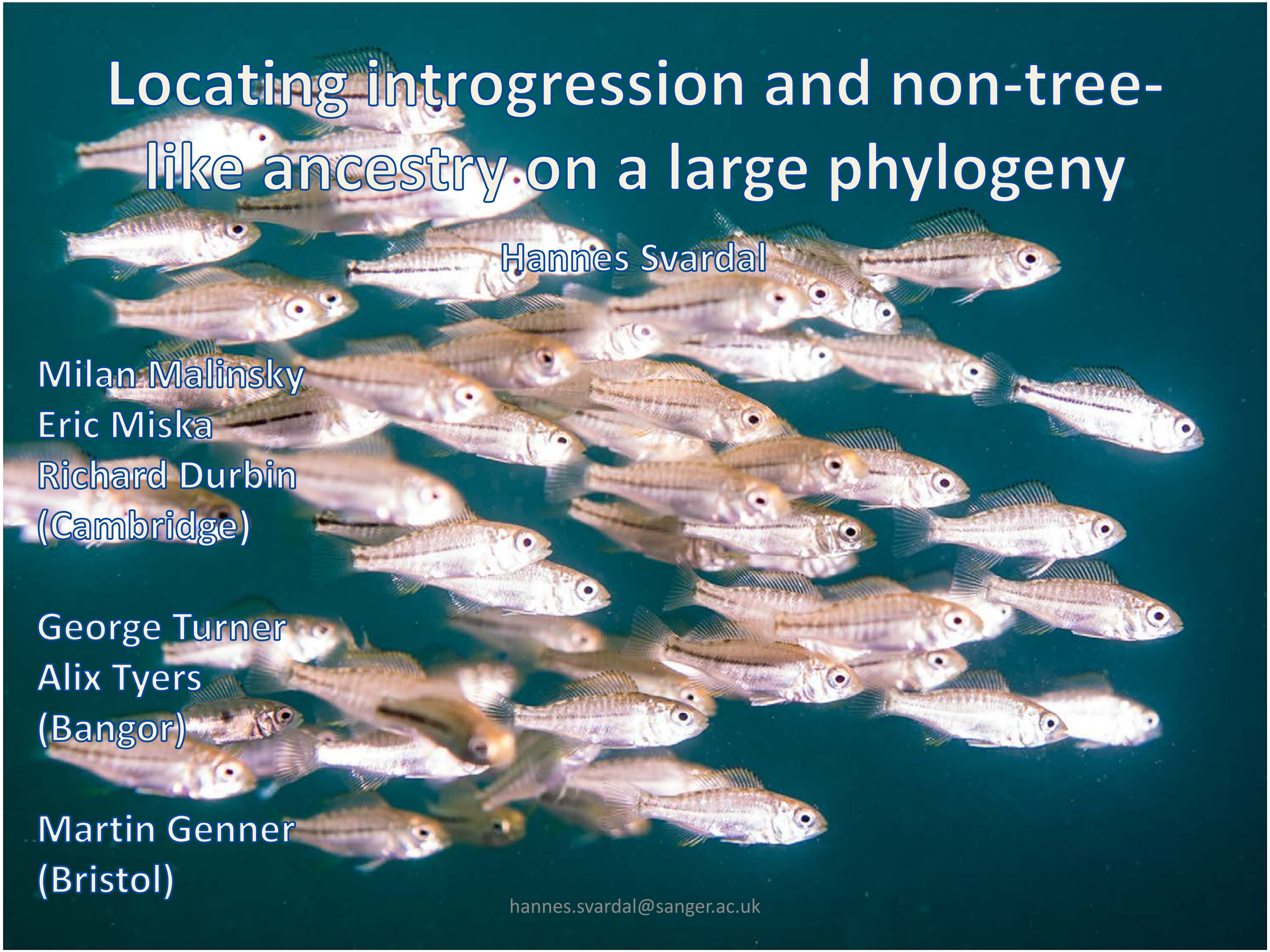


Locating introgression and non-tree-like ancestry on a large phylogeny

Hannes Svardal

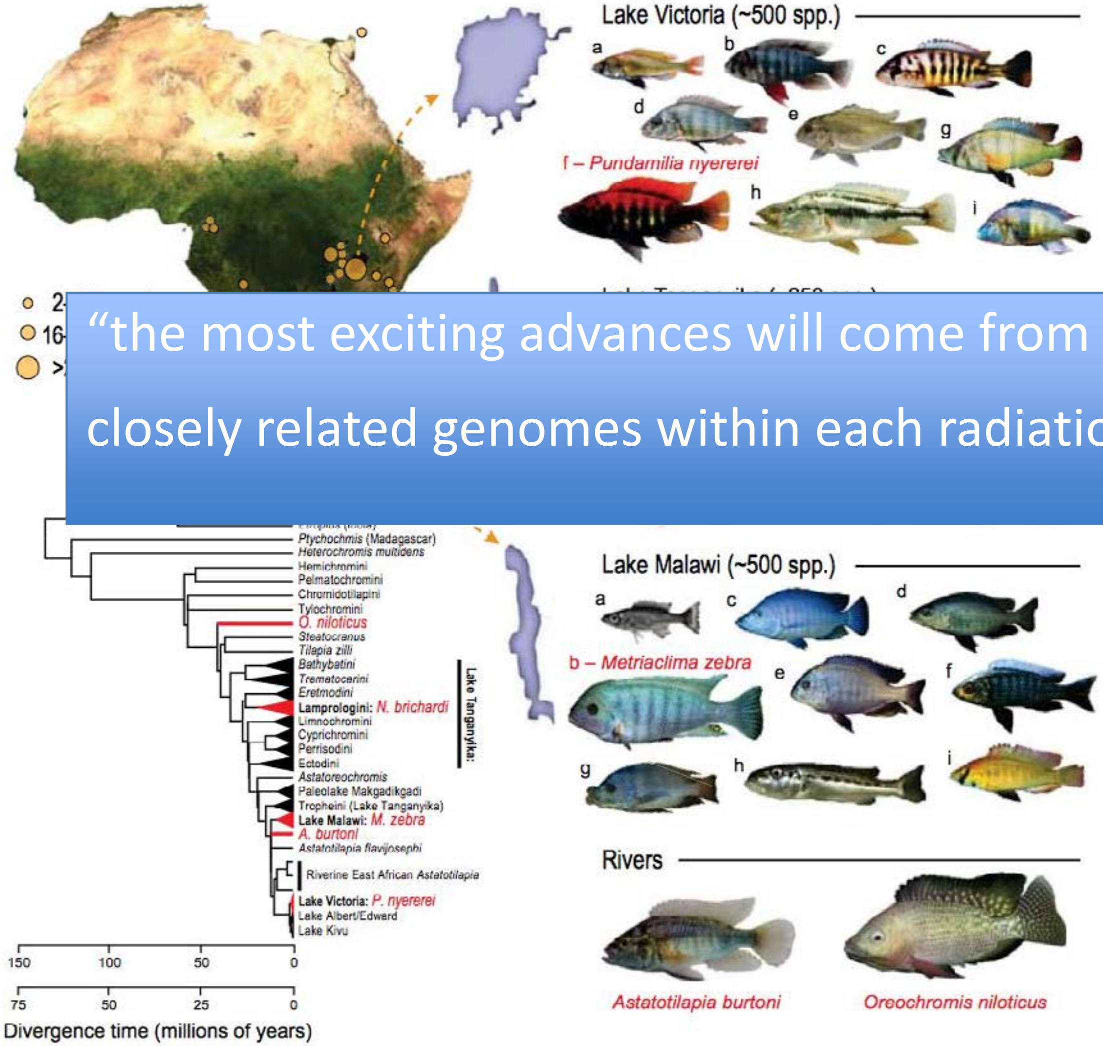
Milan Malinsky
Eric Miska
Richard Durbin
(Cambridge)

George Turner
Alix Tyers
(Bangor)

Martin Genner
(Bristol)

hannes.svardal@sanger.ac.uk

Cichlids like to radiate

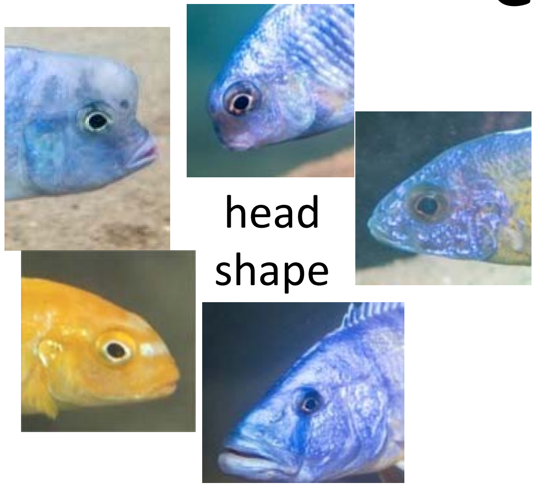


“the most exciting advances will come from analysis of more-closely related genomes within each radiation” (Jiggins 2014)

a – pelagic zooplanktivore;
b – rock-dwelling algae scraper;
c – paedophage;
d – scale eater;
e – snail crusher;
e;

Brawand et al. Nature 2014

Remarkable phenotypic adaptations allowed Lake Malawi cichlids to conquer ecological niches



