Strain-Level Comparative Genomics

Phylogenomics Crash Course
Multiple Genome Alignment & Visualization

STAMPS – Day 5

July 18, 2025

Quick Phylogenetics Refresher

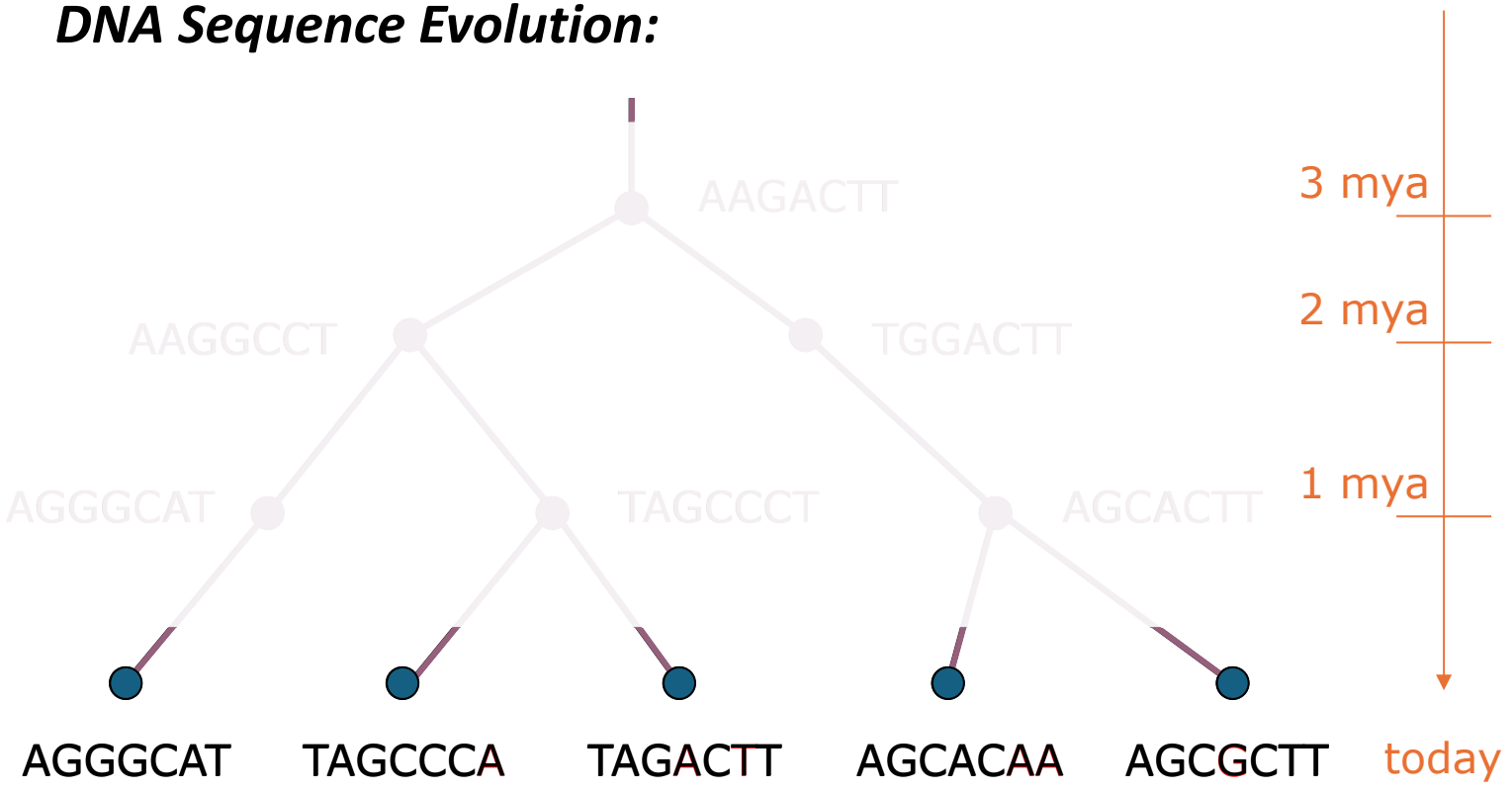
Multiple Sequence Alignment

Phylogeny Estimation

Population-level (“Species”) Tree Estimation

Brief Intro to Molecular Phylogenetics

DNA Sequence Evolution:



Notes:

- insertions and deletions also occur randomly on branches (not shown)
- In statistical terms, the extant sequences are the **observed data**, while the tree shape, tree topology and mutation rates are the **unknown model parameters** to be estimated.

Task is to estimate the tree shape from the extant sequences at the tree leaves...

Two Separate Problems (both NP-Hard):

1. Identify which groups of characters share a common ancestor. (Multiple Sequence Alignment)
2. Find the maximum likelihood tree and model parameters (ML Tree Estimation)

*mya = million years ago

*Sequence Animation Credit: Prof. Tandy Warnow

Multiple Sequence Alignment: Definition & Goal

Input: Sequences from different organisms (or different loci) that evolved from a common ancestor.

Goal: Align sequences so that all sets of positions having a common ancestor are grouped together.

- *Not the same as aligning short sequences (or reads) to a reference ("mapping").*
- *Not the same as **genome** alignment*
- *Typically done before creating a phylogenetic tree…*

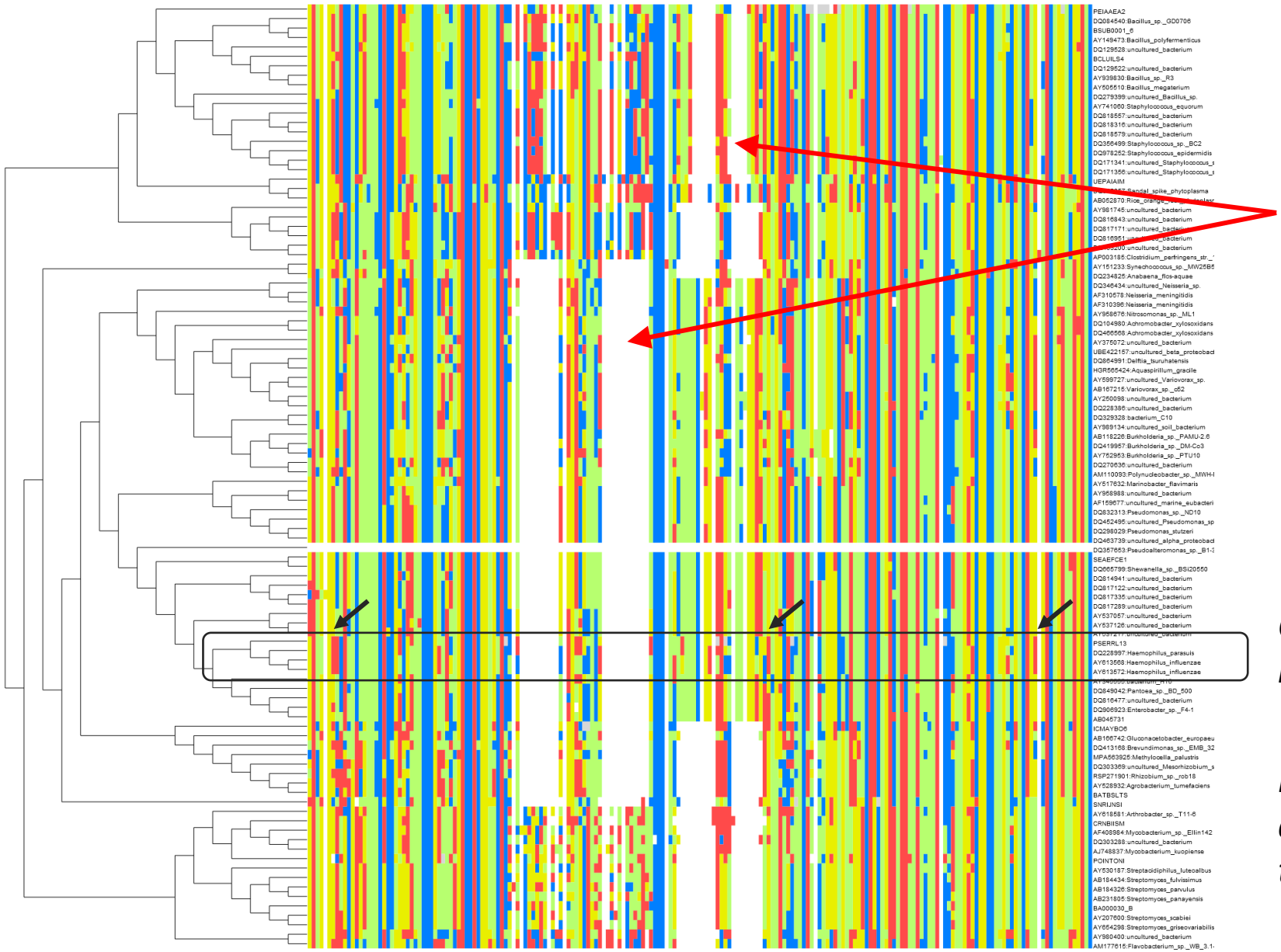
Tools:

- ***MAFFT***
- Muscle
- ***PASTA***
- ClustalW
- DiAlign
- BAli-Phy
- PRANK
- T-COFFEE
- …*et cetera*

Garbage Alignment → ***Garbage Tree***

Bold/Underlined = *My preferred methods*

Multiple Sequence Alignment (Example)



White indicates a “gap” in the alignment, a.k.a. an insertion or deletion (indel)

Conserved mutation patterns indicate evolutionary closeness.

Phylogeny estimation algorithms use this to build a tree…

Legend:

A	T
C	G

Data Properties Affecting Multiple Sequence Alignment

- Avg Sequence Similarity (rate of evolution)
- # of Sequences
- Presence of highly conserved regions
- Sequence Length heterogeneity
- Gap length/frequency
- Sequence fragmentation (*not the same as heterogeneity*)
- Avg Sequence Length

MSA Algorithms & Software (Partial List)

Tool	Use Case	Comments
MAFFT (Katoh et al., 2002)	Single Gene MSA (small N)	• Uses patterns of insertion/deletion to find optimal alignment • Generally pretty accurate in most conditions.
MUSCLE (Edgar, 2004)		• Progressive alignment. Suitable for relatively high overall sequence similarity.
CLUSTAL (Sievers et al., 2011)		• Ideal for protein alignments with structurally important sites.
PASTA (Mirarab et al., 2015)	Single Gene MSA (large N)	• Divide-and-conquer algorithm. Ideal for scaling alignment to large number of sequences (>1000)
HMMER (Eddy, 1998)	*Query sequence alignment to reference*	*• Represents reference alignment as HMM. Query sequence alignment performed using standard HMM algorithms.*

I tend to tell people:

- *Just run MAFFT for anything less than 500 sequences or so*
- *Muscle is fine too if the divergence is low…*
- *…or ClustalW for AA sequences with important structural sites*
- *Use PASTA for over 1k sequences or if avg. %-identity is very low (high rate of evolution).*

This simplistic advice is a STAMPS 2022 exclusive…

Final MSA Points (details in appendix)

- MSA accuracy can drop to 0% *fast*
  - *Especially as diversity & # sequences go up.*
- Bad MSA will lead to a Star-like tree
  - *i.e. no discernable relationships between sequences*
- MSA failure can take many forms (over/under alignment)

