Comparative Genomics of Drosophila

Lior Pachter
Department of Mathematics & Computer Science
UC Berkeley

(on sabbatical at Oxford 2006-2007)

Joint work with A. Caspi, S. Chatterji, C. Dewey, J. Hein, D. Levy and R. Satija

Question: which tree?

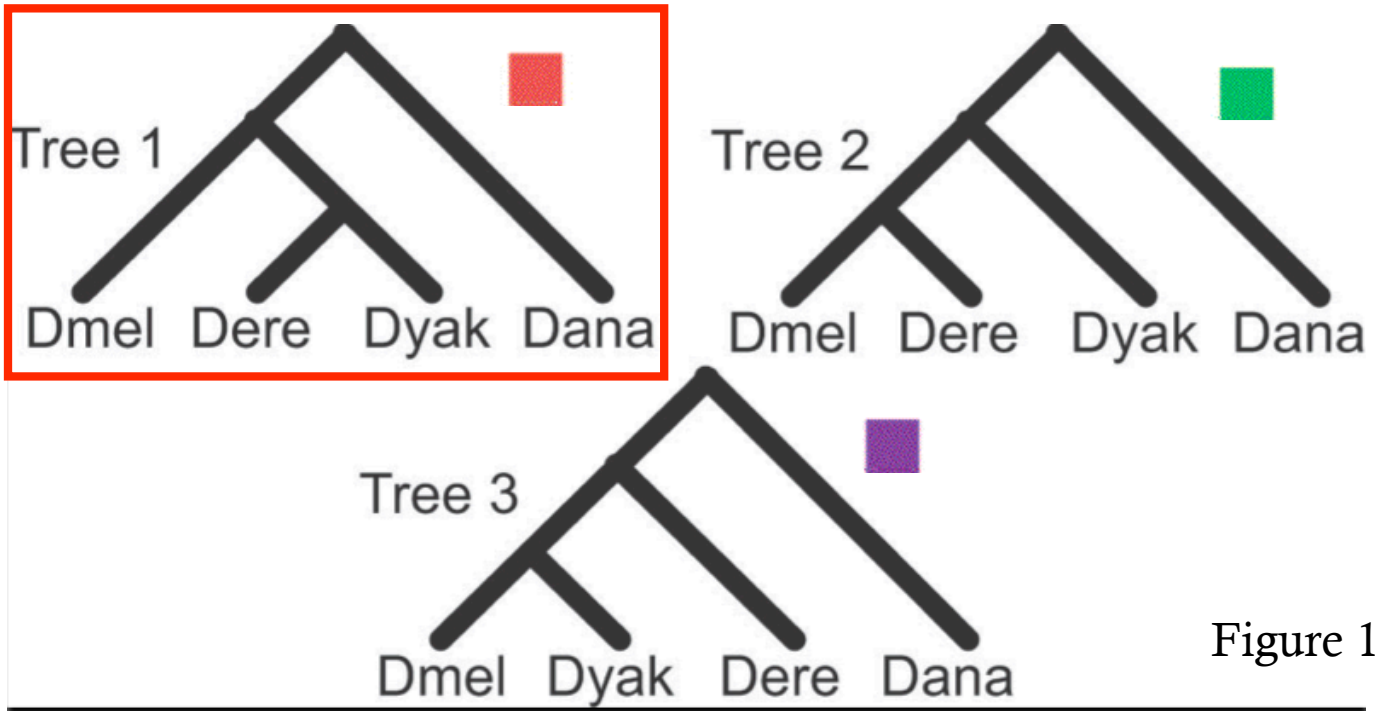
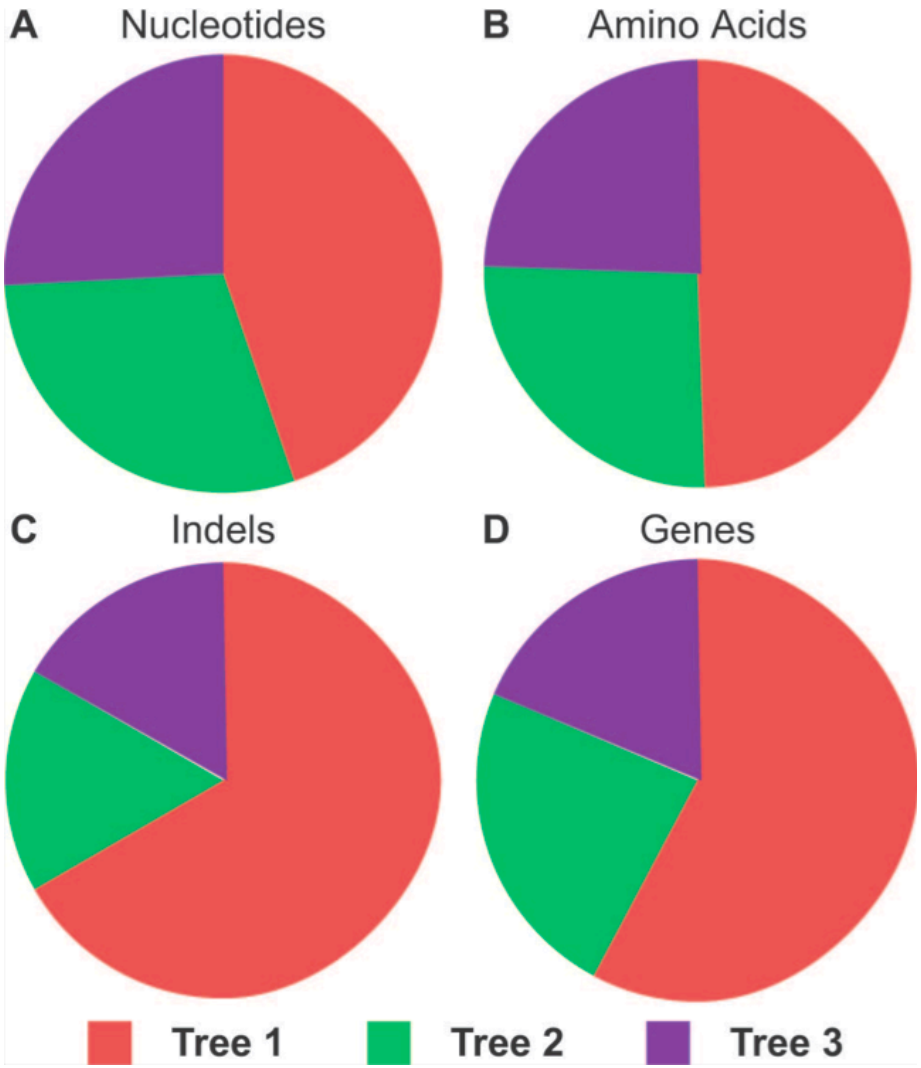


Figure 1

Pollard, D. A., Iyer, V. N., Moses, A. M., Eisen, M. B., Whole Genome Phylogeny of the Drosophila melanogaster Species Subgroup: Widespread Discordance with Species Tree & Evidence for Incomplete Lineage Sorting. PloS Genetics, In Press.

The evidence



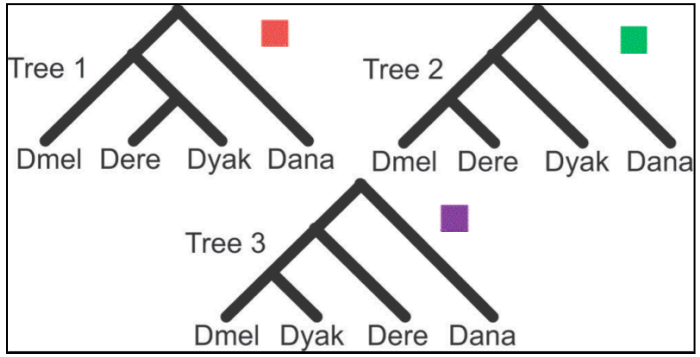


Figure 2

Incomplete lineage sorting?

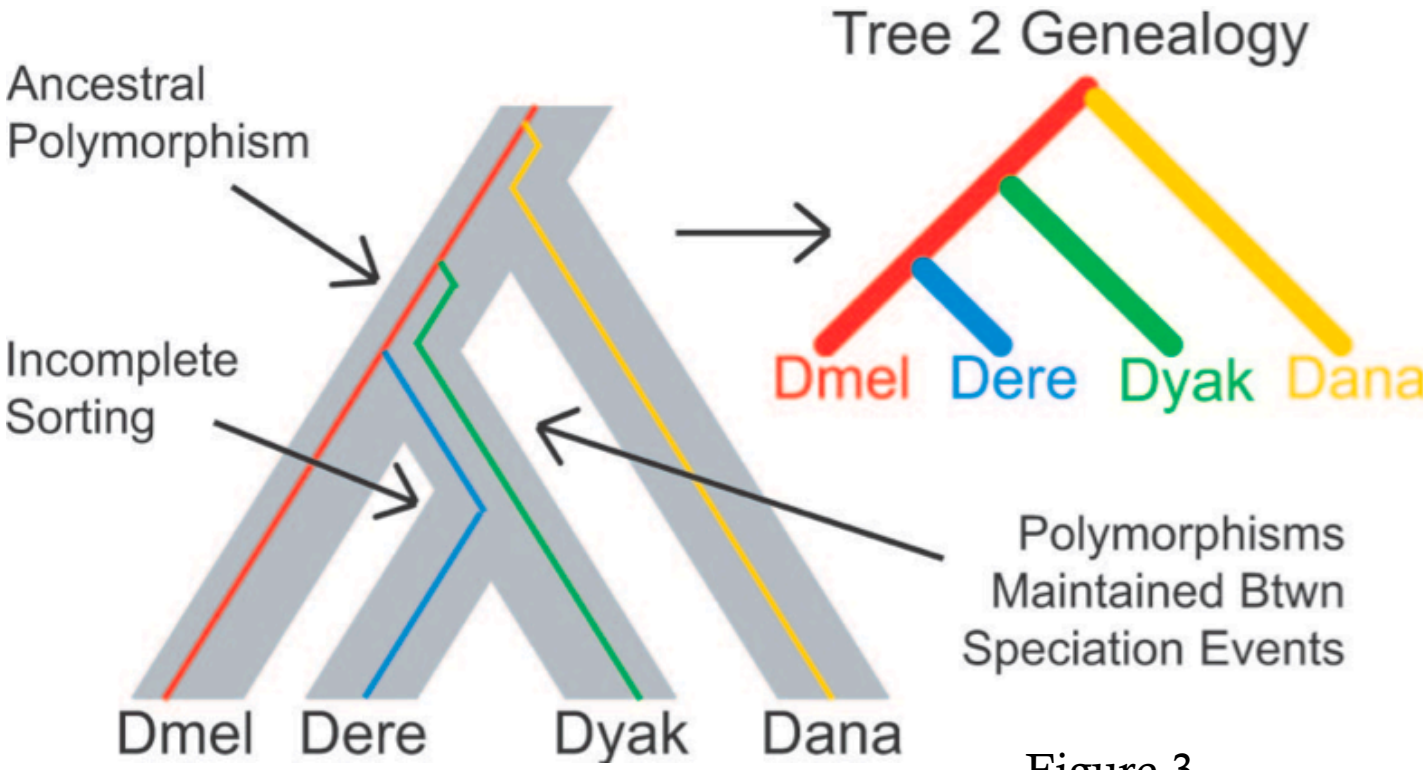
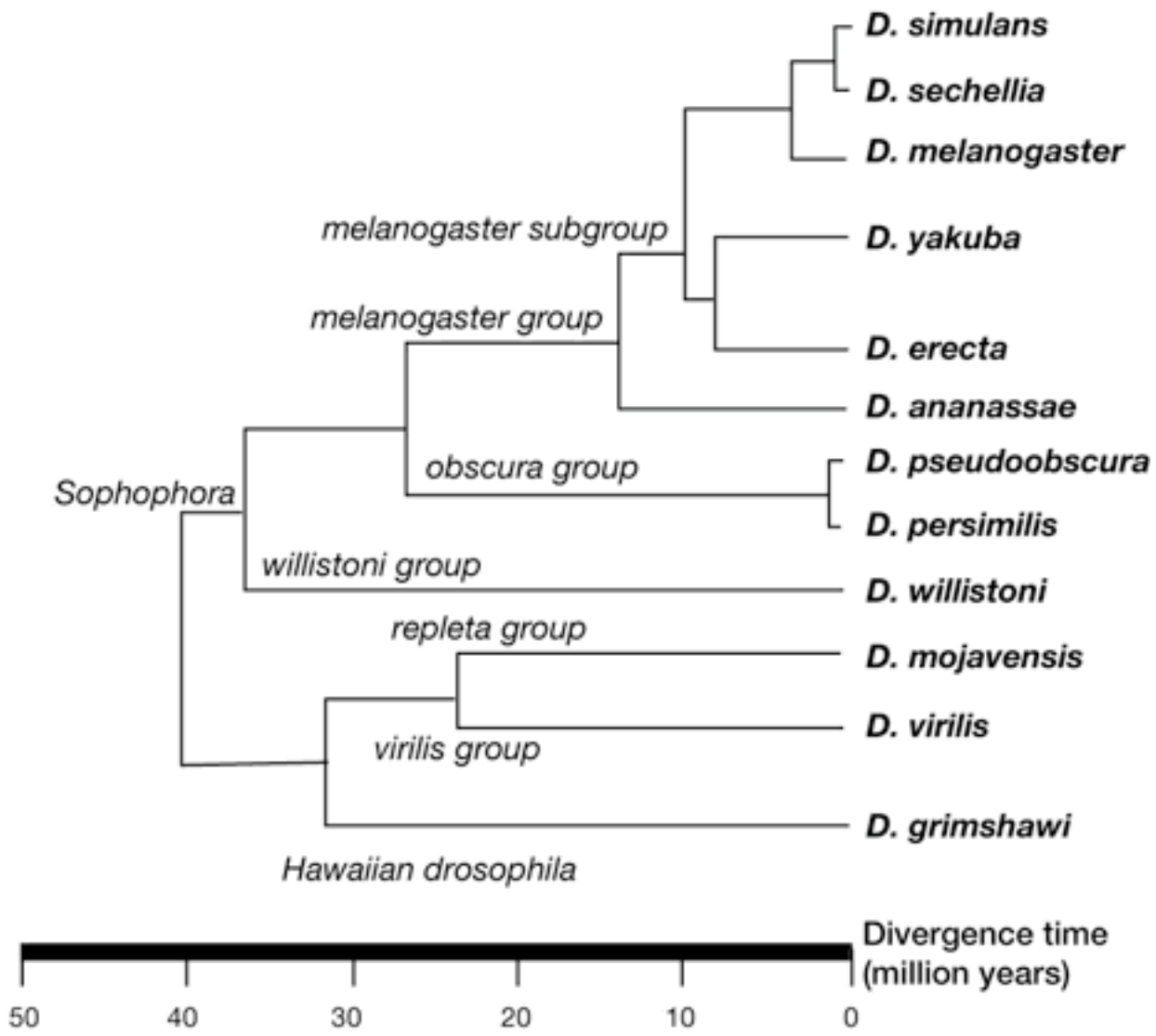