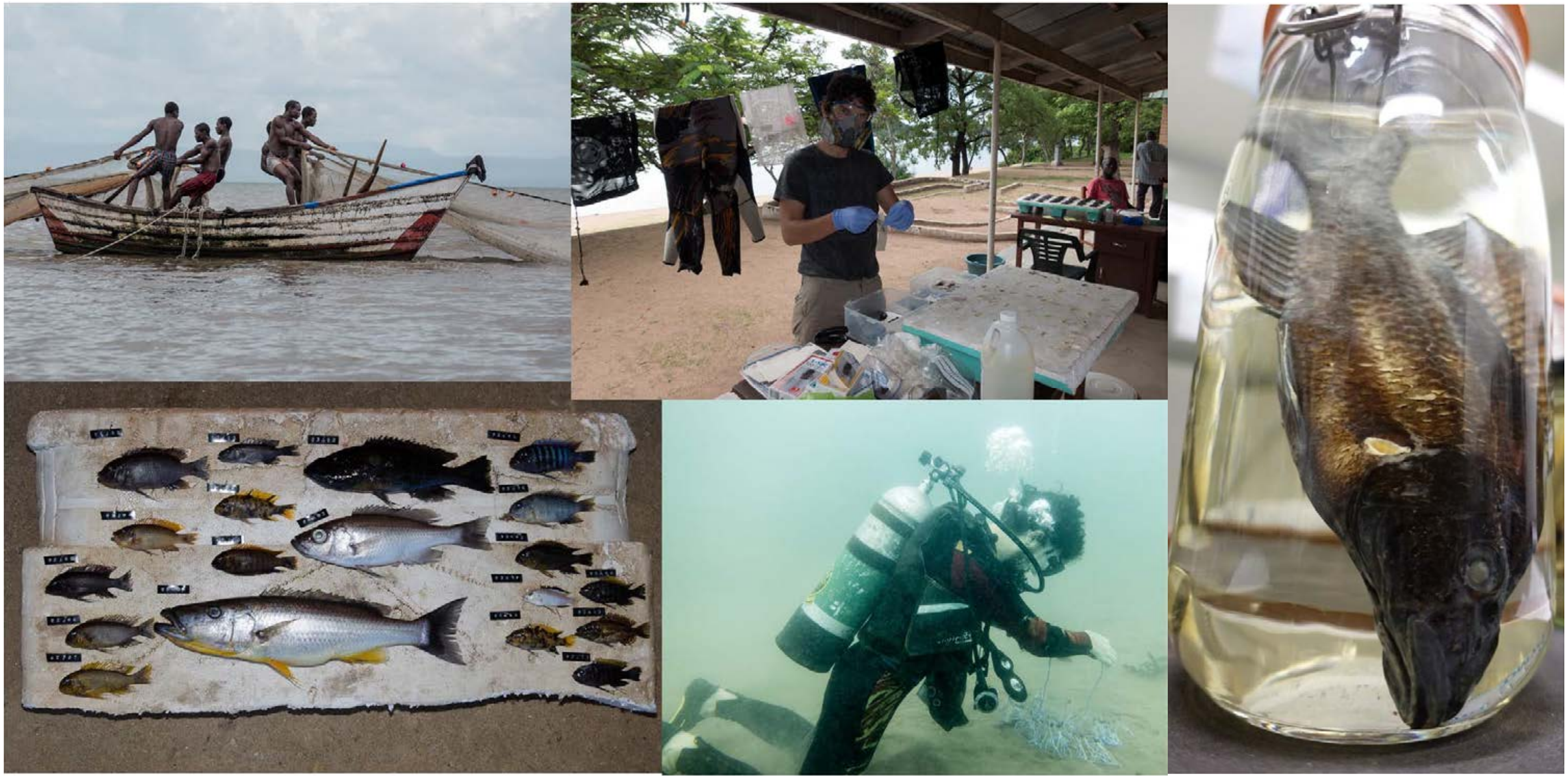
head shape

feeding apparatus

pigmentation

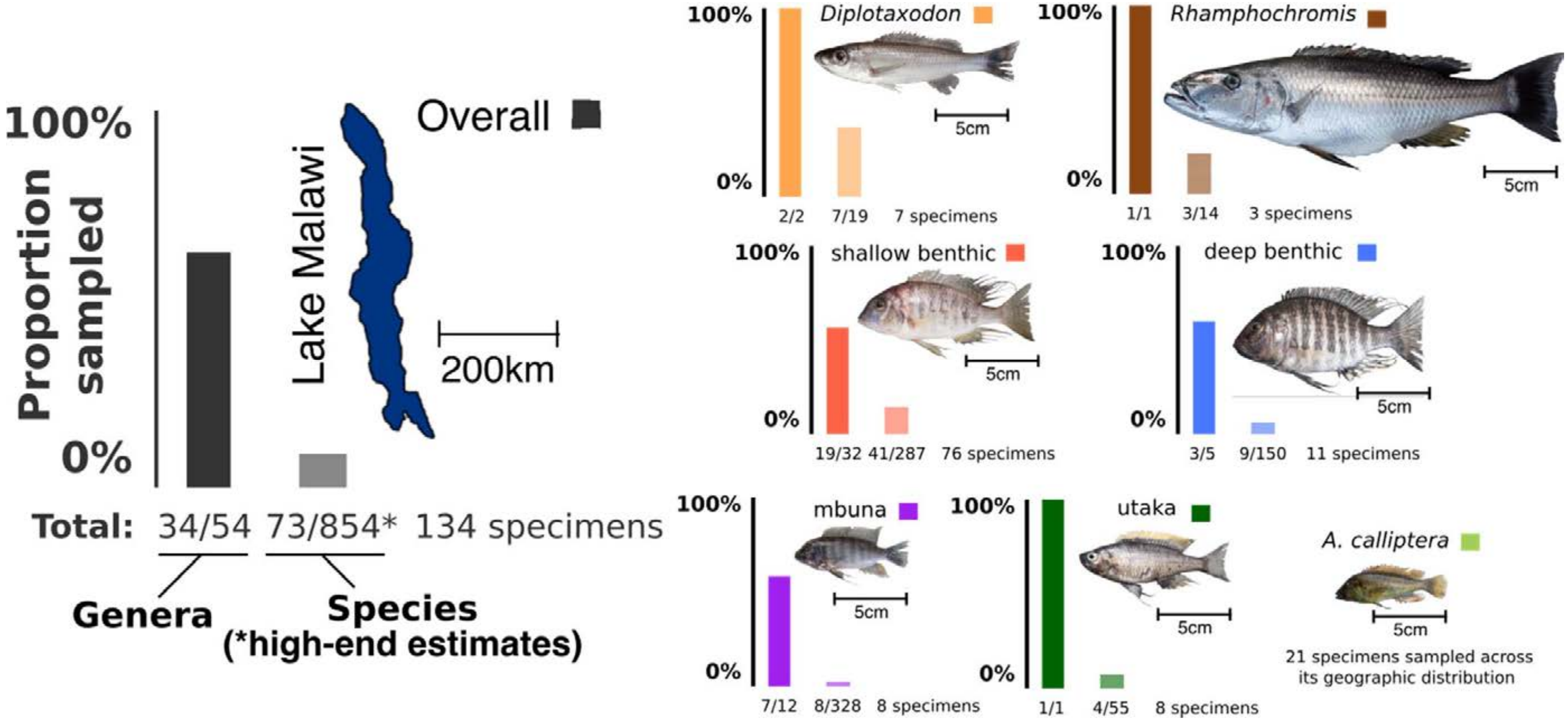
And size, body shape, behaviour…

Sampling: >2500 samples; >250 species; >1800 museum specimen



This talk: ~ 140 samples from 70 species
whole genome sequencing at 15X/5X

Samples from all major morpho-ecological groups



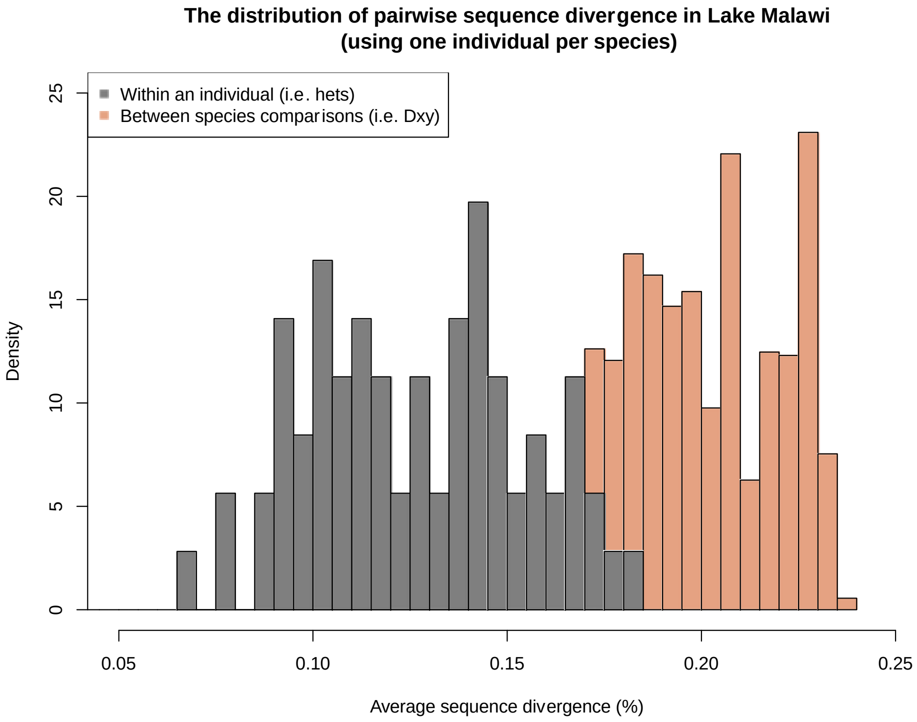
hannes.svardal@sanger.ac.uk

Questions

- Genetic relatedness
- Do species hybridise? Can we quantify gene flow?
- Which genes were under selection?
- Is there any evidence for adaptive introgression?

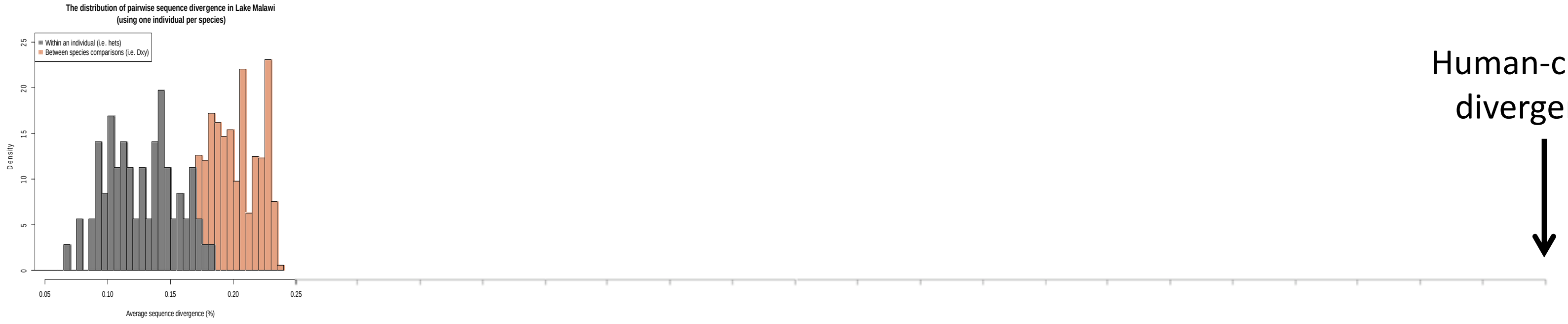
Genetic relatedness

- Lots of phenotypic diversity and hundreds of species with little genetic divergence
  - max. 0.25% sequence divergence, five times less than human-chimp (1.23%)



Genetic relatedness

- Lots of phenotypic diversity and hundreds of species with little genetic divergence
  - max. 0.25% sequence divergence, five times less than human-chimp (1.23%)

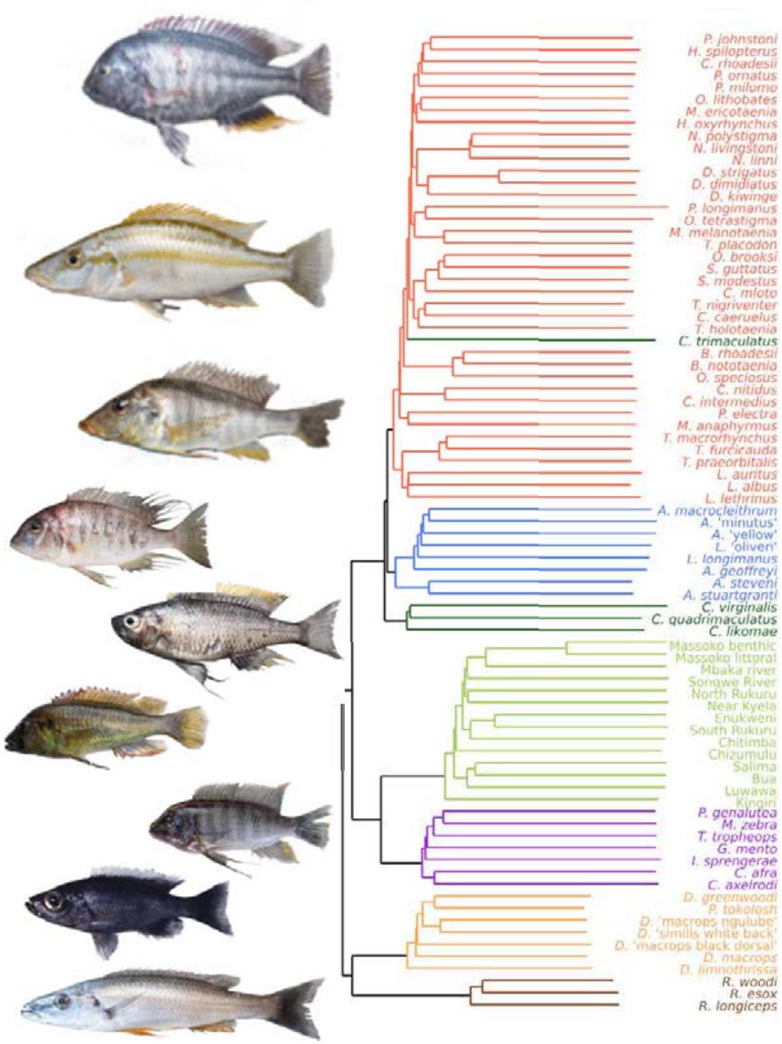


Human-chimp divergence

http://www.spacedaily.com/images-lg/human-child-chimpanzee-baby-lg.jpg

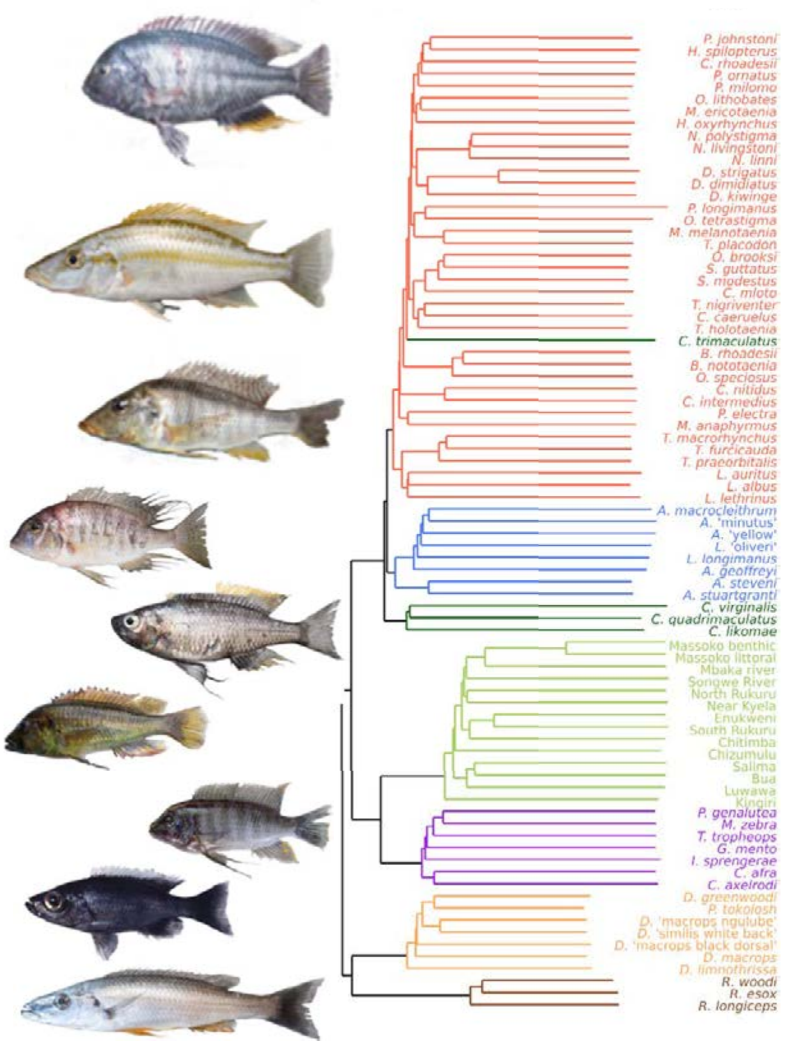
More or less well-defined major clades...

