Experience in assembling genomes with short reads data

Laurent Falquet, Vital-IT
Helsinki, June 2, 2010

Overview

- The Vital-IT team and infrastructure
- Little reminder on Genome Assembly
- Limitations of the techniques
- Limitations of the sequences
- Some examples:
  - *S.aureus* (*de novo* & by mapping)
  - *P.knackmussii* (*de novo*)
  - *S.invicta* (*de novo*)
- Discussion

The assembly team at Vital-IT (www.vital-it.ch)

Oksana Riba-Grognuz, Ariadna Rodriguez, Laurent Falquet, Sandra Calderon
+ Vital-IT and many collaborators:

UNIBE: Joachim Frei, Edi Viley,
UNIL: Nicolas Fasel, Laurent Keller, Yannick Wurm
CHUV: Dominique Blanc, Valérie Vogel, Patrick Basset
DAFL: Keith Harshman, Emmanuel Beaudoing, Sylvain Pradervand
Fasteris, Microsynth
Many more to come… ;-)

List of projects

De novo

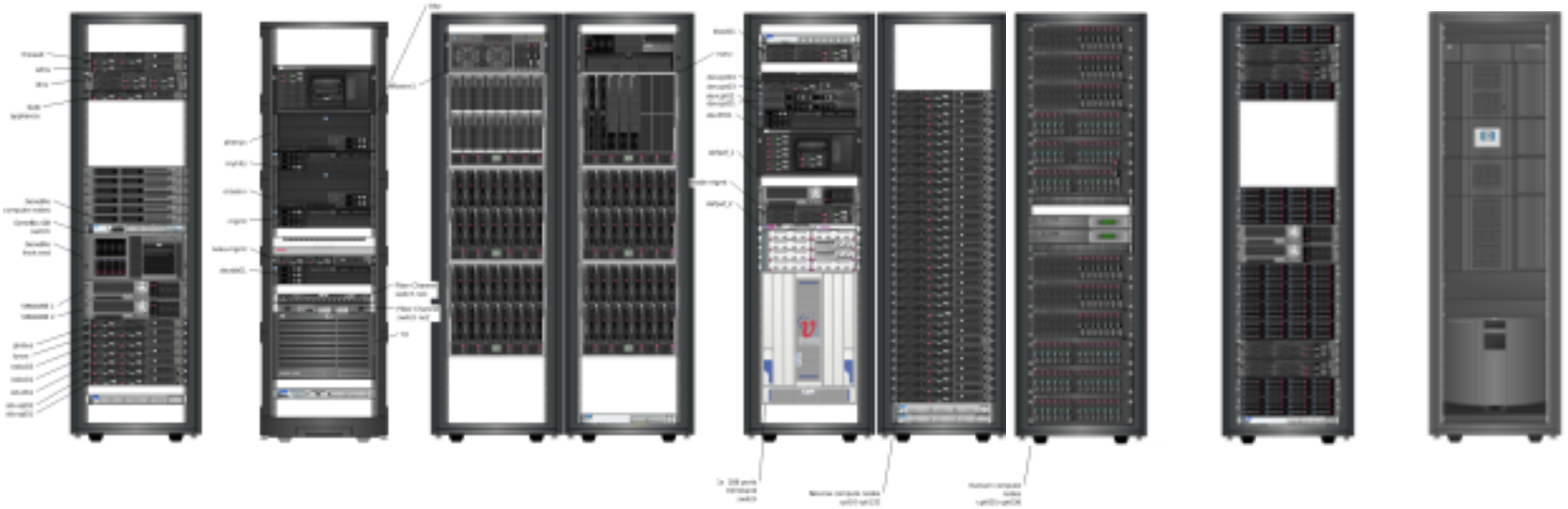
Species name	Kingdom	Size	Status
Mycoplasma conjunctivae	Prokaryote	0.9 Mbp	Finished
Mycoplasma leachii	Prokaryote	1.2 Mbp	Closing
Clostridium chauvoei	Prokaryote	3.2 Mbp	Assembling
Solenopsis invicta (Fire ant)	Eukaryote	600 Mbp	Assembling
Avibacterium paragallinarum	Prokaryote	2.5 Mbp	Assembling
Pseudomonas knackmussii	Prokaryote	6.2 Mbp	Sequencing+Assembling

By mapping

Species name	Kingdom	Size	Status
Staphylococcus aureus	Prokaryote	8 x 2.8 Mbp	Closing
Leishmania guyanensis	Eukaryote	4 x 32 Mbp	Sequencing
Mycoplasma mycoides SC	Prokaryote	2 x 1.0 Mbp	Sequencing

Vital-IT hardware

- Cluster of > 1000 nodes
- 3 dedicated machines with 8-24 CPU + 256Gb RAM
- Large storage capacity (>300Tb)



Vital-IT: Tools installed and used (or tested)

- ABySS
- Velvet
- Euler
- Edena
- Newbler
- MIRA
- WGS(Celera)
- Amos
- Mummer
- Phrap
- Cap3
- SAMtools
- BreakDancer
- SOAPdenovo
- MAQ
- Bowtie
- Mosaik
- BWA
- Eland
- SOAP
- Tagger
- Eagleview
- Hawkeye
- Tablet
- Gambit
- Consed/Gap5
- MAUVE
- Jalview
- ACT/Artemis
- Gepard
- MGCAT
- BAMviewer
- IGVviewer
- ABYSS-explorer

What are NGS short reads data?

Sequencing platform	ABI3730xl Genome Analyzer	Roche (454) FLX	Illumina Genome Analyzer	ABI SOLiD	HeliScope
Sequencing chemistry	Automated Sanger sequencing	Pyrosequencing on solid support	Sequencing-by-synthesis with reversible terminators	Sequencing by ligation	Sequencing-by-synthesis with virtual terminators
Template amplification method	In vivo amplification via cloning	Emulsion PCR	Bridge PCR	Emulsion PCR	None (single molecule)
Read length	700–900 bp	200–500 bp	36-108 bp	35-75 bp	25–55 bp
Sequencing throughput (old numbers)	0.03–0.07 Mb/h	13 Mb/h	25 Mb/h	21–28 Mb/h	83 Mb/h
Advantage by price	700 bp / $	16'000 bp / $	500'000 bp / $	1'000'000 bp / $	1'000'000 bp / $
Nr of installed machines (estimation)	??	179	603	166	10

Application of NGS

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

Ultra High Throughput Sequencing (WGS)

Figure 1. Original genomic DNA is broken into a collection of overlapping fragments

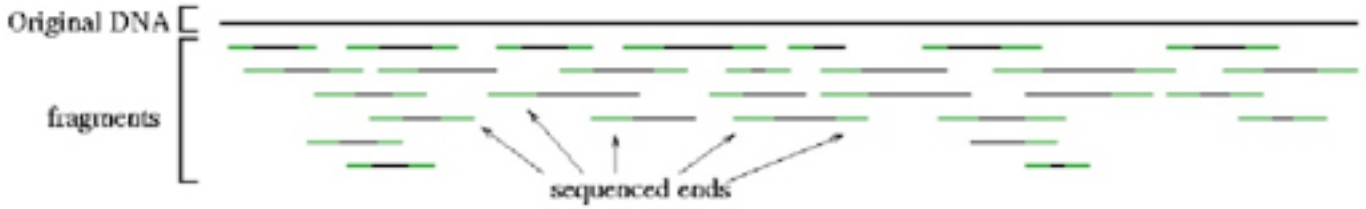


Figure 2. The ends of each fragment (drawn in green) are sequenced

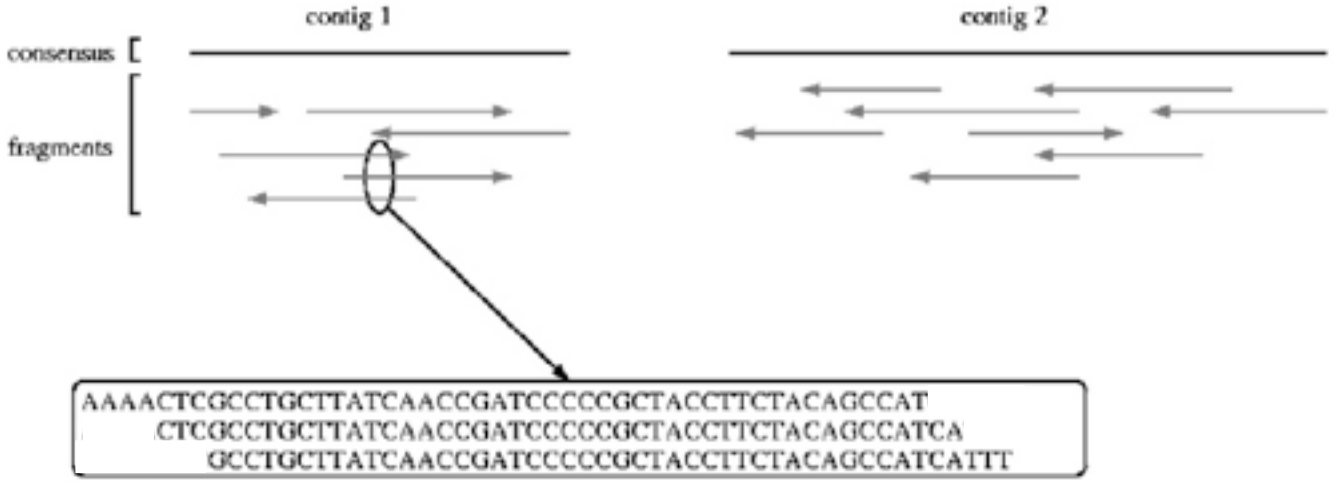
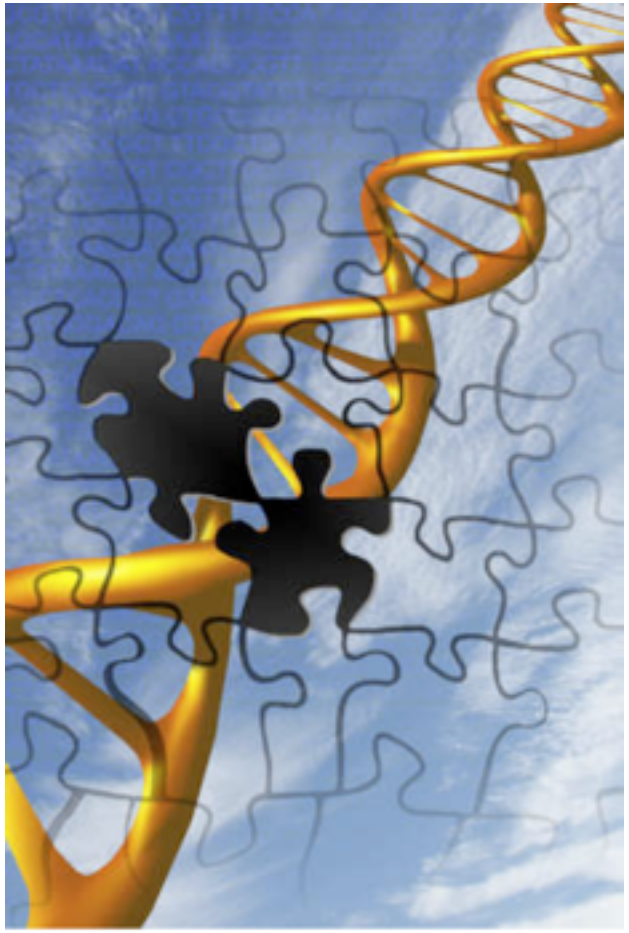


Figure 3

This is also called "Whole Genome Shotgun Sequencing"

Ultra High Throughput Sequencing and Genome Assembly: a Simple Jigsaw Puzzle?

- Yes, but you must deal with
  - Millions of pieces
  - Lots of malformed pieces
  - Often missing pieces
  - Pieces mixed from another puzzle
  - Lots of identical blue sky pieces…
  - If *de novo* you…



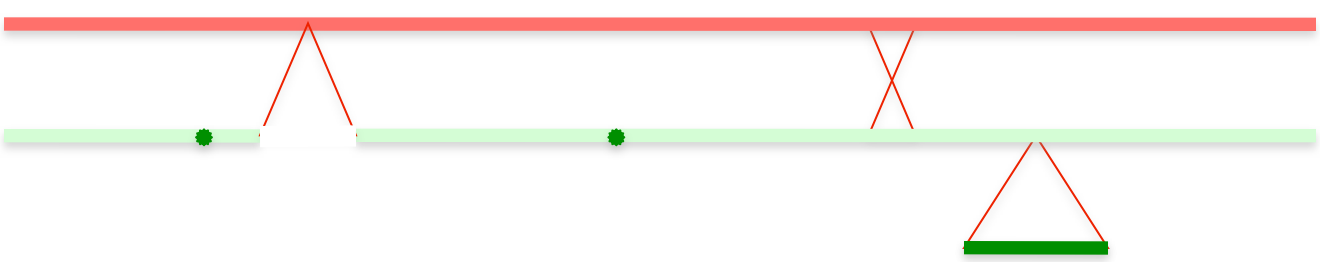
Genome assembly, deep blue…

…don’t even know the final picture…

Ultra High Throughput Sequencing and Genome Assembly: with a reference ?