Tree Estimation Mechanics

- Given the MSA, how to estimate the Tree?
- Statistical Estimation:
  - Tree shape & branch length are ***parameters*** under a model of sequence evolution
  - Called GTR (generalized time reversible)
  - Each site (letter, residue, position, etc…) evolves I.I.D.
  - Mutation along a branch happens according to a continuous-time Markov process ("time" here = branch length).
  - Root is ***not identifiable***.
  - No indels under the model, MSA gaps treated as missing data.
- Under this model
  - Every tree shape/topology has a likelihood given the sequences (the "data")
  - **Maximum Likelihood Tree** will be "statistically consistent" (good)
    - I.e. given enough sites, the ML tree will be the correct one with $p \to 1$
    - Neighbor-joining, UPGMA, etc… are NOT statistically consistent (bad)
- Hence most good Tree estimators use an ML model (e.g. RAxML, FastTree, IQTree)

MGA & Visualization for Strain Analysis

Digging deep…

Introduction

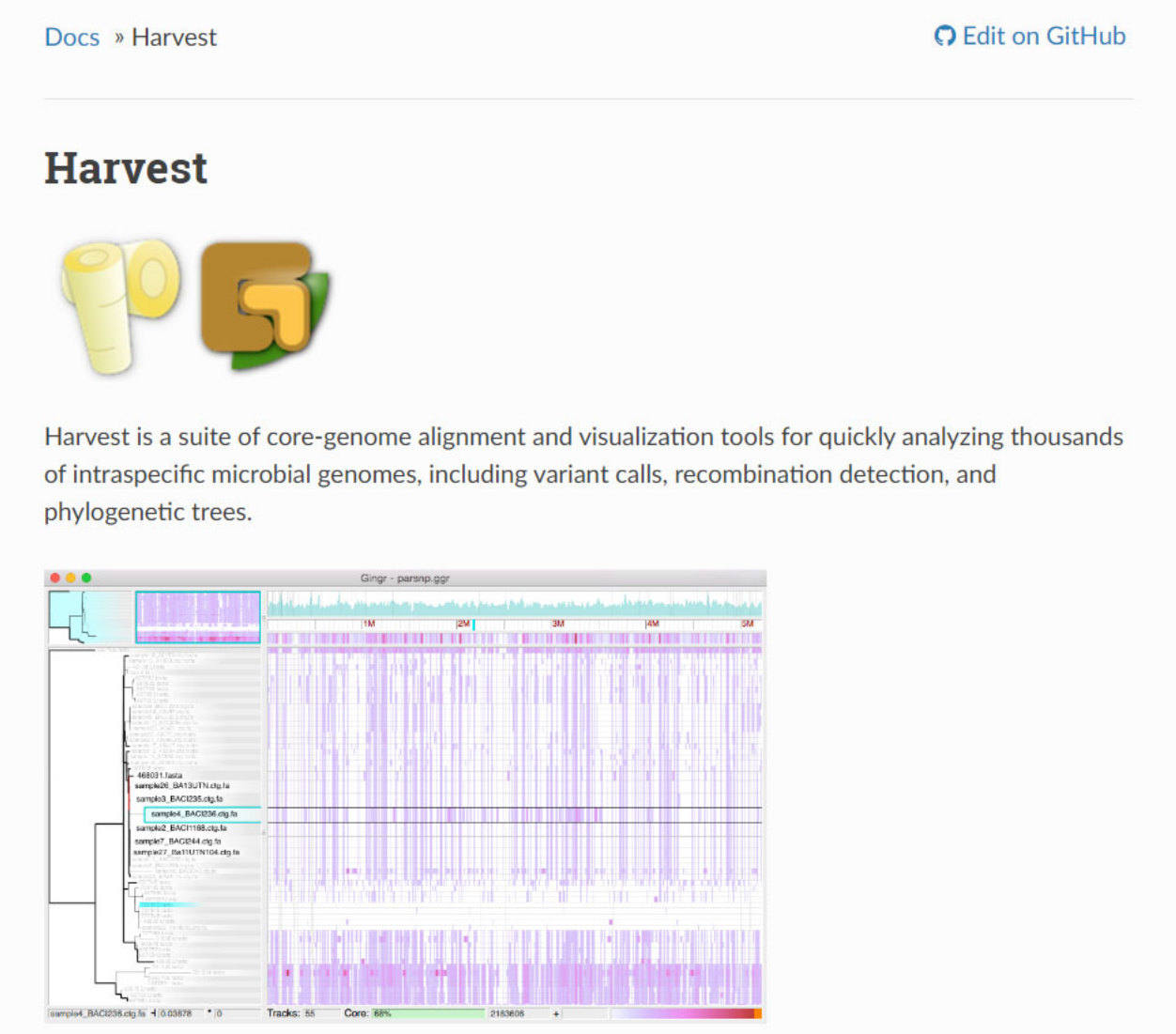
- What does “Strain” mean?
  - Particular SNP?
  - Multiple particular SNPs?
  - Presence/Absence of certain genes?
  - Phenotype?
  - *Other?*
- How do we compare strains?
  - Multiple Genome Alignment
  - Pangenome Analysis

Whole-Genome Alignment

- Idea: align specifically the *shared (“core”)* portion of several genomes.
- Use these aligned segments to identify phylogenetic relationships, etc…
- Visualize what exactly is similar and different…

Tools:

- Parsnp
- Mauve
- SibeliaZ
- (others…)



Whole Genome Alignment: Quick How-To with Parsnp

- Get *assembled* genomes from individual organisms
  - Isolates are nice, MAGs will do
  - Contigs are fine for this, doesn't have to be complete
  - Helps to have at least 1 high-quality, annotated reference genome
  - Useful to run QUAST to QC the assembly
- Run Parsnp:

```
contig_repo=./parsnp_contigs
parsnp_out=./parsnp_output_13
ref_genbank=./ref_assembly_GCF_008121495/Ref_ATCC_29149.gbff

parsnp -g $ref_genbank -d $contig_repo -p 15 -o $parsnp_out
```

Annotated Reference Genome (.gbff format)

Folder with 1 fasta file for each assembly (containing all contigs) *...OR...*
File with a newline-separated list of assembly fasta files (full

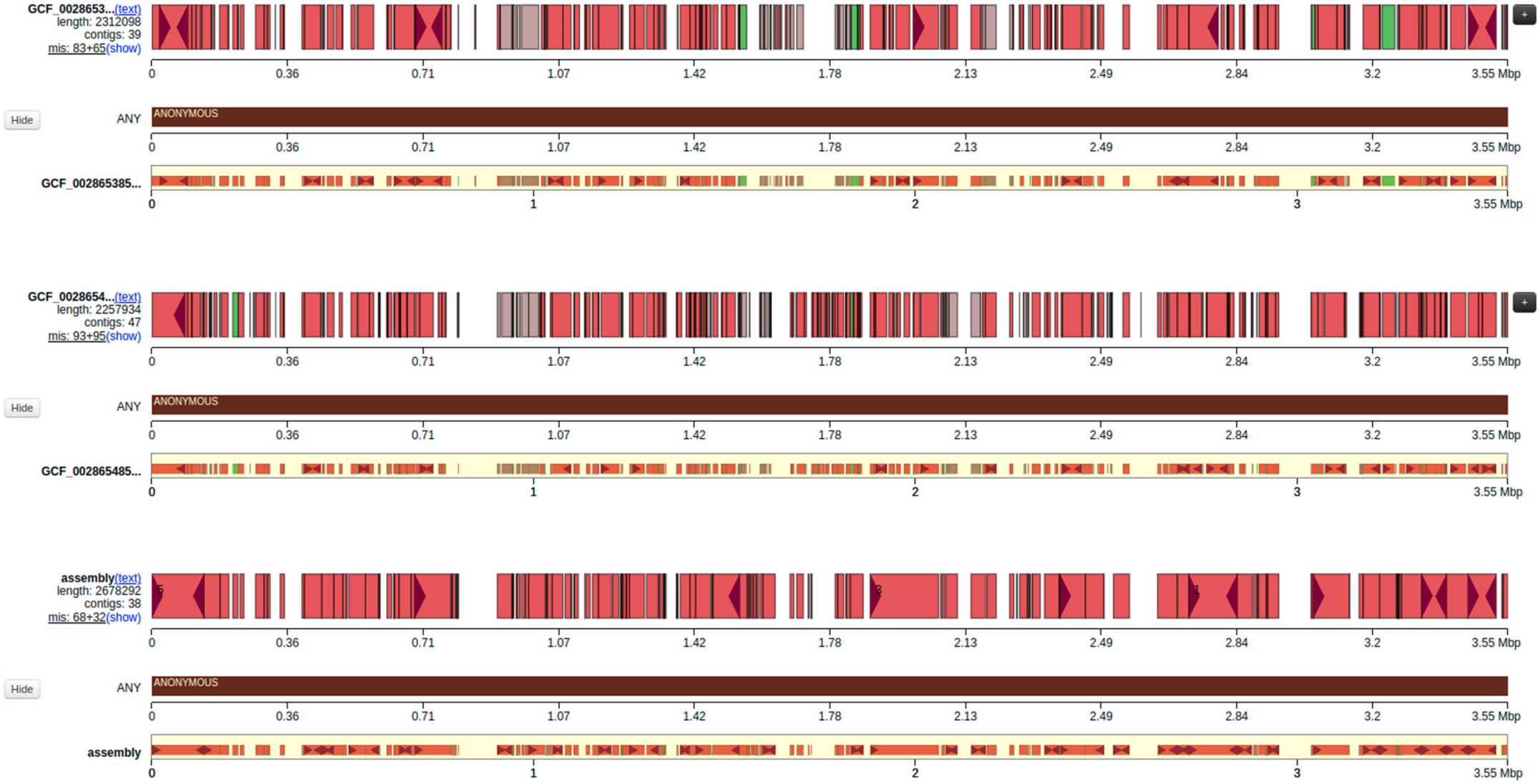
 # processors

Output folder

Interlude: QC-ing an Assembly with QUAST

What can we learn?

- Assembly Quality Issues?
- Issues with Reference?



Notes:

The top two assemblies are SPAdes assemblies done by the original authors of the R. Gnavus paper (citation later).

The bottom is a Unicycler assembly from the same reads.

