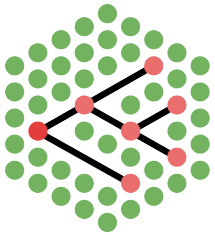
Positive Selection in Mammals

When, where, why?

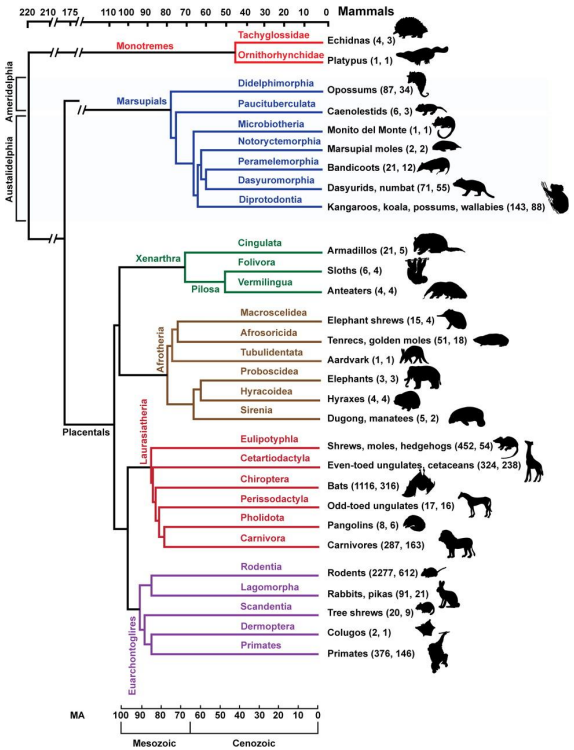
Gregory Jordan



Goldman Group — EMBL-EBI

EBI Postdoc Seminar, February 2012

Evolutionary History of Mammals



adapted from Haussler et al. *J Hered* 2009

- Common ancestor 165-170 Mya
- 100-85 Mya – major extant orders have diverged
- History affects tree shape:
  - Long external branches
  - Deep but short internal branches
  - More mammalian genomes $\Rightarrow$ good bang for the buck

Evolutionary History of Mammals

Sketch of a possible Cretaceous primate ancestor (80 Mya)

adapted from Martin et al. (2007) *Folia Primatol*

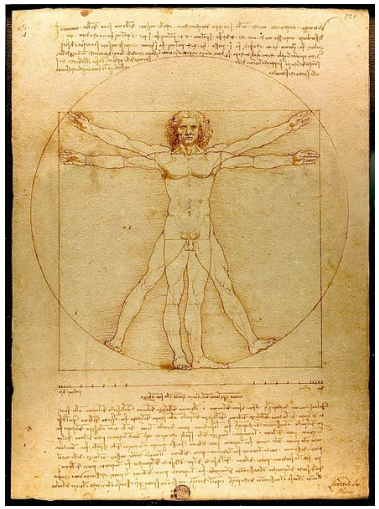
- Ancestral mammals were likely:
  - Smaller than today
  - Largely insectivorous
- General trend towards larger body, more specialized niche
- Life history can affect molecular evolution
  - Body size $\Rightarrow$ mutation rate
  - Sexual behavior $\Rightarrow$ sexual dimorphism, adaptation of gametogenesis genes

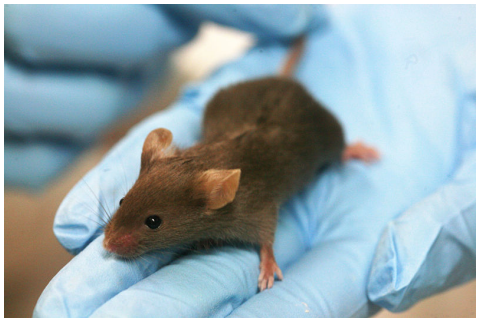
Mammalian Genomes

- Mammalian genomes are big
  - 2.5-4.5 Gb of DNA, 20-80 chromosomes
- Relatively extensive structural evolution
  - Chromosomal rearrangements
  - Gene family expansion / contractions
  - Transposable elements
- Stable set of core genes
  - 20,000 genes
  - 80% one-to-one with mouse, 82% detectable homology with platypus

Wikipedia, public domain.

Mammalian Genomics





Wikipedia, public domain / CC-share-alike.

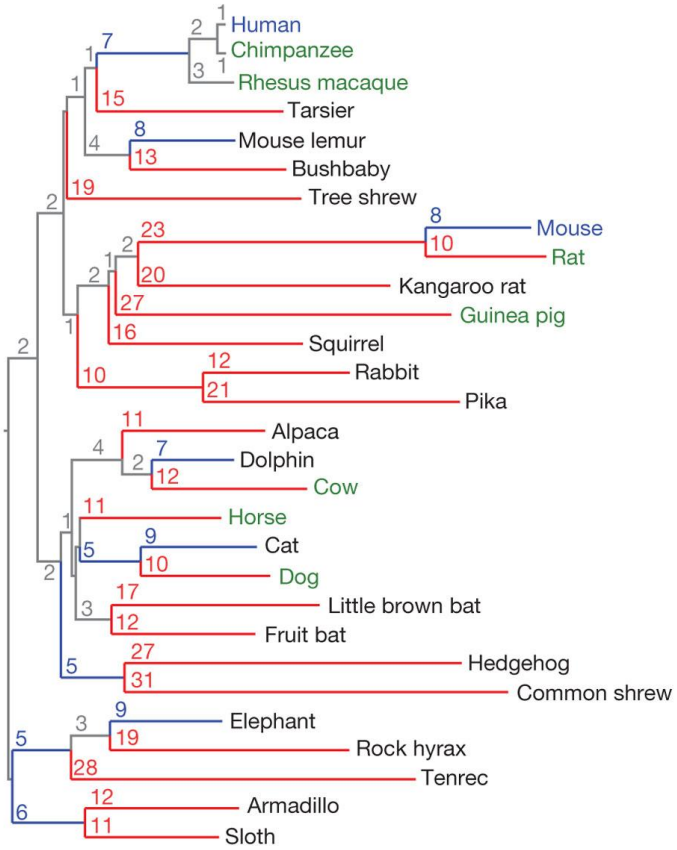
- First human (2001)
- Then, mouse (2002)
- ... and rat (2004), chimpanzee, dog (2005), rhesus (2007), platypus (2008), cow, horse (2009), orangutan (2011)
- Big papers, major insights (5% conserved, genome-wide dN/dS, TE dynamics)

A quote

"Genome comparison is a powerful tool for discovery. It can reveal unknown—and even unsuspected—biological functions, by sifting the records of evolutionary experiments that have occurred over 100 years or over 100 million years. The _____ genome sequence illustrates the range of information that can be gleaned from such studies."

K. Lindblad-Toh et al. (2005). Genome sequence, comparative analysis and haplotype structure of the domestic dog.

The Mammalian Genome Project



from Lindblad-Toh et al. (2011) *Nature*

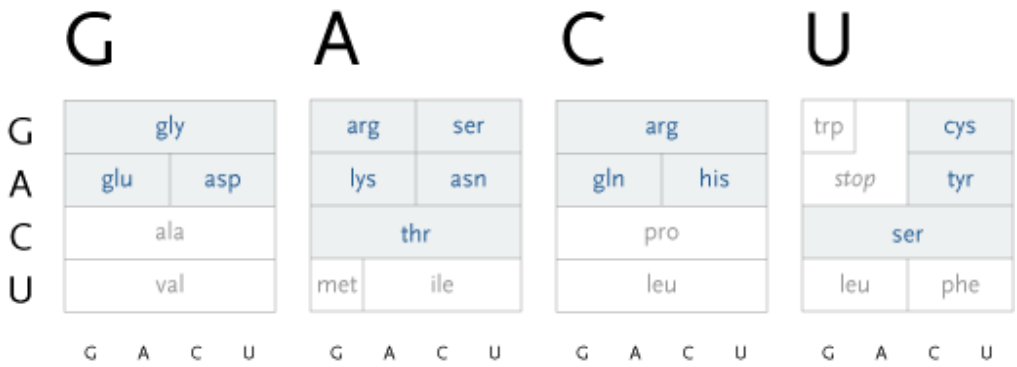
- Massively increase sequenced branch length
- Emphasis on *increased power* resulting from added sequence
- Project was comparative from the start
- BUT: genomes are low-quality (mean 2x read depth, 85% of genome covered)

Mutation and Selection

- DNA mutations occur randomly
  - in humans: $1x10^{-}8$ mutations per site per generation, or 70 per person [1]
- Over time, *Natural selection* in a population:
  1. removes deleterious alleles
  2. promotes beneficial alleles
- To detect beneficial / deleterious alleles: compare evolutionary rate to a *neutral* reference
- Same principle underlying:
  1. GERP / PhastCons elements
  2. Human-accelerated regions
  3. *d*N/*d*S in protein-coding genes

[1] Keightley, PD. (2012) *Genetics* 190: 295-304

Natural Selection in Proteins



- Degeneracy of genetic code
  - ”silent” mutations $\Rightarrow$ *neutral* reference
- Detect non-neutral selection acting on protein-altering mutations
- Compare nonsynonymous rate (dN) to synonymous rate (dS)
  - $d\text{N} \approx d\text{S} \Rightarrow$ neutral evolution
  - $d\text{N} < d\text{S} \Rightarrow$ conservation / negative selection
  - $d\text{N} > d\text{S} \Rightarrow$ positive selection

Detecting Selection with Codon Models

$$q_{ij} = \begin{cases} 0, & \text{if } i \text{ and } j \text{ differ at two or three positions,} \\ \pi_j, & \text{if } i \text{ and } j \text{ differ by a synonymous transversion,} \\ \kappa\pi_j, & \text{if } i \text{ and } j \text{ differ by a synonymous transition,} \\ \omega\pi_j, & \text{if } i \text{ and } j \text{ differ by a nonsynonymous transversion,} \\ \omega\kappa\pi_j, & \text{if } i \text{ and } j \text{ differ by a nonsynonymous transition,} \end{cases}$$

- Model the evolution of DNA sequences in terms of codons
  - Incorporate $\omega = d\text{N}/d\text{S}$ as a model parameter
  - Find parameter values which best fit the data (sequence alignment)
- Likelihood ratio tests (LRTs) for testing when $\omega \neq 1$ produce:
  - The strength of positive or purifying selection (ω)
  - The strength of *evidence* for said selection

Reality Check

- Wait... why are we interested in positive selection anyway?

Previous Genome-Wide Studies

- 2003, 2005: scans in human-chimp-mouse and human-chimp using LRTs

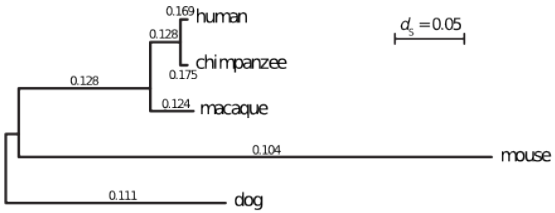
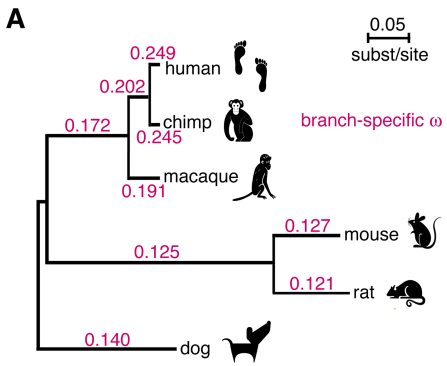


Figure S6.2: An estimate of ω for each branch of a five-species phylogeny. Show is the maximum-likelihood phylogeny for 5286 orthologous quintents, with branch lengths drawn in proportion to the estimated number of synonymous substitutions per synonymous site (d_S). Each branch is labeled with the corresponding estimate of ω.

