Detection of conserved non-coding RNAs in the pig EST data and related mammals

Stefan Seemann

Division of Genetics and Bioinformatics, IBHV, Copenhagen University, Denmark

Bled, Feb 2007

Outline

Non-coding RNAs

Non-coding RNAs (ncRNAs) are transcripts that

- are not translated into proteins
- perform several important regulatory functions
  - $\Rightarrow$ in the geneexpression
  - $\Rightarrow$ in the subcellular distribution of RNAs and proteins
  - $\Rightarrow$ as modulators of the protein functions

In contrast to protein coding genes ncRNAs do not have common primary sequence features.

$\Rightarrow$ Other methods are necessary!

Expressed Sequence Tags

Expressed Sequence Tags (ESTs) are

- short subsequences (500 to 800 nt) of transcripted RNA
- single-stranded
- sequencing results of cDNA library
- poly-A tails of transcribed RNAs are used to initiate cDNA creation
- relatively inaccurate (around 2% error)
- indicate gene expression patterns

Polyadenylated ncRNAs

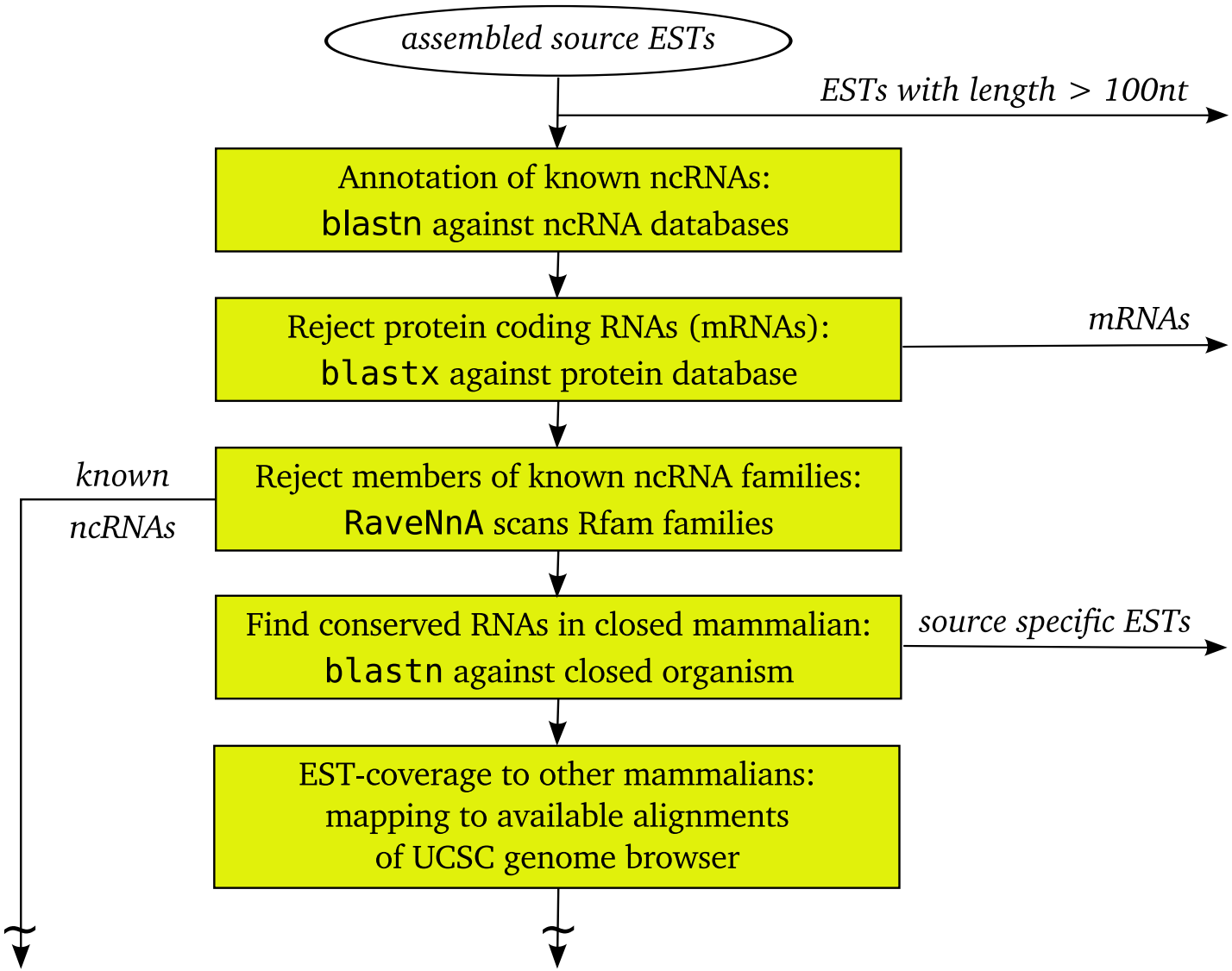
- GO term GO:0043629 verified in yeast
- hybridization signals of ncRNA targets in microarray surveys, which have been used oligo(dT) oligonucleotides to amplify cDNA probes

