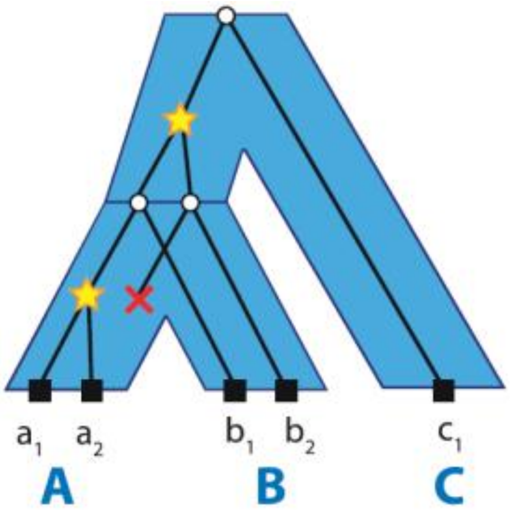
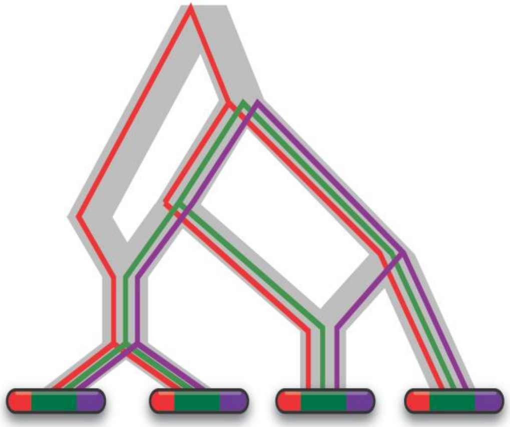
Phylogenomics of closely related species and individuals





Matthew Rasmussen

Siepel lab, Cornell University

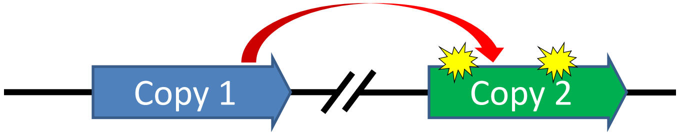
In collaboration with Manolis Kellis, MIT CSAIL

February, 2013

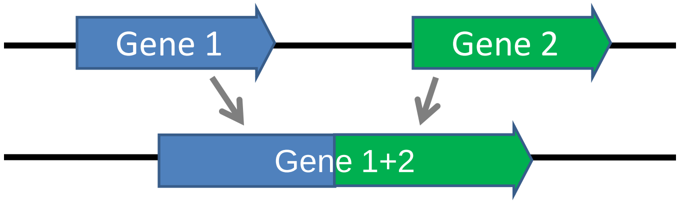
How genomes evolve over short & long time scales

Long time scales
10-100myrs

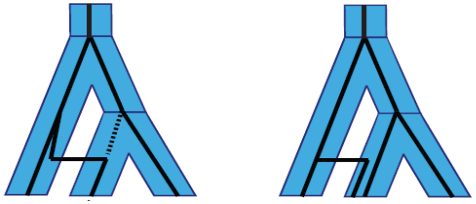
Duplication & Loss



Fusion & Fission

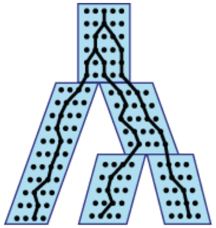


Horizontal gene transfer

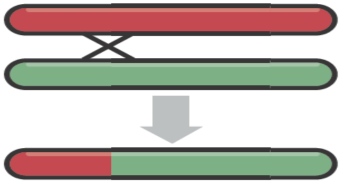


Short time scales
1kyr-1myrs

Drift & Coalescence



Recombination

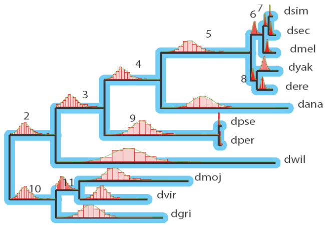


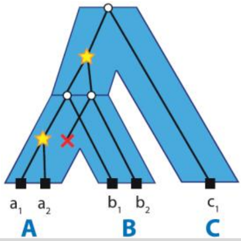
Developing models of evolution

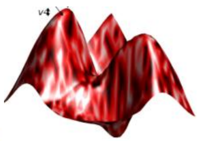
Combining different evolutionary events & processes

Duplication, Loss & Substitution

SPIDIR (GR 2007), SPIMAP (MBE 2010),
TreeFix (SystBio 2012)





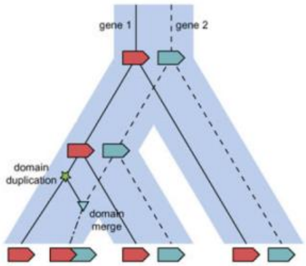


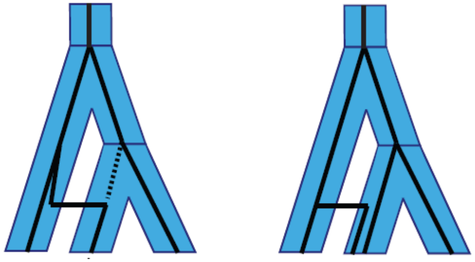
Fusion & Fission

STAR-MP (MBE 2011)

Additive & Replacing transfer

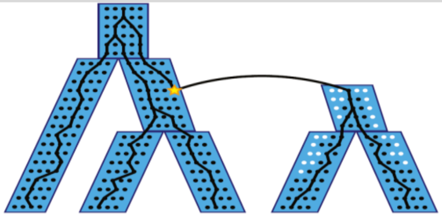
(MBE 2012)





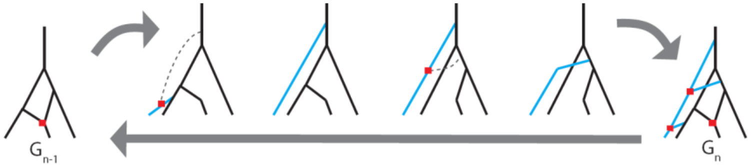
Duplication, Loss & Coalescence

DLCoal (GR 2012)



Coalescence & Recombination

ARGHMM

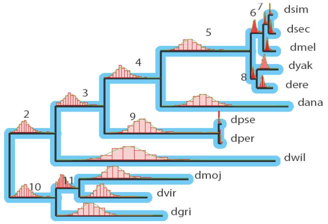


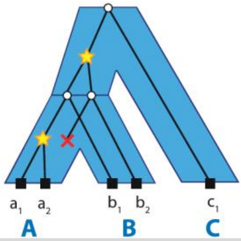
Developing models of evolution

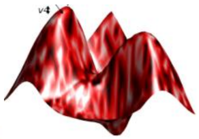
Combining different evolutionary events & processes

Duplication, Loss & Substitution

SPIDIR (GR 2007), SPIMAP (MBE 2010),
TreeFix (SystBio 2012)





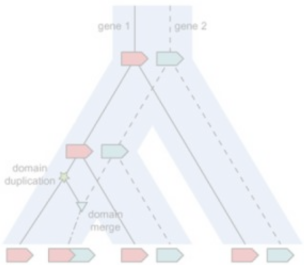


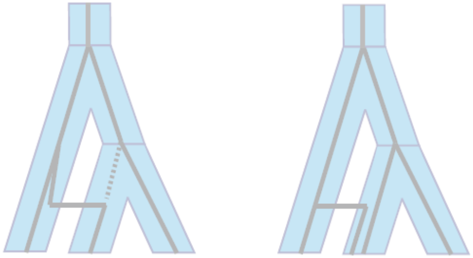
Fusion & Fission

STAR-MP (MBE 2011)

Additive & Replacing transfer

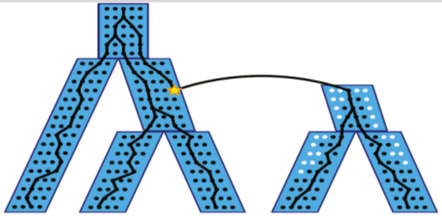
(MBE 2012)





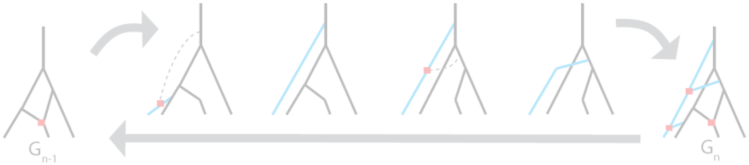
Duplication, Loss & Coalescence

DLCoal (GR 2012)



Coalescence & Recombination

ARGHMM

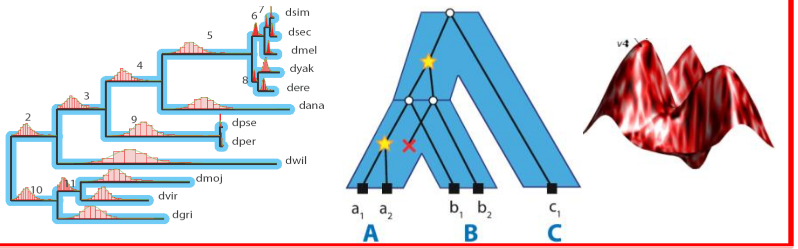


Developing models of evolution

Combining different evolutionary events & processes

Duplication, Loss & Substitution

SPIDIR (GR 2007), SPIMAP (MBE 2010),
TreeFix (SystBio 2012)

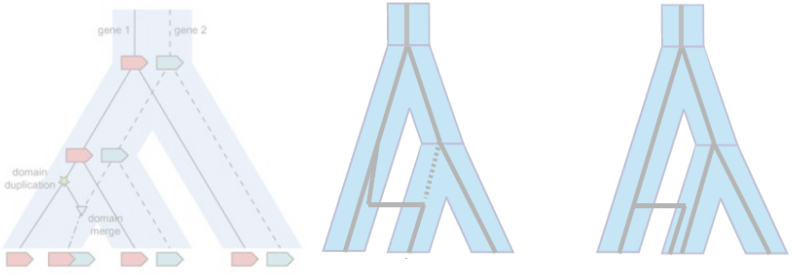


Fusion & Fission

STAR-MP (MBE 2011)

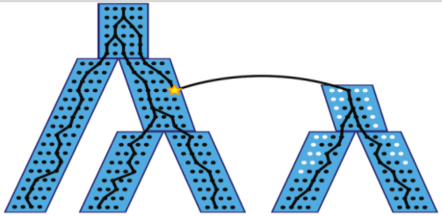
Additive & Replacing transfer

(MBE 2012)



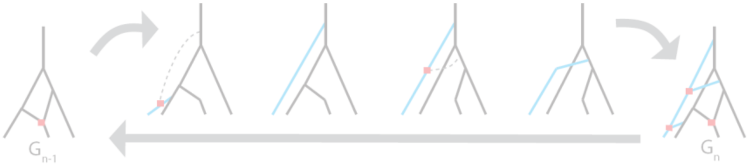
Duplication, Loss & Coalescence

DLCoal (GR 2012)

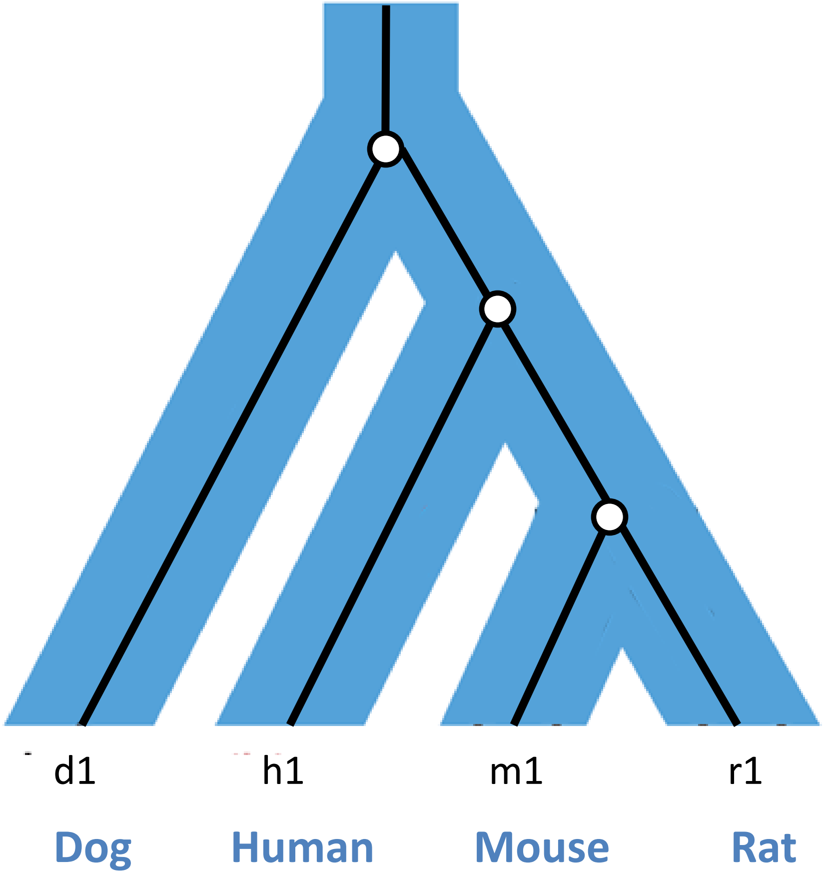


Coalescence & Recombination

ARGHMM



Gene family evolution

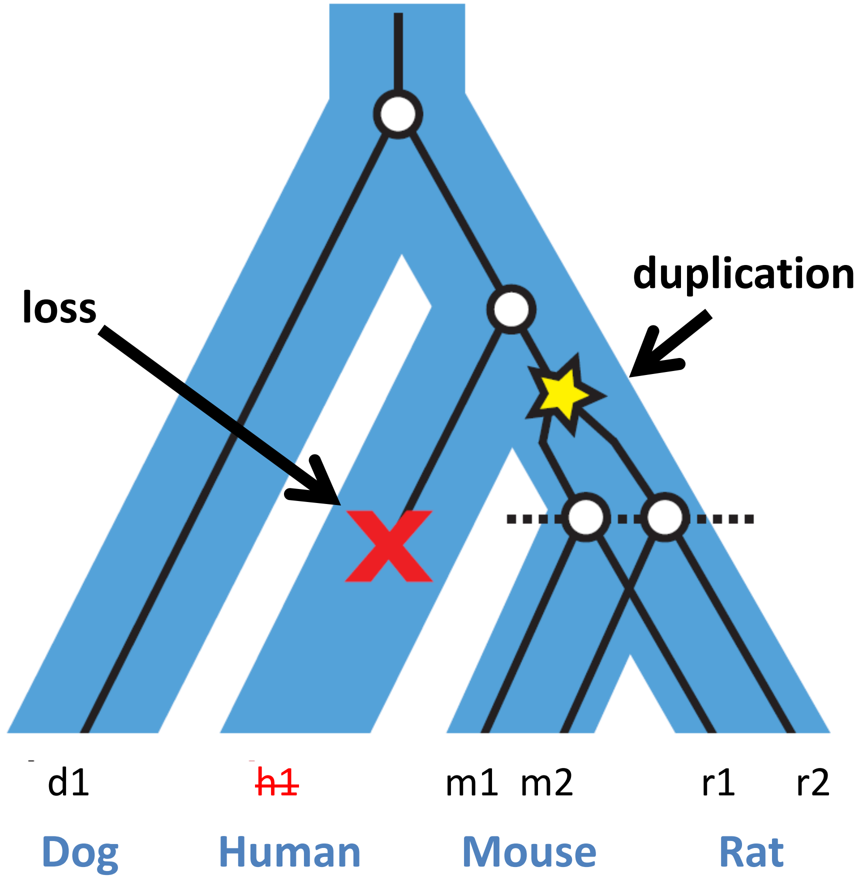


Species tree

Gene tree

d1	ATGCCT**GAA**CCCGTTCTC
h1	ATG**A**CT**GAT**CC**A**GTTCTC
m1	ATGCCTCCCCC**A**GT**AGG**C
r1	ATGCCTCCCCC**A**GT**AGG**C

Gene family evolution



Species tree

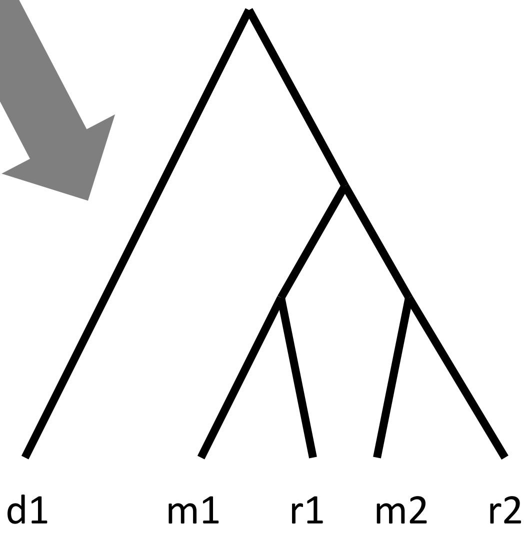
Gene tree

- Two genes are **orthologs** if their MRCA is a *speciation*: ○
- Two genes are **paralogs** if their MRCA is a *duplication*: ★

Inferring events in a gene family

	Gene sequences
d1	ATGCCTGAACCCGTTCTC
m1	ATGACTGATCCAGTTCTC
r1	ATGCCTCCCCCAGTAGGC
m2	ATGCCTCGGGCAGTAGGC
r2	ATGCCTCCCCCAGTAGGC

Phylogenetic method



Gene tree

Obtained previously using a species tree method

Species tree

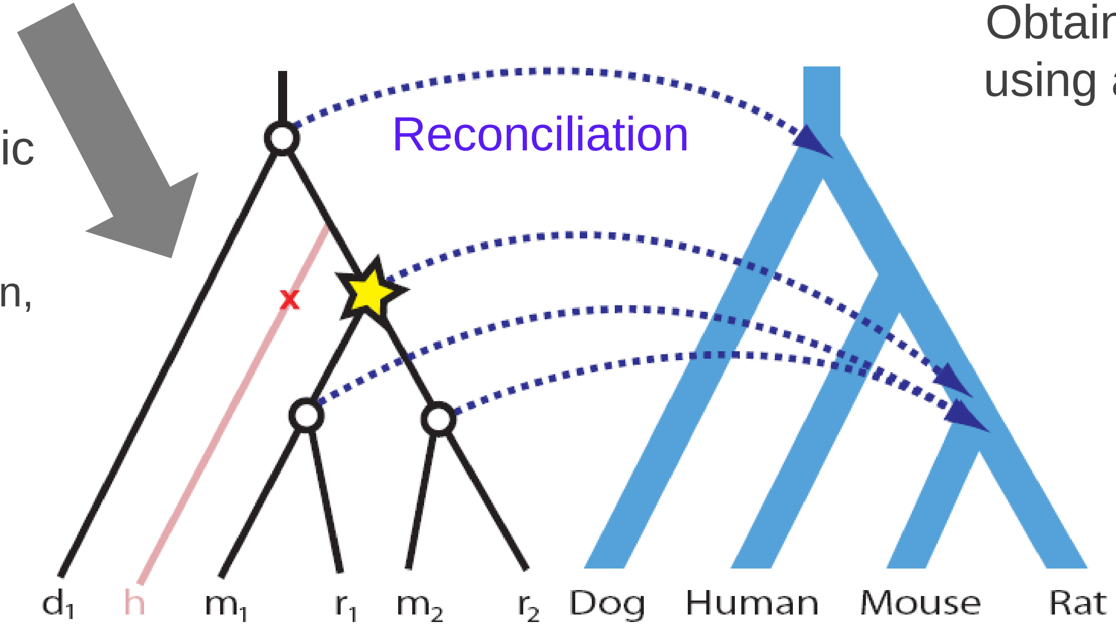
Inferring events in a gene family

Gene sequences

d1	ATGCCTGAACCCGTTCTC
m1	ATGACTGATCCAGTTCTC
r1	ATGCCTCCCCCAGTAGGC
m2	ATGCCTCGGGCAGTAGGC
r2	ATGCCTCCCCCAGTAGGC

Phylogenetic Method

(ML, Bayesian, NJ, etc)

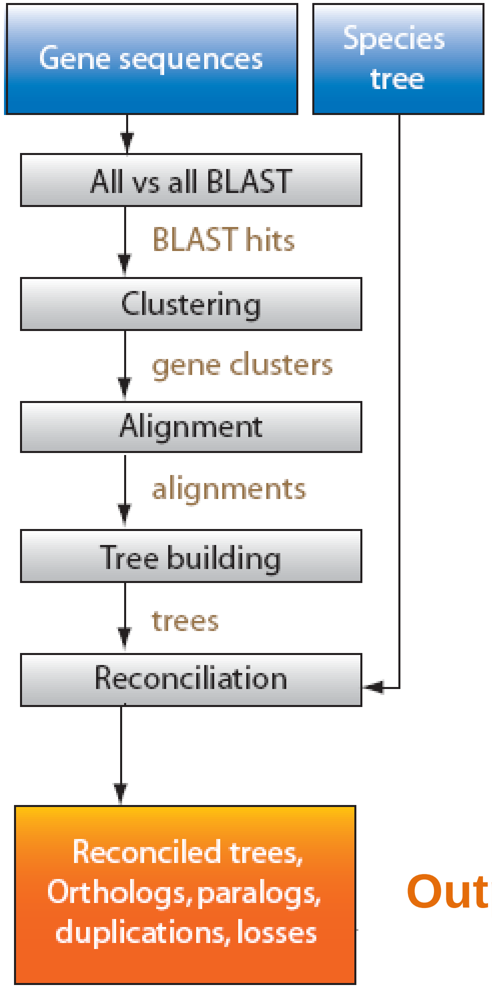


Gene tree

Species tree

Obtained previously using a species tree method

Phylogenomic Pipeline



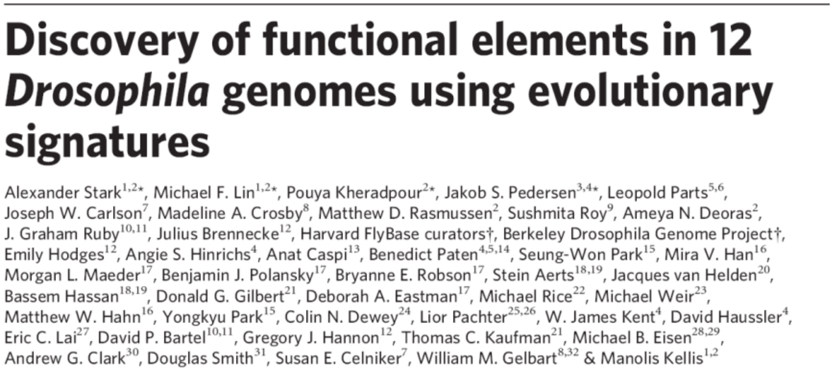
Inputs

Outputs

Evolution of genes and genomes on the *Drosophila* phylogeny

Drosophila 12 Genomes Consortium*

Nature. 2007

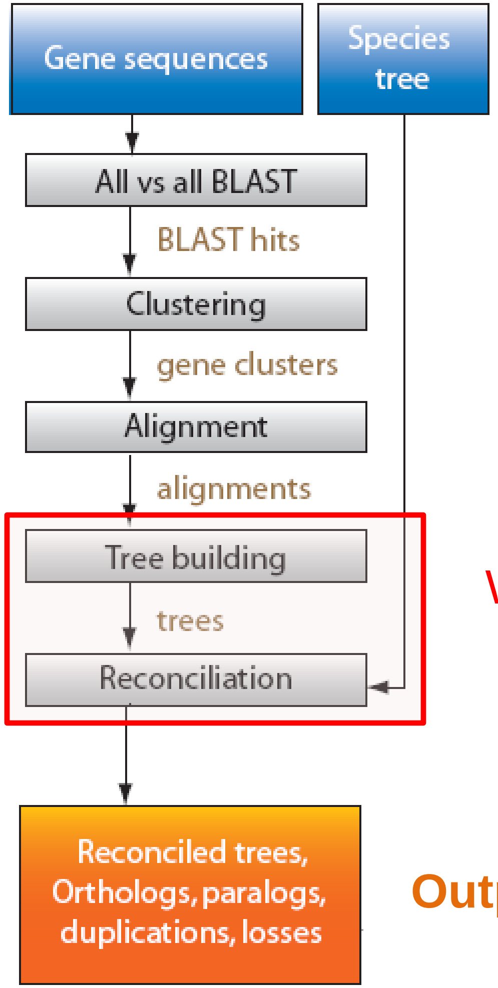


Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures

Alexander Stark[1,2*], Michael F. Lin[1,2*], Pouya Kheradpour[2*], Jakob S. Pedersen[3,4*], Leopold Parts[5,6], Joseph W. Carlson[7], Madeline A. Crosby[8], Matthew D. Rasmussen[2], Sushmita Roy[9], Ameya N. Deoras[2], J. Graham Ruby[10,11], Julius Brennecke[12], Harvard FlyBase curators†, Berkeley Drosophila Genome Project†, Emily Hodges[12], Angie S. Hinrichs[4], Anat Caspi[13], Benedict Paten[4,5,14], Seung-Won Park[15], Mira V. Han[16], Morgan L. Maeder[17], Benjamin J. Polansky[17], Bryanne E. Robson[17], Stein Aerts[18,19], Jacques van Helden[20], Bassem Hassan[18,19], Donald G. Gilbert[21], Deborah A. Eastman[17], Michael Rice[22], Michael Weir[23], Matthew W. Hahn[16], Yongkyu Park[15], Colin N. Dewey[24], Lior Pachter[25,26], W. James Kent[4], David Haussler[4], Eric C. Lai[27], David P. Bartel[10,11], Gregory J. Hannon[12], Thomas C. Kaufman[21], Michael B. Eisen[28,29], Andrew G. Clark[30], Douglas Smith[31], Susan E. Celniker[7], William M. Gelbart[8,32] & Manolis Kellis[1,2]

Nature. 2007

Phylogenomic Pipeline



Inputs

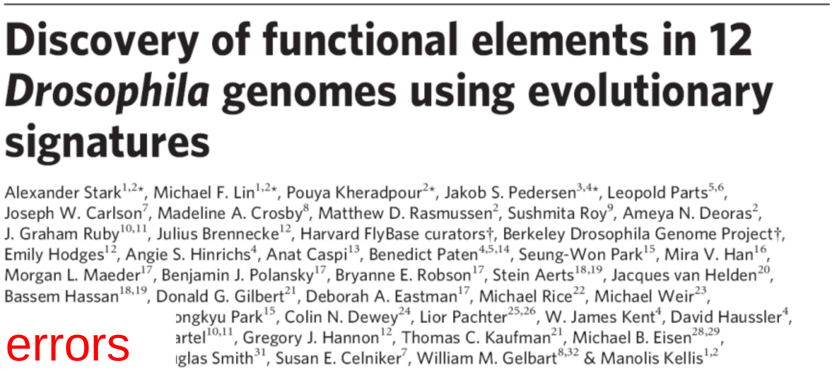
What kind of errors might occur here?

