Canadian Centre for Computational Genomics

C3G Analysis Workshop: RNA-Seq

Day2: Counts & Differential Expression Analysis

22-23 January 2019

Learning objectives

- **Objectives:**
  - **Understand the GenPipes workflow** and how steps relate to each other
  - **Understand the theory** behind each step
  - Be aware of the **differences between gene and transcript level analysis**
  - Know the **different outputs** produced by the pipeline

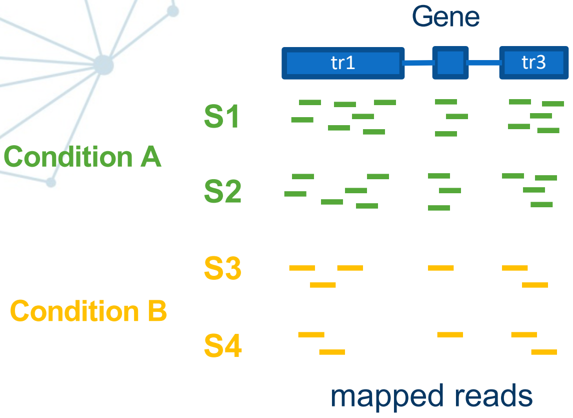
Differential expression analysis?

- What? The **read count at the gene and/or transcript level** in two conditions.
- Why? To identify genes/transcripts that may **play a role in differentiating the groups**.
- How? By **counting the number of reads** assigned to each gene/transcript and by **comparing their average.**

The assumption is that the **number of reads produced** by each gene/transcript is **proportional to its abundance**

There are 3 main steps to the analysis…

Canadian Centre for Computational Genomics



Counts

1

Gene level

	S1	S2	S3	S4
gene1	55	48	12	6
gene2	104	102	247	263
…	…	…	…	…

Transcript level

	S1	S2	S3	S4
tr1	23	17	12	6
tr2	5	6	3	2
…	…	…	…	…

2 Normalization & Filtering

	S1	S2	S3	S4
gene1	53.2	49.1	11.6	5.9
…	…	…	…	…

3

Statistical testing & Multiple testing correction

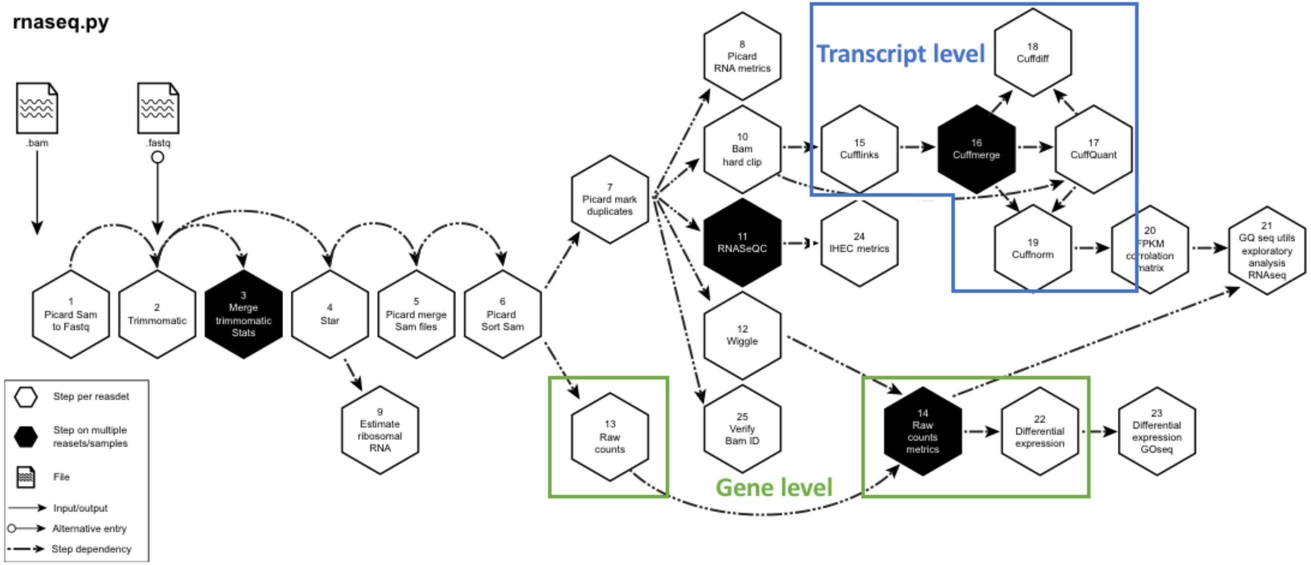
Condition A vs Condition B

	FC	logFC	Pvalue	FDR
gene1	-5	-2.3	0.0012	0.03
…		…	…	…

GenPipes performs both gene level and transcript level analyses

Gene level:
steps 13,14,22

Transcript level:
steps 15,16,17,18,19



RNA-seq data are challenging to analyze

- Technical and biological **variability**
- **Biases**: sequencing depth, composition bias
- **Spliced alignments**, transcript deconvolution
- **Complex** statistical models, low sample size
- **Large amount** of data
- Computationally **intensive**

