Rfam Clans

Steffen Heyne, Sebastian Will,
Rolf Backofen and Paul Gardner

Bled, 19.02.2010

Bioinformatics Group, University of Freiburg

Databases of Sequence Families

Proteins: **Pfam**

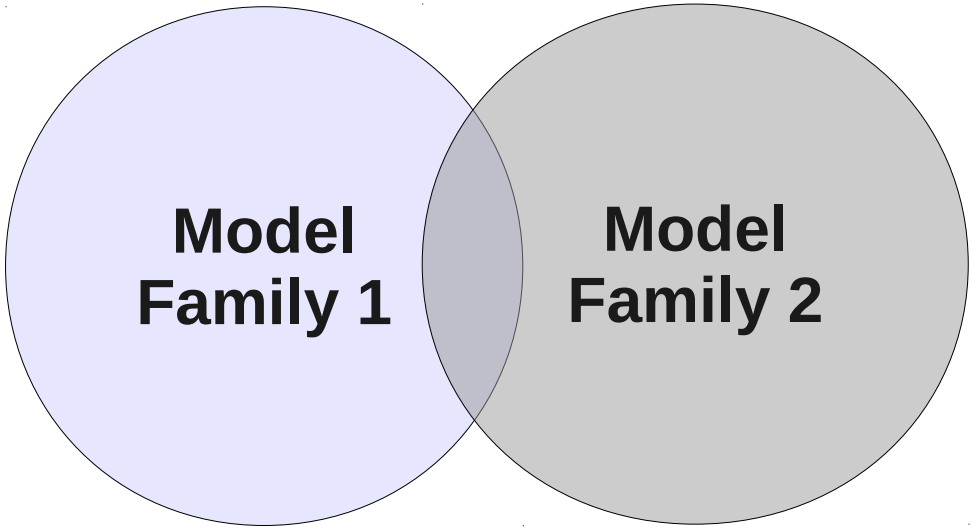
- DB of protein families/Domains
- Pfam 24.0: 11912 families
- HMM models

RNAs: **Rfam**

- DB of RNA families
- Rfam 9.1: 1372 families
- Covariance models/ Infernal

One Family, One Model

- Fundamental “law”: new families are not allowed to overlap with existing families
  → Some families have artifically high tresholds to avoid overlapping
  → For some divergent families a single model is not possible
  → search not recovers full training set

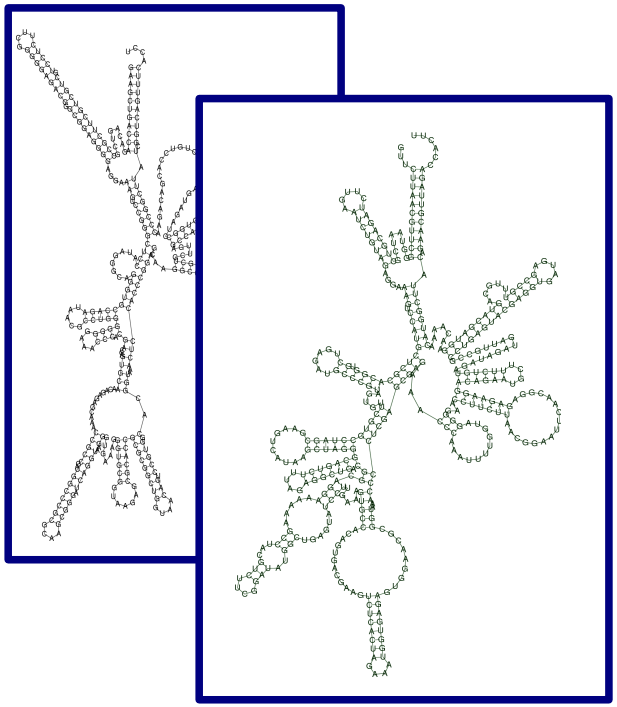


Related Families: Clan

Pfam Clans

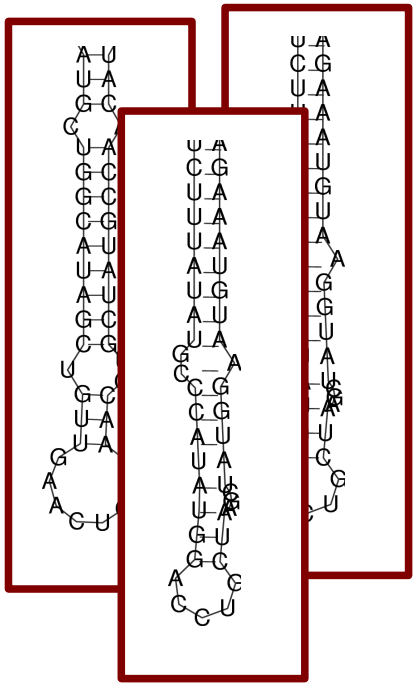
- independent pieces of evidence to assess related families: related structure, related sequence, related function, significant matching of the same sequence to HMMs from different families, profile-profile comparsions
- Significance: E-value < 0.001
- In Pfam introduced in 2005 (Finn et. *al*, NAR 2006)

Related RNA families



- RNaseP families

- small nuclear RNAs (H/ACA, CD-box)
- IRES
- CRISPR
- ...



