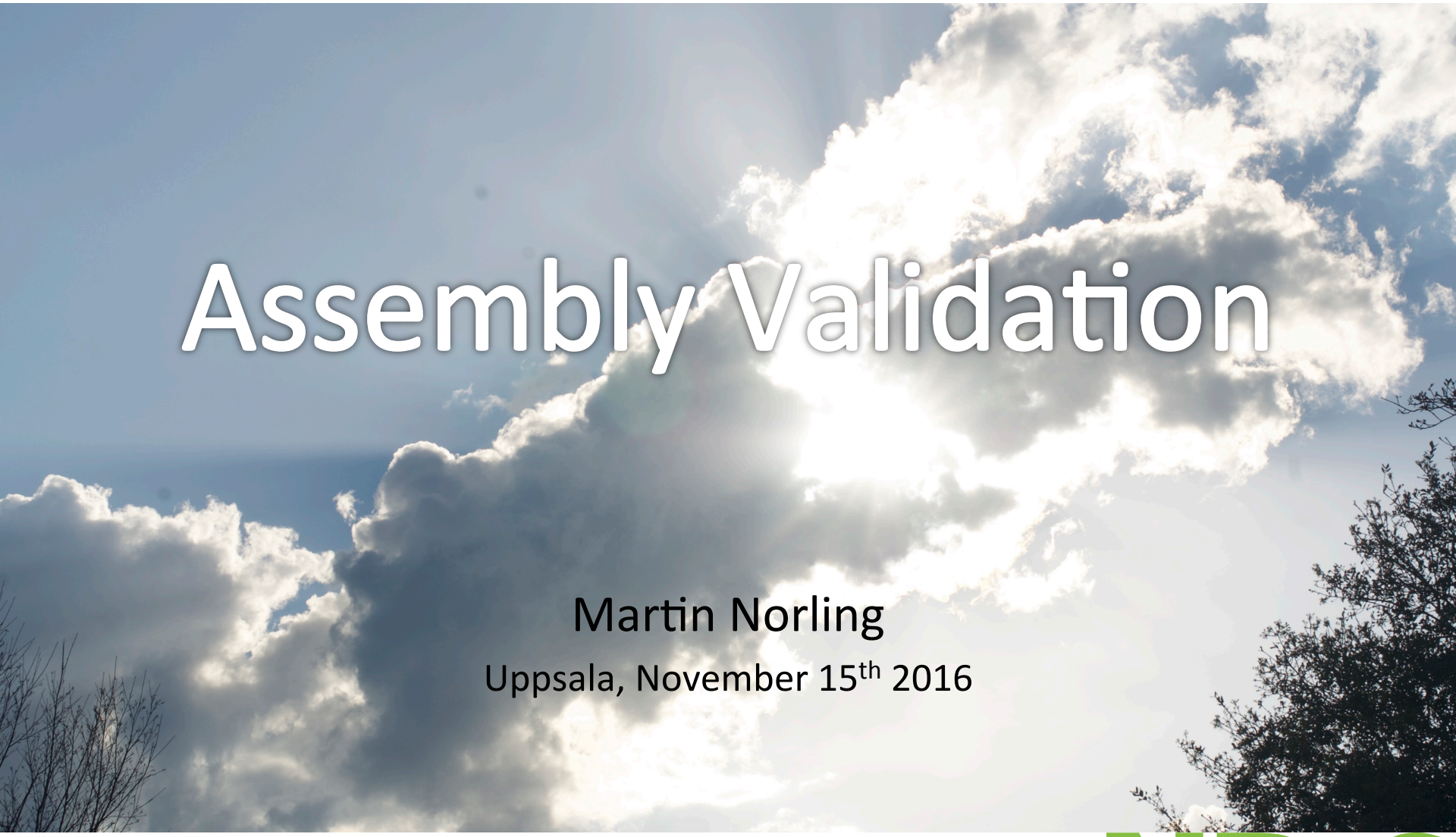
