מכון ויצמן למדע
WEIZMANN INSTITUTE OF SCIENCE

Course Summary

Dena Leshkowitz

An Introduction to Deep-Sequencing Data Analysis For Biologists- 2020

nature REVIEWS GENETICS

nature.com ▸ journal home ▸ archive ▸ issue ▸ review ▸ full text

NATURE REVIEWS GENETICS | REVIEW

ARTICLE SERIES: Applications of next-generation sequencing

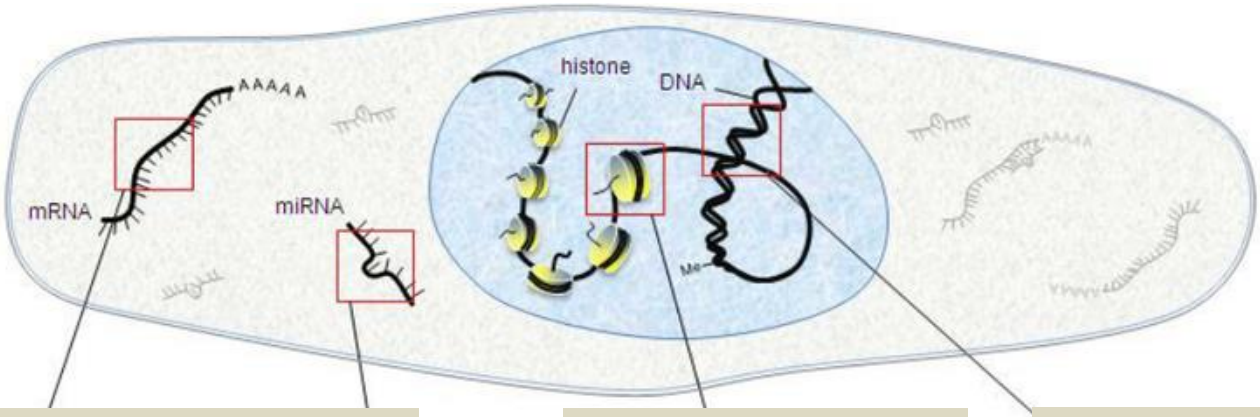
Coming of age: ten years of next-generation sequencing technologies

Sara Goodwin, John D. McPherson & W. Richard McCombie

Affiliations | **Corresponding author**

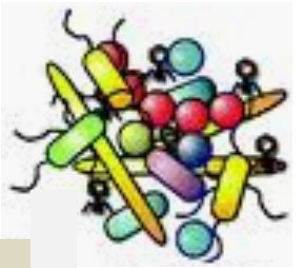
Nature Reviews Genetics **17**, 333–351 (2016) | doi:10.1038/nrg.2016.49
Published online 17 May 2016

Deep Sequencing Applications



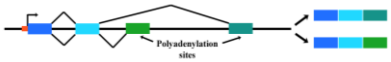
Metagenome

Analyze all genetic material in complex samples



Transcriptome

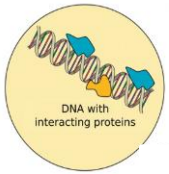
mRNA or microRNA discovery and differential expression analysis



Alternative splicing Analysis

Epigenome

- Identification of open chromatin (ATAC-Seq)
- Protein-DNA interactions (ChIP-Seq)



