TRANSCRIPTOMICS DATA AND DIFFERENTIAL EXPRESSION ANALYSIS

Chandni Desai

Everyone is doing RNA seq !!

nature

Widespread intronic polyadenylation inactivates tumour suppressor genes in

Anterior Pituitary Transcriptome Suggests Differences in ACTH Release in Tame and Aggressive Foxes

J Comput Biol. 2018 Aug 22. doi: 10.1089/cmb.2017.0244. [Epub ahead of print]

Differential Expression Analysis in RNA

Tambonis T[1], Boareto M[2], Leite VBP[1].

The transcriptional landscape of polyploid wheat

R. H. Ramírez-González[1,*], P. Borrill[1,*,†], D. Lang[2], S. A. Harrington[1], J. Brinton[1], L. Venturini[3], M. Davey[4], J. Jacobs[4], F. van...

Human plasma and serum extracellular small RNA reference profiles and their clinical utility

Negative pressure wound therapy in the treatment of diabetic foot ulcers may be mediated through differential

ohendorff, T. Koblik, J. Machlowska, B. Kiec-Wilk, P. Wolkow,

Transcriptome analysis of the response of silkworm to drastic changes in ambient temperature

Authors Authors and affiliations

Huizhen Guo, Chunlin Huang, Liang Jiang, Tingcai Cheng, Tieshan Feng, Qingyou Xia

PeerJ. 2018 Aug 21;6:e5427. doi: 10.7717/peerj.5427. e

RNA-Seq analysis of differential stages of tension wood formatio

Cai M[#1], Huang H[#1], Ni F[1], Tong Z[1], Lin E[1], Zhu M

RNA-seq analysis reveals different gene ontologies and path in rheumatoid arthritis and Kashin–Beck disease

ing Gao, Zongqiang

Changes in circRNA expression profiles related to the antagonistic effects of *Escherichia coli* F17 in lamb spleens

Comparative proteomics and gene expression analysis in *Arachis duranensis* reveal stress response pr to drought tolerance

Lílian S.T. Carmo [a, 1], Andressa C.Q. Martins [a, b, 1], Cinthia C.C. Martins [a], Mário

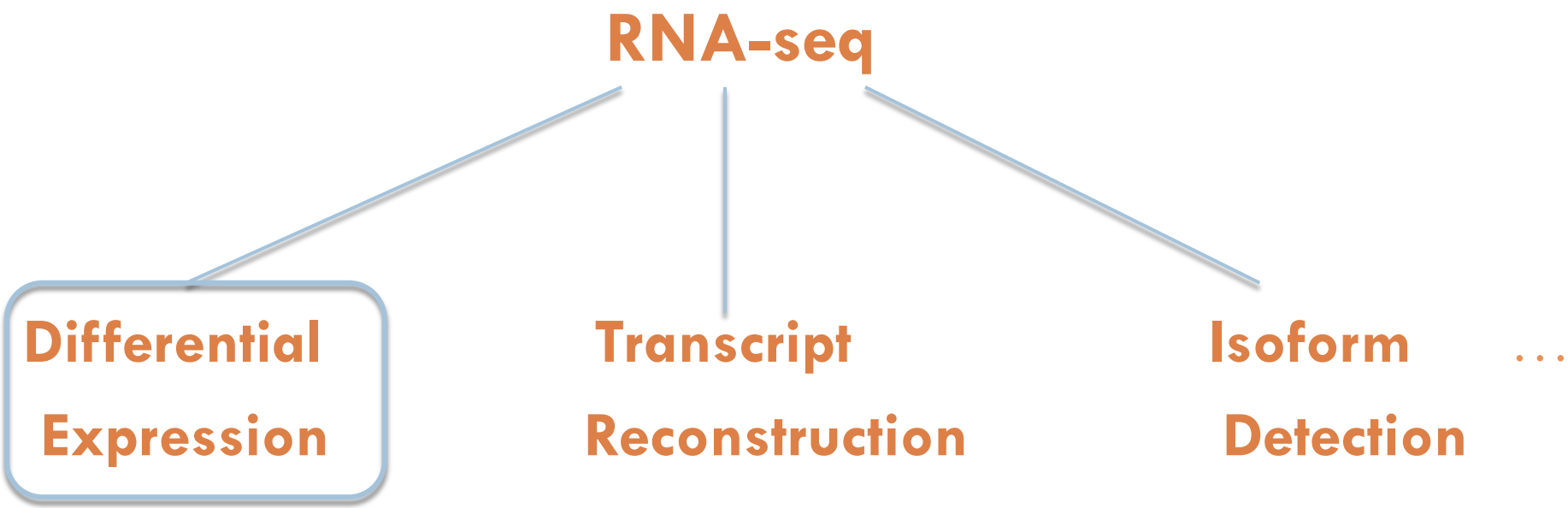
Ana C.G. Araujo [a], Ana C.M. Brasileiro [a], Robert N.G. Miller [b], Patrícia M. Guimar

Mol Cell Proteomics. 2018 Oct 6. pii: mcp.RA118.000832. doi: 10.1074/mcp.RA118.000832. [Epub ahead of print]

Improvements to the rice genome annotation through large-scale analysis of RNA-Seq and proteomics datasets.

Ren Z, Qi D[1], Nina P[2], Li K[1], Wen B[3], Zhou R[1], Xu S[1], Liu S[1], Jones AR[4].

Today,


RNA-seq
Differential Expression
Transcript Reconstruction
Isoform Detection
...

Experiment Design

Sequencing Depths

Replicates

Avoiding bias and batch effects

Sequencing Depth

- coverage will vary drastically between different transcripts depending, most importantly, on their expression
- number of required reads is determined by the **least** abundant RNA species of interest – how can you gauge how much is enough?
- consider :
  - guidelines from literature
  - type of experiment and biological question
  - transcriptome size and complexity
  - error rate of the sequencing platform

Table 1: Recommended sequencing depths for typical RNA-seq experiments for different genome sizes (Genohub, 2015). DGE = differential gene expression, SR = single read, PE = paired-end.

	Small (bacteria)	**Intermediate (fruit fly, worm**	**Large (mouse, human)**
No. of reads for DGE ($x10^6$)	5 SR	10 SR	20–50 SR
No. of reads for *de novo* transcriptome assembly ($x10^6$)	30–65 PE	70–130 PE	100–200 PE
Read length (bp)	50	50–100	>100

http://chagall.med.cornell.edu/RNASEQcourse/

Greater Depth = More Statistical Power

Example: Single gene, reads sampled at different sequencing depths

Reads per sample	Sample A Number of reads	Sample B Number of reads	P-value (Fishers Exact Test)
100,000	1	2	1
1,000,000	10	20	0.099
10,000,000	100	200	8.0e-09

Brian Haas

Replicates – Capturing breadth of variability

- biological replicates allow you to have a better handle on the true mean and variance of expression (of all genes in question) for the biological population of interest
- ideally, there should be enough replicates to capture the breadth of the variability and to identify and isolate sources of noise
- Illumina sequencing data are highly replicable, with relatively little technical variation

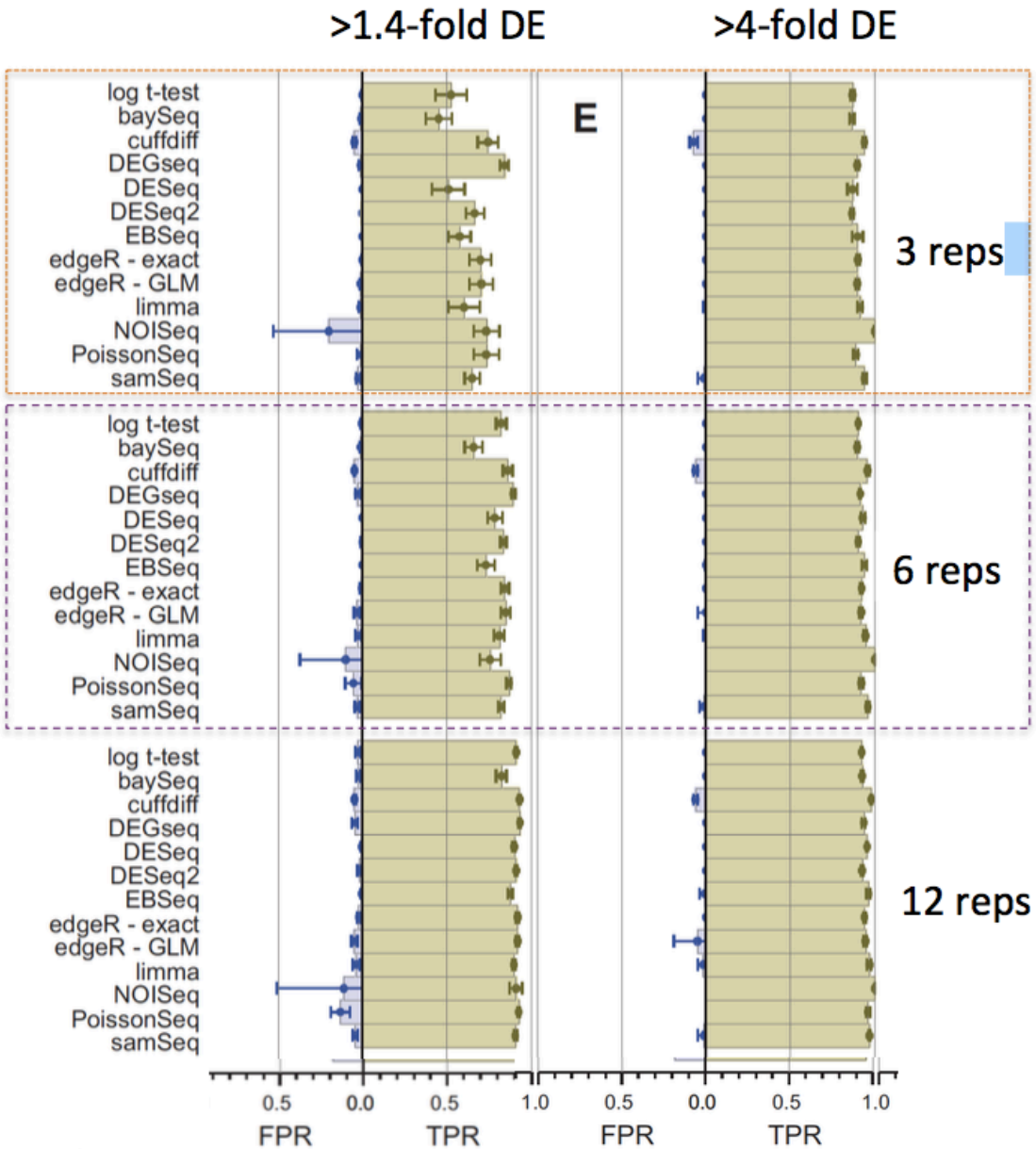
Hypothetical mouse single-cell gene expression RNA sequencing experiment

Replicate types and categories

	Replicate type	Category
Subjects	Colonies	Biological
	Strains	Biological
	Cohoused groups	Biological
	Gender	Biological
	Individuals	Biological
Sample preparation	Organs from sacrificed animals	Biological
	Methods for dissociating cells from tissue	Technical
	Dissociation runs from given tissue sample	Technical
	Individual cells	Biological
	RNA-seq library construction	Technical
Sequencing	Runs from the library of a given cell	Technical
	Reads from different transcript molecules	Variable
	Reads with unique molecular identifier from a given transcript molecule	Technical

Blainey et al. (2014)

Differential Expression Accuracy Improves with Higher Biological Replication



At a minimum, do 3 bio replicates

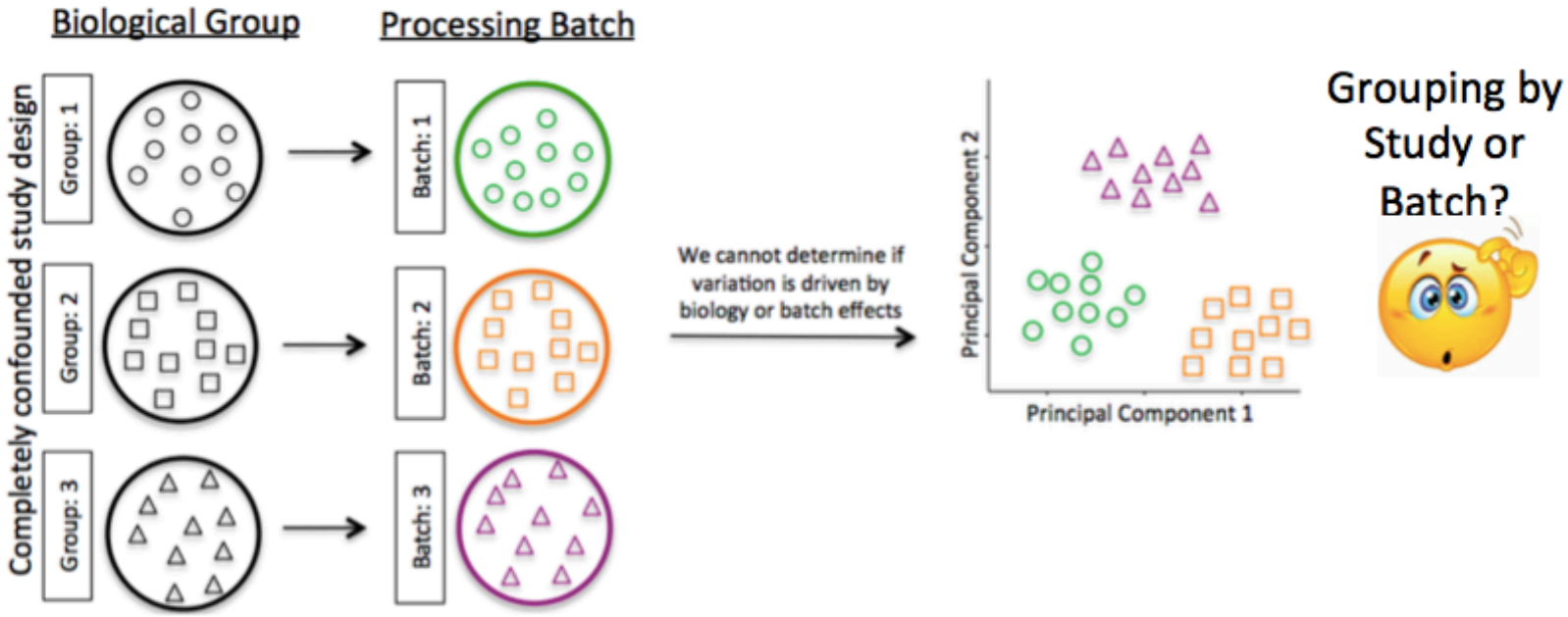
Respectable goal

*Figure taken and adapted from Schurch et al., RNA, 2016

Avoiding Bias and Batch Effects

- Identify the question of interest *(what effect are you truly after?)*
- Attempt to identify possible sources of variability *(nuisance factors)*
- Plan the experiment in a way that reduces the effect of the nuisance factors
- Protect yourself against unknown sources of variation

