



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)

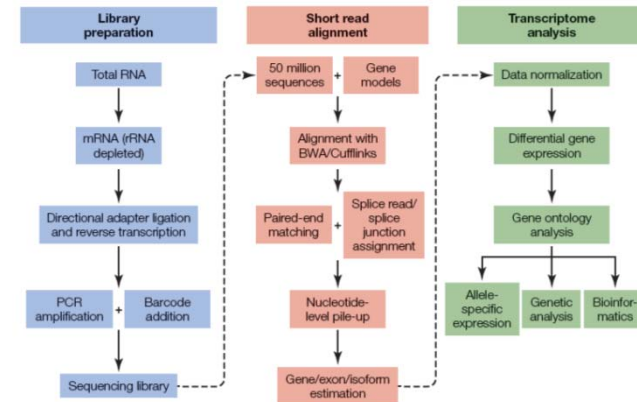
- 4a. Network Analysis (MI)
- 4b. Lab session - WGCNA (MI)

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

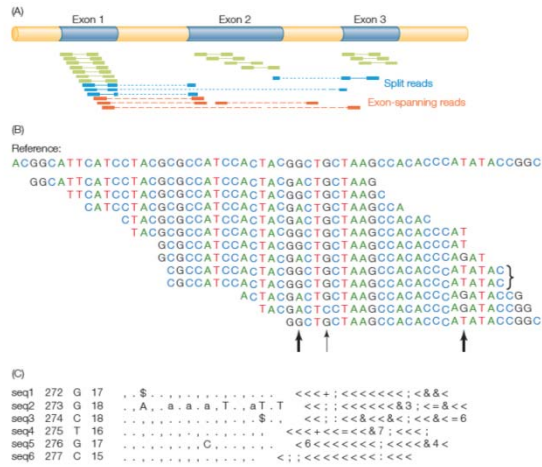
Applications of Expression Analysis

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

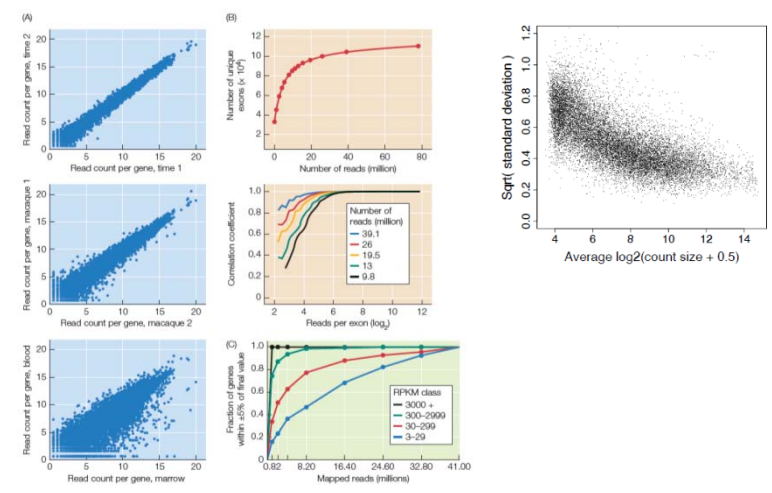
RNASeq workflow



Read alignment



Variability and Depth



Microarray vs RNASeq

Advantages of Microarrays

- Less expensive
- Better sensitivity for low abundance
- Computationally simpler
- Better-defined statistical properties
- Perfectly good for most applications

Disadvantages of Microarrays

- Only for humans, model organisms
- Different platforms give different results
- Large technical batch effects
- Sensitivity to polymorphism
- Low consistency of analytics among groups

Advantages of RNA-Seq

- Disruptive technology
- Unbiased by prior gene knowledge
- Alternative exon usage
- Allele specific expression (ASE)
- High repeatability

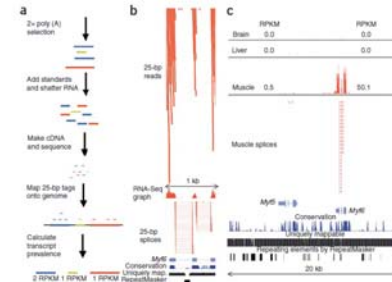
Disadvantages of RNA-Seq

- More opportunity to screw up the analysis
- Oversold resolution of exon level and ASE
- Short read alignment biases
- Sensitivity to polymorphism
- Low consistency of analytics among groups

