Arabidopsis data set we are going to use today

Dataset **SRR074262** and **SRR074285**:  
Arabidopsis thaliana  
Unstranded RNA-Seq  
Illumina single-end  
6 samples in 6 runs:  
  3 from plants infected with a pathogen  
  3 from plants after mock treatment  

Messy fastq files
Human data set we are going to use today

Dataset SRR1039509_1 and SRR1039509_2

Human
Unstranded RNA-Seq
Illumina paired-end

16 samples in 32 runs:
- 4 from cell lines treated with dexamethasone
- 4 from same cell lines treated with albuterol
- 4 from same cell lines treated with both asthma medications
- 4 from same untreated cell lines cultured in parallel

Most RNA-Seq analysis tools only work in Linux/Mac via commands

```bash
cmd="bowtie2 -p ${nthr} \\
    -x ${ref} \\
    -q \\
    -1 ${fq1} \\
    -2 ${fq2} \\
    --phred33 \\
    --fr \\
    -I 0 \\
    -X 500 \\
    --un-gz ${outfolder}/bowtie_unmapped-reads.sam.gz \\
    --end-to-end \\
    --sensitive \\
    --seed 2014 \\
    -S ${outfolder}/${bowtiepe}.sam \\
    -u 100"
```
Galaxy and GenePattern make these tools easy to use

**bowtie index**
- **Select a file**
- **Upload your own file**

Homo_sapiens.Ensembl.GRCh37

A bowtie 2 index. Select a prebuilt index or upload your own as a ZIP file (you can make one using Bowtie.indexer).

**input format**
- **FASTQ**

The format of the reads input files.

**reads pair 1**
- **https://dev.bits.vib.be:8686/gp/data//data/genepattern/users/SHARED_DATA/BITS_trainingdata_RNAseq/SRR1039509_1.fastq**

Unpaired reads file or first mate for paired reads. A file or zip of files containing reads in FASTA or FASTQ format (can be compressed - ie .gz).

**reads pair 2**

Drag Files Here

Second mate for paired reads. A file or zip of files in FASTA or FASTQ format (can be compressed - ie .gz).
RNA-seq workflow

1. Quality control
2. Mapping
3. Processing mapping results
4. Extract count table
5. Test for differential expression
6. Functional analysis
7. Optional: call variants
Quality control of NGS data

Check quality of the reads: FASTQC
Improve quality of the reads: Trimmomatic, cutadapt... steps depend on FASTQC report
Check quality of cleaned reads: FASTQC

FASTQ with raw reads

@ILMN-GA001_3_208HWAAXX_1_1_110_812
ATACAAGCAAGTATAAGTTCTGTTATATGCGGTCTTT
+ILMN-GA001_3_208HWAAXX_1_1_110_812
hhhYhhh]NYhhhhhhhhhhYllhaZT[hYHNNKFSM
@ILMN-GA001_3_208HWAAXX_1_1_111_879
GGAGGCTGGAGTTGGGGACGTATGCGGCATAGAC
+ILMN-GA001_3_208HWAAXX_1_1_111_879
hSWhRN]hFhLdhVOAIB@NFKD@PAB@NBBBB

FASTQ with clean reads

@ILMN-GA001_3_208HWAAXX_1_1_110_812
ATACAAGCAAGTATAAGTTCTGTTATATG
+ILMN-GA001_3_208HWAAXX_1_1_110_812
hhhYhhh]NYhhhhhhhhhhYllhaZT[hY
@ILMN-GA001_3_208HWAAXX_1_1_112_938
ACATCTCATTACTTGGAAAAGTGAAGTCTAGG
+ILMN-GA001_3_208HWAAXX_1_1_112_938
h?\ah\XTfhPhVhh`YKSLJXPUNJZKLJXV
Quality control of Illumina FASTQ files


Checks: quality of the reads

- quality of the bases
- contamination in your libraries

**GUI** for Windows, Linux and Mac

Part of **VIB GenePattern**

Part of **Galaxy**

Always perform QC on your data!

Many errors in report: not necessarily a sign of bad data

Discuss with experts (provider or bioinformatician)
Using our GenePattern server

We have set up our own server

Training server: https://dev.bits.vib.be:8686

We can choose which tools we build in

Login  User: guestX

Password: training

We have a server on the VSC cluster @ KULeuven

We have a server in Ghent
The GenePattern platform: navigation

Available tools

Tools = modules

Fill in parameters forms and view analysis results
The GenePattern platform: navigation

Available files to be used as input in tools

Load files

Manipulate files
The GenePattern platform: Using shared files

`groomer`

Select URL or File Path as Input

Enter URL:

Or select a file from the file system...

- **SHARED_DATA**
  - BITS_trainingdata_RNAseq
- BITS_trainingdata_variant_analysis
- Broad_trainingdata

Drag Modules Here

Drag Files Here

File in fastQ format that must be "groomed".

*Do not forget to precise the correct input format and the desired output format.*

*Files with sequences in color space do need to start with an adaptor base.*
The GenePattern platform: navigation

Result files

Jobs = History

Manipulate result files

Download

View

Save in GP
FASTQC in Genepattern

**Use file from your computer**  **Use file shared on GenePattern server**

**Drag and drop file from Jobs/Files tab**

Run tool on multiple input files = batch mode

**Input files**

*required field

**input format**

* (autodetect)

**Contaminants file**

No input searches built-in library of known contaminants

**Adapters file**

No input searches built-in library of known adapters
FASTQC in GenePattern

FASTQC generates 2 outputs
.html = report
.zip = raw data of report: figures...
Checking the quality of a FASTQ file in GenePattern

Search for FASTQC tool

Run FASTQC on **SRR074262.fastq** (shared data)

Open resulting HTML report in your browser

Discussion of report:


Quality is not so good: **adapter contamination**

**Quality drops** towards 3' end -> Normal for Illumina data

Run FASTQC on **SRR074285.fastq.gz** (shared data)

Check quality of **SRR1039509_1.fastq** and **SRR1039509_2.fastq** (shared data)

Quality of both files is quite high
some tiles have a few low quality bases
RNA-Seq data

SRR074285

Percent of seqs remaining if deduplicated 37.1%

% Deduplicated sequences
% Total sequences

Sequence Duplication Level
1 2 3 4 5 6 7 8 9 >10 >50 >100 >500 >1k >5k >10k
Strong adapter contamination

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Count</th>
<th>Percentage</th>
<th>Possible Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTGAAA</td>
<td>1127154</td>
<td>13.716943124083405</td>
<td>Illumina Single End Adapter 1 (100% over 33bp)</td>
</tr>
<tr>
<td>GGGGAGGAGAGGCGCATTGGTTGAGGCGCCATTGAG</td>
<td>115445</td>
<td>1.4049122825805602</td>
<td>No Hit</td>
</tr>
<tr>
<td>GAAGAGCTCGTATGCCGTCTTCTGCTTGAAAAAAA</td>
<td>76704</td>
<td>0.933452216038745</td>
<td>Illumina Single End Adapter 1 (100% over 28bp)</td>
</tr>
<tr>
<td>AGGGAGGAGAGGAGCGCATTGGTTGAGGCGCCATTGA</td>
<td>33459</td>
<td>0.4071805627169905</td>
<td>No Hit</td>
</tr>
</tbody>
</table>
SRR1039509_1 and SRR1039509_2

8 tiles in second file have 3-4 low quality bases
SRR1039509_1 and SRR1039509_2

Per base N content

This is reflected in the per base N-content plot.

Low quality bases were called as N.
SRR1039509_1 and SRR1039509_2

RNA-Seq data
Too low to be of any concern

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Count</th>
<th>Percentage</th>
<th>Possible Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>GTCGTTAAGGAAGAAGTGAAGATCCTGGTTGTCGCCGATACATTAA</td>
<td>25479</td>
<td>0.12043558742801647</td>
<td>Illumina Single End PCR Primer 1 (100% over 45bp)</td>
</tr>
<tr>
<td>ACACGTCTAACTCCACTGACCAATCCTGGTTGTCCTCTCTTTGCTT</td>
<td>111922</td>
<td>0.5290392800391875</td>
<td>TruSeq Adapter, Index 4 (100% over 51bp)</td>
</tr>
</tbody>
</table>

SRR1039509_1 and SRR1039509_2
Quality control of RNA-Seq data

Check quality of the reads: FASTQC

Improve quality of the reads: Trimmomatic, cutadapt...
steps depend on FASTQC report

Check quality of cleaned reads: FASTQC

FASTQ with raw reads

FASTQ with clean reads
Trimmomatic improves the quality of the data

http://www.usadellab.org/cms/?page=trimmomatic

Java tool that you run at command line

Part of VIB GenePattern

Part of Galaxy

Used to

Remove contamination

Remove low quality reads

Mappers ignore low quality bases at ends of reads -> no need to trim
RNA-Seq mappers perform soft clipping

Alignment: 12345678901234  5678901234567890
Reference: AGCATGTTAGATAA**GATAGCTGTGCTAGTA
Read002+:        aaaAGATAGTAAGATA

aaa are excluded from alignment

Mismatches at ends of reads are not included in alignment score

-> no need to trim low quality bases

-> no need to clip adapters

  unless there’s substantial adapter dimer contamination
How Trimmomatic detects adapter contamination

**Parameters in GenePattern:**
- **Adapter sequence(s)**
  - adapter clip sequence file
    - Built in or own (fasta)
- How many mismatches allowed in seed
  - **adapter clip seed mismatches:**
    - Short reads (< = 60 bp): 2
    - Long reads (> 60 bp): 3
- Sequencing errors occur at 3' end
- Minimum accuracy of complete alignment
- **adapter clip simple clip threshold:**
  - 7 -> 12 bp perfect matches
  - 15 -> 25 bp perfect matches

**Overrepresented sequences**

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Count</th>
<th>Percentage</th>
<th>Possible Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>GATCAGCAAGCTCGATCAGATCGTTGAAG</td>
<td>1870684</td>
<td>19.446980406066654</td>
<td>Illumina Single End Adapter 1 (100% over 33bp)</td>
</tr>
</tbody>
</table>
The adapter.fasta file

Adapter sequences can be found in FASTQC Overrepresented sequences table

Single end:

>illumina_single_end_adapter_1
GATCGGAAGAGCCTCGTATGCCGTCTTCTGCTTGAAA

Paired end:

>prefixPE/1
ACACTCTTTCCCTACACGACGCTCTTCCGATCT Forward adapter

>prefixPE/2
TGACTGGAGTTCAGACGTGTGCTCTTCCGATCT Reverse adapter
Using Trimmomatic to remove polyA tails

To the fasta file containing the adapter sequence(s) add:

>polyA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
>polyT
TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT
Short adapter contamination is hard to detect

Adapter contamination < seed - allowed mismatches is not detected

Paired-end reads: extra info
1. Both reads align to adapter at same position
2. Both reads are reverse complement

Full alignment is scored: reads + adapters = Palindrome mode

Parameters in GenePattern:
Minimum accuracy of complete alignment adapter clip palindrome clip threshold 30 -> 50 perfect matches

! You have to set this parameter even if you don‘t use palindrome mode!

