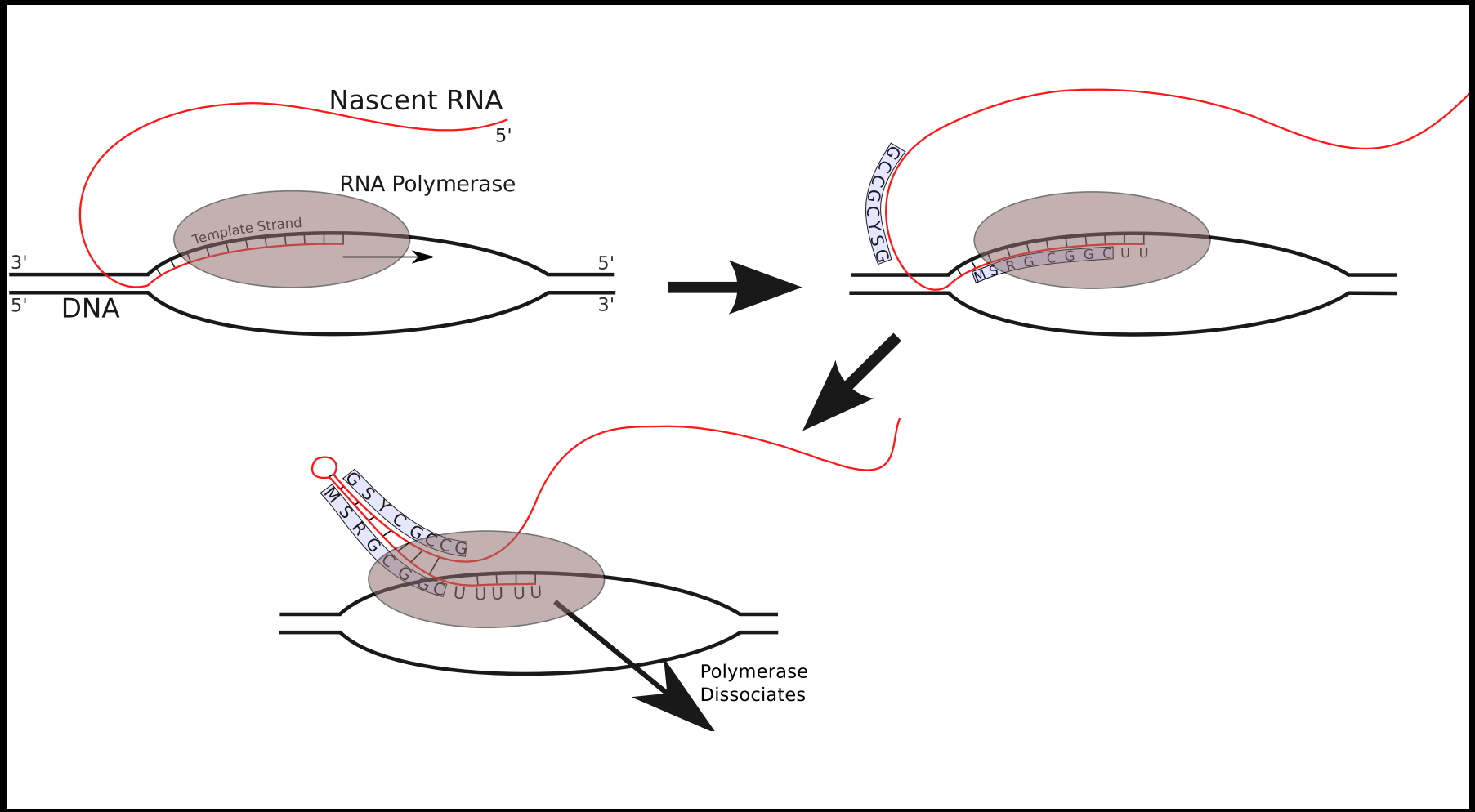
Discovering bacterial transcription termination associated motifs through whole genome clustering

Lars Barquist

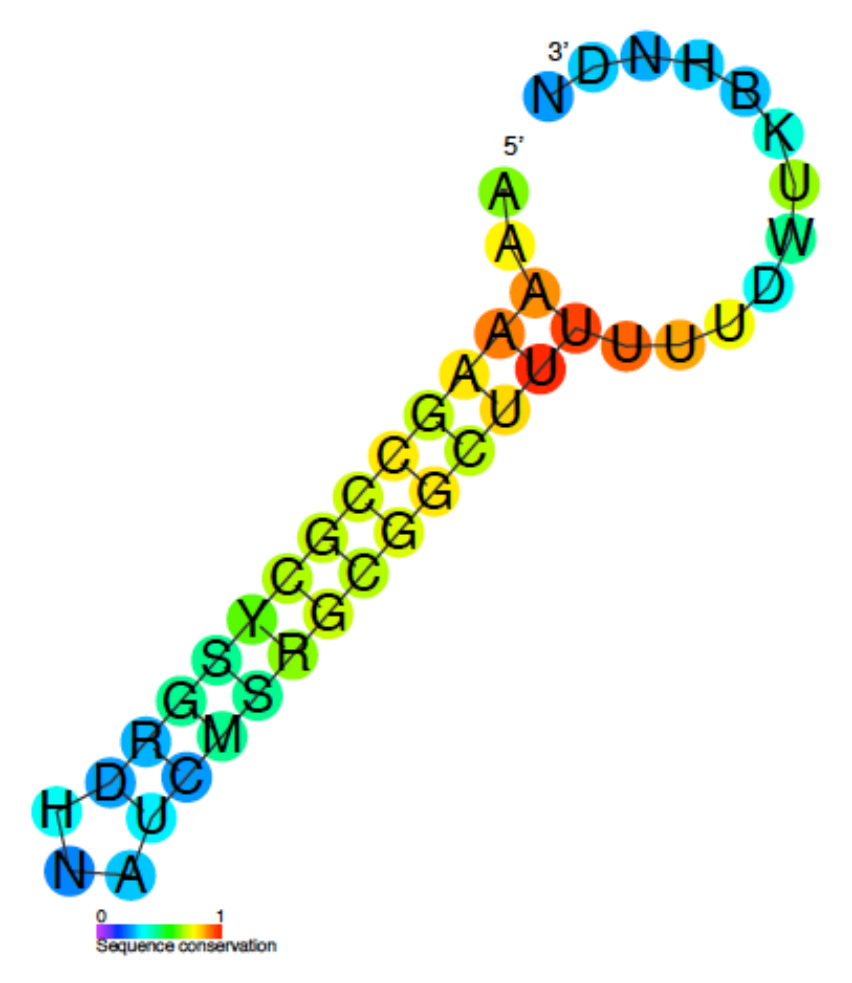
Rho-independent termination



Why care?

- Genomic “landmarks”
- Cis-regulatory components
- Fundamental biological process

A CM for Rho-independent terminators



Constructed from an alignment of 171 E. coli and 891 B. subtilis experimentally validated terminator sequences.