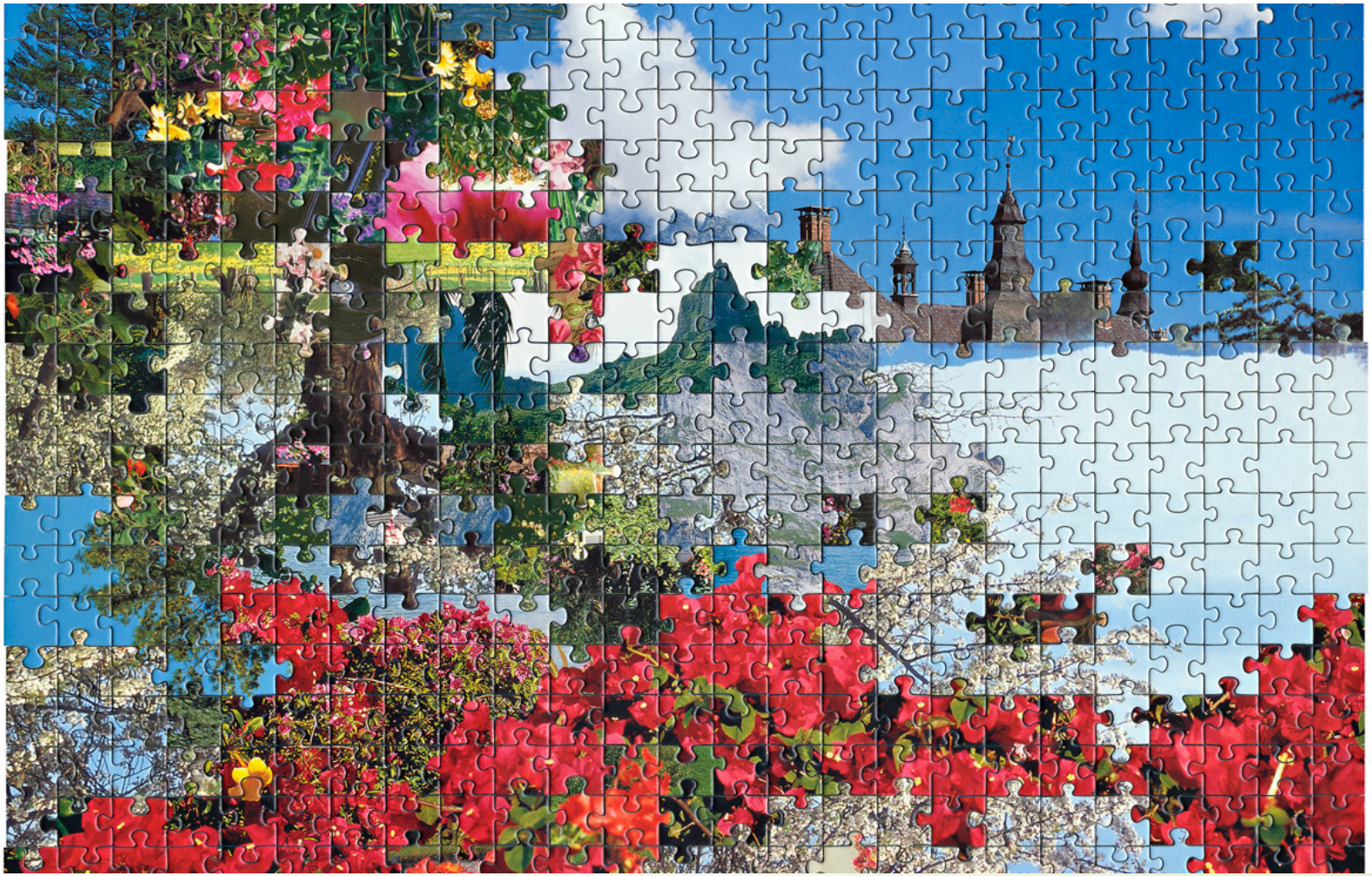


