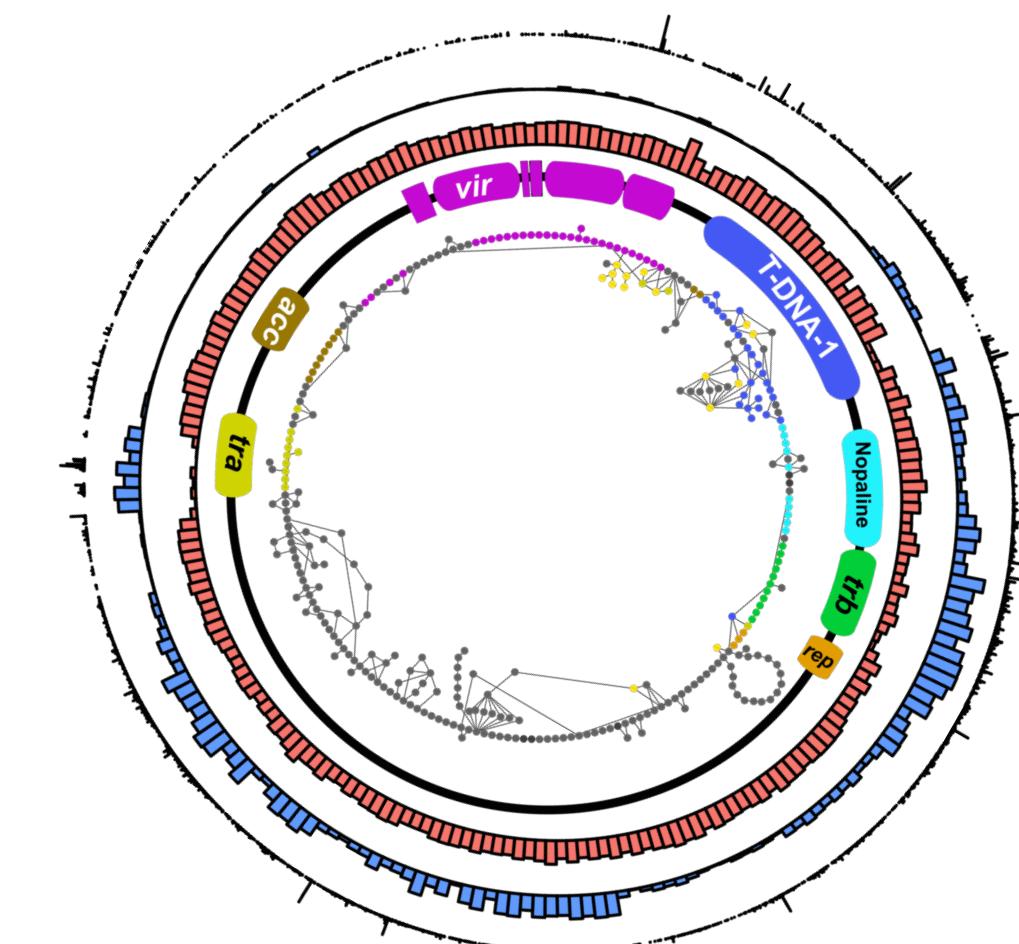


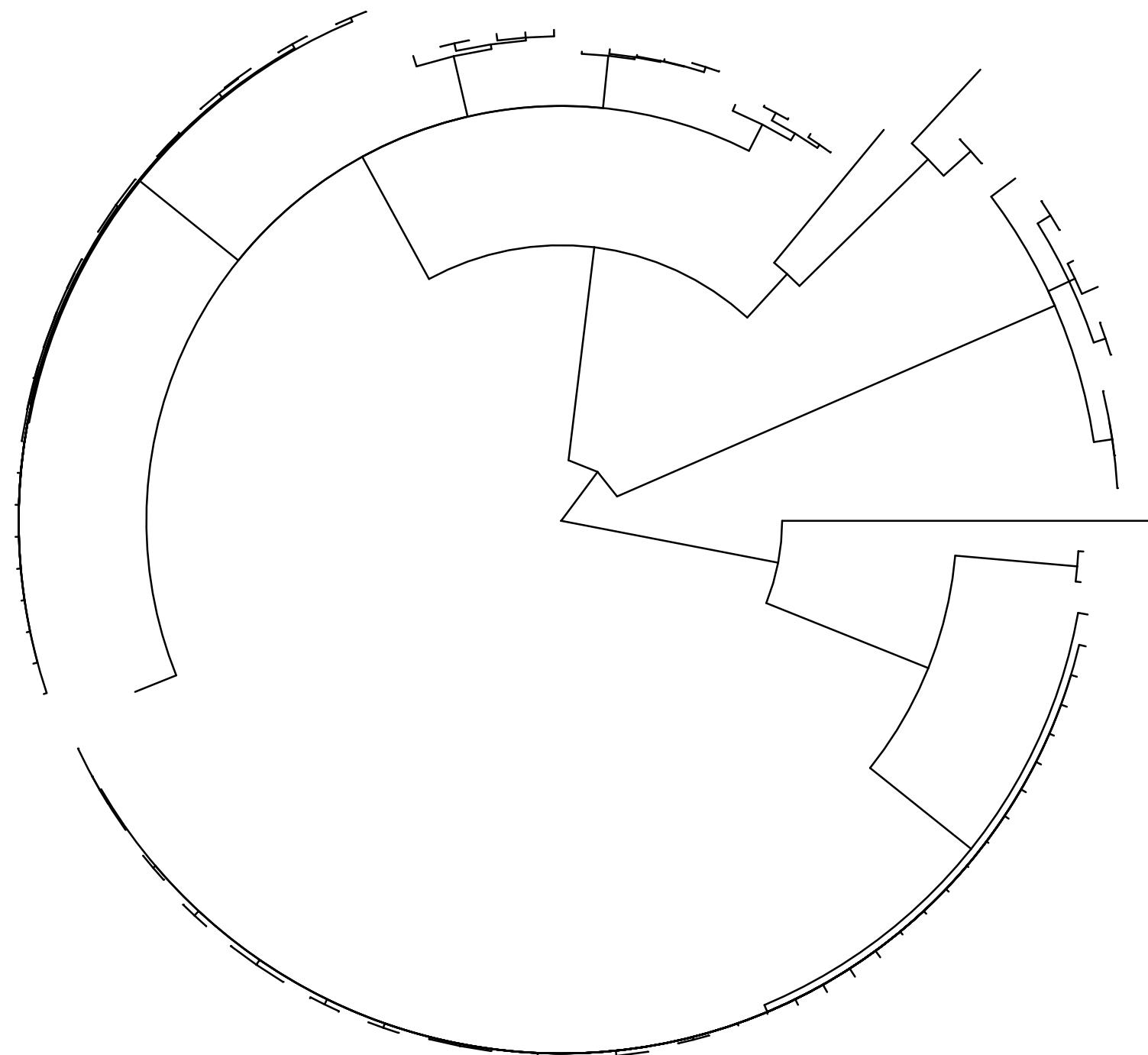
Publication-quality Figures in R

How to make consistent, programmed
figures in R

Alexandra J. Weisberg,
Chang Lab, Oregon State University
github.com/osuchanglab/VTworkshop