Figure 3

The power of comparative genomics: The AAA Project



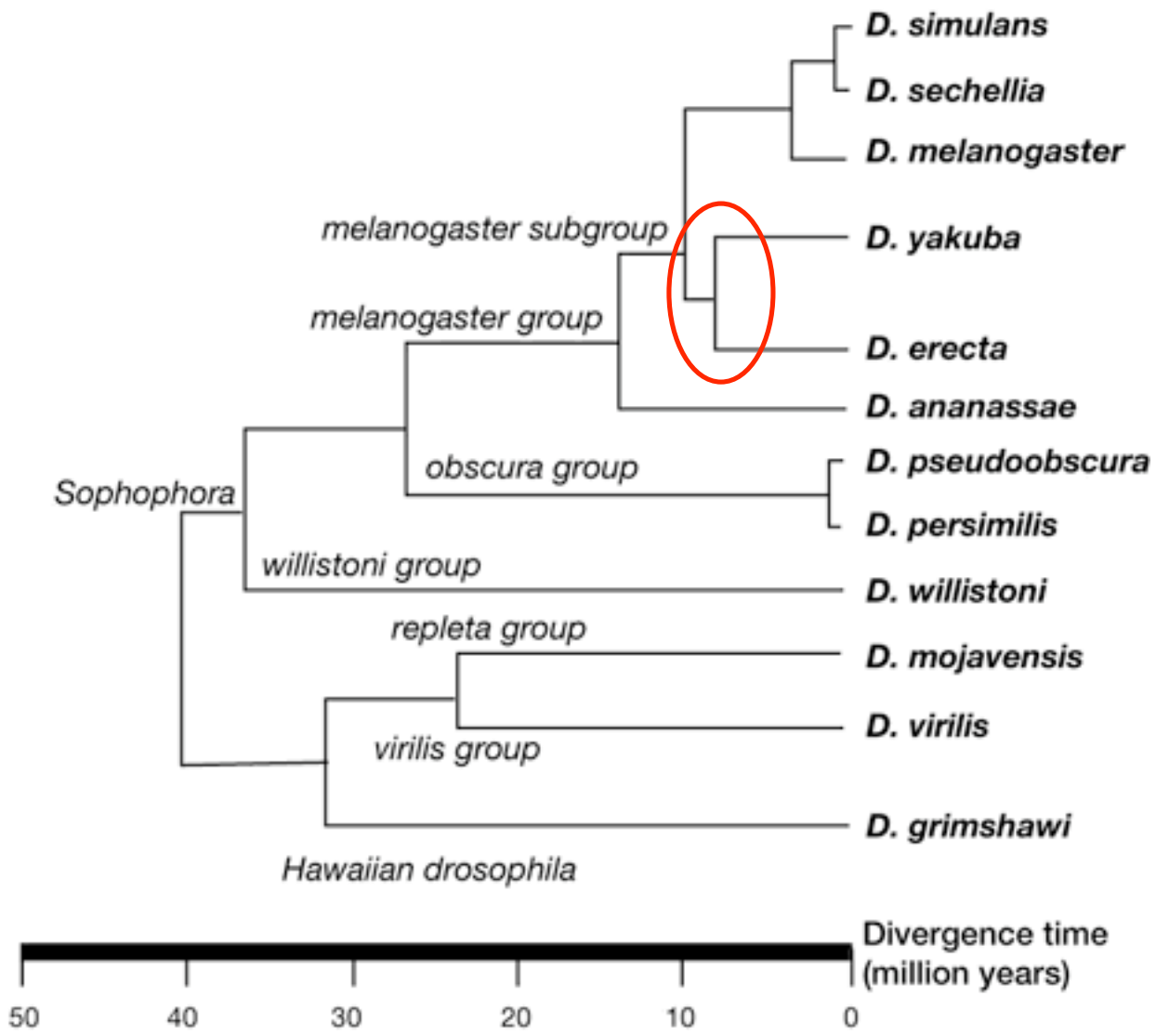
The genomes

Drosophila species genomes, with sizes, abbreviation, Dmel genes found, and sequencing centers for CAF1 assemblies. Total assembly size (T), euchromatin (E) size, and Dmel (Me) DNA homology sizes, in Megabases, total (Nt) and euchromatic scaffold count (Ne), number of inverts to Dmel (In), and percentages of Dmel genes found in total (%Gt) and euchromatic (%Ge) scaffolds. The list of major euchromatic scaffolds matched to Dmel euchromatin are provided at DroSpeGe/maps/muller-elements/.

Abbr.	Species	T	Nt	E	Ne	Me	In	%Gt	%Ge	Sequencing Center
Dmel	*Drosophila melanogaster*	169	14	120	6	120	0	100.0	100.0	Berkeley Dros. Genome Prj. & Celera
Dsim	*D. simulans*	143	17	117	7	117	51	99.2	90.4	Genome Seq. Ctr., Washington U.
Dsec	*D. sechellia*	167	14730	100	20	114	33	99.7	85.8	Broad Institute
Dyak	*D. yakuba*	169	20	127	7	116	67	99.5	96.8	Genome Seq. Ctr, Washington U.
Dere	*D. erecta*	153	5124	125	7	116	70	99.4	98.6	Agencourt Bioscience Corp.
Dana	*D. ananassae*	231	13749	120	25	113	258	97.6	93.7	Agencourt Bioscience Corp.
Dper	*D. persimilis*	188	12838	102	25	110	276	95.7	74.2	Broad Institute
Dpse	*D. pseudoobscura*	153	4896	125	12	111	418	96.1	94.9	H.G.S.C., Baylor Coll. Med.
Dwil	*D. willistoni*	237	14927	140	24	103	397	95.4	89.0	J. Craig Venter Institute
Dmoj	*D. mojavensis*	194	6841	150	9	102	691	94.6	92.7	Agencourt Bioscience Corp.
Dvir	*D. virilis*	206	13530	140	15	104	515	94.8	90.3	Agencourt Bioscience Corp.
Dgri	*D. grimshawi*	200	17440	140	12	102	541	94.1	87.0	Agencourt Bioscience Corp.

Compiled by Don Gilbert, Indiana University

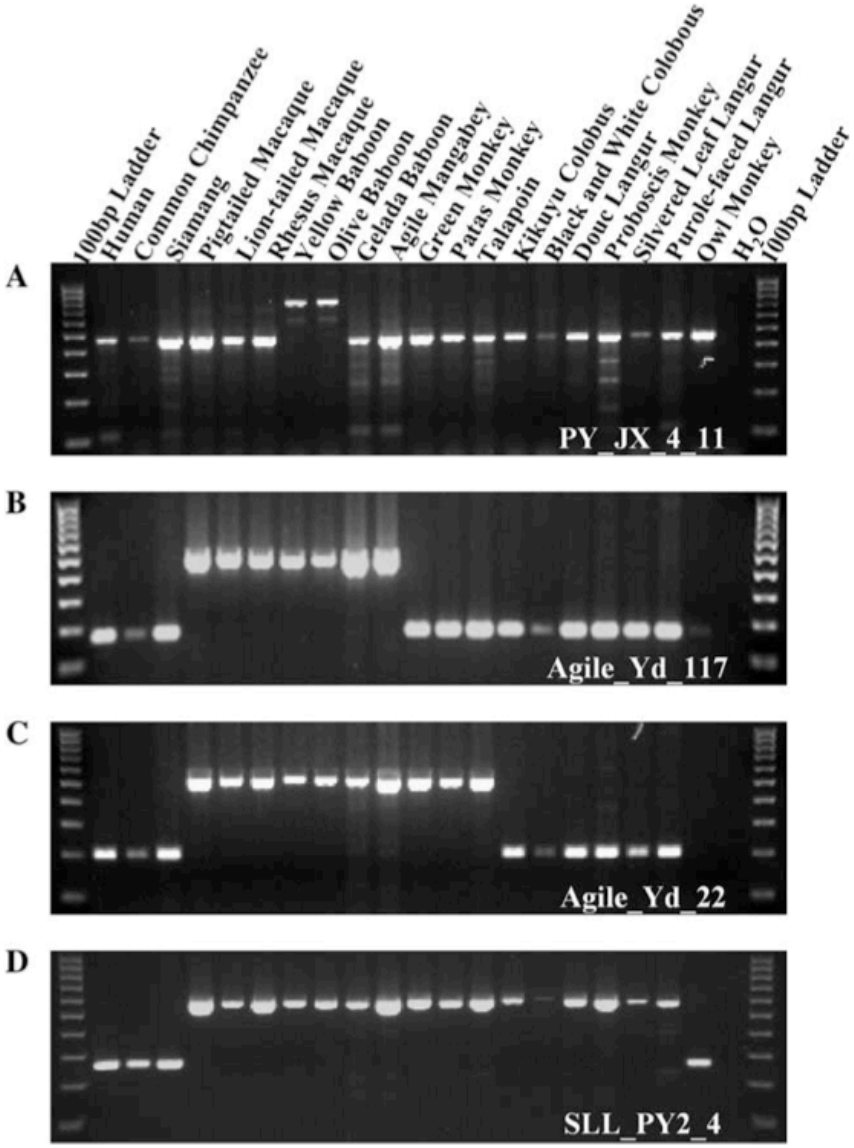
Phylogenetic analysis using Transposable Elements



Phylogenetic analysis using Transposable Elements?

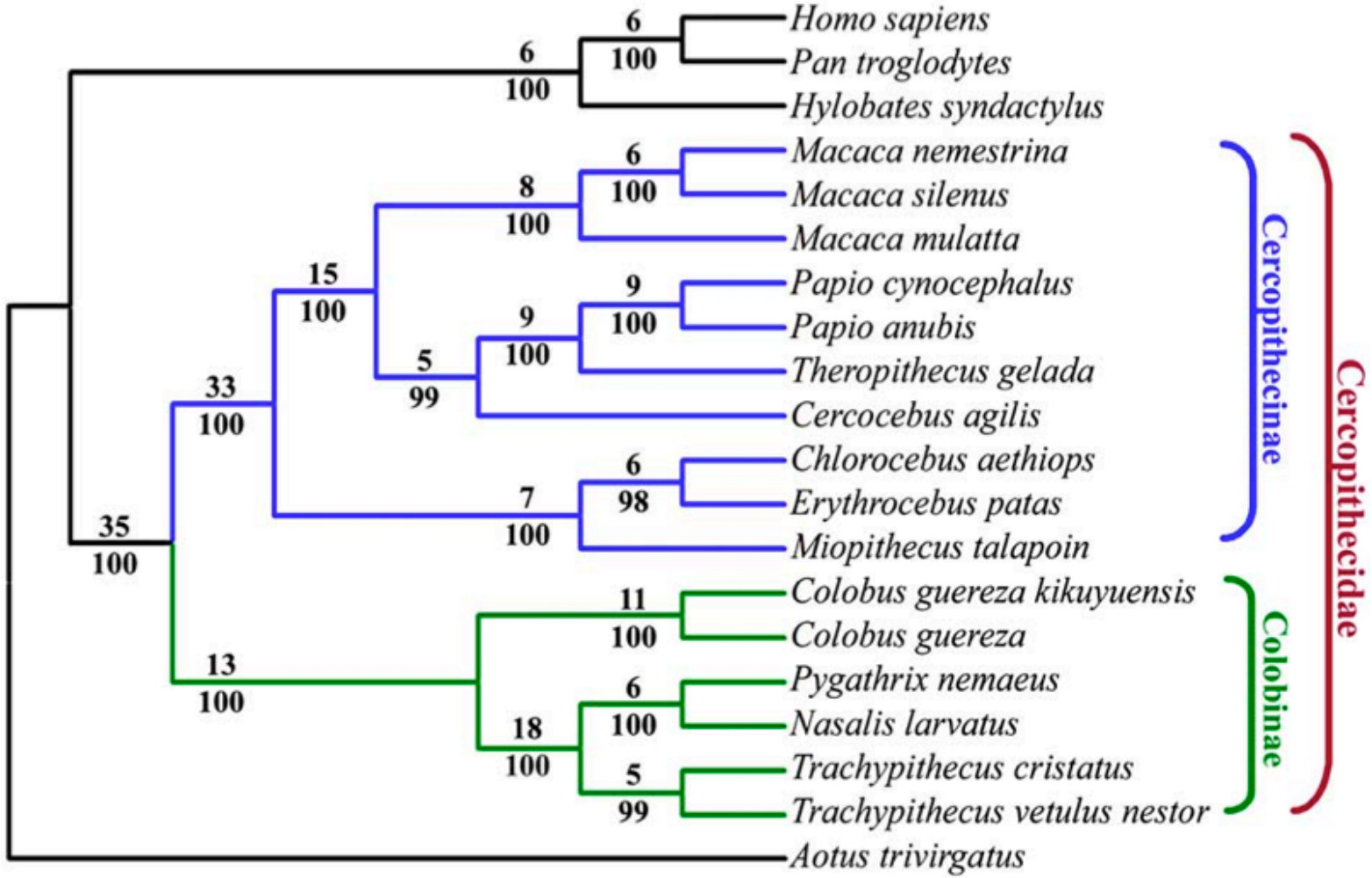
- Many methods for finding transposable elements:
  - BLASTER, ReAS, PILER, RepeatMasker,…
- Difficult to establish homology
- Existing phylogenetic methods mostly based on the analysis of *families.*

Phylogenetic analysis using Transposable Elements Previous Work



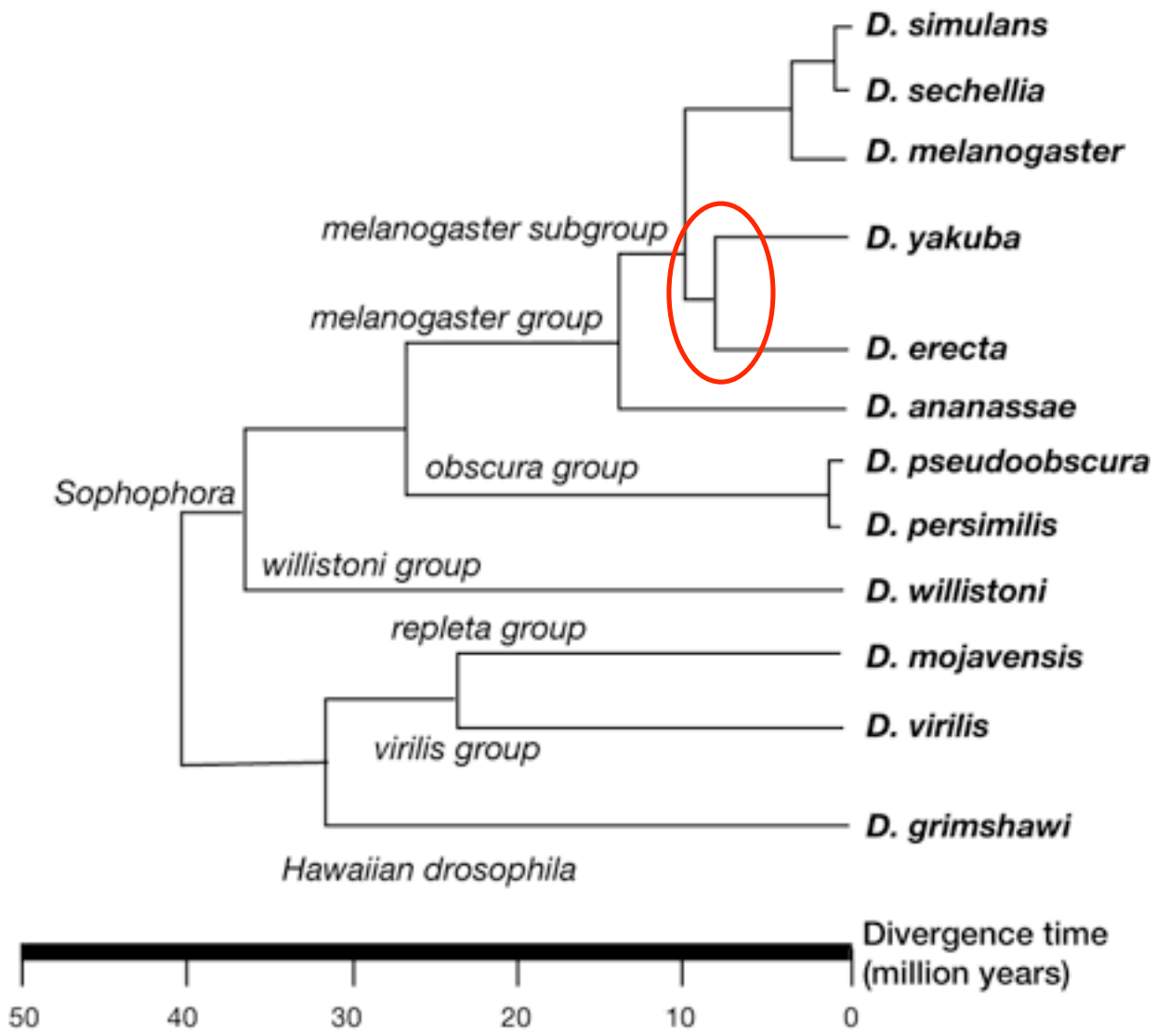
Xing et al., Molecular Phylogenetics nad Evolution, 2005.