# Assembly Validation

Martin Norling

Uppsala, November 15th 2016

# Does the assembly make sense



# Evaluating Gene Space

Finding genes and other genomic regions is generally the domain of annotation, not assembly – we can cheat a bit though!

Making a rough estimate of the gene content compared to what is expected from a complete assembly can be a hint at the state of the assembly (at least of the gene space).

# CEGMA

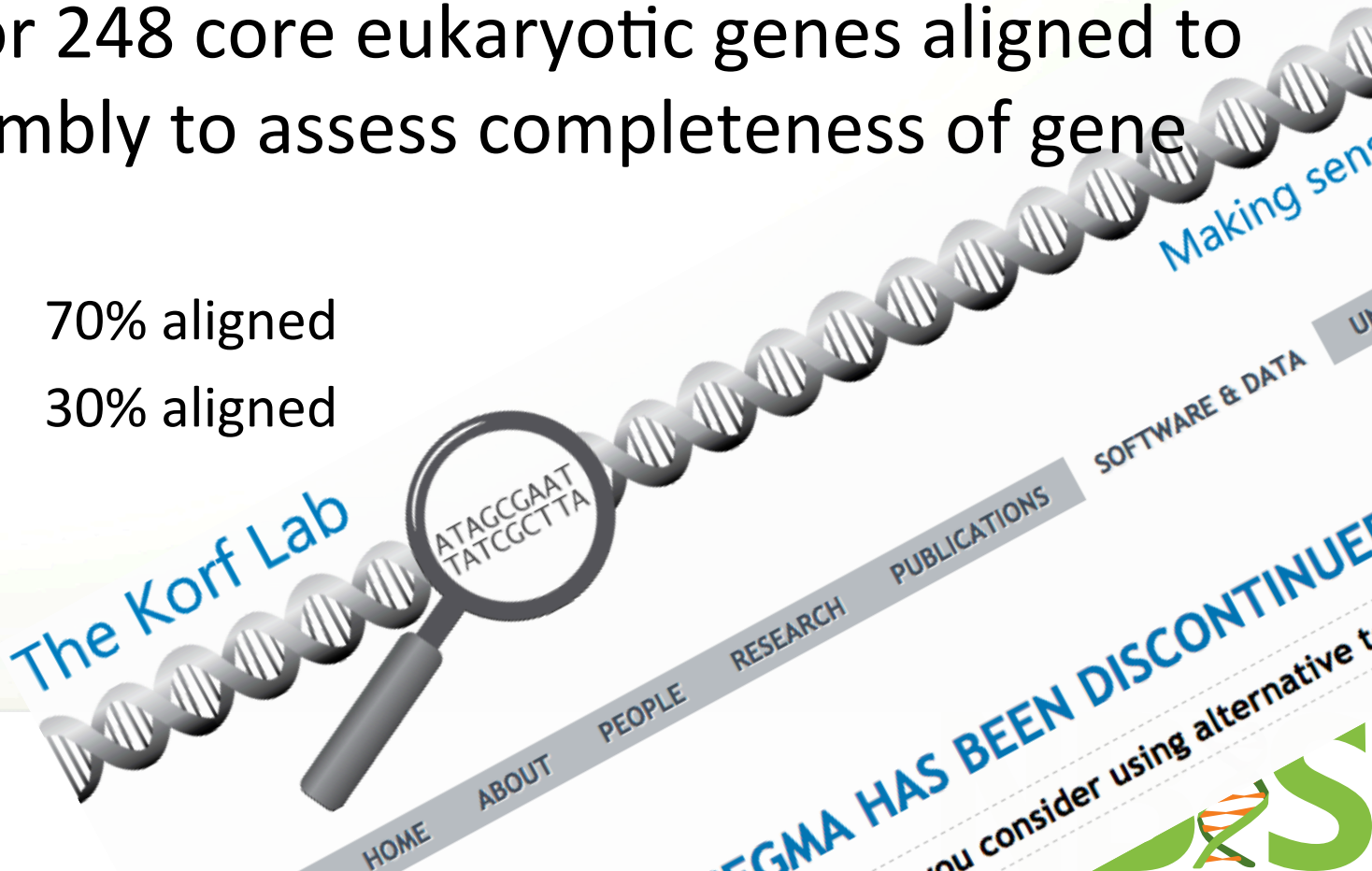
**CEGMA** (Core Eukaryotic Genes Mapping Approach)

(http://korflab.ucdavis.edu/datasets/cegma/)

HMM:s for 248 core eukaryotic genes aligned to your assembly to assess completeness of gene space

