3000788 Intro to Comp Molec Biol

Week 5: RNA sequencing

Fall 2024

Sira Sriswasdi, PhD

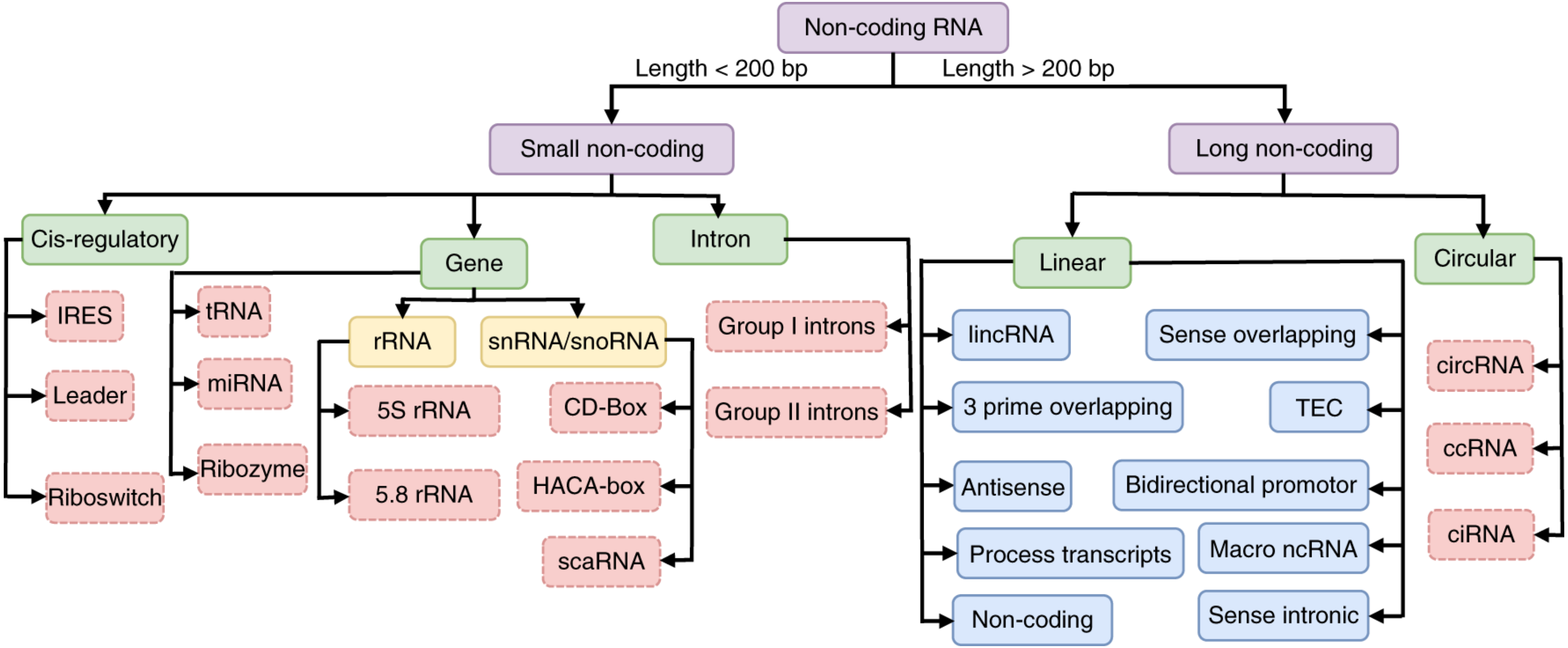
- Research Affairs
- Center of Excellence in Computational Molecular Biology (CMB)
- Center for Artificial Intelligence in Medicine (CU-AIM)

Part I: RNA sequencing & differential expression

- Which class of RNA molecules are you interested in?
- Gene level or isoform level
- Do you want to discover new isoform?
- How to quantify gene expression?
- DESeq2 (read count) vs sleuth (TPM) pipeline

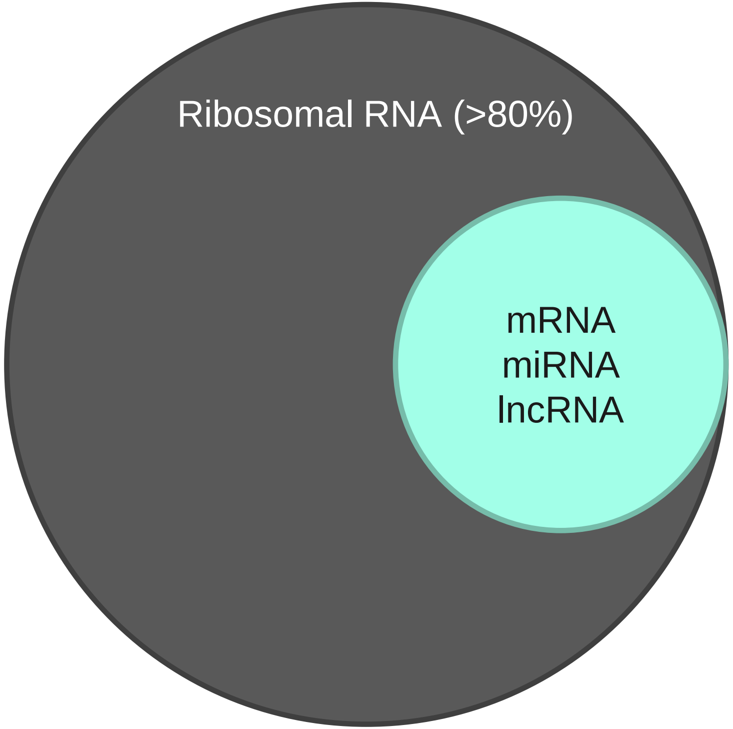
RNA-seq scopes

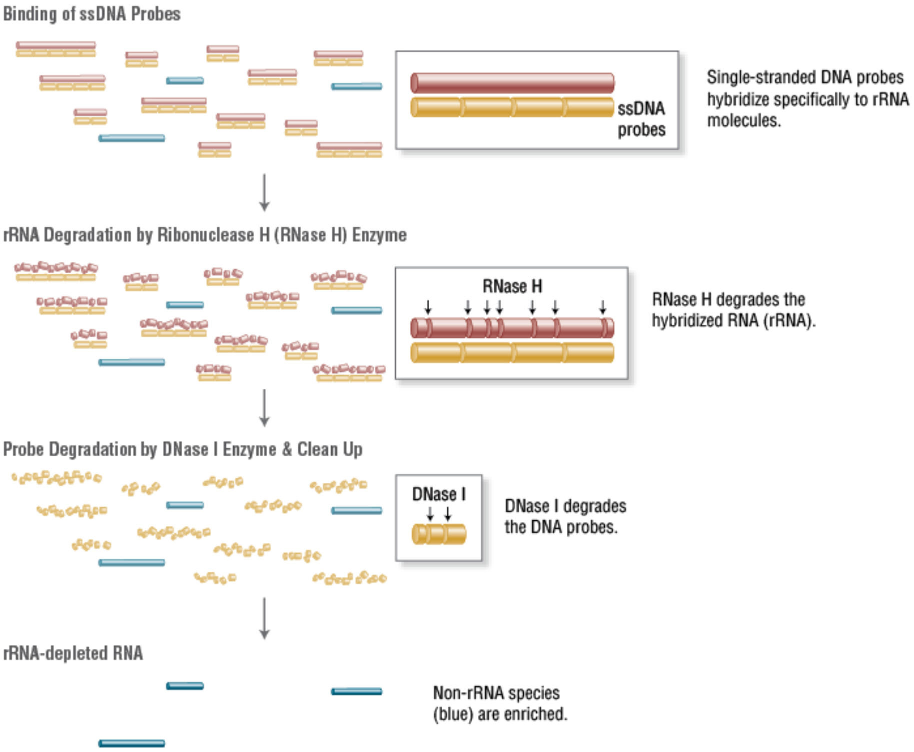
Non-coding RNAs



Amin, N. et al. Nature Machine Intelligence 1:246-256 (2019)

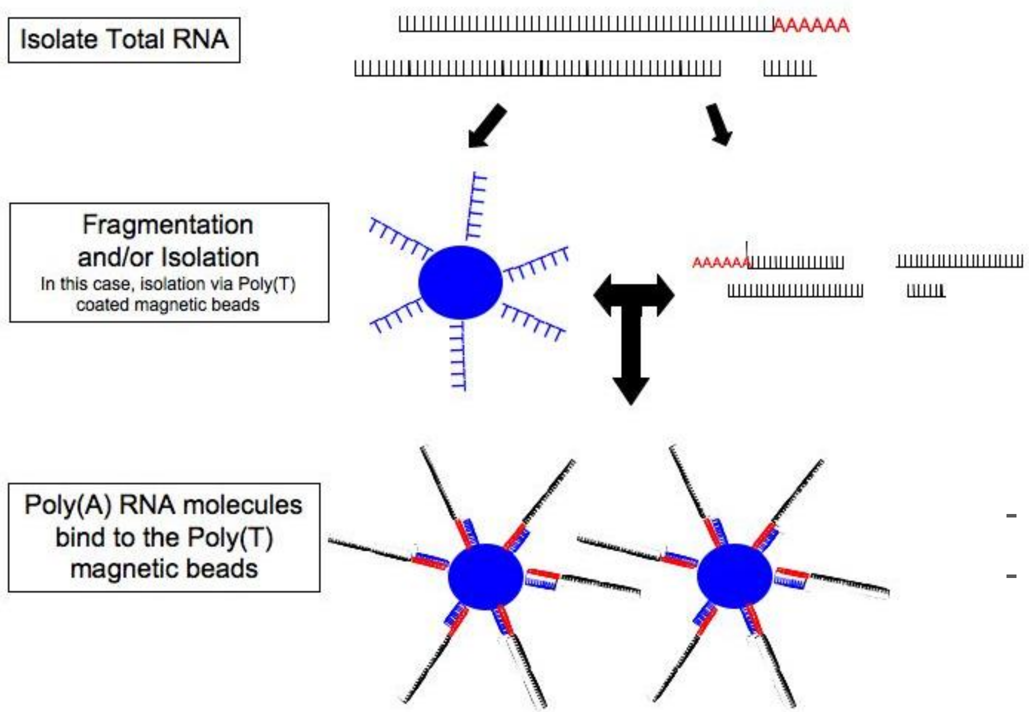
Total RNA sequencing




Binding of ssDNA Probes
ssDNA probes
Single-stranded DNA probes hybridize specifically to rRNA molecules.
rRNA Degradation by Ribonuclease H (RNase H) Enzyme
RNase H
RNase H degrades the hybridized RNA (rRNA).
Probe Degradation by DNase I Enzyme & Clean Up
DNase I
DNase I degrades the DNA probes.
rRNA-depleted RNA
Non-rRNA species (blue) are enriched.

Source: New England BioLabs

mRNA and miRNA sequencing

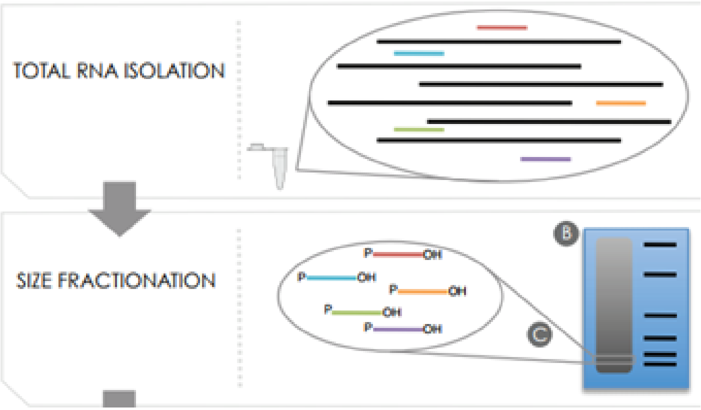


Isolate Total RNA

Fragmentation and/or Isolation

In this case, isolation via Poly(T) coated magnetic beads

Poly(A) RNA molecules bind to the Poly(T) magnetic beads

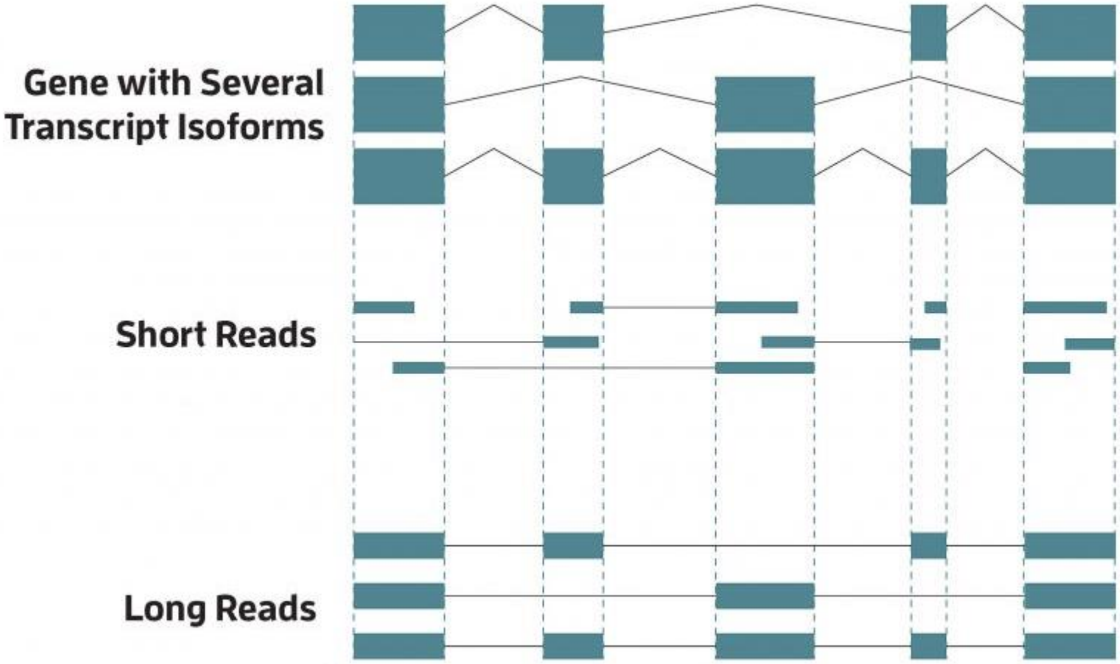


TOTAL RNA ISOLATION

SIZE FRACTIONATION

- Selection by polyT probe
- Size fractionation

Full-length transcript sequencing



https://www.genengnews.com/resources/tutorial/full-length-transcript-sequencing-no-assembly-required/