Phylogenetic analysis using Transposable Elements
Previous Work



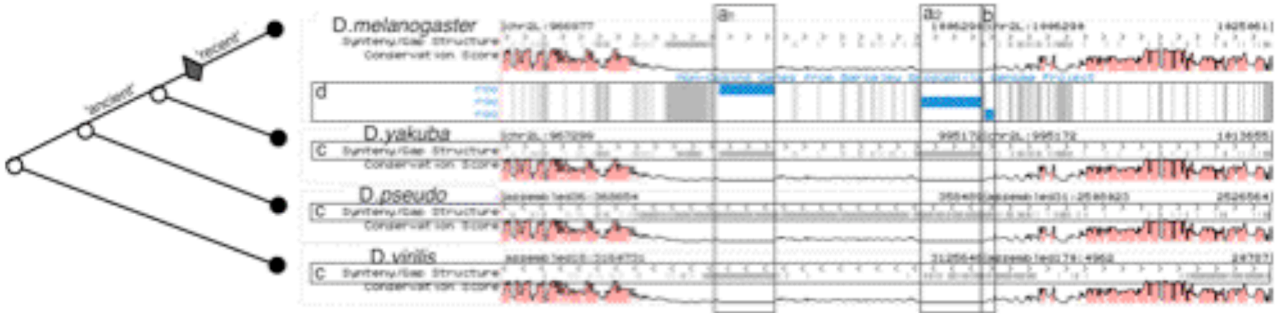
Xing et al., Molecular Phylogenetics nad Evolution, 2005.

Phylogenetic analysis using Transposable Elements **in silico**



Obtaining the data

- **Transposable Element Annotation**



• A. Caspi and L. Pachter, Identification of transposable elements using multiple alignments of related genomes, GR 16 (2006).

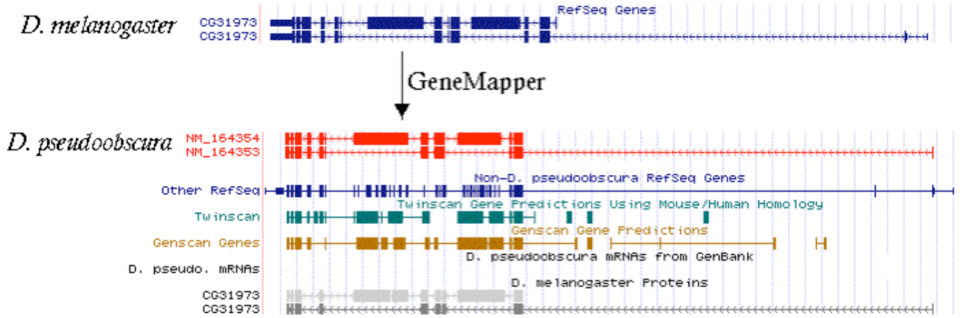
- **Multiple Sequence Alignment**

```
GTCGCTCAACCAGCATTTGCAAAAGTCGCAGAACTTGCGCTCATTGGATTTCCAGTACTC
GTCGCTCAGCCAGCATTTGCAGAAGTCGCAGAACTTGCGCTCGTTTGATTTCCAGTACTC
GTCGCTTAACCAGCATTTACAGAAATCGCAATACTTGCGTTCATTGGATTTCCAGTACTC
GTCGCTCAGCCAGCACTTGCAGAAGTCGCAGTACTTGCGCTCGTTTGATTTCCAGAATTC
GTCGCTCAGCCAGCATTTGCAGAAGTCGCAGAACTTCCGCTCGTTTGACTTCCAGTACTC
****** * ****** ** ** ** *****  **** ** ** ** ** ****** * **
```

• C. Dewey and L. Pachter, Whole Genome Mapping, in preparation.

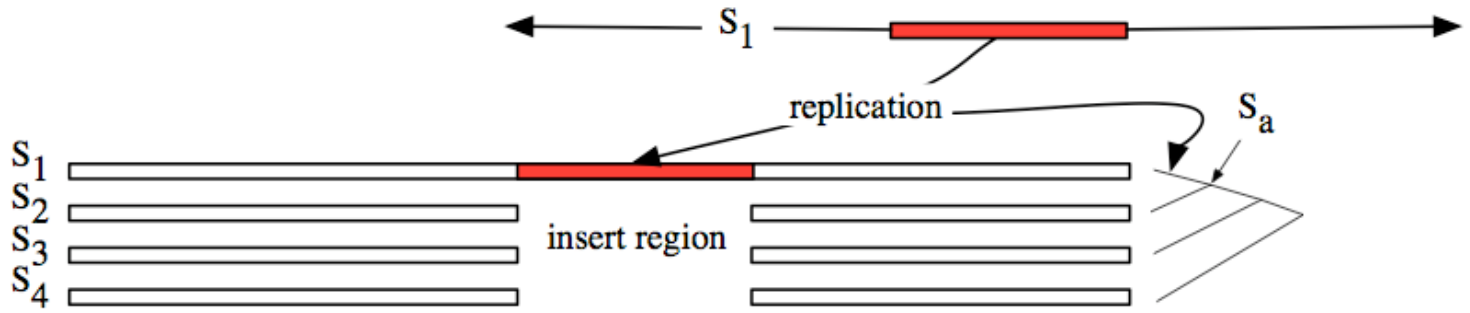
•N. Bray and L. Pachter, MAVID: Constrained ancestral alignment of multiple sequences, GR 14 (2004), p 693--699.

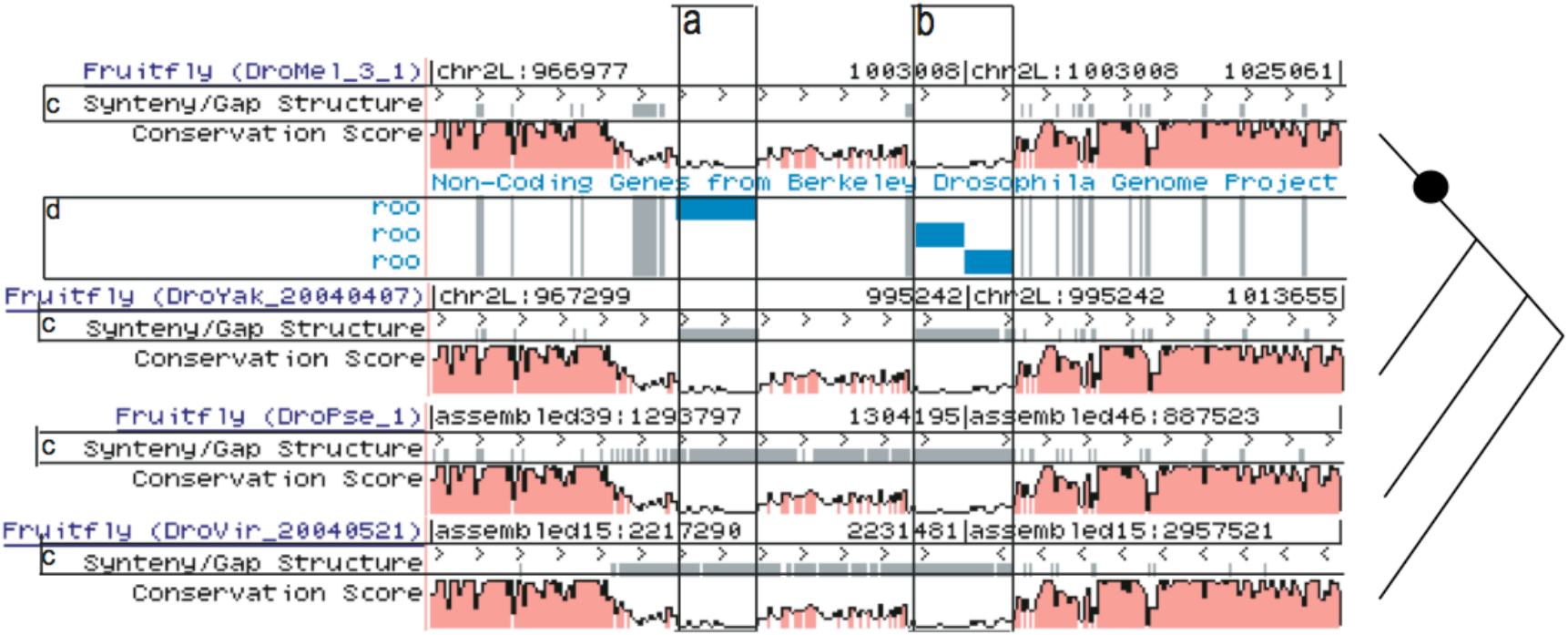
- **Gene Finding**



• S. Chatterji and L. Pachter, GeneMapper: Reference based annotation with GeneMapper, Genome Biology (2006).

Novel approach to finding TEs

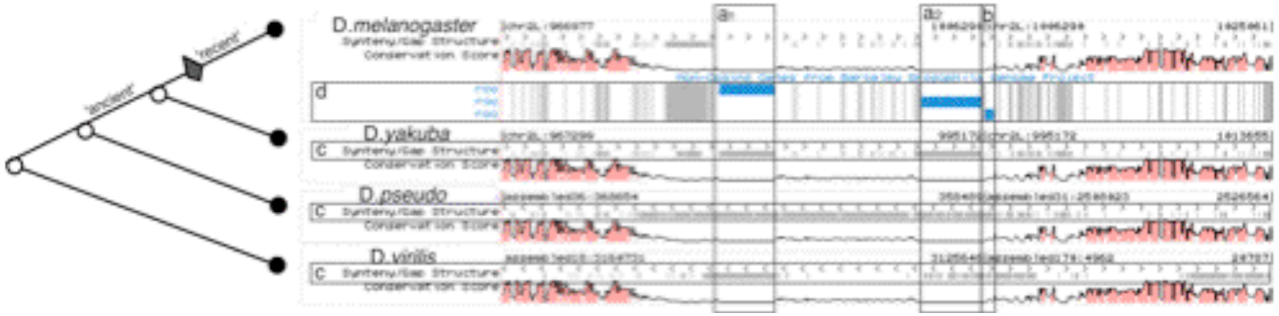




Caspi and Pachter, Genome Research 2006.

Obtaining the data

- **Transposable Element Annotation**



• A. Caspi and L. Pachter, Identification of transposable elements using multiple alignments of related genomes, GR 16 (2006).

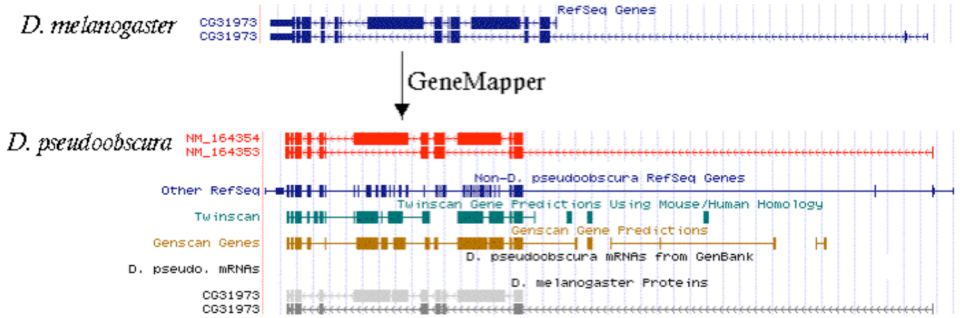
- **Multiple Sequence Alignment**

```
GTCGCTCAACCAGCATTTGCAAAAGTCGCAGAACTTGCGCTCATTGGATTTCCAGTACTC
GTCGCTCAGCCAGCATTTGCAGAAGTCGCAGAACTTGCGCTCGTTTGATTTCCAGTACTC
GTCGCTTAACCAGCATTTACAGAAATCGCAATACTTGCGTTCATTGGATTTCCAGTACTC
GTCGCTCAGCCAGCACTTGCAGAAGTCGCAGTACTTGCGCTCGTTTGATTTCCAGAATTC
GTCGCTCAGCCAGCATTTGCAGAAGTCGCAGAACTTCCGCTCGTTTGACTTCCAGTACTC
****** * ****** ** ** ** *****  **** ** ** ** ** ****** * **
```

• C. Dewey and L. Pachter, Whole Genome Mapping, in preparation.

•N. Bray and L. Pachter, MAVID: Constrained ancestral alignment of multiple sequences, GR 14 (2004), p 693--699.

