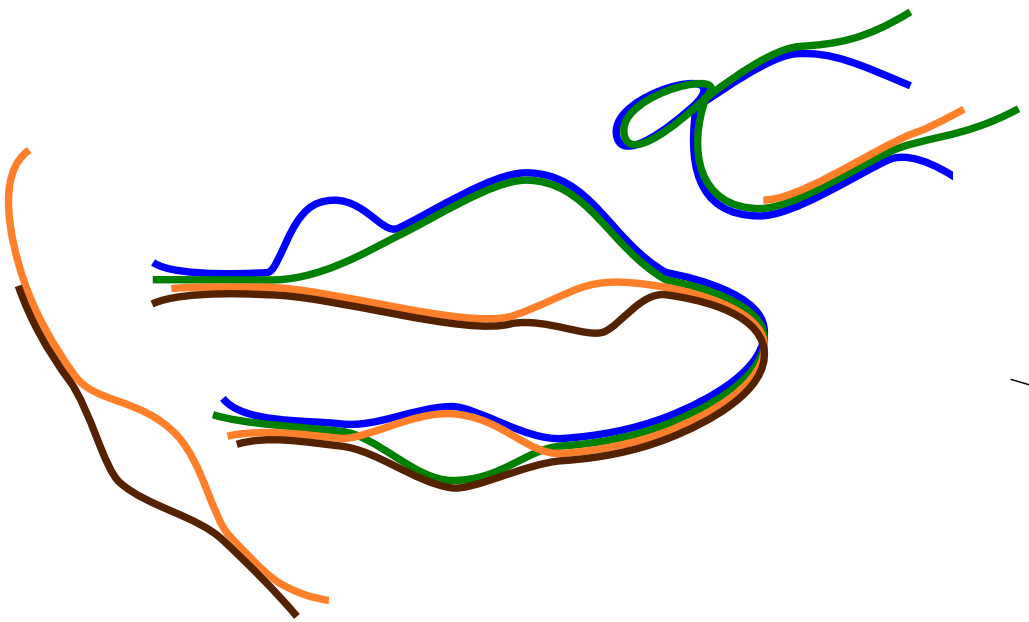
SANS serif: alignment-free, whole-genome based phylogenetic reconstruction

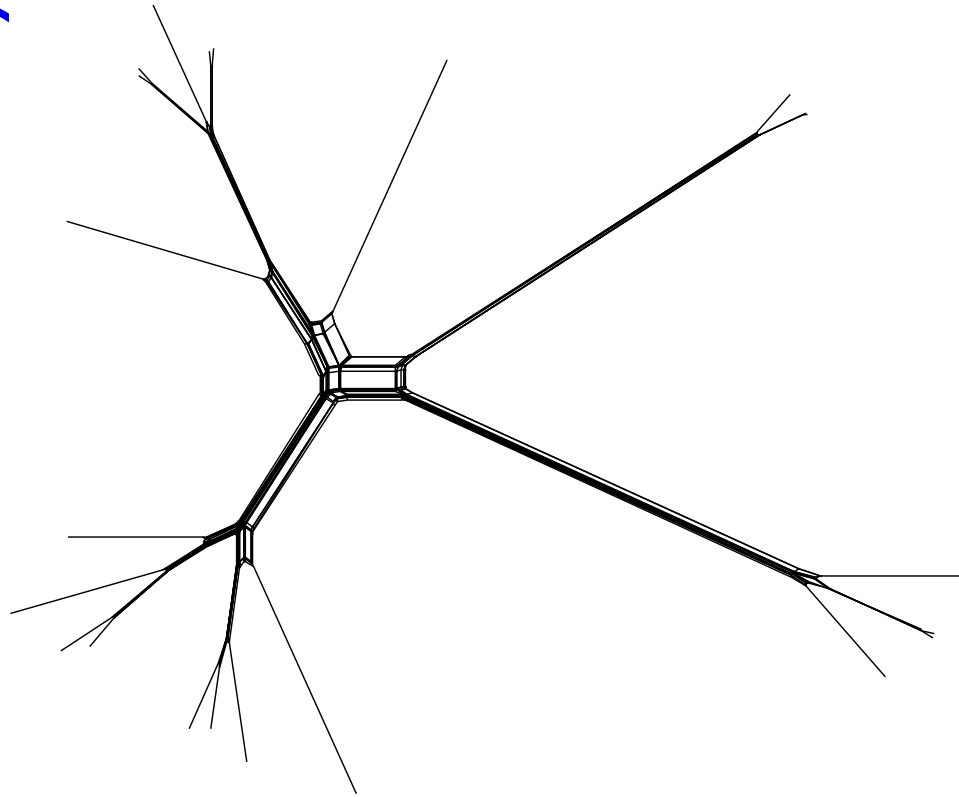
Roland Wittler, Andreas Rempel, Marco Sohn

Bielefeld University

DSB 2021

Outline





Introduction — SANS

Classical approaches

- mult. alignment of marker genes
- alignment to reference $\rightarrow$ SNPs
- pairwise distances or ML
- tree

Introduction — SANS

Classical approaches

- mult. alignment of marker genes
- alignment to reference $\rightarrow$ SNPs
- pairwise distances or ML
- tree

Whole genome

- k-mers,
  spaced k-mers, ...
- pairwise distances
- tree

Introduction — SANS

Classical approaches

- mult. alignment of marker genes
- alignment to reference $\rightarrow$ SNPs
- pairwise distances or ML
- tree

Whole genome

- k-mers,
  spaced k-mers, ...
- pairwise distances
- tree

SANS: Symmetric Alignment-free phylogeNomic Splits

- whole genome (no markers, no alignment, no reference)

Introduction — SANS

Classical approaches

- mult. alignment of marker genes
- alignment to reference $\rightarrow$ SNPs
- pairwise distances or ML
- tree

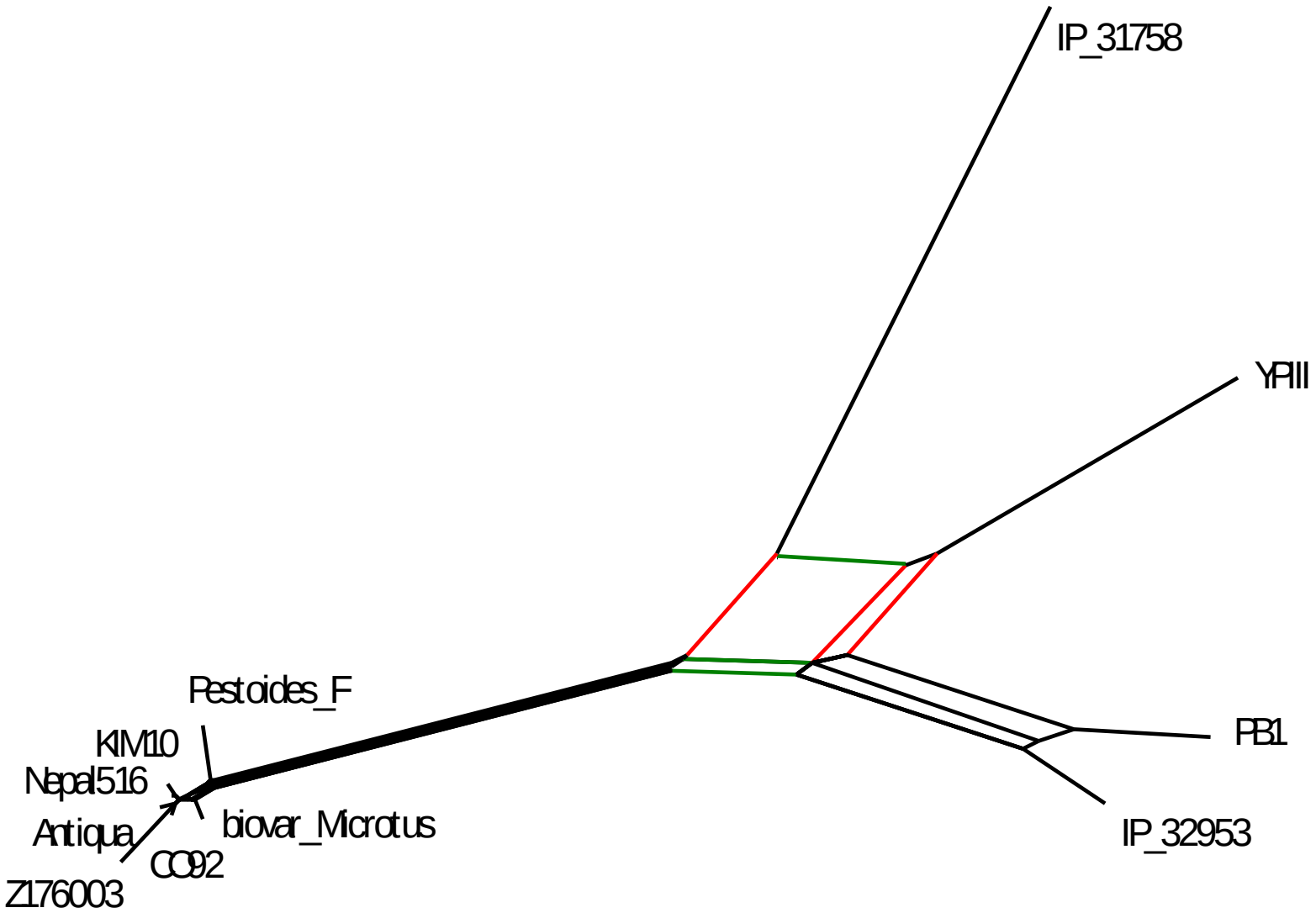
Whole genome

- k-mers,
  spaced k-mers, ...
- pairwise distances
- tree

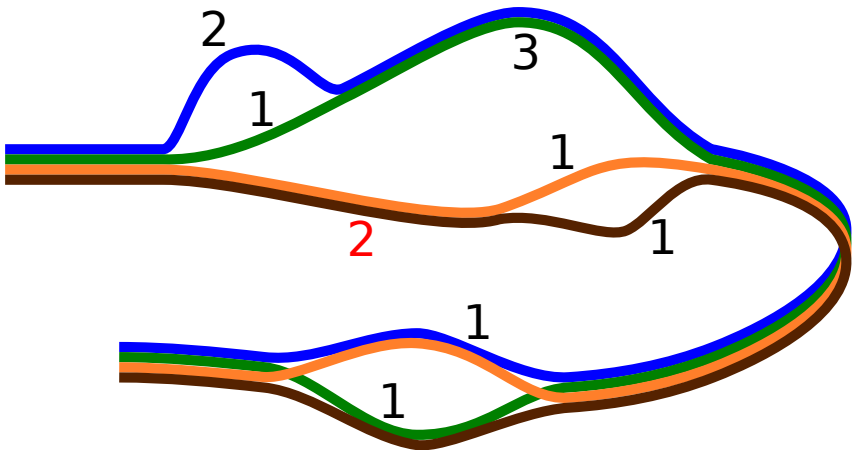
SANS: Symmetric Alignment-free phylogeNomic Splits

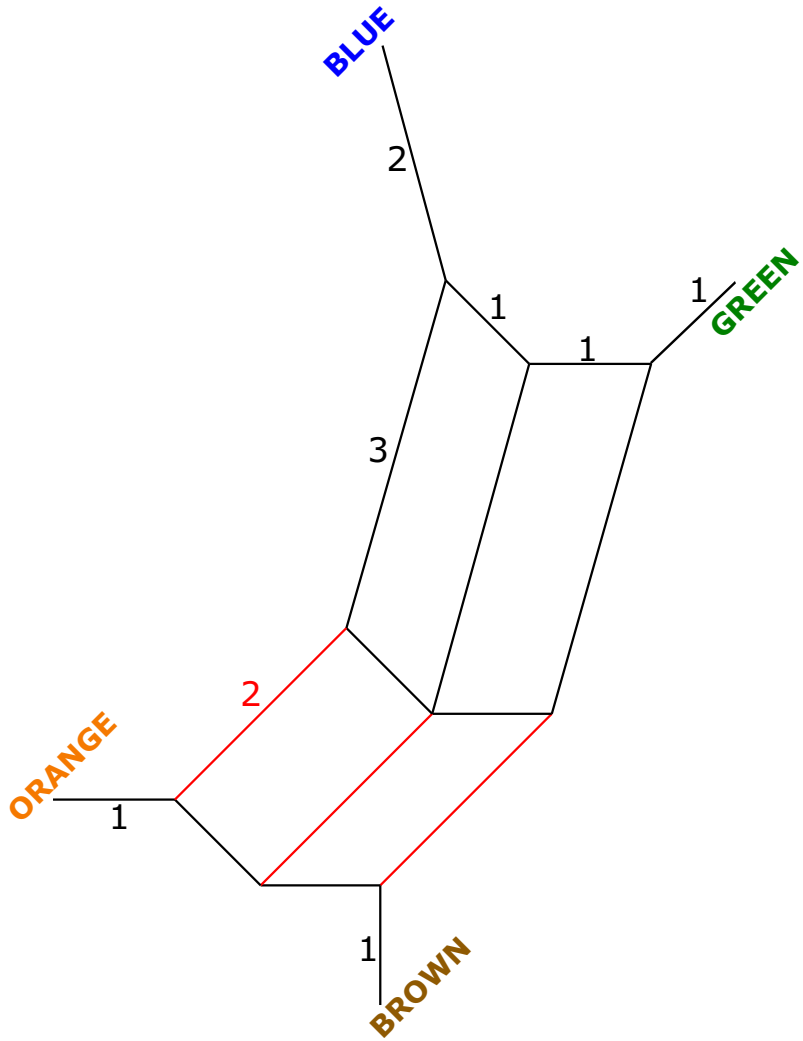
- whole genome (no markers, no alignment, no reference)
- k-mers $\rightarrow$ splits (no pairwise comparison)