- micro RNAs

Clans

Pfam Clans

New: Rfam Clans

- Hirarchical view of Pfam families
- “clan is a collection of families that have arisen from a single evolutionary origin”

?

Clans

Pfam Clans

- Hirarchical view of Pfam families
- “clan is a collection of families that have arisen from a single evolutionary origin”

New: Rfam Clans

- Hirarchical view of RNA families
- Related sequence, structure and function

Rfam Clans

Motivation

- Can we adapt the clan concept to RNAs?
- What can we learn from RNA clans?
- What is an appropriate method to detect RNA clans?

Clans

Pfam Clans

- Hirarchical view of families
- clan is a collection of families that have arisen from a single evolutionary origin
- ~420 clans in Pfam

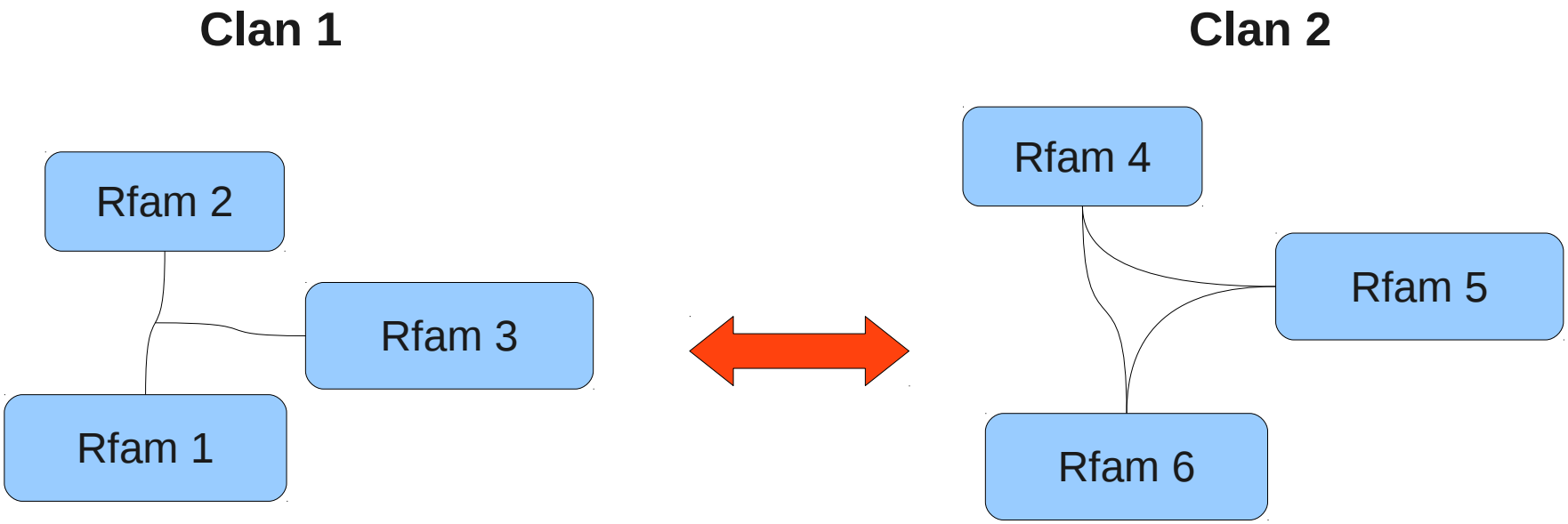
New: Rfam Clans

- Manually created set of ~110 clans, provided by Paul Gardner
- Can we recover them with our methods?

Rfam Clans

- General: identify related RNAs with sequence-structure alignments
- Idea: use LocARNA-P (new probabilistic version) with reliability scores
- LocARNA-P: compute match probabilities of alignment edges via partition function approach → more accurate alignments
- Clan identification: clustering problem!?

