Summer Institute In Statistical Genetics 2017

Integrative Genomics

1a. Experimental Design and Hypothesis Testing

ggibson.gt@gmail.com
http://www.cig.gatech.edu

Georgia Tech

Center for Integrative Genomics

Steps in a Gene Expression Profiling Study

1. Experimental Design (this morning)
2. RNA Sequencing (next)
3. Short read alignment (this afternoon)
4. Normalization (after the break and tomorrow)
5. Hypothesis testing (this morning)
6. Downstream analyses (tomorrow and next module)
7. Genetic analysis (tomorrow afternoon)

Course Outline

1a. Experimental Design and Hypothesis Testing (GG)
1b. Normalization (GG)

2a. RNASeq (MI)
2b. RNASeq practical (MI)

3a. Variance components and Normalization Practical (GG)
3b. eQTL (GG)

4a. Network Analysis (MI)
4b. WGCNA with practical (MI)

5a. Epigenetics and Single Cell RNASeq (GG)
5b. Integrative methods and Microbiome (MI)

RNA Sequencing

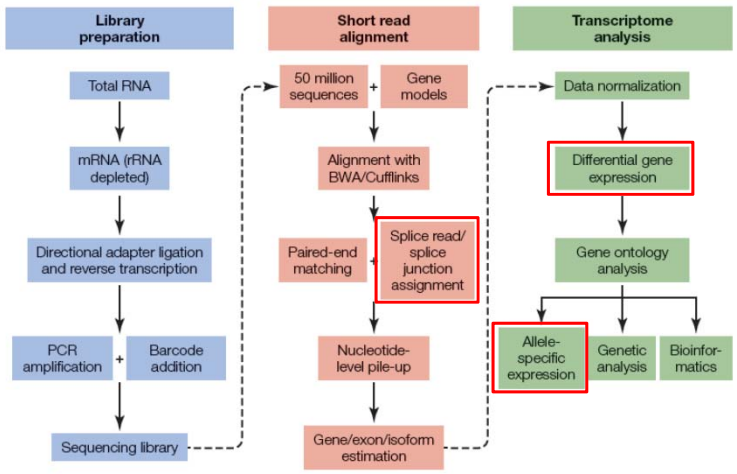
RNA is prepared, mRNA is captured on polyT beads, fragmented, and converted to cDNA using either a stranded or unstranded protocol, usually with 12-24X multiplexing

1. Single-end reads
   - Maximizes the total number of independent reads (50M optimal)
   - When RNA is degraded, eg FFPE specimens

2. Paired-end reads
   - Slightly more accurate alignment
   - But typically lower coverage (25M reads)
   - Better for estimation of alternate splicing and ASE

3. 3' targeted
   - Lexogen protocol is one fifth the cost ($70 vs $350 per sample)
   - Ideal for large sample studies when funds are a concern
   - Single Cell drop digital dd-scRNASeq is also 3' targeted

RNASeq workflow

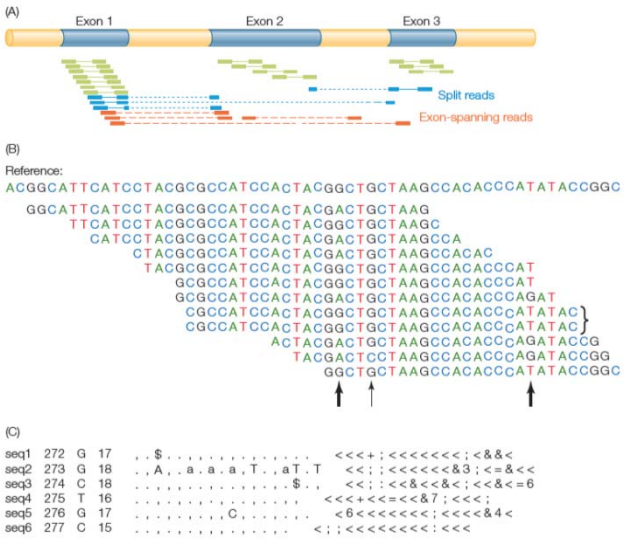


Core Open Source Software

While the Tuxedo protocol (Bowtie, Tophat, Cufflinks, Cuffdiff, CummeRbund) remains popular, we recommend the following open source alternatives.

