Quality Control for Mapped Sequences
Introduction

SAMStat

QualiMap

Conclusions

Fco. Javier López

Quality Control for Mapped Sequences
Introduction

SAMStat

QualiMap

Conclusions
Where are we?

- **NGS pipeline**
  - Sequence preprocessing
    - Mapping
    - Variant Calling
    - Variant prioritization
      - Functional annotation
      - GWAS Analysis
        - Gene-Set Analysis

Fco. Javier López

Quality Control for Mapped Sequences
Why QC on mapped sequences

Acknowledgment: Fernando García Alcalde

- The reads **may look OK** in QC analyses of raw reads but some **issues** only show up **once the reads are aligned**: low coverage, homopolymer biases, experimental artifacts, etc.

- These unwanted biases can be introduced by the selected:
  - Sample extraction process
  - Sequencing technology
  - Sample preparation protocol
  - Mapping algorithm
Why QC on mapped sequences

- SAM/BAM files usually contain information from tens to hundreds of millions of reads
- The **systematic detection** of such biases is a **non-trivial** task that is **crucial** to drive appropriate downstream analyses.
- Look for big biases that really affect the analysis
- Difficult to provide guidelines: general trends
Introduction

SAMStat

QualiMap

Conclusions
Features

- Facilitates the identification of sequencing error biases that may disturb the mapping process
- Provides a concise html page with statistics that highlight problems in the data processing:
  - Reads with an excessive proportion of mapping errors
  - Reads containing contaminants
  - Reads representing novel splice junctions/genomic regions
  - ...
- Easy-to-use command-line tool freely downloadable at:
  
  http://samstat.sourceforge.net

Timo Lassmann, Yoshihide Hayashizaki, and Carsten O. Daub. SAMStat: monitoring biases in next generation sequencing data Bioinformatics (2011) 27(1): 130-131
Running SAMStat

- Input: a BAM/SAM file (other sequence files are also accepted such as fasta or fastq)
- Output: an html report

Run SAMStat with a .bam example

samstat /home/biouser/mda13/mqc-igv/test1.bam

- The html report will be saved at /home/biouser/mda13/mqc-igv/test1.bam.html. Use a web browser (e.g. Firefox) to open it
Concepts

- Mapping quality: an integer in $[0, 254]$ representing $-10 \cdot \log_{10} P(\text{mapping error})$
- Calculated as a function of the quality of the read, and a score that indicates how well the read is aligned
- Algorithm-specific
- The higher it is, the better the alignment. ($\text{MAPQ} = 30 \implies 0.001$ error rate)
- 255 indicates that the mapping quality is not available.
Number of aligned reads and mapping quality

- Proportion of reads mapped in each mapping quality range.
- The “red part” should fill most of the pie chart area.
Number of aligned reads and mapping quality

- Proportion of reads mapped in each mapping quality range.
- The “red part” should fill most of the pie chart area
- WARNING: The appearance of a 0% of unmapped reads does not necessarily mean that there all the raw reads were aligned.
Number of aligned reads and mapping quality

- Proportion of reads mapped in each mapping quality range.
- The “red part” should fill most of the pie chart area.
- WARNING: The appearance of a 0% of unmapped reads does not necessarily mean that there all the raw reads were aligned.
- Why?
Mean base quality

- Mean quality per read base in each mapping quality range
- Higher the base quality $\Rightarrow$ higher mapping quality expected.
SAMStat report

Mean base quality

▶ Mean quality per read base in each mapping quality range
▶ Higher the base quality $\Rightarrow$ higher mapping quality expected.

Error profiles

▶ Number of mismatches at each read position, segregated by the nucleotide causing the mismatch
▶ Should be more or less stable across read positions
▶ More errors are expected at the end of the reads since base qualities tend to be lower at that positions
▶ Nucleotide peaks at different positions may indicate experimental artifacts that disturb read mapping
Over-represented di-nucleotides

- Over-representation scores for each possible di-nucleotide at each read position.
- Significant scores (p-value $\leq 1e^{-100}$) appear in bold.
- Over-represented di-nucleotides may indicate experimental artifacts that disturb read mapping.
Over-represented di-nucleotides

- Over-representation scores for each possible di-nucleotide at each read position.
- Significant scores (p-value \(\leq 1e^{-100}\)) appear in bold.
- Over-represented di-nucleotides may indicate experimental artifacts that disturb read mapping.

Error distribution

- Distribution of the number of errors (mismatches and indels) per read, segregated by mapping quality ranges.
- No more than \(\sim 2\) mismatches should be allowed for short (\(\sim 75b\)) reads.
Nucleotide composition

- Number of As, Cs, Gs and Ts appearing at each read position and segregated by mapping quality
- The counts and proportions should be almost invariant across read positions
Nucleotide composition

- Number of As, Cs, Gs and Ts appearing at each read position and segregated by mapping quality
- The counts and proportions should be almost invariant across read positions

Length distribution

- Distribution of the number of bases per read
Top 5 over-represented 2-mers

- Summary of the “Over-represented di-nucleotides”, including the top-5 2-mers in each position
Top 5 over-represented 2-mers

- Summary of the “Over-represented di-nucleotides”, including the top-5 2-mers in each position

Top 20 over-represented 10-mers

- The 20 most significant 10-mers per quality level
More on SAMStat

Hands-on

- Run SAMStat on
  /home/biouser/mda13/mqc-igv/test2.bam and
  /home/biouser/mda13/mqc-igv/test3.bam

- Interpret the results
Introduction

SAMStat

QualiMap

Conclusions
Qualimap

Aim
Provide an overall view of the data that helps to detect biases in the sequencing and/or mapping of the data

Run QualiMap

qualimap

▶ BAM file needs to be sorted: samtools sort <filename> <fileout>

▶ File → New analysis → BAM/SAM file → /home/biouser/mda13/mqc-igv/HG00096.chrom20.bam

Features

- Fast analysis across the reference of genome coverage and nucleotide distribution
- Easy to interpret summary of the main properties of the alignment data
- Analysis of the reads mapped inside/outside of the regions provided in GFF format
- Insert size mean and median value calculation and plotting statistical distribution
- Analysis of the adequacy of the sequencing depth in RNA-seq experiments
- Clustering of epigenomic profiles

Quality Control for Mapped Sequences
Hands on

▶ Open the online help
▶ Go through the examples
▶ Run qualimap in the previous data (with and without reference annotation http://reports.bioinfomgp.org/external-downloads/refseqgenes.gtf)
▶ Drive conclusions from what you get
▶ BONUS: Run qualimap via de command line
Introduction

SAMStat

QualiMap

Conclusions
Conclusions

- One should always perform QC on the mapped data

The correct interpretation of the QC output may save a lot of time (and money) on downstream analyses.

The expected results are experiment-specific ⇒ Learn from experience
Conclusions

- One should always perform QC on the mapped data
- The correct interpretation of the QC output may save a lot of time (and money) on downstream analyses
Conclusions

▶ One should always perform QC on the mapped data

▶ The correct interpretation of the QC output may save a lot of time (and money) on downstream analyses

▶ The expected results are experiment-specific ⇒ Learn from experience