Visualizing and annotating phylogenies



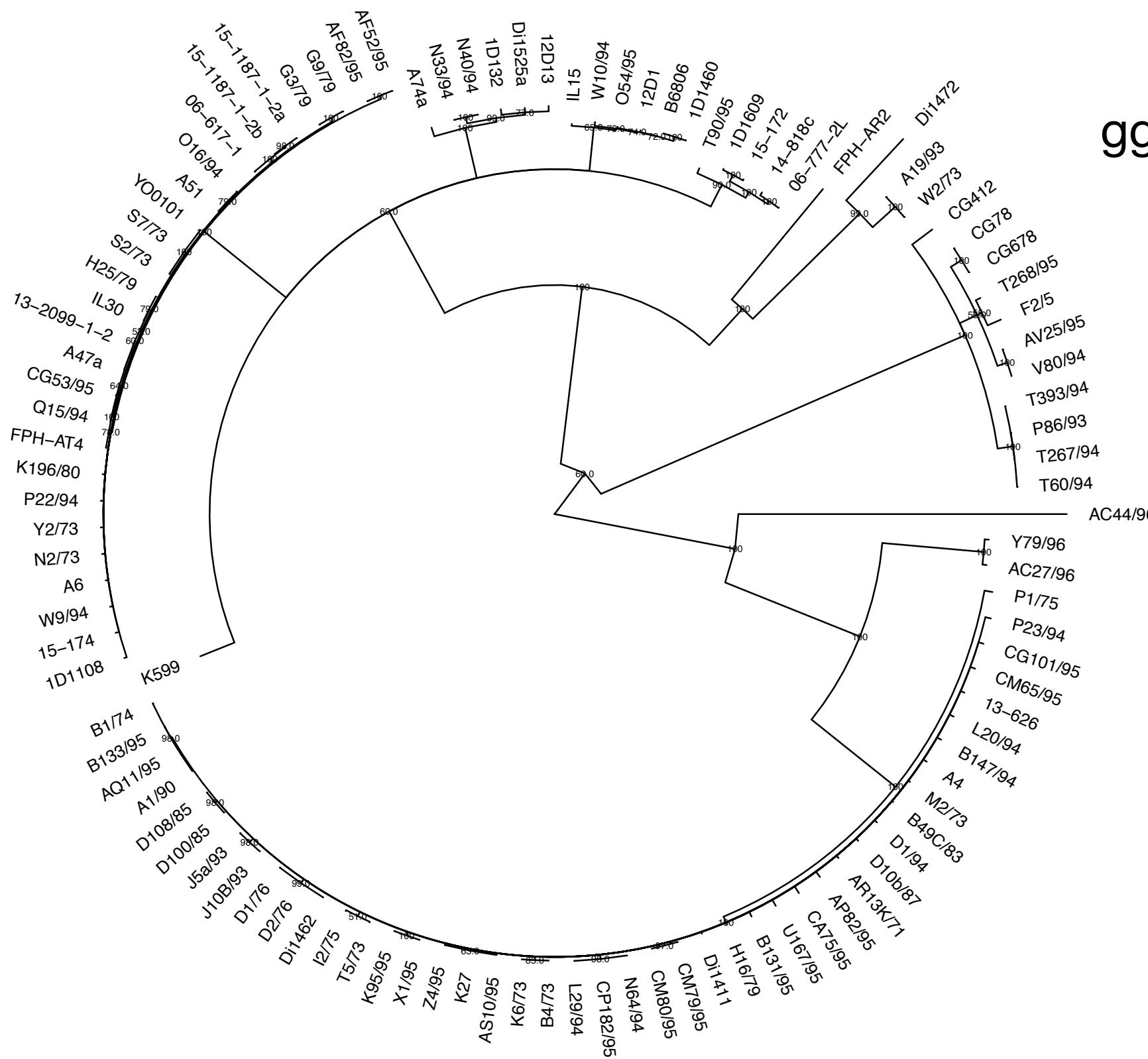
Visualized using R package ggtree:

```
library(ggtree)
```

```
mytree <- read.tree("treefile.tre")
```

```
ggtree(mytree, layout="circular")
```

