

The nuclear genome of *Rhazya stricta* and the evolution of alkaloid diversity in a medically relevant clade of Apocynaceae

Jamal S.M. Sabir^{1†}, Robert K. Jansen^{1,2†}, Dhivya Arasappan², Virginie Calderon³, Emmanuel Noutahi³, Chunfang Zheng⁴, Seongjun Park², Meshaal J. Sabir¹, Mohammed N. Baeshen⁹, Nahid H Hajrah¹, Mohammad A. Khiyami⁵, Nabih A. Baeshen⁶, Abdullah Y. Obaid⁷, Abdulrahman L. Al-Malki⁸, David Sankoff⁴, Nadia El-Mabrouk³ and Tracey A. Ruhlman^{2*}

Supplementary Information_SREP-16-05184

Supplementary Table S1: Data generated for genome assembly

Supplementary Table S2: Improvement of assembly with PacBio and optical mapping.

Supplementary Table S3: OpGen whole genome map statistics.

Supplementary Table S4: Comparison of assembly statistics for four plant nuclear genomes.

Supplementary Table S5: Coding sequence annotation for *Rhazya* nuclear genome.

Supplementary Table S6: Repeated sequences of the *Rhazya* nuclear genome.

Supplementary Table S7: Query sequences used in homolog search.

Supplementary Table S8: Results from reconciliation with different bootstraps cut-off, all possible roots. Results obtained after manual filtering considering both likelihood and reconciliation cost.

Supplementary Table S9: Monoterpenoid indole alkaloids (MIA) isolated from *Rhazya stricta*. Extracts from young *R. stricta* leaves were evaluated for MIA content by mass spectrometry.

Supplementary Text S1. Extended methods for genome assembly and annotation.

Supplementary Figures and Legends S1-S3.

Supplementary Table S1: Data generated for genome assembly

Technology	Library type	Insert size (bp)	Read length (bp)	Number of reads
Illumina	Fragment	160-220	100	291,361,582
Illumina	Mate-pair	3,000	100	78,042,292
Illumina	Mate-pair	6,000	100	26,102,232
PacBio	Continuous long read	10,000	50-22,603	579,535

bp=base pairs

Supplementary Table S2: Improvement of assembly with PacBio and optical mapping.

Refinement steps	Nnumber of scaffolds	N50	Max length	Scaffolds >200KB	
				count	% of assembly (274Mb)
Assembled Illumina reads	1449	1,061,426	4,551,498	n.a.	n.a.
Scaffolding with PacBio (AHA), Close gaps with PacBio (PBjelly)	1099	1,445,001	6,923,550	232	93
Merge large scaffolds with optical mapping	980	5,554,218	16,403,306	113	93

Mb= megabases

Supplementary Table S3: OpGen whole genome map statistics. Data provided by OpGen.

Summary of SMRM data	Maps used in analysis
Total size (Mb)	104,229.97
Number of molecules	291,892
Average size of molecules (kb)	357.08
Minimum molecule size (kb)	250
Average size of fragments (kb)	14.94

Mb= megabases, kb= kilobases

Supplementary Table S4: Comparison of assembly statistics for four plant nuclear genomes.

Taxon	Genome size estimate (Mb)	No. of scaffolds	Assembly size	Scaffold N50 (Mb)
Coffea	710	13,345	568.6	1.26
Cacao	430	4,792	326.9	0.47
Amborella	870	5,745	706.0	4.9
Rhazya	200	980	274.0	5.5

Mb=megabases

Supplementary Table S5: Coding sequence annotation for *Rhazya* nuclear genome.

Number of genes in filtered list	17,041
Additional SNAP genes added to the filtered list	4,123
Total	21,164
Number of exons	107,181
Number of CDS	104,172
Average gene size (bp)	4,452
Average number of exons per gene	6.29

bp=base pairs

Supplementary Table S6: Repeated sequence estimate for the *Rhazya* nuclear genome.

Class	Order	Superfamily	Number of elements	Sequence Occupied (bp)	Proportion of genome
		Copia	19,698	15,260,318	5.6%
		Gypsy	29,986	18,810,088	6.9%
Retro transposons	LTR	DIRS	120	7,332	>0.01%
		ERV	864	56,368	>0.01%
		Others	8,690	5,072,218	1.8%
	LINE		4,060	682,122	0.2%
	SINE		230	17,004	>0.01%
DNA transposons	TIR	Tc1-Mariner	1,126	200,788	0.1%
		hAT	2,492	877,146	0.3%
		Harbinger	676	172,602	0.1%
		Others	7,314	1524,456	0.6%
Low complexity sequence			227,112	10,959,376	4.0%
Total				55498610	20.2%*

*This list of repeated elements is not exhaustive and the actual value for repeat content is likely higher. bp=base pairs, LTR=long terminal repeats, LINE= Long interspersed elements, SINE=Short interspersed elements, TIR= Terminal Inverted Repeats

Supplementary Table S7. Query sequences used in homolog search.

SYMBOL	FULL NAME	*IntEnz	SOURCE	ACCESSION	FIG#
AAE	Acetylajmaline esterase	EC 3.1.1.80	RSA	Q3MKY2	ED_2c
SGD	Strictosidine beta-glucosidase	EC 3.2.1.105	CRA RSA	AAF28800 Q8GU20	ED_2n
PNAE	Polyneuridine aldehyde esterase	EC 3.1.1.78	RSA	Q9SE93	ED_2j
STR	Strictosidine synthase	EC 4.3.3.2	CRA RSA	P18417 P68175	ED_2p
T16H	Tabersonine 16-hydroxylase_CYP71D12	EC 1.14.13.73	CRA	P98183	ED_2q
THAS	tetrahydroalstonine synthase	EC 1.1.1	CRA	AKF02528	ED_2t
VS	Vinorine synthase 70%VS homolog_cathcyc	EC 2.3.1.160	RSA CRA	Q70PR7 GK13-15088	ED_2u
D4H	deacetoxyvindoline 4-hydroxylase	EC 1.14.11.20	CRA	O04847	ED_2e
7DLS	7-deoxyloganic acid hydroxylase_CYP72A224	EC 1.14.13	CRA	AHK60834	ED_2a
160MT	Tabersonine 16-hydroxylase_CYP71D12	EC 1.14.13.73	CRA	P98183	ED_2b
DAT	Deacetylvindoline O-acetyltransferase	EC 2.3.1.107	CRA	Q9ZTK5	ED_2d
G10H	Geraniol 8-hydroxylase_CYP76B6 [†]	EC 1.14.13.152	CRA	Q8VWZ7	ED_2f
LAMT	Loganate O-methyltransferase	EC 2.1.1.50	CRA	AGX93063	ED_2g
MAT	Minovincinine 19-hydroxy-O-acetyltransferase	EC 2.3.1	CRA	AAO13736	ED_2h
NMT	16-hydroxy-2,3-dihydro-3-hydroxytabersonine n-methyltransferase	EC 2.1.1.99	CRA	AHH33092	ED_2i
PR	Perakine reductase	EC 1.1.1.317	RSA	AAX11684	ED_2k
PRX1	Peroxidase 1	EC 1.11.1	CRA	CAJ84723	ED_2l
RO	Reticuline oxidase	EC 1.21.3.3	CRA	AKF02526	ED_2v
TDC	Tryptophan decarboxylase	EC 4.1.1.27	CRA	P17770	ED_2s
SLS	Secologanin synthase_CYP72A1	EC 1.3.3.9	CRA	Q05047	ED_2o
T19H	Tabersonine/lochnericine 19-hydroxylase_CYP71BJ1	EC 1.14.13	CRA	ADZ48681	ED_2r

*<http://www.ebi.ac.uk/intenz/query?cmd=SearchEC&ec=3.2.1.105>.

[†]Gene CYP76B6 encodes a protein with alternate names (Geraniol 8-hydroxylase and 10-hydroxylase).

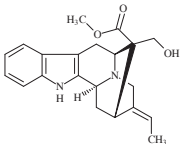
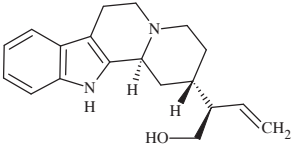
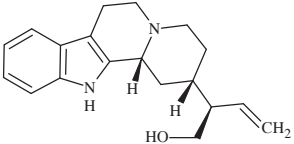
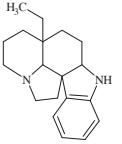
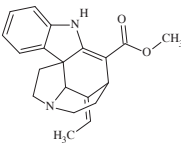
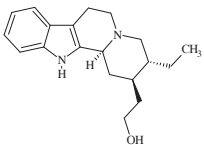
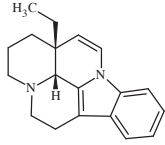
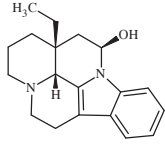
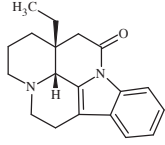
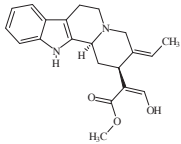
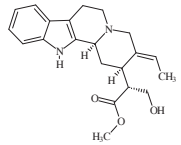
CRA=*Catharanthus roseus*, RSA=*Rauvolfia serpentina*

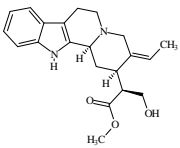
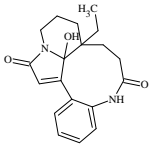
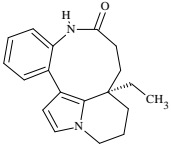
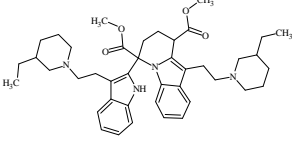
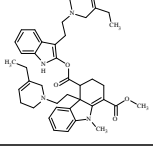
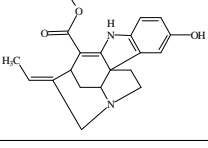
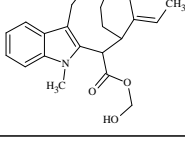
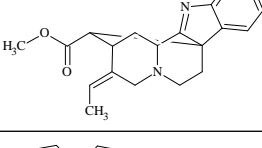
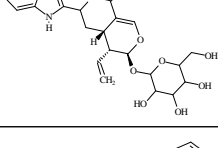
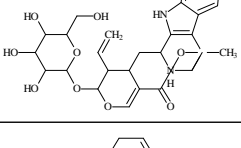
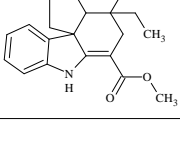
Supplementary Table S8: Results from reconciliation with different bootstrap cut-offs and all possible roots. Results obtained after manual filtering considering both likelihood and reconciliation cost.

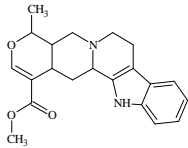
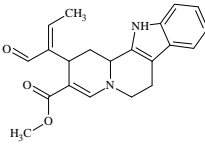
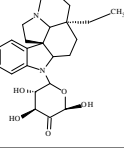
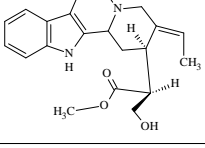
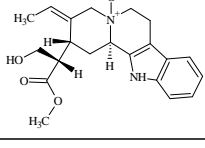
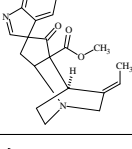
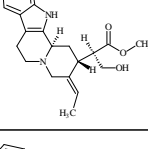
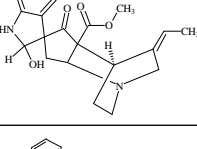
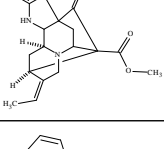
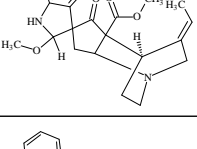
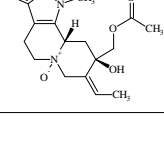
Query	All possible roots		Number of solutions	Lowest reconciliation cost	Reference figure
	Dup	loss			
AAE	26	7-11	11	33	ED_3a
D4E	20-21	14-17	10	34	ED_3b
PNAE	16-18	3-11	12	23	ED_3c
SGD	27	25-29	12	52	ED_3d
STR	28-29	20-25	13	48	ED_3e
T16H	26	14-21	9	40	ED_3f
THAS	11	13-20	13	24	FIG 3
VS	11	6-13	11	19	ED_3g

Dup=duplications, query abbreviations are expanded in Supplementary Table S7.

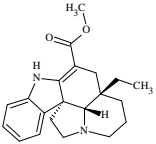
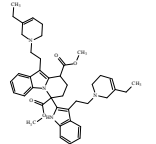
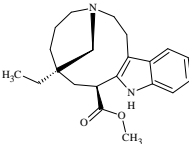
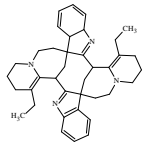
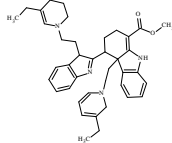
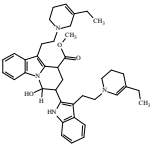
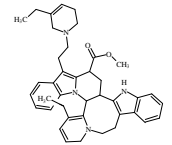
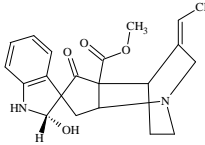
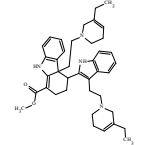
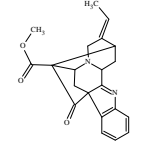
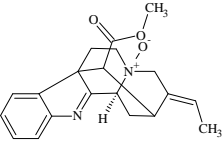
Supplementary Table S9: Monoterpenoid indole alkaloids isolated from *Rhazya stricta* leaves

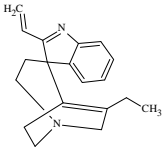
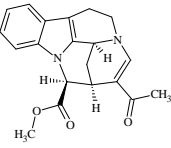
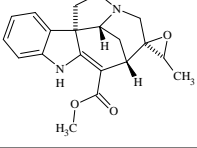
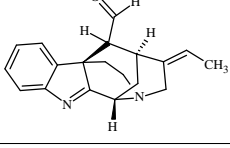
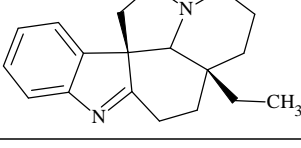
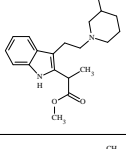
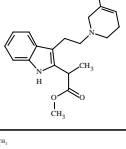
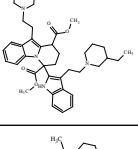
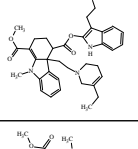
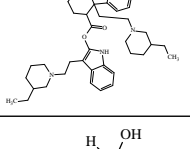
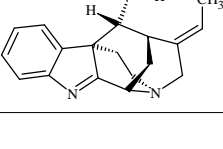
#ID /	Structure	Name	FW
1		Akuammidine	352.4269
2		Antirhine	296.4067
3		3-epi-Antirhine	296.4067
4		Aspidosespermidine	282.4231
5		Condyllocarpine	322.4009
6		dihydrocorynantheol	298.4225
7		eburnamenine	278.3914
8		eburnamine	296.4067
9		Eburnamonine	294.3908
10		geissoschizine	352.4269
11		Isositsirikine	354.4427

