Evolutionary age of mouse brain-specific genes

~ Swagatam Mukhopadhyay

A work in progress
Collaborators

Partha Mitra, CSHL
Anirvan M. Sengupta (BioMaPS, Rutgers)
Pascal Grange, CSHL

Goals:

Use Allen Brain Atlas genes
Gene expression profile + Evolutionary age profile <=> Neuroanatomy
Information on neuron-specific-genes
Integrate data=> novel observations on mouse brain evolution?

Life of a gene

Genes are 'born' by

gene/domain duplication events: DNA repair (double strand breaks)
mobile genetic elements (retrotransposons, transposons)
... ?
Horizontal gene transfer (in prokaryotes)
Intron exon gain/loss by gene fusion/fission <=> alternative splicing

Genes 'grow' by

mutations vs. selective pressure (signature on synonymous vs. non-synonymous mutations)
Evolution rate can be determined for close by species (codon-usage bias complicates matters)

Genes are 'lost' by

gene duplication events
mobile genetic elements
...?

What is a 'gene' when only comparative genomics information is available?

Sequence similarity?

- Doesn't imply active function (but one can hope)
  - Genes share protein domains
  - Gene product = protein (alternative splicing)

Functional definitions

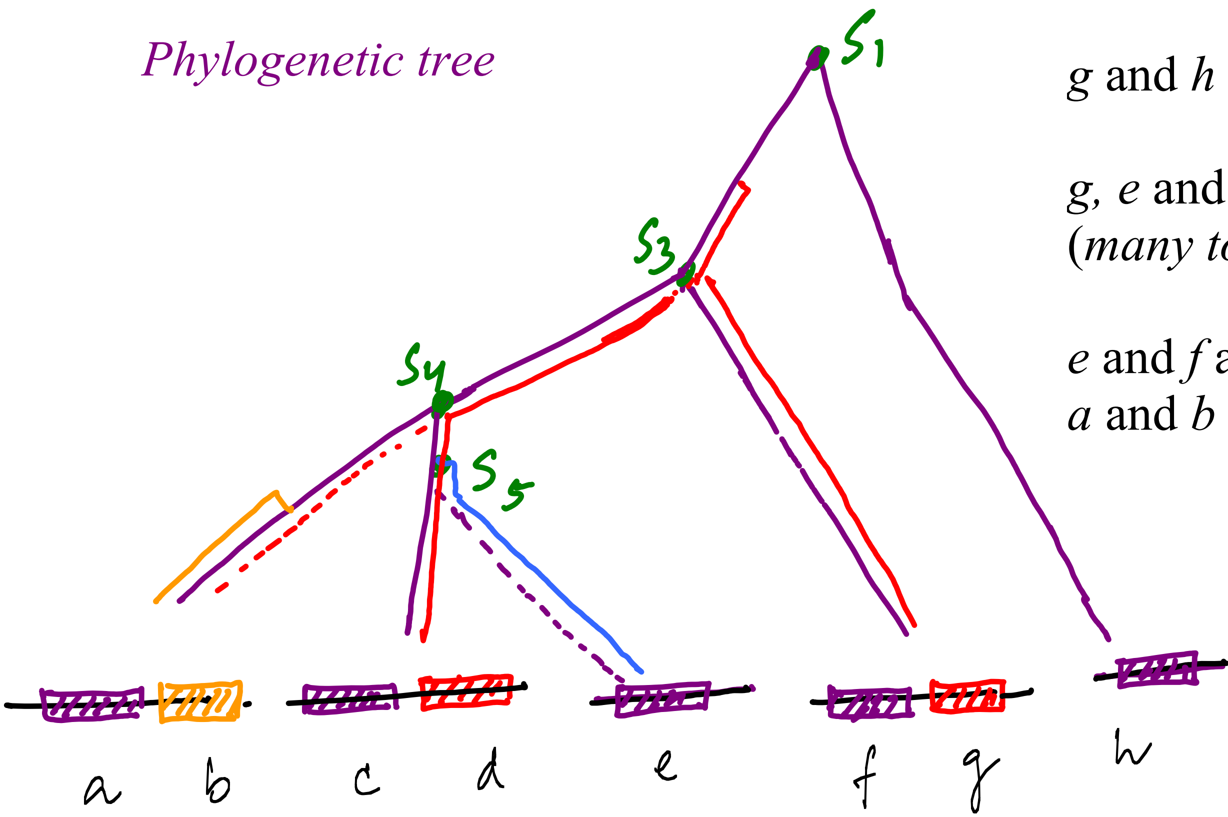
Homolog : atleast 30-35% sequence similarity in protein sequences in different species => common evolutionary ancestry of the genes

