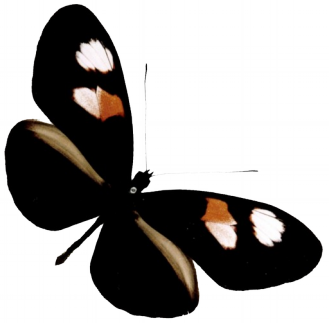
Bioinformatics for Next-Generation Sequencing
ICIPE, Nairobi, November 2014

Evolutionary Genomics 1: Comparative and Phylo-genomics

Simon Martin

Department of Zoology
University of Cambridge



Evolutionary Genomics

- Lecture 1 (this one): Comparative and phylogenomics
  - Genome organisation and synteny
  - Gene family evolution
  - Phylogenomics

- Lecture 2: Population genomics
  - Diversity and population structure
  - Adaptive and neutral evolution
  - Demographic history and hybridisation
  - Speciation

Genome organisation and synteny

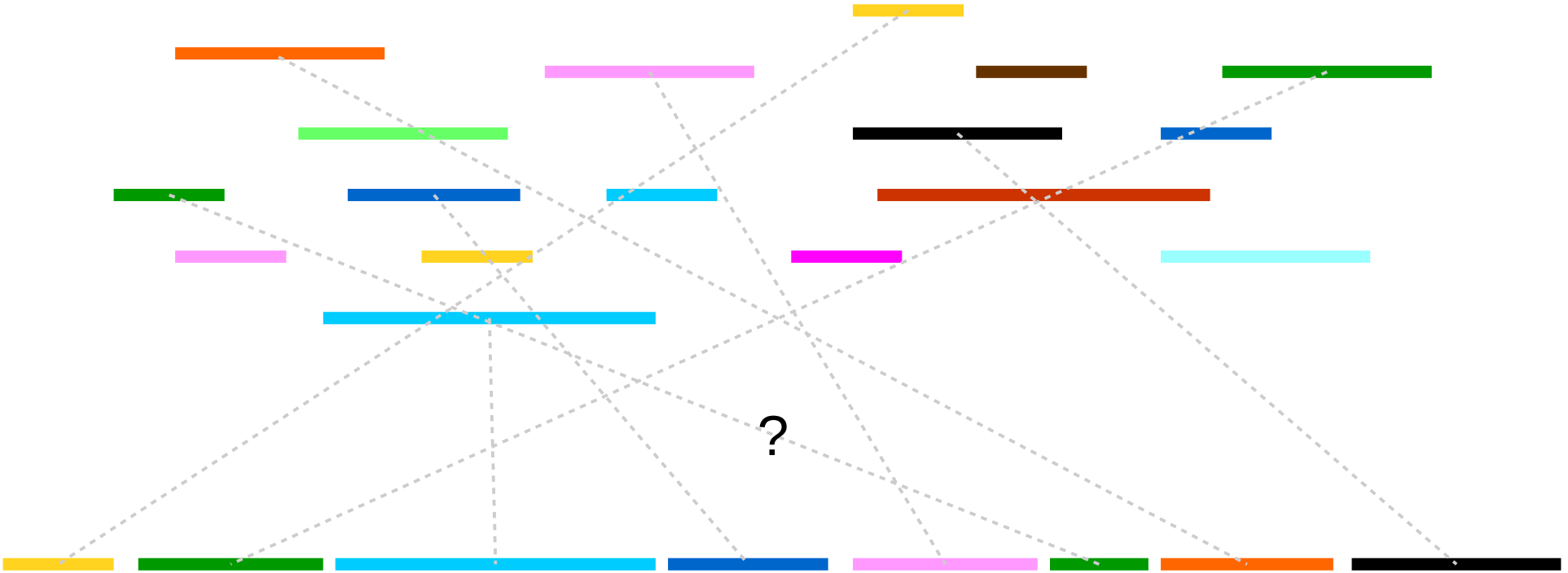
- How has the size and organisation of the genome changed between species?
- Macrosynteny
  - Chromosomal fusions
  - Splits
  - segmental duplications
- Microsynteny
  - Gene order
  - indels
  - inversions

Macrosynteny

- Requires “complete” genome assemblies
- Linkage map (or chromosome-sized scaffolds)
- Annotated genes for both species
- Homologous gene pairs

From contigs to chromosomes

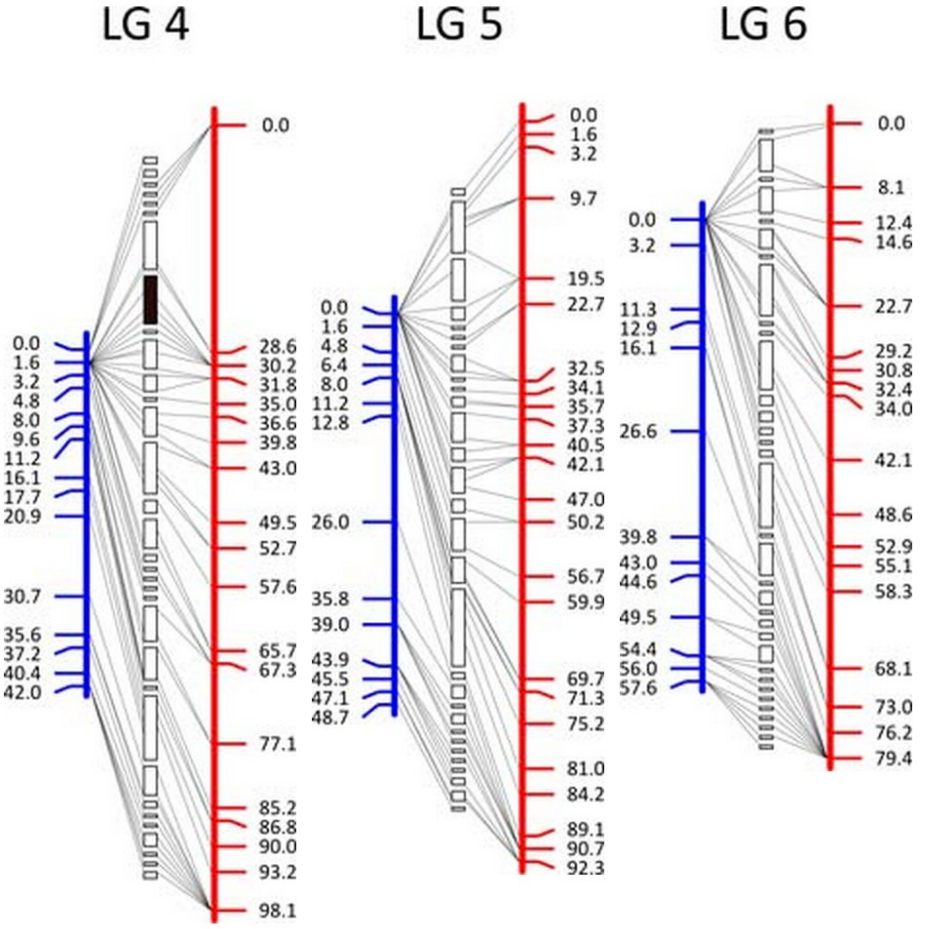
- Next-gen approaches usually produce fairly small contigs.
  - Usually thousands to tens/hundreds of thousands
- So how do we assemble chromosomes?



From scaffolds to chromosomes

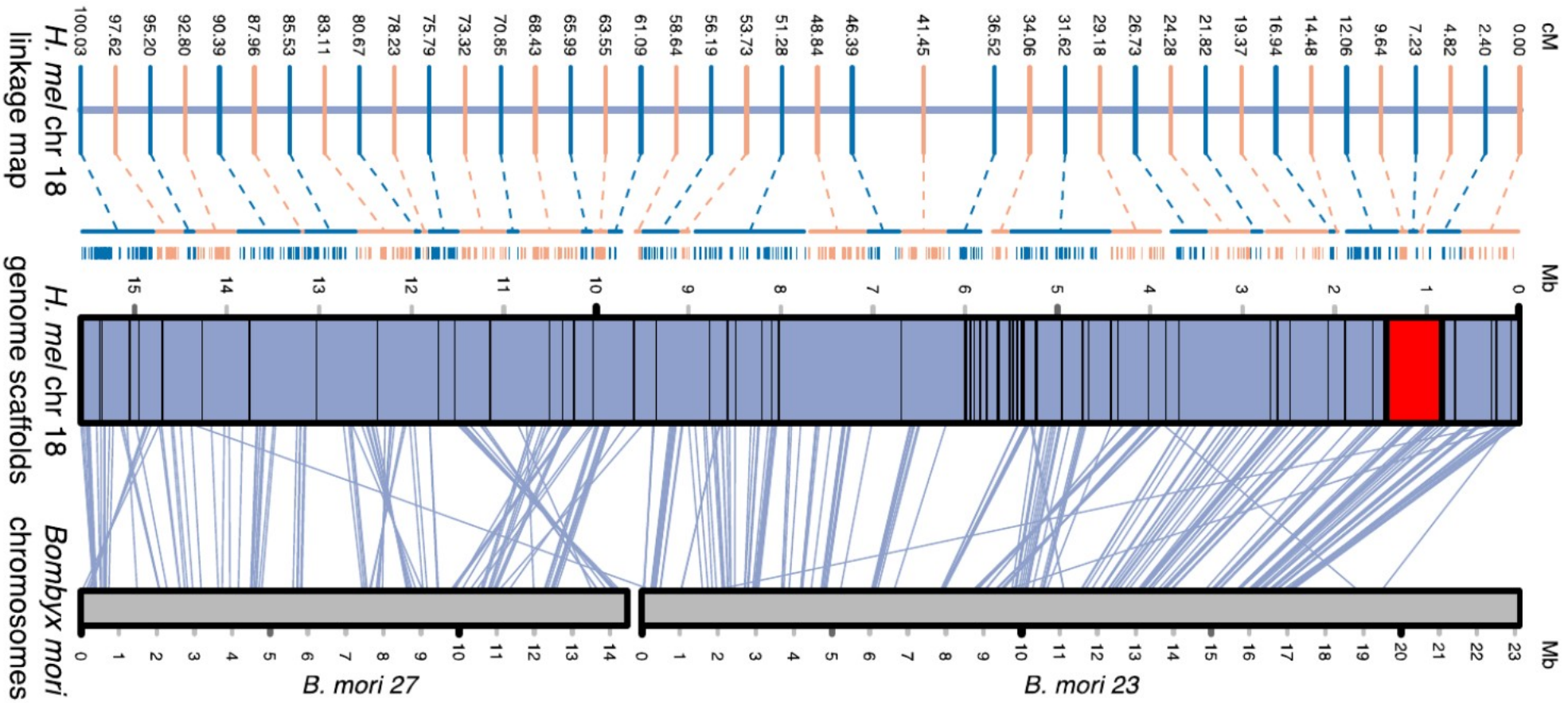
- Connect **genetic linkage maps** to scaffold sequences

