NAPP, a Single-Pass method to detect ncRNA in Bacteria by phylogenetic profiling with a "plus value"

Marchais Antonin

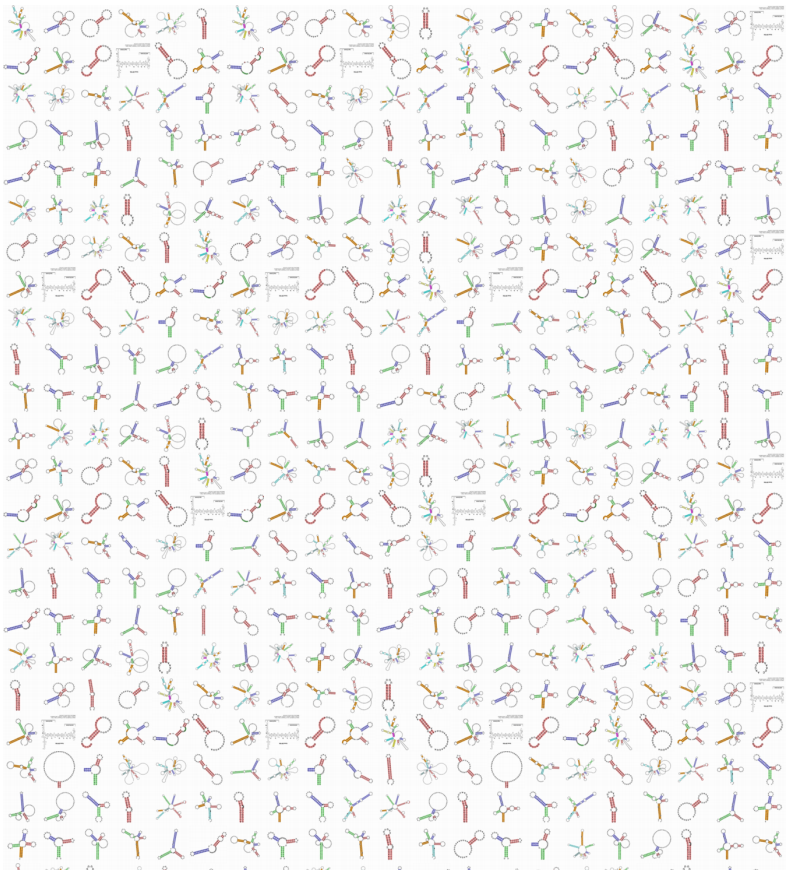
Sequence, structure and function of RNAs – Daniel Gautheret
Institut de Génétique et Microbiologie, Université Paris-Sud

April 16, 2009

Plan

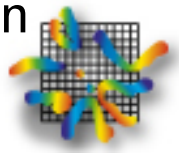
Introduction

The modern ncRNA world



- 1371 ncRNA families in RFAM, twice in two years
- Diversity of function and structure
- Many of them are not functionally characterized

→ With the "genome-by-day" era, automatic ncRNAs detection and function prediction became an absolute necessity



Principal features of ncRNA

- Sequence conservation in intergenic regions (Comparative genomics)
- Structured elements (folding energy)
- Compensatory mutations (Covariance models)
- Terminators in intergenic regions

What about the ncRNA prediction methods ?

Classical pipeline to detect new ncRNAs:

- Detection of intergenic conserved elements
- Filtering of conserved elements using RNAz, QRNA, Evofold
- Time consuming, No information about putative function, Not Automatized

Needed

ncRNA detection tools that:

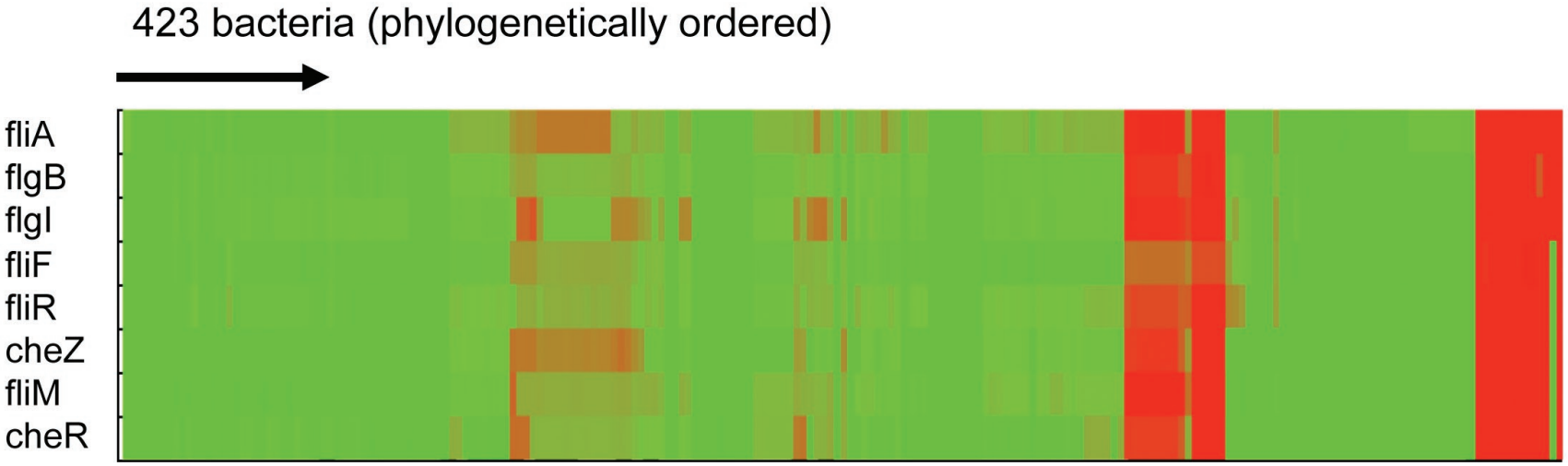
- Deal with current and incoming genome data
- Require as little user input/expertise as possible
- Achieve at least same level of sensitivity/specificity as an expert using own alignments and RNAz/Evofold
- Provide information about the ncRNA putative function ??

NAPP, a new pipeline based on co-inheritance concept

Proposed pipeline

- Take any reference genome
- For each Intergenic region:
  - Blast against all available genomes
  - Collect anything loosely conserved
  - Classify conserved elements using phylogenetic profiling

Phylogenetic profiling



- 8 proteins involved in *E.coli* motility:
  - Similar profiles !

    Pelligrini, 1999

Phylogenetic profiling of noncoding elements may...

- Help distinguish true conservation from dispersed hits
- Functionally distinguish elements based on species distribution
- And it is built by nature to stand up to the deluge of genomic data!

Loosely detect conserved noncoding elements

- For each InterGenic Region in reference sequence, blast it against all available genomes (500, 1000, whatever)
- Pile-up all Blast hits for region
- Measure raw conservation score at each position

```
coli K12      AGCTGACTATGCGTGAC
coli O57:H7   A CTGACTATGCGTGAC
Shigella        TGACTA
Y.pestis          ACCA
Cons. Score   10122332311111111
```

How to cancel noise from multiple highly related sequences?

Loosely detect conserved noncoding elements

- For each InterGenic Region in reference sequence, blast it against all available genomes (500, 1000, whatever)
- Pile-up all Blast hits for region
- Measure raw conservation score at each position

```
coli K12      AGCTGACTATGCGTGAC
coli O57:H7   A CTGACTATGCGTGAC
Shigella        TGACTA
Y.pestis          ACCA
Cons. Score   10122332311111111
```

How to cancel noise from multiple highly related sequences?

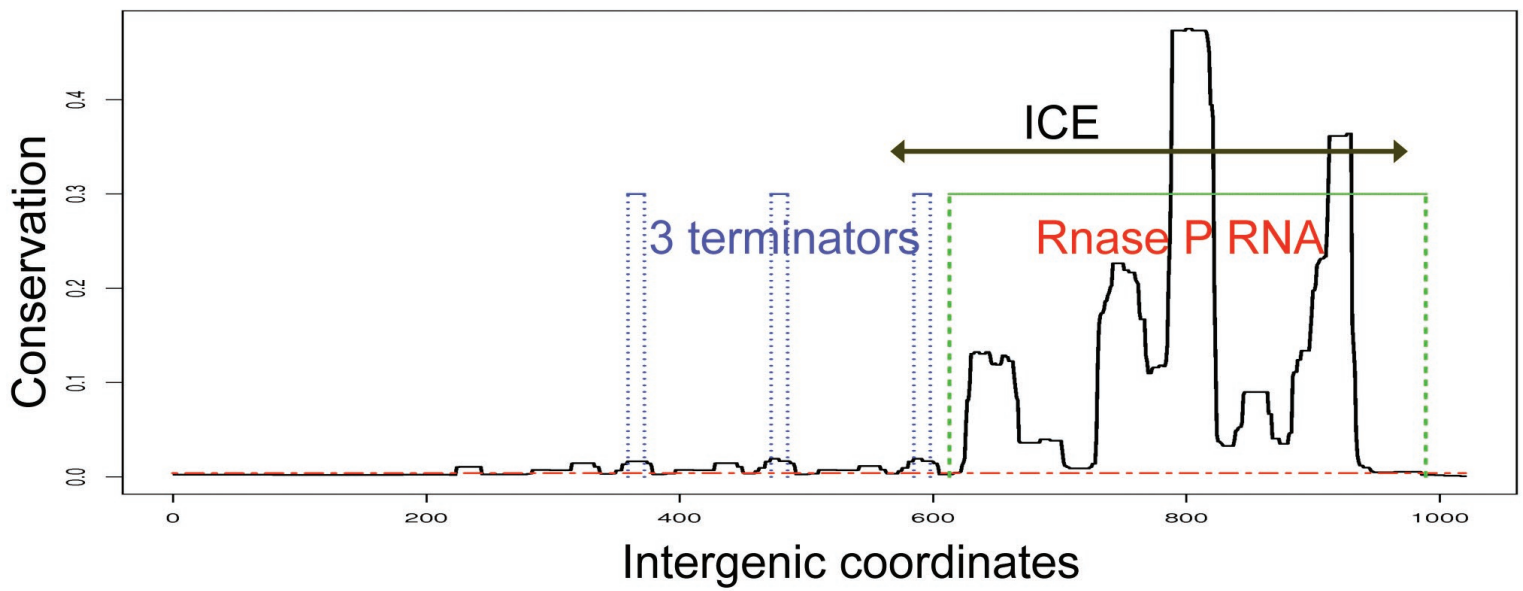
- Weight conservation score by phylogenetic distance using 16S tree

$$\text{Weighted score} = \frac{\sum_{i}^{k}\left(Dist(Oi,Ok) * X_{i->k}\right)}{\sum_{i}^{k} Dist(Oi,Ok)}$$

