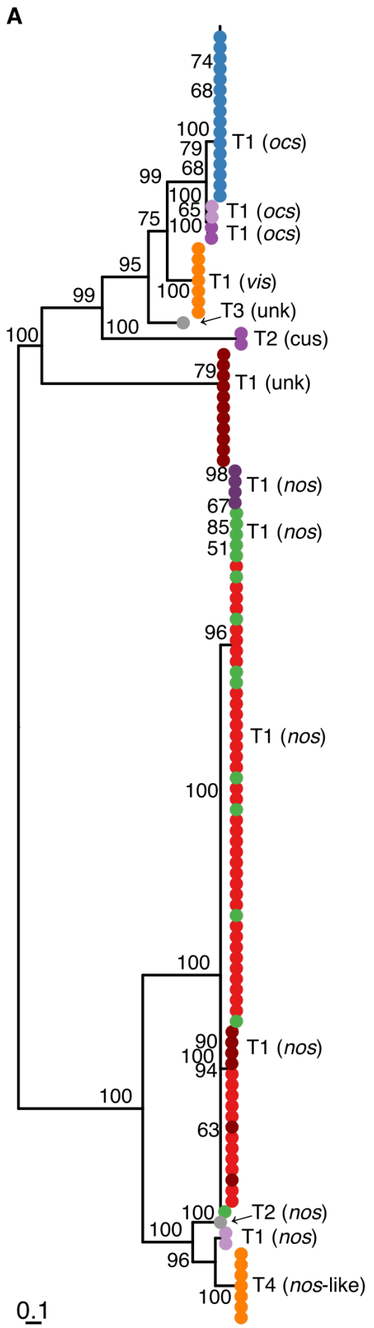
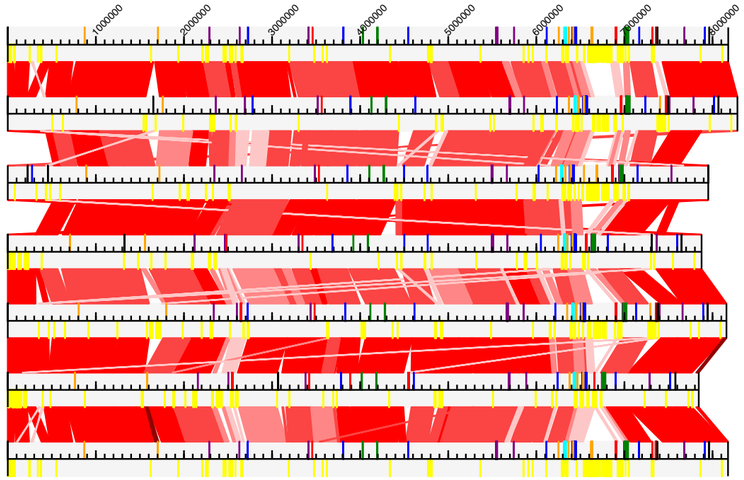
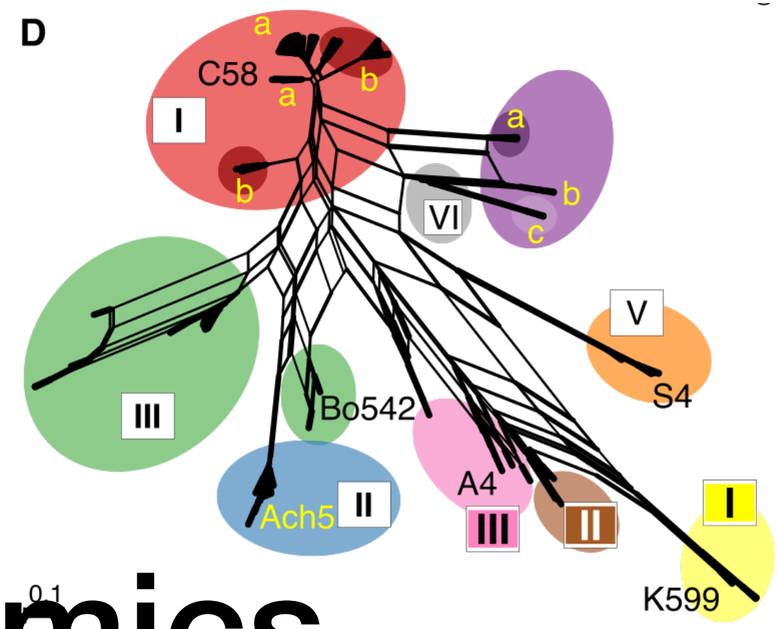


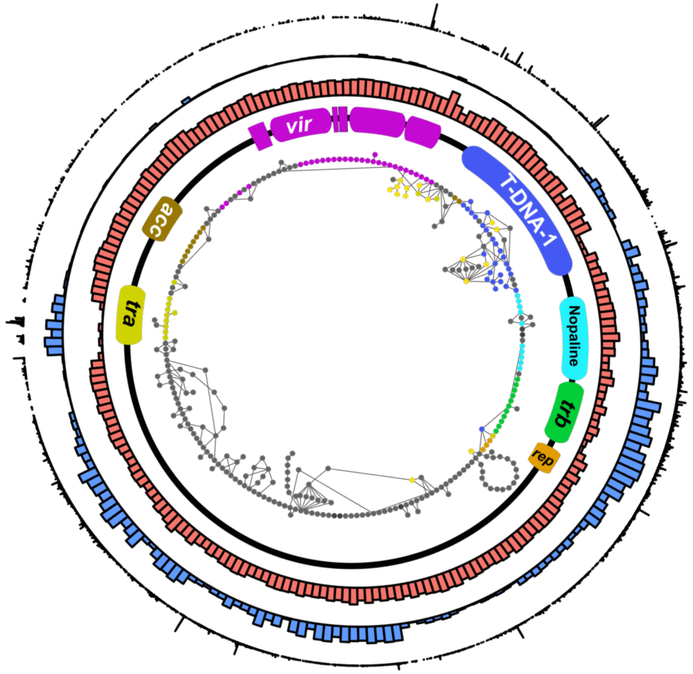




Microbial Genomics Analyses

How do we analyze and visualize data on a genome scale?

Alexandra J. Weisberg,
Chang Lab, Oregon State University



Two genome sequencing technologies

Illumina (HiSeq, MiSeq)

100-250 bp sequenced paired end
Very high base quality (~0.1% error rate)
Very high sequencing depth
Reads are short, can't span repetitive regions
~$80-$100 per genome - 96 per run

Oxford Nanopore (MinION)

Extremely long reads (1 kb – 1 Mbase)
Can assemble complete genomes
Individual read quality is poor (5-10% error rate)
~$100 per genome – 12-24 per flow cell

Hybrid assembly

Complete or near complete assembly (Nanopore) +
High base confidence (Illumina)

I have 10 [100, 1000,…] sequenced genomes, where do I start?

