Prediction and identification of functional cisNATs in plants

Veerendra Gadekar

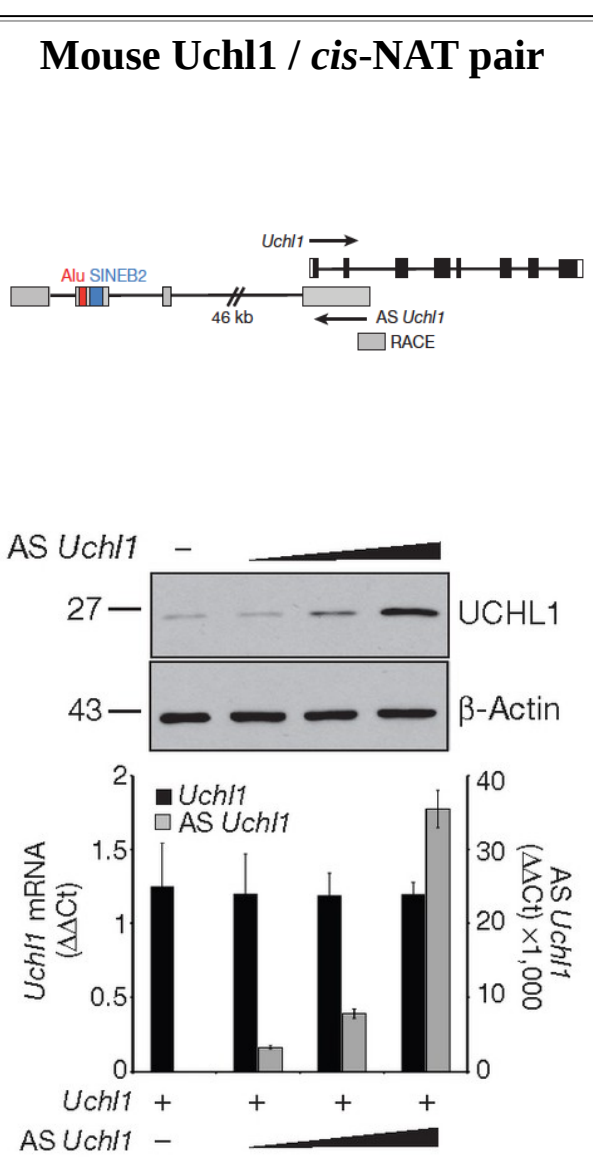
Department of Theoretical Chemistry

University of Vienna

Bled, Feb 17, 2017

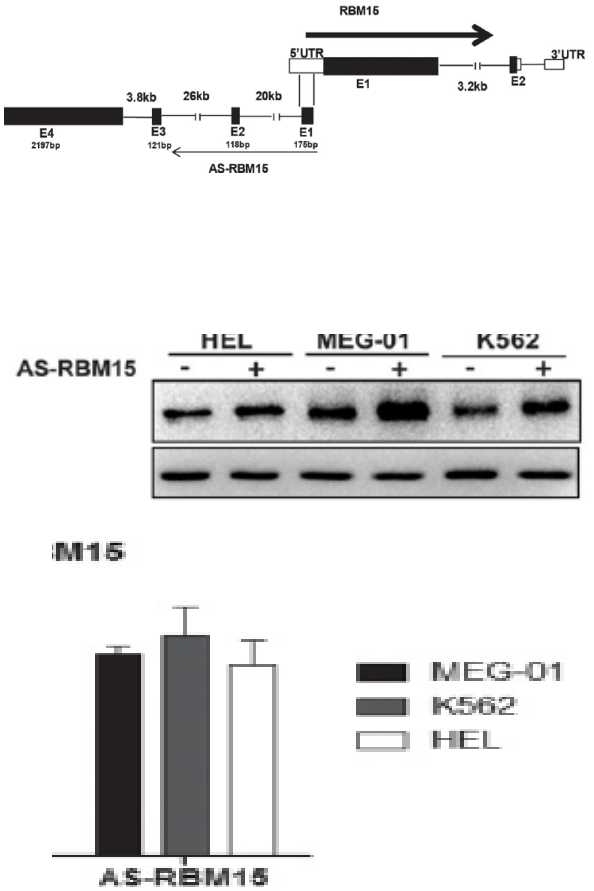
Functional antisense long noncoding RNAs

