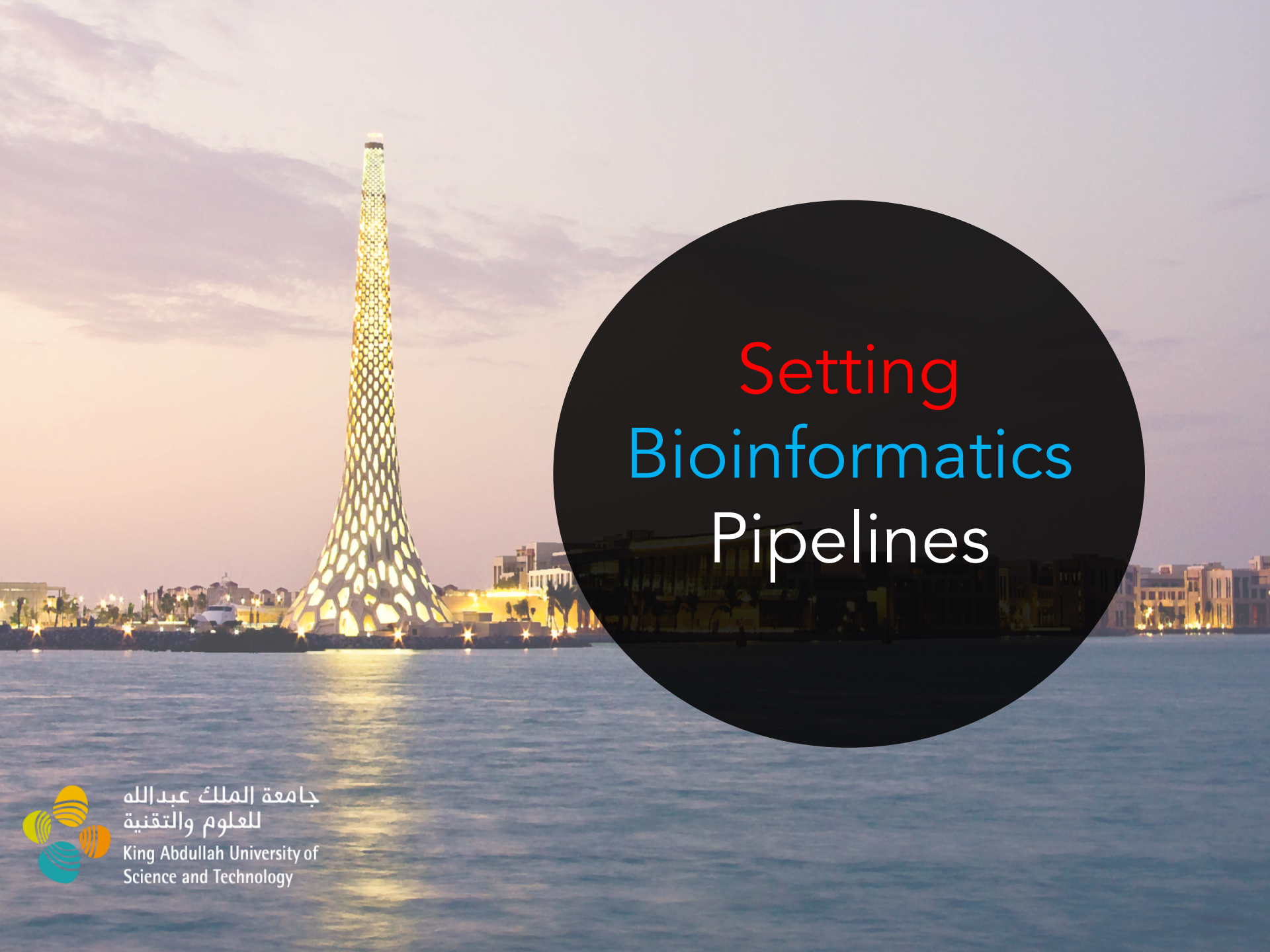

Setting
Bioinformatics
Pipelines
جامعة الملك عبدالله
للعلوم والتقنية
King Abdullah University of
Science and Technology