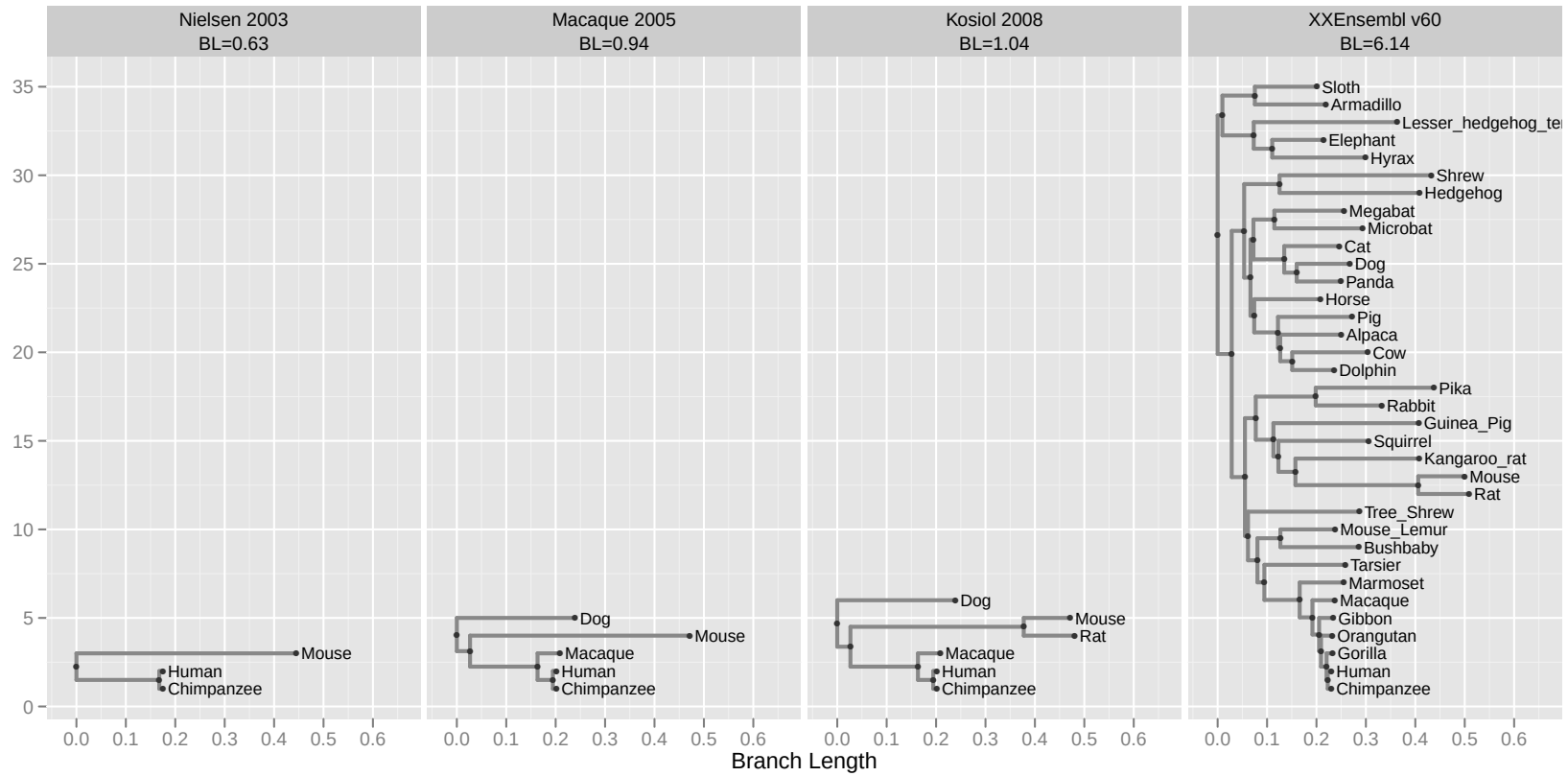


- Rhesus genome analysis (2007) - LRTs for positive selection in 6 mammalian genomes
- Kosiol et al. (2008) - LRTs in 6 mammalian genomes

Patterns of Positive Selection

- Found many positively-selected genes (PSGs) related to:
  - Sensory perception (olfaction, hearing)
  - Apoptosis and spermatogenesis
  - Iron binding and keratin formation
- Presumed causative factors:
  - Host-pathogen ”evolutionary arms race”
  - Sexual selection / genetic conflict
  - Functional adaptations (gene- or organism- level)

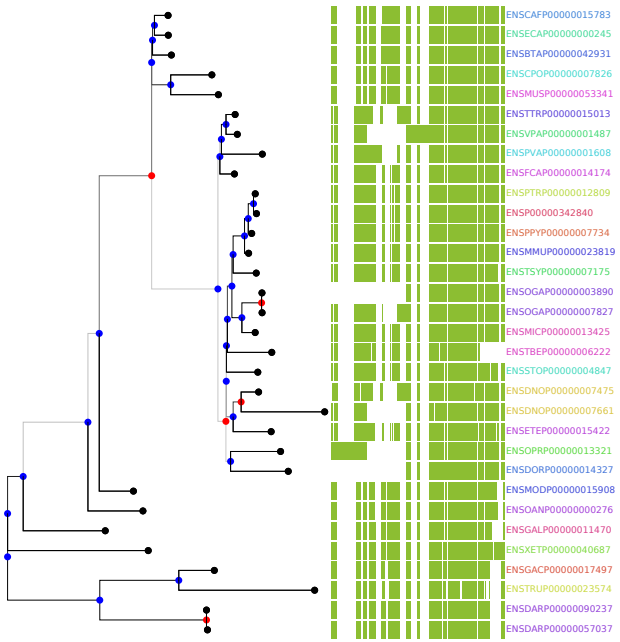
The Major Difference Now: Scale

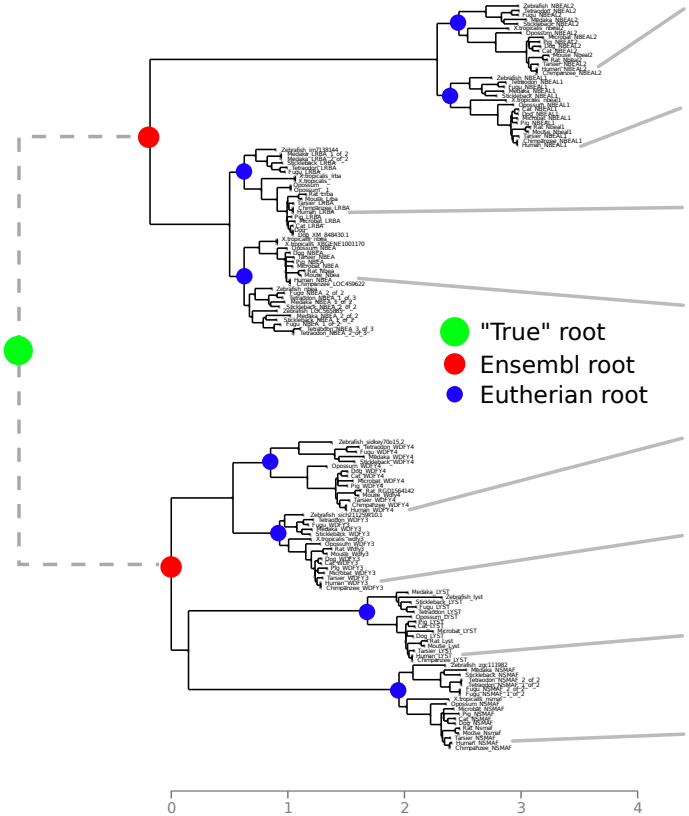


Questions to be Investigated

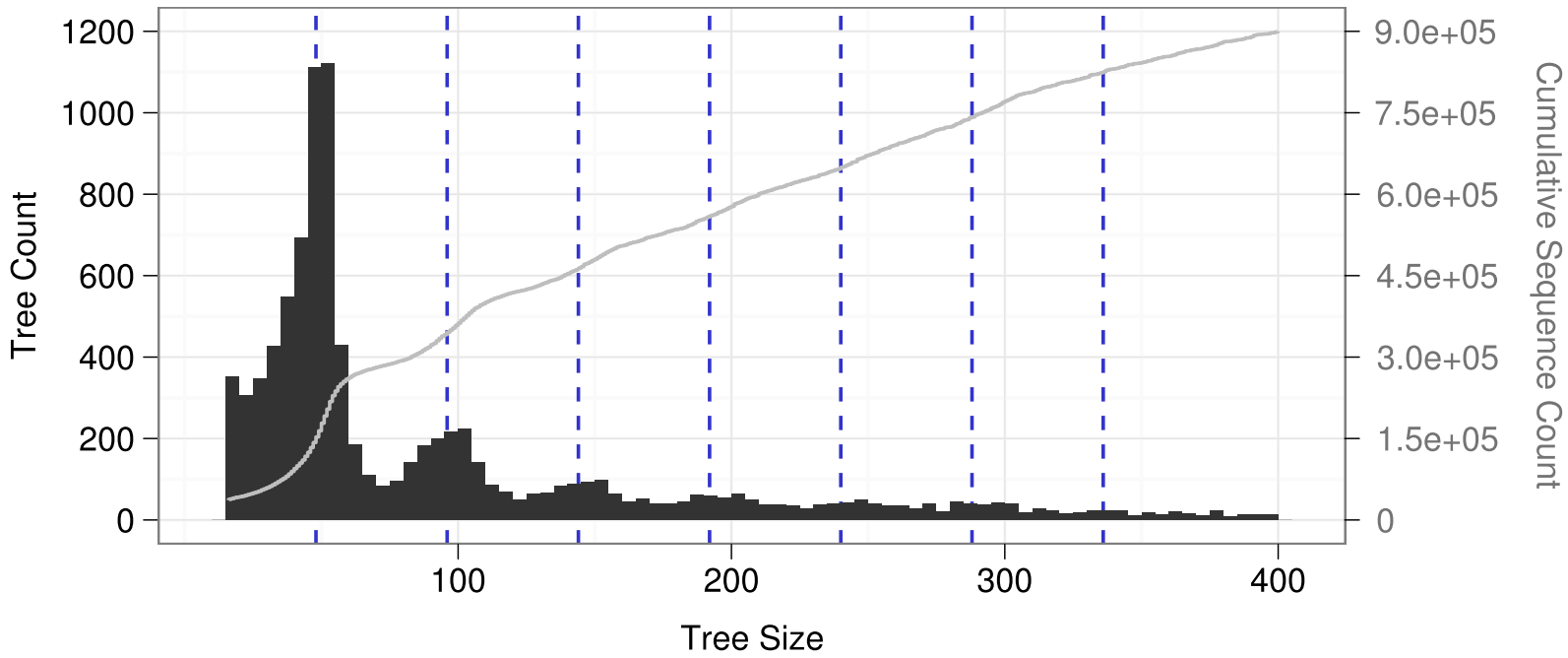
- Technical questions:
  - Are low-coverage genomes a deal-breaker?
  - How much power comes from additional branch length?
- Biological questions:
  - Do different mammalian orders show similar or different patterns?

