From chip to SNP in a zip with Parsnp

Todd J Treangen

Institut Pasteur
Paris, France

September 25th, 2014

Homeland Security

Science and Technology

Scaling genome alignment up to 10K genomes and beyond

Todd J Treangen

Institut Pasteur
Paris, France

September 25th, 2014

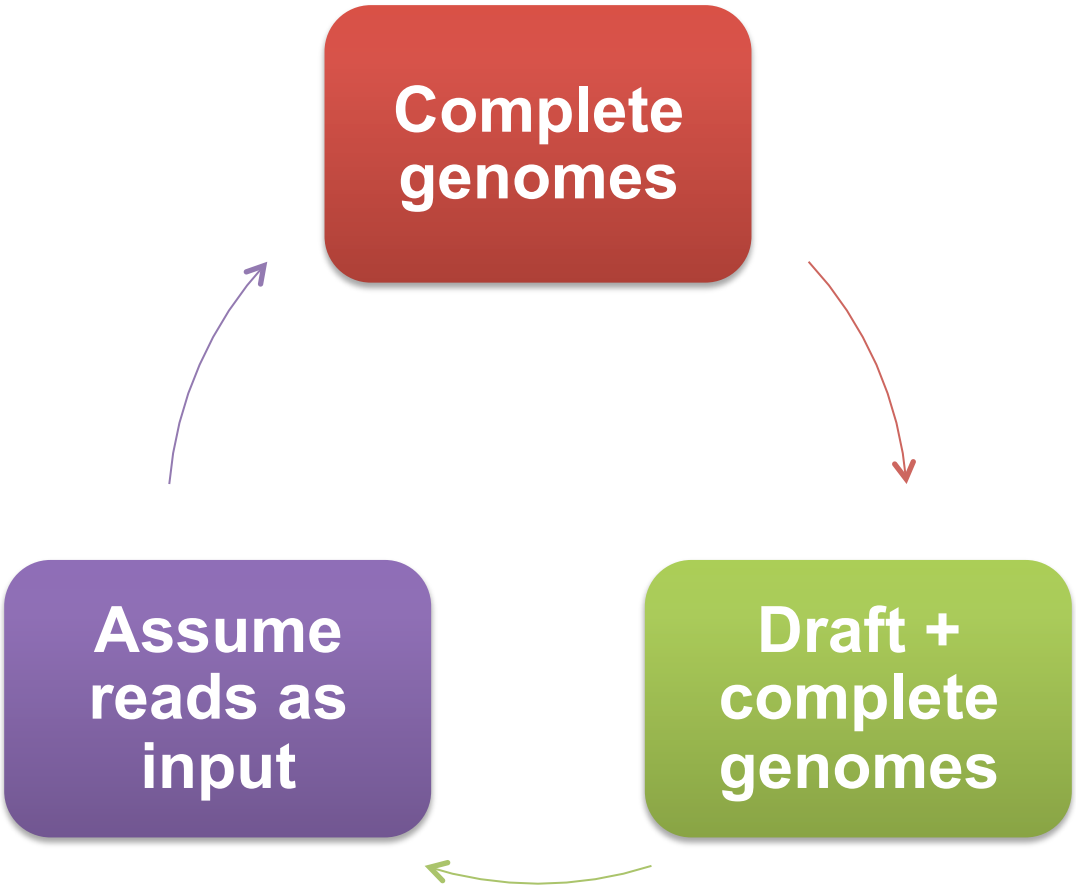
Homeland Security

Science and Technology

Acknowledgement

This work was funded under Agreement No. HSHQDC-07-C-00020 awarded to Battelle National Biodefense Institute by the Department of Homeland Security Science and Technology Directorate (DHS/S&T) for the management and operation of the National Biodefense Analysis and Countermeasures Center a Federally Funded Research and Development Center. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security or the U.S. Government. The Department of Homeland Security does not endorse any products or commercial services mentioned in this presentation.

Historical perspective: expected input



Genome revolution

- **Enabled by:**
  - **Long reads + correction methods:**
    - **PacBio, 5-6kbp+ reads**
    - **Routinely generating single contig assemblies**
  - **Affordable sequencing:**
    - **MiSeq, 2x300bp**
    - **HiSeq, 2x150bp**
  - **Push-button assembly**
- **Soon to be a major factor:**
  - **Oxford Nanopore**

Outline

① **From chip to assembly**

② From assembly to SNP

High-confidence assembly

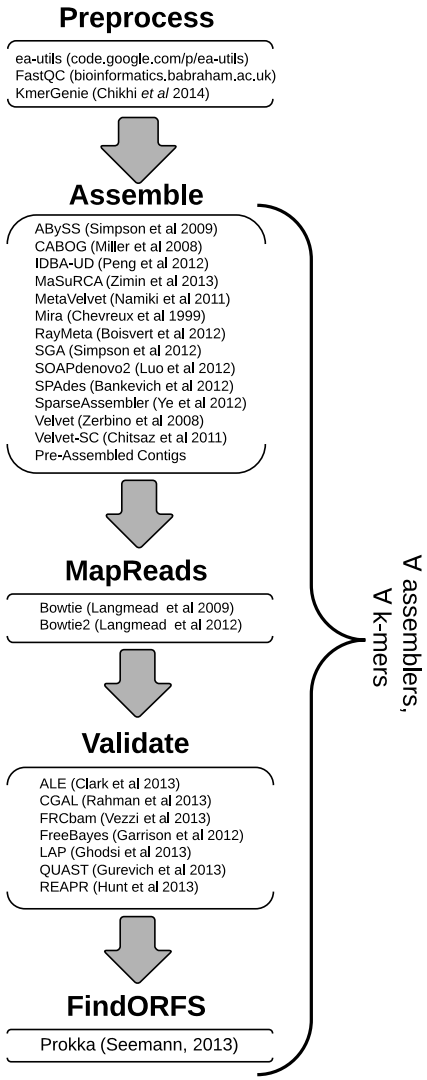
– **Assembly hard, validation easy**
  - Assembly is chaotic
  - Assembly as a hypothesis

– **Ensemble assembly** and validation
  - Each run is an assembly competition
  - Tolerant of tool failures
  - Can beat expert users

:metAMOS:

a metagenomic assembly pipeline for AMOS

- **iMetAMOS: automated sample analysis**
  - Promote and record assembly best practices
    - e.g. Everything is a metagenome
  - Reproducible workflows
  - Low startup cost (frozen binary)



Koren, S., Treangen, T. J., Hill, C. M., Pop, M., & Phillippy, A. M. (2014)
Automated ensemble assembly and validation of microbial genomes. *BMC Bioinformatics*.

GAGE in a box

- GAGE-B *Rhodobacter sphaeroides* MiSeq dataset

Assembler	GAGE-B rank	GAGE-B N50	GAGE-B CN50	iMetAMOS rank	iMetAMOS N50	iMetAMOS CN50
ABySS	5	21,441	21,307	5	38,322	36,101
MIRA	6	15,792	15,190	4	52,034	46,977
MaSuRCA	1	130,714	119,839	1	163,762	139,231
SGA	7	9,108	9,055	7	3,657	3,655
SOAPdenovo2	3	33,491	32,605	3	87,036	65,337
SPAdes	2	118,093	89,065	2	118,214	81,505
Velvet	4	23,979	23,230	6	19,652	19,355

Ensemble Assembly Validation

:metAMOS:

Treangen TJ, Koren S, et. al. Genome Biol. 2013 Jan 15;14(1):R2. PMID: 23320958.

- Preprocess OK
- Assemble OK
- MapReads OK
- Validate OK
- FindORFS OK
- FindRepeats SKIPPED
- Scaffold SKIPPED
- FindScaffoldORFS OK
- Abundance OK
- Annotate OK
- FunctionalAnnotation OK
- Propagate SKIPPED
- Classify OK
- Browse Results OK

Pipeline summary
Run commands

Version: v1.3 Candied Bacon
workflows: core,isolate
Created: 06 Nov 2013, 5:29PM

Selected assembler: spades.47
Selected reference: Mycobacterium_tuberculosis_EAI5_NITR206_uid202218
Sample may have contaminants, only 97.57% contigs/reads >100bp assigned to Mycobacterium. Check Annotate output.