Ortholog : Homologous genes derived by a speciation event from a single ancestral gene in the last common ancestor of the species being compared

Paralog: Homologous genes derived from the duplication event of a single gene

Paralogs can have very different evolutionary rates
May not be functional
Non-functional copies' rate of mutation high because no selective pressure

Phylogenetic tree



g and *h* are orthologs w.r.t. speciation *S2*

g, e and *f* are orthologs w.r.t. speciation *S1*
(*many to one orthology*)

e and *f* are out-paralogs w.r.t *S3*
a and *b* are in-paralogs w.r.t. *S4*

In-paralogs: Paralogs that result from a lineage-specific duplication subsequent to a given speciation event
Out-paralogs: Paralogs resulting from a duplication preceding a given speciation event

Why important? Multiple gene loss can make an out-paralog pose as ortholog, thwarting speciation time estimation from sequence divergence, e.g., *e* and *a* falsely appear to be orthologs w.r.t. speciation *S4*

Orthology non-trivial to ascertain (glossing over issues)
Important for phylogeny

Common approaches for ortholog detection

Reciprocal best hits (RBH)

Two genes Xa and Xb from two genomes Ga and Gb

Xa and Xb are RBH iff

Recognizable similarity exists between them <= threshold similarity score
(Sequence alignment search tool like BLAST or Smith-Waterman)

There is no gene Zb in Gb that is more similar than Xb is to Xa

There is no gene Za in Ga that is more similar than Xa is to Xb

Bidirectional, uses Xa to first query Gb <=> Xb to query Ga.

Reciprocal smallest distance (RSD)

Use sequence alignment tool like BLAST to find a set Hb hits of Xa against Gb

Refine Hb by aligning the protein sequences in Hb with Xa

Alignable region must exceeds a thresholding fraction of the alignment's total length

Use alignment and a suitable algorithm (e.g., PAML) for a maximum likelihood estimate of number of amino acid substitutions (distance) separating protein sequences (use empirical amino acid substitution rate matrix)

Retain only Xb from Hb which has the smallest distance from Xa

Use Xb to perform reciprocal BLAST against genome Ga to obtain hits Ha

If the original query sequence Xa is in Ha *and* if Xa is the sequence with the smallest distance to Xb compared to all other sequences in Ha, then Xa and Xb are declared to be true orthologous pairs

Stable pairs (we use this)

Stable pairs were introduced and defined by the OMA algorithm/project

All pairwise alignments between genome Ga and genome Gb for all genes are performed using Smith-Waterman dynamic programming using a fixed amino-acid substitution (PAM) matrix

Length-tolerance criterion of the shorter aligned sequence

Genes with significant alignment scores are retained as candidate pairs

Evolutionary distance between sequences is computed (PAM estimate)

Similar to the reciprocal smallest distance a distance tolerance is introduced

Further refined to by screening out-paralogs (ancient duplication event -> speciation events + loss of a single duplicate in both species)

Merits of OMA?

Uses Smith Waterman instead of BLAST
One thousand species! Dataset of all orthologs detected for all genes available
Colossal computational effort in checking distance criterions for orthology

oma browser

The OMA Browser is a web-based interface to the data from the OMA project of the CBRG at the ETH Zürich. It offers a comprehensive search and numerous display options for 4.7 million proteins from 1000 species. The main features are the orthologous relationships which can be accessed either group-wise, where all group members are orthologous to all other group members, or on a sequence-centric basis, where for a given protein all its orthologs in all other species are displayed. An application note in *Bioinformatics* describes the main features of the OMA Browser:

A Schneider, C Dessimoz and GH Gonnet (2007): **OMA Browser - Exploring Orthologous Relations across 352 Complete Genomes**, Bioinformatics 23(16), pages 2180-2182. [Open Access]

