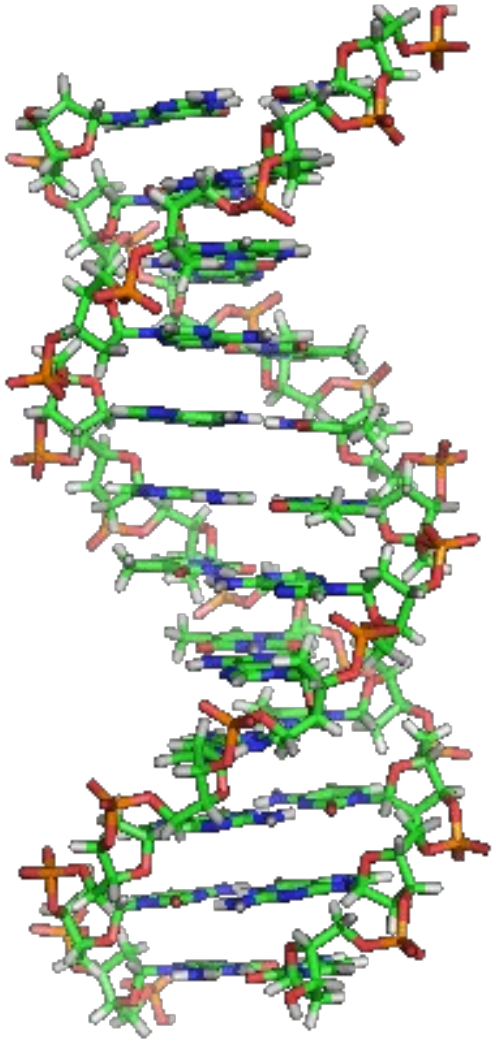


Intro to SNP calling

Ovidiu Paun

ovidiu.paun@univie.ac.at

http://plantgenomics.univie.ac.at

universität wien

Genomics is revolutionising biology

(1969)

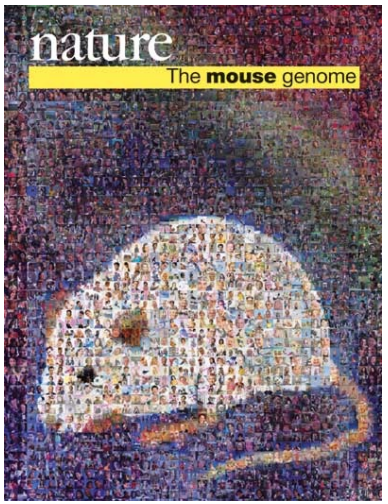
(1998)

(2000)

Human genome (Feb 2001)

Early Comparative Genomics

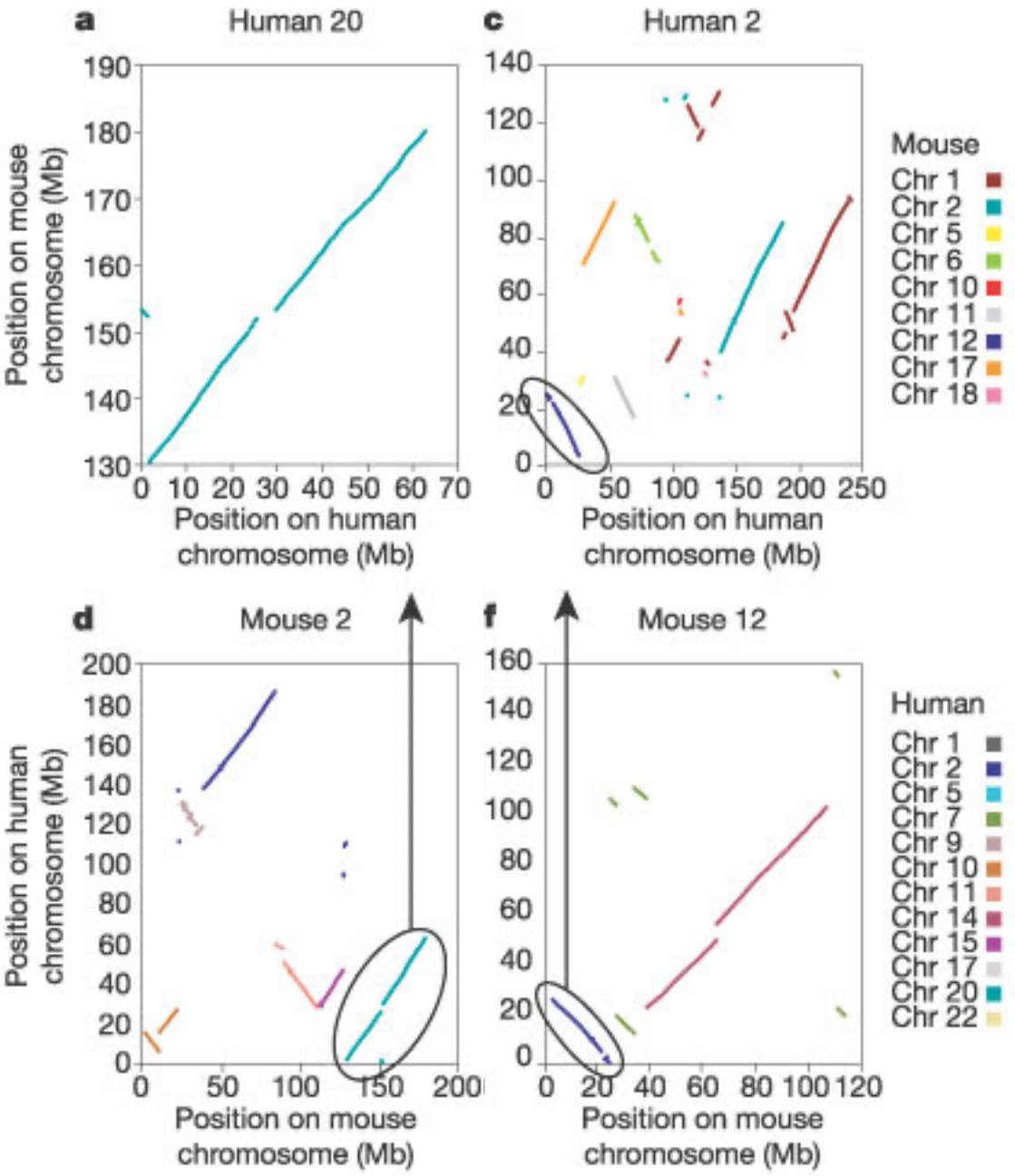
Chinwalla et al. 2002, *Nature* 420



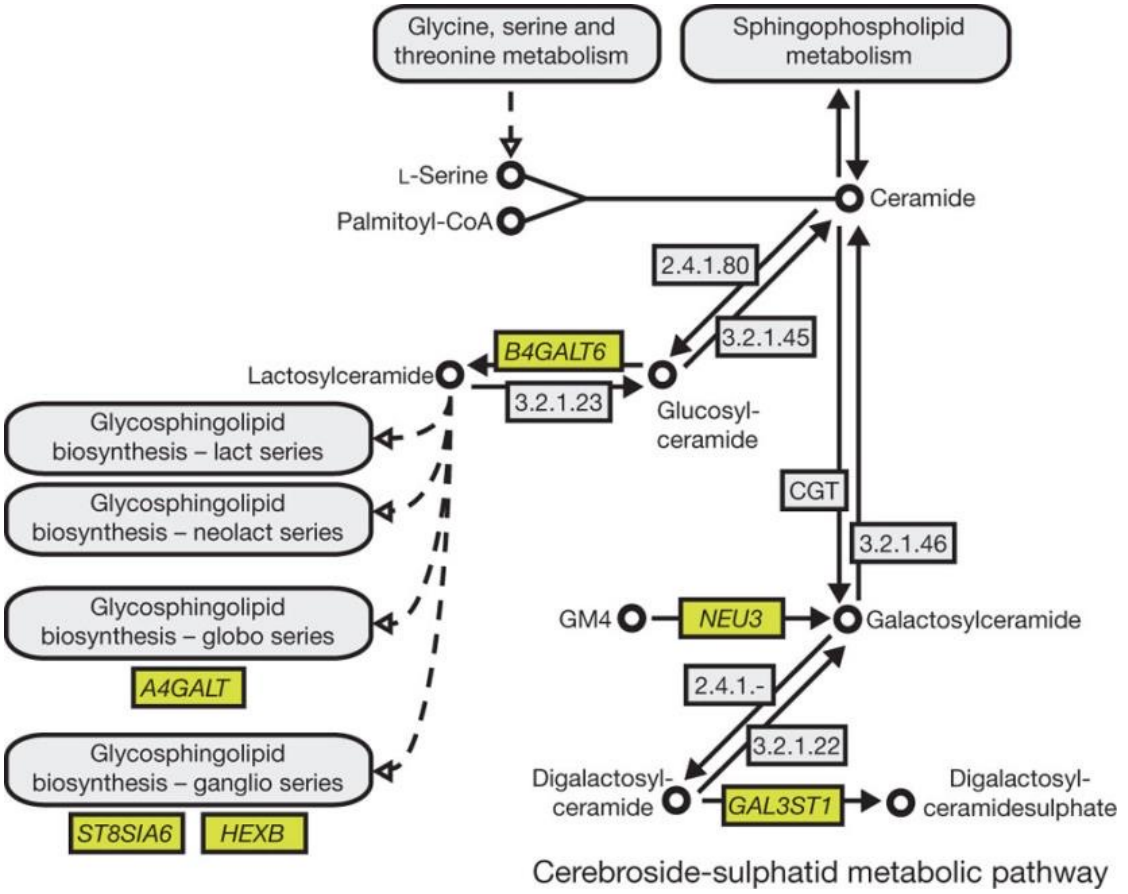
Dot plots comparing mouse and human chromosomes (80MYA divergence)

Large scale synteny

85% (60-99%) identity of protein-coding regions



Orang-utan genome(s)

