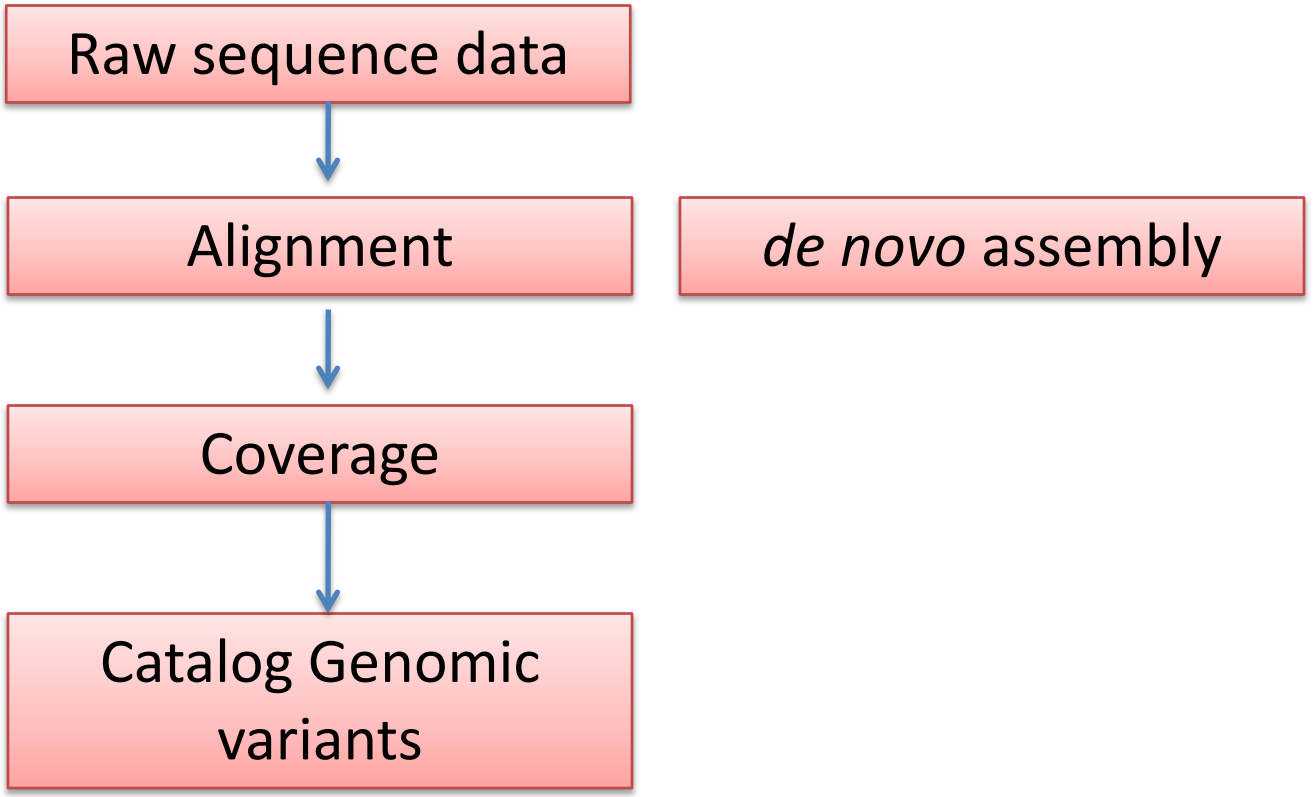
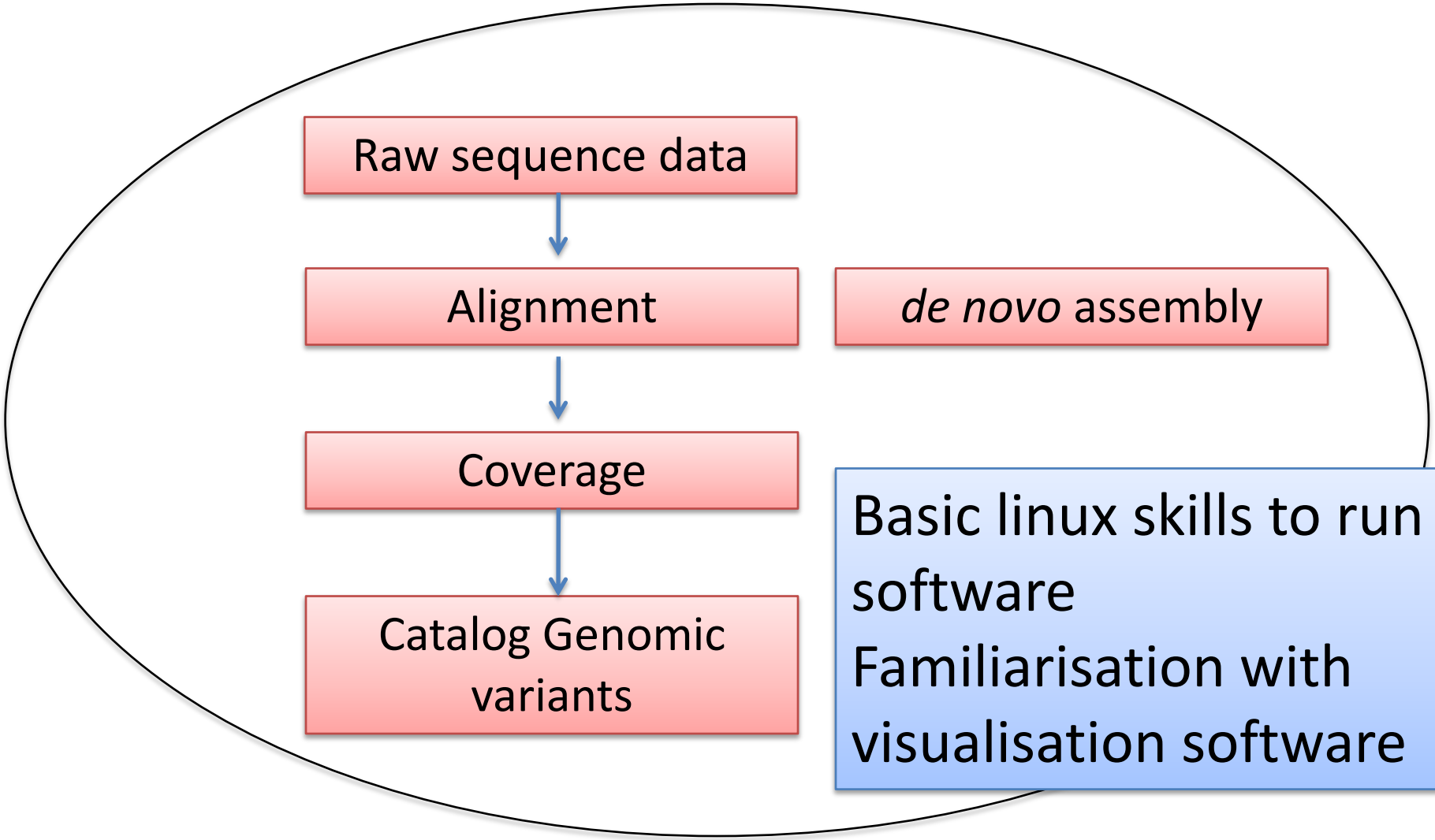
Infectious Disease 'Omics

Course overview

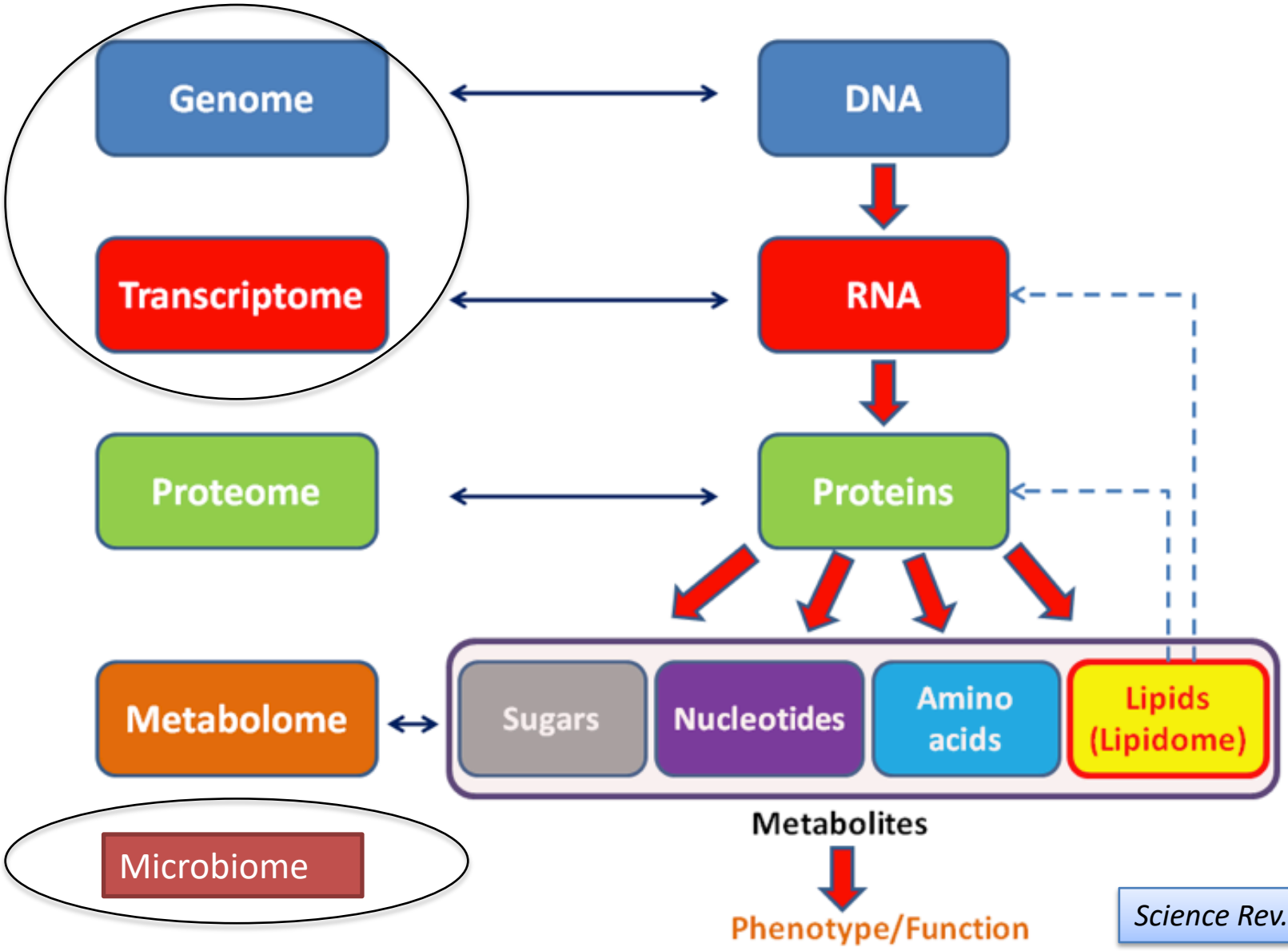
Course outline



Course outline



Omics cascade



Science Rev. 2012

Capillary or “Sanger” Sequencing

1st Generation technology

ABI 3730xl Capillary

- Human (2900Mb; Schuler et al. 1996)
- H37RV (4.4Mb; Cole et al, 1998)
- 3D7 (23Mb; Gardner et al, 2002)

