On the Origin and Fate of Gene Duplicates in Mammalian Genomes

Jin Jun, Edward Hemphill, Ion Mandoiu, Craig Nelson

University of Connecticut

o

Paul Ryvkin

University of Pennsylvania

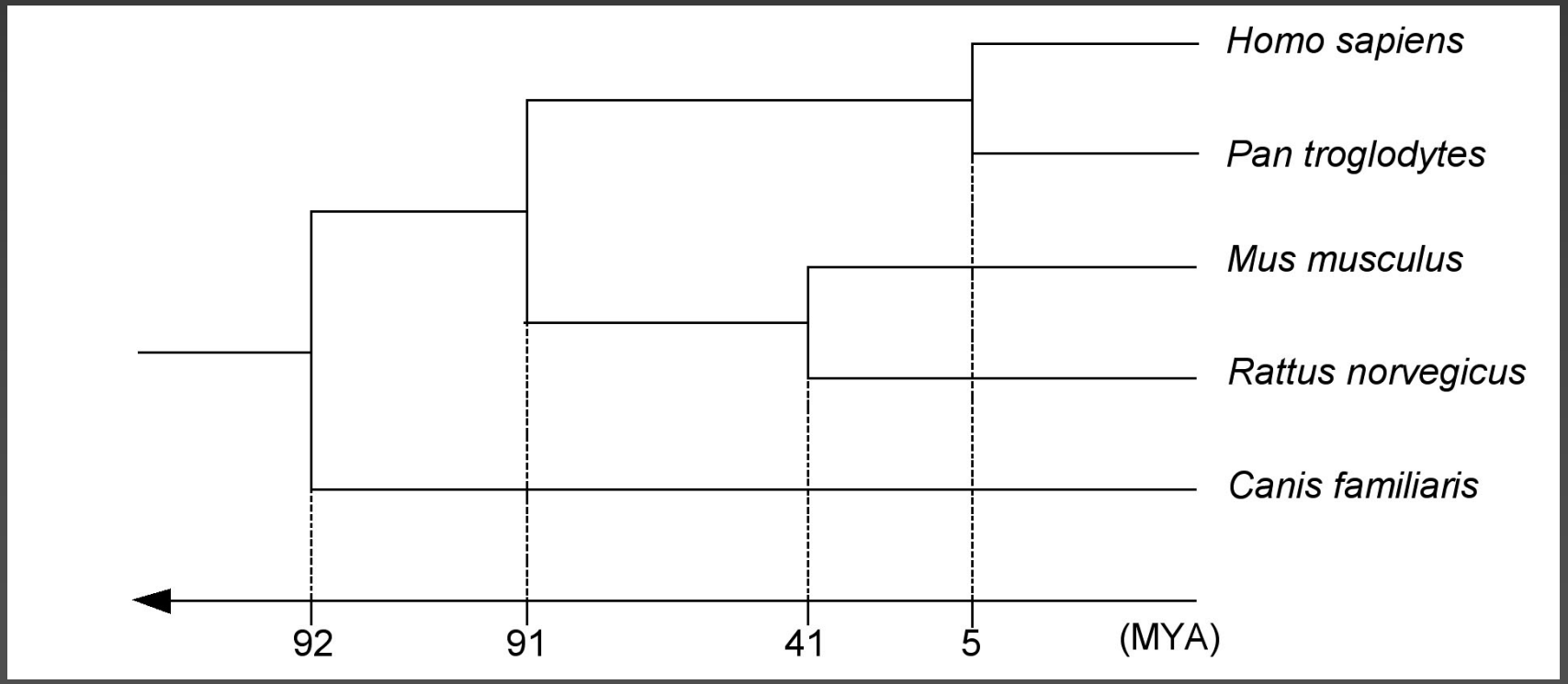
Genome Evolution by Gene Duplication

- Gene duplication has been recognized as a major force in mammalian genome evolution
  - ~720 human-specific gene duplications (Zhang 2002)
- DNA- vs. RNA-mediated duplication mechanisms
  - DNA-mediated
    - Occurs continuously
    - Contributed significantly to the divergence of gene content
  - RNA-mediated
    - Occurs quite frequently
    - Believed to be non-functional (pseudogenes) due to lack of regulatory material of the parental gene
    - Recently functional retro-copies in human/mouse genomes detected (Sakai et al. 2007, Vinckenbosch et al. 2006)

Characteristics of Duplication Mechanisms

- DNA-mediated (segmental duplication or SD) copies
  - Syntenic to each other
  - Share introns
- RNA-mediated (retrotransposed or RT) copies
  - Non-syntenic to parent gene
  - Intronless or share no introns with parent gene
- Pseudogenes
  - Majority of retrotransposition events generate processed pseudogenes
  - Some of SD copies can be degraded to non-processed pseudogenes

Mammalian Genomes in Our Study



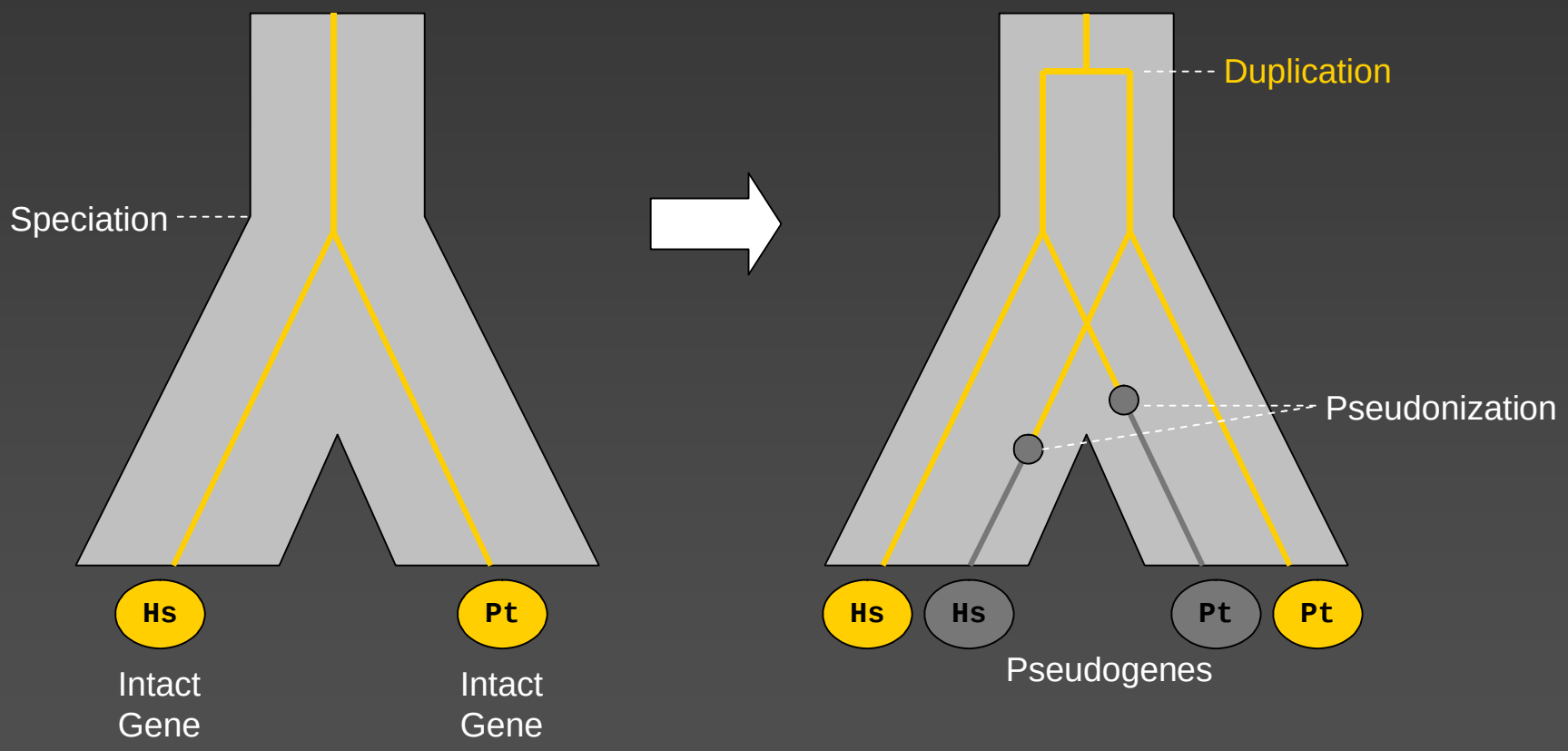
Our Approach

- Included pseudogenes in our analysis to capture more duplication events.

- Avoided using the coding sequence similarity to reconstruct the evolutionary history of gene family.
  - Local synteny information and gene structure information were used to distinguish SD events from RT events.

➔ Able to measure the functional constraint based on the ancestral evolutionary history.