- Simpler by mapping reads onto an existing genome
- Current tools are fast by using the Burrows-Wheeler Transform
- Success depends on the degree of similarity of the reference

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



Software issues

- File formats jungle
  - Each software has its own internal formats, few comply with the emerging standards (FASTQ, SAM/BAM)
- Parameters tuning
  - Several parameters must be tuned, in particular the Kmer
- Large memory requirements
  - Some software might require hundreds of Gbytes
- Often single threads
  - Few of the software are multithreaded
- Unfinished beta software
- Poor visualization

Limitations of the techniques

- Sequencing errors (all methods)
- Roche454 long (>8) mononucleotide repeats
- Illumina and SOLiD, very short reads (20-75bp)
- Missing data (sampling/coverage bias)?

Data issues

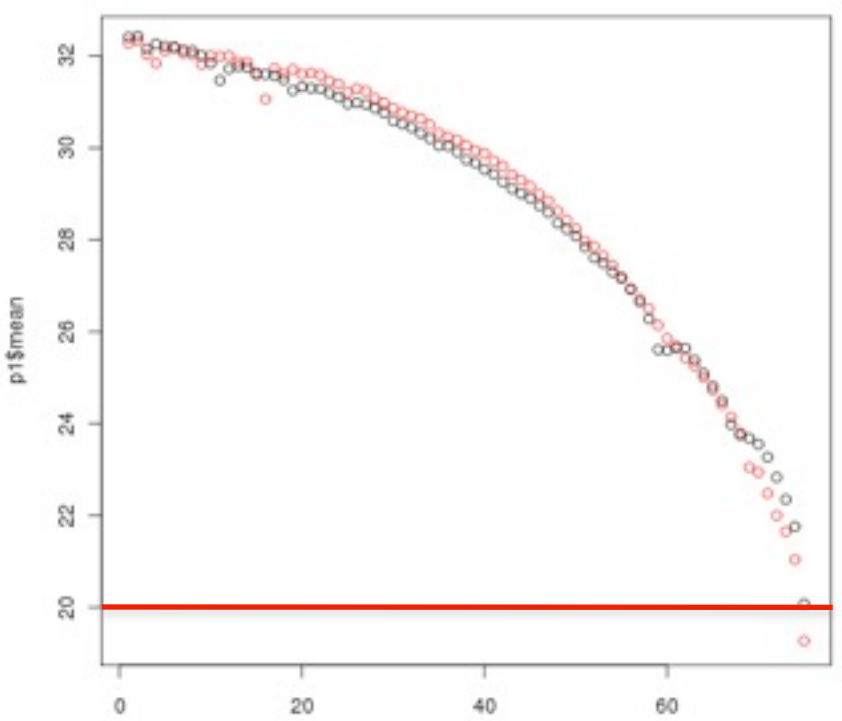
- Large data sets, raw and processed -> storage issue
- Paired-end *vs* Mate pair, interleaved *vs* 2 files
- Quality checks vary

- Filtering examples for Illumina:
  - none
  - Chastity*
  - Remove Ns or ambiguous characters
    - ABySS reads automatically removed
    - Velvet replaces the N with a random nucleotide (A,C, G or T)
  - Remove reads with less than 25 Q30 bases in the first 35 bp.

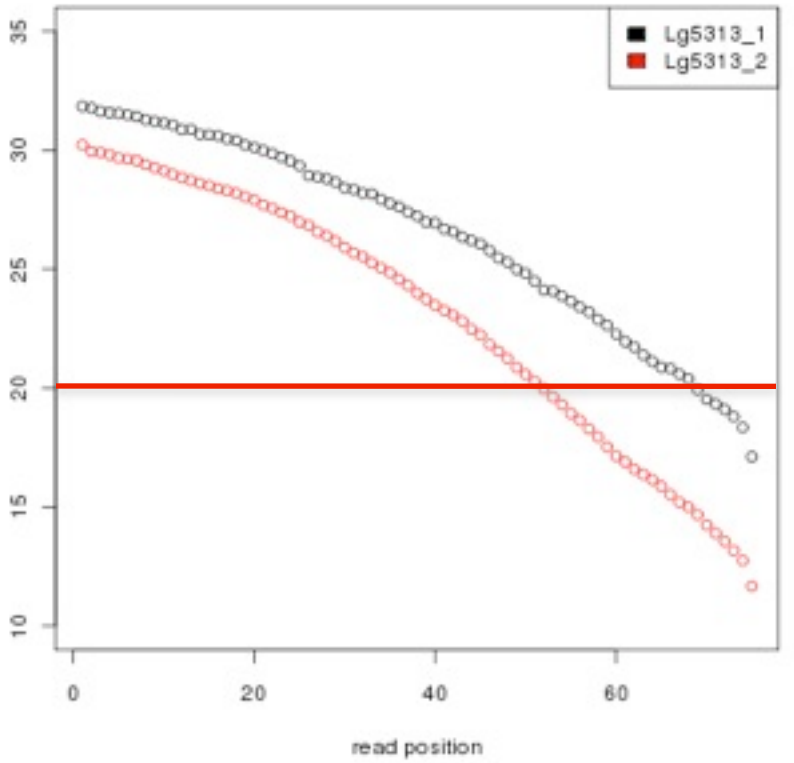
* The chastity of a base call is the ratio of the intensity of the greatest signal divided by the sum of the two greatest signals. Reads do not pass the quality filter if there are two or more base calls with chastity of less than 0.6 in the first 25 cycles. These reads have an "N" in the last column of the GA analysis software export file.

Variability in the quality (mean value per position)

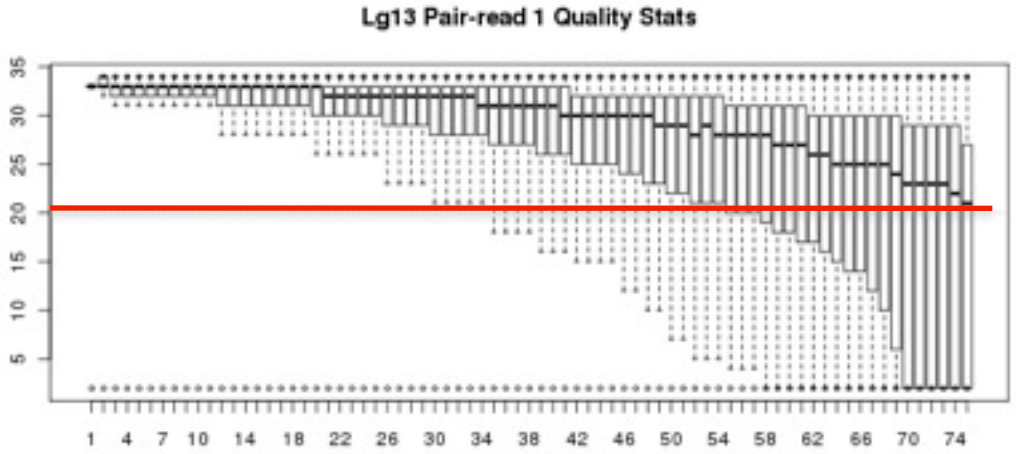
- Good example

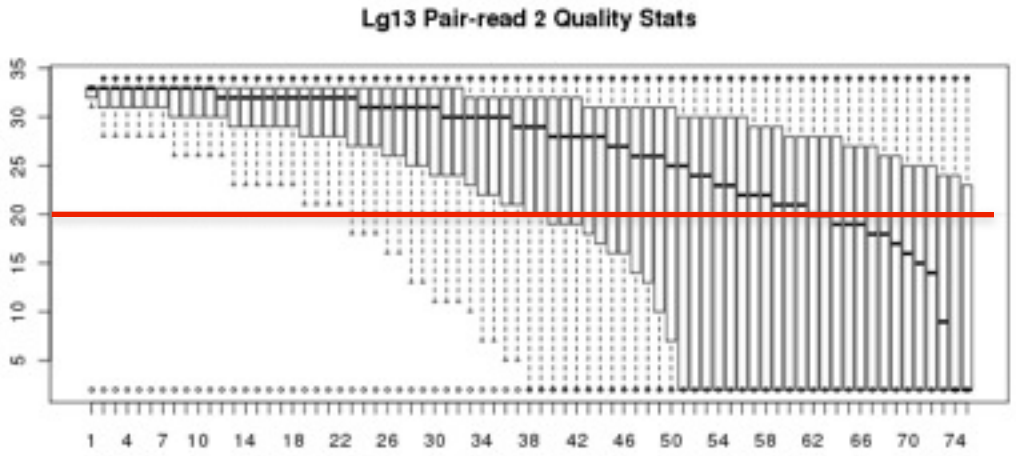


- Less good example…



Variability in the quality (boxplot)





Filtering data can help

http://pathogenomics.bham.ac.uk/blog/2009/09/tips-for-de-novo-bacterial-genome-assembly/

- Illumina reads quality decrease with length
  - Trim 3' ends of reads according to quality
  - Remove reads with average low quality
  - If coverage is high, remove orphan reads
- 454 reads
  - Trim 3' ends of reads according to quality
  - Remove reads with average low quality
  - If possible correct for long mononucleotide repeats
- Check contigs by remapping reads

Limitations of the sequence

- Repeats
  - transposases, IS-elements, retroviruses, duplications, etc.
- Polymorphisms
  - SNPs, CNV, multiploid, sample mixture, etc.
- Sequence bias
  - %GC

Repeats are a major issue for all assemblers

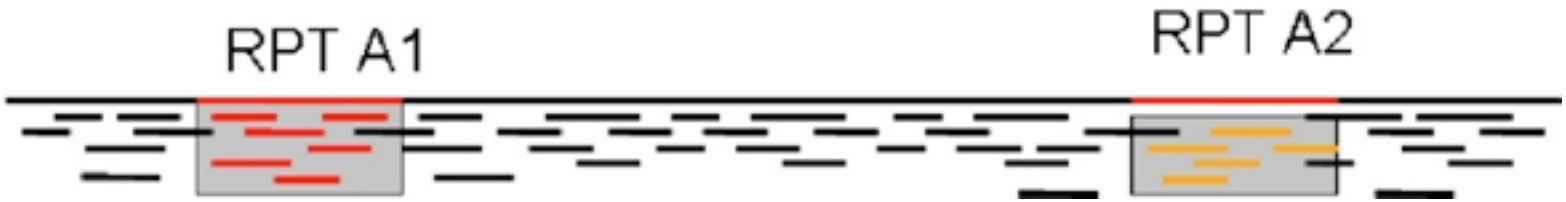


Figure top. Two copies of a repeat along a genome. The reads colored in red and those colored in yellow appear identical to the assembly program.

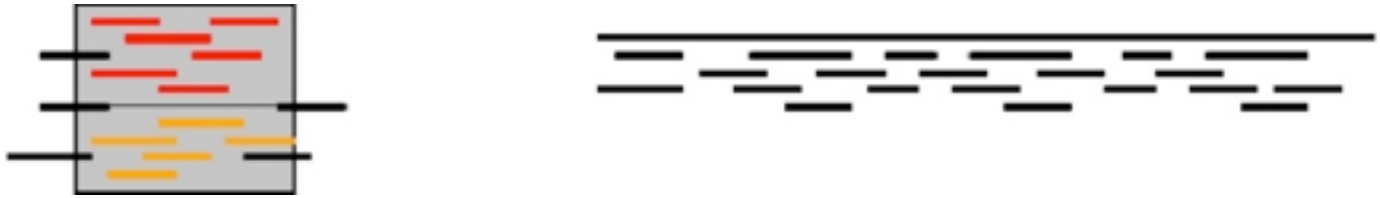


Figure bottom. Genome mis-assembled due to a repeat. The assembly program incorrectly combined the reads from the two copies of the repeat leading to the creation of two separate contigs.