K-mer hashing can quickly approximate relationships or identify contaminants

- K-mer = DNA sequence of length K, can be ”hashed” to quickly compare sequences
- Identify species from short reads or genome assembly in < 2 seconds (**BBtools sendsketch.sh**)
- Pairwise relationships using K-mer hashing (**sourmash**)

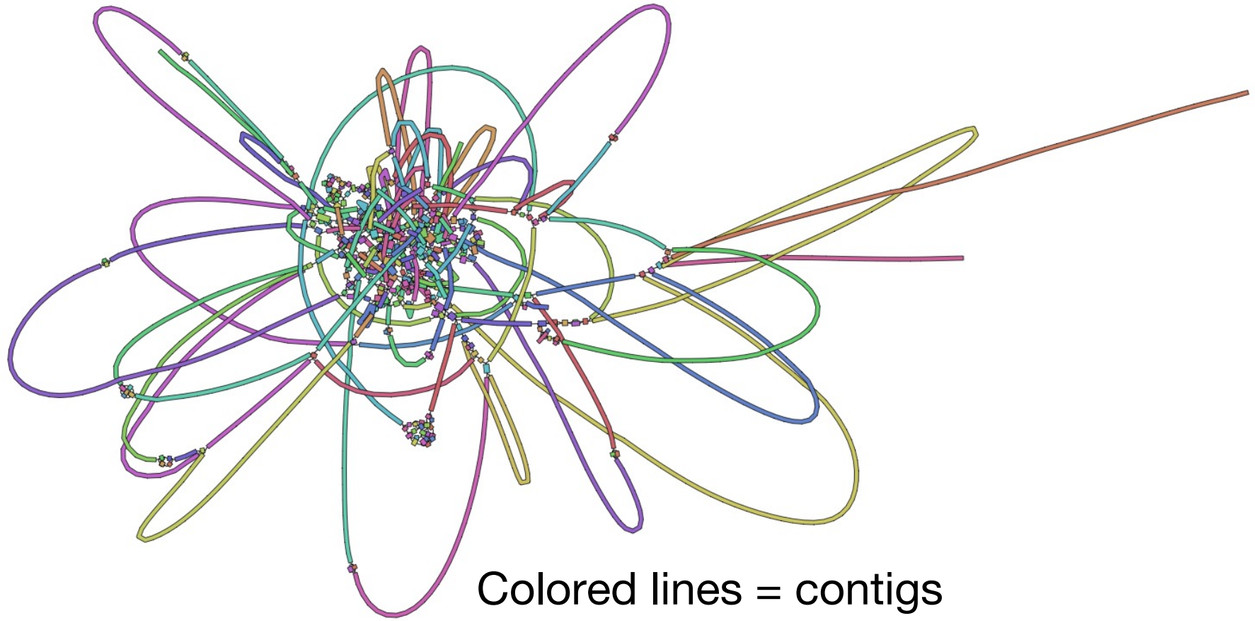
sendsketch.sh myreads.fastq.gz →

Query: lane7-s048-index-ATAACGGT-D12_S48_L007_R1_001.fastq.gz DB: RefSeq SketchLen: 34056 Seqs: 400000 Bases:

WKID	KID	ANI	Complt	Contam	Matches	Unique	TaxID	gSize	gSeqs	taxName	file
52.64%	32.26%	97.69%	94.83%	41.89%	10985	3228	680198	9719205	1	Streptomyces scabiei 87.22	taxa30.sketch
74.98%	24.97%	98.96%	100.00%	49.18%	8504	98	2184002	5263559	25	Bacillus sp. AG236	taxa19.sketch
74.71%	24.34%	98.95%	100.00%	49.81%	8288	94	1286363	5146458	22	Bacillus sp. 171095_106	taxa18.sketch
48.27%	16.63%	97.40%	100.00%	57.52%	5664	37	1292043	5402416	47	Bacillus flexus 27Col1.1E	taxa10.sketch
50.22%	15.75%	97.53%	100.00%	58.40%	5363	2	592022	4943423	1	Bacillus megaterium DSM 319	taxa25.sketch
48.86%	15.64%	97.44%	100.00%	58.51%	5327	0	2052936	5005107	1	Bacillus sp. Y-01	taxa29.sketch
49.29%	15.41%	97.47%	100.00%	58.74%	5247	3	1827146	4911849	1	Bacillus sp. IHB B 7164	taxa28.sketch
45.09%	15.81%	97.16%	100.00%	58.34%	5383	8	1736468	5525916	27	Bacillus sp. Root147	taxa4.sketch
43.63%	15.83%	97.04%	100.00%	58.32%	5391	1	1736389	5719803	34	Bacillus sp. Soil531	taxa18.sketch
42.86%	15.61%	96.98%	100.00%	58.54%	5317	1	2293318	5732190	46	Bacillus sp. ALD	taxa24.sketch
42.48%	15.45%	96.95%	100.00%	58.70%	5262	0	1380110	5733872	65	Bacillus sp. RP1137	taxa14.sketch
41.62%	15.61%	96.88%	100.00%	58.54%	5315	1	1736235	5897614	75	Bacillus sp. Leaf75	taxa6.sketch
41.70%	15.55%	96.88%	100.00%	58.59%	5297	1	1581029	5844297	22	Bacillus sp. FJAT-21351	taxa19.sketch
40.63%	15.69%	96.79%	100.00%	58.46%	5344	1	2293322	6075970	45	Bacillus sp. RC taxa7.sketch	
40.66%	14.68%	96.79%	100.00%	59.47%	4999	7	1736499	5692516	46	Bacillus sp. Root239	taxa16.sketch
26.97%	14.72%	95.36%	74.32%	59.43%	5014	16	1358420	8692093	142	Bacillus aryabhattai B8W22	taxa30.sketch
35.97%	11.46%	96.36%	100.00%	62.69%	3902	4	1292019	5042856	40	Bacillus sp. 278922_107	taxa3.sketch
35.06%	11.31%	96.28%	100.00%	62.84%	3852	3	1202456	5082111	458	Bacillus sp. Aph1	taxa3.sketch
10.76%	7.83%	92.23%	25.88%	64.41%	2968	660	146819	21695K	4178	Streptomyces europaeiscabiei	taxa19.sketch
12.00%	7.69%	92.60%	52.32%	66.46%	2620	18	146820	10102K	951	Streptomyces stelliscabiei	taxa22.sketch

Visualizing *de novo* assemblies

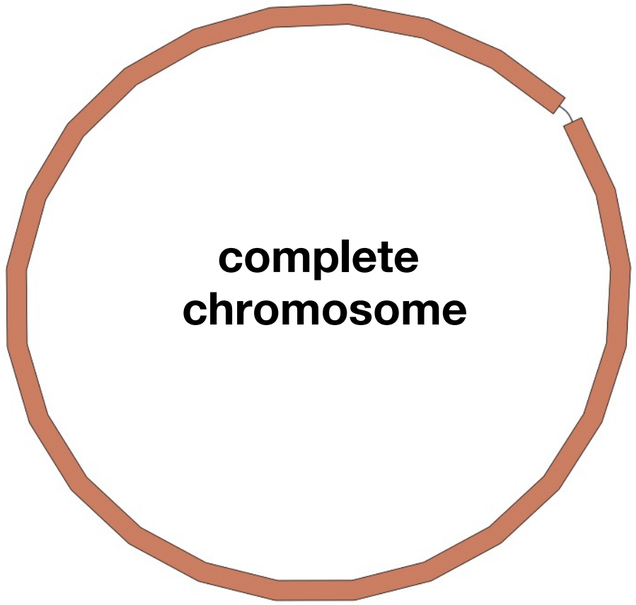
Illumina-only assembly



Colored lines = contigs

(assembled using **SPAdes**)