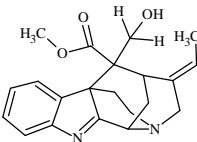
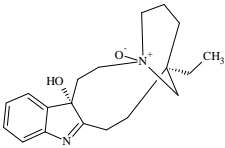
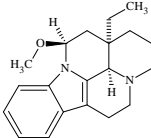
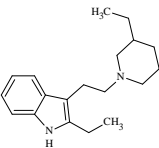
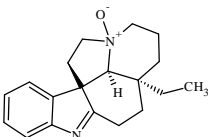
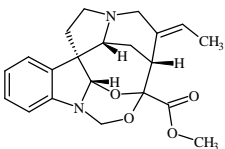
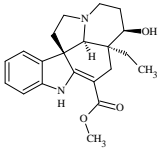
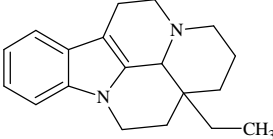
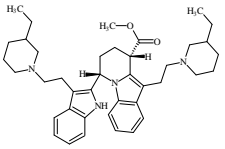
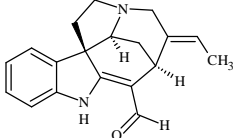
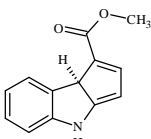
#ID /	Structure	Name	FW
12		16-epi-Z-isositsirikine	354.4427
13		leuconolam	326.3896
14		rhazinilam	294.3908
15		tetrahydrosecamine	680.9184
16		presecamine	676.8867
17		sewarine	338.4003
18		Stemmadenine	354.4427
19		strictamine	322.4009
20		strictosamide	498.5250
21		strictosidine	530.5669
22		tabersonine	336.4275

#ID /	Structure	Name	FW
23		Tetrahydroalstonine	352.4269
24		vallesiachotamine	350.4110
25		Aspidospermiase	428.5213
26		Bhimberine	354.4427
27		Bhimbrine N-Oxide	370.4421
28		Rhazimine	350.4110
29		Rhazimanine	354.4427
30		Rhazicine	368.4263
31		Leepacine	350.4110
32		2-Methoxy-1,2-Dihydrorhazimine	382.4528
33		HR-1	370.4421

#ID /	Structure	Name	FW
34		Vincanicine	322.4009
35		Rhazinaline	350.4110
36		beta-Sitosterol	414.7067
37		Ursolic acid	456.7003
38		Stigmasterol	412.6908
39		Oleanolic acid	456.7003
40		Rhazidigenine (Rhazidine, Strictanol)	298.4225
41		N-methylleuconolam	340.4162
42		(+)-Quebrachamine	282.4231
43		Polyneuridine	350.4110
44		(+)-Vincadiforimine	338.4433

#ID /	Structure	Name	FW
45		(-)-Vincadifformine	338.4433
46		Secamine	676.8867
47		Vincadine	340.4592
48		bis-strictidine	558.7986
49		3,14-dehydrorhazigine	616.8347
50		16-hydrorhazisidine	634.8500
51		rhazisidine	614.8188
52		Isorhazicine	368.4263
53		rhazigine	618.8506
54		strictisidine	348.3951
55		strictamine-N-oxide	338.4003

#ID /	Structure	Name	FW
56		strictigine	278.3914
57		strictine	336.3844
58		stricticine	338.4003
59		strictalamine	292.3749
60		1,2-Dehydroaspidospermidine (eburenine)	280.4073
61		Tetrahydrosecodine	342.4751
62		Dihydrosecodine	340.4592
63		Dihydrosecamine	678.9026
64		Dihydropresecamine	678.9026
65		tetrahydropresecamine	680.9184
66		rhazinol	294.3908

#ID /	Structure	Name	FW
67		Rhazimol	352.4269
68		Rhazidigenine-N-oxide	314.4219
69		(-)-16R,21R-omethyleburnamine	310.4332
70		Decarbomethoxy-15,20,16,17-tetrahydrosecodeine	284.4390
71		1-2-dehydroasidospermidine-N-oxide	296.4067
72		rhazizine	368.4263
74		15-hydroxyvincadifformine	354.4427
76		Dihydroeburnamenine	280.4073
77		16s,16'-decarboxytetra-hydrosecamine	622.8824
78		Nor-C-fluorocurarine	292.3749
79		strictibine	213.2319

Supplementary Text S1

Iteratively varied parameters for AHA assembly

For running the iterative hybrid assembly, these parameters were varied iteratively. Comma separates out the values per iteration.

- Minimum alignment score (aka Z-score) 6,6,6,6
- Minimum number of reads needed to link two contigs. 3,3,2,2
- Minimum subread length to participate in alignment. 75,75,75,75

Protocol file for PBJelly

<jellyProtocol>

<reference>/scratch/01184/daras/jansen/RHA/pbjelly/RHA_small_75per_overlapping/data/reference/final.assembly.fasta</reference>

<outputDir>/scratch/01184/daras/jansen/RHA/pbjelly/RHA_small_75per_overlapping/</outputDir>

<blasr>-minMatch 8 -minPctIdentity 70 -bestn 8 -nCandidates 30 -maxScore -500 -nproc 8 -noSplitSubreads</blasr>

<input
baseDir="/scratch/01184/daras/jansen/RHA/pbjelly/RHA_small_75per_overlapping/data/reads/">

<job>Jansen_Pacbio.cat.fasta</job>

</input>

</jellyProtocol>

Run parameters for MAKER2

Maker_bopts.ctl

#-----BLAST and Exonerate Statistics Thresholds
blast_type=ncbi+ #set to 'ncbi+', 'ncbi' or 'wublast'

pcov_blastn=0.8 #Blastn Percent Coverage Threshold EST-Genome Alignments

pid_blastn=0.85 #Blastn Percent Identity Threshold EST-Genome Alignments

eval_blastn=1e-10 #Blastn eval cutoff

bit_blastn=40 #Blastn bit cutoff

depth_blastn=0 #Blastn depth cutoff (0 to disable cutoff)

pcov_blastx=0.5 #Blastx Percent Coverage Threshold Protein-Genome Alignments

pid_blastx=0.4 #Blastx Percent Identity Threshold Protein-Genome Alignments

eval_blastx=1e-06 #Blastx eval cutoff
bit_blastx=30 #Blastx bit cutoff
depth_blastx=0 #Blastx depth cutoff (0 to disable cutoff)

pcov_tblastx=0.8 #tBlastx Percent Coverage Threshold alt-EST-Genome Alignments
pid_tblastx=0.85 #tBlastx Percent Identity Threshold alt-EST-Genome Alignments
eval_tblastx=1e-10 #tBlastx eval cutoff
bit_tblastx=40 #tBlastx bit cutoff
depth_tblastx=0 #tBlastx depth cutoff (0 to disable cutoff)

pcov_rm_blastx=0.5 #Blastx Percent Coverage Threshold For Transposable Element Masking
pid_rm_blastx=0.4 #Blastx Percent Identity Threshold For Transposable Element Masking
eval_rm_blastx=1e-06 #Blastx eval cutoff for transposable element masking
bit_rm_blastx=30 #Blastx bit cutoff for transposable element masking

ep_score_limit=20 #Exonerate protein percent of maximal score threshold
en_score_limit=20 #Exonerate nucleotide percent of maximal score threshold

Maker_exe.ctl

#-----Location of Executables Used by MAKER/EVALUATOR
makeblastdb=/scratch/projects/tacc/bio/maker/2.28b/blast/bin/makeblastdb
#location of NCBI+ makeblastdb executable
blastn=/scratch/projects/tacc/bio/maker/2.28b/blast/bin/blastn #location of
NCBI+ blastn executable
blastx=/scratch/projects/tacc/bio/maker/2.28b/blast/bin/blastx #location of
NCBI+ blastx executable
tblastx=/scratch/projects/tacc/bio/maker/2.28b/blast/bin/tblastx #location of
NCBI+ tblastx executable
formatdb= #location of NCBI formatdb executable
blastall= #location of NCBI blastall executable
xdformat= #location of WUBLAST xdformat executable
blasta= #location of WUBLAST blasta executable
RepeatMasker=/scratch/projects/tacc/bio/maker/2.28b/RepeatMasker/RepeatMa
sker #location of RepeatMasker executable
exonerate=/scratch/projects/tacc/bio/maker/2.28b/exonerate/bin/exonerate
#location of exonerate executable

#-----Ab-initio Gene Prediction Algorithms
snap=/scratch/projects/tacc/bio/maker/2.28b/snap/snap #location of snap
executable
gmhmm3= #location of eukaryotic genemark executable
gmhmp= #location of prokaryotic genemark executable

augustus=/scratch/projects/tacc/bio/maker/2.28b/augustus/bin/augustus
#location of augustus executable
fgenesh= #location of fgenesh executable

#-----Other Algorithms

probuild= #location of probuild executable (required for genemark)

Maker_opts.ctl

#-----Genome (these are always required)

genome=opgenResult+scaffoldsLengthsLess200.fasta #genome sequence (fasta file or fasta embeded in GFF3 file)

organism_type=eukaryotic #eukaryotic or prokaryotic. Default is eukaryotic

#-----Re-annotation Using MAKER Derived GFF3

maker_gff= #MAKER derived GFF3 file

est_pass=0 #use ESTs in maker_gff: 1 = yes, 0 = no

altest_pass=0 #use alternate organism ESTs in maker_gff: 1 = yes, 0 = no

protein_pass=0 #use protein alignments in maker_gff: 1 = yes, 0 = no

rm_pass=0 #use repeats in maker_gff: 1 = yes, 0 = no

model_pass=0 #use gene models in maker_gff: 1 = yes, 0 = no

pred_pass=0 #use ab-initio predictions in maker_gff: 1 = yes, 0 = no

other_pass=0 #passthrough anything else in maker_gff: 1 = yes, 0 = no

#-----EST Evidence (for best results provide a file for at least one)

est=trinityOut.Trinity.fasta #set of ESTs or assembled mRNA-seq in fasta format

altest= #EST/cDNA sequence file in fasta format from an alternate organism

est_gff= #aligned ESTs or mRNA-seq from an external GFF3 file

altest_gff= #aligned ESTs from a closely related species in GFF3 format

#-----Protein Homology Evidence (for best results provide a file for at least one)

protein=uniprot_sprot.fasta #protein sequence file in fasta format (i.e. from multiple organisms)

protein_gff= #aligned protein homology evidence from an external GFF3 file

#-----Repeat Masking (leave values blank to skip repeat masking)

model_org=all #select a model organism for RepBase masking in RepeatMasker

rmlib= #provide an organism specific repeat library in fasta format for RepeatMasker

repeat_protein=/opt/apps/maker/2.30/data/te_proteins.fasta #provide a fasta file of transposable element proteins for RepeatRunner

rm_gff= #pre-identified repeat elements from an external GFF3 file

prok_rm=0 #forces MAKER to repeatmask prokaryotes (no reason to change this), 1 = yes, 0 = no

softmask=1 #use soft-masking rather than hard-masking in BLAST (i.e. seg and dust filtering)

#----Gene Prediction

snaphmm=RHA.hmm #SNAP HMM file

gmhmm= #GeneMark HMM file

augustus_species= #Augustus gene prediction species model

fgenes_par_file= #FGENESH parameter file

pred_gff= #ab-initio predictions from an external GFF3 file

model_gff= #annotated gene models from an external GFF3 file (annotation pass-through)

est2genome=0 #infer gene predictions directly from ESTs, 1 = yes, 0 = no

protein2genome=1 #infer predictions from protein homology, 1 = yes, 0 = no

unmask=0 #also run ab-initio prediction programs on unmasked sequence, 1 = yes, 0 = no

#----Other Annotation Feature Types (features MAKER doesn't recognize)

other_gff= #extra features to pass-through to final MAKER generated GFF3 file

#----External Application Behavior Options

alt_peptide=C #amino acid used to replace non-standard amino acids in BLAST databases

cpus=24 #max number of cpus to use in BLAST and RepeatMasker (not for MPI, leave 1 when using MPI)

#----MAKER Behavior Options

max_dna_len=100000 #length for dividing up contigs into chunks (increases/decreases memory usage)

min_contig=1 #skip genome contigs below this length (under 10kb are often useless)

pred_flank=200 #flank for extending evidence clusters sent to gene predictors

pred_stats=0 #report AED and QI statistics for all predictions as well as models

AED_threshold=1 #Maximum Annotation Edit Distance allowed (bound by 0 and 1)

min_protein=0 #require at least this many amino acids in predicted proteins

alt_splice=0 #Take extra steps to try and find alternative splicing, 1 = yes, 0 = no

always_complete=0 #extra steps to force start and stop codons, 1 = yes, 0 = no

map_forward=0 #map names and attributes forward from old GFF3 genes, 1 = yes, 0 = no

keep_preds=0 #Concordance threshold to add unsupported gene prediction (bound by 0 and 1)

split_hit=10000 #length for the splitting of hits (expected max intron size for evidence alignments)

single_exon=0 #consider single exon EST evidence when generating annotations, 1 = yes, 0 = no

single_length=250 #min length required for single exon ESTs if 'single_exon is enabled'
correct_est_fusion=0 #limits use of ESTs in annotation to avoid fusion genes

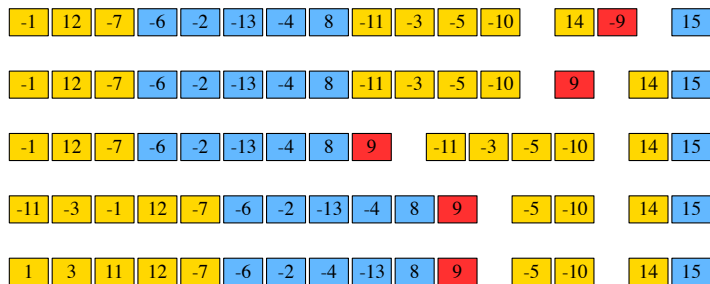
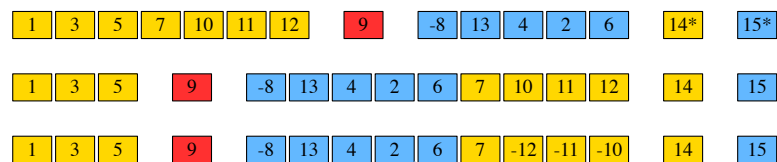
tries=2 #number of times to try a contig if there is a failure for some reason
clean_try=0 #remove all data from previous run before retrying, 1 = yes, 0 = no
clean_up=0 #removes theVoid directory with individual analysis files, 1 = yes, 0 = no
TMP= #specify a directory other than the system default temporary directory for temporary files

Supplementary Figure S1: Divergent structural evolution in *Rhazya stricta*. Genome level rearrangements leading from the ancestral core eudicot, represented by *Vitis* (upper right hand corner) to *Rhazya* and *Coffea*. Each linear arrangement represents successive steps from the eudicot ancestor toward the Gentianales ancestor. An arrow leads from the last of these stages to the relevant portion of the Gentianales ancestor. From this ancestor, two arrows indicate the divergence of *Coffea* and *Rhazya*, and successive rearrangements within each lineage are again indicated by linear arrangements of colored blocks. Inferred rearrangements between lineages are given at the arrows. Extant genomes carry the species name. The circular diagrams at the lower left and right represent the current orthologies between *Rhazya* and *Coffea* and *Vitis*, respectively. Syntenic blocks are colored according to the 21 ancestral core eudicot chromosomes³³ and do not represent biologically significant units. a) inferred orthologies for *Rhazya* superscaffolds 7, 9 and 10; b) superscaffolds 12, 13 and 28.