Helping the assembly with linked reads

- When the **distance** and the **orientation** between 2 reads is known
- First proposed by
  - Edwards, A; Caskey, T (1991). "Closure strategies for random DNA sequencing". Methods: A Companion to Methods in Enzymology 3 (1): 41–47. doi:10.1016/S1046-2023(05)80162-8.
- Also called
  - Double-barreled
  - Mate-pairs
  - Paired-ends

Mate pairs validation example

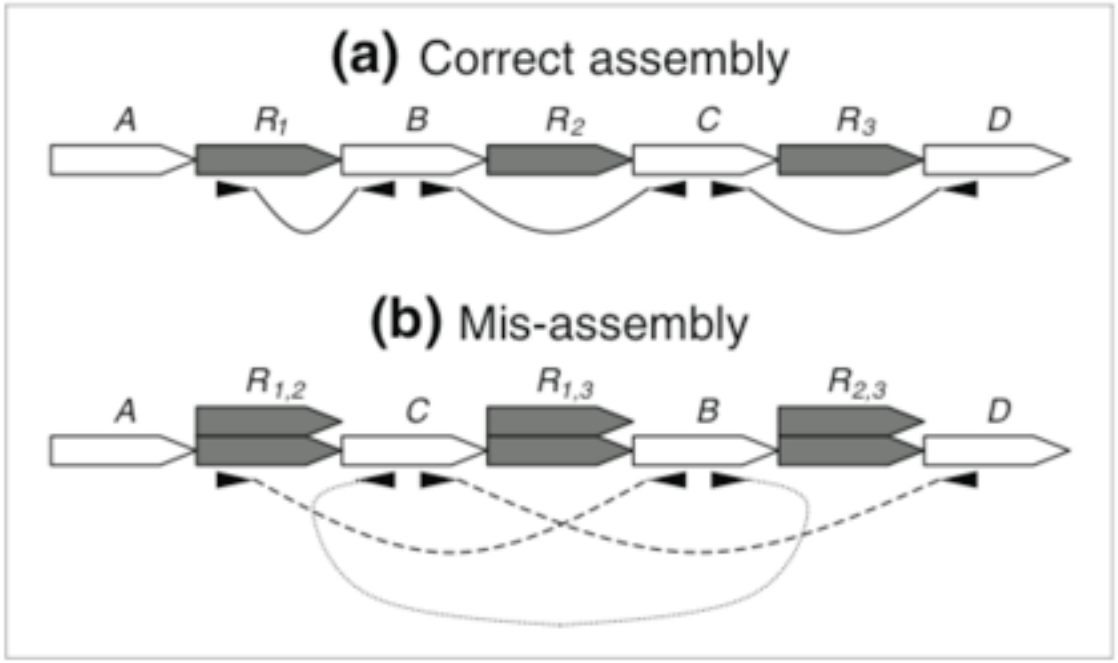
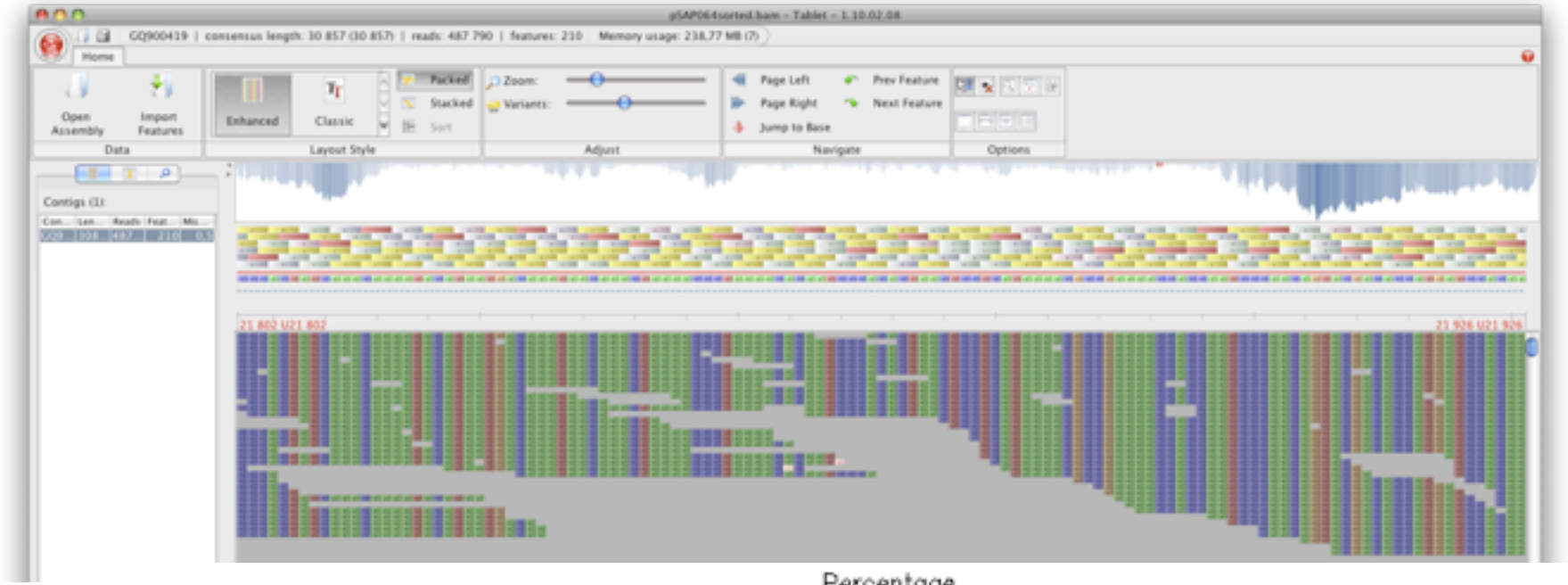
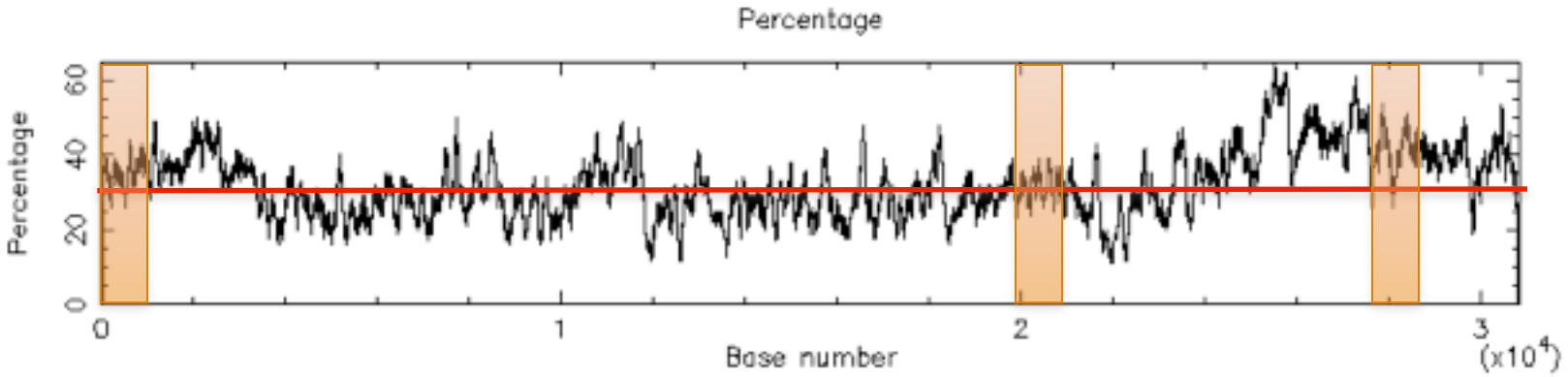


Figure 3
Mate-pair signatures for rearrangement style mis-assemblies. **(a)** Three copy repeat *R*, with interspersed unique sequences *B* and *C*, shown with properly sized and oriented mates. **(b)** Mis-assembled repeat shown with mis-oriented and expanded mate-pairs. The mis-assembly is caused by co-assembled reads from different repeat copies, illustrated by the stacked repeat blocks.

- 4 main criterias
  - Mates too close to each other
  - Mates too far from each other
  - Mates with same orientation
  - Mates pointing away from each other
- Other criterias
  - Mates not present on the assembly (singletons)
  - Mates on different contigs

GC bias = coverage bias ??





Repetitive elements

A real case: assembly and comparative genomics of 8 *Staphylococcus aureus* strains

- Goal: compare and identify genomic differences between the strains in order to explain their various phenotypes
- Means: Illumina sequencing followed by bioinformatics assembly and comparison

Acknowledgements:
A collaboration with Dominique Blanc, Valérie Vogel and Patrick Basset, Service de Médecine Préventive Hospitalière, CHUV

Thanks to Sandra Calderon for her annotation pipeline, SIB

Preliminary results with a *S. aureus* strain

- Each genome is sequenced on a single Illumina lane
  - ~8 mio reads of 36bp paired-end with insert size about 300bp
  - $8x10^6x36x2 = 576$ Mbp for a genome of 2.8 Mbp
  - Theoretical coverage ~200x
- After strictly cleaning low quality reads
  - Only 25%-40% of reads were of acceptable quality.

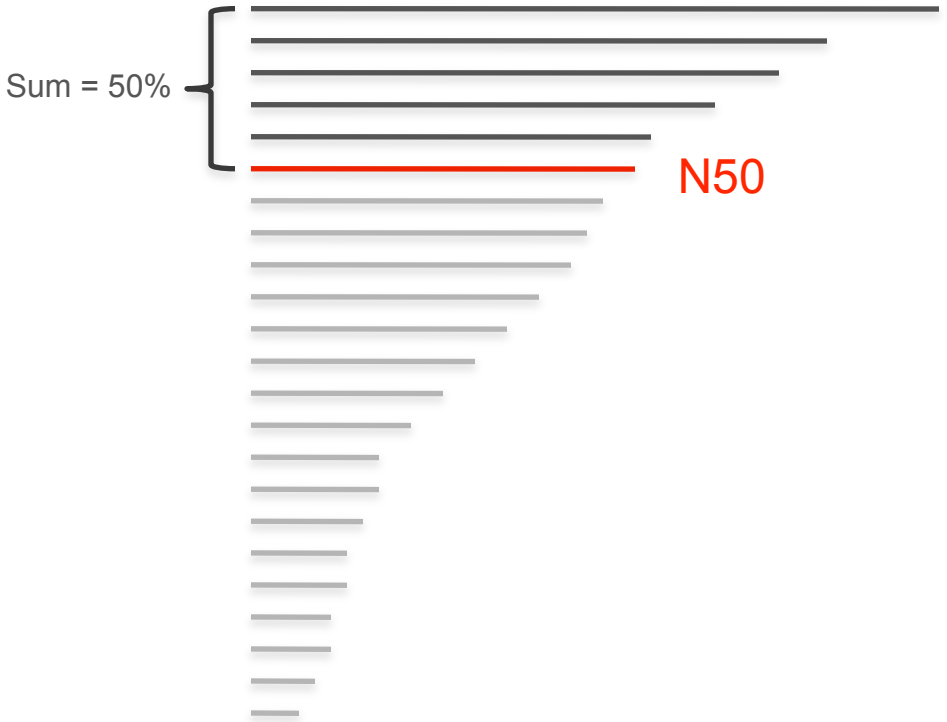
- The strains are very similar to reference genome N315

Assembly quality measurements

- Number of contigs
  - Ideally 1 for a bacterial genome…, but the lower the better
- Contig sizes
  - The larger the better (up to the size of the genome), usually given in maximum, minimum and average lengths.
- Correctness
  - Difficult to assess for a new genome
- N50
  - **The most used quality value for *de novo* assembly**
  - The N50 is the size of the smallest of all the large contigs covering 50% of the genome

N50 what's that?

- Sort the contigs by size
- Sum them starting with the largest until you reach 50% of the estimated genome size
- Last contig added = N50



Velvet for S5

K =	21	23	25	27	29	31
Nr contigs	7012	1906	767	420	325	**252**
Consensus size bp	3103135	2918437	2875773	**2863169**	3619536	2936521
N50	12182	43427	67361	66898	66306	**107440**
Min	41	45	49	53	57	**61**
Max	161171	201664	201396	201389	238872	**369778**

ABySS for S5

K =	21	23	25	27	29	31
Nr contigs	4318	2891	2123	1636	1339	**1113**
Consensus size bp	3220552	3127361	3088161	**3049081**	3078819	3052504
N50	15928	25693	29334	30241	**31596**	29797
Min	21	23	25	27	29	**31**
Max	63797	132812	132816	132992	**132996**	122383

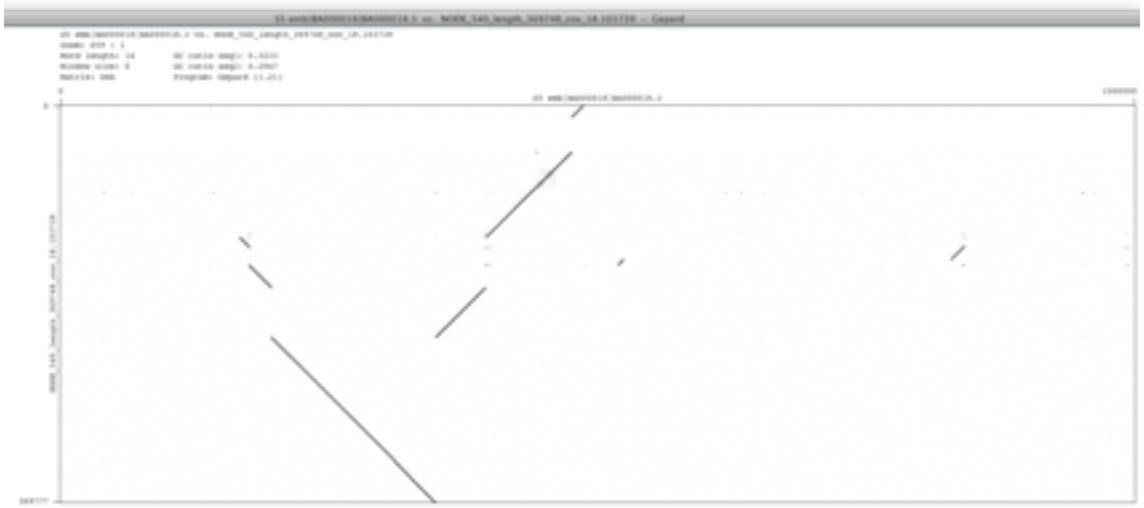
SOAPdenovo for S5

K =	21	23	25	27	29	31
Nr contigs	247	**213**	234	248	280	362
Consensus size bp	2818333	**2825424**	2831502	2833334	2828297	2785031
N50	98956	**99286**	82319	82910	84517	52098
Min	100	100	100	100	100	100
Max	**253458**	252985	181975	182194	182217	141106

Best scores

K =	Velvet31	ABySS29	SOAPdenovo23
Nr contigs	252	1339	**213**
Consensus size bp	2936521	3078819	**2825424**
N50	**107440**	31596	99286
Min	61	29	**100**
Max	**369778**	132996	252985

