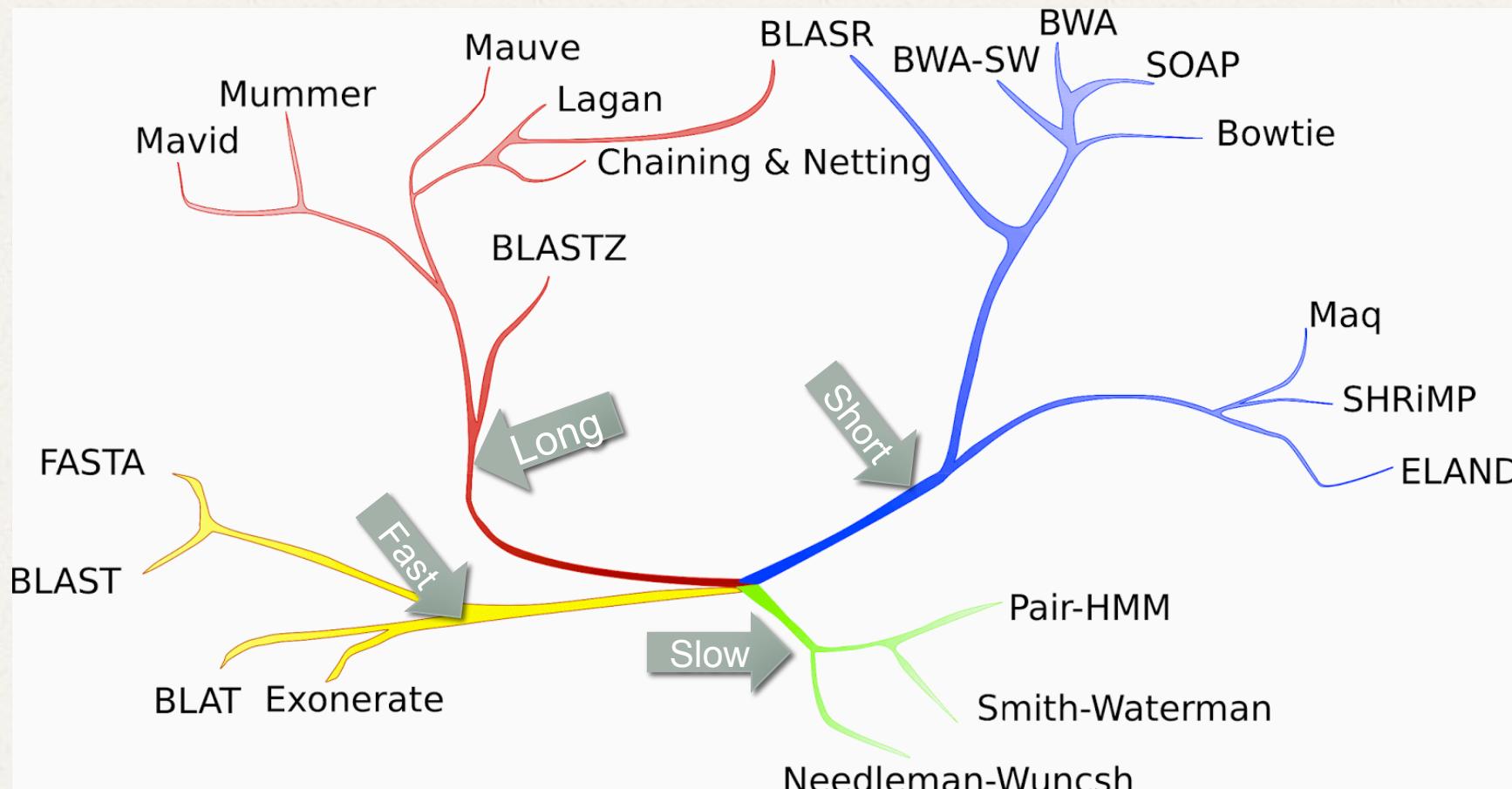


WHOLE GENOME ALIGNMENT

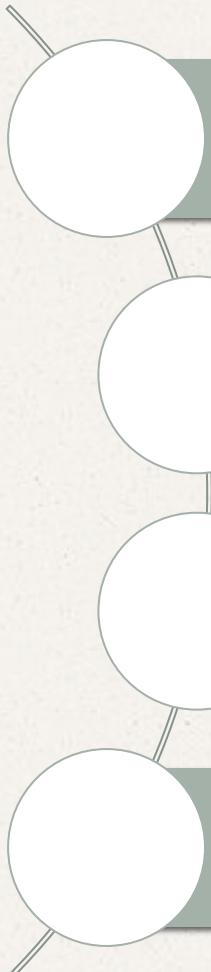


The phylogeny of pairwise alignment



Chaisson & Tesler 2012, *BMC Bioinformatics*

Goals



Why Align Genomes?

Examples from the Literature

Available Tools

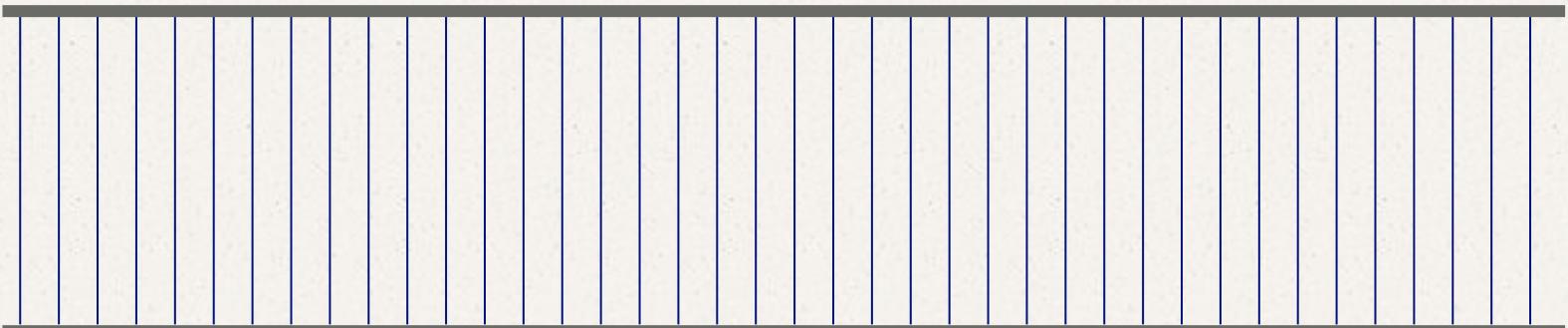
Afternoon: WGA Tutorial



What is a WGA?

- For two genomes, identify the position in genome *A* that corresponds with the same position in genome *B*
 - Predict homologous pairs of positions between multiple genomes

A - CCCGAGGTGCATCGCTGACTGCTGCATGCCGATCGCATCGC

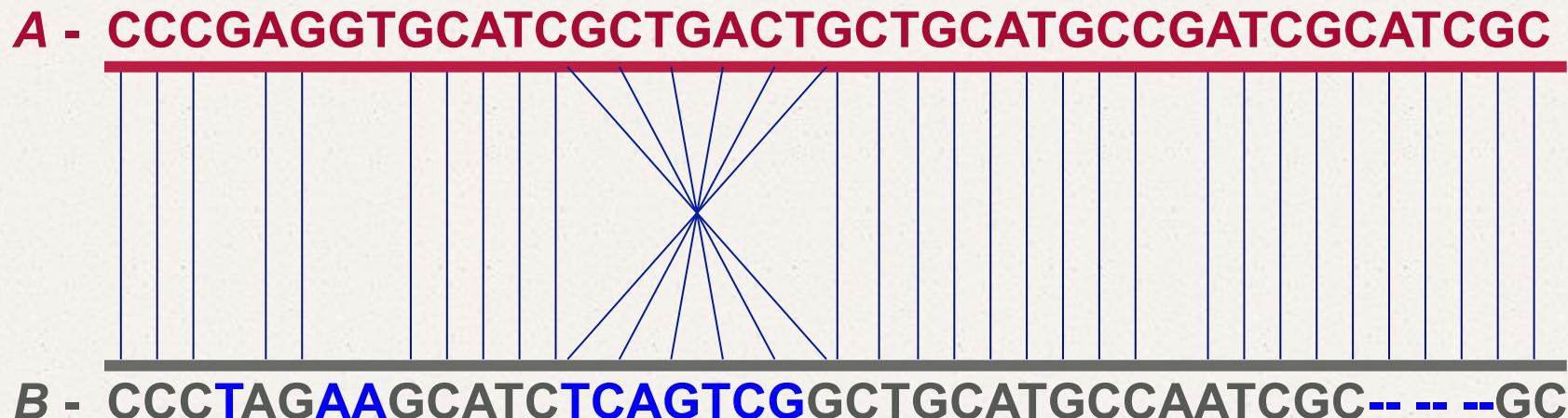


B - CCCGAGGTGCATCGCTGACTGCTGCATGCCGATCGCATCGC



What is a WGA?