Batch effects - example

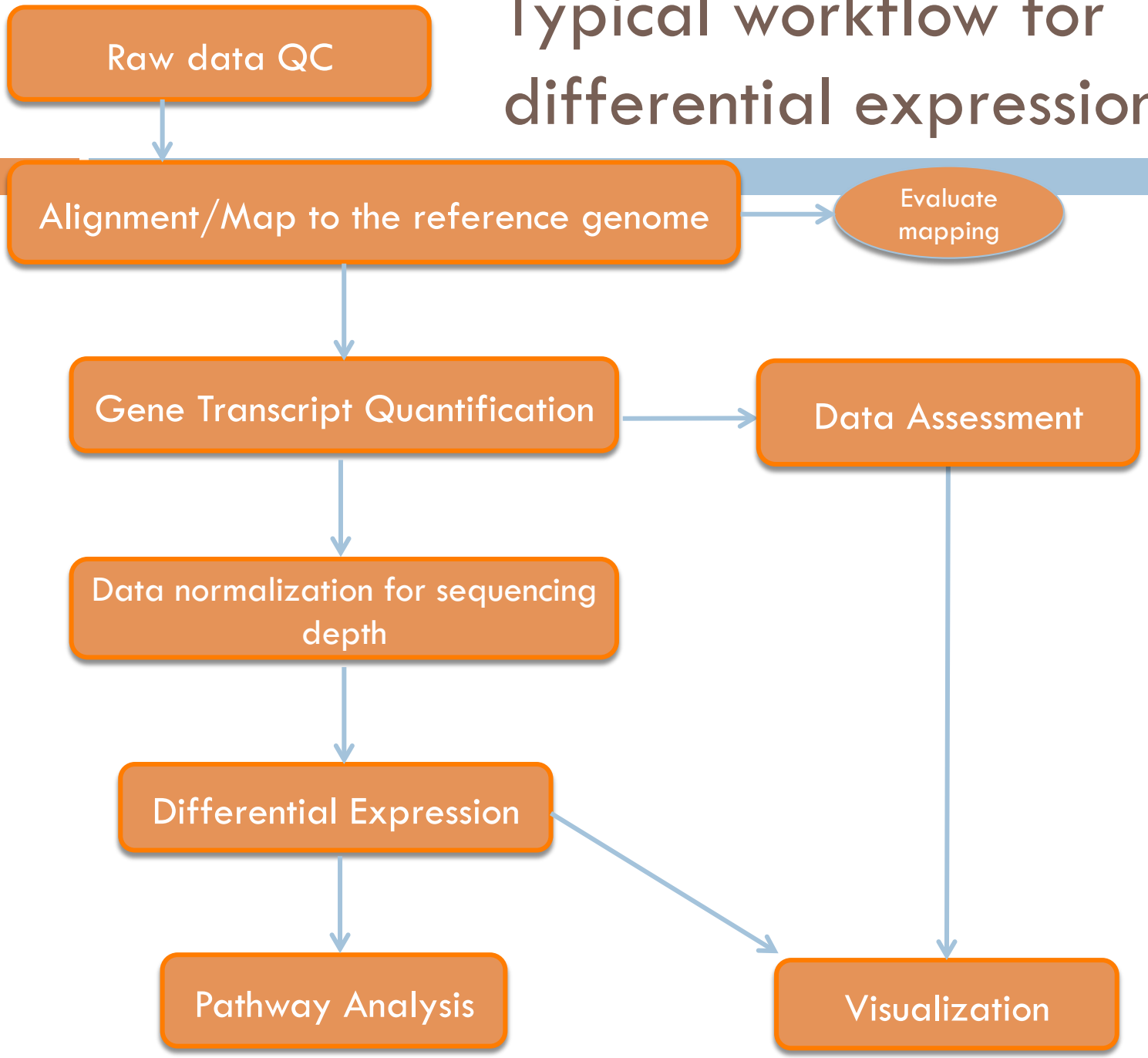


Batch variable types:

- Times and dates
- Technician processing the samples
- Sequencing machine, or flow cell lane (Illumina)

Adapted from: Stephanie C. Hicks, Mingxiang Teng, Rafael A. Irizarry.
https://www.biorxiv.org/content/early/2015/09/04/025528
On the widespread and critical impact of systematic bias and batch effects in single-cell RNA-Seq data.

Typical workflow for differential expression analysis



Reads and the Reference

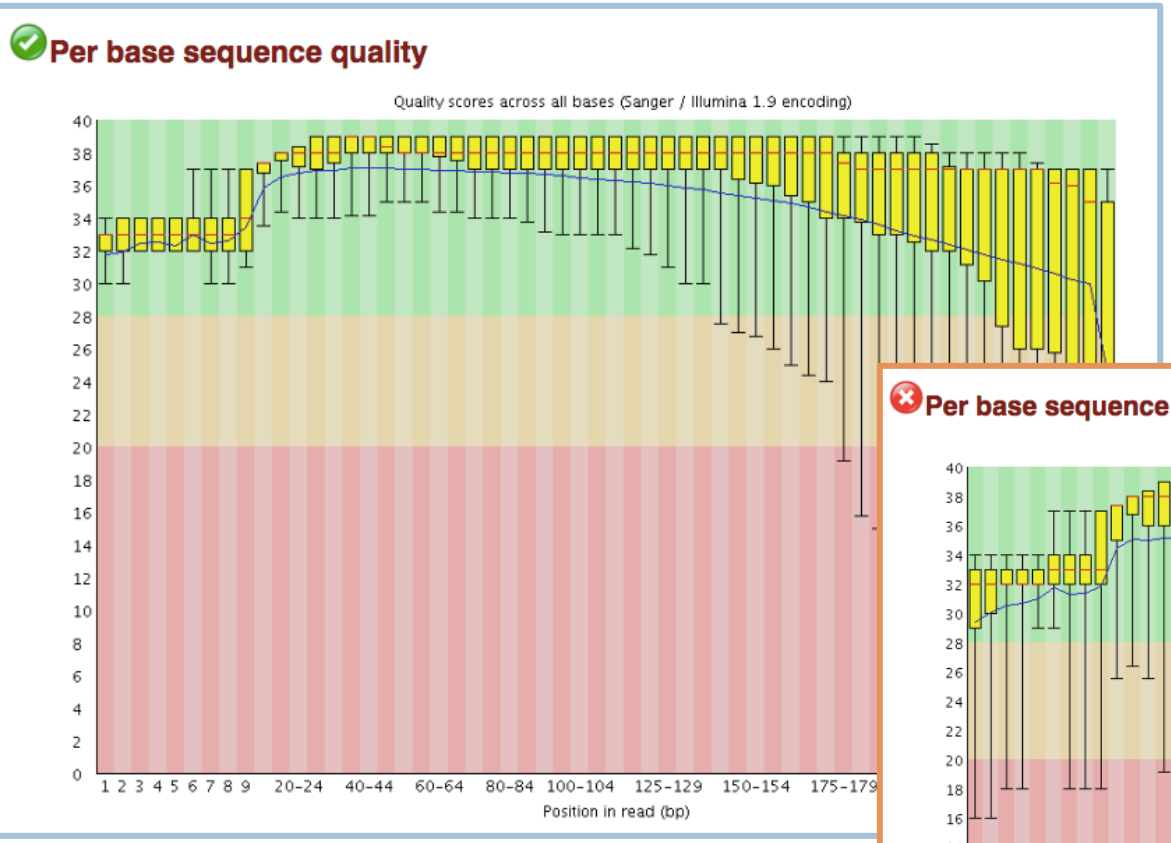
What do I need to get started?

‘Raw Data’/Reads

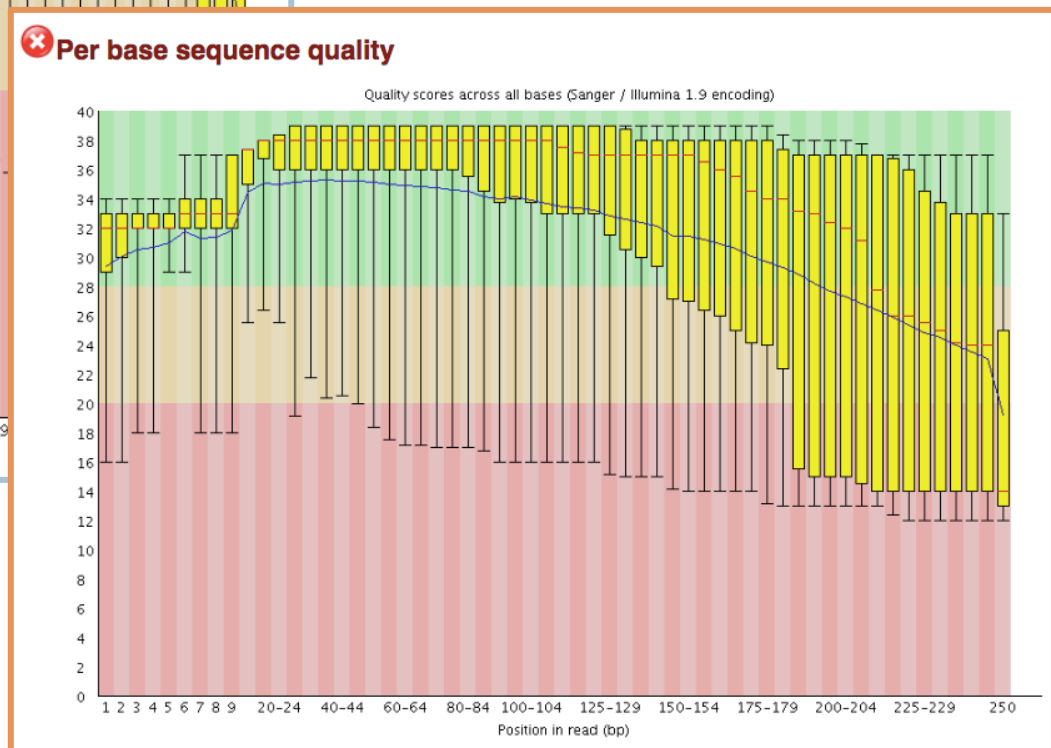
- single-end (SE) or paired-end (PE) reads
- paired reads are preferable for de novo transcript discovery or isoform expression analysis
- longer reads improve mappability and transcript identification
- cheaper, short SE reads are normally sufficient for studies of gene expression levels in well-annotated organisms; longer and PE reads are preferable to characterize poorly annotated transcriptomes

```
run_1868_s_1_3_withindex_sequence.txt_CCTATGCC.fq.gz
run_1868_s_1_3_withindex_sequence.txt_CGTTACCA.fq.gz
run_1868_s_1_3_withindex_sequence.txt_GACAGTAA.fq.gz
run_1868_s_1_3_withindex_sequence.txt_GCACATCT.fq.gz
run_1868_s_1_3_withindex_sequence.txt_GGTCCAGA.fq.gz
run_1868_s_1_3_withindex_sequence.txt_GTATAACA.fq.gz
run_1868_s_1_3_withindex_sequence.txt_GTCGCCTT.fq.gz
run_1868_s_1_3_withindex_sequence.txt_GTCTGATG.fq.gz
run_1868_s_1_3_withindex_sequence.txt_TCCAGCAA.fq.gz
run_1868_s_1_3_withindex_sequence.txt_TCCTTGGT.fq.gz
run_1868_s_1_3_withindex_sequence.txt_TCGCCTTG.fq.gz
run_1868_s_1_3_withindex_sequence.txt_TCGGAATG.fq.gz
run_1868_s_1_3_withindex_sequence.txt_TGCTCGAC.fq.gz
run_1868_s_1_3_withindex_sequence.txt_TTACGCAC.fq.gz
run_1868_s_1_3_withindex_sequence.txt_TTCGCTGA.fq.gz
run_1868_s_1_3_withindex_sequence.txt_TTGAATAG.fq.gz
run_1868_s_2_3_withindex_sequence.txt_CGTTACCA.fq.gz
run_1868_s_2_3_withindex_sequence.txt_GACAGTAA.fq.gz
run_1868_s_2_3_withindex_sequence.txt_GCACATCT.fq.gz
run_1868_s_2_3_withindex_sequence.txt_GGTCCAGA.fq.gz
run_1868_s_2_3_withindex_sequence.txt_GTATAACA.fq.gz
run_1868_s_2_3_withindex_sequence.txt_GTCGCCTT.fq.gz
run_1868_s_2_3_withindex_sequence.txt_GTCTGATG.fq.gz
run_1868_s_2_3_withindex_sequence.txt_TCCAGCAA.fq.gz
run_1868_s_2_3_withindex_sequence.txt_TCCTTGGT.fq.gz
run_1868_s_2_3_withindex_sequence.txt_TCGCCTTG.fq.gz
run_1868_s_2_3_withindex_sequence.txt_TCGGAATG.fq.gz
run_1868_s_2_3_withindex_sequence.txt_TGCTCGAC.fq.gz
run_1868_s_2_3_withindex_sequence.txt_TTACGCAC.fq.gz
run_1868_s_2_3_withindex_sequence.txt_TTCGCTGA.fq.gz
run_1868_s_2_3_withindex_sequence.txt_TTGAATAG.fq.gz
run_1868_s_2_3_withindex_sequence.txt_TTGAGCCT.fq.gz
```

Quality of sequencing reads



Familiar?



Example tools for QC :
bbtools package, trimmomatic

Reference Genome

1. Sequence

- plain text file with full nucleotide genome sequence, long string of A/T/G/C
- file format : fasta

2. Annotation

- table defining genomic regions (exons, introns, etc)
- file format : GFF/GTF (general feature/gene transfer format)

- University of California, Santa Cruz (UCSC; https://genome.ucsc.edu/), or
- the European genome resource, Ensembl (http://www.ensembl.org)
- NCBI/Genbank (ftp://ftp.ncbi.nlm.nih.gov/genomes/)

Reference Genome sequence (fasta)

```
[cdesai@endeavor MouseGenome_GRCm38]$ head mm_ref_GRCm38.p2_chr1.fa
>gi|372099109|ref|NC_000067.6| Mus musculus strain C57BL/6J chromosome 1, GRCm38.p2 C57BL/6J
TAACAATCATATATGTAAAAAGTAGGTTGGCACCGAAAACGGTCGGACCGACCTGGGTACATATATAAAG
AAACCGTAGGGTCAGCAAAGCACACTCATCTATGGGACCGTGCAATCCAACTAATATTTTCCTGCTCTGC
AAGGACTACGAGGTGGACTTTGAGGATTTAAGGCTGACATGTGTATTTTGCAAAAATGAATTAACAACAG
AAGAATTGCTGGCGTTTGCACTAAAGGAGCTAAGCATTGTGTGGAGACATAACTGGCCATTTGGAGTATG
CGCACCATGCTTGGCACGTGAAGTAAAAGTGAGGGAGCTGCGACATTGGGACCATTCCTGTTACGGACCC
ACTGTGGAACAGACAACAGGACGGTCATTAGCTGAACTATATATACGGTGCCATGCATGCAGCAAACCGT
TAAGTATACAGGAAAAGGAGCATCAGGTACAGGCATACATCCACTTCCACTATATAGCTGGACAGTGGAC
GGGAAGGTGTTGCCAGTGTAGAGGGCCATGCACGGCCAGGTGGCAACCATAAAGGACATAGTCCTTGAAG
AGCGTCCTGAGGTGGTTGACCTACATTGCAATGAGCAGTTATTAGACAGCTCAGAGTCAGAGGAGGAGGA
TAGTGTGCGTGAGCAACTTGTTGAACAAGCACAGCAGGCCTACAGGGTGGTTACTACCTGTGGCATTTGT
AAGTGTCCAGTTAGGCTGGTGGTGCAGTGCGGAGACGCAGACCTGAAGGTGCTACATGAACTACTGCTGG
GCGACTTGTCCATAGTGTGTCCTGGCTGCGCATAAGGGACATGGCTGACAGTGAAGGTACAGAAAGCGGG
GATGGGACCGAGGCCGCGGAACGCGCAGGGGGGTGGTTTCTGGTAGAAGCCGTGGTAGACCGCACCACAG
GATACCAGGTGTCCAGTGATGAGGAGGACAATAGCATTGACACAGGGGAAGACCTAGTAGACTTCATAGA
TACAAGGCGCCCCGGGGATGGGCAGGAAGTGCCGCTTGCGTTGTTCGTTCAACAAAATGCACAGGATGAC
GCTGCAACGGTGCAGGCACTAAAACGAAAGTATACATGTAGCCCTGCAAGCAGCACCTGTGTGTCCTTGG
TGGACAGTGAATTAAGTCCCCGGCTGGACGCCATACGGATACACCGGGGACAGGACAGGGCTAGGAGAAG
GCTGTTTGAGCAAGATAGTGGCTATGGCCATACGCAAGTGGAAATTGGAGCATCAGAAAGTCAGGTACCG
GGGGATGCGCAACATGAGGGGGGGGGGAATCCGTGCAGGAAGCAGAGGAGGAGCGTGGGGGGGGGGACG
GAGAGGCCGAGGCCACAGGTAACCAGGAAACGCAAGCGCAGGAGCAGGCGGCAGACATATTAGAGGTGTT
TAAGGTTAGTAATTTAAAAGCAAAATTACTGTACAAATTCAAGGACCTATTTGGACTAGCATTTGGGGAG
CTGGTAAGAAATTTTAAAAGTGATAAGTCAATATGTGGGGACTGGGTAATATGTGCGTTTGGTGTATACC
ATGCTGTGGCAGAGGCTGTAAAAACCTTAATACAACCTATATGTGTGTATGCACACATACAAATACAGAC
ATGCCAGTGGGGAATGGTAATATTAATGCTTGTGCGATACAAATGTGGGAAGAGCAGGGAAACAGTAGCA
CACAGCATGGGAAAACTGTTAAACATACCGGAAAGACAGATGCTAATTGAACCACCAAAGATTAGAAGCG
CACCGTGCGCACTATATTGGTATAGAACAGCCATGGGAAACGCCAGCGAGGTGTATGGCGAAACACCTGA
```

