General sequencing experiment recommendations of the Ottawa Bioinformatics Core Facility

Updated March 23, 2022

While the cost of sequencing has declined over the years, sequencing experiments remain expensive. In practice, most experimental design is constrained by available funding and resources. Effective experimental design is in part an attempt to maximize information while minimizing cost and effort. Developing experimental designs in consultation with a bioinformaticist can be helpful in achieving these goals, particularly if you are intending them to analyze the data after sequencing.

Many experiments analyzed by the Ottawa Bioinformatics Core Facility follow common patterns, and we outline our general recommendations for experimentation below. You may also consider the ENCODE guidelines[1] as generally accepted and citable experimental design standards for all experiment types described below.

We would be happy to discuss your experimental design with you. If you are considering more complex experimental designs such as time series analysis or using novel protocols, please contact a bioinformaticist to discuss your project.

Please note that the recommendations below assume Illumina sequencing platforms. We have limited experience with other sequencing platforms.

RNA-seq

For basic RNA-seq experimental designs (*i.e.,* estimating gene expression differences between conditions) we recommend the following:

At least three biological replicates per condition. Commonly used fold change analyses tools (e.g., DESeq2[2], edgeR[3]) require or recommend replicates to estimate fold change significance. In general, increasing the number of biological replicates will improve the quality of the fold change estimates generated. In cases of cell culture, biological replicates may be generated from separate cultures of the same cell lineage, however the biological variability may be low, and such samples will be more like technical replicates. It may be appropriate to use different cell culture lineages if available to increase biological variability.

At least 30 million mapped and gene-assigned reads per sample. This generally means planning for about 40M sequenced reads per sample as the yield of mapped and gene assigned reads is variable.

If working with samples expected to have a high PCR duplication rate such as small sample protocols, consider using Unique Molecular Identifiers (UMIs) to allow the identification of PCR duplicates. UMIs are added as an identifying sequence to cDNA fragments before PCR amplification, allowing the identification and removal of PCR duplicates.

---

[1] https://www.encodeproject.org/data-standards/

[2] Love, Michael I., Wolfgang Huber, and Simon Anders. "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." *Genome biology* 15.12 (2014): 1-21.

[3] Robinson, Mark D., Davis J. McCarthy, and Gordon K. Smyth. "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data." *Bioinformatics* 26.1 (2010): 139-140.

Single end short reads (e.g. 75bp on the NextSeq 500) are sufficient for gene level expression estimates. Longer and/or paired end reads can be substantially more expensive and do not provide significantly more gene expression information[4], though they can be of value if transcript or exon level expression estimates are needed. The accuracy of such estimates is variable depending on read properties, read depth and gene structure. Please contact a Bioinformaticist to discuss your experimental design if you are interested in transcript or exon level expression estimates.

ERCC spike-ins can be used to correct for absolute differences in numbers of cellular RNA transcripts between conditions by adding these known sequences at known concentration per number of cells for normalization purposes. Though ERCC spike-ins are generally recommended, our experience in using these has been mixed; they can yield questionable values leading us to use other methods for the final analysis. It is however inexpensive (a purchased mix is usable for many samples) and sequencing (<1% of sequenced reads), the ERCC mix usage has a low marginal cost once purchased and their presence can be useful for diagnostic purposes.

Small RNA-seq

These are sequencing experiments optimized for small RNA fragments, usually to identify processed miRNA which are then mapped to the mirBase database of mature miRNA.

As with RNA-seq experiments, we recommend at least 3 biological replicates per condition.

We recommend targeting 10M mapped reads per sample. The yield of mapped mature miRNA from small RNA-seq experiments is highly variable and can be very low for small samples and exosomal extractions. This makes it difficult to reliably predict yields. Planning for deeper sequencing (assuming low yields) can mitigate this.

Spike-ins are recommended as miRNA counts per cell can vary widely, particularly across developmental time series. This is usually an endogenous synthetic processed microRNA, added in known quantities for normalization, commonly *cel-mir-39*[5], a synthetic *C. elegans* processed microRNA.

ChIP-seq

ChIP-seq may be divided into Histone ChIP-seq which identifies histone modifications across the genome, and transcription factor ChIP-seq which identifies transcription factor binding at specific genomic locations.

For both transcription factor and histone modification ChiP-seq we recommend short single end sequencing if price is a limitation. Longer reads have slightly higher mappability, and paired end reads can give an accurate fragment size distribution which can help with peak calling, but both options increase costs. Fragment length may also be estimated from single end reads.[6]

For Transcription Factor ChIP-seq we recommend:

---

[4] Li, Bo, et al. "RNA-Seq gene expression estimation with read mapping uncertainty." *Bioinformatics* 26.4 (2010): 493-500.