Supplementary Figure S2: Maximum likelihood phylogenies for 21 monoterpenoid indole alkaloid (MIA) sequences. Sequences of experimentally verified genes from GenBank for *Catharanthus* and *Rauvolfia* are indicated with a red arrow and accession numbers for these sequences are in Supplementary Table S7. Cra, *C. roseus*; Cc, *Coffea canephora*; rha, *Rhazya stricta*; rsa, *R. serpentina*; tom, *Solanum lycopersicum*; thec, *Theobroma cacao*. Full names for protein abbreviations in upper left corner of each tree are in Supplementary Table S7.

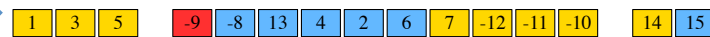
Supplementary Figure S3: Duplication and loss of selected monoterpenoid indole alkaloid (MIA) enzymes across six angiosperms. Gene phylogenies were reconciled

with the species phylogeny to infer duplication and loss history for each enzyme family. Internal and terminal branches in black reflect the amount of change along branch; dashed gray branch extensions are present for branches whose lengths were too short to enable uniform representation with branch support and length information. *cra*, *Catharanthus roseus*; *cof*, *Coffea canephora*; *rha*, *Rhazya stricta*; *rsa*, *Rauvolfia serpentina*; *tom*, *Solanum lycopersicum*; *the*, *Theobroma cacao*. Species indicated in blue font were downloaded from GenBank and those in red are from *Rhazya stricta*. Red square = gene duplication; green circle = speciation event; wide gray dashed line ending with a circle = gene loss. Full names for protein abbreviations in upper left corner of each tree are in Supplementary Table S7. Values above branches indicate bootstrap support for the clade, numbers below nodes give bootstrap support for the event (duplication or speciation) at the node. Bootstrap values less than 50 are not shown; inferred loss events are indicated with an asterisk and do not have associated bootstrap values. Scale bar represents number of amino acid substitutions per site. The maximum likelihood (ML) tree shown is the tree with the best ML score.

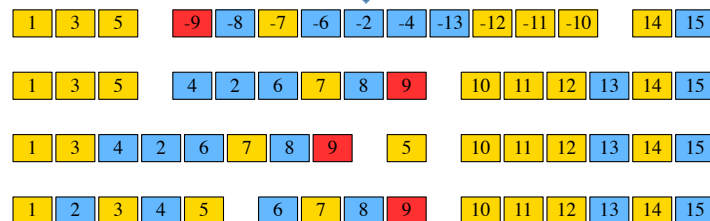
a.*Coffea canephora**Vitis vinifera (post γ eudicot)*

four translocations
one inversion
one transposition

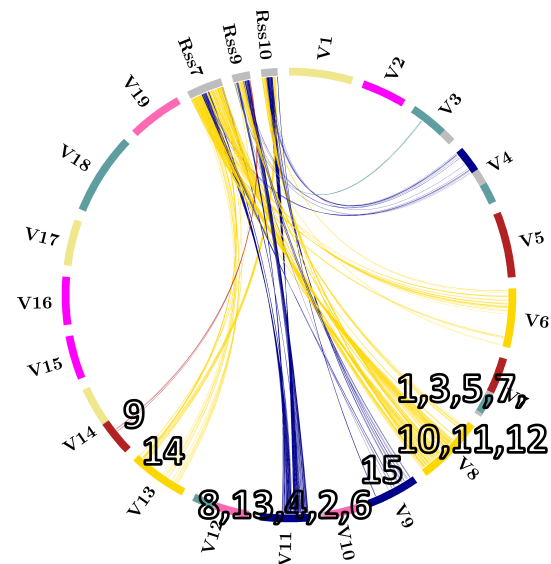
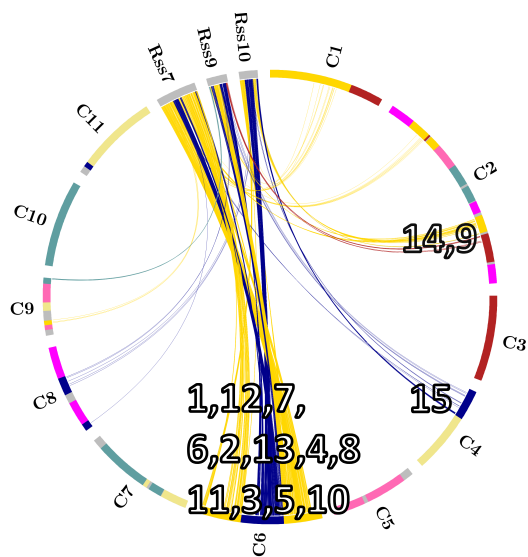
Gentianales ancestor

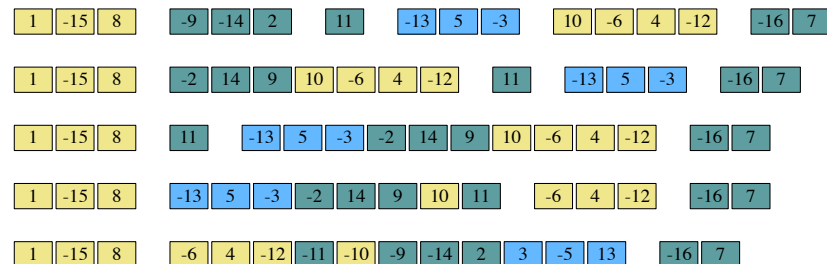
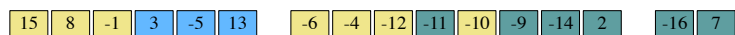


three translocations
one inversion
one transposition

*Rhazya stricta*

one translocation
one inversion
two fusions



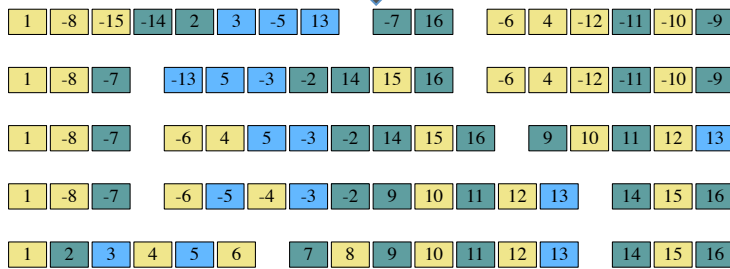
b.*Vitis vinifera* (post γ eudicot)*Coffea canephora*

one translocation
one inversion

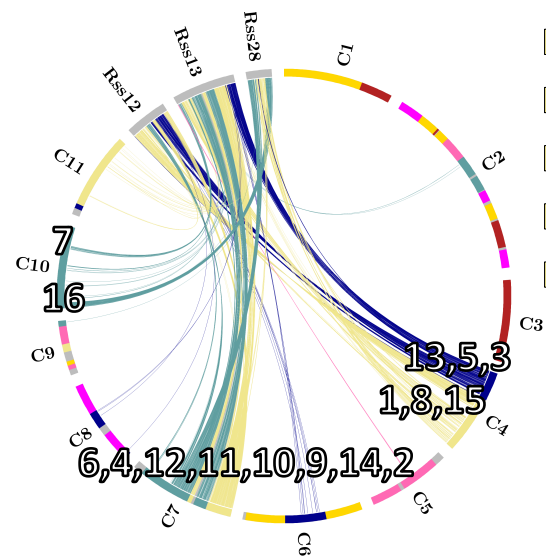
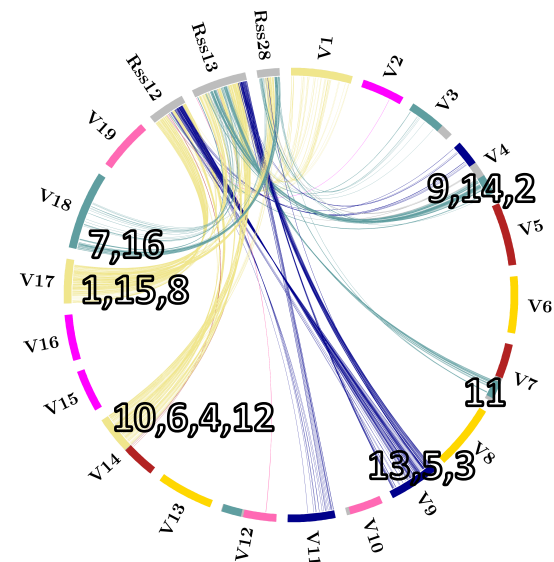
Gentianales ancestor



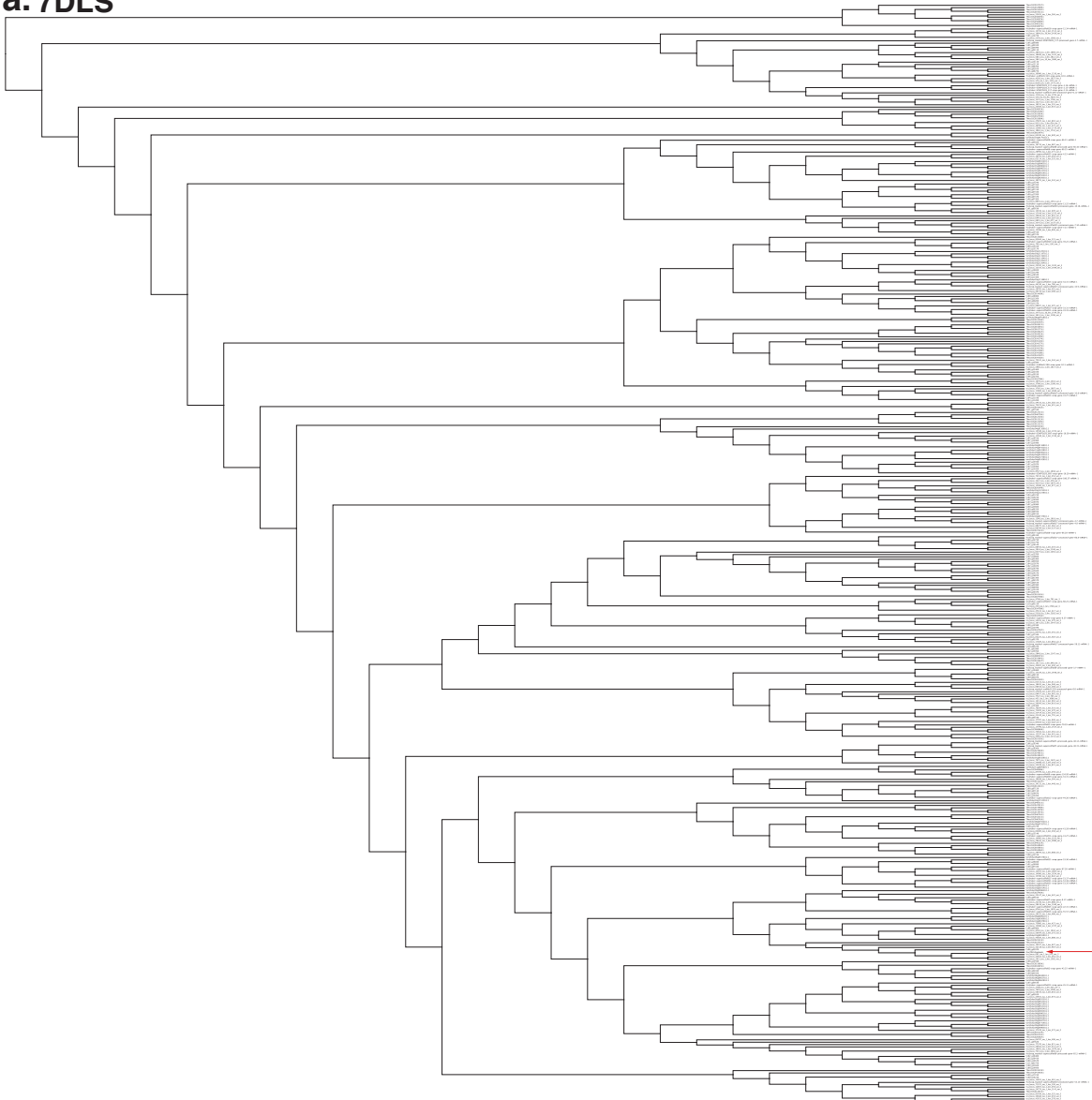
six translocations
one inversion

*Rhazya stricta*

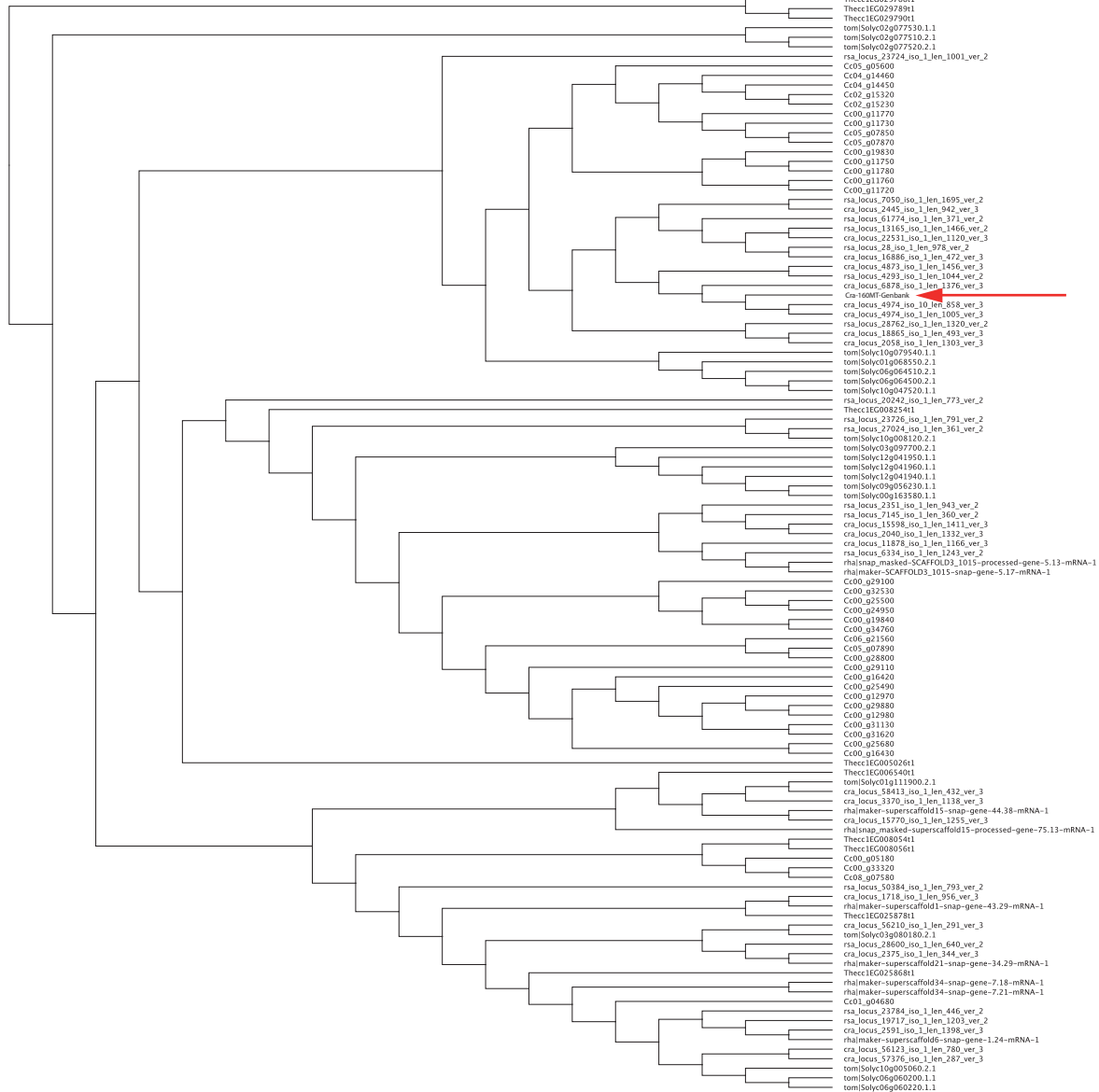
one translocation
one inversion
three fusions
one transposition