Capturing More Duplication Events by Using Pseudogenes

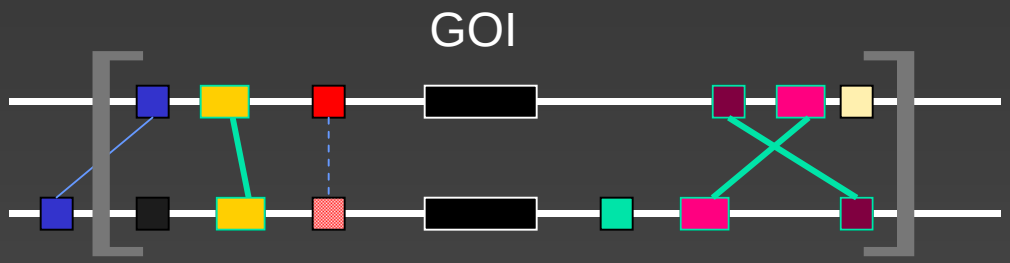


- No duplication event
- Functional orthologs

- One duplication event
- Ancestral orthologs

Measuring Local Synteny

Three neighboring genes on each up/down stream are considered



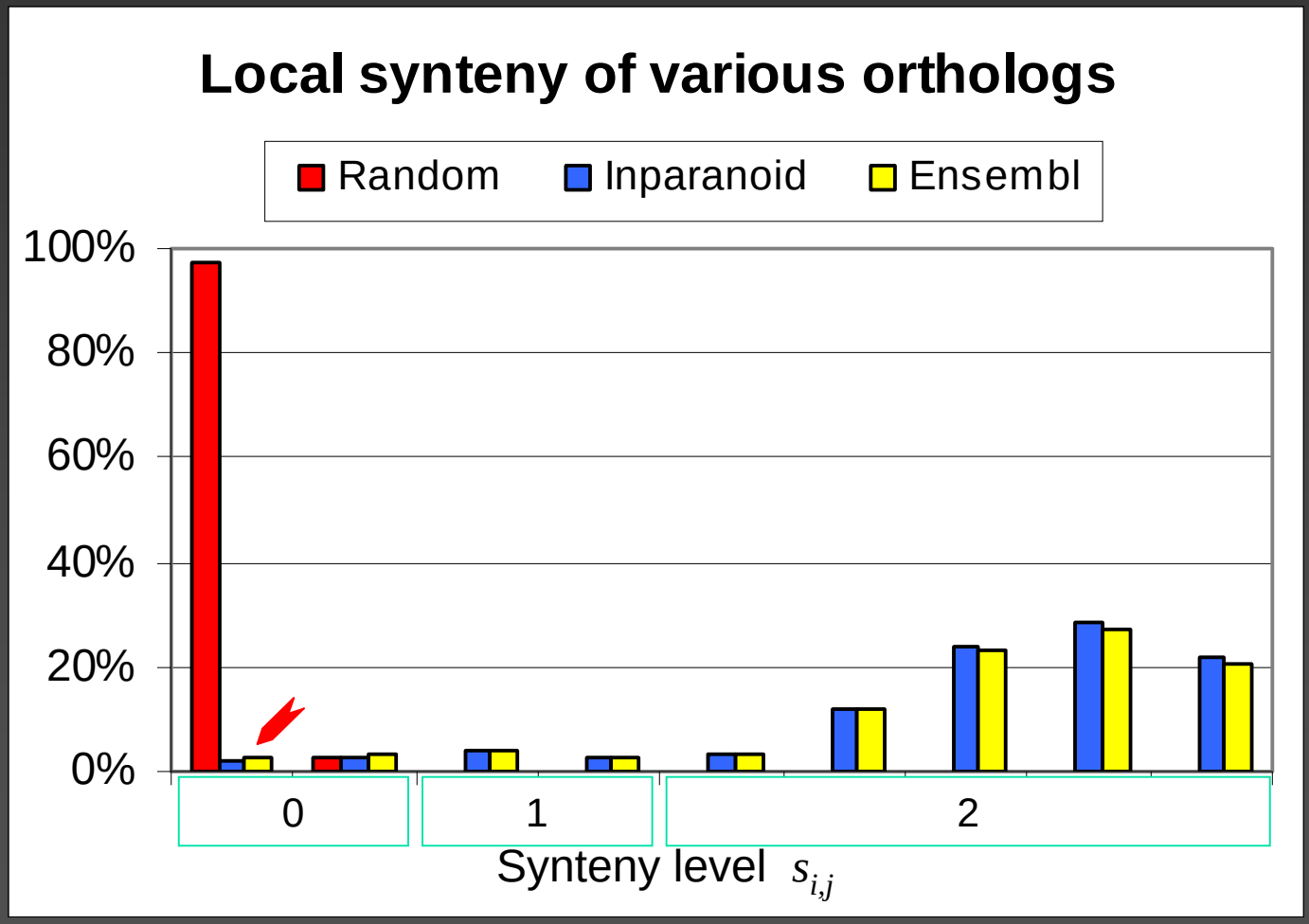
Homologous matches:
BlastP score > 50 and
sequence similarity > 80%

Weak homologous matches

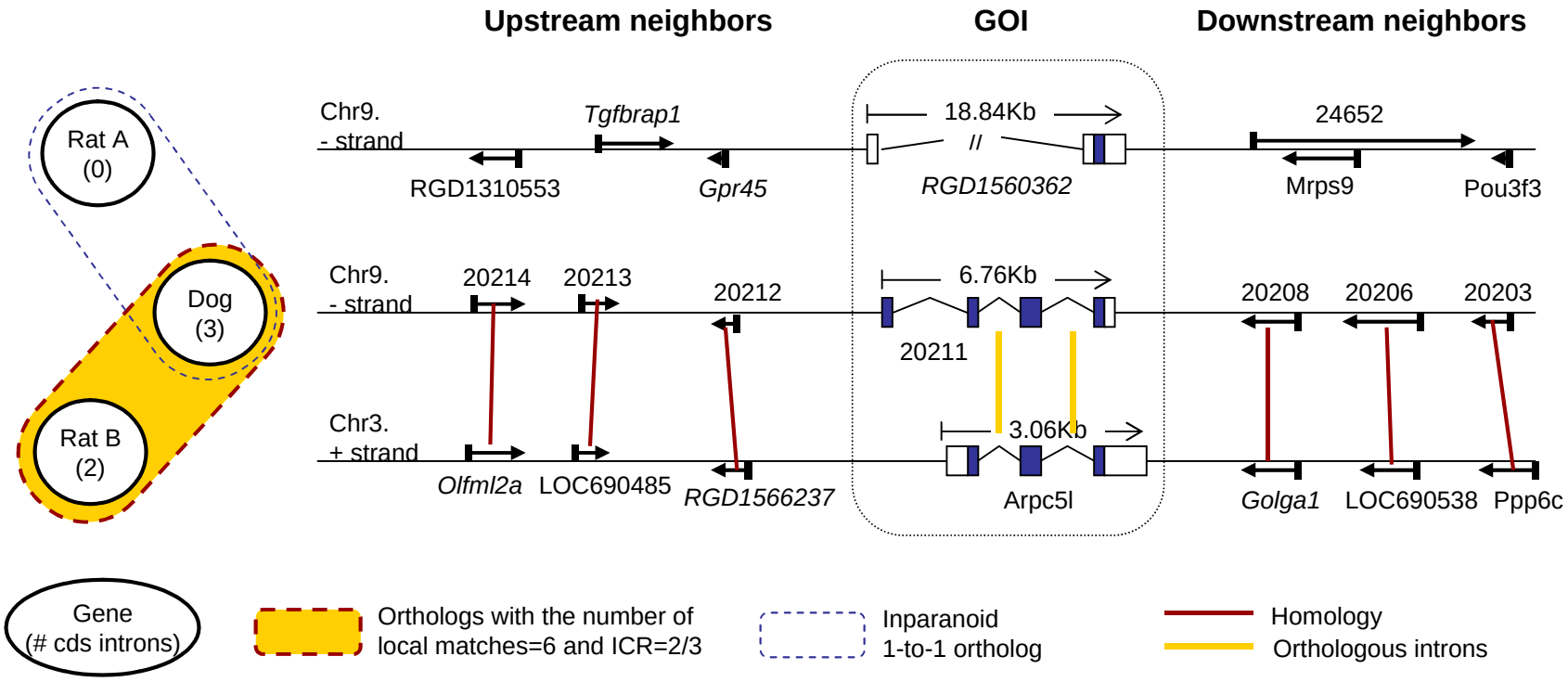
Homologous genes

(# matched sides, # matches) = (2, 3)