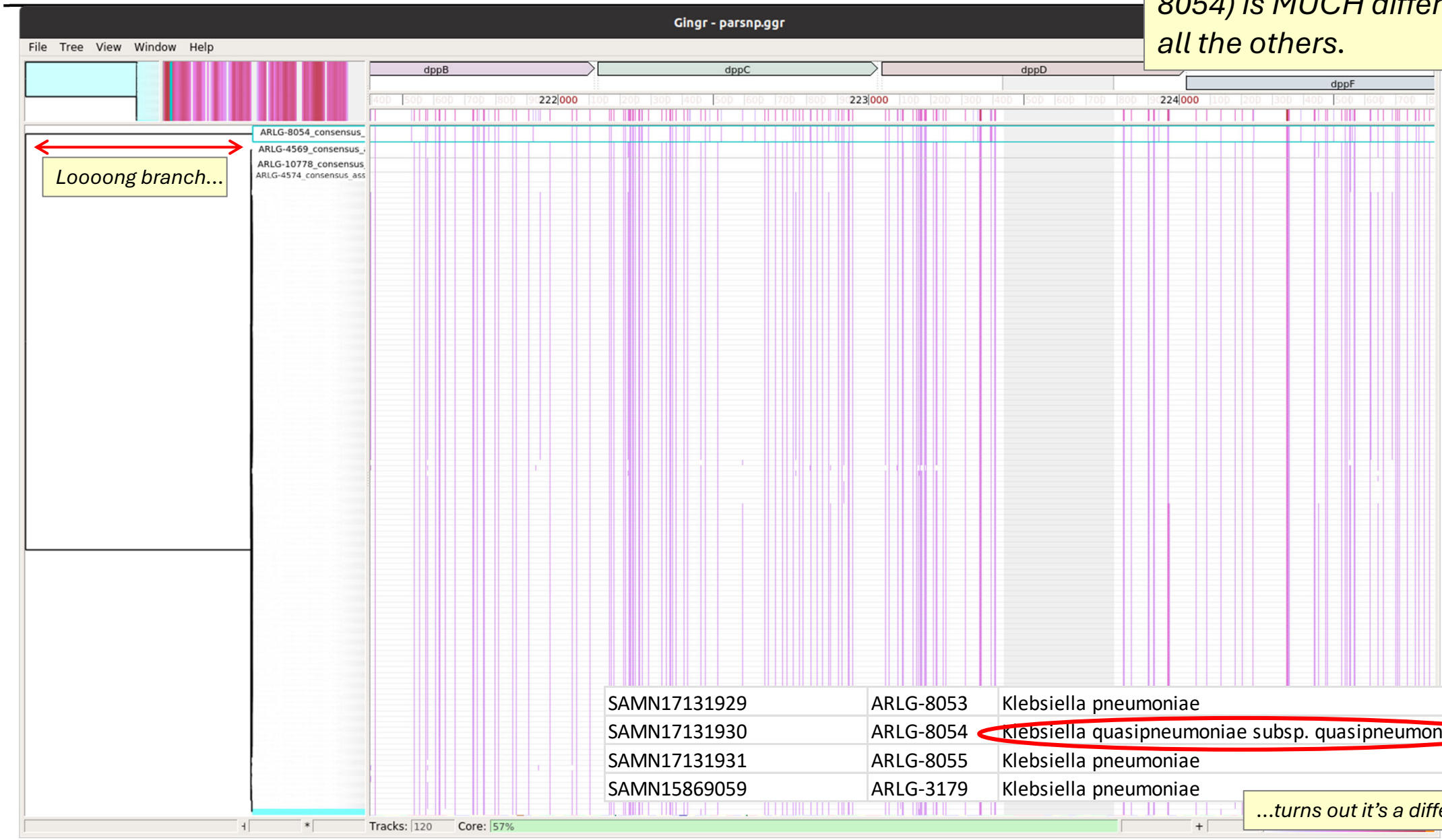
Case Study #1: *Klebsiella pneumoniae*[1]

- 119 Carbapenam-resistant *Kp* isolates (95 from Houston Hospitals)
- Hybrid short/long-read assemblies
  - so they *should* be high-quality
- Focus of study was to show spread of two separate “clonal groups”

[1]Shropshire, et al. (2022), DOI: 10.1128/mbio.00497-22

Case Study #1: Observations…

Observation #1:
One of these genomes (ARLG-8054) is MUCH different from all the others.



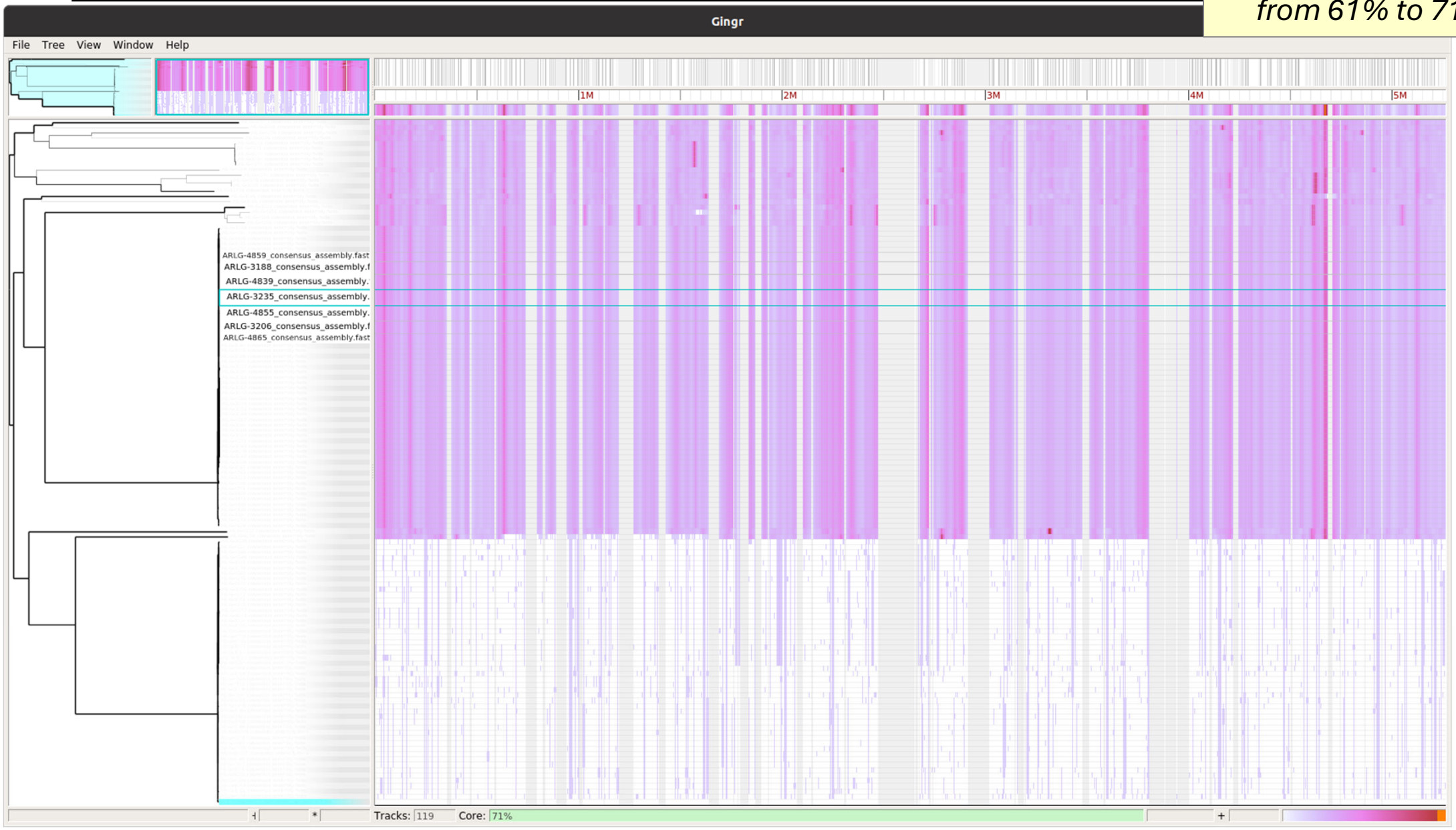
SAMN17131929	ARLG-8053	Klebsiella pneumoniae	TX
SAMN17131930	ARLG-8054	Klebsiella quasipneumoniae subsp. quasipneumoniae	TX
SAMN17131931	ARLG-8055	Klebsiella pneumoniae	TX
SAMN15869059	ARLG-3179	Klebsiella pneumoniae	MI

…turns out it's a different species

Case Study #1: Excluding ARLG-8054

***Observation #2*:**

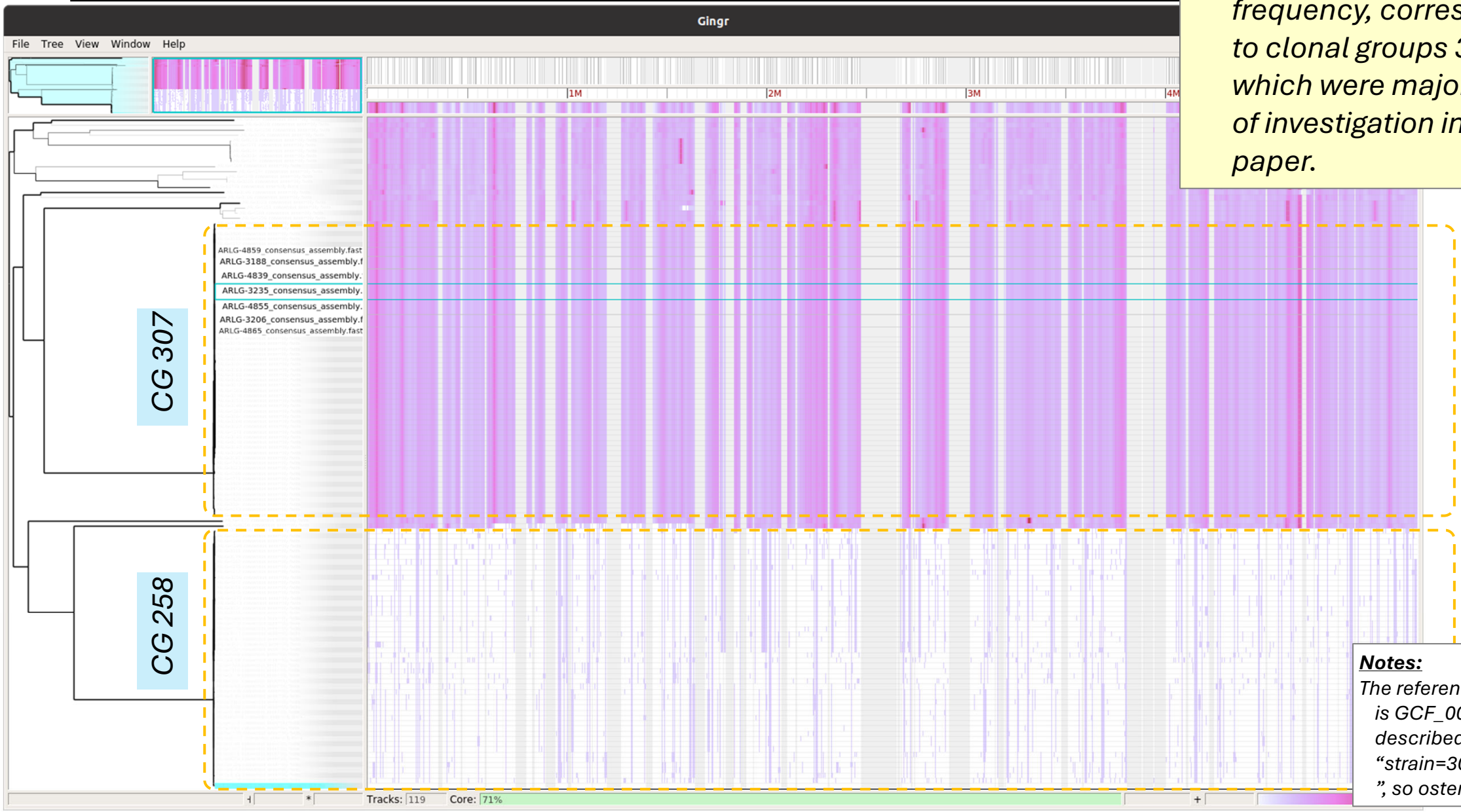
- *Without 8054, core % goes from 61% to 71%*



Case Study #1: Excluding ARLG-8054

Observation #3:

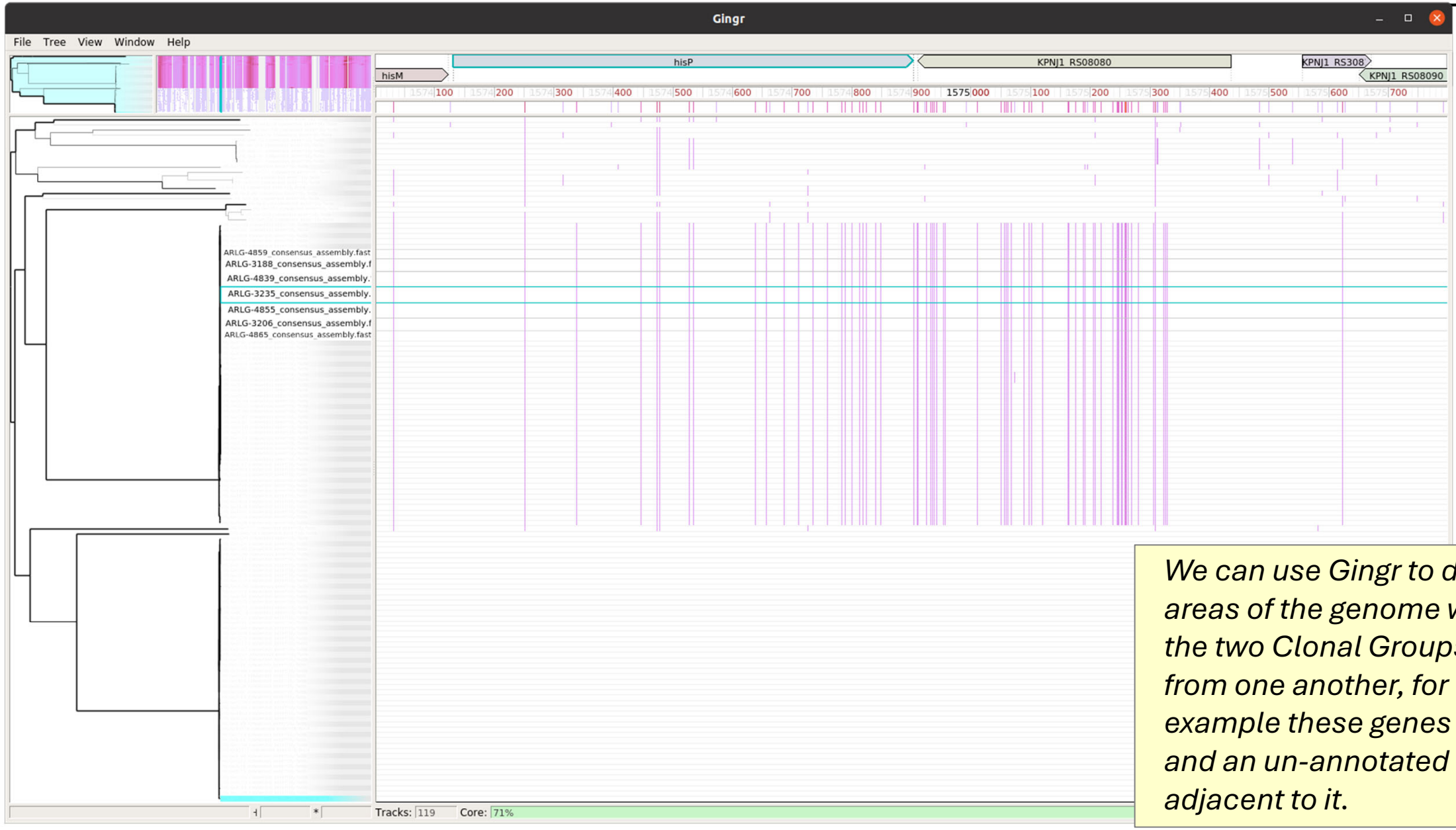
- *2 Clades with highest frequency, corresponding to clonal groups 307, 258 which were major targets of investigation in this paper.*



Notes:

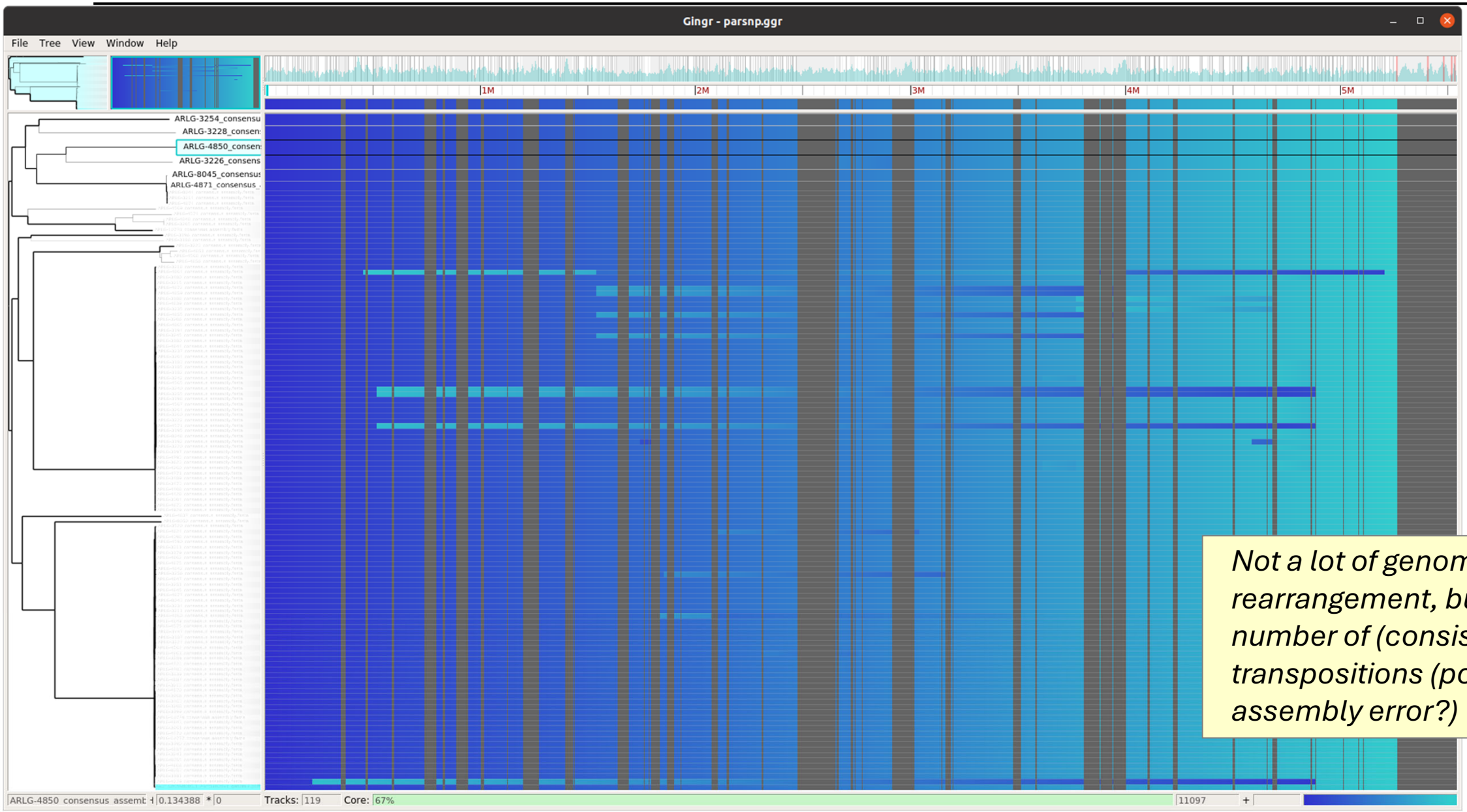
The reference genome here is GCF_000598005.1, described as "strain=30660/NJST258_1", so ostensibly ST258.

Case Study #1: Digging In…



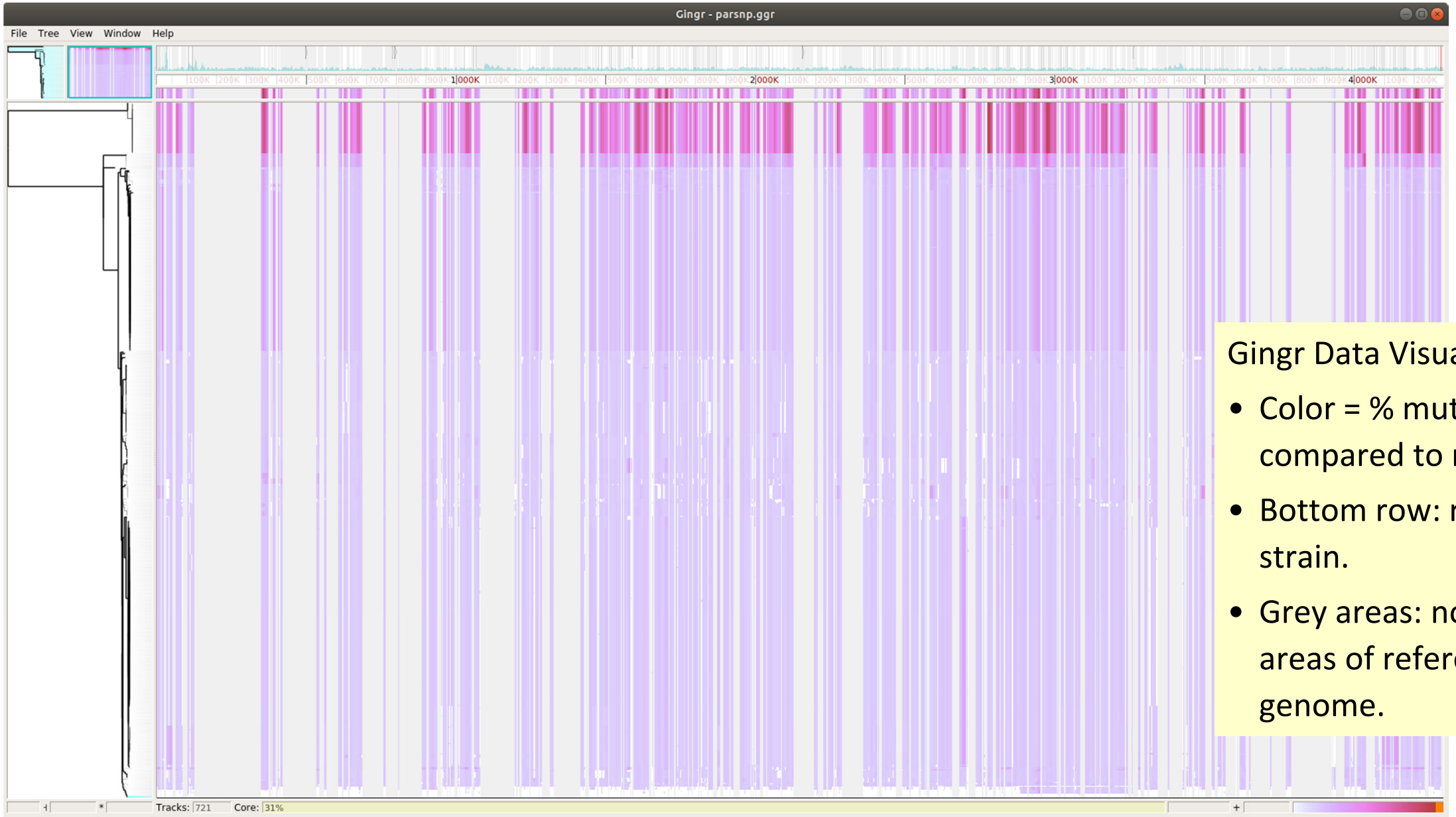
We can use Gingr to dig into areas of the genome where the two Clonal Groups differ from one another, for example these genes "hisP" and an un-annotated one adjacent to it.