Hybrid Illumina/Nanopore assembly



megaplasmids

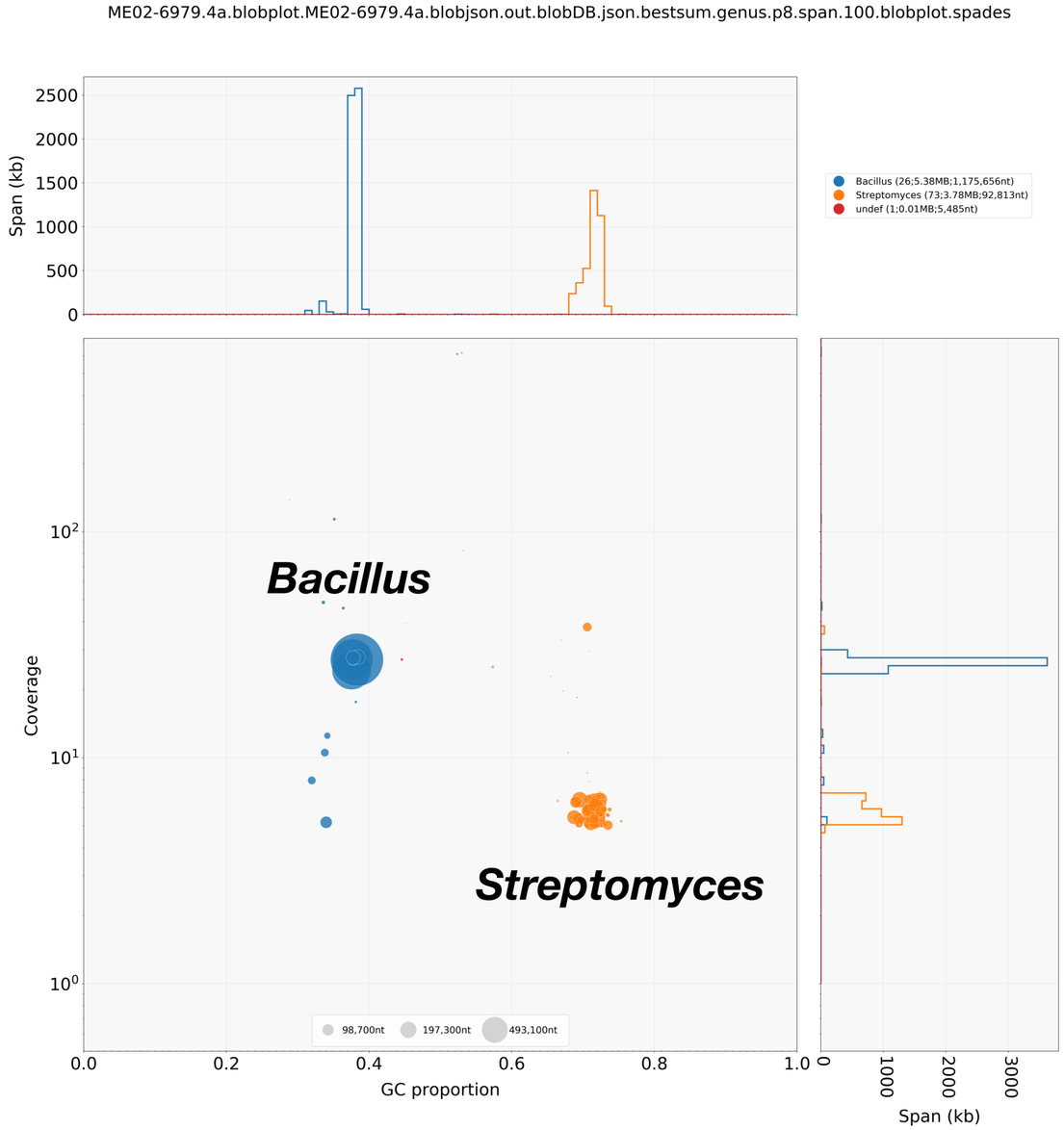
(assembled using **unicycler**)

Bandage – view assembly graph (.gfa output file)

“Bandage image assembly.gfa assembly.png” (from command line)