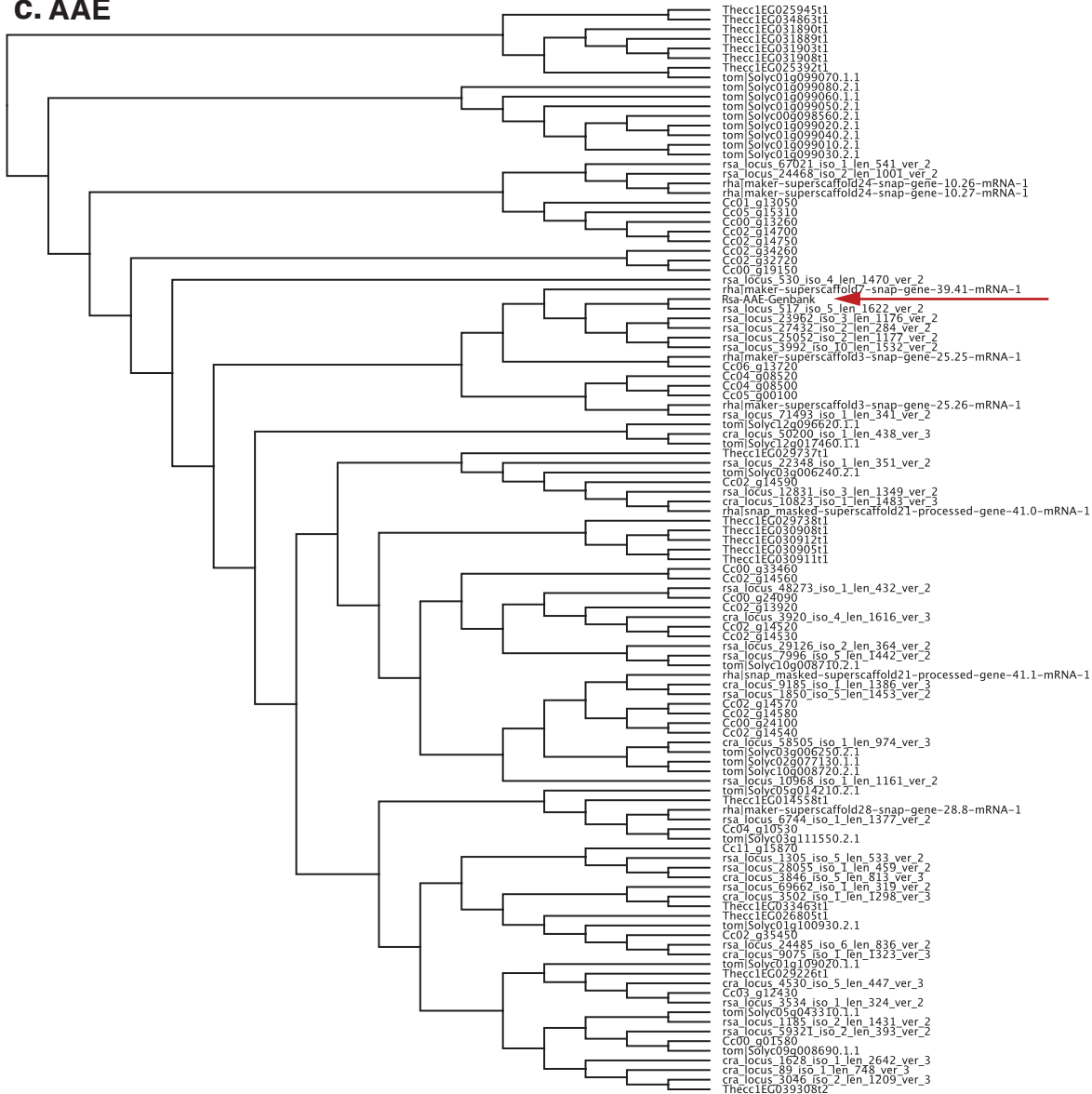
a. 7DLS



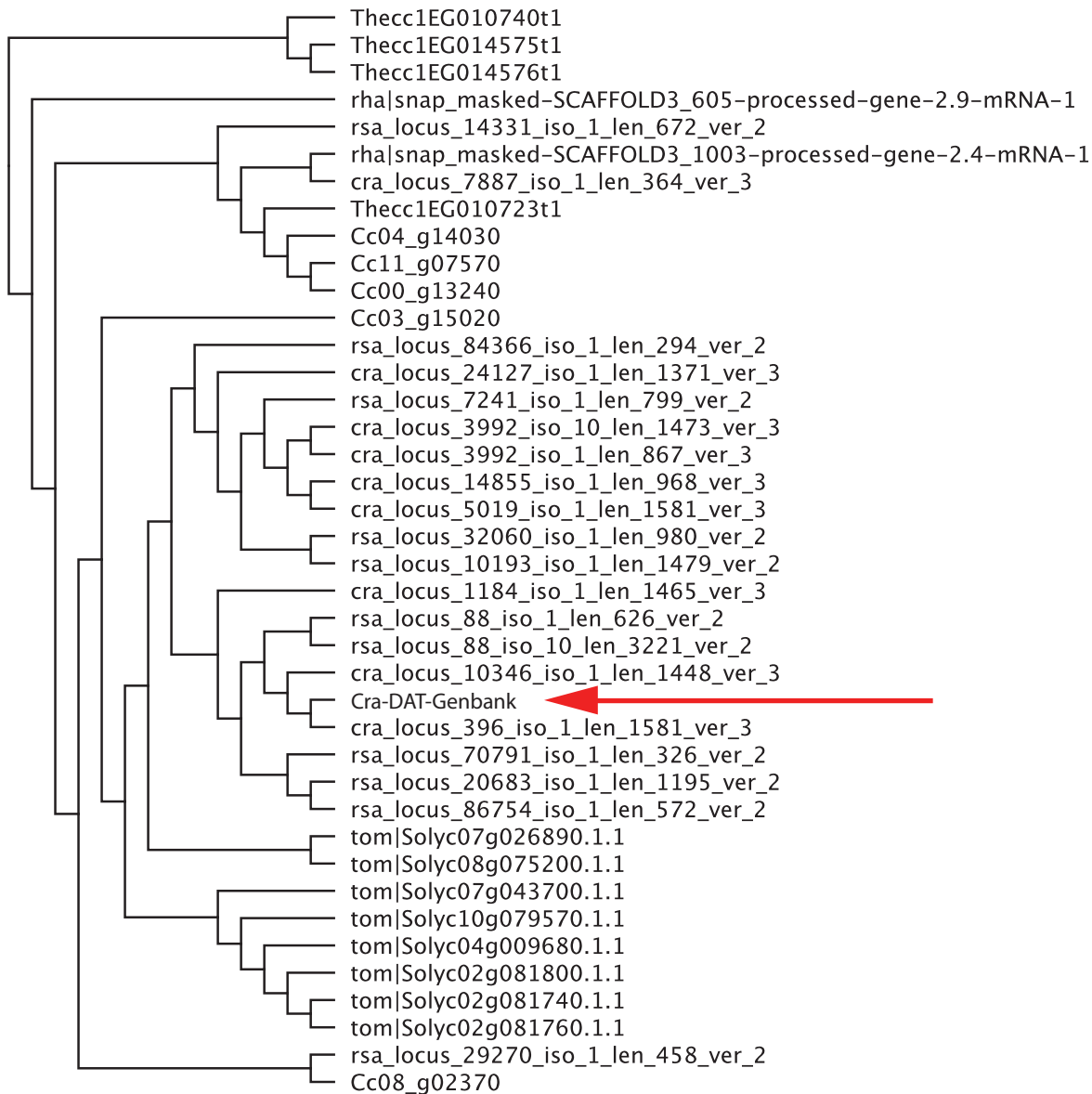
b. 160MT



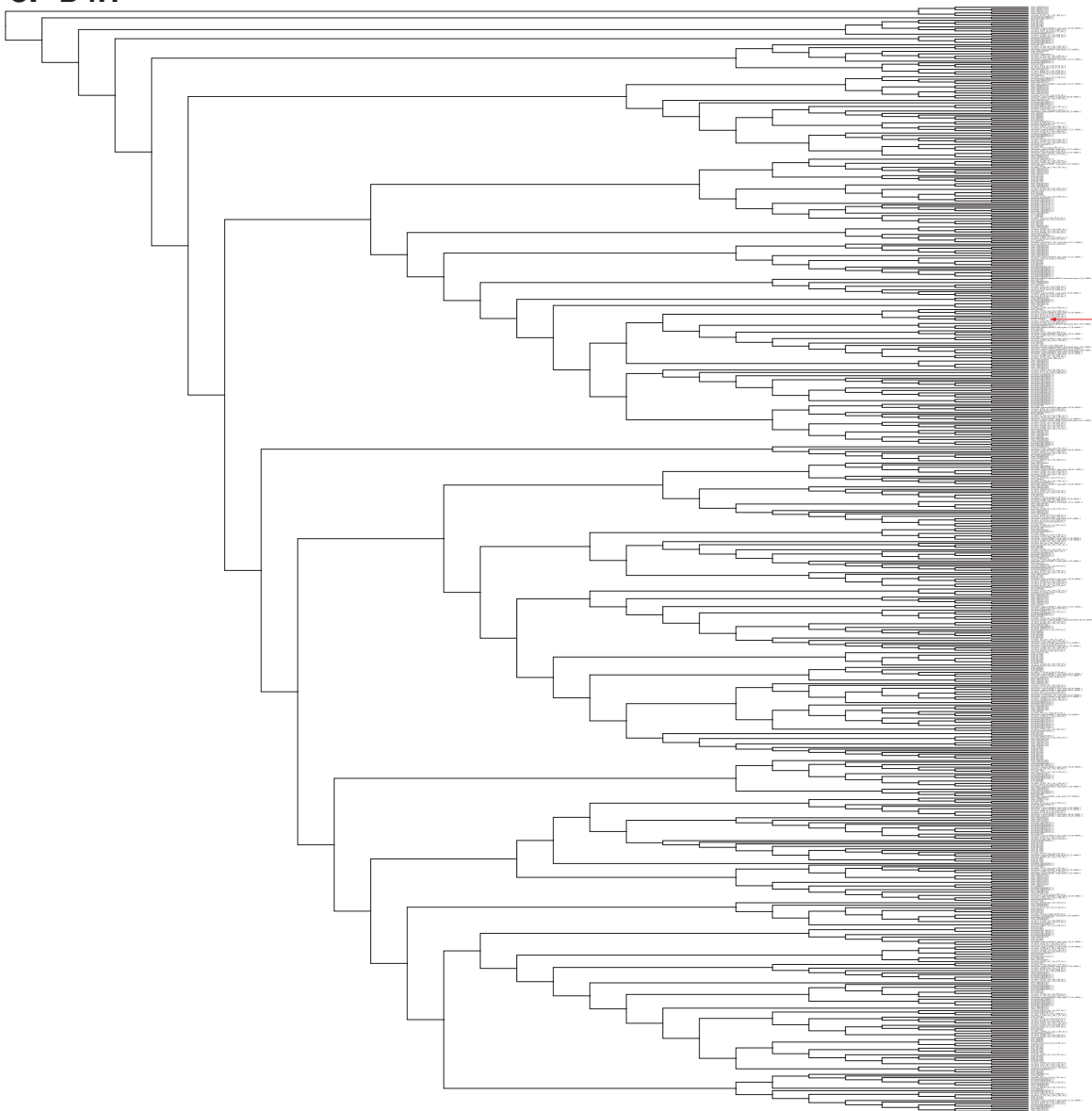
C. AAE



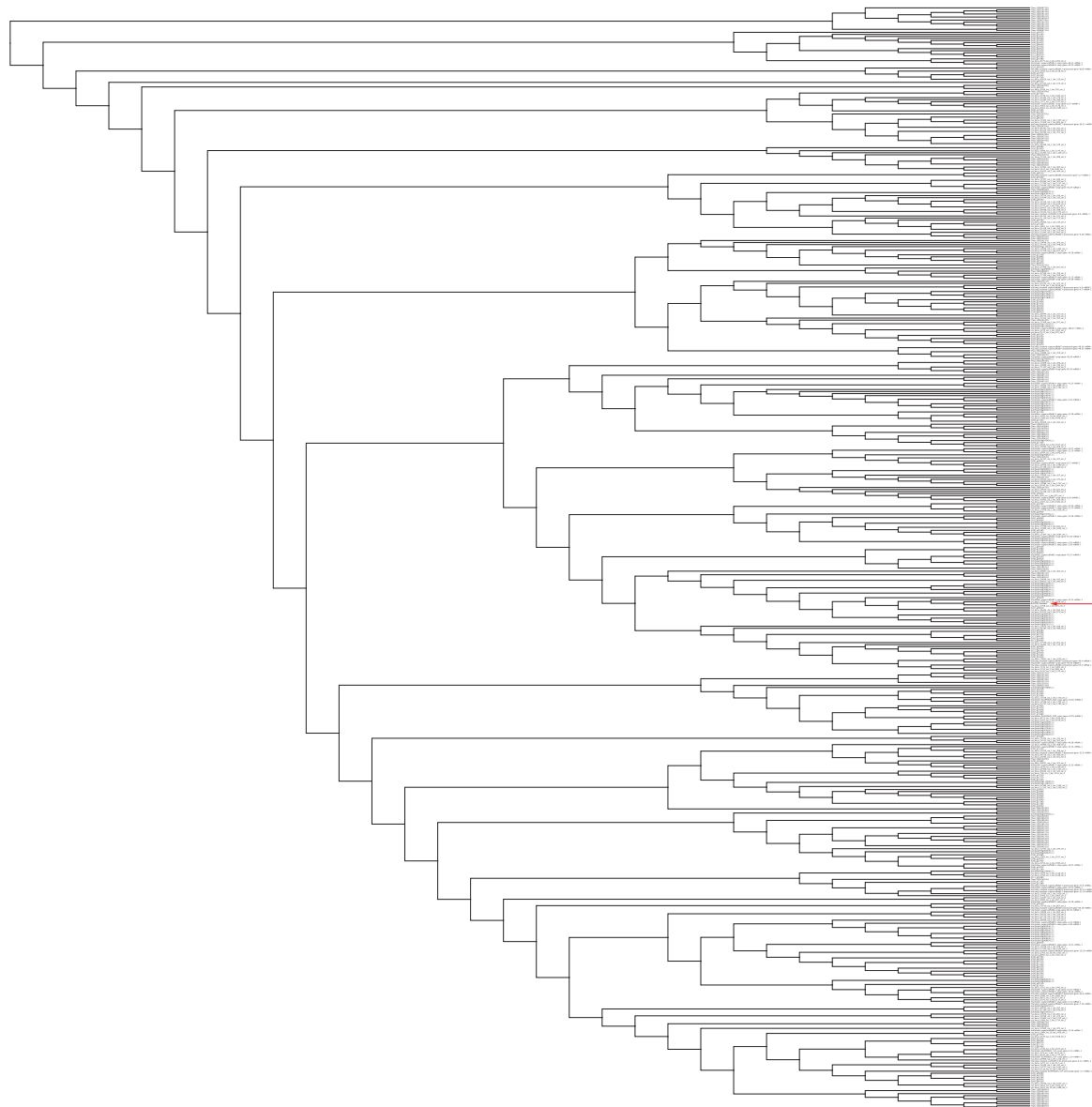
d. DAT



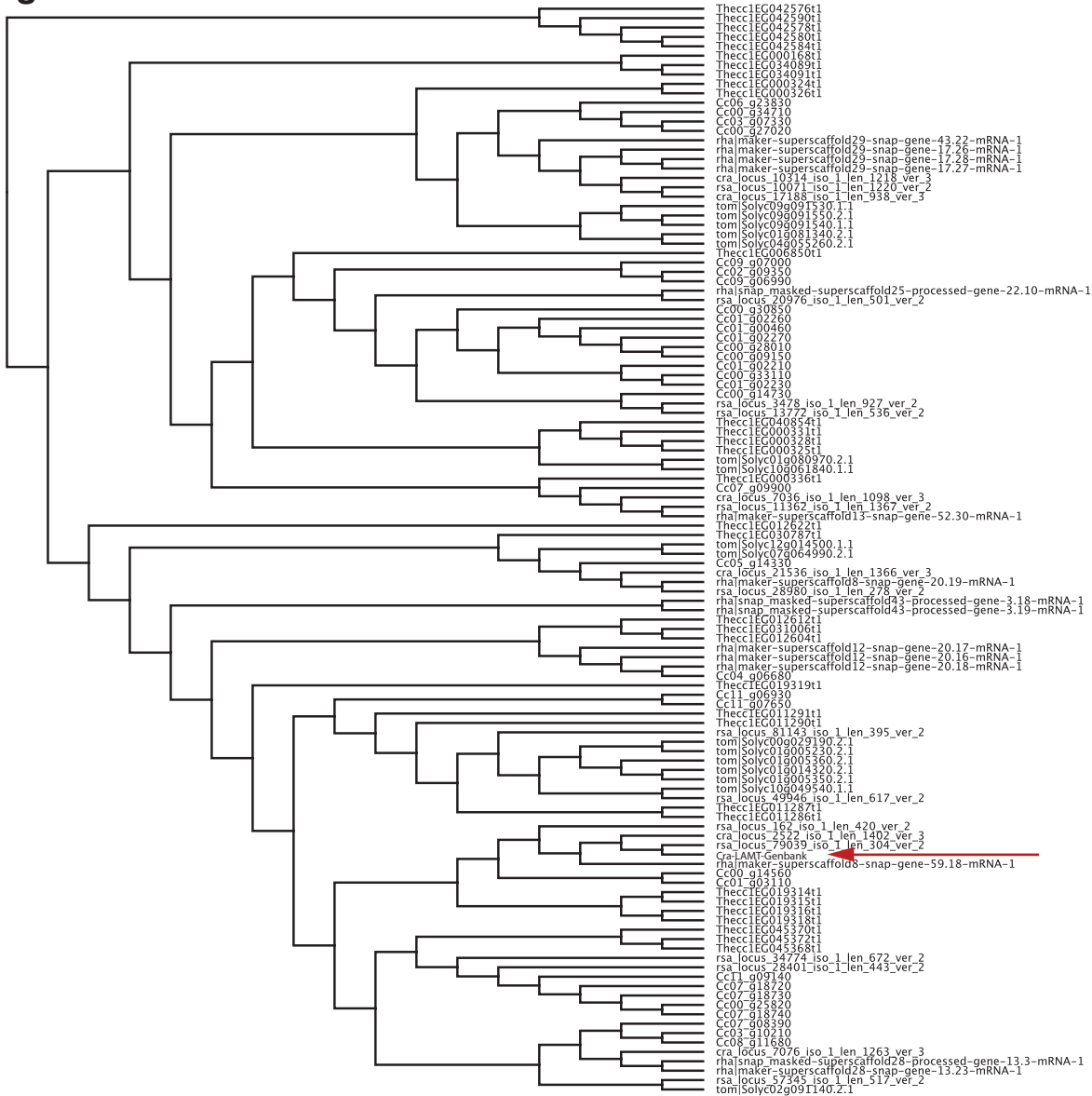
e. D4H