- **Gene Finding**



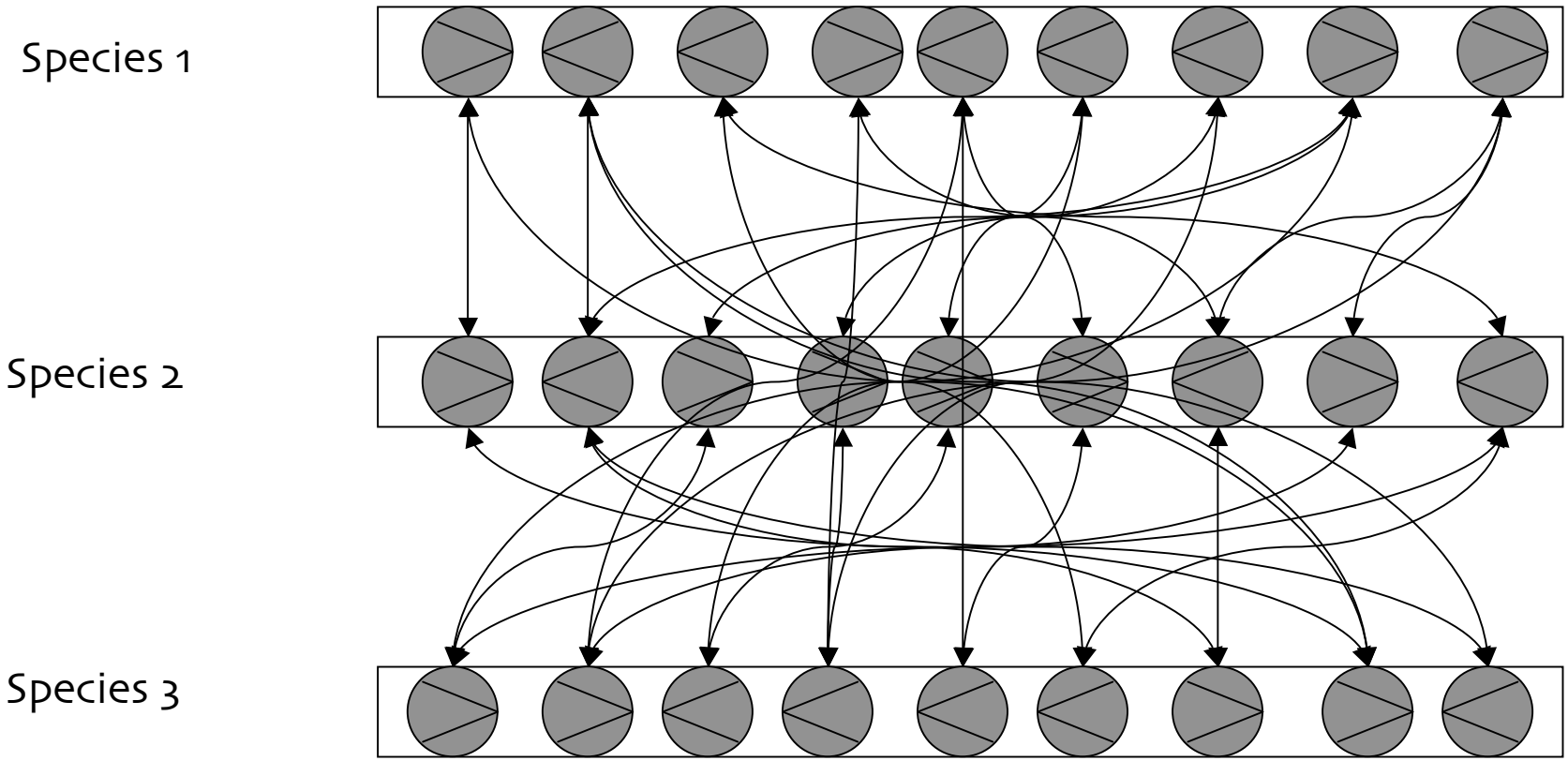
• S. Chatterji and L. Pachter, GeneMapper: Reference based annotation with GeneMapper, Genome Biology (2006).

Mercator: Multiple whole-genome orthology map construction

- Uses coding exons (predicted and experimental) as anchors
- Computes all pairwise similarities between exons using protein sequence
- Finds highest-scoring exon *cliques*: a set of exons, each exon in a different genome and having hits to all of the other exons
- Forms *runs* of cliques that are adjacent and consistent in every genome
- Gives highest priority to large cliques (e.g. one that is in all genomes). Smaller cliques that are inconsistent with adjacent large cliques are filtered out.
- Outputs runs of cliques as *orthologous segments*
- Set of orthologous segments forms an *orthology map*

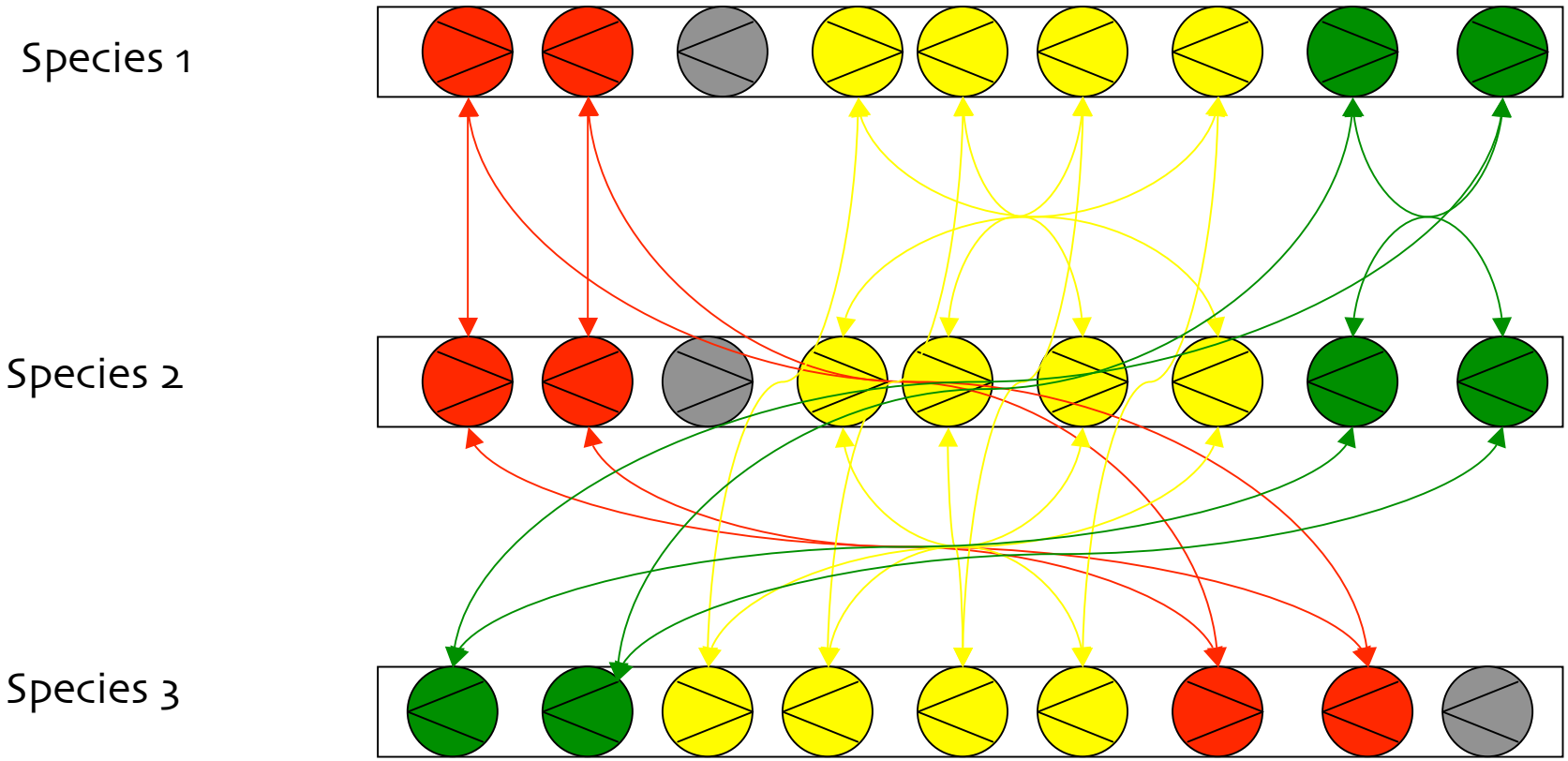
Mercator: Input

anchors (circles) and similarity hits (arrows)



