The new normal: Lab

- Prelab and office hours via Zoom, links on the wiki
  - Instructors can also be reached via email
- Each prelab will have slides posted 1-2 hours prior to the beginning of class
- Instructors and Kevin will be available for entire class time to field questions
  - There will also be a Benchling notebook devoted to questions, especially for R
- Each prelab will be recorded and posted on the wiki for review purposes
  - I'd love to see you in video, but that is optional if you prefer privacy
- To ask or answer questions during class:
  - Use "raise hand" function
  - Can also type questions in the chat box rather than talk if preferred

1

The new normal: Homework/Quizzes

- Kevin will be checking benchling notebooks 24hrs following the beginning of lab to see your progress (i.e. Tues. class is checked at 1pm EDT on Wed.)
- Homework is due via Stellar by 10pm (EDT) the day of the lab session to be on time
- Homework will be returned via Stellar
  - See "comments" tab M1D7 and M2D2 for recent homework graded
- Quizzes will be emailed at the beginning of lab time and must be posted to Stellar by 10pm (EDT) on the same day to be on time

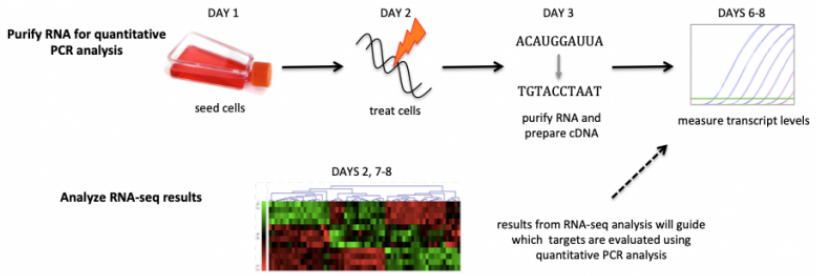
2

M2D6: Analyze RNA-seq data and prepare for qualitative PCR experiment

1. Prelab discussion
2. R.studio.cloud: clustering refresher
3. R.studio.cloud: a549 RNAseq analysis
4. Choose genes to further analyze by qPCR

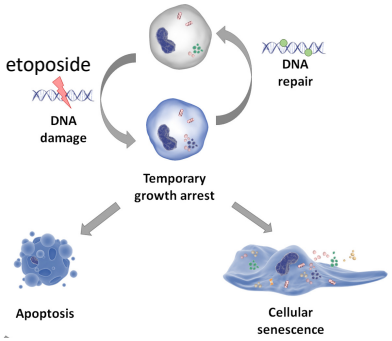
3

Mod2: Experimental overview



4

How does gene expression change upon etoposide treatment?



Soto Gamez et al. Regulation of Survival Networks in Senescent Cells: From Mechanisms to Interventions. JMB July 2019

5

Review Ex2 : RNA-seq data was pre-processed

- Data from sequencer was reads: chopped up cDNA, example "ATTAGAGAACCA"
- Reads were aligned to human genome
- Aligned reads were counted
- RPKM corrects for differences in sequencing depth and gene length

Reads per kilobase million

$$\text{RPKM} = \frac{\text{number of reads mapped to gene sequence}}{\text{gene length (kb)} \times \text{total number of reads (millions)}}$$

- The counts of the aligned RNA-seq reads were loaded to DESeqDataset

6

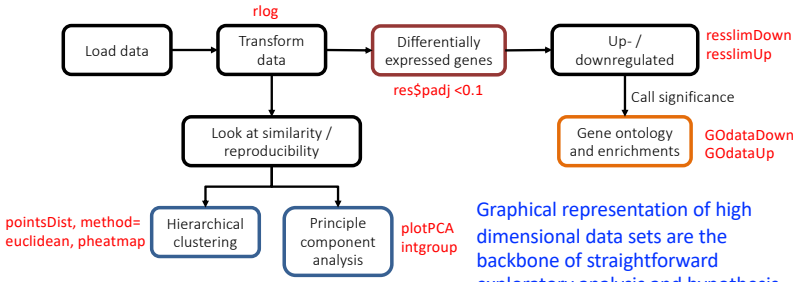
Review Ex2: DESeqDataSet structure

Reads aligned to genes were loaded into data structure called "DESeqDataSet"

- colData: samples
- rowRanges: gene info, e. g. exons
- assay: matrix of counts assigned to each gene for each sample

7

Review Ex2: Workflow RNA-seq analysis

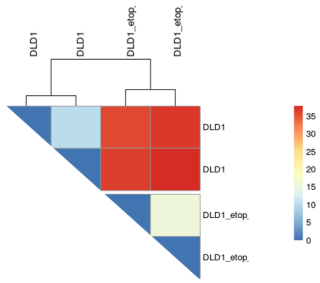


Graphical representation of high dimensional data sets are the backbone of straightforward exploratory analysis and hypothesis generation!