Part1:
Read counts

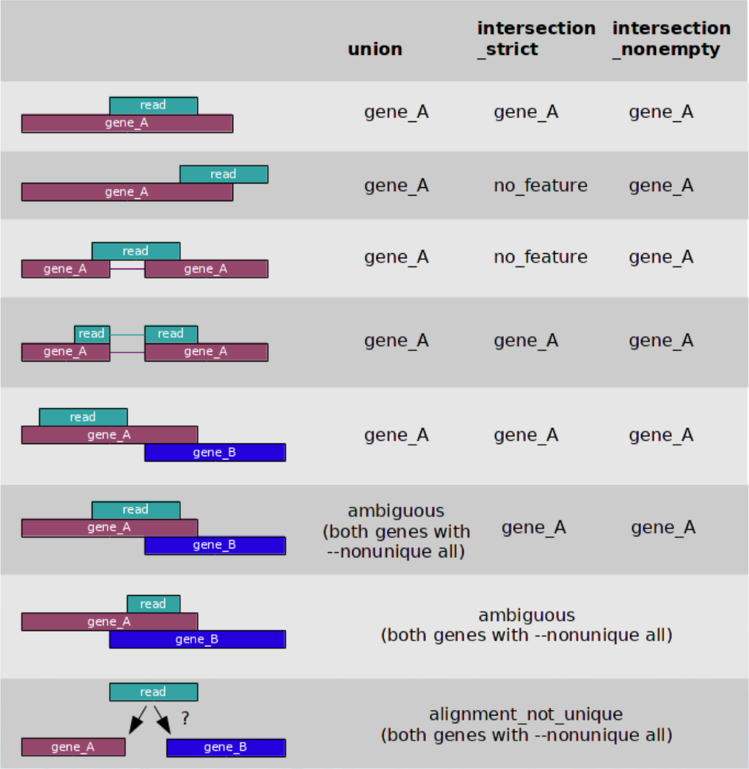
Reads can be assigned to genes or transcripts

- Gene level:
  - count reads falling in genes
  - **HTSeq*, **featureCounts,…
- Transcript level:
  - assign reads to transcripts; more **complex** than for genes!
  - RSEM, StringTie, **Cufflinks package***, Kallisto, Salmon,…

***used by GenPipes**

HTSeq counts the reads falling into coding regions

HTSeq



	union	intersection _strict	intersection _nonempty
read gene_A	gene_A	gene_A	gene_A
read gene_A	gene_A	no_feature	gene_A
read gene_A gene_A	gene_A	no_feature	gene_A
read read gene_A gene_A	gene_A	gene_A	gene_A
read gene_A gene_B	gene_A	gene_A	gene_A
read gene_A gene_B	ambiguous (both genes with --nonunique all)	gene_A	gene_A
read gene_A gene_B	ambiguous (both genes with --nonunique all)		
read ? gene_A gene_B	alignment_not_unique (both genes with --nonunique all)		

http://htseq.readthedocs.io/en/master/count.html

Count reads overlapping genes

rawCountMatrix.tsv

	s1	s2	…
gene1	12	15	
gene2	0	2	
gene3	1643	1352	
…			

Transcript level expression is difficult to calculate

- Genes can have **multiple alternative splicing events** and there is an **unknown number of isoforms**.
- Many possible ways to **reconstruct the gene model** from the data.
- Reads are assigned to an isoforms using **probabilistic methods**.

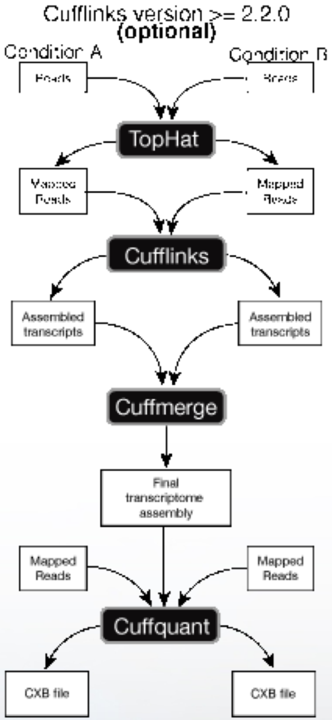
The Cufflinks suite allows transcript level expression