Ancient Duplications in Ensembl Trees

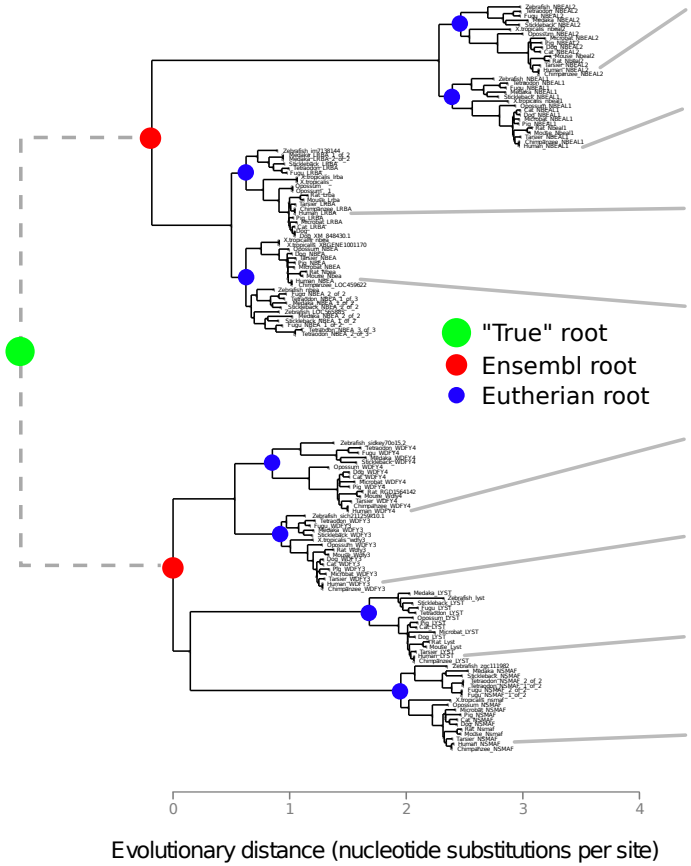




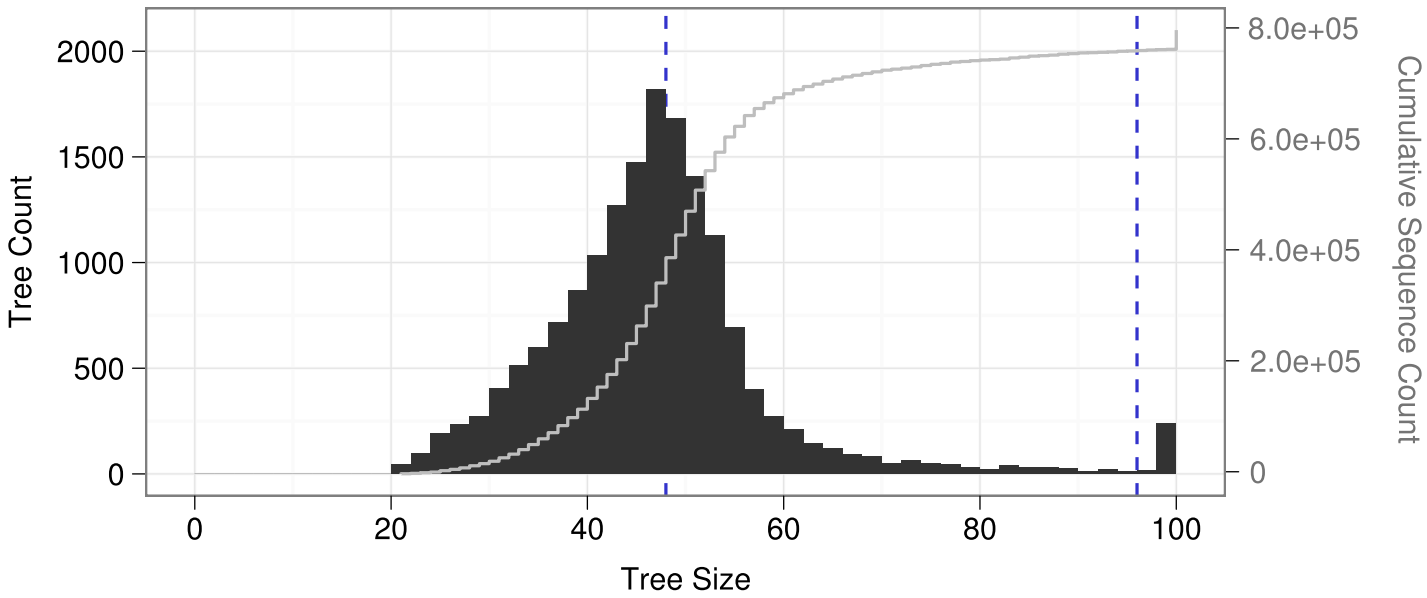
Ancient Duplications in Ensembl Trees



Identifying ”Largely-Orthologous Trees” with Taxonomic Criteria

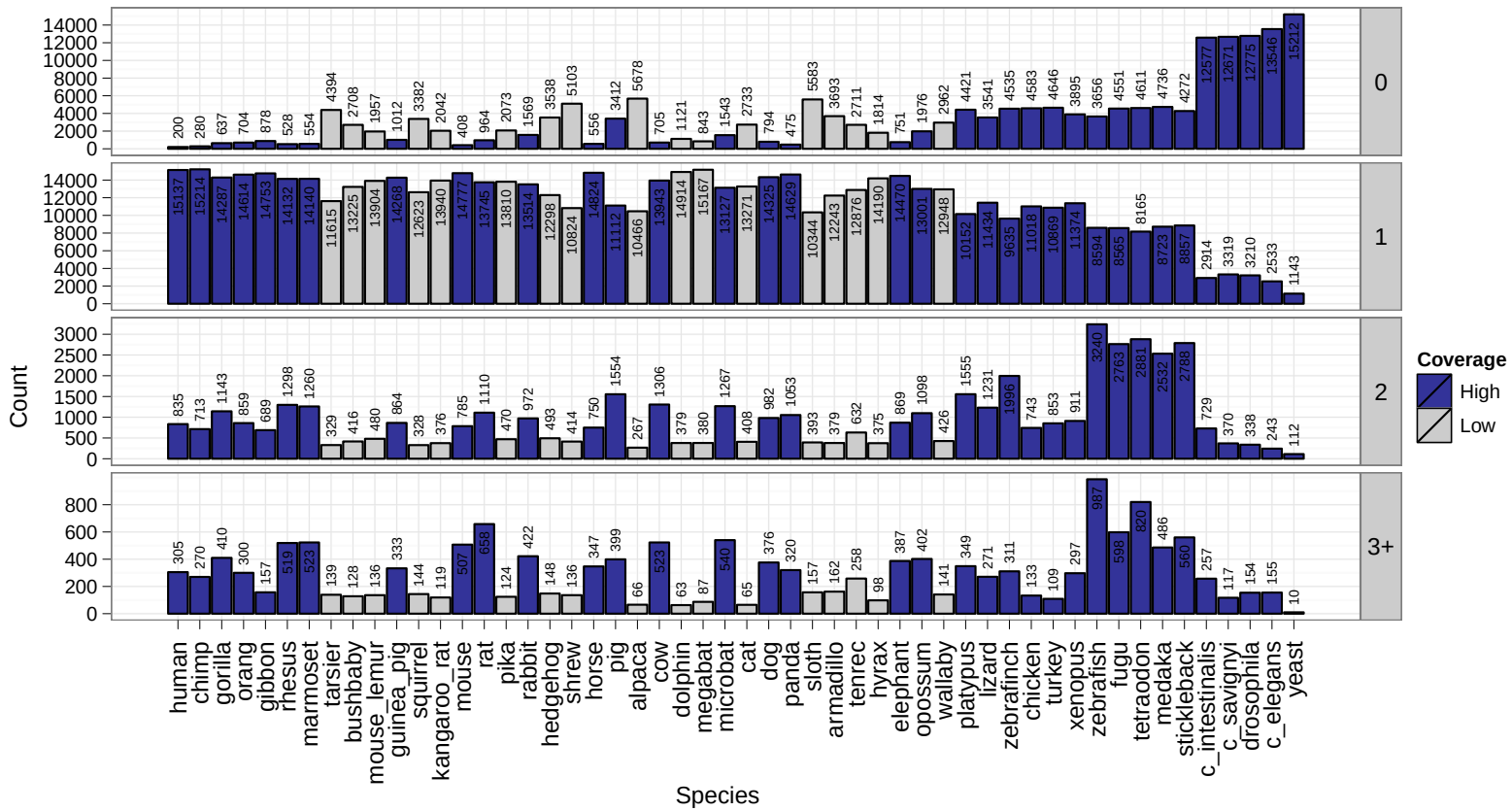


”Largely-Orthologous Trees”: Results

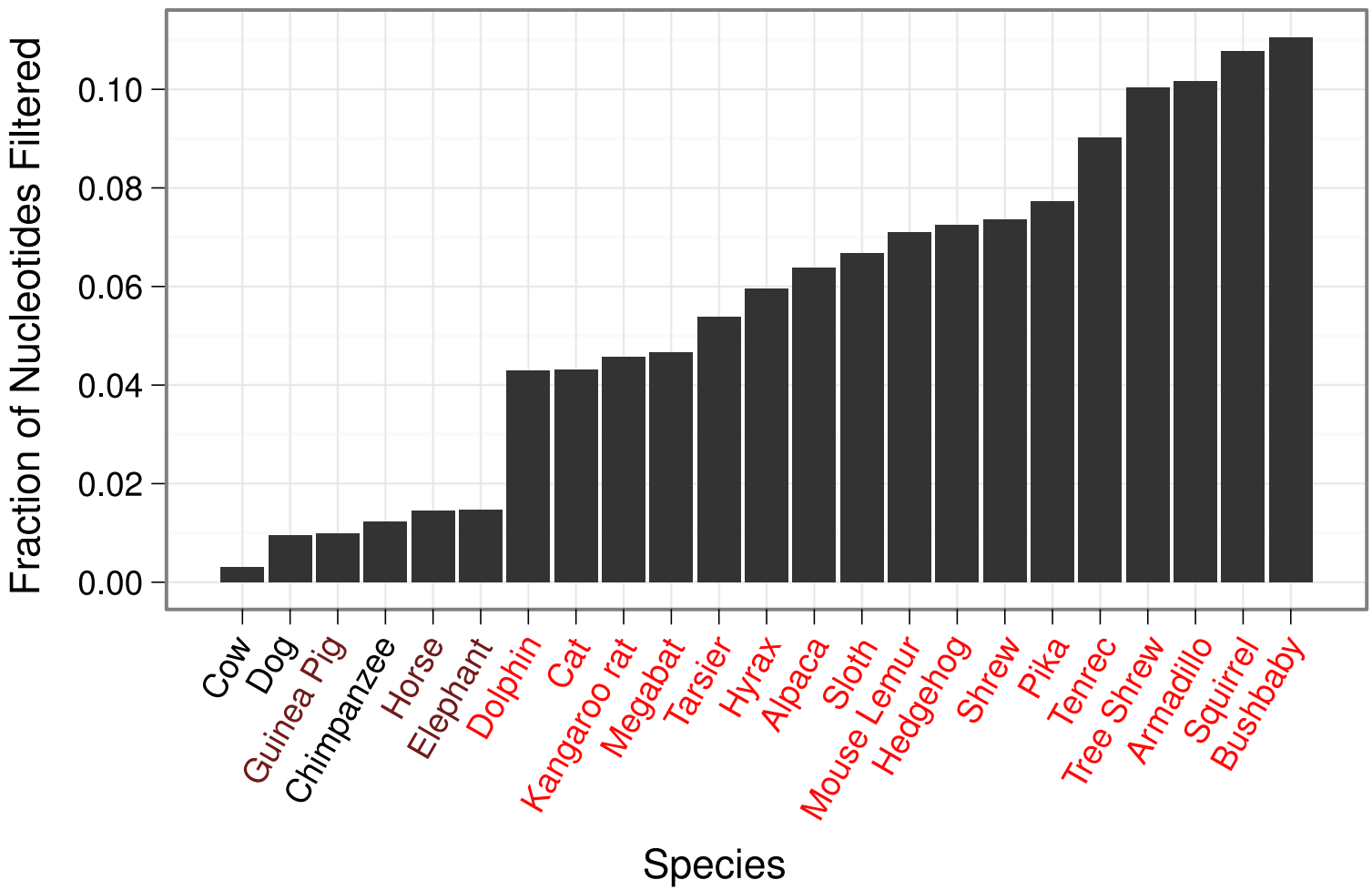


- 16,477 largely-orthologous trees in Eutherian mammals
- Smooth distribution of tree sizes
  - ”Long tail” of trees with family expansions

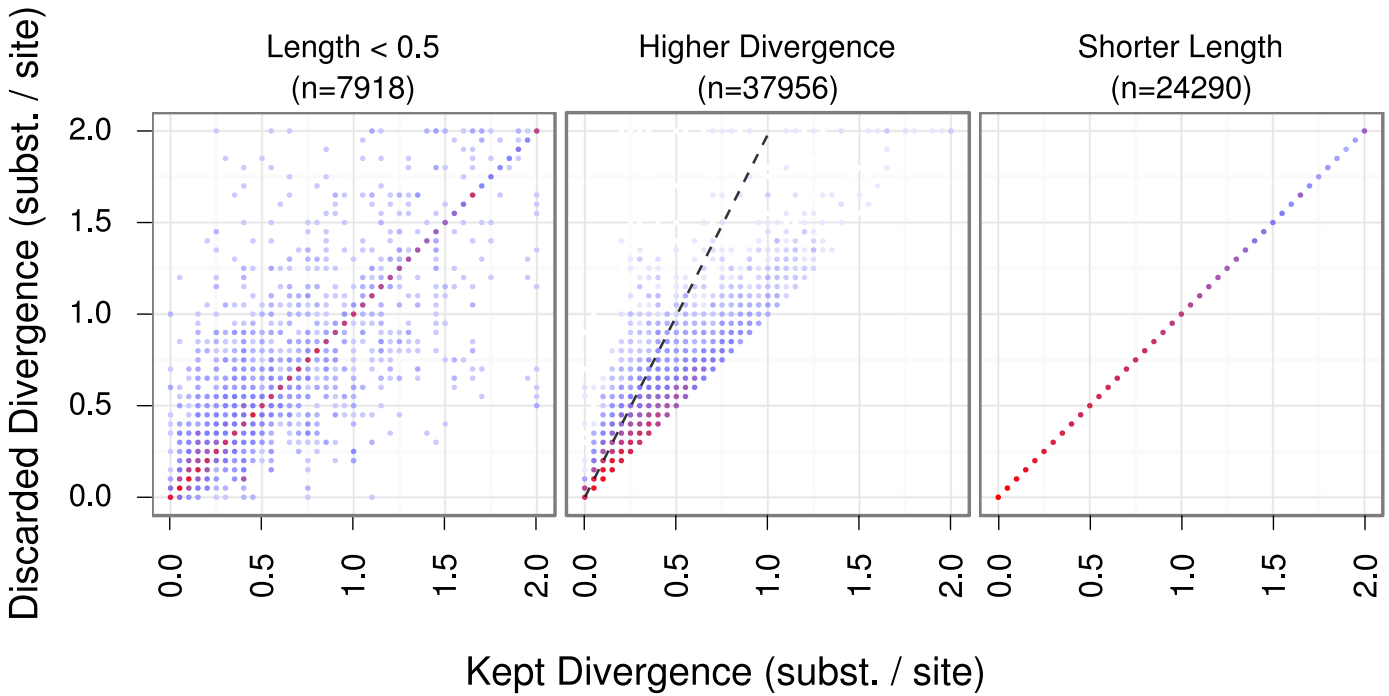
Genome-Specific Duplication and Loss Patterns



