Prediction of structured RNAs: Lessons from the ENCODE pilote project

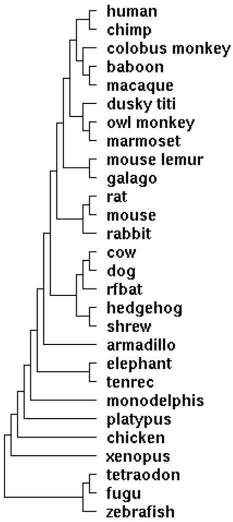
Stefan Washietl

Institute for Theoretical Chemistry
University of Vienna

Bled, February 2007

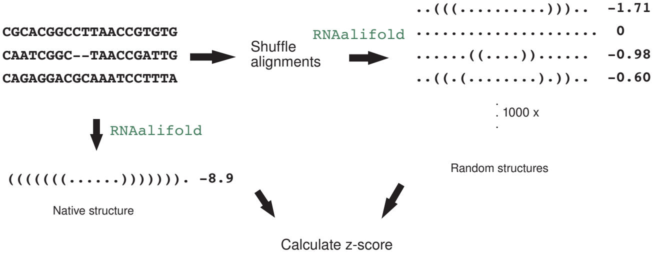
Outline

Mining for structured RNAs in ENCODE data



- 44 ENCODE regions encompassing 1% of the genome
- Targeted sequencing in 28 species
- Multiple alignments created by Multiz/TBA
- Goal: unbiases screen of all non-repeat alignments (10–14 MB) of RNA structures using state-of-the art methods:
  - `AlifoldZ`
  - `RNAz`
  - `EvoFold`

AlifoldZ



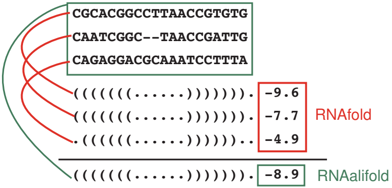
$$z = (m - \mu) / \sigma$$

m ... Consensus minimum free energy of native alignment

μ, σ ... Mean and standard deviation of MFEs of random alignments

RNAz

a) Structural conservation



```
CGCACGGCCTTAACCGTGTG
CAATCGGC--TAACCGATTG
CAGAGGACGCAAATCCTTTA
((((((((......)))))))). -9.6
((((((((......)))))))). -7.7
.(((((((......)))))))).. -4.9
((((((((......)))))))). -8.9
```

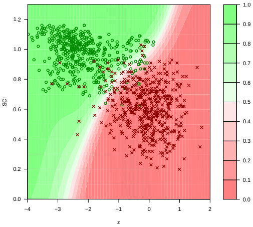
$$\mathrm{SCI} = \frac{\text{Consensus minimum free energy}}{\text{Mean single minimum free energies}}$$

b) Thermodynamic stability

$$z = (m - \mu) / \sigma$$

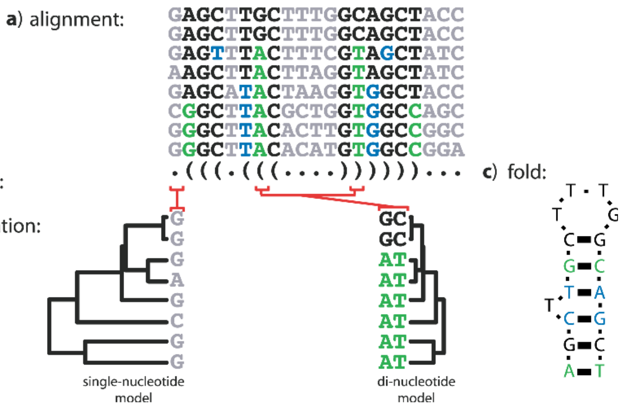
m ... Minimum free energy of native sequences calculated by RNAfold

μ, σ ... Mean and standard deviation of MFEs of random sequences



c) SVM classification

EvoFold



a) alignment:

```
GAGCTTGCTTTGGCAGCTACC
GAGCTTGCTTTGGCAGCTACC
GAGTTTACTTTCGTAGCTATC
AAGCTTACTTAGGTAGCTATC
GAGCATACTAAGGTGGCTACC
CGGCTTACGCTGGTGGCCAGC
GGGCTTACACTTGTGGCCGGC
GGGCTTACACATGTGGCCGGA
.(((.(((....))))))...
```

b) SCFG generated secondary structure:

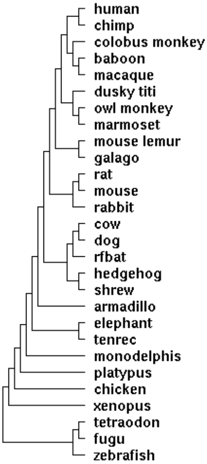
c) fold:

d) Phylogenetic evaluation:

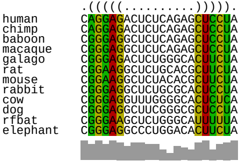
single-nucleotide model

di-nucleotide model

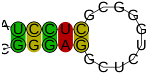
The problem of large genome-wide alignments



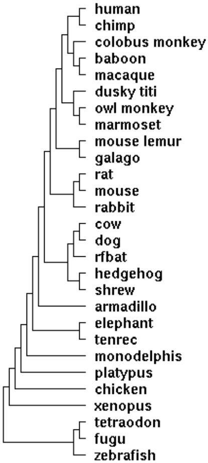
ENCODE species



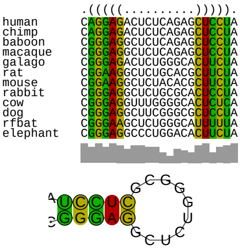
```
              .(((((..........))))).
human         CAGGAGACUCUCAGAGCUCCUA
chimp         CAGGAGACUCUCAGAGCUCCUA
baboon        CGGGAGGCUCUCAGAGCUCCUA
macaque       CGGGAGGCUCUCAGAGCUCCUA
galago        CGGGAGACUCUGGGCACUUCUA
rat           CGGAAGGCUCUGCACGCUUCUA
mouse         CGGAAGGCUCUACACGCUUCUA
rabbit        CGGGAGGCUCUGCGCACUCCUA
cow           CGGGAGGUUUGGGGCACUCUCA
dog           CGGGAGGCUUCGGGCGCUCCUA
rfbat         CGGGAAGCUCUGGGCAUUUUUA
elephant      CGGGAGGCCCUGGACACUUCUA
```



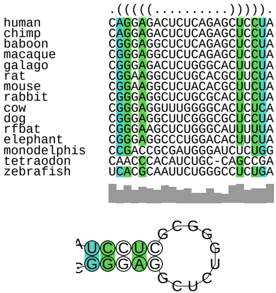
The problem of large genome-wide alignments