FP and FN rates of Local Synteny

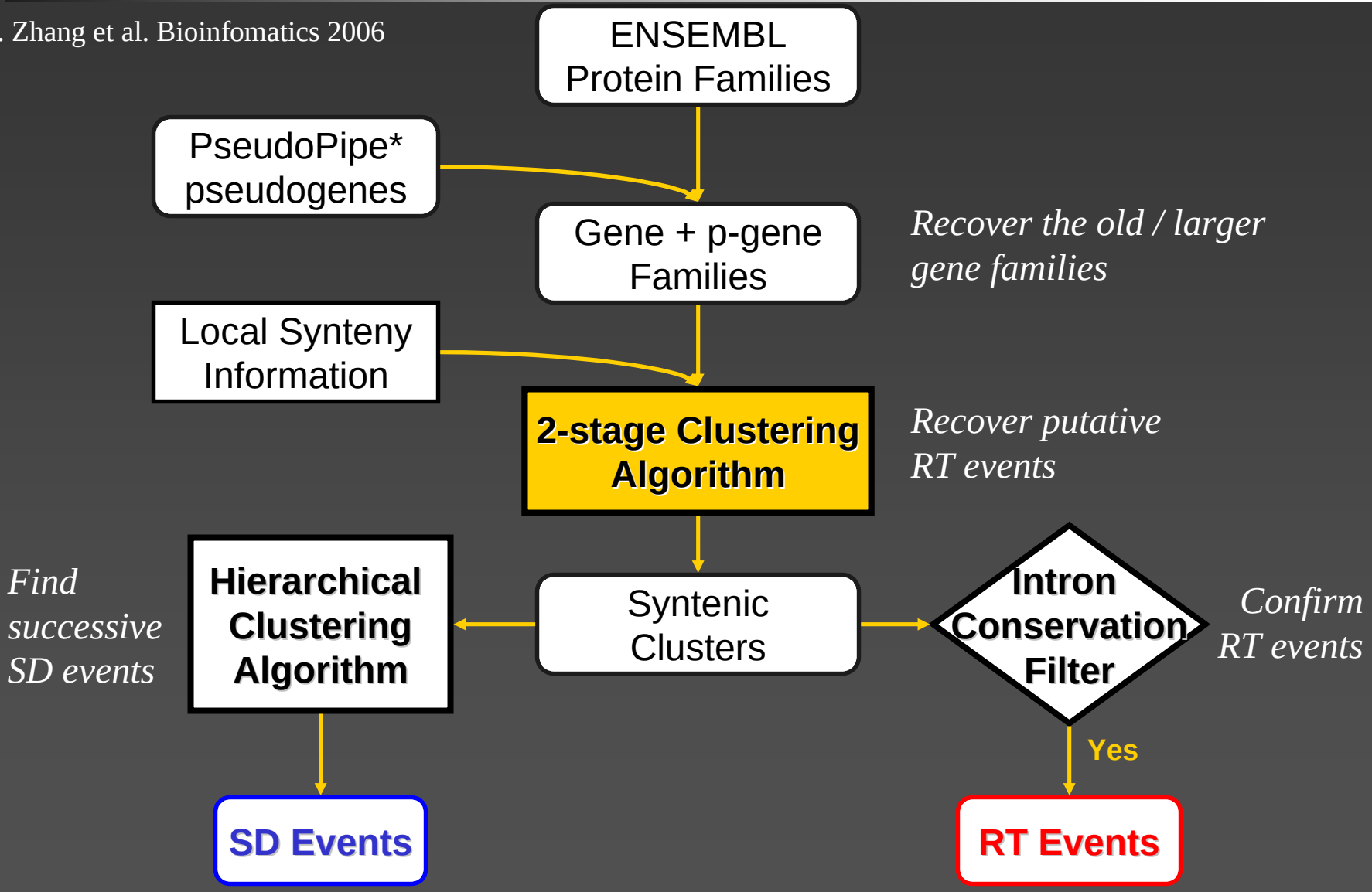


RT Miscall by Inparanoid



Duplication Event Identification

*. Zhang et al. Bioinfomatics 2006



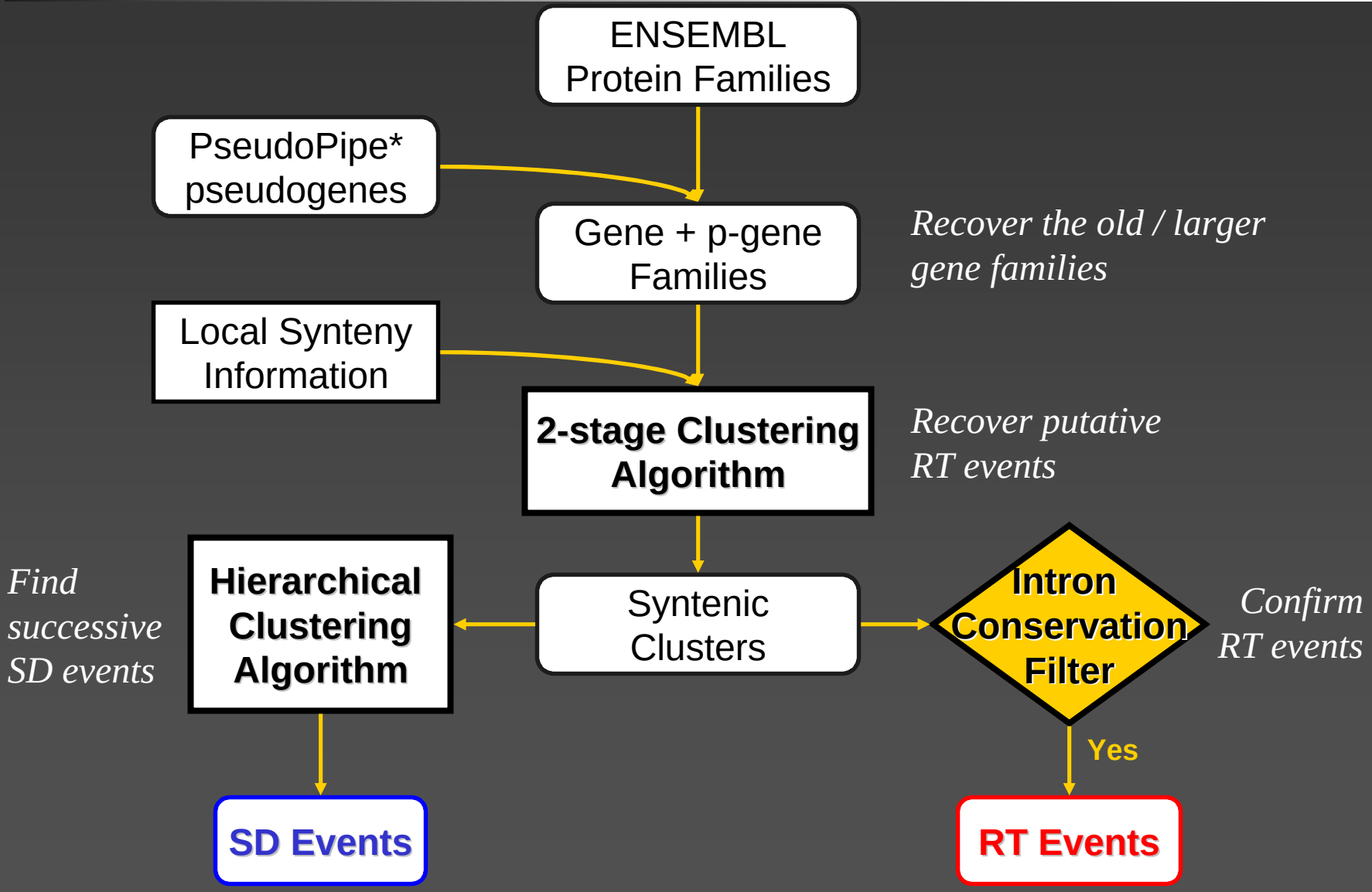
Two-Stage Clustering Algorithm

- Stage 1: Single-linkage clustering with synteny level 2

- Stage 2: Complete-linkage clustering with synteny level 1
  - Considering the phylogenetic structure

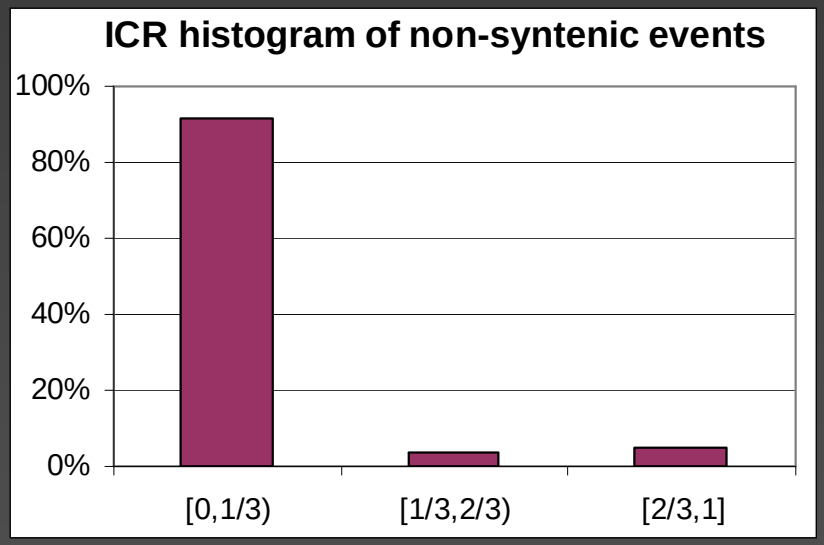
➔ **Result: syntenic clusters**
  - Any member within clusters: from SD or speciation events
  - Between these clusters: RT events or old SD events with loss of local synteny

Duplication Event Identification



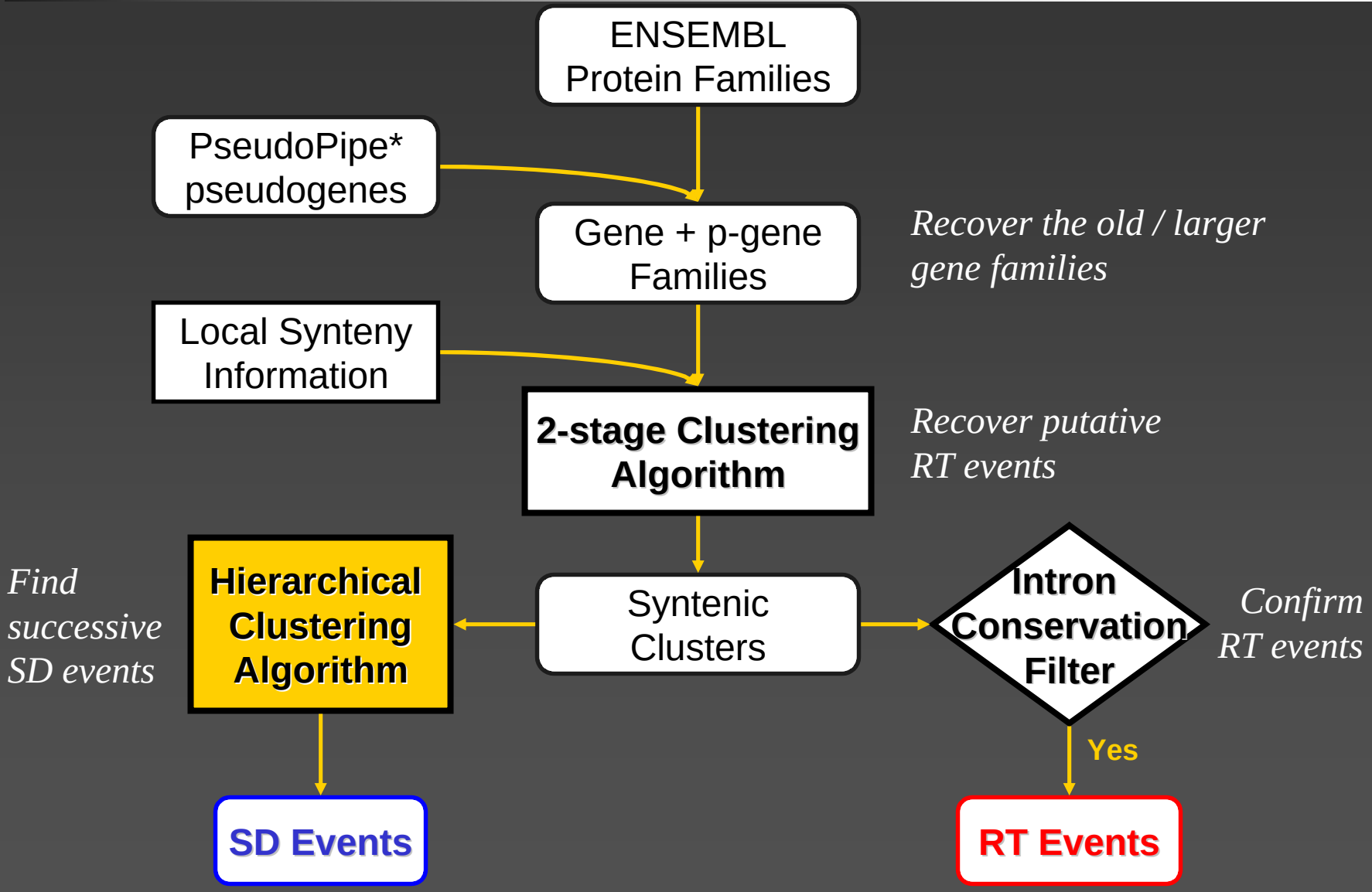
Intron Conservation Filter for RT events

$$\text{ICR (Intron Conservation Rate)} = \frac{\text{\# positional orthologous introns}}{\text{\# total introns positions}}$$