Genome

Mutation discovery

…AACTGGTAC…
…AACTCGTAC …
…AACTGGTAC …
…AACTGGTAC …

scRNA-Seq

Modified DOI: 10.3390/biology2010378

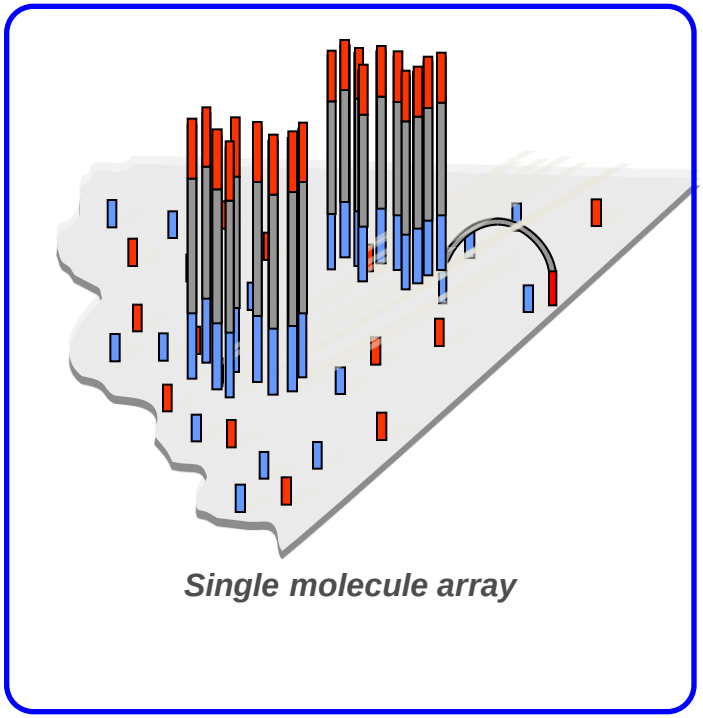
Platforms/collection of tools	UTAP					
Map to Genome	Bowtie	BWA				
RNA-Seq	Mappping: TopHat, STAR	Transcript Assembly: cufflinks, stringtie	Gene level Quantification & differential expression : HTSeq/STAR & DESeq2	Transcript level Quantification & differential expression : Tuxado suite (cufflinks, cuffdiff)	Transcript level Quantification & differential expression : Stringtie & Ballgown	DEXSeq
ChIP /ATAC-Seq	Peak calling: MACS2	Downstream: CEAS, GREAT, Meme-ChIP, Homer	DiffBind			
Functional Analysis	Metascape	GeneAnalytics	https://bbcunit.atlassian.net/wiki/pages/viewpage.action?pageId=27820050			
Genome Browser	IGV					
Clustering	iDEP.90: Hierarchical, KMeans, PCA					
Genetic Variation	GATK	Picard	Annovar			
Single Cell	CellRanger	Seurat	PanglaoDB	GeneAnalytics		
Metagenomics	QIIME					

Summarize the course in Q & A session

You Are All Invited!

Illumina Sequencing- Question 1

- Why do we need to amplify fragments on the flowcell?



Answer

- In order to be able to detect a strong enough emission signal of the dyes.

Illumina Sequencing –Question 2

- Why is there a limitation on the sequence length i.e. why do we sequence short sequences with the Illumina sequencing machines?

Answer

- Due to the drop in quality of sequence that is mostly due to the accumulation of phasing-prephasing in the previous cycles.

Illumina Sequencing –Question 3

- Why do the fragments we sequence need to be of size 200-500 bases?

Answer

- The bridging in the cluster step is sensitive to the fragment size.
- Shorter fragments will not bridge well.
- Longer fragments will cause the clusters not to be separated well from each other.

Illumina Sequencing – Question 4

- Can we read a sequence in NextSeq that starts with GGGGG?

Answer

- We can not read a fragment that starts with GGGGG since G do not produce a signal in NextSeq and therefore the location of the cluster (that is done using the signal in the first five cycles) can’t be determined.

Sequencing - Question 5

In which applications will we prefer to perform sequencing of PE 100 and in which **SE 50** :