Outputs

Evolution of genes and genomes on the *Drosophila* phylogeny

Drosophila 12 Genomes Consortium*

Nature. 2007



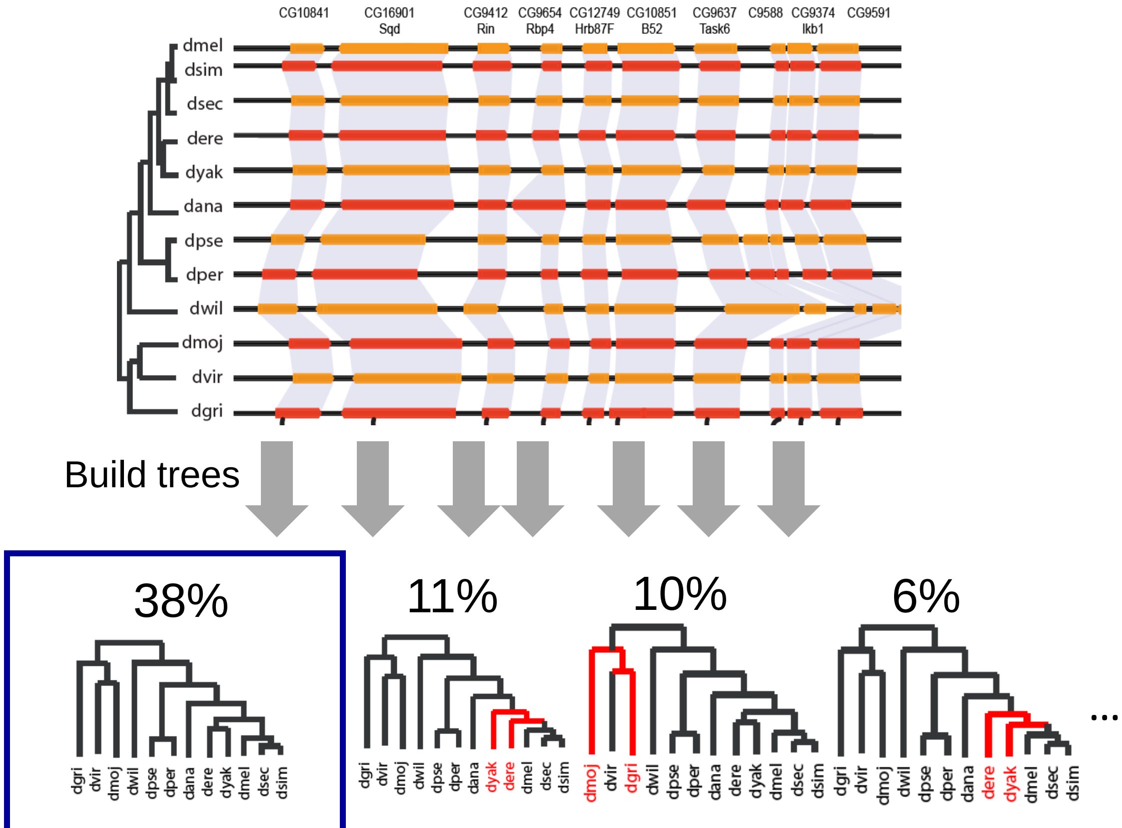
Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures

Alexander Stark[1,2*], Michael F. Lin[1,2*], Pouya Kheradpour[2*], Jakob S. Pedersen[3,4*], Leopold Parts[5,6],
Joseph W. Carlson[7], Madeline A. Crosby[8], Matthew D. Rasmussen[2], Sushmita Roy[9], Ameya N. Deoras[2],
J. Graham Ruby[10,11], Julius Brennecke[12], Harvard FlyBase curators†, Berkeley Drosophila Genome Project†,
Emily Hodges[12], Angie S. Hinrichs[4], Anat Caspi[13], Benedict Paten[4,5,14], Seung-Won Park[15], Mira V. Han[16],
Morgan L. Maeder[17], Benjamin J. Polansky[17], Bryanne E. Robson[17], Stein Aerts[18,19], Jacques van Helden[20],
Bassem Hassan[18,19], Donald G. Gilbert[21], Deborah A. Eastman[17], Michael Rice[22], Michael Weir[23],
ongkyu Park[15], Colin N. Dewey[24], Lior Pachter[25,26], W. James Kent[4], David Haussler[4],
artel[10,11], Gregory J. Hannon[12], Thomas C. Kaufman[21], Michael B. Eisen[28,29],
glas Smith[31], Susan E. Celniker[7], William M. Gelbart[8,32] & Manolis Kellis[1,2]

Nature. 2007

Existing phylogenetic methods are not accurate enough for phylogenomics

~5000 syntenic one-to-one orthologs

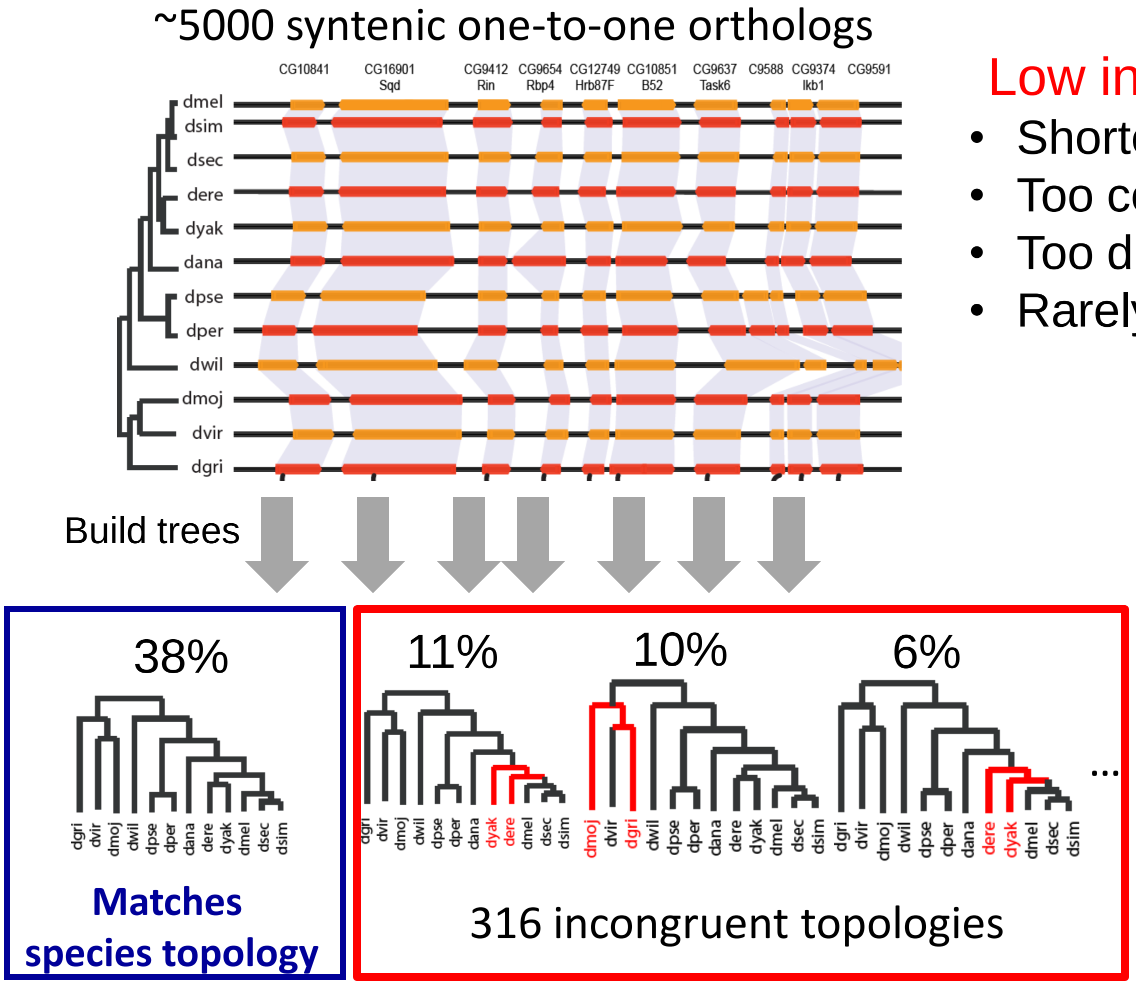


Matches species topology

316 incongruent topologies

Rasmussen, Kellis.
Genome Research 2007

Existing phylogenetic methods are not accurate enough for phylogenomics



~5000 syntenic one-to-one orthologs

Build trees

38%

Matches species topology

11%

10%

6%

…

316 incongruent topologies

Low information ➔ Incongruence

- Shorter sequences (<900bp)
- Too conserved (>90% identity)
- Too diverged (<30% identity)
- Rarely supported by LRT (<5.7%)

Rasmussen, Kellis.
Genome Research 2007

Phylogenomics needs a new approach

Average gene alignment contains too little information for phylogenomic analysis

- Existing algorithms ignore species (species unaware, uses only sequence)

Our approach:

- Use species tree to inform the gene tree reconstruction (species aware)

New Methods:

Drosophila Genomes/Methods

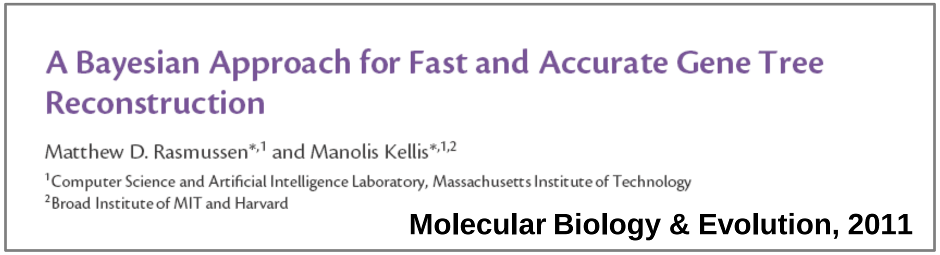
Accurate gene-tree reconstruction by learning gene- and species-specific substitution rates across multiple complete genomes

Matthew D. Rasmussen[1] and Manolis Kellis[1,2,3]

Genome Research, 2007

SPIDIR (Speicies Informed Distance-based Reconstruction)

- Maximum Likelihood



A Bayesian Approach for Fast and Accurate Gene Tree Reconstruction

Matthew D. Rasmussen*[1] and Manolis Kellis*[1,2]

[1] Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology

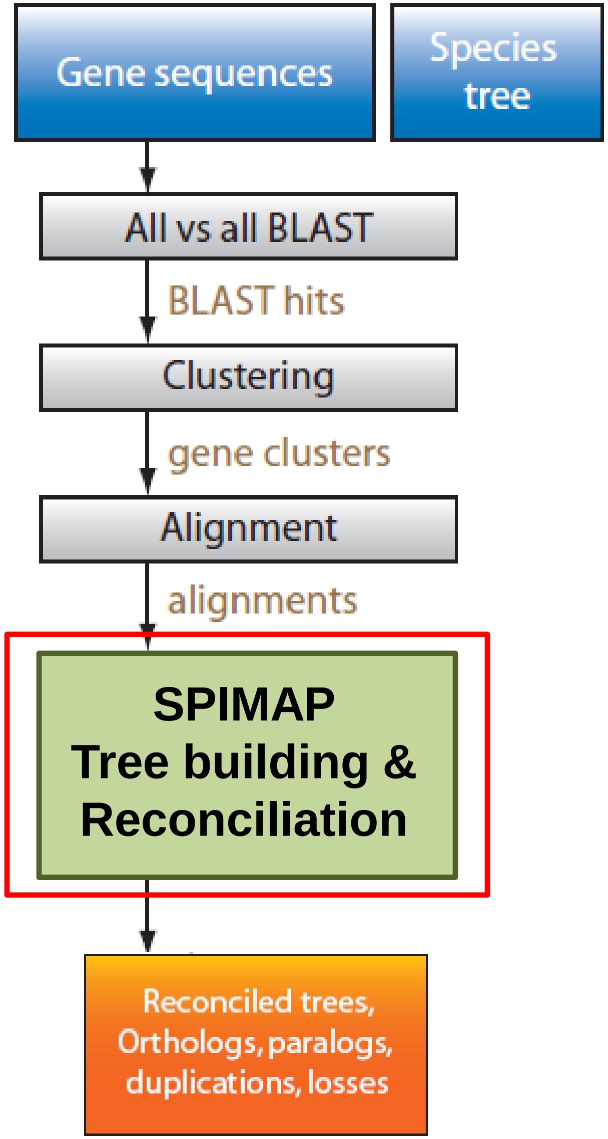
[2] Broad Institute of MIT and Harvard

Molecular Biology & Evolution, 2011

SPIMAP (Species Informed Maximum A Posteriori)

- Bayesian

SPIMAP's Phylogenomic Pipeline



Gene sequences

Species tree

All vs all BLAST

BLAST hits

Clustering

gene clusters

Alignment

alignments

SPIMAP
Tree building & Reconciliation

Reconciled trees, Orthologs, paralogs, duplications, losses

SPIMAP (Species Informed Maximum A Posteriori)

Generative model

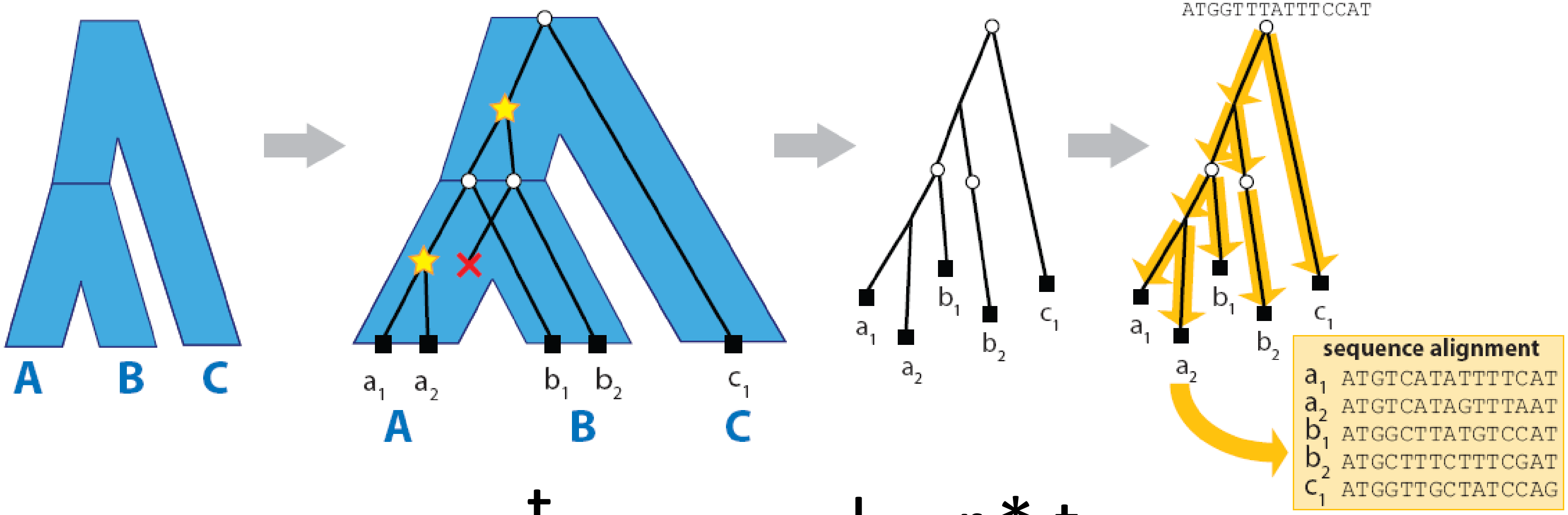
Dup/loss model **Rates model** **Sequence model** (HKY)

Given species tree

Generate times

Generate rates

Generate sequence



sequence alignment

a_1 ATGTCATATTTTCAT
a_2 ATGTCATAGTTTAAT
b_1 ATGGCTTATGTCCAT
b_2 ATGCTTTCTTTCGAT
c_1 ATGGTTGCTATCCAG

$$t_i$$

Time
(myr)

$$l_i = r_i * t_i$$

Divergence
(sub/site) = (sub/site/myr) * (mtr)

Reconstruction using SPIMAP model

We find the maximum *a posteriori* tree

$$\hat{\mathbf{l}}, \hat{T}, \hat{R} = \operatorname*{argmax}_{\mathbf{l},T,R} P(\mathbf{l},T,R|\mathbf{D},\theta) = \operatorname*{argmax}_{\mathbf{l},T,R} \overbrace{P(\mathbf{D}|\mathbf{l},T)}^{\text{likelihood}} \overbrace{P(\mathbf{l}|T,R,\theta)P(T,R|\theta)}^{\text{prior}}$$

Max posterior

Model parameters

Likelihood
Sequence model

Branch prior
Rates model

Topology prior
Dup/loss model

l = vector of branch lengths
T = gene tree topology
R = reconciliation mapping
D = alignment data
θ = (species tree S, and other model parameters)

Reconstruction using SPIMAP model

We find the maximum *a posteriori* tree

$$\hat{\mathbf{l}}, \hat{T}, \hat{R} = \operatorname*{argmax}_{\mathbf{l},T,R} P(\mathbf{l}, T, R|\mathbf{D}, \theta) = \operatorname*{argmax}_{\mathbf{l},T,R} \overbrace{P(\mathbf{D}|\mathbf{l}, T)}^{\text{likelihood}} \overbrace{P(\mathbf{l}|T, R, \theta)P(T, R|\theta)}^{\text{prior}}$$

Max posterior

Model parameters

Novel EM algorithm

Likelihood

Sequence model

Felsenstein's Pruning algorithm ✓

Branch prior

Rates model

Novel gene- & species-specific rates model

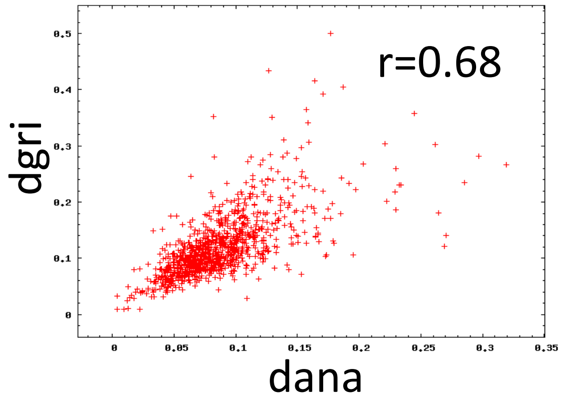
Topology prior

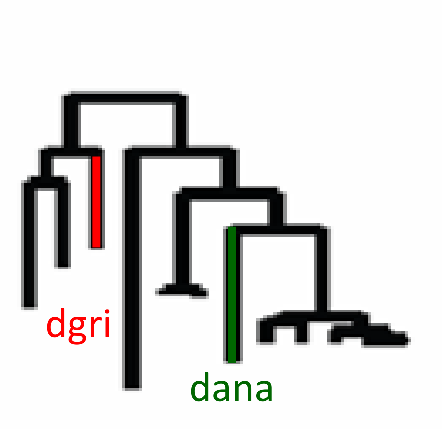
Dup/loss model

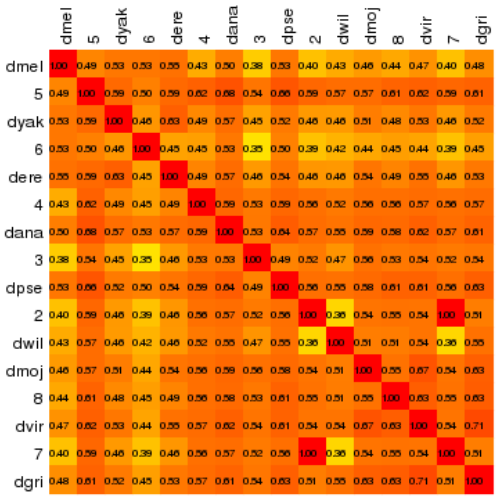
Dynamic Programming Birth-Death [Arvestad 2003] ✓