- Long reads (~700bp) and inserts
- Cloning required
- Low coverage, $500 per Mb
- Run time: months



Platform	Read Length	Throughput	Reads	Runtime	Cost per Gb ($)
Illumina MiSeq v3	300 (PE)	13.2-15Gb	44-50M	21-56 h	110
Illumina HiSeq X	150 (PE)	800–900Gb	2.6–3B	3 d	7
Ion PGM 318	400 (SE)	1–2Gb	4–5.5 M	7.3h	$450–800
PacBio RS II	~20Kb	500 Mb–1 Gb	~55,000*	4 h*	$1,000
MinION	Up to 200Kb	Up to 1.5Gb	100,000	Up to 48h	$750

Compared to Sanger sequencing:

>10^4 increase in data per machine run, < 10^3 fold drop in cost

>10^7 fold ⇧ throughput

Next generation

Oxford Nanopore MinIon
Nature **521,** 15–16 (07 May 2015) doi:10.1038/521015a

Pacific Biosciences
Single molecule real time sequencing; long reads (>10kb), and methylation

Illumina
The HiSeq X Ten
10 HiSeq X ultra-high-throughput instruments
$1000 per human genome.

Company	Read length	Applications	Website
454/Roche	400 bp (single end)	Bacterial and viral genomes, multiplex-PCR products, validation of point mutations, targeted somatic-mutation detection	http://www.454.com/
Illumina	150–300 bp (paired end)	Complex genomes (human, mouse and plants) and genome-wide NGS applications, RNA-seq, hybrid capture or multiplex-PCR products, somatic-mutation detection, forensics, noninvasive prenatal testing	http://www.illumina.com/
ABI SOLiD	75 bp (single end) or 50 bp (paired end)	Complex genomes (human, mouse, plants) and genome-wide NGS applications, RNA-seq, hybrid capture or multiplex-PCR products, somatic-mutation detection	http://www.thermofisher.com/us/en/home/life-science/sequencing/next-generation-sequencing/solid-next-generation-sequencing.html/
Pacific Biosciences	Up to 40 kb (single end or circular consensus)	Complex genomes (human, mouse and plants), microbiology and infectious-disease genomes, transcript-fusion detection, methylation detection	http://www.pacb.com
Ion Torrent	200–400 bp (single end)	Multiplex-PCR products, microbiology and infectious diseases, somatic-mutation detection, validation of point mutations	http://www.thermofisher.com/us/en/home/life-science/sequencing/next-generation-sequencing.html/
Oxford Nanopore	Variable: depends on library preparation (1D or 2D reads)	Pathogen surveillance, targeted mutation detection, metagenomics, bacterial and viral genomes	http://nanoporetech.com/
Qiagen GeneReader	107 bp (single end)	Targeted mutation detection, liquid biopsy in cancer	http://www.genereaderngs.com/

Nature Protocols 12, 213–218 (2017)

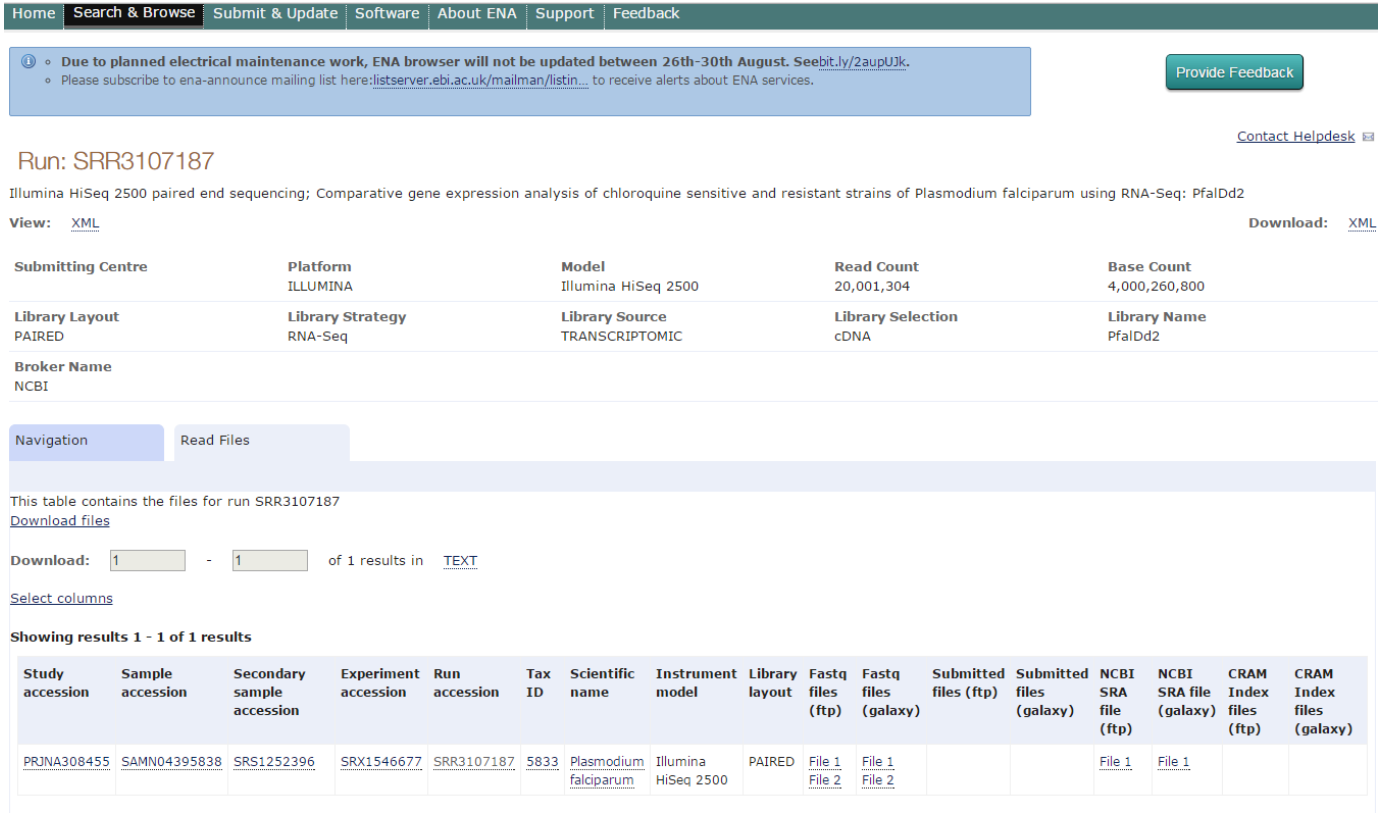
Sequence Read Archive

Next generation sequencing data is accessible for all researchers via the Sequence Read Archive.

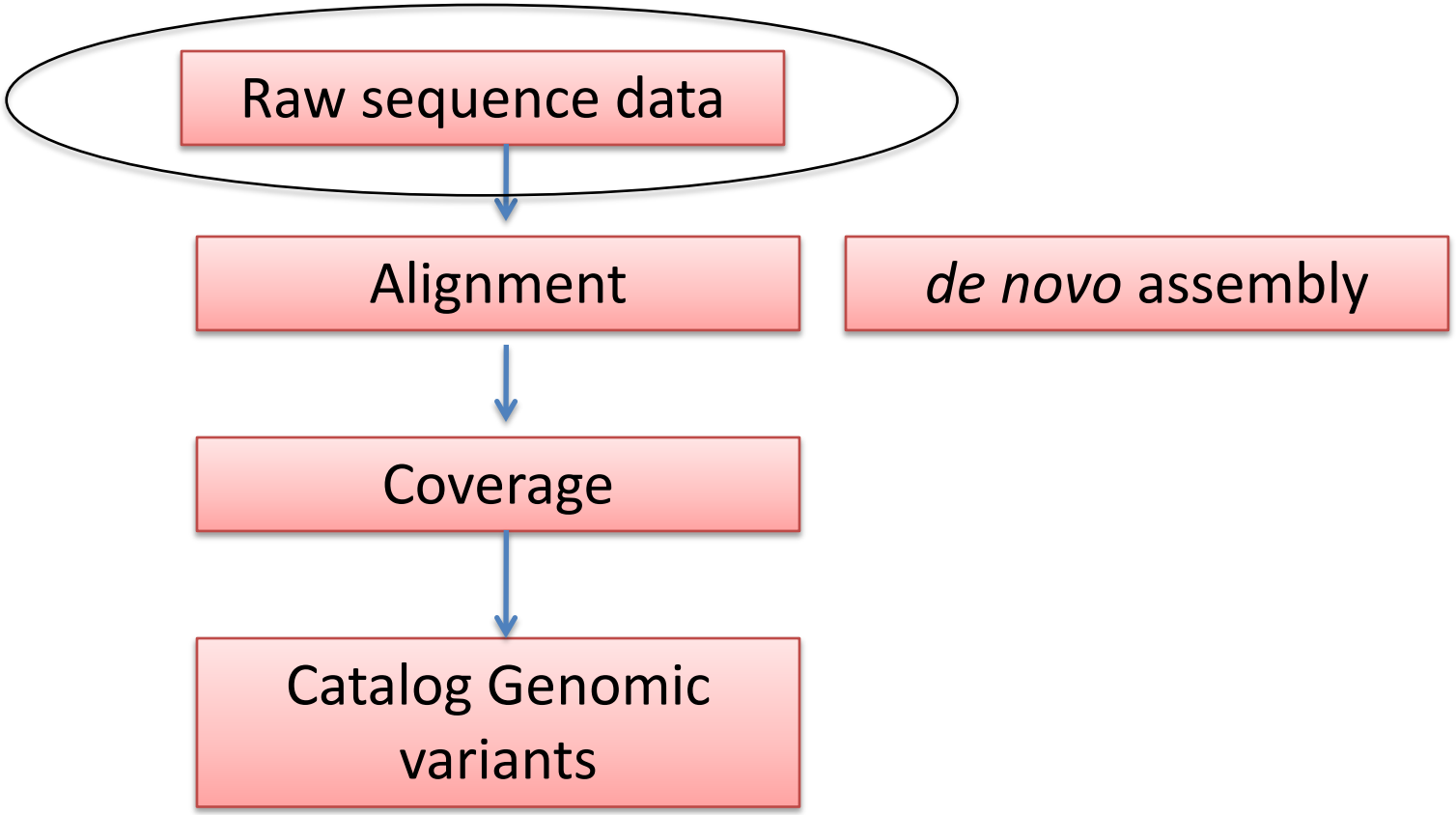
You can access the sequence read archive through the following sites:

