

# Strain-Level Comparative Genomics

Phylogenomics Crash Course

Multiple Genome Alignment & Visualization

(Bonus content: full-length 16S taxon abundance)

---

STAMPS – Day 7

July 25, 2024

Michael Nute & Todd Treangen

[mike.nute@gmail.com](mailto:mike.nute@gmail.com)

[mn56@rice.edu](mailto:mn56@rice.edu)

# Agenda...

---

9:00 to 9:15	(15 min.)	Intro/kickoff (Todd)
9:15 to 9:25	(10 min.)	Advertisement* (3 slides)
9:25 to 9:50	(25 min.)	Binning' (17 slides)
9:50 to 10:10	(20 min.)	Birds of a feather (using contigs from monday)
10:10 to 10:35	(25 min.)	Group Photo (by Lillie?)
10:35 to 10:45	(10 min.)	Phylogenetics + MSA lecture (Mike)
10:45 to 11:25	(40 min.)	Parsnp/strain analysis lecture (Mike)
11:25 to 11:45	(15 min.)	Parsnp tutorial (Todd)
11:45 to 12:00	(15 min.)	Parsnp tutorial discussion (Todd)

**OLD**

# Quick Phylogenetics Refresher

---

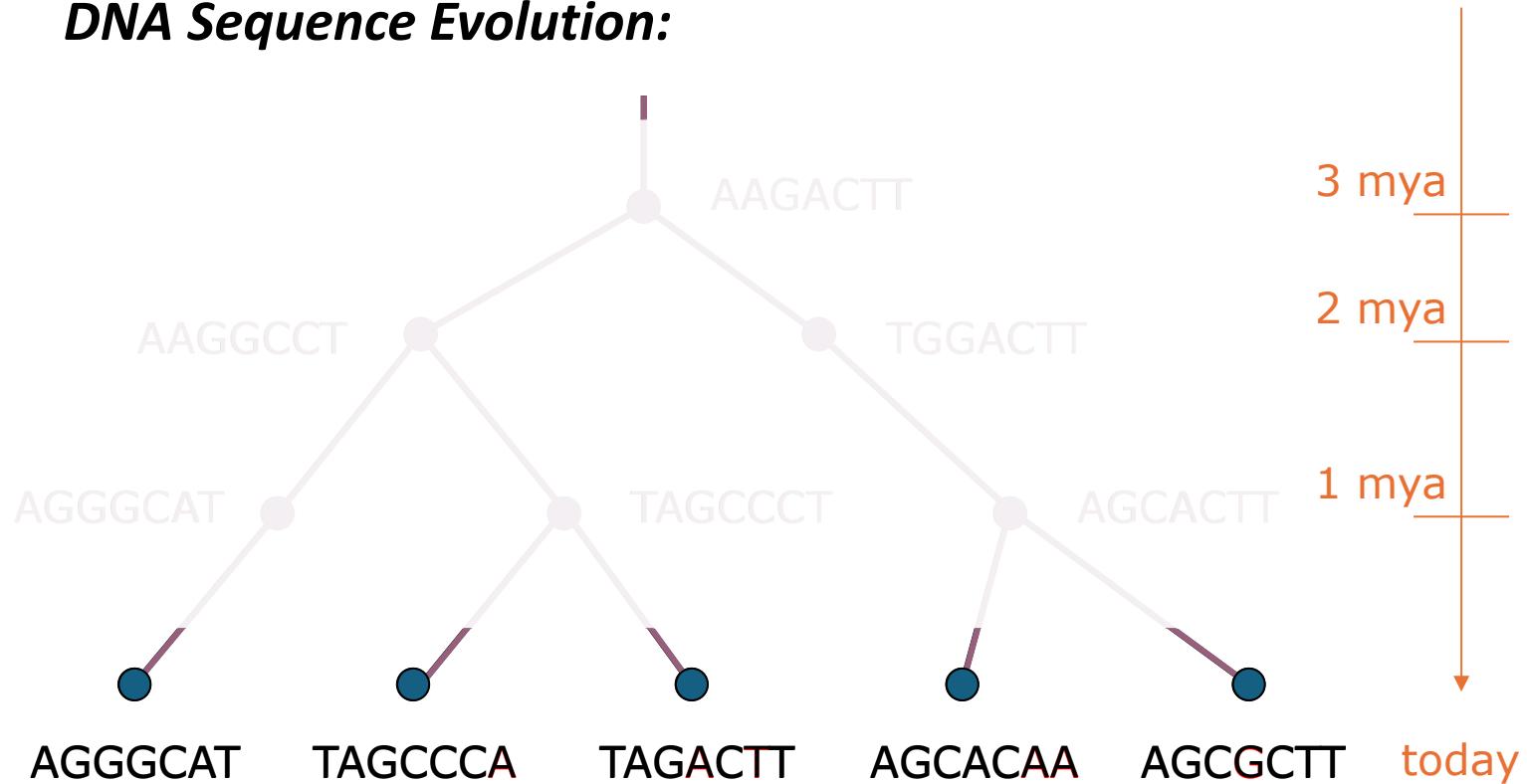
Multiple Sequence Alignment

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

# Multiple Sequence Alignment: Definition & Goal

---

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

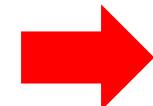
**Goal:** Align sequences so that all sets of positions having a common ancestor are grouped together.

- *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:

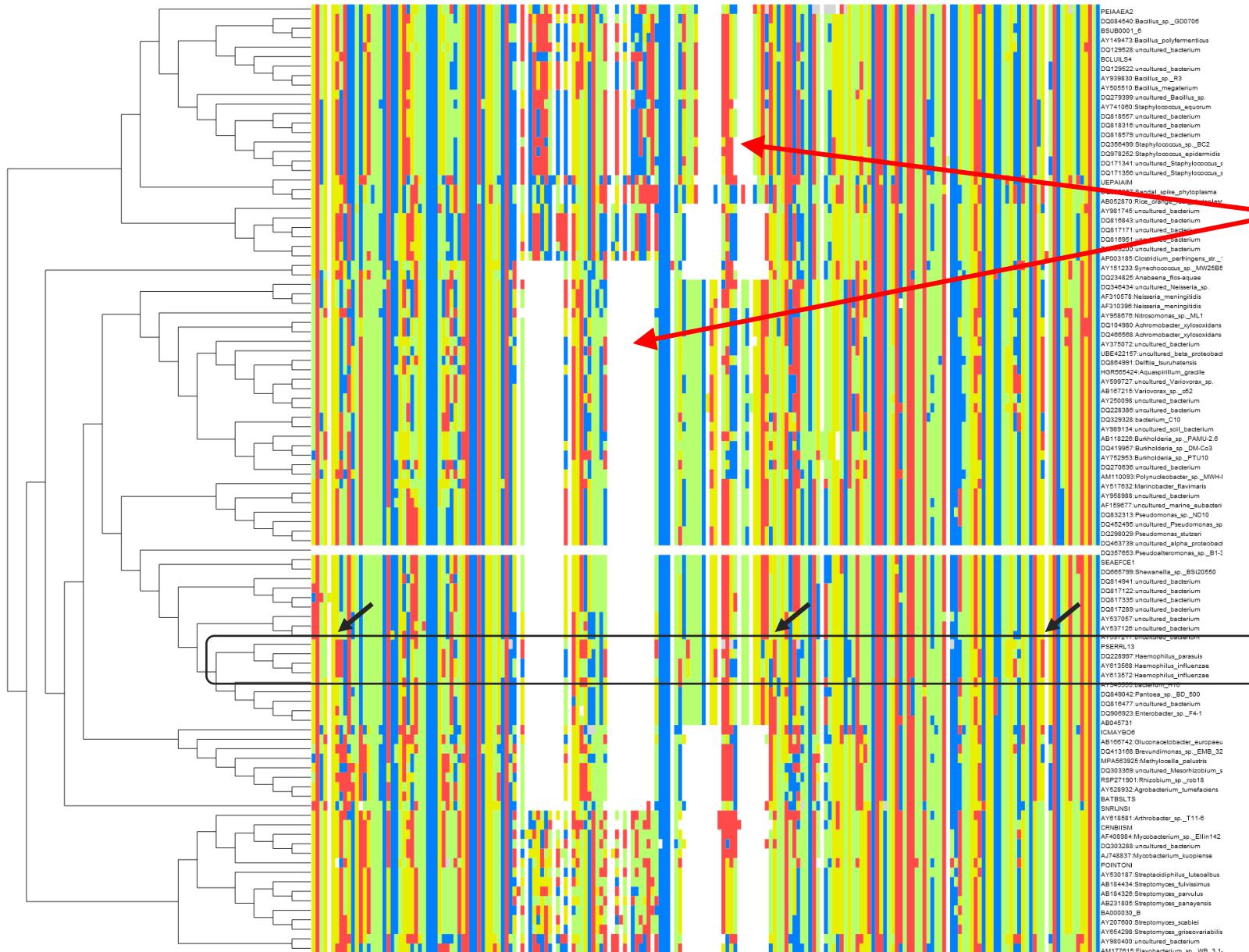
- **MAFFT**
  - Muscle
  - **PASTA**
  - ClustalW
  - DiAlign
  - BAli-Phy
- PRANK
  - T-COFFEE
  - ...et cetera

**Garbage  
Alignment**



**Garbage  
Tree**

# 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...*

# Data Properties Affecting Multiple Sequence Alignment

---

Very Approximate Order of Importance

- 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 $N$ )	<ul style="list-style-type: none"><li>• Uses patterns of insertion/deletion to find optimal alignment</li><li>• Generally pretty accurate in most conditions.</li></ul>
MUSCLE (Edgar, 2004)		<ul style="list-style-type: none"><li>• Progressive alignment. Suitable for relatively high overall sequence similarity.</li></ul>
CLUSTAL (Sievers et al., 2011)		<ul style="list-style-type: none"><li>• Ideal for protein alignments with structurally important sites.</li></ul>
PASTA (Mirarab et al., 2015)	Single Gene MSA (large $N$ )	<ul style="list-style-type: none"><li>• Divide-and-conquer algorithm. Ideal for scaling alignment to large number of sequences (<math>&gt;1000</math>)</li></ul>
HMMER (Eddy, 1998)	Query sequence alignment to reference	<ul style="list-style-type: none"><li>• Represents reference alignment as HMM. Query sequence alignment performed using standard HMM algorithms.</li></ul>

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  $p \rightarrow 1$
    - 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...)

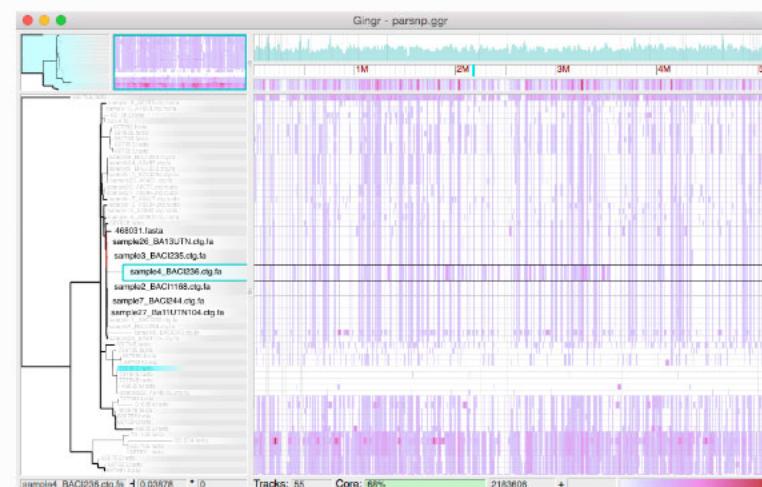
[Docs](#) » Harvest

[Edit on GitHub](#)

## Harvest



Harvest is a suite of core-genome alignment and visualization tools for quickly analyzing thousands of intraspecific microbial genomes, including variant calls, recombination detection, and phylogenetic trees.



# 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  
paths)

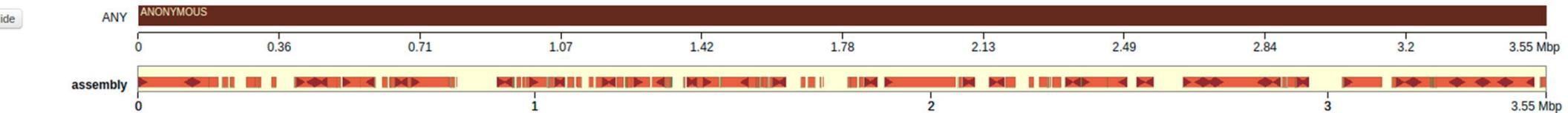
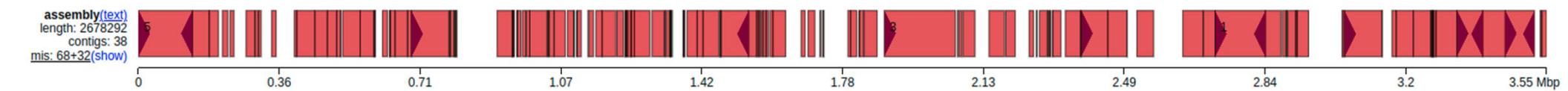
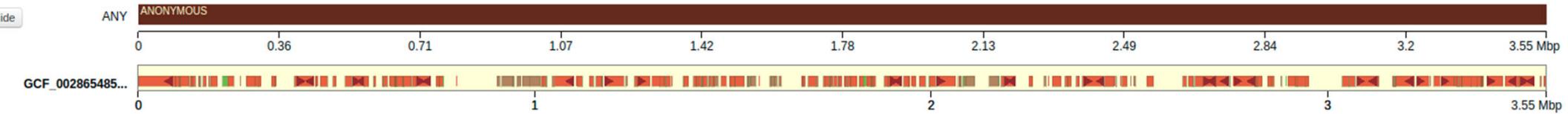
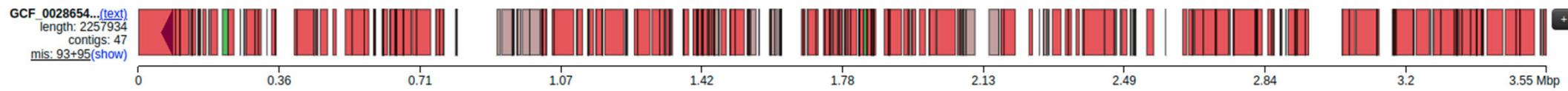
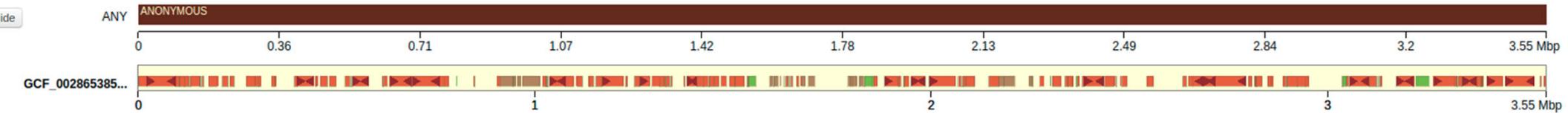
# processors

Output folder

## What can we learn?

- Assembly Quality Issues?
- Issues with Reference?