ENCODE species

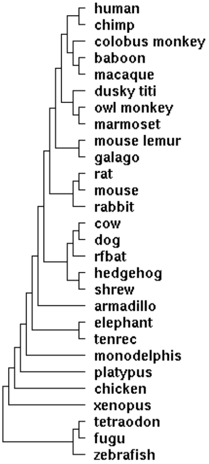


```
             .(((((..........))))).
human        CAGGAGACUCUCAGAGCUCCUA
chimp        CAGGAGACUCUCAGAGCUCCUA
baboon       CGGGAGGCUCUCAGAGCUCCUA
macaque      CGGGAGGCUCUCAGAGCUCCUA
galago       CGGGAGACUCUGGGCACUUCUA
rat          CGGAAGGCUCUGCACGCUUCUA
mouse        CGGAAGGCUCUACACGCUUCUA
rabbit       CGGGAGGCUCUGCGCACUCCUA
cow          CGGGAGGUUUGGGGCACUCUCA
dog          CGGGAGGCUUCGGGCGCUCCUA
rfbat        CGGGAAGCUCUGGGCAUUUUUA
elephant     CGGGAGGCCCUGGACACUUCUA
```

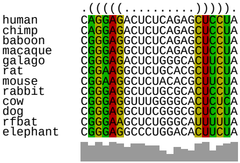


```
             .(((((..........))))).
human        CAGGAGACUCUCAGAGCUCCUA
chimp        CAGGAGACUCUCAGAGCUCCUA
baboon       CGGGAGGCUCUCAGAGCUCCUA
macaque      CGGGAGGCUCUCAGAGCUCCUA
galago       CGGGAGACUCUGGGCACUUCUA
rat          CGGAAGGCUCUGCACGCUUCUA
mouse        CGGAAGGCUCUACACGCUUCUA
rabbit       CGGGAGGCUCUGCGCACUCCUA
cow          CGGGAGGUUUGGGGCACUCUCA
dog          CGGGAGGCUUCGGGCGCUCCUA
rfbat        CGGGAAGCUCUGGGCAUUUUUA
elephant     CGGGAGGCCCUGGACACUUCUA
monodelphis  CCGACCGCGAUGGGAUCUCUGG
tetraodon    CAACCCACAUCUGC-CAGCCGA
zebrafish    UCACGCAAUUCUGGGCCUCUGA
```

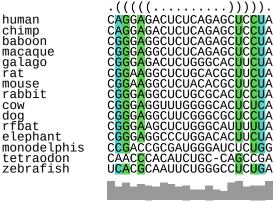
The problem of large genome-wide alignments



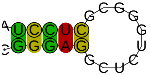
ENCODE species

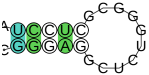


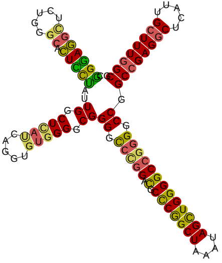
```
          .(((((..........))))).
human     CAGGAGACUCUCAGAGCUCCUA
chimp     CAGGAGACUCUCAGAGCUCCUA
baboon    CGGGAGGCUCUCAGAGCUCCUA
macaque   CGGGAGGCUCUCAGAGCUCCUA
galago    CGGGAGACUCUGGGCACUUCUA
rat       CGGAAGGCUCUGCACGCUUCUA
mouse     CGGAAGGCUCUACACGCUUCUA
rabbit    CGGGAGGCUCUGCGCACUCCUA
cow       CGGGAGGUUUGGGGCACUCUCA
dog       CGGGAGGCUUCGGGCGCUCCUA
rfbat     CGGGAAGCUCUGGGCAUUUUUA
elephant  CGGGAGGCCCUGGACACUUCUA
```

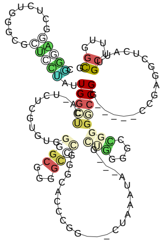


```
             .(((((..........))))).
human        CAGGAGACUCUCAGAGCUCCUA
chimp        CAGGAGACUCUCAGAGCUCCUA
baboon       CGGGAGGCUCUCAGAGCUCCUA
macaque      CGGGAGGCUCUCAGAGCUCCUA
galago       CGGGAGACUCUGGGCACUUCUA
rat          CGGAAGGCUCUGCACGCUUCUA
mouse        CGGAAGGCUCUACACGCUUCUA
rabbit       CGGGAGGCUCUGCGCACUCCUA
cow          CGGGAGGUUUGGGGCACUCUCA
dog          CGGGAGGCUUCGGGCGCUCCUA
rfbat        CGGGAAGCUCUGGGCAUUUUUA
elephant     CGGGAGGCCCUGGACACUUCUA
monodelphis  CCGACCGCGAUGGGAUCUCUGG
tetraodon    CAACCCACAUCUGC-CAGCCGA
zebrafish    UCACGCAAUUCUGGGCCUCUGA
```







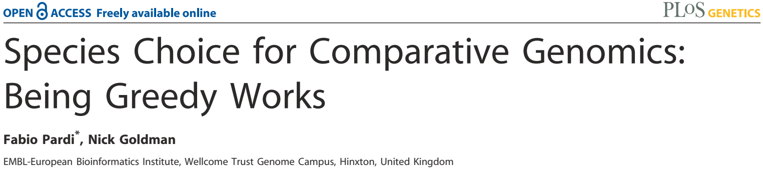


Pragmatic solution: Selecting subsets

- Subsets of 6 and 10 sequences for `RNAz` and `AlifoldZ`, respectively.
- Optimized for a target mean pairwise identity of 85%: reliable alignments and covariation.
- Used greedy algorithm for species selection.

Pragmatic solution: Selecting subsets

- Subsets of 6 and 10 sequences for `RNAz` and `AlifoldZ`, respectively.
- Optimized for a target mean pairwise identity of 85%: reliable alignments and covariation.
- Used greedy algorithm for species selection.



OPEN ACCESS Freely available online

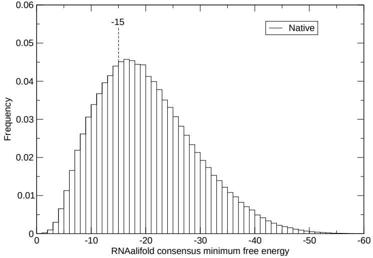
PLoS GENETICS

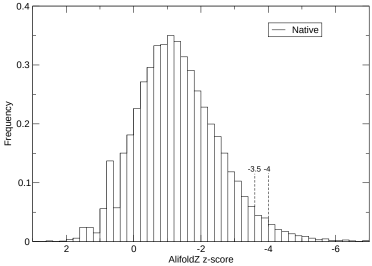
Species Choice for Comparative Genomics: Being Greedy Works

Fabio Pardi[*], Nick Goldman

EMBL-European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, United Kingdom

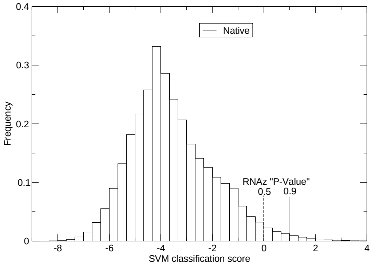
Results of AlifoldZ

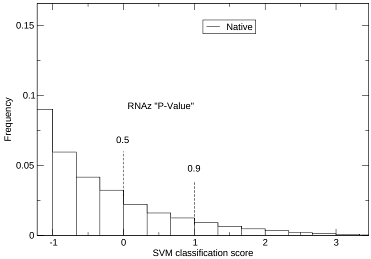




- 660 hits (0.7% of input) with $z < -3.5$
- 384 hits (0.2% of input) with $z < -4$