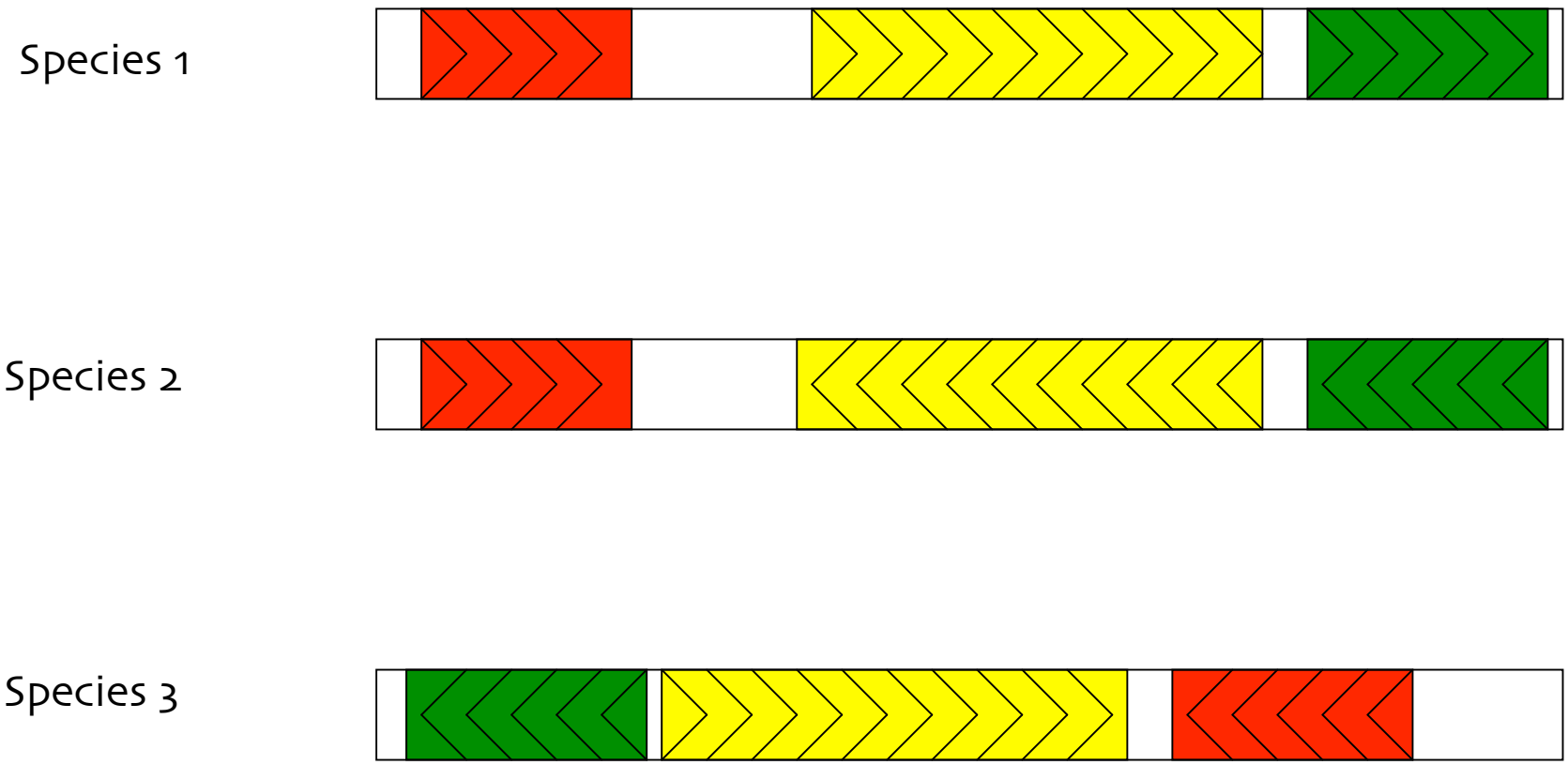
Mercator: Clique/Run Finding

Colored anchors are found to be part of a run

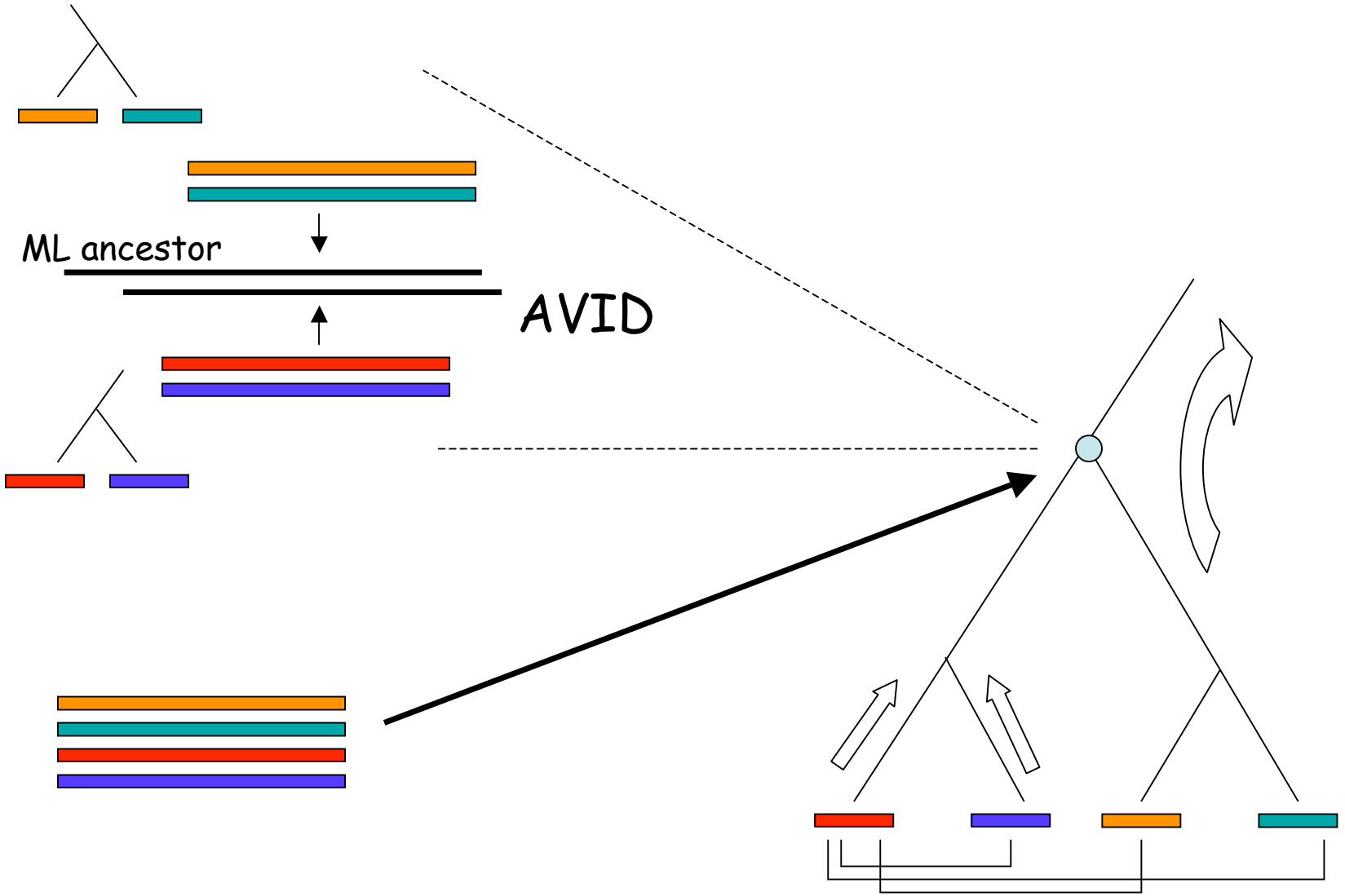


Mercator: Orthologous Segments

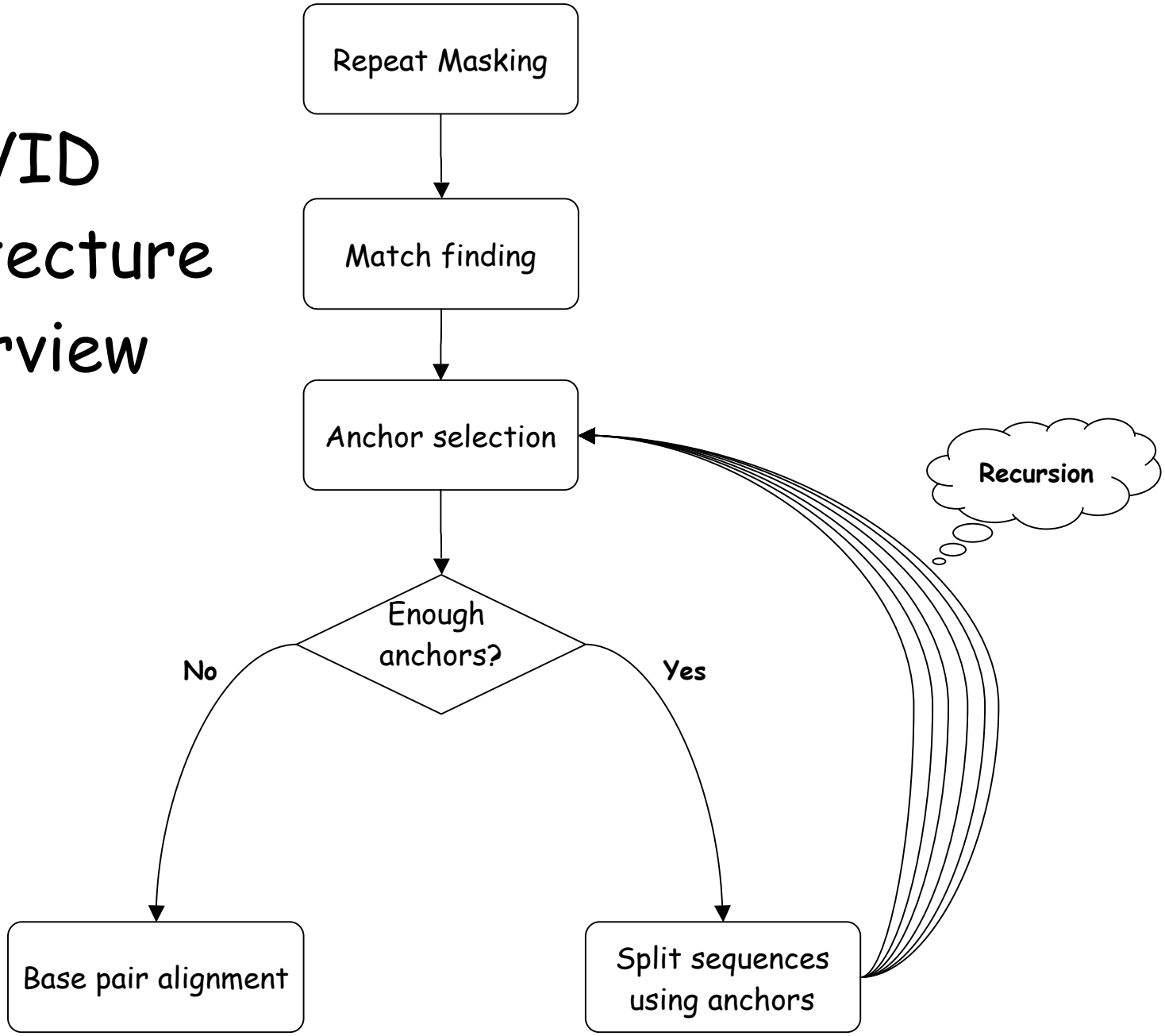
Runs define boundaries of orthologous segments



MAVID architecture overview

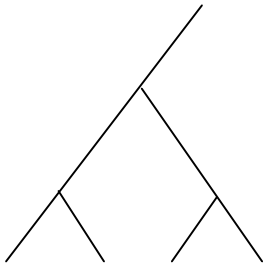


AVID architecture overview

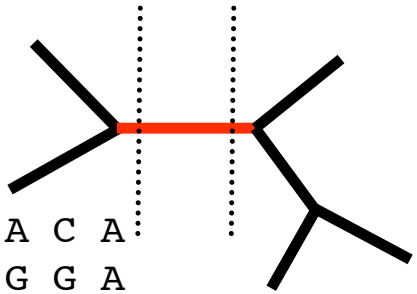


MAVID details

"EM" progressive tree



Iterative refinement



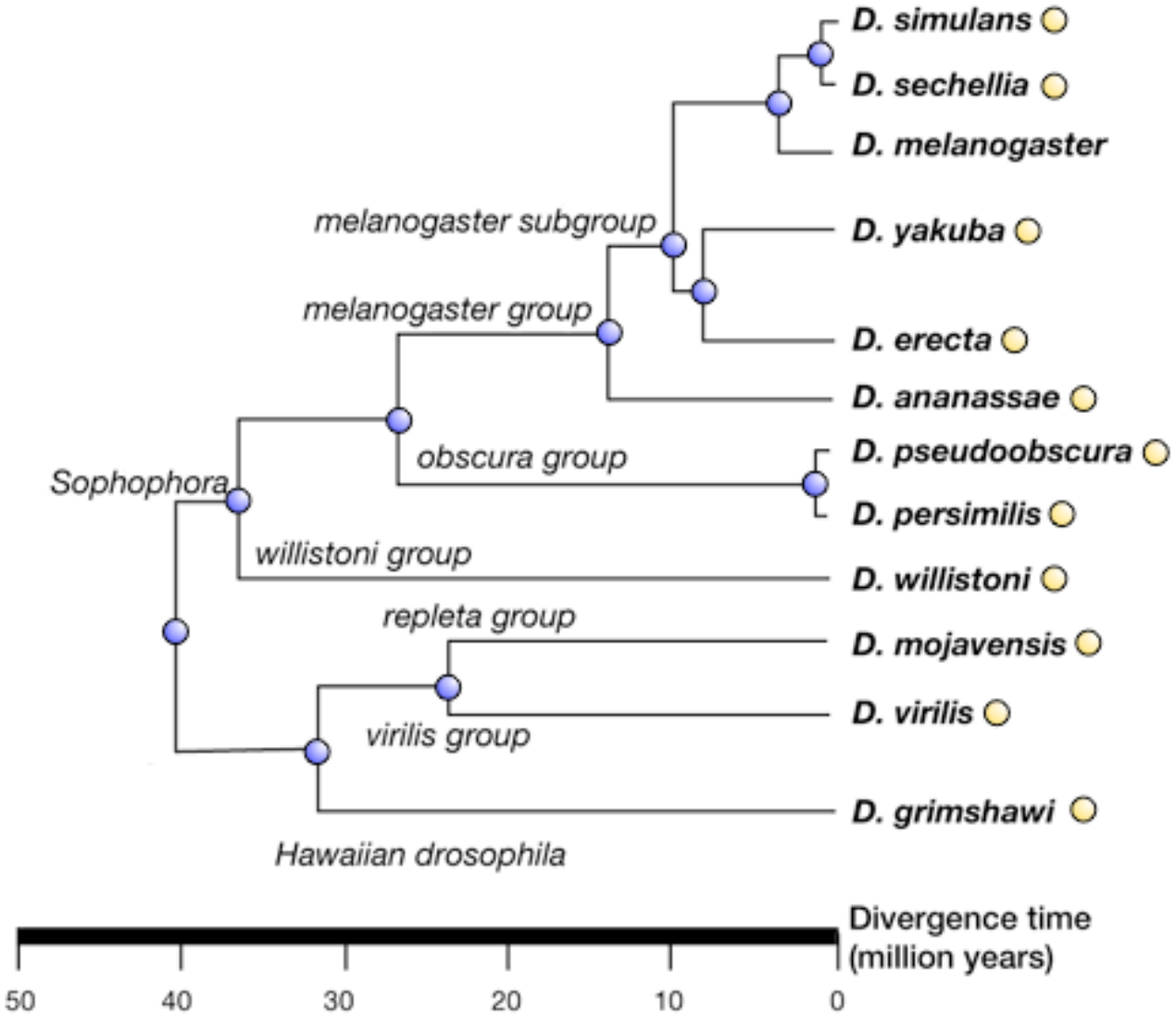
```
A C G G A A C A        A C G G A A C A
A G A - A G G A        A G A - A G G A
G C A - A A A T        G C A - A A A T
A T G G - A A G        A T G G A A G -
A G T G - A G A        A G T G A G A -
```

	A	C	G	T	--
A	−0.008	0.001	0.004	0.001	0.001
C	0.002	−0.009	0.001	0.005	0.001
G	0.005	0.001	−0.009	0.002	0.001
T	0.001	0.005	0.001	−0.009	0.001
--	0.019	0.019	0.018	0.020	−0.076

Ancestral sequences are estimated based on an evolutionary model. The Q-matrix, estimated from exon alignments is shown above.

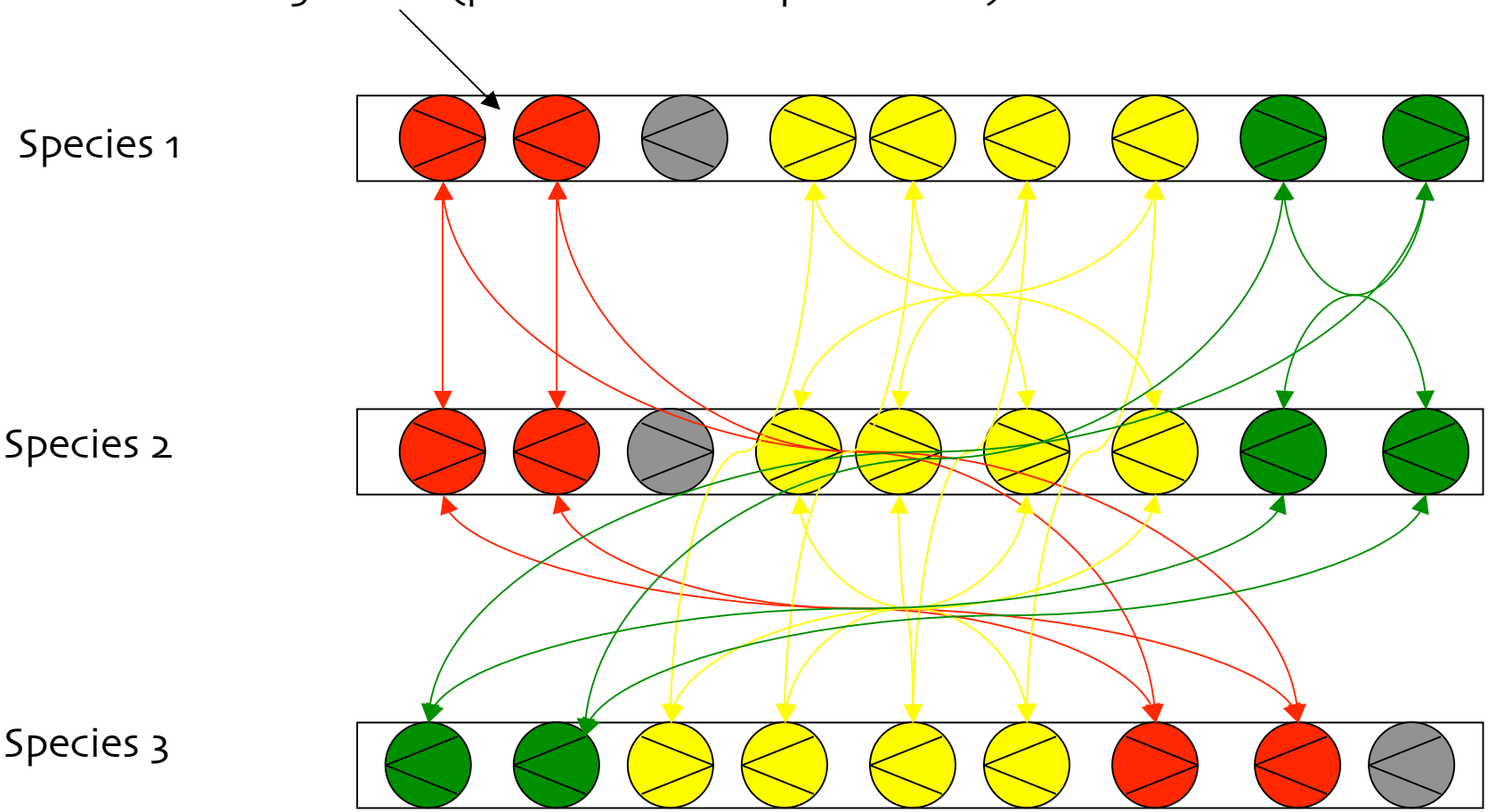
MAVID/MERCATOR alignments

http://www.biostat.wisc.edu/~cdewey/fly_CAF1/



Mercator: Clique/Run Finding

- Uses coding **exons** (predicted and experimental) as anchors



GeneMapper

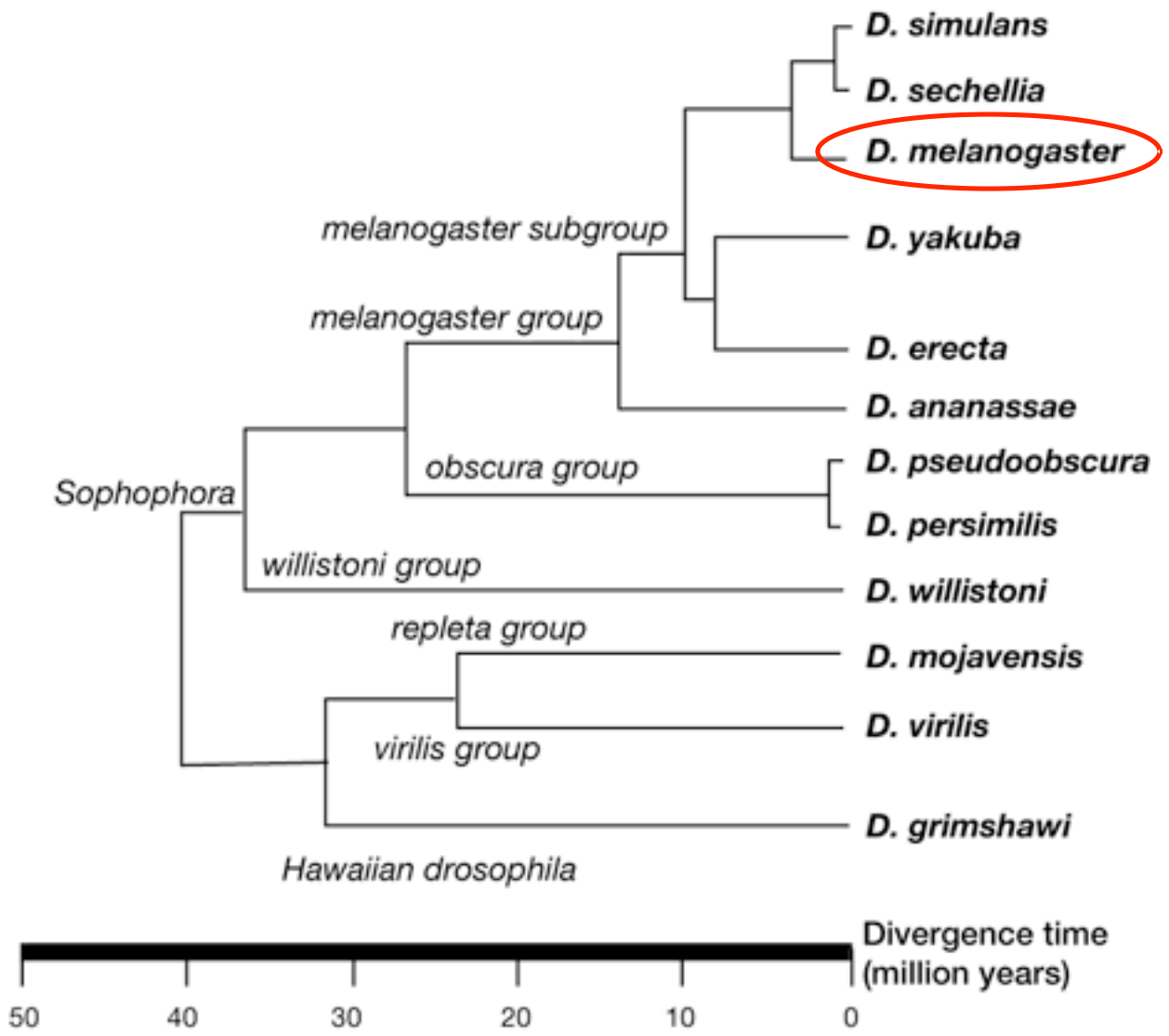
- Transfer annotations from reference to target species
  - Map each exon and then join the exon predictions together (look for exon splitting/fusion later)
  - Makes DP possible (exons much shorter than introns)
  - For multiple species, use profiles to improve accuracy

Chatterji and Pachter, Genome Biology, 2006.