- **Cufflinks suite** includes a number of different programs that work together to **perform transcript level analysis**
- Cufflinks (the program) performs the **transcriptome assembly**
- Cuffmerge creates a **meta-assembly**
- Cuffquant **quantifies transcript expression**

Transcriptome assembly

Meta-assembly

Quantification



http://cole-trapnell-lab.github.io/cufflinks/manual/

Cufflinks assembles the transcriptome

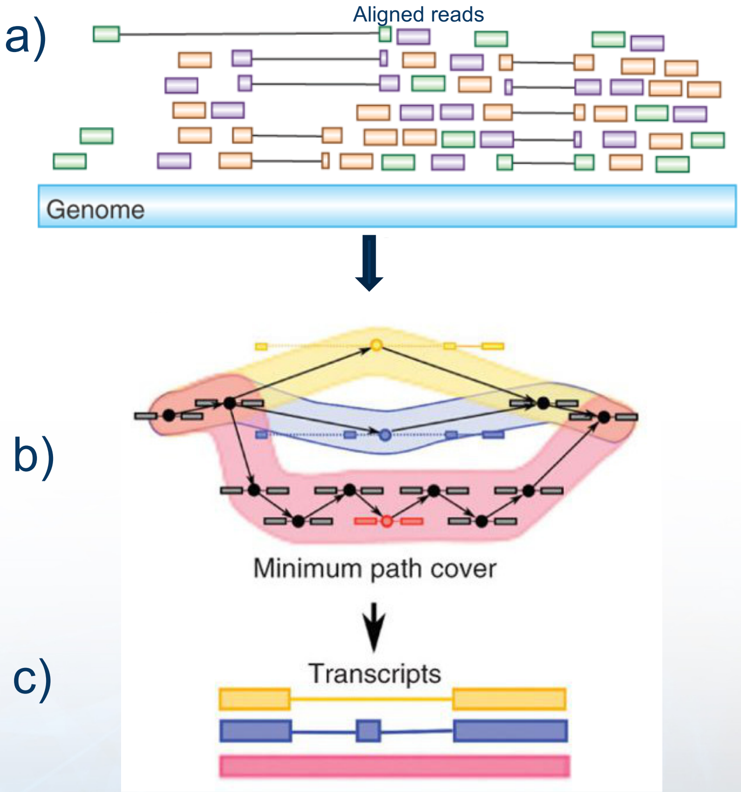
Cufflinks takes the aligned reads and inputs a model of the transcript profile: that’s the **transcriptome assembly**.

a) Cufflinks first **regroups reads into ‘bundles’** of overlapping reads.

b) **Reads are connected** in an ‘overall graph’, forming paths.

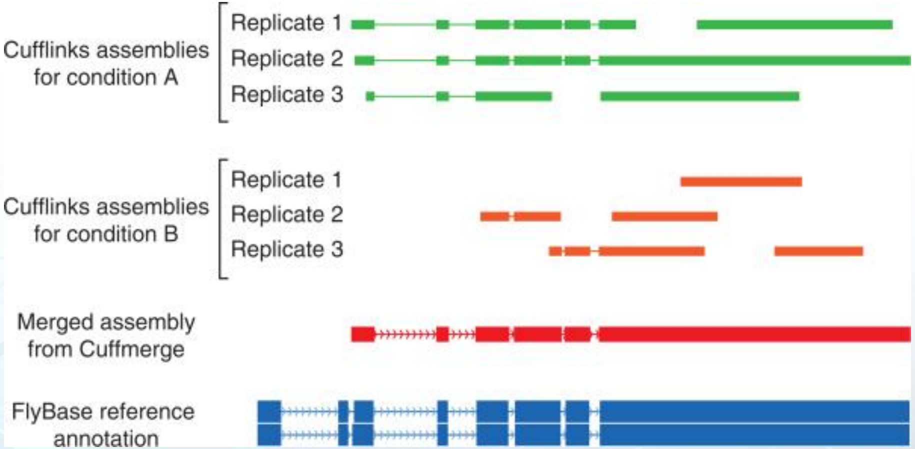
c) Complete path are **merged to form the isoforms.**