Choosing RNA-seq technique

- miRNA or mRNA or total-RNA

- Single-end or paired-end
  - Single-end is ok for gene-level quantification
  - Paired-end is needed to distinguish isoforms

- Illumina or 3rd generation sequencer
  - Long-read data is helpful for genes with complex isoforms and for detecting novel isoforms

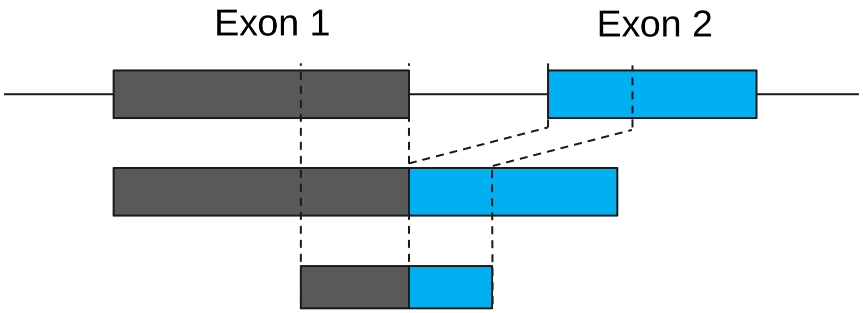
RNA-seq analysis

Three primary pipelines

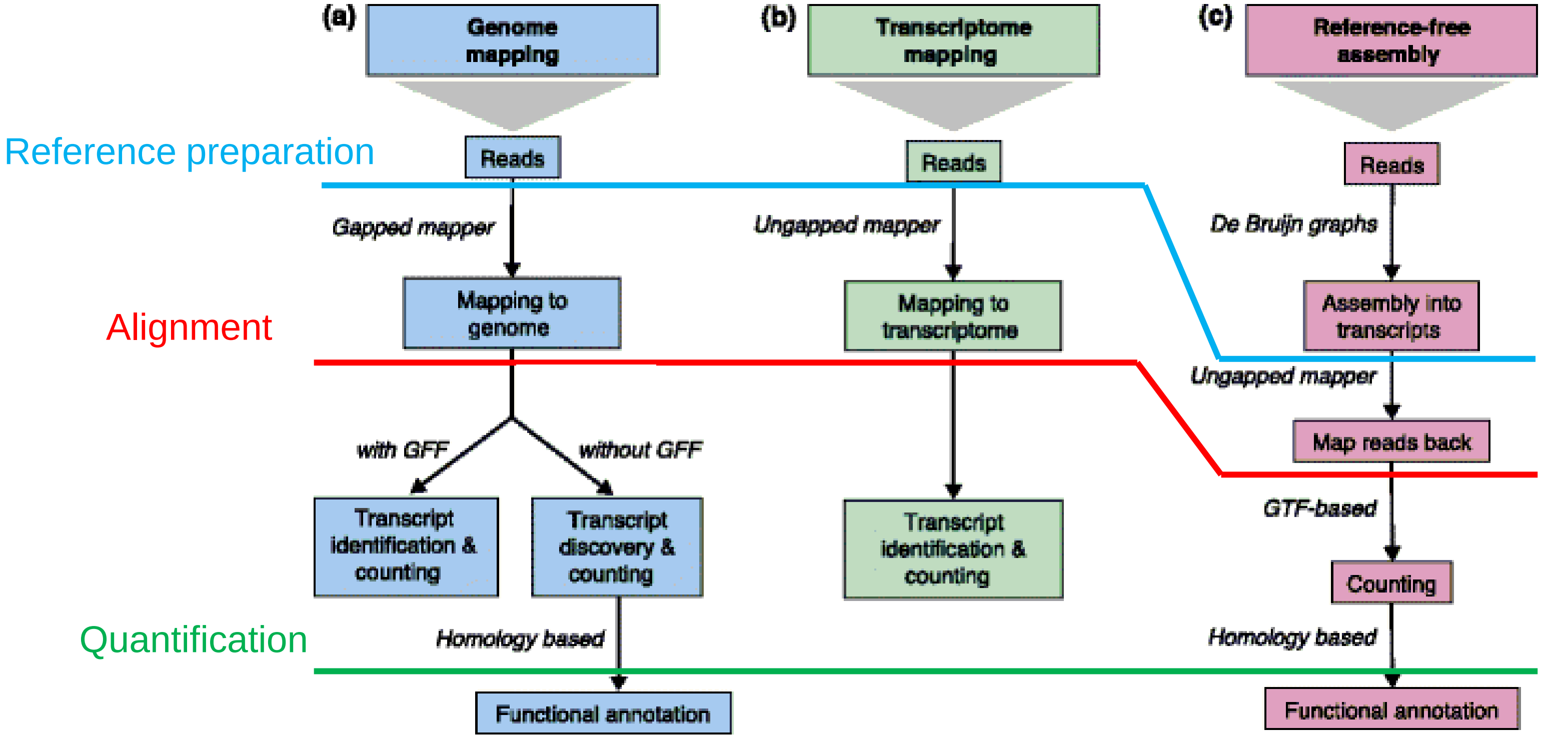
- Reference-free
  - Novel species, rely on *de novo* assembly

- Reference transcriptome
  - Fast, cannot discover new isoform
  - Ungapped, *k*-mer-based alignment

- Reference genome
  - Slow, but can detect new isoforms
  - Gapped alignment

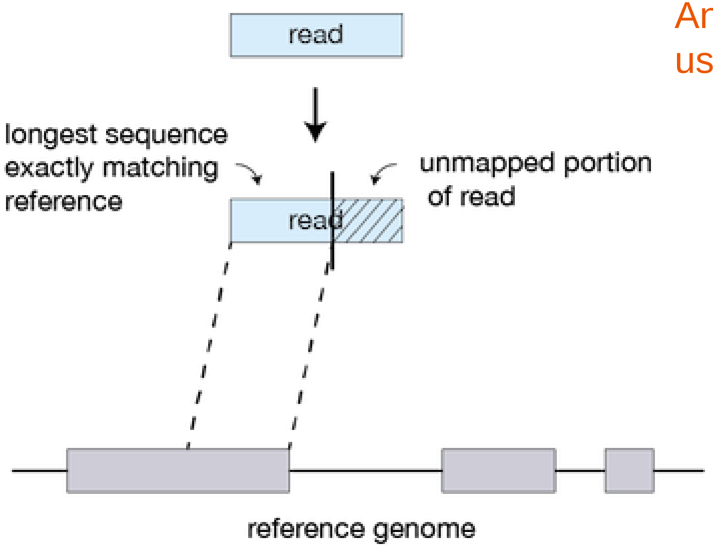


Pipeline overview

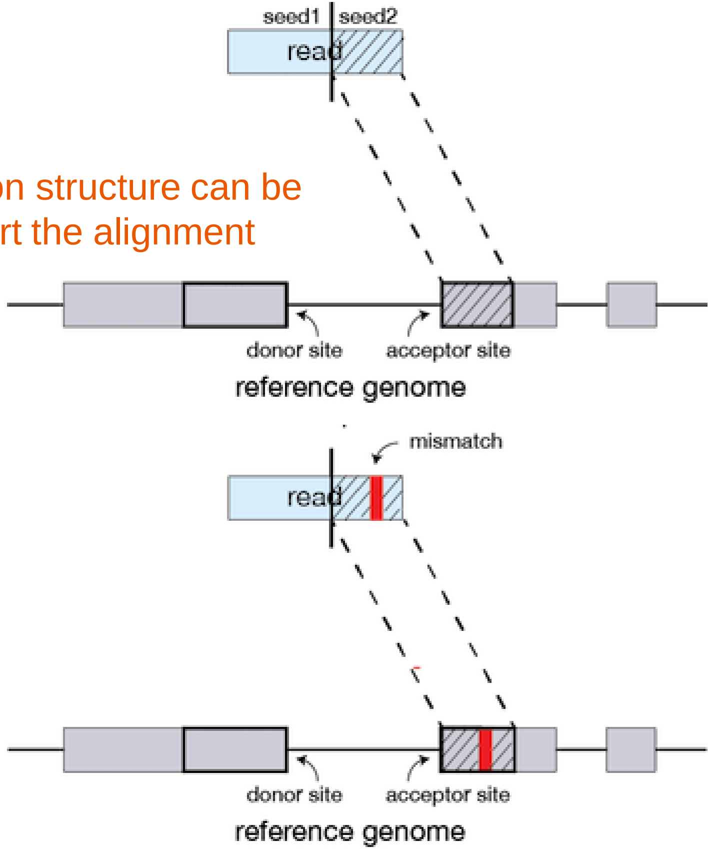


Conesa *et al.* Genome Biology 17:13 (2016)

Gapped alignment

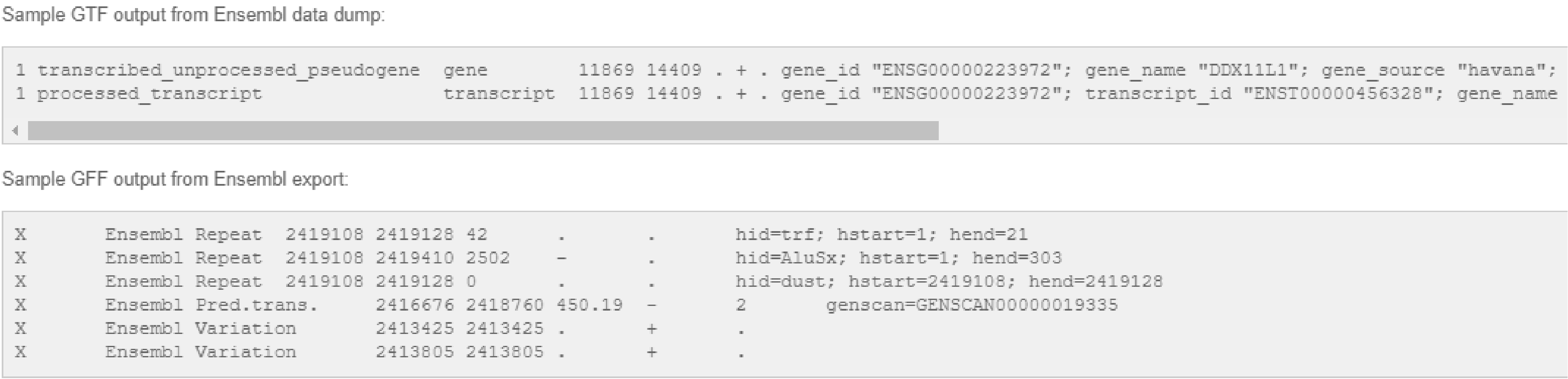


Annotated exon structure can be used to support the alignment



https://hbctraining.github.io/Intro-to-rnaseq-hpc-O2/lessons/03_alignment.html

GTF/GFF genome annotation format



Sample GTF output from Ensembl data dump:

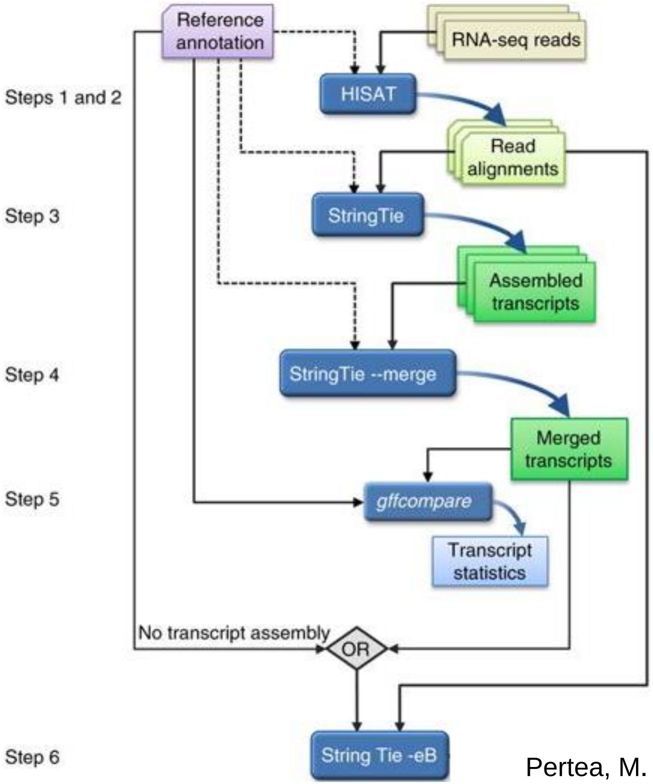
```
1 transcribed_unprocessed_pseudogene  gene        11869 14409 . + . gene_id "ENSG00000223972"; gene_name "DDX11L1"; gene_source "havana";
1 processed_transcript                transcript  11869 14409 . + . gene_id "ENSG00000223972"; transcript_id "ENST00000456328"; gene_name
```

Sample GFF output from Ensembl export:

```
X       Ensembl Repeat  2419108 2419128 42      .       .       hid=trf; hstart=1; hend=21
X       Ensembl Repeat  2419108 2419410 2502    -       .       hid=AluSx; hstart=1; hend=303
X       Ensembl Repeat  2419108 2419128 0       .       .       hid=dust; hstart=2419108; hend=2419128
X       Ensembl Pred.trans.     2416676 2418760 450.19  -       2       genscan=GENSCAN00000019335
X       Ensembl Variation       2413425 2413425 .       +       .
X       Ensembl Variation       2413805 2413805 .       +       .
```

- Tab-separated text file
- Chromosome ID, object name, base pair positions, strand, and other annotation details

