Next generation sequencing: assembly by mapping reads

Laurent Falquet, Vital-IT
Helsinki, June 3, 2010

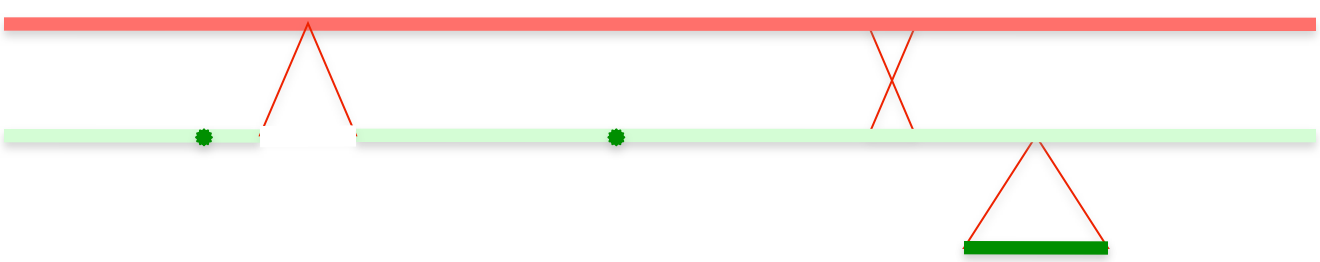
Overview

- What is assembly by mapping?
- Methods
  - BWT
- File formats
- Tools
- Issues
- Visualization
- Discussion

Mapping = alignment of reads on a reference mostly for Ultra High Throughput (re)Sequencing

- Simpler by mapping reads onto an existing genome
  - User must select the most appropriate reference
  - Success depends on the degree of similarity of the reference

- Variations detectable: SNPs and deletions
- Variations difficult to guess: insertions and inversions

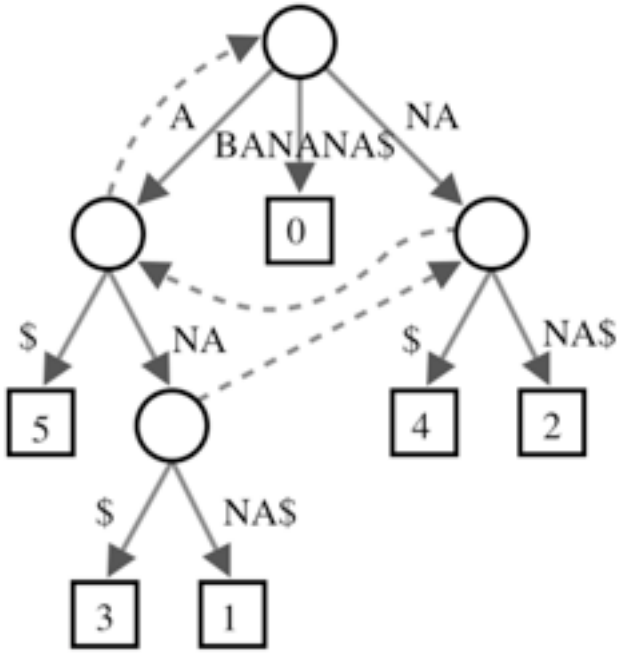


Mapping methods

- By sequence comparison with Smith-Waterman
  - much too slow
- By sequence indexing (e.g., BLAST or BLAT)
  - Conventional tools like Blast or Blat do not work well with short sequence reads.
  - > Modification of existing alignment algorithms to handle short reads.
- Indexing methods
  - Suffix tree
  - Suffix array
  - Seed hash tables
  - BWT (Burrows-Wheeler Transform)

Suffix tree

- The suffix tree for a string S is a tree whose edges are labelled with strings. Suffix trees also provided one of the first linear-time solutions for the longest common substring problem. These speedups come at a cost: storing a string's suffix tree typically requires significantly more space than storing the string itself.



35Gb for the human genome

Suffix array

- Consider the string BANANA$ of length 7. It has 7 suffixes:

index	suffix
0	BANANA$
1	ANANA$
2	NANA$
3	ANA$
4	NA$
5	A$
6	$

sort →

index	suffix
6	$
5	A$
3	ANA$
1	ANANA$
0	BANANA$
4	NA$
2	NANA$

The suffix array is the array of indices: {6,5,3,1,0,4,2}

12Gb for the human genome

Seed hash table

- Given the string ACGTACGTAAG of length 10, extract all substrings length 4 (seeds) and store their starting positions.

index	seed
0,4	ACGT
1,5	CGTA
2	GTAC
3	TACG
6	GTAA
7	TAAG

sort →

index	seed
0,4	ACGT
1,5	CGTA
6	GTAA
2	GTAC
7	TAAG
3	TACG

The size of the hash table depends on the length of the seed and the complexity of the input string