Keep both reads though they‘re reverse complements adapter clip keep both reads default: yes
Calculating adapter alignment scores

Match   + 0.6
Mismatch – phred quality score/10

Mismatches caused by sequencing errors will have less impact

Matches   Score
12 * 0.6  =  7.2
25 * 0.6  =  15
50 * 0.6  =  30
Reads that are too short after clipping will map everywhere

Reads should be at least 20 bases for mapping

Parameter in GenePattern: Min Read Length
Filtering low quality reads

Minimum quality to keep: typically 20

Extra parameter in GenePattern: \texttt{AVGQUAL}: <min average quality>
Trimmomatic cannot demultiplex your data

For this you will need another tool like cutadapt

Most providers do the demultiplexing for you

How to check if your fastq files have been demultiplexed:

Not demultiplexed

```
@SRR074262.1 HWI-EAS121:2:1:14:558/1
TGTTGAATTGAGAGCCTTGTGTTNAGTAGATAGTGTA
+
BCABCBBCBBBCBBBAABB@B@9!7A?ABB@AB=@B@
```

Demultiplexed

```
CCGTCTGTCTCAGCGTTGATGTGTGTCGAAGACAGTCTCTGCTGCCACCTCTCTGCTGCCAGATGGTGAGGGGGA
AGGGCGGTAGAGCTCTGAT
+
BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
```

barcode
Trimmomatic in GenePattern

- **Adapter Clipping**: Filter/trim adapters
- **Trim Leading**: Trim low quality bases from 5' ends
- **Trim Trailing**: Trim low quality bases from 3' ends
- **Max Info**: Trim low quality bases with calculation of threshold
- **Sliding Window**: Trim low quality bases using a sliding window
- **Min Read Length**: Drop sequences that are too short after trimming
- **Extra Steps**: Trim fixed number of bases from ends: e.g. CROP:30
  Filter low quality reads: e.g. AVGQUAL:20
### Trimmomatic adapter clipping in GenePattern

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Built-in or own adapter?</td>
<td><strong>Select a file</strong> or <strong>Upload your own file</strong></td>
</tr>
<tr>
<td>Adapter sequence in fasta format</td>
<td>A FASTA file containing the adapter sequences, PCR sequences, etc. to be clipped. Files are provided for several Illumina pipelines but you can provide your own; see the documentation for details. Be sure to choose a PE file for paired-end an SE file for single-end data.</td>
</tr>
<tr>
<td>Number of mismatches in seed alignment</td>
<td>Specifies the maximum mismatch count which will still allow a full match to be performed. A very high number is recommended. This parameter is required to enable adapter clipping.</td>
</tr>
<tr>
<td>Minimum alignment score for paired end reads</td>
<td>Specifies how accurate the match between the two 'adapter ligated' reads must be for PE palindrome read alignment. This is the log10 probability against getting a match by random chance; value 30 or more are recommended. This parameter is required to enable adapter clipping.</td>
</tr>
<tr>
<td>Minimum alignment score for single end reads</td>
<td>Specifies how accurate the match between any adapter etc. sequence must be against a read probability against getting a match by random chance; values between 7-15 are recommended. This parameter is required to enable adapter clipping.</td>
</tr>
<tr>
<td>Only for paired end reads</td>
<td>adapter clip min length</td>
</tr>
</tbody>
</table>

Only for paired end reads
Trimmomatic in GenePattern

Run the Trimmomatic tool on SRR074262_groomed.fastq

1. Trim overrepresented adapter sequence
2. Remove all reads < 20 bases after clipping
3. Remove reads with average quality below 20

Use simple clip threshold = 8

Recheck the quality of the trimmed data with FastQC

How many reads remain?
Is the length of the reads still 36 bases?
Have the quality scores of the reads improved after trimming?
Has the adapter disappeared from the list of overrepresented sequences?
Why do we still have a large percentage of duplicates even after adapter removal?
FASTQ files are often very messy

- Wrong ASCII offset -> incorrect interpretation of quality scores
- Syntax errors -> downstream tools cannot interpret the data
- Sequences spread over multiple lines -> downstream tools expect 4 lines / read

Downstream tools do not recognize fastq file / crash => groom it!

**Groomer** cleans up messy FASTQ files so downstream tools can process them

Part of **VIB GenePattern**

Part of **Galaxy**
Grooming fastq files in GenePattern

Encoding of quality scores - see FASTQC report:

**Basic Statistics**

<table>
<thead>
<tr>
<th>Measure</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Filename</td>
<td>SRR074262_fastq</td>
</tr>
<tr>
<td>File type</td>
<td>Conventional base calls</td>
</tr>
<tr>
<td>Encoding</td>
<td>Sanger / Illumina 1.9</td>
</tr>
<tr>
<td>Total Sequences</td>
<td>9619406</td>
</tr>
<tr>
<td>Sequences flagged as poor quality</td>
<td>0</td>
</tr>
<tr>
<td>Sequence length</td>
<td>36</td>
</tr>
</tbody>
</table>
Cleaning messy FASTQ files in GenePattern

Search for the Groomer tool

Run the Groomer tool on `SRR074262.fastq`

You can find this data set in the shared data folder

Go to the Files tab and create a new subfolder with your name
Go to the Jobs tab and copy the groomed file to this subfolder
Trimming adapter, **simple clip threshold = 8**

Trimmomatic clips adapters

Score of 8 requires 14 perfect matches: still large contamination

<table>
<thead>
<tr>
<th>Remaining reads</th>
<th>Read length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Sequences</td>
<td>6418717</td>
</tr>
<tr>
<td>Sequences flagged as poor quality</td>
<td>0</td>
</tr>
<tr>
<td>Sequence length</td>
<td>20-36</td>
</tr>
</tbody>
</table>
Mapping RNA-Seq data

**FASTQ with clean reads**

@ILMN-GA001_3_208HWAAXX_1_1_110_812
ATACAAGCAAGTATAAGTTTGCATATATG
+ILMN-GA001_3_208HWAAXX_1_1_110_812
hhhYhh]NYyyyyyyyyYIlhaZT[hY

@ILMN-GA001_3_208HWAAXX_1_1_112_938
ACATCTCATTACTTGGAAGAGTGAAGTCTTAGG
+ILMN-GA001_3_208HWAAXX_1_1_112_938
h?\ah\XTfhPhVhh`YKSLJXPUNJZKLJXV

**Obtain reference genome/transcriptome**

Map to reference: STAR, HISAT2, Kallisto, Salmon...

@SQ  SN:chr20 LN:64444167
@PG  ID:BWA VN:0.7.15 CL:/srv/dna_tools/bwa/bwa mem –t 4 HiSeq_UCSC_hg19.fa
ILMN-GA001:3:208HWAAXX:1:1:110:812   16 chr20 190938 30 18M2I4M1D3M *
ATACAAGCAAGTATAAGTTTGCATATATG  hhhYhh]NYyyyyyyyyYIlhhf_T`h^  
ILMN-GA001_3_208HWAAXX_1_1_112_938  0 chr20 1493131 10 11M2P7l13M *

**BAM or SAM with alignments**
Mapping on genome
Mapping on transcriptome
Built-in indexed genomes for STAR in VIB GenePattern

Each mapper uses a different index format

*Arabidopsis thaliana*  
TAIR 10, TAIR9

*Danio rerio*  
Ensembl GCRz10, Zv9

*Drosophila melanogaster*  
Ensembl BDGP6

*Homo sapiens*  
Ensembl GRCh37, GRCh38 and UCSChg38, hg19

*Mus musculus*  
Ensembl GRCm38 and UCSC mm10, mm9

You can upload genomes but you will have to index them
Where to obtain the reference genome?


Include chromosomes + unlocalized scaffolds (rRNA)

E.g. Ensembl: genome.dna.primary.assembly.fa.gz
Where to obtain the reference genome?

Human + mouse: [https://www.gencodegenes.org/](https://www.gencodegenes.org/)

### Fasta files

<table>
<thead>
<tr>
<th>Content</th>
<th>Regions</th>
<th>Description</th>
<th>Download</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transcript sequences</td>
<td>CHR</td>
<td>Nucleotide sequences of all transcripts on the reference chromosomes</td>
<td>Fasta</td>
</tr>
<tr>
<td>Protein-coding transcript</td>
<td>CHR</td>
<td>Nucleotide sequences of coding transcripts on the reference chromosomes</td>
<td>Fasta</td>
</tr>
<tr>
<td>sequences</td>
<td></td>
<td>Transcript biotypes: protein_coding, nonsense_mediated_decay, non_stop_decay, IG_<em>.gene, TR_</em>.gene, polymorphic_pseudogene</td>
<td></td>
</tr>
<tr>
<td>Protein-coding transcript</td>
<td>CHR</td>
<td>Amino acid sequences of coding transcript translations on the reference chromosomes</td>
<td>Fasta</td>
</tr>
<tr>
<td>translation sequences</td>
<td></td>
<td>Transcript biotypes: protein_coding, nonsense_mediated_decay, non_stop_decay, IG_<em>.gene, TR_</em>.gene, polymorphic_pseudogene</td>
<td></td>
</tr>
<tr>
<td>Long non-coding RNA transcript</td>
<td>CHR</td>
<td>Nucleotide sequences of long non-coding RNA transcripts on the reference chromosomes</td>
<td>Fasta</td>
</tr>
<tr>
<td>sequences</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Genome sequence (GRCh38.p12)</td>
<td>ALL</td>
<td>Nucleotide sequence of the GRCh38.p12 genome assembly version on all regions, including reference chromosomes, scaffolds, assembly patches and haplotypes The sequence region names are the same as in the GTF/GFF3 files</td>
<td>Fasta</td>
</tr>
<tr>
<td>Genome sequence, primary</td>
<td>PRI</td>
<td>Nucleotide sequence of the GRCh38 primary genome assembly (chromosomes and scaffolds) The sequence region names are the same as in the GTF/GFF3 files</td>
<td>Fasta</td>
</tr>
<tr>
<td>assembly (GRCh38)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Where to obtain the annotation?

The **fasta** file only contains the **sequence**

You also need the **location of exons / CDS**

You can download these as **gtf** files from:

- Ensembl

- Human / mouse: GENCODE basic set (= Ensembl + Havana)

Gtf files of built-in genomes are in GenePattern

For other genomes you have to upload a gft file
Ensembl ftp site:

<table>
<thead>
<tr>
<th>Species</th>
<th>DNA (FASTA)</th>
<th>cDNA (FASTA)</th>
<th>CDS (FASTA)</th>
<th>ncRNA (FASTA)</th>
<th>Protein sequence (FASTA)</th>
<th>Annotated sequence (EMBL)</th>
<th>Annotated sequence (GenBank)</th>
<th>Gene sets</th>
<th>Other annotations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>FASTA</td>
<td>FASTA</td>
<td>FASTA</td>
<td>FASTA</td>
<td>FASTA</td>
<td>EMBL</td>
<td>GenBank</td>
<td>GTF</td>
<td>GFF3</td>
</tr>
<tr>
<td>Homo sapiens</td>
<td>Y</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mouse</td>
<td>FASTA</td>
<td>FASTA</td>
<td>FASTA</td>
<td>FASTA</td>
<td>FASTA</td>
<td>EMBL</td>
<td>GenBank</td>
<td>GTF</td>
<td>GFF3</td>
</tr>
<tr>
<td>Mus musculus</td>
<td>Y</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zebrafish</td>
<td>FASTA</td>
<td>FASTA</td>
<td>FASTA</td>
<td>FASTA</td>
<td>FASTA</td>
<td>EMBL</td>
<td>GenBank</td>
<td>GTF</td>
<td>GFF3</td>
</tr>
<tr>
<td>Danio rerio</td>
<td>Y</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Abingdon island giant tortoise</td>
<td>FASTA</td>
<td>FASTA</td>
<td>FASTA</td>
<td>FASTA</td>
<td>FASTA</td>
<td>EMBL</td>
<td>GenBank</td>
<td>GTF</td>
<td>GFF3</td>
</tr>
<tr>
<td>Chelonoidis abingdonii</td>
<td>Y</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## GeneCode website:

### GTF / GFF3 files

<table>
<thead>
<tr>
<th>Content</th>
<th>Regions</th>
<th>Description</th>
<th>Download</th>
</tr>
</thead>
</table>
| Comprehensive gene annotation| CHR     | - It contains the comprehensive gene annotation on the reference chromosomes only  
                              |          | - This is the **main annotation file** for most users                       | GTF GFF3  |
| Comprehensive gene annotation| ALL     | - It contains the comprehensive gene annotation on the reference chromosomes, scaffolds, assembly patches and alternate loci (haplotypes)  
                              |          | - This is a **superset** of the main annotation file                         | GTF GFF3  |
| Comprehensive gene annotation| PRI     | - It contains the comprehensive gene annotation on the primary assembly (chromosomes and scaffolds) sequence regions  
                              |          | - This is a **superset** of the main annotation file                         | GTF GFF3  |
| Basic gene annotation        | CHR     | - It contains the basic gene annotation on the reference chromosomes only  
                              |          | - This is a **subset** of the corresponding comprehensive annotation, including only those transcripts tagged as 'basic' in every gene | GTF GFF3  |
| Basic gene annotation        | ALL     | - It contains the basic gene annotation on the reference chromosomes, scaffolds, assembly patches and alternate loci (haplotypes)  
                              |          | - This is a **subset** of the corresponding comprehensive annotation, including only those transcripts tagged as 'basic' in every gene | GTF GFF3  |
Where to obtain the transcriptome?