Pig EST data

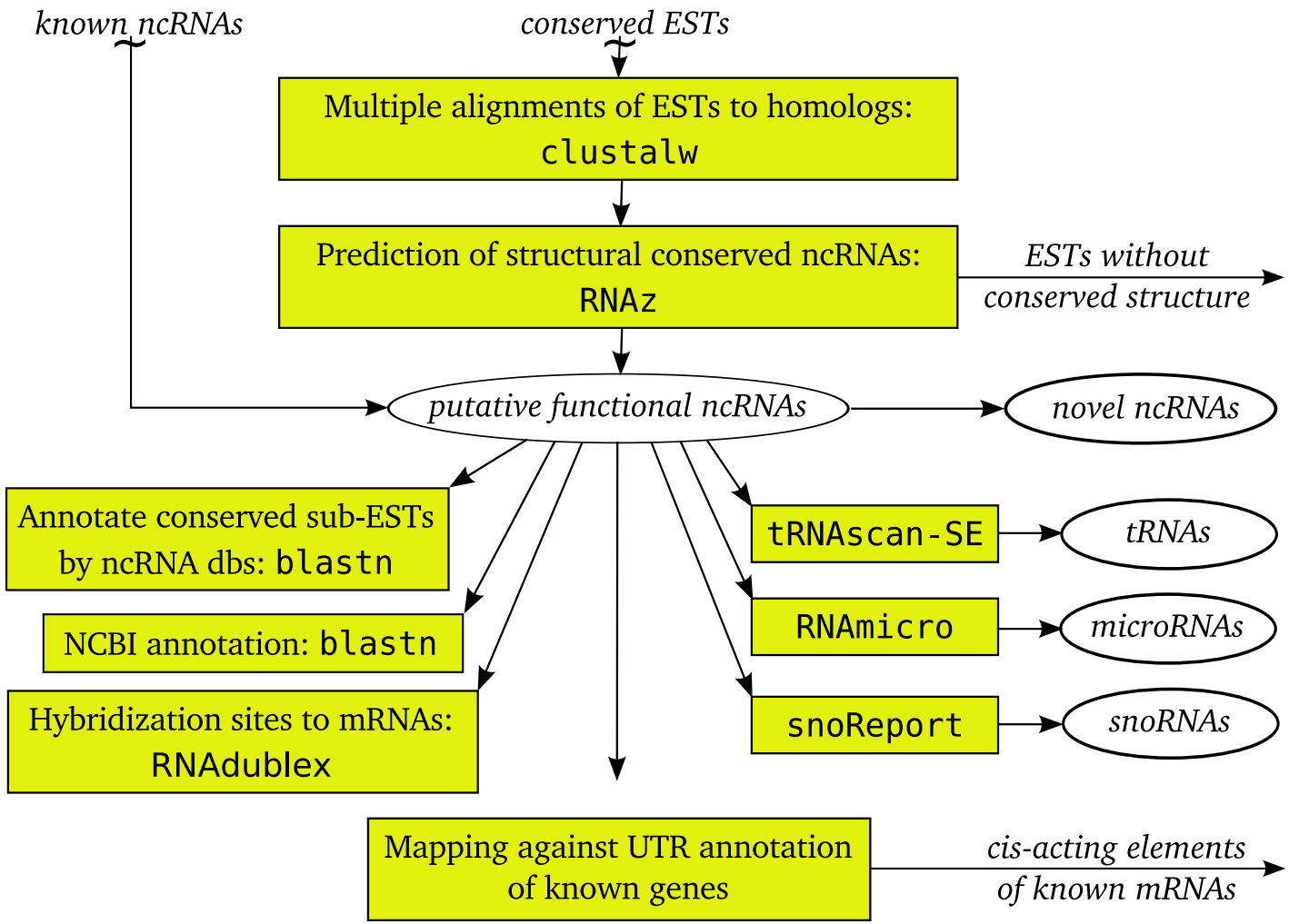
- approximately 1.1 Mio pig ESTs
  $\Rightarrow$ 636.516 from Sino-Danish Pig Genome Project
  $\Rightarrow$ 385.375 from GenBank
- automatic sequence assembly using the `Distiller` package resulted in 121.800 assembled ESTs
  $\Rightarrow$ 48.629 contigs
  $\Rightarrow$ 73.171 singletons

