# BGGN 213

## Genome Informatics (IV)

Barry Grant
UC San Diego

http://thegrantlab.org/bggn213

# Background

- Himes et al. "*RNA-Seq Transcriptome Profiling Identifies CRISPLD2 as a Glucocorticoid Responsive Gene that Modulates Cytokine Function in Airway Smooth Muscle Cells*." PLoS ONE. 2014 Jun 13;9(6):e99625. PMID: 24926665.

- Glucocorticoids inhibit inflammatory processes, used to treat asthma because of anti-inflammatory effects on airway smooth muscle (ASM) cells.

- RNA-seq to profile gene expression changes in 4 ASM cell lines treated with dexamethasone (a common synthetic glucocorticoid).

- Used Tophat and Cufflinks and found many differentially expressed genes. Focus on CRISPLD2 that encodes a secreted protein involved in lung development

- SNPs in CRISPLD2 in previous GWAS associated w/ inhaled corticosteroid resistance and bronchodilator response in asthma patients.

- Confirmed the upregulated CRISPLD2 w/ qPCR and increased protein expression w/ Western blotting.

# Data pre-processing

- Analyzing RNA-seq data starts with sequencing reads.
- Many different approaches, see references on class website.
- Our workflow (previously done):
  - Reads downloaded from GEO (GSE:GSE52778)
  - Quantify transcript abundance (kallisto).
  - Summarize to gene-level abundance (txImport)
- Our starting point is a count matrix: each cell indicates the number of reads originating from a particular gene (in rows) for each sample (in columns).

# Data structure: counts + metadata

## countData

| gene | ctrl_1 | ctrl_2 | exp_1 | exp_1 |
|---|---|---|---|---|
| geneA | 10 | 11 | 56 | 45 |
| geneB | 0 | 0 | 128 | 54 |
| geneC | 42 | 41 | 59 | 41 |
| geneD | 103 | 122 | 1 | 23 |
| geneE | 10 | 23 | 14 | 56 |
| geneF | 0 | 1 | 2 | 0 |
| … | … | … | … | … |

**countData** is the count matrix
(number of reads coming from
each gene for each sample)

## colData

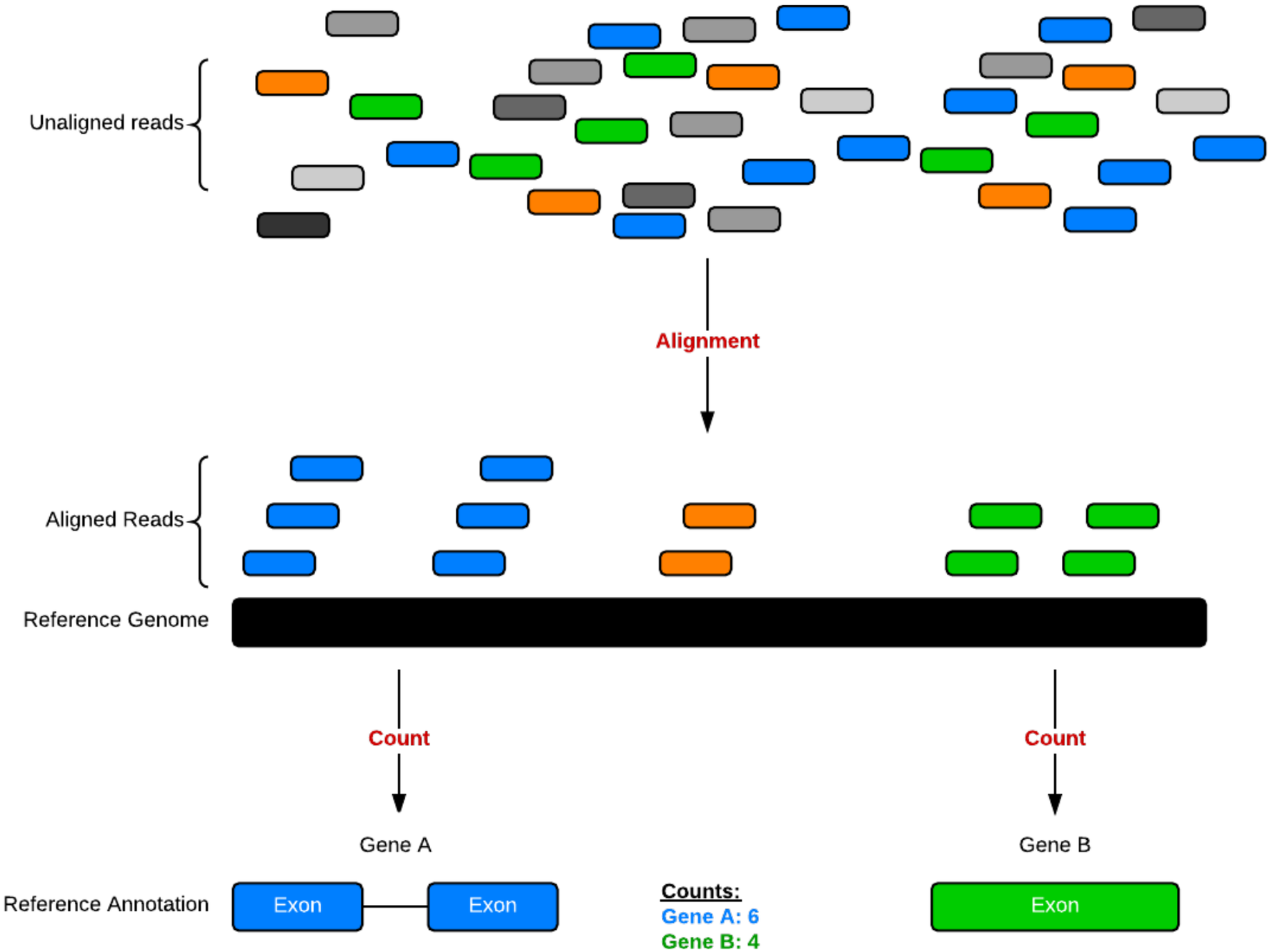
| id | treatment | sex | ... |
|---|---|---|---|
| ctrl_1 | control | male | ... |
| ctrl_2 | control | female | ... |
| exp_1 | treatment | male | ... |
| exp_2 | treatment | female | ... |

