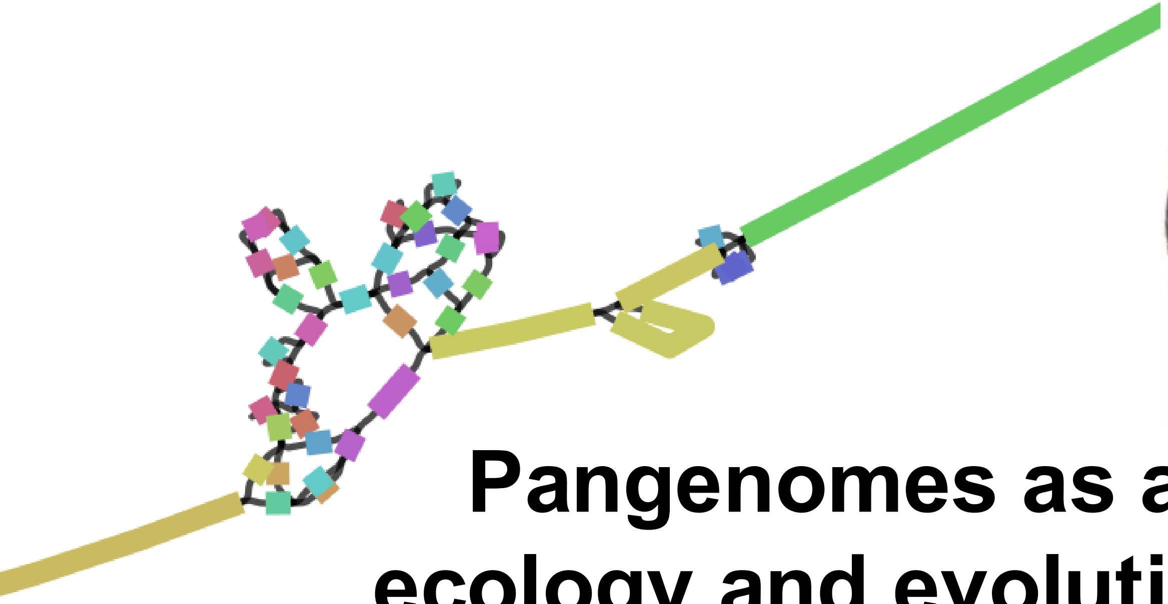
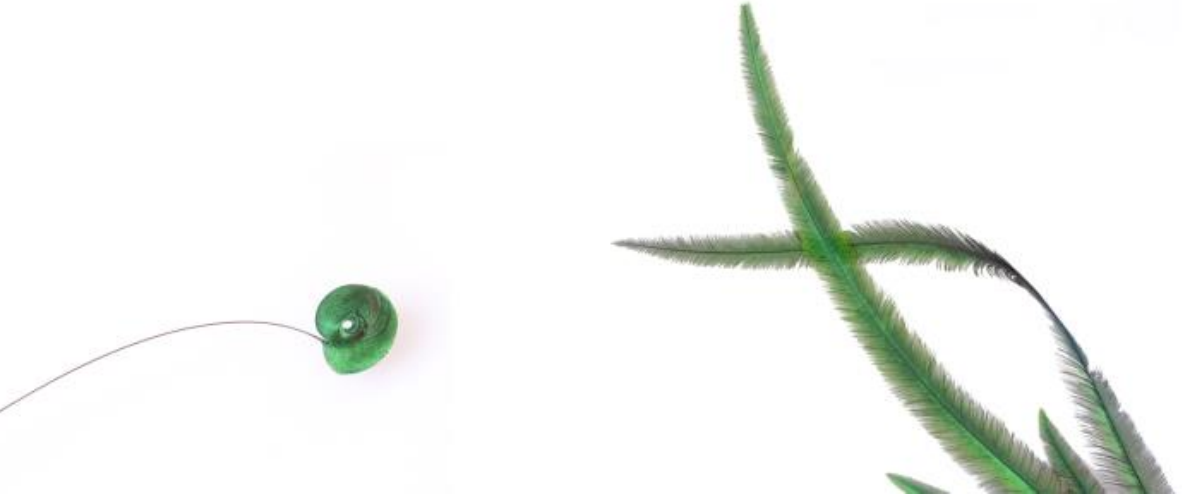
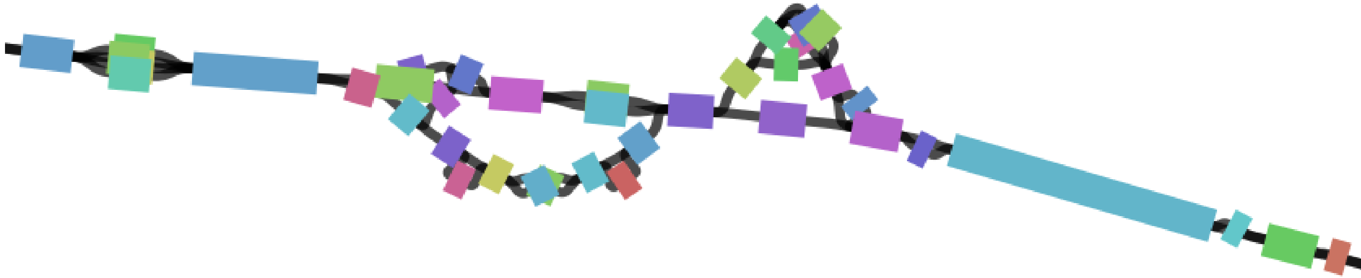
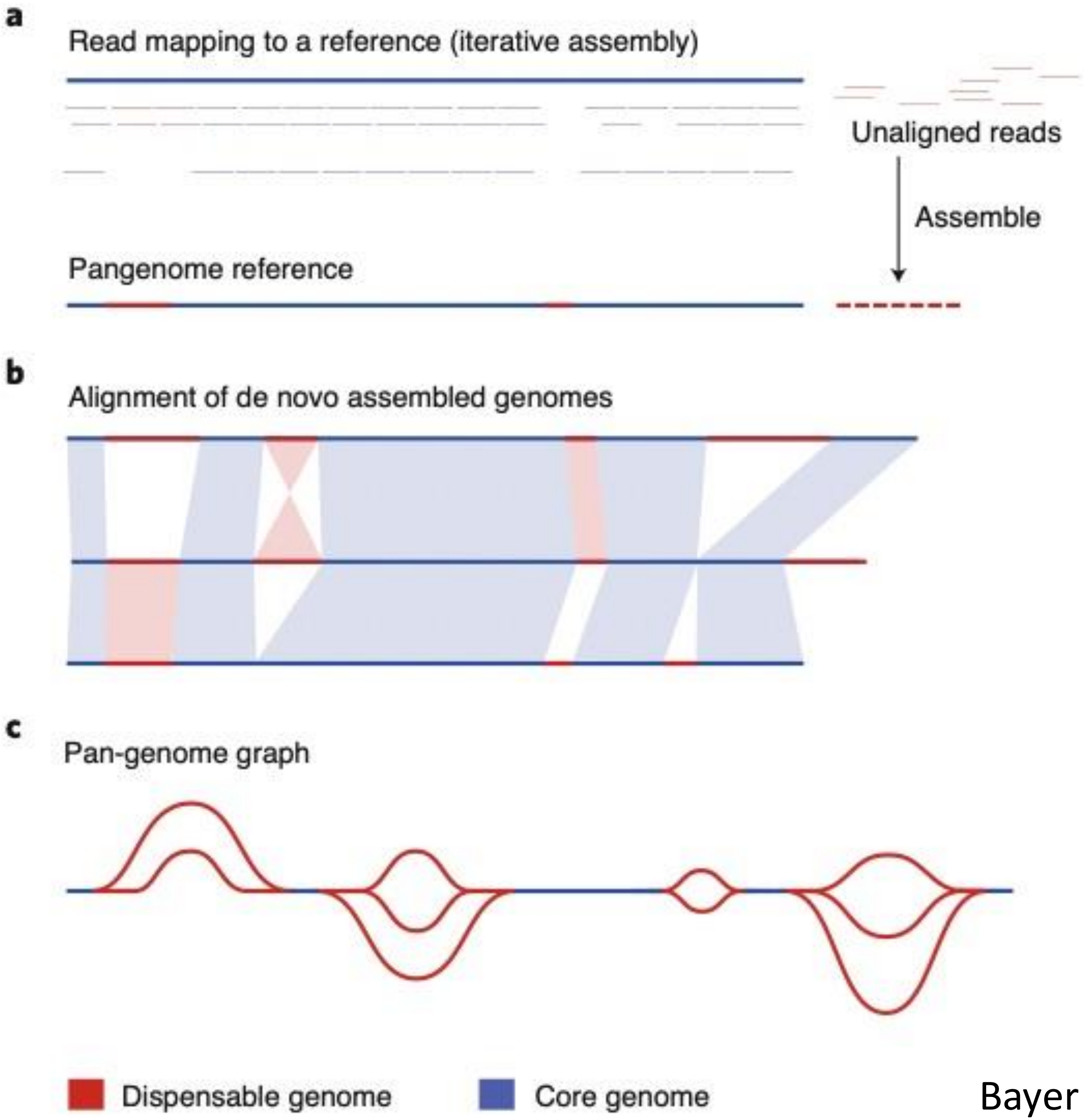