Assembly	Lap	Ale	Cgal	Snp	Frcbam	Orf	N50	Reapr
SPADES.47	-8.528025	-172424658.572994	-1205000000.000000	797	N/A	5109	38,016	N/A
VELVET.47	-8.768155	-226035245.216896	-1050360000.000000	217	N/A	4471	6,610	N/A
MASURCA.47	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

14,627,217	Reads
3,916	Contigs
3,916	Scaffolds
5,109	ORFs
0	Motifs

QUAST report

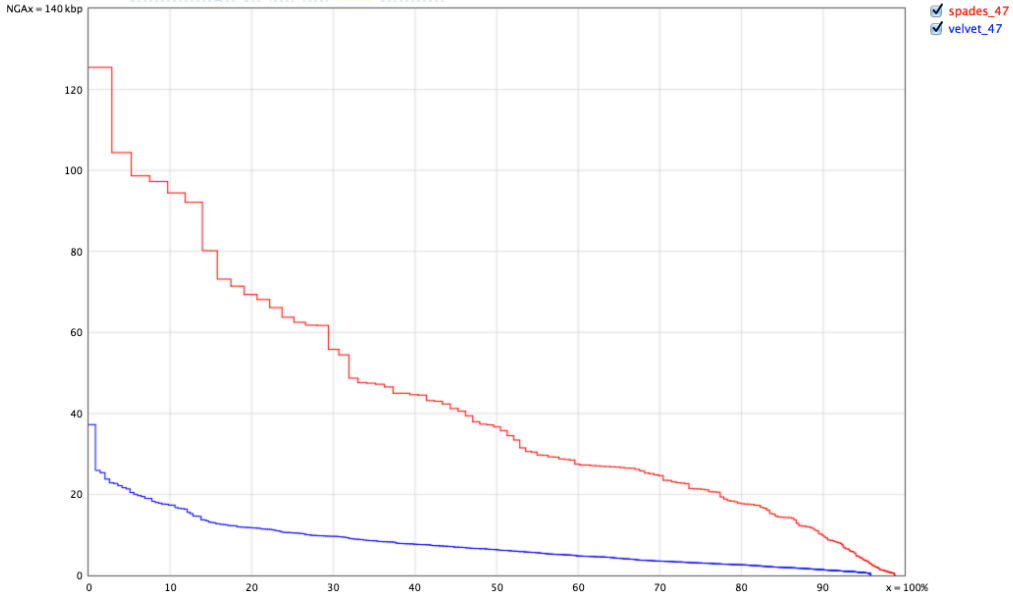
12 November 2013, Tuesday, 17:20:46

All statistics are based on contigs of size >= 500 bp, unless otherwise noted (e.g., "# contigs (>= 0 bp)" and "Total length (>= 0 bp)" include all contigs.)

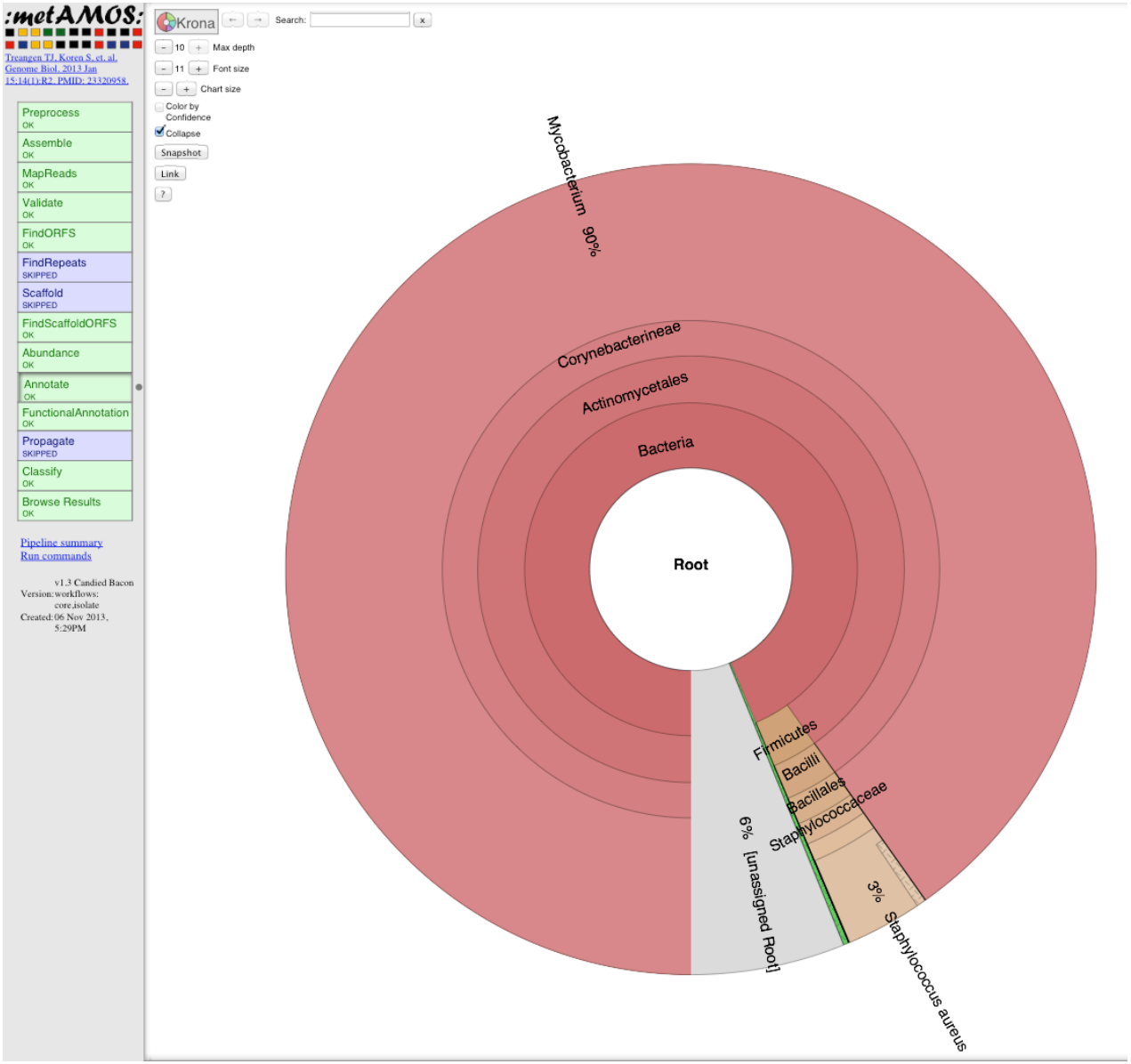
Short report worst.....best

Statistics without reference	spades_47	velvet_47
# contigs	1627	1045
# contigs (>= 0 bp)	1627	1045
# contigs (>= 1000 bp)	531	874
Largest contig	125 443	37 235
Total length	5 541 211	4 248 713
Total length (>= 0 bp)	5 541 211	4 248 713
Total length (>= 1000 bp)	4 792 027	4 120 067
N50	27 266	6656
N75	5902	3379
L50	52	195
L75	137	418
GC (%)	58.52	65.27
Misassemblies		
# misassemblies	16	8
# relocations	16	8
# translocations	0	0
# inversions	0	0
# misassembled contigs	13	7
Misassembled contigs length	475 706	65 772
# local misassemblies	31	19
Unaligned		
# fully unaligned contigs	1359	9
Fully unaligned length	1 173 848	15 804
# partially unaligned contigs	12	16
# with misassembly	1	0
# both parts are significant	3	6
Partially unaligned length	15 870	16 039
Mismatches		
# mismatches	2509	2407
# indels	1303	1058
Indels length	1979	1581
# mismatches per 100 kbp	58.81	57.57
# indels per 100 kbp	30.54	25.31
# short indels (<= 5 bp)	1276	1033
# long indels (> 5 bp)	27	25
# N's	0	0
# N's per 100 kbp	0	0
Genome statistics		
Genome fraction (%)	97.292	95.23
Duplication ratio	1.022	1.009
Largest alignment	125 443	37 235
NG50	37 942	6526
NG75	23 073	3136

Plots: Cumulative length Nx NAx NGx NGAx GC content



ENA TB Contamination



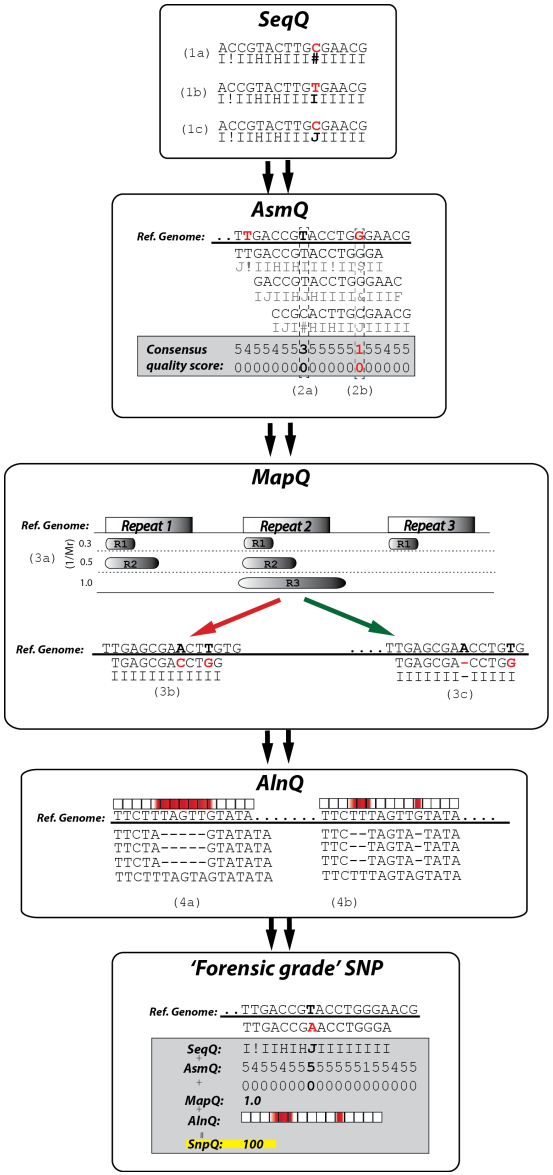
3% of samples

Outline

① From chip to assembly

② **From assembly to SNP**

High-confidence SNPs for outbreak analysis



How can we identify SNPs?