Reference Genome annotation (gtf)

```
[cdesai@endeavor MouseGenome_GRCm38]$ head -20 gencode.vM6.annotation.gtf
##description: evidence-based annotation of the mouse genome (GRCm38), version 6 (Ensembl 81)
##provider: GENCODE
##contact: gencode-help@sanger.ac.uk
##format: gtf
##date: 2015-07-15
chr1    HAVANA  gene    3073253 3074322 .       +       .       gene_id "ENSMUSG00000102693.1"; gene_type "TEC"; gene_status "KNOWN"; gene_name "4933401J01Rik"; level 2; havana_
gene "OTTMUSG00000049935.1";
chr1    HAVANA  transcript      3073253 3074322 .       +       .       gene_id "ENSMUSG00000102693.1"; transcript_id "ENSMUST00000193812.1"; gene_type "TEC"; gene_status "KNOWN
"; gene_name "4933401J01Rik"; transcript_type "TEC"; transcript_status "KNOWN"; transcript_name "4933401J01Rik-001"; level 2; tag "basic"; transcript_support_level "NA"; havana_
gene "OTTMUSG00000049935.1"; havana_transcript "OTTMUST00000127109.1";
chr1    HAVANA  exon    3073253 3074322 .       +       .       gene_id "ENSMUSG00000102693.1"; transcript_id "ENSMUST00000193812.1"; gene_type "TEC"; gene_status "KNOWN"; gene_
name "4933401J01Rik"; transcript_type "TEC"; transcript_status "KNOWN"; transcript_name "4933401J01Rik-001"; exon_number 1; exon_id "ENSMUSE00001343744.1"; level 2; tag "basic";
 transcript_support_level "NA"; havana_gene "OTTMUSG00000049935.1"; havana_transcript "OTTMUST00000127109.1";
chr1    ENSEMBL gene    3102016 3102125 .       +       .       gene_id "ENSMUSG00000064842.1"; gene_type "snRNA"; gene_status "KNOWN"; gene_name "Gm26206"; level 3;
chr1    ENSEMBL transcript      3102016 3102125 .       +       .       gene_id "ENSMUSG00000064842.1"; transcript_id "ENSMUST00000082908.1"; gene_type "snRNA"; gene_status "KNO
WN"; gene_name "Gm26206"; transcript_type "snRNA"; transcript_status "KNOWN"; transcript_name "Gm26206-201"; level 3; tag "basic"; transcript_support_level "NA";
chr1    ENSEMBL exon    3102016 3102125 .       +       .       gene_id "ENSMUSG00000064842.1"; transcript_id "ENSMUST00000082908.1"; gene_type "snRNA"; gene_status "KNOWN"; gen
e_name "Gm26206"; transcript_type "snRNA"; transcript_status "KNOWN"; transcript_name "Gm26206-201"; exon_number 1; exon_id "ENSMUSE00000522066.1"; level 3; tag "basic"; transcr
ipt_support_level "NA";
chr1    HAVANA  gene    3205901 3671498 .       -       .       gene_id "ENSMUSG00000051951.5"; gene_type "protein_coding"; gene_status "KNOWN"; gene_name "Xkr4"; level 2; havan
a_gene "OTTMUSG00000026353.2";
chr1    HAVANA  transcript      3205901 3216344 .       -       .       gene_id "ENSMUSG00000051951.5"; transcript_id "ENSMUST00000162897.1"; gene_type "protein_coding"; gene_st
atus "KNOWN"; gene_name "Xkr4"; transcript_type "processed_transcript"; transcript_status "KNOWN"; transcript_name "Xkr4-003"; level 2; transcript_support_level "1"; havana_gene
 "OTTMUSG00000026353.2"; havana_transcript "OTTMUST00000086625.1";
chr1    HAVANA  exon    3213609 3216344 .       -       .       gene_id "ENSMUSG00000051951.5"; transcript_id "ENSMUST00000162897.1"; gene_type "protein_coding"; gene_status "KN
OWN"; gene_name "Xkr4"; transcript_type "processed_transcript"; transcript_status "KNOWN"; transcript_name "Xkr4-003"; exon_number 1; exon_id "ENSMUSE00000858910.1"; level 2; tr
anscript_support_level "1"; havana_gene "OTTMUSG00000026353.2"; havana_transcript "OTTMUST00000086625.1";
chr1    HAVANA  exon    3205901 3207317 .       -       .       gene_id "ENSMUSG00000                   ein_coding"; gene_status "KN
OWN"; gene_name "Xkr4"; transcript_type "processed_transcript"; transcript_status "KN                    E00000866652.1"; level 2; tr
anscript_support_level "1"; havana_gene "OTTMUSG00000026353.2"; havana_transcript "OT
chr1    HAVANA  transcript      3206523 3215632 .       -       .       gene_id "ENSM                   pe "protein_coding"; gene_st
atus "KNOWN"; gene_name "Xkr4"; transcript_type "processed_transcript"; transcript_status "KNOWN"; transcript_name "Xkr4-002"; level 2; transcript_support_level "1"; havana_gene
 "OTTMUSG00000026353.2"; havana_transcript "OTTMUST00000086624.1";
chr1    HAVANA  exon    3213439 3215632 .       -       .       gene_id "ENSMUSG00000051951.5"; transcript_id "ENSMUST00000159265.1"; gene_type "protein_coding"; gene_status "KN
OWN"; gene_name "Xkr4"; transcript_type "processed_transcript"; transcript_status "KNOWN"; transcript_name "Xkr4-002"; exon_number 1; exon_id "ENSMUSE00000863980.1"; level 2; tr
anscript_support_level "1"; havana_gene "OTTMUSG00000026353.2"; havana_transcript "OTTMUST00000086624.1";
chr1    HAVANA  exon    3206523 3207317 .       -       .       gene_id "ENSMUSG00000051951.5"; transcript_id "ENSMUST00000159265.1"; gene_type "protein_coding"; gene_status "KN
OWN"; gene_name "Xkr4"; transcript_type "processed_transcript"; transcript_status "KNOWN"; transcript_name "Xkr4-002"; exon_number 2; exon_id "ENSMUSE00000867897.1"; level 2; tr
anscript_support_level "1"; havana_gene "OTTMUSG00000026353.2"; havana_transcript "OTTMUST00000086624.1";
chr1    HAVANA  transcript      3214482 3671498 .       -       .       gene_id "ENSMUSG00000051951.5"; transcript_id "ENSMUST00000070533.4"; gene_type "protein_coding"; gene_st
atus "KNOWN"; gene_name "Xkr4"; transcript_type "protein_coding"; transcript_status "KNOWN"; transcript_name "Xkr4-001"; level 2; protein_id "ENSMUSP00000070648.4"; tag "basic";
 transcript_support_level "1"; tag "appris_principal_1"; tag "CCDS"; ccdsid "CCDS14803.1"; havana_gene "OTTMUSG00000026353.2"; havana_transcript "OTTMUST00000065166.1";
```

What does all this mean?

the GTF/GFF file format

Fields

Fields **must** be tab-separated. Also, all but the final field in each feature line must contain a value; "empty" columns should be denoted with a '.'

1. **seqname** - name of the chromosome or scaffold; chromosome names can be given with or without the 'chr' prefix. **Important note**: the seqname must be one used within Ensembl, chromosome name or an Ensembl identifier such as a scaffold ID, without any additional content such as species or assembly. See the example GFF output below.
2. **source** - name of the program that generated this feature, or the data source (database or project name)
3. **feature** - feature type name, e.g. Gene, Variation, Similarity
4. **start** - Start position of the feature, with sequence numbering starting at 1.
5. **end** - End position of the feature, with sequence numbering starting at 1.
6. **score** - A floating point value.
7. **strand** - defined as + (forward) or - (reverse).
8. **frame** - One of '0', '1' or '2'. '0' indicates that the first base of the feature is the first base of a codon, '1' that the second base is the first base of a codon, and so on..
9. **attribute** - A semicolon-separated list of tag-value pairs, providing additional information about each feature.

- tab-delimited
- one line per feature
- 9+ columns
- Sample gtf -

```
1 transcribed_unprocessed_pseudogene  gene        11869 14409 . + . gene_id "ENSG00000223972"; gene_name "DDX11L1";
1 processed_transcript                transcript  11869 14409 . + . gene_id "ENSG00000223972"; transcript_id "ENST0
```

Keep in mind …

- UCSC and Ensembl use slightly different naming conventions; Try to stick to one source
- Know exactly which version of genome and annotation you are working with. And make sure they match!
- Ensure your GTF file is correctly formatted
- At every stage, Get to know your data!

Alignment


Raw data QC
Alignment/Map to the reference genome
Evaluate mapping

Mapping reads to the reference genome

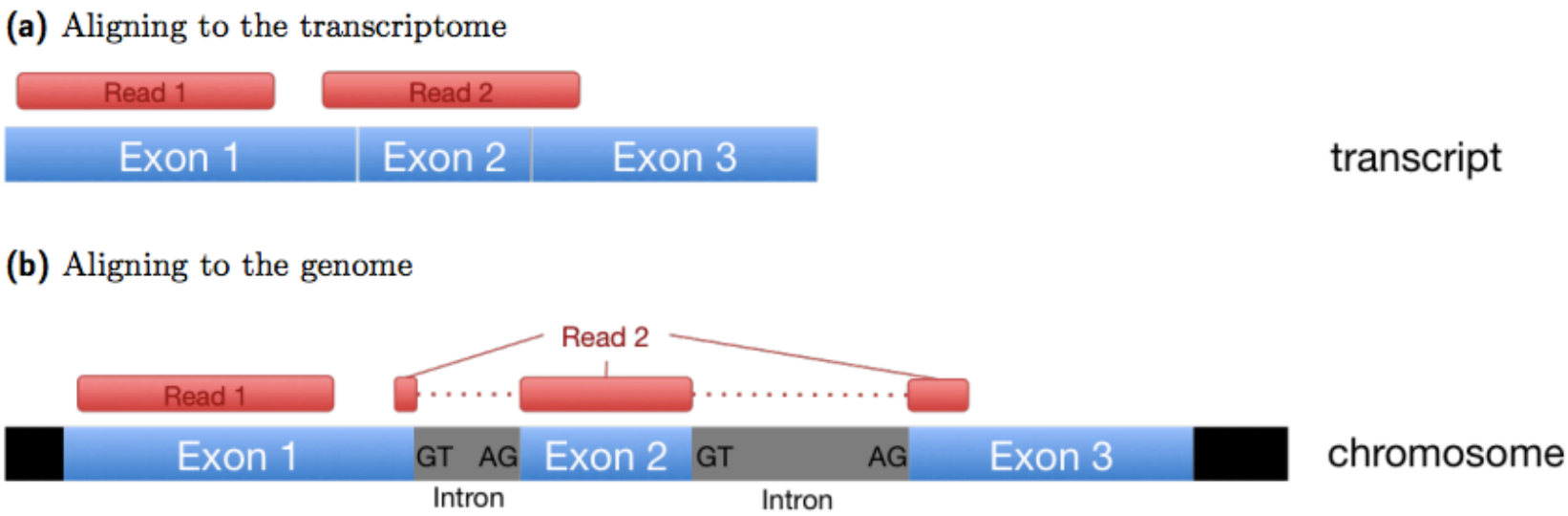


Figure 8: RNA-seq of mRNAs produces two kinds of reads: single exon reads (Read 1) and exon-exon-spanning reads (Read 2). While single exon reads can be aligned equally easily to the genome and to the transcriptome, exon-exon-spanning reads have to be split in order to be aligned properly if only the genome sequence is used as a reference (b).

Examples of software : STAR, HISAT, TopHat

http://chagall.med.cornell.edu/RNASEQcourse/

Example tool : STAR

Bioinformatics. 2013 Jan 1;29(1):15-21. doi: 10.1093/bioinformatics/bts635. Epub 2012 Oct 25.

STAR: ultrafast universal RNA-seq aligner.

Dobin A[1], Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR.

Author information

Abstract

MOTIVATION: Accurate alignment of high-throughput RNA-seq data is a challenging an
contiguous transcript structure, relatively short read lengths and constantly increasing th
available RNA-seq aligners suffer from high mapping error rates, low mapping speed, re

RESULTS: To align our large (>80 billon reads) ENCODE Transcriptome RNA-seq datas
to a Reference (STAR) software based on a previously undescribed RNA-seq alignmen
seed search in uncompressed suffix arrays followed by seed clustering and stitching pro
of >50 in mapping speed, aligning to the human genome 550 million 2 × 76 bp paired-en
at the same time improving alignment sensitivity and precision. In addition to unbiased c
discover non-canonical splices and chimeric (fusion) transcripts, and is also capable of
454 sequencing of reverse transcription polymerase chain reaction amplicons, we exper
junctions with an 80-90% success rate, corroborating the high precision of the STAR ma

AVAILABILITY AND IMPLEMENTATION: STAR is implemented as a standalone C++ co
under GPLv3 license and can be downloaded from http://code.google.com/p/rna-star/.

STAR manual 2.6.1a

Alexander Dobin
dobin@cshl.edu

August 14, 2018

Contents

https://github.com/alexdobin/STAR

Genome Index Generation

- index files will include
  - the genome sequence
  - suffix arrays (table of k-mers)
  - chromosome names and lengths
  - splice junction coordinates
  - gene information, eg strand

Book Index

The alignment

STAR Input

```
STAR \
--genomeDir $GENOME_DIR \
--readFilesIn $READS_DIR$file \
--outFileNamePrefix $OUTPUT_DIR${sample}/${sample}. \
--outReadsUnmapped Fastx \
--outSAMtype BAM SortedByCoordinate
```

Reference genome

Input file

Output file directory

Output file type

Think about -

- handling of multi-mapped reads (how is the best alignment score assigned? How are secondary alignments reported?)
- optimization for very small genomes
- defining min and max intron sizes allowed
- handling genomes with large number of scaffolds (draft genomes)
- using STAR for the detection of chimeric and circular transcripts

STAR Output

```
AACTTGAC.Aligned.sortedByCoord.out.bam
AACTTGAC.Log.final.out
AACTTGAC.Log.out
AACTTGAC.Log.progress.out
AACTTGAC.SJ.out.tab
AACTTGAC.Unmapped.out.mate1
```

