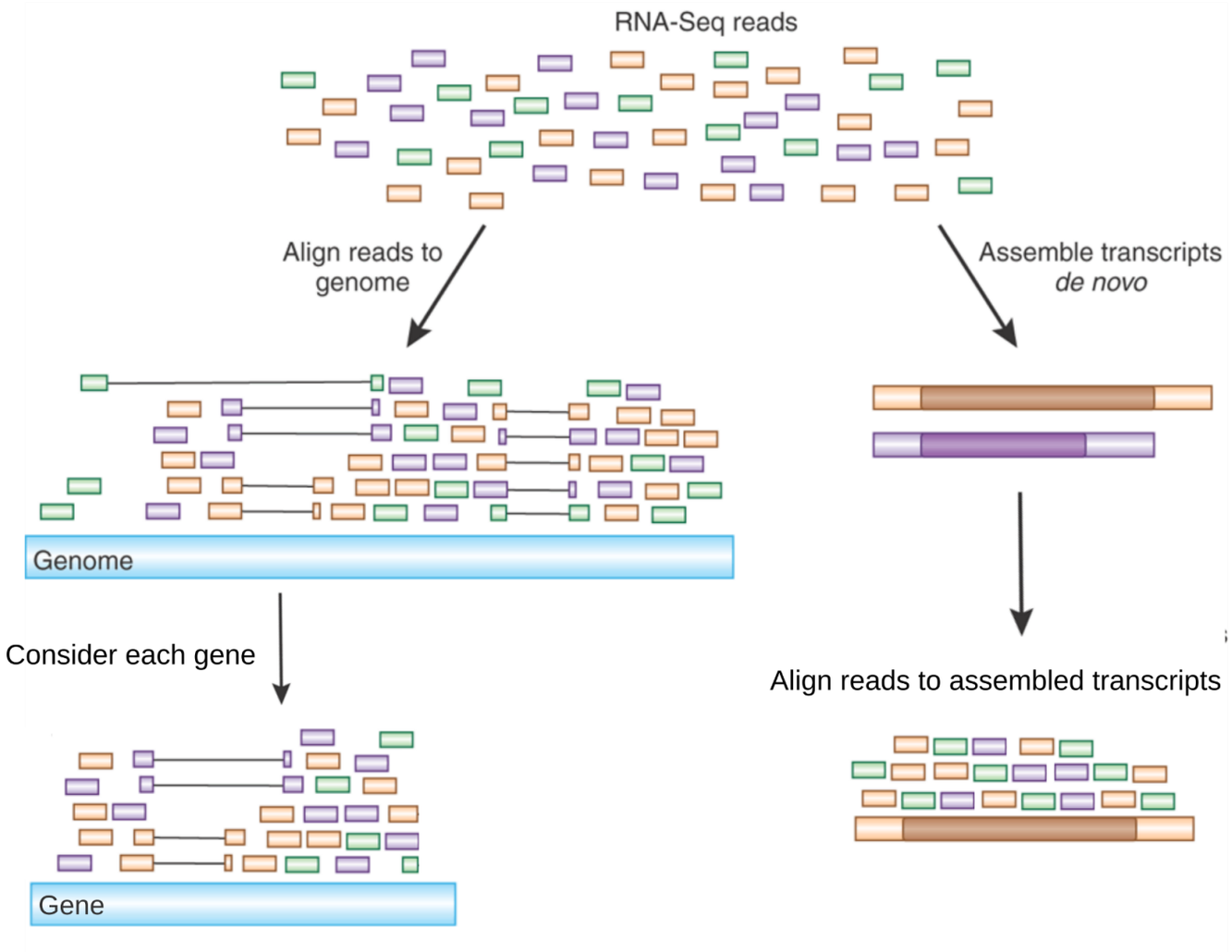
Differential Gene Expression Analysis

DECRyPT Workshop

Fantin Mesny - October 30th, 2019

From RNAseq reads to read-counts



The output of a RNAseq reads mapper

abundance.tsv
~/Desktop/SPPworkshop

target_id	length	eff_length	est_counts	tpm
AT1G19090.1	2509	2310	0	0
AT1G18320.1	699	500	7.03866	3.70058
AT5G11100.1	1709	1510	23	4.00407
AT4G35335.1	1280	1081	112	27.236
AT1G60930.1	3876	3677	26	1.85879
AT1G17000.1	2351	2152	0	0
AT3G55270.1	3130	2931	236	21.1664
AT4G16150.1	3181	2982	224	19.7465
AT3G18150.1	1370	1171	0	0
AT3G52310.1	2443	2244	39	4.56869
AT3G08943.1	3165	2966	435.709	38.6168
AT4G15950.1	814	615	24	10.2586
AT5G60900.1	2873	2674	19	1.86785
AT5G18710.1	1919	1720	3	0.458504
AT1G64460.1	2240	2041	63	8.11424
AT2G31100.1	1247	1048	0	0
AT1G60260.1	1634	1435	57	10.4417
AT2G06210.1	3579	3380	221	17.188
AT3G57180.1	2232	2033	101	13.0597
ATMG01320.1	922	723	0	0
AT4G15130.1	1234	1035	57	14.4772
AT5G39100.1	808	609	0	0

One file per sample. If 3 replicates: 3 files per condition.