Pangenomes as a new tool for studying ecology and evolution of natural populations

Scott V. Edwards
Museum of Comparative Zoology, Harvard University, Cambridge, USA

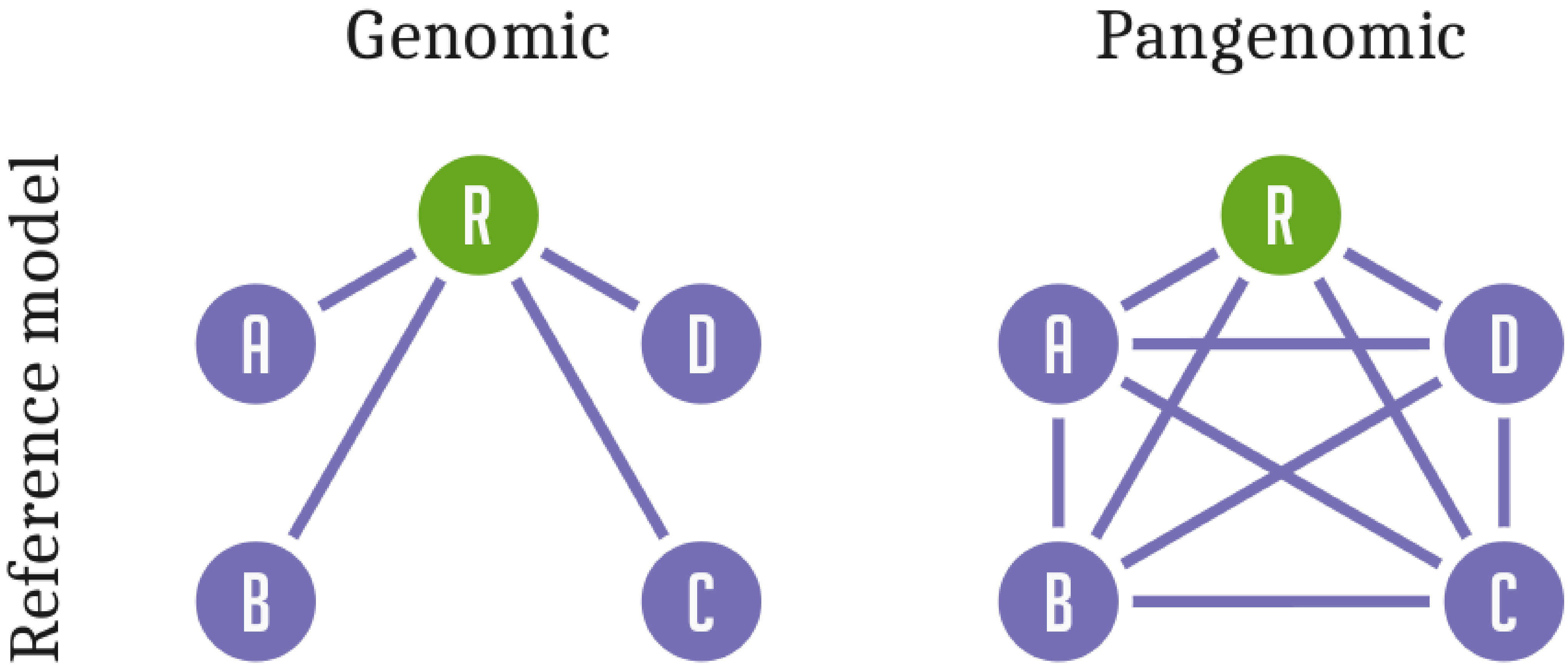



Pangenomes: moving beyond reference-based genomics



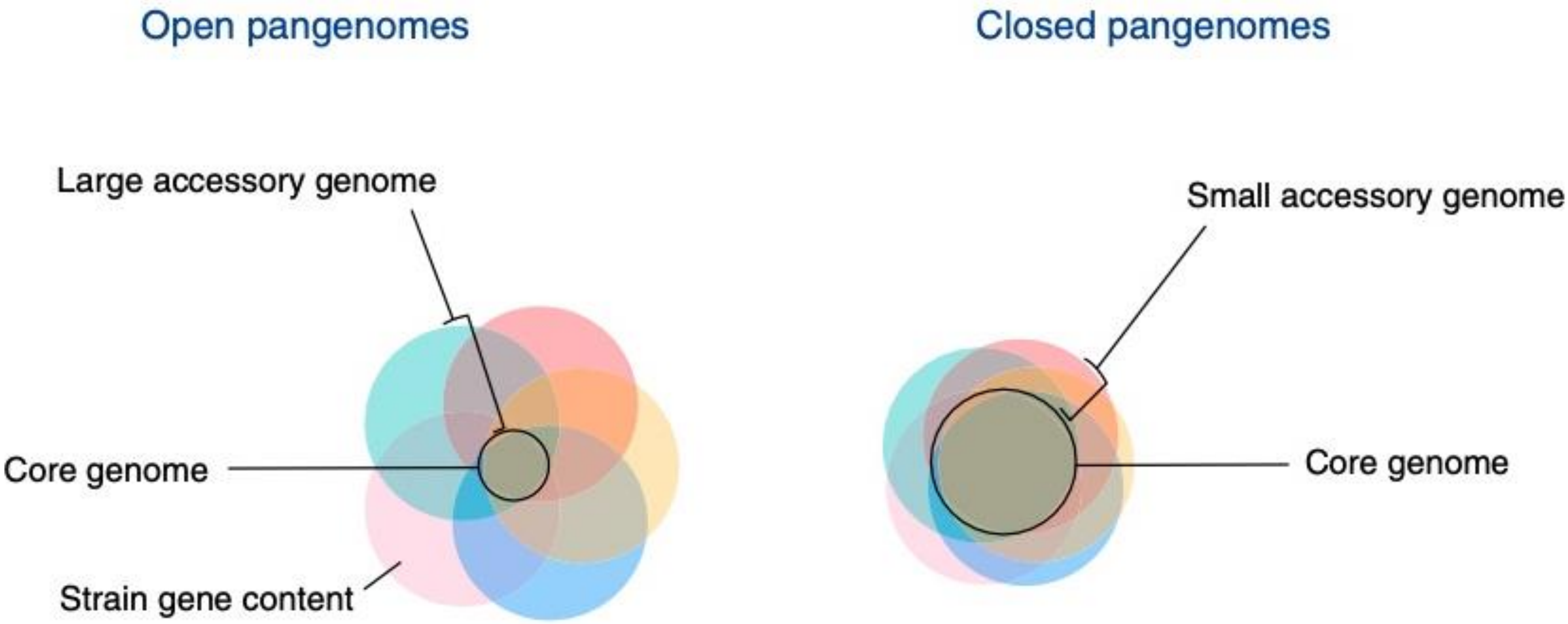
Bayer et al. 2020. *Nature Plants* 6: 914-920.