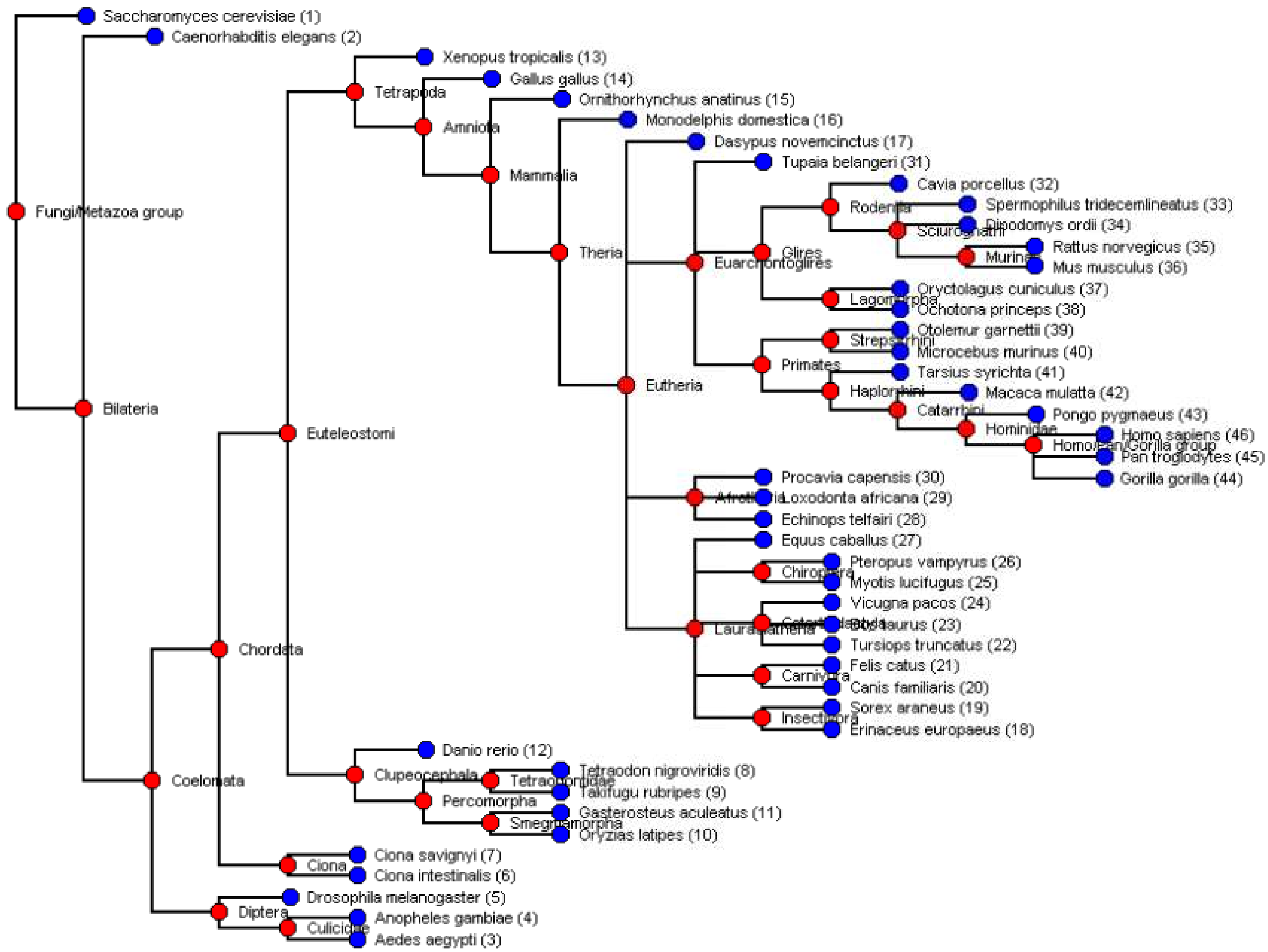
Recent developments are described in 2011's *Nucleic Acids Research Database Issue*:

AM Altenhoff, A Schneider, GH Gonnet and C Dessimoz (2011): **OMA 2011: orthology inference among 1000 complete genomes**, Nucl. Acids Res., 2011, 39(suppl 1): D289-D294. [NAR: Open Access]

But how do we put a measure of age on genes from knowing their orthologs in other species?