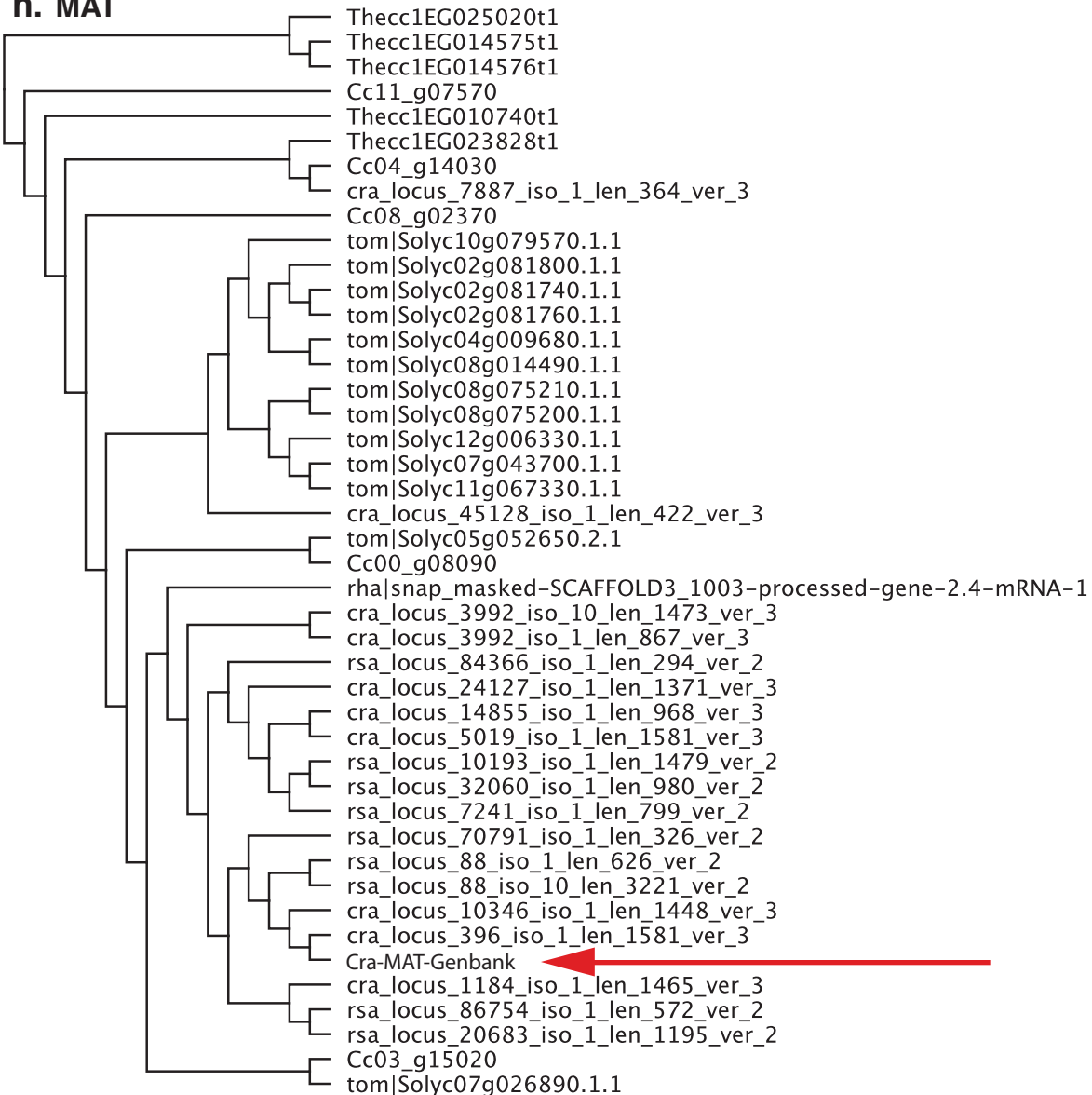
f. G10H



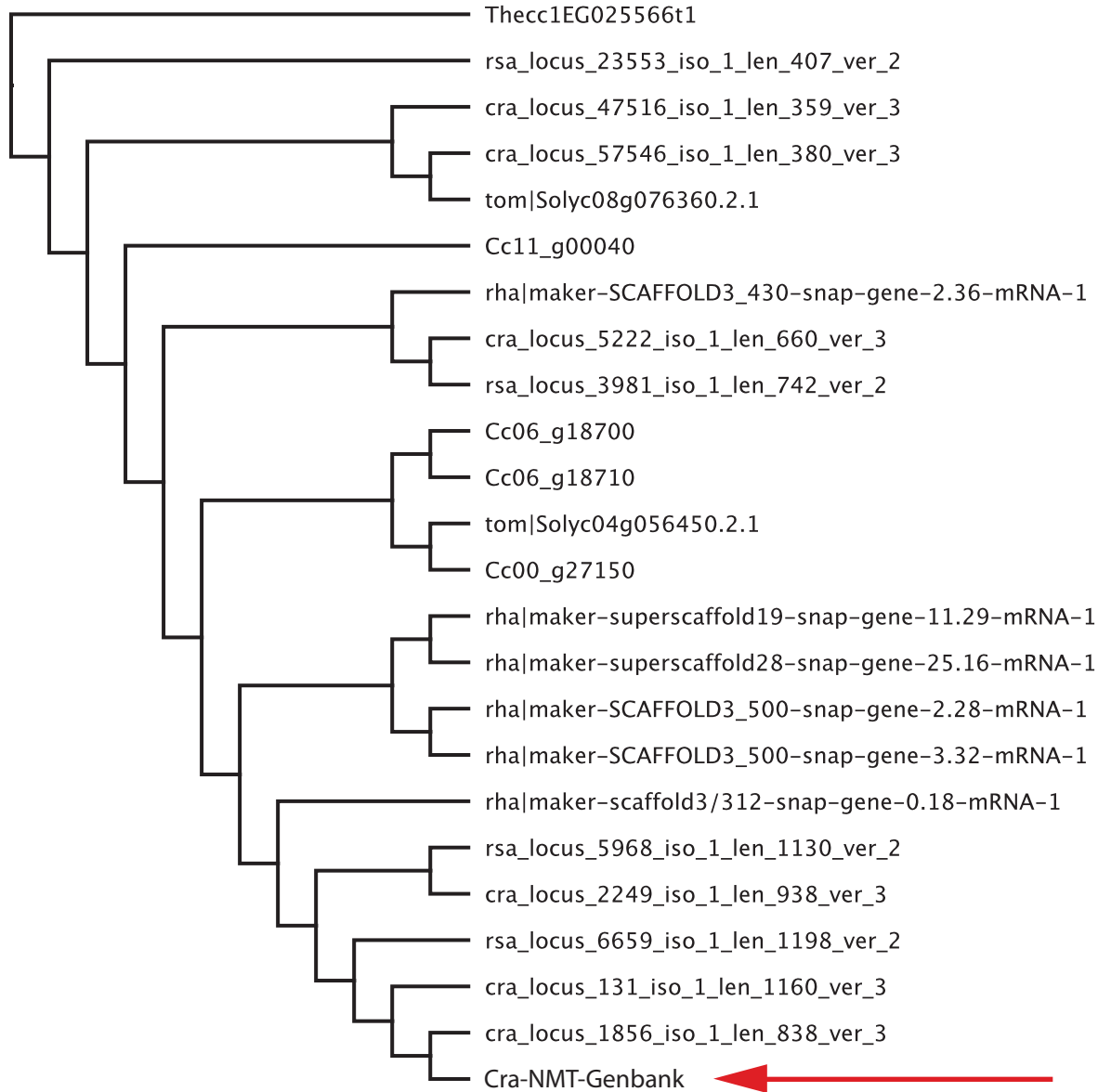
g. LAMT



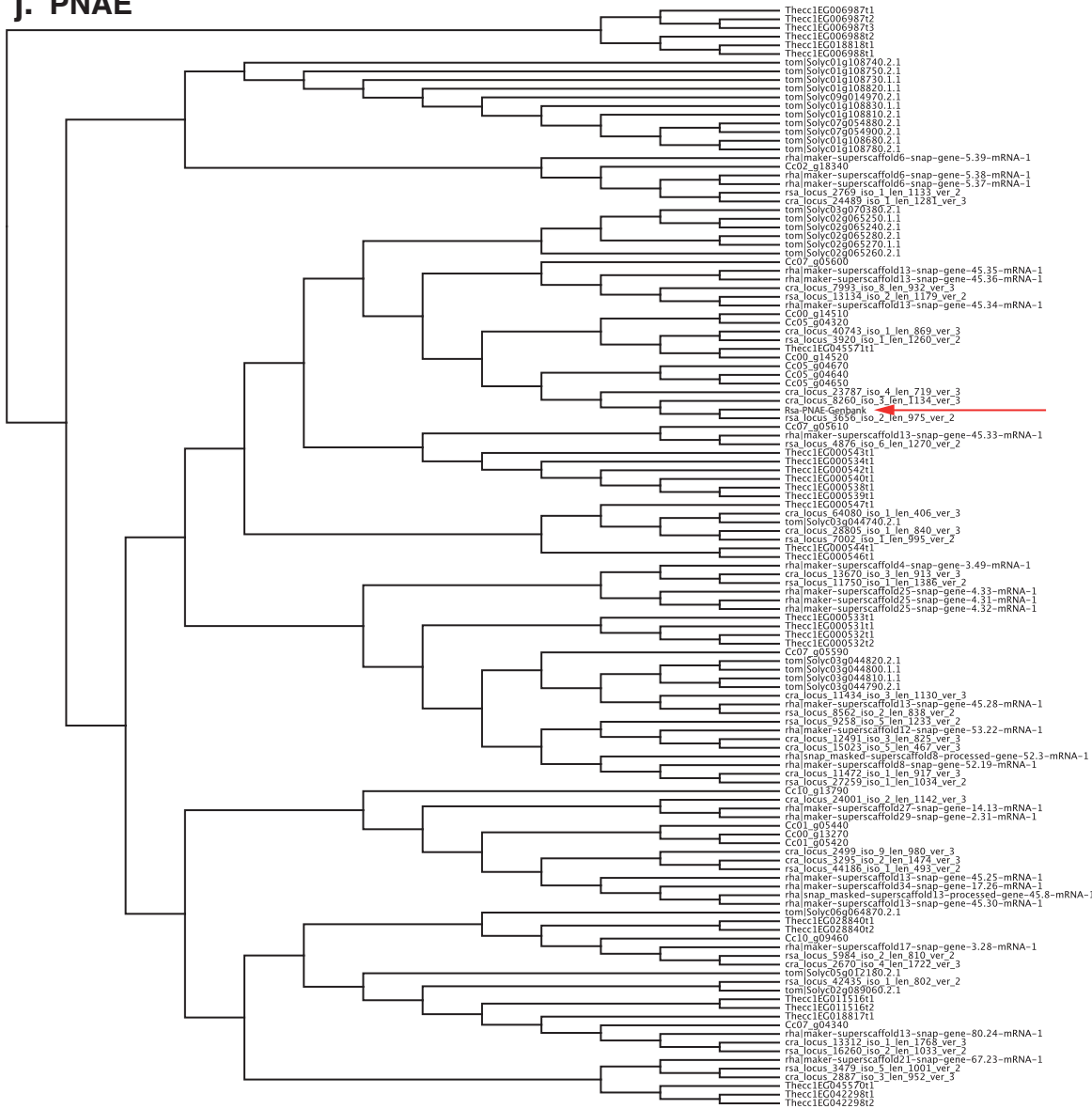
h. MAT



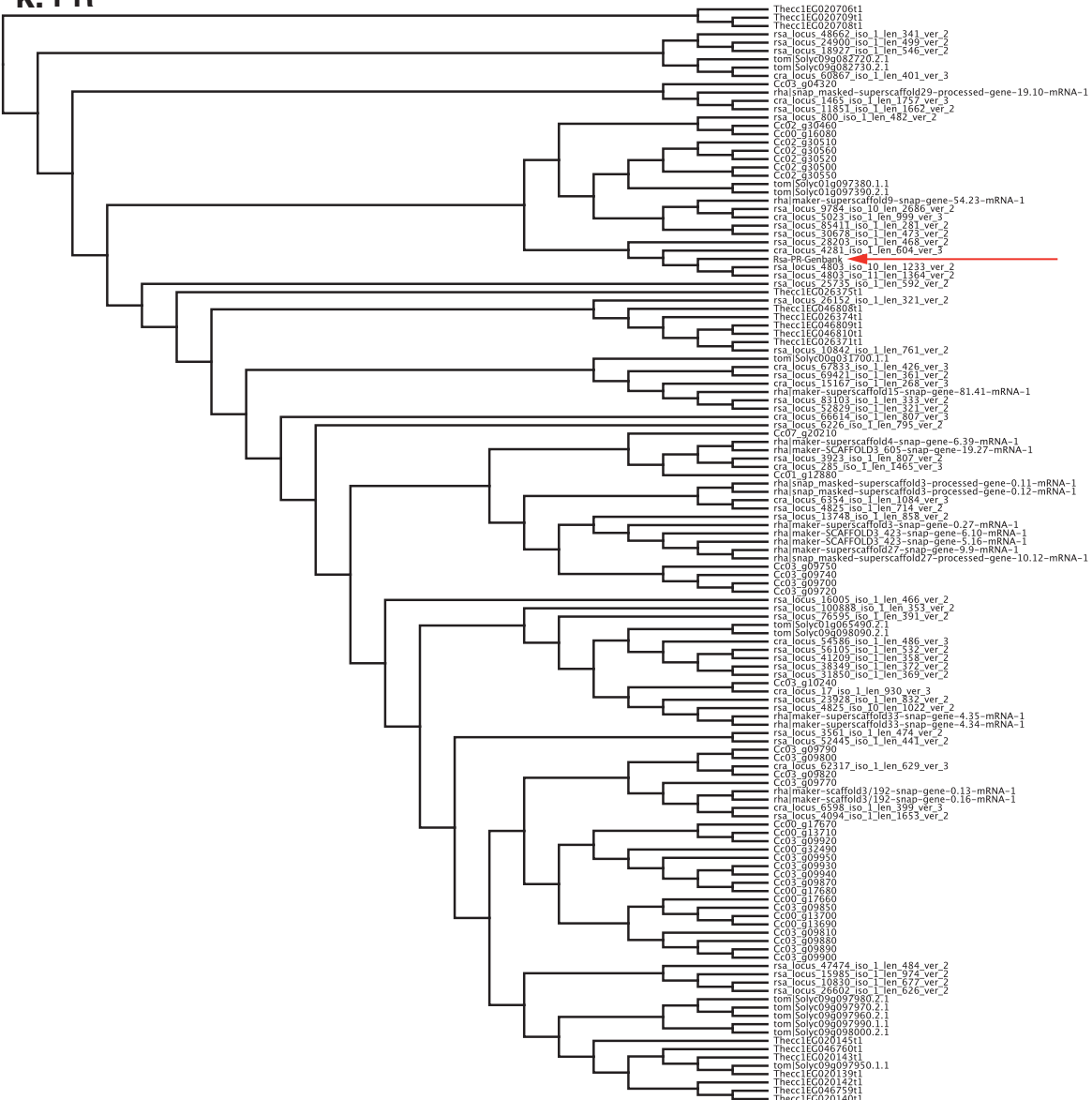
i. NMT



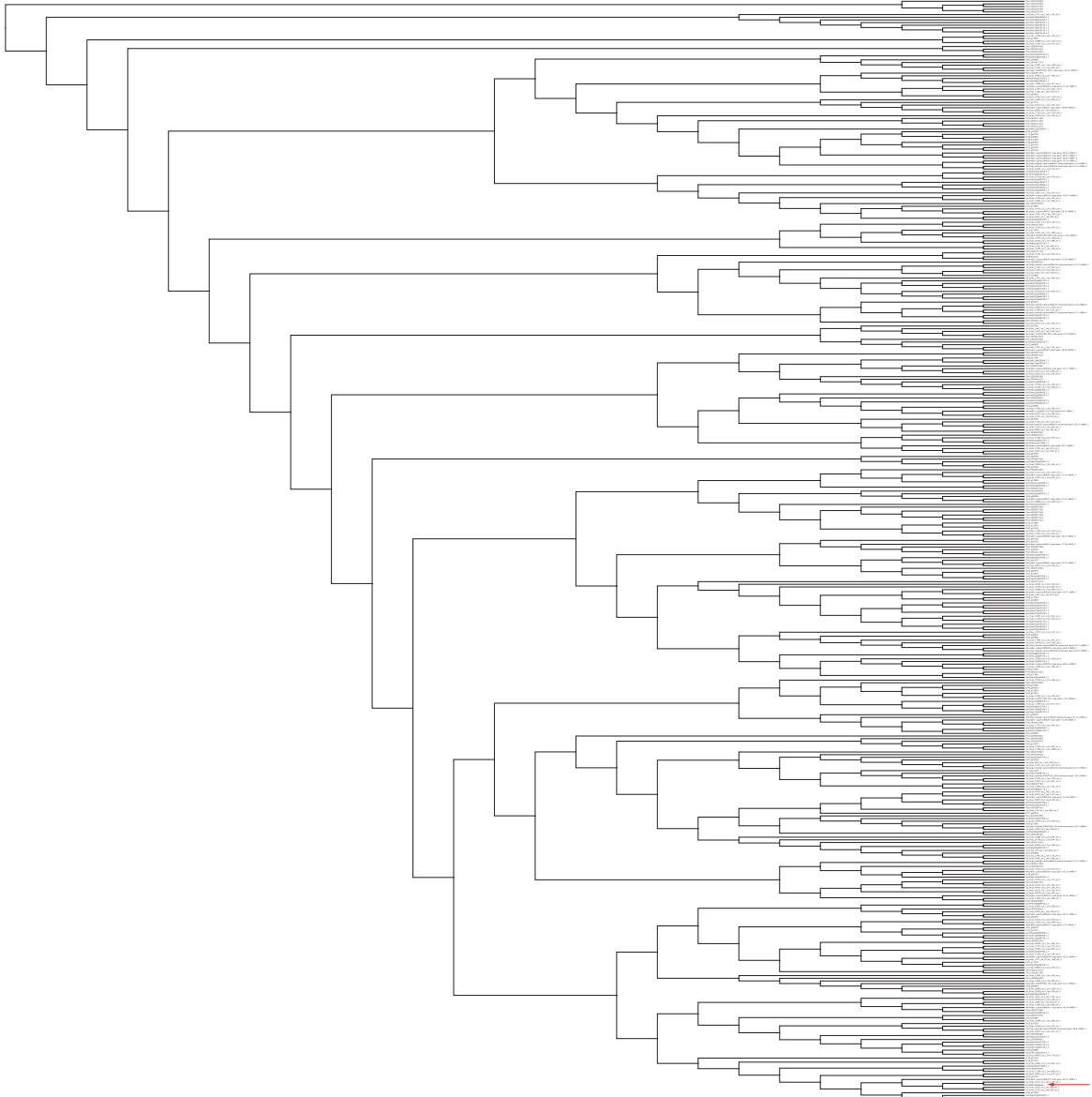
J. PNAE



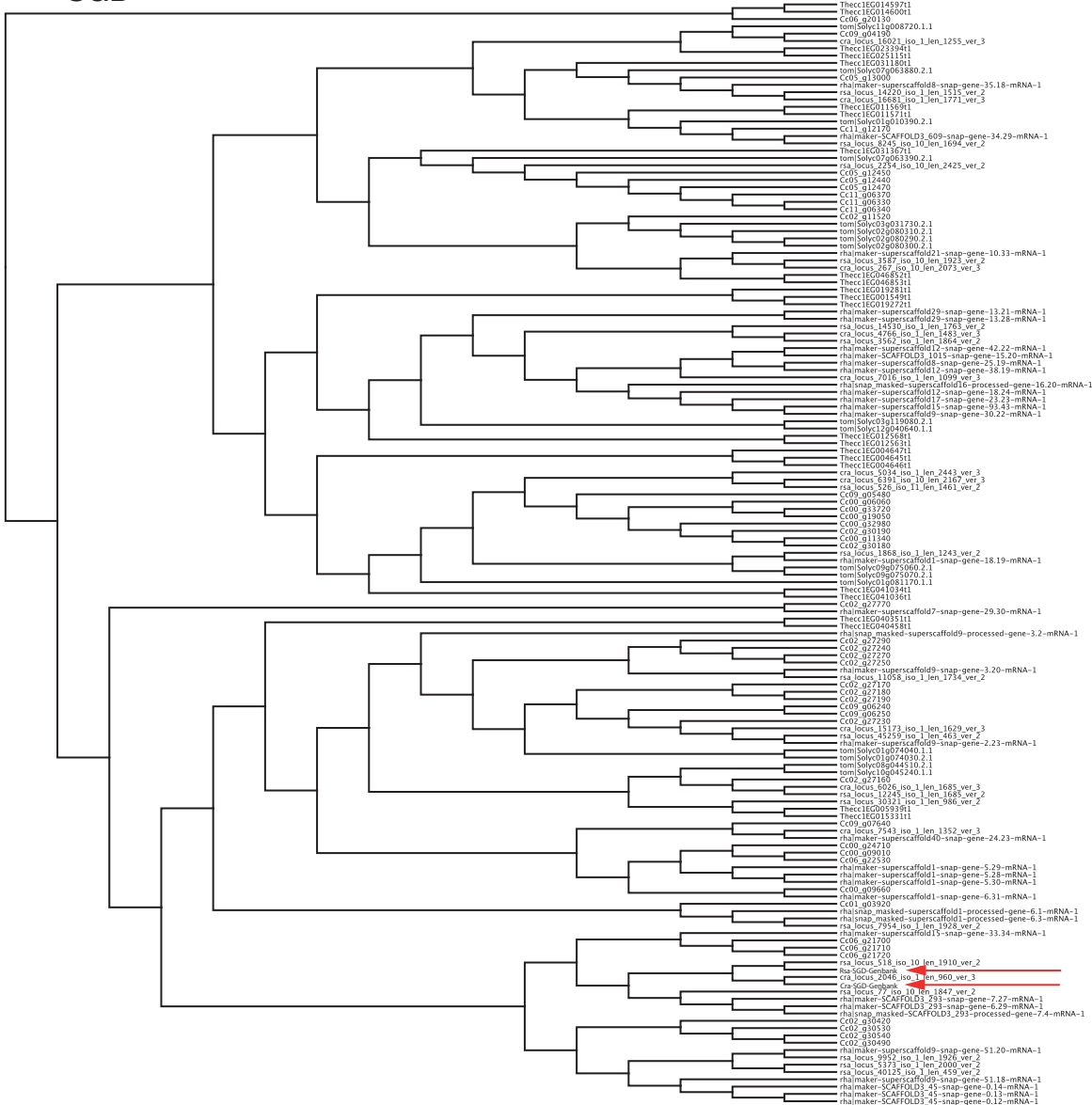