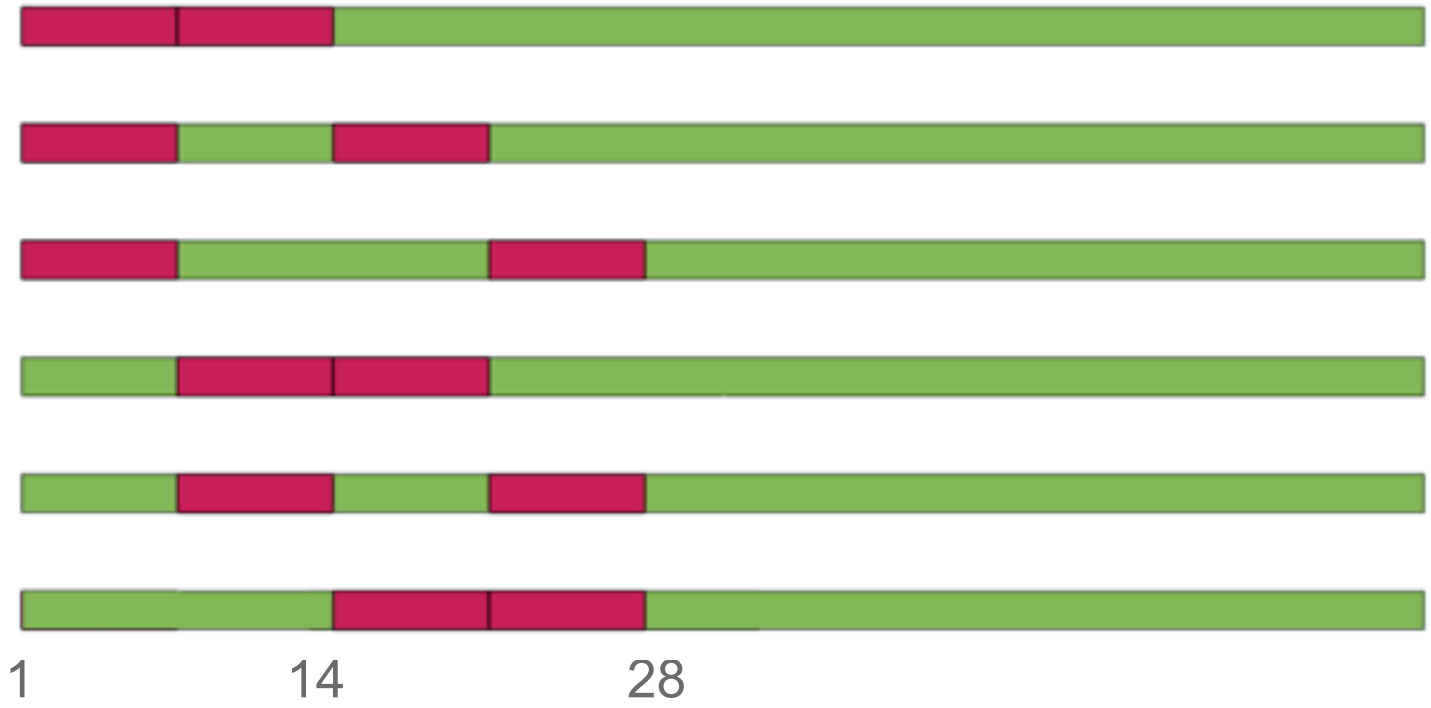
12Gb for the human genome

Seed hash table

- Hash tables can be generated according to different hash functions and using various seeds. Example for the sequence AGTGACAGT
  - Continuous seed (length 4): AGTG, GTGA, TGAC…
  - Non continuous seed (length 4): AGTG, ACAG
  - Spaced seed (length 4, weight 3, path1101): AG*G, GT*A, TG*C…
  - Periodic spaced seed: path=n*(1101)
- Hash tables have been extensively used in mapping programs.

Spaced seed hash table indexing (MAQ)

- MAQ build 6 hash tables, each indexing 14 of the first 28 bases

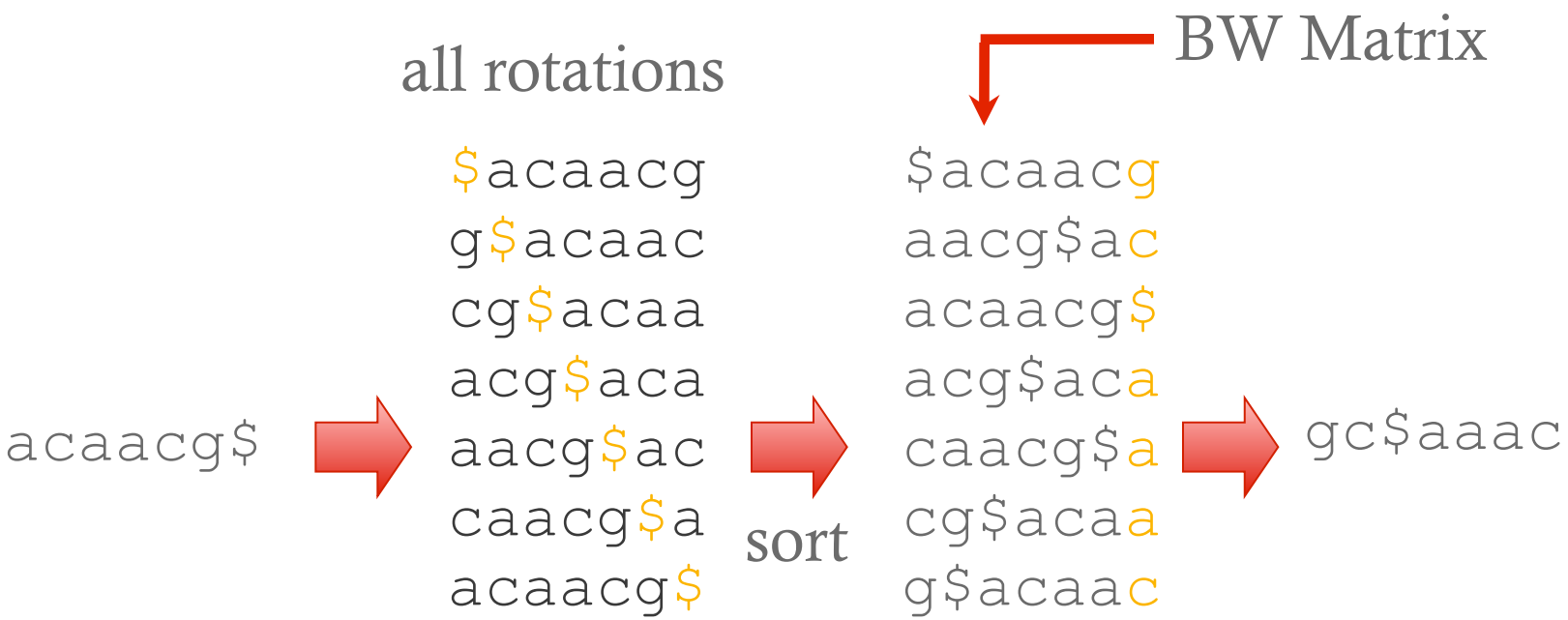


Hence, Maq finds all alignments with at most 2 mismatches in the first 28 bases.

Why Burrows-Wheeler?

- BWT very compact:
  - Approximately ½ byte per base
  - As large as the original text, plus a few “extras”
  - Can fit onto a standard computer with 2GB of memory
- Linear-time search algorithm
  - proportional to length of query for exact matches

Burrows-Wheeler Transform (BWT)



Langmead et al. 2009 Genome Biology

Burrows-Wheeler Matrix

$acaacg
aacg$ac
acaacg$
acg$aca
caacg$a
cg$acaa
g$acaac

See the hidden suffix array?

Burrows-Wheeler Transform "LF mapping" property:

The i^{th} occurrence of character X in the **L**ast column corresponds to the same text character as the i^{th} occurrence of X in the **F**irst column

acaacg$

$acaacg
aacg$ac
2nd acaacg$
acg$aca
caacg$a 2nd
cg$acaa
g$acaac

Burrows-Wheeler Transform “LF mapping” property:

Using LF the UNPERMUTE algorithm can recreate the original string

Burrows-Wheeler Transform “LF mapping” property:

Using LF the EXACTMATCH algorithm from Ferragina and Manzini can find occurrence of a substring from right to left (! greedy)

File formats

- Input
  - FASTA
  - FASTQ (various versions)
  - csFASTA
  - QSEQ
- Paired-end
  - 2 files
  - crossbow style

- Output
  - map
  - bwt
  - pileup
  - SAM
  - BAM

Example of FASTQ Illumina 1.5

```
@C3PO_0001:2:1:17:1499#0/1
TGAATTCATTGACCATAACAATCATATGCATGATGCAAATTATAATATCATTTTTAGTGACGTCGTGAATCGTTT
+C3PO_0001:2:1:17:1499#0/1
abaaaaaaaaaaa`a`aa_aaaaaaaaaaaaaaaa_a__aaa`aaaaa^aaaaa`a]^`a__YZYZ^`NJDJ\_Z
@C3PO_0001:2:1:17:1291#0/1
TGTTTGAGCAAATGATTCATAATAATGTATTTCAATATTTTTAGGAATATCTCCCAATATTGCGCGTGCTGAATT
+C3PO_0001:2:1:17:1291#0/1
a`_`_\a_aaaa_a^Z^^a[a^aa]a_^_a_``aa__`aa`X^X^^`aa_\_]VR`\a_]W\_`_a]a]][\RZV
@C3PO_0001:2:2:1452:1316#0/1
GTCCATCCGCAGCAGCGAATTTTTGACGTCCCCCCCCCGAANGGANGNGANNNNGNNGNNNTNTNNAAANGNNNNN
+C3PO_0001:2:2:1452:1316#0/1
_U__a\__`]_`ZP\\_Z^[]aa^a_]XNBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
…
```