Alignment output file

Alignment logs

splice junctions details

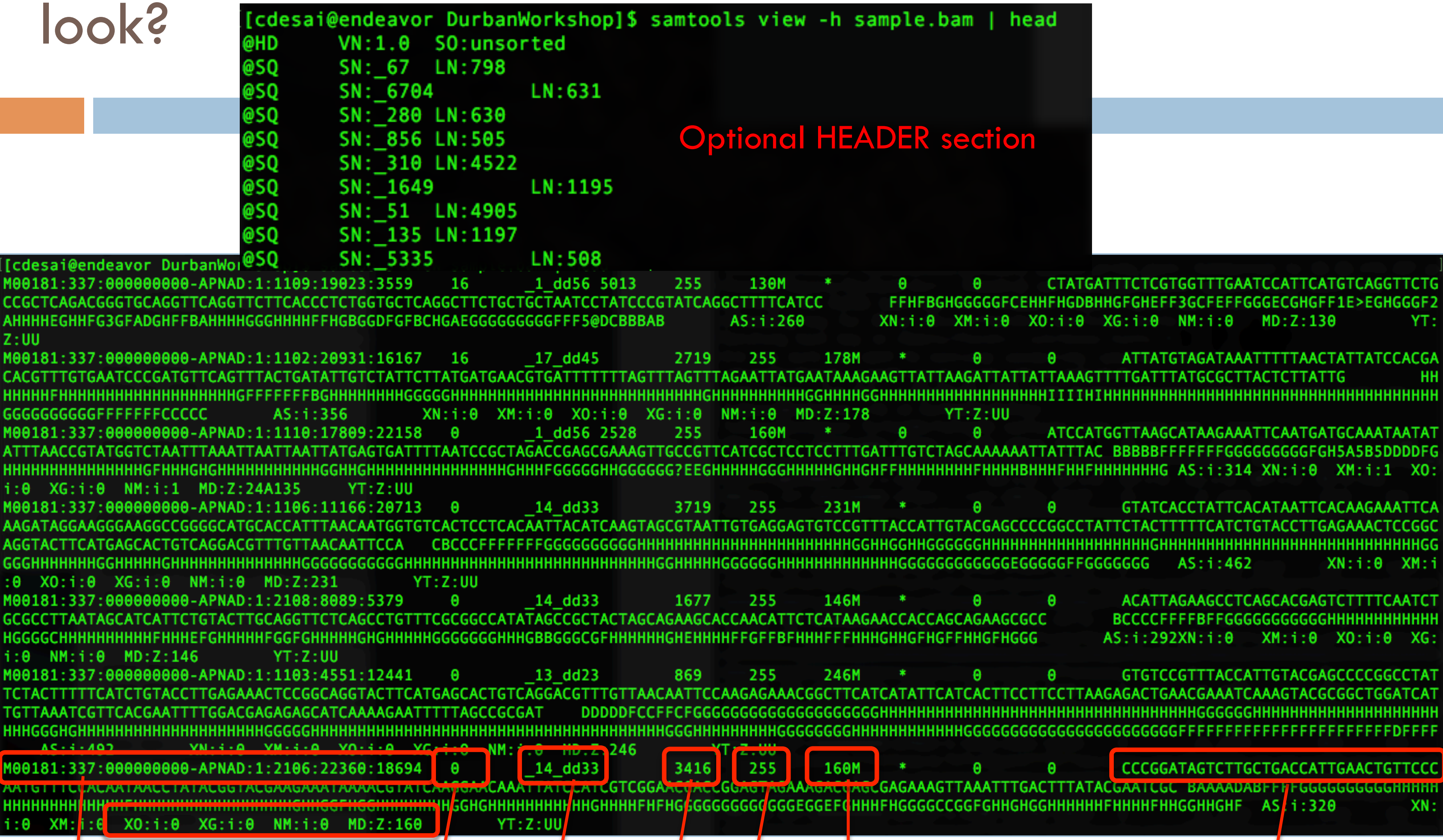
unmapped reads

Alignment log file –
How well did my reads align to the reference?

```
[cdesai@endeavor AACTTGAC]$ more AACTTGAC.Log.final.out
                                 Started job on |       Sep 24 19:33:33
                             Started mapping on |       Sep 24 19:35:59
                                    Finished on |       Sep 24 19:36:35
       Mapping speed, Million of reads per hour |       285.19

                          Number of input reads |       2851948
                      Average input read length |       50
                                    UNIQUE READS:
                   Uniquely mapped reads number |       2302024
                        Uniquely mapped reads % |       80.72%
                          Average mapped length |       49.82
                       Number of splices: Total |       238207
            Number of splices: Annotated (sjdb) |       235646
                       Number of splices: GT/AG |       235942
                       Number of splices: GC/AG |       1420
                       Number of splices: AT/AC |       385
               Number of splices: Non-canonical |       460
                      Mismatch rate per base, % |       0.13%
                         Deletion rate per base |       0.00%
                        Deletion average length |       1.31
                        Insertion rate per base |       0.00%
                       Insertion average length |       1.12
                             MULTI-MAPPING READS:
        Number of reads mapped to multiple loci |       465987
             % of reads mapped to multiple loci |       16.34%
        Number of reads mapped to too many loci |       20831
             % of reads mapped to too many loci |       0.73%
                                  UNMAPPED READS:
       % of reads unmapped: too many mismatches |       0.00%
                 % of reads unmapped: too short |       1.37%
                     % of reads unmapped: other |       0.84%
```

Alignment output file - How does an alignment output look?



```
[cdesai@endeavor DurbanWorkshop]$ samtools view -h sample.bam | head
@HD     VN:1.0  SO:unsorted
@SQ     SN:_67  LN:798
@SQ     SN:_6704        LN:631
@SQ     SN:_280 LN:630
@SQ     SN:_856 LN:505
@SQ     SN:_310 LN:4522
@SQ     SN:_1649        LN:1195
@SQ     SN:_51  LN:4905
@SQ     SN:_135 LN:1197
@SQ     SN:_5335        LN:508
```

Optional HEADER section

```
[cdesai@endeavor DurbanWo
M00181:337:000000000-APNAD:1:1109:19023:3559    16      _1_dd56 5013    255     130M    *       0       0       CTATGATTTCTCGTGGTTTGAATCCATTCATGTCAGGTTCTG
CCGCTCAGACGGGTGCAGGTTCAGGTTCTTCACCCTCTGGTGCTCAGGCTTCTGCTGCTAATCCTATCCCGTATCAGGCTTTTCATCC       FFHFBGHGGGGGFCEHHFHGDBHHGFGHEFF3GCFEFFGGGECGHGFF1E>EGHGGGF2
AHHHHEGHHFG3GFADGHFFBAHHHHGGGHHHHFFHGBGGDFGFBCHGAEGGGGGGGGGFFF5@DCBBBAB       AS:i:260        XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0   MD:Z:130        YT:
Z:UU
M00181:337:000000000-APNAD:1:1102:20931:16167   16      _17_dd45        2719    255     178M    *       0       0       ATTATGTAGATAAATTTTTAACTATTATCCACGA
CACGTTTGTGAATCCCGATGTTCAGTTTACTGATATTGTCTATTCTTATGATGAACGTGATTTTTTTAGTTTAGTTTAGAATTATGAATAAAGAAGTTATTAAGATTATTATTAAAGTTTTGATTTATGCGCTTACTCTTATTG          HH
HHHHHFHHHHHHHHHHHHHHHHHHHGFFFFFFBGHHHHHHHHGGGGHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHGHHHHHHHHHHGGHHHHGGHHHHHHHHHHHHHHHHHHIIIHIHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH
GGGGGGGGGGFFFFFFFCCCCC        AS:i:356        XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:178        YT:Z:UU
M00181:337:000000000-APNAD:1:1110:17809:22158   0       _1_dd56 2528    255     160M    *       0       0       ATCCATGGTTAAGCATAAGAAATTCAATGATGCAAATAATAT
ATTTAACCGTATGGTCTAATTTAAATTAATTAATTATGAGTGATTTTAATCCGCTAGACCGAGCGAAAGTTGCCGTTCATCGCTCCTCCTTTGATTTGTCTAGCAAAAAATTATTTAC BBBBBFFFFFFFGGGGGGGGGFGH5A5B5DDDDFG
HHHHHHHHHHHHHHHGFHHHGHGHHHHHHHHHHHHGGHHGHHHHHHHHHHHHHHHHHGHHHFGGGGGHHGGGGGG?EEGHHHHHHGGGHHHHHHGHHGHFFHHHHHHHHHFHHHHBHHHFHHFHHHHHHHHG AS:i:314 XN:i:0  XM:i:1  XO:
i:0  XG:i:0  NM:i:1  MD:Z:24A135      YT:Z:UU
M00181:337:000000000-APNAD:1:1106:11166:20713   0       _14_dd33        3719    255     231M    *       0       0       GTATCACCTATTCACATAATTCACAAGAAATTCA
AAGATAGGAAGGGAAGGCCGGGGCATGCACCATTTAACAATGGTGTCACTCCTCACAATTACATCAAGTAGCGTAATTGTGAGGAGTGTCCGTTTACCATTGTACGAGCCCCGGCCTATTCTACTTTTTCATCTGTACCTTGAGAAACTCCGGC
AGGTACTTCATGAGCACTGTCAGGACGTTTGTTAACAATTCCA   CBCCCFFFFFFFGGGGGGGGGGHHHHHHHHHHHHHHHHHHHHHHHHHHHHGGHHGGHHGGGGGGHHHHHHHHHHHHHHHHHHHHGHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHGG
GGGHHHHHHHHGGHHHHHGHHHHHHHHHHHHHHHGGGGGGGGGGGHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHGGHHHHHGGGGGGHHHHHHHHHHHHHHHGGGGGGGGGGGGEGGGGGFFGGGGGGG   AS:i:462        XN:i:0  XM:i
:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:231        YT:Z:UU
M00181:337:000000000-APNAD:1:2108:8089:5379     0       _14_dd33        1677    255     146M    *       0       0       ACATTAGAAGCCTCAGCACGAGTCTTTTCAATCT
GCGCCTTAATAGCATCATTCTGTACTTGCAGGTTCTCAGCCTGTTTCGCGGCCATATAGCCGCTACTAGCAGAAGCACCAACATTCTCATAAGAACCACCAGCAGAAGCGCC        BCCCCFFFFBFFGGGGGGGGGGHHHHHHHHHH
HGGGGCHHHHHHHHHHHFHHHEFGHHHHHFGGFGHHHHHGHGHHHHHGGGGGGGHHHGBBGGGCGFHHHHHHGHEHHHHFFGFFBFHHHFFFHHHGHHGFHGFFHHGFHGGG        AS:i:292XN:i:0   XM:i:0  XO:i:0  XG:
i:0  NM:i:0  MD:Z:146        YT:Z:UU
M00181:337:000000000-APNAD:1:1103:4551:12441    0       _13_dd23        869     255     246M    *       0       0       GTGTCCGTTTACCATTGTACGAGCCCCGGCCTAT
TCTACTTTTTCATCTGTACCTTGAGAAACTCCGGCAGGTACTTCATGAGCACTGTCAGGACGTTTGTTAACAATTCCAAGAGAAACGGCTTCATCATATTCATCACTTCCTTCCTTAAGAGACTGAACGAAATCAAAGTACGCGGCTGGATCAT
TGTTAAATCGTTCACGAATTTTGGACGAGAGAGCATCAAAAGAATTTTTAGCCGCGAT      DDDDDFCCFFCFGGGGGGGGGGGGGGGGGGGGGHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHGGGGGHHHHHHHHHHHHHHHHHHHH
HHHGGGHGHHHHHHHHHHHHHHHHHHHHHGGGGGHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHGGGHHHHHHHHHHGGGGGGGHHHHHHHHHHHHGGGGGGGGGGGGGGGGGGGGGGGGFFFFFFFFFFFFFFFFFFFFFDFFFF
    AS:i:492        XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:246        YT:Z:UU
M00181:337:000000000-APNAD:1:2106:22360:18694   0       _14_dd33        3416    255     160M    *       0       0       CCCGGATAGTCTTGCTGACCATTGAACTGTTCCC
AATGTTTCGACAATAACCTATACGGTACGAAGAAATAAAACGTATCAAGGAACAAATTATCCATCGTCGGAACGGACCGGAGTAGAAAGACGAGCGAGAAAGTTAAATTTGACTTTATACGAATCGC  BAAAADABFFFFGGGGGGGGGGHHHHH
HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHGHHGGFHGGHHHHHHHHHGGHGHHHHHHHHHHHHGHHHHHFHFHGGGGGGGGGGGGEGGEFGHHHFHGGGGCCGGFGHHGHGGHHHHHHFHHHHHFHHGGHHGHF  AS:i:320        XN:
i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:160        YT:Z:UU
```

query

optional fields

flag

refname

pos

mapq

CIGAR

Sequence of aligned read

What is this 'flag' column?

Bit		Description
1	0x1	template having multiple segments in sequencing
2	0x2	each segment properly aligned according to the aligner
4	0x4	segment unmapped
8	0x8	next segment in the template unmapped
16	0x10	SEQ being reverse complemented
32	0x20	SEQ of the next segment in the template being reverse complemented
64	0x40	the first segment in the template
128	0x80	the last segment in the template
256	0x100	secondary alignment
512	0x200	not passing filters, such as platform/vendor quality controls
1024	0x400	PCR or optical duplicate
2048	0x800	supplementary alignment

Confusing? I have some examples!

77	$(= 1 + 4 + 8 + 64)$	The read is paired, is the first read in the pair, both are unmapped.
83	$(= 1 + 2 + 16 + 64)$	The read is paired, mapped in a proper pair, is the first read in the pair, and it is mapped to the reverse strand.
99	$(= 1 + 2 + 32 + 64)$	The read is paired, mapped in a proper pair, is the first read in the pair, and its mate is mapped to the reverse strand.
133	$(= 1 + 4 + 128)$	The read is paired, is the second read in the pair, and it is unmapped.
137	$(= 1 + 8 + 128)$	The read is paired, is the second read in the pair, and it is mapped while its mate is not.

Still not happy? Go to –
https://broadinstitute.github.io/picard/explain-flags.html

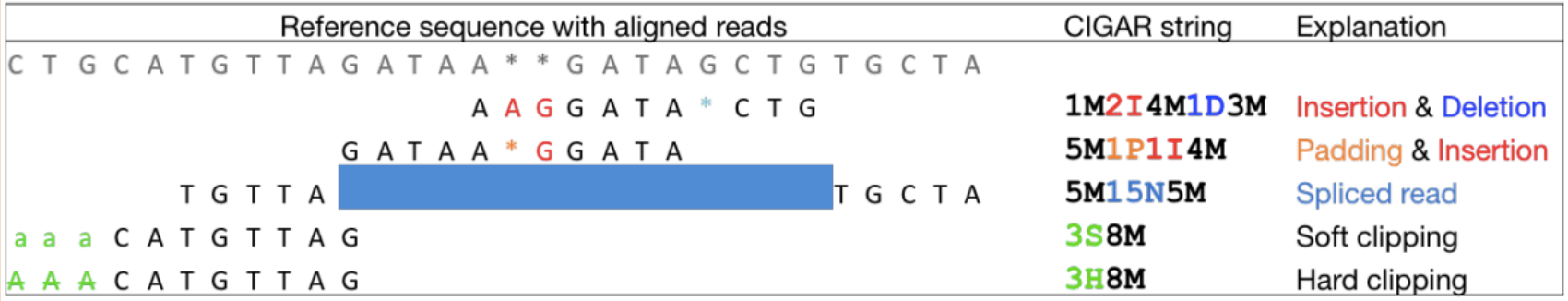
Okay, what about CIGAR?

CIGAR [Concise Idiosyncratic Gapped Alignment Report] String The sixth field of a `SAM` file contains a so-called `CIGAR` string indicating which *operations* were necessary to map the read to the reference sequence at that particular locus.

The following operations are defined in `CIGAR` format (also see Figure 10):

- `M` Alignment (can be a sequence match or mismatch!)
- `I` Insertion in the read compared to the reference
- `D` Deletion in the read compared to the reference
- `N` Skipped region from the reference. For mRNA-to-genome alignments, an N operation represents an intron. For other types of alignments, the interpretation of `N` is not defined.
- `S` Soft clipping (clipped sequences are present in read); `S` may only have `H` operations between them and the ends of the string
- `H` Hard clipping (clipped sequences are NOT present in the alignment record); can only be present as the first and/or last operation
- `P` Padding (silent deletion from padded reference)
- `=` Sequence match (not widely used)
- `X` Sequence mismatch (not widely used)

The sum of lengths of the `M`, `I`, `S`, `=`, `X` operations must equal the length of the read.



