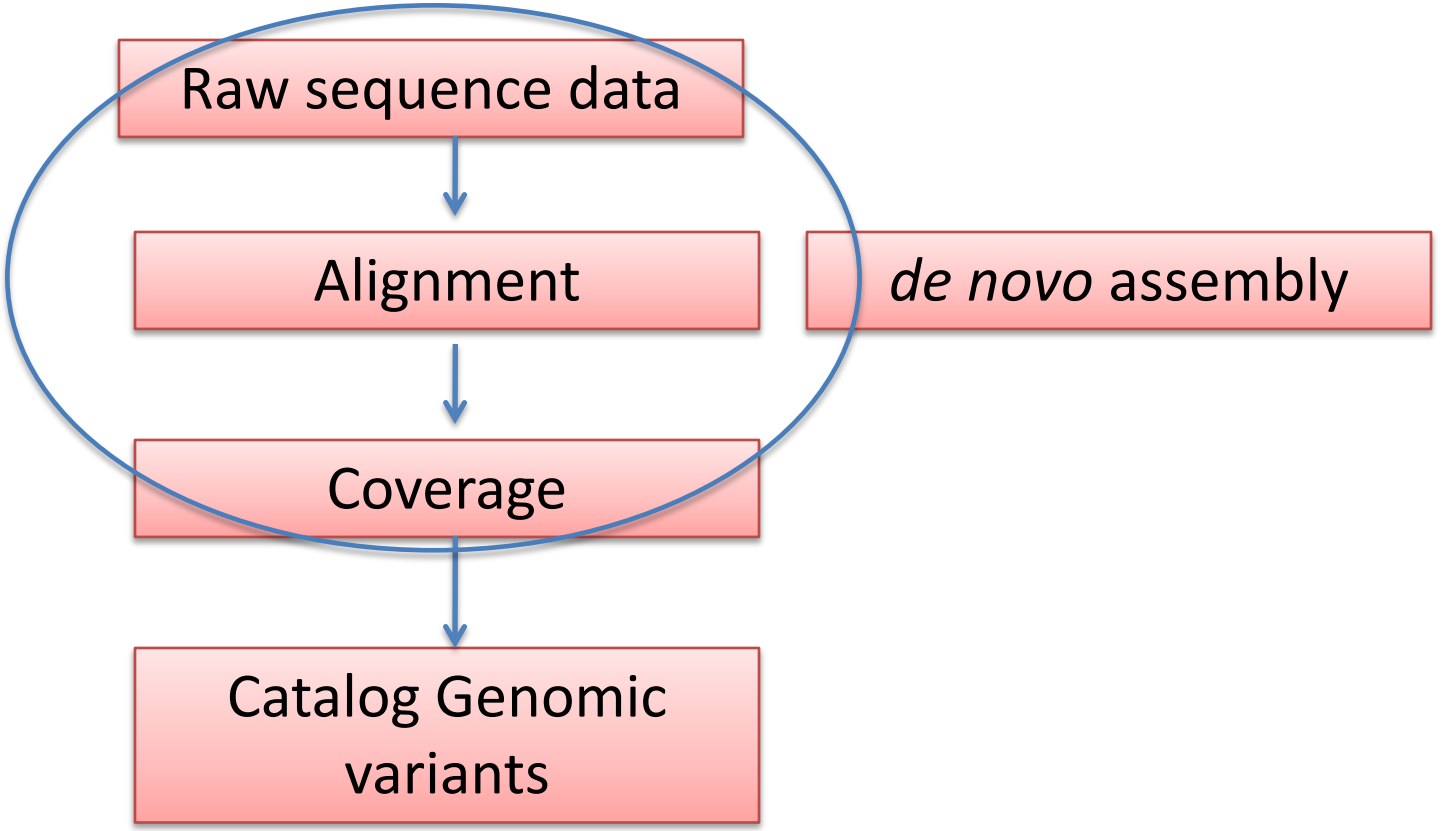
Sequence mapping and visualisation

London School of Hygiene and Tropical Medicine

Some aspects of the course



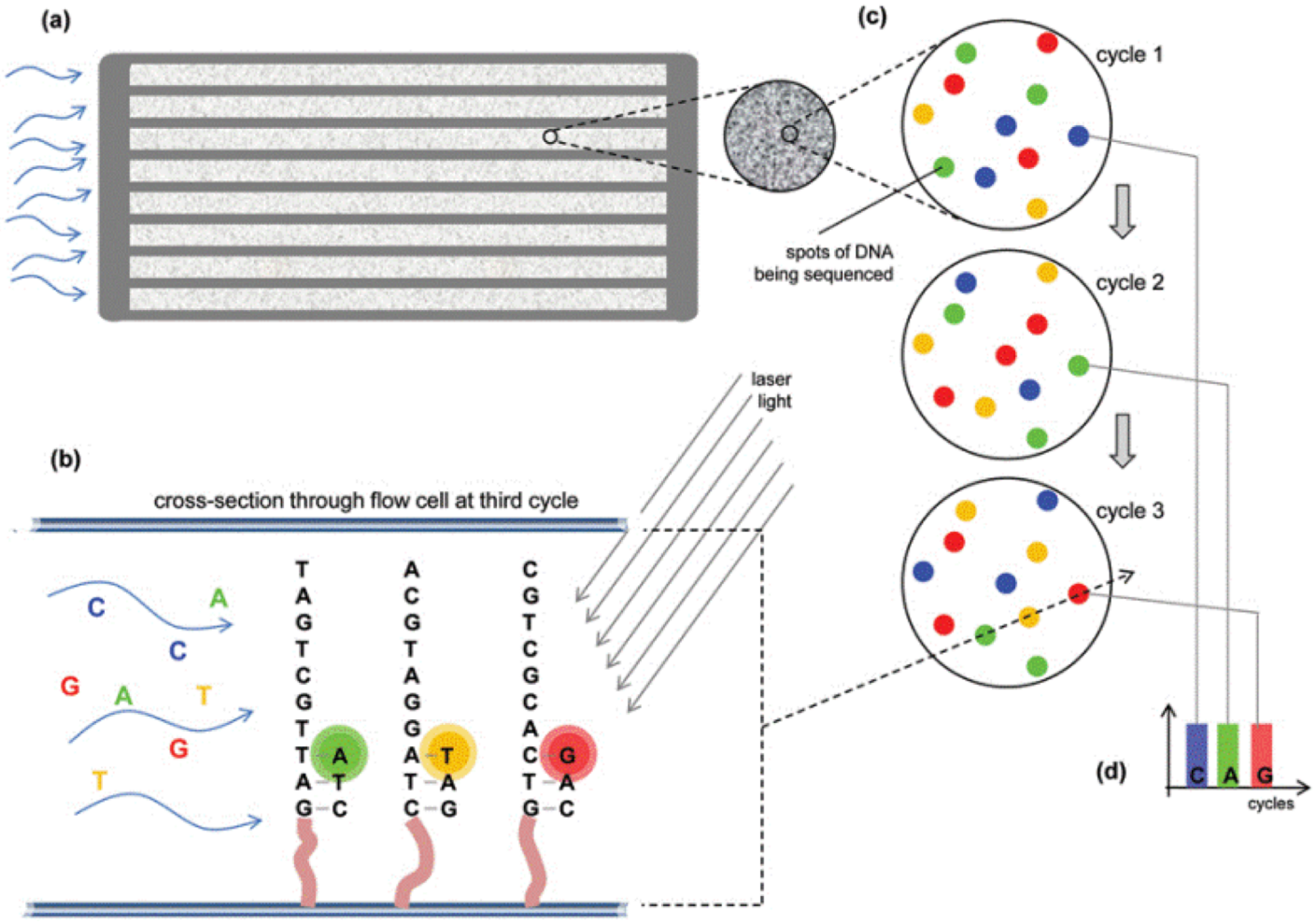
Population genetics

Whole genome Association studies

Outline

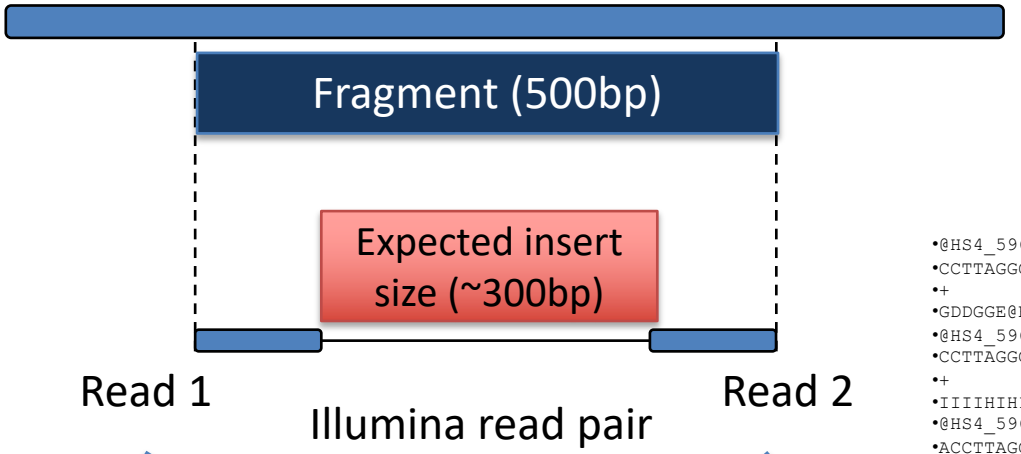
- Mapping to a reference genome
- Assessing the quality of the alignment
- Practical

Illumina Sequencing



http://www.nimr.mrc.ac.uk/mill-hill-essays/bringing-it-all-back-home-next-generation-sequencing-technology-and-you

Raw sequence data (Illumina) – “fastq”