RNA Sequencing

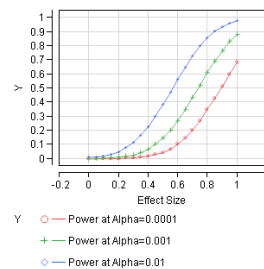
Platforms: Illumina HiSeq / LifeTech ProtonTorrent / ABI SOLiD

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 or 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

The image displays two screenshots of biological data repositories. The left screenshot shows the NCBI GEO (Gene Expression Omnibus) interface, featuring a search bar, filters for 'Data type' (e.g., Gene expression, miRNA expression) and 'Platform' (e.g., Affymetrix, Illumina), and a 'Submit' button. The right screenshot shows the ArrayExpress interface, which includes a search bar, filters for 'Experiment type' and 'Platform', and a 'Submit' button. Both interfaces provide detailed information about the data being submitted or searched, including sample IDs, accession numbers, and associated metadata.

A GEO record

2: GSE17065 record: Geographical Genomics of Human Leukocytes Gene Expression Variation [[Links](#)]
Homo sapiens]

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.

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

Platform GPL6947 Query Database for GPL6947

Status: Public on Jun 10, 2008
 Title: Illumina HumanHT-12 V3.0 expression beadchip
 Technology type: oligonucleotide beads
 Distribution: commercial
 Organism: Homo sapiens
 Manufacturer: Illumina Inc.
 Manufacturing protocol: see manufacturer's website

Description: The HumanHT-12 V3.0 Expression BeadChip features up-to-date content derived from the National Center for Biotechnology Information Reference Sequence (NCBI RefSeq) database (Build 36.2, Release 22). Please use the GEO Data Submission Report Plug-in v1.0 for Gene Expression - which can be downloaded from https://www.illumina.com/software/illumina_downloads.html to format the readfiles and raw data. These should be submitted as part of a GEO dataset. Instructions for assembling a GEO dataset may be found at http://www.ncbi.nlm.nih.gov/geo/doc/geo_submission.html.

Submission date: Jun 10, 2008
 Last update date: Jun 10, 2008
 Organization: Illumina Inc.
 E-mail(s): expression@illumina.com, techsupport@illumina.com
 Phone: (619) 594-9900/9909
 URL: <http://www.illumina.com>
 Street address: 9500 Towne Centre Drive
 City: San Diego
 State/Province: CA
 ZIP/Postal code: 92121
 Country: USA

Data table header descriptions:

- ID: Unique identifier for the probe (across all products and species)
- Source: Transcript sequence source name
- Source_Key: Internal ID used for custom design array
- Transcript: Internal transcript ID
- SNP_Gene: Internal gene symbol
- Source_Reference_ID: ID in the source database
- RefSeq_ID: RefSeq ID
- Unigene_ID: Unigene ID
- Ensembl_ID: Ensembl gene ID
- Accession: GenBank accession number
- RefSeq: Gene symbol from the source database
- Protein_Product: GenBank protein accession number
- Gene_Accession_ID: Decipher ID
- Probe_Type: Information about what the probe is targeting
- Probe_Start: Position of the probe relative to the 5' of the source transcript sequence
- Probe_End: Probe End
- SEQUENCE: Chromosome
- Probe_Chromosome: Chromosome
- Probe_Chromosome_coordinates: Coordinates on the NCBI genome build
- Probe_Coordinates: genomic position of the probe on the NCBI genome build
- Labelled: Gene description from the source
- Ontology_Annotation: Cellular component annotations from Gene Ontology project
- Ontology_Process: Biological process annotations from Gene Ontology project
- Ontology_Function: Molecular function annotations from Gene Ontology project
- Sequence: Gene symbol sequence from RefSeq
- Checksum_Probe_ID: Identifier of probe ID before lga time
- CGI_AGC

