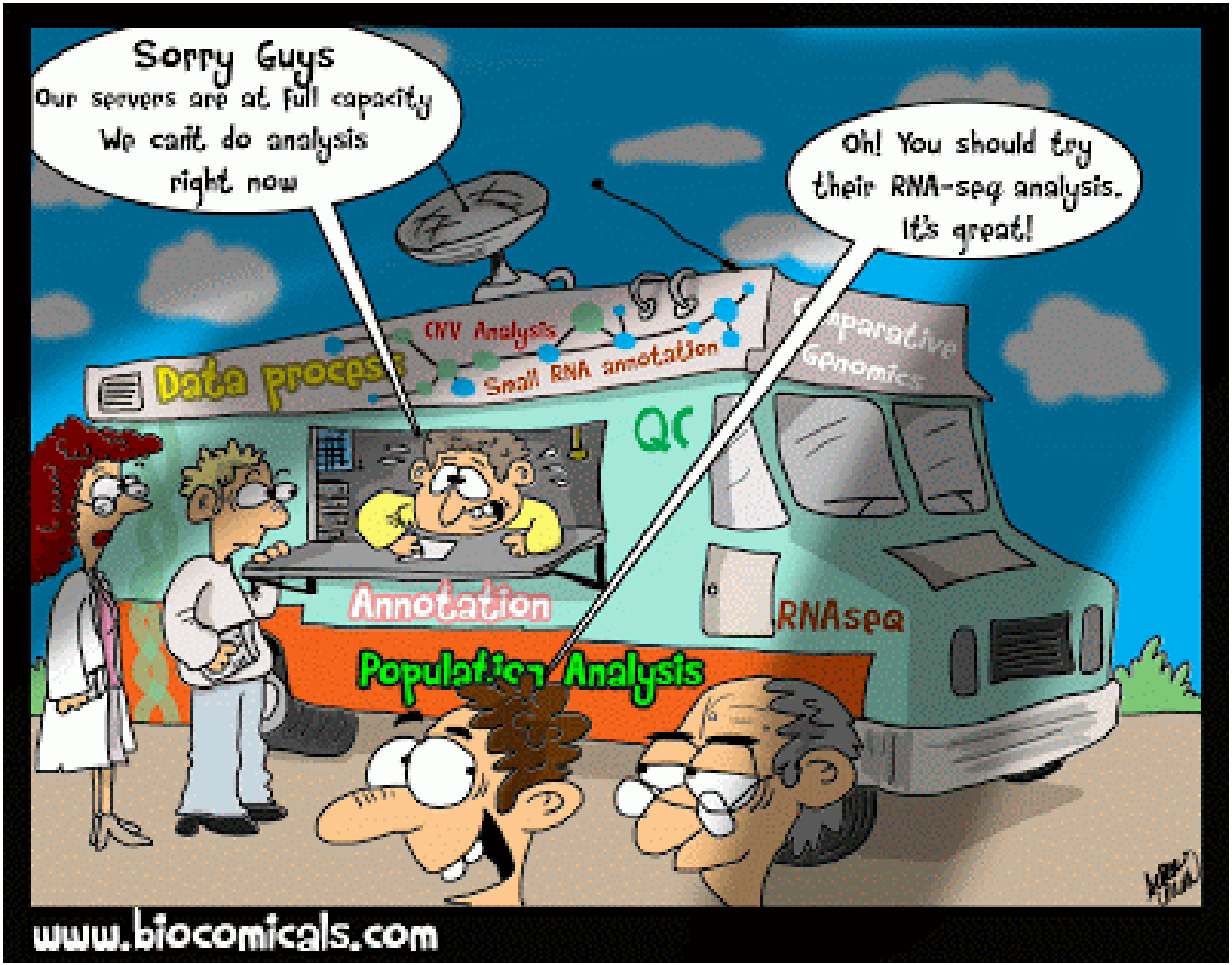
PhylOligo & ContaLocate

Contamination identification, location and surrogate metagenomics

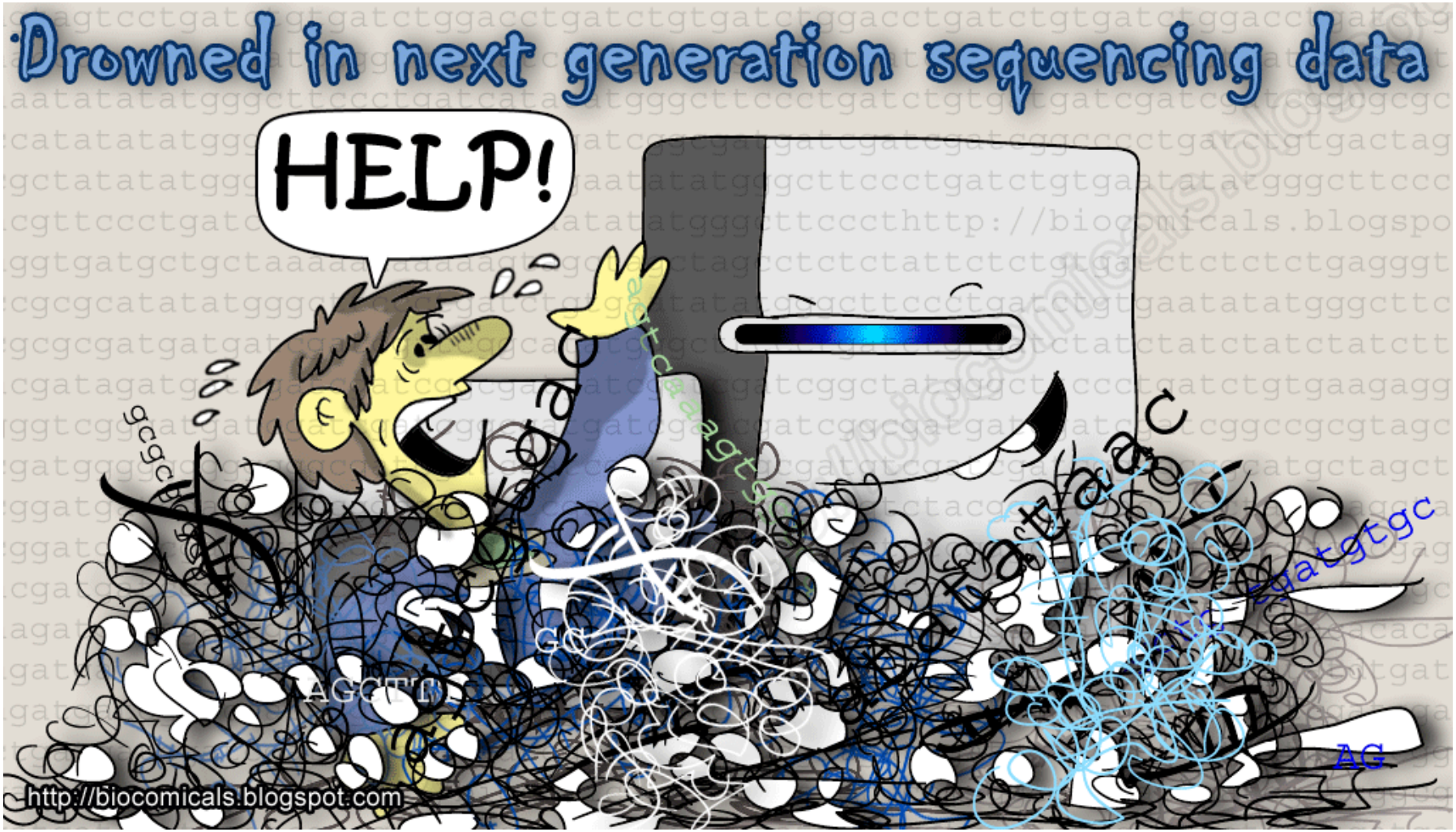
Ludovic Mallet, Franck Cerruti, Tristan Bitard-Feildel and Hélène Chiapello

November 25, 2016

Breakfast of champions



Don’t drop it on your feet



Genomics on a black friday

Soo, what now?

"Data don't make any sense,
we will have to resort to statistics."

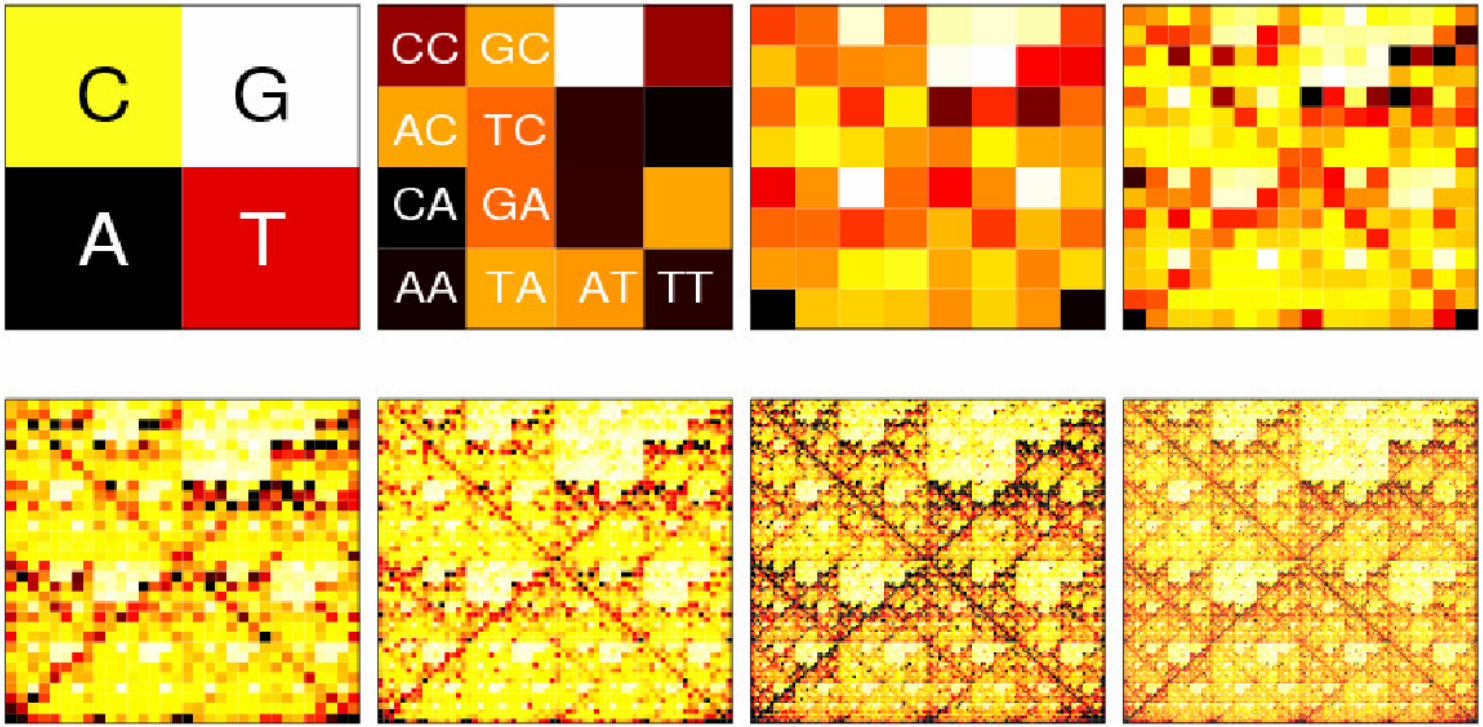
Oligonucleotides

Oligonucleotide $\equiv$ k-mers (in DNA) $\equiv$ words

Frequency of -overlapping- fixed-lenght words in a sequence

```
ACGACTGC ⇒ {ACGA, CGAC, GACT, ACTG, CTGC}
 CGAC
  GACT
   ACTG
    CTGC
```