Case Study #1: Synteny Comparison



Not a lot of genome rearrangement, but a small number of (consistent) large transpositions (possible assembly error?)

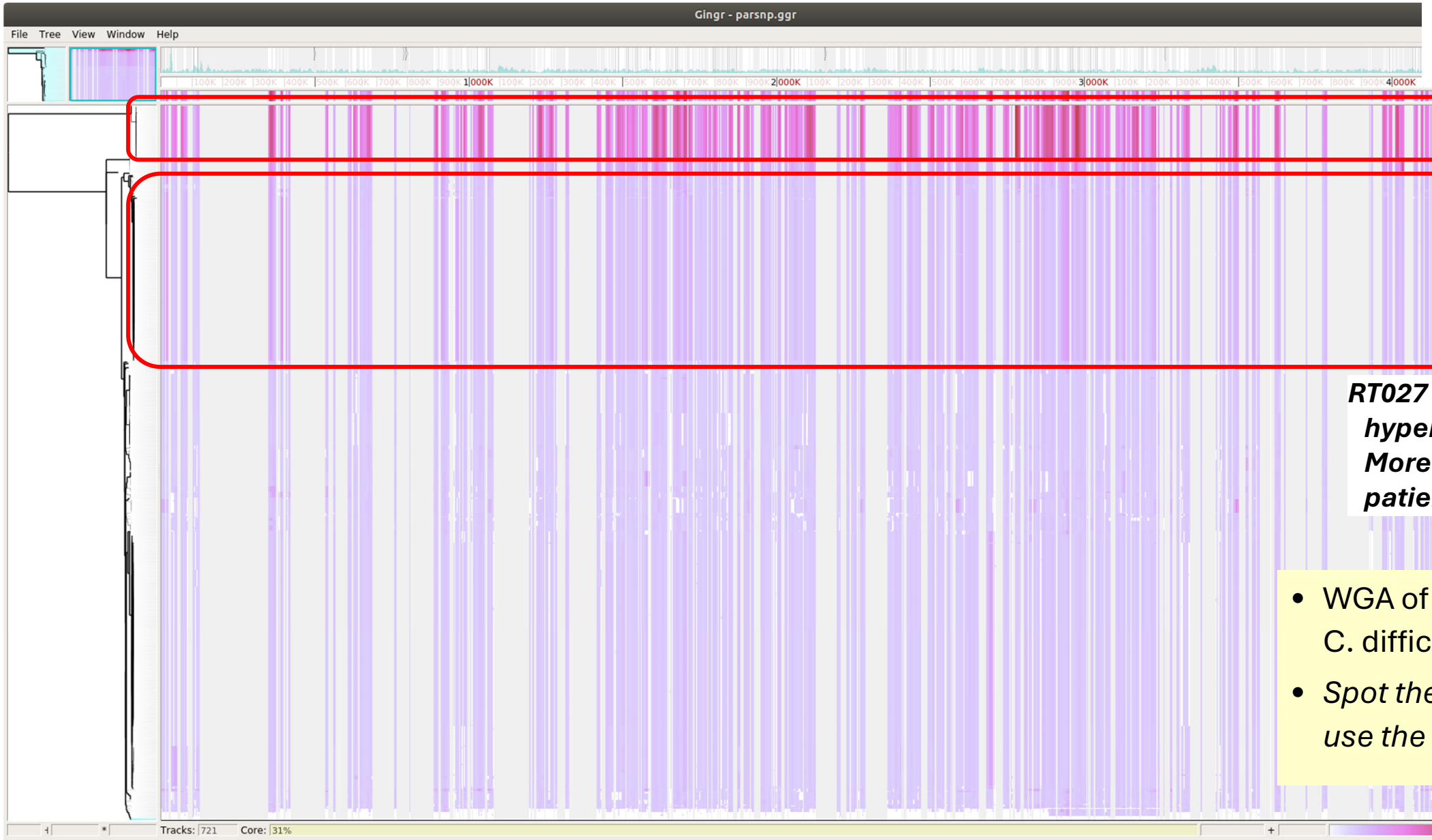
Case-Study #2: *C. difficile* Genomes



Gingr Data Visualization:

- Color = % mutation compared to reference
- Bottom row: reference strain.
- Grey areas: non “core” areas of reference genome.

Case-Study #2: *C. difficile* Genomes

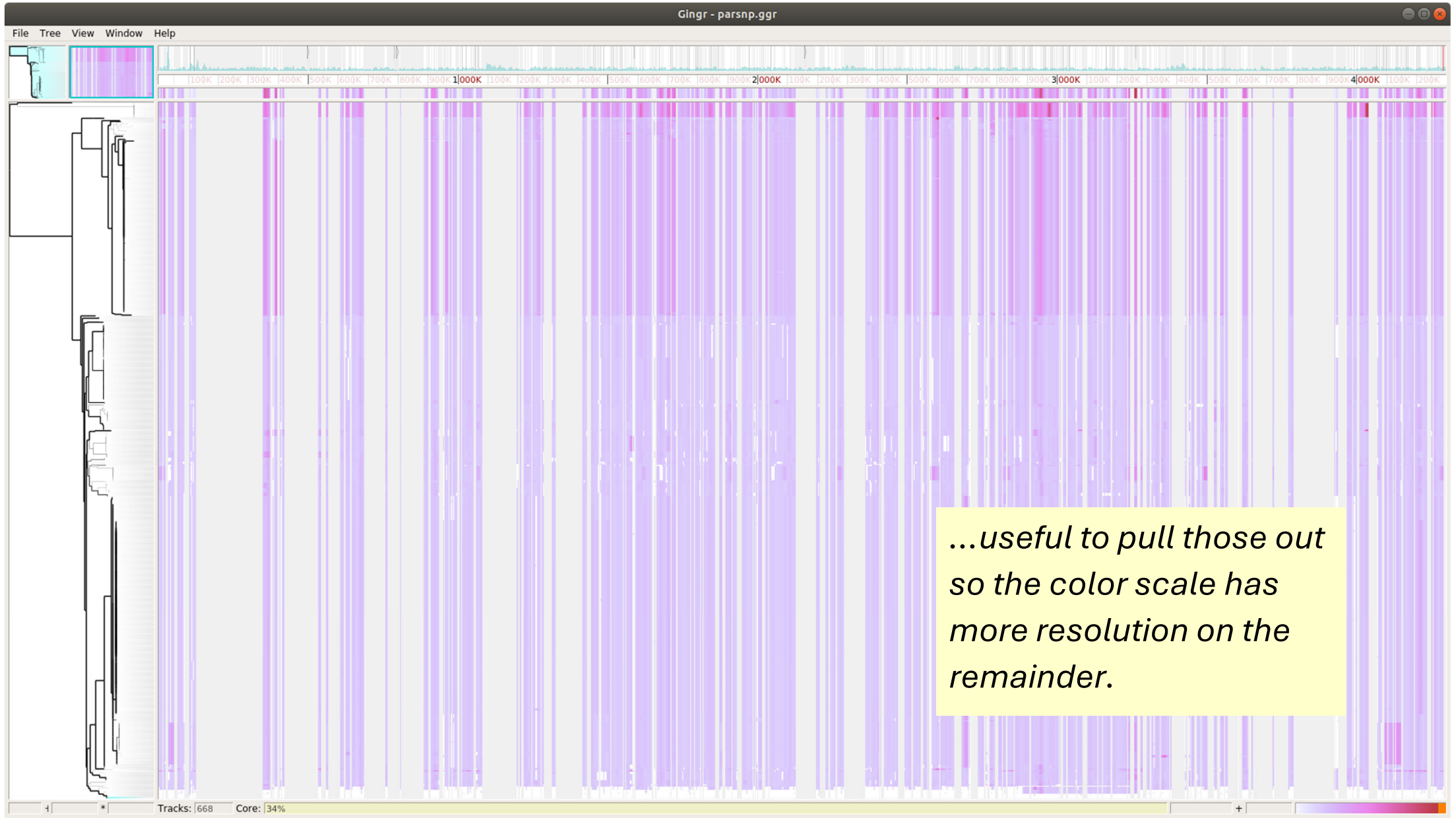


RT078 – Originated in animal host, crossed over

RT027 – Known hypervirulent strain. More recurrent, nastier patient outcomes.

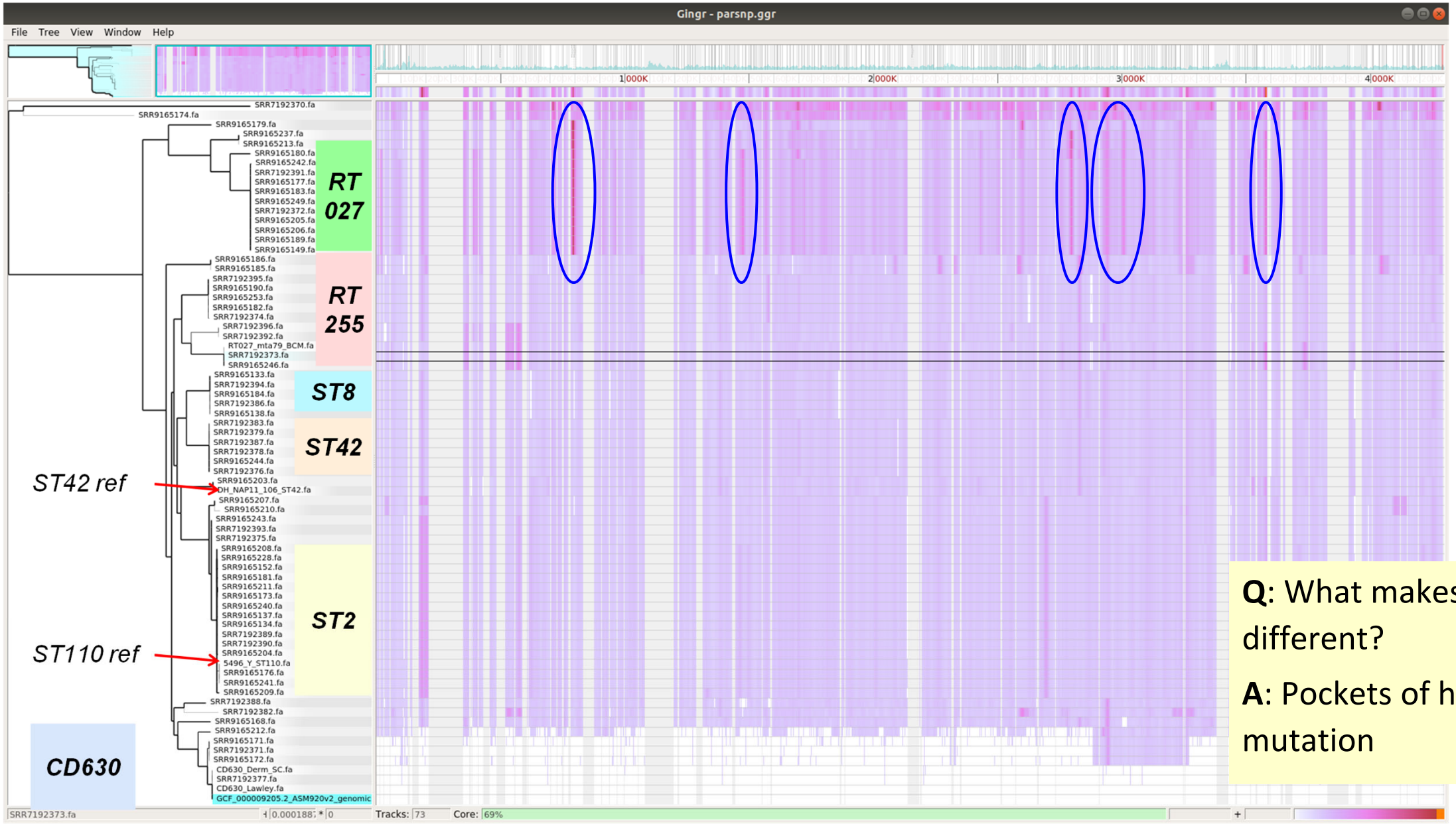
- WGA of 720 assembled C. difficile genomes
- *Spot the strains… (hint: use the tree)*

Case-Study #2: *C. difficile* Genomes (excluding RT078 samples)



Case-Study #2

Subset of Genomes w/ST annotation



Q: What makes RT027 different?

