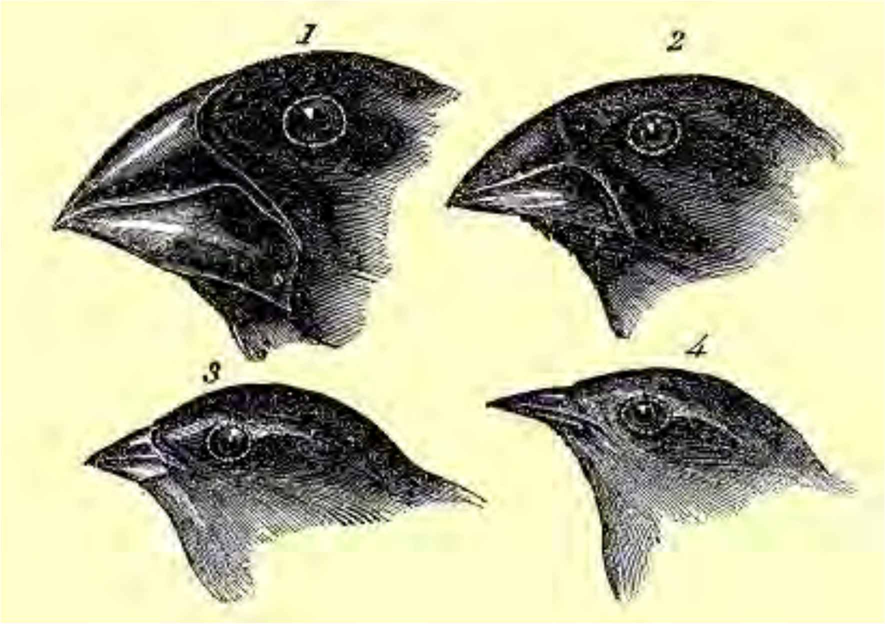
Genomic studies of speciation and gene flow



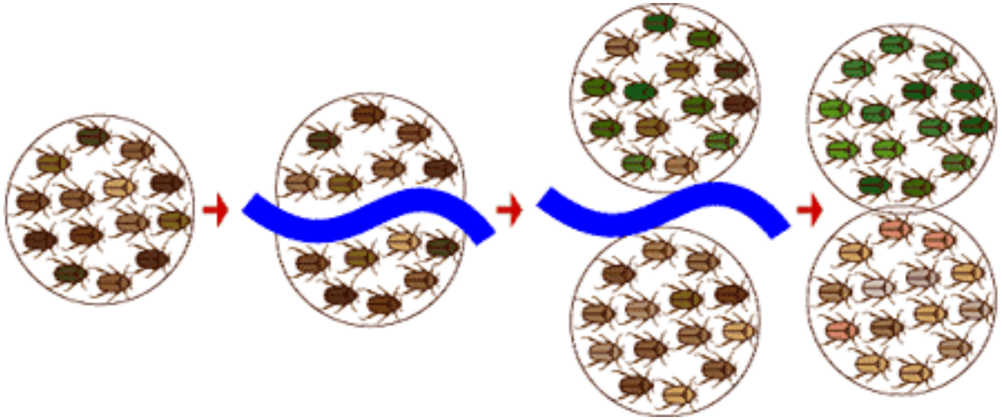
Why study speciation genomics?

Long-standing questions (role of geography/gene flow)

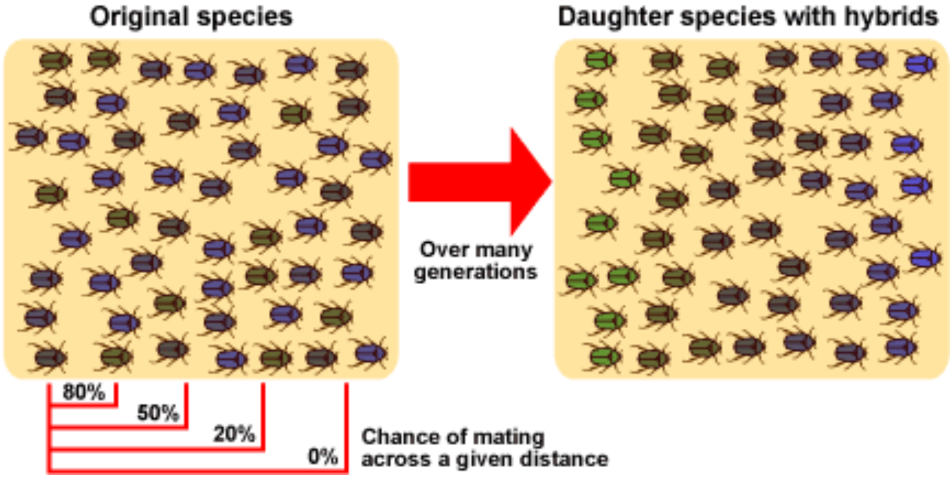
How do genomes diverge?

Find speciation genes

Genomic divergence during speciation



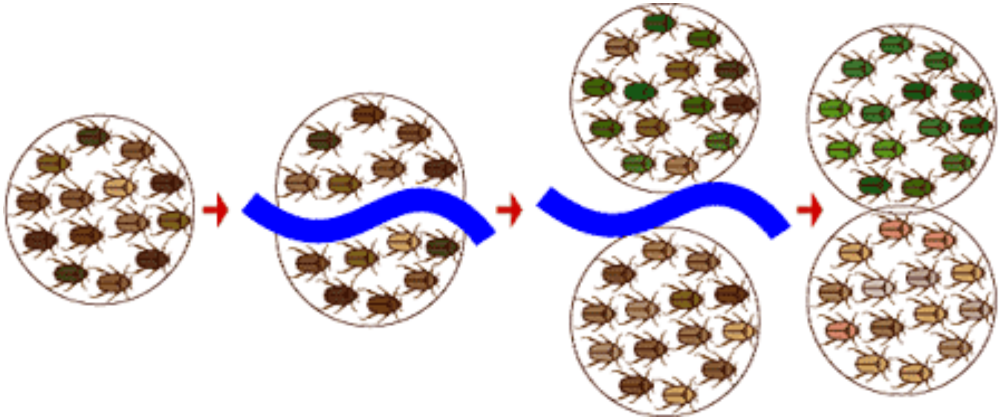
1. Speciation as a bi-product of physical isolation



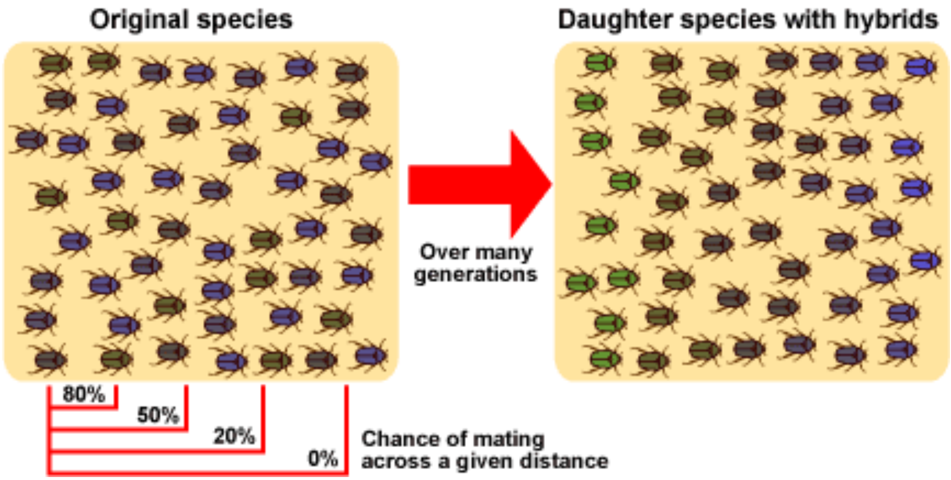
2. Speciation due to selection – without isolation

evolution.berkeley.edu

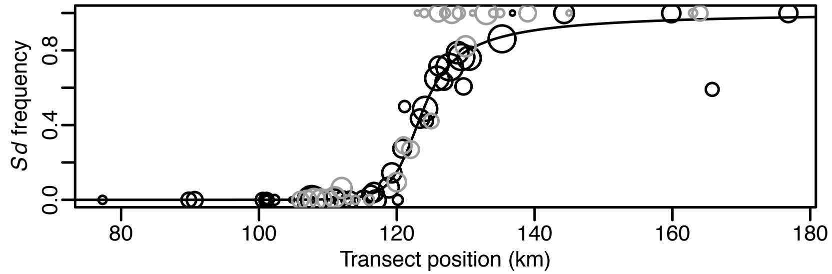
Genomic divergence during speciation



1. Speciation as a bi-product of physical isolation



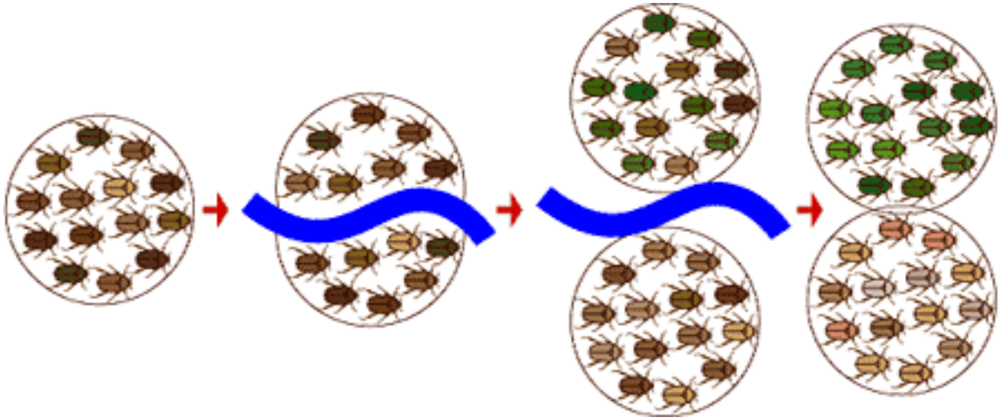
2. Speciation due to selection – without isolation



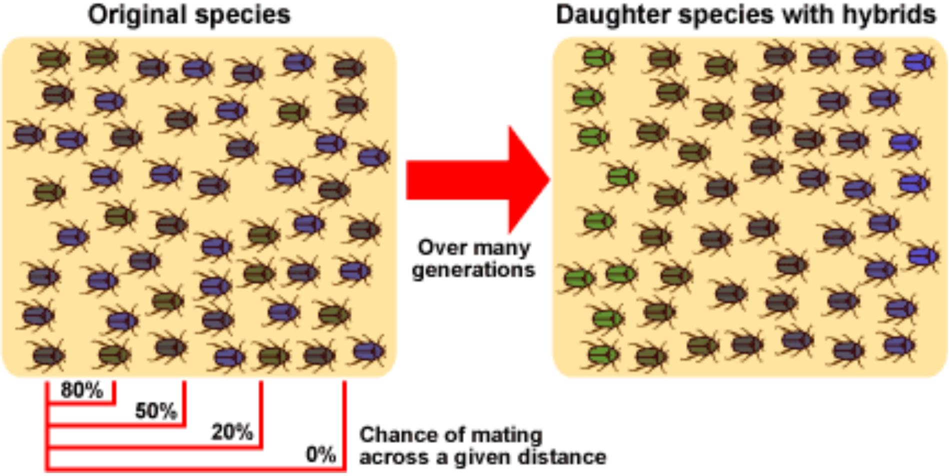
Cline theory - e.g. Barton and Gale 1993

evolution.berkeley.edu

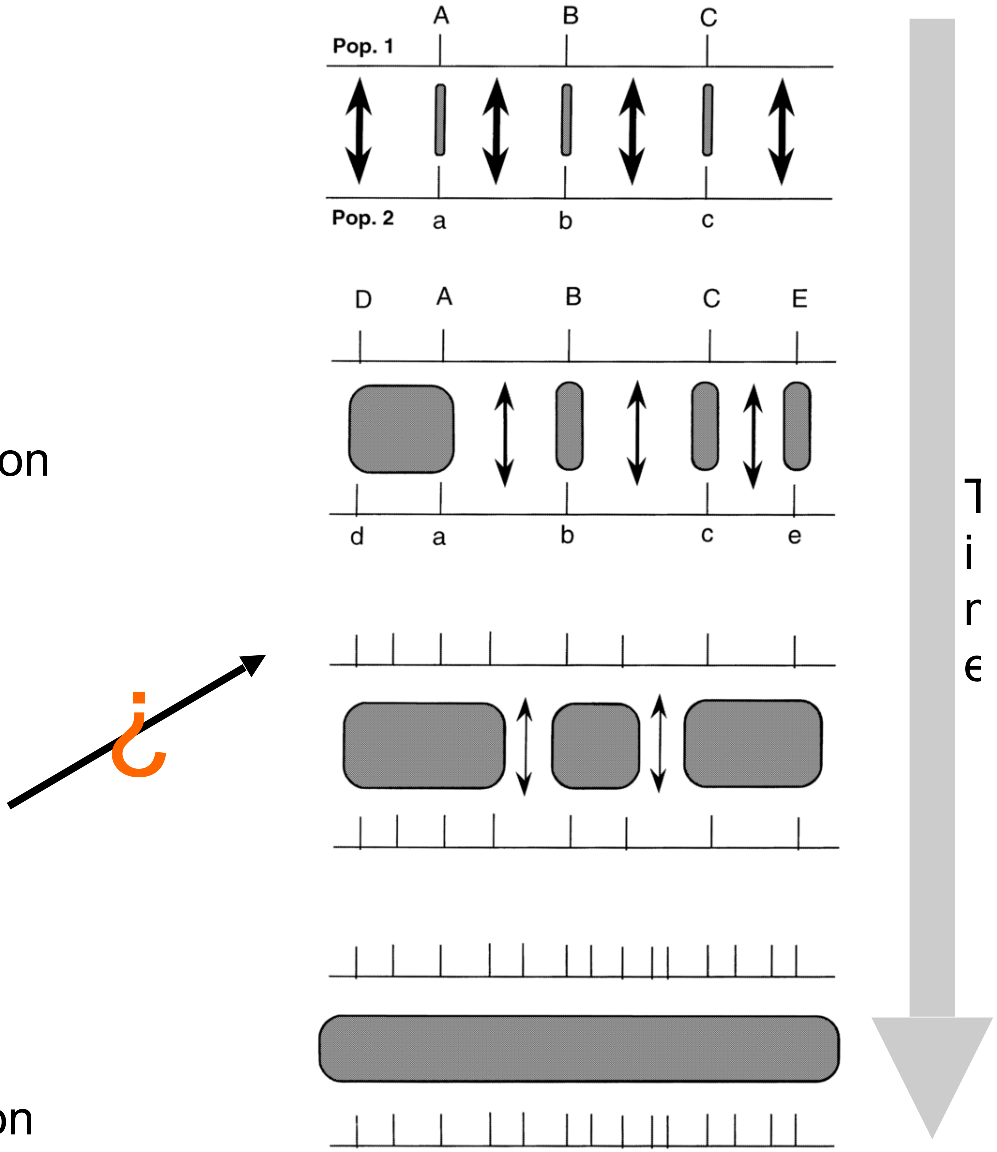
Genomic divergence during speciation



1. Speciation as a bi-product of physical isolation



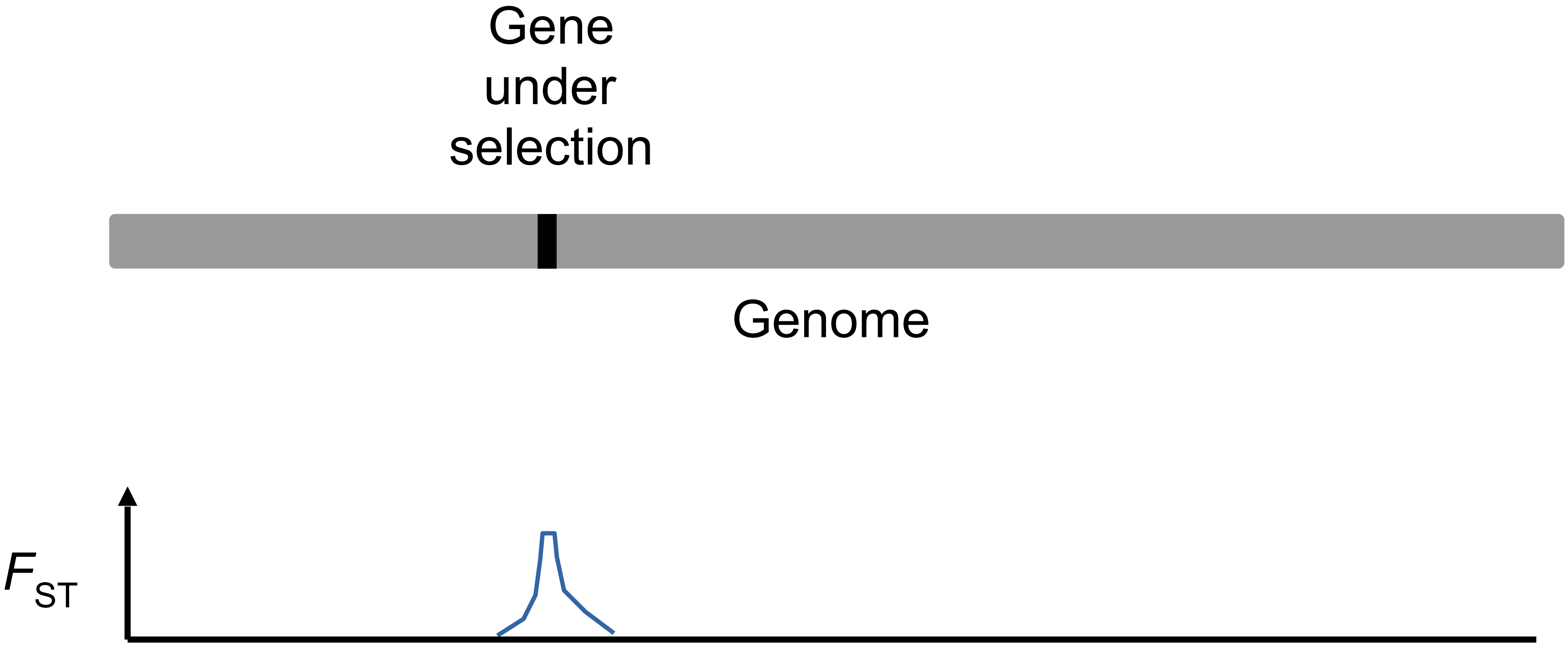
2. Speciation due to selection – without isolation



evolution.berkeley.edu

Wu 2001, JEB

Stage 1 - one or few loci under disruptive selection



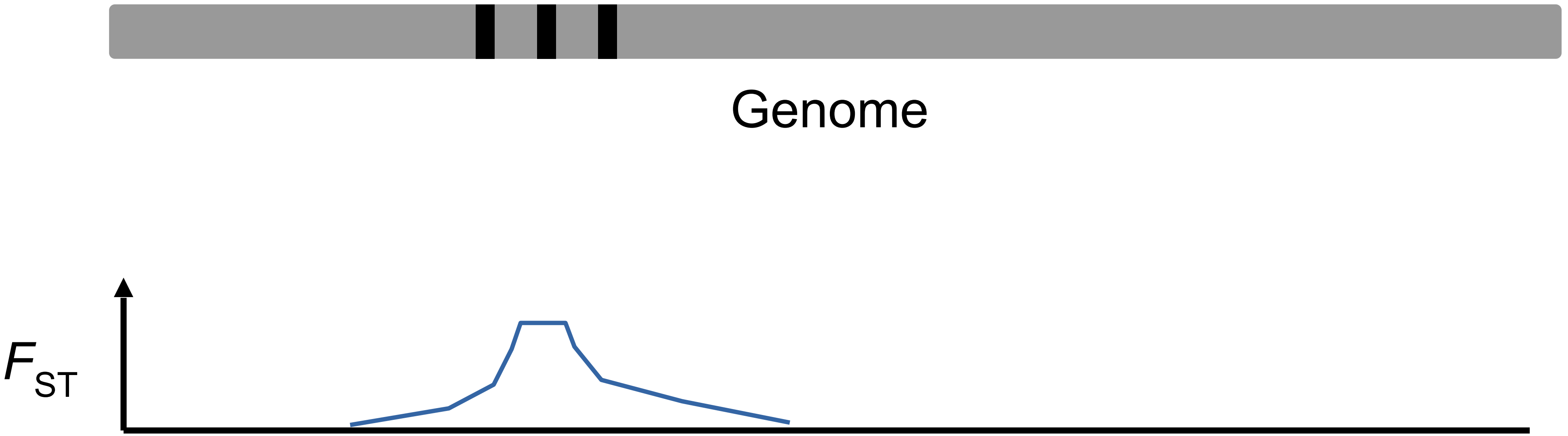
Gene under selection

Genome

F_{ST}

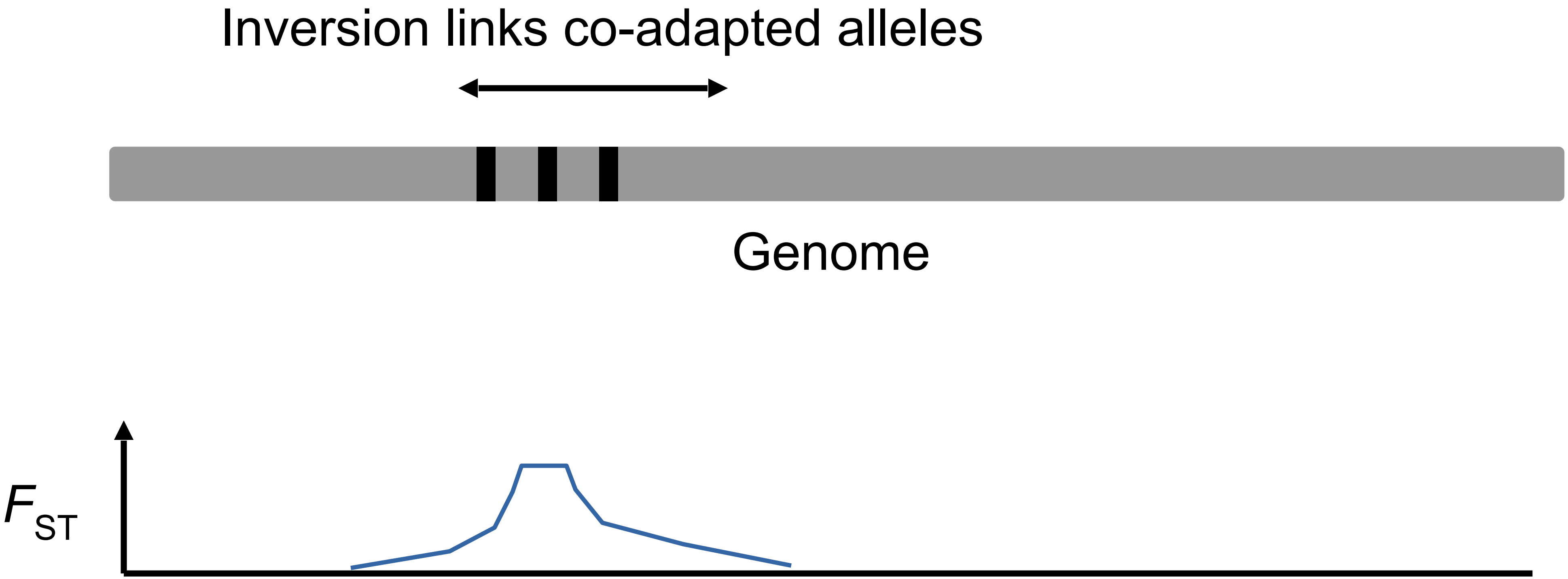
Feder, Egan and Nosil TiG

Stage 2 - Divergence hitchhiking



Feder, Egan and Nosil TiG

Stage 2b - Inversion



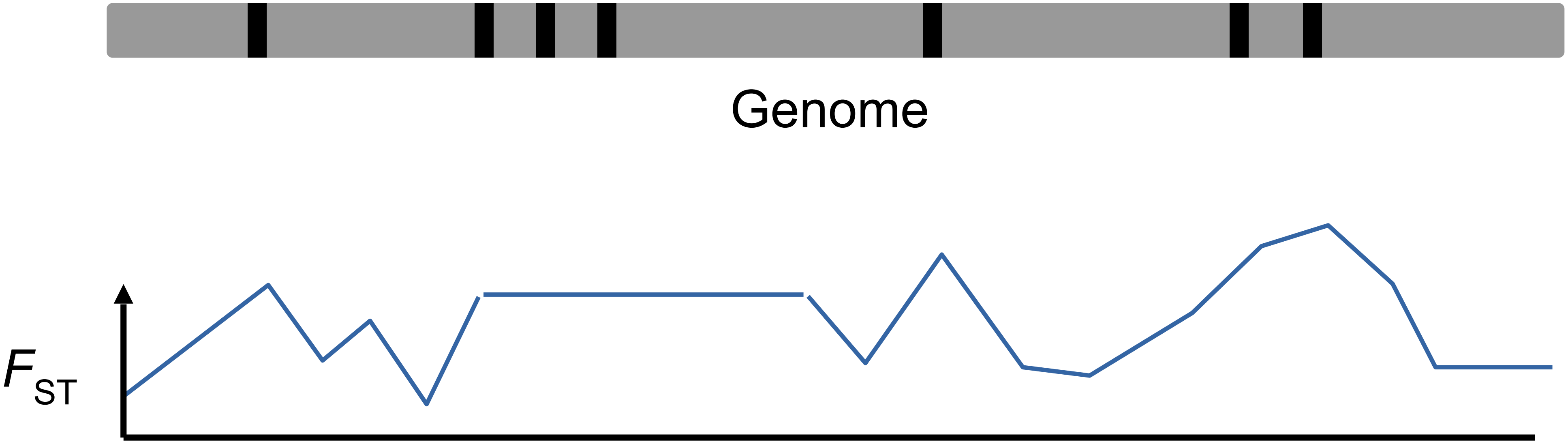
Inversion links co-adapted alleles

Genome

F_{ST}

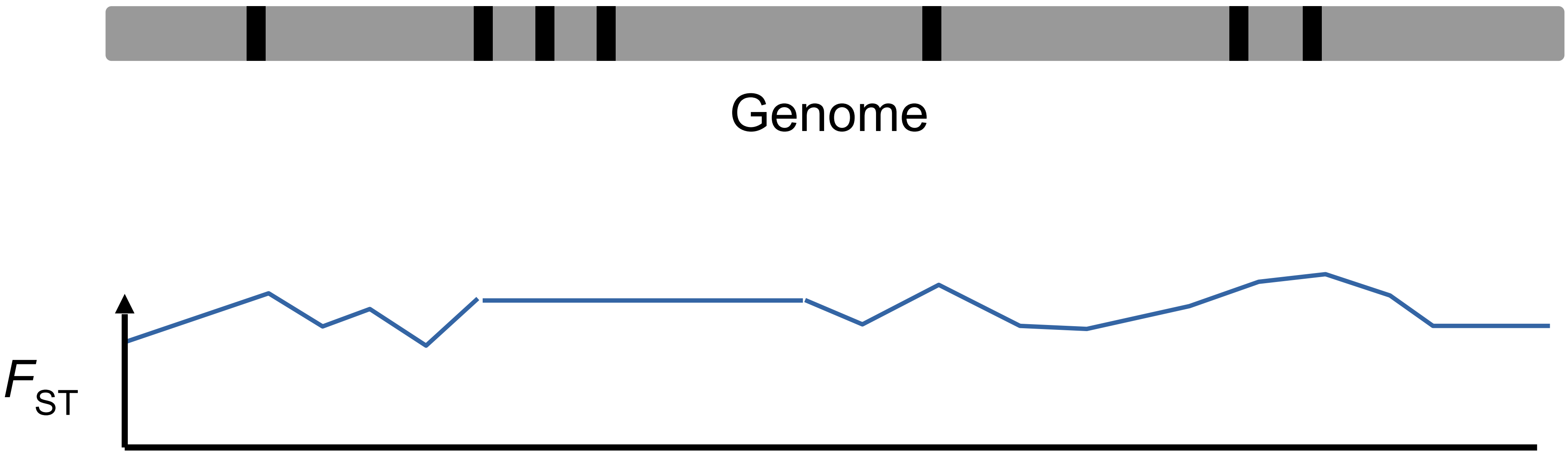
Feder, Egan and Nosil TiG

Stage 3 - Genome hitchhiking



Feder, Egan and Nosil TiG

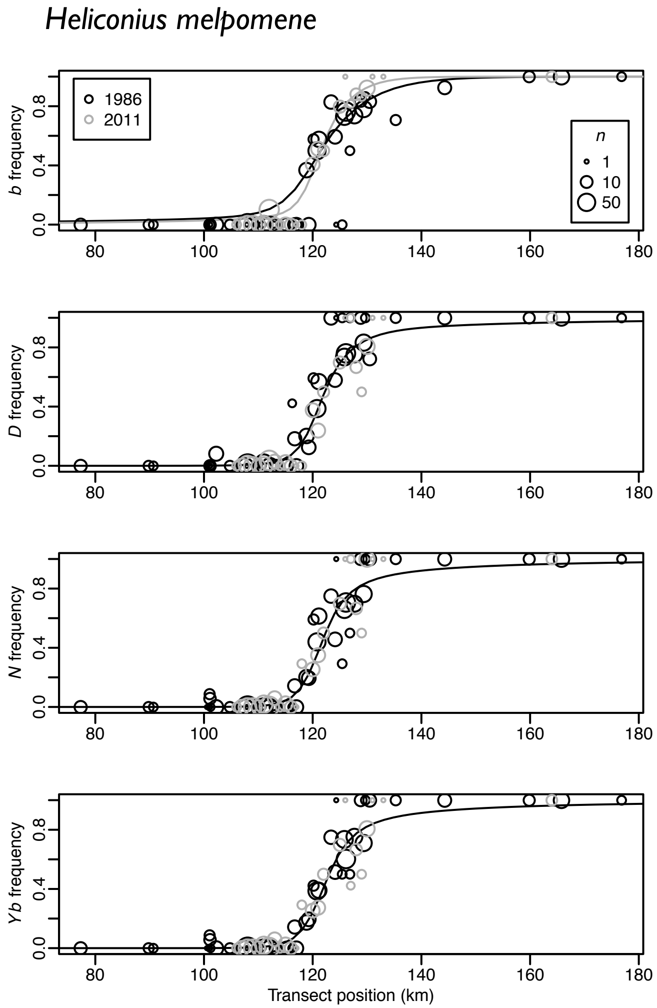
Stage 4 - Genome wide isolation

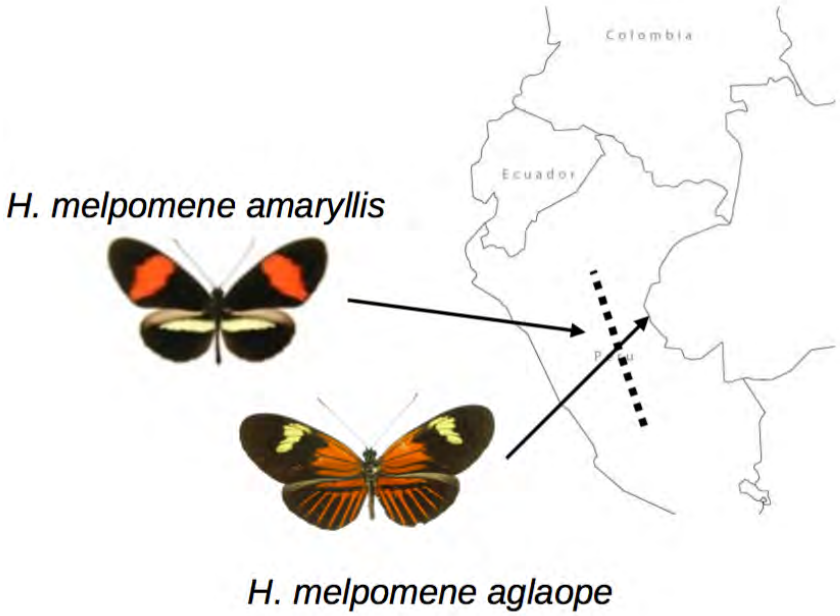


Feder, Egan and Nosil TiG

Some sub-species clearly in stage 1

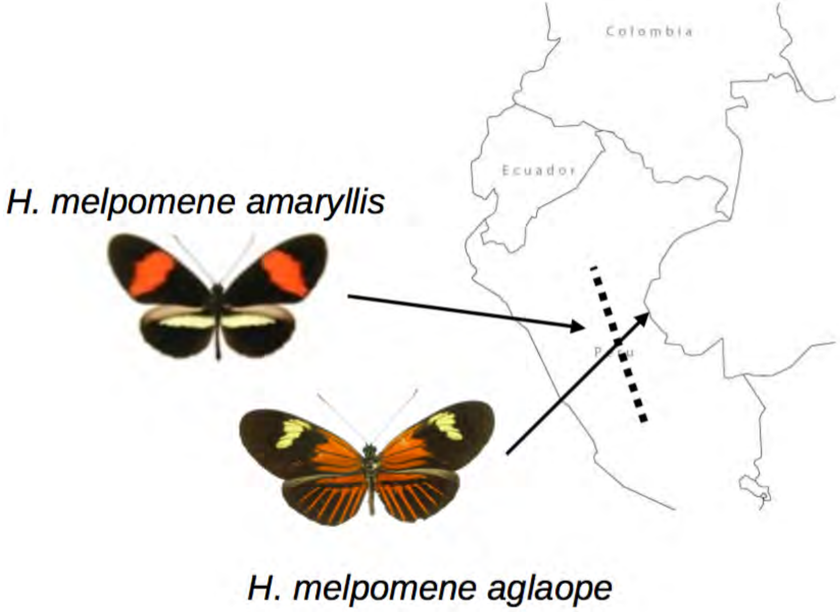
- Wing pattern “races” of *Heliconius melpomene*

