Quality Control for Raw Data

Francisco García García
Feb 29th 2016
Contents

- **Data formats**
  - Sequence capture
  - Fasta and fastq formats
  - Sequence quality encoding

- **Quality Control**
  - Evaluation of sequence quality
  - Quality control tools
  - Identification of artefacts & filtering
Contents

- **Data formats**
  - Sequence capture
  - Fasta and fastq formats
  - Sequence quality encoding

- **Quality Control**
  - Evaluation of sequence quality
  - Quality control tools
  - Identification of artifacts & filtering
Sequence Capture
Where are we?

Sequence preprocessing
  ↓
  Mapping
  ↓
  Variant calling
  ↓
  Variant prioritization

NGS pipeline
Contents

- **Data formats**
  - Sequence capture
  - **Fasta and fastq formats**
  - Sequence quality encoding

- **Quality Control**
  - Evaluation of sequence quality
  - Quality control tools
  - Identification of artifacts & filtering
From sequencers to digital data

- **What structure does the data have?**
  - Text-based formats (easy to use!)
  - If not compressed, it can be huge

- **Different data formats:**
  - Different sequencers output different files (sff, fasta, csfasta, qual file, fastq...)
  - There are some data formats widely accepted (e.g., FastQ format)
Fasta format

- Two lines per sequence:
  1. Header lines starts with ">" followed by a sequence ID
  2. Sequence (string of nt or peptides)

```text
>gi|5524211|gb|AAD44166.1| cytochrome b [Elephas maximus maximus]
LCLYTHIGRNIIYYGSLYSETWNTGIMLLHLLTMZAFMYVYLPGQMSFWGATVTNLFSAIPYIGTNLV
EWIWGGFSVNDKATLNRFFAFHFLPFTMTVALAGVHLTFLHGETGNNPLGLTSDSDKIPFHPYTYIKDFLG
LLILIIIIIIIIALLSPMLGDNPNDMPADPLNTPHLIKPEWYFLFAVAILRSVPNLGGLVIALFLSIVIL
GLMPFLHTSKHRSMMLRPLSQAFLWTLTMDLLTLTWIGSQPVEYPYTIIGQMASYLIFSIIAFLPIAGX
IENY
```

```text
>BBTBSCRYR
tgcaccaaaacatgtctaaagctggaaccacaattacttttttttgagaagacaaaaaactttca
agccgcctactatgacagqcattgcaactgtgcagatttcacatgtacctgagccgctg
caatcctcataagttggaaggagccacctgggctgtgtatgaaaggccavaattttgctgg
gtacatgtatcatctacctacccggggcagtgattcctgagttaccagcactggatggcctcaa
```

- Typical file extensions (.fasta, .fa, .fna, .fnn, .faa, ...)
Fastq format

- We could say “it is a fasta with **qualities**”:
  1. Header (like the fasta but starting with “@”)
  2. Sequence (string of nt)
  3. “+” and sequence ID (optional)
  4. Encoded quality of the sequence

```
@SEQ_ID
GATTTGGGGTTCAAAAGCAGTATCGATCAAATAGTAAATCCATTGTGTCAACTCACAGTTT
+
!'*(((**++))%%%%++)(%%%%).1**+-++')**)55CCF>>>>>>>>CCCCCCCC65
```
Quality codification

- Phred quality score
  - Error probability
  - ASCII encoded
- Phred +33
  - Sanger [0,40]
  - Illumina 1.8 [0,41]
  - Illumina 1.9 [0,41]
- Phred +64
  - Illumina 1.3 [0,40]
  - Illumina 1.5 [3,40]
Contents

- **Data formats**
  - Sequence capture
  - Fasta and fastq formats
  - Sequence quality encoding

- **Quality Control**
  - Evaluation of sequence quality
  - Quality control tools
  - Identification of artifacts & filtering
Sequence quality evaluation

- If we evaluate our sequence in depth we will know how reliable our results are.

- **Problem:** huge files $\rightarrow$ need of a tool to do it.
Sequence quality evaluation

- Quality control tools:
  - **Fastx-toolkit**
  - **NGS QC Toolkit**

http://hannonlab.cshl.edu/fastx_toolkit/download.html
http://www.nipgr.res.in/ngsqctoolkit.html
Other quality control tool: FastQC

http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
Sequence quality per base position

- **Good data**
  - Consistent
  - High quality along the read

- The central red line is the median value
- The yellow box represents the inter-quartile range (25-75%)
- The upper and lower whiskers represent the 10% and 90% points
- The blue line represents the mean quality
Per sequence quality distribution

Quality score distribution over all sequences

Low Quality Reads

Average Quality per read

Bad data
- Non-uniform distribution
Per base sequence content

- **Good data**
  - Smooth over length
  - Organism dependent (GC)
Per base sequence content

- **Bad data**
  - Sequence position bias
Per sequence GC content

- **Good data**
  - Fits with expected
  - Organism dependent

- **Bad data**
  - Does not fit with expected
  - Library contamination?
Per base N content

- **Good data**

- **Bad data**

  *Not good if there are N bias per base position*
Just descriptive:
- Some sequencers output sequences of different length (e.g. 454)
- In **Transcriptomics**, you expect higher number of duplicated sequences.
- In **Genomics** you should be worried if this happens → PCR artifact?
Question:
  • If we obtain the exact same sequences too many times → **Do we have a problem?**

Answer:
  • **Sometimes!**

Examples:
  • PCR primers, adapters ...
Typical artifacts

- Sequence adapters
Typical artifacts

- Platform dependent
Improving sequence quality

- Removing bad quality data will improve our confidence on downstream analysis
Improving sequence quality

- Sequence filtering
  - Mean quality
  - Read length
  - Read length after trimming
  - Percentage of bases above Q
  - Adapter trimming
  - Adapter reads
Improving sequence quality

- Sequence filtering tools
  - Fastx-toolkit
  - Galaxy (https://main.g2.bx.psu.edu/)
  - SeqTK (https://github.com/lh3/seqtk)
  - Cutadapt (http://code.google.com/p/cutadapt/)
  - Trimmomatic (http://www.usadellab.org/cms/?page=trimmomatic)
  - And more....
Improving sequence quality

Working plan:

1) Use **FastQC** to evaluate quality of sequences.

2) Use **Fastx-toolkit** to optimize different datasets and then visualize the result with **FastQC** to prove your success!

You can work from other more tools to filter sequences.
Installing FastQC

http://www.bioinformatics.babraham.ac.uk/projects/download.html#fastqc

**FastQC** A quality control application for high throughput sequence data

- [README](#)
- [Installation and setup instructions](#)
- [Release Notes](#) Please read these before using the program.
- [FastQC v0.10.1 (Win/Linux zip file)](#)
- [FastQC v0.10.1 (Mac DMG image)](#)
- [Source Code for FastQC v0.10.1 (zip file)](#)
Final remarks

- After preprocessing sequences, it is important to **evaluate the quality for raw data**

- **Fastq is the standard format** for NGS raw data. This format includes a quality score for each position.

- **NGS Genomic Data Analysis Pipeline** produces a **control quality report** to control quality for sequences.
Any questions?