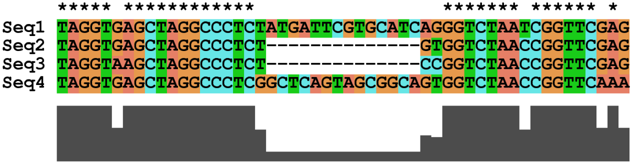
Results of RNAz





- 7,093 hits (7.7% of input) with $P > 0.5$
- 3,707 hits (4.2% of input) with $P > 0.9$

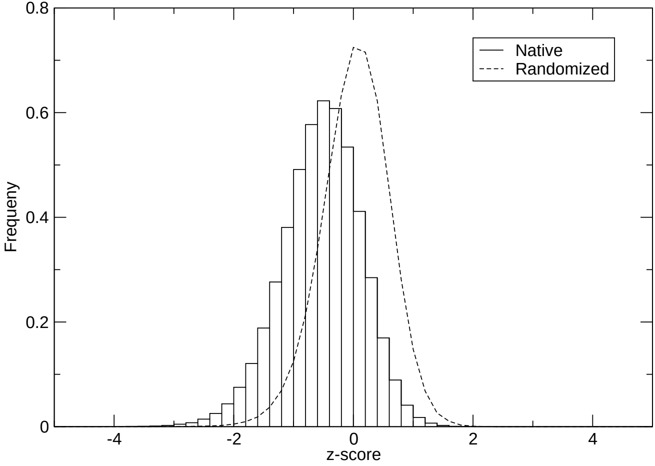
Estimating false positives by shuffling



***** ************ ******* ****** *
Seq1 TAGGTGAGCTAGGCCCTCTATGATTCGTGCATCAGGGTCTAATCGGTTCGAG
Seq2 TAGGTGAGCTAGGCCCTCT--------------GTGGTCTAACCGGTTCGAG
Seq3 TAGGTAAGCTAGGCCCTCT--------------CCGGTCTAACCGGTTCGAG
Seq4 TAGGTGAGCTAGGCCCTCGGCTCAGTAGCGGCAGTGGTCTAACCGGTTCAAA

- Current protocol shuffles columns preserving
  - Mean pairwise identity
  - Base composition
  - Local conservation pattern
  - Gap pattern
- Problems:
  - Limiting for large alignments
  - Dinucleotide content

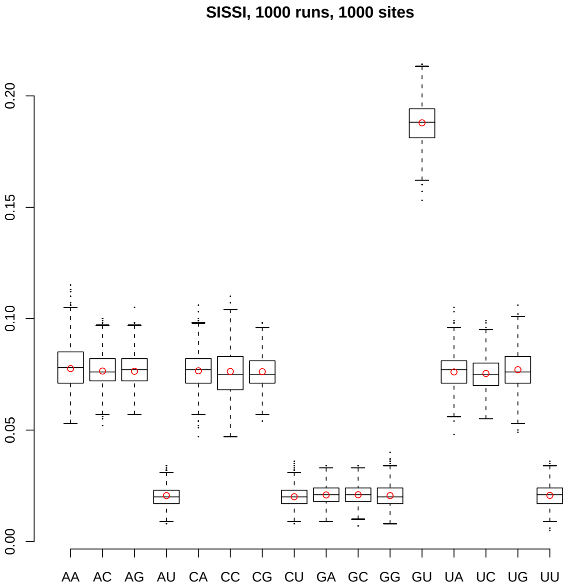
Genomic dinucleotide bias



Solution

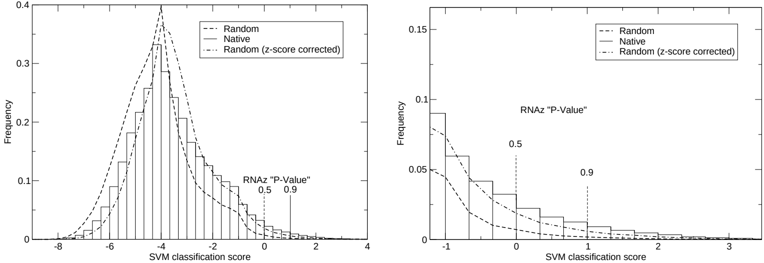
- Simulate alignments rather than shuffling it.
- Simulation produces **on average** alignments with the desired properties.
- Possible strategy:
  1. Choose evolutionary model
  2. Estimate tree under this model
  3. Simulate along this tree using the model
  4. Use rate heterogenities to achieve different divergence levels of sites.
  5. Estimate history of gap pattern formation using maximum parsimony and re-introduce gaps accordingly during the simulation.
- Dinucleotide model: **SISSI** with overlapping dependencies

Simulating alignments with given dinucleotide content



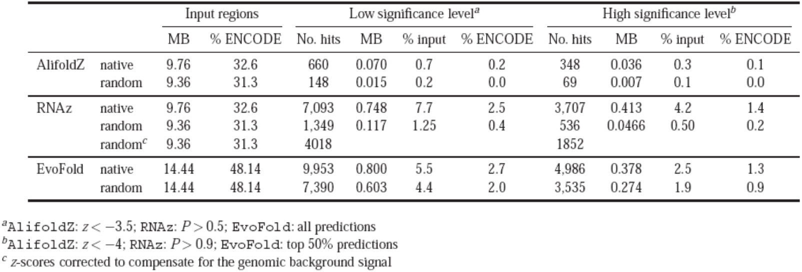
SISSI, 1000 runs, 1000 sites

Simple correction of dinucleotid bias



- Correct all *z*-scores by the background bias of 0.5, re-classify using the SVM
- Estimated false positives for $P > 0.9$:
  - Mononucleotide shuffled: 536
  - Dinucleotide-corrected: 1852

Summary of results



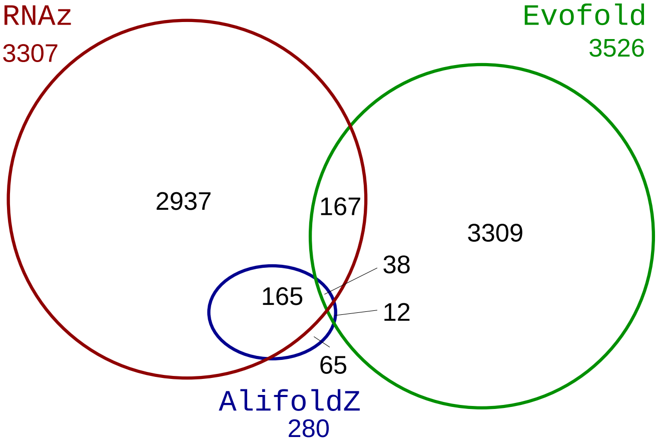
		Input regions		Low significance level[a]				High significance level[b]			
		MB	% ENCODE	No. hits	MB	% input	% ENCODE	No. hits	MB	% input	% ENCODE
AlifoldZ	native	9.76	32.6	660	0.070	0.7	0.2	348	0.036	0.3	0.1
	random	9.36	31.3	148	0.015	0.2	0.0	69	0.007	0.1	0.0
RNAz	native	9.76	32.6	7,093	0.748	7.7	2.5	3,707	0.413	4.2	1.4
	random	9.36	31.3	1,349	0.117	1.25	0.4	536	0.0466	0.50	0.2
	random[c]	9.36	31.3	4018				1852			
EvoFold	native	14.44	48.14	9,953	0.800	5.5	2.7	4,986	0.378	2.5	1.3
	random	14.44	48.14	7,390	0.603	4.4	2.0	3,535	0.274	1.9	0.9