"complete": 70% aligned

"partial": 30% aligned



# BUSCO

**BUSCO**(http://busco.ezlab.org/)

Assessing genome assembly and annotation completeness with Benchmarking Universal Single-Copy Orthologs

**Similar idea based on aa or nt alignments of**

- Golden standard genes from own species
- Transcriptome assembly
- Reference species protein set

Use e.g. GSNAP/BLAT (nt), exonerate/SCIPIO (aa)

# BUSCO

Provides quantitative assessment of assembly completeness based on evolutionarily informed expectations of gene content from near-universal single-copy orthologs.

```
$ cat short_summary_spades
#Summarized BUSCO benchmarking for file: spades.fasta
#BUSCO was run in mode: genome

Summarized benchmarks in BUSCO notation:
          C:97%[D:2.5%],F:0.0%,M:2.5%,n:40

          39        Complete BUSCOs
          38        Complete and single-copy BUSCOs
          1         Complete and duplicated BUSCOs
          0         Fragmented BUSCOs
          1         Missing BUSCOs
          40        Total BUSCO groups searched
```

# Comparing to close relatives

If available, another way of assessing the completeness and sanity of an assembly is to compare your assembly to a published reference.

# Synteny

Synteny is the conservation of gene order between species – that can sometimes be quite distant.

Evaluating synteny can be a way of looking for possible misassemblies, or a way to detect genomic rearrangements.

# … if there is synteny

Then again – some phyla, such as *Apicomplexa*, have very low synteny.