ICE: Intergenic Conserved Elements

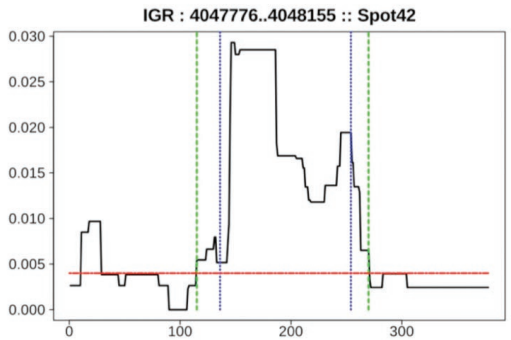
Example of ICE detected in *E.coli K12* intergenic fragment 3267625..3268646

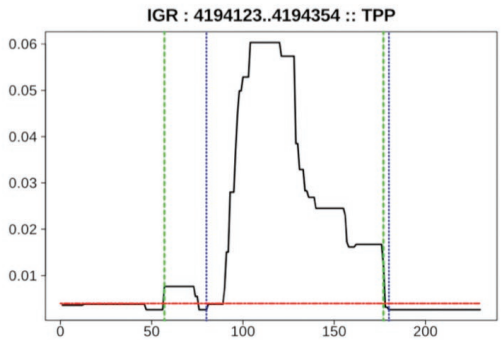
Empirical criteria for Cons $>$ 0.04; ICE: $>$ 15 nt length / $<$ 35 nt spacing

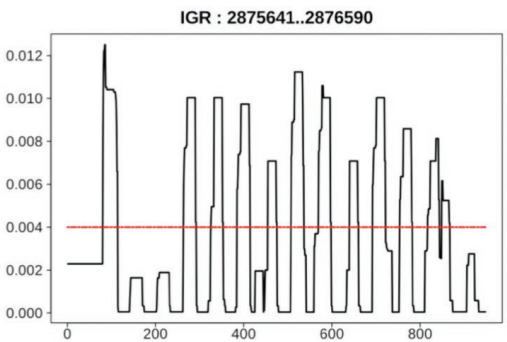


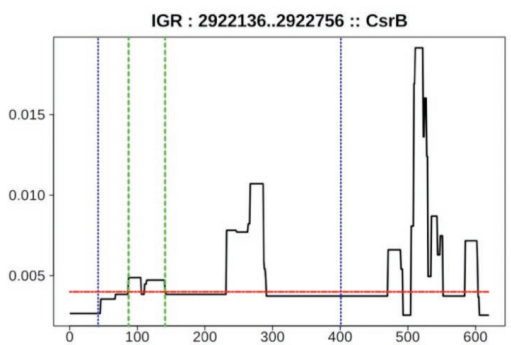
ICE Gallery

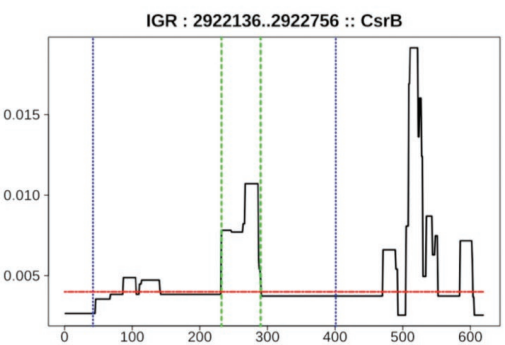
ICE gallery...