Around 3000 brain specific genes that has high correlation in coronal and sagittal sections of mouse brain

Use established phylogenetic tree
Use robust analysis

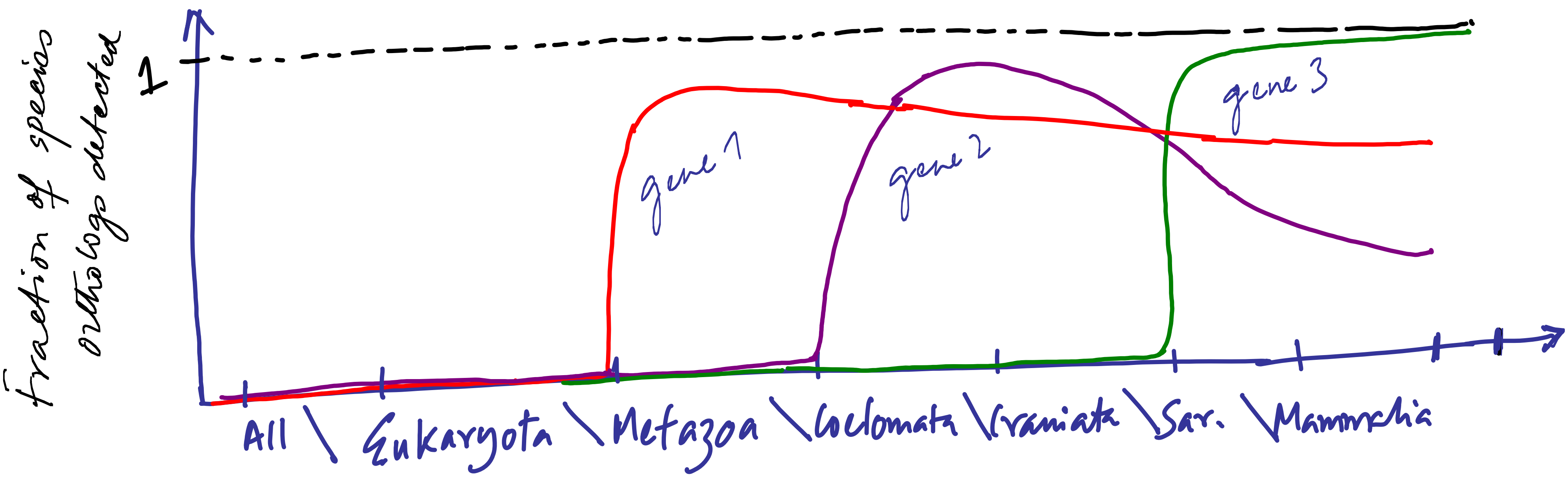


OMA has 108 Eukaryotes

Define an order of progressively refining phylogenetic clades

All > Eukaryota > Metazoa > Coelomata > Craniata > Sarcopterygii > Mammalia > Euarchontoglires > Rodentia

Look for fraction of hits of each mouse gene on all species of each such *subsets*
The difference subset for which the fraction of hits is significant translates to a gene age



=> gene1 older than gene2 and gene2 older than gene3

All > Eukaryota > Metazoa > Coelomata > Craniata > Sarcopterygii > Mammalia > Euarchontoglires > Rodentia

Coelomata has coelom
Chordata has notochord
Sarcopterygii are fleshy- and lobe-finned vertebrates (include tetrapods)
Craniata has skull
Euarchontoglires is a clade of mammals: living members of are rodents, lagomorphs, treeshrews, colugos and primates

#total species = 1054
#total non-Eukaryota = 946
#total Eukaryota / Metazoa = 38
#total Metazoa / Coelomata = 8
#total Coelomata / Chordata = 13
#total Chordata / Vertebrata = 3
#total Vertebrata / Sarcopterygii = 5
#total Sarcopterygii / Mammalia = 5
#total Mammalia / Eutheria = 3
#total Eutheria / Euarchontoglires = 16
#total Euarchontoglires / Rodentia = 12
#total Rodentia = 5

Complications?