Composition in oligonucleotides

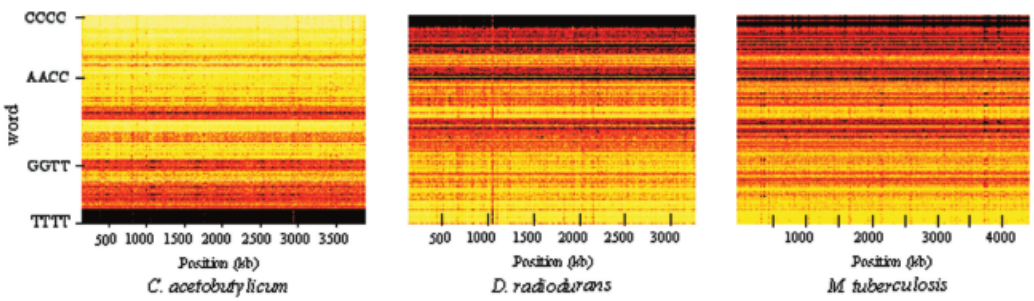


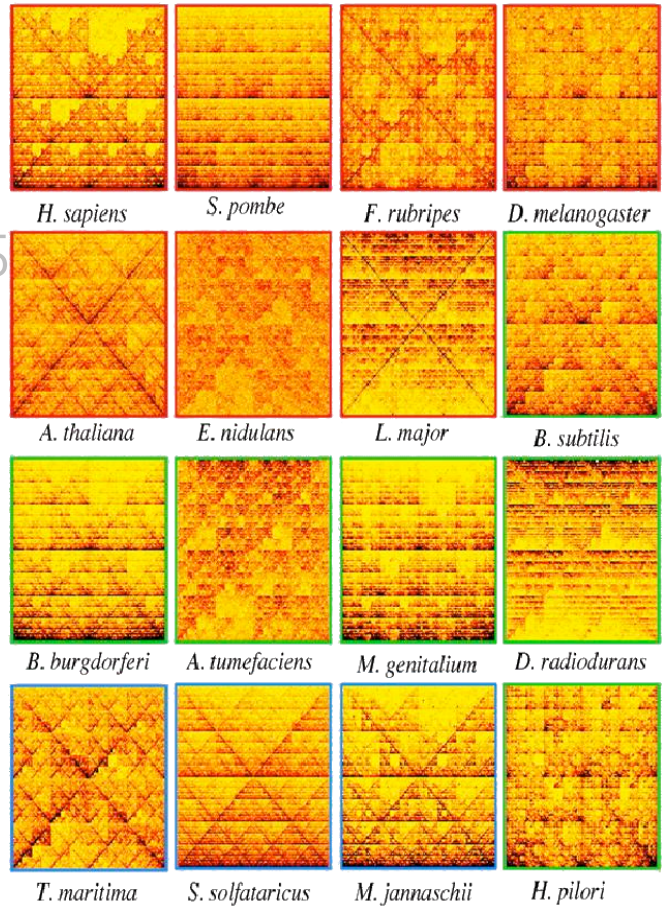
4 nucleotide word frequencies ⇒ Genomic signature

Chromosome 22, *Homo sapiens*, Becq J, PC

The genomic signature

- Species-specific
- Quantitatively comparable
- Suitable for phylogeny - Chapus C., 2005
- Globally homogeneous

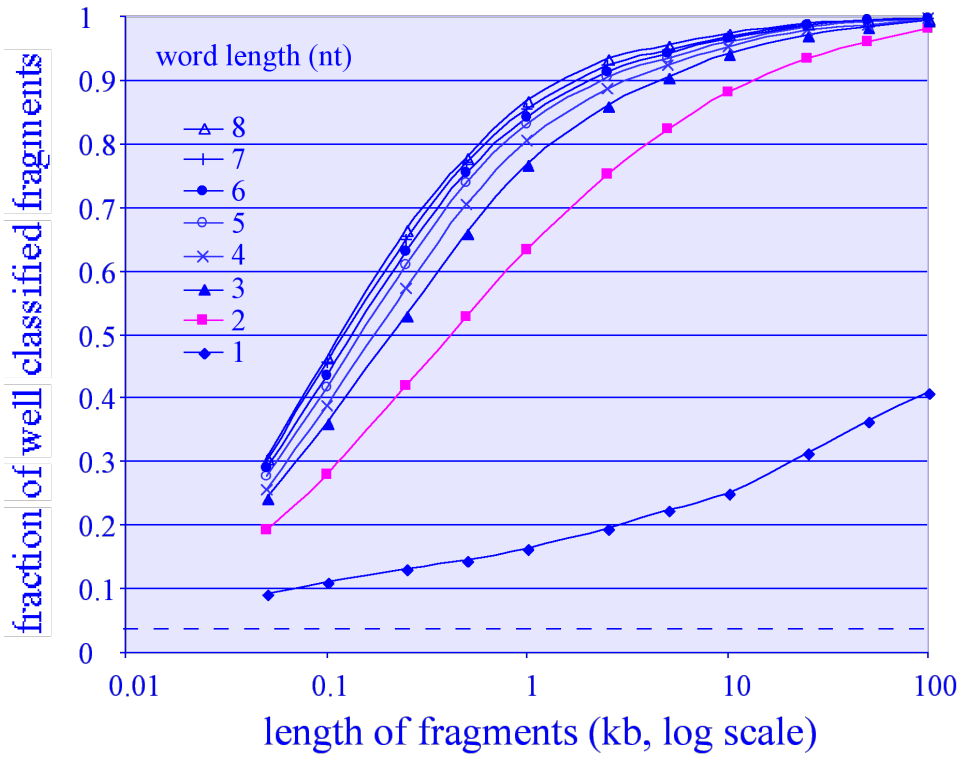


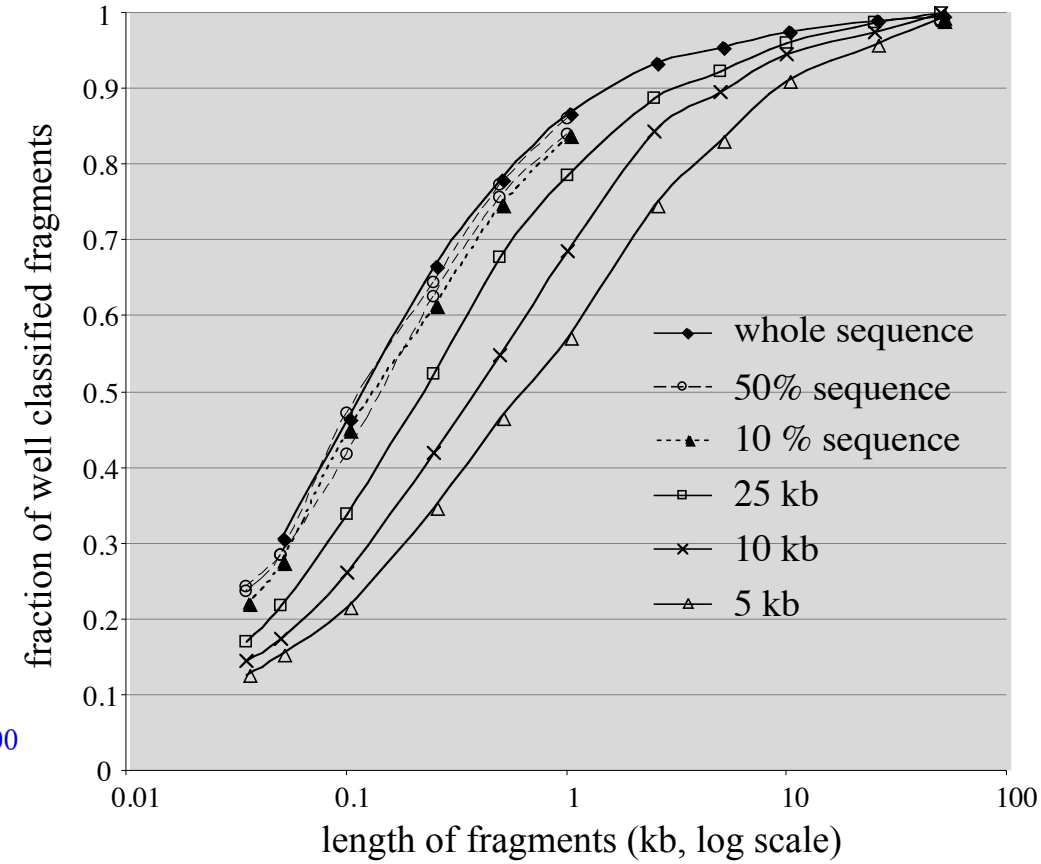


Dufraigne C. *et al.*, *NAR*, 2005

Deschavanne P. *et al.*, *Mol Biol Evol.*, 1999

Size DO matters





Deschavanne P, PC

Horizontal Gene Transfer (HGT): Parametric methods

«Sequences that are new to a bacterial genome,

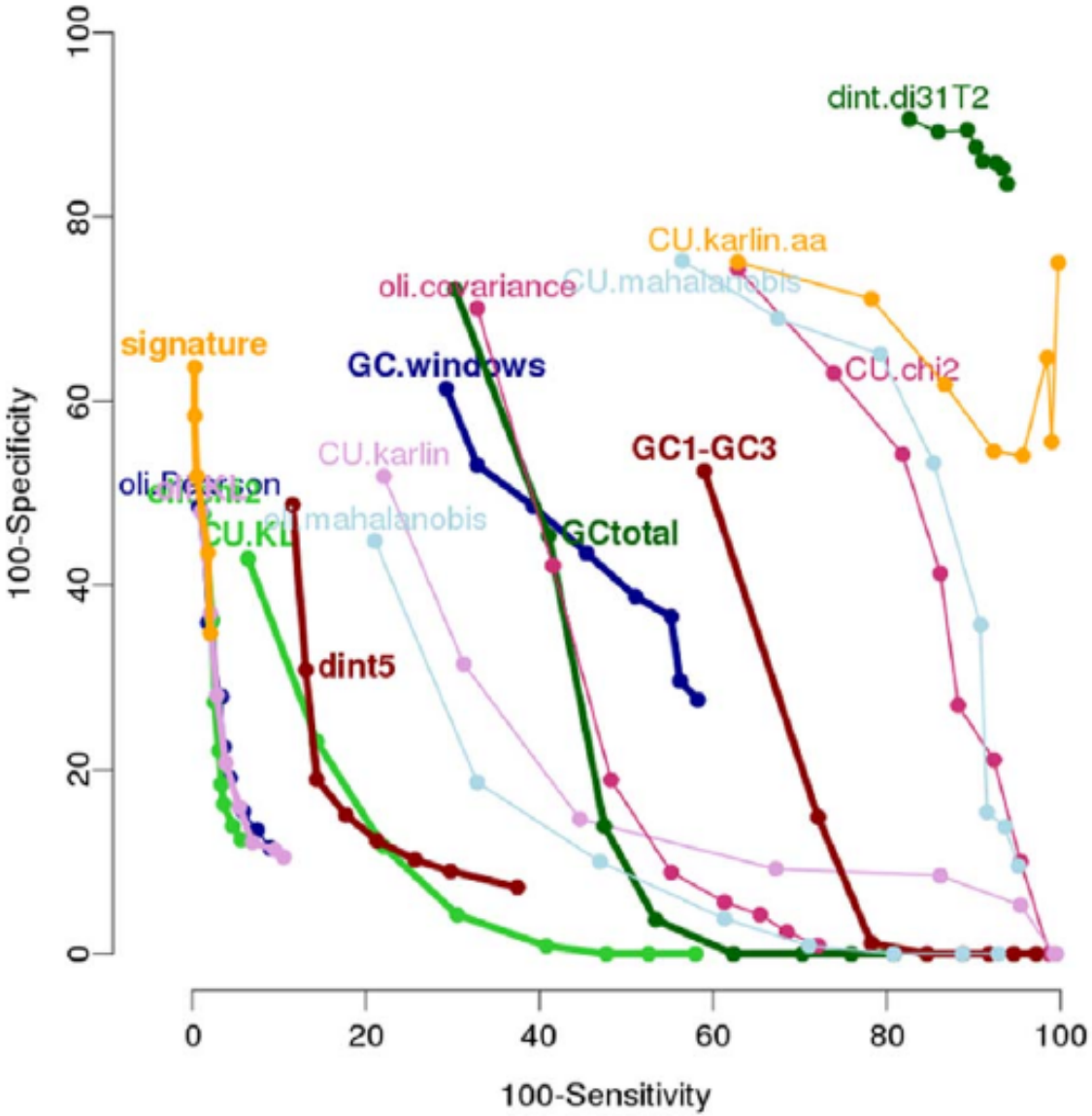
i.e. those introduced through horizontal transfer,

often bear unusual sequence characteristics

and thus can be distinguished from ancestral DNA»

