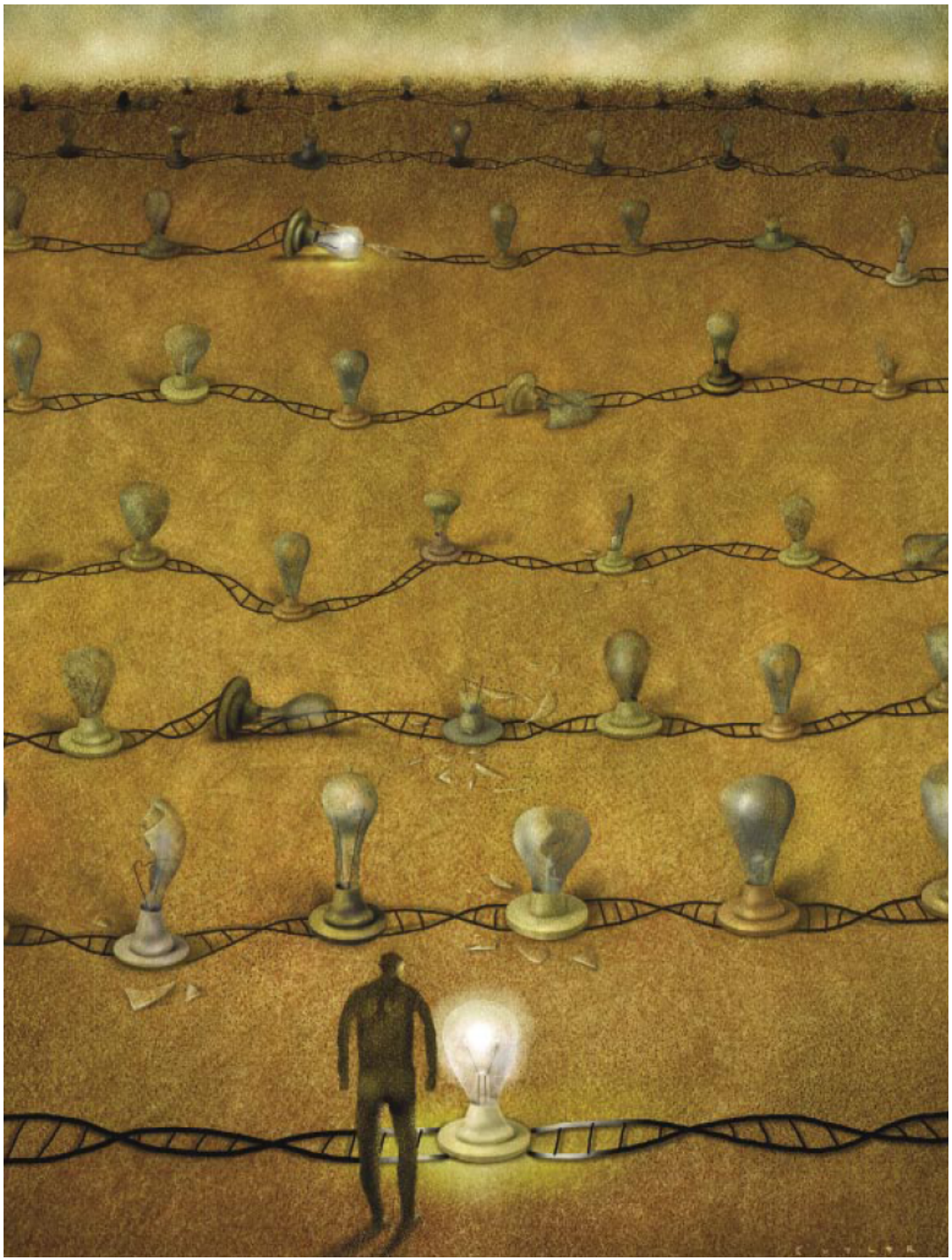


Gencode Mar '10 Meeting: Pseudogene Project Update

Mark Gerstein

Illustration from Gerstein & Zheng (2006). Sci Am.

Overall Flow: Pipeline Runs, Coherent Sets, Annotation, Transfer to Sanger

- Overall Approach
  1. Overall Pipeline runs at Yale and UCSC, yielding raw pseudogenes
  2. Extraction of coherent subsets for further analysis and annotation
  3. Passing to Sanger for detailed manual analysis and curation
  4. Incorporation into final GENCODE annotation
  5. Pipeline modification

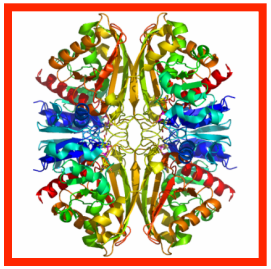
- Chronology of Sets
  1. Encode Pilot 1%
  2. Ribosomal Protein pseudogenes
  3. **Unitary pseudogenes (Hard)**
  4. **Glycolytic Pseudogenes**
  5. **Polymorphic Pseudogenes**
  6. **Pseudogenes Associated with SDs**

Specific Pseudogene Assignments: Glycolytic Pseudogenes (completed)

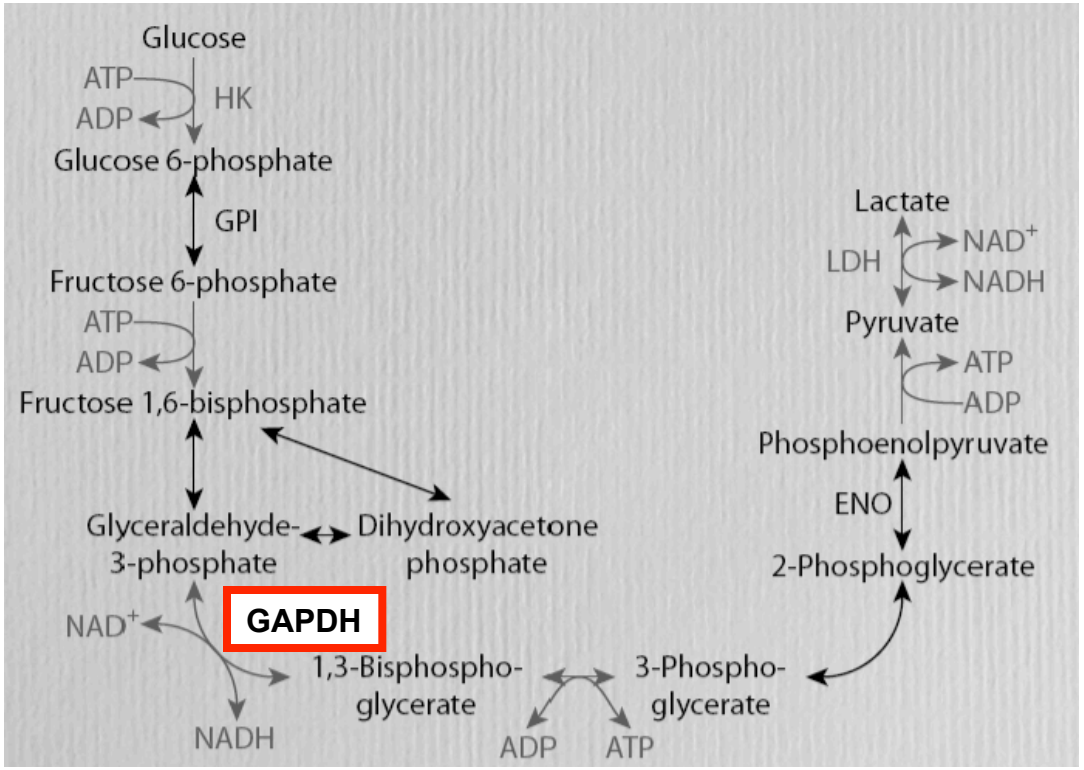
Number of pseudogenes for each glycolytic enzyme

[Liu et al. BMC Genomics ('09)]

Large numbers of processed GAPDH pseudogenes in mammals comprise one of the biggest families but numbers not obviously correlated with mRNA abundance.