Mouse Uchl1 / *cis*-NAT pair



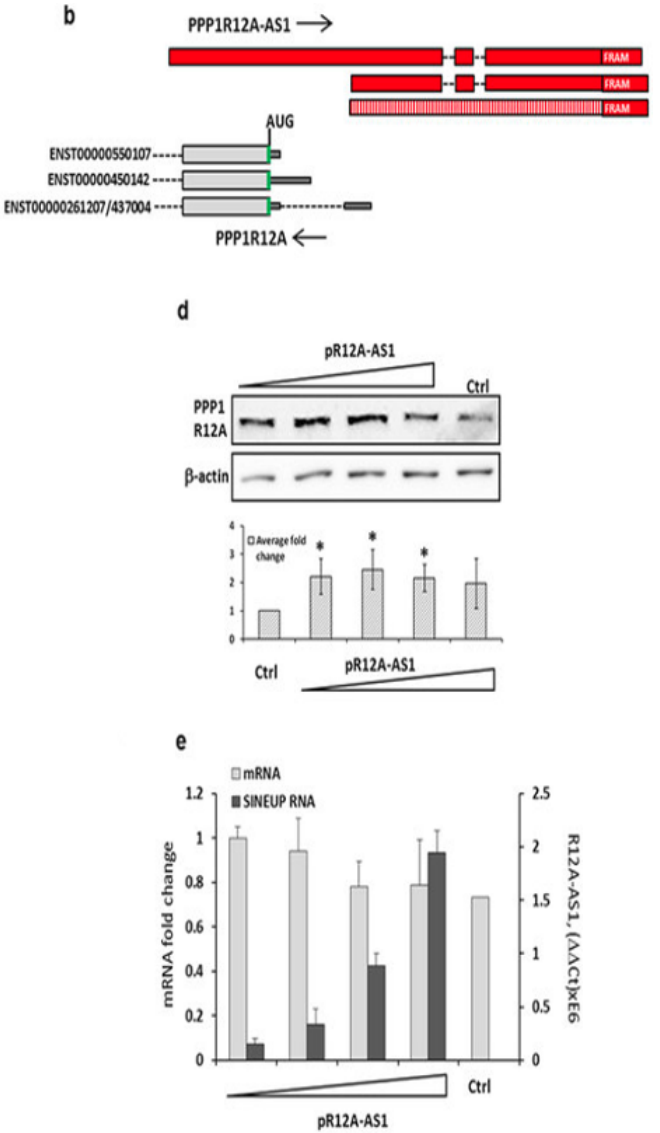
Carrieri et al. 2012

Human RBM15 / *cis*-NAT pair



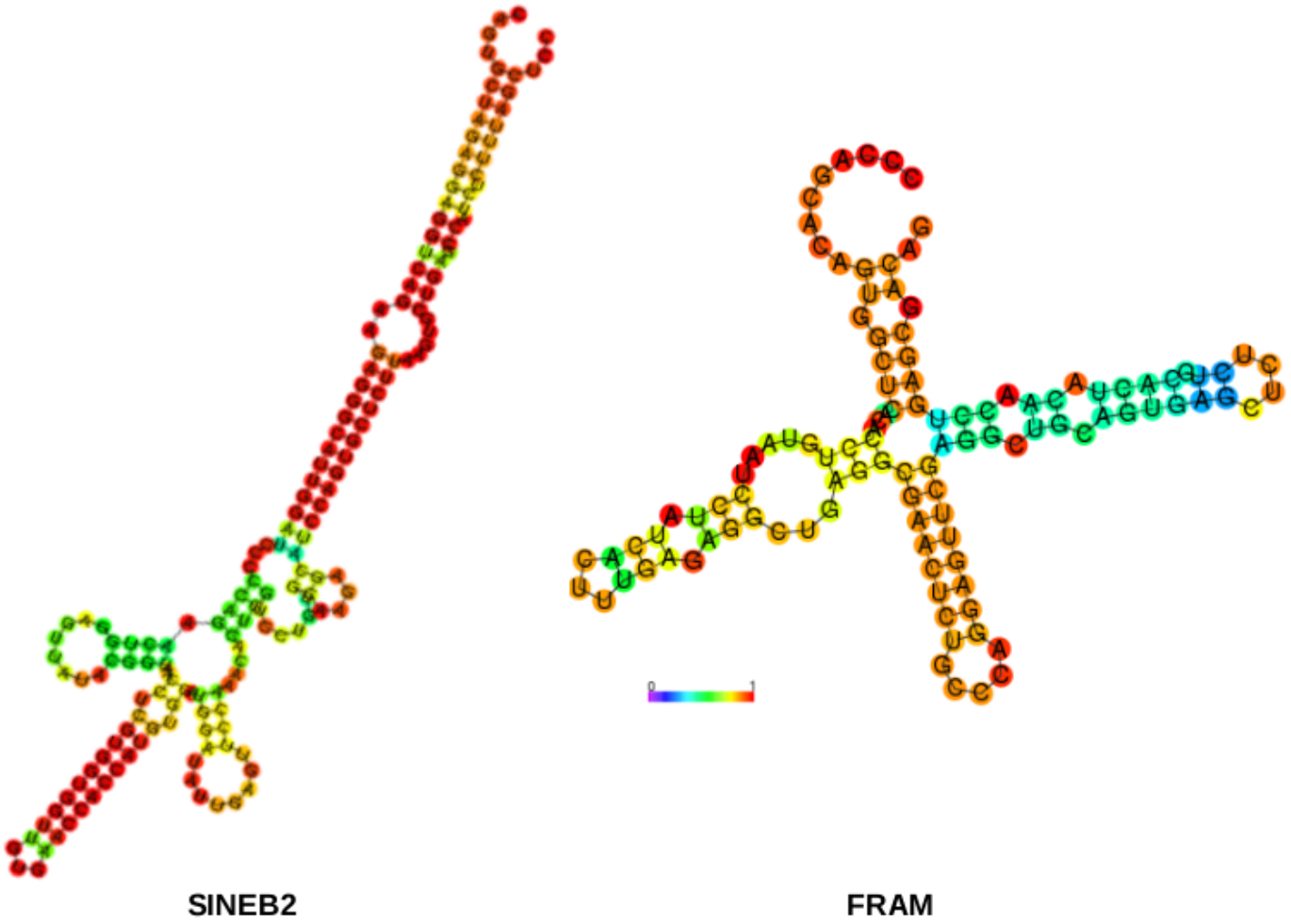
Tran et al. 2016

Human PPP1R12A / *cis*-NAT pair



Schein et al. 2016

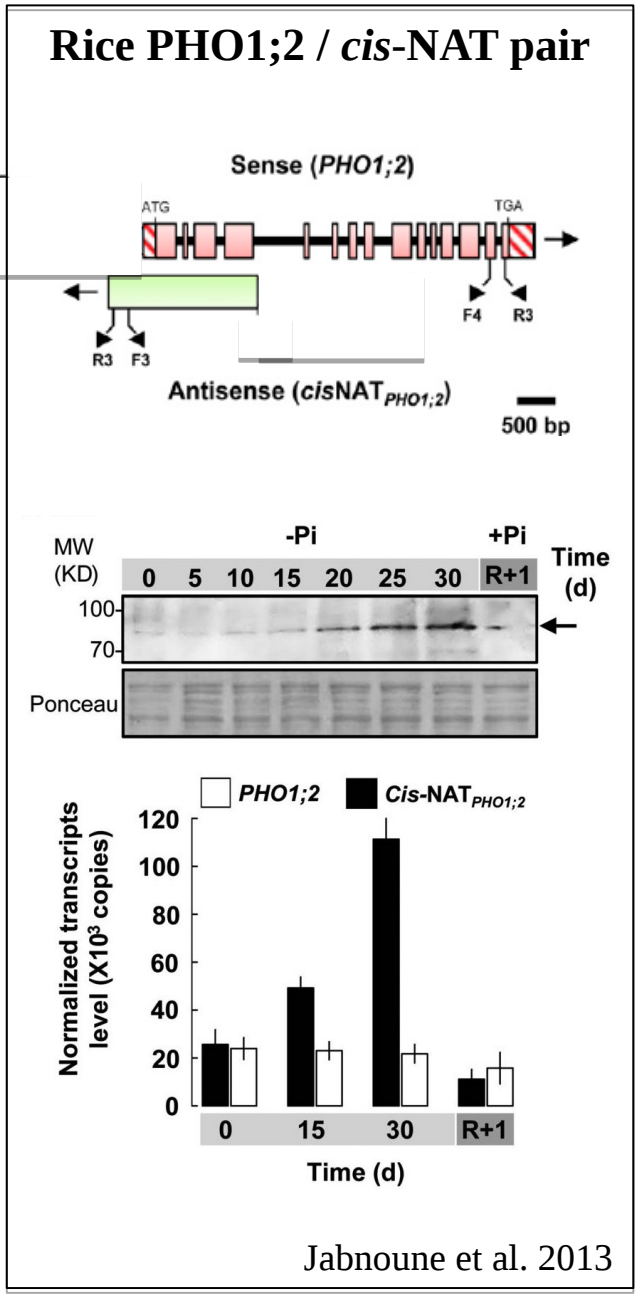
SINEB2 and FRAM are predicted to fold into a different structures



Supplementary Figure S5. Human FRAM and mouse SINEB2 elements obtain different secondary structures
Secondary structure of SINEs, predicted by RNAfold program. Prediction confidence for each base is indicated by the color code.

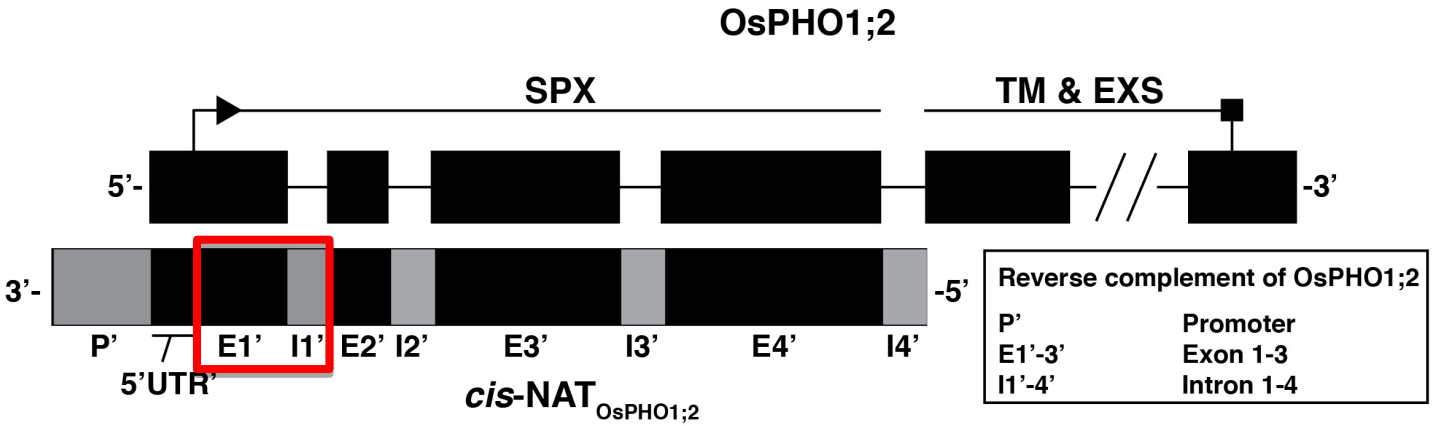
Schein et al. 2016

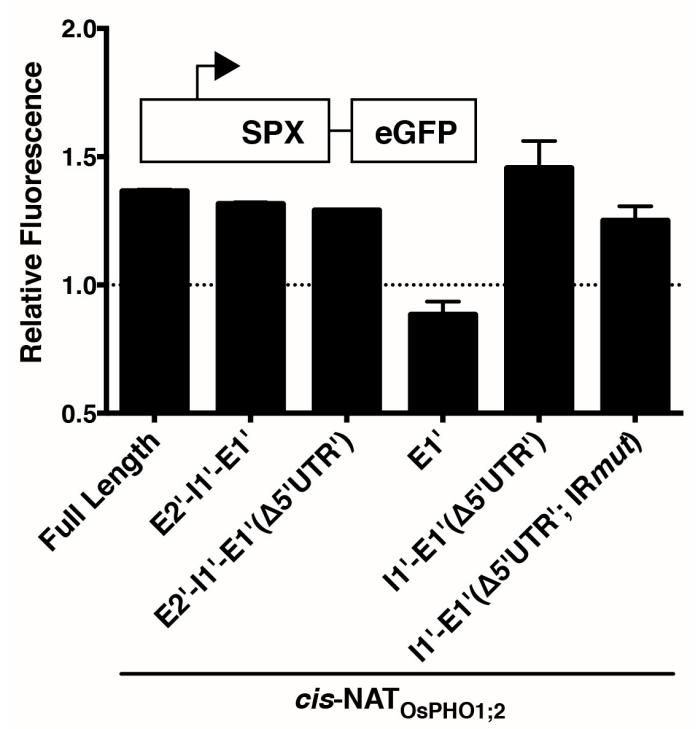
A Rice cis-Natural Antisense RNA

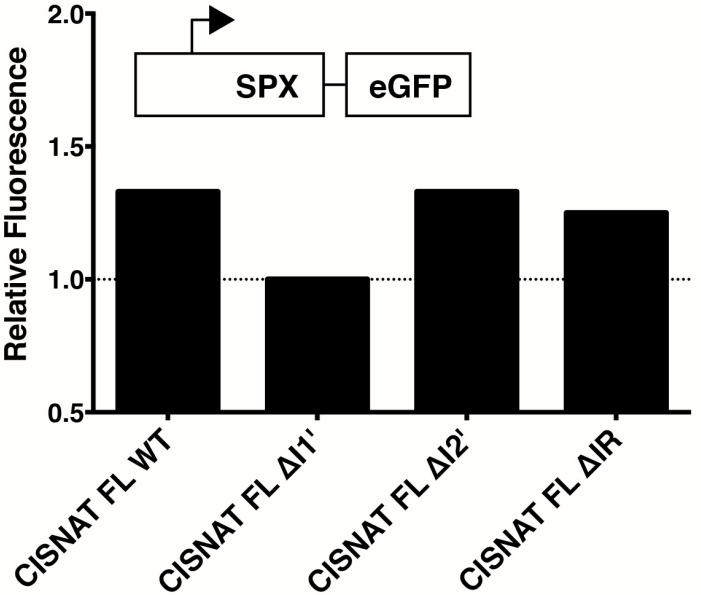


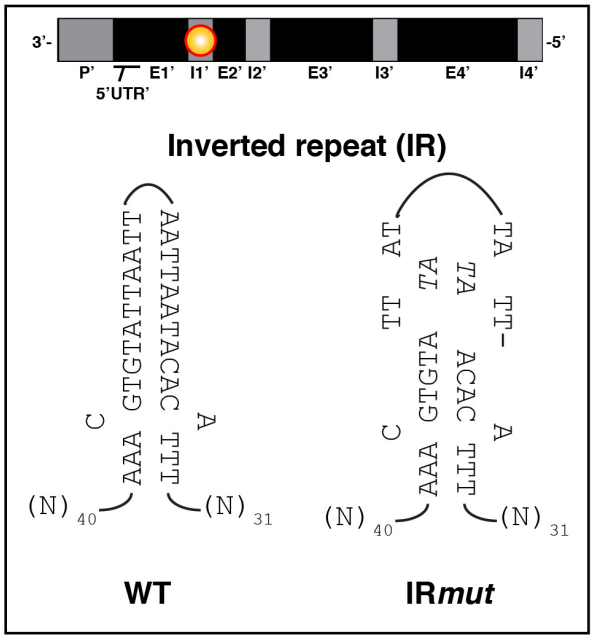
Jabnoune et al. 2013

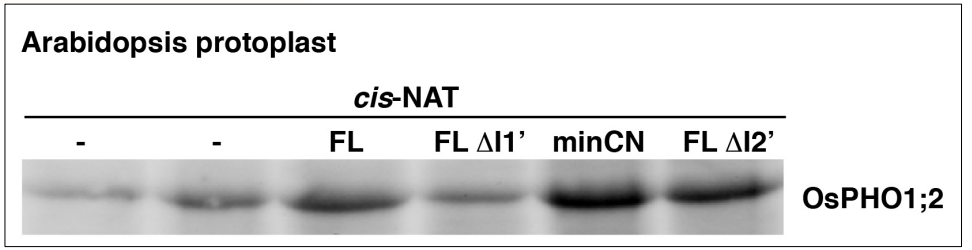
Dissection of *cis*-NAT$_{OsPHO1;2}$ translation enhancement of OsPHO1;2