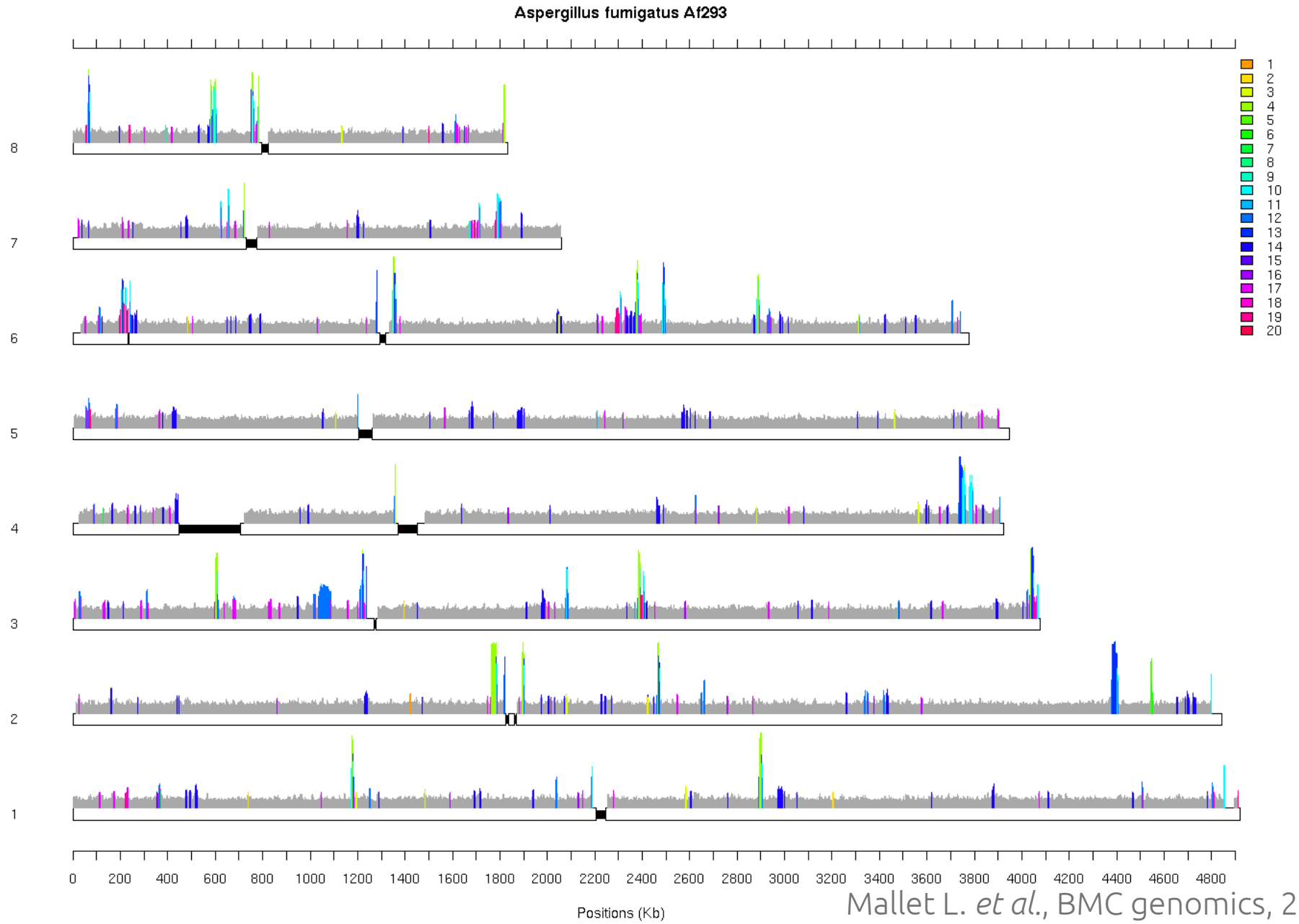
Lawrence & Ochman, *PNAS*, 1998

Benchmark of features and metrics



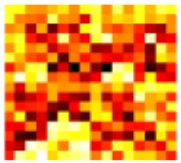
Becq *et al. Plos One* 2010

HGTs in *Aspergillus fumigatus*

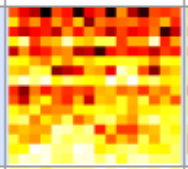
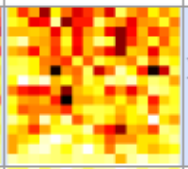
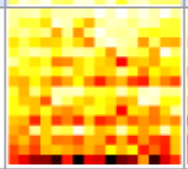
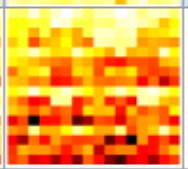
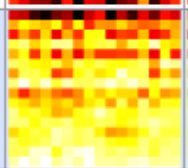
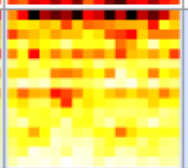
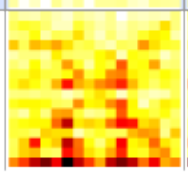


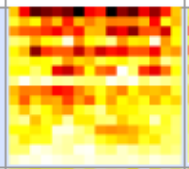
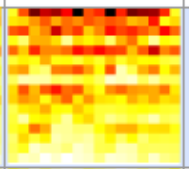
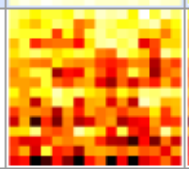
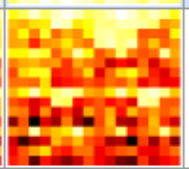
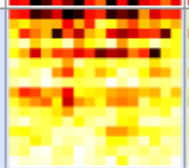
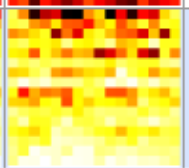
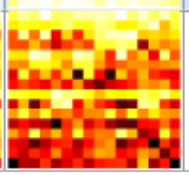
Mallet L. *et al.*, BMC genomics, 2010

Transfer likelihood and potential donors



Host genome signature

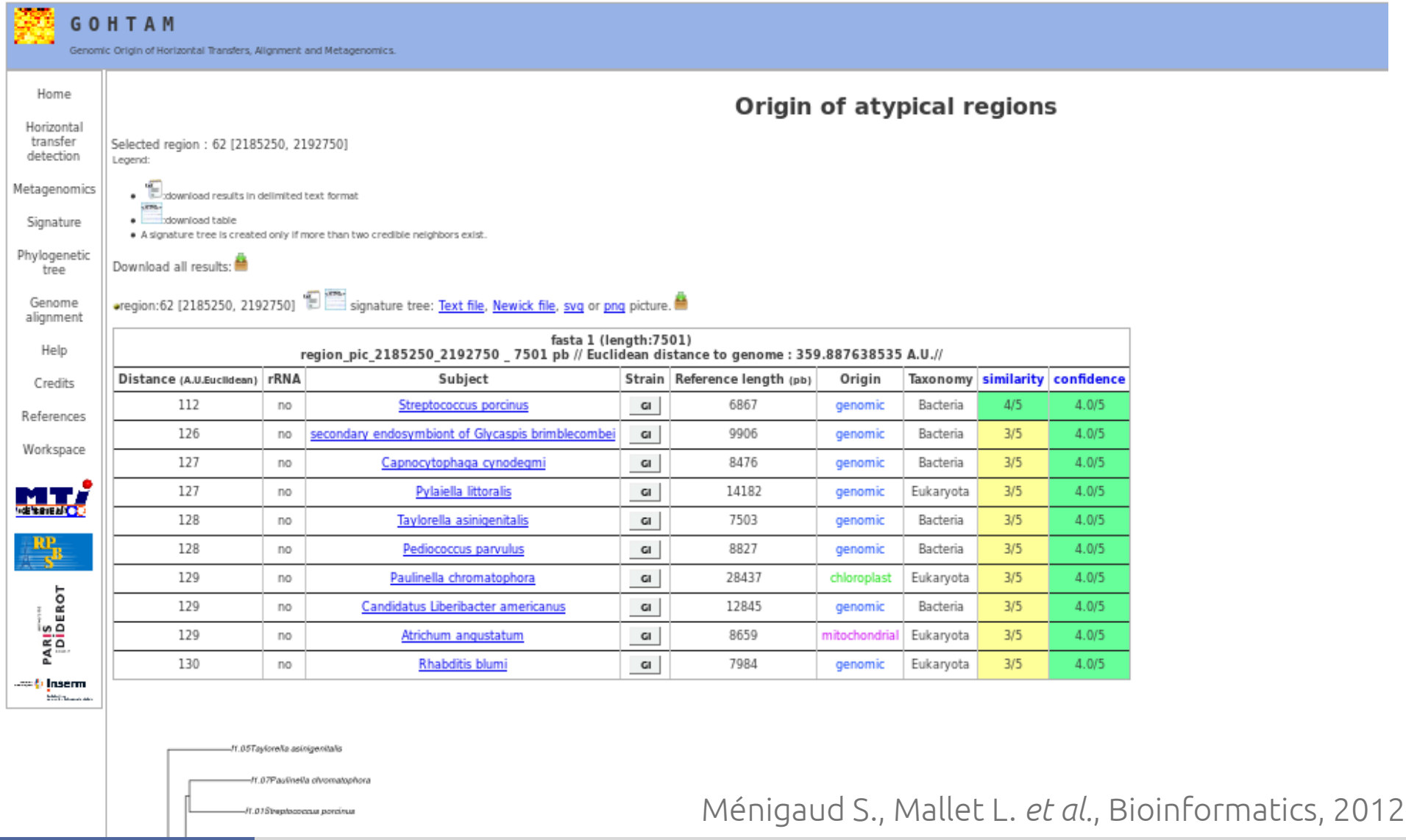
Atypical region (size)(pb)	Signature		Domain (distance)
	Region	Neighbor	
c1r25 (9500)			Eukaryotes (130)
c1r28 (5000)			Viruses (160)
c2r3 (4500)			Viruses (140)
c2r22 (4500)			Archae (190)

Atypical region (size)(pb)	Signature		Domain (distance)
	Region	Neighbor	
c3r17 (9000)			Bacteria (140)
c5r9 (4000)			Viruses (150)
c3r10 (3500)			Bacteria (170)
c3r9 (3500)			Bacteria (160)

Mallet L. *et al.*, *BMC Genomics*, 2010

GOHTAM (PROGECT)

- Database & webserver for genomic signature



GOHTAM

Genomic Origin of Horizontal Transfers, Alignment and Metagenomics.

Home
Horizontal transfer detection
Metagenomics
Signature
Phylogenetic tree
Genome alignment
Help
Credits
References
Workspace

Origin of atypical regions

Selected region : 62 [2185250, 2192750]
Legend:

- download results in delimited text format
- download table
- A signature tree is created only if more than two credible neighbors exist.

Download all results:

region:62 [2185250, 2192750] signature tree: Text file, Newick file, svg or png picture.