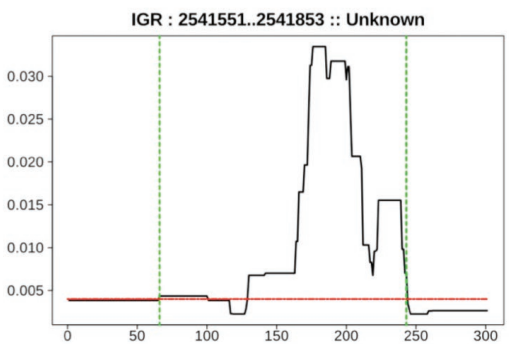


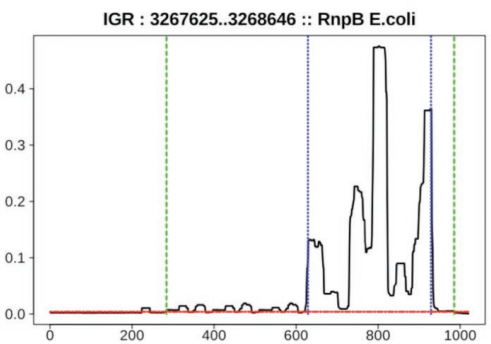


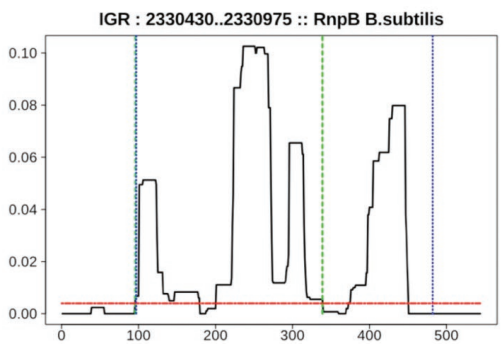




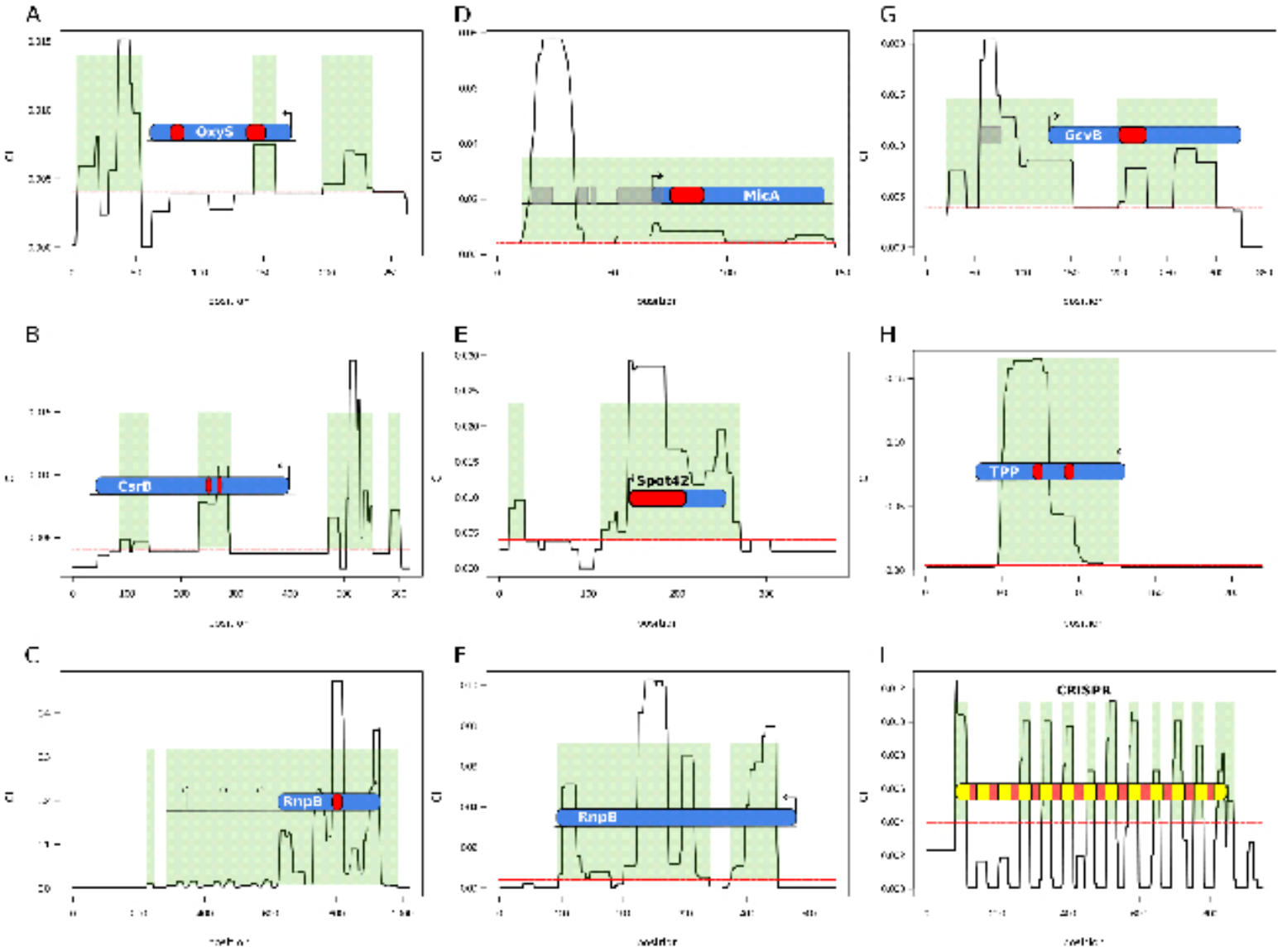


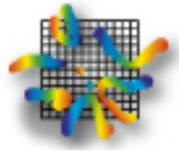






ICE Gallery

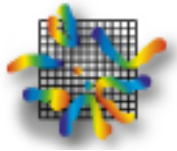




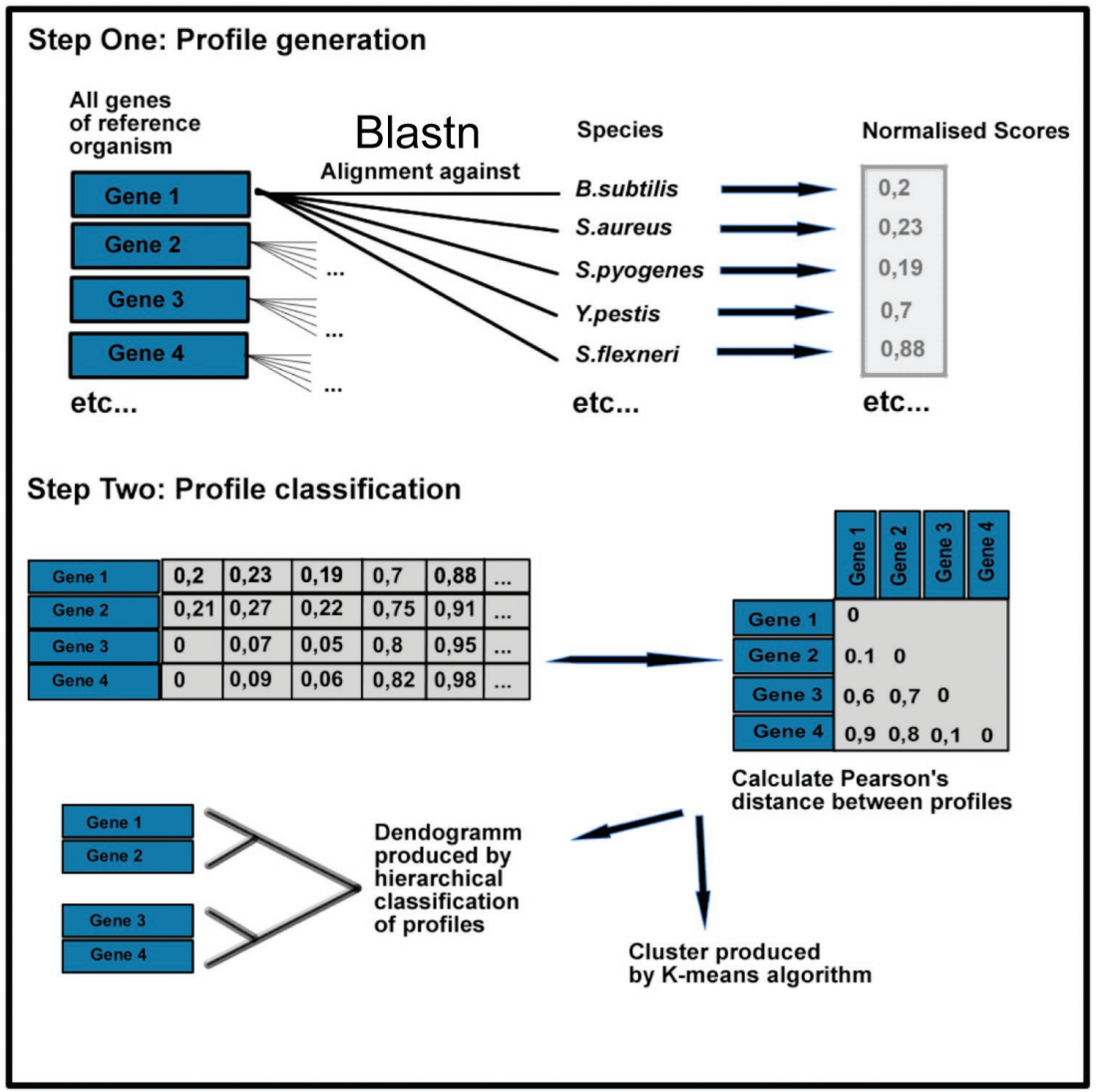
ICEs stats

	E. coli	***B. subtillis***
# ICEs	3483	2714
ICE min/max size	15/701	15/284
ICE mean size	39	34
Fraction of known ncRNA captured (except tRNA/rRNA)	56/74	61/70