A: Pockets of heavy mutation

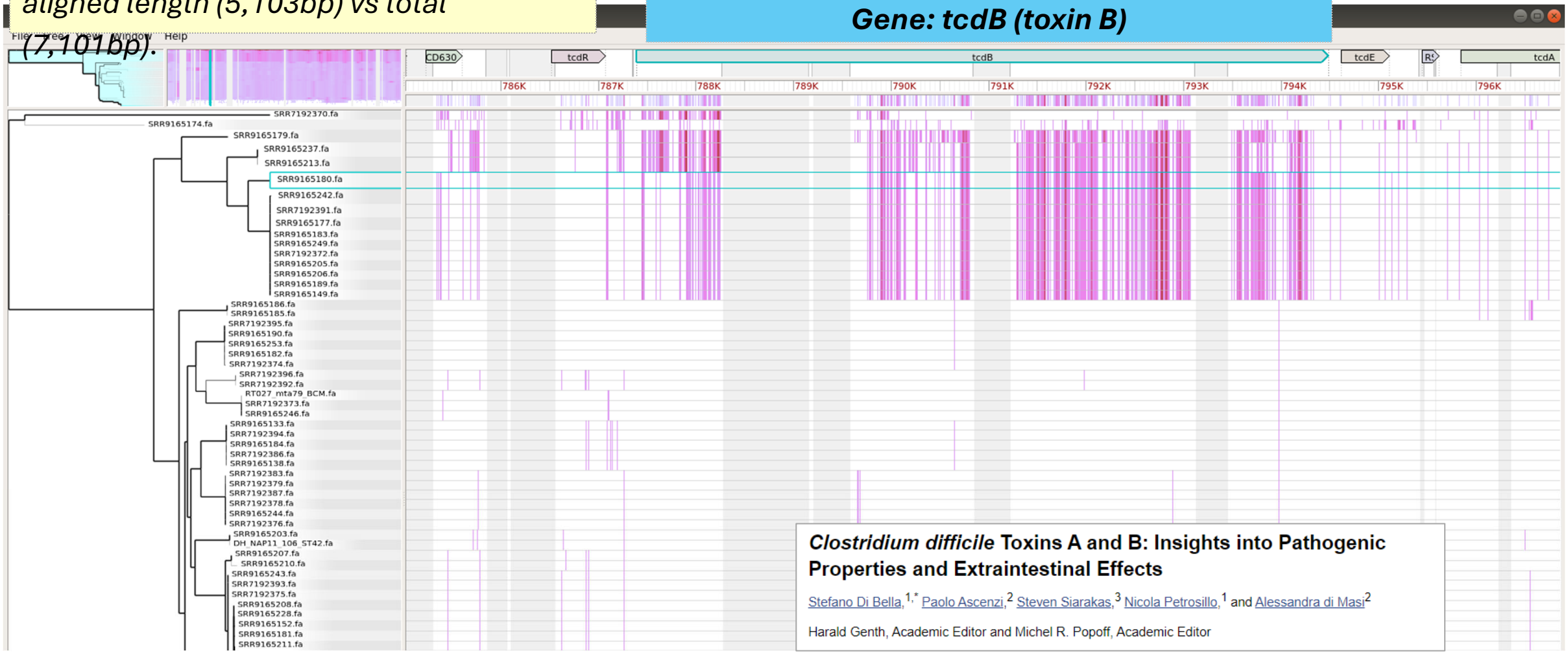
Case-Study #2

Digging Deeper…

Note: *not all of the tcdB gene was aligned by Parsnp, so this table represents the aligned length (5,103bp) vs total (7,101bp).*

- *This particular region is precisely the coding locus for Toxin B.*
- *RT027 carries a variant tcdB gene with altered function that contributes to its virulence.*

Gene: tcdB (toxin B)



***Clostridium difficile* Toxins A and B: Insights into Pathogenic Properties and Extraintestinal Effects**

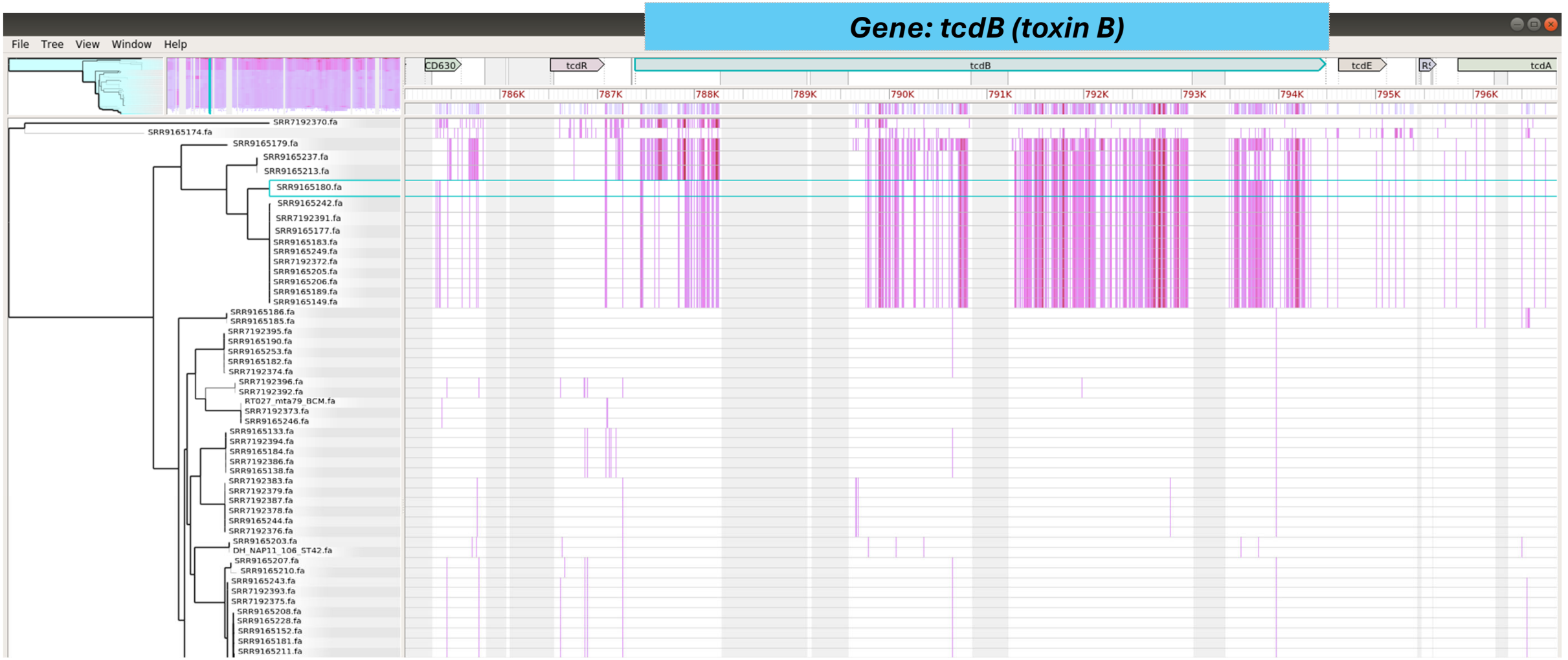
Stefano Di Bella,[1,*] Paolo Ascenzi,[2] Steven Siarakas,[3] Nicola Petrosillo,[1] and Alessandra di Masi[2]

Harald Genth, Academic Editor and Michel R. Popoff, Academic Editor

Case-Study #2

Remark...

Gene-level phylogenetic signal largely matches up with whole-genome phylogeny...

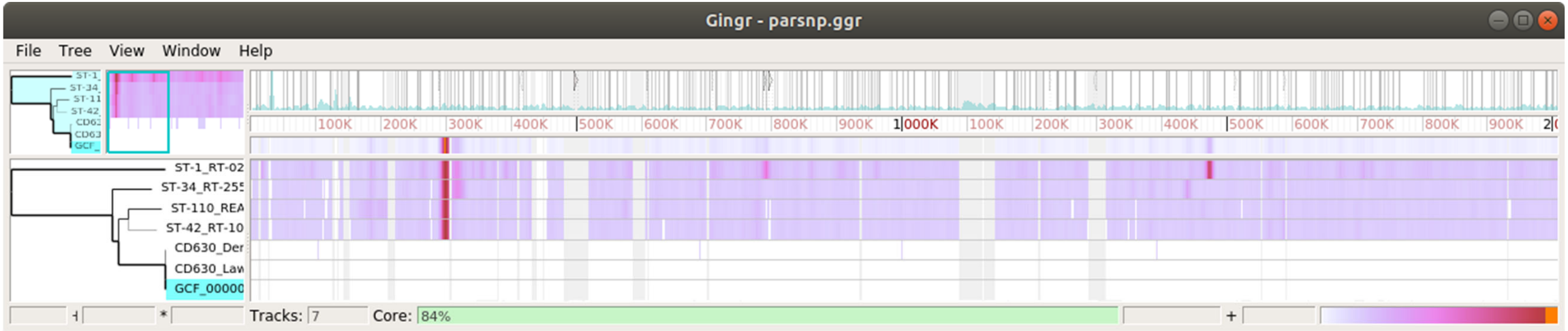


Case-Study #2

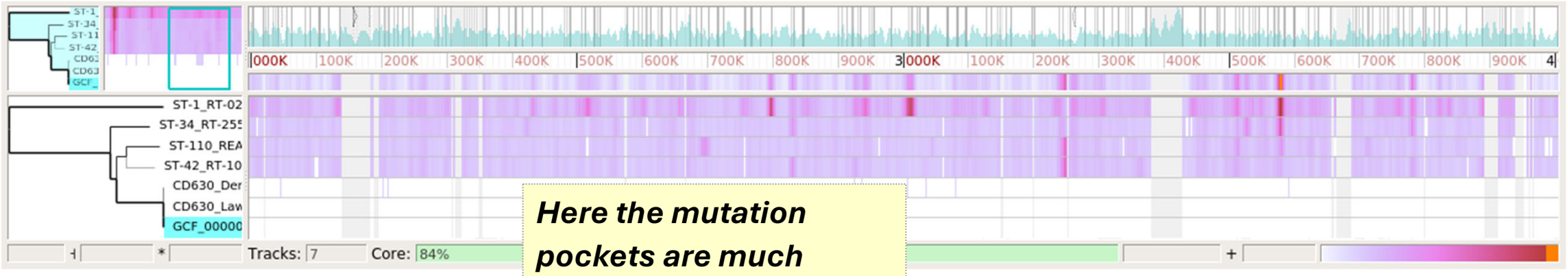
Comparing Reference Genomes for Some Strains