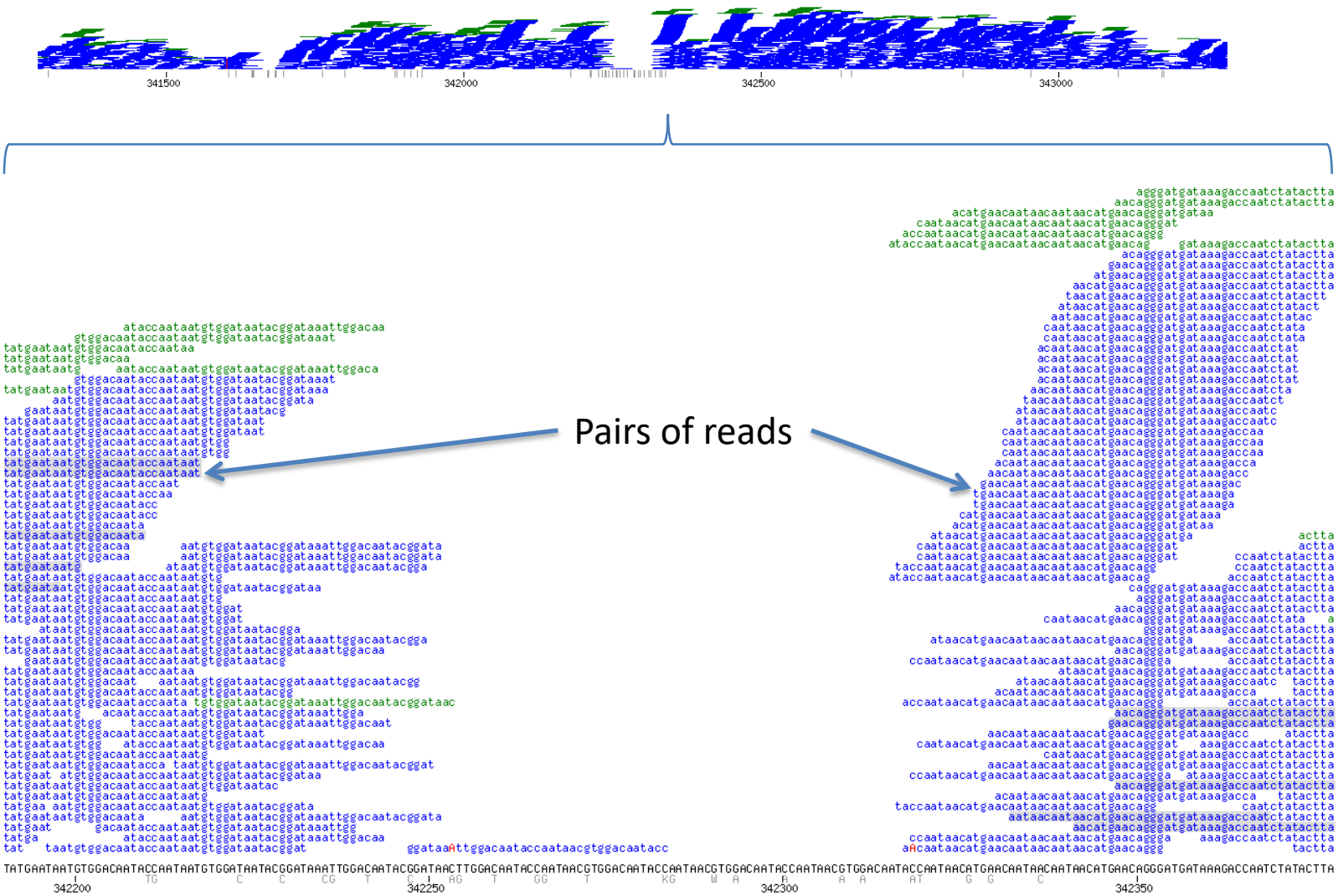
We get (50-100) millions of small fragments (around **100bp** each).

How can we make a genome from it?

