Analysis of the porcine transcriptome from 1 million EST sequences and then something very different namely something *brief* about RNA

Jan Gorodkin, gorodkin@bioinf.kvl.dk

Outline:

- Sequence assembly
- The Porcine EST project
  - Assembly
  - Basic analysis (cDNA lib content / expression)
  - Identifying housekeeping genes
  - Expression analysis
- Perspectives
- The ncRNA stuff

Acknowledgements

University of Copenhagen:

- Susanna Cirera
- Claus B. Jørgensen
- Elfar Torarinsson
- Milena Sawera
- Karsten Scheibye-Knudsen
- Jakob H. Havgaard
- Stefan Seemann
- Troels Arvin
- Steen Lumholdt
- Merete Fredholm

Collaborators:

- Mike Gilchrist (Cambridge University)
- Sino-Danish Pig genome consortium: DIAS, BGI
- Ebbe S. Andersen
- Allan Lind-thomsen
- Rune Lyngsø
- Peter Sestoft
- Niels Tommerup
- Jørgen Kjems
- Gary Stormo
- Ivo Hofacker
- Peter Stadler

Funding:

- Danish Slaughterhouses
- The Danish Research Councils
- The Danish Center for Scientific Computing
- The Ministry of food, agriculture and fisheries

Sequence assembly

First there were sequences ....

Sequence assembly

Usefull observations related to assembly:

- SNP detection.
- Expression level (for ESTs, number of reads in contig).

Associated problems with sequence assembly:

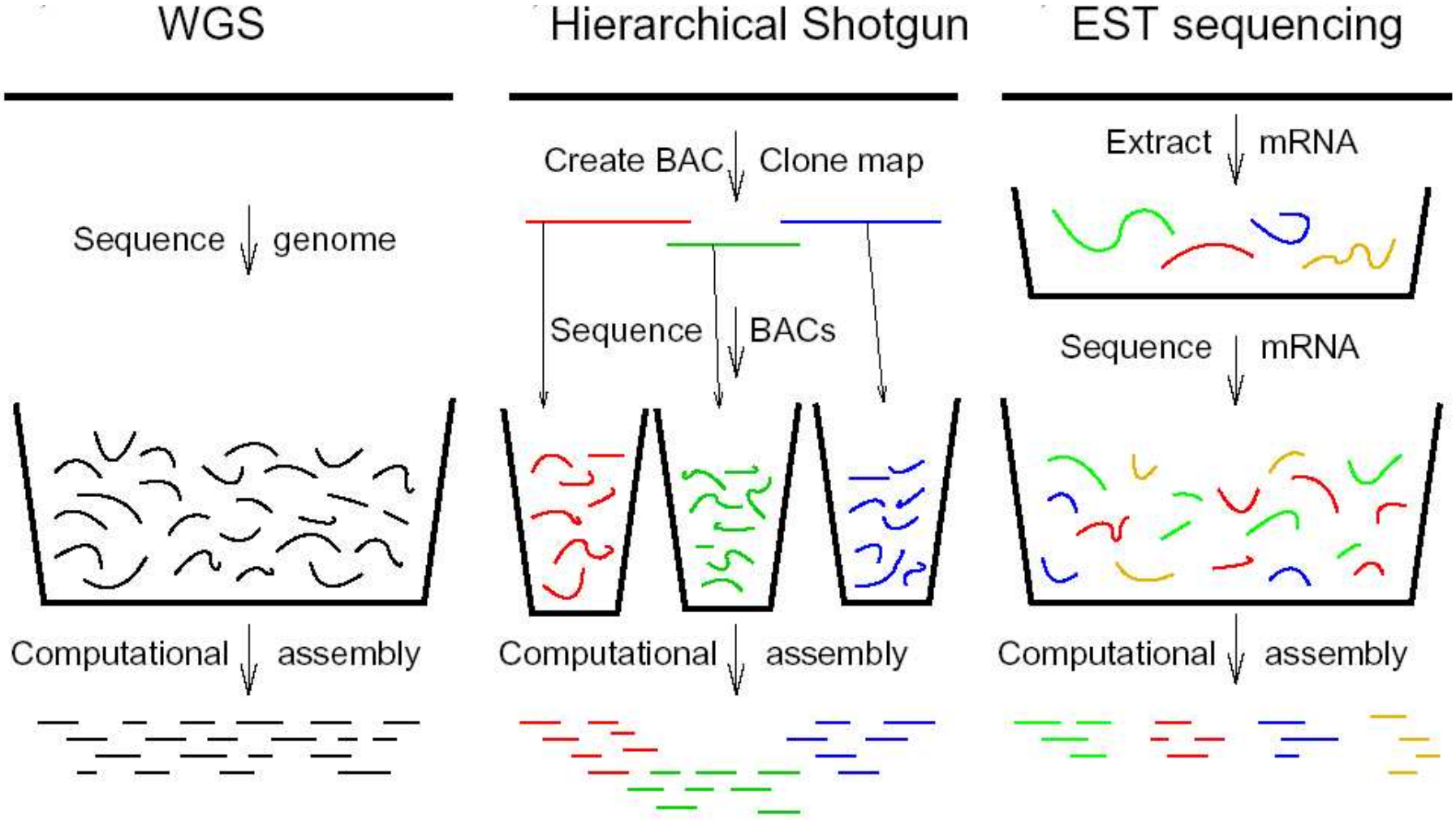
- Repetative sequences.
- Highly expressed sequences (housekeeping genes).
- Alternatively spliced sequences.
- Different proteins using the same domain in different contexts.

Sequence assembly

Strategies

- Bacterial articifical chromosomes. Sequence bit by bit.
- Assemble large pieces. [Kent, GigAssembler].
- Shotgun. Chop up the genome sequence it and assemble it by computer. [you will be trouble if your computer program can not assemble your sequences in reasonable time].
- Size 500–600 nucleotides.
- Human genome: jigsaw puzzle of 30 million pieces.
- Today: Tiling BAC clones and shotgun within each BAC clone.
- Or today: Non-sanger sequencing (journal club).

Sequence assembly scenarios



Sequence assembly

Assembly methods

- Exist a range many not mentioned in the textbook.
- Some known ones: Phrap, TGICL, GigAassembler, celeras assembler, Distiller.
- Most use single linkage realtionship to merged sequence reads into the same cluster (and then assemble them into a contig).
- Distiller, use double linkage.

Porcine genome analysis

- 98 (97 non-normalized) cDNA libraries generated from $\sim$ 200 pigs.
- Libraries covering 35 tissues.
- Pigs: DK: Landrace, Yorkshire, Duroc and Hampshire; CN: Taihu/Erhualian.
- Sequences: 636516 (cleaned from 970000 trace files) and 385375 (cleaned) public available sequences.
- Assembly: Distiller (Gilchrist et al., 2004): 48000 clusters, 73000 singletons.
- Approx. 27% of clusters match UniProt (at least 50% coverage and 60% identity).

Assembly by Distiller

The Distiller software (Gilchrist et al., Dev. Biol 271, 2004.)

- Ungapped alignment in clustering from seed alignment.
- Use BLAST at several levels (outside db framework).
- Detection of alternative splice variants.
- 12-mer scans to find consensus sequence.
- Double linkage clustering (reduce super clusters and chimera problems).
- Post processesing in cluster joining.
- Assembly scales linearly with data size.
- The assembly program is written in SQL.