http://www.ncbi.nlm.nih.gov/sra

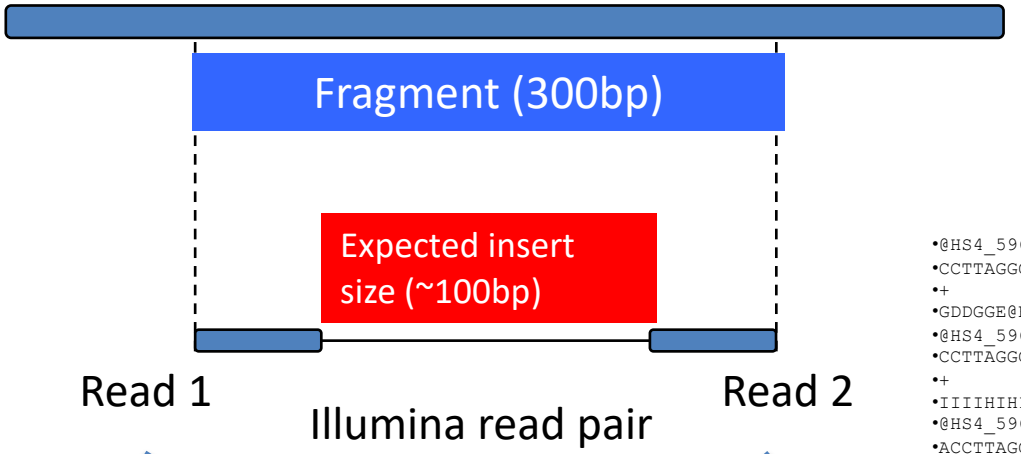
http://www.ebi.ac.uk/ena



The course


Raw sequence data
Alignment
de novo assembly
Coverage
Catalog Genomic variants

Raw sequence data (Illumina) – “fastq”



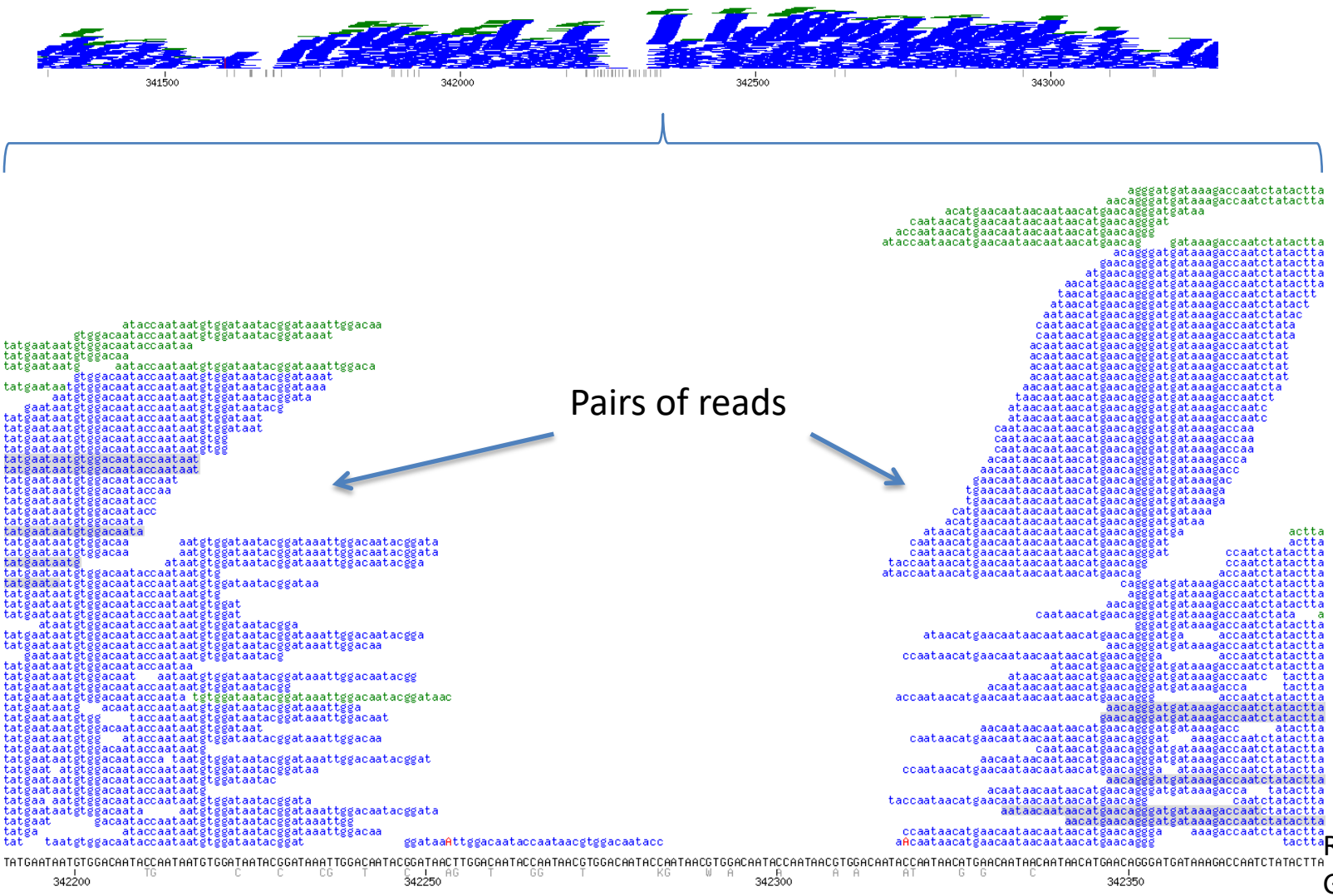
We get (50-100) millions of small fragments (around 100bp each).

How can we make a genome from it?

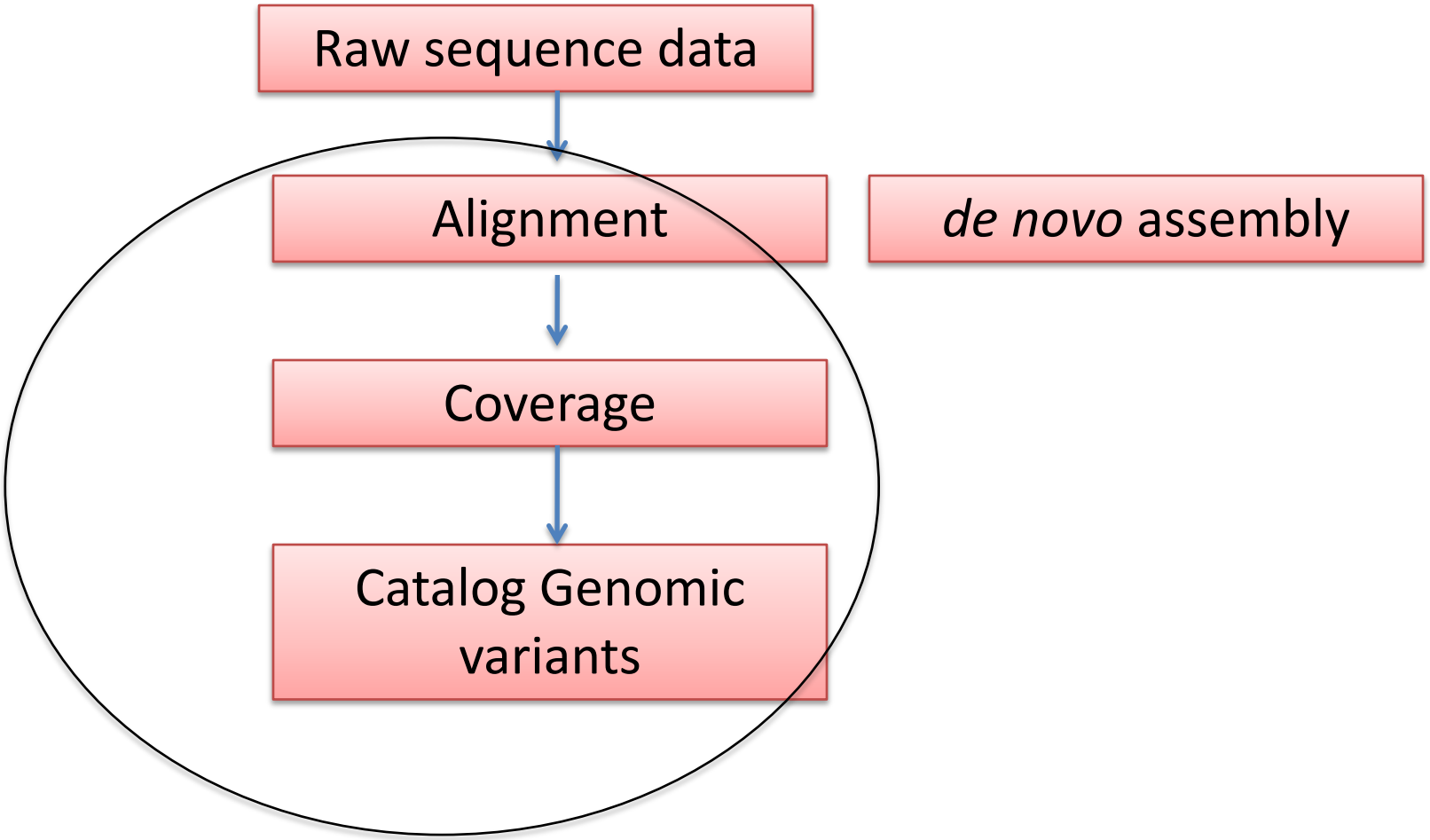
DNA Sequence
(genetic code of A, C, G, and T alleles)