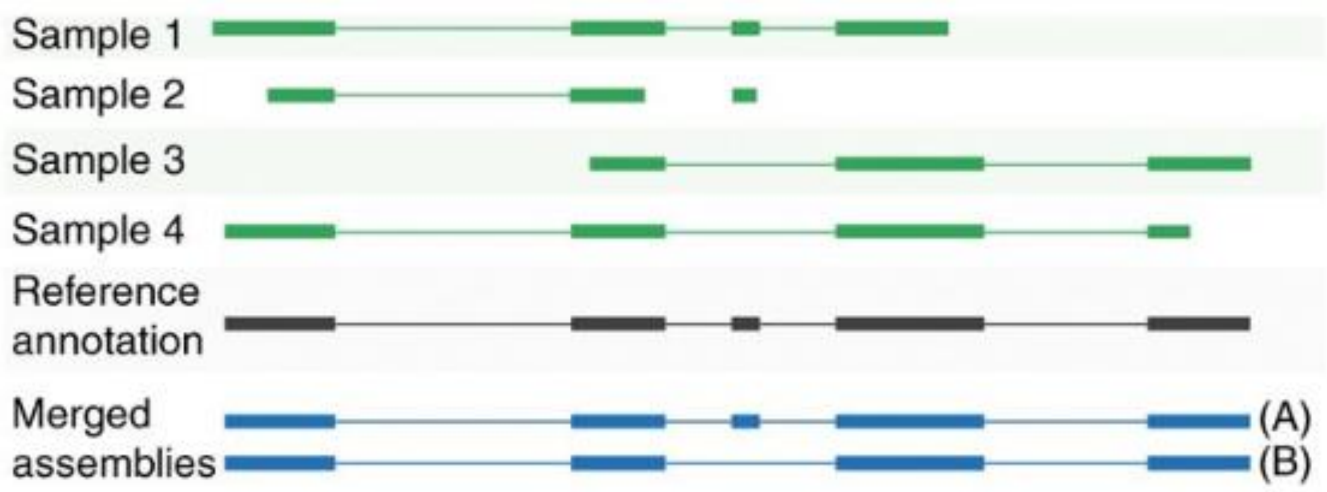
Multi-step post-alignment processing



- Initial alignment
- Assemble potential novel isoforms
- Merge isoforms across samples
- Re-quantify isoform abundances using the merged database of isoforms

Pertea, M. et al. Nature Protocols 11:1650-1667 (2016)

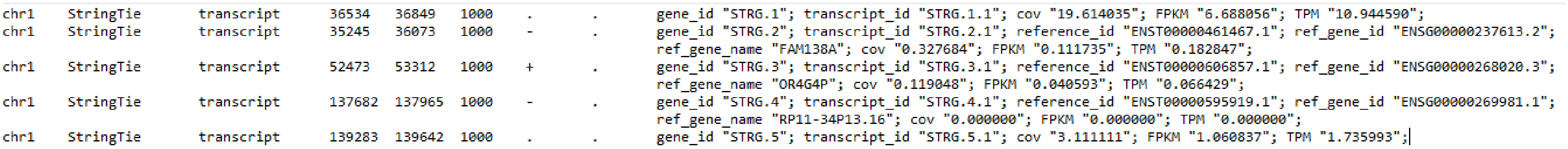
Importance of merging isoforms



Pertea, M. et al. Nature Protocols 11:1650-1667 (2016)

- Rare isoform may be missing in some samples
- Reads can get misinterpret if the correct isoform is not in the reference

GTF with abundance annotation



```
chr1    StringTie       transcript      36534   36849   1000    .       .       gene_id "STRG.1"; transcript_id "STRG.1.1"; cov "19.614035"; FPKM "6.688056"; TPM "10.944590";
chr1    StringTie       transcript      35245   36073   1000    -       .       gene_id "STRG.2"; transcript_id "STRG.2.1"; reference_id "ENST00000461467.1"; ref_gene_id "ENSG00000237613.2";
                                                                                ref_gene_name "FAM138A"; cov "0.327684"; FPKM "0.111735"; TPM "0.182847";
chr1    StringTie       transcript      52473   53312   1000    +       .       gene_id "STRG.3"; transcript_id "STRG.3.1"; reference_id "ENST00000606857.1"; ref_gene_id "ENSG00000268020.3";
                                                                                ref_gene_name "OR4G4P"; cov "0.119048"; FPKM "0.040593"; TPM "0.066429";
chr1    StringTie       transcript      137682  137965  1000    -       .       gene_id "STRG.4"; transcript_id "STRG.4.1"; reference_id "ENST00000595919.1"; ref_gene_id "ENSG00000269981.1";
                                                                                ref_gene_name "RP11-34P13.16"; cov "0.000000"; FPKM "0.000000"; TPM "0.000000";
chr1    StringTie       transcript      139283  139642  1000    .       .       gene_id "STRG.5"; transcript_id "STRG.5.1"; cov "3.111111"; FPKM "1.060837"; TPM "1.735993";
```

- Different tool outputs transcript abundance in different format

- GTF can accommodate abundance annotation in the last columns
  - Coverage (cov) = fraction of transcript length with mapped read
  - FPKM = Fragment per kilobase of exon per million reads mapped
  - TPM = Transcript per million

Units for transcript abundance

$$\text{FPKM} = \frac{\text{Read Count}}{\frac{\text{Transcript Length}}{1{,}000} \times \frac{\text{Total Read Count}}{1{,}000{,}000}}$$

Long transcript generates more fragments and more read counts

Experiment with higher sequencing depth generates more read counts

Similar to percentage (but per million)

$$\text{TPM} = \frac{\text{FPKM}}{\sum \text{FPKM}} \times 1{,}000{,}000$$

- Read count (number of mapped reads)
- FPKM = Fragment per kilobase of exon per million reads mapped
- TPM = Transcript per million

Alignment-based pipeline summary

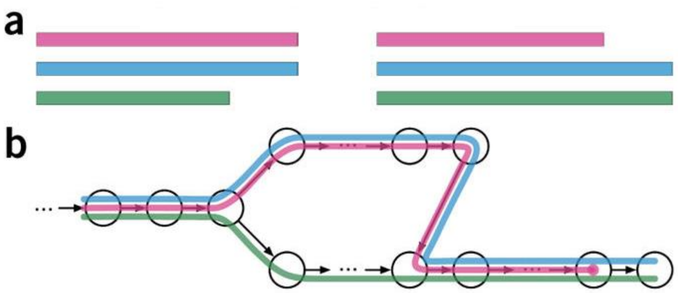
- Initial alignment to reference genome (with annotated gene structure)
  - STAR / HISAT2

- [Optional]
  - Identify novel isoforms
  - Merge isoforms across samples

- Quantify transcript abundances
  - Read count / FPKM / TPM
  - StringTie2 / htseq-count

k-mer pseudoalignment

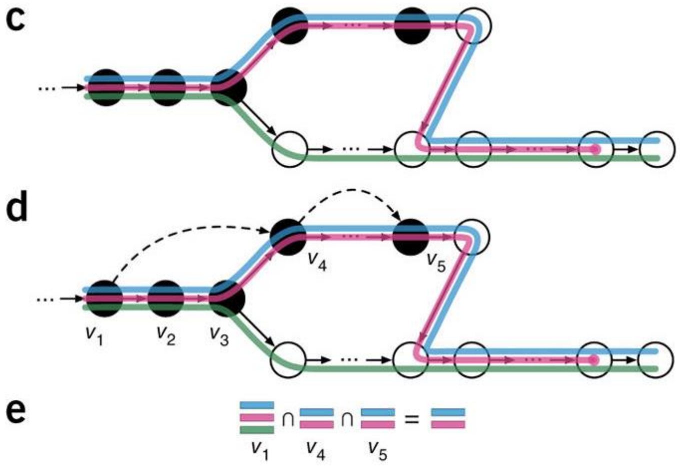
k-mer database for transcriptome



Bray *et al.* Nat Biotech 34:525-527 (2016)

- Create de Bruijn graph with *k*-mer as nodes
- Map node to transcripts with that *k*-mer
- Contig = a path on de Bruijn graph that mapped to the same transcript

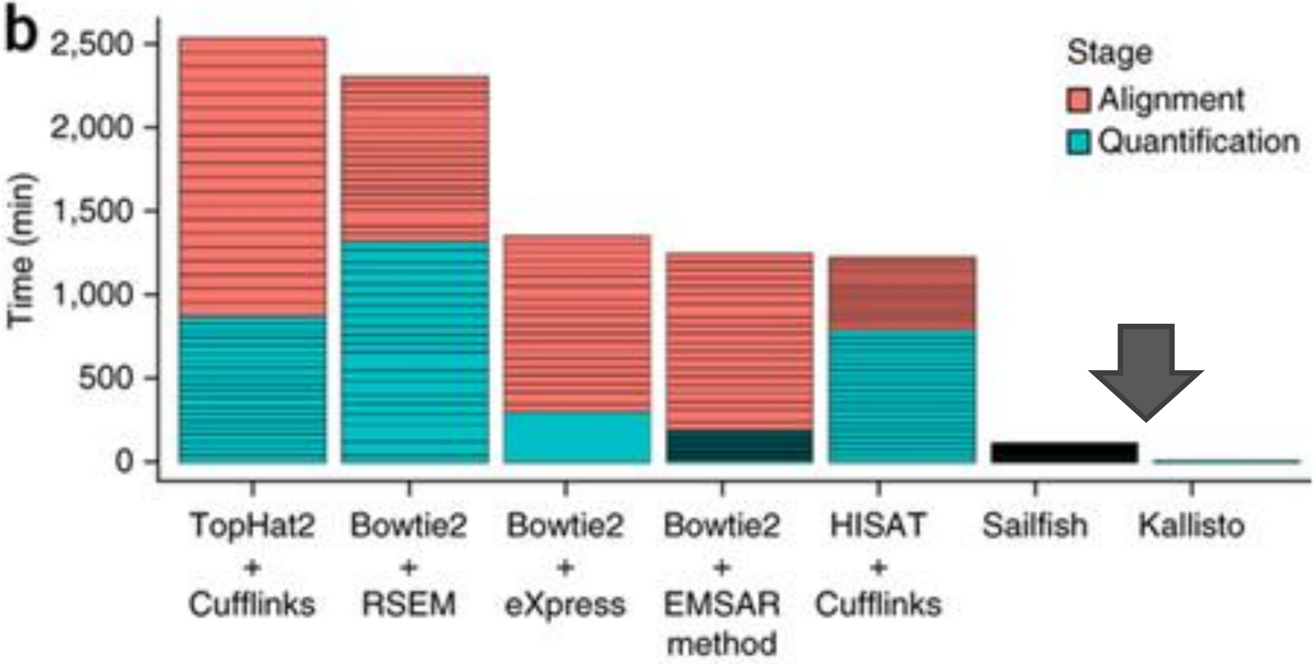
k-mer pseudoalignment



Bray *et al.* Nat Biotech 34:525-527 (2016)

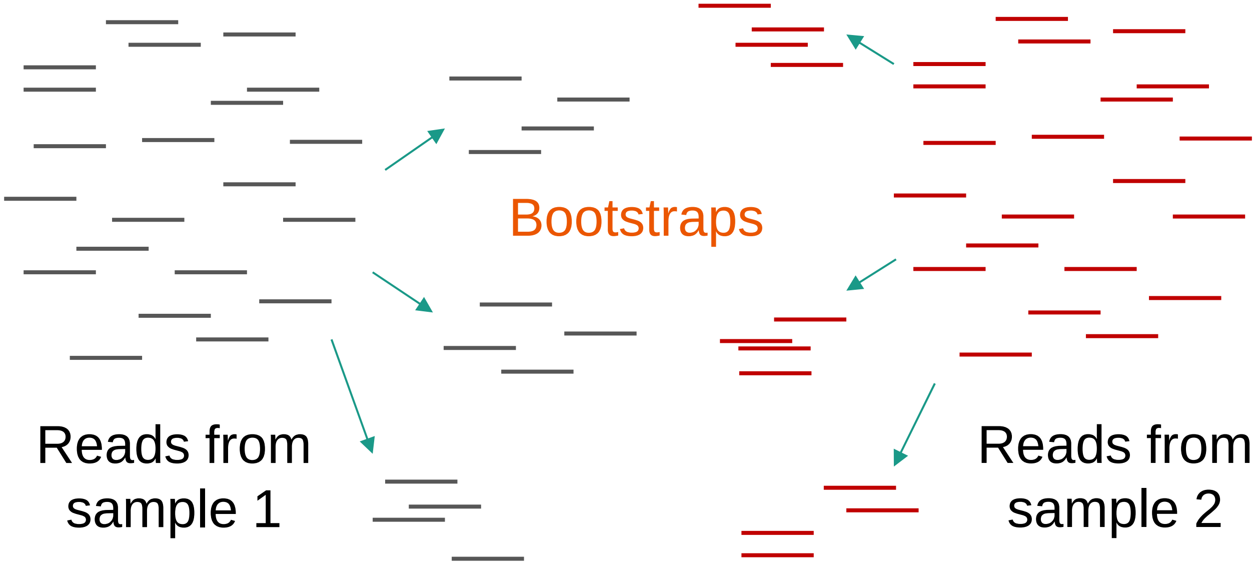
- For a new read, all of its *k*-mers are mapped against contigs
  - Ignore the ordering of *k*-mers on the read
- Report only contigs that are compatible with all *k*-mers
- Speed up by skipping uninformative *k*-mer
  - V_2 and V_3 regions
  - Only 2-4 *k*-mer lookups are enough