linkage group 4					
Marker	Scaffold Number in Fugu ver5	Map position (cM)		Primer sequences	
		Male Map	Female Map	Forward (5'-3')	Reverse (5'-3')
f799	357	0.0	-	ggtttgggcacattctgact	ttagcatttaggccgacgag
f756	357	0.0	-	gtaaagggctggatggactg	ggctggaaaaactgctgact
f1414	588	-	0.0	catatgggcctggaacactc	gagggcgaggaagagatttt
f1424	599	0.0	0.0	cgaaaggccacagacagatt	ttgccccgaatatctggtta
f1468	655	-	0.0	cacgctgctcagttctgata	ctcgtgatagggctggtgat
f1486	676	0.0	0.0	aaaagcacccagatgtgacc	acttggccctggaggactat
f209	52	1.6	28.6	cccctctttccttctgtcatcatt	tctctgtggggtcatcgtgtt
f338	52	1.6	28.6	ggtgtggagccttcagagat	ttgcatgaattggtttcctg
f560	53	1.6	-	cggcccctatgagtattgaa	gggagcagaaaggtaaagtgaa
f351	53	-	30.2	gctgactgaaaactgcacca	atgaaccatccaacccacat
f551	53	-	30.2	tacttcatgcaggcacttgg	gagtcgcaagattggtcaca
f91	53	1.6	30.2	tgcatgacagaaatgaaaggac	tcattctttccggcttgttt
f1217	53	1.6	30.2	atctggccttgtgtttcacc	gagctcagaaaaccagtgtgg
f1594	822	1.6	30.2	aggaatgtggctcaagttgg	gccagtgctgcatagaaaca
f1747	121	1.6	30.2	ccacctgttacaaacacacacag	agtcatttgccttgtctaccc
f76	121	1.6	-	ggcccacttttcacacatct	tctaaaactggccagcagaa
f1746	121	1.6	30.2	agactgaggtggtgatcggta	cgtgtgtgcaaagggttgat
f32	121	1.6	-	ttatcagctcctcaaact	agggtcggctgtgggtaa
f1745	121	-	31.8	gaggcattcactggctcttc	agggggaaattaggatgcac
f650	201	1.6	31.8	gccaacgctacttgaaaggt	aaaccagcagttttcccaga
f1322	201	1.6	31.8	cctcaacaaagccaaactgt	cgcttcacacaaatgatgct



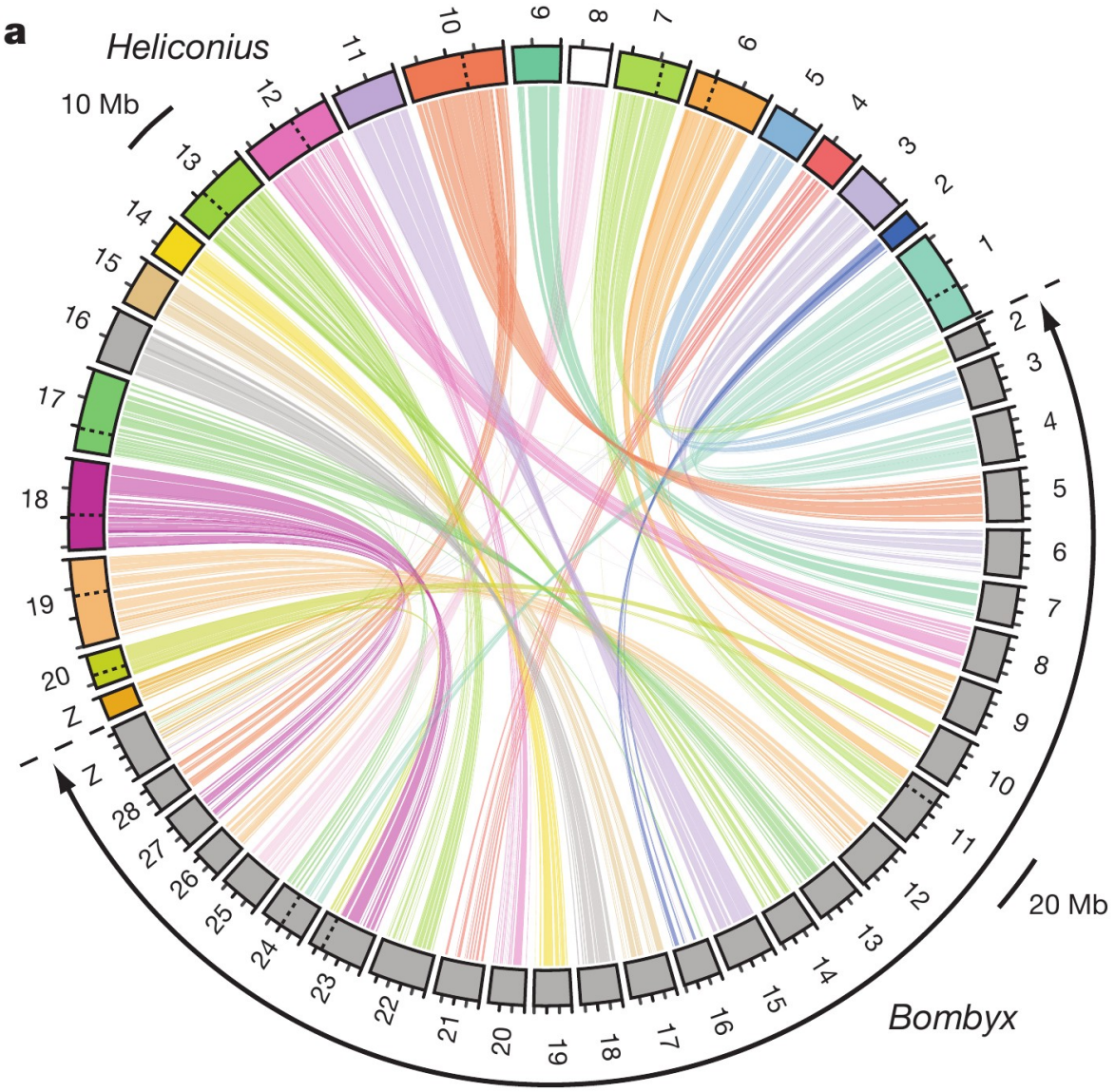
Kikuchi et al. 2011 Genome Biology and Evolution

Chromosomal synteny in *Heliconius* and *Bombyx*



The Heliconius Genome Consortium 2012 Nature

Chromosomal synteny in *Heliconius* and *Bombyx*



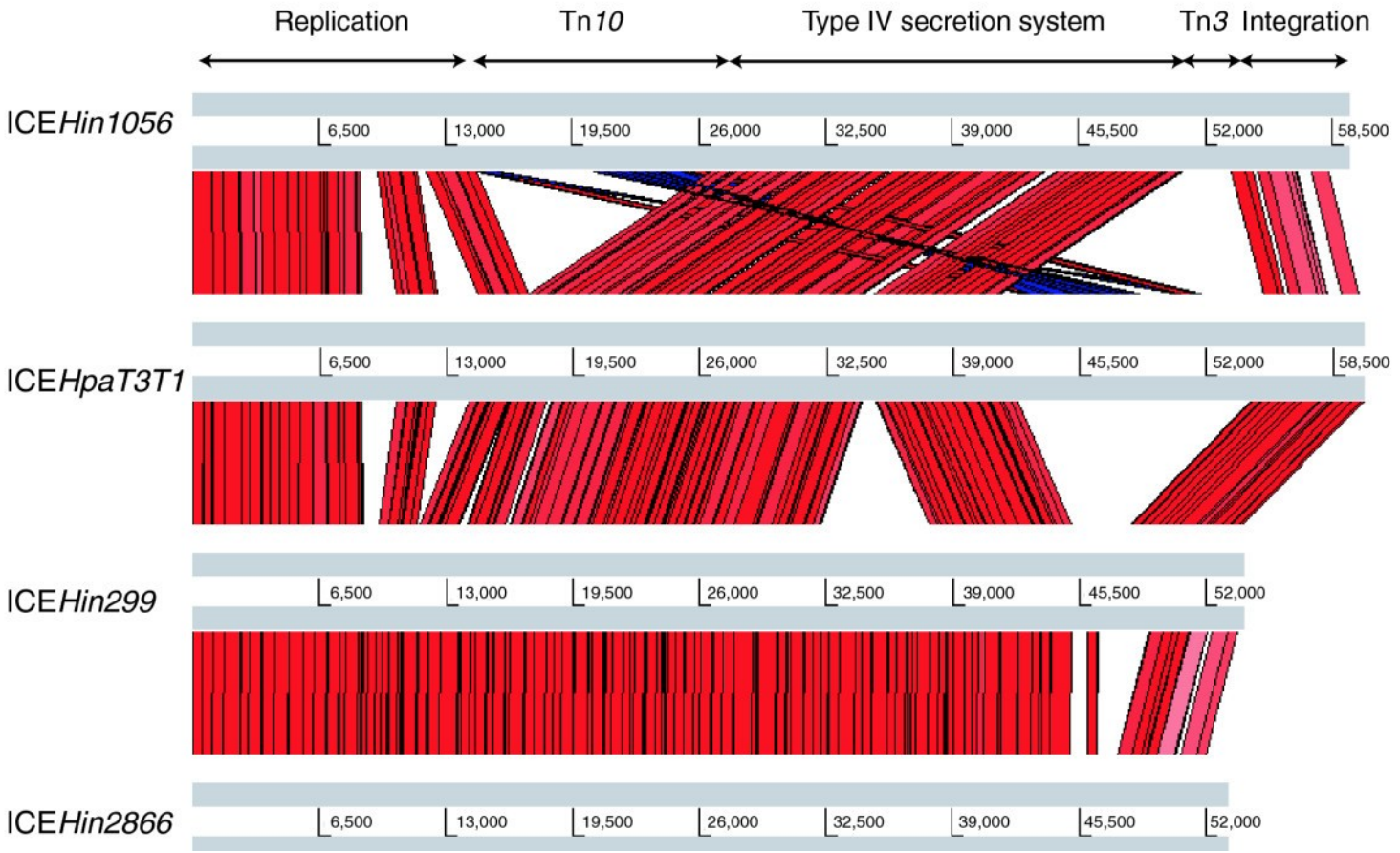
The Heliconius Genome Consortium 2012 Nature

Some software for macrosynteny

- InParanoid – find orthologous gene pairs
- OrthoCluster – identify synteny blocks and chromosomal rearrangements
- CIRCOS – beautiful chromosome plots!