However longest contig is not always the best…



- Velvet: 369’778bp

- ABySS: 88’077bp and 132’996bp

Beware of unwanted options…

- E.g., Velvet scaffolding (ON by default in the last version 0.7.55)…

```
>NODE_1931_length_5525_cov_11.776470
TTTTTAAGTGCATGTGTATAATTTTCTACTGGGATAGGATCTGATGTTGCTGAACCTTCA
AATATAGTTATTTCTGGCAATCTTTCCTCTGCATAGTTAAAAGCTTTATTTAAAATTTCA
TCTATGTCTACATATATTTTTGTATACAGTCTCTTACCTAATTGAGTATTTAAATAACAA
TATTCACACATCCCCATGCACCCACTAACTAAAGGTAACTGATAATGTGCGGATGGTTTN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
…
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNTATACATCTGTAGATATCCATTCTCTAGAAACTTCT
ACTCCATTTTCATATGTTATTTGGTAAGAACTTGCCTTATAGCCTGGCATACCATTAATT
TTTACTTTTTGCTCTCCTTCTTTTAAAGCTGGATTATCAATATATTCAACTTCTGGATTA
TATTTTTCGTGTACTTCATTTACTAGTTGATATGTCTTTCCACTTAAAGCCTCTTTATTT
CCGTAAATATTAAATCCTACAGTTCCTCCACCAATATATCCTTCTATATATACTGGAAAA
TCATATGAATTAACAAATTTATAATCTACTACTCCATAAGCTACTGTAGCATCTAAACCT
GGATCTGAATATGAAACTGTTAATGTGTGGTTAGCTCTTTCTGTTGATTTTAAATTAGCA
```

Assembly by mapping onto a reference

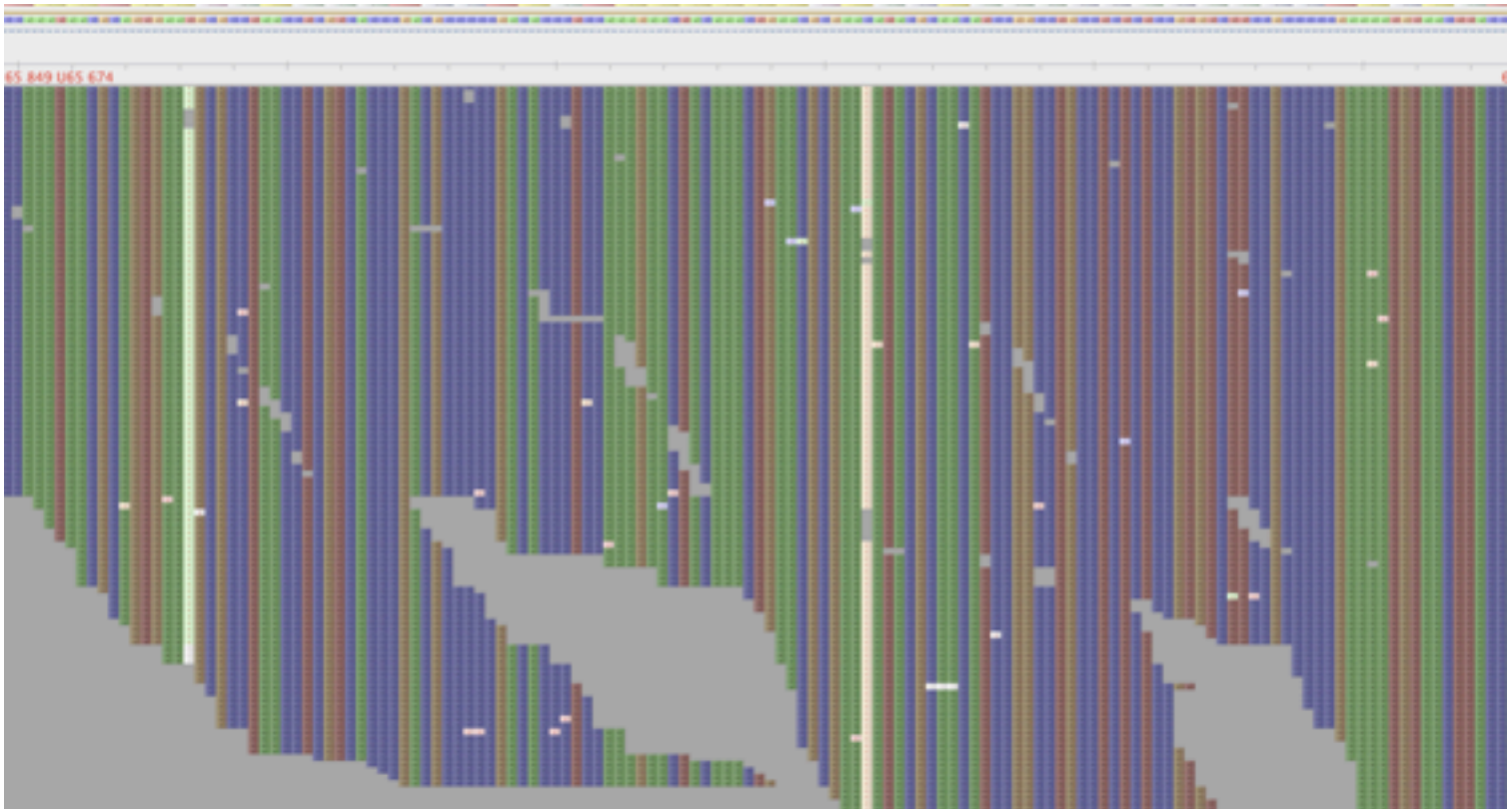
- Easier with MAQ, Bowtie, BWA, Mosaik…
- SNPs are well identified if coverage is sufficient (20x or more)
- Difficult to identify insertions, deletions and inversions
- Warning some software give unexpected results (e.g., MAQ)
- Limitation in the number of mismatches allowed

- Example with *S. aureus*
  - Using the closest reference genome *S.aureus* N315
  - MAQ gives a nice consensus with the gaps and the SNPs
  - Bowtie gives also the non-matching reads

MAQ Pileup example of S6

```
…
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   @TTTTttTTTTTTTTTtTTTTTtTttttTtTTTTtttttttttTtttttTTTTTtTTTttttttTttTTTTtttTTtTTTTttttTttTTtTTttttTttTTtTT
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.,,,,,,,,,,,.,,...,..,,,.,,,,,,,,,,,.,,,.........,,,
```

Visualization of the mapping and the SNPs



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

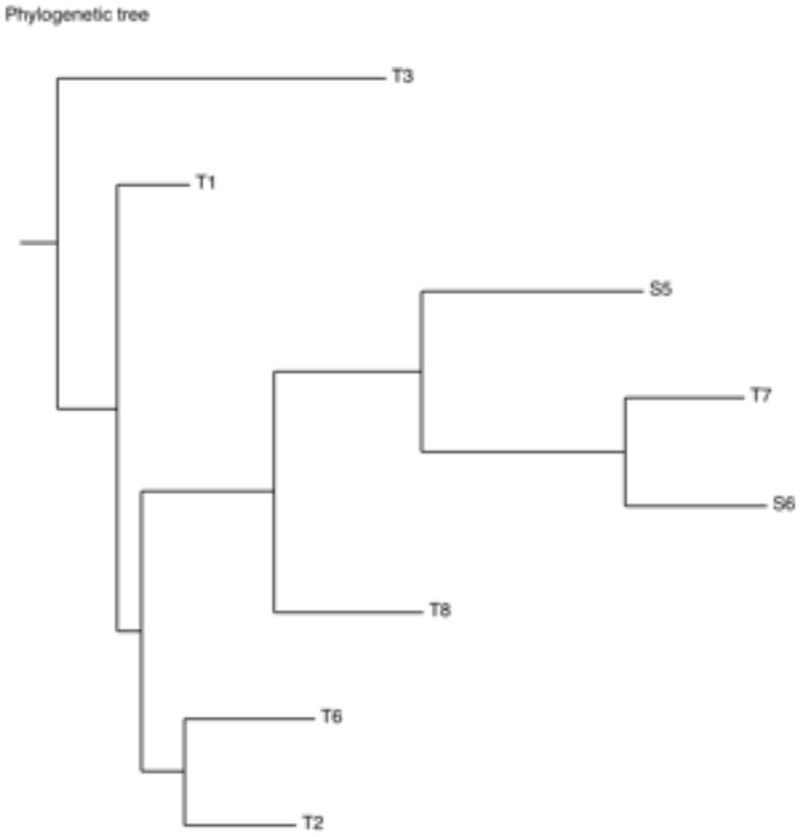
Comparison obtained for matching reads

- Most of the two genomes match perfectly with the reference
- Few SNPs (single nucleotide polymorphisms) are found

Genome	SNPs vs 315	SNPs vs S5	SNPs vs S6	Total SNPs
S5	775	0	74	849
S6	775	88	0	863

- Unfortunately those SNPs are not found in significant genes
- They are often due to sequencing ambiguities

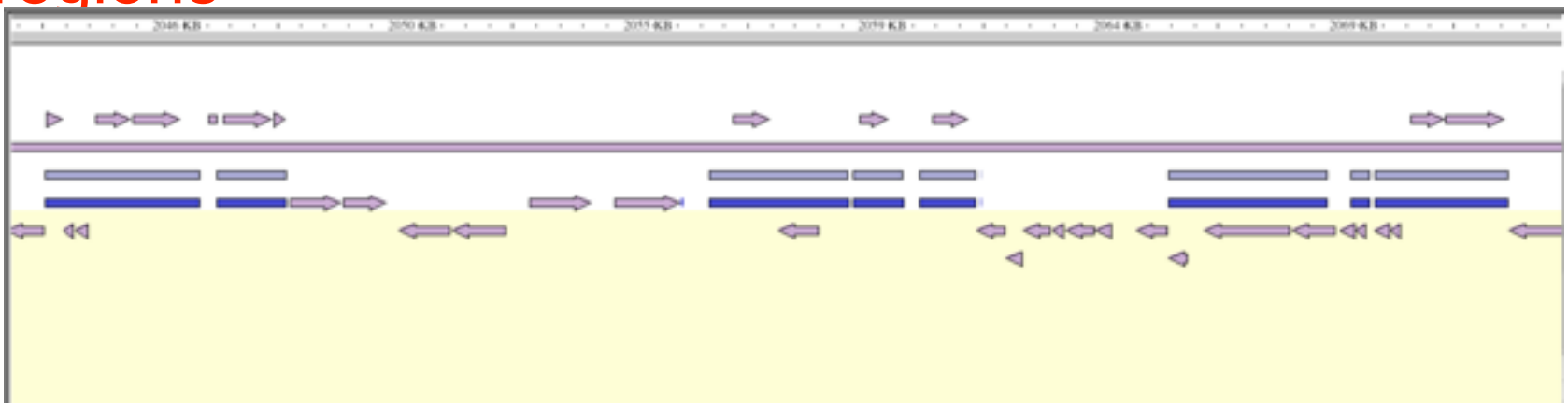
Tree and SNPs



Genome	SNPs vs ref N315	Unique SNPs
T3	904	48
T1	853	32
S5	849	44
T7	898	22
S6	863	25
T8	703	24
T6	879	57
T2	895	34

Of these 557 SNPs are identical in all genomes.

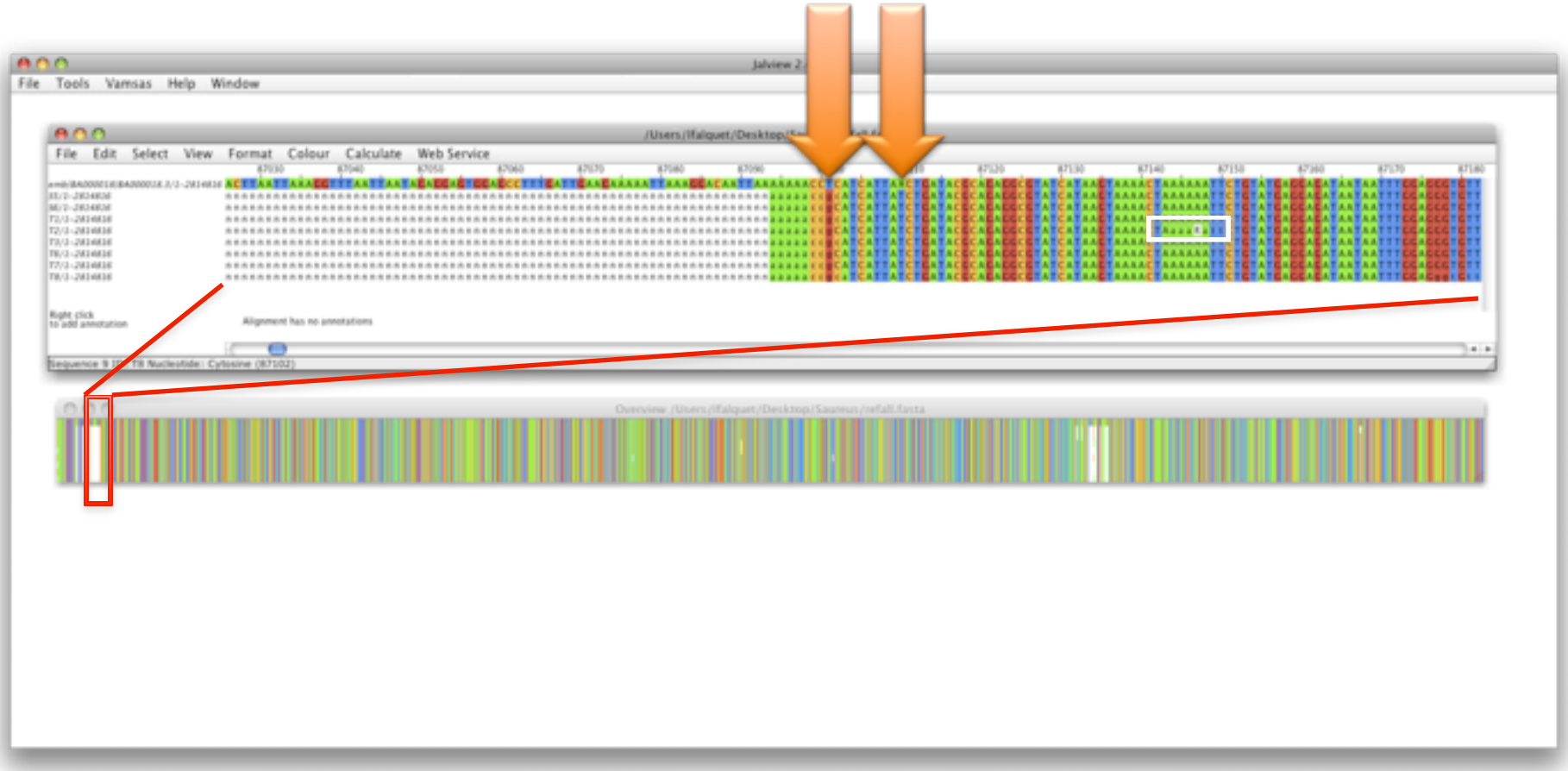
Several gaps *vs* N315 are identified in a few regions



- Overall **S5** gap size 65’436 bp or 0.2%
- Overall **S6** gap size 65’332 bp or 0.2%

=> Gaps are small (a few Kbp) and nearly identical in both genomes!