Processed/Duplicated

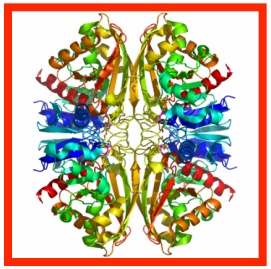


	Human	Chimp	Mouse	Rat	Chicken	Zebrafish	Pufferfish	Fruitfly	Worm
HK	1/0	1/2	0/1	-	0/2	-	-	-	-
GPI	-	-	1/0	-	-	-	-	-	-
PFK	-	-	-	-	-	0/1	-	-	-
ALDO	1/1	1/1	11/0	7/0	0/1	-	-	-	-
TPI	3/0	2/1	6/1	3/1	-	-	-	-	-
GAPDH	60/2	47/3	285/46	329/35	0/1	-	-	-	-
PGK	1/1	1/2	2/0	12/0	-	-	-	-	-
PGM	12/0	13/1	9/0	3/0	-	-	-	-	-
ENO	1/0	1/2	12/1	36/3	-	-	-	-	-
PK	2/0	3/0	10/3	4/1	-	-	-	-	-
LDH	10/2	9/1	27/7	25/4	-	-	-	-	-
Total	**97**	**91**	**422**	**463**	**4**	**1**	**0**	**0**	**0**

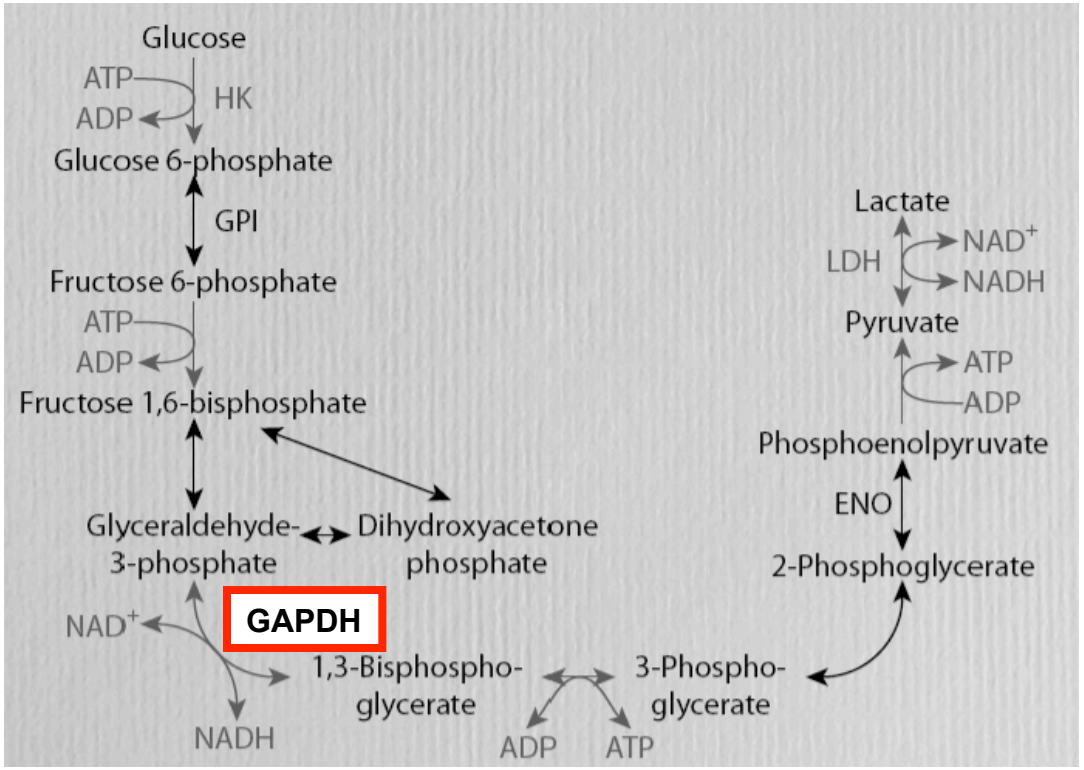
Number of pseudogenes for each glycolytic enzyme

[Liu et al. BMC Genomics ('09)]

Large numbers of processed GAPDH pseudogenes in mammals comprise one of the biggest families but numbers not obviously correlated with mRNA abundance.



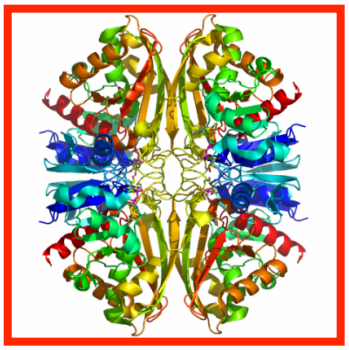
Processed/Duplicated



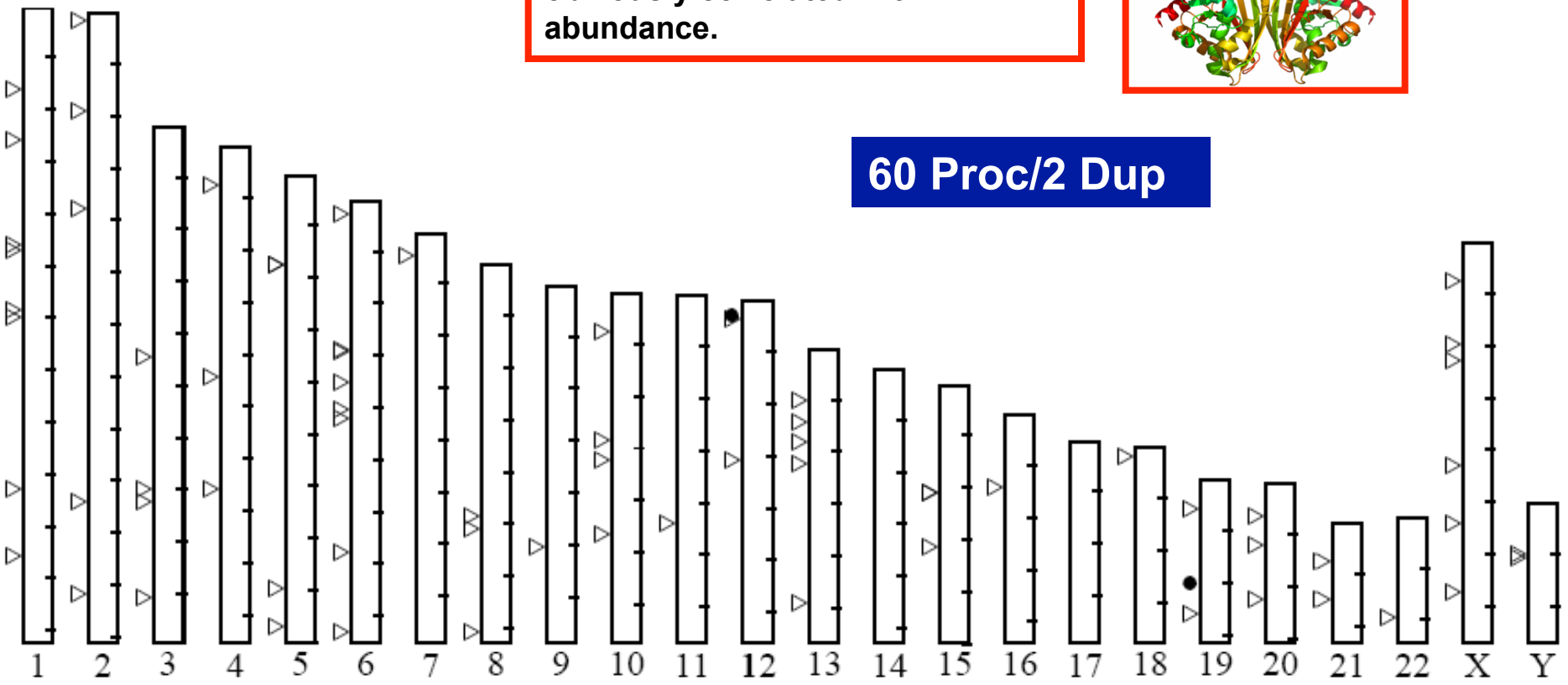
	Human	Chimp	Mouse	Rat	Chicken	Zebrafish	Pufferfish	Fruitfly	Worm
HK	1/0	1/2	0/1	-	0/2	-	-	-	-
GPI	-	-	1/0	-	-	-	-	-	-
PFK	-	-	-	-	-	0/1	-	-	-
ALDO	1/1	1/1	11/0	7/0	0/1	-	-	-	-
TPI	3/0	2/1	6/1	3/1	-	-	-	-	-
GAPDH	60 Proc/2 Dup	7/3	285/46	329/35	0/1	-	-	-	-
PGK	1/1	1/2	2/0	12/0	-	-	-	-	-
PGM	12/0	13/1	9/0	3/0	-	-	-	-	-
ENO	1/0	1/2	12/1	36/3	-	-	-	-	-
PK	2/0	3/0	10/3	4/1	-	-	-	-	-
LDH	10/2	9/1	27/7	25/4	-	-	-	-	-
Total	**97**	**91**	**422**	**463**	**4**	**1**	**0**	**0**	**0**

