Hardware requirements for analysis of NGS data

**Variant analysis**
- 1 human full genome sample
- 4 GB RAM
- 4 Intel i5 CPUs
- 100 GB

**RNA-Seq**
- 6 human single chromosome samples
- 16 GB RAM
- 4 Intel i5 CPUs
- 60 GB

---

Not feasible / very slow on a normal laptop
We use computer with 6 cores and 48 GB RAM
What you get from your provider: Illumina raw data files

Sequence files: _seq.txt  
raw sequences

Probability files: _prb.txt  
quality score of each base in each cycle

**FASTQ file:** combines raw sequences and quality scores

Qseq files: same info as FASTQ but different format

SCARF files: sequences and quality scores on a single line

http://nar.oxfordjournals.org/content/38/6/1767.long
http://en.wikipedia.org/wiki/FASTQ_format
Illumina sequence files contain the reads

<table>
<thead>
<tr>
<th>Lane number</th>
<th>Tile number</th>
<th>X pos</th>
<th>Y pos</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>137</td>
<td>689</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>87</td>
<td>649</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>121</td>
<td>642</td>
</tr>
</tbody>
</table>

You don’t know which genomic region the read comes from
You only know the location of the cluster that gave rise to the read
Illumina uses **cluster location** as IDs
Probability files contain confidence of base calls

Quality scores calculated by Illumina software

How confident is software about base call?

According to Illumina: score > 20 = OK

<table>
<thead>
<tr>
<th>score</th>
<th>confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>40</td>
<td>99,99%</td>
</tr>
<tr>
<td>30</td>
<td>99,9%</td>
</tr>
<tr>
<td>20</td>
<td>99%</td>
</tr>
<tr>
<td>10</td>
<td>90%</td>
</tr>
<tr>
<td>0</td>
<td>0%</td>
</tr>
</tbody>
</table>
Illumina FASTQ files are used for analysis

Divided into blocks of **4 lines**

<table>
<thead>
<tr>
<th>Machine ID</th>
<th>Run ID</th>
<th>Lane</th>
<th>Tile</th>
<th>Location of cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>@ILMN-GA001 3 208HWAAAXX</td>
<td>1 1 110 812</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATACAAGCAAGTATAAGTTCTGATGCCTCTT</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+ILMN-GA001 3 208HWAAAXX 1 1 110 812</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hhhYhh]NYhhhhhhhYIhhaZT [hYHNSPKXR</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>@ILMN-GA001 3 208HWAAAXX 1 1 111 879</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GGAGGCTGGAGTGGGGGACGTATGCCTGCGCATAG</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+ILMN-GA001 3 208HWAAAXX 1 1 111 879</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hSWhRNJ\hfLdhVOhAIb@NFKD@PAB?N?</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Illumina quality scores in FASTQ files

String of same length as sequence: 1 score for each nucleotide

Quality scores represented as ASCII characters (ASCII-64)

http://en.wikipedia.org/wiki/ASCII

-> Phred score + 64   e.g. Phred score = 8   ->   8 + 64 = 72
-> ASCII character that corresponds to 72 = H
Why different ASCII offsets?

<table>
<thead>
<tr>
<th>Binary</th>
<th>Oct</th>
<th>Dec</th>
<th>Hex</th>
<th>Glyph</th>
</tr>
</thead>
<tbody>
<tr>
<td>010 0000</td>
<td>040</td>
<td>32</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>010 0001</td>
<td>041</td>
<td>33</td>
<td>21</td>
<td>!</td>
</tr>
<tr>
<td>010 0010</td>
<td>042</td>
<td>34</td>
<td>22</td>
<td>&quot;</td>
</tr>
<tr>
<td>010 0011</td>
<td>043</td>
<td>35</td>
<td>23</td>
<td>#</td>
</tr>
<tr>
<td>010 0100</td>
<td>044</td>
<td>36</td>
<td>24</td>
<td>$</td>
</tr>
<tr>
<td>010 0101</td>
<td>045</td>
<td>37</td>
<td>25</td>
<td>%</td>
</tr>
<tr>
<td>010 0110</td>
<td>046</td>
<td>38</td>
<td>26</td>
<td>&amp;</td>
</tr>
<tr>
<td>010 0111</td>
<td>047</td>
<td>39</td>
<td>27</td>
<td>'</td>
</tr>
<tr>
<td>010 1000</td>
<td>050</td>
<td>40</td>
<td>28</td>
<td>(</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Binary</th>
<th>Oct</th>
<th>Dec</th>
<th>Hex</th>
<th>Glyph</th>
</tr>
</thead>
<tbody>
<tr>
<td>100 0000</td>
<td>100</td>
<td>64</td>
<td>40</td>
<td>@</td>
</tr>
<tr>
<td>100 0001</td>
<td>101</td>
<td>65</td>
<td>41</td>
<td>A</td>
</tr>
<tr>
<td>100 0010</td>
<td>102</td>
<td>66</td>
<td>42</td>
<td>B</td>
</tr>
<tr>
<td>100 0011</td>
<td>103</td>
<td>67</td>
<td>43</td>
<td>C</td>
</tr>
<tr>
<td>100 0100</td>
<td>104</td>
<td>68</td>
<td>44</td>
<td>D</td>
</tr>
<tr>
<td>100 0101</td>
<td>105</td>
<td>69</td>
<td>45</td>
<td>E</td>
</tr>
<tr>
<td>100 0110</td>
<td>106</td>
<td>70</td>
<td>46</td>
<td>F</td>
</tr>
<tr>
<td>100 0111</td>
<td>107</td>
<td>71</td>
<td>47</td>
<td>G</td>
</tr>
<tr>
<td>100 1000</td>
<td>110</td>
<td>72</td>
<td>48</td>
<td>H</td>
</tr>
</tbody>
</table>