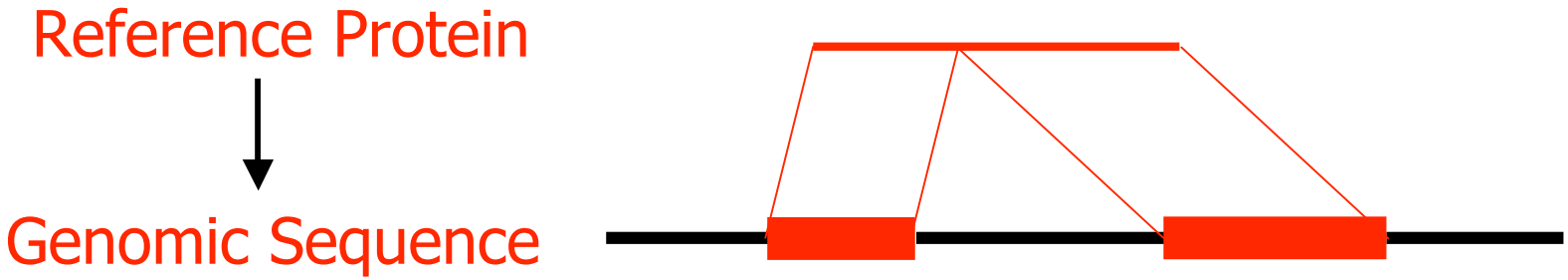
Species	ESTs	mRNAs	RefSeq
D. melanogaster	383407	19931	19967
D. simulans	5013	80	None
D. yakuba	11015	808	None
D. erecta	None	6	None
D. ananassae	None	11	None
D. pseudoobscura	35042	40	None
D. mojavensis	361	2	None
D. virilis	663	41	None
D. grimshawi	None	None	None

Source : UCSC browser

The reference genome



Protein Alignment Approach



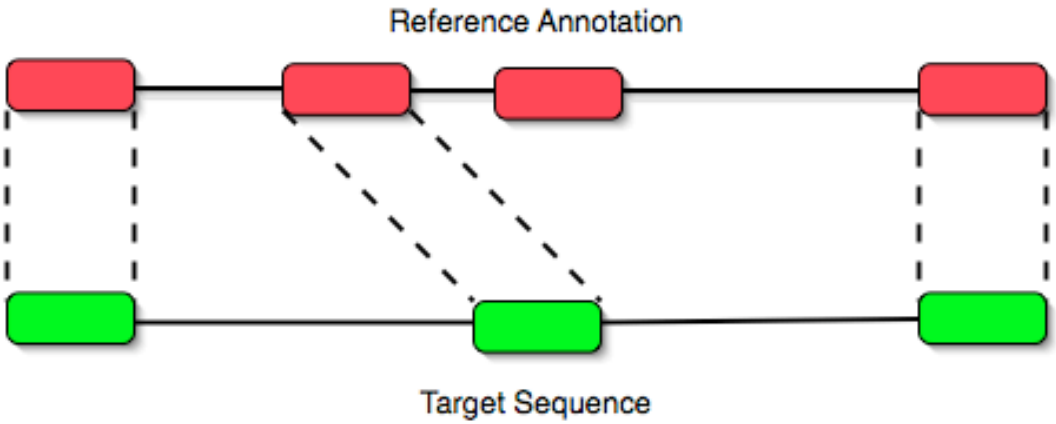
- **Procrustes** [Gelfand at al. 1996]
- **GeneWise** [Birney at al. 2004]
  - Integral part of the ENSEMBL gene annotation pipeline.
- Not aware of exon/intron boundaries.
- Accuracy decreases when sequence identity is low.

Similarity Based Approach

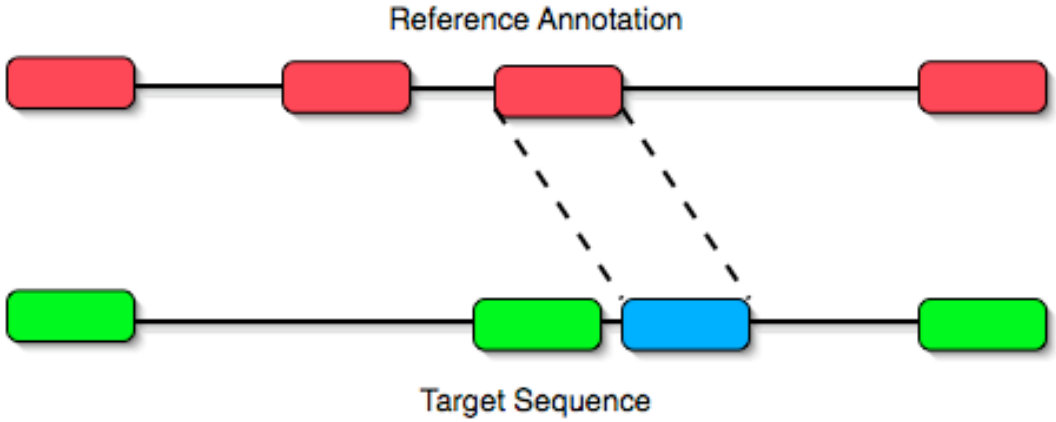
- **Projector** [Meyer and Durbin 2004]
  - Predicts the global gene structure using a pair HMM.
  - Uses heuristics to decrease the search space.
- **GeneMapper**
  - Uses a bottom up algorithm for predicting the gene structure.
- Not suitable if the exon/intron structure of the gene has diverged a lot.

The GeneMapper Algorithm

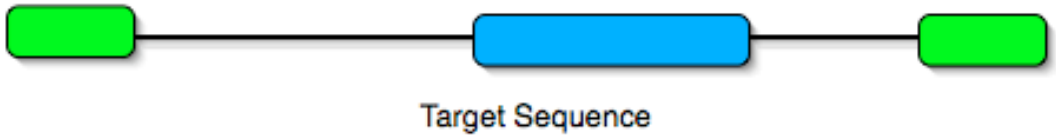
- Bottom Up Algorithm
  - Predict the ortholog of each reference exon in the target sequence.
  - Join exon predictions together to predict gene structure.
- Multiple Species GeneMapper
  - Uses all available information if the gene has to be mapped into multiple target species.
  - Uses a profile based approach to get more accurate annotations in evolutionary distant species.



(a) Step 1: Map the highly conserved exons

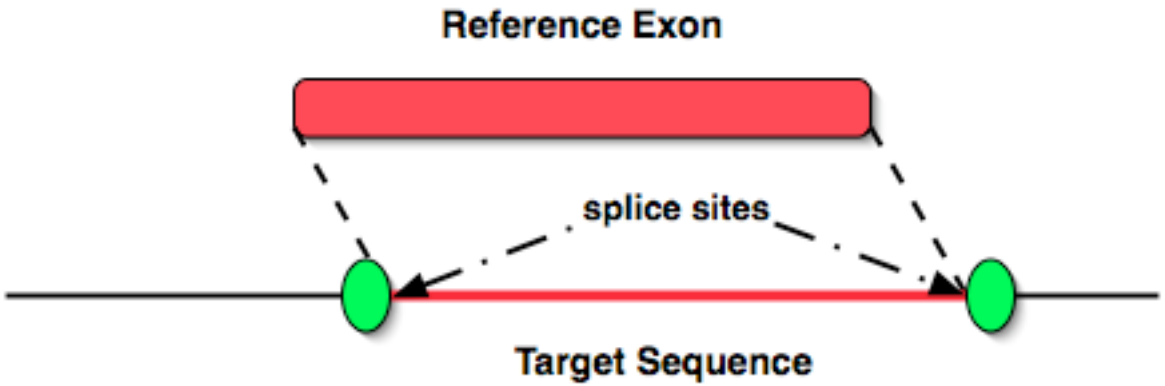


(b) Step 2: Use extrapolation to map less conserved exons



(c) Step 3: Find cases of exon splitting and exon fusion

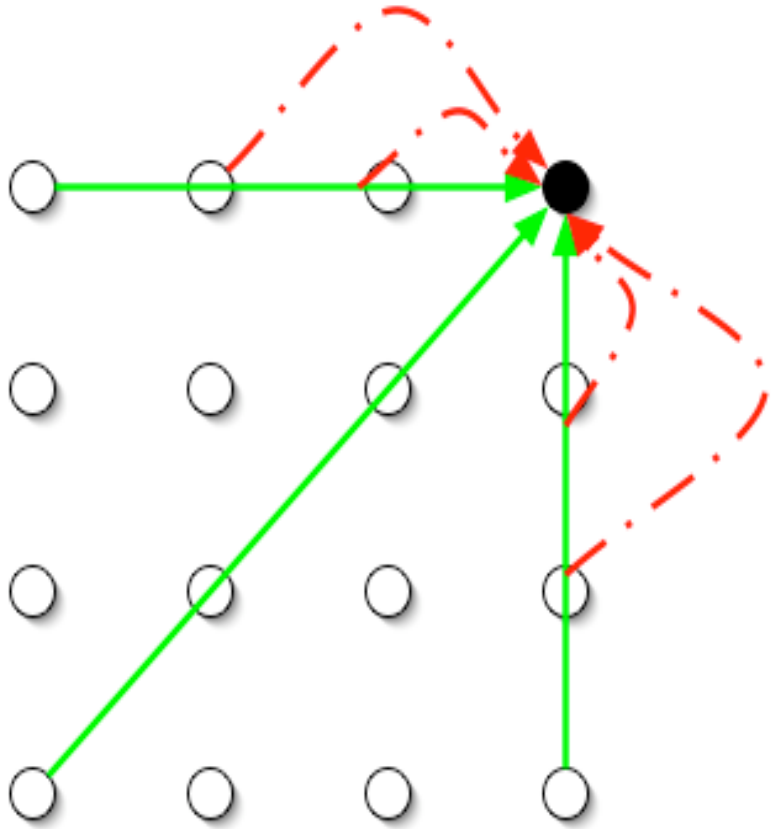
Mapping Exons Accurately



- Predicting the ortholog of a reference exon in the target sequence
  - Accurately model the evolution of exons.
  - Use ***StrataSplice*** to model splice sites.

Mapping Exons Accurately

- Use version of Smith Waterman algorithm.
  - Exact Optimization feasible.
- Green edges to model the evolution of codons.
  - Uses 64 * 64 **COD** distance matrices.
- Red edges to allow for frameshifts.



Multiple Species GeneMapper

Human : **AGT TTG** *GGA* GAA TCG *TCC* **TTT GGG** *AGC* **CAT** CTG **CCT GAC**

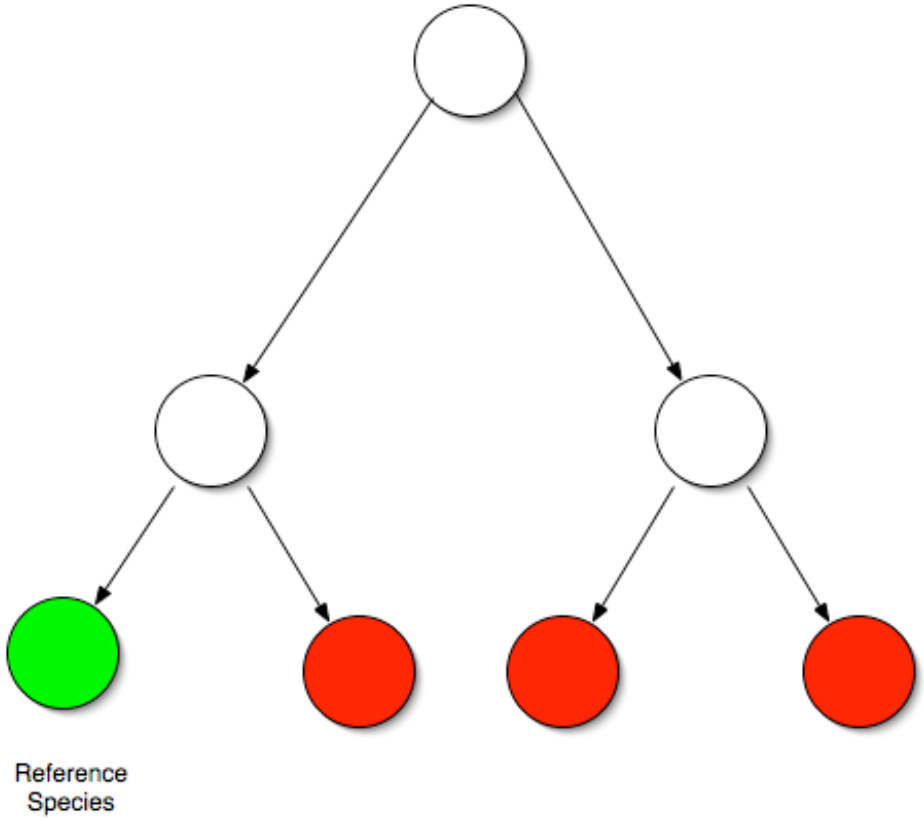
Chimp : **AGT TTG** *GGA* GAA TCG *TCC* **TTT GGG** *AGT* **CAT** CTG **CCT GAC**

Mouse : **AGT TTG** *GGT* GAC ___ *TCT* **TTT GGG** *AGC* **CAT** CCA **CCT GAC**

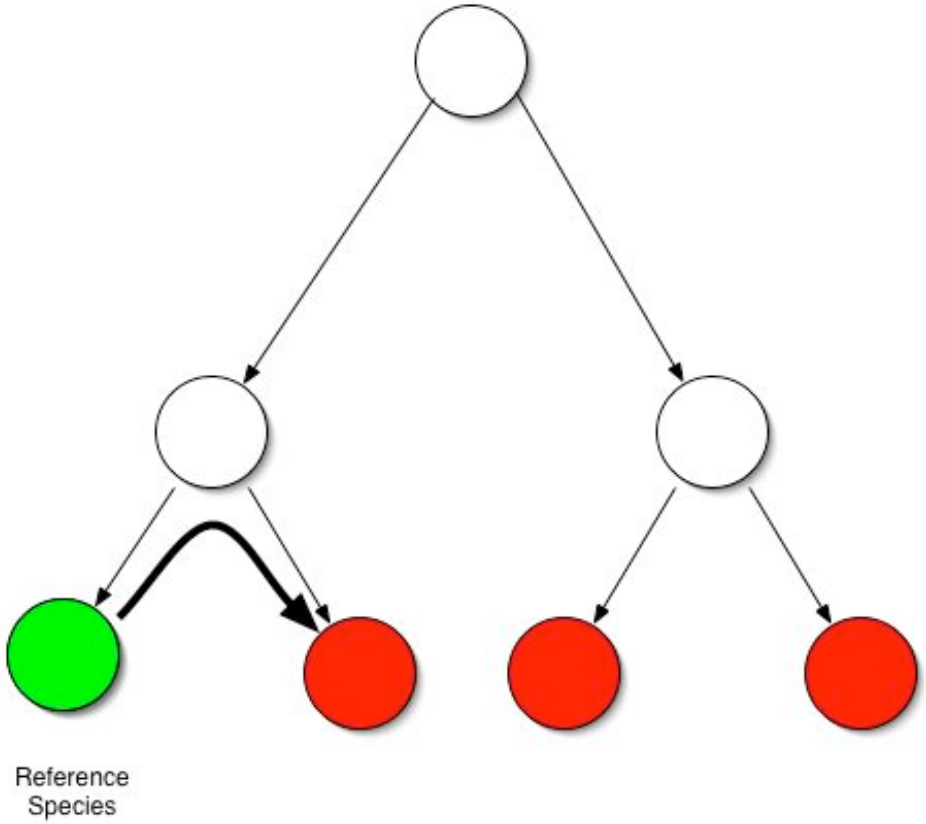
Rat : **AGT TTG** *GGA* GAC ___ *TCT* **TTT GGG** *AGC* **CAT** CCA **CCT GAC**

- Generates a **gene profile** of orthologous genes.
  - A more complete characterization than a single gene.
- Each column contains an alignment of orthologous codons.
- Special columns of width 1 are allowed to account for frameshifts and sequencing errors.