Ensembl ftp site:

- **Human (Homo sapiens)**
  - DNA (FASTA)
  - cDNA (FASTA)
  - CDS (FASTA)
  - ncRNA (FASTA)
  - Protein sequence (FASTA)
  - Annotated sequence (EMBL)
  - Annotated sequence (GenBank)
- **Mouse (Mus musculus)**
  - DNA (FASTA)
  - cDNA (FASTA)
  - CDS (FASTA)
  - ncRNA (FASTA)
  - Protein sequence (FASTA)
  - Annotated sequence (EMBL)
  - Annotated sequence (GenBank)

Genencode website:

**Fasta files**

- **Transcript sequences**
  - Description: Nucleotide sequences of all transcripts on the reference chromosomes
  - Download: Fasta

- **Protein-coding transcript sequences**
  - Description: Nucleotide sequences of coding transcripts on the reference chromosomes
  - Description: Transcript biotypes: protein_coding, nonsense_mediated_decay, non_stop_decay, IG_*.gene, TR_*.gene, polymorphic_pseudogene
  - Download: Fasta

- **Protein-coding transcript translation sequences**
  - Description: Amino acid sequences of coding transcript translations on the reference chromosomes
  - Description: Transcript biotypes: protein_coding, nonsense_mediated_decay, non_stop_decay, IG_*.gene, TR_*.gene, polymorphic_pseudogene
  - Download: Fasta
Mapping NGS data

@ILMN-GA001_3_208HWAAXX_1_1_110_812
ATACAAGCAAGTATAAGTTCGTATATG
+ILMN-GA001_3_208HWAAXX_1_1_110_812
hhhYhh]NYhhhhhhhhYIhhaZT[hY
@ILMN-GA001_3_208HWAAXX_1_1_112_938
ACATCTCATTACTTGGAAAAAGTGAAGTCTAGG
+ILMN-GA001_3_208HWAAXX_1_1_112_938
h?\ah\XTfhPhVhh`YKSLJXPUNJZKLJXV

Obtain reference genome
Map to reference: STAR / HISAT2 / Kallisto / Salmon

@SQ  SN:chr20  LN:64444167
@PG  ID:BWA  VN:0.7.15  CL:/srv/dna_tools/bwa/bwa mem –t 4 HiSeq_UCSC_hg19.fa
ILMN-GA001:3:208HWAAXX:1:1:110:812  16 chr20 190938 30 18M2l4M1D3M *
ATACAAGCAAGTATAAGTTCGTATATG  hhhYhh]NYhhhhhhhhYIhhf_T`h^
ILMN-GA001_3_208HWAAXX_1_1_112_938  0 chr20 1493131 10 11M2P7l13M *

BAM or SAM with alignments
Tools for mapping RNA-Seq data

Alignment based mappers:
STAR: [https://code.google.com/p/rna-star/](https://code.google.com/p/rna-star/)
HISAT2: [http://ccb.jhu.edu/software/hisat/index.shtml](http://ccb.jhu.edu/software/hisat/index.shtml)

Alignment free mappers:
Kallisto: [https://pachterlab.github.io/kallisto](https://pachterlab.github.io/kallisto)
Salmon: [https://github.com/COMBINE-lab/salmon](https://github.com/COMBINE-lab/salmon)

Tools to run at command line

Part of **VIB GenePattern** (not Salmon)
Part of **Galaxy**

References have to be indexed

Every aligner (mapper) has a corresponding indexer
Comparison of performance of alignment based mappers

<table>
<thead>
<tr>
<th>Program</th>
<th>Run time (min)</th>
<th>Memory usage (GB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HISATx1</td>
<td>22.7</td>
<td>4.3</td>
</tr>
<tr>
<td>HISATx2</td>
<td>47.7</td>
<td>4.3</td>
</tr>
<tr>
<td>HISAT</td>
<td>26.7</td>
<td>4.3</td>
</tr>
<tr>
<td>STAR</td>
<td>25</td>
<td>28</td>
</tr>
<tr>
<td>STARx2</td>
<td>50.5</td>
<td>28</td>
</tr>
<tr>
<td>GSNAP</td>
<td>291.9</td>
<td>20.2</td>
</tr>
<tr>
<td>OLeggo</td>
<td>989.5</td>
<td>3.7</td>
</tr>
<tr>
<td>TopHat2</td>
<td>1,170</td>
<td>4.3</td>
</tr>
</tbody>
</table>

Run times and memory usage for HISAT and other spliced aligners to align 109 million 101-bp RNA-seq reads from a lung fibroblast data set. We used three CPU cores to run the programs on a Mac Pro with a 3.7 GHz Quad-Core Intel Xeon E5 processor and 64 GB of RAM.

Kim et al., Nature Methods, 2015
STAR mapping approach

Parameters in GenePattern:

- star index
- max number mismatches: 10 per read / pair
- better: 5% of read length
- depends on quality of reads
- max fraction mismatches: 0.3

chop reads into pieces (=seeds)
map seeds to genome
STAR mapping approach

min overhang annotated read: 3
min overhang not annotated read: 5
map only reported junctions

minimum length of the anchors
minimum overhang annotated read: 3
minimum overhang not annotated read: 5
map only reported junctions

accuracy of reported novel junctions

Anchor 1
Anchor 2

Intron?
Splice junctions

Two-pass mode

repeat mapping with extended gtf file
add

gtf file
novel junctions

Intron?
Fortunately splice signals are strongly conserved
Mapping to transcriptome: single end versus paired end

Single-end

Short reads will map to multiple locations
Mapper takes the alignment with the highest score
> 1 alignment with highest score => mapper makes a random pick

Paired-end

Two reads of a pair align to opposite strands
Distance between reads should fall in fragment length range that was selected
Mapping to transcriptome

Max number of locations a read may map to

Expected distance between reads of a pair

RNA Fragment size

Read_1

Inner distance or insert size

Read_2

Parameters in GenePattern:

max multimapping: 10

> X alignments => unmapped

mates max gap: 500000

(determined by data)

Ideally: max fragment size
Mapping to genome: single-end versus paired-end

Single-end

Short reads will map to multiple locations
Mapper takes the alignment with the highest score
>1 alignment with highest score => mapper makes a random pick

Paired-end

Two reads of a pair align to opposite strands
There can be an intron between the reads increasing the mapping distance
Mapping reads to the genome

Expected length of introns

Parameters in GenePattern:
- min intron length
- max intron length
- gap = intron if gap length < max intron length
- mate inner distance = max intron length
  - default setting: 500000bp
  - ok for vertebrates
  - not ok for non-vertebrates!

Parameters for soft clipping

Alignment: 12345678901234 5678901234567890
Reference: AGCATGTTAGATAA**GATAGCTGTGCTAGTA
Read002+: aaaAGATAA*GGATA

aaa are excluded from alignment: no need to trim low quality bases before mapping

These mismatches are not included in the alignment score

Parameters in GenePattern:

Align reads end to end (no: with soft-clipping)
Output of STAR

STAR.Aligned.out.sam  all alignments
                       multimappers get printed at each location
STAR.SJ.out.tab       splice junctions
STAR.Log.final.out     mapping statistics
STAR.Log.progress.out  mapping statistics
STAR.log.out           translation of GP settings to command executed by STAR

Parameters in GenePattern:

output format (SAM)
BAM / BAM SortedByCoordinate
output unmapped reads (No)
The splice junctions file

<table>
<thead>
<tr>
<th>chrom</th>
<th>start</th>
<th>end</th>
<th>intron</th>
<th>strand</th>
<th>splice motif</th>
<th>novel?</th>
<th>#reads</th>
<th>#multimappers</th>
<th>max overlap</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4606</td>
<td>4705</td>
<td>1</td>
<td>1</td>
<td>1 = GT/AG</td>
<td>0 = novel</td>
<td>10</td>
<td>0</td>
<td>17</td>
</tr>
<tr>
<td>1</td>
<td>5327</td>
<td>5438</td>
<td>1</td>
<td>1</td>
<td>1 = GT/AG</td>
<td>0 = novel</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>7070</td>
<td>7156</td>
<td>2</td>
<td>2</td>
<td>2 = CT/AC</td>
<td>1 = known</td>
<td>1</td>
<td>1</td>
<td>13</td>
</tr>
<tr>
<td>1</td>
<td>8297</td>
<td>8416</td>
<td>2</td>
<td>2</td>
<td>2 = CT/AC</td>
<td>1 = known</td>
<td>2</td>
<td>1</td>
<td>17</td>
</tr>
<tr>
<td>1</td>
<td>8326</td>
<td>8416</td>
<td>2</td>
<td>2</td>
<td>2 = CT/AC</td>
<td>1 = known</td>
<td>2</td>
<td>1</td>
<td>17</td>
</tr>
<tr>
<td>1</td>
<td>8465</td>
<td>8593</td>
<td>2</td>
<td>2</td>
<td>2 = CT/AC</td>
<td>1 = known</td>
<td>2</td>
<td>2</td>
<td>13</td>
</tr>
<tr>
<td>1</td>
<td>24963</td>
<td>25040</td>
<td>1</td>
<td>1</td>
<td>1 = GT/AG</td>
<td>0 = novel</td>
<td>1</td>
<td>1</td>
<td>14</td>
</tr>
<tr>
<td>1</td>
<td>26453</td>
<td>26542</td>
<td>1</td>
<td>1</td>
<td>1 = GT/AG</td>
<td>0 = novel</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>27013</td>
<td>27098</td>
<td>1</td>
<td>1</td>
<td>1 = GT/AG</td>
<td>0 = novel</td>
<td>2</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>1</td>
<td>29081</td>
<td>29159</td>
<td>1</td>
<td>1</td>
<td>1 = GT/AG</td>
<td>0 = novel</td>
<td>1</td>
<td>1</td>
<td>12</td>
</tr>
</tbody>
</table>

1 = +
2 = -

1 = GT/AG
2 = CT/AC
3 = GC/AG
4 = CT/GC
5 = AT/AC
6 = GT/AT
0 = non-canonical
All alignment based mappers perform ± equally well

You have to run STAR in two-pass mode!