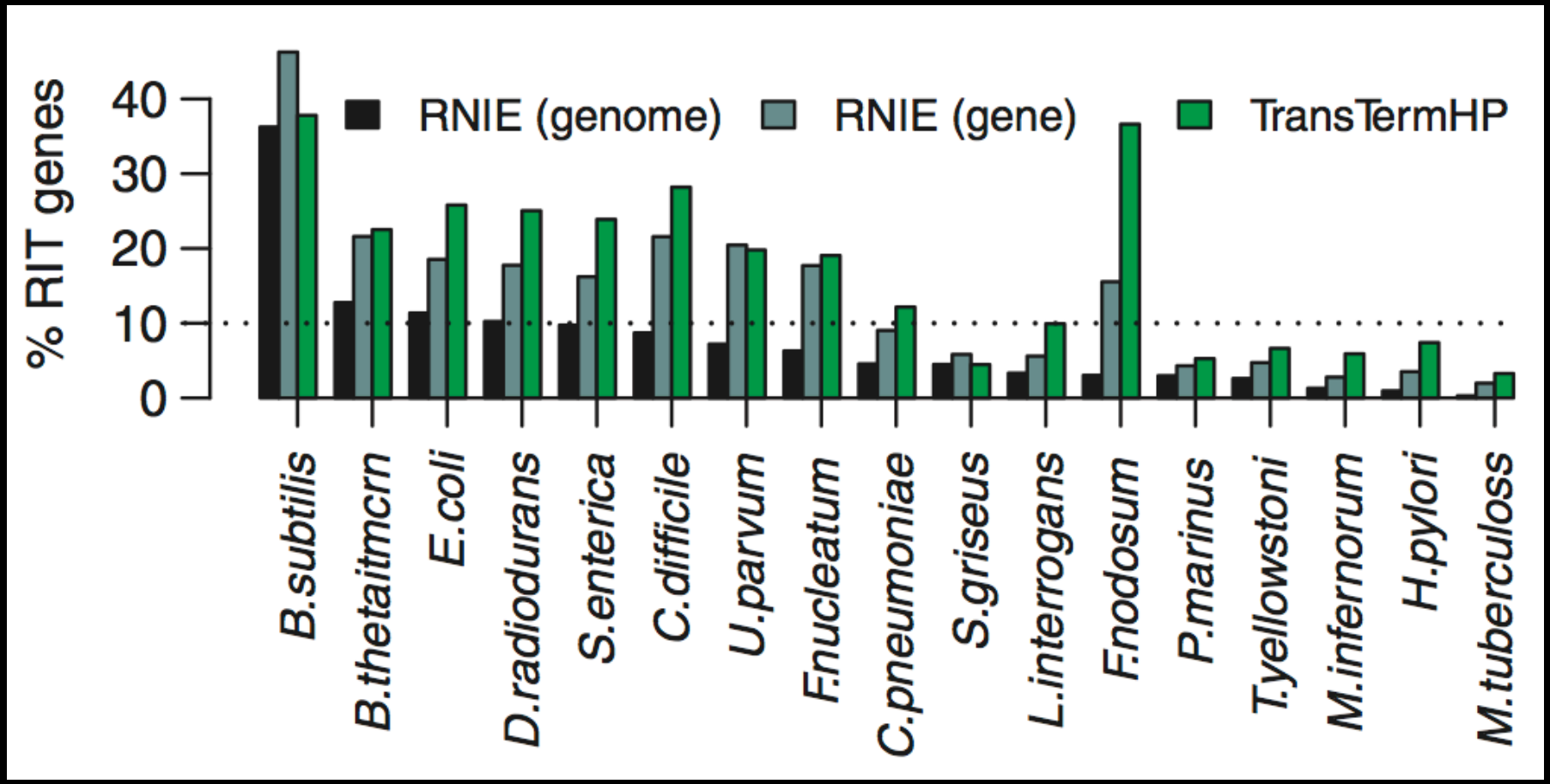
Published online 7 April 2011 *Nucleic Acids Research, 2011, Vol. 39, No. 14* **5845–5852**
doi:10.1093/nar/gkr168

RNIE: genome-wide prediction of bacterial intrinsic terminators

Paul P. Gardner[1,*], Lars Barquist[1], Alex Bateman[1], Eric P. Nawrocki[2] and Zasha Weinberg[3]

[1]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton CB10 1SA0, UK, [2]Janelia Farm Research Campus, Howard Hughes Medical Institute, Ashburn, VA 20147 and [3]Howard Hughes Medical Institute, Yale University, Box 208103, New Haven, CT 06520, USA

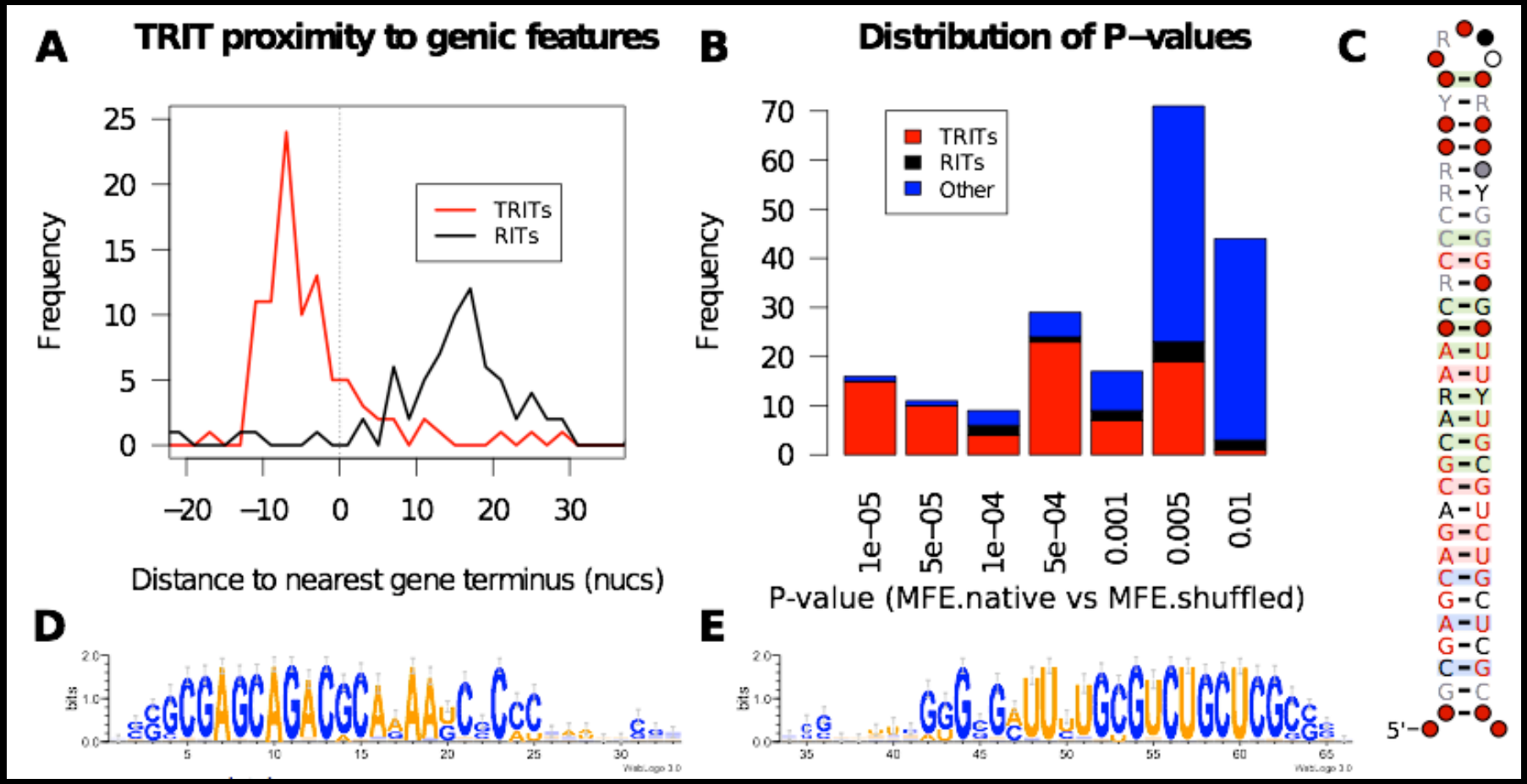
Missing terminators?