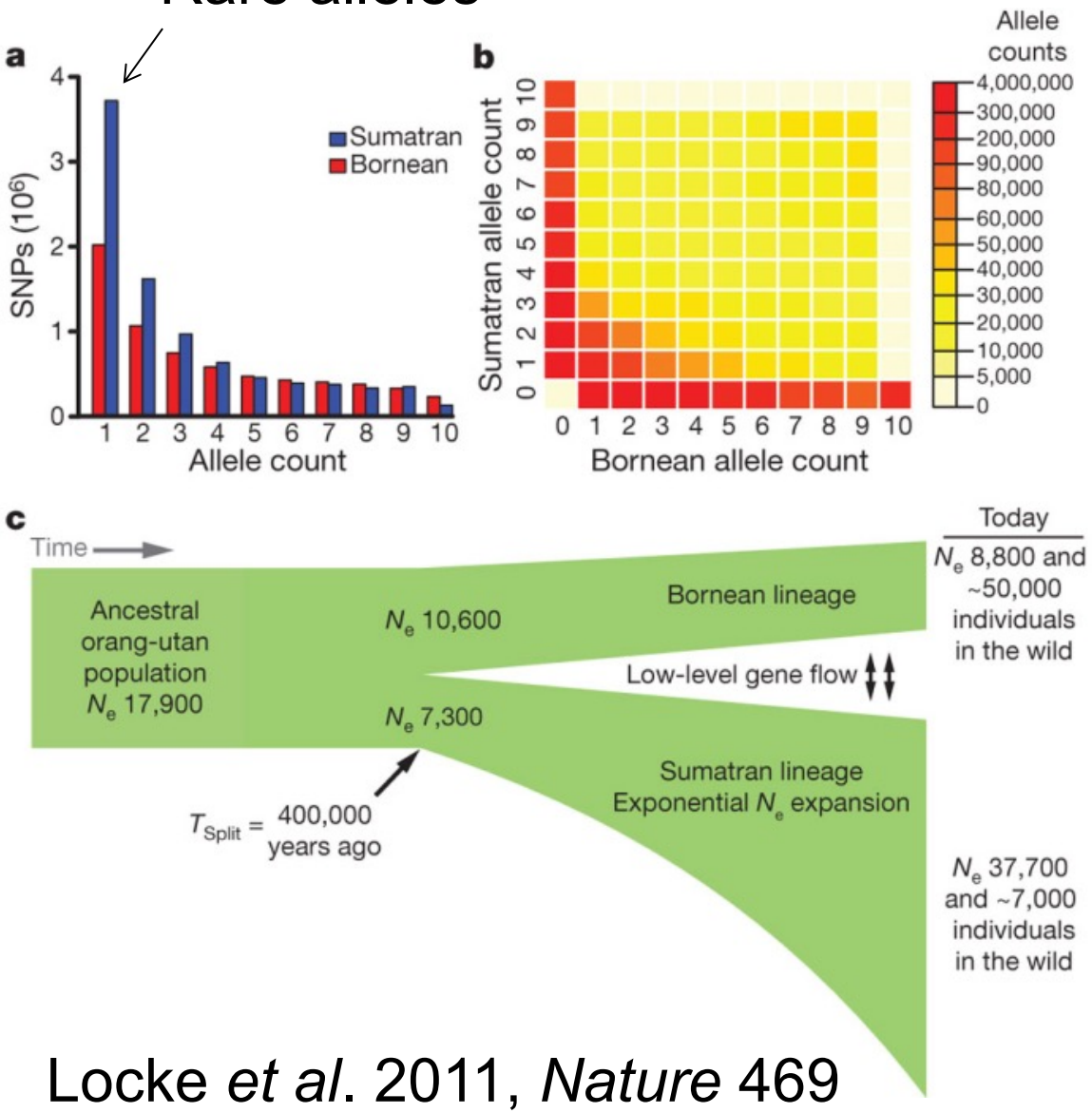


Locke *et al*. 2011, *Nature* 469

Evidence for positive selection in primates: ‘visual perception’ and ‘glycolipid metabolic processes’, important for the nervous system

Orang-utan genome(s)

Rare alleles



More recent split than thought

Census and N_e show opposing tendencies

Locke *et al*. 2011, *Nature* 469

NGS on fossil DNA deciphers human evolution

7 MAY 2010 VOL 328 **SCIENCE** www.sciencemag.org

A Draft Sequence of the Neandertal Genome

Richard E. Green,[1]*†‡ Johannes Krause,[1]†§ Adrian W. Briggs,[1]†§ Tomislav Maricic,

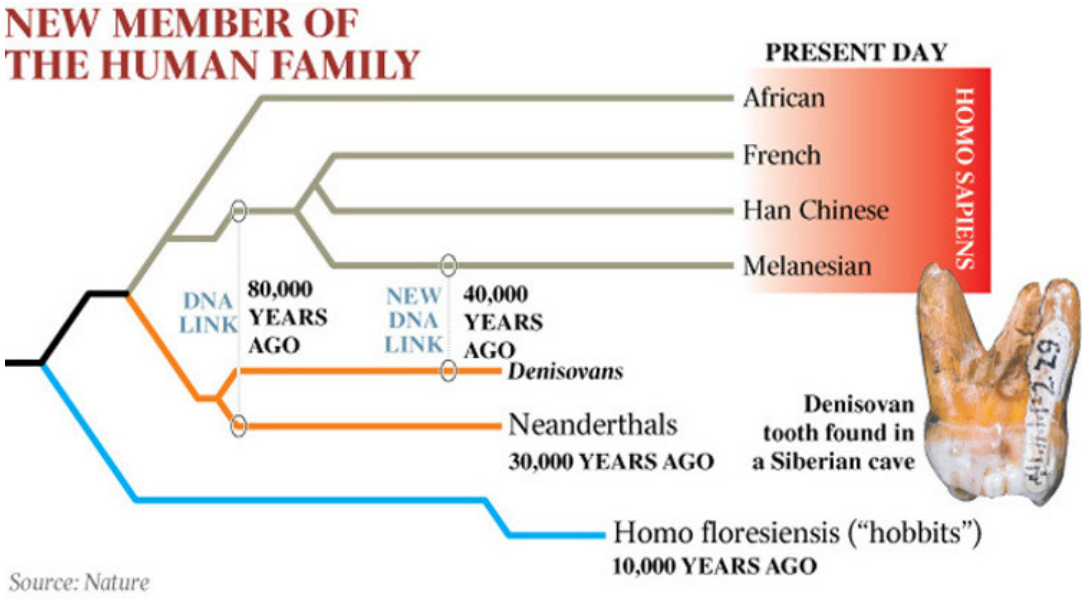
Sciencexpress Res

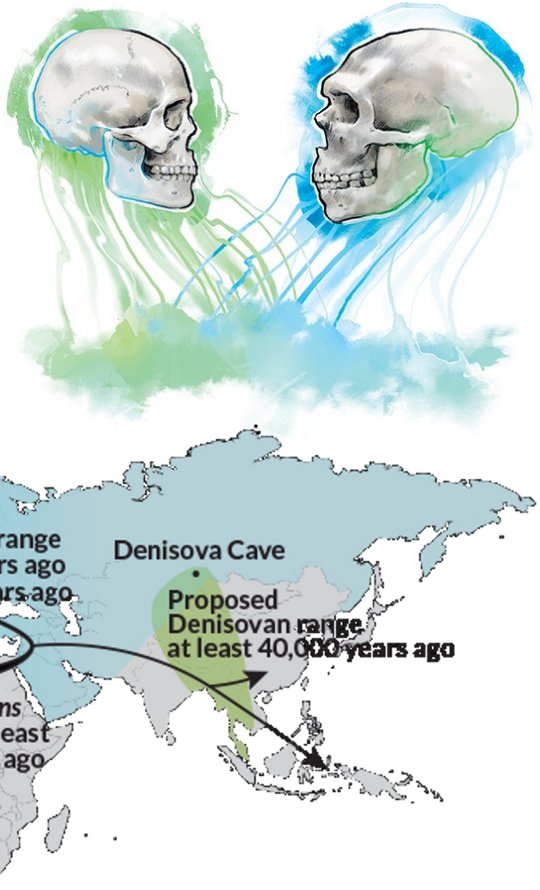
Corrected 31 August 2012

A High-Coverage Genome Sequence from an Archaic Denisovan Individual

Matthias Meyer,[1*‡] Martin Kircher,[1*†] Marie-Theres Gansauge,[1] Heng Li,[2]

NGS on fossil DNA deciphers human evolution





Source: S. Vattathil and J. M. Akey/*Cell* 2016


Sciencexpress Res

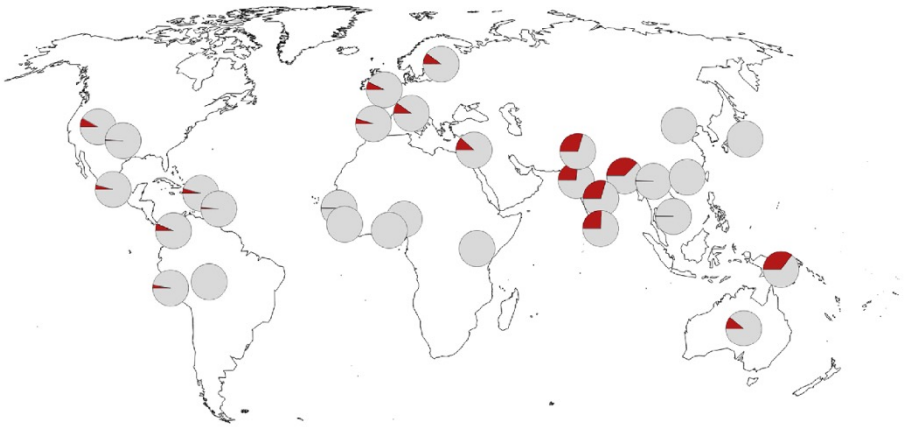
Corrected 31 August 2012

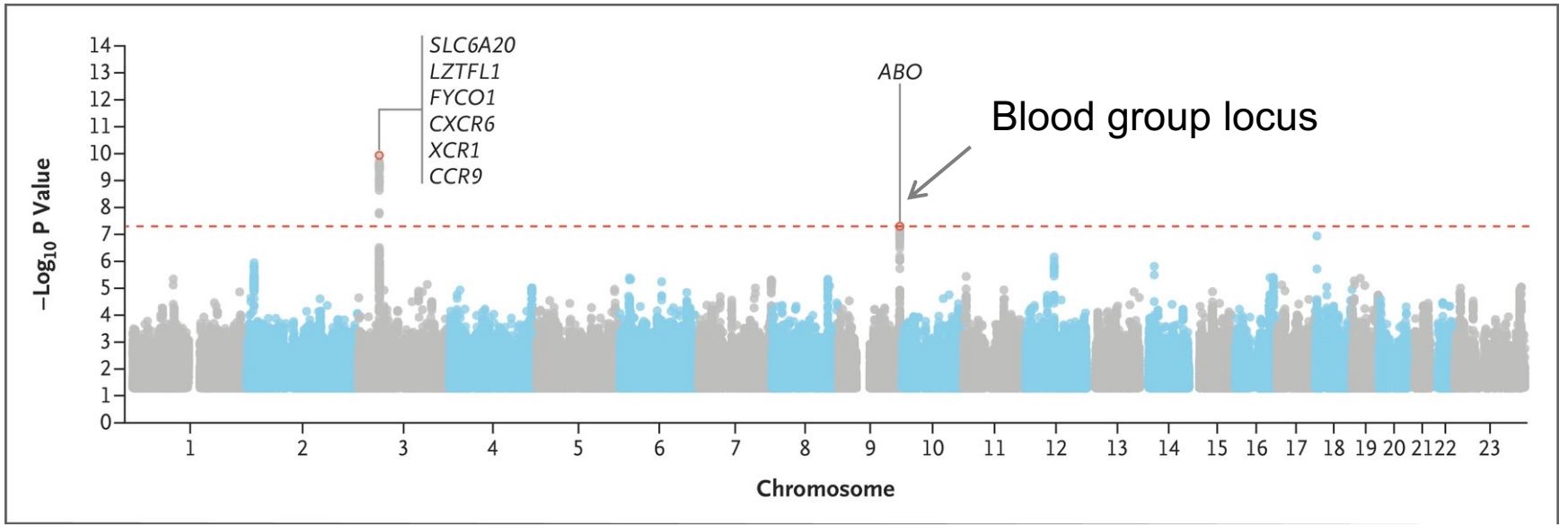
A High-Coverage Genome Sequence from an Archaic Denisovan Individual