Rfam Clans:
Clustering of Rfam Families

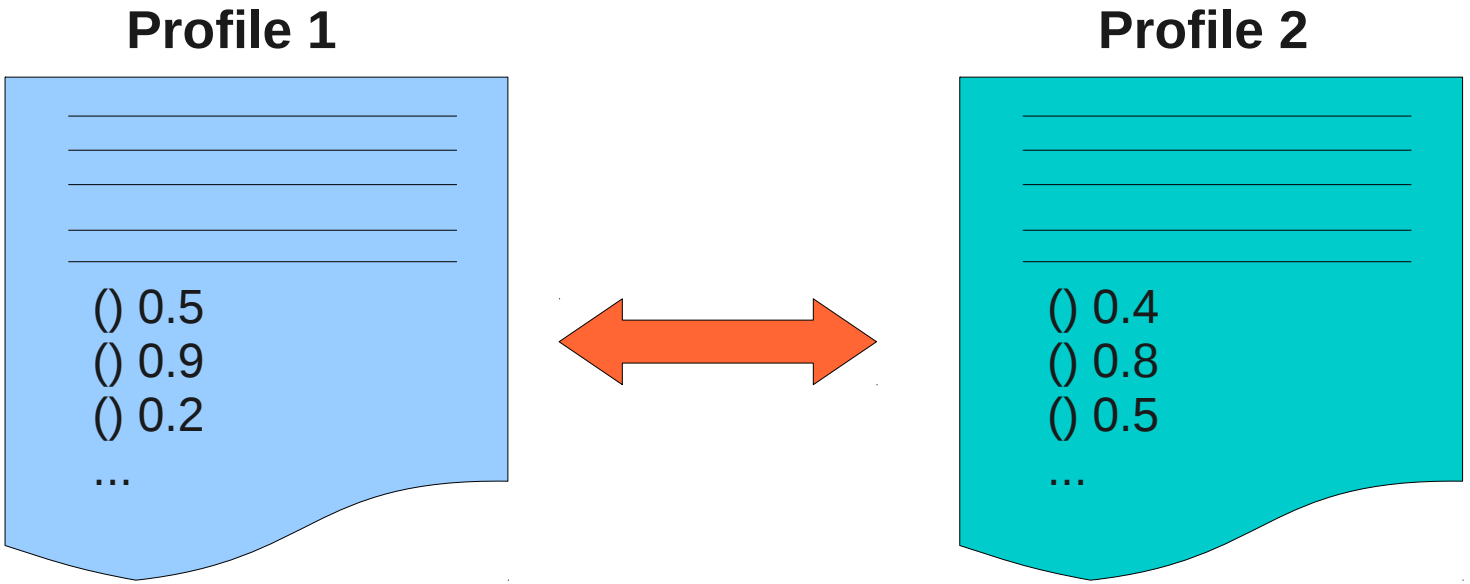


Clustering of RNA families

- all-to-all pairwise comparsion
- Problem: a single RNA family is MSA
- Need: “RNA Family Profile”
- Approach for consensus dot-plot:
    1. Take Rfam seed alignment of a family
    2. Eliminate columns with >50% gaps
    3. Reduce #sequences <500
    4. RNAalifold with consensus structure of seed

Clustering of RNA families

- LocARNA-P align two “Rfam profiles” to get pairwise scores



- Assume manually created clans as clusters --> evaluation

Manually Created Rfam Clans

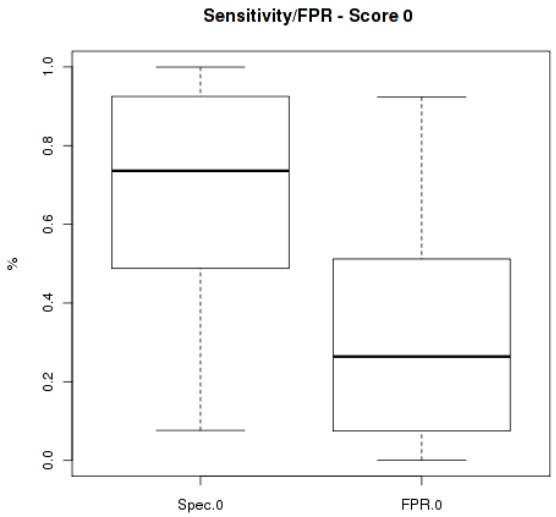
7SK_clan	RF00100,RF01052
CRISPR-1_clan	RF01315,RF01317,RF01325,RF01327,RF01338
CRISPR-2_clan	RF01320,RF01377
FinP-traJ_clan	RF00107,RF00243
Glm_clan	RF00083,RF00128
GP_knot_clan	RF01073,RF01092
Hammerhead_clan	RF00008,RF00163
IRES1_clan	RF00061,RF00209
mir-105_clan	RF00670,RF01033
mir-137_clan	RF00694,RF00859
mir-15_clan	RF00254,RF00455
mir-182_clan	RF00663,RF00702,RF00706,RF00843
mir-2_clan	RF00047,RF00143,RF00813,RF00844,RF00854
mir-216_clan	RF00654,RF00747
mir-279_clan	RF00754,RF00948
mir-28_clan	RF00655,RF00917
mir-290_clan	RF00639,RF00665,RF00668,RF01413
mir-3_clan	RF00716,RF00818
mir-34_clan	RF00456,RF00711
mir-36_clan	RF00685,RF00794
mir-50_clan	RF00672,RF00824
mir-640_clan	RF00985,RF01042
mir-73_clan	RF00830,RF00831
mir-81_clan	RF00727,RF00728
mir-BART_clan	RF00363,RF00866
MIR169_clan	RF00645,RF00865
MIR171_clan	RF00643,RF00692
MIR806_clan	RF01058,RF01062
PK-rep_clan	RF01087,RF01089
RF_site_clan	RF01074,RF01079