[a] `AlifoldZ`: $z < -3.5$; `RNAz`: $P > 0.5$; `EvoFold`: all predictions

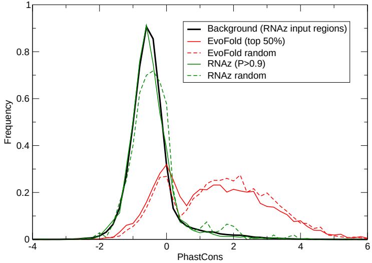
[b] `AlifoldZ`: $z < -4$; `RNAz`: $P > 0.9$; `EvoFold`: top 50% predictions

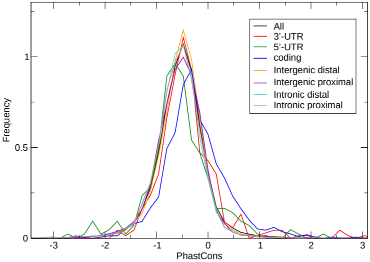
[c] z-scores corrected to compensate for the genomic background signal

Overlap of predictions



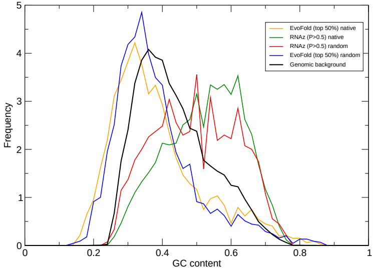
Sequence conservation of predictions

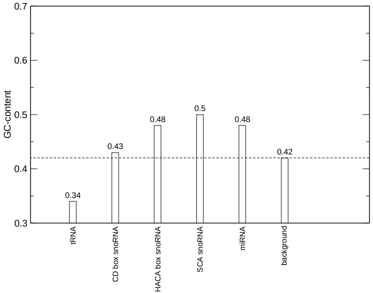




- Both programs have higher false positive rate in regions of high conservation
- `RNAz` predictions essentially follow the background
- `EvoFold` is highly biased for extremely conserved regions.

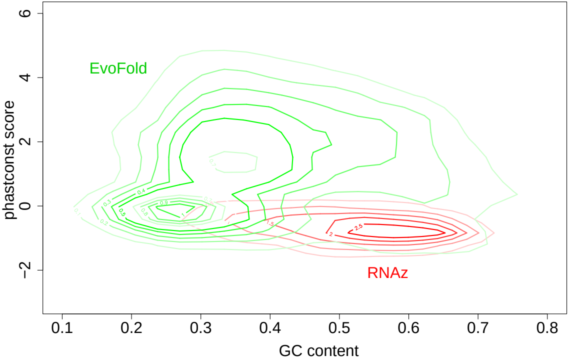
Base composition of predictions



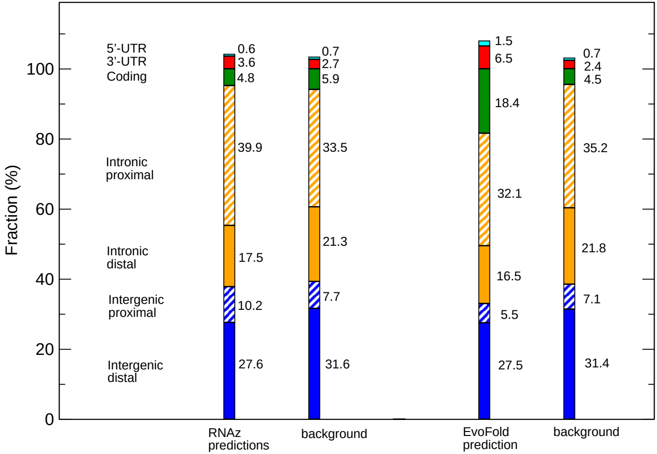


- RNAz favours GC rich regions, EvoFold AT rich regions
- There are known ncRNAs in both ends of the spectrum.

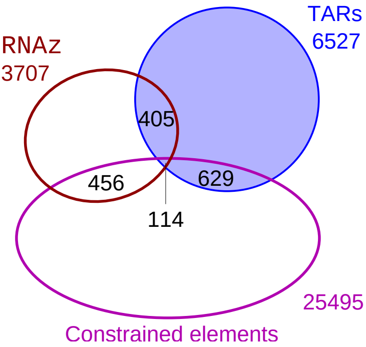
Both programs essentially predict complementary RNA structures

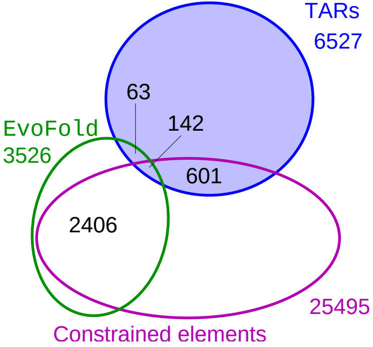


Genomic loction of hits

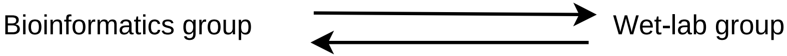


Overlap with other ENCODE data





Experimental verification: ideal case



Bioinformatics group

Wet-lab group

The ENCODE genes and transcripts way

Jan (Yale)

Peter [...], Leipzig

Matt, UCSC

Todd, UCSC

Stefan [...], Vienna

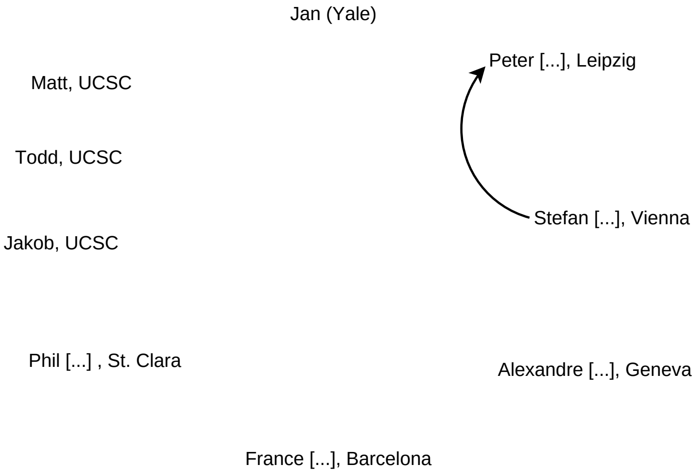
Jakob, UCSC

Phil [...] , St. Clara

Alexandre [...], Geneva

France [...], Barcelona

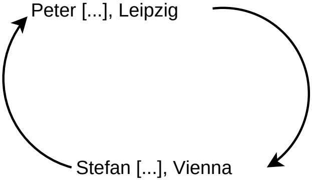
The ENCODE genes and transcripts way



The ENCODE genes and transcripts way

Jan (Yale)