Basic SAM/BAM manipulations

```
# I want to take a peek at how my alignment file looks
samtools view alignment_file.bam | head

# I want to turn my BAM file into a human-readable SAM
samtools view -h alignment_file.bam > alignment_file_bam2sam.sam

# I want to compress my SAM file into BAM format
samtools view -Sb alignment_file.sam > alignment_file_sam2bam.bam

# I want to convert SAM to a SORTED BAM file
samtools view -Sb alignment_file.sam | samtools sort - alignment_file.sorted

# I need to generate an index for a BAM file (needs to  be sorted)
samtools index alignment_file.sorted.bam

# Count how many reads mapped  to each reference fasta/contig
samtools idxstats alignment_file.sorted.bam

# Show me Unmapped reads only!
samtools view -h -b -f 4 alignment_file.sorted.bam > unmapped_reads.bam
```

show header output BAM include only reads where 0x4 bit is set)

```
# Show me Mapped reads only!
samtools view -hb -F 4 alignment_file.sorted.bam > mapped_reads.bam
```

show header output BAM include only reads where 0x4 bit is NOT set)

Evaluate your Alignment - metrics

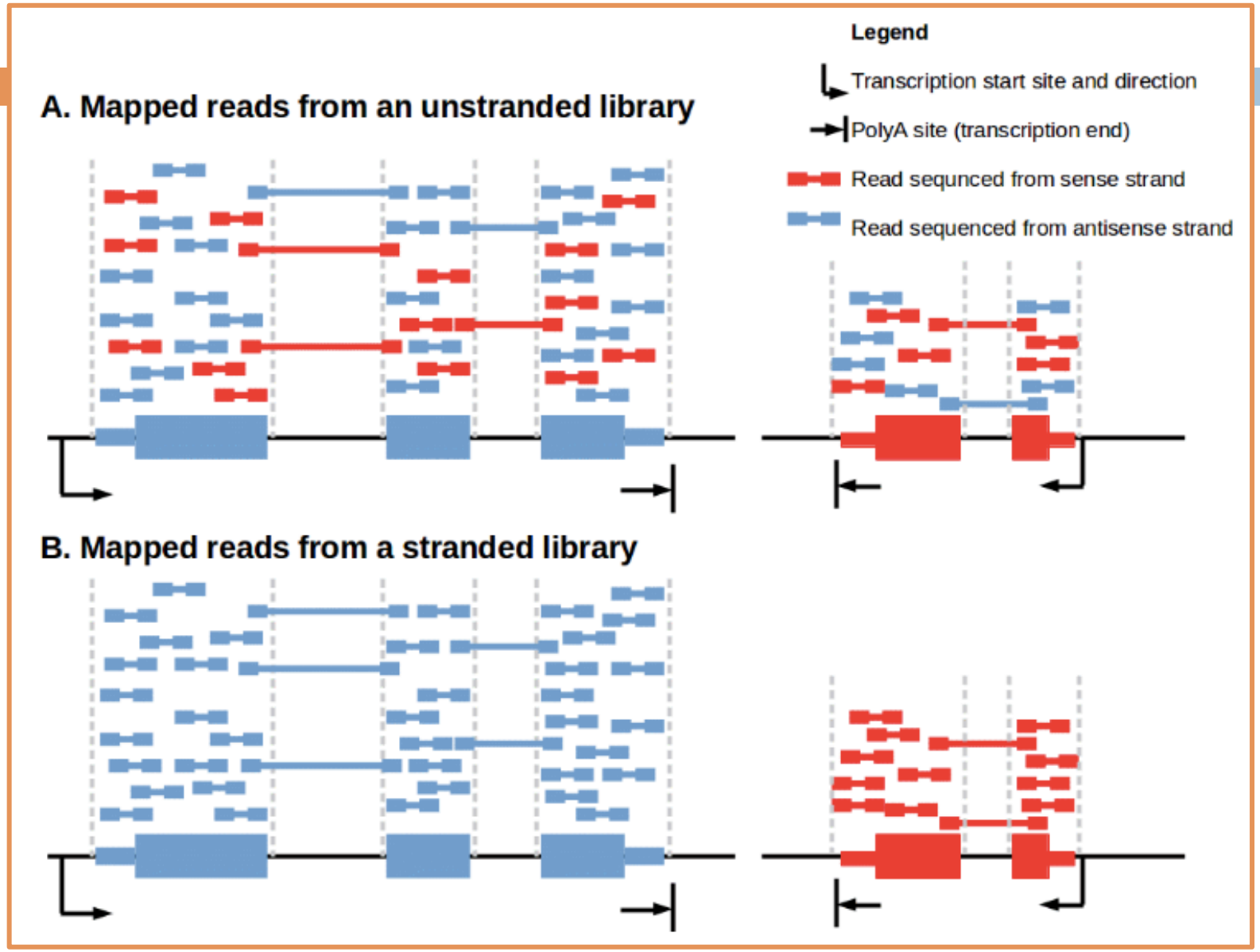
- distribution of the bases in the alignment file
- fractions of nucleotides within specific genomic regions UTRs, introns, intergenic sequences, and exons
- Examples of software : pileup from bbtools, CollectRnaSeqMetrics from GATK, samtools flagstat, RSeQC
- what to look out for –
  - Intron coverage: if many reads align to introns, this is indicative of incomplete poly(A) enrichment or abundant presence of immature transcripts
  - Intergenic reads: if a significant portion of reads is aligned outside of annotated gene sequences, this may suggest genomic DNA contamination (or abundant non-coding transcripts)
  - 3' bias: over-representation of 3' portions of transcripts indicates RNA degradation

Get to know your data!

Visualize your alignment – Example tool : IGV

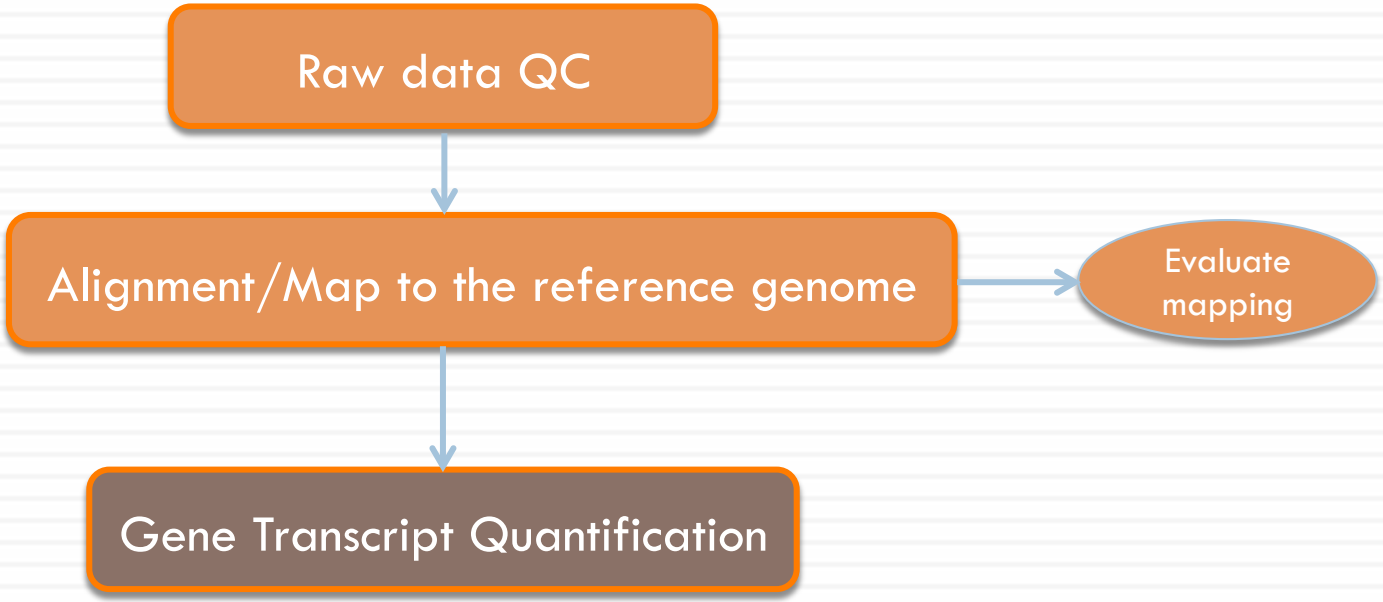
http://software.broadinstitute.org/software/igv/

Stranded or not?



https://www.ecseq.com/support/ngs/how-do-strand-specific-sequencing-protocols-work

Gene transcript Quantification

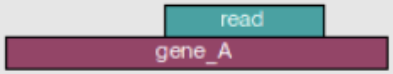
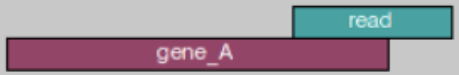
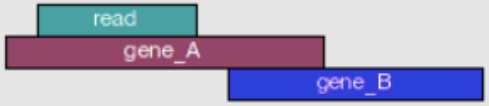
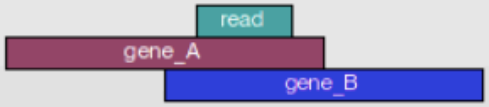


Gene-based read counting

- to compare expression of single genes between different conditions
- to think about while shopping for quantification software; how does the program handle –
  - overlap size (full read vs. partial overlap)
  - multi-mapping reads
  - reads overlapping multiple genomic features of the same kind
  - reads overlapping introns
- Answer will depend on nature of your experiment and the desired outcome
- Examples of software : htseq, featureCounts, cufflinks

	Sample1.Day_3.	Sample2.Day_3	Sample3.Day_3	Sample4.Day_3.	Sample5.Day_3.	Sample6.Day_3.	Sample7.Day_3.	Sample8.Day_3.	Sample9.Day_1.	Sample10.Day_1
ENSMUSG00000000001.4	378	361	501	114	220	167	141	247	274	270
ENSMUSG00000000003.13	0	0	0	0	0	0	0	0	0	0
ENSMUSG00000000028.12	43	14	9	2	7	0	9	0	7	11
ENSMUSG00000000031.13	4	0	5	0	0	0	3	0	1	9
ENSMUSG00000000037.14	1	0	0	0	0	0	0	0	0	1
ENSMUSG00000000049.9	1	5	1	0	6	4	0	0	2	14
ENSMUSG00000000056.7	110	92	150	43	96	60	79	43	129	125
ENSMUSG00000000058.6	5	4	3	20	5	0	14	2	6	3
ENSMUSG00000000078.6	1880	3639	4136	737	2396	1312	1178	6902	2727	2455
ENSMUSG00000000085.14	20	12	14	6	22	14	8	7	18	26

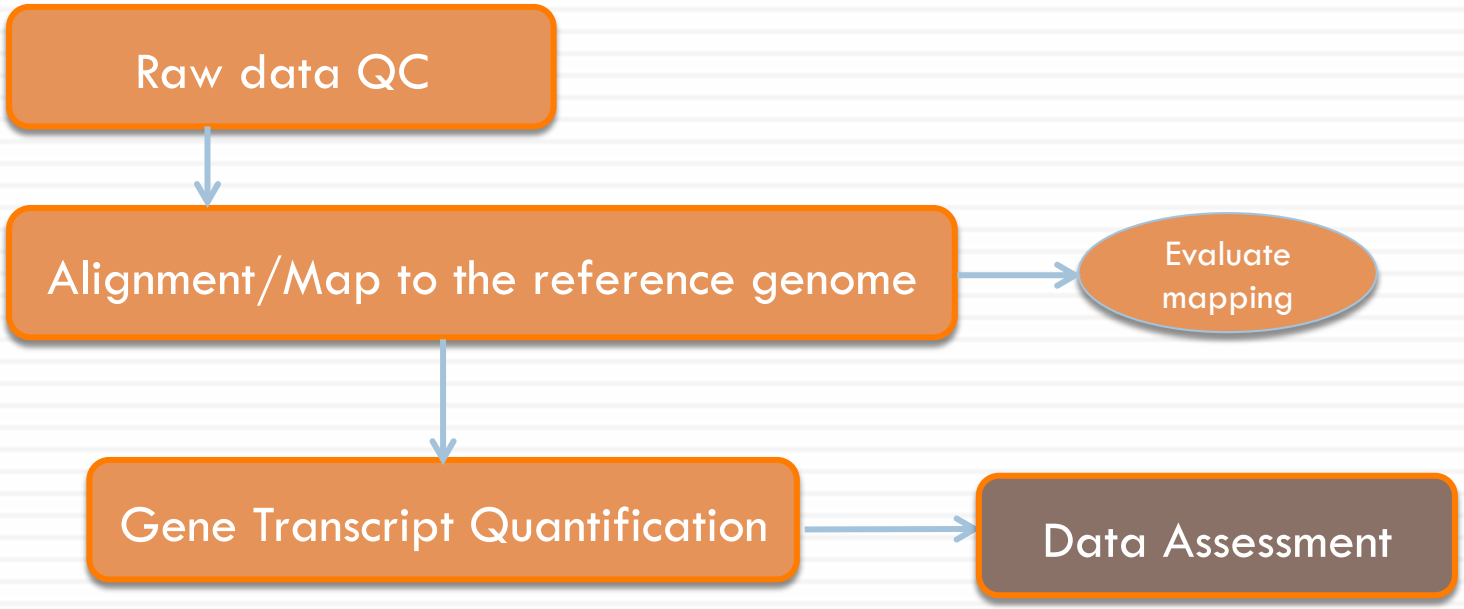
Example tool : htseq-count

	union	intersection _strict	intersection _nonempty
	gene_A	gene_A	gene_A
	gene_A	no_feature	gene_A
	gene_A	no_feature	gene_A
	gene_A	gene_A	gene_A
	gene_A	gene_A	gene_A
	ambiguous	gene_A	gene_A
	ambiguous	ambiguous	ambiguous

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

Data Assessment

Are expectations about global patterns met?



Examining global read count patterns

- Pairwise correlation

ENCODE recommends - for messenger RNA, biological replicates [should] display >0.9 correlation for transcripts/features

- Hierarchical clustering

Clusters the group of samples into a dendogram. Different experimental conditions/treatment groups should fall into well-separated clusters

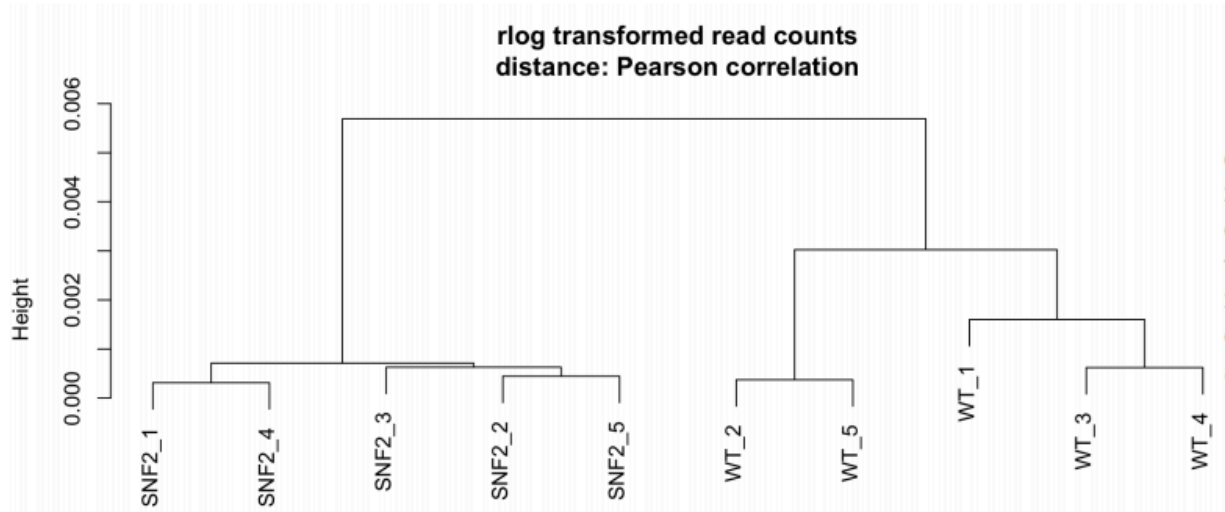
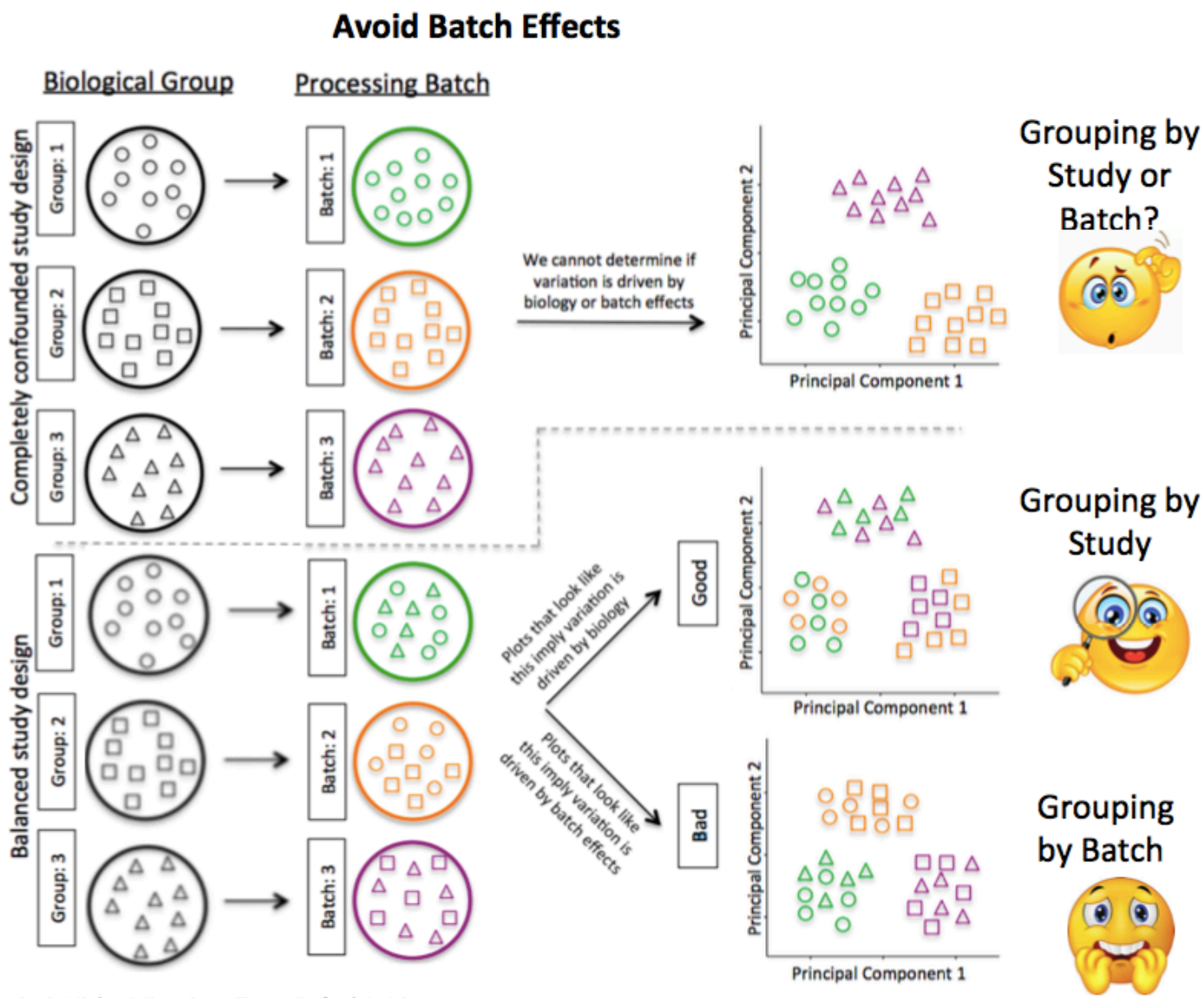


Figure 17: Dendrogram of *rlog*-transformed read counts for ten different samples, using the "complete" linkage function. The two conditions, SNF2 and WT, are well separated.