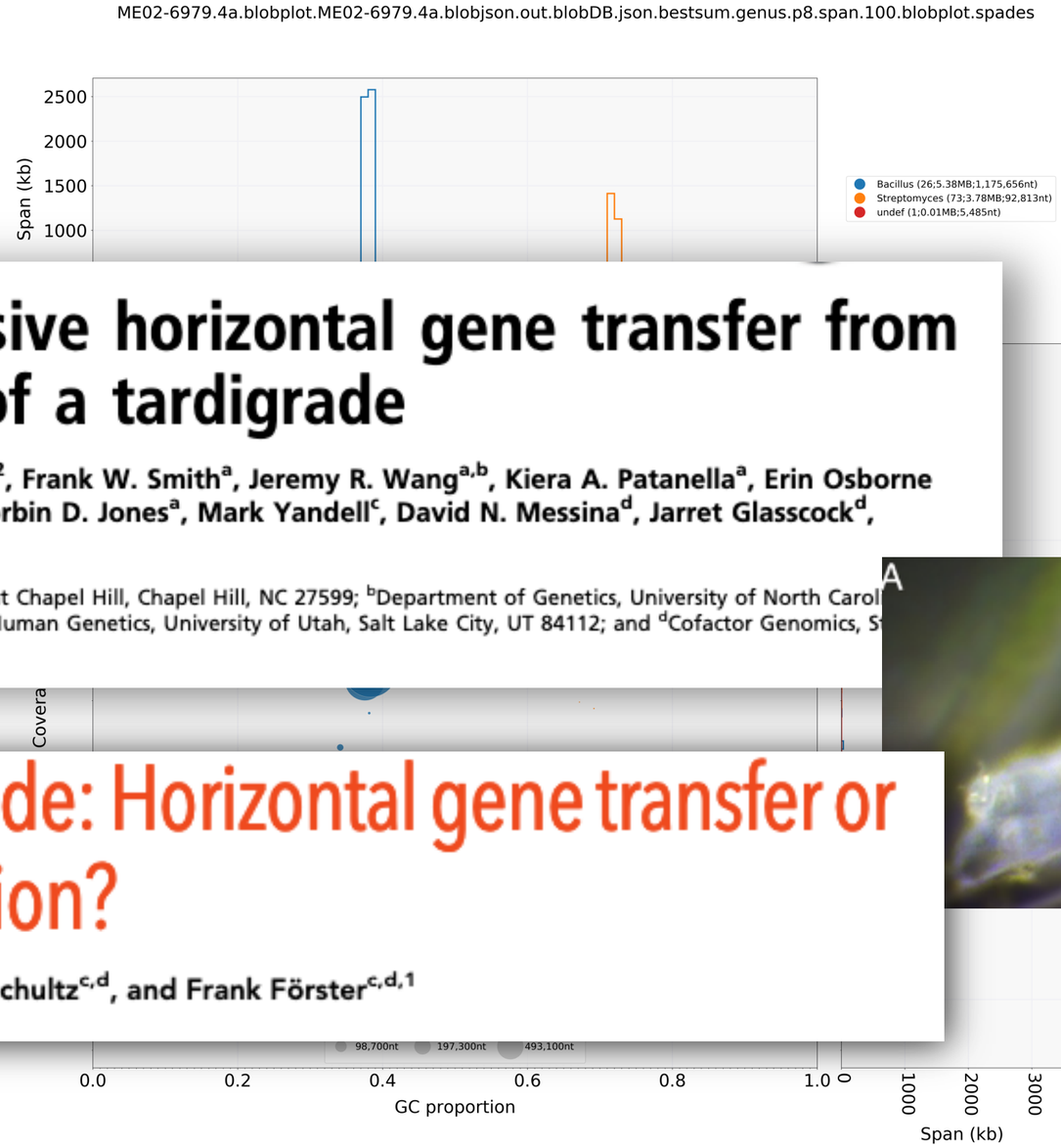
Verify a genome assembly with blobplots

- **Blobtools** blobplots
- GC content vs read coverage
- Circles = contigs
  - Color = species (megablast)
  - Size = contig length



Verify a genome assembly with blobplots

- **Blobtools** blobplots
- GC content vs read coverage
- Circ
- Colo

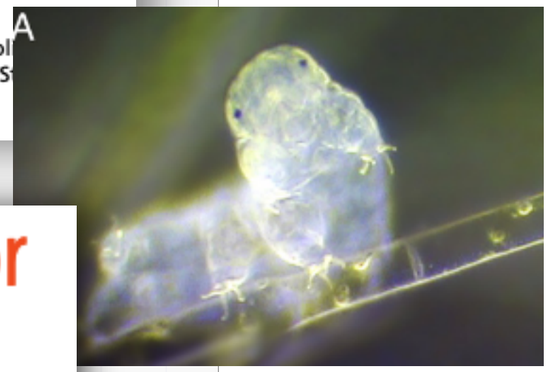


PNAS

Evidence for extensive horizontal gene transfer from the draft genome of a tardigrade

Thomas C. Boothby[a,1], Jennifer R. Tenlen[a,2], Frank W. Smith[a], Jeremy R. Wang[a,b], Kiera A. Patanella[a], Erin Osborne Nishimura[a], Sophia C. Tintori[a], Qing Li[c], Corbin D. Jones[a], Mark Yandell[c], David N. Messina[d], Jarret Glasscock[d], and Bob Goldstein[a]

[a]Department of Biology, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599; [b]Department of Genetics, University of North Carol
Chapel Hill, Chapel Hill, NC 27599; [c]Eccles Institute of Human Genetics, University of Utah, Salt Lake City, UT 84112; and [d]Cofactor Genomics, S
MO 63110



Genome of a tardigrade: Horizontal gene transfer or bacterial contamination?

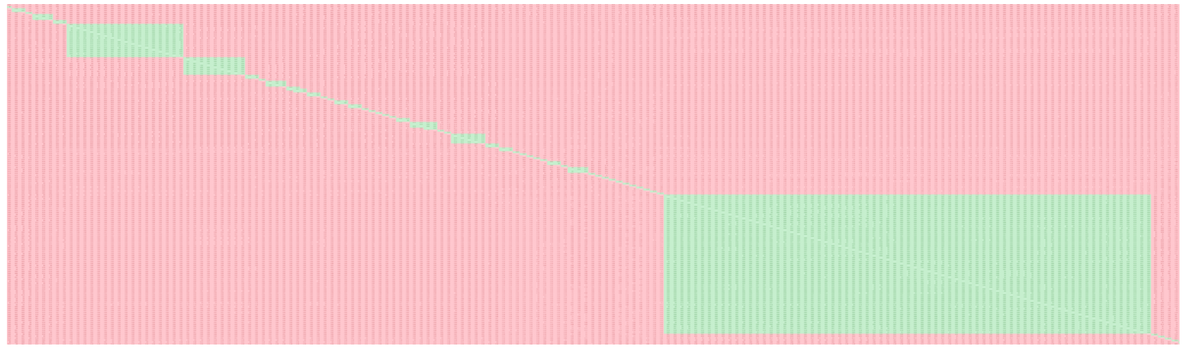
Felix Bemm[a], Clemens Leonard Weiß[b], Jörg Schultz[c,d], and Frank Förster[c,d,1]

How diverse are my sequenced strains?

Taxonomic names are confusing, contentious, and often incorrect

How many species are represented?

- Average Nucleotide Identity (**ANI**)
- >95% pairwise ANI for same species-level group
- Include strains from NCBI to provide context
- **autoANI**, **fastANI**



Pairwise ANI table, green: >95% ANI

How many genera?

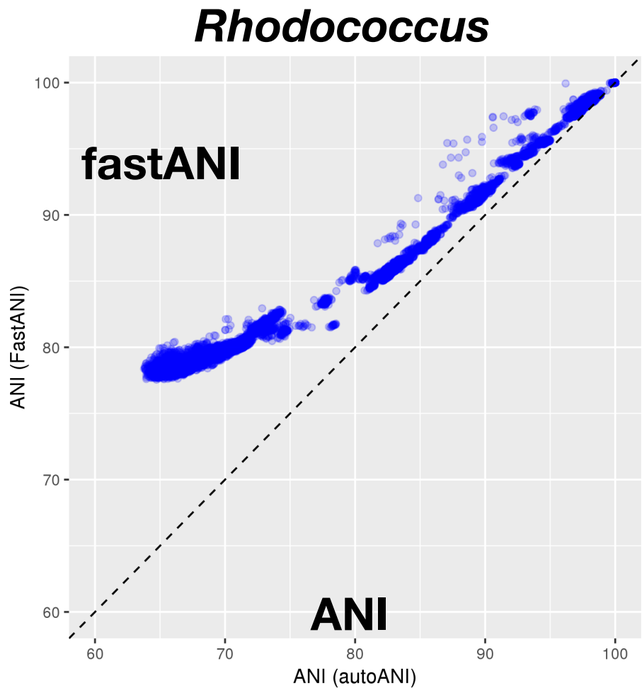
- Percentage of Conserved Proteins (**POCP**)
- >50% POCP = same genus-level group (Qin, 2014)
- **autoPOCP**, **get_homologues -P**