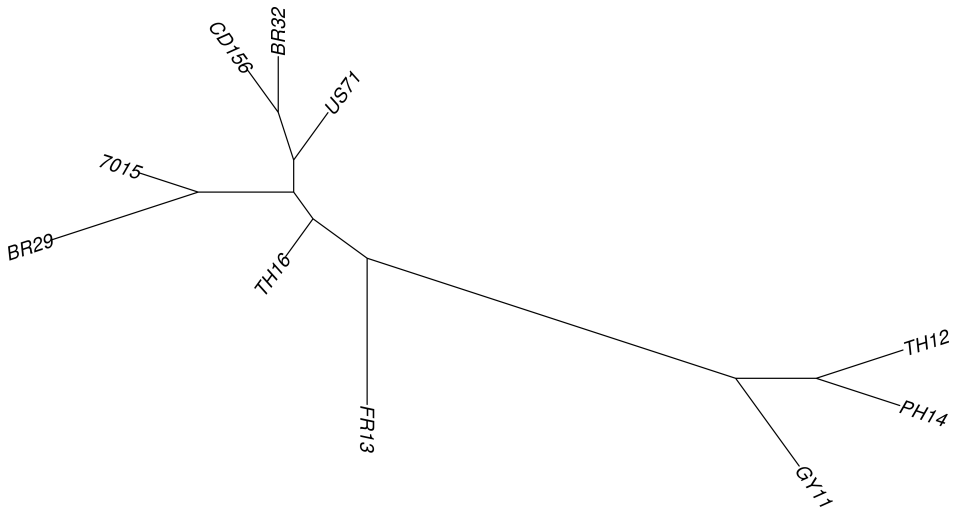
fasta 1 (length:7501)
region_pic_2185250_2192750 _ 7501 pb // Euclidean distance to genome : 359.887638535 A.U.//

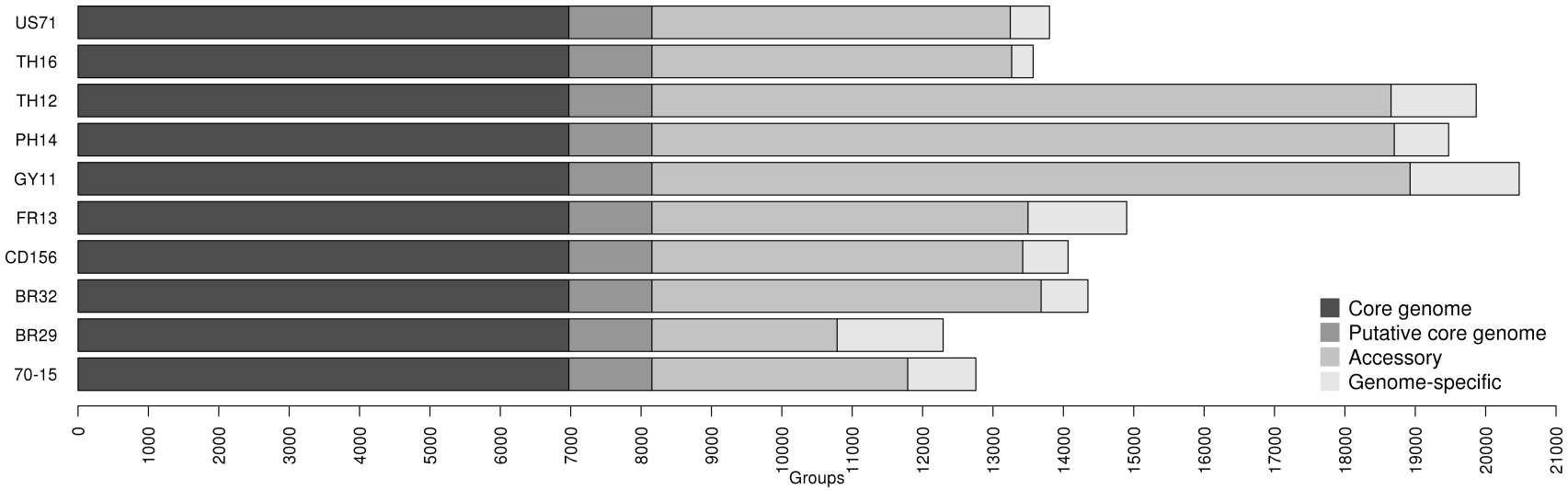
Distance (A.U.Euclidean)	rRNA	Subject	Strain	Reference length (pb)	Origin	Taxonomy	similarity	confidence
112	no	Streptococcus porcinus	GI	6867	genomic	Bacteria	4/5	4.0/5
126	no	secondary endosymbiont of Glycaspis brimblecombei	GI	9906	genomic	Bacteria	3/5	4.0/5
127	no	Capnocytophaga cynodegmi	GI	8476	genomic	Bacteria	3/5	4.0/5
127	no	Pylaiella littoralis	GI	14182	genomic	Eukaryota	3/5	4.0/5
128	no	Taylorella asinigenitalis	GI	7503	genomic	Bacteria	3/5	4.0/5
128	no	Pediococcus parvulus	GI	8827	genomic	Bacteria	3/5	4.0/5
129	no	Paulinella chromatophora	GI	28437	chloroplast	Eukaryota	3/5	4.0/5
129	no	Candidatus Liberibacter americanus	GI	12845	genomic	Bacteria	3/5	4.0/5
129	no	Atrichum angustatum	GI	8659	mitochondrial	Eukaryota	3/5	4.0/5
130	no	Rhabditis blumi	GI	7984	genomic	Eukaryota	3/5	4.0/5

Ménigaud S., Mallet L. *et al.*, Bioinformatics, 2012

Suspicion

- OrthoMCL
- Core genome = singletons
- Putative = All - 1
- Accessory = 2 ∼ 8





Need for tools, efficient ones

Blobology - Kumar S. *et al.*, Front Genet. 2013 $\sim$ 20000 contigs

MetaWatt - Strous M., *et al.*, Front Microbiol. 2012 struggles with eukaryotes

Our own scripts - Of course

- Take on metagenomics
- Take on long noisy reads

pro·gram·mer

n. an organism that turns caffeine into software

github.com/itsmeludo

PhylOligo - Olignonucleotide frequencies and distance matrix

PhyloSelect - Identify prototype sequences from species mixes

- PhyloSelect.R - NJ tree and interactive exploration of branching
- PhyloSelect.py - clustering (hdbscan) qnd visualisation (t-SNE)

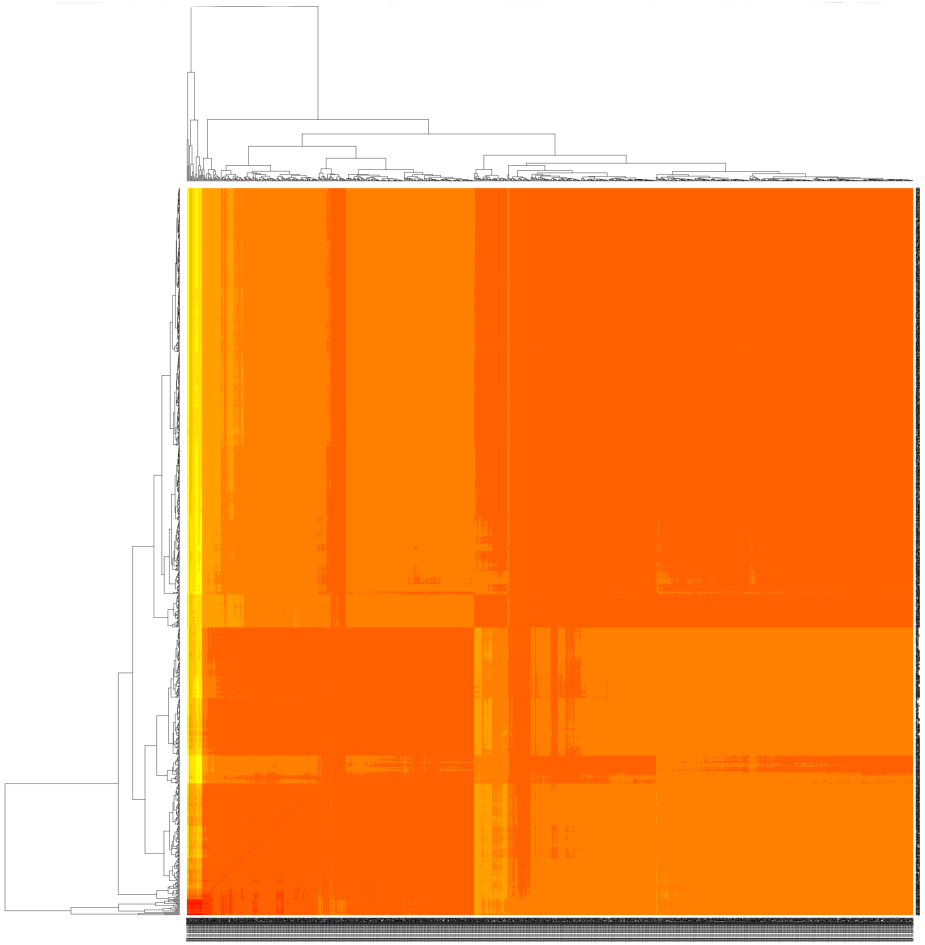
Contalocate - Learns composition and locates species-specific regions

Distance matrix

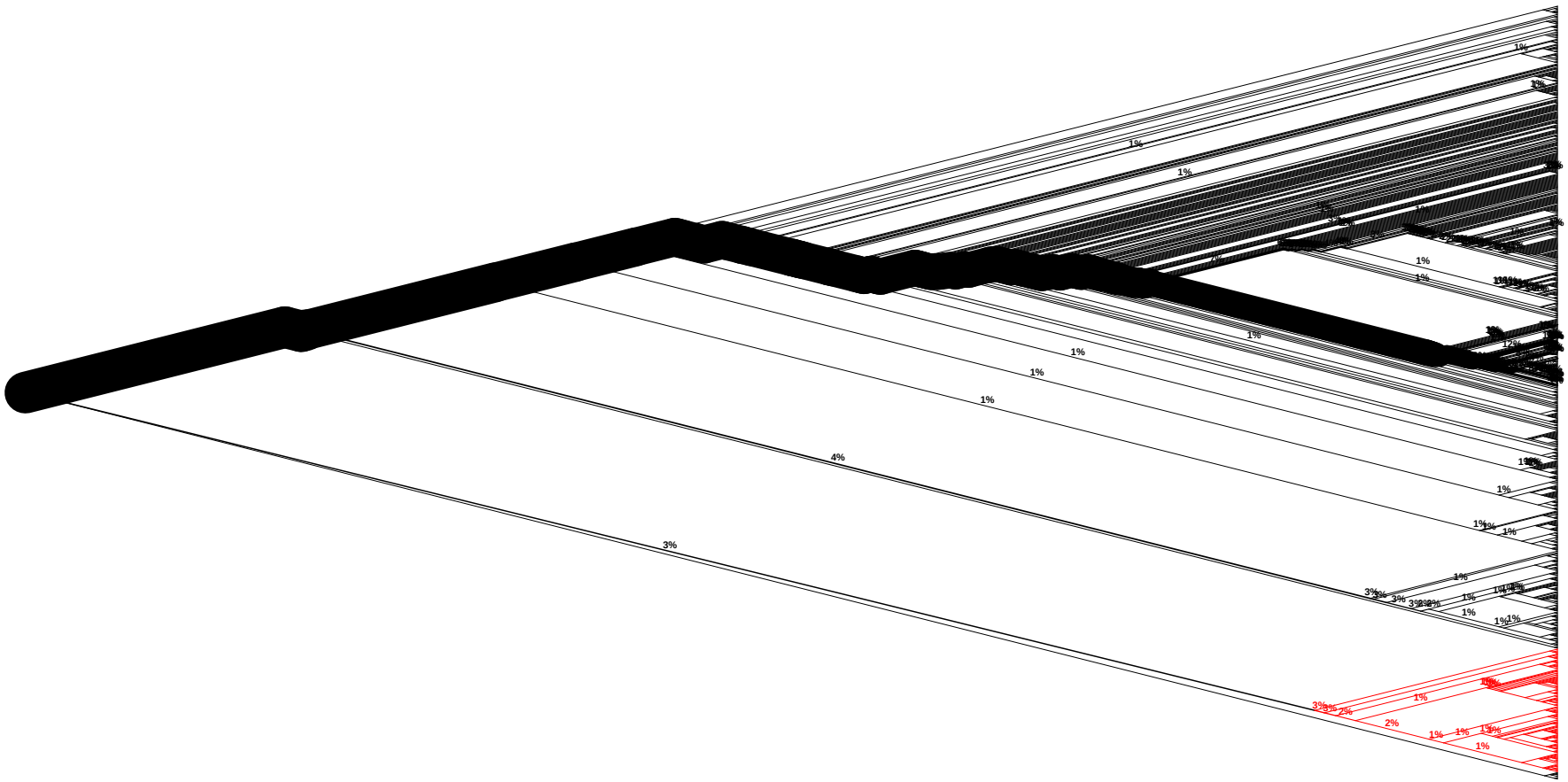
Genomic signature of each contig
All vs. all contigs
Euclidean distance or Jensen-Shannon divergence