Get to know your data!

Principal Component Analysis



Adapted from: Stephanie C. Hicks, Mingxiang Teng, Rafael A. Irizarry.
https://www.biorxiv.org/content/early/2015/09/04/025528
On the widespread and critical impact of systematic bias and batch effects in single-cell RNA-Seq data.

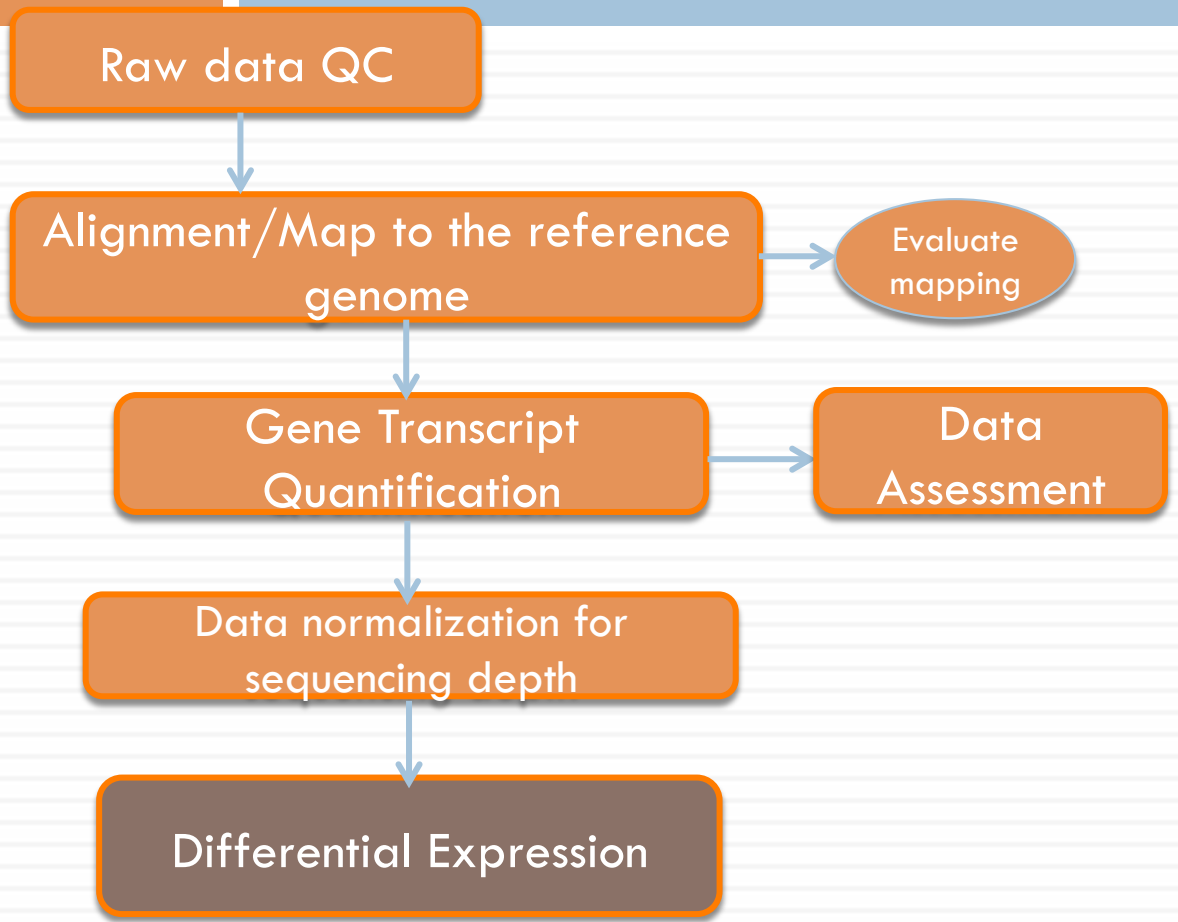
(Explore Batch Removal Techniques)

Table 6: Comparison of unsupervised classification and clustering techniques. The following table was adapted from Karimpour-Fard et al. (2015); see that publication for more details on additional (supervised) classification methods such as support vector machines. Classifiers try to reduce the number of features that represent the most prevalent patterns within the data. Clustering techniques aim to group similar features.

	Method	What does it do?	How?	Strengths	Weaknesses	Sample size
Classification	PCA	Separates features into groups based on commonality and reports the weight of each component's contribution to the separation	Orthogonal transformation; transfers a set of correlated variables into a new set of uncorrelated variables	Unsupervised, nonparametric, useful for reducing dimensions before using supervision	Number of features must exceed number of treatment groups	Number of features must exceed number of treatment groups
	ICA	Separates features into groups by eliminating correlation and reports the weight of each componentâĂŹs contribution to the separation	Nonlinear, non-orthogonal transformation; standardizes each variable to a unit variance and zero mean	Works well when other approaches do not because data are not normally distributed	Features are assumed to be independent when they actually may be dependent	Unlimited sample size; data non-normally distributed
Clustering	K-means	Separates features into clusters of similar expression patterns	Compares and groups magnitudes of changes in the means into K clusters where K is defined by the user	Easily visualized and intuitive; greatly reduces complexity; performs well when distance information between data points is important to clustering	Sensitive to initial conditions and user-specified number of clusters (K)	Best with a limited dataset, i.e., ca. 20 to 300 features
	Hierarchical	Clusters treatment groups, features, or samples into a dendrogram	Compares all samples using either agglomerative or divisive algorithms with distance and linkage functions	Unsupervised; easily visualized and intuitive	Does not provide feature contributions; not iterative, thus sensitive to cluster distance measures and noise and outliers	Best with a limited dataset, i.e., ca. 20 to 300 features or samples

http://chagall.med.cornell.edu/RNASEQcourse/

Differential Expression

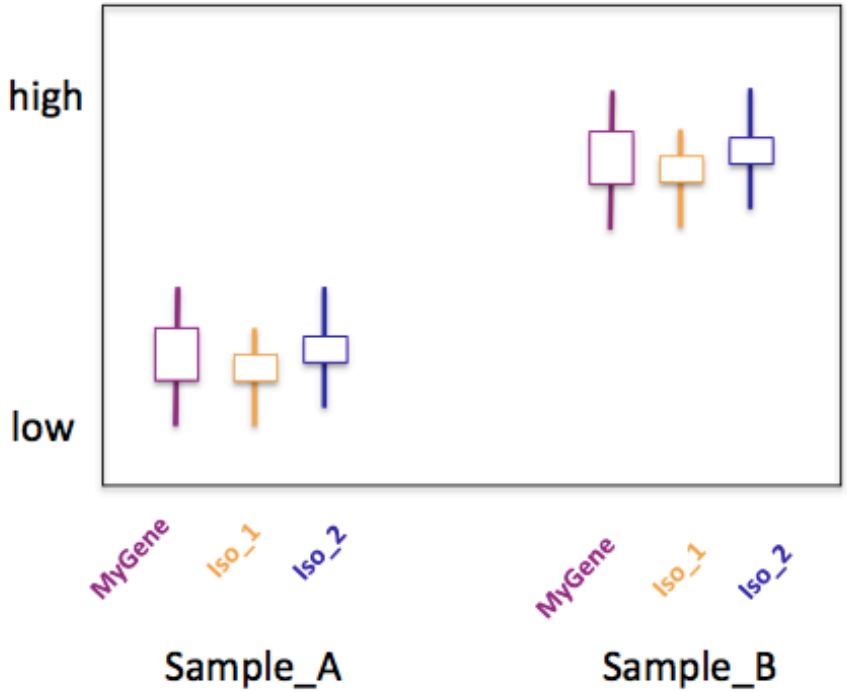


What is Differential ‘Expression’?

Flavours of Differential Expression

- differential gene expression
- differential transcript expression
- differential transcript usage
- differential exon usage

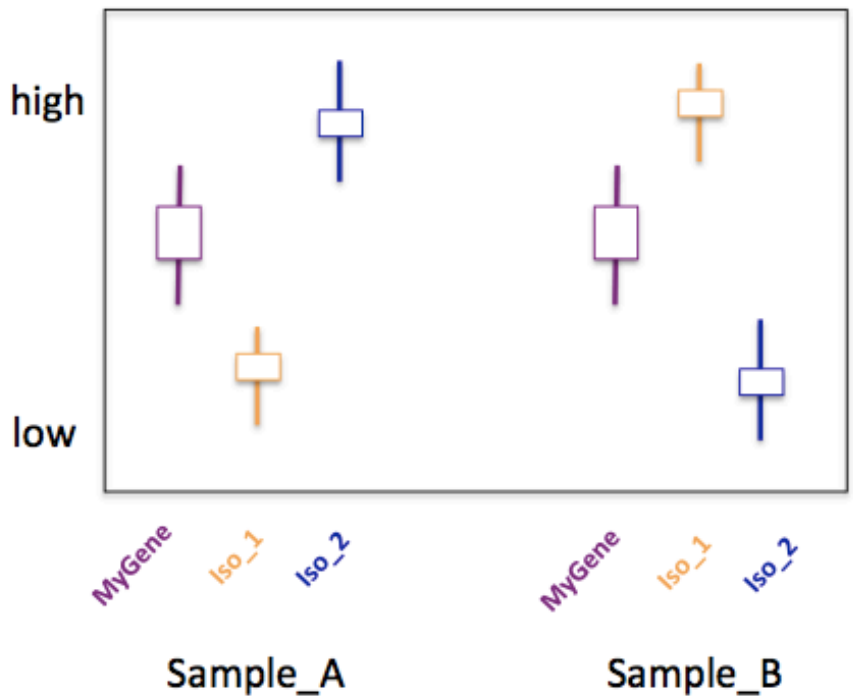
Example1



Feature	Diff Expressed?
MyGene	Yes
Iso_1	Yes
Iso_2	Yes
Diff. Transcript Usage ? (eg. Isoform switching)	No

Brian Haas

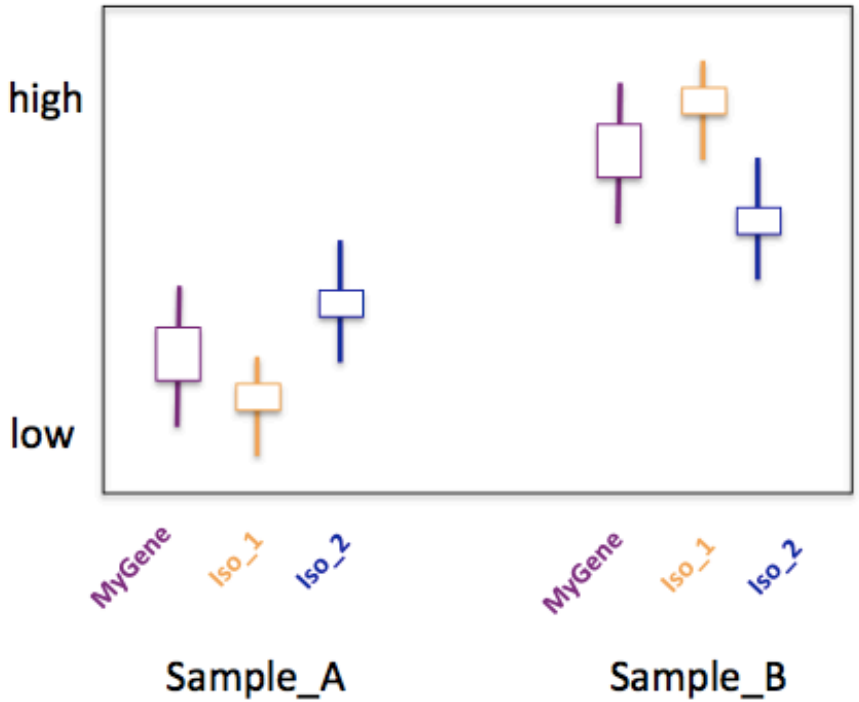
Example2



Feature	Diff Expressed?
MyGene	No
Iso_1	Yes
Iso_2	Yes
Diff. Transcript Usage ? (eg. Isoform switching)	Yes

Brian Haas

Example3



Feature	Diff Expressed?
MyGene	Yes
Iso_1	Yes
Iso_2	Yes
Diff. Transcript Usage ? (eg. Isoform switching)	Yes

Brian Haas

Differential Gene Expression

Phase1. estimate the magnitude of differential expression between two or more conditions based on read counts from replicated samples (taking into account the differences in sequencing depth and variability)

Phase2. Estimate the significance of the difference and correct for multiple testing

	Sample1.Day_3.	Sample2.Day_3	Sample3.Day_3	Sample4.Day_3.	Sample5.Day_3.	Sample6.Day_3.	Sam
ENSMUSG00000000001.4	378	361	501	114	220	167	
ENSMUSG00000000003.13	0	0	0	0	0	0	
ENSMUSG00000000028.12	43	14	9	2	7	0	
ENSMUSG00000000031.13	4	0	5	0	0	0	
ENSMUSG00000000037.14	1	0	0	0	0	0	
ENSMUSG00000000049.9	1	5	1	0	6	4	
ENSMUSG00000000056.7	110	92	150	43	96	60	
ENSMUSG00000000058.6	5	4	3	20	5	0	
ENSMUSG00000000078.6	1880	3639	4136	737	2396	1312	
ENSMUSG00000000085.14	20	12	14	6	22	14	
ENSMUSG00000000088.6	1365	1353	834	376	755	665	

SampleName	condition
Sample1.Day_3.E	Day_3_B6_no_abx
Sample2.Day_3.E	Day_3_B6_no_abx
Sample3.Day_3.E	Day_3_B6_+_abx
Sample4.Day_3.E	Day_3_B6_+_abx
Sample5.Day_3.E	Day_3_B6_no_abx
Sample6.Day_3.E	Day_3_B6_no_abx
Sample7.Day_3.E	Day_3_B6_+_abx
Sample8.Day_3.E	Day_3_B6_+_abx
Sample9.Day_1.E	Day_1_B6_no_abx
Sample10.Day_1	Day_1_B6_no_abx
Sample11.Day_1	Day_1_B6_+_abx
Sample12.Day_1	Day_1_B6_+_abx
Sample13.Day_1	Day_1_B6_no_abx
Sample15.Day_1	Day_1_B6_+_abx
Sample16.Day_1	Day_1_B6_+_abx

Examples of software : DESeq2, edgeR, limma-voom

Comparison of programs for differential gene expression identification

Feature	DESeq2	edgeR	limmaVoom	Cuffdiff
Seq. depth normalization	Sample-wise size factor	Gene-wise trimmed median of means (TMM)	Gene-wise trimmed median of means (TMM)	FPKM-like or DESeq-like
Assumed distribution	Neg. binomial	Neg. binomial	*log*-normal	Neg. binomial
Test for DE	Exact test (Wald)	Exact test for over-dispersed data	Generalized linear model	*t*-test
False positives	Low	Low	Low	High
Detection of differential isoforms	No	No	No	Yes
Support for multi-factored experiments	Yes	Yes	Yes	No
Runtime (3-5 replicates)	Seconds to minutes	Seconds to minutes	Seconds to minutes	Hours

http://chagall.med.cornell.edu/RNASEQcourse/

Example tool : DESeq2

Genome Biol. 2014; 15(12): 550.
Published online 2014 Dec 5. doi: 10.1186/s13059-014-0550-8
PMCID: PMC4302049
PMID: 25516281

Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2

Michael I Love, Wolfgang Huber, and Simon Anders

Author information ▶ Article notes ▶ Copyright and License information ▶ Discla

This article has been cited by other articles in PMC.