Data table

ID	Species	Source	Source_Key	Transcript	SNP_Gene	Source_Reference_ID	RefSeq_ID	Unigene_ID	Ensembl
ILMN_177083	Homo sapiens	RefSeq	SNRPB	SNRPB	LOC21517	HS_050504.1	SNRPB	23127	23127
ILMN_193180	Homo sapiens	Unigene	SNRPB	SNRPB	SNRPB	HS_575038	SNRPB	23127	23127
ILMN_193414	Homo sapiens	RefSeq	SNRPB	SNRPB	SNRPB	HS_575038	SNRPB	23127	23127
ILMN_176952	Homo sapiens	RefSeq	SNRPB	SNRPB	SNRPB	HS_575038	SNRPB	23127	23127

A GEO sample

Sample GSM428853 [Query Statistics for GSM428853](#)

Status: Public on Dec 01, 2009
 Title: GSE4982
 Sample type: RNA

Source Name: Leukocytes, Agadir, Urban
 Organism: Homo sapiens
 Characteristics: geographic location: Agadir
 gender: urban
 tissue: peripheral blood cell from leukocytes

Extracted molecule: total RNA
 Extraction protocol: Total RNA was extracted from leukocyte samples using Ambion's Leukocyte RNeasy Lysis Reagent. Quality control was performed using Agilent's Bioanalyzer.

Label: Biotin
 Label protocol: cDNA and mRNA labeling and amplification were all performed using a single kit: Ambion's Illumina TotalSeq RNA Amplification Kit.

Hybridization protocol: Illumina's BeadChip protocol
 Scan protocol: Illumina's BeadChip protocol and BeadStudio (Gene expression Module)
 Description: GSE4982
 Data processing: Raw data was log2 transformed and median-centered using RPK Genomics (SAS, Cary NC).

Submission date: Jul 11, 2009
 Last update date: Nov 12, 2009
 Contact name: Yousef Idaghdour
 E-mail(s): idiaghd@biu.edu
 Organization name: Biu
 Department: Genetics
 Lab: Gibson
 Street address: South Gardner Hall
 City: Raleigh
 State/province: NC
 ZIP/Postal code: 27695
 Country: USA

Platform ID: GPL1047
 Series (1): [GSE17061](#) Geographical Genomics of Human Leukocytes Gene Expression Variation

Data table header descriptions

ID_REF	VALUE
ILMH_1762337	-0.270341813
ILMH_2055271	0.212533766
ILMH_1736007	-0.065211006
ILMH_2385226	-0.294733746
ILMH_1806310	-0.179540575
ILMH_1779670	-0.158079383
ILMH_2321282	-0.070913534
ILMH_1671474	0.082277201
ILMH_1772582	0.227444623
ILMH_1735698	0.000532764
ILMH_1653355	0.197030952
ILMH_1717783	-0.284910419
ILMH_1705025	0.025293968
ILMH_1814316	-0.022732656
ILMH_2359168	0.04036411
ILMH_1731507	-0.480941688
ILMH_1787689	-0.046885413
ILMH_1745607	-0.099316314
ILMH_2136495	0.043699207
ILMH_1688111	-0.218402127

Total number of rows: 48803
 Table truncated, full table size 1210 Kbytes.