Microsynteny

ACT (Artemis Comparison tool)



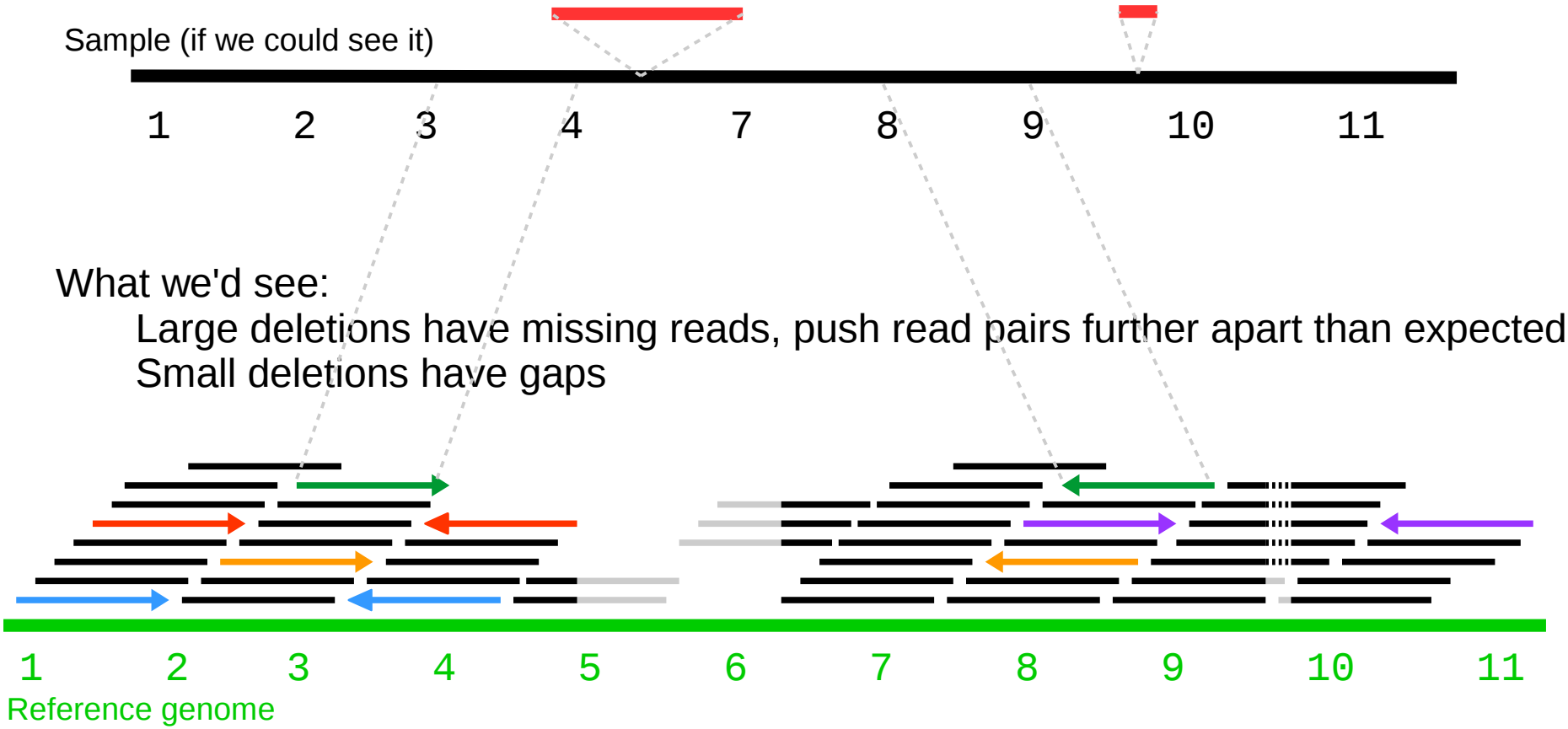
Juhas et al. 2007 Genome Biology

Can you study genome structure evolution with just one assembled genome?

- Yes, you can study “structural variants” (SVs) among individuals from the **same species or closely related species**.
- By resequencing single individuals and mapping the reads to the reference you can identify
  - Deletions
  - Insertions
  - Duplications
  - Inversions

Using next gen data to study structural variation

Deletions



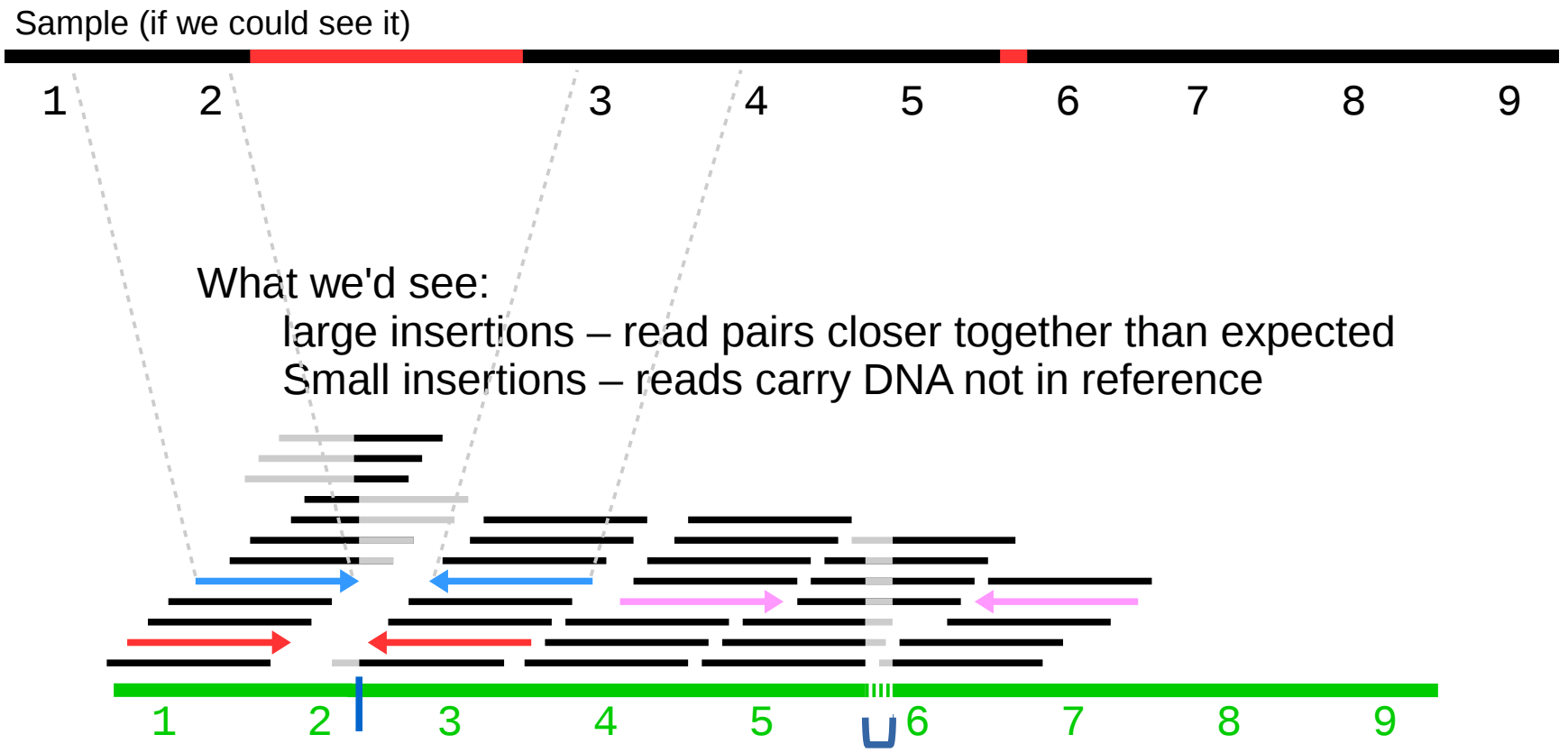
What we'd see:

Large deletions have missing reads, push read pairs further apart than expected

Small deletions have gaps

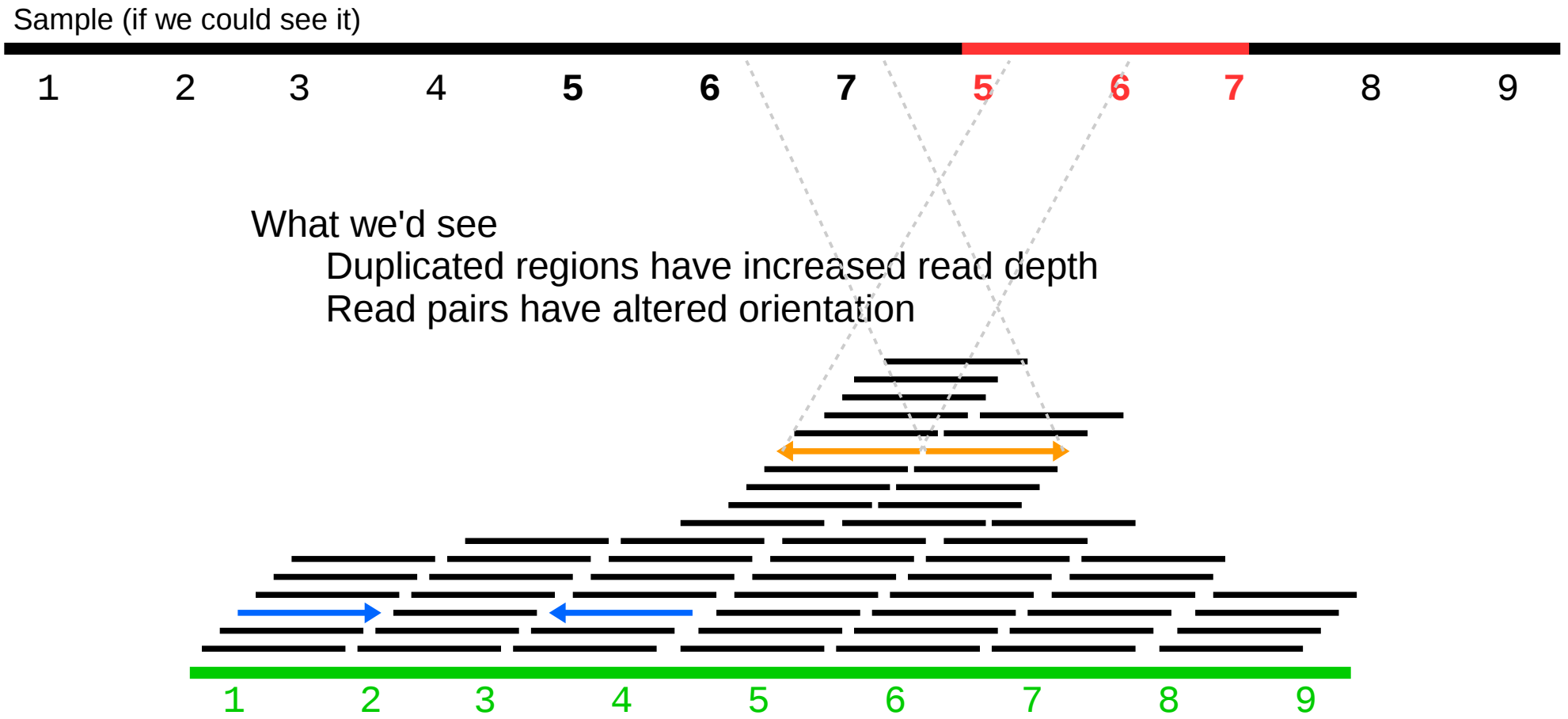
Using next gen data to study structural variation

Insertions



Using next gen data to study structural variation

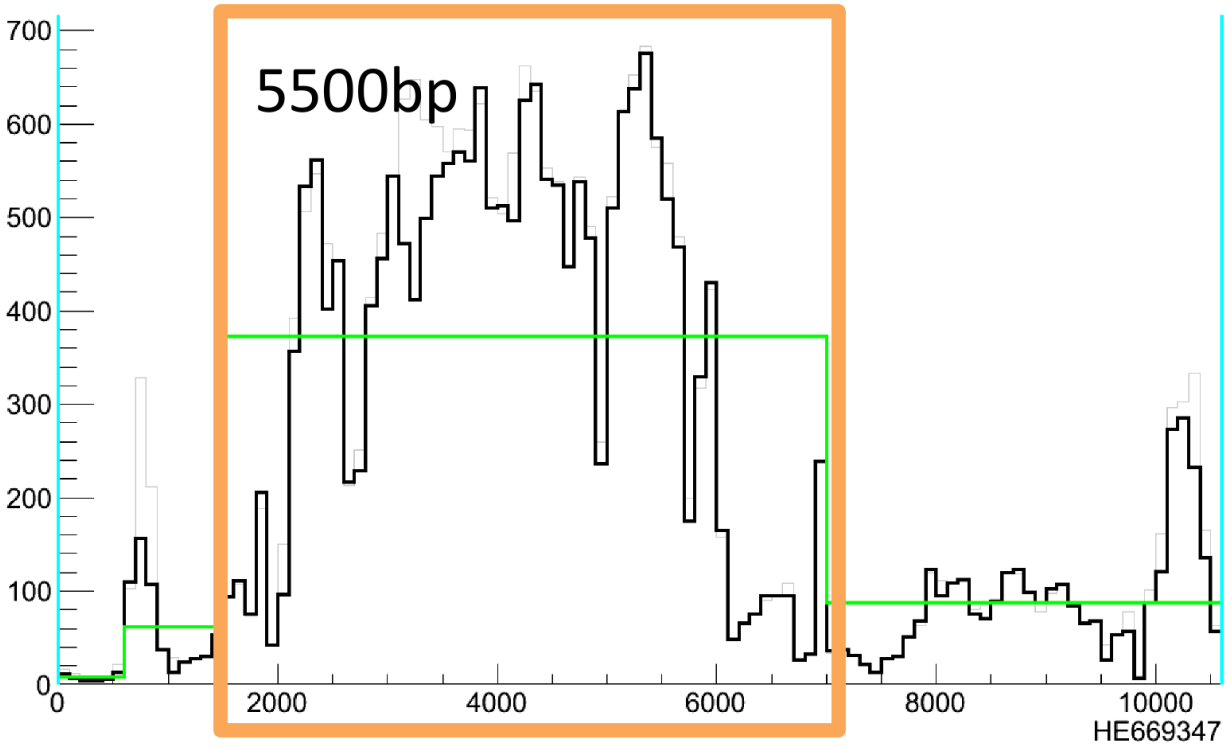
Tandem duplications



Using next gen data to study structural variation

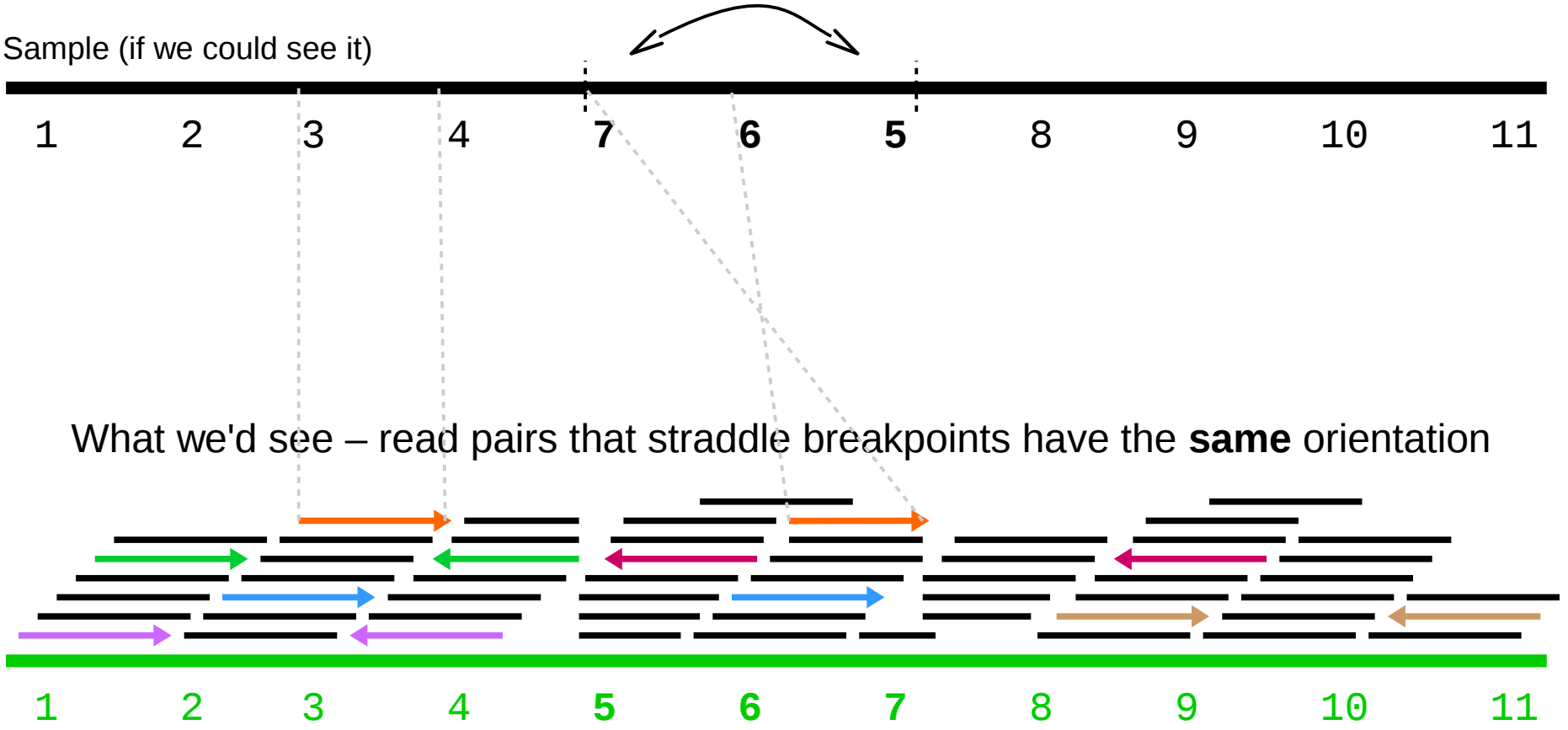
Duplications relative to reference

In reality, depth of coverage is quite variable, so only large duplications are detectable