★ **ANI** = similarity of **shared** regions ★

POCP = **proportion** of genome that is shared

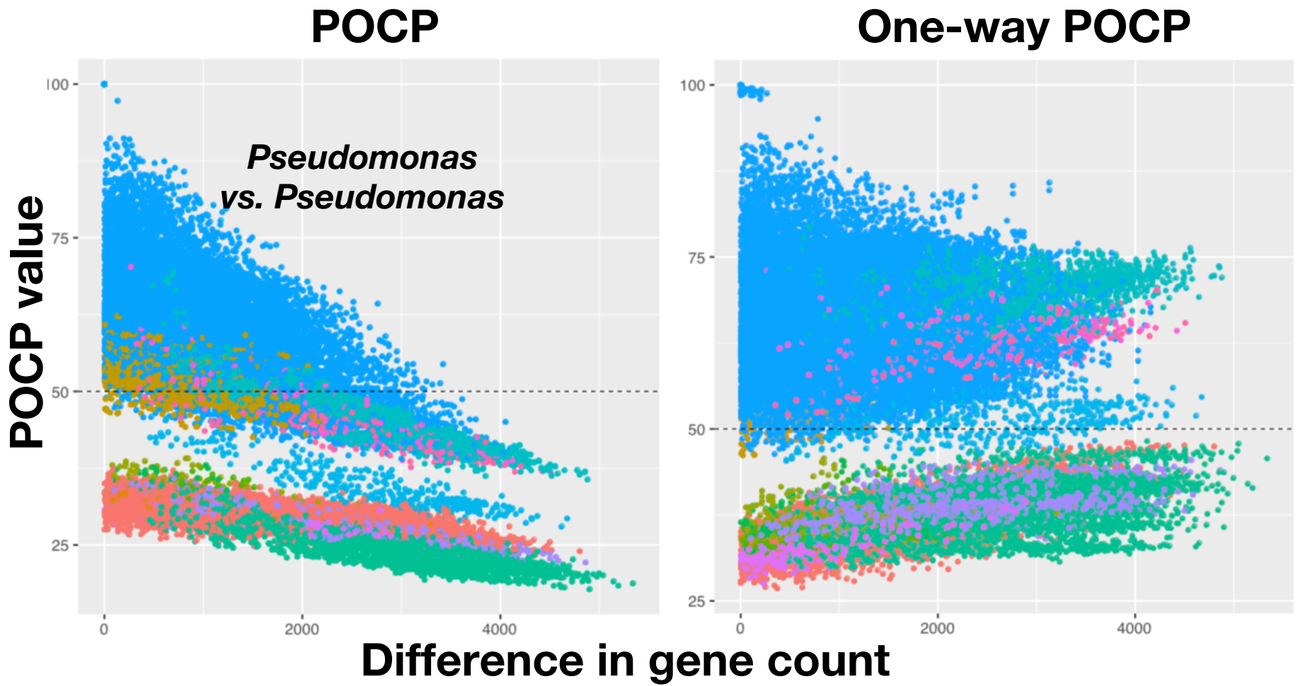
fastANI overestimates ANI

- ANI for 200 genomes takes days on a cluster, exponentially increases with more genomes
- fastANI estimates ANI using k-mer composition and runs in less than an hour



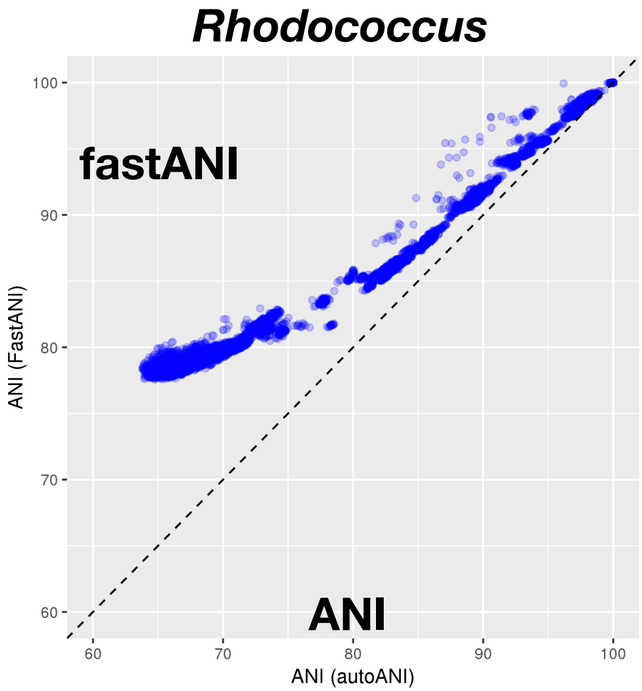
Dramatic difference in genome size skews POCP

- POCP = (A_in_B + B_in_A) /(TotalA + TotalB)
  conservative
- Consider one-way/inclusive POCP = max(A_in_B, B_in_A)/(TotalA + TotalB)
  but tends to overestimate



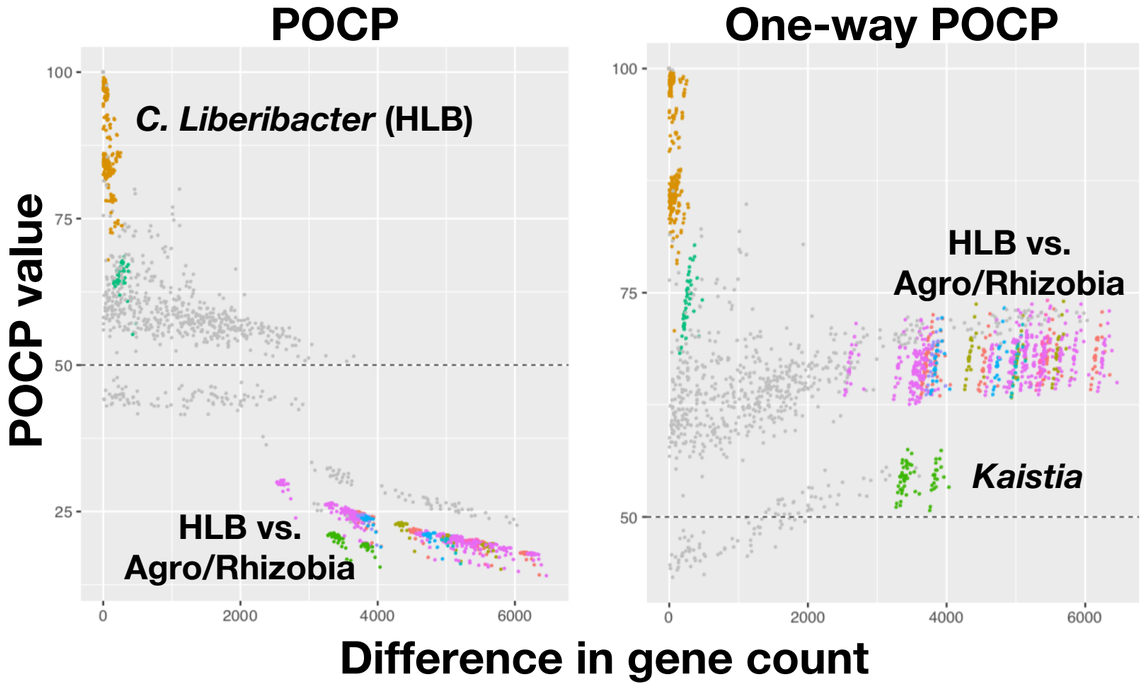
fastANI overestimates ANI

- ANI for 200 genomes takes days on a cluster, exponentially increases with more genomes
- fastANI estimates ANI using k-mer composition and runs in less than an hour