Reference-free genomics



Eizenga et al. 2021. *Ann. Rev. Genomics and Human Genetics*

Open and closed pangenomes



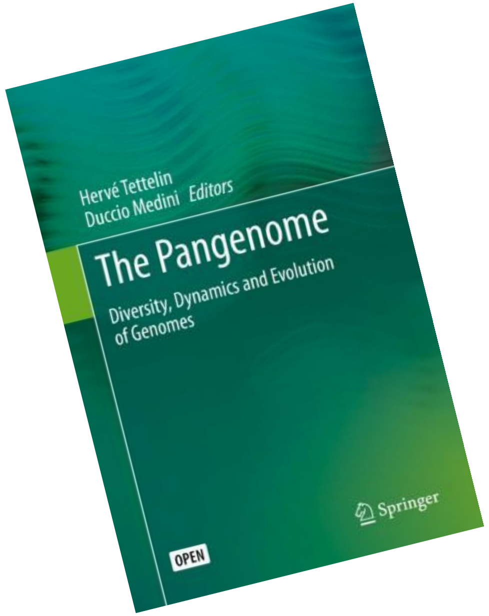
Open pangenomes

Closed pangenomes

Brokhurst et al. 2019. *Curr. Biol.*

The eukaryotic pangenome

- “The existence of pangenomes in eukaryotes is debated…Pangenome studies in eukaryotes are challenging due to their more complex genome and architectures and a lack of replete genome-level sampling” (Brockhurst et al. 2019. *Current Biology)*