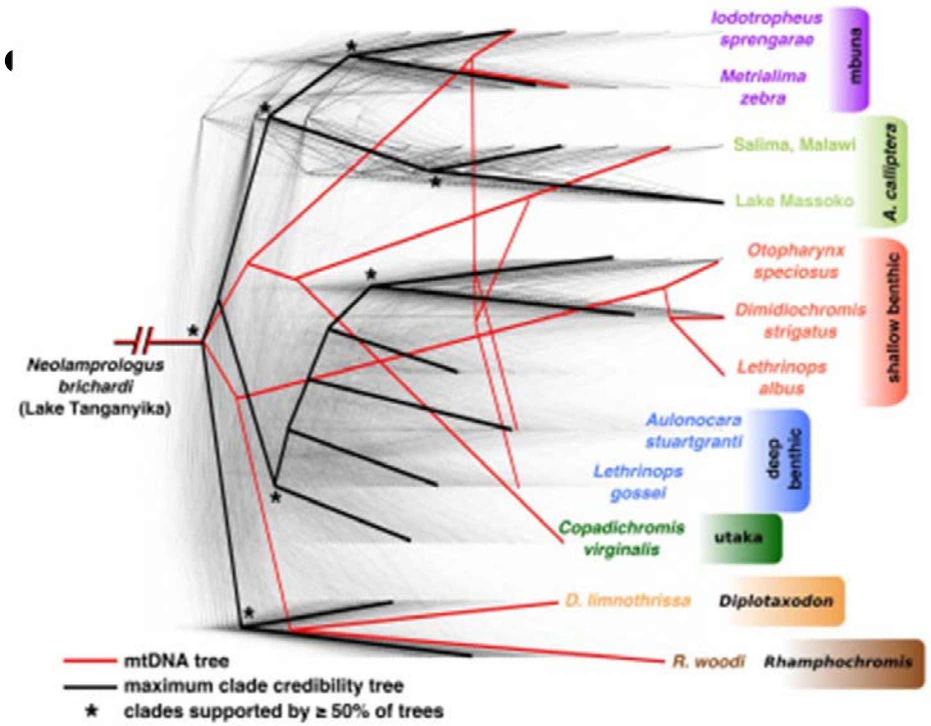
NJ tree



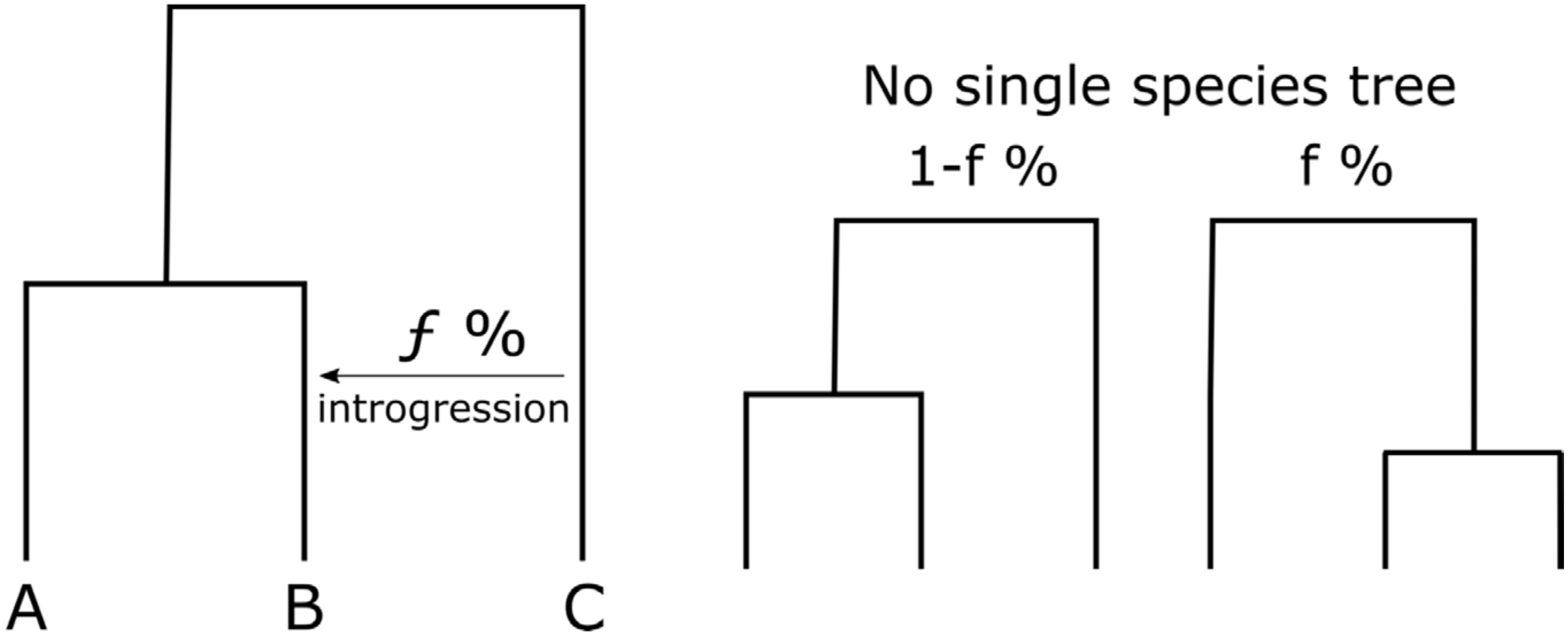
Interesting observations

- Tree topology sensitive to method used, samples and loci included

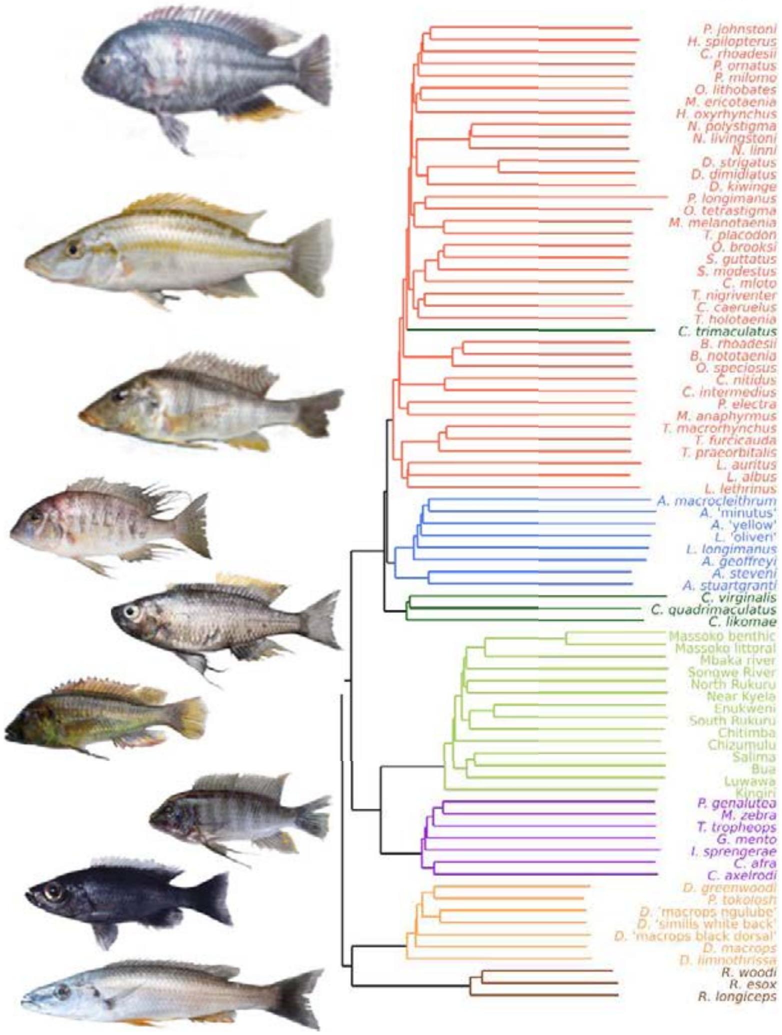




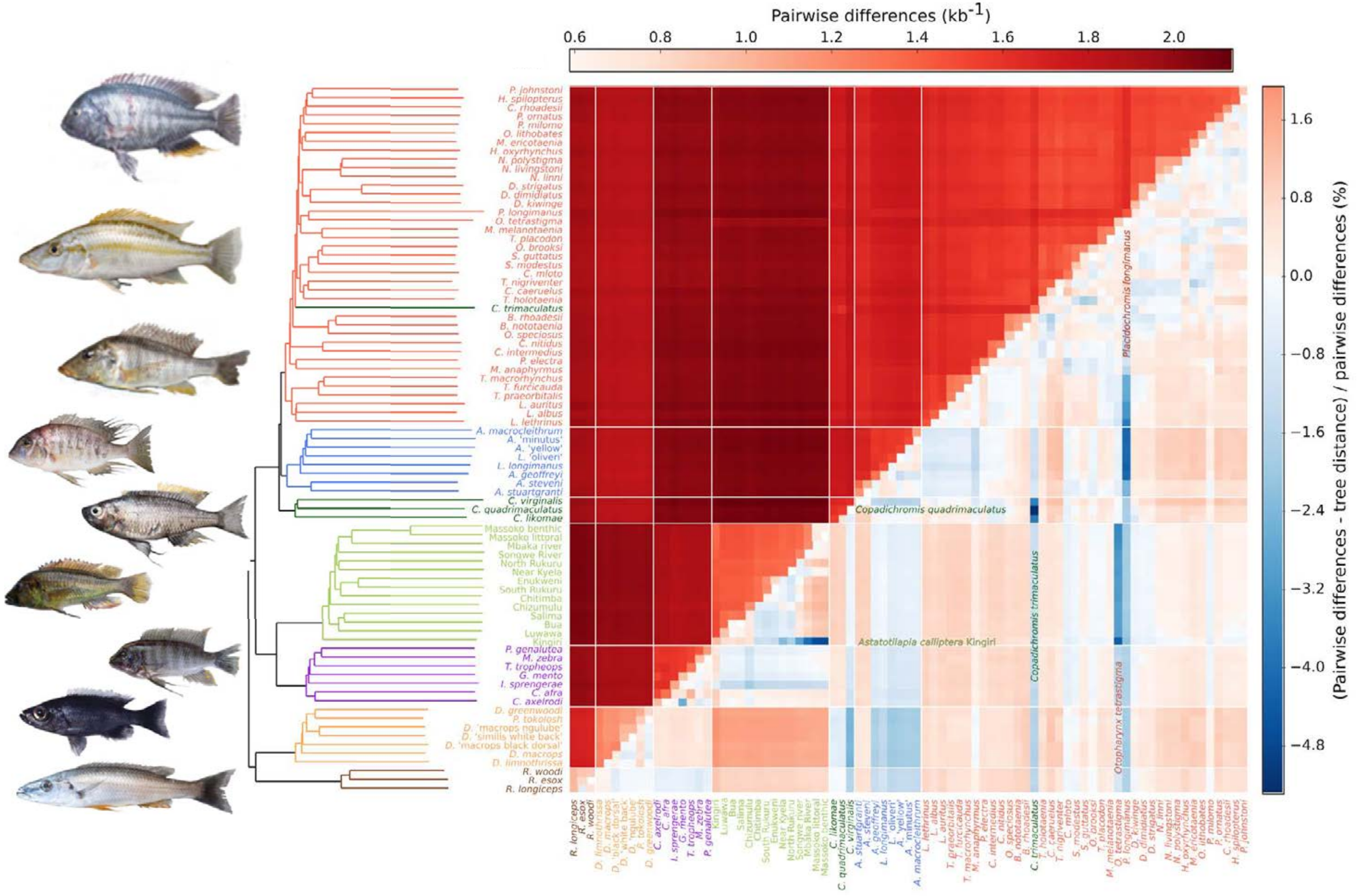
Under introgression, a phylogeny cannot capture the data.



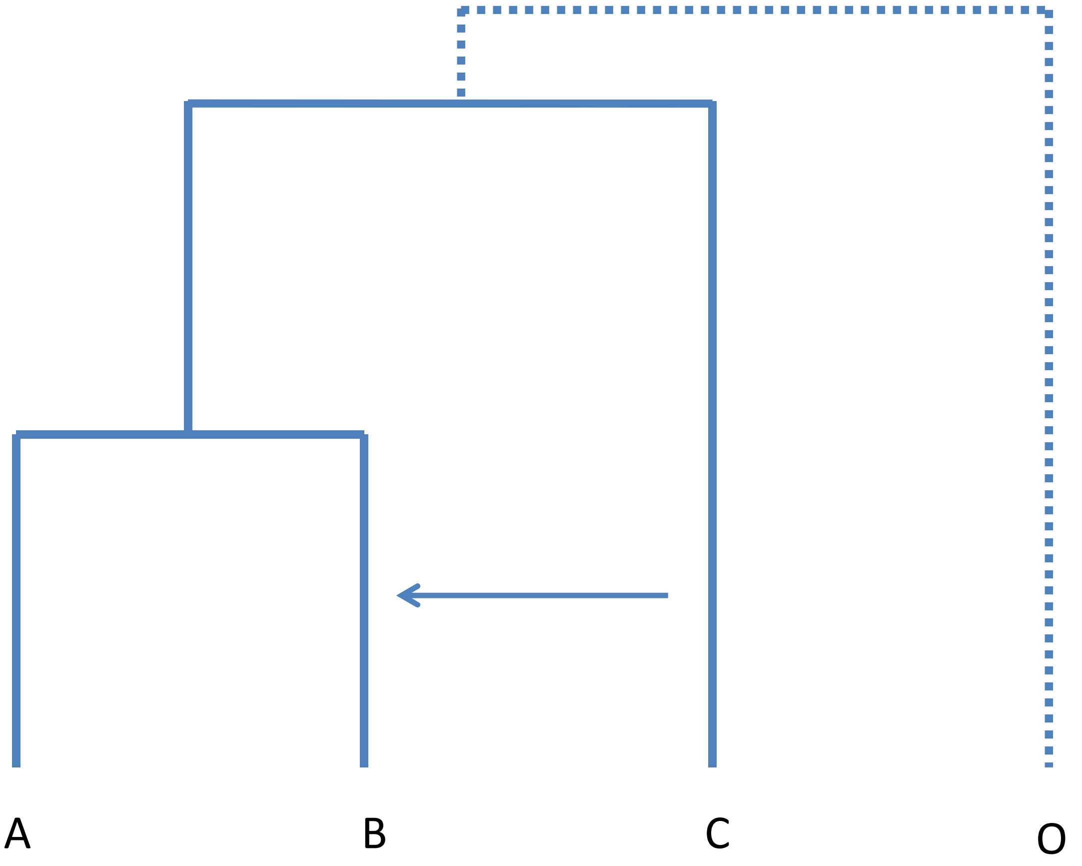
More or less well-defined major clades...



Relatedness not tree-like



Measuring gene-flow: D/F4-statistic

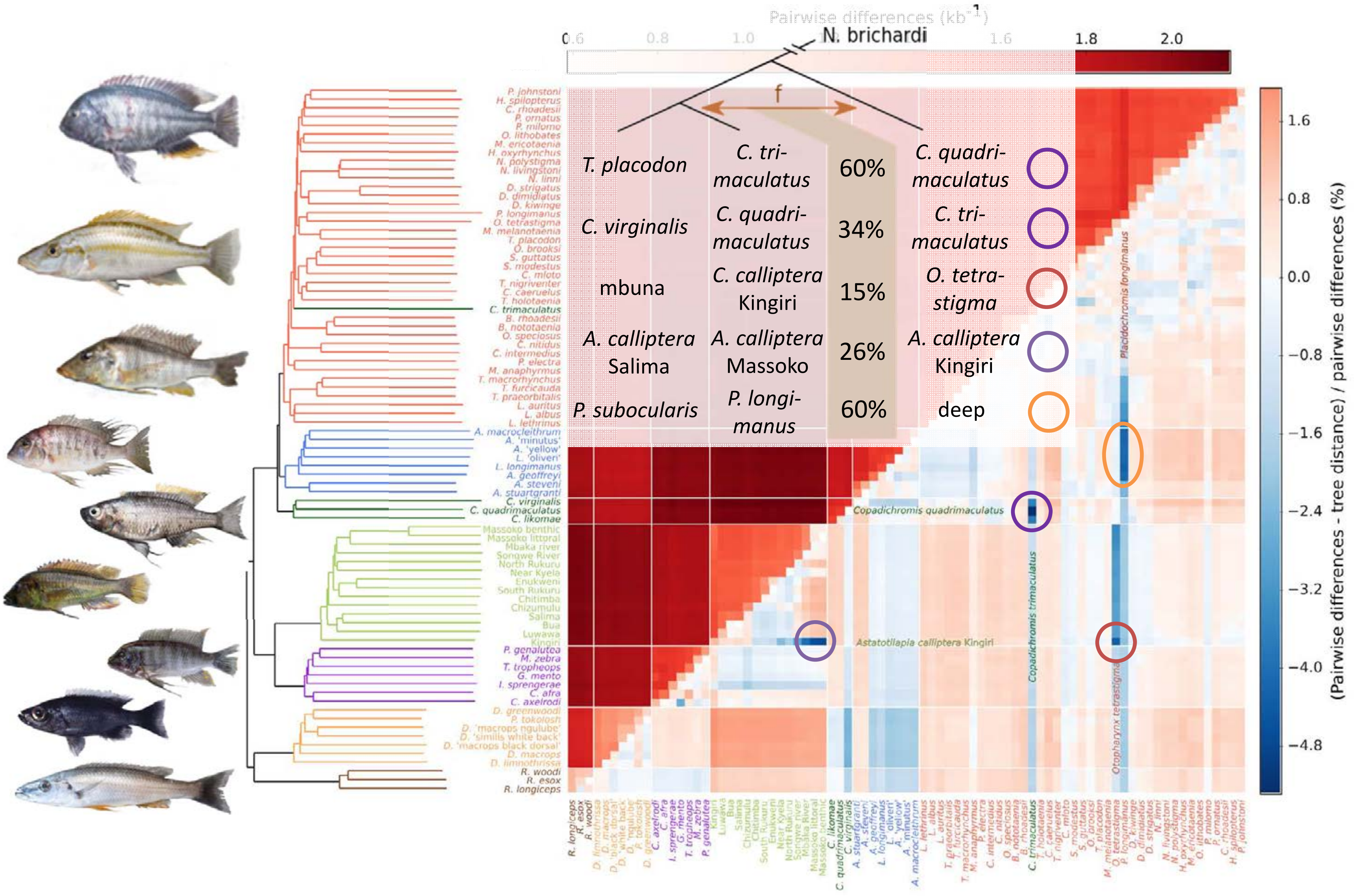


- A and B should be equally closely related to C
- However, if there was gene flow between B and C then B should be (a bit) more closely related to C than A is to C