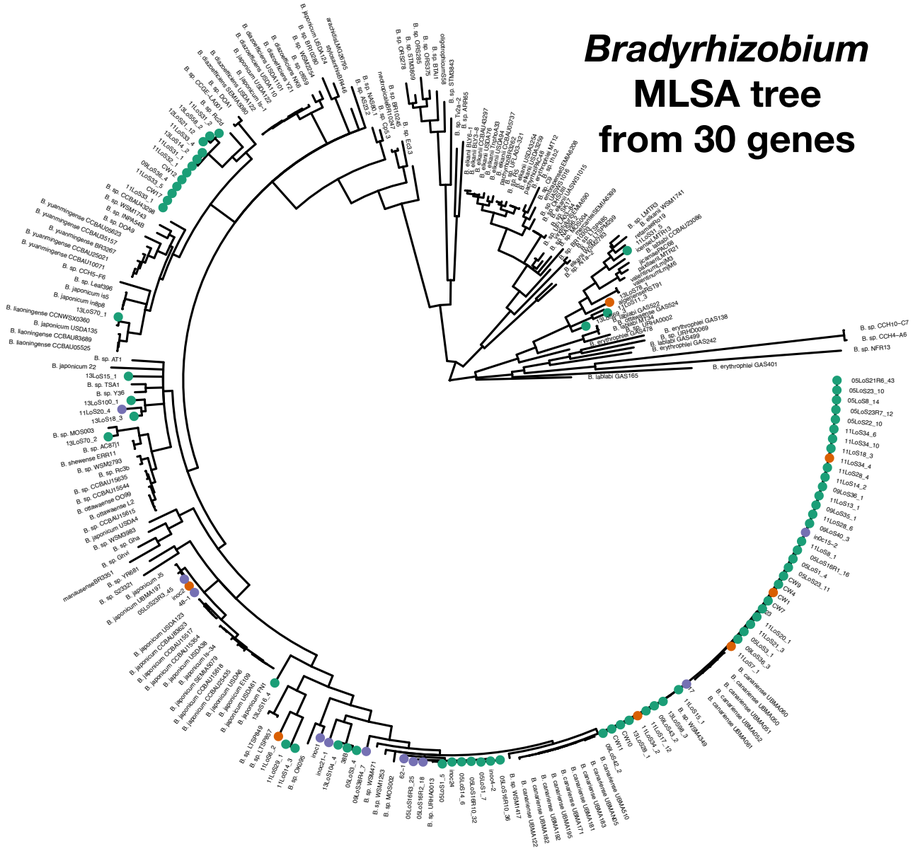
Dramatic difference in genome size skews POCP

- POCP = (A_in_B + B_in_A) /(TotalA + TotalB)
  conservative
- Consider one-way/inclusive POCP = max(A_in_B, B_in_A)/(TotalA + TotalB)
  but tends to overestimate



How are my strains related?

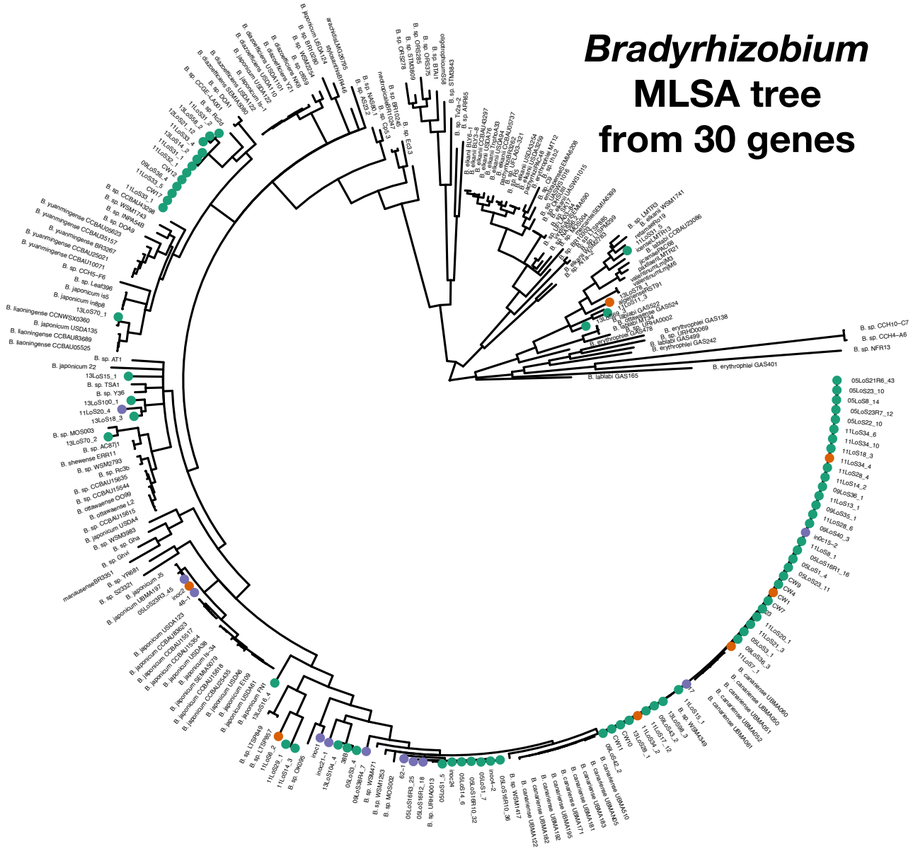
- 16S – conserved in all bacteria, poor resolution
- MLSA - multiple housekeeping genes, informative in a lineage
- Time tree - Neutrally evolving genes + Molecular Clock + BEAST analysis
- Core gene phylogeny – high resolution, requires comparative genomics to identify core genome
- Whole genome SNPs – highest resolution, difficult and highly lineage specific



Multi-locus Sequence Analysis (MLSA)

Phylogeny based on multiple (~5-30) housekeeping genes

- Reference genes chosen on a per-lineage basis
  - Identify gene/protein homologs (**blastn** or **tblastn**/**blastp**)
  - Filter strains that lack one or more genes
  - Multiple sequence alignment per gene (**MAFFT**)
  - Identify best fit evolutionary model per gene (**ModelTest**/**RAxML**)
  - Concatenate alignments and generate alignment partition (R package evobiR superMatrix)
  - Maximum likelihood (ML) phylogenetic analysis (100 searches) and bootstrapping analysis (”autoMRE” criteria or 1000 replicates) (**RAxML**, **RAxML-ng**, **IQtree**)



