Summer Institute in Statistical Genetics 2016

Integrative Genomics

1a. Introduction

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

Course Outline

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

2a. RNASeq (MI)
2b. Clustering and Pathways (MI)

3a. Lab session – qvalue, PCA (GG)
3b. Lab session – SNM, edgeR (GG)

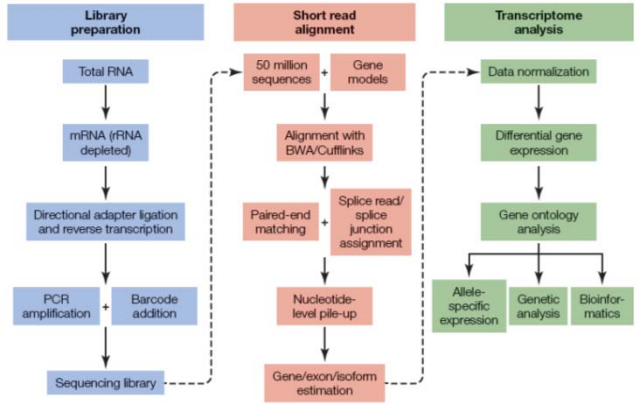
4a. Network Analysis (MI)
4b. Lab session - WGCNA (MI)

5a. Integrative methods (MI)
5b. eQTL Analysis (GG)

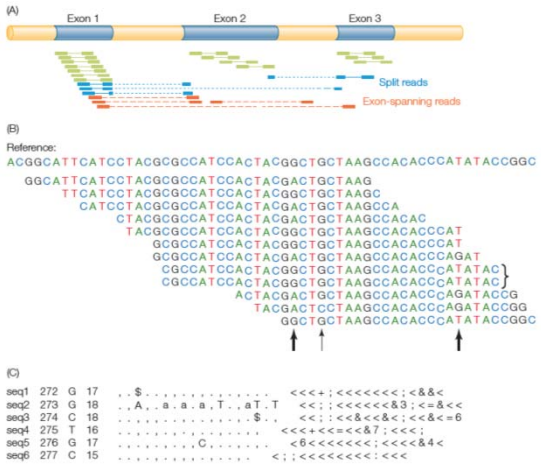
Applications of Expression Analysis

1. Atlases of gene expression for functional annotation
2. Identification of differentially expressed genes
3. Assembly of networks of co-regulated genes
4. Investigation of regulatory mechanisms
5. Evolutionary and ecological genomics
6. Clinical genomics
7. Quantitative basis of complex traits