Making sense of read-counts

Differential gene expression analysis:

Statistics to look for over- or under-expressed genes in the test condition (in comparison to the control condition)

👉 Which genes are differentially expressed in presence/absence of treatment ?

R packages dedicated to DGE analysis:

- DESeq2
- edgeR
- Limma

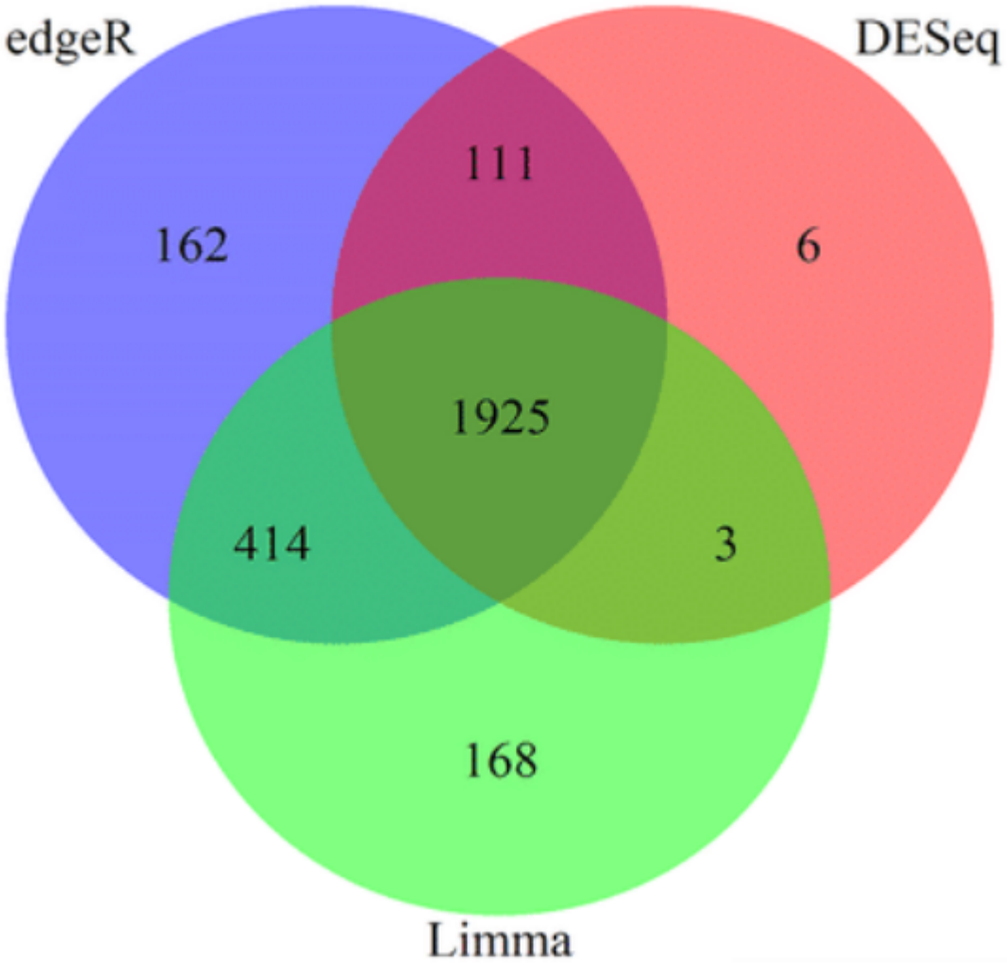
What do the DEG-dedicated packages do ?

- Correct the read counts for the different sequencing depths between samples
- (Filter out noise)
- Statistical testing for differential expression

Method	Normalization	Read counts distribution	Differential Expression Test
edgeR	TMM	Negative Binomial distribution	Exact test
DESeq	DESeq sizeFactors	Negative Binomial distribution	Exact test
Limma	TMM	Voom transformation of counts	Empirical Bayes method

Laura Elo, 2014

Results comparison



Shahjaman et al., 2019

Example of a DESeq output file

Open ▾ | DESeq_output ~/Desktop/SPPworkshop | Save

	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
FBgn0000008	95.1442917575889	-0.262373083152848	0.218504501036476	-1.20076740711647	0.229841438140096	0.536181830631843
FBgn0000014	1.05652281859341	3.28988530366209	2.0527863031	1.6026438303363	0.109013311107605	NA
FBgn0000017	4352.55356876647	-0.100020130655258	0.11209123269922	-0.892310025027978	0.372226782328688	0.683195420721474
FBgn0000018	418.61048415965	0.229049127470739	0.130260862036797	1.75838792933855	0.0786815255463772	0.291788667835051
FBgn0000024	6.406199980976	0.306050642950857	0.751285635377863	0.40736921956045	0.683736830521498	0.880472446679203
FBgn0000032	989.720216813761	0.237413396111431	0.120285782752783	1.97374444991037	0.048410808136915	0.217658242771846

Plain Text ▾ Tab Width: 8 ▾ Ln 9, Col 1 ▾ INS

👉 Adjusted P-values reveal the genes that are differentially expressed

About Log2FoldChanges...