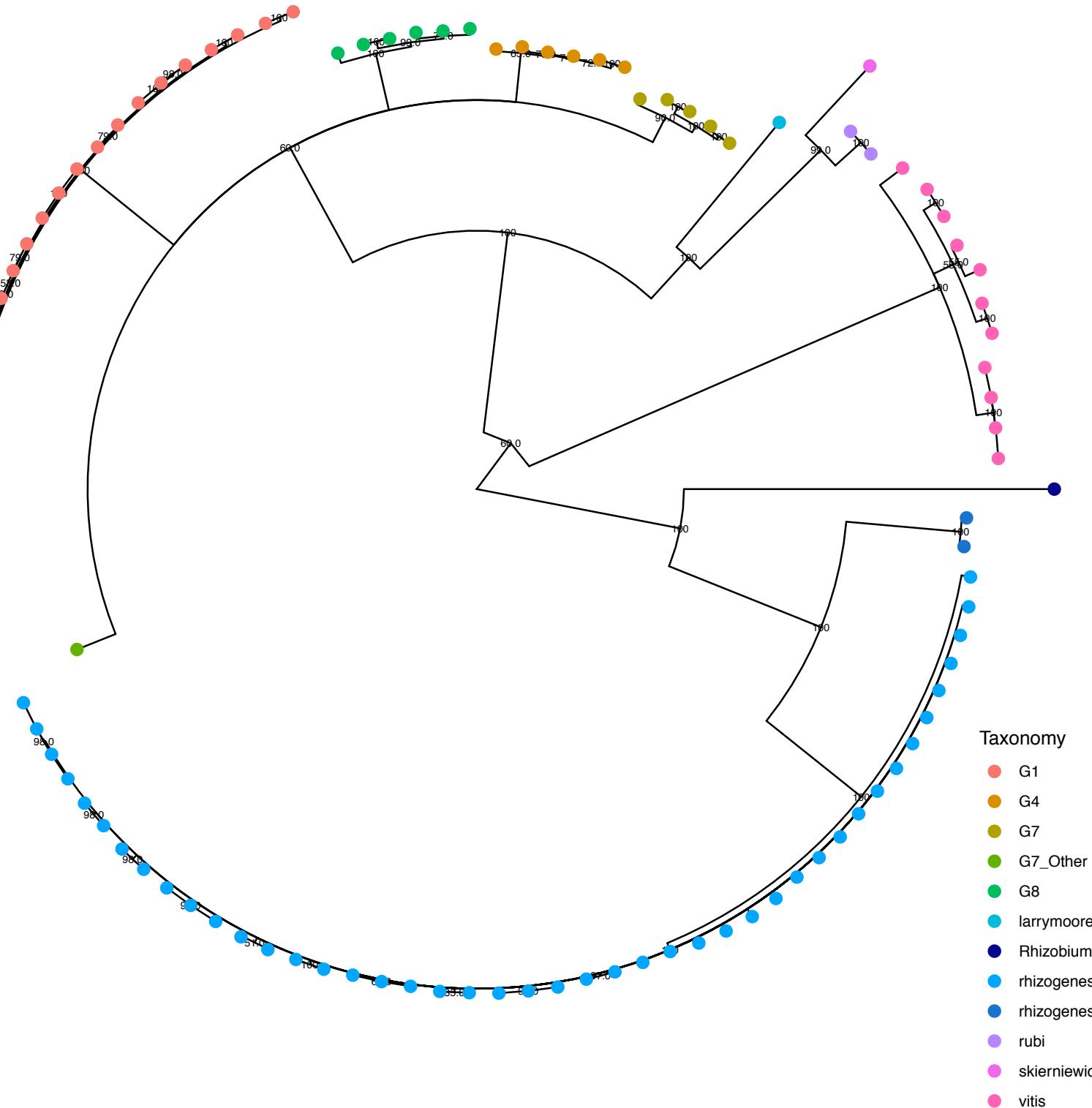
Visualizing and annotating phylogenies



```
ggtree(mytree, layout="circular") + geom_tiplab()
```

Visualizing and annotating phylogenies

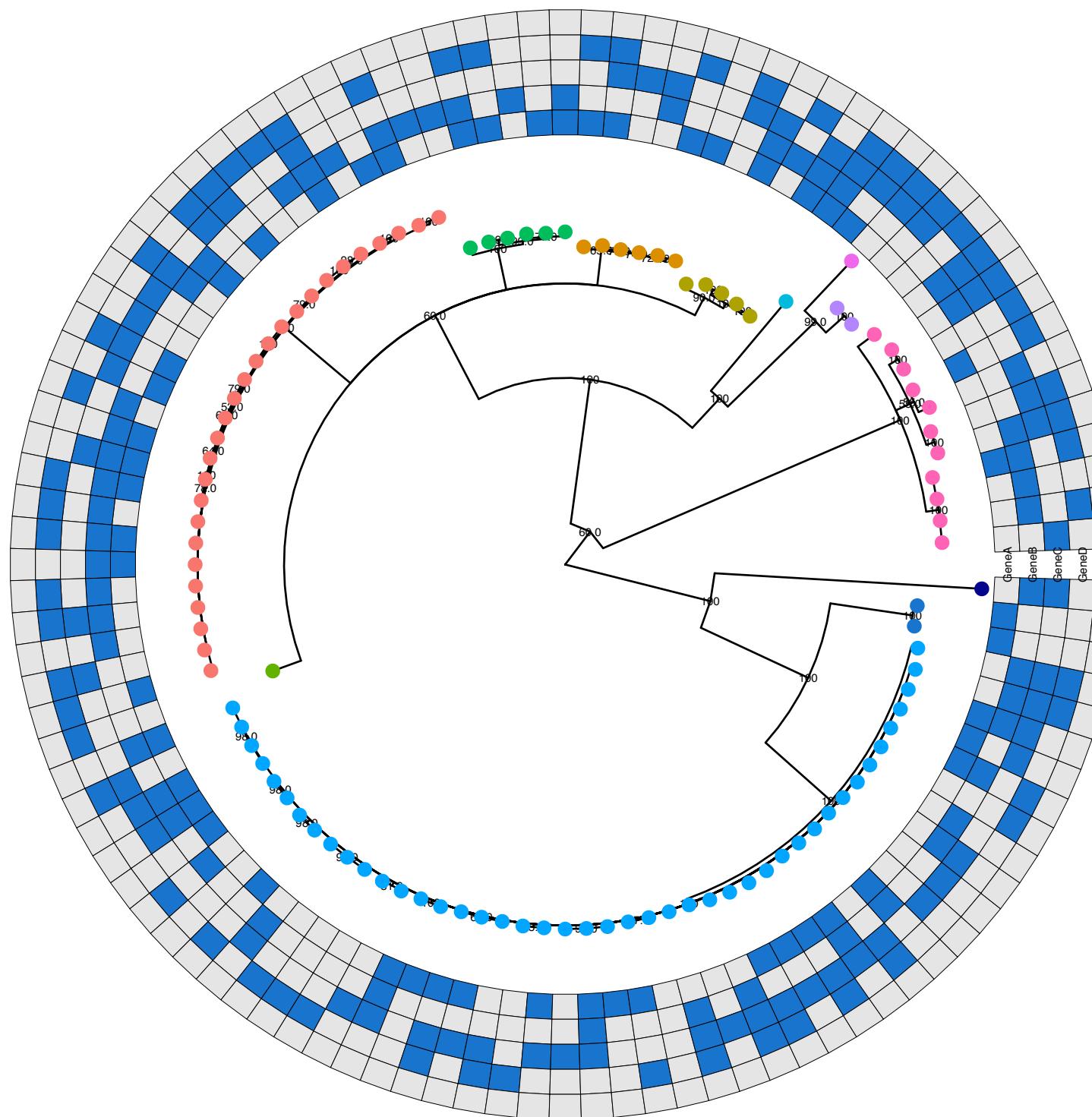
`ggtree(mytree, layout="circular") + geom_tippoint(aes(color=Species))`



Color by a phenotype

A legend is generated automatically

Visualizing and annotating phylogenies



```
p <- ggtree(mytree, layout="circular") +  
  geom_tipoint(aes(color=Species))
```

```
gheatmap(p, myphenotype_table)
```