Sequencing an unknown bacteria

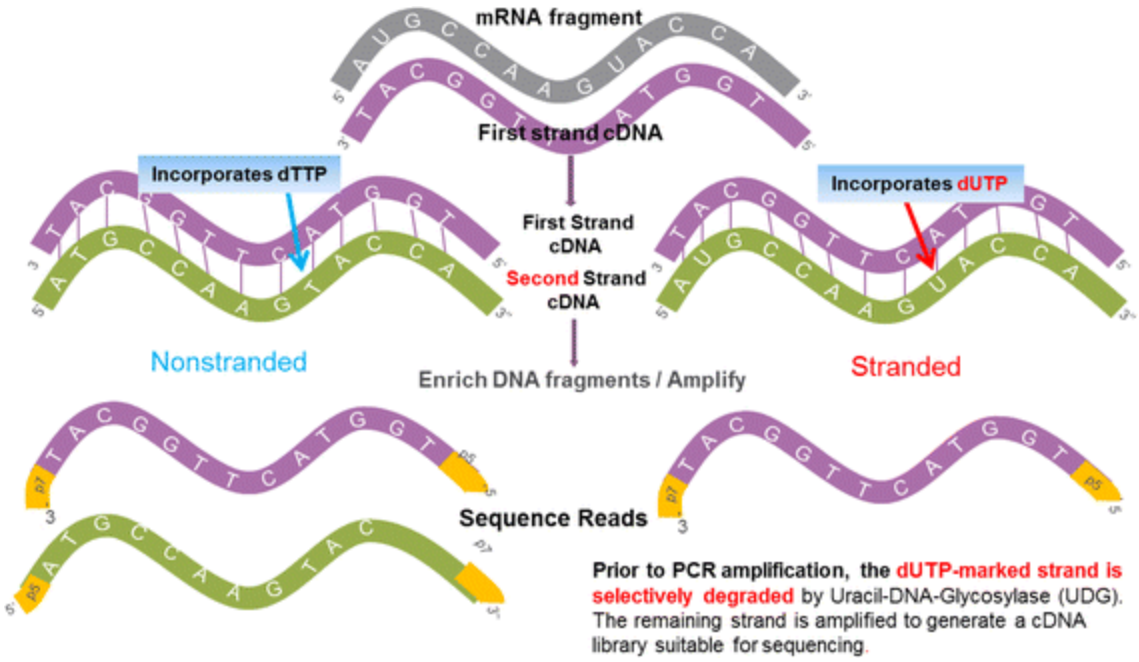

Human disease related genetic variation study


RNA-Seq of mouse samples


Small RNA-Seq

Sequencing - Question 6

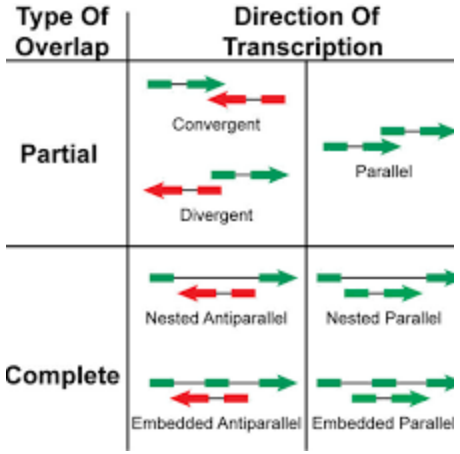
Which RNA-Seq experiment will we prefer to perform using stranded RNA-Seq protocol?



Zhao et al. BMC Genomics volume 16, Article number: 675 (2015)

Answer

- Discriminate between overlapping genes



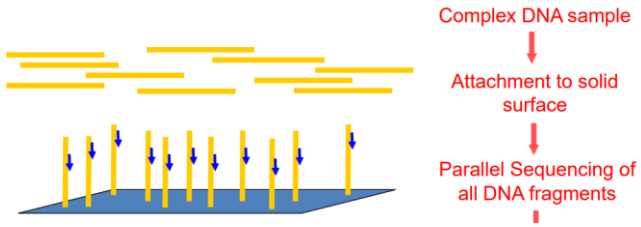
- Define new transcripts

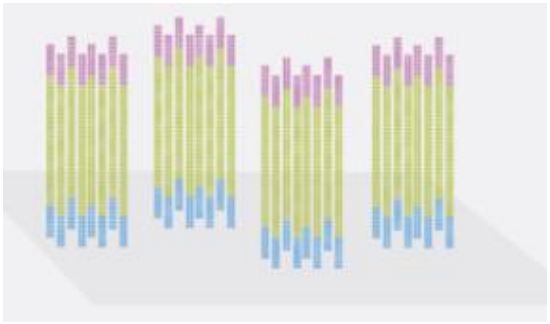
Sequencing Question 7

- Name two technological requirements for massive and parallel Illumina sequencing.

Answer

1. Attachment to a solid surface, to keep track of each of sequence independently by determining a physical location for each sequence





2. Amplify in situ to detect the fluorescence signal

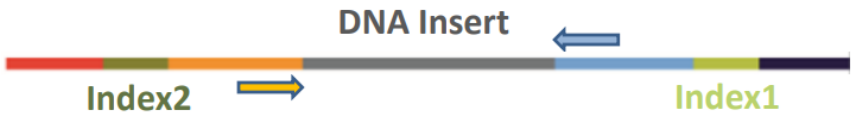
Sequencing Question 8

- What elements should be in an Illumina library so we can sequence it and get results?