جامعة الملك عبدالله
للعلوم والتقنية
King Abdullah University of
Science and Technology

Setting
Bioinformatics
Pipelines
Reasoning with data

DAY 2

David Gomez-Cabrero
Jesper N Tegner
Vincenzo Lagani
Robert Lehmann

جامعة الملك عبدالله
للعلوم والتقنية
King Abdullah University of
Science and Technology

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

Tentative Course Schedule

(Time, topic/emphasis & resources)

Week	Lectures	Topic
1	Tue 01/23/2024	Pipeline analysis: what is it and general guidelines for mastering pipelines. Investigating our first pipeline: RNA-seq. RNA-seq hands-on.
2	Tue 01/30/2024	Setting a pipeline: code, Github, tracing back analysis, Reimplement the RNA-seq pipeline over the new concepts. Review the statistics behind count data I.
3	Tue 02/06/2024	ATAC-seq. Understanding and implementing the pipeline Review the statistics behind count data II.
4	Tue 02/13/2024	Single-cell RNA-seq. Understanding and implementing the pipeline.
5	No schedule	No class
6	Tue 02/27/2024	Single-cell multi-ome. Understanding and implementing the pipeline.
7	Tue 03/05/2024	No class
8	Tue 03/12/2024	Setting an integrative pipeline.
9	Tue 03/19/2024	DNA Methylation. Understanding and implementing the pipeline.
10	Tue 03/26/2024	No class
11	Tue 04/02/2024	No class
12	No schedule	No class
13	Tue 04/16/2024	No class
14	Tue 04/23/2024	No class
15	Tue 04/30/2024	No class
16	Tue 05/07/2024	Per groups a single-pipeline: HiC, CITE-seq, CyToF, etc.

Lecture time:

- 8:30 - 11:50 then lunch break.
- 13:10 – 15:30
- A homework will be described.

Please join the slack channel

bese394a_setting_bioinfopipelines

Setting
Bioinformatics
Pipelines

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

Setting a pipeline: code, Github, tracing back analysis,
Reimplement the RNA-seq pipeline over the new concepts.
Review the statistics behind count data I.

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

8:30 - 10:30 Review the activities.

10:40 - 11:50 Experimental Design and Statistics.

break

13:10 - 13:55 ORA and GSEA

14:00 - 15:00 GeneSetCluster.

15:00 - 15:30 Introduction to Github.

Outline Day 2

Extending our first pipeline: *review* and **add**.

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

Review the exercises: statistics and tutorials.

1. Reviewed the normalization and differential expression

   Experimental design
   Limma: Normalize and set design
   Limma: Voom or Trend?
   Activity 1 -> How to compare -> Exercise
   Activity 2 -> Plan the analysis -> Exercise

Extending our first pipeline: *review* and **add**.

2. Gene Set Analysis: ORA and GSEA

   ORA and Gene Set Enrichment analysis
   Tools required
   Run ORA
   Run GSEA
   GeneSetCluster -> Exercise

Generate a report for a manuscript: Github.

3. Github

   Brief introduction
   Set a Github related to the GSE198256 analysis -> Exercise

Off-class exercises.

4. Comment additional exercises

Extending our first pipeline: *review* and **add**.

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

Review the exercises: statistics and tutorials.

1. Reviewed the normalization and differential expression

   Experimental design
   Limma: Normalize and set design
   Limma: Voom or Trend?
   Activity 1 -> How to compare -> Exercise
   Activity 2 -> Plan the analysis -> Exercise

Review the exercises: statistics and tutorials.

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

Part 1

1. **Explain TMM**
2. **Explain Voom**
3. **Explain limma Trend**
4. **Explain Wald-test in DESeq2**
5. **Explain LTR in DESeq2**

*What is the important aspect**s** to have into consideration?*

Review the exercises: statistics and tutorials.

