Transcriptomics (II)

BBMS 3009: Genome Science (First Semester, 2021)

Dr. Yuanhua Huang
School of Biomedical Sciences &
Department of Statistics and Actuarial Science

香 港 大 學
THE UNIVERSITY OF HONG KONG

Today's learning objectives

1. Read count: bias correction & normalization
2. Single gene analysis: differentially expressed genes
3. Multiple genes analysis: gene set enrichment & pathway analysis

Reading list

1) Chapter 8 in Modern Statistics for Modern Biology: https://web.stanford.edu/class/bios221/book/Chap-CountData.html
2) A survey of best practices for RNA-seq data analysis, Genome Biology, 2016

Quantification and read counts

- Direct counting
  - Gene level or exon level
  - Aligning reads to genome reference

- Isoform quantification: assigning ambiguous reads
  - Transcript / splicing isoform level
  - Maximum likelihood assignment of the reads (e.g., Kallisto, Salmon)
  - Bayesian modelling; the whole posterior distribution (e.g., BitSeq, MISO)

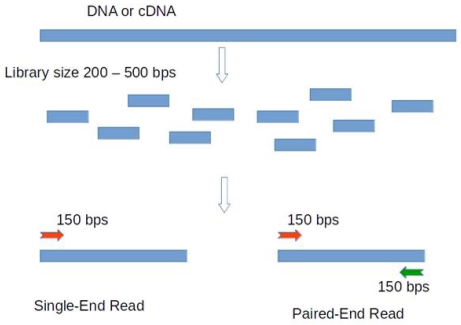
Raw read counts

Bias correction and normalization are needed

	untreated1	untreated2	untreated3	untreated4	treated1	treated2	treated3
FBgn0020369	3387	4295	1315	1853	4884	2133	2165
FBgn0020370	3186	4305	1824	2094	3525	1973	2120
FBgn0020371	1	0	1	1	1	0	0
FBgn0020372	38	84	29	28	63	28	27

Data set on Drosophila: https://bioconductor.org/packages/release/data/experiment/html/pasilla.html

Gene expression metrics



- Commonly used metrics
  - Raw counts: directly from the quantification step
  - RPKM = reads per kilo-base per million
  - FPKM = fragments per kilo-base per million (paired-end)
  - TPM = Transcripts per million

$$\text{RPKM} = 10^9 * \frac{\text{Reads mapped to the transcript}}{\text{Total reads} * \text{Transcript length}}$$

$$\text{TPM} = 10^6 * \frac{\text{reads mapped to transcript/transcript length}}{\text{Sum(reads mapped to transcript/transcript length)}}$$

$$\text{TPM} = 10^6 * \frac{\text{RPKM}}{\text{Sum(RPKM)}}$$

CPM = count per million (more used in 3’ or 5’ reads in single-cell RNA-seq)

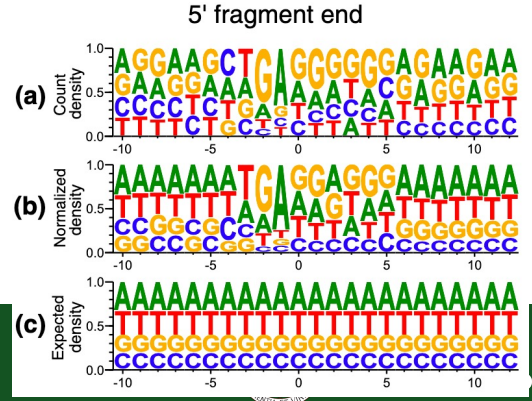
These metrics contains both **bias correction** and **normalization**

Gene expression bias correction

- For one sample, the transcriptome with N transcripts
  - raw counts vector $[c_1, c_2, \ldots, c_N]$
  - Bias corrected vector $[c_1/l_1, c_2/l_2, \ldots, c_N/l_N]$
- Bias correction on transcript length
  - Per kilo-base correction: $l_t = \frac{length\ of\ transcript\ t}{1000}$
  - Transcript level in TPM: $l_t = length\ of\ transcript\ t$
  - In RNA-Seq, longer transcripts have a higher chance of being sequenced

Different transcripts **within** sample will be comparable

- More bias corrections (less commonly used)
  - Position bias, sequence bias
  - GC content bias correction
  - $l_t = length\ of\ transcript\ t \times other\ bias\ b_t$



Roberts, Adam, et al. Genome biology 2011
Love, Hogenesch, & Irizarry. Nature Biotech, 2016