Using next gen data to study structural variation

Inversions



Using next gen data to study structural variation

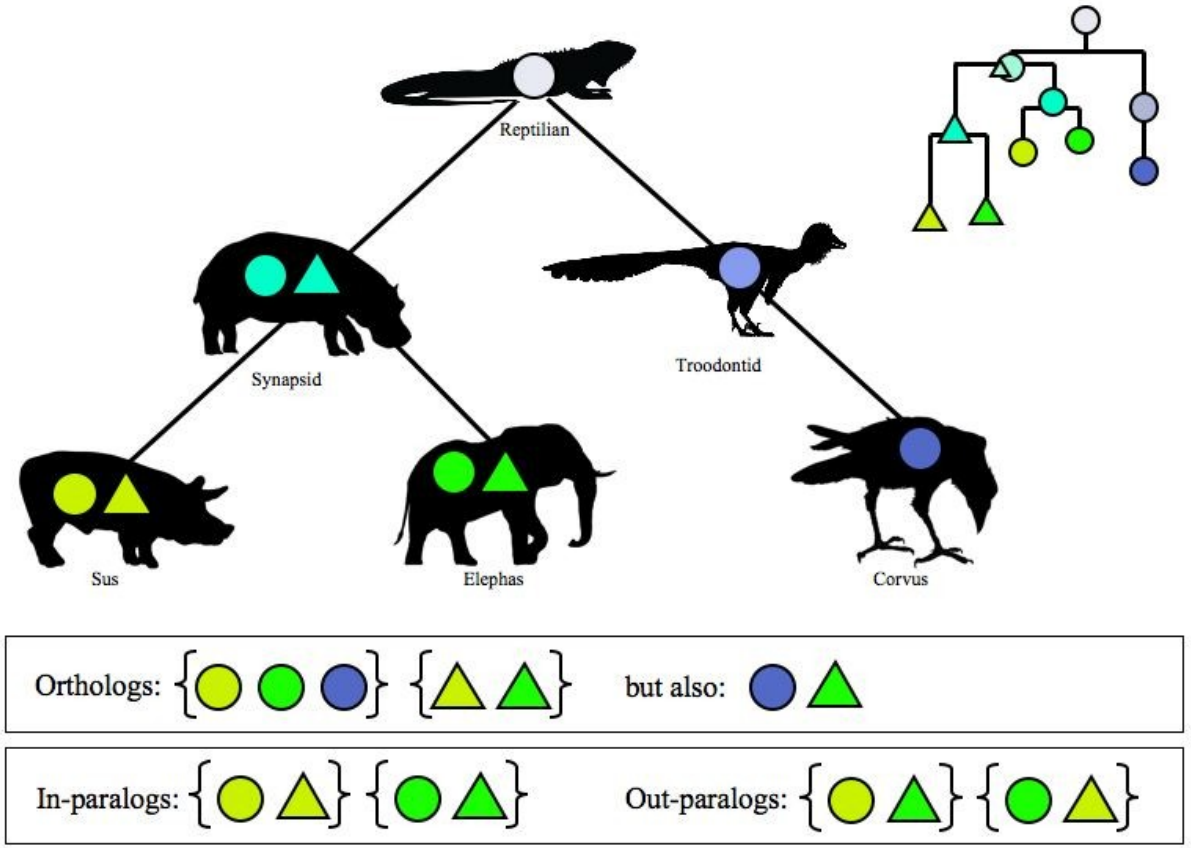
Some software

- CNVnator (Abyzov et al. 2012 Genome Research)
- Pindel (Ye et al. 2009 Bioinformatics)
- Genome STRiP (Handsaker et al. 2011 Nature Genetics)
- DELLY (Rausch et al. 2012 Bioinformatics)

Gene family evolution

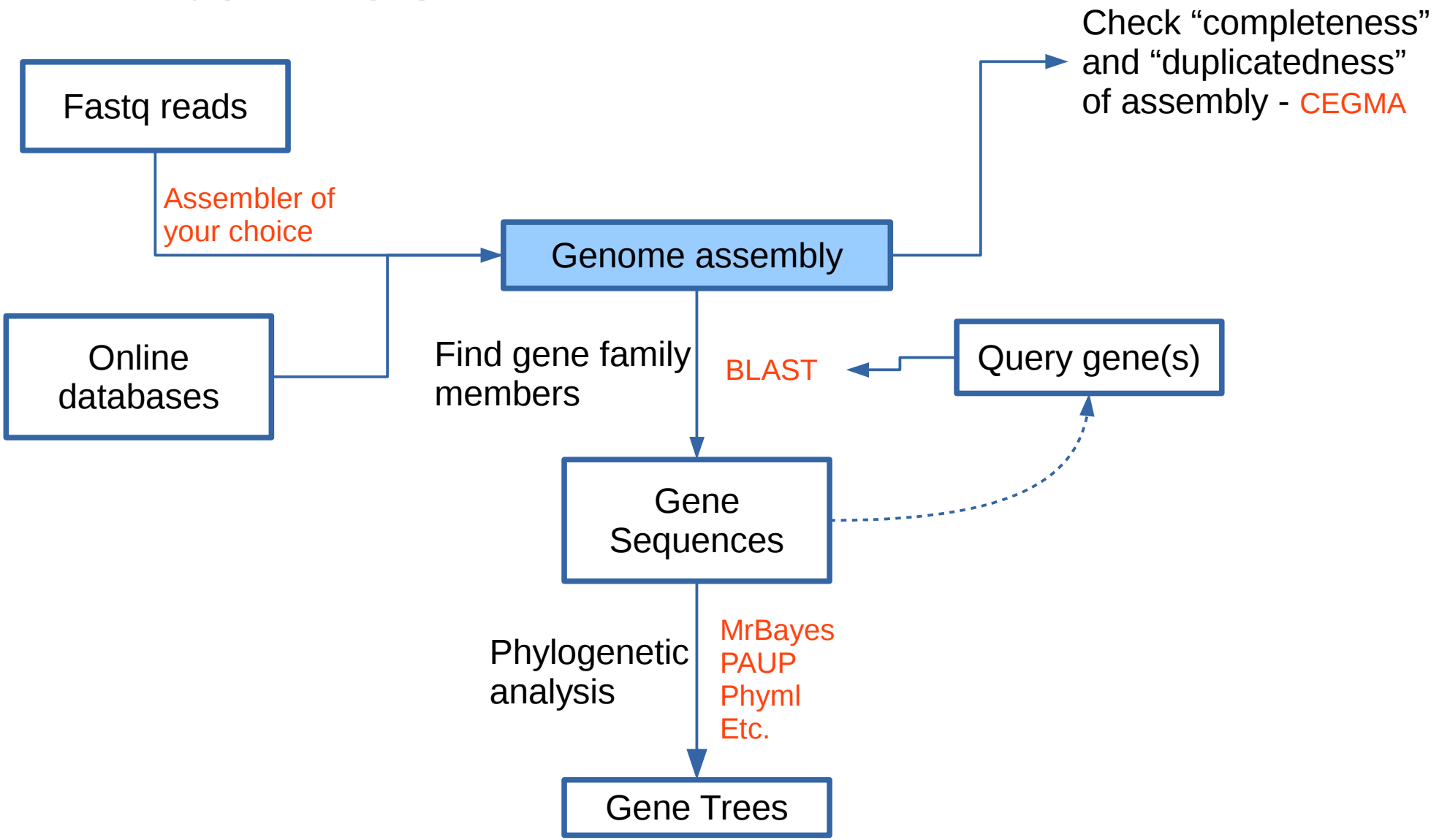
- Gene duplication leads to “families” of related genes
- E.G. There are ~80 Cytochrome P450 genes in the *Drosophlia* genome
- We can study the relationship among homologous genes in one or more species