- **K-mers**
  - **No assembly required**
  - **Typically 21-31 length sequence with a wobble position in middle**
  - **Missing context, good for broad brush analysis**

How can we identify SNPs?

- **Read-mapping**
  - Many excellent tools for mapping reads to a reference genome
    - BWA, Bowtie-2, NUCmer, etc
  - No assembly required
  - Can have issues with multi-mapping & Indels
  - Primary input shifting back to genomes
  - Multiple alignment is not directly available

How can we identify SNPs?

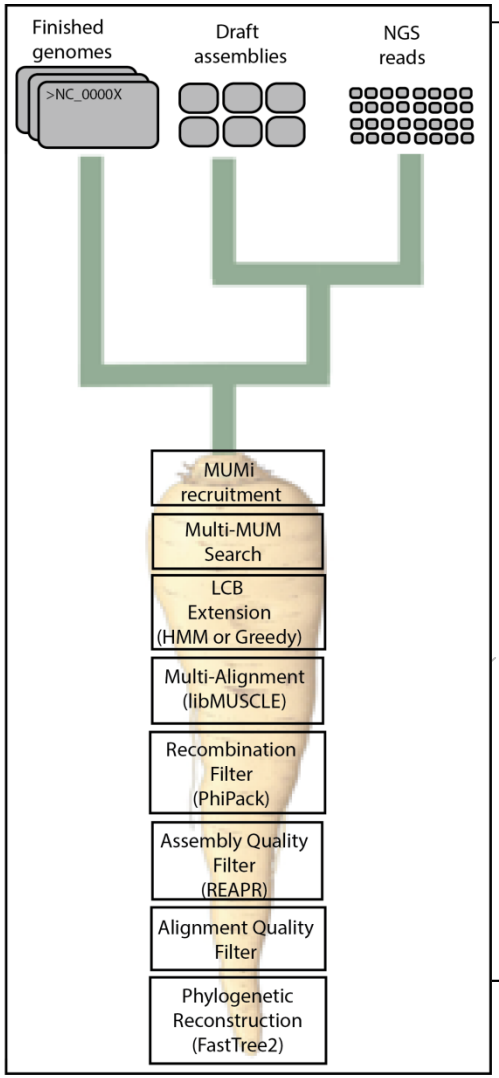
- **Whole genome alignment**
  - 1 to 1 relationship for each and every base pair
  - Strategy is typically all pairs of genomes
  - Exponential increase in runtime, not feasible to align many

How can we identify SNPs?

- **Core genome alignment**
  - 1 to 1 relationship for each and every base pair, within commonly conserved regions
  - Eliminates subset problem, focuses on well defined subproblem of genome alignment
  - Useful for phylogenetic reconstruction and rapid analyses of outbreaks
- **Parsnp built around core genome detection**

ParSNP

1. NGS reads, draft assemblies, and/or finished genomes as input
2. Near-neighbor genomes are recruited
3. **Efficient Multi-MUM search to identify locally collinear blocks**



4. Multi-Alignment of LCBs with Muscle
5. HMM-based LCB extension
6. Apply quality filters
7. Phylogenetic reconstruction with FastTree2
8. Done !

400 600 800 1000

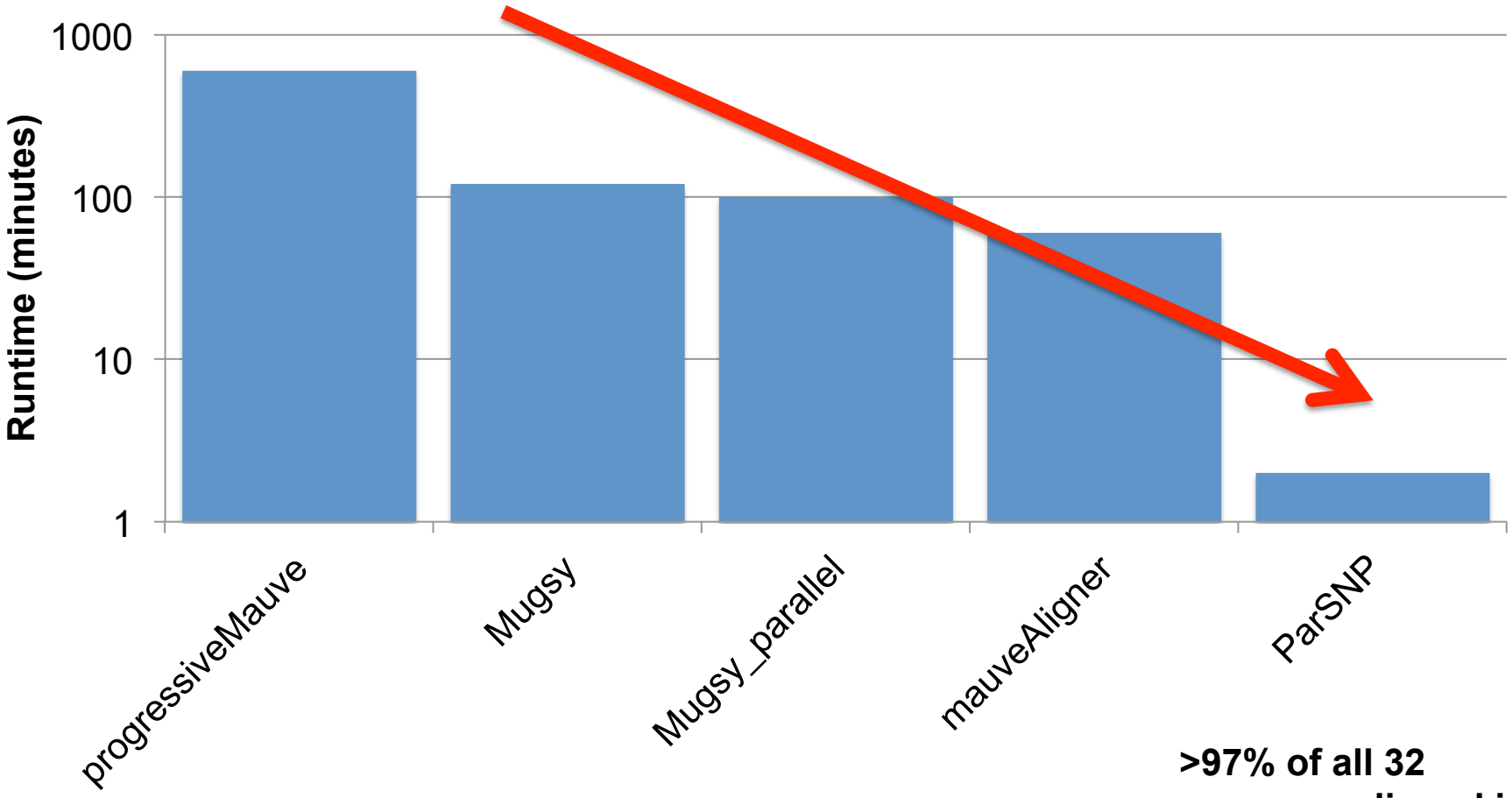
# of genomes

OK, but..

- is it efficient?