Gene expression normalization (re-scaling)

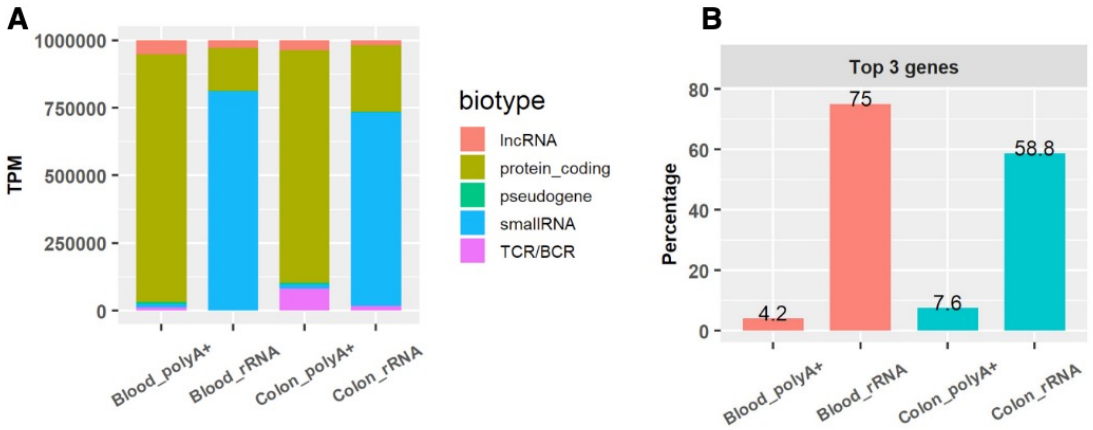
- Normalization (confusing term):
  - Simply re-scaling to a common scale
  - Count vector in sample 1: $[c_{11}, c_{21}, \ldots, c_{N_1}]$
  - Count vector in sample 2: $[c_{12}, c_{22}, \ldots, c_{N2}]$

Same transcripts **across** samples will be comparable

- Normalization over library size
  - Per million reads: $a_s = 1000000 \,/\, \sum_{t=1}^{N} c_{t,s}$ for sample $s \in \{1,2\}$
  - $[c_{11}, c_{21}, \ldots, c_{N1}] * a_1$ vs $[c_{12}, c_{22}, \ldots, c_{N2}] * a_2$
  - In RNA-Seq, each sample has different number of reads (called library size)

Gene expression normalization (re-scaling)

- Be careful
  - The metric means proportional measure
  - It is based on assumption: no major compositional change between samples