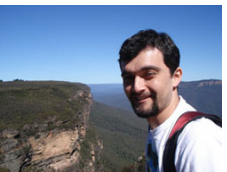












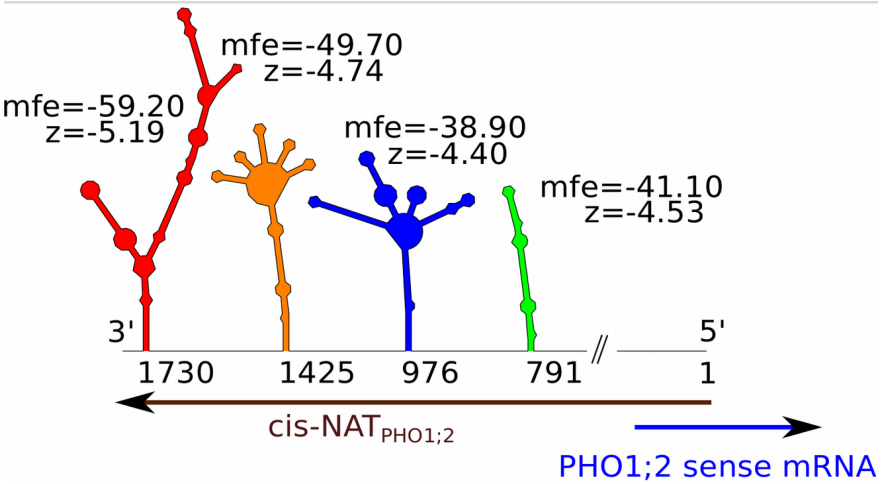
Rodrigo S. Reis Poirier Lab @ UniL

Aim

Mechanism underlying the translation enhancement of PHO1 by *cis*-NAT in rice

Objectives

- Identification of similar putative *cis*-NATs in plant genomes
- Application of comparative genomics to identify evolutionary conserved structure elements
- Select candidate structures for experimental testing



Data set

Species	Genome Version	Genome size in Mbp	No. of chromosomes
Oryza sativa	IRGSP-1.0	~374	12
Oryza rufipogon	OR_W1943	~338	12
Oryza punctata	AVCL00000000	~394	12
Oryza nivara	AWHD00000000	~338	12
Oryza meridionalis	Oryza_meridionalis_v1.3	~336	12
Oryza longistaminata	O_longistaminata_v1.0	~326	60198 scaffold
Oryza indica	ASM465v1	~412	12
Oryza glumaepatula	ALNU02000000	~372	12
Oryza glaberrima	AGI1.1	~316	12
Oryza brachyantha	Oryza_brachyantha.v1.4b	~260	12
Oryza barthii	O.barthii_v1	~308	12
Brassica rapa	IVFCAASv1	~284	10
Brassica oleracea	v2.1	~489	9
Brassica napus	AST_PRJEB5043_v1	~738	20899 supercontigs
Arabidopsis thaliana	TAIR10	~135	5
Arabidopsis lyrata	v.1.0	~206	8

Phylogenetic tree

No. of annotated transcripts

Species	Protein coding	ncRNA	Other	Total
Oryza sativai	42132	55619	0	97751
Oryza rufipogon	47441	2020	758	50219
Oryza punctata	41060	28	5167	46255
Oryza nivara	48360	24	1648	50032
Oryza meridionalis	43455	2283	0	45738
Oryza longistaminata	31686	0	0	31686
Oryza indica	40745	2232	45461	88438
Oryza glumaepatula	46893	19	3264	50176
Oryza glaberrima	33164	1997	38717	73878
Oryza brachyantha	32037	18	2100	34155
Oryza barthii	41595	1681	665	43941
Brassica rapa	41025	1	1827	42853
Brassica oleracea	59220	5	0	59225
Brassica napus	101040	0	0	101040
Arabidopsis thaliana	48321	3917	1775	54013
Arabidopsis lyrata	32667	0	0	32667

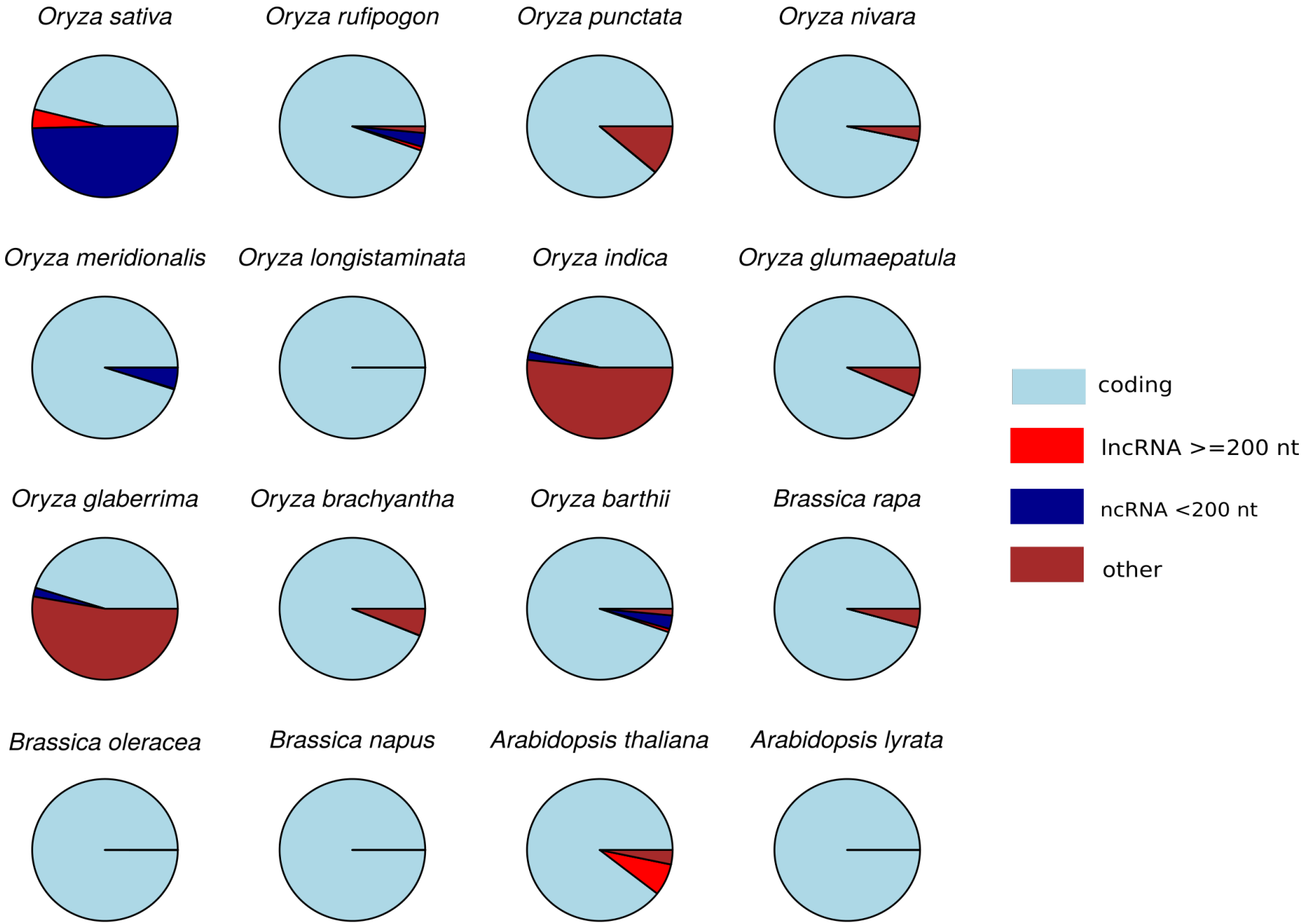
No. of annotated transcripts

Species	Protein coding	ncRNA	Other	Total
Oryza sativa	42132	55619	0	97751
Oryza rufipogon	47441	2020	758	50219
Oryza punctata	41060	28	5167	46255
Oryza nivara	48360	24	1648	50032
Oryza meridionalis	43455	2283	0	45738
Oryza longistaminata	31686	0	0	31686
Oryza indica	40745	2232	45461	88438
Oryza glumaepatula	46893	19	3264	50176
Oryza glaberrima	33164	1997	38717	73878
Oryza brachyantha	32037	18	2100	34155
Oryza barthii	41595	1681	665	43941
Brassica rapa	41025	1	1827	42853
Brassica oleracea	59220	5	0	59225
Brassica napus	101040	0	0	101040
Arabidopsis thaliana	48321	3917	1775	54013
Arabidopsis lyrata	32667	0	0	32667

