THE UNIVERSITY OF CHICAGO MEDICINE & BIOLOGICAL SCIENCES
Center for Research Informatics

Bioinformatics Advice on Experimental Design

Where do I start?

Please refer to the following guide to better plan your experiments for good statistical analysis, best suited for your research needs. **Statistics cannot rescue a bad experimental design.**

Please contact our Bioinformatics team for a consultation when in doubt.

Next Generation Sequencing (NGS) experiments

Many steps in the experimental process can introduce various biases and errors, and careful consideration must be given to the following aspects:

- **Platform choice:**

Platform	Platform				
	Genome Sequencer FLX Titanium System	Genome Analyzer IIx	Hiseq 2000	SOLiD 4 system	HeliScope
Company	Roche	Illumina	Illumina	Applied Biosystems	Helicos Biosciences
Read length	400-600bp	2x100bp	2x100-150bp	50 +25bp	~30bp
Samples per run	16	8	16	16	50
Reads per run	~1 million	~300million	~800 million	>700 million	~500 million
Run time	10 h	8 days	8 days	11-13 days	8 days
Website	www.454.com	www.illumina.com	www.illumina.com	www.appliedbiosystems.com	www.helicosbio.com

These numbers change rapidly as technology improves. Please note that these numbers are based on data from Oct. 2010. Please refer to the websites listed under each platform for the latest numbers.

- **Type of Run – Paired End (PE) or Single End (SE):**

The following table provides a guide to what type to run is recommended for typical applications of various NGS assays.

Paired End	Single End
RNASeq - De novo Assembly	RNASeq - Counting
RNASeq - Splicing	ChIP-Seq - Counting
ChIP Seq – Epigenetic modifications	
DNA – SNP Identification	
DNA – Indel identification	
DNA – Structural variants	

- **Read Length:**
  50bp reads are typically sufficient for read mapping to the reference genome, and RNASeq counting experiments. >100bp reads are useful for whole genome and transcriptome studies based on the application.

- **Replication:**
  Samples must be sequenced with replicates to identify sources of variance and increase statistical power to separate true biological variance from technical variance. Biological replicates are critical whereas technical replicates are typically not required.

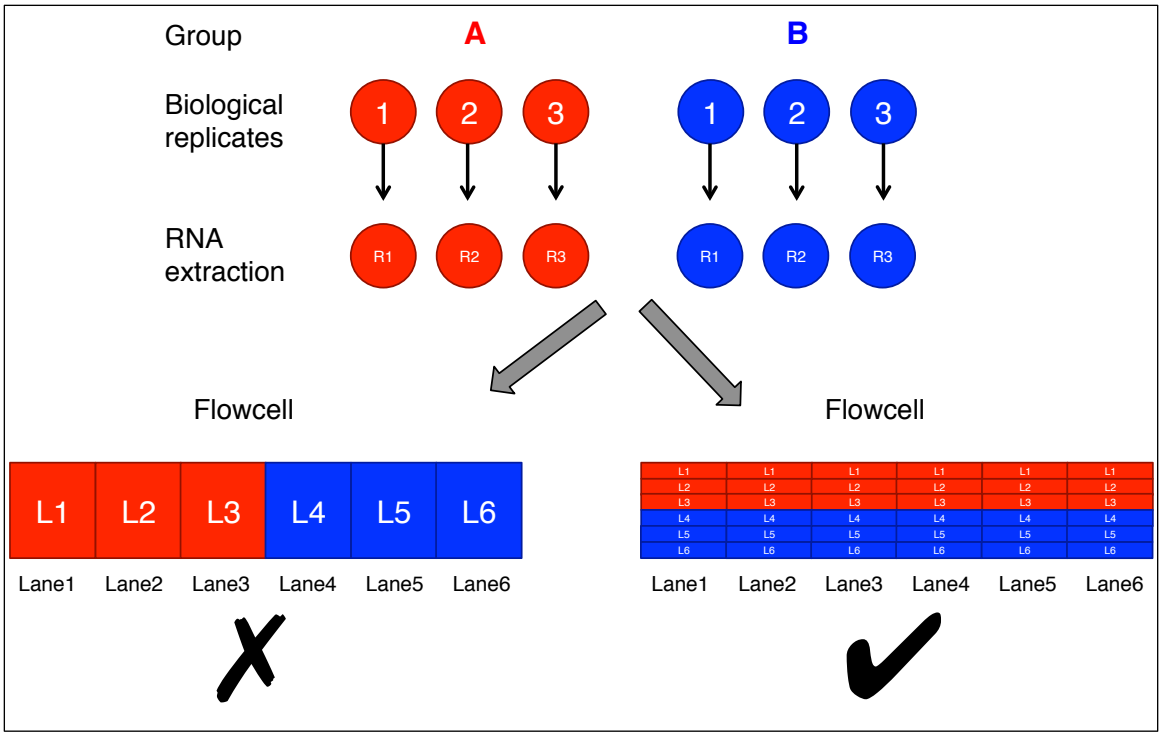
  Cutting back replicates to reduce cost might seem like a good option, but remember: A sample or sequencing run can fail, and lead to repeating the experiment.