Matt, UCSC



Peter [...], Leipzig

Stefan [...], Vienna

Todd, UCSC

Jakob, UCSC

Phil [...] , St. Clara

Alexandre [...], Geneva

France [...], Barcelona

The ENCODE genes and transcripts way

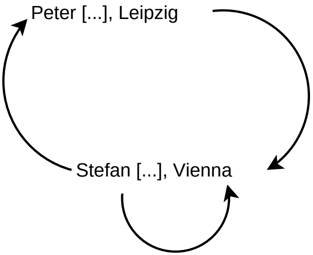
Jan (Yale)

Matt, UCSC

Todd, UCSC

Jakob, UCSC

Phil [...] , St. Clara



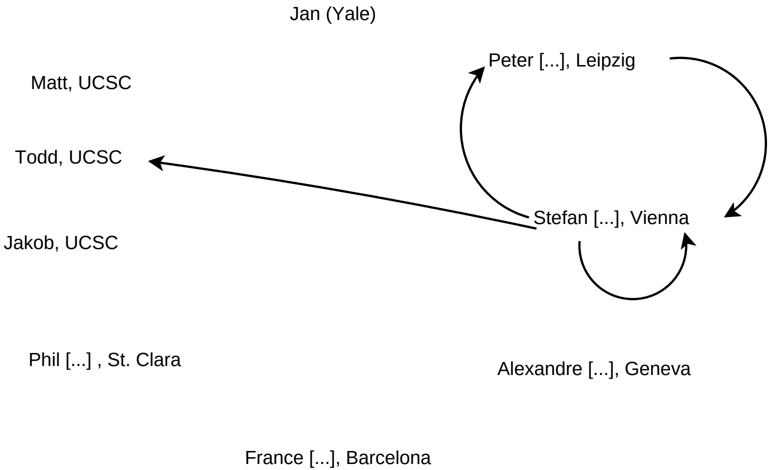
Peter [...], Leipzig

Stefan [...], Vienna

Alexandre [...], Geneva

France [...], Barcelona

The ENCODE genes and transcripts way



Jan (Yale)

Peter [...], Leipzig

Matt, UCSC

Todd, UCSC

Stefan [...], Vienna

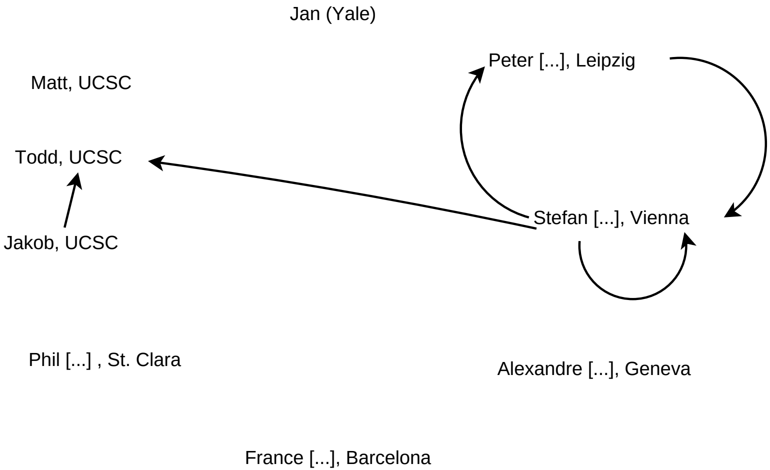
Jakob, UCSC

Phil [...] , St. Clara

Alexandre [...], Geneva

France [...], Barcelona

The ENCODE genes and transcripts way



Jan (Yale)

Peter [...], Leipzig

Matt, UCSC

Todd, UCSC

Stefan [...], Vienna

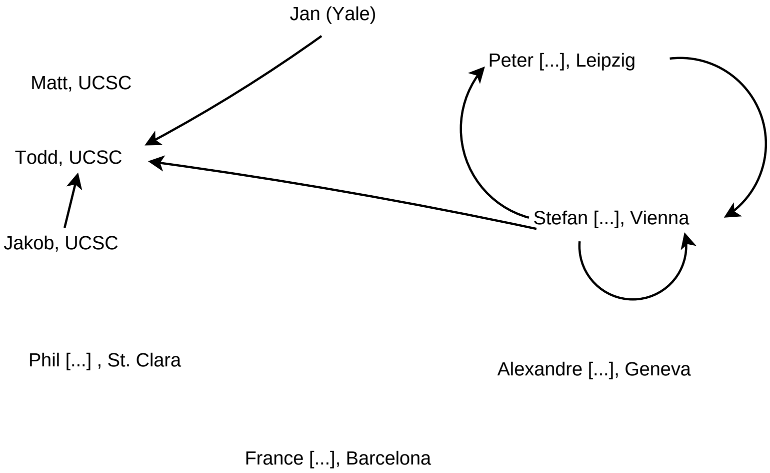
Jakob, UCSC

Phil [...] , St. Clara

Alexandre [...], Geneva

France [...], Barcelona

The ENCODE genes and transcripts way



Jan (Yale)

Peter [...], Leipzig

Matt, UCSC

Todd, UCSC

Stefan [...], Vienna

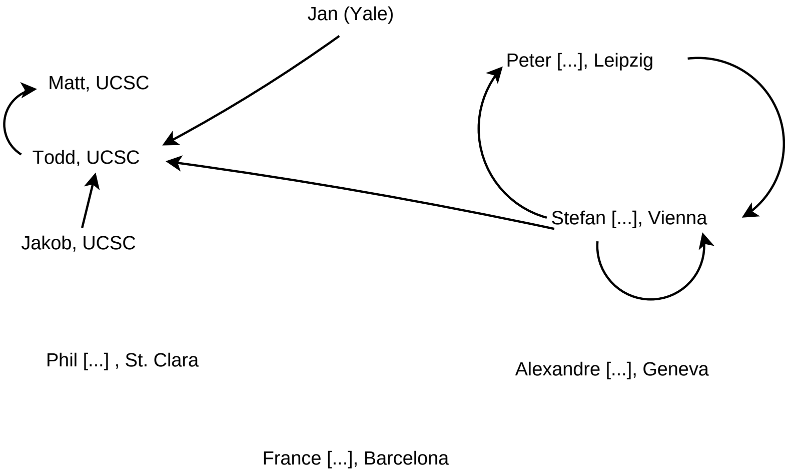
Jakob, UCSC

Phil [...] , St. Clara

Alexandre [...], Geneva

France [...], Barcelona

The ENCODE genes and transcripts way



Jan (Yale)

