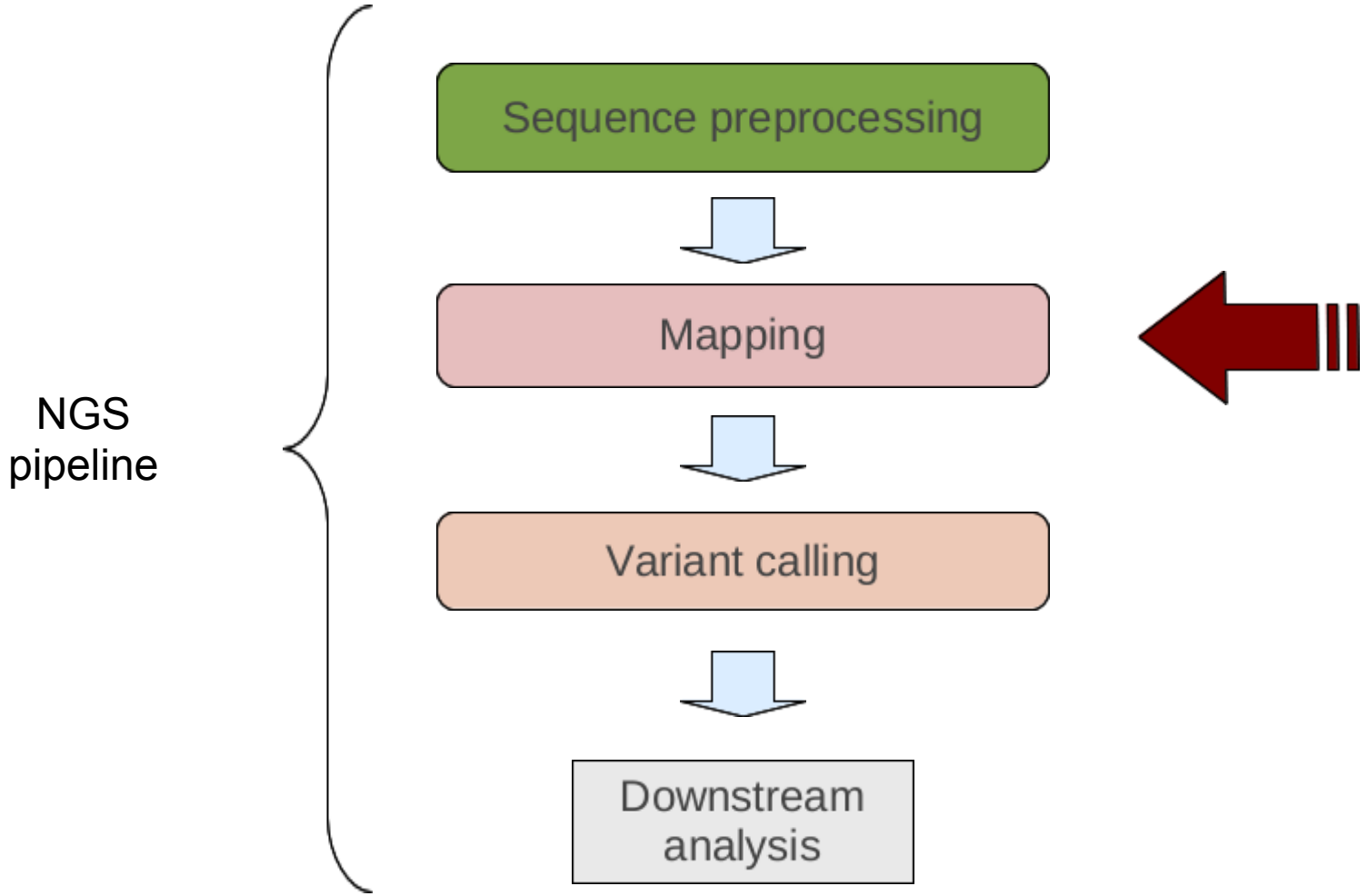
Sequence alignment

Enrique Vidal
evidal@cipf.es

Roadmap


NGS
pipeline
Sequence preprocessing
Mapping
Variant calling
Downstream
analysis

Index

- Intro
- Reference genome
- Aligners
- Results
- Quality Control
- Visualization
- Extra
- Computing needs

... why?