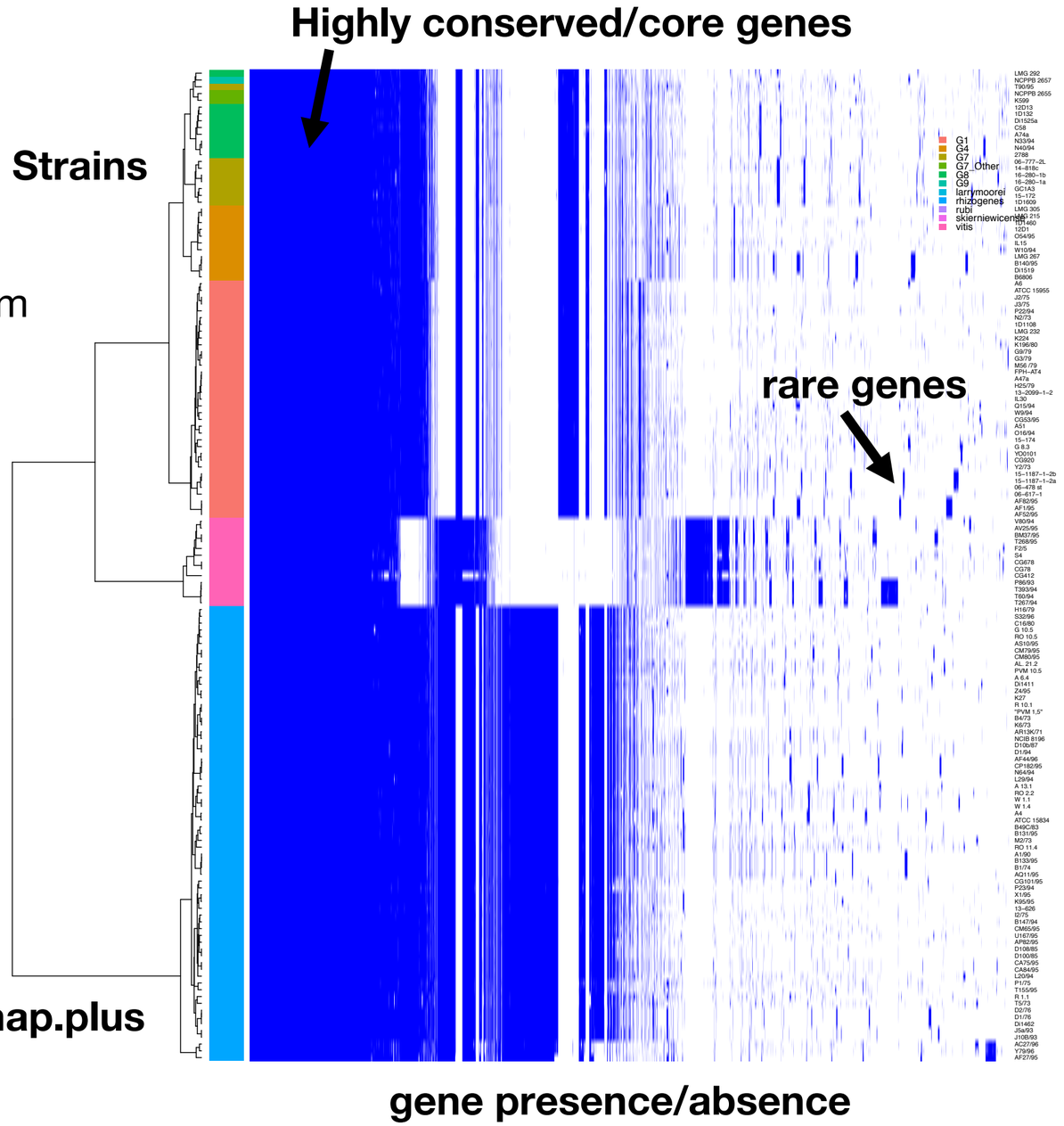
autoMLSA - automates this given a set of reference genes and genomes to search + NCBI (Davis et al., 2016)

How to select housekeeping genes for MLSA

- Reference genes (~5-30) chosen on a per-lineage basis – read the literature!
- Key criteria:
  - Universally conserved
  - Phylogenetically informative
  - Vertically inherited / consistent with whole genome SNP/core gene tree
- Nucleotide vs Protein sequences
  - Protein for deeper splits/more distant relationships – AA seq. more conserved
    - 21 states/position = more information, random chance/“Long branch attraction” less likely
  - Nucleotide for recent divergence/closely-related strains, has diversity

What genes are shared between genomes?

- Identify homologs/orthologs from a set of genomes
- Core genes and larger pan-genome
- Reciprocal BLAST searches + clustering with MCL algorithm
- Genome x gene matrix = which genes are found where
- **get_homologues, ROARY, PIRATE**