EST sequence data

Sino-Danish resource

- 970404 initial reads.
- 146533 completely *empty* files!
- Remove unk seqs and seqs < 100 nt: 685851 reads.
  (vector and linker seqs also removed).

Public available ESTs: 398837

EST sequence data

Distiller masking:

SD resource:

Remaining sequences	637270
combination of all other masks	114
mitochondrial sequence	38930
ribosomal sequence	131
repeat sequence	6800
very low complexity sequences	2600
length < 100 nt	0
vector	6

Public ESTs:

Remaining sequences	385375
combination of all other masks	46
mitochondrial sequence	2751
ribosomal sequence	1886
repeat sequence	6755
very low complexity sequences	1242
length < 100 nt	329
vector	453

EST sequence data

Assembly:

Clusters (ex. splice variants)	48629
Singletons	73171

- 430 Clusters should be discarded (repeat sequences).
- Approximately 2645 clusters have alternative splice variants.
- Number of SNPs not extracted yet.
- Total number of used reads for assembly: 1021891
  (seqs in cls: 948720; sing: 73171).

EST sequence data

SD resource vs. public:

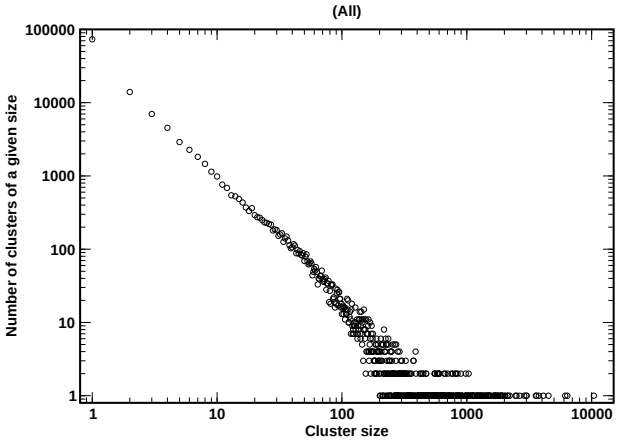
# contigs with SD pigests	35234
# contigs with only SD pigests. (ie. without pubests)	6312
# contigs with pubests.	42048
# contigs with only pubests (ie. without SD pigests)	13126
# singletons SD pigests	26429
# singletons pubests	46742

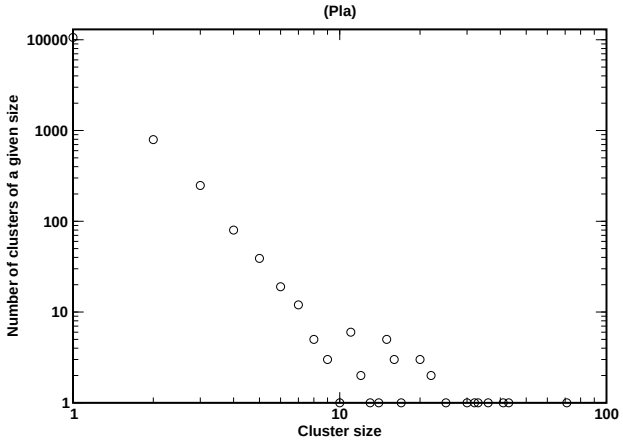
SD ESTs more redundant than public ESTs.

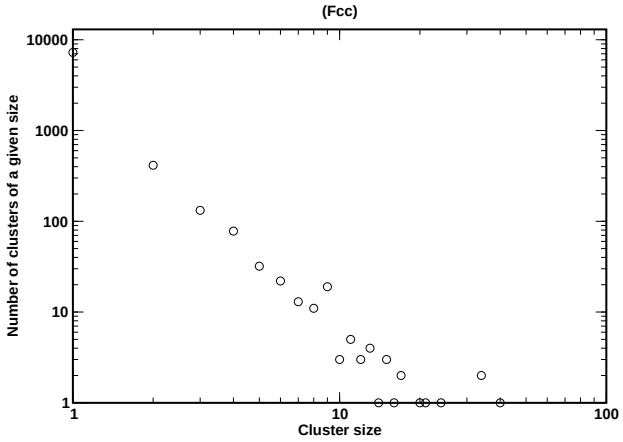
Blast matches

	Contigs		Singletons	
Match level (ID/Sbj)	**UniProt**	**ncRNAdb**	**UniProt**	**ncRNAdb**
M0 (98% / 100%)	1982	21	173	6
M1 (95% / 95%)	1304	18	101	12
M2 (85% / 90%)	2517	72	236	20
M3 (70% / 70%)	3480	—	749	—
M4 (60% / 50%)	3603	—	1355	—
M5 (20% / 20%)	11973	—	12337	—