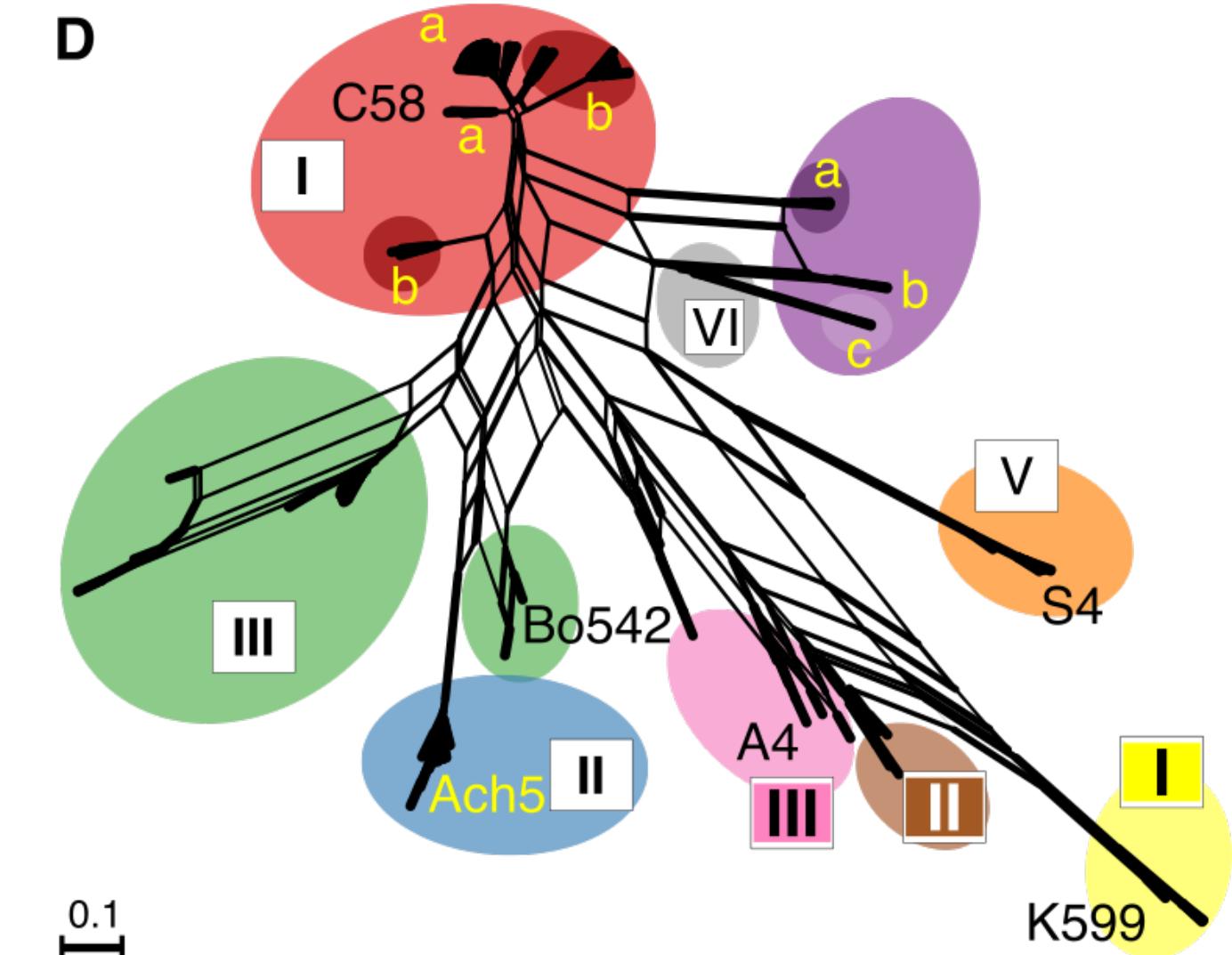


Code for this example on GitHub:
github.com/osuchanglab/VTworkshop

See **ggtree vignettes** for details and more examples:
<http://www.bioconductor.org/packages/3.1/bioc/vignettes/ggtree/inst/doc/ggtree.html>

Phylogenetic networks

- Network from set of gene alignments
NeighborNet
- Network from set of trees **SuperQ**
- Visualize with R **phangorn** package,
SplitsTree



Network of 142
Agrobacterium Ti plasmids
based on 125 conserved gene trees

Visualizing structural variation

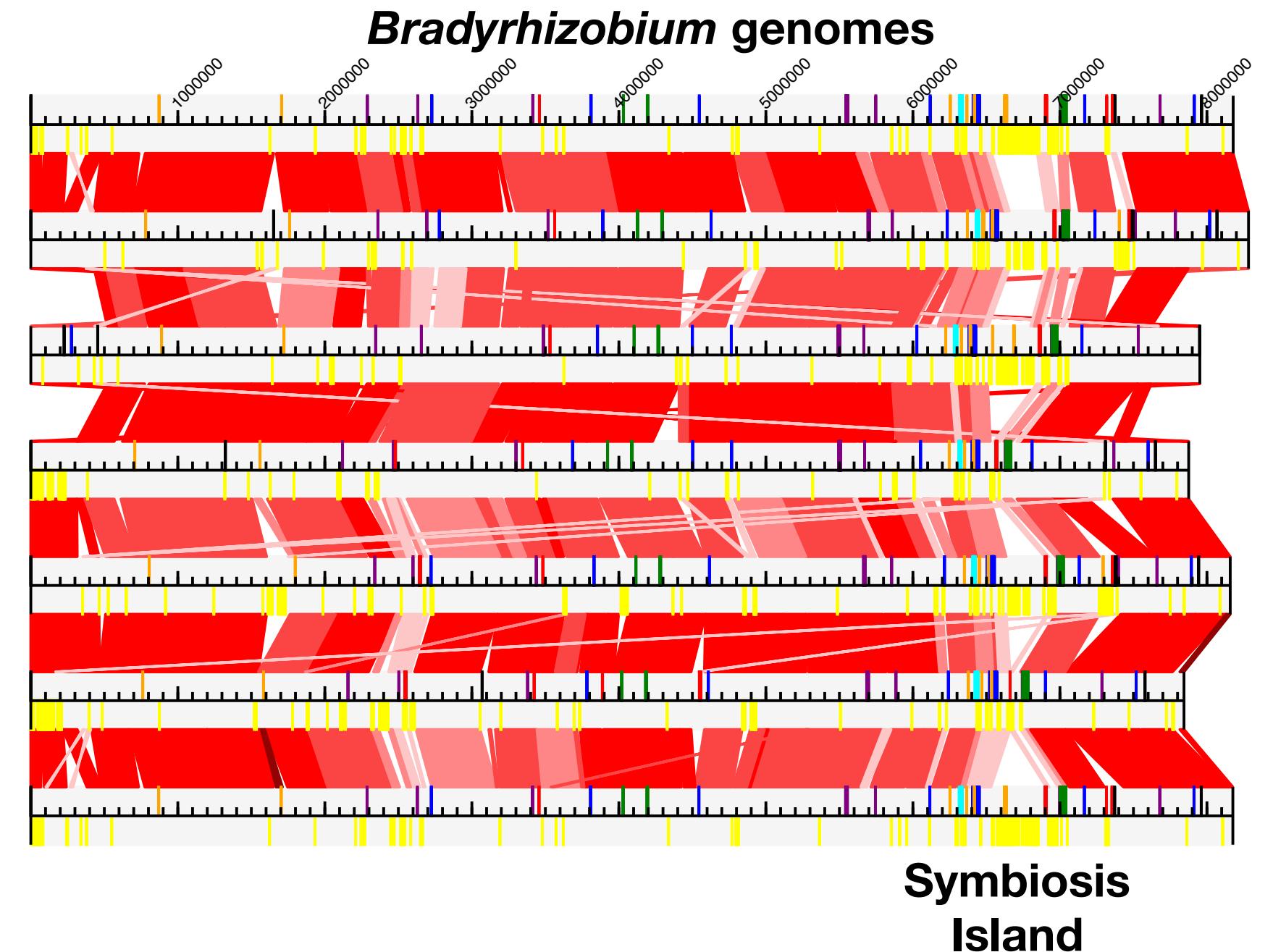
Given a set of genomes/plasmids/regions of interest:

- What gene diversity is present, and where is it?
- Where are hotspots of recombination/HGT?
- How do genomes differ in non-shared regions?

Zero SNP differences: Clones! ...maybe, check POCP/gene content

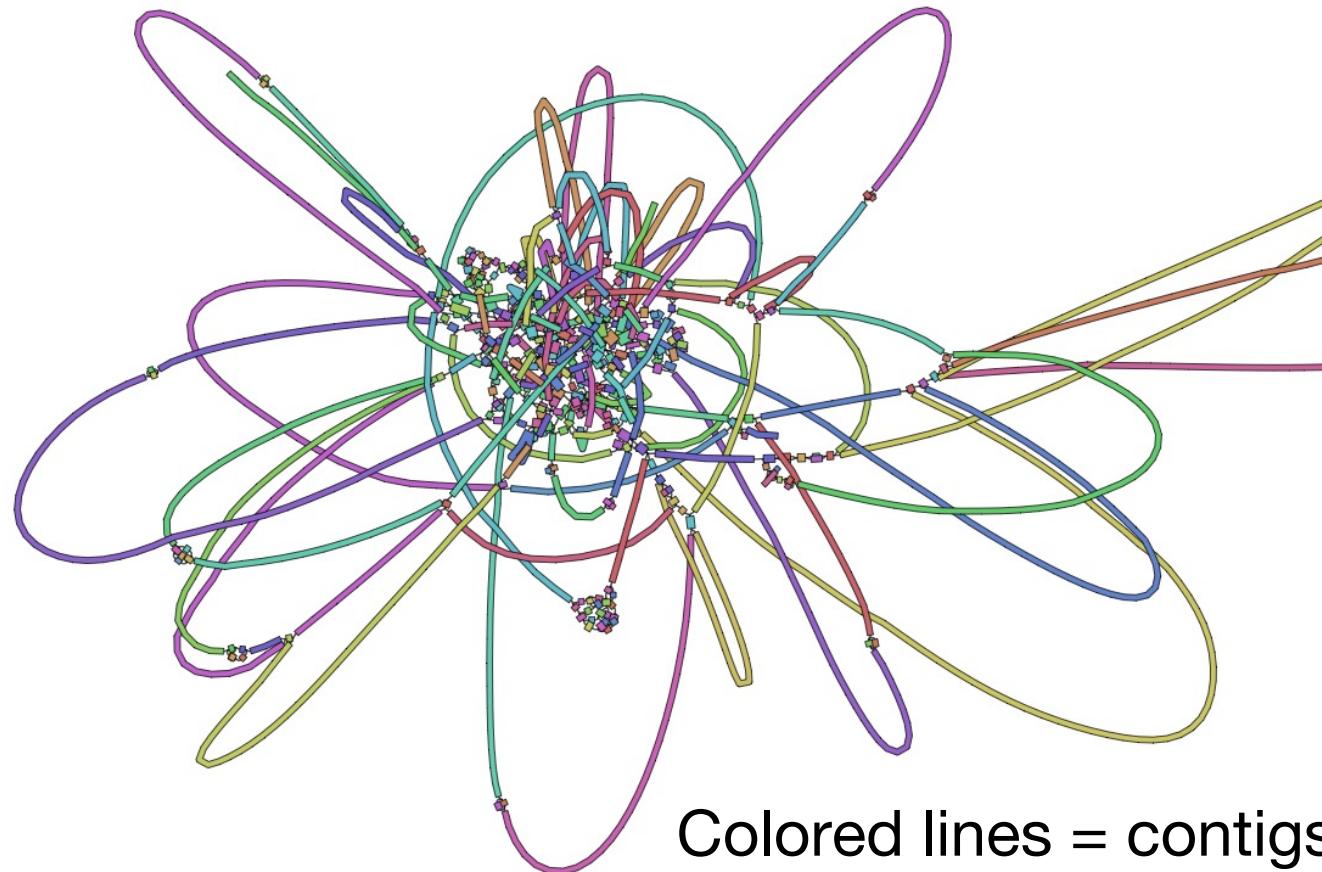
Visualizing structural variation

- Identify shared regions
 - **minimap2** and visualize with BioPython **GenomeDiagram**
 - **Mauve**
 - Hybrid Illumina/Nanopore assemblies provide whole genome structure