Matt, UCSC

Todd, UCSC

Jakob, UCSC

Peter [...], Leipzig

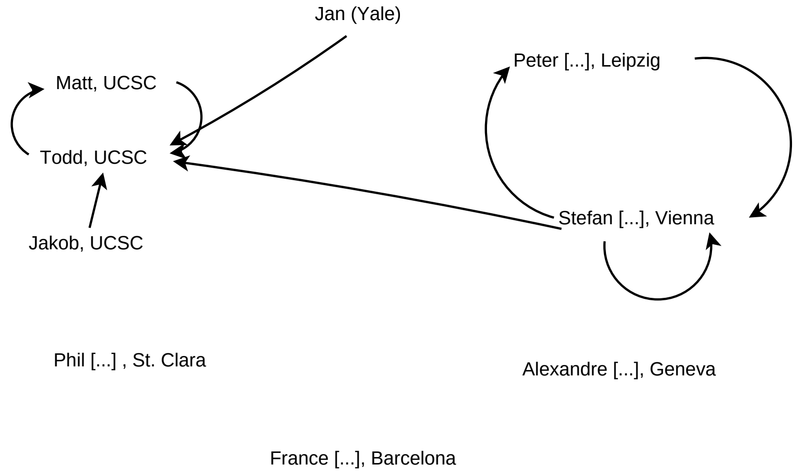
Stefan [...], Vienna

Phil [...] , St. Clara

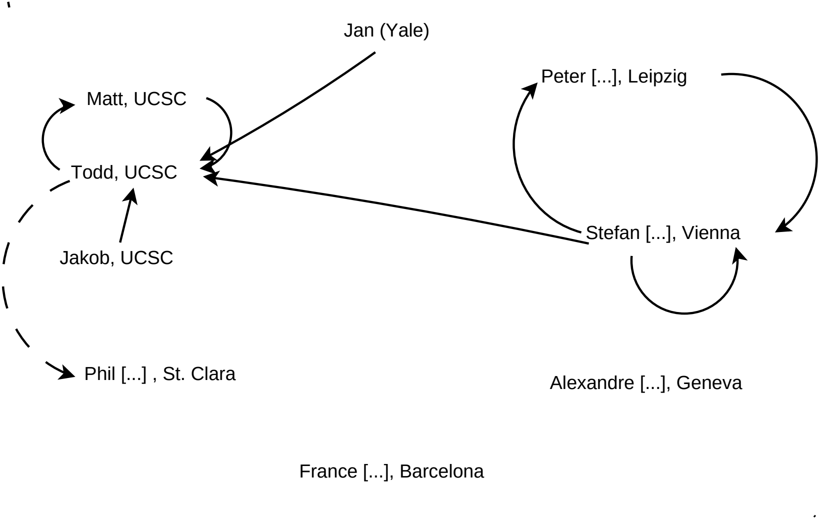
Alexandre [...], Geneva

France [...], Barcelona

The ENCODE genes and transcripts way



The ENCODE genes and transcripts way



Jan (Yale)