Before 64: invisible characters, symbols, numbers

Before 33: only invisible characters

Always check which ASCII version is used!
In FASTQ files reads of a pair are linked

<table>
<thead>
<tr>
<th>Machine ID</th>
<th>Run ID</th>
<th>x:y coord.</th>
<th>Read pair #</th>
</tr>
</thead>
<tbody>
<tr>
<td>@HWI-ST395_0083:3:1:3429:2628#0/1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Pair: read in other file: @HWI-ST395_0083:3:1:3429:2628#0/2
What you get from your provider: single cell raw data files

Illumina sequencer's base call files .bcl

Binary representation of base calls and quality scores

10x genomics software transforms bcl into **multiple fastq files** for every sample:

Index file:

I1.fastq sample indexes

Read files:

R1.fastq cell barcodes + UMIs

R2.fastq files contain reads

General quality metrics file
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 expect standard FASTQ files -> crashes / errors

**Groomer** can clean messy FASTQ files

Galaxy / BITS GenePattern

*Definition of FASTQ format and description of all variants*

[http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2847217/](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2847217/)
NGS data analysis

Most analysis tools only work in Linux / Mac via commands

```
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**
  - **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).
<table>
<thead>
<tr>
<th>RNA-Seq workflow</th>
<th>DNA-Seq workflow</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Quality control</td>
<td>1. Quality control</td>
</tr>
<tr>
<td>3. Processing mapping results</td>
<td>3. Processing mapping results</td>
</tr>
<tr>
<td>4. Extract count table</td>
<td>4. Calling variants</td>
</tr>
<tr>
<td>5. Test for differential expression</td>
<td>5. Processing variant files</td>
</tr>
<tr>
<td>7. Optional: call variants</td>
<td></td>
</tr>
</tbody>
</table>
NGS analysis workflow

Checking quality of the reads

Improving quality of the reads

Checking quality of cleaned reads

FASTQ with raw reads

@ILMN-GA001_3_208HWAAXX_1_1_110_812
ATACAAGCAAGTATAAGTTCGTATATGCCCCCC
+ILMN-GA001_3_208HWAAXX_1_1_110_812
hhhYhh]NYhhhhhhhhYIhhaZT[hY
@ILMN-GA001_3_208HWAAXX_1_1_111_879
GGAGGCTGGAGTTGGGGACGTATGCGCATAGAC
+ILMN-GA001_3_208HWAAXX_1_1_111_879
hSWhRNJ\hFhLdhVOAIB@NFKDEPAB@NBBB
@ILMN-GA001_3_208HWAAXX_1_1_112_938
ACATCTCATTACTTGGAAAAGTGAAGTCTAGG
+ILMN-GA001_3_208HWAAXX_1_1_112_938
h?\ah\XTfhpVhh`YKSLJXPUNJZKLJXLBBB

FASTQ with clean reads
FASTQC checks the quality of your data

http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Easy to use tool that works on Windows, Mac and Linux

Checks: quality of the reads
    quality of the bases
    contamination in your libraries

Always perform QC on your data!

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

    discuss with experts (provider or bioinformatician)
FASTQC: Basic Statistics

**Encoding:** ASCII encoding of quality values

**Total Sequences:** total number of sequences
should correspond to coverage ordered from provider

**Filtered Sequences:** Illumina does no filtering

**Sequence Length:** length of shortest and longest sequence
Illumina: all reads have same length
FASTQC: Sequence Length Distribution

**Illumina**: all sequences have $= \text{length}$

**Warning** if sequences have $\neq \text{length}$
FASTQC: Per base sequence quality

Distribution of quality scores for each position

For each position: 1 box plot

- Median score
- Mean score
- Center 50% of the scores
- Center 80% of the scores
FASTQC checks the quality of the bases

Per base sequence quality

Warning if end of box < 10 or median < 25

Failure if end of box < 5 or median < 20
Sequence quality: Good versus Bad quality

You always see a drop in quality scores towards 3’ ends

FastQC guesses quality encoding but it can make an error
(especially when data is very good!)

Using Groomer prior to FASTQC solves this problem

http://dwheelerau.com/tag/fastqc/
FASTQC guesses the ASCII encoding of your FASTQ file

Sanger -> ASCII33
Solexa and Illumina 1.0 -> ASCII64
Illumina 1.3 and 1.5 -> ASCII64
Illumina 1.8 and higher -> ASCII33

Definition of FASTQ format and description of all variants
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2847217/
Why is the quality of the 3′ bases always lower?

There’s always a drop in quality towards the 3′ end

**Cycle 1: clear signal – A**

Sequencing primer

Cycle 1
Why is the quality of the 3’ bases always lower?

Cycle 2: sometimes no base is added

still clear signal – T (1000 fragments/cluster)

Sequencing primer

Cycle 1
Cycle 2
Why is the quality of the 3' bases always lower?

Cycle 3: signal gets diluted – a lot of A but also T

More cycles -> more errors -> more dilution of signal

Fix: Trim low quality bases from 3' end
FASTQC checks the quality of the reads