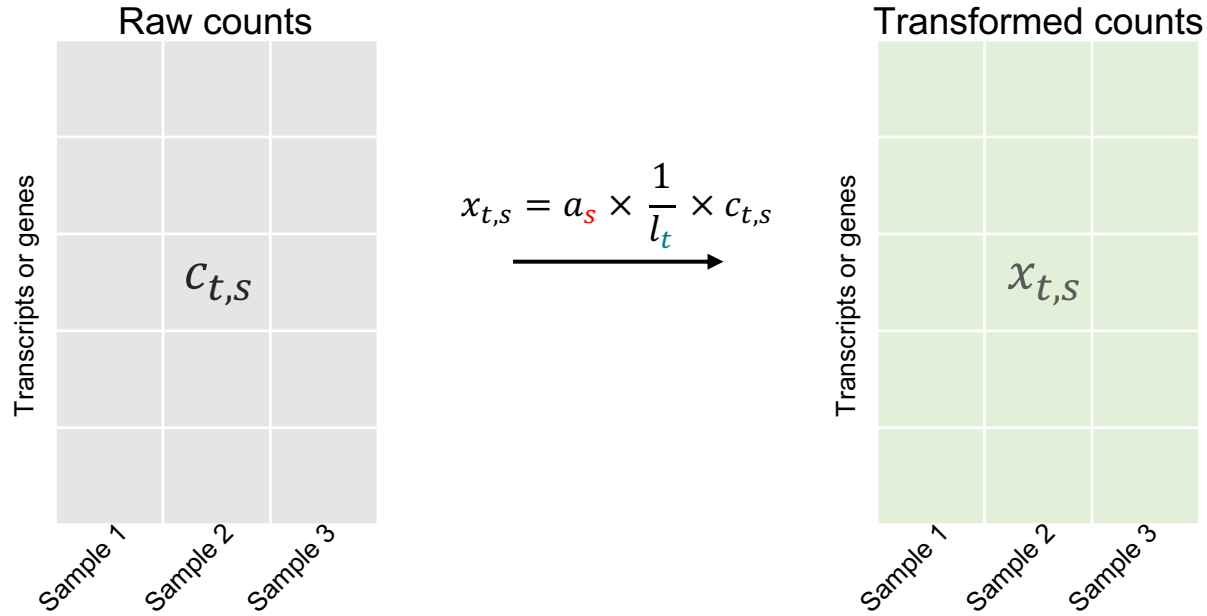


This is a negative example: composition changes substantially

Zhao, et al. RNA, 2020

Bias correction vs normalization

- General use
  - Bias correction at each transcript t: $\frac{1}{l_t}$ -- transcript length, GC, sequence
  - Normalization at each sample s: a_s -- library size, other sample level factors



Examples

- You want to compare the splicing efficiency between genes and find the regulatory patterns
  - Bias correction? Normalization?
- Now you have multiple time points (i.e., multiple samples)
  - Bias correction? Normalization?

- You want to find differentially expressed genes between a tumor sample and the adjacent normal sample
  - Bias correction? Normalization?

- Note, when people talk normalization, they may refer to both bias correction and normalization, as this term is already confusing.

Today's learning objectives

1. Read count: bias correction & normalization
    1) Read count on gene or transcript level (ambiguous reads)
    2) Bias correction: transcript length, GC content, sequence bias, etc
    3) Normalization: library size

2. Single gene analysis: differentially expressed genes

3. Multiple genes analysis: gene set enrichment & pathway analysis

Raw count data

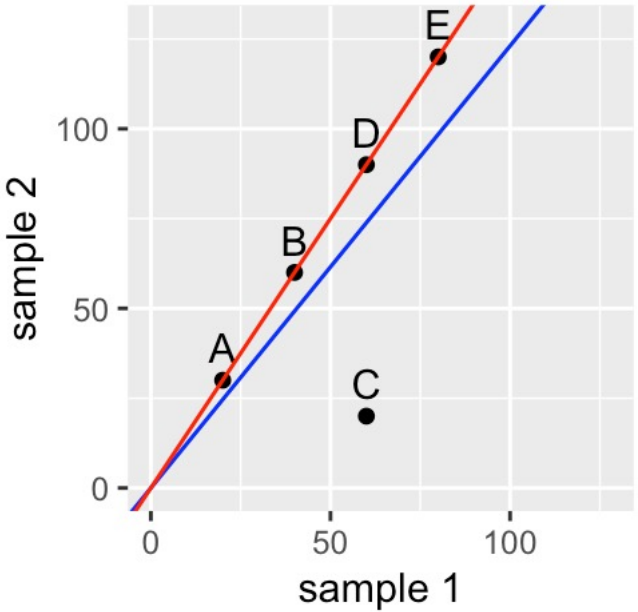
- Aim: detect differentially expressed genes between untreated samples (4 replicates) and treated samples (3 replicates)
- Normalization is necessary, but often contained in the software package, e.g., DESeq and edgeR, which prefer raw counts as input
- Some Gaussian based model may prefer logarithm transformation but be careful with the 0s. Also Log(x+1) is not always appropriate for small values

	untreated1	untreated2	untreated3	untreated4	treated1	treated2	treated3
FBgn0020369	3387	4295	1315	1853	4884	2133	2165
FBgn0020370	3186	4305	1824	2094	3525	1973	2120
FBgn0020371	1	0	1	1	1	0	0
FBgn0020372	38	84	29	28	63	28	27

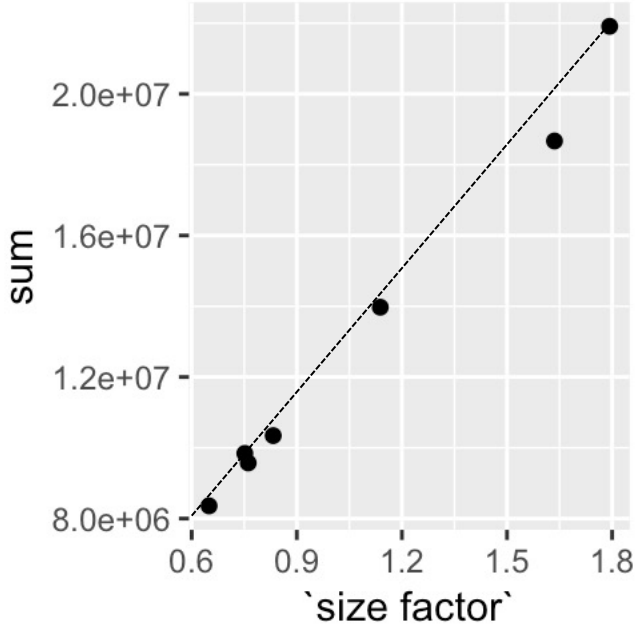
Data set on Drosophila: https://bioconductor.org/packages/release/data/experiment/html/pasilla.html