Answer

1. Two different adaptors in the 5’ and 3’ (to bind to the flow cell and hybridize primers for the sequencing)
2. DNA insert (the sequence you are interested in sequencing)
3. One or two barcodes if you intend to multiplex your samples

Sequencing Paired End Libraries with Dual Index Read

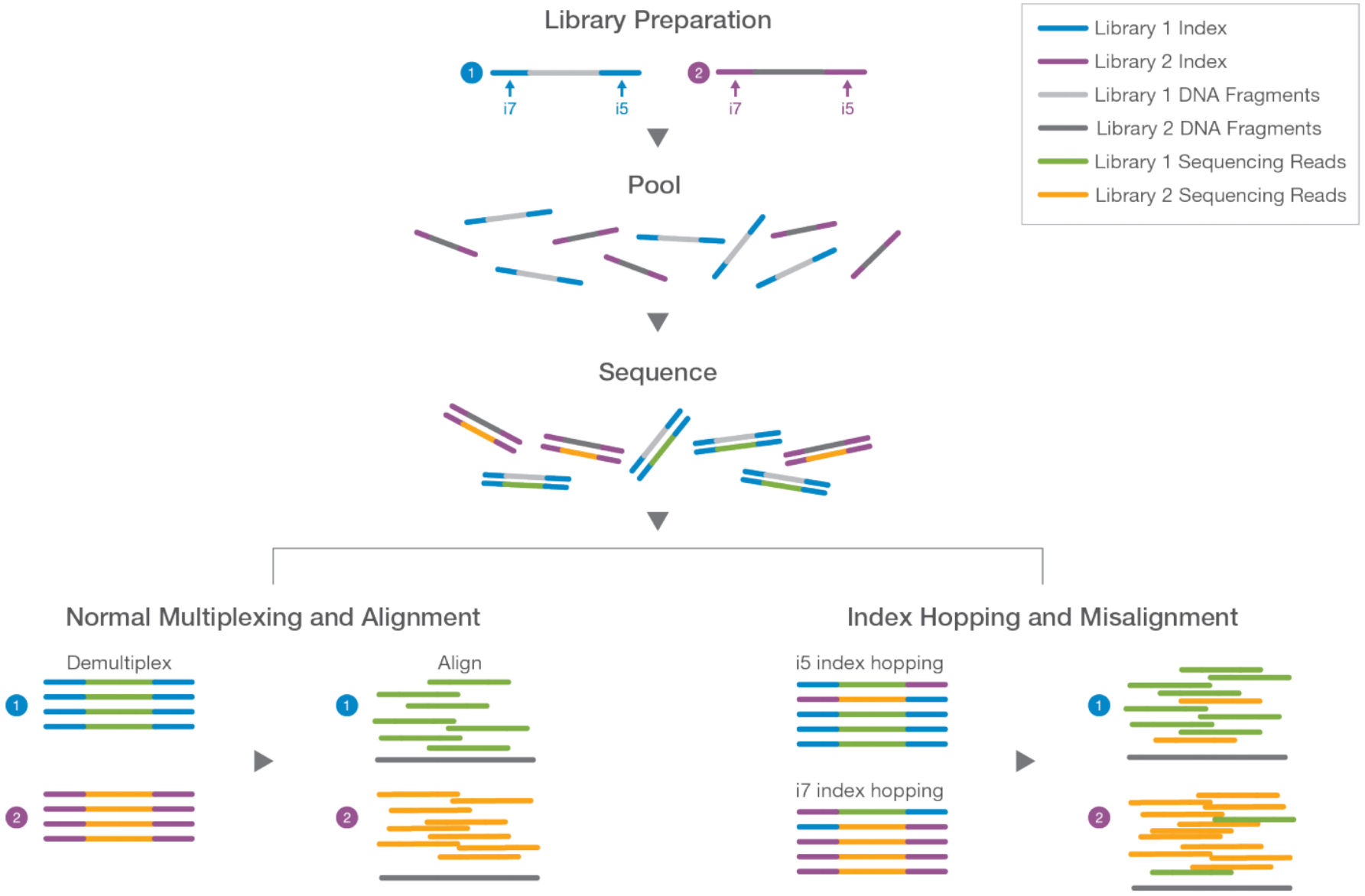


Sequencing Question 9

- What are the differences between random and patterned flow cells?