This can of course be interesting, but not very useful for validation…

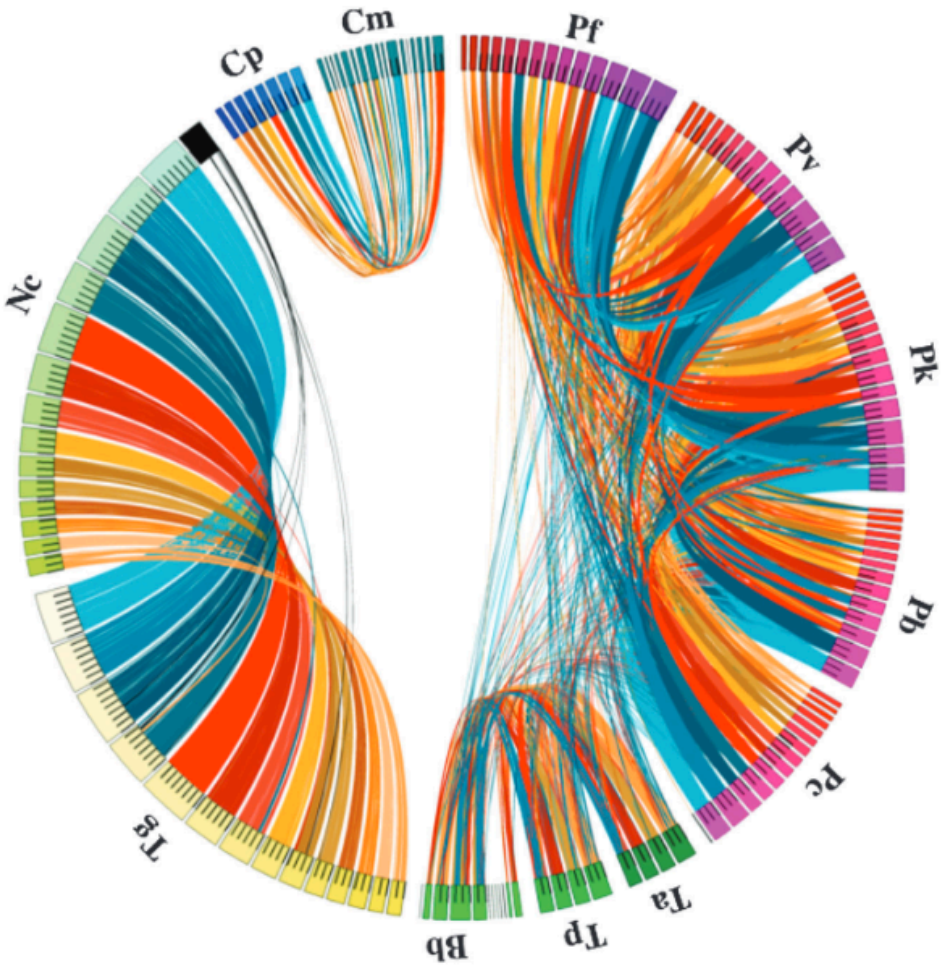



FIG. 2. Detected synteny across the Apicomplexa. The circle is a graphical representation of the annotated chromosomes and contigs in each genome. Each species' genome is labeled with the genus species abbreviation. Scaffolds/Contigs that are not assigned to chromosomes but contain syntenic regions are shown in black. Tick marks represent 1 Mb. Lines that span the interior of the circle connect syntenic regions as detected by MCSCAN. "Twisted" spans represent inversions. Different colors represent different chromosomes within each species.

# Whole genome alignment

Aligning whole genomes is, if you think of the algorithms we've covered this far, not at all like short-read assembly or read mapping.

# MUMmer and nucmer

The MUMmer package, which include the program *nucmer* for nucleotide alignments is quite the old friend. It was first published in 1999, with the current version (version 3.0) published in 2004.

*Nucmer* is based on finding matching locations, “seeds”, in both sequences, and then extending these by *smith-waterman alignment*. This is extremely fast, but sometimes not that exact.

# Satsuma

Satsuma, “seedless” alignment, is a whole genome aligner based on the *Fast Fourier Transform*, as well as a refining algorithm they call “Battleship Search”.

Satsuma can be more sensitive than nucmer, but has significantly longer runtime.

# mauve

Mauve is a third program for doing whole genome alignment. Mauve uses a seed matches similar to MUMmer, but uses a recursive algorithm over a phylogenetic guide tree to extend alignments.

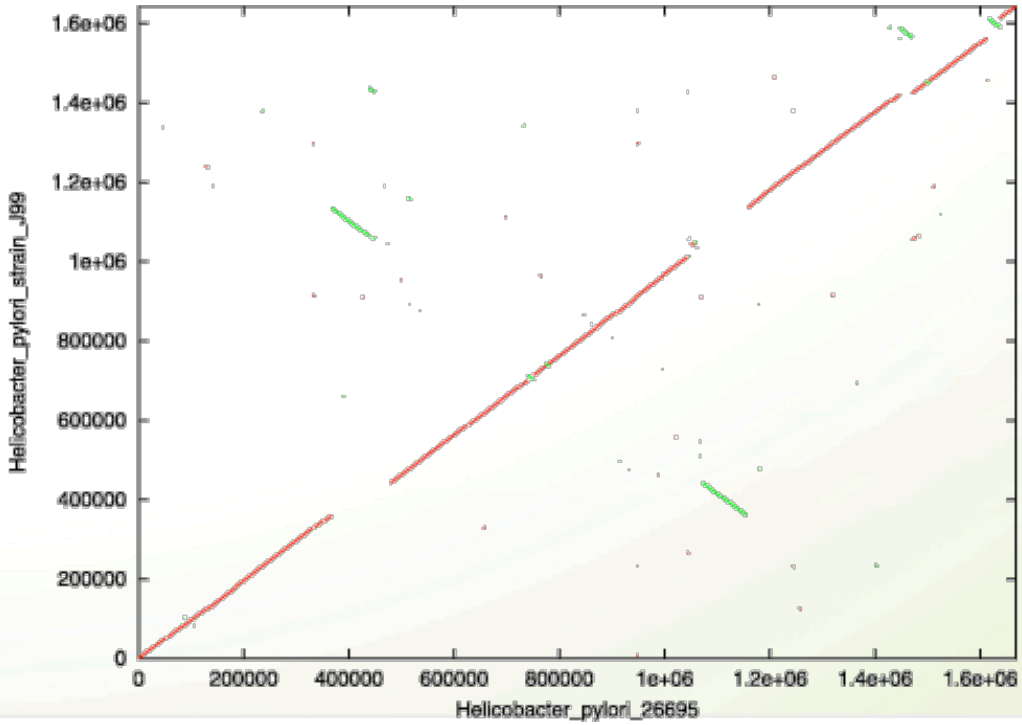
# visualizing

While commonly being output in plain text, whole genome alignments are about as easy as SAM files to read.