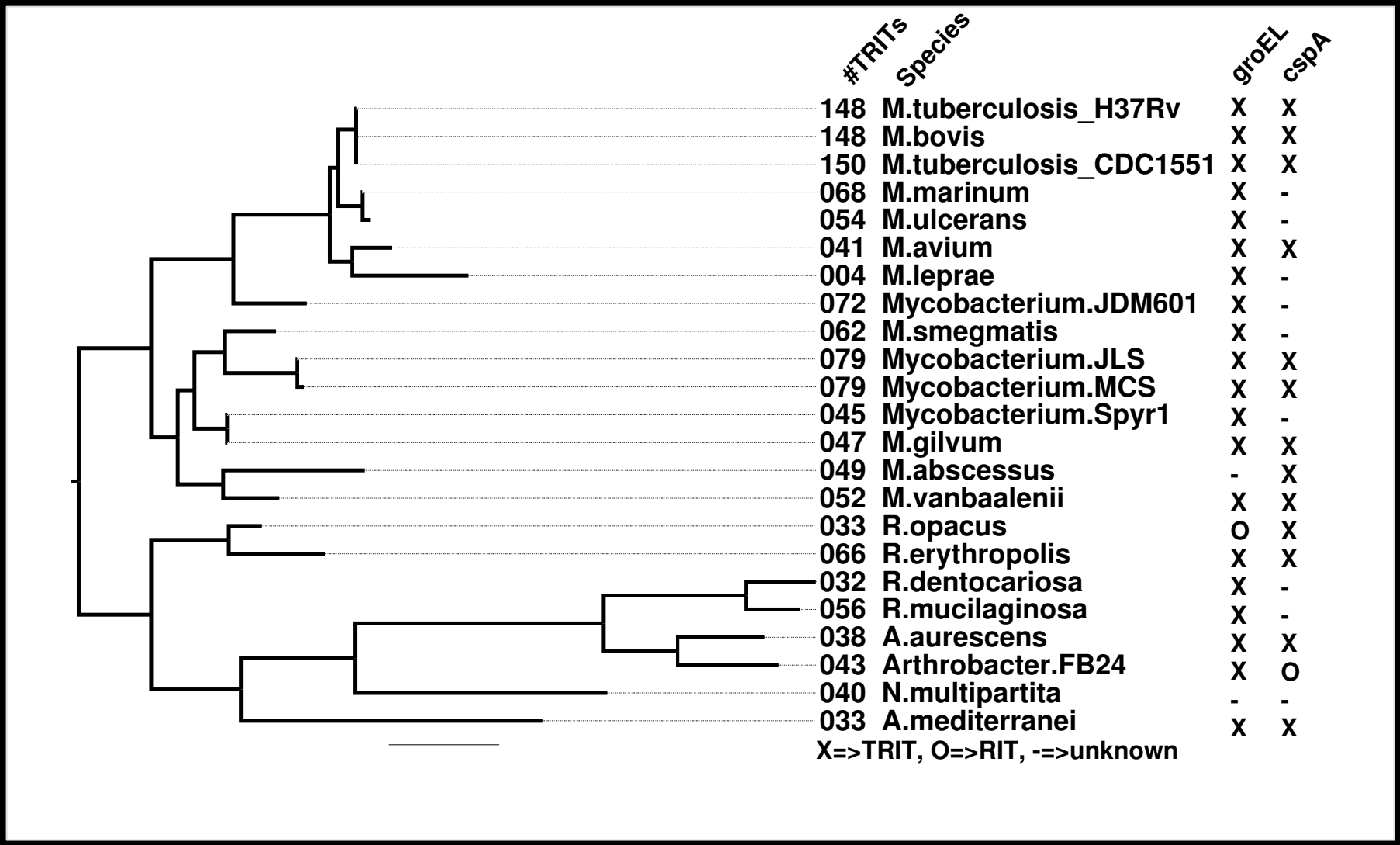
>MT1104|100-bp terminator sequence|1198311,1198410|strand:-1
UGAUCGAGCGCCUCGCGUGAGCGAGGCGCUCGAGUGCGUUGACUCUGCGUCCACCACGCAAAAGUGCGAGUAGGACACGUGGUGGGCGCAGAGUCAACGU
>MT2818|100-bp terminator sequence|3054490,3054589|strand:1
GGAUGCUGCUGGUGCUGUAGCCCGGCGAGCAGACGCAAAAUCGCCUCAUUUCGGCACGAAAUGGGCGAUUUUGCGUCUGCUCGGCGGGCUACUCGCCGCC
>MT0690|100-bp terminator sequence|757228,757327|strand:-1
GCCUGCUCGGCGACCGCUGACGGCCCCGGGCGAGCAGUCGCAUAAGCCCCCGACACGCCGAGCGUGCGGGGGCUUAUGCGUCUGCUCGCCGGCCGACAAU
>MT0884|100-bp terminator sequence|958341,958440|strand:-1
UGCUGGGCCCGGCAAUUUAGGUUGCGCGAGCAGACGUAAAAGCCCCCGACACGCCGRGCGUGCGGGGGCUUUUACGUCUGCUCGCGCUCAGCUUAGCAGC
>MT2121|100-bp terminator sequence|2319422,2319521|strand:-1
AGGGCGACCUGGAAGCCUGAGUACGGCGAGCAGACGCAGAAUCGCAUGCGCGGGACCCCGCGCAGUGCGAUUCUGCGUCUGCUCGCGCGGUAGGUUGGCA
>MT1437|100-bp terminator sequence|1567875,1567974|strand:1
ACCUCAAGCGCGCCAUCUAGCGUCGAGGGCGCGAGCAGACGCAGAAUCGCACGCGGAAAGGCUUCCGCGUGCGAUUCUGCGUCUGCUCGGCGCUAGCUGC
>MT1127|100-bp terminator sequence|1223826,1223925|strand:1
AGAUCACCGGGCCGCGCUGAGUGCGCCUCCCGCGAGCAGACACAGAAUCGCACUGCGCCGGCCCGGCGCGUGCGAUUCUGUGUCUGCUCGCCGGUAGACU
>MT0244|100-bp terminator sequence|279624,279723|strand:1
UGGCAGCGGCCUCGGCCUAGGCCUGGCGAGCAGACGCAAAAUCGCCCAAUUUCGUGCCGAAUUGGGCGAUUUUGCGUCUGCUCGCCAGGGGAACGCUAGG
>MT2819|100-bp terminator sequence|3054498,3054597|strand:-1
GGGGUGACGGCGGCGAGUAGCCCGCCGAGCAGACGCAAAAUCGCCCAUUUCGUGCCGAAAUGAGGCGAUUUUGCGUCUGCUCGCCGGGCUACAGCACCAG
>MT3684|100-bp terminator sequence|4013504,4013603|strand:1
CCCAGGUUCUGGGGAUCUAGCCGCAAGGGCGCGAGCAGACGCAGAAUCGCAUGAUUUGAGCUCAAAUCAUGCGAUUCUGCGUCUGCUCGCGAGGCUCGCG
>MT1438|100-bp terminator sequence|1567887,1567986|strand:-1
ACUACCGCAUCAGCAGCUAGCGCCGAGCAGACGCAGAAUCGCACGCGGAAGCCUUUCCGCGUGCGAUUCUGCGUCUGCUCGCGCCCUCGACGCUAGAUGG
>MT0208|100-bp terminator sequence|234550,234649|strand:-1
GGGUCCGGAUCUGGAACUAGGCUAAGACCCCGCGAGCAGACGCAGAAUCGCCCAUUUGGGCUACCAAAUGGGCGAUUCUGCGUCUGCUCGGCGGAUGGCG
>MT2865|100-bp terminator sequence|3099695,3099794|strand:-1
AACCCUCGCCGACCCCGUAGCAGUUGGGCUAGGUUGGCCGAGCAGACGCAAAAGCCCCCAUUUCGGGCCCGAAAUGGGGGCUUUUGCGUCUGCUCACGCC
>MT3420|100-bp terminator sequence|3702432,3702531|strand:1
CGCUGAUGUUCACCCGCUGAGGGCUUGCGCGAGCAGACGCAAAAUCGCCCGAAAACCAGUGGUUUUGGGCGAUUUUGCGUCUGCUCGCGCAGCCGGGUCU
>MT0670|100-bp terminator sequence|737658,737757|strand:-1
AGCCGGGUGCCGCGGCCUAAGCCCGGCGAGCAGACGCAUAAGCCCCCGACACGCCGUGCGUGCGGGGGCUUAUGCGUCGCCGGGGAAACUACUCCCCCGC
>MT0669.1|100-bp terminator sequence|737650,737749|strand:1
GCAACUUCGCGGGGGAGUAGUUUCCCCGGCGACGCAUAAGCCCCCGCACGCACGGCGUGUCGGGGGCUUAUGCGUCUGCUCGCCGGGCUUAGGCCGCGGC
>MT3441|100-bp terminator sequence|3722137,3722236|strand:1
CGCAUGUCCGGGCCAGUUAGCGGGCGCGAGCAGACGCAAAAUCGCCCAUUUCGGCACGAAAUUGGGCGAUUUUGCGUCUGCUCGCCCUAAUUGGCCAGCU
>MT0245|100-bp terminator sequence|279639,279738|strand:-1
AGACCGUCUGGAUCGCCUAGCGUUCCCCUGGCGAGCAGACGCAAAAUCGCCCAAUUCGGCACGAAAUUGGGCGAUUUUGCGUCUGCUCGCCAGGCCUAGG
>MT0883|100-bp terminator sequence|958334,958433|strand:1
CGGAGUCGCUGCUAAGCUGAGCGCGAGCAGACGUAAAAGCCCCCGCACGCYCGGCGUGUCGGGGGCUUUUACGUCUGCUCGCGCAACCUAAAUUGCCGGG