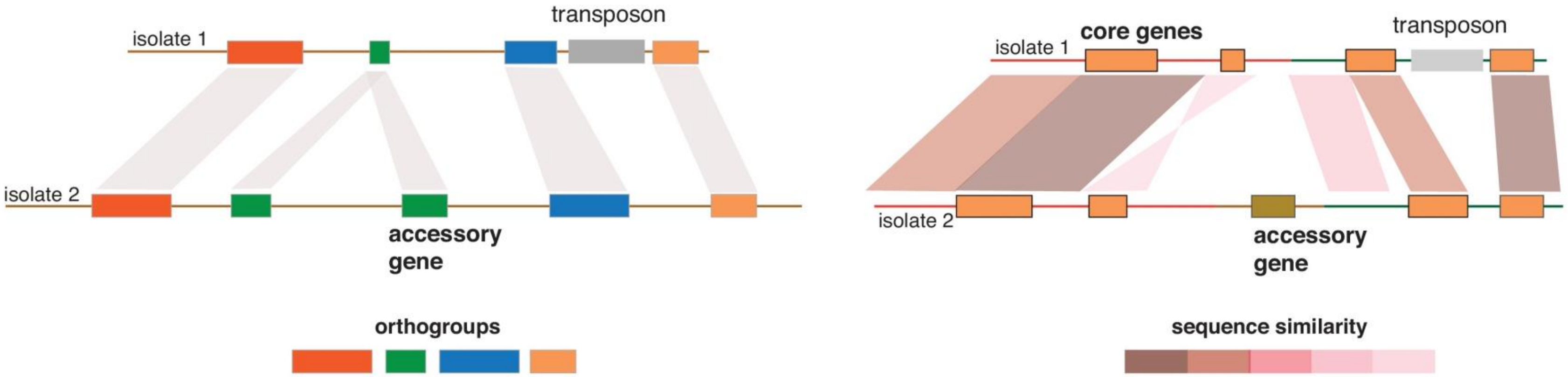
Hervé Tettelin
Duccio Medini *Editors*

The Pangenome

Diversity, Dynamics and Evolution of Genomes

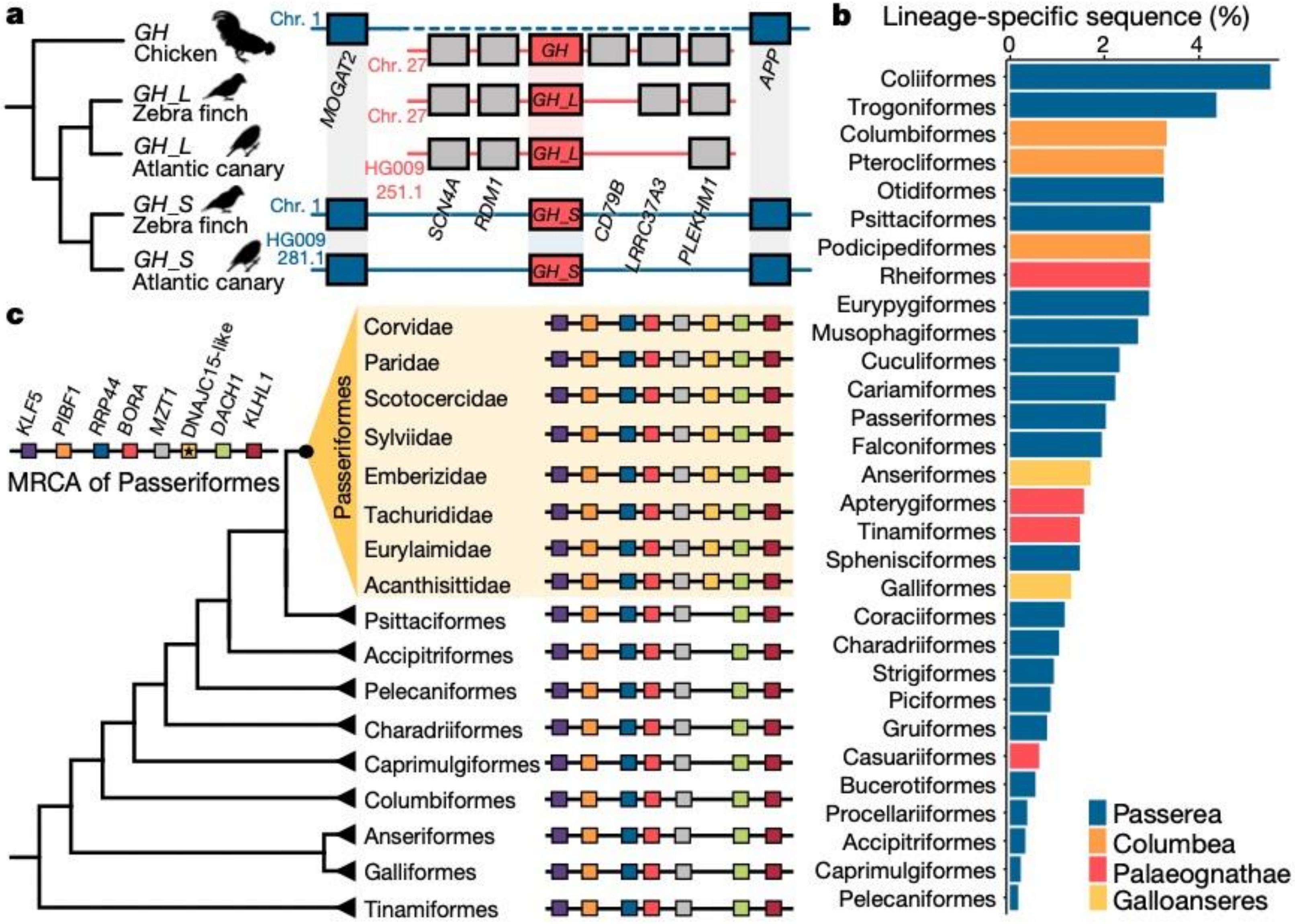
Springer

OPEN



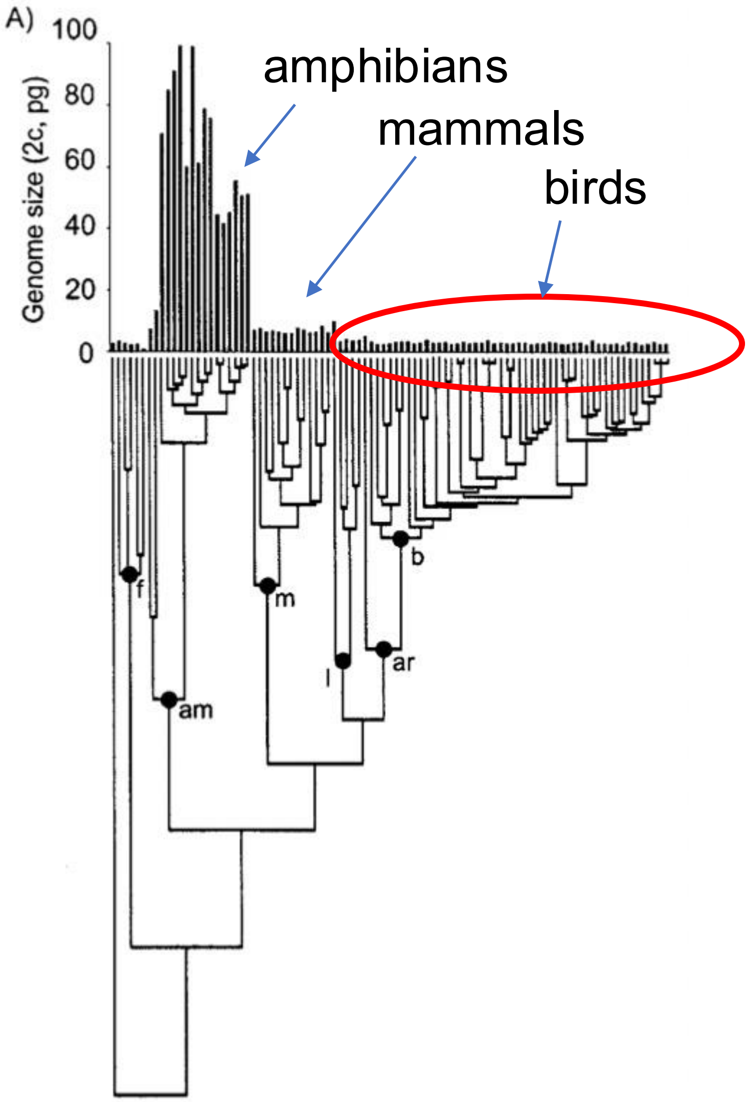
https://pathogen-genomics.org/research/

Pangenome approach to comparative genomics

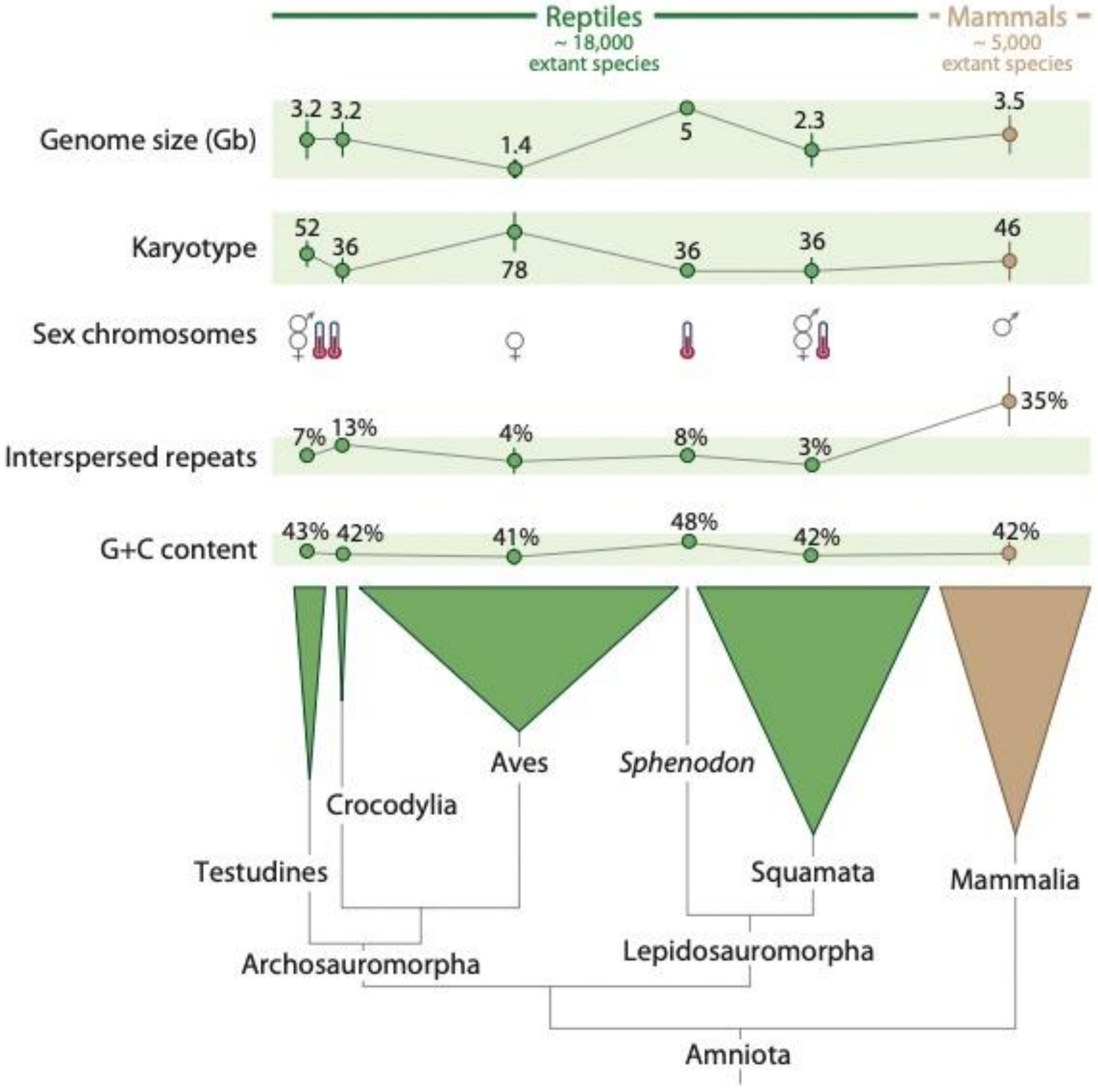


Feng et al. 2020. *Nature* 587:252-257.