Run time performance

(32 simulated E. coli W3110 genomes)

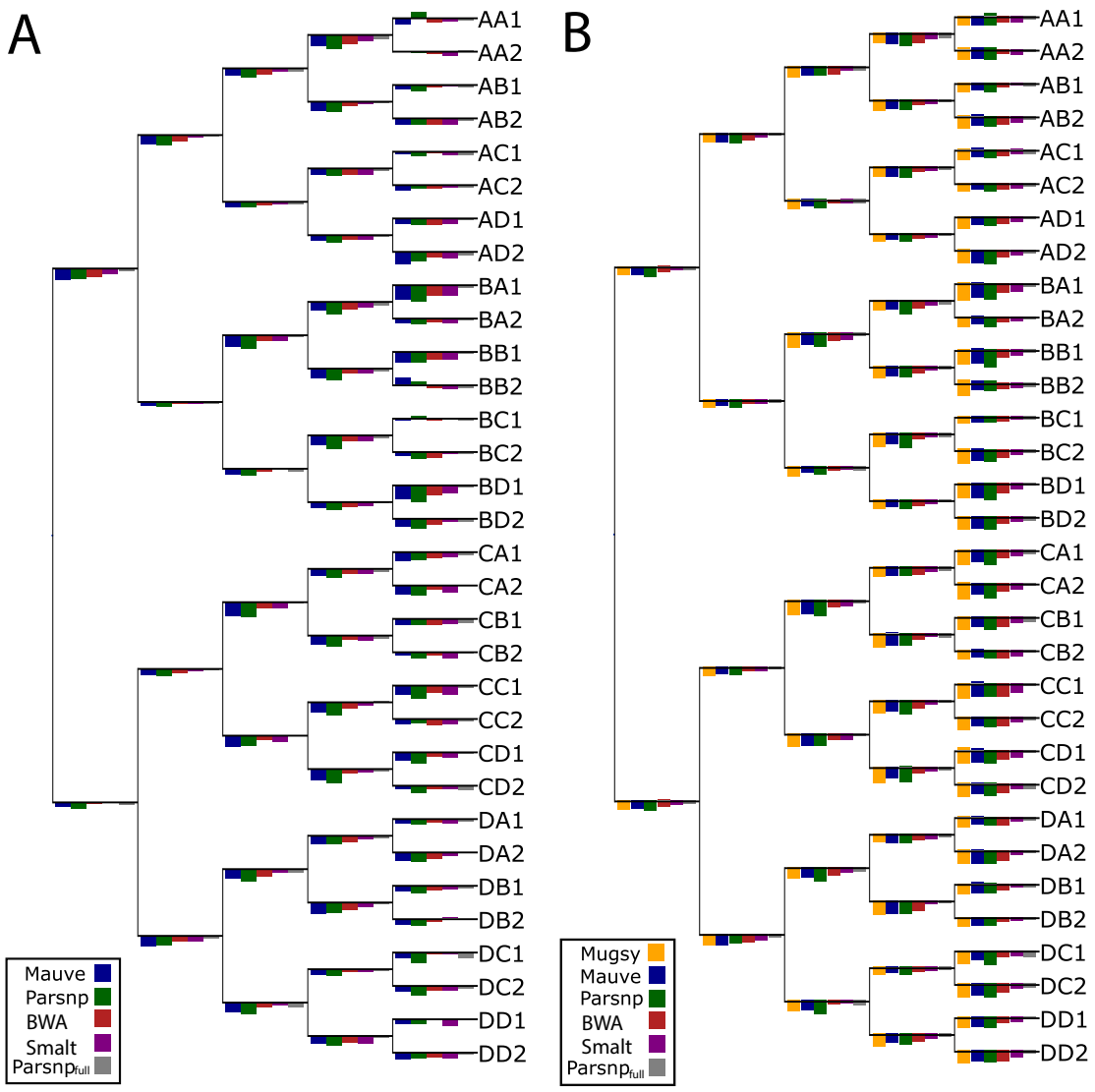


>97% of all 32 genomes aligned in less than 3 minutes!

OK, but..

- is it accurate?

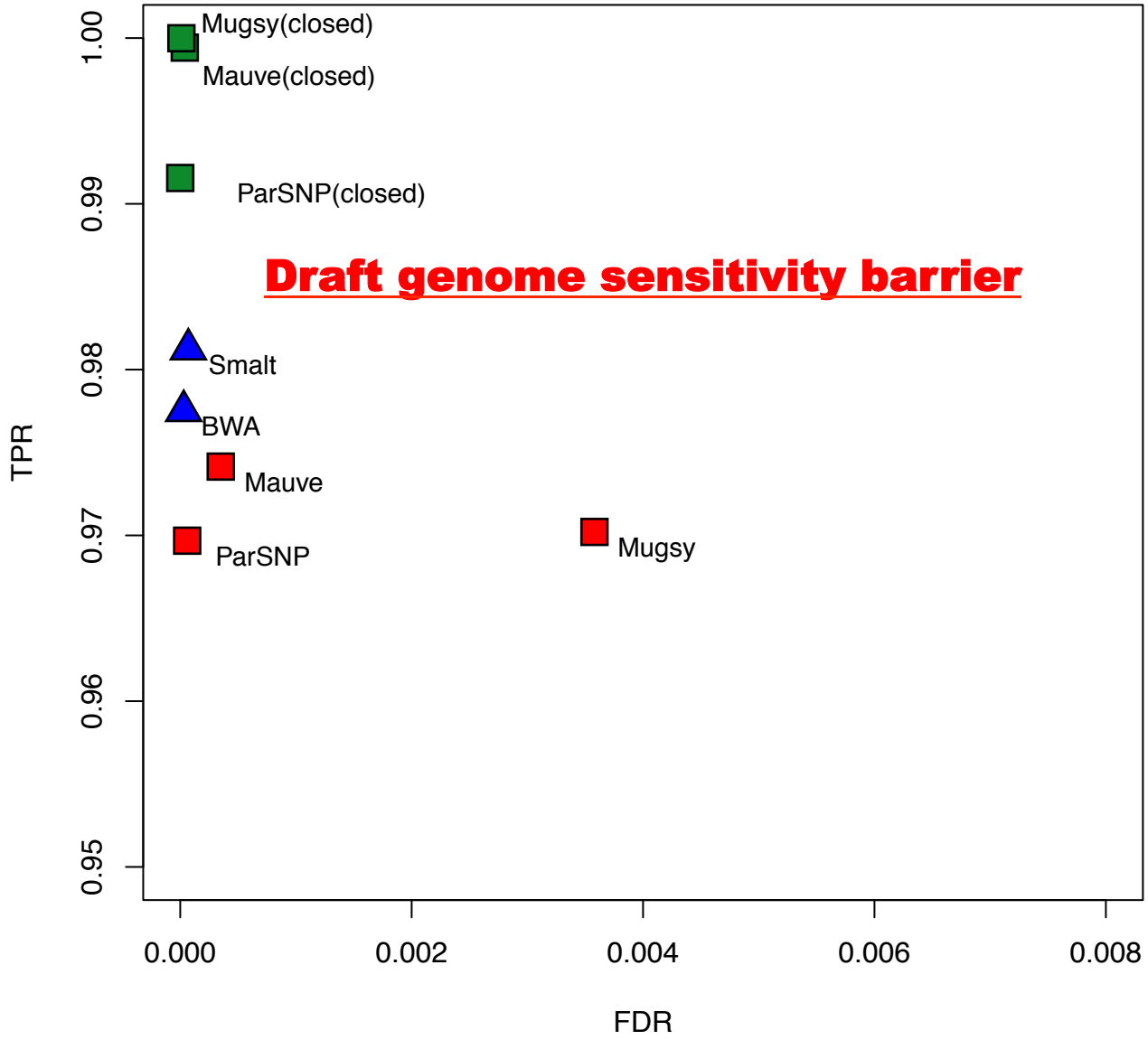
Branch SNP/length performance



OK, but..

- is it sensitive?