Abstract

In comparative high-throughput sequencing assays, a fundame
as read counts per gene in RNA-seq, for evidence of systemat
Small replicate numbers, discreteness, large dynamic range an
statistical approach. We present *DESeq2*, a method for differe
estimation for dispersions and fold changes to improve stabili
enables a more quantitative analysis focused on the strength r
expression. The *DESeq2* package is available at
http://www.bioconductor.org/packages/release/bioc/html/DES

Analyzing RNA-seq data with DESeq2

Michael I. Love, Simon Anders, and Wolfgang Huber

04/30/2018

Abstract

A basic task in the analysis of count data from RNA-seq is the detection of differentially expressed genes. The count data are presented as a table which reports, for each sample, the number of sequence fragments that have been assigned to each gene. Analogous data also arise for other assay types, including comparative ChIP-Seq, HiC, shRNA screening, mass spectrometry. An important analysis question is the quantification and statistical inference of systematic changes between conditions, as compared to within-condition variability. The package DESeq2 provides methods to test for differential expression by use of negative binomia generalized linear models; the estimates of dispersion and logarithmic fold changes incorporate data-driven prior distributions This vignette explains the use of the package and demonstrates typical workflows. An RNA-seq workflow on the Bioconductor website covers similar material to this vignette but at a slower pace, including the generation of count matrices from FASTQ files. DESeq2 package version: 1.20.0

https://bioconductor.org/packages/release/bioc/html/DESeq2.html

Normalizing for sequencing depth differences

- DESeq's default method to normalize read counts to account for differences in sequencing depths is implemented in estimateSizeFactors

```
# calculate the size factor and add it to the data set
> DESeq.ds <- estimateSizeFactors(DESeq.ds)
> sizeFactors(DESeq.ds)

# counts() allows you to immediately retrieve the _normalized_ read counts
> counts.sf_normalized <- counts(DESeq.ds, normalized = TRUE)
```

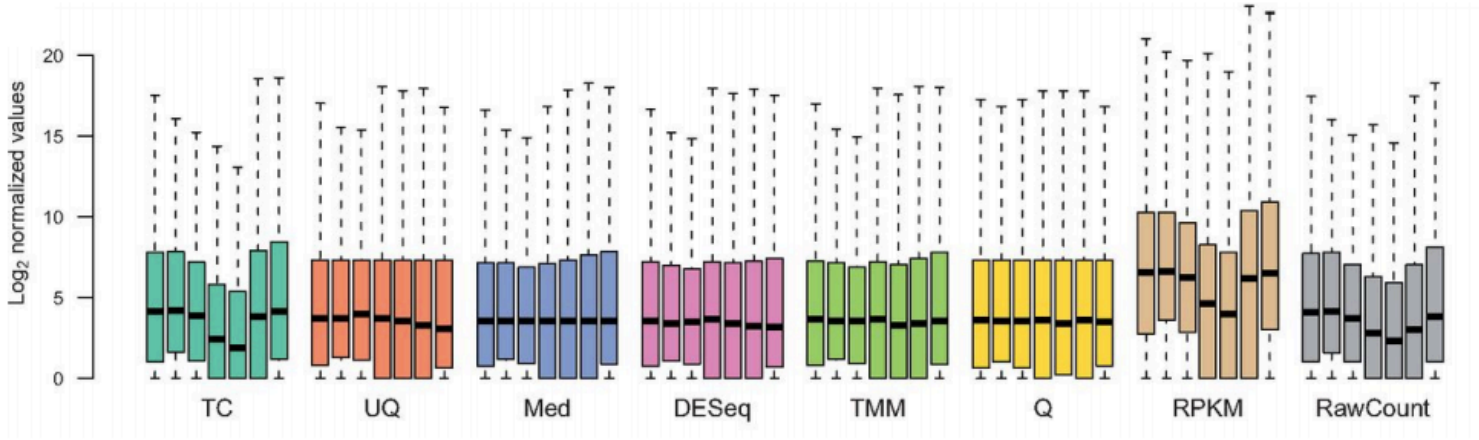


Figure 14: Figure from Dillies et al. (2013) that shows the effects of different approaches to normalize for read count differences due to library sizes (TC, total count; UQ, upper quartile; Med, median; DESeq, size factor; TMM, Trimmed Mean of M-values; Q, quantile) or gene lengths (RPKM). See Tables 13 and 14 for details of the different normalization methods.

http://chagall.med.cornell.edu/RNASEQcourse/

Finding differentially expressed genes

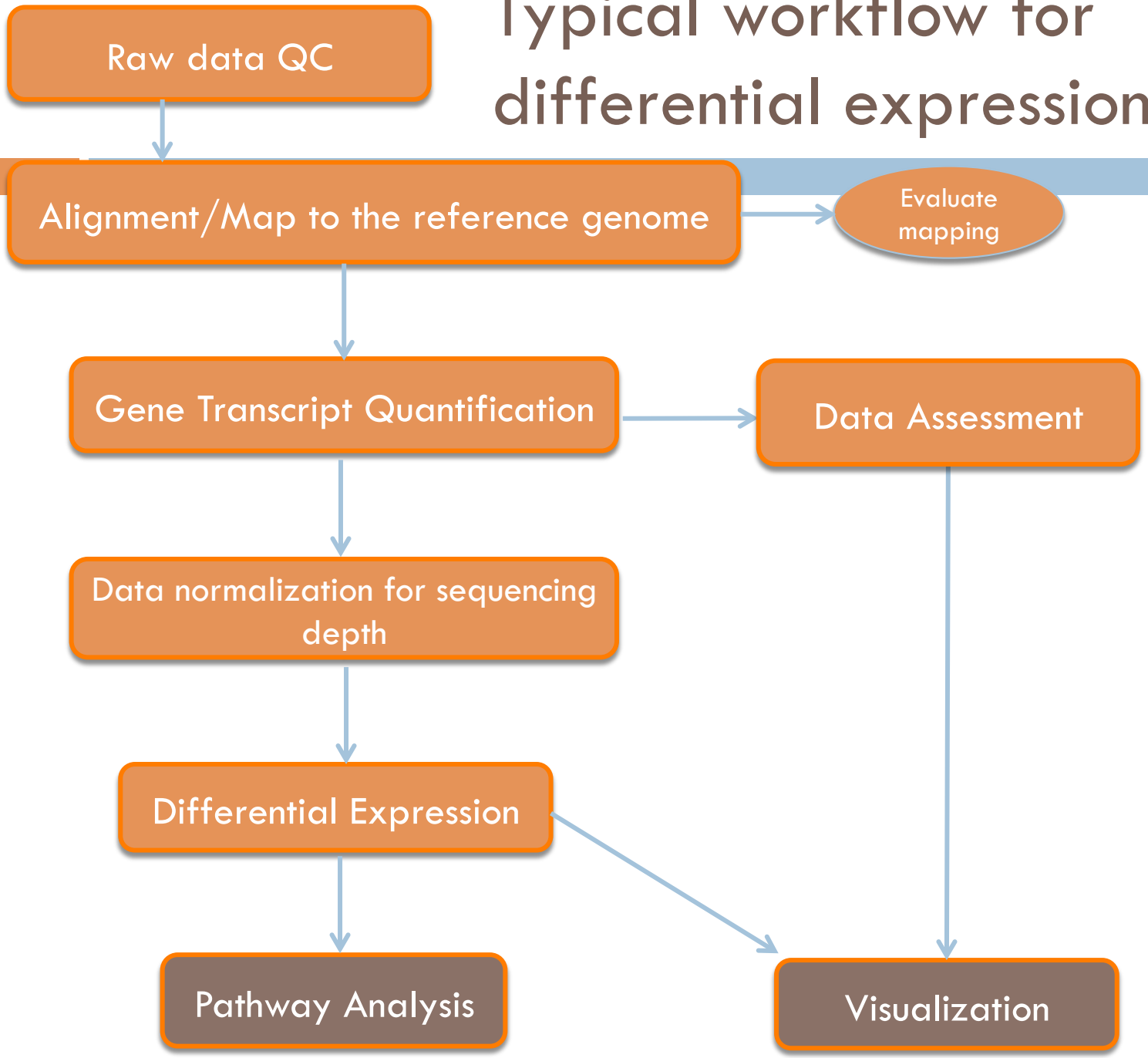
```
dds <- DESeq(dds)
res <- results(dds)
res
```

```
## log2 fold change (MLE): condition treated vs untreated
## Wald test p-value: condition treated vs untreated
## DataFrame with 9921 rows and 6 columns
##                     baseMean      log2FoldChange             lfcSE
##                    <numeric>           <numeric>         <numeric>
## FBgn0000008 95.1442917575889 0.00227644123005389 0.223728651436475
## FBgn0000014 1.05652281859341  -0.495120386382503  2.14318579455575
## FBgn0000017 4352.55356876647  -0.239918943537385 0.126336905277886
## FBgn0000018  418.61048415965  -0.104673911941152 0.148489059621453
## FBgn0000024   6.406199980976   0.210847791726071 0.689587552519466
## ...                      ...                 ...               ...
## FBgn0261570 3208.38861003698   0.295532889721694 0.127350479150082
## FBgn0261572 6.19718814545467  -0.958822964551161 0.775314665308774
```

- what do these columns exactly mean?

1. Gene identifier (from the annotation file)
2. baseMean : mean normalized counts, averaged over all samples from both conditions
3. log2FoldChange : log2 of the fold change in expression of that gene
4. lfcSE : standard error estimate for the log2 fold change estimate
5. stat: : Wald statistic
6. pval : p value for the statistical significance of this change
7. padj : p value adjusted for multiple testing with the Benjamini-Hochberg procedure which controls the FDR

Typical workflow for differential expression analysis



What more can I do with this data?

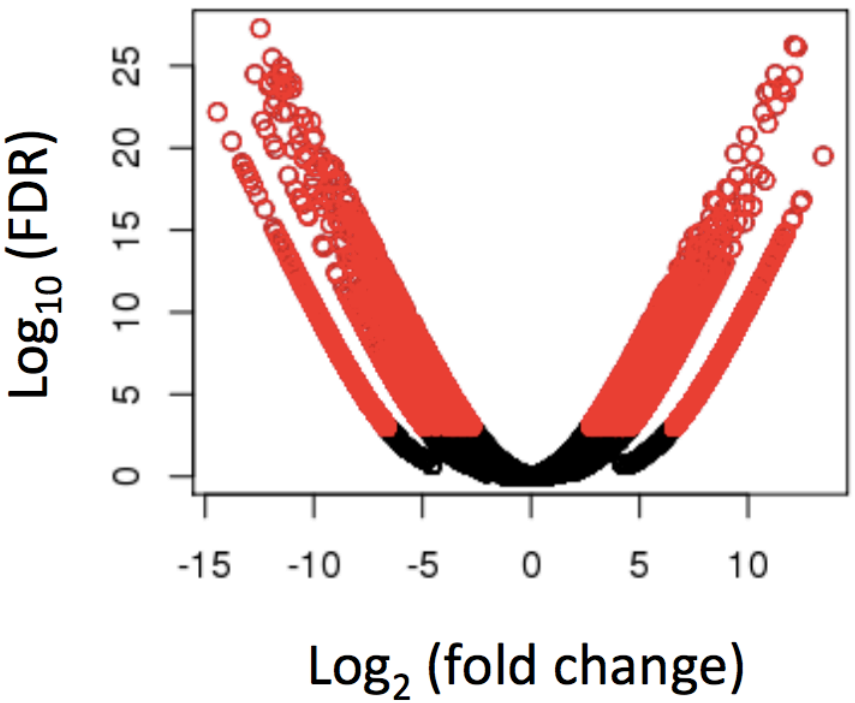
Volcano plots

Heatmaps

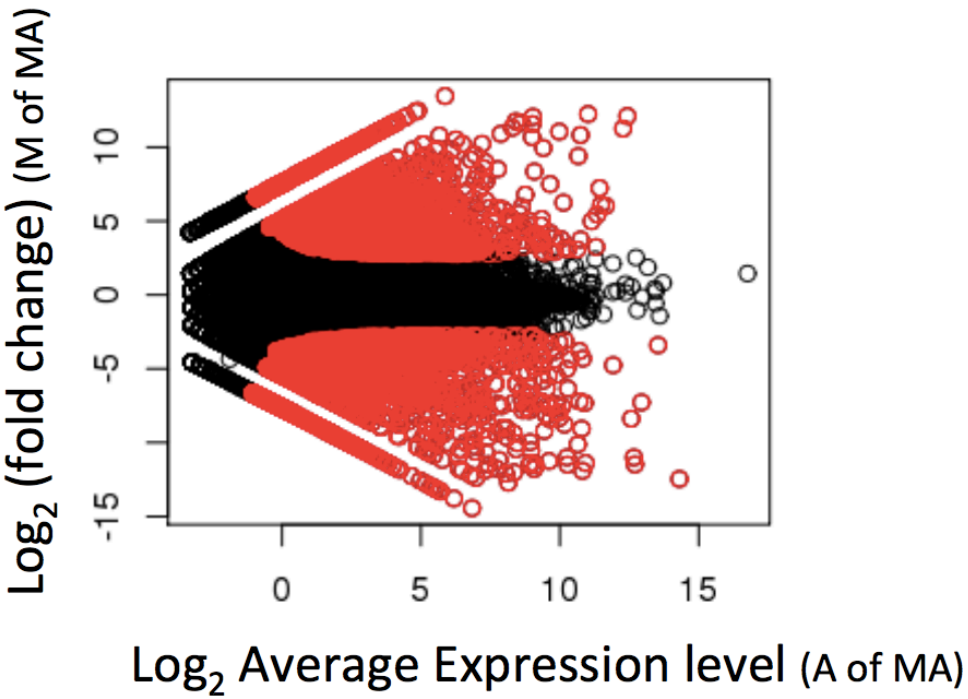
Pathway Analysis

Plotting pairwise differential expression

Volcano plot
(fold change vs significance)



MA plot
abundance vs fold change)