Rates model: branch length correlations

Absolute lengths

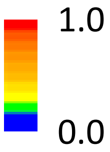






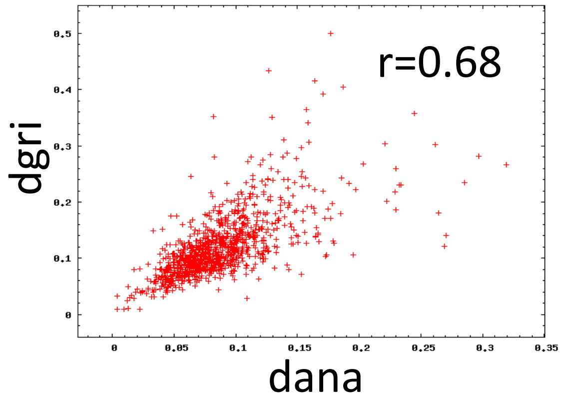
	dmel	5	dyak	6	dere	4	dana	3	dpse	2	dwil	dmoj	8	dvir	7	dgri
dmel	1.00	0.49	0.53	0.53	0.55	0.43	0.50	0.38	0.53	0.40	0.43	0.46	0.44	0.47	0.40	0.48
5	0.49	1.00	0.59	0.50	0.59	0.62	0.68	0.54	0.66	0.59	0.57	0.57	0.61	0.62	0.59	0.61
dyak	0.53	0.59	1.00	0.46	0.63	0.49	0.57	0.45	0.52	0.46	0.46	0.51	0.48	0.53	0.46	0.52
6	0.53	0.50	0.46	1.00	0.45	0.45	0.53	0.35	0.50	0.39	0.42	0.44	0.45	0.44	0.39	0.45
dere	0.55	0.59	0.63	0.45	1.00	0.49	0.57	0.46	0.54	0.46	0.46	0.54	0.49	0.55	0.46	0.53
4	0.43	0.62	0.49	0.45	0.49	1.00	0.59	0.53	0.59	0.56	0.52	0.56	0.56	0.57	0.56	0.57
dana	0.50	0.68	0.57	0.53	0.57	0.59	1.00	0.53	0.64	0.57	0.55	0.59	0.58	0.62	0.57	0.61
3	0.38	0.54	0.45	0.35	0.46	0.53	0.53	1.00	0.49	0.52	0.47	0.56	0.53	0.54	0.52	0.54
dpse	0.53	0.66	0.52	0.50	0.54	0.59	0.64	0.49	1.00	0.56	0.55	0.58	0.61	0.61	0.56	0.63
2	0.40	0.59	0.46	0.39	0.46	0.56	0.57	0.52	0.56	1.00	0.36	0.54	0.55	0.54	1.00	0.51
dwil	0.43	0.57	0.46	0.42	0.46	0.52	0.55	0.47	0.55	0.36	1.00	0.51	0.51	0.54	0.36	0.55
dmoj	0.46	0.57	0.51	0.44	0.54	0.56	0.59	0.56	0.58	0.54	0.51	1.00	0.55	0.67	0.54	0.63
8	0.44	0.61	0.48	0.45	0.49	0.56	0.58	0.53	0.61	0.55	0.51	0.55	1.00	0.63	0.55	0.63
dvir	0.47	0.62	0.53	0.44	0.55	0.57	0.62	0.54	0.61	0.54	0.54	0.67	0.63	1.00	0.54	0.71
7	0.40	0.59	0.46	0.39	0.46	0.56	0.57	0.52	0.56	1.00	0.36	0.54	0.55	0.54	1.00	0.51
dgri	0.48	0.61	0.52	0.45	0.53	0.57	0.61	0.54	0.63	0.51	0.55	0.63	0.63	0.71	0.51	1.00

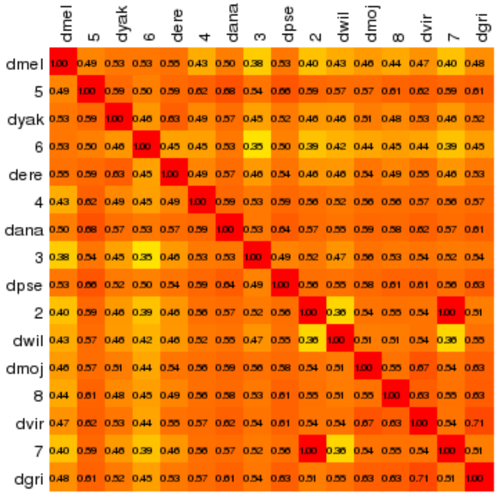
average
r = 0.56



Rates model: branch length correlations

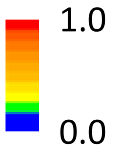
Absolute lengths

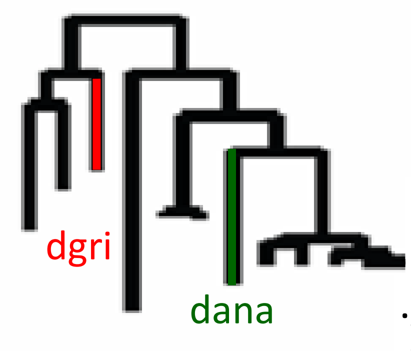




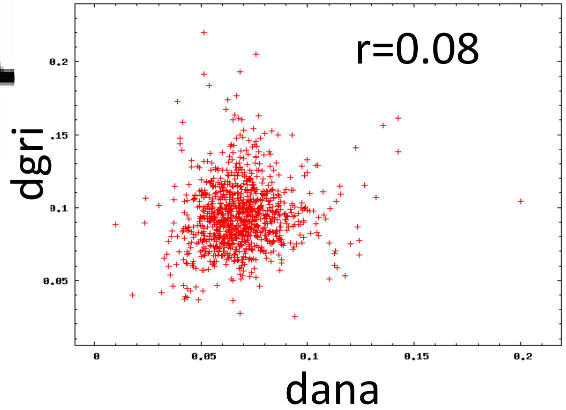
	dmel	5	dyak	6	dere	4	dana	3	dpse	2	dwil	dmoj	8	dvir	7	dgri
dmel	1.00	0.49	0.53	0.53	0.55	0.43	0.50	0.38	0.53	0.40	0.43	0.46	0.44	0.47	0.40	0.48
5	0.49	1.00	0.59	0.50	0.59	0.62	0.68	0.54	0.66	0.59	0.57	0.57	0.61	0.62	0.59	0.61
dyak	0.53	0.59	1.00	0.46	0.63	0.49	0.57	0.45	0.52	0.46	0.46	0.51	0.48	0.53	0.46	0.52
6	0.53	0.50	0.46	1.00	0.45	0.45	0.53	0.35	0.50	0.39	0.42	0.44	0.45	0.44	0.39	0.45
dere	0.55	0.59	0.63	0.45	1.00	0.49	0.57	0.46	0.54	0.46	0.46	0.54	0.49	0.55	0.46	0.53
4	0.43	0.62	0.49	0.45	0.49	1.00	0.59	0.53	0.59	0.56	0.52	0.56	0.56	0.57	0.56	0.57
dana	0.50	0.68	0.57	0.53	0.57	0.59	1.00	0.53	0.64	0.57	0.55	0.59	0.58	0.62	0.57	0.61
3	0.38	0.54	0.45	0.35	0.46	0.53	0.53	1.00	0.49	0.52	0.47	0.56	0.53	0.54	0.52	0.54
dpse	0.53	0.66	0.52	0.50	0.54	0.59	0.64	0.49	1.00	0.56	0.55	0.58	0.61	0.61	0.56	0.63
2	0.40	0.59	0.46	0.39	0.46	0.56	0.57	0.52	0.56	1.00	0.36	0.54	0.55	0.54	1.00	0.51
dwil	0.43	0.57	0.46	0.42	0.46	0.52	0.55	0.47	0.55	0.36	1.00	0.51	0.51	0.54	0.36	0.55
dmoj	0.46	0.57	0.51	0.44	0.54	0.56	0.59	0.56	0.58	0.54	0.51	1.00	0.55	0.67	0.54	0.63
8	0.44	0.61	0.48	0.45	0.49	0.56	0.58	0.53	0.61	0.55	0.51	0.55	1.00	0.63	0.55	0.63
dvir	0.47	0.62	0.53	0.44	0.55	0.57	0.62	0.54	0.61	0.54	0.54	0.67	0.63	1.00	0.54	0.71
7	0.40	0.59	0.46	0.39	0.46	0.56	0.57	0.52	0.56	1.00	0.36	0.54	0.55	0.54	1.00	0.51
dgri	0.48	0.61	0.52	0.45	0.53	0.57	0.61	0.54	0.63	0.51	0.55	0.63	0.63	0.71	0.51	1.00

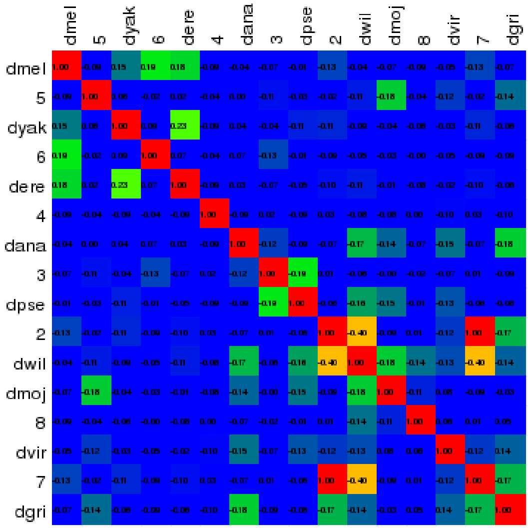
average
r = 0.56





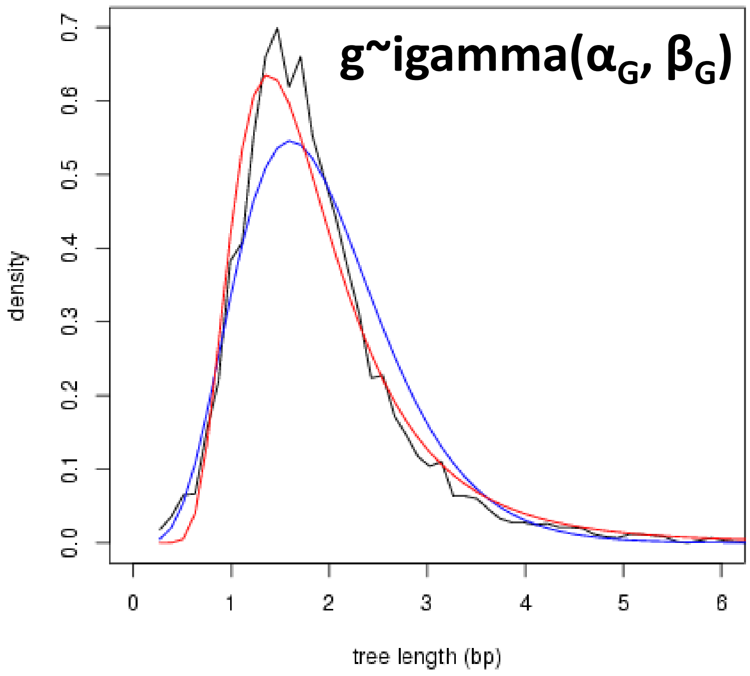
Relative lengths



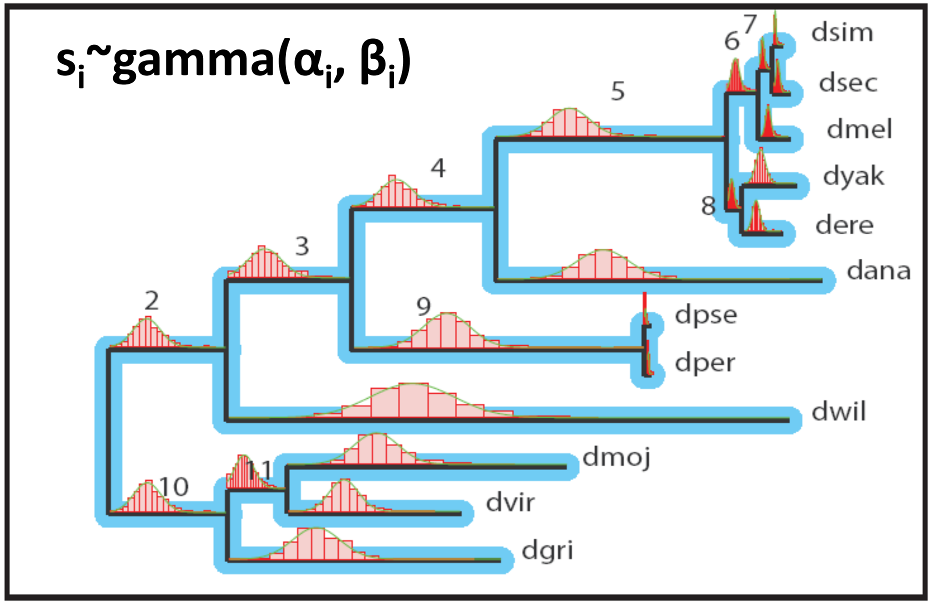


average
r = 0.09

Rates model: rate distributions



Gene rate distribution
(inverse-gamma)



Species rate distributions
(gamma)

$$l_i = r_i * t_i$$

$$r_i = g * s_i$$

l_i (sub/site)

r_i (sub/site/myr)

t_i (myr)

Method Evaluation

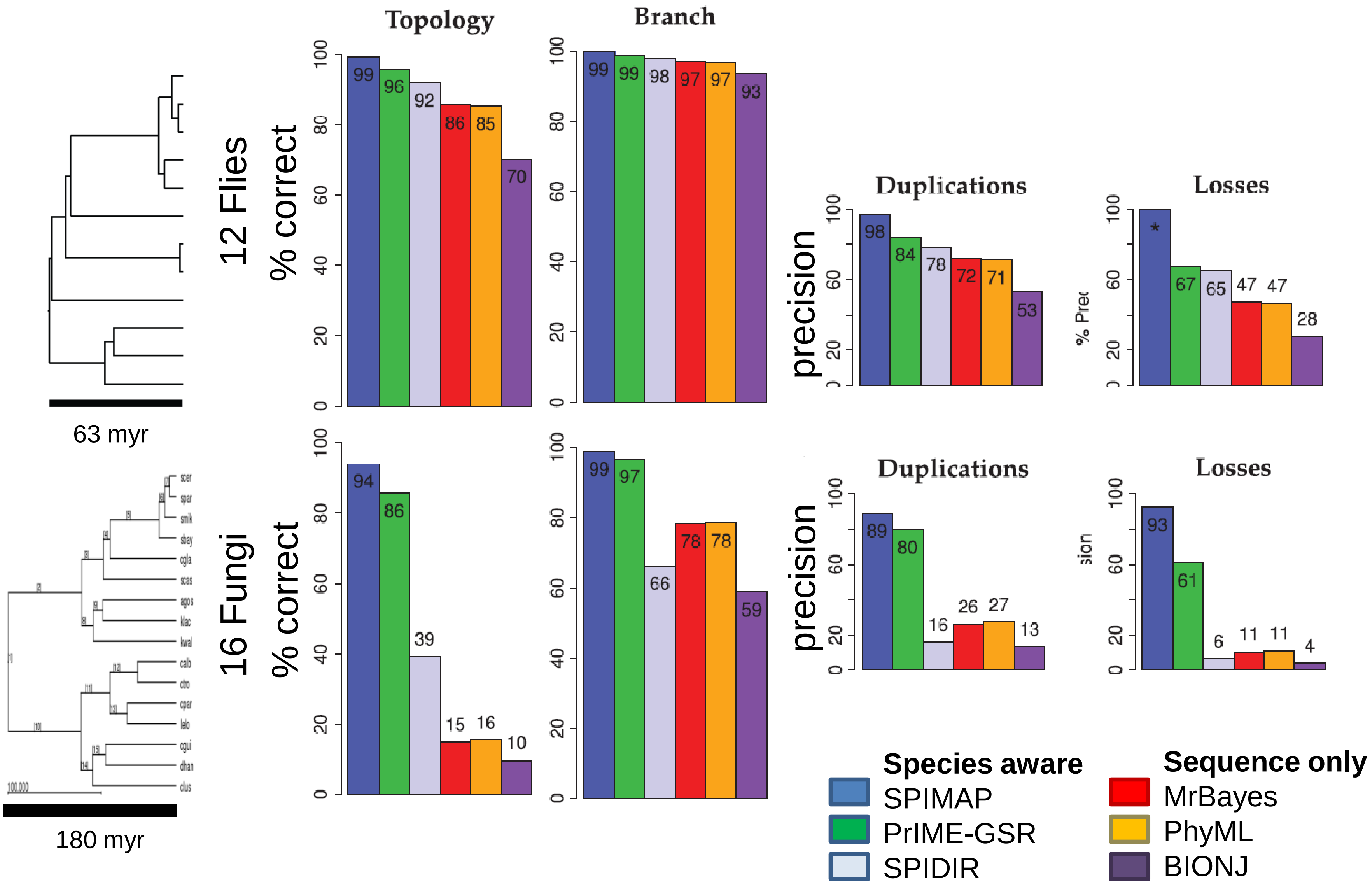
“Sequence-only” methods

- PHYML (ML) Guidon 2003
- RAxML (ML) Stamatakis 2006
- MrBayes (Bayesian) Ronquist 2003
- BIONJ (NJ) Gascuel 1997

Species-aware methods:

- SPIMAP (Bayesian) Rasmussen, Kellis 2010
- SPIDIR (dist. ML) Rasmussen, Kellis 2007
- SYNERGY (dist/parsimony) Wapinski 2007
- PrIME-GSR (Bayesian, i.i.d.) Akerborg 2009

Evaluation on fly & fungi simulated data



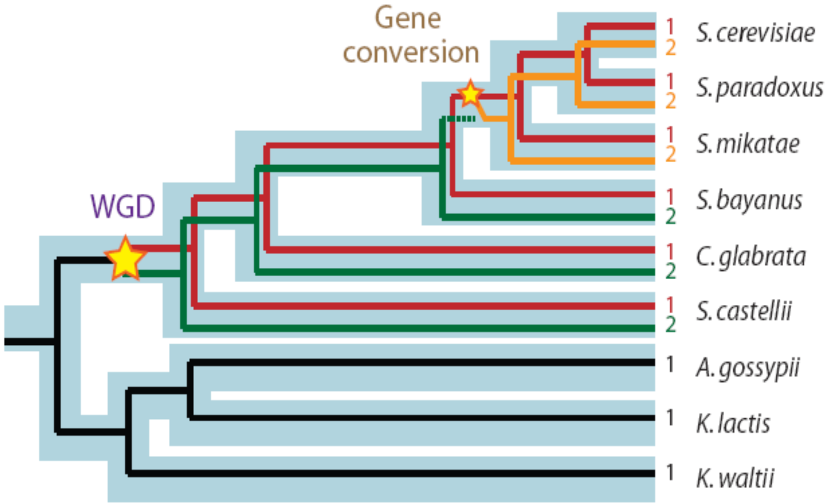
Evaluation on 16 fungal species

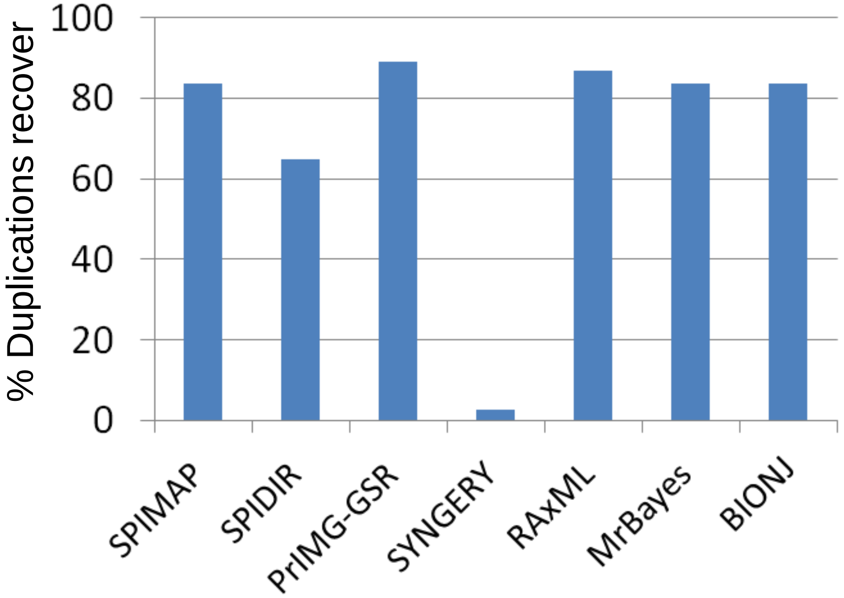
	Program	Orthologs	# Dup	# Loss	Run time
Species aware	SPIMAP	96.2%	5,541	10,884	1.0 min
	SPIDIR	83.3%	10,177	33,550	2.2 min
	PrIMG-GSR	90.7%	7,951	21,099	53.1 min
	SYNGERY	99.2%	4,604	8,179	---
Sequence only	RAxML	63.3%	21,485	65,392	18.4 s
	MrBayes	64.2%	21,307	65,238	43.2 s
	BIONJ	60.4%	22,396	71,231	0.5 s

(1) Species-aware methods recover far more syntenic orthologs

(2) Sequence-only methods infer far more (likely spurious) duplications-losses

(3) High recovery of duplications due to gene conversion





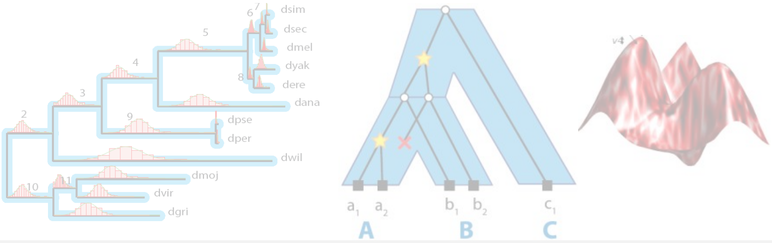
Methods that **balance** sequence information vs dup-loss and synteny information perform best

Developing models of evolution

Combining different evolutionary events & processes

Duplication, Loss & Substitution

SPIDIR (GR 2007), SPIMAP (MBE 2010),
TreeFix (SystBio 2012)

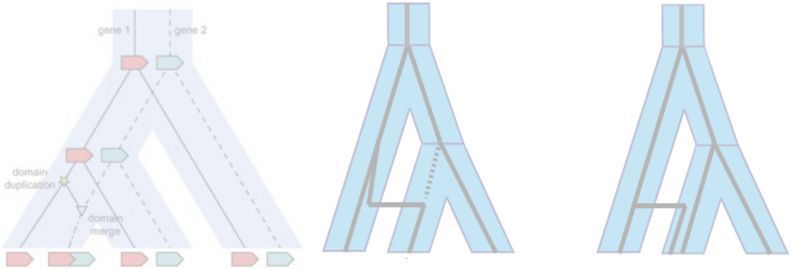


Fusion & Fission

STAR-MP (MBE 2011)

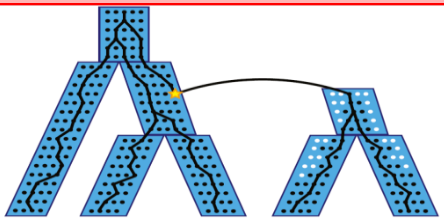
Additive & Replacing transfer

(MBE 2012)



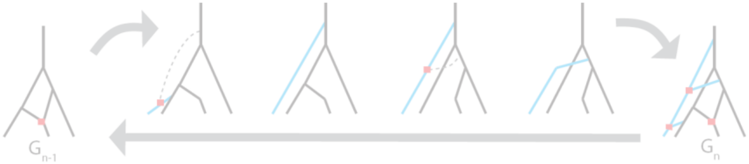
Duplication, Loss & Coalescence

DLCoal (GR 2012)