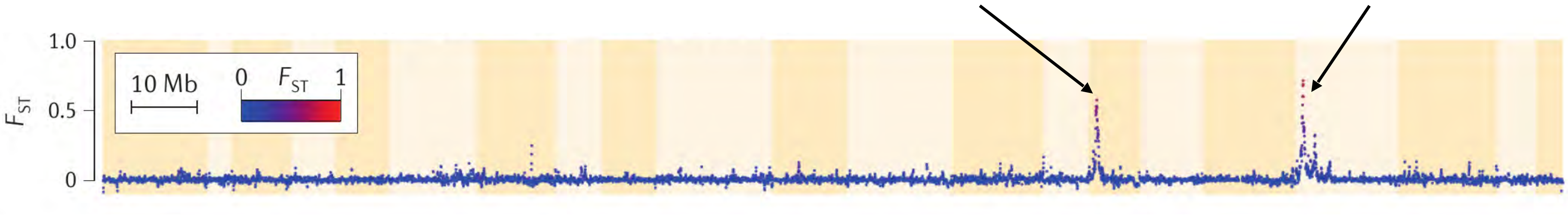




Some sub-species clearly in stage 1

- Wing pattern “races” of *Heliconius melpomene*

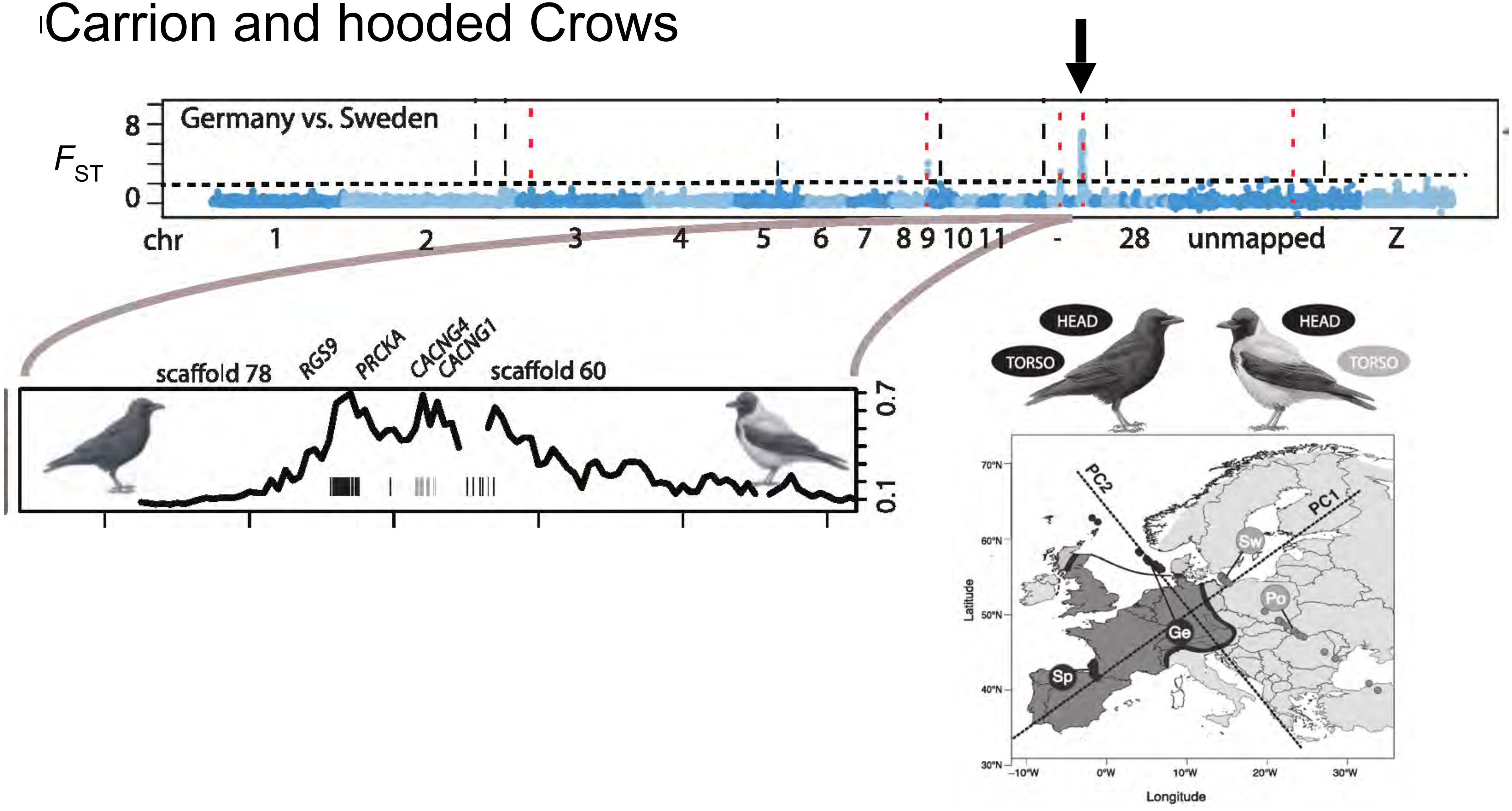




S. H. Martin et al. Genome Res. 23, 1817–1828 (2013).
O. Seehausen et al. Nat. Rev. Genet. 15, 176–92 (2014).

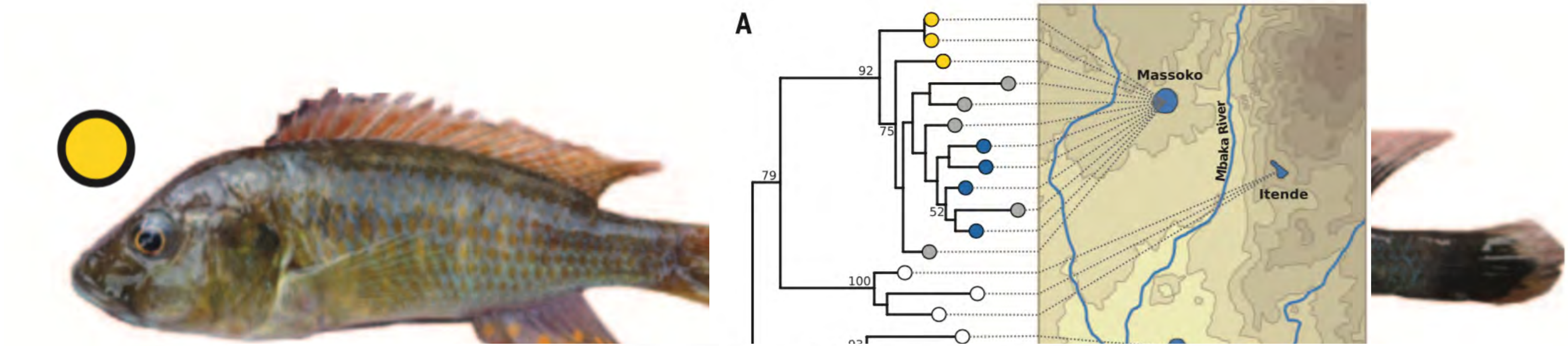
Some sub-species clearly in stage 1

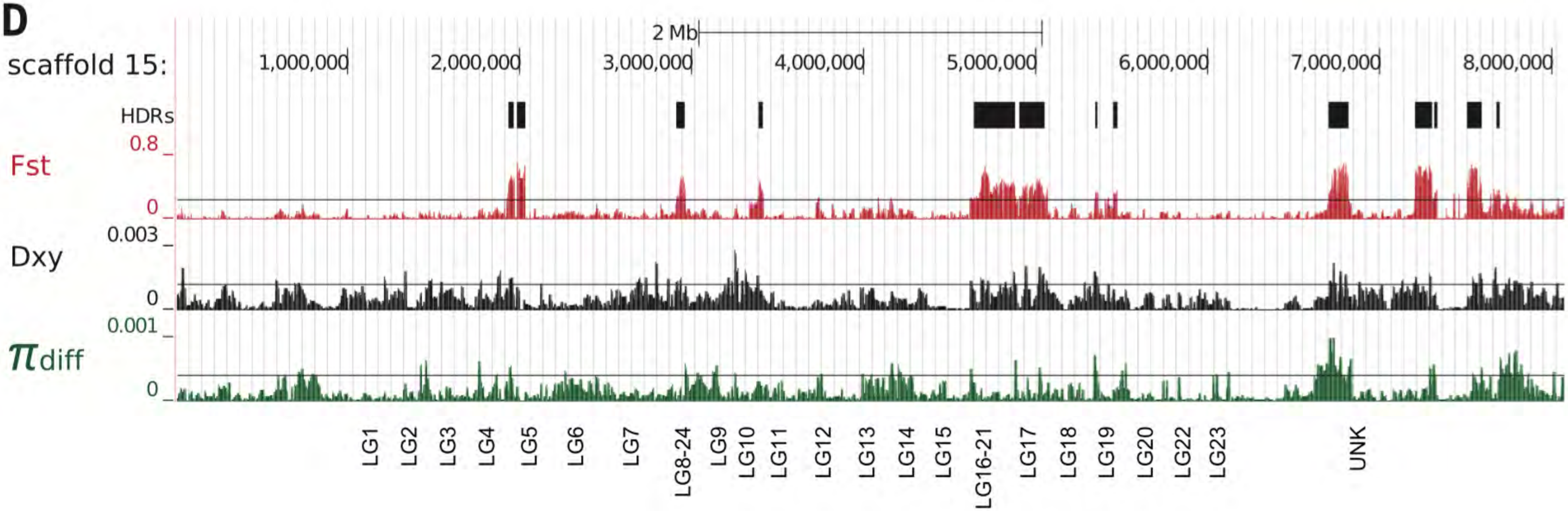
- Carrion and hooded Crows



Poelstra, J. W. et al. Science 344, 1410–4 (2014).

And an example with multiple islands?

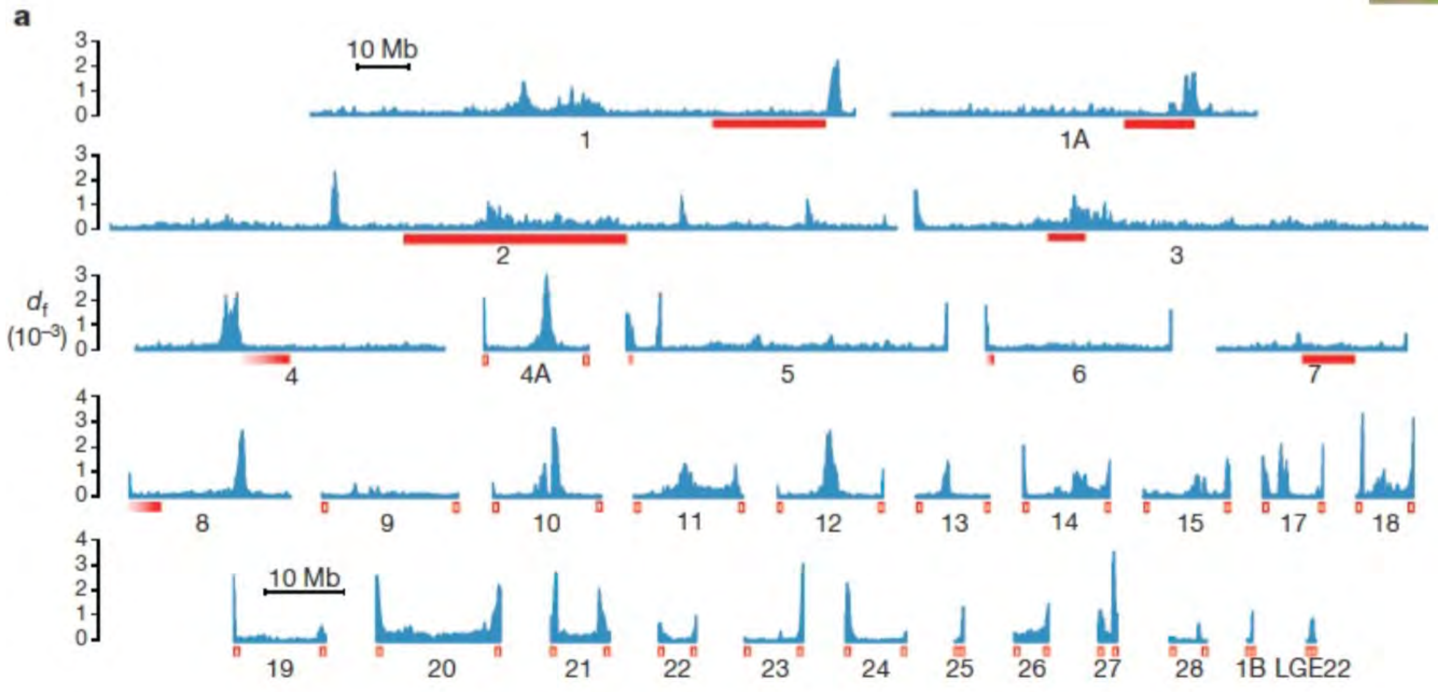




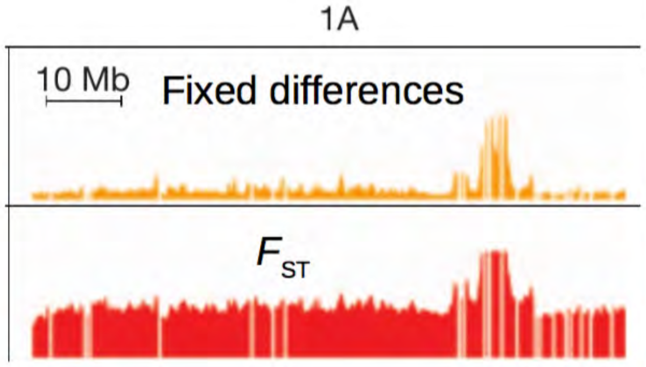
Malinsky et al., Science 350, 1493 (2015).

Other species have islands…but are they real?

Collared and Pied Flycatchers



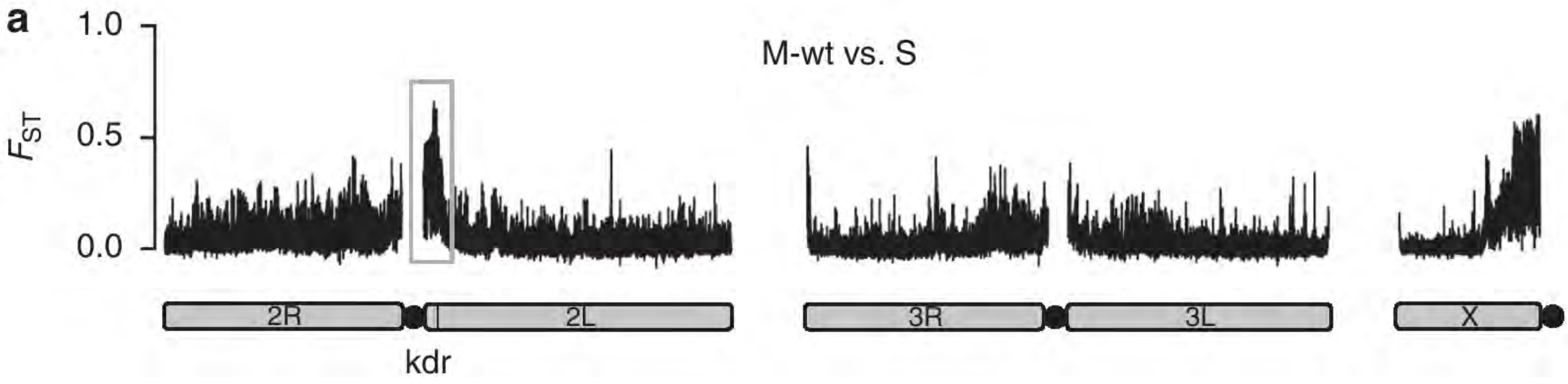
Fixed differences



Ellegren, et al. Nature 491, 756- (2012).

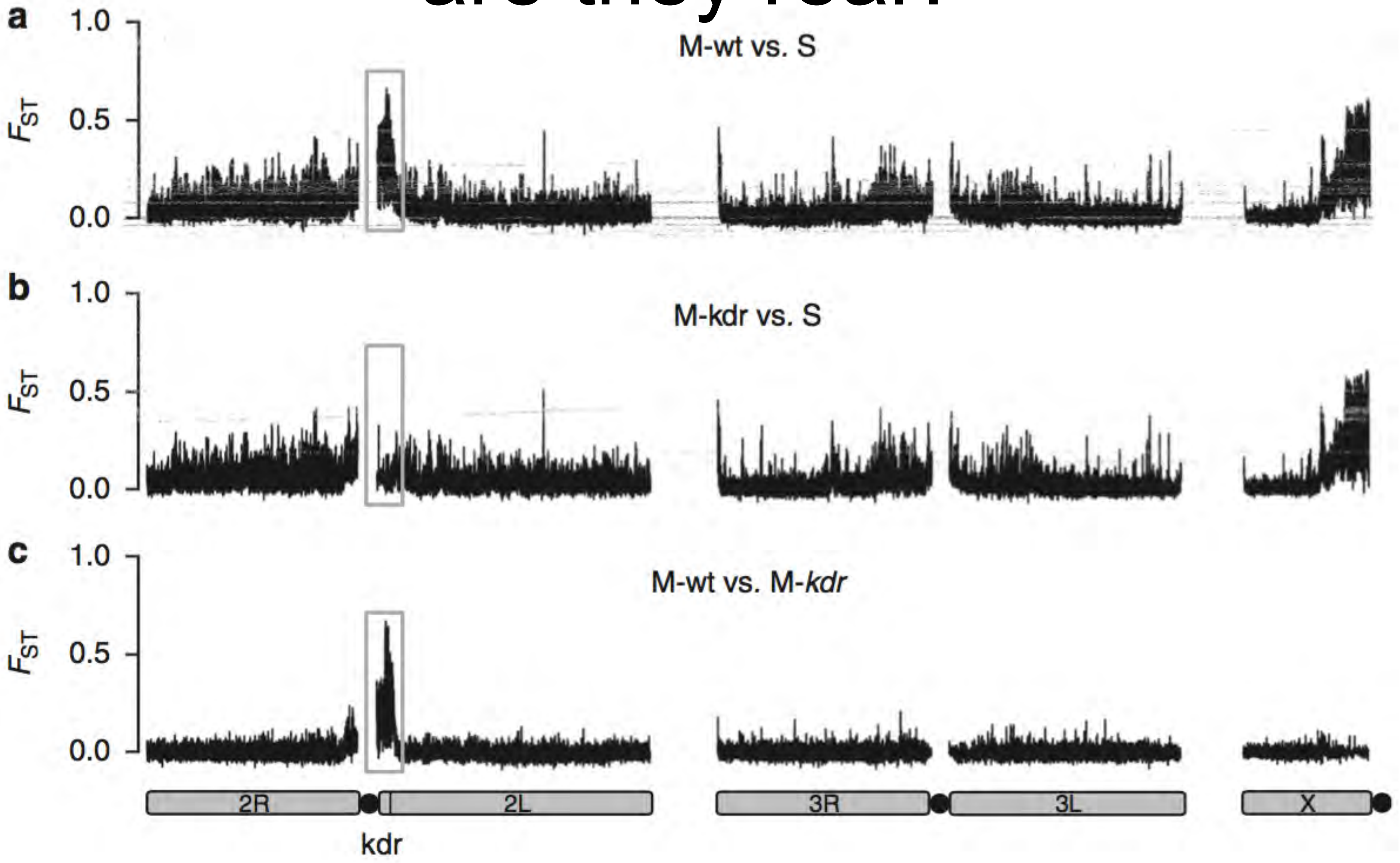
Other species have islands…but are they real?

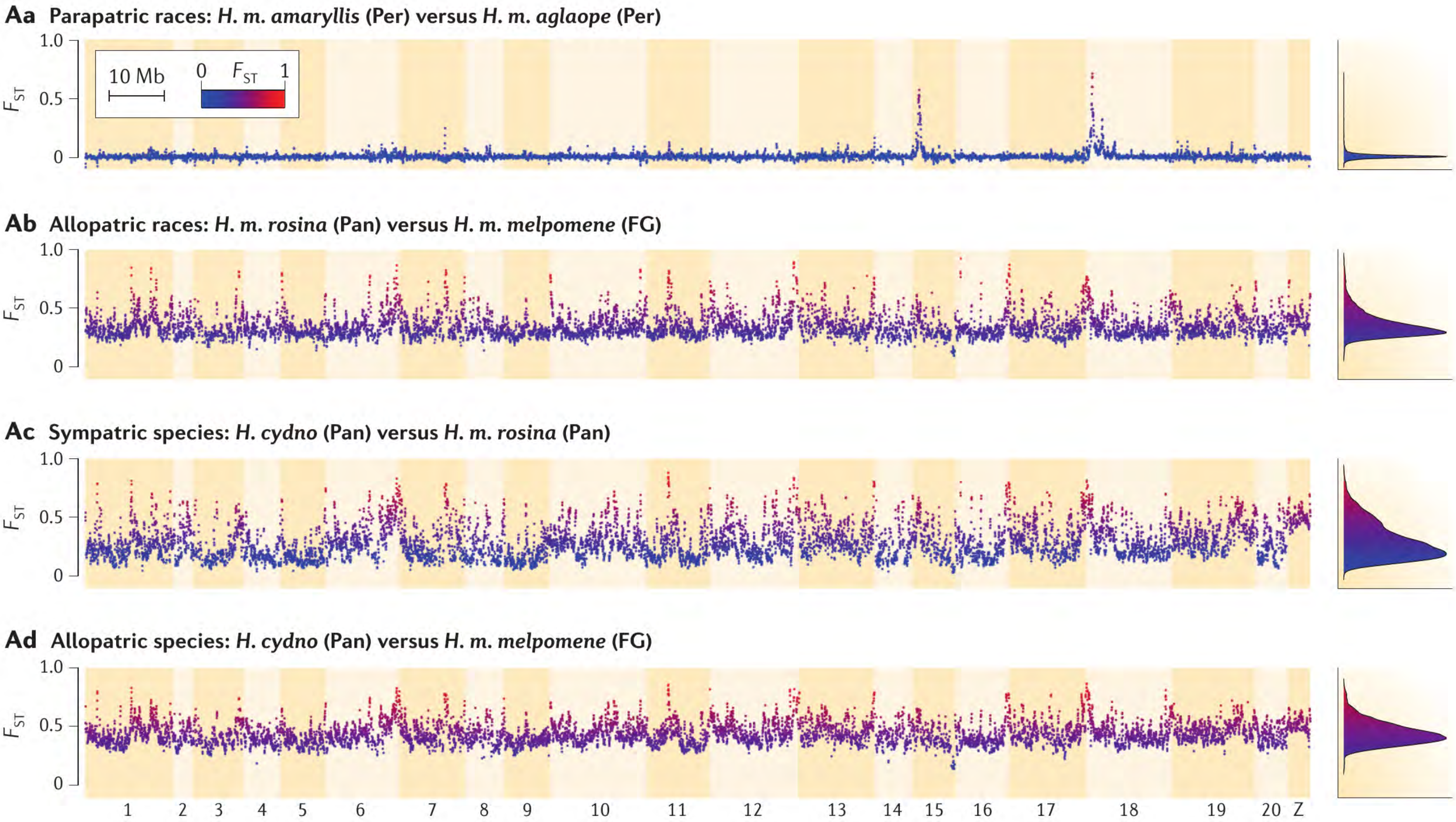
Anopheles gambiae and *A. coluzzi*
Formerly M and S forms of A. gambiae



Clarkson et al. 2014 Nature Communications

Other species have islands…but are they real?





Aa Parapatric races: *H. m. amaryllis* (Per) versus *H. m. aglaope* (Per)

Ab Allopatric races: *H. m. rosina* (Pan) versus *H. m. melpomene* (FG)

Ac Sympatric species: *H. cydno* (Pan) versus *H. m. rosina* (Pan)

Ad Allopatric species: *H. cydno* (Pan) versus *H. m. melpomene* (FG)

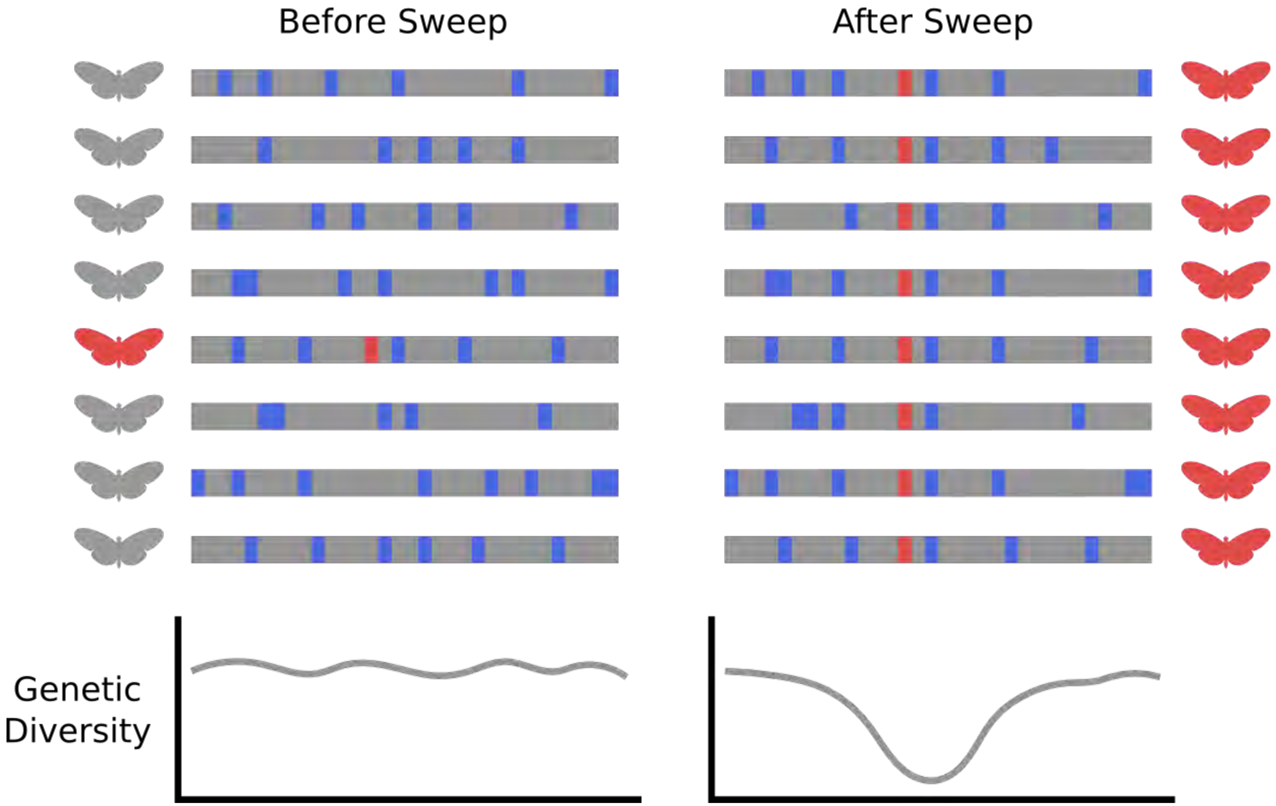
Seehausen et al., Nature Reviews Genetics, 2014

What do patterns of F_{st} really mean?

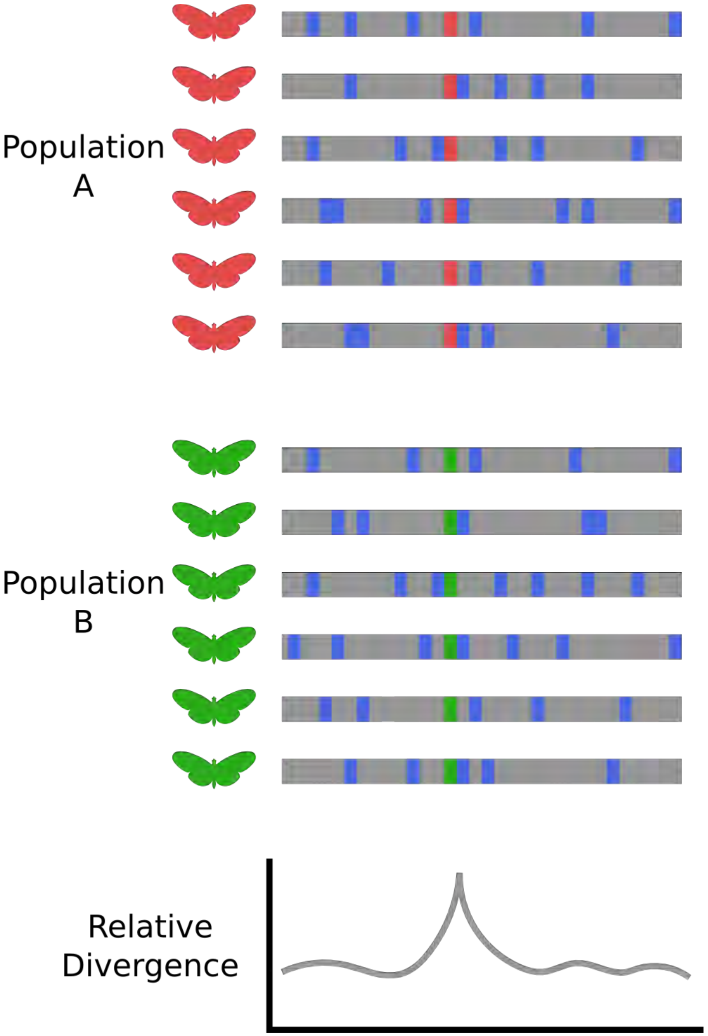
- F_{st} measures relative divergence

$$F_{ST} = \frac{H_T - H_S}{H_T},$$

- Peaks indicate regions of higher than expected between population divergence, given the within population divergence
- Peaks can therefore result from reduced diversity within species
- This could be due to lower Ne within species (selective sweeps, background selection)
- So peaks NOT NECESSARILY due to reduced gene flow



Note that sometimes sweeps within species = speciation genes



Sweeps across the species barrier can also lead to Fst peaks

Double peaks??

Nicolas Bierne, Daniel Berner and others

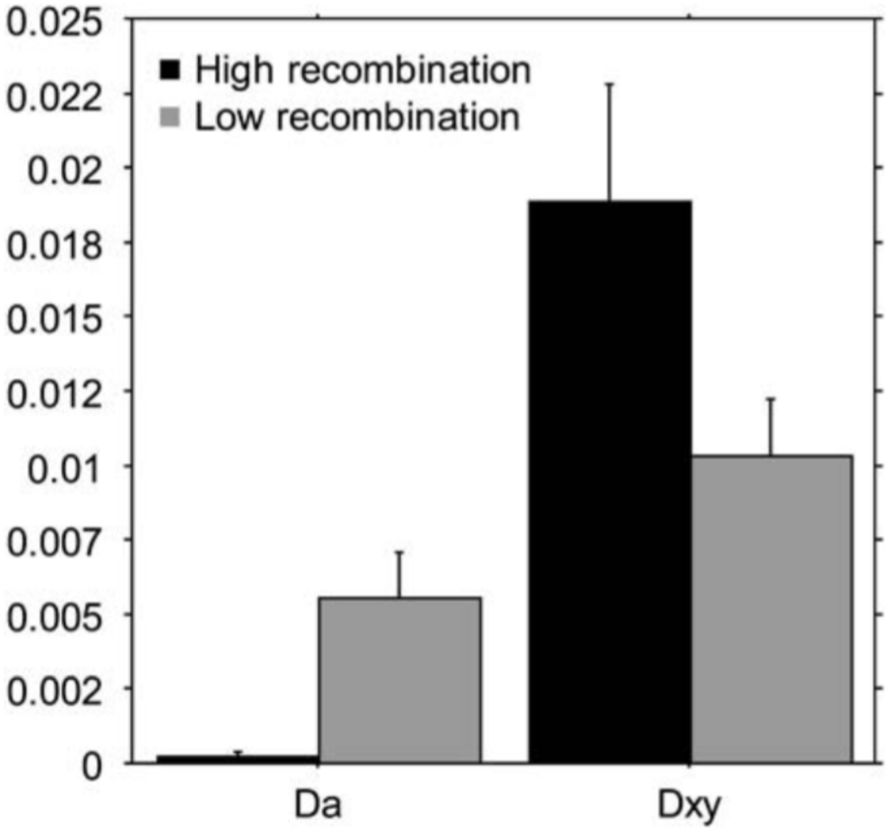
Heredity (2009) 103, 439–444

www.nature.com/hdy

REVIEW

Islands of speciation or mirages in the desert? Examining the role of restricted recombination in maintaining species

MAF Noor and SM Bennett
Biology Department, Duke University, Durham, NC, USA



Anopheles M-S divergence

Relative divergence higher in low recombination regions - not significant for absolute divergence

see also: Charlesworth 1998 MBE Measures of divergence…
More recently see papers by Reto Burri

MOLECULAR ECOLOGY

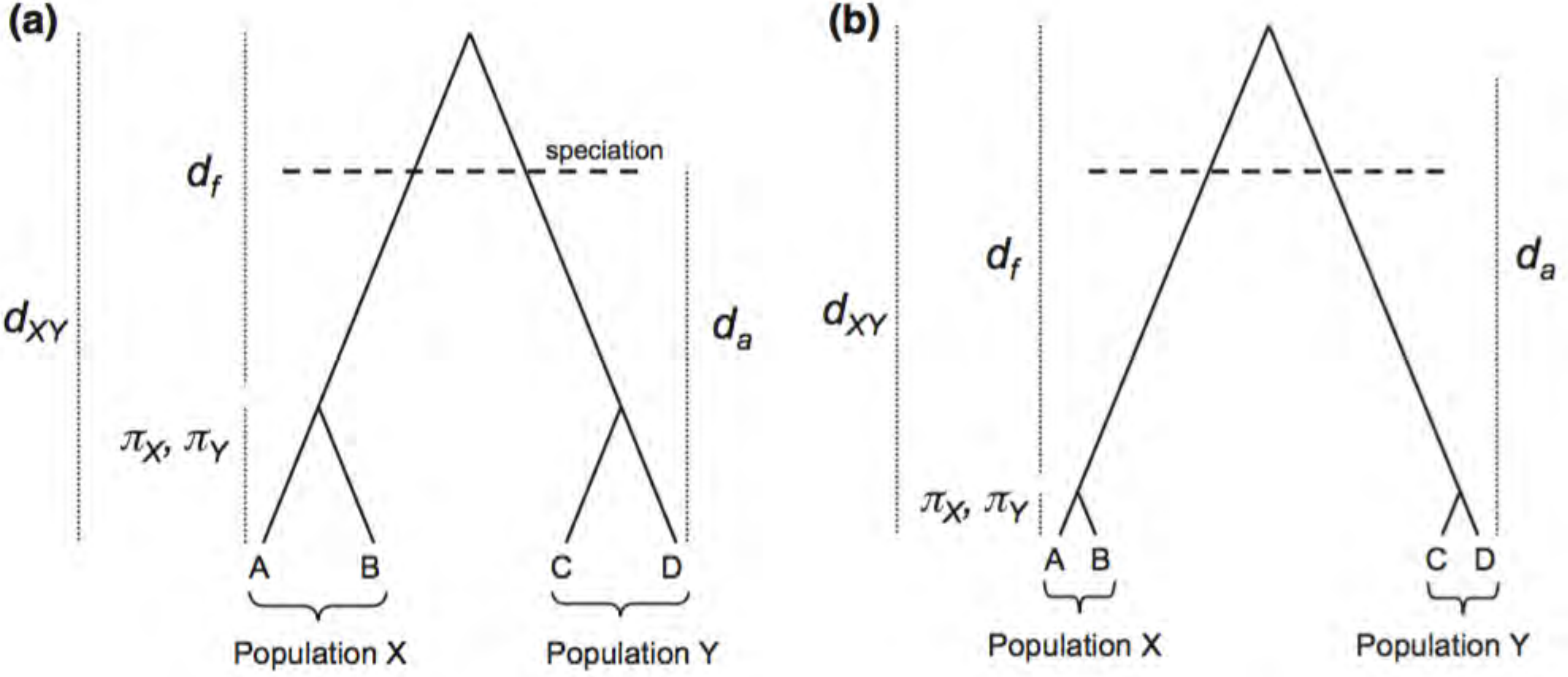
Molecular Ecology (2014) **23**, 3133–3157 doi: 10.1111/mec.12796

INVITED REVIEWS AND SYNTHESES

Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow

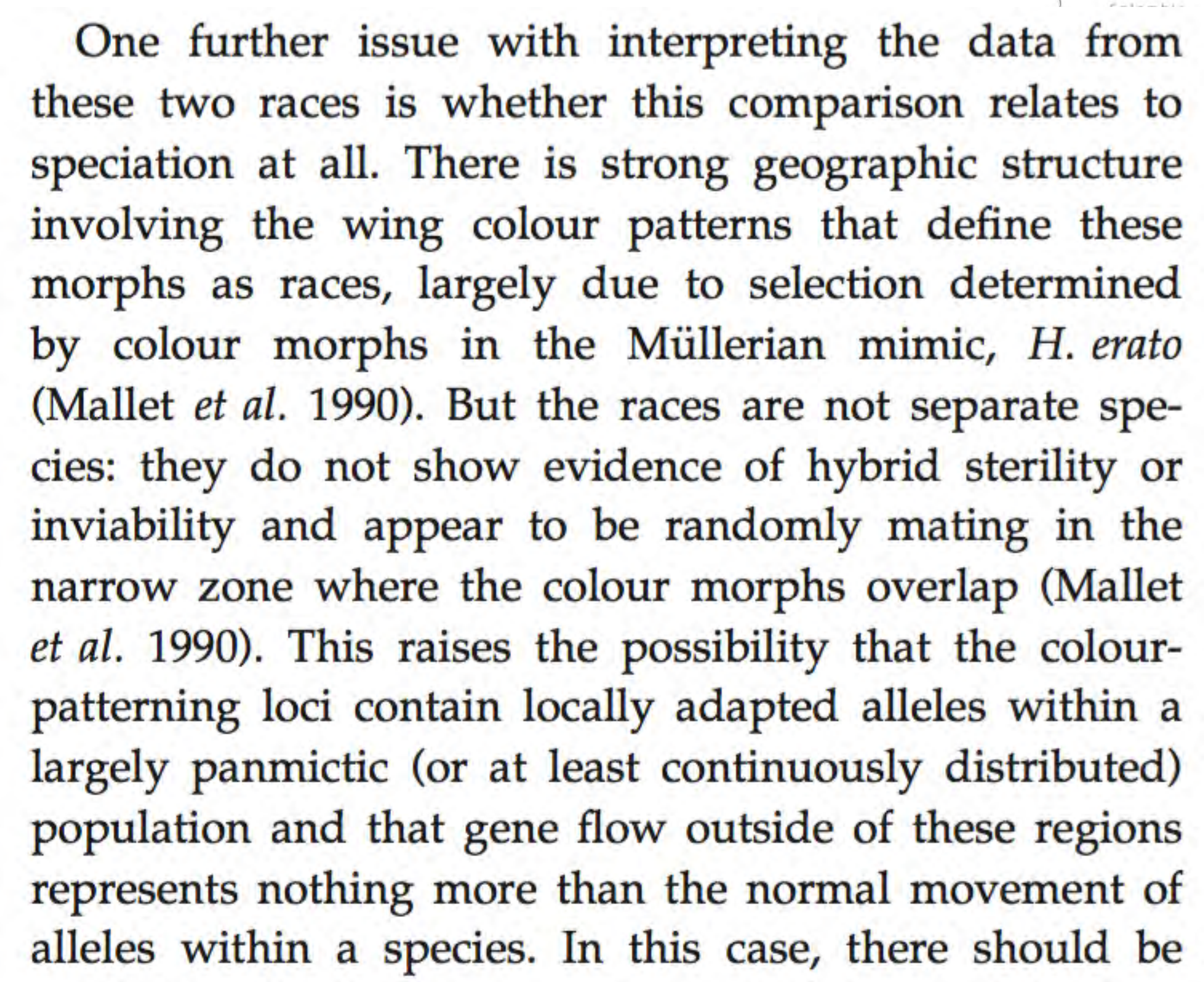
TAMI E. CRUICKSHANK* and MATTHEW W. HAHN*†

**Department of Biology, Indiana University, Bloomington, IN 47405, USA, †School of Informatics and Computing, Indiana University, Bloomington, IN 47405, USA*



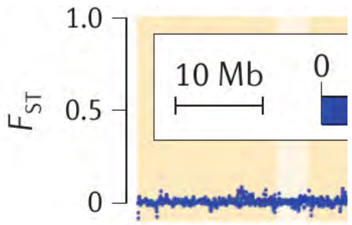
No evidence for higher Dxy in wing pattern loci

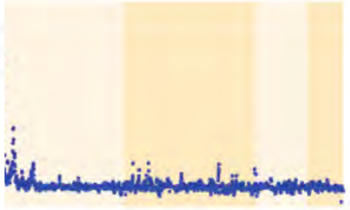
Wing pat
Heliconiu



One further issue with interpreting the data from these two races is whether this comparison relates to speciation at all. There is strong geographic structure involving the wing colour patterns that define these morphs as races, largely due to selection determined by colour morphs in the Müllerian mimic, *H. erato* (Mallet *et al.* 1990). But the races are not separate species: they do not show evidence of hybrid sterility or inviability and appear to be randomly mating in the narrow zone where the colour morphs overlap (Mallet *et al.* 1990). This raises the possibility that the colour-patterning loci contain locally adapted alleles within a largely panmictic (or at least continuously distributed) population and that gene flow outside of these regions represents nothing more than the normal movement of alleles within a species. In this case, there should be

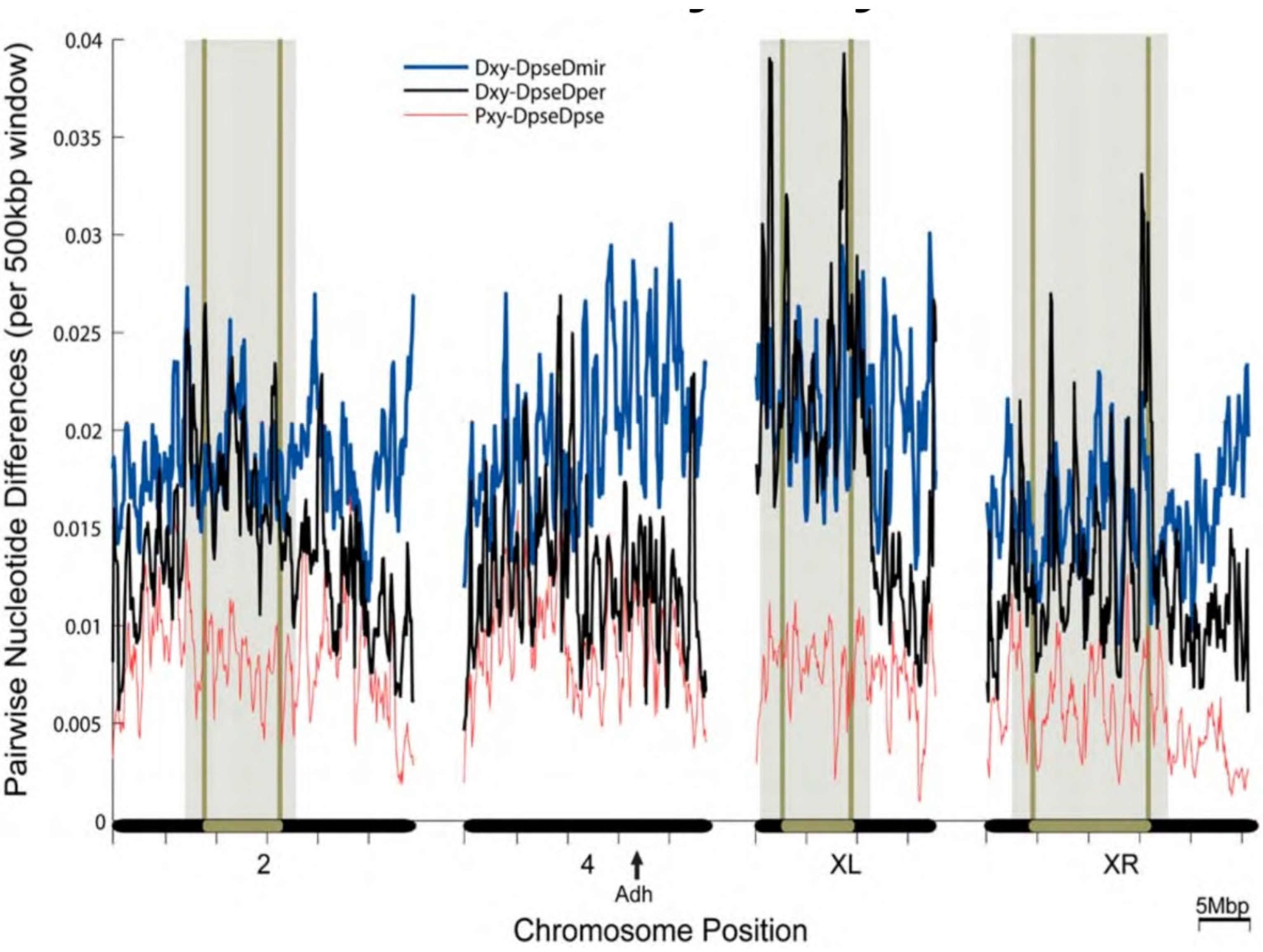
/white patterns)





8 (2013).
92 (2014).

