What is de novo assembly? – The Sequencing Center

by Richard Casey

Created OnJanuary 16, 2019

byRichard Casey

De novo assembly is a method for constructing genomes from a large number of (short- or long-) DNA fragments, with no *a priori* knowledge of the correct sequence or order of those fragments.

The terminology for *de novo* assembly is sometimes inconsistent so we'll use the definitions below:

- **Reads**. DNA fragments. "short-reads" typically range in size 35 – 1,000 bp (nucleotide base pairs). "long-reads" typically range in size 1,000 – 500,000 bp. For our purposes we'll assume read length is 150 bp, although read length depends on the sequencer model and library prep protocol used for a particular sequencing run. Raw reads generated by sequencers are generally stored in FastQ files.
- **Contigs**. Set of overlapping oriented reads. A single contig is constructed from two or more overlapping and oriented reads. The reads share a subset or all of their nucleotide base pairs. The reads may have to be reversed ("flipped") to yield a matching orientation, although this is rarely necessary.
- **Scaffolds**. Set of joined oriented contigs. A single scaffold is constructed from two or more joined and oriented contigs. The contigs may have to be reversed ("flipped") to yield a matching orientation. The contigs may be overlapping or non-overlapping.
- **Chromosomes**. Set of joined oriented scaffolds. A single chromosome is constructed from two or more joined and oriented scaffolds. The scaffolds may have to be reversed ("flipped") to yield a matching orientation. The scaffolds may be overlapping or non-overlapping.

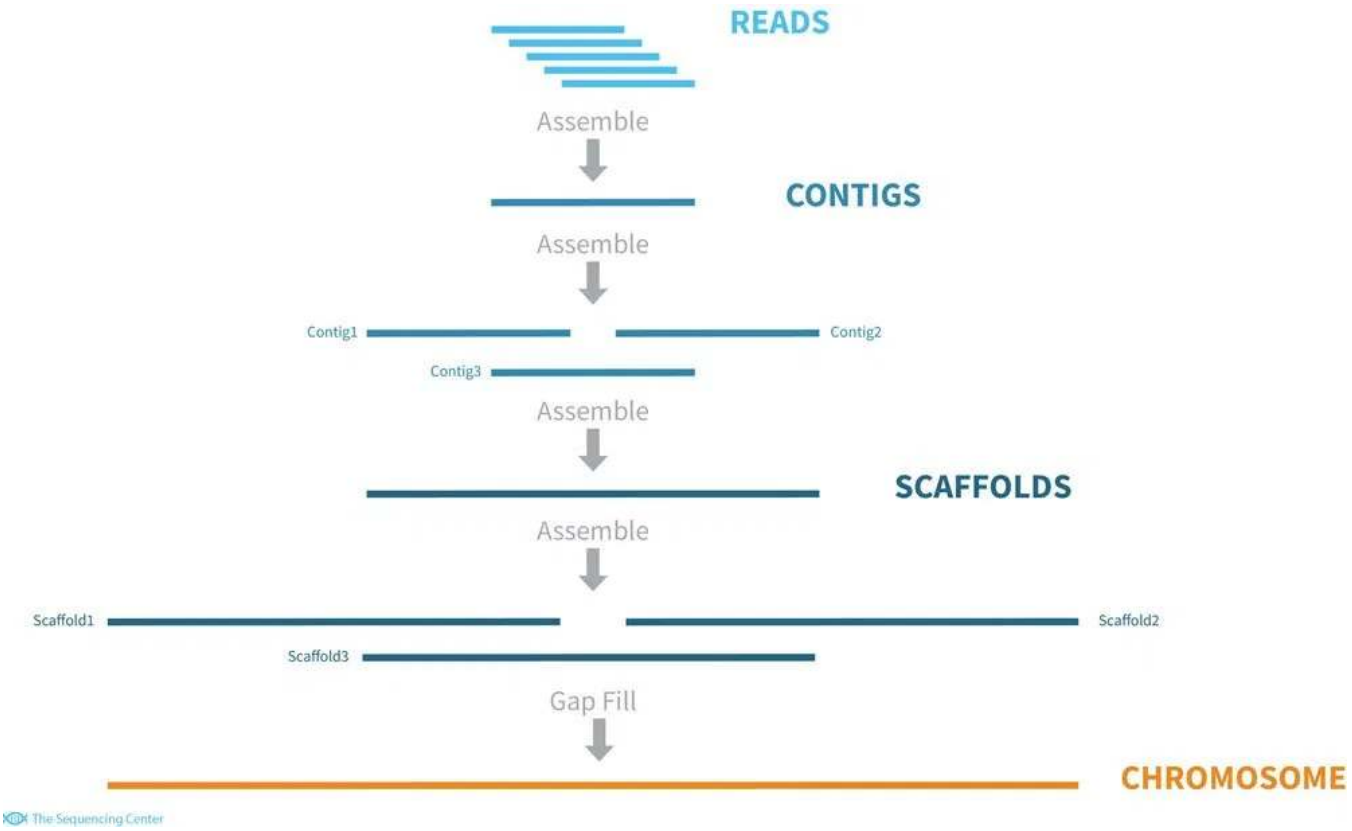

Fig. 1

Fig. 1 shows an overview of the *de novo* assembly process. Partially, or sometimes fully, overlapping reads are assembled into one or more contigs. Sets of overlapping or non-overlapping contigs are joined into one or more scaffolds. Sets of overlapping or non-overlapping scaffolds are joined into a single chromosome.

In the contig assembly step, reads must overlap by a minimum number of base pairs, or k-mers, before they can be mapped together. In the scaffold assembly step, contigs do not necessarily have to overlap in order to be joined together. This can be attributed to paired-end sequencing, which we'll cover in another article. In the chromosome assembly step, scaffolds are joined together in a gap-filling, gap-closing or genome finishing process. This final step is difficult, and sometimes impossible, to complete using only short-read technology. The presence of repetitive sequences especially can inhibit gap-filling using only short-reads, although some progress is being made in this area.

Finishing complete chromosomes often requires the use of multiple sequencing technologies and hybrid assembly protocols. You'll

often see short-read technology combined with long-read technology, optical maps, Bionano maps, etc. to generate fully finished genomes. Employing multiple sequencing technologies on a per sample basis can be costly.

Note. At the present time we can provide *de novo* assembly services to the **scaffold level** for many smaller genomes, such as bacteria, bacteriophage, virus, some yeast, some fungi, etc. In general we cannot *de novo* assemble short-reads to the finished chromosome level, as this would require using multiple sequencing technologies and possibly a small team of researchers assigned to each assembly.

Assembly algorithms. There are many *de novo* assembly algorithms and software applications available for Next Generation sequencing projects. For small genome assembly (i.e. bacterial scale genomes) we often use Spades and Geneious but may use other tools if it's more appropriate.

Assembly quality. In general we use Quast to report on the quality of *de novo* assembled scaffolds.