Outline

Detection pipeline (part 1)



Detection pipeline (part 2)

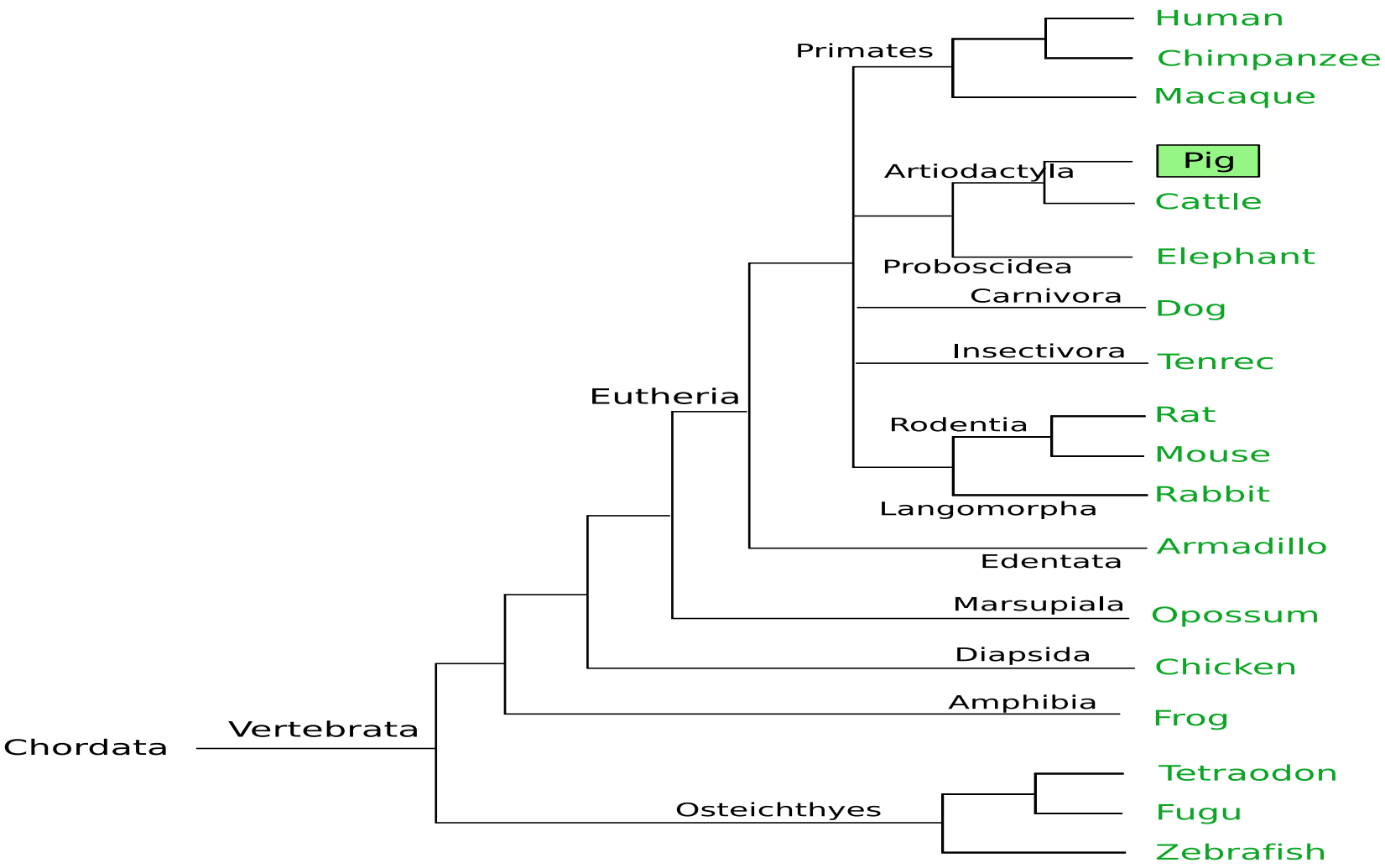


Several parameter settings

- `blastn` with EST specific parameters [1] (advanced): -r 1 -q -1 -G 1 -E 2 -W 9 -F ”m D;V” -U -e 1e-20 -b 100 -v 1000
- filtering of blastn results: the best (lowest evalue) of non-overlapping blast hits with length$>$100nt and identity$>$75%
- UCSC over.chain files for pairwise alignments, filtering 80% coverage
- UCSC maf file for multiple alignments, filtering 60% coverage

[1] I.Korf, M.Yandell, J.Bedell. BLAST. *o’reilly*, p.137, 2003

Pig genome and related mammals



RNAz 1.0pre[2]

- detection of evolutionary conserved (SCI) and thermodynamically stable (z-score) RNA sec.struct.