➔ **Non-syntenic + low ICR → RT event**

Duplication Event Identification

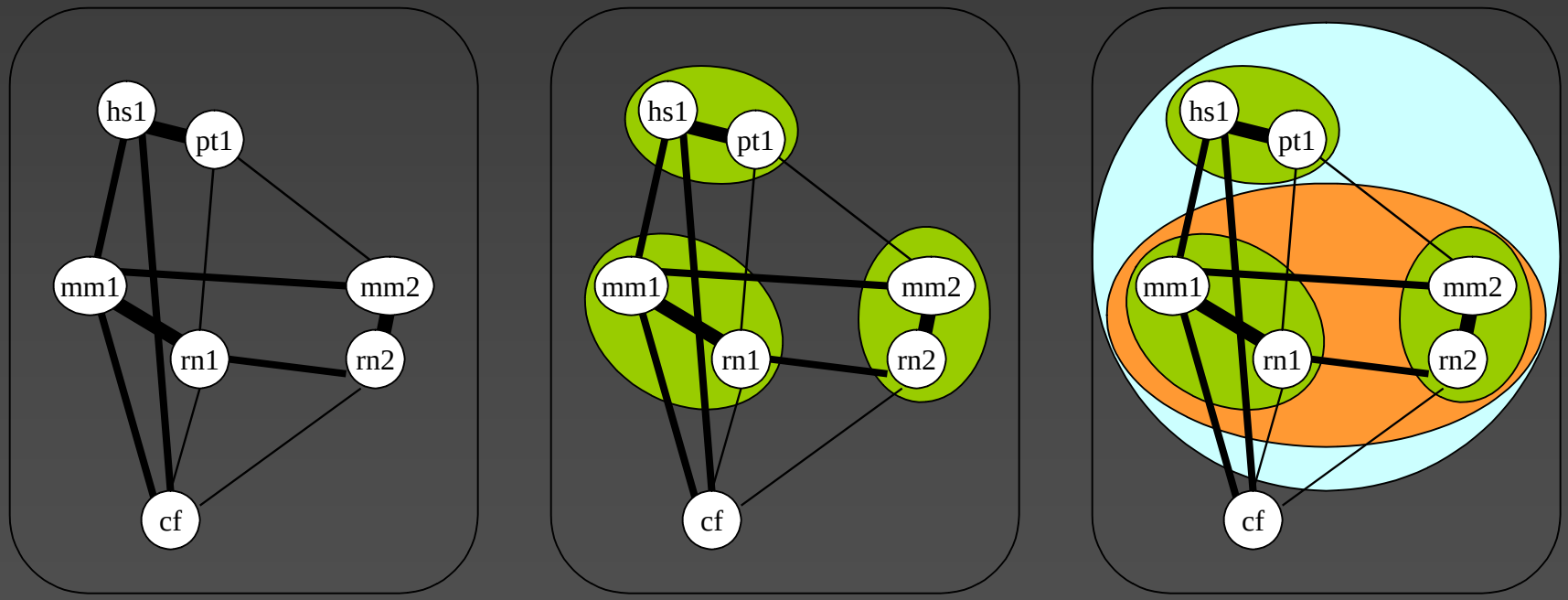


Hierarchical Clustering for Successive SD events

- Pearson's correlation coefficients
- Hierarchical clustering (UPGMA)

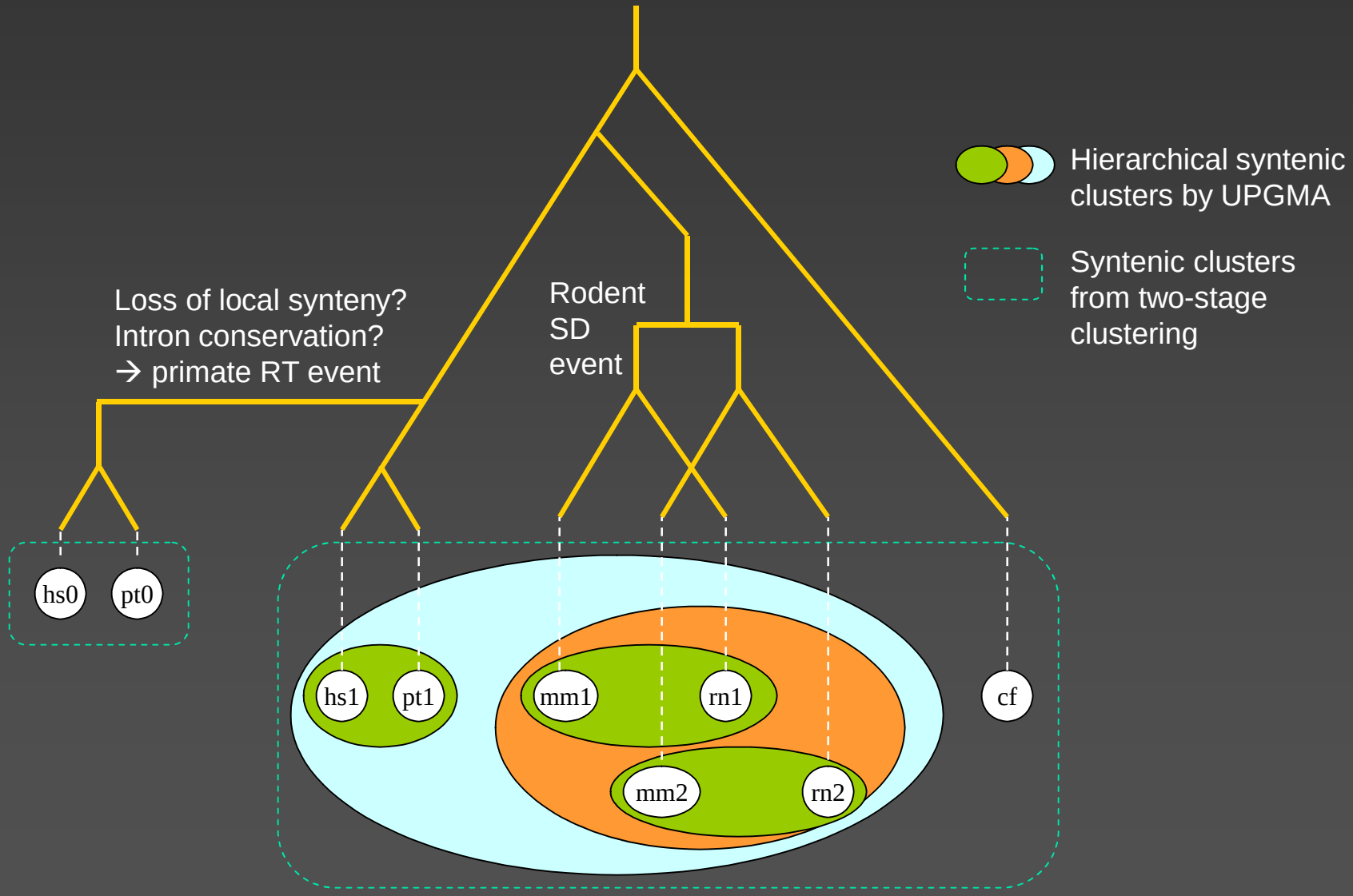
Line thickness corresponds to Pearson's correlation coefficient

Hierarchical syntenic clusters from UPGMA

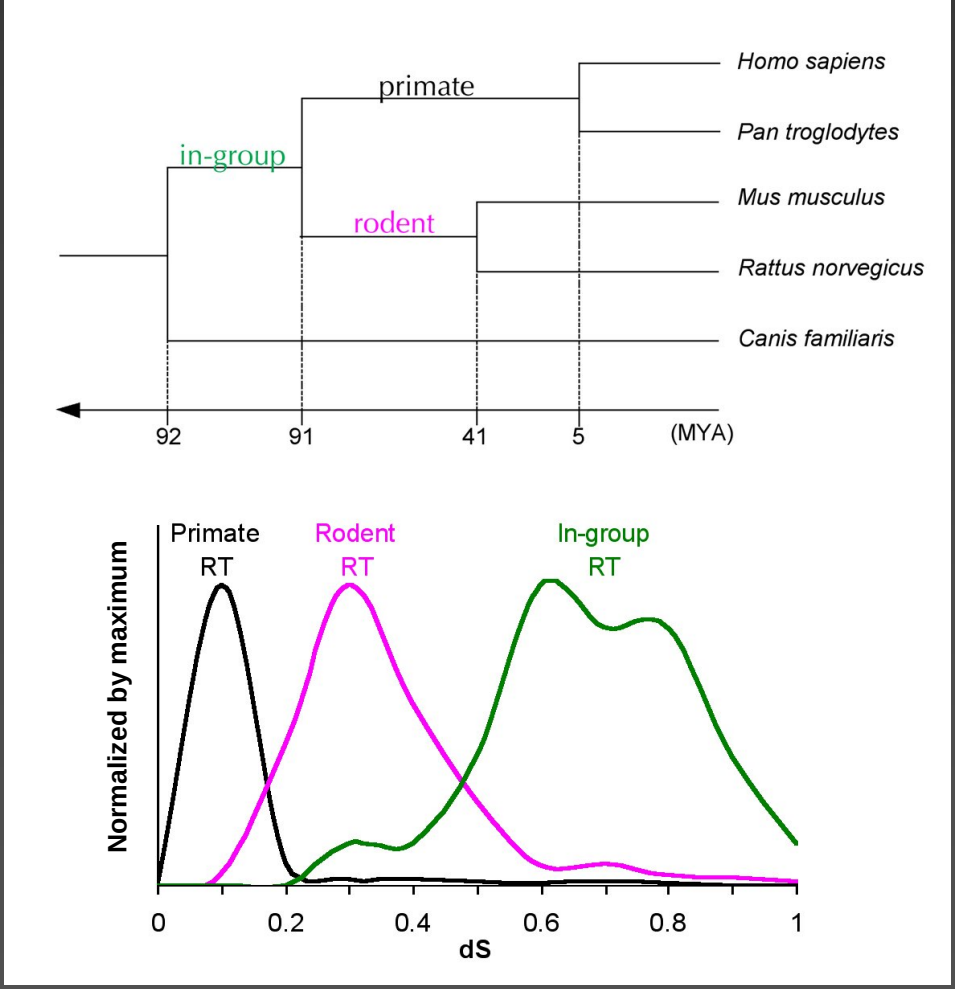


➔ **Result: hierarchy of SD/speciation events**
shown by different degree of synteny