$$S(A,B,C,O) = \sum_{loci} (\hat{p}_A - \hat{p}_B)(\hat{p}_C - \hat{p}_O)$$

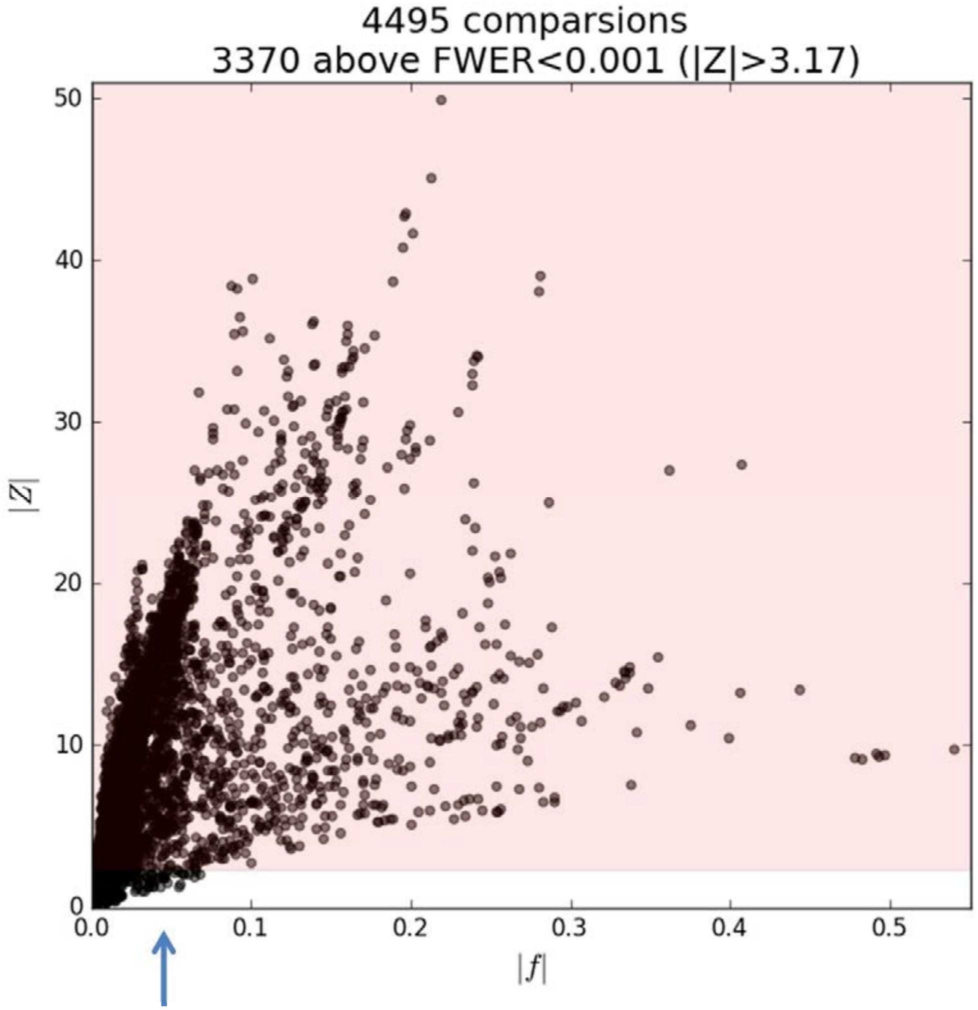
$$f = \frac{S(A,B,C,O)}{S(A,C_1,C_2,O)}$$

Relatedness not tree-like



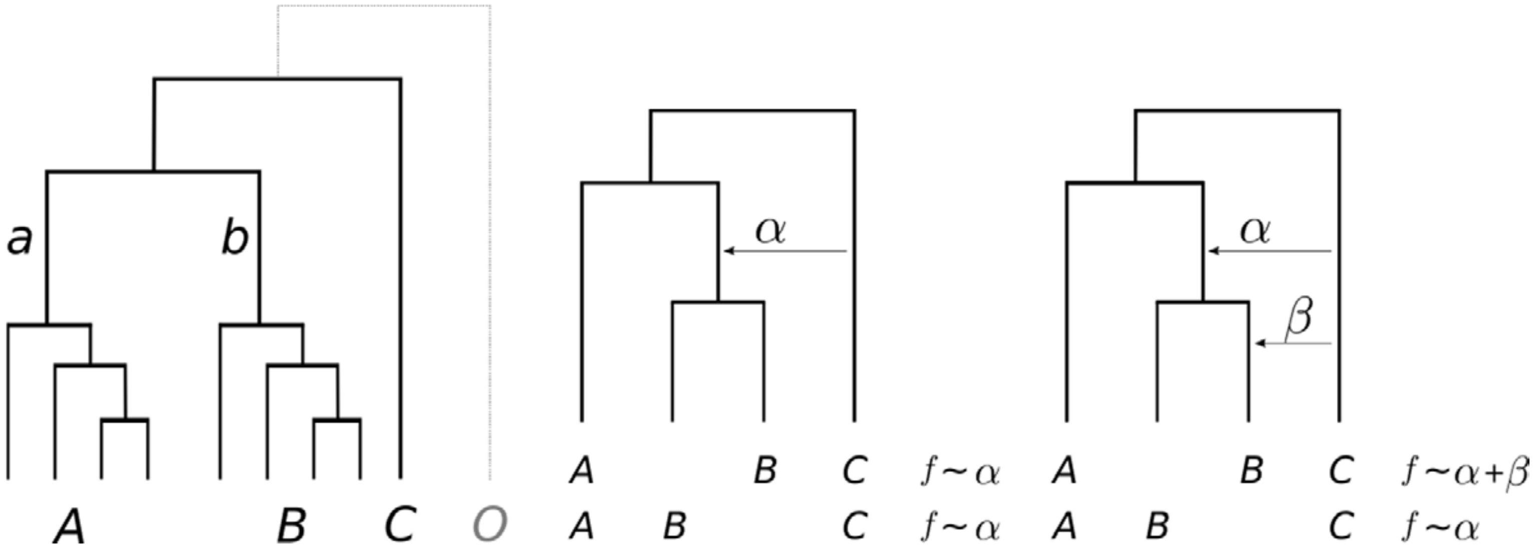
T. placodon	*C. trimaculatus*	60%	*C. quadrimaculatus*
C. virginalis	*C. quadrimaculatus*	34%	*C. trimaculatus*
mbuna	*C. calliptera* Kingiri	15%	*O. tetrastigma*
A. calliptera Salima	*A. calliptera* Massoko	26%	*A. calliptera* Kingiri
P. subocularis	*P. longimanus*	60%	deep

Many ABBA BABA tests are highly significant



human -- Neanderthal

ABBA-BABA tests not independent

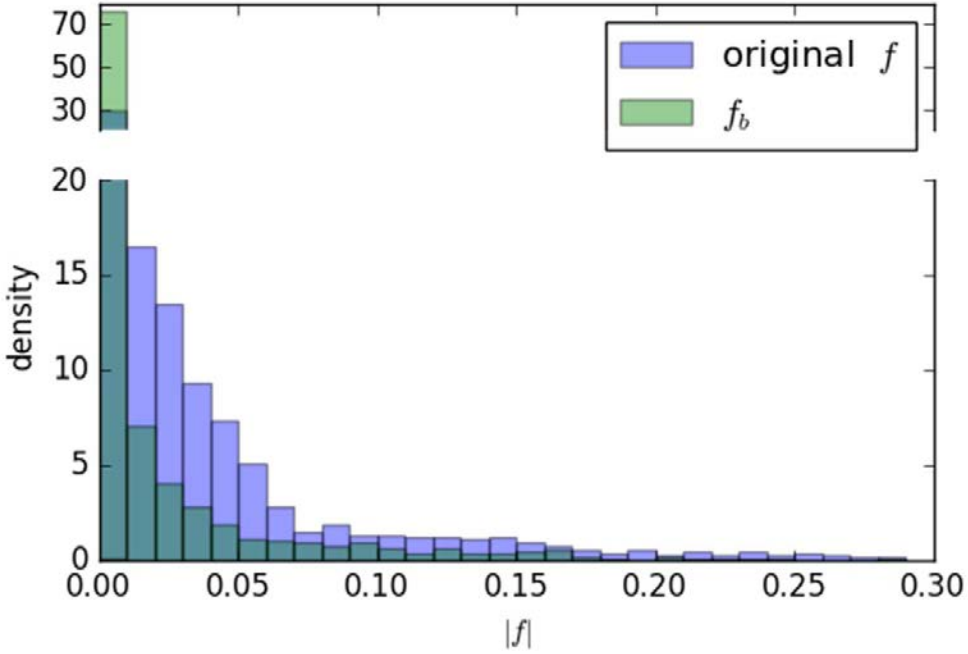


Branch specific f-score f_b:

$$f_b(C) = median_A[min_B[f(A,B,C)]]$$

hannes.svardal@sanger.ac.uk

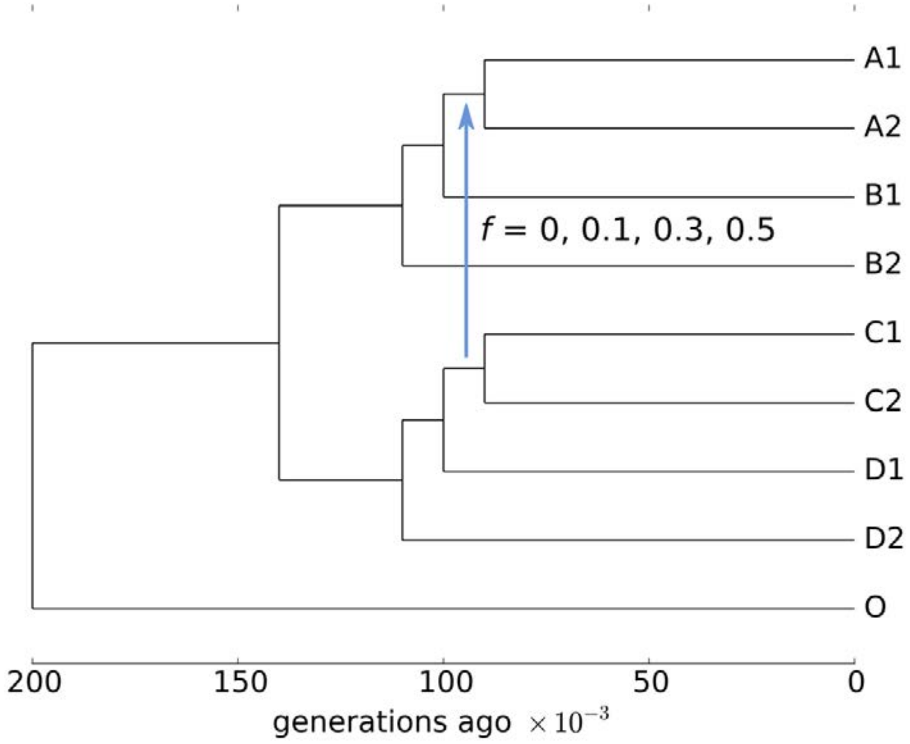
Branch specific f-score f_b



- Reduces the amount of tests
- Removes some of the correlation
- Allows to infer non-tree like relationships at internal and terminal branches.