# Interlude: QC-ing an Assembly with QUAST



### 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*<sup>1</sup>

---

- 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”

# Case Study #1: Observations...

## Observation #1:

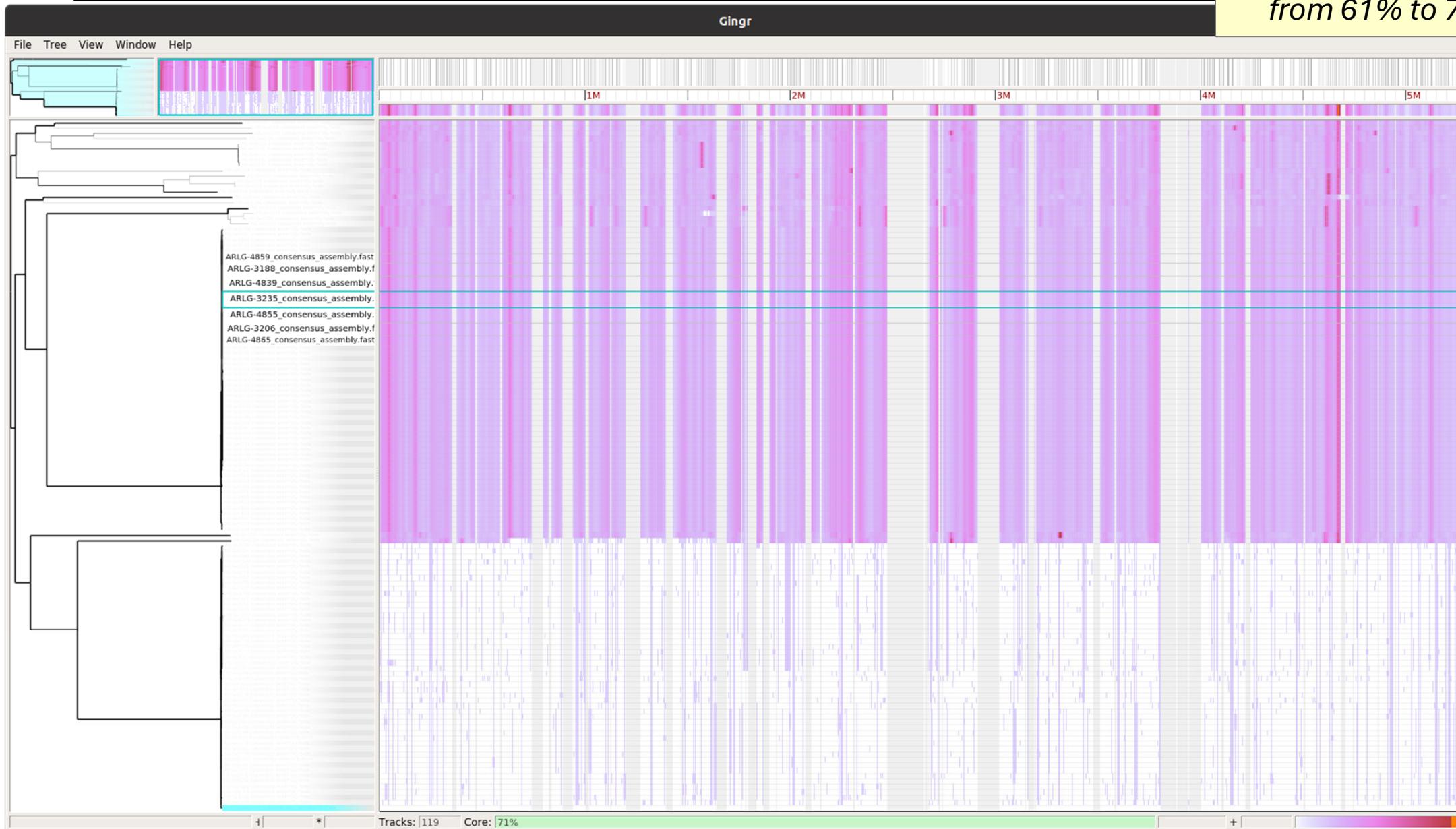
One of these genomes (ARLG-8054) is **MUCH** different from all the others.



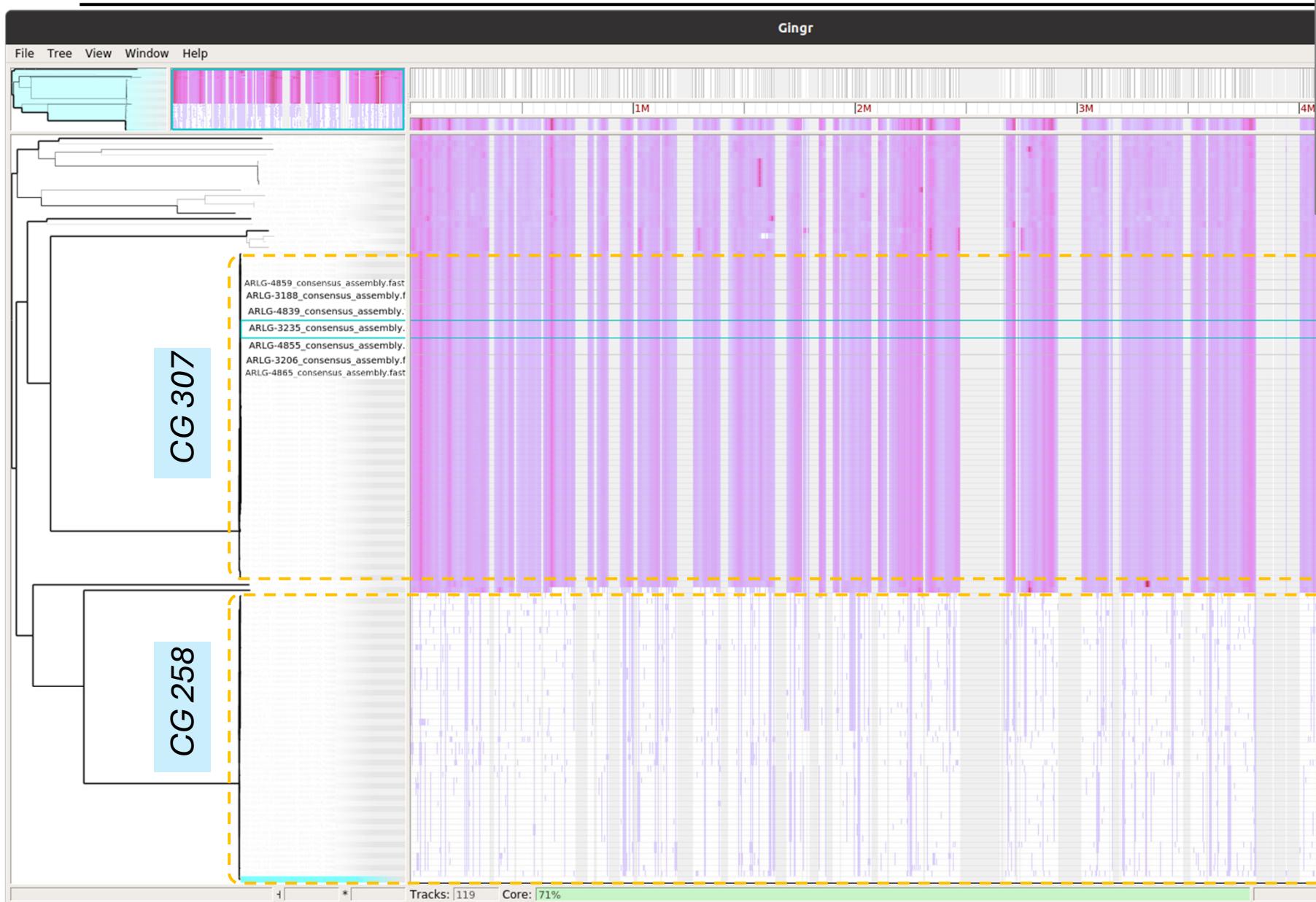
# Case Study #1: Excluding ARLG-8054

## Observation #2:

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



# Case Study #1: Excluding ARLG-8054



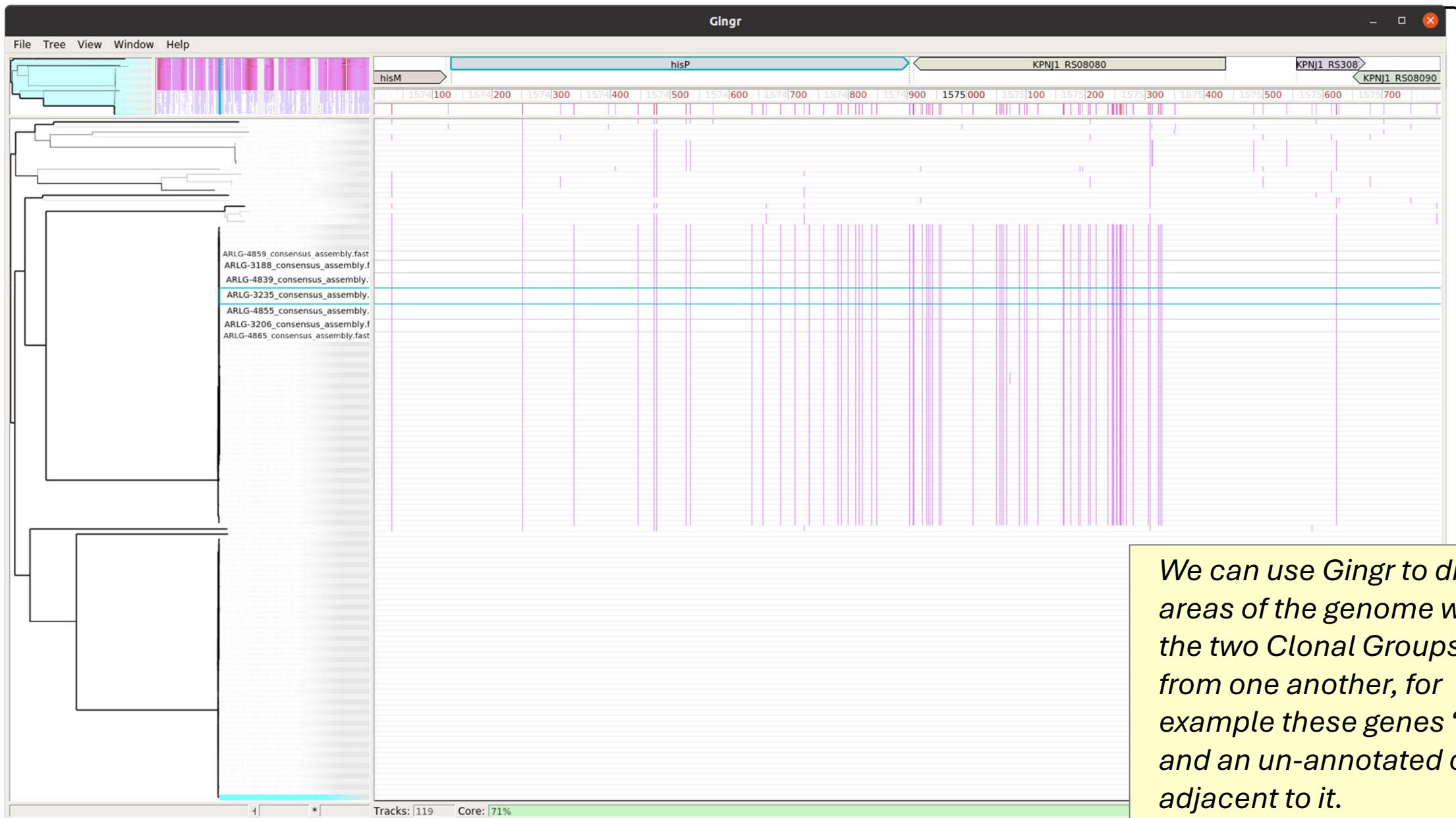
## Observation #3:

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

## Notes:

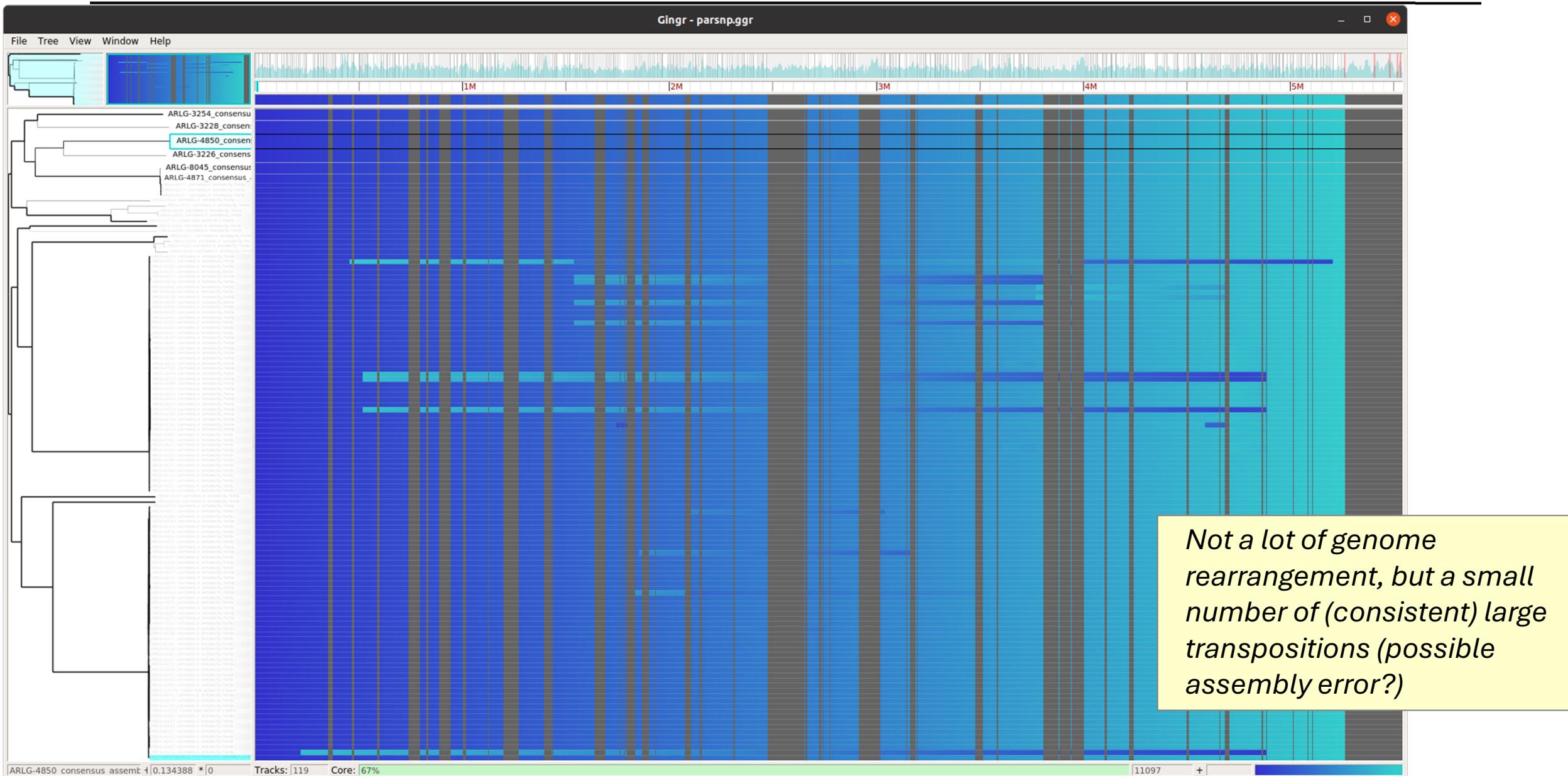
The reference genome here is GCF\_000598005.1, described as "strain=30660/NJST258\_1", so ostensibly ST258.