Filtering on Sequence Quality

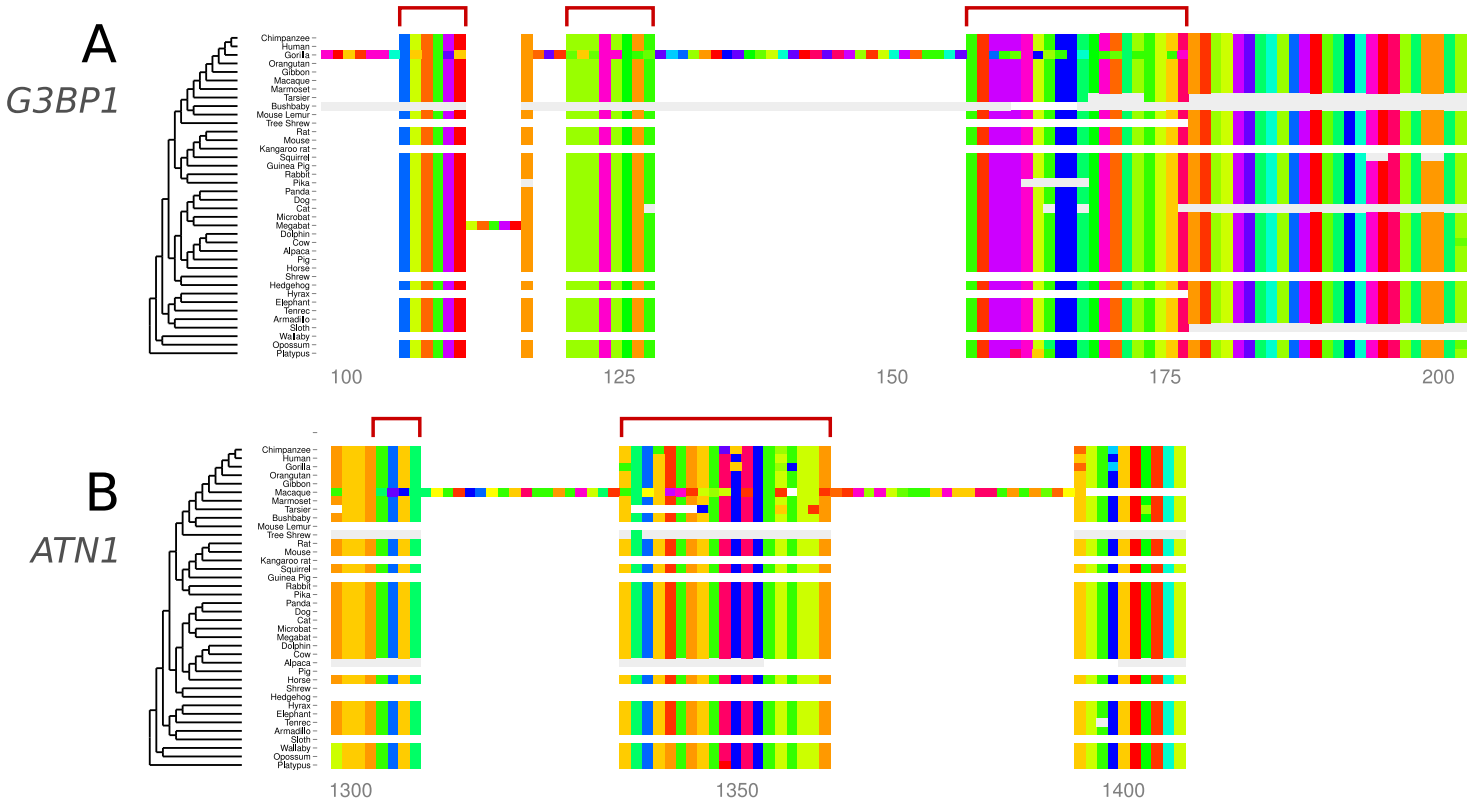


Removing Apparent Paralogs

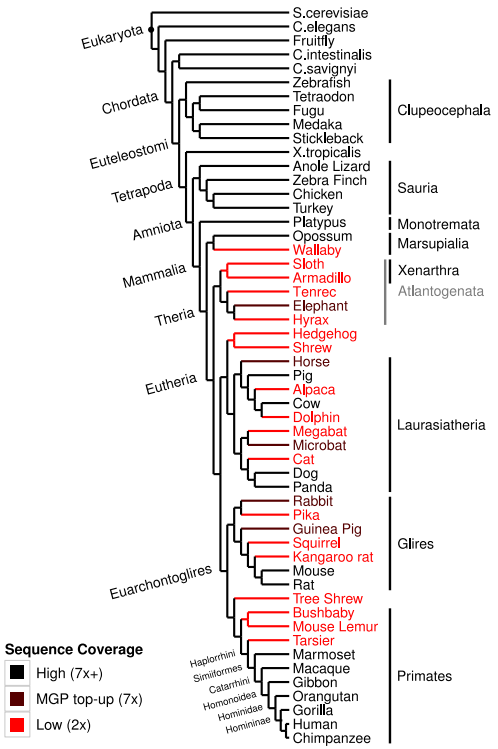


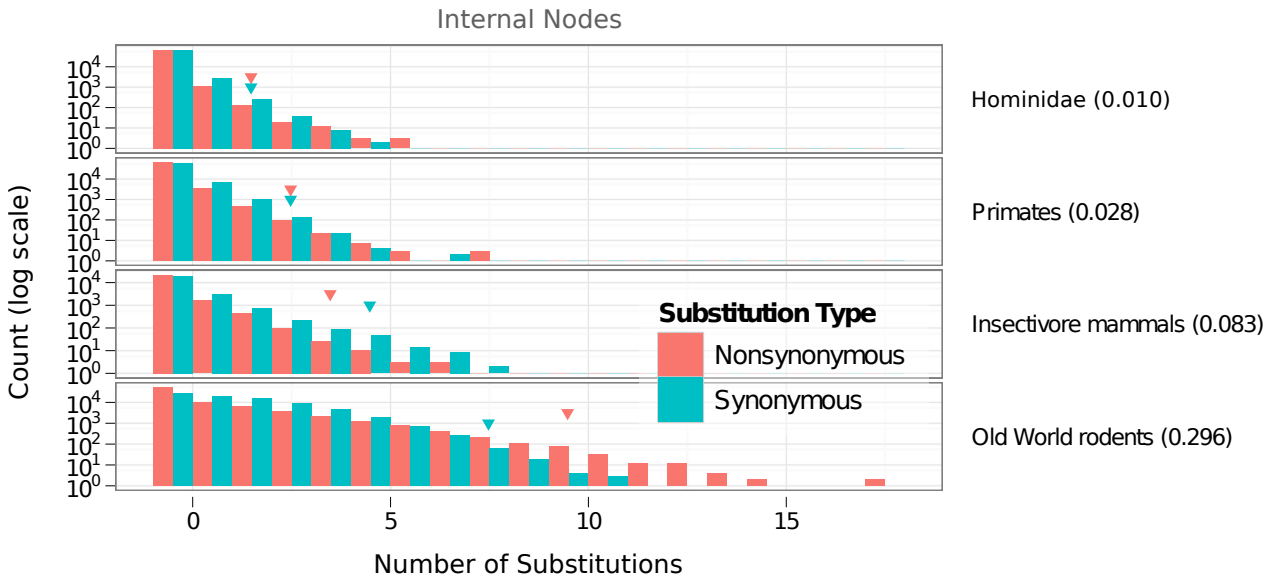
- Why remove paralogs?
  - ”Split” paralogs: erroneous annotations
  - Elevated adaptive evolution after duplication-divergence

Clusters of Nonsynonymous Mutations: Example

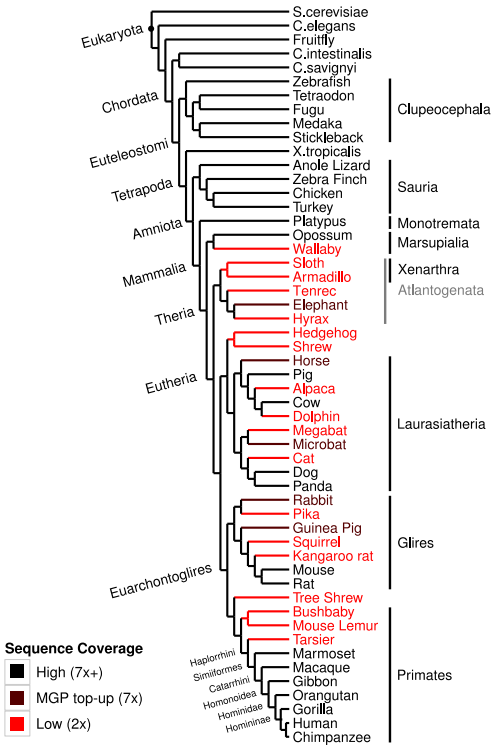


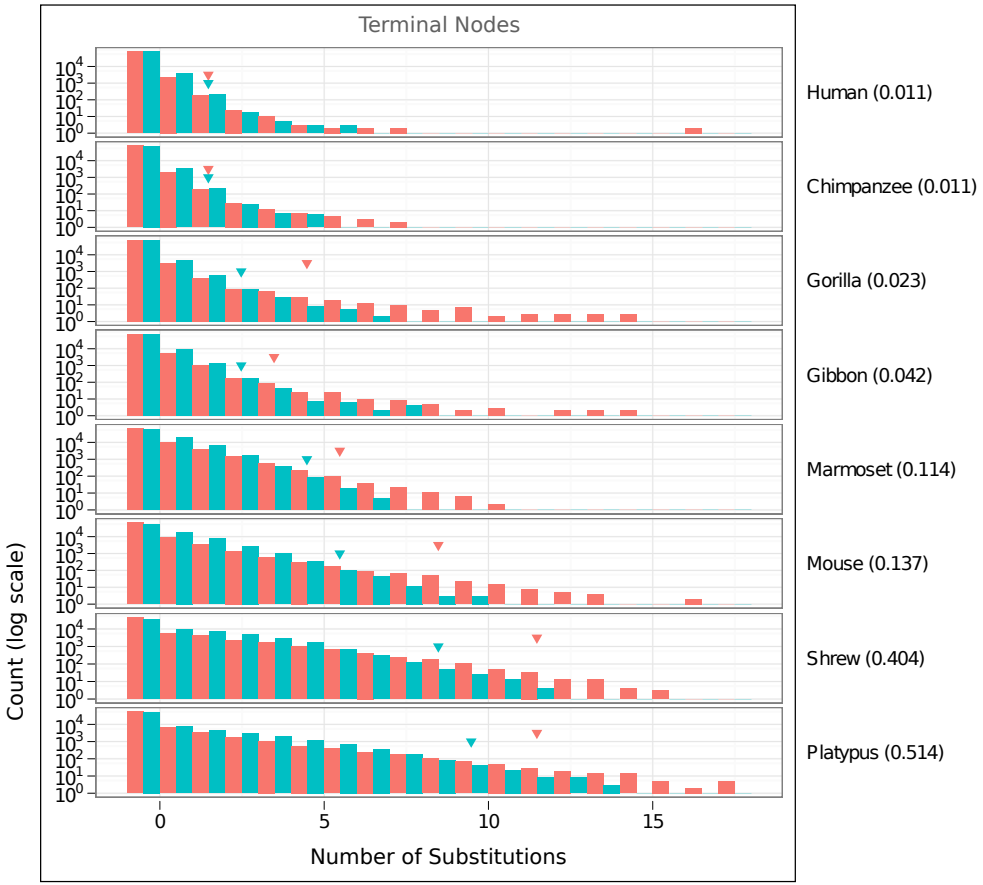
Clusters of Nonsynonymous Mutations: Results