Introduction — SANS

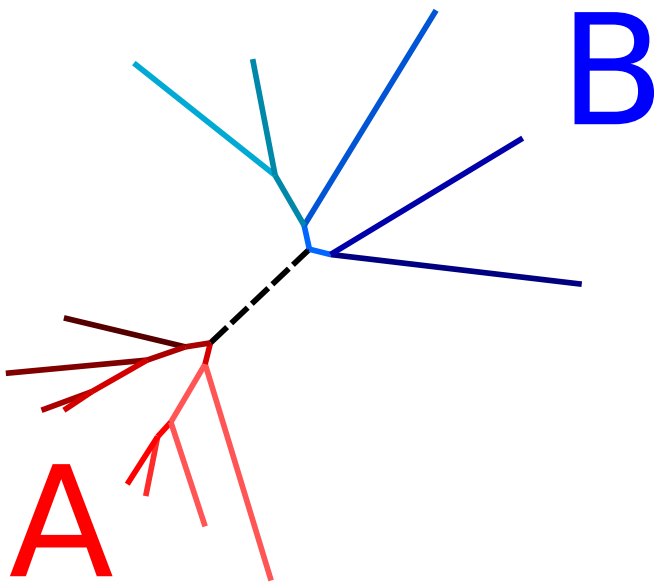


Introduction — SANS





Introduction — SANS



A and B separated in phylogeny

$\Rightarrow$ mutations on that edge

$\Rightarrow$ sequences unique to A and
sequences unique to B

$\Rightarrow$ A and B k-mers
of about same total amount $w \approx w'$

$\Rightarrow$ $\sqrt{w \cdot w'}$ to downweight asymmetry
or $\sqrt{(w+1) \cdot (w'+1)}$ to not loose any

Outline

SANS serif

Andreas Rempel

First SANS version [DSB 2019, WABI 2019]:

1. build C-DBG (Bifrost, Melsted&Holley)
2. extract splits from unitigs and store in trie

Re-implementation:

- hash tables
- no dependencies
- space and time efficient
- includes filtering of splits (e.g. tree filter)

First SANS version [DSB 2019, WABI 2019]:

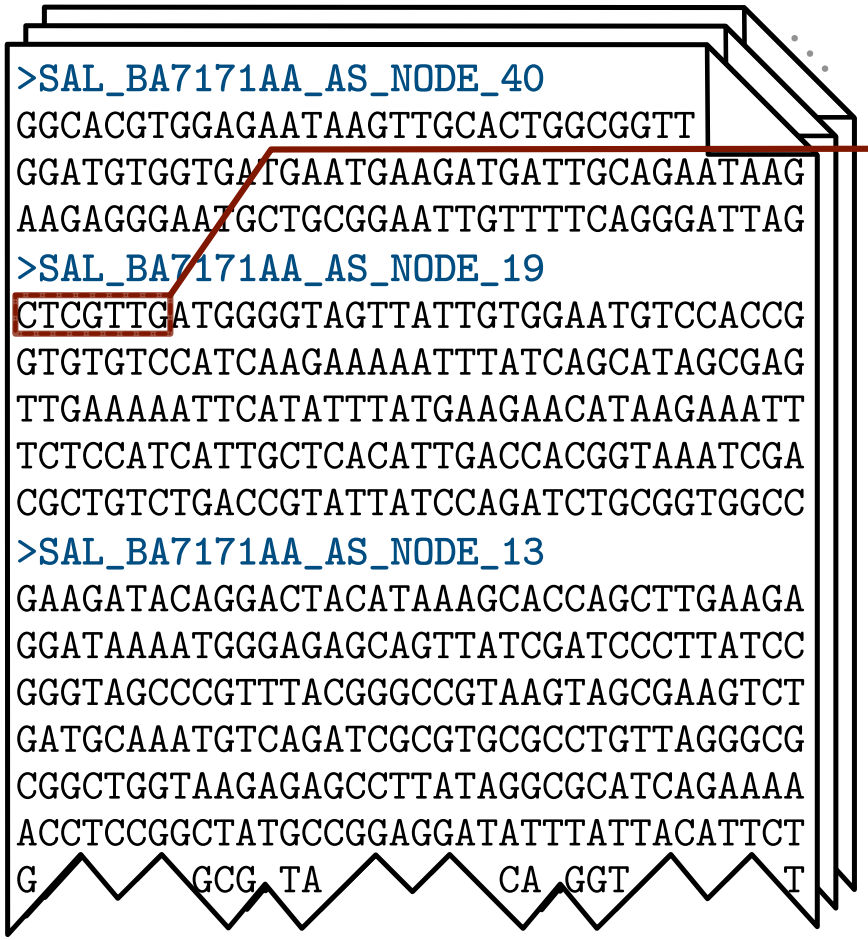
1. build C-DBG (Bifrost, Melsted&Holley)
2. extract splits from unitigs and store in trie

Space and time **E**fficient **R**e-**I**mplementation incl. **F**ilters: SANS **serif**

- hash tables
- no dependencies
- space and time efficient
- includes filtering of splits (e.g. tree filter)

SANS serif

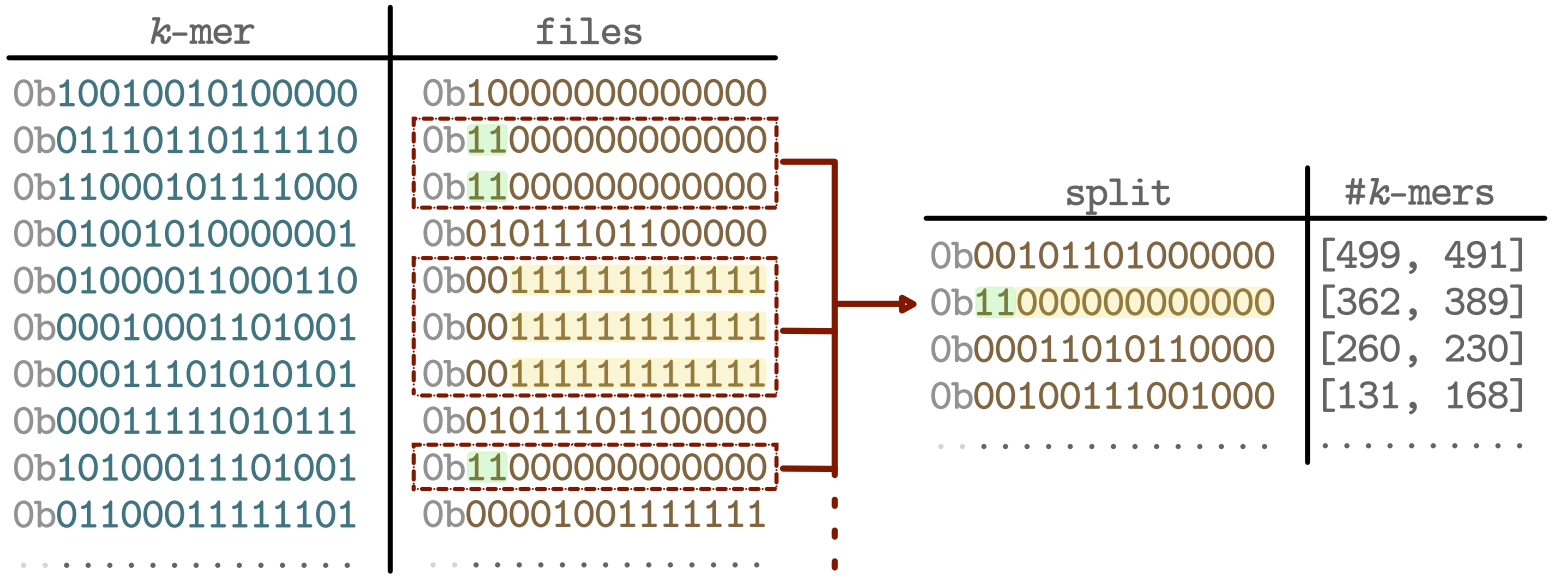
Andreas Rempel



k-mer	files
0b10010010100000	0b10000000000000
0b01110110111110	0b11000000000000
0b11000101111000	0b11000000000000
0b01001010000001	0b01011101100000
0b01000011000110	0b00111111111111
0b00010001101001	0b00111111111111
0b00011101010101	0b00111111111111
0b00011111010111	0b01011101100000
0b10100011101001	0b11000000000000
0b01100011111101	0b00001001111111
. . . . .	. . . . .

SANS serif

Andreas Rempel



SANS serif

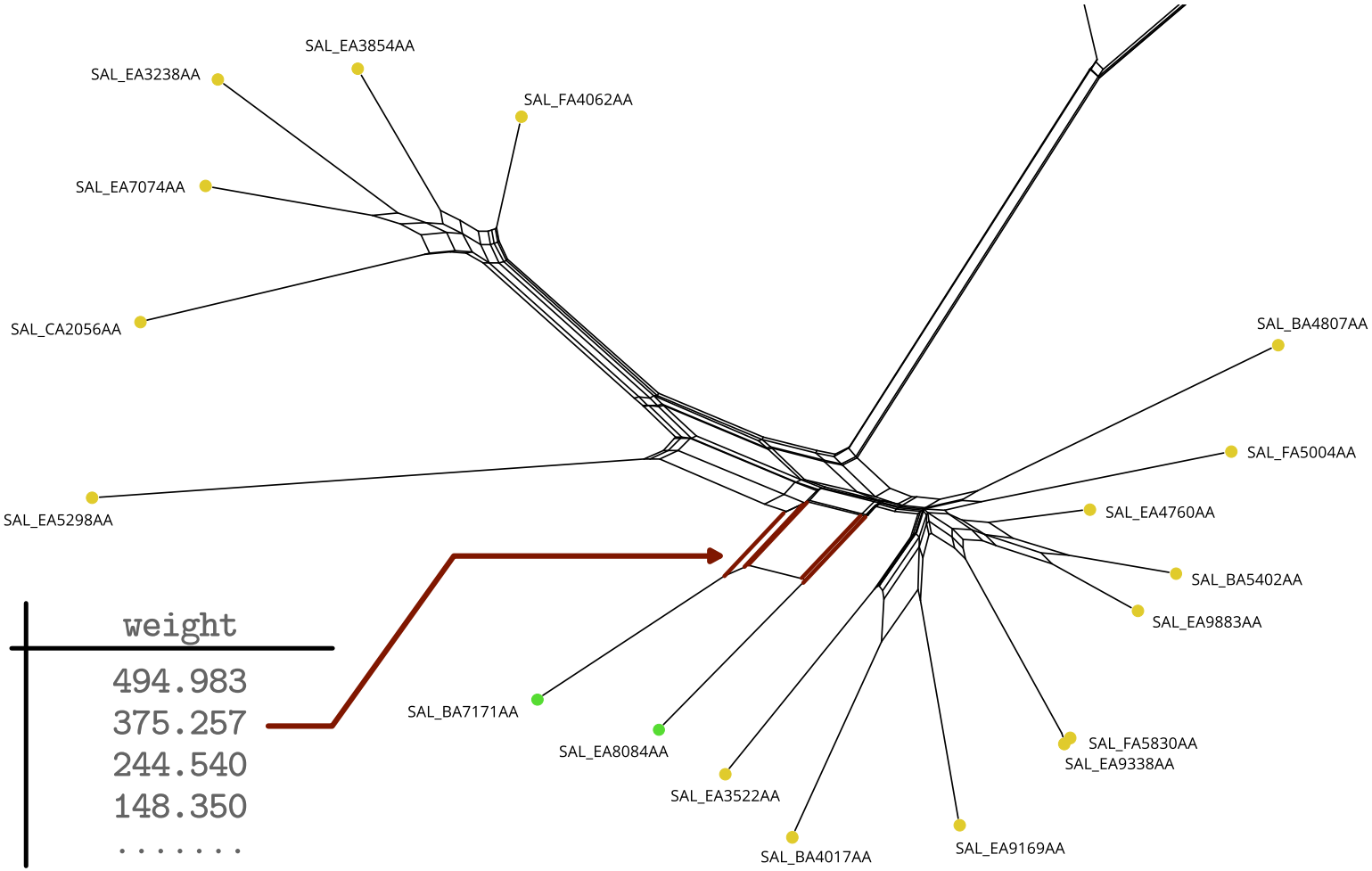
Andreas Rempel

split	#k-mers
0b00101101000000	[499, 491]
0b11000000000000	[362, 389]
0b00011010110000	[260, 230]
0b00100111001000	[131, 168]
.....	.....

weight
494.983
375.257
244.540
148.350
.....