**Per sequence quality**

**Warning** if peak < 27 (0.2% error rate)

**Failure** if peak < 20 (1% error rate)
FASTQC checks the quality of the reads

Good quality = blue - Not OK = other colors

Dirt, smudges, something getting stuck in the flow cell
Too many clusters in the tile
Remove low quality reads
Low quality in specific areas over a few cycles

Air bubbles or dirt passing through and washing out
Low quality calls in the center of the reads -> cannot be fixed
Quality control: Duplicate sequences

Not biological copies but result of technical issues:

- e.g. overamplification by PCR
  - same read was detected twice (borders of tiles)
  - contamination of the library

Remove because duplicates will distort results

RNASEq: normal (highly expressed genes)

Do not remove

http://proteo.me.uk/2011/05/interpreting-the-duplicate-sequence-plot-in-fastqc/
10 single reads  
\[10 \times 1 = 10 \rightarrow 50\%

5 reads duplicated twice:  
\[5 \times 2 = 10 \rightarrow 50\%

Deduplicated:  
Single: 10 reads  
\[\rightarrow 66\%

Twice: 5 reads  
\[\rightarrow 33\%

10 single reads  
\[10 \times 1 = 10 \rightarrow 50\%

1 read duplicated 10x:  
\[1 \times 10 = 10 \rightarrow 50\%

Deduplicated:  
Single: 10 reads  
\[\rightarrow 91\%

10x: 1 read  
\[\rightarrow 9\%
FASTQC: Duplicate Sequences

Percentage of sequences with different levels of duplication

**Warning** if > 20% of reads would be lost in case of deduplication

**Failure** if > 50% of reads would be lost in case of deduplication

63% of reads lost if library is deduplicated

Percent of seqs remaining if deduplicated: 36.77%
FASTQC: Interpretation of duplicate sequences plot

**Strong peak left:** random, diverse library

**Peak at the right:** small genome

  contamination with adapter dimers

  -> look at FASTQC overrepresented sequences module

**Flattened lines:** library with low diversity (ChIP-Seq)

  library from RNA-Seq

! If sequence quality is very low you will also have a lot of errors in first 50 bases!
FASTQC: Per Base Sequence Content – Good quality

Sequence content across all bases

Nucleotide frequencies for each position

OK -> no difference between positions -> 4 parallel lines

Some variation at 5’ends of reads is normal for Illumina data
FASTQC: Per Base Sequence Content – Bad quality

Most reads have same sequence at 5’ -> PROBLEM !
First cycles are used for calibration of base calling
Bad calibration due to biased sequence composition
FASTQC: Per Base Sequence Content – Bad quality

Presence of identical reads distorts the plot

**Warning** if A-T or G-C > 10%

**Failure** if A-T or G-C > 20%
GC plot shows the same info as Per Base Sequence Content plot

Identical reads introduce sharp spike in GC plot
FASTQC checks if the library was contaminated

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Count</th>
<th>Percentage</th>
<th>Possible Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>GATCGGAAGAGCTCGTATGCCGTCTTCTGCTG...</td>
<td>1870684</td>
<td>19.447</td>
<td>Illumina Single End Adapter 1...</td>
</tr>
<tr>
<td>GAAGAGCTCGTATGCCGTCTTCTGCTTGAA...</td>
<td>95290</td>
<td>0.991</td>
<td>Illumina Single End Adapter 1...</td>
</tr>
<tr>
<td>GGGGAGGAGAGCCATTGTGGAGCCGGCC...</td>
<td>46910</td>
<td>0.488</td>
<td>No Hit</td>
</tr>
<tr>
<td>AGGGAGGAGAGCCATTGTGGAGCCGGCC...</td>
<td>21029</td>
<td>0.219</td>
<td>No Hit</td>
</tr>
<tr>
<td>GTGGGAGGCTTTTTAAAAGCTGGGAGAGGGT...</td>
<td>15355</td>
<td>0.16</td>
<td>No Hit</td>
</tr>
<tr>
<td>GGGGGAGAGAGCCATTGTGGAGCCGGCC...</td>
<td>15262</td>
<td>0.159</td>
<td>No Hit</td>
</tr>
</tbody>
</table>

Reads that occur in many identical copies are BLASTed against a library of known contaminants

Adapter dimers: adapters ligated to each other – no fragment

Fix: Trim adapter sequences
Sources of overrepresented sequences

**Adapter dimers**

PhiX virus sequences: spiked in by sequencing center as a control in organisms with extreme GC content (Arabidopsis, Plasmodium, bacteria...)

- low diversity libraries

DNA-Seq: satellites (= millions of highly similar copies of a given repeated sequence)

RNA-Seq: reads from rRNA genes (even if you remove rRNA during library prep)

No Hit: variations on contaminating sequences with sequencing errors
Exercises: RNA-Seq data set for this training

**Dataset SRR074262 – SRR074263**

*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 file
Exercises: ChIP-Seq data set for this training

Dataset SRR576933 – SRR576938

Escherichia coli str. K-12 substr. MG1655

ChIPSeq

Illumina single-end

9 samples in 9 runs: 2 after IP with antibody against TF FNR

4 after IP with antibody against TF sigma 70

3 without IP

Wiki exercises:
Quality control of NGS data
**NGS analysis workflow**

- **Checking quality of the reads**
- **Improving quality of the reads**
- **Checking quality of cleaned reads**

**FASTQ with raw reads**