Warning: various FASTQ formats…

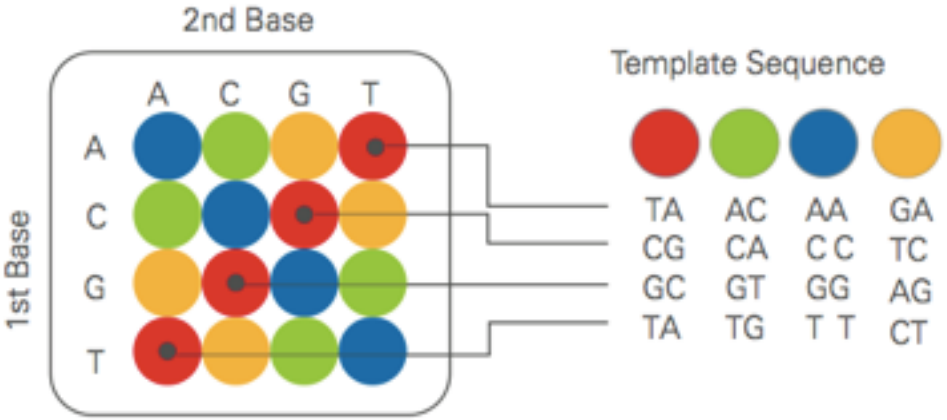
```
SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.....................
...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII.....................
.................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....................
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
|                         |    |        |                              |                     |
33                                     59   64       73                            104
126
S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 40) with 0 and 1=unused, 2=Read Segment
Quality Control Indicator (bold)
```

http://en.wikipedia.org/wiki/FASTQ_format

SOLiD color space FASTA format

```
>1_51_64_F3
T1030103123033323320333300002112222 3
>1_51_127_F3
T2010323233203132310110100200310310 2
```

Each number can be replaced according to this table



MAQ Pileup example

…

emb|BA000018.3	36129	A	102	@...,,,..,,,,,,,,,,.,...,,.,,,..,,..,,.,,,,....,.,,,,..,,,,,,,,,.,,,,.,....,.,,,,,........,.......,,,,,,.,,,..
emb|BA000018.3	36130	A	103	@,,,..,,,,,,,,,,.,...,,.,,,..,,..,,.,,,,....,.,,,,..,,,,,,,,,,,,,,,,,,,,,,....,.,,,,,........,.....,,,,,.,,,......
emb|BA000018.3	36131	T	100	@,,,,,,,,,.,...,.,,,..,,..,,.,,,,....,.,,,,..,,,,,,,,,,,,,,,,,,,,,....,.,,,,,,..........g.,,,,,.,,,.......,.
emb|BA000018.3	36132	T	93	@,.,....,.,,,,,..,,..,,.,,,,....,.,,,,..,,,,,,,,,,,,,,,,,....,,,,,,.........,.......,,,,,.,,,,,,,,......,..
emb|BA000018.3	36133	A	95	@,...,.,,,..,,..,,.,,,,....,.,,,,..,,,,,,,,,,,,,,,,,....,.,,,,,........,.......,,,,,.,,,.,,,,......,.,.,,,,
emb|BA000018.3	36134	G	98	@,...,.,,,..,,..,,.,,,,....,.,,,,..,,,,,,,,,,,,,,,,,,....,.,,,,,........,.......,,,,,.,,,.,,,......,.,,,,,,...
emb|BA000018.3	36135	T	99	@,...,.,,,..,,..,,.,,,,....,.,,,,..,,,,,,,,,,,,,,,,,....,.,,,,,...........,......G,G,,.,,,......,.,,,,,,....,
emb|BA000018.3	36136	C	97	@,.,,,..,,..,,.,,,,....,.,,,,..,,,,,,,,,,,,,,,t,....,.,,,,,........,.......,,,,,.,,,.,,,,......,.,,,,,,....,,.
emb|BA000018.3	36137	T	96	@,,,...,.,,,....,,,,,.,,,,,,,,.,,,,.,....,.,,,,,........,.......,,,,,,.,,,.,,,,......,.,,,,,,.....,,,,..,,,
emb|BA000018.3	36138	A	96	@,,..,,,,....,,,,,.,,,,,,,,.,,,,.,....,.,,,,,........,.......,,,,,.,,,.,,,,.,,,......,.,,,,,,.....,,,,,,....
emb|BA000018.3	36139	T	93	@,.,,,....,.,,,..,,,,,.,,,,,,,,.,,,....,.,,,,,........,.......,,,,,.,,,,,,,,.......,.,,,,,,,,.....,,,,,,....
emb|BA000018.3	36140	C	94	@,,,....,.,,,,,..,,,,,.,,,,,,,,.,,,....,.,,,,,........,.......,,,,,.,,,,,,,,.......,.,,,,,,,,......,,,,,..
emb|BA000018.3	36141	A	97	@,,,....,,,,,,..,,,,,.,,,,,,,,.,,,....,.,,,,,........,.......,,,,,.,,,,,,,,.......,.,,,,,,,,.,,,,,......,,,,.
emb|BA000018.3	36142	A	100	@....,.,,,..,,,,,,.,,,,.,....,,,,,,,,.,,,,,........,.......,,,,,,.,,,,,,,,.......,.,,,,,,,,.,,,,,....,.,,,,,,....
emb|BA000018.3	36143	A	102	@...,.,,,,..,,,,,,.,,,,.,....,.,,,,........,.......,,,,,,.,C,......,..,,,,...,,,,.,,,,,....,.,,.,,.....,,..
emb|BA000018.3	36144	A	102	@,.,,,,...,,,,,,,.,,,,.,....,.,,,,..........,.......,,,,,.,,,,,,,,.......,.,,,,,,,,.,,,,,.......,.,,,,.....,,,.
emb|BA000018.3	36145	G	102	@TTTTttTTTTTTTTtTTTTtTttttTtTTTTttttttttttTttttTTTTTtTTTTtttttttTttTTTTtttTTtTTTTtttttTttTTtTTtttttTttTTTtTT
emb|BA000018.3	36146	A	103	@,,..,,,,,,,,,,,.,,,,,.,....,.,,,,,............,.......,,,,,,,.,,,,,,,,,.......,.,,,,,,,,.,,,,,,,.......,.,,,,,.,,,,,,
emb|BA000018.3	36147	A	105	@,,..,,,,,,,,,,,.,,,,,....,.,,,,,............,.......,,,,,,,.,,,,,,,,,.......,.,,,,,,,,.,,,,,,,.,,,,g.,..,,.,,,,,,,
emb|BA000018.3	36148	A	108	@,..,,,,,,,.,,,,.,....,.,,,,,,............,C,,.,,,......,.,,,,,..,,,,,,,,,,,t.,,.,,,.....,..,,,.,,,,,,,,,,,.
emb|BA000018.3	36149	G	110	@..,,,,,,,,.,,,,.,....,.,,,,,.........c,........,,,,,,,,.......,.,,,,,..,,,,,,,,,,,,.,,,,.....,..,,,.,,,,,,,,,.,..
emb|BA000018.3	36150	G	113	@,,,,,,,,,,,,.,....,.,,,,,...............,,,,,,,,......,.,,,,,..,,,,,,,,,,,,.,,,,.....,..,,,.,,,,,,,,,,,.,,,,,....,,,....
emb|BA000018.3	36151	G	109	@,.,,,,.,....,.,,,,,............,,,,,,,,,.......,.,,,,,..,,,,,,,,,,,,.,,,,.....,..,,,.,,,,,,,,,,,.,,,,,..,.....,..
emb|BA000018.3	36152	G	110	@,,,,.,....,.,,,,,............,,,,,,,,,.......,.,,,,,..,,,,,,,,,,,,.,,,,.....,..,,,.,,,,,,,,,,,.,,,,,,.......,..,,.
emb|BA000018.3	36153	T	111	@,.,....,.,,,,,............,,,,,,,,,,.......,..,,,,,..C.,,,,,,,,,,...,,,,,.....,..,,,.,,,,,,,,,,,.,,,,,,.....,,,,....
emb|BA000018.3	36154	T	110	@,....,.,,,,,............,,,,,,,,,,.......,..,,,,,..,,,,,,,,,,,,....,,,,,.....,..,,,.,,,,,,,,,,,.,,,,,,......,,,,...,
emb|BA000018.3	36155	G	111	@,.,,,,........,....,,,,,,,,,,.......,..,,,,,..,,,,,,,,,,,,....,,,,,.....,..,,,.,,,,,,,,,,,.,,,,,,.......,,,,,,,,,....
emb|BA000018.3	36156	G	116	@,,,,,,........,....,,,,,,,,,,.......,..,,,,,..,,,,,,,,,,,,....,,,,,.....,..,,,.,,,,,,,,,,,.,,,,,,.......,,,,,,,,,,......
emb|BA000018.3	36157	G	112	@.........,,,,,,,,,,.......,..,,,,,..,,,,,,,,,,,,....,,,,,.....t.,,,,,,,,,,,,,.,,,,,,.......,,,,,,,,,,,,.....