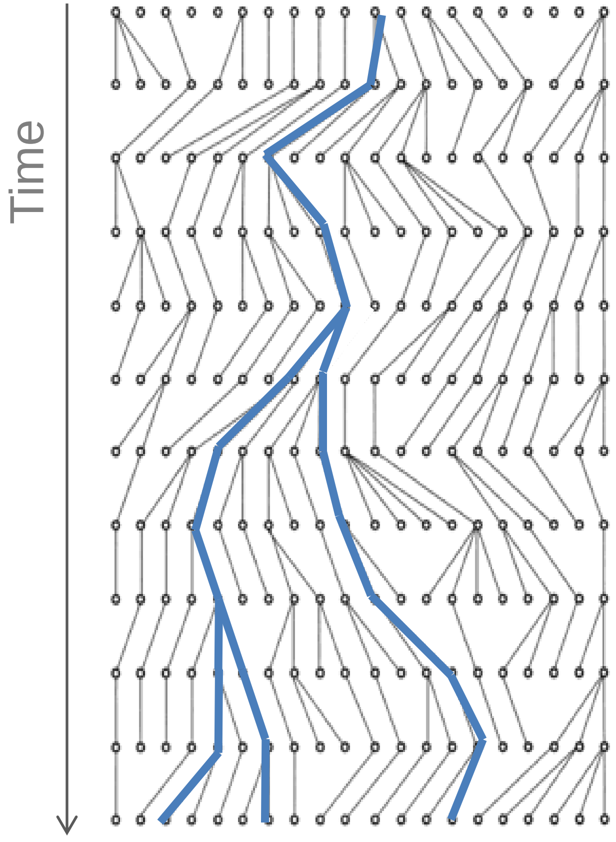
Coalescence & Recombination

ARGHMM



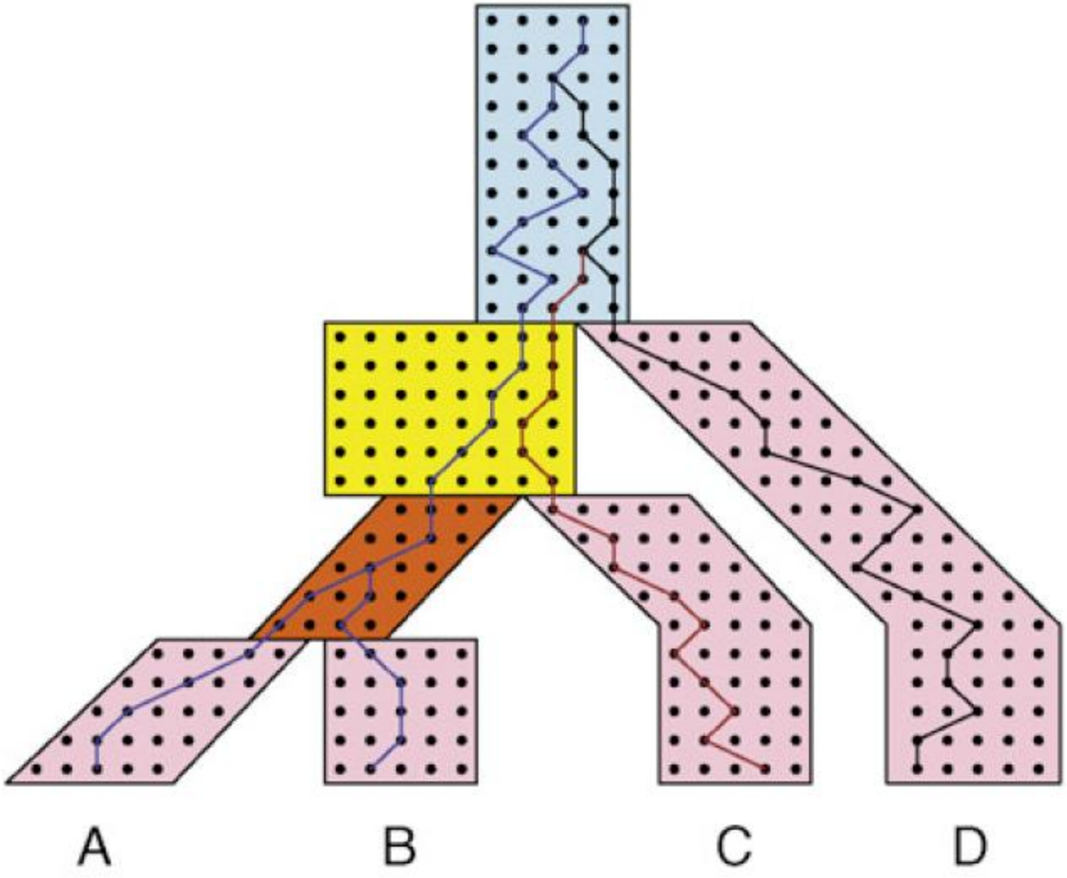
Modeling drift with the coalescent

Wright-Fisher Process
Coalescent Process



Population of chromosomes

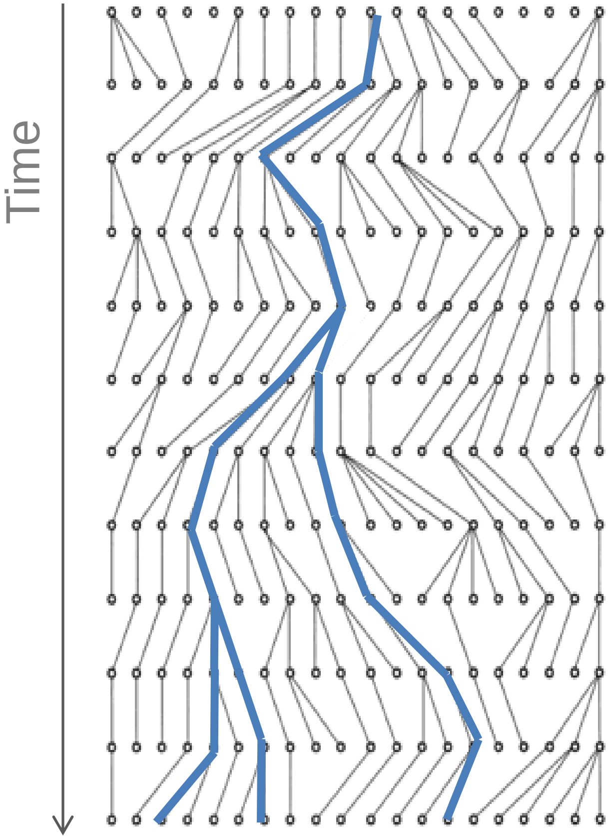
Multispecies coalescent



Degnan 2009

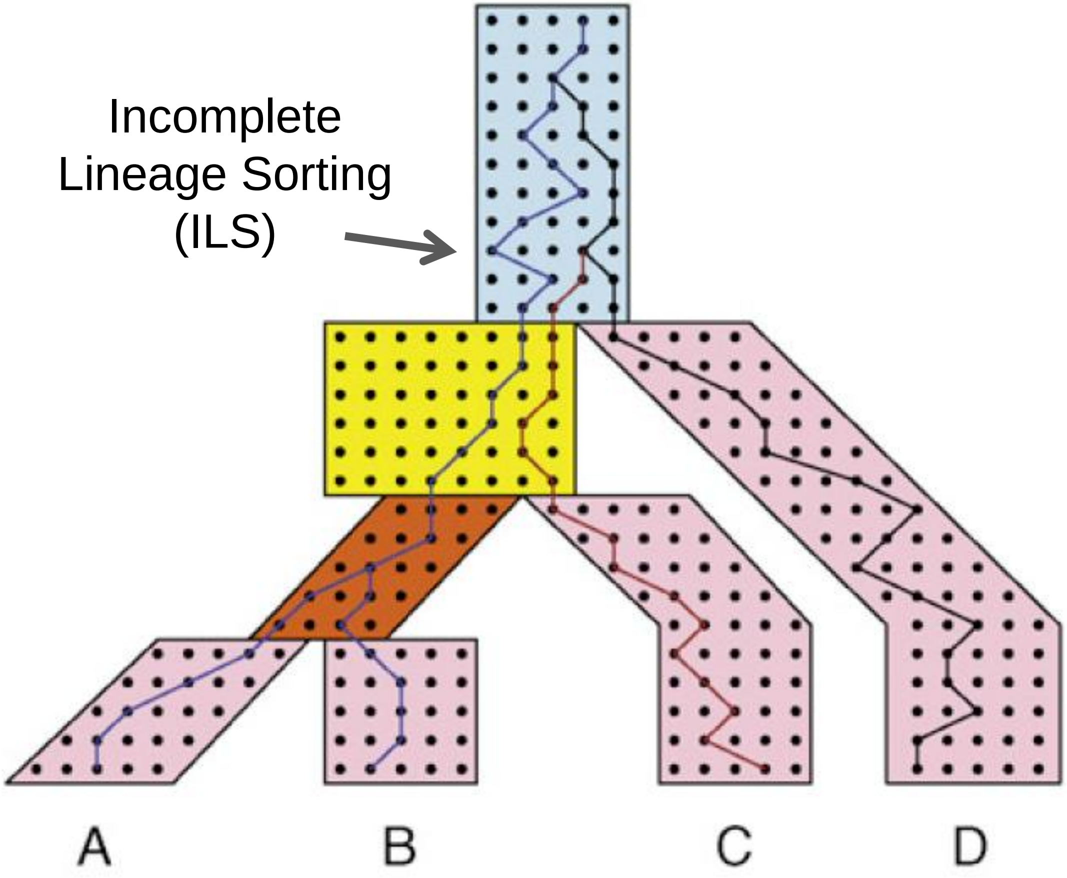
Modeling drift with the coalescent

Wright-Fisher Process
Coalescent Process



Population of chromosomes

Multispecies coalescent

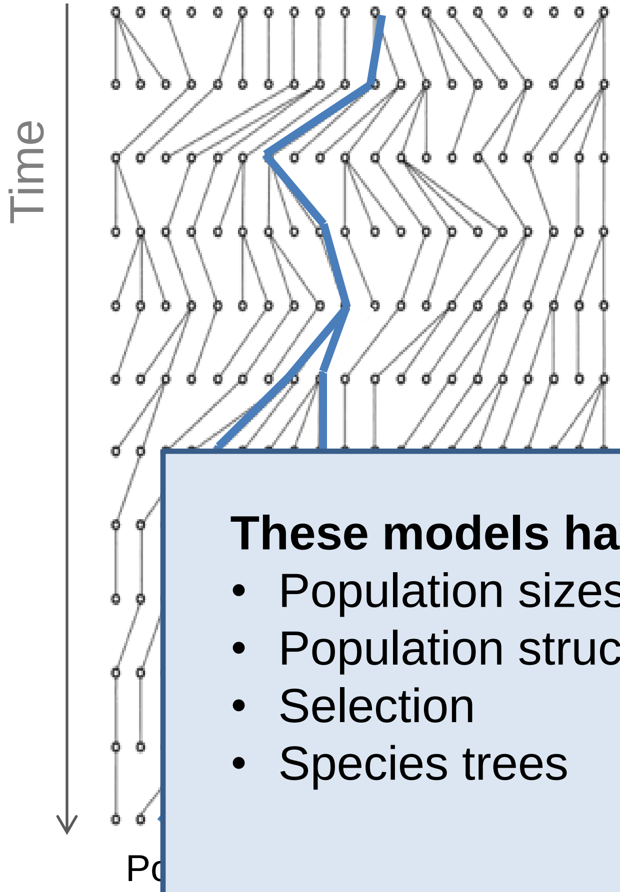


Incomplete
Lineage Sorting
(ILS)

Degnan 2009

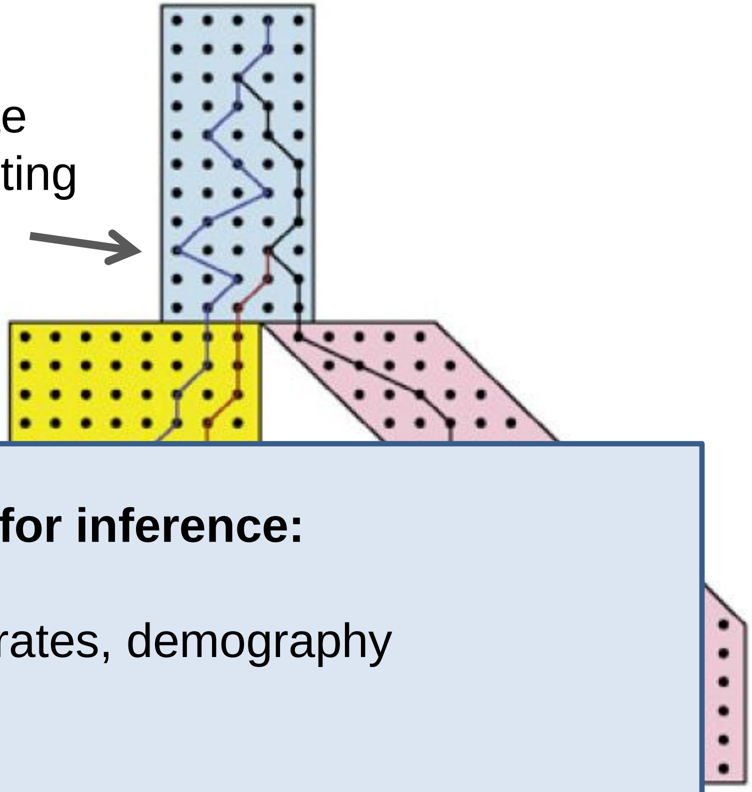
Modeling drift with the coalescent

Wright-Fisher Process
Coalescent Process



Multispecies coalescent

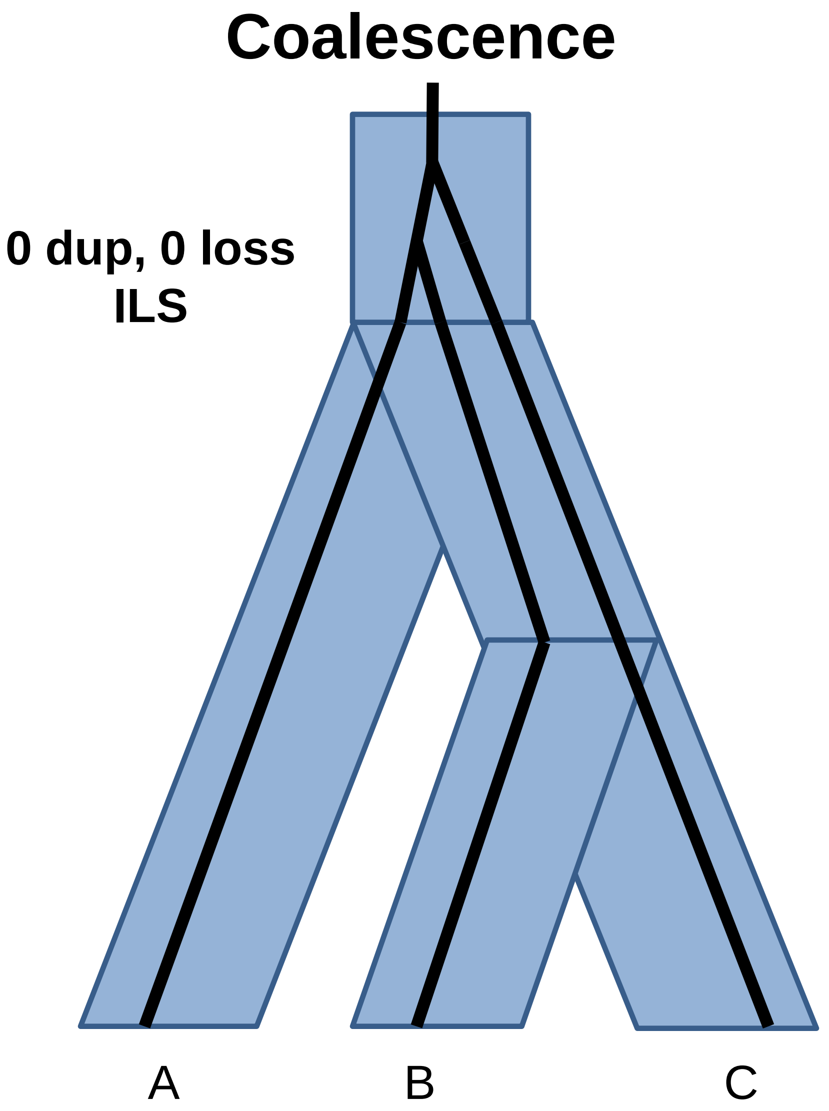
Incomplete
Lineage Sorting
(ILS)



These models have many uses for inference:
- Population sizes, bottle necks
- Population structure, migration rates, demography
- Selection
- Species trees