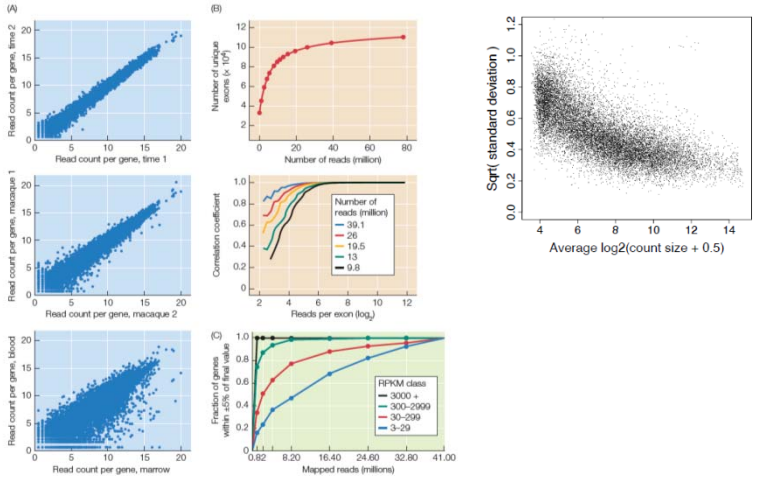
RNASeq workflow



Read alignment



Variability and Depth



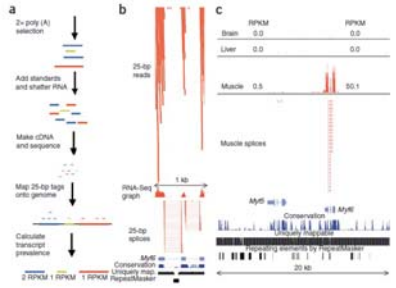
Microarray vs RNASeq

Advantages of Microarrays	Advantages of RNA-Seq
Less expensive	Disruptive technology
Better sensitivity for low abundance	Unbiased by prior gene knowledge
Computationally simpler	Alternative exon usage
Better-defined statistical properties	Allele specific expression (ASE)
Perfectly good for most applications	High repeatability
Disadvantages of Microarrays	**Disadvantages of RNA-Seq**
Only for humans, model organisms	More opportunity to screw up the analysis
Different platforms give different results	Oversold resolution of exon level and ASE
Large technical batch effects	Short read alignment biases
Sensitivity to polymorphism	Sensitivity to polymorphism
Low consistency of analytics among groups	Low consistency of analytics among groups

RNA Sequencing

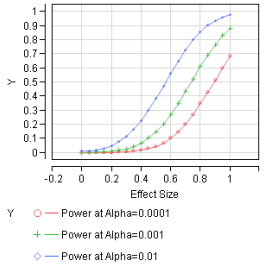
Platforms: Illumina HiSeq / LifeTech ProtonTorent / ABI SOLiD

Analytical Steps:
- Short read alignment (BWA, TopHat)
- Inference of abundance (Cufflinks, DESeq, etc)
  - Gene, Exon, and Isoform level analysis
- Normalization (as for microarrays)
- Inference of differential expression



Mortazavi et al (2008) *Nat. Methods*. **5**: 621

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 microarray 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.

Levels of 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 and labeling (pooling minimizes outlier effects)
- 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

Design Biases

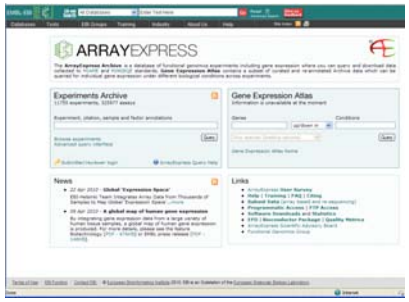
At the design step, avoid confounding biological factors:
- don't contrast bloods from young males and old females
- don't contrast hearts from normal mice and livers from obese ones

as far as possible, balance all biological factors

Be aware of the potential for technical confounding:
- date of RNA extraction of hybridization
- batch of arrays
- person who did the hybridization
- scanning software

For 2-color arrays, recognize that the *order of hybridization affects power*:
- Reference designs
- Loop designs
- Split-plot designs
- Molecular biologist's designs

GEO and ArrayExpress



A GEO record

☐ **2:** GSE17065 record: **Geographical Genomics of Human Leukocytes Gene Expression Variation** [*Homo sapiens*] Links

Summary: (Submitter supplied) Genome-wide association studies of transcript abundance in peripheral blood samples or derivative cell lines have demonstrated a preponderance of eSNP effects that, for the most part, involve regulatory polymorphisms in the differentially expressed gene. Several of these highlight associations that contribute to a variety of disease conditions, but the question arises as to how the associations are affected by the environment. Here we address the robustness of eSNP associations to environmental geography and population structure in a comparison of 194 Arab and Amazigh individuals from a city and two villages in southern Morocco.
1 related Platform