- But...
 - Indels
 - Repetitive elements
 - Variants (SNPs, errors)
- Duplications
- Inversion
- Translocations
- All of the above

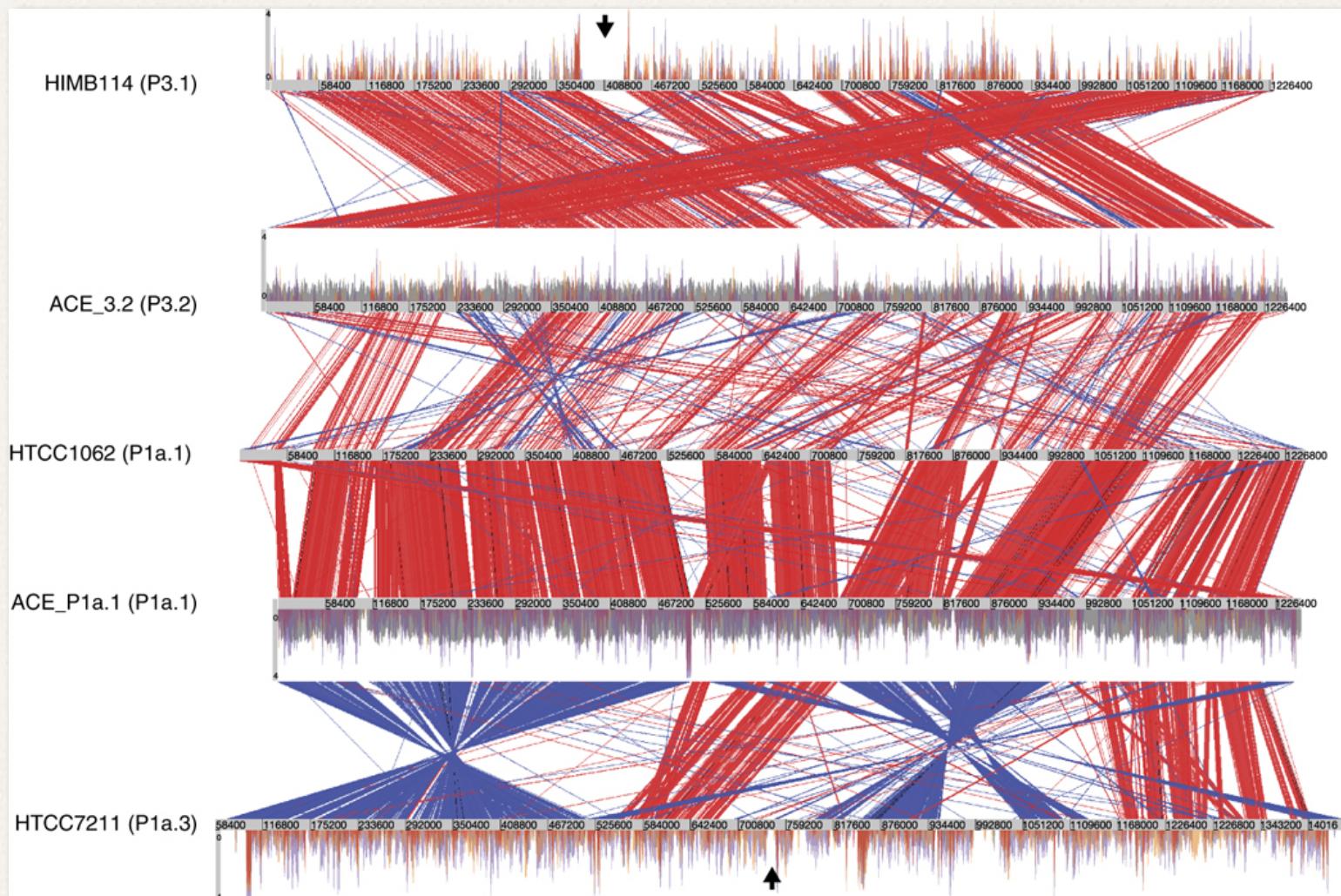


Why align genomes?

- Find syntenic regions (genomic structure)
 - Synteny: conservation of the order of regions



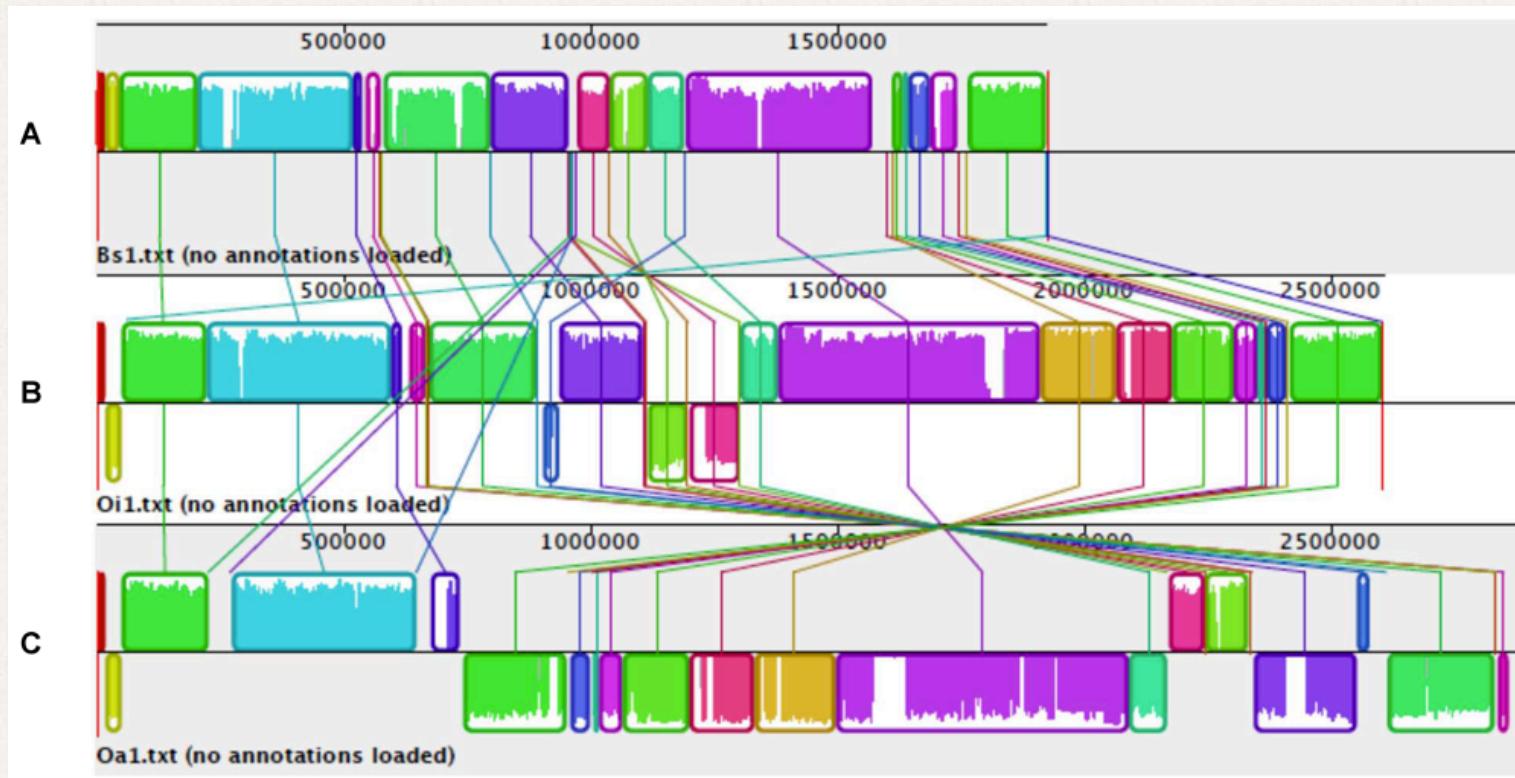
WGA: Genome synteny and recruitment plots of SAR11 genomes.



