Guideline on how to get started

Considerations when designing a whole genome sequencing (WGS) service: From Sample to Result

Nucleic Acid Extraction

- Bacteria will require DNA extraction from isolates. Development of a robust protocol for nucleic acid extraction is critical but already available from many sources, e.g. ENGAGE (see Appendix C in this report). A key component to this is the extraction method. Many manual kits (e.g. Promega Wizard genomic DNA purification and Qiagen/Stratec genomic DNA purification spin columns) are suitable but it is critical to check that the resulting DNA is of sufficient quantity (Illumina recommendation 8-100 ng/µl, Illumina Nextera recommendation 1ng/µl then diluted to 0.2ng/µl). If more than a few tens of samples are expected to be processed per week a high throughput DNA purification system such as Qiasymphony, EZ1, SP/AS, Qiacube HT (Qiagen company) is recommended.
- Viruses present a greater problem for extraction and either an amplicon strategy (Quick et al., 2016) or bait-based enrichment protocol (Depledge et al., 2011) is required.
- In both of these cases, work to assess yields from these protocols is essential so that a standard operating procedure (SOP) can be produced which, if followed, results in a high probability of the amount and quality of DNA being sufficient for subsequent library preparation.

Quantification

- Although it is possible that following a SOP generated from the previous step results in a consistent amount of DNA that does not require quantification, it is recommended that prior to library preparation, quantification is performed.
- Recommended instruments for quantification include the GloMax (high throughput) from Promega and Qubit (single tube) from ThermoFisher. The NanoDrop (ThermoFisher) is not recommended, due to lack of sufficient accuracy and consistency of the readings for the purposes of library preparation.

Library Preparation

- There are two main alternatives for library preparation:
  - Nextera – a kit from Illumina that makes the number of hands on steps minimal but has the disadvantages of giving slightly less uniform coverage compared to physical sheering methods (see below) and having a higher per sample cost. In addition, it is susceptible to being less efficient for genomes with a %GC content significantly different from 50%. Furthermore, Nextera is recommended for bacterial genomes but is probably not suitable for smaller viral genomes.
  - Physical sheering of DNA (e.g. from Covaris) and adaptor ligation. This is more technically challenging and the upfront cost is greater. However, the per sample cost is less and the uniformity of sequence coverage is better and less susceptible to variations in %GC.
- After preparing libraries for each sample including the addition of unique indices per sample, normalization of the quantity of DNA added per sample into the pooled tube (PAL) that will be sequenced is essential in order to ensure each sample has adequate coverage. There are again at least two possible methods:

- Using the Nextera XT kit Guide 150319425031942 following the protocol revision C (http://support.illumina.com/downloads/nextera_xt_sample_preparation_guide_15031942.html) where a bead-based method ensures simple normalization of sample quantities that are added. Other bead based normalisation kits are available.
- Measurement of the concentration of each sample.

- Ideally the fragment size of the libraries should also be measured before addition to the PAL tube in order to ensure the correct range for efficient sequencing (250 - 1000 bp). This can be achieved by fragment analysers such as LabChip from Perkin Elmer or BioAnalyser from Agilent.
- During library preparation a positive control comprising DNA from a known isolate (to check the effectiveness of library preparation and the absence of sample transposition) and ideally a negative control (to check for lack of contamination) should be included.

Sequencing

There are several short read sequencing technologies that are currently on the market, including Illumina and Ion Torrent™ Personal Genome Machine™ (from Thermo Fisher Scientific company) both of which have 'desktop' machines, the MiSeq and Ion Torrent™ Personal Genome Machine™ (PGM), respectively. These technologies also have larger capacity high throughput machines. When deciding which technology to use it is important to consider capacity and speed. Is turnaround time crucial and, if so, how much of a sequencing plate is required to be filled before it is sufficiently cost effective?

Whatever the technology used one critical step that should always be carried out and audited is the on-machine quality metrics calculation. It implies that the quality of the data as assessed by the machine is recorded as well as the quality of the final output fastq files. On the Illumina platform this will include metrics such as cluster density and percentage of clusters that pass/fail.