1. Short Read Alignment
   - STAR https://github.com/alexdobin/STAR/releases
   - HISAT2 https://ccb.jhu.edu/software/hisat2/index.shtml
2. Read counting
   - HTseq http://www-huber.embl.de/HTSeq/doc/overview.html
3. Differential Expression
   - DESeq https://bioconductor.org/packages/release/bioc/html/DESeq2.html
   - DExSeq https://www.bioconductor.org/packages/release/bioc/html/DEXSeq.html
   - edgeR https://bioconductor.org/packages/release/bioc/html/edgeR.html
   - Voom http://web.mit.edu/~r/current/arch/i386_linux26/lib/R/library/limma/html/voom.html
4. Data Normalization
   - SVASeq https://www.bioconductor.org/packages/release/bioc/html/sva.html
   - Combat https://www.rdocumentation.org/packages/sva/versions/3.20.0/topics/ComBat
   - PEER http://www.sanger.ac.uk/science/tools/peer
   - SNM https://www.bioconductor.org/packages/release/bioc/html/snm.html

Read alignment



Basic Experimental Design: Replication

Often you will have a fixed budget that constrains how many arrays can be processed. So your first task is to determine what levels of replication you can afford, and how they will impact statistical power.

Technical Replication:

- RNA preparation (eg. from adjacent biopsies)
- cDNA synthesis (pooling minimizes outlier effects)
- library preparation
- array hybridization (with commercial arrays, quality generally very high)
- duplicate "probes" for the same gene

Biological Replication:

Fixed effects:
- gender
- treatment (drug, growth regimen, tissue)
- time of sampling (repeated measures in some cases)
- genotype (IF specifically chosen and resampled)

Random effects
- individual from a population
- field plot

Basic Experimental Design: Contrasts

At the same time, you need to be aware of the contrasts you wish to make since by tweaking the design you may gain a lot in terms of what you can infer.

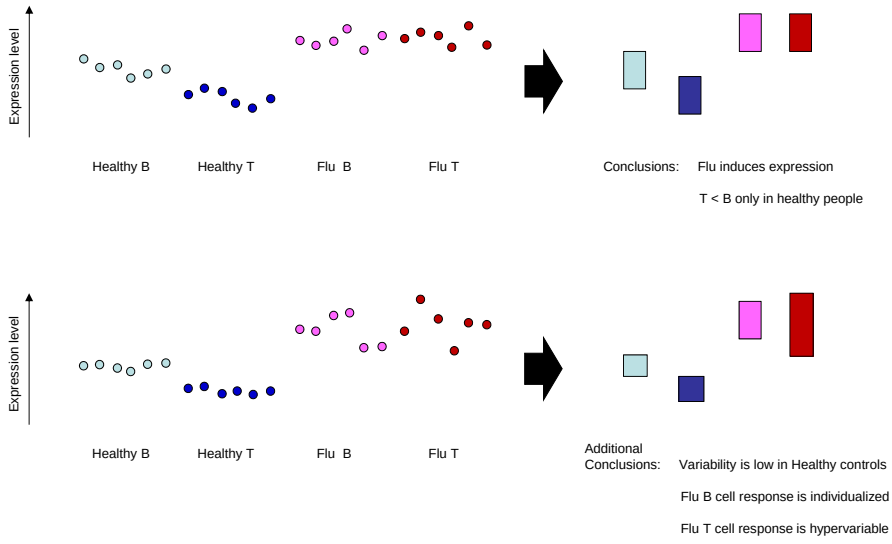
Suppose you want to compare B cells and T cells from Healthy controls and Flu patients, and you have the funds to generate 24 RNASeq profiles

What is the best design?
- 6 controls and 6 patients, each donating both a B and a T cell sample
- 12 controls and 12 patients, each donating either a B or a T cell sample
- 3 controls and 3 patients, each donating a B and a T cell sample, processed twice
- 3 controls and 3 patients, each donating 2 B and 2 T cell samples, on separate days
- same as above, but only men or only women
- 12 controls and 12 patients, each donating either a B or a T cell sample, but pooling two visits

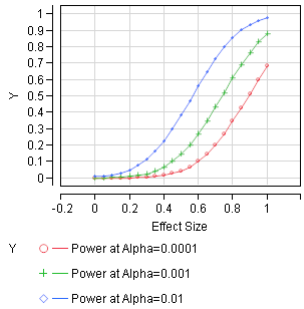
Main effects can only be contrasted if you have biological replicates:
reducing the number of individuals may allow you to address intra-individual variability

Interaction effects allow you to ask questions like whether B cells and T cells differ more between healthy volunteers or sick patients

Two hypothetical sets of results



Basic Experimental Design: Statistical Power



Power is a function of:

- the sample size
- the magnitude of the difference between classes
- the variance within the classes being compared

Since two of these parameters vary for each gene, Power in a GEP experiment is usually assessed in terms of the effect size (amount of variance explained), not as a magnitude of difference.