Kim et al., Nature Methods, 2015
STAR in GenePattern

Search for the STAR aligner tool and run it on the trimmed and groomed file
1. Run STAR with default parameters (single pass) on the trimmed file
2. Run STAR setting Maximum intron length for Arabidopsis: 3000 (single pass)
3. 2 + maximum fraction of mismatches: 0.05 (single pass)
4. 3 + two-pass mode, output = bam sorted by coordinates

Store the resulting bam file of 4 in your uploads folder

View the **Log.final.out** file in your browser

Mapping rate? Multimappers? Unmapped reads? Mismatch rate per base?
2. Why less multimappers?
3. Why more unmapped reads?

View **Log.out** in your browser
2. Less junctions to map to -> decrease reference sequence space
3. Discard alignments with many mismatches
4. Second pass: few novel junctions found?

<table>
<thead>
<tr>
<th></th>
<th>Default</th>
<th>Intron size</th>
<th>Intron size + Mismatch fraction</th>
<th>Two-pass</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uniquely mapped</td>
<td>86.94%</td>
<td>87.11%</td>
<td>86.84%</td>
<td>86.84%</td>
</tr>
<tr>
<td>Multimappers</td>
<td>7.19%</td>
<td>6.97%</td>
<td>6.84%</td>
<td>6.84%</td>
</tr>
<tr>
<td>Unmapped</td>
<td>5.86%</td>
<td>5.92%</td>
<td>6.32%</td>
<td>6.32%</td>
</tr>
<tr>
<td>Mismatch rate</td>
<td>0.43%</td>
<td>0.44%</td>
<td>0.35%</td>
<td>0.35%</td>
</tr>
</tbody>
</table>
Kallisto mapping approach

Transcriptome sequence

Parameters in GenePattern:
Kallisto index

Strand specificity: none

Correct bias: no
Take into account differences in GC content between fragments when estimating counts?

Pseudo alignment: transcriptome
Reference?

http://tinyheero.github.io/2015/09/02/pseudoalignments-kallisto.html
Kallisto fragment length distribution parameter

Parameters in GenePattern:

**Fragment length distribution:**

SE: you have to provide mean + SD

Impossible to know for public data unless published

PE: calculated based on data

Outcome of Kallisto is very sensitive to this parameter

=> Unless you know distribution, not really suited for SE reads
Processing mapping results

Quality control: RNA-Seq: RSeQC
Index, sort: Picard, samtools
Visualize: IGV

BAM or SAM with alignments

```
@SQ SN:chr20 LN:64444167
@PG ID:BWA VN:0.7.15 CL:/srv/dna_tools/bwa/bwa mem -t 4 HiSeq_UCSC_hg19.fa
ILMN-GA001:3:208HWAAXX:1:1:110:812 16 chr20 190938 30 18M2I4M1D3M *
ATACAAGCAAGTATAAGTTCGTATATG hhhYhh]NYhhhhhhhhYlhhf_T`h^
ILMN-GA001_3_208HWAAXX_1_1_112_938 0 chr20 1493131 10 11M2P7l13M *
```

Sorted BAM + BAI with alignments

```
@SQ SN:chr20 LN:64444167
@PG ID:BWA VN:0.7.15 CL:/srv/dna_tools/bwa/bwa mem -t 4 HiSeq_UCSC_hg19.fa
ILMN-GA001:3:208HWAAXX:1:1:110:812 16 chr20 190938 30 18M2I4M1D3M *
ATACAAGCAAGTATAAGTTCGTATATG hhhYhh]NYhhhhhhhhYlhhf_T`h^
ILMN-GA001_3_208HWAAXX_1_3_88_990 16 chr20 191056 34 19M2l16M *
```
Tools for thoroughly checking the quality of the mapping

Qualimap: http://qualimap.bioinfo.cipf.es/

**GUI** on Windows / Mac / Linux (~ FASTQC)

**General** bam analysis and **RNASEq specific** bam analysis

Needs **gtf** files for RNASEqQC: built-in in GenePattern

fetch from Ensembl/UCSC ftp site:

http://grch37.ensembl.org/info/data/ftp

http://ensemblgenomes.org/info/access/ftp
Tools for thoroughly checking the quality of the mapping


**Specific for RNASEq** bam analysis

Set of Python scripts to run at command line

Part of **VIB GenePattern**

Part of **Galaxy**

Needs *bed* files: **VIB GenePattern** contains **gtf2bed** tool for conversion
RSeQC on GenePattern/Galaxy

Statistics of the mapping: bam_stat

Distribution of mapped reads over features (CDS, UTR, intron...): read_distribution

Number of duplicated reads (same sequence/map to same location): read_duplication

Is sequencing deep enough for alternative splicing analyses?: junction_saturation

Do you find novel splice junctions?: junction_annotation

Are reads equally distributed over full length of transcripts?: geneBody_coverage

Distribution of detected inner distances (only for paired end reads): inner_distance

Parameters

<table>
<thead>
<tr>
<th>Parameters</th>
<th>input file</th>
<th>bam file</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BED file</td>
<td>bed file with annotation</td>
</tr>
<tr>
<td>output prefix</td>
<td></td>
<td>name output file</td>
</tr>
<tr>
<td>minimum map quality</td>
<td></td>
<td>to remove multimappers from analysis</td>
</tr>
<tr>
<td>default (30) is ok for this</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Alignment section:
Read001   163   chr20   7    20    8M2I4M1D3M ...
QNAME    FLAG   RNAME   POS   MAPQ   CIGAR

QNAME       Name of the read: same as in FASTQ file
FLAG        Bitwise flag: gives details of mapping e.g. strand...
RNAME       Name of the reference sequence
POS         Position of start alignment on the reference sequence
MAPQ        MAPping Quality: same scale as quality scores of bases
unique mapper: 255
read that maps twice: 3
read that maps > twice: 1
CIGAR       CIGAR string: represents actual alignment
...

RSeQC in GenePattern


Run the following tools on the SRR074262 bam file:

**bam_stat**: How many secondary alignments?

How many alignments with mapping quality below cutoff (30)?

**read_distribution**: How many reads map in 5’ and 3’UTRs?

**read_duplication**: Open the plot in pdf format

What percentage of reads is not duplicated? Why difference between blue and red?

**junction_saturation**: Open the plot in pdf format

Was coverage high enough in this experiment?
Multimapper: complete read maps to multiple locations
best alignment is chosen as primary (random if more than 1)
other alignments are secondary
**SRR074262 read duplication**

**sequence** based: reads with identical sequence = duplicates

**mapping** based: reads that map to same location = duplicates

22% of reads occur once
4% of reads occur twice
...

Mapping: 2 mismatches allowed
Sequence: no mismatches allowed

=> Slightly different numbers
SRR074262 clipping profile

Some soft clipping at ends of reads
SRR074262 read distribution

distribution of mapped reads over genome features

e.g. CDS exons, 5’UTR exons, 3’UTR exons, introns, intergenic regions

<table>
<thead>
<tr>
<th>Group</th>
<th>Total bases</th>
<th>Tag_count</th>
<th>Tags/Kb</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDS_Exons</td>
<td>33775031</td>
<td>5528803</td>
<td>163.69</td>
</tr>
<tr>
<td>5’UTR_Exons</td>
<td>205309627</td>
<td>2758</td>
<td>0.01</td>
</tr>
<tr>
<td>3’UTR_Exons</td>
<td>205278010</td>
<td>15125</td>
<td>0.07</td>
</tr>
<tr>
<td>Introns</td>
<td>22839</td>
<td>67</td>
<td>2.93</td>
</tr>
<tr>
<td>TSS_up_1kb</td>
<td>20458</td>
<td>12</td>
<td>0.59</td>
</tr>
<tr>
<td>TSS_up_5kb</td>
<td>43381</td>
<td>18</td>
<td>0.41</td>
</tr>
<tr>
<td>TSS_up_10kb</td>
<td>52970</td>
<td>18</td>
<td>0.34</td>
</tr>
<tr>
<td>TES_down_1kb</td>
<td>18702</td>
<td>1</td>
<td>0.05</td>
</tr>
<tr>
<td>TES_down_5kb</td>
<td>45282</td>
<td>1</td>
<td>0.02</td>
</tr>
<tr>
<td>TES_down_10kb</td>
<td>60933</td>
<td>2</td>
<td>0.03</td>
</tr>
</tbody>
</table>

TSS = transcription start site, TES = transcription end site
Is coverage high enough to perform splice variant analysis?

Make subsamples of bam file < 5%, 10%... of the alignments

Count # annotated splice sites in each subsample

If data is saturated you should go to a plateau

Still increasing: more reads might detect more known splice sites
SRR074262 junction annotation

Divides splice junctions into 3 categories:

**known**: 5' and 3' splice site are annotated

**partial novel**: 5' or 3' splice site is annotated

**complete novel**: neither 5' nor 3' splice site is annotated
SRR074262: genebody coverage

Equal distribution of the reads over the full length of the transcripts: OK
Other plots generated by RSeQC: Inner distance

Only valid for **paired-end data**

low quality, duplicates, multimappers are excluded from calculations

read2_start - read1_end (mapped on same exon)

read2_start - read1_end – intron size (mapped on different exons)

insert size = fragment length – 2*read length

![Graph showing Inner distance distribution](image)
BAM files have to be processed to improve accessibility

STAR produces a bam file with alignments sorted according to genomic location
For visualization and downstream processing you also need .bai index file

PICARD:
 BuildBamIndex
Indexing alignments in GenePattern

Search for the BuildBamIndex tool
Run BuildBamIndex on the bam file
Download the bam and bai file to your computer
Counting expression levels = number of mapped reads

Gene-level counts, often obtained by genome alignment + overlap counting

Splice variants are ignored
Counting expression levels = number of mapped reads

**Exon-level** counts, often obtained by genome alignment + overlap counting

Splice variants are **not** ignored
Counting how many reads map to a certain gene

**STAR** can count reads after mapping

**HTSeq-count** after mappers that cannot count themselves

Input: mapped reads + annotation

Output: count of reads per gene

**Rsubread::featureCounts** in R (only Mac/Linux)

**GenomicAlignments::summarizeOverlaps** in R
Counting how many reads map to a certain transcript

**STAR**

**HTSeq-count**

### Parameters in GenePattern:
- Quantify genes (no)/yes
- Input format: bam or sam?
- Strandedness: Count on both strands or not?
- Min qual: Which reads to count? Minimum quality of alignment

<table>
<thead>
<tr>
<th>Output of HTSeq-count</th>
<th>Output of STAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>AT1G01010</td>
<td>AT1G01010</td>
</tr>
<tr>
<td>AT1G01020</td>
<td>AT1G01020</td>
</tr>
<tr>
<td>AT1G01030</td>
<td>AT1G01030</td>
</tr>
<tr>
<td>AT1G01040</td>
<td>AT1G01040</td>
</tr>
<tr>
<td>AT1G01046</td>
<td>AT1G01046</td>
</tr>
<tr>
<td>AT1G01050</td>
<td>AT1G01050</td>
</tr>
<tr>
<td>AT1G01060</td>
<td>AT1G01060</td>
</tr>
<tr>
<td>AT1G01070</td>
<td>AT1G01070</td>
</tr>
<tr>
<td>AT1G01073</td>
<td>AT1G01073</td>
</tr>
<tr>
<td>AT1G01080</td>
<td>AT1G01080</td>
</tr>
<tr>
<td>88</td>
<td>92</td>
</tr>
<tr>
<td>89</td>
<td>95</td>
</tr>
<tr>
<td>32</td>
<td>32</td>
</tr>
<tr>
<td>134</td>
<td>144</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>99</td>
<td>101</td>
</tr>
<tr>
<td>9</td>
<td>8</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>289</td>
<td>306</td>
</tr>
</tbody>
</table>
Stranded RNA sequencing

In standard protocols we don’t know from which strand a read comes.
In stranded RNASeq you do know.
Lowers number of multimappers.