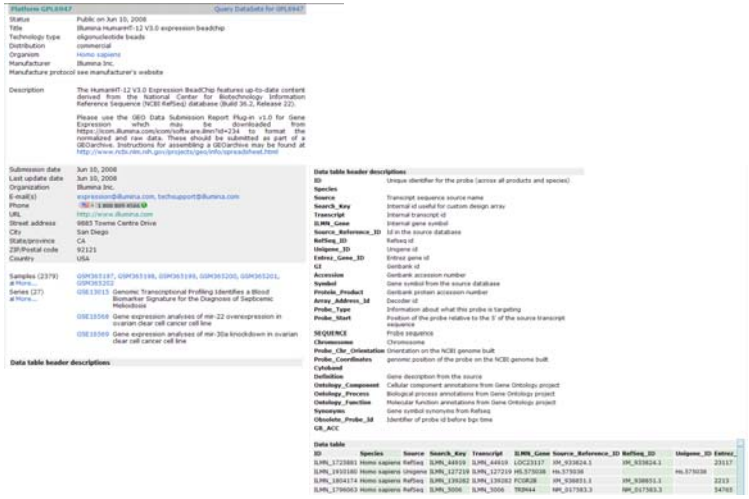
Type: Expression profiling by array

Supplementary Files: TXT download...

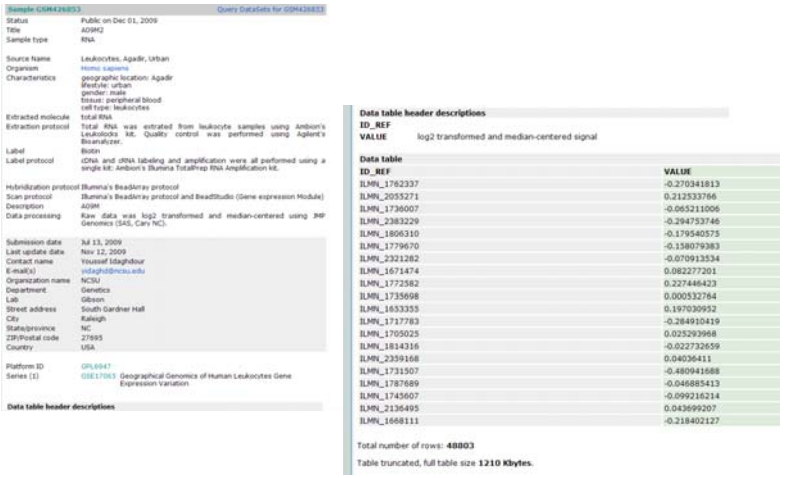
Samples: 194

GSM426853: A09M2
GSM426856: A106M5
GSM426859: A110M8
GSM426866: A135M15
GSM426869: A139M18
GSM426872: A147M21

A GEO platform



A GEO sample



Status	Public on Dec 01, 2009
Title	AGMK2
Sample type	RNA
Source Name	Leukocytes, Agadir, Urban
Organism	Homo sapiens
Characteristics	geographic location: Agadir lifestyle: urban gender: male tissue: peripheral blood cell type: leukocytes
Extracted molecule	total RNA
Extraction protocol	Total RNA was extrated from leukocyte samples using Ambion's LeukoLocks kit. Quality control was performed using Agilent's Bioanalyzer.
Label	Biotin
Label protocol	cDNA and cRNA labeling and amplification were all performed using a single kit: Ambion's Illumina TotalPrep RNA Amplification kit.
Hybridization protocol	Illumina's BeadArray protocol
Scan protocol	Illumina's BeadArray protocol and BeadStudio (Gene expression Module)
Description	AGMK
Data processing	Raw data was log2 transformed and median-centered using JMP Genomics (SAS, Cary NC).
Submission date	Jul 13, 2009
Last update date	Nov 12, 2009
Contact name	Youssef Idaghdour
E-mail(s)	yidaghd@ncsu.edu
Organization name	NCSU
Department	Genetics
Lab	Gibson
Street address	South Gardner Hall
City	Raleigh
State/province	NC
ZIP/Postal code	27695
Country	USA
Platform ID	GPL6947
Series (1)	GSE17065 Geographical Genomics of Human Leukocytes Gene Expression Variation