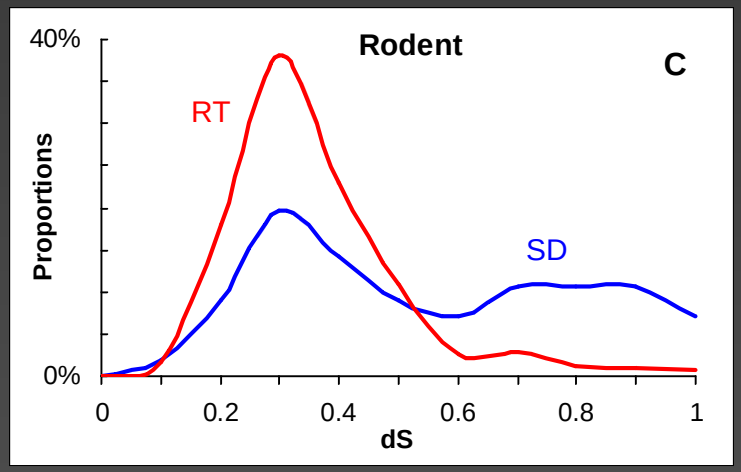
Inferring SD and RT events

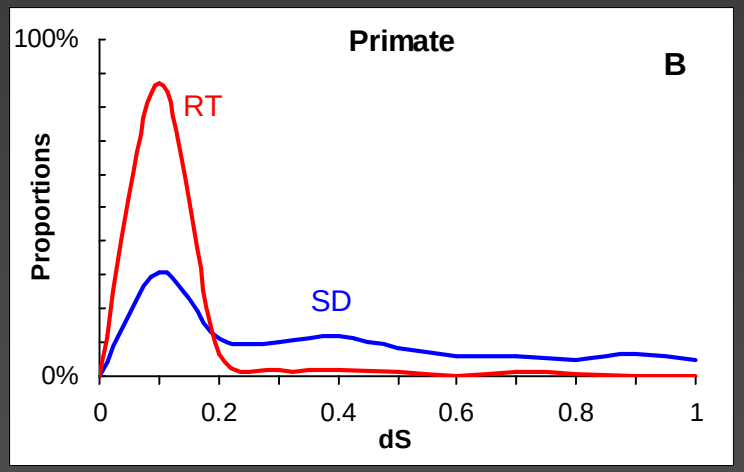


Age Distribution of RT Events Are Consistent With the Bursts of Retrotransposition Events on Mammalian Genomes

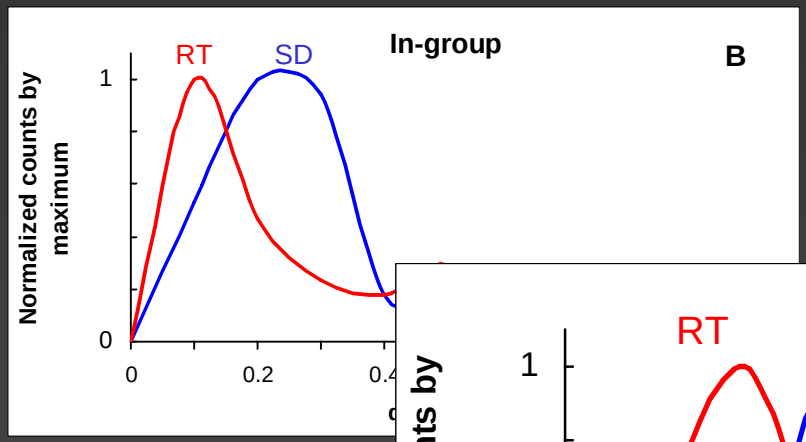


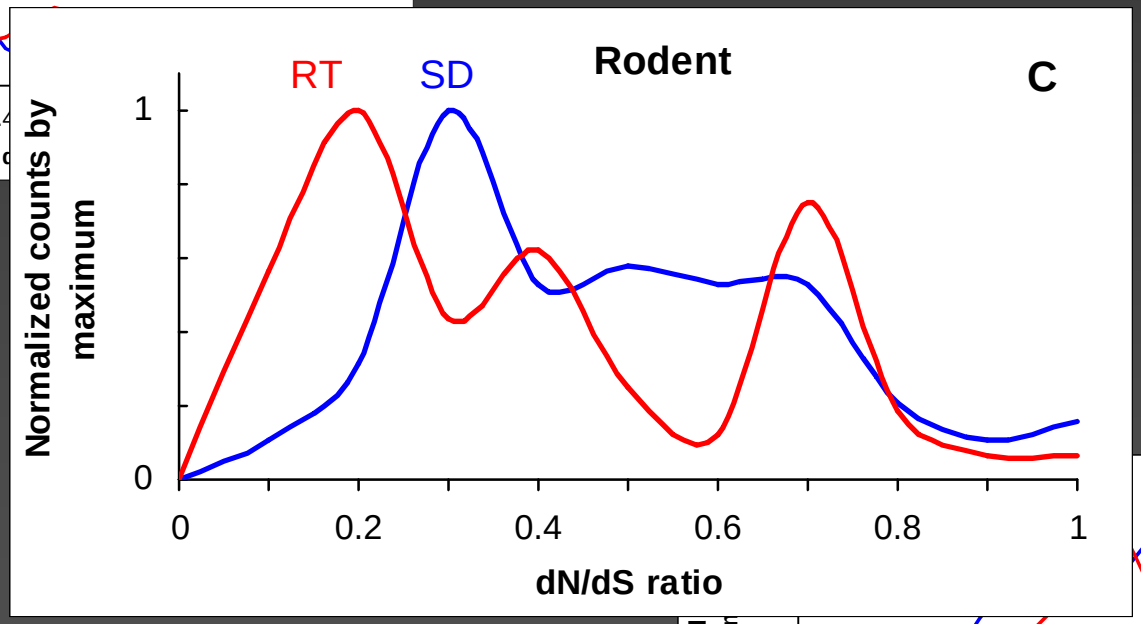
SD Events Occur at a More Stable Rate Than RT Events

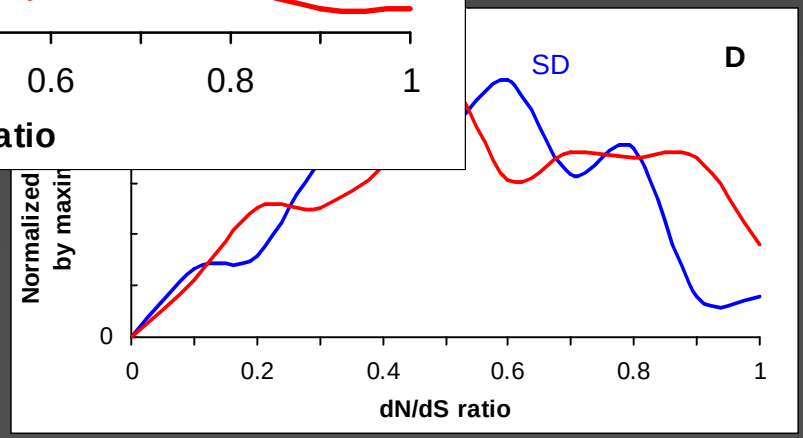




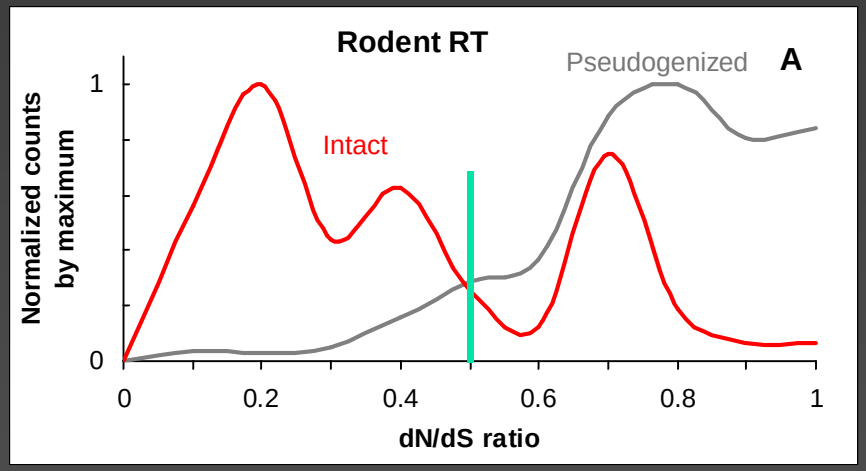
Each Mechanism Is Under Similar Levels of Constraint on Their Protein Coding Regions

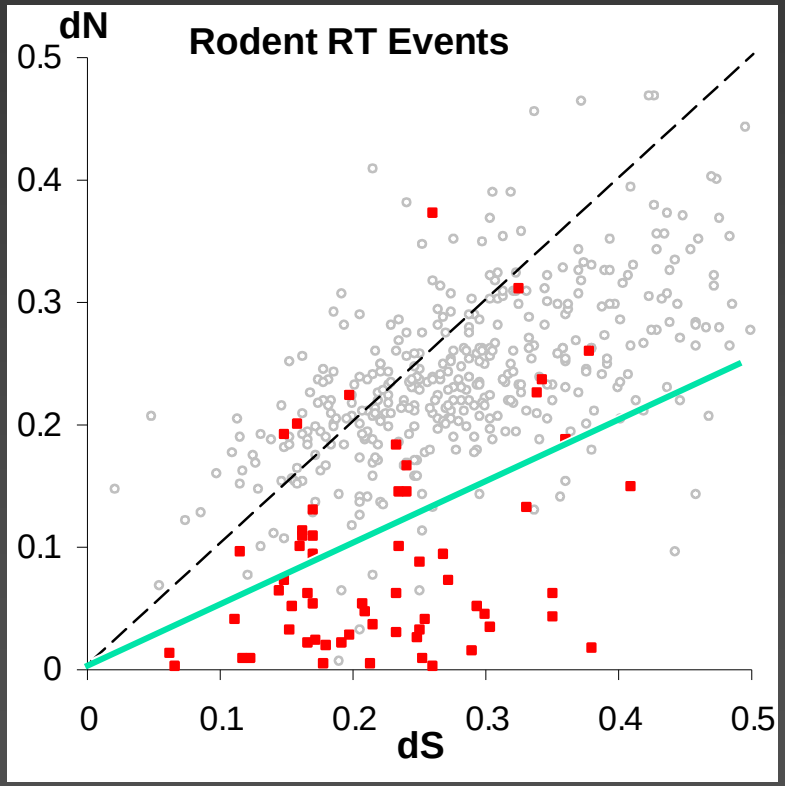






Many Non-disabled Gene Predictions Show No Evidence of Purifying Selective Pressure