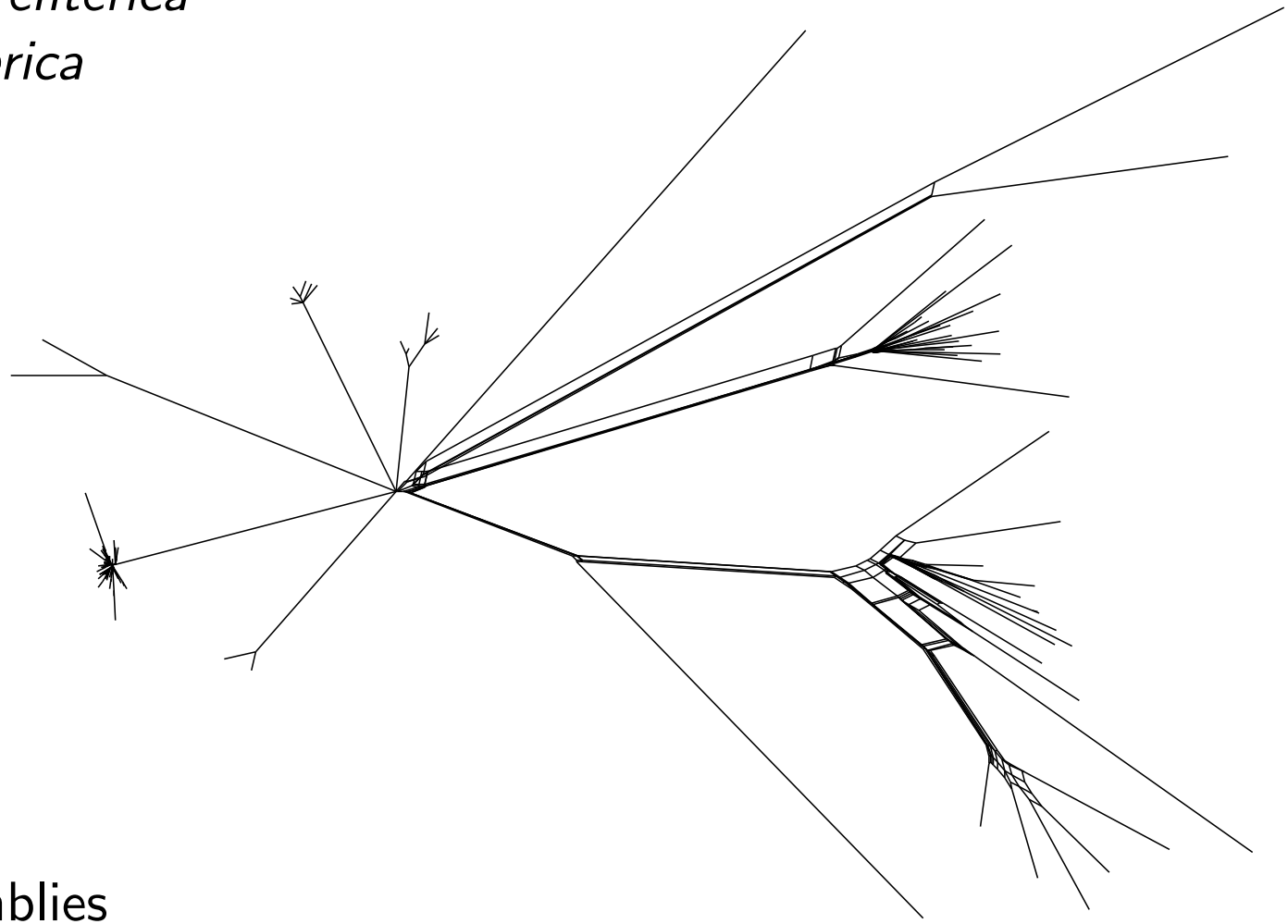
SANS serif

Andreas Rempel



[SplitsTree]

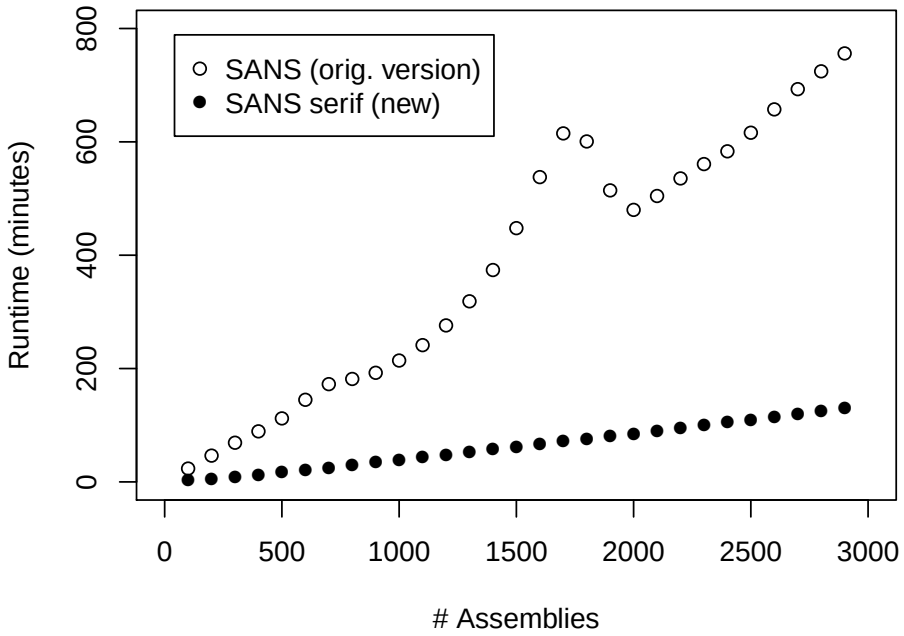
Salmonella enterica
subsp. *enterica*

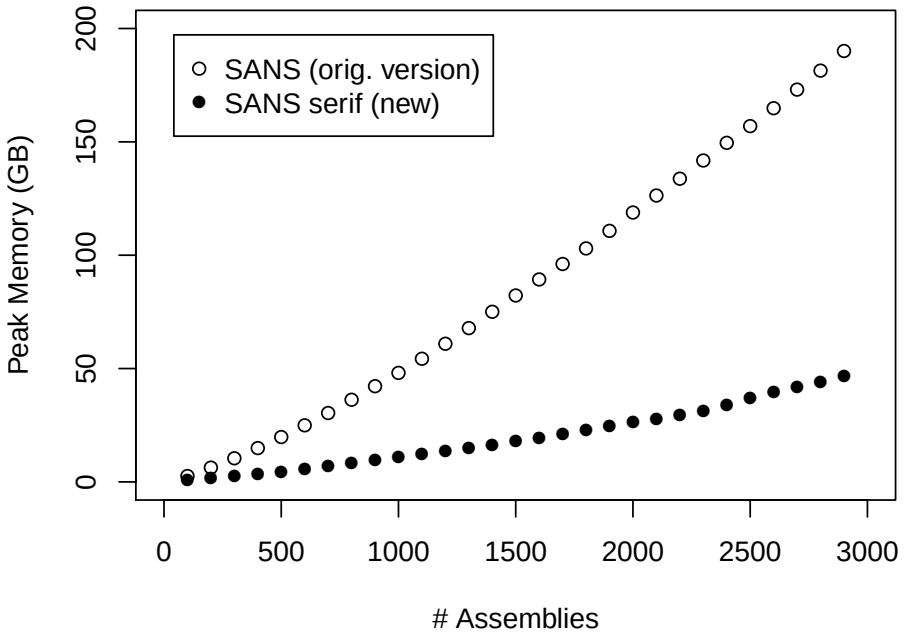


2964 assemblies
[Zhou et al., 2017]

SANS serif

Andreas Rempel





250 assemblies: andi: 110 min, Co-phylog: 9 h, FSWM: 50 h, ...

Outline

IUPAC Characters

Andreas Rempel

R → A or G
Y → C or T
S → G or C
W → A or T
K → G or T
M → A or C
B → C or G or T
D → A or G or T
H → A or C or T
V → A or C or G
N → any base

Consider all variants:

$$\text{AARCGYA} \Rightarrow \begin{cases} \text{AAACGCA} \\ \text{AAACGTA} \\ \text{AAGCGCA} \\ \text{AAGCGTA} \end{cases}$$

IUPAC Characters

Andreas Rempel

R → A or G
Y → C or T
S → G or C
W → A or T
K → G or T
M → A or C
B → C or G or T
D → A or G or T
H → A or C or T
V → A or C or G
N → any base

Consider all variants:

$$\text{AARCGYA} \Rightarrow \begin{cases} \text{AAACGCA} \\ \text{AAACGTA} \\ \text{AAGCGCA} \\ \text{AAGCGTA} \end{cases}$$

True variant:

supports / is supported by further k-mers

False variant:

cancelled out due to missing inverse split

IUPAC Characters

Andreas Rempel

Simulated phylogeny, 100 leaf genomes, length ∼96 kb, 5 PAM to the root
simulated with ALF [Dalquen et al., 2012]

	original	with 0.1 % `N`'s	
		skipped	replaced
unweighted			
precision	0.90	0.63	0.78
recall	0.70	0.38	0.58
weighted			
precision	0.98	0.95	0.98
recall	0.88	0.68	0.88

Outline

Coding Sequences
Marco Sohn

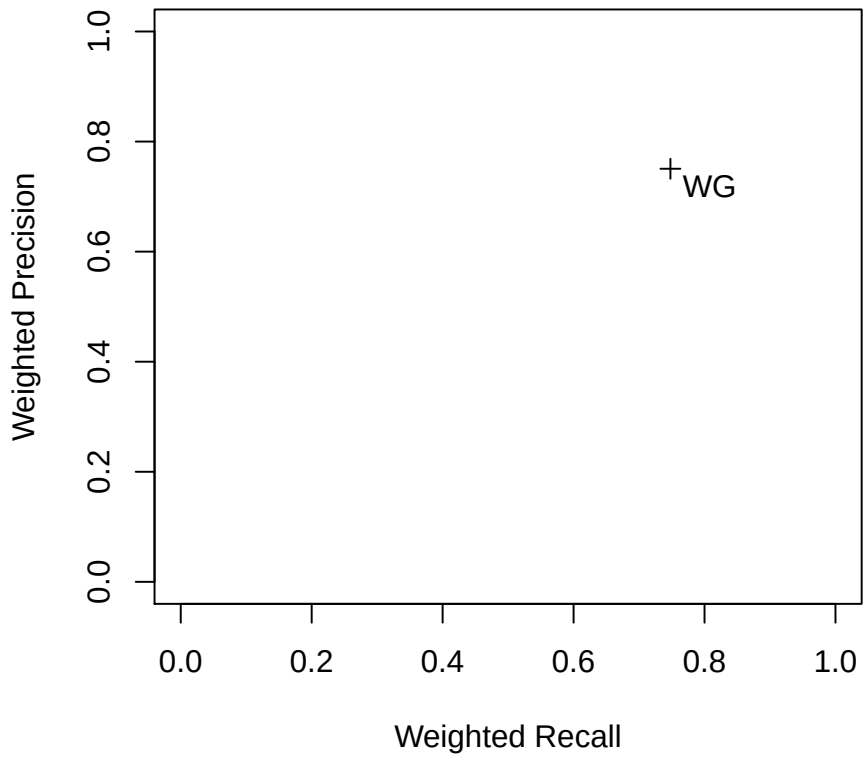
whole genome	$\Rightarrow$	coding sequnences
DNA	$\Rightarrow$	amino acids

```
-a, --amino Consider amino acids:
            --input provides amino acid sequences
            Implies --norev and a default k of 10

-c, --code  Translate DNA: --input provides coding sequences
            Implies --norev and a default k of 10
            optional: ID of the genetic code to be used
            Default: 1 (The Standard Code)
            Use 11 for Bacterial, Archaeal, and Plant Plastid Code
```