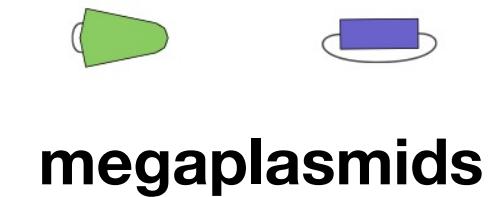
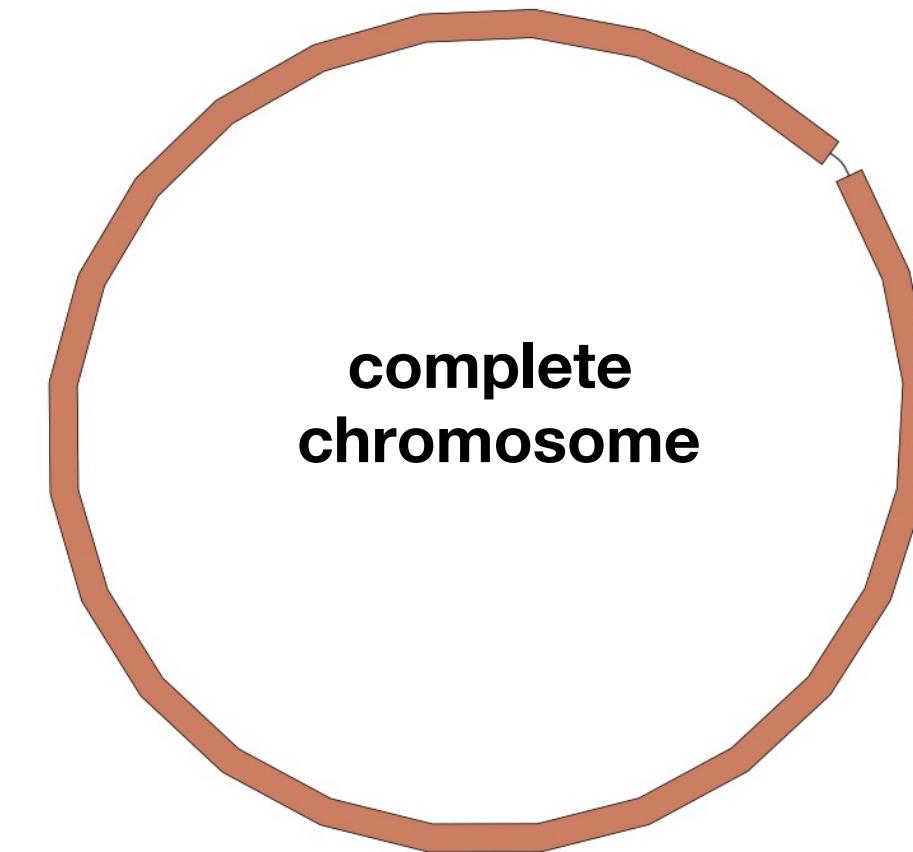
Visualizing *de novo* assemblies

Illumina-only assembly



(assembled using **SPAdes**)

Hybrid Illumina/Nanopore assembly



(assembled using **unicycler**)

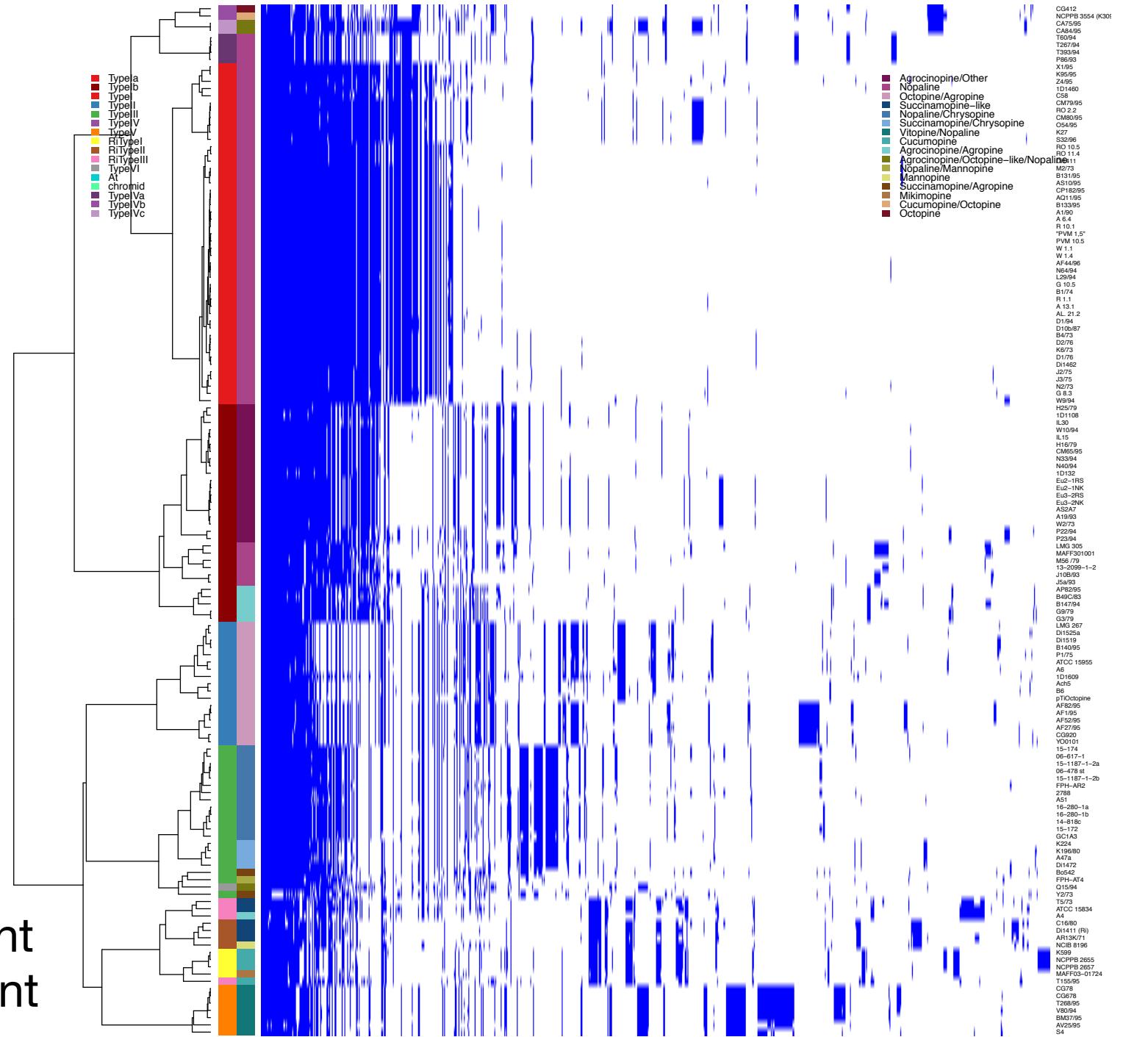
Bandage – view assembly graph (.gfa output file)

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

What gene diversity is present and where is it?