Sensitive, high precision SNP calls



. kSNP

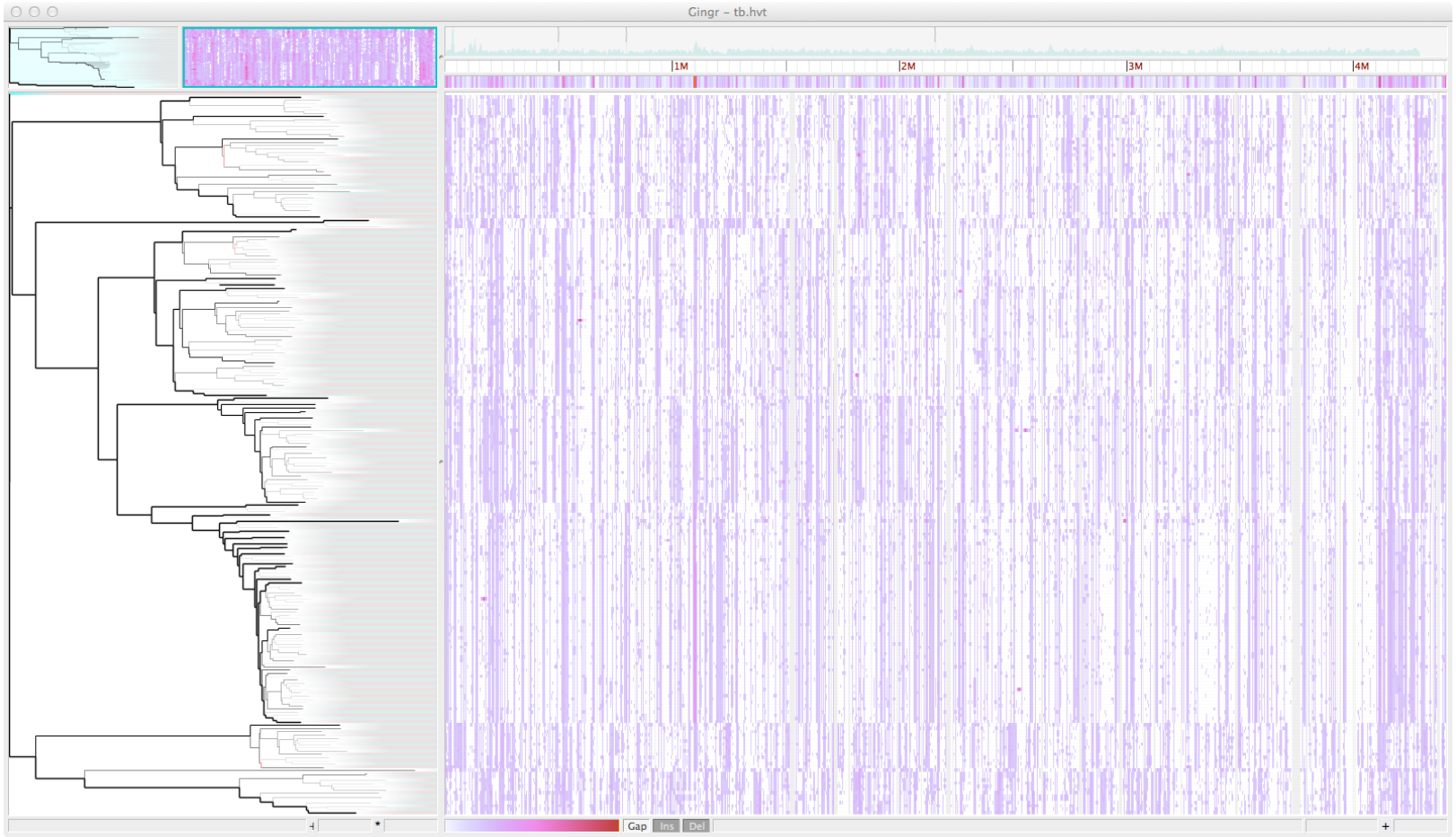
From simulated to real

- **Mycobacterium tuberculosis, "Out of Africa" study involving 200+ isolates (Comas et al 2013)**

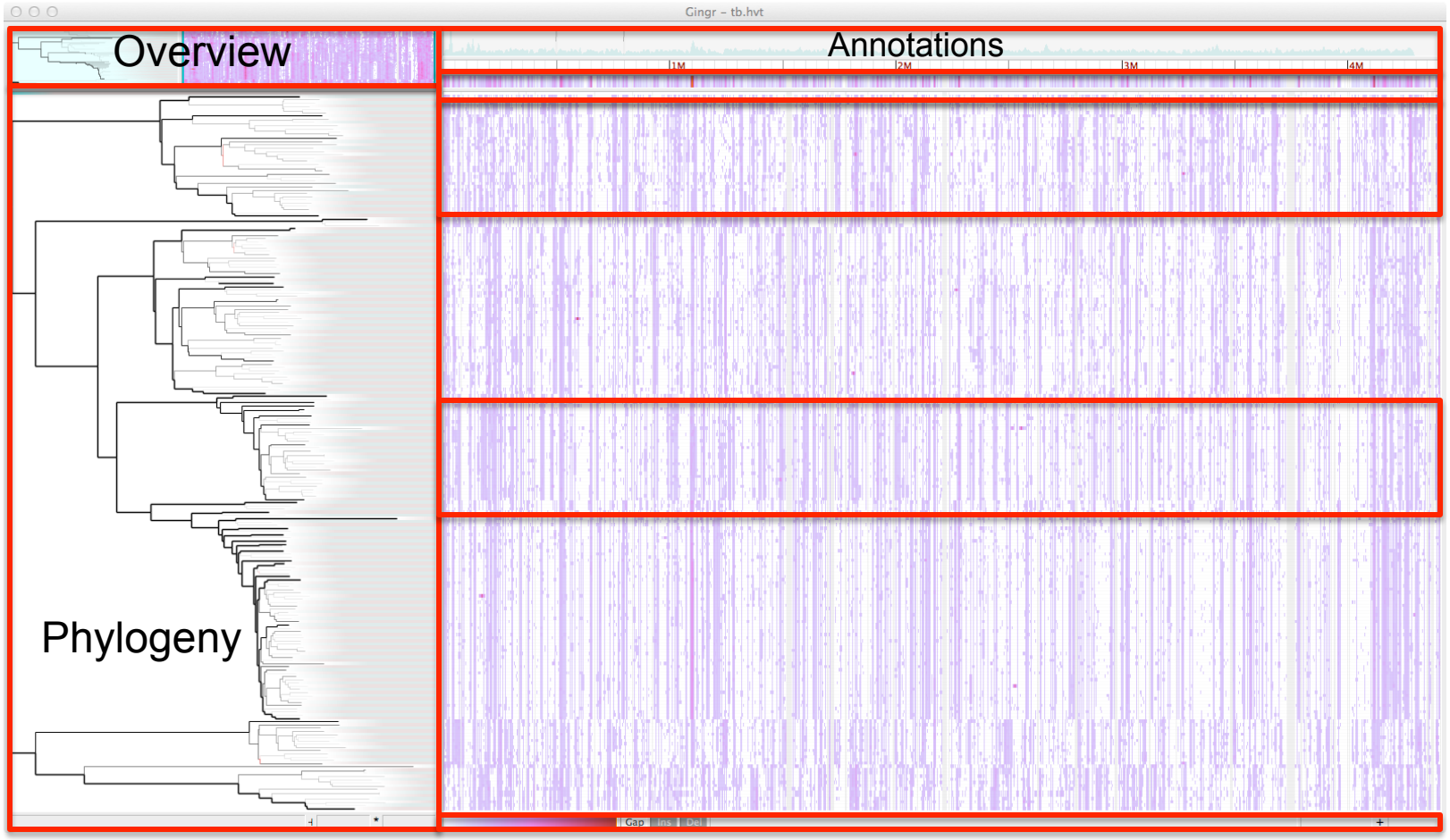
- **Downloaded all data associated with publication**
  - **Data ranged from 51bp SE reads to 100bp PE reads**

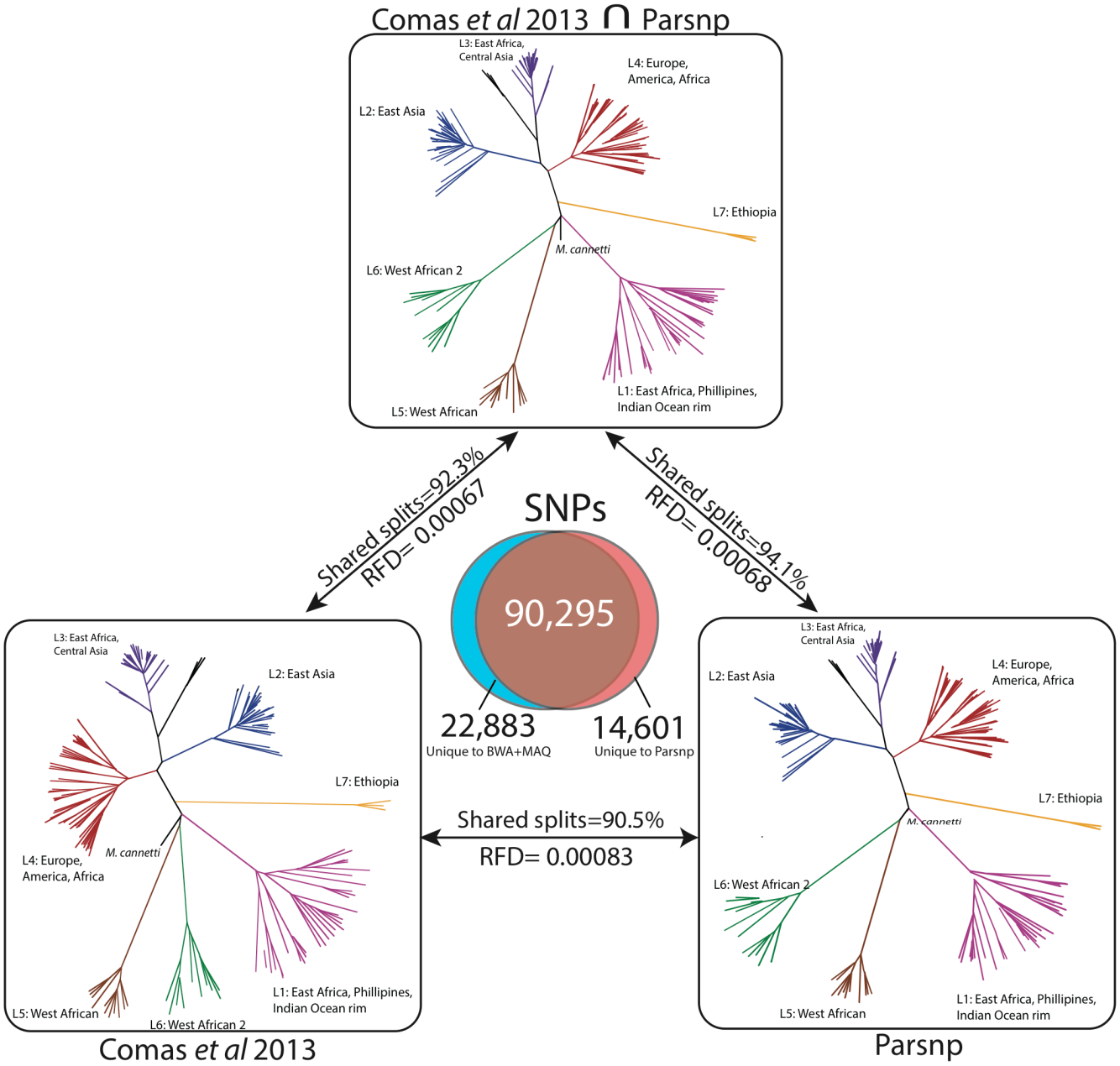
- **Questions to answer:**
  - **How does our phylogeny compare to the Comas et al study?**
  - **How sensitive are we in this "worst case" scenario?**