108 species: under- and biased-sampling in genome sequenced (we love our primates, farm and pet animals)
Poorly sampled: birds, fishes, reptiles
Poorly sampled lower vertebrates
We choose the clades with a balance between an effort to sample well and brain evolutionary nodes of interest

Genes are born and lost => not `step function' profile
Fit the profiles?

$$f(s) = \frac{\mathrm{Exp}\left[-\mu\,(s - T)\right]}{1 + \mathrm{Exp}\left[-(s - T)/b\right]},$$

Exponential decay function captures gene loss, multiplying a sigmoidal function that captures rate of 'fixation' of gene on the tree

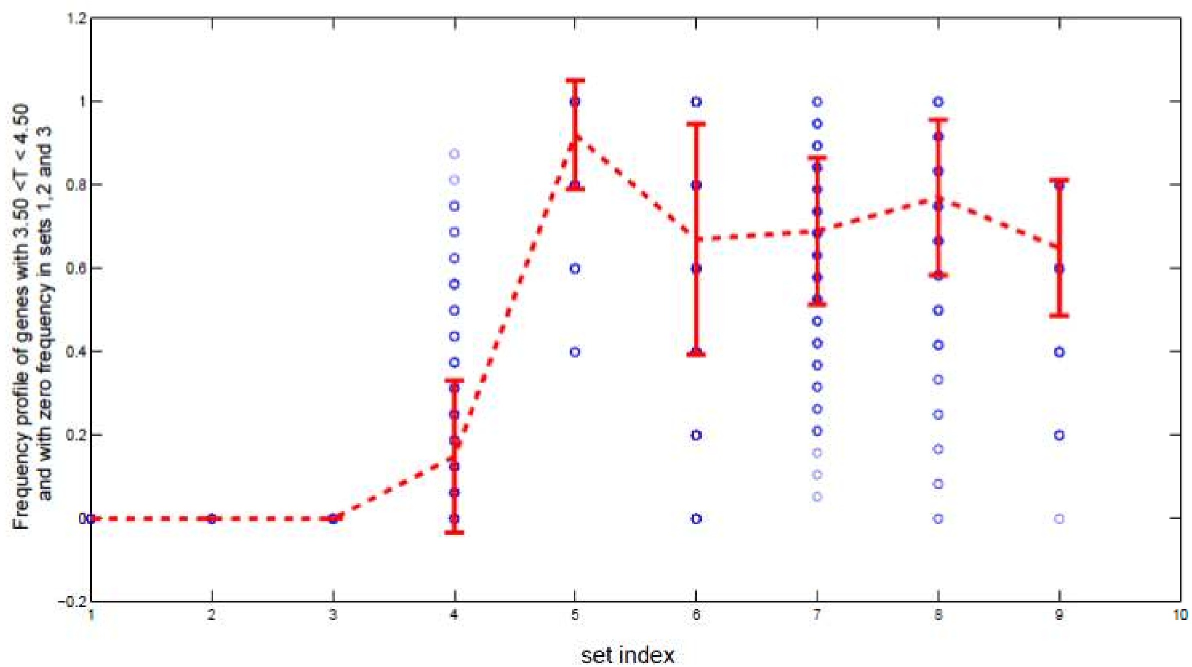
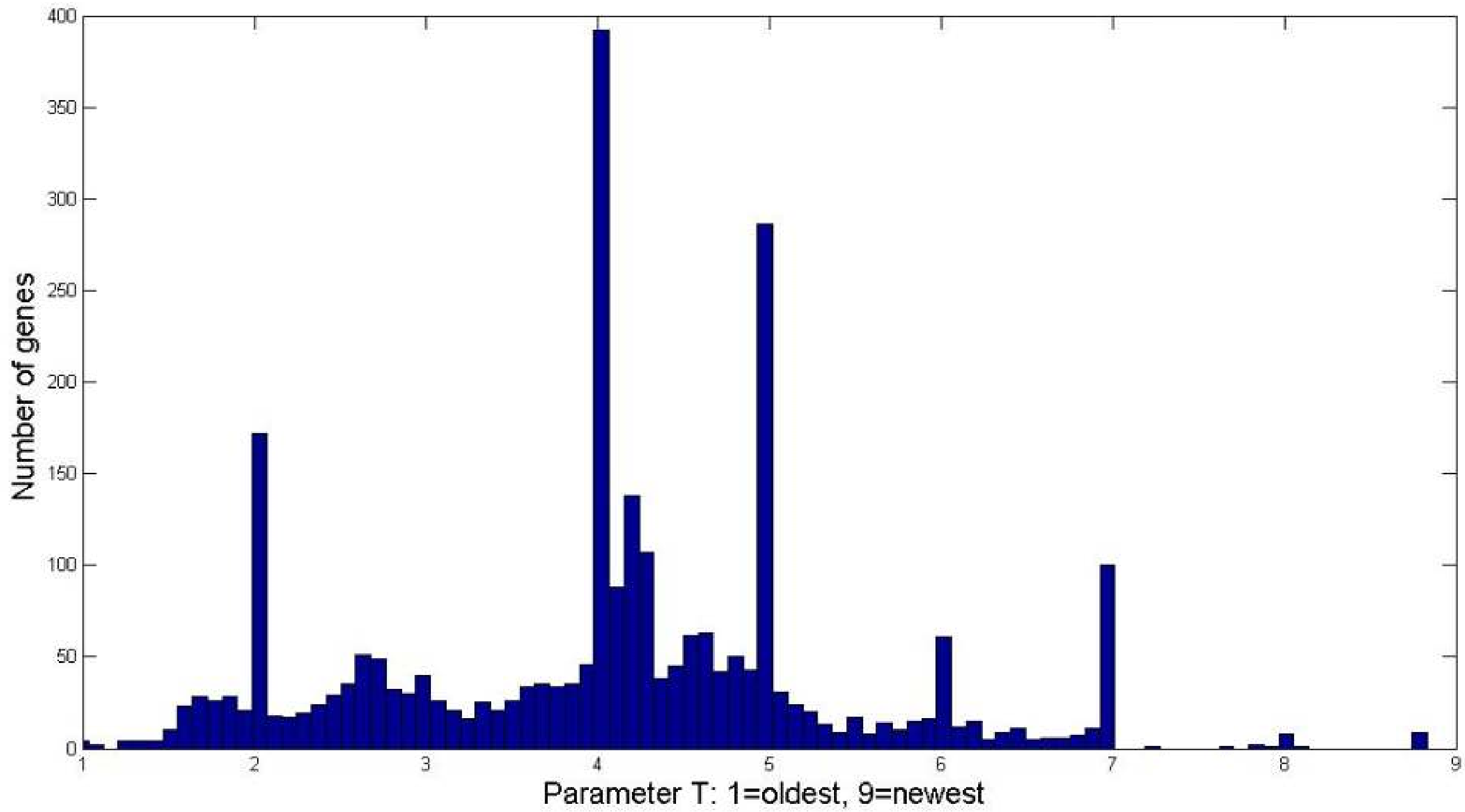
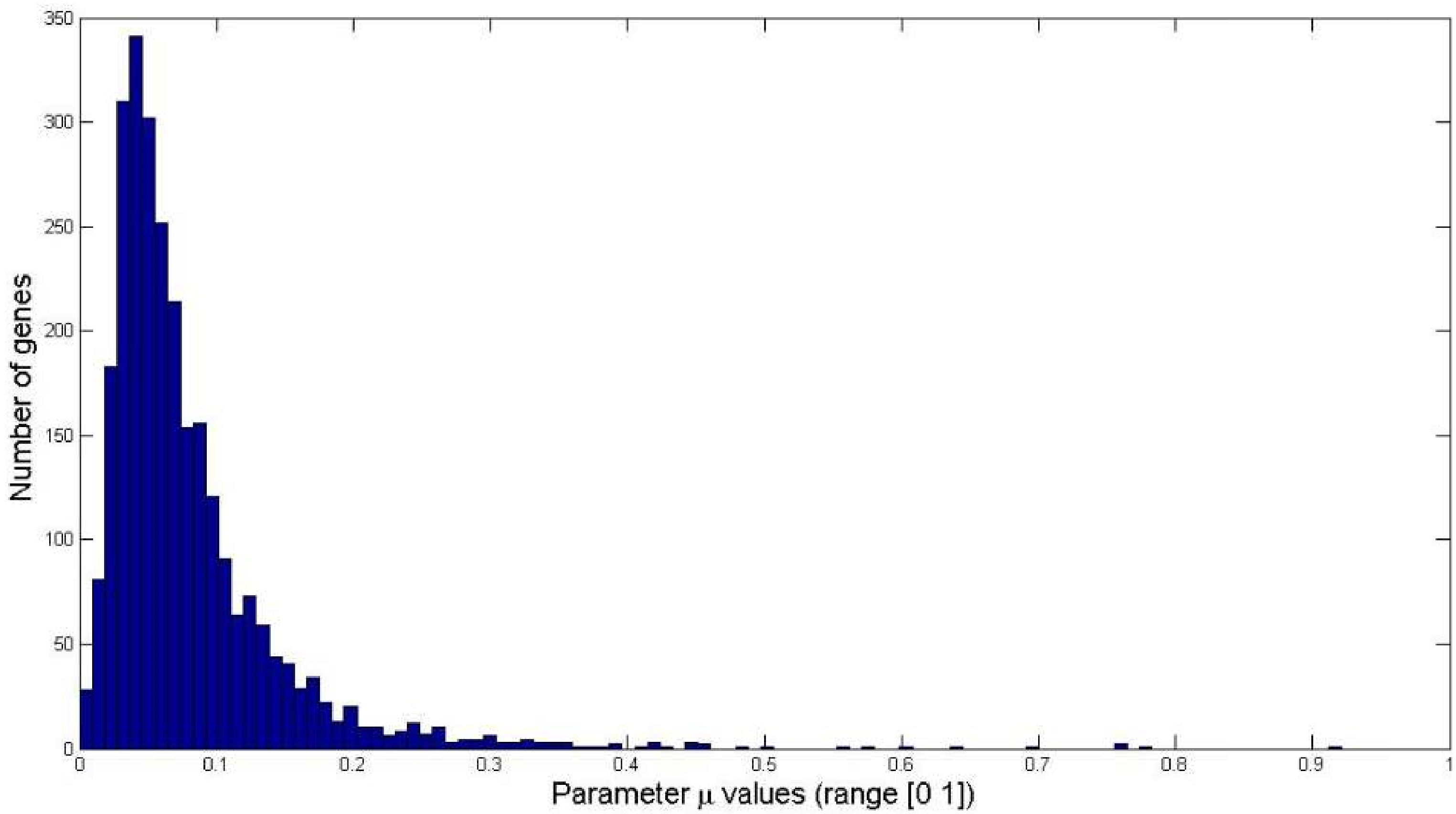