Image from Casper Enghuus, Sp17 TA 20.109

8

Review Ex2: Hierarchical clustering groups similar objects



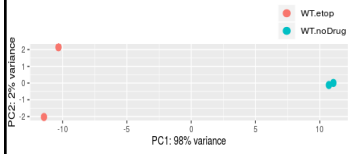
-Plot: heatmap with dendrogram
-I blocked the info below the diagonal (it's a mirror image) and copied labeled to the other axis

- Analysis end point= cluster, where clusters are distinct from
- Objects (box) in each cluster (two grouped boxes are similar
- Color of each cluster distance between compared cluster

What are the objects in this cluster? Rlog transformed counts
What is scale? Euclidean distance
What makes the biggest difference in clustering? Etoposide treatment

9

Review Ex2: Principle component analysis (PCA) shows relatedness of objects



-Plot of the samples ~20,000 genes across 4 samples
- Two replicates of DLD-1 and two replicates of DLD-1+ etoposide (color=ColData)
-PCA: multidimensional data are projected onto 2D plane such that the spread is in two directions that explains the most difference

- Analysis results in components that capture the largest possible variance
- PC1 (principle component 1) captures the most variance, here 98%
- PC2 (principle component 2) captures the second most, here 2%

What difference captures most of the variance (98%)? Etoposide treatment

10

Review Ex2: Gene ontology (GO) terms based on gene product properties

	GO.ID	Term	Annotated	Significant	Expected	Rank in classicFisher	classicFisher	classicKS
1	GO:0051301	cell division	145	16	21.52	952	0.97383	1.0e-07
2	GO:0031668	cellular response to extracellular stimu...	12	8	1.78	1	4.2e-05	0.00013
3	GO:0010389	regulation of G2/M transition of mitotic...	30	7	4.45	260	0.13535	0.00019

- GO table terms:
  - GO ID: Numerical identifiers of GO group
  - Term: GO term describes our knowledge w/i 3 aspects: molecular function, cellular component, bio. process
  - Annotated: number of genes in our gene list annotated with this term
  - Significant: number of significantly differentially expressed genes (DEGs) annotated with that term
  - Expected: under random chance, number of DEGs expected in that term
  - Classic Fisher: p value determined with Fisher's test
  - Classic KS: p value determined with Kolmogorov–Smirnov test

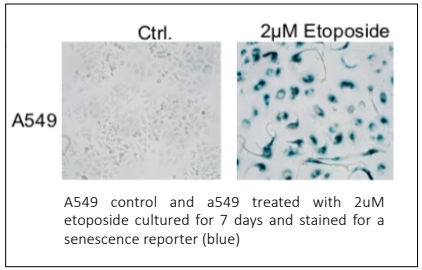
11

Clarification of P values, Classic Fisher vs. ClassicKS

- Fisher's exact test compares the expected number of significant genes at random to the observed number of significant genes to arrive at a probability. (In our scripts, this p-value is 0.1, which is given in the line geneSel = function(allScore) {return(allScore<0.1)} used to generate the topGO table.)
- The KS test compares the distribution of gene p-values expected at random to the observed distribution of the gene p-values to arrive at a probability (we're comparing our gene distributions against a random reference). KS is theoretically the better choice because it does not require an arbitrary p-value threshold.
- In the field people rank annotations based on either Fisher or KS p-values, and they select the ranking method that identify the most biologically relevant GO terms
- Note about P values in RNA-seq analysis (adapted from page 7 of Ex2):
  - A p value indicates the probability that a fold change as strong as the observed one, or even stronger, would be seen under the situation described by the null hypothesis (null= that there is no difference between gene expression in DLD-1 vs DLD-1 etoposide). A low probability that the data fits the "there is no expression change" hypothesis, i.e. a p value <5%, means that you can reject the null hypothesis, and claim with high confidence that the gene does show expression difference between groups. In high-throughput biology, we use the adjusted p-value ("padj") to minimize false positives in our list of differentially expressed genes. We consider a fraction of 10% false positives acceptable; we can consider all genes with an adjusted p value below 10% = 0.1 as significant, meaning the gene is differentially expressed upon etoposide treatment.
  - P value lower than your threshold, "Yes, this gene is differentially expressed with etoposide treatment."
  - P value higher than your threshold, "We can't say whether this gene is differentially expressed with this treatment."