Distribution of human GAPDH pseudogenes

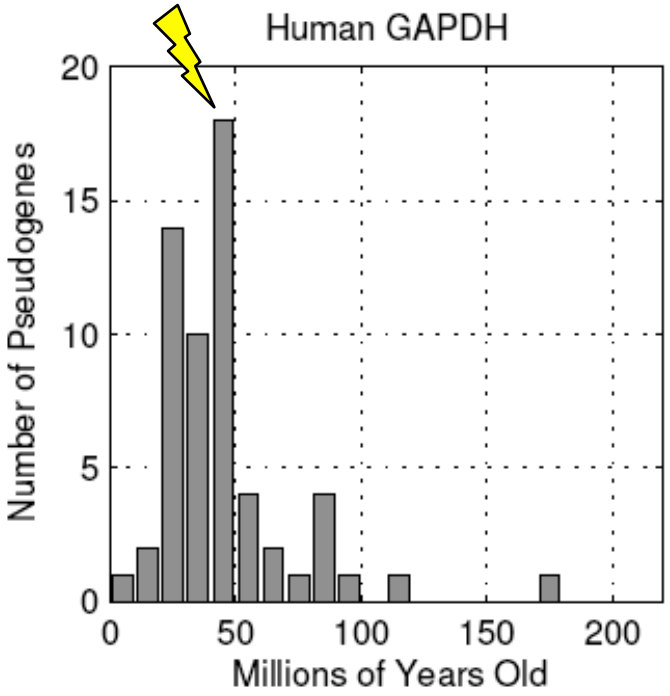
Large numbers of processed GAPDH pseudogenes in mammals comprise one of the biggest families but numbers not obviously correlated with mRNA abundance.



60 Proc/2 Dup



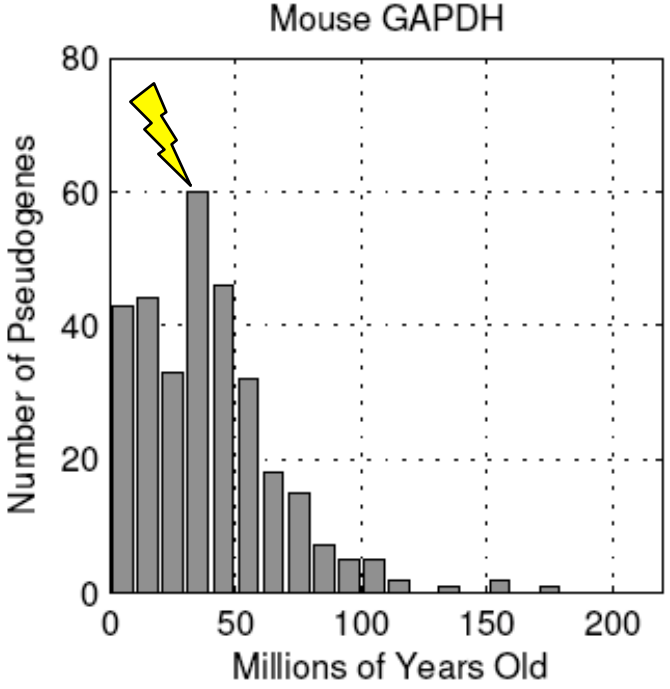
[Liu et al. BMC Genomics ('09, in press)]