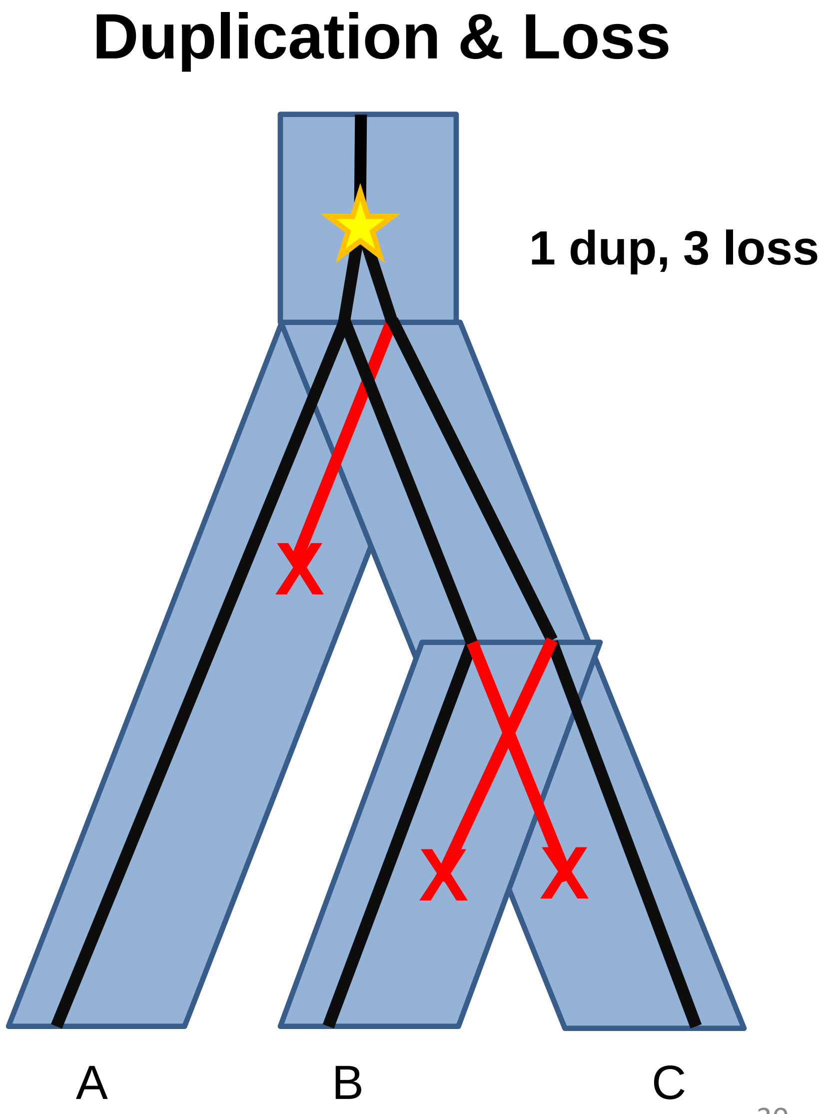
Degnan 2009

The interpretation of a gene tree depends on your model



Coalescence

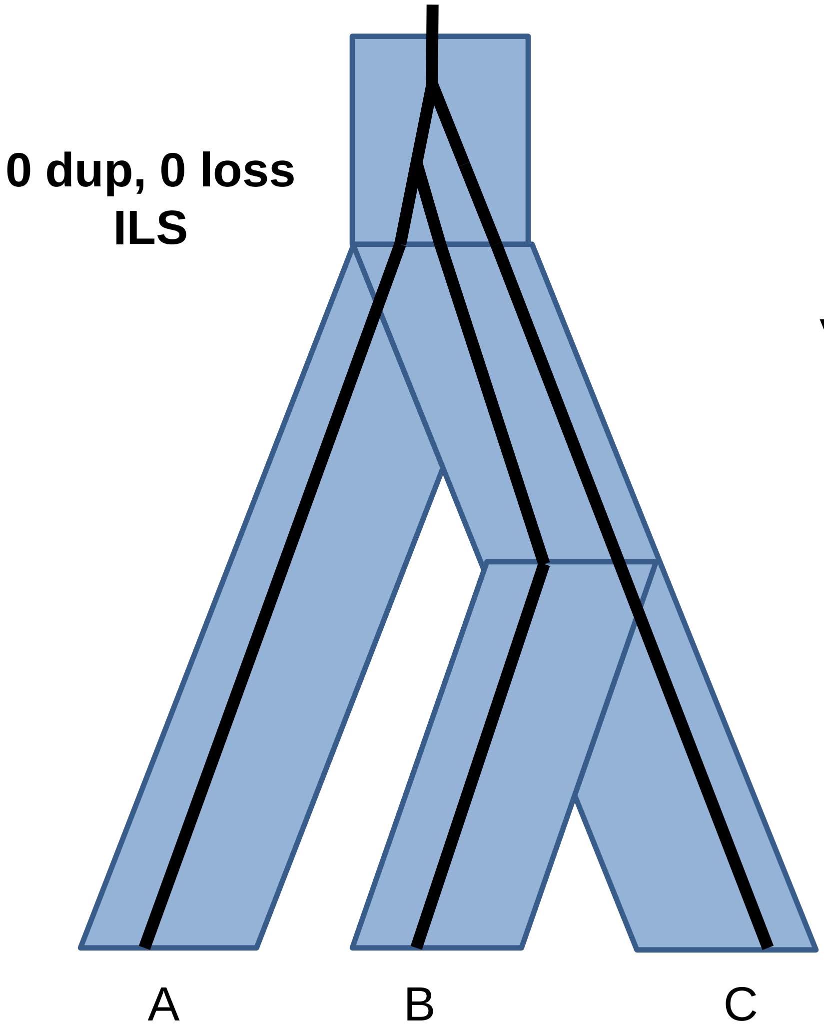
0 dup, 0 loss
ILS



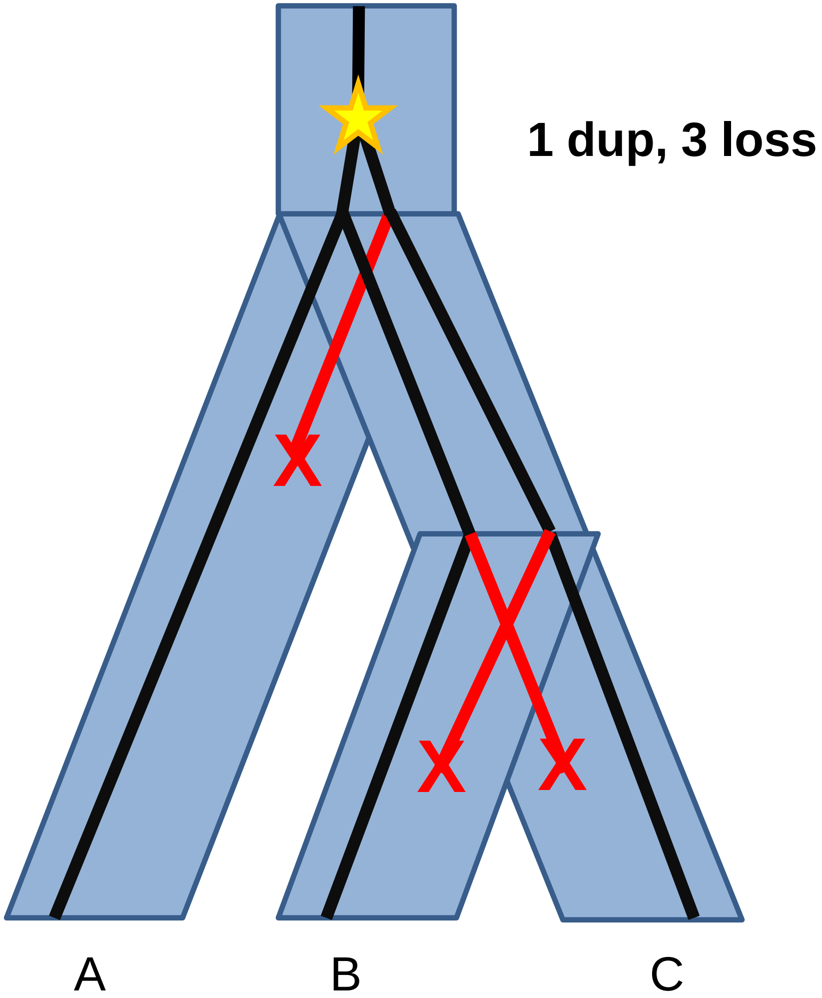
Duplication & Loss

1 dup, 3 loss

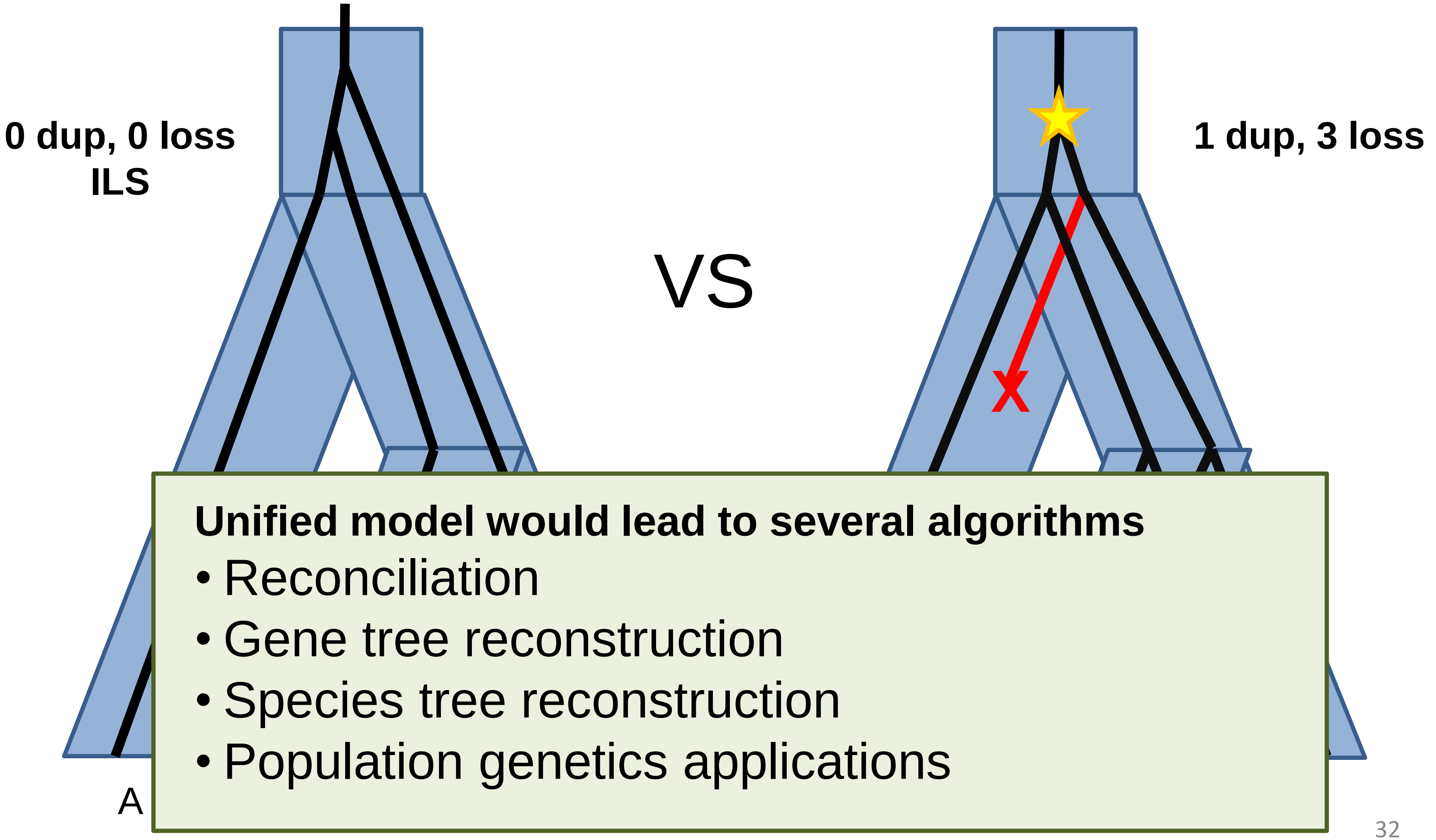
Unified model could capture both



VS



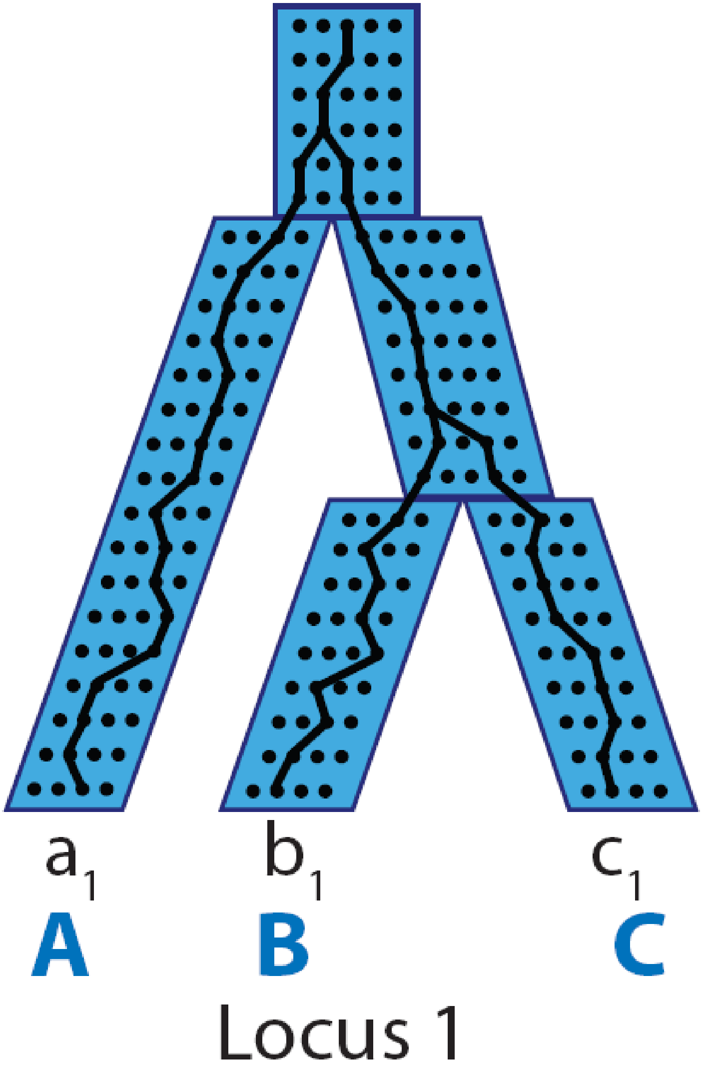
Unified model could capture both



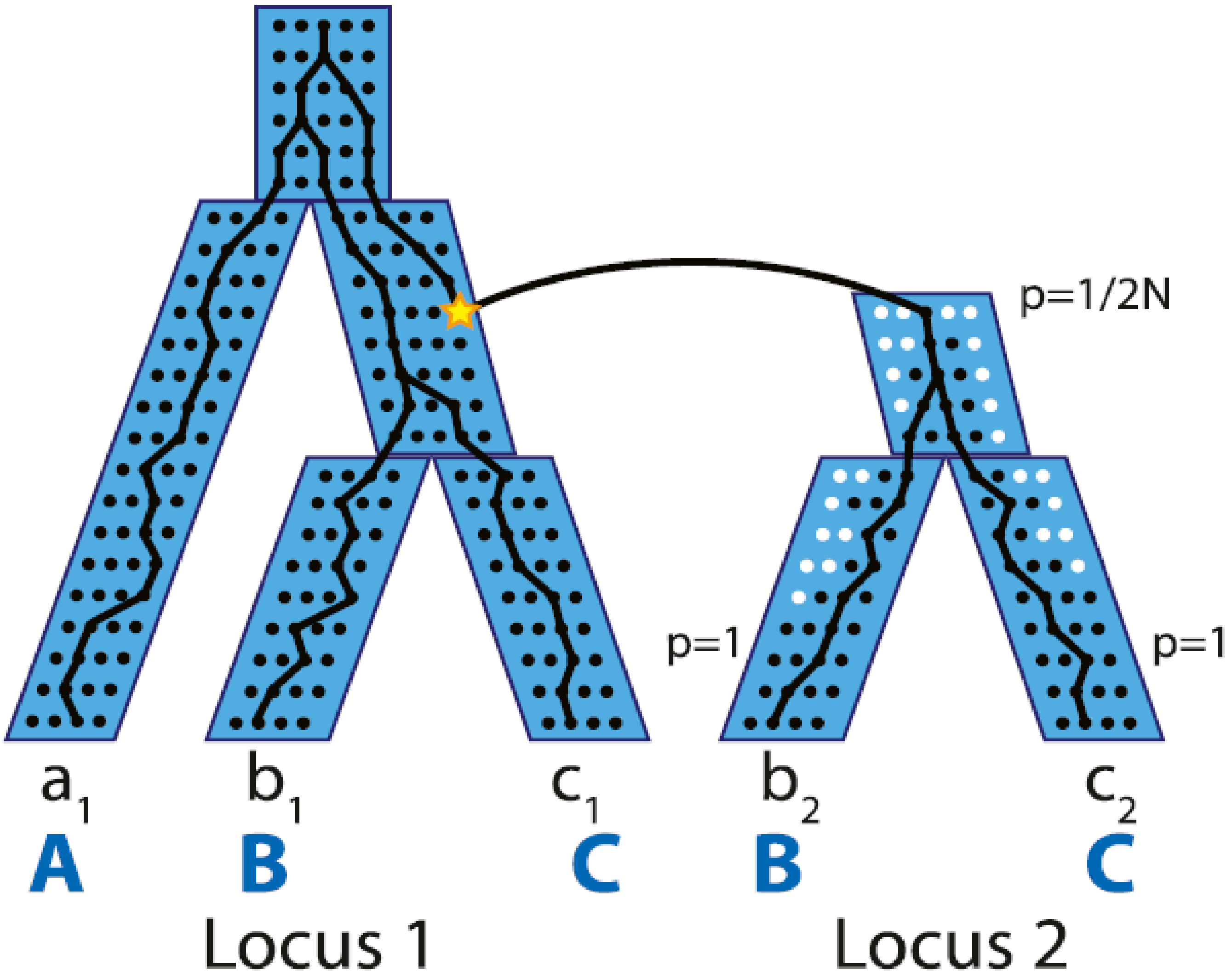
Unified model would lead to several algorithms

- Reconciliation
- Gene tree reconstruction
- Species tree reconstruction
- Population genetics applications

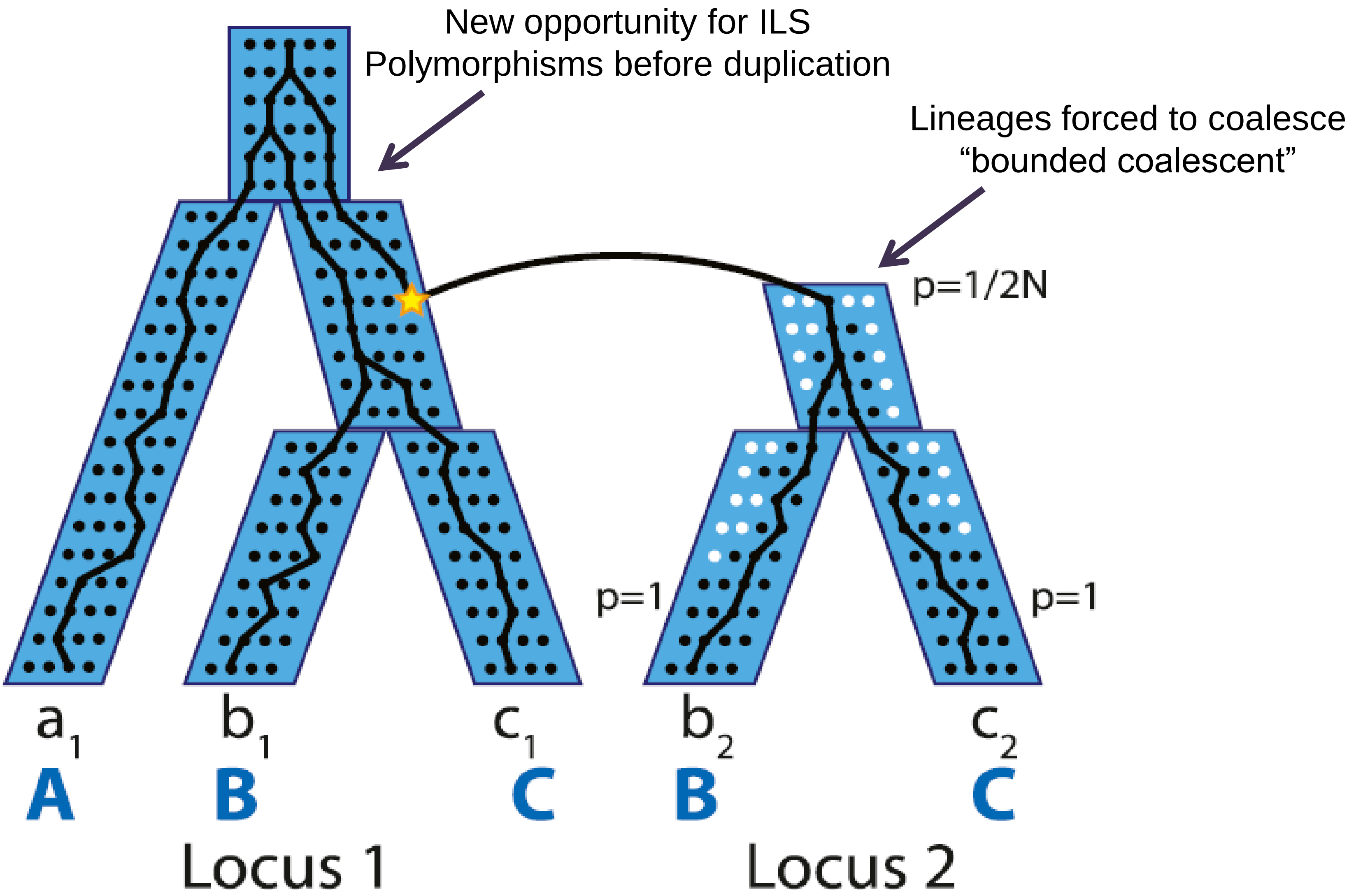
Example of duplication in a population



Example of duplication in a population



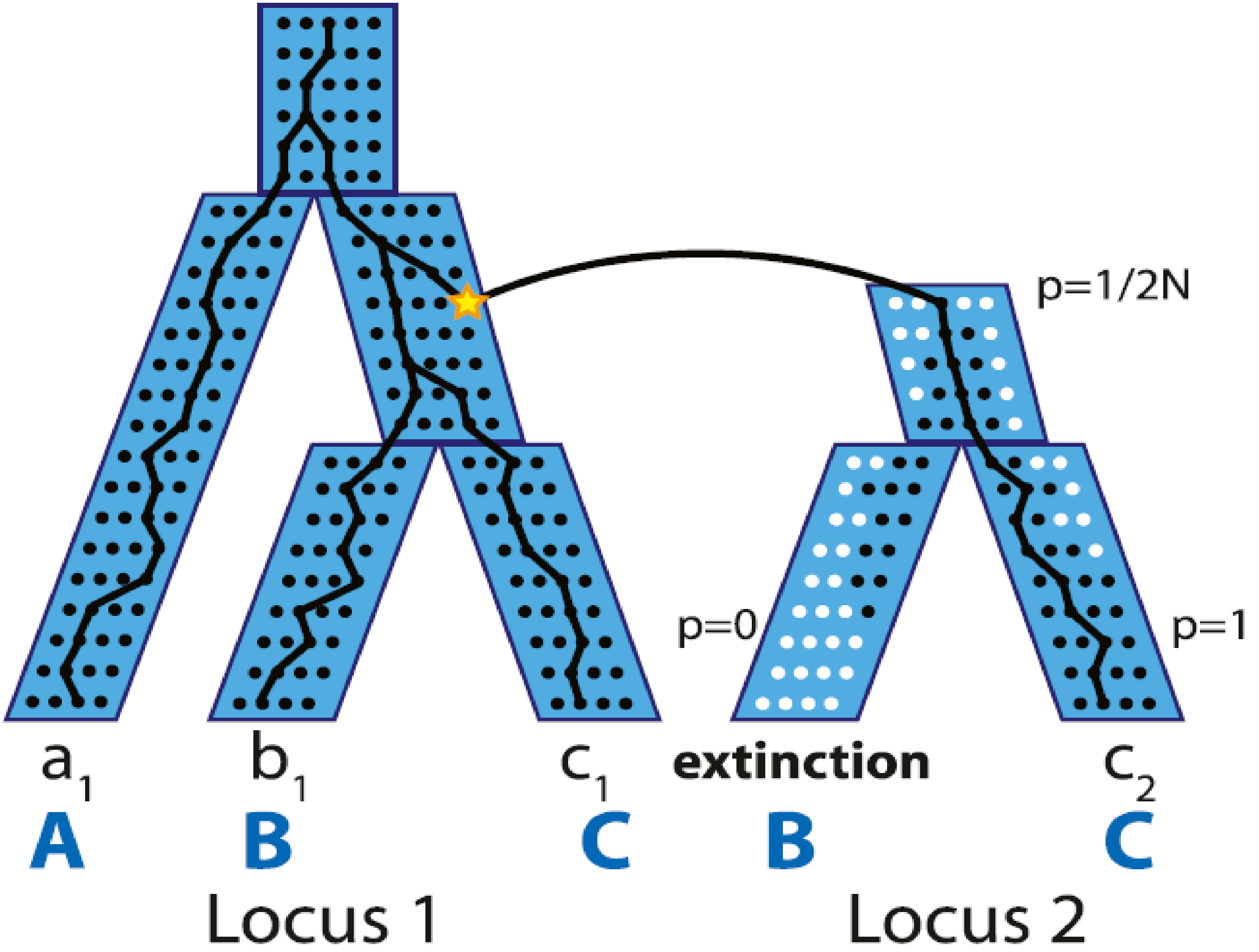
Example of duplication in a population



New opportunity for ILS
Polymorphisms before duplication

Lineages forced to coalesce
“bounded coalescent”

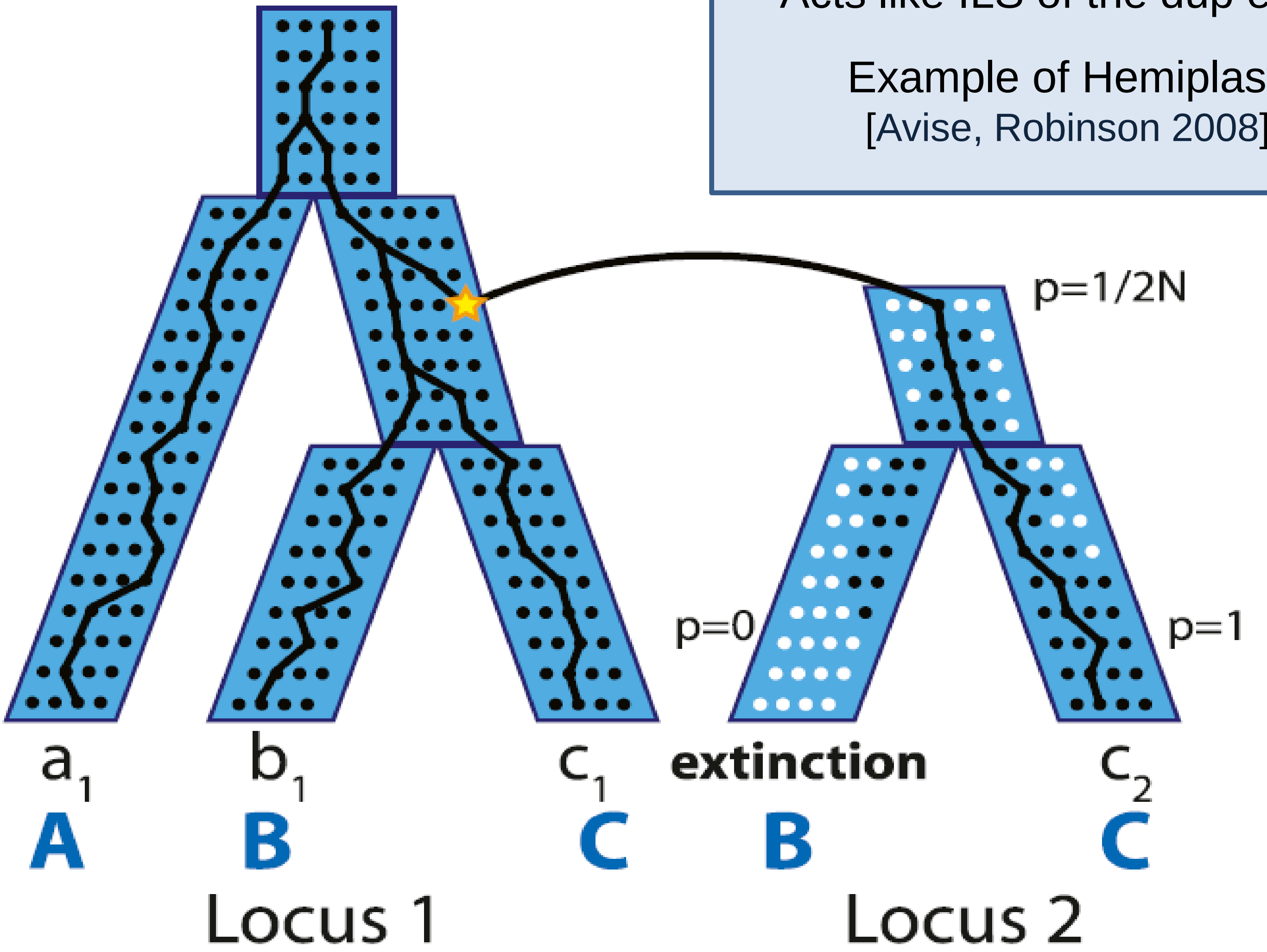
Duplications can fail to fix



Duplications can fail to fix

Acts like ILS of the dup event

Example of Hemiplasy
[Avise, Robinson 2008]

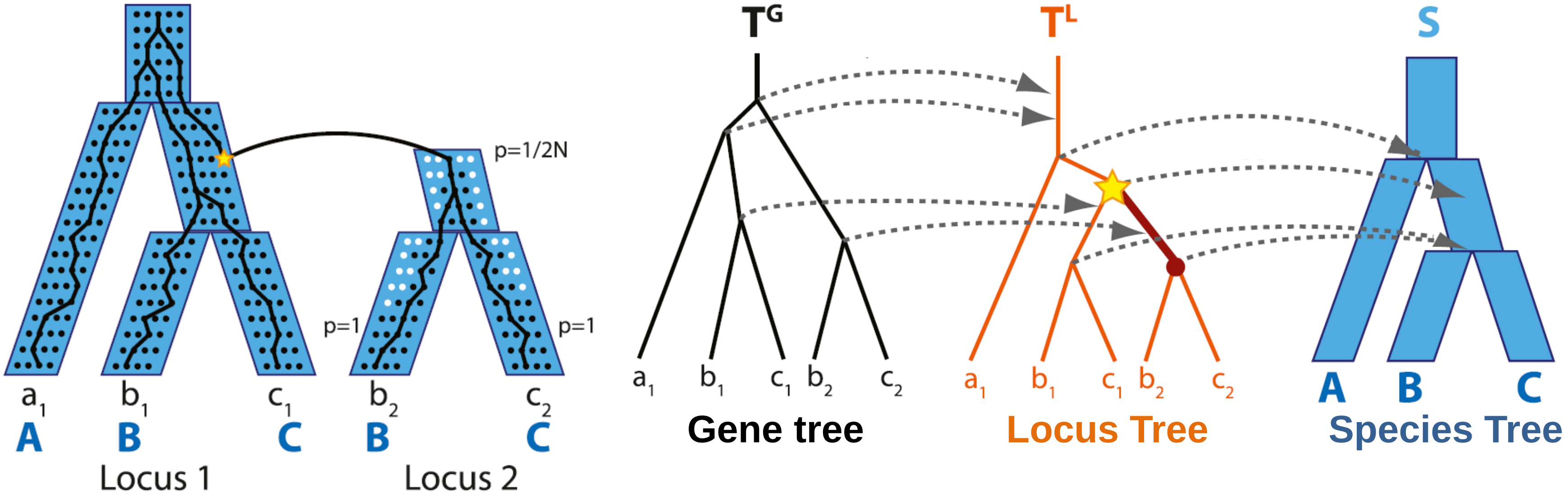


How to unify dup-loss and coalescent models?