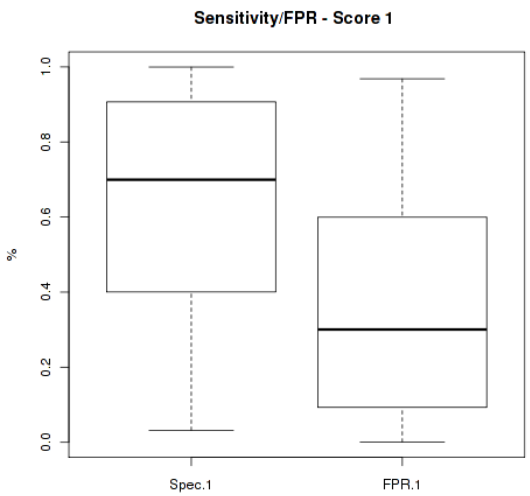
Evaluation

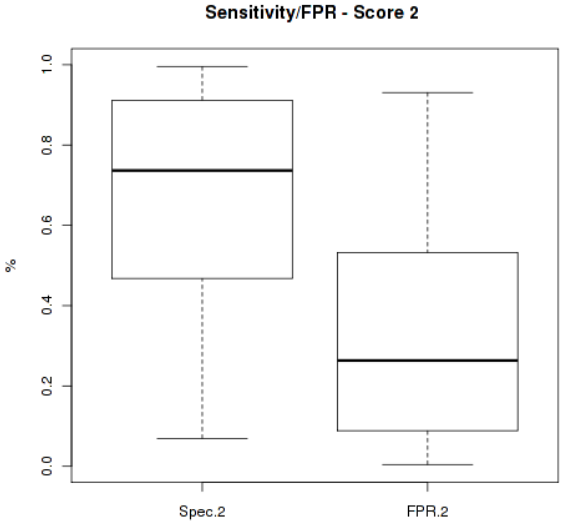
Initial Test: Single Linkage Clustering

- Compute pairwise scores “intra-clan”
- Compute pairwise scores against Rfam
- Lowest intra-clan score is threshold to discriminate Rfam scores
- Compute specificity

Evaluation - Scores







Score 0:
normalized over alignment length

```
Median :0.7360
Mean   :0.6804
```

Score 1:
normalized over matched columns

```
Median :0.69949
Mean   :0.62438
```