Sample names:
ctrl_1, ctrl_2, exp_1, exp_2

**colData** describes metadata
about the *columns* of countData

First column of **colData** must match column names of **countData** (-1st)

# Counting is (relatively) easy:



Do it Yourself!

# Hands-on time!

https://tinyurl.com/bggn213-class17

# Fold change (log ratios)

- To a statistician fold change is sometimes considered meaningless. Fold change can be large (e.g. >>two-fold up- or down-regulation) without being statistically significant (e.g. based on probability values from a t-test or ANOVA).

- To a biologist fold change is almost always considered important for two reasons. First, a very small but statistically significant fold change might not be relevant to a cell's function. Second, it is of interest to know which genes are most dramatically regulated, as these are often thought to reflect changes in biologically meaningful transcripts and/or pathways.

# Inferential statistics

- Inferential statistics are used to make inferences about a population from a sample.

- Hypothesis testing is a common form of inferential statistics. A null hypothesis is stated, such as:
    - “There is no difference in signal intensity for the gene expression measurements in normal and diseased samples.” The alternative hypothesis is that there is a difference.

- We use a test statistic to decide whether to accept or reject the null hypothesis. For many applications, we set the significance level a to $p < 0.05$.

# false discovery rate

- The false discovery rate (FDR) is a popular multiple corrections correction. A false positive (also called a type I error) is sometimes called a false discovery.

$$\mathrm{FDR} = \frac{\#\,\text{false positives}}{\#\,\text{called significant}}$$

- The FDR equals the p value of the t-test times the number of genes measured (e.g. for 10,000 genes and a p value of 0.01, there are 100 expected false positives).