[5] https://norgenbiotek.com/product/microrna-cel-mir-39-spike-kit

[6] Ramachandran, Parameswaran, et al. "MaSC: mappability-sensitive cross-correlation for estimating mean fragment length of single-end short-read sequencing data." *Bioinformatics* 29.4 (2013): 444-450.

30 million mapped reads; this may require planning for a higher numbers of raw reads as not all reads will map to the genome.

At least 3 Biological replicates are recommended to ensure that called peaks are not spurious.

Histone ChIP-seq may be further subdivided into experiments looking at narrow-peak histone modifications and broad peak histone modifications.

Narrow peak histone modifications are those which are restricted to discrete, narrow regions such as H3K4me3. For such histones we recommend at least 30 million mapped reads.

Broad peak histone modifications are those which tend to be found in more dispersed, wider regions, such as H3K27me3; these require more reads to identify enriched regions.

For broad peak histone ChIP-seq we recommend at least 50 million mapped reads per sample.

For ATAC-seq:

At least 50 million mapped fragments are required for detection of open chromatin, and 200 million are required for TF footprinting. Paired end short reads are recommended for ATAC-seq to determine fragment sizes for nucleosome footprinting, and to aid in detection and removal of PCR duplicates.[7] We also recommend at least three biological replicates.

10X single cell

Single cell analysis design is experiment specific; we recommend reviewing Dr. David Cook's presentation on single cell RNA-seq analysis which has some general guidelines. [8]

Guidelines from 10X Genomics recommend at least 20,000 reads per cell, so total sequencing depth required will depend on the number of cells captured in the experiment. Read lengths are specified by the single cell protocol being used.

General issues

Prototyping analyses

Prototyping analyses before generating experimental data can help supply insight to guide more effective experimental design. The NCBI Gene Expression Omnibus (GEO)[9] and dbGAP[10] are rich sources of raw experimental data to help you clarify your hypotheses. Though the exact experimental conditions of interest may not be available in the archived data, analogous data sufficient to provide insight may

[7] Yan, Feng, et al. "From reads to insight: a hitchhiker's guide to ATAC-seq data analysis." *Genome biology* 21.1 (2020): 1-16.

[8] https://gitlab.com/ohri/2021-scn-rnaseq-workshop/-/raw/main/Files/scRNAseq_workshop_ppt.pptx?inline=false

[9] Barrett, Tanya, et al. "NCBI GEO: archive for functional genomics data sets—update." *Nucleic acids research* 41.D1 (2012): D991-D995.

[10] Tryka, Kimberly A., et al. "NCBI's Database of Genotypes and Phenotypes: dbGaP." *Nucleic acids research* 42.D1 (2014): D975-D979.

well be, for example similar experiments in another species or different treatments for the disease of interest.

UMIs

The use of UMIs is becoming increasingly common as a means of correcting for PCR duplication. UMIs are already a part of 10X single cell sequencing protocols but may be considered for small samples in general or other experiments where PCR duplicates are likely to have a significant effect.

De-risking large scale projects

Large scale experiments involving the sequencing of many samples are costly in labor, consumables, and sequencing. Failure of individual samples to meet sequencing goals, or the loss of a sequencing run due to technical issues can lead to data insufficient to meet test the desired hypotheses. Careful experimental design can help reduce these risks.

Before starting the experiment, discuss the experimental design with a bioinformaticist, ideally the same one who will analyze the data. It's common for bioinformaticists to be asked to analyze data after completion of the experiment only to find that the hypotheses of interest are not testable with the experimental design used.

For experimental protocols that are new to the lab or implementor, test sequencing runs on a small number of samples can help decide how well the protocol is working and may supply evidence as to how to optimize.

When multiple samples are sequenced on the same run using multiplexing, it can be difficult to balance reads per sample due to the nature of multiplexing. Also, a bad sample can consume many sequencing reads. Multiplexing many samples across multiple runs is a strategy which allows incremental analysis of results as they are generated, removal of bad samples based on quality analyses and removal of samples for incremental runs when sequencing read threshold are achieved.

Exogenous Controls

Exogenous controls are available for most sequencing experiment types and should be considered for all experimental designs.[11] In the absence of spike-in controls, most sequencing experiments are implicitly normalized at the sample level and so do not capture global differences in cellular abundance of the molecule in question which can be biologically important.

[11] Chen, Kaifu, et al. "The overlooked fact: fundamental need for spike-in control for virtually all genome-wide analyses." *Molecular and cellular biology* 36.5 (2015): 662-667.