https://home.cc.umanitoba.ca/~frist/PLNT7690/lec12/lec12.3.html



Cuffmerge creates a meta-assembly

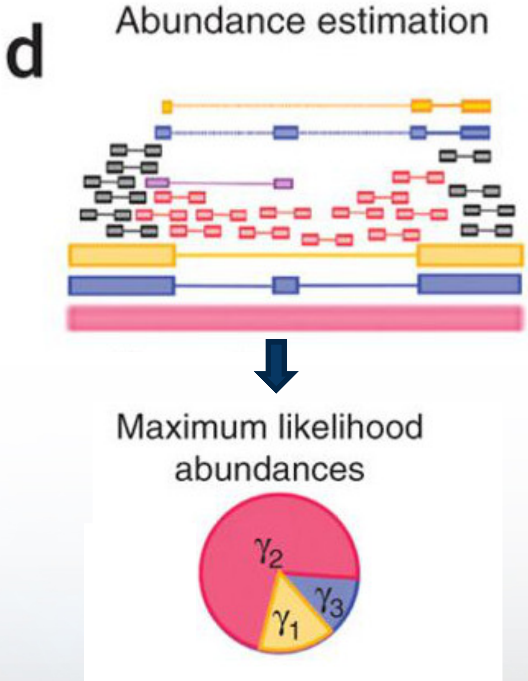
- Merge assemblies to create single **merged transcriptome annotation**
  - Genes with low expression don't permit full reconstruction in each sample => merging often **recovers complete gene**
  - **Newly discovered isoforms** integrated with known ones to form more complete gene model


Cufflinks assemblies for condition A
Replicate 1
Replicate 2
Replicate 3
Cufflinks assemblies for condition B
Replicate 1
Replicate 2
Replicate 3
Merged assembly from Cuffmerge
FlyBase reference annotation

https://www.nature.com/articles/nprot.2012.016

Cuffquant quantifies the expression

- **Quantifying gene and transcript level expression** for known and novel transcripts
- Fragments are **matched to the transcripts** from which they could have originated.
- Estimates **transcript abundances** using a statistical model.

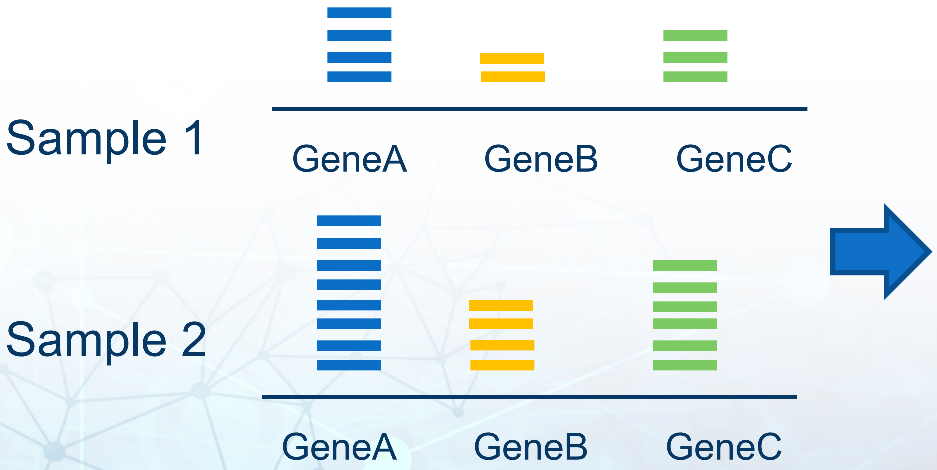


https://www.nature.com/articles/nbt.1621/figures/1

Part 2:
Normalization and filtering

Library size affects the number of counts

- There are several factors influencing the read counts. We are mostly concerned with **sample-specific effects.**
- The most common bias is coming from differences in library size.
- Samples have different number of total reads: the **number of reads assigned to a gene is dependent on the total number** of reads generated.

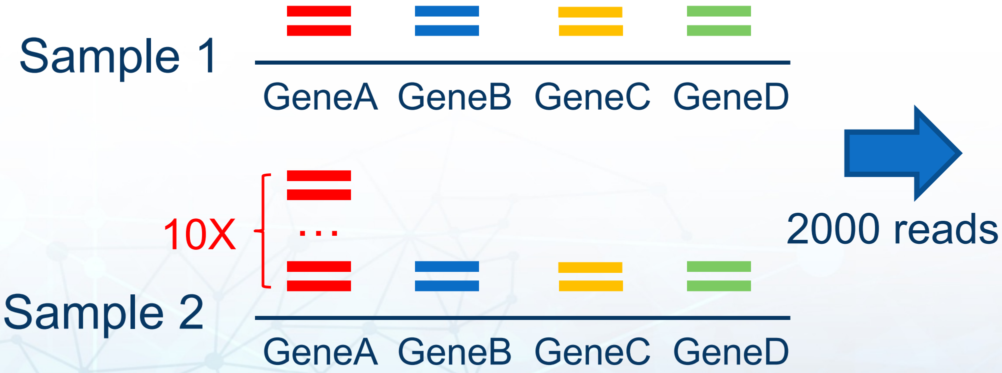


	Sample1	Sample2
geneA	4	8
geneB	2	4
geneC	3	6
Total size	**9**	**18**

Coefficients: **X1** **X0.5**

Composition bias towards high counts genes overshadows the rest

- **Highly expressed** genes “**consume**” a substantial portion of the reads (number of reads is finite)
- Resulting in **remaining genes being under-sampled**
- **Normalization factor minimizes the log-fold** changes between the samples for most genes (this assumes they are not diff. exp.)