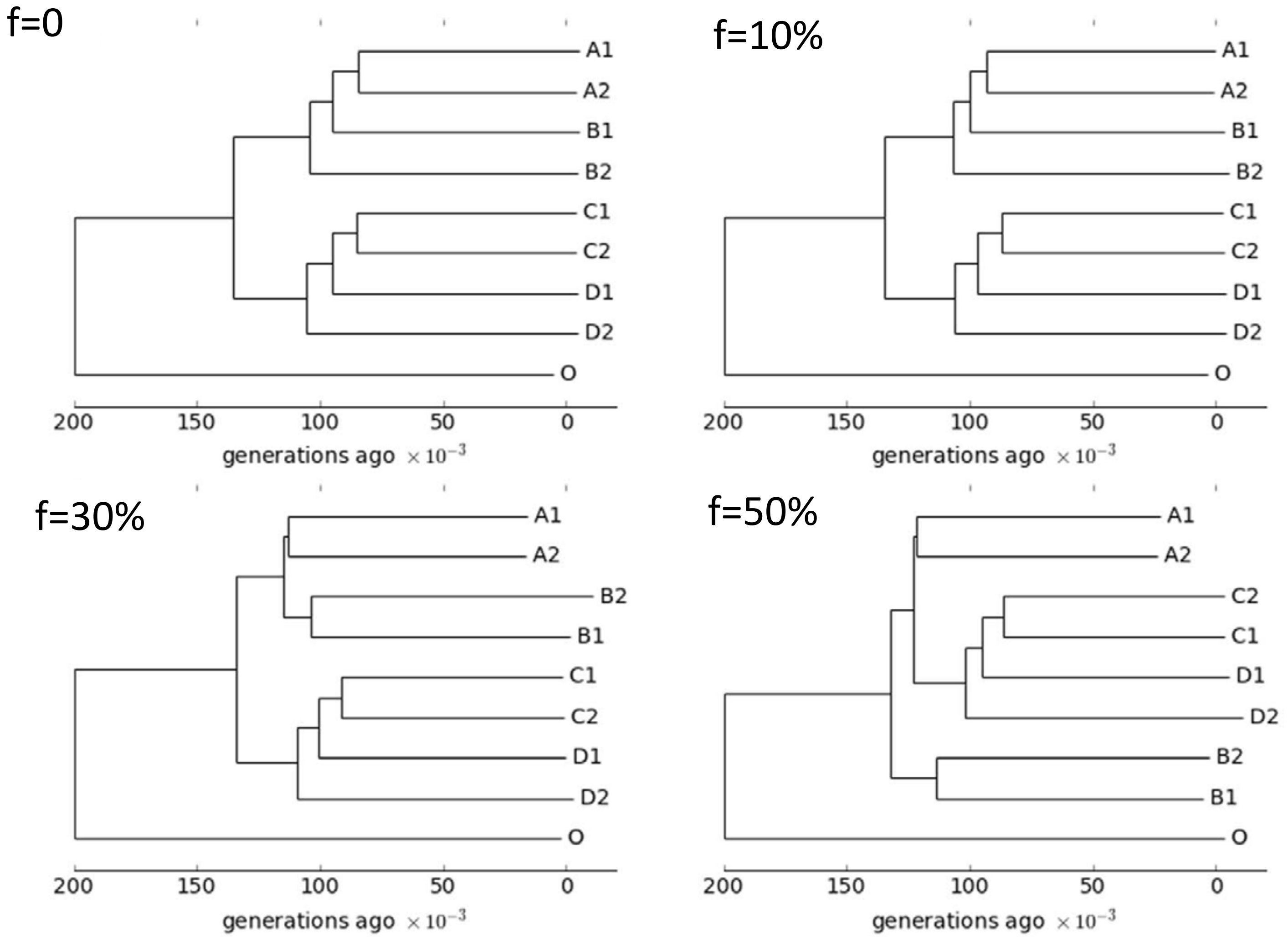
hannes.svardal@sanger.ac.uk

Test f_b with Simulations



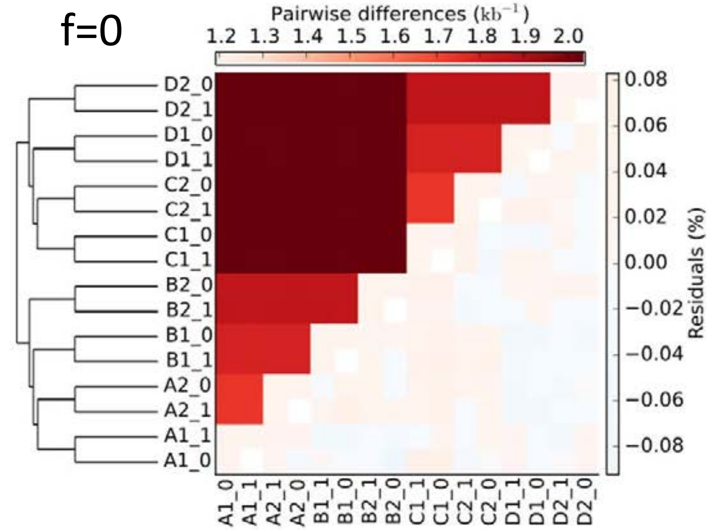
A1, A2, …, D2 ... 2 diploid individuals
outgroup O ... single diploid individual
Effective population size constant at 10^5.
Recombination rate $2*10^{-8}$, mutation rate $3*10^{-9}$.
120 independent stretches of $5*10^6$ bp (600 10^9 bp in total)