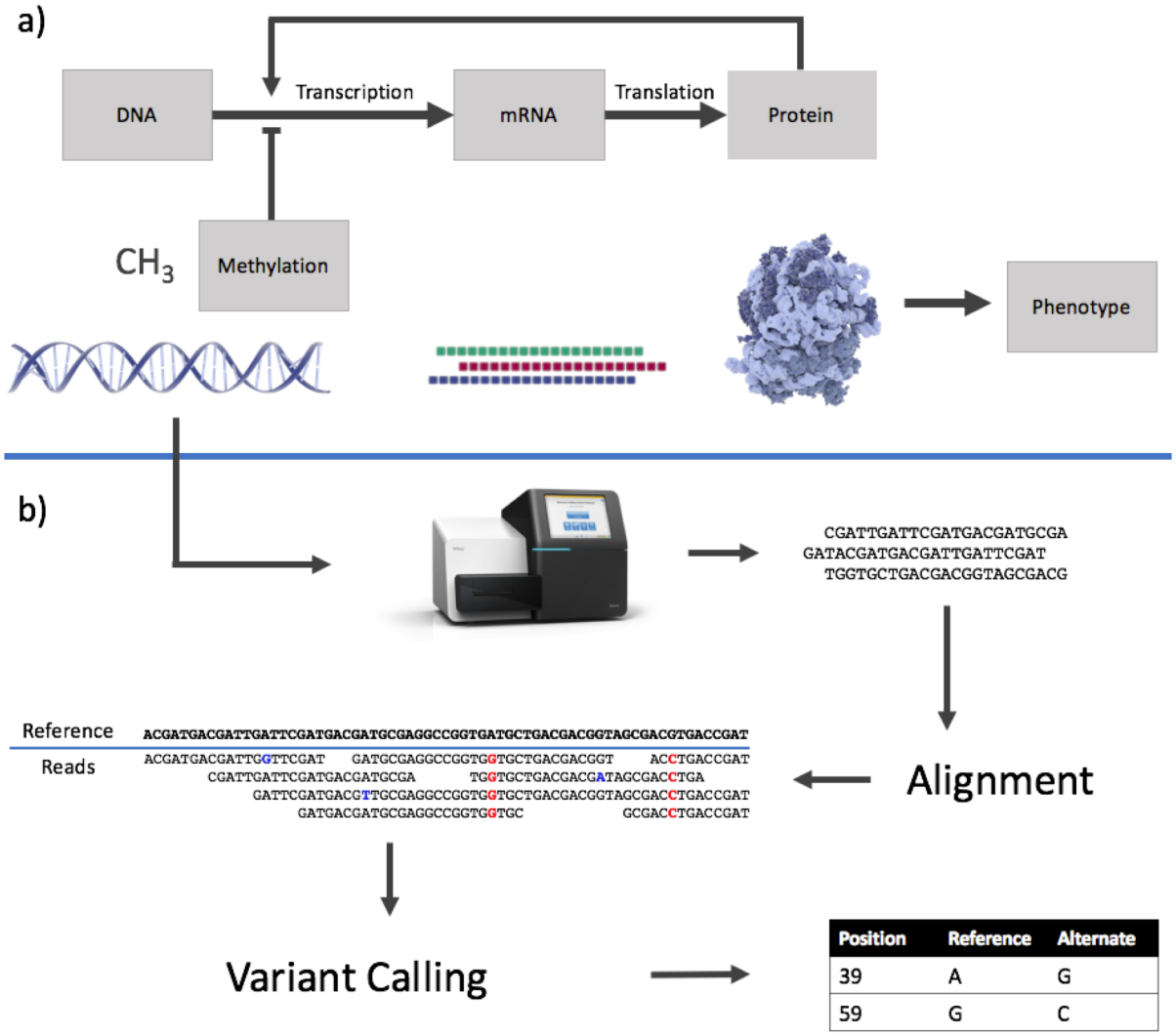
```
•@HS4_5964:1:1106:17017:101018#12/1
•CCTTAGGGTCGCCGTTAAGTTCGGAGACGACCGCGTTCCACACTGTGGTGAAGCCTGAACCGGGGTCATCGGTCA
•+
•GDDGGE@E?B????BDEEEE?=DBBGDGGDD?BBBDGAGDB:=B=CE?9EDAGD@===:=292,/9=:=B=566;
•@HS4_5964:1:2205:13272:35605#12/1
•CCTTAGGGTCGCCGTTAAGTTCGGAGACGACCGCGTTCCACACTGTGGTGAAGCCTGAACCGGGGTCATCGGTCA
•+
•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
•+
•IIIIIIIIIIIIIIIIHIIIIIIIIHIGIIIIIIHFIIGIEGIIIHHIIDIIIIIIIIIHIIHHIIIIHIGFIIII
•@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)



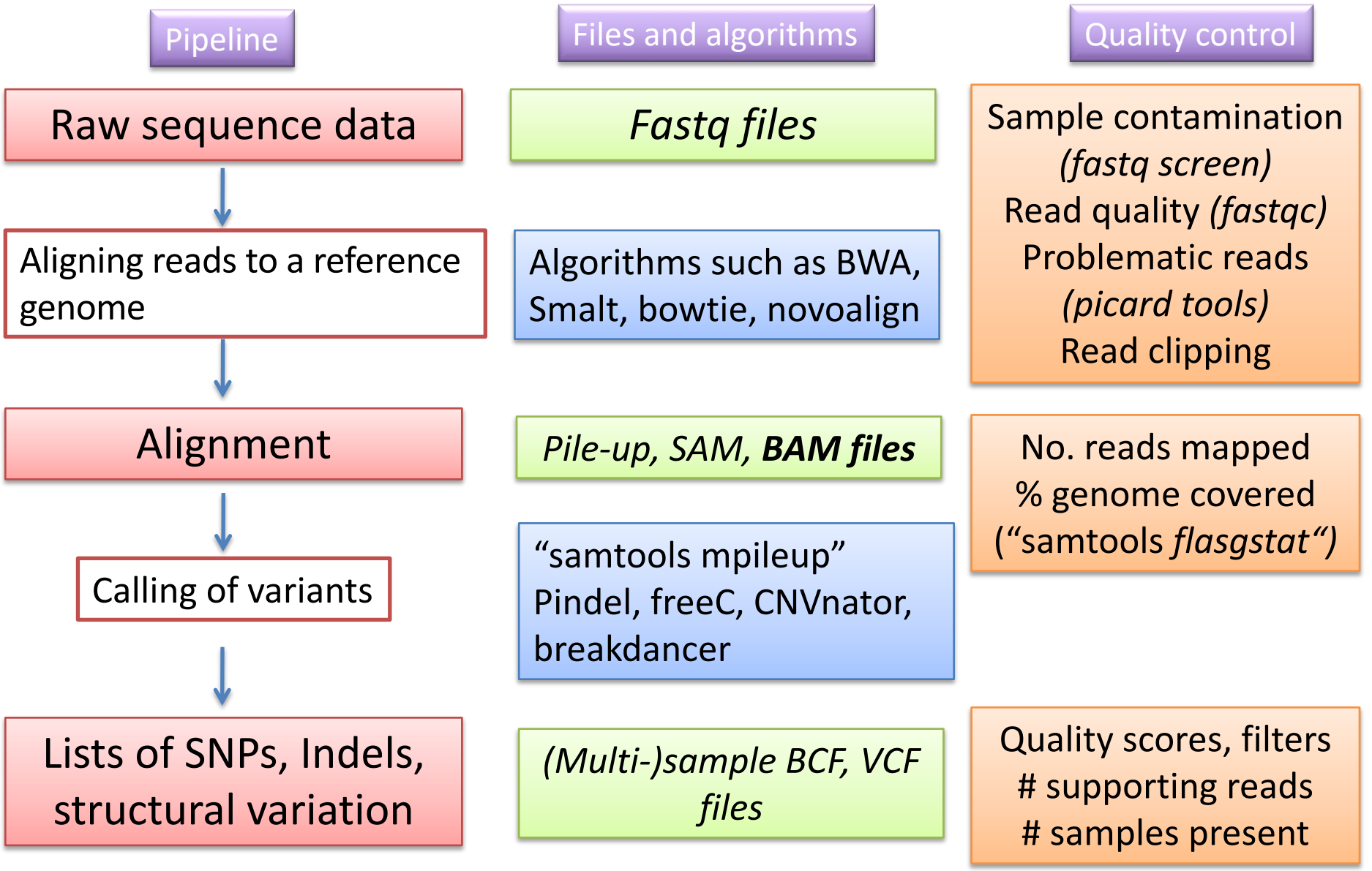
The course


Raw sequence data
Alignment
Coverage
Catalog Genomic variants
de novo assembly



Position	Reference	Alternate
39	A	G
59	G	C

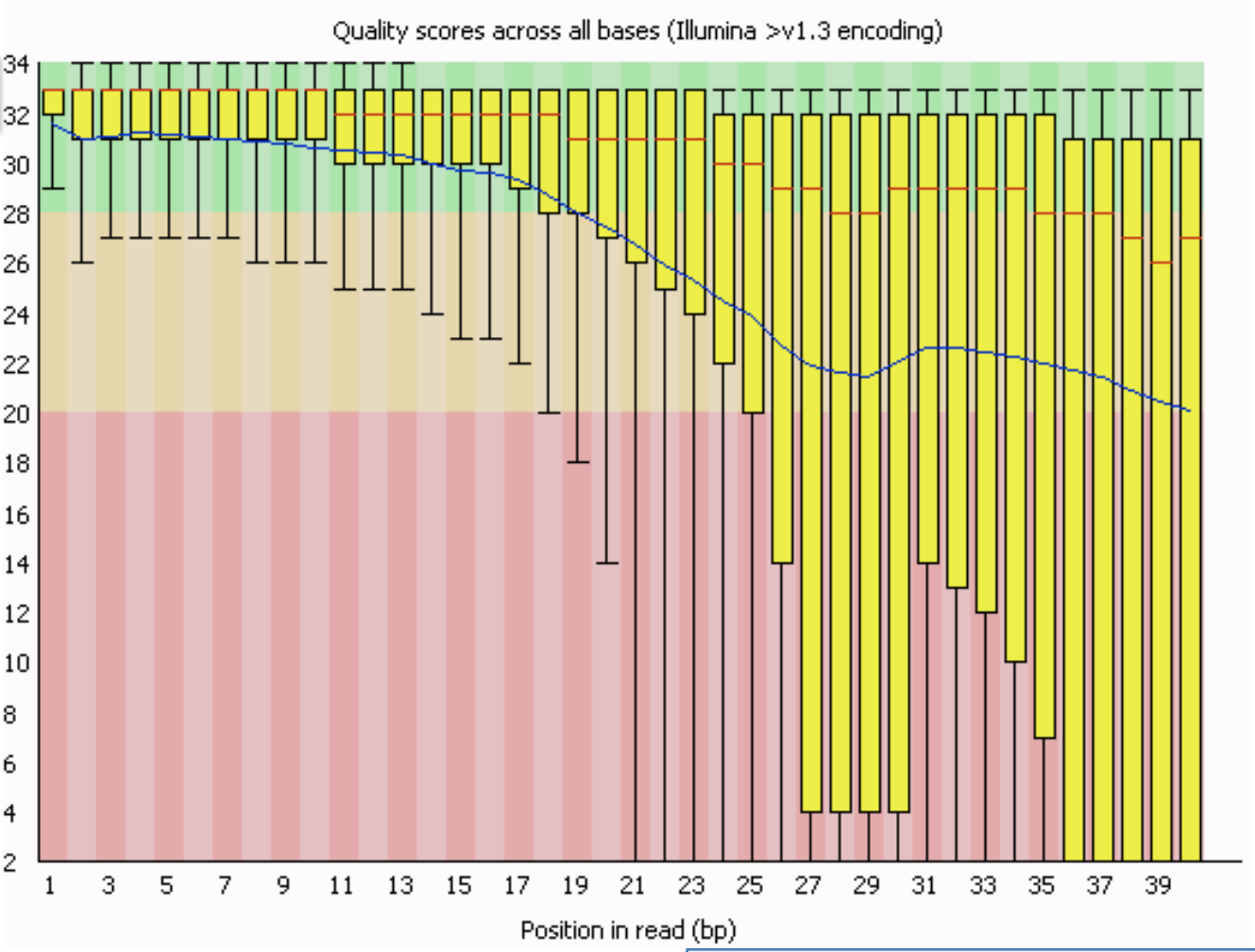
Alignment and beyond



FastQC: Quality score per read base

Good quality

Poor quality



Q20/30/40 = 1 error per 100/1000/10000 bp

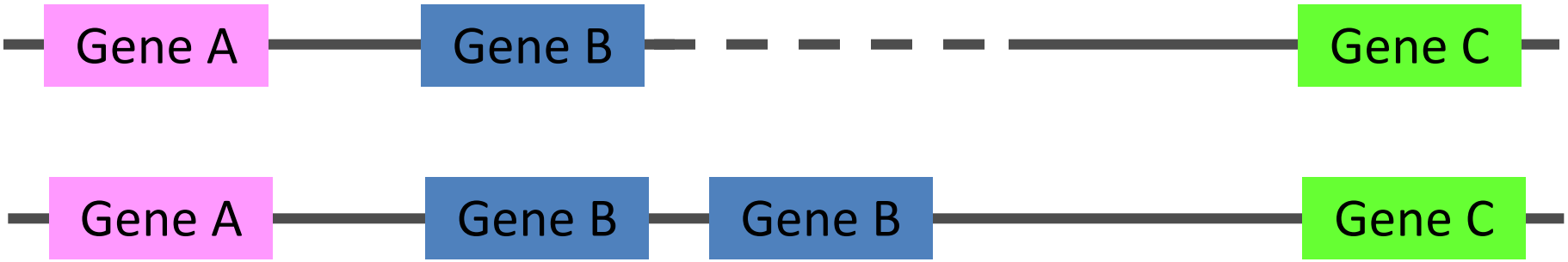
Potentially trim reads from 26bp onwards –
"trimmomatic" – hard clipping
Or within alignment software – soft clipping

Examples of variants detectable from alignments