Longest elements: *E.coli* Rnase P and *B.subtilis* SRP

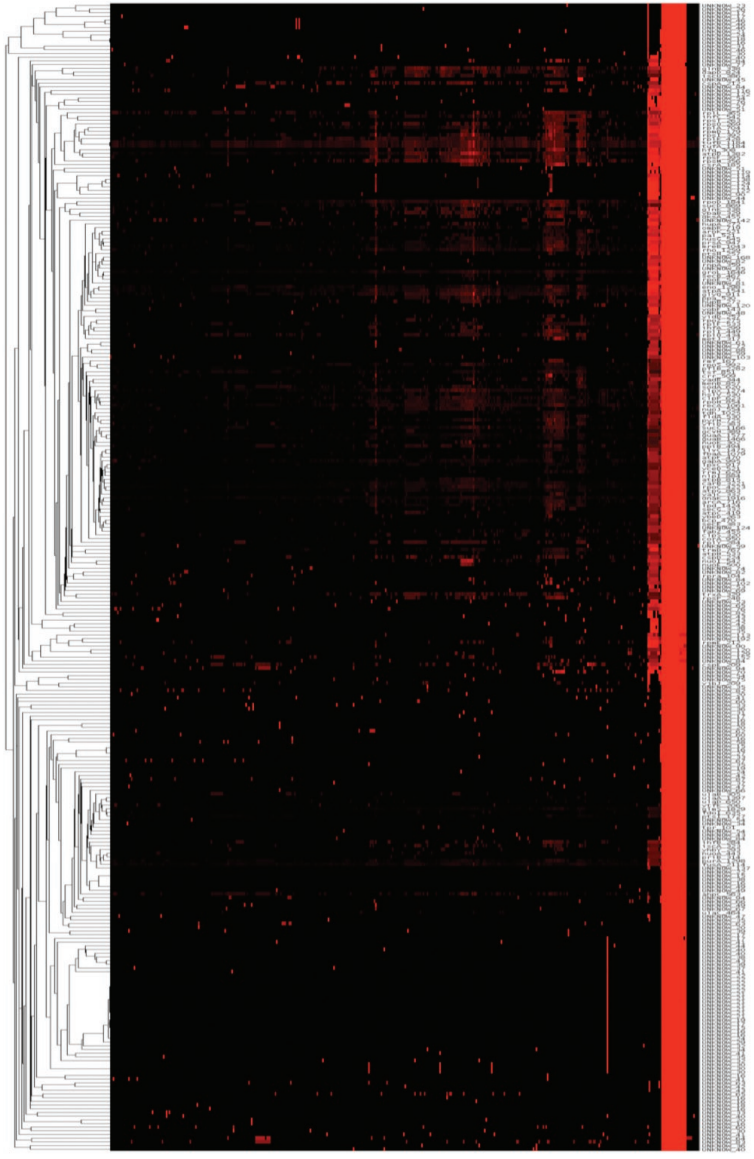


From ICE to profiles

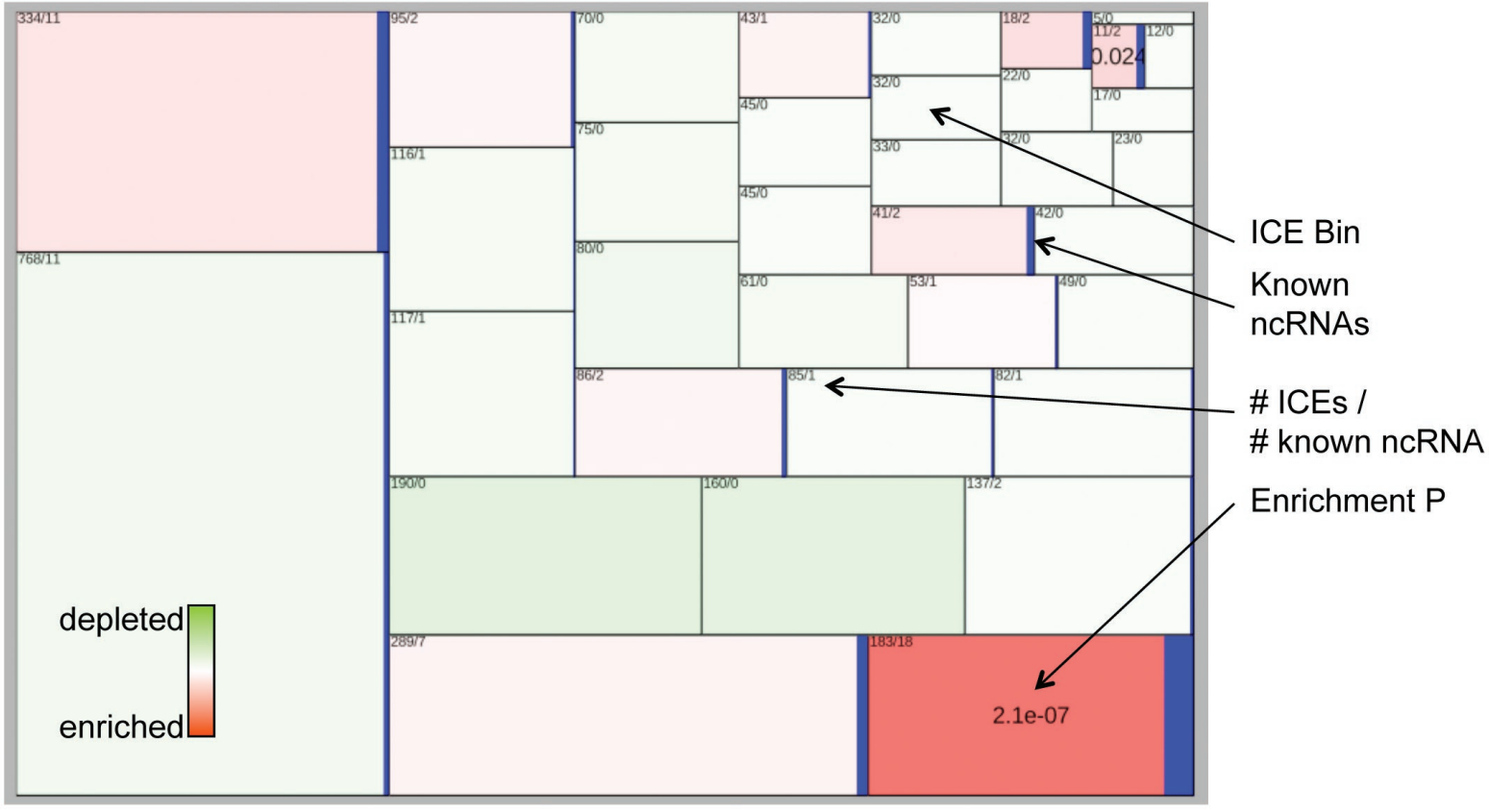


ICE phylogenetic profile

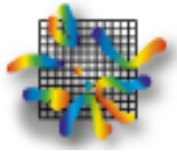
The hierarchical clustering view



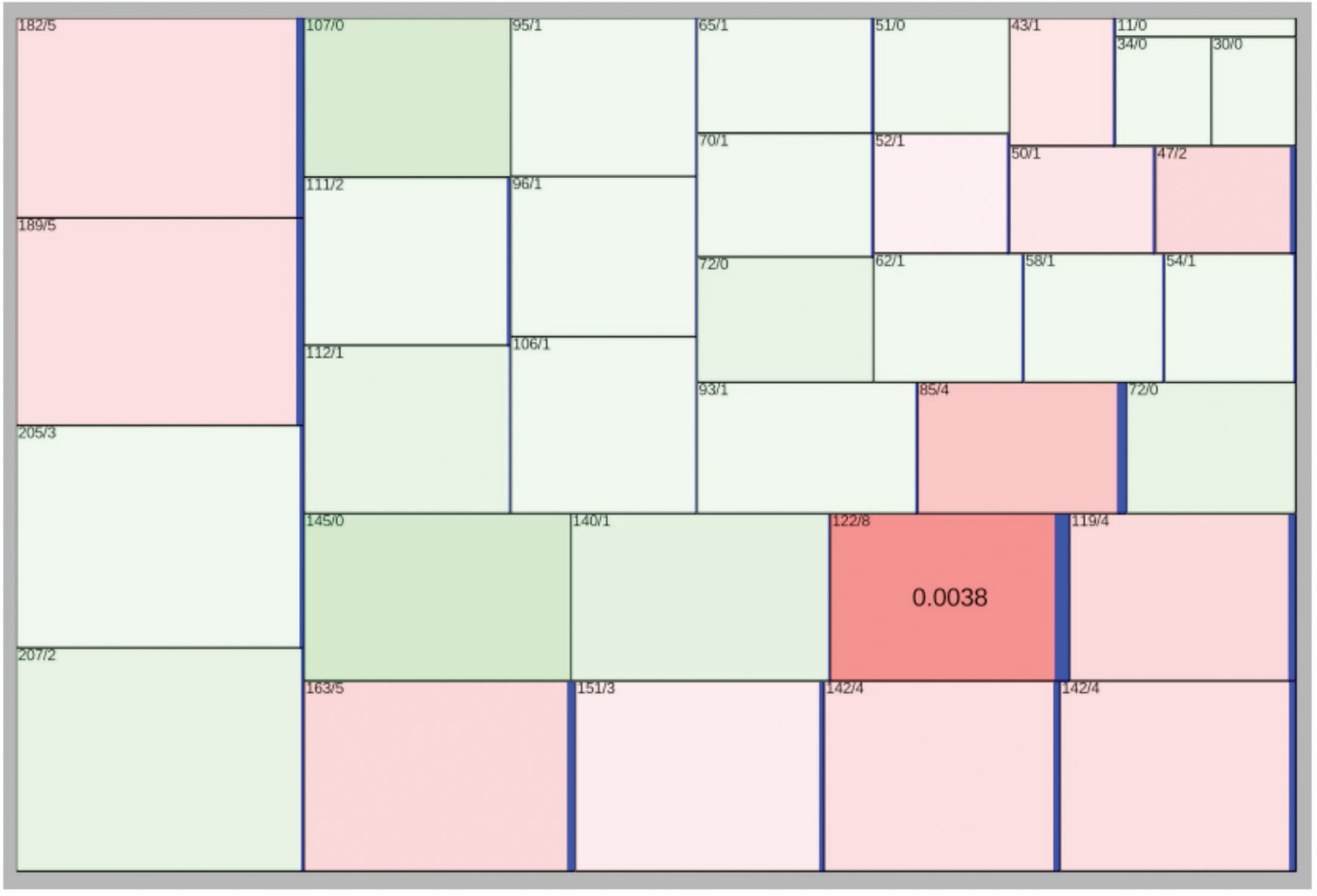
K-means Clustering (ICEs only)