DNA Sequence
(genetic code of A, C, G, and T alleles)

```
•@HS4_5964:1:1106:17017:101018#12/1
•CCTTAGGGTCGCCGTTAAGTTCGGAGACGACCGCGTTCCACACTGTGGTGAAGCCTGAACCGGGGTCATCGGTCA
•+
•GDDGGE@E?B????BDEEEE?=DBBGDGGDD?BBBDGAGDB:=B=CE?9EDAGD@===:=292,/9=:=B=566;
•@HS4_5964:1:2205:13272:35605#12/1
•CCTTAGGGTCGCCGTTAAGTTCGGAGACGACCGCGTTCCACACTGTGGTGAAGCCTGAACCGGGGTCATCGGTCA
•+
•IIIIHIHIIIIIIIHIIIHIIFIFHIIIIHIIIIIIIIDIDBGGGGDG>DGBD@DGGGDFG>E?C2>D<8B(*,46
•@HS4_5964:1:2108:7021:12911#12/1
•ACCTTAGGGTCGCCGTTAAGTTCGGAGACGACCGCGTTCCACACTGTGGTGAAGCCTGAACCGGGGTCATCGGTC
•+
•IIIIIIHIIHIIIIIIHIHIIIIIIGIDIFIIIIIIIIIHIIHIHIIEIDFIHIGFIIIIIG>GEE?2CCEFG8BD
•@HS4_5964:1:1206:21270:179616#12/1
•CAANCTTAGGGTCGCCGTTAAGTTCGGAGACGACCGCGTTCCACACTGTGGTGAAGCCTGAACCGGGGTCATCGG
•+
•>>7%<??;?8DGGBGEGGDCDGG>GHHHHHGD@@@DGGGGHDHBHBGDECDDGD<EE??<?=?A:+744'=;947
•@HS4_5964:1:1103:11932:160767#12/1
•AAACGGCACTCGACAATCAAGCGAGGATGGCGGATTGACTAGCGGGCCCGACAACCTGGACCCGGGGGTTTCAAC
•+
•GGGDGGGEGGGEGDG2BBB=BBD?DGGGGG4<=/<'/18-+550-('1)-+4-1.',,)6(.&11&2)(7.&',4
•@HS4_5964:1:2206:3766:101157#12/1
•CGTCGTCAACCTTAGGGTCGCCGTTAAGTTCGGAGACGACCGCGTTCCACACTGTGGTGAAGCCTGAACCGGGGT
•+
•IIIIIIIIIIIIIIIIIIIHIIGIIIIIIIIIIGIBIFHIHIIIIIHIIIGIIGGIHHIDBDIIIIHHIDIGGID
•@HS4_5964:1:2104:10206:46786#12/1
•GGGTGTTTTCAACACGAGGATCACGAGCCGTTGCCGGTAGGTTGCCGCTGGGTTTTGTAGGGGAGGTCTACCAAT
•+
•BC?CAFGEAAGGGGDDG@BGEFBFEGD:GGGGG:B7;;?3;3;31+:32/>+)'')'&&*4'*))&)'&&1')'7
•@HS4_5964:1:1206:14745:64142#12/1
•GCTGGGTCCGTCGTCAACCTTAGGGTCGCCGTTAAGTTCGGAGACGACCGCGTTCCACACTGTGGTGAAGCCTGA
•+
•IIIIIIIIIIIIIIIIHIIIIIIIIHIGIIIIIIHFIIGIEGIIIHHIIDIIIIIIIIHIIHHIIIIHIGFIIII
•@HS4_5964:1:1101:16277:45982#12/1
•AGCATCACTGCTGGGTCCGTCGTCAACCTTAGGGTCGCCGTTAAGTTCGGAGACGACCGCGTTCCACACTGTGGT
•+
•BDDGGHHGGHHHHHEEGEEGDGGGGDGGGEFFBFEHHHGHGGDGGFFFFEGGDGGGGGD>GBEGG@G2GGBCD8>
•@HS4_5964:1:1207:5095:179812#12/1
•AGCATCACTGCTGGGTCCGTCGTCAACCTTAGGGTCGCCGTTAAGTTCGGAGACGACCGCGTTCCACACTGTGGT
```

Alignment to a reference (pile-up view)



Alignment and beyond

- Alignment to a reference (e.g. BWA, Bowtie)
- Multiple alignment format (e.g. BAM, SAM)
- Visualising read alignment and variants in browsers (e.g. Tablet, Artemis, IGV)
- Algorithms for calling variants
  - Small variants (e.g. SNPs, indels) SAM/BCFtools → VCF
  - Larger variants (e.g. large indels) Pindel, Delly, Lumpy
- Reference free or *de novo* assembly (e.g. Velvet)

Commonly used “short read” mappers

BWA

“Burrows-Wheeler Alignment tool“ (Li, 2009)

Perhaps the most widely used aligner, 20x faster than MAQ

Bowtie 2

Builds upon Bowtie, allowing for gaps (Langmead et al, 2009/12)

Works best when aligning short reads to large genomes

Forms the basis of other tools (e.g. Crossbow, Tophat, Cufflinks)

Li and Homer, 2010

Technical issues

- Maximum number of mismatches allowed
  - E.g. MAQ -n 3 specifies no more than 3 mismatches.
- What if reads map to multiple locations?
  - Discard those reads (e.g. SMALT)
  - Pick a random location (e.g. BWA/MAQ).