Comparison of strand-specific RNA-Seq protocols
http://www.nature.com/nmeth/journal/v7/n9/full/nmeth.1491.html
Stranded RNA sequencing

e.g. ligate blocked adapter to 3′ end of transcripts
transcripts from + strand
transcripts from - strand

In stranded RNASeq reads are clearly stratified between the two strands

image from GATC Biotech
**Mode: When is read counted for a gene?**

Union is best choice.

<table>
<thead>
<tr>
<th></th>
<th>union</th>
<th>intersection.strict</th>
<th>intersection.nonempty</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image" alt="Diagram" /></td>
<td>gene_A</td>
<td>gene_A</td>
<td>gene_A</td>
</tr>
<tr>
<td><img src="image" alt="Diagram" /></td>
<td>gene_A</td>
<td>no_feature</td>
<td>gene_A</td>
</tr>
<tr>
<td><img src="image" alt="Diagram" /></td>
<td>gene_A</td>
<td>no_feature</td>
<td>gene_A</td>
</tr>
<tr>
<td><img src="image" alt="Diagram" /></td>
<td>gene_A</td>
<td>gene_A</td>
<td>gene_A</td>
</tr>
<tr>
<td><img src="image" alt="Diagram" /></td>
<td>gene_A</td>
<td>gene_A</td>
<td>gene_A</td>
</tr>
<tr>
<td><img src="image" alt="Diagram" /></td>
<td>ambiguous</td>
<td>gene_A</td>
<td>gene_A</td>
</tr>
<tr>
<td><img src="image" alt="Diagram" /></td>
<td>ambiguous</td>
<td>ambiguous</td>
<td>ambiguous</td>
</tr>
<tr>
<td><img src="image" alt="Diagram" /></td>
<td>ambiguous</td>
<td>ambiguous</td>
<td>ambiguous</td>
</tr>
</tbody>
</table>
Counting approaches

**STAR** = htseq-count with default parameters

**htseq-count vs featureCounts**

SE data: featureCounts and htseq-count are almost equivalent

PE data: featureCounts is better for organisms with many overlapping genes

htseq-count: ambiguous pair

featureCounts: gene B
What about multimappers?

Gene A  

Gene B  

Gene A  

Gene B  

<table>
<thead>
<tr>
<th>Gene A</th>
<th>Gene B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treat</td>
<td>Control</td>
</tr>
<tr>
<td>Real counts</td>
<td>10</td>
</tr>
<tr>
<td>--nonunique none</td>
<td>5</td>
</tr>
<tr>
<td>--nonunique all</td>
<td>13</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Treat</th>
<th>Control</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real counts</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>--nonunique none</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>--nonunique all</td>
<td>11</td>
<td>7</td>
</tr>
</tbody>
</table>

Same principle for secondary and supplementary alignments: you should not count them.
<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>count nonunique*</td>
<td>no</td>
<td>Whether to count reads that are not uniquely aligned or are ambiguous.</td>
</tr>
<tr>
<td>count secondary*</td>
<td>no</td>
<td>Whether to count secondary alignments (which are marked in the SAM).</td>
</tr>
<tr>
<td>count supplementary*</td>
<td>no</td>
<td>Whether to count supplementary alignments (which are marked in the SAM).</td>
</tr>
<tr>
<td>feature type*</td>
<td>exon</td>
<td>Name in the 3rd column of the GTF/GFF input file that is used to identify the features that must be counted.</td>
</tr>
<tr>
<td>id type*</td>
<td>gene_id</td>
<td>GTF/GFF attribute used to group features.</td>
</tr>
</tbody>
</table>

**Count multimappers?**

**Count secondary alignments?**

**Count chimeric alignments?**

**How is gene defined?**

**secondary**

**supplementary (read maps over inversion)**
Counting reads in GenePattern


Run HTSeq Count on the sorted bam file in union mode
Minimum quality of mapping should be 10
Open counts file in your browser
Scroll to the bottom of the page
How many reads could not be assigned to a gene?
How many reads were assigned to overlapping genes?
How many multimappers were there?
Why are multimappers discarded?
Save the counts file to your computer
Output bamstat:

Non primary hits: 551833
Unmapped reads: 0
mapq < mapq cut (non-unique): 437073

Alignments of multimappers that were not chosen as primary Multimappers

Output htseq-count

<table>
<thead>
<tr>
<th>Category</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>__no_feature</td>
<td>174140</td>
</tr>
<tr>
<td>__ambiguous</td>
<td>281684</td>
</tr>
<tr>
<td>__too_low_aQual</td>
<td>0</td>
</tr>
<tr>
<td>__not_aligned</td>
<td>0</td>
</tr>
<tr>
<td>__alignment_not_unique</td>
<td>437073</td>
</tr>
</tbody>
</table>

Not mapped to an exon
Mapped on overlapping genes

If a read maps to two genes htseq-count cannot decide which gene it originates from. Discarding does not affect expression ratio. Including in both genes would...
Counting with STAR

Run STAR with the following parameters:
Maximum intron length: 3000
Maximum fraction of mismatches: 0.05
Output bam file sorted by coordinate
2-pass mode
Count reads
Open counts file
Compare with htseq-count results
Compare statistics with those of bamstat and htseq-count

To do this on the human paired end data what would you change?
## STAR count files

<table>
<thead>
<tr>
<th></th>
<th>both strands</th>
<th>fw</th>
<th>rev</th>
</tr>
</thead>
<tbody>
<tr>
<td>N_unmapped</td>
<td>446936</td>
<td>446936</td>
<td>446936</td>
</tr>
<tr>
<td>N_multimapping</td>
<td>517090</td>
<td>517090</td>
<td>517090</td>
</tr>
<tr>
<td>N_noFeature</td>
<td>190313</td>
<td>4776188</td>
<td>1025292</td>
</tr>
<tr>
<td>N_ambiguous</td>
<td>295211</td>
<td>3411</td>
<td>5061</td>
</tr>
<tr>
<td>AT1G01010</td>
<td>92</td>
<td>11</td>
<td>81</td>
</tr>
<tr>
<td>AT1G01020</td>
<td>94</td>
<td>16</td>
<td>78</td>
</tr>
<tr>
<td>AT1G03987</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>AT1G01030</td>
<td>32</td>
<td>4</td>
<td>28</td>
</tr>
<tr>
<td>AT1G01040</td>
<td>108</td>
<td>36</td>
<td>108</td>
</tr>
<tr>
<td>AT1G03993</td>
<td>0</td>
<td>31</td>
<td>5</td>
</tr>
<tr>
<td>AT1G01046</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>AT1G01050</td>
<td>101</td>
<td>10</td>
<td>91</td>
</tr>
<tr>
<td>AT1G03997</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>AT1G01060</td>
<td>9</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>AT1G01070</td>
<td>4</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>AT1G04003</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>AT1G01080</td>
<td>306</td>
<td>82</td>
<td>224</td>
</tr>
<tr>
<td>AT1G01090</td>
<td>870</td>
<td>101</td>
<td>769</td>
</tr>
<tr>
<td>AT1G01100</td>
<td>487</td>
<td>63</td>
<td>424</td>
</tr>
</tbody>
</table>
Counting expression levels

Equivalence class counts, often obtained by “alignment-free” estimation methods

- **Salmon** (Patro et al, Nat Methods 2017)
- **kallisto** (Bray et al, Nat Biotechnol 2016)

Splice variants are **not** ignored

Kallisto and Salmon automatically count reads after mapping
Equivalence class counts to transcript level counts

Transcript-level counts, often obtained by “alignment-free” estimation methods

Splice variants are not ignored

8.53
11.26
10.21
Alignment free mappers: the good and the bad....

+
Counts per transcript – splice variants
Very fast

-
Mapping to transcriptome => you will not find novel genes!
Not suited for SE reads unless you know fragment length distribution
No real alignments => no visualization in IGV
Output of Kallisto and Salmon

### Kallisto

<table>
<thead>
<tr>
<th>target_id</th>
<th>length</th>
<th>eff_length</th>
<th>est_counts</th>
<th>tpm</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENST00000406070</td>
<td>2025</td>
<td>1874.91</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ENST00000446844</td>
<td>2227</td>
<td>2076.91</td>
<td>3.37465</td>
<td>0.129755</td>
</tr>
<tr>
<td>ENST00000599620</td>
<td>686</td>
<td>535.97</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ENST00000471557</td>
<td>505</td>
<td>355.404</td>
<td>2.84168</td>
<td>0.638509</td>
</tr>
<tr>
<td>ENST00000338761</td>
<td>1456</td>
<td>1305.91</td>
<td>1.3122e-05</td>
<td>8.02414e-07</td>
</tr>
<tr>
<td>ENST00000417509</td>
<td>1444</td>
<td>1293.91</td>
<td>5.15988</td>
<td>3.018455</td>
</tr>
<tr>
<td>ENST00000484946</td>
<td>610</td>
<td>460.029</td>
<td>17.4159</td>
<td>3.02326</td>
</tr>
<tr>
<td>ENST00000490656</td>
<td>660</td>
<td>509.97</td>
<td>7.51996</td>
<td>1.17756</td>
</tr>
<tr>
<td>ENST00000439537</td>
<td>1161</td>
<td>1010.91</td>
<td>14.432</td>
<td>1.14006</td>
</tr>
<tr>
<td>ENST00000493251</td>
<td>641</td>
<td>491.006</td>
<td>2.63203</td>
<td>0.428073</td>
</tr>
<tr>
<td>ENST00000460127</td>
<td>408</td>
<td>259.526</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

### Salmon

<table>
<thead>
<tr>
<th>Name</th>
<th>Length</th>
<th>EffectiveLength</th>
<th>TPM</th>
<th>NumReads</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENST00000406070</td>
<td>2025</td>
<td>1869.81</td>
<td>0.137334</td>
<td>3.71695</td>
</tr>
<tr>
<td>ENST00000446844</td>
<td>2227</td>
<td>2071.81</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ENST00000599620</td>
<td>686</td>
<td>530.936</td>
<td>0.731211</td>
<td>3.3457</td>
</tr>
<tr>
<td>ENST00000471557</td>
<td>505</td>
<td>350.256</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ENST00000338761</td>
<td>1456</td>
<td>1300.81</td>
<td>7.58582e-08</td>
<td>1.27717e-06</td>
</tr>
<tr>
<td>ENST00000484946</td>
<td>610</td>
<td>455.039</td>
<td>2.87905</td>
<td>17.1142</td>
</tr>
<tr>
<td>ENST00000490656</td>
<td>660</td>
<td>504.969</td>
<td>1.46703</td>
<td>9.67744</td>
</tr>
<tr>
<td>ENST00000439537</td>
<td>1161</td>
<td>1005.81</td>
<td>1.47611</td>
<td>19.3952</td>
</tr>
<tr>
<td>ENST00000493251</td>
<td>641</td>
<td>485.994</td>
<td>0.597774</td>
<td>3.79512</td>
</tr>
<tr>
<td>ENST00000460127</td>
<td>408</td>
<td>253.708</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Alignment based and alignment free mappers give similar counts
Alignment free mappers map slightly more reads
Using pipelines in GenePattern

Find tool to unzip SRR074285.fastq.gz (in shared_data folder)
Save the extracted file to your files tab
Make a pipeline in GenePattern and repeat the analysis on SRR074285.fastq

Demo Guy
Visualization of mapping results: IGV

IGV: https://software.broadinstitute.org/software/igv/home

GUI on Windows / Mac / Linux

Genome is displayed as bottom track
Visualization of mapping results: Loading data into IGV

You can load BAM files provided BAI file is in the same folder

You have to zoom in to see the reads
Visualization of mapping results: IGV

Hover your mouse over reads to get extra info: start alignment, mapping quality...
Visualization of mapping results: IGV

Differences with reference are coloured
Right click tracks to adjust visualization

Easier to compare coverage across genes
Right click tracks to adjust visualization
Visualize the mapping results in IGV


Run IGV on your computer

Select the correct reference and open the sorted SRR074262 bam file

Zoom in on AT1G02930 until you see the nucleotides of the reference

Open the sorted SRR074285 bam file

Hover your mouse over the highest peak for AT1G02930 in the Coverage track

Do you think AT1G02930 is differentially expressed?