>MT1104|100-bp terminator sequence|1198311,1198410|strand:-1
UGAUCGAGCGCCUCGCGUGAGCGAGGCGCUCGAGUGCGUUGACUCUGCGUCCACCACGCAAAAGUGCGAGUAGGACACGUGGUGGGCGCAGAGUCAACGU
>MT2818|100-bp terminator sequence|3054490,3054589|strand:1
GGAUGCUGCUGGUGCUGUAGCCCGGCGAGCAGACGCAAAAUCGCCUCAUUUCGGCACGAAAUGGGCGAUUUUGCGUCUGCUCGGCGGGCUACUCGCCGCC
>MT0690|100-bp terminator sequence|757228,757327|strand:-1
GCCUGCUCGGCGACCGCUGACGGCCCCGGGCGAGCAGUCGCAUAAGCCCCCGACACGCCGAGCGUGCGGGGGCUUAUGCGUCUGCUCGCCGGCCGACAAU
>MT0884|100-bp terminator sequence|958341,958440|strand:-1
UGCUGGGCCCGGCAAUUUAGGUUGCGCGAGCAGACGUAAAAGCCCCCGACACGCCGRGCGUGCGGGGGCUUUUACGUCUGCUCGCGCUCAGCUUAGCAGC
>MT2121|100-bp terminator sequence|2319422,2319521|strand:-1
AGGGCGACCUGGAAGCCUGAGUACGGCGAGCAGACGCAGAAUCGCAUGCGCGGGACCCCGCGCAGUGCGAUUCUGCGUCUGCUCGCGCGGUAGGUUGGCA
>MT1437|100-bp terminator sequence|1567875,1567974|strand:1
ACCUCAAGCGCGCCAUCUAGCGUCGAGGGCGCGAGCAGACGCAGAAUCGCACGCGGAAAGGCUUCCGCGUGCGAUUCUGCGUCUGCUCGGCGCUAGCUGC
>MT1127|100-bp terminator sequence|1223826,1223925|strand:1
AGAUCACCGGGCCGCGCUGAGUGCGCCUCCCGCGAGCAGACACAGAAUCGCACUGCGCCGGCCCGGCGCGUGCGAUUCUGUGUCUGCUCGCCGGUAGACU
>MT0244|100-bp terminator sequence|279624,279723|strand:1
UGGCAGCGGCCUCGGCCUAGGCCUGGCGAGCAGACGCAAAAUCGCCCAAUUUCGUGCCGAAUUGGGCGAUUUUGCGUCUGCUCGCCAGGGGAACGCUAGG
>MT2819|100-bp terminator sequence|3054498,3054597|strand:-1
GGGGUGACGGCGGCGAGUAGCCCGCCGAGCAGACGCAAAAUCGCCCAUUUCGUGCCGAAAUGAGGCGAUUUUGCGUCUGCUCGCCGGGCUACAGCACCAG
>MT3684|100-bp terminator sequence|4013504,4013603|strand:1
CCCAGGUUCUGGGGAUCUAGCCGCAAGGGCGCGAGCAGACGCAGAAUCGCAUGAUUUGAGCUCAAAUCAUGCGAUUCUGCGUCUGCUCGCGAGGCUCGCG
>MT1438|100-bp terminator sequence|1567887,1567986|strand:-1
ACUACCGCAUCAGCAGCUAGCGCCGAGCAGACGCAGAAUCGCACGCGGAAGCCUUUCCGCGUGCGAUUCUGCGUCUGCUCGCGCCCUCGACGCUAGAUGG
>MT0208|100-bp terminator sequence|234550,234649|strand:-1
GGGUCCGGAUCUGGAACUAGGCUAAGACCCCGCGAGCAGACGCAGAAUCGCCCAUUUGGGCUACCAAAUGGGCGAUUCUGCGUCUGCUCGGCGGAUGGCG
>MT2865|100-bp terminator sequence|3099695,3099794|strand:-1
AACCCUCGCCGACCCCGUAGCAGUUGGGCUAGGUUGGCCGAGCAGACGCAAAAGCCCCCAUUUCGGGCCCGAAAUGGGGGCUUUUGCGUCUGCUCACGCC
>MT3420|100-bp terminator sequence|3702432,3702531|strand:1
CGCUGAUGUUCACCCGCUGAGGGCUUGCGCGAGCAGACGCAAAAUCGCCCGAAAACCAGUGGUUUUGGGCGAUUUUGCGUCUGCUCGCGCAGCCGGGUCU
>MT0670|100-bp terminator sequence|737658,737757|strand:-1
AGCCGGGUGCCGCGGCCUAAGCCCGGCGAGCAGACGCAUAAGCCCCCGACACGCCGUGCGUGCGGGGGCUUAUGCGUCGCCGGGGAAACUACUCCCCCGC
>MT0669.1|100-bp terminator sequence|737650,737749|strand:1
GCAACUUCGCGGGGGAGUAGUUUCCCCGGCGACGCAUAAGCCCCCGCACGCACGGCGUGUCGGGGGCUUAUGCGUCUGCUCGCCGGGCUUAGGCCGCGGC
>MT3441|100-bp terminator sequence|3722137,3722236|strand:1
CGCAUGUCCGGGCCAGUUAGCGGGCGCGAGCAGACGCAAAAUCGCCCAUUUCGGCACGAAAUUGGGCGAUUUUGCGUCUGCUCGCCCUAAUUGGCCAGCU
>MT0245|100-bp terminator sequence|279639,279738|strand:-1
AGACCGUCUGGAUCGCCUAGCGUUCCCCUGGCGAGCAGACGCAAAAUCGCCCAAUUCGGCACGAAAUUGGGCGAUUUUGCGUCUGCUCGCCAGGCCUAGG
>MT0883|100-bp terminator sequence|958334,958433|strand:1
CGGAGUCGCUGCUAAGCUGAGCGCGAGCAGACGUAAAAGCCCCCGCACGCYCGGCGUGUCGGGGGCUUUUACGUCUGCUCGCGCAACCUAAAUUGCCGGG

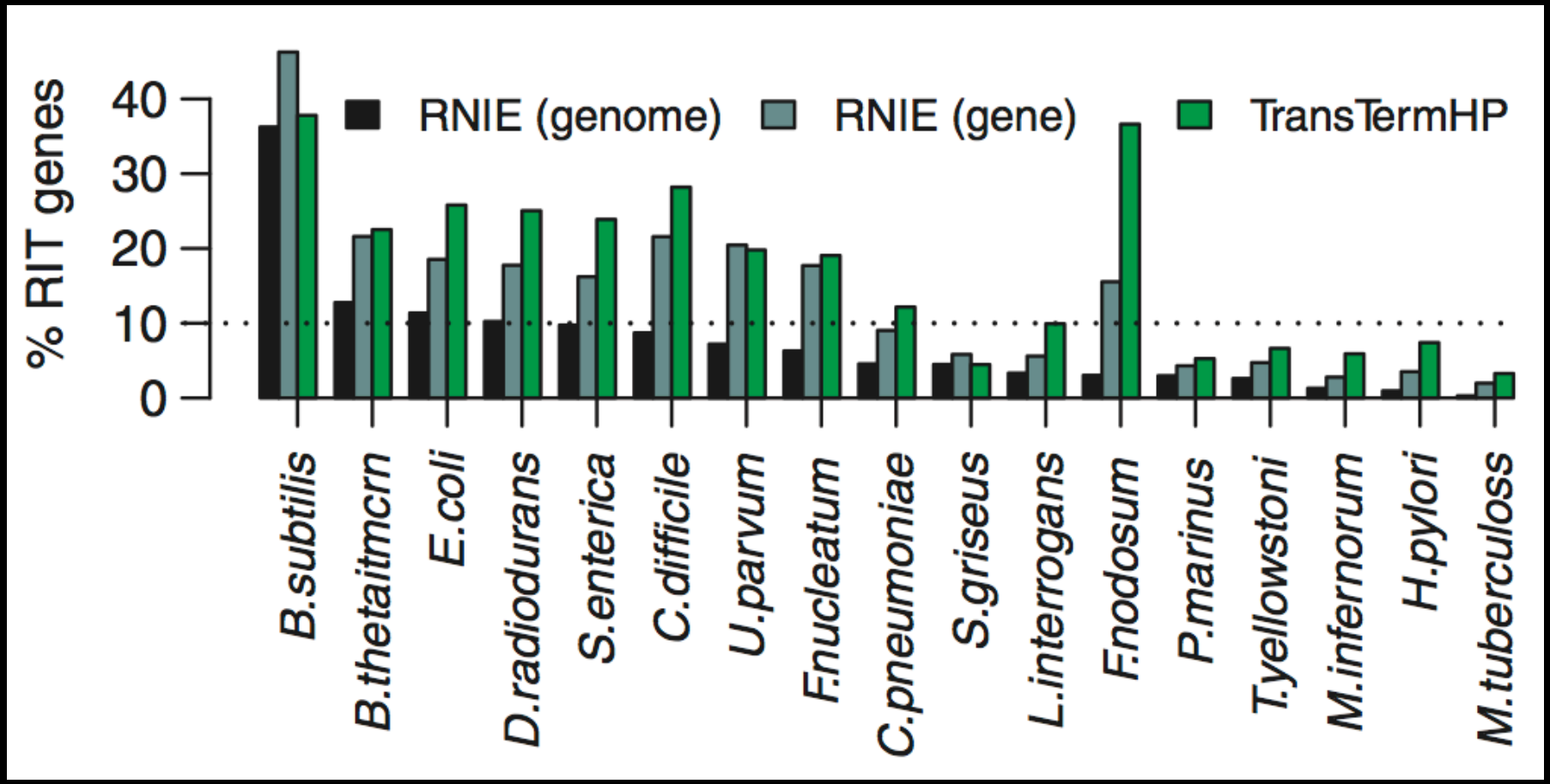
Mycobacterium tuberculosis



Missing terminators?



