components of a standard commercially available kit optimally designed and prepared for this purpose by the vendor. In brief, the DNA is sheared mechanically, targeted fragments captured by probe hybridization, and then amplified before sequencing on an Illumina HiSeq2000 instrument. NISC continually evaluates improvements in these technologies, and implements those that represent reduction in cost and time or increase in exon coverage.

Q3. What material should I send for whole exome sequencing?

A3. We need 1.5 µg of highly-purified genomic DNA in a volume of 50 µl or less for whole exome sequencing. To ensure that each sample is uniformly pure and free of infectious agents, we require that all DNAs be phenol:chloroform extracted before submission. A simple protocol is available from NISC. The investigator must submit an image of a QC agarose gel as evidence the DNA is of good integrity. We highly recommend Qubit for quantitation of the DNA sample, since it uses a double-strand DNA-specific method. UV absorption methods, e.g., using a NanoDrop spectrophotometer, can drastically over estimate the concentration of DNA due to RNA contamination.

Q4. How long do the reads need to be for whole exome analysis?

A4. Typically, fragment read lengths are 100 bases. Paired-end reads are generated, so overall 200 bases of sequence are read from each fragment.

Frequently Asked Questions – Illumina HiSeq2000 Sequencing

Q5. How many lanes of a flow cell are used for a whole exome analysis?

A5. Each lane of a sequencing flow cell on a HiSeq2000 yields ~280 million paired-end reads. For our current method, we typically need 1/3rd lane of paired-end 100 base reads. This amount of reads is usually sufficient to meet our minimum coverage requirement to call genotypes with MPG score ≥ 10 for at least 85% of the targeted bases.

Q6. How are variants called?

A6. Sequence reads produced for a sample are aligned to the human reference sequence and the results stored in BAM format. A custom analysis program MPG (Most Probable Genotype) processes this information using a probabilistic Bayesian algorithm, calling genotypes at all reference positions at which there are high quality bases from the aligned sequence reads [2]. The likelihood of each possible genotype given the observed sequence data is calculated and given an MPG score, where $\text{MPG} \geq 10$ is considered accurate. These genotype calls have been compared against Illumina Human 1M-Quad genotype chips, and genotypes with a MPG score of 10 or greater show >99.89% concordance with SNP chip data [2].

Q7. What data are returned by NISC?

A7. All variants, genotypes, and annotations are delivered to the investigator in tab-delimited “VS” format suitable for input to a java-based genotype viewer called VarSifter [3]; and available from NISC. The file can also be input to Excel. The VS file contains all discovered variants with genotypes of all samples sequenced, as well as gene locations (5' UTR, 3' UTR, coding--synonymous, nonsynonymous, or stop, splice site, or intron).

References:

1. Biesecker L (2010) “Exome sequencing makes medical genomics a reality.” *Nature Gen.* 42, 13-14.
2. Teer JK, *et al.* (2010) “Systematic comparison of three genomic enrichment methods for massively parallel DNA sequencing.” *Genome Res.* 20, 1420-1431.
3. Teer JK, manuscript submitted.