- We have a bunch of (HQ) reads
- We want to make some sense out of them
- Reconstruct the puzzle!

Reference genome: What?

- Reference as a compound
- Guide to compare samples ...
- ... or comparison against compound

Reference genome: Who? Where?

- Genome Reference Consortium
- Current Assembly: GRCh37
- Next Assembly: July 2013

Tools: How?

- BFAST
- Bowtie
- BWA

BWA

- Index reference genome
  - Suffix array
  - Burrows-Wheeler transform
- Backward search
- Mismatches, gaps
- Sed region

Coming soon ...

- High Performance Genomics - aligner
- CIPF (Computational Biology Unit)
- Faster & More accurate
- Flexible (DNA, RNA, BS-DNA, ...)

Result: Overall

- Mean coverage
- Mapping efficiency
- Mean mapping quality
- Single and multiple hits

Result: How?

- SAM/BAM format (standard!)
- Information about:
  - Genomic position(chromosome, position, strand)
  - Reference agreement (mapping quality, mismatches)

SAM: fixed fields

QNAME	read name
FLAG	bitwise flag (http://picard.sourceforge.net/explain-flags.html)
RNAME	chromosome
POS	leftmost genomic position
MAPQ	mapping quality
CIGAR	CIGAR string (gaps, clipping)
RNEXT	paired read name
PNEXT	paired read position
TLEN	total length of template
SEQ	read base sequence
QUAL	read base quality

SAM: fixed fields

:185815#6@ 83	20	68307 60	76M	=	68245 -138	TTCTGTATTC...	BEJHGD@GEL...
:185815#6@ 163	20	68245 60	76M	=	68307 138	TCTGGTTCAT...	DEFIECCDCD...
:111763#2@ 99	20	68246 60	76M	=	68315 145	CTGGTTCATC...	BGCFCCEEEI...
:111763#2@ 147	20	68315 60	76M	=	68246 -145	TCCTCAGGAC...	##########...
:182649#6@ 83	20	68320 60	76M	=	68249 -147	AGGACACAGA...	EEDCGD?0D=...
:182649#6@ 163	20	68249 60	76M	=	68320 147	GTTCATCACC...	EFDGEDGFGG...
:164917#2@ 99	20	68254 60	76M	=	68323 145	NCACCCATGA...	#4@ADEHFJF...
:164917#2@ 147	20	68323 60	76M	=	68254 -145	ACACAGAGCT...	##########...
:182092#2@ 99	20	68263 60	76M	=	68328 141	ATAGACCAGT..	FCDHFIHHJC...

SAM: optional fields

NM	Edit distance
MD	Mismatching positions/bases
AS	Alignment score
BC	Barcode sequence
X0	Number of best hits
X1	Number of suboptimal hits found by BWA
XN	Number of ambiguous bases in the referenece
XM	Number of mismatches in the alignment
XO	Number of gap opens
XG	Number of gap extentions
XT	Type: Unique/Repeat/N/Mate-sw
XA	Alternative hits; format: (chr,pos,CIGAR,NM;)*
XS	Suboptimal alignment score
XF	Support from forward/reverse alignment
XE	Number of supporting seeds