Brown et al. Mol Syst Biol 2012;8:595



WGA: *Brucella suis*, *Ochtrabactrum intermedium* & *O. anthropi*

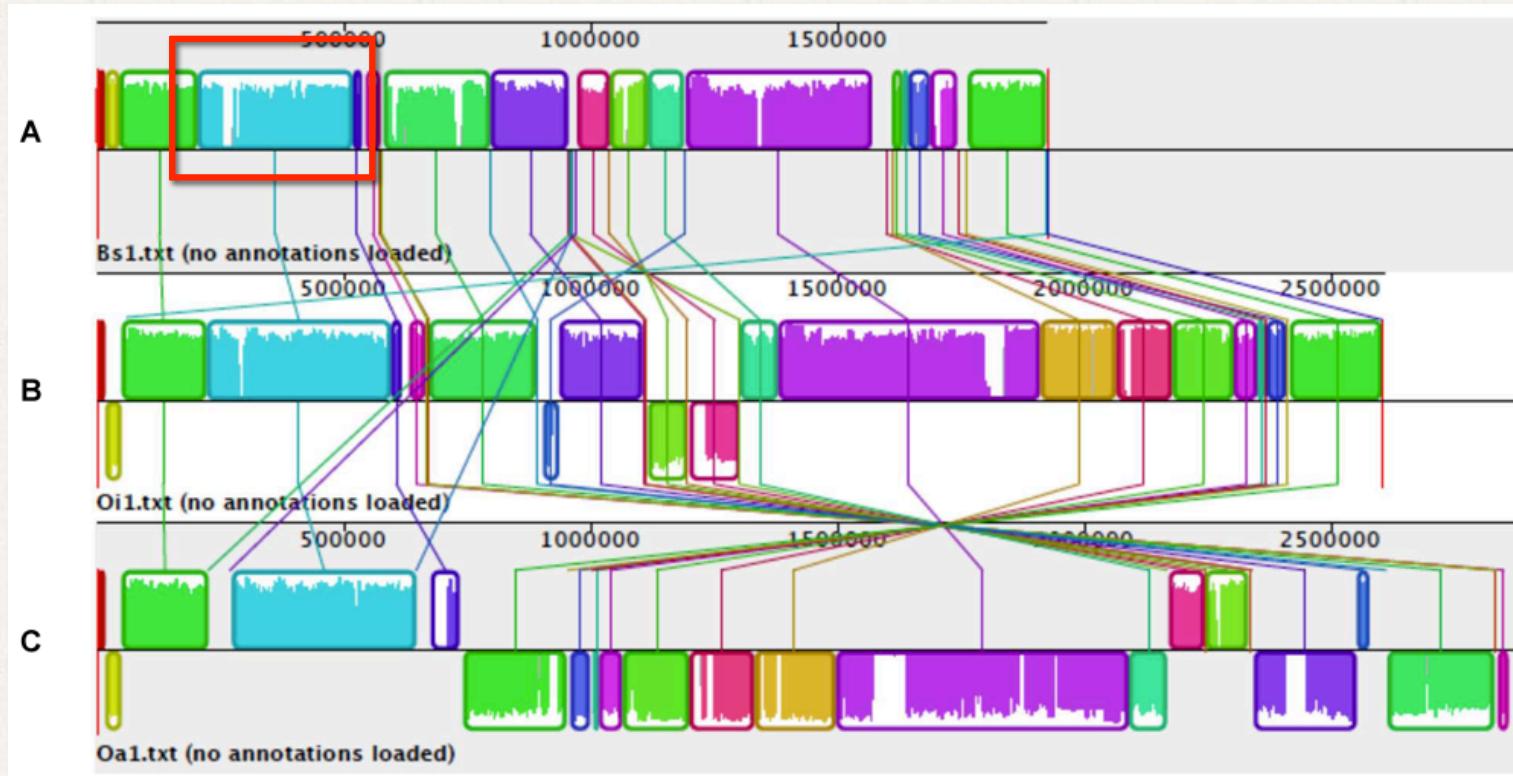


Aujoulat et al. 2012, Genes



WGA: *Brucella suis*, *Ochtrabactrum intermedium* & *O. anthropi*

Block: homologous and co-linear



Aujoulat et al. 2012, Genes



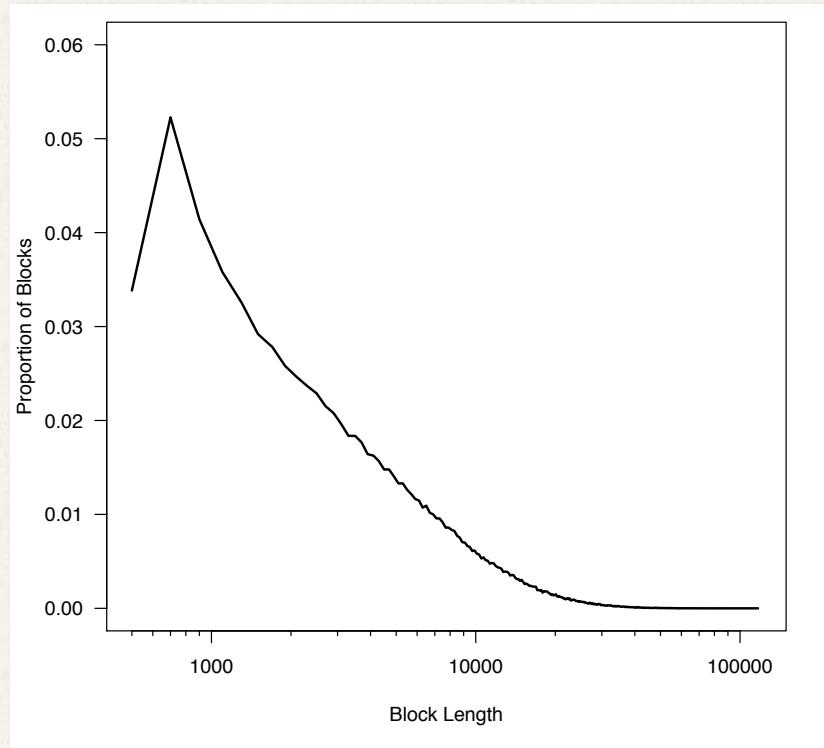
Why align genomes?

- Find syntenic regions (genomic structure)
 - Synteny: conservation of the order of regions
- Detect intergenomic variants (polymorphisms)