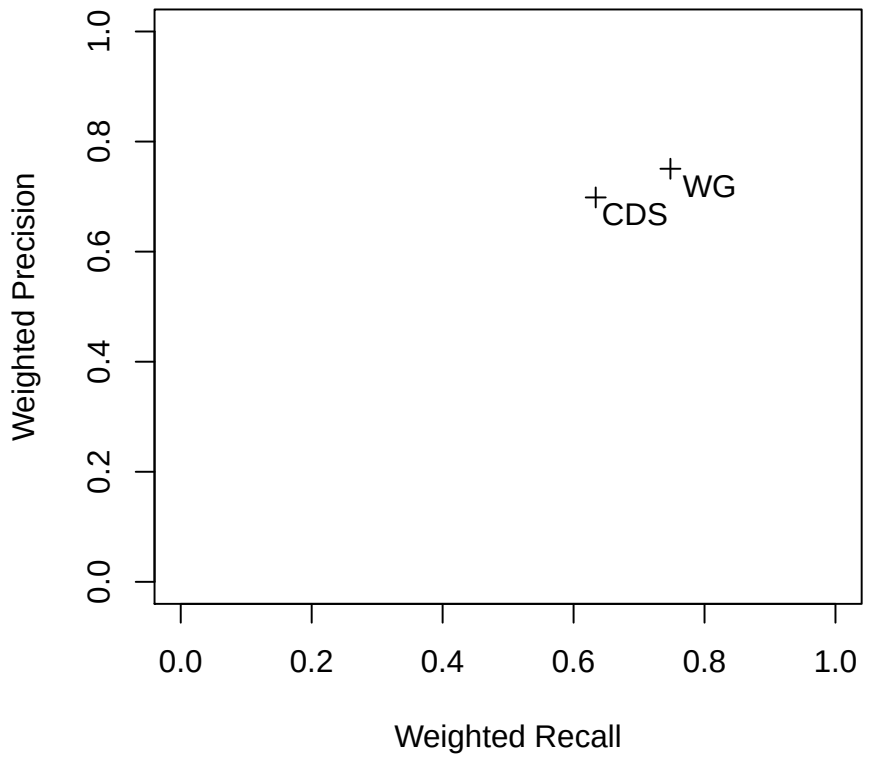
Coding Sequences
Marco Sohn

90 Pseudomonas genomes



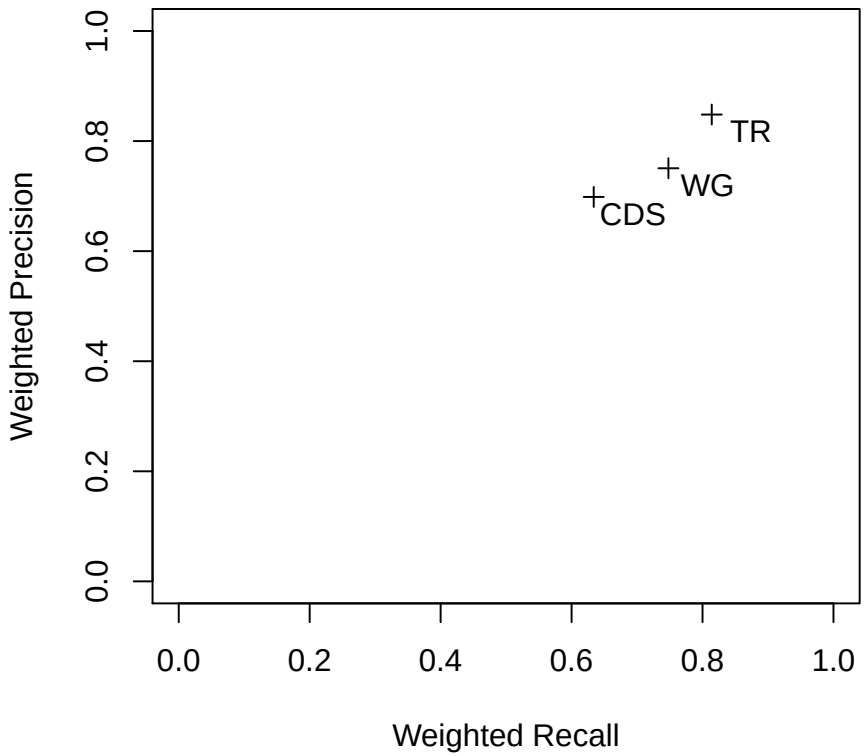
Coding Sequences
Marco Sohn

90 Pseudomonas genomes



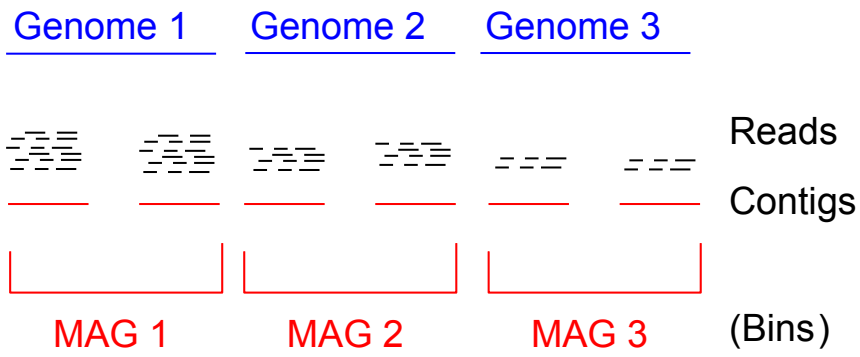
Coding Sequences
Marco Sohn

90 Pseudomonas genomes



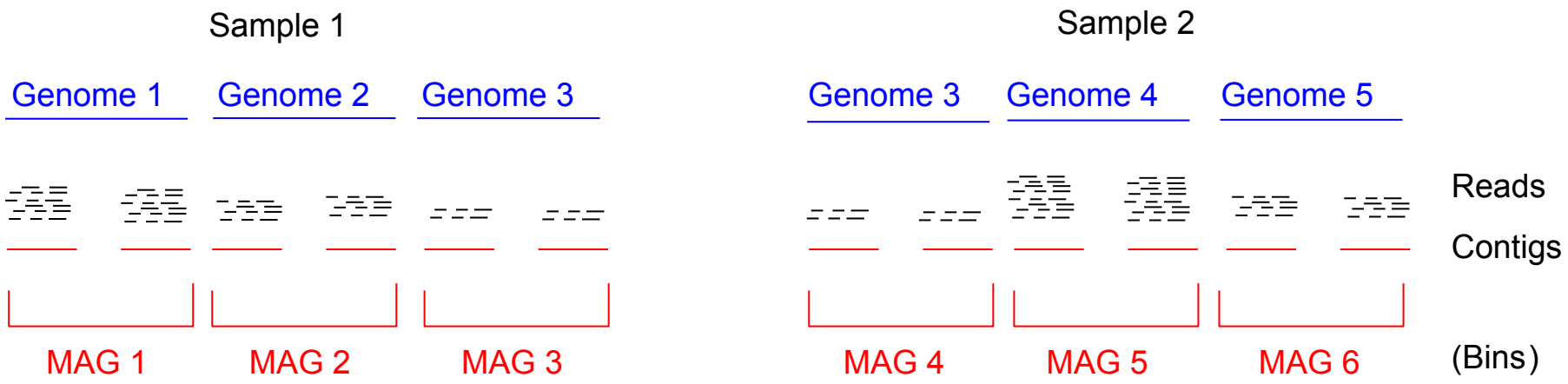
Outline

Clustering MAGs

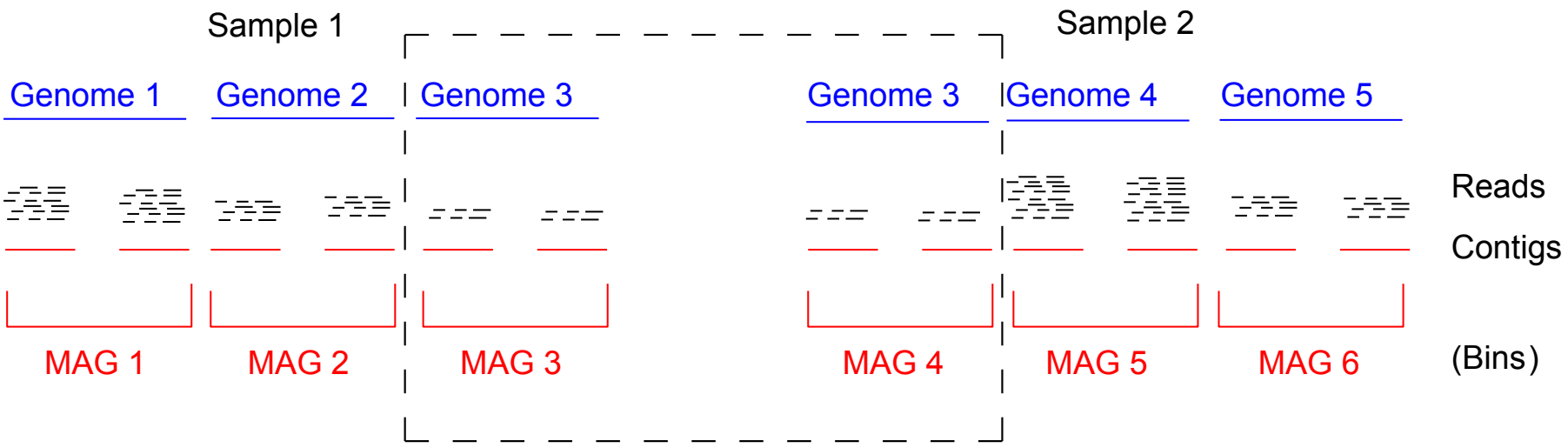


Metagenome Assembled Genomes

Clustering MAGs



Clustering MAGs



Given: Many MAGs from many samples.

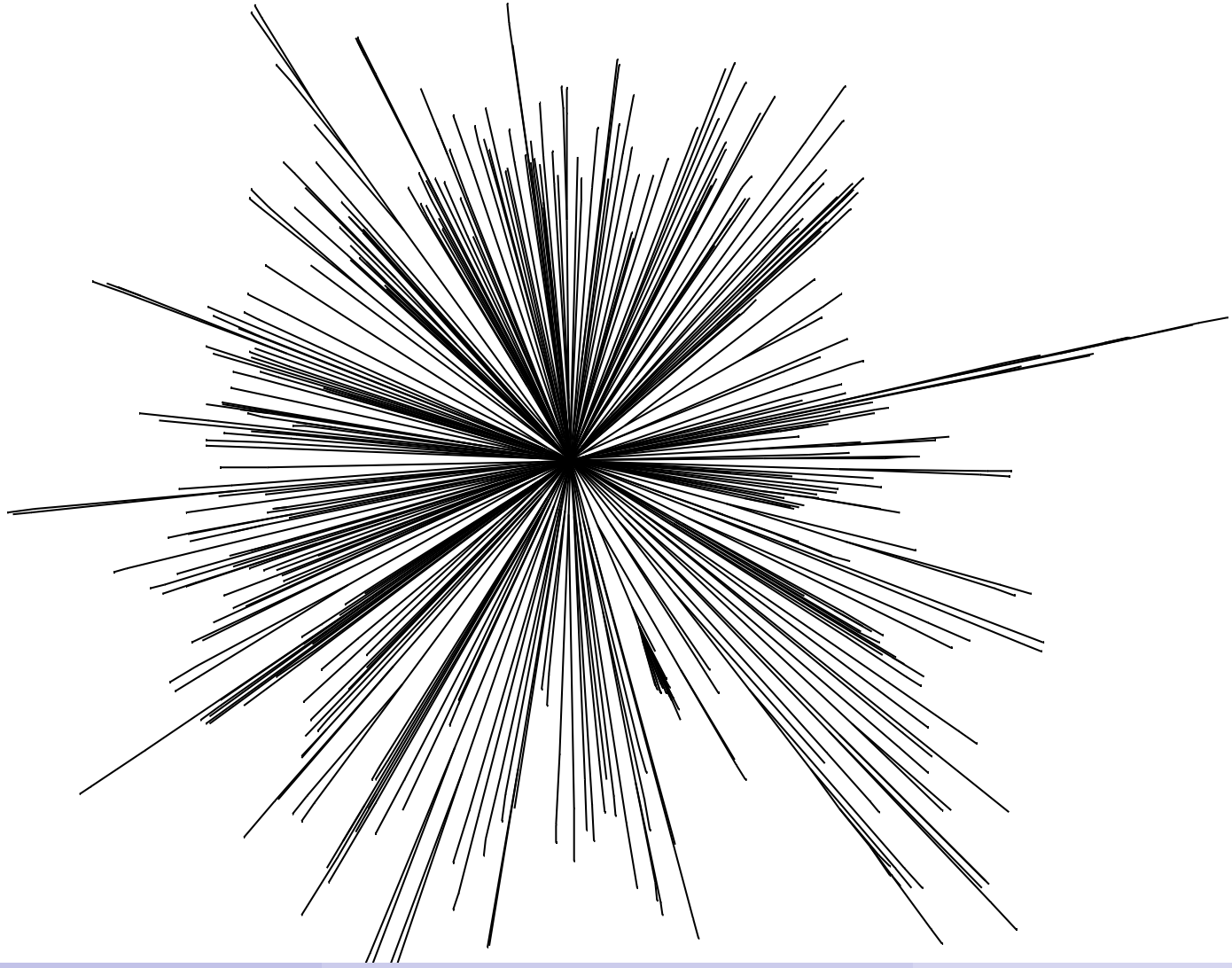
Problem: Cluster MAGs such that: cluster = genome