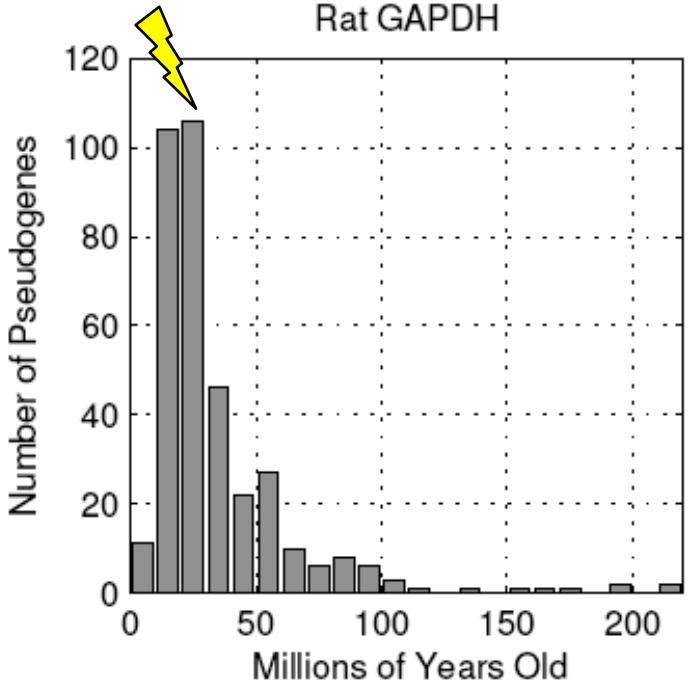


Burst of Retrotran-spositional Activity

Aproximate Age of GAPDH pseudogenes

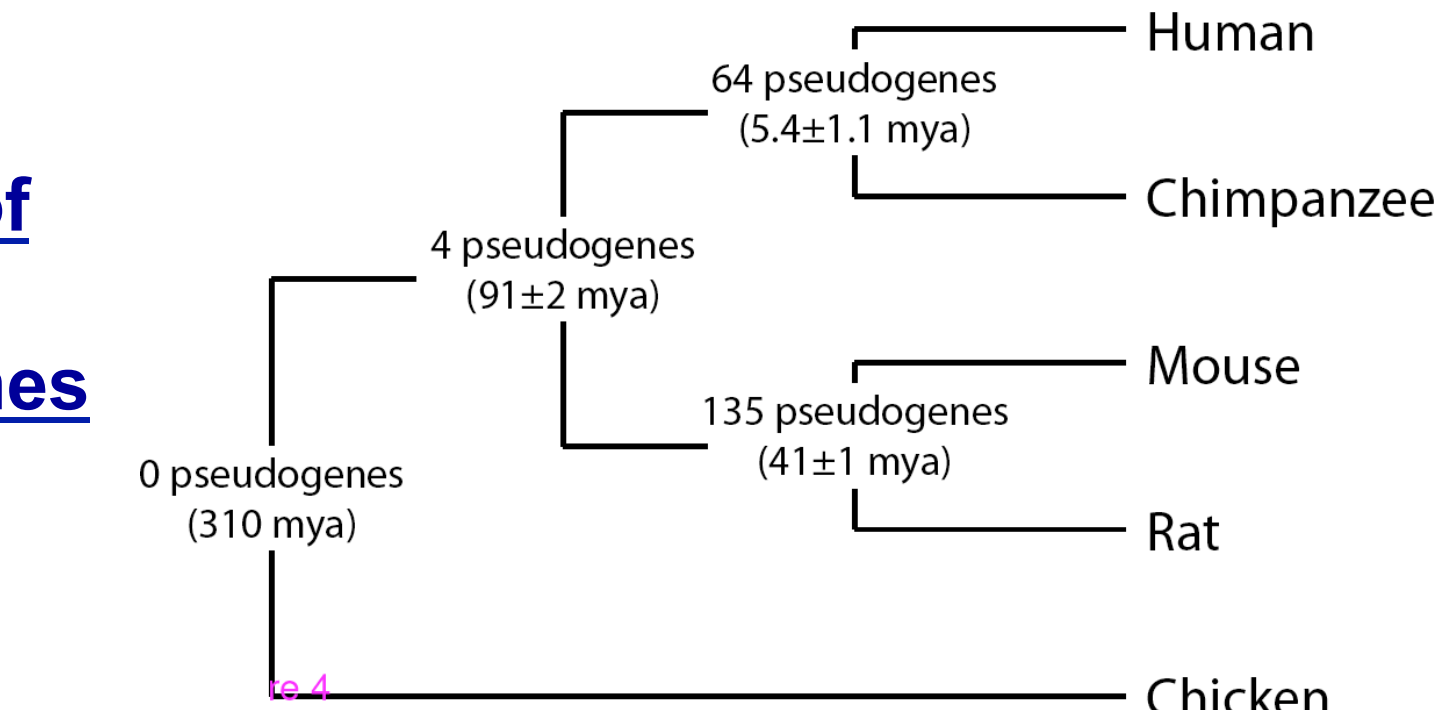
Age calculated based on Kimura-2 parameter model of nucleotide substitution

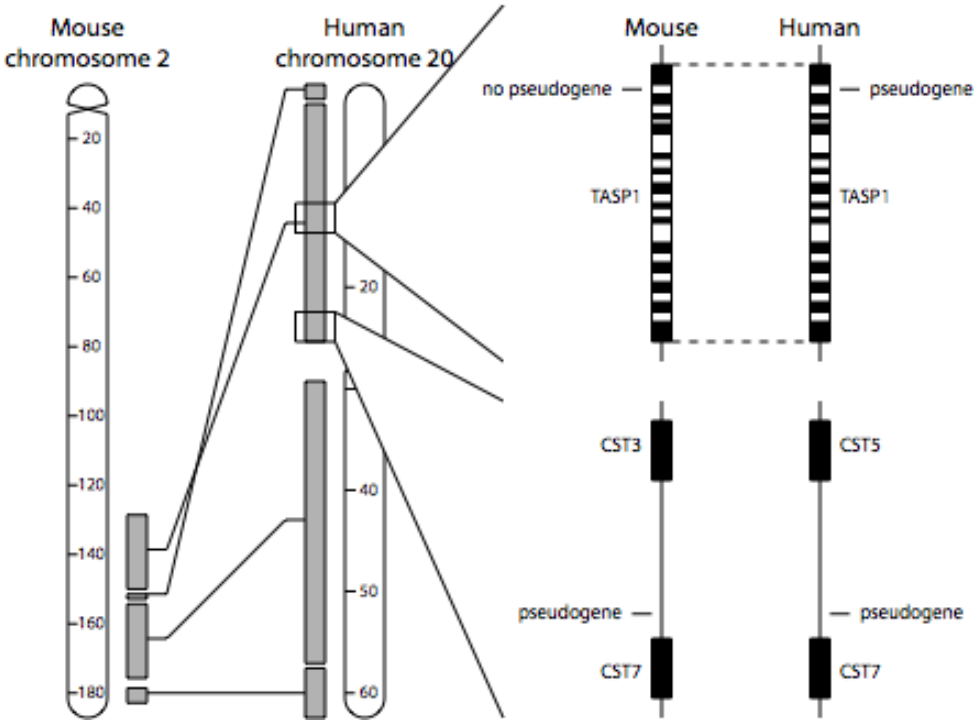




[Liu et al. BMC Genomics ('09)]

Synteny of GAPDH pseudogenes





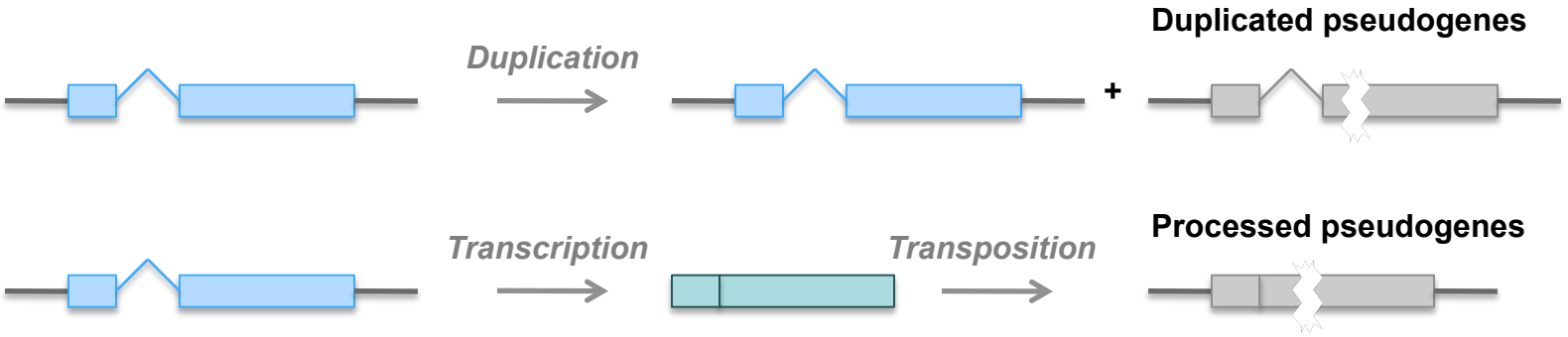
Synteny derived based on local gene orthology

[Liu et al. BMC Genomics ('09)]

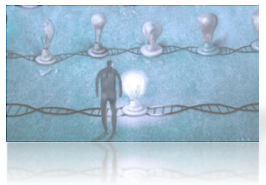
Specific Pseudogene Assignments: Unitary Pseudogenes (completed)