Gingr view of MTB alignment

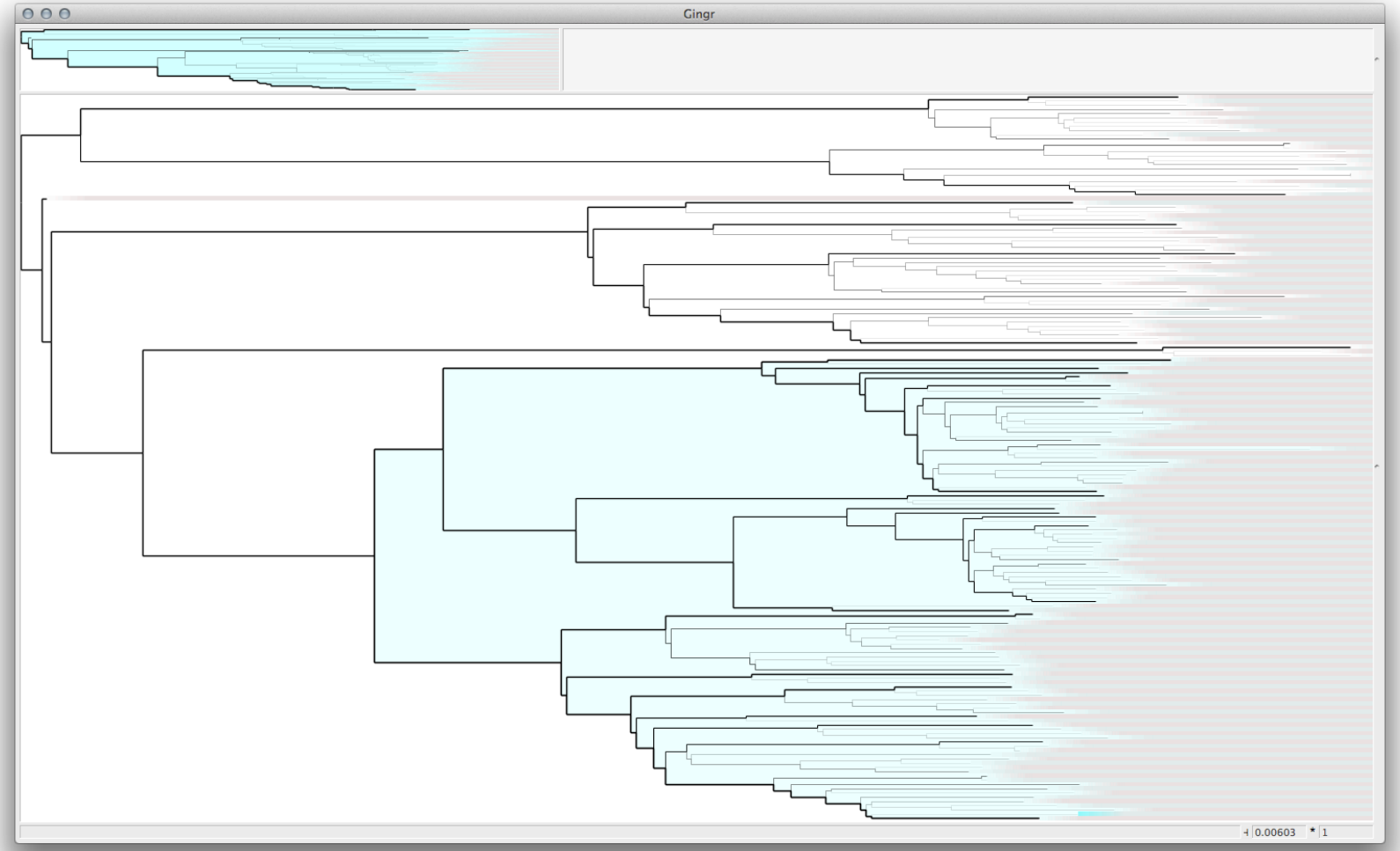


Gingr view of MTB alignment




Comas et al 2013 ∩ Parsnp
L3: East Africa, Central Asia
L4: Europe, America, Africa
L2: East Asia
L7: Ethiopia
M. cannetti
L6: West African 2
L5: West African
L1: East Africa, Phillipines, Indian Ocean rim
Shared splits=92.3%
RFD= 0.00067
Shared splits=94.1%
RFD= 0.00068
SNPs
90,295
22,883
Unique to BWA+MAQ
14,601
Unique to Parsnp
Shared splits=90.5%
RFD= 0.00083
Comas et al 2013
Parsnp