Scaling invariance of contig sizes

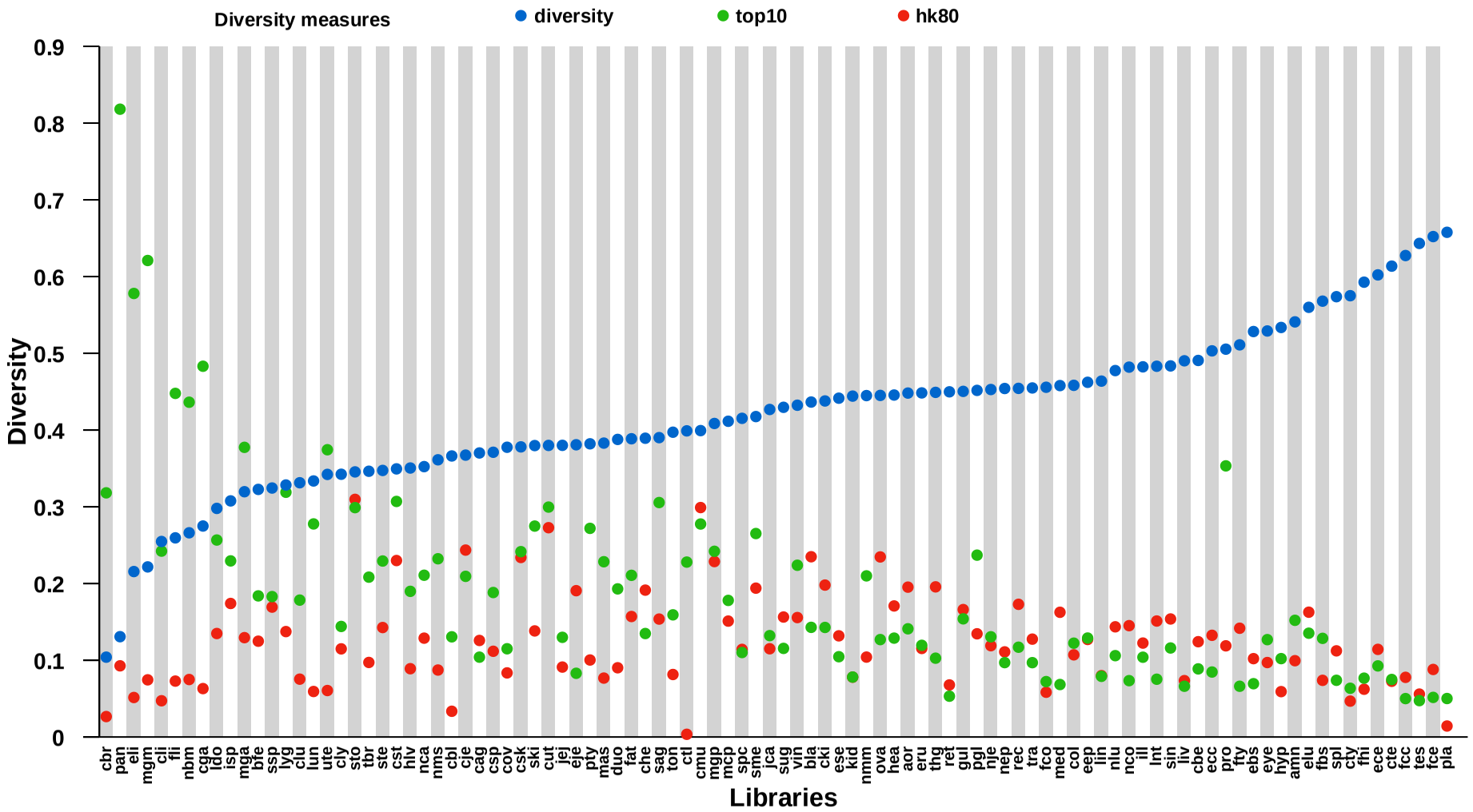






The diversity of porcine cDNA libraries

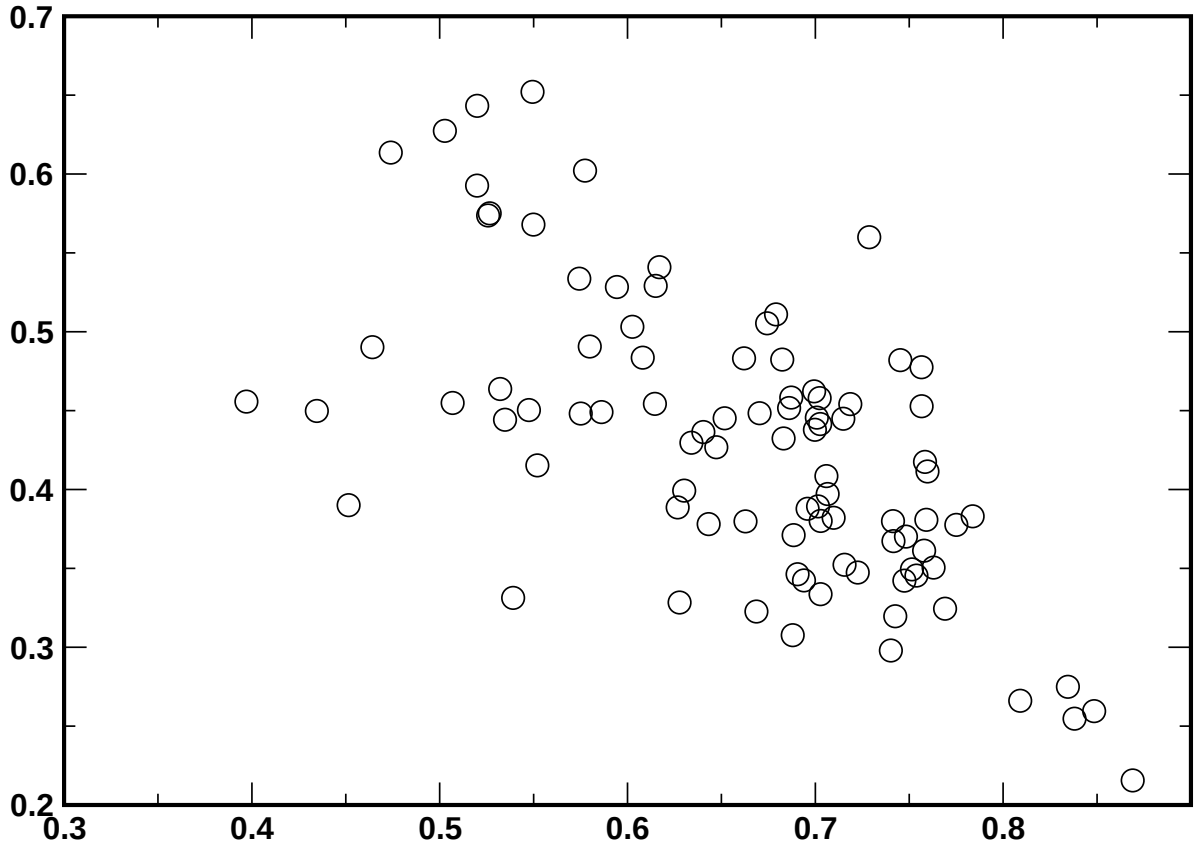
Diversity: The fraction of reads going into clusters/singletons



Not an artifact of library size. Correlation of diversity and lib size: –0.21.

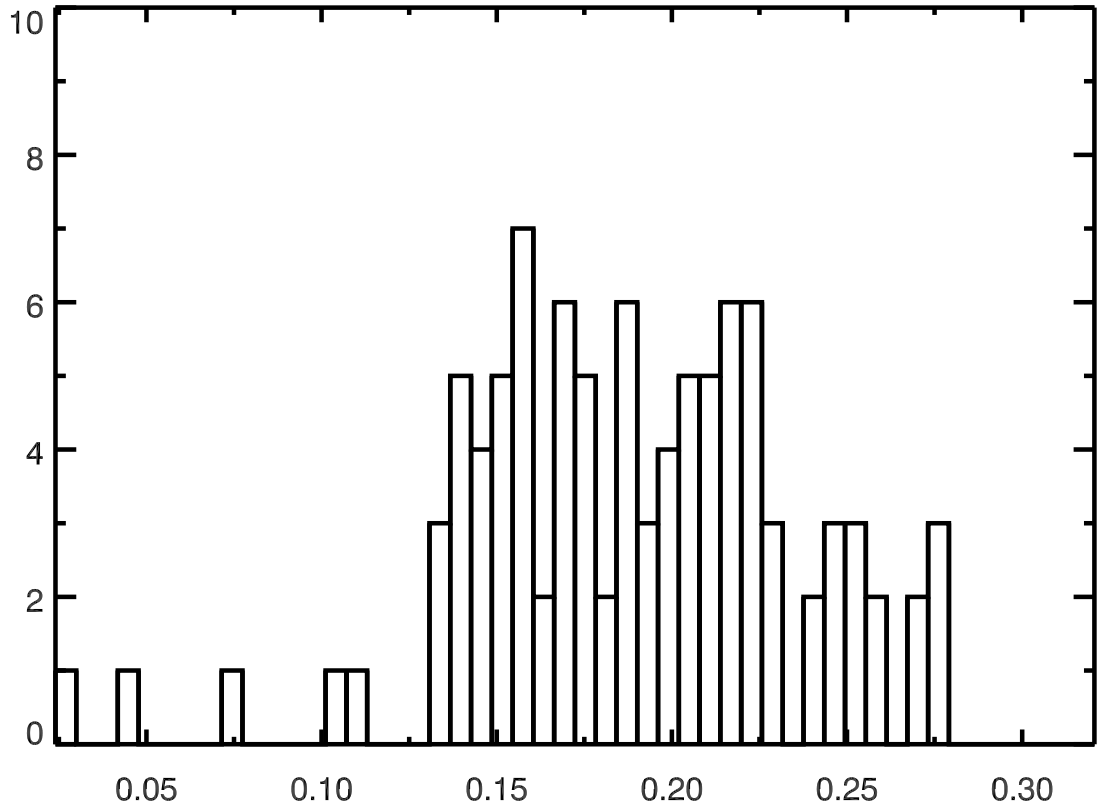
The diversity of porcine cDNA libraries

Diversity versus BLAST matches



The PigEST project

Quality measure: Joint degree of diversity and BLAST matches.
More reliable extraction of highly expressed and uniquely expressed genes.