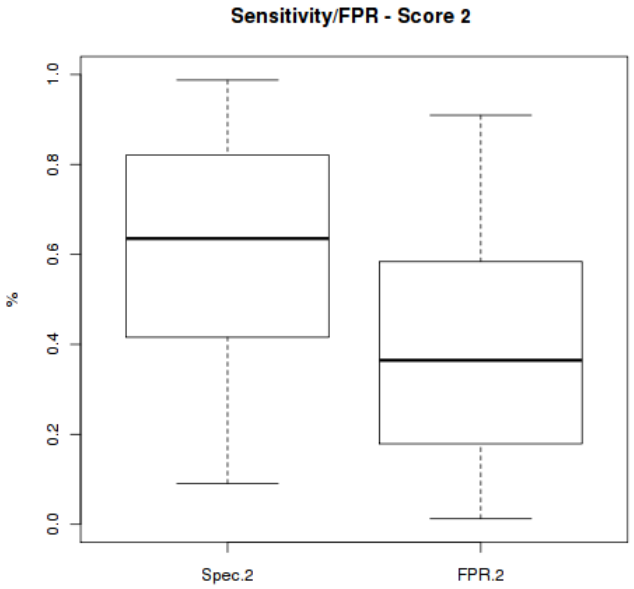
Score 2:
max. reliable structure

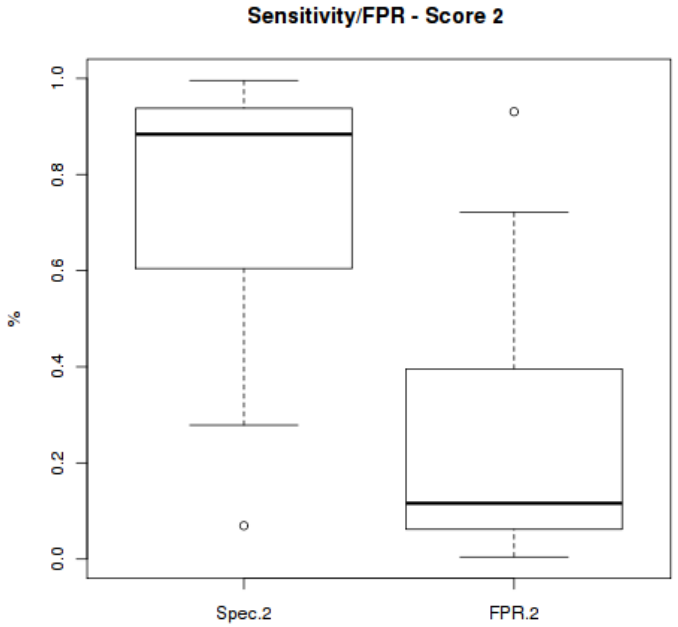
```
Median :0.73659
Mean   :0.66639
```

Evaluation

Clan	Spec-2	FPR-2	TP-2	FP-2	TN-2
CRISPR-2_clan	0,9956	0,0044	1	12	2730
IRES1_clan	0,9949	0,0051	1	14	2728
RF_site_clan	0,9916	0,0084	1	23	2719
SNORA30_clan	0,9876	0,0124	1	34	2708
SNORD16_clan	0,9865	0,0135	1	37	2705
SNORD18_clan	0,9840	0,0160	1	44	2698
SNORA17_clan	0,9832	0,0168	1	46	2696
snoU85_clan	0,9821	0,0179	1	49	2693
7SK_clan	0,9810	0,0190	1	52	2690
GP_knot_clan	0,9799	0,0201	1	55	2687
SNORD26_clan	0,9741	0,0259	1	71	2671
Hammerhead_clan	0,9734	0,0266	1	73	2669
mir-279_clan	0,9690	0,0310	1	85	2657
SNORA74_clan	0,9679	0,0321	1	88	2654
mir-3_clan	0,9679	0,0321	1	88	2654
SNORA20_clan	0,9668	0,0332	1	91	2651
SCARNA3_clan	0,9588	0,0412	1	113	2629
mir-BART_clan	0,9581	0,0419	1	115	2627
RNaseP_clan	0,9478	0,0522	10	357	6483
SNORA35_clan	0,9312	0,0688	6	377	5099
mir-15_clan	0,9282	0,0718	1	197	2545
mir-640_clan	0,9263	0,0737	1	202	2540
CRISPR-1_clan	0,9250	0,0750	10	513	6327
mir-137_clan	0,9227	0,0773	1	212	2530
mir-36_clan	0,9220	0,0780	1	214	2528
U2_clan	0,9212	0,0788	1	216	2526
SNORA5_clan	0,9205	0,0795	1	218	2524
mir-34_clan	0,9023	0,0977	1	268	2474
U1_clan	0,9005	0,0995	3	409	3701
mir-81_clan	0,8920	0,1080	1	296	2446
SNORA7_clan	0,8899	0,1101	1	302	2440
SL_clan	0,8877	0,1123	1	308	2434

Evaluation – clan subset





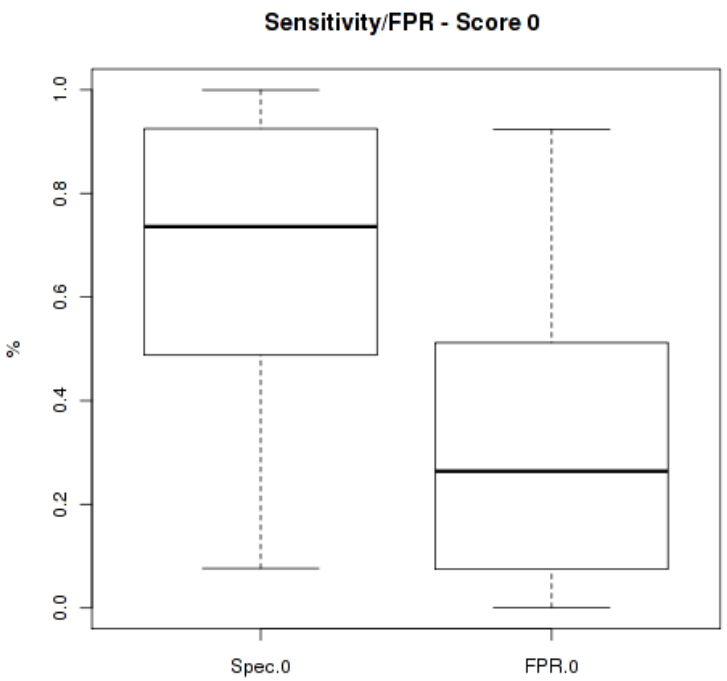
Score 2 – all sn|SN* clans (64)

```
Min.   :0.09051
1st Qu.:0.41639
Median :0.63535
Mean   :0.60106
3rd Qu.:0.81975
Max.   :0.98760
```

Score 2 – all without sn|SN* clans (44)

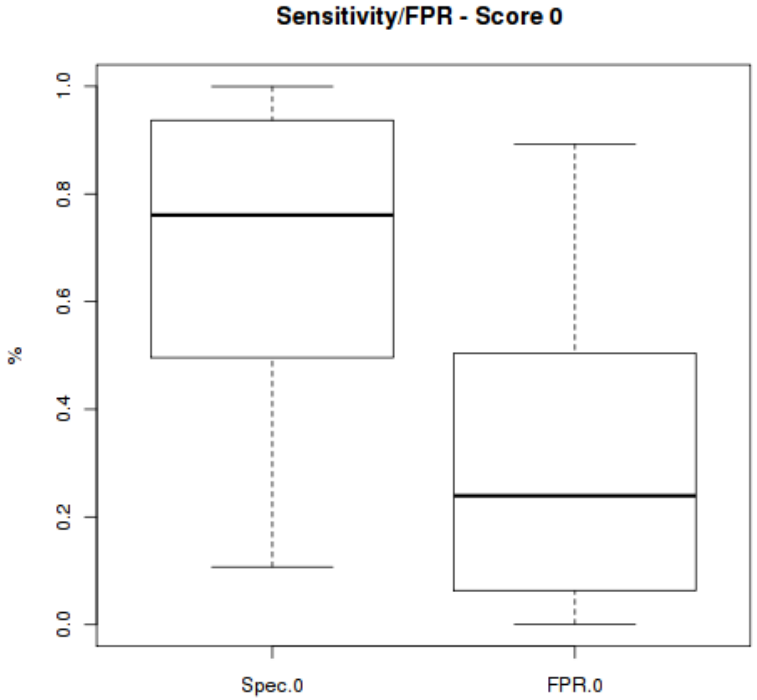
```
Min.   :0.06934
1st Qu.:0.62372
Median :0.88384
Mean   :0.76142
3rd Qu.:0.93307
Max.   :0.99562
```