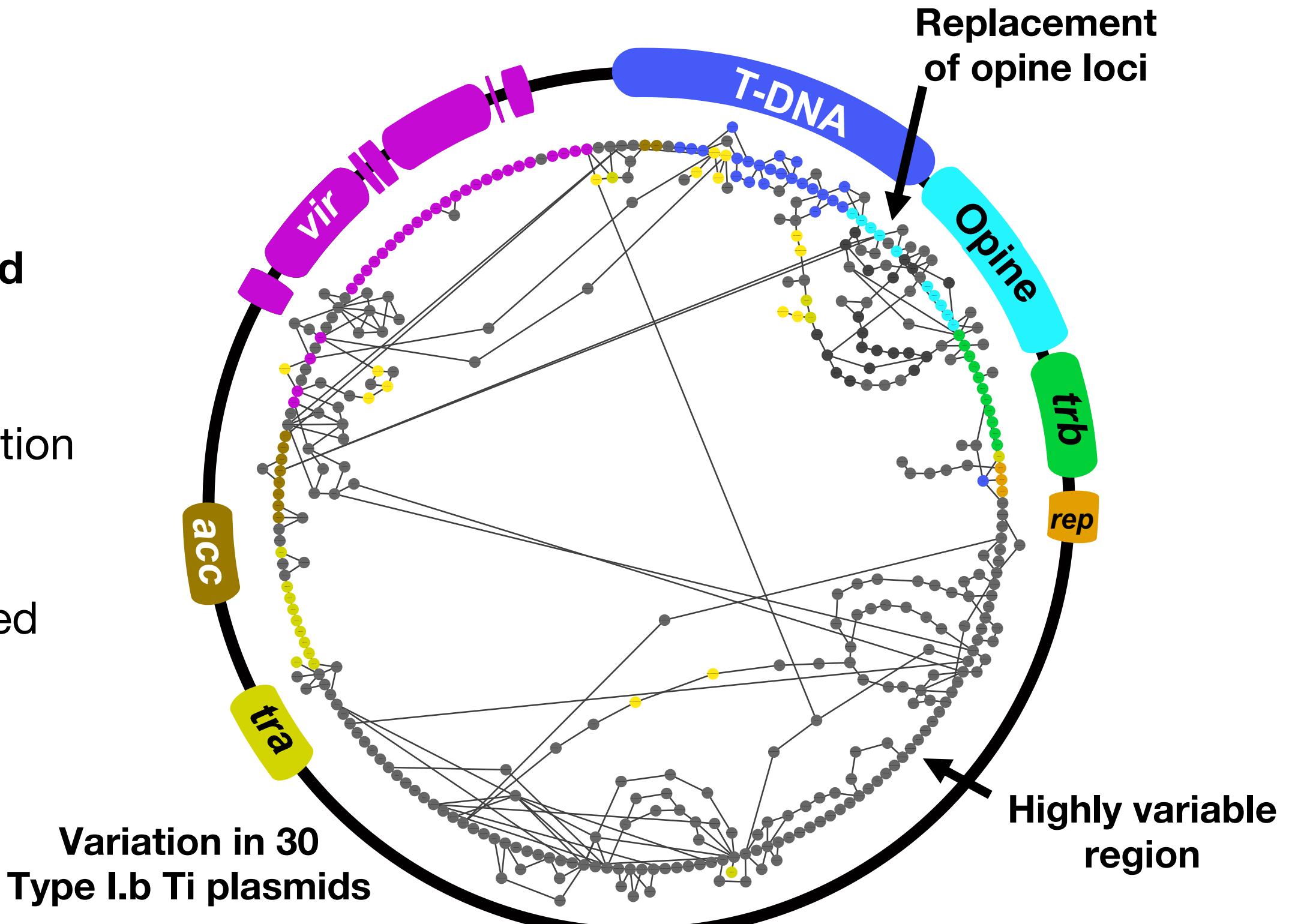
- Visualize gene presence/absence variation
- *Agrobacterium* Ti plasmid gene content
- Heatmap shows that variation exists, but not where, how they are related
- No structural context

Blue: gene present
White: gene absent



Gene synteny graphs

- What gene diversity is present **and** where is it?
- Visualize gene P/A and HGT variation with a gene synteny graph
- Nodes = genes, connections based on synteny in at least 1 strain