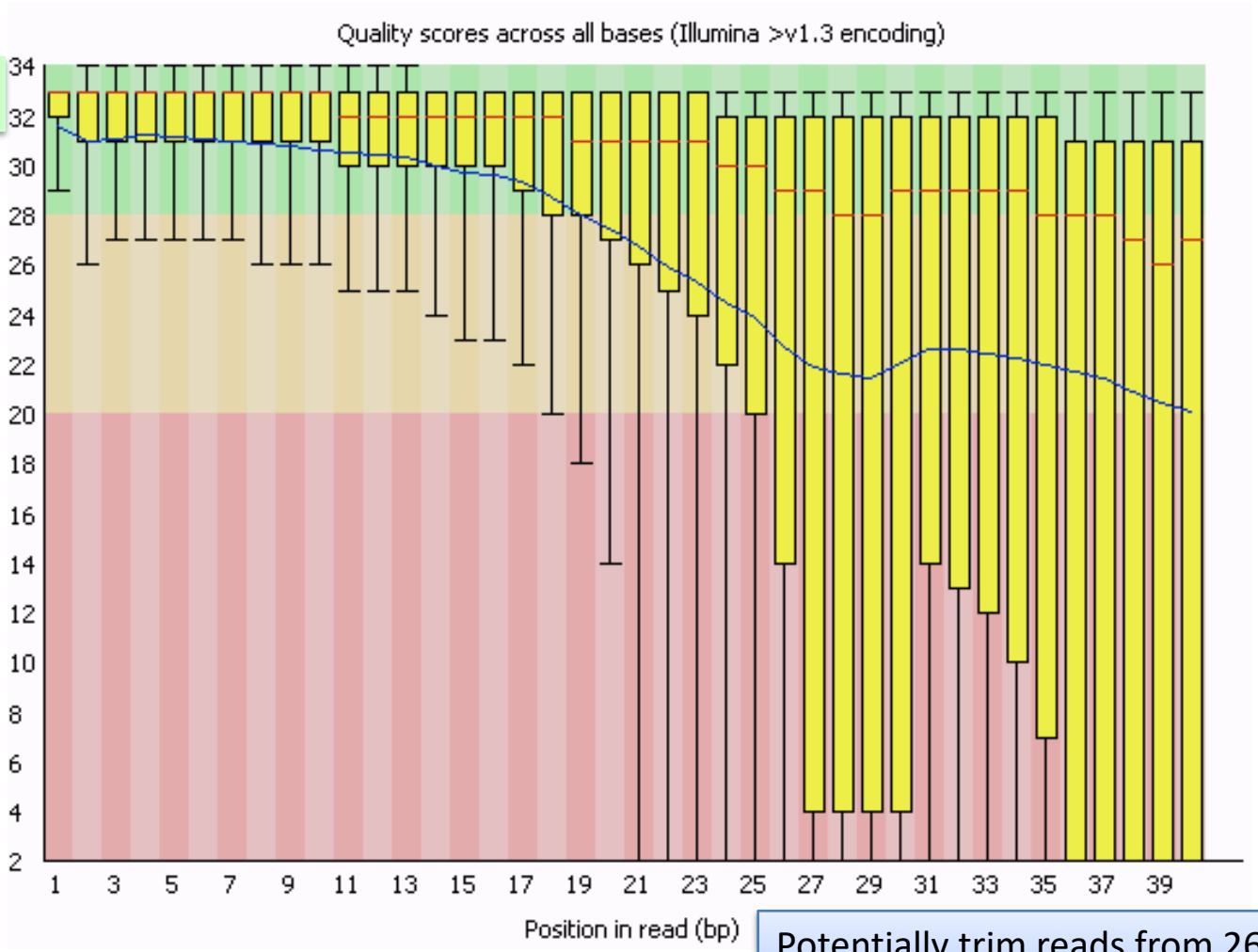
Quality control

- Prior to mapping, screening for contamination
  - E.g. search a large sequence dataset against a panel of different databases (*FastQ Screen*)
- Unreliable read ends can lead to over-calling of indels
  - Quality scores across bases / sequences can be calculated (*fastQC*)
- Reads can be trimmed / clipped
  - *Hard clipping* does not store the clipped sequence of the read, and is performed when the read is first processed;
  - *Soft clipping* keeps the full read, usually performed within an alignment algorithm
- Duplicate or problematic reads can be filtered (*Picard tools*)
- Mapping statistics
  - % reads mapped: single and paired (*samtools*)
  - % genome that is covered, and to what level (*samtools*)