MSA of all genomes *vs* N315

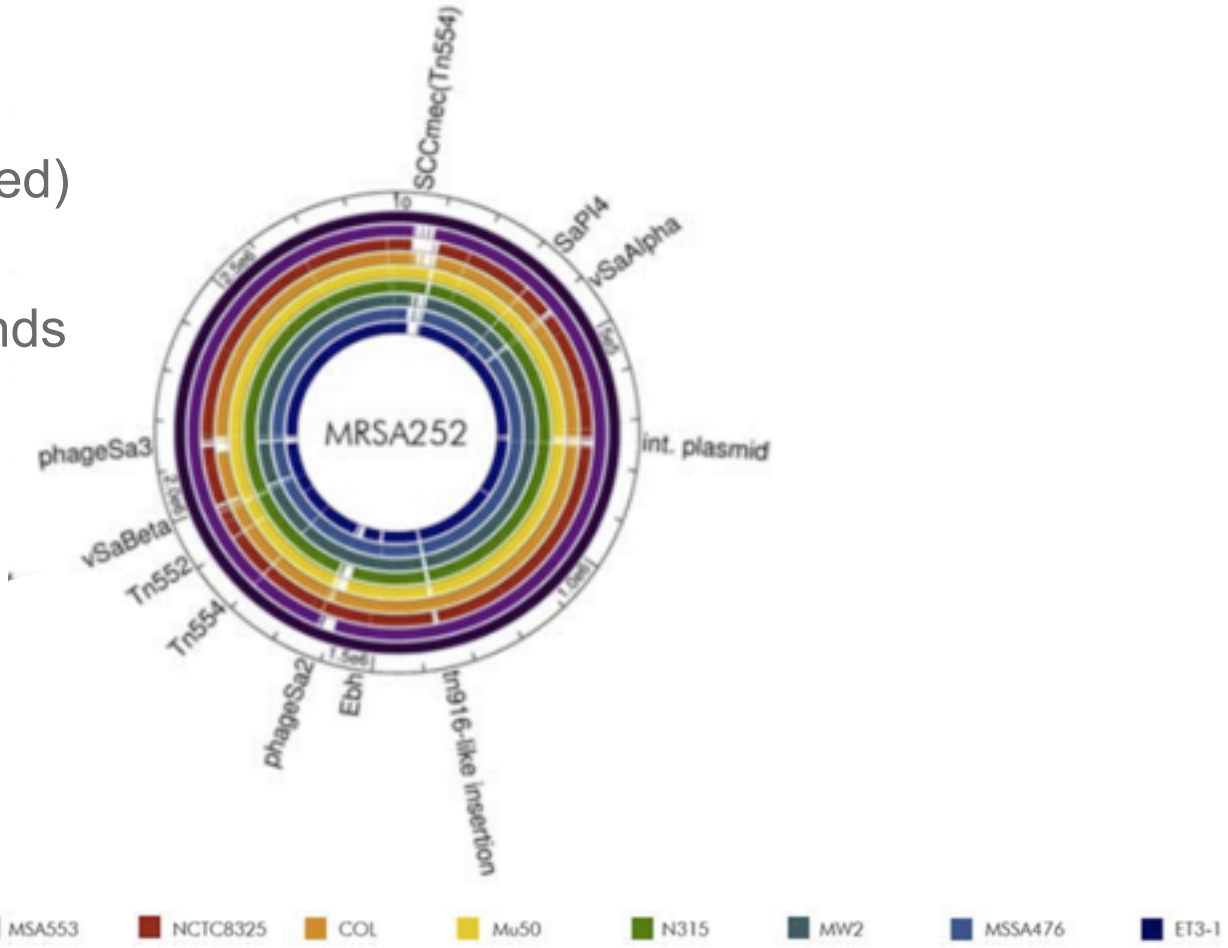


Where is the difference?

- 99.8 % of the two genomes are identical to the reference
- 0.2 % of the two genomes are gaps, but at exact same positions
- Only 162 SNPs are found, but not in significant genes

Known differences between *S.aureus* strains

- Phages
- Plasmids (integrated)
- Genomic islands
- Pathogenicity islands
- SCCmec
- Transposases
- Other



Comparison of non-matching reads

Genome	Matched reads	Unmatched reads (many of low quality)	Total
S5	5'094'413	2'927'239	8'021'652
S6	5'490'459	3'212'370	8'702'829

- We used Bowtie to obtain the list of non-matching reads
- => We assembled the unmatched reads of good quality with ABySS

Comparison of non-matching contigs

Genome	Contig1	Contig2	Contig3	Contig4	Contig5	Contig N
S5	**28'398**	17'936	16'007	15'949	8'298	…
S6	**20'698**	17'938	15'951	14'801	11'482	…

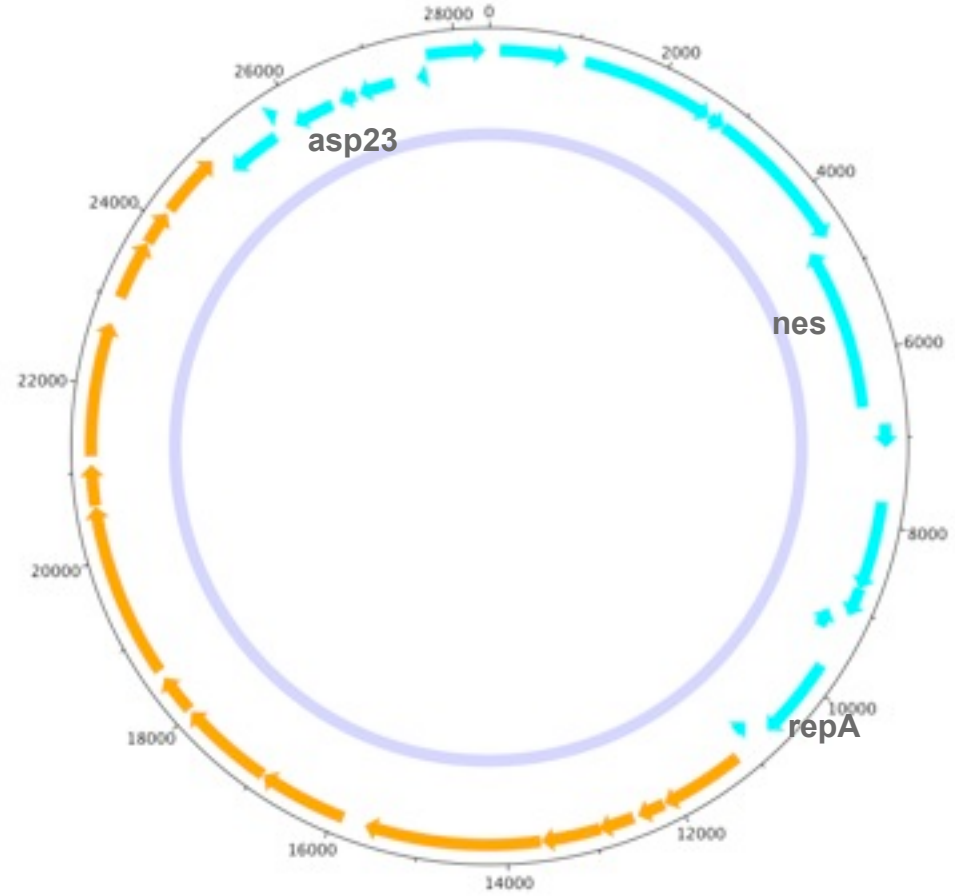
Differences between S5 or S6, and the reference

Main difference between S5 and S6

- Analysis of the contigs
  - **1st contigs** seems to represent two different plasmids!!
  - 2nd contigs SCCmec region likely to originate from *S.aureus* COL
  - Other smaller contigs: resemble integrated phages

Plasmid pS5 close to pUSA03

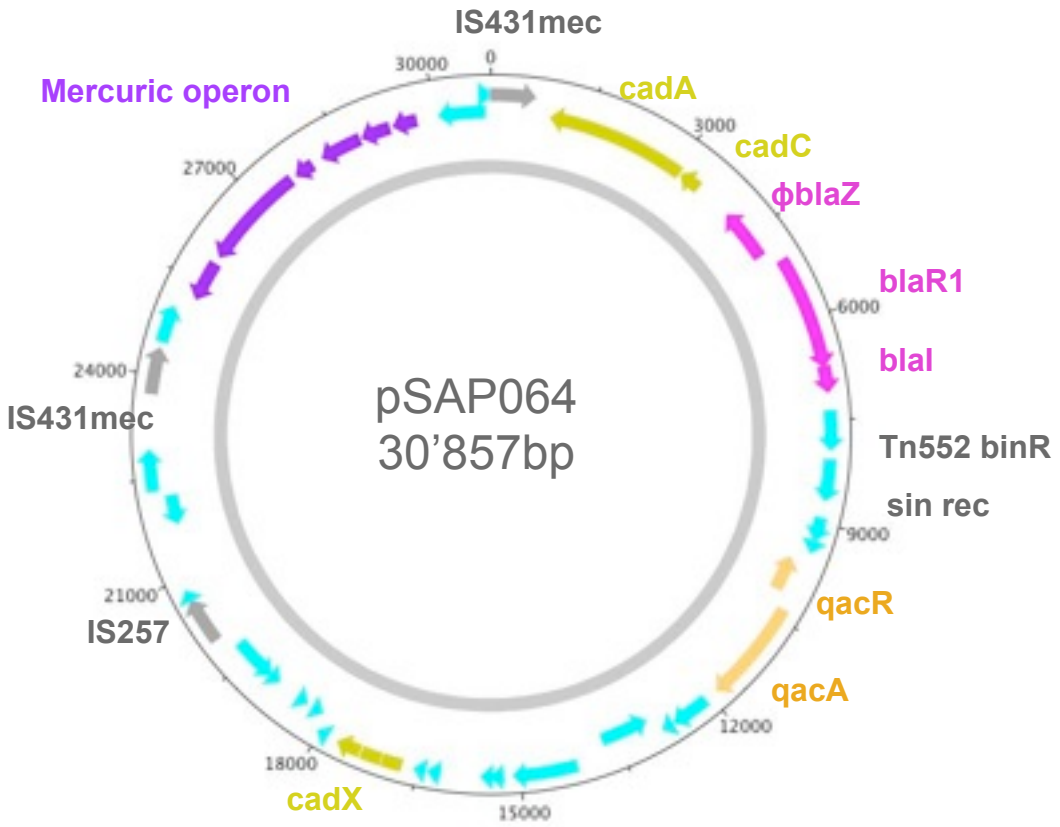
- But lacks the mupirocin resistance gene
- This plasmid contains an interesting Tra operon also called the F sex factor



Tra operon: Genes coding the F-Pilus and DNA transfer pores (F sex factor)