>100 fold speed up with pseudoalignment



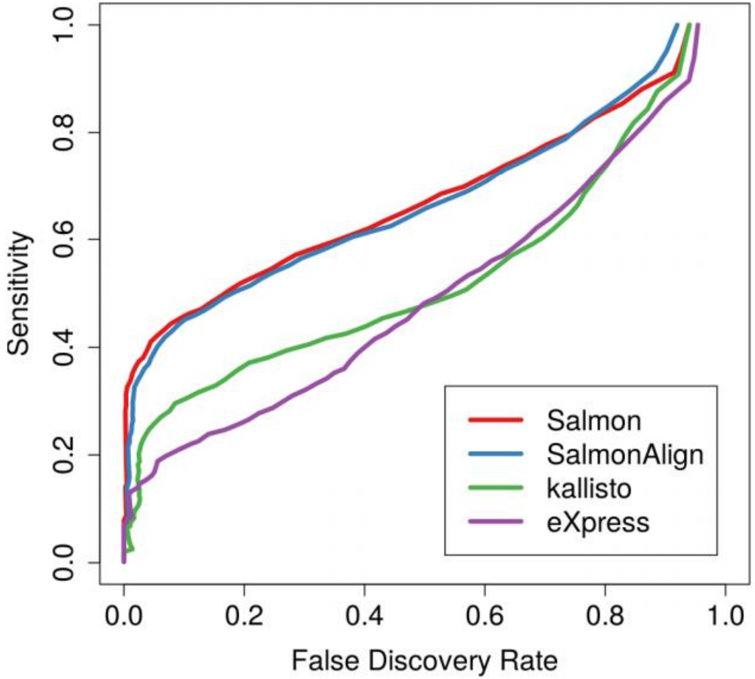
Bray *et al.* Nat Biotech 34:525-527 (2016)

Bootstrapping enabled by pseudoalignment



- Bootstrapping estimates technical variances

Salmon: improved *k*-mer alignment



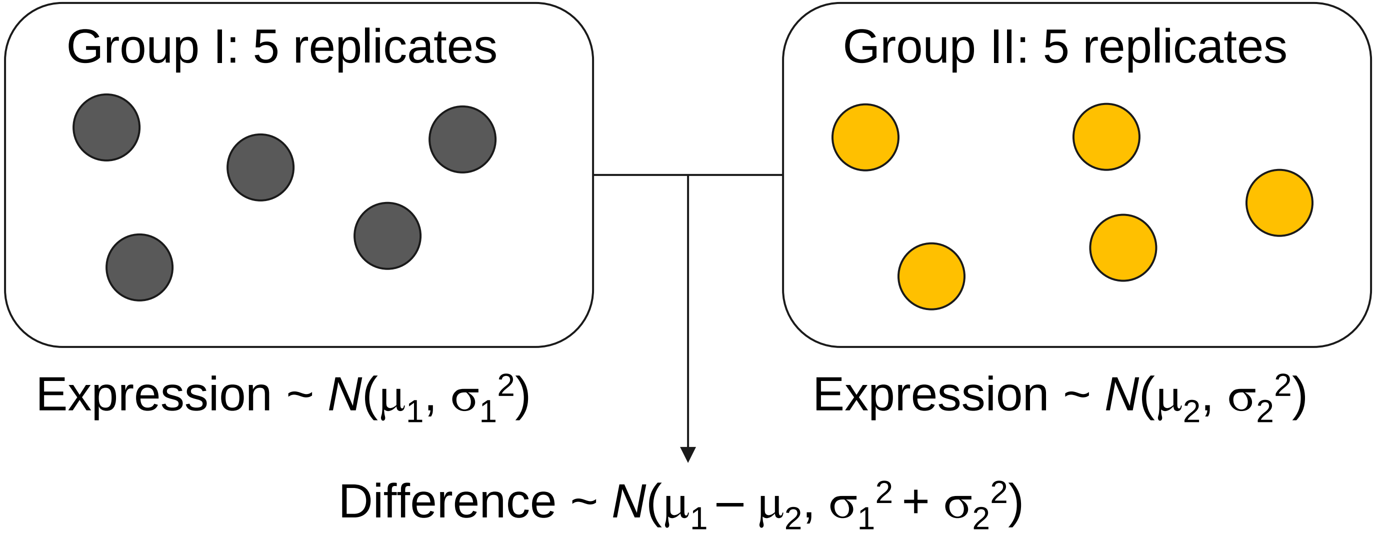
Patro *et al.* Nat Methods 14:417-419 (2017)

- Also track the location of *k*-mer on the input read → semi-alignment

- Correct quantification based on GC content and 3’ / 5’ amplification biases

Differential expression analysis

DE with *t*-test



Group I: 5 replicates

Group II: 5 replicates

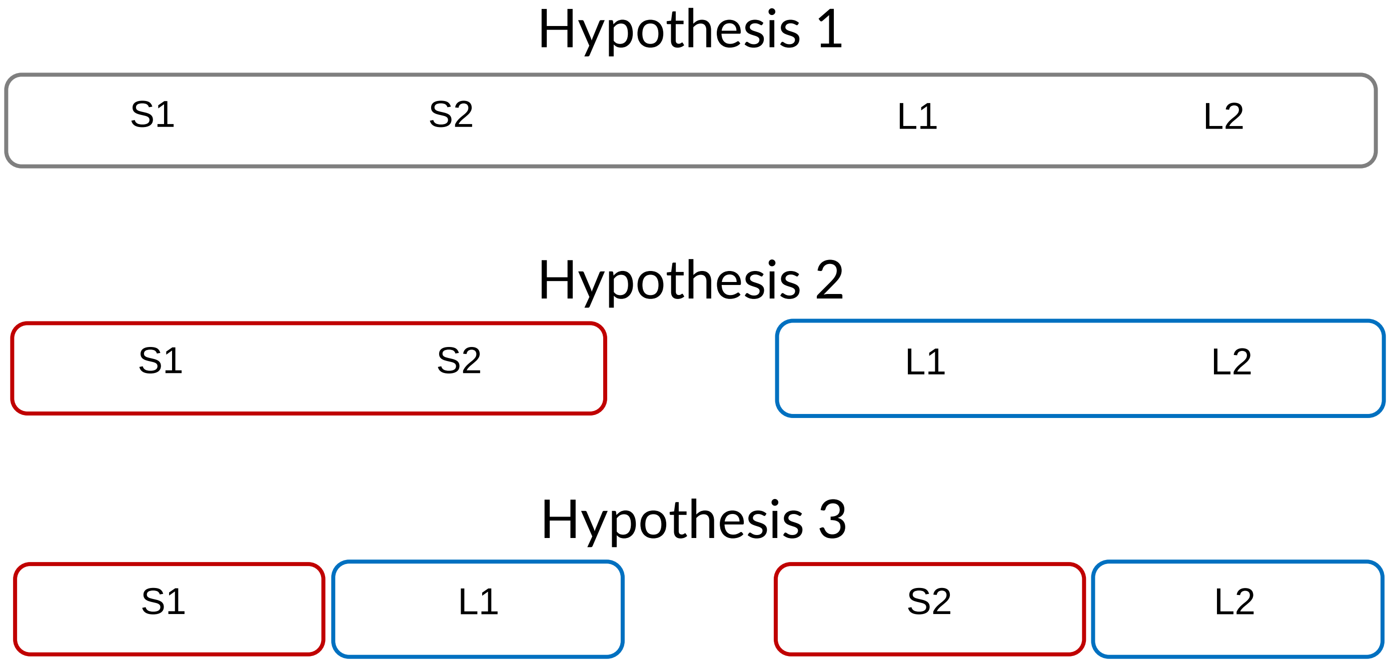
Expression ~ $N(\mu_1, \sigma_1^2)$

Expression ~ $N(\mu_2, \sigma_2^2)$

Difference ~ $N(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2)$

- Simple *t*-test for normally distributed abundance data

DE as nested model testing / likelihood ratio test



Hypothesis 1

S1 S2 L1 L2

Hypothesis 2

S1 S2

L1 L2

Hypothesis 3

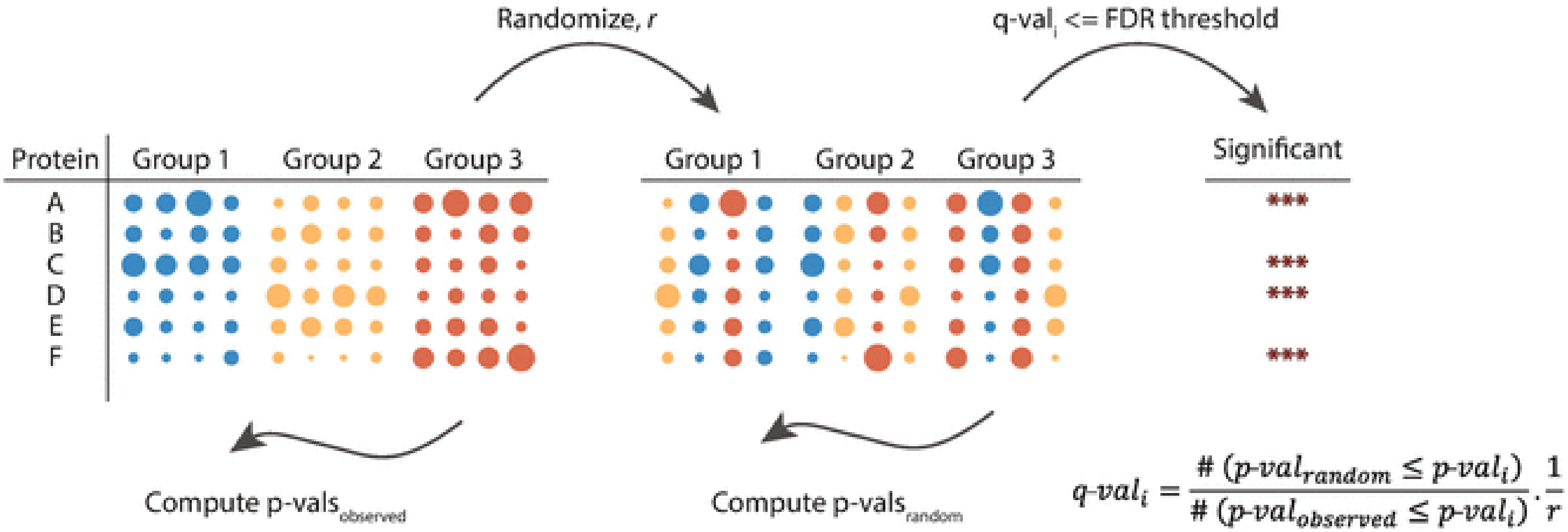
S1

L1

S2

L2

DE as permutation test

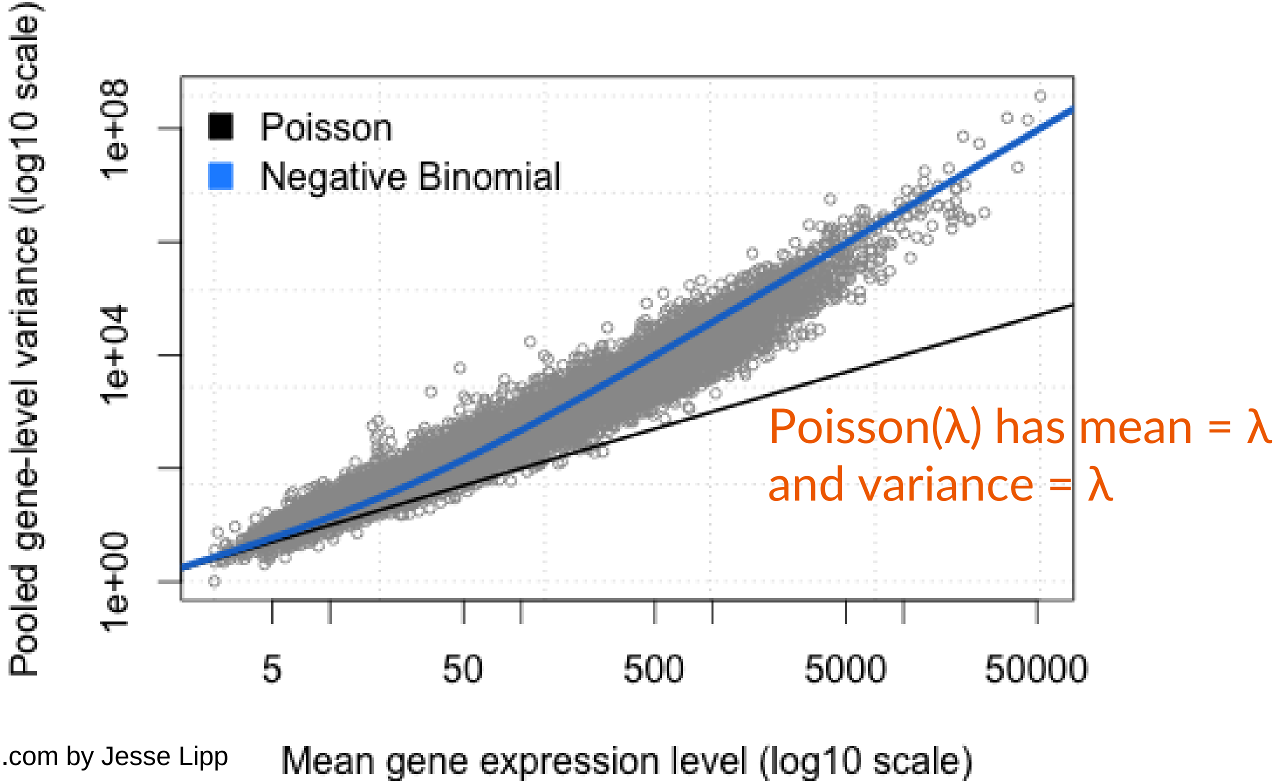

Randomize, r
q-val$_i$ <= FDR threshold
Protein
Group 1
Group 2
Group 3
A
B
C
D
E
F
Group 1
Group 2
Group 3
Significant
***
***
***
***
Compute p-vals$_{observed}$
Compute p-vals$_{random}$
$q\text{-}val_i = \frac{\#\,(p\text{-}val_{random} \leq p\text{-}val_i)}{\#\,(p\text{-}val_{observed} \leq p\text{-}val_i)} \cdot \frac{1}{r}$

Tyanova and Cox. Cancer Systems Biology pp 133-148. (2018)

- Permuting sample labels = remove condition-specificity

DESeq2 model for read count

The distribution of RNA-seq read count



Poisson(λ) has mean = λ and variance = λ

Image from bioramble.wordpress.com by Jesse Lipp

Negative binomial model

- NB(*r*, *p*) = the number of failures that we will see in a series of Bernoulli trials with probability of success *p* until we obtain *r* successes
  - X O O X X X O O X O O = 5 failures until 6 successes

- $P_{\mathrm{NB}}(k; r, p) = \binom{k+r-1}{k}(1-p)^k p^r$
  - k + r – 1 locations to place k failures (the last location must be success)