Evaluation – scoring scheme



Score 0 - std

```
Min.   :0.0764
1st Qu.:0.4923
Median :0.7360
Mean   :0.6804
3rd Qu.:0.9231
Max.   :0.9996
```



Score 0

--indel-open -400 --indel -250 --struct-weight 180

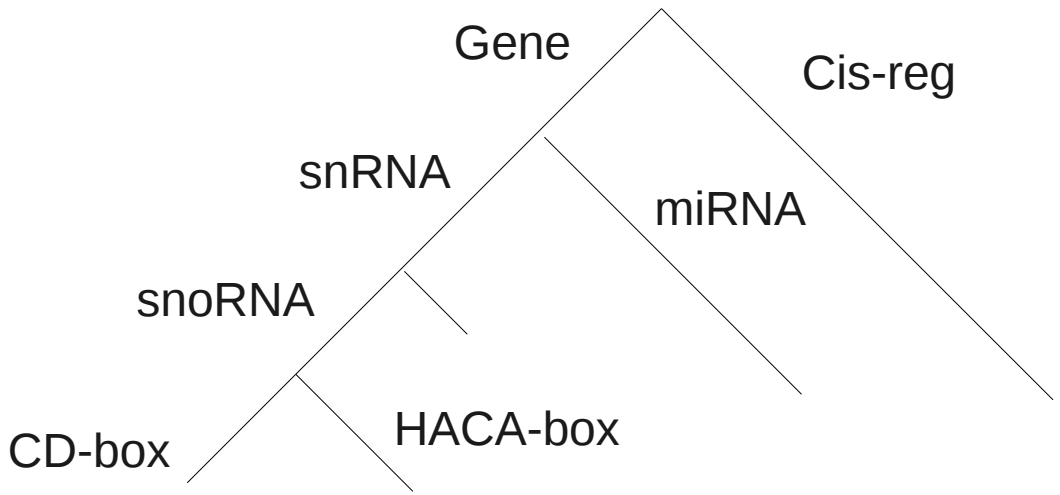
```
Min.   :0.1071
1st Qu.:0.4960
Median :0.7608
Mean   :0.6928
3rd Qu.:0.9367
Max.   :0.999
```

Summary and Outlook

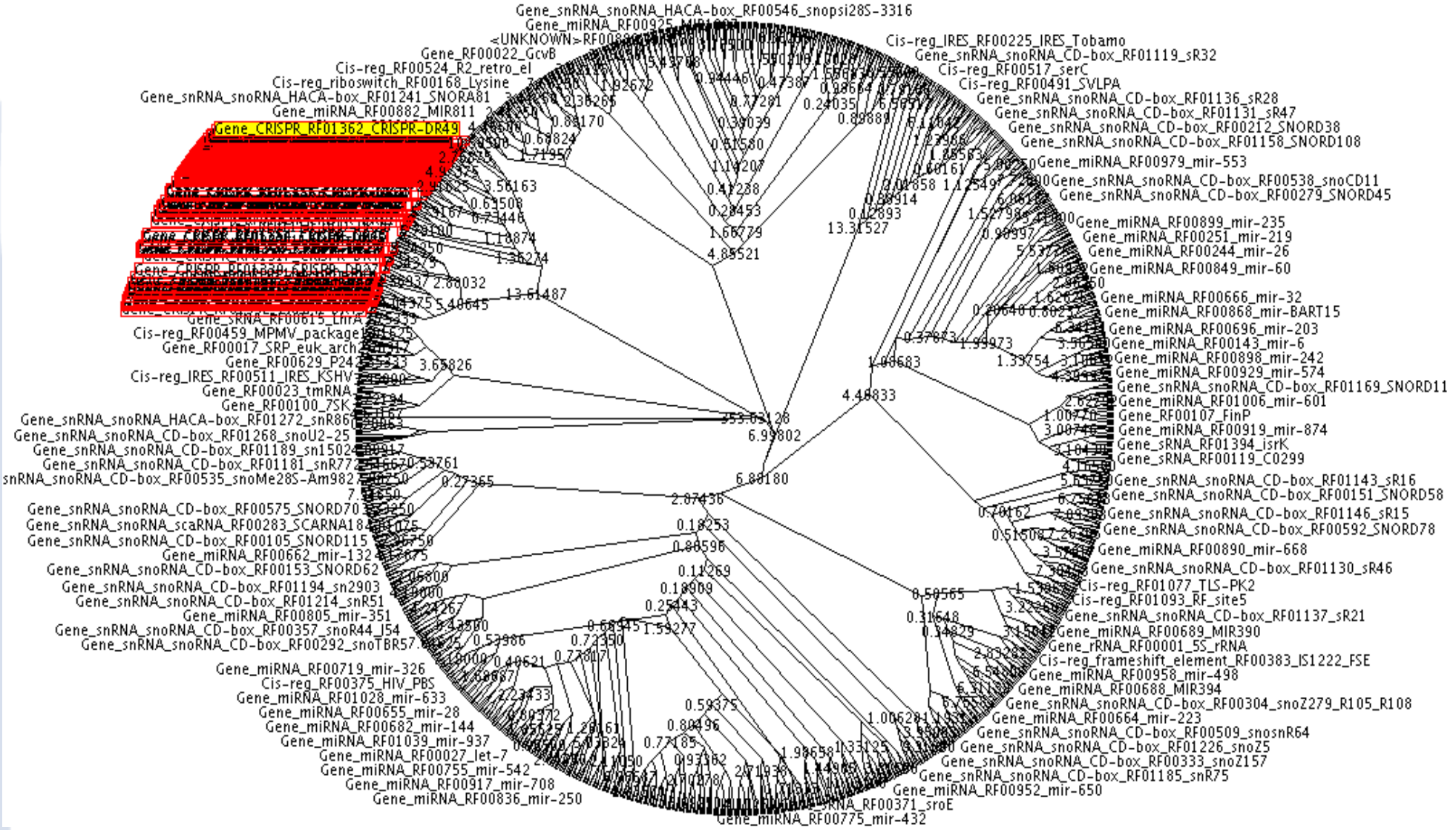
- Specificity ~76% for all provided clans currently
- ongoing: PGMA tree for clan evaluation
- SVM to learn clans (combine features of CM and alignment)
- Evaluation against artificial/random data
  random alignments – SISSIz...
- Bugs in Bioperl fixed :-)