Answer

1. Random positioning of oligos on the surface versus oligos only inside nanowells (patterned)
2. Higher density of clusters in the patterned flow cell
3. Percent of passed filter clusters is higher in the random flow cell
4. Higher yield in the patterned flow cell
5. Index hopping in the patterned flow cell

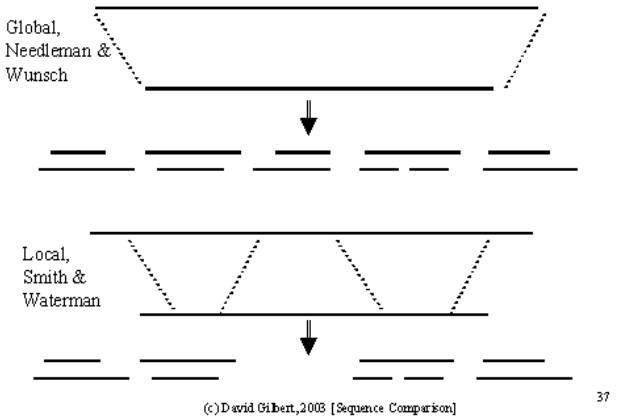

Library Preparation
1
i7
i5
2
i7
i5
Pool
Sequence
Library 1 Index
Library 2 Index
Library 1 DNA Fragments
Library 2 DNA Fragments
Library 1 Sequencing Reads
Library 2 Sequencing Reads
Normal Multiplexing and Alignment
Demultiplex
Align
Index Hopping and Misalignment
i5 index hopping
i7 index hopping

Mapping Question 1

What is the difference between local and global alignment?

Answer

- **Global alignments** attempt to align every residue in every sequence
- **Local alignments** look for regions of similarity or similar sequence motifs within their larger sequence context



Mapping – Question 2

- What are the main differences between BLAST vs Bowtie aligner?

Answer

- Speed: Bowtie is faster
- By default NGS aligners are stringent in their alignment algorithm, while BLAST was designed to find weak similarities
- Bowtie can align only short reads