- Mean = $\frac{pr}{(1-p)}$
- Variance = $\frac{pr}{(1-p)^2} = \frac{pr}{(1-p)}\frac{1}{(1-p)} = \frac{pr}{(1-p)}\left(1+\frac{p}{1-p}\right)$

$$= \frac{pr}{(1-p)} + \frac{p^2 r}{(1-p)^2} = \frac{pr}{(1-p)} + \left(\frac{pr}{1-p}\right)^2 \frac{1}{r}$$

Another view of negative binomial model

- $\mathrm{P}_{\mathrm{NB}}(k; r, p) = \int_0^\infty \mathrm{P}_{\mathrm{Poisson}(\lambda)}(k) \cdot \mathrm{P}_{\mathrm{Gamma}\left(r, \frac{1-p}{p}\right)}(\lambda)\, d\lambda$
- Negative binomial distribution is a continuous mixture of Poisson distribution, with mixing weights Gamma-distributed
  - Same as Gamma site-specific mutation rates

- Bulk gene expression is an average over many cells
- Mixture of read counts from multiple cells, each following Poisson(λ)

DESeq2 model of gene expression

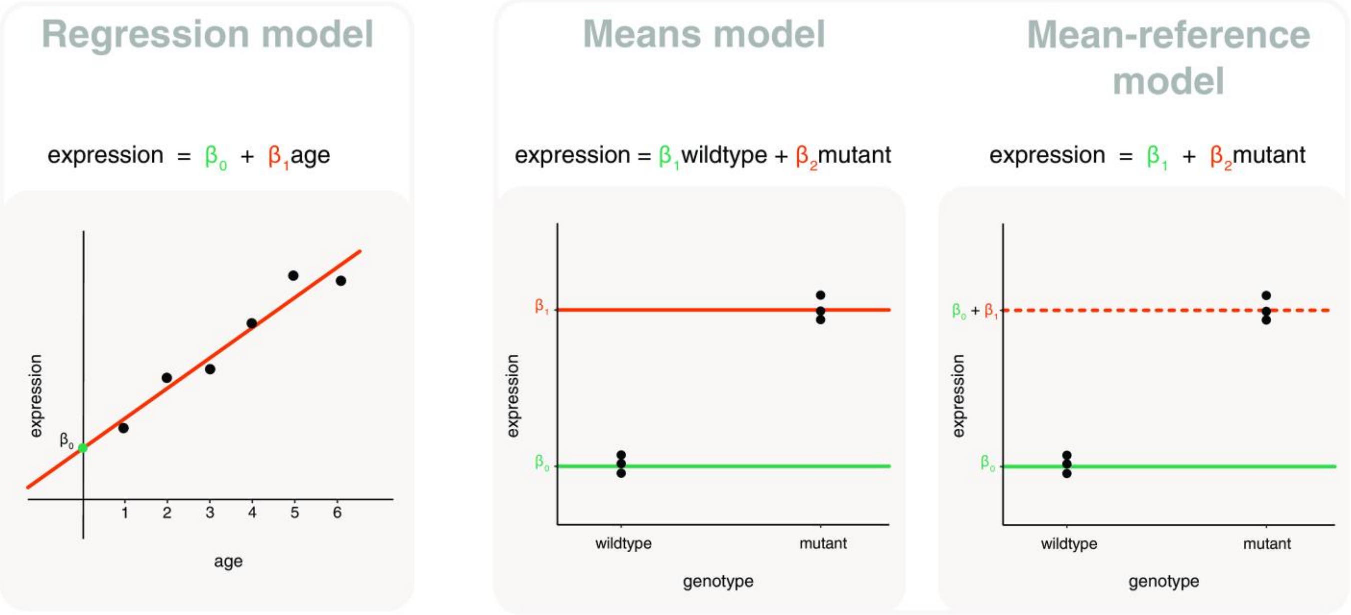
- Read count $K_{i,j}$ ~ NB($\mu_{i,j}$, $\sigma_{i,j}^2$), for gene *i* from sample *j*
  - $\mu_{i,j}$ = sample effects x sequencing effects
  - $\sigma_{i,j}^2 = \mu_{i,j}$ + gene-specific effects x $\mu_{i,j}^2$

- Sample effects
  - Control / Treatment
  - Confounding factors: age, time after treatment, etc.
  - Log FC = $\sum_r x_{j,r}\beta_{i,r}$ where $x_{j,r}$ are design parameters for sample *j* and β_{ir} are the effect sizes
    - Linear effect model

Linear effect models

Covariates: quantitative measurements (e.g. age)

Factors: categorical variables (e.g. genotype)



Regression model

expression $= \beta_0 + \beta_1$age

Means model

expression $= \beta_1$wildtype $+ \beta_2$mutant

Mean-reference model

expression $= \beta_1 + \beta_2$mutant

Law, C.E. et al. F100Res 9:1444 (2020)

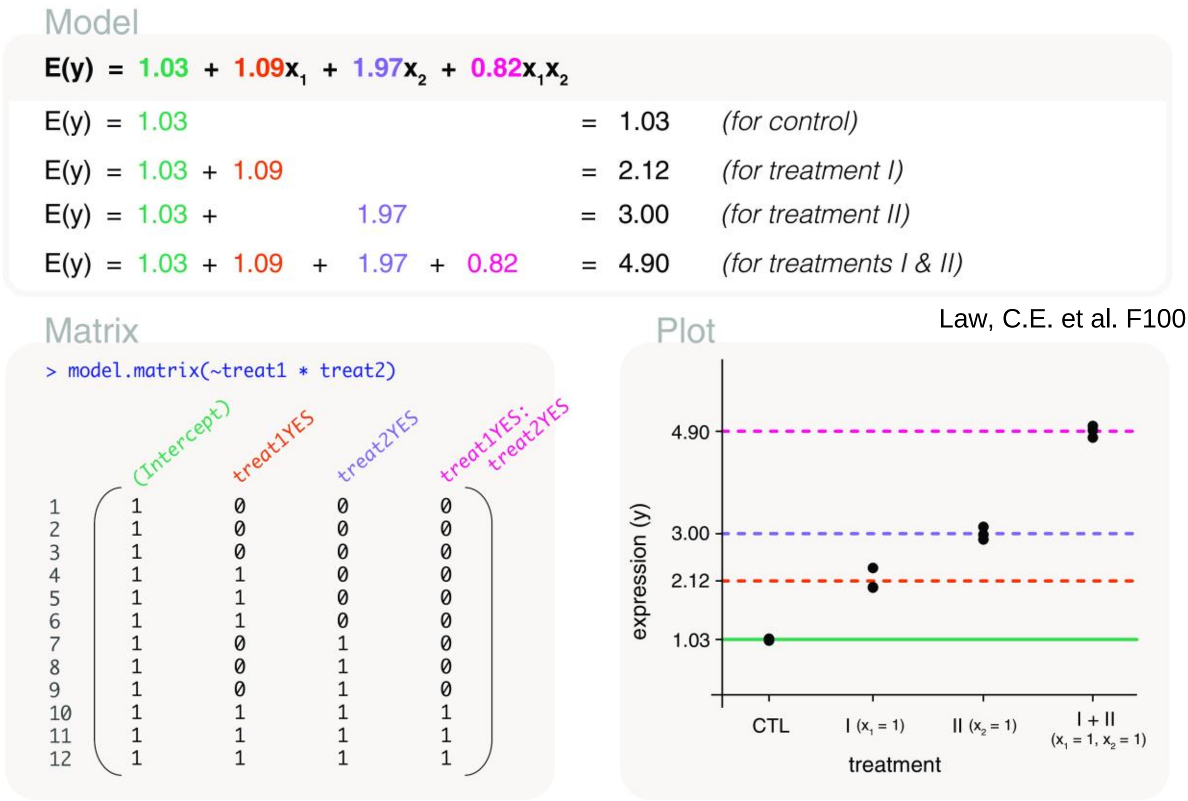
Legend

- Original data points

🟩 🟥 Expected gene expression *(based on model)*

- - - Expected gene expression *(of non-reference levels in mean-reference model)*

Linear model for multiple effects



Model

$$E(y) = 1.03 + 1.09x_1 + 1.97x_2 + 0.82x_1x_2$$

$$E(y) = 1.03 = 1.03 \quad \textit{(for control)}$$

$$E(y) = 1.03 + 1.09 = 2.12 \quad \textit{(for treatment I)}$$

$$E(y) = 1.03 + 1.97 = 3.00 \quad \textit{(for treatment II)}$$

$$E(y) = 1.03 + 1.09 + 1.97 + 0.82 = 4.90 \quad \textit{(for treatments I \& II)}$$

Law, C.E. et al. F100Res 9:1444 (2020)

Matrix

```
> model.matrix(~treat1 * treat2)
```

$$\begin{array}{c|cccc}
 & \text{(Intercept)} & \text{treat1YES} & \text{treat2YES} & \text{treat1YES:treat2YES} \\
\hline
1 & 1 & 0 & 0 & 0 \\
2 & 1 & 0 & 0 & 0 \\
3 & 1 & 0 & 0 & 0 \\
4 & 1 & 1 & 0 & 0 \\
5 & 1 & 1 & 0 & 0 \\
6 & 1 & 1 & 0 & 0 \\
7 & 1 & 0 & 1 & 0 \\
8 & 1 & 0 & 1 & 0 \\
9 & 1 & 0 & 1 & 0 \\
10 & 1 & 1 & 1 & 1 \\
11 & 1 & 1 & 1 & 1 \\
12 & 1 & 1 & 1 & 1
\end{array}$$

Plot

DESeq2 model of gene expression

- Read count $K_{i,j} \sim \mathrm{NB}(\mu_{i,j}, \sigma_{i,j}^2)$, for gene *i* from sample *j*
  - $\mu_{i,j}$ = sample effects x sequencing effects
  - $\sigma_{i,j}^2 = \mu_{i,j}$ + gene-specific effects x $\mu_{i,j}^2$

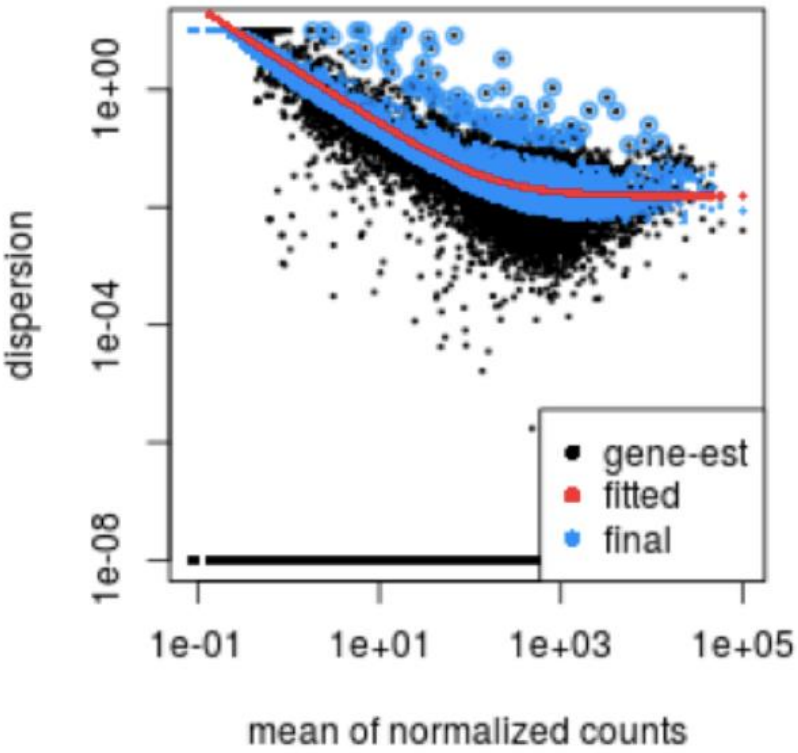
- Sequencing effects
  - Sequencing depth (sample-specific)
  - GC content (gene-specific)
  - Gene length (gene-specific)

DESeq2 model of gene expression

- Read count $K_{i,j} \sim \mathrm{NB}(\mu_{i,j}, \sigma_{i,j}^2)$, for gene *i* from sample *j*
  - $\mu_{i,j}$ = sample effects x sequencing effects
  - $\sigma_{i,j}^2 = \mu_{i,j}$ + gene-specific effects x $\mu_{i,j}^2$

- Gene-specific effects on variance
  - **Assumption**: Genes with similar expression should have similar variances
  - Regression of gene-specific effects versus $\mu_{i,j}$
  - Also called dispersion

Two-step Bayesian approach for dispersion fitting



- Dispersion = $\frac{\sigma_{i,j}^2 - \mu_{i,j}}{\mu_{i,j}^2} = \left(\frac{\sigma_{i,j}}{\mu_{i,j}}\right)^2 - \frac{1}{\mu_{i,j}}$
- For genes with high expression level,

$$\mathrm{Log}(\mathrm{Dispersion}) \approx 2 \cdot \mathrm{Log}\left(\frac{\sigma_{i,j}}{\mu_{i,j}}\right)$$

- Fit trend using local regression
  - Similar to moving average

Love, Huber, and Anders. Genome Biol. 15:550 (2014)