Pseudogenes

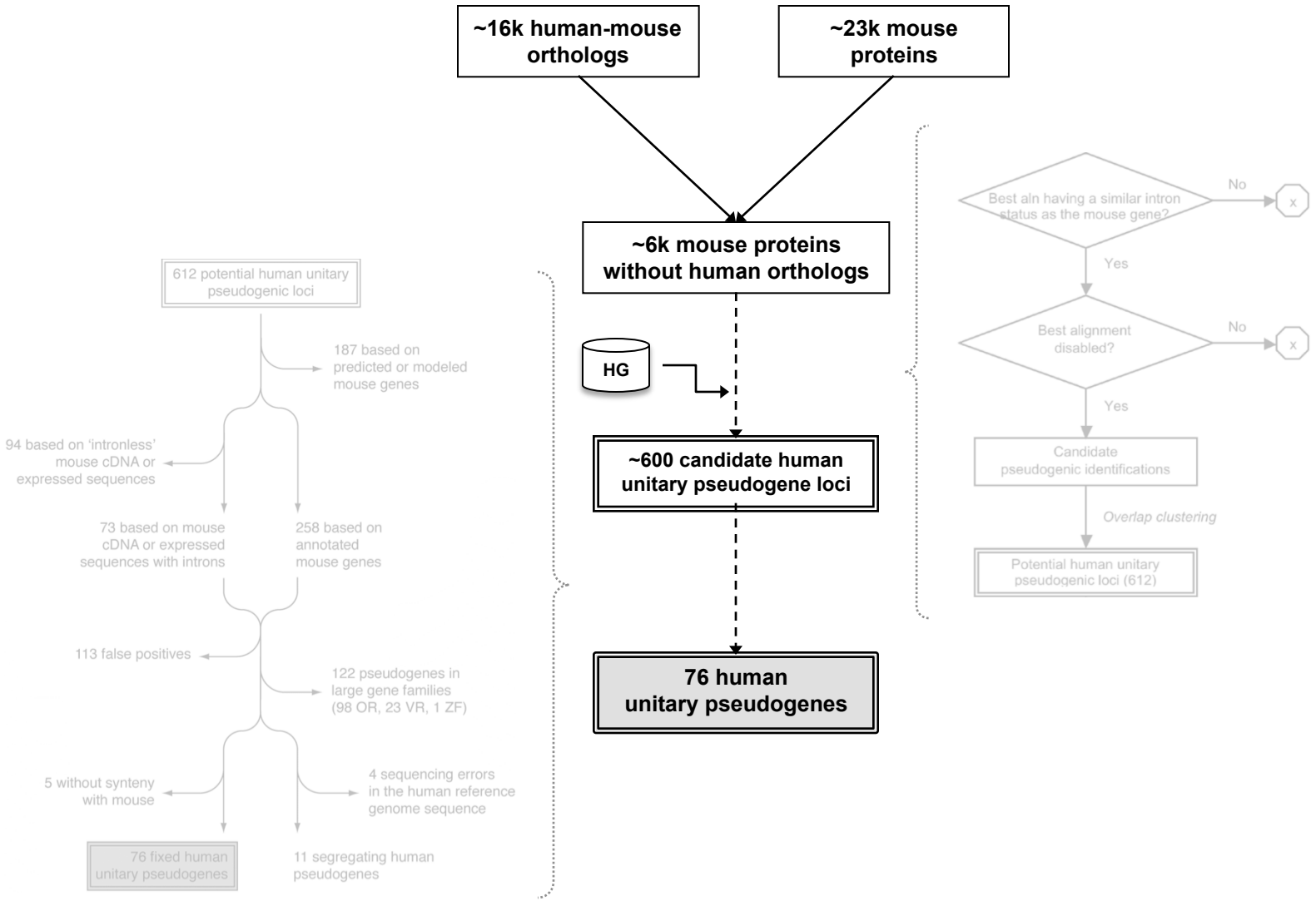
- Pseudogenes: nongenic DNA segments with high sequence similarity to functional genes



- **Unitary pseudogenes: unprocessed pseudogenes with no functional counterparts**

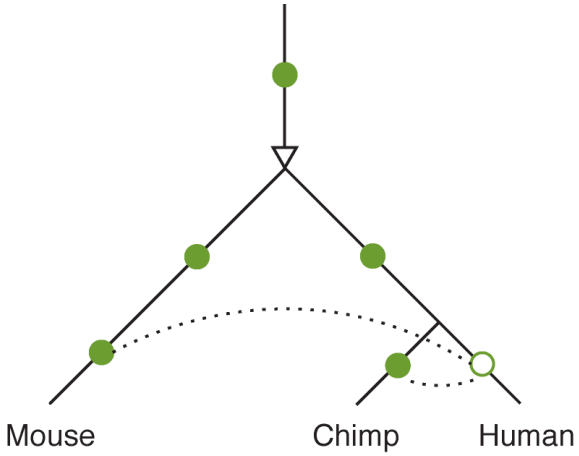


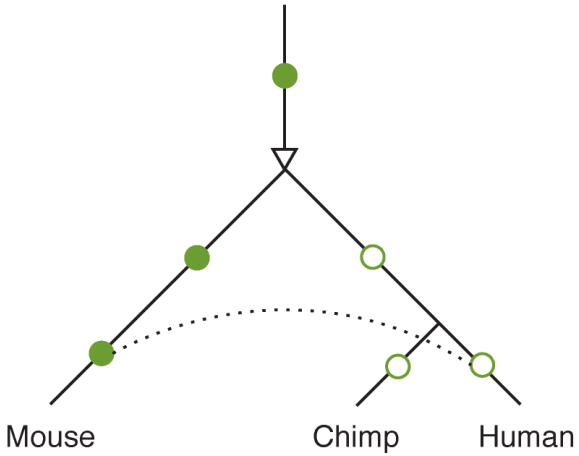
Identification pipeline

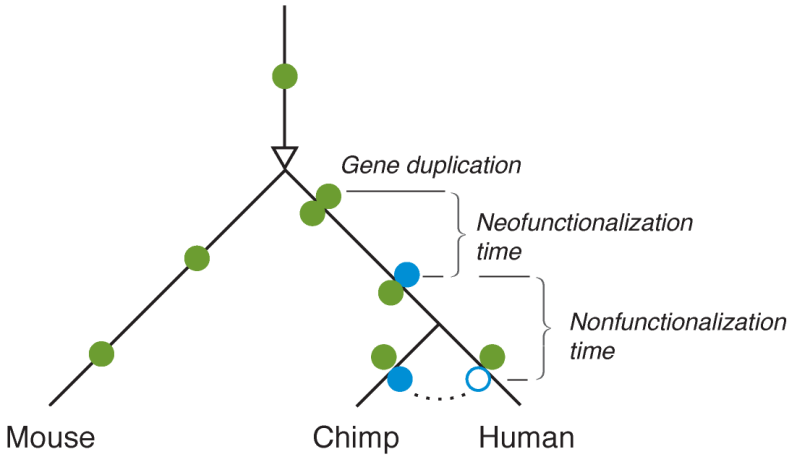


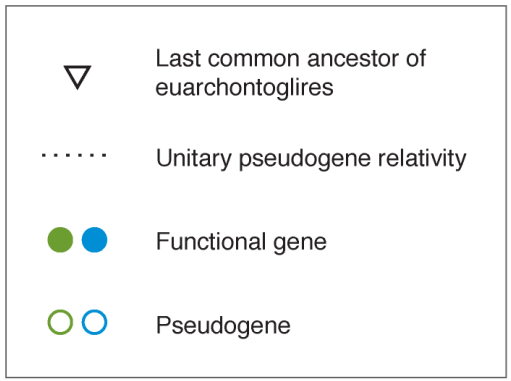
[Zhang et al. GenomeBiology (in press, '10)]