Matthias Meyer,[1*‡] Martin Kircher,[1*†] Marie-Theres Gansauge,[1] Heng Li,[2]

The major genetic risk factor for severe COVID-19 is inherited from Neanderthals

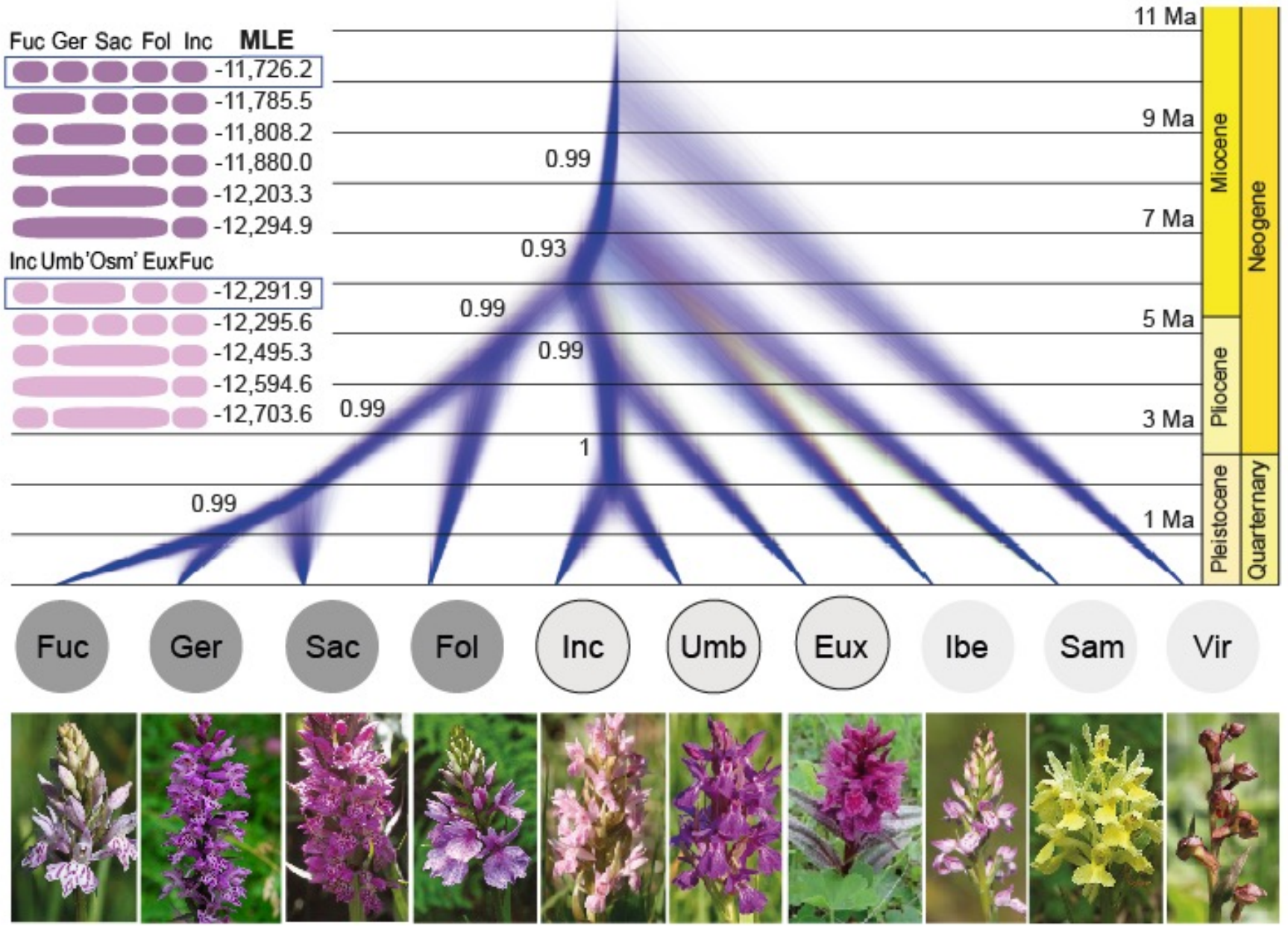
https://doi.org/10.1038/s41586-020-2818-3
Hugo Zeberg [1,2] ✉ & Svante Pääbo [1,3] ✉
Published online: 30 September 2020





The Severe Covid-19 GWAS Group 2020, New England J of Medicine

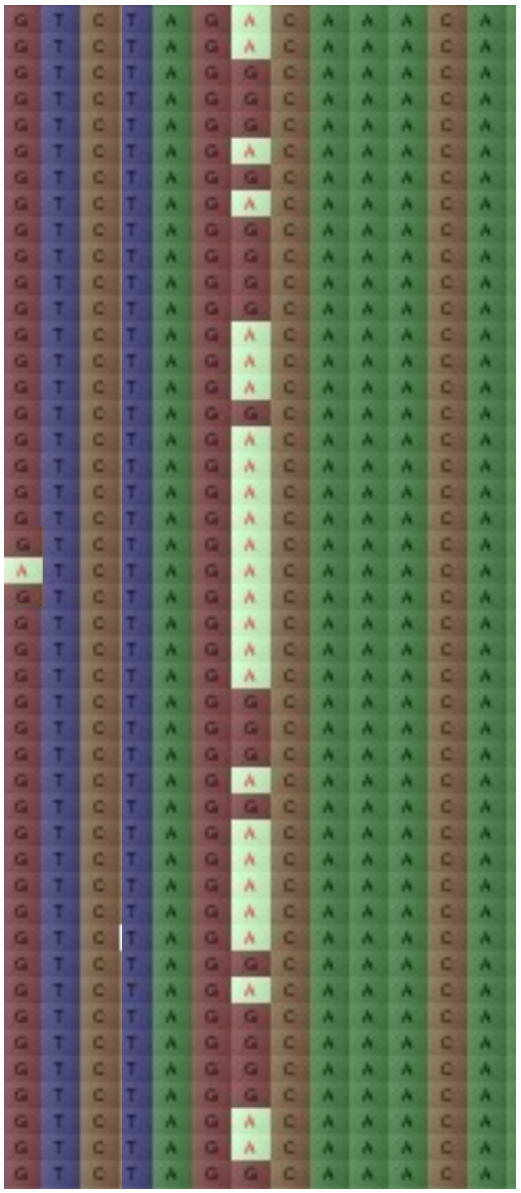
Pyhlogenomics (eg species tree of *Dactylorhiza*)



Brandrud, .., Paun 2020 *Syst Biol*

All this is inferred based on variation in genomes

Aim of SNP calling in a nutshell:
decide which variation represents
true polymorphism

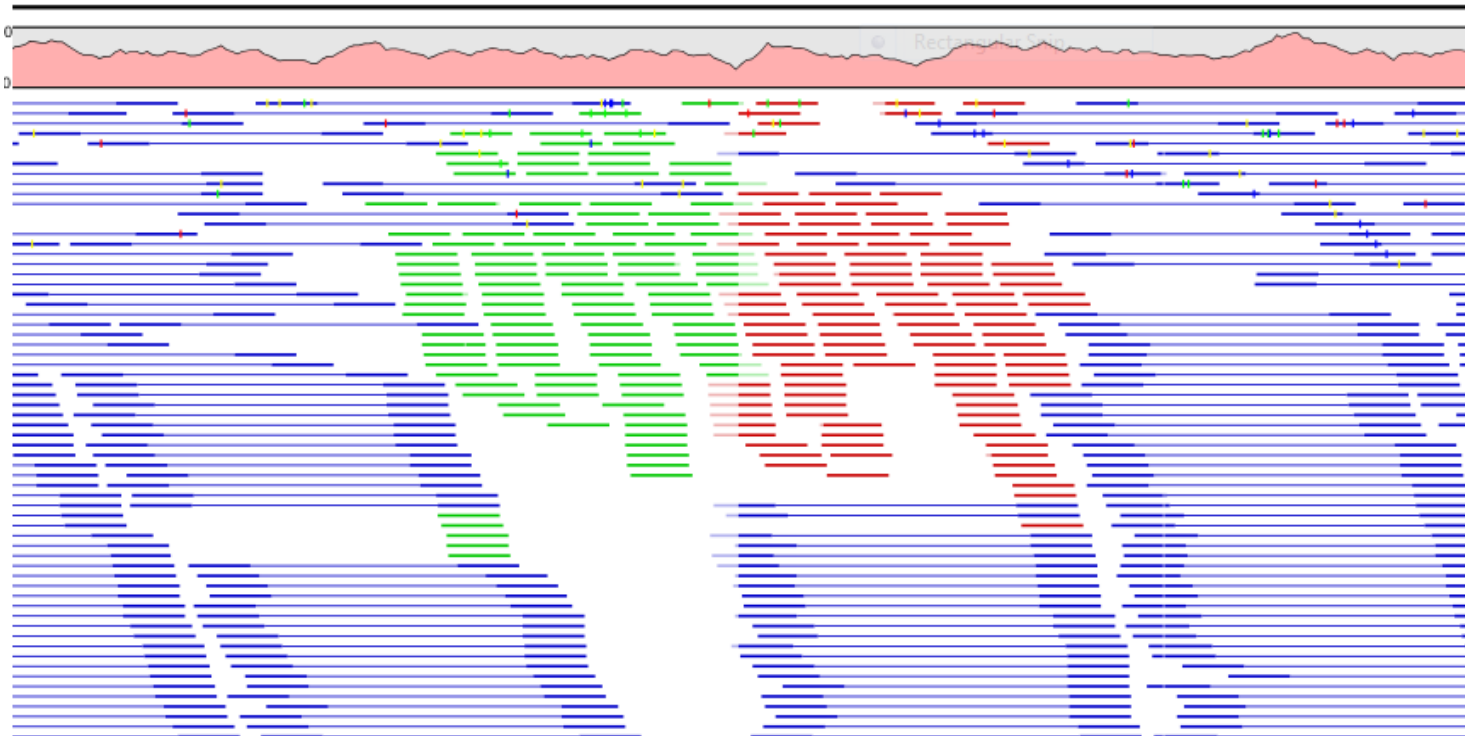


Simple example reads mapping

Find the aligning position of short reads within a reference genome

Mapping software are getting pretty good, but might miss crucial differences (repeats…)