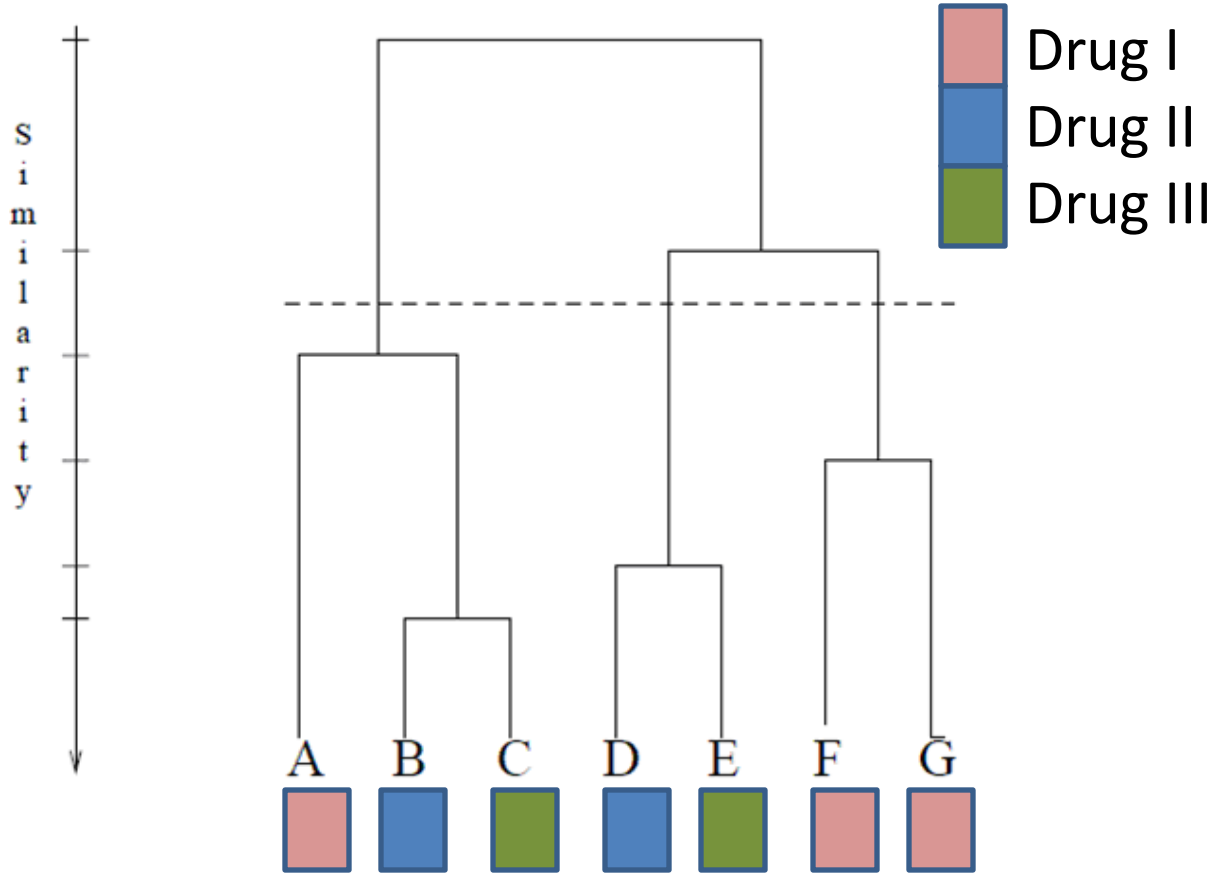
Mapping - Question 3

In which applications do we perform de novo assembly and in which mapping to a genome:

- DNA-Seq of an unknown bacteria
- RNA-Seq of mouse samples
- RNA-Seq of non-model organism

RNA-Seq – question 1

What can be the reason for the following result when performing hierarchical clustering of samples treated with three drugs (I, II and III)?



Answer

- Batch affect stronger than the true biological variance between the conditions
- Or that the drugs have no biological effect

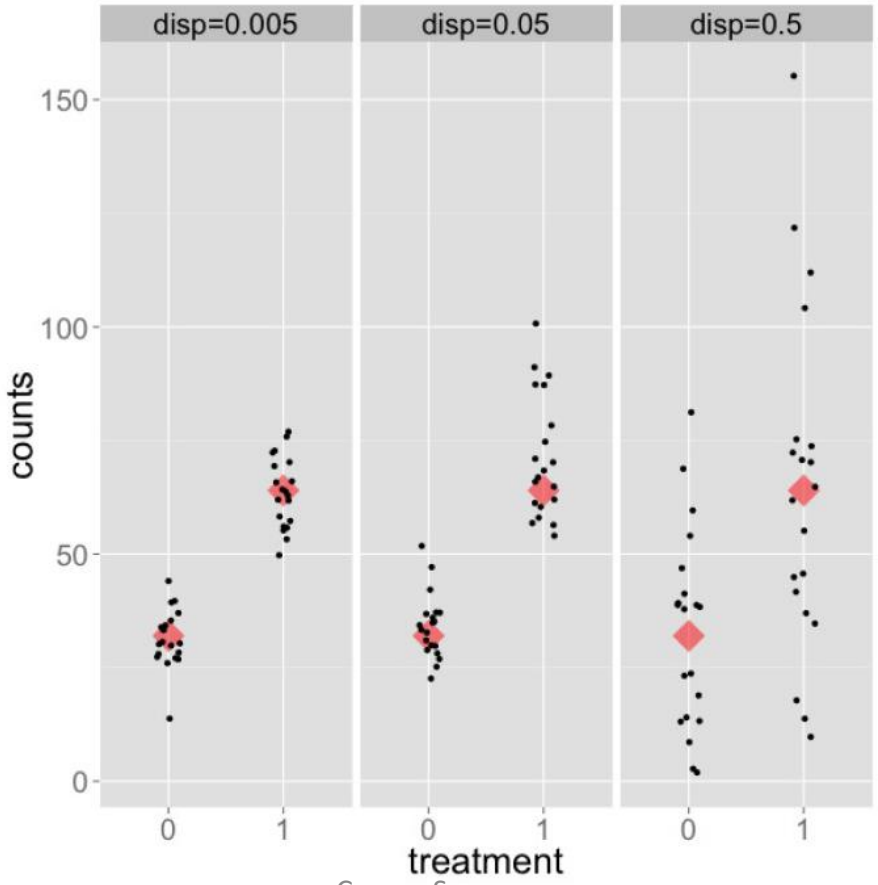
RNA-Seq – question 2

- You performed a differential expression analysis and found a gene with a fold change value of 2 but with a p value of 0.3, what could be the reason for the high p value?