Birds have small, streamlined genomes

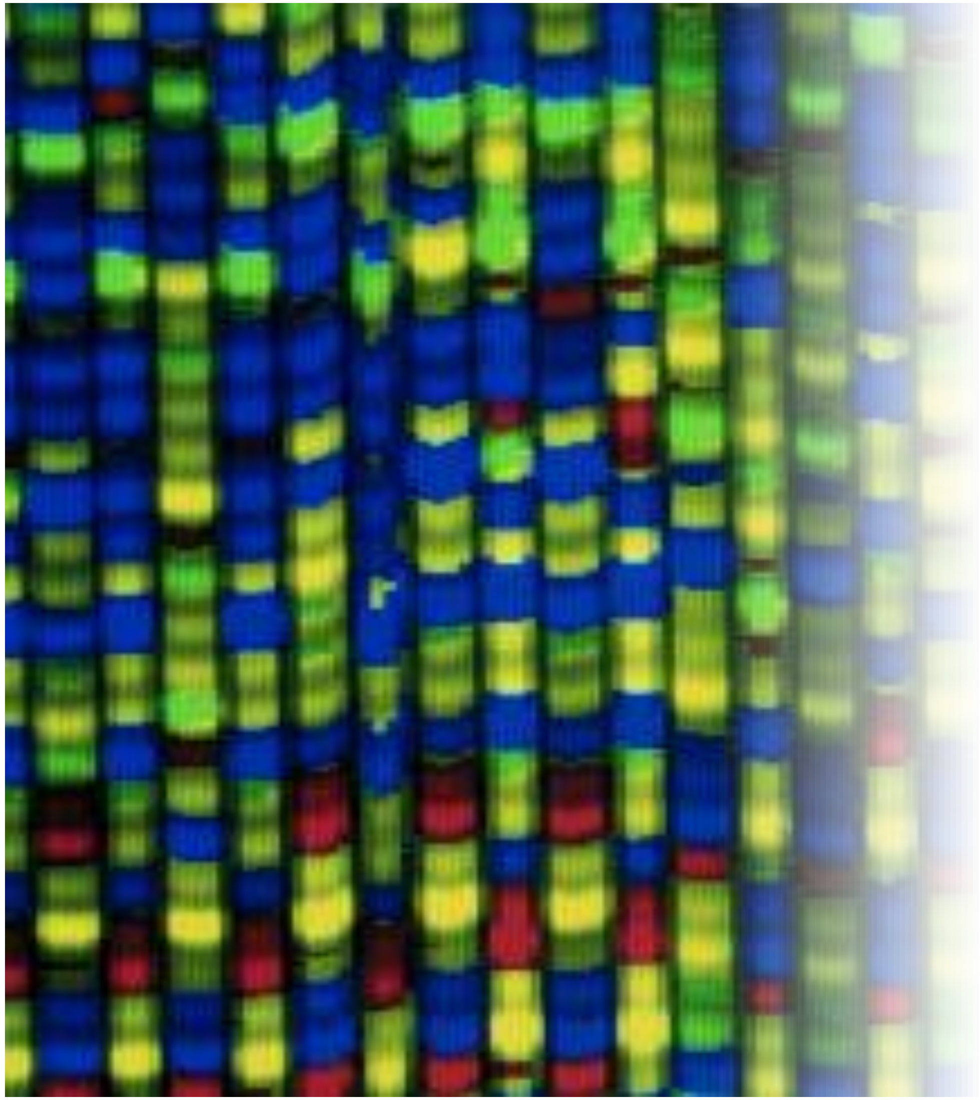


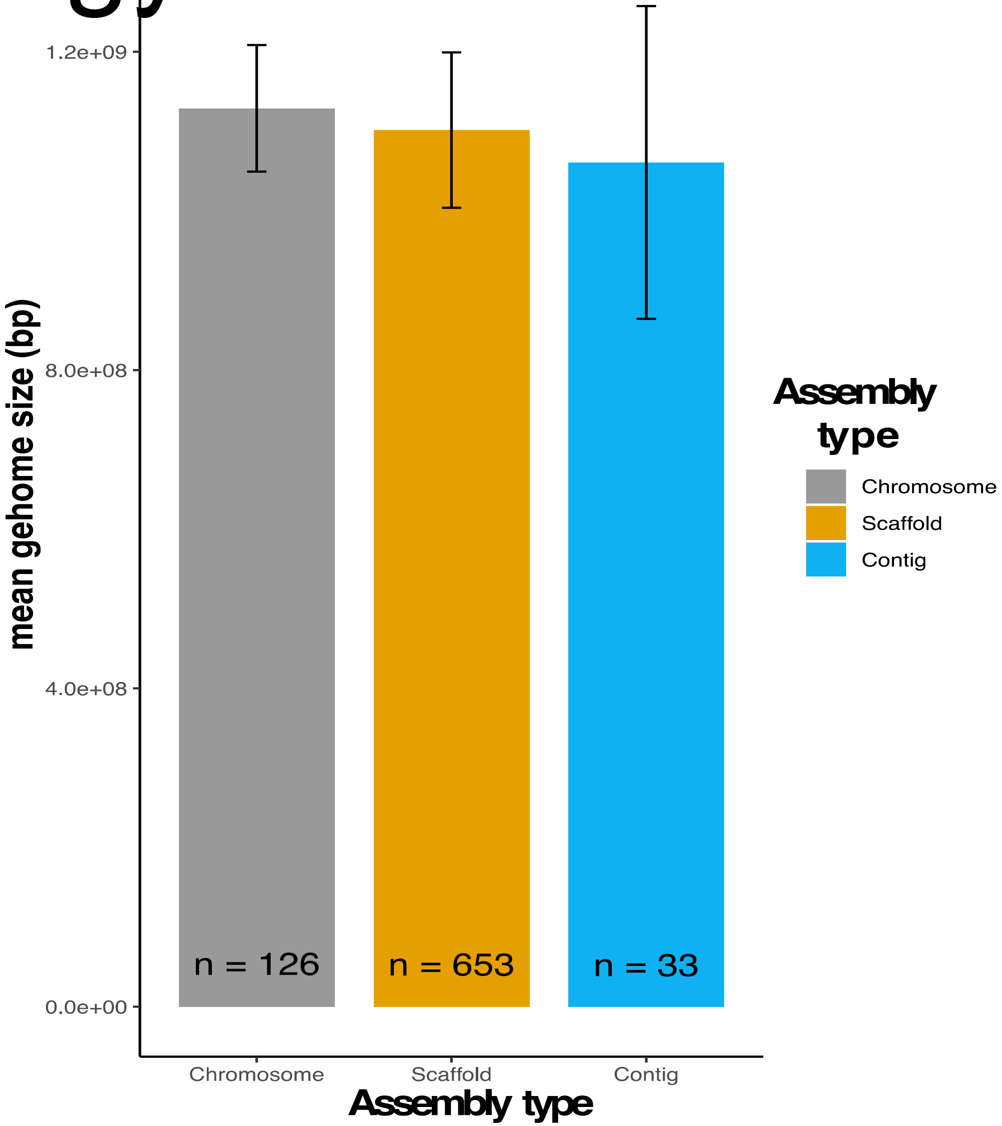
Waltari & Edwards. 2002. *Am. Nat.*



Organ et al. 2010. *Ann. Rev. Genom. Hum. Genet.*

Avian genomes are growing with each new technology





Data from NCBI, accessed 13 Nov. 2021

Scrub Jay Pangenome Consortium

Colorado team - Island Scrub Jay
Chris Funk
Rebecca Cheek
Paul Hohenlohe
Cameron Ghalambor