Visualization can make this problem a lot simpler!

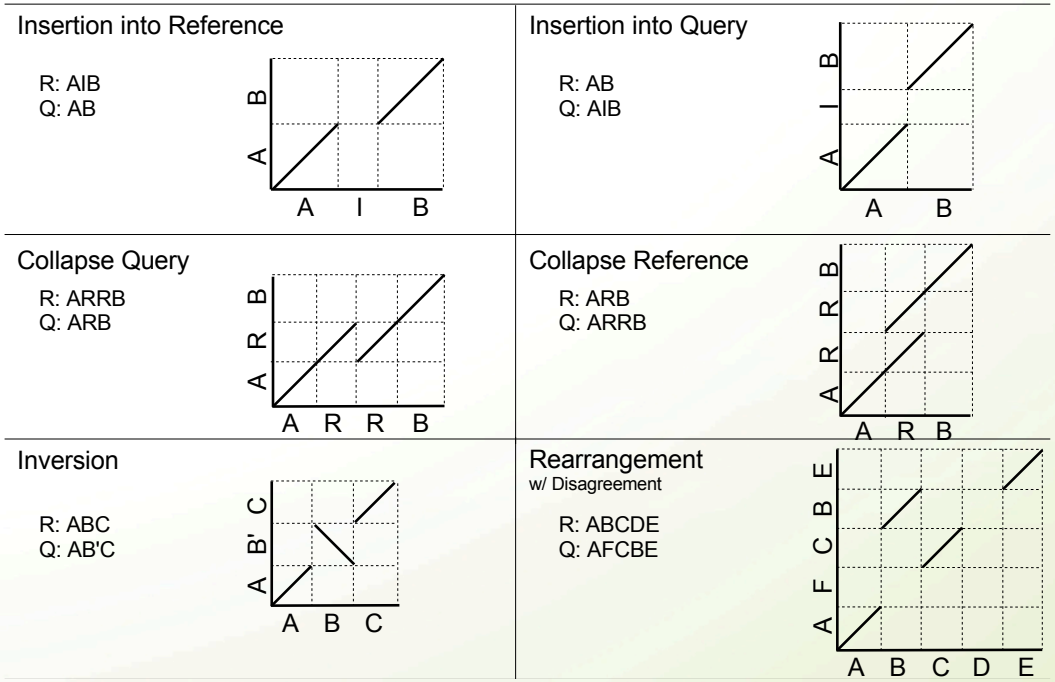
# DotPlots

Used to show overlapping parts of assemblies. Patterns show differences between query and reference.



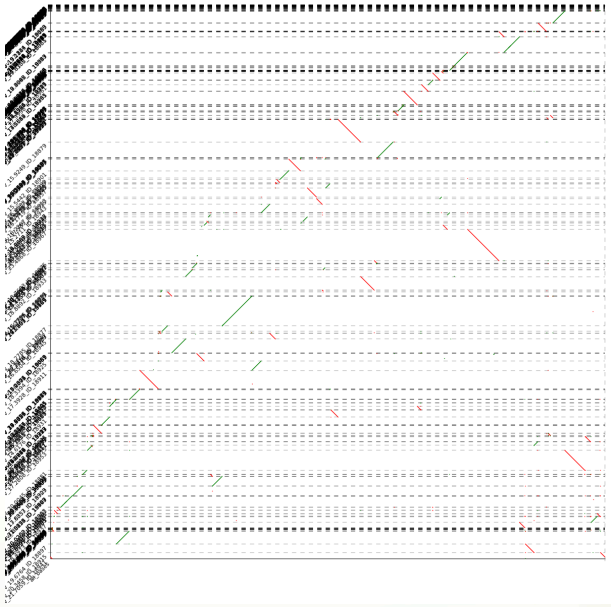
# DotPlots

Used to show overlapping parts of assemblies. Patterns show differences between query and reference.