```
ACTCTACGATTTACGGTACTTAGGAGCATATGCTACT
ACTGTACGATTTACGGTACTTAG.AGCATATGCTACT
```

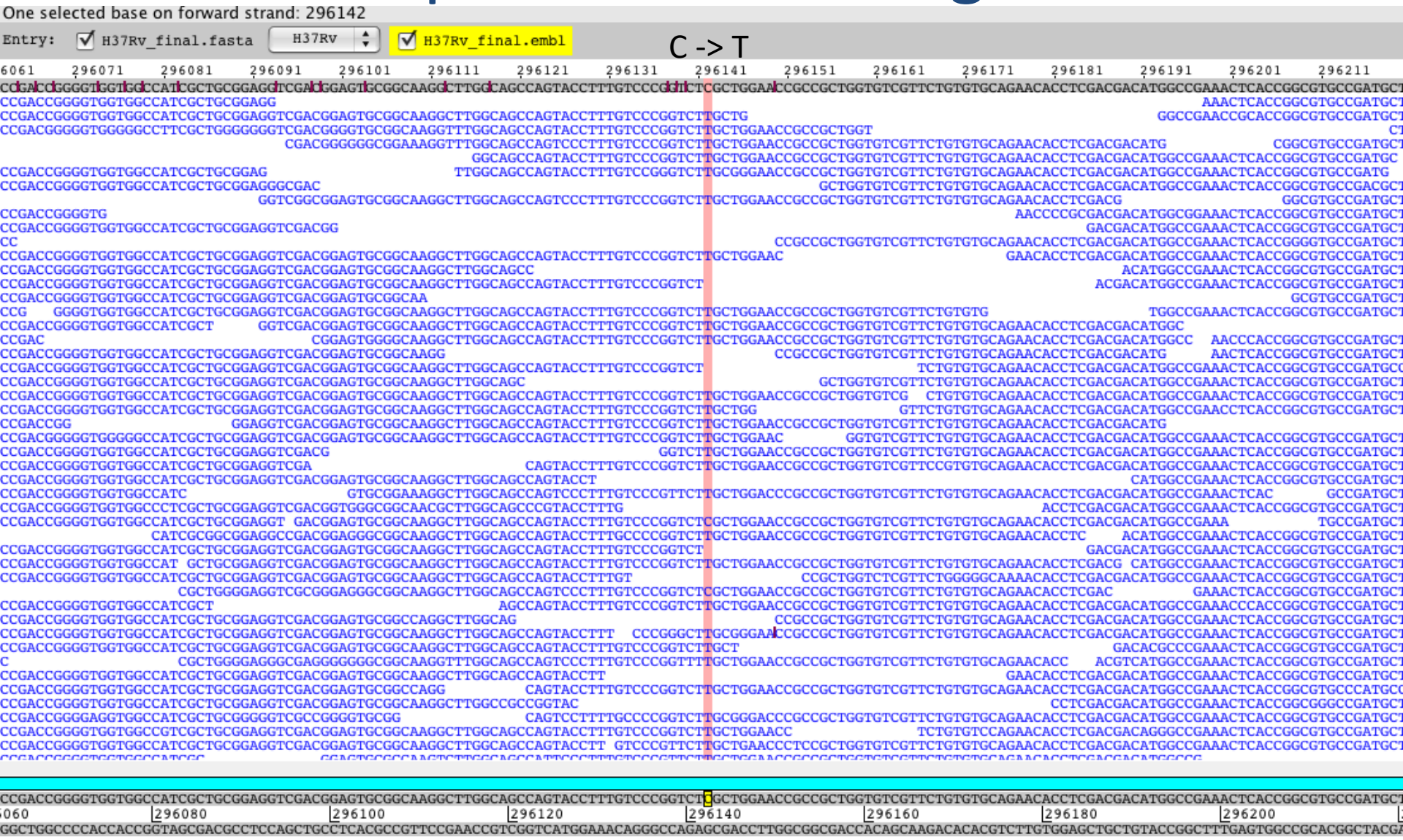
SNP
single nucleotide polymorphism

indel
insertion / deletion



Copy number variation involving a large chunk of DNA that includes the whole of gene B

Pile-up view - detecting SNPs



Lisboa *Mtb* strain

Rv0245

Useful view for observing multiplicity of infection – heterozygous positions

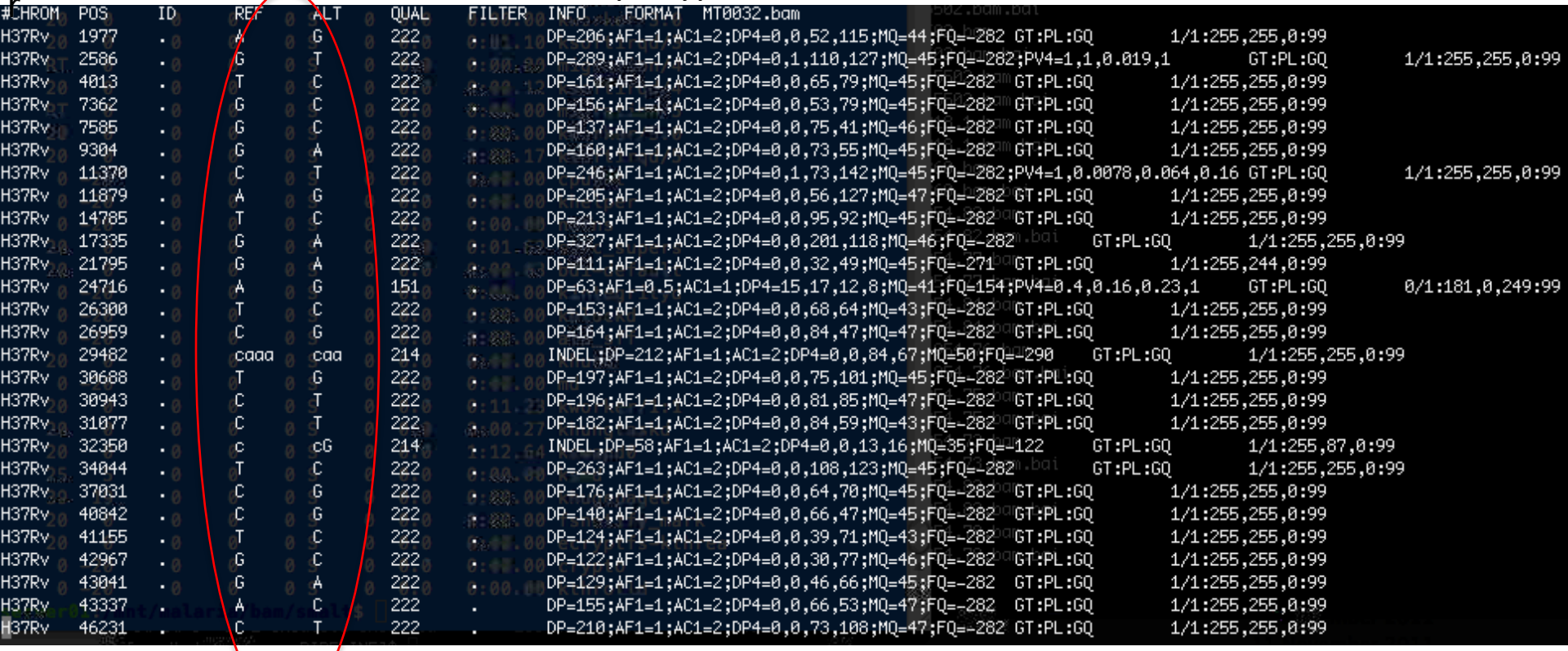
SNPs and small indels in VCF format

Chr

ID

Alleles

Information concerning read depth and genotype / haplotype calls



#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	MT0032.bam
H37Rv	1977	.	A	G	222	.	DP=206;AF1=1;AC1=2;DP4=0,0,52,115;MQ=44;FQ=-282	GT:PL:GQ	1/1:255,255,0:99
H37Rv	2586	.	G	T	222	.	DP=289;AF1=1;AC1=2;DP4=0,1,110,127;MQ=45;FQ=-282;PV4=1,1,0.019,1	GT:PL:GQ	1/1:255,255,0:99
H37Rv	4013	.	T	C	222	.	DP=161;AF1=1;AC1=2;DP4=0,0,65,79;MQ=45;FQ=-282	GT:PL:GQ	1/1:255,255,0:99
H37Rv	7362	.	G	C	222	.	DP=156;AF1=1;AC1=2;DP4=0,0,53,79;MQ=45;FQ=-282	GT:PL:GQ	1/1:255,255,0:99
H37Rv	7585	.	G	C	222	.	DP=137;AF1=1;AC1=2;DP4=0,0,75,41;MQ=46;FQ=-282	GT:PL:GQ	1/1:255,255,0:99