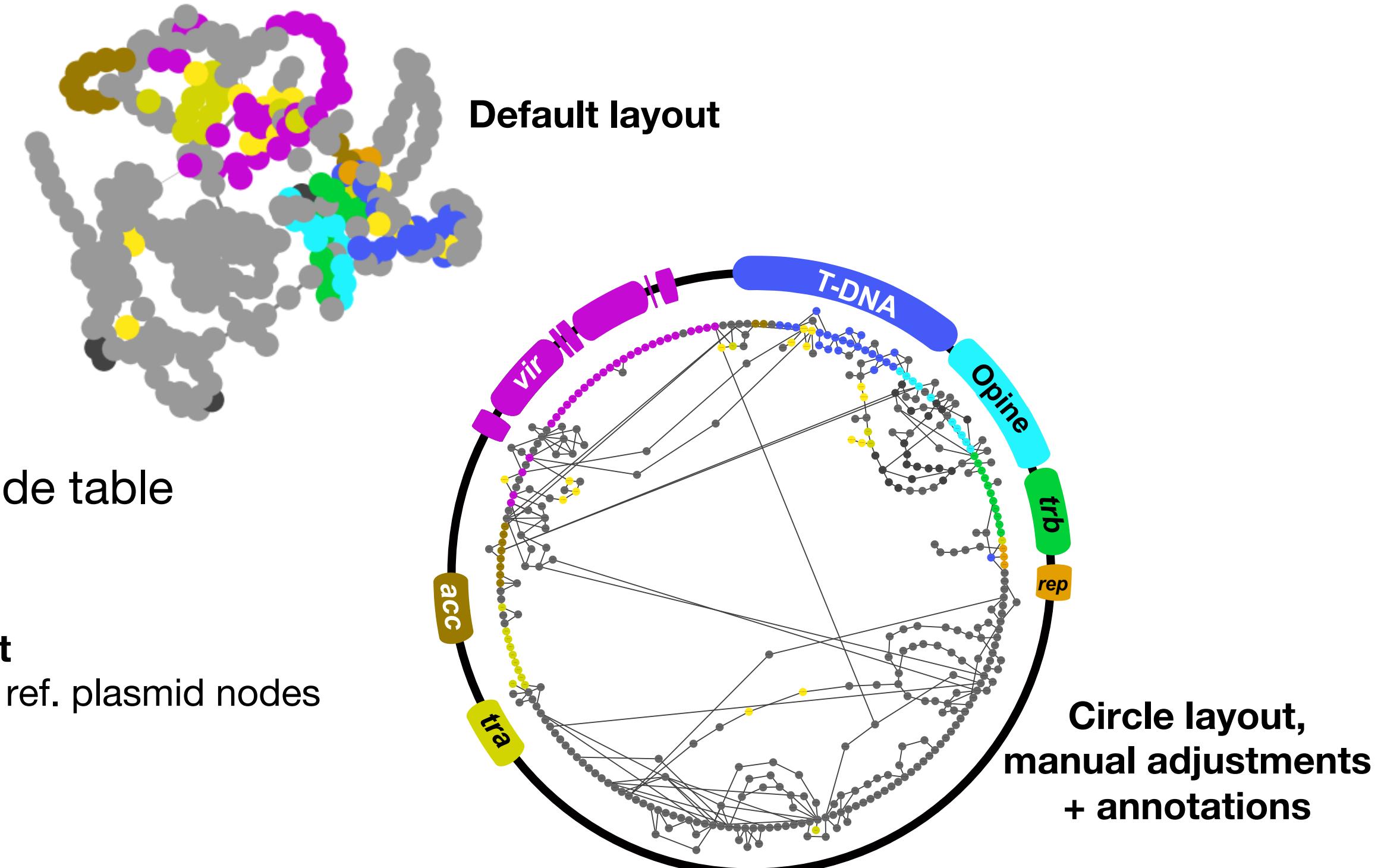
Gene synteny graphs

- For each sample (i.e. plasmid):
 - For each gene :
 - Find the ortholog cluster containing that gene **GeneB = OrthologB**
 - Print pairs of syntenic genes as their ortholog clusters “OrthologA” “OrthologB”
 “OrthologB” “OrthologC”
 - If the replicon is circular, print the last cluster connected to the first “OrthologZ” “OrthologA”
 “OrthologA” “OrthologQ”
Variation results in alternate paths “OrthologA” “OrthologQ”

Gene synteny graphs

Output (sif format):

OrthologA	OrthologB
OrthologB	OrthologC
OrthologC	OrthologD
OrthologD	... etc.

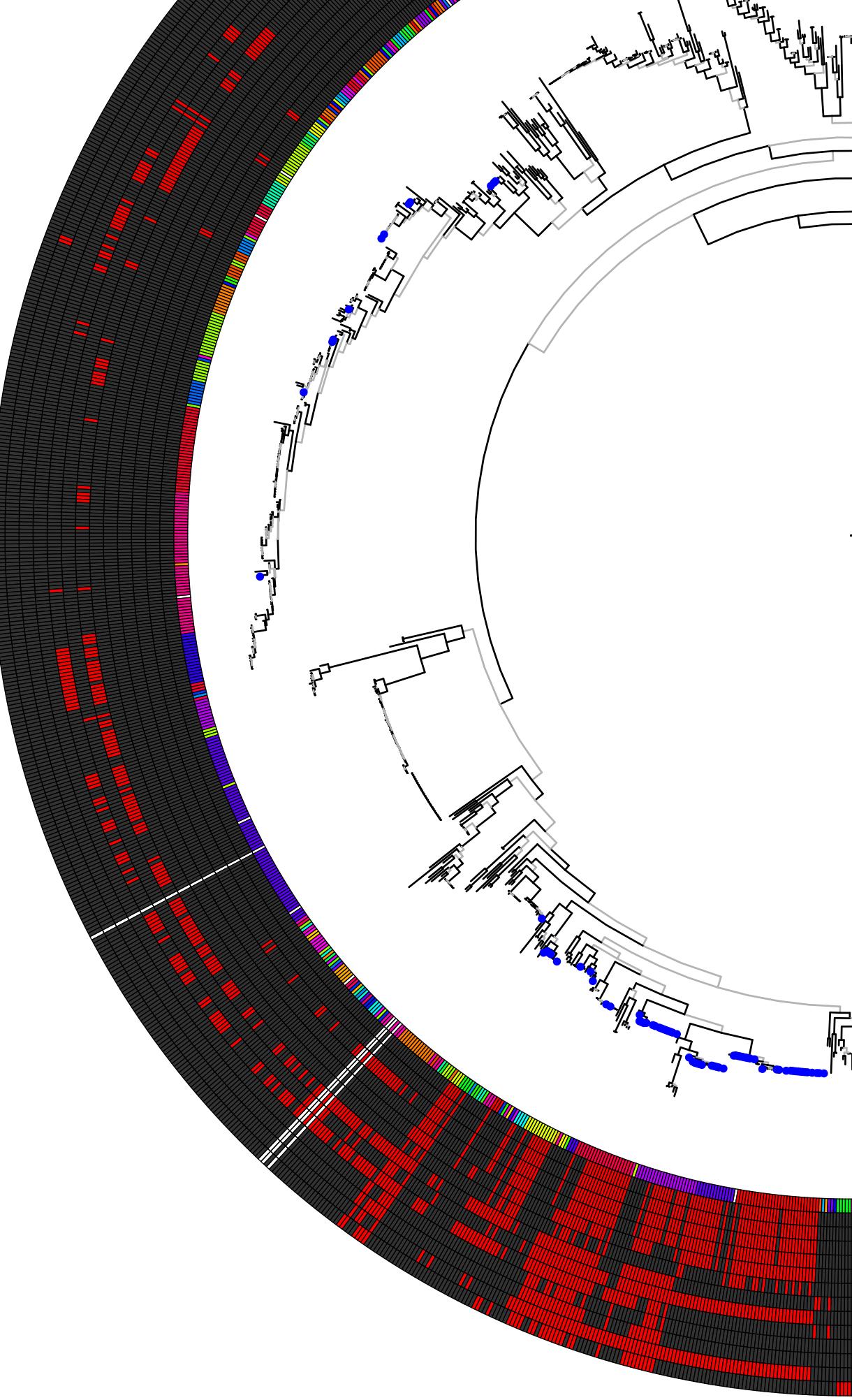


***Step-by-step instructions for example data on github**

github.com/osuchanglab/VTworkshop

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**



Acknowledgements

Jeff Chang Lab
Nik Grunwald Lab

Department of Botany and Plant Pathology, Oregon State University

USDA ARS, Oregon State University

Funding
USDA NIFA Postdoctoral Fellowship
#2017-67012-26126



Twitter: @weisbergaj



github.com/osuchanglab/VTworkshop

For slides and example data