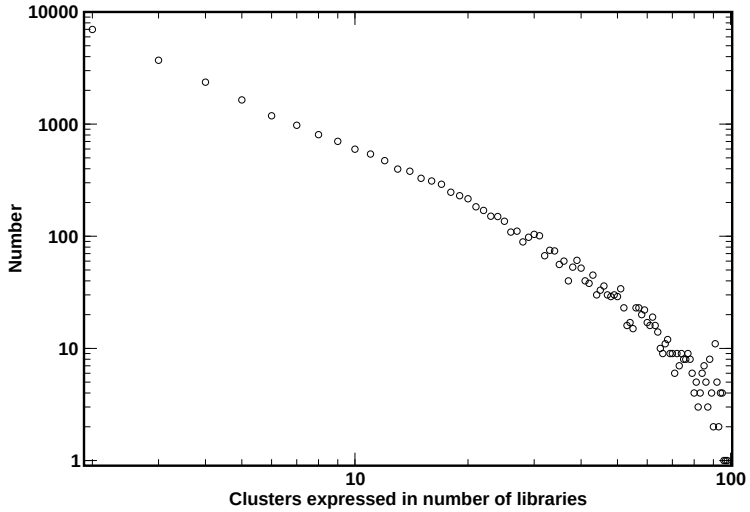


Discard: ctl; cbr; pan; cbl; mgm. We also discard a normalized lib (pla).

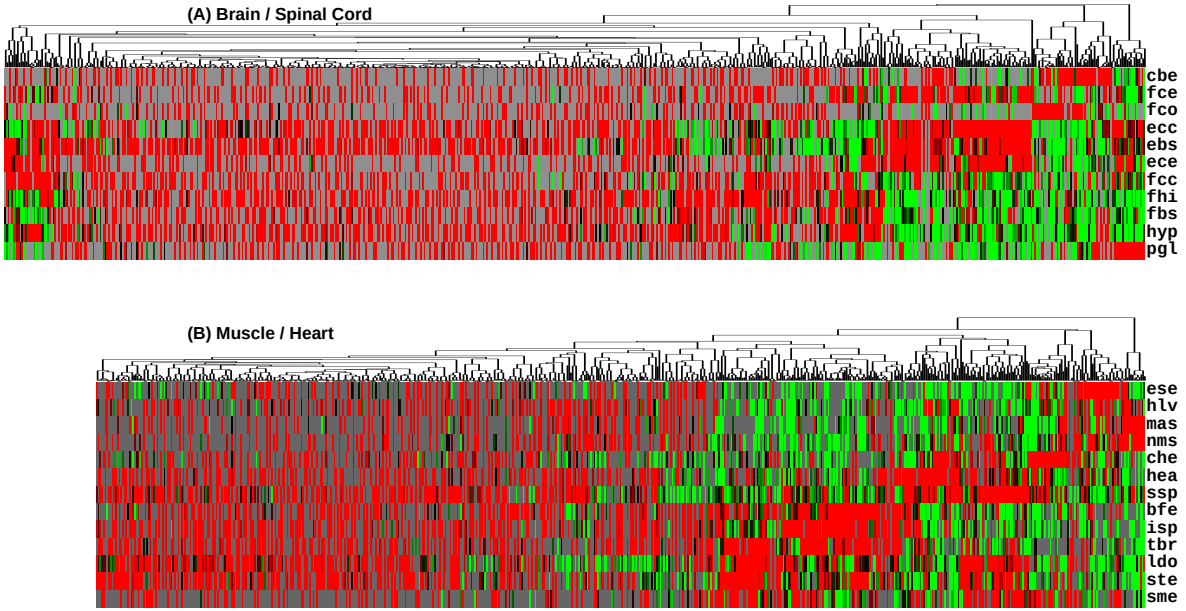
Patterns of differential expression

Compare expression levels of genes (clusters) in non-normalized cDNA libraries.

Not all genes are reprented in each cDNA library.

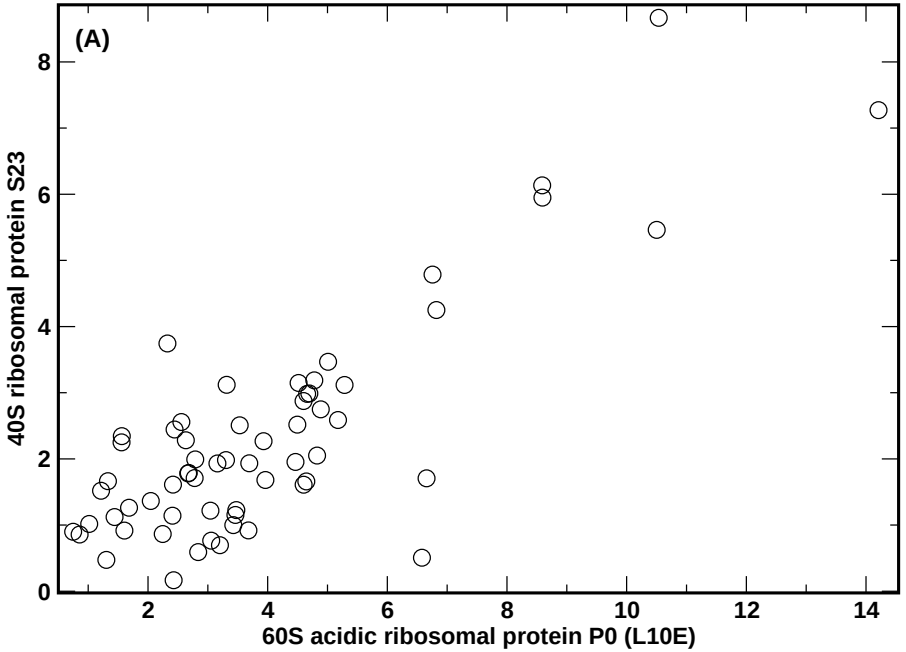


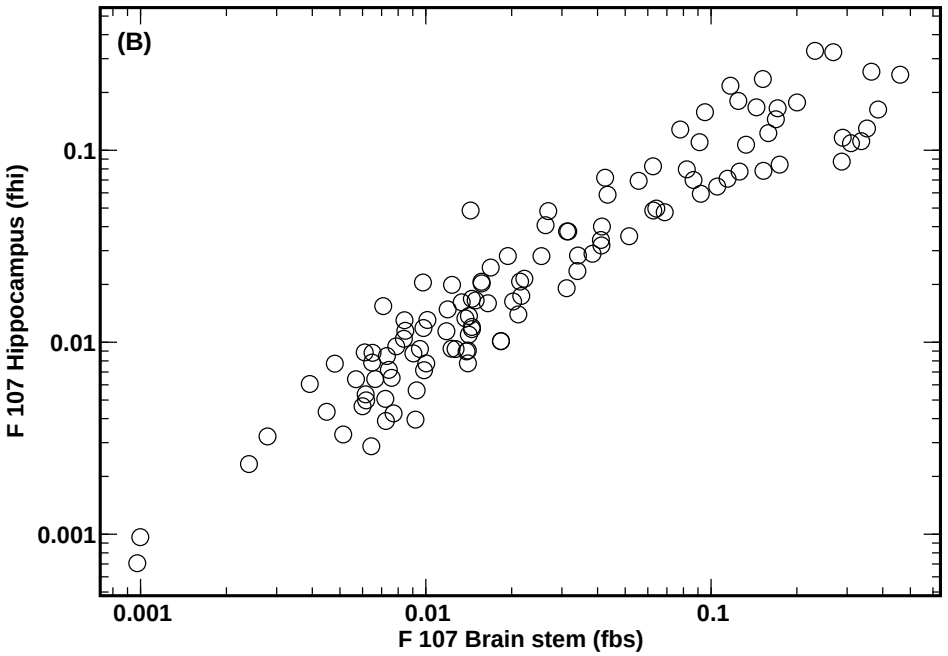
Hierarchical clustering [Eisen, deHoon] of gene-lib subtables