	Sample1	Sample2
gene1	500	1538
gene2	500	154
gene3	500	154
gene4	500	154

→

X3.25

Scaling factor

Sample2
5000
500
500
500

Low expressed genes/transcripts are not informative

- Biologically, a gene must be **expressed at some minimal level** before it is likely to be translated into a protein or to be **biologically important**
- **Remove low expressed genes/isoforms** as they provide little evidence for differential expression
- **Improve statistical analysis** (less tests to perform)
- **No standard** threshold!
- GenPipes “loose” filtering:
  - Genes: **at least 1 read per sample**
  - Transcripts: remove if **<10% of the most abundant transcript**

Part 3:
Differential expression analysis

DEA consists of comparing the expression level

- Taking the **normalized read count** data and performing **statistical analysis**
- Identify quantitative **changes in expression levels** between experimental groups
- Gene level: **edgeR***, **DESeq2***,…
- Transcript level: **Cuffdiff***, Sleuth,…
- Only **pair-wise comparisons** supported by GenPipes

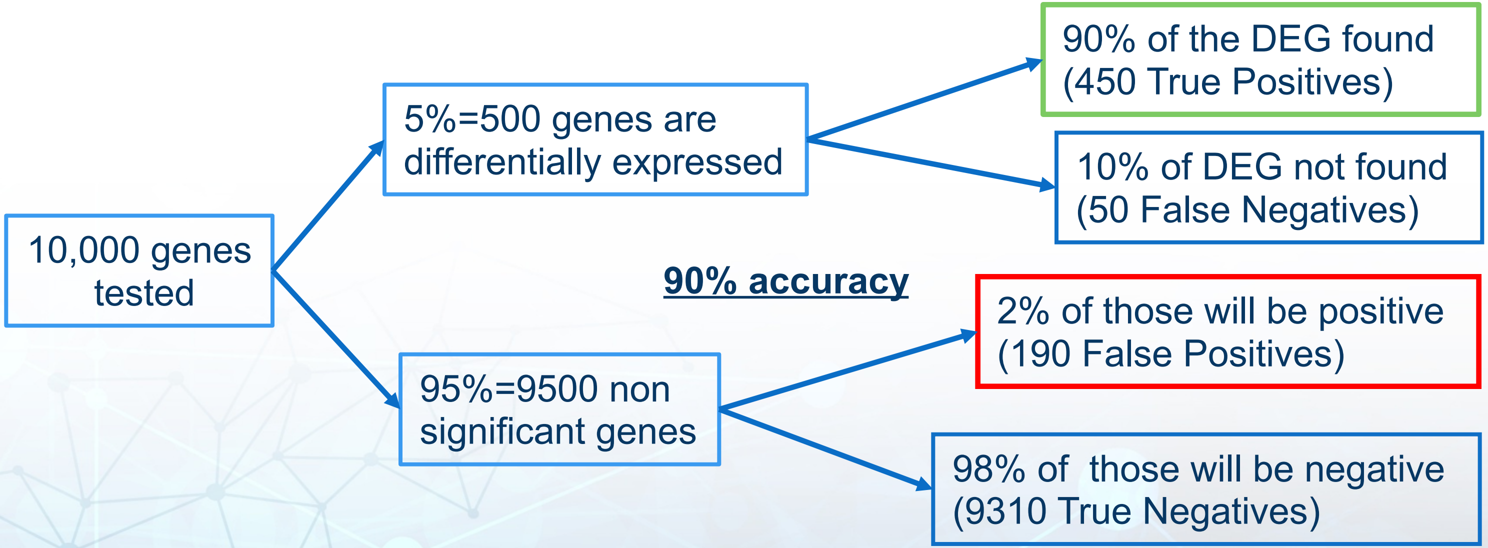
***used by GenPipes**

Statistical tests compare the expression between groups

- The replicates are used **to estimate the variance** and calculate the significance of observed **changes in expression** (logFC) between groups.
- **Many different statistical tests** exist depending of the tool and the experimental design (e.g. Fisher's exact test).
- A **p-value** reflecting the confidence that a **gene is differentially expressed** is then computed.
- An adjusted p-value is computed to account for **False Discoveries.**

False Positives are a big concern when working with large datasets

- When performing millions of tests (one per gene), **some will be positive** by chance only (**False Positive**).
- E.g. an analysis with 90% accuracy:



10,000 genes tested

5%=500 genes are differentially expressed

95%=9500 non significant genes

90% accuracy

90% of the DEG found (450 True Positives)

10% of DEG not found (50 False Negatives)

2% of those will be positive (190 False Positives)

98% of those will be negative (9310 True Negatives)

$$\frac{190}{450 + 190}$$

= 30% of False Positive!

