Sequencing entire genomes

Despite substantial improvements, sequencing reads are still substantially smaller than the typical sizes of genomes. Genome sizes, range from a couple of million bp for bacterial genomes up to several billion base pairs for some eukaryotes (Fig. 1).

Figure 1: Distribution of genome sizes across the tree of life (Figure taken from here).

The challenge to sequence genomes of such sizes is typically addressed with a whole genome shotgun sequencing strategy. This approach was introduced in 1995 when a research team set out to sequence the genome of *Haemophilus influenzae RD* by random sequencing (Fleischmann, et al. 1995), and replaced the traditional hierarchical shotgun sequencing approach, and the prier. The same lab used this technology to sequence five years later the genome of a eukaryote, *Drosophila melanogaster*, which is about 100 times larger than the bacterial genome (Adams, et al. 2000). Nowadays, virtually all genome sequencing efforts are whole genome shotgun approaches.

In a nutshell, the genomic DNA is initially shredded at random positions into overlapping fragments, which are later often referred to as inserts. Short DNA segments with known sequence, so called *adapters*, are then added to these *inserts* to provide the starting points necessary for the sequencing. Among others, this can be primer binding sites for both fragment amplification via PCR, and for the sequencing itself (e.g. Illumina adapter or the bell-shaped adapters used by PacBio), or adapters providing the motor that is necessary to pull the DNA molecule through a Nanopore (cf. Figure ##REF:longread##). The collection of resulting fragments is referred to as a *shotgun library*. Depending on whether only one end or both ends of these shotgun fragments are sequenced, they are referred to as single- or paired-end shotgun libraries. Once a shotgun library has been sequenced – typically up to a coverage between 60 and 100 – the genome is reconstructed from this data.

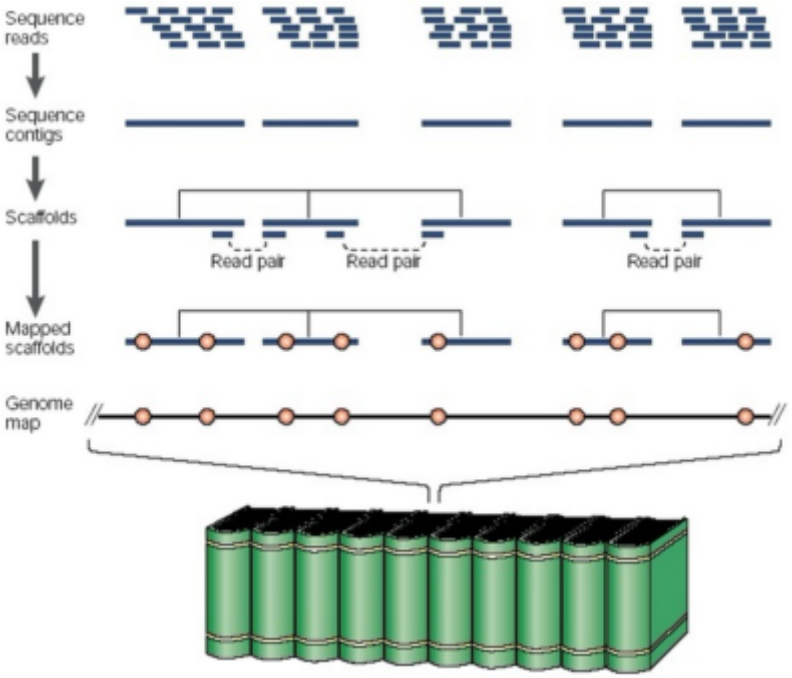

Figure 2: Workflow of a typical whole genome shotgun sequencing approach. Figure 1: Long-range sequence assembly in whole-genome shotgun sequencing. Individual sequence reads generated in a whole-genome shotgun-sequencing project are initially assembled into sequence contigs. Groups of sequence contigs are then organized into scaffolds on the basis of linking information provided by read pairs (in each case, with one sequence read from a pair assembling into one contig and the other read into another contig). In turn, the scaffolds can be aligned relative to the source genome (represented by an encyclopedia

set) by the identification of already mapped, sequence-based landmarks (for example, STSs, genetic markers and genes; depicted as red circles) in the sequence contigs, thereby associating them with a known location on the genome map. © 2001 Nature Publishing Group Green, E. D. Strategies for the systematic sequencing of complex genomes. Nature Reviews Genetics 2, 580 (2001)