Most used software BWA and BOWTIE

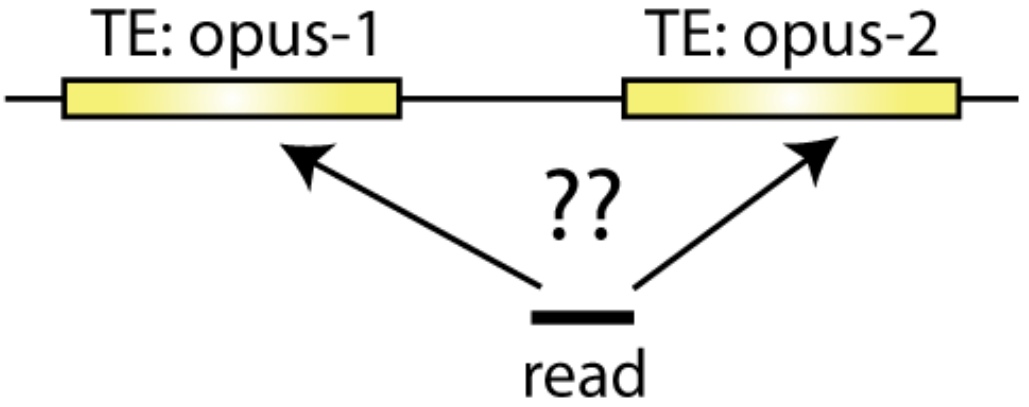


Mapping quality

SAM/BAM files contain mapping quality. Similar to base quality, mapping quality is in general the log scaled probability that the position of the read is wrongly mapped

0 -> every read is wrongly mapped
20 -> 1 in 100 reads is wrongly mapped
30 -> 1 in 1000 reads is wrongly mapped

A major cause for incorrect or ambiguous mappings are repetitive regions in the genome



Coverage and coverage bias

Coverage/read depth
Coverage: The average number of times each bp of a region is sequenced
Depends very much on application/question
De novo: >30X
Mapping: 5X sometimes enough, sometimes 20X or even more

Coverage bias
Extremely important for quantitative gene expression (eg RNA-seq)

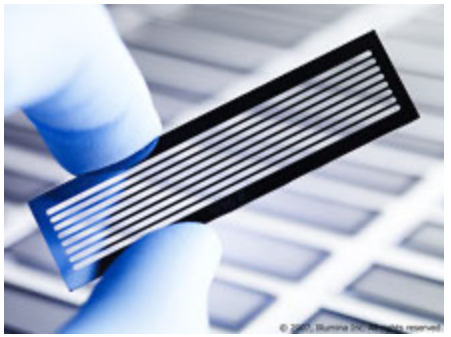
Uneven coverage can be due to biases at several levels

- DNA extraction
- PCR amplification (better PCR free libraries; PCR is completely avoided in single-molecule sequencing!)
- Sequencing

SAM/BAM data formats

ID	flag	alignment info							seq	qual	addit. info
HWI-EAS225_37:3:108:40:58#0/1	0	ChrC	1	255	76M	*	0	0	ATG...	III...	XA:i:1 ….

BAM is the SAM compressed binary format. Details on the specifications are available here:
http://samtools.sourceforge.net/SAM-1.3.pdf



Before Variant calling

Sorting by Coordinates (listing the reads according to the coordinates in the reference genome)

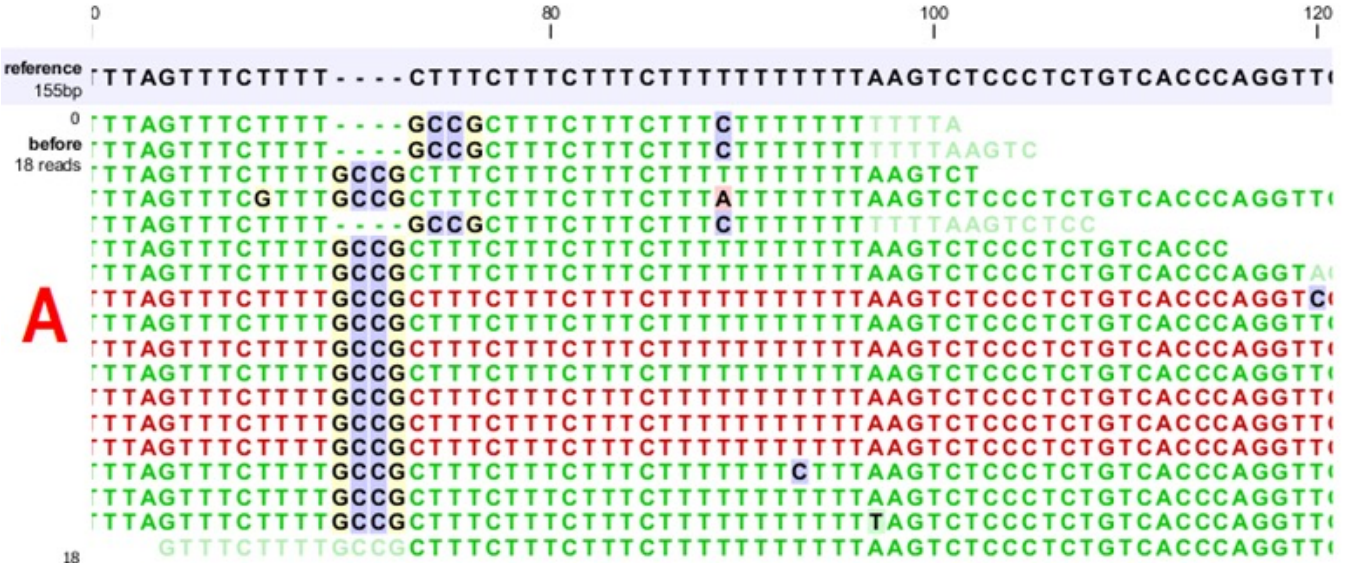
AddOrReplaceReadGroups (adding accession information in the BAM header)

Mark PCR duplicates

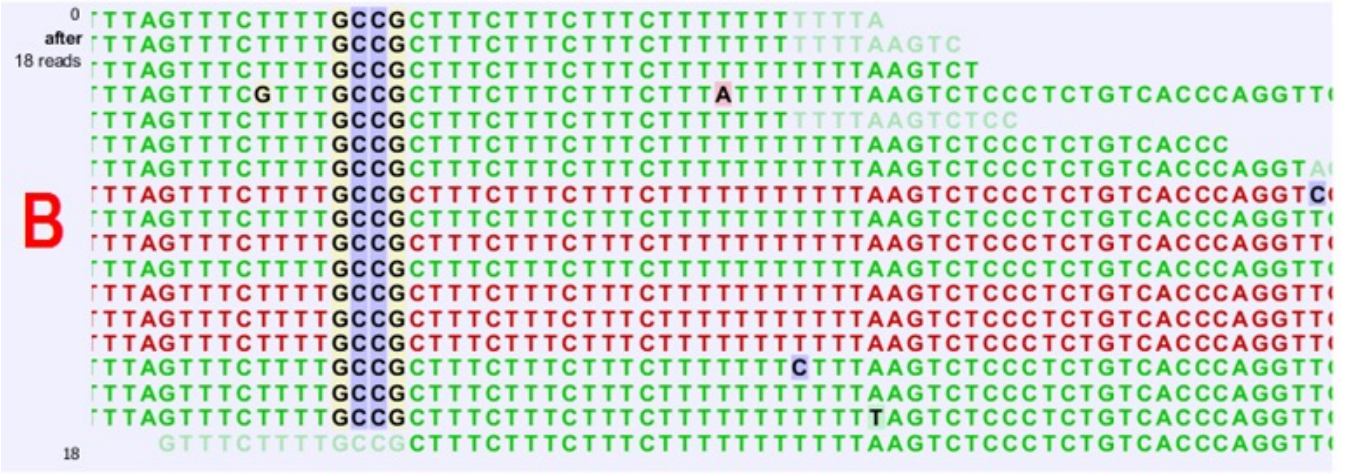
Indel realignment (refining mapping)

Refining the Mapping – indel realignment

Before



After



Calling variants - objective

For each site of the genome determine

- Whether or not there is **any variation** at this site (across all samples)
- If there is variation, **assign a genotype** (for diploids: pair of alleles) to each sample
- Report variant sites (positions, alleles, sample genotypes, …)

Calling variants - approach

Classical way: based on thresholds (e.g., ratio of reference/non-reference bases)

Most recent: probabilistic framework (eg, ANGSD https://www.popgen.dk/angsd/index.php/ANGSD)

- Calculate and report **probability** of a site being a variant (given read alignments)
- Calculate and report probabilities (or **likelihoods**) of various sample genotypes (given read alignments)

Variant calling

Different types of variants:

SNVs (single nulceotide variants)

MNVs (multiple, neighbouring SNVs)

indels

For a diploid organism there are 2 alleles:

- expected are A/A, A/C, A/G, A/T, A/-, C/C, etc

Variant Call Format (.vcf)

```
##fileformat=VCFv4.0
##fileDate=20151119
##source="Stacks v1.29"
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=AF,Number=.,Type=Float,Description="Allele Frequency">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=AD,Number=1,Type=Integer,Description="Allele Depth">
##FORMAT=<ID=GL,Number=.,Type=Float,Description="Genotype Likelihood">
#CHROM  POS   ID   REF  ALT  QUAL  FILTER  INFO                FORMAT       ind10                  ind12
un      6059  321  C    T    67    PASS    NS=2;AF=0.759,0.241 GT:DP:AD:GL  0/1:83:41,42:.,115.06,.  0/0:104:52,52:.,144.17,.
un      6457  341  G    A    47    PASS    NS=2;AF=0.983,0.017 GT:DP:AD:GL  0/0:90:45,45:.,124.77,.  1/0:125:62,63:.,173.29,.
```

- text file format (likely stored in compressed manner)
- encodes SNPs and other structural variants
- contain meta-information lines (##) - describe symbols found later on
- a header line (#), and data lines each containing information about a position in the genome
- could contain information from a single or multiple samples
- could contain genotype information on samples or not
- is produced by most commonly used NGS analysing tools, e.g., GATK, STACKS

Variant Call Format (.vcf)

```
##fileformat=VCFv4.0
##fileDate=20151119
##source="Stacks v1.29"
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=AF,Number=.,Type=Float,Description="Allele Frequency">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=AD,Number=1,Type=Integer,Description="Allele Depth">
##FORMAT=<ID=GL,Number=.,Type=Float,Description="Genotype Likelihood">
#CHROM  POS   ID   REF  ALT  QUAL  FILTER  INFO                FORMAT       ind10                  ind12
un      6059  321  C    T    67    PASS    NS=2;AF=0.759,0.241 GT:DP:AD:GL  0/1:83:41,42:.,115.06,.  0/0:104:52,52:.,144.17,.
un      6457  341  G    A    47    PASS    NS=2;AF=0.983,0.017 GT:DP:AD:GL  0/0:90:45,45:.,124.77,.  1/0:125:62,63:.,173.29,.
```