Patterns of differential expression

- Correlations of expression between genes accross cDNA libraris.
- Correlations between accumaleted gene expression of pairs of libraries [Identification of 128 pairs of libraries.]





Identifying cDNA specific genes

- Pre-select lib with highest expression factor 25 times higher expression than lib with second highest expression.
- At least 11 reads these libraries.
- Stat test by Audic and Clevarie ($P<0.05$).
- Obtain 876 candidates.

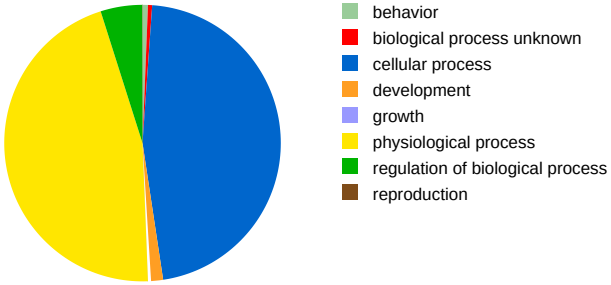
Identifying cDNA specific genes

- Inspection of 50 candiates.
- Expression from Unigene (mouse, human or cow).
- Using the description line of matching blast sjbct seq.
- 25 cases with agreement. 11 cases of disagreement. In 14 cases comparison could not be made.
- For M0 cases only: 16/22 (70%) agreement.
- Eight additional PCR experiments showed agreement.

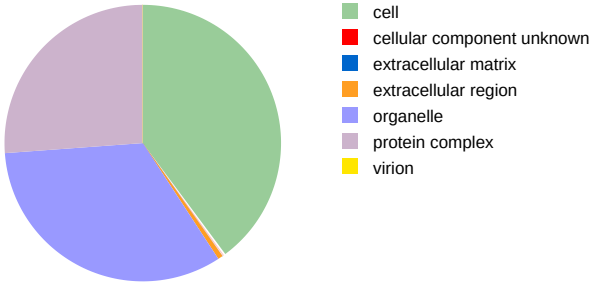
Gene ontology patterns

Comparing the gene ontology content between the libraries.
Ecc (cortex foetus 50 days) versus Fco (adult frontal cortex)

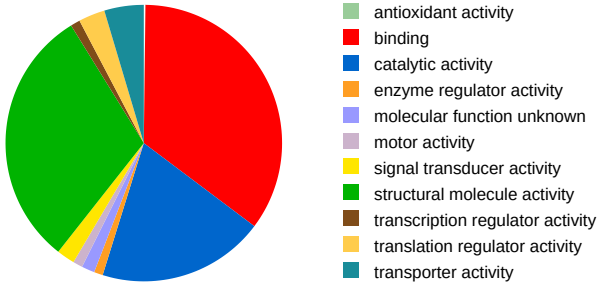
ecc − biological process



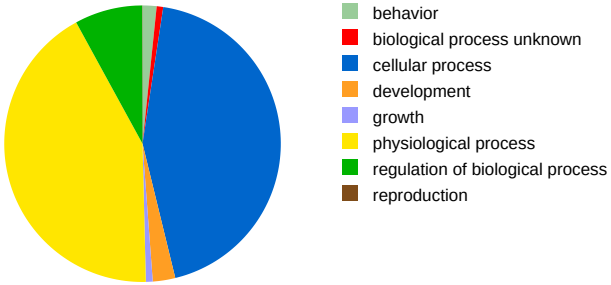
ecc − cellular component



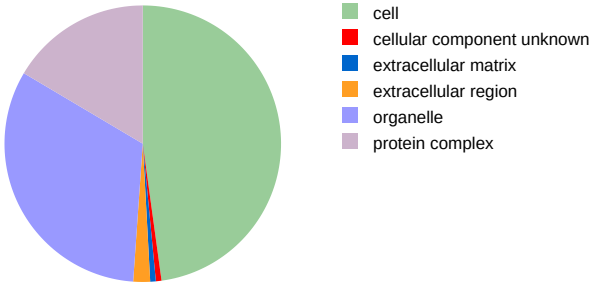
ecc − molecular function



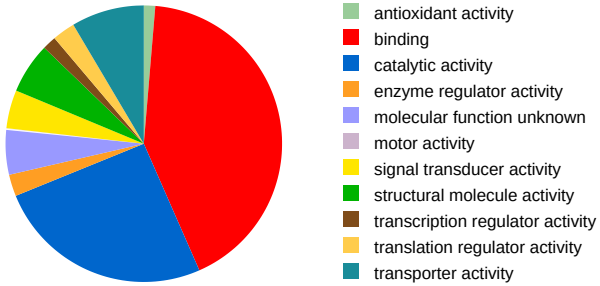
fco − biological process



fco − cellular component



fco − molecular function

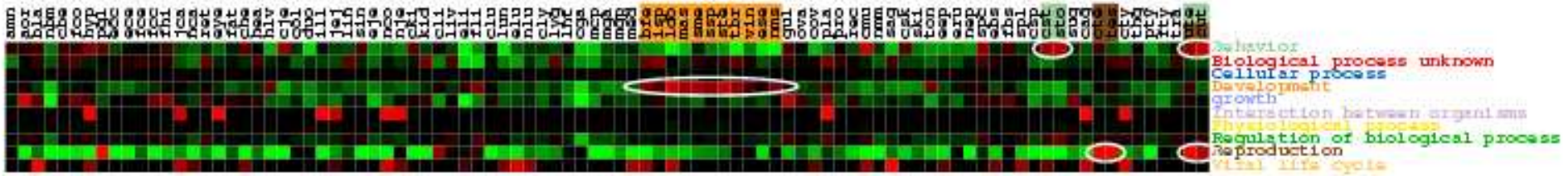


Gene ontology patterns

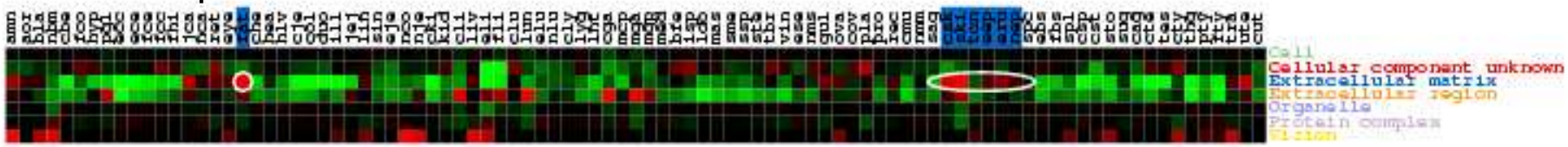
Comparing each library to an average pie using the relative entropy:

