# 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 very long length reads, for example, sequence data from PacBio or Oxford Nanopore can greatly improve assembly contiguity and can generate complete finished genomes for bacteria.

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.

- For alignment to reference – PCR-free small insert library sequenced on Illumina platform
- For de novo assembly – large insert library sequenced on Pacific Biosciences Sequel IIe and/or Oxford Nanopore PromethION. For an assembly approaching T2T quality a combination of technologies may be employed. Data from a small insert library sequenced on Illumina platform can be helpful for error correction.
- Haplotype level assembly will require additional data, such as parental sequence (trios) or Hi-C data.

Q3. What material should I send for WGS ?

A3. We need 1.5 µg DNA for a PCR-free small insert library. 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 sufficiently pure, we strongly suggest that all DNAs be cleaned up by phenol:chloroform extraction before submission. A simple protocol is available from NISC. Ref: www.nisc.nih.gov/docs/gDNA_submission_exome_cc.pdf

For a large insert library to be sequenced on the Pacific Biosciences Sequel IIe, the amount of DNA needed will depend on the genome size, DNA fragment length, and depth of coverage requested. Since those libraries are not amplified, the amount of DNA needed is much higher. For a mammalian sized genome, we request 50 µg of high molecular weight DNA (40 kb or larger). For microbial isolates, we request 1-3 µg of high molecular weight DNA (20 kb or larger) depending on the number of samples to be multiplexed for sequencing.

Oxford Nanopore offers two options. The ultra-long sequence option produces very long reads, up to a megabase, but the DNA must be extremely high weight. For a mammalian sized genome, we request 50 µg of very high molecular weight DNA (>50 kb). For a standard library we request 5 µg of high molecular weight DNA (>20 kb).

Q4. How should the DNA be qualified ?

A4. For short read sequencing the investigator must submit an image of an analytical agarose gel or trace as evidence the DNA is of good integrity and the appropriate molecular weight for the sequencing approach. 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.

For long read sequencing a pulse-field gel is needed. If needed NISC can QC the DNA for you.

Q5. How long will the reads be from WGS ?

A5. Typically, NISC generates read lengths of 150 bases on an Illumina NovaSeq X Plus. Paired-end reads generate a total of 300 bases of sequence from each fragment in the library.

Pacific Biosciences Sequel IIe generates HiFi reads averaging ~20 kb.

Oxford Nanopore offers two options. An ultra-long sequence option that produces very long reads, up to a megabase, but the DNA must be extremely high weight. These reads are less accurate but are quite useful for scaffolding contigs. The second option produces average read lengths of ~20-30kb but produces much more data with higher accuracy.

Q6. How many reads are required for WGS ?

A6. We typically recommend enough sequence reads to yield 30× coverage of the genome [2], but more may be needed for de novo assembly.

Q7. What data are returned by NISC ?

A7. For human genomes, we return aligned BAM files and variant calls. Raw fastq sequence files are returned for other mammalian as well as mid-sized genomes. These files contain basecalls and quality scores. For microbial genomes, we also generate a preliminary assembly. 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.