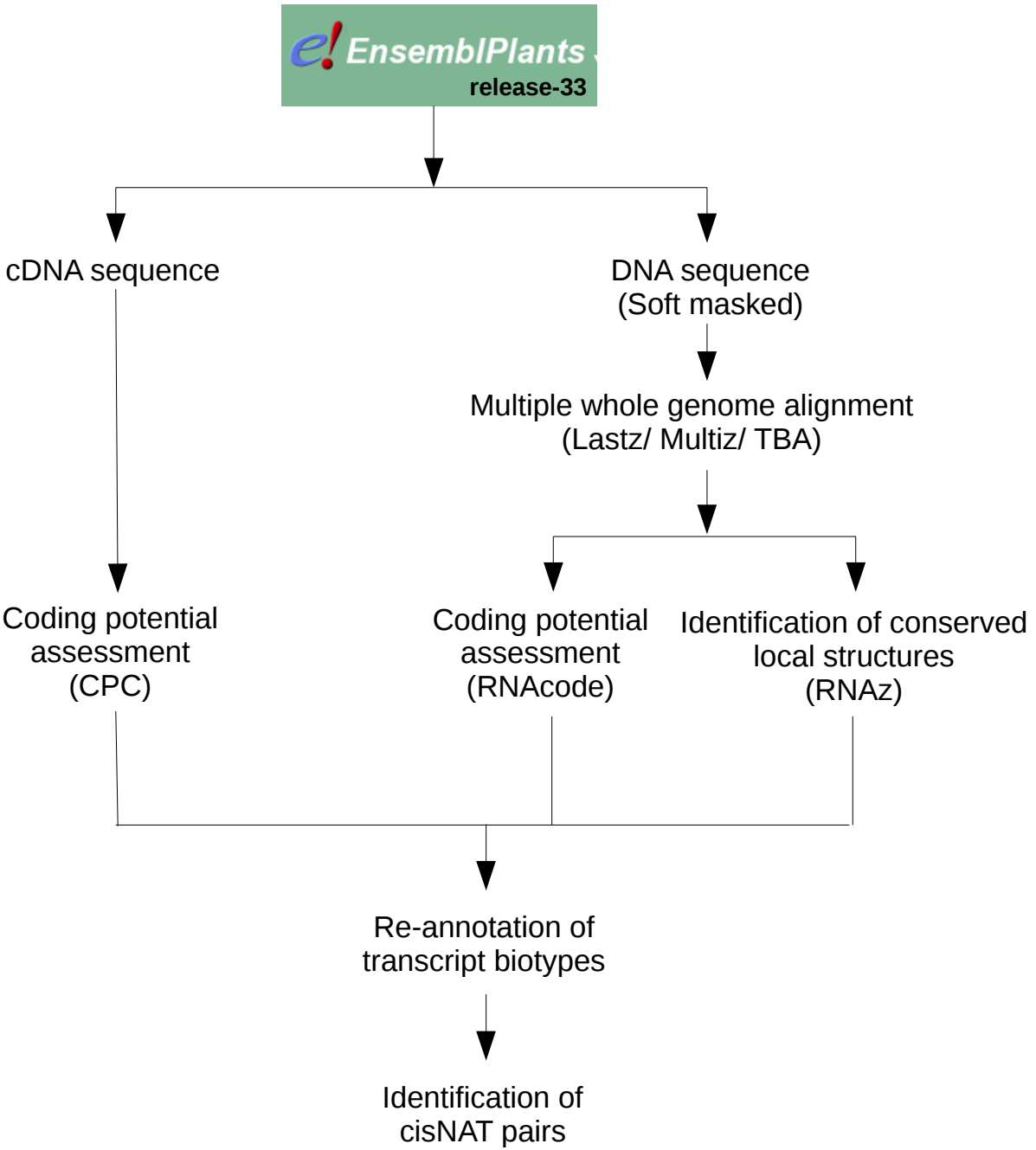
No. of annotated transcripts

Species	Protein coding	lncRNA nt >= 200	ncRNA nt < 200	Other
Oryza sativa	42132	3923	45310	0
Oryza rufipogon	47441	440	1580	758
Oryza punctata	41060	4	24	5167
Oryza nivara	48360	1	23	1648
Oryza meridionalis	43455	82	2201	0
Oryza longistaminata	31686	0	0	0
Oryza indica	40745	15	1669	45461
Oryza glumaepatula	46893	2	17	3264
Oryza glaberrima	33164	18	1459	38717
Oryza brachyantha	32037	1	17	2100
Oryza barthii	41595	326	1355	665
Brassica rapa	41024	0	1	1827
Brassica oleracea	59219	5	0	0
Brassica napus	101040	0	0	0
Arabidopsis thaliana	48321	3916	1	1775
Arabidopsis lyrata	32663	0	0	0

Percentage of different transcript biotypes



Work flow


EnsemblPlants
release-33
cDNA sequence
DNA sequence
(Soft masked)
Multiple whole genome alignment
(Lastz/ Multiz/ TBA)
Coding potential
assessment
(CPC)
Coding potential
assessment
(RNAcode)
Identification of conserved
local structures
(RNAz)
Re-annotation of
transcript biotypes
Identification of
cisNAT pairs

Coding potential Calculator (CPC)

Quality of the ORF

- Implements *framefinder* software to identify the longest reading frame in the three forward frames
- Accesses the integrity of the ORF by checking if the ORF begins with a start codon and ends with an in-frame stop codon

BLASTX against uniref

- Extracts the number of hits
- Quality of HSPs (Lower E-values)
- Distribution of HSPs with in the frame

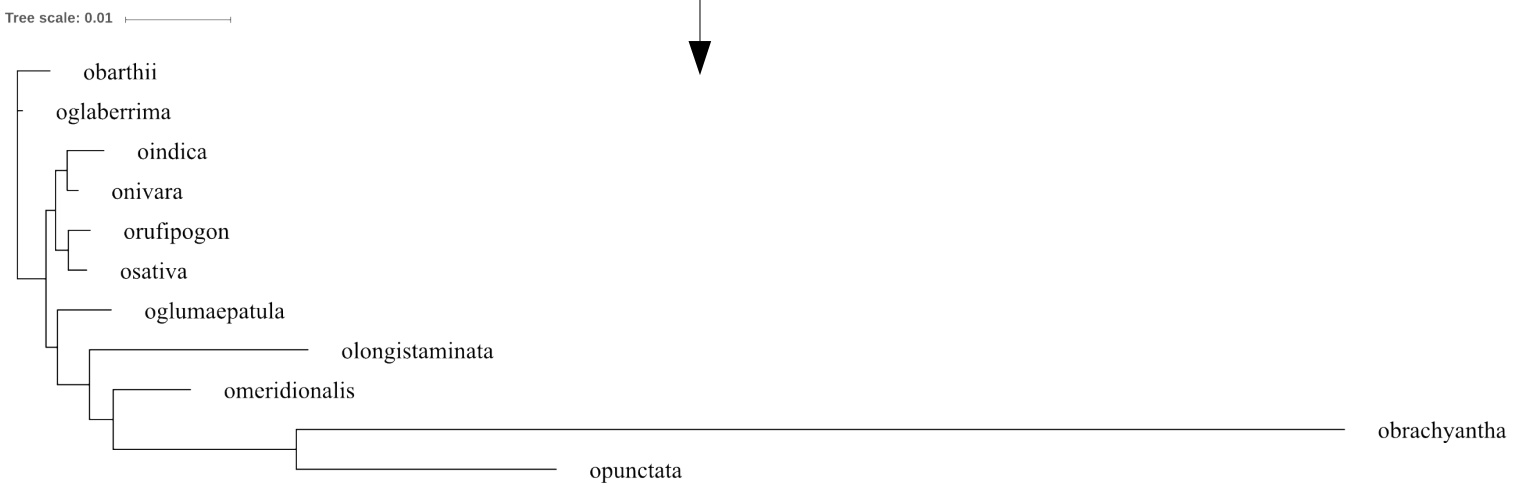
Support vector machine (SVM) machine learning classifier

- Transcripts with Score <1.5 is classified as noncoding

CPC : Kong et al., 2007; framefinder: Slater et al., 2000

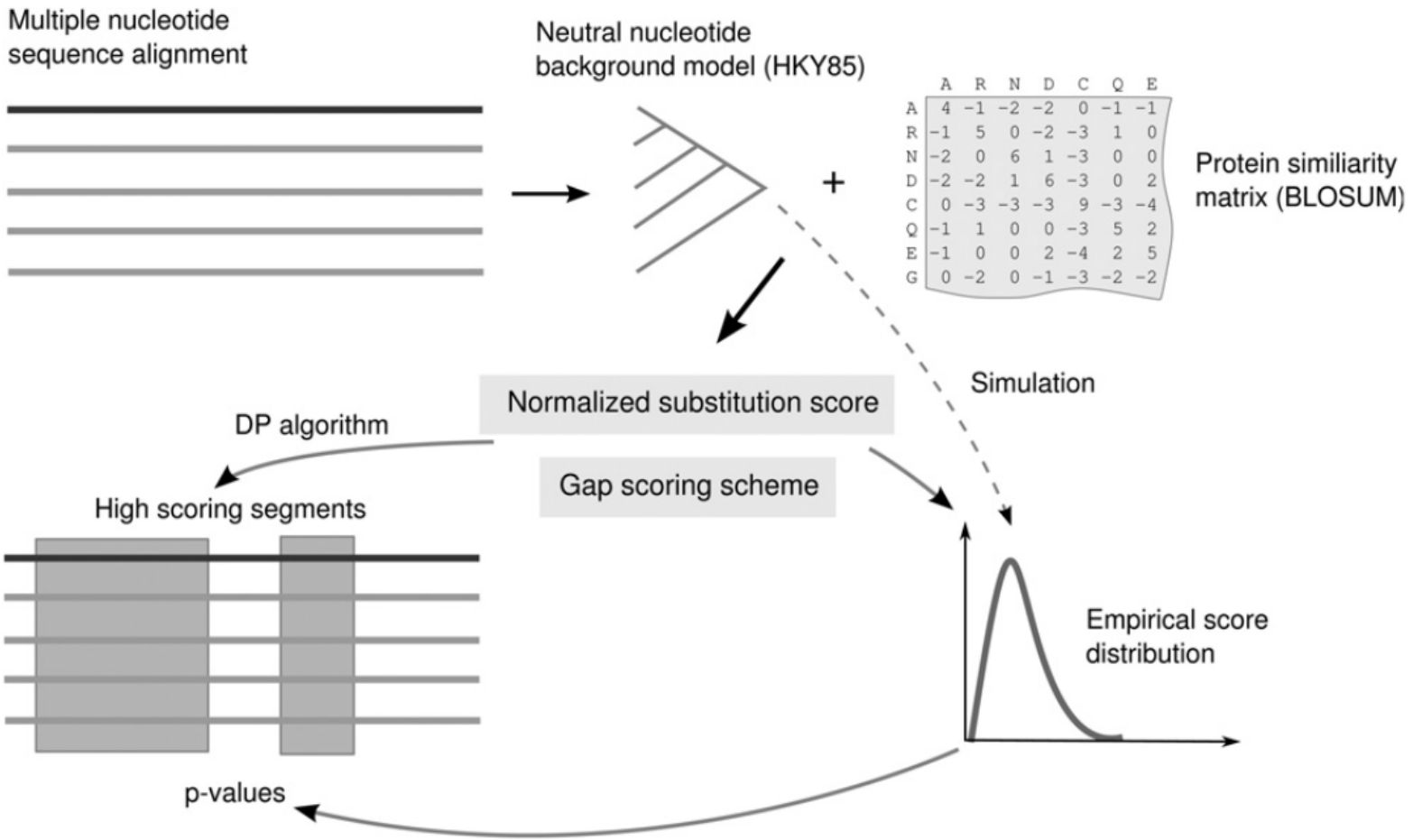
Multiple whole genome alignment

orthagogue → 100 randomly selected **one-to-one common ortholog gene** cDNA sequences → **Mafft / trimAl** → Guide tree generation **IQ-TREE** (v 1.4.4)