- **Realization:** the “gene trees” in dup-loss and coalescent models are distinct objects
  - In coalescent: describes the history of gene sequences
  - In dup-loss: describes history of loci (i.e. changes in copy number)
- **Resolution: “three tree model”**

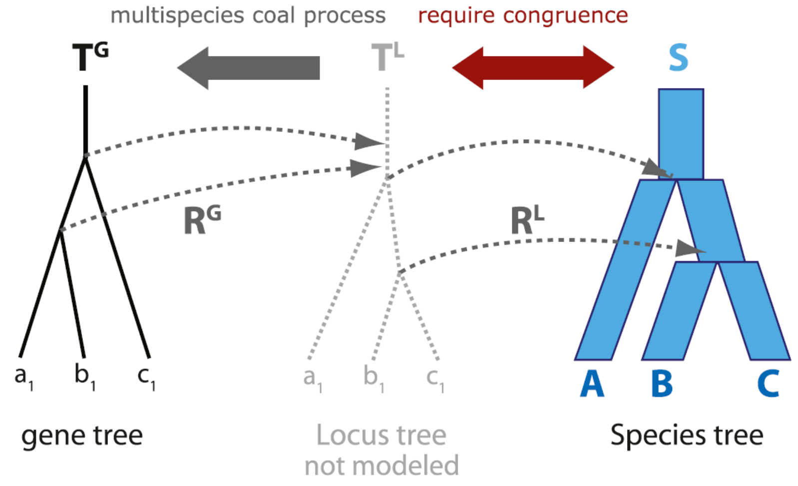
How to unify dup-loss and coalescent models?

- **Realization:** the “gene trees” in dup-loss and coalescent models are distinct objects
  - In coalescent: describes the history of gene sequences
  - In dup-loss: describes history of loci (i.e. changes in copy number)
- **Resolution: “three tree model”**
  - Can track lineages across: individuals, loci, and species
  - **Genes** evolve within **loci** according to the coalescent process
  - **Loci** evolve within **species** according to a birth-death process

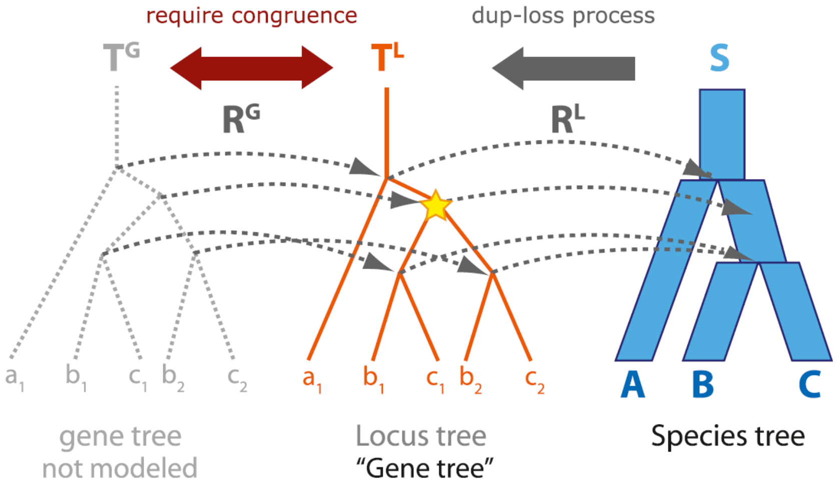


DLCoal model generalizes dup-loss and multispecies coalescent models

- **Multispecies coalescent** assumes locus tree congruent to species tree (i.e. no paralogs)

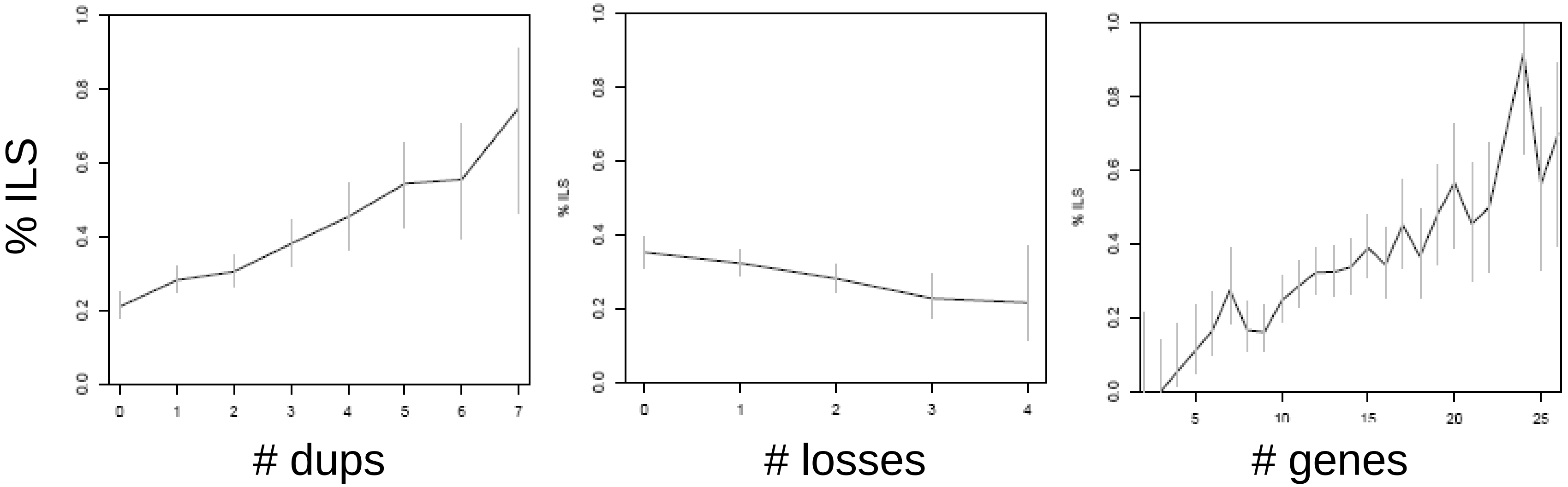


- Previous **dup-loss models** assume gene tree congruent to locus tree (i.e. no ILS)



Simulation with dup,loss,coal:

Large families (more dups) have higher ILS rate

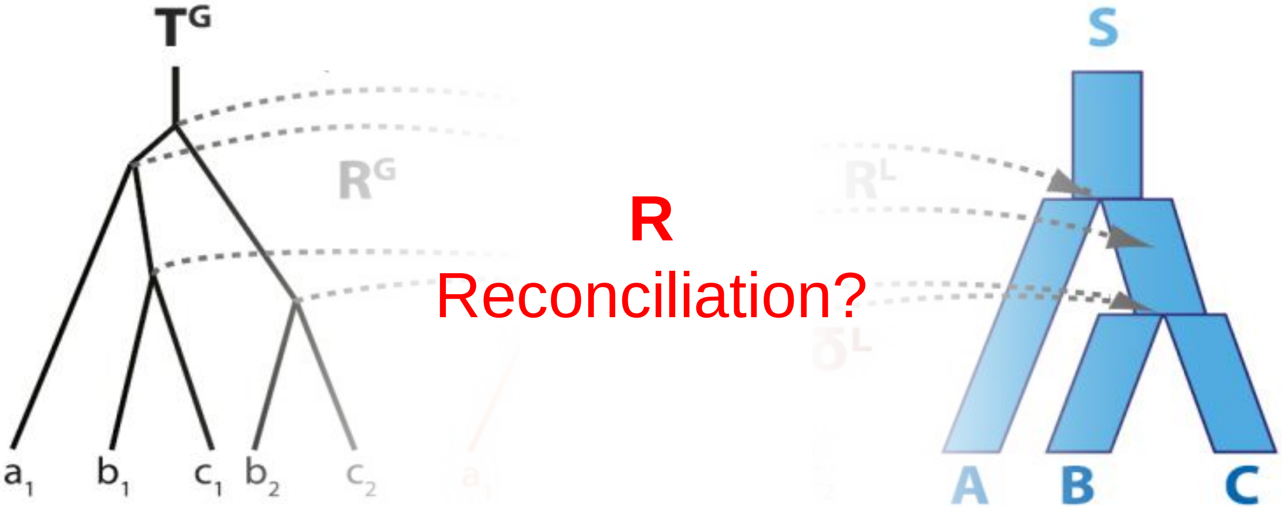


- Duplications break up long branches in locus tree ➔ ILS more likely
- Losses do the reverse, joining branches in the locus tree ➔ ILS less likely

Rasmussen and Kellis. Genome Research. 2012

A new reconciliation method: **DLCoalRecon**

Problem: Given gene tree and species tree, find most likely duplications and losses in the presence of ILS.



R

Reconciliation?

Gene tree

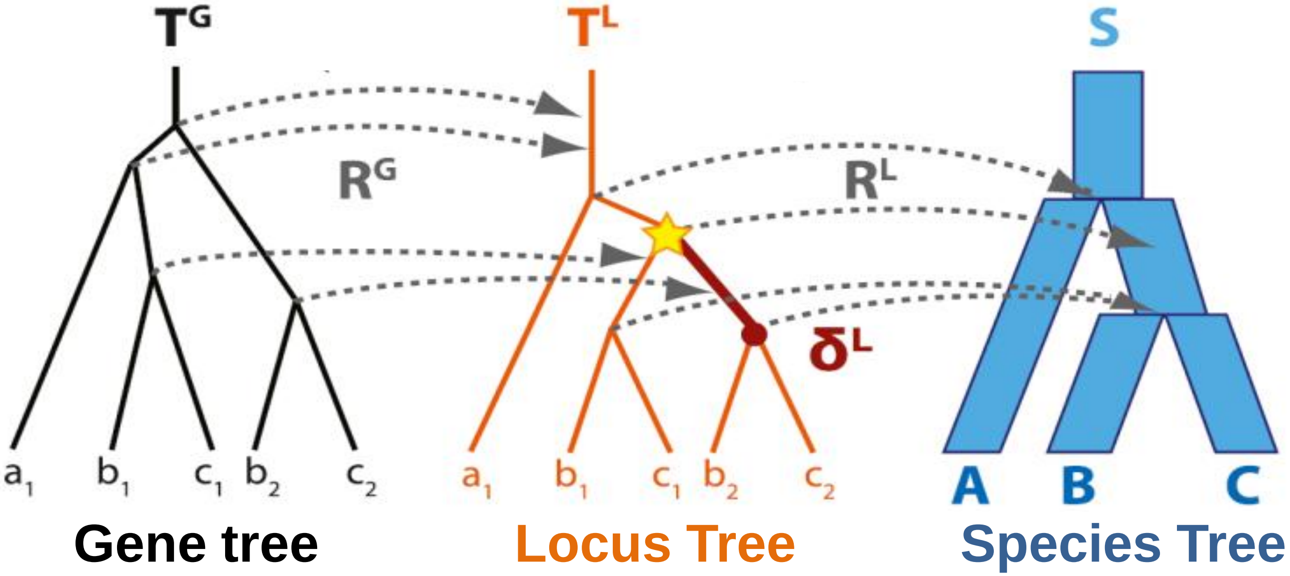
Species Tree

Input: Gene tree $\mathbf{T^G}$, Species tree **S**, model parameters **θ** (pop size, dup/loss rates)

Output: R = **?**

A new reconciliation method: **DLCoalRecon**

Problem: Given gene tree and species tree, find most likely duplications and losses in the presence of ILS.



Input: Gene tree **T^G**, Species tree **S**, model parameters **θ** (pop size, dup/loss rates)

Output: **R** = (Locus tree **T^L**, daughters **δ^L**, mappings **R^G**, and **R^L**)

DLCoalRecon: Maximum A Posterior reconciliation

Input: Gene tree $\mathbf{T^G}$, Species tree **S**, model parameters $\boldsymbol{\theta}$ (pop size, dup/loss rates)

Output: **R** = (Locus tree $\mathbf{T^L}$, daughters $\boldsymbol{\delta^L}$, mappings $\mathbf{R^G}$, and $\mathbf{R^L}$)

Goal: $\hat{\mathbb{R}} = \underset{\mathbb{R}}{\operatorname{argmax}} P(\mathbb{R}|T^G, S, \theta) = \underset{\mathbb{R}}{\operatorname{argmax}} P(\mathbb{R}, T^G|S, \theta)$

Factor probability into previously derived terms:

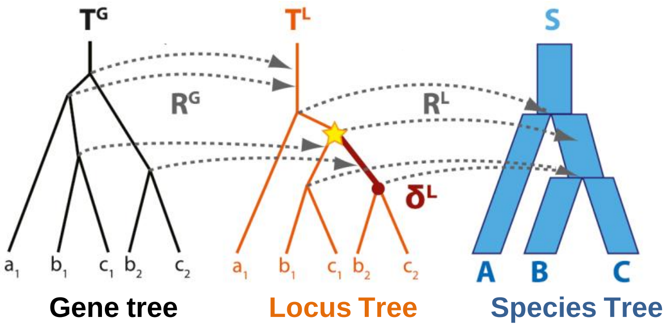
$$P(T^G, \mathbb{R}|S, \theta) = P(\delta^L|T^L, R^L, S) P(T^L, R^L|S, \theta) \int P(T^G, R^G|T^L, \mathbf{t}^L, \delta^L, \mathbf{N}^L) P(\mathbf{t}^L|T^L, R^L, S, \theta) d\mathbf{t}^L$$

Daughters

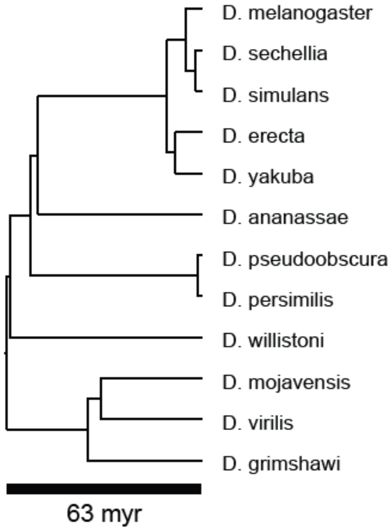
Topology prior
Dup/loss model

Topology prior
Coal model

Branch length prior
Dup/loss model



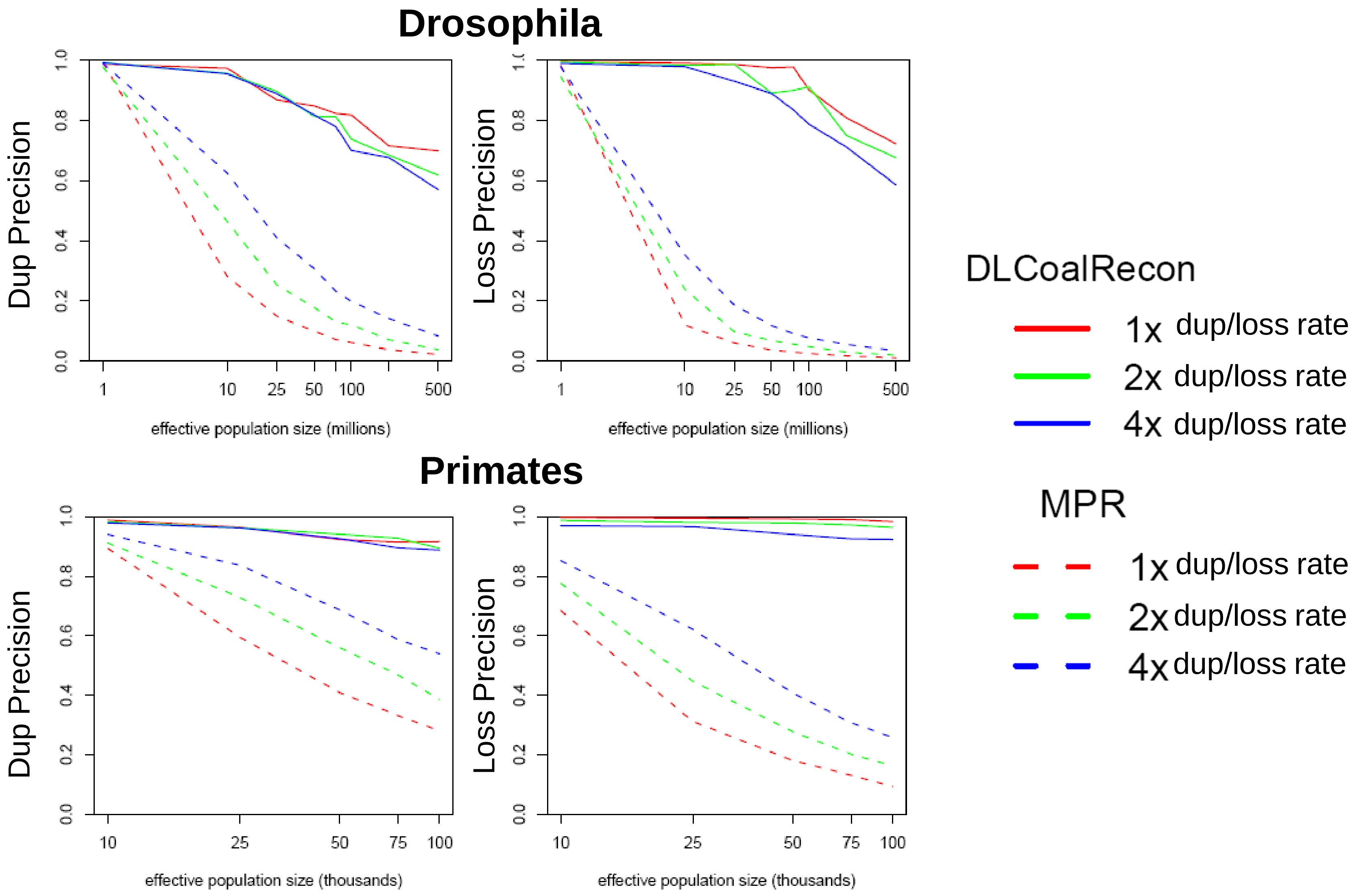
DLCoalRecon outperforms on 500 simulated fly gene families



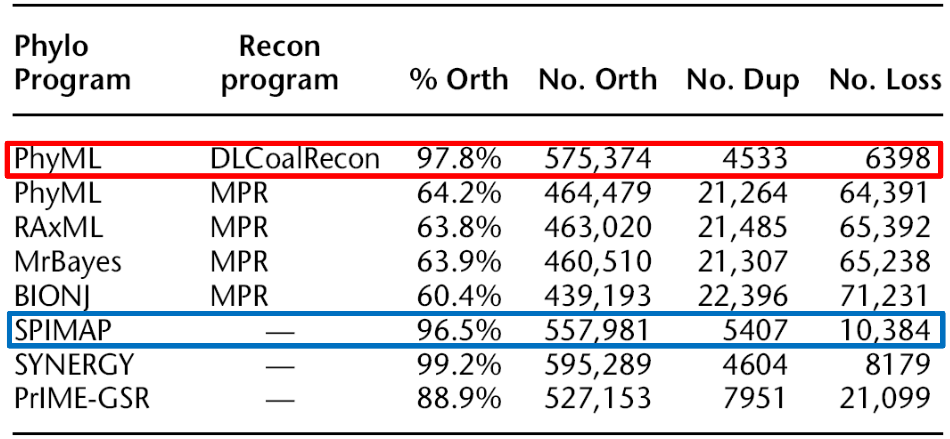
	Actual	MPR	DLCoal Recon
# dup	218	1562	231
dup sensitivity		71.6%	89.9%
dup precision		10.0%	84.8%
# loss	202	4873	200
loss sensitivity		88.6%	96.5%
loss precision		3.7%	97.5%
# orths	32,625	17,680	32,708
orth sensitivity		54.2%	99.6%
orth precision		100%	99.4%

- 0.1 years/generation
- Ne=50x10^{6}
- Dup-loss = 0.0012 event/gene/myr
  - Hahn et al 2007

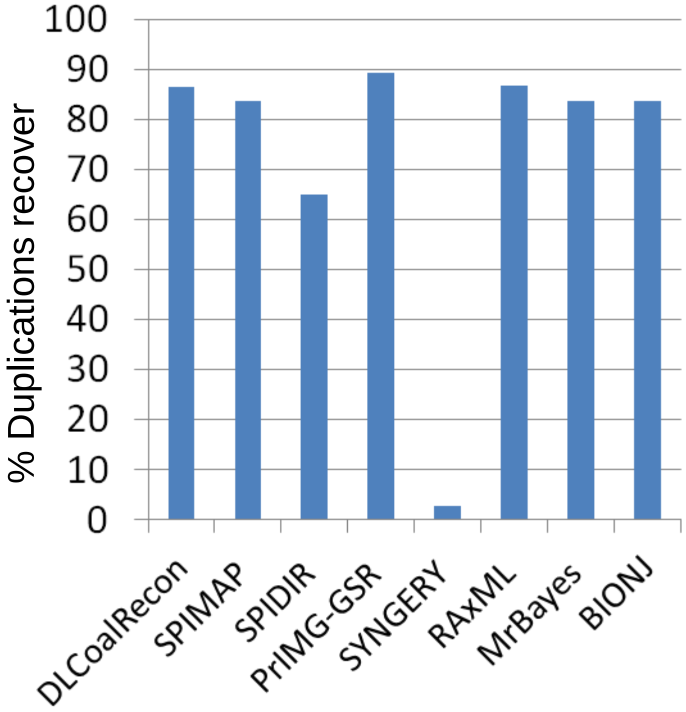
Performance holds over a variety of populations sizes and dup/loss rates



DLCoalRecon outperforms on 16 fungi genomes



Phylo Program	Recon program	% Orth	No. Orth	No. Dup	No. Loss
PhyML	DLCoalRecon	97.8%	575,374	4533	6398
PhyML	MPR	64.2%	464,479	21,264	64,391
RAxML	MPR	63.8%	463,020	21,485	65,392
MrBayes	MPR	63.9%	460,510	21,307	65,238
BIONJ	MPR	60.4%	439,193	22,396	71,231
SPIMAP	—	96.5%	557,981	5407	10,384
SYNERGY	—	99.2%	595,289	4604	8179
PrIME-GSR	—	88.9%	527,153	7951	21,099