Fieldwork
Greg and Donna Schmitt

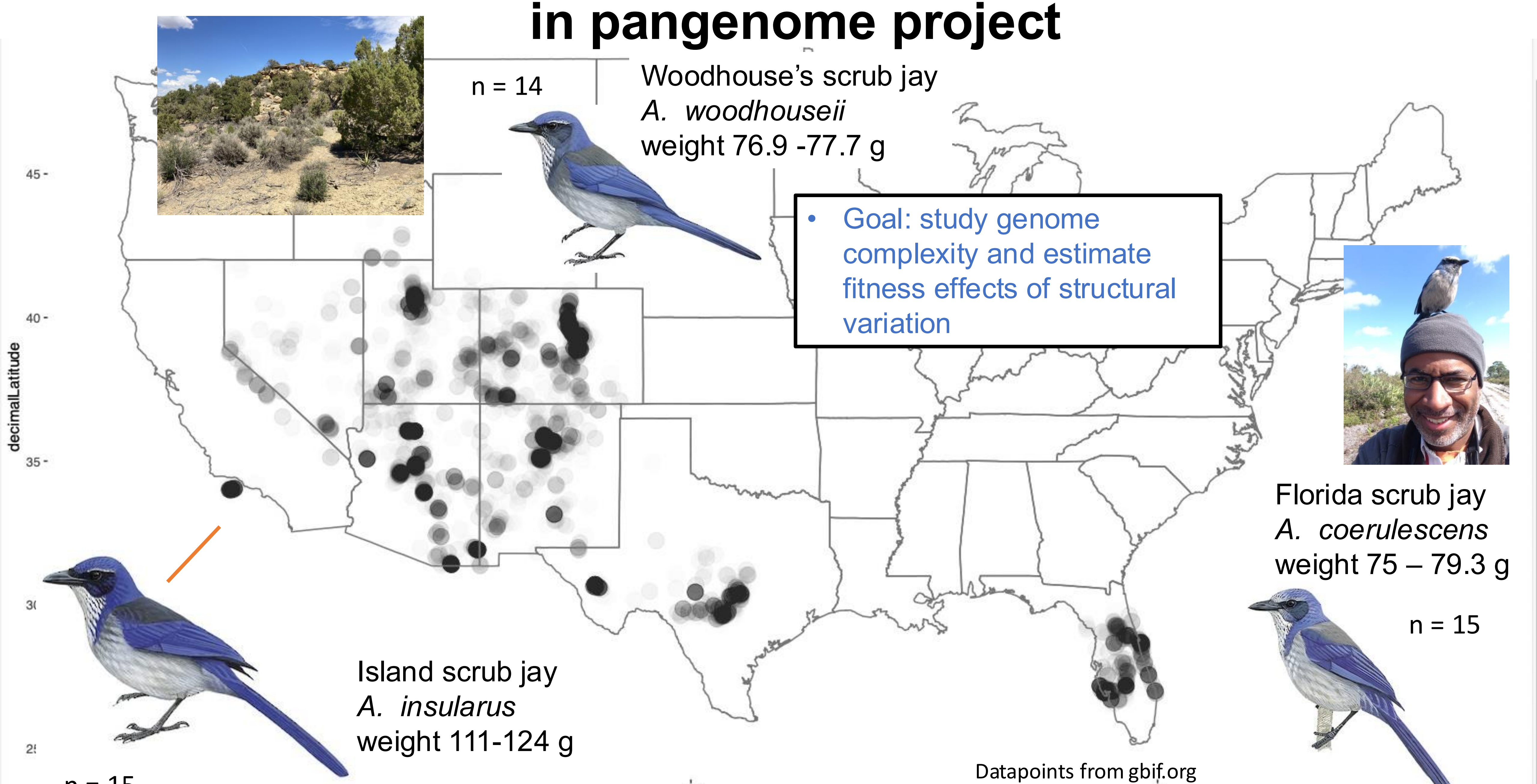
Florida team - Florida Scrub Jay
Nancy Chen
Reed Bowman
John Fitzpatrick

Pangenome informatics
Erik Garrison
Andrea Guarracino

Harvard team - Woodhouse's Scrub Jay and Informatics
Tim Sackton
Bohao Fang
George Kolyfetis
Danielle Khost
Heng Li

San Juan County, New Mexico

Three scrub-jay (*Aphelocoma*) species in pangenome project



n = 14

Woodhouse’s scrub jay
A. woodhouseii
weight 76.9 -77.7 g

- Goal: study genome complexity and estimate fitness effects of structural variation

Florida scrub jay
A. coerulescens
weight 75 – 79.3 g

n = 15

Island scrub jay
A. insularus
weight 111-124 g

n = 15

Datapoints from gbif.org

The Evolution of Comparative Phylogeography: Putting the Geography (and More) into Comparative Population Genomics

Scott V. Edwards[1,2,*], V. V. Robin[3], Nuno Ferrand[4], and Craig Moritz[5]

GBE

Table 1

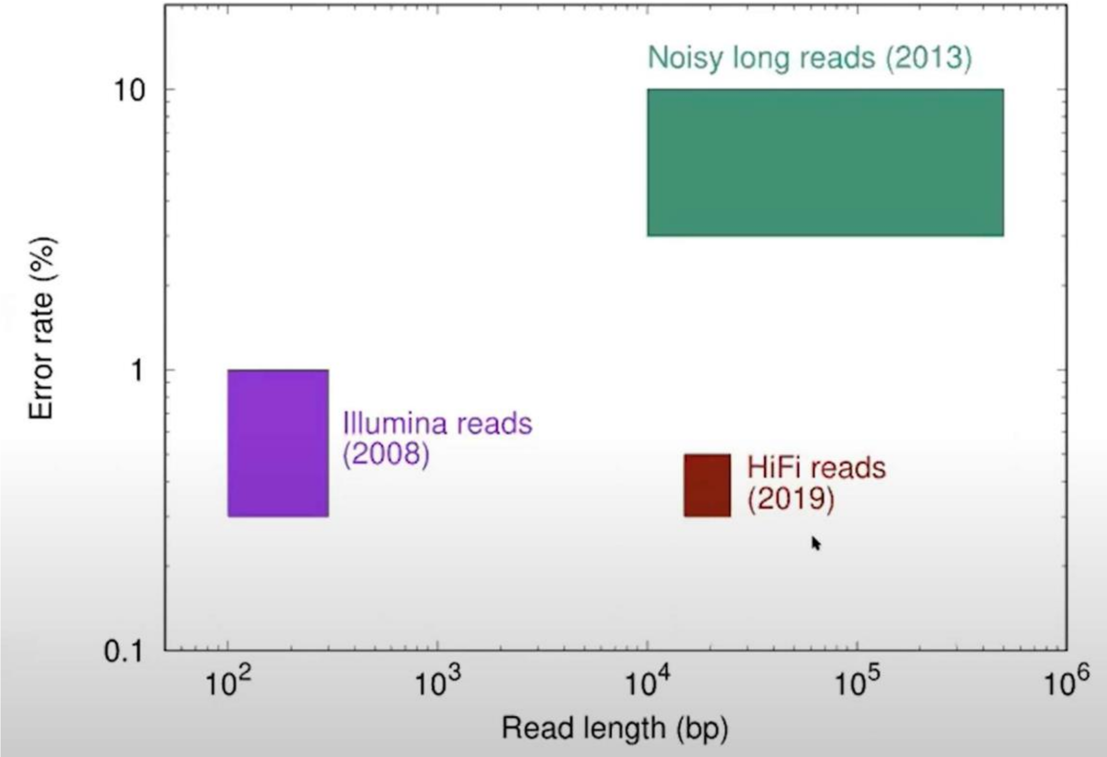
Conceptual Relationships between the Fields of Comparative Population Genomics, Landscape Genomics, and Comparative Phylogeography