Exploiting Phylogeny : Species Hopping

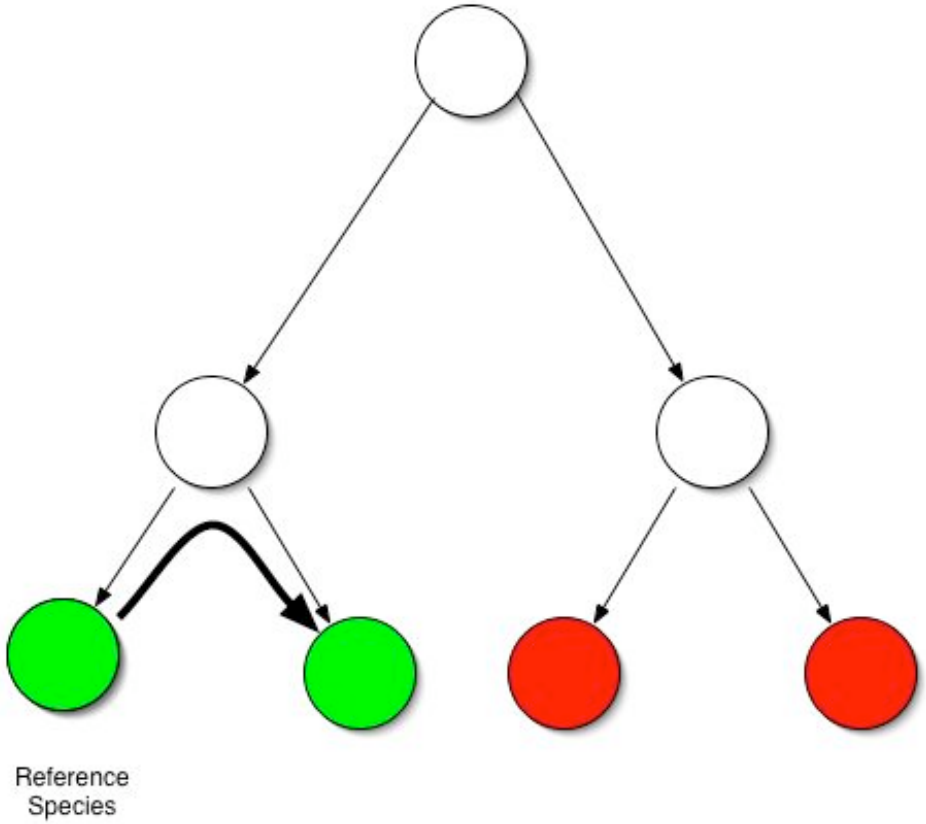


Exploiting Phylogeny : Species Hopping



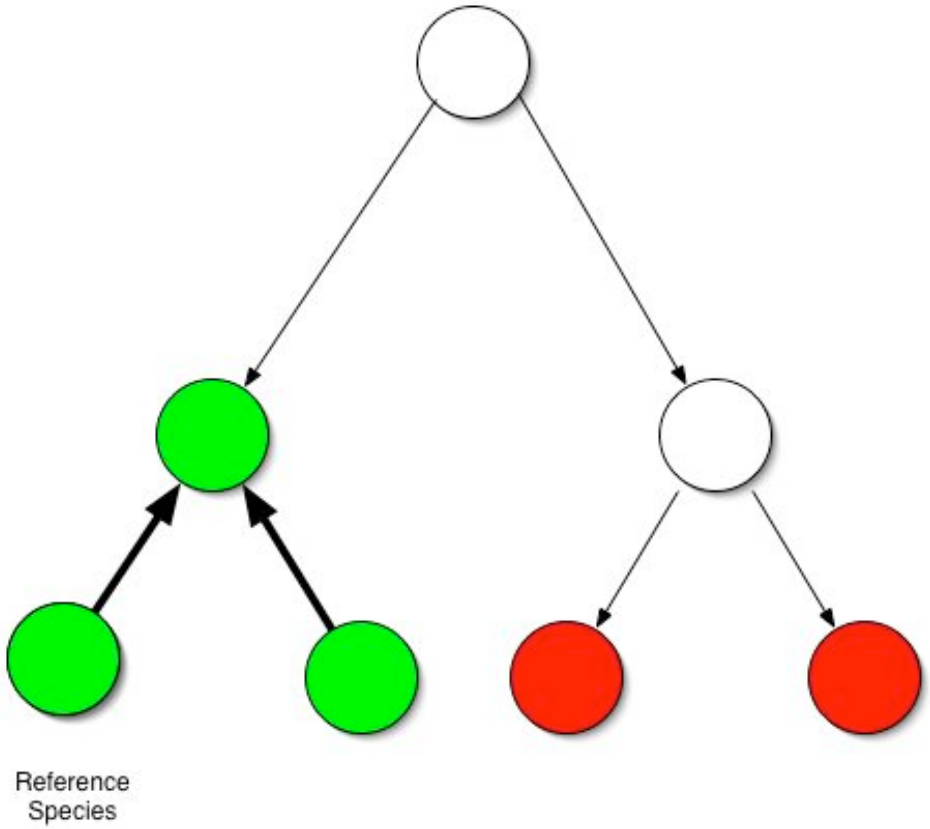
Map gene into closest species

Exploiting Phylogeny : Species Hopping



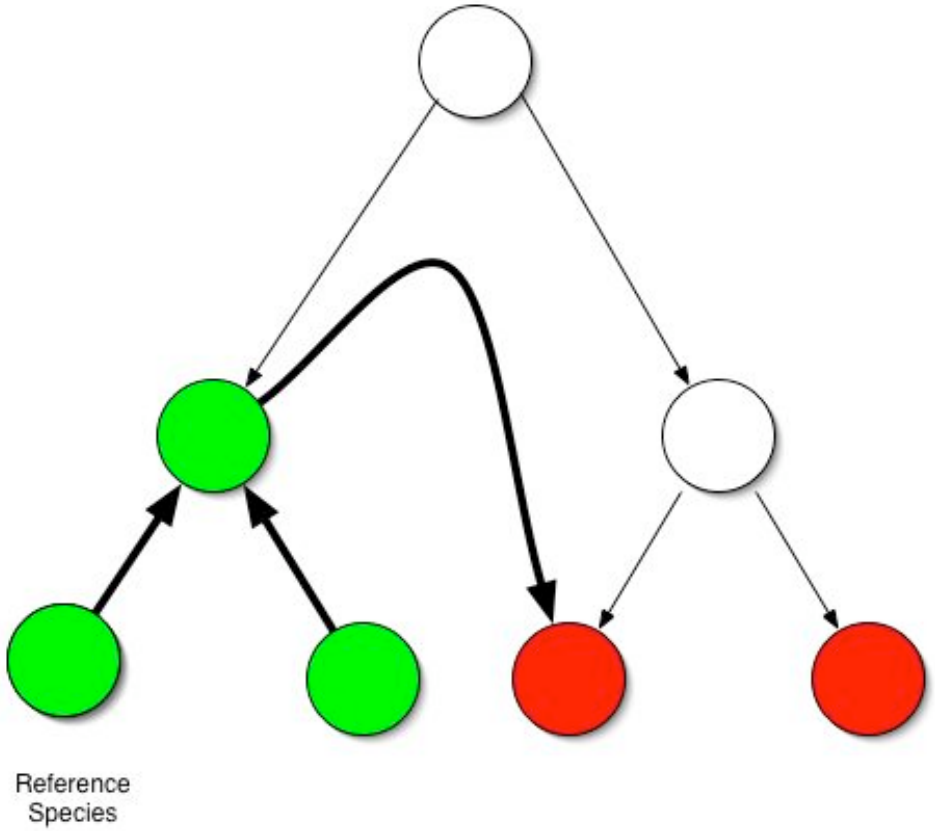
Map gene into closest species

Exploiting Phylogeny : Species Hopping



Add the prediction to the profile

Exploiting Phylogeny : Species Hopping



Use profile to map gene into the next closest species

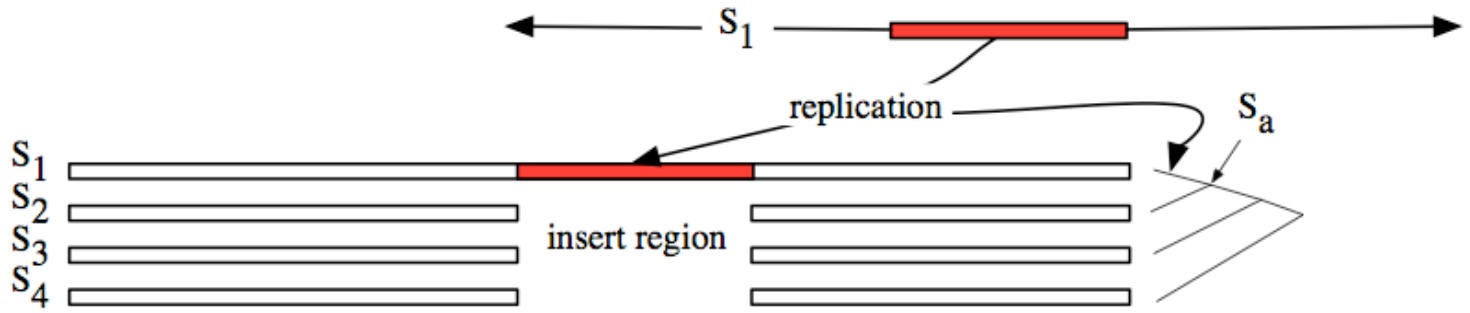
Annotation Statistics

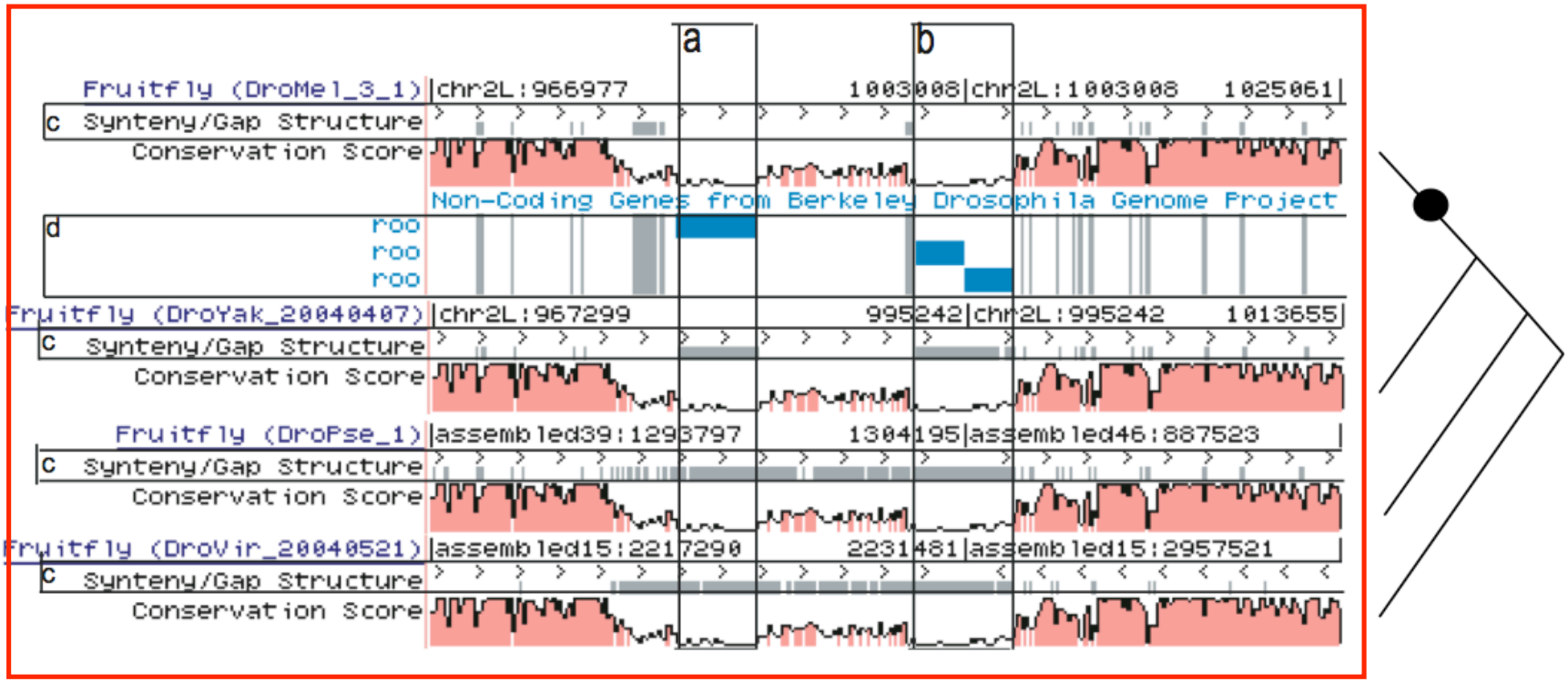
Species	Transcripts	Unique	Complete
D. melanogaster	19697	13488	N/A
D. simulans	18274	12353	17074
D. yakuba	18551	12594	17614
D. erecta	18700	12682	18203
D. ananassae	17398	11561	15858
D. pseudoobscura	16651	10867	14595
D. mojavensis	15908	10214	13192
D. virilis	16032	10305	13451
D. grimshawi	15700	10063	13107

Comparison with Existing Programs

	Projector	GeneWise	GeneMapper
Gene Sn	59.13	60.79	**81.54**
Gene Sp	59.13	60.79	**81.54**
Exon Sn	94.19	92.72	**97.15**
Exon Sp	90.46	93.44	**97.79**
Nucl Sn	99.07	99.30	**99.99**
Nucl Sp	99.29	99.71	**99.99**