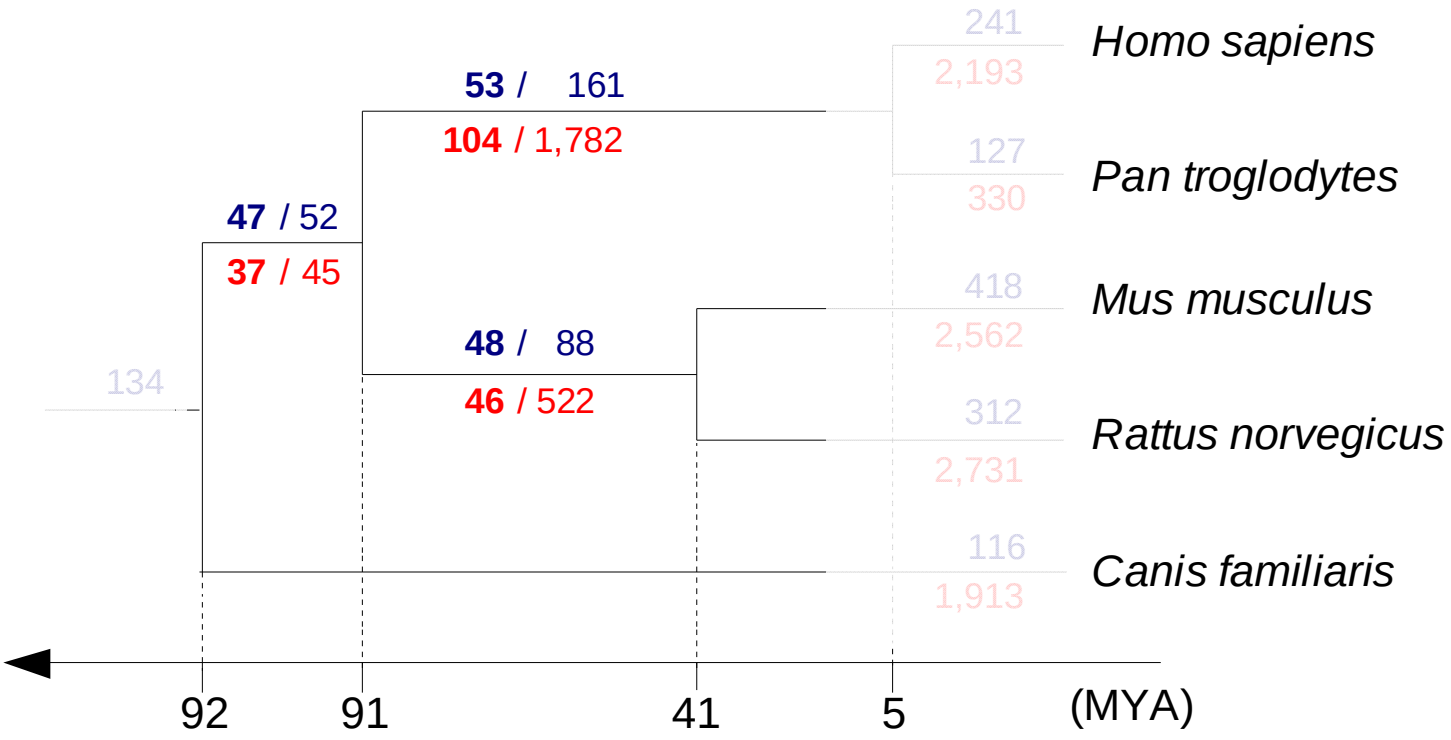


dN/dS = 0.5

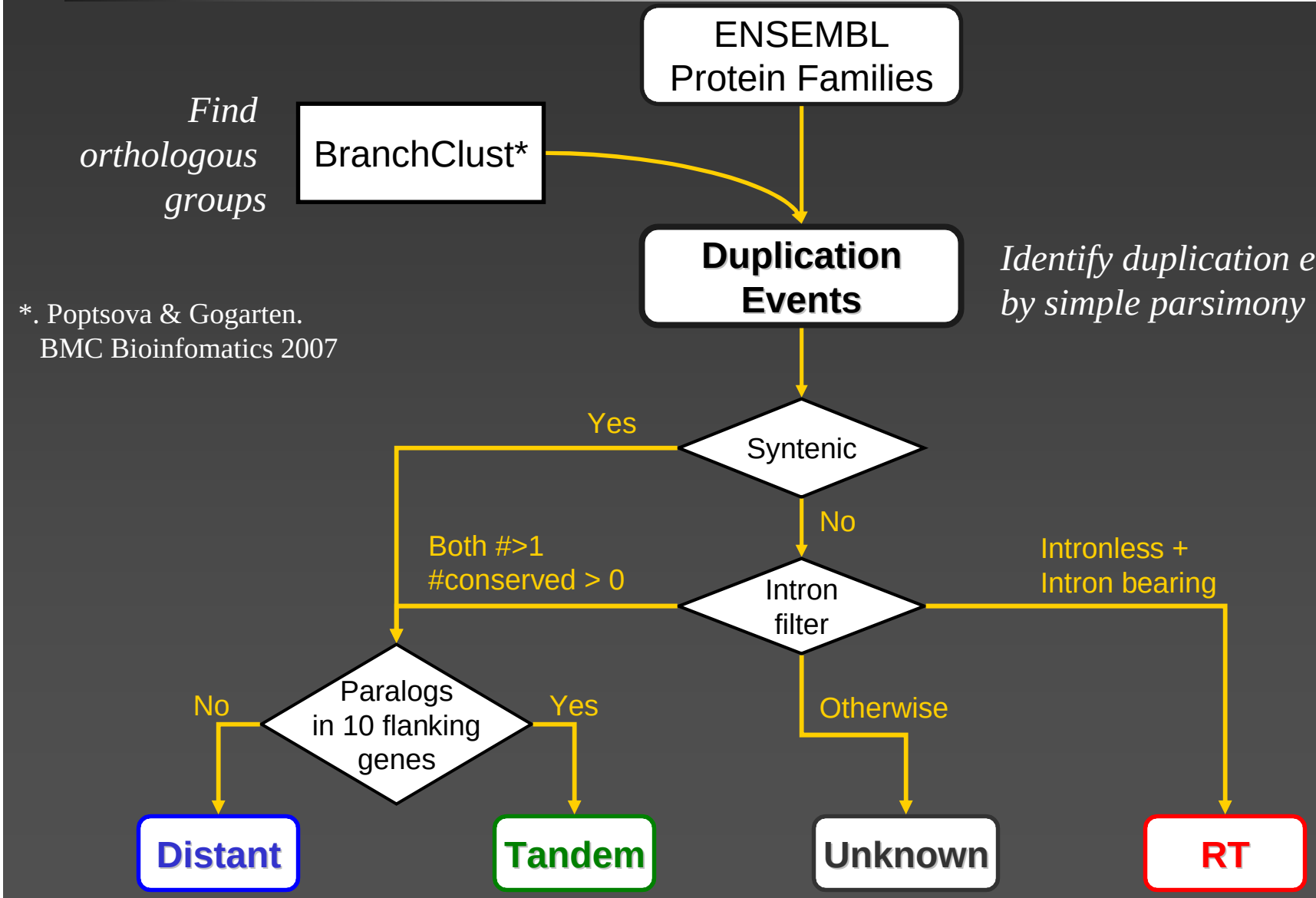
“Functional” events : dN/dS ratios < 0.5

Retrotransposition Contributes Roughly Equal Numbers of New Functional Genes

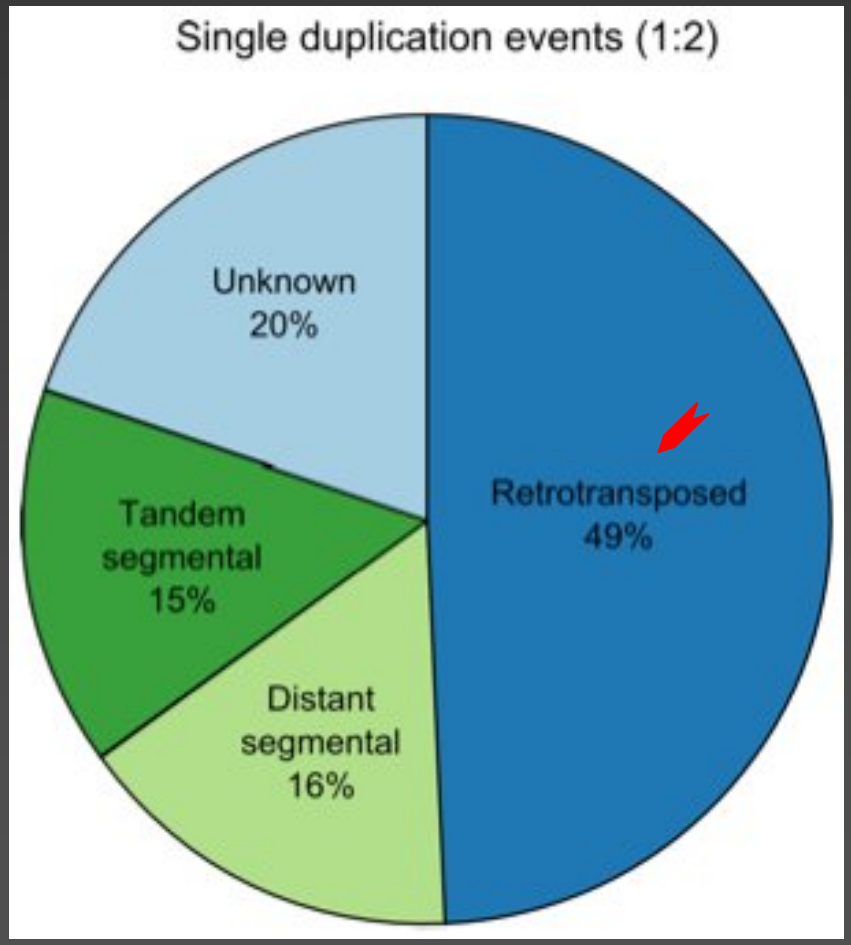
		Three internal branches	Whole tree
SD	**functional** / total events	**148** / 301 = 49.17%	1,649
RT	**functional** / total events	**187** / 2,349 = 7.96%	12,078