Part 2

1. **Generate a tutorial for (beautiful) Heatmaps.**
2. **Generate a tutorial for PCA in RNA-seq data.**
3. **Generate a tutorial for ORA**
4. **Generate a tutorial for GSEA**
5. **How to select the genes to be included in the analysis based on # reads**

How do we bring this into our analysis?

Extending our first pipeline: *review* and **add**.

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

Some thoughts on statistics…

Theory behind DESeq2

The DESeq2 model

The DESeq2 model and all the steps taken in the software are described in detail in our publication (Love, Huber, and Anders 2014), and we include the formula and descriptions in this section as well. The differential expression analysis in DESeq2 uses a generalized linear model of the form:

$$K_{ij} \sim \mathrm{NB}(\mu_{ij}, \alpha_i)$$

$$\mu_{ij} = s_j q_{ij}$$

$$\log_2(q_{ij}) = x_{j\cdot}\beta_i$$

where counts K_{ij} for gene i, sample j are modeled using a negative binomial distribution with fitted mean μ_{ij} and a gene-specific dispersion parameter α_i. The fitted mean is composed of a sample-specific size factor s_j and a parameter q_{ij} proportional to the expected true concentration of fragments for sample j. The coefficients β_i give the log2 fold changes for gene i for each column of the model matrix X. Note that the model can be generalized to use sample- and gene-dependent normalization factors s_{ij}.

The dispersion parameter α_i defines the relationship between the variance of the observed count and its mean value. In other words, how far do we expected the observed count will be from the mean value, which depends both on the size factor s_j and the covariate-dependent part q_{ij} as defined above.

$$\mathrm{Var}(K_{ij}) = E[(K_{ij} - \mu_{ij})^2] = \mu_{ij} + \alpha_i \mu_{ij}^2$$

An option in DESeq2 is to provide maximum *a posteriori* estimates of the log2 fold changes in β_i after incorporating a zero-centered Normal prior (`betaPrior`). While previously, these moderated, or shrunken, estimates were generated by *DESeq* or *nbinomWaldTest* functions, they are now produced by the *lfcShrink* function. Dispersions are estimated using expected mean values from the maximum likelihood estimate of log2 fold changes, and optimizing the Cox-Reid adjusted profile likelihood, as first implemented for RNA-seq data in edgeR (Cox and Reid 1987,edgeR_GLM). The steps performed by the *DESeq* function are documented in its manual page `?DESeq`; briefly, they are:

1. estimation of size factors s_j by *estimateSizeFactors*
2. estimation of dispersion α_i by *estimateDispersions*
3. negative binomial GLM fitting for β_i and Wald statistics by *nbinomWaldTest*

For access to all the values calculated during these steps, see the section above.

Changes compared to DESeq

The main changes in the package *DESeq2*, compared to the (older) version *DESeq*, are as follows:

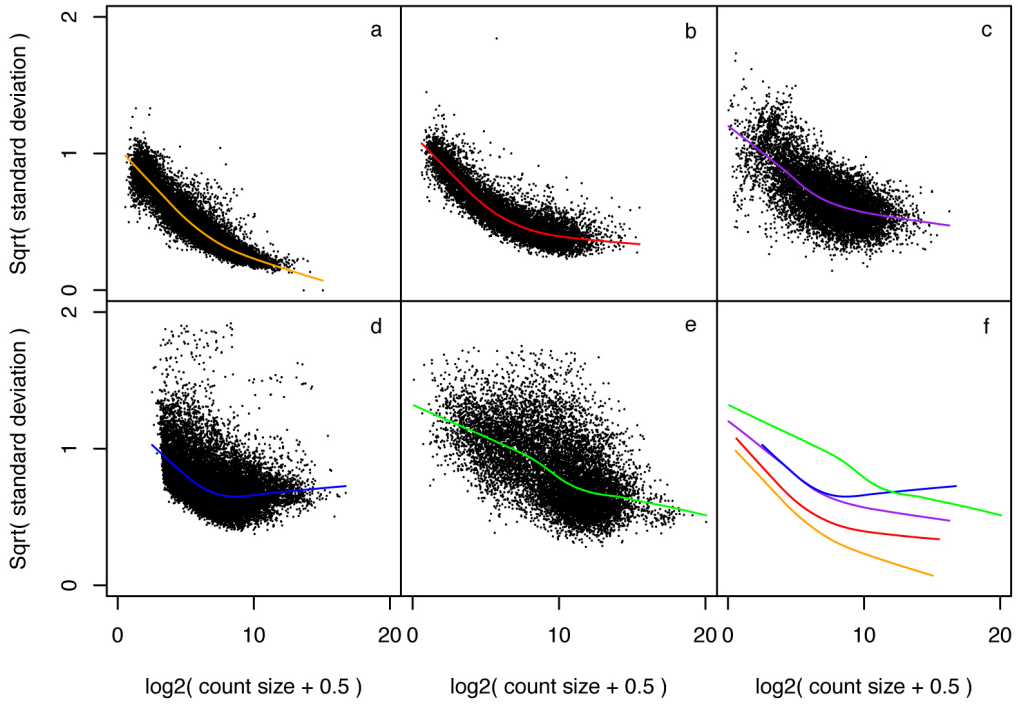
- *RangedSummarizedExperiment* is used as the superclass for storage of input data, intermediate calculations and results.
- Optional, maximum *a posteriori* estimation of GLM coefficients incorporating a zero-centered Normal prior with variance estimated from data (equivalent to Tikhonov/ridge regularization). This adjustment has little effect on genes with high counts, yet it helps to moderate the otherwise large variance in log2 fold change estimates for genes with low counts or highly variable counts. These estimates are now provided by the *lfcShrink* function.
- Maximum *a posteriori* estimation of dispersion replaces the `sharingMode` options `fit-only` or `maximum` of the previous version of the package. This is similar to the dispersion estimation methods of DSS (Wu, Wang, and Wu 2012)

https://bioconductor.org/packages/devel/bioc/vignettes/DESeq2/inst/doc/DESeq2.html#theory

Extending our first pipeline: *review* and **add**.

Some thoughts on statistics…

A simple approach to analyzing RNA-seq data would be input the log-cpm values into a well established microarray analysis pipeline such as that provided by the limma software package [3, 9]. This would be expected to behave well if the counts were all reasonably large, but it ignores the mean-variance trend for lower counts.



Extending our first pipeline: *review* and **add**.

Some thoughts on statistics…

A simple approach to analyzing RNA-seq data would be input the log-cpm values into a well established **microarray analysis pipeline such as that provided by the limma software** package [3, 9]. This would be expected to behave well if the counts were all reasonably large, but it ignores the **mean-variance** trend for lower counts.

Extending our first pipeline: *review* and **add**.

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

Some thoughts on statistics…

Limma for microarrays

requires one or two matrices to be specified.
design matrix + **contrast matrix**

You have to start by **fitting a linear model to your data which fully models the systematic part of your data**. The model is specified by the **design matrix**. Each row of the design matrix corresponds to an array in your experiment and each column corresponds to a coefficient that is used to describe the RNA sources in your experiment.

The **main purpose of this step is to estimate the variability in the data**, hence the systematic part needs to be modeled so it can be distinguished from random variation.

Extending our first pipeline: *review* and **add**.

Some thoughts on statistics…

Limma for microarrays

requires one or two matrices to be specified.
design matrix + **contrast matrix**

```
design <- model.matrix(~ 0+factor(c(1,1,1,2,2,3,3,3)))
colnames(design) <- c("group1", "group2", "group3")

contrast.matrix <- makeContrasts(group2-group1, group3-group2,
                                 group3-group1, levels=design)
```

Extending our first pipeline: *review* and **add**.