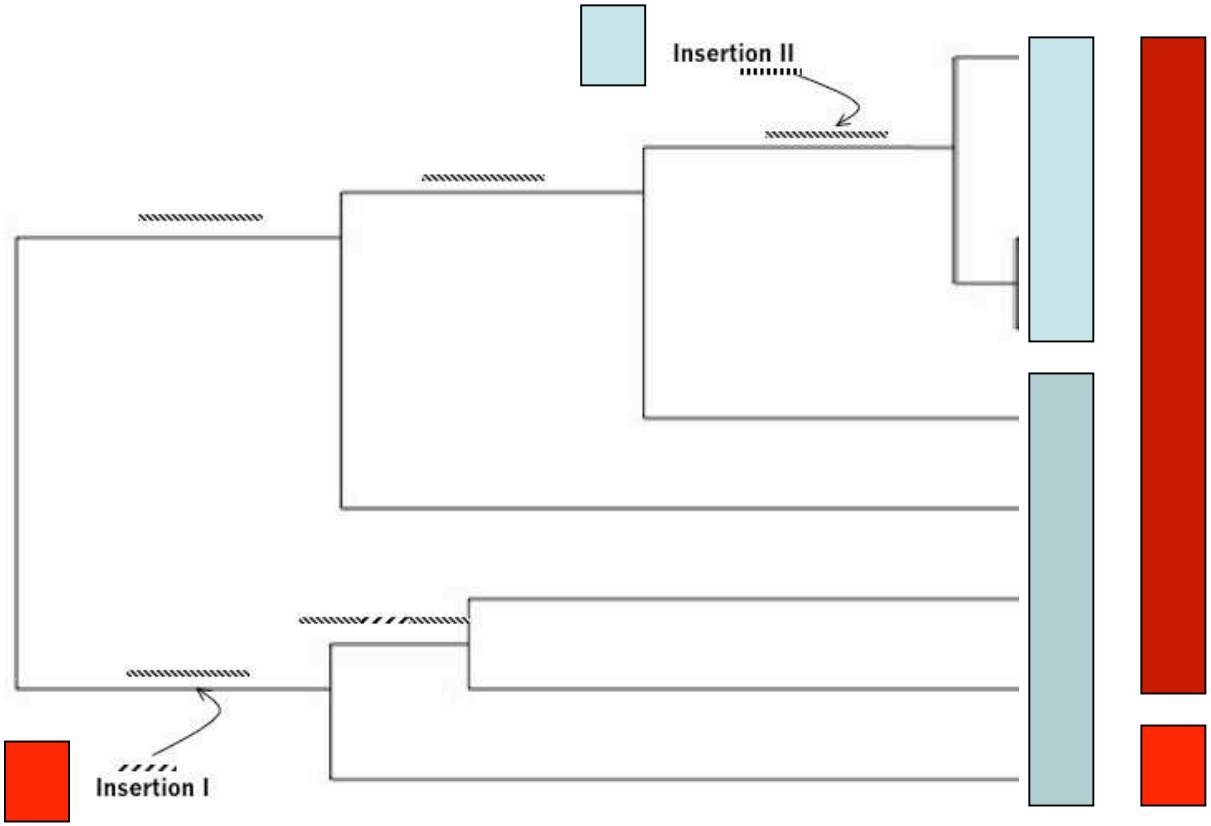
Novel approach to finding TEs





Caspi and Pachter, Genome Research 2006.

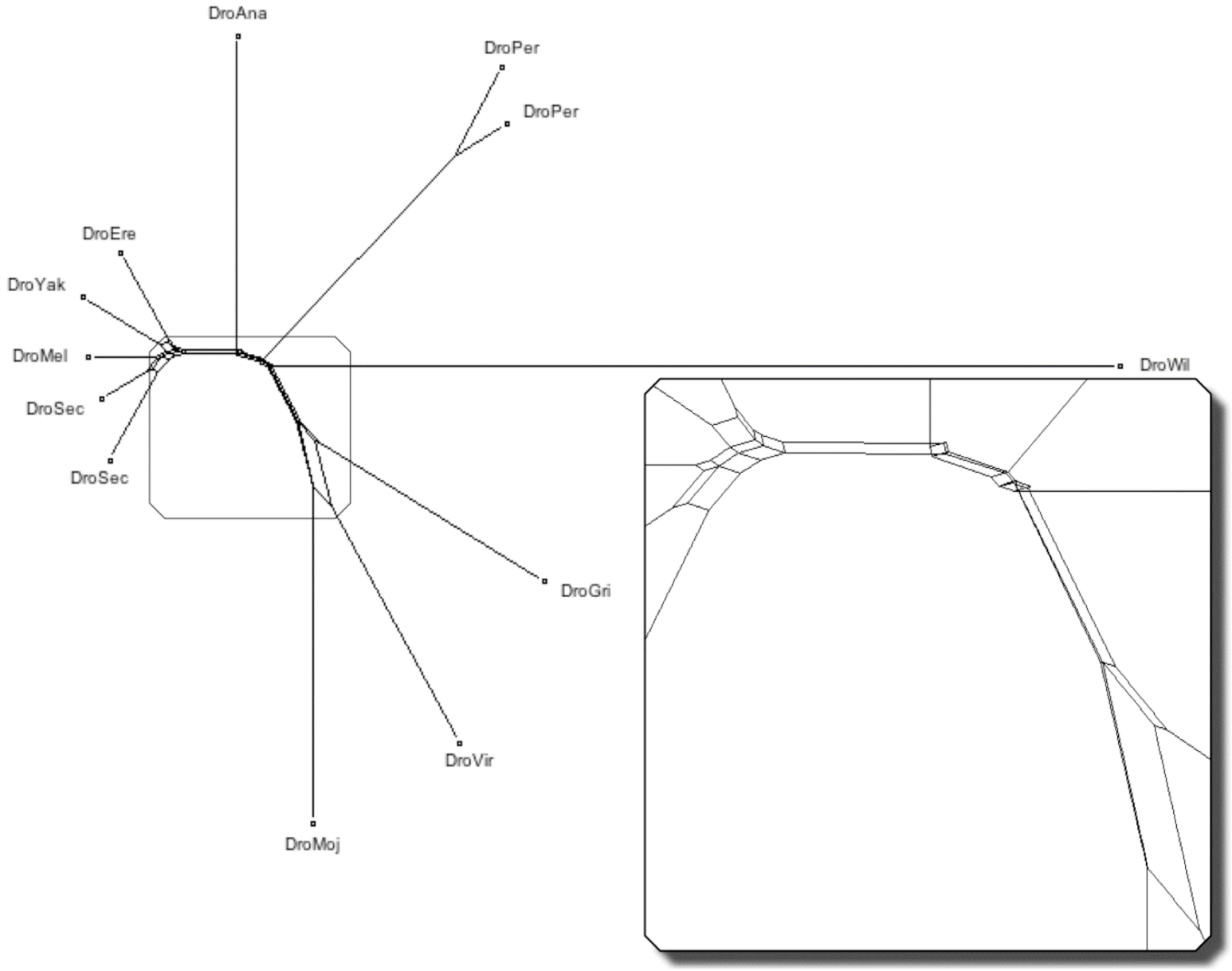
Data consists of *splits*

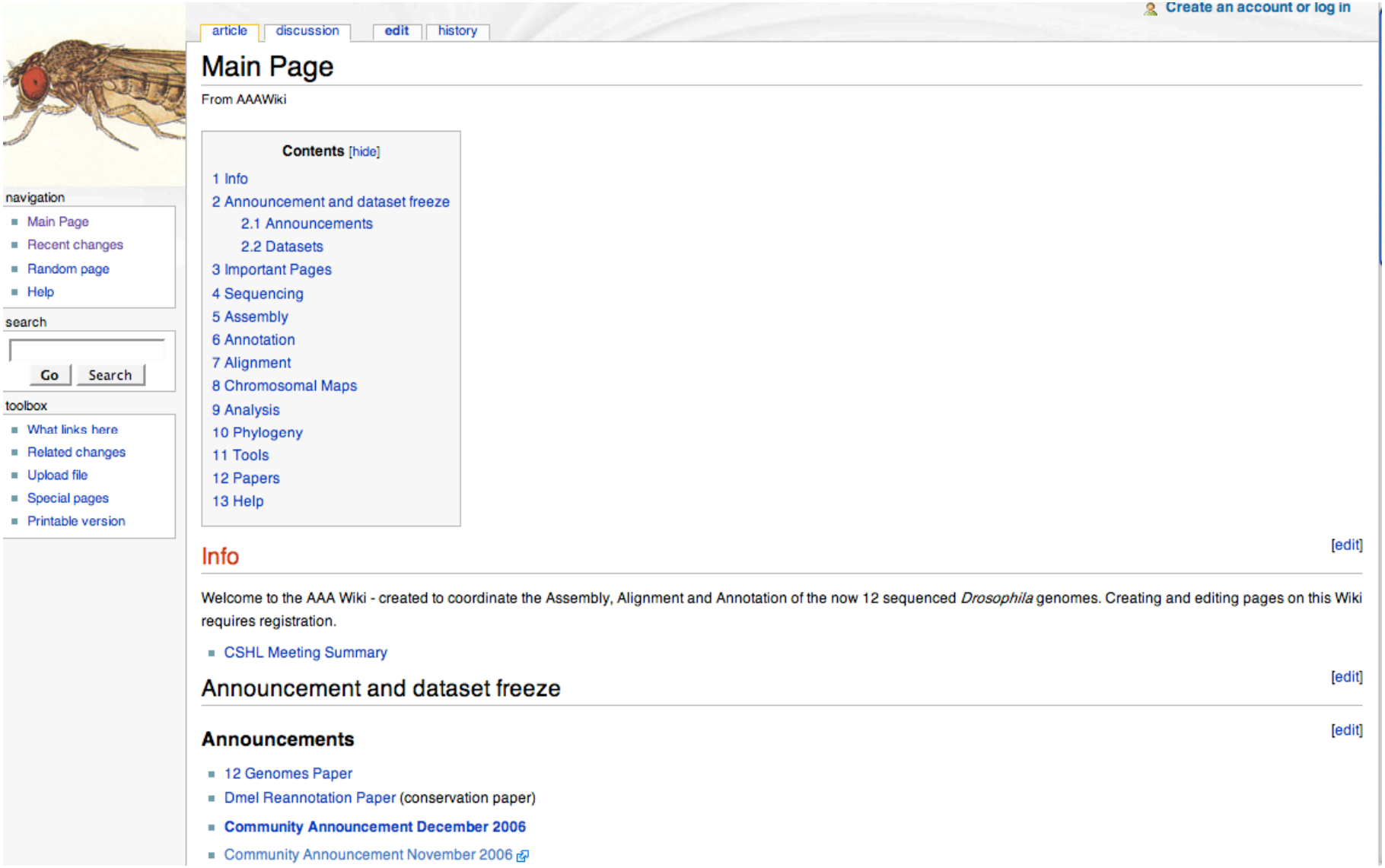


The data

```
  .
  .
  .
  3  DroAna_CAF1_DroPer_CAF1_DroPse_CAF1_DroSim_CAF1
  1  DroAna_CAF1_DroPer_CAF1_DroPse_CAF1_DroSim_CAF1_DroWil_CAF1
  3  DroAna_CAF1_DroPer_CAF1_DroPse_CAF1_DroSim_CAF1_DroWil_CAF1_DroYak_CAF1
  3  DroAna_CAF1_DroPer_CAF1_DroPse_CAF1_DroSim_CAF1_DroYak_CAF1
  6  DroAna_CAF1_DroPer_CAF1_DroPse_CAF1_DroVir_CAF1
 16  DroAna_CAF1_DroPer_CAF1_DroPse_CAF1_DroVir_CAF1_DroWil_CAF1
 61  DroAna_CAF1_DroPer_CAF1_DroPse_CAF1_DroWil_CAF1
  4  DroAna_CAF1_DroPer_CAF1_DroPse_CAF1_DroWil_CAF1_DroYak_CAF1
  8  DroAna_CAF1_DroPer_CAF1_DroPse_CAF1_DroYak_CAF1
  1  DroAna_CAF1_DroPer_CAF1_DroSec_CAF1
  1  DroAna_CAF1_DroPer_CAF1_DroSec_CAF1_DroSim_CAF1
  1  DroAna_CAF1_DroPer_CAF1_DroVir_CAF1
  1  DroAna_CAF1_DroPer_CAF1_DroVir_CAF1_DroYak_CAF1
  2  DroAna_CAF1_DroPer_CAF1_DroWil_CAF1
 62  DroAna_CAF1_DroPse_CAF1
  1  DroAna_CAF1_DroPse_CAF1_DroSim_CAF1_DroYak_CAF1
  2  DroAna_CAF1_DroPse_CAF1_DroVir_CAF1
  1  DroAna_CAF1_DroPse_CAF1_DroVir_CAF1_DroWil_CAF1
  5  DroAna_CAF1_DroPse_CAF1_DroWil_CAF1
  2  DroAna_CAF1_DroPse_CAF1_DroYak_CAF1
 59  DroAna_CAF1_DroSec_CAF1
 13  DroAna_CAF1_DroSec_CAF1_DroSim_CAF1
  1  DroAna_CAF1_DroSec_CAF1_DroSim_CAF1_DroWil_CAF1
  1  DroAna_CAF1_DroSec_CAF1_DroSim_CAF1_DroWil_CAF1_DroYak_CAF1
  3  DroAna_CAF1_DroSec_CAF1_DroSim_CAF1_DroYak_CAF1
  1  DroAna_CAF1_DroSec_CAF1_DroVir_CAF1_DroWil_CAF1
  1  DroAna_CAF1_DroSec_CAF1_DroVir_CAF1_DroYak_CAF1
  2  DroAna_CAF1_DroSec_CAF1_DroWil_CAF1
  1  DroAna_CAF1_DroSec_CAF1_DroYak_CAF1
 35  DroAna_CAF1_DroSim_CAF1
  2  DroAna_CAF1_DroSim_CAF1_DroWil_CAF1
  7  DroAna_CAF1_DroSim_CAF1_DroYak_CAF1
 70  DroAna_CAF1_DroVir_CAF1
  6  DroAna_CAF1_DroVir_CAF1_DroWil_CAF1
403  DroAna_CAF1_DroWil_CAF1
  9  DroAna_CAF1_DroWil_CAF1_DroYak_CAF1
  .
  .
  .
```

Phylogenetics with splits


DroAna
DroPer
DroPer
DroEre
DroYak
DroMel
DroSec
DroSec
DroWil
DroGri
DroVir
DroMoj



navigation
- Main Page
- Recent changes
- Random page
- Help

search
Go Search

toolbox
- What links here
- Related changes
- Upload file
- Special pages
- Printable version

Create an account or log in

article | discussion | edit | history

Main Page

From AAAWiki

Contents [hide]

Info [edit]

Welcome to the AAA Wiki - created to coordinate the Assembly, Alignment and Annotation of the now 12 sequenced *Drosophila* genomes. Creating and editing pages on this Wiki requires registration.

- CSHL Meeting Summary

Announcement and dataset freeze [edit]

Announcements [edit]

- 12 Genomes Paper
- Dmel Reannotation Paper (conservation paper)
- **Community Announcement December 2006**
- Community Announcement November 2006

http://rana.lbl.gov/drosophila/wiki/index.php/Main_Page