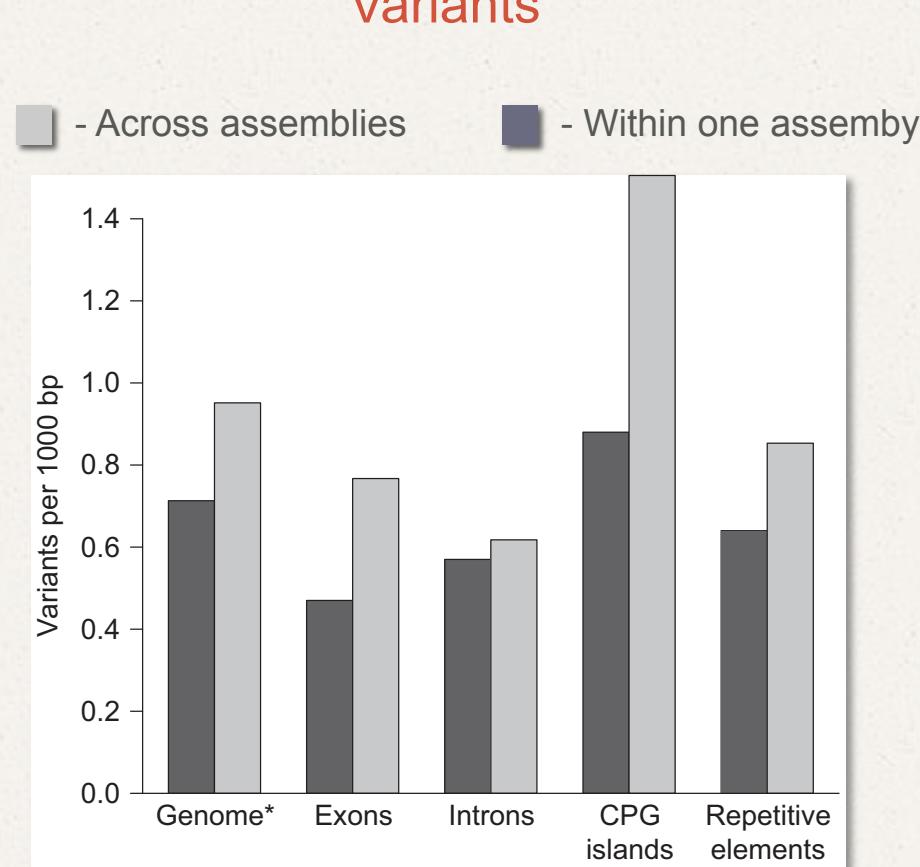


WGA: Two dromedary reference genomes

Block length histogram



Variants



Fitak et al. 2016, Mol Ecol Res 16:314-324



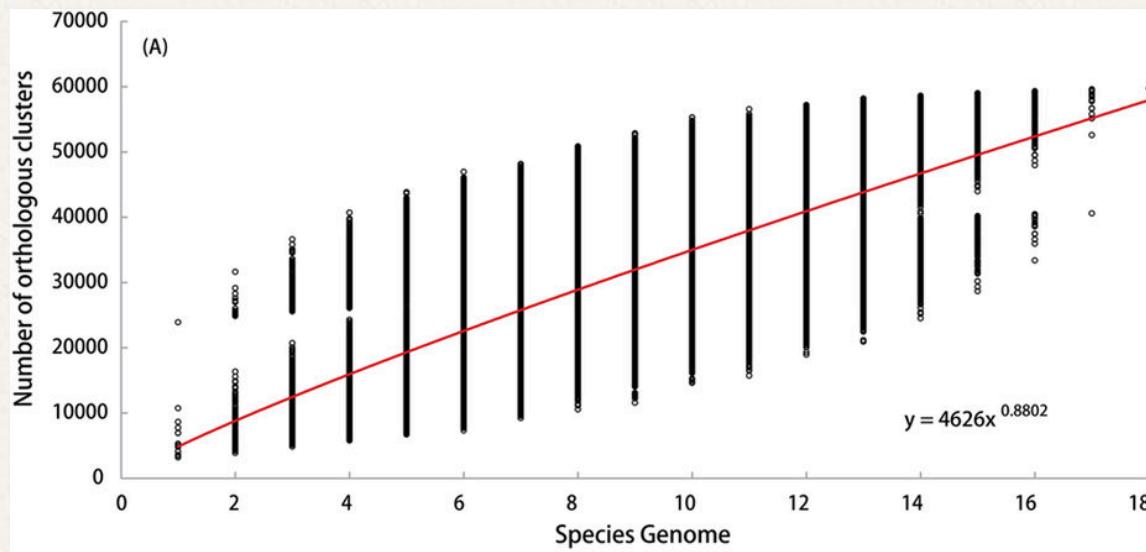
Why align genomes?

- Find syntenic regions (genomic structure)
 - Synteny: conservation of the order of regions
- Detect intergenomic variants (polymorphisms)
- Identify conserved regions
- Identify unique regions

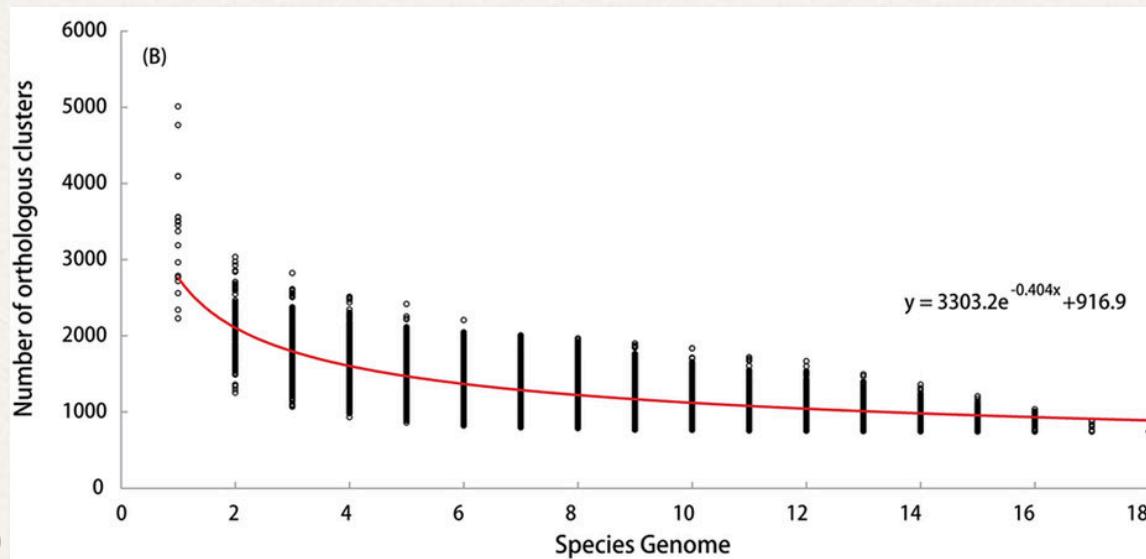


WGA: 18 *Leptospira* species genomes

Pan-genome:



Core genome:



Xu et al. 2016, Sci Rep 6:20020

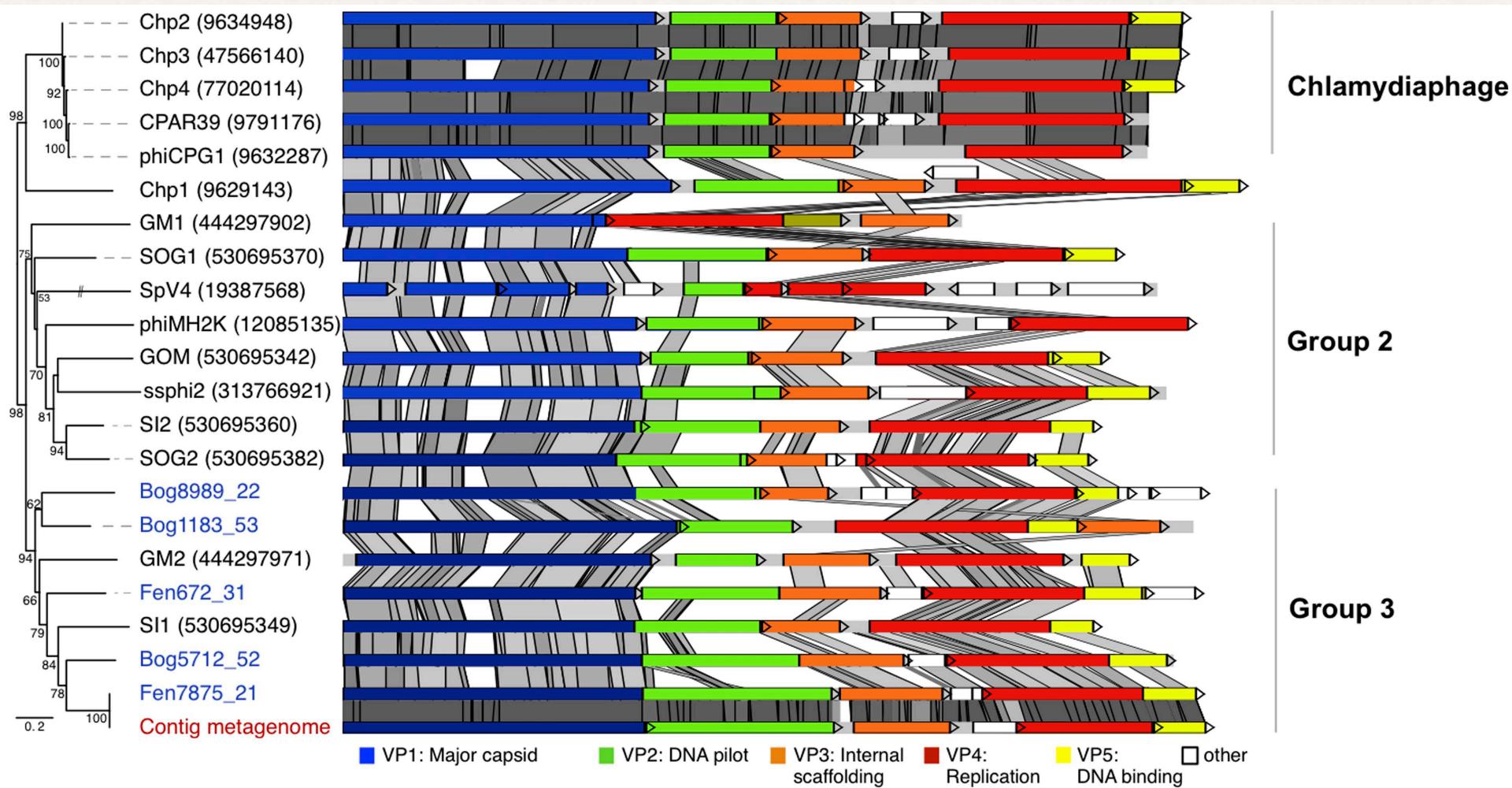


Why align genomes?

- Find syntenic regions (genomic structure)
 - Synteny: conservation of the order of regions
- Detect intergenomic variants (polymorphisms)
- Identify conserved regions
- Identify unique regions
- Phylogenomics



WGA: *Gokushovirinae* bacteriophages



Quaiser et al. 2015, Front Microbiol 6:375

WGA Tools

Software	Type	Pairwise vs Multiple
LastZ	Local	pairwise
MUMer	Local	pairwise
Mugsy	Hierarchical	multiple
Mauve	Hierarchical	multiple
MultiZ	Local	multiple
LAGAN	Global	both
Mavid	Global	multiple
progressiveCactus	Hierarchical/Cactus graphs	multiple

A Great Review:

Dewey 2012. Whole-Genome Alignment.