- input are multiple sequence alignments
- usage of a support vector machine (SVM), trained by well known ncRNAs, as classificator (p-value)
- distinguish between conserved structural RNA (ncRNA) and not conserved structural RNA

[2] Stefan Washietl, University Vienna

Framework EST2ncRNA and Database

The pipeline is available as reusable framework called `EST2ncRNA`:

- object oriented `Perl` version 5.8.5
- 10 modules controlled by executable `est2ncrna.pl`
- `http://www.bioinf.uni-leipzig.de/~seemann/EST2ncRNA/EST2ncRNA.tar.gz`

The data is stored and maintained by a RDBMS:

- `MySQL` database server version 4.1.12
- 15 tables

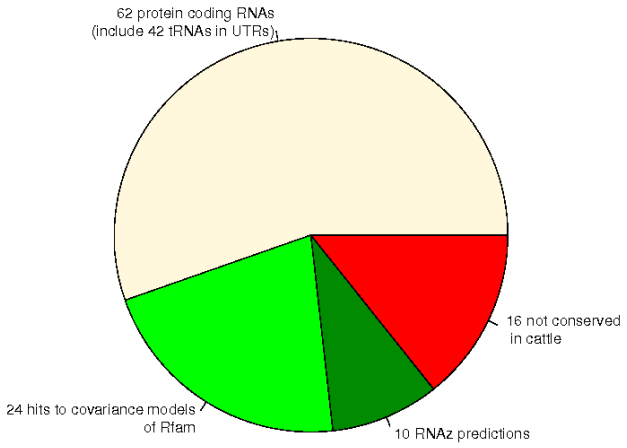
Integration is planned in the distiller database of KVL PigEST resource[3].