12

Apply R workflow from Ex2 to new RNA-seq dataset in Ex3



A549 control and a549 treated with 2uM etoposide cultured for 7 days and stained for a senescence reporter (blue)

- Authors studying senescence induction as an approach to cancer treatment
- A549, model cell line for lung cancer
- Treated with 2uM etoposide, harvested RNA for sequencing after 7 days
- RNA-seq read counts were made available as a public data set

Wang et al. High-Throughput Functional Genetic and Compound Screens Identify Targets for Senescence Induction in Cancer. Cell Reports 2017.

13

Getting help with R:

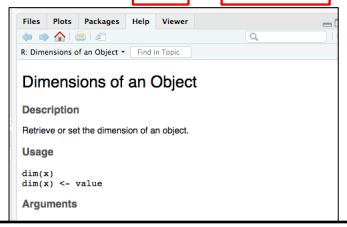
- Ask questions during lab. Anne will log into zoom from 3:30-4:30EST
- Review 20.109.Ex2.codeExplained.pdf under Ex2
- Ask questions on the Mod2 R.studio.cloud benchling page
- Make an appt with *new* BE data lab! mit.mywconline.net
- Use R help function

Method 1
`?function`

or

Method 2
`help(function)`

Example: Type `?dim` or `help(dim)`



14

R Studio Cloud Ex3 Checklist

- ❑ Complete Exercise3_clustering_refresher.R
- ❑ Generate PCA plot of A549 data
- ❑ Generate GO tables of top upregulated and downregulated A549 genes in response to etoposide treatment, with statistical tests
- ❑ Generate PCA plot comparing etoposide treatment in DLD1 and A549 cells
- ❑ Create heatmap of DLD1 and A549 datasets

15

M2D6 "Lab" Checklist

1. Ask questions and understand the RNA-seq data analysis
   – this analysis will translate to figures in your research article
2. You must choose genes for qPCR analysis, note this in your benchling notebook

- Homework due M2D7: Methods M2D1-M2D3 and draft Introduction

16

Methods Reminders:

- Include enough information to replicate the experiment
  - list manufacturers name, like (Qiagen)
- Organize methods into subsections with descriptive titles
  - Put in logical order
  - Begin with topic sentence to introduce purpose
  - R subsection, include package and version, **DESeq2 (v. 1.26.0)**
- Use clear and concise full sentences
  - NO tables and lists
  - Passive voice and past tense
- Use the most flexible units
  - Write concentrations (when known) rather than volumes
- Eliminate 20.109 specific details
  - Example "labeled Row A, Row B…"
  - Do not include details about tubes and water!
  - Assume reader has some biology experience
  - Include steps teaching faculty carried out for you

M2D7 Methods HW should include experiments from M2D1-M2D3

17

Tissue Culture:

TK6 cells were grown in a flask with 12ml RPMI supplemented with FBS. The cells were kept in an incubator at 37°C. A stain was used to assess if the cells were alive or dead.

18

Improving the Methods paragraph

Maintaining lymphoblastoid cell line(s):

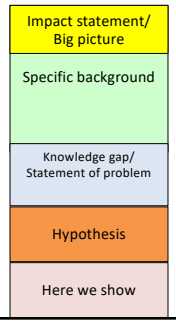
TK6 human lymphoblastoids (gift of the Engelward Lab, MIT) were cultured at 1-9 x 10^5 cells/mL, cell number calculated via hemocytometer and trypan blue stain. Cells were grown in RPMI medium 1640 (Invitrogen) supplemented with 10% fetal bovine serum (Atlanta Biologicals) and 100 units/mL penicillin-streptomycin (Invitrogen). Culture conditions were maintained at 37°C, 5% CO2 and 95% relative humidity.

19

Mod2 Introduction Reminders

M2D7 homework should include:

- Draft the entire first big picture paragraph
- Topic sentence (first sentence) of each additional paragraph
- References in text and brief summary of each reference at the end



Motivation, why should the reader care?

What does a scientist need to know to understand your research?
What is your experimental approach?

What is unknown?

What do you predict the result will be?

What do you report in this research article?

20

M2D2HW feedback: for journal club presentations

- edit the figures / data you are presenting. Take time to describe one or two plots or images rather than list many
- identify color coding on slide in text if space allows
- Verbally transition to next experiment, what did the result motivate the authors to do next?

21