Relativity of unitary pseudogenes



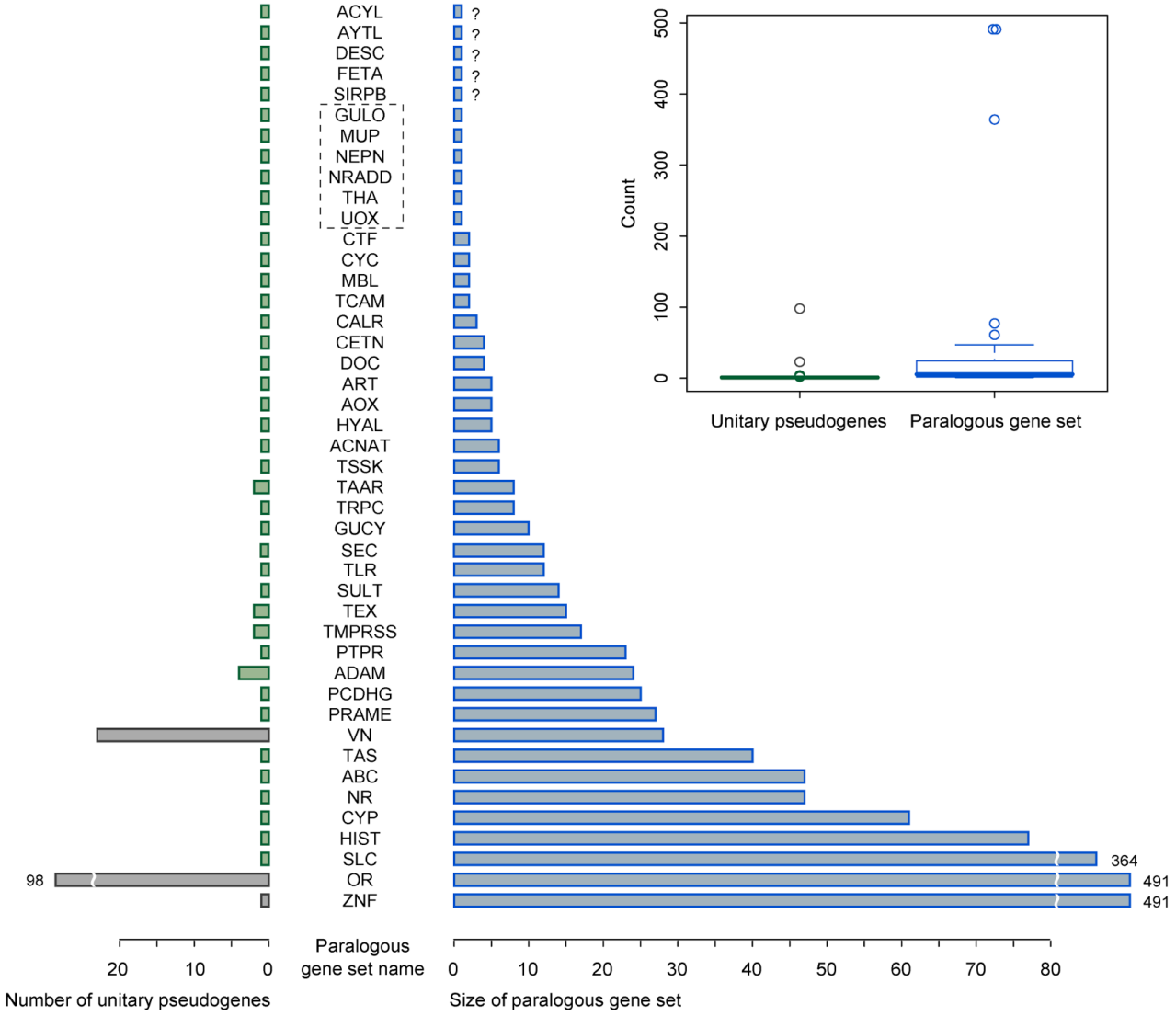




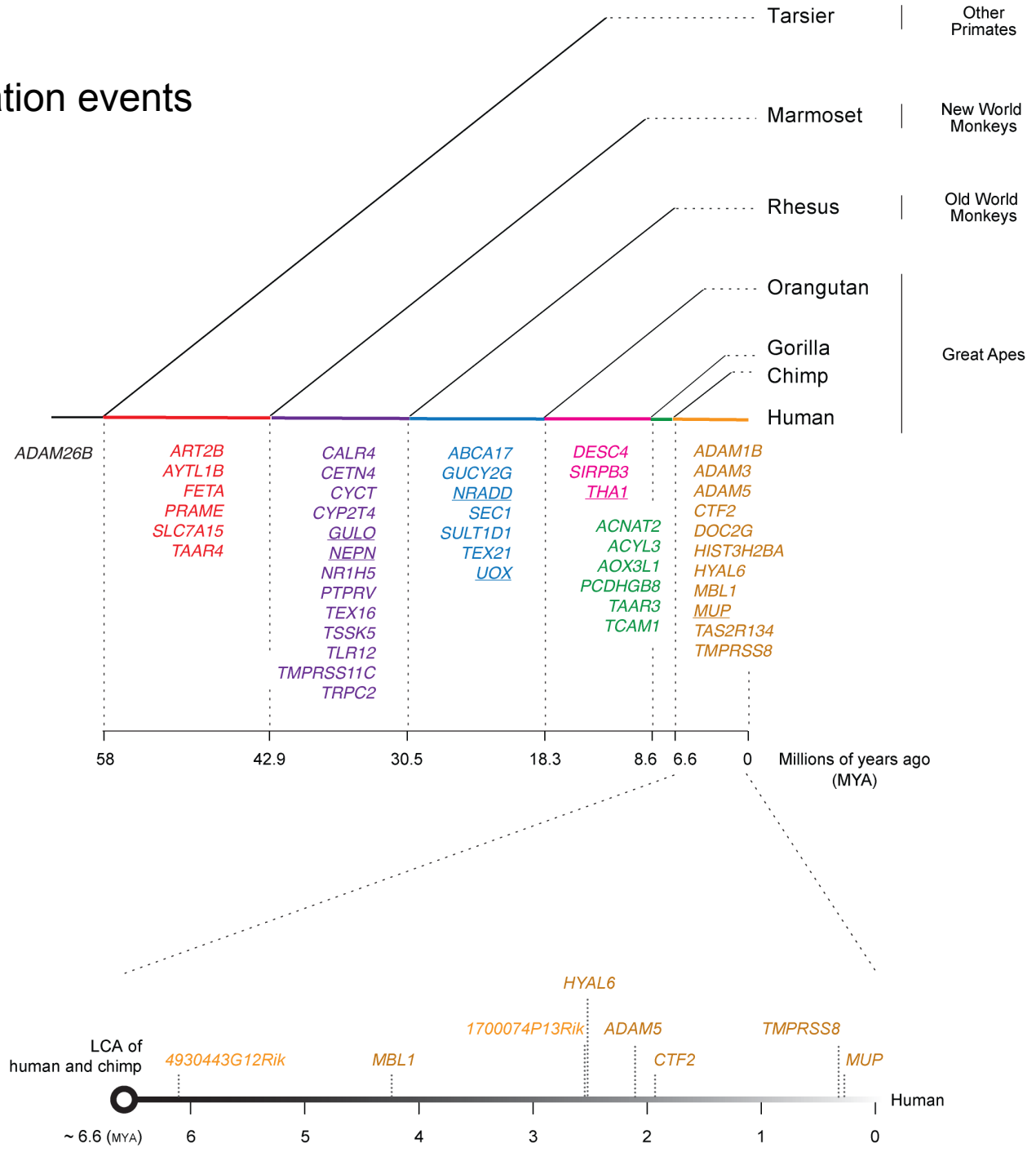


[Zhang et al. GenomeBiology (in press, '10)]

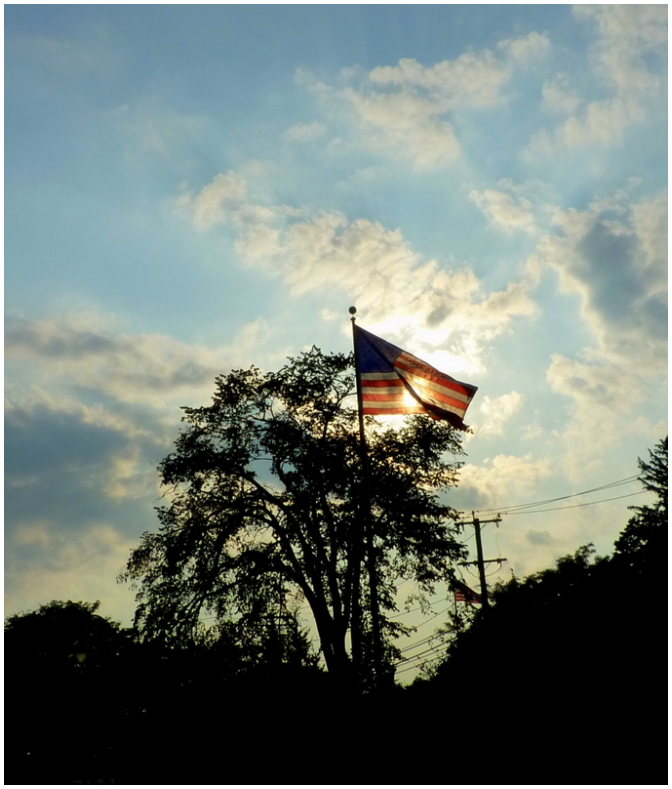
Unitary Pseudogene Families



Dating the pseudogenization events



Specific Pseudogene Assignments: Polymophic Pseudogenes (in process)