SAM/BAM formats

- Here is an example of an SAM file:

```
@HD VN:1.0
@SQ SN:chr20 LN:62435964
@RG ID:L1 PU:SC_1_10 LB:SC_1 SM:NA12891
@RG ID:L2 PU:SC_2_12 LB:SC_2 SM:NA12891
read_28833_29006_6945 99  chr20 28833 20 10M1D25M = 28993 195 AGCTTAGCTAGCTACCTATATCTTGGTCTTGGCCG
     <<<<<<<<<<<<<<<<<<<<:<9/,&,22;;<<< NM:i:1 RG:Z:L1
read_28701_28881_323b 147 chr20 28834 30 35M      = 28701 -168 ACCTATATCTTGGCCTTGGCCGATGCGGCCTTGCA
     <<<<<;<<<<7;:<<<6;<<<<<<<<<<<7<<<< MF:i:18 RG:Z:L2
```

- BAM is the binary version of the same data

http://samtools.sourceforge.net/SAM1.pdf

Mapping Tools list (non-exhaustive)

Tool	Open Source	Max read Length	Algorithm	SOLiD colorspace?
BFAST	Yes		Hash genome + SW	Yes
Bowtie	Yes		BWT	No
BWA	Yes	200 + more	BWT + SW	Yes
ELAND	Com.		Hash reads	No
MAQ	Yes	127	Hash reads	Yes
Mosaik	Yes		Hash genome + SW	Yes
Novoalign	Com.		Hash reads	No
RMAP	Yes	64	Hash reads	No
SHRiMP	Yes		Hash reads + SW	Yes
SOAP2	No	60	2way BWT	No
Zoom!	Com.	240	Hash reads	Yes

MAQ

```
#index the reference
maq fasta2bfa reference.fa genomeref.bfa
#index the reads
maq fastq2bfq S6out_1.fastq S6out_1.bfq
maq fastq2bfq S6out_2.fastq S6out_2.bfq
#map the reads in paired-end
maq map -a 600 -1 36 -2 36 S6.map genomeref.bfa S6out_1.bfq S6out_2.bfq
# get the consensus
maq assemble -p S6.cns genomeref.bfa S6.map
maq cns2fq S6.cns > S6consensus.fq  ## warning need to convert to fasta
### Find SNPs
maq cns2snp S6.cns > S6.snp
## filter quality
maq.pl SNPfilter -d 50 -w 20 -q 40 S6.snp > S6.fil.snp
### pileup the reads
maq pileup -p -m 2 genomeref.bfa S6.map > S6.pileup
```

Bowtie

```
# index reference
bowtie-build -f reference.fa Saureus
# map reads
bowtie -n 1 -l 36 -I 400 -X 700 –un unmapped -p 10 Saureus -1
    s_1_1_sequence.txt -2 s_1_2_sequence.txt > mybest.bwtmap

# convert to SAM & BAM
# index reference
samtools faidx reference.fa
# convert bowtie to SAM
bowtie2sam.pl mybest.bwtmap > mybest.sam
# convert SAM to BAM
samtools view -T reference.fa -b mybest.sam > mybest.bam
# sort BAM
samtools sort mybest.bam mybest.sorted.bam
# index BAM
samtools index mybest.sorted.bam
```