Plasmid pS6 99.99% identical to pSAP064

- qac = antiseptic resistance
- bla = ß-lactamase (phenicillin resistance)
- **cad** = Cadmium resistance
- mer = mercuric operon



pT1, pT2, pT3, pT6, pT7, pT8 and pS6 share the same plasmid

A real case: assembly of *Pseudomonas knackmussii*

Acknowledgements:
A collaboration with Prof. J. van der Meer, UNIL

Thanks to Sandra Calderon, SIB

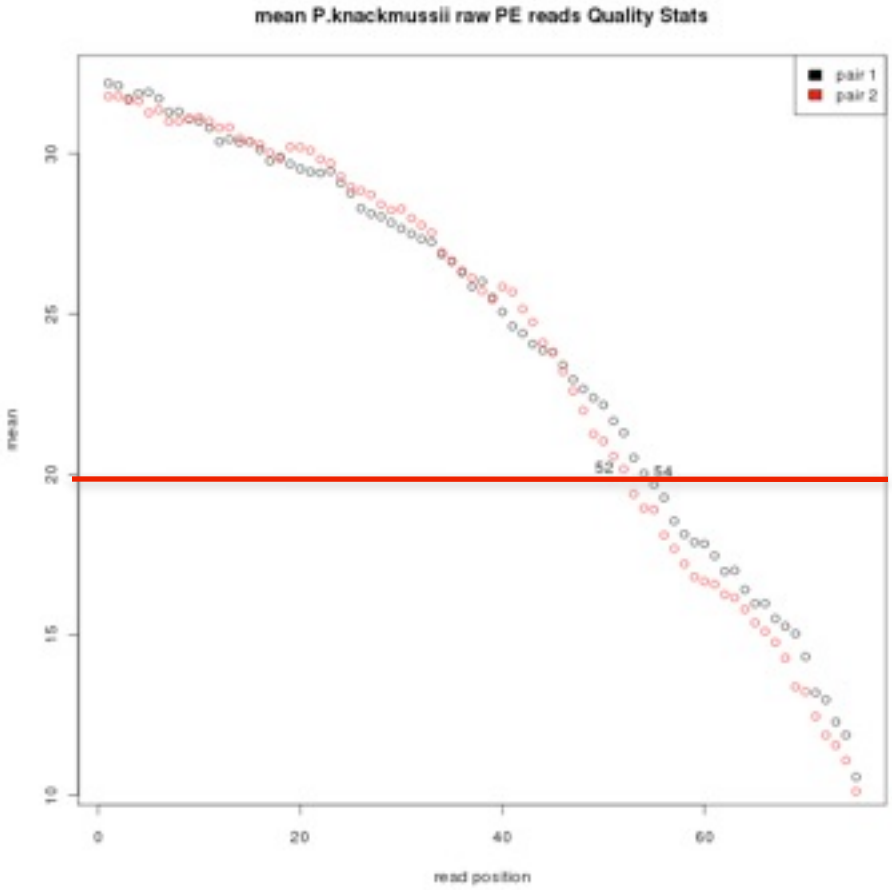
Two interesting examples in the first run

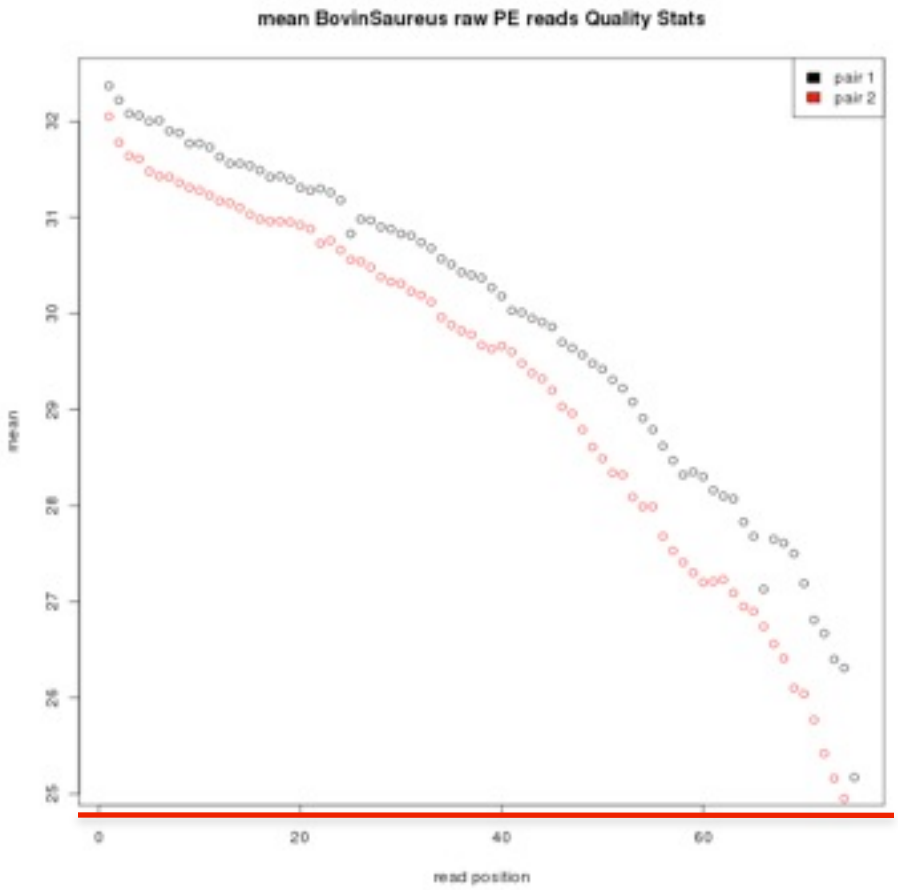
- 2 different genomes sequenced on the same Illumina plate
- one lane of 75bp paired-end each

- Pseudomonas knackmussii B13
  - 6.2 Mb
  - 39’442’350 reads
  - Coverage ~950 X

- Staphylococcus aureus
  - 2.8 Mb
  - 26’000’722 reads
  - Coverage ~700 X

Quality control examples





Pseudomonas knackmussii B13

soft	n	n:100	n:N50	min	median	mean	N50	max	sum	K
ABySS raw	8'186	2'928	440	100	798	1'396	2'529	23'042	4'088'965	45
ABySS filtered	6'751	3'855	525	100	316	585	986	25'334	2'256'952	25
Velvet trimmed	5'889	5'889	729	100	294	653	1'179	34'426	3'845'785	47
SOAPdenovo raw	2'682	2'682	314	100	835	1'717	**3'670**	**39'122**	4'606'315	29
SOAPdenovo trimmed	6'608	6'608	828	100	292	701	1'444	15'745	4'633'289	31

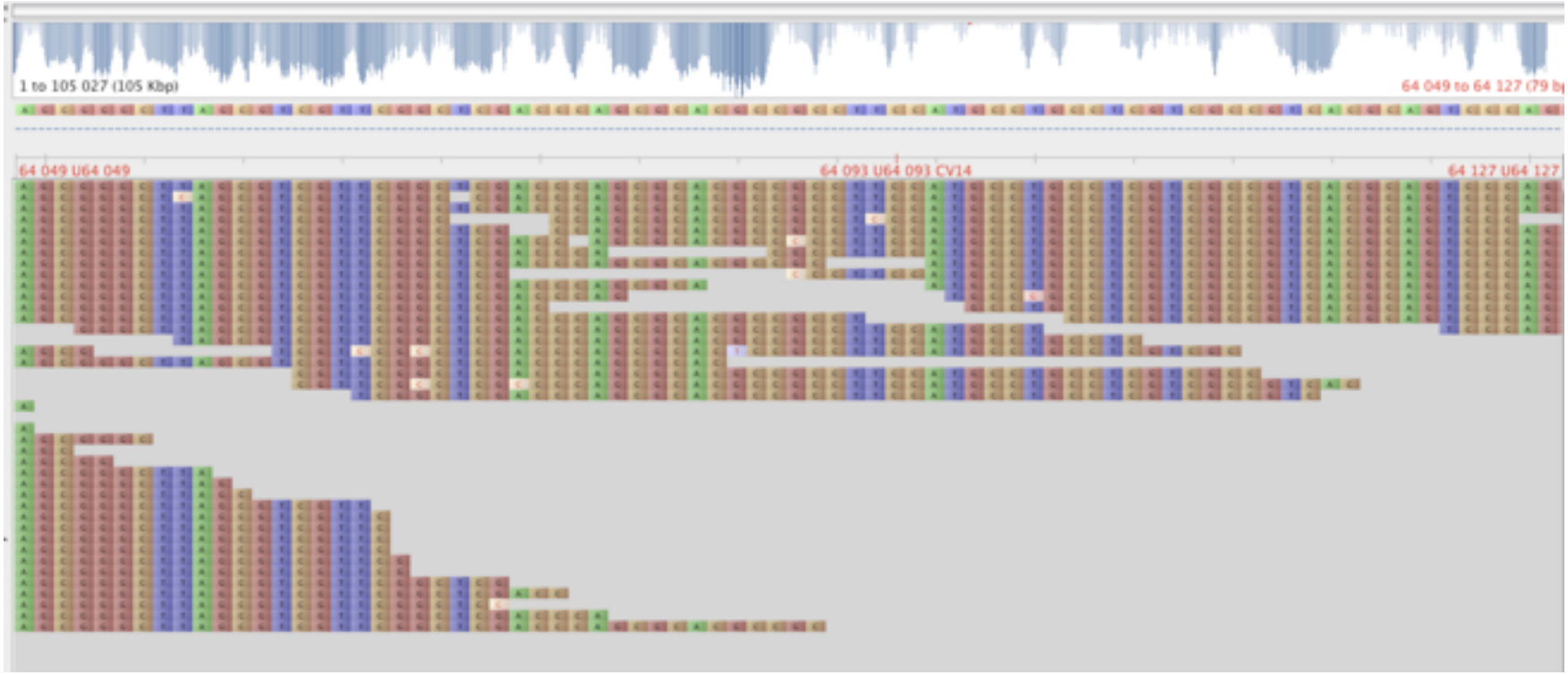
The size of the contigs and the N50 are poor.
The size of the genome is strangely lower than expected.

Staphylococcus aureus

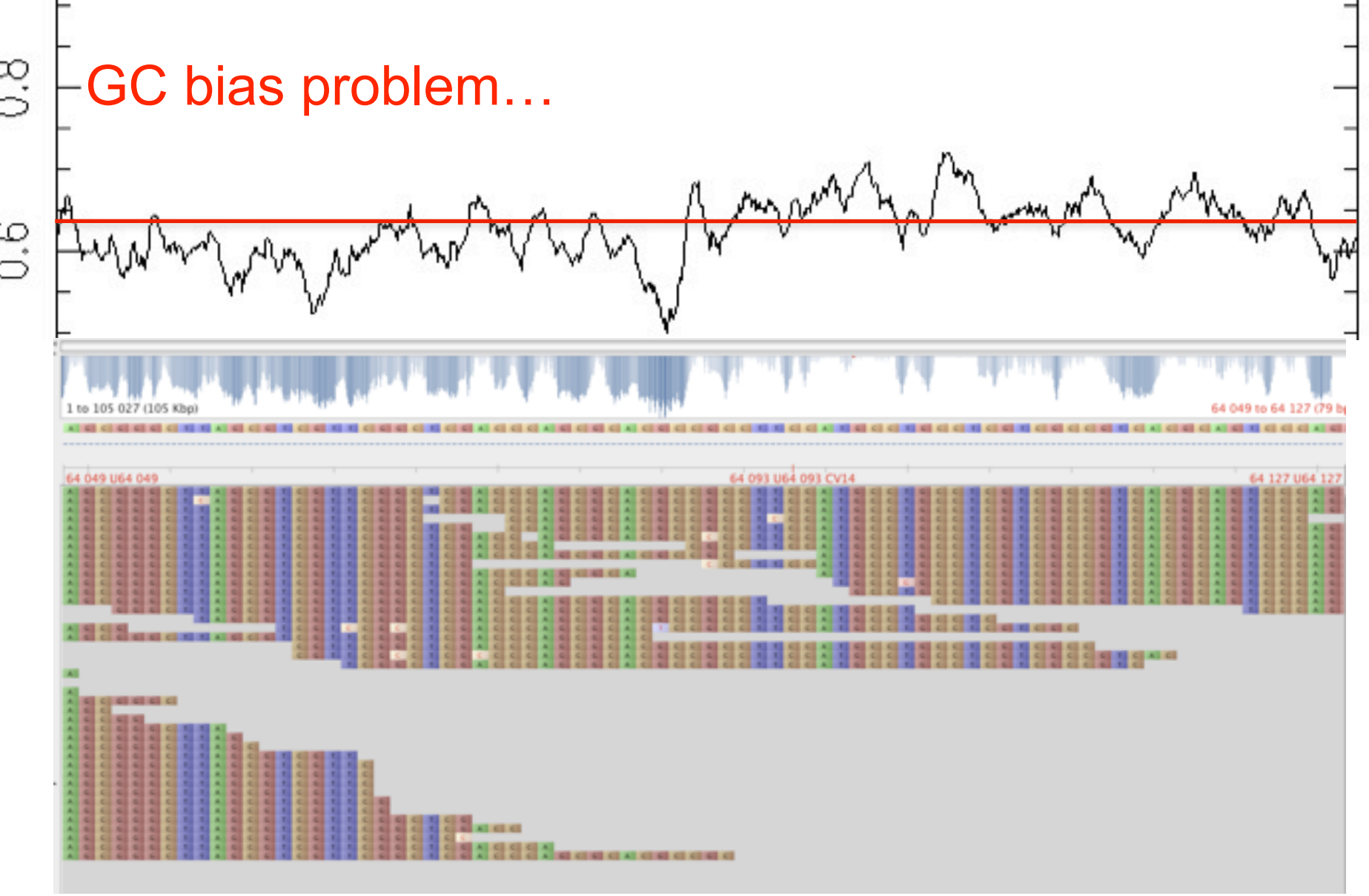
soft	n	n:100	n:N50	min	median	mean	N50	max	sum	K
ABySS raw	410	237	6	101	127	12’718	**145’727**	**399’819**	3’014’238	64
Velvet raw	452	452	27	117	190	6’016	31’599	137’324	2’719’549	59
SOAPdenovo raw	1’985	1’985	218	100	512	1’305	3’021	21’251	2’590’721	31

The size of the contigs and the N50 are correct.
The expected size of the genome is correct.