Create a Sashimi plot for this region
You can load multiple samples in IGV and compare them.

AT1G02930 is DE.

Peaks look same height in Coverage track.

When you hover over them you see that coverage is different.
Sashimi plot is more informative
Finding differentially expressed genes

Genes whose expression changes between conditions

RNASeq counts

<table>
<thead>
<tr>
<th>ENSG000000000003</th>
<th>SRR1039508</th>
<th>SRR1039509</th>
<th>SRR1039512</th>
<th>SRR1039513</th>
<th>SRR1039516</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>679</td>
<td>448</td>
<td>873</td>
<td>408</td>
<td>1138</td>
</tr>
<tr>
<td>ENSG000000000005</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ENSG000000000419</td>
<td>467</td>
<td>515</td>
<td>621</td>
<td>365</td>
<td>587</td>
</tr>
<tr>
<td>ENSG000000000457</td>
<td>260</td>
<td>211</td>
<td>263</td>
<td>164</td>
<td>245</td>
</tr>
<tr>
<td>ENSG000000000460</td>
<td>60</td>
<td>55</td>
<td>40</td>
<td>35</td>
<td>78</td>
</tr>
<tr>
<td>ENSG000000000938</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

How many reads for gene A in sample 1?
featureCounts in R

```r
> featureCounts(files = bamfiles,
    annot.ext = "my_genes.gtf",
    isGTFAnnotationFile = TRUE,
    GTF.featureType = "exon",
    GTF.attrType = "gene_id",
    useMetaFeatures = TRUE,
    isPairedEnd = TRUE,
    strandSpecific = 0)
```

check your GTF file!
Input for DE analysis

HTSeq count -> folder with count files
FeatureCounts -> count matrix
Kallisto, Salmon -> count matrix
To find DE genes DESeq2 needs metadata

<table>
<thead>
<tr>
<th>Sample names</th>
<th>File containing the counts</th>
<th>Pairing factor</th>
<th>Grouping factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>SRR1039508</td>
<td>SRR1039508_all_counts.txt</td>
<td>N61311</td>
<td>untreated</td>
</tr>
<tr>
<td>SRR1039509</td>
<td>SRR1039509_all_counts.txt</td>
<td>N61311</td>
<td>Dex</td>
</tr>
<tr>
<td>SRR1039512</td>
<td>SRR1039512_all_counts.txt</td>
<td>N052611</td>
<td>untreated</td>
</tr>
<tr>
<td>SRR1039513</td>
<td>SRR1039513_all_counts.txt</td>
<td>N052611</td>
<td>Dex</td>
</tr>
<tr>
<td>SRR1039516</td>
<td>SRR1039516_all_counts.txt</td>
<td>N080611</td>
<td>untreated</td>
</tr>
<tr>
<td>SRR1039517</td>
<td>SRR1039517_all_counts.txt</td>
<td>N080611</td>
<td>Dex</td>
</tr>
<tr>
<td>SRR1039520</td>
<td>SRR1039520_all_counts.txt</td>
<td>N061011</td>
<td>untreated</td>
</tr>
<tr>
<td>SRR1039521</td>
<td>SRR1039521_all_counts.txt</td>
<td>N061011</td>
<td>Dex</td>
</tr>
</tbody>
</table>

Which samples come from same cell line?
Which samples are biological replicates?

Wiki: GSE52778_metadata.txt

Himes et al. 2014
To find DE genes DESeq2 needs design

design = ~ cells + treatment

is modeled by Pairing factor Grouping factor

Which data need to be combined? Counts from same cell line

Which data need to be compared? Dex versus untreated

Variable of interest comes last

Factors have levels: e.g. treatment: dex, alb, control...

By default R chooses reference in alphabetical order -> alb will be chosen as ref

You need to tell R what the reference is:

> dds$treatment <- relevel(dds$treatment, ref="control")

Analyze all groups together, do not split them up in separate analyses!
To find DE genes DESeq2 needs design

More factors?

design = ~ cells + treatment + time

=> Assumption: effect treatment is same on all time points

Interactions?

design: ~ treatment + time + treatment:time

=> Assumption: effect treatment is different on different time points
Simpler solution for interactions

```r
> sampleTable$group <- factor(paste(dds$treatment, dds$time))

combine factors into a single factor with all combinations of original factors

design: ~ group

Use contrasts to see results:

Time point 2 treated compared to time point 1 treated

```r
> results(dds, contrast=c("group","2T","1T"))

Time point 1 treated compared to time point 1 control

```r
> results(dds, contrast=c("group","1T","1C"))
```
The SummarizedExperiment() to store omics data

Gene info:
- rowRanges(se)
- rowData(se)
- subsetByOverlaps(se, roi)
- assays(se)
  - assay(se, n = 2)
  - assay(subsetByOverlaps(se, roi))
  - assay(se[, se$dex == "trt"])
- colData(se)
  - se[, se$dex == "trt"]
- metadata(se)
  - metadata(se)$modelFormula

Description of samples
- Samples (Columns)
- Features (Rows)

The actual data: counts
DESeq2 uses a DESeqDataSet to store RNASeq data

Data from htseq-count:

```r
> folder <- "htseq_counts"
> dds <- DESeqDataSetFromHTSeqCount(sampleTable=sampleTable,
+  directory=folder,
+  design= ~cells+treatment)

> dds
class: DESeqDataSet
dim: 57773 8
metadata(1): version
assays(1): counts
rownames(57773): ENSG000000000003 ENSG000000000005 ...
   ENSG00000273492 ENSG00000273493
rowData names(0): 
colnames(8): SRR1039508 SRR1039509 ... SRR1039520
   SRR1039521
colData names(2): cells treatment
```

Sample names, file names + grouping info
Folder with count files
Statistics info: what do you want to compare?
Access counts via `counts(dds)`
Access grouping info via `colData(dds)`
DESeq2 uses a DESeqDataSet to store RNASeq data

Data from featureCounts:

```r
dds <- DESeqDataSetFromMatrix(countData = cts,
    colData = coldata,
    design = ~ condition)
```

- **Matrix with counts**: cts
- **Sample names, file names + grouping info**: coldata
- **Statistics info: what do you want to compare?**: design

Access counts via `counts(dds)`

Access grouping info via `colData(dds)`
EdgeR uses a list to store RNASeq data

```r
> dge <- DGEList(counts = countdata, 
+     samples = coldata, 
+     genes = genetable)
> dge
An object of class "DGEList"
$counts
      SRR1039508  SRR1039509  SRR1039512  SRR1039513
ENSG000000000003    667     434       862       401
ENSG000000000005       0       0         0         0
ENSG000000000419     430     488       556       334

$samples
        group     lib.size norm.factors  cells treatment
SRR1039508  1  18104038  14432.1458 1   N61311  untreated
SRR1039509  1  16545934  13917.9945 1   N61311        Dex
SRR1039512  1  22630840  19117.8148 1   N052611  untreated

$genes
  gene.id
ENSG000000000003 ENSG000000000003
ENSG000000000005 ENSG000000000005
ENSG000000000419 ENSG000000000419
```

Matrix with counts
Sample names, grouping info
Gene names
Access counts via `dge$counts`
Access sample info via `dge$samples`
Access gene names via `dge$genes`
Counts are a special type of data

Discrete, Positive, Skewed

Variance is not stable => high counts have higher variance

Normal distributions and classic statistics are not valid for RNASeq data
Prefiltering counts

Remove genes with very low counts

Not required but speeds up analysis

```r
> dds <- dds[rowSums(counts(dds)) >= 10, ]
```
How to model RNA-Seq data

RNA-Seq data are modeled using **negative binomial distribution**:  

2 parameters  

- **mean**  
- **dispersion**  

Variance (gene X) = mean + dispersion * mean^2  

Dispersion = biological variation within a group of samples  

=> For each gene you need to calculate mean + dispersion in each group of samples
EdgeR and DESeq2 use a GLM to model the counts

RNA-Seq data are modeled using **Generalised Linear Model:**

1. Distribution = negative binomial distribution
2. Linear predictor
3. Link function = log link: transforms the mean **not the actual counts**

Transformed counts would not follow negative binomial distribution

So you would have to search for the appropriate distribution again

Model used by DESeq2
http://genomebiology.com/2014/15/12/550

Model used by EdgeR
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2796818/
You need to calculate the dispersions

For each gene:

Dispersion is calculated on **counts** of biological replicates
You need many replicates to accurately calculate dispersion.

**Example:**
estimate variance of normally distributed variable

True value = 3
Not enough replicates for accurate calculation of dispersion

Dispersion is very important: variation between groups > dispersion => DE

You don’t have enough replicates but you do have a lot of genes
Dispersion is re-estimated based on information of all genes (= data distribution)

Default dispersion estimation in DESeq2 not ok when you have < 4 replicates
=> use the Loess fit to estimate the dispersions in DESeq2:
> dds <- DESeq(dds, fitType="local")
Not enough replicates for accurate calculation of dispersion

Data cloud = actual data distribution

Fit model to data cloud = theoretical distribution

Genes with similar expression levels should have similar dispersions
Dispersions are shrunken based on info from similar genes

Shrink actual dispersions towards the curve

Detect outlier dispersions and do not shrink them
Library size influences read counts

Sample 1: all genes same expression

Sample 2: 9 genes same expression as in sample 1
  gene 10 10x higher expression level
Library size influences read counts

Sample1: you sequence 1000 reads

Sample2: you sequence 1000 reads

All genes will appear DE!
How to normalize for library size?

Divide counts by total number of mappable reads (Total Count normalization)

Not applied anymore!

