Konstantin Okonechnikov

Robust quality control of Next Generation Sequencing alignment data

Max Planck Institute For Infection Biology
Molecular Biology department
Bioinformatics Unit

BOSC 2013 Berlin
Next Generation Sequencing: amazing discovery tool

High speed & and deep coverage

Technologies:
• Whole genome or exome
• Whole transcriptome (RNA-seq)
• Histone modifications (ChIP-seq)
• Much more ...

Various applications

Fast developing

But...
... there are some things to take into account

- Small read size vs repetitive genome
- High price (even now up to 3K $)
- Computational requirements:
  - Sophisticated bioinformatics analysis
  - Storage and computational performance
- Platform specific and protocol errors
  - Optical duplicates, PCR artifacts...
- Algorithm induced biases
Quality Control of NGS data

The systematic detection of the biases is crucial -> saves time and money

Some packages exist:

- FastQC
- Samtools
- Picard tools
- RNA-seq QC

However there is room for improvement: more comprehensive and user-friendly tool could be useful.
Quality Control of NGS data

Our solution:

![QualiMap](image)

A Java application, which allows computing statistics and presenting different graphs for the evaluation of NGS alignment data.

Provides both GUI and command line interfaces
Qualimap features

- Supported types of experiments: WG-seq, RNA-seq, exome seq, methylation studies, ...
- 3 modes of analysis: **BAM QC**, Counts QC, RNA-seq QC
- Analysis possible for whole alignment or for arbitrary regions
- Input:
  - BAM/SAM alignment
  - GFF/GTF/BED annotations
- Output:
  - Interactive visualization
  - PDF or HTML report
Qualimap features: BAM QC

Summary:
- Global data (reference size, number of reads)
- Coverage (mapped, paired, per chromosome)
- Reads info (insert size, quality, homopolymers, duplication rate)
Qualimap features: BAM QC

- Coverage across reference
- Genome Fraction Coverage
- Insert Size Distribution
- Insert Size Histogram
- GC content (%)
Qualimap features: Counts QC

- 2 samples comparison
- Sequencing saturation
- Feature by biotype classification
Qualimap features: RNA-seq QC

- Transcript coverage
- 5‘-3‘ bias calculation
Some more features

• Tools: counts computation, epigenetic clustering

• Command line interface: easy integration

• Performance: runs in parallel on multicore systems

• Manuscript:
Further development

• New features are suggested by users

• Discussion forum: google-groups

• Source code on bitbucket

• Early builds are available as snapshots

• There is a Galaxy wrapper developed by Joachim Jacob available from the Galaxy Tool Shed
Thank you for attention!
Please provide your questions.

Useful links:

• Web-site:  http://qualimap.bioinfo.cipf.es/

• Bitbucket:  https://bitbucket.org/kokonech/qualimap

• Google-groups:  http://groups.google.com/group/qualimap

• Galaxy repo:  http://toolshed.g2.bx.psu.edu/repos/joachim-jacob/qualimap_suite