Inspired by http://www.statisticshowto.com/false-discovery-rate/

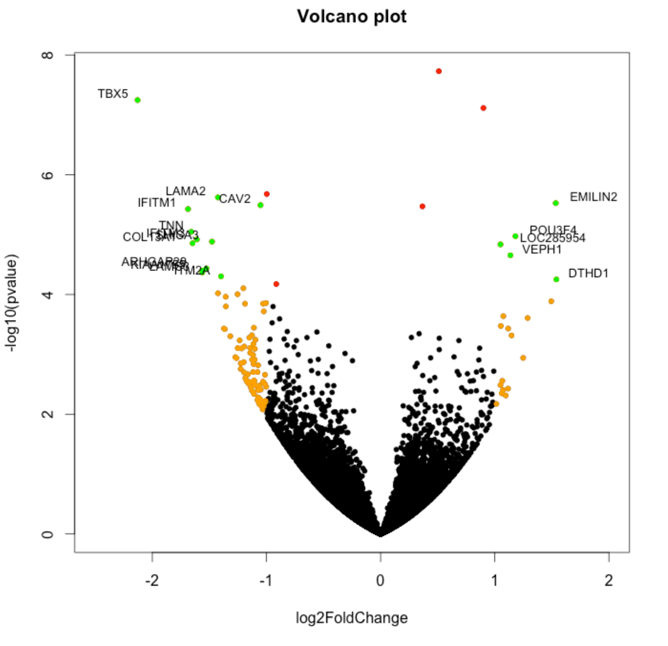
The number of False Discoveries can be controlled

- We need to account for that => **Multiple testing correction**
- Benjamin-Hochberg method known as **FDR** (**False Discovery Rate**) most commonly used
- This allows us to set the **rate of False Positive** (usually 5%)

A FDR of 5% means that **5% of *significant* results** will be false positives!

What constitutes a differentially expressed gene isn't well established

- **No clear definition** of a "differentially expressed gene"
- Common approach is to use log Fold Change and FDR: **logFC>1.5** and **FDR<0.05**
- LogFC threshold is **arbitrary** and depends of the **sensitivity of the technology**
- **Small logFC** might **not be biologically relevant**, but the exact definition of "small" is open to interpretation



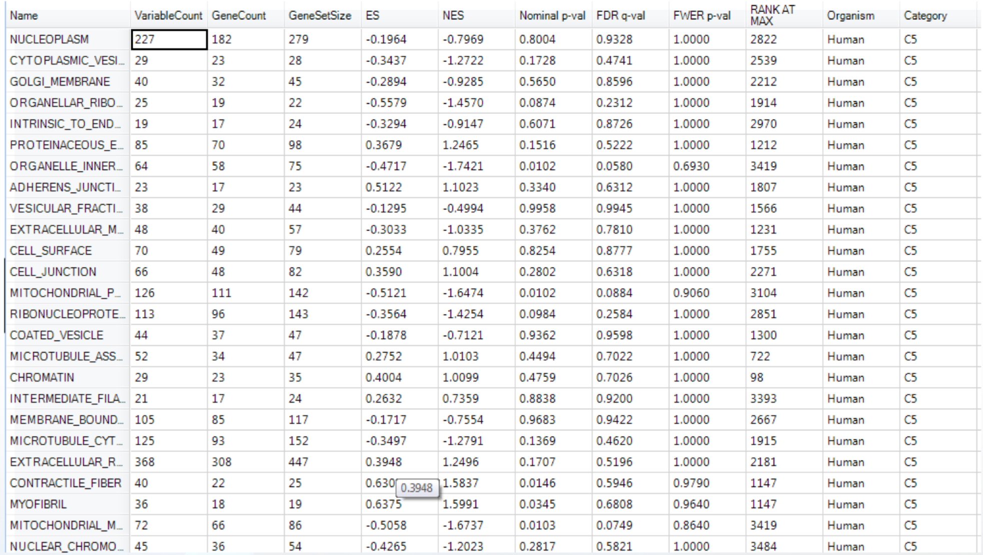
http://www.gettinggeneticsdone.com/2014/05/r-volcano-plots-to-visualize-rnaseq-microarray.html

