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...





MAX-PLANCK-GESELLSCHAFT

... there are some things to take into account





-Algorithm induced biases

p123 p123	• Cr#2016.	p11.23 p11.2	- all	911.22	-	al2	E 2	2 0131		91331	913.33
, FE FE	073,040 by		62,873,050	10 Бр — Ме			62,471	868 by			62,472,479
			0.0.0		000				0.000	ri mt mi	
							1				
								<u>p</u> i	1.1.1		
					Шî.	i i i			i ini ini		





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:



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 regions¹⁰
- 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)

🖄 BAM QC: saliva.sorted.bam 🗴	🖄 BAM QC: ERR089819.bam 🗴 🖄 Counts QC	C: Data Analysis 🗴							
⊘ Results Summa Input	Summary								
 Coverage across reference Coverage Histogram Coverage Histogram (0-50X) Duplication Rate Histogram Genome Fraction Coverage Mapped Reads Nucleotide Content Mapped Reads GC-content Distributi Mapping Quality Across Reference Mapping Quality Histogram Insert Size Across Reference Insert Size Histogram 	Globals Reference size Number of reads Number/percentage of mapped reads Number/percentage of unmapped reads Number/percentage of paired reads Number/percentage of reads both mates paired Number/percentage of singletons	100,286,002 35,576,180 30,983,200 / 87.09% 4,592,980/12.91% 30,983,200/87.09% 30,983,200/87.09%							
	Read min/max/mean length ACGT Content Number/percentage of A's Number/percentage of C's	100/100/100 1,003,585,776 / 32.43% 541 692 060 / 17 5%							
	Number/percentage of T's Number/percentage of G's Number/percentage of N's GC Percentage	1,006,351,599 / 32.52% 543,136,344 / 17.55% 0 / 0% 35.05%							
	Coverage Mean Standard Deviation Mapping Quality	30.86 25.01							

AX-PLANCK-GESELLSCHAI

Qualimap features: BAM QC



Coverage (X)



tut

Qualimap features: Counts QC

- 2 samples comparison
- Sequencing saturation
- Feature by biotype classification



Qualimap features: RNA-seq QC

- Transcript coverage
- 5'-3' bias calculation



-S しり Infekti



Some more features

- Tools: counts computation, epigenetic clustering
- Command line interface: easy integration
- Performance: runs in parallel on multicore systems
- Manuscript:

García-Alcalde F, Okonechnikov K, Carbonell J, Cruz L, Götz S, Tarazona S, Dopazo J, Meyer T, Conesa A. "Qualimap: evaluating next-generation sequencing alignment data." Bioinformatics 28, no. 20 (2012): 2678-2679.



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: <u>https://bitbucket.org/kokonech/qualimap</u>
- Google-groups: <u>http://groups.google.com/group/qualimap</u>
- Galaxy repo: <u>http://toolshed.g2.bx.psu.edu/repos/joachim-jacob/qualimap_suite</u>