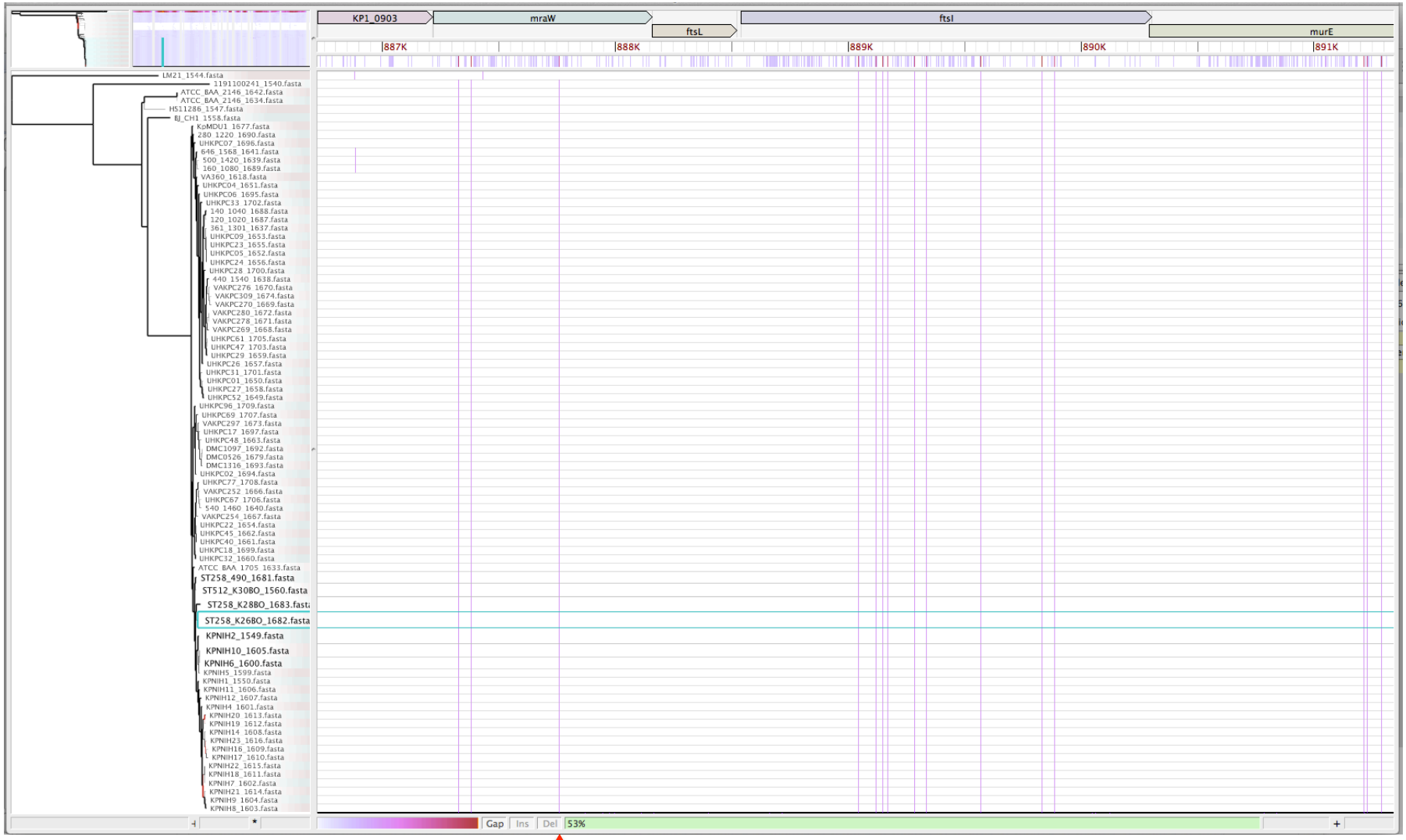
SNPs unique to each method



Gingr interactive viewer

159 *Klebsiella pneumoniae genomes*

At the gene level



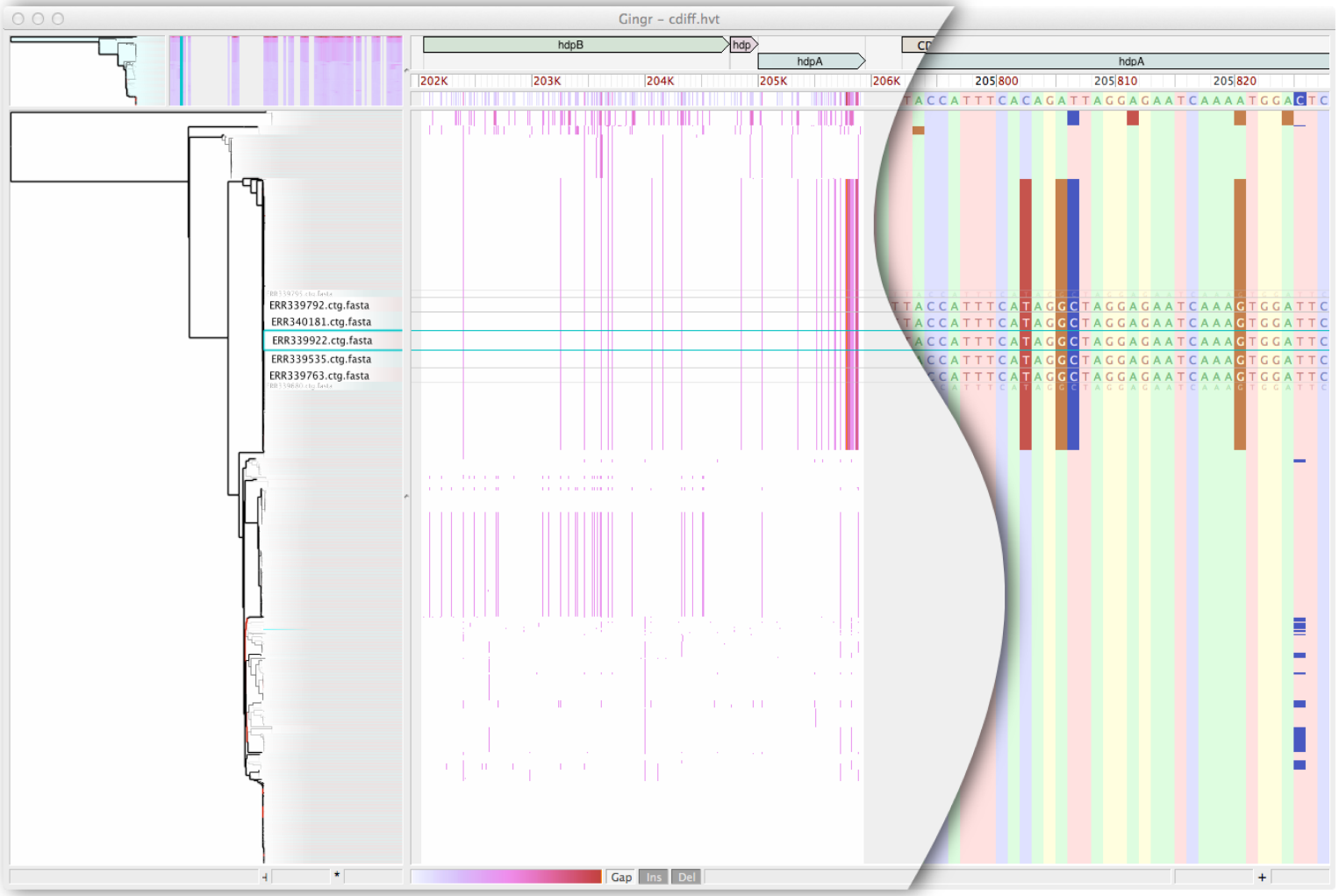
Inspecting individual SNPs (*mraW* gene)

P. difficile

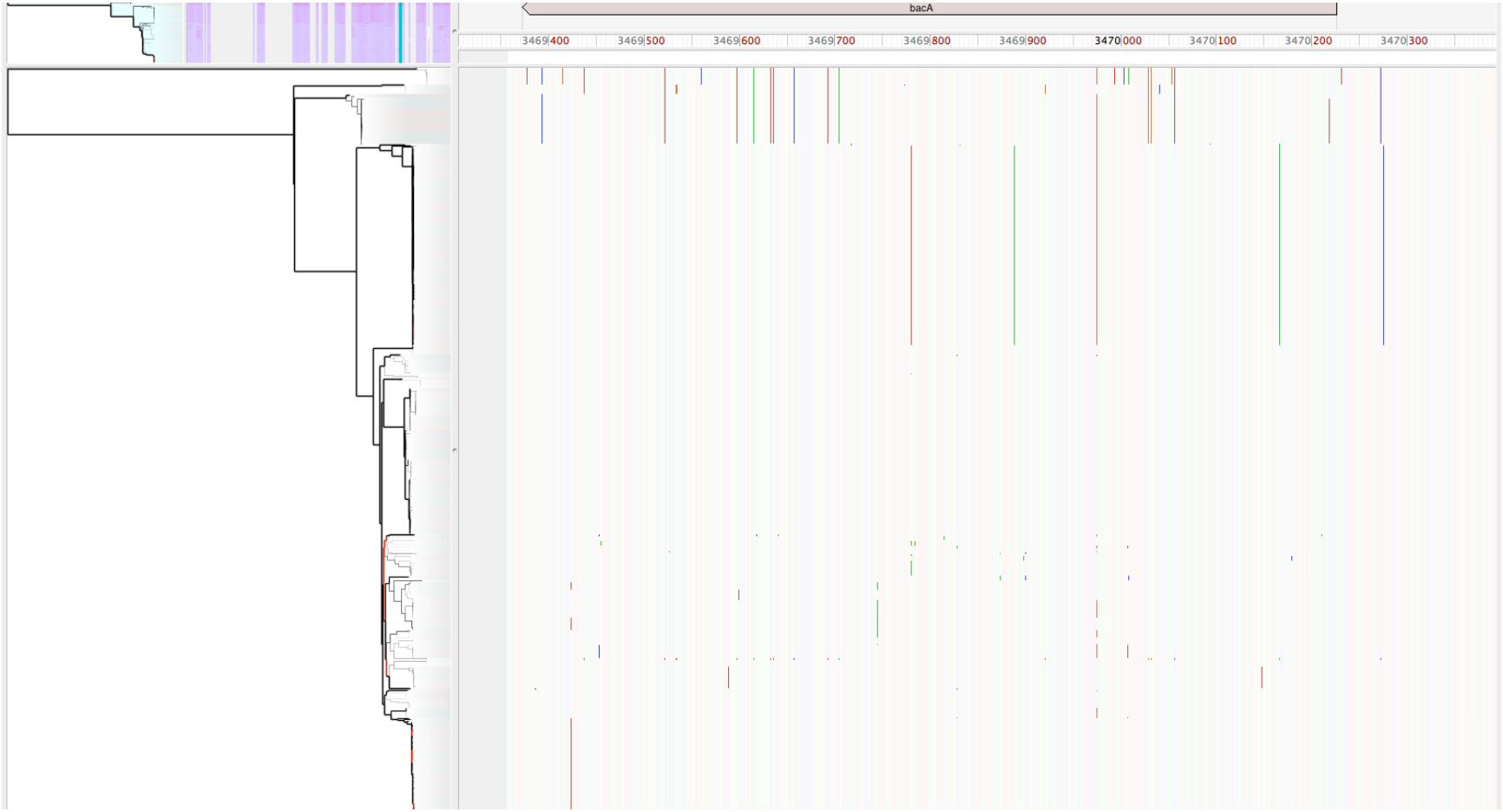
- ***Clostridium difficile → Peptoclostridium difficile***
  - ***Yitin et al, Environ Micro, 2013***

