UQ-Brisbane SISG

Module 9: Gene Expression and Epigenetic Profiling

Monday February 13, 2017

"Transcriptomics Study Design"

Greg Gibson

Georgia Institute of Technology

greg.gibson@biology.gatech.com

Content of the Course

Monday Morning	1a.	Experimental Design (GG)
	1b.	RNASeq (JP)
Monday Afternoon	2a.	Data Normalization (GG)
	2b.	Hypothesis testing (JP)
Tuesday Morning	3a.	Epigenomics (GG)
	3b.	eQTL analysis (JP)
Tuesday Afternoon	4a.	Colocalization and Immunogenomics (GG)
	4b.	Single Cell RNASeq (JP)

Content of the Lecture

Fundamentals of Gene Expression Profiling

Accessing and Depositing data in GEO or ArrayExpress

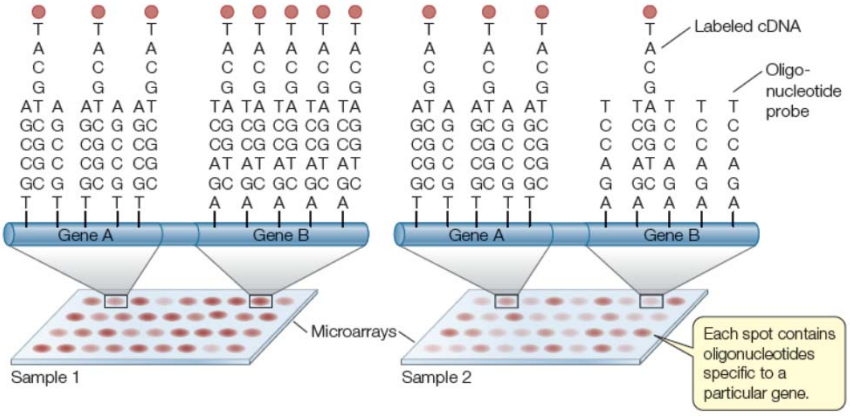
Fundamentals of Hypothesis Testing

Volcano Plots and False Discovery Rates

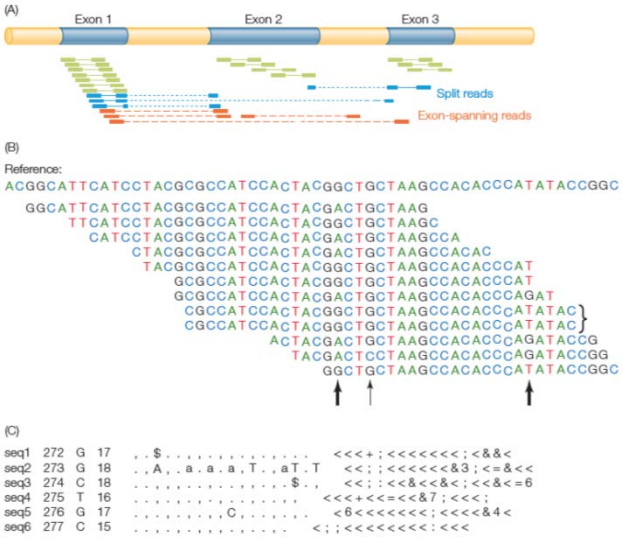
Applications of Gene Expression Analysis

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

Microarrays: The Hybridization principle

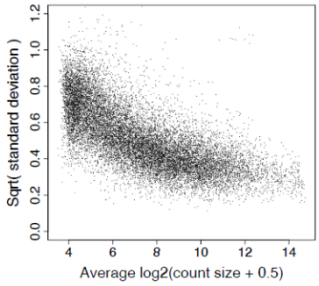


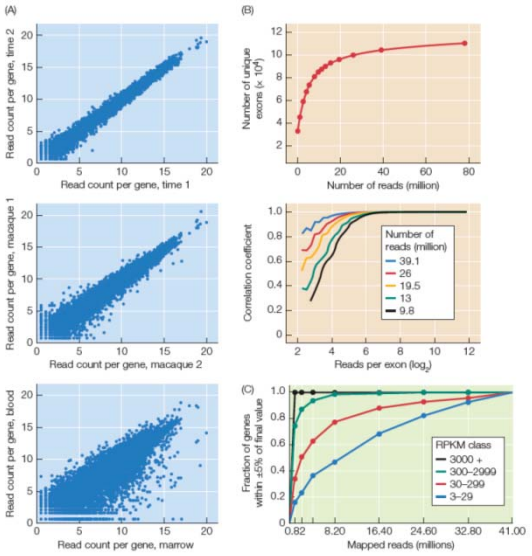
RNASeq: Digital counting of aligned reads



3 Additional Analytical Concerns with RNASeq

1. Low abundance transcripts have high variability
   (the distribution approximates a negative binomial)
2. Abundance estimates may be a function of read depth
3. Use of Counts per Million (cpm) requires adjustment for over-representation of high abundance transcripts
   (eg using TMM in EdgeR)





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

Design Biases

Molecular biologists tend not to be used to thinking about things like:
- statistical power: how many replicates to I need to see an effect?
- batch effects: there are lots of ways to obscure what you want to see
- cost: each sample typically costs $500 or more

think about these things before you start the project!!