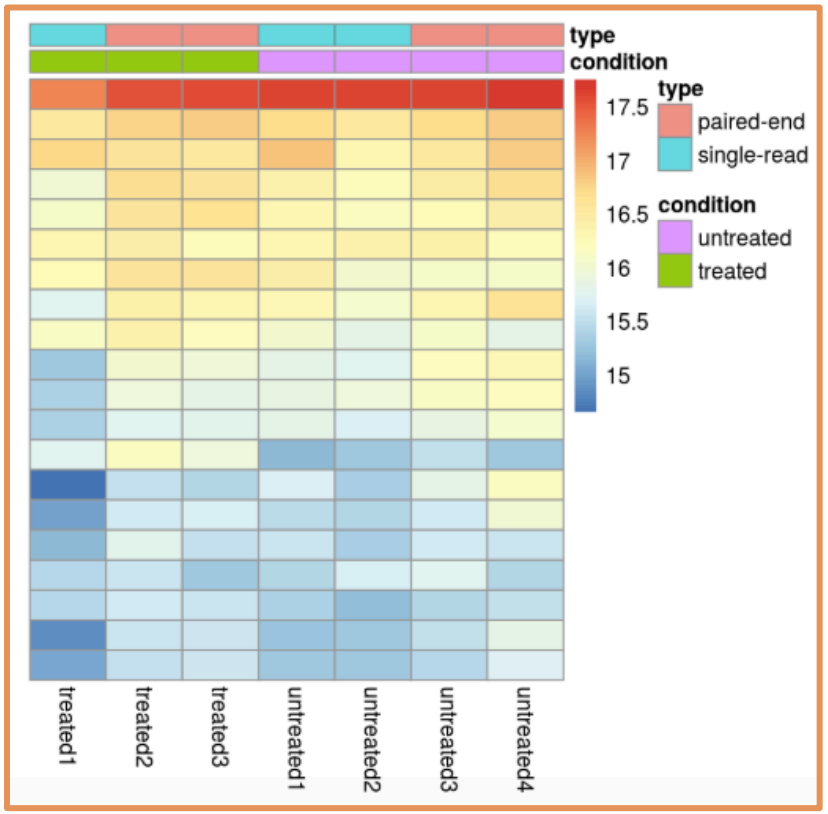


Significantly differently expressed transcripts have FDR <= 0.001 (shown in red)

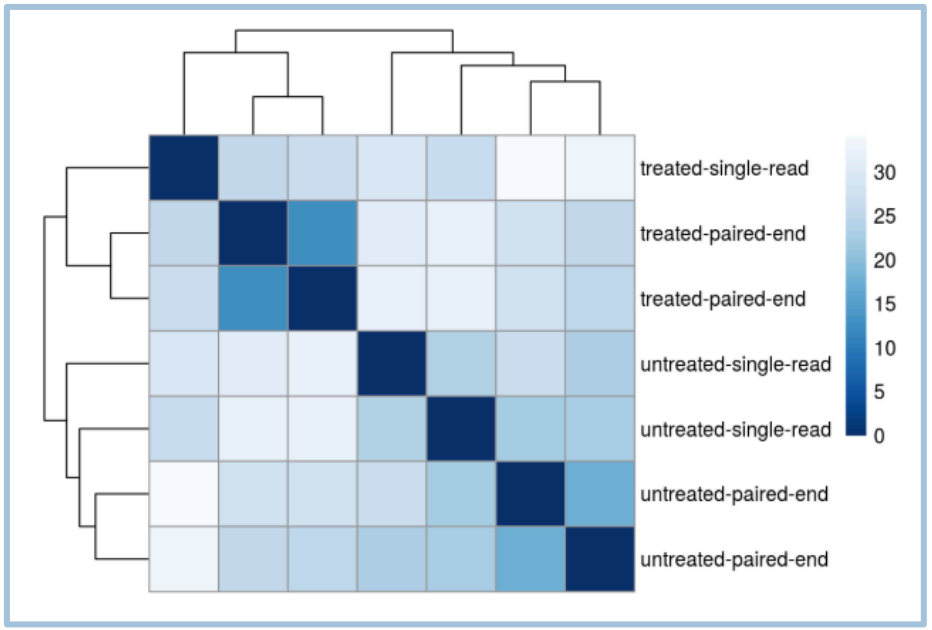
https://bioconductor.org/packages/release/bioc/html/DESeq2.html

Heatmaps

Heatmap of the expression counts



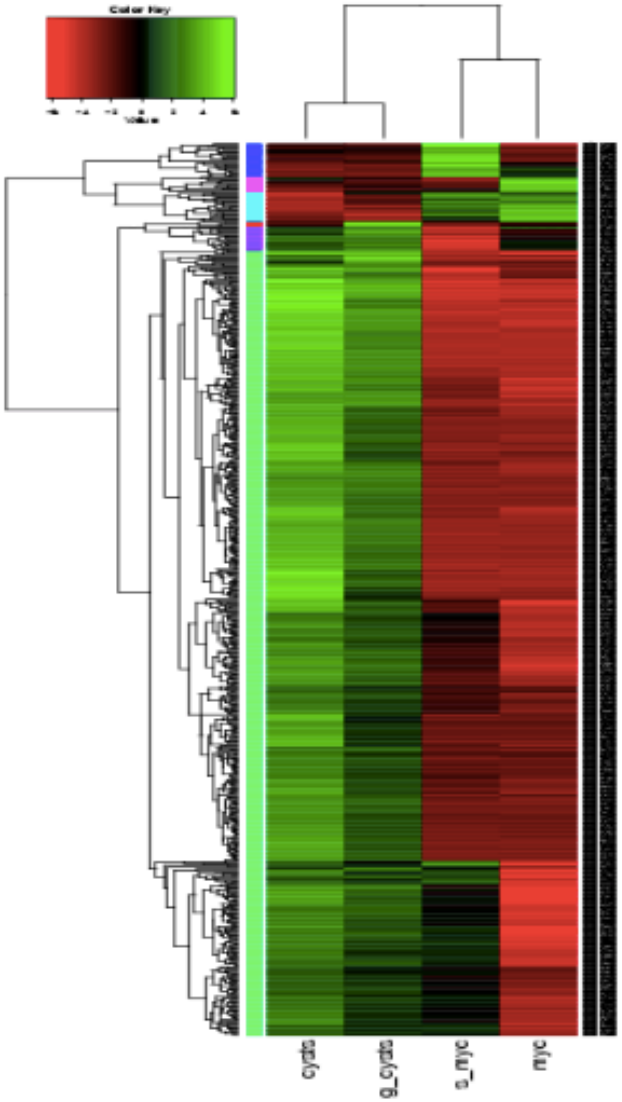
Heatmap of sample-to-sample distances



https://bioconductor.org/packages/release/bioc/html/DESeq2.html

Note: PCA and clustering should be done on normalized and preferably transformed read counts, so that the high variability of low read counts does not occlude potentially informative data trends

Multiple samples, multiple genes, clustering



Heatmaps provide an effective tool for navigating differential expression across multiple samples.

Clustering can be performed across both axes:

-cluster transcripts with similar expression patters.

-cluster samples according to similar expression values among transcripts.

Brian Haas

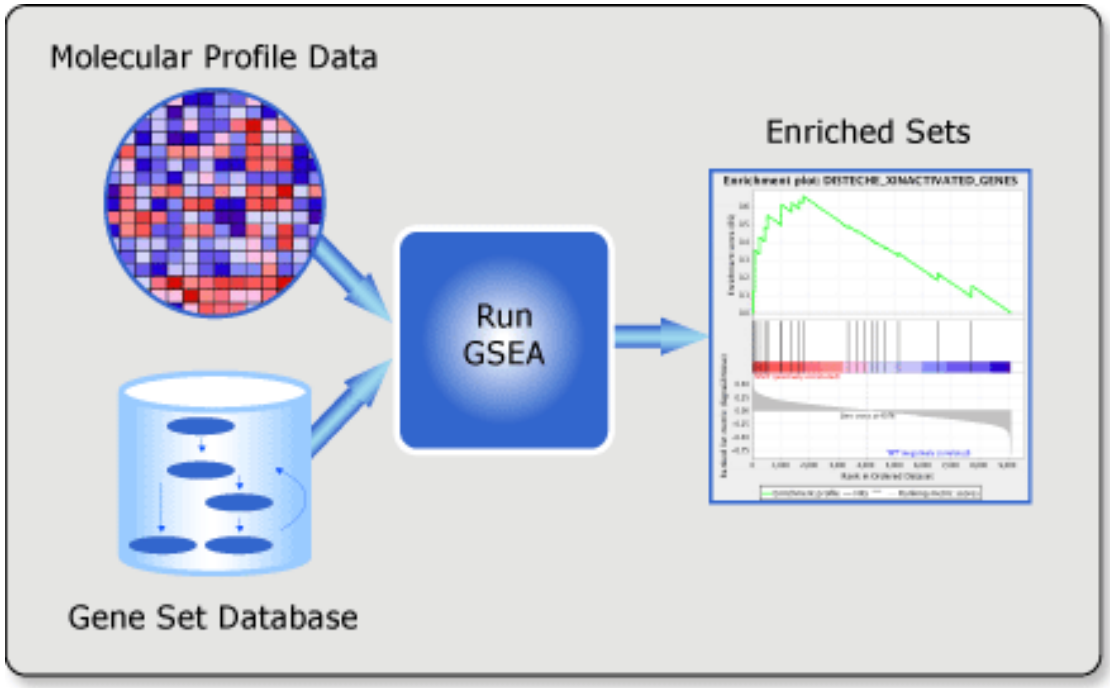
Pathway Analysis

- common approach to interpreting gene expression data
- pathways - simple sets of genes and an enrichment p-value is calculated for each
- gene set enrichment analysis
  - based on the functional annotation of the genes
  - useful for finding out if the expressed genes in your dataset are associated with a certain biological process or molecular function

Limitations

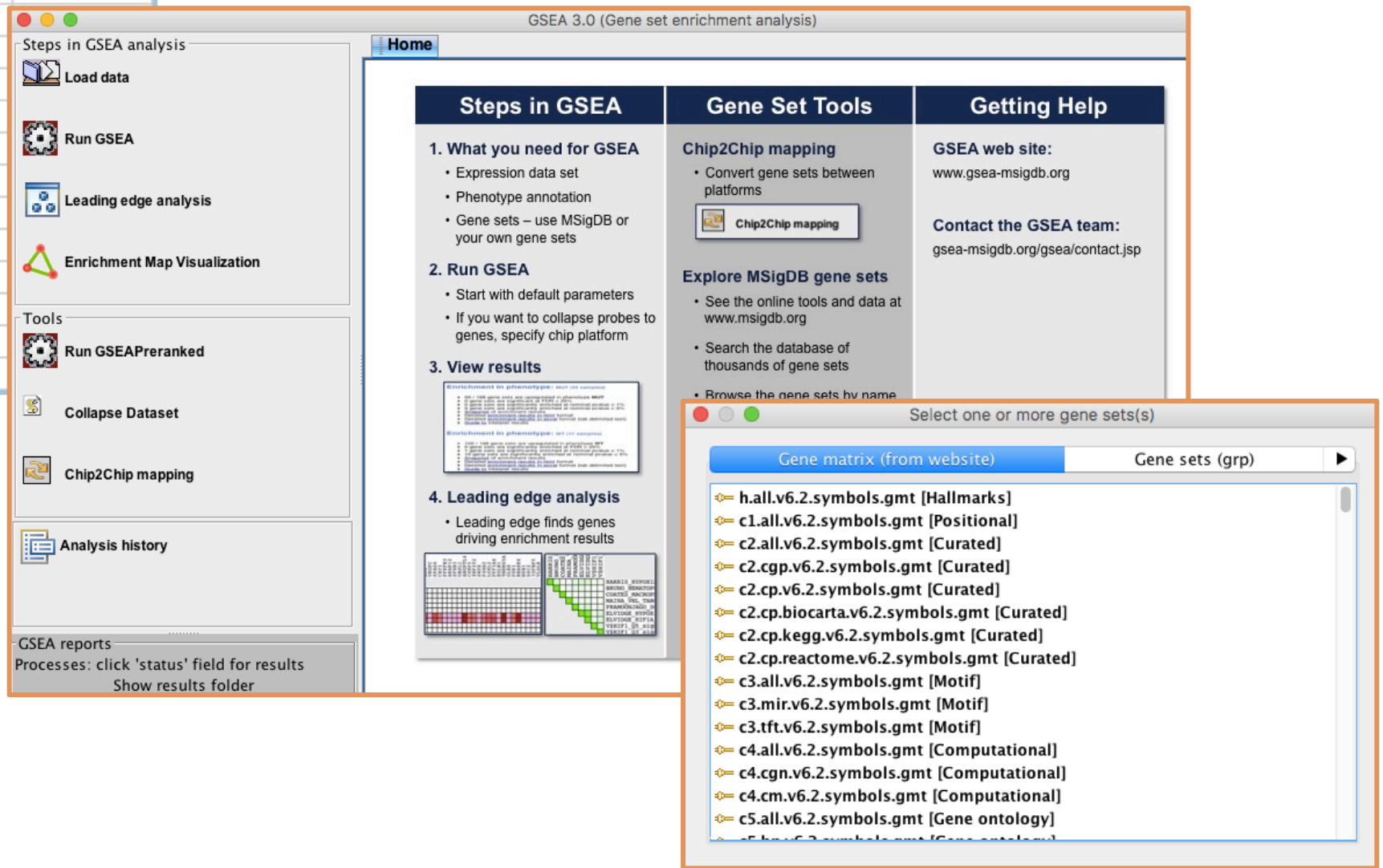
- p-values are calculated based on the assumption that all variables (genes) are **independent** (while the pathways are there precisely to tell you how these genes influence each other)
- pathways are treated as simple bags of genes, disregarding all the phenomena and interactions between genes that they describe.

Examples of software : GSEA, Ingenuity, DAVID



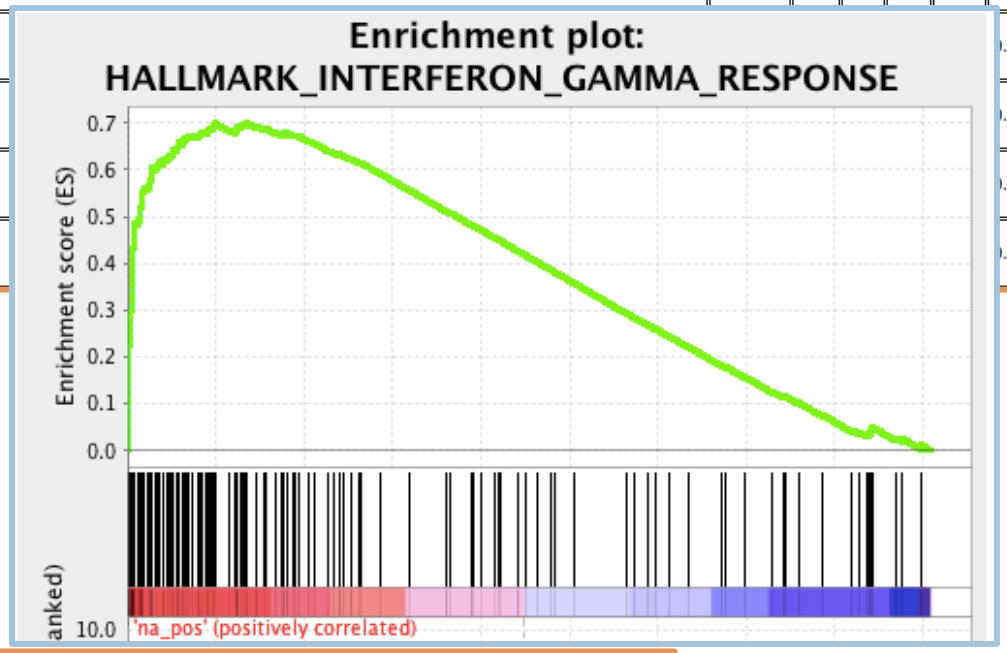
Example tool : GSEA

SLFN4	18.1360813
IFI44	16.1983401
OAS3	15.2804733
LY6a	14.5654609
BST2	14.0001159
IFIT3	13.9294652
RSAD2	13.9285877
USP18	13.7588896
IFIT2	
TRIM30a	
ISG15	
OAS2	
CMPK2	
APOL9b	
APOL9a	
IFIT3b	
IRF7	
IFIT1	
DHX58	
PHF11b	



GSEA results

	GS follow link to MSigDB	GS DETAILS	SIZE	ES	NES	NOM p-val	FDR q-val	FWER p-val
1	HALLMARK_INTERFERON_GAMMA_RESPONSE	Details ...	158	0.70	2.97	0.000	0.000	0.000
2	HALLMARK_INTERFERON_ALPHA_RESPONSE	Details ...	79	0.73	2.80	0.000	0.000	0.000
3	HALLMARK_ALLOGRAFT_REJECTION						.000	0.000
4	REACTOME_IMMUNE_SYSTEM						.000	0.000
5	REACTOME_INTERFERON_ALPHA_BETA_SIGNALING						.000	0.000
6	REACTOME_INTERFERON_SIGNALING						.000	0.000



Gene Set: HALLMARK_INTERFERON_GAMMA_RESPONSE

Standard name	HALLMARK_INTERFERON_GAMMA_RESPONSE
Systematic name	M5913
Brief description	Genes up-regulated in response to IFNG [GeneID=3458].

Congratulations!
You have successfully completed this course!