A good classifier for known ncRNAs

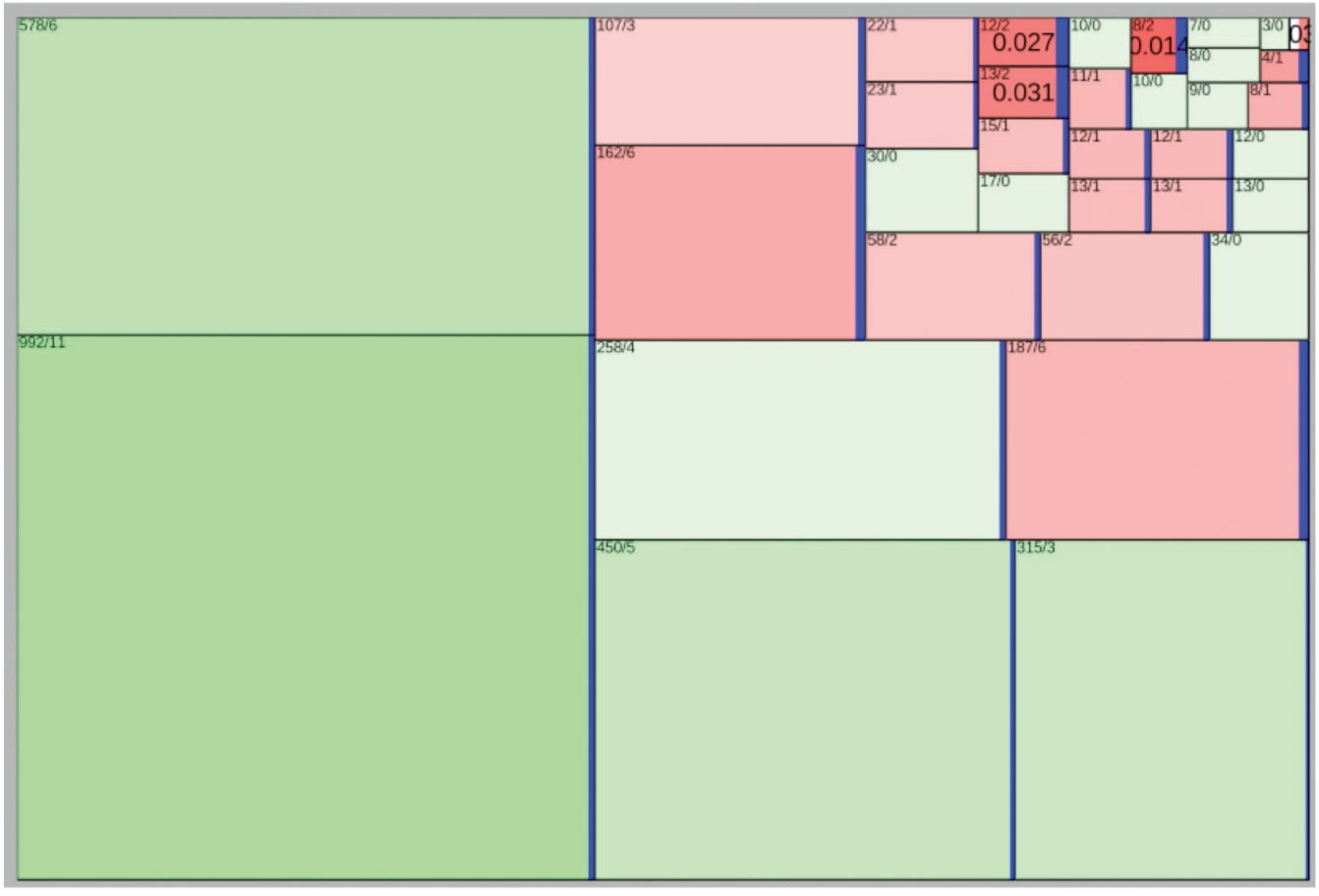


Clustering based on conservation



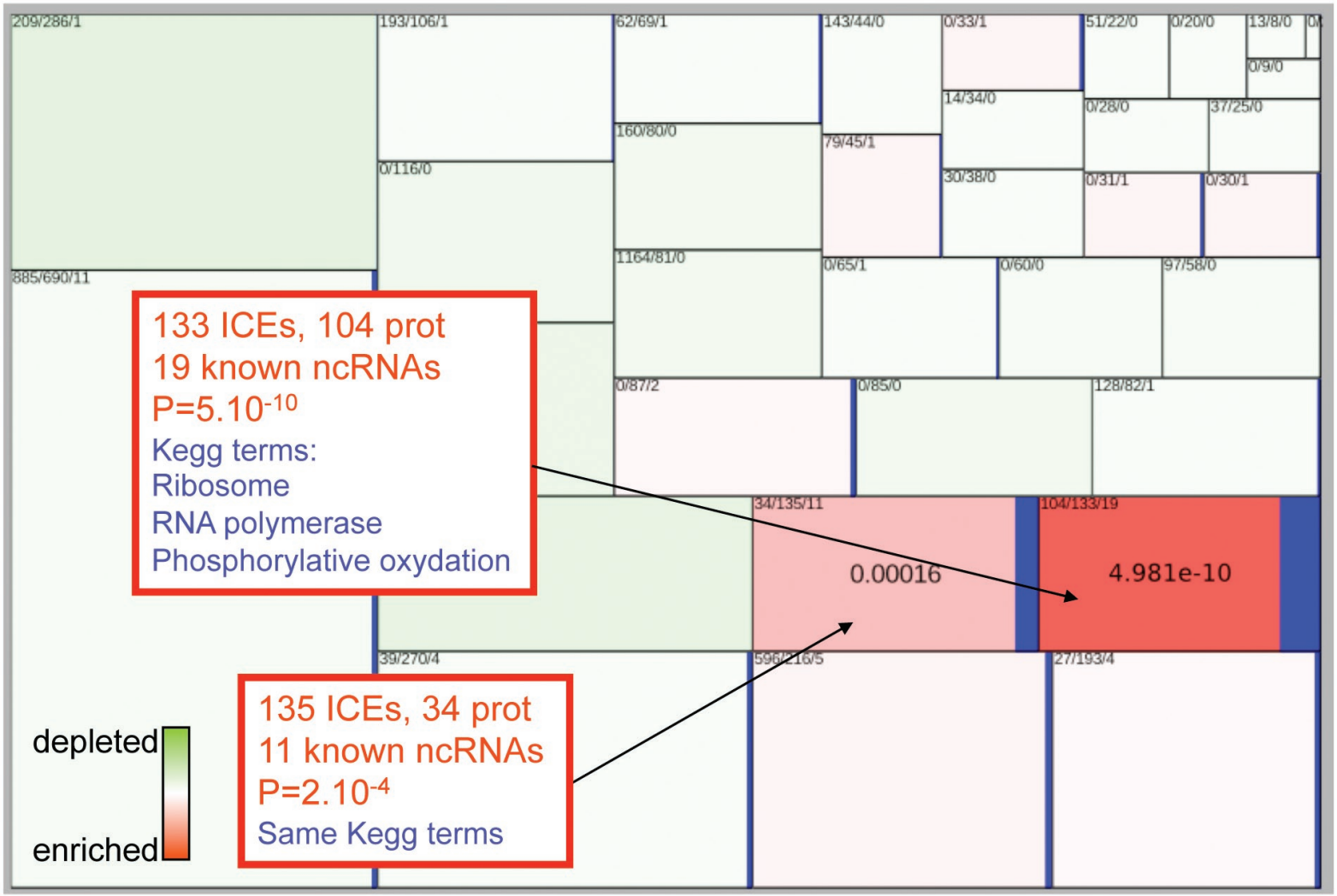
Best enrichment in known ncRNA: 4e-3

Clustering based on size



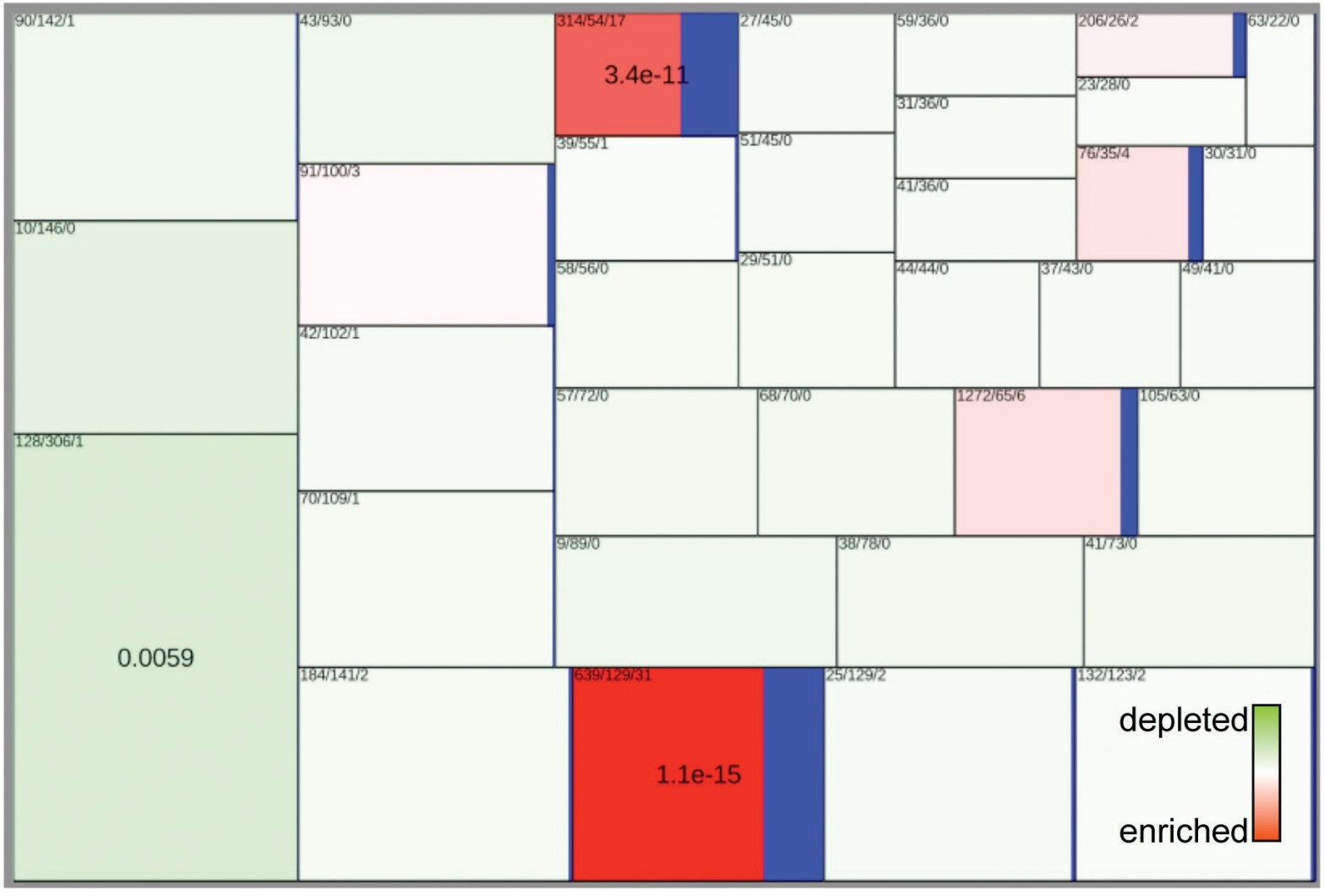
Size-based clustering. Best enrichment in known ncRNA: 1e-2