- You can adjust the false discovery rate. For example:

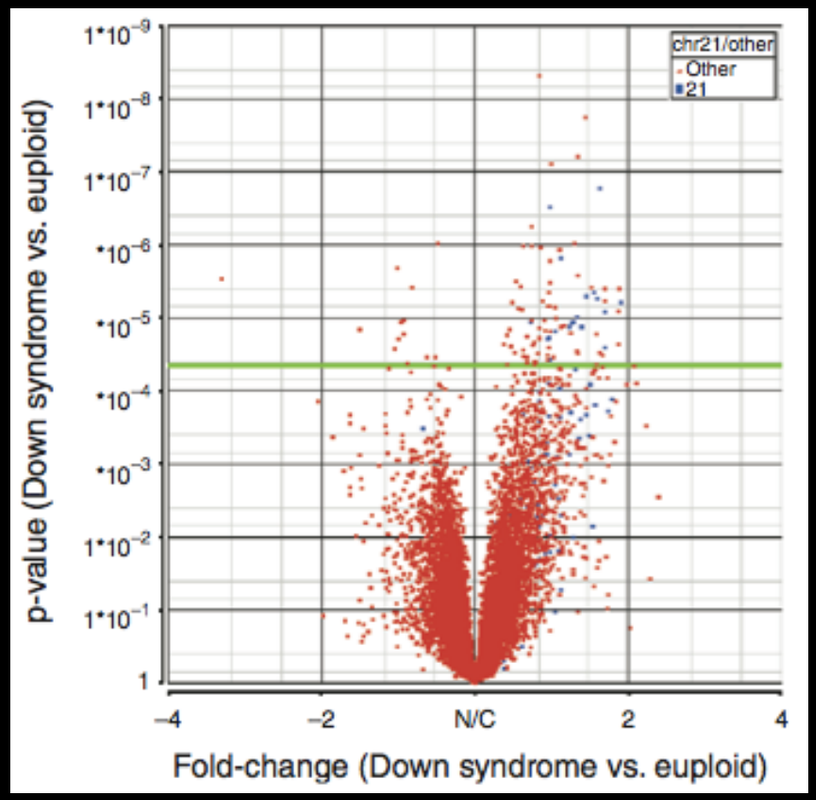
- Would you report 100 regulated transcripts of which 10 are likely to be false positives, or 20 transcripts of which one is likely to be a false positive?

- You can adjust the false discovery rate. For example:

- Would you report 100 regulated transcripts of which 10 are likely to be false positives, or 20 transcripts of which one is likely to be a false positive?

| FDR | # regulated transcripts | # false discoveries |
|---|---|---|
| 0.1 | 100 | 10 |
| 0.05 | 45 | 3 |
| 0.01 | 20 | 1 |

# Volcano plot: significantly regulated genes vs. fold change




- A volcano plot shows fold change (x-axis) versus p value from ANOVA (y-axis). Each point is the expression level of a transcript. Points high up on the y-axis (above the pale green horizontal line) are significantly regulated.

# Principal Components Analysis (PCA)

- An exploratory technique used to reduce the dimensionality of the data set to 2D or 3D

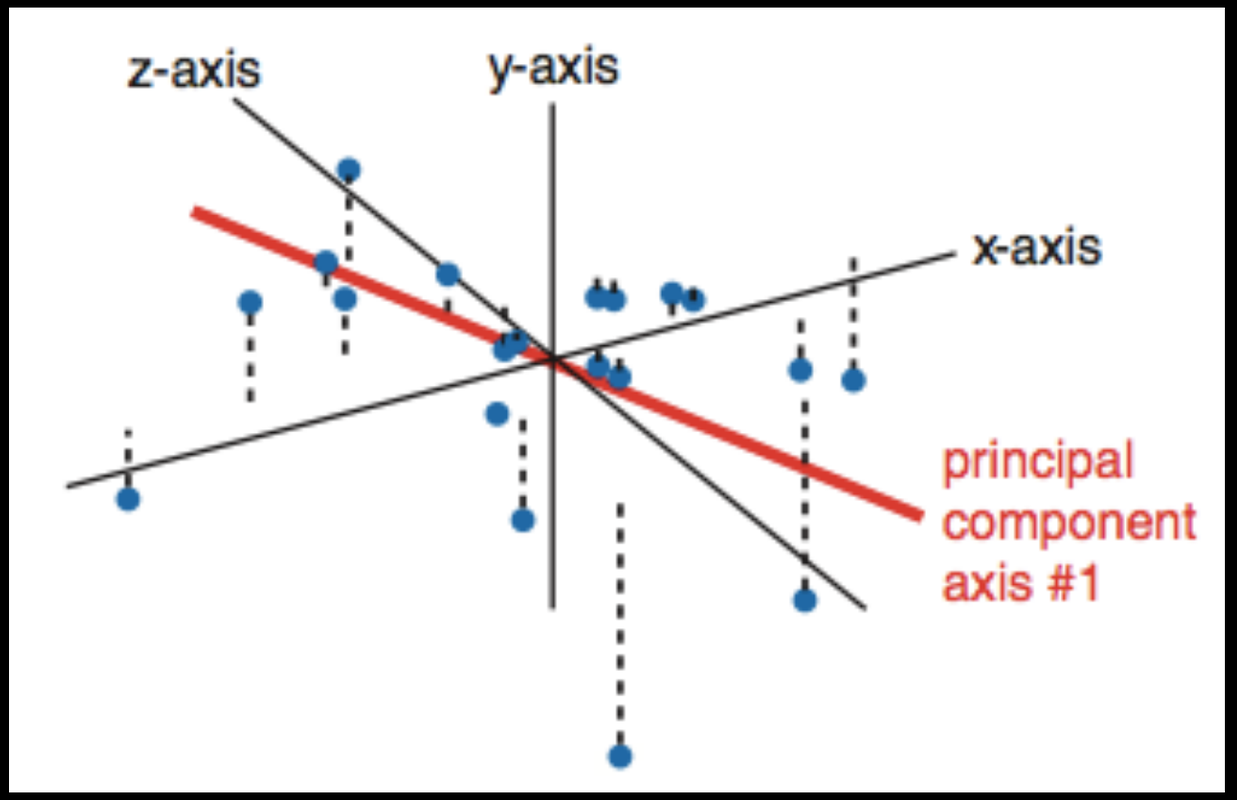
- For a matrix of m genes x n samples, create a new covariance matrix of size n x n