- ID: some ID for the variant, if known
- REF, ALT: reference and alternative alleles (on forward strand of reference)
- QUAL: -10*log(1-p), where p is the probability of a variant being present given the read data
- FILTER: whether the variant failed a filter (filters defined by the user or program processing the file)
- INFO: annotations defined in the meta-information

Variant Call Format (.vcf)

```
##fileformat=VCFv4.0
##fileDate=20151119
##source="Stacks v1.29"
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=AF,Number=.,Type=Float,Description="Allele Frequency">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=AD,Number=1,Type=Integer,Description="Allele Depth">
##FORMAT=<ID=GL,Number=.,Type=Float,Description="Genotype Likelihood">
#CHROM  POS   ID   REF  ALT  QUAL  FILTER  INFO                FORMAT       ind10                  ind12
un      6059  321  C    T    67    PASS    NS=2;AF=0.759,0.241 GT:DP:AD:GL  0/1:83:41,42:.,115.06,.  0/0:104:52,52:.,144.17,.
un      6457  341  G    A    47    PASS    NS=2;AF=0.983,0.017 GT:DP:AD:GL  0/0:90:45,45:.,124.77,.  1/0:125:62,63:.,173.29,.
```

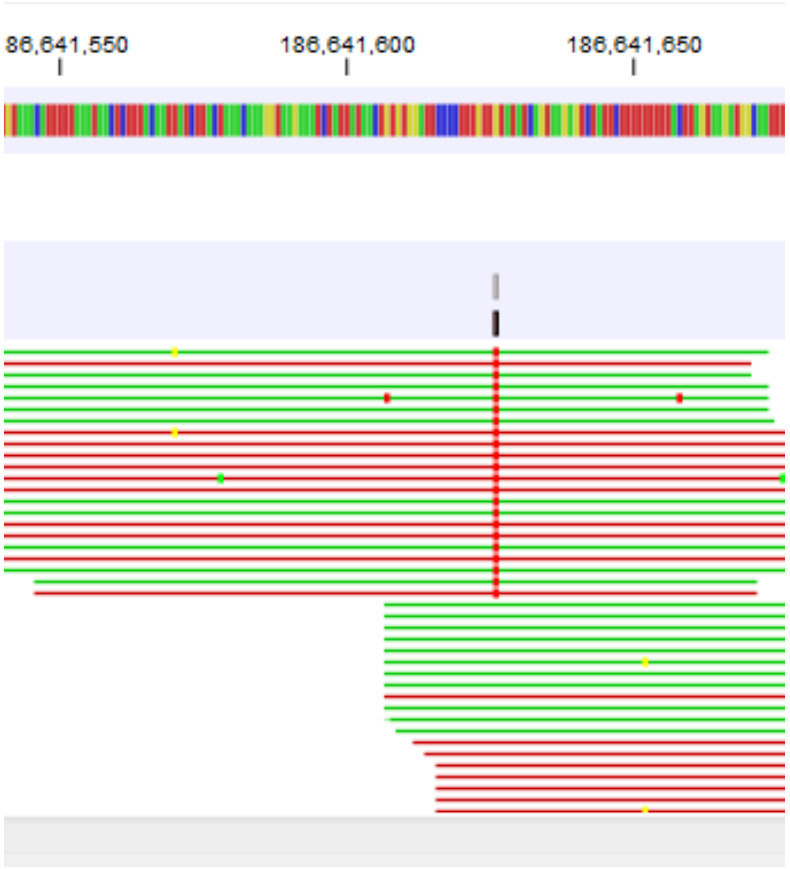
- GT: genotype
  - 0/0 reference homozygote
  - 0/1 reference-alternative heterozygote
  - 1/1 alternative homozygote
- AD: allele depths
- DP: total depth (may be different from sum of AD, as the latter include only reads supporting alleles)
- GL: genotype likelihood (phred-scaled), normalized to the best genotype

Variant filtering

Filters high-confidence variants to be retained for further analyses

Could include filtering steps on:

- Mapping direction - strand bias: Fisher strand value (FS) > 60
- Correction by coverage: Quality by depth values (QD) < 2
- Mapping quality (MQ) < 40

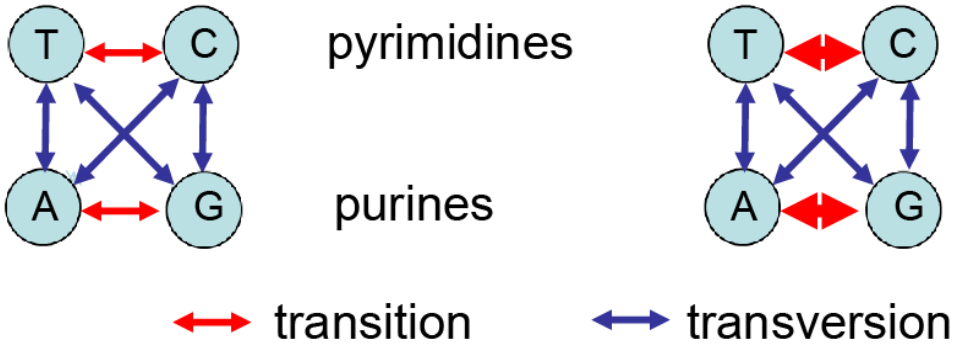


Variant filtering

How to know if we have a high quality SNP dataset?

A good quality criteria for a vcf set is that Ts/Tv should be higher than 1.5 genome-wide, or higher than 2 for exonic regions

e.g., Ts/Tv = 2.71 for *Drosophila* GstD1 gene

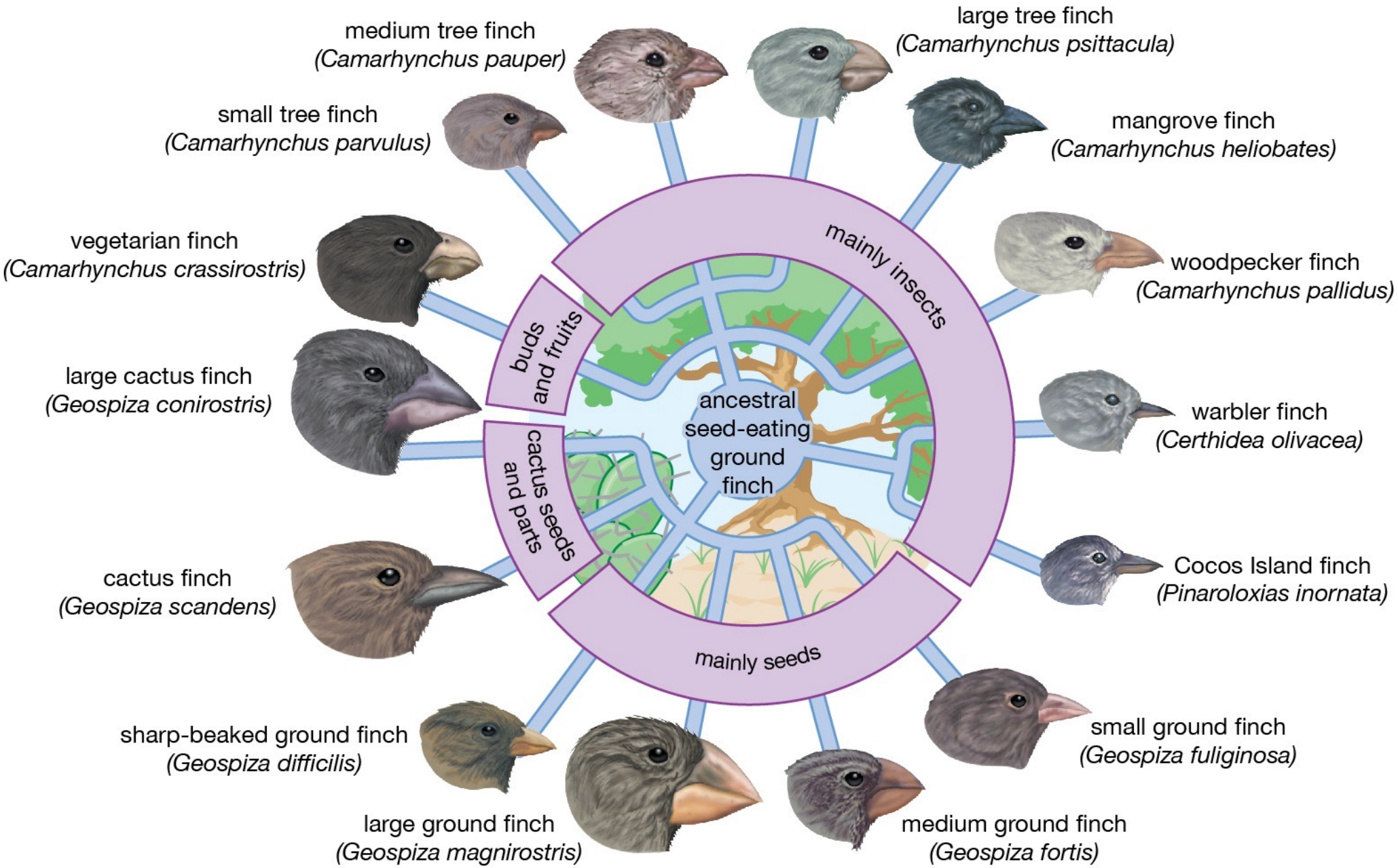


Adaptive radiation of *Tillandsia* subgenus *Tillandsia* (Bromeliaceae)