ICE + Protein clustering

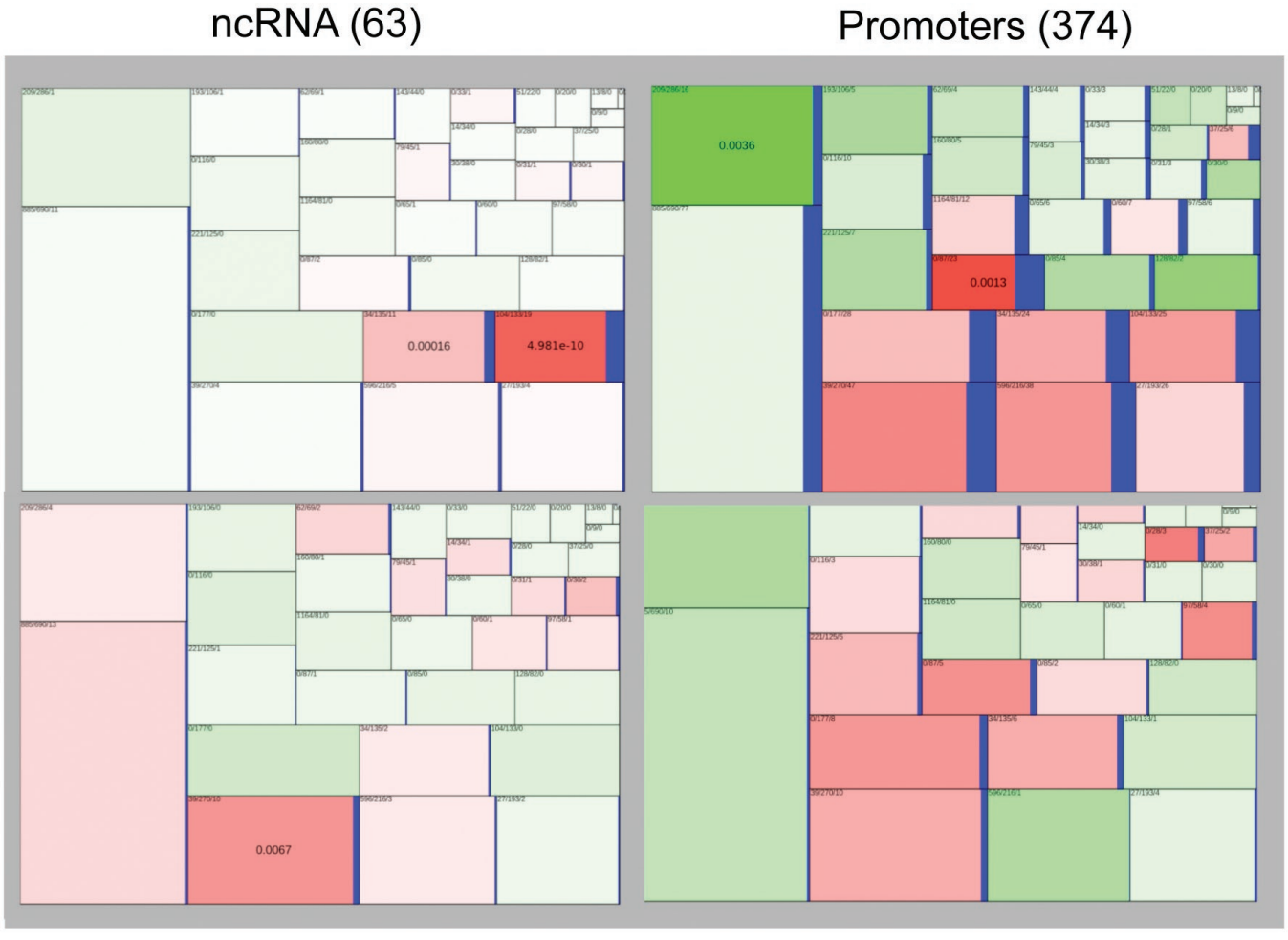


Better than ICEs alone

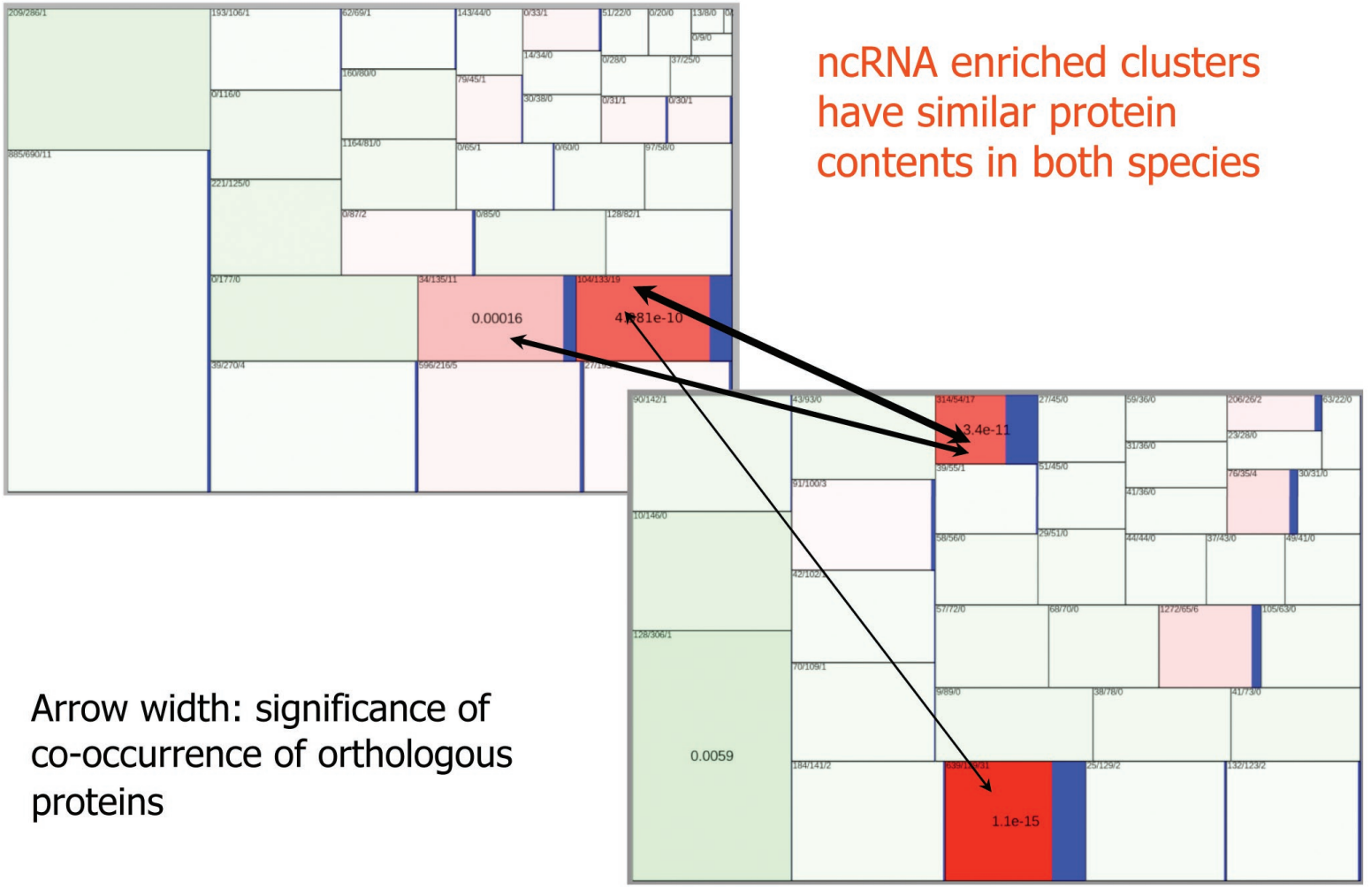
Results in *Bacillus subtilis*



Enrichment in other non-coding elements



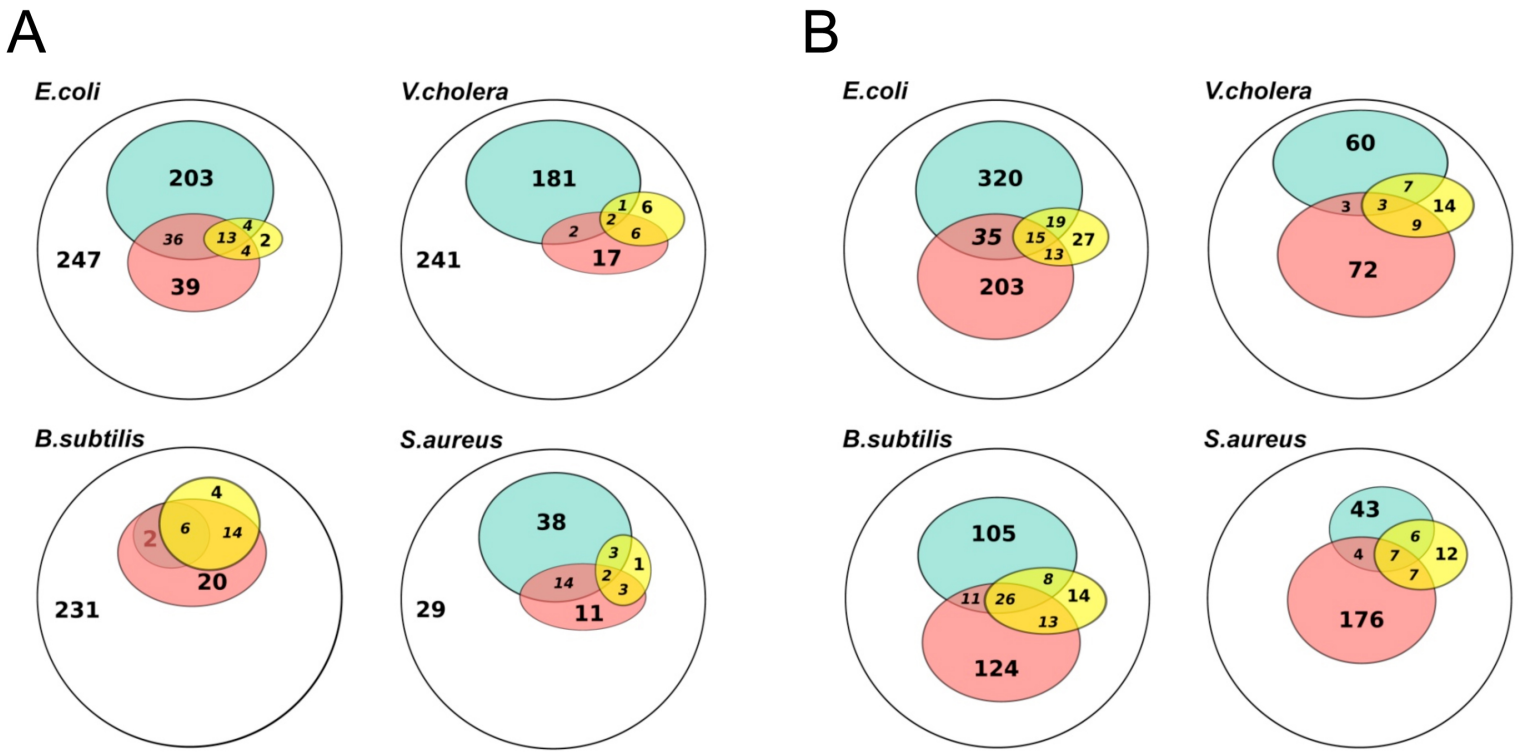
E.coli vs *B.subtilis*



ncRNA enriched clusters have similar protein contents in both species

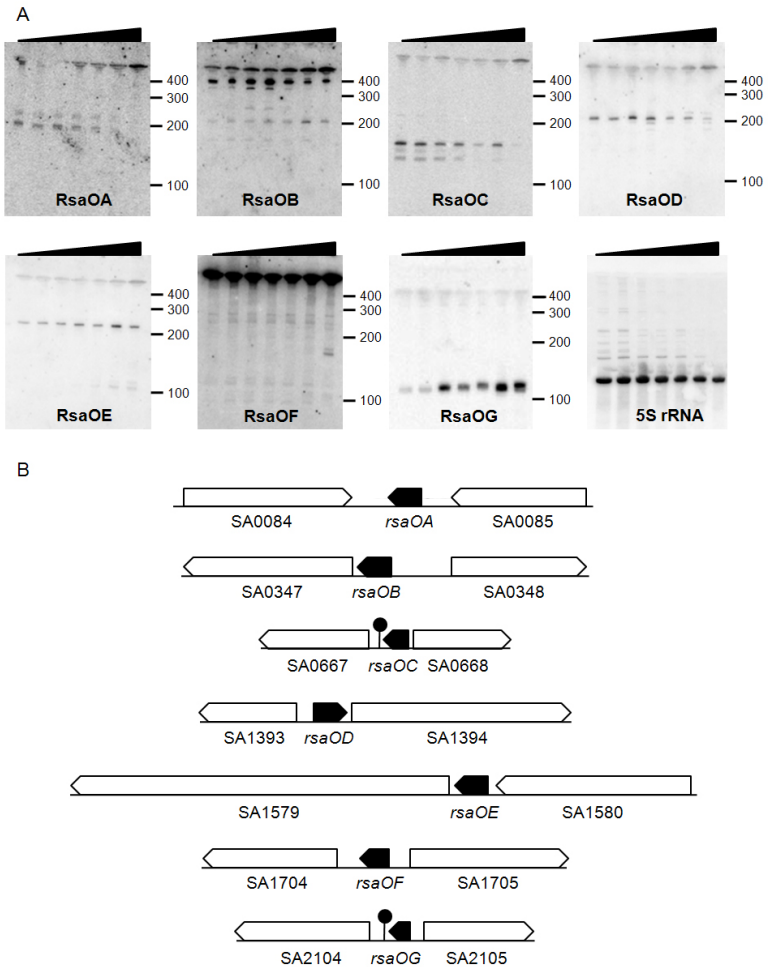
Arrow width: significance of co-occurrence of orthologous proteins

Performance of NAPP as a ncRNA classifier



These results show that NAPP clustering of ncRNAs compares favorably with two recent, specialized ncRNA predictions systems.