# Case Study #1: Digging In...

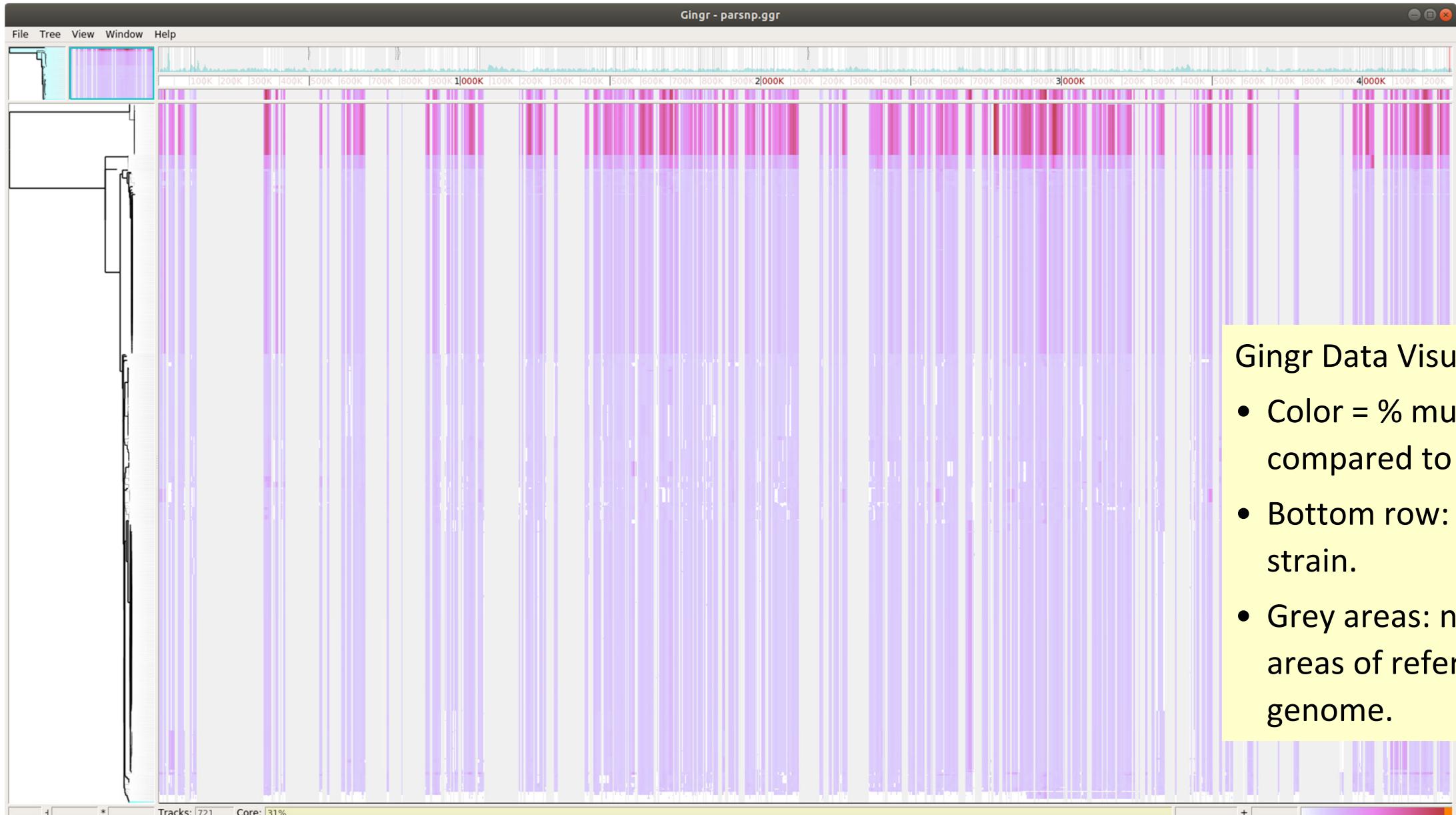


We can use Gingr to dig into areas of the genome where the two Clonal Groups differ from one another, for example these genes “hisP” and an un-annotated one adjacent to it.

# Case Study #1: Synteny Comparison



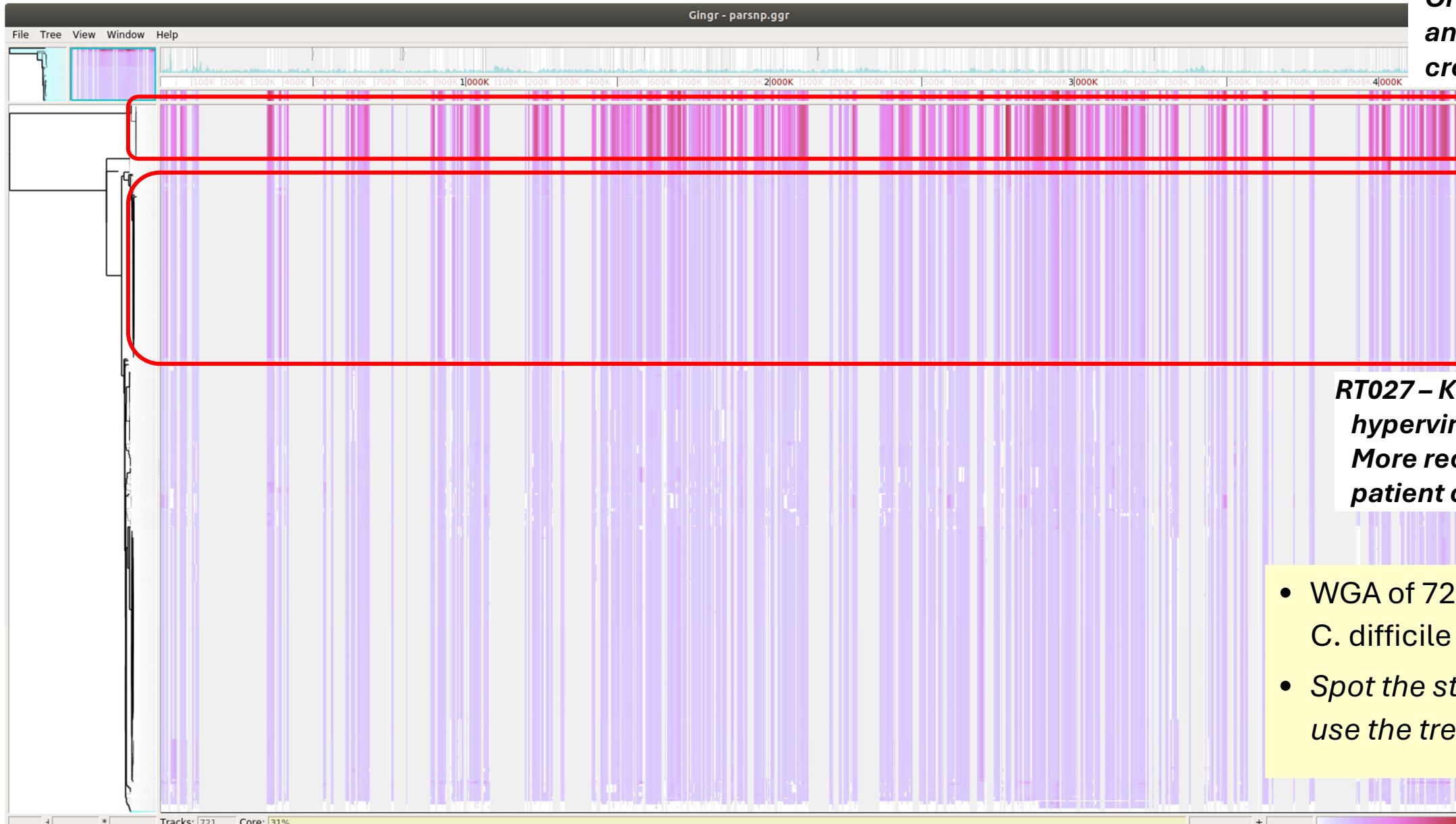
# Case-Study #2: *C. difficile* Genomes



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

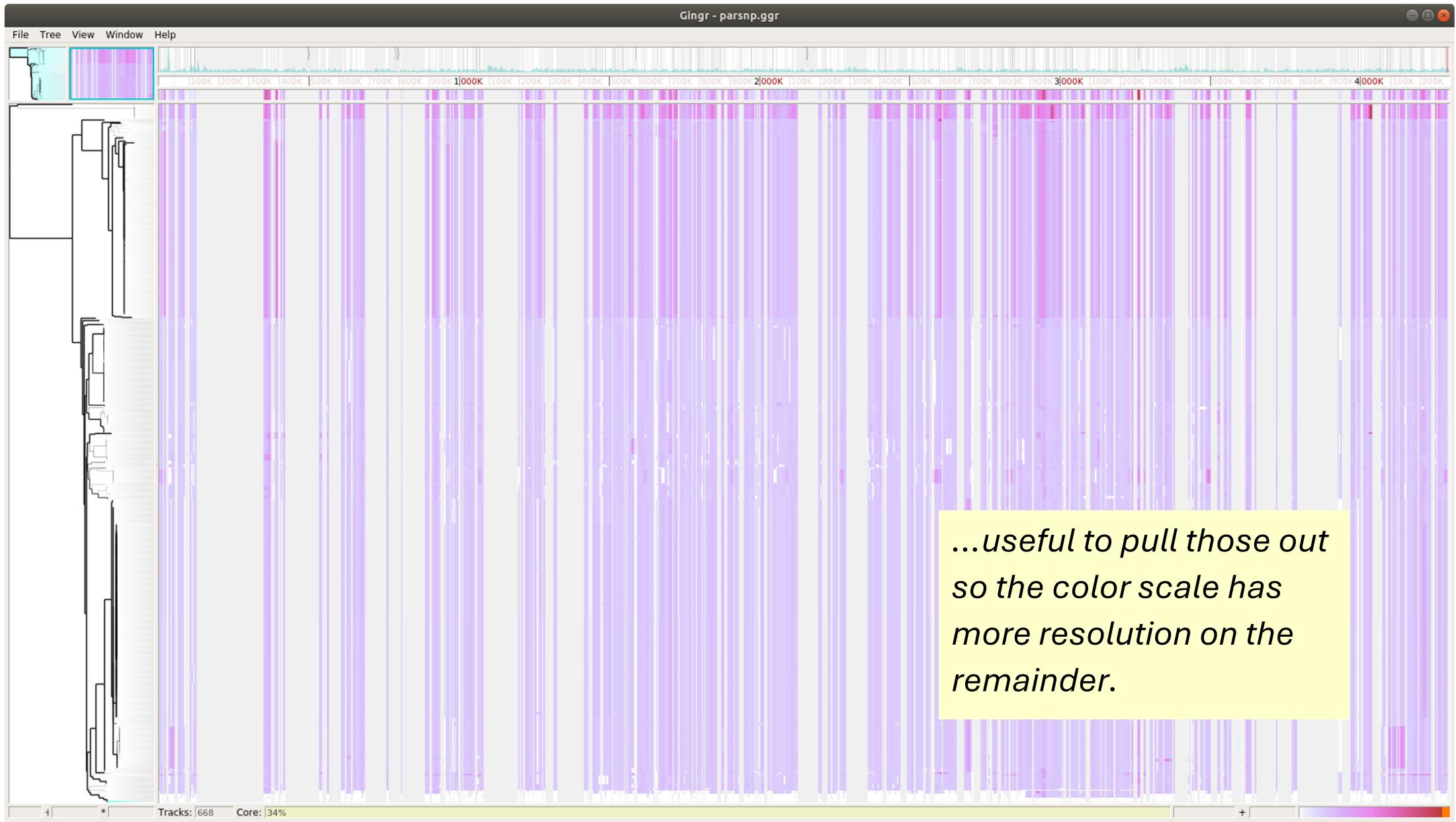


- WGA of 720 assembled *C. difficile* genomes
- Spot the strains... (hint: use the tree)

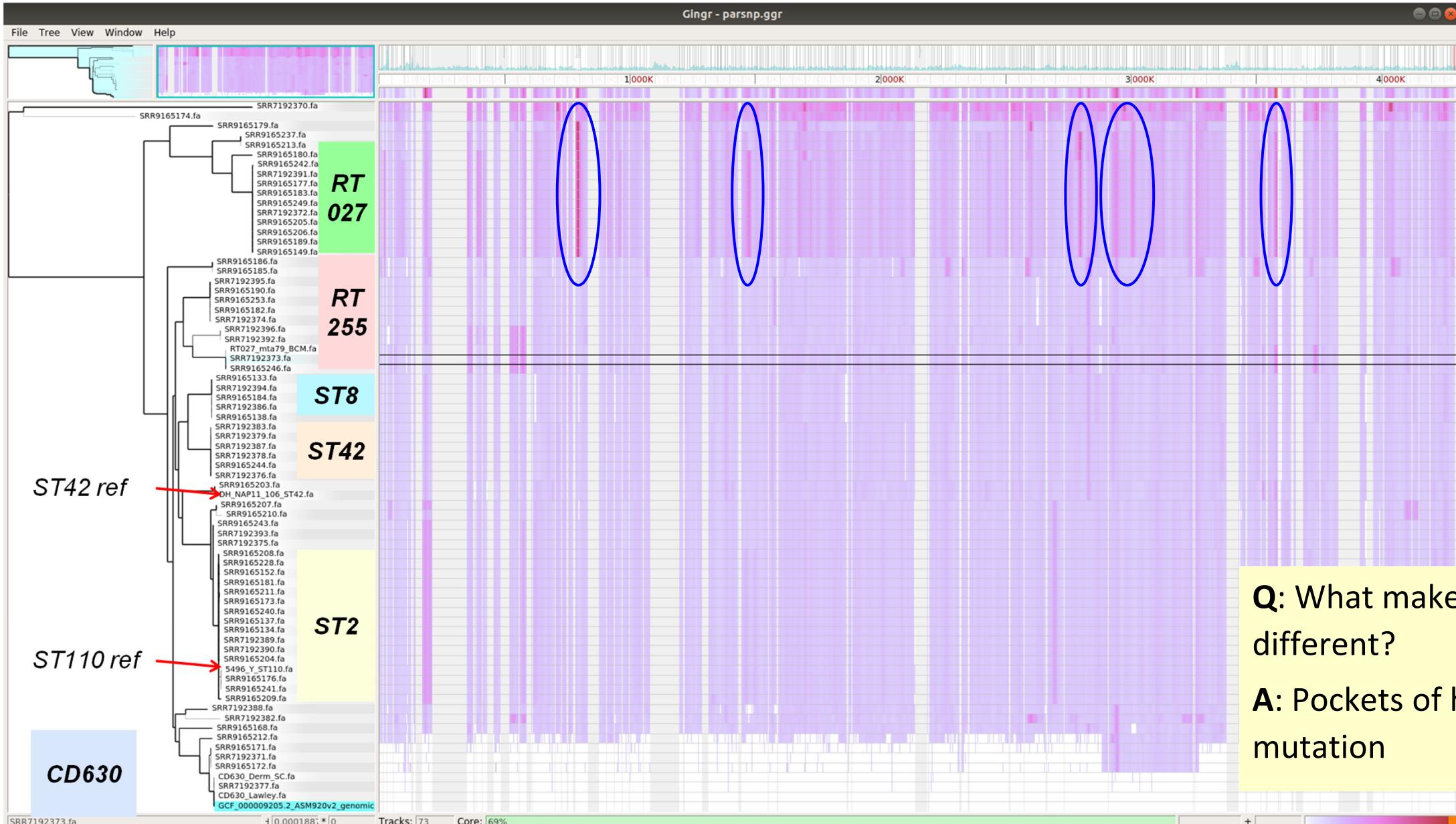
**RT078 –**  
*Originated in animal host, crossed over*

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

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



# Subset of Genomes w/ST annotation



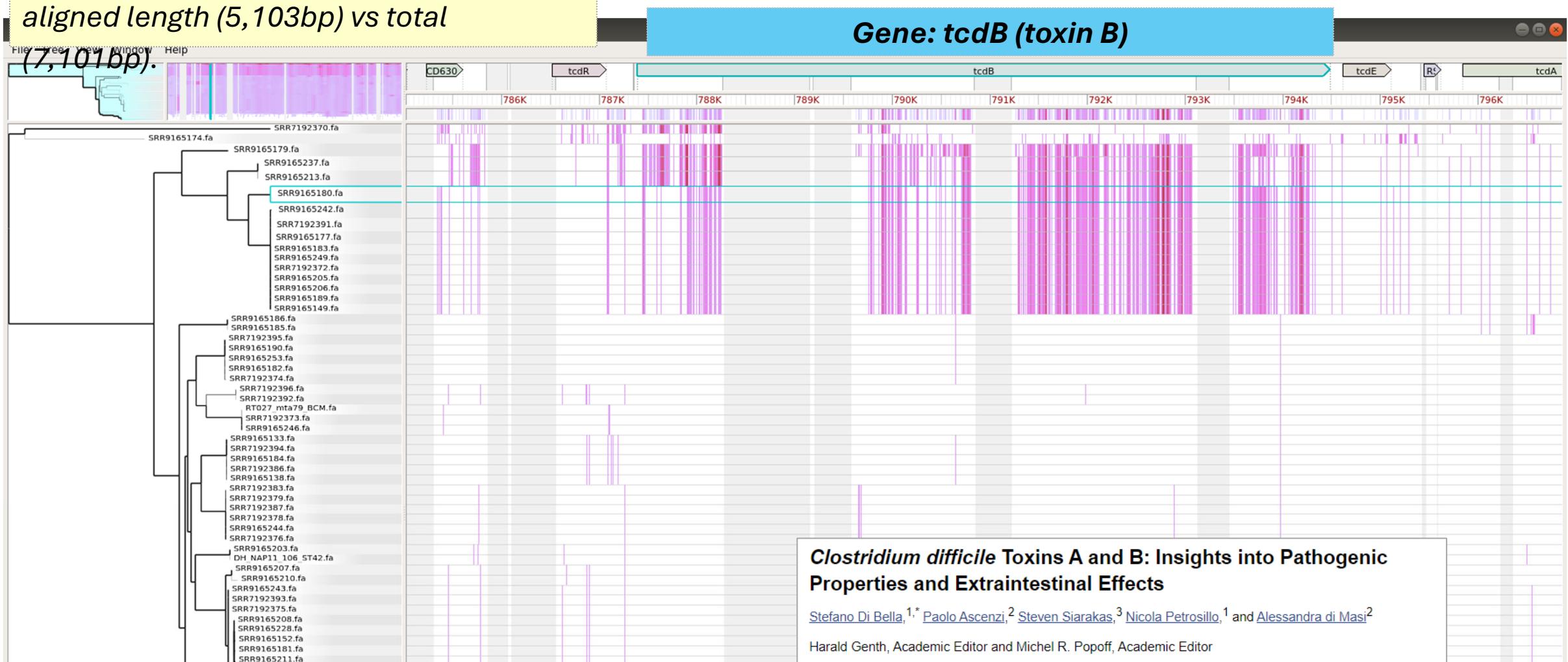
**Q:** What makes RT027 different?