Matt, UCSC

Peter [...], Leipzig

Todd, UCSC

Stefan [...], Vienna

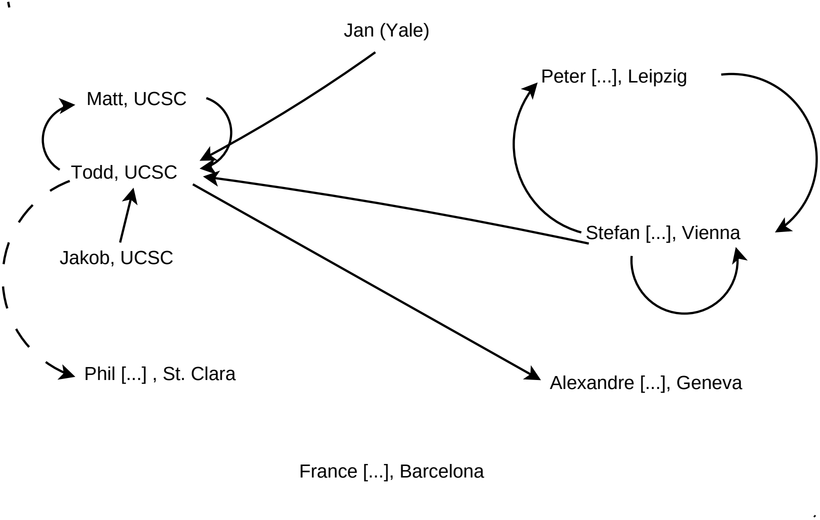
Jakob, UCSC

Phil [...] , St. Clara

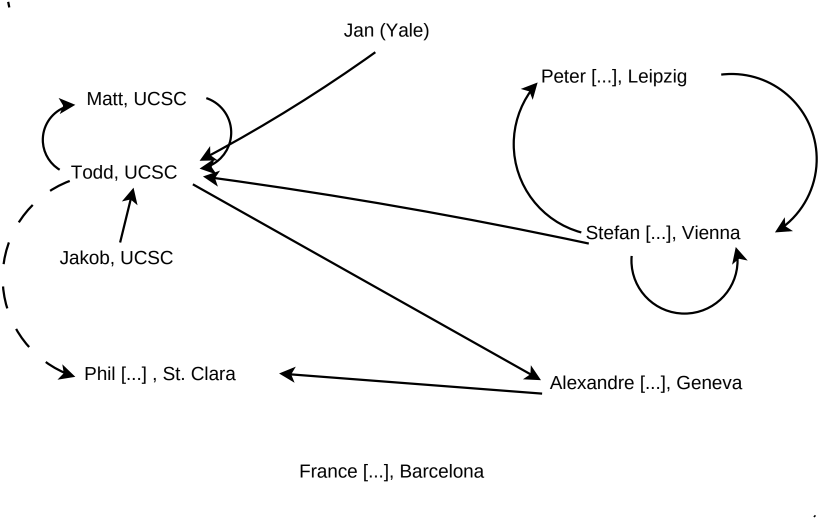
Alexandre [...], Geneva

France [...], Barcelona

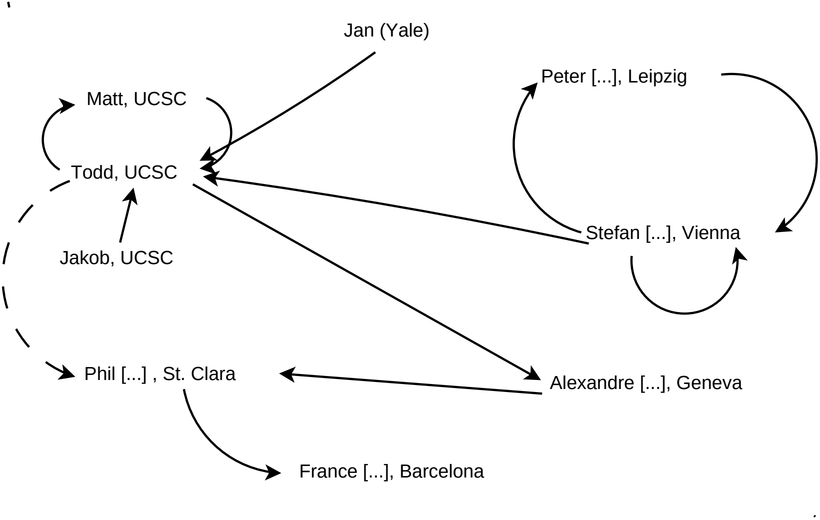
The ENCODE genes and transcripts way



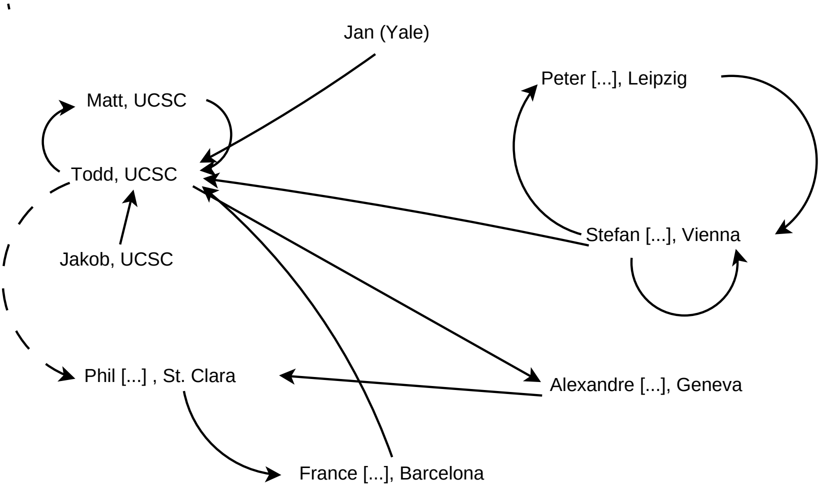
The ENCODE genes and transcripts way



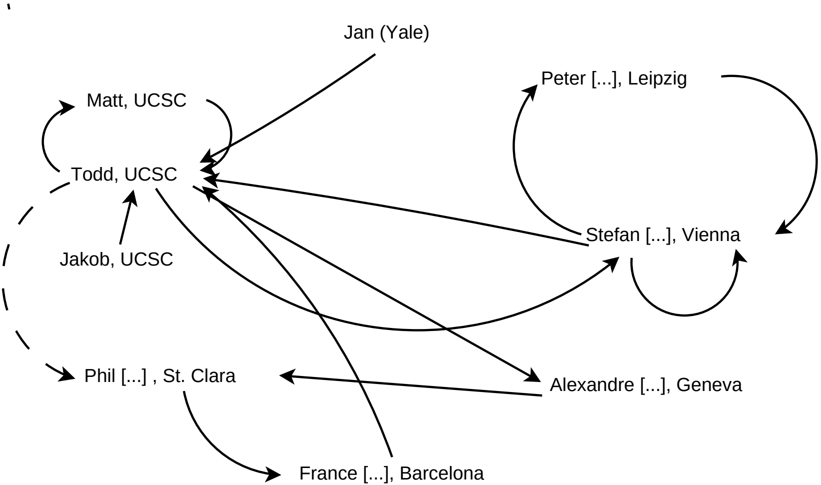
The ENCODE genes and transcripts way



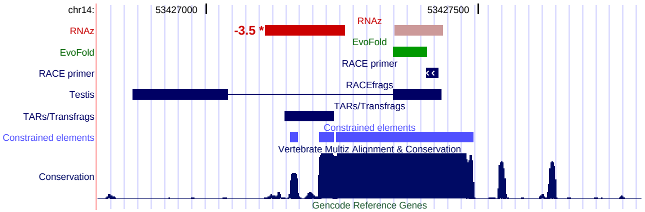
The ENCODE genes and transcripts way

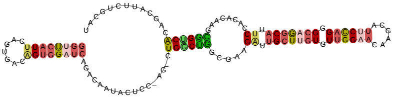


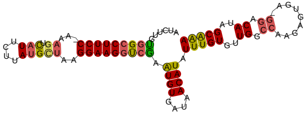
The ENCODE genes and transcripts way



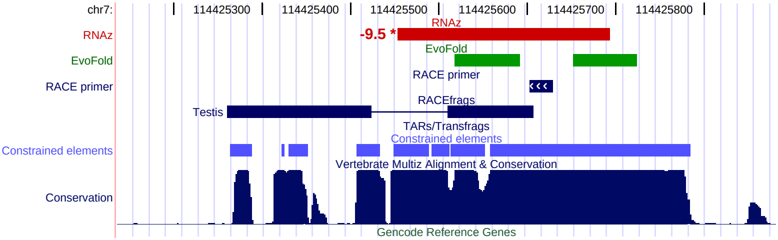
Intergenic RNAs

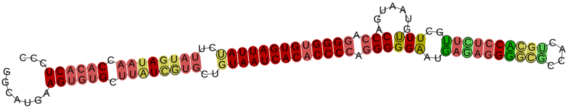




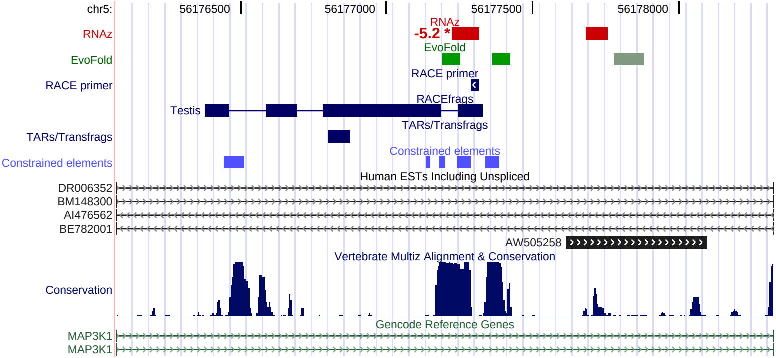


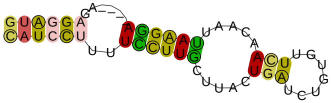
Intergenic RNAs



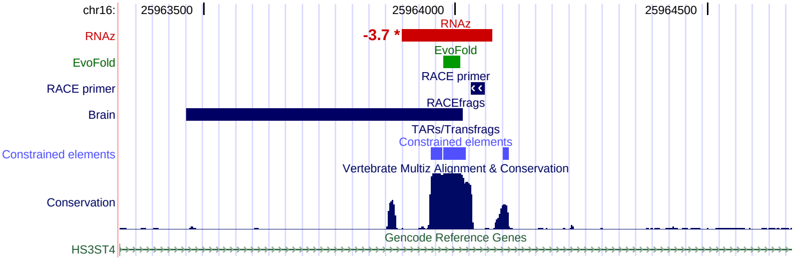


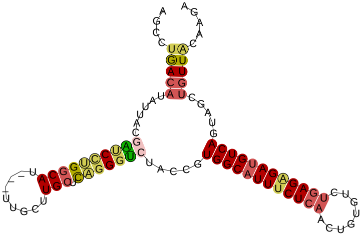
Intronic RNAs



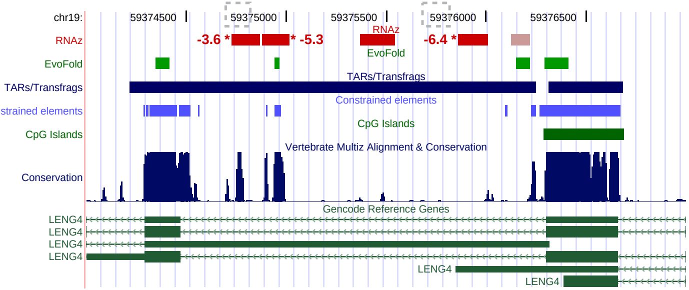


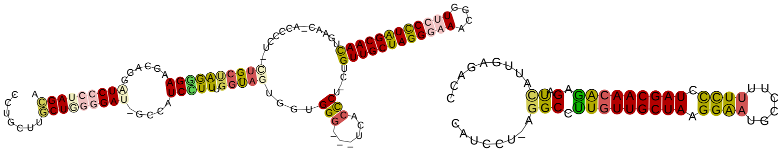
Intronic RNAs

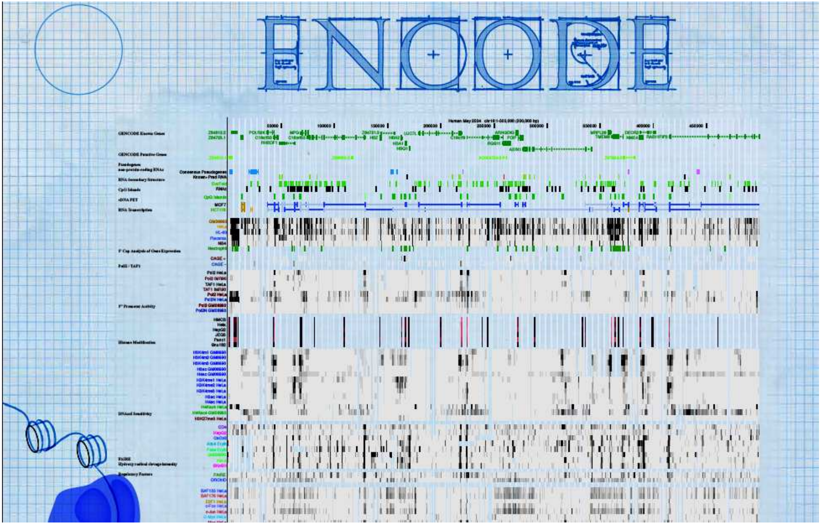


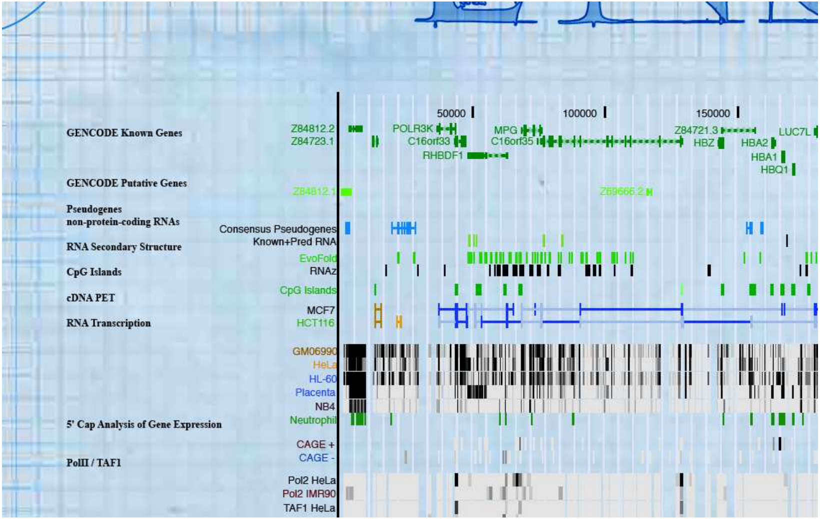


Alternative spliced loci






ENCODE
Human May 2004 chr16:1-500,000 (500,000 bp)
50000
100000
150000
200000
250000
300000
350000
400000
450000
GENCODE Known Genes
GENCODE Putative Genes
Pseudogenes
non-protein-coding RNAs
Consensus Pseudogenes
Known+ Pred RNA
RNA Secondary Structure
EvoFold
RNAz
CpG Islands
CpG Islands
ditPET
MCF7
HCT116