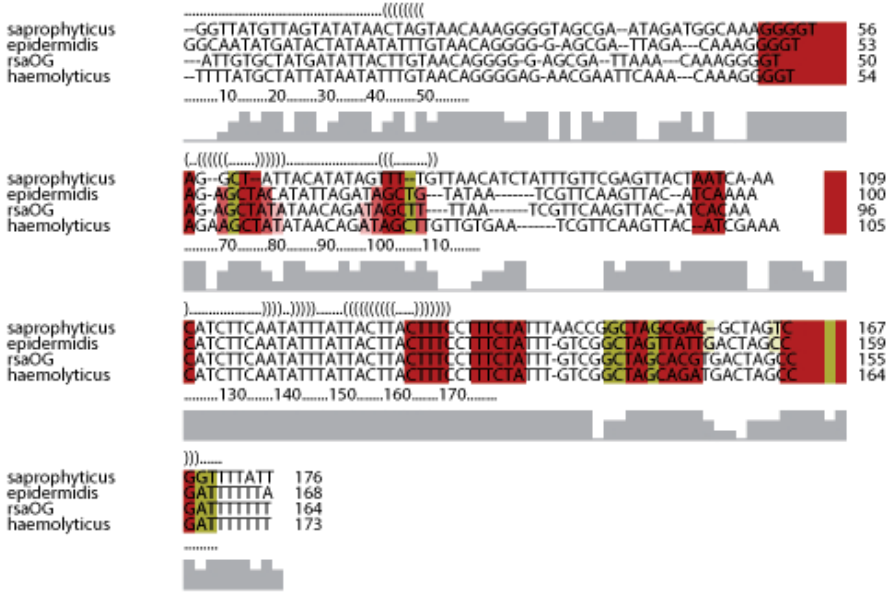
Experimental Validation of ncRNA candidates in *S.aureus*

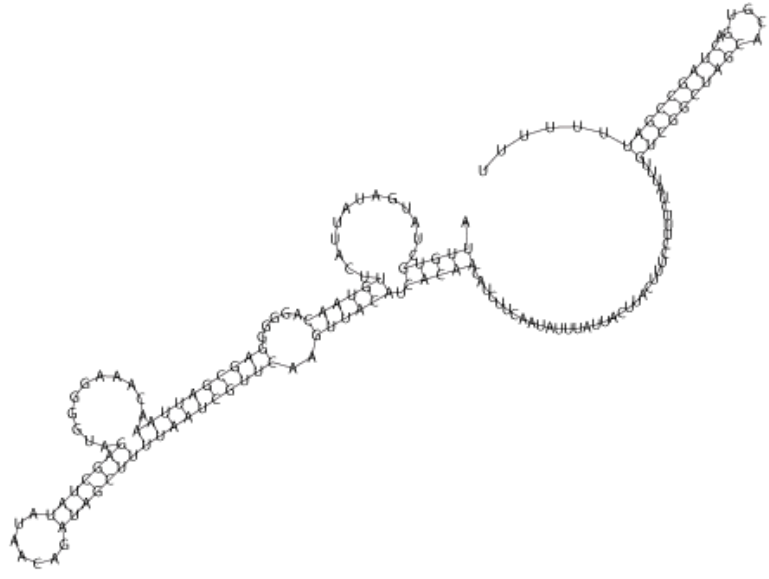


- Of 48 ncRNA predicted ICEs, we randomly selected 24 to be tested. 7 showed a transcript signal between 100 and 300 bp
- 4 putative riboswitches (T-box, SAM, PreQ1) and 3 putative non-coding RNAs acting in trans