NOTE: ***Orthologs*** and ***paralogs*** are both types of ***homologs***



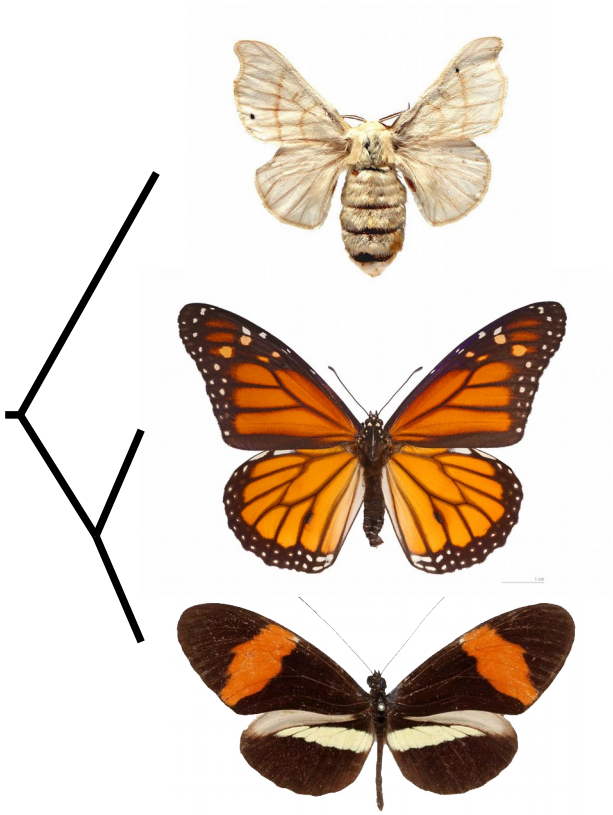
Gene family evolution

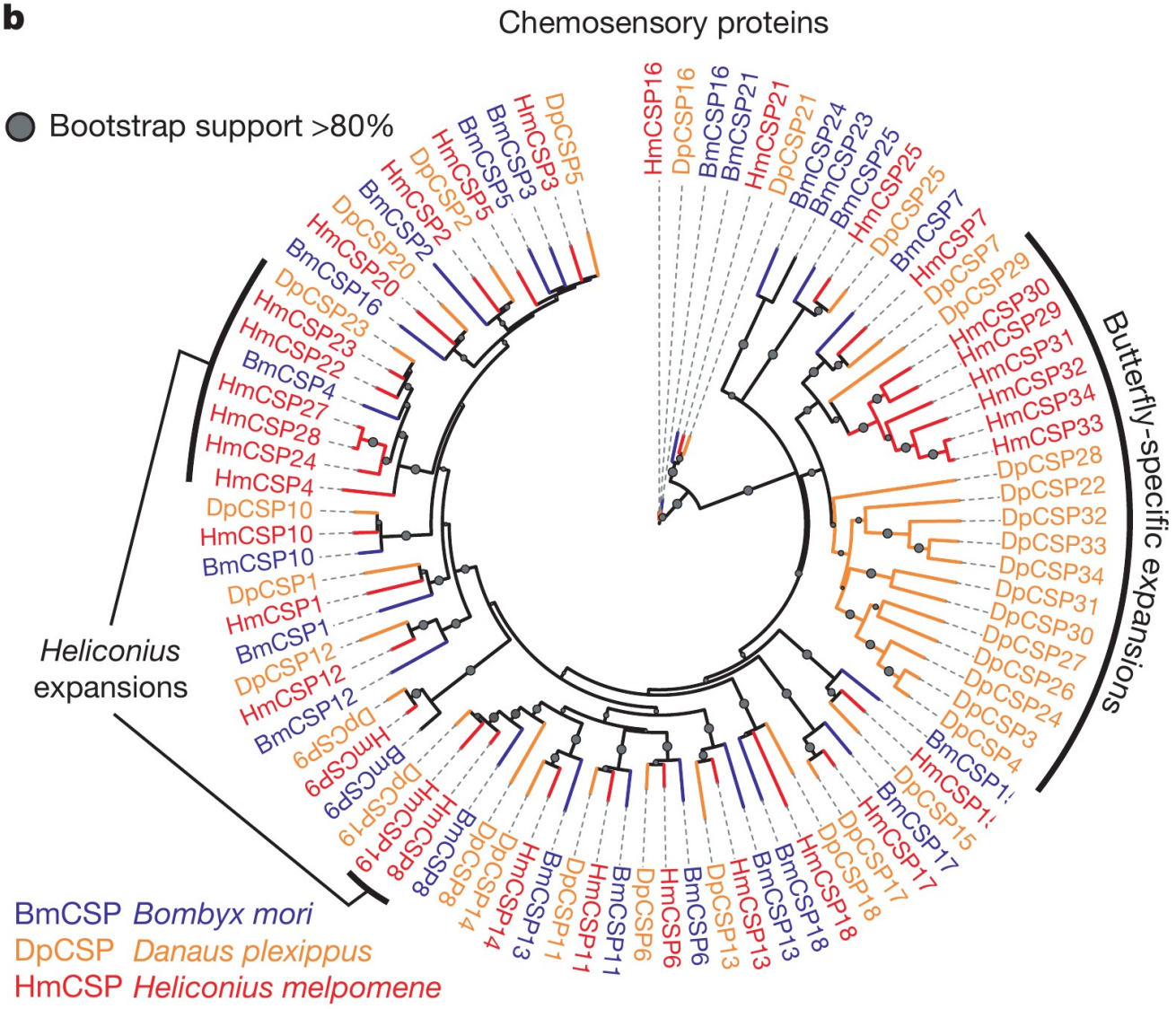
A typical pipeline



Gene family evolution

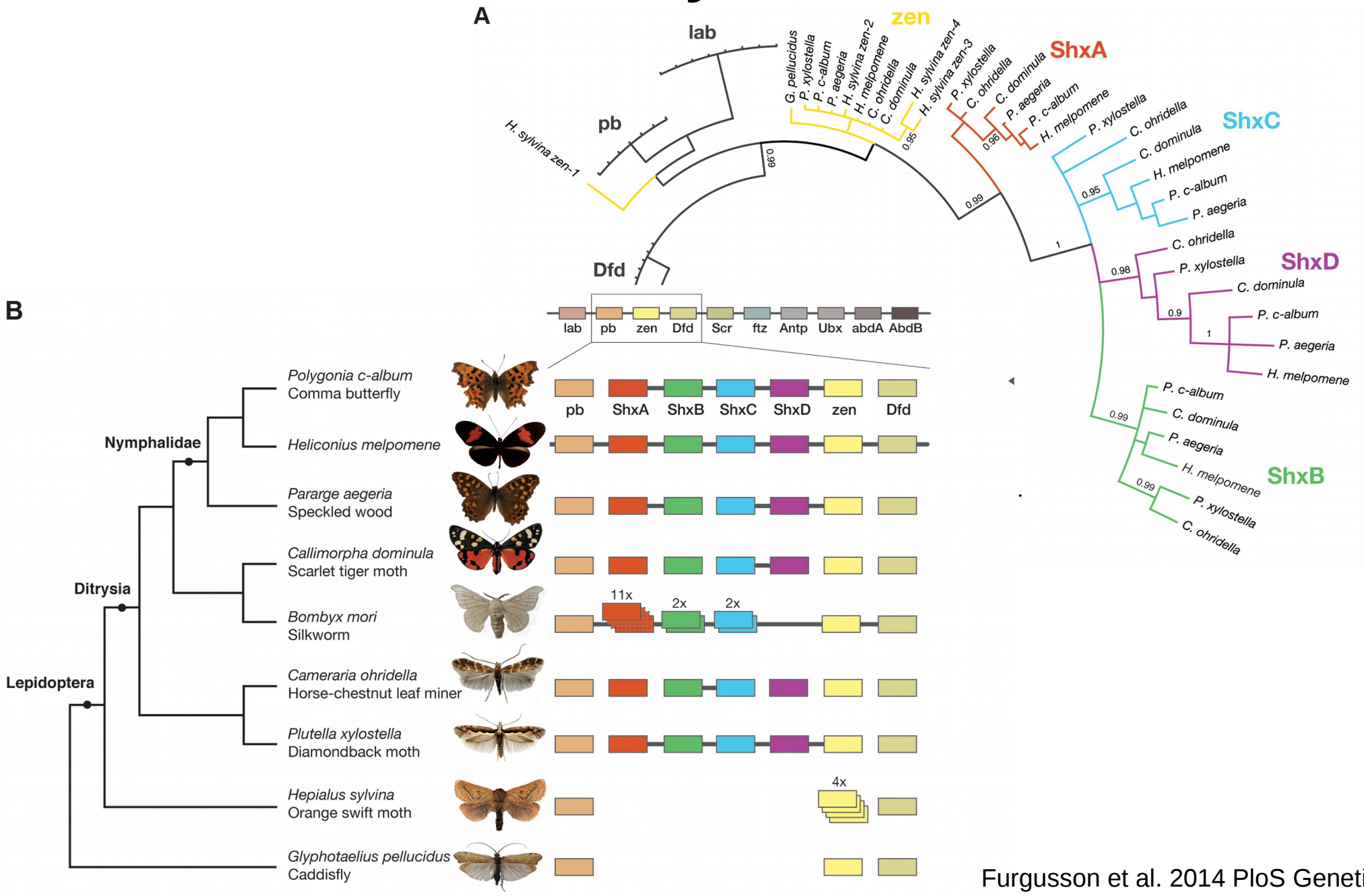
Chemosensory gene family expansions in butterflies.





The Heliconius Genome Consortium 2012 Nature

Gene family evolution

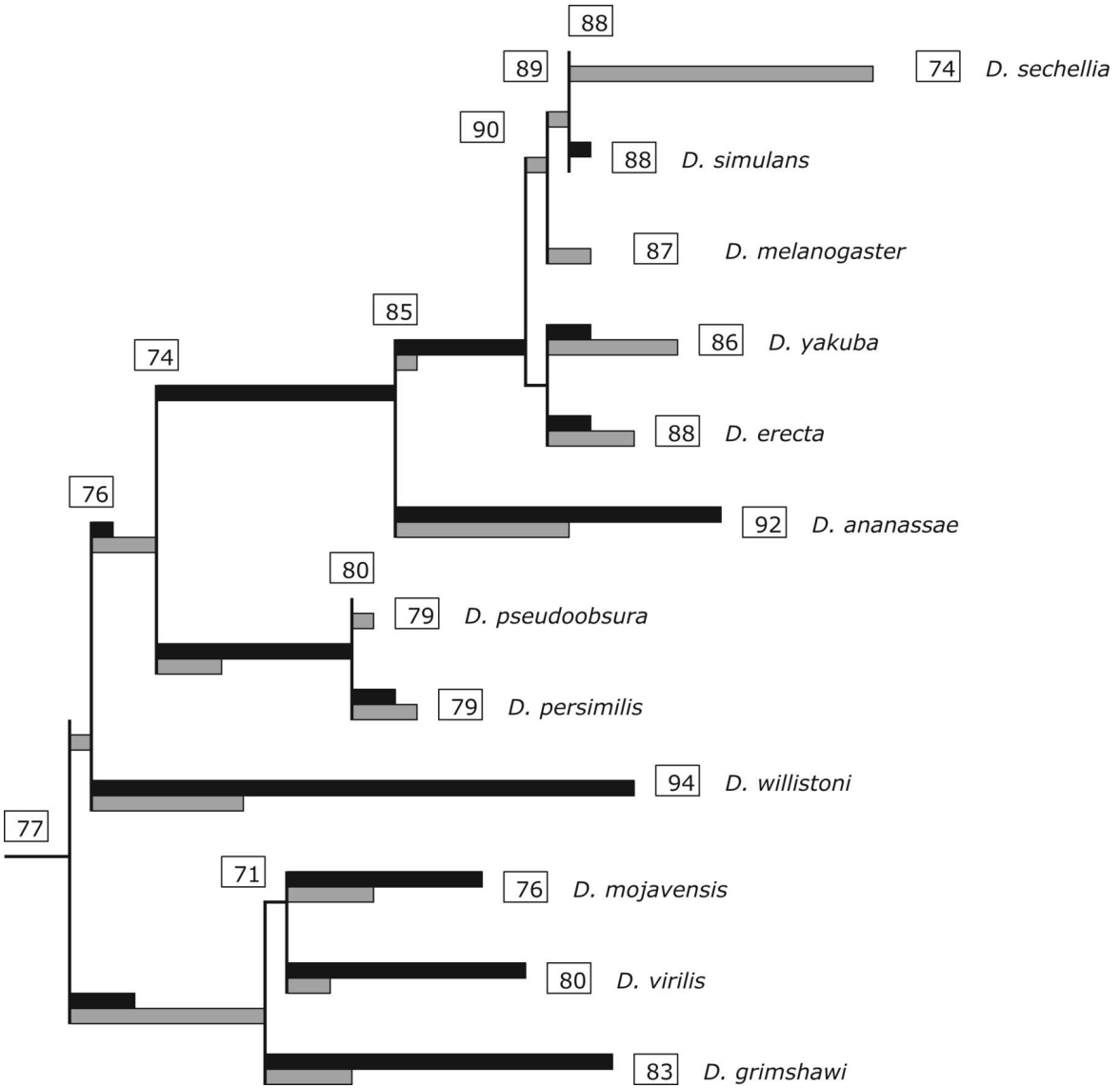


Furgusson et al. 2014 PloS Genetics

Gene family evolution

P450 gene gain and loss across the *Drosophila* phylogeny.