```
@ILMN-GA001_3_208HWAAXX_1_1_110_812
ATACAAGCAAGTATAATGCGTATATGCCGCTTT
+ILMN-GA001_3_208HWAAXX_1_1_110_812
hhhyYhh]NYhhhhhhhYIhhaZT[yHYNNKFSM
@ILMN-GA001_3_208HWAAXX_1_1_111_879
GGAGGCTGGAGTGGGACGTATGCGGCATAGAC
+ILMN-GA001_3_208HWAAXX_1_1_111_879
hSWhRNJ\hFhLdhVOAIB@NFKD@PAB@NBBBB
@ILMN-GA001_3_208HWAAXX_1_1_112_938
ACATCTCATTACTTGGAATGAGCTAGCTAGCG
+ILMN-GA001_3_208HWAAXX_1_1_112_938
h?\ah\XTfhhPhVhh`YKSLJXPUNJZKLJXLB
```

**FASTQ with clean reads**

```
@ILMN-GA001_3_208HWAAXX_1_1_110_812
ATACAAGCAAGTATAATGCGTATATG
+ILMN-GA001_3_208HWAAXX_1_1_110_812
hhhyYhh]NYhhhhhhhYIhhaZT[yY
@ILMN-GA001_3_208HWAAXX_1_1_112_938
ACATCTCATTACTTGGAATGAGCTAGCTAGG
+ILMN-GA001_3_208HWAAXX_1_1_112_938
h?\ah\XTfhhPhVhh`YKSLJXPUNJZKLJXV
```
Improving the quality of the reads: Filtering / Trimming

Filtering: removal of sequences

Trimming: removal of bases in sequences

Remove contamination

Cut low quality bases from the 3’end of the reads

Remove low quality reads

Remove polyA tails

Demultiplex: split reads over multiple fastq files based on barcodes

Filtering: removal of sequences

Trimming: removal of bases in sequences
Quality control: Filtering and trimming

**RNASeq** mappers **ignore** mismatches at the end of reads = soft clipping
FASTQC detects bad tiles, reads => **filter low quality reads**
FASTQC detects **heavy** adapter contamination => **trim adapters**

**DNASeq** mappers **do not ignore** mismatches at the end of reads
**Trim low quality bases** at 3’ ends
FASTQC detects bad tiles, reads => **filter low quality reads**
FASTQC detects adapter contamination => **trim adapters**

No filtering/trimming => lower percentage of mapped sequences
Trimming low quality bases from the ends of the reads

5’ trimming bases with quality below 20

3’ trimming bases with quality below 20
Filtering low quality reads
### Trimming adapter sequences

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Count</th>
<th>Percentage</th>
<th>Possible Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>GATCGGAAGAGCTCGTATGCCGTCTTCTGCG...</td>
<td>1870684</td>
<td>19,447</td>
<td>Illumina Single End Adapter 1...</td>
</tr>
<tr>
<td>GAAGAGCTCGTATGCCGTCTTCTGCTTTGAA...</td>
<td>95290</td>
<td>0,991</td>
<td>Illumina Single End Adapter 1...</td>
</tr>
<tr>
<td>GGGGAGGGAGAGGAGAGCCATTGTTGAGGCCGCGC...</td>
<td>46910</td>
<td>0,488</td>
<td>No Hit</td>
</tr>
<tr>
<td>AGGGGAGGGAGAGGAGAGCCATTGTTGAGGCCGCGC...</td>
<td>21029</td>
<td>0,219</td>
<td>No Hit</td>
</tr>
<tr>
<td>GTGGGGAGTGGTATAACAACTGGAGGAGGTTG...</td>
<td>15355</td>
<td>0,16</td>
<td>No Hit</td>
</tr>
<tr>
<td>GGGGGGGAGAGAGAGGAGCCATTGTTGAGGCCGCGC...</td>
<td>15262</td>
<td>0,159</td>
<td>No Hit</td>
</tr>
<tr>
<td>GTCGGATTAAAACAGCTCGGACAGTGAAAGGA...</td>
<td>10002</td>
<td>0,104</td>
<td>No Hit</td>
</tr>
</tbody>
</table>
Removing duplicates?

Genome sequencing / ChIP-Seq:
Remove duplicates so they do not result in false variant / peak calling

RNASeq:
Remove duplicates while keeping track of counts

Removing duplicates can be done based on:

**Sequence:** duplicates = exact copies of reads

**Location on flow cell:** duplicates = reads with same ID

Mapping location: duplicates = reads that map to exactly same location

How to see if duplicates are PCR duplicates?
https://ethanomics.wordpress.com/2012/01/06/to-filter-or-not-to-filter-duplicate-reads-chip-seq/
Tools for Filtering / Trimming

**FASTX**: Galaxy

**Trimmomatic**: GenePattern

**Cutadapt**: Python script that you can run at command line

**Kraken**: command line
## Functionality of trimming tools

<table>
<thead>
<tr>
<th>Function</th>
<th>Cutadapt</th>
<th>FASTX</th>
<th>Kraken</th>
<th>Trimmomatic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adapter/primer removal</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Quality trimming</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Removal of bases with B scores</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Removal of polyA tails</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Removal of Ns</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Demultiplex: filter reads based on barcodes</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Removing duplicates</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Identify adapter sequences</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Convert offset of Phred scores</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
</tbody>
</table>
Parameters of trimming tools

Adapter sequence (in fasta format): see FASTQC

Minimum phred score to keep:
- 3: you should at least do this
- 10: 90% confidence - short reads
- 20: 99% confidence - longer reads

Minimum size after trimming -> very short reads map everywhere: 15 - 20