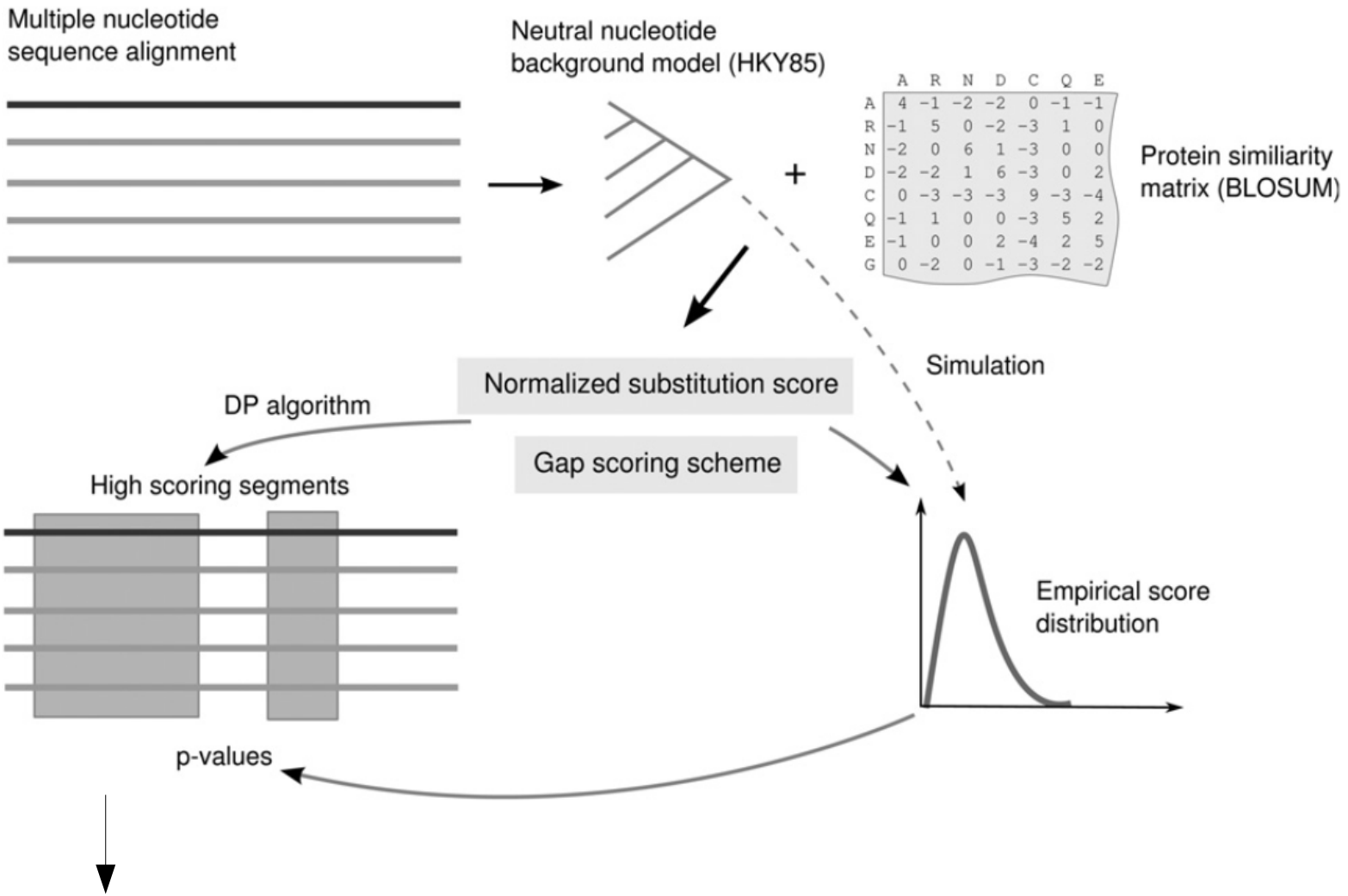


Pairwise genome alignments **Lastz** → generating the multiple alignment **Multiz/TBA** → “projecting” the alignment onto a reference sequence **maf_project**

RNAcode

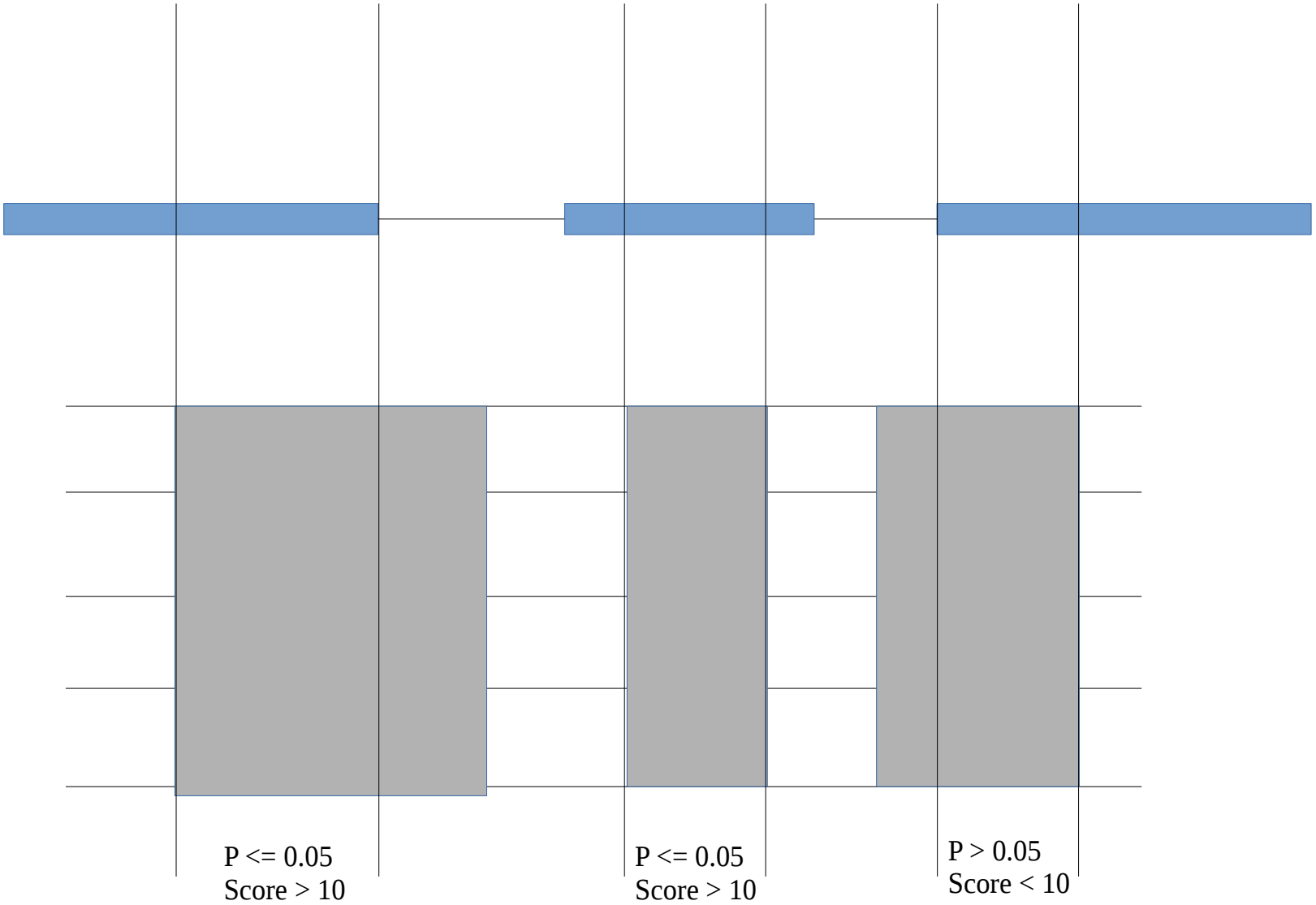

Multiple nucleotide sequence alignment
Neutral nucleotide background model (HKY85)
+
Protein similiarity matrix (BLOSUM)
Simulation
Normalized substitution score
DP algorithm
Gap scoring scheme
High scoring segments
Empirical score distribution
p-values