DE as a test of effect size

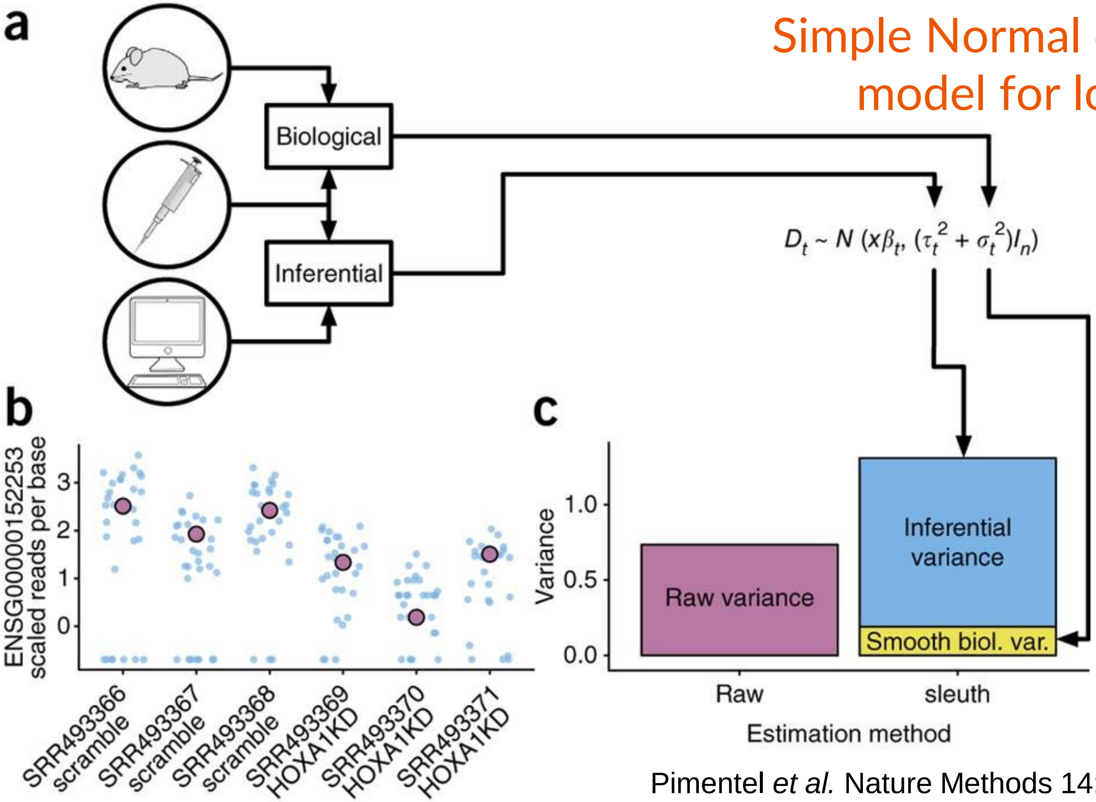
- Sample effects
  - Log FC = $\sum_r x_{j,r}\beta_{i,r}$ where $x_{j,r}$ are design parameters for sample *j* and β_{ir} are the effect sizes
- Wald test for each $\beta_{i,r}$: $\frac{\beta_{ir}}{\mathrm{SE}(\beta_{ir})} \sim$ Standard Normal

sleuth model for TPM

Making use of bootstrap to estimate variance

Simple Normal distribution model for log TPM

Technical variance estimates from bootstrapping



Pimentel *et al.* Nature Methods 14:687-690 (2017)

Normal distribution model for log TPM

- True expression: $y_{t,i} = x_i^T \beta_t + \varepsilon_{t,i}$ for sample *i* and transcript *t*
- Observed expression: $D_{t,i} = y_{t,i} + \zeta_{t,i}$

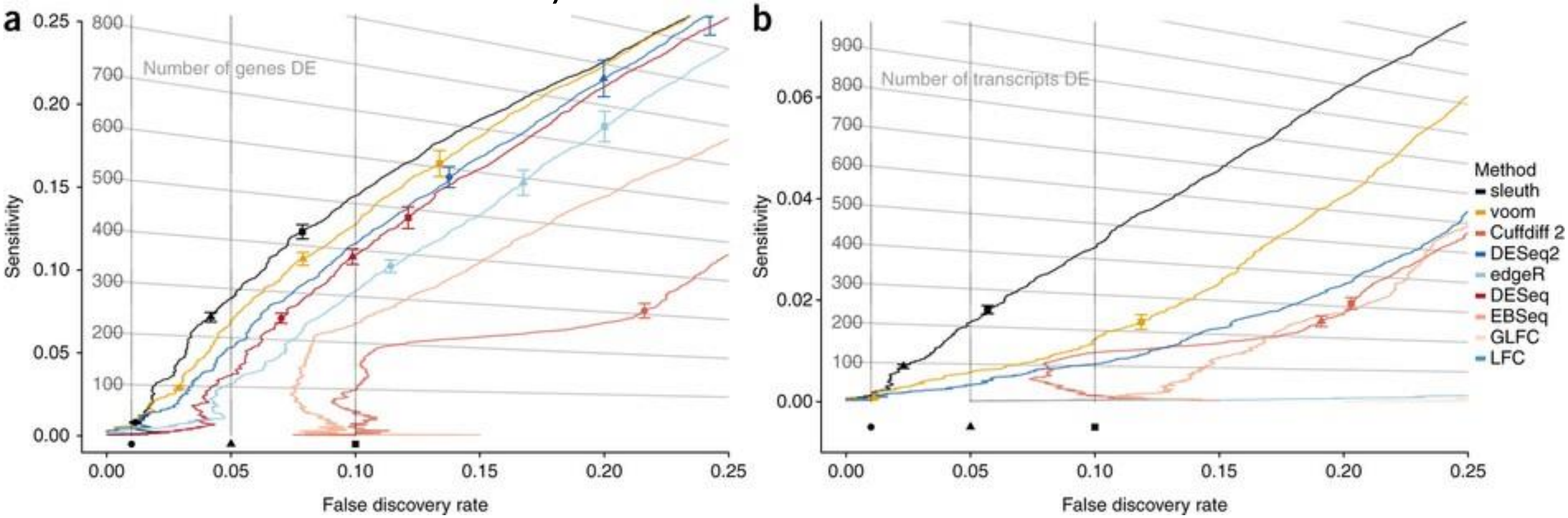
- Noises are normally distributed: $\varepsilon_{t,i} \sim N(0, \sigma_t^2)$ and $\zeta_{t,i} \sim N(0, \tau_t^2)$
  - Transcript-specific

- Full model: $D_t \sim N(x^T \beta_t, (\sigma_t^2 + \tau_t^2) I_n)$

Variance estimates improve sensitivity of DE

Gene-level analysis

Transcript-level analysis



Pimentel *et al.* Nature Methods 14:687-690 (2017)

Differential expression summary

- DE can be formulated in multiple ways but depend heavily on the model of gene expression distribution

- Read count model using Negative Binomial distribution
  - Bayesian update to improve the estimate of variance
  - Tied to genome-based pipeline: STAR

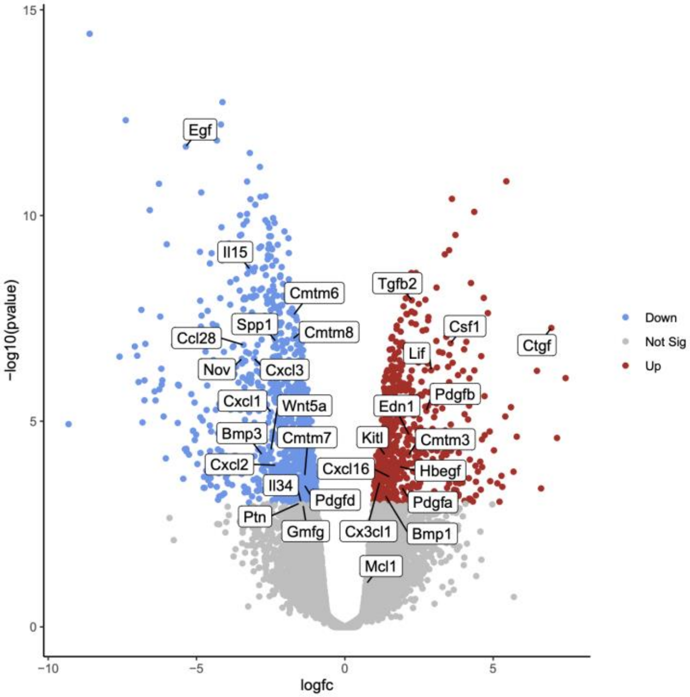
- Log TPM model using Normal distribution
  - Estimate technical variance directly using bootstrapping
  - Tied to transcriptome-based pipeline with *k*-mer pseudoalignment
  - kallisto / salmon

Any question?

Part II: Functional enrichment analysis

- From individual genes to sets of genes with common characteristics
- Overrepresentation = Fisher’s exact test = hypergeometric distribution
- Gene Set Enrichment Analysis (GSEA)
- Network topology-based

Differential expression result



- Statistical significance (p-values) and effect size (fold-changes)

- Do these genes correspond to specific biological characteristics?

https://training.galaxyproject.org/training-material/topics/transcriptomics/tutorials/rna-seq-viz-with-volcanoplot/tutorial.html

Overrepresentation analysis

Enrichment fold

Gene group	Kinase	Not kinase	Total
Differentially expressed	50	350	400
Not differentially expressed	150	5450	5600
Total	200	5800	6000

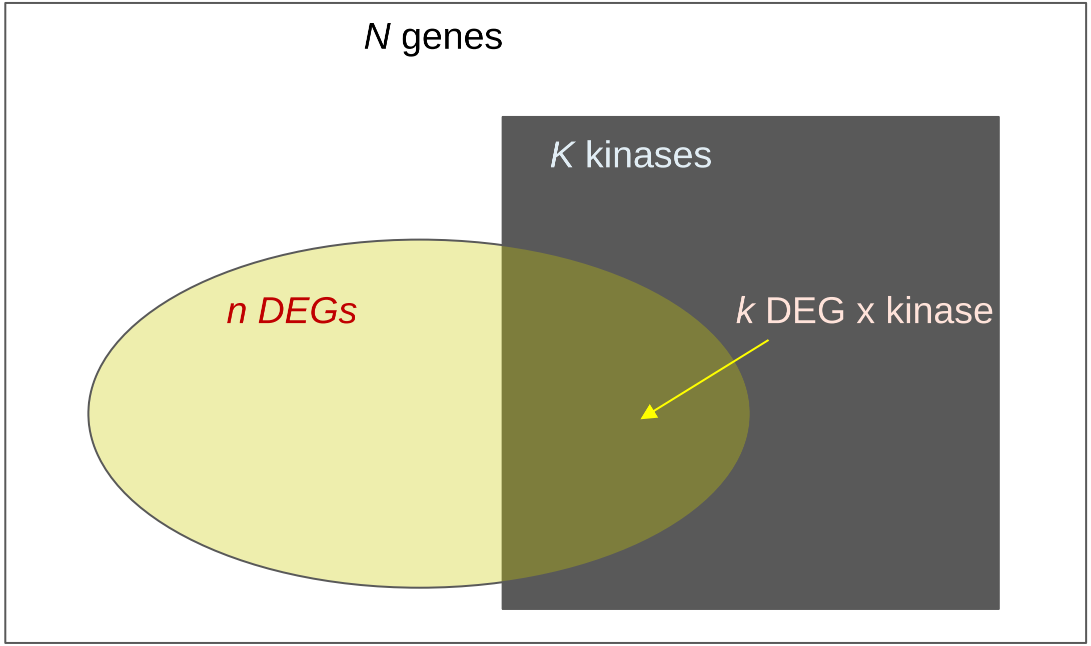
- There are 200 kinases among 6000 genes
- Expected 400 x 200 / 6000 = 13 kinases to be differentially expressed
- Enrichment = 50 / 13 = 3.85 folds

Fisher’s Exact Test

Gene group	Kinase	Not kinase	Total
Differentially expressed	$k \geq 50$	$400 - k$	400
Not differentially expressed	$200 - k$	$5400 + k$	5600
Total	200	5800	6000

- P-value for this observation = P(Kinase & DE ≥ 50)
- P(Kinase & DE = k) = Hypergeometric(N = 6000, K = 200, n = 400, k)

Hypergeometric distribution



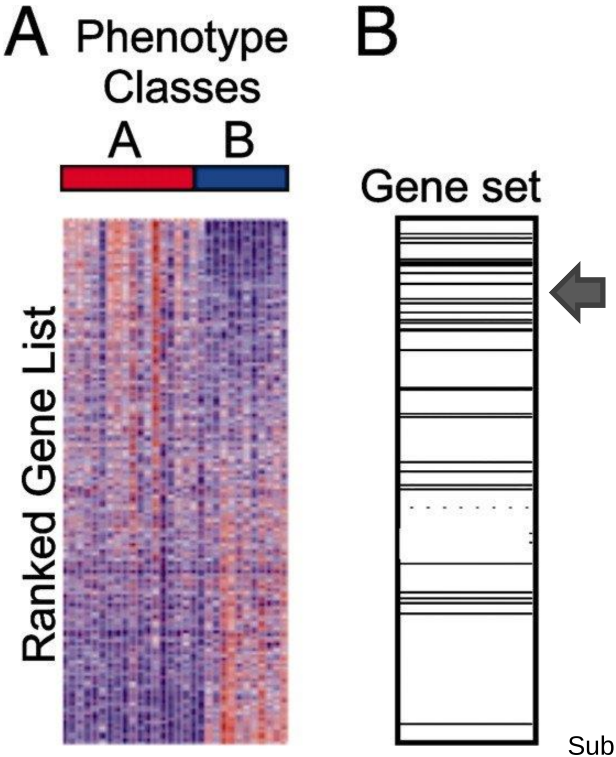
- $\binom{K}{k}$ ways to select the intersected *k* genes
- $\binom{N-K}{n-k}$ ways to select the remaining $n-k$ non-kinase genes

- Total of $\binom{N}{n}$ ways

- Probability = $\frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$

Gene Set Enrichment Analysis (GSEA)