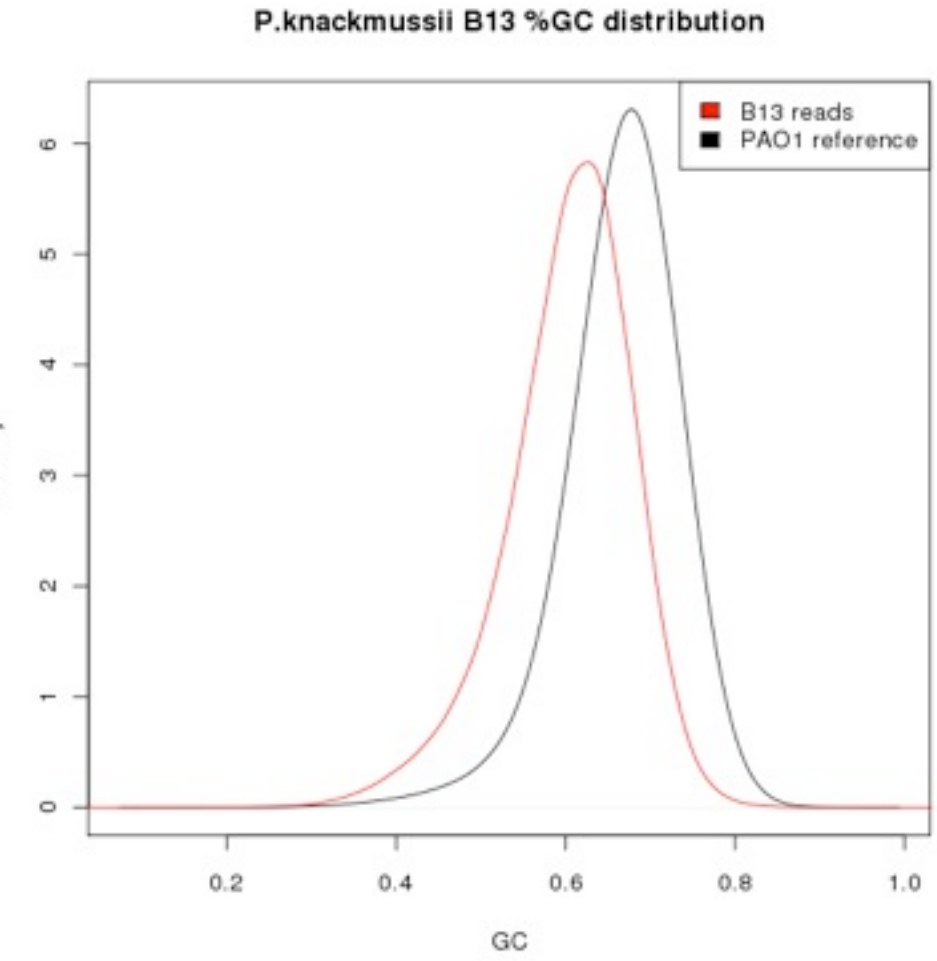
Coverage problem ? (mapping on genomic island ICEclc 100kbp)

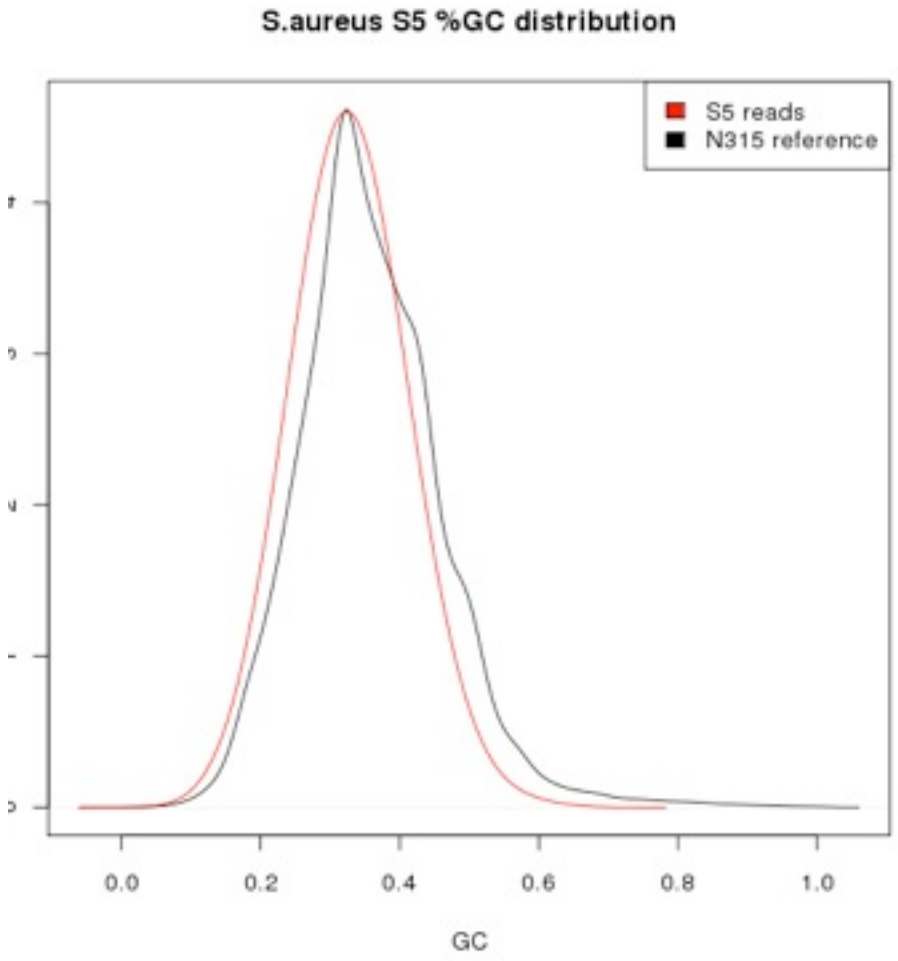


GC bias problem…



Comparison of experimental vs theoretical GC content



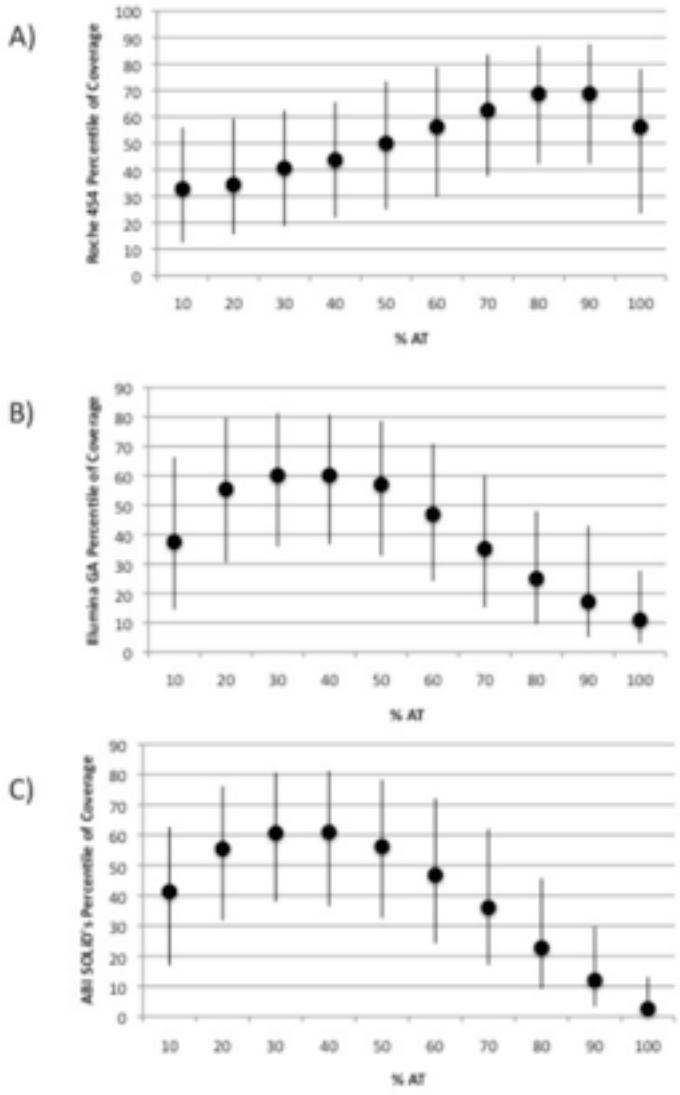


GC bias can create uneven coverage issues

- Similar results for other *P.knackmussii* strains… ☹
- With Illumina we observe low coverage of the reads in regions:
  - above 55% GC
  - below 30% GC
- Potential solutions in the literature:
  - Low GC: use a protocol designed by Sanger Center, extract DNA from gel at low temperature
  - High GC: use a protocol designed by the Broad Institute, run PCR with Betaine or DMSO, and use a different polymerase

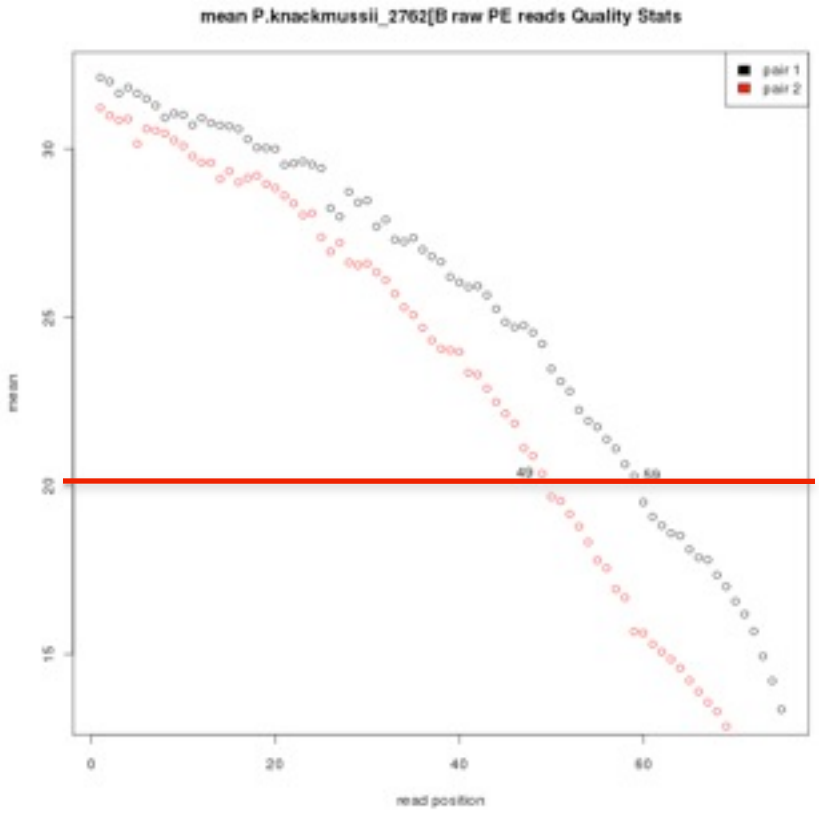
GC bias in 3 platforms contradictory results…

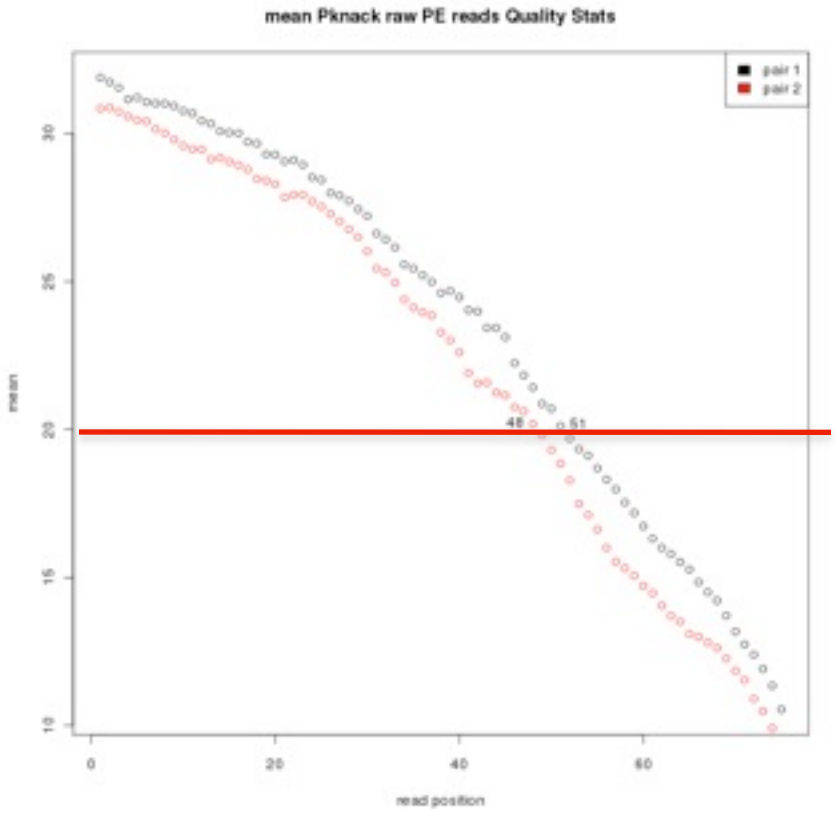
- A) 454
- B) Illumina
- C) SOLiD



Harismendy *et al. Genome Biology 2009, 10:R32*

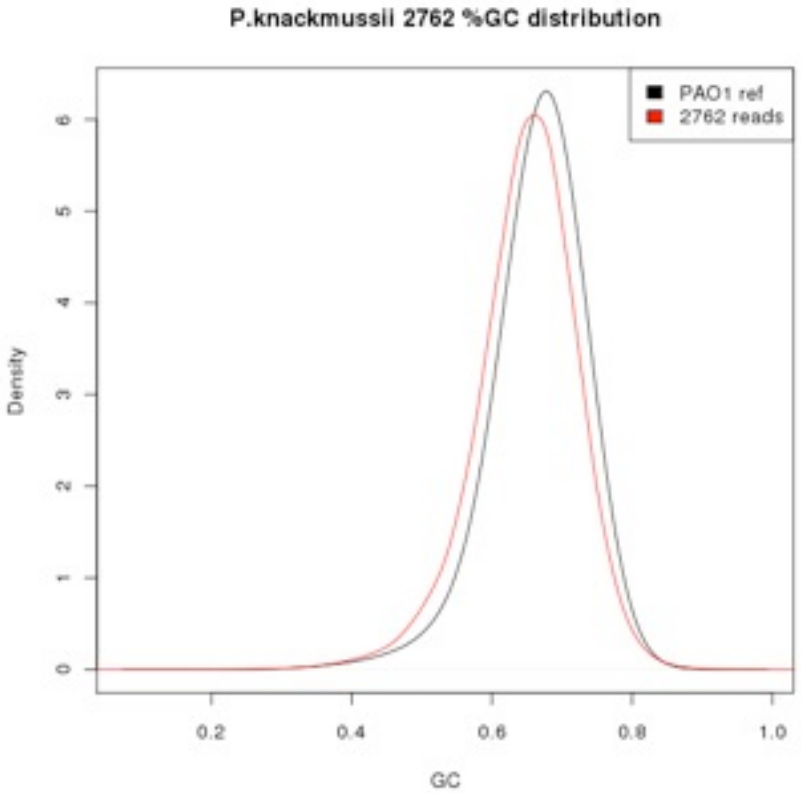
New sequencing with AccuPrime and E-select gel for P.knackmussii 2762 (left) vs B13 (right)