BWA

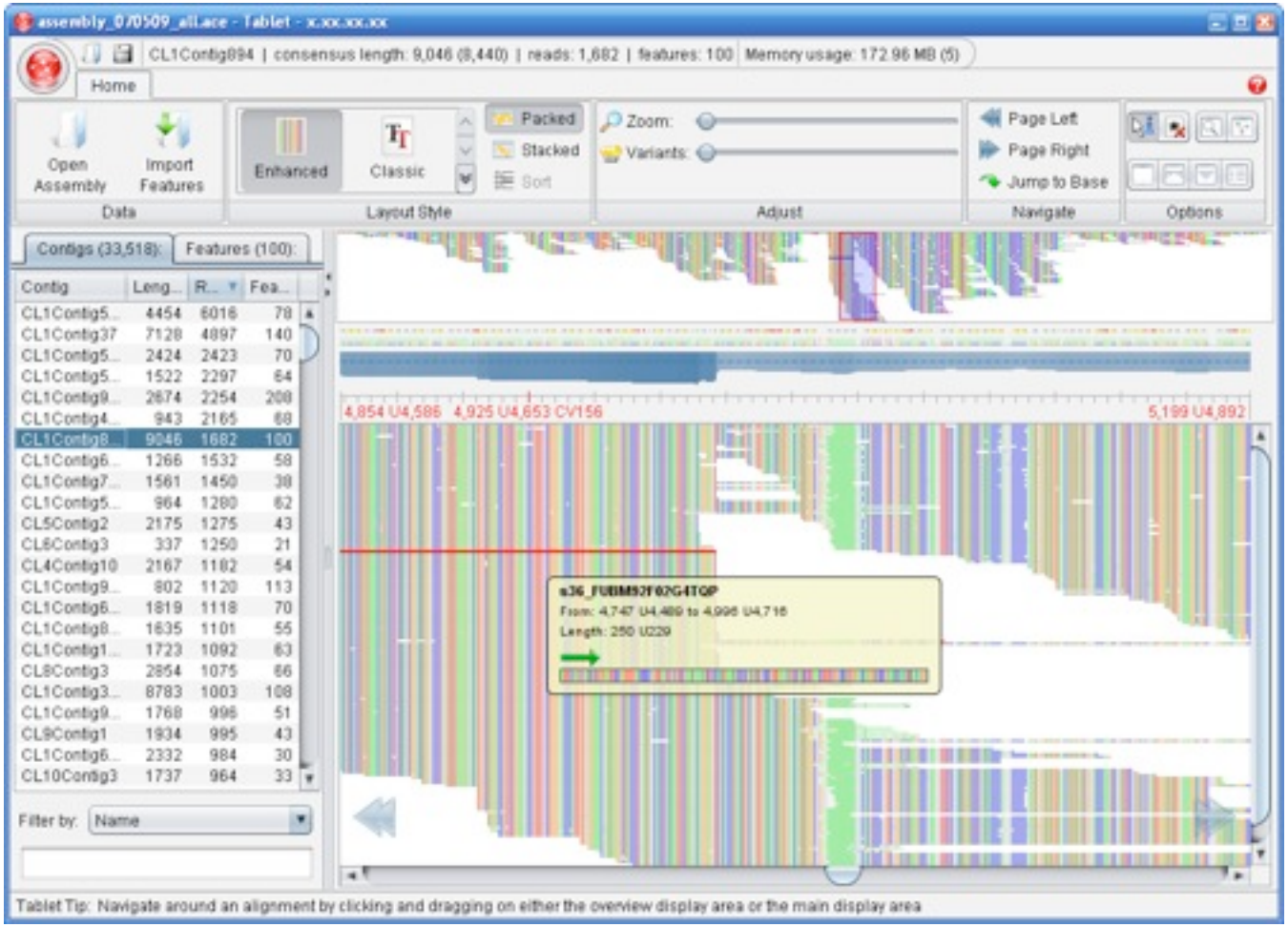
```
# index reference
bwa index reference.fa
# index reads
bwa aln -t 4 mybest.fa ../saureus_1.fq > saureus_1.sai
bwa aln -t 4 mybest.fa ../saureus_2.fq > saureus_2.sai
# map reads
bwa sampe -a 600 -P reference.fa saureus_1.sai saureus_2.sai ../
   saureus_1.fq ../saureus_2.fq > mybest.sam

# index reference
samtools faidx reference.fa
# convert SAM to BAM
samtools view -T reference.fa -b mybest.sam > mybest.bam
# sort BAM
samtools sort mybest.bam mybest.sorted.bam
# index BAM
samtools index mybest.sorted.bam
```

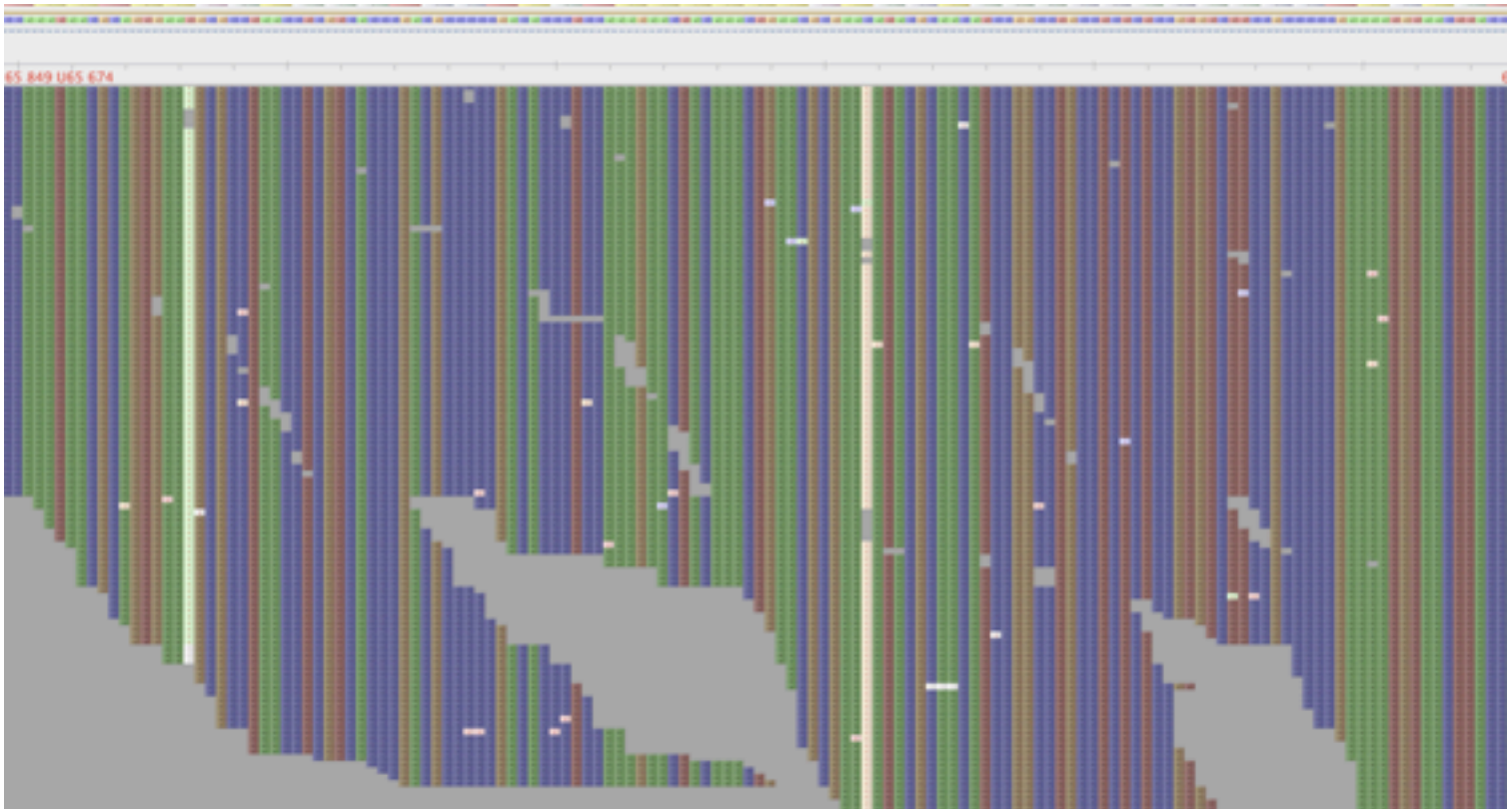
Visualization tools for assemblies

Tool	Windows	Linux	Mac	Input format
BAMview	Y	Y	Y	BAM
Consed/Gap5	N	Y (X11)	Y (X11)	ACE, MAQ, BAM
Eagleview	Y	Y	Y	ACE
Gambit	Y	Y	Y	BAM
Hawkeye	Y (cigwin)	Y	(Y)	afg (AMOS)
IGVviewer	Y	Y	Y	BAM, SAM, …
Tablet	Y	Y	Y	ACE, MAQ, BAM, afg, SAM, …