***Note**: RT027 is in the top row. CD630 is a lab strain used as a common reference.*

Segment 1 (positions 0-2mbp)



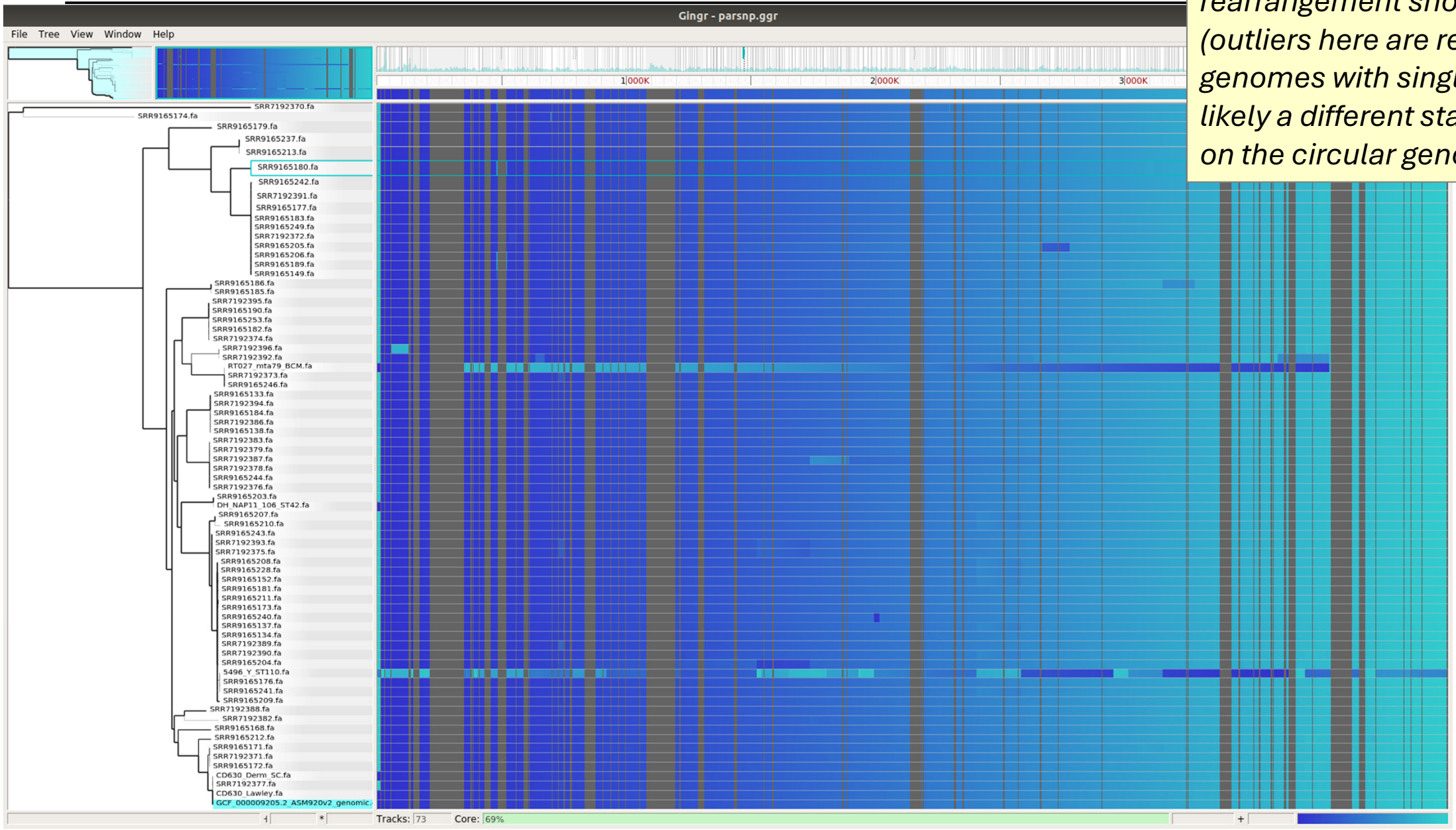
Segment 2 (positions 2-4mbp)



Here the mutation pockets are much clearer.

Case-Study #2

Synteny Comparison: *C. difficile* Isolates



For C. diff, even across a huge number of isolates, very little rearrangement shows up (outliers here are reference genomes with single contig, likely a different starting point on the circular genome.)

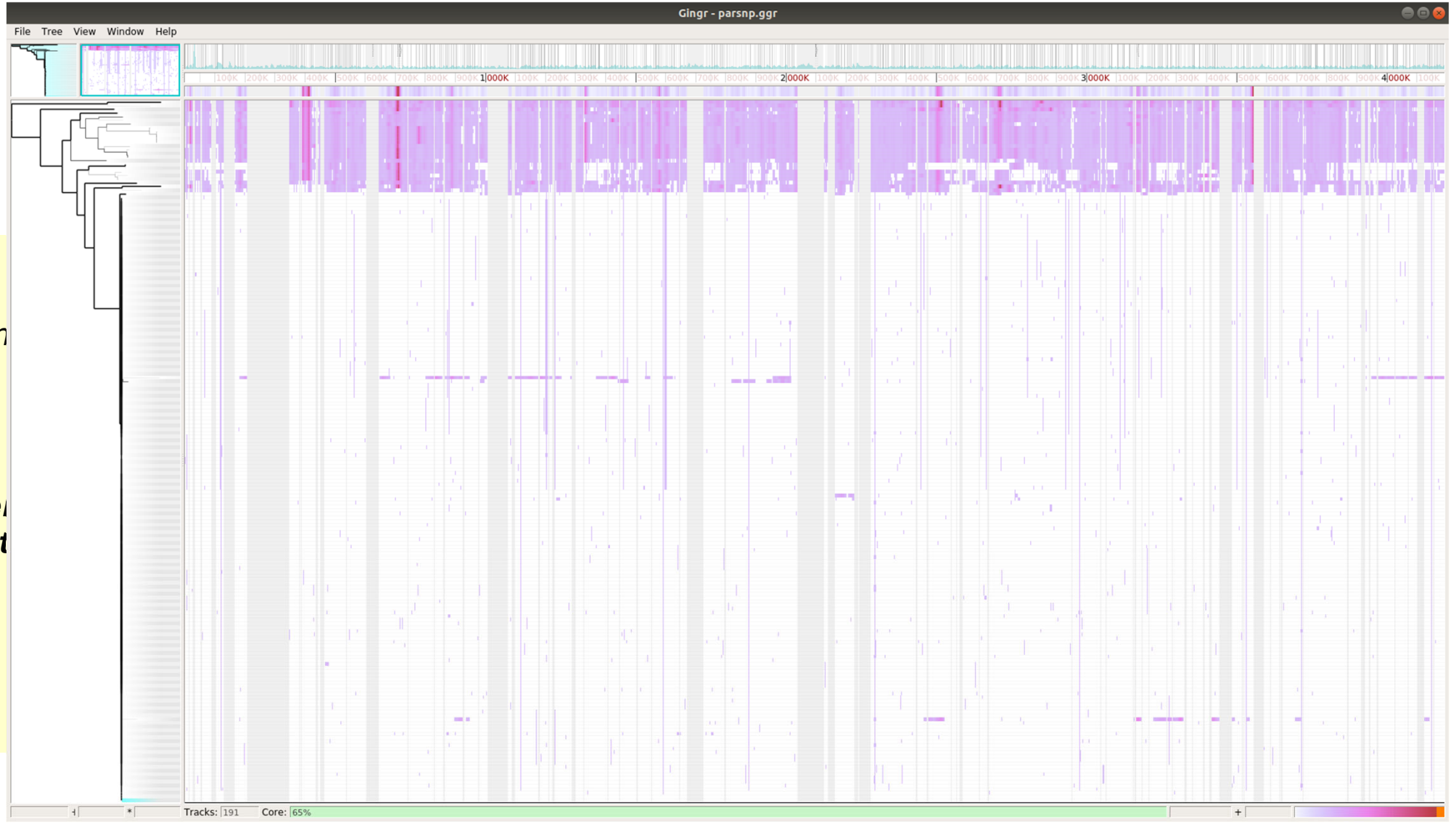
Case-Study #2

Alignment of RT027 isolates (and near relatives) to RT027 ref.

Q: *Does the RT027 Reference match the genomes from the clin*

A: *...Yes*

- ***Very little to see, ve high match level wit all RT027 isolates except 3.***



Case Study #3: *R. gnavus* Isolates from IBD Patients

14 Genomes:

- Reference: ATCC 29149 (RefSeq GCF_008121495)
- ATCC 29149 *de novo* assembly (by me)
- ATCC 35913 (GenBank GCA_900036035)
- 12 Genomes from Hall et al. (2017) (table at right)

RJX1118*	Stool from infant treated with antibiotics
RJX1119*	Stool from infant treated with antibiotics
RJX1120*	Biopsy from IBD patient
RJX1121*	Biopsy from IBD patient
RJX1122*	Biopsy from IBD patient
RJX1123*	Biopsy from IBD patient
RJX1124*	Biopsy from IBD patient
RJX1125*	Biopsy from IBD patient
RJX1126*	Biopsy from IBD patient
RJX1127*	Biopsy from IBD patient
RJX1128*	Stool from IBD patient

Case-Study #3

R. gnavus Isolates from IBD Patients

Game 1 : Spot the 2 Genomes from Infant Stool (non-IBD)

Case-Study #3

R. gnavus Isolates from IBD Patients

Game 1 : Spot the 2 Genomes from Infant Stool (non-IBD)

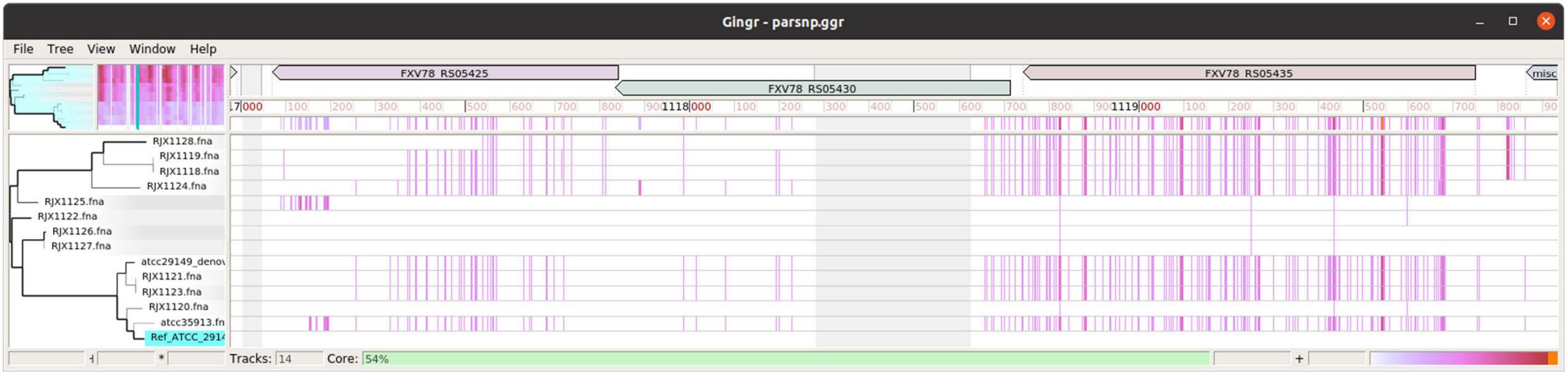
Game 2 : Spot the 2nd ATCC 29149 genome (supposedly the same as the reference)

