Week 3, Lecture 6

István Albert

Bioinformatics Consulting Center
Penn State, 2014
NGS sequencing read formats

**Reads**: short sequences produced by the instrument

**Illumina** → FastQ format (.fastq or .fq)

**Solid** → colorspace fasta (.xsq or .csfasta + .qual)

**454** → standard flowgram format (.sff)
Random DNA fragment sequencing with Illumina

Fragmentation

For each fragment -> adapter ligation -> separate by strands -> some pieces get sequenced

Sequencer

Single end sequencing

sequencing direction
Extending the FASTA format

• The sequences are measurements

• There needs to be a way to associate quality measures to each base

• FASTQ ➔ .fq, .fastq (FASTA with qualities)
The structure of the FASTQ file

Four lines per FASTQ record

1. @ indicates the sequence identifier
2. The sequence content of the read
3. + optionally repeat the sequence id (often left empty)
4. Sequence quality string

An encoding is a transformation from one representation to another

- The information is not changed
- The optimization method changes

i.e: pig latin is a type of encoding
Ordinal (numerical) value of a character (ord)

```
ialbert@grit ~/edu/lec6
$ python -c 'print ord("A")'
65

ialbert@grit ~/edu/lec6
$ python -c 'print chr(65)'
A

ialbert@grit ~/edu/lec6
$ python -c 'print chr(ord("A"))'
A

ialbert@grit ~/edu/lec6
$ 
```
Encoding

One character $\rightarrow$ one byte space
ABCa = 4 bytes long
65 66 67 97 = 11 bytes long

**Good**: three characters are turned into one, saves space
**Bad**: not readable, hinders understanding
Remapping an encoding

• Problems: only some types of characters can be printed.

• So the encoding must start at a character that can be printed, that won’t be zero and it needs to represent zero

• Say character “A” has a code of 65. If we were to choose “A” as the minimum of our scale then we need to shift the scale by 65

```
ialbert@grit ~/edu/lec6
$ python -c 'print ord("A") - 65'
0
```

ialbert@grit ~/edu/lec6
$
Quality Scores

- A quality score is a **number** that usually has limits, a low (say 0) to a high (say 40)

- A quality score represents an error probability.

- It characterizes a single step of the process and the **NOT** the entire experimental procedure

- Quality scores are used to represent base calling accuracy, alignment accuracy and other probabilities
Connecting a quality score to a probability

For a quality score $Q$ the error probability is

$$P = 10^{-Q/10}$$

Examples:

- $Q = 10 \rightarrow P = 10^{-1} = 1/10 = 0.1 \Rightarrow P = 10\%$
- $Q = 40 \rightarrow P = 10^{-4} = 1/10000 = 0.0001 \Rightarrow P = 0.01\%$
There are multiple encodings: shifts

- Illumina used to switch around the encoding every once in a while.

- Finally they settled on the Sanger for encoding/Phred quality representation. Since 2011 or so.

- There are plenty of datasets/tools out there that may use different encodings!
Sanger Encoding (shift by 33)

- Quality Value range between 0 and 93
- Start the scale at character 33
- End the scale at character $33 + 93 = 126$

(currently most instruments only produce qualities in the range is 0 to 40)
Illumina 1.3 encoding (shift by 64) (obsolete but still often observed in the wild)

• Quality range between 0 to 62

• Start scale at character 64

• End scale at character 64 + 62 = 126
If you understand how to read this you’ll understand the FASTQ format.
Understanding encodings

```python
# the encoded letter that represents 
# the quality 10, (0.1% error rate)
# when shifted by the Sanger encoding (+33)
print chr( 10 + 33 )

# the quality value represented by the letter +
# when shifted by the Sanger encoding (+33)
print ord("+") - 33
```

```
$ python -c 'print ord("+") - 33'
10

$ python -c 'print chr(10 + 33)'
+
```
More information may be present

De-facto standard for producing sequencing reads. The vast majority of current tools expect this format.

Storing data in SRA removes the extra header information in the FASTQ record! That is unfortunate! Some information is now lost and available only to the original authors!
Illumina FASTQ header format

1. Instrument name: **HWI-ST1342** (unique for every sequencer)
2. Run id: **96**
3. Flowcell id: **H0NP9ADXX** (unique for every flowcell)
4. Flowcell lane: **2**
5. Tile number within the flowcell: **1115**
6. X-coordinate of the cluster in the tile: **13393**
7. Y-coordinate of the cluster in the tile: **59201**

More fields are may also be present (not shown above):
1. Mate pair 1 or 2
2. Flag: Y or N
... control bits, index sequences, usually defined in the Illumina manuals
Homework 6

• What characters in the Sanger encoding represent base calling error probabilities of:
  
  – 100%
  
  – 0.01%
  
  – 0.001%

• Create a Sanger encoded **FASTQ** file that a single record with the sequence **ATGC** and has the qualities of **40, 35, 36** and **32**