RNAcode

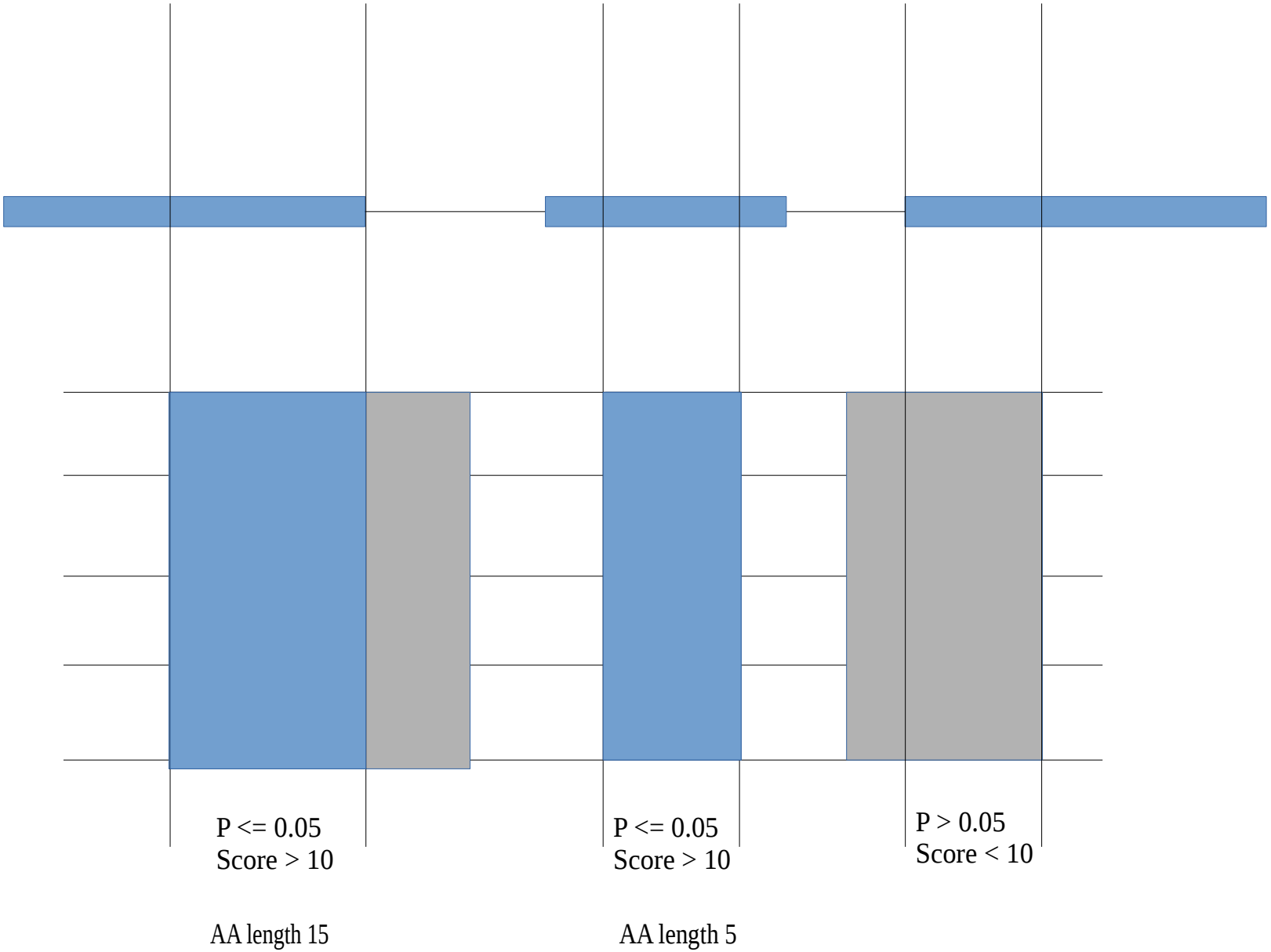


Map against transcript cds/exons

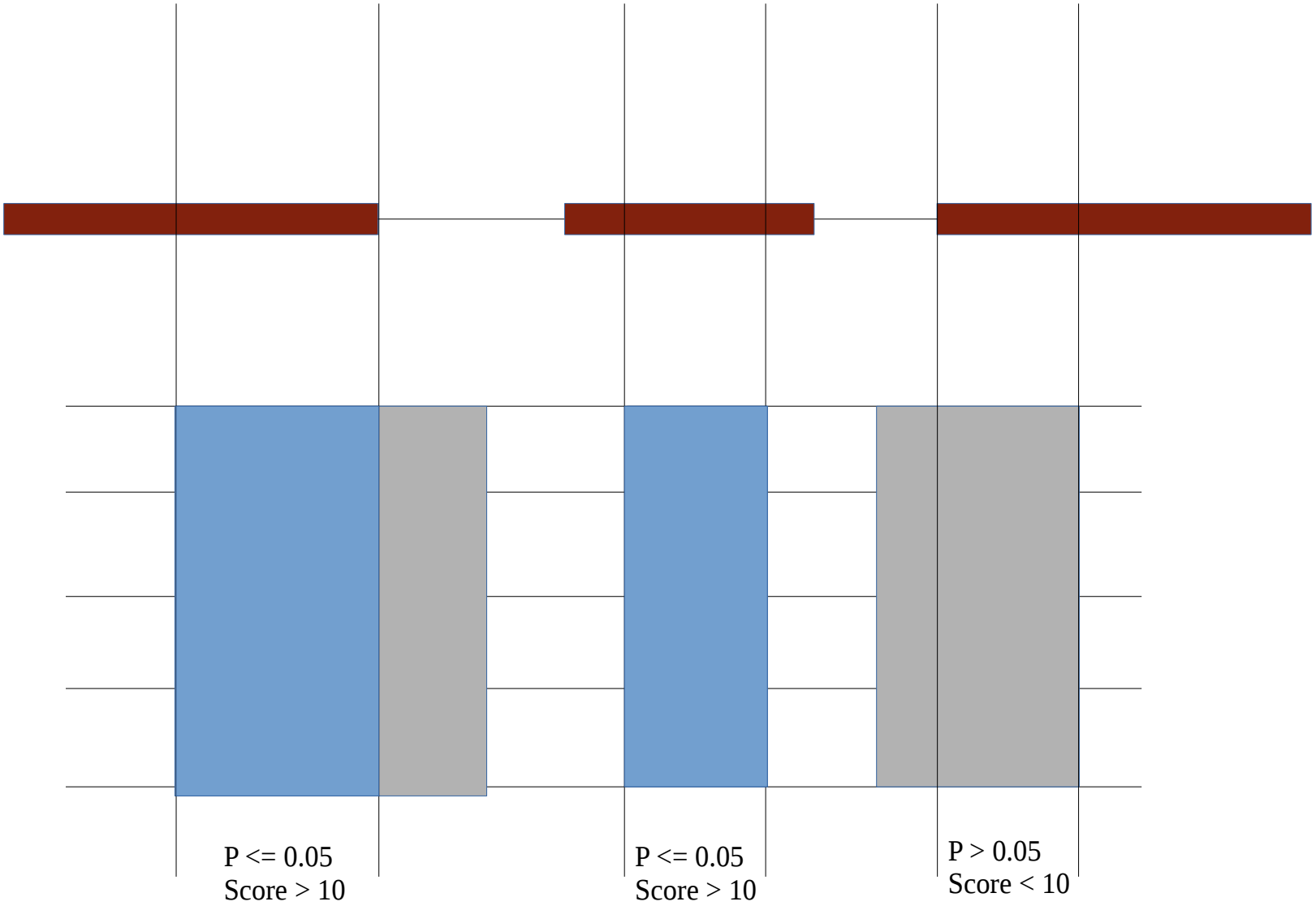
RNAcode



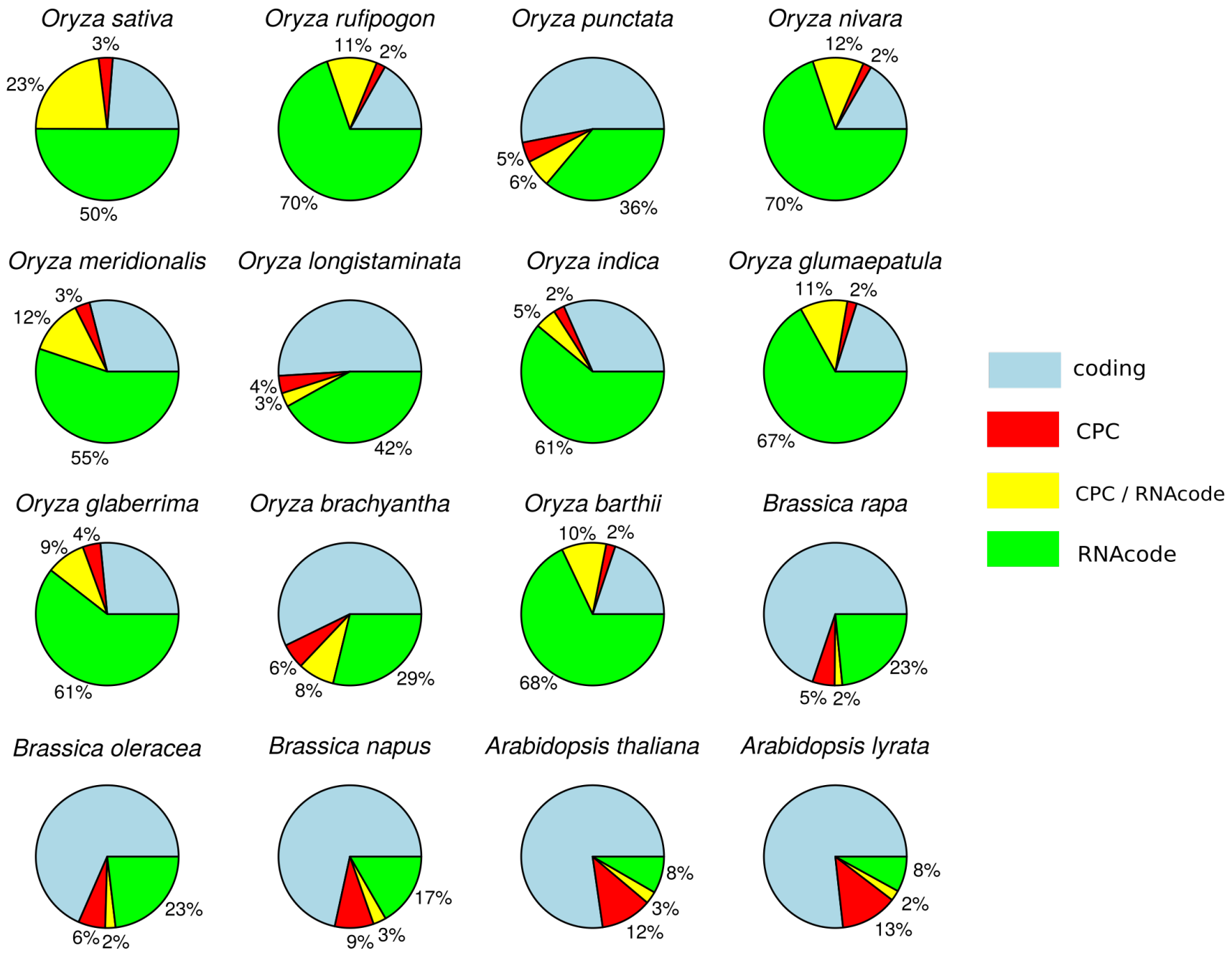
RNAcode


P <= 0.05
Score > 10
P <= 0.05
Score > 10
P > 0.05
Score < 10
AA length 15
AA length 5