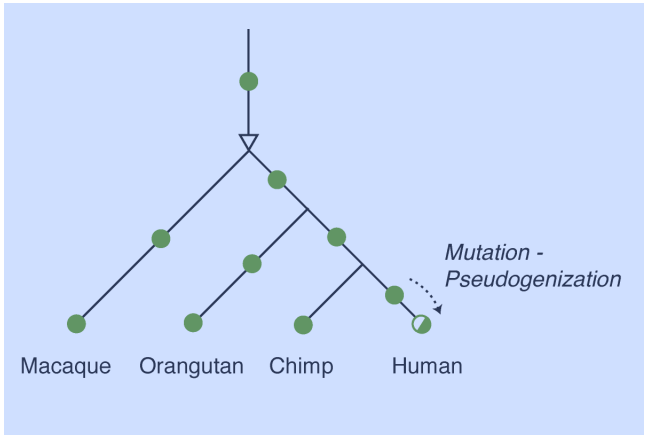
11 Polymorphic Pseudogenes

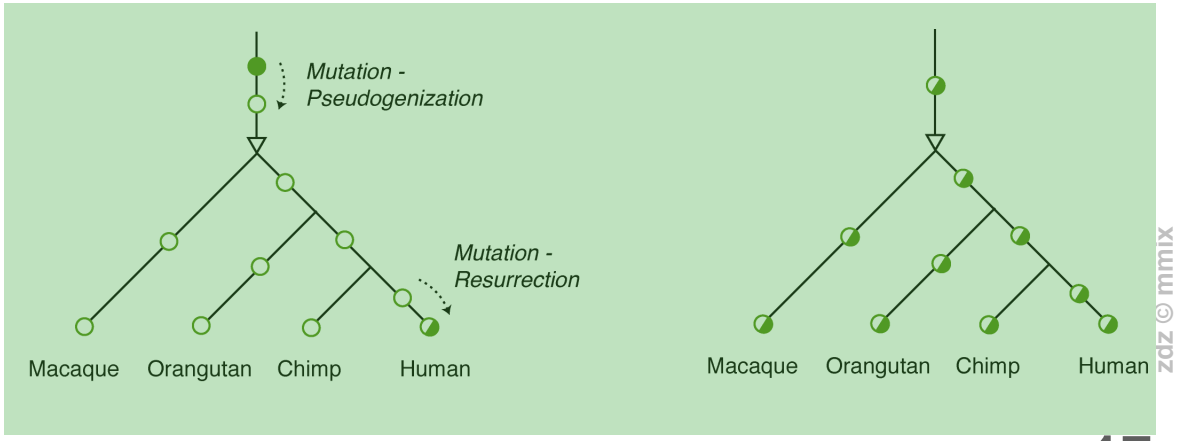
Table 2. Human polymorphic pseudogenes

Gene	CDS-disruptive mutation		dbSNP ID [3]	HapMap SNP ID
	Change [1]	Location [2]		
Nonsense mutation				
FBXL21	taT (Y) → taA	chr5+:135,300,350	rs17169429 (+27)	rs17169429 (+27)
FCGR2C	Cag (Q) → Tag	chr1+:159,826,011	rs3933769 (−60)	rs3933769 (−60)
GPR33	Cga (R) → Tga	chr14−:31,022,505	rs17097921	rs17097921
SEC22B	Caa (Q) → Taa	chr1+:143,815,304	rs2794062	rs16826061 (+95)
SERPINB11	Gaa (E) → Taa	chr18+:59,530,818	rs4940595	rs4940595
TAAR9	Aaa (K) → Taa	chr6+:132,901,302	rs2842899	rs2842899
Frame-shift mutation				
CASP12	ΔCA	chr11−:104,268,394-5	rs497116 (−67)	rs497116 (−67)
KRTAP7-1	ΔT	chr21−:31123841	rs35359062	rs9982775 (−20)
PSAPL1	∇A	chr4−:7,487,457	rs58463471	rs4484302 (+441)
TMEM158	∇A	chr3−:45,242,396	rs11402022	rs33751 (+725)
TPSB2	ΔC	chr16−:1,219,240	rs2234647	rs2745145 (−1771)

Polymorphic pseudogenes (3 with allele frequency data)

CDS-disrupted gene	GPR33	SERPINB11	TAAR9
Disruptive mutation [1]	Cga (R) → Tga	Gaa (E) → Taa	Aaa (K) → Taa
dbSNP ID	rs17097921	rs4940595	rs2842899
Genomic location	chr14−:31,022,505	chr18+:59,530,818	chr6+:132,901,302
Disrupted codon position [2]	140 (332)	89 (388)	61 (344)
Reference allele in human	T	T	T
Reference allele in other primates [3]	C	T	T
Allele frequency [4]			
Test statistic for HWE in the meta-population [5]	0.285 ($P = 0.867$)	8.659 ($P = 0.013$)	0.071 ($P = 0.965$)