Answer

The within-group variation is high

Or a low read count



RNA-Seq – question 3

- You are planning an experiment and can not process all your samples together but only 6 samples at once (a batch)
- You have a cell culture whom you treat with different drugs (5 different drugs + control with no drug) and plan to have three replicates of each (total 18 samples) .
- Explain which samples you would take for each batch

Answer

- Each batch should include all the drugs as well as the control. Label the samples from which batch they were derived.

Clustering Question 1

- Are hierarchical and k-means clustering a supervised analysis?

Answer

No

- Discriminant analysis (supervised, i.e. ANOVA)

CLASSES KNOWN

- Cluster analysis (unsupervised)

CLASSES NOT KNOWN

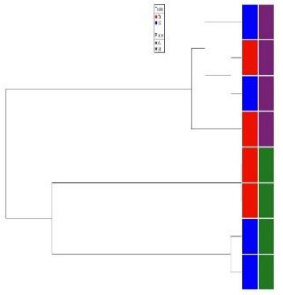
Clustering Question 2

- Which are frequent uses of clustering in gene expression experiments?

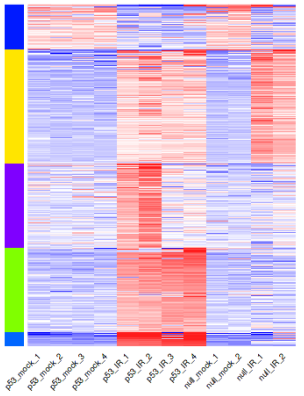
Answer

- Exploratory analysis

  Use all genes to estimate the similarity between samples



- Find common expression patterns of DE genes and visualize



Functional analysis – Question 1

- What are the differences between GO annotation and pathways?

Answer

- A GO term indicates a relationship between the gene product and the function, process or cellular component.
- Pathways (such as represented KEGG db) describe molecular interactions and reactions allowing for mass flux (metabolism) or information flux (signal transduction).
- The GO system is built as an acyclic graph (3 graphs, one for each subsection) while pathways are networks.

Functional analysis – Question 2

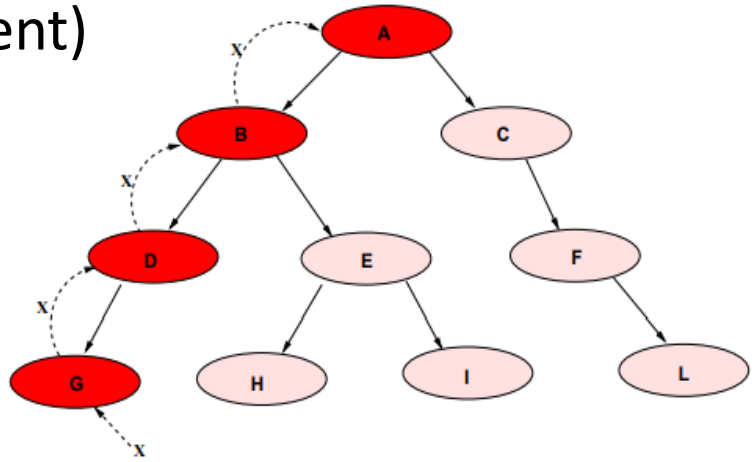
Question:
True or False?

If a child term describes the gene product, then all its ancestors (parent) terms must also apply to that gene product.

Answer: True

Every GO term obeys “the true path rule”:

If a child term describes the gene product, then all its ancestors (parent) terms must also apply to that gene product.



Functional analysis – Question 3

Question:

Give an example of a gene attribute that cannot be described using GO?

What is not GO?

Answer:

- Gene products: e.g. cytochrome c is not in the ontologies, but attributes of cytochrome c, such as oxidoreductase activity, are
- Processes, functions or components that are unique to mutants or diseases: e.g. oncogenesis
- Attributes of sequence such as intron/exon parameters
- Protein domains or structural features
- Protein-protein interactions
- Environment, evolution and expression
- A pathway

Functional analysis – Question 4

You analyzed the same gene list using different pathway enrichment software and got different results.
What could be the reasons for that?

Answer

- Pathway analysis tools differ from each other because of:
- **Database** (different functional categories, organism and cell types, level of database **curation**)
- **Statistical** methodology for calculating enrichment (control thresholds, FDR, gene background, term size)