Which end to trim: 3' or 5' or both?
- In most cases 3' is sufficient: this is the side where quality drops
NGS analysis workflow

Checking quality of the reads
Improving quality of the reads
Checking quality of cleaned reads

FASTQ with raw reads

@ILMN-GA001_3_208HWAAXX_1_1_110_812
ATACAGCAAGTATAAGTTTGATATATGCAGCCTTT

+ILMN-GA001_3_208HWAAXX_1_1_110_812

@ILMN-GA001_3_208HWAAXX_1_1_111_879
GGAGGCTGAGTGTGGGACGTATGCGGCATAGAC

+ILMN-GA001_3_208HWAAXX_1_1_111_879

@ILMN-GA001_3_208HWAAXX_1_1_112_938
ACATCTCATTACCTTGGAAAAAGTGAAGTCTAGCGG

+ILMN-GA001_3_208HWAAXX_1_1_112_938

FASTQ with clean reads

@ILMN-GA001_3_208HWAAXX_1_1_110_812
ATACAGCAAGTATAAGTTTGATATATGCAGCCTTT

+ILMN-GA001_3_208HWAAXX_1_1_110_812

@ILMN-GA001_3_208HWAAXX_1_1_111_879
GGAGGCTGAGTGTGGGACGTATGCGGCATAGAC

+ILMN-GA001_3_208HWAAXX_1_1_111_879

@ILMN-GA001_3_208HWAAXX_1_1_112_938
ACATCTCATTACCTTGGAAAAAGTGAAGTCTAGCGG

+ILMN-GA001_3_208HWAAXX_1_1_112_938
NGS generates a lot of errors !!

NGS has a 10x higher error rate than Sanger sequencing

Underrepresentation of AT- and GC-rich regions in the sequence reads

Why?

Denaturation of GC-rich regions is difficult

Primer annealing to AT-rich regions is difficult
Oversampling to lower the influence of these errors

Sequence the same region multiple times -> High coverage

**Coverage** = *average* number of times a base is represented in the reads

ATACGAACGATAAAATATGATTACCATTATACCTCTTTCAAAAAAGTTCTCTAGTTTCAG
ATACGAACGATAACTATGATTACCATTATACTATTTCAAAAAAGTTCTCTAGTTTCAG
ATACGAACGATAAAATATGATTACCATTATACTCTTTCAAAAAAGTTCTCTAGTTTCAG
ATACGAACGATAAAATATGATTACCATTATACTATTTCAAAAAAGTTCTCTAGTTTCAG
ATACGAACGATAAAATATGATTACCATTATACTCTTTCAAAAAAGTTCTCTAGTTTCAG
ATACGAACGATAAAATATGATTACCATTATACTATTTCAAAAAAGTTCTCTAGTTTCAG
ATACGAACGATAAAATATGATTACCATTATACTCTTTCAAAAAAGTTCTCTAGTTTCAG
ATACGAACGATAAAATATGATTACCATTATACTATTTCAAAAAAGTTCTCTAGTTTCAG

Consensus

ATACGAACGATAAAATATGATTACCATTATACTATTTCAAAAAAGTTCTCTAGTTTCAG

http://www.slideshare.net/flxlex/ngs-techniques-and-data-relevant-for-metagenomics-analyses
What coverage do you need?

What you can pay for

Not a miracle solution: missing fragments will never be covered

Different definitions:

What you order from your provider are **millions of reads**

What you get after mapping is **coverage**
Which coverage do you choose for variant analysis?

For variant analysis very crucial question

Few samples, high coverage

- e.g. 30x

You find all variation in the samples but not between samples

More samples, low coverage

- e.g. 1000 genomes project > 1000 persons, 4x coverage

You see variation between samples but you are less confident about the sequence of each sample e.g. you don’t know if person is homo/heterozygous

For genotyping

For population genetics
Coverage for DNA-Seq used at the Nucleomics core

**Variant calling:** > 20-fold for haploid

- 40-fold for diploid
- 100-fold for cancer samples

> 8 reads must show variant to be called

**Human ChIP-Seq:**

**Minimum:** 5-10 million reads per sample

**Ideal:** 30-50 million reads per sample

Depends on antibody used for immunoprecipitation

ChIP and control samples are sequenced at different depths!

More reads for control sample to ensure sufficient coverage of the genome
Which coverage do you choose for RNA-Seq?

Depends on goal of the experiment:

(Splice) variant analysis: a lot of reads

Differential expression: less reads

but risk of missing genes with very low expression levels

Prefer biological replicates over coverage!
Nucleomics core guidelines on coverage for RNA-Seq

**DE:** 15 million single end Illumina reads per sample

Depends on state of RNA

<table>
<thead>
<tr>
<th>mRNA</th>
<th>total RNA</th>
</tr>
</thead>
<tbody>
<tr>
<td>good</td>
<td>15</td>
</tr>
<tr>
<td>degraded</td>
<td>30</td>
</tr>
<tr>
<td>bad</td>
<td>60</td>
</tr>
</tbody>
</table>

**Fusion detection:** 50 million paired end reads
JOINING OF TWO GENES BY TRANSLOCATION OR INVERSION

**Splice variants:** 80 million paired end reads

Comparison of RNA-Seq methods for degraded or low input samples

http://www.nature.com/nmeth/journal/v10/n7/full/nmeth.2483.html
NGS analysis workflow

*FASTQ with clean reads*

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

Obtaining reference sequence

Mapping to reference

@SQ SN:chr20 LN:64444167
@PG ID:BWA VN:0.7.15 CL:/srv/dna_tools/bwa/bwa mem -M -t 4 HiSeq_UCSC_hg19.fa
ILMN-GA001:3:208HWAAXX:1:1:110:812 16 chr20 190938 30 18M2I4M1D3M *
ATACAAGCAAGTATAAGTTCGTATATG hhhYhh]NYhhhhhhhhYIhhaZT[hY
ILMN-GA001_3_208HWAAXX_1_1_112_938 0 chr20 1493131 10 11M2P7I13M *

*BAM or SAM with alignments*
Where do you find the reference genome?

Built-in in Galaxy/GenePattern

Igenomes + Bowtie website:


http://support.illumina.com/sequencing/sequencing_software/igenome.html

You can immediately use them in Bowtie/HISAT mappers

Ensembl: http://www.ensembl.org/

UCSC: http://genome.ucsc.edu/


You have to index them first but Galaxy automatically handles this for you.
Indexing makes genomes easier to access

Genomes are available in fasta format

These are very large files

Aligning a read to such a file is comparable to searching for a word in a phone book

Indexing the file make the search faster
Ensembl versus UCSC: what’s the difference?

Sequences should be the same but this is not always true

**mouse**: UCSC only strain C57BL/6J

NCBI/Ensembl also other strains

**human** previous version (hg19): UCSC full genome

NCBI/Ensembl forgot mitochondrial DNA

Annotations and IDs are different


Comparison Ensembl – UCSC – Refseq annotation for RNA-Seq

=> For RNA-Seq Ensembl is the best choice
Different names for the same genome release

Mapping between UCSC names and NCBI + Ensembl names:

https://genome.ucsc.edu/FAQ/FAQreleases.html

<table>
<thead>
<tr>
<th>SPECIES</th>
<th>UCSC VERSION</th>
<th>RELEASE DATE</th>
<th>RELEASE NAME</th>
<th>STATUS</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAMMALS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Human</td>
<td>hg38</td>
<td>Dec. 2013</td>
<td>Genome Reference Consortium GRCh38</td>
<td>Available</td>
</tr>
<tr>
<td></td>
<td>hg19</td>
<td>Feb. 2009</td>
<td>Genome Reference Consortium GRCh37</td>
<td>Available</td>
</tr>
<tr>
<td></td>
<td>hg18</td>
<td>Mar. 2006</td>
<td>NCBI Build 36.1</td>
<td>Available</td>
</tr>
<tr>
<td></td>
<td>hg17</td>
<td>May 2004</td>
<td>NCBI Build 35</td>
<td>Available</td>
</tr>
<tr>
<td></td>
<td>hg16</td>
<td>Jul. 2003</td>
<td>NCBI Build 34</td>
<td>Available</td>
</tr>
<tr>
<td></td>
<td>hg15</td>
<td>Apr. 2003</td>
<td>NCBI Build 33</td>
<td>Archived</td>
</tr>
</tbody>
</table>

GRCh38 or GRCh37?
- GRCh37 no mitochondrial genome while GRCh38 and hg38 are same full sequence
- GRCh38 solved assembly gaps and incorrect alleles
- Annotation is slow: more annotations available for GRCh37
To map RNA-Seq reads you also need annotation

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

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

You can download these as **gtf** files from Ensembl / UCSC

<table>
<thead>
<tr>
<th>chr21</th>
<th>transcript</th>
<th>10862622</th>
<th>10863067</th>
<th>+</th>
<th>gene_id &quot;ENSG00000169..&quot;</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr21</td>
<td>exon</td>
<td>10862622</td>
<td>10862667</td>
<td>+</td>
<td>gene_id &quot;ENSG00000169..&quot;</td>
</tr>
<tr>
<td>chr21</td>
<td>CDS</td>
<td>10862622</td>
<td>10862667</td>
<td>+</td>
<td>gene_id &quot;ENSG00000169..&quot;</td>
</tr>
<tr>
<td>chr21</td>
<td>start_codon</td>
<td>10862622</td>
<td>10862624</td>
<td>+</td>
<td>gene_id &quot;ENSG00000169..&quot;</td>
</tr>
<tr>
<td>chr21</td>
<td>exon</td>
<td>10862751</td>
<td>10863067</td>
<td>+</td>
<td>gene_id &quot;ENSG00000169..&quot;</td>
</tr>
<tr>
<td>chr21</td>
<td>CDS</td>
<td>10862751</td>
<td>10863064</td>
<td>+</td>
<td>gene_id &quot;ENSG00000169..&quot;</td>
</tr>
<tr>
<td>chr21</td>
<td>stop_codon</td>
<td>10863065</td>
<td>10863067</td>
<td>+</td>
<td>gene_id &quot;ENSG00000169..&quot;</td>
</tr>
<tr>
<td>chr21</td>
<td>UTR</td>
<td>10863065</td>
<td>10863067</td>
<td>+</td>
<td>gene_id &quot;ENSG00000169..&quot;</td>
</tr>
</tbody>
</table>
To map targeted DNAsSeq reads you also need annotation

You need **the location of the exons / targeted regions**

You can get these as **bed** files from the provider of the PCR/capture kit

=> you will get the file that is valid for their oligos

<table>
<thead>
<tr>
<th>chr7</th>
<th>127471196</th>
<th>127472363</th>
<th>gene1</th>
<th>0</th>
<th>+</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr7</td>
<td>127472363</td>
<td>127473530</td>
<td>gene2</td>
<td>2</td>
<td>+</td>
</tr>
<tr>
<td>chr7</td>
<td>127473530</td>
<td>127474697</td>
<td>gene3</td>
<td>20</td>
<td>+</td>
</tr>
<tr>
<td>chr7</td>
<td>127474697</td>
<td>127475864</td>
<td>gene4</td>
<td>3</td>
<td>+</td>
</tr>
<tr>
<td>chr7</td>
<td>127475864</td>
<td>127477031</td>
<td>gene5</td>
<td>100</td>
<td>-</td>
</tr>
<tr>
<td>chr7</td>
<td>127477031</td>
<td>127478198</td>
<td>gene6</td>
<td>3</td>
<td>-</td>
</tr>
<tr>
<td>chr7</td>
<td>127478198</td>
<td>127479365</td>
<td>gene7</td>
<td>5</td>
<td>-</td>
</tr>
<tr>
<td>chr7</td>
<td>127479365</td>
<td>127480532</td>
<td>gene8</td>
<td>1</td>
<td>+</td>
</tr>
<tr>
<td>chr7</td>
<td>127480532</td>
<td>127481699</td>
<td>gene9</td>
<td>3</td>
<td>-</td>
</tr>
</tbody>
</table>
**NGS analysis workflow**

**FASTQ with clean reads**