Inferred split trees

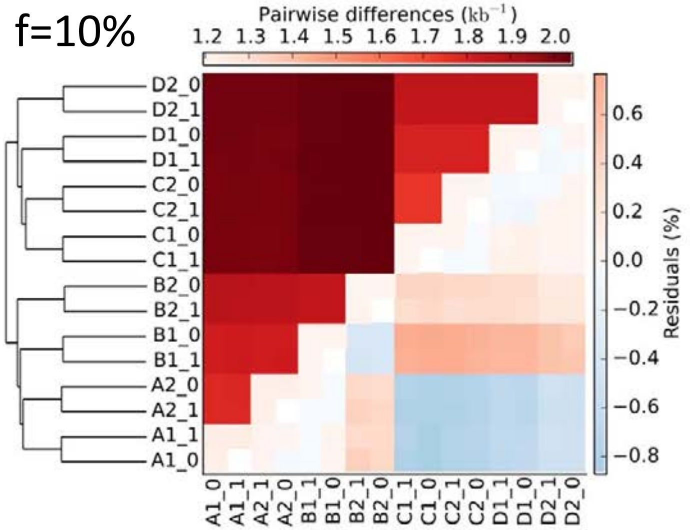


Pairwise differences and residuals

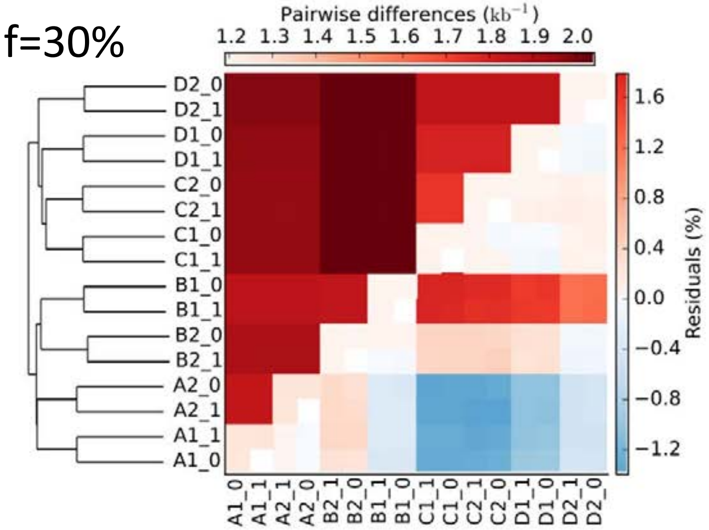
f=0



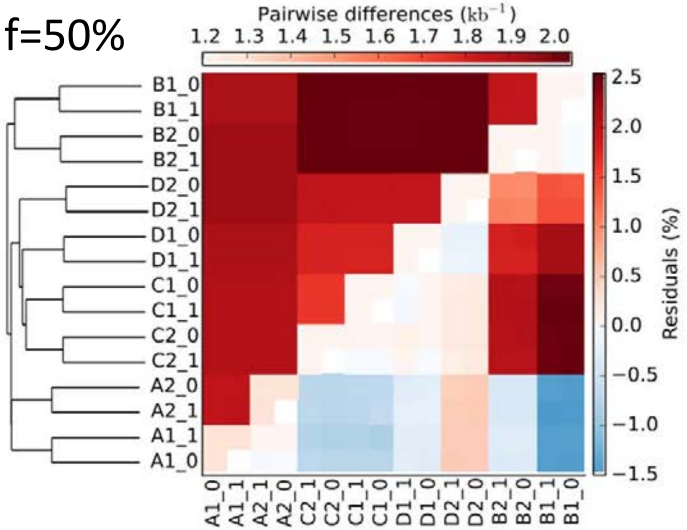
f=10%



f=30%

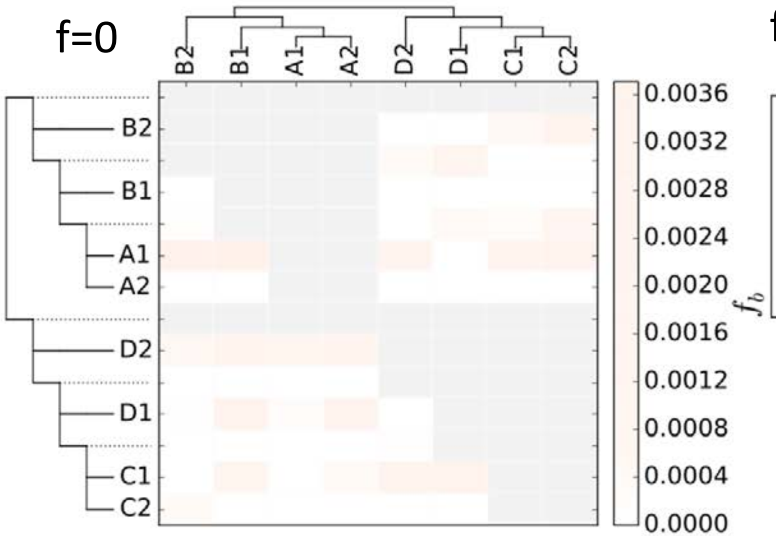


f=50%

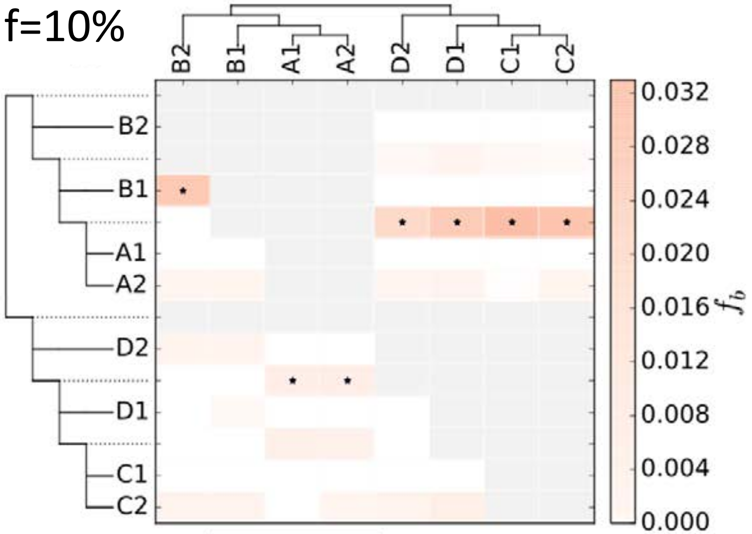


f-branch

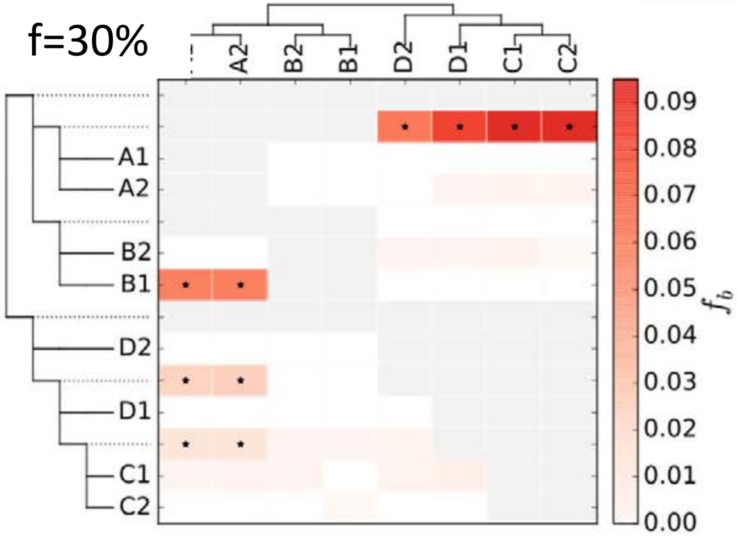
f=0



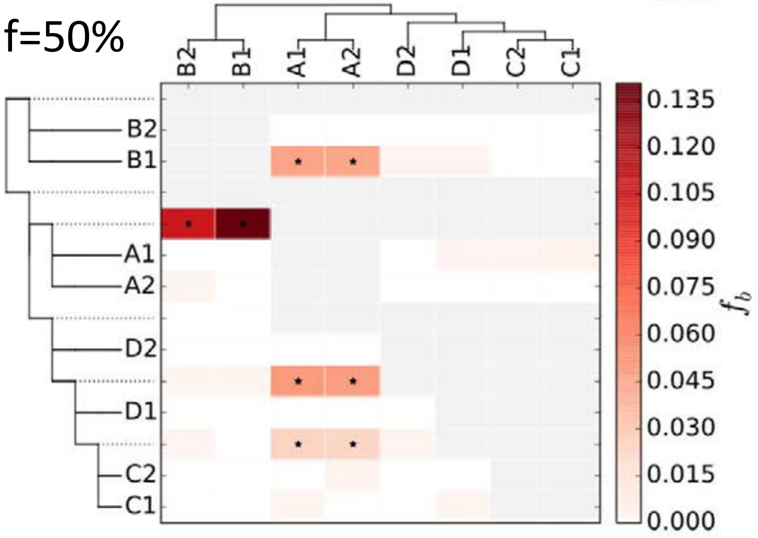
f=10%



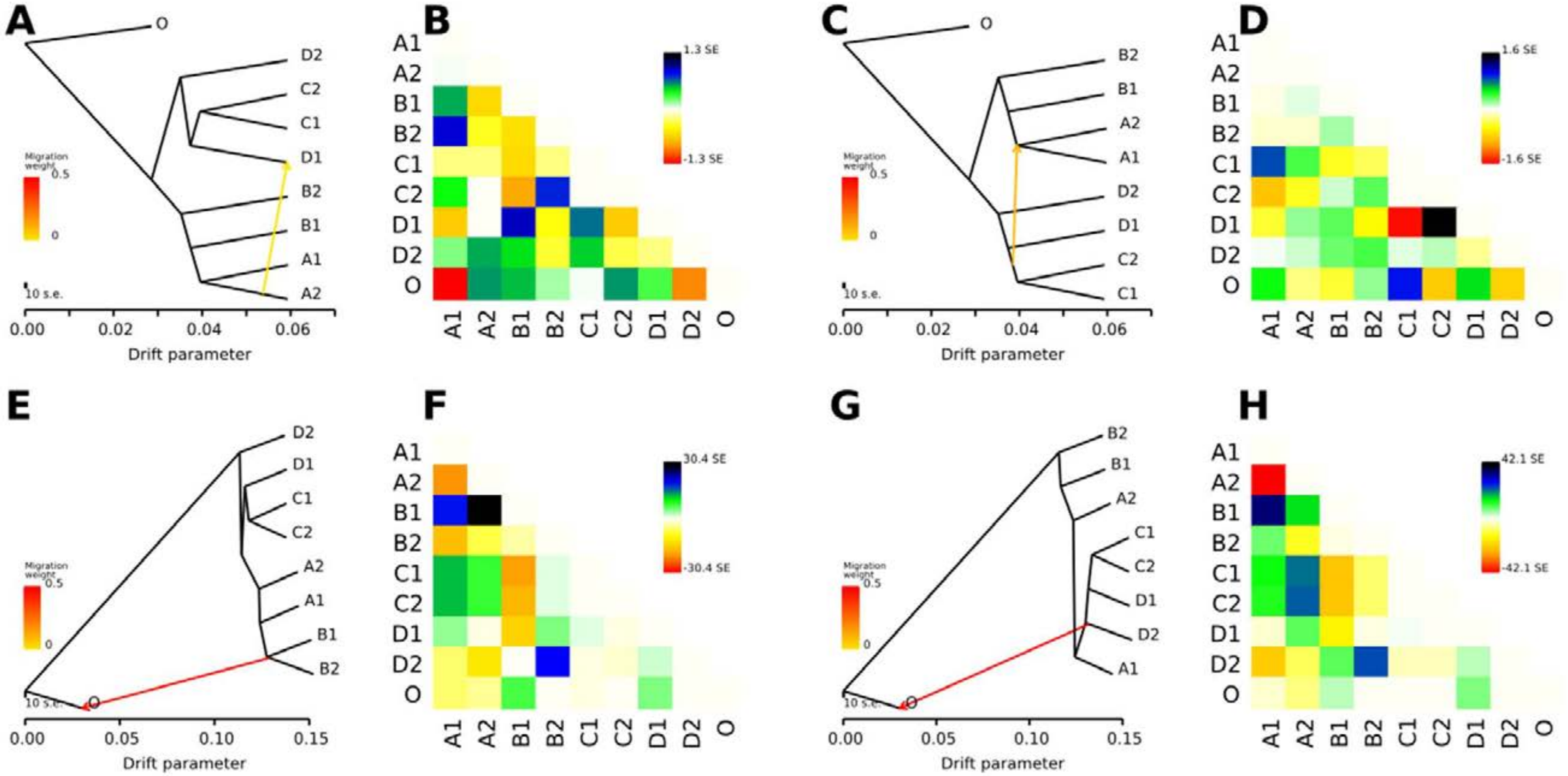
f=30%



f=50%

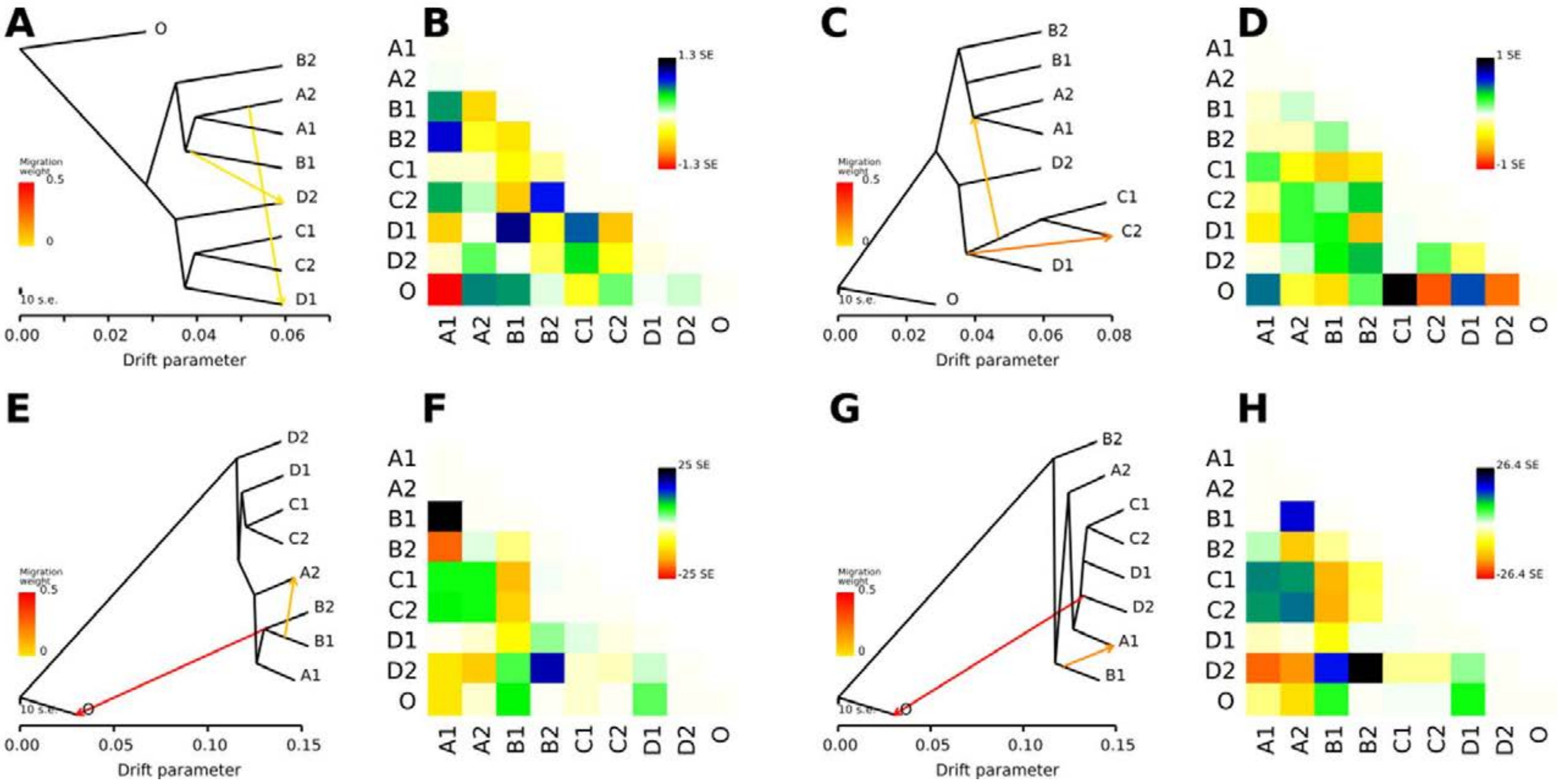


Comparison to treemix: m=1

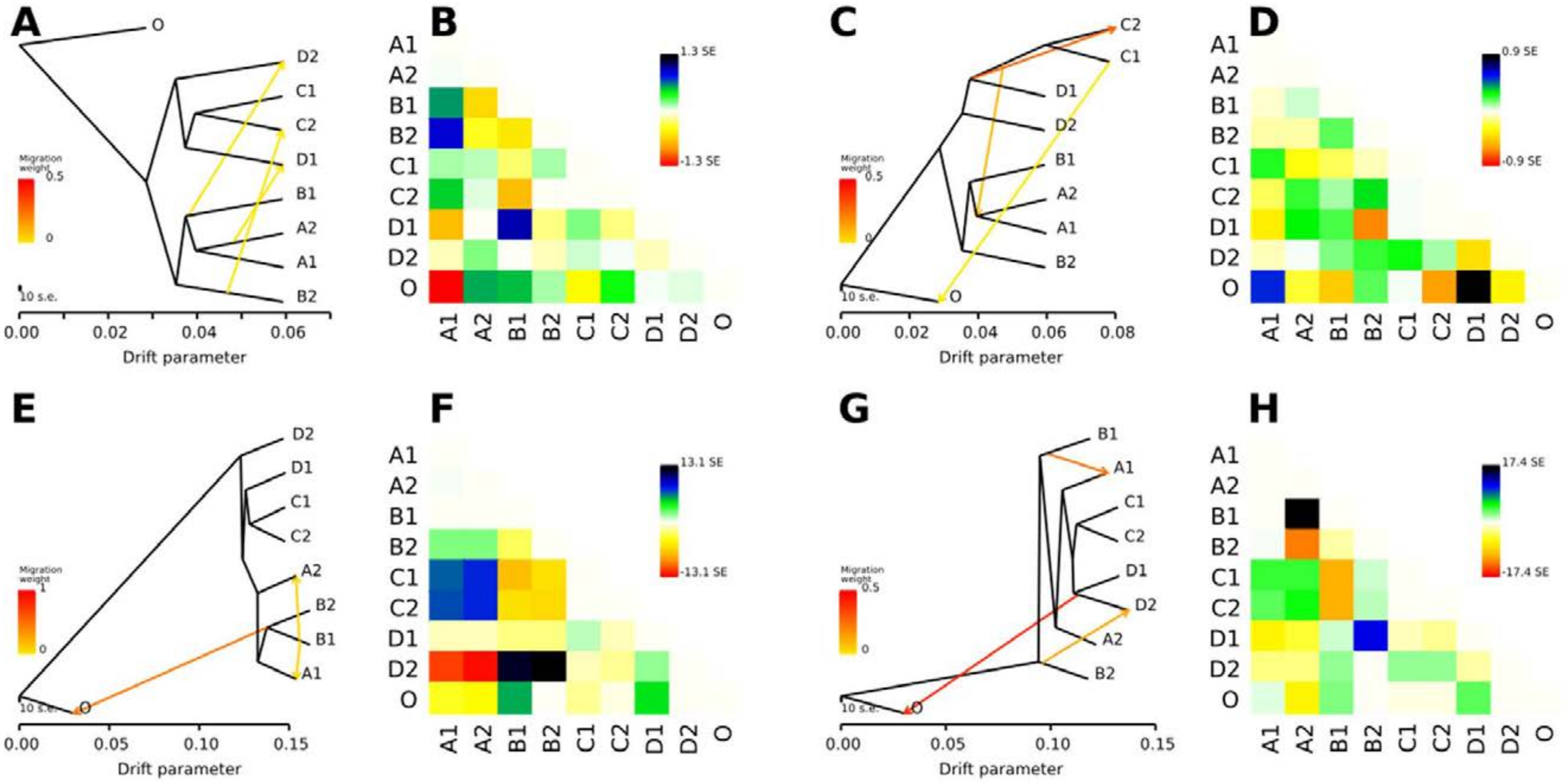


Pickrell and Pritchard 2012, PLOS Genetics

treemix m=2


A
O
B2
A2
A1
B1
D2
C1
C2
D1
Migration weight
0.5
0
10 s.e.
0.00
0.02
0.04
0.06
Drift parameter
B
A1
A2
B1
B2
C1
C2
D1
D2
O
1.3 SE
-1.3 SE
C
B2
B1
A2
A1
D2
C1
C2
D1
O
Migration weight
0.5
0
10 s.e.
0.00
0.02
0.04
0.06
0.08
Drift parameter
D
1 SE
-1 SE
E
D2
D1
C1
C2
A2
B2
B1
A1
O
Migration weight
0.5
0
10 s.e.
0.00
0.05
0.10
0.15
Drift parameter
F
25 SE
-25 SE
G
B2
A2
C1
C2
D1
D2
A1
B1
O
Migration weight
0.5
0
10 s.e.
0.00
0.05
0.10
0.15
Drift parameter
H
26.4 SE
-26.4 SE

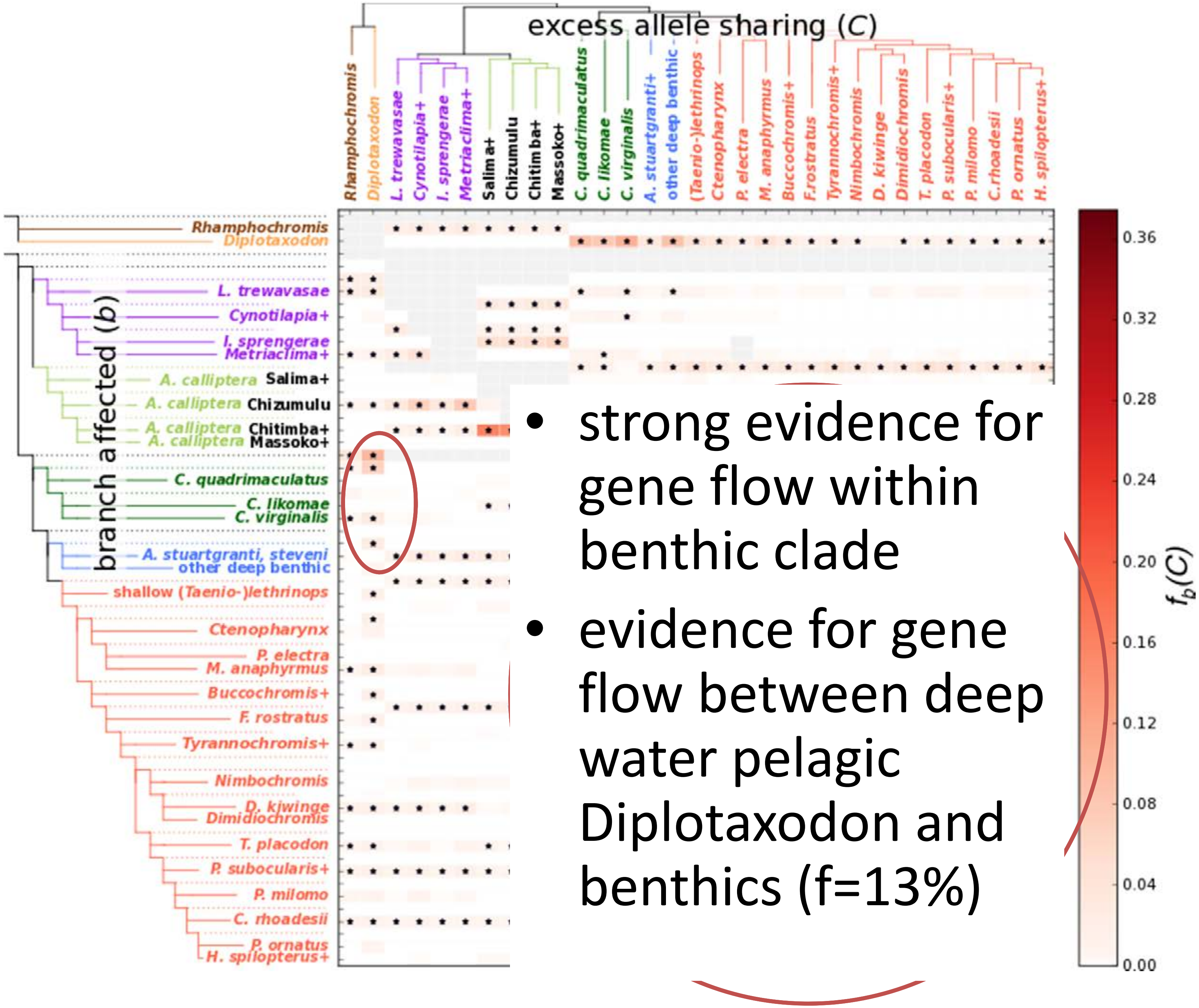
treemix m=3


A
B
C
D
E
F
G
H
O
D2
C1
C2
D1
B1
A2
A1
B2
Migration weight
0.5
0
1
10 s.e.
Drift parameter
0.00
0.02
0.04
0.06
0.08
0.05
0.10
0.15
1.3 SE
-1.3 SE
0.9 SE
-0.9 SE
13.1 SE
-13.1 SE
17.4 SE
-17.4 SE