SAM: optional fields

XT:A:U	NM:i:0	SM:i:37	AM:i:37	X0:i:1	X1:i:0	XM:i:0	XO:i:0	XG:i:0	MD:Z:76
XT:A:U	NM:i:0	SM:i:37	AM:i:37	X0:i:1	X1:i:0	XM:i:0	XO:i:0	XG:i:0	MD:Z:76
XT:A:U	NM:i:0	SM:i:37	AM:i:37	X0:i:1	X1:i:0	XM:i:0	XO:i:0	XG:i:0	MD:Z:76
XT:A:U	NM:i:0	SM:i:37	AM:i:37	X0:i:1	X1:i:0	XM:i:0	XO:i:0	XG:i:0	MD:Z:76
XT:A:U	NM:i:0	SM:i:37	AM:i:37	X0:i:1	X1:i:0	XM:i:0	XO:i:0	XG:i:0	MD:Z:76
XT:A:U	NM:i:0	SM:i:37	AM:i:37	X0:i:1	X1:i:0	XM:i:0	XO:i:0	XG:i:0	MD:Z:76
XT:A:U	NM:i:1	SM:i:37	AM:i:37	X0:i:1	X1:i:0	XM:i:1	XO:i:0	XG:i:0	MD:Z:0T75
XT:A:U	NM:i:0	SM:i:37	AM:i:37	X0:i:1	X1:i:0	XM:i:0	XO:i:0	XG:i:0	MD:Z:76
XT:A:U	NM:i:0	SM:i:37	AM:i:37	X0:i:1	X1:i:0	XM:i:0	XO:i:0	XG:i:0	MD:Z:76

Extra format: BED

- Another standard!
- Genomic regions
- chromosome, start, end, whatever

File handling

samtools (http://samtools.sourceforge.net/)

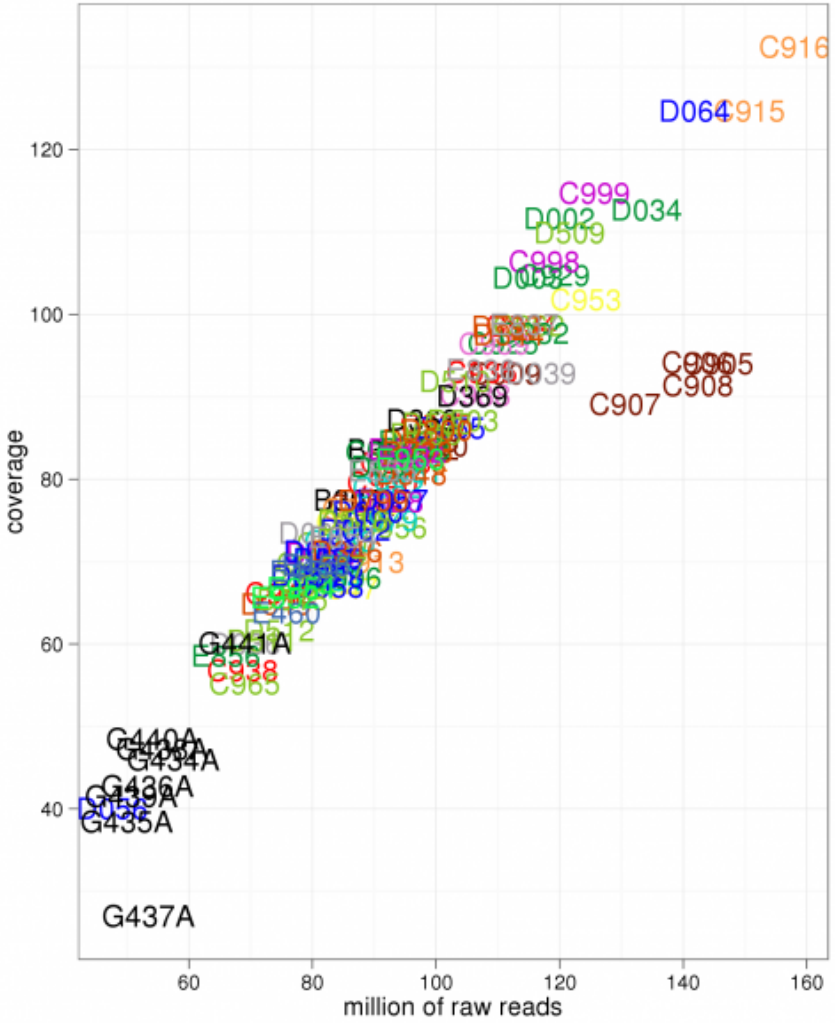
- very powerful and flexible :)
- command-line :(