K. PR



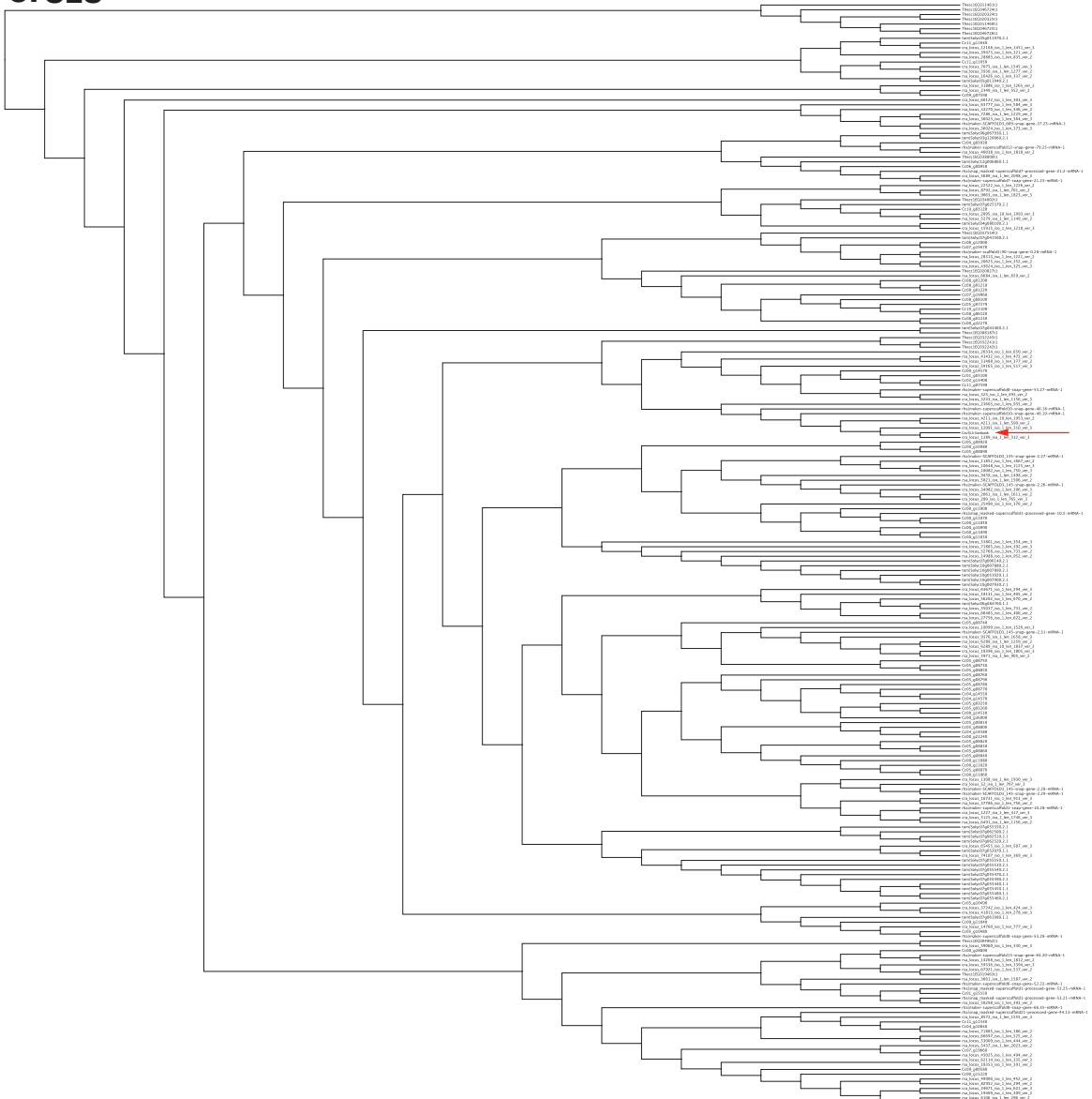
I. PRX1



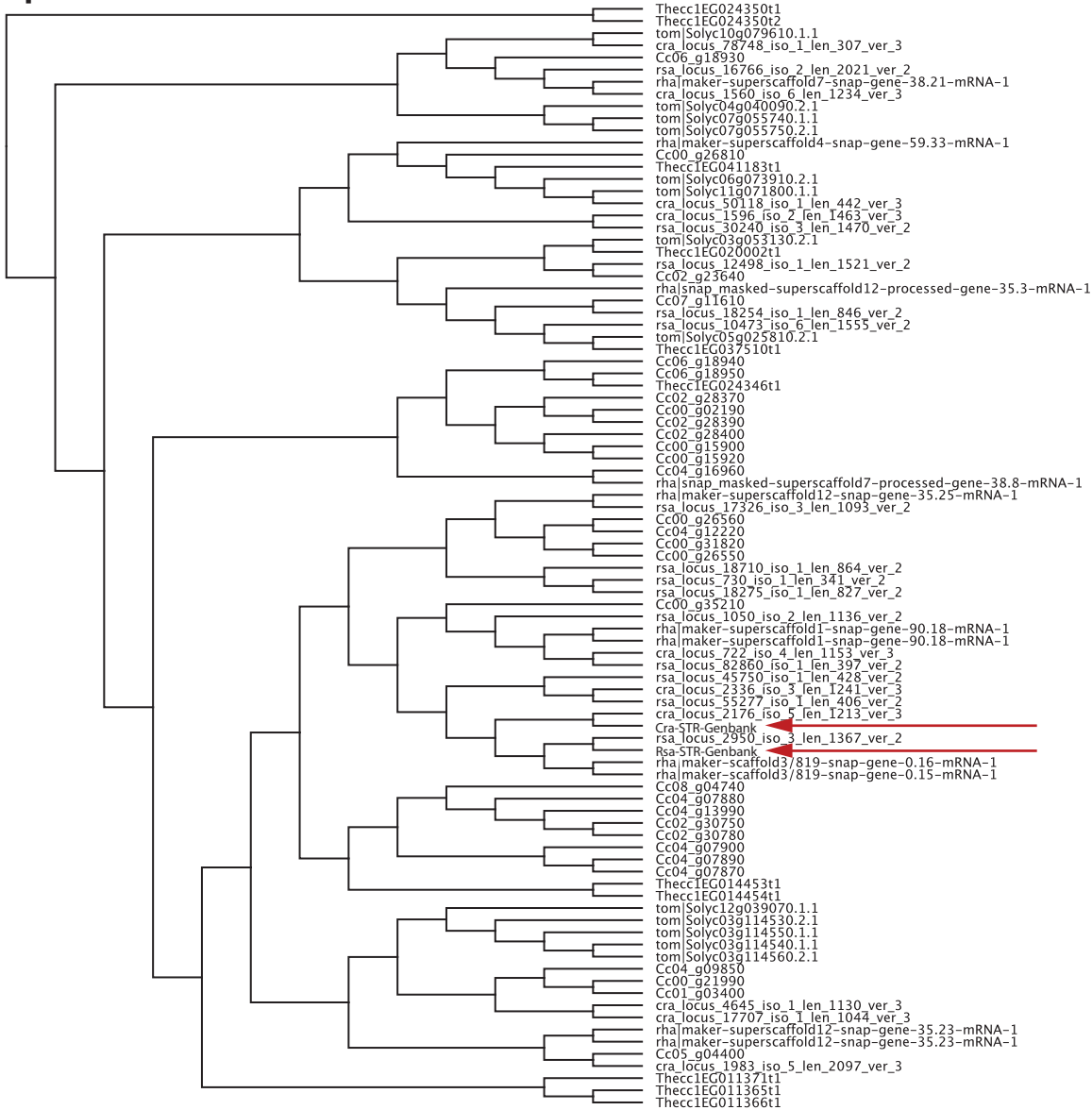
n. SGD



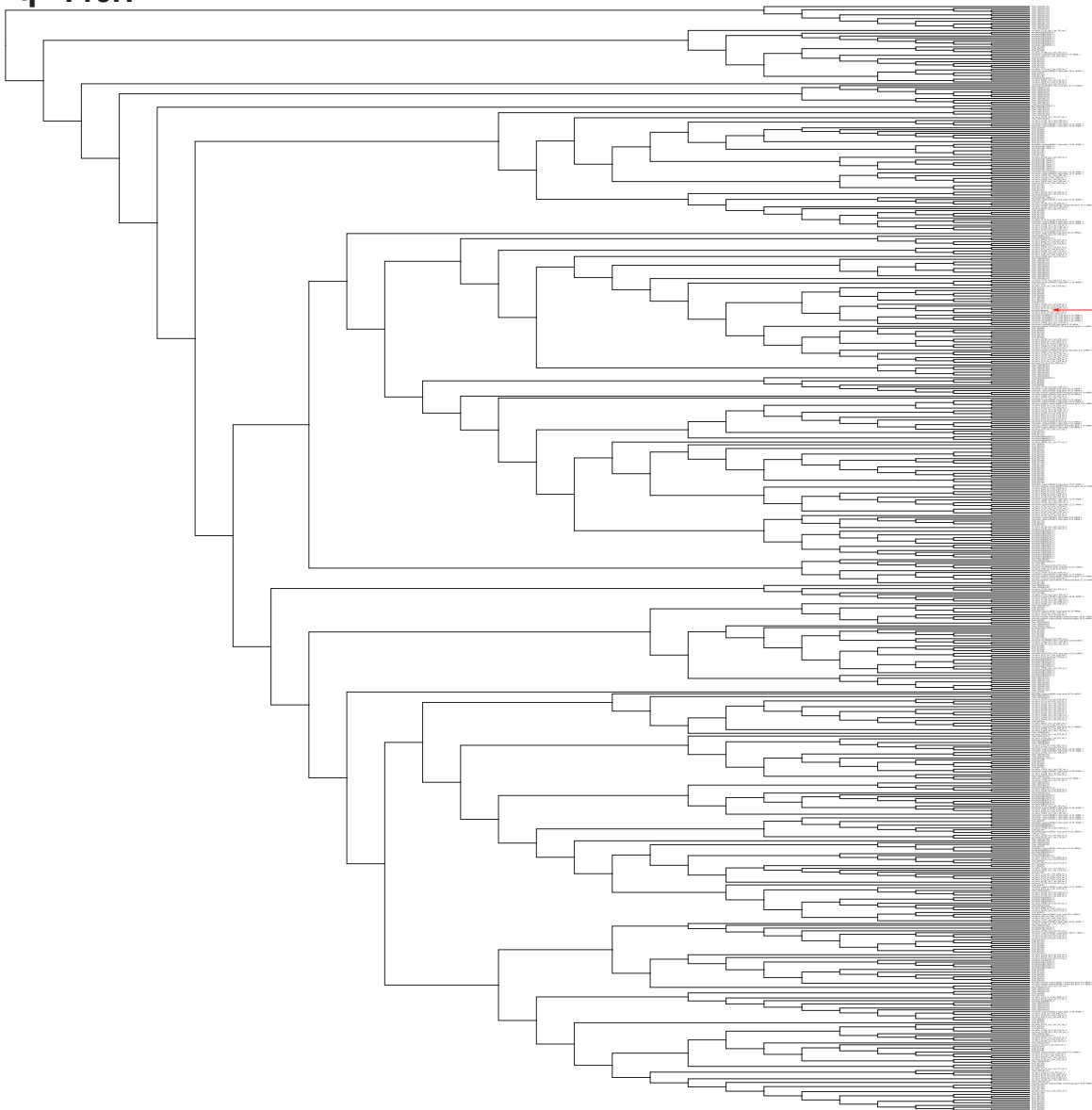
O. SLS



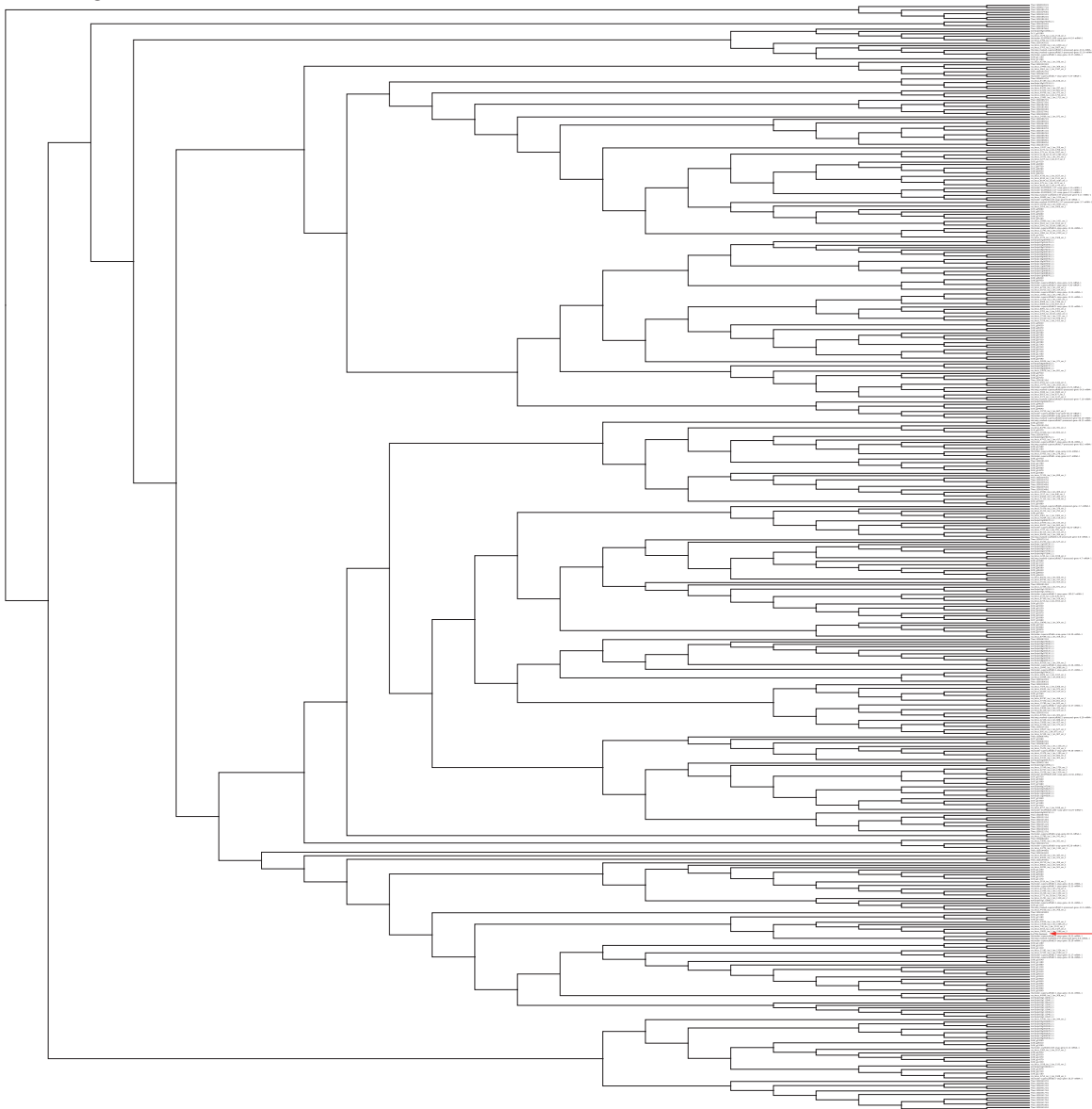
p. STR



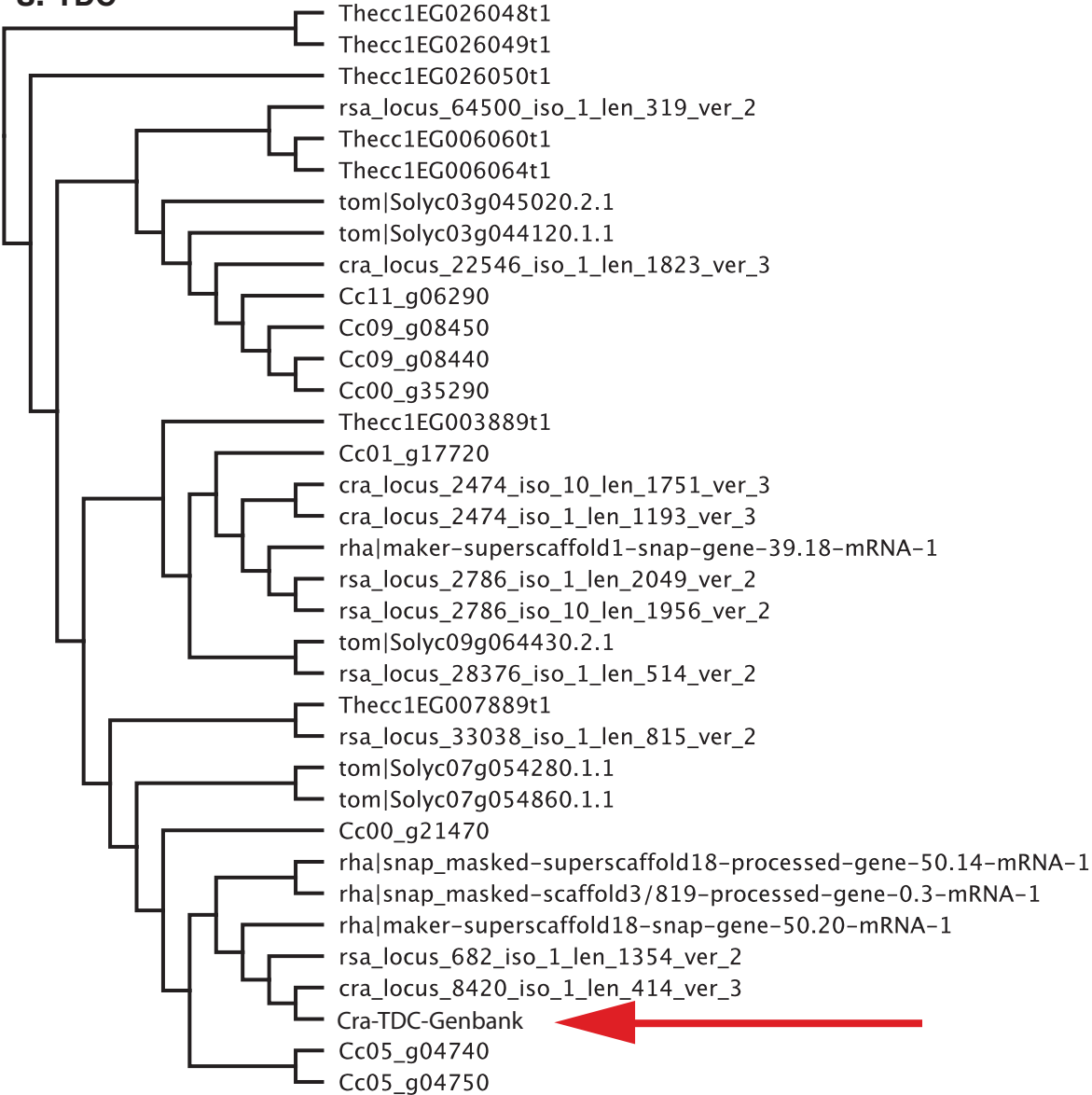