Figure 4: Average profile of genes novel in fourth clade



Paramter T determines where the knee of the sigmoidal function is. A lot of genes show a sharp integral T

$$f(s) = \frac{\mathrm{Exp}\left[-\mu\,(s-T)\right]}{1+\mathrm{Exp}\left[-(s-T)/b\right]},$$



The variations in gene loss rate

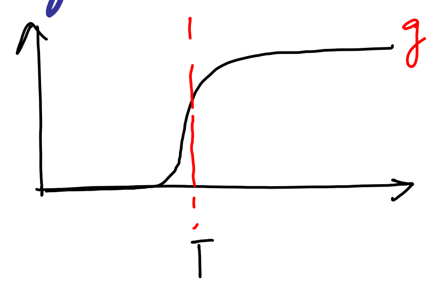
$$f(s) = \frac{\mathrm{Exp}\left[-\mu\,(s-T)\right]}{1+\mathrm{Exp}\left[-(s-T)/b\right]},$$

Allen dataset poor-man's summary

Mouse brain partitioned into 50,000 voxels of 200 microns size
Gene expression of >20000 genes, *in situ* hybridization
Registration to a reference atlas, three dimensional frame acheived through coronal and sagittal section
Anatomical annotation of each voxel allows exploration of region specific gene expression