Case-Study #3

R. gnavus strain-level phylogenetic signal is a mes

Depending on the operon, the phylogenetic appearance is completely different
⇒ ***Highly plastic genome***

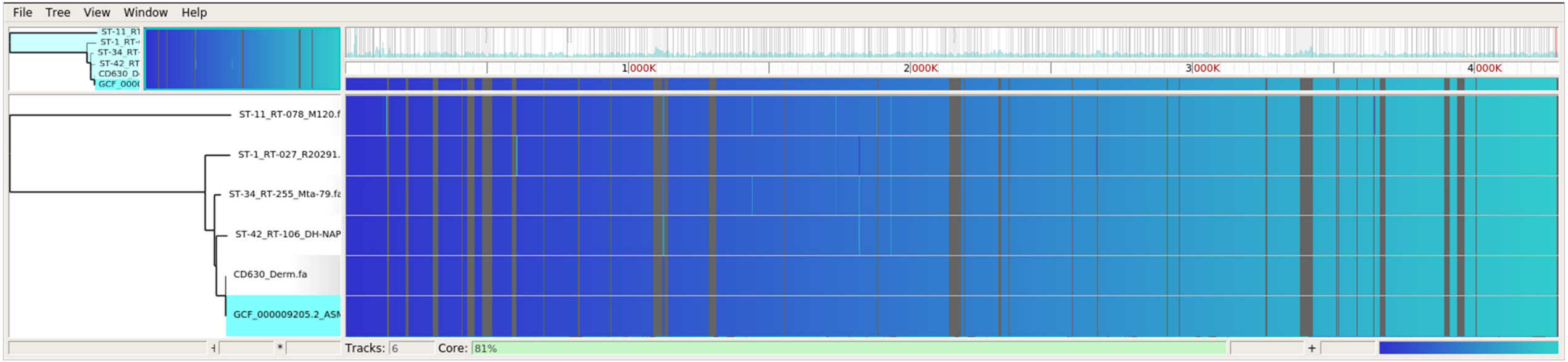
This does NOT show genome rearrangements (synteny)

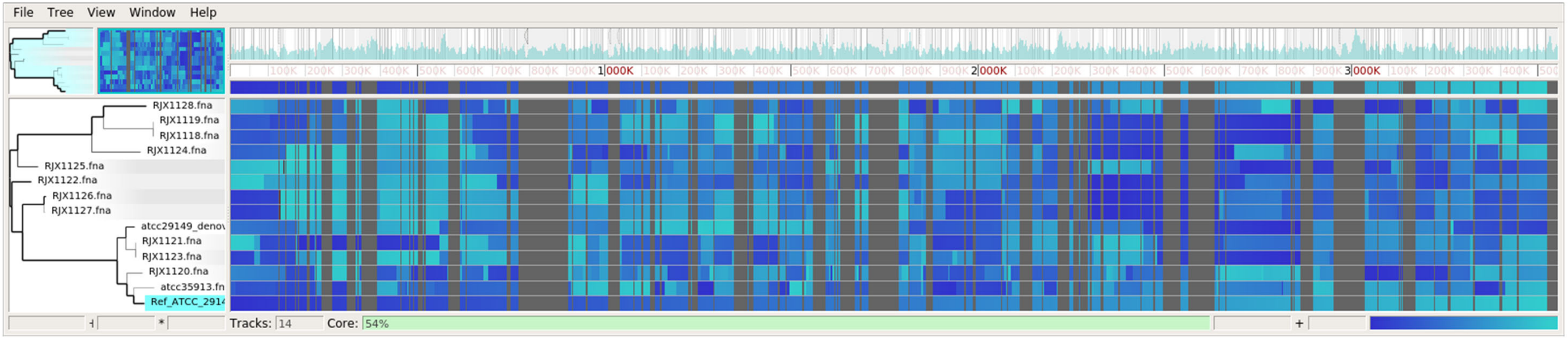


Case-Study #3

Synteny Comparison: *R. gnavus* & *C. difficile*

These two organisms have very different types of genome plasticity.





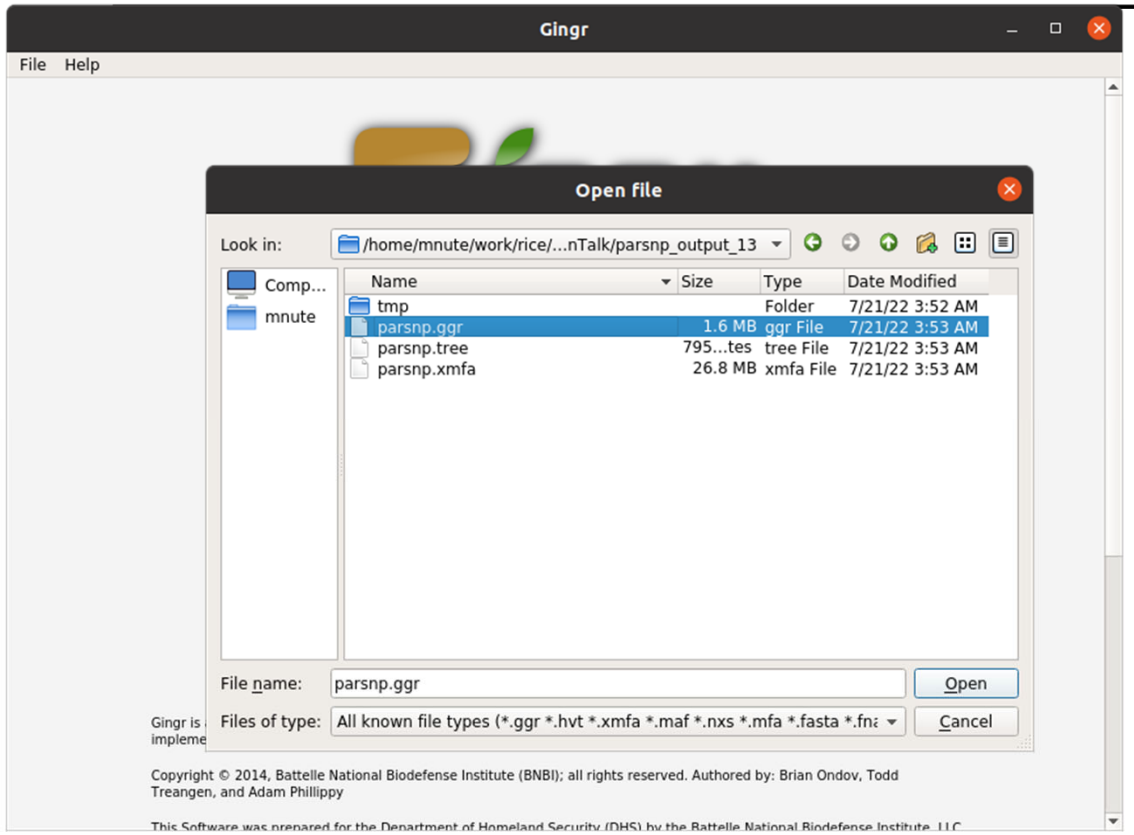
Conclusions

- Whole-genome alignment will give a detailed comparison specifically of the *core* genome
  - Maybe also auxiliary genes (*pan*-genome)
- Visualization can get you up close and personal with the data
  - (This statement applies to almost everything, not just genomes)
- Strains can differ from one another in weird ways.
  - Selective mutation at points of interest
  - Gene gain/loss depending on environment
  - Genome-wide phylogenetic signal vs. Locus-specific signal
  - Etc…?

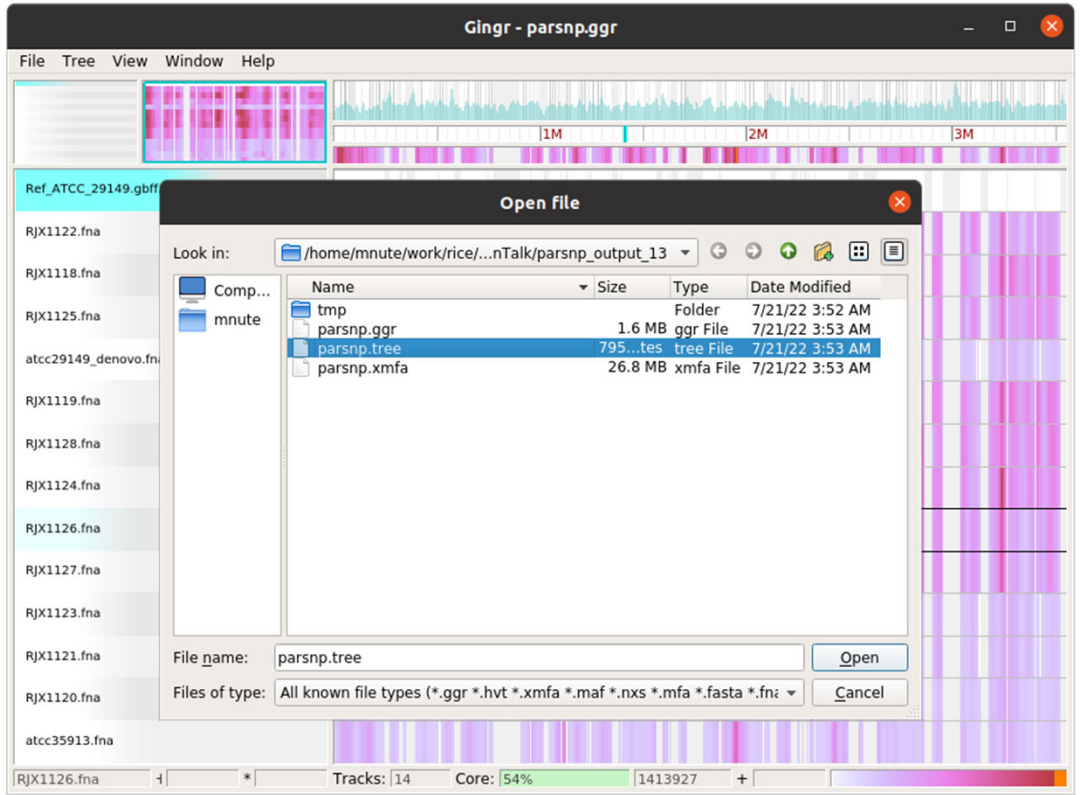
Special Thanks To:

- The Treangen Lab (Rice)
  - Todd Treangen
  - Bryce Killie
  - Kristen Curry
  - Nick Sapoval
  - Yunxi Liu
  - Yilei Fu
  - Advait Balaji
- The Savidge Lab (Baylor College of Medicine)
  - Qinglong Wu
  - Charlie Seto
- Taylor Reiter (for the *R. gnavus* idea)

Appendix: Quick How-to with Gingr (1 of 2)



*1.) Open the *.ggr file created in the parsnp output folder.*



*2.) Once it is open, go back to the "Open" dialogue and open the *.tree file in the same folder.*

Appendix: Quick How-to with Gingr (2 of 2)

3.) This will give you the standard Gingr view. Other options to re-root the tree or to switch to Synteny view are available under the "Tree" and "View" menus.