Learned scaling factor (normalization)

- DESeq learns scaling factor, i.e., library size factor
- Same purpose as normalization; so only need normalization or size scaling



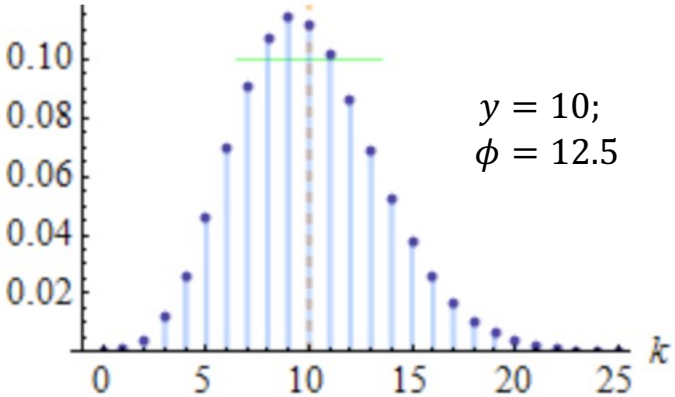
Size factor: library size vs regression based



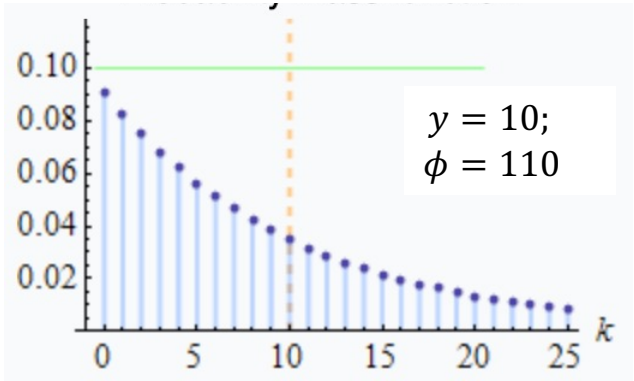
Size factor learned for the 7 samples

Hypothesis test – generalized linear model

- Detection of differentially expressed genes: hypothesis test
- Generalized linear model with negative binomial distribution (raw count)
  - $P(c_{t,s}) = NB(c_{t,s}|y_{t,s}, \phi_t)$ -- Negative Binomial likelihood: mean $y_{t,s}$, variance ϕ_t
  - $\mathbb{E}(c_{t,s}) = y_{t,s} = \beta_0 + \beta_1 x_{1,s} + \beta_2 x_{2,s}$ -- generalized linear model



Negative binomial distribution:
mean; standard deviation



Higher variance

mean	Size factor	treated
x_0	x_1	x_2
1	0.6	0
1	1.3	0
1	0.9	1
1	1.1	1

Negative Binomial another parameterization: with success rate p, the number of trials K needed to reach r success. $K \sim NB(r, p)$

Hypothesis test – generalized linear model

- Detection of differentially expressed genes: hypothesis testing
- Option 1: Wald test
  - Only fit alternative model and estimate mean and standard error of β_2
  - Wald test statistic: z = mean / std err
  - Under the null (standard normal distribution): $z \sim \mathcal{N}(0,1)$

- Option 2: Likelihood ratio test
  - Fit **both** alternative and null models
  - Calculate both likelihoods L_0 and L_1
  - Likelihood ratio test statistic: $r = 2 * \log(L_1 / L_0)$
  - Under the null (chi-square distribution) : $r \sim \chi^2(df = 1)$

Alternative model (Likelihood L_1) $\boldsymbol{y = \beta_0 + x_1\beta_1 + x_2\beta_2}$