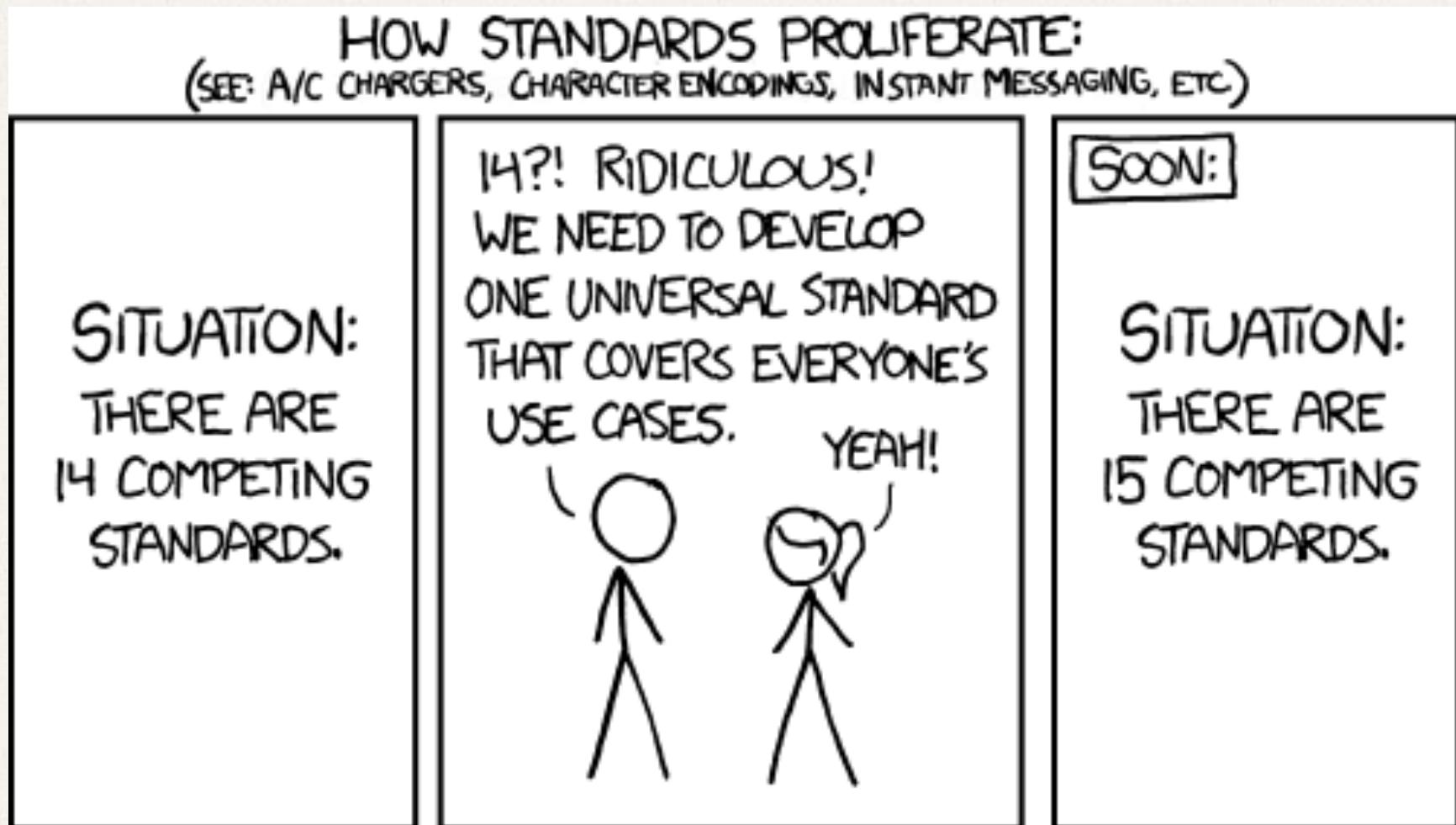
In *Evolutionary Genomics: Statistical and Computational Methods*. p. 237-257



WGA: File formats



WGA: File formats



WGA: File formats

- MAF (multiple alignment format)

```
##maf version=1 scoring=tba.v8
# tba.v8 (((human chimp) baboon) (mouse rat))
# multiz.v7
# maf_project.v5 _tba_right.maf3 mouse _tba_C
# single_cov2.v4 single_cov2 /dev/stdin

a score=23262.0
s hg16.chr7    27578828 38 + 158545518 AAA-CGGAATGTTAACCAAATGA---ATTGTCTCTTACGGTG
s panTro1.chr6  28741140 38 + 161576975 AAA-CGGAATGTTAACCAAATGA---ATTGTCTCTTACGGTG
s baboon        116834 38 +   4622798 AAA-CGGAATGTTAACCAAATGA---GTTGTCTCTTATGGTG
s mm4.chr6      53215344 38 + 151104725 -AATGGGAATGTTAAGCAAACGA---ATTGTCTCTCAGTGTG
s rn3.chr4      81344243 40 + 187371129 -AA-GGGGATGCTAAGCCAATGAGTTGTTCTCAATGTG
```



WGA: File formats

- MAF (multiple alignment format)

```
##maf version=1 scoring=tba.v8
# tba.v8 (((human chimp) baboon) (mouse rat))
# multiz.v7
# maf_project.v5 _tba_right.maf3 mouse _tba_C
# single_cov2.v4 single_cov2 /dev/stdin
```

Header

```
a score=23262.0
s hg16.chr7    27578828 38 + 158545518 AAA-CGGAATGTTAACCAAATGA---ATTGTCTCTTACGGTG
s panTro1.chr6 28741140 38 + 161576975 AAA-CGGAATGTTAACCAAATGA---ATTGTCTCTTACGGTG
s baboon       116834 38 +   4622798 AAA-CGGAATGTTAACCAAATGA---GTTGTCTCTTATGGTG
s mm4.chr6     53215344 38 + 151104725 -AATGGGAATGTTAAGCAAACGA---ATTGTCTCTCAGTGTG
s rn3.chr4     81344243 40 + 187371129 -AA-GGGCATGCTAACCAATGAGTTGTTCTCTCAATGTG
```



WGA: File formats

- MAF (multiple alignment format)

```
##maf version=1 scoring=tba.v8
# tba.v8 (((human chimp) baboon) (mouse rat))
# multiz.v7
# maf_project.v5 _tba_right.maf3 mouse _tba_C
# single_cov2.v4 single_cov2 /dev/stdin
```

Block

```
a score=23262.0
s hg16.chr7    27578828 38 + 158545518 AAA-CGGAATGTTAACCAAATGA---ATTGTCTCTTACGGTG
s panTro1.chr6 28741140 38 + 161576975 AAA-CGGAATGTTAACCAAATGA---ATTGTCTCTTACGGTG
s baboon       116834 38 +   4622798 AAA-CGGAATGTTAACCAAATGA---GTTGTCTCTTATGGTG
s mm4.chr6     53215344 38 + 151104725 -AATGGGAATGTTAAGCAAACGA---ATTGTCTCTCAGTGTG
s rn3.chr4     81344243 40 + 187371129 -AA-GGGGATGCTAAGCCAATGAGTTGTTCTCAATGTG
```



WGA: File formats

- MAF (multiple alignment format)

```
##maf version=1 scoring=tba.v8
# tba.v8 (((human chimp) baboon) (mouse rat))
# multiz.v7
# maf_project.v5 _tba_right.maf3 mouse _tba_C      Score
# single_cov2.v4 single_cov2 /dev/stdin

a score=23262.0
s hg16.chr7    27578828 38 + 158545518 AAA-CGGAATGTTAACCAAATGA---ATTGTCTCTTACGGTG
s panTro1.chr6  28741140 38 + 161576975 AAA-CGGAATGTTAACCAAATGA---ATTGTCTCTTACGGTG
s baboon        116834   38 +  4622798 AAA-CGGAATGTTAACCAAATGA---GTTGTCTCTTATGGTG
s mm4.chr6     53215344 38 + 151104725 -AATGGGAATGTTAAGCAAACGA---ATTGTCTCTCAGTGTG
s rn3.chr4     81344243 40 + 187371129 -AA-GGGGATGCTAAGCCAATGAGTTGTTCTCAATGTG
```