Conclusion simulations

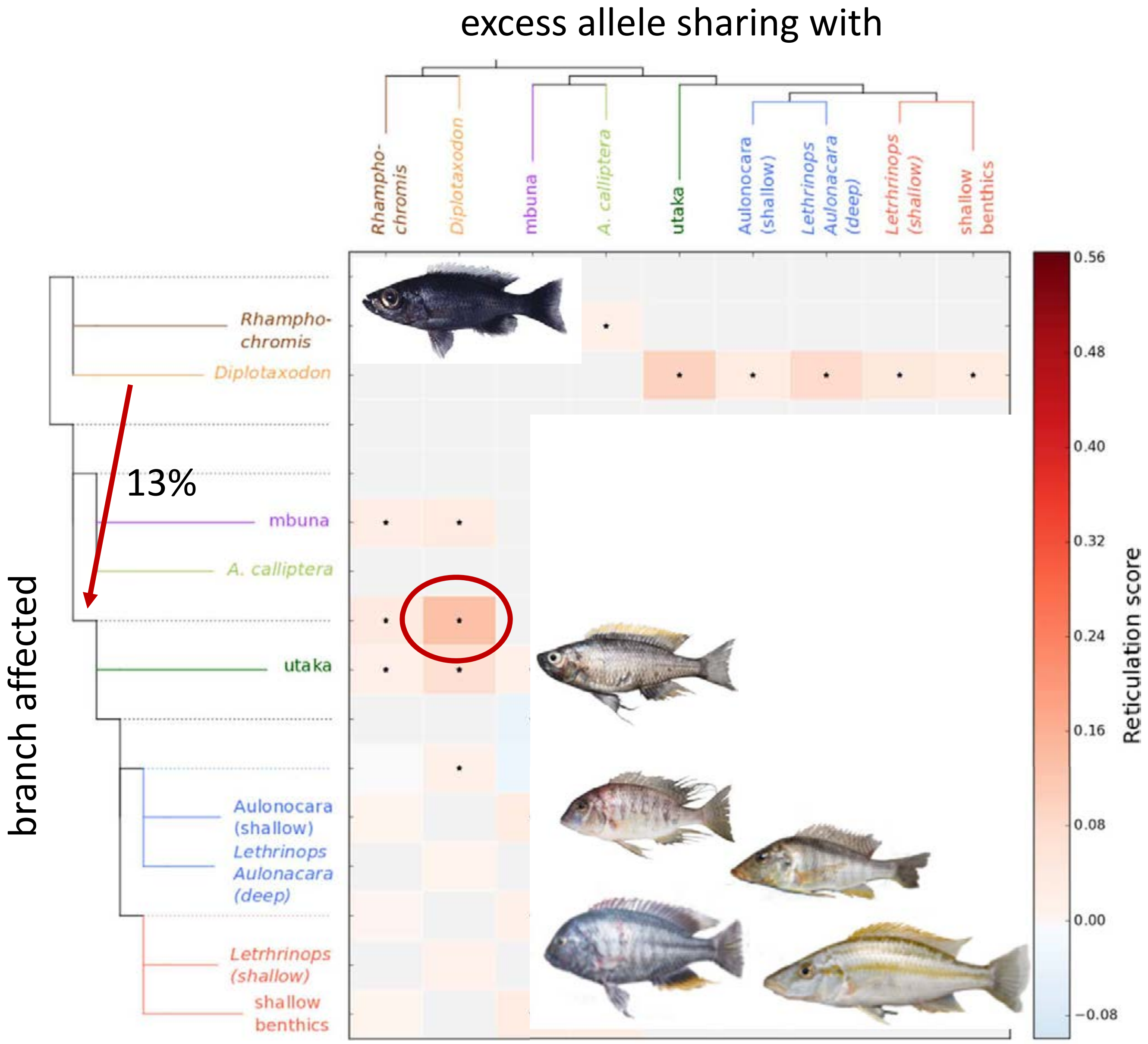
- A single gene-flow event generally leads to multiple significant f_b
- f_b can identify problematic branches that are not consistent with tree like ancestry
- f_b scores have a biologically sensible interpretation even if the inferred tree is wrong
- treemix is sensitive to correct inference of the tree --> results not sensible in scenarios with strong gene flow

hannes.svardal@sanger.ac.uk

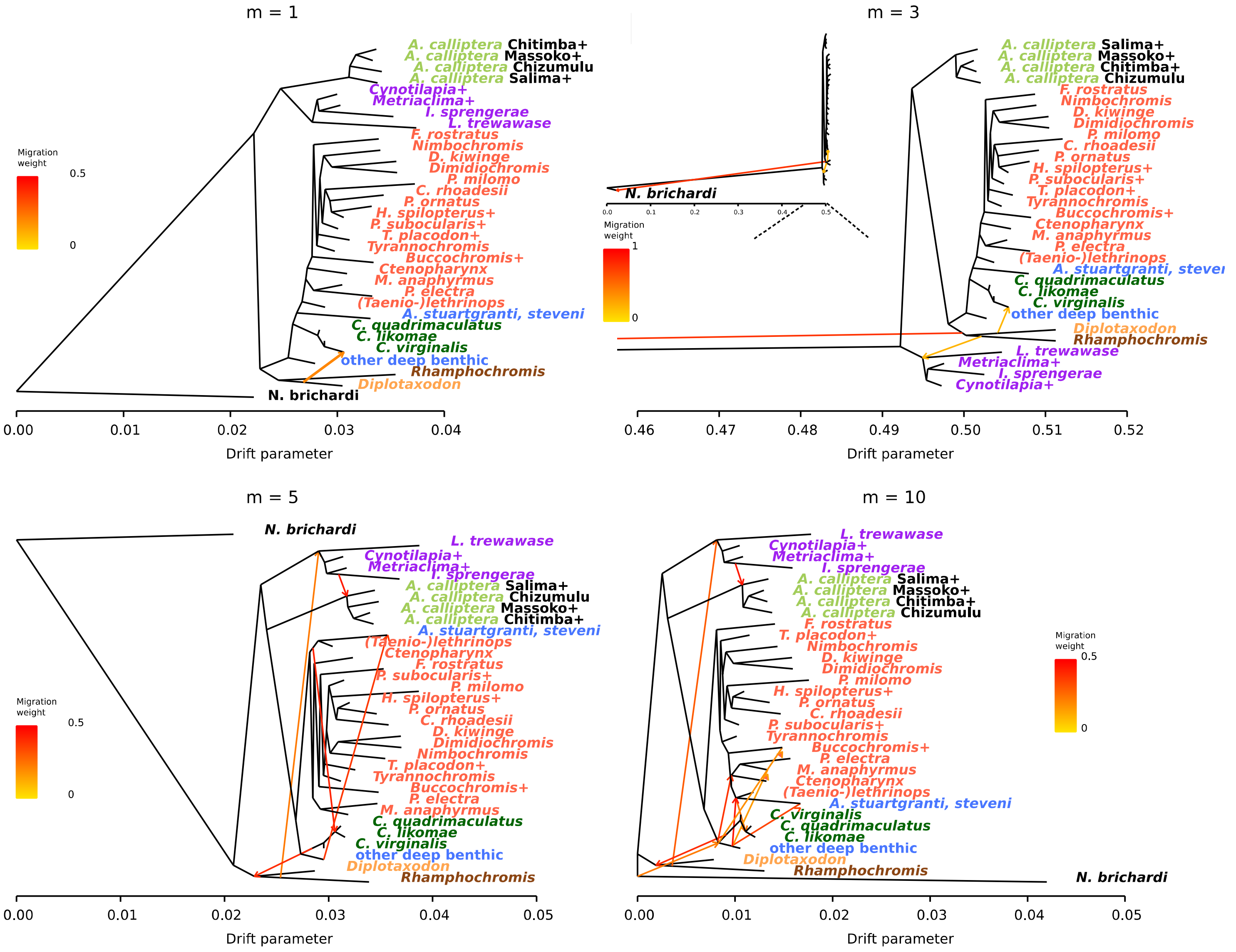


- strong evidence for gene flow within benthic clade
- evidence for gene flow between deep water pelagic Diplotaxodon and benthics (f=13%)