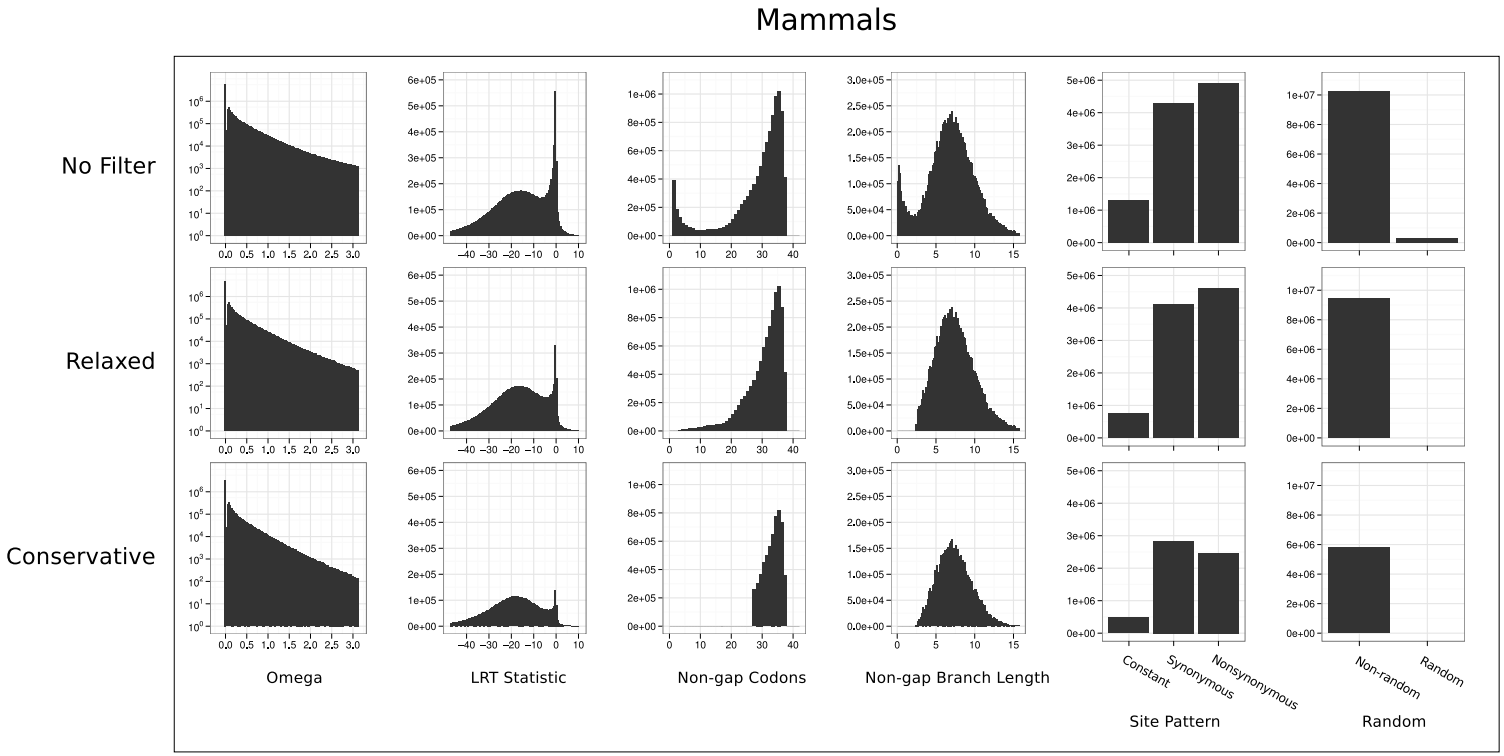


Clusters of Nonsynonymous Mutations: Results

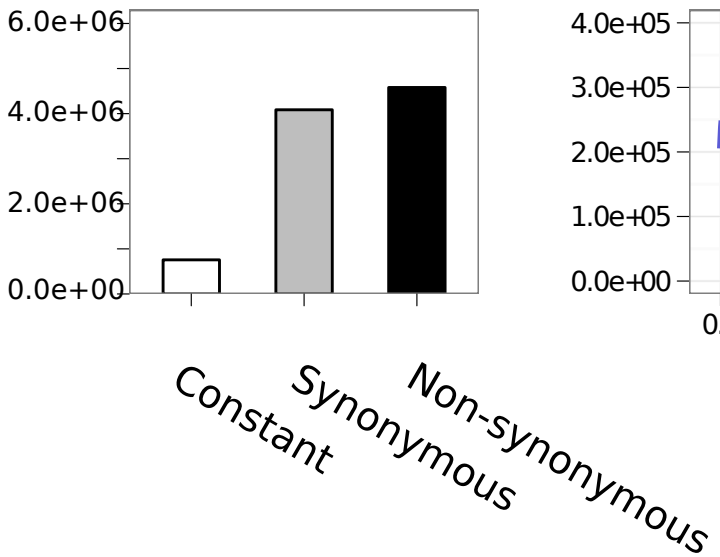


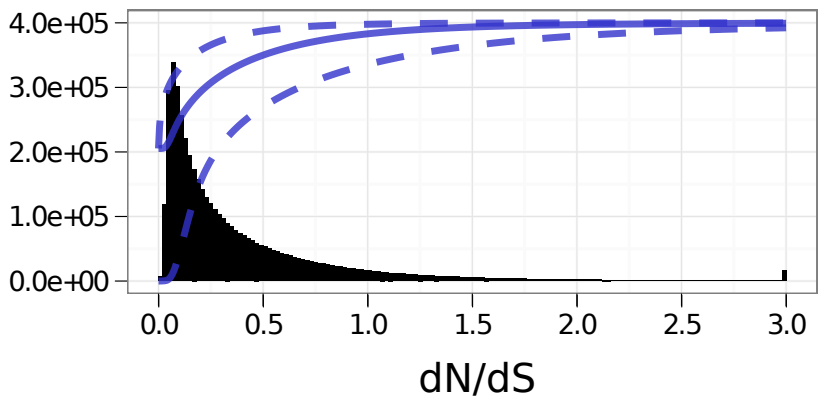


Additional Filters



Global Distribution of dN/dS Mammals

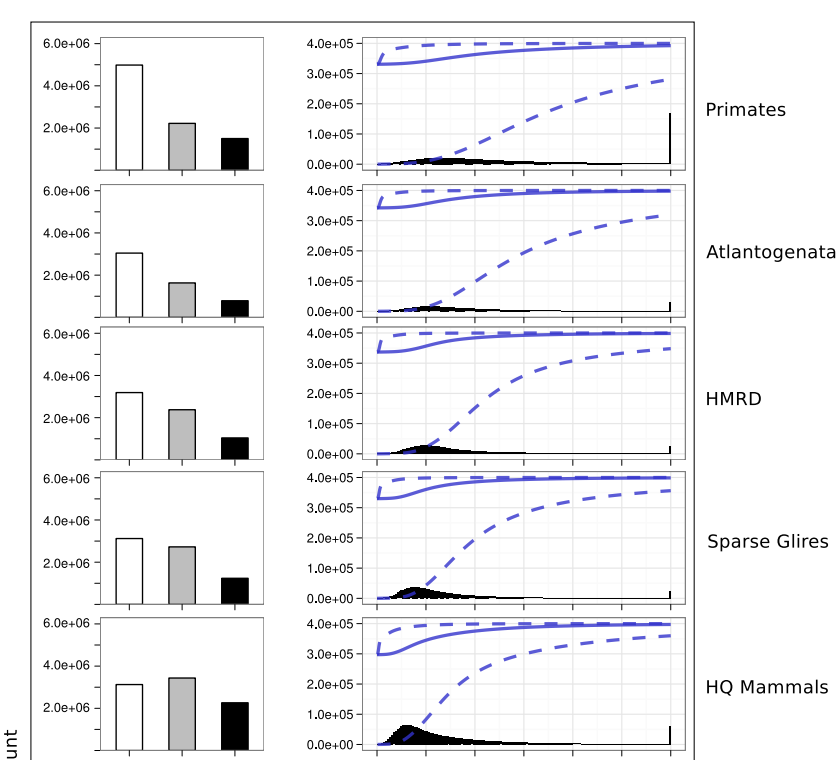


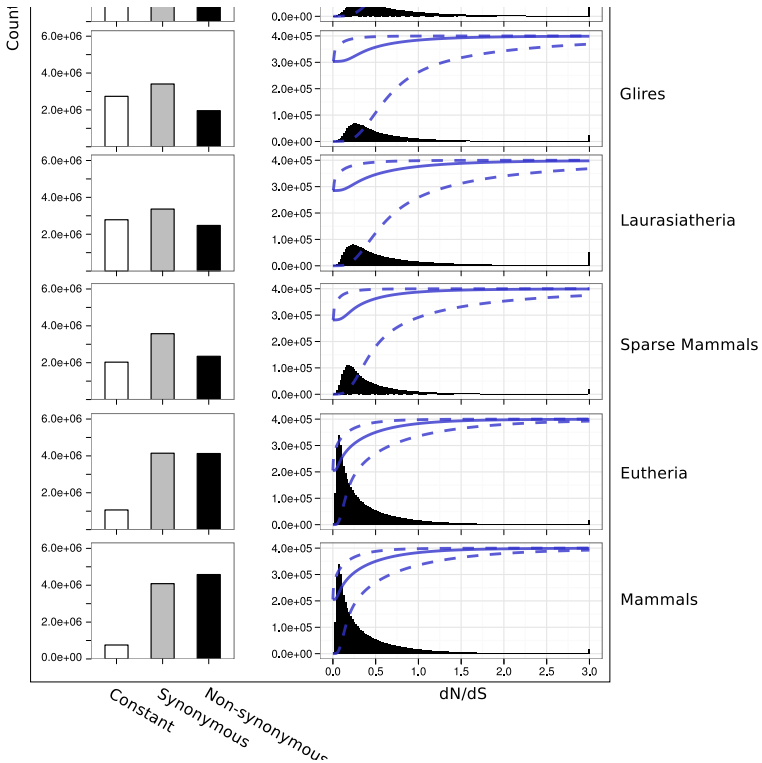


Mammals

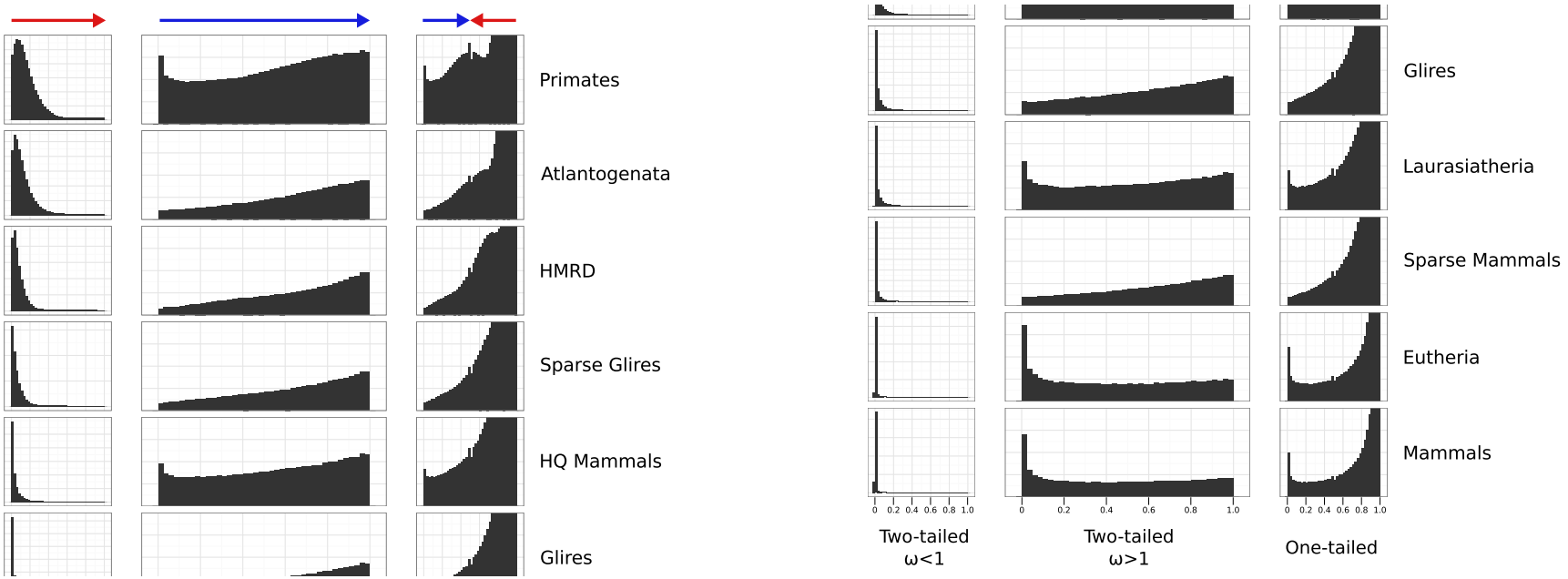
- Ca. 6M protein-coding sites
- ML ω estimate
- Lower / upper bounds