(c) Bugguide.net

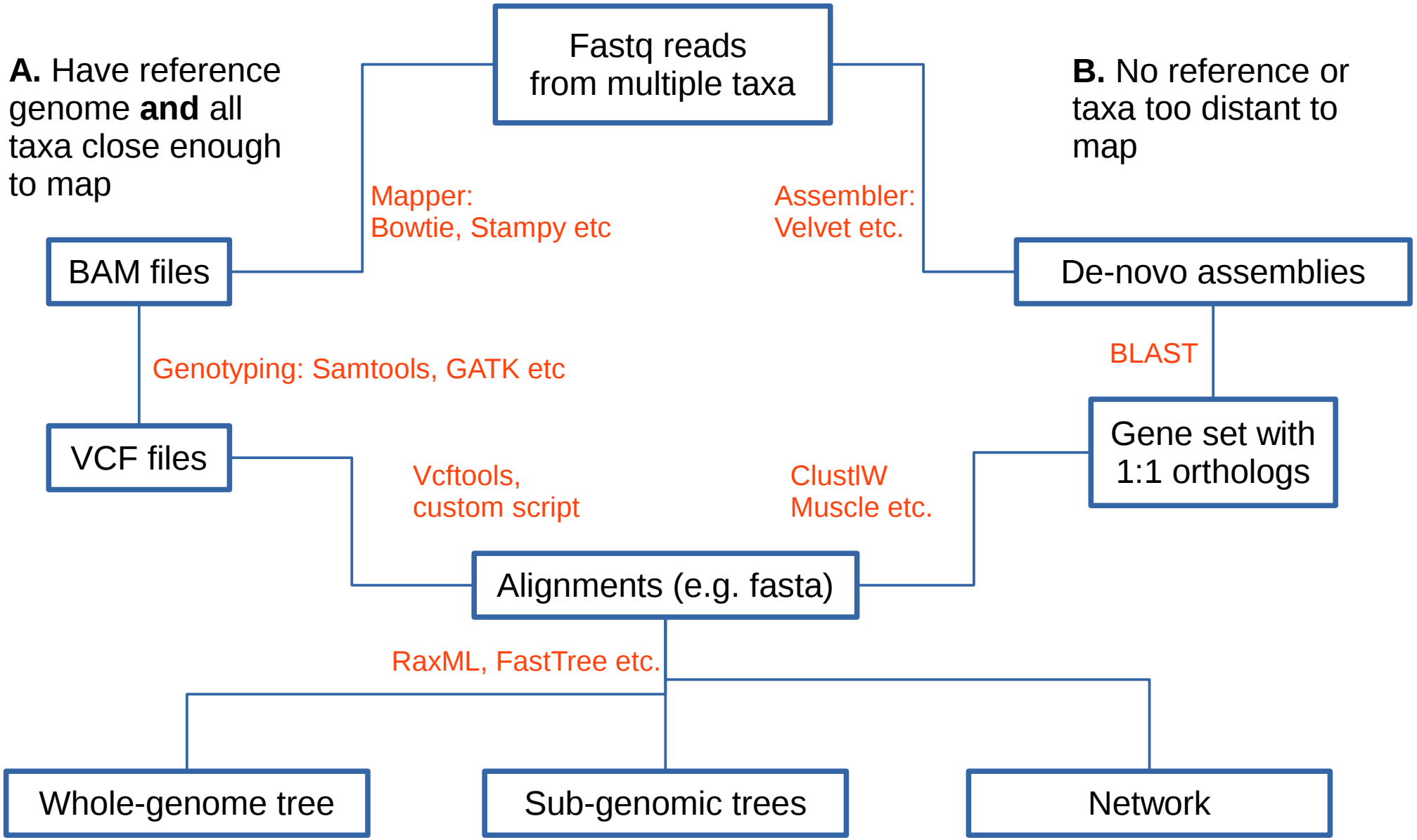


Good R T et al. 2014 Genome Biol Evol

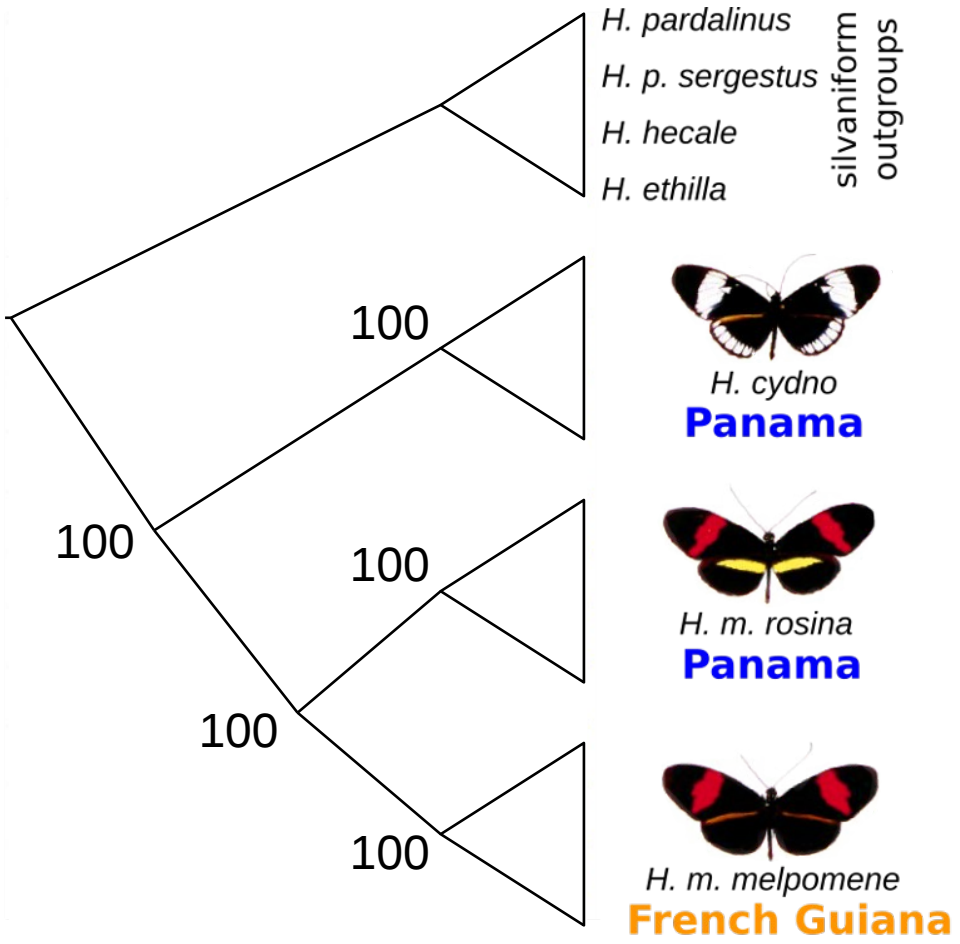
Phylogenomics

- Phylogenetics is a HUGE field – will not be covered in detail
- Instead we will focus on one misconception:
  - Genome-scale data *should* reveal the “true” relationships between species right?
  - Maybe not
- Relationships can differ in different parts of the genome
  - Lateral gene transfer and gene flow
  - Incomplete lineage sorting

Phylogenomics with next-gen data - pipeline



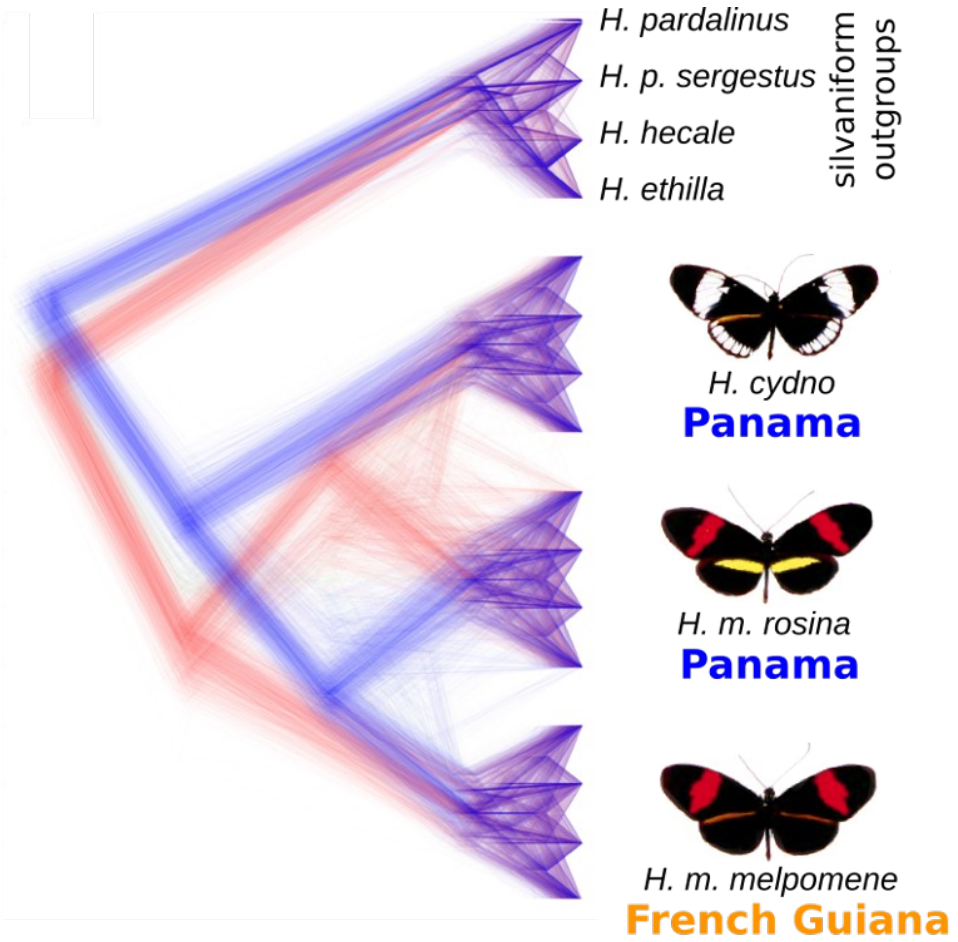
Phylogenomics: beware bootstrapping



Whole Genome Tree

100% bootstrap support for grouping by species.