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

Some thoughts on statistics…

Limma for microarrays

requires one or two matrices to be specified.
design matrix + **contrast matrix**

Differential expression?

The **basic statistic** used for significance analysis is the **moderated t-statistic**, which is computed for each probe and for each contrast. This has the same interpretation as an **ordinary t-statistic except that the standard errors have been moderated across genes**, i.e., squeezed towards a common value, using a simple Bayesian model. This has the effect of **borrowing information from the ensemble of genes to aid with inference about each individual gene** [35, 21].

The use of global parameters is a simple means of sharing information between genes that can be used even for the smallest experiments, because the global parameters can be estimated from the entire data set involving all the genes at once

The estimated variance for each gene then becomes a compromise between the gene-wise estimator, obtained from the data for that gene alone, and the global variability across all genes, estimated by pooling the ensemble of all genes. This has the effect of increasing the effective degrees of freedom with which the gene-wise variances are estimated. It was an innovation of the *limma* package to show that exact small-sample inference could be conducted using the empirical Bayes posterior variance estimators (16).

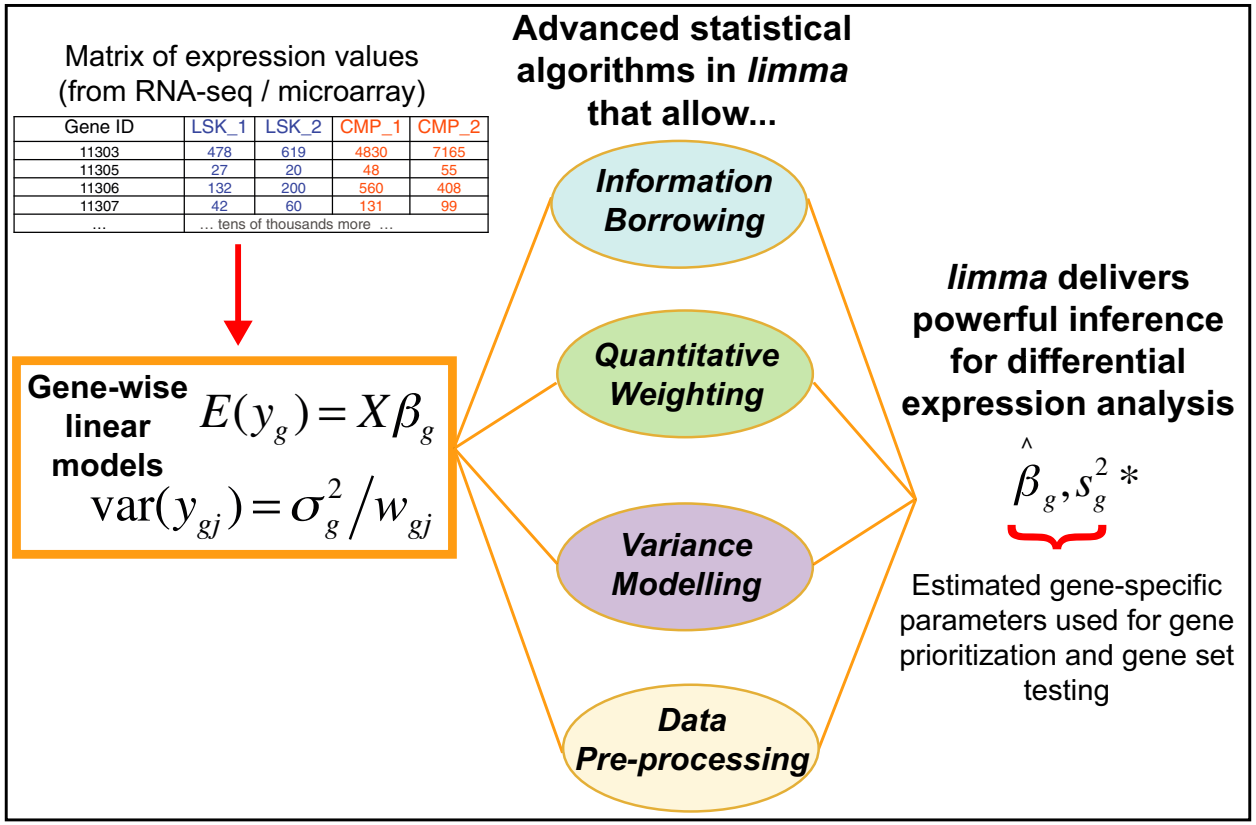
Extending our first pipeline: *review* and **add**.

