

Sequence Assembly

The essential problem that sequence assembly is trying to overcome is that the average microbial genome is 2,000,000 bp, while the typical sequence read length is 150-300 bp for Illumina sequences and upto 50,000 bp for PacBio or Nanopore sequences.

With such (relatively) short sequences, how can we assemble a whole, or nearly whole, genome?

The answer is by repetitively sequencing the same thing over and over again! If we start each sequence at a random location, and we have enough sequences, eventually we can join those sequences together to form what we call **contigs**.

There are four types of sequence assembly algorithms:

1. Naive assemblers which just try and find all matching pairs of reads
2. Greedy assemblers which start with one read and keep adding reads until you can not find any more matches, and then start with the next read.
3. Overlap-layout-consensus assemblers which layout the reads looking for overlaps between them. The overlaps are usually refined by a Smith-Watermann search, and then a consensus constructed.
4. de Bruijn graph assemblers

This table describes some of the common sequence assemblers that you will run across.

Sequencing					
Name	Type	Tech	Citation	Doc	Homepage
SPAdes	genomes	Illumina	Narayanan et al, 2014	version 3.12	SPAdes
	single cell	Solexa	et al, 2013	man-454, and 2013	ual
	metagenome	Sanger	et al, 2013	ual	
	ESTs	Ion			
		Tor-			
		rent,			
		PacBio,			
		Ox-			
		ford			
		Nanopore			
Velvet	genomes	Sanger	Zerbino et al, 2008	version 1.12	EBI
		Solexa	Bir-	man-	
		SOLi	De-	ual	
			2008		

Sequencing					
Name	Type	Tech	Citation	Doc	Repo
Canu	genomic	PacBio	Koren et al. 2017	book	GitHub
MaSuRCA	genomic	all	Shen et al. 2016	book	GitHub
SPAdes	genomic	PacBio, Oxford Nanopore	Antony et al. 2015	book	GitHub
Hifiasm	genomic	PacBio	Cheng et al. 2017	book	GitHub

We use the St. Petersburg genome assembler, SPAdes and the version installed on the AWS instances is 3.12.0 for which the manual is [here](#)

For Nanopore reads we typically use the CANU assembler.

SPAdes output files

SPAdes makes a lot of files and directories in the output, and this summarizes what those files are. Of course, more details can be found in the SPAdes manual

- **scaffolds.fasta** contains the scaffolds generated by SPAdes and is the **output file you want to use**.
- the directory **/corrected/** contains reads corrected by BayesHammer in compressed fastq format
- **contigs.fasta** contains the contigs before they are scaffolded into scaffolds. Often this is similar to the scaffolds.fasta depending on how much scaffolding information there is
- **assembly_graph.gfa** contains the assembly graph and scaffolds paths in GFA 1.0 format
- **assembly_graph.fastg** contains the assembly graph in FASTG format

- `contigs.paths` contains paths in the assembly graph corresponding to `contigs.fasta`. This is how the graph is resolved into contigs.
- `scaffolds.paths` contains paths in the assembly graph corresponding to `scaffolds.fasta`.
- `K21`, `K33`, `K55`, etc are directories containing the de Bruijn graph assemblies for different lengths of k
- `before_rr.fasta` are the assembled contigs before repeat resolution has been applied.
- `dataset.info` and `input_dataset.yaml` contain information about the sequence read files that were supplied.
- `params.txt` is a summary of all the spades parameters
- `spades.log` is the log that was printed to the screen while SPAdes was running. This contains lots of information about the assembly process.