H37Rv	9304	.	G	A	222	.	DP=160;AF1=1;AC1=2;DP4=0,0,73,55;MQ=45;FQ=-282	GT:PL:GQ	1/1:255,255,0:99
H37Rv	11370	.	C	T	222	.	DP=246;AF1=1;AC1=2;DP4=0,1,73,142;MQ=45;FQ=-282;PV4=1,0.0078,0.064,0.16	GT:PL:GQ	1/1:255,255,0:99
H37Rv	11879	.	A	G	222	.	DP=205;AF1=1;AC1=2;DP4=0,0,56,127;MQ=47;FQ=-282	GT:PL:GQ	1/1:255,255,0:99
H37Rv	14785	.	T	C	222	.	DP=213;AF1=1;AC1=2;DP4=0,0,95,92;MQ=45;FQ=-282	GT:PL:GQ	1/1:255,255,0:99
H37Rv	17335	.	G	A	222	.	DP=327;AF1=1;AC1=2;DP4=0,0,201,118;MQ=46;FQ=-282	GT:PL:GQ	1/1:255,255,0:99
H37Rv	21795	.	G	A	222	.	DP=111;AF1=1;AC1=2;DP4=0,0,32,49;MQ=45;FQ=-271	GT:PL:GQ	1/1:255,244,0:99
H37Rv	24716	.	A	G	151	.	DP=63;AF1=0.5;AC1=1;DP4=15,17,12,8;MQ=41;FQ=154;PV4=0.4,0.16,0.23,1	GT:PL:GQ	0/1:181,0,249:99
H37Rv	26300	.	T	C	222	.	DP=153;AF1=1;AC1=2;DP4=0,0,68,64;MQ=43;FQ=-282	GT:PL:GQ	1/1:255,255,0:99
H37Rv	26959	.	C	G	222	.	DP=164;AF1=1;AC1=2;DP4=0,0,84,47;MQ=47;FQ=-282	GT:PL:GQ	1/1:255,255,0:99
H37Rv	29482	.	caaa	caa	214	.	INDEL;DP=212;AF1=1;AC1=2;DP4=0,0,84,67;MQ=50;FQ=-290	GT:PL:GQ	1/1:255,255,0:99
H37Rv	30688	.	T	G	222	.	DP=197;AF1=1;AC1=2;DP4=0,0,75,101;MQ=45;FQ=-282	GT:PL:GQ	1/1:255,255,0:99