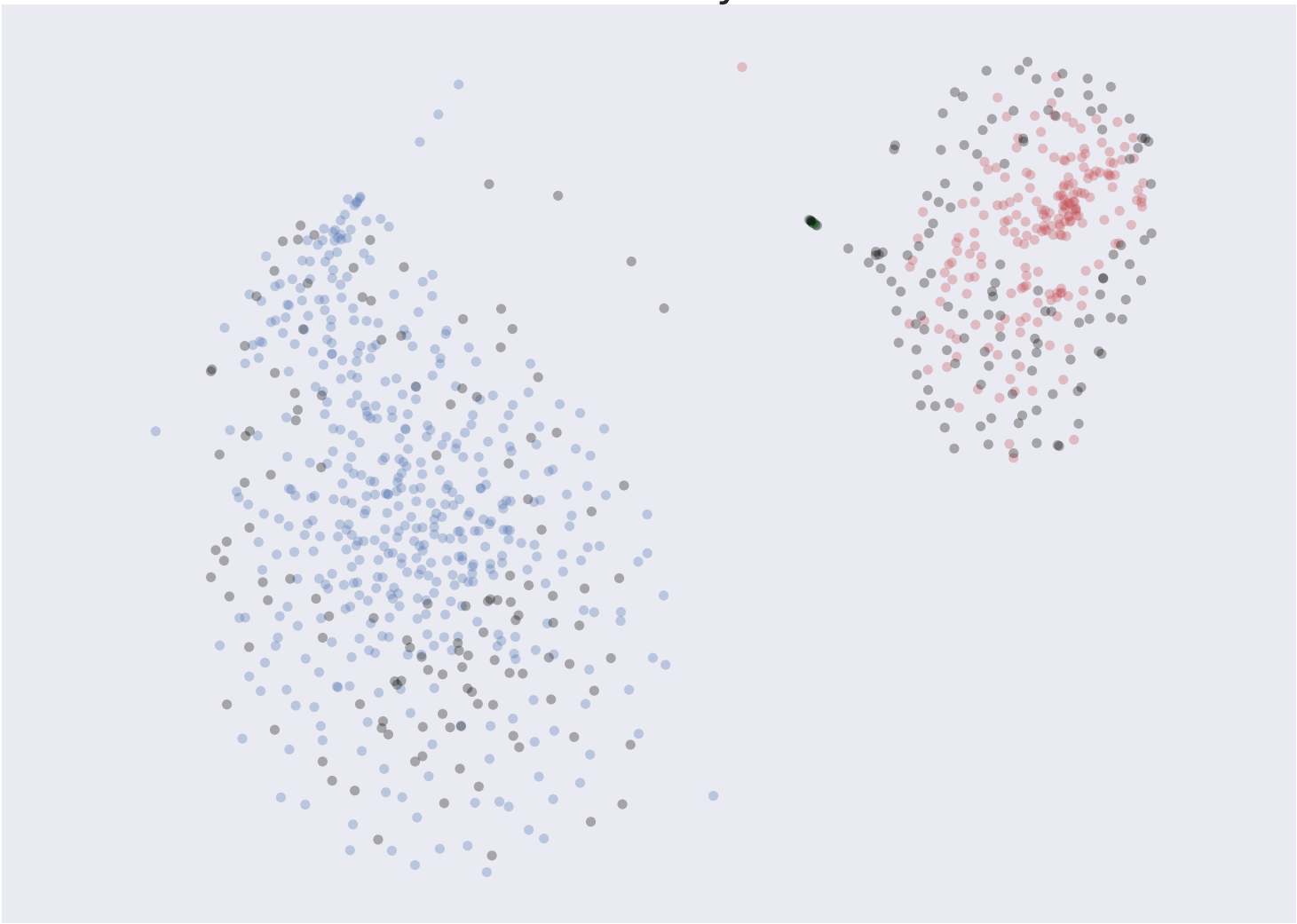
Distance matrix

Genomic signature of each contig
All vs. all contigs
Euclidean distance or Jensen-Shannon divergence



PhyloSelect.R




Clusters found by hdbscan

Sliding window genome scan

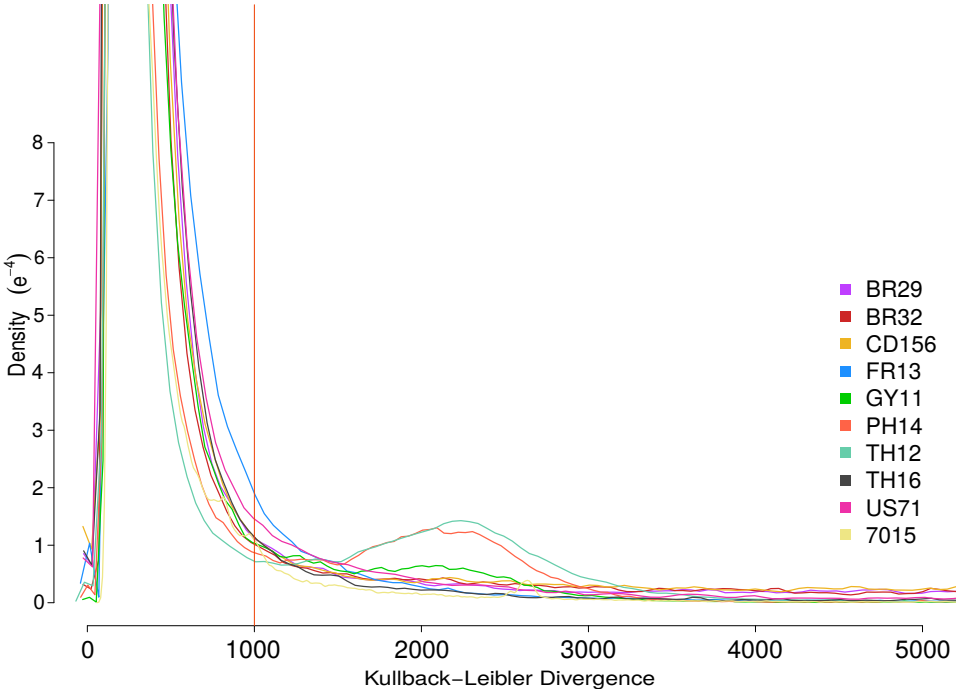
Genomic signature of 5kb windows, 100bp step

JSD or Euclidean distances

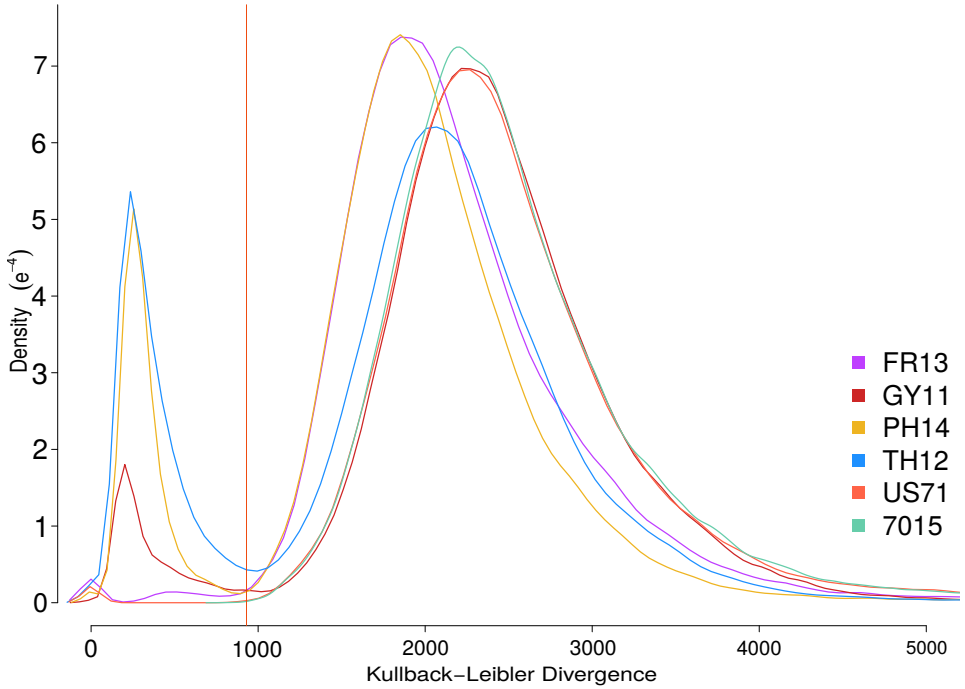
- Host subset vs. whole assembly
- Conta subset vs. whole assembly

Distribution of distances $\Rightarrow$ double thresholds

Thresholds determination



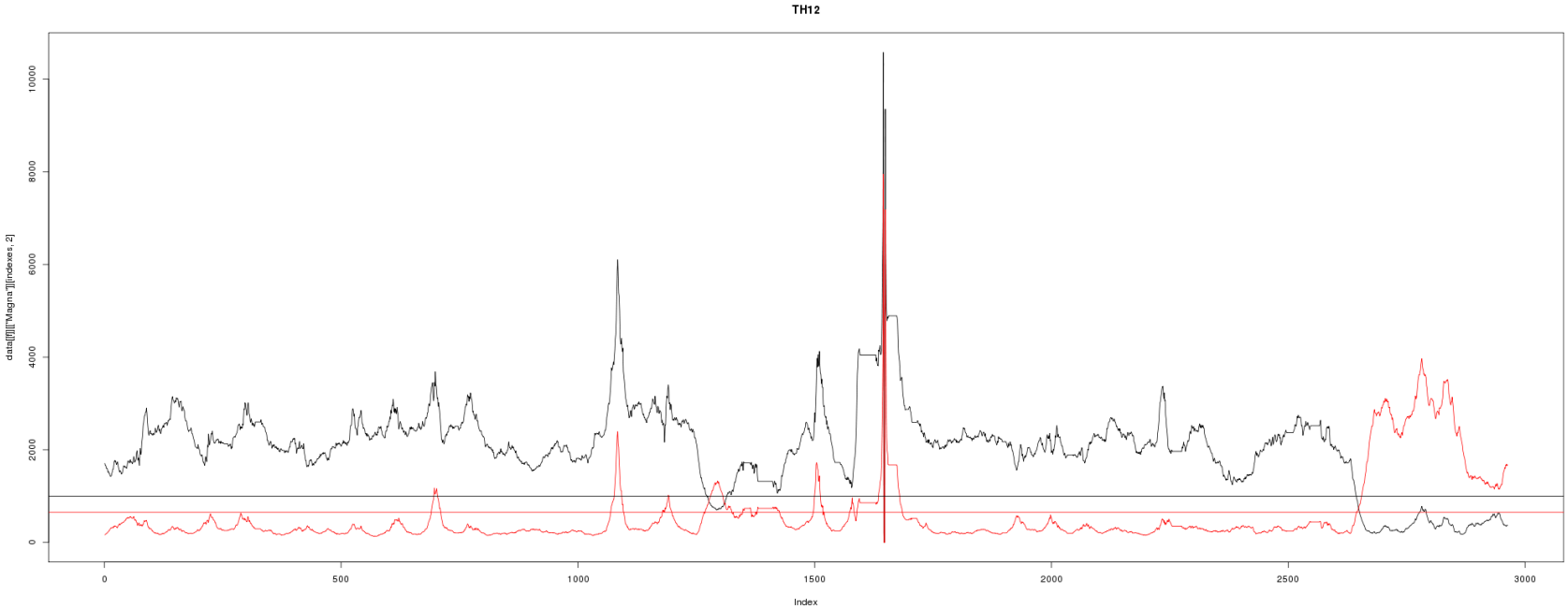
Distance Kullback leibler: 5Kb windows over scaffolds v.s. Prototype of *Magnaporthe*