GSEA algorithm sketch



- Sort genes by the extent of up-/down-regulation across conditions

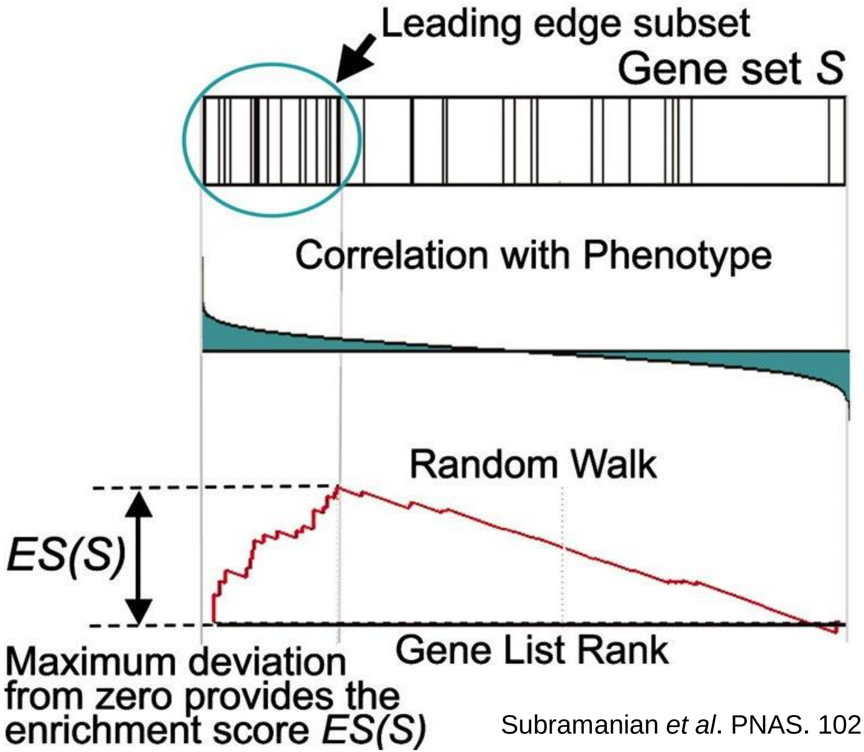
- Label genes annotated with a function

- If these genes are clustered together at the top, then this function is up-regulated

- If these genes are clustered together at the bottom, then this function is down-regulated

Subramanian *et al*. PNAS. 102:15545-15550 (2005)

GSEA scoring



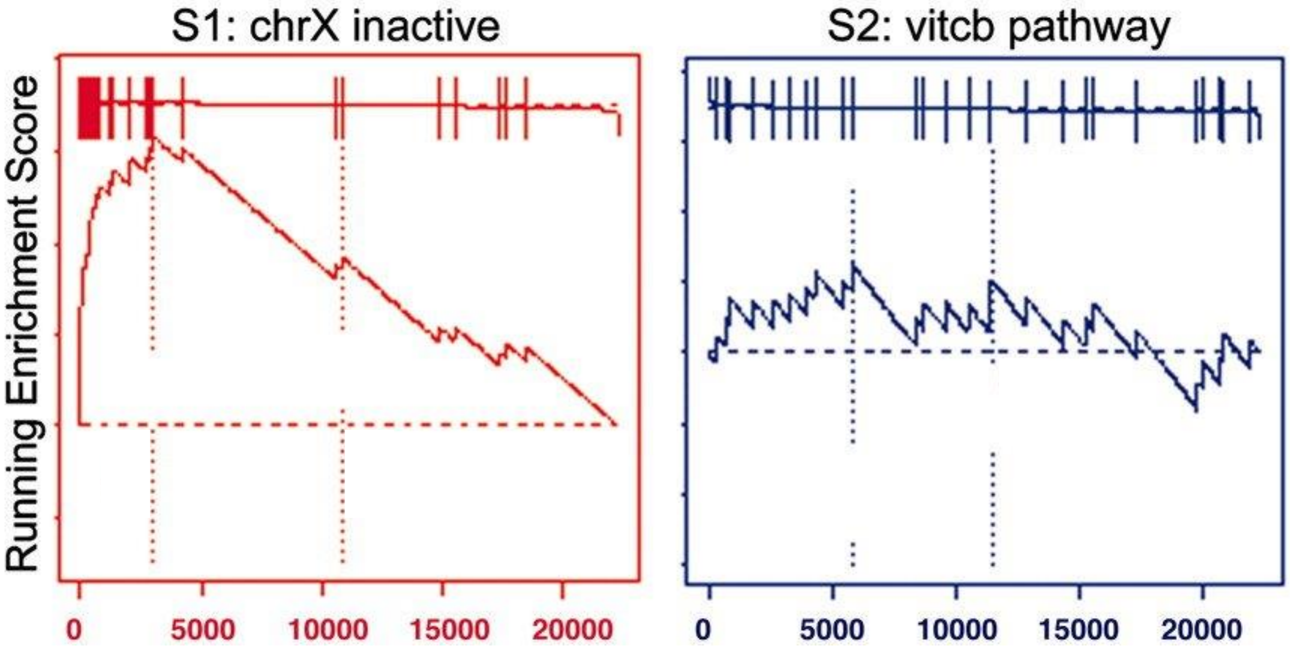
Subramanian *et al*. PNAS. 102:15545-15550 (2005)

- Starting at score = 0 from the top of the sorted gene list

- If encounter gene from *S*, +score
- Otherwise, −score

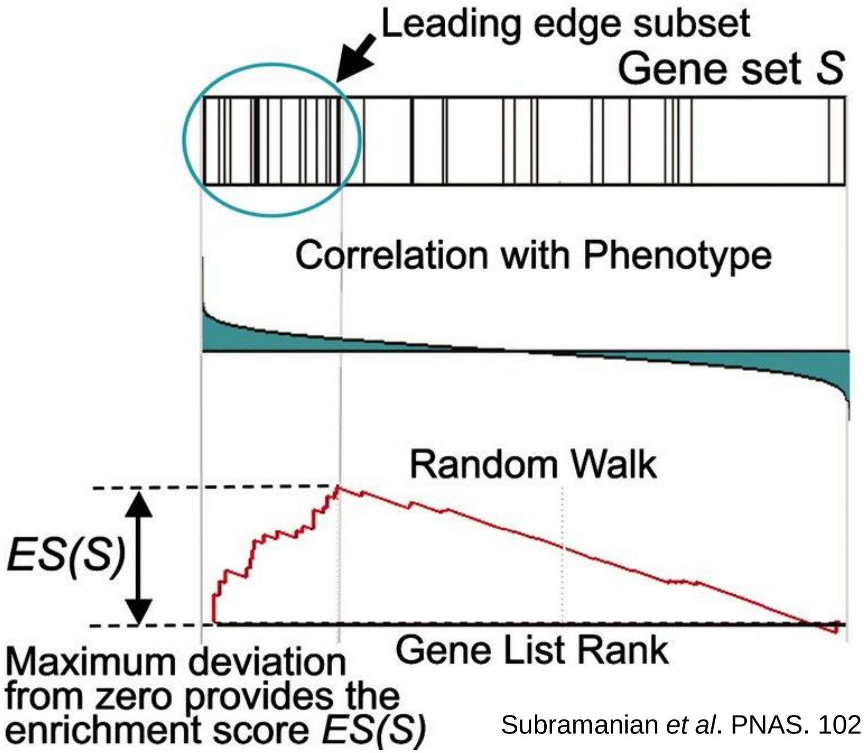
- Score indicates the extent of up-/down-regulation
  - Correlation with conditions
  - Log fold-change

Up-regulated versus unchanged pathways



Subramanian *et al*. PNAS. 102:15545-15550 (2005)

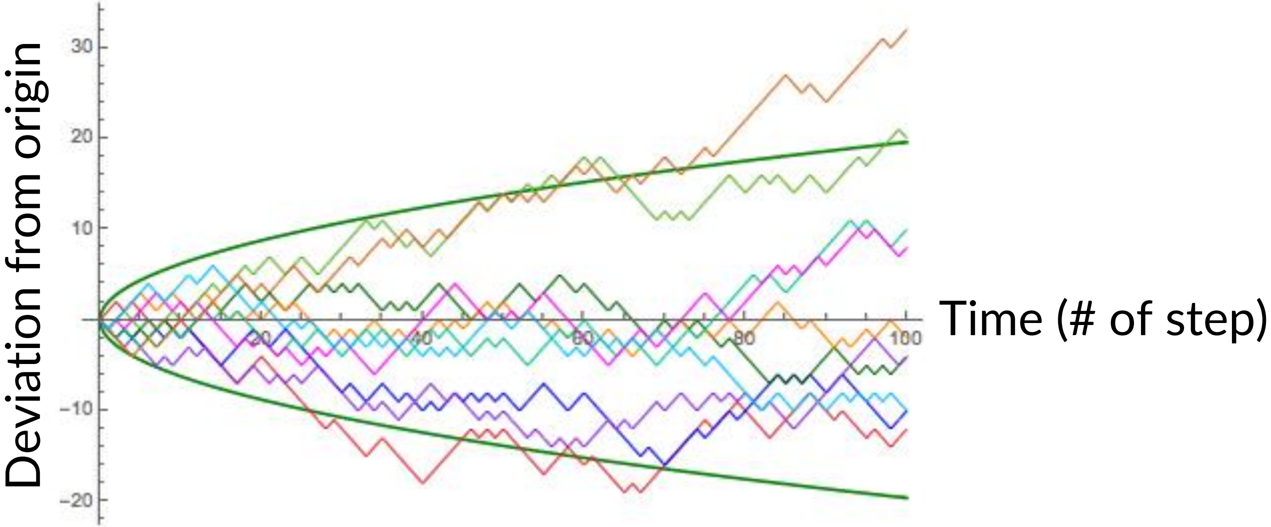
Null hypothesis for GSEA



Subramanian *et al*. PNAS. 102:15545-15550 (2005)

- **Null hypothesis**: Genes from *S* are uniformly distributed in the list
- +score and −score are uniformly distributed in the list
- This is a Random Walk

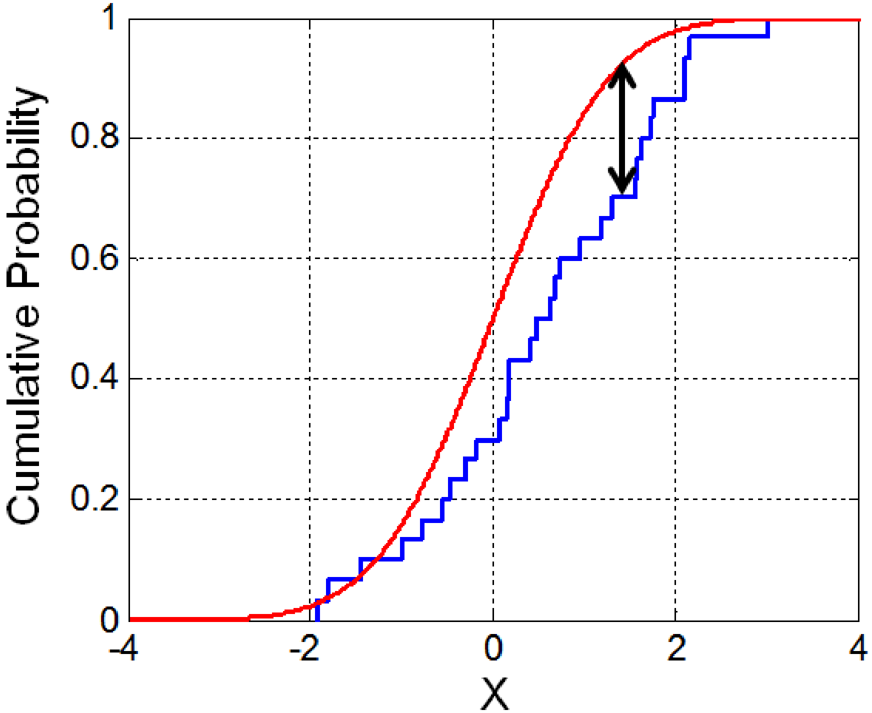
Statistical behaviors of random walks



https://demonstrations.wolfram.com/SimulatingTheSimpleRandomWalk/

- P(maximal deviation > d) $\approx 2\sum_{k=1}^{\infty}(-1)^{k-1}\, e^{-2(kd)^2}$

Kolmogorov-Smirnov test



https://en.wikipedia.org/wiki/Kolmogorov%E2%80%93Smirnov_test

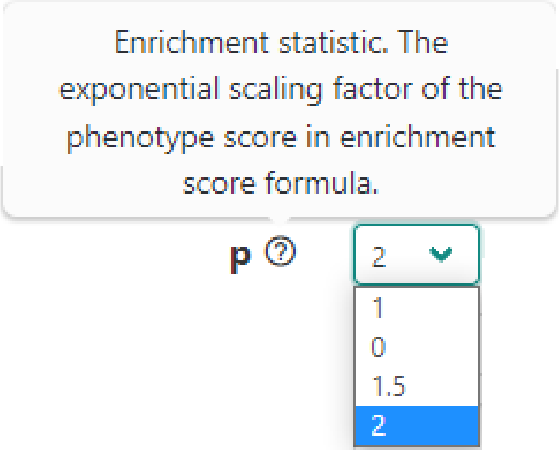
- Test whether two probability distribution are equal

- Compare cumulative density (red and blue trends)

- If they are equal, the two curves should stay close to each other

- **Null hypothesis:** random walk

Setting the score for GSEA



Enrichment statistic. The exponential scaling factor of the phenotype score in enrichment score formula.

