Sequencing data Analysis of Samplicons

In this document we outline some standard data analysis that can be run on sequencing data generated from Samplicons (DNA enriched and amplified with Samplix Xdrop™ technology).

Oxford nanopore technology (ONT)

We usually perform fast5 raw ONT data **basecalling** with Guppy (provided by ONT) and obtain fastq files.

We perform **quality** evaluation of the flowcell run and data using Nanopore summary statistics and basic QC tool.

For **mapping** ONT data to a reference genome, we recommend using minimap2 (1) with default parameters. We usually output as sam file (-s option) and compress it into bam using samtools (2) for downstream analysis.

For ***de novo* assembly** we recommend using *Canu* (*3*) with default settings and reduced genome size. That's because *de novo* assemblers, including *Canu*, are not designed to deal with enriched genomic data, characterised by uneven coverage (much higher in the enriched region), we recommend reducing the genome size when running this software, as we have observed that it improves the contig length and the overall *de novo* assembly. Following *Canu* assembly we recommend doing polishing with ***minimap2*** **(*1*)** and *Racon* (*4*), this process may be repeated a number of times.

Additionally, for de novo assembly it may be helpful to look specifically for long reads spanning the region to be assembled. This may provide input for a reference guided assembly approach.

Visualisation of reads mapped to reference or the *de novo* assemblies can be performed in a variety of software, including the freely available *IGV* software (*5*).

Illumina

We recommend **trimming** Illumina adapters using compatible software, e.g. *trimmomatic* (*6*).

We perform quality evaluation of the flowcell run and raw data using FastQC software.

For **mapping** Illumina data to a reference genome, we recommend using *BWA-MEM* (*7*) with default parameters. Other software such as Bowtie2 (*8*) should be also suitable.

Visualisation of mapped reads can be performed in a variety of software, including the freely available *IGV* software (*5*).

Note

Note that Samplicons are generated using Multiple Displacement Amplification in droplets, which can create chimeric reads, consisting of concatemers of inverted repeats of the same sequence (as each sequence is compartmentalized). For this reason, **pre-processing** reads to split chimeric reads could be beneficial, especially prior to *de novo* assembly. This may be done by using software such as *Pacasus* (*9*), which splits palindromic reads, including naturally occurring inverted repeats. This is especially relevant for long read sequencing approaches, as only few short reads from e.g. Illumina sequencing will be chimeric, given that long fragments are broken up prior to sequencing.

Citations

1. H. Li, Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*. **34**, 3094–3100 (2018).

2. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. **25**, 2078–2079 (2009).

3. S. Koren, B. P. Walenz, K. Berlin, J. R. Miller, N. H. Bergman, A. M. Phillippy, Canu: Scalable and accurate long-read assembly via adaptive κ-mer weighting and repeat separation. *Genome Res.* (2017), doi:10.1101/gr.215087.116.

4. R. Vaser, I. Sović, N. Nagarajan, M. Šikić, Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* (2017), doi:10.1101/gr.214270.116.

5. J. T. Robinson, H. Thorvaldsdóttir, W. Winckler, M. Guttman, E. S. Lander, G. Getz, J. P. Mesirov, Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).

6. A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. **30**, 2114–2120 (2014).

7. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. **25**, 1754–1760 (2009).

8. B. Langmead, S. Salzberg, Bowtie2. *Nat. Methods* (2013), doi:10.1038/nmeth.1923.Fast.

9. S. Warris, E. Schijlen, H. van de Geest, R. Vegesna, T. Hesselink, B. te Lintel Hekkert, G. Sanchez Perez, P. Medvedev, K. D. Makova, D. de Ridder, Correcting palindromes in long reads after whole-genome amplification. *BMC Genomics*. **19**, 798 (2018).