- **Study published in NEJM, downloaded sequencing data**
  - **Eyre et al NEJM, 2013**

Phylogeny, annotation & nucleotiodes



BacA gene conservation



Conclusions

① **iMetAMOS for automated assembly of MiSeq data, PBcR, PacBio data for closed genomes**

② **Align hundreds of closely-related microbial strains in *minutes***

- **32 *E. coli* genomes in 4 minutes on 8 cores**
- **200 *K. pneumoniae genomes in 20 minutes on 8 cores***
- ***800 P. dificile* genomes in 40 minutes on 32 cores**

③ **Interactive interface for simultaneous visualization of:**

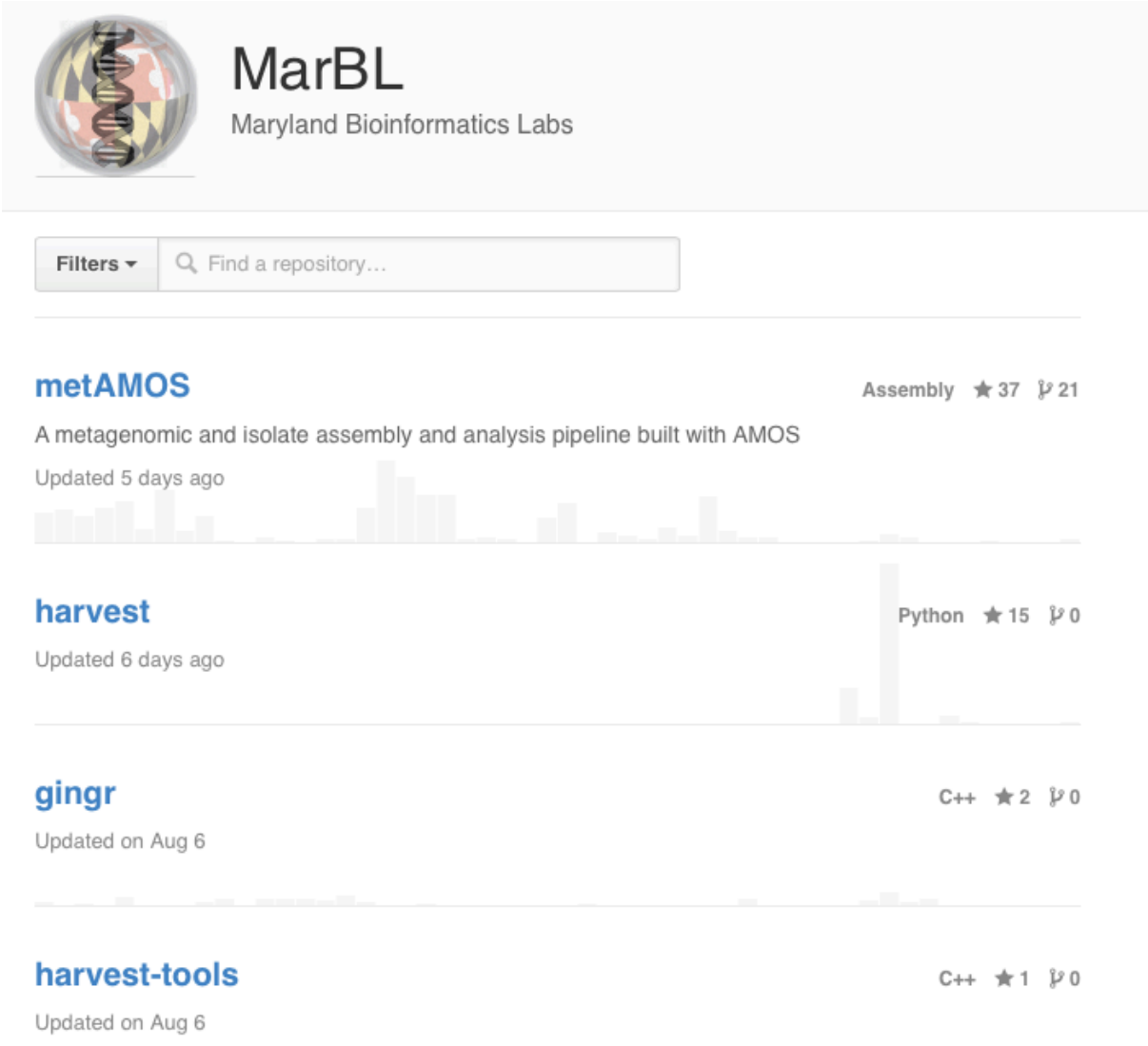
- ***core genomes***
- ***SNPs***
- ***clade phylogeny***

④ **Assembly + multiple alignment as a sustainable path to core genome SNP typing:**

- **825 genomes → 1500 GB (reads) → 3.3 GB (asms) → 0.13 GB (aln)**

Software download

- github.com/marbl

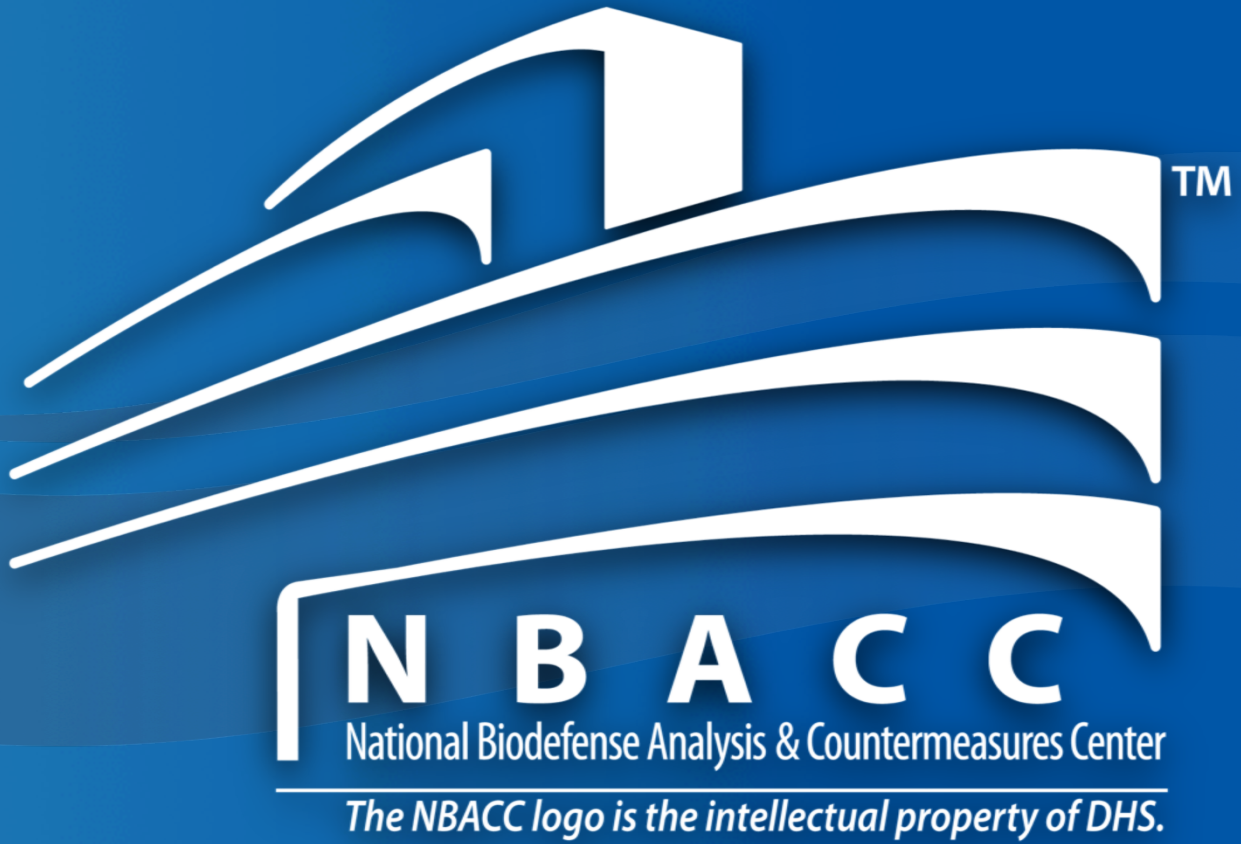


Acknowledgments

- *National Biodefense Analysis and Countermeasures Center (NBACC)*
  - Adam Phillippy
  - Brian Ondov (→Gingr!)
  - Sergey Koren

- Centre for Public Health Research (CSSIP), Valencia, Spain
  - Iñaki Comas (MTB data)

Questions?



Homeland Security

Science and Technology