Suggestion that we use absolute measures of divergence?



Understanding genomic divergence

No single statistic will capture the complex history of mutation, migration and selection

Patterns need to be interpreted in the specific context of the study species

Much better to use explicit tests for gene flow

Need to design sampling so the expectations in the absence of gene flow are clear and testable

The key is to identify ‘control’ populations that are not influenced by admixture

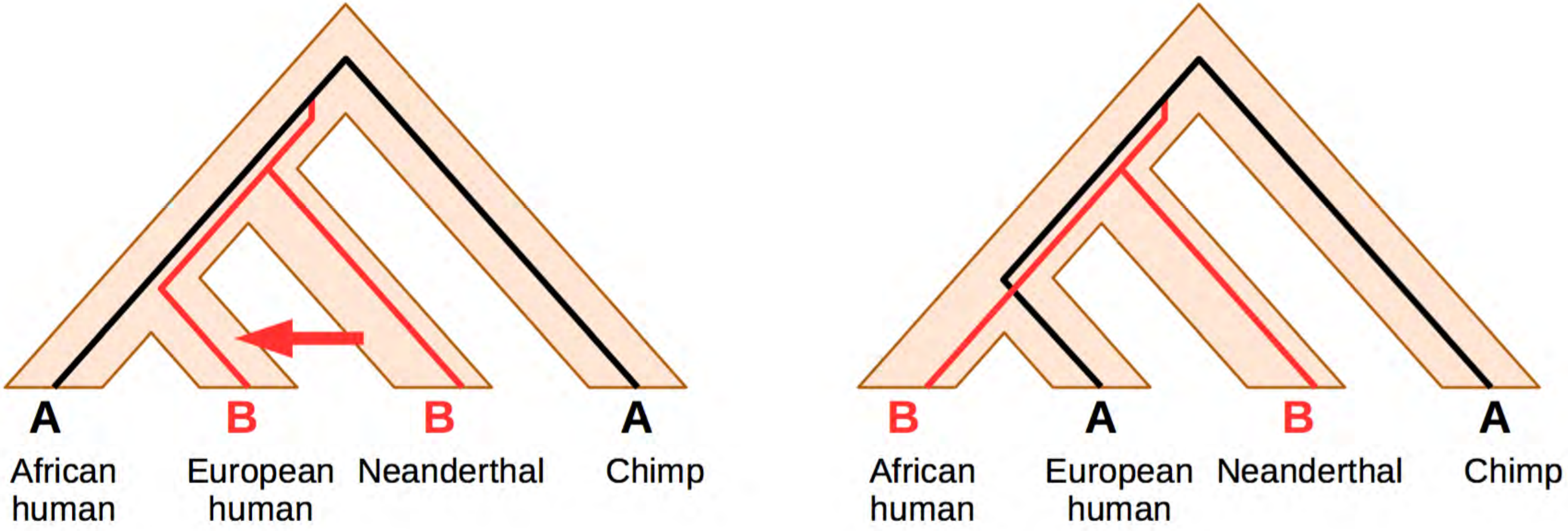
Explicit tests for gene flow: Neanderthal genome

- Isolated DNA from bones 38,000 yrs old in Croatia
- We diverged from Neanderthals around 270-440,000yrs ago

- Evidence for gene exchange with humans (1-4% of genome?)

Green et al., 328:710 Science 2010

Explicit tests for gene flow: ABBA-BABA test

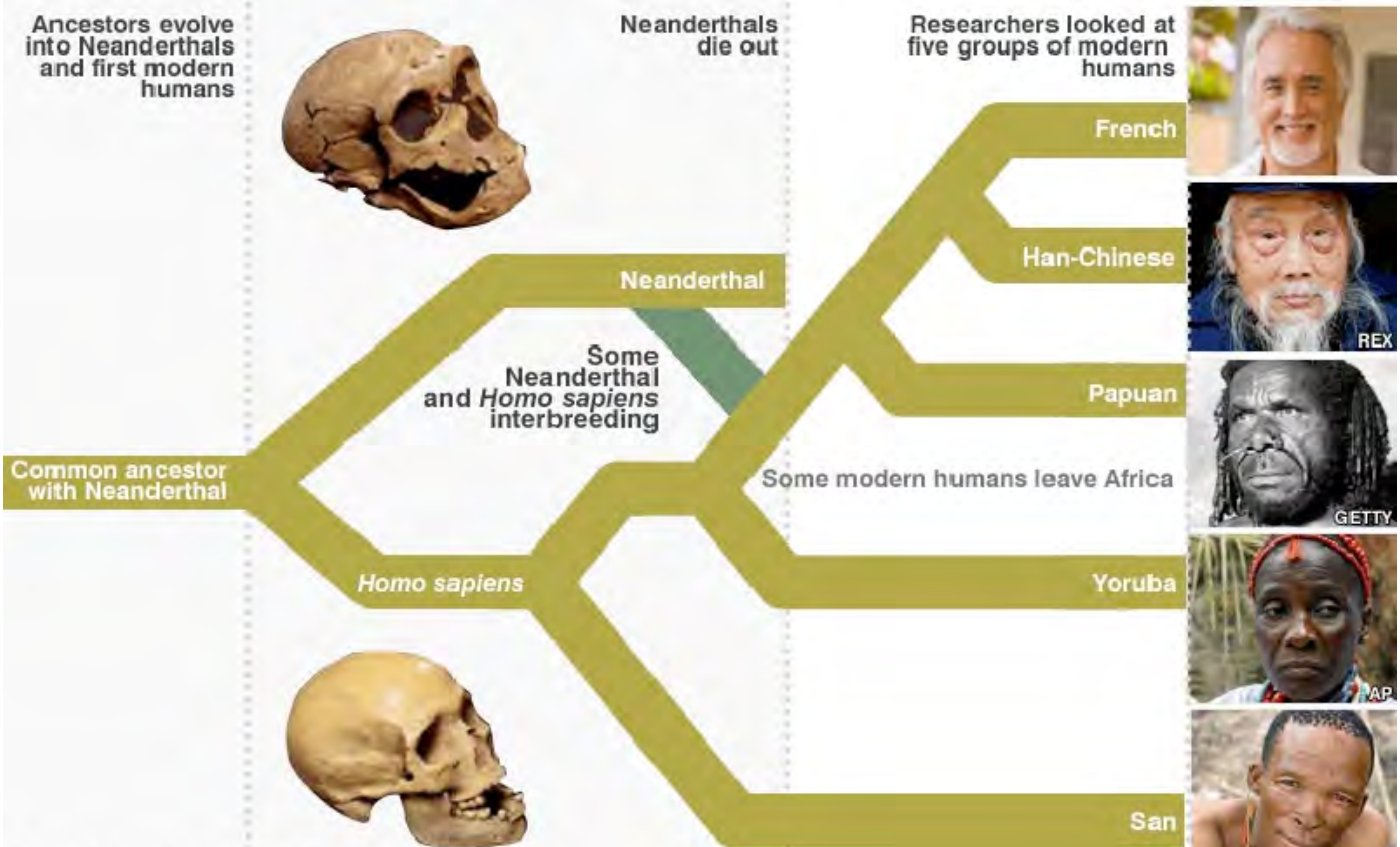


$$D(P_1, P_2, P_3, O) = \frac{\sum C_{ABBA}(i) - C_{BABA}(i)}{\sum C_{ABBA}(i) + C_{BABA}(i)} \quad (1)$$

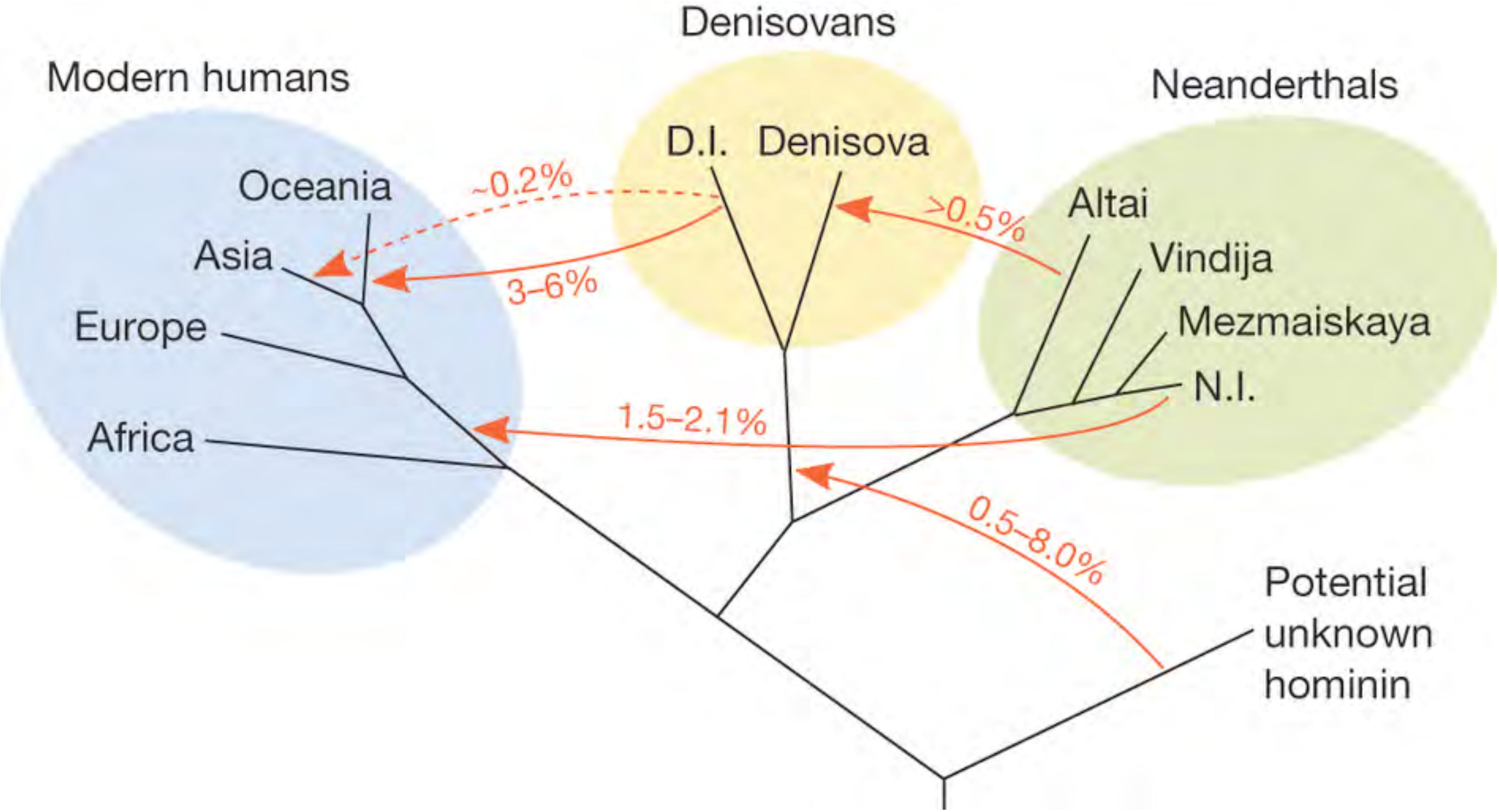
EXPECT:	OBSERVE:
50% ABBA	103612 ABBA
50% BABA	94029 BABA

Green *et al.* 2010 *Science* 328:710-722

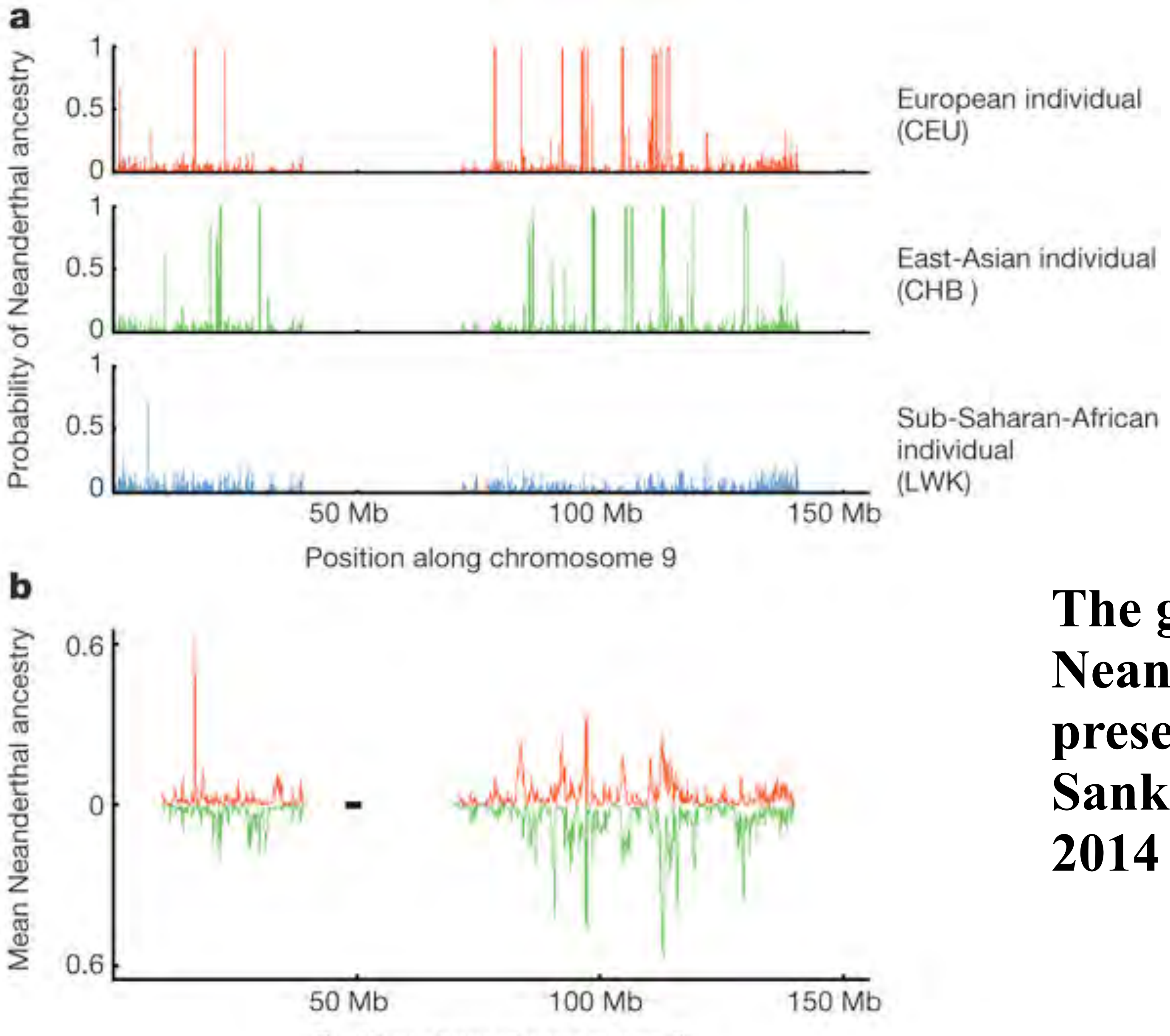
Explicit tests for gene flow: ABBA-BABA test



Explicit tests for gene flow: ABBA-BABA test

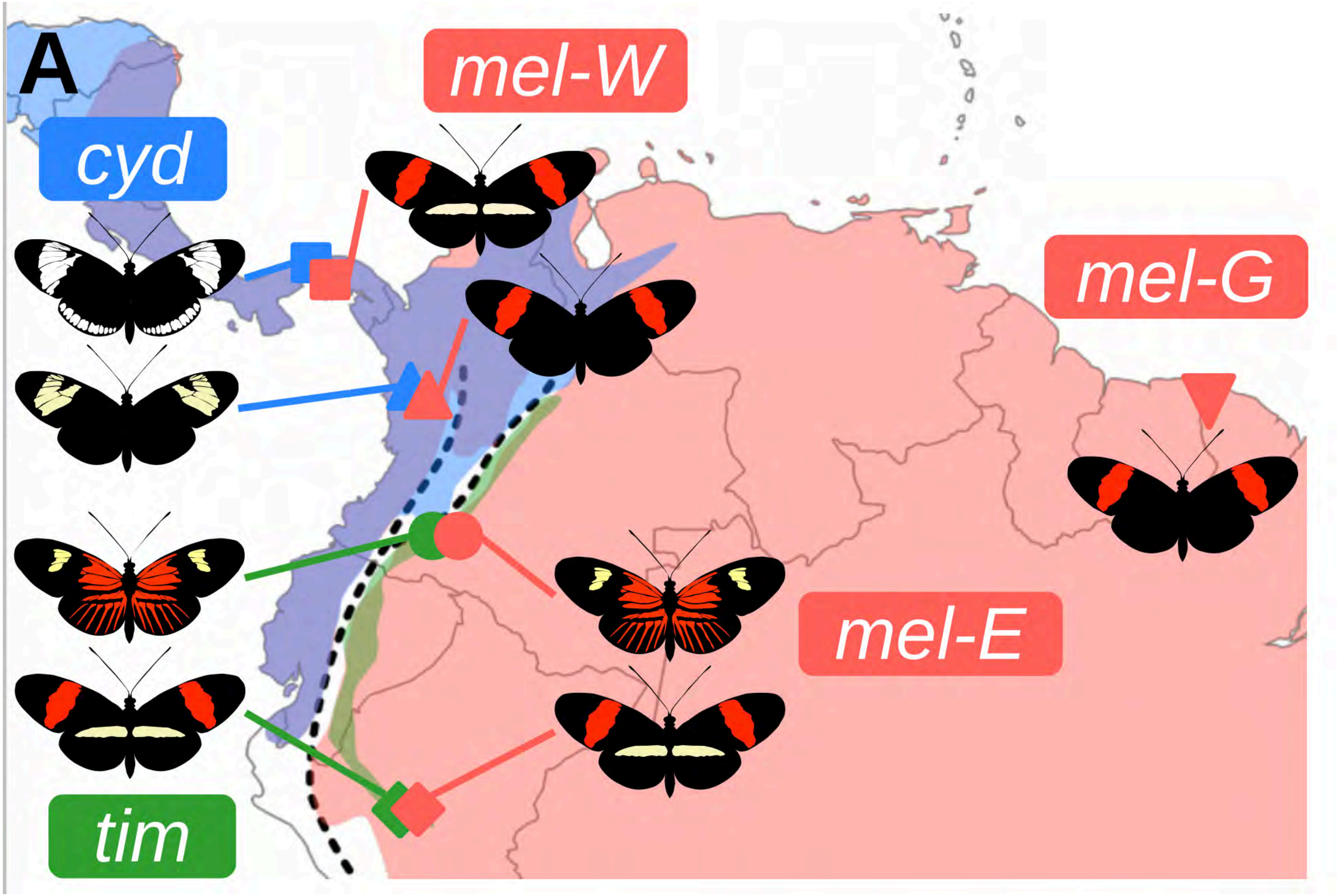


Explicit tests for gene flow: Combining multiple signals



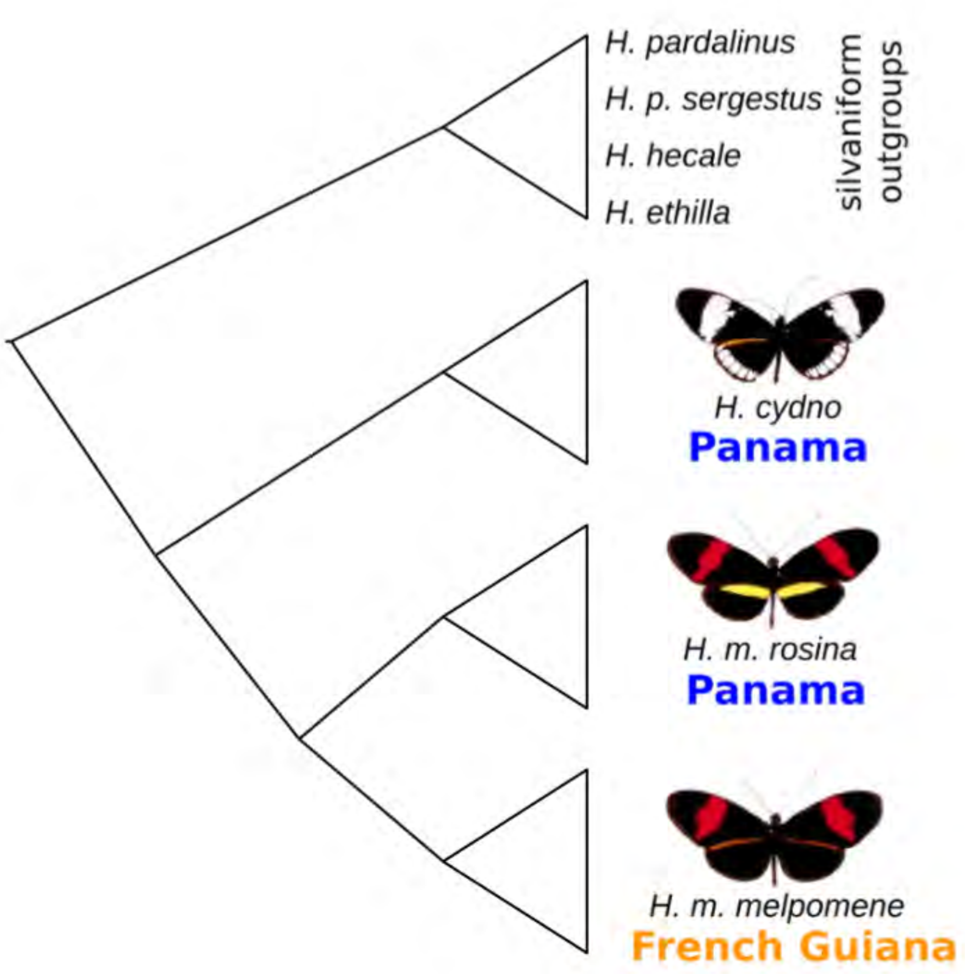
1) Derived alleles at high frequency shared with Neanderthal
2) High divergence to Africa but low to Neanderthal
3) Long haplotype blocks

The genomic landscape of Neanderthal ancestry in present-day humans - Sankararaman et al. Nature 2014



Sampled 10 complete high coverage genomes per population

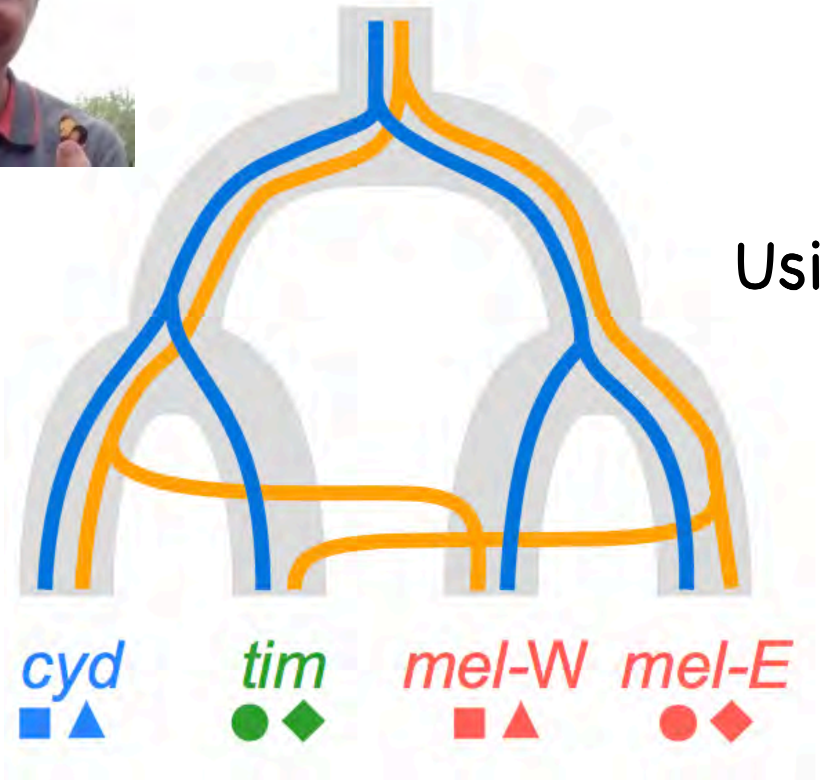
Explicit tests for gene flow: *Heliconius* butterflies



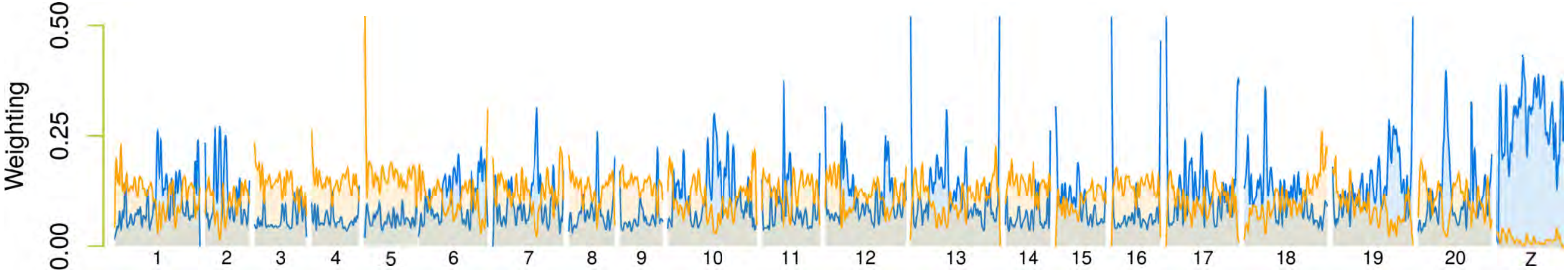
Whole-genome phylogeny supports grouping by species

Many sources of reproductive isolation:

Female hybrids are sterile
Different host plant use
Different habitat preference
Strong assortative mating

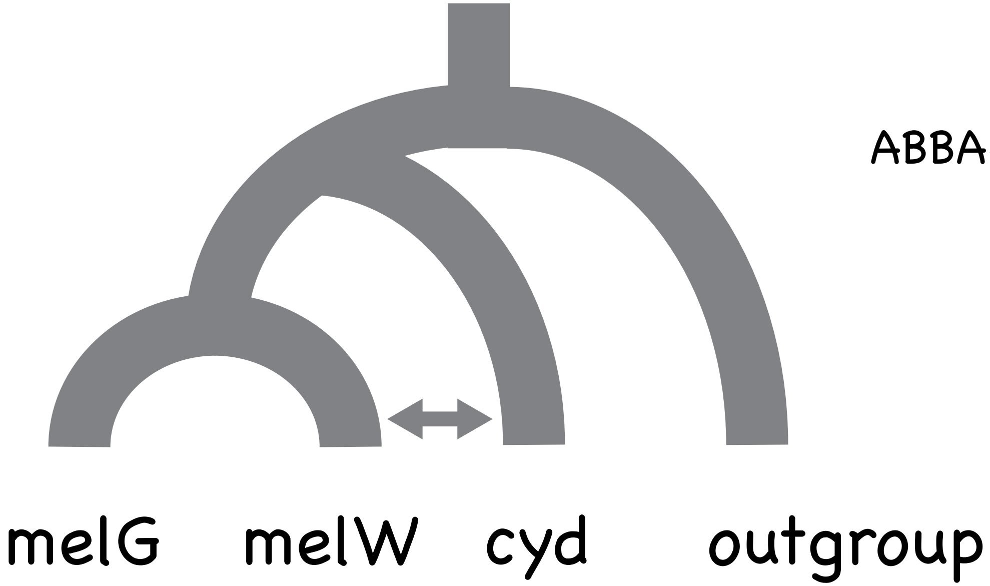


Using Simon Martin's Twisst method to characterise relationships

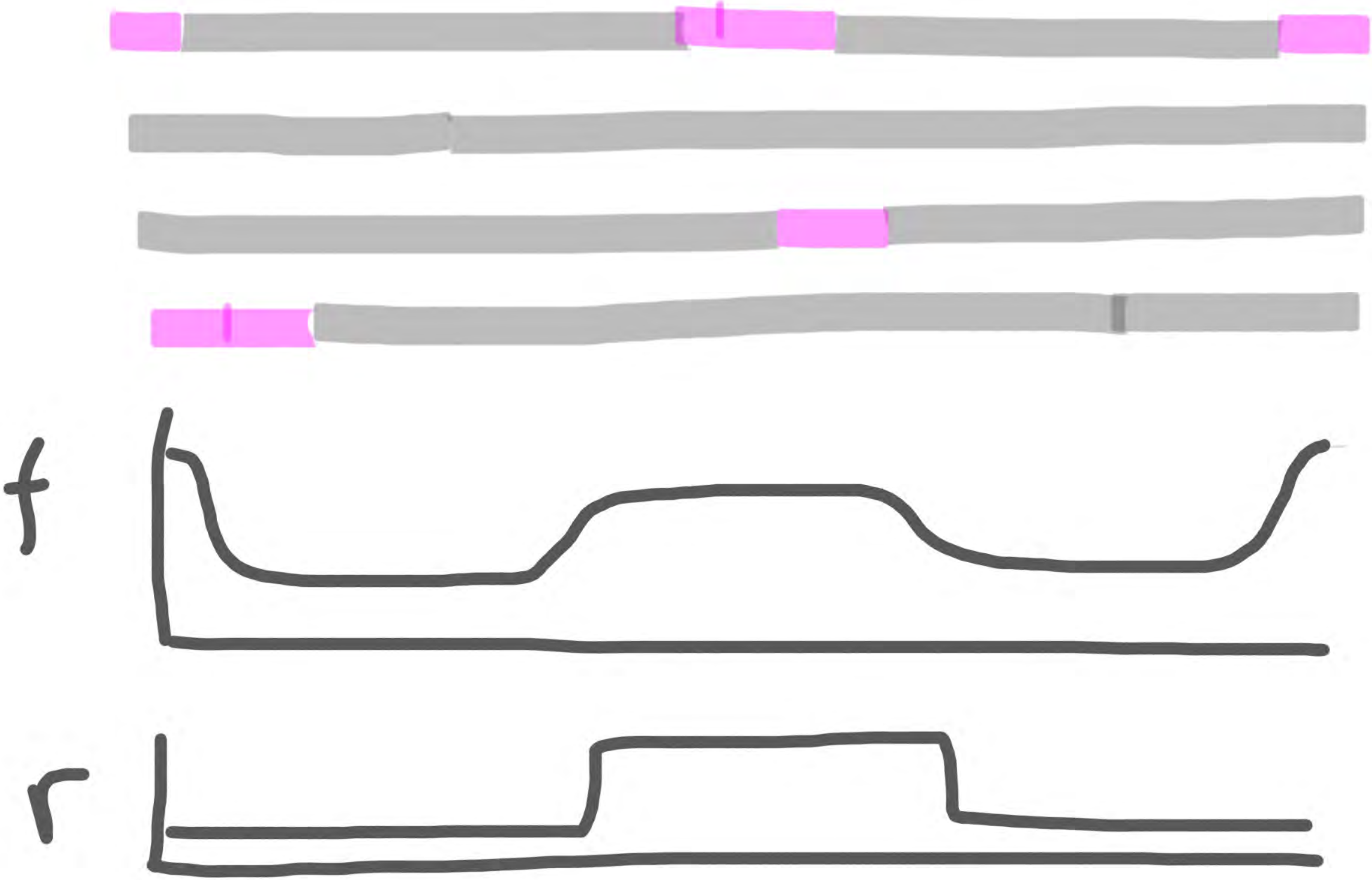


Simon Martin

ABBA-BABA statistics



Martin et al., MBE 2015


f
r

John Davey

300+ offspring for each cross type

H. cydno **Panama** x *H. cydno* **Panama**

H. m. rosina **Panama** x *H. cydno* **Panama**

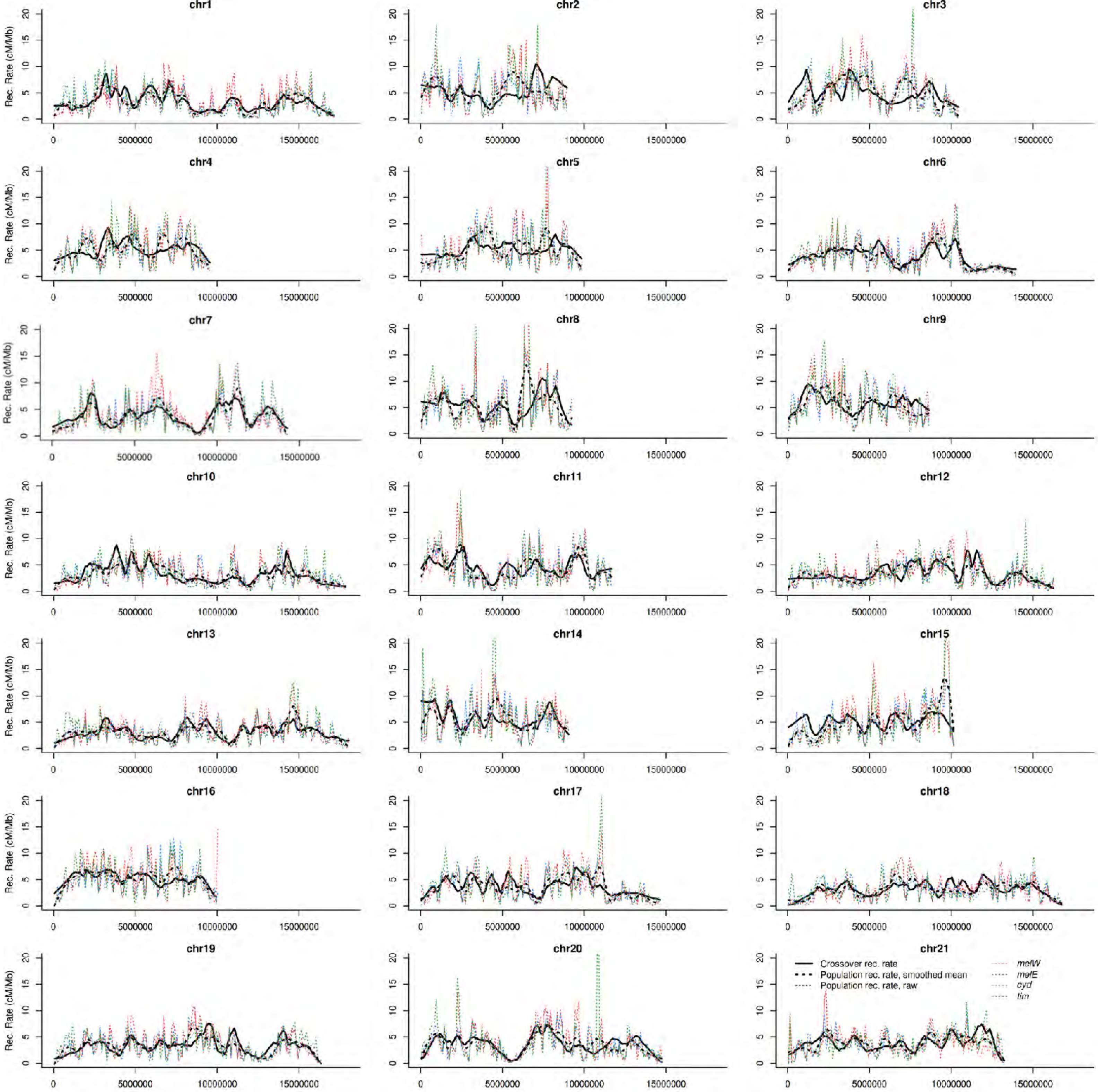
H. m. rosina **Panama** x *H. m. rosina* **Panama**