WGA: File formats

- MAF (multiple alignment format)

```
##maf version=1 scoring=tba.v8
# tba.v8 (((human chimp) baboon) (mouse rat))
# multiz.v7
# maf_project.v5 _tba_right.maf3 mouse _tba_C
# single_cov2.v4 single_cov2 /dev/stdin

a score=23262.0
s hg16.chr7 27578828 38 + 158545518 AAA-CGGAATGTTAACCAAATGA---ATTGTCTCTTACGGTG
s panTro1.chr6 28741140 38 + 161576975 AAA-CGGAATGTTAACCAAATGA---ATTGTCTCTTACGGTG
s baboon 116834 38 + 4622798 AAA-CGGAATGTTAACCAAATGA---GTTGTCTCTTATGGTG
s mm4.chr6 53215344 38 + 151104725 -AATGGGAATGTTAAGCAAACGA---ATTGTCTCTCAGTGTG
s rn3.chr4 81344243 40 + 187371129 -AA-GGGCATGCTAACCAATGAGTTGTCTCTCAATGTG
```

Sequence IDs



WGA: File formats

- MAF (multiple alignment format)

```
##maf version=1 scoring=tba.v8
# tba.v8 (((human chimp) baboon) (mouse rat))
# multiz.v7
# maf_project.v5 _tba_right.maf3 mouse _tba_C      Start
# single_cov2.v4 single_cov2 /dev/stdin

a score=23262.0
s hg16.chr7    27578828 38 + 158545518 AAA-CGGAATGTTAACCAAATGA---ATTGTCTCTTACGGTG
s panTro1.chr6 28741140 38 + 161576975 AAA-CGGAATGTTAACCAAATGA---ATTGTCTCTTACGGTG
s baboon       116834   38 +  4622798 AAA-CGGAATGTTAACCAAATGA---GTTGTCTCTTATGGTG
s mm4.chr6     53215344 38 + 151104725 -AATGGGAATGTTAAGCAAACGA---ATTGTCTCTCAGTGTG
s rn3.chr4     81344243 40 + 187371129 -AA-GGGGATGCTAAGCCAATGAGTTGTTCTCAATGTG
```



WGA: File formats

- MAF (multiple alignment format)

```
##maf version=1 scoring=tba.v8
# tba.v8 (((human chimp) baboon) (mouse rat))
# multiz.v7
# maf_project.v5 _tba_right.maf3 mouse _tba_C
# single_cov2.v4 single_cov2 /dev/stdin

a score=23262.0
s hg16.chr7    27578828 38 + 158545518 AAA-CGGAATGTTAACCAAATGA---ATTGTCTCTTACGGTG
s panTro1.chr6 28741140 38 + 161576975 AAA-CGGAATGTTAACCAAATGA---ATTGTCTCTTACGGTG
s baboon       116834   38 + 4622798  AAA-CGGAATGTTAACCAAATGA---GTTGTCTCTTATGGTG
s mm4.chr6     53215344 38 + 151104725 -AATGGGAATGTTAAGCAAACGA---ATTGTCTCTCAGTGTG
s rn3.chr4     81344243  40 + 187371129 -AA-GGGCATGCTAACCAATGAGTTGTTCTCTCAATGTG
```



WGA: File formats

- MAF (multiple alignment format)

```
##maf version=1 scoring=tba.v8
# tba.v8 (((human chimp) baboon) (mouse rat))
# multiz.v7
# maf_project.v5 _tba_right.maf3 mouse _tba_C
# single_cov2.v4 single_cov2 /dev/stdin

a score=23262.0
s hg16.chr7    27578828 38 + 158545518 AAA-CGGAATGTTAACCAAATGA---ATTGTCTCTTACGGTG
s panTro1.chr6 28741140 38 + 161576975 AAA-CGGAATGTTAACCAAATGA---ATTGTCTCTTACGGTG
s baboon       116834   38 +  4622798 AAA-CGGAATGTTAACCAAATGA---GTTGTCTCTTATGGTG
s mm4.chr6     53215344 38 + 151104725 -AATGGGAATGTTAACCAAACGA---ATTGTCTCTCAGTGTG
s rn3.chr4     81344243 40 + 187371129 -AA-CGGGATGCTAACCAATGAGTTGTTCTCTCAATGTG
```

Source length



WGA: File formats

- MAF (multiple alignment format)

```
##maf version=1 scoring=tba.v8
# tba.v8 (((human chimp) baboon) (mouse rat))
# multiz.v7
# maf_project.v5 _tba_right.maf3 mouse _tba_c
# single_cov2.v4 single_cov2 /dev/stdin
```

```
a score=23262.0
s hg16.chr7    27578828 38 + 158545518
s panTro1.chr6 28741140 38 + 161576975
s baboon       116834 38 +   4622798
s mm4.chr6     53215344 38 + 151104725
s rn3.chr4     81344243 40 + 187371129
```

Aligned sequence block

```
AAA-CGGAATGTTAACCAAATGA---ATTGTCTCTTACGGTG
AAA-CGGAATGTTAACCAAATGA---ATTGTCTCTTACGGTG
AAA-CGGAATGTTAACCAAATGA---GTTGTCTCTTATGGTG
-AATGGGAATGTTAAGCAAACGA---ATTGTCTCTCAGTGTG
-AA-GGGGATGCTAAGCCAATGAGTTGTTCTCTCAATGTG
```



CoGe: [www.genomevolution.org/coge](https://genomevolution.org/coge/)

The screenshot shows the homepage of the CoGe platform. At the top, there is a browser toolbar with several tabs open, including "Inbox (1) -", "lastz tutorial", "Animal Mag", "www.tucso", "stenglein-la", "genome ali", "12864_201", "Comparativ", and "CoGe: Com". The URL in the address bar is <https://genomevolution.org/coge/>. The main header features the "CoGe" logo, a search bar with an "advanced" link, and navigation links for "My Data", "Tools", "Help", and "Log in". A green banner at the top encourages users to take a "Future Directions Survey" with a link to <http://goo.gl/nizYYz>. Below this, key statistics are displayed: "Organisms: 17,713", "Genomes: 32,316", "Features: 1,254,888,775", and "Experiments: 8,509". The page is divided into several sections: "New to CoGe?", "Tools", "Latest News", and "Worldwide Usage". The "New to CoGe?" section contains a brief introduction and links to "Get started", "Create an Account", "Tutorials", "Documentation", and "FAQ". The "Tools" section lists "OrganismView" (with a DNA helix icon) and "EPIC-CoGe" (with a bar chart icon). The "Latest News" section highlights recent updates: "NGS Analysis Performance Improvements" (March 13th 2017), "GATK HaplotypeCaller GVCF Now Available" (March 13th 2017), "BBDuk Trimmer Now Available" (March 13th 2017), "Improved Search Results Interface" (January 5th 2017), and "EPIC-CoGe Genome Browser Update" (December 12th 2016). A "...more..." link is also present. The "Worldwide Usage" section shows a horizontal bar with various file icons and names: "fmicb-06-00375.pdf", "srep20020-s9.xls", "Norway_Lecture4.ppt", "wga.pdf", and "F3.ppt". A "Show All" button and a close button are located at the end of this bar.



CoGe: [www.genomevolution.org/coge/](https://genomevolution.org/coge/)

The screenshot shows the CoGe website interface. At the top, there is a toolbar with various browser tabs and a search bar. Below the toolbar, the URL <https://genomevolution.org/coge/> is displayed.

Tools

- OrganismView**: Search for organisms and get an overview of their genomic make-up. Example - Documentation
- EPIC-CoGe**: Visualize genomes and experiments using a dynamic, interactive genome browser. Example - Documentation
- CoGeBlast**: Blast sequences against any number of organisms in CoGe. Example - Documentation
- SynMap**: Compare any two genomes to identify regions of synteny. Example - Documentation
- SynMap3D**: Compare any three genomes to identify regions of synteny. Example - Documentation
- SynFind**: Search CoGe's annotation database for homologs. Example
- GEvo**: Compare sequences and genomic regions to discover patterns of genome evolution. Example - Documentation
- Load Genome**: Load your own genome from NCBI or a FASTA file. Documentation
- Load Experiment (LoadExp+)**: Load experimental data from various standard input formats (such as BED, WIG, BAM, and FASTQ) and run downstream analyses including read mapping, expression measurement, and SNP identification. Documentation

BBDuk Trimmer Now Available
March 13th 2017

Improved Search Results Interface
January 5th 2017

EPIC-CoGe Genome Browser Update
December 12th 2016

[...more...](#)

Worldwide Usage

A world map where countries are shaded in green, indicating usage of the CoGe platform. The map shows high usage density in North America, Europe, and parts of Asia and Australia.

Tutorials

Links to various tutorials and resources are shown below the map.