Distance Kullback leibler: 5Kb windows over scaffolds v.s. prototype of *Burkholderia*

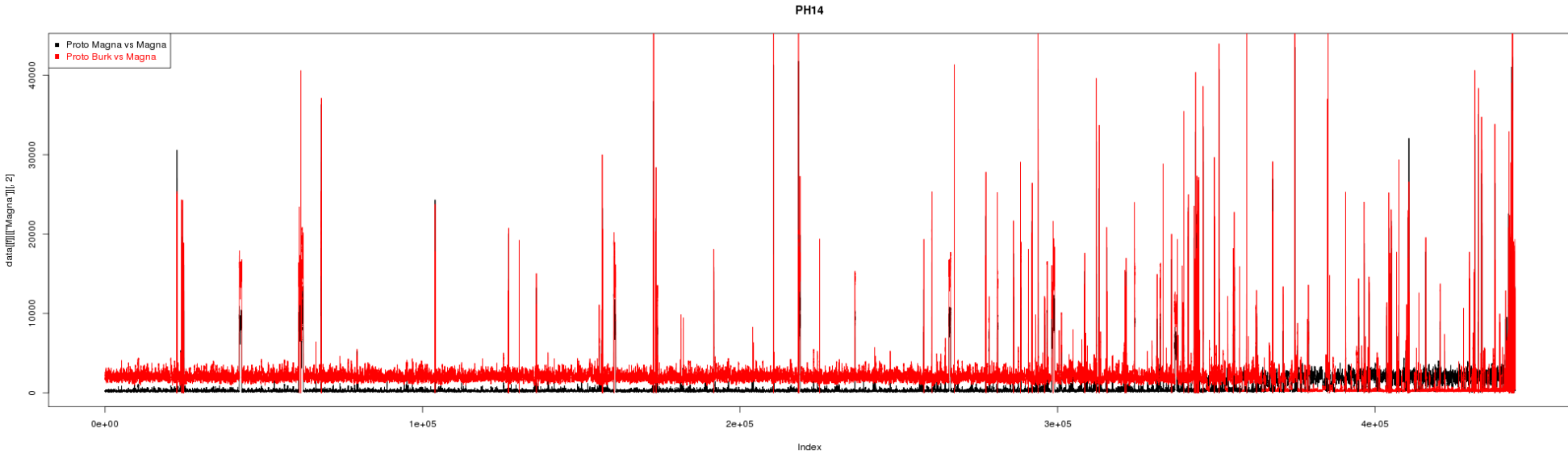
Chiapello H., *et al.*, GBE 2015

Example of filtering



red: distance to the *Burkholderia* prototype
black: distance to the *Magnaporthe* prototype

Example of filtering

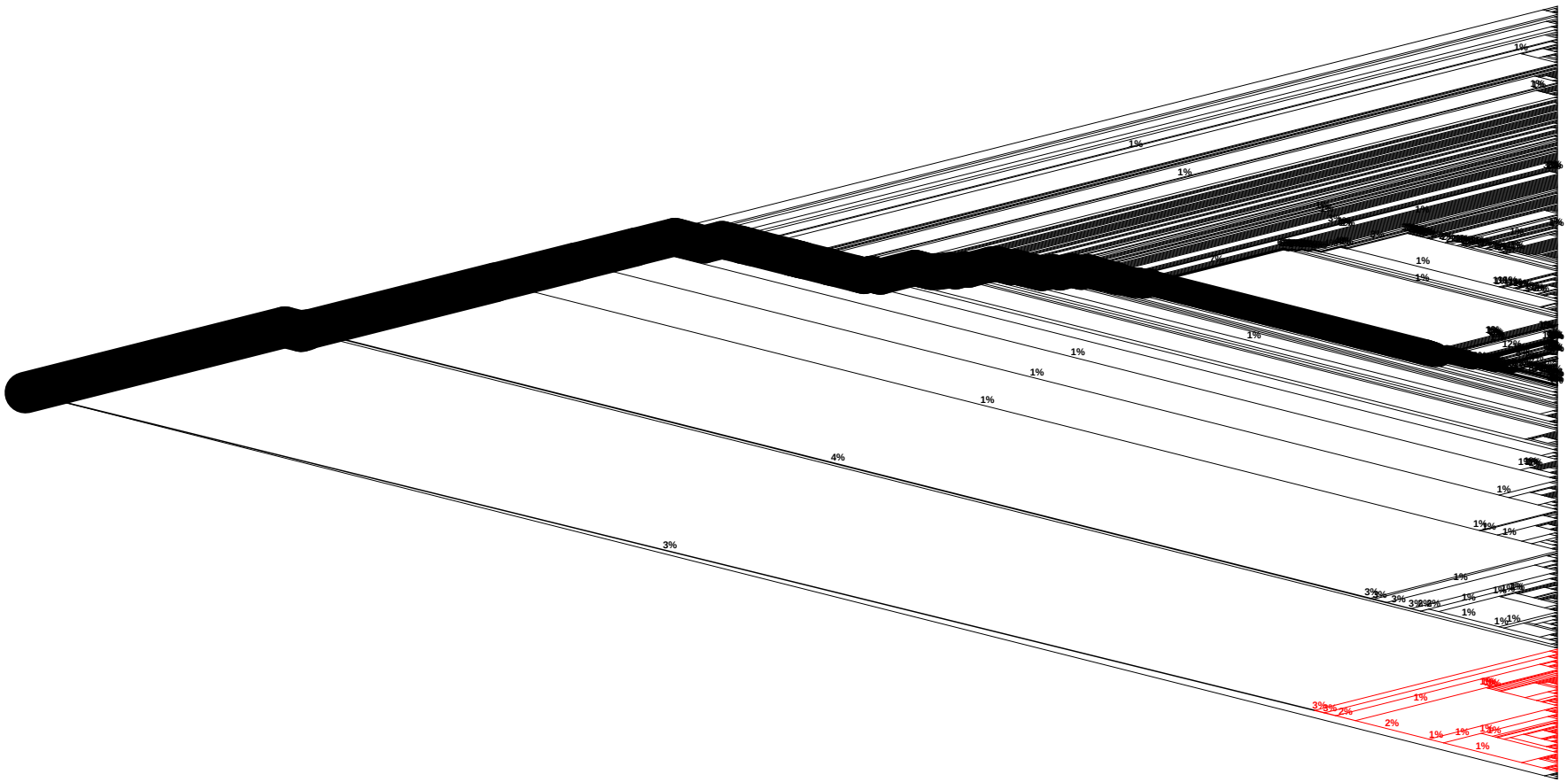


red: distance to the *Burkholderia* prototype
black: distance to the *Magnaporthe* prototype