PstI RAD sequencing
(site every ~10kb)

Linkage maps built
with Lep-MAP

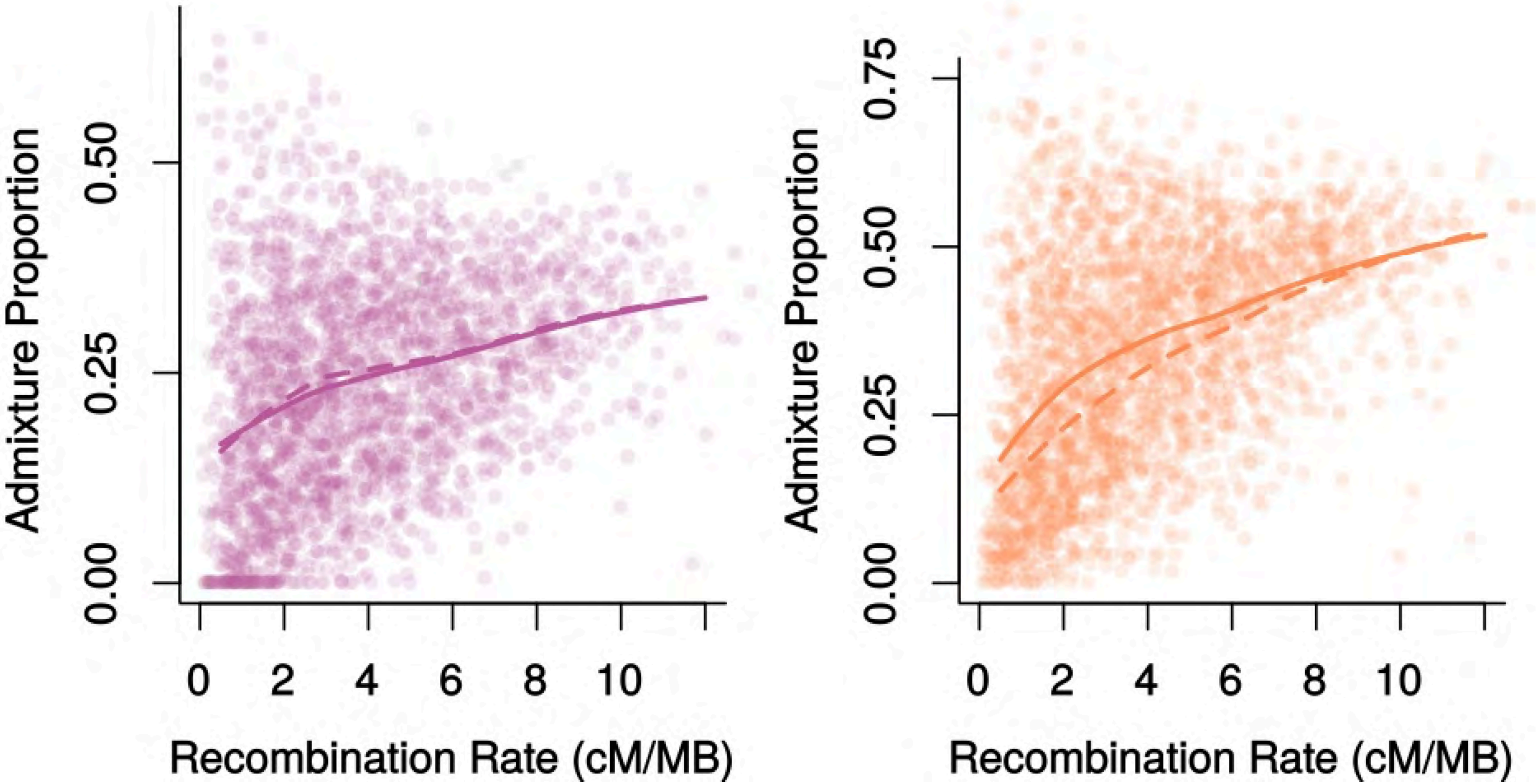
Davey et al., Evol Letters 2017

Pop gen vs actual estimates of recombination rate

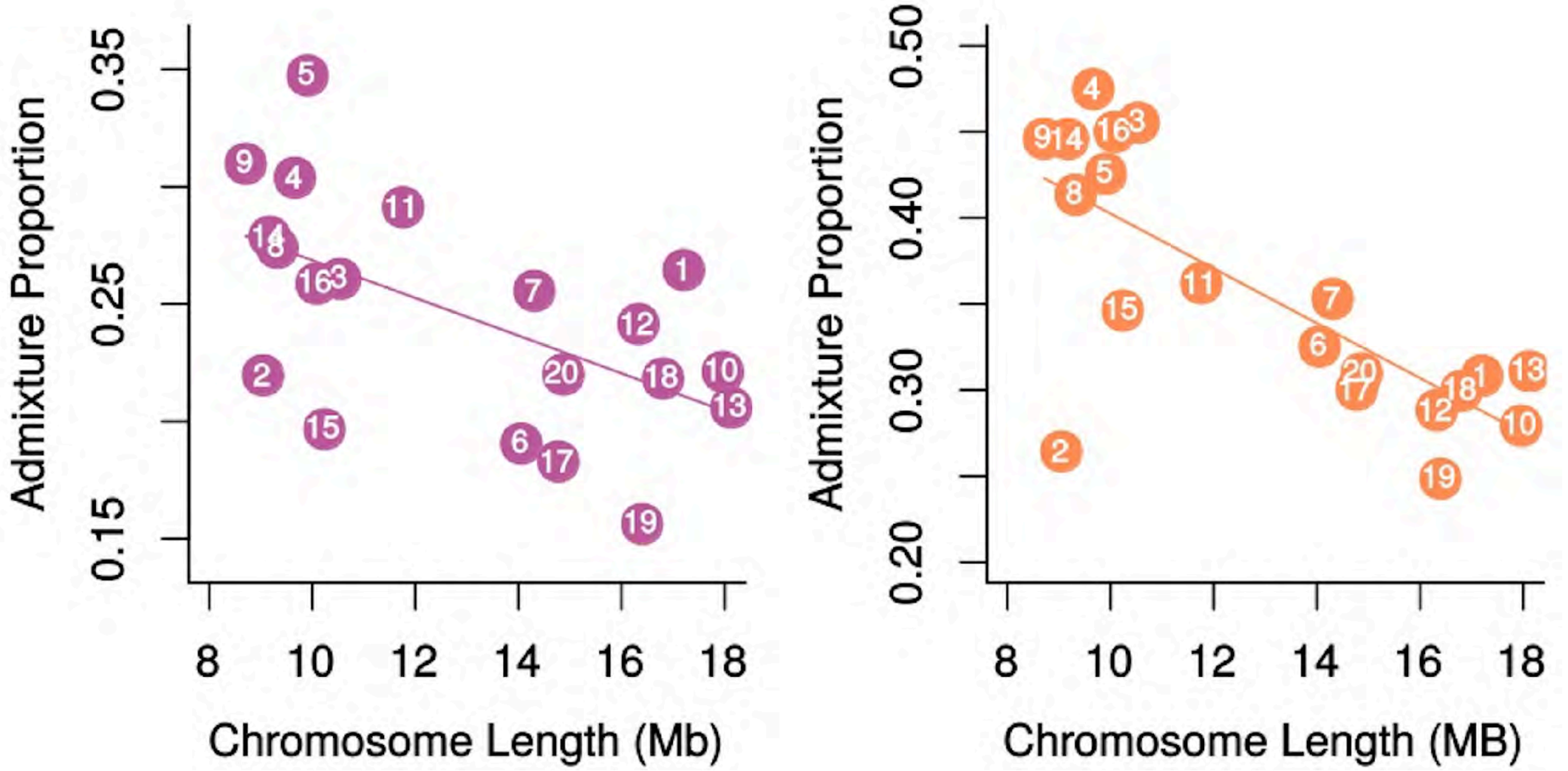


Recombination rate strongly correlated with admixture proportion

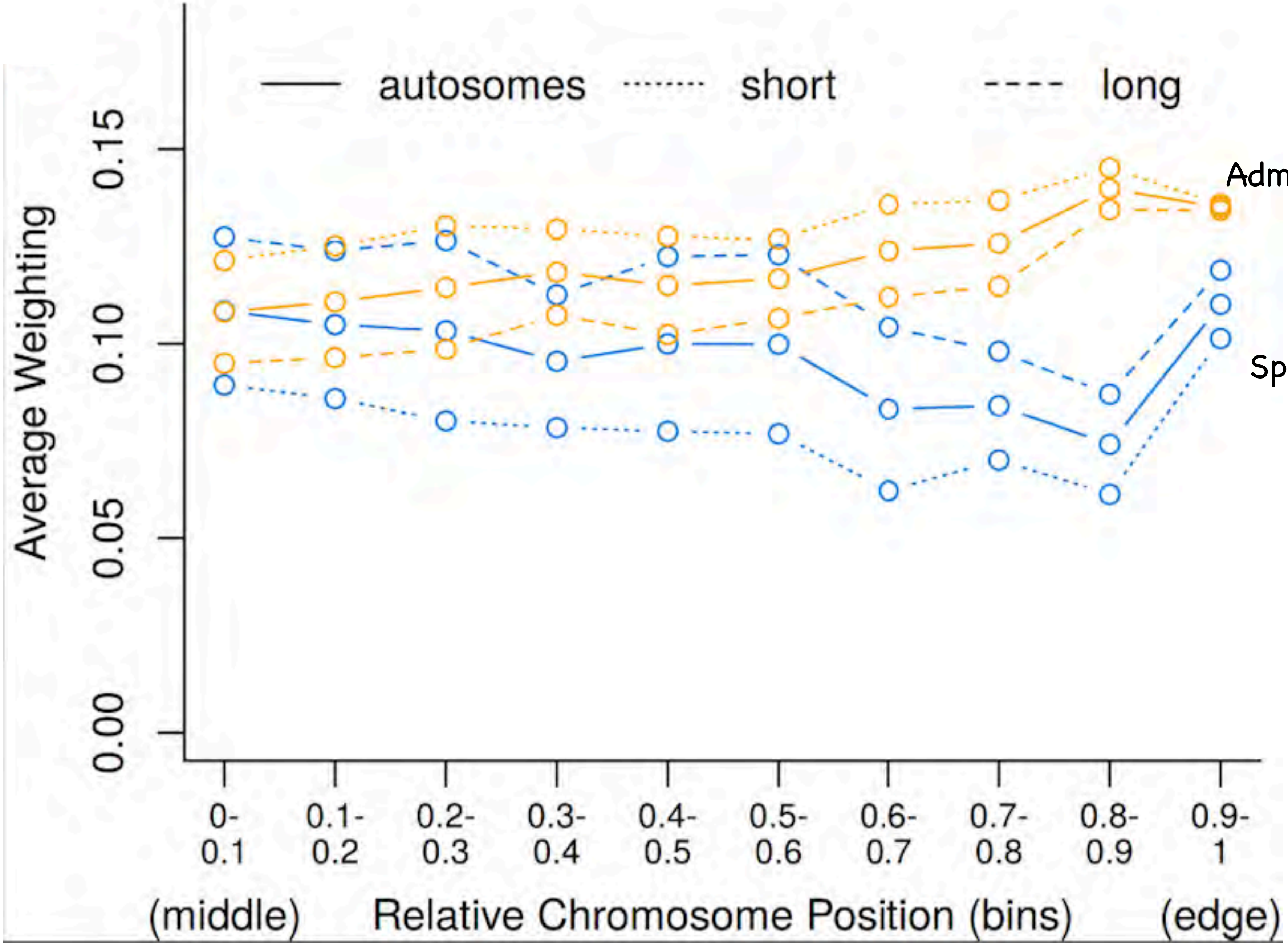
Simon Martin



Short chromosomes have more admixture:

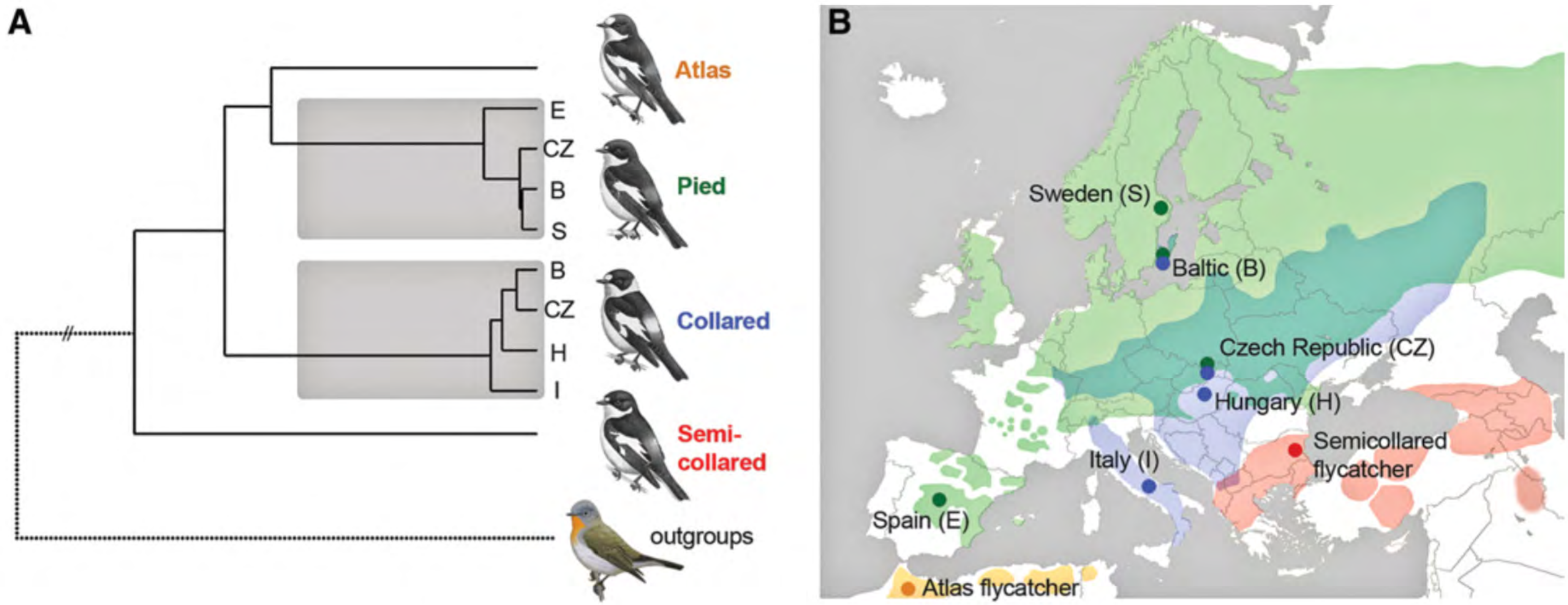


And chromosome ends have more admixture:



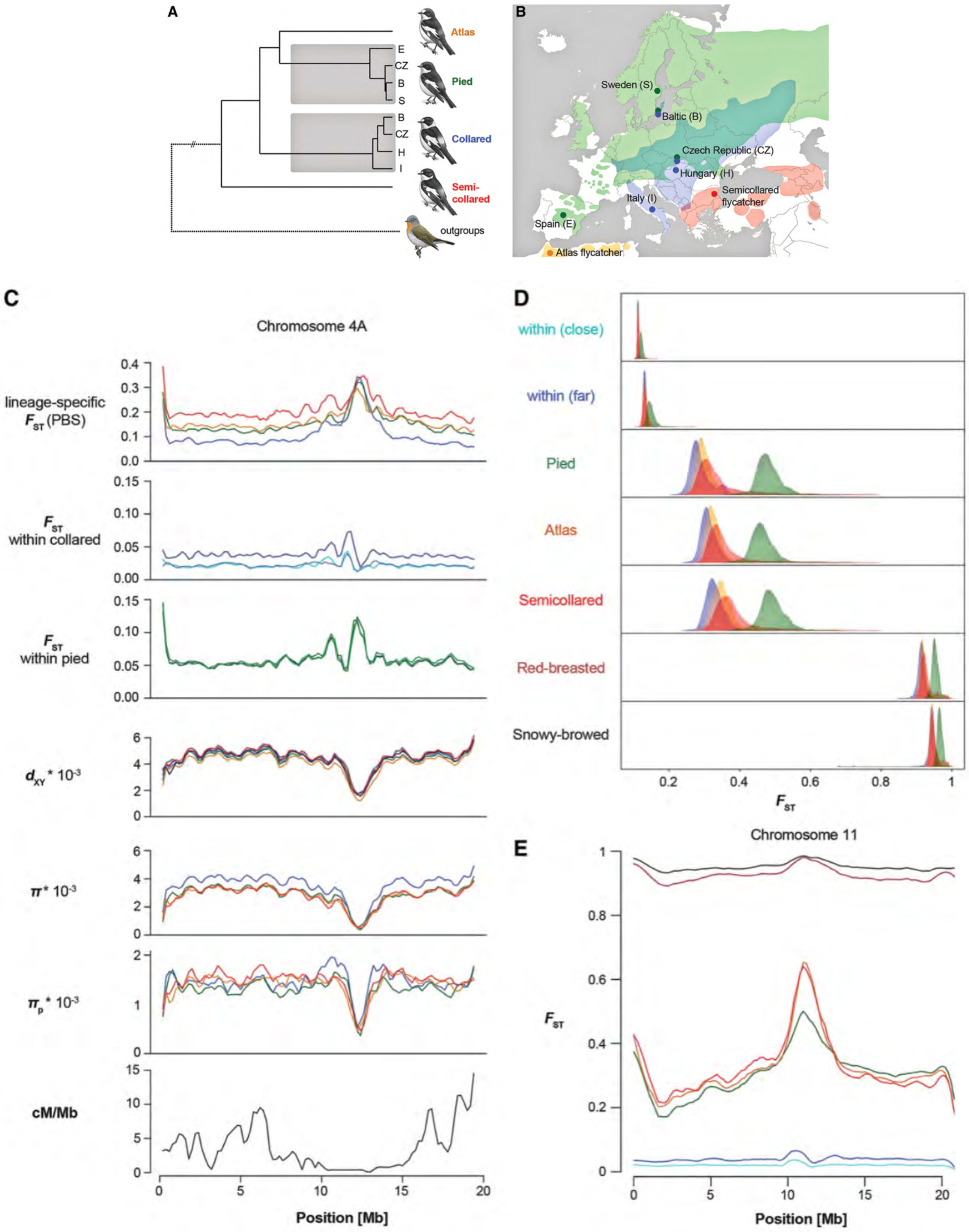
Admixture tree

Species tree



Sequenced 20 individuals per population at 20x coverage

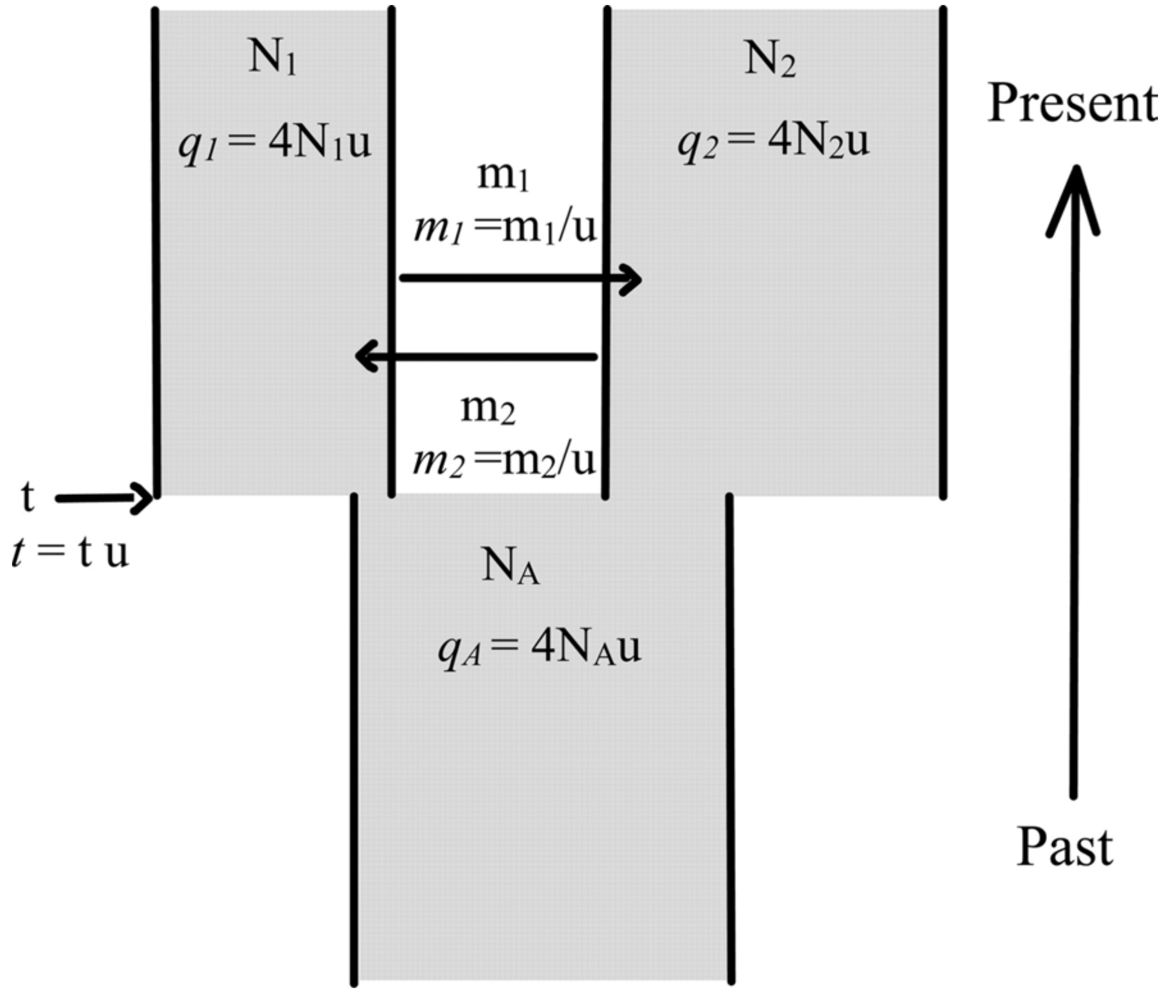
Burri et al., Genome Research 2015


A
E
CZ
B
S
B
CZ
H
I
Atlas
Pied
Collared
Semi-collared
outgroups
B
Sweden (S)
Baltic (B)
Czech Republic (CZ)
Hungary (H)
Italy (I)
Semicollared flycatcher
Spain (E)
Atlas flycatcher
C
Chromosome 4A
lineage-specific F_{ST} (PBS)
0.4
0.3
0.2
0.1
0.0
F_{ST} within collared
0.15
0.10
0.05
0.00
F_{ST} within pied
0.15
0.10
0.05
0.00
d_{XY} * 10^{-3}
6
4
2
0
π * 10^{-3}
6
4
2
0
π_p * 10^{-3}
2
1
0
cM/Mb
15
10
5
0
0
5
10
15
20
Position [Mb]
D
within (close)
within (far)
Pied
Atlas
Semicollared
Red-breasted
Snowy-browed
0.2
0.4
0.6
0.8
1
F_{ST}
E
Chromosome 11
1
0.8
0.6
0.4
0.2
0
F_{ST}
0
5
10
15
20
Position [Mb]

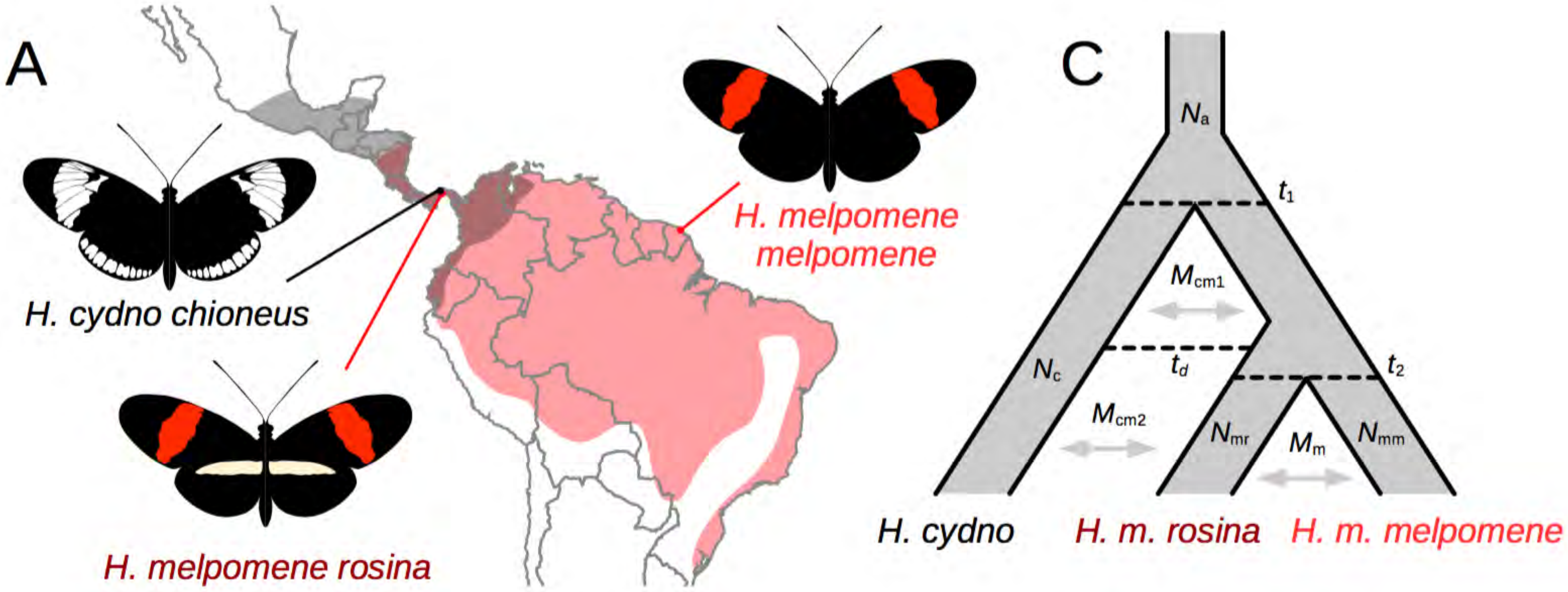
Supplemental Table S4. ABBA-BABA tests for gene flow. Populations/species among which the test indicates gene flow are highlighted in bold.

1. Inner	2. Inner	1. Outgroup	Mean(D)	SE(D)	p-value
collared Italy	collared CZ	pied CZ	0.0010	0.0010	0.3344
pied Spain	pied CZ	collared CZ	0.0004	0.0005	0.4186
pied Spain	Atlas	**collared Italy**	-0.1648	0.0027	$<10^{-4}$
pied Spain	Atlas	**semicollared**	-0.0108	0.0016	$<10^{-4}$
pied Spain	**collared Italy**	**semicollared**	0.1162	0.0018	$<10^{-4}$
Atlas	**collared Italy**	**semicollared**	0.1242	0.0016	$<10^{-4}$

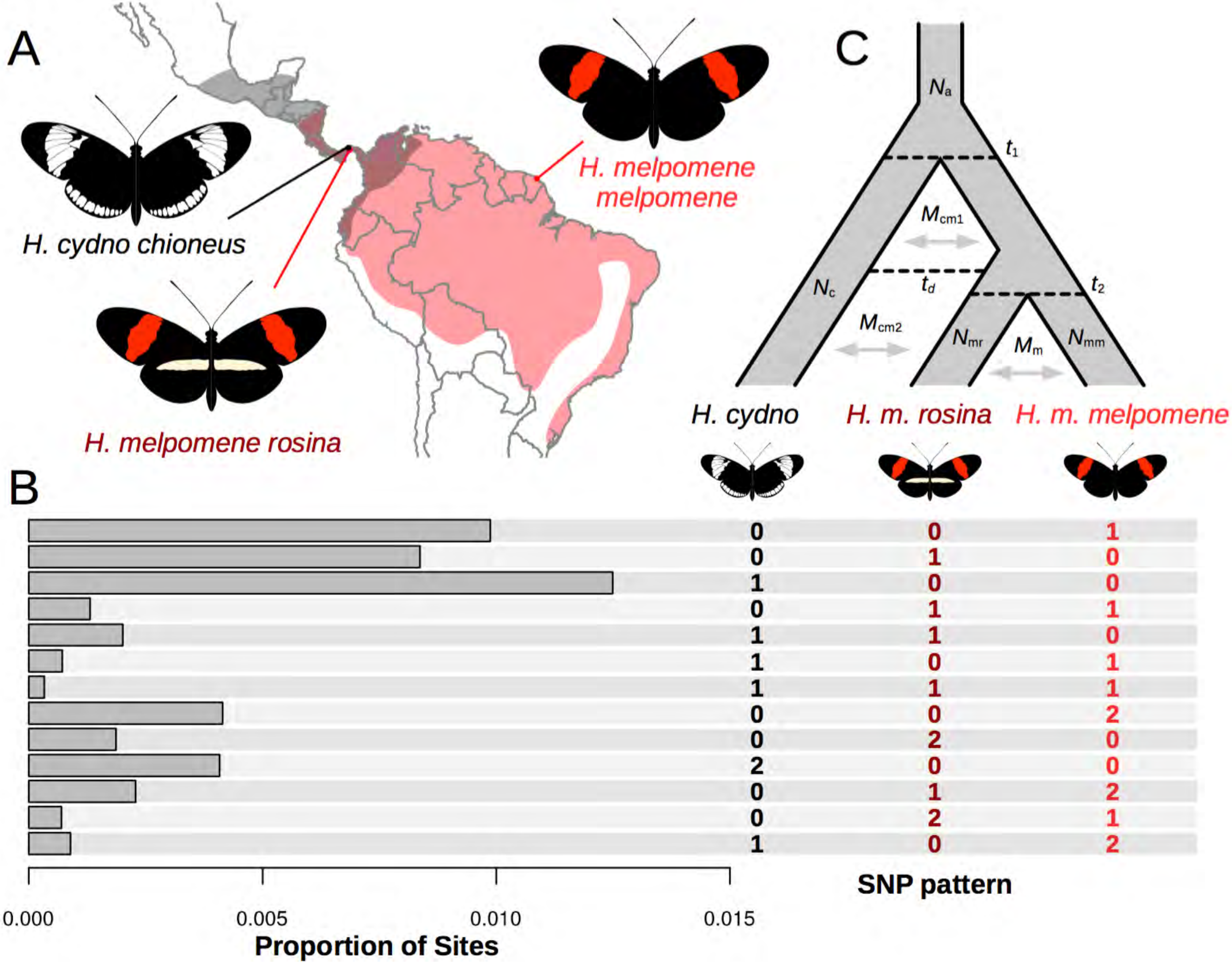
An alternative is to take an explicit modelling approach