**A:** Pockets of heavy mutation

# Digging Deeper...

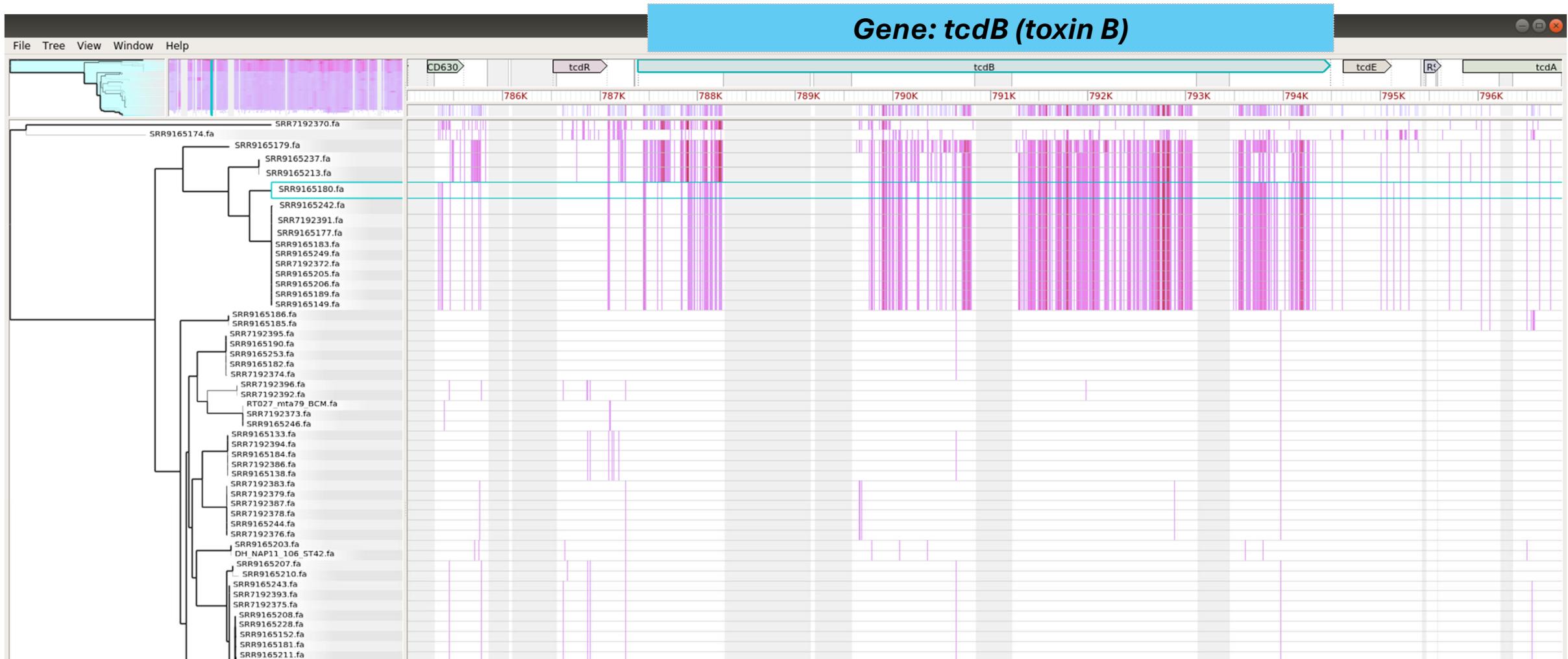
**Note:** not all of the *tcdB* gene was aligned by Parsnp, so this table represents the aligned length (5,103bp) vs total (7,101bp).

- 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.



# Remark...

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

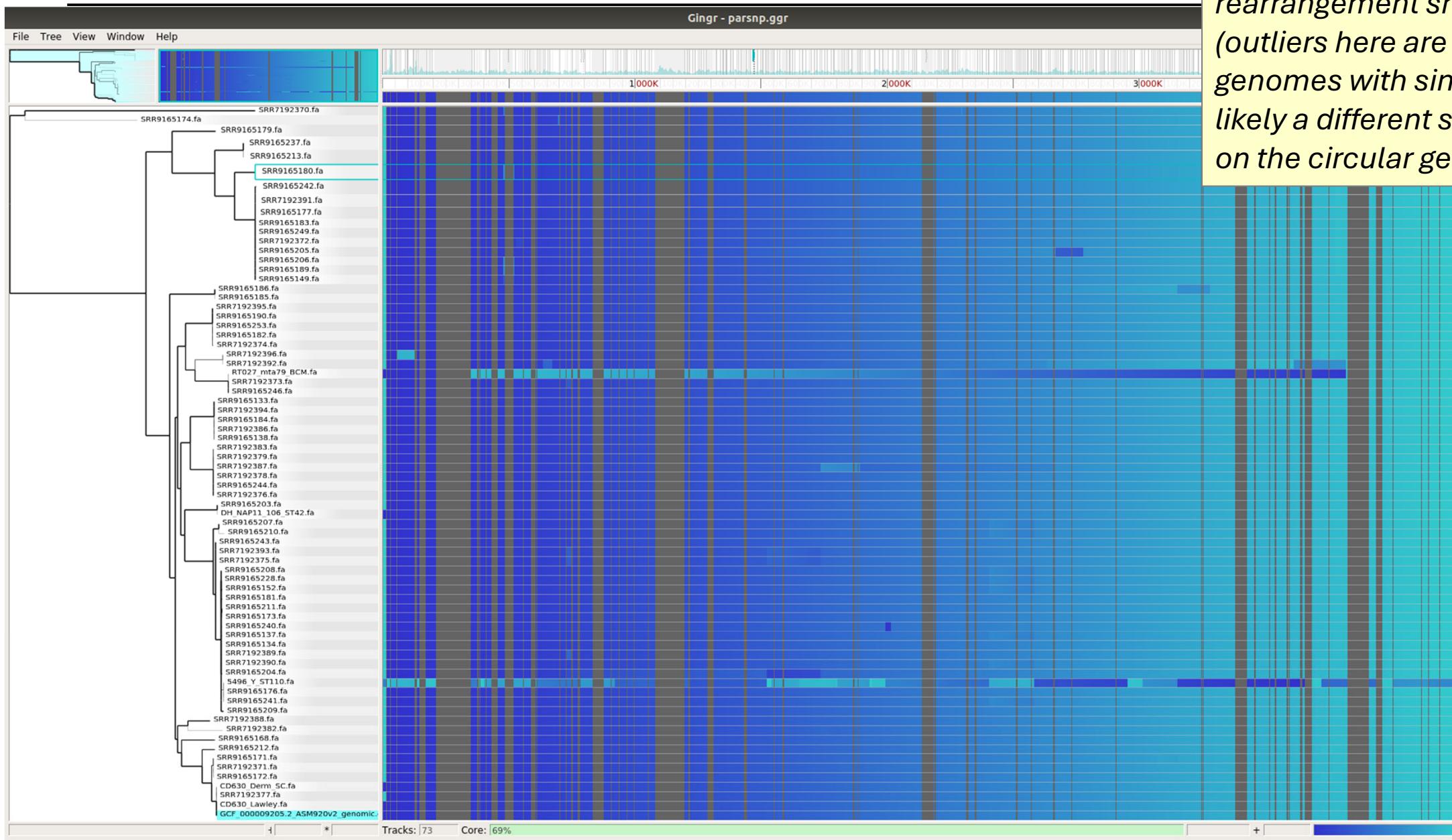


# Comparing Reference Genomes for Some Strains

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



# Synteny Comparison: *C. difficile* Isolates



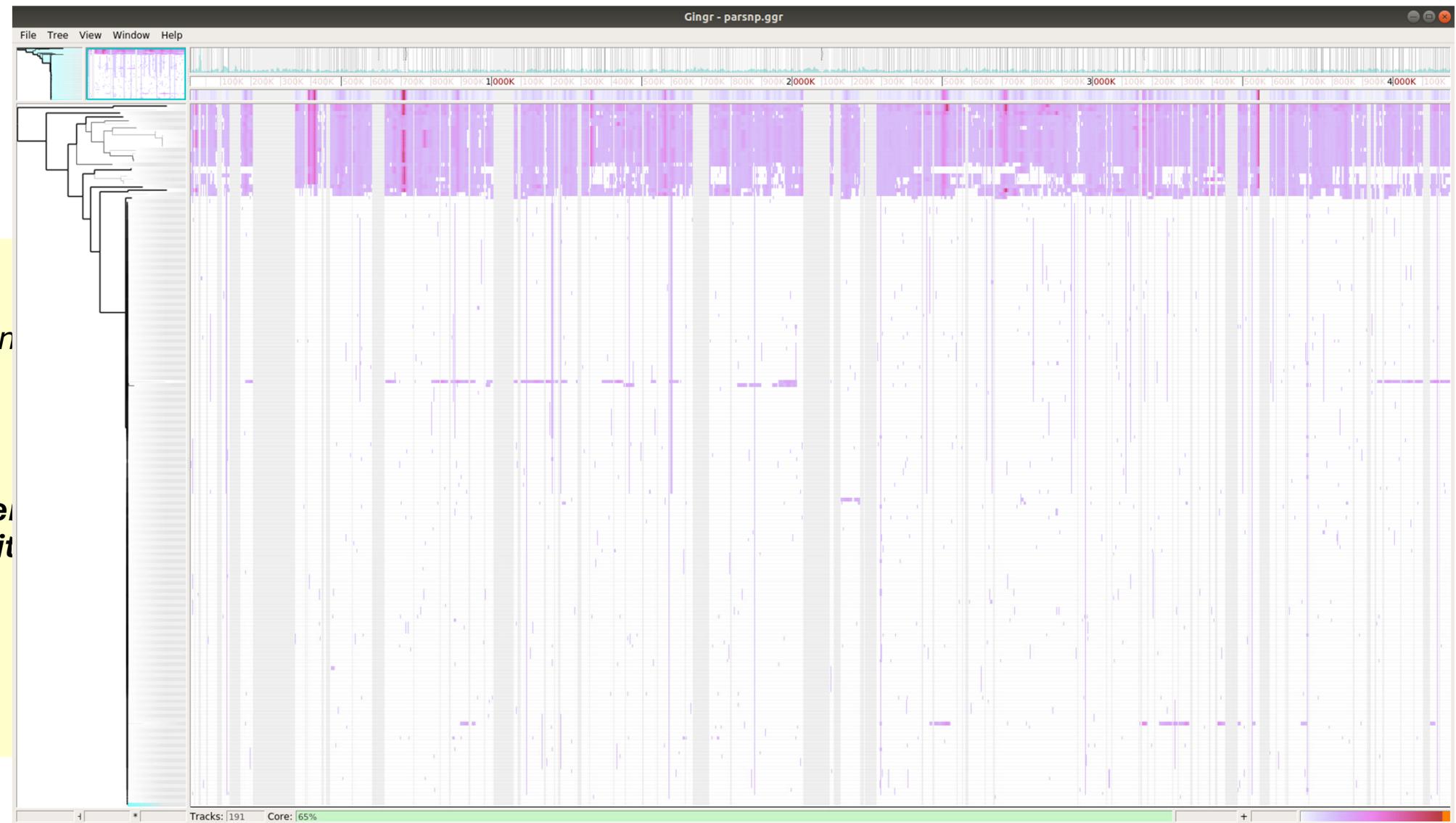
For *C. diff*, even across a huge number of isolates, very little rearrangement shows up (outliers here are reference genomes with single contig, likely a different starting point on the circular genome.)

# 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, very high match level with all RT027 isolates except 3.**

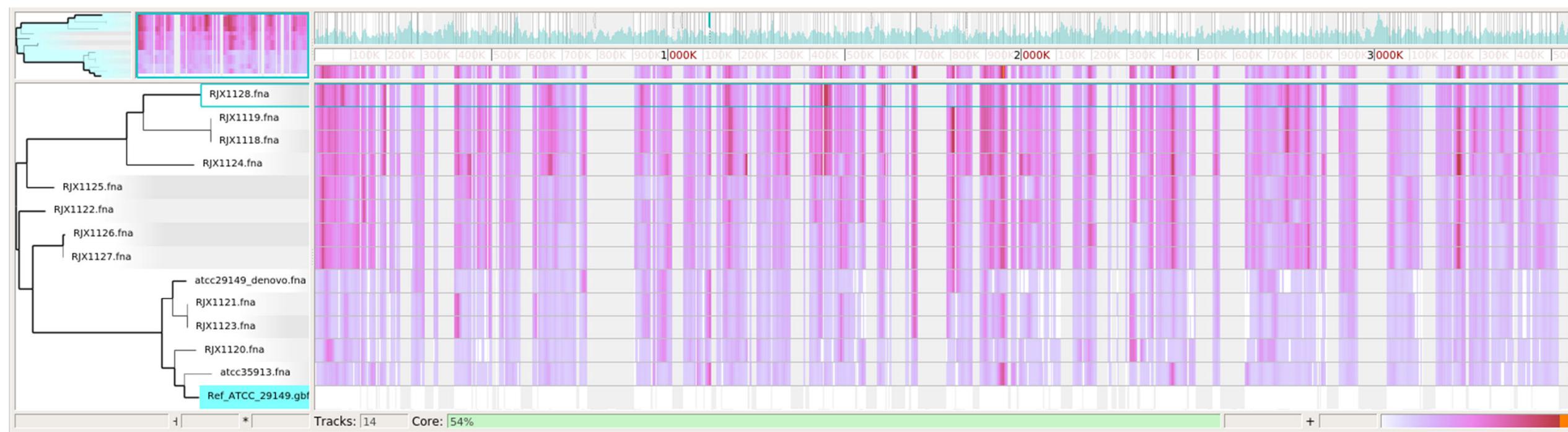


# Case Study #3: *R. gnavus* Isolates from IBD Patients

## 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*	<a href="#">Stool from infant treated with antibiotics</a>
RJX1119*	<a href="#">Stool from infant treated with antibiotics</a>
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