H37Rv	30943	.	C	T	222	.	DP=196;AF1=1;AC1=2;DP4=0,0,81,85;MQ=47;FQ=-282	GT:PL:GQ	1/1:255,255,0:99
H37Rv	31077	.	C	T	222	.	DP=182;AF1=1;AC1=2;DP4=0,0,84,59;MQ=43;FQ=-282	GT:PL:GQ	1/1:255,255,0:99
H37Rv	32350	.	c	cG	214	.	INDEL;DP=58;AF1=1;AC1=2;DP4=0,0,13,16;MQ=35;FQ=-122	GT:PL:GQ	1/1:255,87,0:99
H37Rv	34044	.	T	C	222	.	DP=263;AF1=1;AC1=2;DP4=0,0,108,123;MQ=45;FQ=-282	GT:PL:GQ	1/1:255,255,0:99
H37Rv	37031	.	C	G	222	.	DP=176;AF1=1;AC1=2;DP4=0,0,64,70;MQ=45;FQ=-282	GT:PL:GQ	1/1:255,255,0:99
H37Rv	40842	.	C	G	222	.	DP=140;AF1=1;AC1=2;DP4=0,0,66,47;MQ=45;FQ=-282	GT:PL:GQ	1/1:255,255,0:99
H37Rv	41155	.	T	C	222	.	DP=124;AF1=1;AC1=2;DP4=0,0,39,71;MQ=43;FQ=-282	GT:PL:GQ	1/1:255,255,0:99
H37Rv	42967	.	G	C	222	.	DP=122;AF1=1;AC1=2;DP4=0,0,30,77;MQ=46;FQ=-282	GT:PL:GQ	1/1:255,255,0:99
H37Rv	43041	.	G	A	222	.	DP=129;AF1=1;AC1=2;DP4=0,0,46,66;MQ=45;FQ=-282	GT:PL:GQ	1/1:255,255,0:99
H37Rv	43337	.	A	C	222	.	DP=155;AF1=1;AC1=2;DP4=0,0,66,53;MQ=47;FQ=-282	GT:PL:GQ	1/1:255,255,0:99
H37Rv	46231	.	C	T	222	.	DP=210;AF1=1;AC1=2;DP4=0,0,73,108;MQ=47;FQ=-282	GT:PL:GQ	1/1:255,255,0:99