[3] http://pigest.kvl.dk/server/index.html

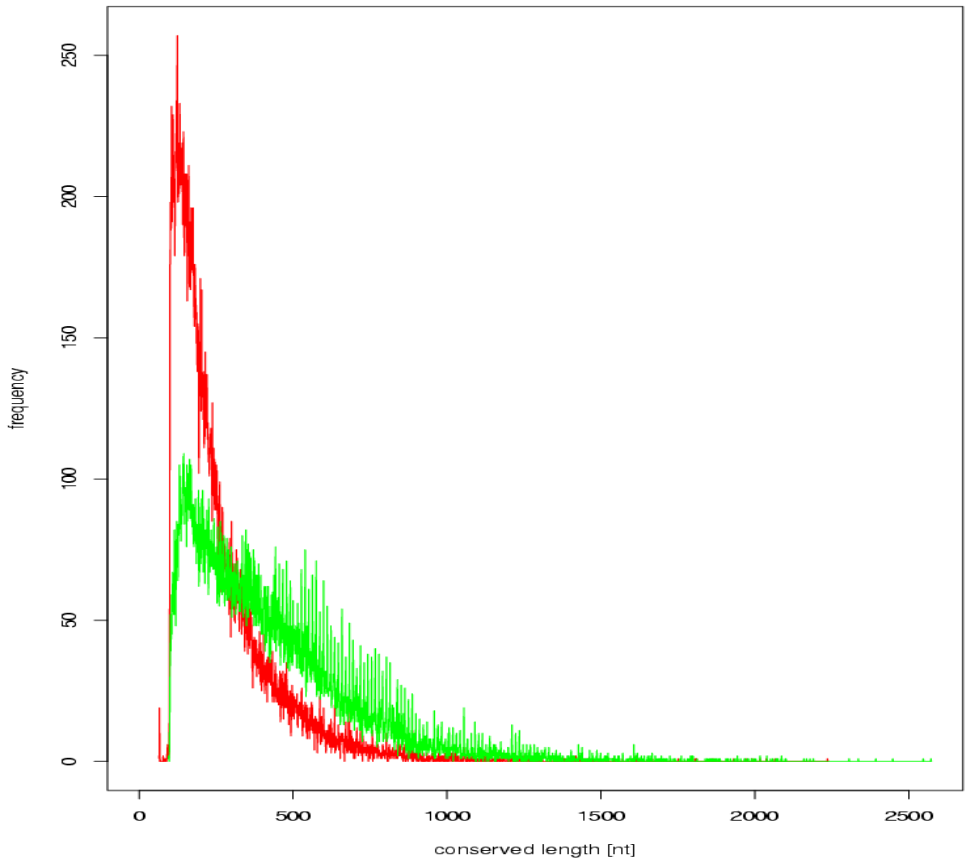
Outline

Pipeline results

- 18.000 protein coding genes
- 177 members of known ncRNA families
  - $\Rightarrow$ 111 tRNAs are found by `RaveNnA`
  - $\Rightarrow$ tRNAs are almost completely verified by `tRNAscan_SE`
- 112 ESTs annotated as ncRNAs by sequence similarity to `RNAdb`, `Rfam`, `miRBase`, `fantom_nc`
  - $\Rightarrow$ 62 cis-regulated mRNAs (42 times tRNA in UTR)
  - $\Rightarrow$ 50 ncRNAs (14 miRNAs, 19 tRNAs, 2 snoRNAs)



Pig to Cattle alignment



Distribution of alignment lengths of pig ESTs to cattle. `Blastn` results applied with EST specific parameters (green curve) and with standard parameters (red) are presented.

Orthologous pig sequences

30.926 pig ESTs are conserved at least in cattle (around 4.000 ones only in cattle) and 71.885 ESTs are pig specific.

Pig EST-coverage generated by UCSC pairwise alignments	
aligned organism	**conserved EST subsequences**
cow (bosTau2)	37.123
human (hg17)	32.437
mouse (mm7)	24.540

Pig EST-coverage generated by UCSC multiple alignments			
aligned organism	**cons EST subseq**	**aligned organism**	**cons EST subseq**
human (hg17)	9.796	tenrec (echTel1)	7.189
chimp (panTro1)	9.574	armadillo (dasNov1)	7.178
dog (canFam2)	9.531	opossum (monDom2)	6.815
macaque (rheMac2)	9.504	chicken (galGal2)	3.884
mouse (mm7)	8.972	frog (xenTro1)	3.251
rat (rn3)	8.819	zebrafish(danRer3)	3.043
elephant (loxAfr1)	7.679	fugu (fr1)	3.000
rabbit (oryCun1)	7.668	tetraodon (tetNig1)	2.917

1.136 high confidence ncRNA candidates

- the high confidence novel ncRNAs consist of 802 contigs and 334 singletons
- 61% of all putative ncRNAs are conserved in cow, human and mouse
- mean length is 143 nts, std dev is 56 nts
- all candidates are also identified by CONC

Putative ncRNAs predicted by RNAz					
p	**z**	**EST-cov.**	**Contigs**	**Singletons**	**Loci**
< 0.50	–	–	41.758	68.590	-
> 0.50	–	–	6.871	4.581	15.534
> 0.90	–	–	3.956	2.229	7.465
> 0.90	< -3	–	802	334	1.217
> 0.90	< -3	$> 80\%$	19	17	36