Data table header descriptions

Data table header descriptions

ID_REF	
VALUE	log2 transformed and median-centered signal

Data table

ID_REF	VALUE
ILMN_1762337	-0.270341813
ILMN_2055271	0.212533766
ILMN_1736007	-0.065211006
ILMN_2383229	-0.294753746
ILMN_1806310	-0.179540575
ILMN_1779670	-0.158079383
ILMN_2321282	-0.070913534
ILMN_1671474	0.082277201
ILMN_1772582	0.227446423
ILMN_1735698	0.000532764
ILMN_1653355	0.197030952
ILMN_1717783	-0.284910419
ILMN_1705025	0.025293968
ILMN_1814316	-0.022732659
ILMN_2359168	0.04036411
ILMN_1731507	-0.480941688
ILMN_1787689	-0.046885413
ILMN_1745607	-0.099216214
ILMN_2136495	0.043699207
ILMN_1668111	-0.218402127

Total number of rows: **48803**

Table truncated, full table size **1210 Kbytes**.

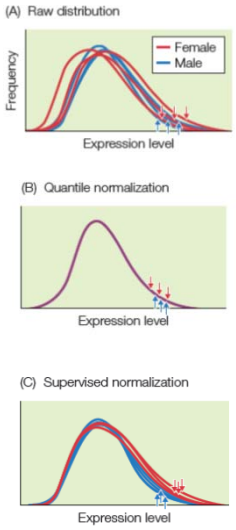
Summer Institute in Statistical Genetics 2016

Integrative Genomics

1b. Hypothesis Testing and Normalization

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

Rank vs Absolute Expression

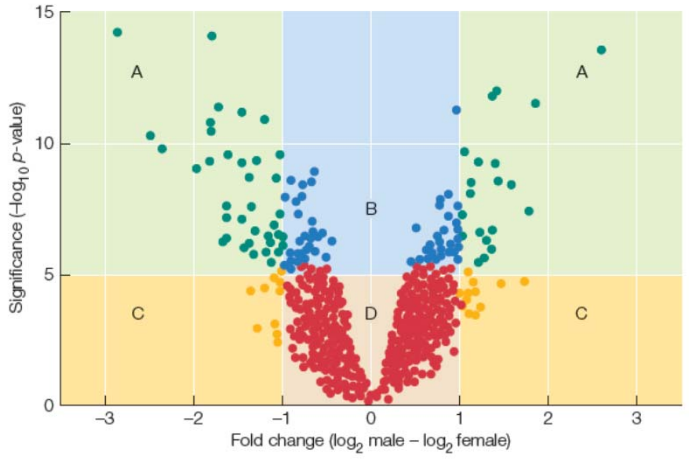


Raw data:
no effect

Variance transformed:
no effect

Mean centered:
significant effect

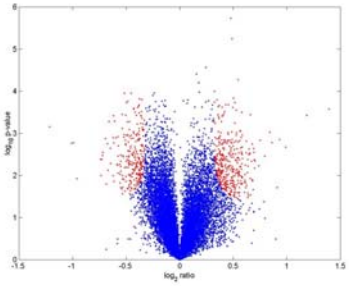
Transcriptome Volcano plots



Russ Wolfinger, Greg Gibson

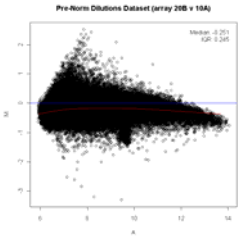
Variance estimators

- Gene-specific approach means that the power for each gene varies, but shrinkage can equilibrate the variance
- Permutation approach may be more appropriate where you have many treatments with low replication