$$I = \sum_i q_i \log_2(q_i / p_i)$$

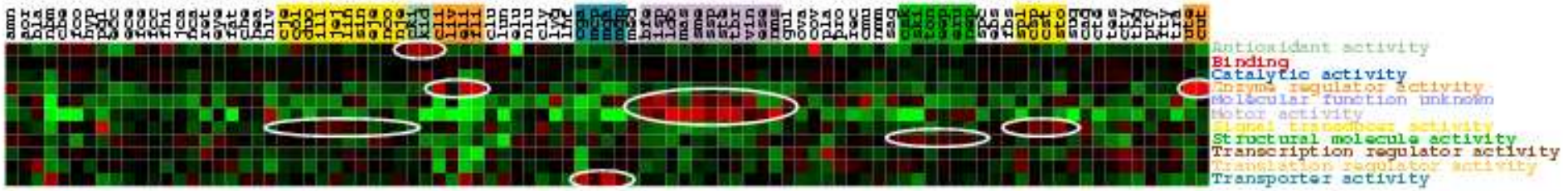
Biological process



Cellular component

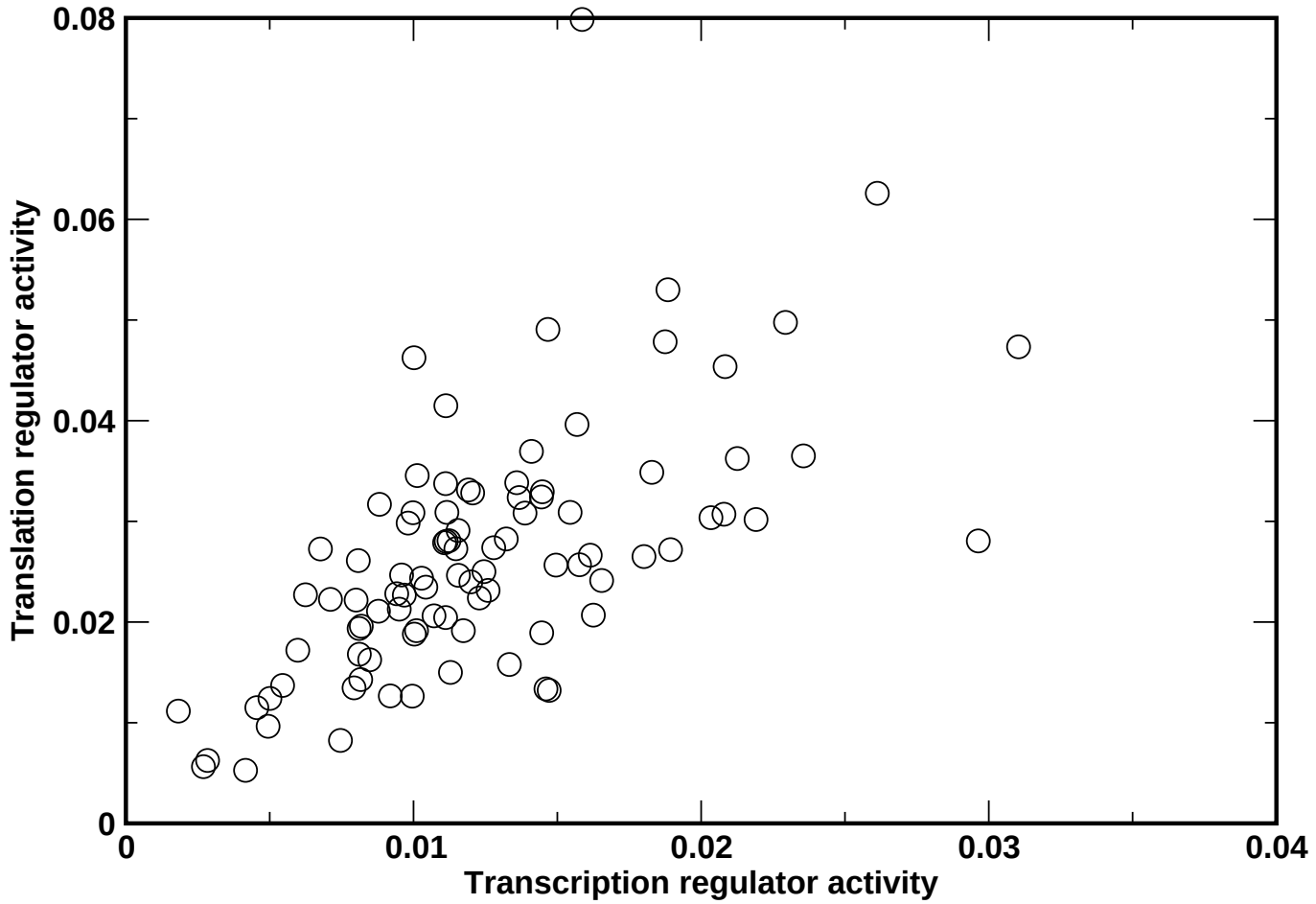


Molecular function



Gene ontology patterns

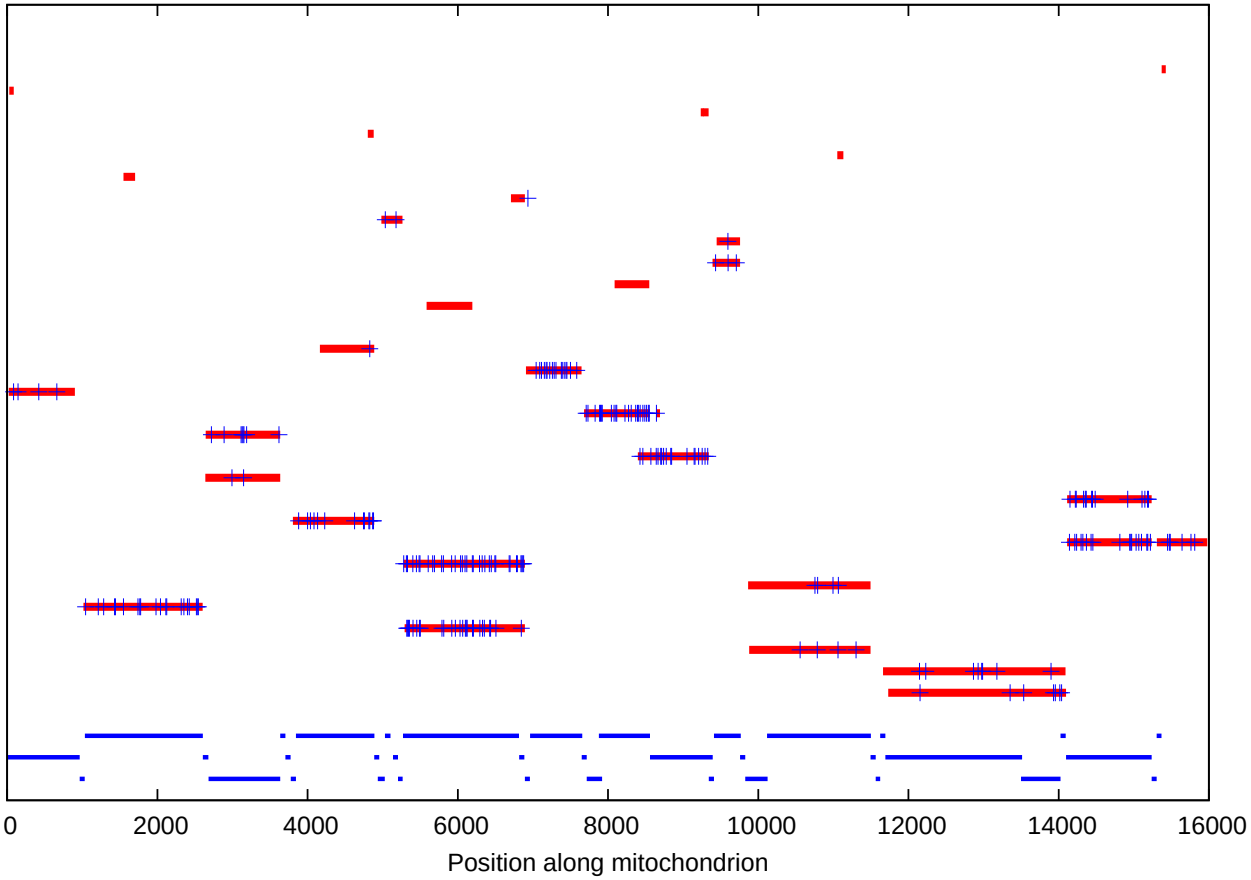
Correlations between gene ontology categories across libraries