At the bottom, there is a horizontal navigation bar with links like Home, Help, Support, Contact, and Log In. Below this, a file menu bar shows several open files: fmcb-06-00375.pdf, srep20020-s9.xls, Norway_Lecture4.ppt, wga.pdf, F3.ppt, and a Show All button.



CoGe: [www.genomevolution.org/coge/](https://genomevolution.org/coge/)

The screenshot shows the CoGe website interface. On the left, there is a sidebar titled "Tools" containing several icons and descriptions:

- OrganismView**: Search for organisms and get an overview of their genomic make-up. Example - Documentation
- EPIC-CoGe**: Visualize genomes and experiments using a dynamic, interactive genome browser. Example - Documentation
- CoGeBlast**: Blast sequences against any number of organisms in CoGe. Example - Documentation
- SynMap**: Compare any two genomes to identify regions of synteny. Example - Documentation
- SynMap3D**: Compare any three genomes to identify regions of synteny. Example - Documentation
- SynFind**: Search CoGe's annotation database for homologs. Example
- GEvo**: Compare sequences and genomic regions to discover patterns of genome evolution. Example - Documentation
- Load Genome**: Load your own genome from NCBI or a FASTA file. Documentation
- Load Experiment (LoadExp+)**: Load experimental data from various standard input formats (such as BED, WIG, BAM, and FASTQ) and run downstream analyses including read mapping, expression measurement, and SNP identification. Documentation

On the right side, there is a "Worldwide Usage" section featuring a world map where green shading indicates active users. Below the map is a "Tutorials" section with three video thumbnail links.

At the bottom, there is a toolbar with several document icons and a "Show All" button.

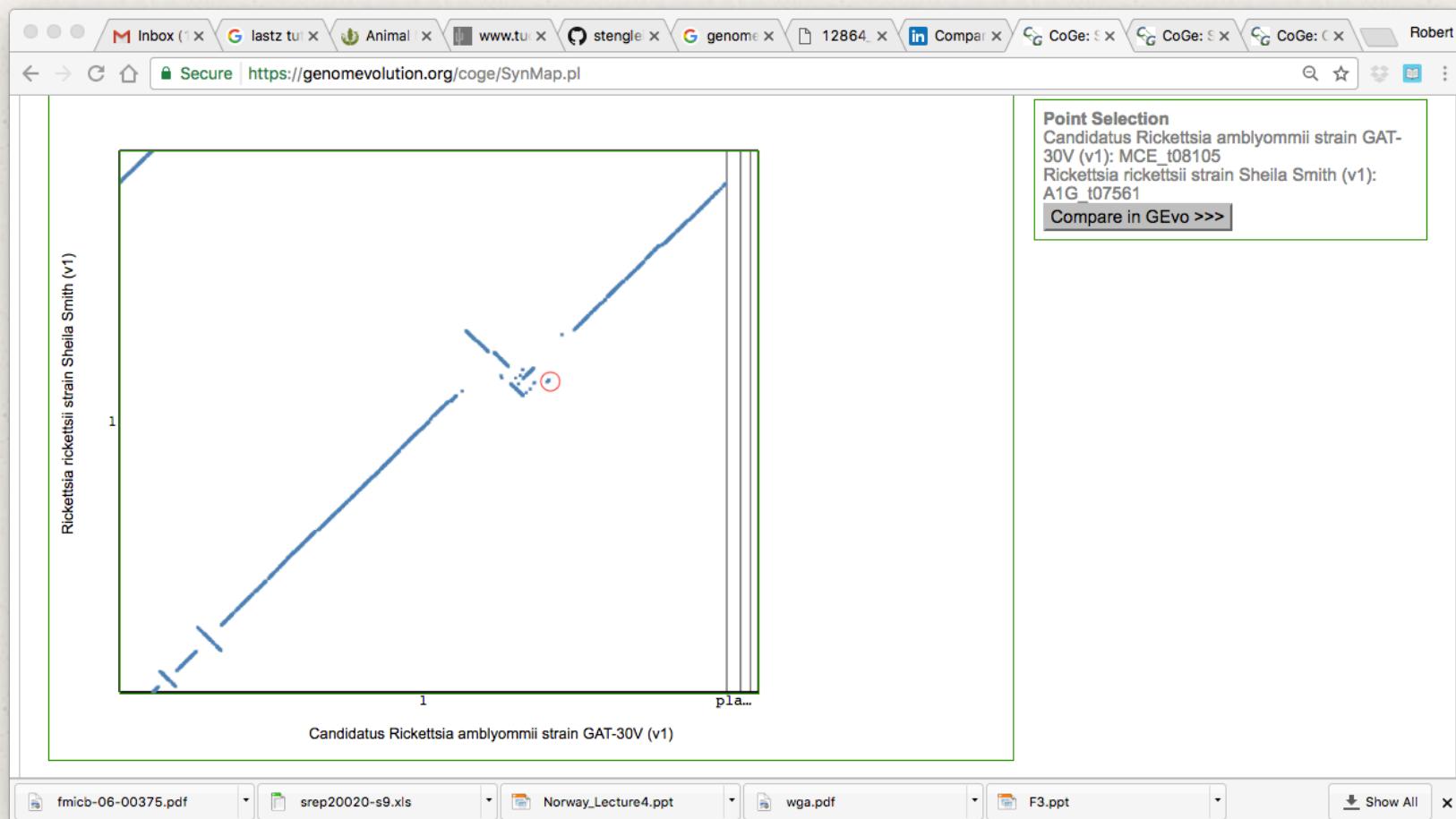


CoGe: www.genomevolution.org/coge

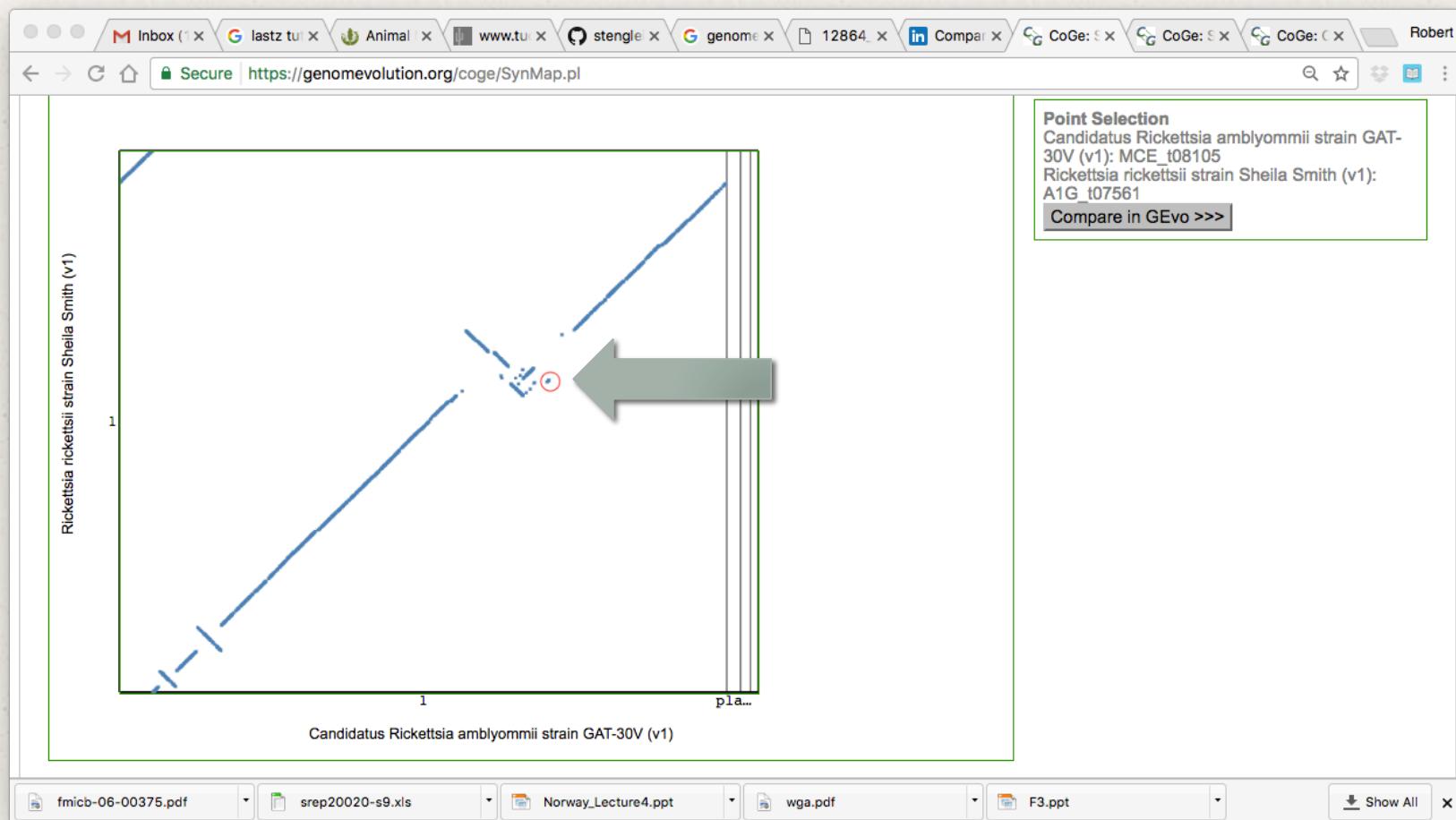
The screenshot shows the CoGe SynMap interface. At the top, there are tabs for "Select Organisms", "Analysis Options", and "Display Options". The "Select Organisms" tab is active. In the "Organism 1" section, the search bar contains "Ricket" and the results list includes "Rickettsia rickettsii strain Hino (id36438)", "Rickettsia rickettsii strain Hlp#2 (id36396)", "Rickettsia rickettsii strain Iowa (id25023)", "Rickettsia rickettsii strain Sheila Smith (id25022)" (which is highlighted), and "Rickettsia sibirica strain 246 (id25024)". Below this, there are dropdown menus for "Genomes" set to "unmasked (v1,id18653)" and "CDS". A detailed description panel shows the following information for "Rickettsia rickettsii strain Sheila Smith":
Description: Rickettsia rickettsii strain Sheila Smith (v1, id18653): unmasked
Taxonomy: Bacteria; Proteobacteria; Alphaproteobacteria; Rickettsiales; Rickettsiaceae; Rickettsiae; Rickettsia; spotted fever group
Source: NCBI
Dataset: CP000848.gbk; LOCUS: CP000848, ACCESSION: CP000848, VERSION: 1
Chromosomes: 1
DNA content: GC: 32.47%, AT: 67.53%, N: 0%, X: 0%
Total length: 1,257,710
In the "Organism 2" section, the search bar contains "rickettsia" and the results list includes "Candidatus Pelagibacter ubique strain HTCC1062 (id23945)", "Candidatus Rickettsia amblyommii strain AaR/SC (id30886)", "Candidatus Rickettsia amblyommii strain GAT-30V (id36315)" (which is highlighted), and "Cycloclasticus sp. strain MCCC 1A01040; P1 (id39305)". Below this, there are dropdown menus for "Genomes" set to "unmasked (v1,id16940)" and "CDS".
At the bottom, there is a toolbar with file icons and labels: fmicb-06-00375.pdf, srep2020-s9.xls, Norway_Lecture4.ppt, wga.pdf, F3.ppt, Show All, and a close button.