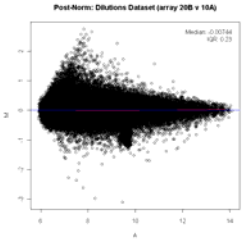
Gary Churchill, Katie Kerr

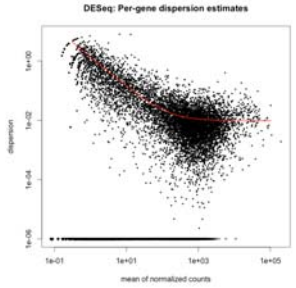
Deviation, Abundance, and Dispersion



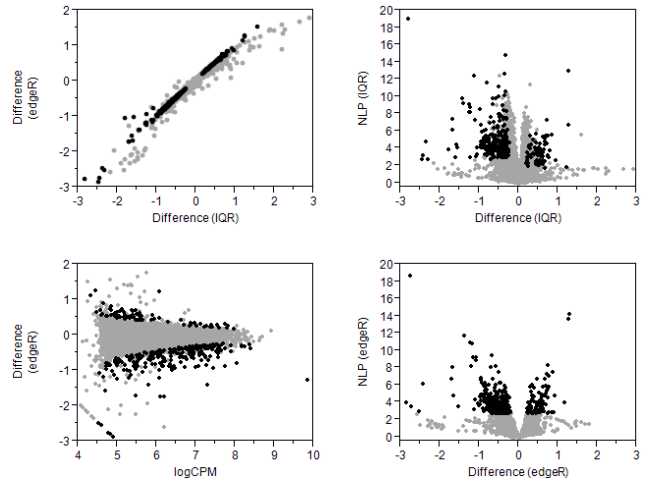
$$M = \log_2(R/G) = \log_2(R) - \log_2(G)$$

$$A = \frac{1}{2}\log_2(RG) = \frac{1}{2}(\log_2(R) + \log_2(G))$$





Effect of variance estimation



Affymetrix Probe-set normalization

- MAS 5.0 is Affymetrix' weighted average of probes
- DChip (Li and Wong) is an invariant set procedure
- RMA (Irizarry) is a probe-set quantile normalization
- GCRMA also adjusts for probe GC content

- CEL files contain the raw probe intensities
- CDF files match the probe locations to probe-sets

- Popular Bioconductor package is affy

Two-step Analysis

1. Normalize the samples

   log(fluorescence) = μ + Array + Residual

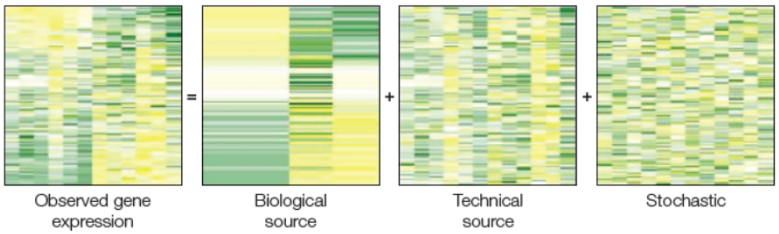
   OR variance transforms, OR supervised methods

2. For each gene, assess significance of treatment effects on the Residual (ie. relative expression level)

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

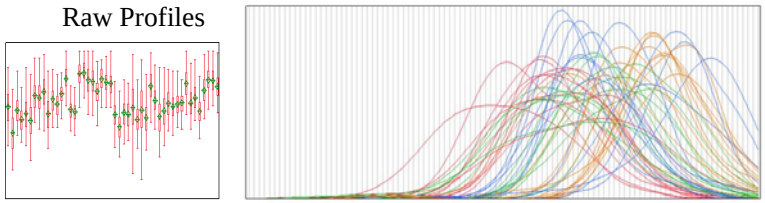
Wolfinger et al, 2001. J Comput Biol 8: 625-637