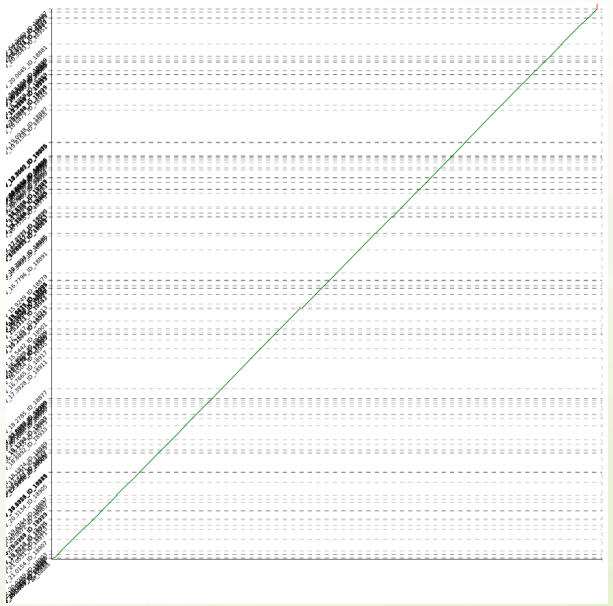


# DotPlots

Note that nucmer output isn't aligned by itself, so what could look like a bad alignment might just be badly sorted!



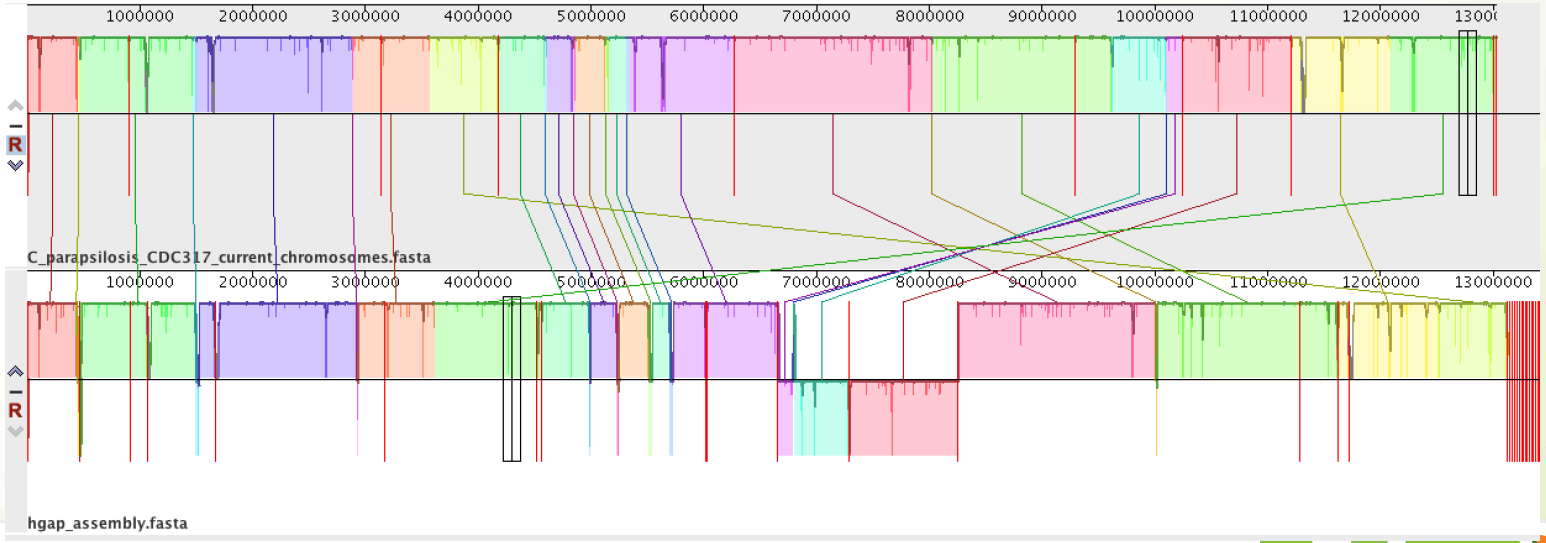
unaligned



aligned

# mauve

Mauve is one of few bioinformatic programs that comes with a graphical interface. This is extremely helpful, but this is another tool that is getting old.



# Questions?