- Thus transform some large number of variables into a smaller number of uncorrelated variables called principal components (PCs).

# PCA objectives

- to reduce dimensionality
- to determine the linear combination of variables
- to choose the most useful variables (features)
- to visualize multidimensional data
- to identify groups of objects (e.g. genes/samples)
- to identify outliers

# Principal Components Analysis




- The first principal component (PC) follows a “best fit” through the data points. Other PCs must cross the origin of the plot, and must be orthogonal.

# Normalization

- Normalization is required to make comparisons in gene expression
    - Between 2+ genes in one sample
    - Between genes in 2+ samples

- Genes will have more reads mapped in a sample with high coverage than one with low coverage
    - 2x depth ≈ 2x expression

- Longer genes will have more reads mapped than shorter genes
    - 2x length ≈ 2x more reads

# **Normalization**: RPKM, FPKM and TPM

- **N.B.** Some tools for differential expression analysis such as edgeR and DESeq2 want raw read counts - i.e. non normalized input!
- However, often for your manuscripts and reports you will want to report normalized counts - e.g. plots of Log(FoldChange) vs Transcripts Per Million (or TPM)
- RPKM, FPKM and TPM all aim to normalize for sequencing depth and gene length. For the former:
    - Count up the total reads in a sample and divide that number by 1,000,000 – this is our “per million” scaling factor.
    - Divide the read counts by the “per million” scaling factor. This normalizes for sequencing depth, giving you reads per million (RPM)
    - Divide the RPM values by the length of the gene, in kilobases. This gives you RPKM.

- FPKM was made for paired-end RNA-seq
- With paired-end RNA-seq, two reads can correspond to a single fragment
- The only difference between RPKM and FPKM is that FPKM takes into account that two reads can map to one fragment (and so it doesn’t count this fragment twice).

- TPM is very similar to RPKM and FPKM. The only difference is the order of operations. Here's how you calculate TPM:
  - Divide the read counts by the length of each gene in kilobases. This gives you reads per kilobase (RPK).
  - Count up all the RPK values in a sample and divide this number by 1,000,000. This is your "per million" scaling factor.
  - Divide the RPK values by the "per million" scaling factor. This gives you TPM.
- So you see, when calculating TPM, the only difference is that you normalize for gene length first, and then normalize for sequencing depth second. However, the effects of this difference are quite profound.

- When you use TPM, the sum of all TPMs in each sample are the same.
- This makes it easier to compare the proportion of reads that mapped to a gene in each sample.
- In contrast, with RPKM and FPKM, the sum of the normalized reads in each sample may be different, and this makes it harder to compare samples directly.