```plaintext
@ILMN-GA001_3_208HWAAXX_1_1_110_812
ATACAAGCAAGTATAAGTTCGTATATG
+ILMN-GA001_3_208HWAAXX_1_1_110_812
hhhYhh]NYhhhhhhhhYIhhhaZT[hY
@ILMN-GA001_3_208HWAAXX_1_1_112_938
ACATCTCATTACTTGGAAAAGTGAAGTCTAGG
+ILMN-GA001_3_208HWAAXX_1_1_112_938
h?\ah\XTfhPhVhh`YKSLJXPUNJZKLJXV
```

**Obtaining reference sequence**

**Mapping to reference**

```plaintext
@SQ SN:chr20 LN:64444167
@PG ID:BWA VN:0.7.15 CL:/srv/dna_tools/bwa/bwa mem -M -t 4 HiSeq_UCSC_hg19.fa
ILMN-GA001:3:208HWAAXX:1:1:110:812 16 chr20 190938 30 18M2I4M1D3M *
ATACAAGCAAGTATAAGTTCGTATATG hhhYhh]NYhhhhhhhhYIhhf_T`h^
ILMN-GA001_3_208HWAAXX_1_1_112_938 0 chr20 1493131 10 11M2P7I13M *
```

**BAM or SAM with alignments**
Aligning reads to a reference sequence

Reference genome/EST sequence(s)

Variant analysis:

Compare the reads and the reference sequence
Identify sequence differences

Other goals:

Map the reads to find out

Which location on the reference sequence each fragment comes from?
Which gene? Which exon? ...
Overview of mappers of short DNA reads


reads < 200nt

**Bowtie:**  [http://bowtie-bio.sourceforge.net/bowtie/](http://bowtie-bio.sourceforge.net/bowtie/)

Same algorithm as BWA

Faster

Good at aligning short reads (< 50 nt)

**Bowtie2:**  [http://bowtie-bio.sourceforge.net/bowtie2/](http://bowtie-bio.sourceforge.net/bowtie2/)

Better for longer reads (> 50 nt)
Aligning short reads is a challenge

Single end sequencing

Reference

read 1A

read 1

read 2

read 3

Many short reads map to multiple locations -> which is right location?

Paired end sequencing

Reference

read 1A

read 1B

read 1A

read 1B

Reads from a pair have to map to opposite strands

In each others vicinity: separated by fragment length
Mapping RNA-Seq reads is an even greater challenge.

Mappers must allow gapped alignments.

Gaps = intron length.
Overview of mappers for RNA-Seq

**STAR:** [https://code.google.com/p/rna-star/](https://code.google.com/p/rna-star/)
- current standard
- very fast but requires a lot of memory
- cannot work with alternative alleles in reference

**HISAT2:** [http://ccb.jhu.edu/software/hisat/index.shtml](http://ccb.jhu.edu/software/hisat/index.shtml)
- fast and requires less resources (RAM, CPU)
- can take into account alternative alleles e.g. hg38

Both can handle splice variant processing, fusions, polyA tails...

Both can do soft clipping

Parameters of mappers

Indexed reference sequence (+ annotation)

Number of mismatches allowed in alignments:

5% of read length unless FASTQC tells you otherwise (low quality positions)

Paired end reads

Distance between reads of a pair = max. fragment length + 50

RNA-Seq

Map to what? Genome / Transcriptome / Both

both: faster than genome but you do find novel transcripts

Default maximum intron length = 500000 => not OK for plants/yeast...

What is SAM format?

Sequence Alignment Map

Standardized format for storing reads aligned to reference

Tab-delimited text: human readable


**header** section:

lines start with @

**alignment** section:

each line has 11 mandatory fields

+ optional fields
How to read a SAM file?

Alignment section:
Read001  163  chr20  7  20  8M2I4M1D3M ...
QNAME  FLAG  RNAME  POS  MAPQ  CIGAR

Required fields:
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
CIGAR CIGAR string: represents actual alignment
...
How to read a SAM file?

Read001  163  ref  7  20  8M2I4M1D3M = 37
QNAME  FLAG  RNAME  POS  MAPQ  CIGAR  MRNM  MPOS

Flag decoder: [https://broadinstitute.github.io/picard/explain-flags.html](https://broadinstitute.github.io/picard/explain-flags.html)
How to read a CIGAR string?

Alignment: 12345678901234  5678901234567890  
Reference: AGCATGTTAGATAA**GATAGCTGTGCTAGTA  
Read001+:   TTCGATAAAGGATA*CTG

Corresponding line in SAM file:
Read001   163   chr20   7   30   8M2I4M1D3M
QNAME    FLAG   RNAME   POS   MAPQ   CIGAR

CIGAR:
8M: first 8 bases of Read001 match reference
2I: then two insertions take place
4M: then again 4 matches
1D: then a deletion
3M: and finally again 3 matches
What is a BAM file?

Binary Alignment Map

**Binary representation of SAM:** not human readable

Exactly same information as SAM

More compact in size
NGS analysis workflow

**BAM or SAM with alignments**

Sorting and indexing

Quality control

Sorted BAM + BAI with alignments
SAM and BAM files have to be processed to improve accessibility

**Sorted:** reads are in same order as in fastq file

visualization needs reads sorted according to chromosome location

**Indexed:** easy access is provided by .bai file (= index)

For visualization and downstream processing you need sorted .bam AND .bai file
Software for processing SAM/BAM files

PICARD: set of tools

**SortSam:**  
**input:** bam file sorted according to location on flow cell  
**output:** bam file sorted according to genomic coordinates  
bai file containing the index

**BuildBamIndex:**  
**input:** bam file sorted according to genomic coordinates  
**output:** bai file containing the index

Samtools: set of tools

**sort:** sorting + indexing  
**index:** indexing  
**flagstat:** general mapping statistics
NGS analysis workflow

**BAM or SAM with alignments**

| @SQ | SN: chr20 | LN: 64444167 |
| @PG | ID: BWA | VN: 0.7.15 | CL: /srv/dna_tools/bwa/bwa mem -M -t 4 HiSeq_UCSC_hg19.fa |
| ILMN-GA001:3:208HWAAAXX:1:1:110:812 | 16 | chr20 | 190938 | 30 | 18M2I4M1D3M * |
| ATACAAGCAAGTATAAGTTGATATATG | hhhYhh]NYhhhhhhhhhYIhhf_T`h^ |
| ILMN-GA001_3_208HWAAAXX_1_1_112_938 | 0 | chr20 | 1493131 | 10 | 11M2P7I13M * |

**Sorting and indexing**

**Quality control**

| @SQ | SN: chr20 | LN: 64444167 |
| @PG | ID: BWA | VN: 0.7.15 | CL: /srv/dna_tools/bwa/bwa mem -M -t 4 HiSeq_UCSC_hg19.fa |
| ILMN-GA001:3:208HWAAAXX:1:1:110:812 | 16 | chr20 | 190938 | 30 | 18M2I4M1D3M * |