In the example you divide all counts by 1000

Sample 1:
All genes: 0.1

Sample 2:
Genes 1-9: 0.05
Gene 10: 0.5

You still think all genes are DE

RPKM is a variation on this that also normalizes for gene length
EdgeR uses trimmed mean normalization for library size

You filter out genes with very high counts

You divide by mean read count for genes with low count number

Sample 1: you divide by 100
Sample 2: you divide by 50

All genes: 1
Genes 1-9: 1
Gene 10: 10

It truly reflects actual expression numbers

### DESeq2 normalization for library size

#### Table

<table>
<thead>
<tr>
<th>Sample1</th>
<th>Sample2</th>
<th>Geomean</th>
<th>Quotient1</th>
<th>Quotient2</th>
<th>Normalized1</th>
<th>Normalized2</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>C2</td>
<td>(\sqrt{C_1 \times C_2})</td>
<td>C1/geomean</td>
<td>C2/geomean</td>
<td>C1/median</td>
<td>C2/median</td>
</tr>
<tr>
<td>100</td>
<td>50</td>
<td>70,71</td>
<td>1,41</td>
<td>0,71</td>
<td>70,71</td>
<td>70,71</td>
</tr>
<tr>
<td>100</td>
<td>50</td>
<td>70,71</td>
<td>1,41</td>
<td>0,71</td>
<td>70,71</td>
<td>70,71</td>
</tr>
<tr>
<td>100</td>
<td>50</td>
<td>70,71</td>
<td>1,41</td>
<td>0,71</td>
<td>70,71</td>
<td>70,71</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>100</td>
<td>50</td>
<td>70,71</td>
<td>1,41</td>
<td>0,71</td>
<td>70,71</td>
<td>70,71</td>
</tr>
<tr>
<td>100</td>
<td>500</td>
<td>223,61</td>
<td>0,45</td>
<td>2,24</td>
<td>70,71</td>
<td>707,11</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1,41</td>
<td>0,71</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Median: Median

1. For each gene: calculate geometric mean of counts
2. For each sample: calculate quotients of counts / geometric mean
3. For each sample: calculate median of quotients
4. For each sample: divide counts by median

DESeq2 uses only 1 gene (median)

It truly reflects actual expression numbers
EdgeR / DESeq2 normalization assume that most genes are not DE

If this is the real biological situation then EdgeR / DESeq2 normalization fail

In most experiments the number of DE genes is low

If there are many DE genes: normalization with good housekeeping genes

use Genevestigator RefGenes tool for HK genes
Library normalization is only used for visualization

EdgeR and DESeq2 model raw counts not normalized counts

Normalized counts do not follow negative binomial distribution

So you would have to search for the appropriate distribution again

Negative binomial distribution

Same data after multiplying by 2
Factors influencing read counts are tackled by normalization

**Biological variation**: dispersions estimated based on all genes

**Total number of reads**: normalization for library size

**Length of gene**: longer transcripts generate more reads

  Not an issue since you compare conditions for each gene separately

  Only an issue when you want to compare genes

**Sequencing biases**: GC content of genes influences number of reads

  Not an issue since you compare conditions for each gene separately

  Only an issue when you want to compare genes

**Expression level** = what you’re interested in: compare expression between conditions
Tests for finding DE genes

**Wald test** default in DESeq2

\[ \frac{lfc}{lfcSE} = 0 \]?

Pairwise comparisons: treated at time 2 vs. control at time 2

Less reliable for small samples

**Likelihood Ratio Test** option in DESeq2, default in EdgeR

full model (multiple factors) = reduced model (less factors) ?

Find genes that are affected by Dex over all cell lines

Find genes that behave different in treated over all time points
DE tests in DESeq2

> dds <- DESeq(dds)

Results are extracted into a results table:

> res <- results(dds)

Two factors, LRT test

> DESeq(dds, test="LRT", full = ~ time + treatment, reduced = ~ time)

Find genes that behave different in treated over all time points

One factor, very few replicates, LRT test

> DESeq(dds, test="LRT", full = ~ treatment, reduced = ~ 1)
Results table in DEseq2

<table>
<thead>
<tr>
<th></th>
<th>baseMean</th>
<th>log2FoldChange</th>
<th>lfcSE</th>
<th>stat</th>
<th>pvalue</th>
<th>padj</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>165995</td>
<td>514.2841</td>
<td>3.321662</td>
<td>0.1307366</td>
<td>25.40728</td>
<td>2.095444e-142</td>
</tr>
<tr>
<td>2</td>
<td>152583</td>
<td>985.5593</td>
<td>4.340812</td>
<td>0.1760858</td>
<td>24.65169</td>
<td>3.529480e-134</td>
</tr>
<tr>
<td>3</td>
<td>120129</td>
<td>3325.4027</td>
<td>2.873150</td>
<td>0.1167734</td>
<td>24.60448</td>
<td>1.131211e-133</td>
</tr>
<tr>
<td>4</td>
<td>101347</td>
<td>13616.9348</td>
<td>3.606558</td>
<td>0.1517550</td>
<td>23.76566</td>
<td>7.569343e-125</td>
</tr>
</tbody>
</table>

baseMean: average of normalized counts, divided by size factors, over all samples

log2FC: fold change estimate – log2!

lfcSE: uncertainty of log2FC

Statistic for Wald test: lfc / lfcSE

Compare statistic to a standard normal distribution -> p-value

You perform multiple tests (one for each gene) -> adjust p-values

by Benjamini and Hochberg's method to control FDR

padj can be NA: genes with zero counts or an extreme outlier count

Always report adjusted p-values!
Results table in DEseq2

This is the result for the last factor (treatment)

Comparison last level – first level (Dex – control)

What if you have multiple levels?

> results(dds, **contrast**=c("treatment","Alb","control")) comparison albuterol - control


http://bioconductor.org/packages/devel/bioc/vignettes/DESeq2/inst/doc/DESeq2.html#contrasts
Results table in DEseq2

What if you have 2 factors?

Pairwise comparisons with Wald test:
> results(dds, contrast=c("group"," N052611Dex"," N052611control"))

<table>
<thead>
<tr>
<th>factor</th>
<th>level</th>
<th>control level</th>
</tr>
</thead>
</table>

Overall comparisons with LRT:
> results(dds, name="treatment_Dex_vs_control")

Comparison over all cell lines
Values for name can be found via
> resultsNames(dds)
Log ratios are only used for visualization

EdgeR and DESeq2 model raw counts not log ratios

Ratios and log ratios do not follow negative binomial distribution

So you would have to search for the appropriate distribution again

<table>
<thead>
<tr>
<th></th>
<th>Sample1</th>
<th>Sample2</th>
<th>lfc</th>
</tr>
</thead>
<tbody>
<tr>
<td>gene1</td>
<td>400</td>
<td>50</td>
<td>3</td>
</tr>
<tr>
<td>gene2</td>
<td>8</td>
<td>1</td>
<td>3</td>
</tr>
</tbody>
</table>

Model ratios => both genes will be called DE

Model raw counts =>

GLM takes into account that gene2 has low mean and gene1 has high mean =>
gene1 will be called DE and gene2 probably not
Calculation of log2 fold changes

Based on normalized counts

Genes with low read counts have unrealistically high log fold changes (lfc)

<table>
<thead>
<tr>
<th>Gene</th>
<th>Sample1</th>
<th>Sample2</th>
<th>lfc</th>
<th>shrunken lfc</th>
</tr>
</thead>
<tbody>
<tr>
<td>gene1</td>
<td>400</td>
<td>50</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>gene2</td>
<td>8</td>
<td>1</td>
<td>3</td>
<td>0.31</td>
</tr>
</tbody>
</table>

Lfc's are shrunken toward 0 - Stronger shrinkage when counts are low

For visualization and ranking genes use shrunken lfc's:

> lfcShrink(dds, coef=5)

Values for coef can be found via:

> resultsNames(dds)

EdgeR also shrinks lfc's but less strong than DESeq2
For visualization normalized counts are transformed

vst: variance stabilizing transformation
rlog: regularized logarithm
⇒ log2 normalized counts
⇒ produces constant variance along the range of mean count values
blind: TRUE do not take into account info provided by design formula
Set to FALSE for downstream analysis especially when you expect a lot of DE genes:
> vsd <- vst(dds, blind=FALSE)
> rld <- rlog(dds, blind=FALSE)
Comparison of DE finding methods

https://academic.oup.com/bib/article/14/6/671/189645

<table>
<thead>
<tr>
<th></th>
<th>RC</th>
<th>TC</th>
<th>RPKM</th>
<th>EdgeR</th>
<th>DESeq</th>
</tr>
</thead>
<tbody>
<tr>
<td>RC</td>
<td>184</td>
<td>175</td>
<td>149</td>
<td>184</td>
<td>184</td>
</tr>
<tr>
<td>TC</td>
<td>548</td>
<td>399</td>
<td>547</td>
<td>543</td>
<td></td>
</tr>
<tr>
<td>RPKM</td>
<td></td>
<td>417</td>
<td>416</td>
<td>413</td>
<td></td>
</tr>
<tr>
<td>EdgeR</td>
<td></td>
<td></td>
<td>1190</td>
<td>1169</td>
<td></td>
</tr>
<tr>
<td>DESeq2</td>
<td></td>
<td></td>
<td>1249</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

# DE genes in common

RC = raw counts
TC = total counts

False positive rates over 10 independent data sets simulated with varying proportions of DE genes (from 0% to 30%)

EdgeR and DESeq2 work well
They work equally well
Characterization of the DE genes

Finding gene names: BioMart

http://www.ensembl.org/biomart/martview/

Access online or you can install the software on your computer

BioMart works in three steps:

1. select the database you want to query

2. set filters: requirements which the genes you want to retrieve need to match

3. set attributes: which data you want to retrieve about these genes

   e.g. sequence, names, domains...
Step 1. Select database and data set to search in

Dataset
[None selected]

- CHOOSE DATASET -
- CHOOSE DATASET -
  Chicken genes (Gallus_gallus-5.0)
  Human genes (GRCh38.p12)
  Mouse genes (GRCm38.p6)
  Rat genes (Rnor_6.0)
  Zebrafish genes (GRCz11)

-----------------------------

Algerian mouse genes (SPRET_EiJ_v1)
Alpaca genes (vicPac1)
Amazon molly genes (Poecilia_formosa-5.1.2)
Angola colobus genes (Cang.pa_1.0)
Anole lizard genes (AnoCar2.0)
Armadillo genes (Dasnov3.0)
Black snub-nosed monkey genes (ASM169854v1)
Bolivian squirrel monkey genes (SaiBol1.0)
Bonobo genes (panpan1.1)
Step 2. Set filters to define the genes you want

Fields you can filter on
- Region: location in the genome
- **Gene**: define a set of genes using IDs, gene type...
- Phenotype: involvement in a disease
- Gene Ontology: functional annotation
- Expression: the organ a gene is expressed in
- Protein domains: presence of a domain
...

Max 500 genes at a time!
Step 3. Set attributes: specify what you want to retrieve
Step 3. Select what you want to retrieve

Please select columns to be included in the output and hit 'Results' when ready

- **Features**
- **Homologs**
- **Structures**
- **Variation**
- **Transcript Event**
- **Sequences**

**GENE:**

<table>
<thead>
<tr>
<th>Ensembl</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Ensembl Gene ID</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ensembl Transcript ID</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ensembl Protein ID</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ensembl Exon ID</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Description</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chromosome Name</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gene Start (bp)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gene End (bp)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Strand</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Band</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Transcript Start (bp)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Transcript End (bp)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Retrieve the results of your search

Mapping Ensembl IDs to gene symbols

Use Biomart to map the Ensembl IDs of the upregulated genes to gene symbols

Database?
Dataset?
Filter?
How many genes?
Attribute?
Download file with gene symbols to your home folder.
Functional characterization of a set of genes

What’s the biology behind a list of genes?

We have a list of almost 400 upregulated genes.

Impossible to evaluate each gene individually.

See what the genes in the list have functionally in common.

E.g., 60% are involved in MAP kinase pathway.

- MAP kinase pathway is affected by dexamethasone treatment.

Gene enrichment analysis

Gene set enrichment analysis.
Functional characterization of a set of genes

Find functions and pathways that are enriched in potential targets

1. Add functional annotations to the genes in the list
2. Define background: full set of all genes measured by the platform
3. Perform test to identify enriched 'functions', 'diseases', 'pathways' ...