Tablet interface

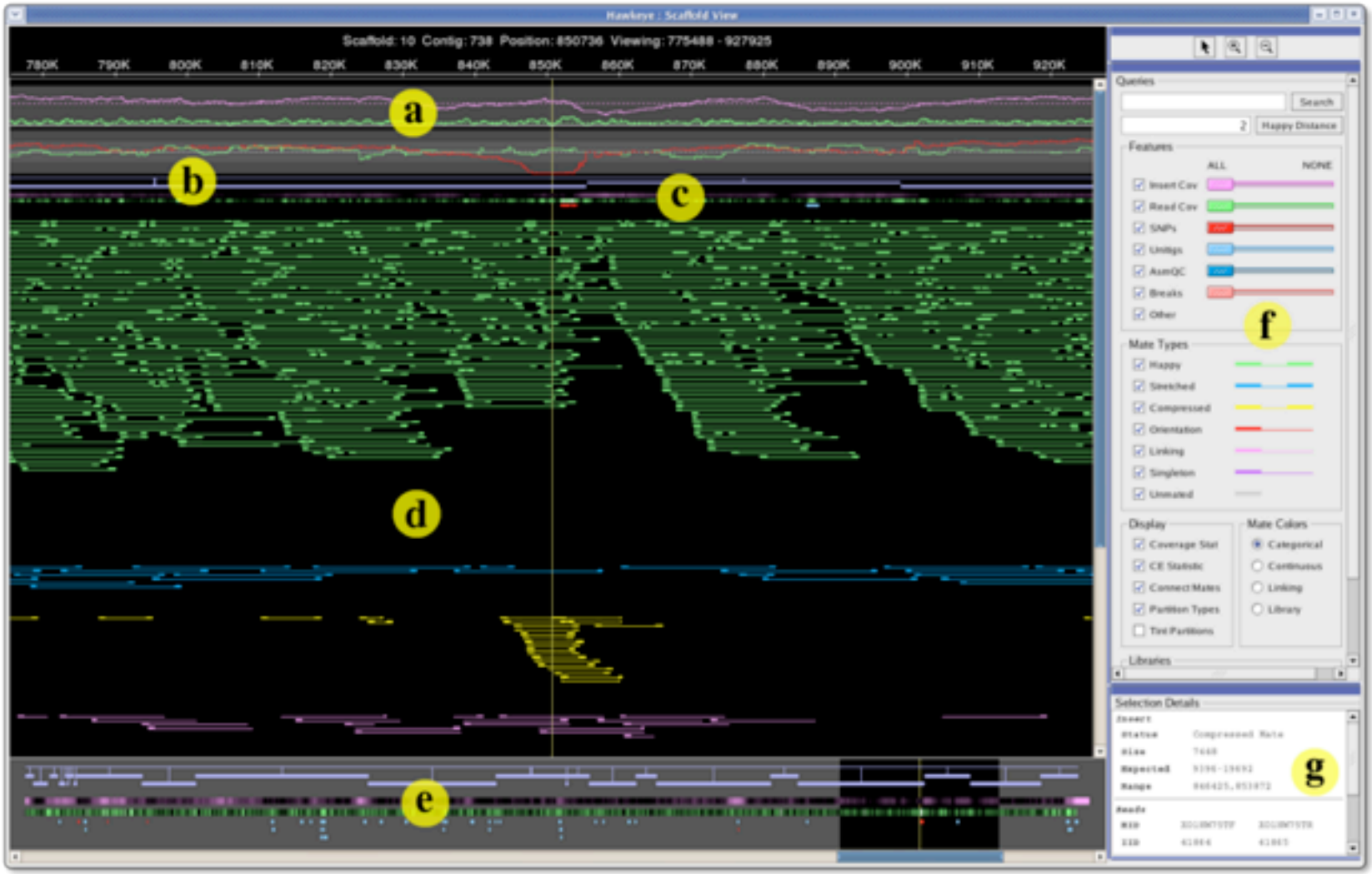


Tablet
visualization of the mapping and the SNPs

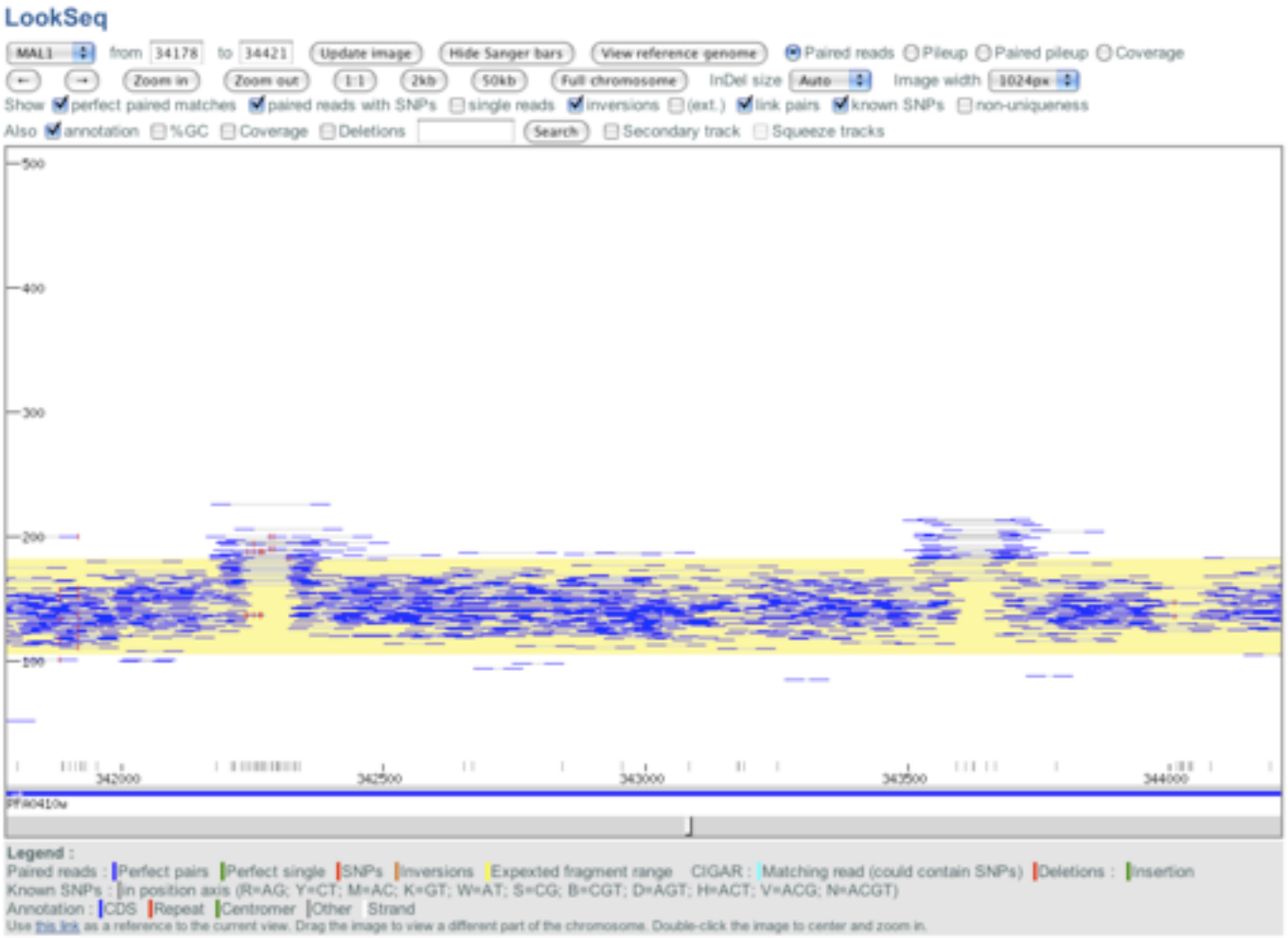


Mapping of the reads of a Staphylococcus aureus sequencing, showing 2 SNPs vs the reference genome.

Hawkeye scaffold view



Best view: LookSeq



Software issues

- File formats jungle
  - Each software has its own internal formats, few comply with the emerging standards