Summer Institute
 In Statistical Genetics

2016

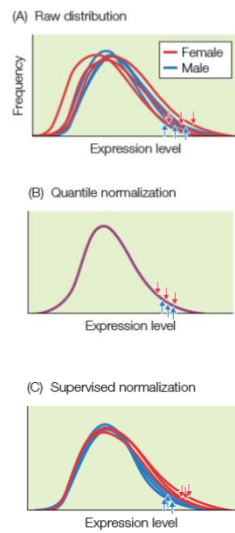
Integrative Genomics

1b. Hypothesis Testing and Normalization



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

Rank vs Absolute Expression

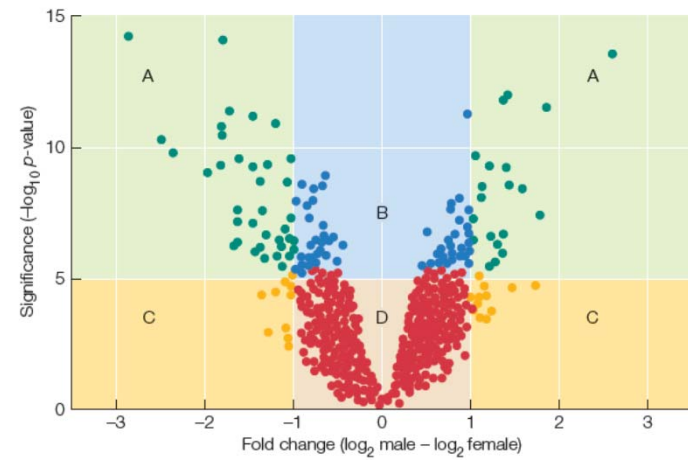


Raw data:
no effect

Variance transformed:
no effect

Mean centered:
significant effect

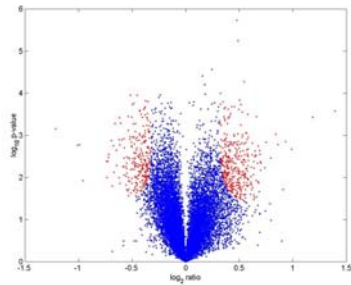
Transcriptome Volcano plots



Russ Wolfinger, Greg Gibson

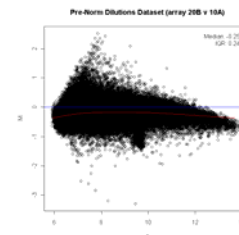
Variance estimators

- 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



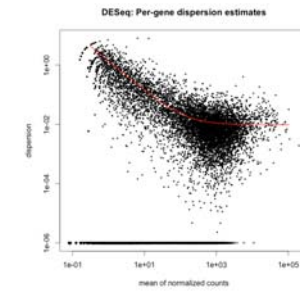
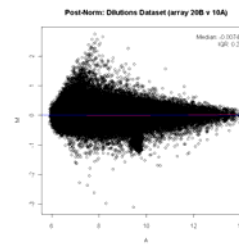
Gary Churchill, Katie Kerr

Deviation, Abundance, and Dispersion

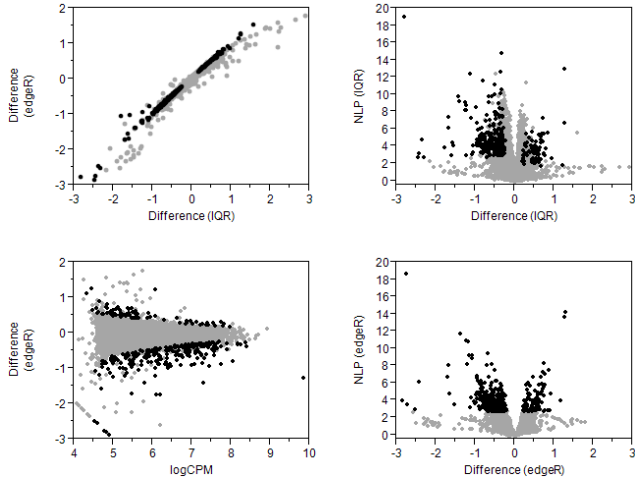


$$M = \log_2(R/G) = \log_2(R) - \log_2(G)$$

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



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

$$\log(\text{fluorescence}) = \mu + \text{Array} + \text{Residual}$$

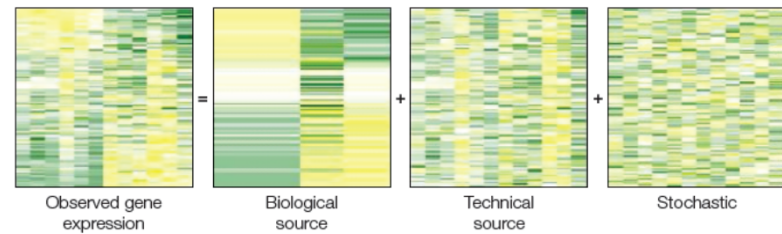
OR variance transforms, OR supervised methods