# *R. gnavus* Isolates from IBD Patients

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



# *R. gnavus* Isolates from IBD Patients

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

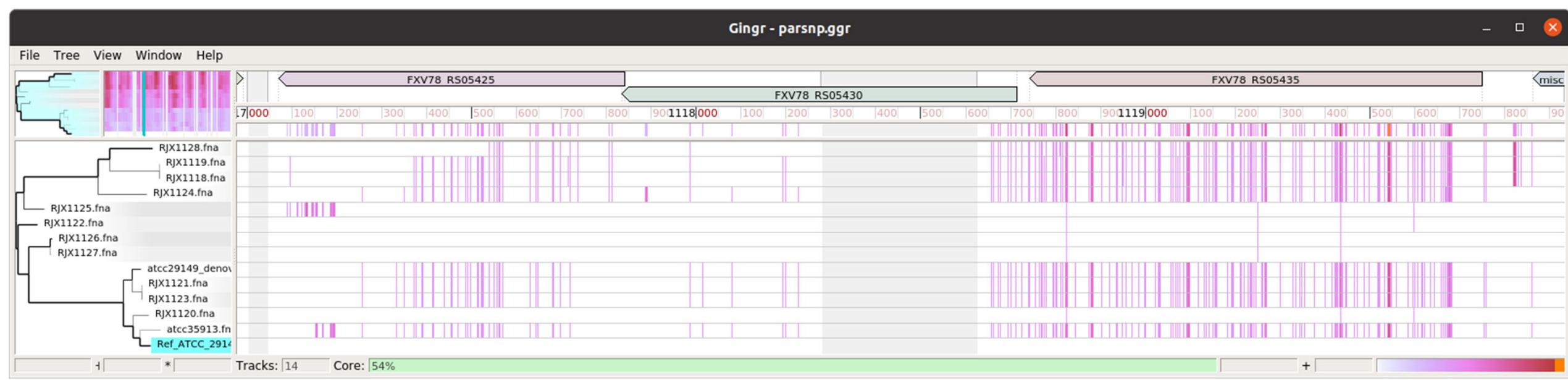
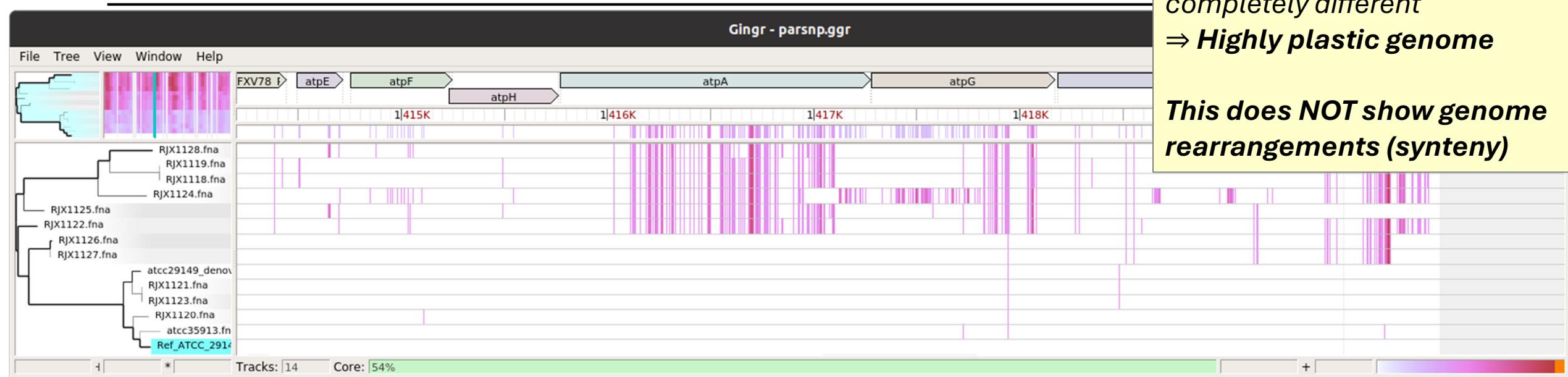
**Game 2 : Spot the 2<sup>nd</sup> ATCC 29149 genome (supposedly the same as the reference)**



# *R. gnavus* strain-level phylogenetic signal is a mess

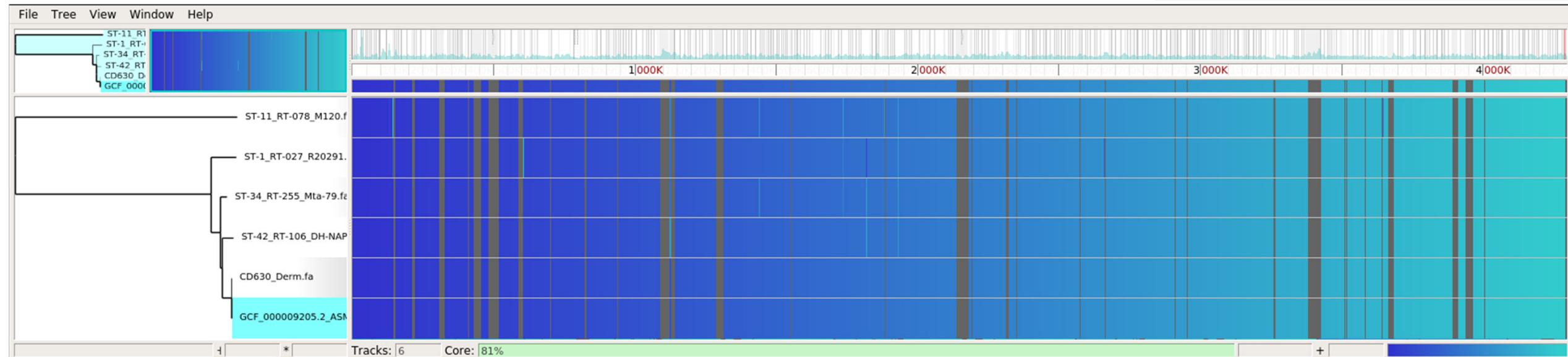
Depending on the operon, the phylogenetic appearance is completely different  
**⇒ Highly plastic genome**

This does NOT show genome rearrangements (synteny)



# 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)

## Bonus Topic:

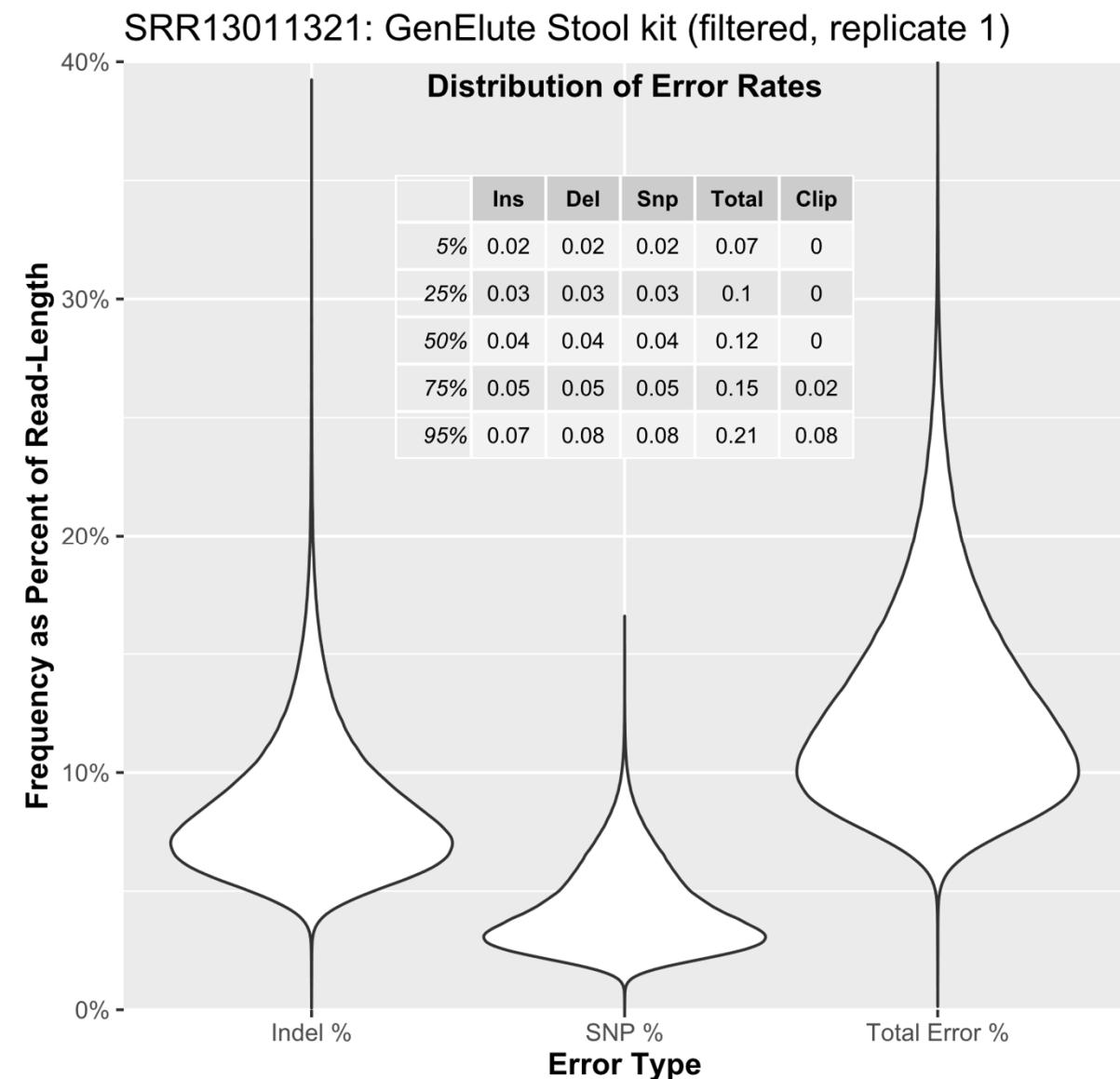
# Taxon Abundance with Full-Length 16S ONT Reads

---

Not the HiFi ones, the noisy stuff...

# ONT reads still have high error

- Data from Mann, et al. (2023)<sup>1</sup>
  - 16S amplicons with 27F/1492R primers
  - Multiple extraction protocols
    - Only GenElute shown at left<sup>2</sup>
    - Others have nearly identical error rates
  - Zymo standard, sequenced on MinION
  - R9.4.1 chemistry
  - Base-called with Guppy 3.2.4
  - Reads filtered with NanoFilt
- Median total error rate: 12%
  - 90% of reads between 7-21% error as percent of read-length
  - Not including soft-clipping, which remains in some reads even post-filtering.



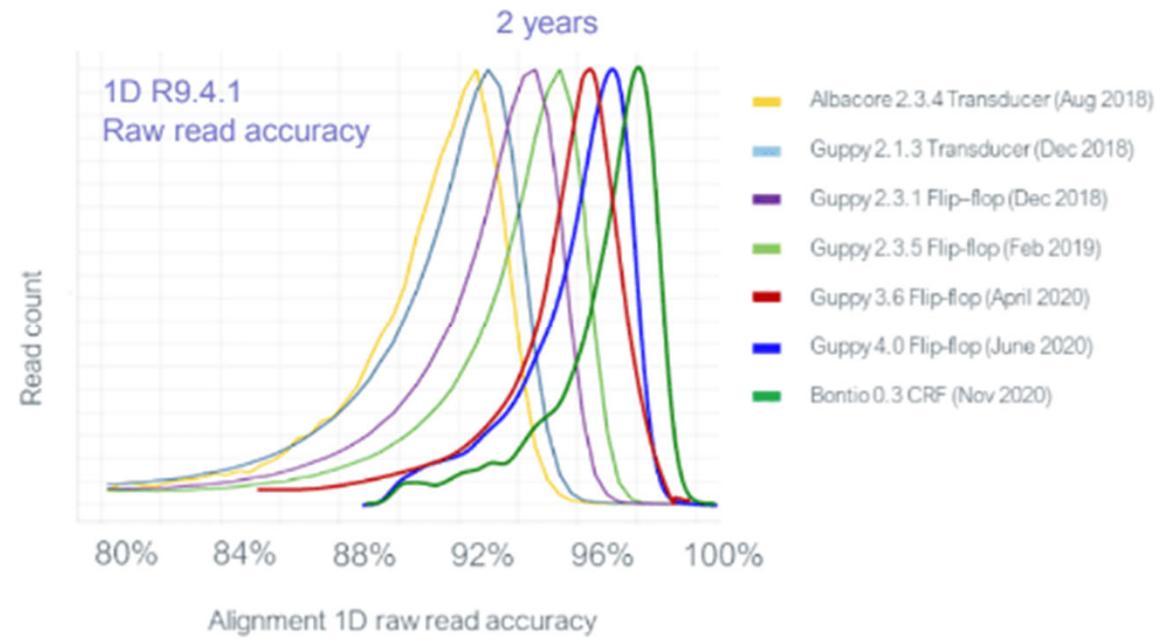
<sup>1</sup>DOI: [10.1016/j.dib.2021.107036](https://doi.org/10.1016/j.dib.2021.107036)

<sup>2</sup>As a sidenote, one of the kits was the QIAamp DNA Microbiome Kit which appeared to **really** mess up the profile of microbes recovered somehow.

# Errors coming down, but still material

- R10 vs. R9 chemistry:
  - Zhang, et al (2023)<sup>3</sup> reports error rates on Zymo Standard
  - Qiagen DNeasy PowersoilPro, PromethION, Guppy v6,
  - R9.4.1 and R10 chemistry
  - (*Data is non-public*)
- Error Rates<sup>4</sup>:
  - 9% total error with Guppy v6 vs. 11-12% ca. 2020

Chem.	Samp.	SNP	Ins.	Del.	Total
R9.4.1	S1	3.1%	1.5%	2.6%	7.2%
	S2	3.5%	1.5%	2.6%	7.6%
	Zymo	4.5%	1.6%	2.8%	8.8%
	All	3.6%	1.5%	2.7%	7.9%
R10.4.1	S1	1.5%	0.4%	0.6%	2.4%
	S2	1.8%	0.4%	0.6%	2.9%
	Zymo	2.9%	0.6%	0.7%	4.2%
	All	2.1%	0.5%	0.6%	3.2%



*Basecalling has been improving, but true error rates still have a ways to go...*

*Image credit: Clive Brown via @nanopore on TwitterX*

<sup>3</sup>DOI: [10.1128/aem.00605-23](https://doi.org/10.1128/aem.00605-23)

<sup>4</sup>This is a subset of Table 2 from Zhang, et al. Zymo is the Zymo Standard D6305, S1/S2 are other samples

# 16S Taxonomic Profiling: Read-Length vs. Error-Rate

---

- Relative Abundance from 16S Amplicons:
  - First 16S taxon identification was accomplished by the ancient Romans *citation needed*
- Illumina Reads:
  - **Short reads** ⇒ limited power to discriminate at species level (at least without extra effort)
  - **Low error** ⇒ classification generally accurate
    - \* *Recall/Sensitivity problem*
- Nanopore Reads:
  - **Full length 16S** ⇒ Species-level classification
  - **High error rate** ⇒ Cross confusion common with near relatives
    - \* *Precision/Specificity problem*

➤ [Microbiome](#). 2020 May 15;8(1):65. doi: 10.1186/s40168-020-00841-w.