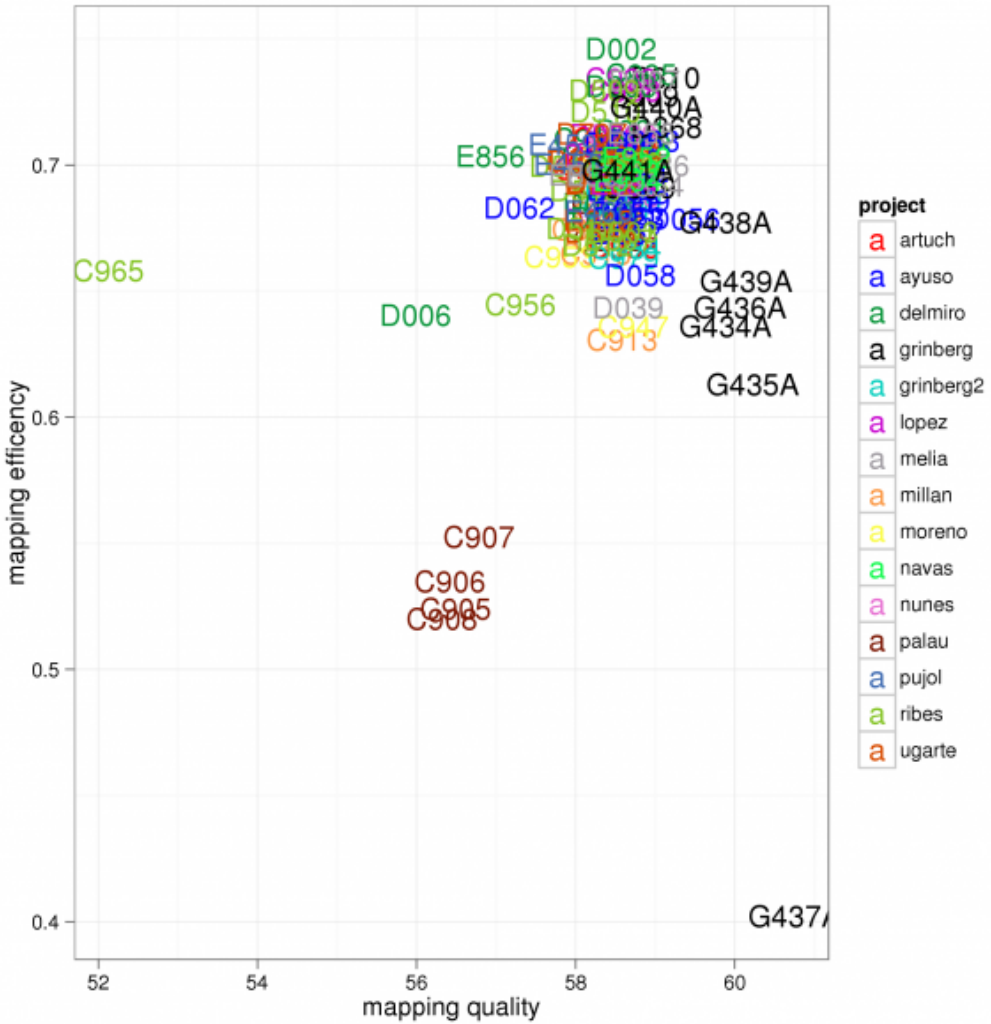
File handling

bedtools (http://code.google.com/p/bedtools/)