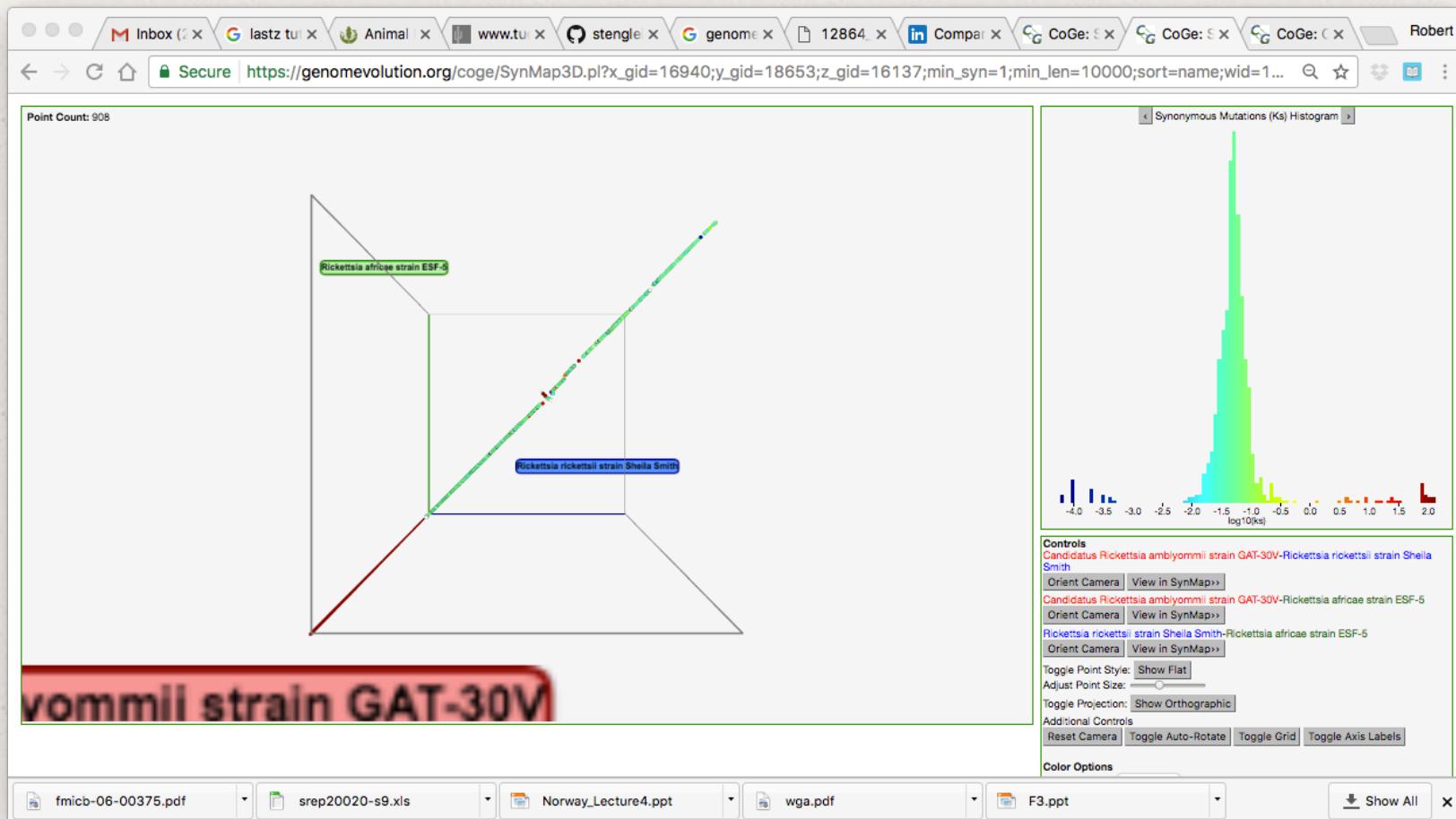
CoGe: [www.genomevolution.org/coge](https://genomevolution.org/coge/SynMap.pl)



CoGe: [www.genomevolution.org/coge](https://genomevolution.org/coge/SynMap.pl)



CoGe: www.genomevolution.org/coge



CoGe: [www.genomevolution.org/coge/](https://genomevolution.org/coge/)

The screenshot shows the homepage of the CoGe website (<https://genomevolution.org/coge/>). The page features a navigation bar at the top with various tabs and links. On the left, there is a sidebar titled "Tools" containing eight entries, each with an icon and a brief description:

- OrganismView**: Search for organisms and get an overview of their genomic make-up. Example - Documentation
- EPIC-CoGe**: Visualize genomes and experiments using a dynamic, interactive genome browser. Example - Documentation
- CoGeBlast**: Blast sequences against any number of organisms in CoGe. Example - Documentation
- SynMap**: Compare any two genomes to identify regions of synteny. Example - Documentation
- SynMap3D**: Compare any three genomes to identify regions of synteny. Example - Documentation
- SynFind**: Search CoGe's annotation database for homologs. Example
- GEvo**: Compare sequences and genomic regions to discover patterns of genome evolution. Example - Documentation
- Load Genome**: Load your own genome from NCBI or a FASTA file. Documentation
- Load Experiment (LoadExp+)**: Load experimental data from various standard input formats (such as BED, WIG, BAM, and FASTQ) and run downstream analyses including read mapping, expression measurement, and SNP identification. Documentation

To the right of the tools, there are several news items:

- BBDuk Trimmer Now Available** (March 13th 2017)
- Improved Search Results Interface** (January 5th 2017)
- EPIC-CoGe Genome Browser Update** (December 12th 2016)

A link "...more..." is also present. Below these news items is a section titled "Worldwide Usage" featuring a world map where green shading indicates active usage across continents.

At the bottom of the page, there is a "Tutorials" section with three video icons. The bottom navigation bar shows several open tabs with file names: fmicb-06-00375.pdf, srep20020-s9.xls, Norway_Lecture4.ppt, wga.pdf, F3.ppt, and a "Show All" button.