Applications

New ncRNA family : RsaOG

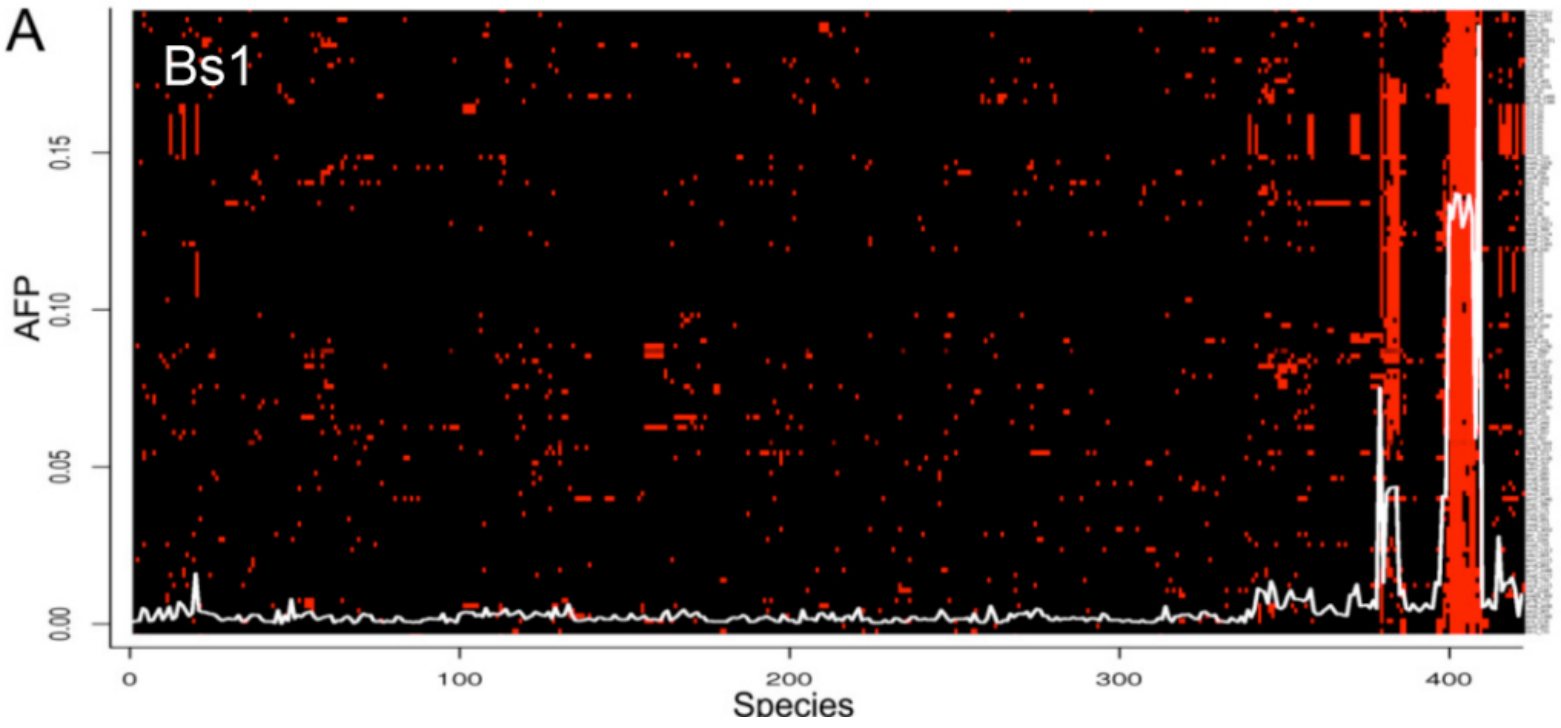




A new ncRNA family conserved in all *Staphylococcus*

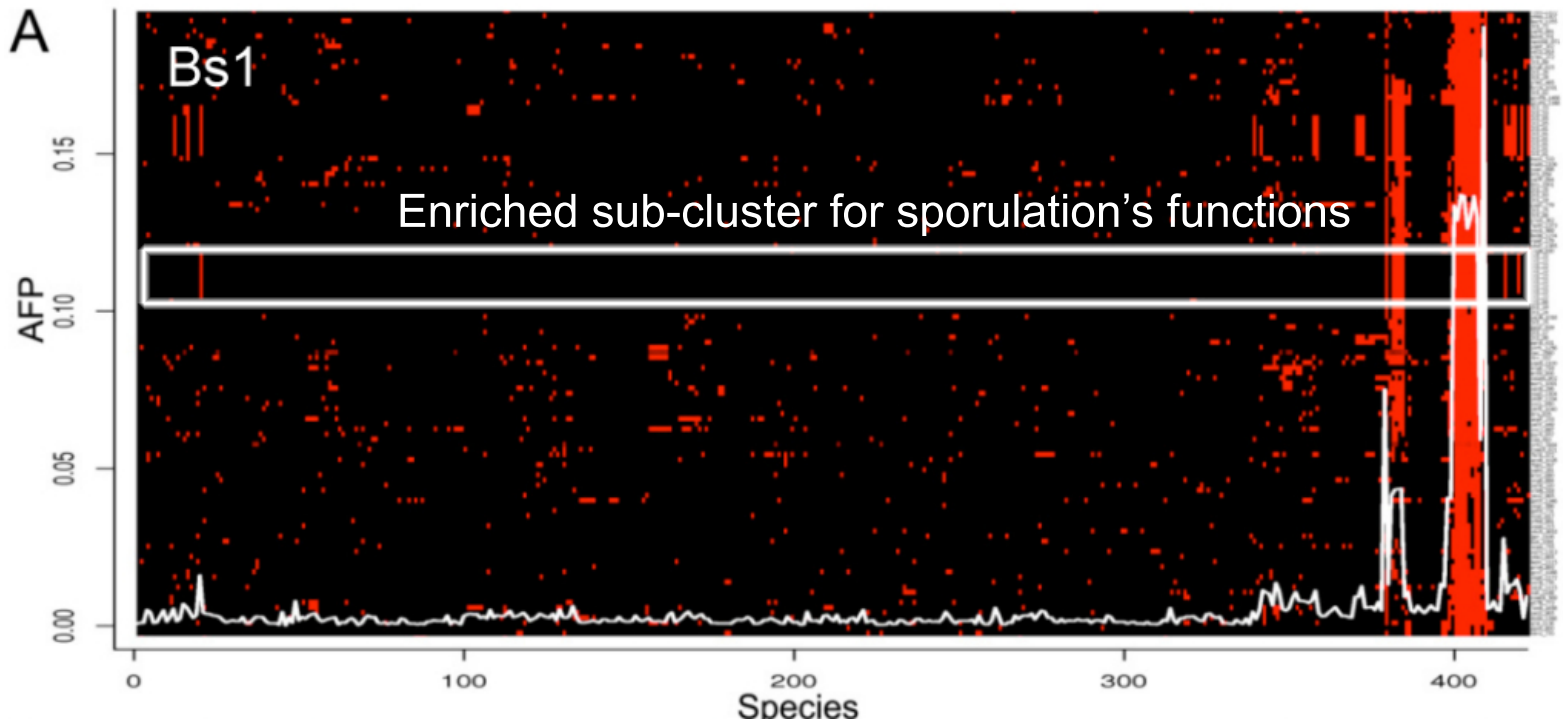
Find ncRNA by co-inheritance with a specific pathway or function

A ncRNA enriched cluster of *Bacillus subtilis*



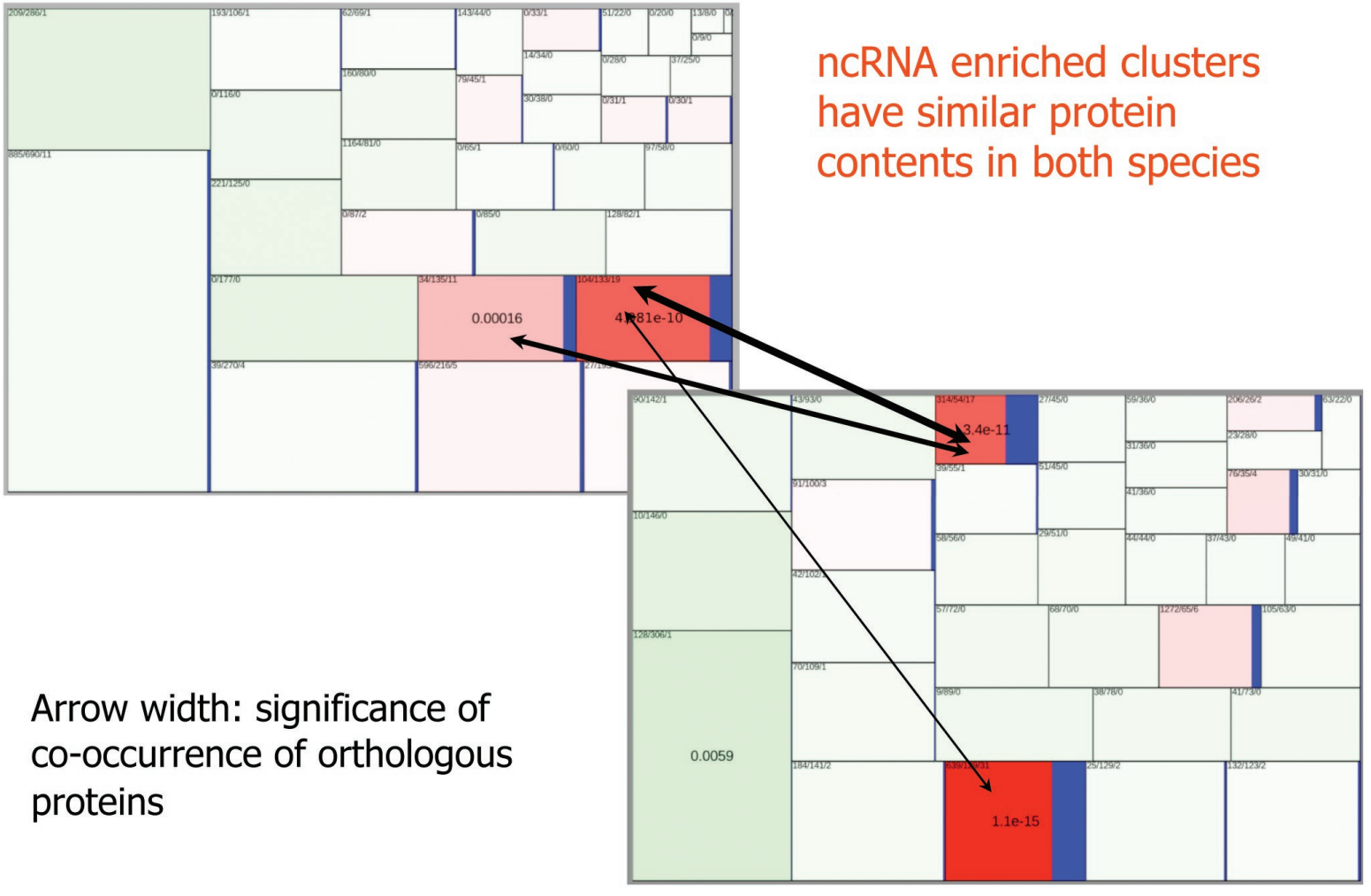
Find ncRNA by co-inheritance with a specific pathway or function

A ncRNA enriched cluster of *Bacillus subtilis*



ICEs clustering in a ncRNA enriched sub-group presenting a functional enrichment may be involved in the same function.

Search of homologous ncRNA related proteins



ncRNA enriched clusters have similar protein contents in both species

Arrow width: significance of co-occurrence of orthologous proteins

Conclusion and Perspectives

Currently, our method can

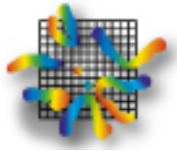
- Predict ncRNAs in any bacterial genome
- Processes a new genome in just a couple of hours
- With no expert intervention
- With a predictive performance that equals or beats structure/covariation based methods
- No sequence content analysis, no covariation, no folding energy
- Add information about functional cohineritance

Perspectives

- Improve of the alignment of ICEs against genome (Other alignment software)

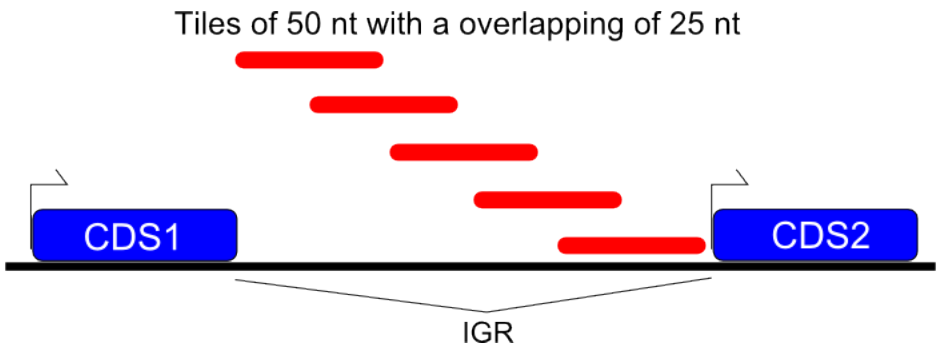
Perspectives

- Improve of the alignment of ICEs against genome (Other alignment software)
- Optimization of the empirical parameters



Perspectives

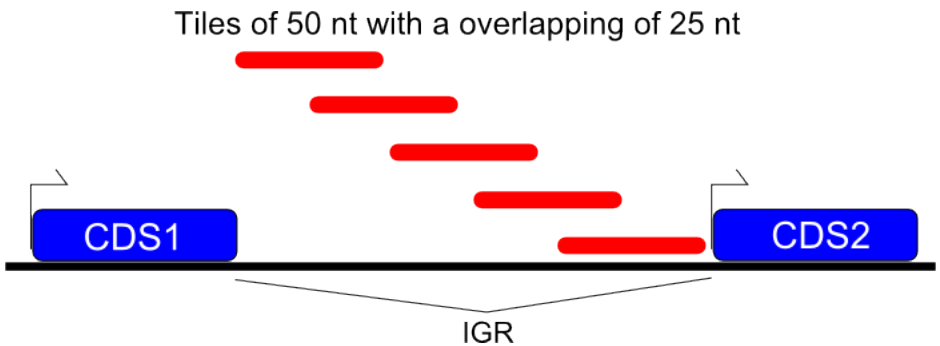
- Improve of the alignment of ICEs against genome (Other alignment software)
- Optimization of the empirical parameters
- Use Tilling as an alternative strategy to improve covering of genome:



Construct phylogenetic profiles for each Tile

Perspectives

- Improve of the alignment of ICEs against genome (Other alignment software)
- Optimization of the empirical parameters
- Use Tilling as an alternative strategy to improve covering of genome:



Construct phylogenetic profiles for each Tile

- Update the number of bacterial genomes in database and compute for all

Remerciements

- Sequence, Structure et Fonction des ARN (IGM):
  - Daniel Gautheret
  - Magali Naville
  - Chongjian Chen
  - Claire Toffano-Nioche
- Signalisation et Reseaux de regulations bacteriens (IGM):
  - Philippe Bouloc
  - Chantal Bohn
  - Patricia Skorski
- Laboratoire de regulation de l'expression genetique chez les microorganismes (IBPC):
  - Patrick Stragier
- More information in Marchais et al, Genome Res, 2009