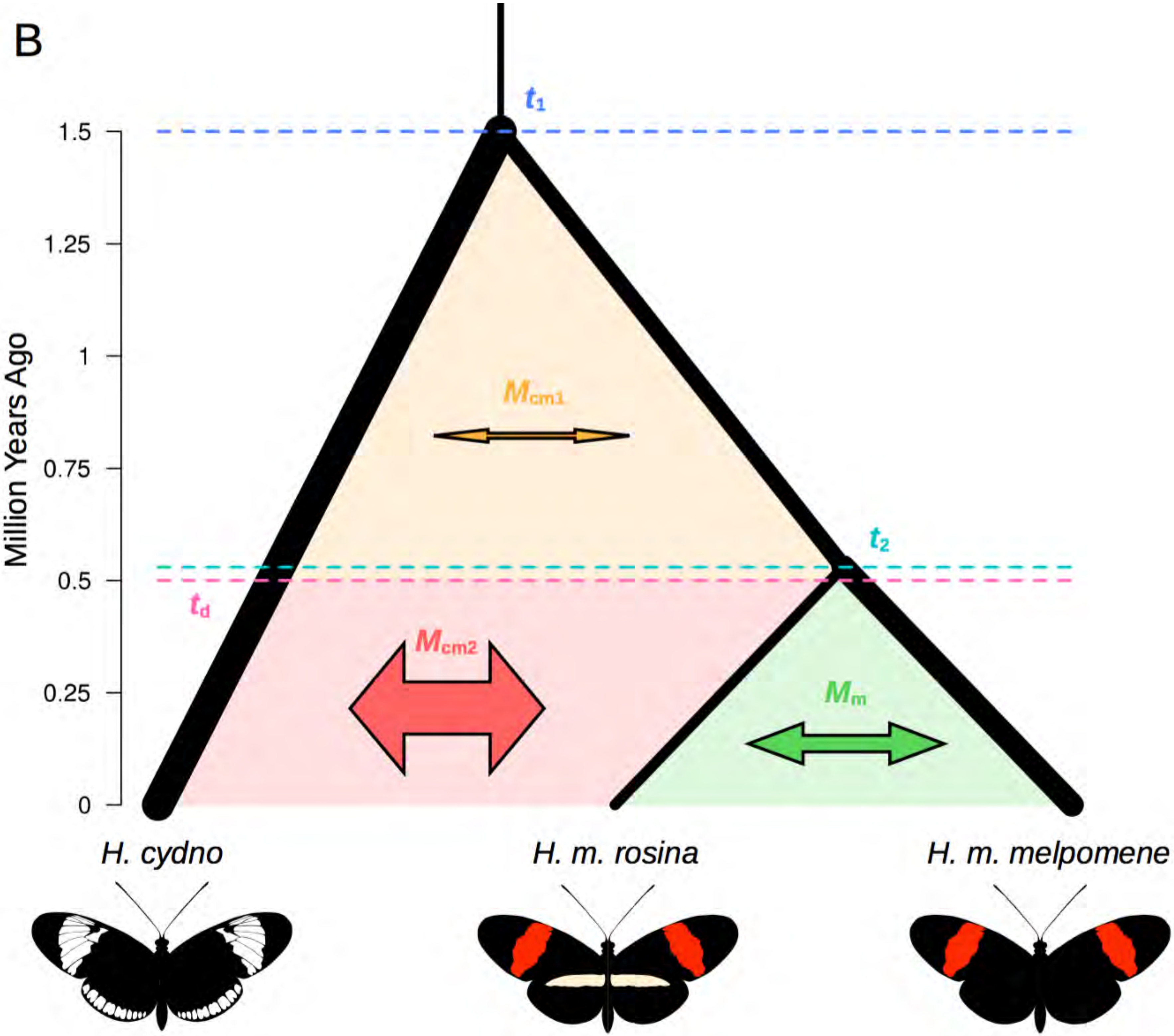


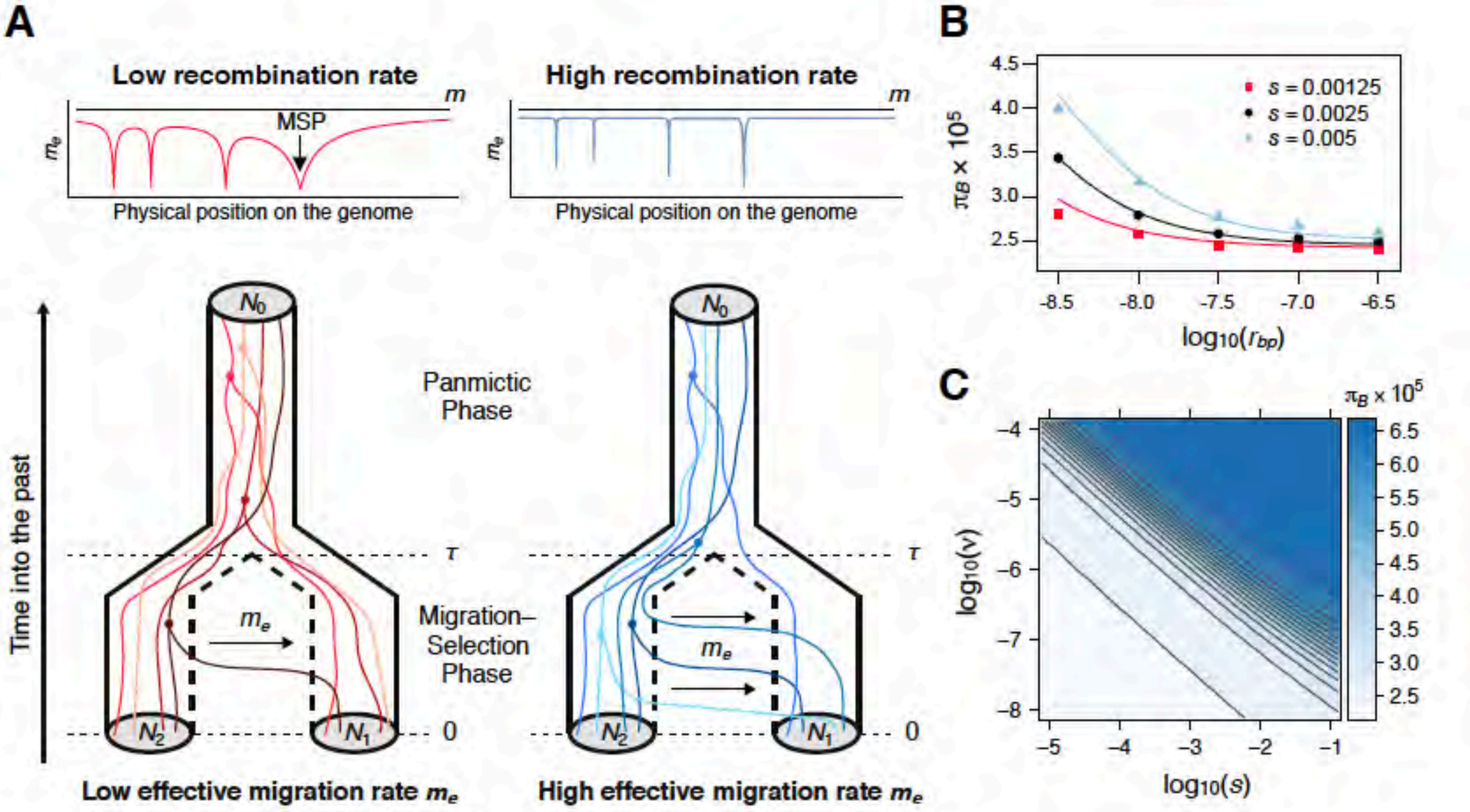
IM and IMa
Jody Hey



Martin et al., Biorxiv 2015


A
H. melpomene
melpomene
H. cydno chioneus
H. melpomene rosina
C
N_a
t_1
M_{cm1}
N_c
t_d
t_2
M_{cm2}
N_{mr}
M_m
N_{mm}
H. cydno
H. m. rosina
H. m. melpomene
B
0 0 1
0 1 0
1 0 0
0 1 1
1 1 0
1 0 1
1 1 1
0 0 2
0 2 0
2 0 0
0 1 2
0 2 1
1 0 2
SNP pattern
0.000
0.005
0.010
0.015
Proportion of Sites

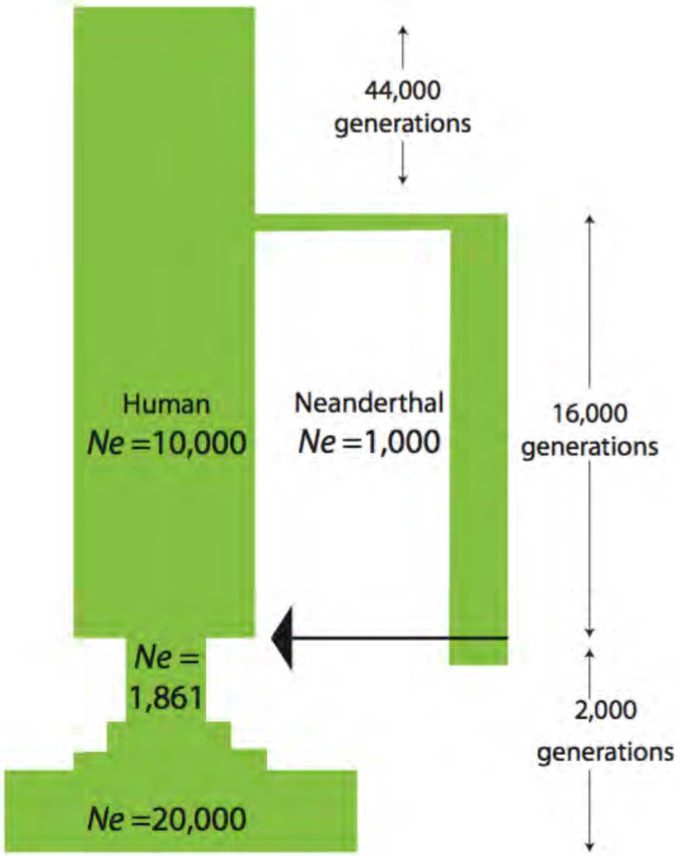

B
Million Years Ago
1.5
1.25
1
0.75
0.5
0.25
0
t_1
t_2
t_d
M_{cm1}
M_{cm2}
M_m
H. cydno
H. m. rosina
H. m. melpomene

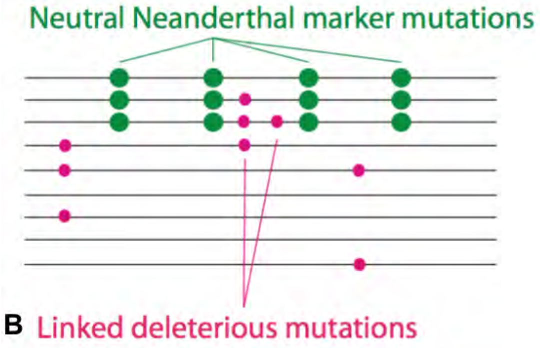


Population-genomic inference of the strength and timing of selection against gene flow

Simon Aeschbacher[1,a], Jessica P. Selby[2], John H. Willis[2], and Graham Coop[1]

The effect of background selection on introgression in humans





Admixture is less in gene rich regions supporting this model…..

Harris and Nielson Genetics 2016,
Juric, Aeschbacher and Coop 2016

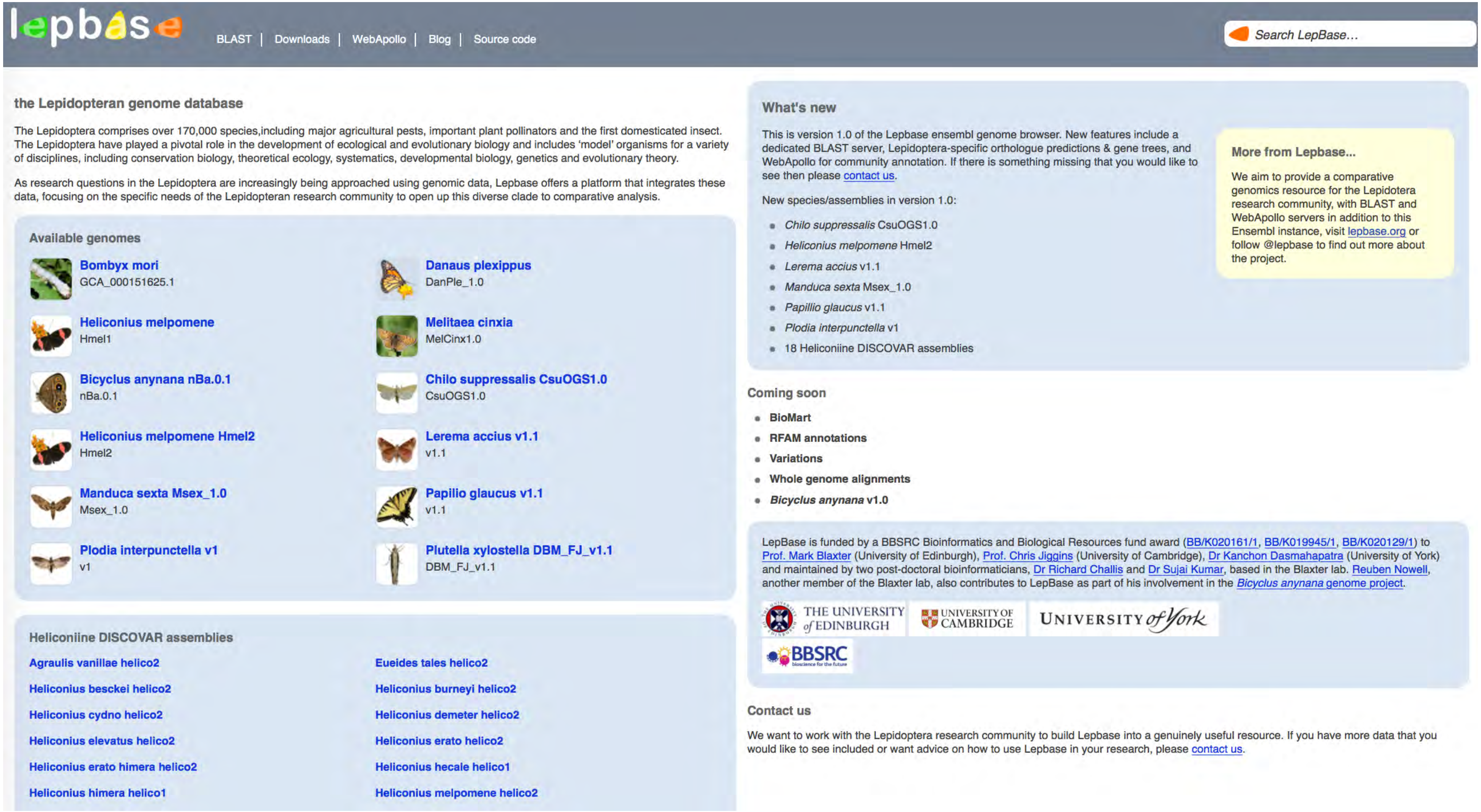
Population and speciation genomics: Conclusions

- Great power to detect subtle signals of selection and gene flow
- Can make more general observations about genes and regions involved in adaptation
- BUT genomic processes complicate the picture

- Best approaches combine multiple signals to infer process

- Eventually we need to combine background selection, recombination, positive selection

And finally a plug….



the Lepidopteran genome database

The Lepidoptera comprises over 170,000 species,including major agricultural pests, important plant pollinators and the first domesticated insect. The Lepidoptera have played a pivotal role in the development of ecological and evolutionary biology and includes 'model' organisms for a variety of disciplines, including conservation biology, theoretical ecology, systematics, developmental biology, genetics and evolutionary theory.

As research questions in the Lepidoptera are increasingly being approached using genomic data, Lepbase offers a platform that integrates these data, focusing on the specific needs of the Lepidopteran research community to open up this diverse clade to comparative analysis.

Available genomes

Bombyx mori
GCA_000151625.1

Danaus plexippus
DanPle_1.0

Heliconius melpomene
Hmel1

Melitaea cinxia
MelCinx1.0

Bicyclus anynana nBa.0.1
nBa.0.1

Chilo suppressalis CsuOGS1.0
CsuOGS1.0

Heliconius melpomene Hmel2
Hmel2

Lerema accius v1.1
v1.1

Manduca sexta Msex_1.0
Msex_1.0

Papilio glaucus v1.1
v1.1

Plodia interpunctella v1
v1

Plutella xylostella DBM_FJ_v1.1
DBM_FJ_v1.1

Heliconiine DISCOVAR assemblies

Agraulis vanillae helico2
Heliconius besckei helico2
Heliconius cydno helico2
Heliconius elevatus helico2
Heliconius erato himera helico2
Heliconius himera helico1
Eueides tales helico2
Heliconius burneyi helico2
Heliconius demeter helico2
Heliconius erato helico2
Heliconius hecale helico1
Heliconius melpomene helico2

What's new

This is version 1.0 of the Lepbase ensembl genome browser. New features include a dedicated BLAST server, Lepidoptera-specific orthologue predictions & gene trees, and WebApollo for community annotation. If there is something missing that you would like to see then please contact us.

New species/assemblies in version 1.0:

- *Chilo suppressalis* CsuOGS1.0
- *Heliconius melpomene* Hmel2
- *Lerema accius* v1.1
- *Manduca sexta* Msex_1.0
- *Papillio glaucus* v1.1
- *Plodia interpunctella* v1
- 18 Heliconiine DISCOVAR assemblies

More from Lepbase...

We aim to provide a comparative genomics resource for the Lepidotera research community, with BLAST and WebApollo servers in addition to this Ensembl instance, visit lepbase.org or follow @lepbase to find out more about the project.

Coming soon

- **BioMart**
- **RFAM annotations**
- **Variations**
- **Whole genome alignments**
- ***Bicyclus anynana* v1.0**

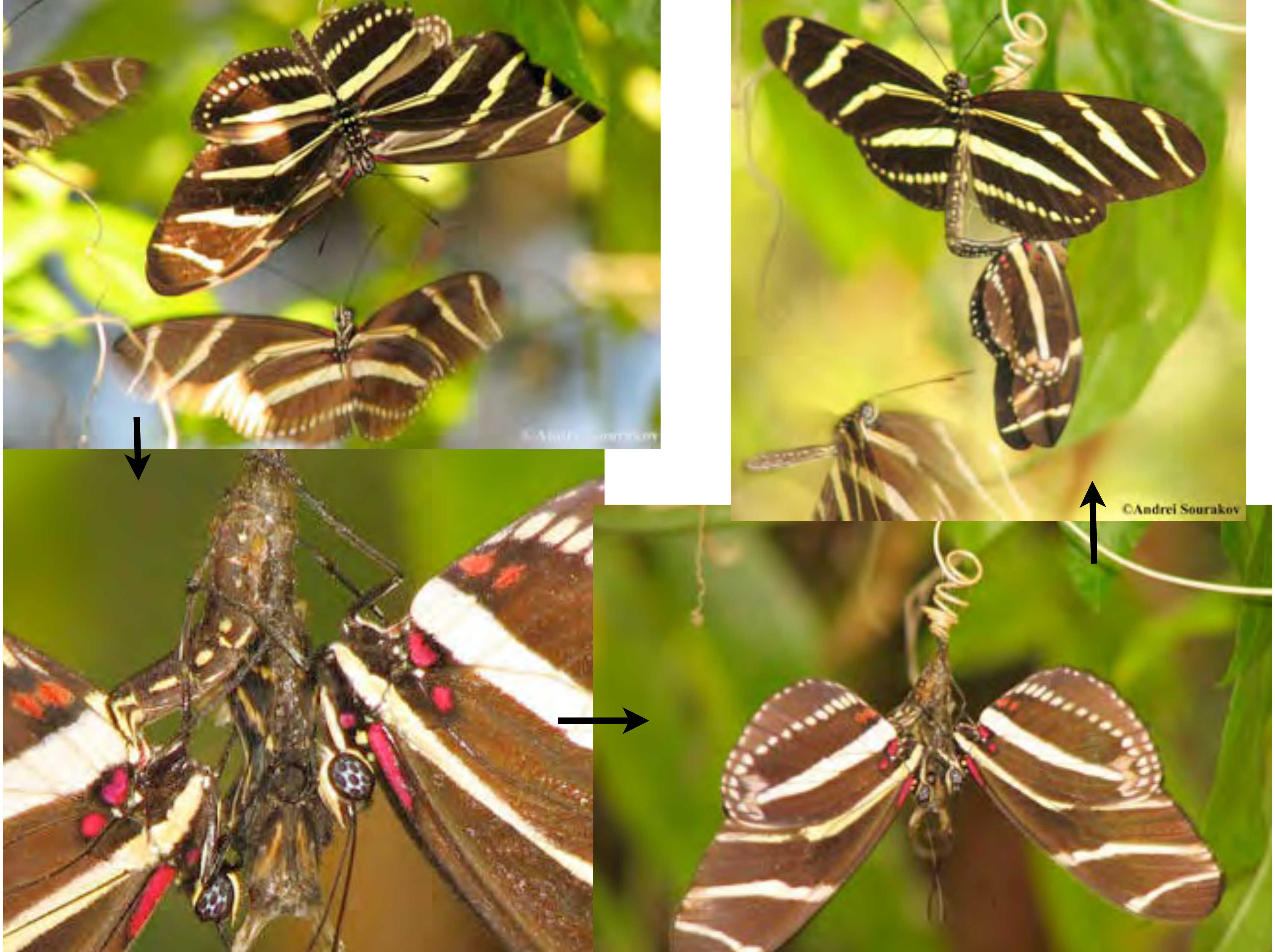
LepBase is funded by a BBSRC Bioinformatics and Biological Resources fund award (BB/K020161/1, BB/K019945/1, BB/K020129/1) to Prof. Mark Blaxter (University of Edinburgh), Prof. Chris Jiggins (University of Cambridge), Dr Kanchon Dasmahapatra (University of York) and maintained by two post-doctoral bioinformaticians, Dr Richard Challis and Dr Sujai Kumar, based in the Blaxter lab. Reuben Nowell, another member of the Blaxter lab, also contributes to LepBase as part of his involvement in the *Bicyclus anynana* genome project.

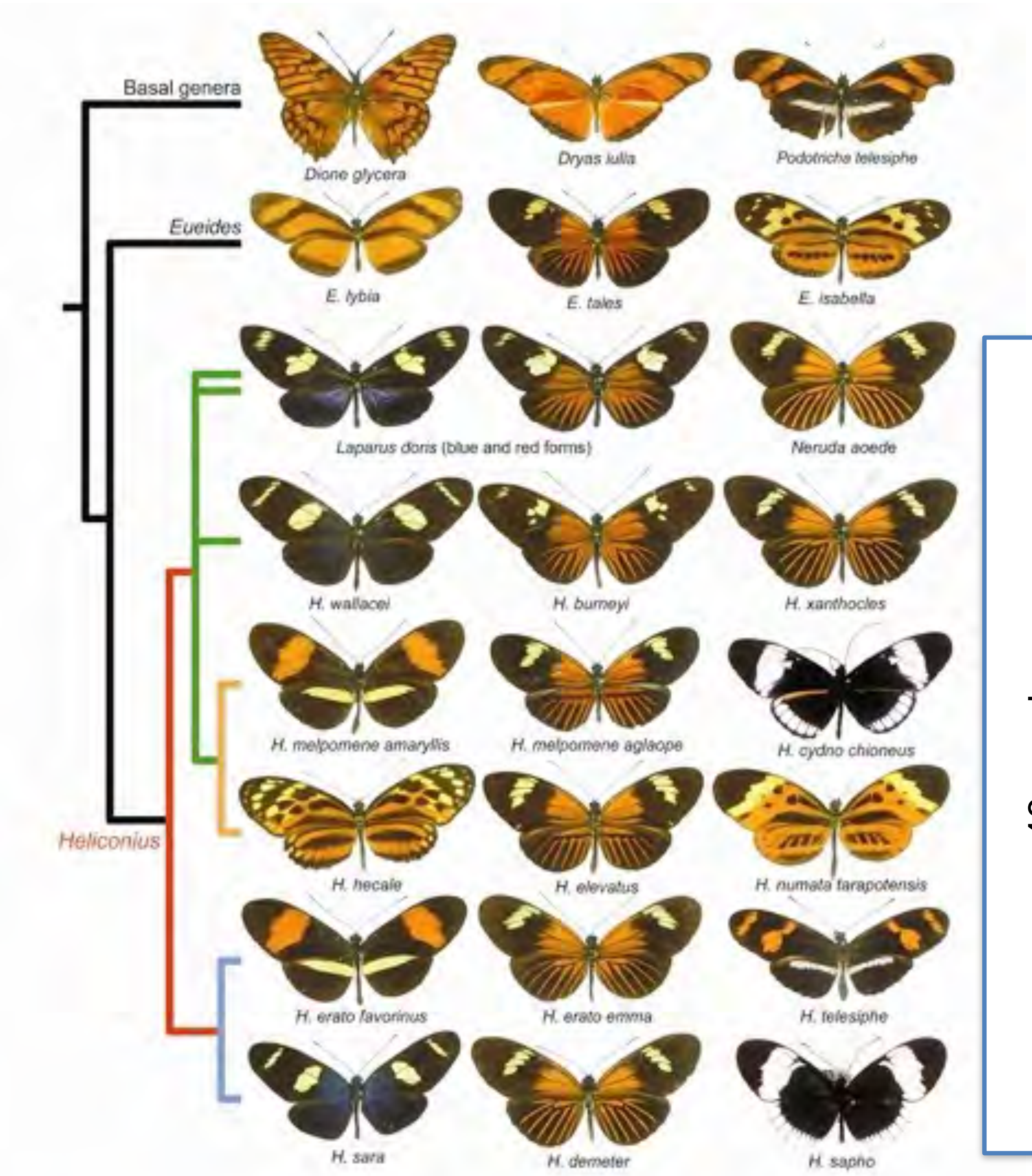
THE UNIVERSITY of EDINBURGH
UNIVERSITY OF CAMBRIDGE
UNIVERSITY of York
BBSRC bioscience for the future

Contact us

We want to work with the Lepidoptera research community to build Lepbase into a genuinely useful resource. If you have more data that you would like to see included or want advice on how to use Lepbase in your research, please contact us.