Concept/Parameter	Comparative Population Genomics	Landscape Genomics	Comparative Phylogeography
Comparative perspective	Growing	Nascent	Mature
Emphasis on space	No	Yes	Yes
Geographic scale	Random mating population	Region	Biome
Temporal scale	Arbitrary	Recent	Deep
Focus on:			
selection versus neutrality	Both	Both	Neutrality
recombination	Yes	Not yet considered	Not yet considered
geography versus environment	Nuisance parameters	Environment	Both
Future use of whole-genome sequencing	Yes	Likely	Unlikely
Growth out of museum collections community	No	No	Partial

Edwards et al. 2021. *Genome Biology and Evolution* 14: 10.1093/gbe/evab176

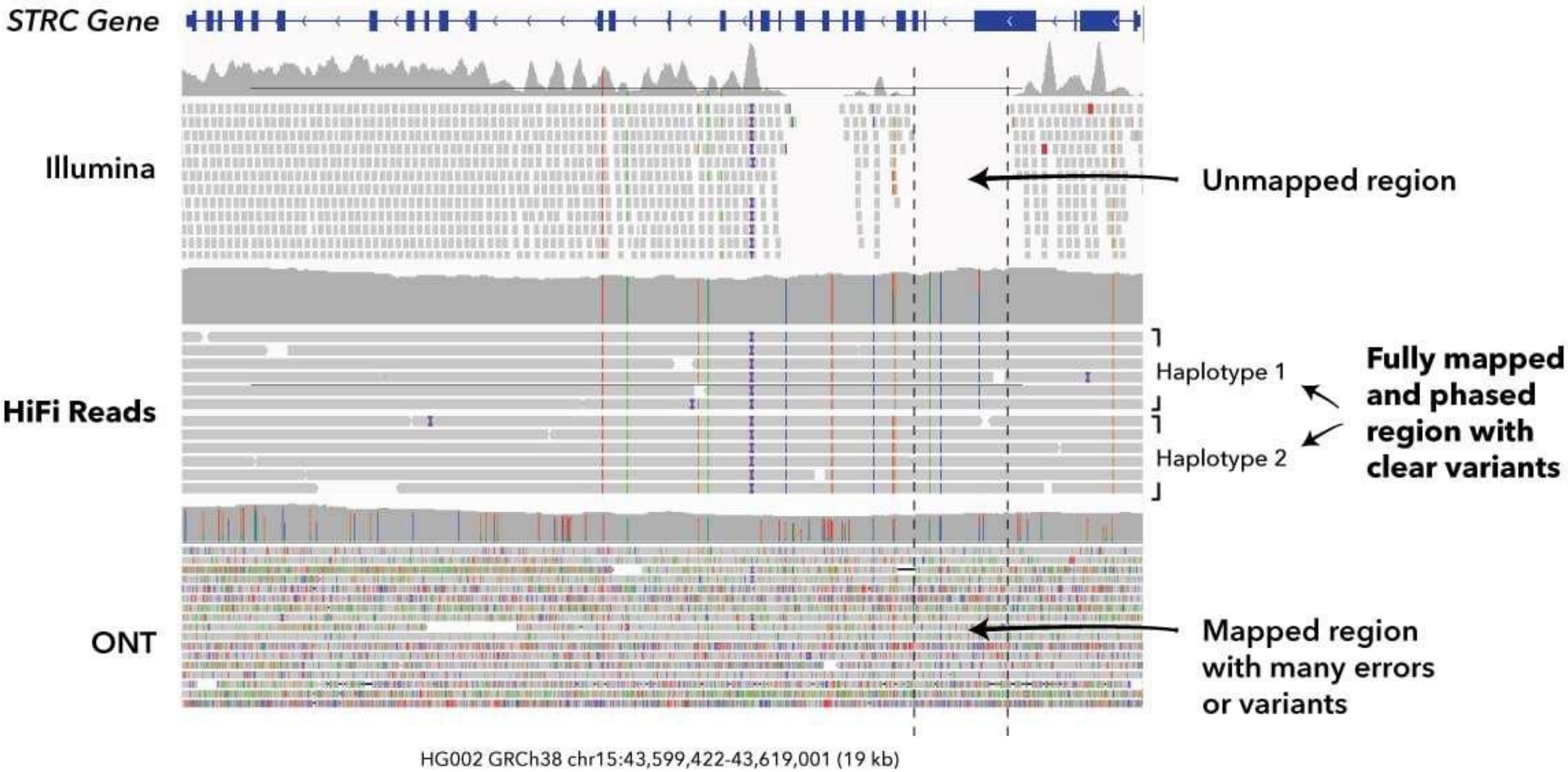
PacBio HiFi reads are long and accurate

- HiFi reads: long & accurate
- A breakthrough every ~5 years
- Most existing assemblers cannot make full use of the accuracy

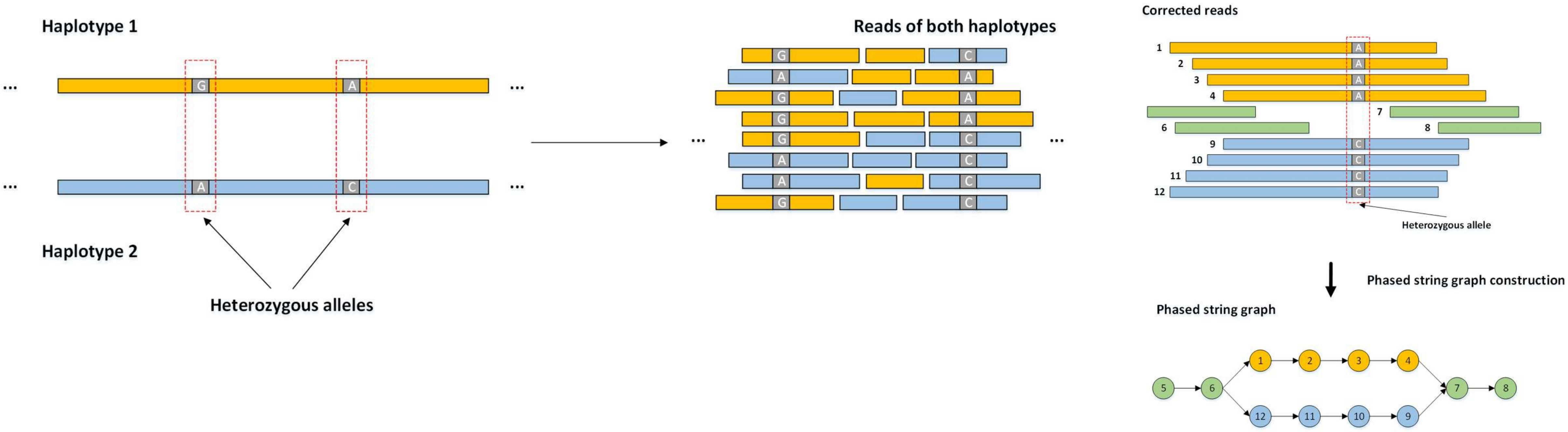


Coutesy Haoyu Cheng, Dana Farber Cancer Institute

PacBio HiFi reads are long <u>and</u> accurate



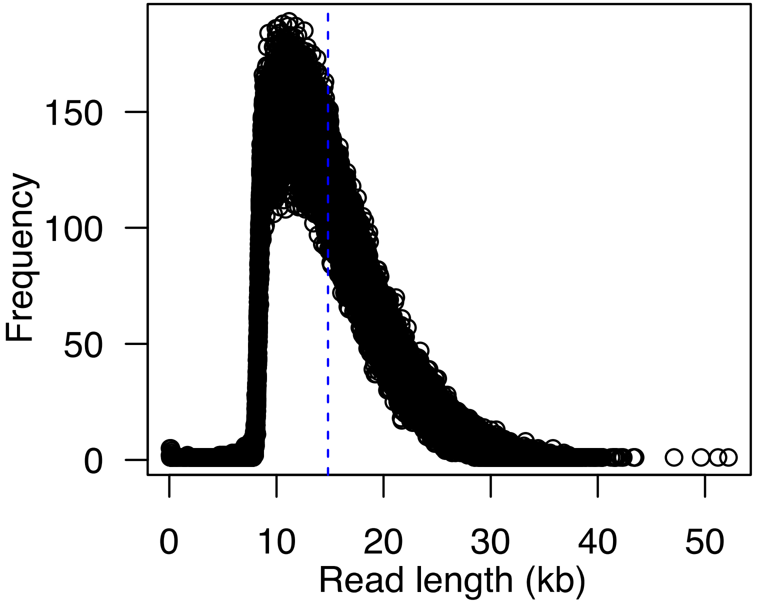
Hifiasm – a HiFi accurate read assembler that resolves haplotypes