Team:
Clara Groot Crego, Gil Yardeni, Michael HJ Barfuss, Jaqueline Hess, Marylaure de La Harpe, Neil McNair, Margot Paris, Walter Till, Christian Lexer, Thibault Leroy, Ovidiu Paun

universität wien

Adaptive radiation in Galapagos finches



Rapid radiation of *Diospyros* on New Caledonia

30+ *Diospyros* species radiate on New Caledonia after a long distance dispersal

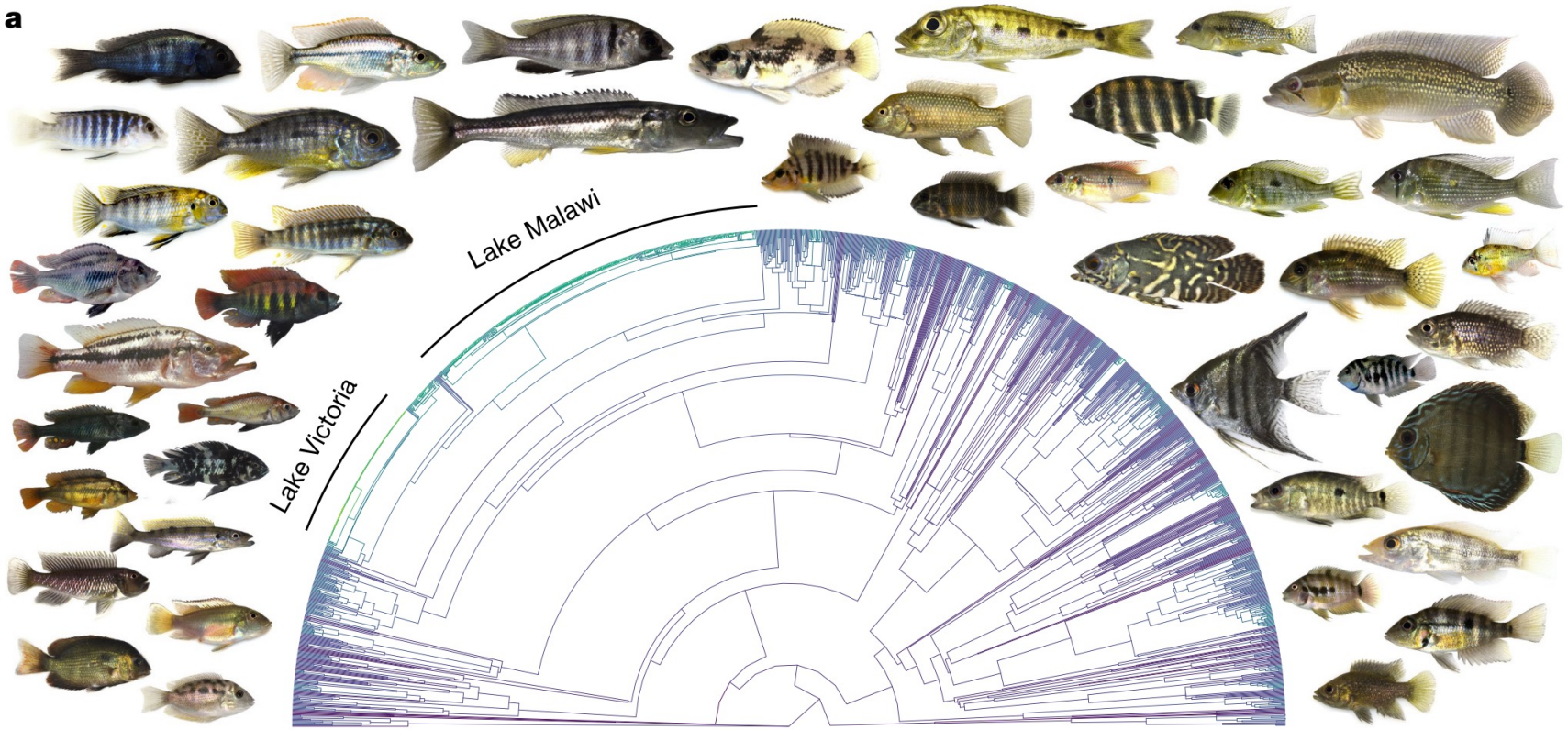
New Caledonia:
- a biodiversity hotspot
- highest level of endemism worldwide
- highly heterogeneous environment

D. veilonii

D. minimifolia

Adaptive radiation:

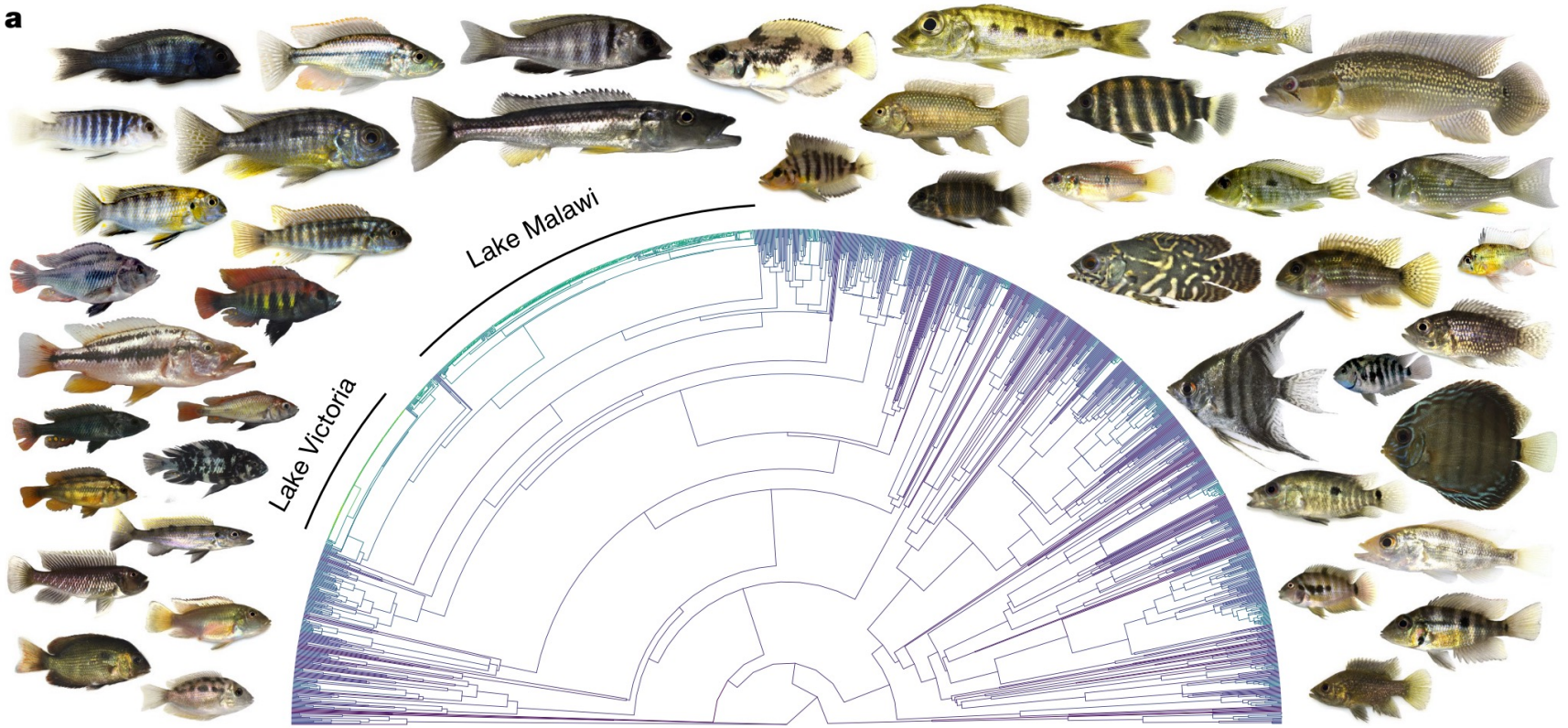
radiation through ecological opportunity



McGee et al, 2020

Adaptive radiation:

radiation through ecological opportunity



Recent studies show:

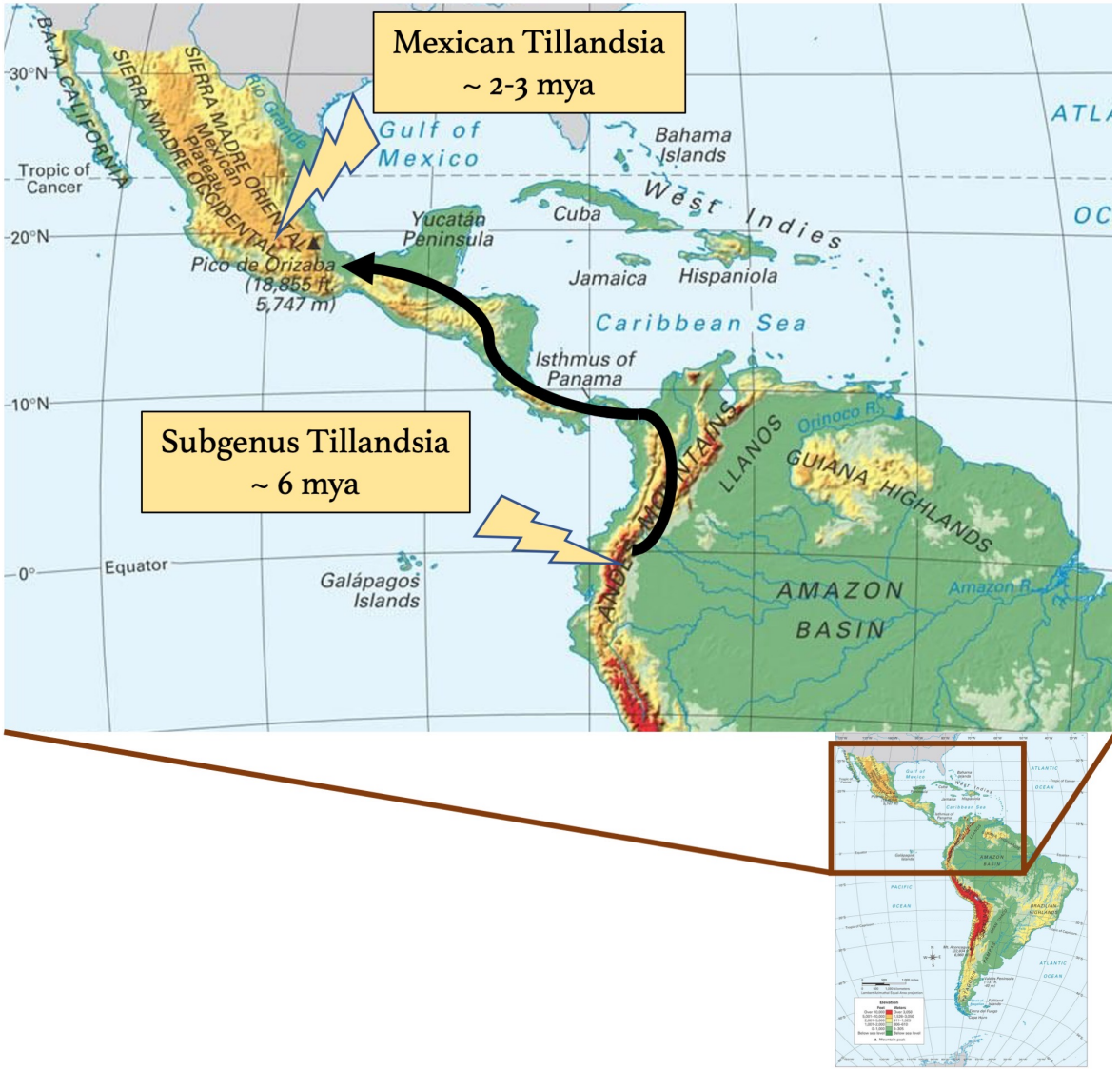
The story is not so simple

Tillandsia as a plant model for adaptive radiation

~ 350 species in total

~ 30 species in ancestral area of S America

~ 300 species in Central America



Mendoza et al. 2017

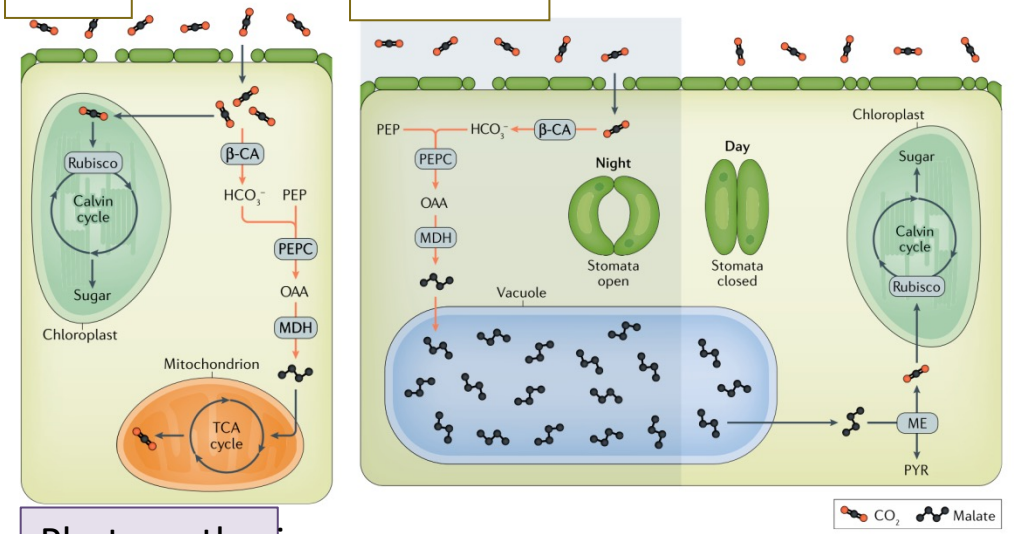
Key innovation traits

Epiphytism

Tank formation

C3

CAM



Photosynthesis

From Heyduk et al. 2019 (Nat. Genetics)

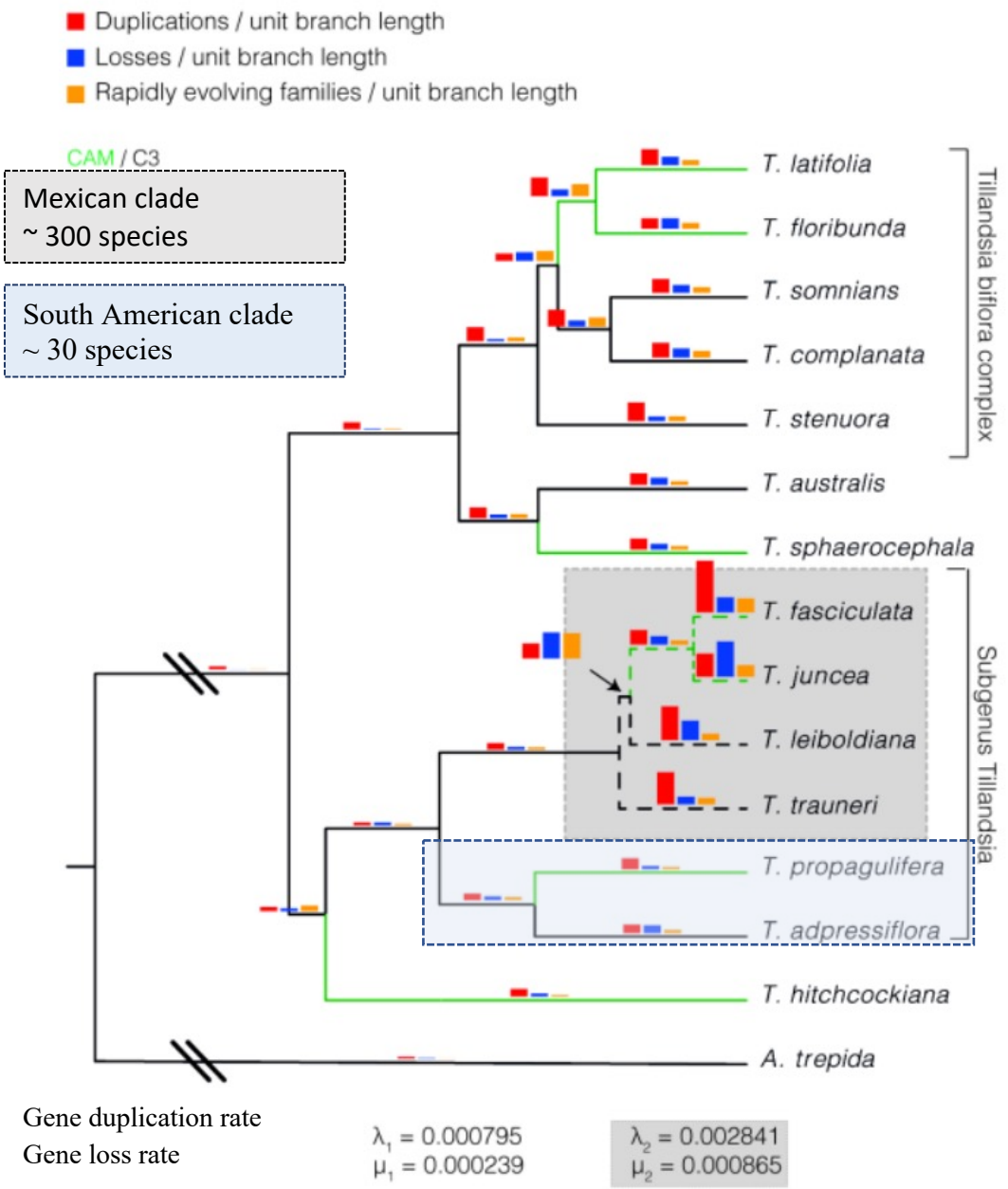
Pollination syndromes

Genome dynamics in *Tillandsia*

- Comparatively high rates of gene loss and duplication in radiating *Tillandsia*

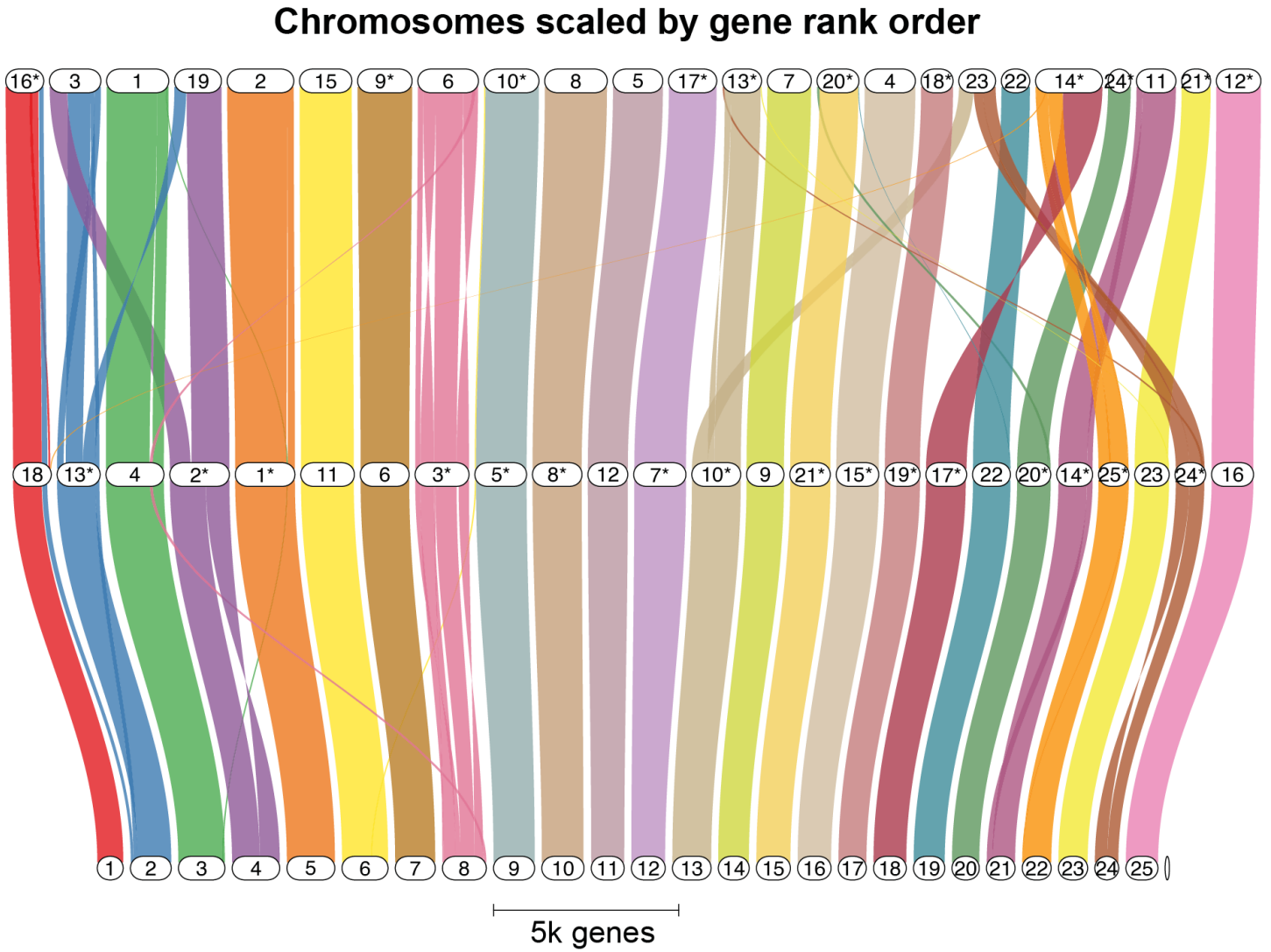
What are the sources of variation driving the divergent evolution of Tillandsia?

- Structural variation?
- Regulatory evolution?
- Adaptive sequence evolution?