Clustering MAGs

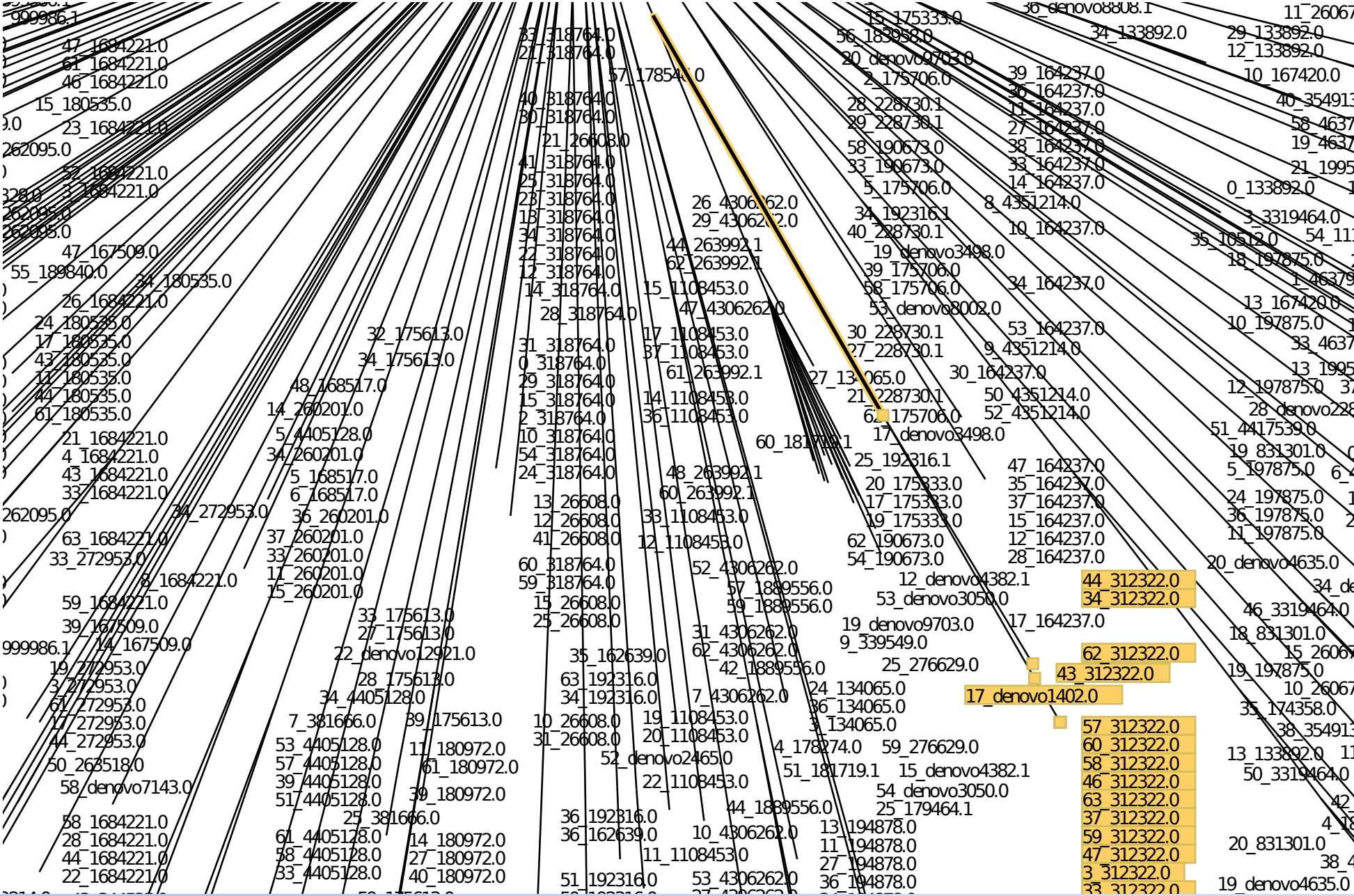
1. infer phylogeny using SANS
2. chop tree into clusters

Clustering MAGs

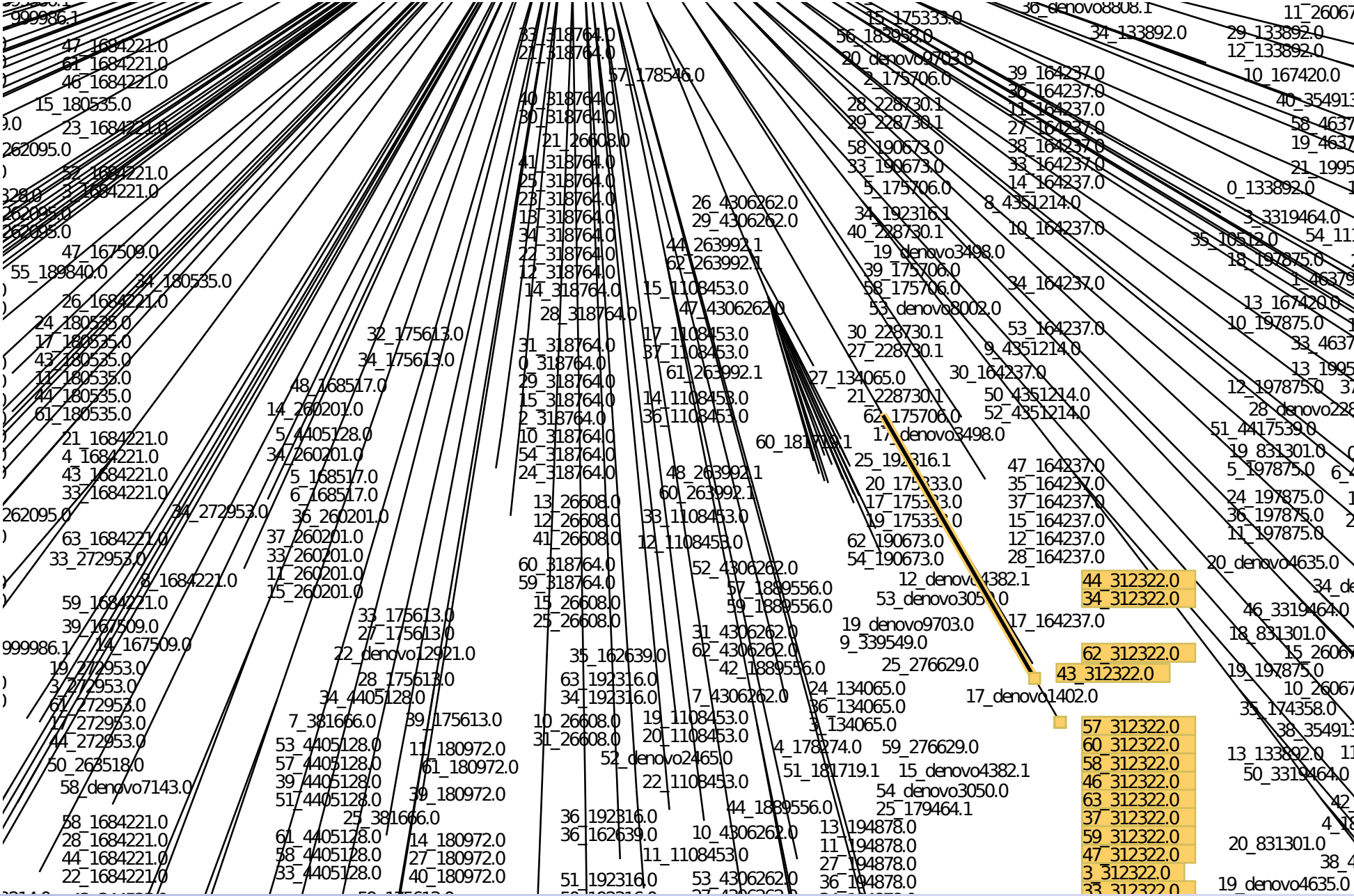
1. infer phylogeny using SANS
2. chop tree into clusters