q. T16H



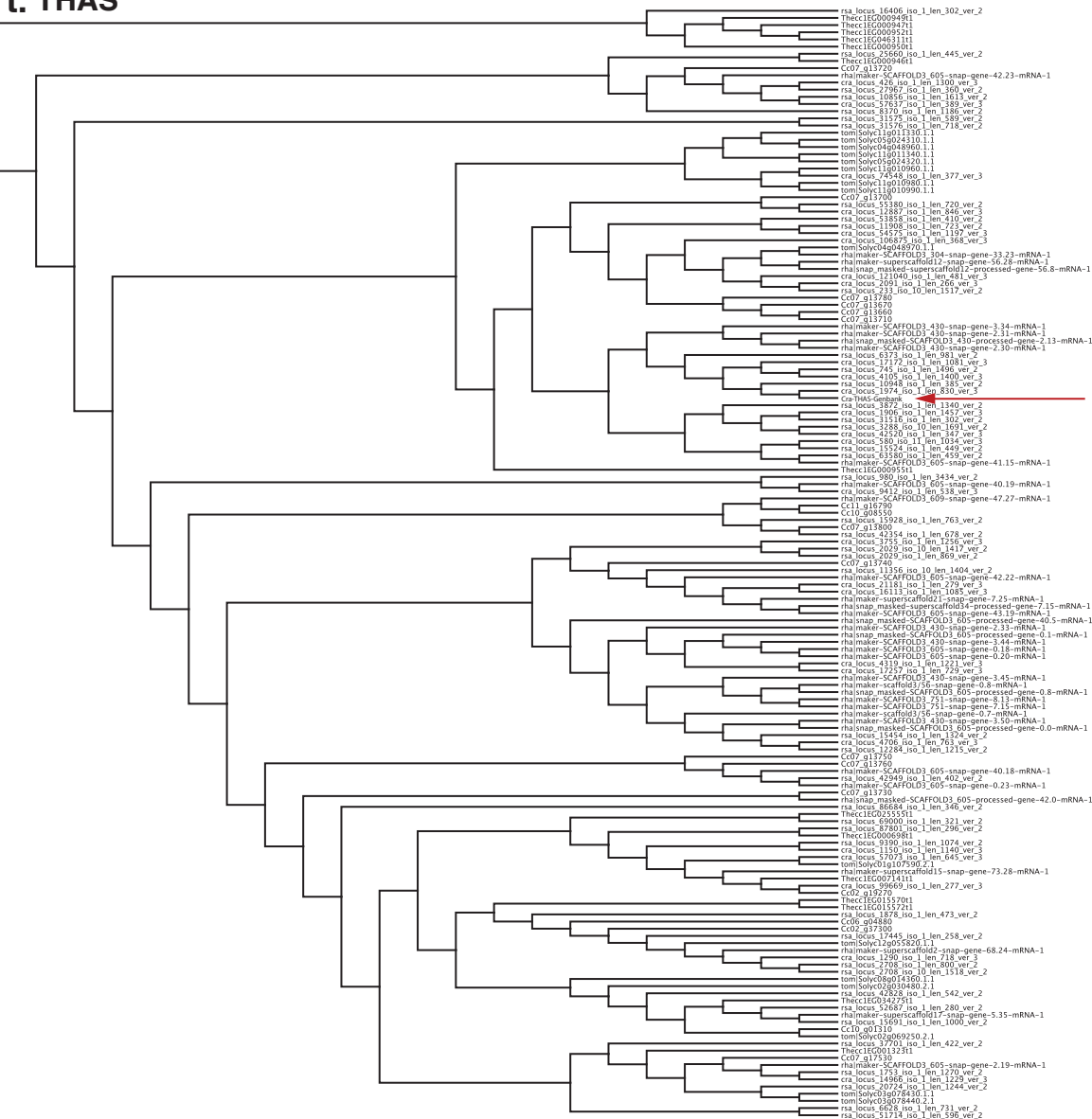
r. T19H



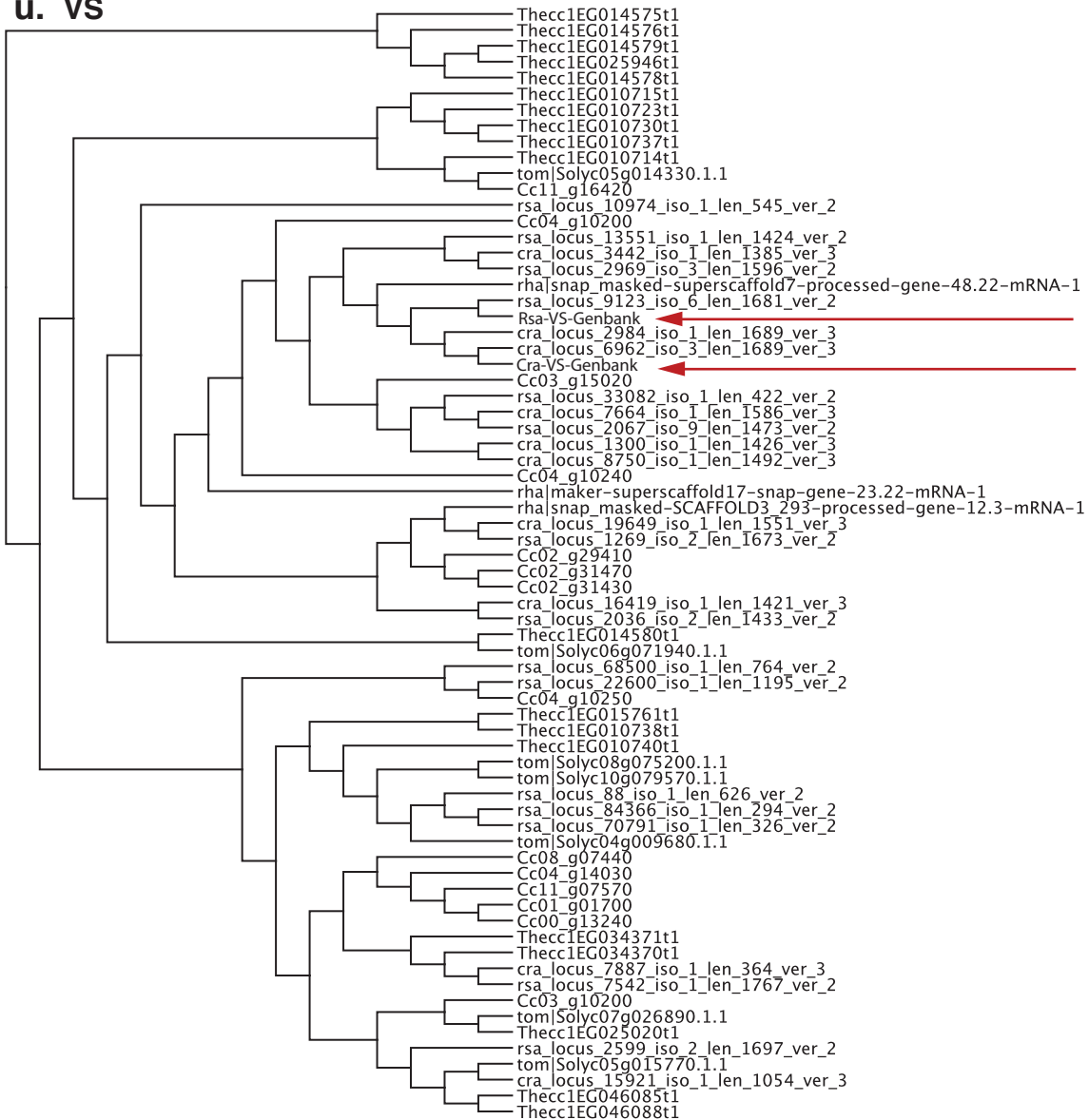
s. TDC



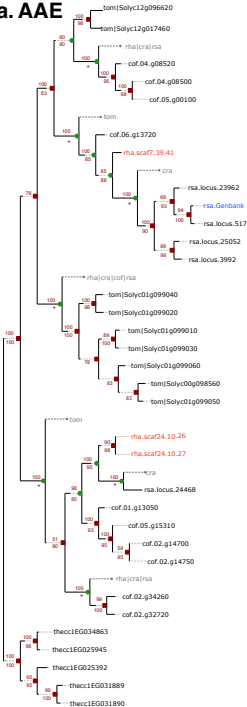
t. THAS



u. VS



a. AAE

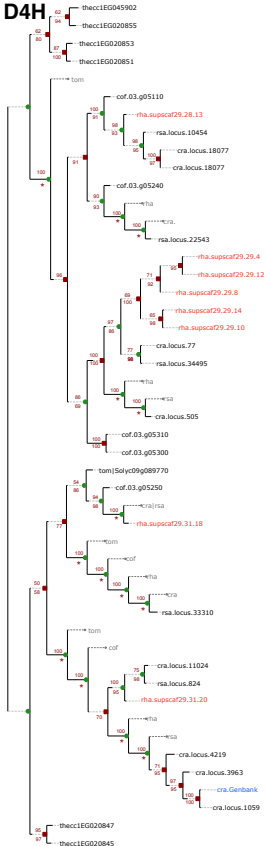


1.14

Duplications : 26-1

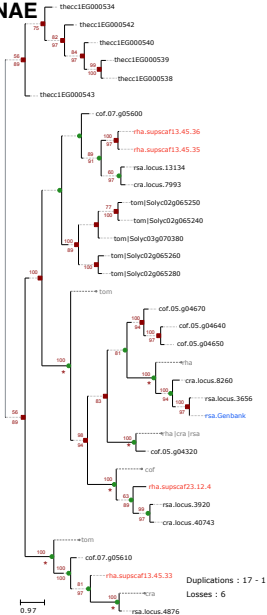