The normalization challenge

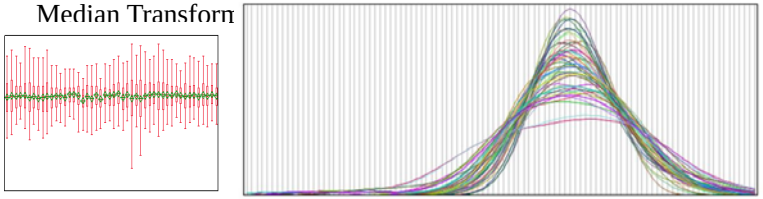


Simple normalization

Raw Profiles



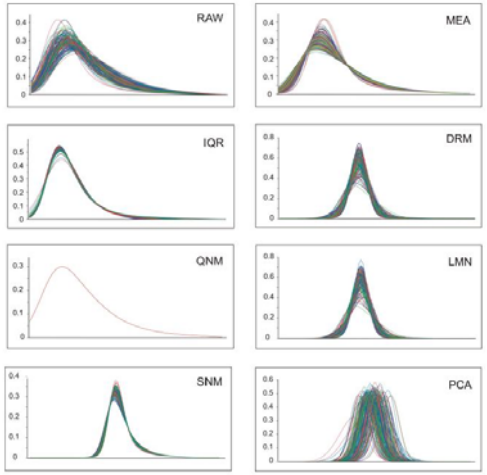
Median Transform



Types of normalization

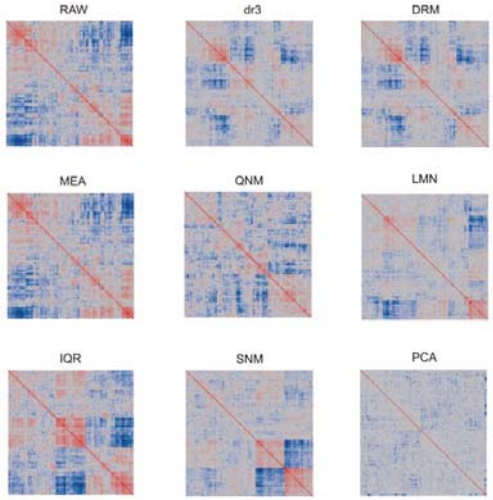
- Mean or Median transform, simply centers the distribution
  - Something like this is essential to control for overall distributional effects (eg RNA concentration)
- Variance transforms, such as standardization or inter-quartile range
  - Depends on whether you think the overall distributions should have similar variance
- Quantile normalization
  - Transforms the ranks to the average expression value for each rank
- Gene-level model fitting
  - Remove technical or biological effects before model fitting on the residuals
- Supervised normalization
  - Optimally estimate the biological effect while fitting technical factors across the entire experiment

Effect of Normalization on Distributions

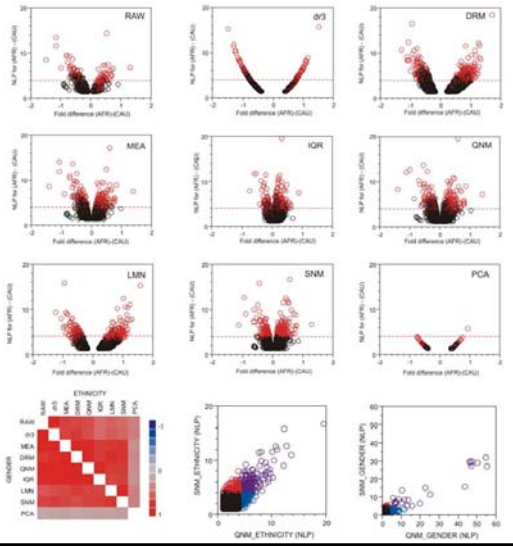


Qin et al, 2013. Frontiers in Genetics **3**: 160

Effect of Normalization on Covariance



Effect of Normalization on Significance



Analytical strategy

1. Normalize the samples
2. Extract the Principal components of gene expression
3. Ask whether the major PC are correlated with technical covariates such as Batch or RNA quality; or with Biological variables of interest
4. If they are, renormalize to remove those effects
   (PEER factor normalization is a Bayesian approach to fitting Surrogate Variables;
   SVA is a linear modeling approach often performed with COMBAT;
   SNM is a supervised approach that allows you to retain Biological factors while fitting or removing technical ones)
5. As much as possible, analyze the dataset in several different ways to (i) confirm that the findings are not sensitive to your analytical choice, and (ii) gain insight into what may cause differences, eg find confounding factors