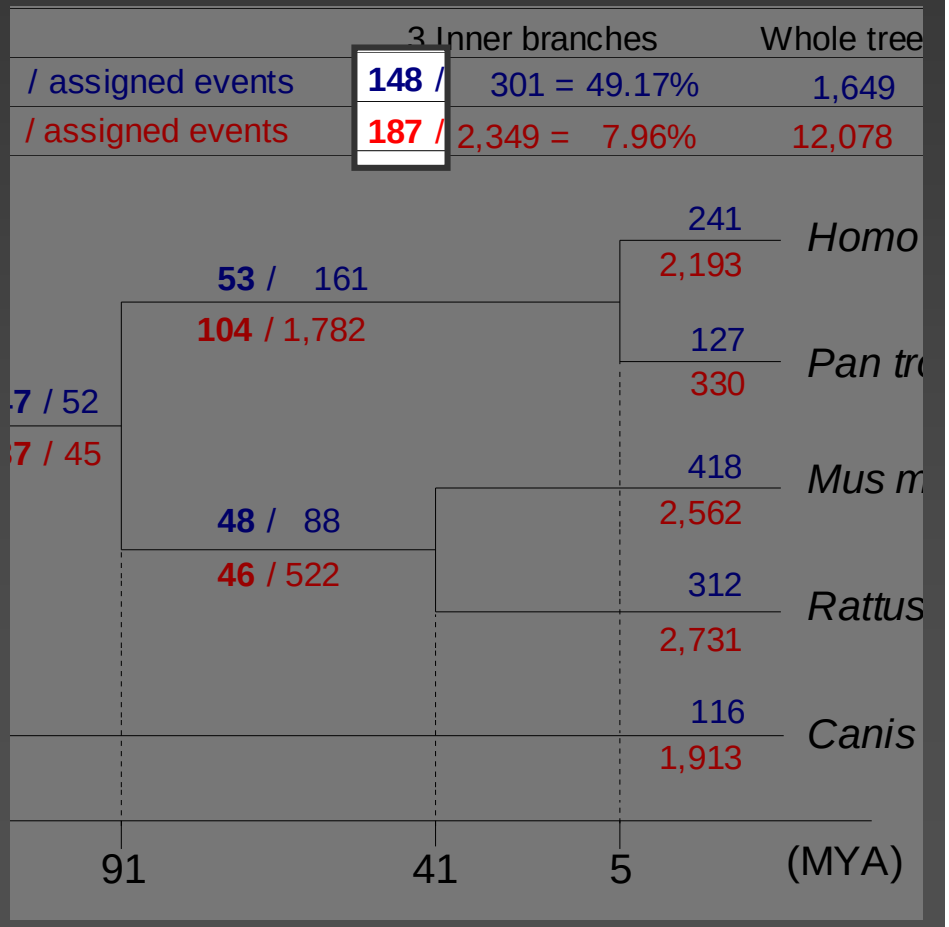


Duplication Event Identification II

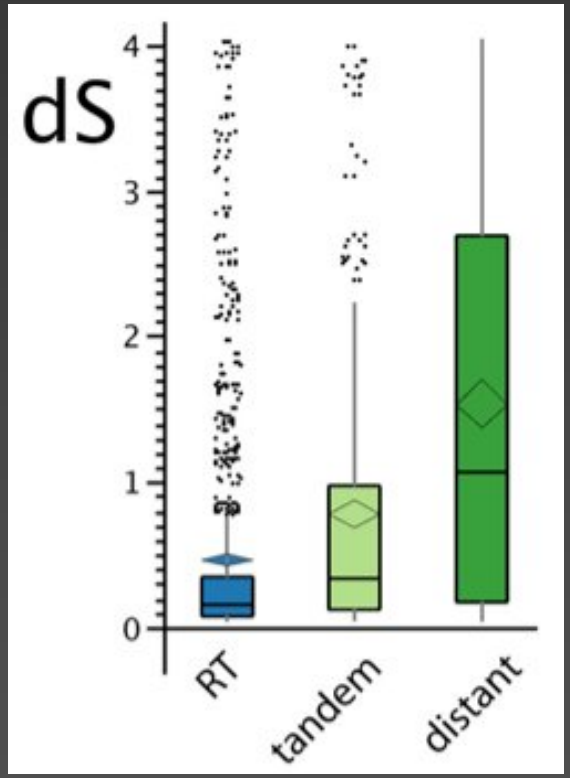


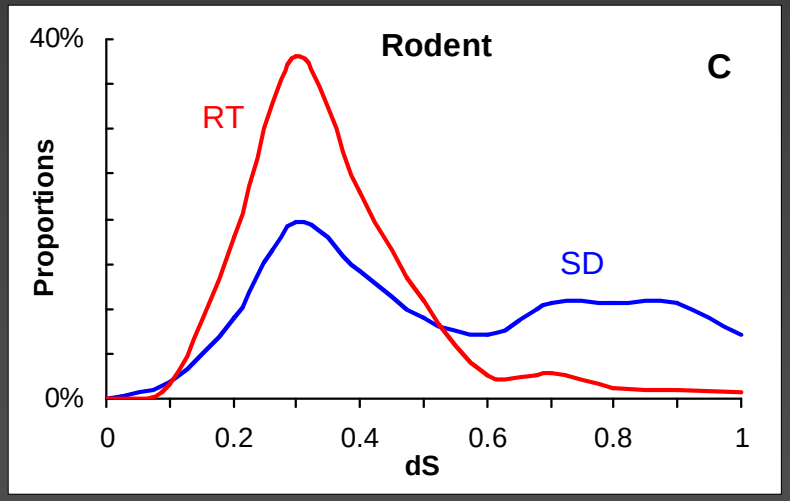
Retrotransposition Produces Nearly Half of All Predicted Gene Duplicates



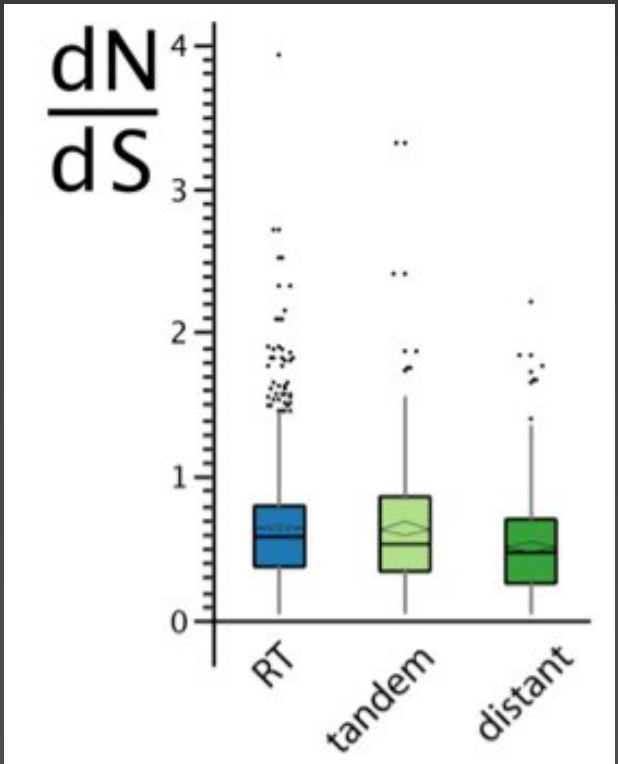


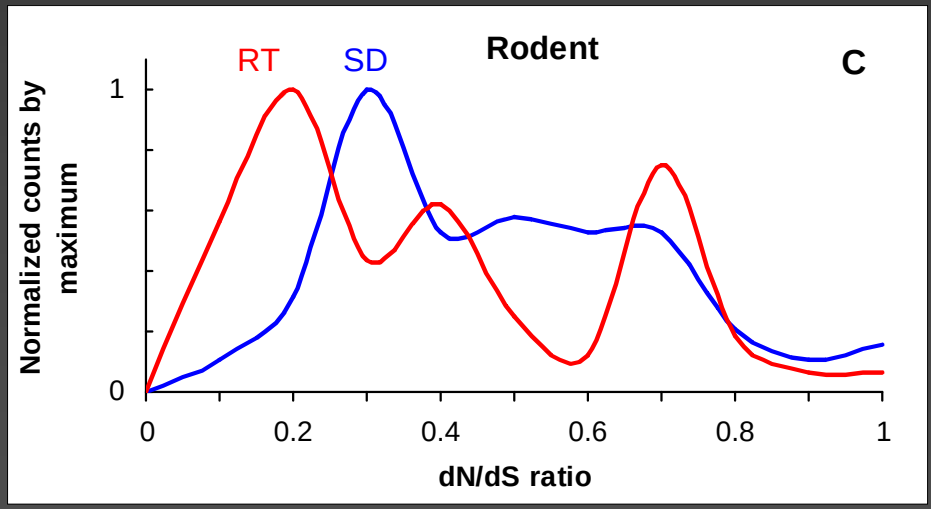
Pairwise Comparisons Reveal Age Differences Between Duplication Types