Global Distributions of dN/dS in Ten Species Groups

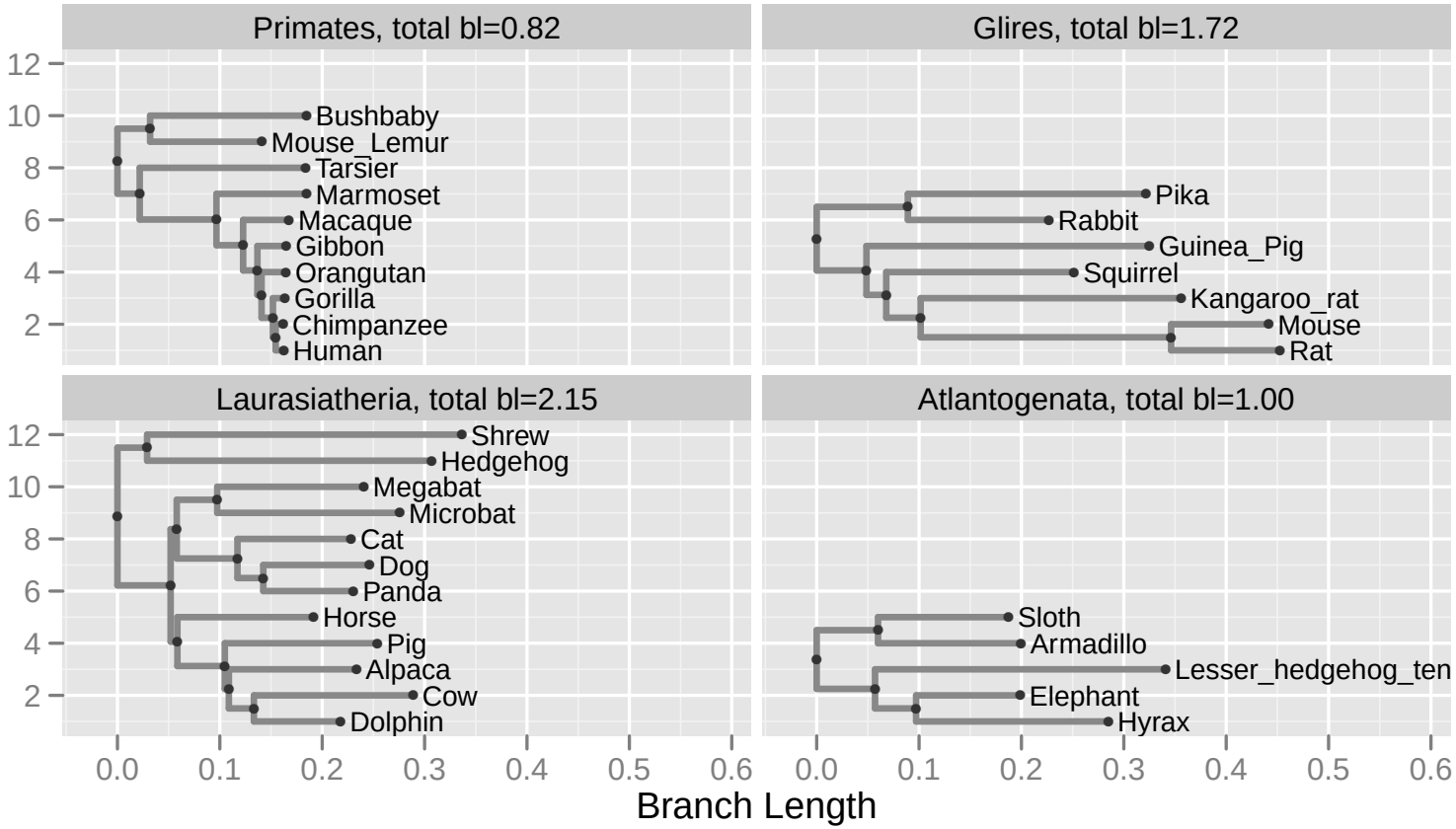




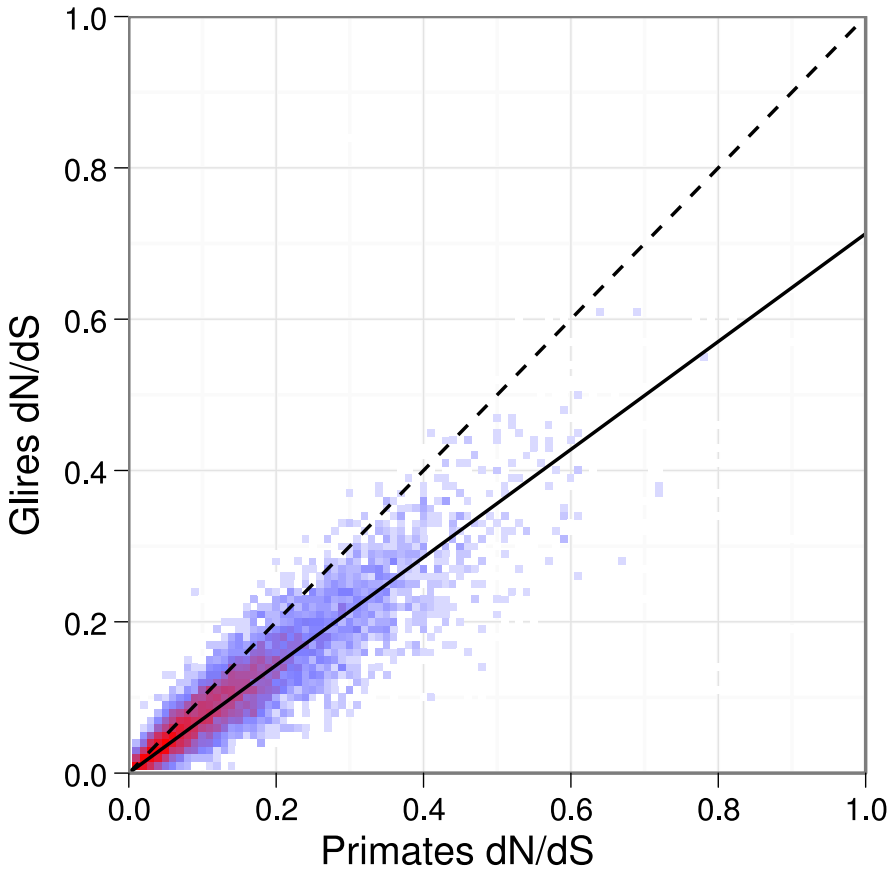
Sitewise p-values for Positive Selection



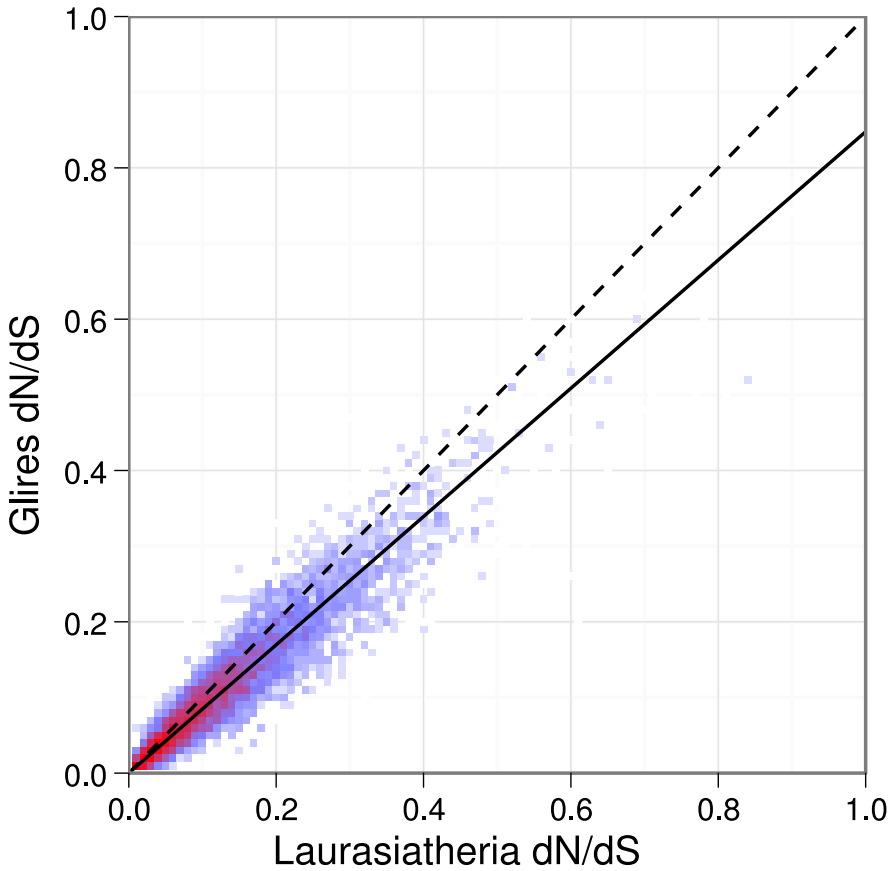
The Mammalian Superorders



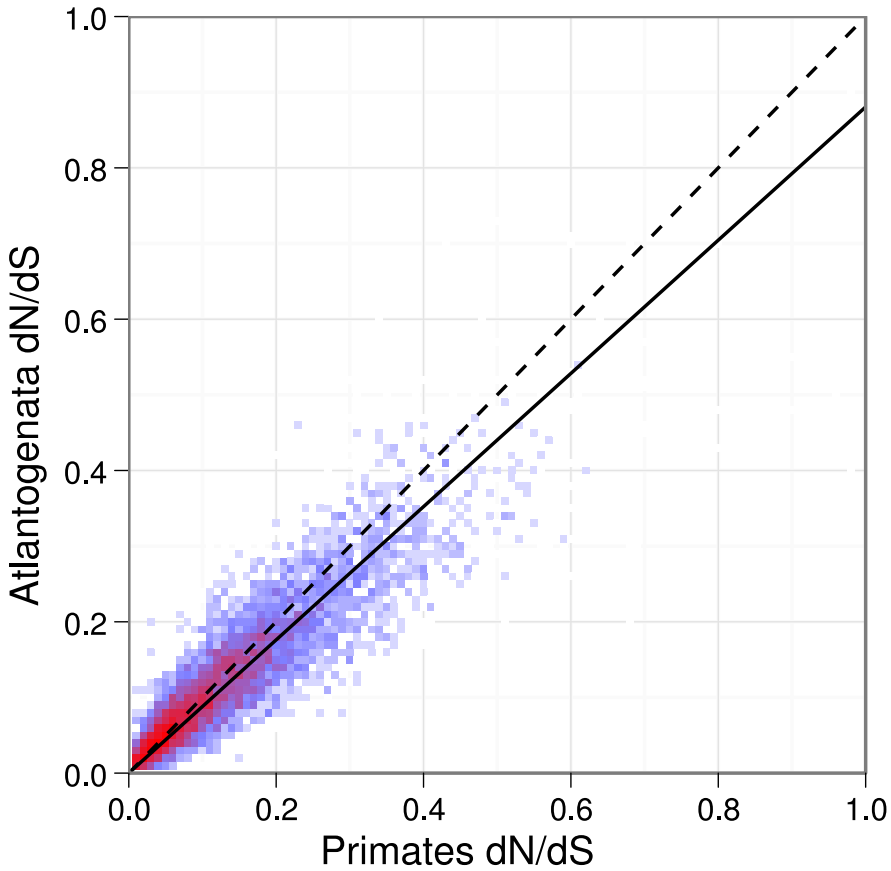
Gene-by-Gene *d*N/*d*S in Species Groups — Effecs of N_E



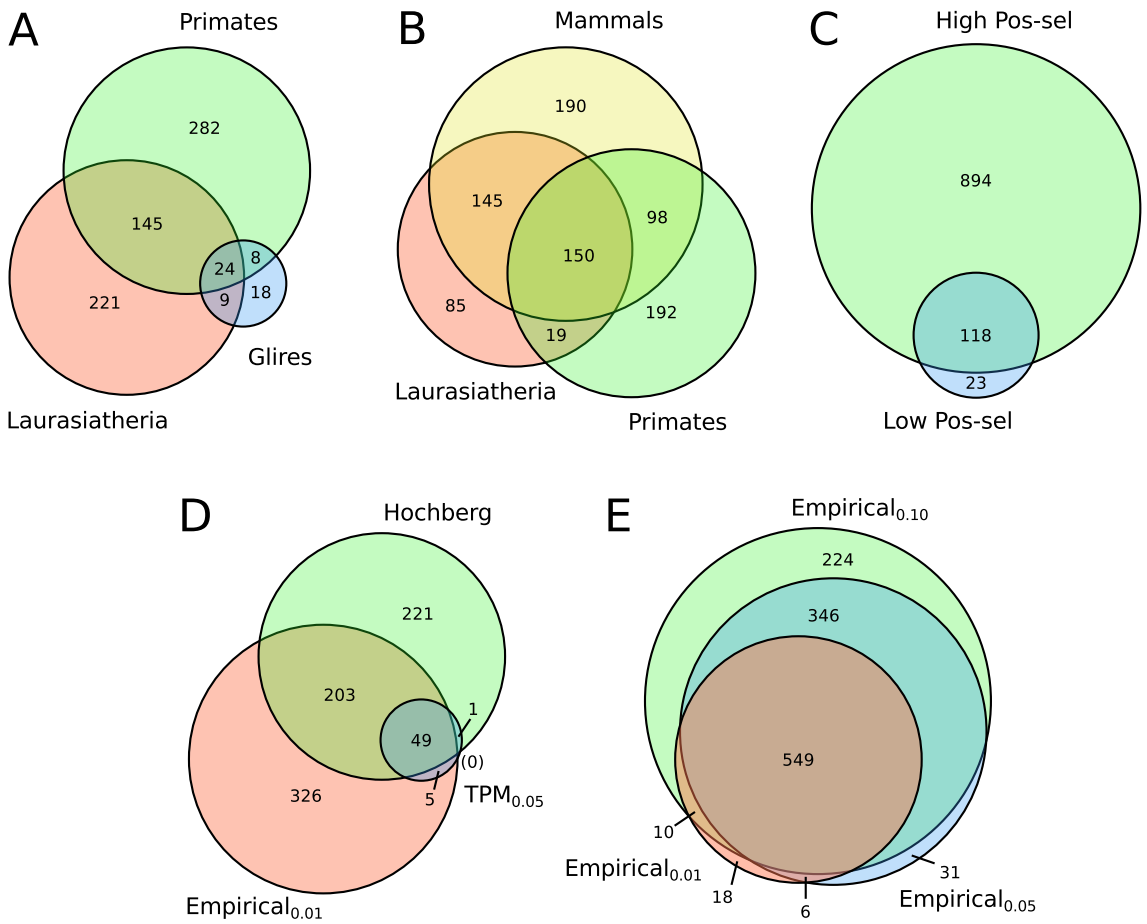
Gene-by-Gene *d*N/*d*S in Species Groups — Effecs of N_E



