Frequently Asked Questions

Whole Genome Sequencing

Q1. What is whole genome sequencing ?

A1. Whole genome sequencing (WGS) is simply the sequencing of the entire genome of an organism at one time [1]. The purpose may be to determine the genome sequence of a previously unsequenced species to extend evolutionary biology studies or to look for difference between similar samples, for example, to determine sequence variations that may cause phenotype differences between cancerous and normal tissue cells. Almost any type of cell can be the source of DNA for WGS, including human, mouse, jellyfish and bacteria. *Note, DNA samples derived from living humans must be consented for WGS before acceptance at NISC for sequencing.*

Most commonly WGS data are aligned to a suitable reference sequence for analysis. Alternatively, sequence reads can be assembled de novo, although incompletely, since without a suitable scaffold the short length of Illumina reads produces numerous contigs that are orders of magnitude smaller than the chromosomes from which they are derived. Incorporating mate-pair and very long length reads, for example, sequence data from MiSeq, 454 or PacBio, can greatly improve assembly contiguity.

Q2. How is WGS performed at NISC ?

A2. A "library" of DNA fragments is prepared for sequencing from the purified DNA sample. There are several strategies for WGS depending on the goals of the project and the size of the genome.

- Microbial genome sequence for alignment to a reference – small insert library
- Microbial genome sequence for de novo assembly – small insert library plus mate-pair library (end reads from a large insert library)
- Mid-sized genome (50-500 Mb) sequence for alignment to reference – PCR-free small insert library
- Mid-sized genome (50-500 Mb) sequence for de novo assembly – PCR-free small insert library plus mate-pair library (end reads from a large insert library)
- Mammalian-sized genome for alignment to reference – PCR-free small insert library

Q3. What material should I send for WGS ?

A3. We need 1.5 μg DNA for a small insert library, 2.5 μg for a PCR-free small insert library, and 1.5-4.5 μg for a mate-pair library (depending on the distance separating the ends to be sequenced). Samples should be submitted in 1.5-1.7 ml microfuge tubes (example: VWR cat. no.89000-028) or 2 ml screw cap tubes (example: Sarstedt cat. no. 72.694.007). Please DO NOT send samples in 0.5 or 0.2 ml tubes. 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. Ref: www.nisc.nih.gov/docs/gDNA_submission_exome_cc.pdf

Q4. How should the DNA be qualified ?

A4. The investigator must submit an image of an analytical agarose gel as evidence the DNA is of good integrity, i.e., not just a 'fuzzy blob' of low molecular weight. 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 overestimate the concentration of DNA due to RNA and small molecule contamination.

Q5. How long do the reads need to be for WGS analysis ?

A5. Typically, NISC generates read lengths of 125 bases on a HiSeq for mid-size and mammalian genomes. Paired-end reads generate a total of 250 bases of sequence from each fragment in the library. For microbial genomes the sequencing is performed on a MiSeq so read lengths are 300 bases, thus paired-end reads generate 600 bases of sequence from each fragment.

Q6. How many reads are required for WGS ?

A6. We recommend enough sequence reads to yield 30-100× coverage of the genome. Since reads are paired end, we generate two reads for each fragment sequenced [2].

Q7. What data are returned by NISC ?

A7. Sequence reads produced for a sample are aligned to the appropriate reference sequence, if available, and the results stored in BAM format. This file also contains basecalls and quality scores. Data files can be quite large, for example 30× coverage of a human genome is about 90 GB of compressed data. The investigator is expected to provide data analyses; this is not offered by NISC.

References :

1. Illumina (2013) "An Introduction to Next-Generation Sequencing Technology." www.illumina.com/documents/products/Illumina_Sequencing_Introduction.pdf
2. Sims, D., *et al.* (2014) "Sequencing depth and coverage: key considerations in genomic analyses." *Nature Rev. Genetics* **15**: 121-132.