Preprocess Quality

100 Amerindian Genome Project

Cristóbal Fresno Joshua Haase

2017-05-08

Experiment

- 95 samples
    - paired end reads
    - length 150bp
    - experimental protocol?
    - sequencer?
- fastq
    - raw data
    - bgi
        - remove adapters
        - drop reads with 10% N
        - drop reads with $Q_{average} < 18$
    - inmegen
        - remove adapters
        - remove bases with $Q < 28$ from begining
        - trim using sliding window of size 5 where $Q_{average} < 28$
        - drop reads where $length < 70$

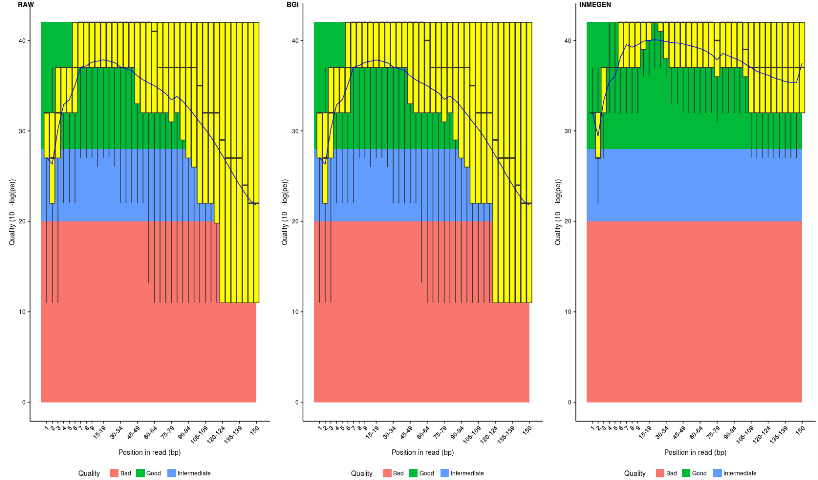

Figure 1: Quality per base summary.

Not the best quality. BGI preprocess too permisive.

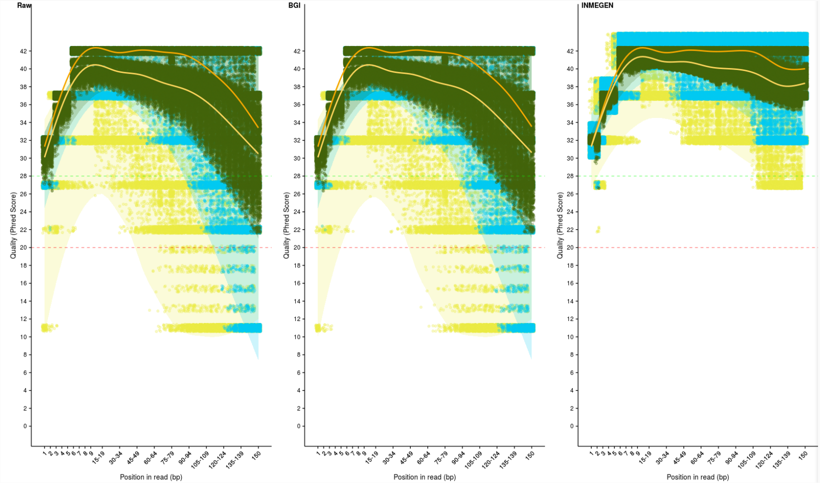

Figure 2: Quality per base detailed.

Experimental bias shown. Potencially from sequencer and/or flowcell.

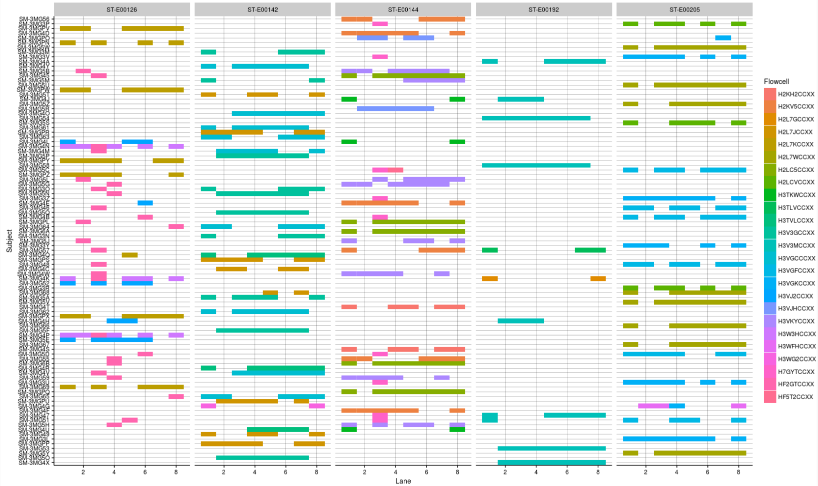

Figure 3: Experimental design.

Not random at all!! Subjects confused with flowcell and sequencer.

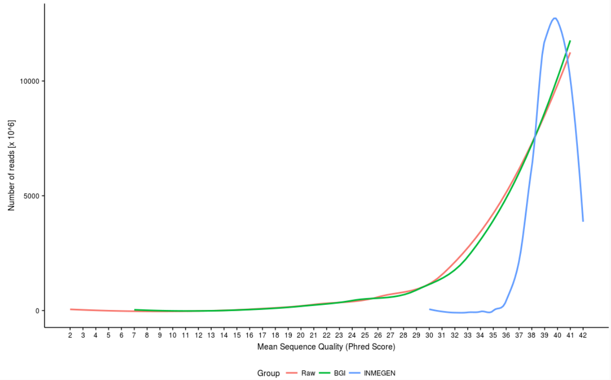

Figure 4: Read density per quality.

Almost no difference between RAW and BGI. Clear quality improvement with our preprocessing.

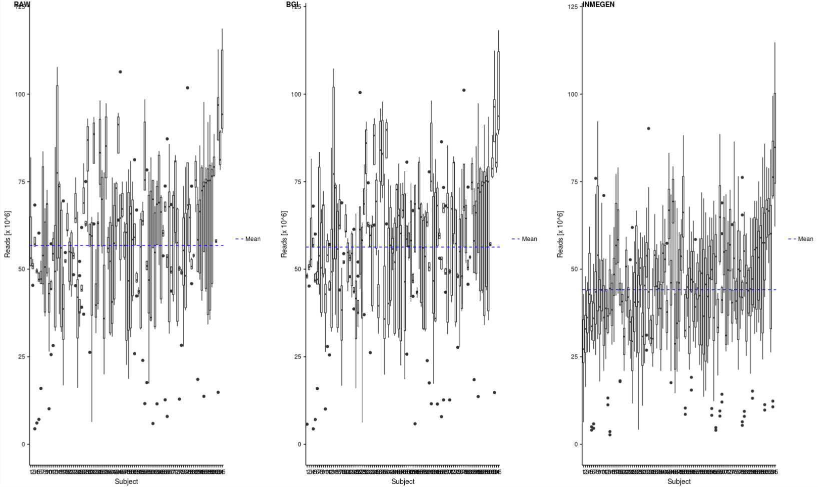

Figure 5: Reads per subject.

~ 85% of reads kept.

Next steps

Next steps

- Read alignment for HG38.
  - Pilot test with SNAP → GATK.
  - Alignment with BWA MEM.
- Alignment Quality Control using qualimap.
- Variant calling.
  - GATK variant anotation.
  - Phasing using 1000 genome reference.
  - Comparison with microarrays.
- Structural variant search.
- Positive selection.