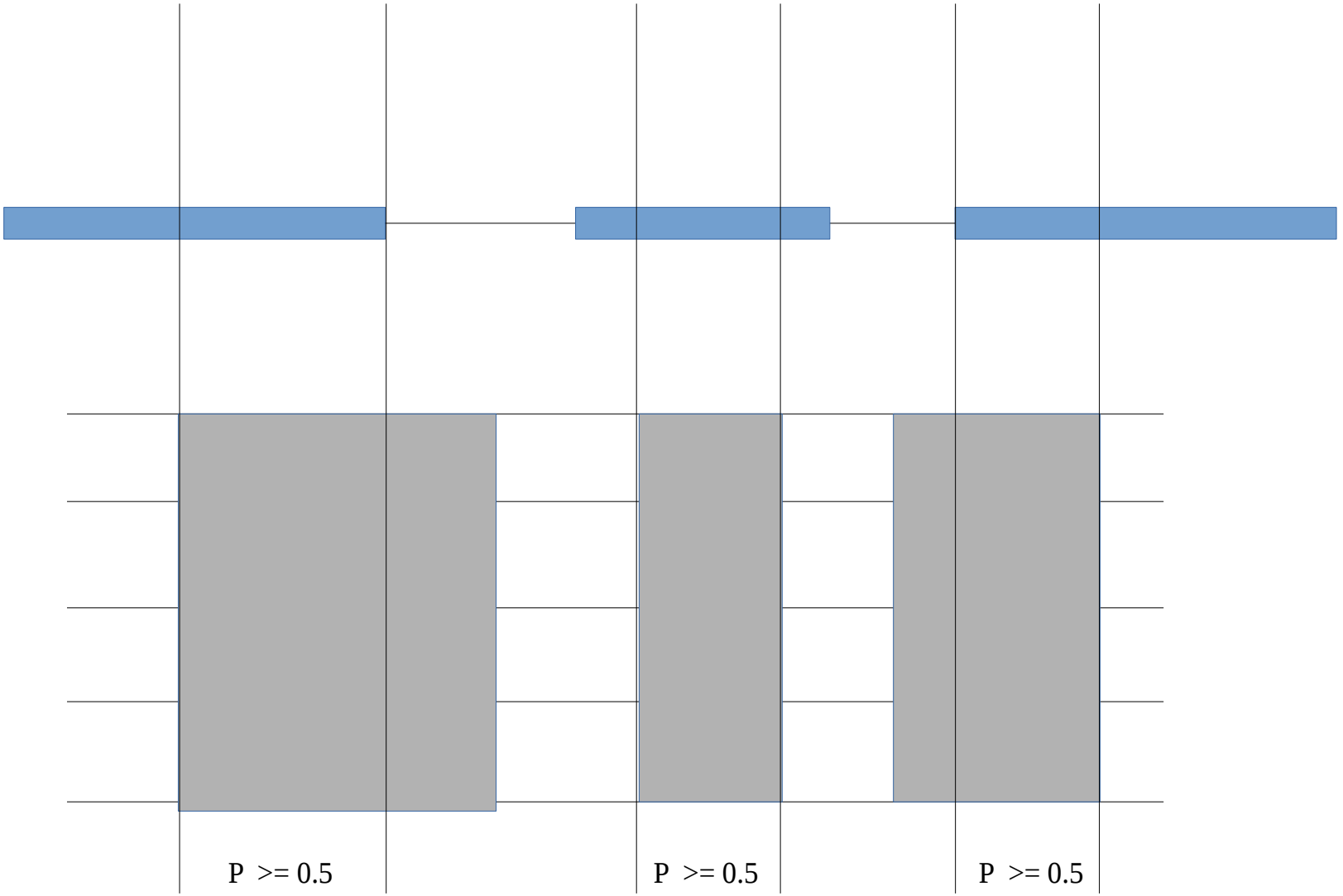
RNAcode



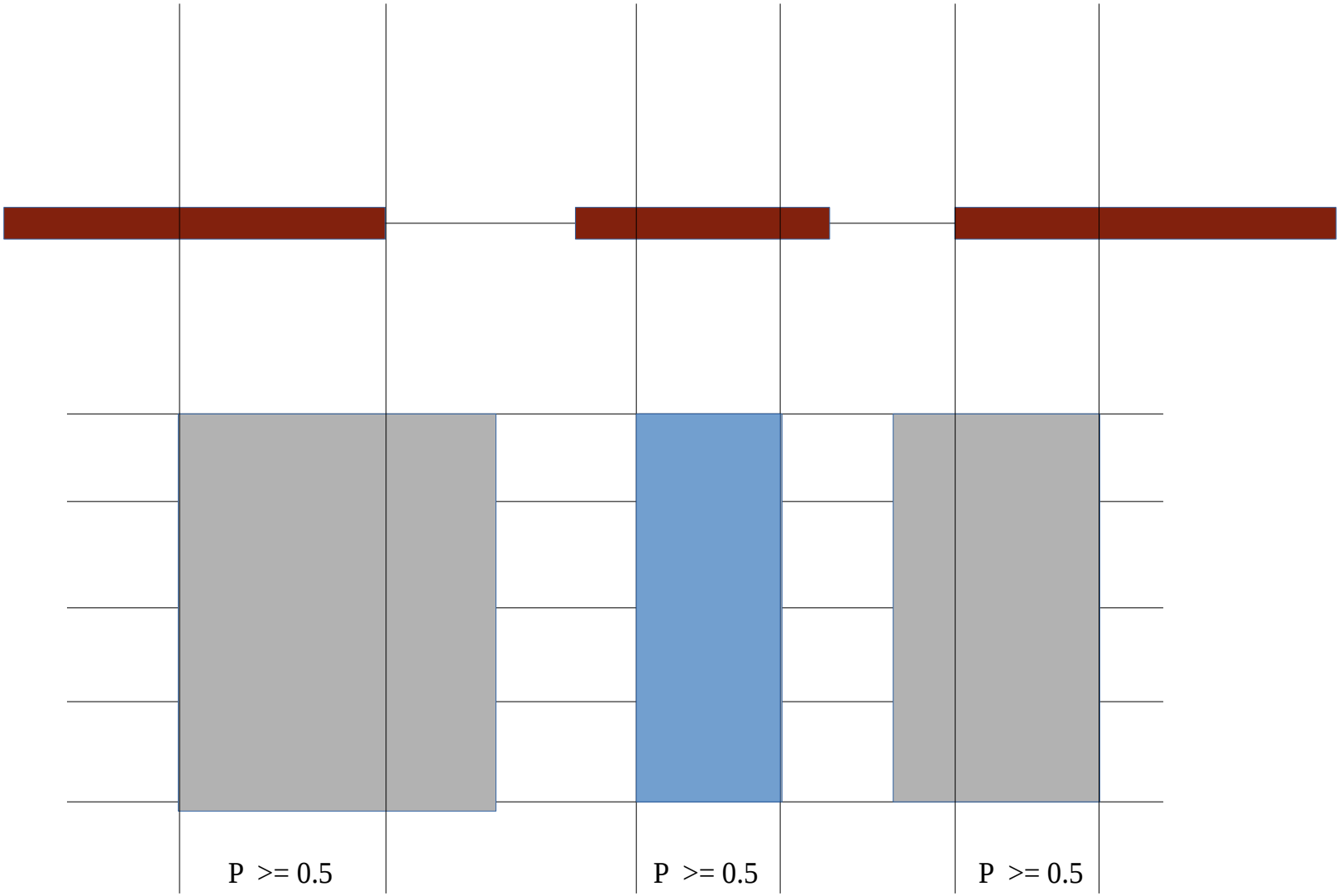
Coding potentials identified by RNAcode vs CPC