Missing terminators?



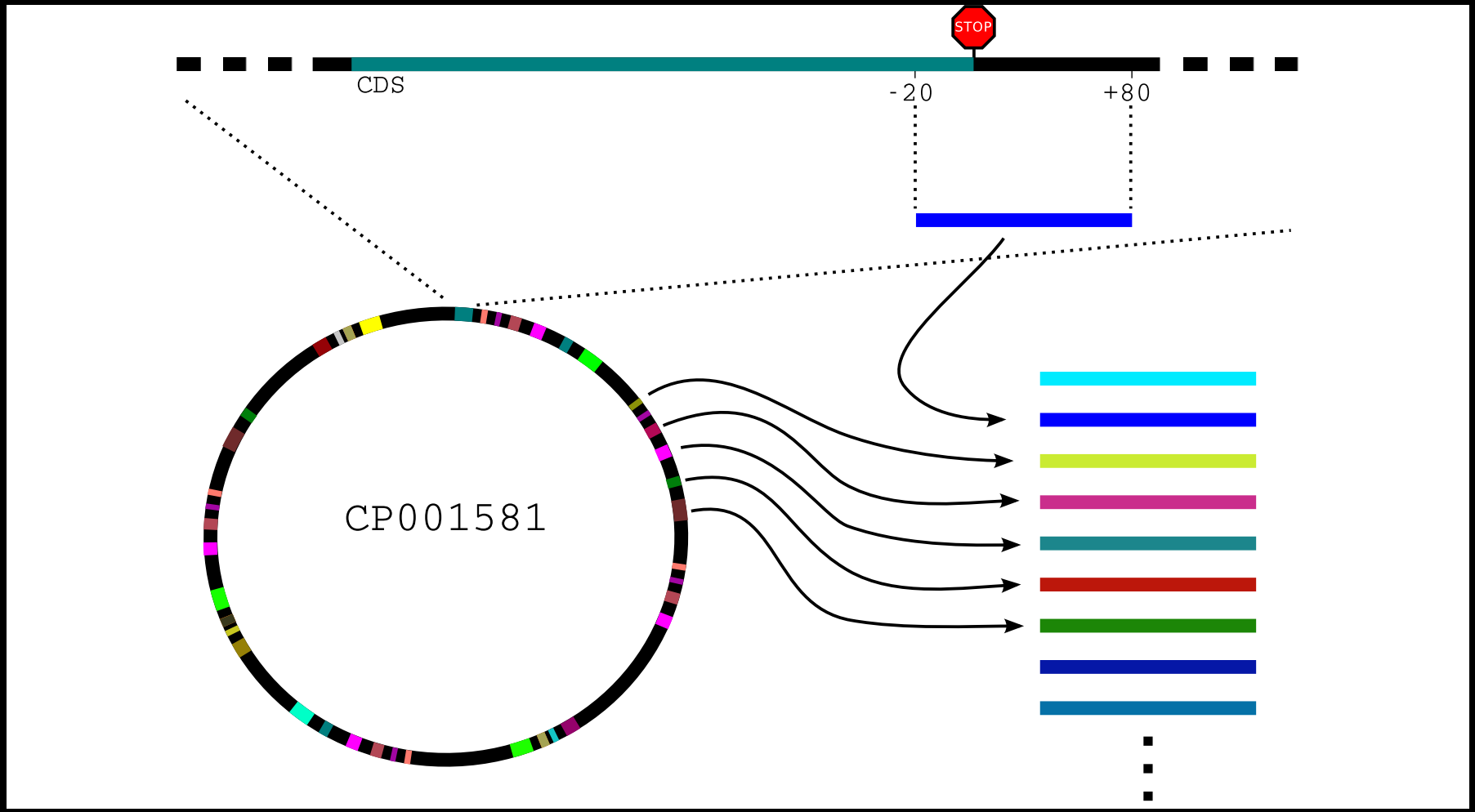
Questions

Are there terminator “families”?

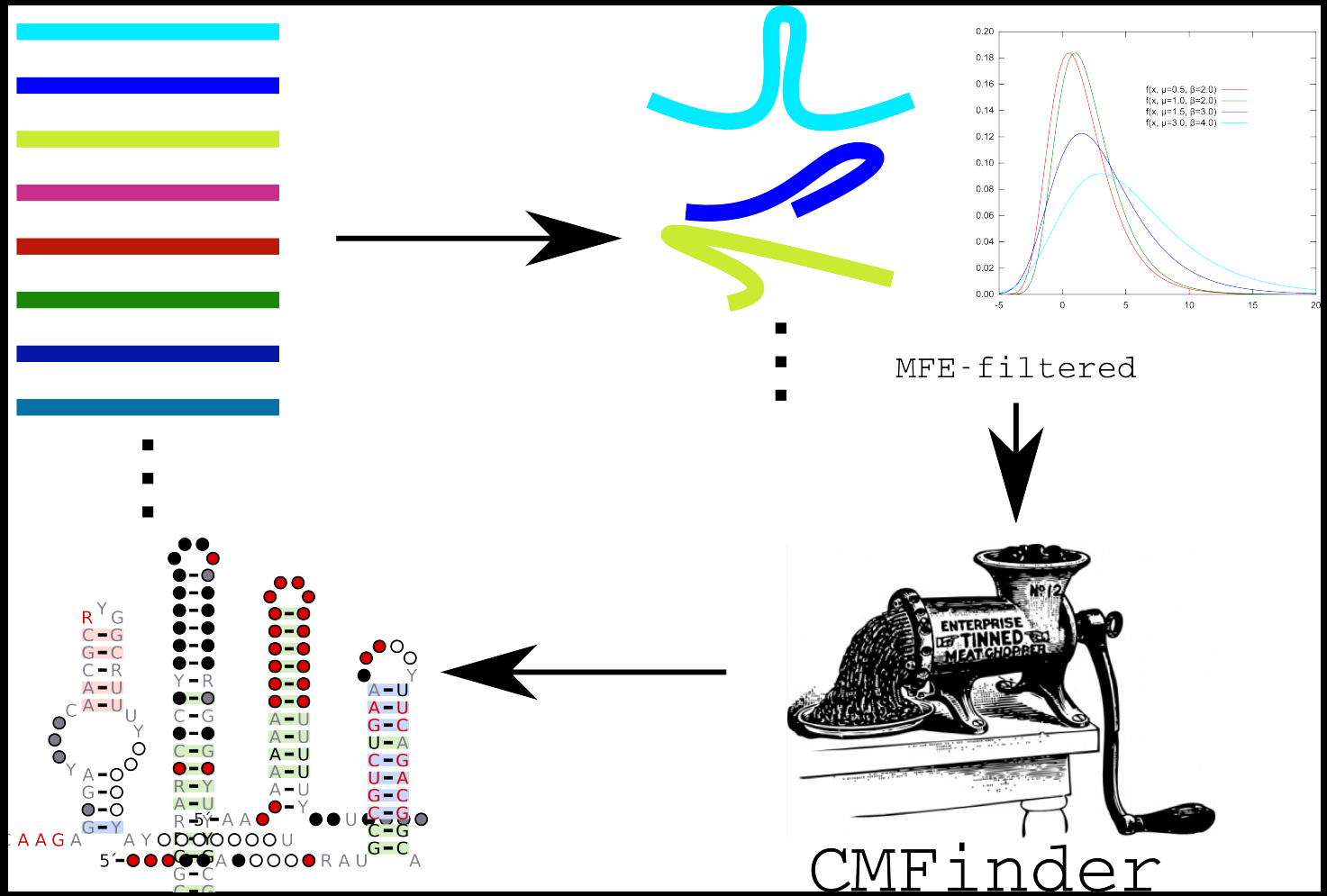
How many are there?

Do these families have a shared origin?