Construction of habitat-specific training sets to achieve species-level assignment in 16S rRNA gene datasets

Isabel F Escapa <sup>1 2 3</sup>, Yanmei Huang <sup>1 2</sup>, Tsute Chen <sup>1 2</sup>, Maoxuan Lin <sup>1</sup>, Alexis Kokaras <sup>1</sup>, Floyd E Dewhurst <sup>1 2</sup>, Katherine P Lemon <sup>4 5 6 7</sup>

Affiliations + expand

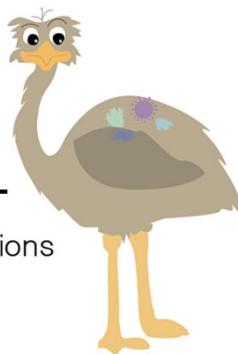
PMID: 32414415 PMCID: [PMC7291764](#) DOI: 10.1186/s40168-020-00841-w

Free PMC article

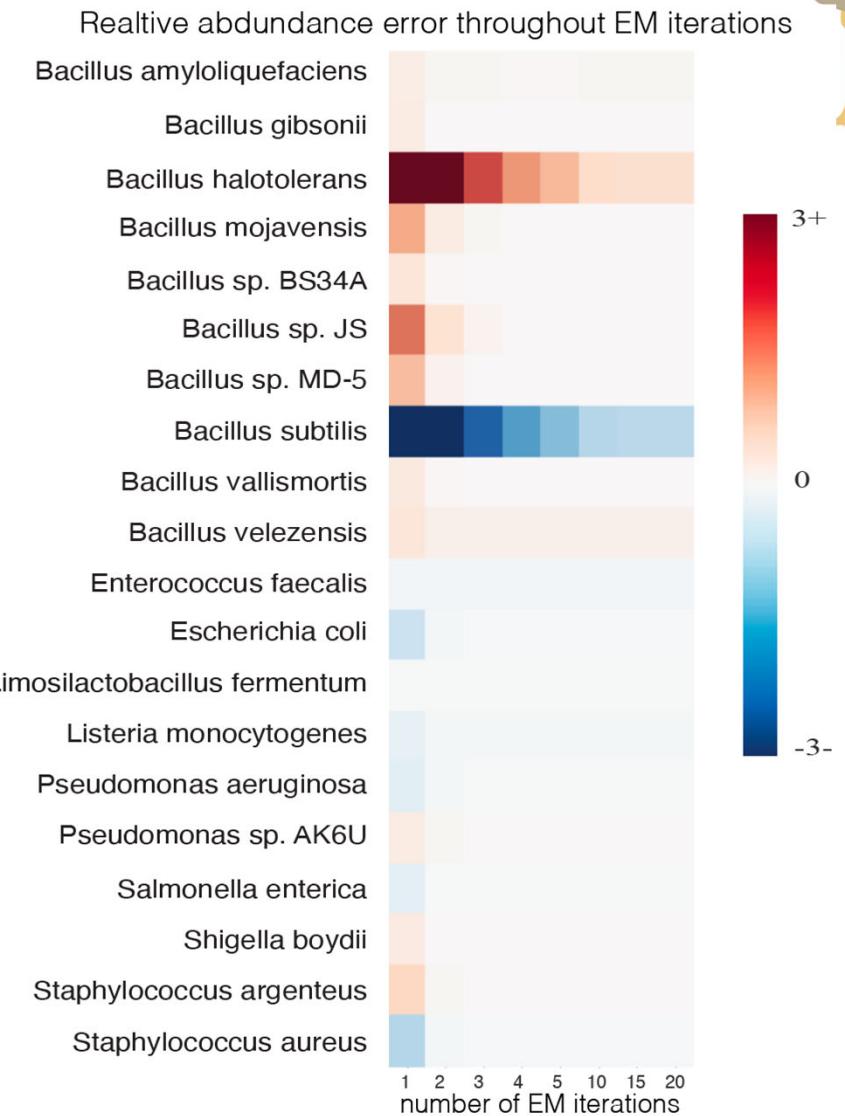
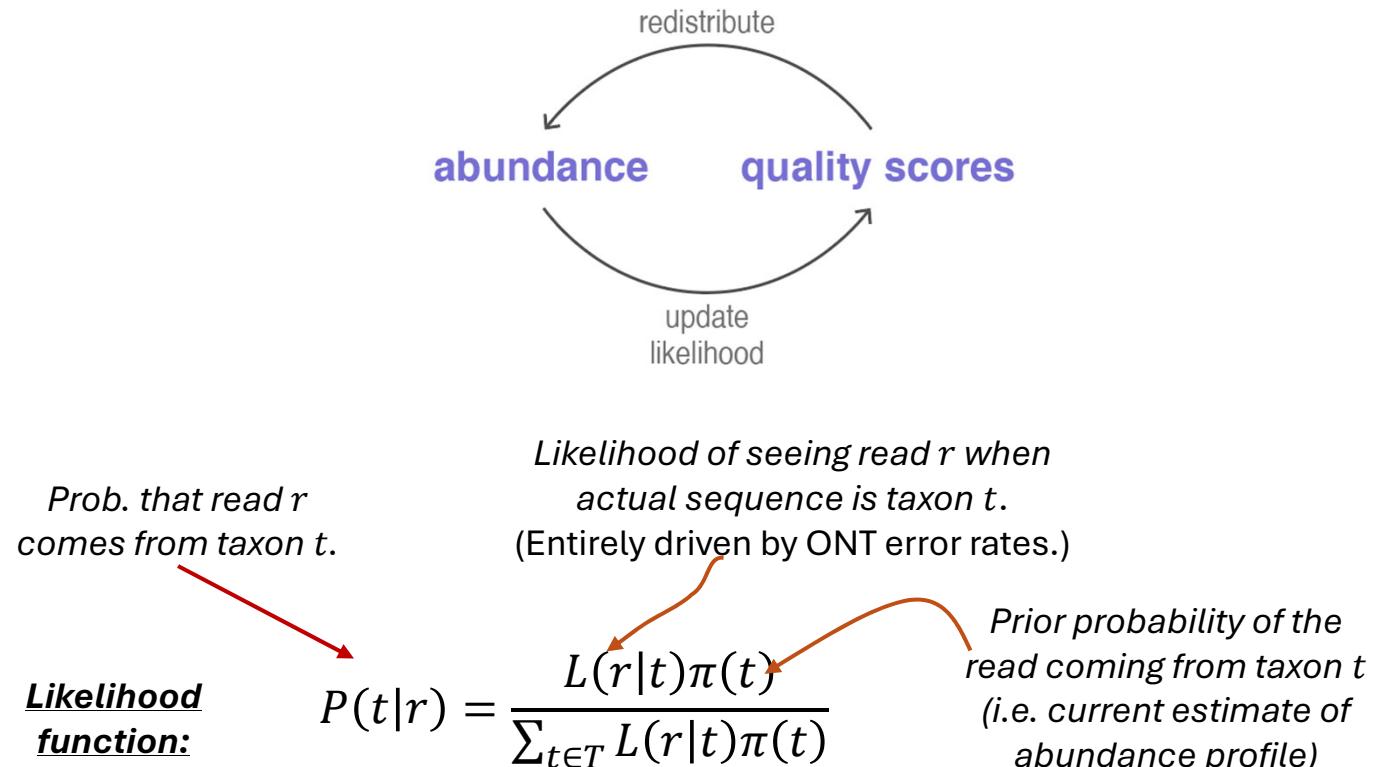
*This is one of the few references that attempts species-level classification based on 16S sequencing using Illumina sequences...*

*...does it by designing domain-specific database.*

# Emu: Species-level Taxon ID for full-length 16S reads



The EM algorithm corrects taxon assignment errors by considering prior likelihood that read was mis-assigned to a close relative.



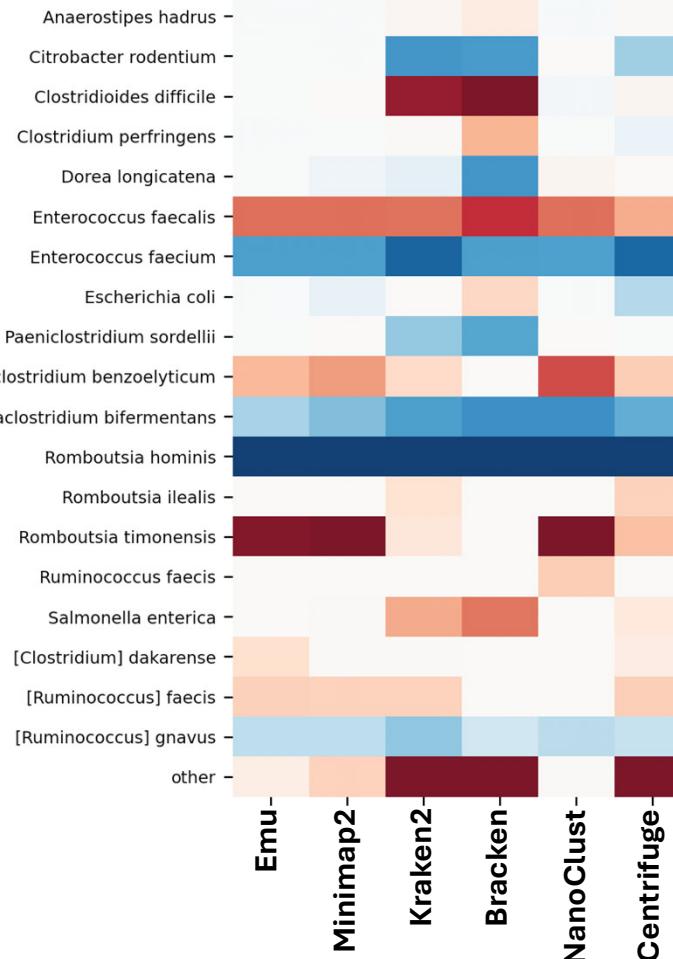
# Example Results: Species-Level Relative Abundance

## Synthetic Gut Community

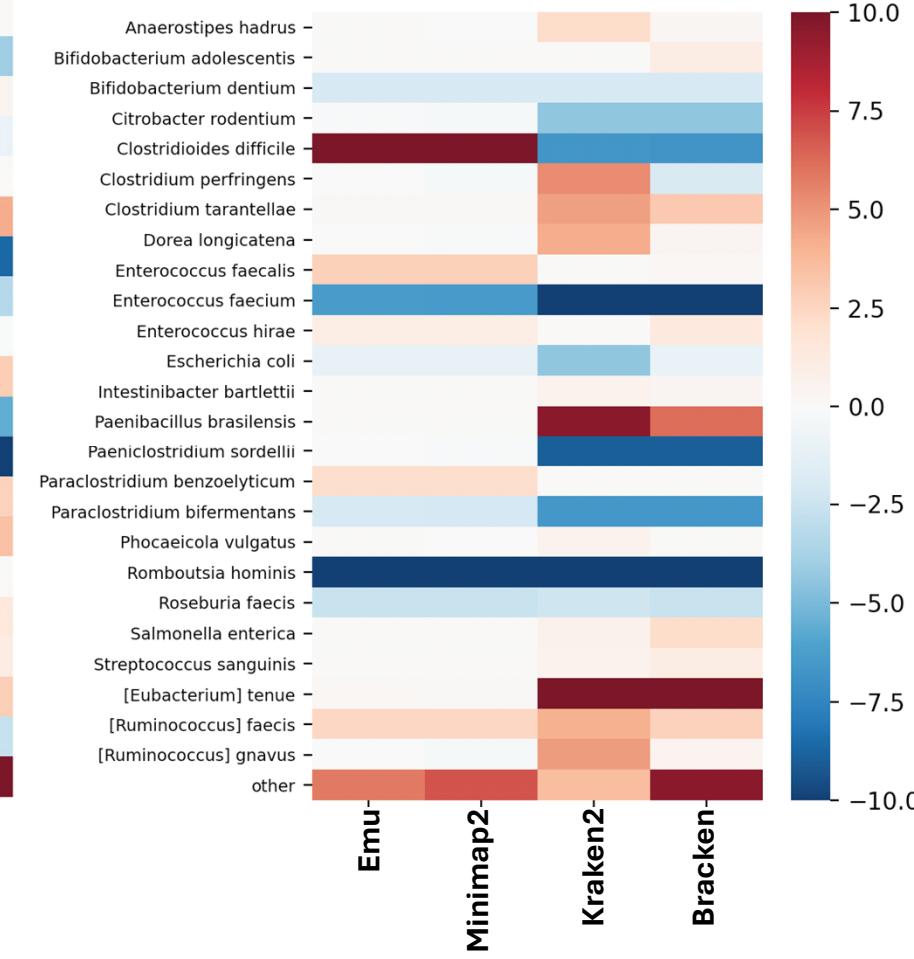
### Approximate Ground Truth Abundance

Strain	CFU	Pct
Bifidobacterium dentum ATCC 27678	20000	18.9
Enterococcus faecium S66643	20000	18.9
Citribacter rodentium ATCC 51459	13000	12.3
Gemmiger formicilis ATCC 27749	10000	9.4
Escherichia coli MG1655	10000	9.4
Clostridium perfringens MT676	9000	8.5
Romboutsia hominis FRIFI	7000	6.6
Clostridium leptum ATCC 29065	5000	4.7
Clostridium scindens ATCC 35704	4000	3.7
Ruminococcus gnavus ATCC 29149	3000	2.8
Bacteroides vulgatus PC510	2000	1.9
Clostridium bartlettii DSM 16795	600	0.5
Clostridium innocuum ATCC 14501	440	0.4
Clostridium bifermentans ERIN_30100	300	0.2
Clostridium sordellii ATCC 9714	240	0.2
Bacteroides thetaiotaomicron VPI -5482	200	0.19
Eubacterium hadrum DSM 3319	200	0.19
Clostridium difficile 630	200	0.19
Dorea longicatena DSM 13814	100	0.095
Roseburia faecis DSM 16840 (M72)	4	0.004

## Nanopore



## Illumina

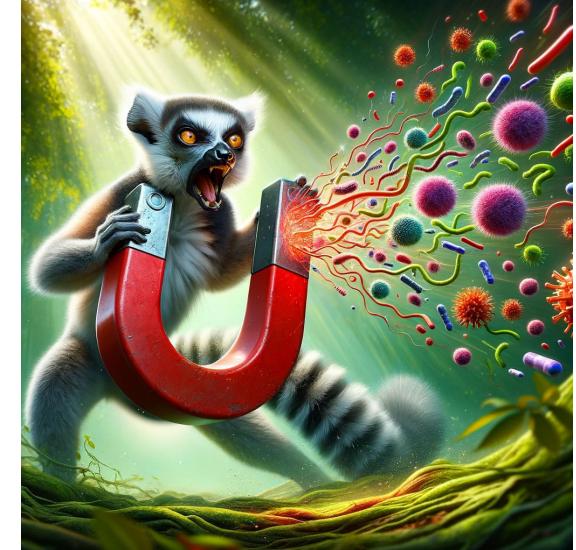
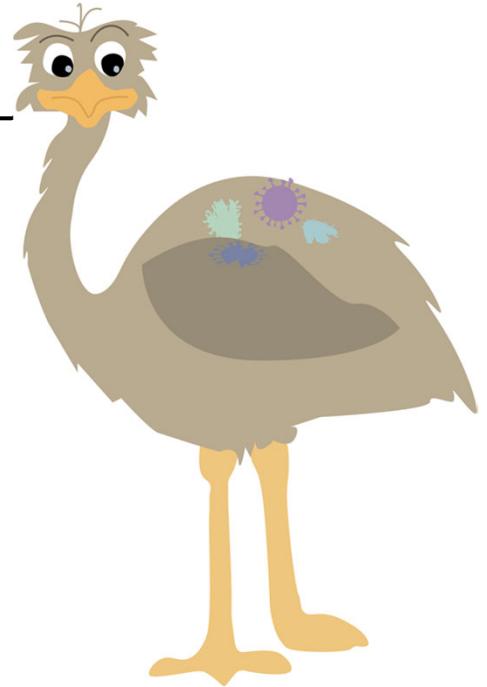


...note that Emu does best here but it was also the only purpose-built 16S ONT method at the time (2021)

# Emu Conclusions

---

- Download: <https://github.com/treangenlab/emu>
- Purpose-built for full-length, high-error 16S amplicon reads
  - I.e. ONT reads, especially R9 chemistry, even with old base-caller
  - Was the first algorithm *really* for this
    - Some newer ones now but we haven't evaluated
- Advantages (big):
  - Relative abundance accuracy (★ ★ ★)
  - Lower false positives (★ ★ ★ ★)
- Disadvantages (no big deal):
  - Some extra run time vs. just minimap2
  - Abundance below some minimum (e.g. 0.5%) is set to 0,
    - So not ideal for estimating very-low abundance members
- Extension to shotgun data via Lemur/Magnet<sup>5</sup> (pre-printed)



# Thank You

---

Please don't hesitate to follow up with me after STAMPS if I can help or go through any of this material again.