Gene-by-Gene *d*N/*d*S in Species Groups — Effecs of N_E

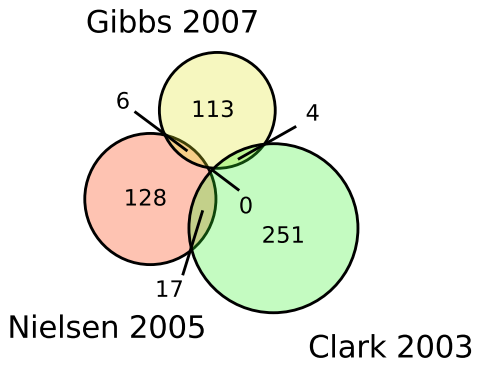


PSG Overlap Between Species Groups and Filters

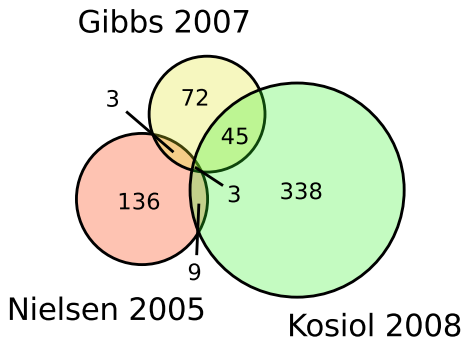


PSG Overlap Between Different Studies

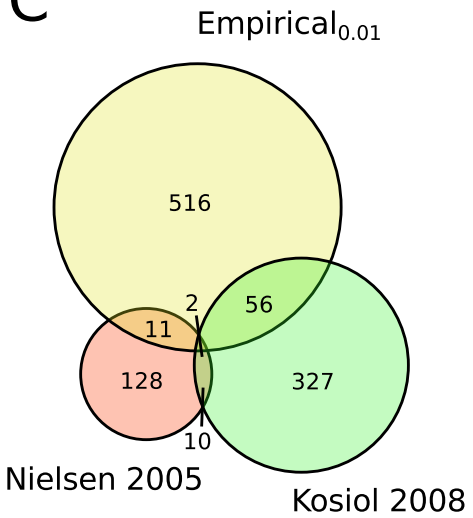
A



B



C



ROC Plot — Enrichment for Strongest PSGs

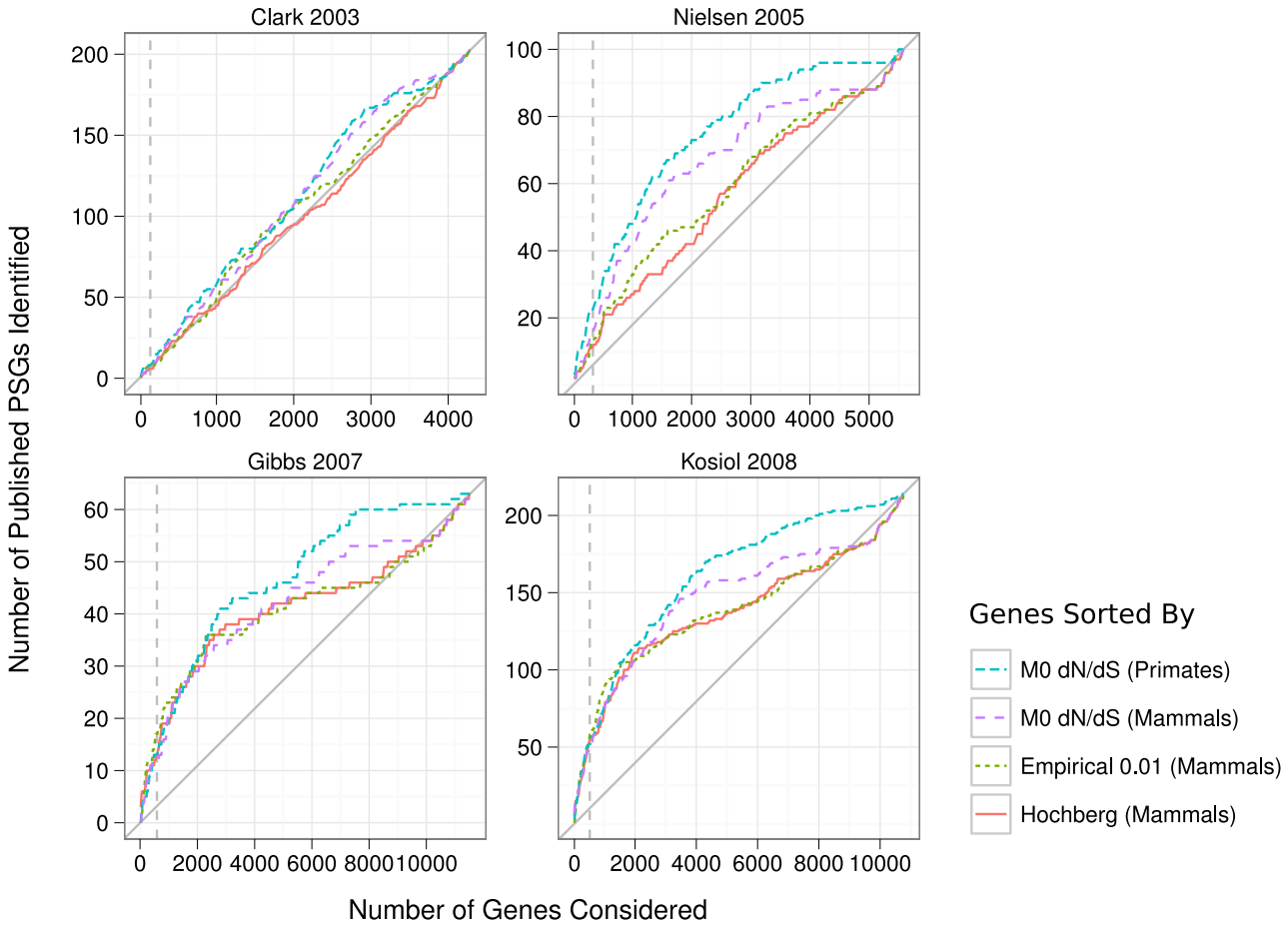


Table — Enrichment for Functional Groups

GO Term		Enriched in		Values for Mammals $Emp_{0.05}$				
ID	Description	This Study	Lit.	FET	topGO	Ann.	Sig.	Exp.
Top 10 Enriched Terms								
GO:0006954	inflammatory response	P LMmHD	RK	**9.4e-11**	**2.0e-06**	202	35	10.2
GO:0045087	innate immune response	P LMmHDi	RK	**2.4e-09**	**2.7e-04**	144	27	7.3
GO:0051607	defense response to virus	P LMmH i	K	**4.2e-05**	**5.4e-03**	43	10	2.2
GO:0042742	defense response to bacterium	PgLMmHDi	K	**8.8e-05**	**1.2e-04**	38	9	1.9
GO:0000236	mitotic prometaphase	PgLMm Di		**3.7e-04**	**3.7e-04**	55	10	2.8
GO:0019221	cytokine-mediated signaling	Mm	K	**1.3e-03**	**4.1e-02**	98	13	5.0
GO:0050900	leukocyte migration	P LMm		**1.5e-03**	**1.4e-03**	112	14	5.7
GO:0007067	mitosis	Pg Mm Di		**3.4e-03**	**4.7e-02**	218	21	11.0
GO:0002576	platelet degranulation	Mm		**4.7e-03**	**4.7e-03**	42	7	2.1
GO:0051297	centrosome organization	gLMm		**5.6e-03**	**1.2e-02**	33	6	1.7
Other Terms Commonly Identified in the Literature								
GO:0006952	defense response	P LMmHDi	RK	**2.0e-15**	1.5e-01	376	59	19.0
GO:0006955	immune response	PgLMmHDi	RK	**2.2e-12**	**3.3e-03**	415	57	21.0
GO:0009611	response to wounding	P LMmHD	RK	**3.2e-07**	5.8e-01	553	56	28.0
GO:0050896	response to stimulus	P LMmHD	RK	**4.0e-06**	6.0e-01	2343	160	118.7
GO:0009607	response to biotic stimulus	P LMmHDi	RK	**2.7e-05**	1.0e+00	265	30	13.4
GO:0050909	sensory perception of taste		RK	5.7e-01	5.7e-01	16	1	0.8
GO:0007600	sensory perception		C K	7.4e-01	1.0e+00	274	12	13.9
GO:0007606	sensory perception of chemical stimulus		RK	8.5e-01	1.0e+00	36	1	1.8
GO:0007166	cell surface receptor linked signaling		CR	1.0e+00	8.2e-01	1107	37	56.1
GO:0007608	sensory perception of smell		C K	1.0e+00	1.0e+00	17	0	0.9

Table — Enrichment for Functional Groups

Other Terms Identified in This Study but Not in the Literature					
GO:0006302	double-strand break repair	Mm i	**8.4e-03**	**2.6e-02**	
GO:0051301	cell division	g Mm	**1.5e-02**	**7.2e-03**	2
GO:0031295	T cell costimulation	LMm D	**2.1e-02**	**2.1e-02**	
GO:0007059	chromosome segregation	Mm	**2.4e-02**	**1.6e-02**	
GO:0015711	organic anion transport	mH	7.6e-02	9.1e-02	
GO:0071706	TNF superfamily cytokine production	L mH	7.6e-02	9.8e-02	
GO:0007283	spermatogenesis	P L Di	8.3e-02	5.4e-02	1

Table — Enrichment for Positively-Selected Protein Domains

Pfam Domain		FDR < 0.1		All Sites		p < 0.01 Sites		
Accession	Description	Cons.	Relaxed	Genes	Sites	Genes	Sites	Top 5 Genes w
Immune Related Domains								
PF07686	Immunoglobulin V-set domain	MmH	PgLMmH	220	10851	58	210	TMIGD1, TREM
PF00047	Immunoglobulin domain	H	P MmH	240	30486	51	180	Pecam1, CD4, PI
PF00084	Sushi domain (SCR repeat)		P LMmH	37	7957	17	125	C1S, CD46, C4B
PF00059	Lectin C-type domain	g Mm	PgLMmH	51	4798	29	111	CD72, KLRB1, P
PF00048	Small cytokines (intecrine/chemokine), IL-8 like	LMmH	PgLMmH	26	1231	20	53	CXCL13, CXCL9
PF00530	Scavenger receptor cysteine-rich domain	H	P MmH	17	2865	8	42	CD5L, MARCO,
PF01823	MAC/Perforin domain	P MmH	P LMmH	9	1176	5	28	C9, C8A, C6, C7
PF00021	u-PAR/Ly-6 domain		P LMmH	16	775	7	27	CD59, TEX101,
PF00340	Interleukin-1 / 18		PgLMm	8	523	6	27	IL1A, IL18, IL1F
PF00969	Class II histocompatibility antigen, beta domain		P MmH	5	266	4	22	HLA-DMB, HLA
PF02841	Guanylate-binding protein, C-terminal domain		PgLMm	4	970	4	21	GBP4, GBP5, GI
PF00074	Pancreatic ribonuclease		Mm	5	338	5	17	RNASE7, RNASI
PF00993	Class II histocompatibility antigen, alpha domain		gLMmH	5	364	4	16	HLA-DQA1, HLA
PF00354	Pentaxin family		Pg MmH	6	677	3	14	CRP, APCS, SVI
PF00062	C-type lysozyme/alpha-lactalbumin family		MmH	7	462	5	13	LALBA, LYZ, LY
Protease Domains								

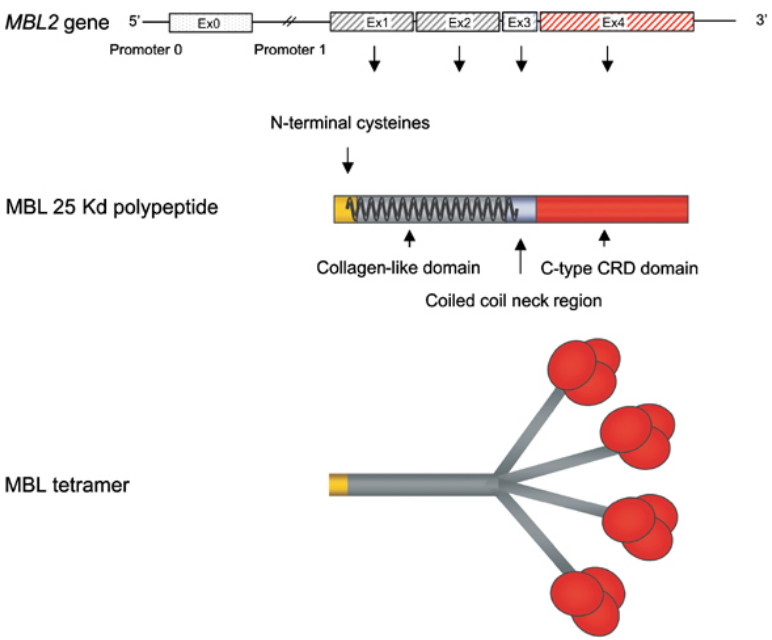
Table — Enrichment for Positively-Selected Protein Domains