- Often single threads
  - Few of the software are multithreaded
- Difficult to identify insertions/deletions/inversions
- Unfinished beta software or not maintained
- Poor visualization tools

The practicals

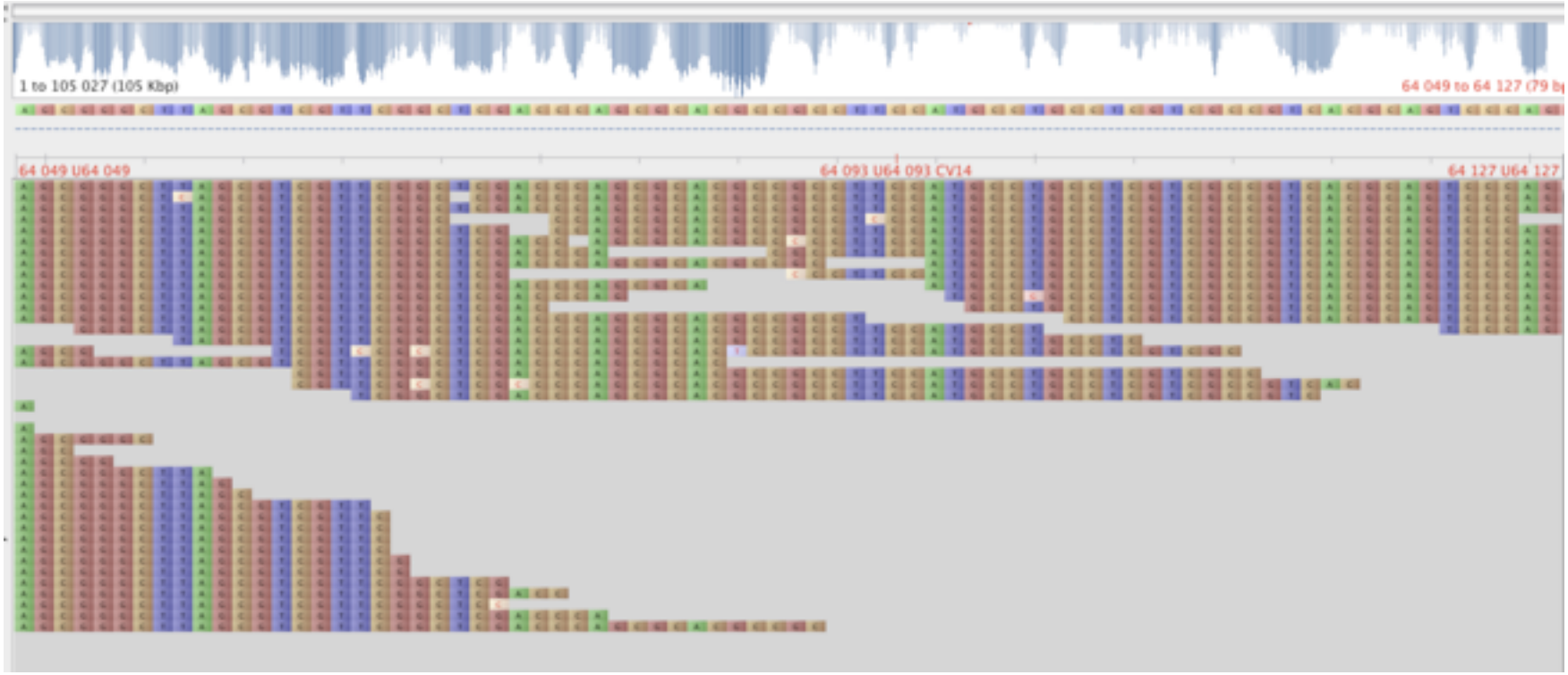
- http://edu.isb-sib.ch

- select «workshops» on the left menu
- select «Helsinki NGS workshop» at the top
- Enrol yourself with the key «EMBRACE2010»