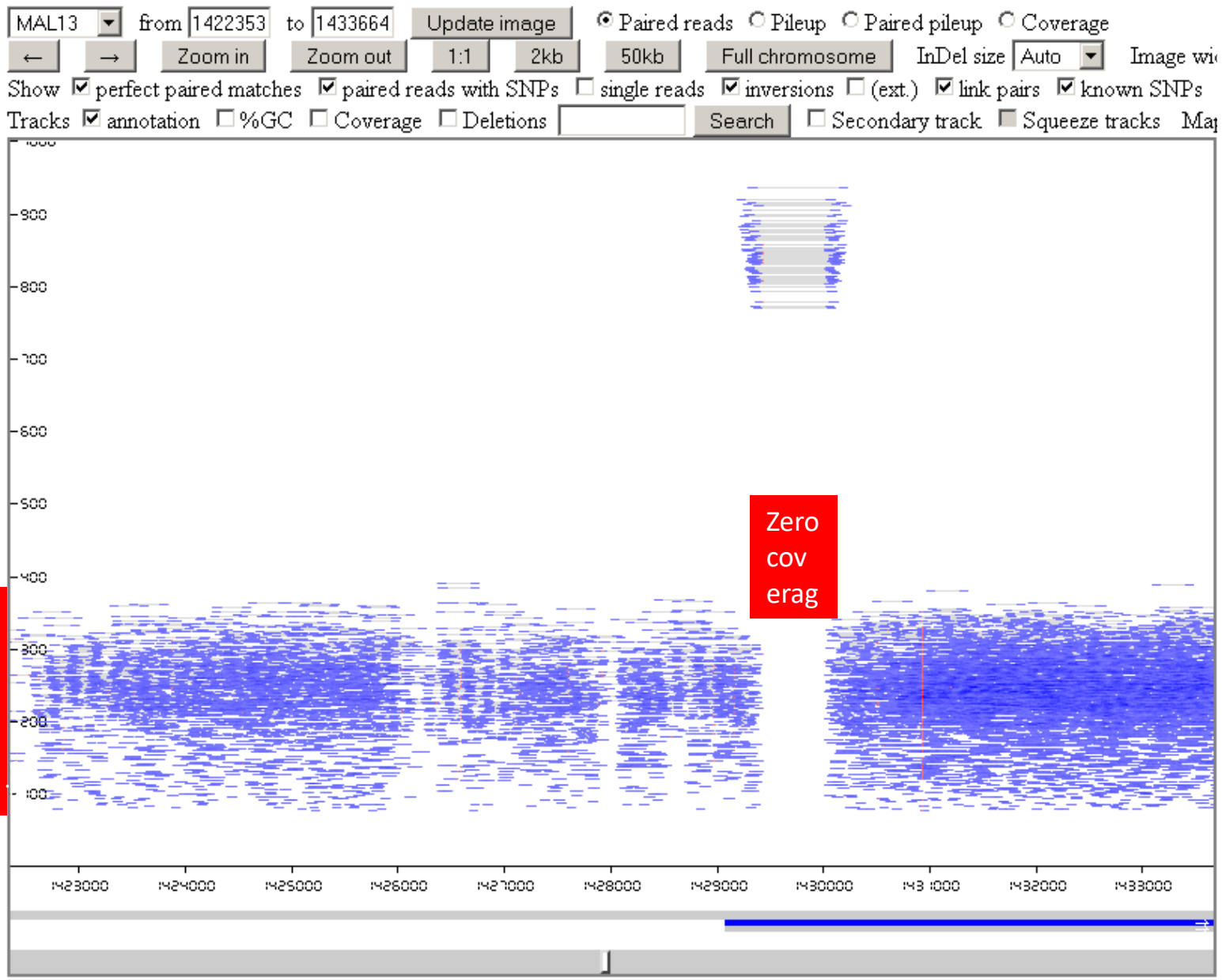
Position

Quality

VCF files generated directly BAM files using Samtools & BCF/VCFtools

Using paired end mapping

P. falciparum



Expected mapping span

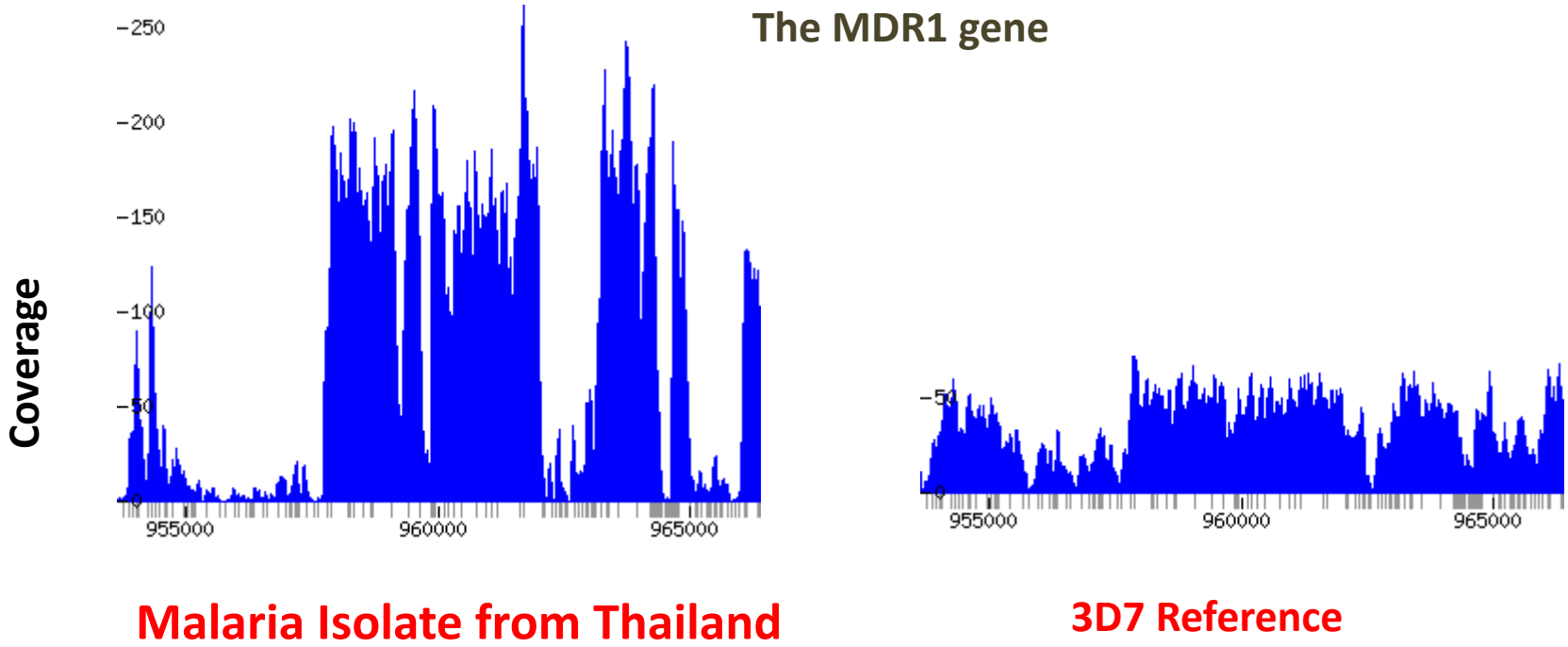
Zero cov erag

Deletion in Pfrbp2b gene

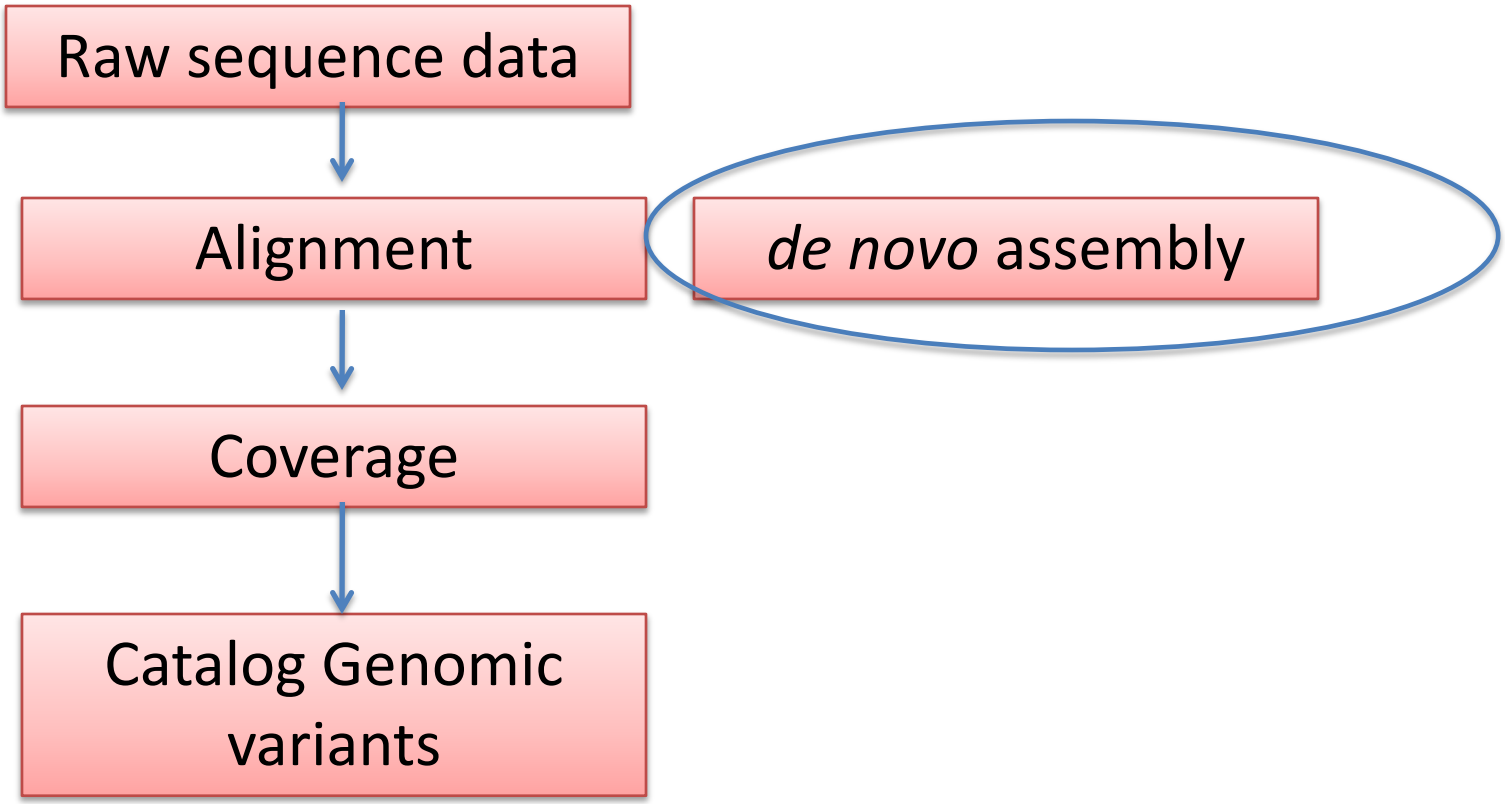
Using Read depth

- Low or zero coverage may indicate the presence of a deletion
- Excess coverage may imply duplications

(Yoon et al, 2009; Boeva et al, 2011; Sepulveda et al, 2012)



Course

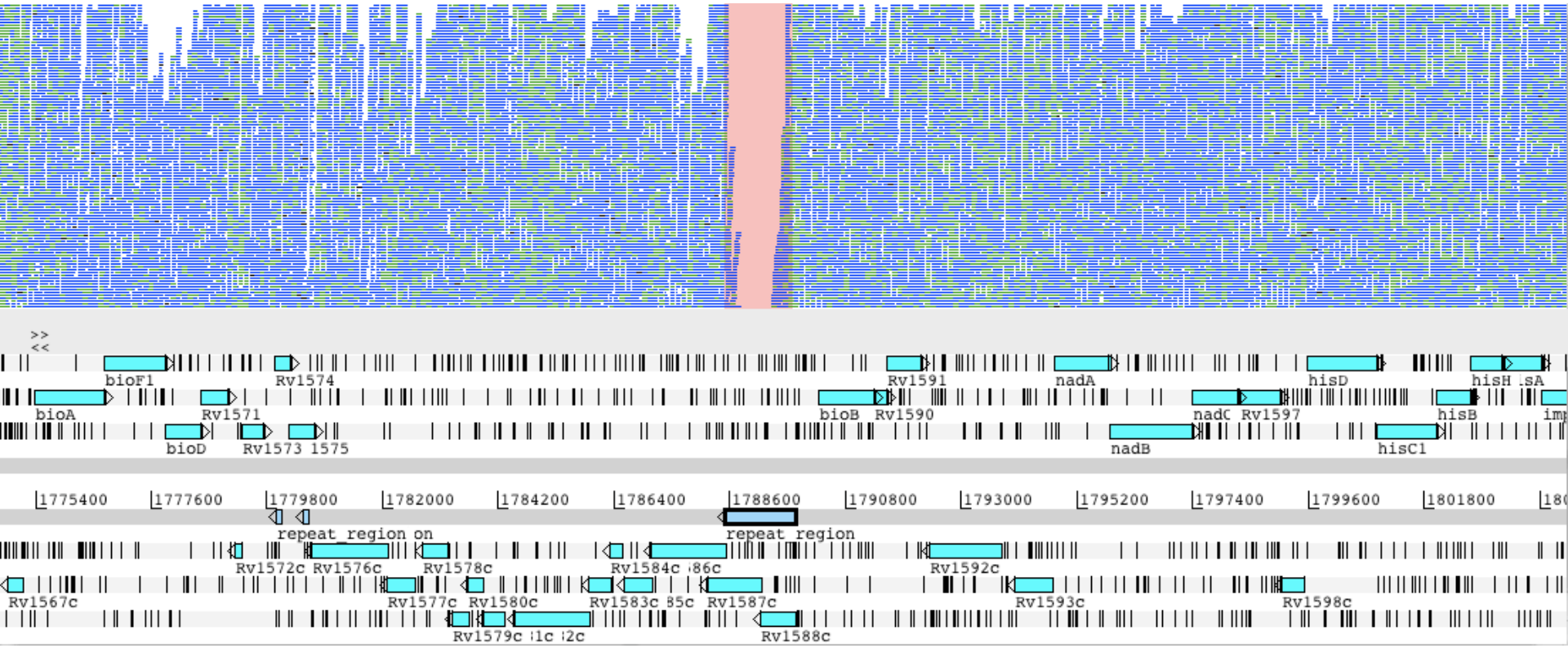

Raw sequence data
Alignment
de novo assembly
Coverage
Catalog Genomic variants

What if we cannot map to a reference?

Lisboa Mtb strain

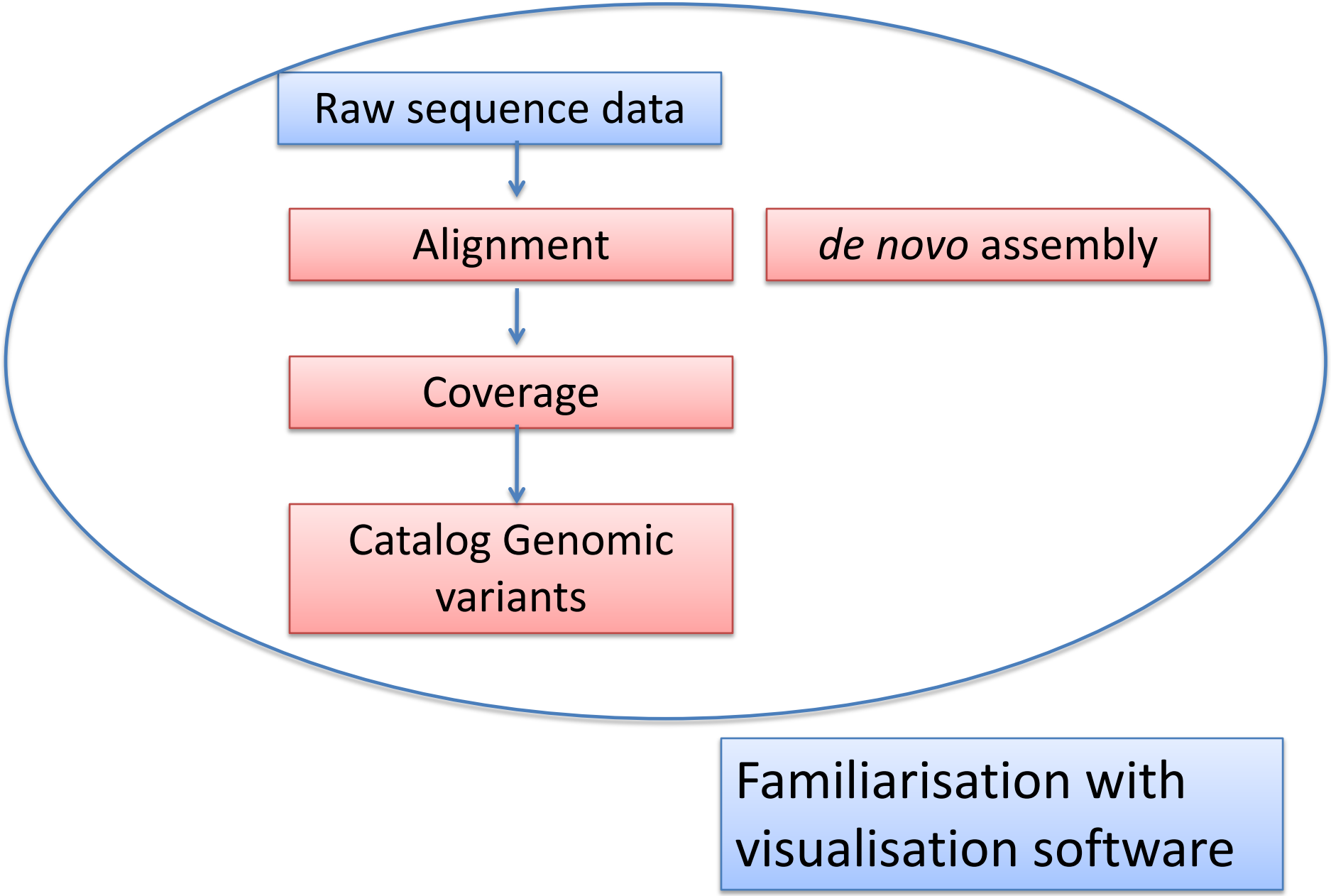
Repeat region (Rv1587/8c)

Artemis software



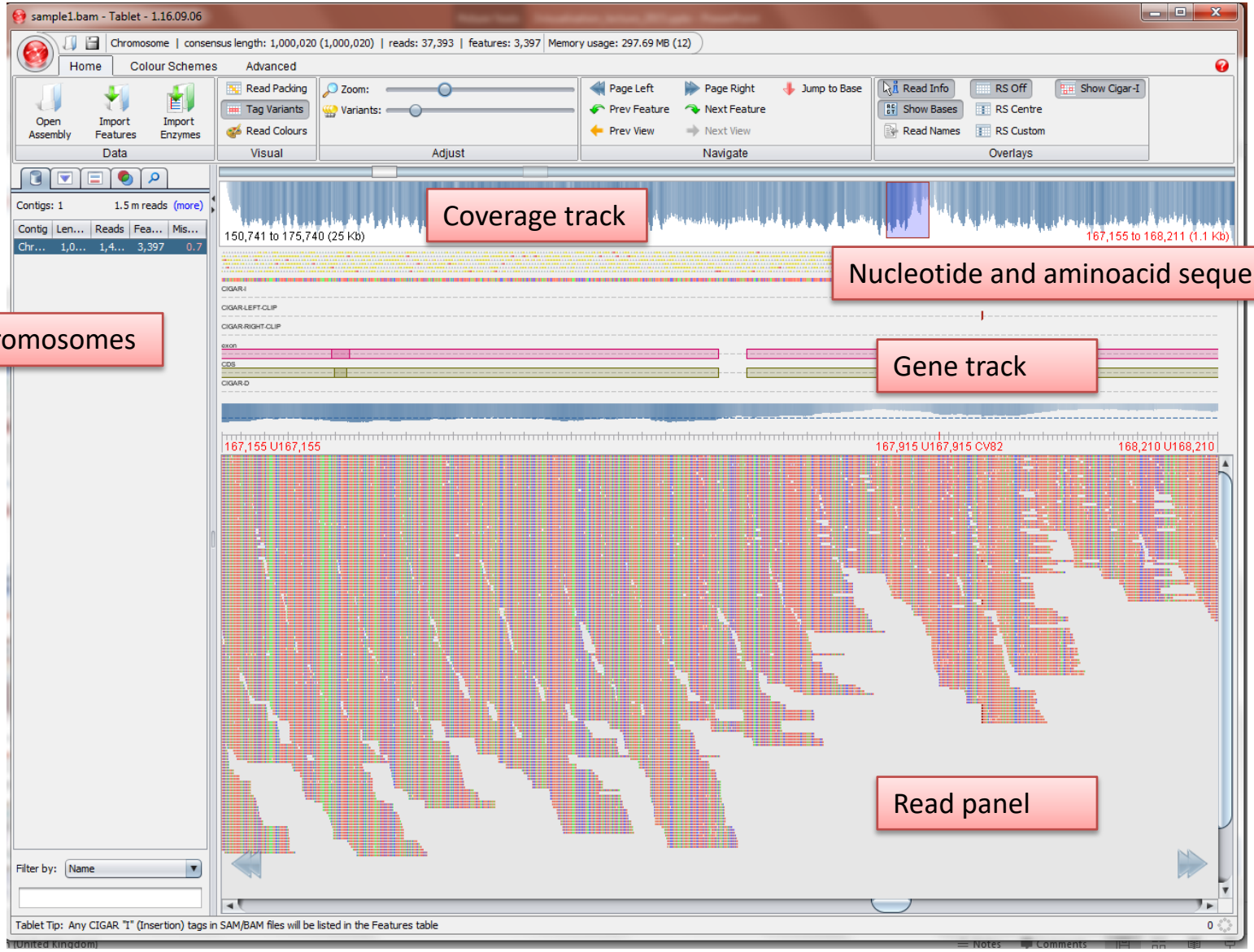
An answer: *De novo or* reference-free assembly

Course



Familiarisation with visualisation software

Tablet



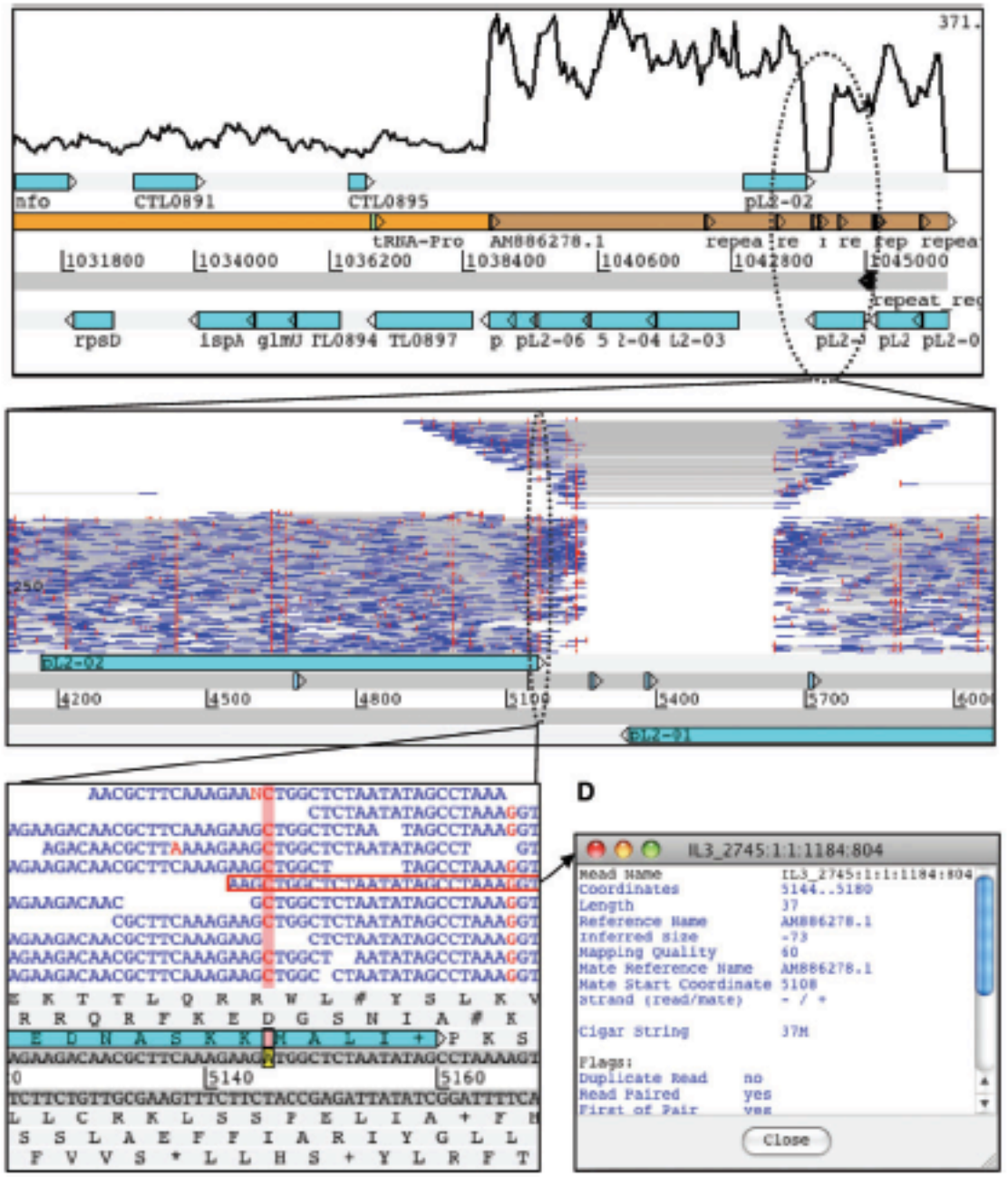
Milne et al, 2010

Artemis

Coverage by region

Inferred mapping distance to visualise indels

Mapped read information



Carver et al, 2010