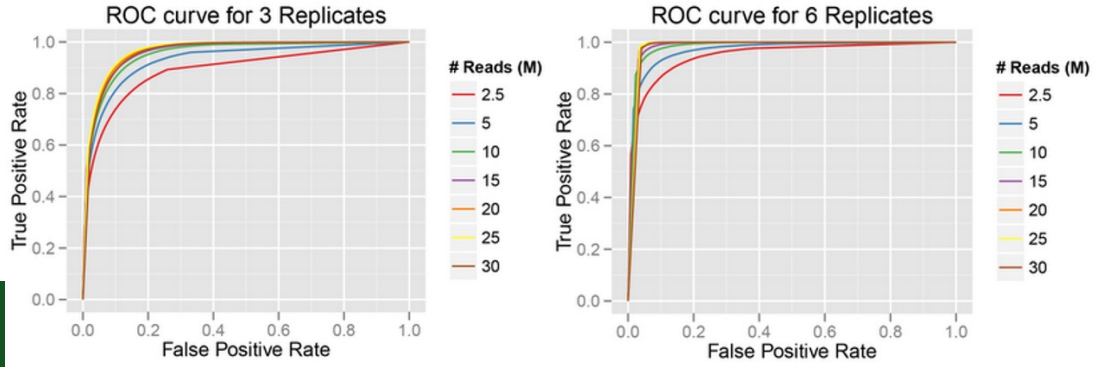
Null model (Likelihood L_0) $\boldsymbol{y = \beta_0 + x_1\beta_1}$

mean	Size factor	treated
x_0	x_1	x_2
1	0.6	0
1	1.3	0
1	0.9	1
1	1.1	1

Experiment design: coverages vs replicates

- Sequencing depths (vs number of samples)
  - Read length: 75bp, 100bp, 150bp, (possibly) 250 bp
  - Rough pricing: 2x150 bp & 300 million reads: 2,500 USD
  - Balance between number of samples and depths
    - 2x150 bp: 100 million x 3 samples
    - 2x150 bp: 25 millions x 12 samples
    - 1x150 bp: 50 millions x 12 samples

- Now think about how it affects the estimation of dispersion?



Liu, Zhou & White, Bioinfo, 2014

Differentially expressed gene analysis

- DESeq2 (DESeq normalization method)
  - http://bioconductor.org/packages/release/bioc/html/DESeq.html
- edgeR (TMM normalization method)
  - http://www.bioconductor.org/packages/release/bioc/html/edgeR.html
- limma (voom normalization method)
  - http://bioconductor.org/packages/release/bioc/html/limma.html

- Cuffdiff
  - Not too accurate.
  - It can only accept .bam files.

Today's learning objectives

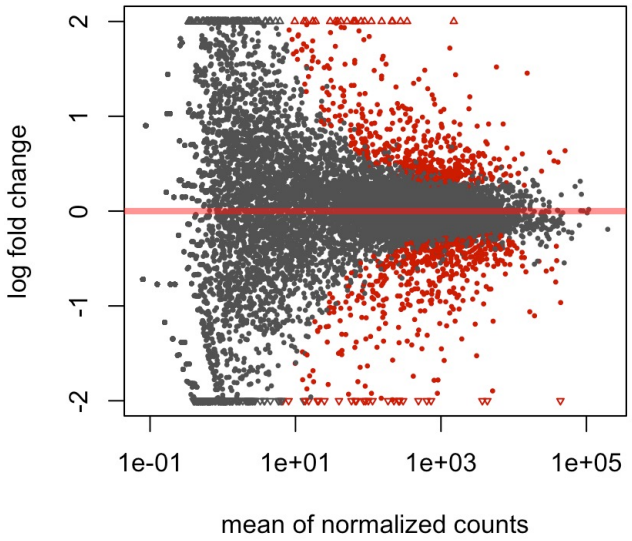
1. Read count: bias correction & normalization

2. Single gene analysis: differentially expressed genes
    1) Learning library scaling factor
    2) Likelihood ratio test; generalized linear model
    3) Estimate dispersion by sharing between genes

