One slide biology summary

- Cells contain genetic material → DNA → large molecules called nucleic acids of them: Adenine, Thymine, Guanine, Cytosine
- Stored in long segments → chromosomes → sum of chromosomes → genome
- Parts of the genome get transcribed into RNA some RNA (mRNA) gets translated into proteins
- Proteins are made of amino acids (23 of them): Alanine, Threonine, Glycine, Cysteine, ...
- Sequencing technologies can be used to identify nucleic acids (DNA)
- Important: we cannot directly sequence DNA of a cell → there is always a laboratory protocol of substantial complexity → library preparation
- Sequencing instruments sequence libraries.

Sequencing Technologies - perspective

1st generation: Frederic Sanger develops DNA sequencing technology. Latest versions 3 million bases/day, 1500bp long reads

2nd generation: (next-gen) sequencing started 2005 with the release of the 454 sequencing platform. 600 billion bases/week, 150bp long reads

3rd generation: single molecule (no DNA amplification required), these are not replacing but augmenting 2nd generation systems, longer reads, shorter turnarounds
Sequencing Instruments

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Runs (cycles)</th>
<th>Million Reads</th>
<th>Reads/seq</th>
<th>Yield</th>
<th>Reads</th>
<th>Cost/MB</th>
<th>Cost/MG</th>
<th>Min cost (MB)</th>
<th>Min cost (MG)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3730xL (448 holds)</td>
<td>2 h</td>
<td>37000000</td>
<td>200000</td>
<td>0.04</td>
<td>86</td>
<td>$13000</td>
<td>$6500</td>
<td>$6.00</td>
<td>$6.00</td>
</tr>
<tr>
<td>4200xL (96 holds)</td>
<td>2 h</td>
<td>37000000</td>
<td>200000</td>
<td>0.04</td>
<td>86</td>
<td>$13000</td>
<td>$6500</td>
<td>$6.00</td>
<td>$6.00</td>
</tr>
<tr>
<td>4800xL (32 holds)</td>
<td>2 h</td>
<td>37000000</td>
<td>200000</td>
<td>0.04</td>
<td>86</td>
<td>$13000</td>
<td>$6500</td>
<td>$6.00</td>
<td>$6.00</td>
</tr>
<tr>
<td>480xL (16 holds)</td>
<td>2 h</td>
<td>37000000</td>
<td>200000</td>
<td>0.04</td>
<td>86</td>
<td>$13000</td>
<td>$6500</td>
<td>$6.00</td>
<td>$6.00</td>
</tr>
<tr>
<td>440xL (4 holds)</td>
<td>2 h</td>
<td>37000000</td>
<td>200000</td>
<td>0.04</td>
<td>86</td>
<td>$13000</td>
<td>$6500</td>
<td>$6.00</td>
<td>$6.00</td>
</tr>
<tr>
<td>440xL (2 holds)</td>
<td>2 h</td>
<td>37000000</td>
<td>200000</td>
<td>0.04</td>
<td>86</td>
<td>$13000</td>
<td>$6500</td>
<td>$6.00</td>
<td>$6.00</td>
</tr>
<tr>
<td>440xL (1 hold)</td>
<td>2 h</td>
<td>37000000</td>
<td>200000</td>
<td>0.04</td>
<td>86</td>
<td>$13000</td>
<td>$6500</td>
<td>$6.00</td>
<td>$6.00</td>
</tr>
<tr>
<td>440xL (0.5 holds)</td>
<td>2 h</td>
<td>37000000</td>
<td>200000</td>
<td>0.04</td>
<td>86</td>
<td>$13000</td>
<td>$6500</td>
<td>$6.00</td>
<td>$6.00</td>
</tr>
<tr>
<td>440xL (0.25 holds)</td>
<td>2 h</td>
<td>37000000</td>
<td>200000</td>
<td>0.04</td>
<td>86</td>
<td>$13000</td>
<td>$6500</td>
<td>$6.00</td>
<td>$6.00</td>
</tr>
<tr>
<td>440xL (0.125 holds)</td>
<td>2 h</td>
<td>37000000</td>
<td>200000</td>
<td>0.04</td>
<td>86</td>
<td>$13000</td>
<td>$6500</td>
<td>$6.00</td>
<td>$6.00</td>
</tr>
<tr>
<td>440xL (0.0625 holds)</td>
<td>2 h</td>
<td>37000000</td>
<td>200000</td>
<td>0.04</td>
<td>86</td>
<td>$13000</td>
<td>$6500</td>
<td>$6.00</td>
<td>$6.00</td>
</tr>
<tr>
<td>440xL (0.03125 holds)</td>
<td>2 h</td>
<td>37000000</td>
<td>200000</td>
<td>0.04</td>
<td>86</td>
<td>$13000</td>
<td>$6500</td>
<td>$6.00</td>
<td>$6.00</td>
</tr>
<tr>
<td>440xL (0.015625 holds)</td>
<td>2 h</td>
<td>37000000</td>
<td>200000</td>
<td>0.04</td>
<td>86</td>
<td>$13000</td>
<td>$6500</td>
<td>$6.00</td>
<td>$6.00</td>
</tr>
<tr>
<td>440xL (0.0078125 holds)</td>
<td>2 h</td>
<td>37000000</td>
<td>200000</td>
<td>0.04</td>
<td>86</td>
<td>$13000</td>
<td>$6500</td>
<td>$6.00</td>
<td>$6.00</td>
</tr>
</tbody>
</table>

FASTA format

- Seemingly trivial but it is also “under-specified”, there are many “custom” extensions
- Tools may make assumptions on the structure of a FASTA file
- Surprising number of problems can arise

Alphabets

- International Union of Pure and Applied Chemistry (IUPAC) codes
- Nucleic acid sequences
- Peptide sequences → polypeptides could be proteins
It is not clear what the sequence above contains nucleic acids or amino acids
(feels like a nucleic acids because of having so many ACTG both those are also valid amino acids)

First step of any sequence processing step understand your FASTA file

- What’s is what in this file, what does the fasta header look like?

1. How many sequences do we have
2. Are sequences all on a single line or over multiple lines
3. What is the identifier, what is embedded in the description
Creating scripts

You can run it with bash script-name.sh

Special variable names: $1, $2, $3 ... command line parameters

A more advanced version of the script

Comments, Why, what, where. Make them full coherent sentences.

Bash has lots of features

We will slowly introduce some features along the way. It is not a particularly good programming language. Don’t try to write overly complex programs. There are better solutions for that.

Homework 5

Write a script that selects the first 10 bases of every ebola virus assembly submitted to the PRJNA257197 project

1. How many sequences have you obtained?
2. How many unique sequence starts are there?
3. Are there small or big differences between the genome starts?
4. What is the most common sequence start?
5. Which one do you think is the most dissimilar sequence start compared to the most common one?