Losses : 7

b. D4H

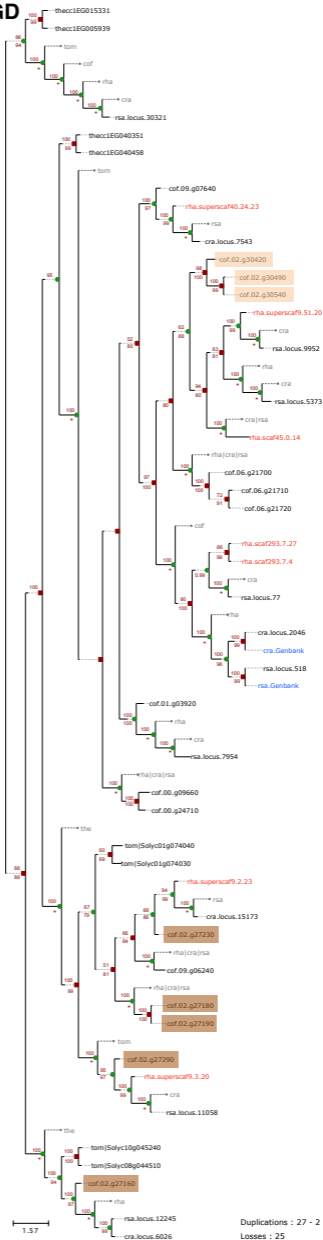


 Duplications : 20 - 1
 1.43
 Losses : 14

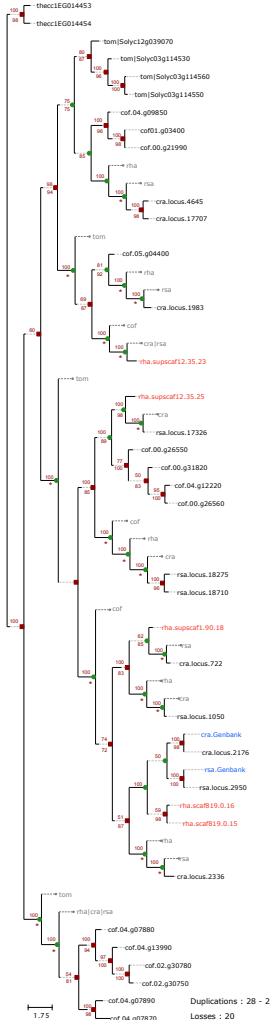
C.PNAE



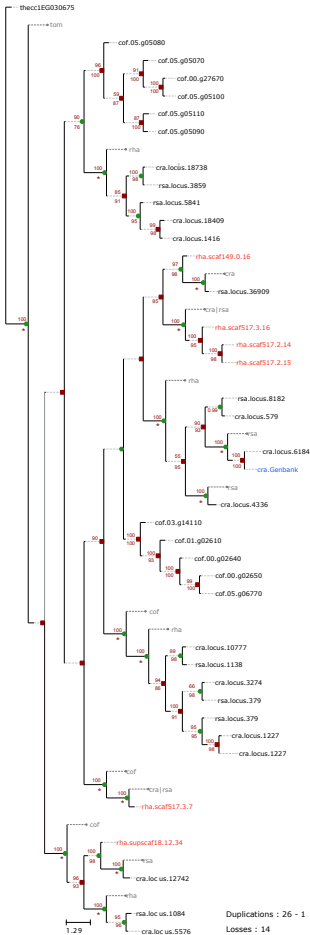
d. SGD



e. STR



f. T16H



g.VS

thecc1EG014580