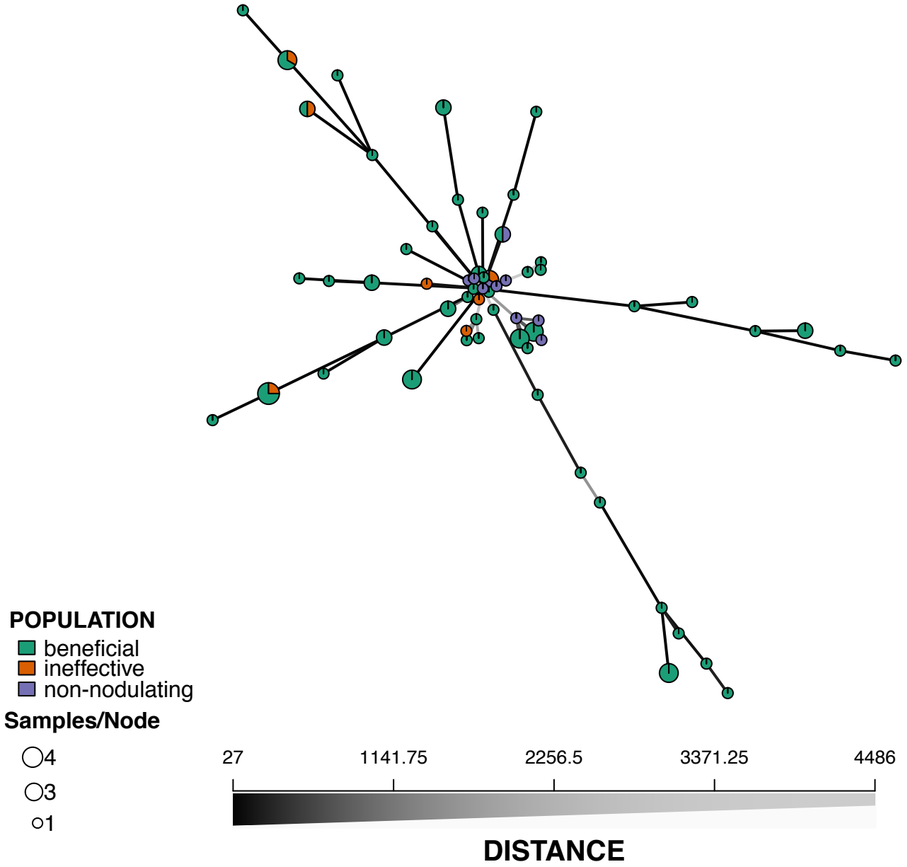


Blue: gene present
White: gene absent

Visualized using R **heatmap.plus**

Whole genome SNPs

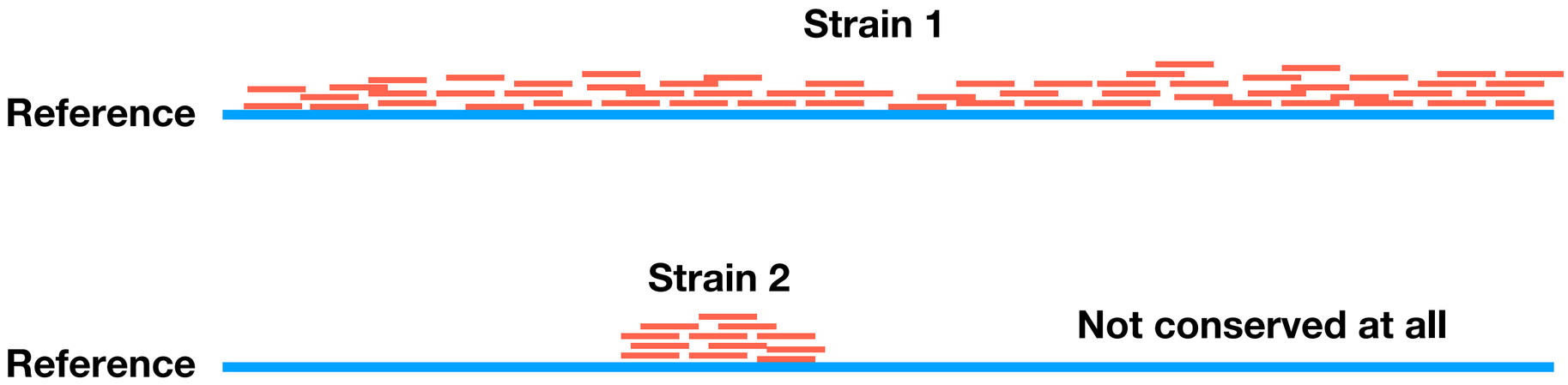
- Single nucleotide polymorphism (SNP) calling
- Much higher resolution than 16S, MLSA
- More time consuming, requires a lot of computational resources
- Many ways to go wrong, requires a lot of verification



Minimum spanning network
of SNP genotypes
(R package **poppr**)

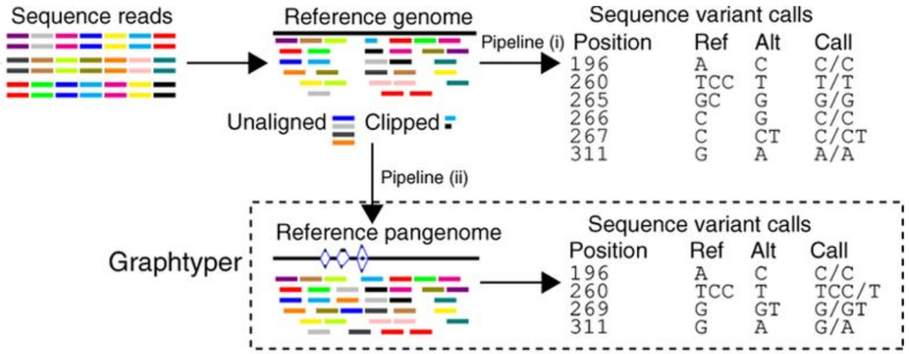
SNP calling caveats - coverage

- Align filtered reads to a reference genome then identify differences (SNPs)
- Only looks at regions that are homologous
- If small % of genome is homologous and few SNPs, erroneously think closely related
- Might see one strain that appears closely-related to everything, verify
- Strains may be similar or differ in regions not included in the reference genome



SNP calling caveats – population structure

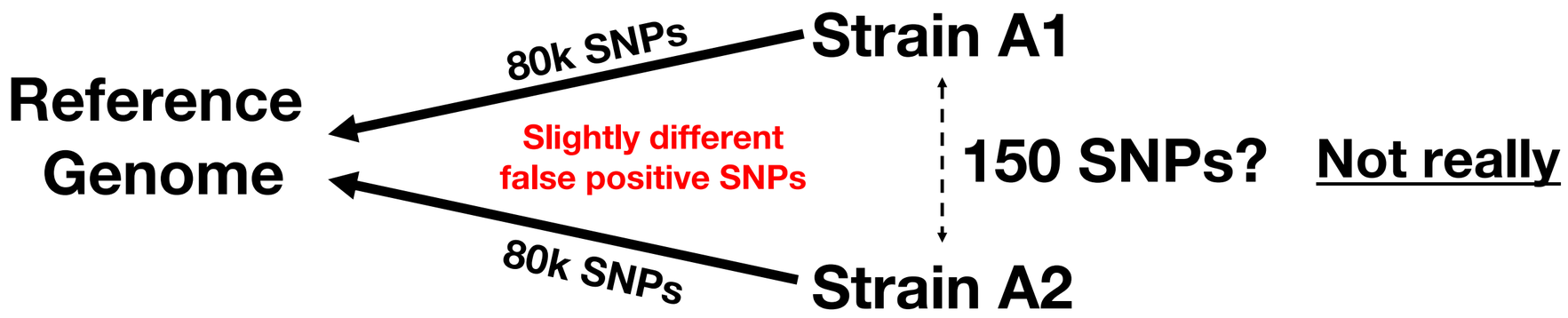
- Choosing good reference genome is **extremely important**
- Distant references = false positive SNPs from poor alignment
- Poor alignment to the reference can skew results (pick a close reference)
- If you have reads for your reference strain, call SNPs for it against itself
- Call SNPs for species level groups (ANI) or even specific lineages (**rhierBAPS**)
- Pangenome-graph based SNP calling can mitigate some of these issues



Graphtyper (Eggertsson et al., 2017)

Example SNP calling issue

- SNPs called using a single reference genome for known lineage
- Lineage contains 4 ANI species groups
- We know these are clones however they differed by 150-200 SNPs?
- SNP calling with a closer reference or a pangenome graph method both resulted in zero pairwise SNP differences



A reasonable SNP pipeline (based on GATK best practices)

Align trimmed (**BBduk**) reads with **BWA mem**

Picard tools to sort alignments (**SortSam**) and **MarkDuplicates**

GATK **HaplotypeCaller** on individual samples

GATK **GenotypeGVCFs** to call SNPs on all samples together

Filter SNPs with **vcffilter** (see GATK docs for recs., its complicated)

Graphtyper for graph-based SNP calling

No haploid option yet, filter out heterozygous SNPs

Filter SNPs with **vcffilter**

pairwise SNP-distance table (custom script, see github)

Genotypes, minimum spanning network (R package **poppr**)

ML Phylogeny (**RAxML** with ascertainment bias correction)

Summary

- ANI, MLSA, WGS SNPs = only compares regions that are shared
- POCP, gene P/A, synteny maps = only shows variation in presence, not relatedness of shared regions
- Be aware of the level of resolution, appropriateness, and potential pitfalls of each method
- **No single method is sufficient to explain all variation**