Part 4:
Further analyses

GSEA determines if a set of genes is statistically different

- **Gene Set Enrichment Analysis** (GSEA) is a computational method that helps answer the question "**Are genes related to _____ significantly differentially expressed?**"
- Input: list of gene sets, expression matrix
- Gene sets can be molecular signatures (MSigDB) including gene ontology gene set (c5), immunologic signature gene set (c7), etc.
- Output: pvalues and FDRs for each gene set

http://software.broadinstitute.org/gsea/index.jsp

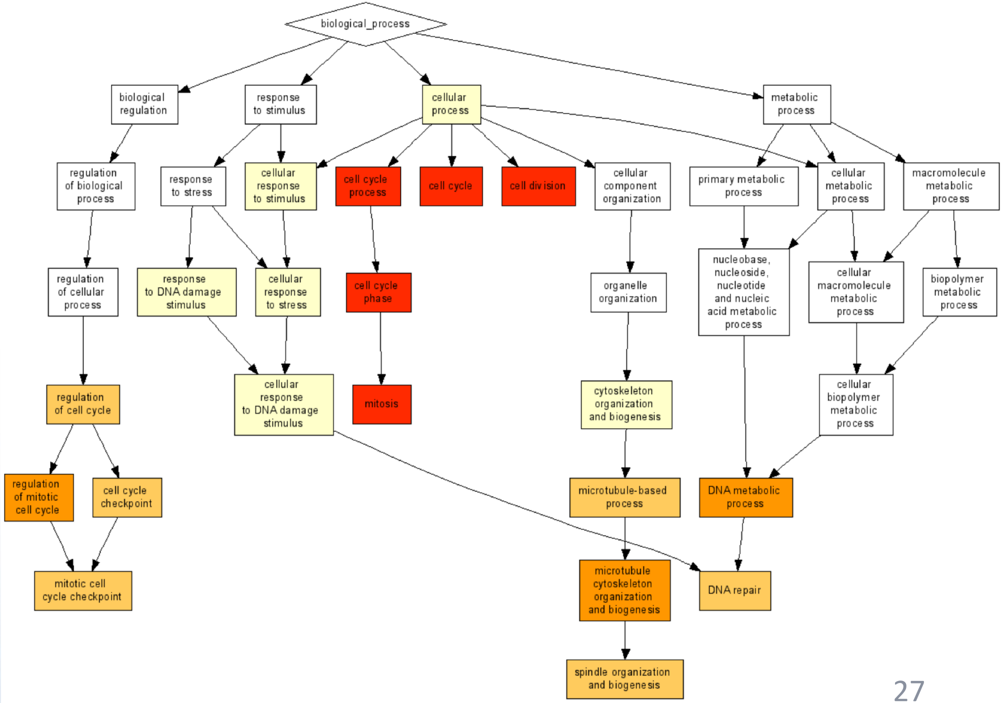


Name	VariableCount	GeneCount	GeneSetSize	ES	NES	Nominal p-val	FDR q-val	FWER p-val	RANK AT MAX	Organism	Category
NUCLEOPLASM	227	182	279	-0.1964	-0.7969	0.8004	0.9328	1.0000	2822	Human	C5
CYTOPLASMIC_VESI...	29	23	28	-0.3437	-1.2722	0.1728	0.4741	1.0000	2539	Human	C5
GOLGI_MEMBRANE	40	32	45	-0.2894	-0.9285	0.5650	0.8596	1.0000	2212	Human	C5
ORGANELLAR_RIBO...	25	19	22	-0.5579	-1.4570	0.0874	0.2312	1.0000	1914	Human	C5
INTRINSIC_TO_END...	19	17	24	-0.3294	-0.9147	0.6071	0.8726	1.0000	2970	Human	C5
PROTEINACEOUS_E...	85	70	98	0.3679	1.2465	0.1516	0.5222	1.0000	1212	Human	C5
ORGANELLE_INNER...	64	58	75	-0.4717	-1.7421	0.0102	0.0580	0.6930	3419	Human	C5
ADHERENS_JUNCTI...	23	17	23	0.5122	1.1023	0.3340	0.6312	1.0000	1807	Human	C5
VESICULAR_FRACTI...	38	29	44	-0.1295	-0.4994	0.9958	0.9945	1.0000	1566	Human	C5
EXTRACELLULAR_M...	48	40	57	-0.3033	-1.0335	0.3762	0.7810	1.0000	1231	Human	C5
CELL_SURFACE	70	49	79	0.2554	0.7955	0.8254	0.8777	1.0000	1755	Human	C5
CELL_JUNCTION	66	48	82	0.3590	1.1004	0.2802	0.6318	1.0000	2271	Human	C5
MITOCHONDRIAL_P...	126	111	142	-0.5121	-1.6474	0.0102	0.0884	0.9060	3104	Human	C5
RIBONUCLEOPROTE...	113	96	143	-0.3564	-1.4254	0.0984	0.2584	1.0000	2851	Human	C5
COATED_VESICLE	44	37	47	-0.1878	-0.7121	0.9362	0.9598	1.0000	1300	Human	C5
MICROTUBULE_ASS...	52	34	47	0.2752	1.0103	0.4494	0.7022	1.0000	722	Human	C5
CHROMATIN	29	23	35	0.4004	1.0099	0.4759	0.7026	1.0000	98	Human	C5
INTERMEDIATE_FILA...	21	17	24	0.2632	0.7359	0.8838	0.9200	1.0000	3393	Human	C5
MEMBRANE_BOUND...	105	85	117	-0.1717	-0.7554	0.9683	0.9422	1.0000	2667	Human	C5
MICROTUBULE_CYT...	125	93	152	-0.3497	-1.2791	0.1369	0.4620	1.0000	1915	Human	C5
EXTRACELLULAR_R...	368	308	447	0.3948	1.2496	0.1707	0.5196	1.0000	2181	Human	C5
CONTRACTILE_FIBER	40	22	25	0.630 0.3948	1.5837	0.0146	0.5946	0.9790	1147	Human	C5
MYOFIBRIL	36	18	19	0.6375	1.5991	0.0345	0.6808	0.9640	1147	Human	C5
MITOCHONDRIAL_M...	72	66	86	-0.5058	-1.6737	0.0103	0.0749	0.8640	3419	Human	C5
NUCLEAR_CHROMO...	45	36	54	-0.4265	-1.2023	0.2817	0.5821	1.0000	3484	Human	C5