| ATACAAGCAAGTATAAGTGATATATG | hhhYhh]NYhhhhhhhhhYIhhf_T`h^ |
| ILMN-GA001_3_208HWAAAXX_1_3_88_990 | 16 | chr20 | 191056 | 34 | 19M2I16M * |

**Sorted BAM + BAI with alignments**
Tools for thoroughly checking the quality of the mapping

Qualimap: [http://qualimap.bioinfo.cipf.es/](http://qualimap.bioinfo.cipf.es/)

Command line / GUI on Windows / Mac / Linux (~ FASTQC)

<table>
<thead>
<tr>
<th>Summary</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Globals</strong></td>
<td></td>
</tr>
<tr>
<td>Reference size</td>
<td>4,639,675</td>
</tr>
<tr>
<td>Number of reads</td>
<td>6,605,032</td>
</tr>
<tr>
<td>Mapped reads</td>
<td>6,332,518 / 95,87%</td>
</tr>
<tr>
<td>Unmapped reads</td>
<td>272,514 / 4,13%</td>
</tr>
<tr>
<td>Mapped paired reads</td>
<td>0 / 0%</td>
</tr>
<tr>
<td>Read min/max/mean length</td>
<td>35 / 35 / 35</td>
</tr>
<tr>
<td>Duplicated reads (estimated)</td>
<td>3,122,607 / 47,28%</td>
</tr>
<tr>
<td>Duplication rate</td>
<td>55,25%</td>
</tr>
<tr>
<td>Clipped reads</td>
<td>0 / 0%</td>
</tr>
</tbody>
</table>

Duplicates: reads that map to the same position
Mean Coverage: every base of reference is on average represented 48x in the reads
General error rate: total number of mismatches + gaps / number of mapped bases
computed based on CIGAR strings
Factors that influence the quality of the mapping

Preprocessing:
Remove contamination => higher percentage of mapped reads
Trim low quality bases at the 3’ end => higher percentage of mapped reads

Mapping parameters:
Number of mismatches allowed in alignments
Paired end reads: distance between reads of a pair
RNA-Seq: intron length

Biology
Cancer cell lines: many rearrangements => 40% mapped reads is normal
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
Tools for viewing alignment results

**Grey:** nucleotide matches reference
**Colored:** nucleotide does not match reference
Arrows mark direction of the reads
Free servers for NGS data analysis

Galaxy:
Work on public server: https://usegalaxy.eu/
Install local: http://wiki.galaxyproject.org/Admin/Get%20Galaxy
Tutorials: http://wiki.galaxyproject.org/Learn

GenePattern:
Install local: http://www.broadinstitute.org/cancer/software/genepattern/download
Tutorials:
The Galaxy platform: navigation

Available tools

Fill in parameters forms and view analysis results

History: Overview outputs from previous tools
The Galaxy platform: history

Overview of all generated histories: automatically saved

Creating a new history

Sharing data with others

Saving all steps (tools) + parameters of current history to be able to do exactly the same analysis on other data

Manage history

Green: upload finished
Colors of data sets in the history reflect their status

2: DRR000542_1.fastq.gz

- Uploading is busy

3: Map with Bowtie for Illumina on data 1: mapped reads

- A tool is queued: the resulting dataset does not exist yet

3: Map with Bowtie for Illumina on data 1: mapped reads

- A tool is running: the resulting data is being generated

8: Map with Bowtie for Illumina on data 2 and data 1: mapped reads

- A tool has finished: the dataset is ready

17: Map with Mosaik on data 1

- A tool has encountered an error: the dataset is not to trust
The Galaxy platform: actions in the history

- displays the content of the data set in the middle pane
- edit the description of the data set e.g. the type of data, species the data come from...
- deletes the data set
- download the data set to your computer
- view more info, e.g. which parameters were used for generating the data...
- run this step again, displaying the parameter settings in the middle pane
- view in Trackster, the build-in genome browser
- tag the data set, so you can easily retrieve it by searching
- add comments, especially useful if sharing this data set with colleagues
GenePattern: free server for NGS data analysis

You can use public server: http://genepattern.broadinstitute.org/gp/pages/login.jsf

We have set up our own server

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

You can choose which tools you build in

We have set up a server on the VSC cluster @ KULeuven

We plan to set up a server on the VSC cluster @ UGent
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

Manipulate files

Load files
The GenePattern platform: navigation

- **Jobs = History**
- **Result files**
- **Manipulate result files**
- **Download**
- **View**
- **Save in GP**
Servers versus command line tools: PROs and CONs

PRO:  graphical user interface is very easy to use
      easy to share data, workflows and results
      history makes analyses reproducible

CON:  slower than command line tools
      Galaxy: less parameters to set than command line
      getting your data on the server can be hell
      not all tools are available
      data are on server: confidentiality ?
      -> local server

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2945788/

Group exercises
Galaxy and GenePattern @ VIB

VIB Bioinformatics Core GenePattern server Ghent:
https://dev.bits.vib.be:8686
Free for everyone, contact bits@vib.be for account

VIB Bioinformatics Core GenePattern server Leuven:
https://genepattern.hpc.kuleuven.be/gp
Free, you need KULeuven account, contact bits@vib.be for account

PSB Galaxy server: contact frederik.coppens@ugent.vib.be

Main Galaxy server: https://usegalaxy.eu or https://usegalaxy.org
All questions regarding NGS data analysis

Seqanswers: http://seqanswers.com/

http://seqanswers.com/wiki/SEQanswers

Biostar: http://www.biostars.org/

OMIC tools: http://omictools.com/

BARC: http://barcwiki.wi.mit.edu/wiki/SOPs

Wikibook on NGS:

http://en.wikibooks.org/wiki/Next_Generation_Sequencing_%28NGS%29