Examples

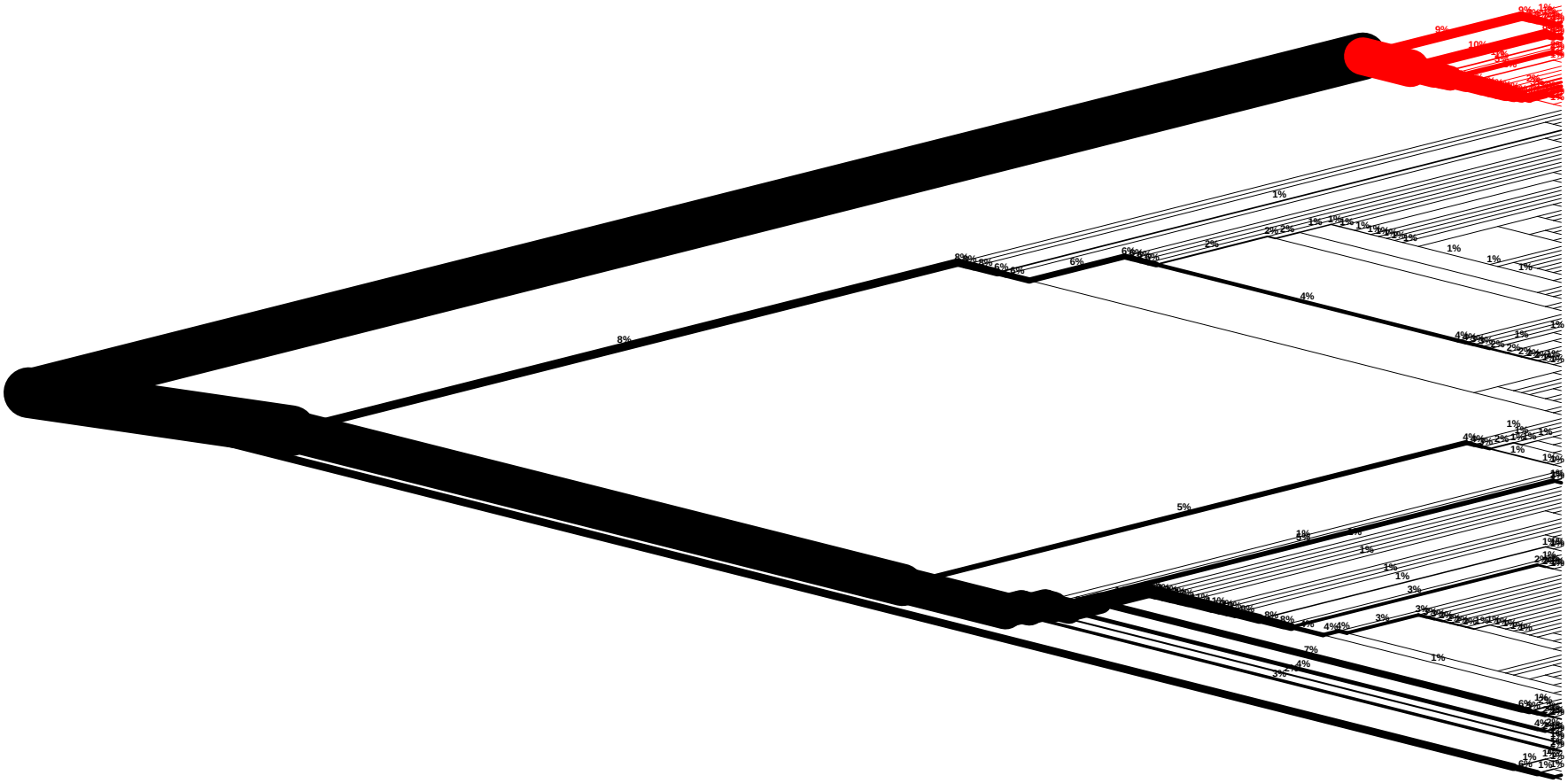
Magnaporthe oryzae TH12

Phytopathogen fungus. ~40Mb

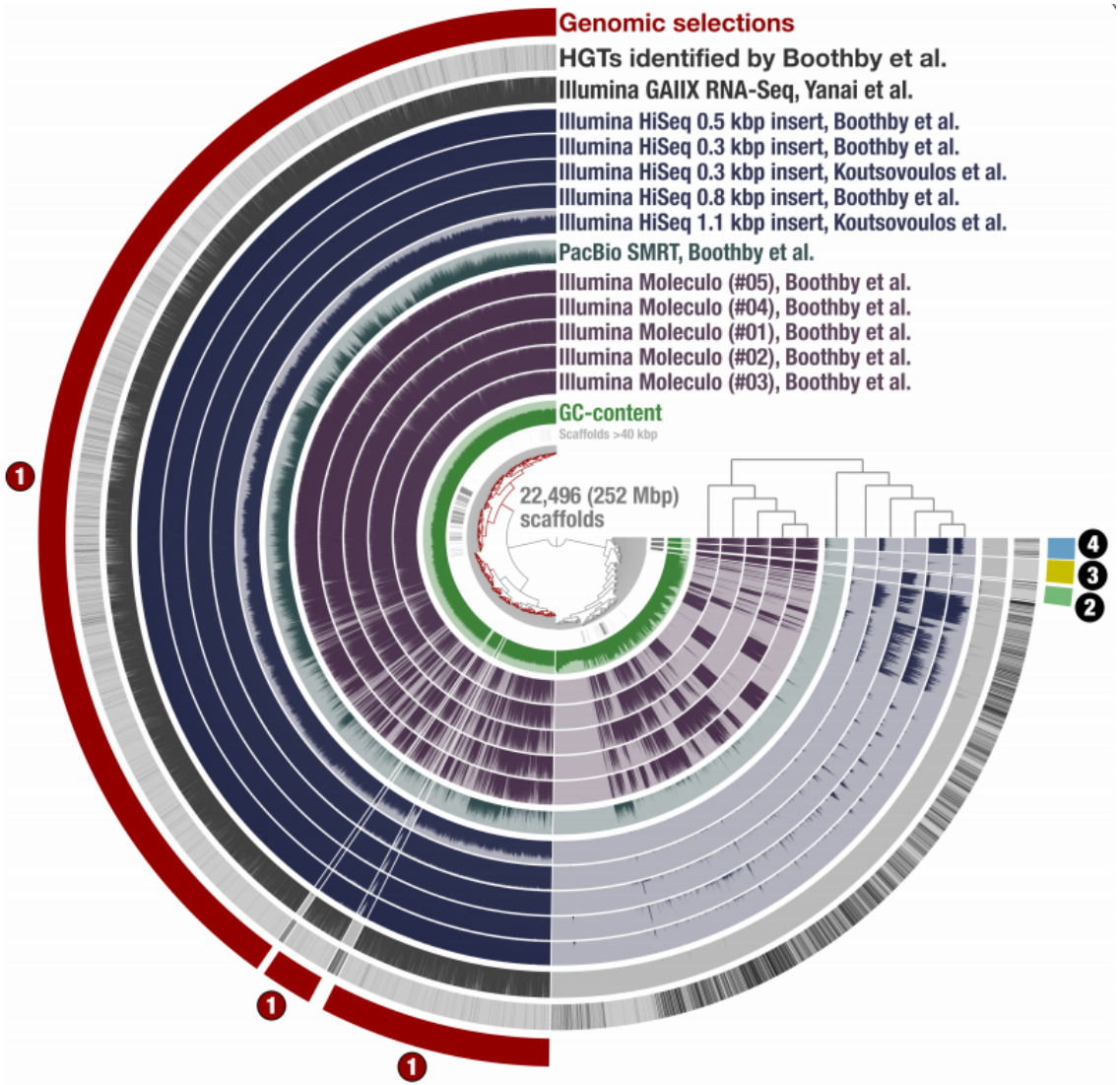


Ganoderma lucidum

Fungus used in traditional chinese medicine. ~43Mb



Hypsibius dujardini - The infamous Tardigrade



1 ***Hypsibius dujardini* curated draft genome. 182.2 Mbp** (14,961 scaffolds).

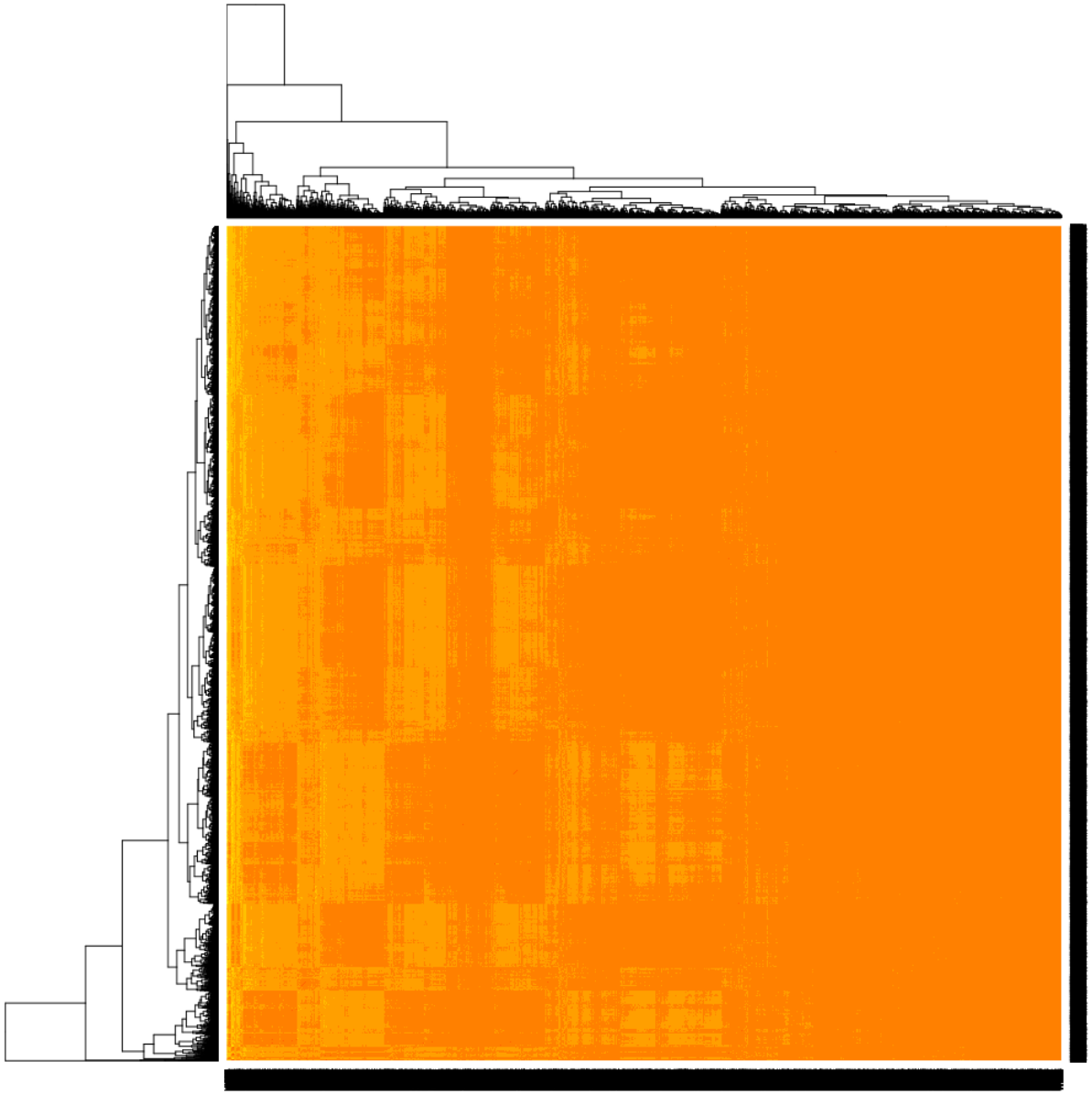
2 **Bacterial draft genome #1. 4.8 Mbp** (4 scaffolds; 100% complete with 5.9% redundancy. RAST Taxonomy: *Chitinophaga pinensis*).

3 **Bacterial draft genome #2. 4.5 Mbp** (29 scaffolds; 97% complete with 0% redundancy. RAST Taxonomy: *Chitinophaga pinensis*).

4 **Bacterial draft genome #3. 3.8 Mbp** (5 scaffolds; 97% complete with 5.9% redundancy. RAST Taxonomy: *Thermosinus carboxydivorans*).