Gorilla identifies enriched GO terms

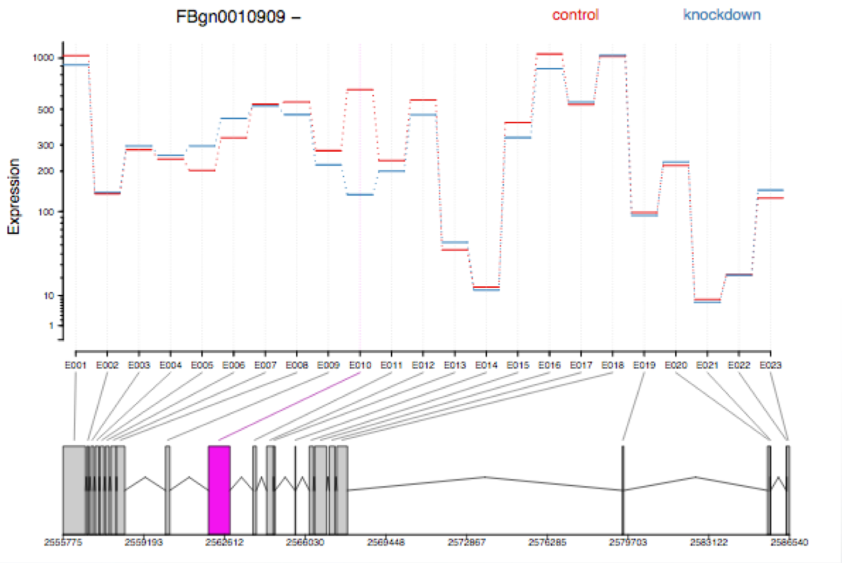
- GOrilla is a tool for identifying and visualizing enriched GO terms in ranked lists of genes.
- What gene ontologies and pathways do my DGE share?
- Input: list(s) of genes
- Output: pvalues and FDR for enriched GO terms, GO chart

http://cbl-gorilla.cs.technion.ac.il/


biological_process
biological regulation
response to stimulus
cellular process
metabolic process
regulation of biological process
response to stress
cellular response to stimulus
cell cycle process
cell cycle
cell division
cellular component organization
primary metabolic process
cellular metabolic process
macromolecule metabolic process
regulation of cellular process
response to DNA damage stimulus
cellular response to stress
cell cycle phase
organelle organization
nucleobase, nucleoside, nucleotide and nucleic acid metabolic process
cellular macromolecule metabolic process
biopolymer metabolic process
regulation of cell cycle
cellular response to DNA damage stimulus
mitosis
cytoskeleton organization and biogenesis
cellular biopolymer metabolic process
regulation of mitotic cell cycle
cell cycle checkpoint
microtubule-based process
DNA metabolic process
mitotic cell cycle checkpoint
microtubule cytoskeleton organization and biogenesis
DNA repair
spindle organization and biogenesis

There are more analyses you can do!

- Alternative splicing
- Gene fusion analysis
- Differential exon usage
- …



Canadian Centre for Computational Genomics

McGill

MiCM McGill initiative in Computational Medicine

compute canada | calcul canada

GenomeCanada

GenomeQuébec

Ontario Genomics

SickKids

bioinformatics.ca

canarie

IHEC

Calcul Québec

GenAP

Terry Fox PROFYLE

HPC4Health