Some thoughts on statistics…

Limma for microarrays

requires one or two matrices to be specified.
design matrix + **contrast matrix**

Differential expression?



Extending our first pipeline: *review* and **add**.

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

Review the exercises: statistics and tutorials.

1. Reviewed the normalization and differential expression

   Experimental design
   Limma: Normalize and set design
   Limma: Voom or Trend?
   Activity 1 -> How to compare -> Exercise
   Activity 2 -> Plan the analysis -> Exercise

Extending our first pipeline: *review* and **add**.

2. Gene Set Analysis: ORA and GSEA

   ORA and Gene Set Enrichment analysis
   Tools required
   Run ORA
   Run GSEA
   GeneSetCluster -> Exercise

Extending our first pipeline: *review* and **add**.

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

Review the exercises: statistics and tutorials.

1. Reviewed the normalization and differential expression

   Experimental design
   Limma: Normalize and set design
   Limma: Voom or Trend?
   Activity 1 -> How to compare -> Exercise
   Activity 2 -> Plan the analysis -> Exercise

Extending our first pipeline: *review* and **add**.

2. Gene Set Analysis: ORA and GSEA

   ORA and Gene Set Enrichment analysis
   Tools required
   Run ORA
   Run GSEA
   GeneSetCluster -> Exercise

Generate a report for a manuscript: Github.

3. Github

   Brief introduction
   Set a Github related to the GSE198256 analysis -> Exercise

Extending our first pipeline: *review* and **add**.

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

Review the exercises: statistics and tutorials.

1. Reviewed the normalization and differential expression

   Experimental design
   Limma: Normalize and set design
   Limma: Voom or Trend?
   Activity 1 -> How to compare -> Exercise
   Activity 2 -> Plan the analysis -> Exercise

Extending our first pipeline: *review* and **add**.

2. Gene Set Analysis: ORA and GSEA

   ORA and Gene Set Enrichment analysis
   Tools required
   Run ORA -> Exercises
   Run GSEA -> Exercises
   GeneSetCluster -> Exercises

Generate a report for a manuscript: Github.

3. Github

   Brief introduction
   Set a Github related to the GSE198256 analysis -> Exercise
   Update past GSE198256 based on the new elements today and upload to GitHub -> Exercise
   Your RMarkDown of (1) and (2) upload to GitHub -> Exercise

Off-class exercises.

4. Comment additional exercises

EXERCISES PER GROUP DAY 2

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

1. **Explain ATAC-seq wet-lab + data.**
2. **Explain MACS2**
3. **Explain Footprints**
4. **Explain Motifs: JASPAR DB.**
5. **Explain TF Gene Regulatory Networks**

MUST READ FOR ALL

https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-1929-3

https://genomebiology.biomedcentral.com/articles/10.1186/gb-2008-9-9-r137

https://academic.oup.com/nar/article/50/D1/D165/6446529

https://www.nature.com/articles/nature11212

Generate a report for a manuscript: Github.

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

Lets review the *thinking* process

1. **Define and understand the biological/biomedical question.**
2. **How to address the question?**
   What data? How to set an experimental design?...
3. **Data generation**
4. **Data analysis: set a plan**
5. **Validation experiments.**
6. **Manuscript. + submitting data and code**