Adaptive introgression


© Andrei Sourakov
©Andrei Sourakov



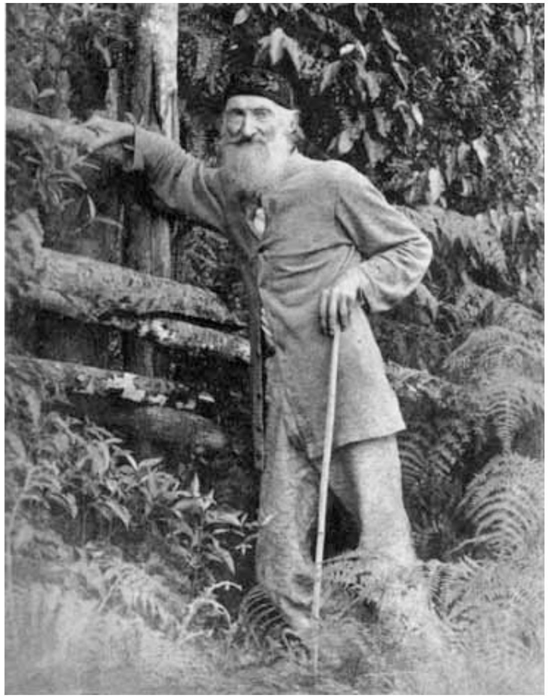
43 species

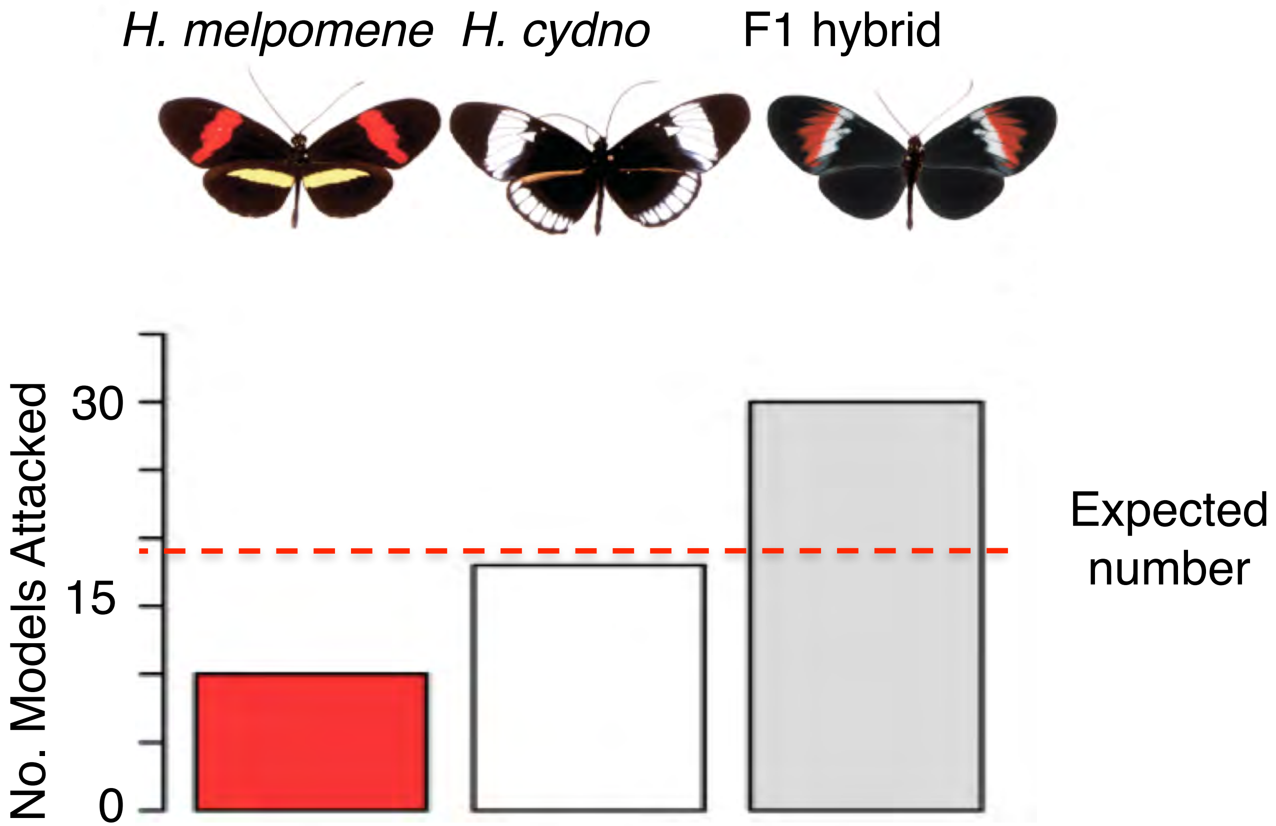
77 species


© Richard G

Rufous-tailed Jacamar

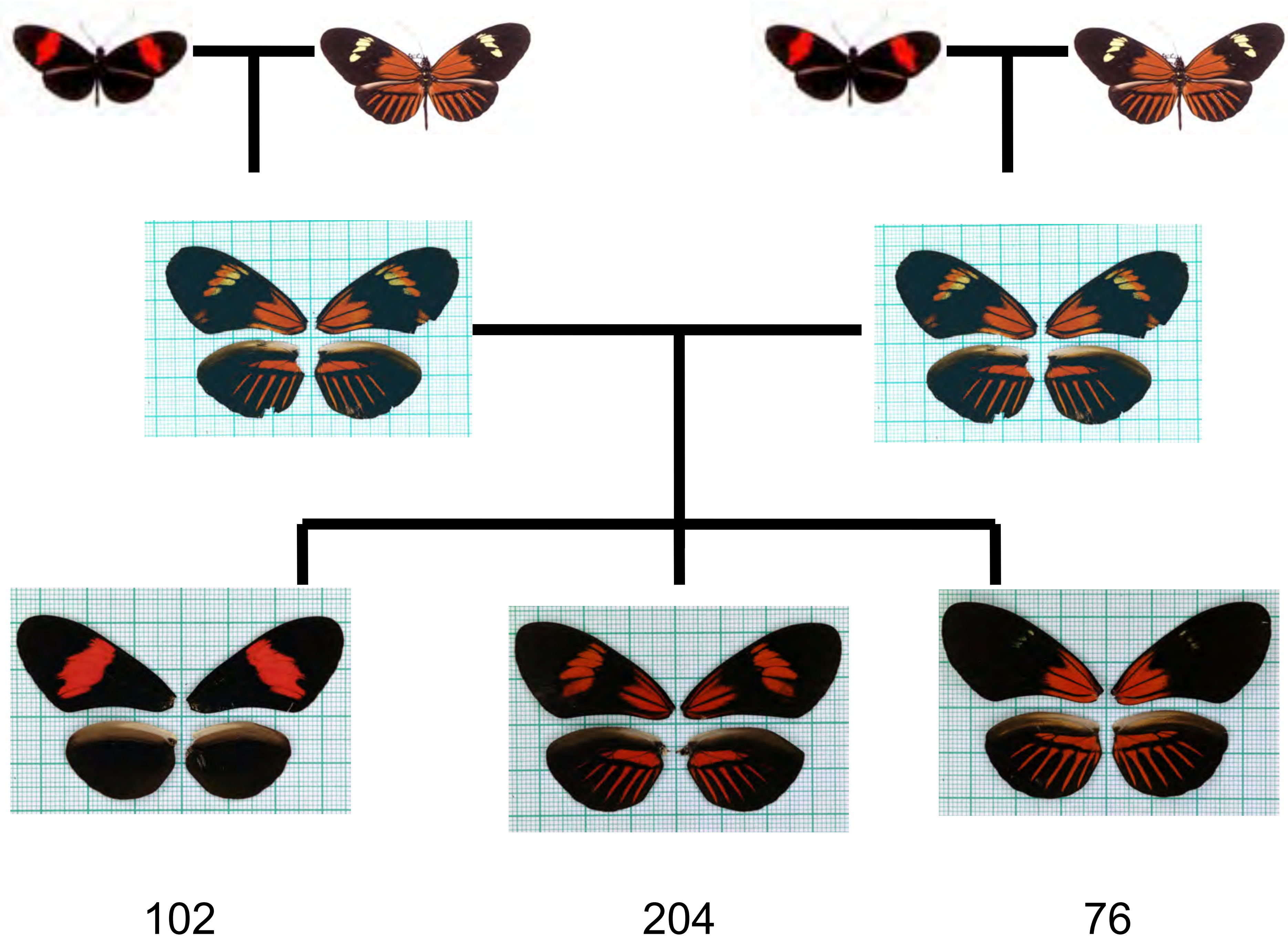
Fritz Müller

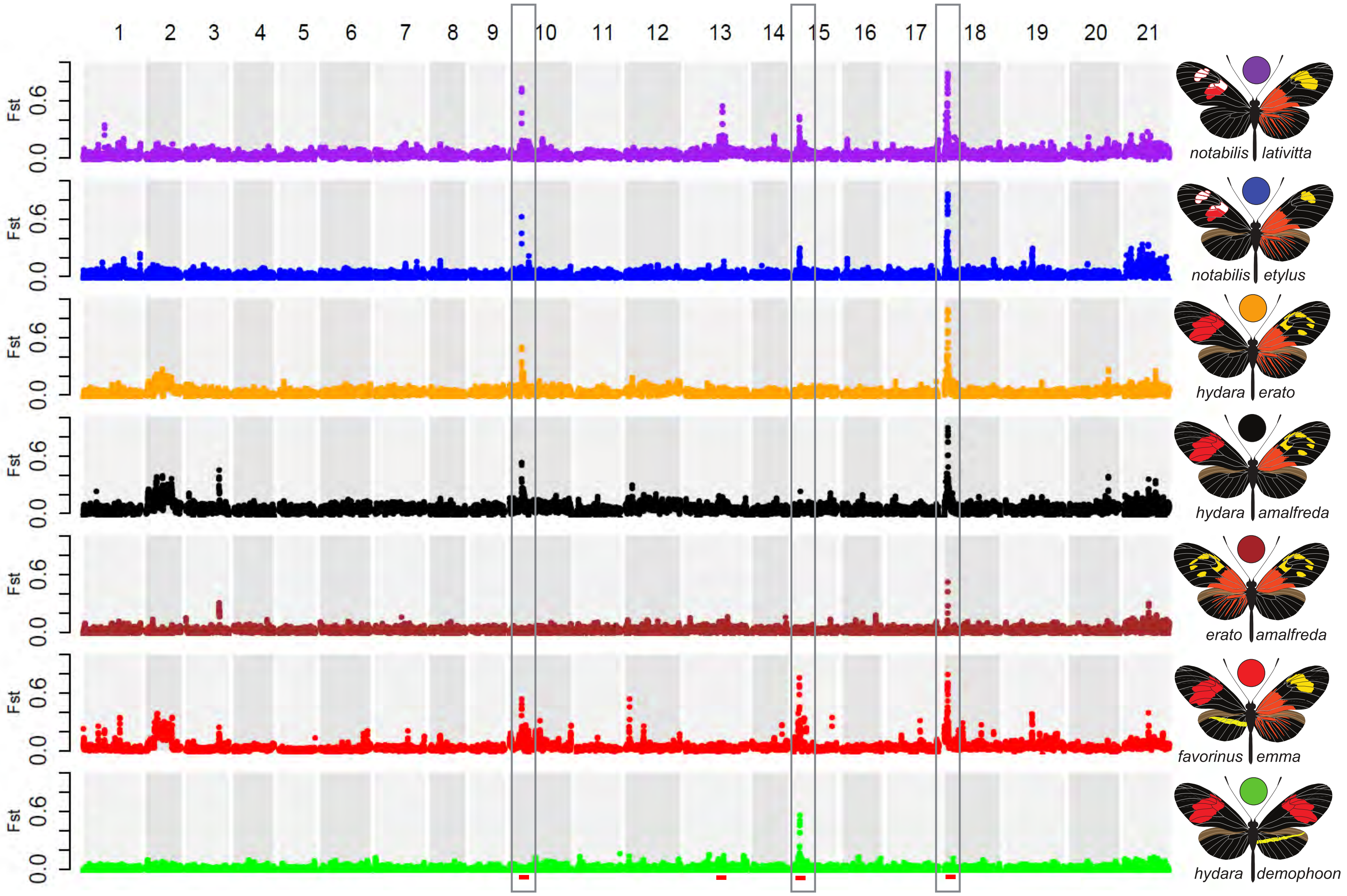




– *G*-test: $G = 7.25$, d.f. = 1, $p = 0.007$

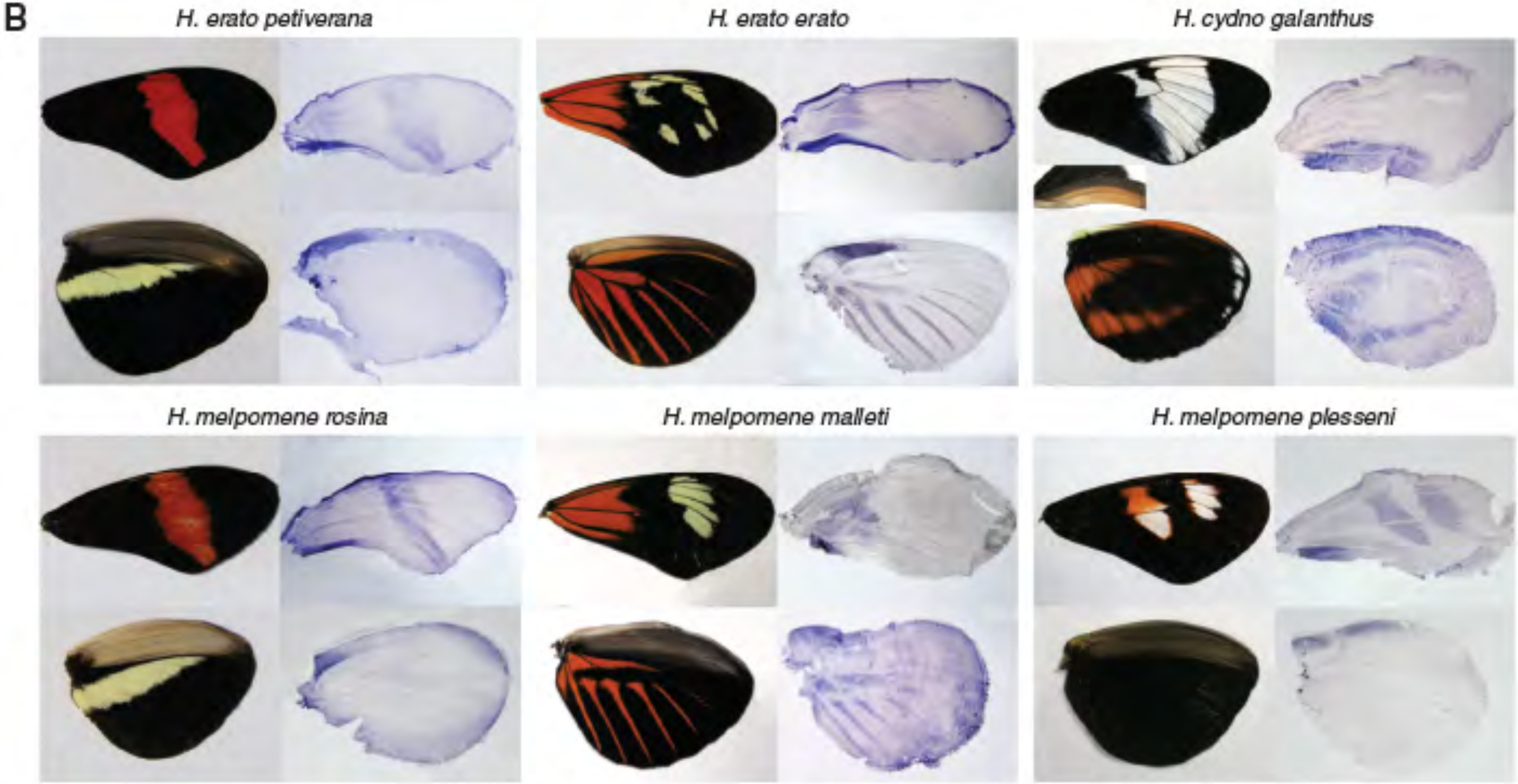
Merrill et al., Proc. Roy. Soc 2012


102
204
76



Peaks of divergence correspond to wing pattern genes

van Belleghem et al., Nature Ecol Evol



H. erato petiverana

H. erato erato

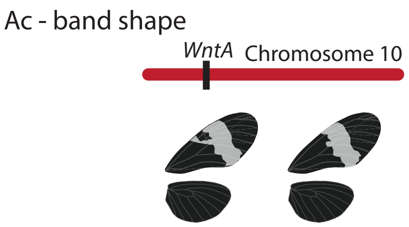
H. cydno galanthus

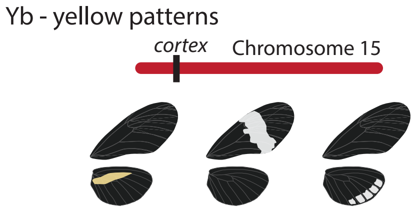
H. melpomene rosina

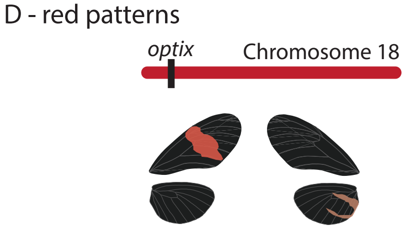
H. melpomene malleti

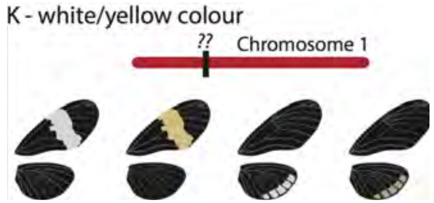
H. melpomene plesseni

Reed et al., 2011 Science









Wing pattern controlled almost entirely by large effect loci

Wild type

knockout

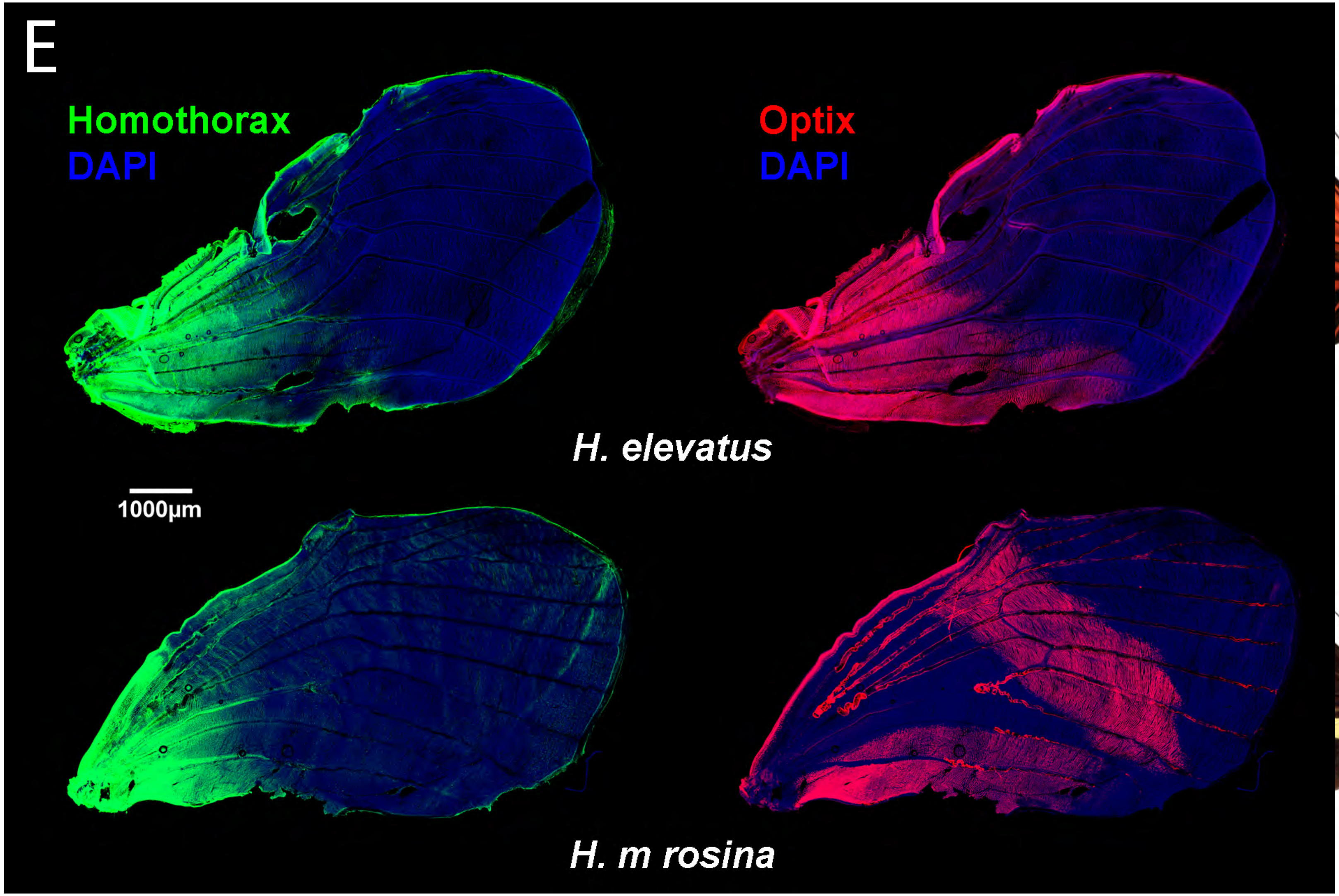
ription

tor,

red in

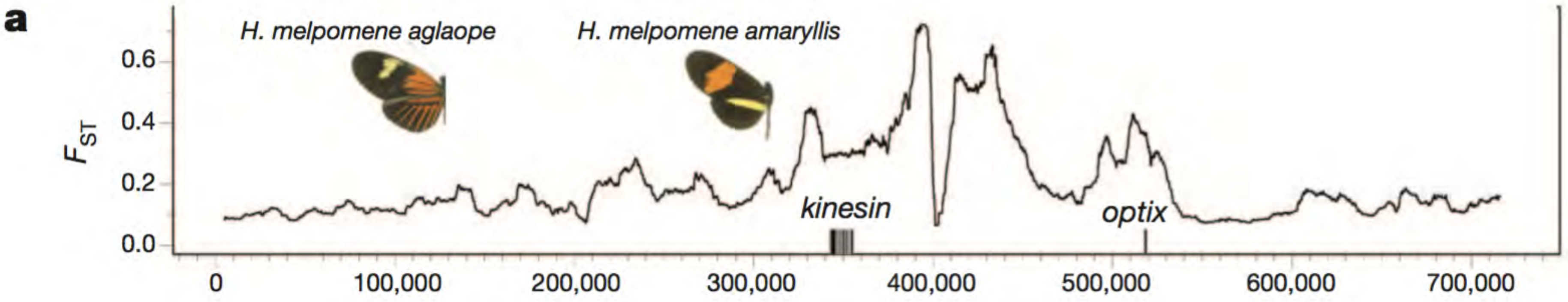
wing

SC (Arnaud



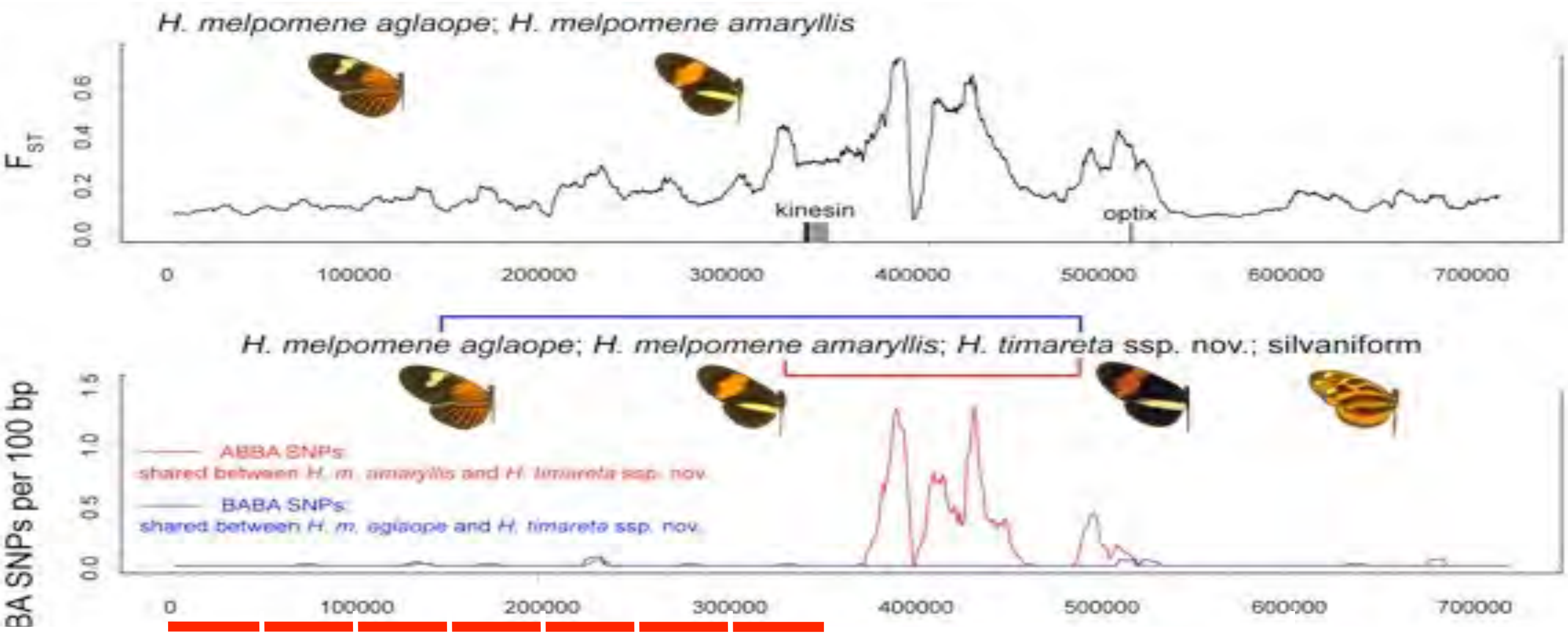
Richard Wallbank

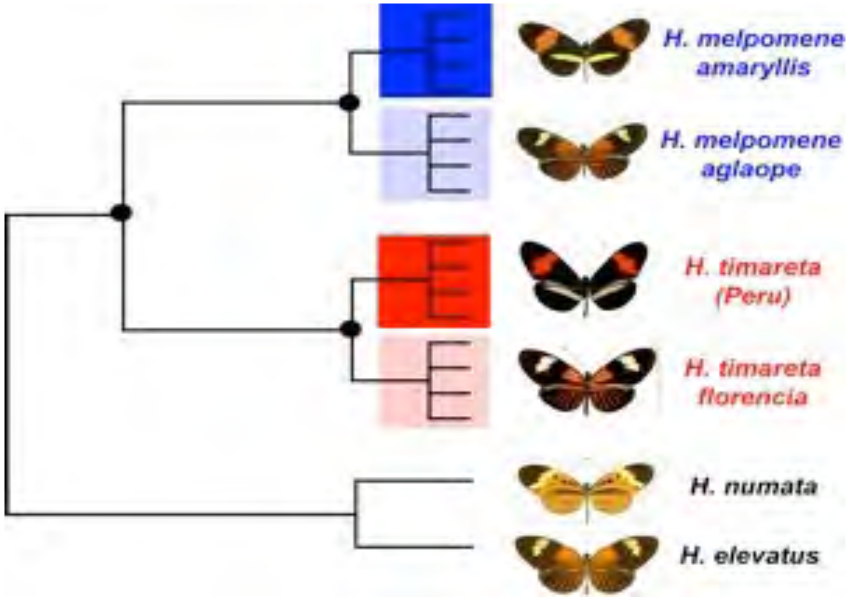
Adaptive introgression



Heliconius Genome Consortium Nature 2012

Phylogenies across *B/D*

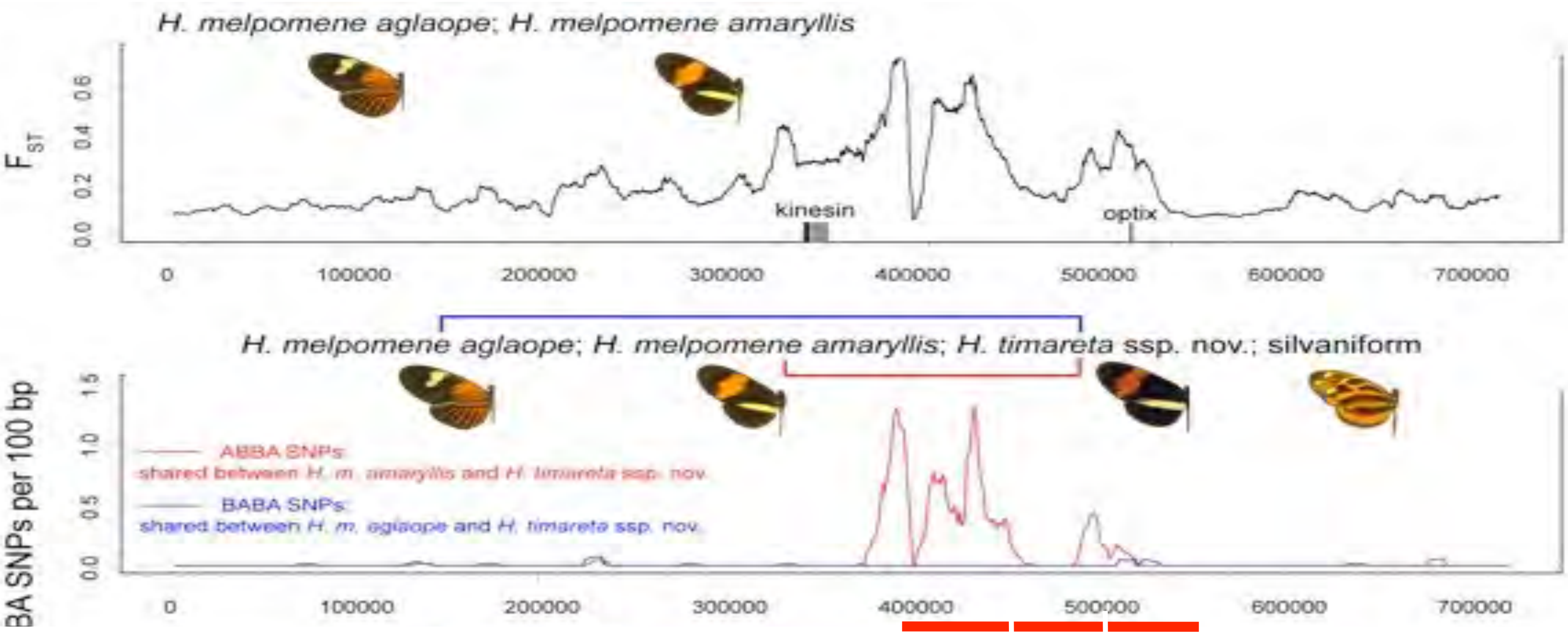


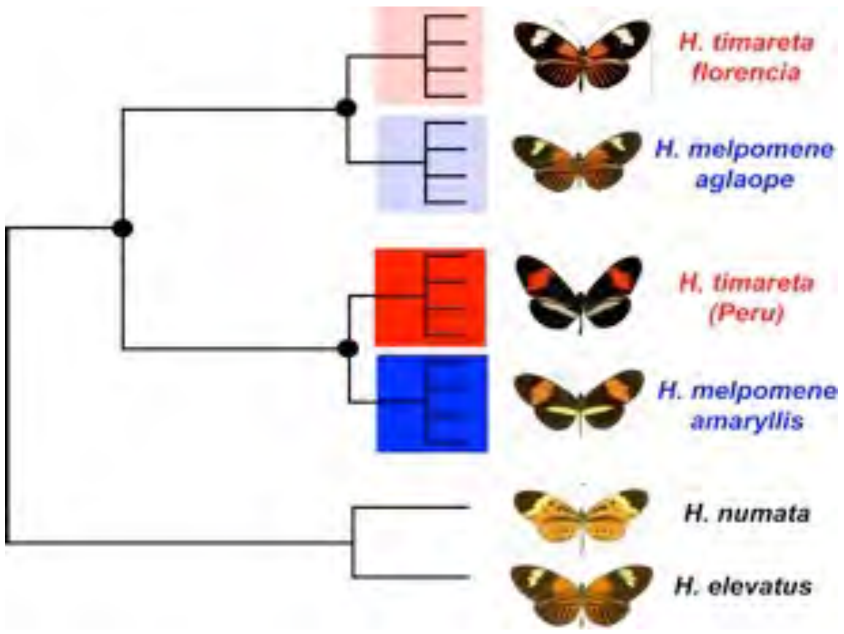


ML tree based
on
50,000 bp

Heliconius Genome Consortium Nature 2012

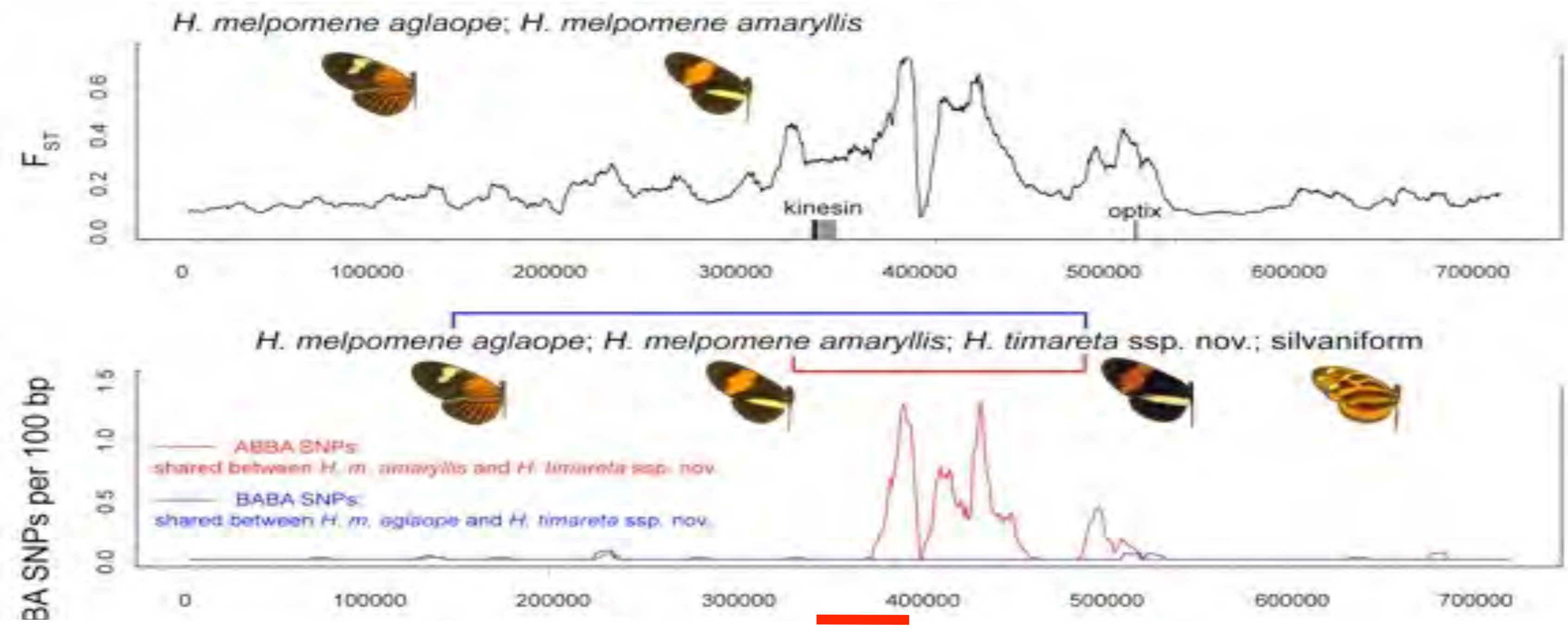
Phylogenies across *B*/*D*

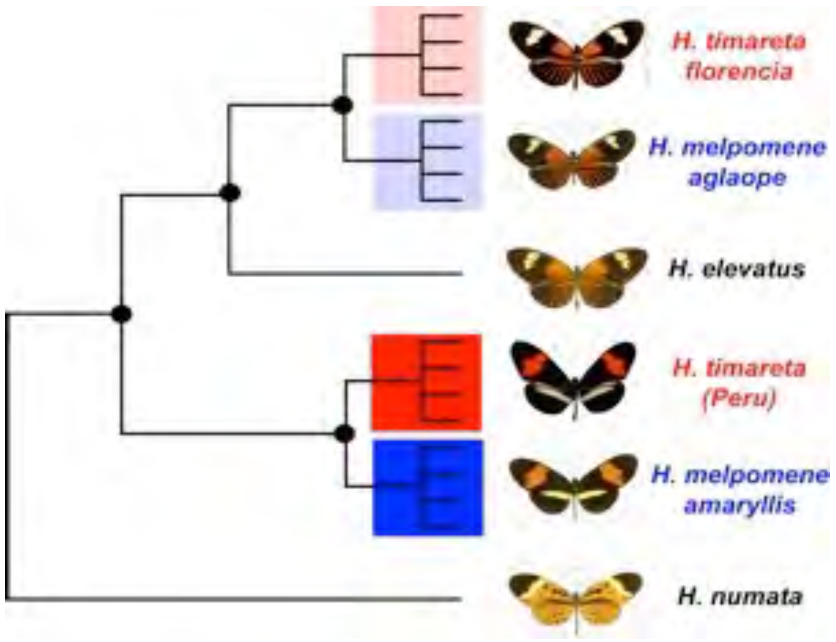




ML tree based on 50,000 bp

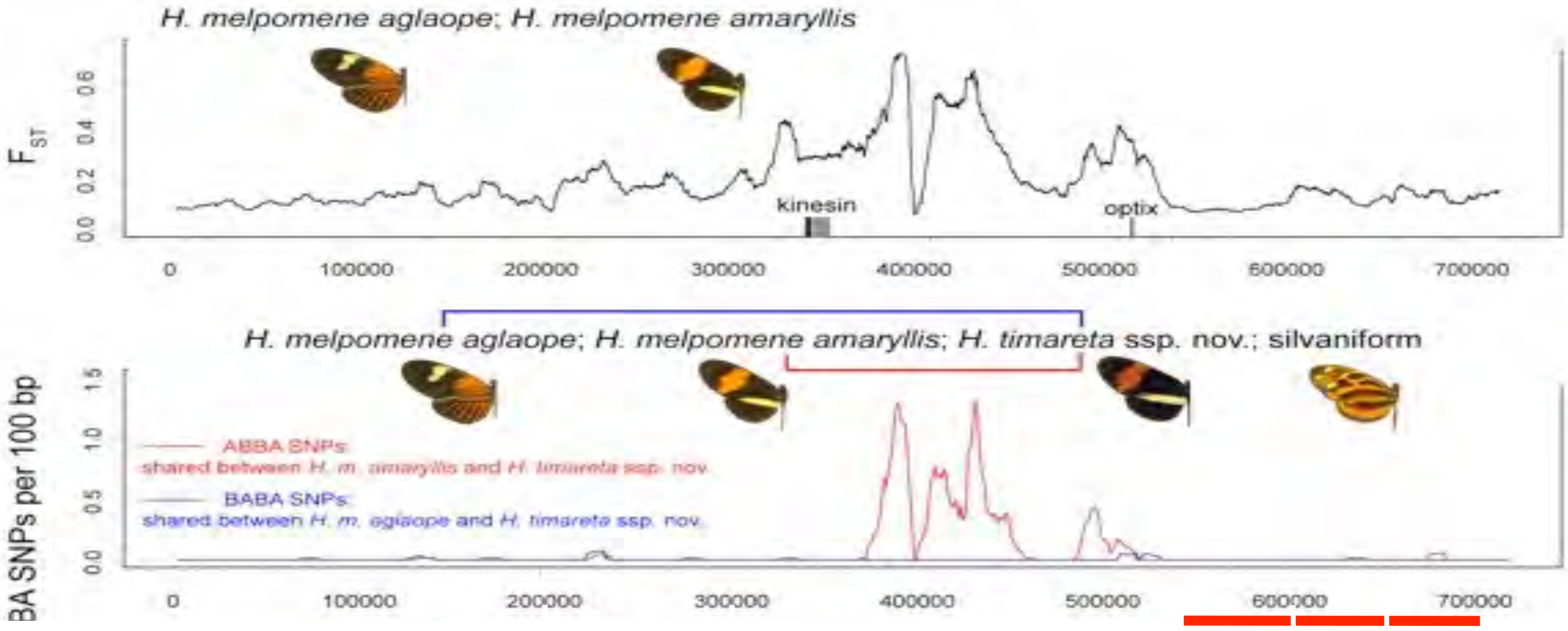
Phylogenies across *B*/*D*

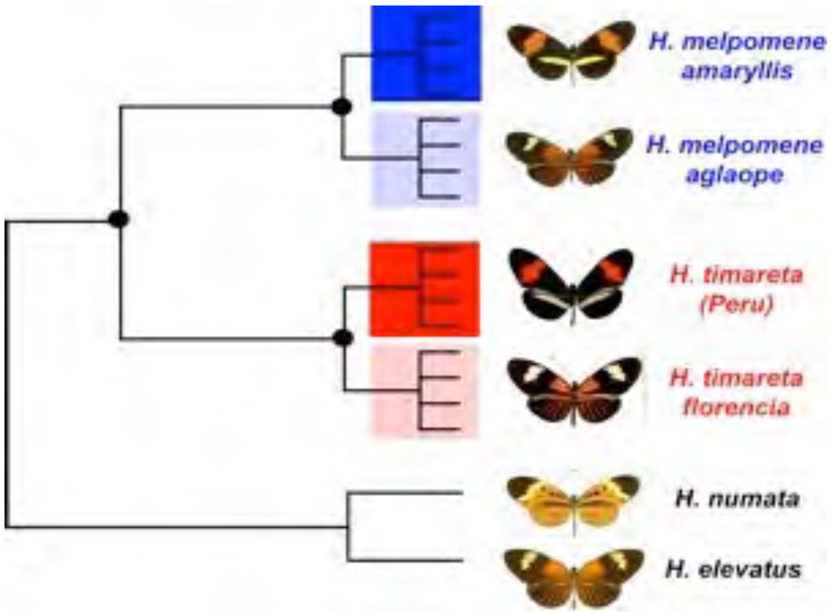




ML tree based
on
50,000 bp

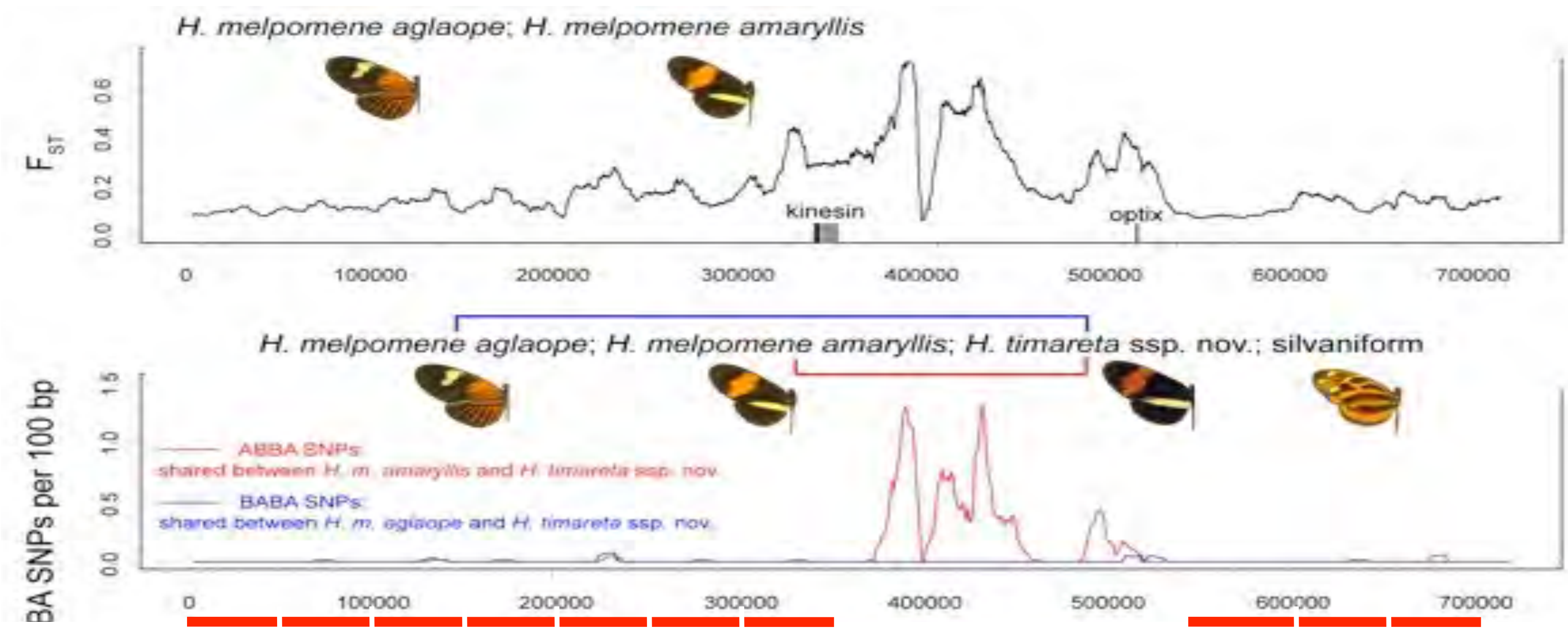
Phylogenies across *B*/*D*

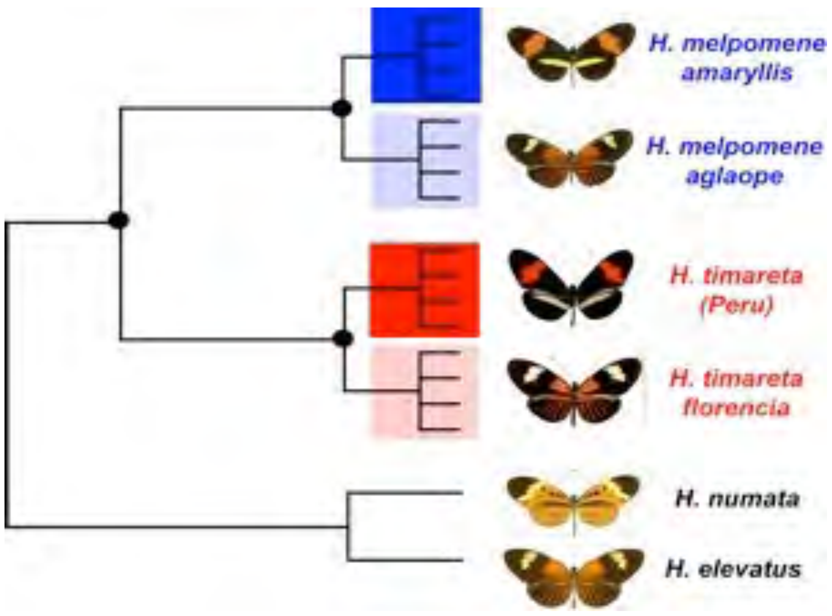




ML tree based on 50,000 bp

Phylogenies across *B*/*D*





ML tree based on 50,000 bp

Okay, so introgression causes mimicry

But mimicry is weird, right?