- very powerful and flexible :)
- command-line :(

So far BIER results



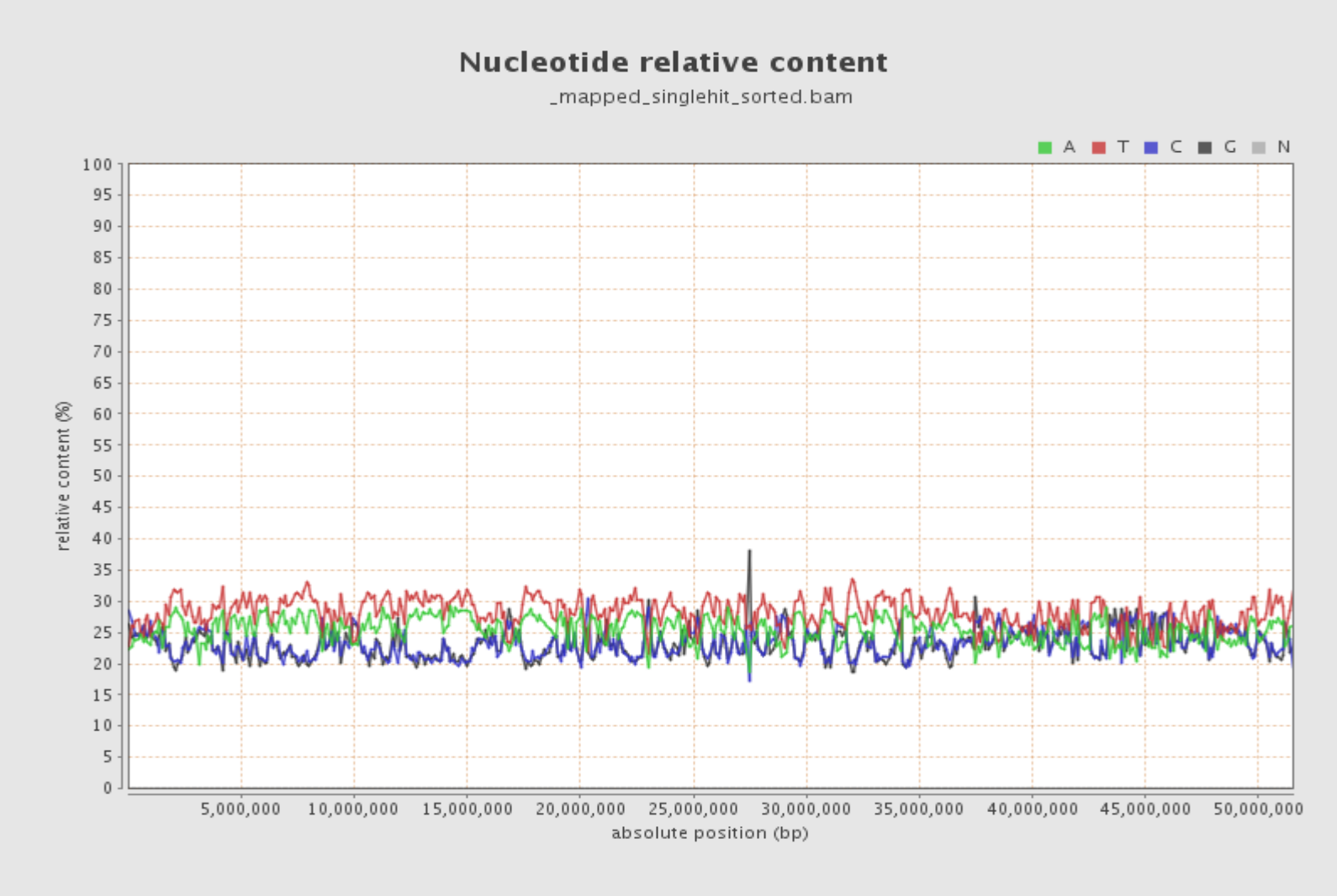


Quality Control

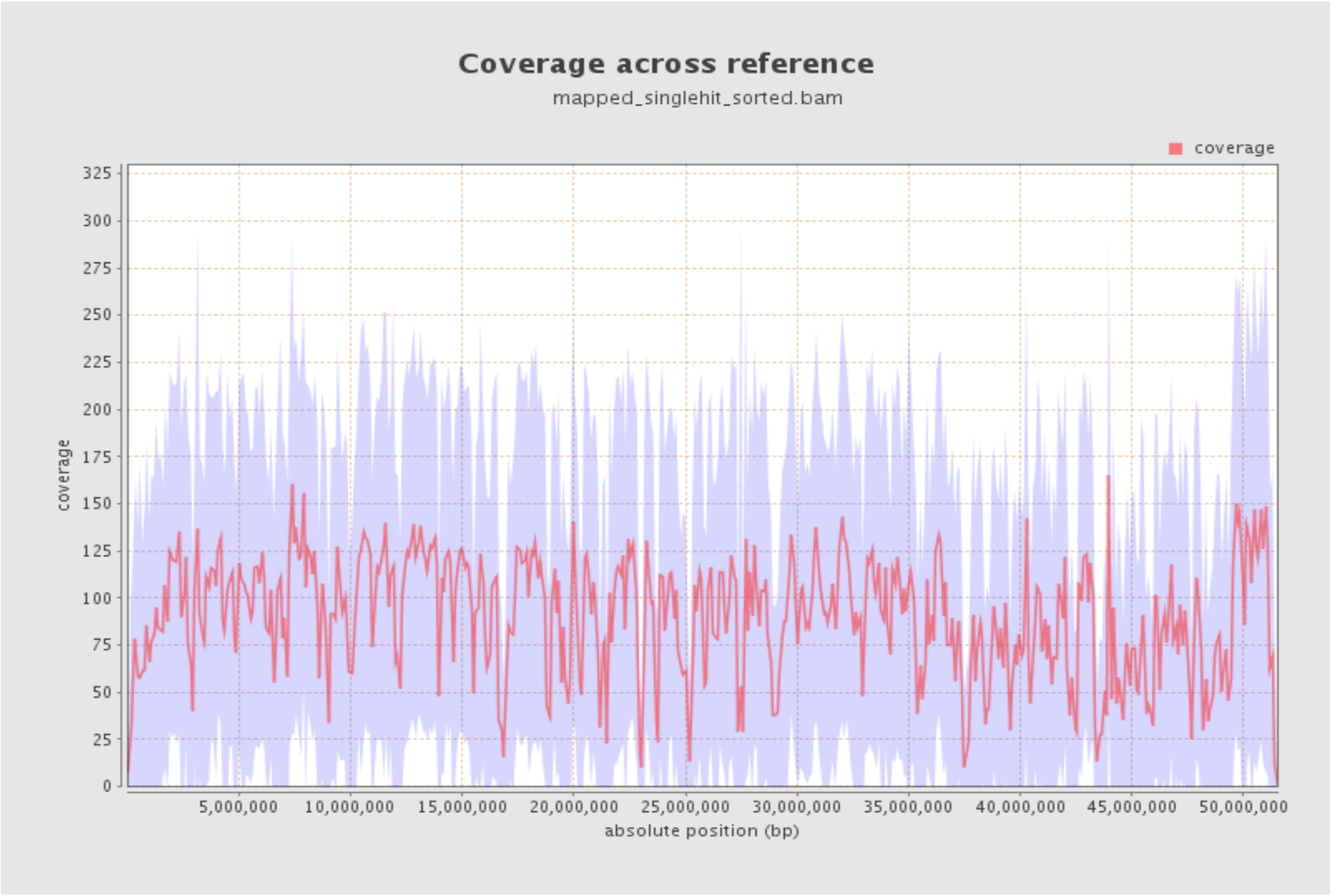
QualiMap