  In general, 4 biological replicates per experiment are recommended, however, 3 replicates if also reasonable. Please consult with us with further questions. You can also use http://bioinformatics.bc.edu/marthlab/scotty/help.html for calculation of power from your pilot data.

- **Randomization:**
  Assign individuals at random to different groups to reduce bias. We recommend randomization of samples such that each sequencing lane contains samples from all experimental groups. Please refer to *Blocking and Multiplexing* below to understand how to do this.

- **Blocking & Multiplexing:**
  Distribute samples across various lanes on the flowcell to avoid lane effects. Use multiplexing effectively for balanced block designs. (Fig.1) But all samples cannot be sequenced on each lane as the number of unique barcodes for each lane also limits us. Solution: Balanced incomplete block design.

  "Block what you can and randomize what you cannot." – Box, Hunter, & Hunter (1978)



If,
I= Number of groups/treatments
J= Number of biological replicates per treatment
s= Number of unique barcodes that can be added in one lane
L= Number of lanes sequenced
T=Total number of technical replicates

$$T = \frac{sL}{JI}$$

If s<I, complete block design is not possible. [1]

- **Sequencing depth:**

The following table provides general recommendations for coverage/reads (https://genohub.com/recommended-sequencing-coverage-by-application/) for typical read lengths for the HUMAN genome. Please visit https://genohub.com/next-generation-sequencing-guide/#reads for typical number of reads/lane for various commonly used NGS platforms.

A useful resource from Illumina for specific coverage estimates for various Illumina instruments and genomes of different sizes is http://support.illumina.com/downloads/sequencing_coverage_calculator.html

DNA:

Category	Application	Recommended coverage (X) or reads (in millions)
Whole genome	Re-sequencing	30-80X
	De novo assembly	100X
	SNP detection	10-30X
	Indel detection	60X
	Genotype calls	35X
	CNVs	1-8X
Whole exome sequencing	SNVs detection	100x (3-13x local depth)
	Indel detection	NOT recommended
	De novo assembly	>100M
DNA Target-Based Sequencing	ChIP-Seq	10-40X [10-14M (sharp peaks); 20-40M (broad marks)]
	Hi-C	100M
DNA Methylation Sequencing	CAP-Seq	>20M
	RRBS (Reduced Representation Bisulfite Sequencing)	10X
	Bisulfite-Seq	5-15X; 30X

RNA (for human/mouse genome):
Please note that the number of reads your need for any type of RNASeq also depends on the desired dynamic range of expression.

Category	Application	Recommended of mapped reads (in millions)
Transcriptome Sequencing (RNA-Seq)	Differential expression	10-25M
	Alternative splicing	50-100M
	Allele specific expression	50-100M
RNA-Target-Based Sequencing	CLIP-Seq	10-40M
	PAR-CLIP	5-15M
	RIP-Seq	5-20M
Small RNA (microRNA) Sequencing	Differential Expression	~1-2M
	Discovery	~5-8M

Microarray Experiments

A very useful resource for microarray design is:
http://discover.nci.nih.gov/microarrayAnalysis/Experimental.Design.jsp

- **Balanced samples**
  - Same amount of cases and controls
  - Matched phenotypes: gender, age, etc.
- **Biological replicates**
  - Pure background to avoid biological variation
  - More replicates are needed if there is larger variation between individuals and small difference between groups
- **Avoid technical variation**
  - Process sample at same condition as much as possible
  - Technician, reagents, time, procedures
- **Randomize samples on array**
  - Avoid confounding technical and biological factors
  - Randomly put samples on different array slides and positions

Frequently Asked Questions

- **What if I do not have replicates of data points?**
  Understand the limitations of un-replicated data! You cannot separate technical variance from biological variance, thus, the results only apply to the data points sequenced but cannot be extrapolated to the population.

- **What is difference between Biological replicates and technical replicates?**
  Technical replicates: measure quantity from 1 source. This measures the reproducibility of the results. The differences are based only on technical issues in the measurement. (I weigh myself three times, do I get different weights? How different?)
  Biological replicates: measure a quantity from difference sources under the same conditions. Tumors from 5 different people with lung cancer may show similar gene expression patterns. These replicates are useful to show what is similar in your replicates and how they are different from a different set of conditions (ie. treated, normal).

  Biological variation is intrinsic to all organisms; it may be influenced by genetic or environmental factors, as well as by whether the samples are pooled or individual. Technical variation is introduced during the extraction, labeling and hybridization of samples. Measurement is associated with reading the fluorescent signals, which may be affected by factors such as dust on the array.

References

1. P. L. Auer and R. W. Doerge. 2010. *Statistical design and analysis of RNA sequencing data.* Genetics 185:405-416.