Mining EMBL

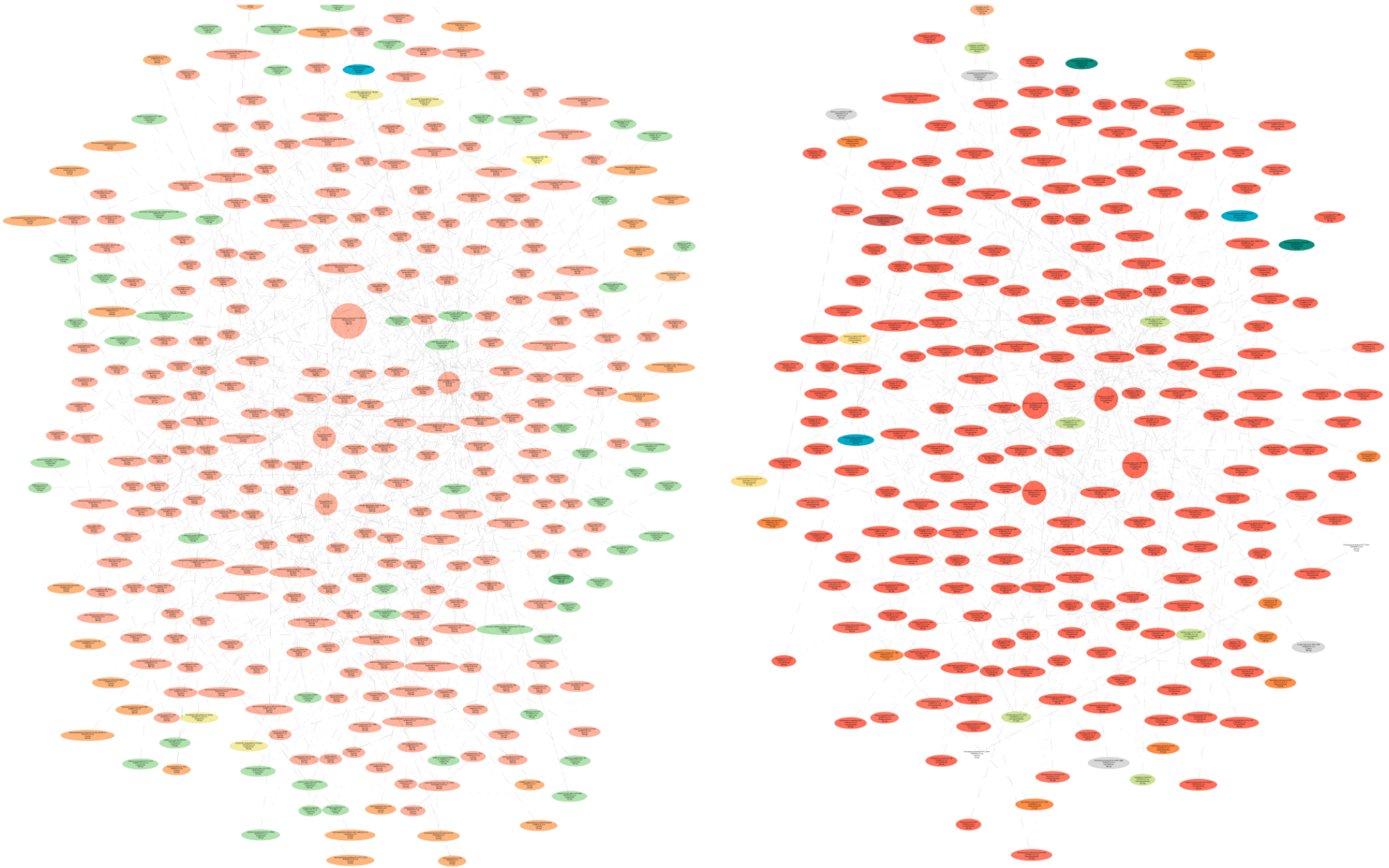


Mining EMBL



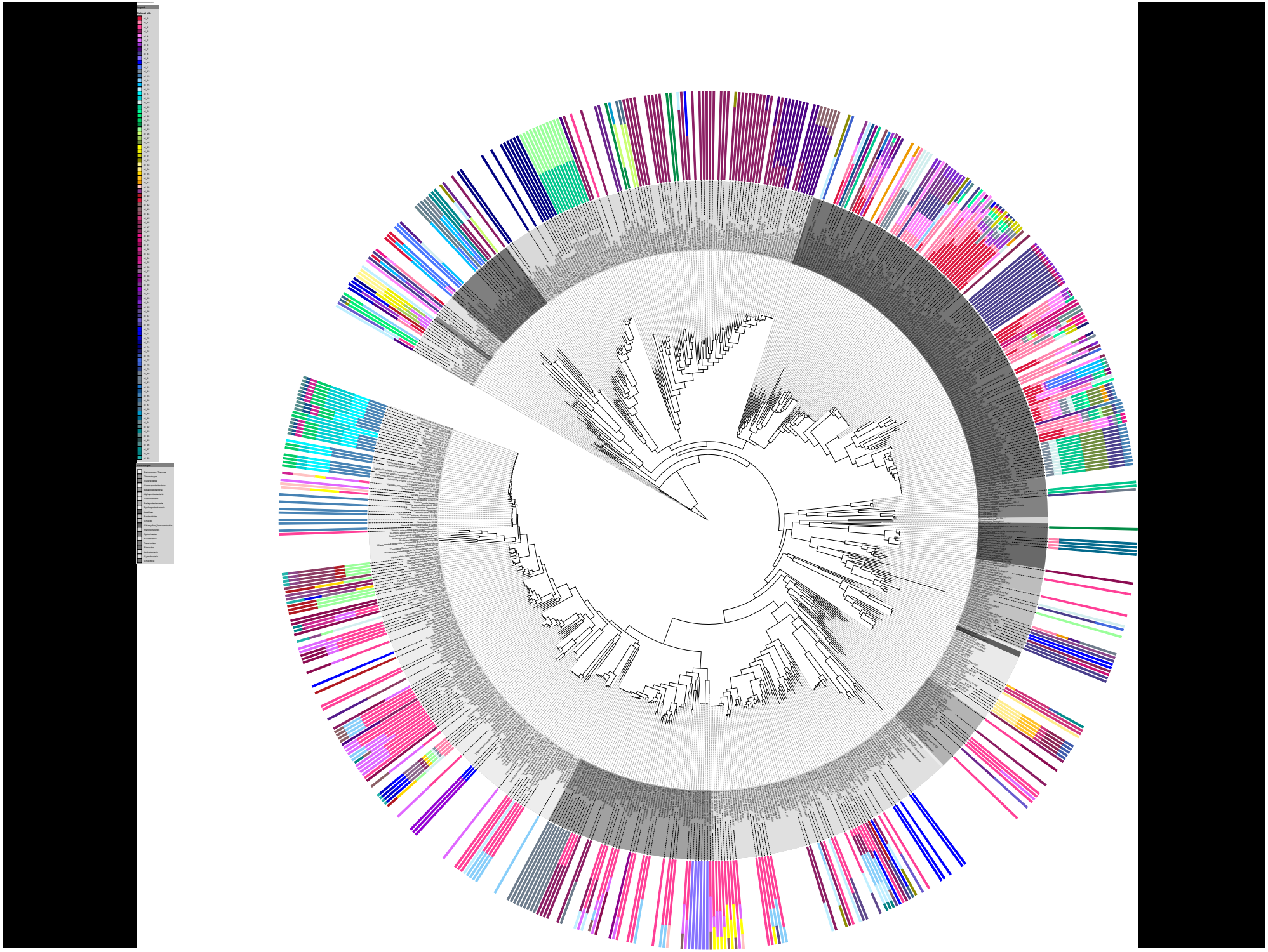
Clustering CMs

- Emit 1000 sequences from one model, score with the other & vice versa
- Average -lg(E-values) & shift positive
- Feed in to MCL (see Enright et al. TRIBE-MCL)
- Results in ~100 (non-trivial) clusters