Novelty can arise through introgression and recombination

NNBB

Heliconius heurippa

NNbb

Heliconius cydno cordula

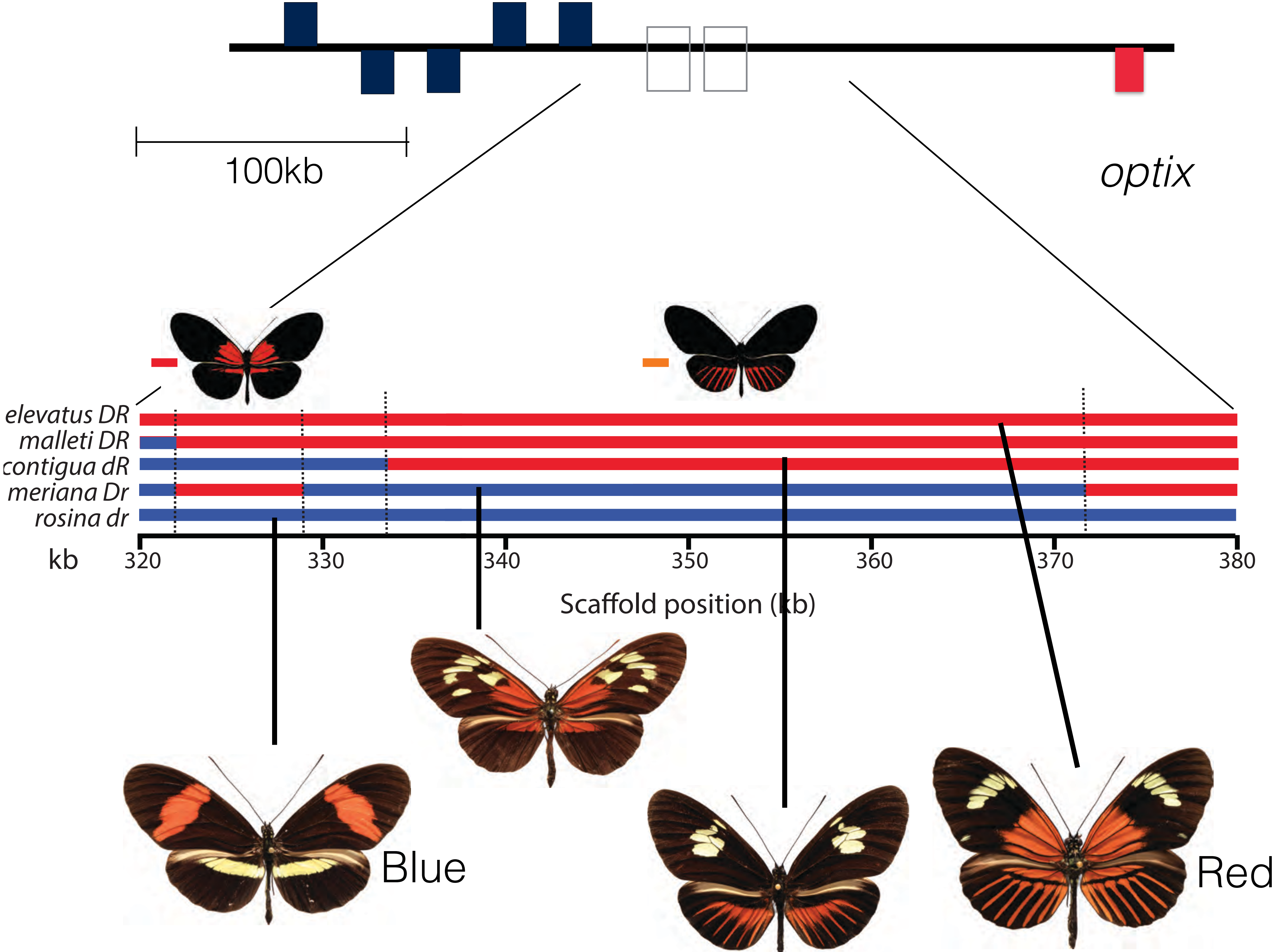
nnBB

Heliconius melpomene melpomene

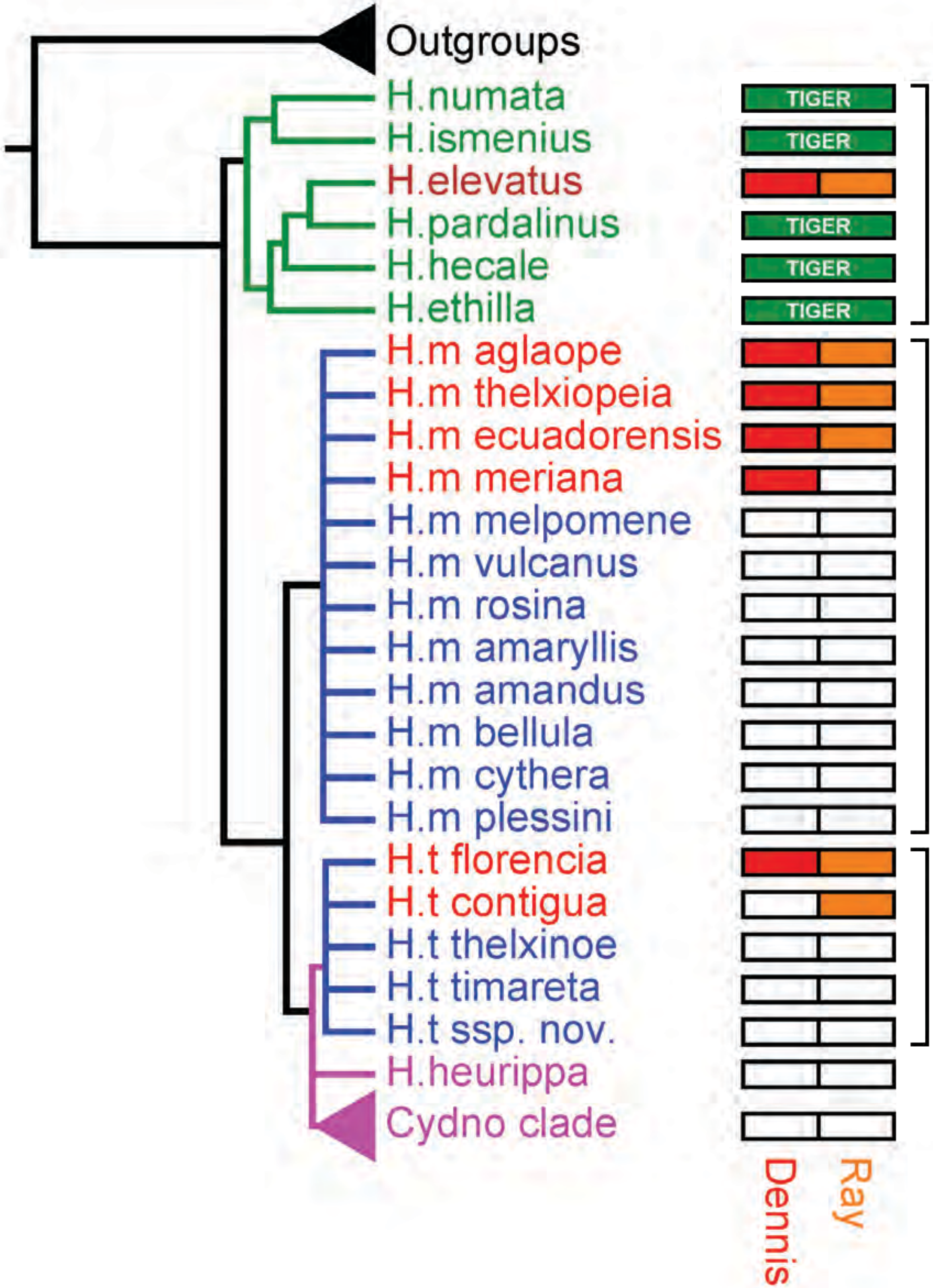
Camilo Salazar

Mavarez et al., Nature 2006

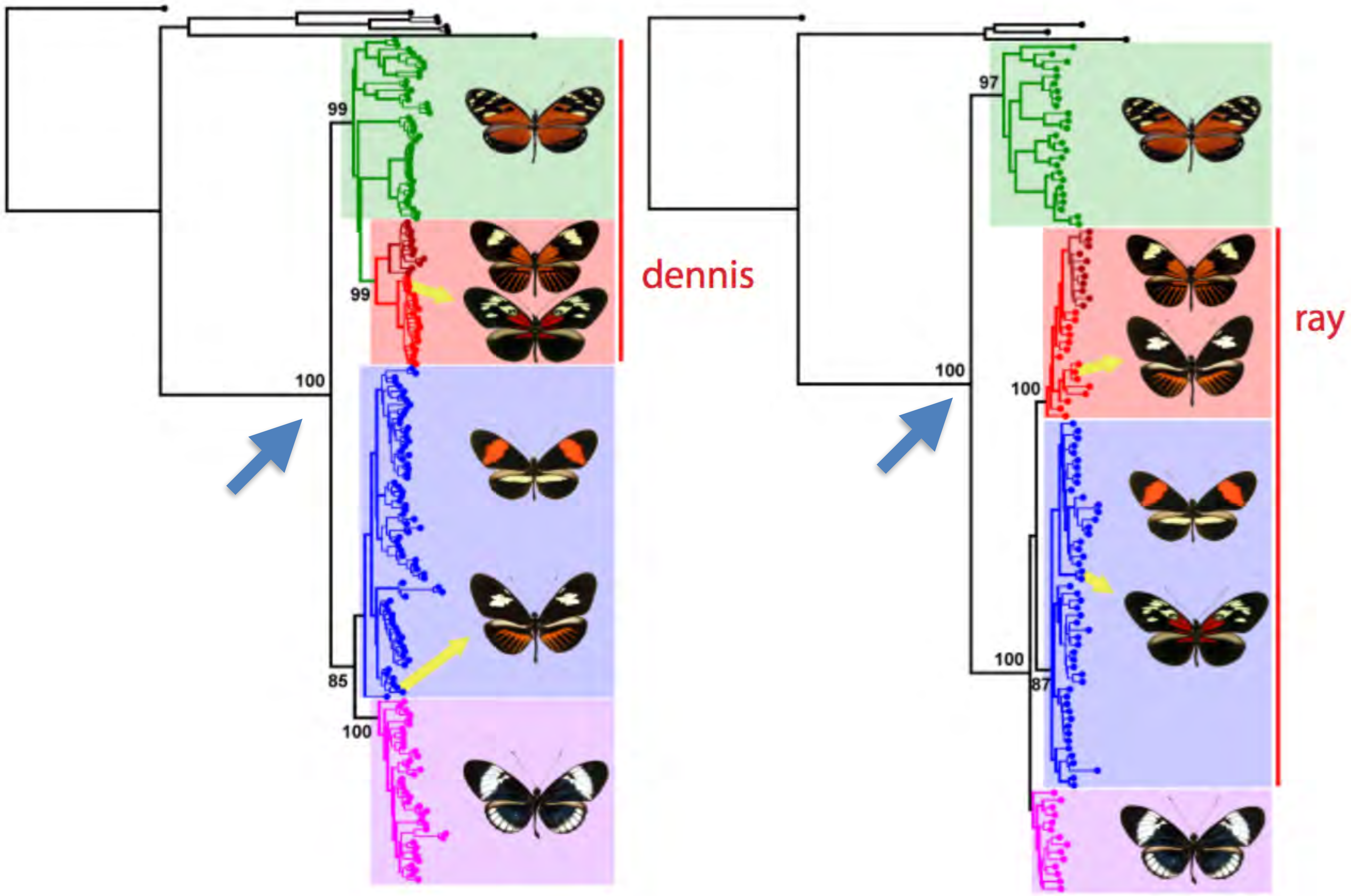
Wallbank et al., PLoS Biology 2016



Wallbank et al., PLoS Biology 2016

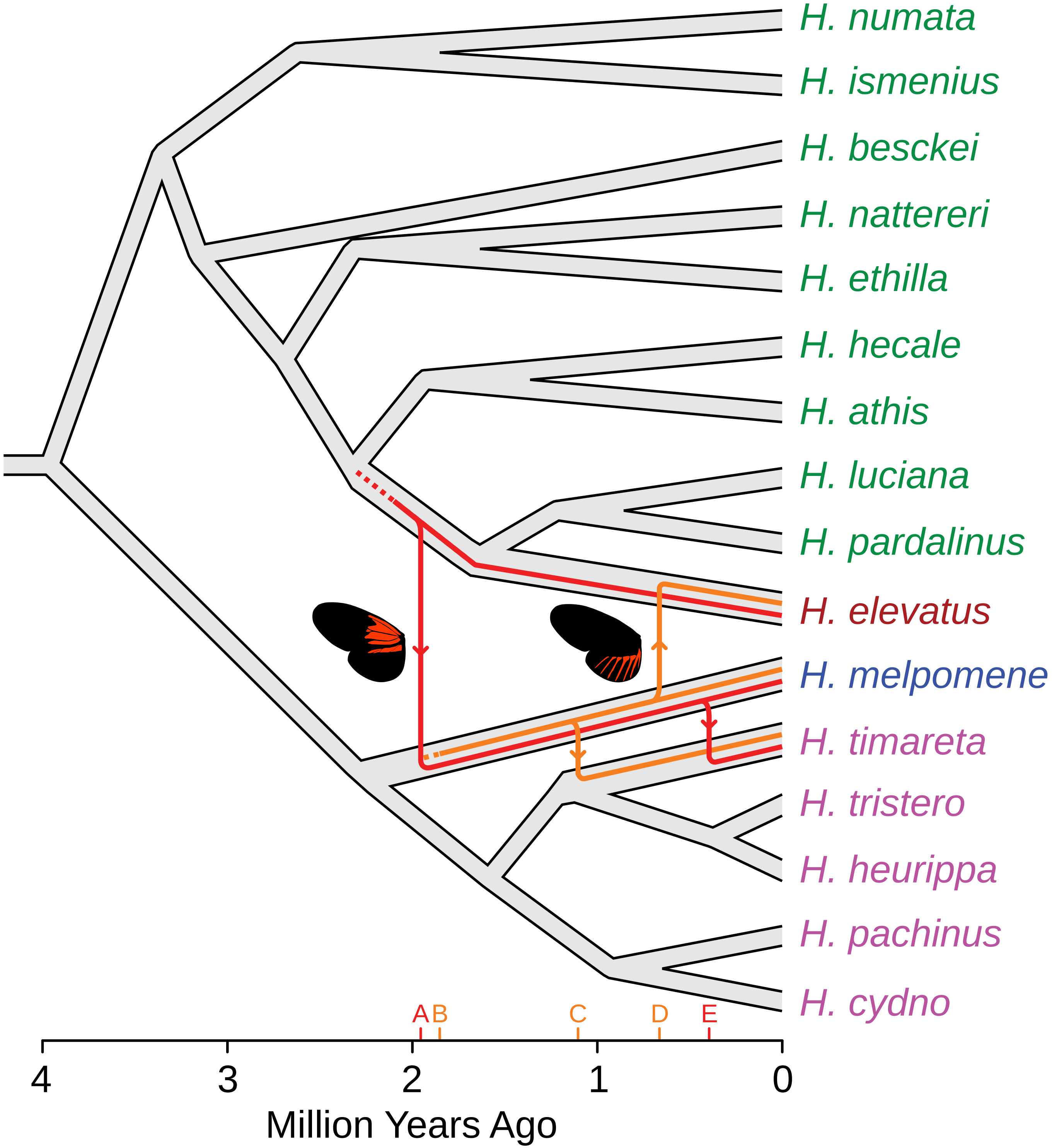


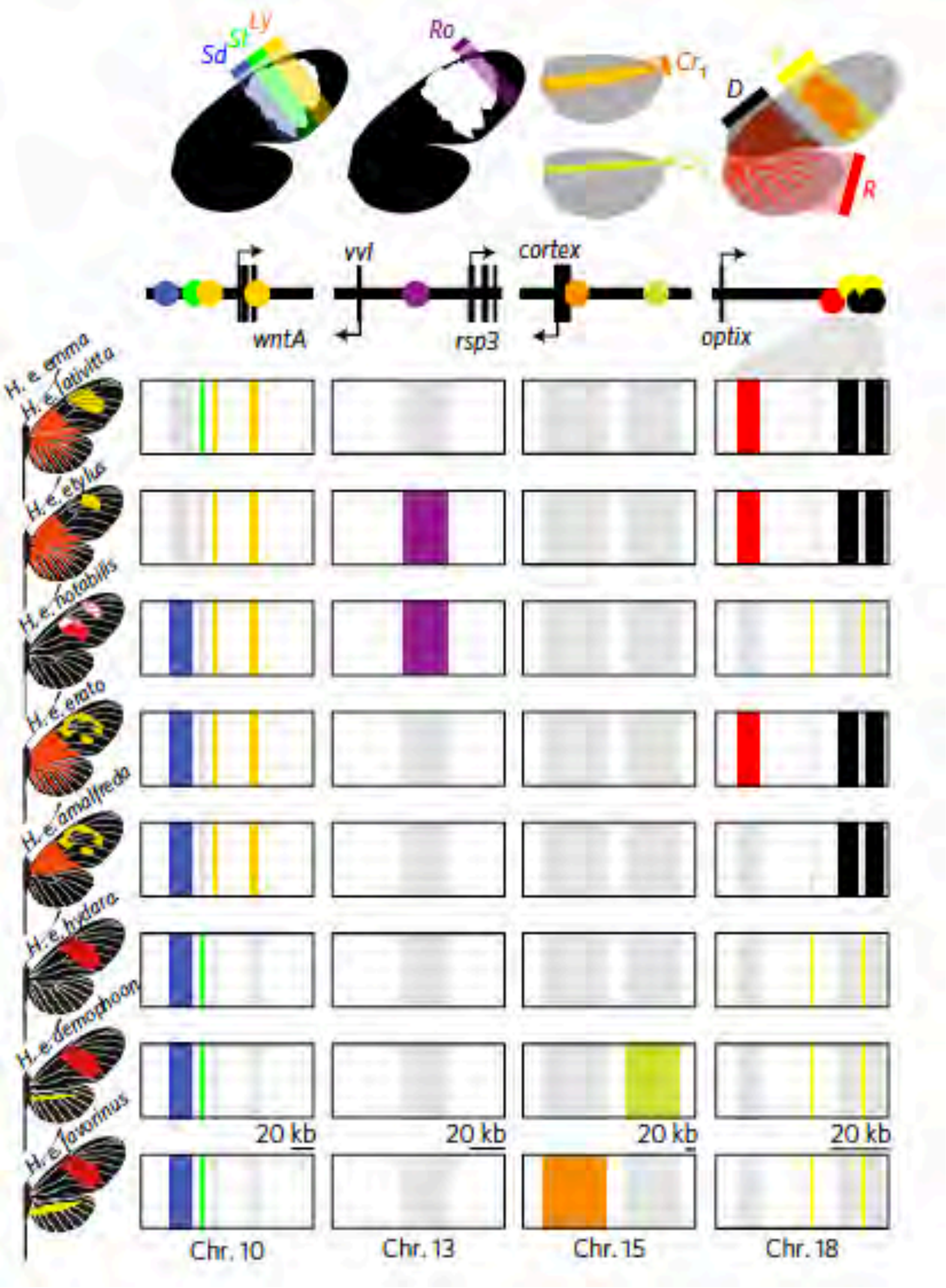
Wallbank et al., PLoS Biology 2016

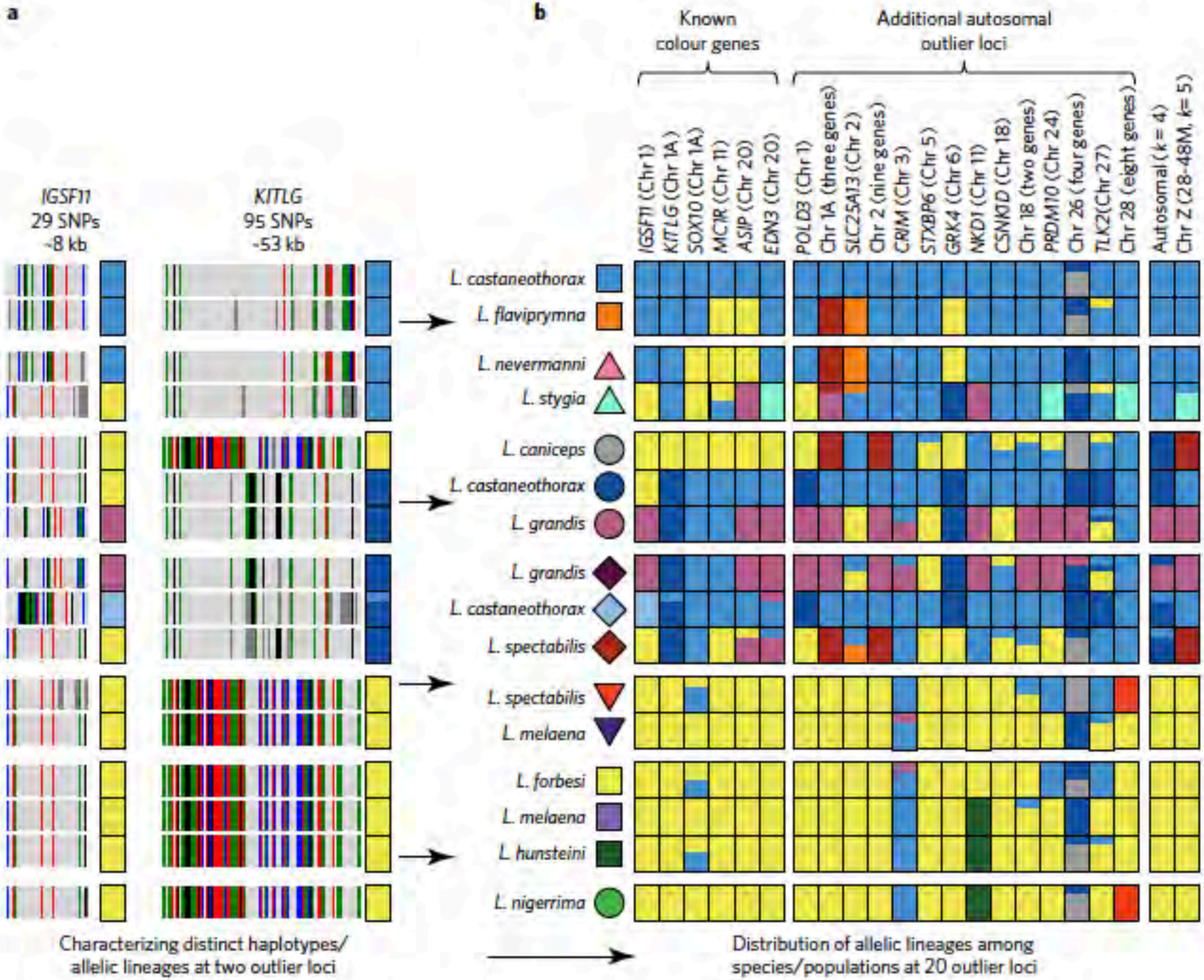


Generate dated trees using this node as a reference point

Wallbank et al., PLoS Biology 2016



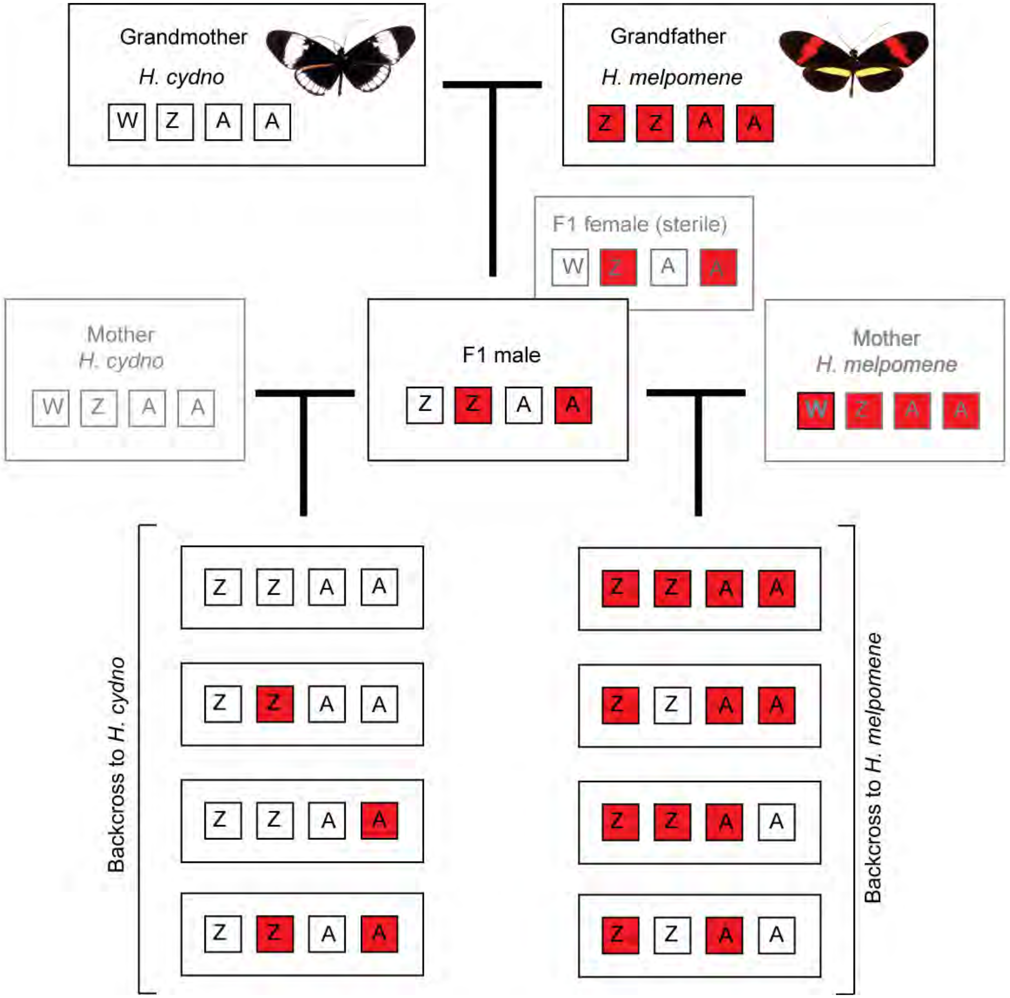

Sd
Sl
Ly
Ro
Cr1
D
Y
R
vvl
cortex
wntA
rsp3
optix
H. e. emma
H. e. lativitta
H. e. etylus
H. e. notabilis
H. e. erato
H. e. amalfreda
H. e. hydara
H. e. demophoon
H. e. favorinus
20 kb
20 kb
20 kb
20 kb
Chr. 10
Chr. 13
Chr. 15
Chr. 18

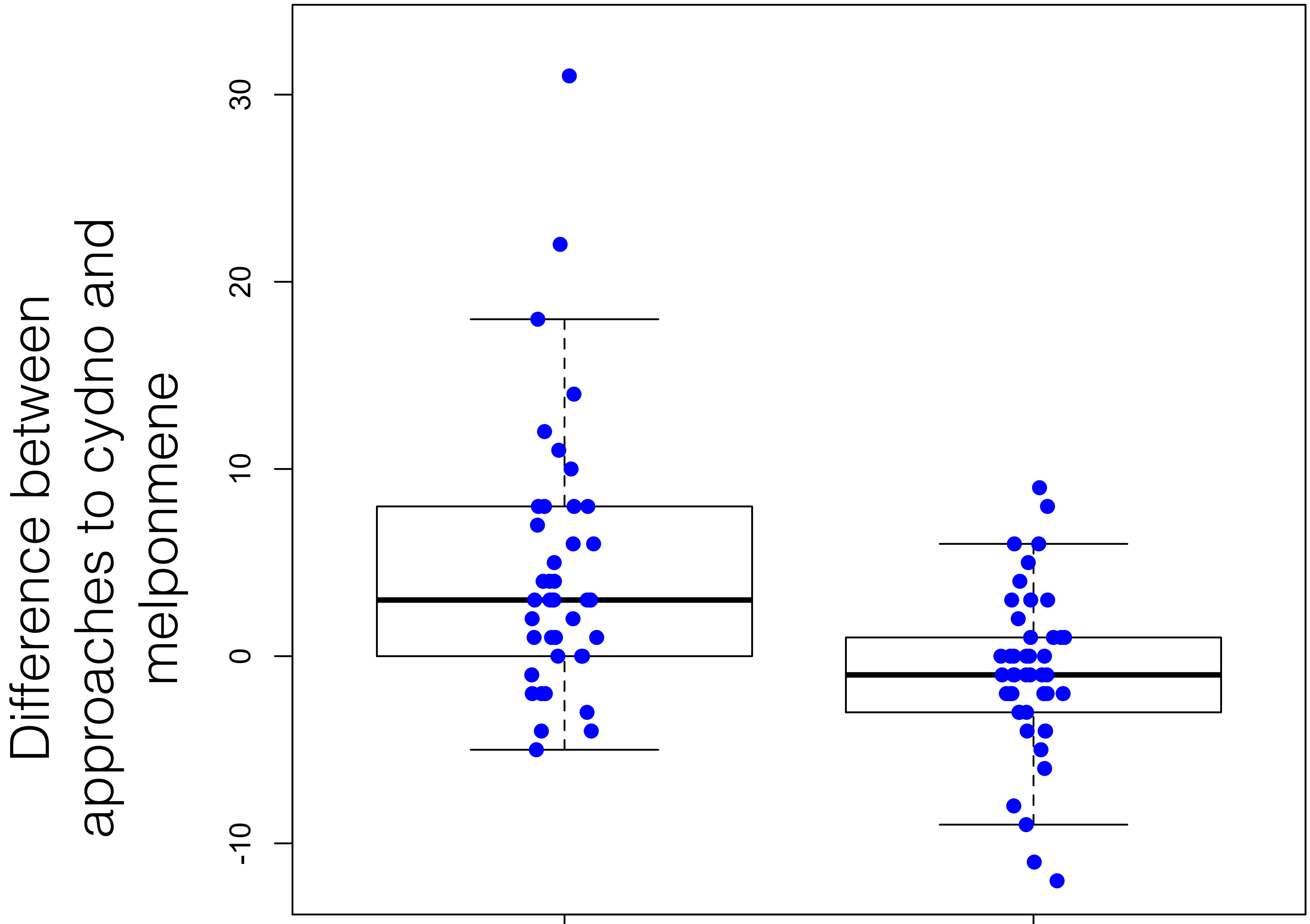

a
b
Known
colour genes
Additional autosomal
outlier loci
IGSF11 (Chr 1)
KITLG (Chr 1A)
SOX10 (Chr 1A)
MC1R (Chr 11)
ASIP (Chr 20)
EDN3 (Chr 20)
POLD3 (Chr 1)
Chr 1A (three genes)
SLC25A13 (Chr 2)
Chr 2 (nine genes)
CRIM (Chr 3)
STXBP6 (Chr 5)
GRK4 (Chr 6)
NKD1 (Chr 11)
CSNK1D (Chr 18)
Chr 18 (two genes)
PRDM10 (Chr 24)
Chr 26 (four genes)
TLK2(Chr 27)
Chr 28 (eight genes)
Autosomal (k = 4)
Chr Z (28-48M, k = 5)
IGSF11
29 SNPs
~8 kb
KITLG
95 SNPs
~53 kb
L. castaneothorax
L. flaviprymna
L. nevermanni
L. stygia
L. caniceps
L. castaneothorax
L. grandis
L. grandis
L. castaneothorax
L. spectabilis
L. spectabilis
L. melaena
L. forbesi
L. melaena
L. hunsteini
L. nigerrima
Characterizing distinct haplotypes/
allelic lineages at two outlier loci
Distribution of allelic lineages among
species/populations at 20 outlier loci

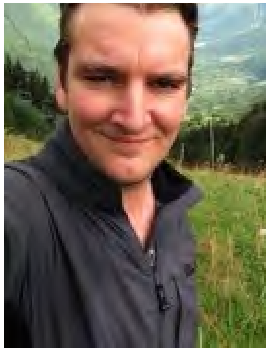
What about behaviour?

H. melpomene X H. cydno

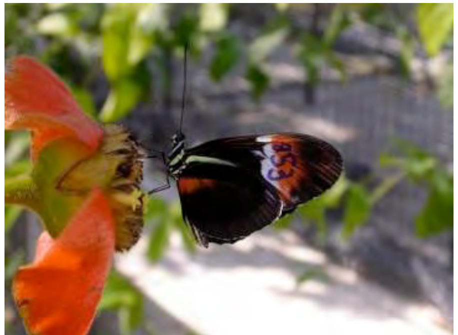
Backcross design:

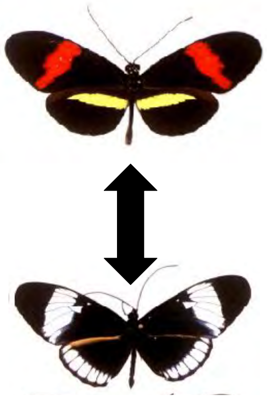


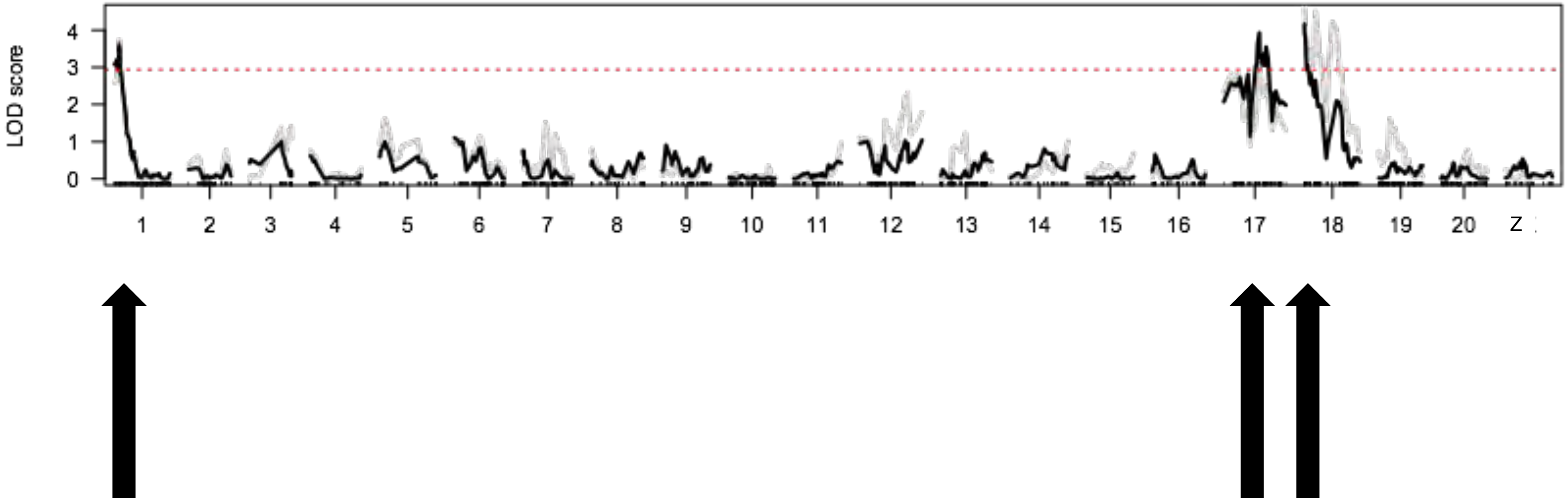




Richard
Merrill



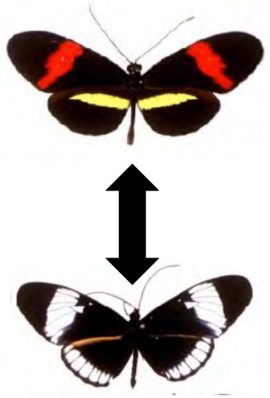


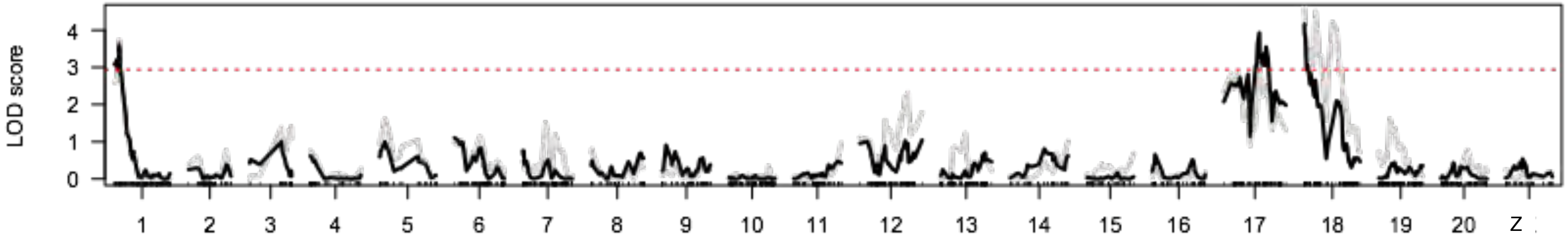


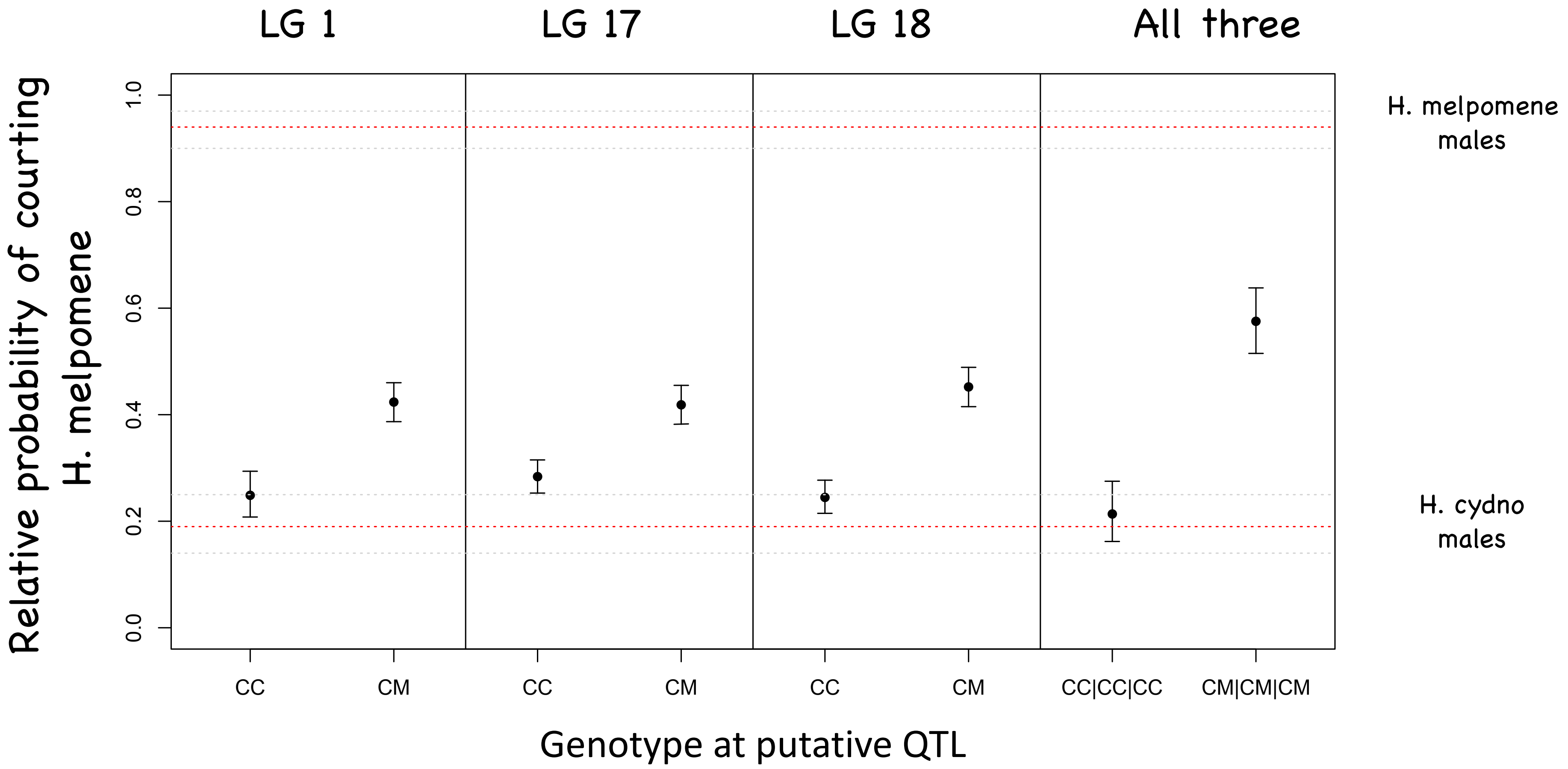
Significant QTL detected on three linkage groups

5% genome-wide significance threshold

Richard Merrill unpub.