# Appendix

---

Multiple Sequence Alignment: Additional Slides

# Contents

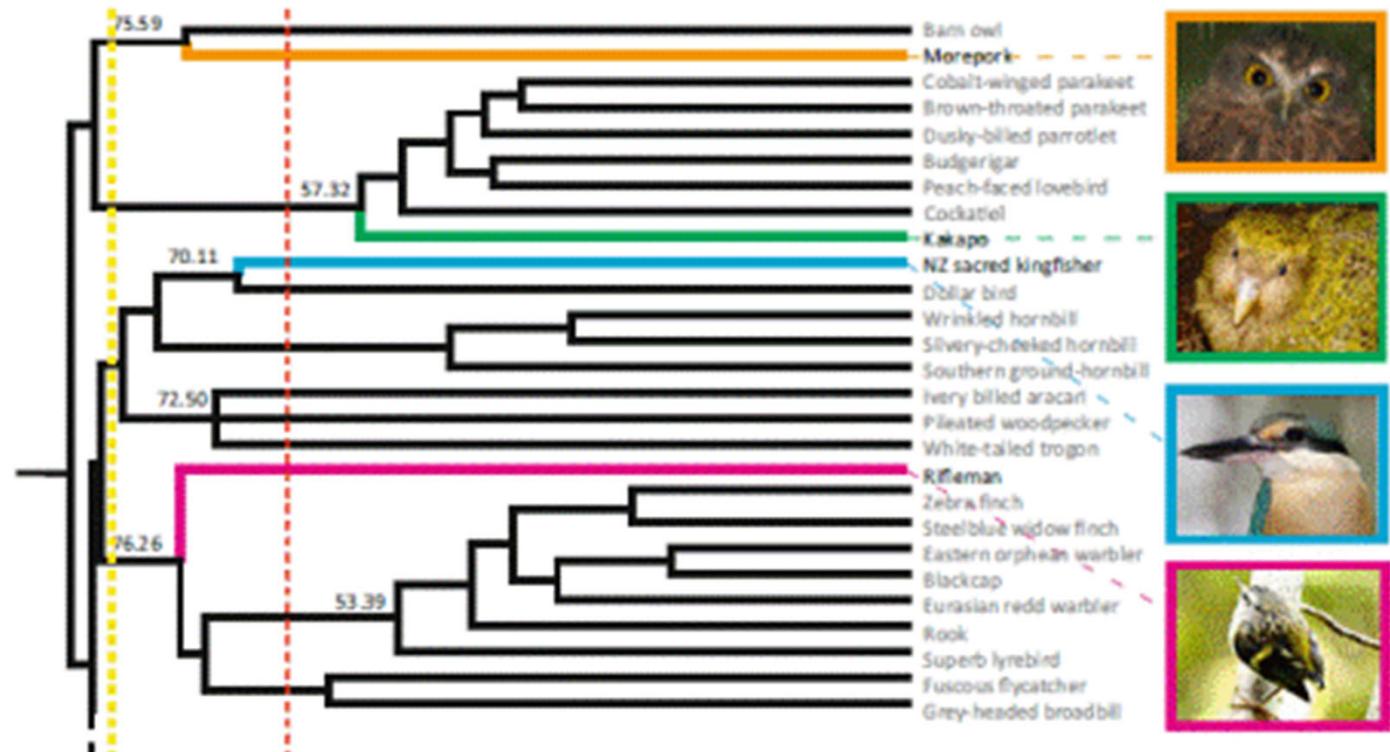
---

- MSA Extras
  - Star Trees, Failure Modes, Accuracy Dropoff
- Tree Estimation Extras
  - GTR details, Phylogenetic Placement
- Species Tree Estimation
- Gingr:
  - Quick How-to & Example

# What Happens If the Alignment is Garbage?

A poor alignment will give a tree with **very deep branches**

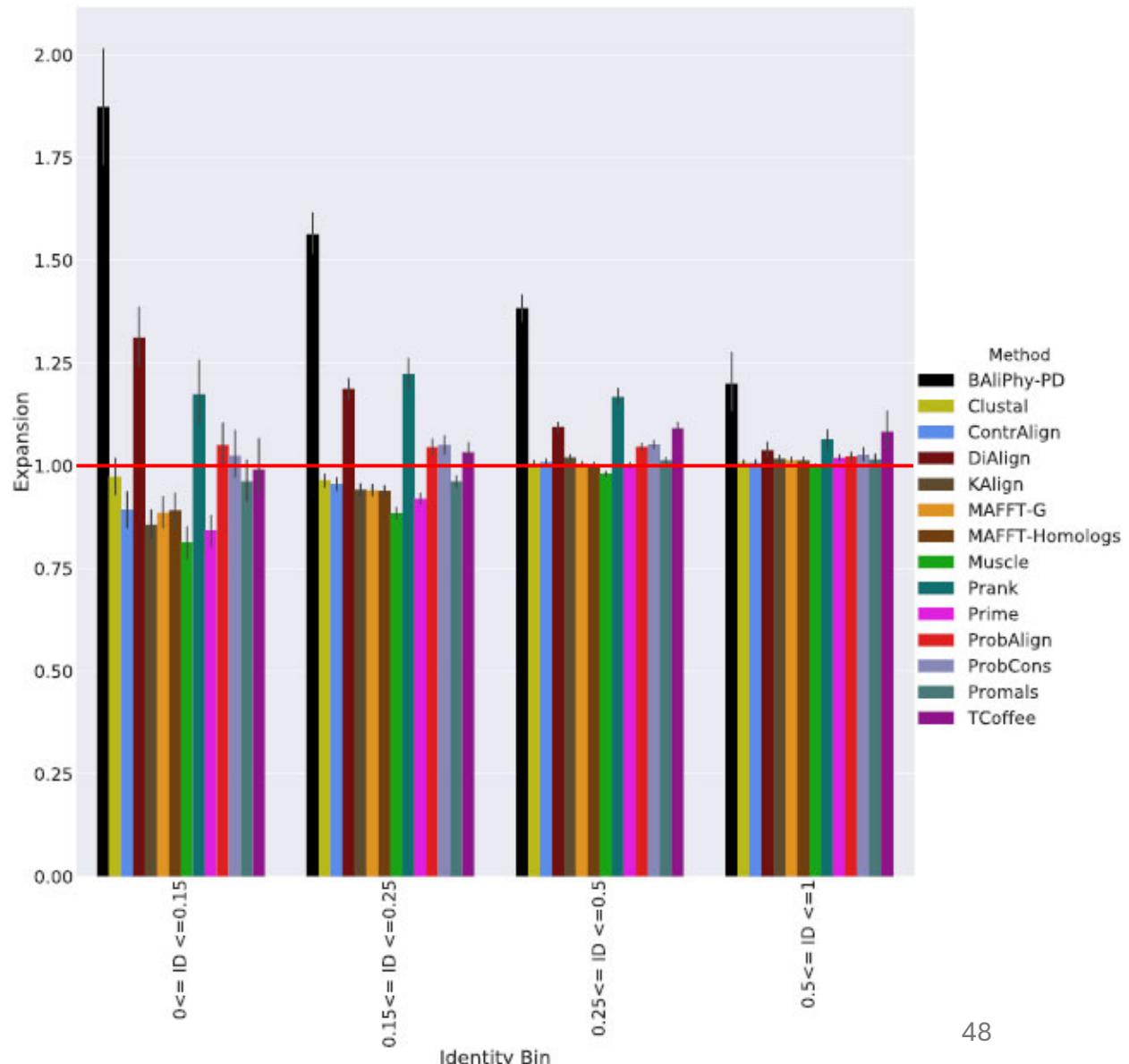
- I.e. long branches above leaves.
  - A “Star”-like tree
- “Nothing in this tree is systematically related to anything else in any significant capacity.”
  - *Could be because relationships were nuked by bad alignment*
- Of course, star-like trees *can* be real! (e.g. birds)



# Failure Modes: Under/Over-Alignment

---

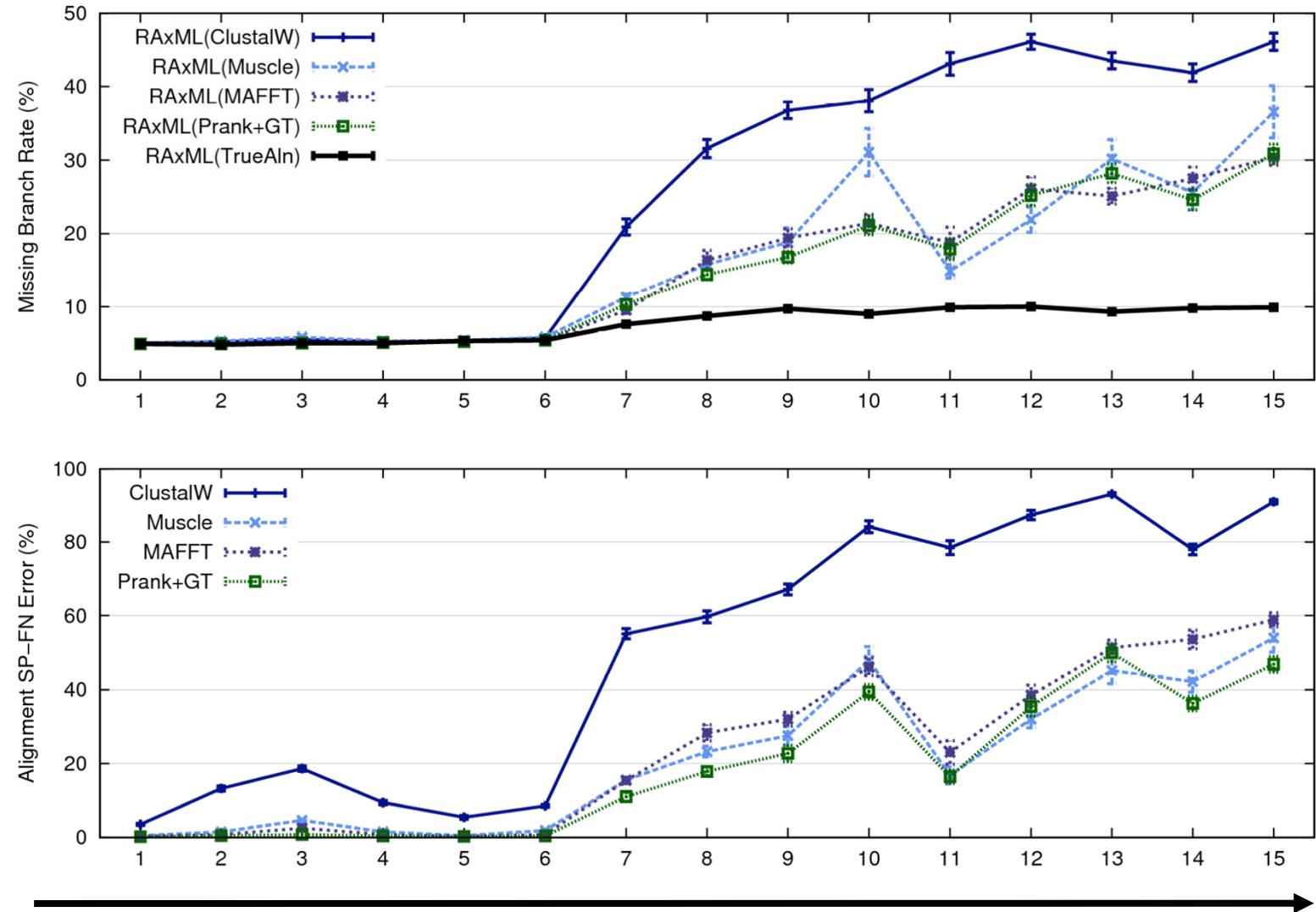
- Most of the commonly used alignment methods will tend to over-align.
- ...*better than aligning just the right amount incorrectly!*
- ...*but can lead to some weird internal branches...*



# Large N, Low %ANI → Very Hard Alignment

*It is far easier than  
widely appreciated  
to get an alignment  
with  $\approx 0\%$   
accuracy.*

**BE CAREFUL  
OUT THERE!**



1000-taxon models, ordered by difficulty (Liu et al., 2009)

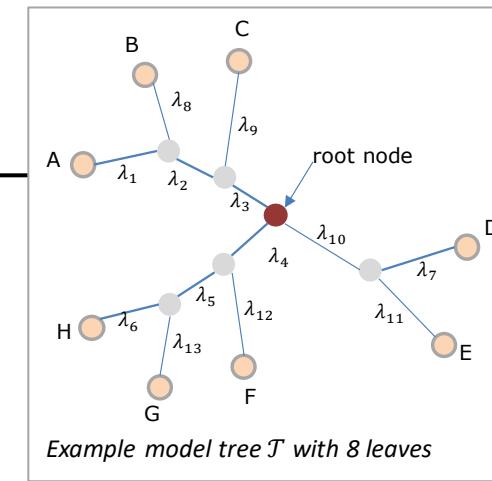
49

# Generalized Time Reversible (GTR) Model

- A probability model for sequence evolution.
  - Characters drawn randomly at the root, then follows a CTMC on each branch in sequence.
- Parameters:
  - Binary Tree  $\mathcal{T} = (T, \Lambda, \rho)$  with  $n$  leaves, topology  $T$ , branch lengths  $\Lambda = (\lambda_1, \dots, \lambda_{2n-3})$  and root node  $\rho$  (actually a nuisance parameter).
  - $\boldsymbol{\pi} = [\pi_A \ \pi_C \ \pi_G \ \pi_T]$  is the probability vector of each character state at the root.
  - Transition rate matrix:

$$Q = \begin{bmatrix} - & a\pi_C & b\pi_G & c\pi_T \\ a\pi_A & - & d\pi_G & e\pi_T \\ b\pi_A & d\pi_C & - & f\pi_T \\ c\pi_A & e\pi_C & f\pi_G & - \end{bmatrix}$$

where the values on the diagonal are set so the row sums to 0.



- Notes:
  - Assumes all sites (columns) evolve *i.i.d.*
  - No indels.
  - Given sequences generated according to the GTR, ML estimation of  $(T, \Lambda, \boldsymbol{\pi}, Q)$  is statistically consistent.
  - Model is equivalent for any choice of root location.
  - In practice, gaps in the alignment are treated as missing data.
  - RAxML<sup>1</sup> and FastTree<sup>2</sup> are two off-the-shelf programs for ML phylogeny estimation.
  - ML topology estimation is NP-Hard.
  - **Branch lengths are denominated in expected mutations per site.**

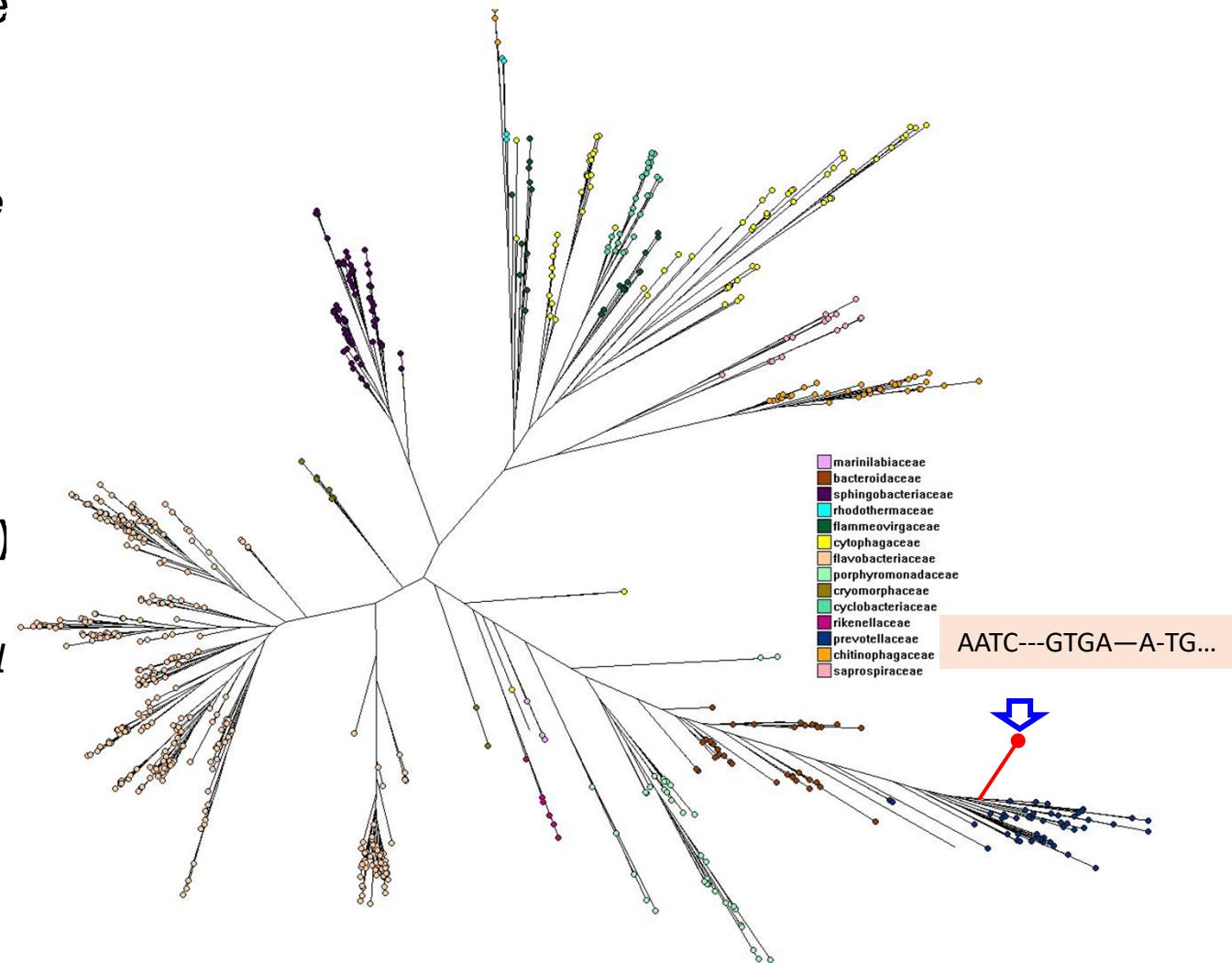
<sup>1</sup><https://cme.h-its.org/exelixis/web/software/raxml/index.html>

<sup>2</sup><http://www.microbesonline.org/fasttree/>

# Related Problem: Phylogenetic Placement

- **Problem:** consider a fixed “reference” group of aligned sequences and a phylogeny  $T$ .
  - given a new sequence  $q$  find the best **new branch** in  $T$  with  $q$  at the leaf (via ML).