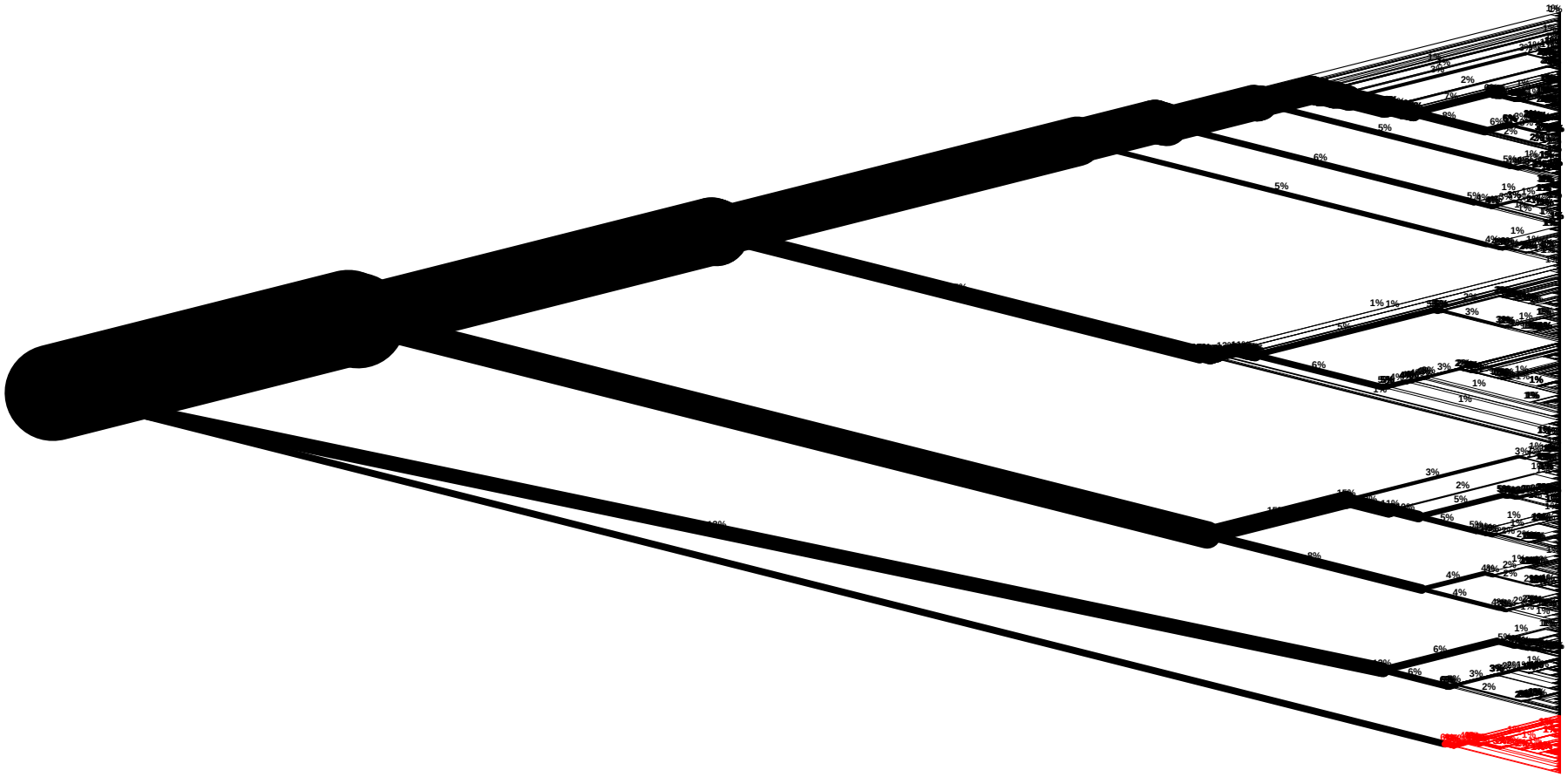
Hypsibius dujardini

Survival winner. Genome size debated 55∼350 Mb



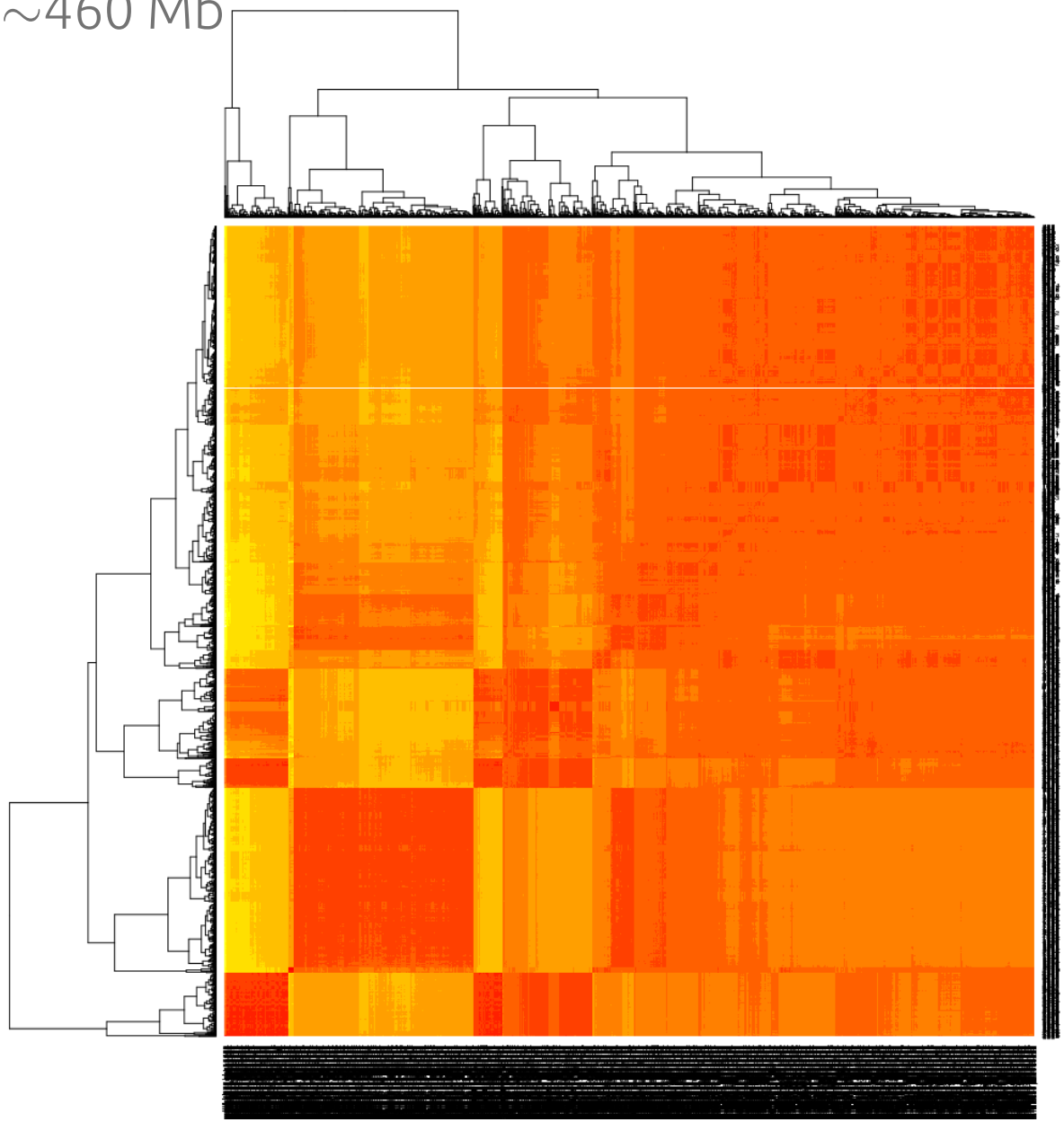
Hypsibius dujardini

Survival winner, Genome size debated 55∼350 Mb



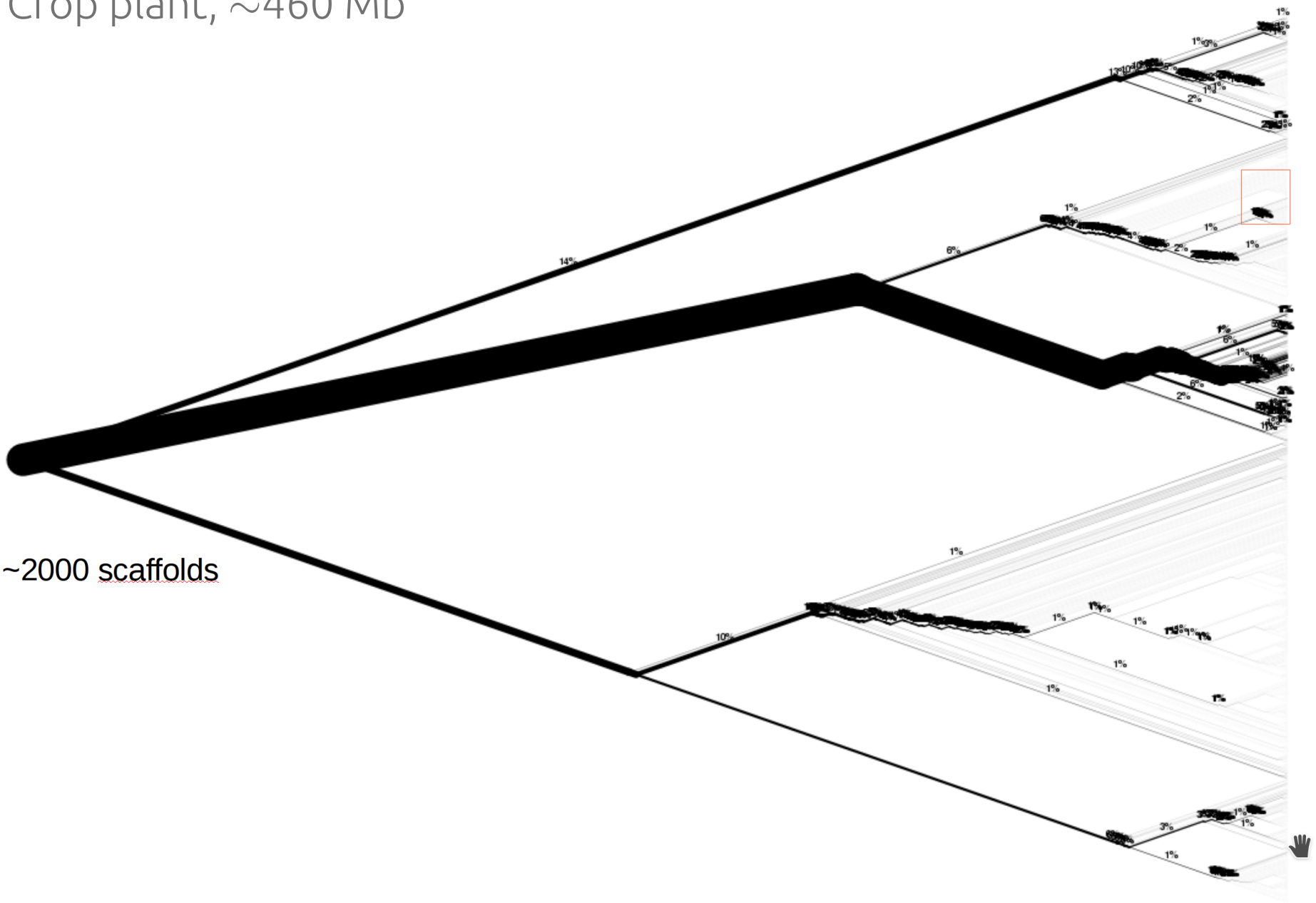
Aeschynomene evenia

Crop plant, ~460 Mb



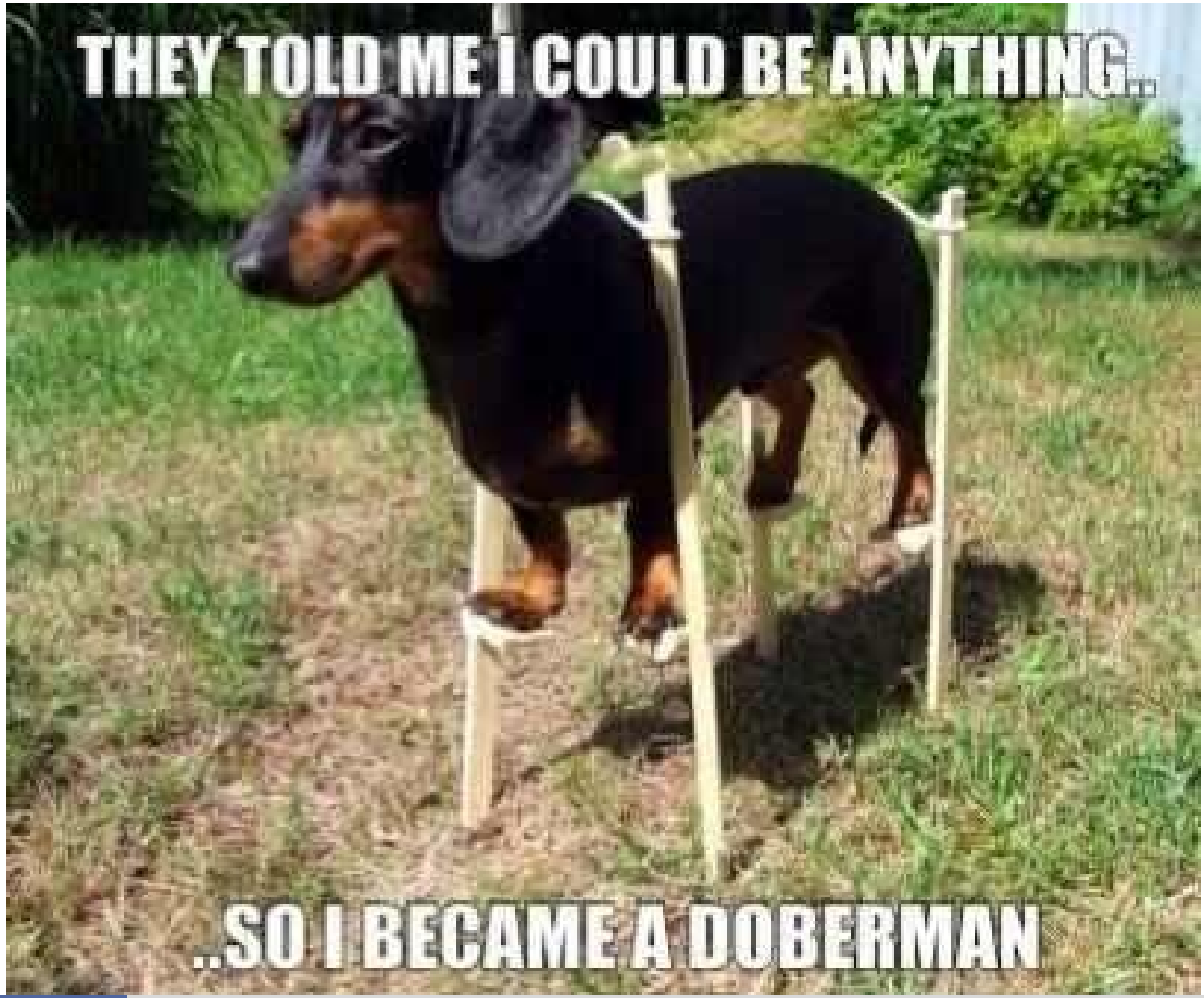
Aeschynomene evenia

Crop plant, ~460 Mb



~2000 scaffolds

Same player go again



PacBio Metagenome

Human Microbiome Project MockB Shotgun

Benchmark synthetic sample Set 1 (7 SMRT Cells using P4-C2) (43 Gb tar.gz)

- 20 known species
- Equimolar ribosomal RNA operon counts

~ 350000 raw reads (13~17% error)

PhylOligo ran. It dumped a 450G compressed matrix.

PhylSelect.py awaits its turn on the cluster

At last

Almost completely functional toolbox

Highly efficient implementation

Work on assembled data

- Filter species-specific regions (keeps HGTs in place)
- Do not fragment the assembly at repeated regions
- Can be launched iteratively for multiple contaminations

Assemblies might exhibit complex structures

PacBio might makes it harder to find contaminants

Might work on unassembled PacBio data

Admins hate it, click here to know why!

Go back to work!