FastQC: Quality score per read base

Good quality

Poor quality

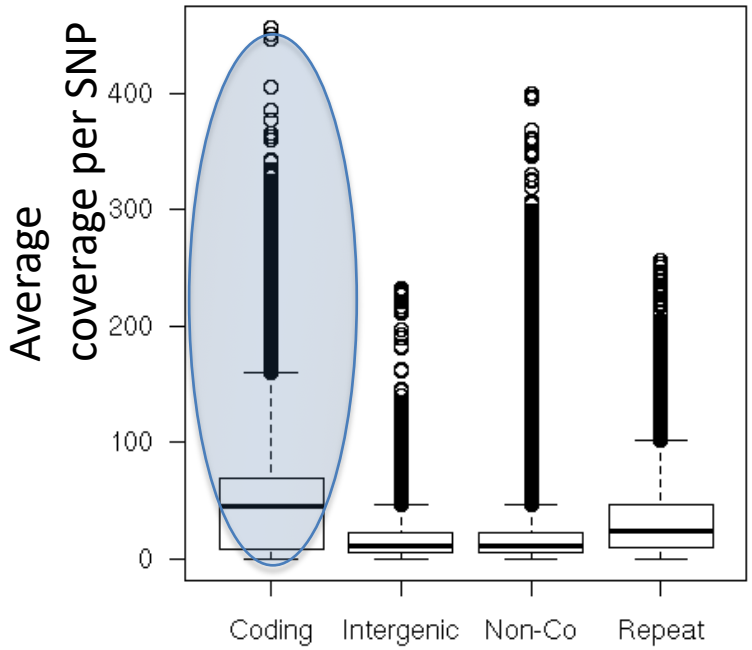


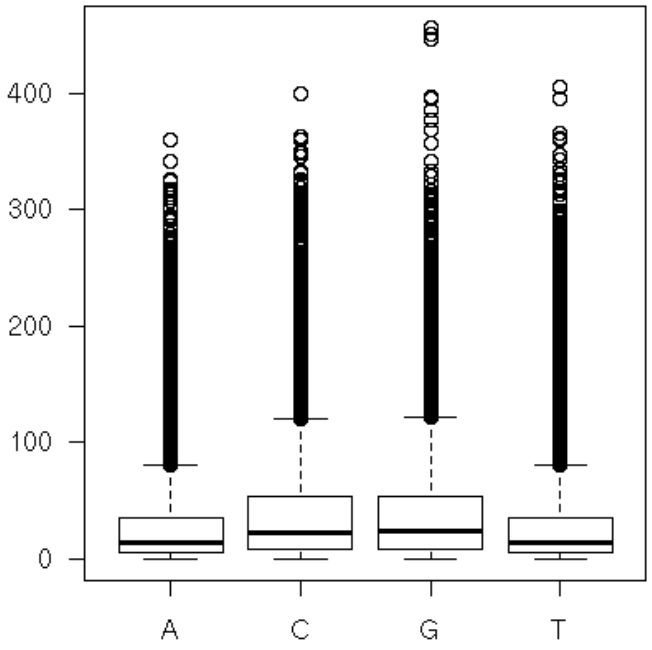
Q30/40 = 1 error per 1000/10000 bp

Potentially trim reads from 26bp onwards – “trimmomatic” – hard clipping

Factors contributing to poor mapping

- DNA quality and sequencing errors
- Inappropriate reference genome
- Non-unique regions (e.g. repeat regions)
- GC content (e.g. malaria genome – 81% AT)





Different types of alignment format

- Pile-up
  - Summarizes the number of reads matching to each position, and the alleles associated with these reads.
  - It can also contain information about mapping qualities
- SAM "Sequence Alignment/Map" format.
  - See http://genome.sph.umich.edu/wiki/SAM for a description
  - It is a TAB-delimited text format consisting of an optional header section, and an alignment section.
  - Minimum format agreed on to report sequencing results, and includes all the data in a *fastq* file
- BAM files
  - Binary version of the SAM format, and much more compact in term of storage

Samtools

- Main software to process and analyse alignments

  (Li et al., 2009, samtools.sourceforge.net)
- Processes BAM (not SAM) files
- Some of the options

  index to index the BAM file for rapid analysis access

  view to extract the genomic region or a section of interest

  Idxstats & *flagstat* summary statistics of the mapping

  mpileup generates (multiple) sample alignments in BCF format for variant detection using BCF/VCFtools

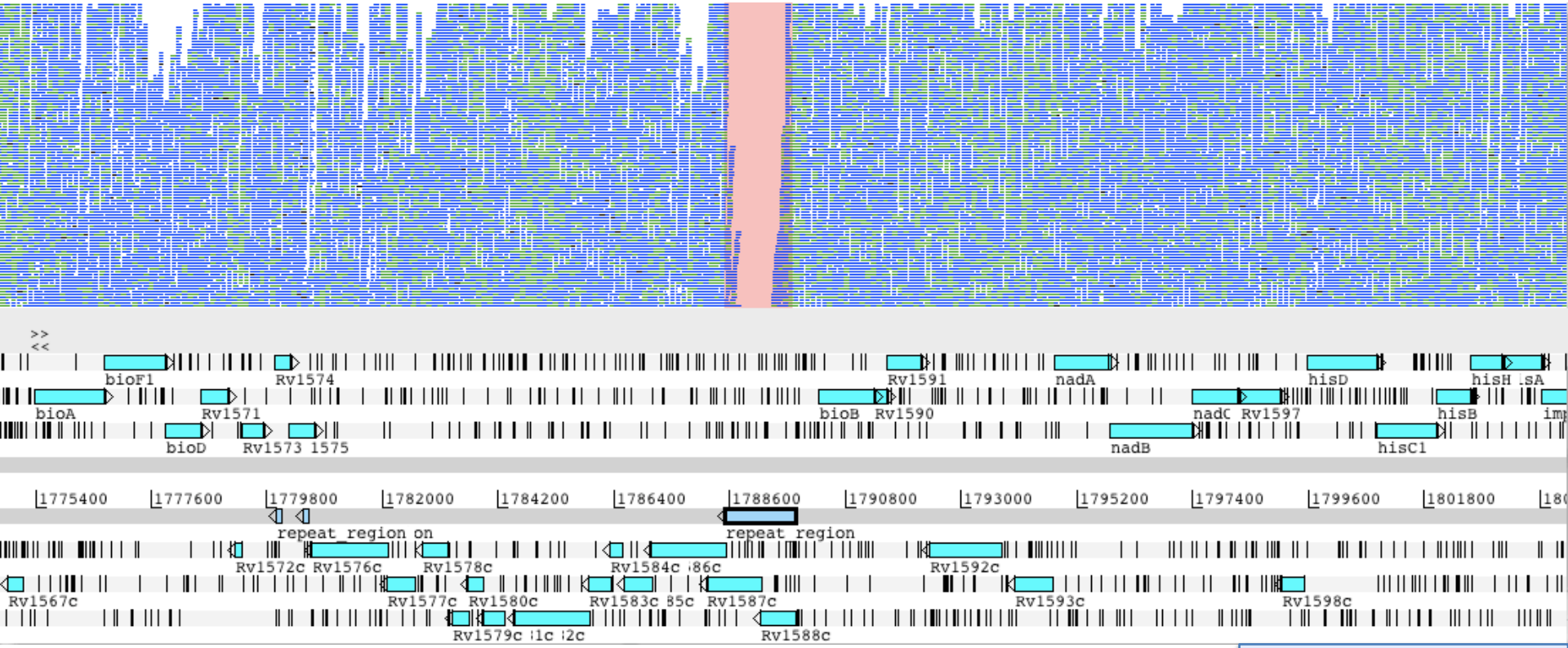
Aligning the reads in *M.tb*

Mapping uniquely to a modified H37Rv reference (GC rich ~65.6%)