Outlook

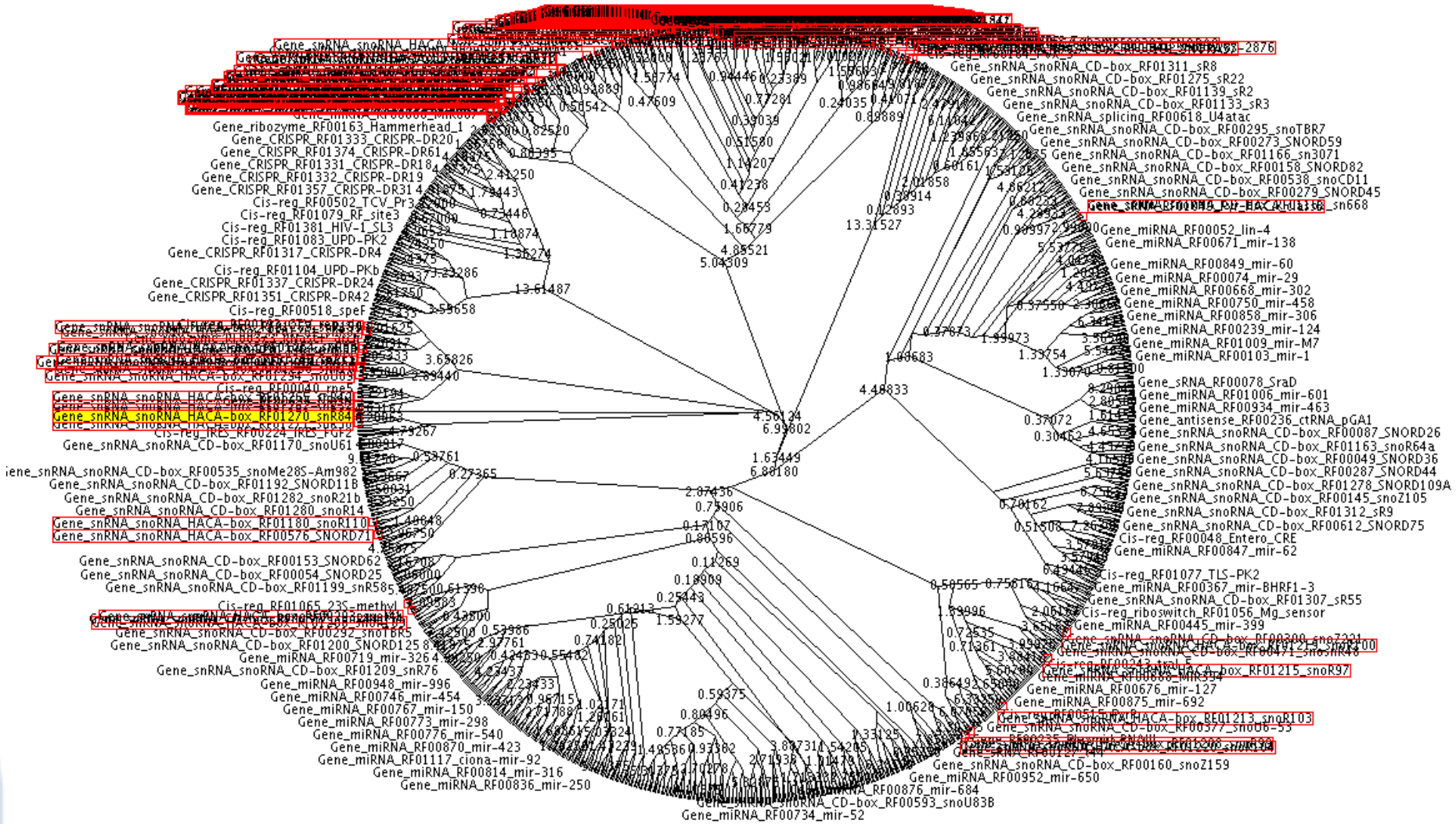
- Use Rfam “types” for evaluation:
  - Gene;snRNA;snoRNA;HACA-box
  - Gene;snRNA;snoRNA;CD-box
  - Gene;miRNA
  - Cis-reg;riboswitch ...