3. Multiple genes analysis: gene set enrichment & pathway analysis

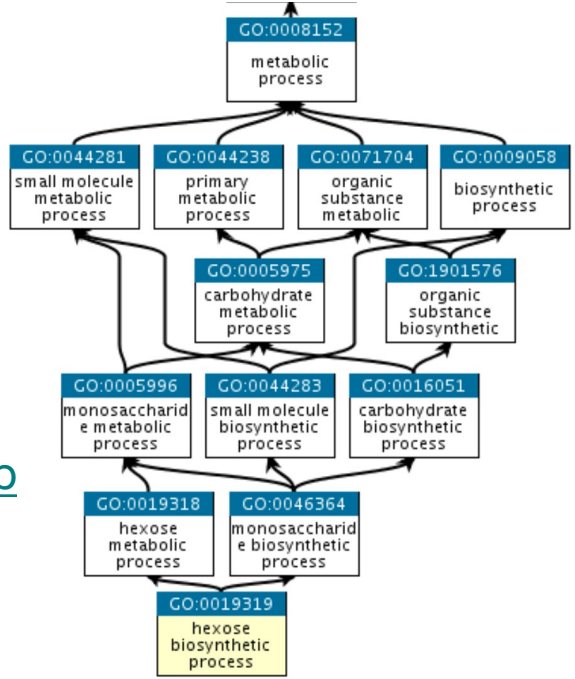
DE Genes

- Let say, 400 differentially expressed genes are detected between treated and untreated conditions
- Option 1: examine these genes individually
- Option 2: examine them jointly, e.g., in key pathways or cellular processes



Gene ontology enrichment

- Gene ontology: http://geneontology.org
  - Molecular Function
  - Cellular Component
  - Biological Process
- Each GO term contains a gene set
- Test if the DE genes enrich in any of the GO terms
  - Important to choose background gene list
  - http://pantherdb.org/webservices/go/overrep.jsp
  - https://david.ncifcrf.gov



Other annotated gene sets

- KEGG pathway annotation
  - https://www.genome.jp/kegg/
  - https://david.ncifcrf.gov
- Hallmark gene set annotation (molecular signature)
  - https://www.gsea-msigdb.org/gsea/msigdb/collections.jsp

HALLMARK_ADIPOGENESIS
HALLMARK_ALLOGRAFT_REJECTION
HALLMARK_ANDROGEN_RESPONSE
HALLMARK_ANGIOGENESIS
HALLMARK_APICAL_JUNCTION
HALLMARK_APICAL_SURFACE
HALLMARK_APOPTOSIS
HALLMARK_BILE_ACID_METABOLISM
HALLMARK_CHOLESTEROL_HOMEOSTASIS
HALLMARK_COAGULATION
HALLMARK_COMPLEMENT
HALLMARK_DNA_REPAIR
HALLMARK_E2F_TARGETS
HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION
HALLMARK_ESTROGEN_RESPONSE_EARLY
HALLMARK_ESTROGEN_RESPONSE_LATE
HALLMARK_FATTY_ACID_METABOLISM
HALLMARK_G2M_CHECKPOINT
HALLMARK_GLYCOLYSIS
HALLMARK_HEDGEHOG_SIGNALING
HALLMARK_HEME_METABOLISM
HALLMARK_HYPOXIA
HALLMARK_IL2_STAT5_SIGNALING
HALLMARK_IL6_JAK_STAT3_SIGNALING
HALLMARK_INFLAMMATORY_RESPONSE
HALLMARK_INTERFERON_ALPHA_RESPONSE
HALLMARK_INTERFERON_GAMMA_RESPONSE
HALLMARK_KRAS_SIGNALING_DN
HALLMARK_KRAS_SIGNALING_UP
HALLMARK_MITOTIC_SPINDLE
HALLMARK_MTORC1_SIGNALING
HALLMARK_MYC_TARGETS_V1
HALLMARK_MYC_TARGETS_V2
HALLMARK_MYOGENESIS
HALLMARK_NOTCH_SIGNALING
HALLMARK_OXIDATIVE_PHOSPHORYLATION
HALLMARK_P53_PATHWAY
HALLMARK_PANCREAS_BETA_CELLS
HALLMARK_PEROXISOME
HALLMARK_PI3K_AKT_MTOR_SIGNALING
HALLMARK_PROTEIN_SECRETION
HALLMARK_REACTIVE_OXYGEN_SPECIES_PATHWAY
HALLMARK_SPERMATOGENESIS
HALLMARK_TGF_BETA_SIGNALING
HALLMARK_TNFA_SIGNALING_VIA_NFKB
HALLMARK_UNFOLDED_PROTEIN_RESPONSE
HALLMARK_UV_RESPONSE_DN
HALLMARK_UV_RESPONSE_UP
HALLMARK_WNT_BETA_CATENIN_SIGNALING
HALLMARK_XENOBIOTIC_METABOLISM

Questions

1. Read count: bias correction & normalization
2. Single gene analysis: differentially expressed genes
3. Multiple genes analysis: gene set enrichment & pathway analysis

Reading list

1) Chapter 8 in Modern Statistics for Modern Biology: https://web.stanford.edu/class/bios221/book/Chap-CountData.html
2) A survey of best practices for RNA-seq data analysis, Genome Biology, 2016