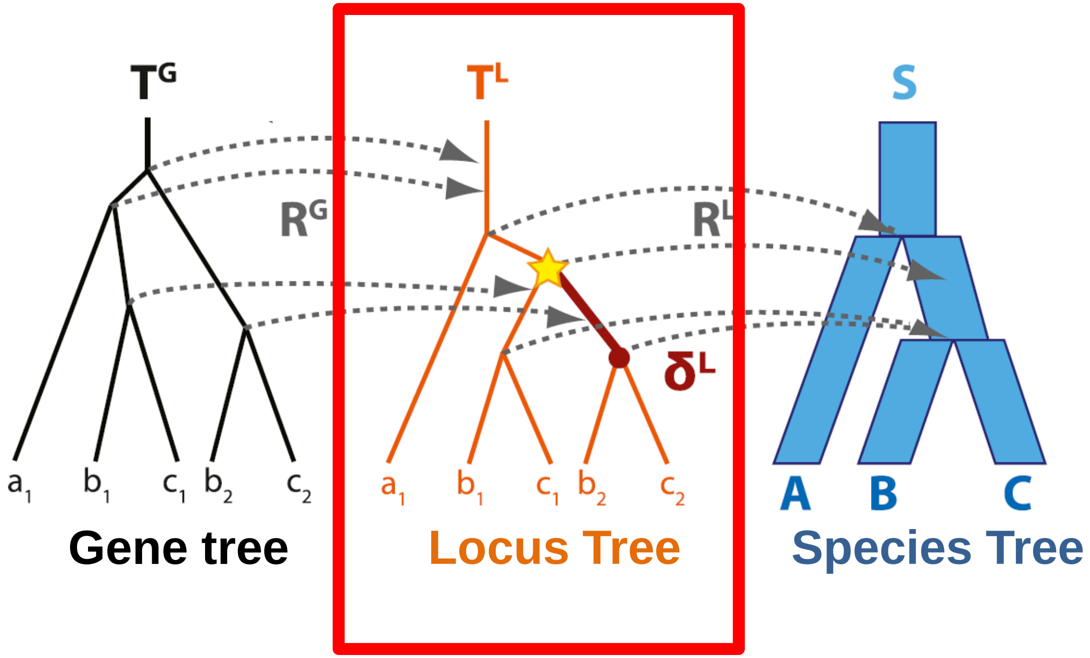


(1) Further improved syntenic ortholog recovery

(2) Even fewer dup-losses inferred

(3) Improved gene conversion recovery

Importance of the locus tree



- Dup & losses can only meaningfully be annotated on the locus tree.
- Locus tree is a gene tree with the ILS removed and is often more meaningful
- DLCoal has been critical for accurate orthology determination in the modENCODE project

Acknowledgements

Adam Siepel
Cornel
Postdoc advisor

Manolis Kellis
MIT
Postdoc & PhD advisor

Candida sequencing and analysis
Geraldine Butler (UC Dublin)
Mike Lorenz (Texas)
Neil Gow (Aberdeen)
Christina Cuomo (Broad Institute)
Bruce Birren (Broad Institute)
Mike Lin (MIT)

STAR-MP: domain evolution
Yi-chieh Wu
Manolis Kellis

TreeFix: gene tree correction
Yi-chieh Wu
Mukul Bansal

Horizontal transfer in Strep
Sang Chul Choi
Melissa Hubisz
Ilan Gronau
Michael Stanhope

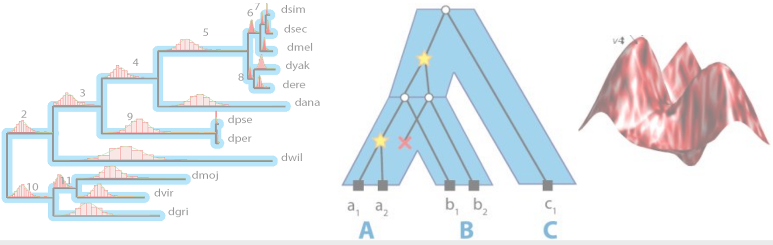
Funding sources
Ruth L. Kirschstein National Research Service Award
Cornell Center for Comparative and Population Genomics Fellowship

Developing models of evolution

Combining different evolutionary events & processes

Duplication, Loss & Substitution

SPIDIR (GR 2007), SPIMAP (MBE 2010),
TreeFix (SystBio 2012)

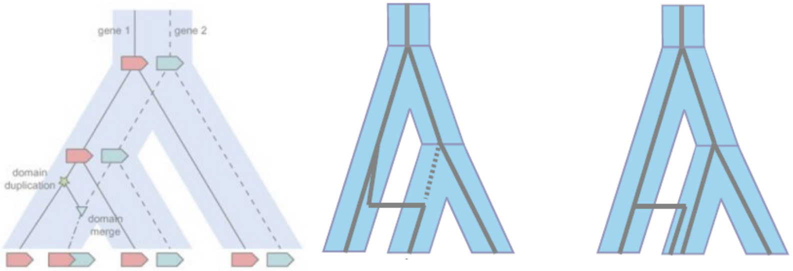


Fusion & Fission

STAR-MP (MBE 2011)

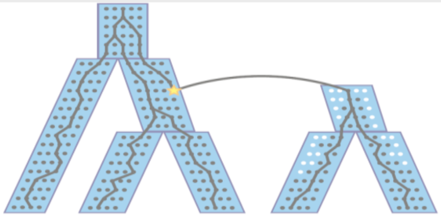
Additive & Replacing transfer

(MBE 2012)



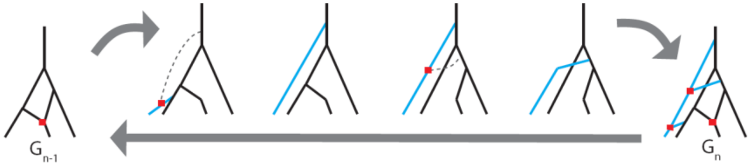
Duplication, Loss & Coalescence

DLCoal (GR 2012)

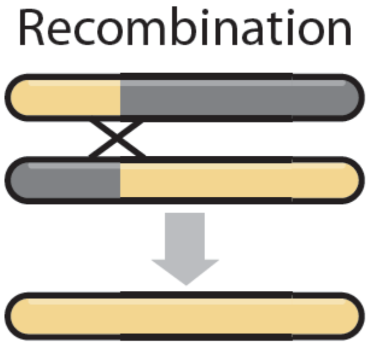


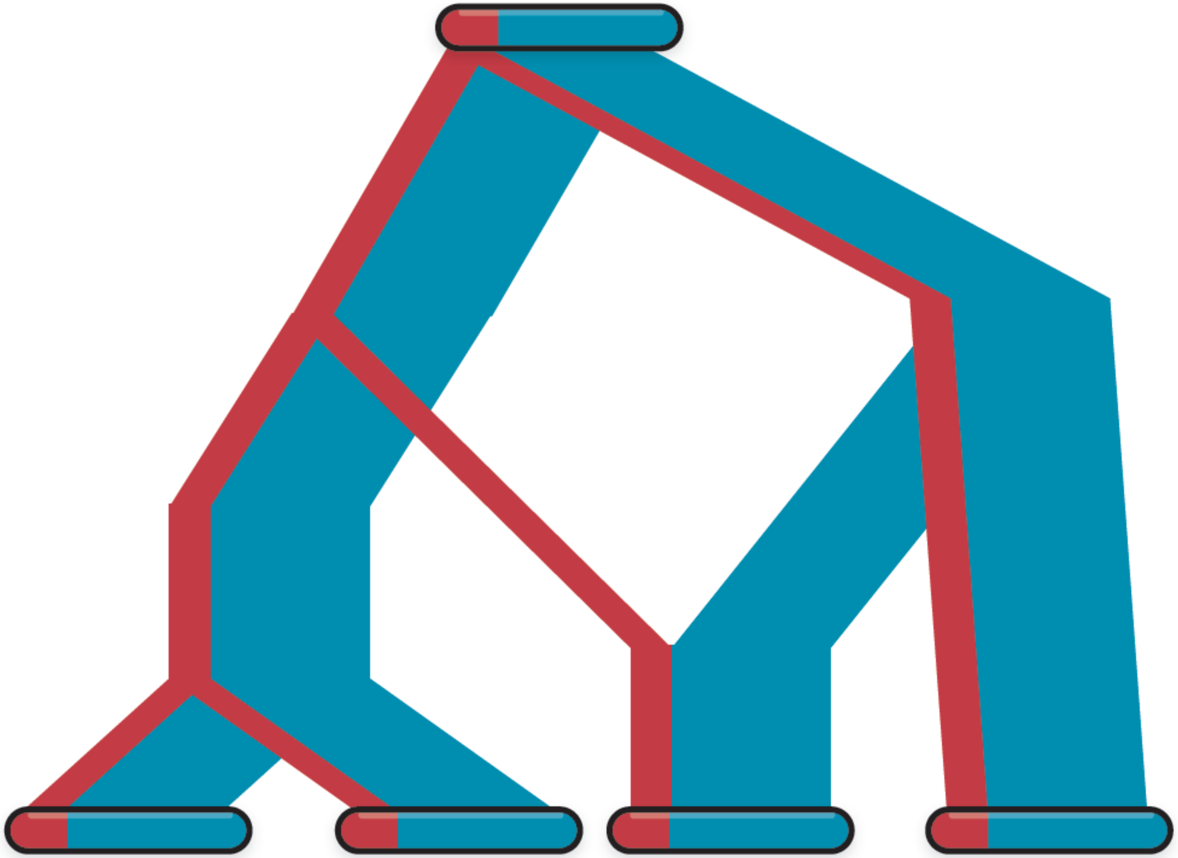
Coalescence & Recombination

ARGHMM

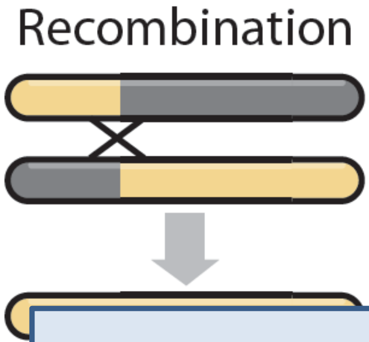


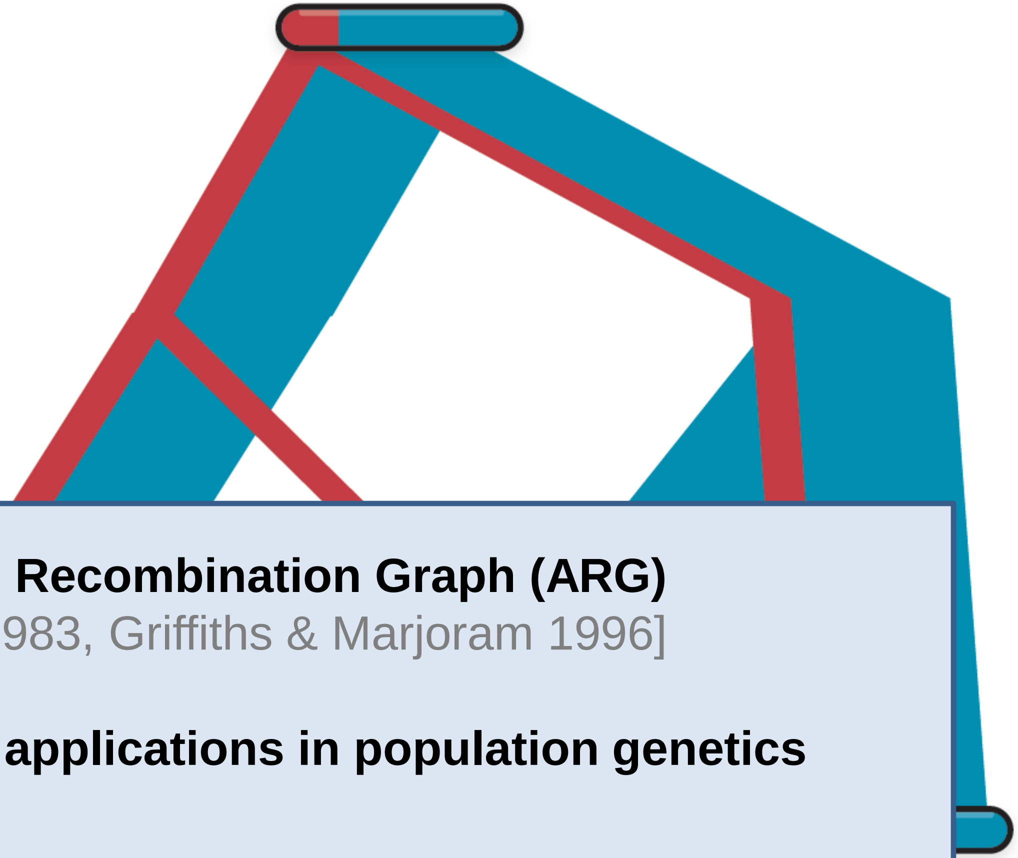
With recombination history is no longer a simple tree





With recombination history is no longer a simple tree



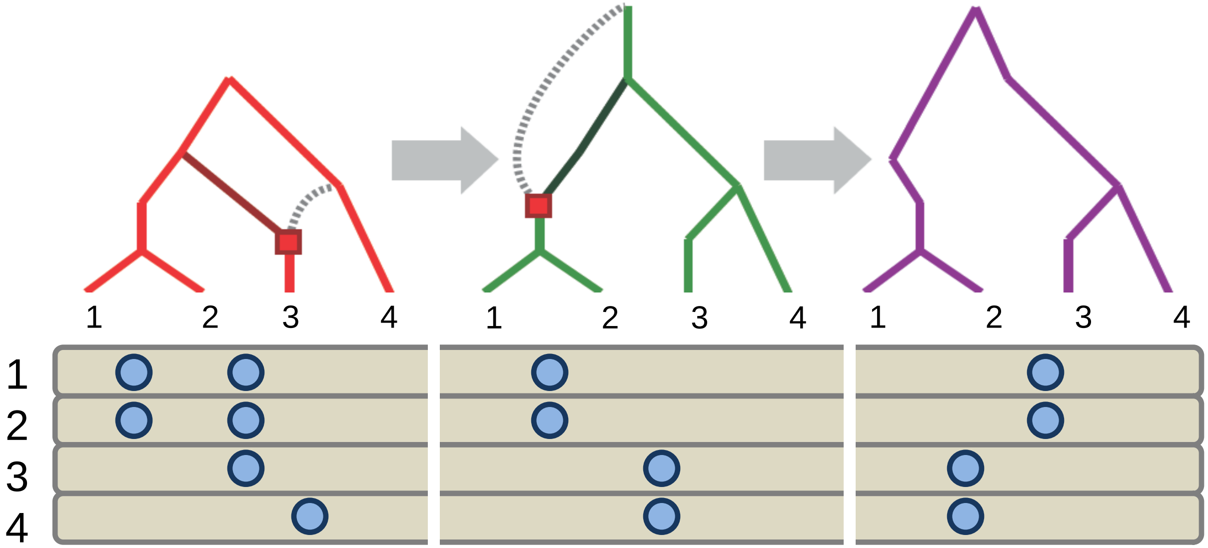


Ancestral Recombination Graph (ARG)
[Hudson 1983, Griffiths & Marjoram 1996]

ARGs have many applications in population genetics

- Selection
- Demography
- Trait mapping

Breaking an alignment into local trees

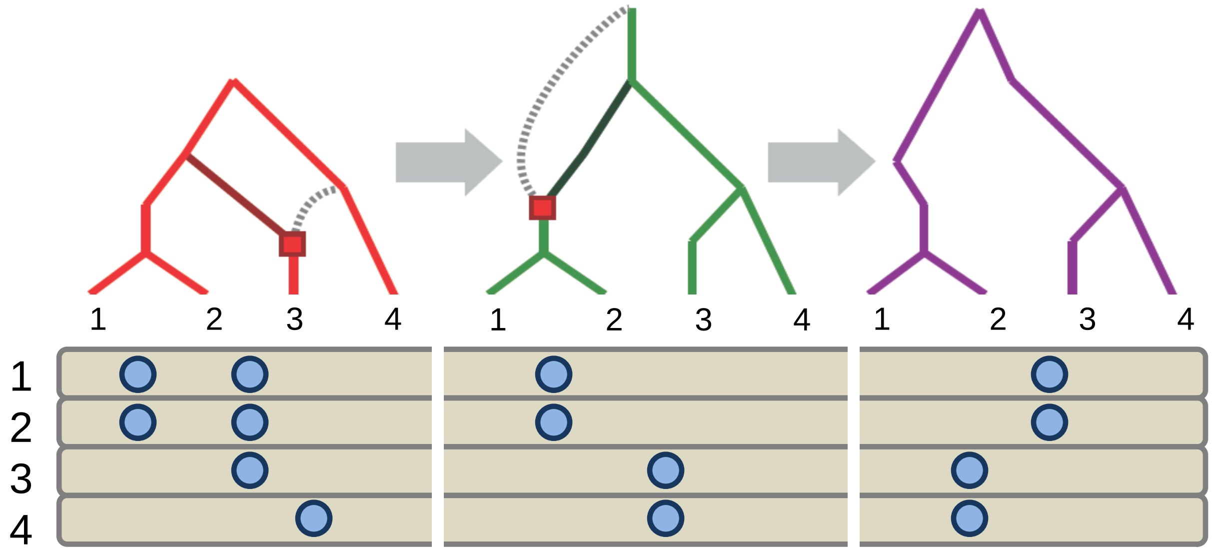


- **Local trees** contain same information as the **ARG**

- **However,** too little information per block to build directly

- **Idea:** trees are correlated ➔ pool information across blocks
  - Recombinations break and re-coalesce a single branch (e.g. SPR operations)

Approximation: sequential Markov coalescent



Assume a **Markov process** for local trees [McVean and Cardin 2005]

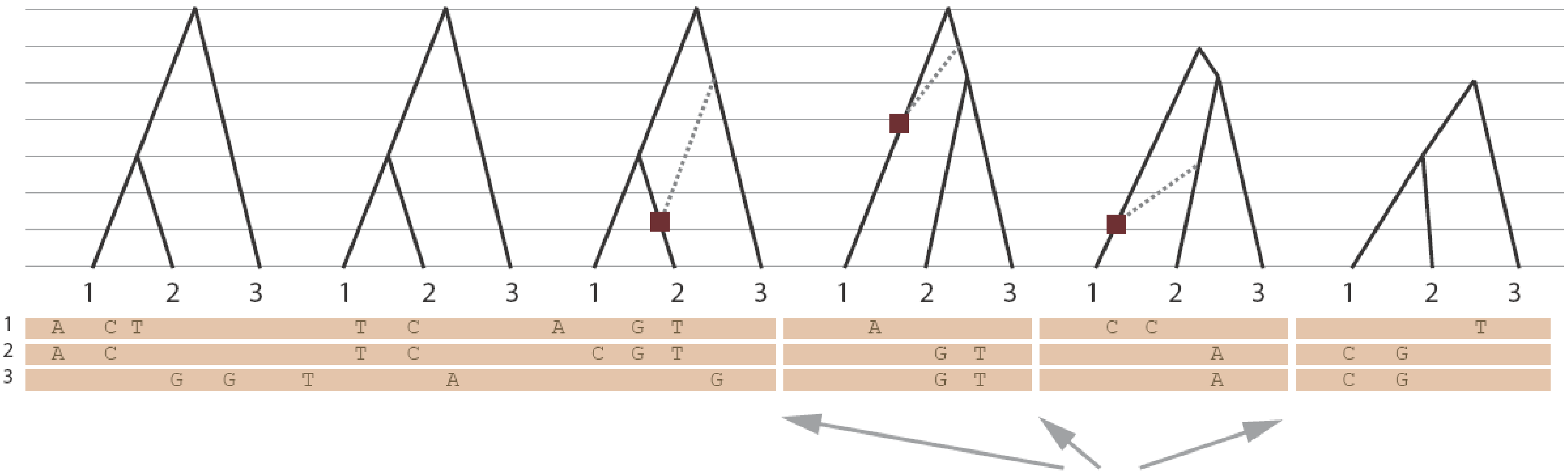
$$P(T_i|T_1,...,T_{i-1}) = P(T_i|T_{i-1})$$

Applications:

- Enabled many efficient sequence **simulation** programs
    - FastCoal, Fastsimcoal, MACS, etc

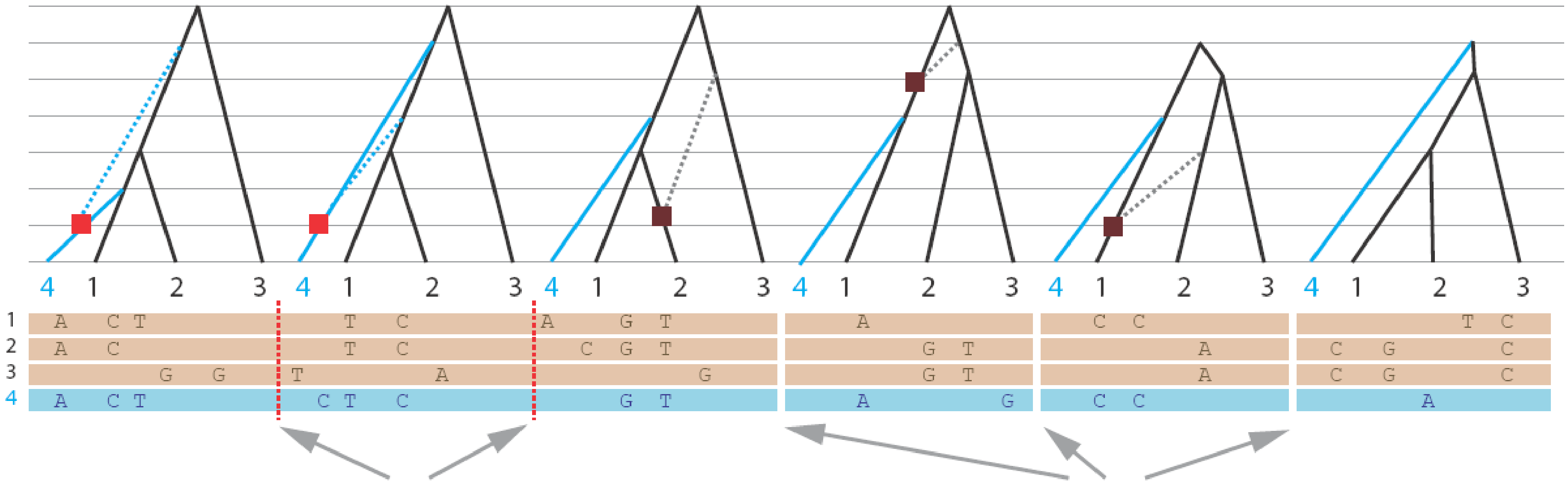
- Can it be used for **inference**?