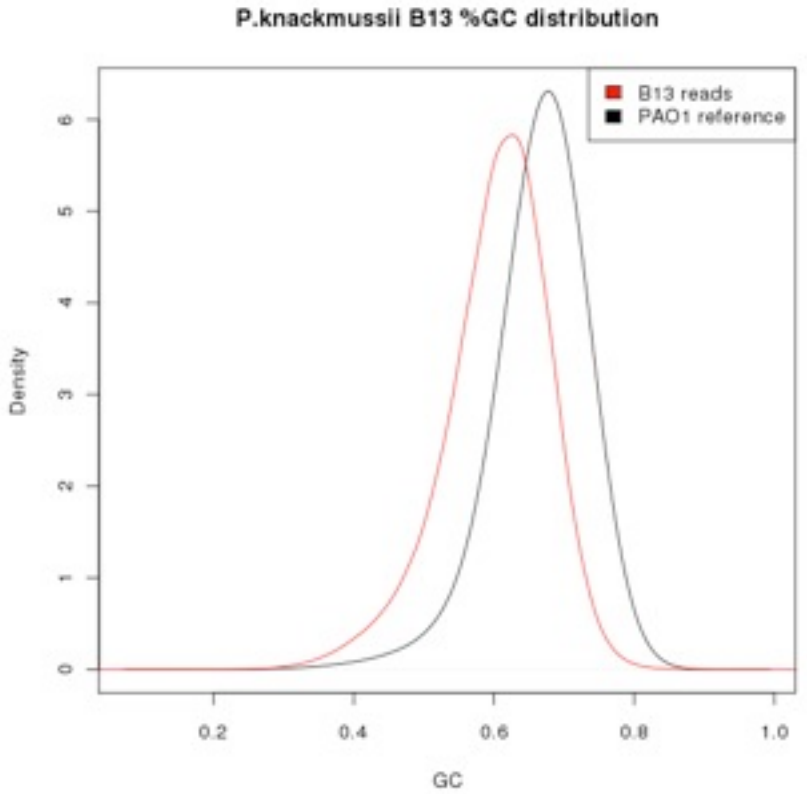




Only slight improvement apparently…

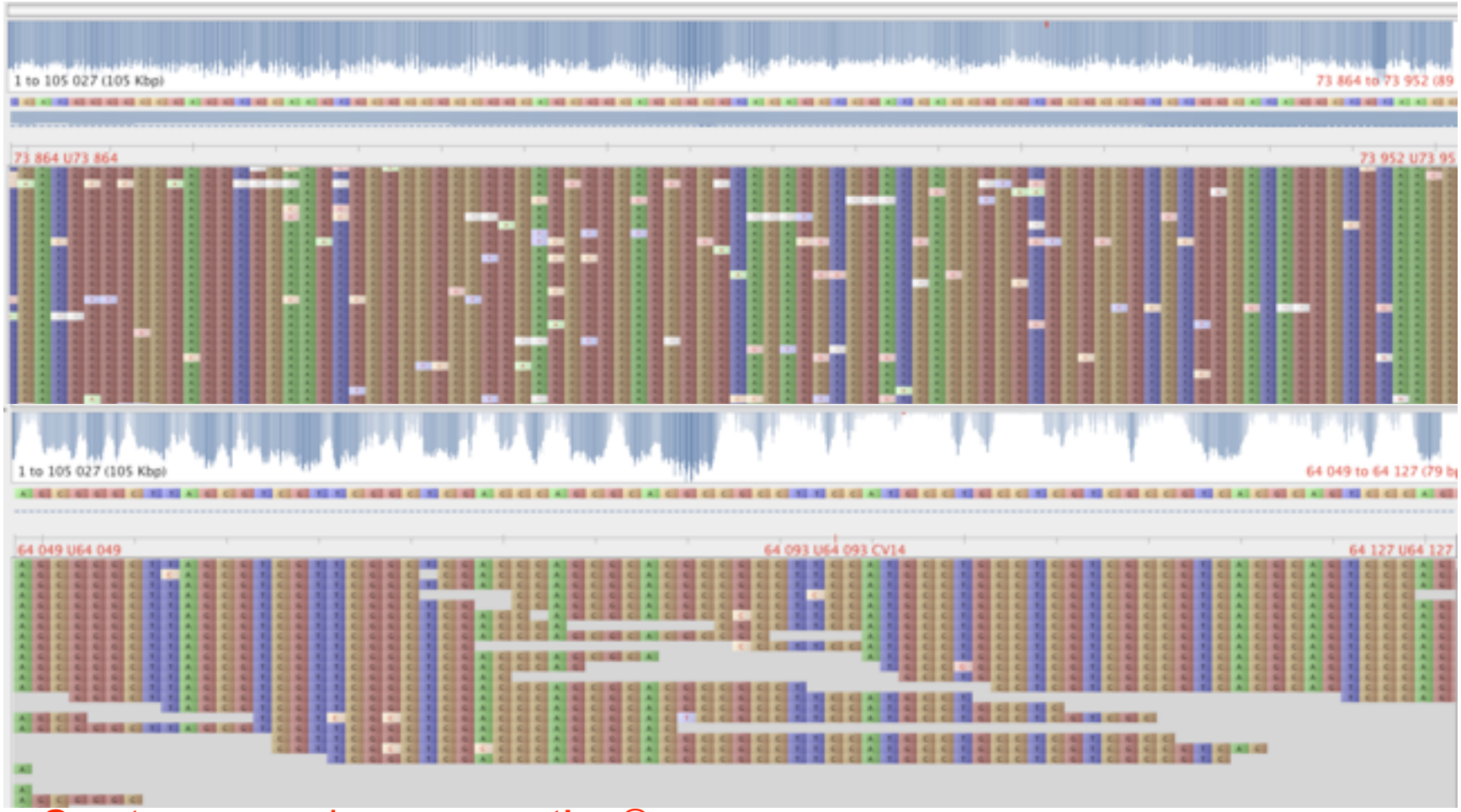
New sequencing GC content for P.knackmussii 2762 (left) vs B13 (right)





Much less GC bias!

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



Great coverage improvement!... ☺

Pseudomonas knackmussii improved assembly 2762 (top) vs B13 (bottom)

soft	n	n:100	n:N50	min	median	mean	N50	max	sum	K
ABySS raw	250	133	18	102	24’087	46’811	108’423	233’781	**6’225’940**	49
ABySS trimmed	248	122	15	101	23’817	49’660	119’517	**426’707**	6’058’631	39
Velvet trimmed	773	773	81	100	2’151	7’507	21’874	110’157	5’803’151	47
SOAPdenovo raw	199	199	14	100	182	28’502	**133’061**	417’556	5’672’003	31
SOAPdenovo trimmed	368	368	30	100	816	15’621	62’344	163’765	5’748’813	31

soft	n	n:100	n:N50	min	median	mean	N50	max	sum	K
ABySS raw	8'186	2'928	440	100	798	1'396	2'529	23'042	4'088'965	45
ABySS filtered	6'751	3'855	525	100	316	585	986	25'334	2'256'952	25
Velvet trimmed	5'889	5'889	729	100	294	653	1'179	34'426	3'845'785	47
SOAPdenovo raw	2'682	2'682	314	100	835	1'717	**3'670**	**39'122**	4'606'315	29
SOAPdenovo trimmed	6'608	6'608	828	100	292	701	1'444	15'745	4'633'289	31

Conclusions

- Be careful with biased GC content genomes (low or high!)
- SOLiD and Illumina share the same bias sensitivity

Summary

- Lessons from the genome assembly
  - Easy to map reads onto a closely related reference (always better than *de novo*)
  - Less easy to find non-matching reads and what they are (plasmids, insertion sequences, phages)
  - Contigs obtained by *de novo* assembly must be verified
  - Repeats are a nightmare in any case
  - Paired-ends help especially for *de novo* assembly!

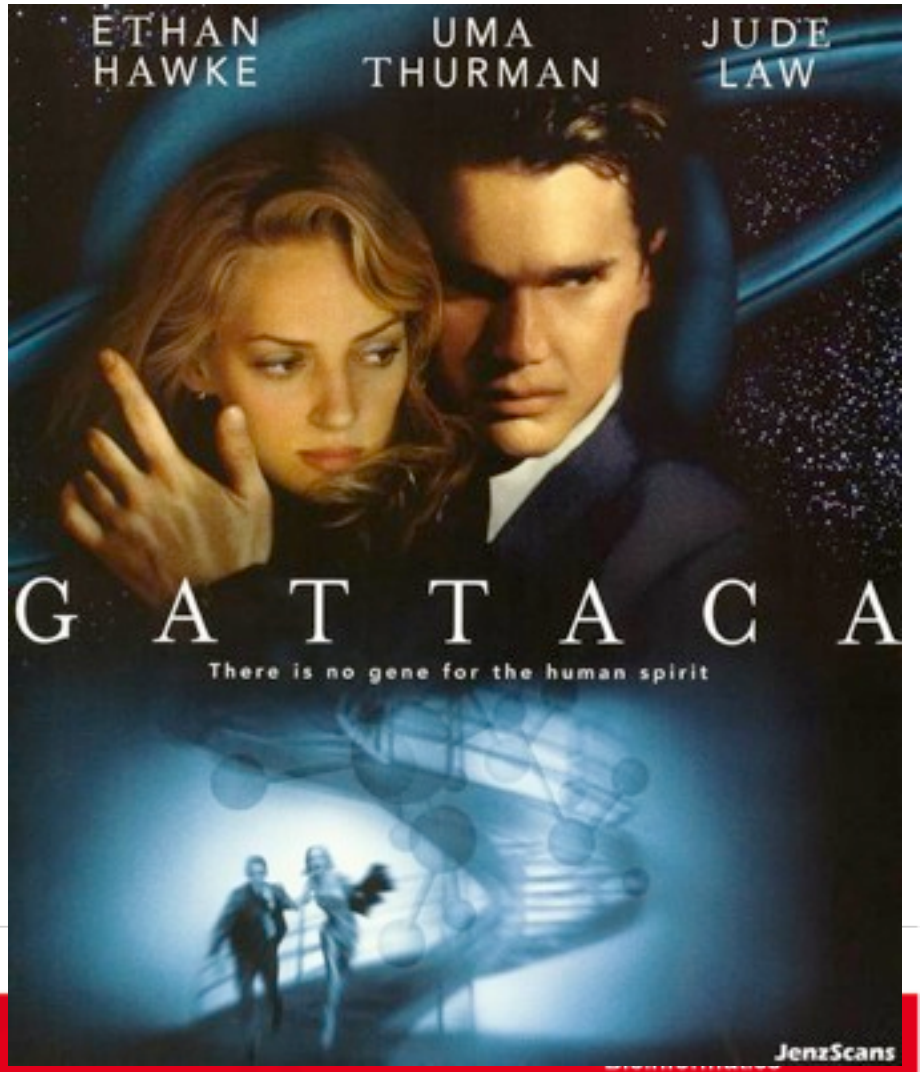
- But… should we care??

Welcome to GATTACA !

- Next-gen sequencing = last year…

- Next-next-gen = this year
- Illumina HiSeq2000
- SOLiD 4hq

- Next-next-next-gen…
- Pacific Bioscience
- Zero-mode waveguide (ZMW) nano structure
- Up to 50’000 bp reads of single molecule at 99.3% accuracy
- 2011: 160’000 ZMW per chip
- By 2014: 1 Mio ZMW per chip!! (99,999% accurate human genome in 30 minutes!)



Thank You