Diplotaxodon—benthics introgression



treemix on cichlid data

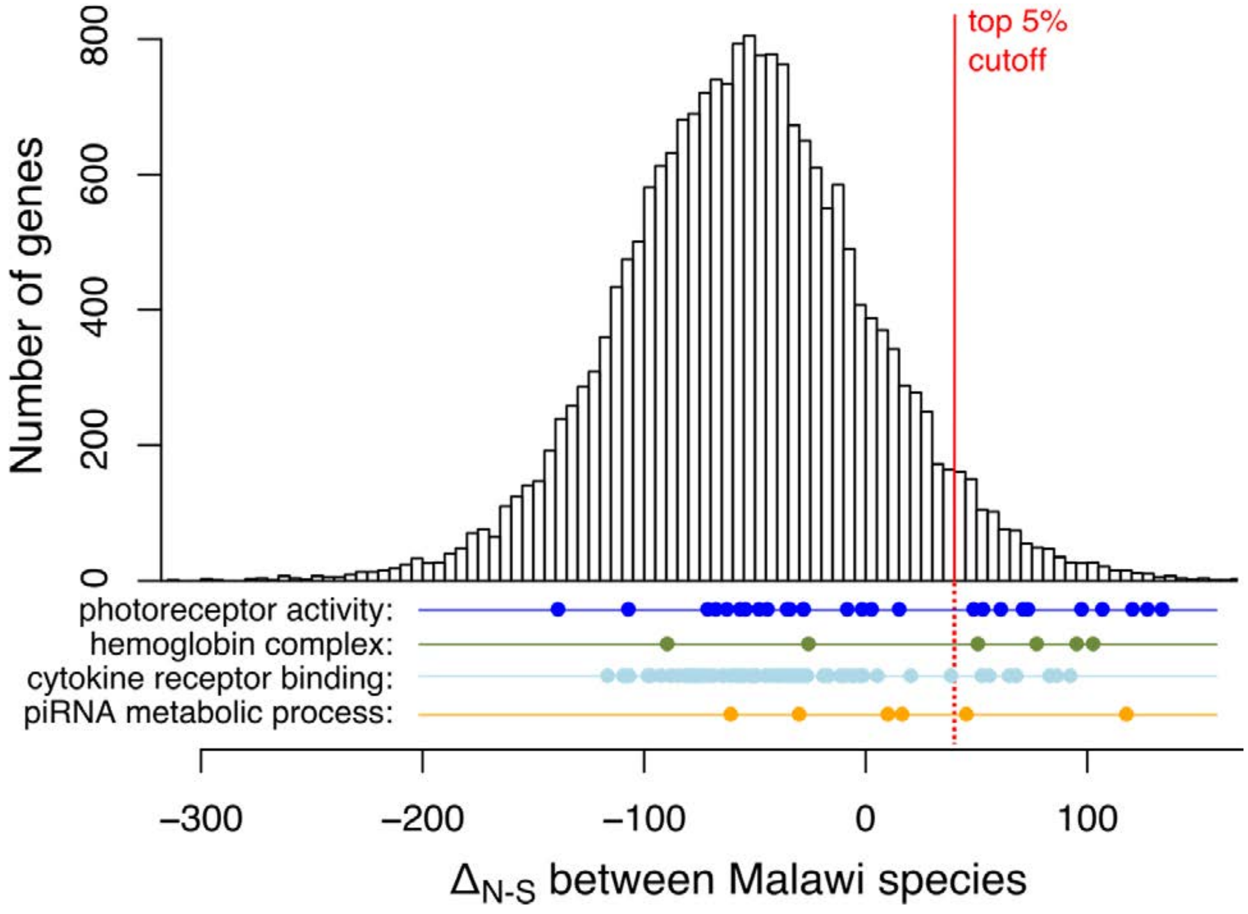


hannes.svardal@sanger.ac.uk

SELECTION

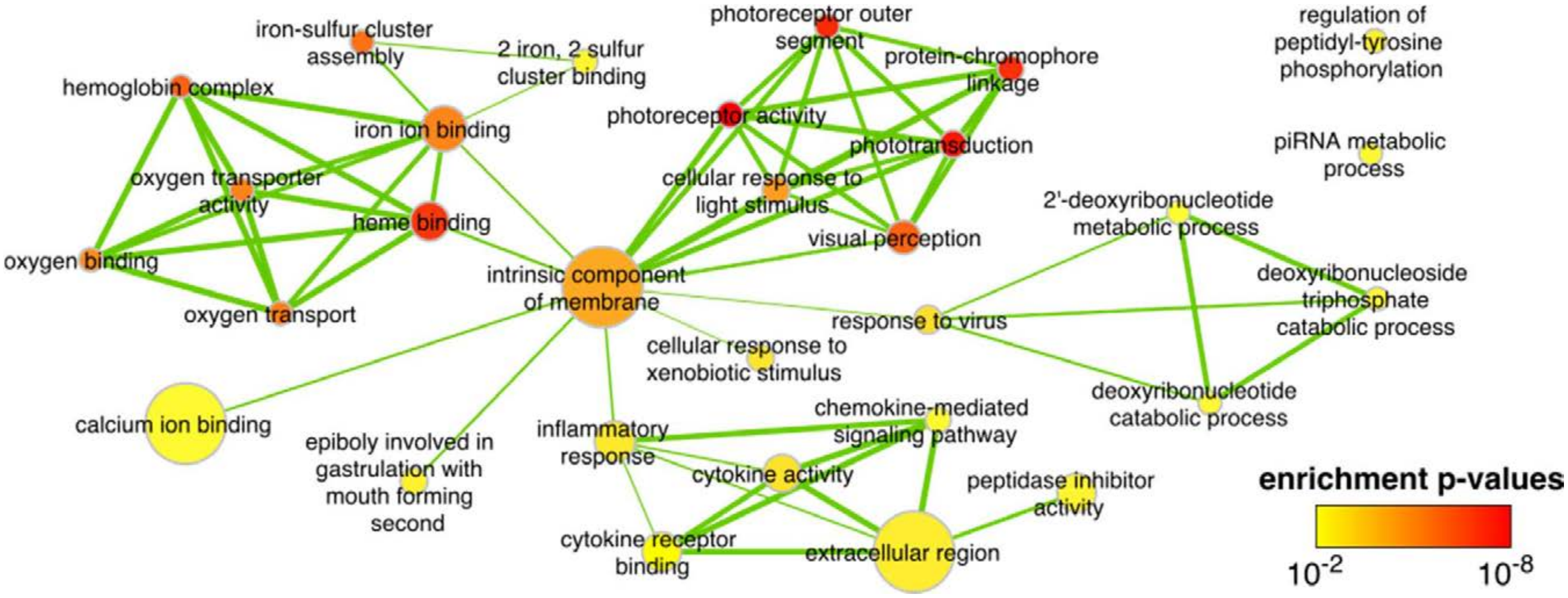
hannes.svardal@sanger.ac.uk

Nonsynonymous vs synonymous diversity



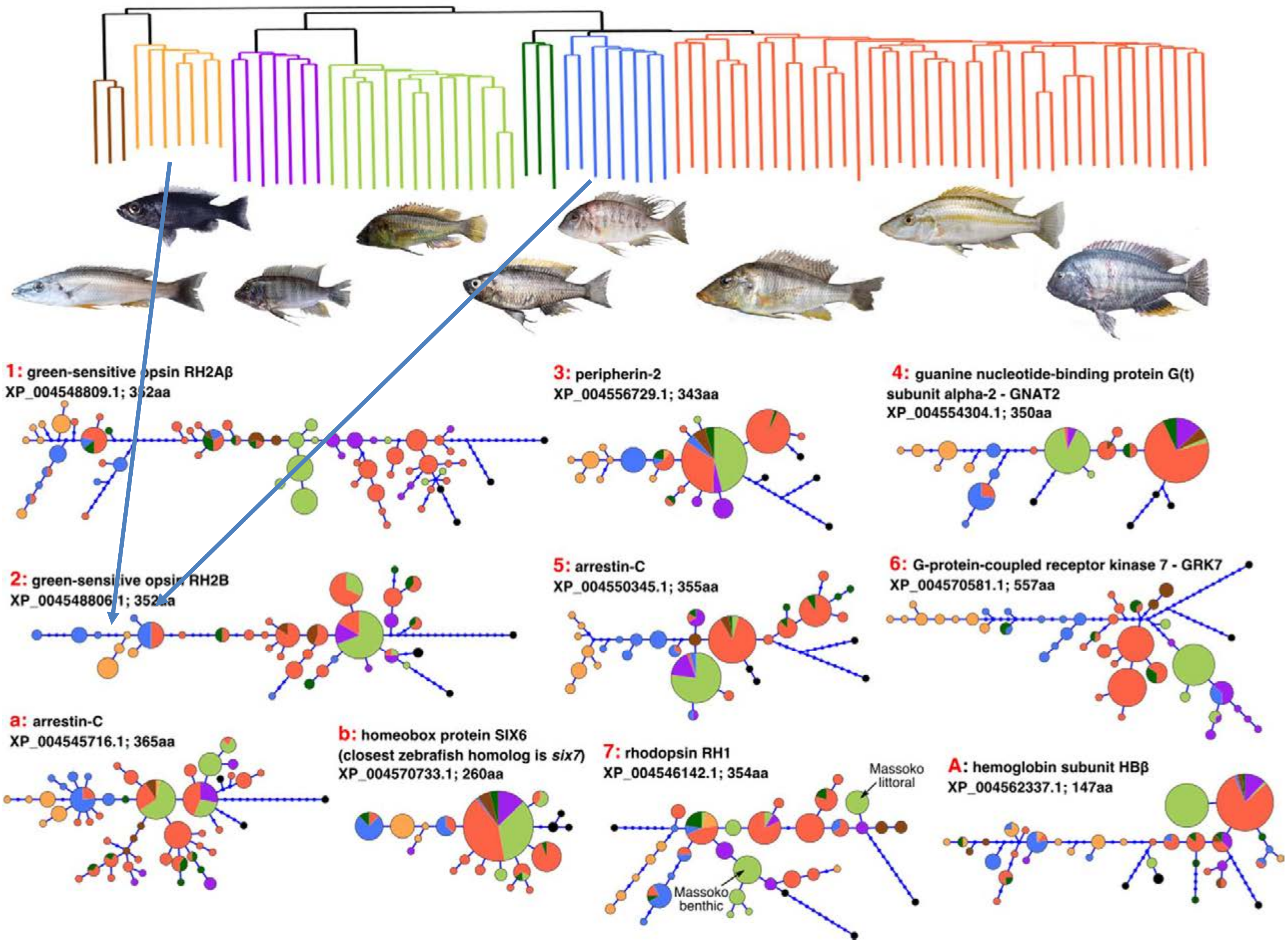
hannes.svardal@sanger.ac.uk

Visual system and oxygen transport pathways enriched for high selection scores



hannes.svardal@sanger.ac.uk

Haplotype networks of these genes show deep benthic – Diolotaxodon allele sharing

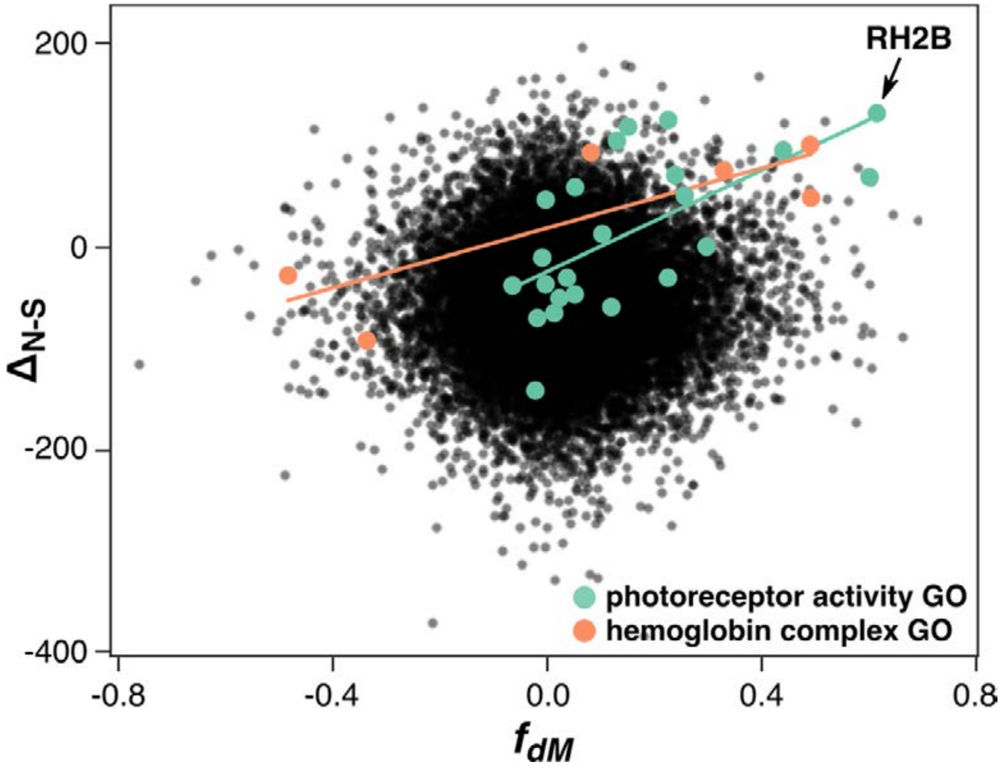


What is the mechanism behind deep benthic – Diolotaxodon allele sharing at these loci?

- Independent de-novo mutations?
- Selection on ancestral polymorphism?
- Adaptive introgression

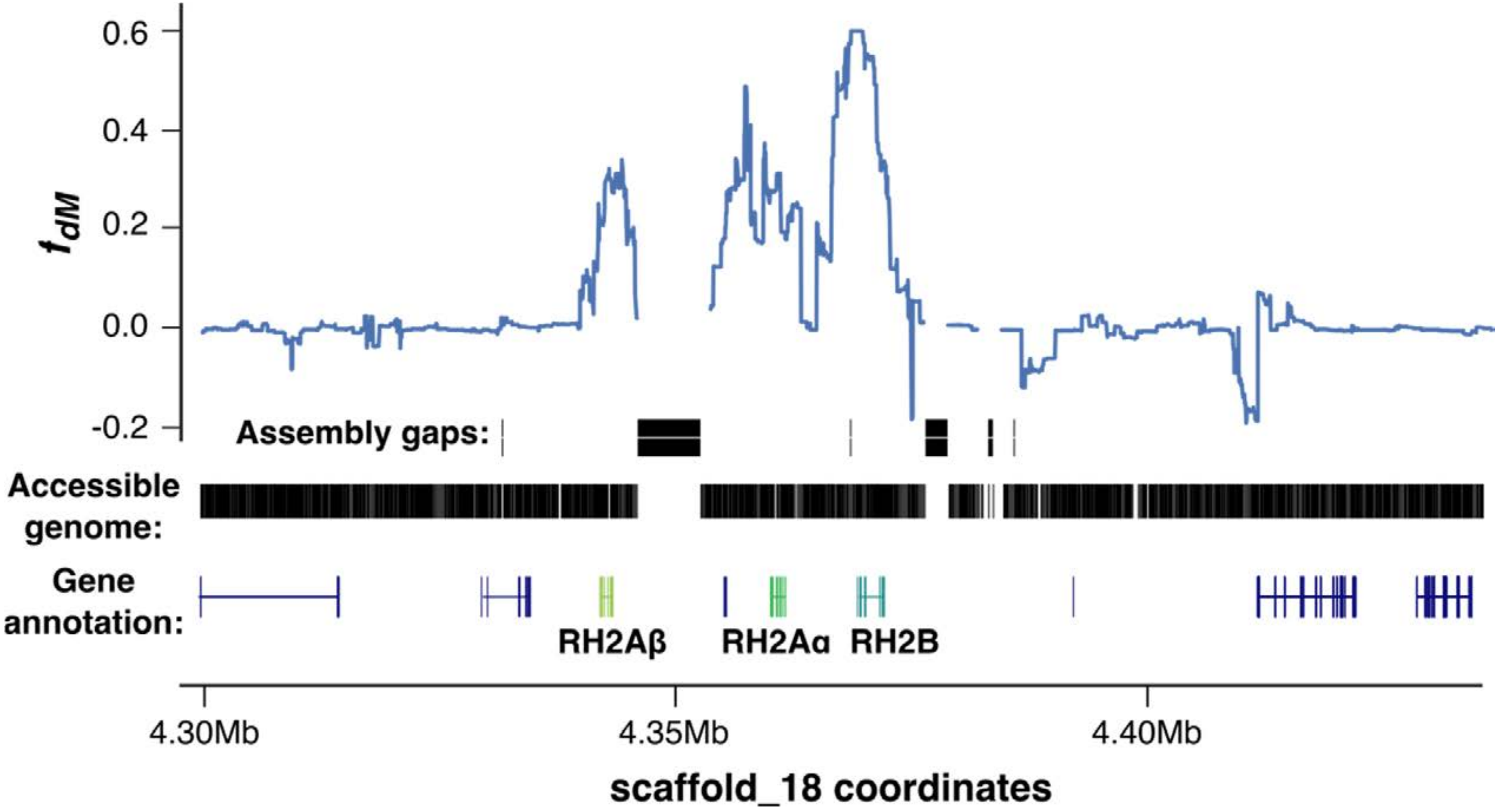
hannes.svardal@sanger.ac.uk

Elevated f-statistic for photoreceptor genes



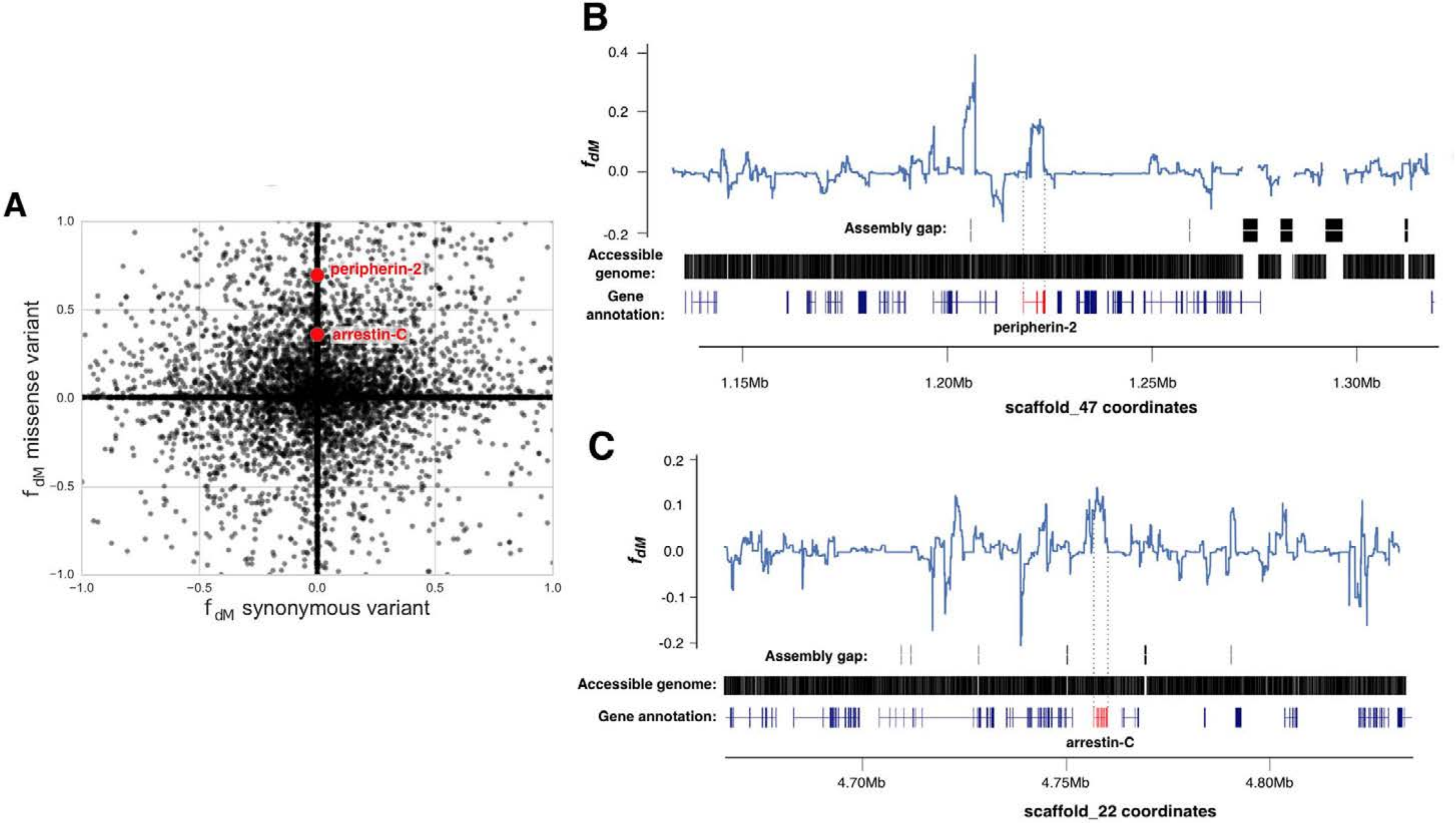
hannes.svardal@sanger.ac.uk

Some genes show long introgression haplotypes



hannes.svardal@sanger.ac.uk

For others f is only elevated for non-synonymous variants



hannes.svardal@sanger.ac.uk

Conclusion

- Developed branch specific f-statistic f_b
- Gene flow within and between major clades
- Evidence for selection on specific gene categories
- Excess allele sharing in ecologically relevant genes between two genetically distant groups sharing the same habitat (the deep)
- For some genes, the pattern is most consistent with adaptive introgression, for others with de-novo mutations or ancestral variation

hannes.svardal@sanger.ac.uk

Acknowledgements

Thank you!

Cichlids
Milan Malinski (Cambridge)
Richard Durbin
Eric Miska
Gregoire Vernaz
Mingliu Du
George Turner (Bangor)
Alexandra Tyers
Martin Genner (Bristol)

Data available at
cambridgecichlids.org

Preprint available soon.

Wellcome Trust Sanger Institute
Wellcome Trust
Alborada Cambridge-Africa Trust