But, biologically it is not clear what effect size is important for any given gene.

Confounding Design Biases

At the design step, avoid confounding biological factors:

- don't contrast bloods from young males and old females
- don't contrast hearts from normal mice and livers from obese ones

as far as possible, balance all biological factors

Be aware of the potential for technical confounding:

- date of RNA extraction or hybridization
- batch of samples (particularly for microarray studies)
- person who prepared the libraries
- SE or PE, read length and quality of reads
- quality of RNA (RIN = Bioanalyzer RNA Integrity Number)

Linear Modeling

1. Normalize the samples:

   log(fluorescence) = μ + Array + Residual

   OR variance transforms, OR supervised methods

2. For each gene, assess significance of treatment effects on the Residual (ie. Expression level) with a linear model:

   Residual = μ + Sex + Geno + Treat + Interact + Error

Wolfinger *et al*, 2001. *J Comput Biol* **8**: 625-637

Hypothesis testing

1. Generally we are interested in asking whether there is a significant difference between two or more treatment group(s) on a gene-by-gene basis

2. For a simple contrast, we can use a t-test to test the hypothesis. Significance is always a function of:
   1. The difference between the two groups: [5,6,4] vs [7,5,6] has a diff of 1
   2. The variance within the groups: [2,5,8] vs [3,6,9] does as well, but is less obvious
   3. The sample size: [5,6,4,4,6,5] vs [7,5,6,5,6,7] is better

3. For contrasts involving multiple effects, we usually use General Linear Models in the ANOVA framework (analysis of variance: significance is assessed as the F ratio or between sample to residual sample variance.

4. edgeR uses limma to perform One-Way ANOVAs. This likelihood framework is very powerful, but constrains you to contrast treatments with a reference

5. Robust statistics (eg using lme4) also allow you to evaluate INTERACTION EFFECTS, namely not just whether two treatments are individual significant, but also whether one depends on the other

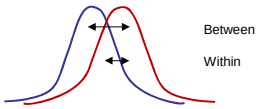
6. Given a list of p-values and DE estimates, we need to evaluate a significance threshold, which is usually done using False Discovery Rate (FDR) criteria, either B-H or a qvalue

T-tests and F-ratios

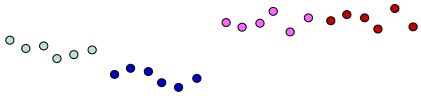
A two-sample t-test evaluates whether the two population means differ one another, given the pooled standard deviations of the sample and the number of observations:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p\sqrt{2/n}} \qquad s_p = \sqrt{\frac{s_{X_1}^2 + s_{X_2}^2}{2}}$$

F-statistics generalize this approach by asking whether the deviations between samples samples are large relative to the deviations within samples. The larger the difference between the means, the larger the F-ratio, hence significance is evaluated by contrasting variances.



Generally we start by evaluating whether there is variation among all the groups, and then test specific post-hoc contrasts. With multifactorial designs you also should be aware of the difference between Type I and Type III sums of squares, namely univariate and conditional tests.



Hypothesis testing in edgeR

If you compare the output of edgeR to those of a t-test or F-test, particularly for small n, you will often get *very* different results. There are at least 4 reasons for this.

1. edgeR performs a TMM normalization. Since RNASeq generates counts, we adjust for library size by computing cpm (counts per million). If a few high abundance transcripts vary by several percent, they throw off all the other estimates. TMM fixes this.

2. edgeR shrinks the variance of low abundance transcripts by fitting the distribution to the "negative binomial" expectation. Basically it adds values to account for sampling error at the low end so that comparing 0, 1 and 2 is more like comparing 10, 11 and 12.

3. edgeR also employs a powerful within-sample variance adjustment in its GLM fitting, with the result that it puts much more weight on fold-change than standard F-tests.

4. For a one-way ANOVA the approaches are similar (though you need to be careful about whether you fit an intercept, which means you compare multiple samples to a reference rather than to one another). For more complicated analyses involving two or more factors, nesting, and random effects, the modeling frameworks are really quite different.

I prefer to run ProcGLM or ProcMIXED in SAS or JMP-Genomics, but there is no clear equivalent in R – lme4 is close, but needs looping.

GEO and ArrayExpress