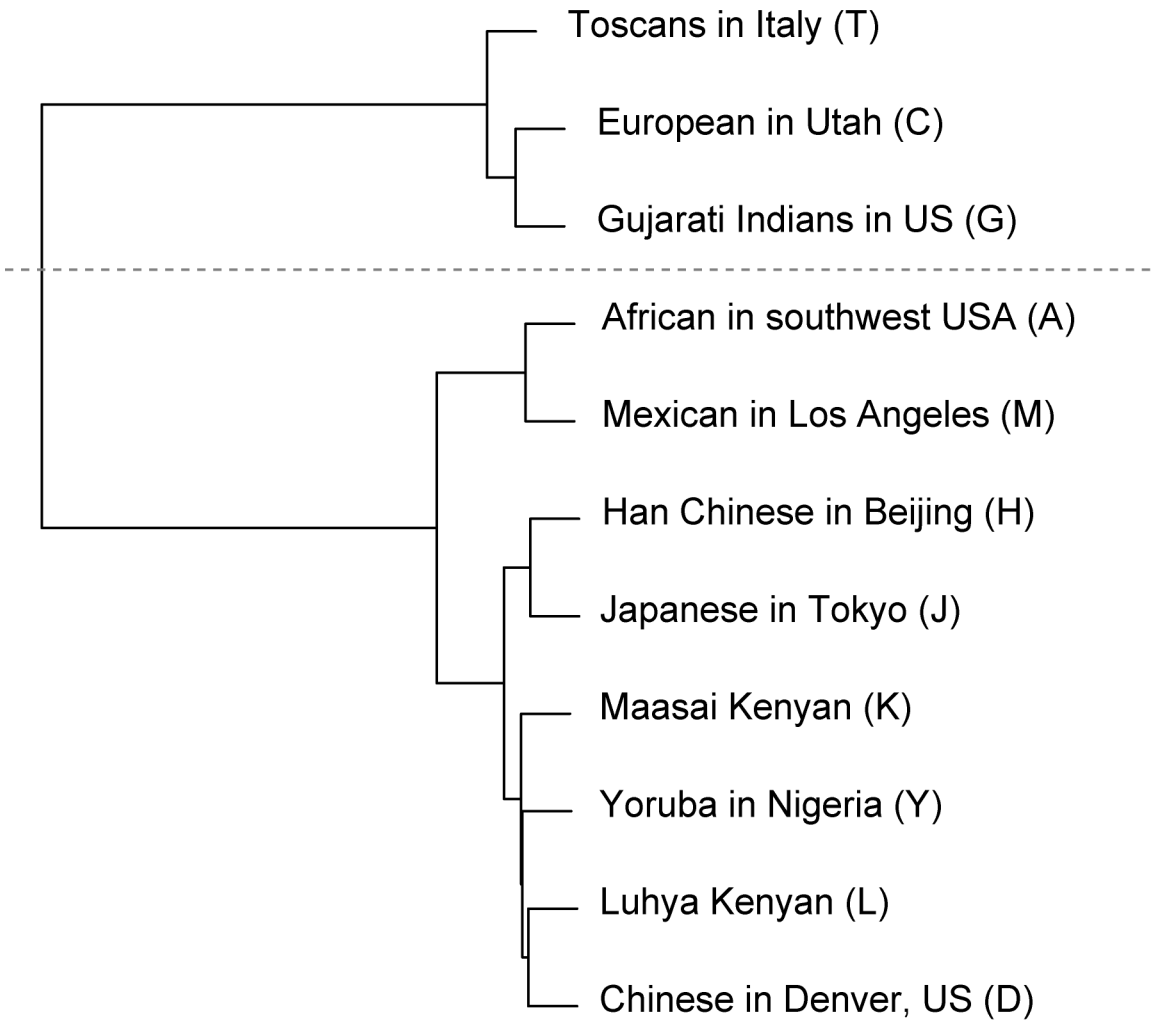




[Zhang et al. GenomeBiology (in press, '10)]

3 SNPs not found to be under recent positive selection....

F_{st} hierarchical clustering for rs4940595 in SERPINB11



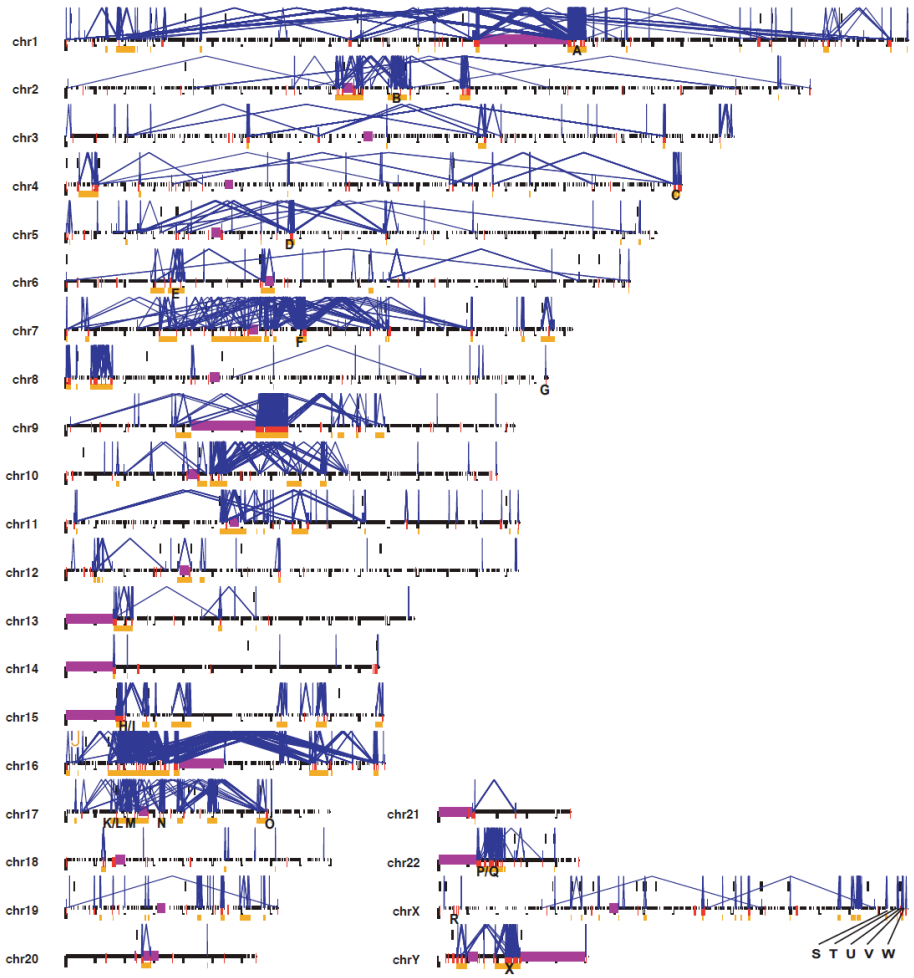
▪▪▪▪but population structure at rs4940595—the difference in the allelic frequencies in different populations—could be result of different selective regimes that the same allele at rs4940595 is subjected to in different population subdivisions.

Specific Pseudogene Assignments: SD-associated Pseudogenes (in process)

Segmental duplications (SDs)

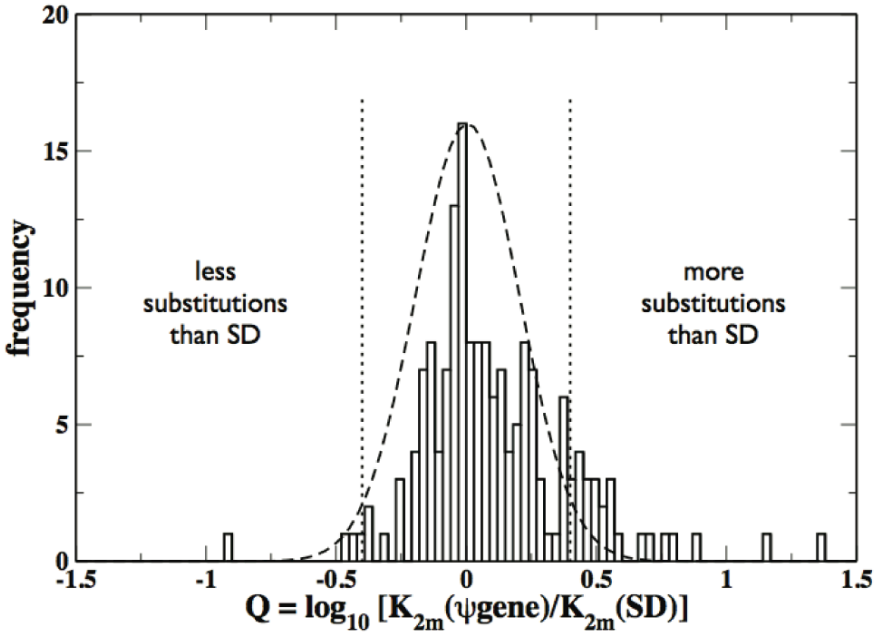
- Regions of the genome with ≥ 90% sequence identity and ≥ 1kb in length
- Based on neutral divergence correspond to last ~40 million years of human evolution
- Comprise ~5-6% of the human genome
- Enriched with genes (~18%) and pseudogenes (duplicated ~45%, processed ~22%)

Can the study of ψgenes in SDs provide information not obvious from individual dataset ?



Bailey et al, Science, 2002

Nucleotide substitutions in ψ genes and SDs containing them



Parent gene

Duplicated ψgene

K_{2m} : Nucleotide substitutions per site computed using Kimura's two parameter model

Most ψgenes show the same number of substitutions as larger SD region containing them

- Duplication accompanied by disablement
- Followed by neutral rate of evolution

Acknowledgements

Z Zhang
E Khurana
Y J Liu
YK Lam
S Balasubramanian
G Fang
N Carriero
R Robilotto
P Cayting
M Wilson
A Frankish
M Diekhans
R Harte
T Hubbard
J Harrow

Pseudogene.org