$E(v,g)$ ← gene expression strength of gene g at voxel v

G_T = genes with age frequency profile T



$$E_T(v) = \sum_{g \in G_T} E(g,v)$$

$$E_{tot}(v) = \sum_{g=1}^{G} E(g, v)$$

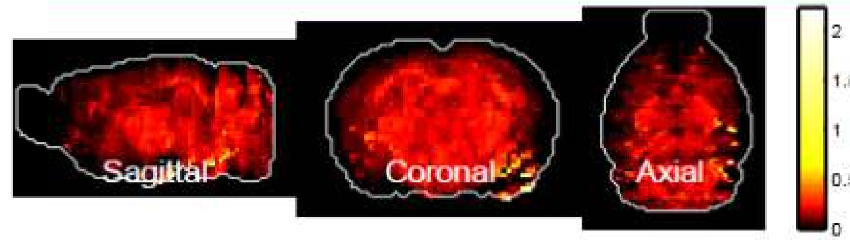
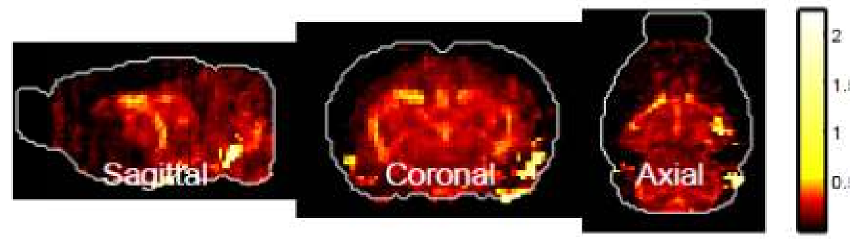
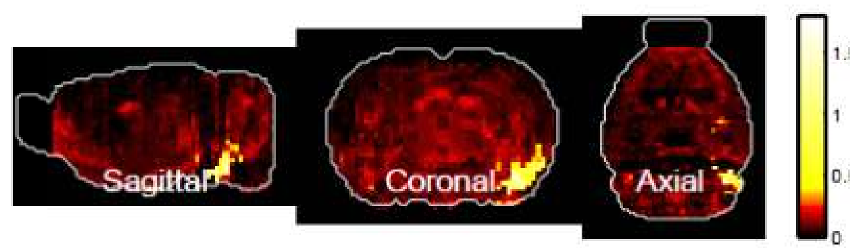
all genes

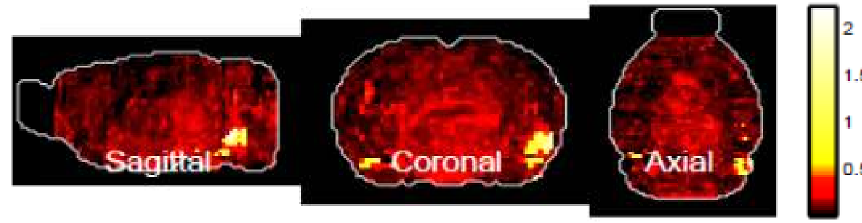
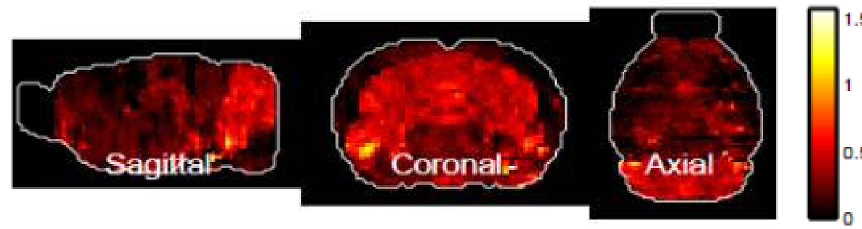
$$L_T(v) = \log\left(\frac{E_T(v)}{E_{tot}(v)}\right)$$

We present this voxel by voxel measure of age specific genes expressions over the background pattern

If this has any structure, then the corresponding region developed around that evolutionary time

Preliminary results of age-selected-gene expression profile log ratios in Allen dataset

Coelomata Non Chordata (13)	52	
Chordata Non Vertebrata (3)	88	
Vertebrata Non Sarcopterygii (5)	442	
Sarcopterygii Non Mammalia (5)	328	

Set of species (number)	♯ genes	log of ratio to sum
Mammalia Non Eutheria (3)	70	
Eutheria Non Euarchontoglires (16)	35	
Euarchontoglires Non Rodentia (12)	47	

Ask Pascal Grange!

Works that remain to be done

Distilling qualitative information from analysis

Check noise: robustness of patters observed?

Similar analysis of neuron type specific genes and integrate these three sources

Once other gene expression atlas datasets are available, comparative neuroanatomy and brain evolution possible at the gene expression level