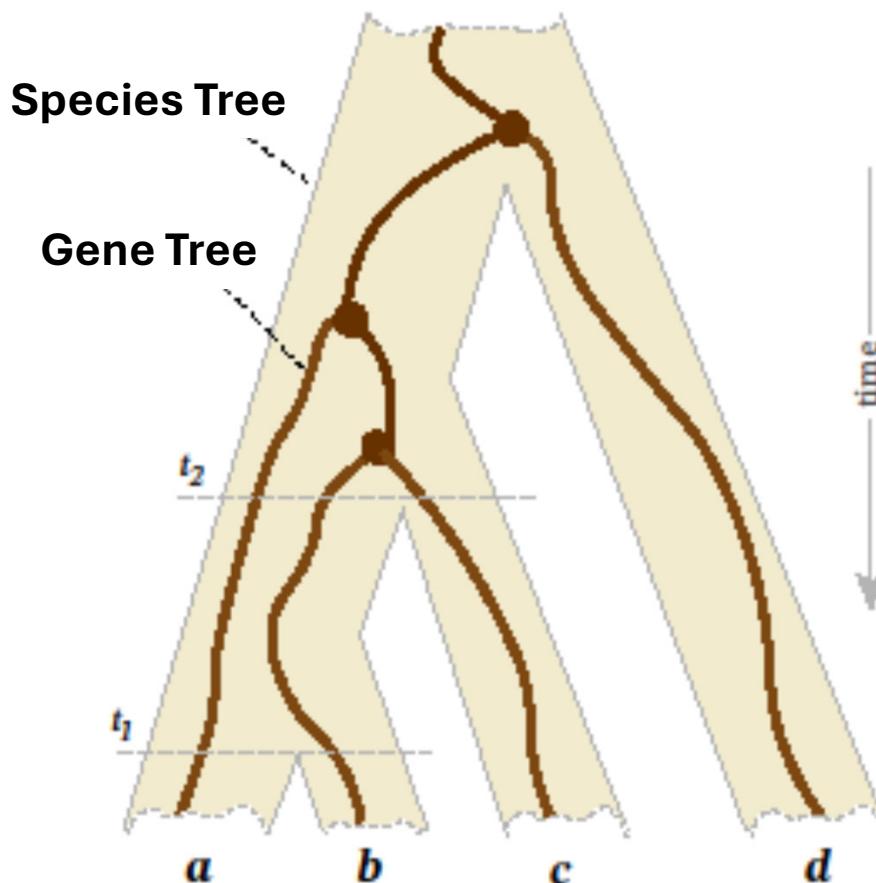
- New branch includes:
  1. **Attachment Point**
    - *Identifies the closest relative in the data.*
  2. **New Branch Length (“pendant length”)**
    - *Distance from its closest relative.*
    - *A measure of novelty relative to the original sequences*



# Species Trees: The Multi-Species Coalescent

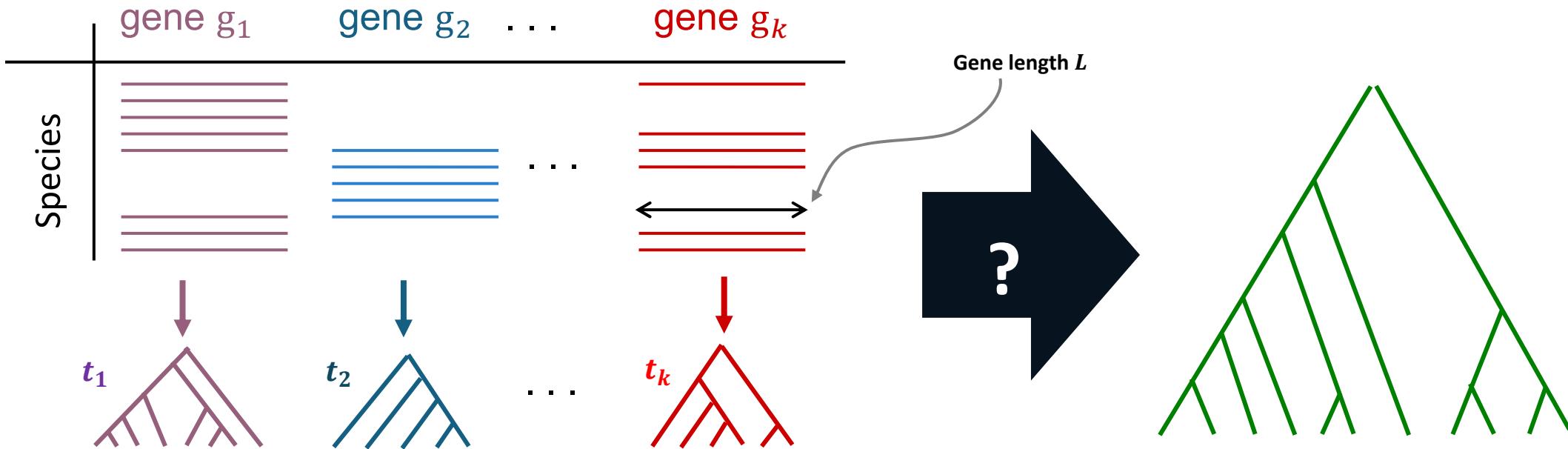
---

- Genes evolve within a *population*
  - So gene branching events can happen without full population speciation
  - Can give rise to genes with a different tree shape than the species



# Species Tree Estimation Problems

Problem: given multiple different genes generated from a single population species tree according to the multi-species coalescent (MSC), how do we estimate the population phylogeny for the species? And when is our estimate consistent?

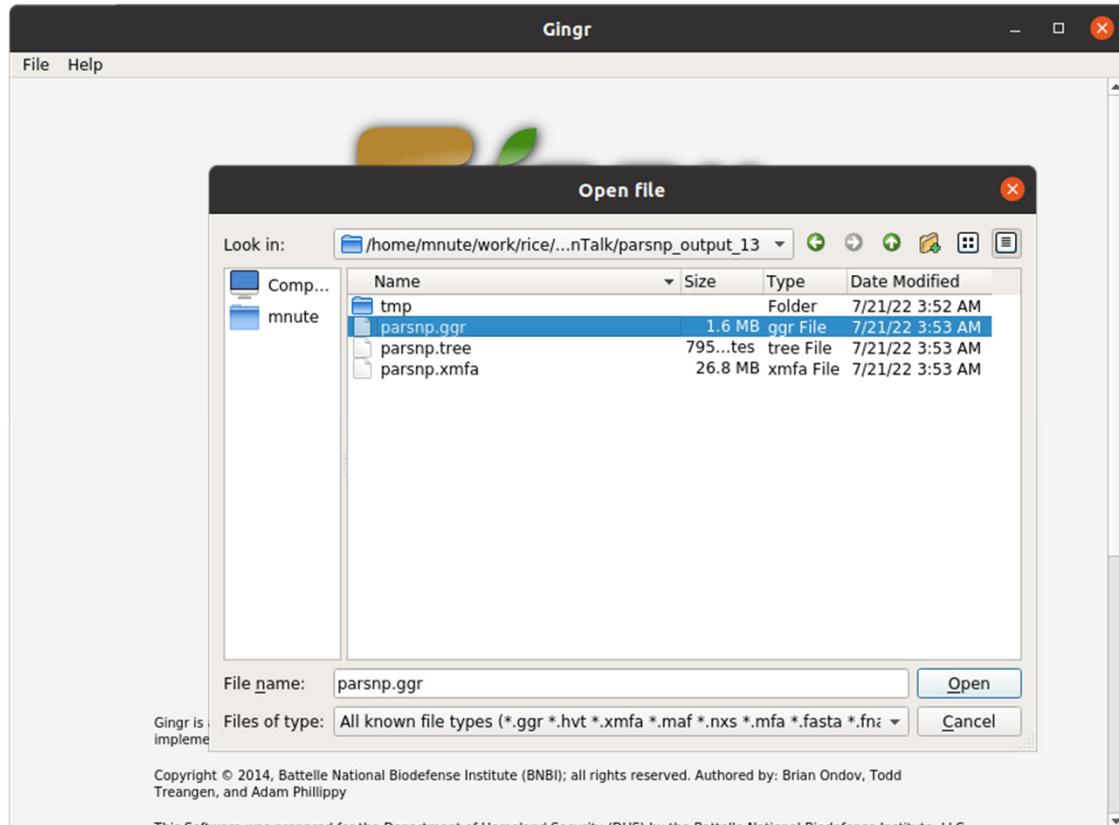


# Species Tree Estimation Options

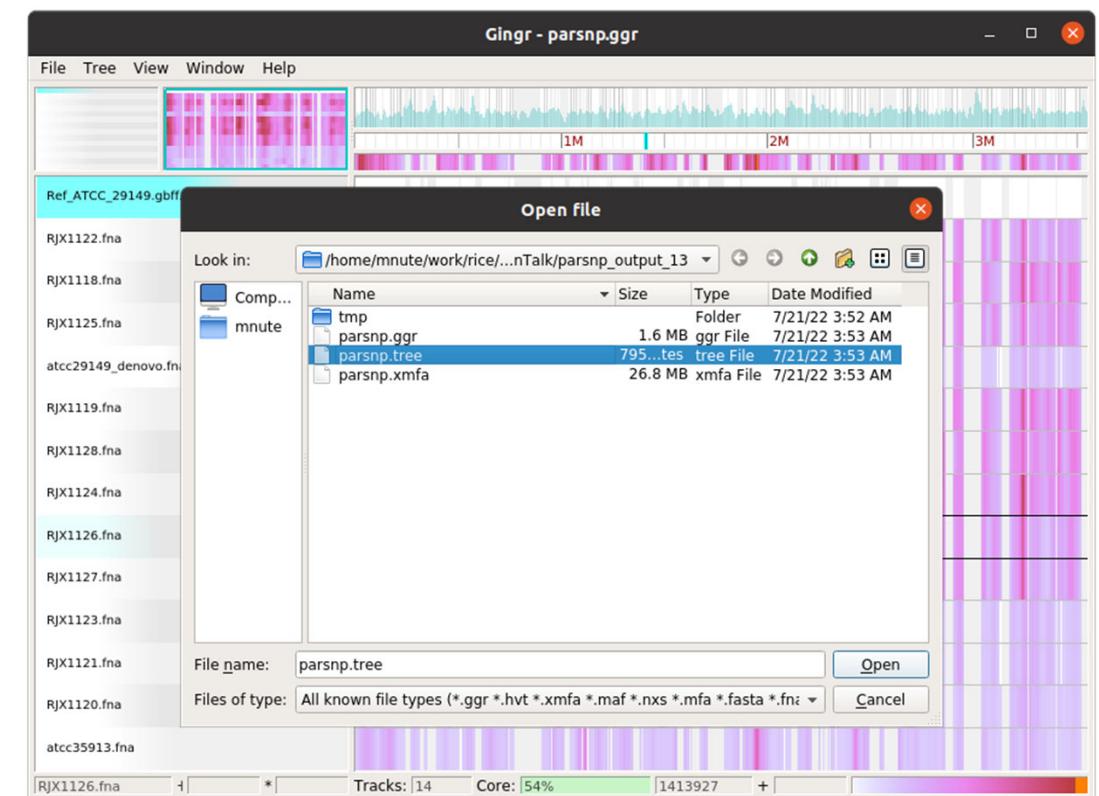
---

- What can we do to estimate a species tree when all the individual genes have their own topology?
  - Concatenation: join the individual gene-alignments together, in order, and estimate a tree under GTR.
    - Advantages: simple, RAxML can handle long sequences,
    - Disadvantages: ***not*** statistically consistent.
      - Not always a problem in practice though unless there is a lot of heterogeneity in gene trees.
      - In other words, species tree has many *relatively* recent speciation events.
  - Quartet methods: estimate the tree separately for every gene, then use a software specifically to combine the gene trees
    - E.g. ASTRAL
    - **IS** statistically consistent
    - Not quite as simple of a pipeline.
- My advice: use ASTRAL
  - ...*but I know that most people don't bother.*

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

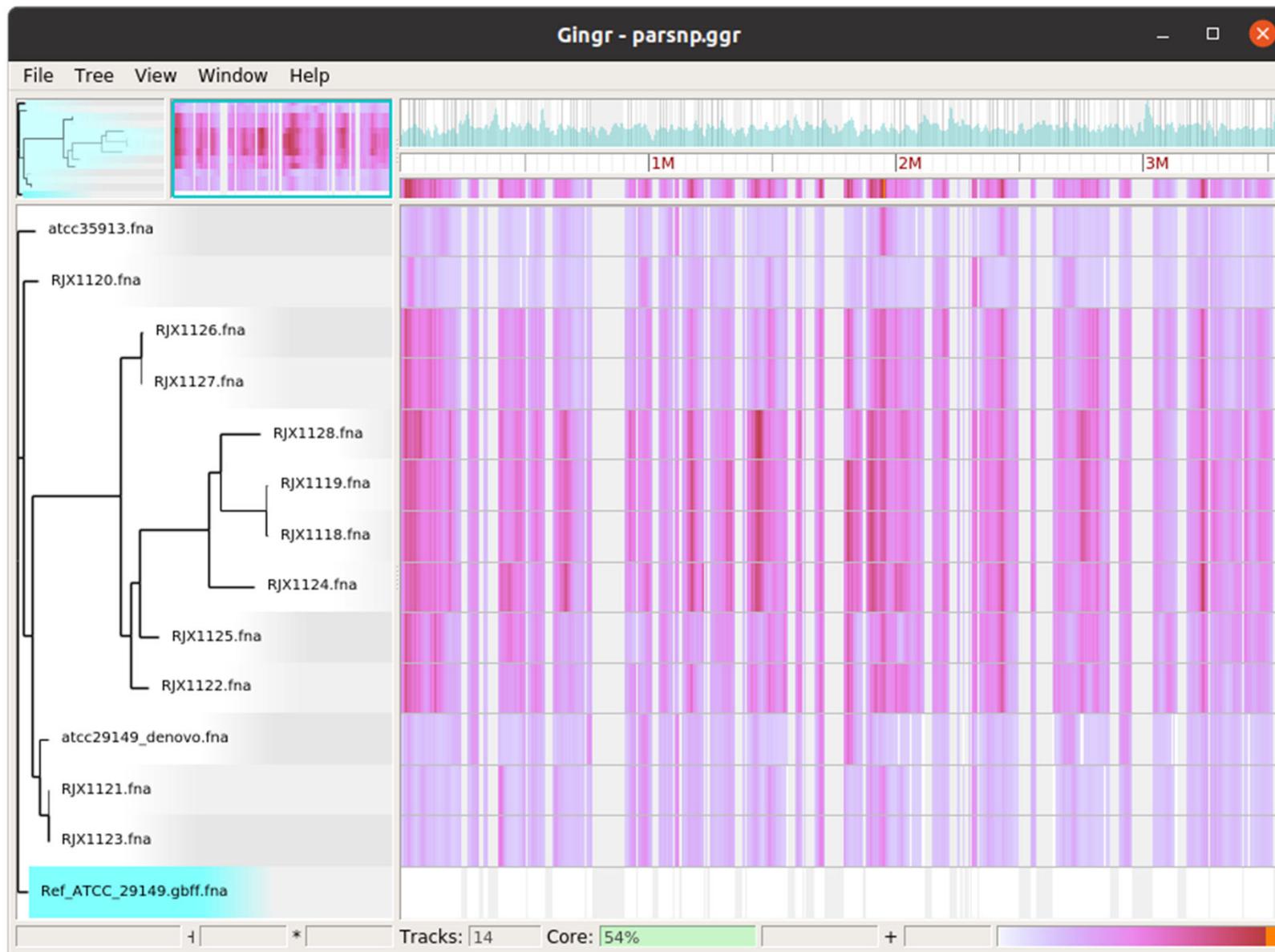


1.) Open the \*.ggr file created in the parsnp output folder.



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.