De La Harpe et al (2020)

Despite high genome dynamics and a difference in chromosome number, synteny largely retained with notable exceptions



C3

T. leiboldiana

CAM

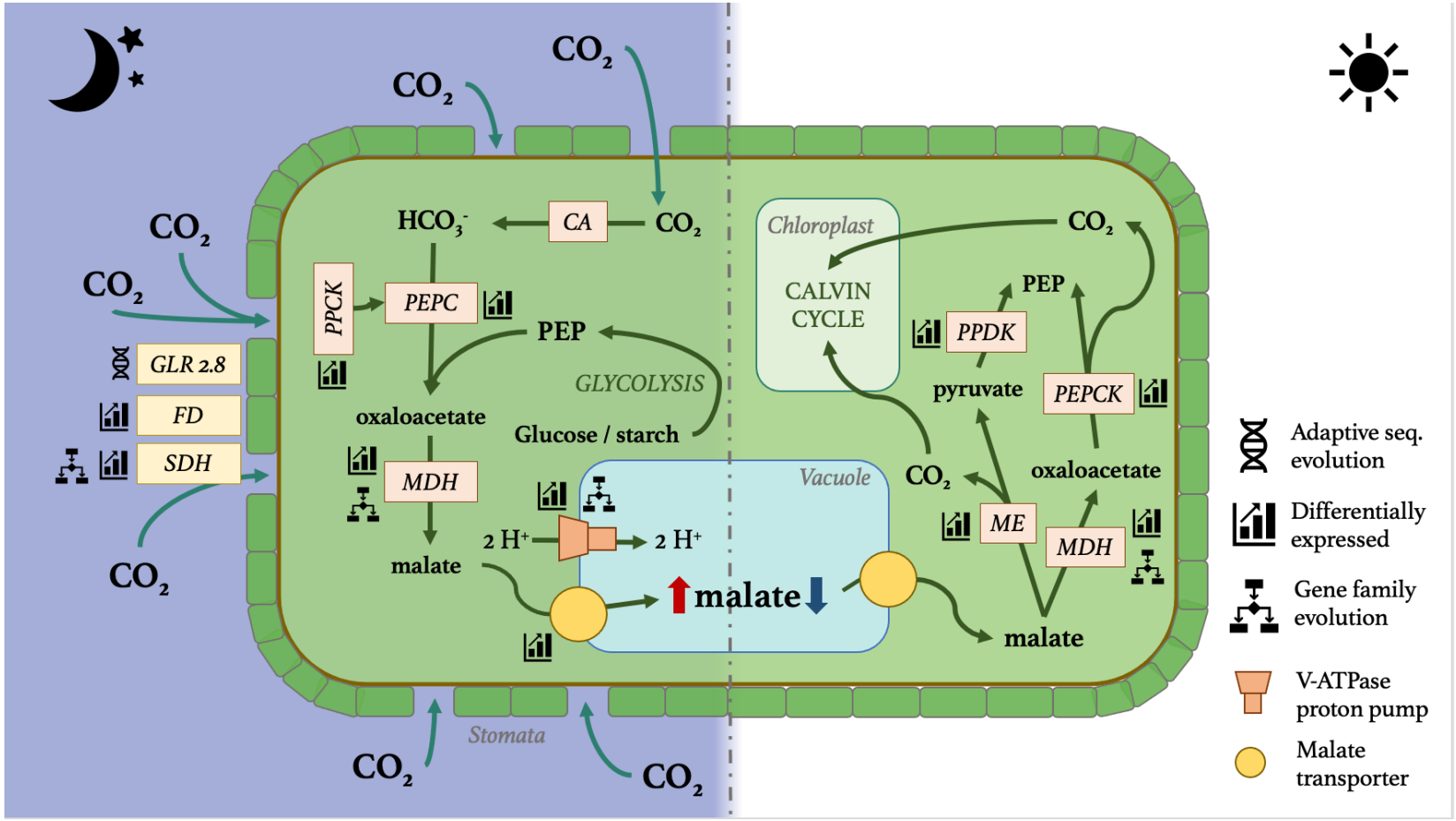
T. fasciculata

A. comosus

(Groot Crego *et al.* 2023 *BioRxiv*)

Evolution of genes with CAM/C3 functional relevance

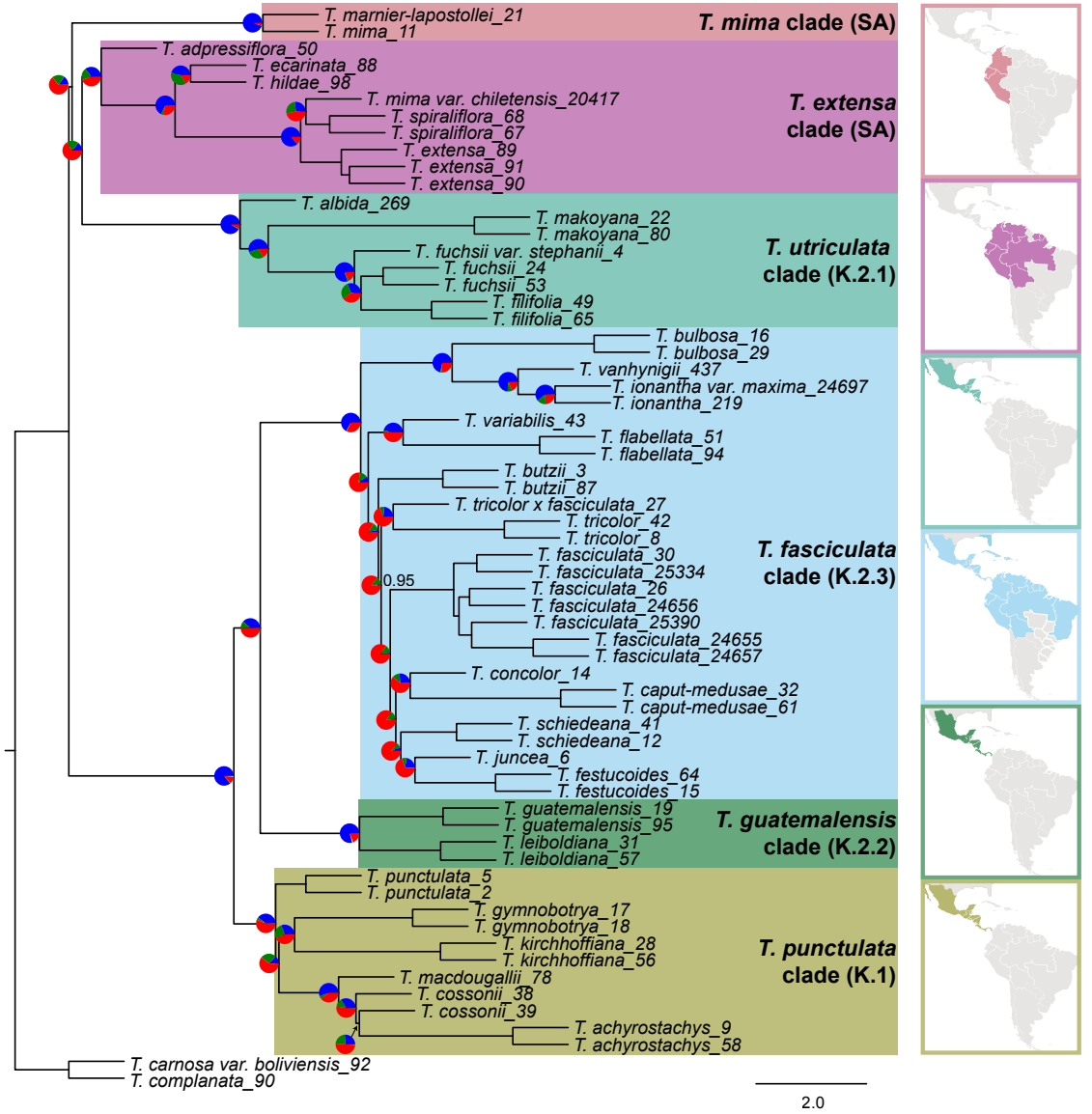
Clara Groot Crego



Studying *Tillandsia* subg *Tillandsia*'s evolutionary history

Whole-genome re-sequencing

Gil Yardeni



Coalescent-based species trees generated with ASTRAL-III. Posterior probabilities are 1 unless noted otherwise. Pie charts at the nodes show levels of gene tree discordance.

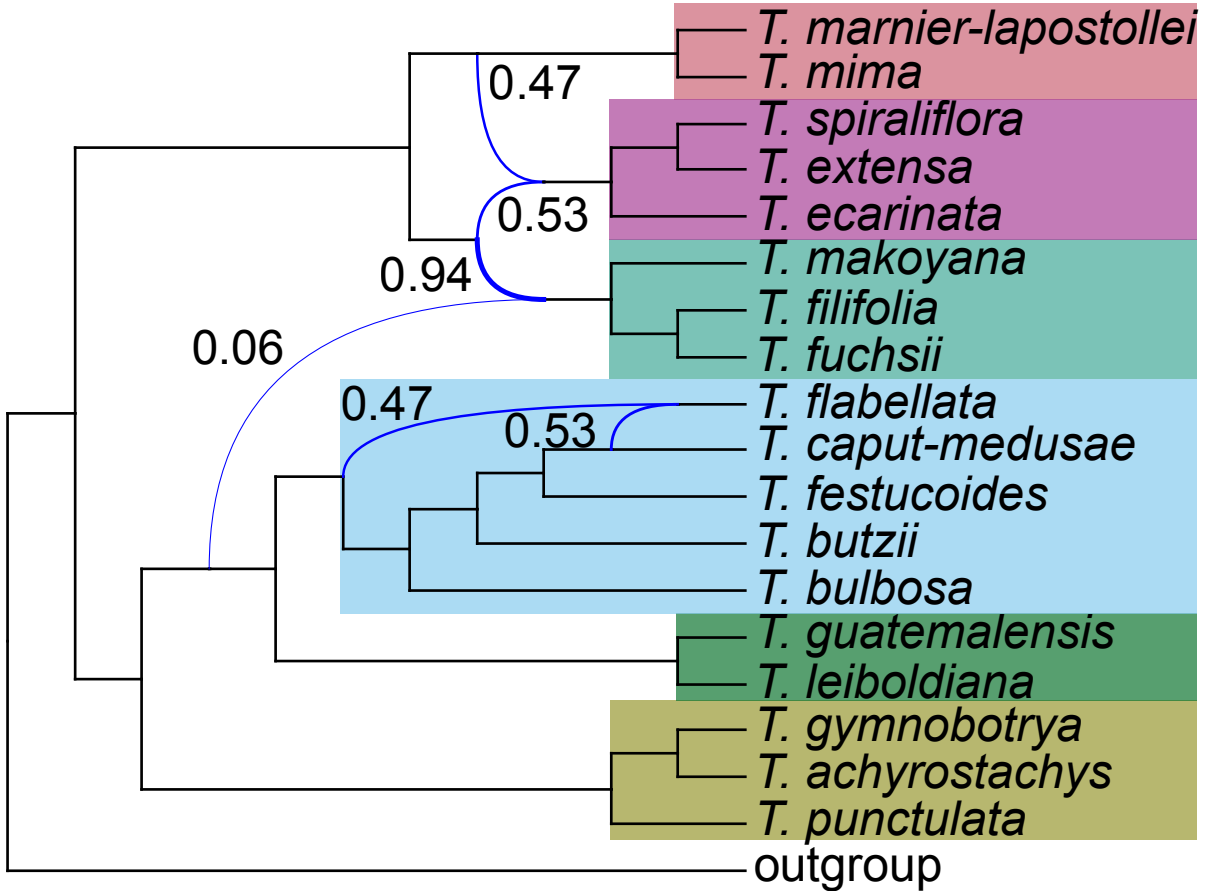
At least 2 dispersal events to C. America!

(Yardeni et al. 2023 *BioRxiv*)

Studying *Tillandsia* subg *Tillandsia*'s evolutionary history

Whole-genome re-sequencing

Gil Yardeni



Species network constructed with PhyloNet.

At least 2 dispersal events to C. America.

Complex evolutionary history.

(Yardeni et al. 2023 *BioRxiv*)