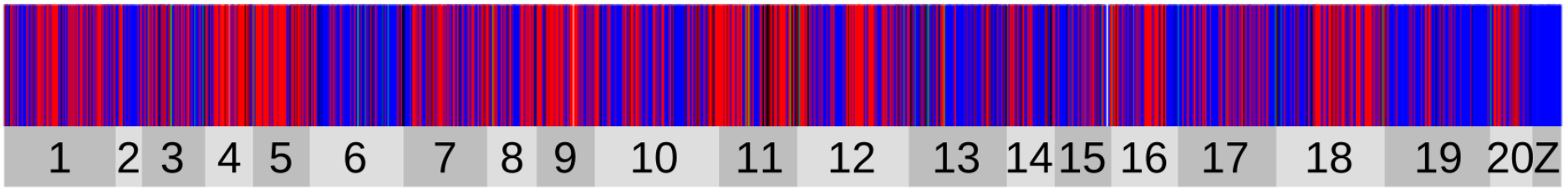
Phylogenomics: beware bootstrapping



Separate trees for 100 kb blocks

Only **53%** group populations by species. **42%** group populations by geography!

Bootstrapping over millions of sites will only ever support the most common tree.



Martin et al. 2013 Genome Research

Phylogenomics: YESTERDAY!

Published Online November 27 2014

Science DOI: 10.1126/science.1258524

< *Science* Express Index

Leave a comment (0)

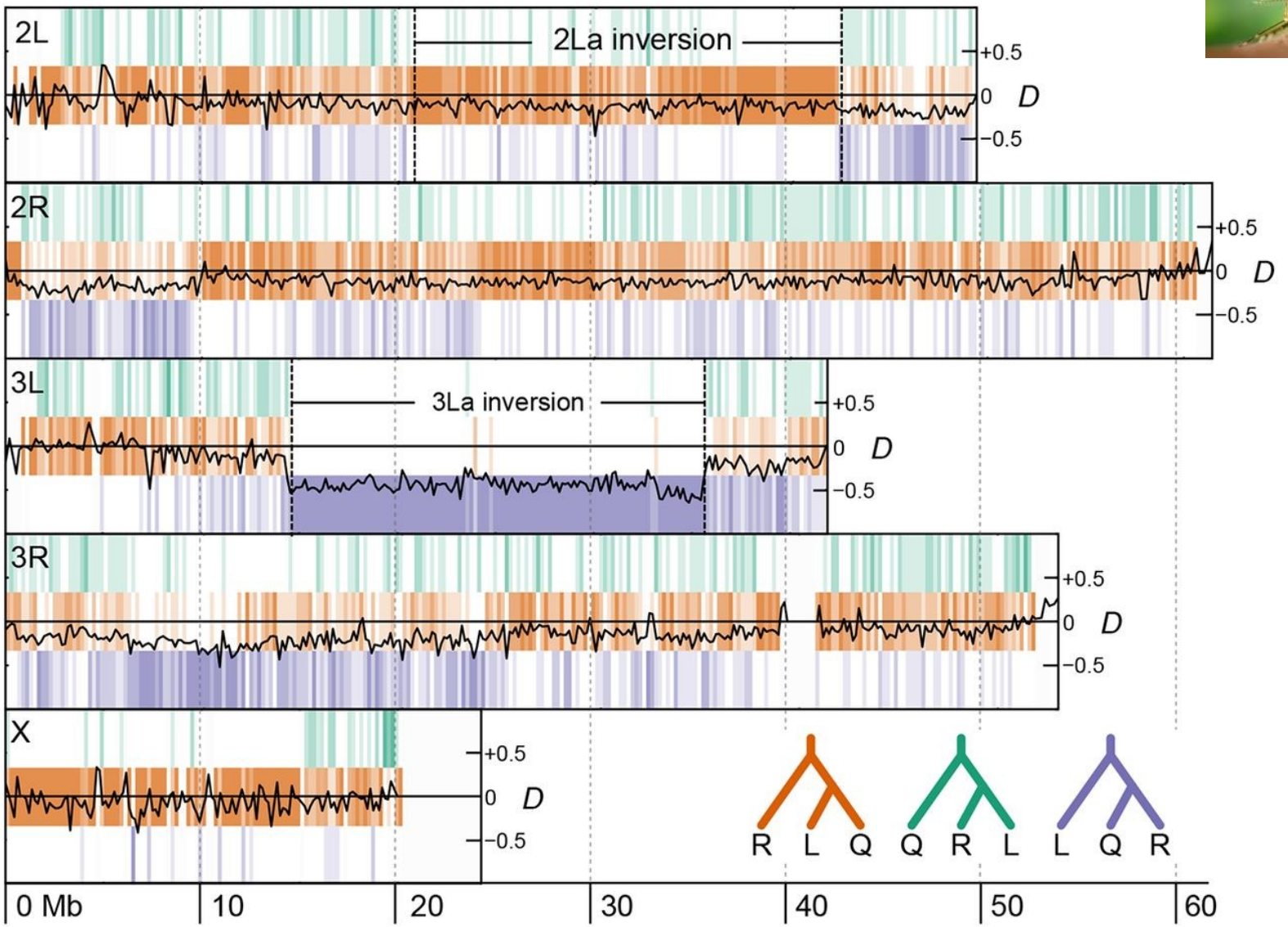
RESEARCH ARTICLE

Extensive introgression in a malaria vector species complex revealed by phylogenomics

Michael C. Fontaine[1,2,*,†], James B. Pease[3,*], Aaron Steele[4], Robert M. Waterhouse[5,6,7,8], Daniel E. Neafsey[6], Igor V. Sharakhov[9,10], Xiaofang Jiang[10], Andrew B. Hall[10], Flaminia Catteruccia[11,12], Evdoxia Kakani[11,12], Sara N. Mitchell[11], Yi-Chieh Wu[5], Hilary A. Smith[1,2], R. Rebecca Love[1,2], Mara K. Lawniczak[13,‡], Michel A. Slotman[14], Scott J. Emrich[2,4], Matthew W. Hahn[3,15,§], Nora J. Besansky[1,2,§]

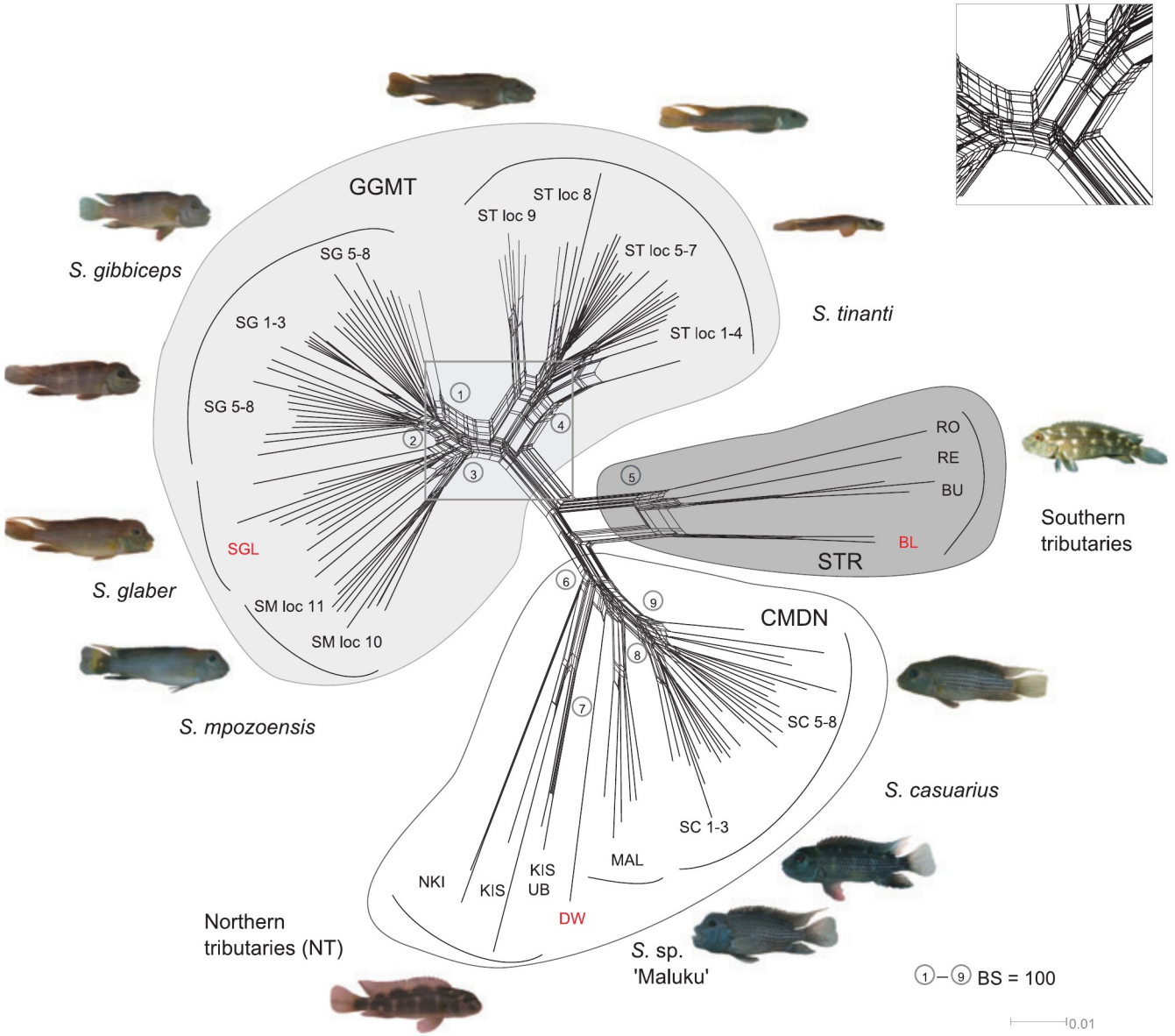
Phylogenomics: YESTERDAY!

wikipedia.org



Fontaine et al. 2014 Science

Phylogenomics: networks are more honest



Schwarzer et al. 2012 JEB