Pig Mitochodrial ESTs

Expression and SNPs of the porcine mitochondrial genome

- Mitochondrial genome: 13 protein coding genes; 21 tRNAs; 2 rRNAs.
- 42000 ESTs (39K SD and 3K pub).
- Assembly (distiller): 32 Clusters.
- 255 SNPs (present in min 2 ESTs).
- Phylogenetic decomposition Euro-pean (Landerace) vs. Asian (Taihu).

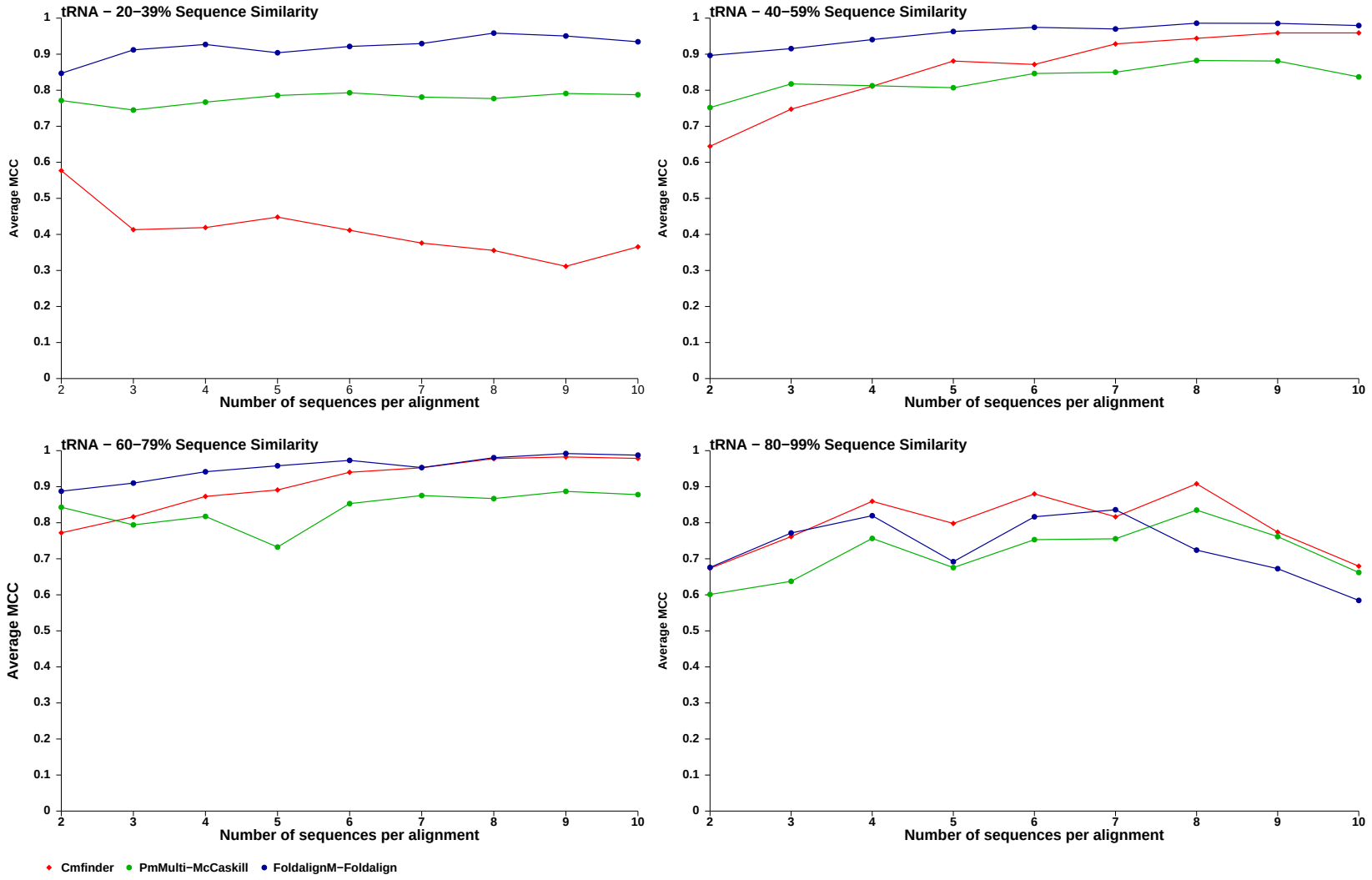


Multiple structural RNA alignments

- Multiple version of FOLDALIGN.
- Better use of FOLDALIGN pairwise scans.
- Aim for aligning a few low sequence similarity sequences.
- Based on the PMcomp Sankoff style implementation (Hofacaker,Stadler).
- Constrain sequences by their putative base pairs
  (base pair probability matrices).
- Use FOLDALIGN constraints from pairwise version (δ and pruning).
- Interested in clustering as well.