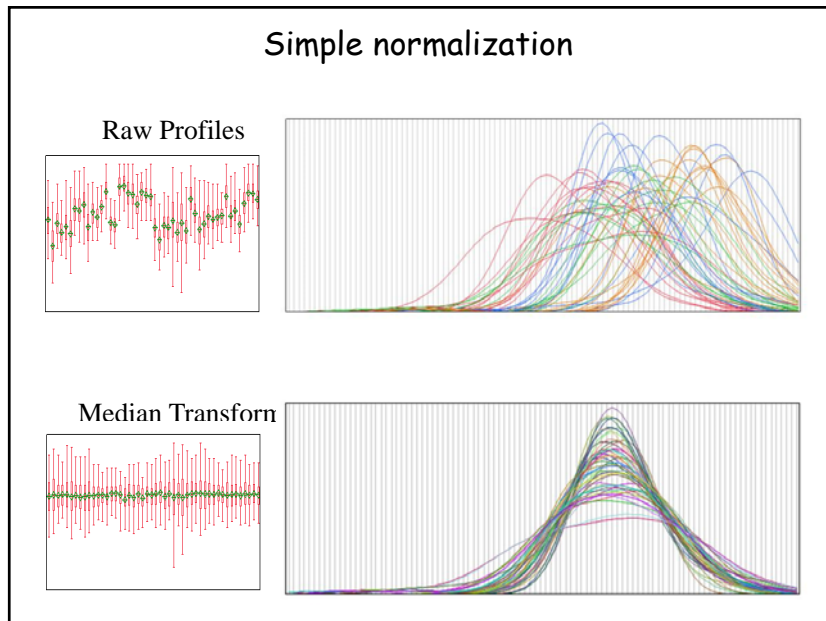
2. For each gene, assess significance of treatment effects on the Residual (ie. relative expression level)

$$\text{Residual} = \mu + \text{Sex} + \text{Geno} + \text{Treat} + \text{Interact} + \text{Error}$$

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

The normalization challenge

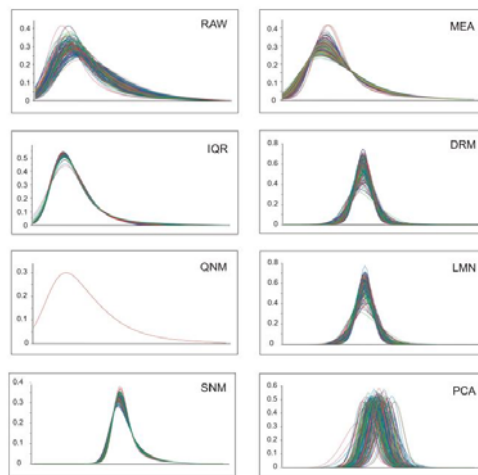




Types of normalization

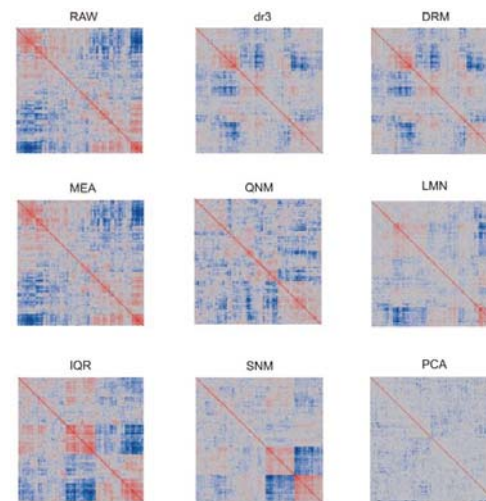
- Mean or Median transform, simply centers the distribution
 - Something like this is essential to control for overall distributional effects (eg RNA concentration)
- Variance transforms, such as standardization or inter-quartile range
 - Depends on whether you think the overall distributions should 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

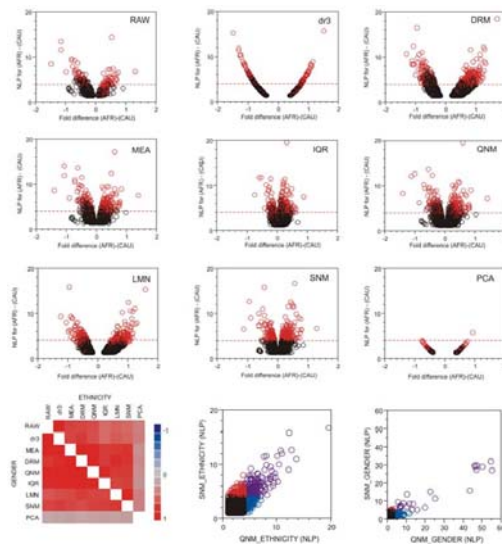


Qin et al, 2013. *Frontiers in Genetics* 3: 160

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