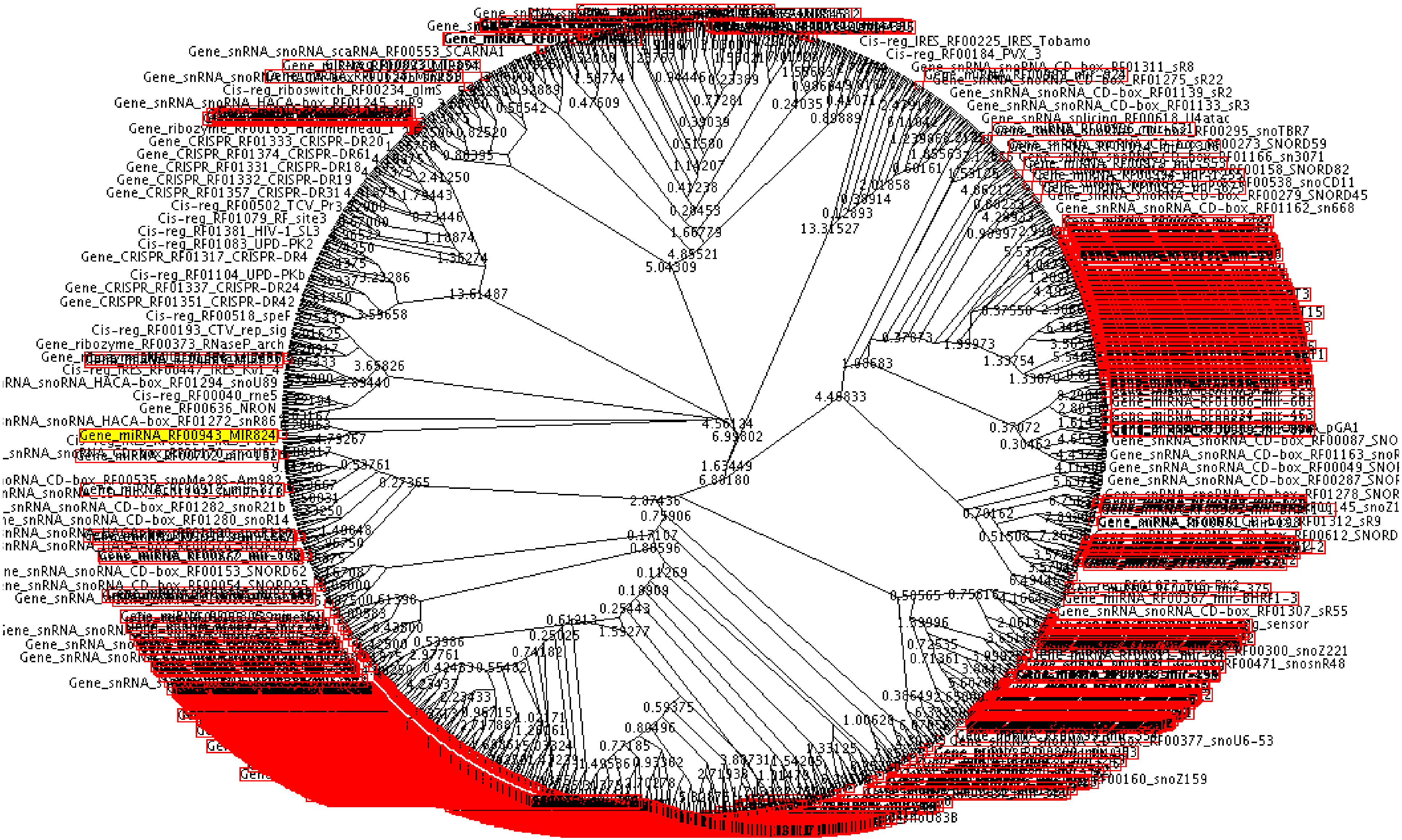
Rfam type: CRISPR



Rfam Type: HACA-box



Rfam Type: miRNA



Acknowledgement

Sebastian Will
Rolf Backofen
Paul Gardner

Freiburg Bioinformatics Group

Thank you!

...and the Bled 2010 Team