RNAz



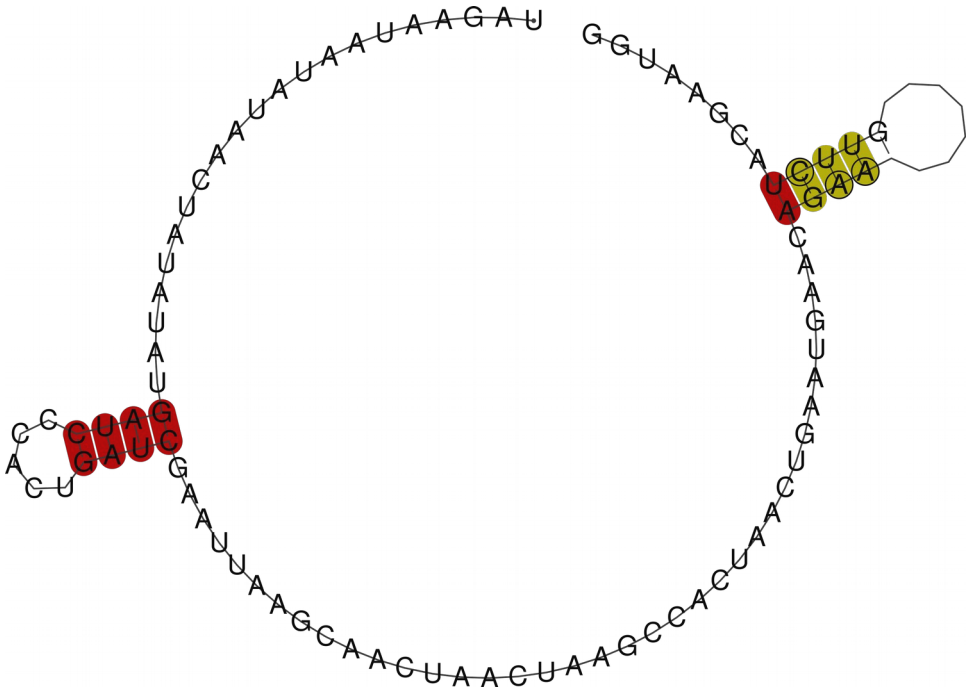
RNAz



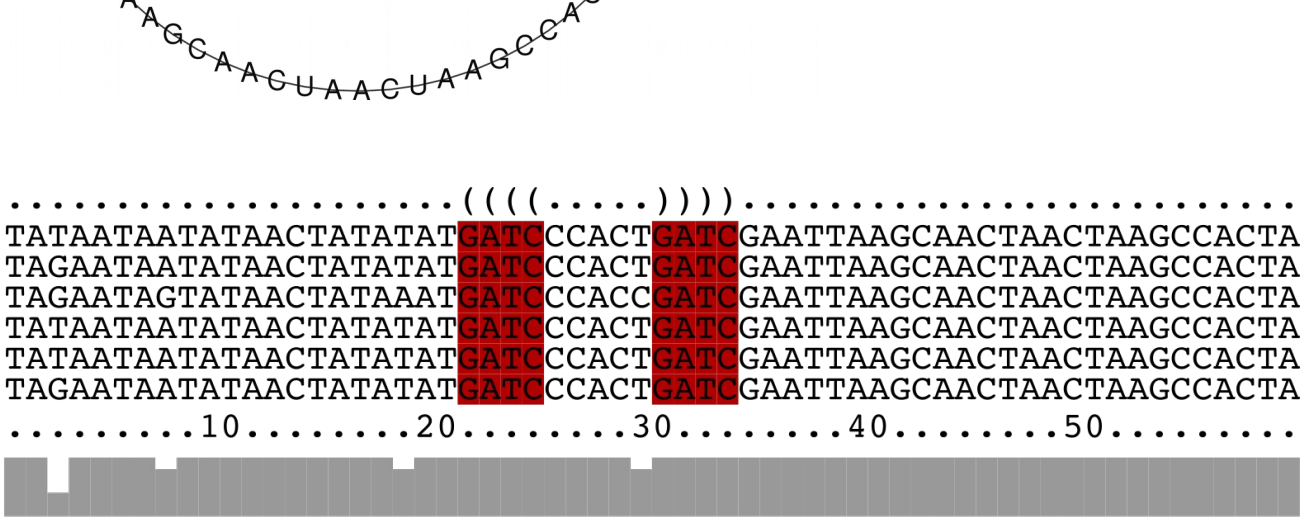
100 % overlap

No. of identified lncRNA transcripts with conserved local structure

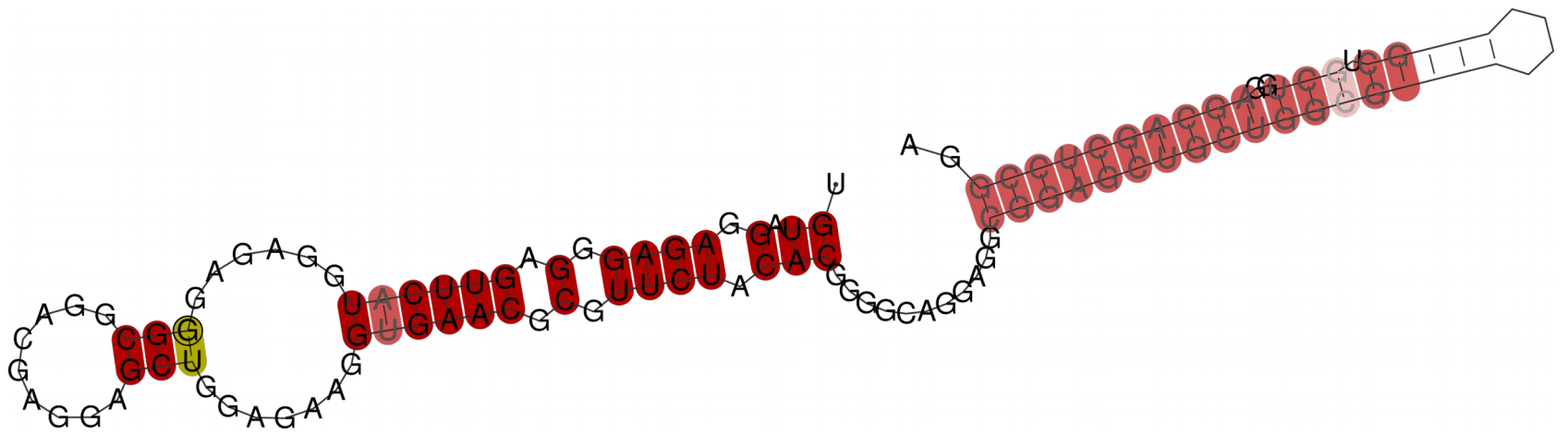
Species	Protein coding	RNAcode classified lncRNA	RNAz
Oryza sativa	42132	30779	14835
Oryza rufipogon	47441	33111	13690
Oryza punctata	41060	14828	3807
Oryza nivara	48360	33786	14531
Oryza meridionalis	43455	23971	9995
Oryza longistaminata	31686	13276	4257
Oryza indica	40745	24887	9102
Oryza glumaepatula	46893	31398	12618
Oryza glaberrima	33164	20073	6912
Oryza brachyantha	32037	9224	1258
Oryza barthii	41595	28263	10675
Brassica rapa	41024	9604	42
Brassica oleracea	59219	13690	46
Brassica napus	101040	16841	87
Arabidopsis thaliana	48321	4033	131
Arabidopsis lyrata	32663	2617	104



```
                                    .......................((((.....))))..........................
osativa.chr2                        TATAATAATATAACTATATATGATCCCACTGATCGAATTAAGCAACTAACTAAGCCACTA 60
omeridionalis.chr2                  TAGAATAATATAACTATATATGATCCCACTGATCGAATTAAGCAACTAACTAAGCCACTA 60
opunctata.chr2                      TAGAATAGTATAACTATAAATGATCCCACCGATCGAATTAAGCAACTAACTAAGCCACTA 60
oglumaepatula.chr2                  TATAATAATATAACTATATATGATCCCACTGATCGAATTAAGCAACTAACTAAGCCACTA 60
oglaberrima.chr2                    TATAATAATATAACTATATATGATCCCACTGATCGAATTAAGCAACTAACTAAGCCACTA 60
olongistaminata.scaKN540772         TAGAATAATATAACTATATATGATCCCACTGATCGAATTAAGCAACTAACTAAGCCACTA 60
                                    .........10........20........30........40........50.........
```



```
                                    ...........((((.......))))).........
osativa.chr2                        ACTGAATGAACAGGACACAAACAGTTCTACGAATGG 96
omeridionalis.chr2                  ACTGAATGAACAGAGCACAAACAGTTCTACGAATGG 96
opunctata.chr2                      ACTGAATGAACAGAACACAAACAGTTTTAAAAATAG 96
oglumaepatula.chr2                  ACTGAATGAACAGGACACAAACAGTTCTACGAATGG 96
oglaberrima.chr2                    ACTGAATGAACAGGACACAAACAGTTCTACGAATGG 96
olongistaminata.scaKN540772         ACTGAATGAACAGAGCACAAACAGTTCTACAAACTG 96
                                    .........70........80........90.....
```



```
                                  .((.(.((((.(.((((((.....(((.........))).......)))))).).)))).
osativa.chr2                      TGTAGGAGAGGGAGTTCATGGAGAGGGCGGACGAGGAGCTGGAGAAGGTGAACGCGTTCT 60
obrachyantha.chr2                 GGTAGGAGAGGGAGTTCTTGGAGCGAGCGGACGAGGAGCTGGAGAAGGTGAACGCGTTCT 60
opunctata.chr2                    TGTAGGAGAGGGAGTTCATGGAGCGGGCGGACGAGGAGCTGGAGAAGGTGAACGCGTTCT 60
oglumaepatula.chr2                TGTAGGAGAGGGAGTTCATGGAGAGGGCAGACGAGGAGCTGGAGAAGGTGAACGCGTTCT 60
oglaberrima.chr2                  TGTAGGAGAGGGAGTTCATGGAGAGGGCGGACGAGGAGCTGGAGAAGGTGAACGCGTTCT 60
olongistaminata.scaKN540772       TGTAGGAGAGGGAGTTCATGGAGAGGGCGGACGAGGAGCTGGAGAAGGTGAACGCGTTCT 60
                                  .........10........20........30........40........50.........

                                  .)))..........(((((((((((((((....)))))).))))..)))))))))..
osativa.chr2                      ACACGGGGCAGGAGGCGGAGCTGCTGGCGCGCGGCGACGCGCTGCTGGAGCAGCTCCGGA 120
obrachyantha.chr2                 ACACGGGGCAGGAGGCGGAGCTGCTGGCGCGCGGCGAGGCGCTGCTGGAGCAGCTCCGGA 120
opunctata.chr2                    ACACGGGGCAGGAAGCGGAGCTGCTGGCGCGCGGCGACGCGCTCCTGGAGCAGCTCCGGA 120
oglumaepatula.chr2                ACACGGGGCAGGAGGCGGAGCTGCTGGCGCGCGGCGACGCGCTGCTGGAGCAGCTCCGGA 120
oglaberrima.chr2                  ACACGGGGCAGGAGGCGGAGCTGCTGGCGCGCGGCGACGCGCTGCTGGAGCAGCTCCGGA 120
olongistaminata.scaKN540772       ACACGGGGCAGGAGGCGGAGCTGCTGGCGCGCGGCGACGC------------------- 100
                                  .........70........80........90........100.......110........

```

Total number of identified cisNAT pairs

Species	RNAcode based	Annotation based
Oryza sativa	970	445
Oryza rufipogon	293	1
Oryza punctata	329	0
Oryza nivara	384	0
Oryza meridionalis	622	18
Oryza longistaminata	0	0
Oryza indica	26	3
Oryza glumaepatula	420	0
Oryza glaberrima	7	7
Oryza brachyantha	346	1
Oryza barthii	162	1
Brassica rapa	4	0
Brassica oleracea	0	0
Brassica napus	0	0
Arabidopsis thaliana	839	552
Arabidopsis lyrata	6	0

Future work

- ✔ **Identification of true protein coding and non-coding transcripts**
- ✔ **Genome wide screen for local conserved structures**
- ✔ **Identification of overlapping features**
- **Re-annotation of transcript biotypes considering the results from CPC, RNAcode and RNAz**
- **Enrichment analysis of local conserved structures in cisNATs**
- **Application of GraphClust / RNAscClust on the candidate cisNATs with conserved local structures for grouping them together for genome wide screen for similar regions**
- **Selection of candidate cisNATs for experimental testing**
- **Annotation of missing protein-coding genes**

Acknowledgments

Ivo Hofacker
Andrea Tanzer

Bernhard Thiel
Roman Ochsenreiter
Richard Neuboeck

Uni Lausanne, Switzerland

Yves Poirier
Jules Deforges
Rodrigo Siqueira Reis

ETH Zurich, Switzerland

Wilhelm Gruissem
Katja Baerenfaller
Julia Svozil

Funding

No. of annotated lncRNA transcripts classified as coding

Species	Annotated ncRNA	CPC Score >= 1.5	RNAcode Score>10; pvalue<0.05
Oryza sativa	55619	65	337
Oryza rufipogon	2020	12	34
Oryza punctata	28	0	0
Oryza nivara	24	0	0
Oryza meridionalis	2283	7	16
Oryza longistaminata	0	0	0
Oryza indica	2232	5	23
Oryza glumaepatula	19	0	0
Oryza glaberrima	1997	6	18
Oryza brachyantha	18	1	0
Oryza barthii	1681	3	15
Brassica rapa	1	0	0
Brassica oleracea	5	3	1
Brassica napus	0	0	0
Arabidopsis thaliana	3917	61	660
Arabidopsis lyrata	0	0	0