Little Difference in Evolutionary Constraint Between Duplication Types

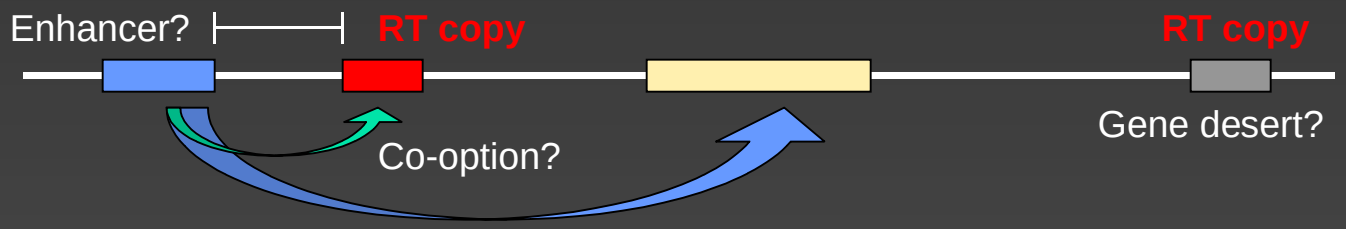


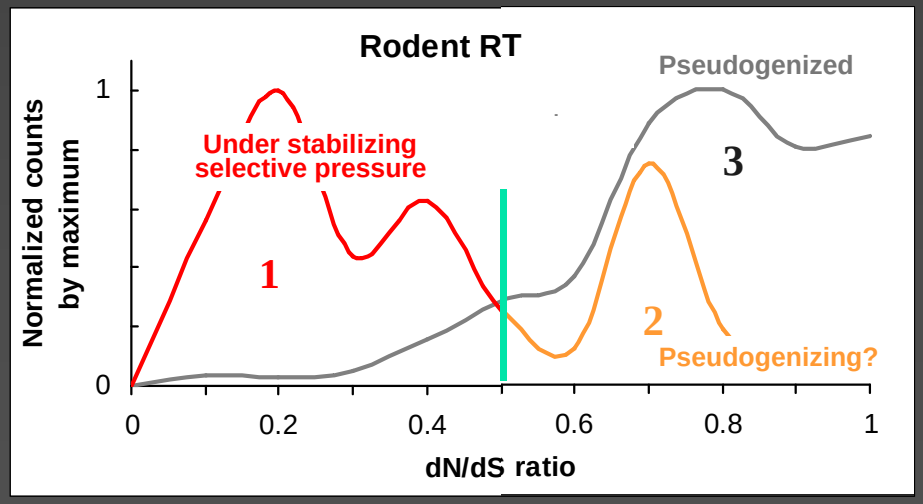


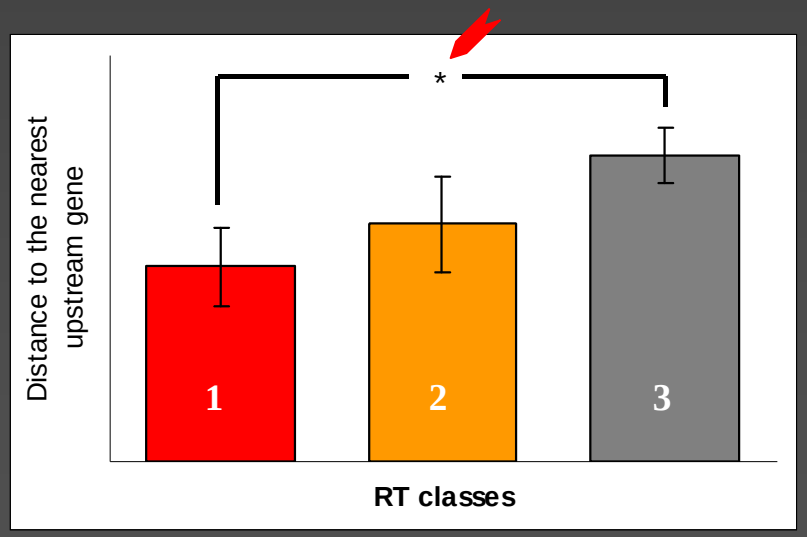
Fate of Duplicates

- Some RT copies are under stabilizing selective pressure
- Any asymmetry of duplicates in selective pressure?
- Can we gain some insight into regulatory element of the duplicates?
  - Do RT copies co-opt the preexisting enhancers?
  - Does any disruption in flanking regions affect the fate of SD duplicates?

The RT Copies Under Stabilizing Selective Pressure are Significantly Closer to the Nearest Upstream Genes



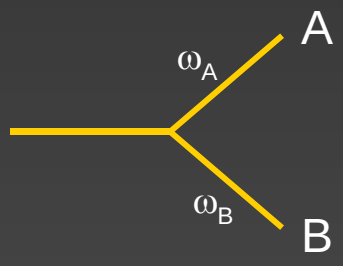




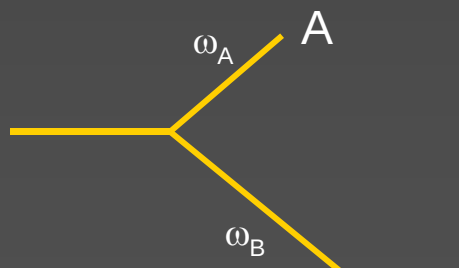
Gain of regulatory elements by co-opting

Disruption in Flanking Region Affects the Fate of SD Duplicates

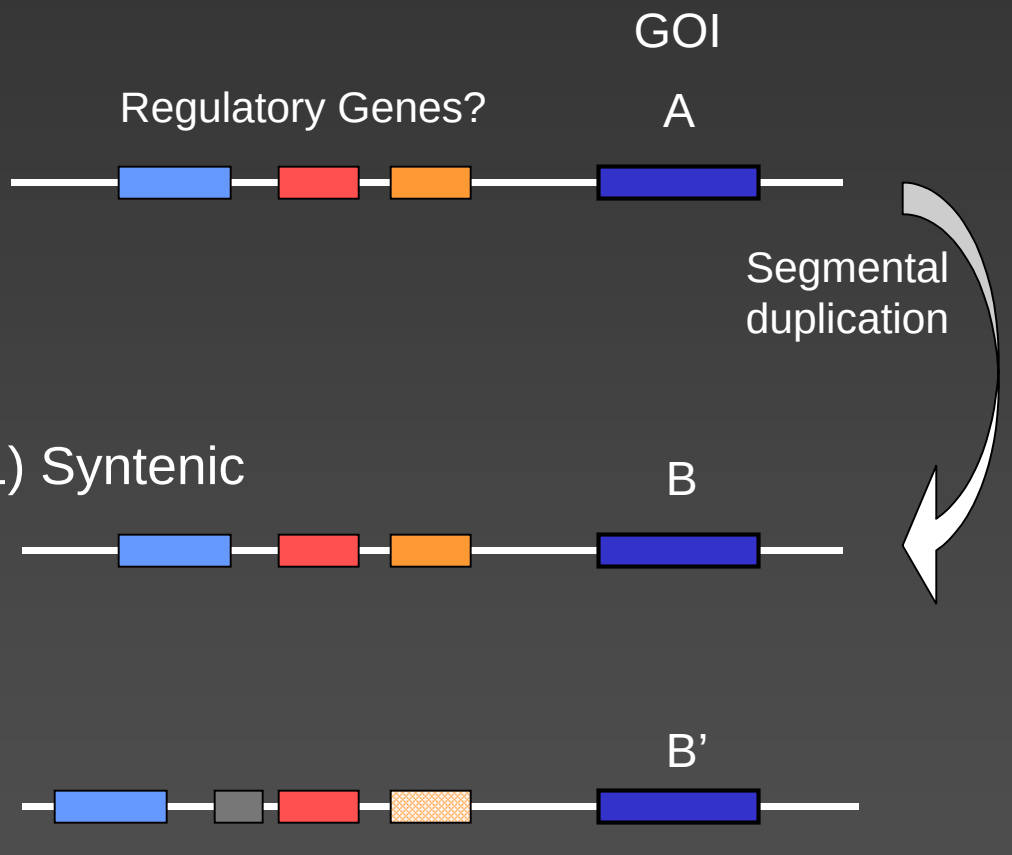
1) Symmetry



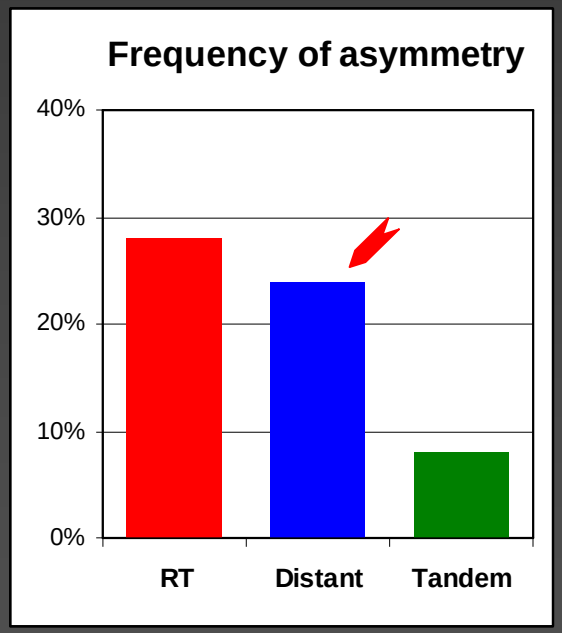
2) Asymmetry

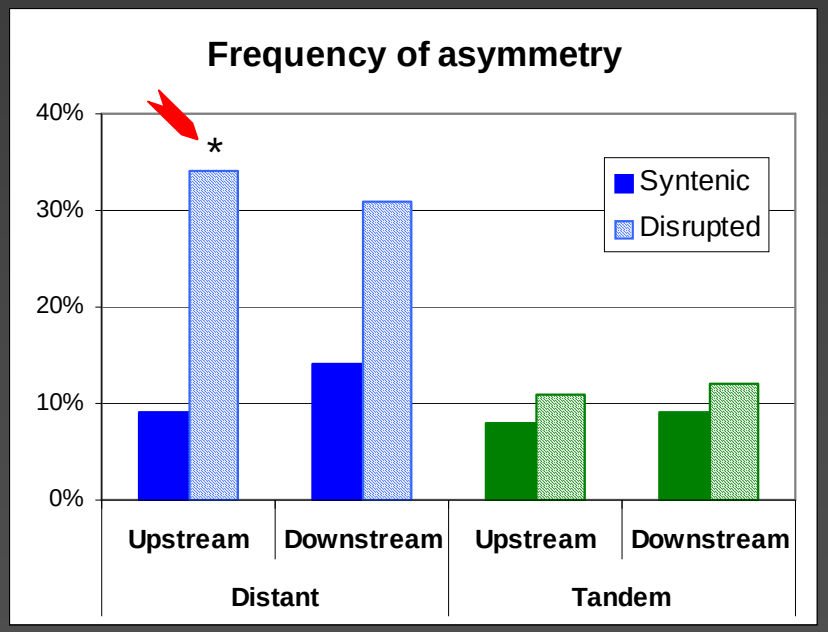


Relaxation of evolutionary constraint?



Disruptions in Flanking Regions of Distant Segmental Duplicates Correlate With an Increase in a Relaxation of Constraint





Loss of regulatory elements by disruption in flanking region

Conclusion

- A method to reconstruct ancestral relations independent of functional relations
- Roughly equal contributions of new genes by SD and RT duplication mechanisms
- Gain/loss of regulatory elements affect the fate of duplicates
  1. Co-opting preexisting enhancers by RT copies
  2. Correlation between the disruptions in flanking regions and an increasing asymmetry on distant SD copies

Thanks

- Dr. Craig Nelson
- Dr. Ion Mandoiu

- Paul Ryvkin
- Edward Hemphill
- Matthew Kozachek

- Gerstein Lab
- Gogarten Lab

- ENSEMBL
- InParanoid

RECOMB-CG ’08