Post-sequencing data processing

- **Demultiplexing**
  When processing samples through the desktop machines, e.g. MiSeq, the processing of raw reads into per sample fastq files can occur on board the machine itself. However, for the higher throughput machines a server/computing infrastructure will likely be required.
- **Quality Control**
  Once the sequencing data has been demultiplexed it is critical that quality assessment is carried out (fastQC, see Appendix D, is recommended) and that, if necessary, poor quality data is removed using software such as Trimmomatic (Appendix D). At the very least adapter removal should be performed.
- **Analytical processes**
  Before embarking on the process of analysing samples to obtain results it is critical to think what the end point is and what result needs to be reported. Then a literature search can be carried out to assess how this can be best achieved. The list of software provided as part of the ENGAGE project (Appendix D) or the tools listed here (https://omictools.com/whole-genome-resequencing-category) will be a good place to start. A key consideration will be throughput. For any more than a few samples per week, a web based solution will probably not be suitable since it will be too person-hour expensive and difficult to audit and record. Alternatives for running analytic pipelines include:

- **Web services**

  A good example of this are the services at the Centre for Genomic Epidemiology (https://cge.cbs.dtu.dk/services/all.php). These allow sample by sample processing and in some cases batch processing. However, tracking the result outcomes and version of software is challenging.

- **Galaxy**

  A wide range of tools as listed in this report (Appendix D) are available via the Galaxy website (https://usegalaxy.org/) and these can be chained together into pipelines. This offers a lot of flexibility although it is likely that downstream processing of the outputs will be required in order to make them ready for interpretation.

- **Infrastructure**

  If processing any more than a few samples a dedicated server running best practice software is desirable. However, this will require ongoing dedicated IT support and programmatic bioinformatics skills.

Whichever solution is chosen from the options listed above, the location for the long term storage of the data should be considered. Although data can be uploaded to the public nucleotide archives (e.g. EMBL-EBI (https://www.ebi.ac.uk/ena/submit/sra/#home) or NCBI https://www.ncbi.nlm.nih.gov/sra/docs/submitportal/), it is likely that local storage of the files will be necessary. The amount of storage space required will be in the order of several terabytes. A resilient storage system recommended by local IT should be purchased unless they can give assurance of being able to store data of this magnitude.

Reporting

It is critical to think about the format and content of the final report that contains results derived from WGS. At an early stage consultation with the end-users (microbiologists, clinicians and epidemiologists) should be carried out in order to discuss what should be reported. The process by which the final outputs from the analytical pathways can be converted into a report should be planned at an early stage, to enable automatization.

Sample tracking and auditing

Throughout all these processes good record keeping and tracking of sample progress should be employed in order to allow construction of a full audit trail. A NGS sample LIMS (Laboratory Information and Management Systems) would be recommended such as the one listed here (https://omictools.com/lims-category).

References

Depledge, D.P., Palser, A.L., Watson, S.J., Lai, I.Y.-C., Gray, E.R., Grant, P., Kanda, R.K., Leproust, E., Kellam, P., and Breuer, J. (2011). Specific Capture and Whole-Genome Sequencing of Viruses from Clinical Samples. PLOS ONE 6, e27805.

Quick, J., Loman, N.J., Duraffour, S., Simpson, J.T., Severi, E., Cowley, L., Bore, J.A., Koundouno, R., Dudas, G., Mikhail, A., et al. (2016). Real-time, portable genome sequencing for Ebola surveillance. Nature 530, 228–232.

Companies’ main websites

- QIAGEN: https://www.qiagen.com/gb/
- STRATEC: https://www.molecular.stratec.com/home
- THERMO FISHER: https://www.thermofisher.com
- PROMEGA: https://www.promega.co.uk/
- COVARIS: http://covaris.com/
- ILLUMINA: https://www.illumina.com/
- PERKIN ELMER: http://www.perkinelmer.com
- AGILENT: https://www.agilent.com/

--- --- ---