Together explain ~ 50% measured differences between H. melpomene and H. cydno

NNbb

Heliconius cydno cordula

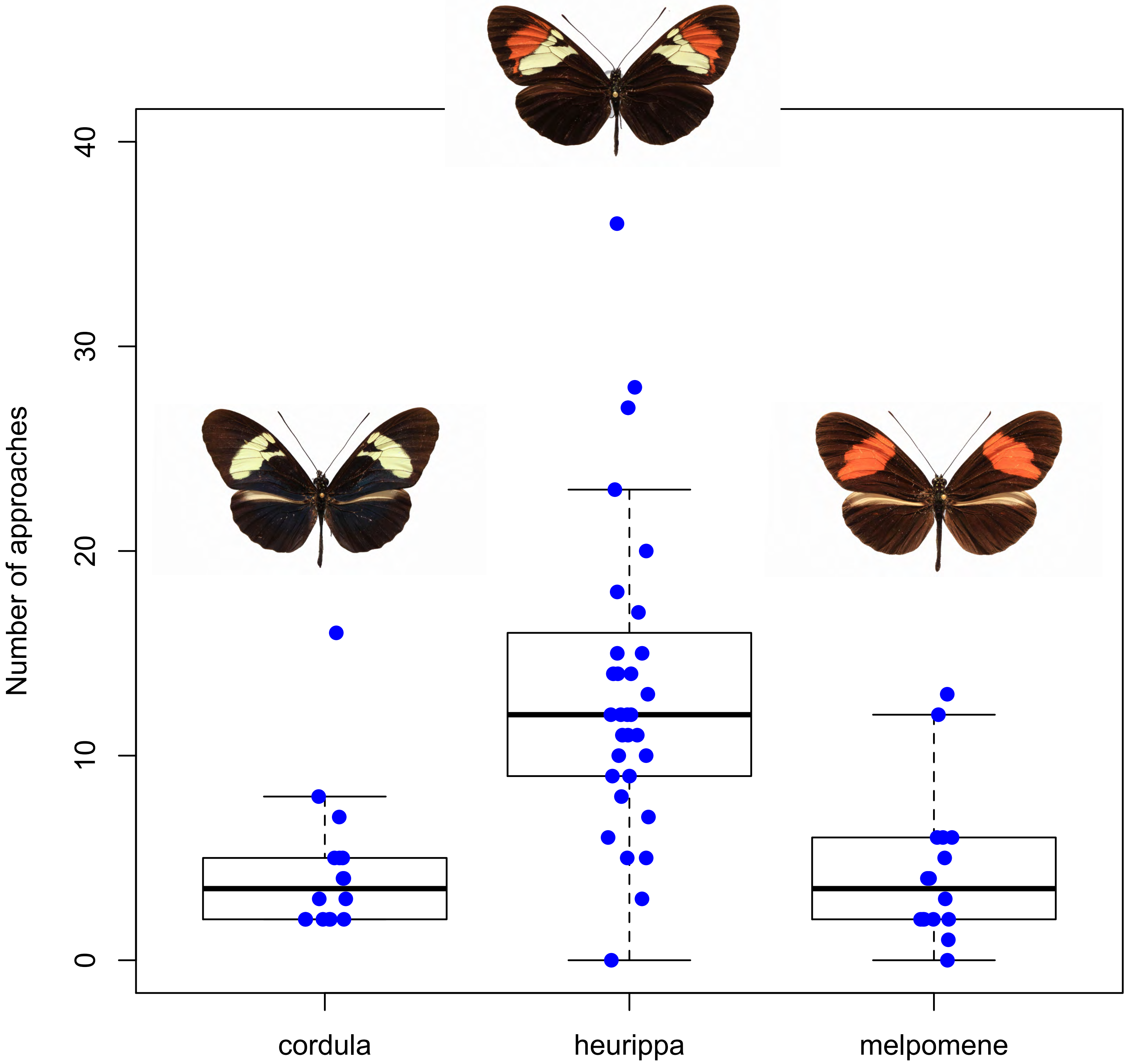
NNBB

Heliconius heurippa

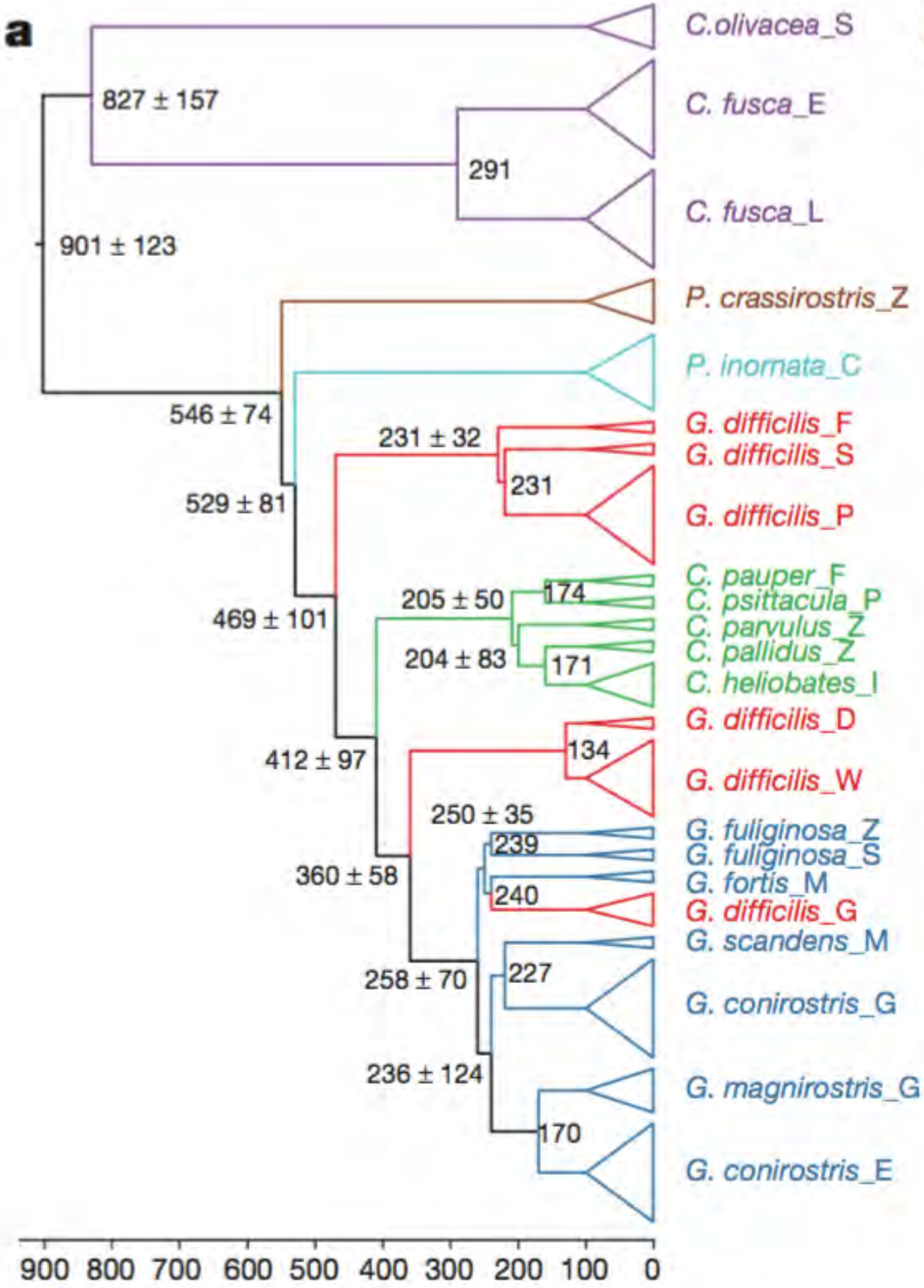
nnBB

Heliconius melpomene melpomene

Mavarez et al., Nature 2006

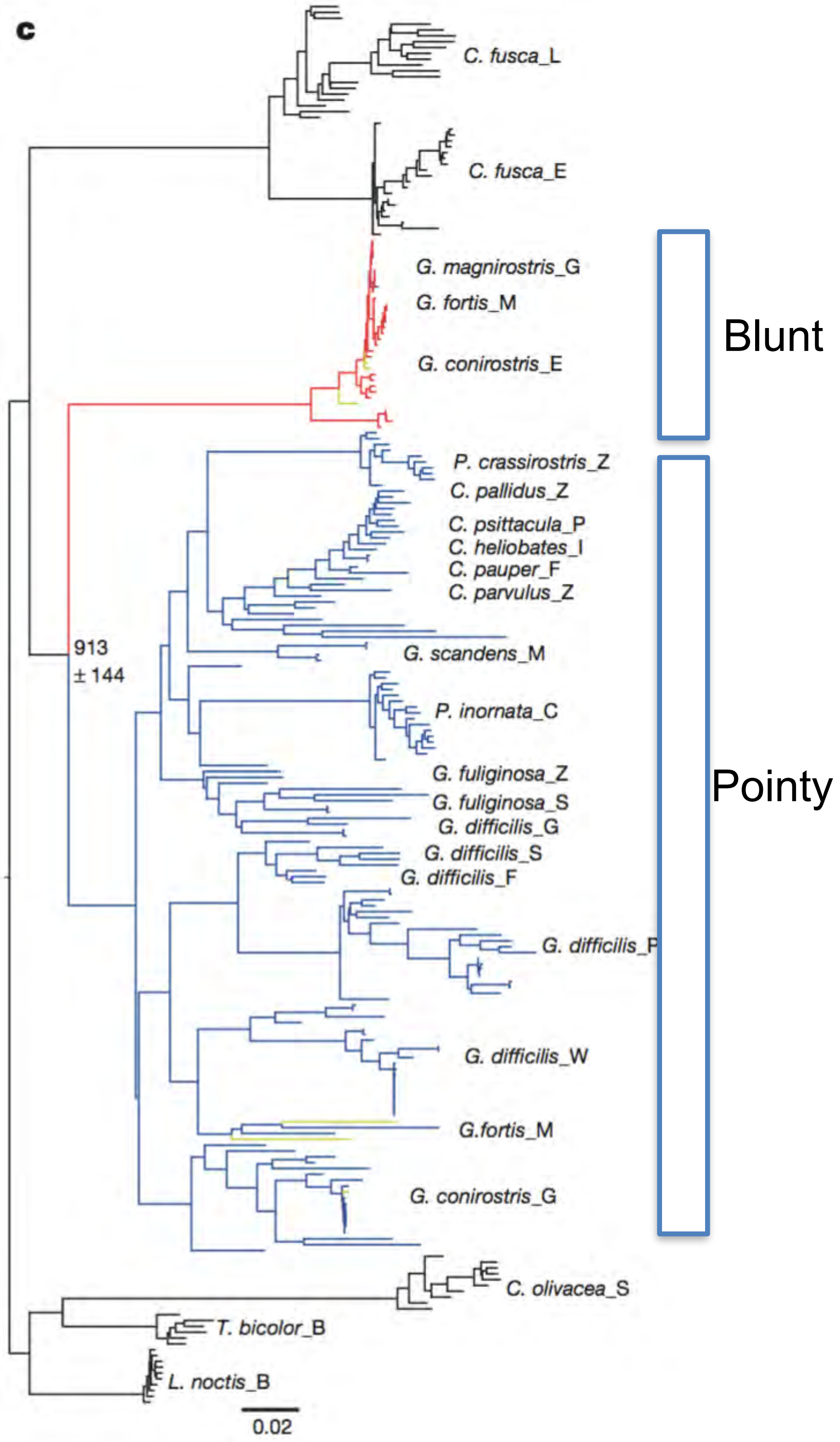


Melo et al., Evolution 2009



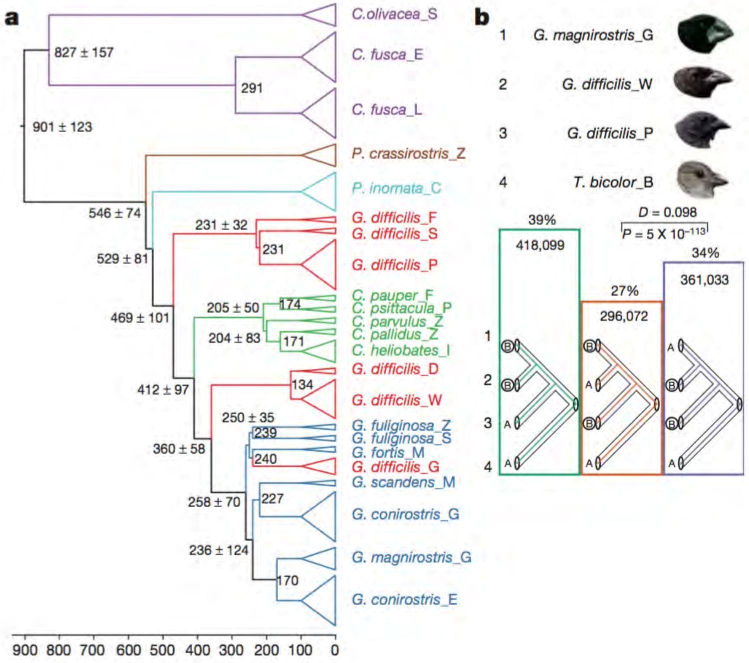
Lamichhaney et al., Nature 2015

ALX1 associated with beak shape



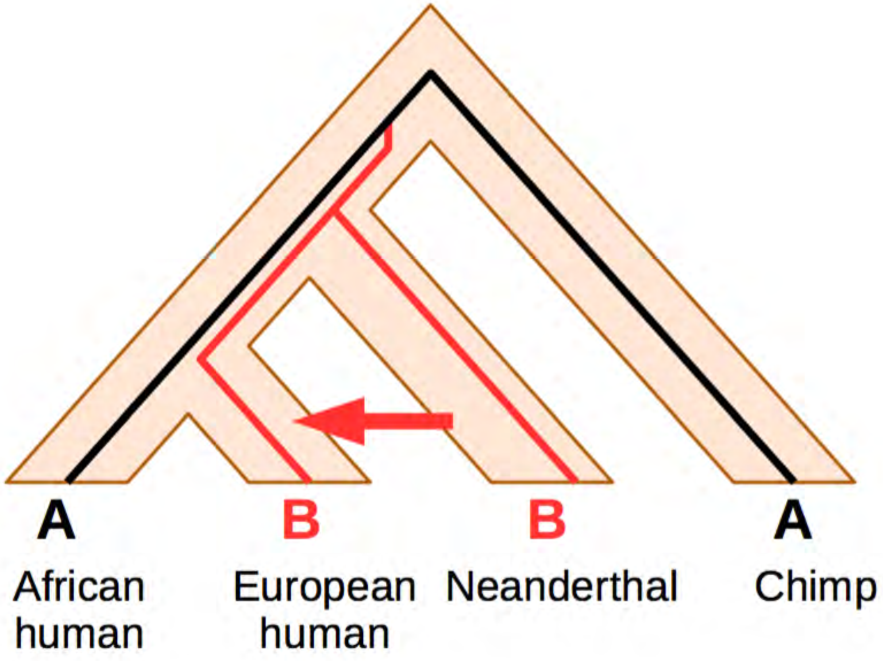
Blunt

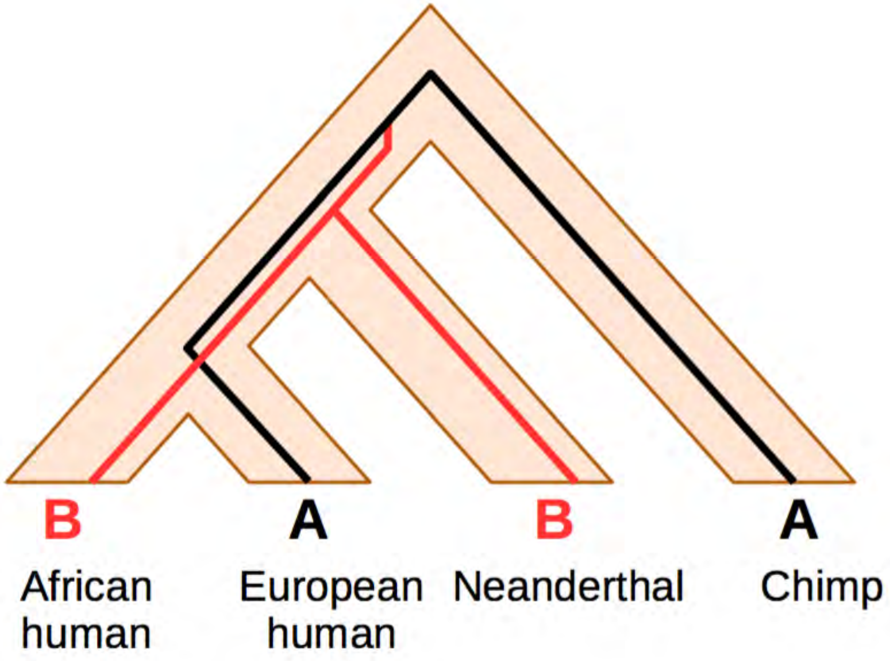
Pointy



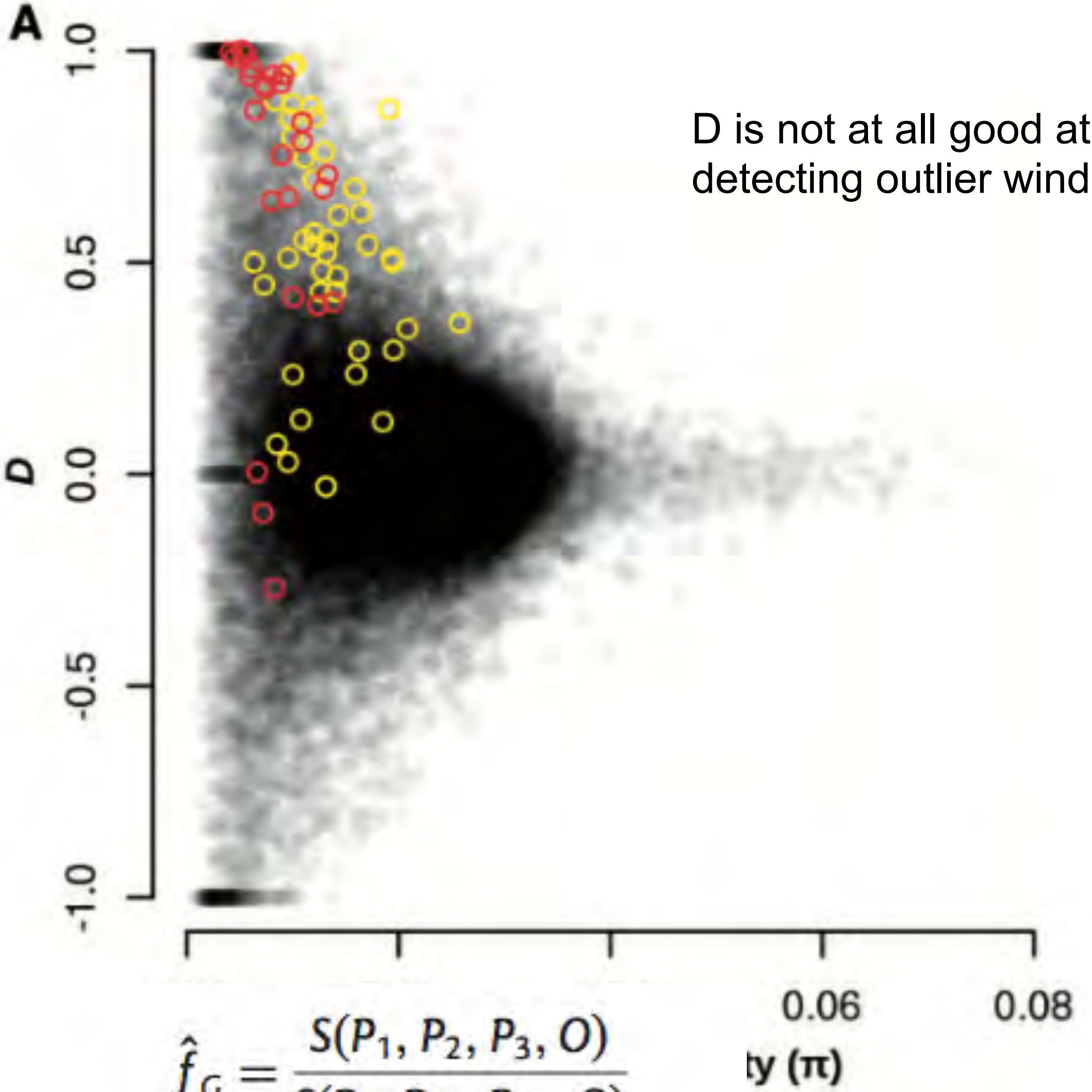
Most of these studies use phenotype associations to identify introgressed loci

But can we identify them a priori using the ABBA-BABA method?





Green *et al.* 2010 *Science* 328:710-722



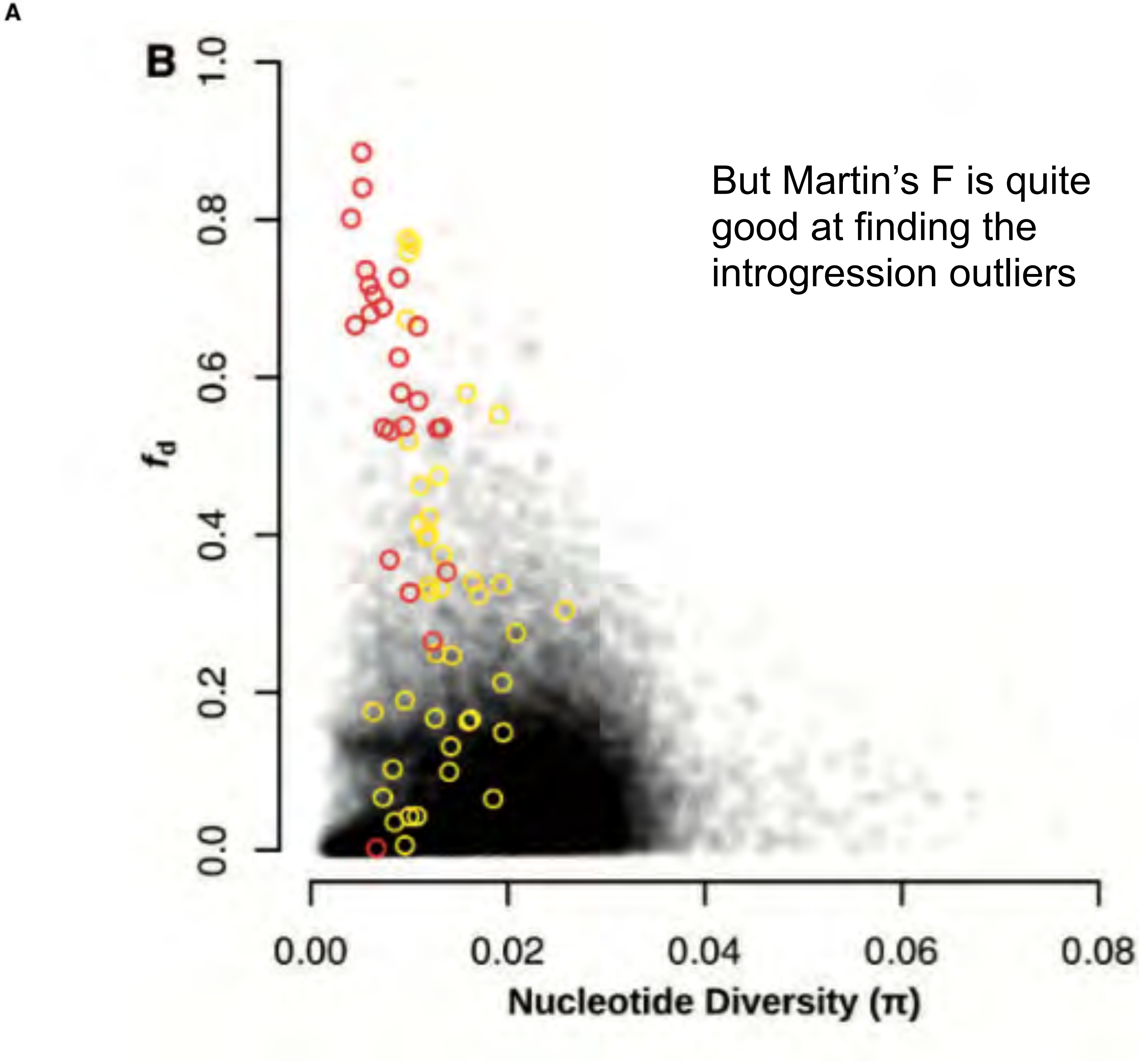
D is not at all good at
detecting outlier windows

ites (the denominator)

$$\hat{f}_G = \frac{S(P_1, P_2, P_3, O)}{S(P_1, P_{3a}, P_{3b}, O)}$$

Where s is numerator from the D equation
f is the fraction of introgression compared to maximum possible

Martin et al., MBE 2014



But Martin’s F is quite good at finding the introgression outliers

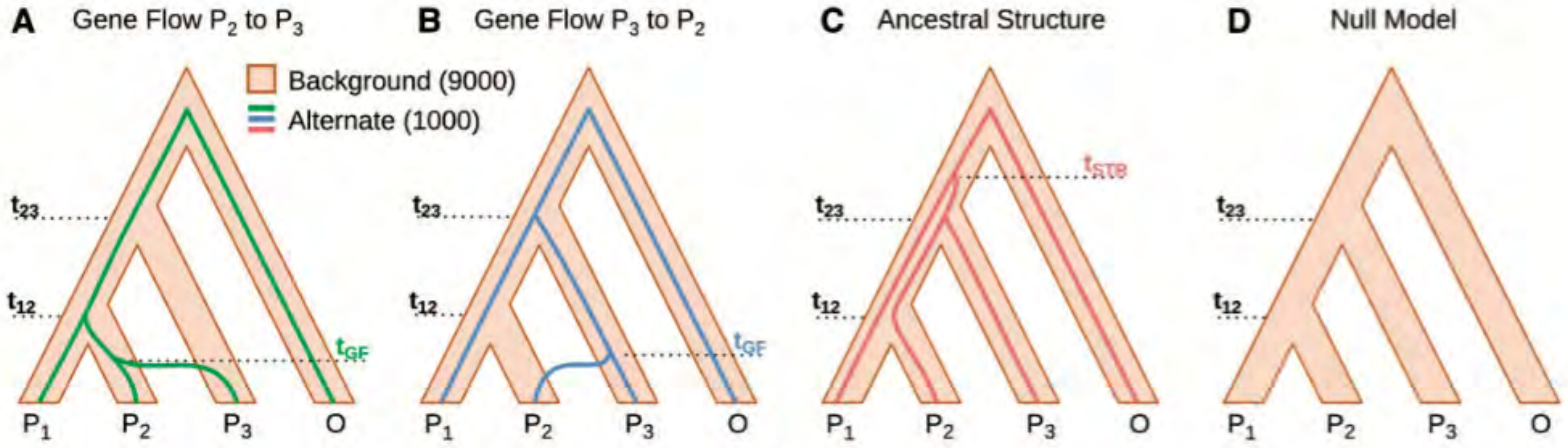
Martin et al., MBE 2014

Do *Heliconius* butterfly species exchange mimicry alleles?

Joel Smith, Marcus R. Kronforst

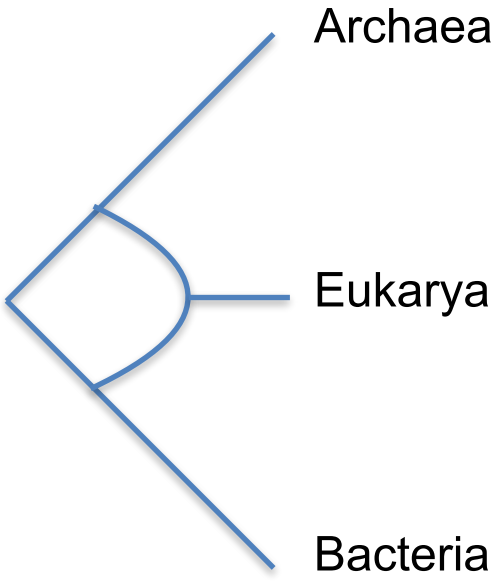
Published 17 July 2013. DOI: 10.1098/rsbl.2013.0503

- Smith and Kronforst argued that introgression could be inferred where ABBA-BABA outliers showed lower Dxy compared to genome-wide average


A
Gene Flow P_2 to P_3
Background (9000)
Alternate (1000)
t_{23}
t_{12}
t_{GF}
P_1
P_2
P_3
O
B
Gene Flow P_3 to P_2
t_{23}
t_{12}
t_{GF}
P_1
P_2
P_3
O
C
Ancestral Structure
t_{STR}
t_{23}
t_{12}
P_1
P_2
P_3
O
D
Null Model
t_{23}
t_{12}
P_1
P_2
P_3
O

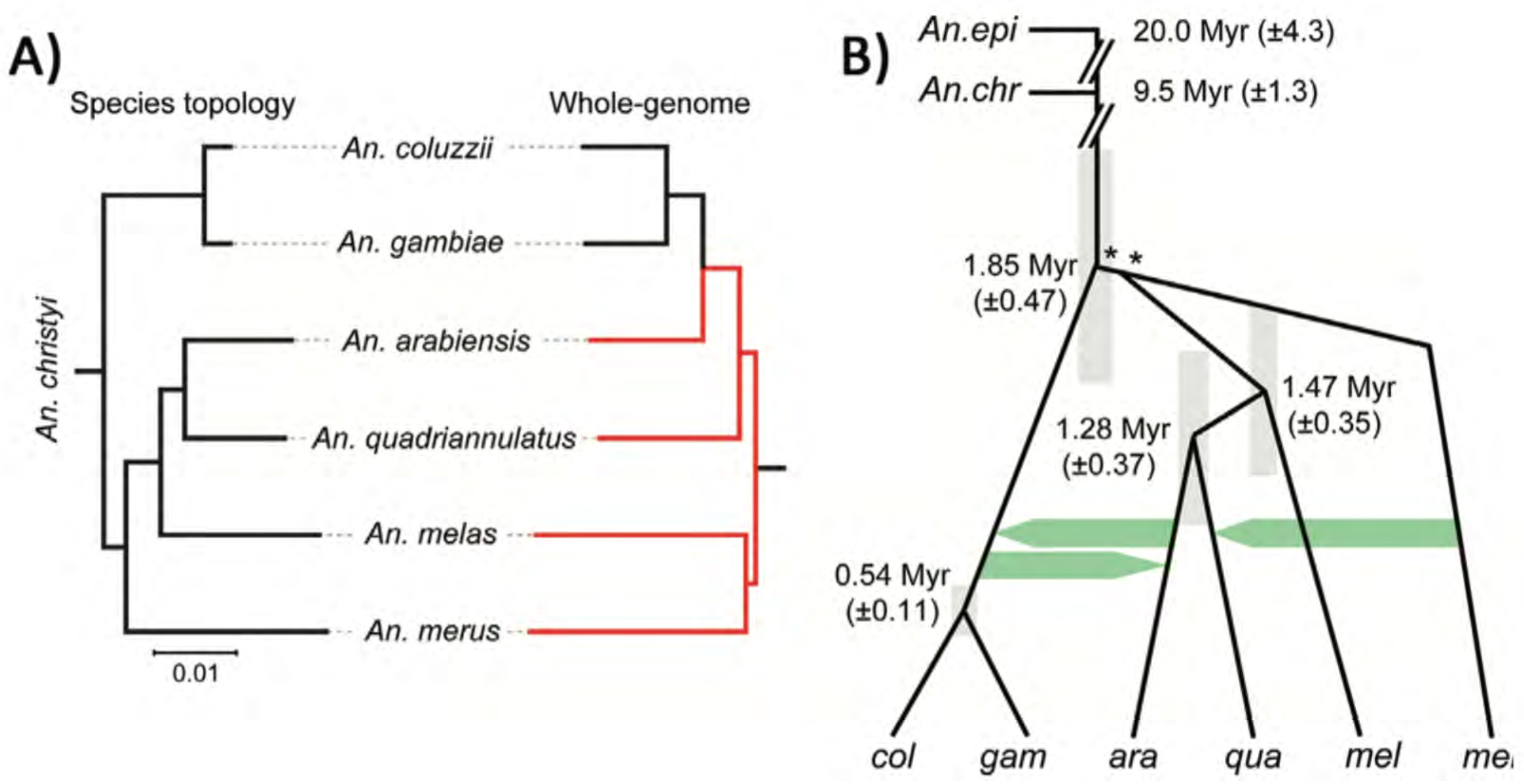
- Be wary of window based D statistics
- F is better than D…
- Sampling design is very important!

Implications for tree-thinking

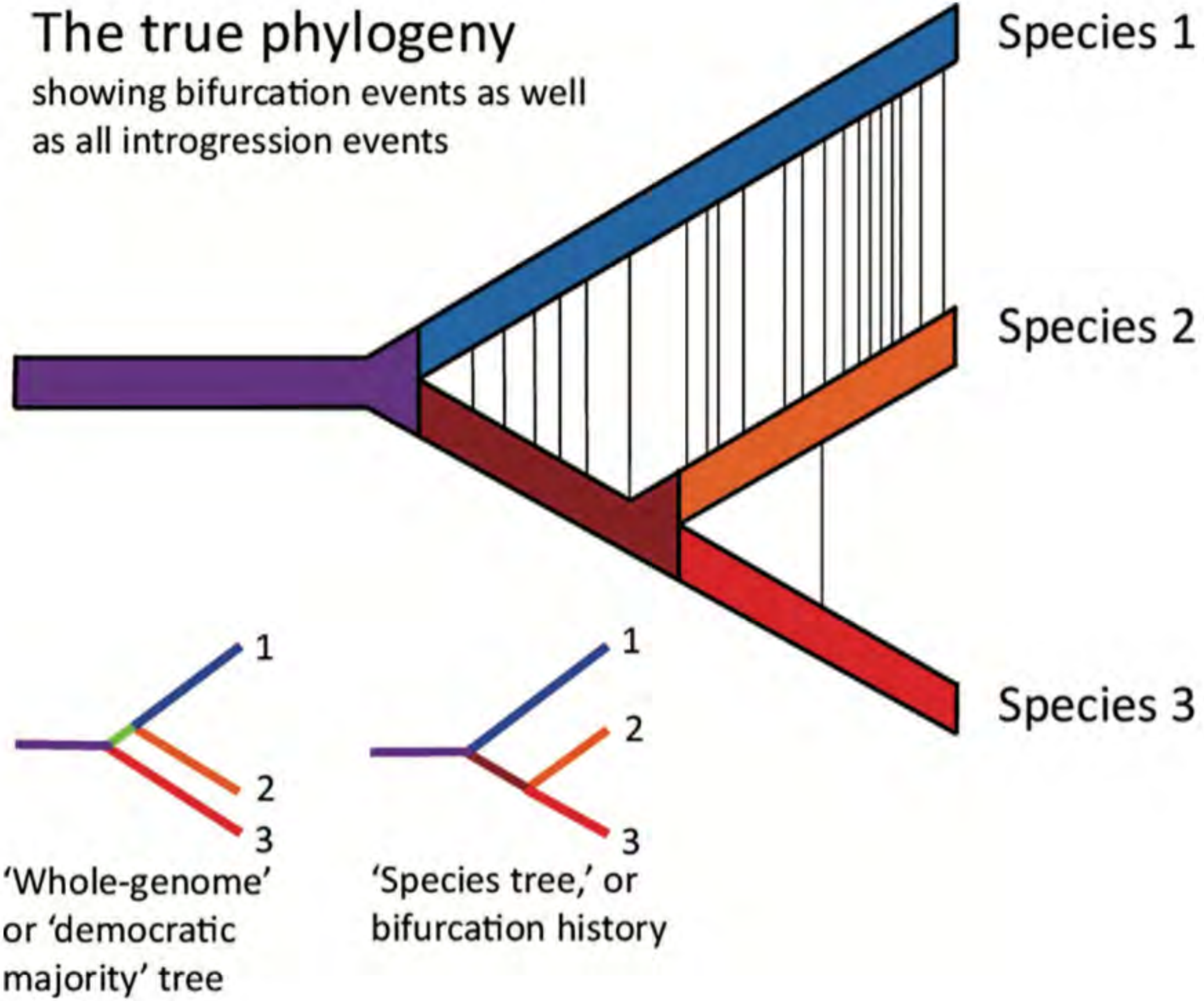


The tree of life is reticulated

Implications for tree-thinking



Mallet, Hahn and Besansky BioEssays 2015
Hahn and Nakhleh Evolution 2015


The true phylogeny
showing bifurcation events as well
as all introgression events
Species 1
Species 2
Species 3
1
2
3
'Whole-genome'
or 'democratic
majority' tree
1
2
3
'Species tree,' or
bifurcation history

Okay, so what have we learnt
and where do we go from here?