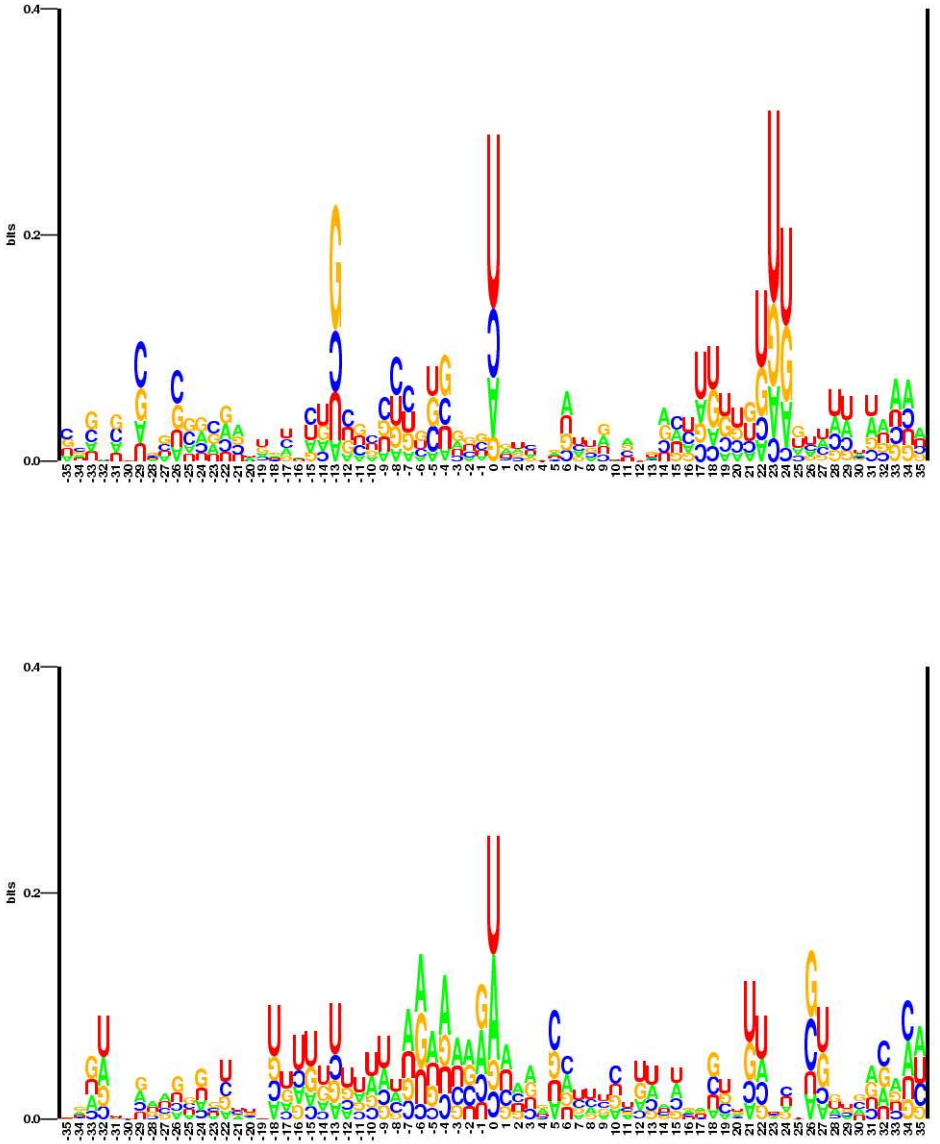
Multiple structural RNA alignments

Putative results (tRNA)



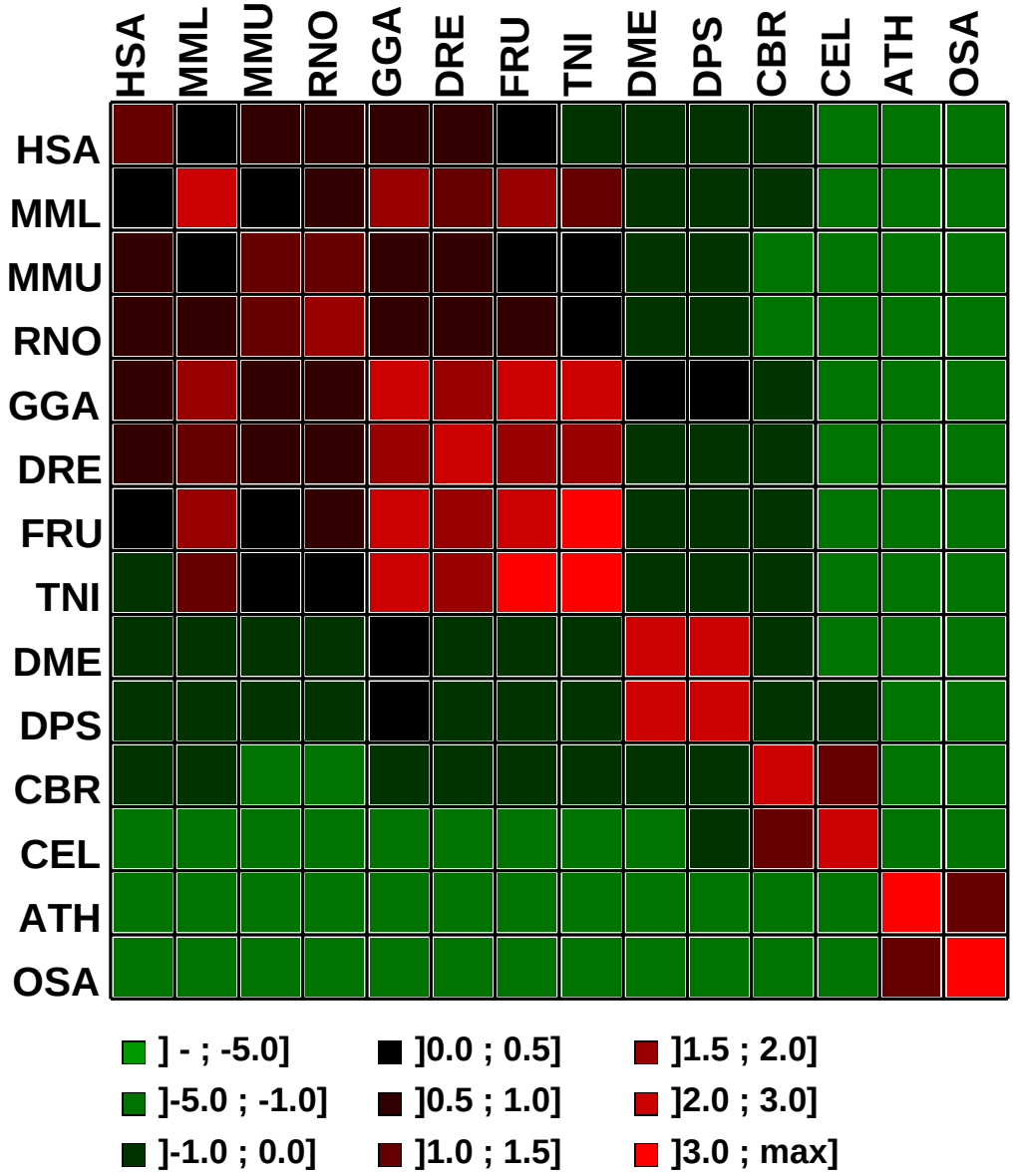
MicroRNA motifs

- 5’ and 3’ stem arm mature miRNAs are asymmetric (Gorodkin, *et al.*, 2006).
- Different using the ALLR score of Wang and Stormo.
- Many other observations about asymmetric processing of miRNA.



MicroRNA motifs

5’ arm mature



Scanning the ENCODE regions for ncRNAs using CMfinder

- CMfinder (Yao et al). Local RNA structure search on the ENCODE set.
- Motivation: If there are RNAs out there (not close to secret us military bases) to what extend is realignment needed to find the structure?
- Shuffling: Retain coarse grained pattern of conservation (only columns with mean pairwise id >0.5 and <0,5 were shuffled).
- Obtain 10,000 candidates (false positive rate of 50%).
- 5000 candidates show revisions in more than 50% of their aligned positions.
- The lower sequence similarity, the more revision.

Perspectives

- Still much to with PigEST.
- Better understanding of scaling invariance.
- Better handling of expression data (new methods...)
- Also SNPs (which I didn’t mention).
- For the other stuff RNA: Also stuff to do...