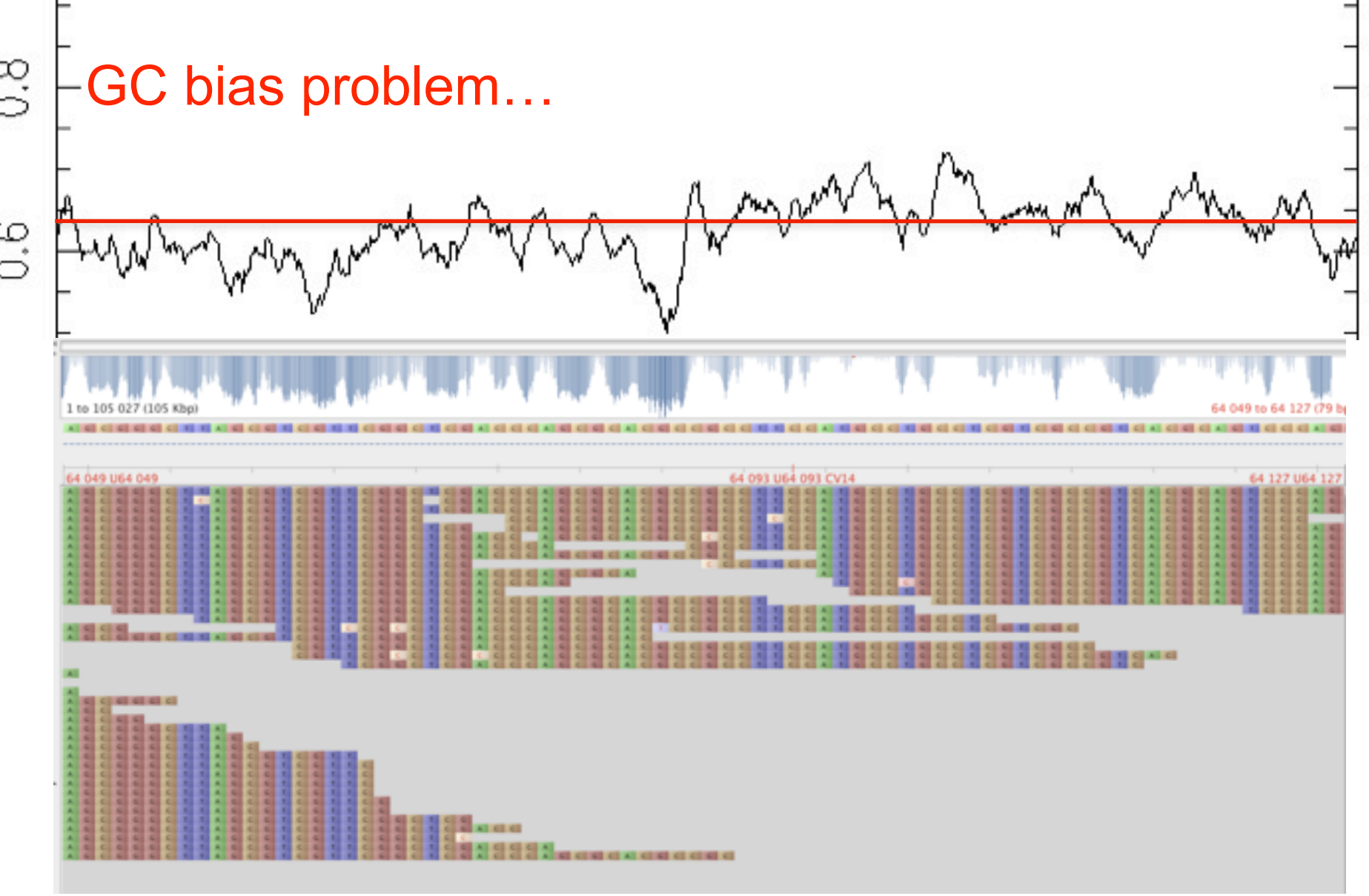
- Login to hippu server, then follow the instructions of the exercises

Thank You

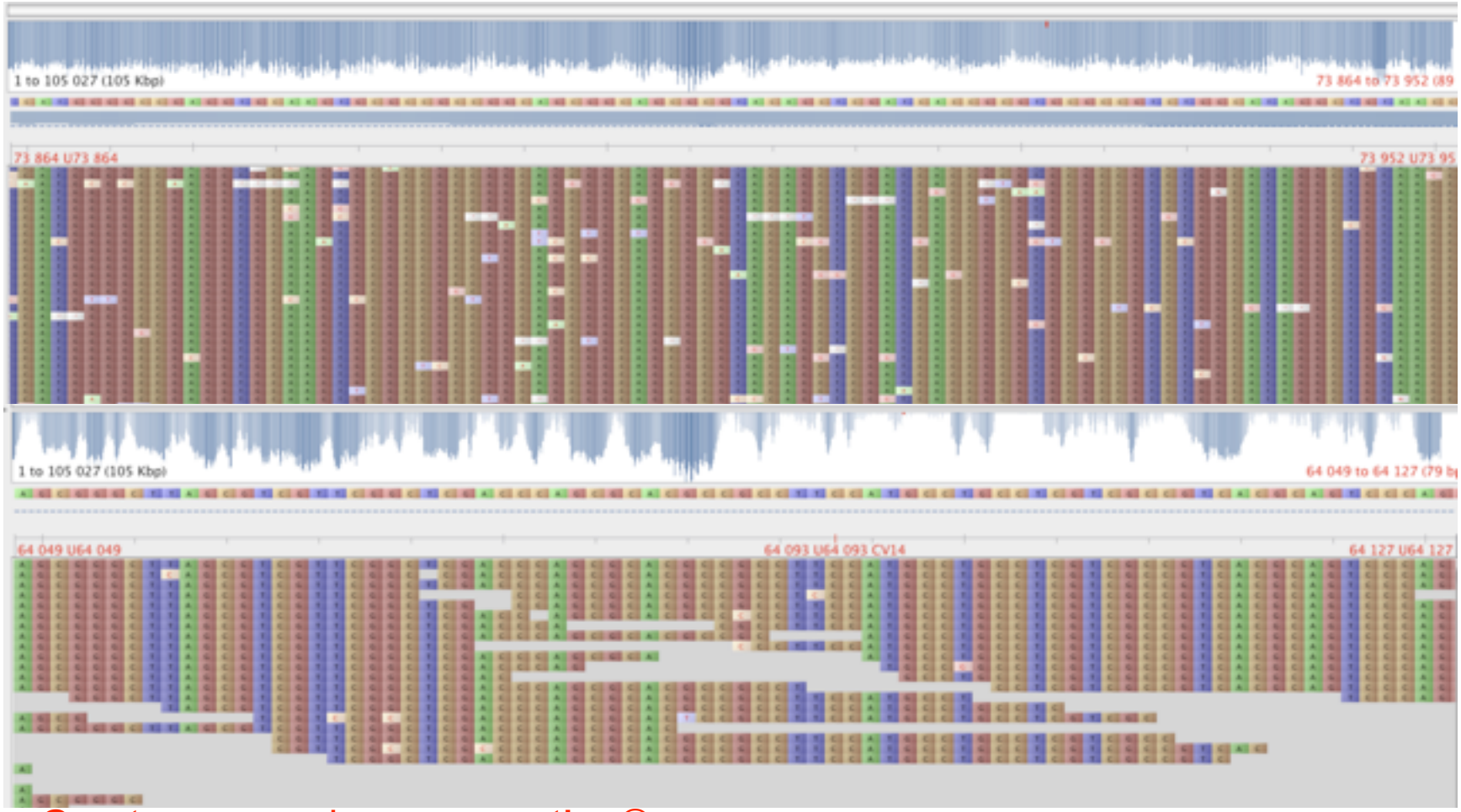
Coverage problem ? (mapping on genomic island ICEclc 100kbp)



GC bias problem…



New sequencing coverage of CLC 100kbp for P.knackmussii 2762 (top) vs B13 (bottom)



Great coverage improvement!... ☺

Application of NGS

- Genome sequencing (denovo and resequencing)
- Transcriptome (RNAseq)
- ChIPseq
- Metagenomics
- Genotyping
- Comparative genomics
- Systems Biology
- …