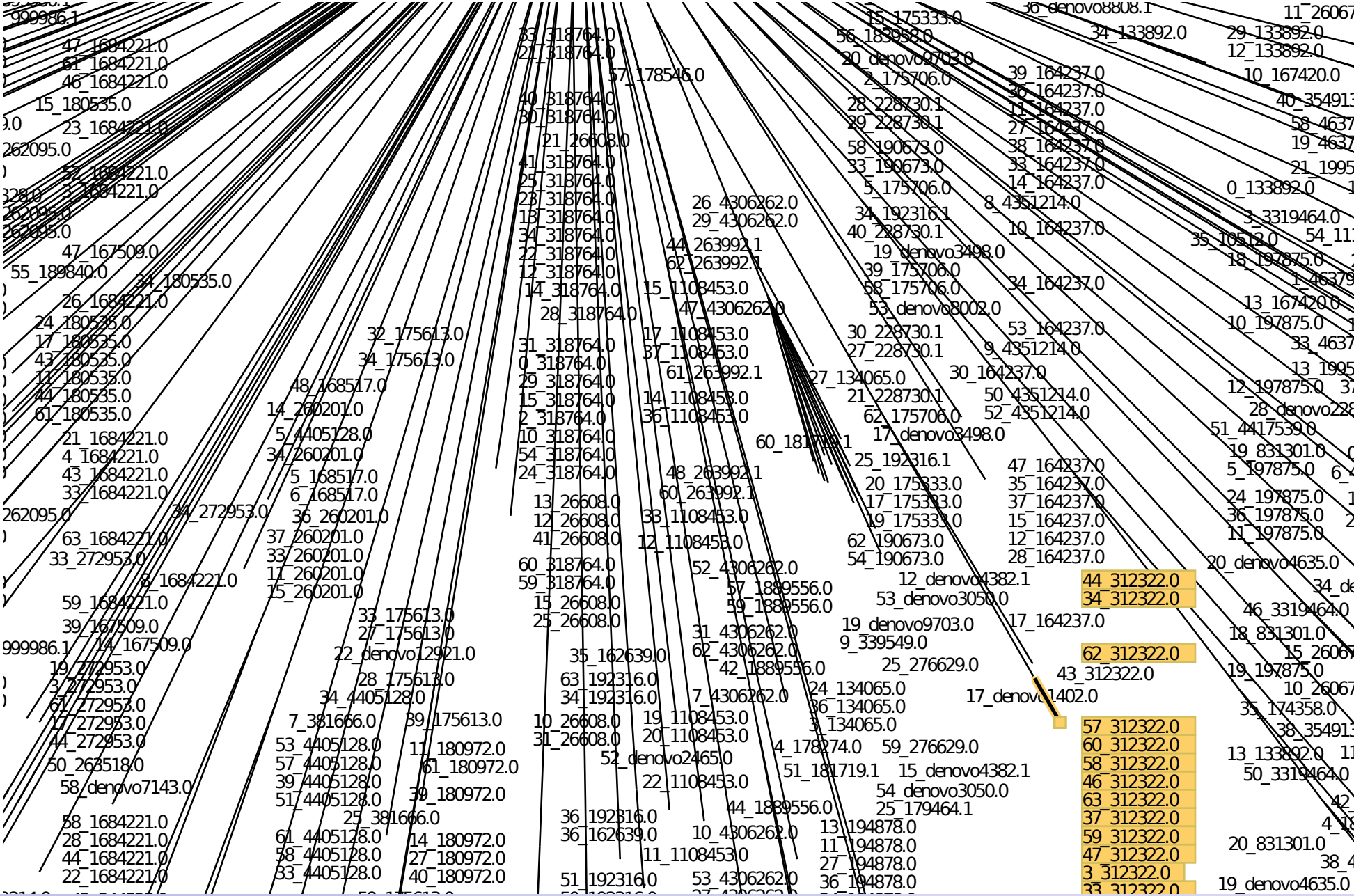
Clustering MAGs



Clustering MAGs

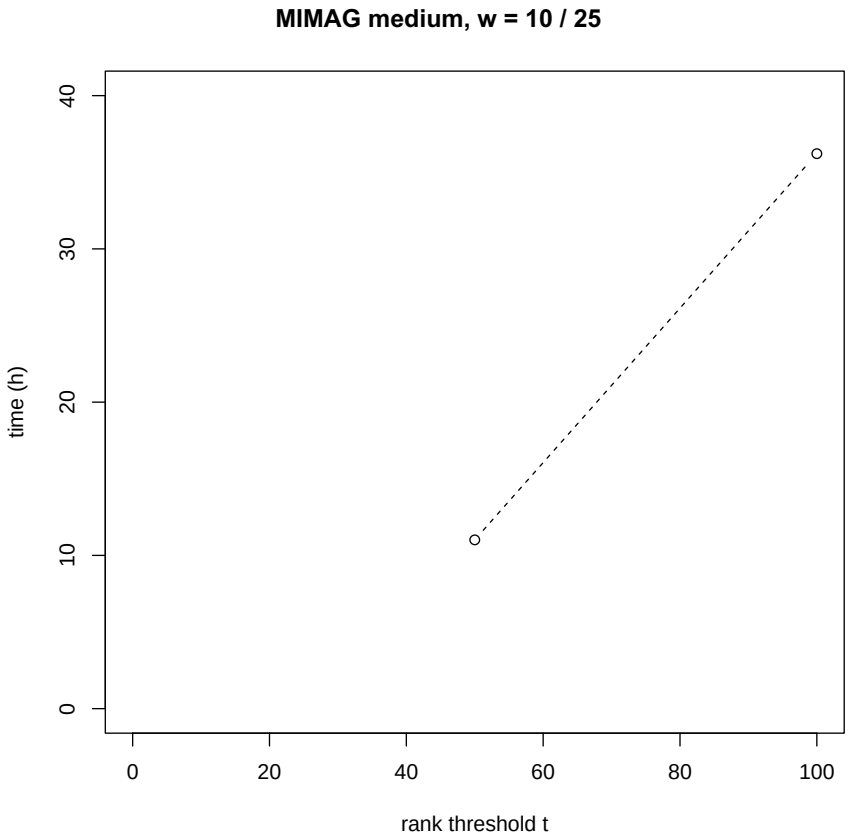


Clustering MAGs

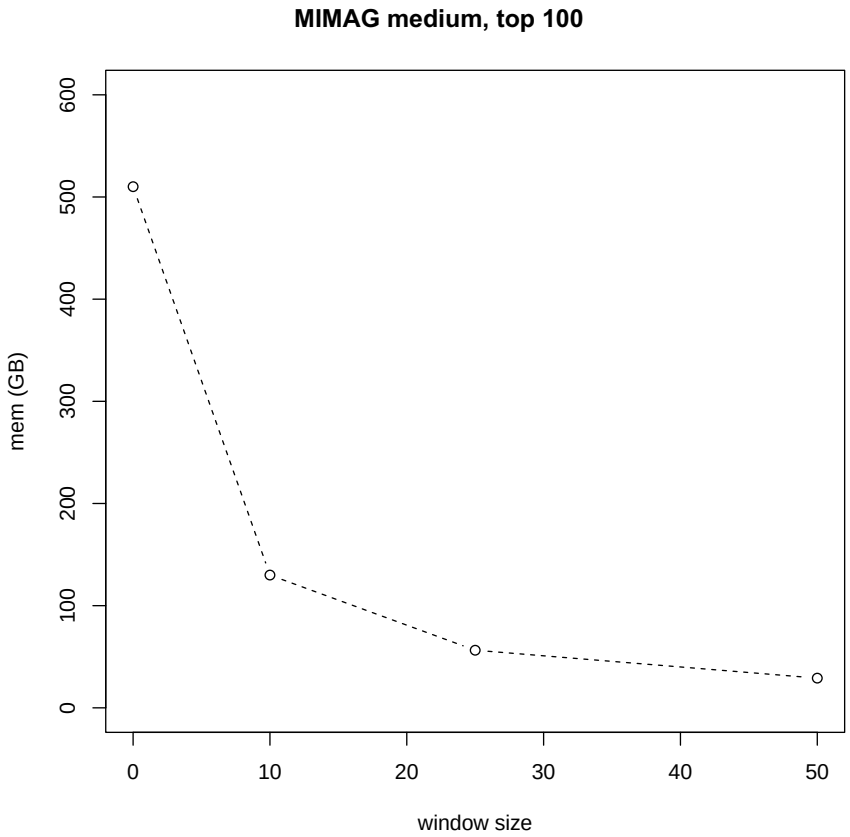


Clustering MAGs

Rank threshold



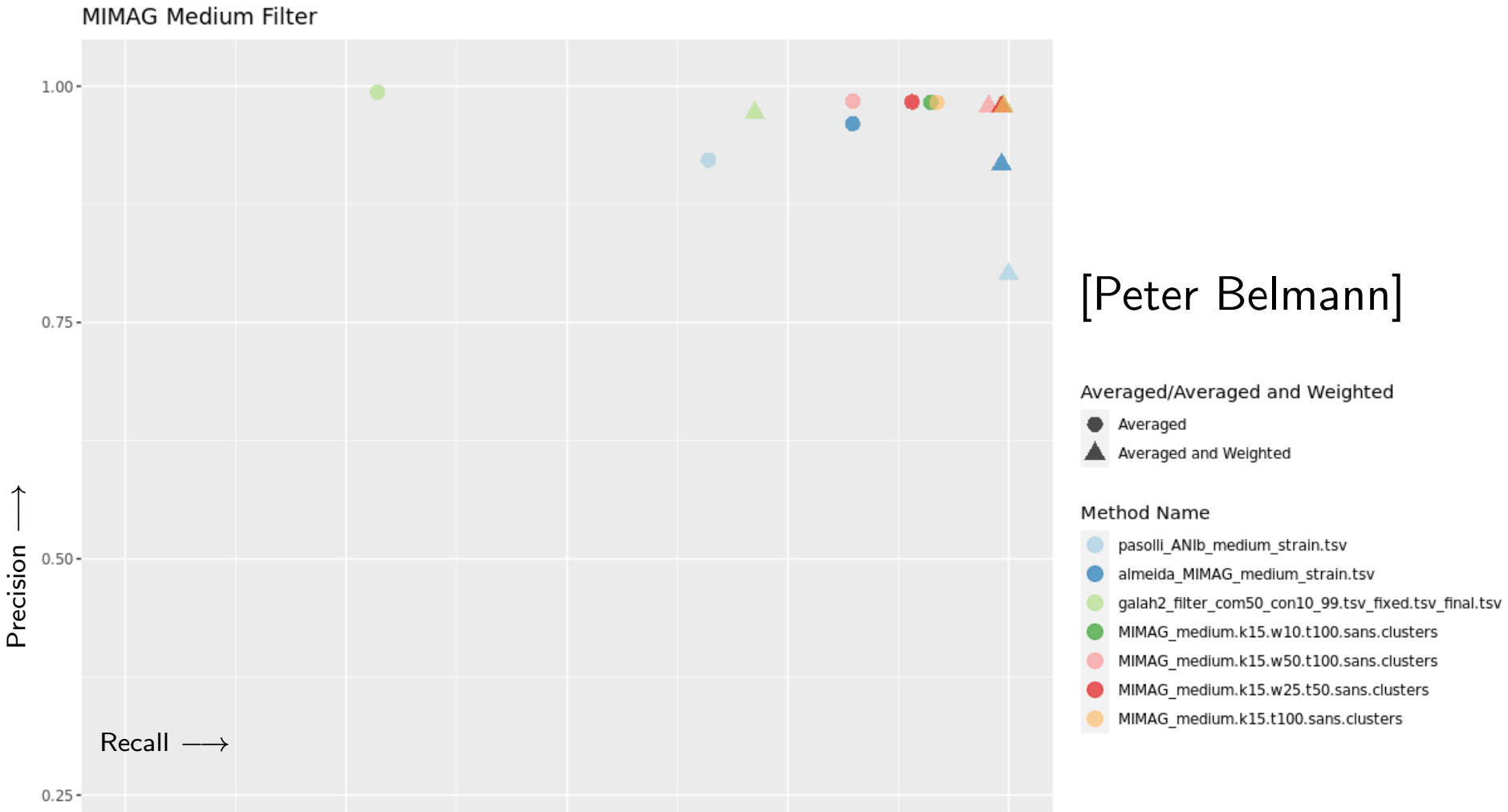
Minimizers *[Andreas Rempel]*



Clustering MAGs

CAMI mouse gut metagenome [Sczyrba et al.]

5 786 MAGs, 686 genomes (simulated)



[Peter Belmann]

Outline

Outlook

- parallelization [Fabian Kolesch]
  in particular reading the input files

Outlook

- parallelization [Fabian Kolesch]
  in particular reading the input files
- counting k-mers [Ann-Cathrin Groba]
  1 1 0 0 vs. 9 9 0 0 vs. 9 9 1 0

Outlook

- parallelization [Fabian Kolesch]
  in particular reading the input files
- counting k-mers [Ann-Cathrin Groba]
  1 1 0 0 vs. 9 9 0 0 vs. 9 9 1 0
- estimating k [Rebecca Pfeil]
  Shannon entropy(?), sampling

Outlook

- parallelization [Fabian Kolesch]
  in particular reading the input files
- counting k-mers [Ann-Cathrin Groba]
  1 1 0 0 vs. 9 9 0 0 vs. 9 9 1 0
- estimating k [Rebecca Pfeil]
  Shannon entropy(?), sampling
- reference free quality measure [Marco Sohn, Rebecca Pfeil]
  tree-likeliness

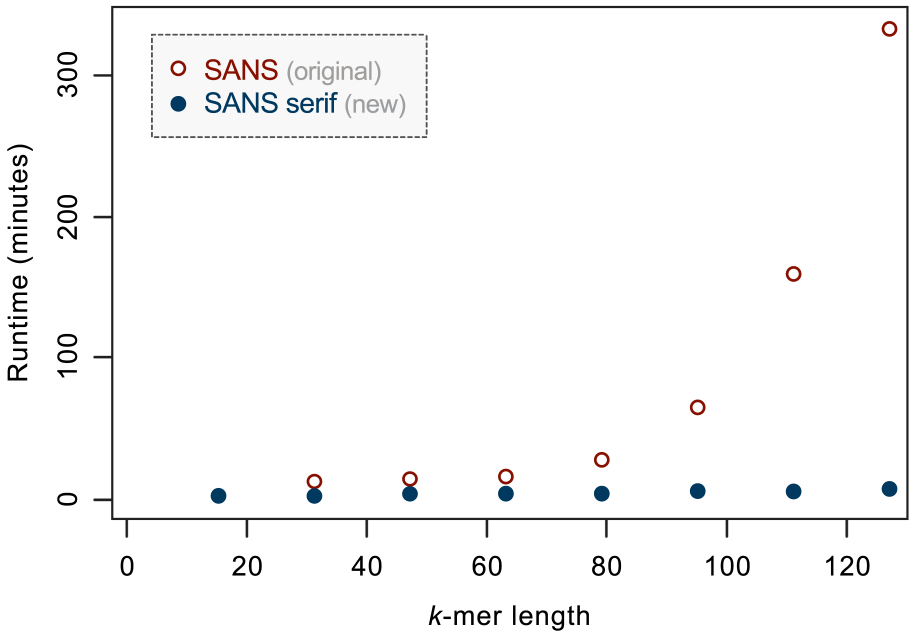
Outlook

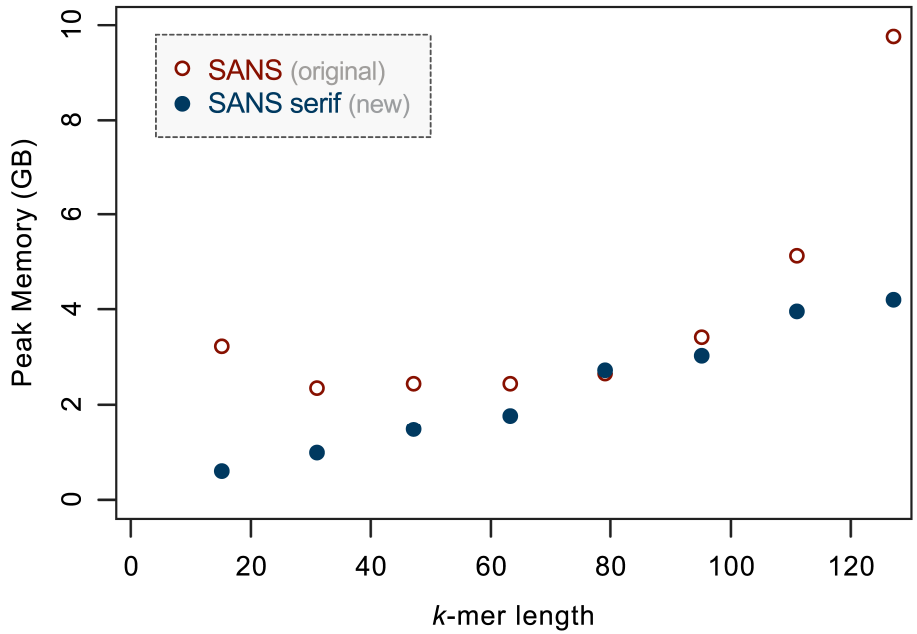
- parallelization [Fabian Kolesch]
  in particular reading the input files
- counting k-mers [Ann-Cathrin Groba]
  1 1 0 0 vs. 9 9 0 0 vs. 9 9 1 0
- estimating k [Rebecca Pfeil]
  Shannon entropy(?), sampling
- reference free quality measure [Marco Sohn, Rebecca Pfeil]
  tree-likeliness

Thank you!

