Sequencing Technologies - perspective

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

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

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

Random DNA fragment sequencing with Illumina

For each fragment $\rightarrow$ adapter ligation $\rightarrow$ separate by strands $\rightarrow$ some pieces get sequenced

NGS sequencing read formats

Reads: short sequences produced by the instrument

New instruments produce FastQ format (.fastq or .fq). Current platforms:

- HiSeq/MiSeq (Illumina)
- Ion Torrent (Thermo Fisher)
- PacBio (Pacific Biosciences)
- MinION (Oxford Nanopore)

Obsolete:

- Solid $\rightarrow$ colorspace fasta (.xsq or .csfasta + .qual)
- 454 $\rightarrow$ standard flowgram format (.sff)
Quality Control

Quality control is the first step of any data analysis.

This is where you first meet your data.

The first step is to identify problems then improve of them.

Bad data is like poison → it can turn even good data into bad

Compressed data formats

We commonly need to download data or install a new tool. You may also need to share your data with others.

You will need to be very familiar with how to deal with compressed files (archives)

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Run time</th>
<th>Million reads</th>
<th>Base per read</th>
<th>Yield Mb</th>
<th>Reagent cost</th>
<th>Reagent cost/Mb</th>
<th>Minimum unit cost (Mb)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3730xl (capillary)</td>
<td>2 h</td>
<td>0.850956</td>
<td>620</td>
<td>0.06</td>
<td>$96</td>
<td>$0.02</td>
<td>$0.17</td>
</tr>
<tr>
<td>Ion Torrent 314 chip</td>
<td>2 h</td>
<td>0.10</td>
<td>500</td>
<td>&gt;10</td>
<td>$90</td>
<td>$0.20</td>
<td>$0.18</td>
</tr>
<tr>
<td>454 FLX Titanium</td>
<td>10 h</td>
<td>0.10</td>
<td>400</td>
<td>50</td>
<td>$118</td>
<td>$0.22</td>
<td>$0.14</td>
</tr>
<tr>
<td>454 FLX</td>
<td>10 h</td>
<td>0.10</td>
<td>400</td>
<td>50</td>
<td>$118</td>
<td>$0.22</td>
<td>$0.14</td>
</tr>
<tr>
<td>Fluidigm</td>
<td>0.5-2 h</td>
<td>0.01</td>
<td>600-1000</td>
<td>5-10</td>
<td>$110-900</td>
<td>$0.1-0.90</td>
<td>$0.1-0.90</td>
</tr>
<tr>
<td>454 FLX Titanium</td>
<td>10 h</td>
<td>1</td>
<td>400</td>
<td>50</td>
<td>$280</td>
<td>$2.80</td>
<td>$0.18</td>
</tr>
<tr>
<td>Ion Torrent 314 chip</td>
<td>2 h</td>
<td>1</td>
<td>&gt;100</td>
<td>&gt;100</td>
<td>$76</td>
<td>&gt;0.76</td>
<td>&gt;0.100</td>
</tr>
<tr>
<td>HiSeq</td>
<td>N/A</td>
<td>50</td>
<td>20,000</td>
<td>N/A</td>
<td>NA</td>
<td>NA</td>
<td>$100-120 (100%)</td>
</tr>
<tr>
<td>Ion Torrent 314 chip</td>
<td>2 h</td>
<td>4-4</td>
<td>&gt;100</td>
<td>&gt;100</td>
<td>&gt;25</td>
<td>&gt;0.25</td>
<td>&gt;0.120 (100%)</td>
</tr>
<tr>
<td>Illumina MiSeq</td>
<td>26 h</td>
<td>5.4</td>
<td>500-1500</td>
<td>500</td>
<td>$750</td>
<td>$0.07</td>
<td>$0.150</td>
</tr>
<tr>
<td>Illumina HiSeq</td>
<td>8 days</td>
<td>250</td>
<td>500-1000</td>
<td>500</td>
<td>$10,200</td>
<td>$0.02</td>
<td>$0.103</td>
</tr>
<tr>
<td>SOLiD - 4</td>
<td>14 days</td>
<td>320</td>
<td>120-150</td>
<td>90</td>
<td>$11,320</td>
<td>$0.01</td>
<td>$0.150</td>
</tr>
<tr>
<td>Illumina HiSeq 1000</td>
<td>8 days</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
<td>$20,200</td>
<td>$0.01</td>
<td>$0.150</td>
</tr>
<tr>
<td>SOLiD - 500150</td>
<td>8 days</td>
<td>&gt;300</td>
<td>&gt;75</td>
<td>&gt;75</td>
<td>&gt;100</td>
<td>&gt;0.10</td>
<td>&gt;0.080</td>
</tr>
<tr>
<td>Illumina HiSeq 2000</td>
<td>8 days</td>
<td>&gt;300</td>
<td>&gt;75</td>
<td>&gt;75</td>
<td>&gt;100</td>
<td>&gt;0.10</td>
<td>&gt;0.080</td>
</tr>
<tr>
<td>SOLiD - 5000150</td>
<td>8 days</td>
<td>&gt;300</td>
<td>&gt;75</td>
<td>&gt;75</td>
<td>&gt;100</td>
<td>&gt;0.10</td>
<td>&gt;0.080</td>
</tr>
<tr>
<td>Illumina HiSeq 2000 - v2</td>
<td>10 days</td>
<td>&gt;5000</td>
<td>&gt;1000</td>
<td>&gt;1000</td>
<td>&gt;25,400</td>
<td>&gt;0.01</td>
<td>&gt;0.050</td>
</tr>
</tbody>
</table>