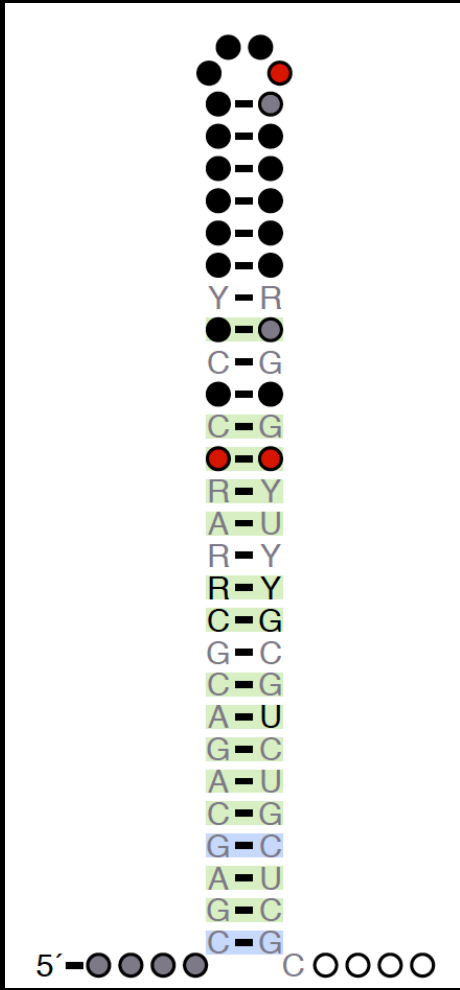


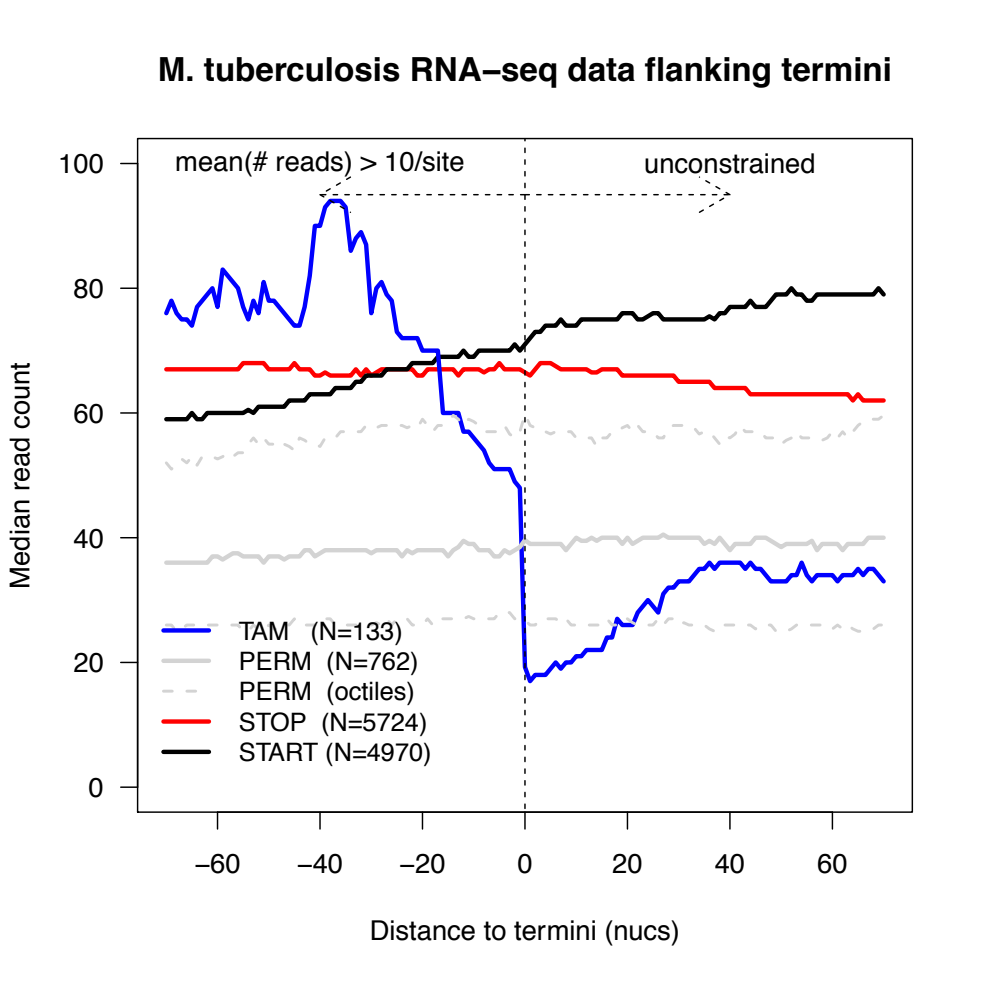
Cluster 0
73% Bacillales

Cluster 3
81% Actinobacteria



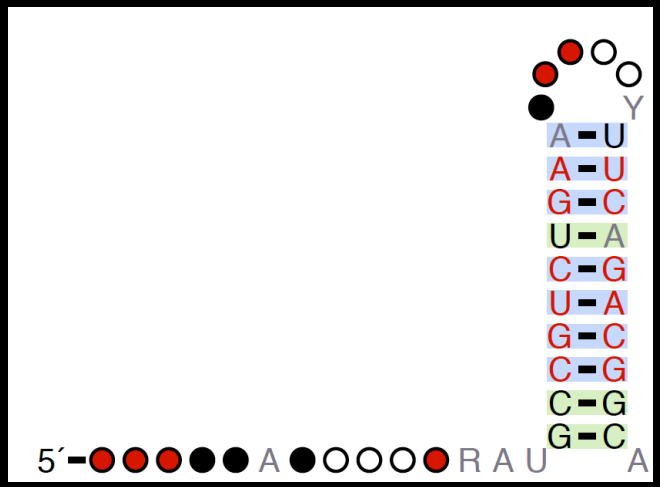
Actinobacterial Terminator



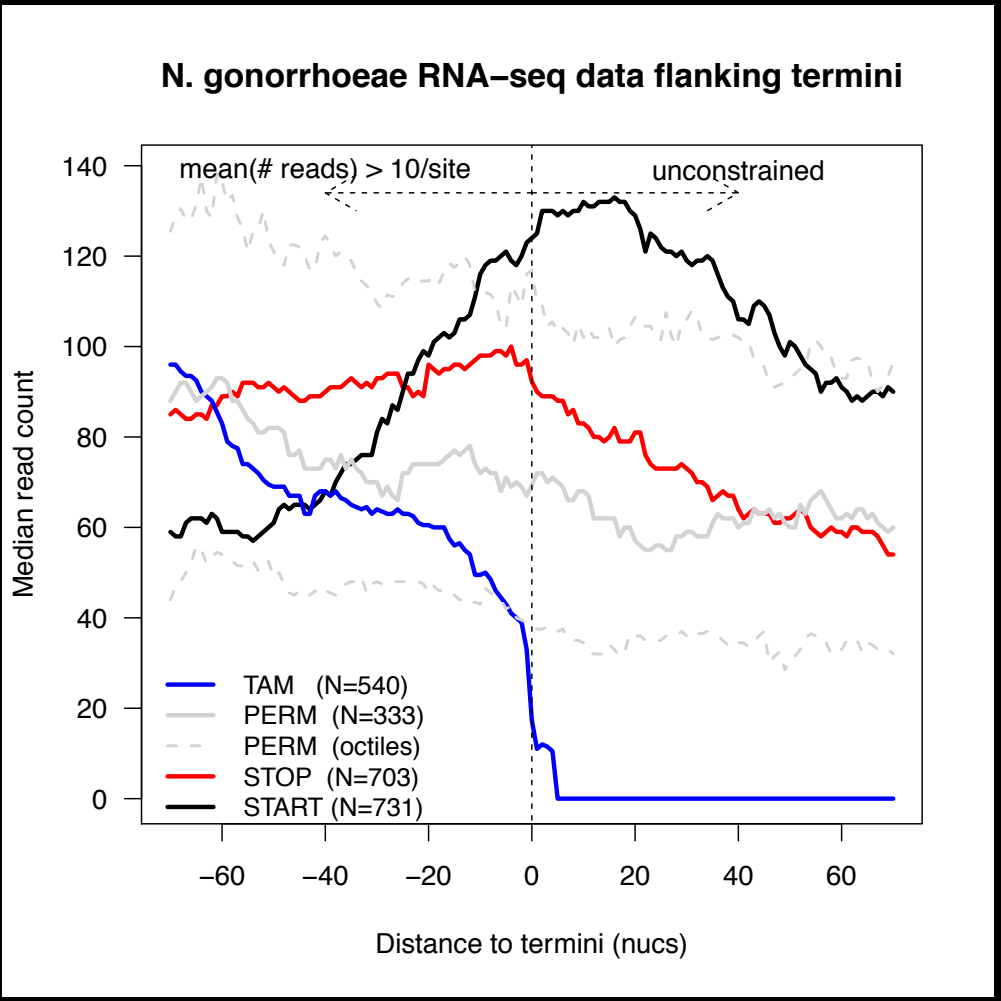


Data courtesy of KB Arnvig, NIMH

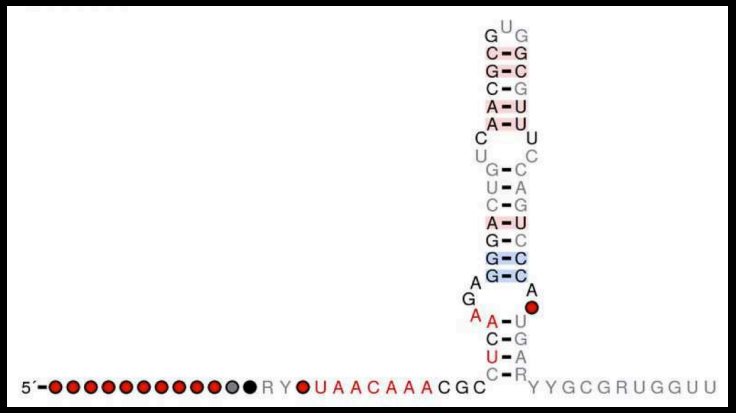
Neisseria DNA Uptake Terminator



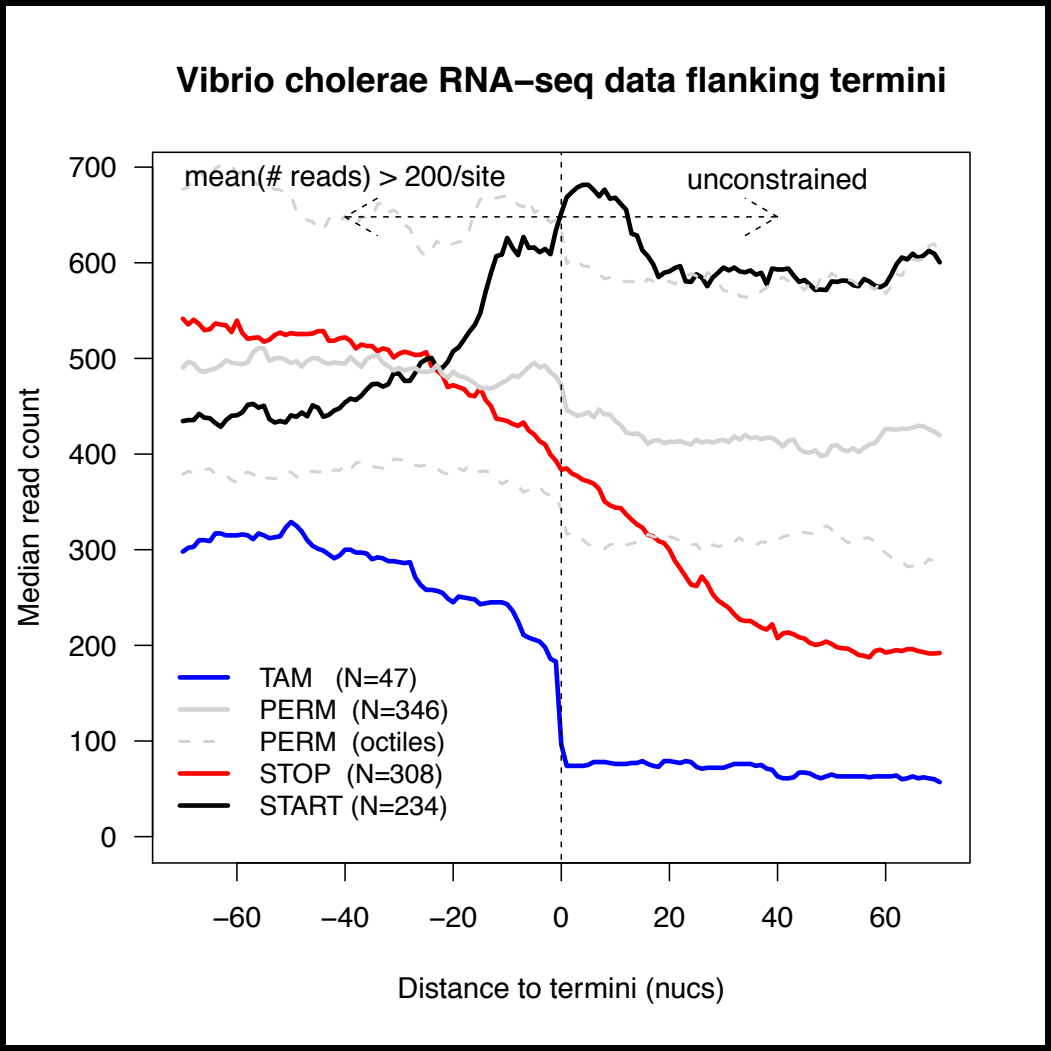
~600/genome, high percent id
Patchy poly-U tail
Exapted DNA Uptake sequence?



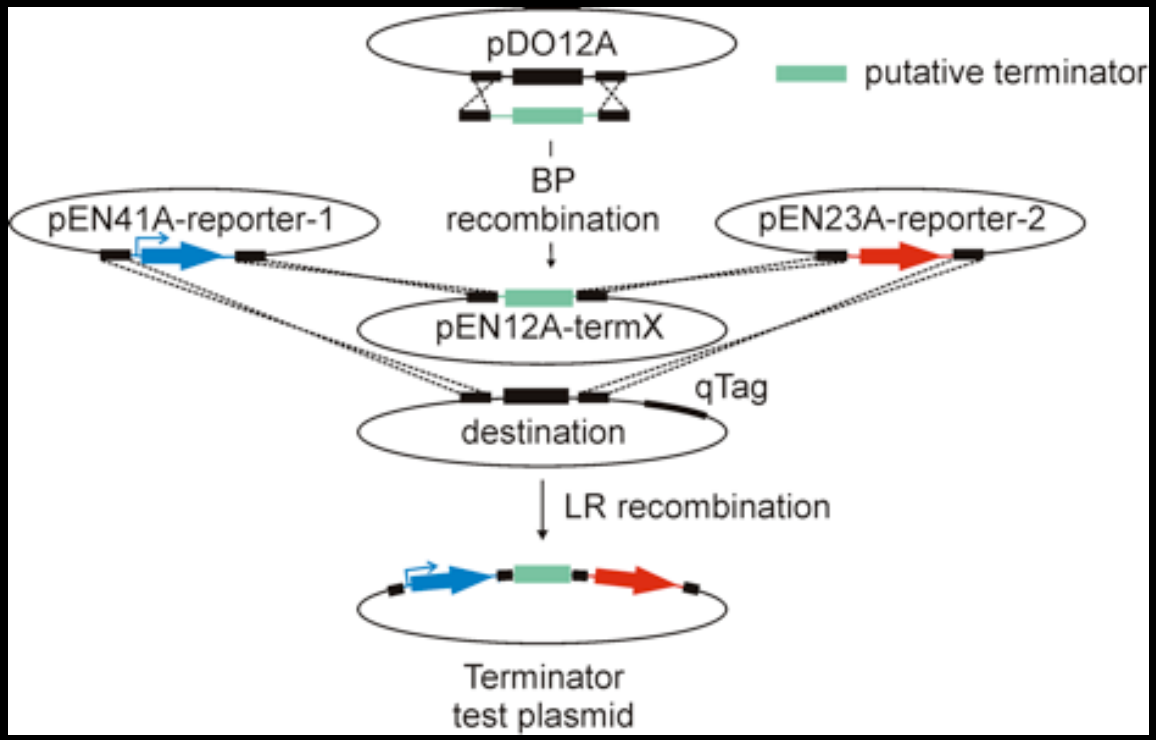
Vibrio Superintegron Terminus-Associated Motif