RNA Transcription
GM06990
HeLa
HL-60
Placenta
NB4
Neutrophil
5' Cap Analysis of Gene Expression
CAGE +
CAGE -
PolII / TAF1
Pol2 HeLa
TAF1 HeLa
Pol2 HeLa
5' Promoter Activity
Pol2 GM06990
HMCB
HeLa
HepG2
JEG3
Panc1
Histone Modification
H3K4me1 GM06990
H3K4me2 GM06990
H3K4me3 GM06990
H3ac GM06990
H4ac GM06990
H3K4me1 HeLa
H3K4me2 HeLa
H3K4me3 HeLa
H3ac HeLa
H4ac HeLa
DNAseI Sensitivity
H3K27me3 HeLa
CD4
HepG2
FAIRE
Hydroxy radical cleavage intensity
Regulatory Factors
FAIRE
ORChID
BAF155 HeLa
BAF170 HeLa
E2F1 HeLa
c-Fos HeLa
c-Jun HeLa
c-Myc HeLa


GENCODE Known Genes
GENCODE Putative Genes
Pseudogenes
non-protein-coding RNAs
RNA Secondary Structure
CpG Islands
cDNA PET
RNA Transcription
5' Cap Analysis of Gene Expression
PolII / TAF1
50000
100000
150000
Z84812.2
Z84723.1
POLR3K
C16orf33
RHBDF1
MPG
C16orf35
Z84721.3
HBZ
HBA2
HBA1
HBQ1
LUC7L
Z84812.1
Z69666.2
Consensus Pseudogenes
Known+Pred RNA
EvoFold
RNAz
CpG Islands
MCF7
HCT116
GM06990
HeLa
HL-60
Placenta
NB4
Neutrophil
CAGE +
CAGE -
Pol2 HeLa
Pol2 IMR90
TAF1 HeLa

Acknowledgements

- Ivo L. Hofacker, Andreas Gruber, Andrea Tanzer (Vienna)
- Peter Stadler, Jörg Hackermüller, Jana Hertel, Manja Lindemeyer, Kristin Reiche, Claudia Stocsits (Leipzig)
- Jakob S. Pedersen (USA, UCSC)
- Jan O. Korbel, Mark B. Gerstein, Michael Snyder (USA, Yale)
- France Denoeud, Julien Lagarde,Roderic Guigo (Barcelona)
- Jorg Drenkow, Philipp Kapranov, Thomas R. Gingeras (USA, Affymetrix)
- Catherine Ucla, Carine Wyss, Stylianos E. Antonarakis, Alexandre Reymond (Switzerland)