Appendix

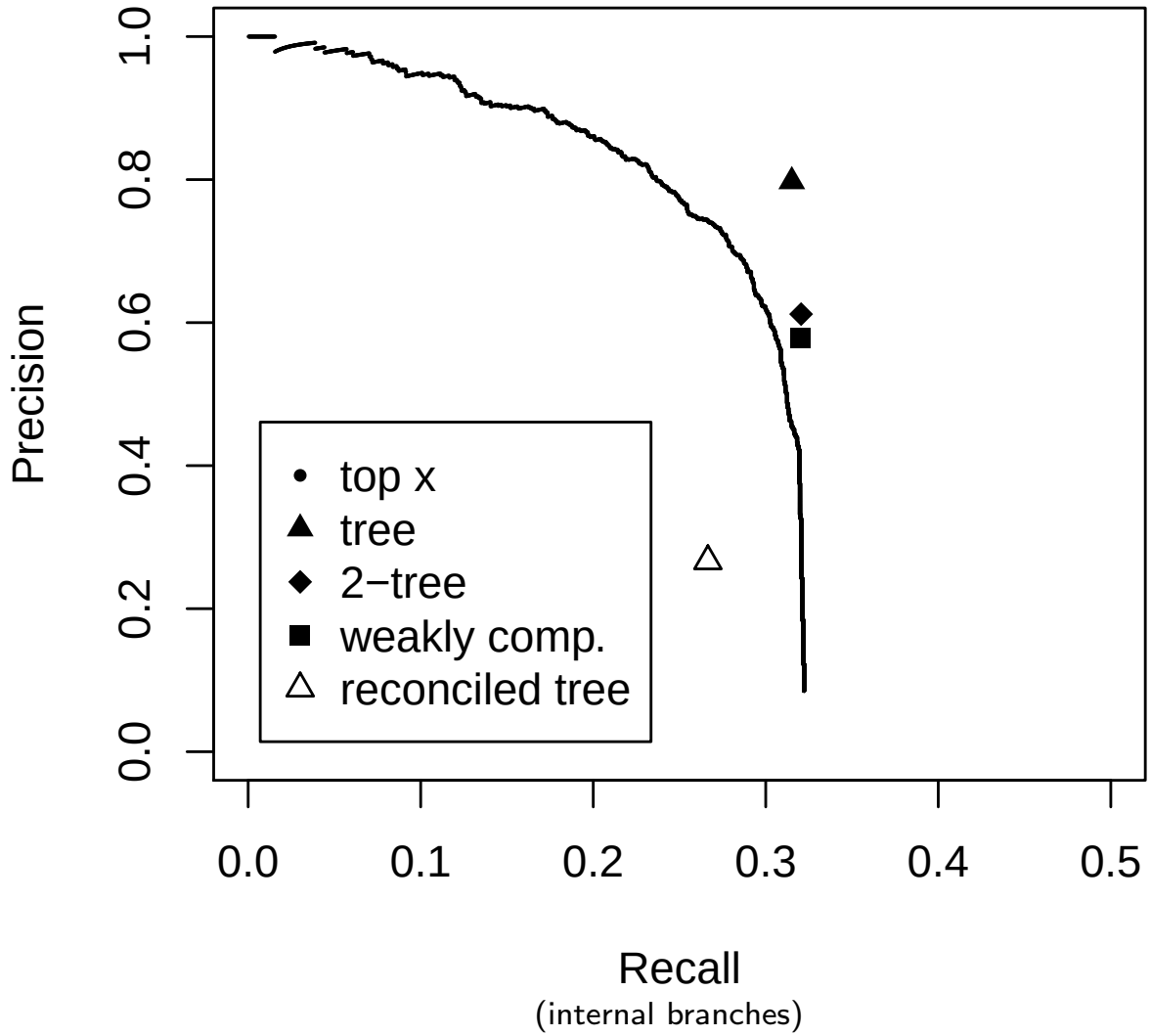
Appendix





(100 assemblies)

Appendix


Precision
1.0
0.8
0.6
0.4
0.2
0.0
top x
tree
2−tree
weakly comp.
reconciled tree
0.0
0.1
0.2
0.3
0.4
0.5
Recall
(internal branches)

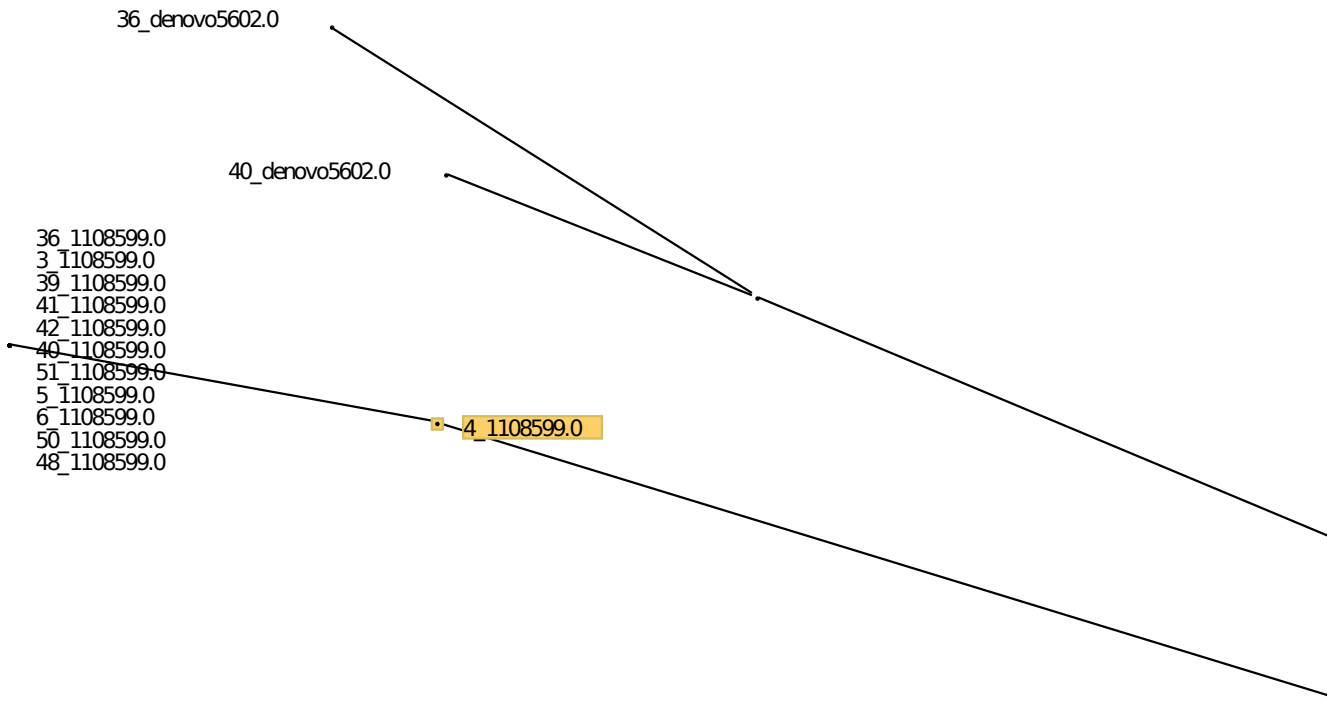
Appendix

90 Pseudomonas genomes

	Whole Genome	Coding Sequences	
	DNA	DNA	AA
k	25	25	8
time	17min	6.9s	1.5s
memory	12G	140M	17M

Appendix

1. SANS, filter for tree, re-root
2. post-order traversal:
   - 2.1 ignore non-branching node
   - 2.2 get clusters from sub-trees (recursively)
     item if edge $\geq$ parent edge:
     remove found clusters from current leaf set
     remaining leaf set $=:$ new cluster



Appendix

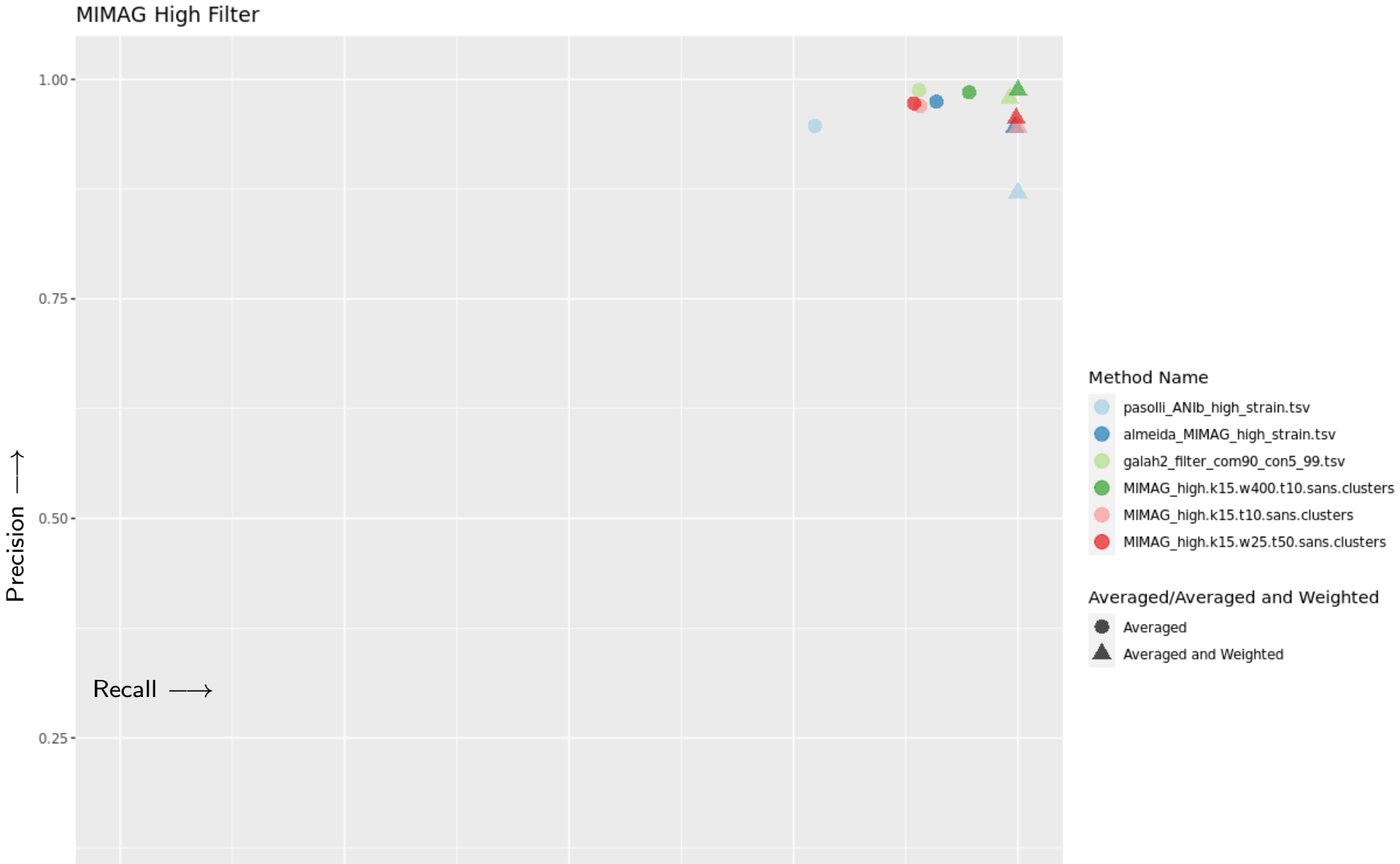
CAMI mouse gut metagenome [Sczyrba et al.]