Compressed files/archives

Compressed formats → reduce the space requirements.

Same information but now optimized for size. Downside – needs to be decompressed to access the content.

1. Compressed File → a file reduced in size
2. Compressed Archive → multiple files combined then reduced in size

Important to remember:

Compression requires substantially more computational resources than decompression. So it is always much faster to restore file from compressed form.
Three major compression formats

- **ZIP** → .zip → zip/unzip
  Used if you keep seeing Windows people

- **GZIP** → .gz → gzip/gunzip
  The standard compression format, best tradeoff between speed vs compression

- **BZIP2** → .bz/.bz2 → bzip2/bunzip2
  Used by programming prima donnas: “Look how special I am, I’m even using a different compression format!”

- **Compress** → .Z → compress/uncompress
  Used by people that don’t know what they are doing

Archives

- **.zip** file → may contain one or multiple files or entire directory trees

- **.gz** → always is uncompressed to a single file

In the Unix world to compress multiple files you would need to create an archive then compress that archive.

Extensions: .tar.gz or .tgz

See the code archive for lecture 7 for details.

Creating archives with **tar** (tape archive)

The `tar` command can collect multiple files/directories into one file.

```
tar <commands> output-file <input files>
```

Commands: `create`, `extract`, `gzip`, `file`, `list`, `verbose`

```
tar cvf myfile.tar sample1.fq sample2.fq
```

Major annoyance: accidentally listing the file name as the archive name will destroy the file that you are trying to archive!
**tar can handle entire directories**

- Goes through and collects everything and packs it into one file, then compresses the file.
- Check the file extensions, it will tell you what it is.
- Sign of an inexperienced software developer.

**tar-bomb** → a tarball that contains the files without a directory name, the contents "explode" over your directory (see **tarbomb.tar.gz** on the website).

**Defensive measures:**
- List the content of the archive and make decision
- Create a new directory, expand the file there (this is what Mac/Windows does by default regardless what is in the archive).

---

**Quality control tools: FastQC (a great tool but with numerous flaws)**

FastQC

- **Function:** A quality control tool for high-throughput sequence data.
- **Language:** Java
- **Requirements:** A suitable Java Runtime Environment
- **Download:** [download](http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/)

**Per base quality scores**

- Averaging over columns.
- The average quality value at a given position.

**Per sequence quality score**

- Computes an average quality for each read.
- How many of a given average were found.
Sequence duplication: misleading plot

What does this plot actually mean: See Biostar threads (linked from the course page)

Adapter Content

Checks known contamination. You can add your own contaminants

Kmer enrichment

The plot can be greatly misleading. A longer motif would be counted multiple times.

Homework 6

- Get the sequencing platform data from the course website and unpack it. It contains data from four different sequencing platforms.
- For each file create a command line that counts the number of FASTQ records in that file.
- Generate a fastqc report on each of the datasets. The MinilON data run may fail with memory error. (We will fix this next time - or if you are feeling adventurous find the fastqc program, edit it as text and in that jungle of options change the amount of memory allocation from -Xm250m to -Xm750m)
- Now choose two sequence platforms, compare the attributes of the data that they produce – discuss the results.