~97% of reads mapped (uniquely), ~300-500x coverage

Some non-unique regions (e.g. repeat regions) are problematic

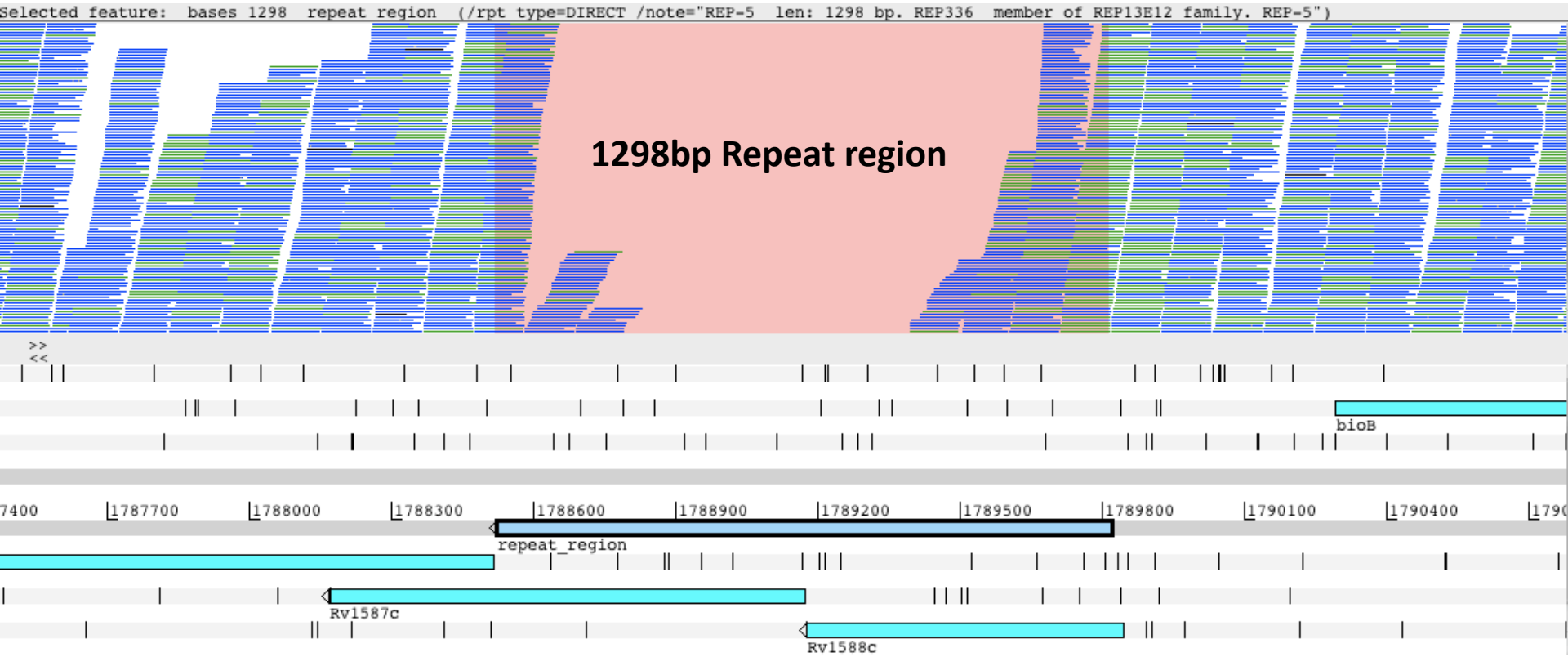
Repeat region (Rv1587/8c)



Lisboa strain

Artemis software

What if we cannot map to a reference?