http://qualimap.bioinfo.cipf.es/

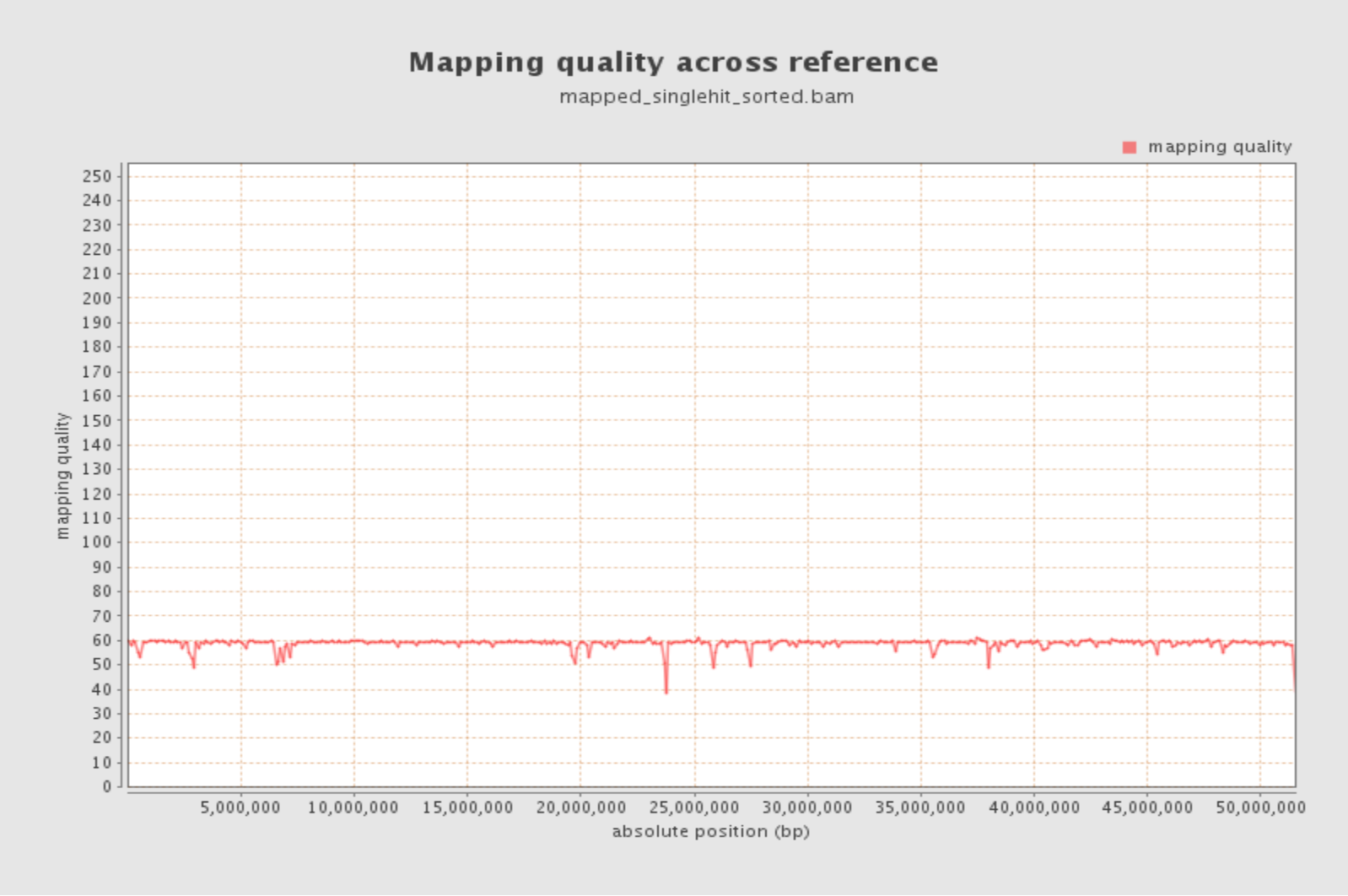
Quality control



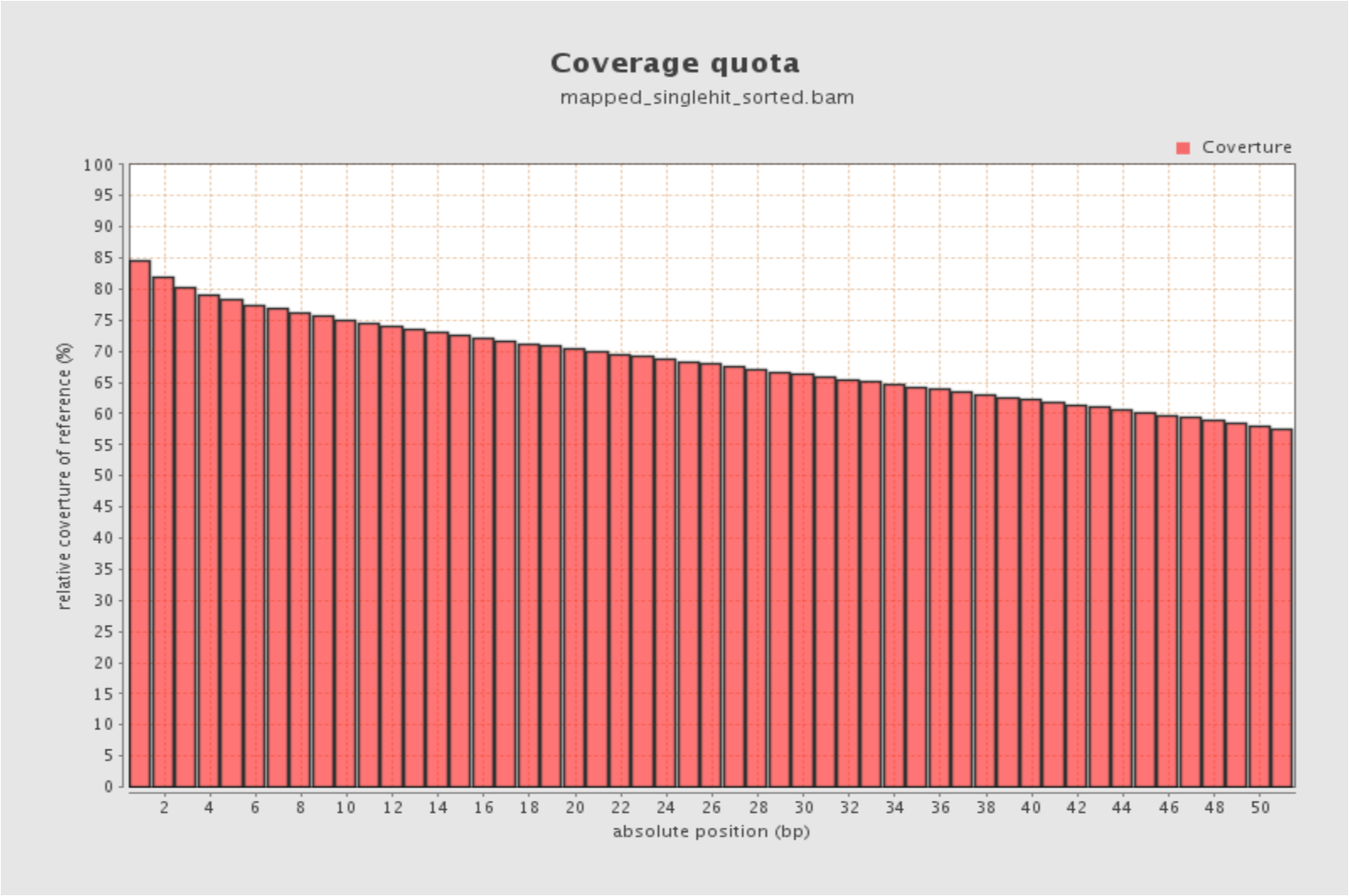
Quality control



Quality control



Quality control



Visualization

IGV

http://www.broadinstitute.org/igv/

Extra: Local realignment

Problematic loci (SNPs, indels, mismatches)

Redo the reference (haplotype)

- Realign
- Check
- Correct

Extra: Base quality recalibration

Empirical mismatches (dbSNP)

Correct for error covariates

Computational needs

- Machine: processor, RAM and disk
- Time
  - ~ 20 h/sample
- Errors!

THE END