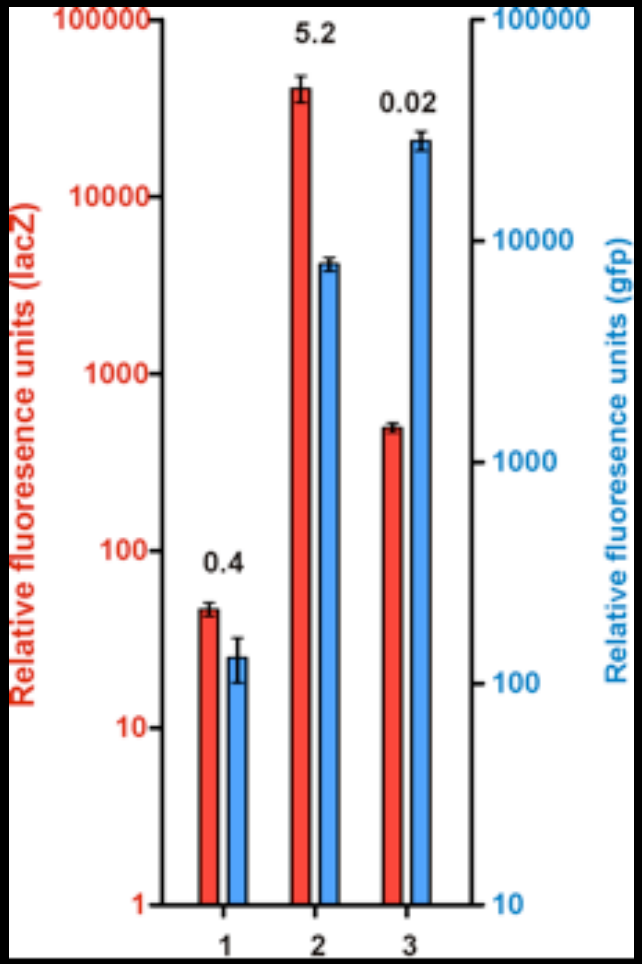


attC SI recombination site



Experimental Validation





Courtesy of D Schnappinger,
Weill Cornell Medical College

Conclusions so far...

- Rho-independent termination has arisen independently multiple times
- G/C-rich hairpin + poly-U tail appears to be the dominant (but not the only!) model for RI termination
- Connections to horizontal transfer?

Acknowledgements

Paul Gardner — University of Canterbury
James Hadfield — New Zealand

Tine Arnvig — MRC National Institute for Medical Research, UK

Dirk Schnappinger — Weill Cornell Medical College, USA