Building up local trees one sequence at a time



Given recombination break points

Building up local trees one sequence at a time

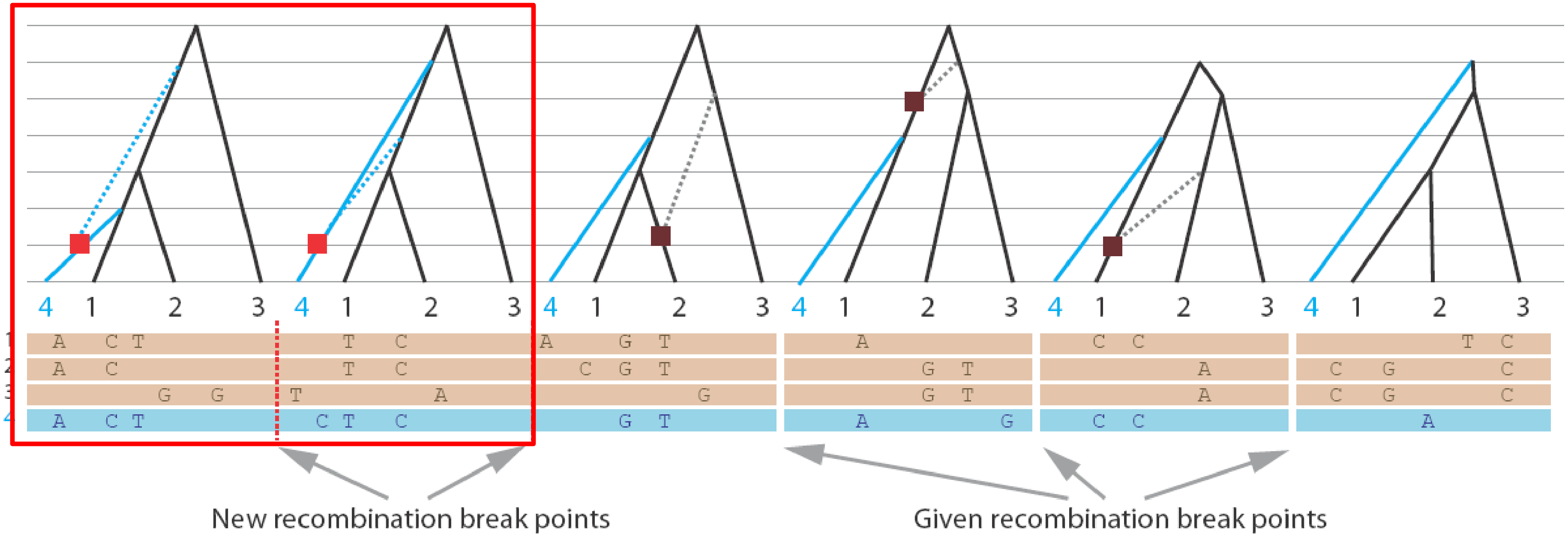


New recombination break points

Given recombination break points

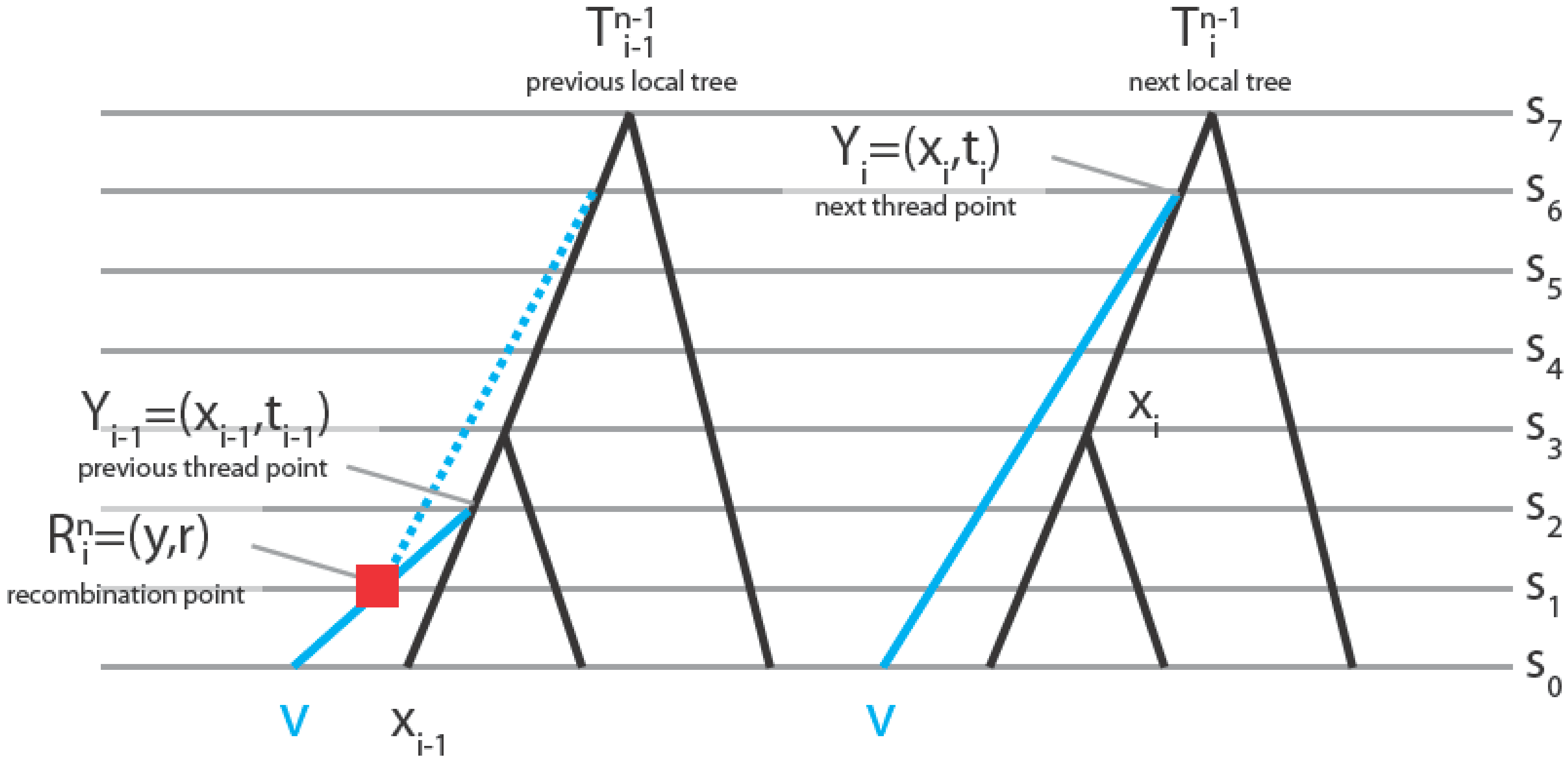
- Adding sequence ➔ adding one more branch
- We call this “**threading**” a sequence into local trees

Building up local trees one sequence at a time



- Adding sequence ➔ adding one more branch
- We call this “**threading**” a sequence into local trees

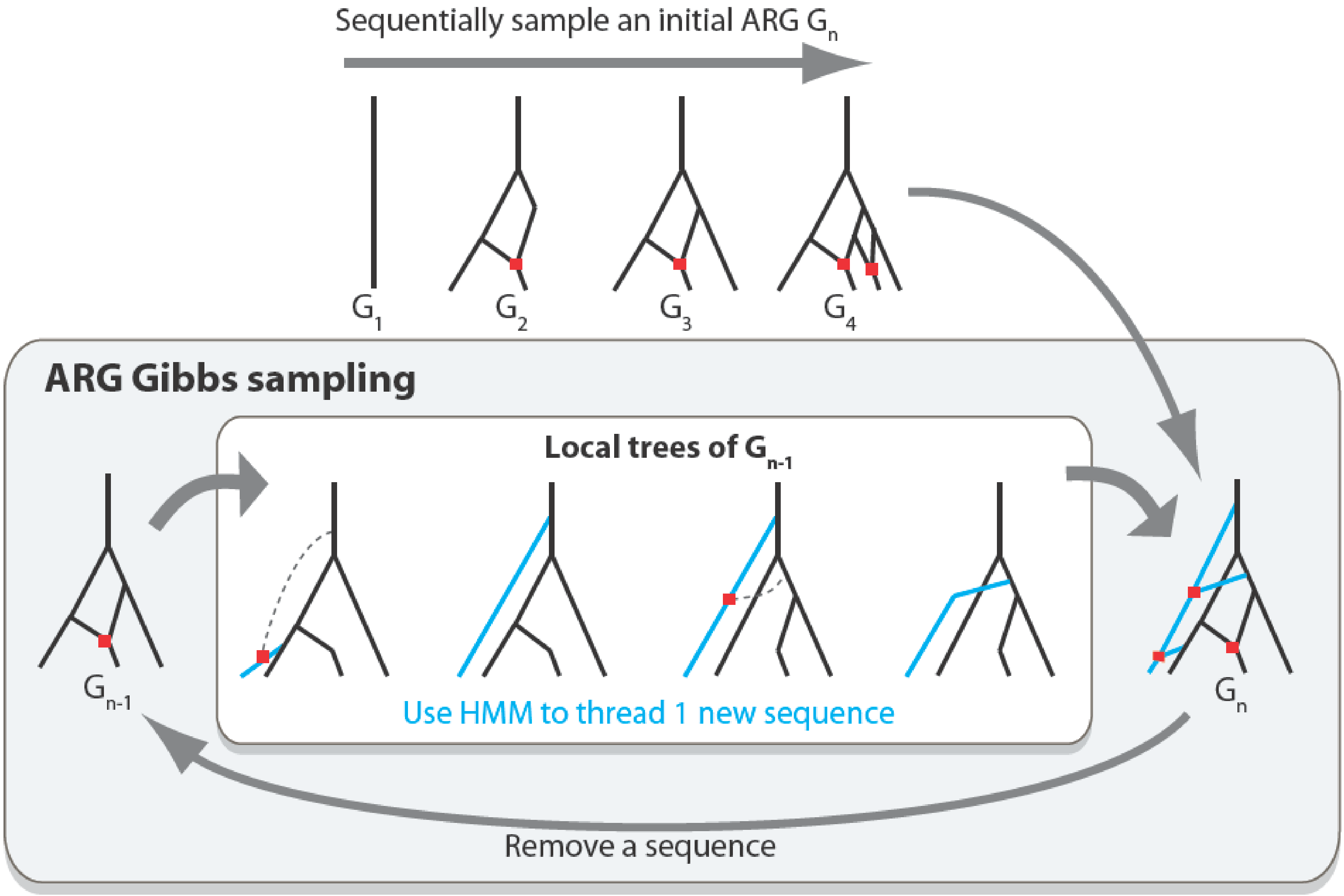
Use Hidden Markov Model to add new sequence



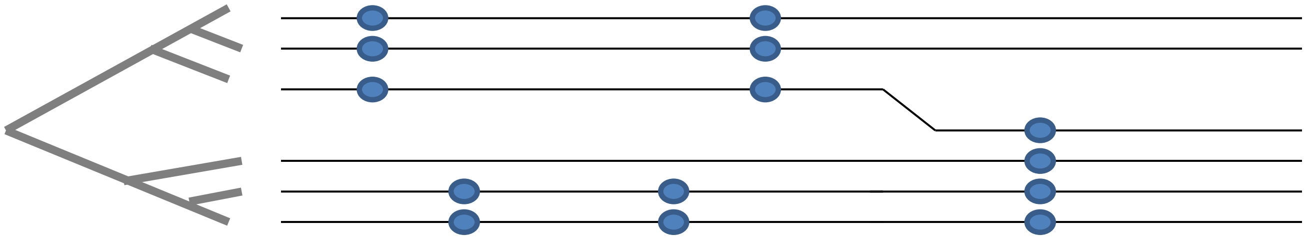
Problem: Given ARG $\mathbf{G_{n-1}}$ and sequence data **D**, sample the **threading** $\mathbf{Y_1,\ldots,Y_L}$

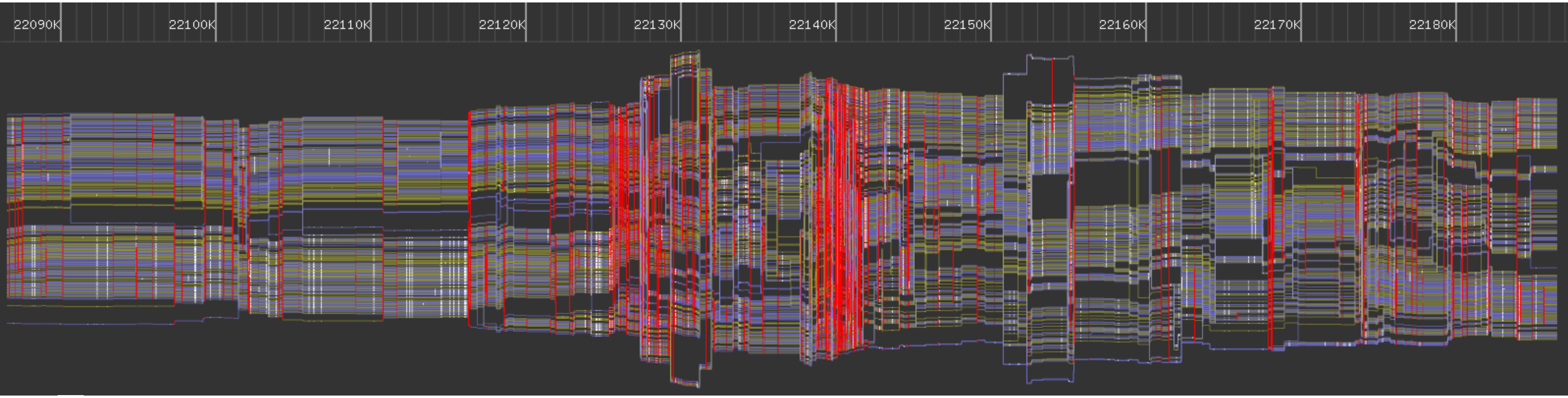
- **HMM** definition:
  - **Hidden states:** threading path $\mathbf{Y_1,\ldots,Y_L}$
  - **Transitions:** derived from DSMC model
  - **Emissions:** columns in sequence alignment **D**
- Use forward algorithm and stochastic traceback to sample $\mathbf{P(G_n \mid G_{n-1}, D)}$

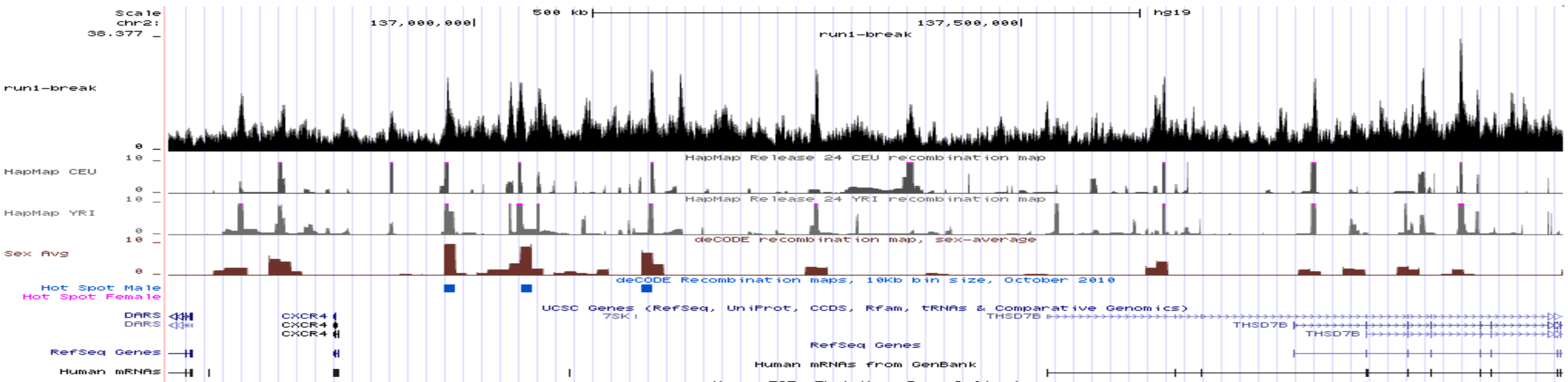
Use threading to build full ARG



Reconstructing ancestral haplotype structure and recombination hotspots







Advantages over other approaches

- **Scales to many more sequences**
  - *CoalHMM:* Hobolth, Christensen, Mailund, Schierup. Genome Research. 2007
    - 4-6 sequences
  - *PSMC:* Li and Durbin. Nature. 2011
    - 2 sequences

- **Scales to longer sequences**
  - *LAMARC:* Kuhner. Bioinformatics. 2006

- **Captures more information by using full local trees**
  - *PAC:* Paul, Steinrucken, Song. Genetics. 2011
    - Considers only “trunk” genealogies

- **Correctly samples from the posterior distribution**
  - *MARGARITA:* Minichiello and Durbin. AJHG. 2006
    - Heuristic sampling approach

Future directions

- Estimate genome-wide ARGs for a core set of human genomes
  - "Reference panel" for ancestry
  - Would allow coalescence-based:
    - Phasing, imputation, local ancestry

- Large-scale ARG-based inference of demography
  - Estimate smaller older tracks of IBD
  - Infer-based on local genealogies

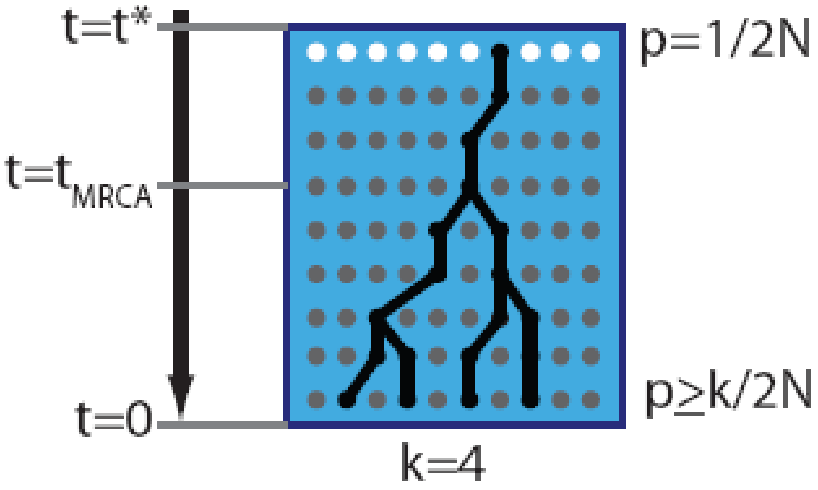
- ARG-based inference of selection
  - Estimate allele ages
  - Regions of recent purifying selection

First version of a dup,loss,coal model

For a first version, we make the following assumptions

- New duplicates begin at unlinked loci
- We model no gene conversion
- **Hemiplasy assumption:**
    - no event (duplications or losses) under goes hemiplasy
    - Namely, full extinction or no extinctions

Building up the model: Bounded coalescent



- **Given:**
  - new mutation (black) at t*
  - k lineages at t=0, all of them have mutation

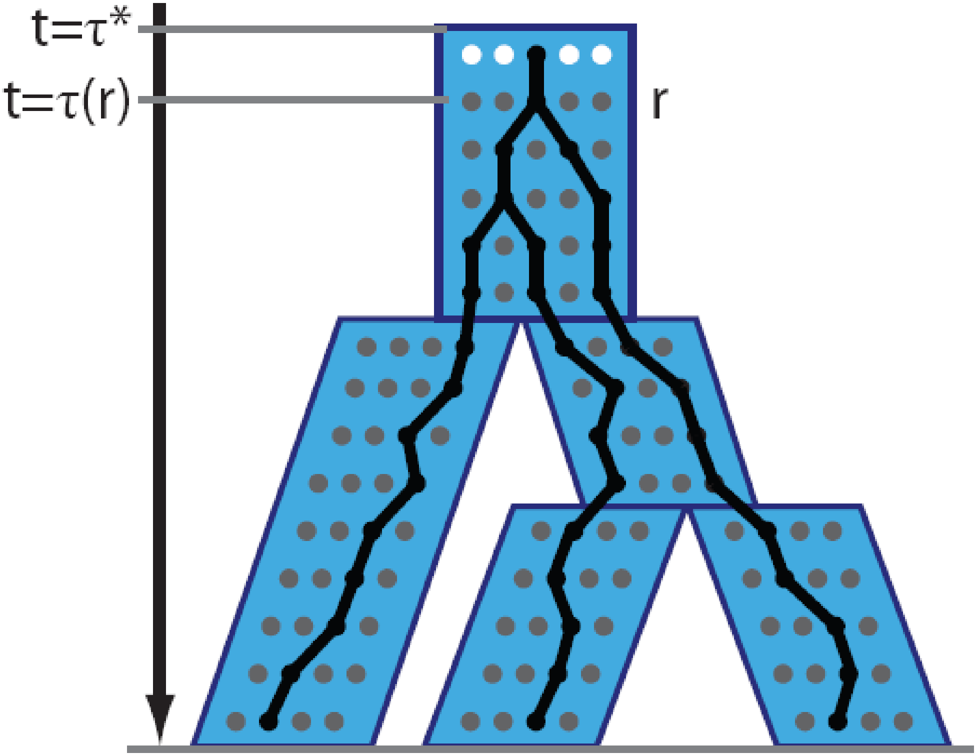
- Equivalent to conditioning t_{MRCA}<t*

$$P(t|t_{MRCA} < t^*, k, N) = \frac{P(t, t_{MRCA} < t^*|k, N)}{P(t_{MRCA} < t^*|k, N)}$$

$$= \begin{cases} \dfrac{P(t|k, N)}{P(t_{MRCA} < t^*|k, N)}, & \text{if } t_{MRCA} < t^* \\ 0, & \text{otherwise} \end{cases}$$

Building up the model:

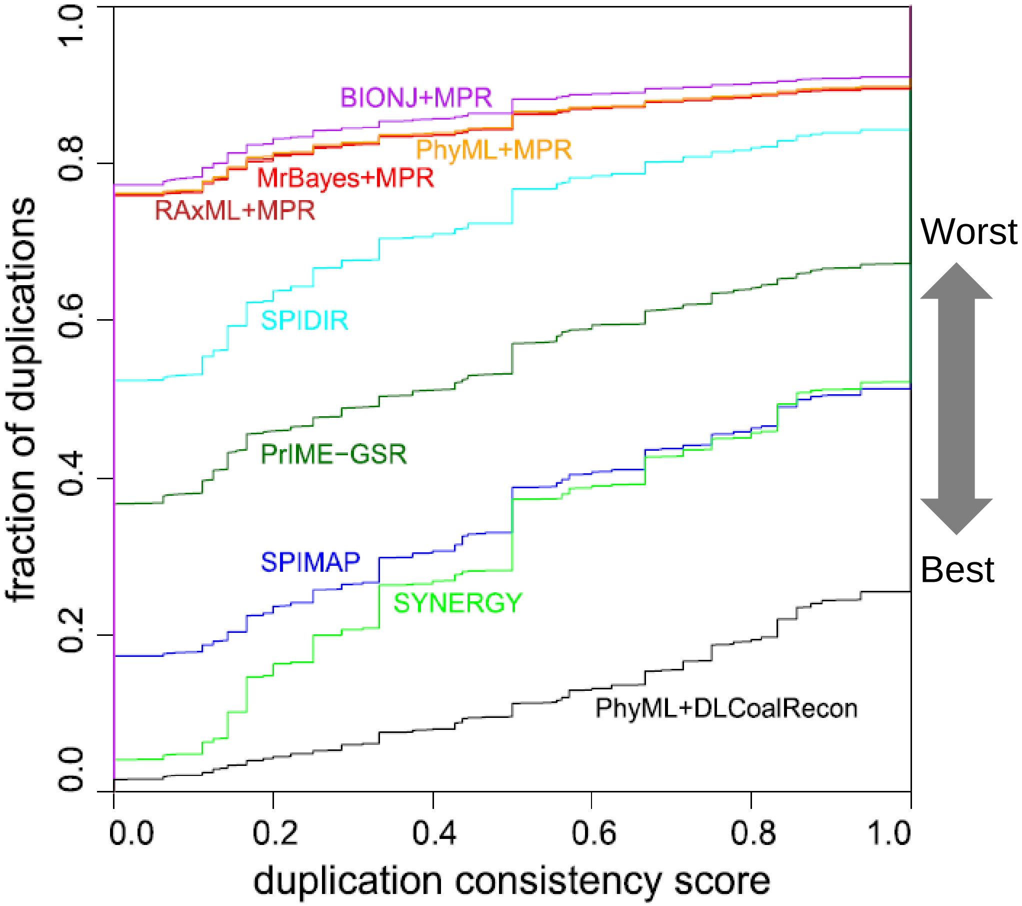
Bounded multispecies coalescent



- Condition t(r)<t*
- Use time of MRCA of MC [Efromovich & Kubatko 2009]

DLCoalRecon outperforms on 16 fungi genomes

(3) Improved duplication consistency



(4) Improved gene conversion recovery