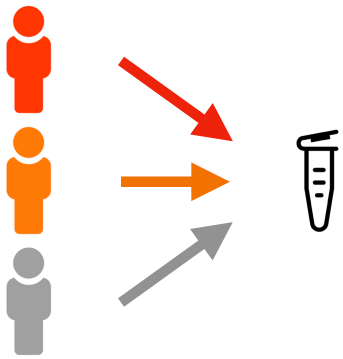
Pooling samples for NGS library preparation

Is pooling the same as multiplexing?

There are similarities between pooling and the multiplexing of samples in NGS experiments, but they aren't the same. We are going to describe pooling as the combination of samples prior to any experimental process.



Multiplexing involves the addition of a barcode to each sample, which is pooled together for sequencing. The key difference between multiplexing and pooling is that the barcode allows the data to be separated again for analysis.

Why are samples pooled?

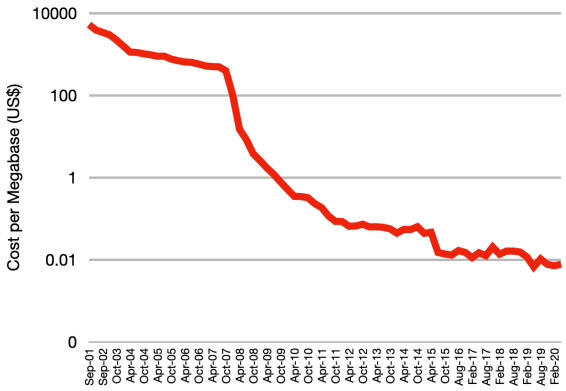
There are three main reasons why samples are pooled:

1) Sample availability – there are difficulties in collecting enough material for achieve the minimum input for a particular technology.

2) Cost savings – if enough quality data can be achieved by pooling samples together there can be a cost saving by running fewer experiments.

3) Time savings by having fewer sample handing and library preparation steps in an experiment.

In the early days of omic-scale technologies, these reasons were real barriers to the uptake of sequencing techniques. However, with sequencing costs reducing each year and techniques now being applicable at the single cell level, the decision to pool samples becomes an increasingly difficult one to justify to your peers as part of the funding or publication process.



Source: https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data

When might pooling be a good idea?

- Pooling samples may be beneficial when your study design aims to detect a simple or binary signal amongst the data.

- Pooling can be useful when identifying pathogens or diseases present in populations when individual testing is prohibited by cost or time constraints.

Pooling for variant discovery experiments

Anand et al. (2016) review how a pool-seq approach can be used for calling variants. They conclude:

"pooling of DNA creates new problems and challenges for accurate variant call and allele frequency (AF) estimation. In particular, sequencing errors confound with the alleles present at low frequency in the pools possibly giving rise to false positive variants."

However, their work does also conclude that:

"*Pool-seq AFs are robust and reliable by comparing them with public variant databases and in-house SNP-genotyping data of individual subjects of pools."*

The decision to employ a pooling approach should therefore be based on a careful review of the experimental purpose and is likely to be most useful when looking for population rather than sample focused outputs.

Pooling in RNAseq experiments

Pooling in RNAseq experiments is rarely a good idea. The main reason being that the statistical approaches used in RNAseq to identify differentially expressed genes are going to be influenced by any sample outliers introduced to your pool. The outlier will drag the other sample values towards the outlier value resulting in an incorrect determination of expression differences.

Even in the absence of gross outliers, there are issues of balancing the pool to ensure that all samples are contributing the same amount of signal to the pool which will, again, bias any analysis.

Assefa et al. (2020) do however point out a scenario where pooling may be beneficial in an RNAseq experiment, stating:

"for scenarios with a high biological variability, a small pool size such as 2 can be effective to optimize the cost of the experiment and maintain the power that one would attain without pooling."

References

Anand, S., Mangano, E., Barizzone, N. et al. Next Generation Sequencing of Pooled Samples: Guideline for Variants' Filtering. Sci Rep 6, 33735 (2016). https://doi.org/10.1038/srep33735

Takele Assefa, A., Vandesompele, J. & Thas, O. On the utility of RNA sample pooling to optimize cost and statistical power in RNA sequencing experiments. BMC Genomics 21, 312 (2020). https://doi.org/10.1186/s12864-020-6721-y