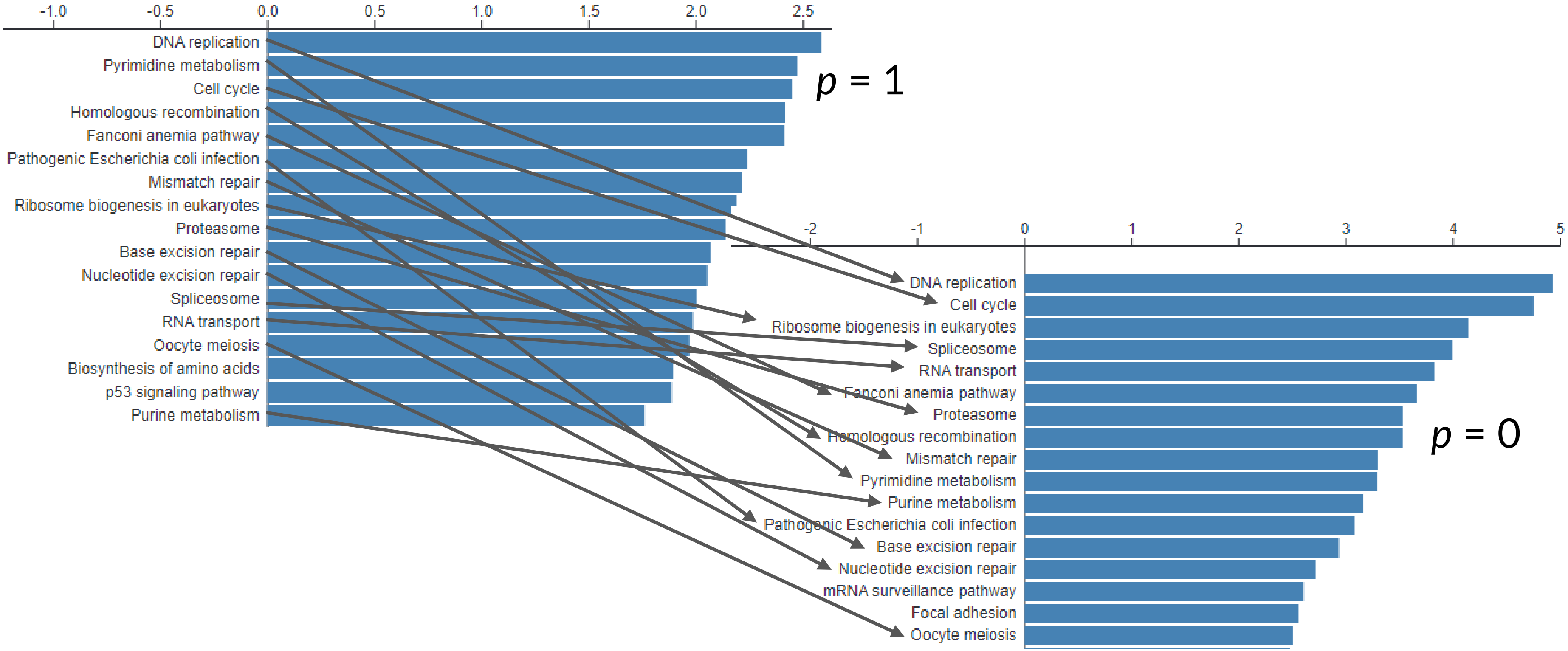
p ⓘ 2

- Originally developed for microarray data

- Adapted to RNA-seq
  - Log fold-change
  - No score (simply rank genes)

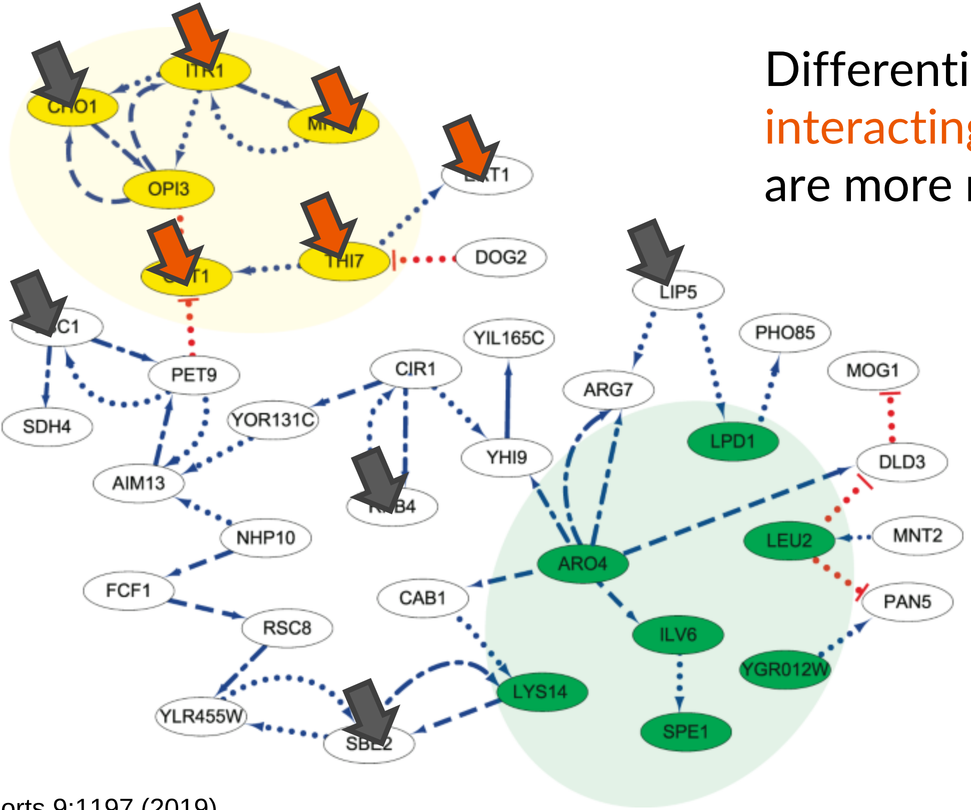
- Weighted score = (score)^{p}
  - Default: p = 1
  - No score: p = 0
  - More weights for top genes: p > 1

Comparing the impact of *p* = 0 and 1



Network topology-based analysis

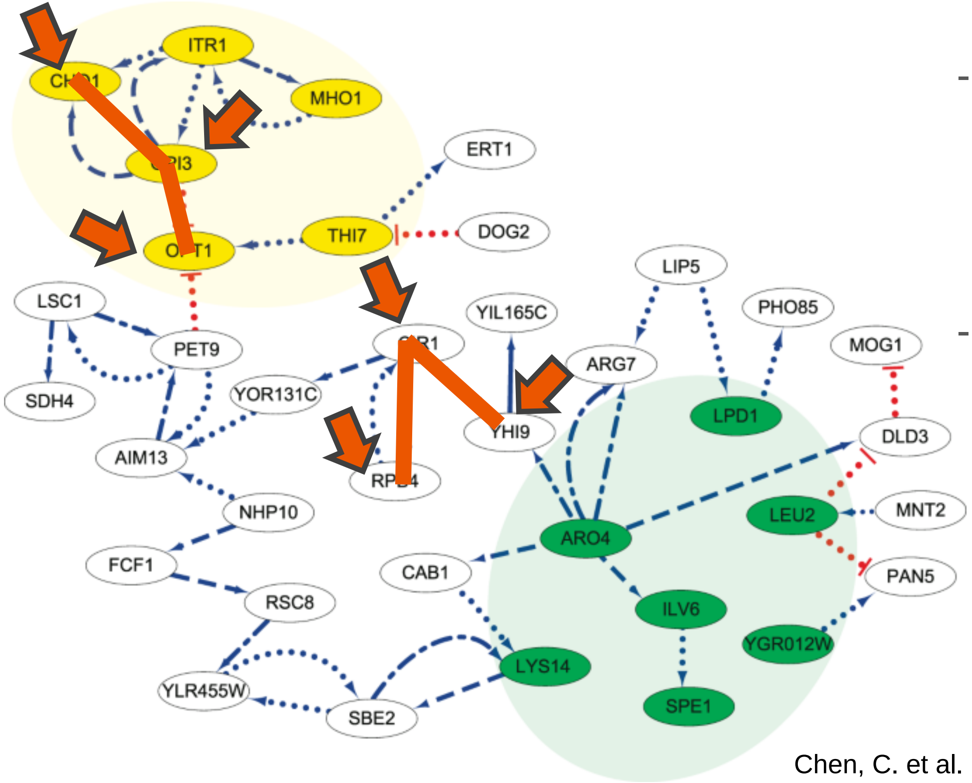
Gene and protein interaction networks



Differential expression of interacting genes/proteins are more meaningful

Chen, C. et al. Scientific Reports 9:1197 (2019)

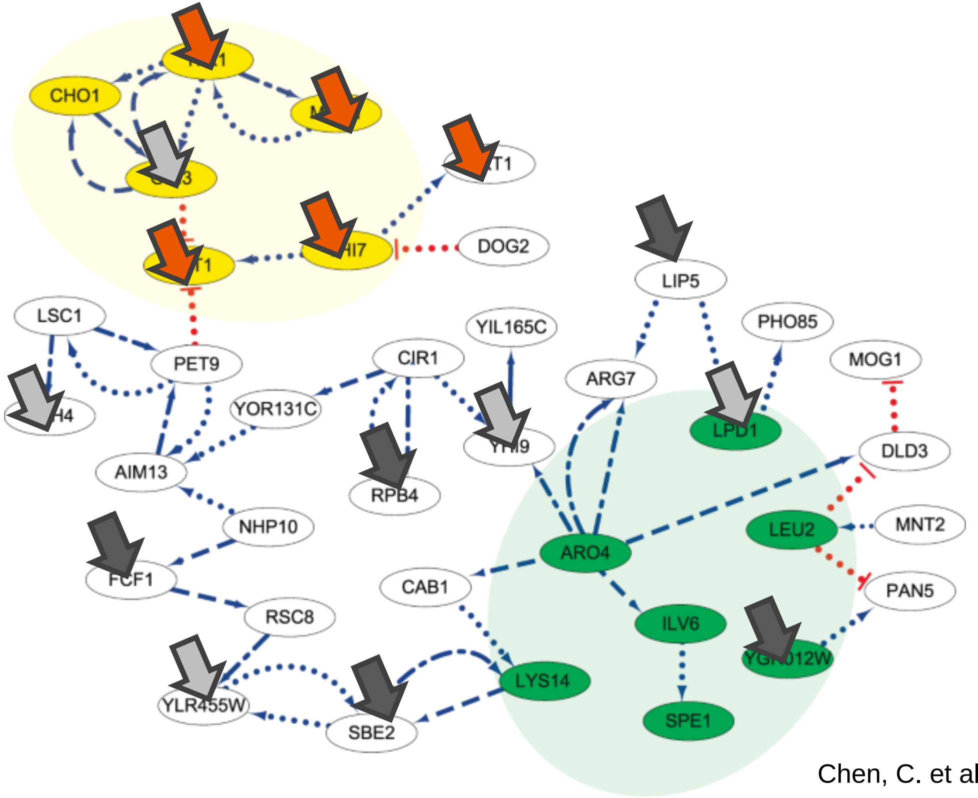
Network coherence scores



- Connectedness
  - Number of components
  - Number of edges
- Path length between genes
  - Unweighted
  - Weighted by fold changes

Chen, C. et al. Scientific Reports 9:1197 (2019)

Permutation test: Gene set



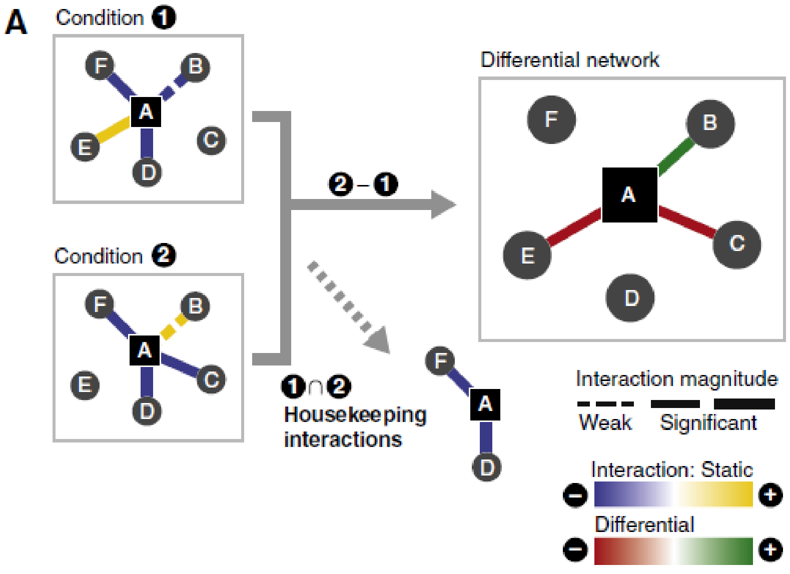
Chen, C. et al. Scientific Reports 9:1197 (2019)

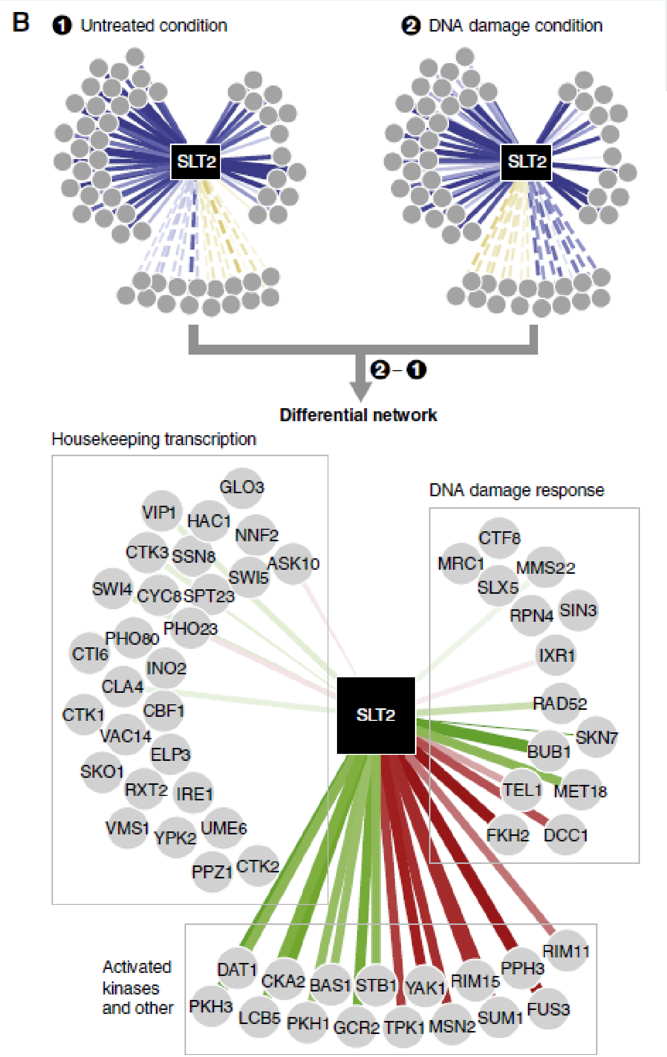
- Randomly select the same number of genes

- Recalculate network coherence scores

- P-value = fraction of samplings that the score is ≥ the original

Differential network





Ideker *et al*. Mol Syst Biol, 8:565 (2012)

- Detect gain/loss gene co-expression
- Unaffected interactions remain the same

Pros and cons

- Overrepresentation
  - Easy and fast to calculate
  - Depend on p-value cutoff

- GSEA
  - No p-value cutoff
  - Distinguish up- and down-regulated functions

- Network-based
  - Most biologically meaningful
  - Network data is incomplete

Any question?