Simulated dataset representing 64 metagenome samples.

	All	MIMAG medium	MIMAG high
contamination		$< 10\,\%$	$< \ 5\,\%$
completeness		$\geq 50\,\%$	$> 90\,\%$
# MAGs	11 602 (23 GB)	5 786	2 510
# genomes	791	686	349
# species	509	448	271

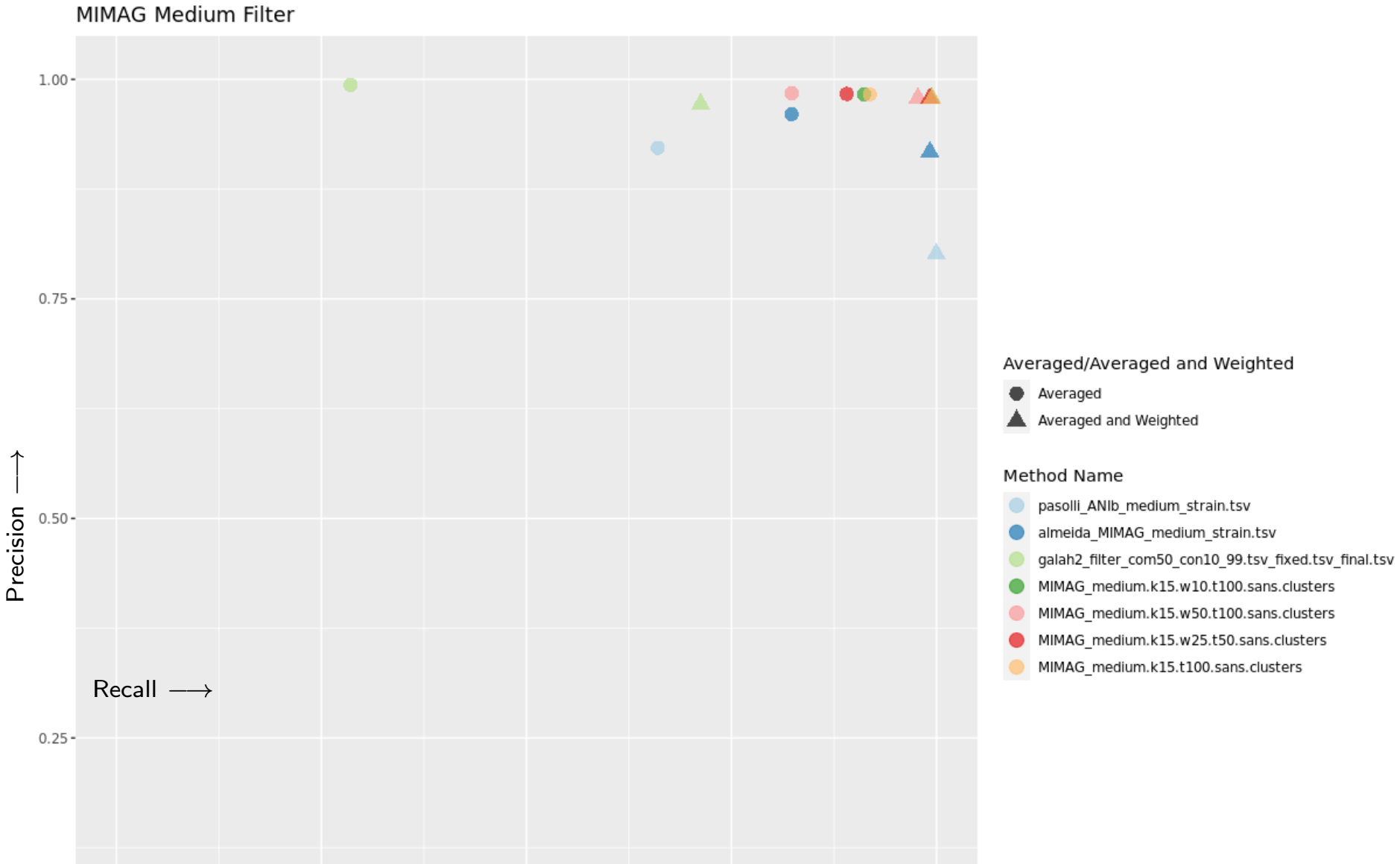
Clustering MAGS – Strain level

[Peter Belmann]



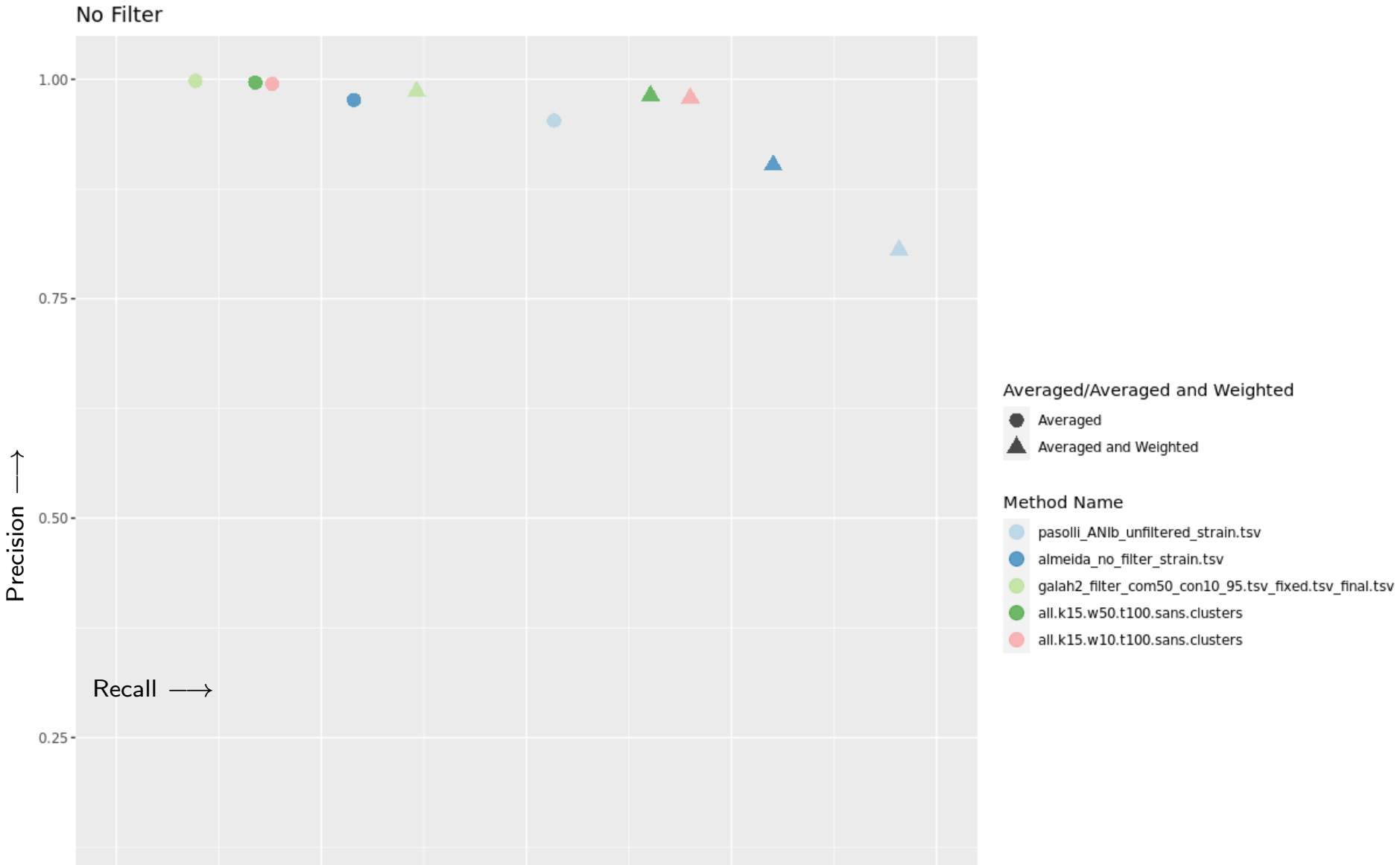
Clustering MAGS – Strain level

[Peter Belmann]



Clustering MAGS – Strain level

[Peter Belmann]

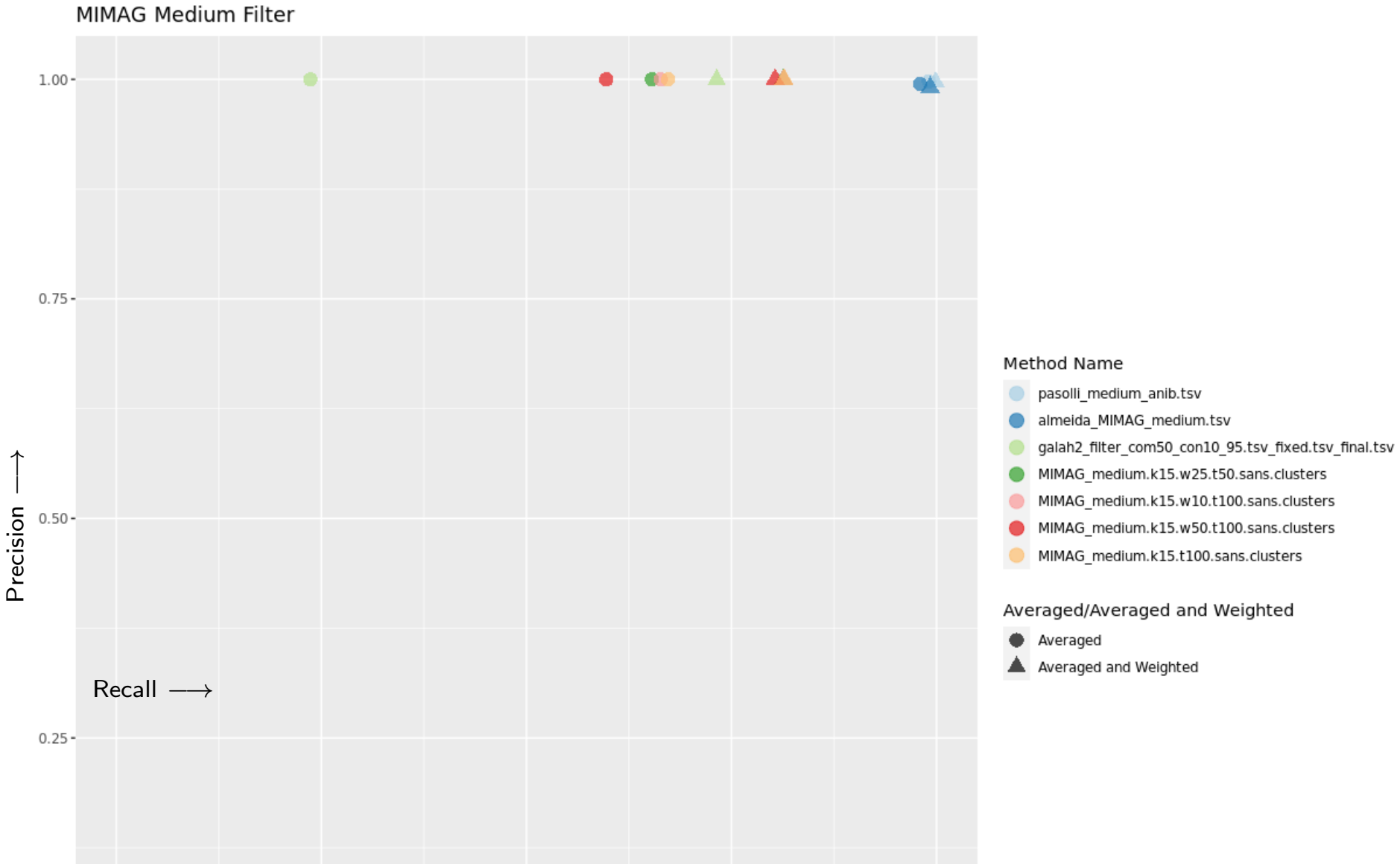


Clustering MAGS – Species level

[Peter Belmann]

Clustering MAGS – Species level

[Peter Belmann]



Clustering MAGS – Species level

[Peter Belmann]