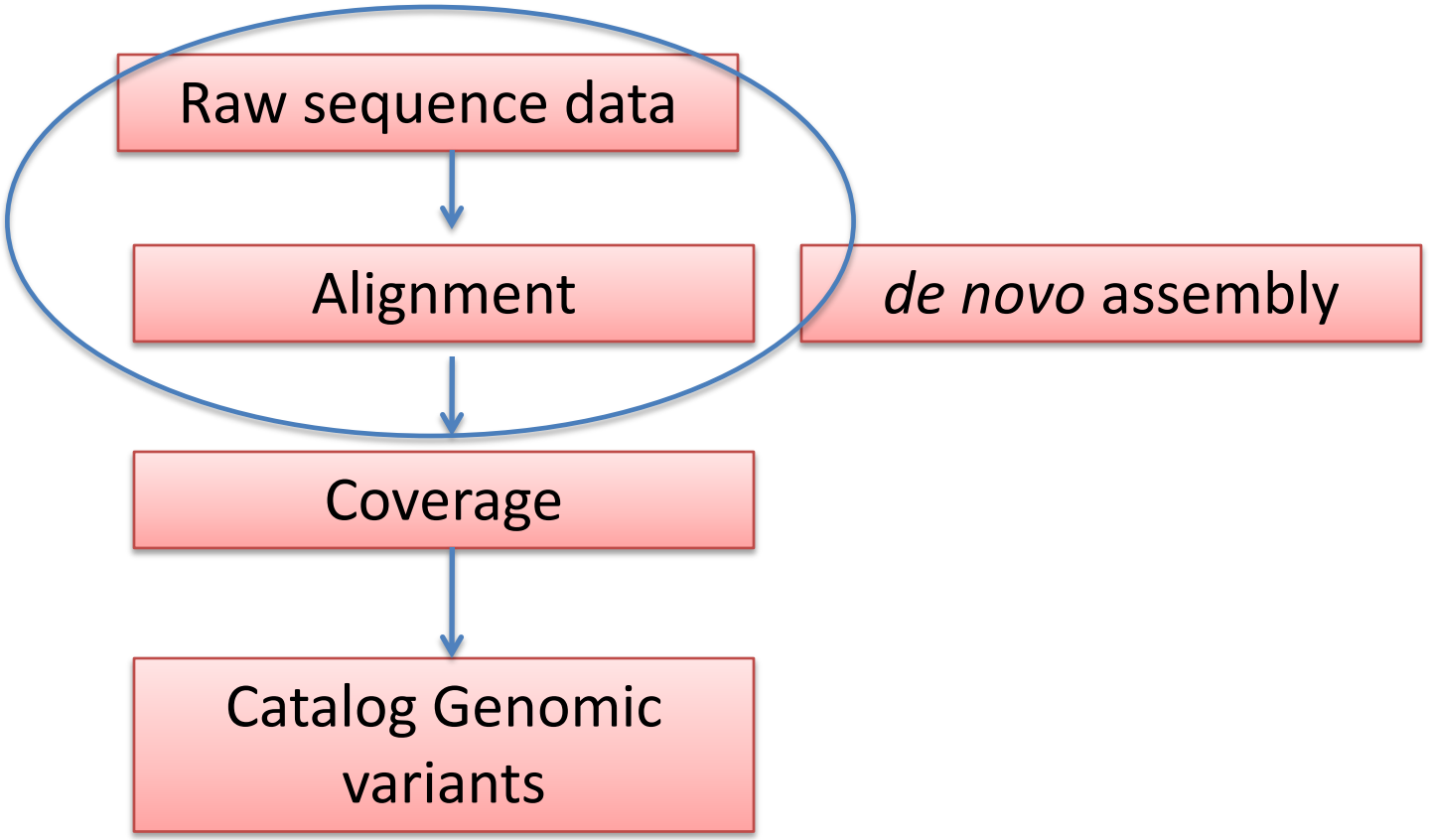
Solutions:

- Perform *de novo* assembly on non-mapped reads
- Use technologies with longer reads (e.g. 454 technology)
- Population-specific reference genomes

Browsing, visualizing and interpreting data

- Visualization of genome assemblies
  - EagleView [1]
  - HawkEye [2]
  - **Tablet** [3]
- Visualising read alignments with genome annotation
  - LookSeq[4]
  - Integrative Genome Viewer (IGV) **[5]**
  - Integrated Genome Browser (IGB) [6]
  - BamView (integrated with **Artemis**, ACT) [9-11]
  - GenoViewer [12]
- Browsers also with genetic variation detection and analysis
  - MagicViewer [7]
  - Savant [8]

Alignment



Population genetics

Whole genome Association studies

References

1. Huang W, Marth GT. EagleView: a genome assembly viewer for next-generation sequencing technologies. Genome Res 2008;18:1538–43.
2. Schatz MC, Phillippy AM, Shneiderman B, Salzberg SL. Hawkeye: an interactive visual analytics tool for genome assemblies. Genome Biol 2007;8:R34.
3. Milne I, Bayer M, Cardle L, et al. Tablet—next generation sequence assembly visualization. Bioinformatics 2010;26: 401–2.
4. Manske HM, Kwiatkowski DP. LookSeq: a browser-based viewer for deep sequencing data. Genome Res 2009;192125–32.
5. Robinson JT, Thorvaldsdottir H, Winckler W, et al. Integrative genomics viewer. Nat Biotech 2011;29:24–6.
6. Nicol JW, Helt GA, Blanchard SG, et al. The Integrated Genome Browser: free software for distribution and exploration of genome-scale datasets. Bioinformatics 2009;25:2730–1.
7. Hou H, Zhao F, Zhou L, et al. MagicViewer: integrated solution for next-generation sequencing data visualization and genetic variation detection and annotation. Nucleic Acids Res 2010;38:W732–6.
8. Fiume M, Williams V, Brook A, et al. Savant: genome browser for high-throughput sequencing data. Bioinformatics 2010;26:1938–44.
9. Carver T, Bo¨hme U, Otto TD, et al. BamView: viewing mapped read alignment data in the context of the reference sequence. Bioinformatics 2010;26:676–7.
10. Rutherford K, Parkhill J, Crook J, et al. Artemis: sequence visualization and annotation. Bioinformatics 2000; 16:944–5.
11. Carver T, Berriman M, Tivey A, et al. Artemis and ACT: viewing, annotating and comparing sequences stored in a relational database. Bioinformatics 2008;24:2672–6.
12. http://www.genoviewer.com