Enriched = Over-represented

Occurring more frequently than expected based on background data

I characterize up- and down regulated genes separately
Tools for gene enrichment analysis

ToppGene: https://toppgene.cchmc.org/
Enrichr: http://amp.pharm.mssm.edu/Enrichr/
Webgestalt: http://www.webgestalt.org/
DAVID: http://david.abcc.ncifcrf.gov/
### Comparison of tools for gene enrichment analysis

<table>
<thead>
<tr>
<th></th>
<th>DAVID</th>
<th>EnrichR</th>
<th>ToppGene</th>
<th>Webgestaltt</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Organisms</strong></td>
<td>Many</td>
<td>Human, mouse, rat</td>
<td>Human, mouse, rat</td>
<td>12 organisms</td>
</tr>
<tr>
<td><strong>Visuals</strong></td>
<td>table</td>
<td>table + charts</td>
<td>table</td>
<td>table</td>
</tr>
<tr>
<td><strong>Up-to-date</strong></td>
<td>2016</td>
<td>2017</td>
<td>2018</td>
<td>2017</td>
</tr>
<tr>
<td><strong># resources</strong></td>
<td>Many</td>
<td>Many</td>
<td>Most</td>
<td>Many</td>
</tr>
<tr>
<td><strong>Ease of use</strong></td>
<td>Website, some steps</td>
<td>Website, very easy</td>
<td>Website, very easy</td>
<td>Website, one db at a time</td>
</tr>
</tbody>
</table>
ToppGene: very up-to-date but human (mouse/rat) only

ToppGene Suite

A one-stop portal for gene list enrichment analysis and candidate gene prioritization based on functional annotations and protein interactions network

<table>
<thead>
<tr>
<th>GO Biological Process</th>
<th>Human Phenotype</th>
<th>Mouse Phenotype</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Annotations</strong>:</td>
<td>15,171</td>
<td>8,400</td>
</tr>
<tr>
<td><strong>Genes</strong>:</td>
<td>18,623</td>
<td>4,707</td>
</tr>
</tbody>
</table>

**Updated 11-dec-2017**

<table>
<thead>
<tr>
<th>GO Cellular Component</th>
<th>Human Phenotype</th>
<th>Mouse Phenotype</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Annotations</strong>:</td>
<td>1,801</td>
<td>10,346</td>
</tr>
<tr>
<td><strong>Genes</strong>:</td>
<td>19,061</td>
<td>9,294</td>
</tr>
</tbody>
</table>

**Updated 11-feb-2018**

<table>
<thead>
<tr>
<th>GO Molecular Function</th>
<th>Human Phenotype</th>
<th>Mouse Phenotype</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Annotations</strong>:</td>
<td>4,961</td>
<td>18,661</td>
</tr>
<tr>
<td><strong>Genes</strong>:</td>
<td>18,661</td>
<td></td>
</tr>
</tbody>
</table>

Gene symbols of human, mouse and rat orthologs are the same

So you can use it for mouse too but analysis is based on data from human

Tutorial: https://toppgene.cchmc.org/help/supplimental.jsp#ToppFun
Functional characterization of the upregulated genes

Use ToppGene to find enriched functional categories in the list of upregulated genes

Set p-value cutoff to 0.01

Enriched functional processes?
Enriched domains?
Enriched miRNAs?
Enriched diseases?
### 2: GO: Biological Process

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Source</th>
<th>pValue</th>
<th>FDR B&amp;H</th>
<th>FDR B&amp;Y</th>
<th>Bonferroni</th>
<th>Genes from Input</th>
<th>Genes in Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>GO:0007155 cell adhesion</td>
<td></td>
<td>6.372E-7</td>
<td>1.273E-3</td>
<td>1.144E-2</td>
<td>2.853E-3</td>
<td>58</td>
<td>1530</td>
</tr>
<tr>
<td>2</td>
<td>GO:0022610 biological adhesion</td>
<td></td>
<td>8.208E-7</td>
<td>1.273E-3</td>
<td>1.144E-2</td>
<td>3.675E-3</td>
<td>58</td>
<td>1542</td>
</tr>
<tr>
<td>3</td>
<td>GO:0009725 response to hormone</td>
<td></td>
<td>8.530E-7</td>
<td>1.273E-3</td>
<td>1.144E-2</td>
<td>3.619E-3</td>
<td>44</td>
<td>1033</td>
</tr>
<tr>
<td>4</td>
<td>GO:0030198 extracellular matrix organization</td>
<td></td>
<td>1.833E-6</td>
<td>1.719E-3</td>
<td>1.545E-2</td>
<td>8.207E-3</td>
<td>22</td>
<td>354</td>
</tr>
</tbody>
</table>

### 6: Domain

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Source</th>
<th>pValue</th>
<th>FDR B&amp;H</th>
<th>FDR B&amp;Y</th>
<th>Bonferroni</th>
<th>Genes from Input</th>
<th>Genes in Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>PF00090 TSP 1</td>
<td>Pfam</td>
<td>3.296E-7</td>
<td>2.254E-4</td>
<td>1.846E-3</td>
<td>6.661E-4</td>
<td>10</td>
<td>63</td>
</tr>
<tr>
<td>2</td>
<td>PS50092 TSP1</td>
<td>PROSITE</td>
<td>4.461E-7</td>
<td>2.254E-4</td>
<td>1.846E-3</td>
<td>9.016E-4</td>
<td>10</td>
<td>65</td>
</tr>
<tr>
<td>3</td>
<td>IPR000884 TSP1 rpt</td>
<td>InterPro</td>
<td>4.461E-7</td>
<td>2.254E-4</td>
<td>1.846E-3</td>
<td>9.016E-4</td>
<td>10</td>
<td>65</td>
</tr>
<tr>
<td>4</td>
<td>SM00209 TSP1</td>
<td>SMART</td>
<td>4.461E-7</td>
<td>2.254E-4</td>
<td>1.846E-3</td>
<td>9.016E-4</td>
<td>10</td>
<td>65</td>
</tr>
</tbody>
</table>

### 7: Pathway

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Source</th>
<th>pValue</th>
<th>FDR B&amp;H</th>
<th>FDR B&amp;Y</th>
<th>Bonferroni</th>
<th>Genes from Input</th>
<th>Genes in Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Ensemble of genes encoding core extracellular matrix including ECM glycoproteins, collagens and proteoglycans</td>
<td>MSigDB C2 BIOCARTA (v6.0)</td>
<td>1.884E-8</td>
<td>2.623E-5</td>
<td>2.050E-4</td>
<td>2.623E-5</td>
<td>23</td>
<td>275</td>
</tr>
<tr>
<td>2</td>
<td>Ensemble of genes encoding extracellular matrix and extracellular matrix-associated proteins</td>
<td>MSigDB C2 BIOCARTA (v6.0)</td>
<td>7.988E-7</td>
<td>5.560E-4</td>
<td>4.346E-3</td>
<td>1.112E-3</td>
<td>46</td>
<td>1028</td>
</tr>
<tr>
<td>3</td>
<td>Genes encoding structural ECM glycoproteins</td>
<td>MSigDB C2 BIOCARTA (v6.0)</td>
<td>4.032E-6</td>
<td>1.871E-3</td>
<td>1.462E-2</td>
<td>5.612E-3</td>
<td>16</td>
<td>196</td>
</tr>
</tbody>
</table>

Report adjusted p-values
Do the enrichment results make sense?

1: GO: Molecular Function [Display Chart] 396 input genes in category / 1166 annotations before applied cutoff / 18661

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Source</th>
<th>pValue</th>
<th>FDR B&amp;H</th>
<th>FDR B&amp;Y</th>
<th>Bonferroni</th>
<th>Genes from Input</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0005539</td>
<td>glycosaminoglycan binding</td>
<td></td>
<td>1.081E-6</td>
<td>1.260E-3</td>
<td>9.624E-3</td>
<td>1.260E-3</td>
<td>18</td>
</tr>
</tbody>
</table>

Dexamethasone regulation of glycosaminoglycan synthesis in...
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC425408/ ▼
by TJ Smith - 1984 - Cited by 107 - Related articles

2: GO: Biological Process [Display Chart] 403 input genes in category / 4807 annotations before applied cutoff / 18623 genes in category

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Source</th>
<th>pValue</th>
<th>FDR B&amp;H</th>
<th>FDR B&amp;Y</th>
<th>Bonferroni</th>
<th>Genes from Input</th>
</tr>
</thead>
</table>

3: GO: Cellular Component [Display Chart] 405 input genes in category / 464 annotations before applied cutoff / 19061 genes in category

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Source</th>
<th>pValue</th>
<th>FDR B&amp;H</th>
<th>FDR B&amp;Y</th>
<th>Bonferroni</th>
<th>Genes from Input</th>
<th>Genes in Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0031012</td>
<td>extracellular matrix</td>
<td>4.842E-10</td>
<td>2.247E-7</td>
<td>1.509E-6</td>
<td>2.247E-7</td>
<td>33</td>
<td>444</td>
<td></td>
</tr>
</tbody>
</table>


Dexamethasone enhances cell resistance to chemotherapy by increasing adhesion to extracellular matrix in human ovarian cancer cells.

6: Domain [Display Chart] 404 input genes in category / 2215 annotations before applied cutoff / 18735 genes in category

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Source</th>
<th>pValue</th>
<th>FDR B&amp;H</th>
<th>FDR B&amp;Y</th>
<th>Bonferroni</th>
<th>Genes from Input</th>
<th>Genes in Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>PF00090</td>
<td>TSP 1</td>
<td>Pfam</td>
<td>8.974E-7</td>
<td>6.698E-4</td>
<td>5.547E-3</td>
<td>1.988E-3</td>
<td>10</td>
<td>63</td>
</tr>
</tbody>
</table>

Invoking the Power of Thrombospondins: Regulation of ... - NCBI - NIH
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4143502/ ▼
by O Stenina-Adognravi - 2014 - Cited by 24 - Related articles
Feb 25, 2014 - All TSPs share a high degree of homology at the protein level, especially in their "signature domain" - the C-terminal half of the human protein that ..... Increased levels of TSP-1 mRNA were detected upon stimulation with dexamethasone (162); progesterone (163); high glucose (35, 73),
Functional characterization of the upregulated genes

Use EnrichR to find enriched functional categories in the list of upregulated genes.

Enriched GO biological processes? Use p-value cutoff of 0.01

Does it correspond to the results of ToppGene?
I checked with QuickGO and ToppGene list seems to be the correct one.
You can revert the question: Gene set enrichment analysis

Determines if an *a priori* defined set of genes shows statistically significant concordant differences between two biological sources

Instead of starting from a set of DE genes and looking for overrepresented categories, you start from a set of genes belonging to a category and see if they are DE.
Analyzing the pathways in your list via IPA / Cytoscape

http://www.cytoscape.org/

Free, many apps for additional analyses in Cytoscape

e.g. iRegulon to identify regulatory networks (TF + targets) in your list

GeneMania to identify functional networks in your list

Tutorials: http://opentutorials.cgl.ucsf.edu/index.php/Portal:Cytoscape3

VIB scientists can use Ingenuity Pathway Analysis for free!


Heavily curated, high quality

Human, mouse, rat

Custom training: bits@vib.be
Functional characterization of the upregulated genes

Use iRegulon to find TF that might regulate the upregulated genes

See:
#iRegulon_detects_regulatory_networks_in_a_set_of_genes
Tools for gene set enrichment analysis

GeneTrail: http://genetrail.bioinf.uni-sb.de/enrichment_analysis.php?js=1&cc=1

WebGestalt: http://www.webgestalt.org/

Can do both types of analyses
Splice variant identification

Go for long reads via Nanopore or PacBio

PacBio generates reads of a few kb (= full transcripts)

Short reads:

Kallisto/Salmon are faster and more accurate than STAR + Cufflinks

https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-017-4002-1

=> Counts/transcript instead of /gene

Check out IPA – Isoprofiler tool for visualization and analysis:

https://tv.qiagenbioinformatics.com/v.ihtml/player.html?source=embed&token=40736f30cc24ce800162df1063afe51a&photo_id=14619975