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

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.

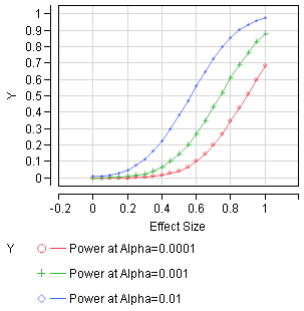
Technical Replication:
- RNA preparation (eg. from adjacent biopsies)
- cDNA synthesis and labeling (pooling minimizes outlier effects)
- array hybridization (with commercial arrays, quality generally very high)
- duplicate probes for the same gene

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

Statistical Power

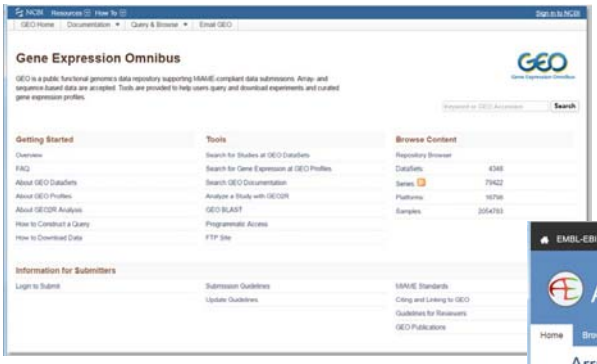


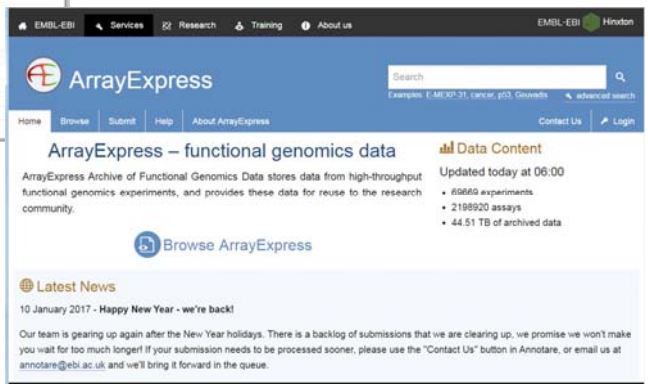
Power is the concept of how often you expect to detect an effect at a certain significance level, given a number of samples. It is a function of:

- 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.

GEO and ArrayExpress





A GEO record

https://www.ncbi.nlm.nih.gov/geo/browse/

Description of the study

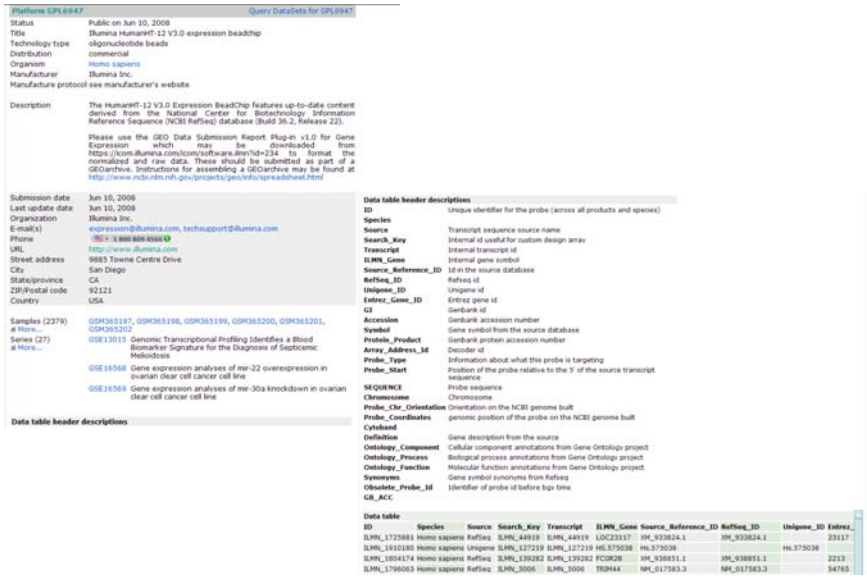
Citation

Address when uploaded

List of samples

Downloadable files

A GEO platform



A GEO sample

Sample GSM426853 Query DataSets for GSM426853

Status	Public on Dec 01, 2009
Title	A09M2
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	A09M
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**.