Further annotation of ncRNA candidates

- 39 high confident microRNAs (35 contigs, 4 singletons) predicted by `RNAmicro`
- 73% of high confidence `RNAz` predictions are similar to ESTs discovered in other organisms (NCBI dbEST)
- mapping to the first long continuous stretch of the pig genome (SSC17, homologous to human HSA20)
  - 42 protin coding ESTs, 78 non protein coding ESTs
  - 36% of non protein coding ESTs have conserved secondary RNA structure (p$>$0.5)

False discovery rate (FDR) of SVM based methods [%]:

	$p > 0.5$	$p > 0.9$
`RNAz`	18.2	11.8
`RNAmicro`	4.9	4.4
`snoReport`	110.8	81.8

Usage of human genome annotation

86% of predicted ncRNAs are human aligned.
The high confidence RNAz predictions ($p > 0.9$, $z < -3$) include

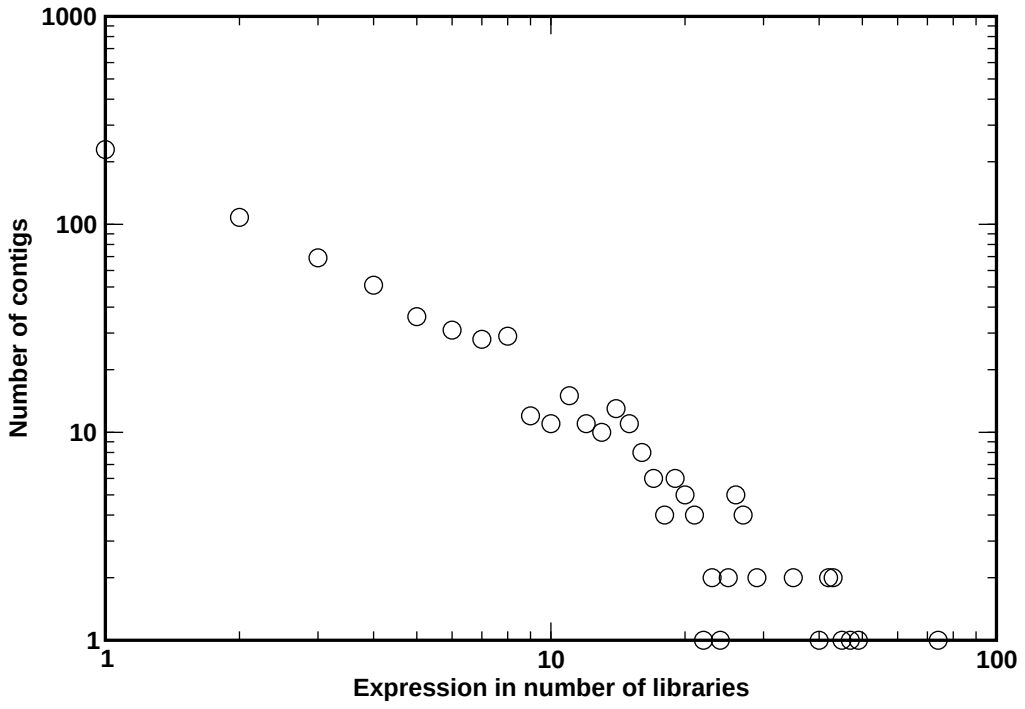
- about 50% of cis-regulated mRNAs
- 15% ncRNAs binding mRNAs (level of significance is 0.01 quantile of shuffled input data)
- 21% ncRNAs covers at least 50% of an existing human structure alignment
- about 800 pig ESTs hold RNA structures which are not predicted before

Conserved secondary structure of novel ncRNA candidate

- Ss1.1-Pig1-116007.5 (166-314, +)
- RNAz p=1, z=-3.06, mid=88.28%
- RNAmicro p=0.91
- 5'-UTR (human CENTG1)

Source distribution of expressed ncRNAs

# tissues	10	15	20	25	30
# ncRNAs with at least 1 read	132	72	37	24	11
# ncRNAs with at least 2 reads	29	8	5	2	1



Outline

Further work

- prediction of reading direction of structured RNAs by `RNAstrand`
- novel ncRNA candidates being conserved in special organisms
- structural clustering of ncRNA candidates to identify new families
- analysis of source distribution of expressed ncRNAs $\Rightarrow$ functional roles, interactions with proteins
- quality analysis

Thank you!!!

My supervisor

- Peter F. Stadler
- Jan Gorodkin