PF00062	C-type lysozyme/ alpha-lactalbumin family			7	402	5	15	LALBA, LYZ, LY
Protease Domains								
PF00089	Trypsin			82	10426	44	184	CFI, C1S, PRSS
PF00656	Caspase domain			12	1474	10	40	CASP8, CASP10
PF00246	Zinc carboxypeptidase			22	4505	9	23	CPB1, CPB2, CP
PF07859	alpha/beta hydrolase fold			6	804	5	16	AADAC, AADAC
Protease Inhibitor Domains								
PF00079	Serpin (serine protease inhibitor)			33	7358	19	73	SERPINA3, SER
PF00031	Cystatin domain			11	780	7	33	HRG, KNG1, FE
PF01835	MG2 domain			8	1615	7	21	C5, C4A, PZP, A
PF07678	A-macroglobulin complement component			8	1483	7	14	C5, C4A, A2M, A

Putative Gene Conversion between Nearby Paralogous Pairs

Gene Family	Genes	NPPs	PSGs	NPP–PSGs	Top 4 NPP–PSGs
Ensembl Families with 3PSGs					
ENSFM00600000921151	6	6	6	6	COL4A6, COL4A2, COL4A4, COL4A5
ENSFM00250000001219	5	5	5	5	CD1D, CD1C, CD1A, CD1E
ENSFM00250000000804	4	4	4	4	C6, C7, C8A, C8B
ENSFM00250000000852	4	4	4	4	GBP6, GBP4, GBP5, GBP3
ENSFM00500000269596	11	11	4	4	SERPINB3, SERPINB12, SERPINB13, SERPINB9
ENSFM00500000269665	4	4	4	4	CTSG, GZMB, GZMH, CMA1
ENSFM00250000000002	89	68	4	3	ZFP37, ZNF473, ZNF677
ENSFM00250000000948	4	3	4	3	ACOT6, ACOT4, ACOT2
ENSFM00350000105388	5	5	3	3	CES5A, CES2, CES1
ENSFM00400000131714	7	6	3	3	SLC22A25, SLC22A14, SLC22A8
ENSFM00400000131728	7	7	3	3	MMP3, MMP8, MMP1
ENSFM00470000251442	4	3	4	3	EMR2, EMR3, CD97
ENSFM00500000269709	5	4	4	3	ITGAL, ITGAX, ITGAM
ENSFM00500000269927	4	4	3	3	SLC17A3, SLC17A1, SLC17A4
ENSFM00570000851010	5	4	4	3	NLRP9, NLRP4, NLRP5
Manually Curated Families					FET p-value
Toll-like Receptors	8	4	5	3	0.50 TLR8, TLR6, TLR1
Collagen	30	7	22	7	0.08 COL4A6, COL4A2, COL4A4, COL4A5
ADAM Family	42	11	7	4	0.06 ADAM32, ADAM2, ADAM28, ADAM7
Solute Carrier Family	338	51	37	10	0.03 SLC26A3, SLC17A3, SLC17A1, SLC17A4
All Genes	15946	1150	1898	200	0.00 AC090098.1, CDKN2A, FAM26F, ACOT6

Case Study: Mannose-Binding Lectin 2



- Collectin family member
  - Binds carbohydrates of invading pathogens
  - Facilitates phagocytosis, activates complement pathway
- High frequencies of collagen-disrupting alleles
  - At odds with supposed important immune function
  - Recent selection favoring low MBL levels?
  - Longer-term host-immune conflict?

Case Study: Mannose-Binding Lectin 2

"MBL2 is well conserved in agreement with its important role in the immune system."

Verga F et al. (2004). Evolution of the mannose-binding lectin gene in primates. *Genes and Immunity*, 5, 653-661.

But...

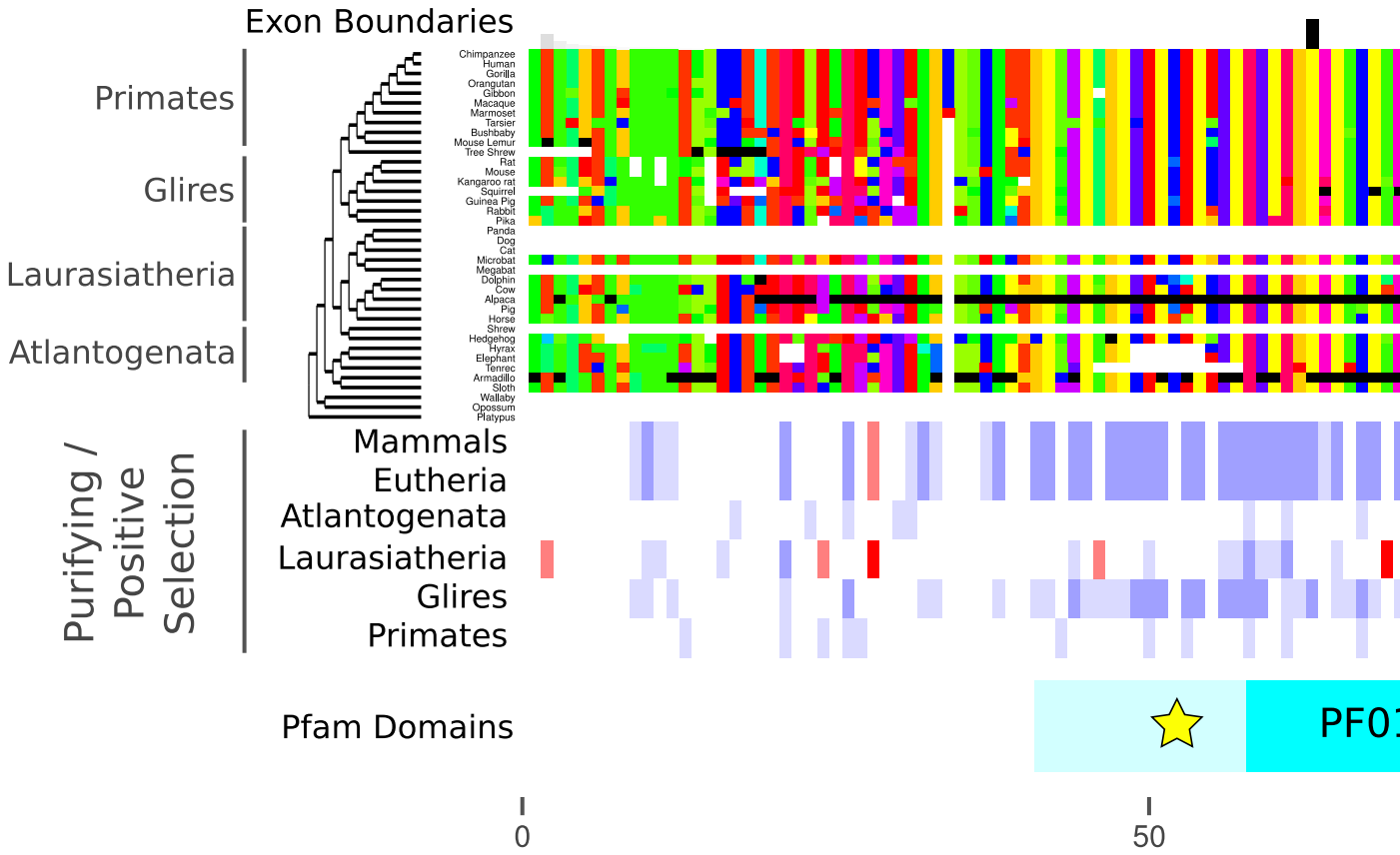
"The evolutionary neutrality of MBL2 strongly supports [...] that this lectin is largely redundant in host human defences"

Verdu P et al. (2006). Evolutionary insights into the high worldwide prevalence of *MBL2* deficiency alleles. *Human Molecular Genetics*, 15, 2650-2658.

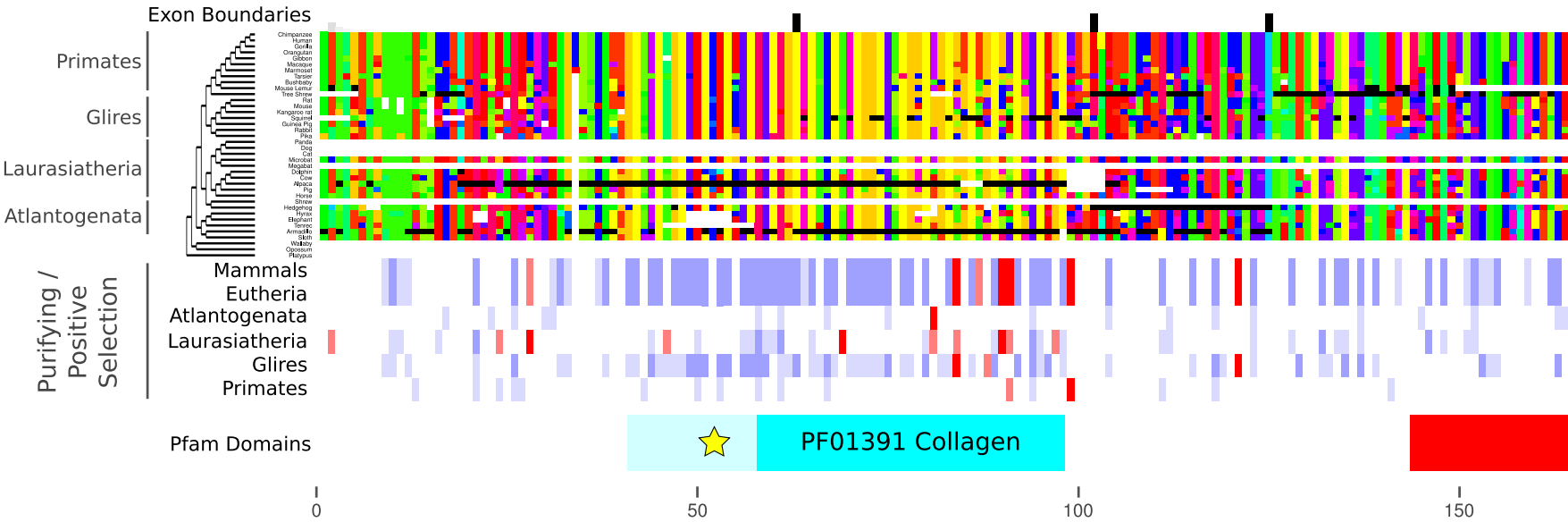
Case Study: Mannose-Binding Lectin 2

With sitewise estimates, a more balanced picture arises

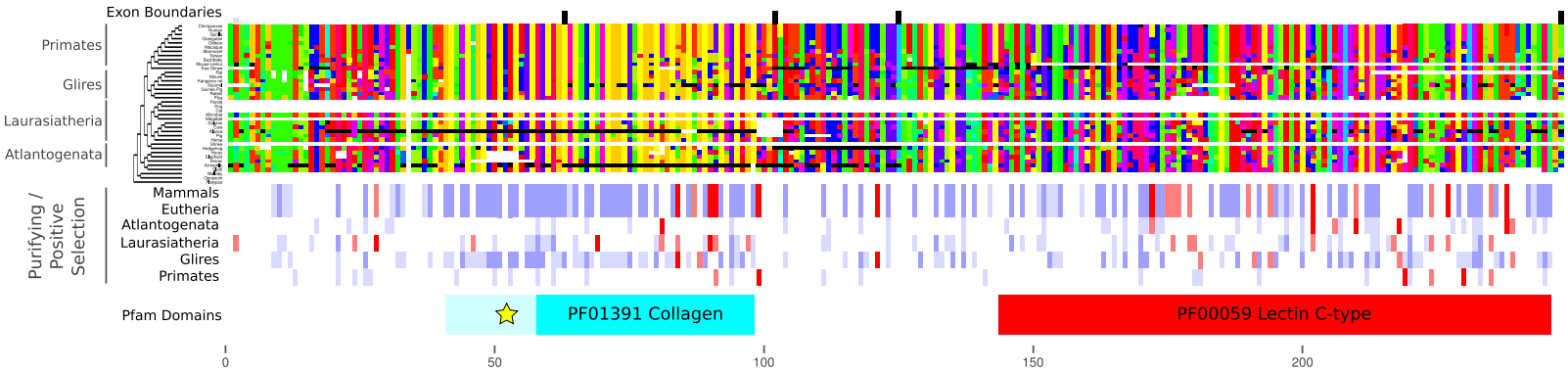
Case Study: Mannose-Binding Lectin 2



Case Study: Mannose-Binding Lectin 2



Case Study: Mannose-Binding Lectin 2



- Large number of purifying sites; *MBL2* is functional across mammals
  - Strongest conservation tract within collagen domain
- Some regions show positive selection
  - No overlap with location of human SNPs (starred)
  - Potential protein-protein or protein-ligand interacting regions

Blah Blah

Acknowledgements