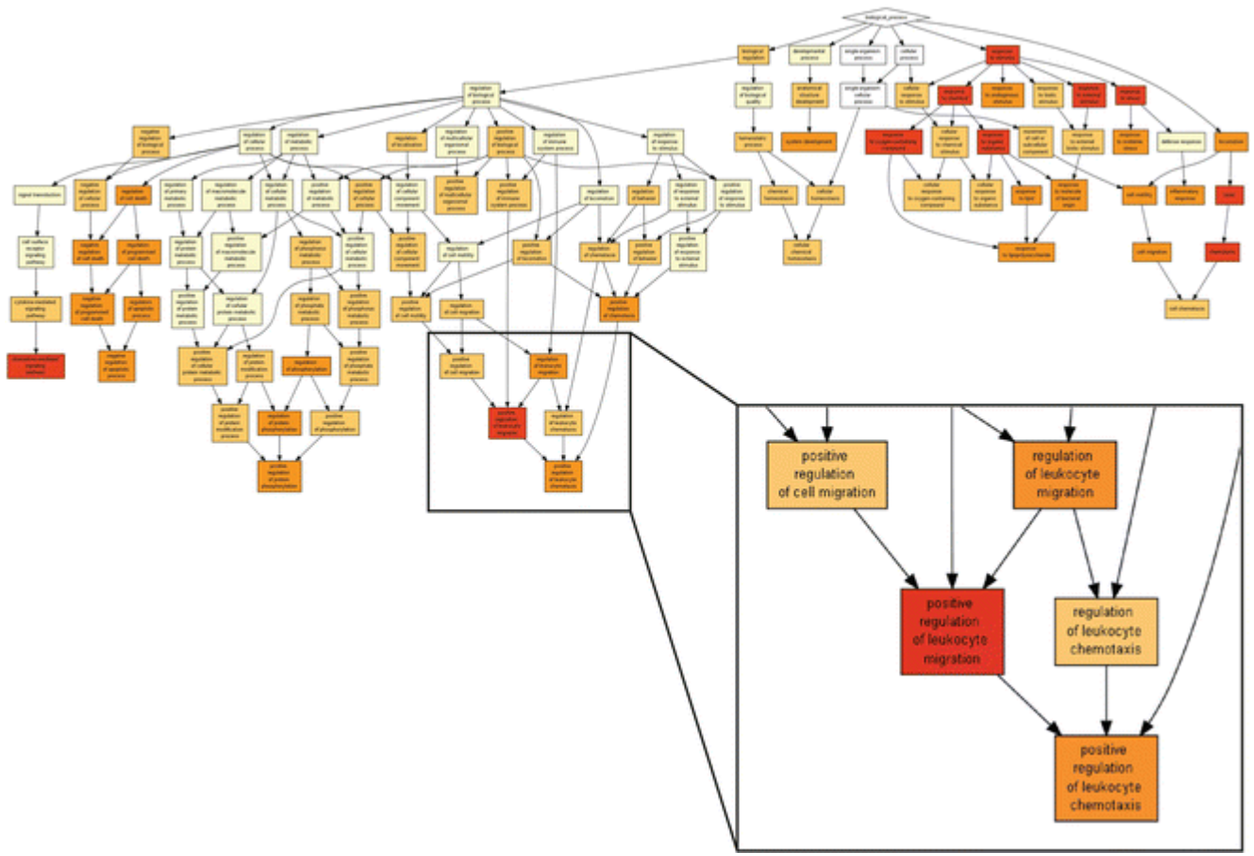
- Calculated on the read counts
- Biased for genes with low read counts
- ⚠️ Can only be interpreted after a correction !
  Multiple LFC-shrinkage algorithms have been developed.

What's the function of over-expressed genes ?

- 👉 **Linking DEGs to gene annotation**
- 👉 **Functional enrichment**

Gene ontology

- Hierarchical functional annotation
- Three independant categories:
  - Biological process
  - Molecular function
  - Cellular components



GO enrichment:

What are the GO terms that are significantly enriched ?

enriched= more present than what we would expect by chance

GO enrichment analysis

- Dedicated websites (for model organisms): *PANTHER*, *DAVID*...
- R packages (*topGO*)
  - need to provide a GO-based gene annotation as an input:

```
gene1       GO:0001478,GO:000784;GO:040411
gene2       GO:0007482,GO:000994;GO:040971
...
```