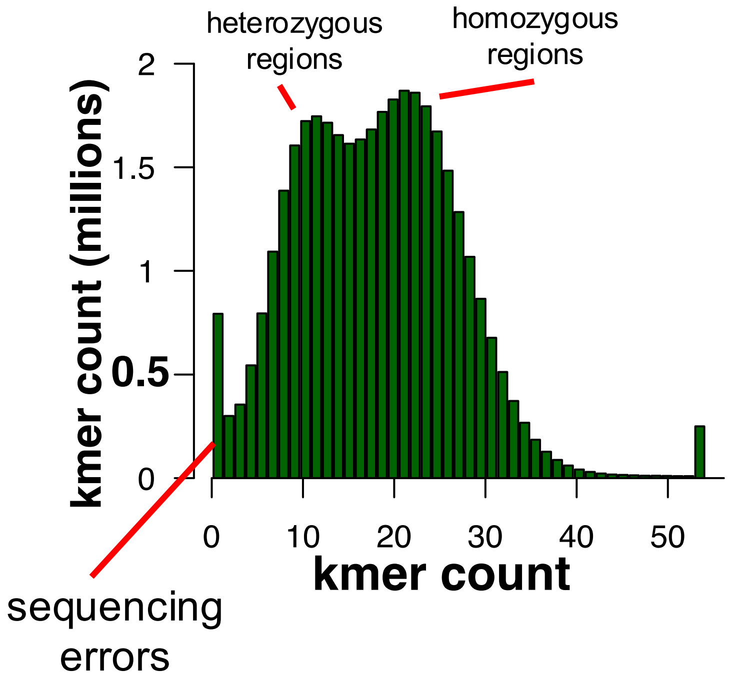
Coutesy Haoyu Cheng, Dana Farber Cancer Institute

Scrub-jay PacBio HiFi data characteristics

Average read length 14.8 kb



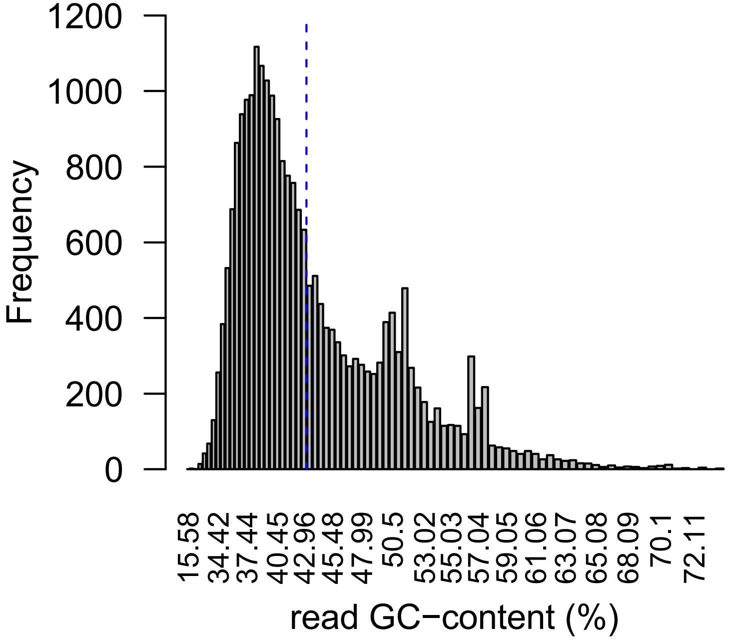
Average coverage ~40X


heterozygous regions

homozygous regions

sequencing errors

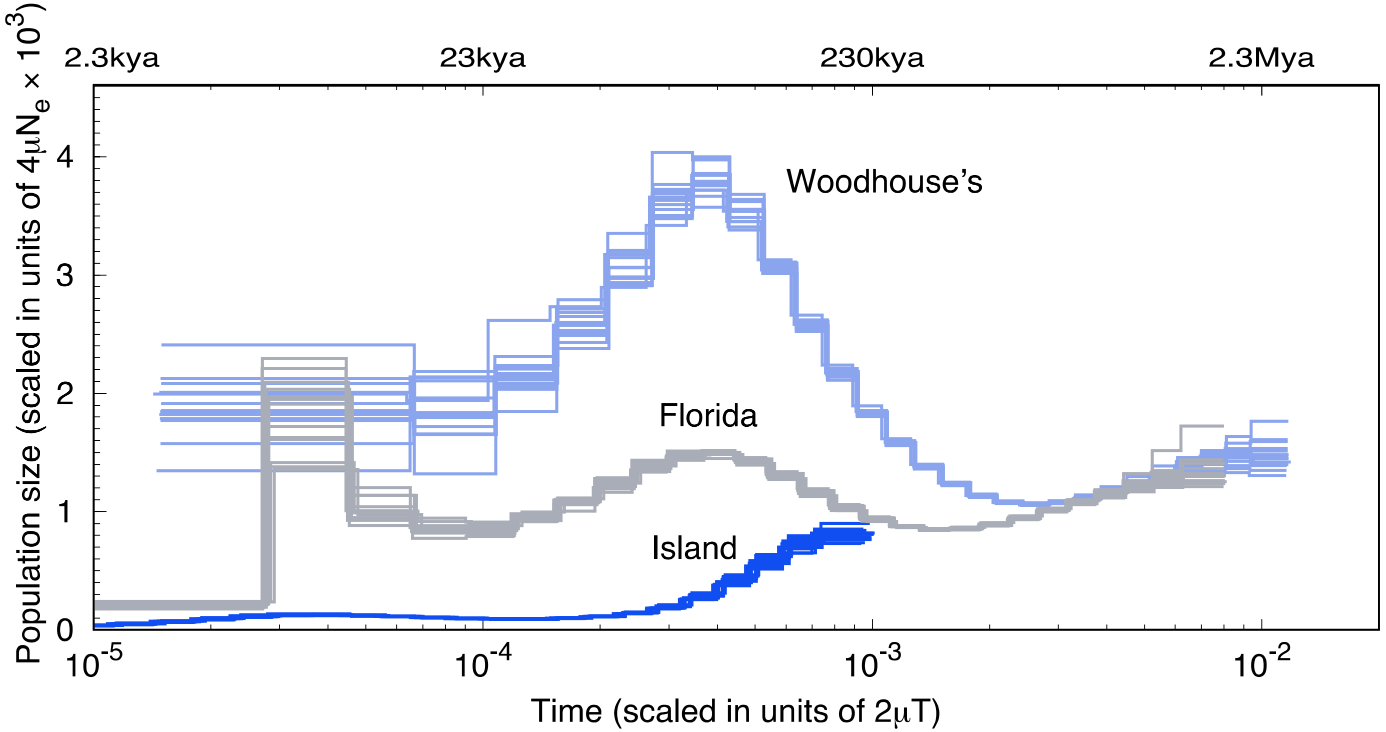
GC-content



60-fold range in effective population size across species

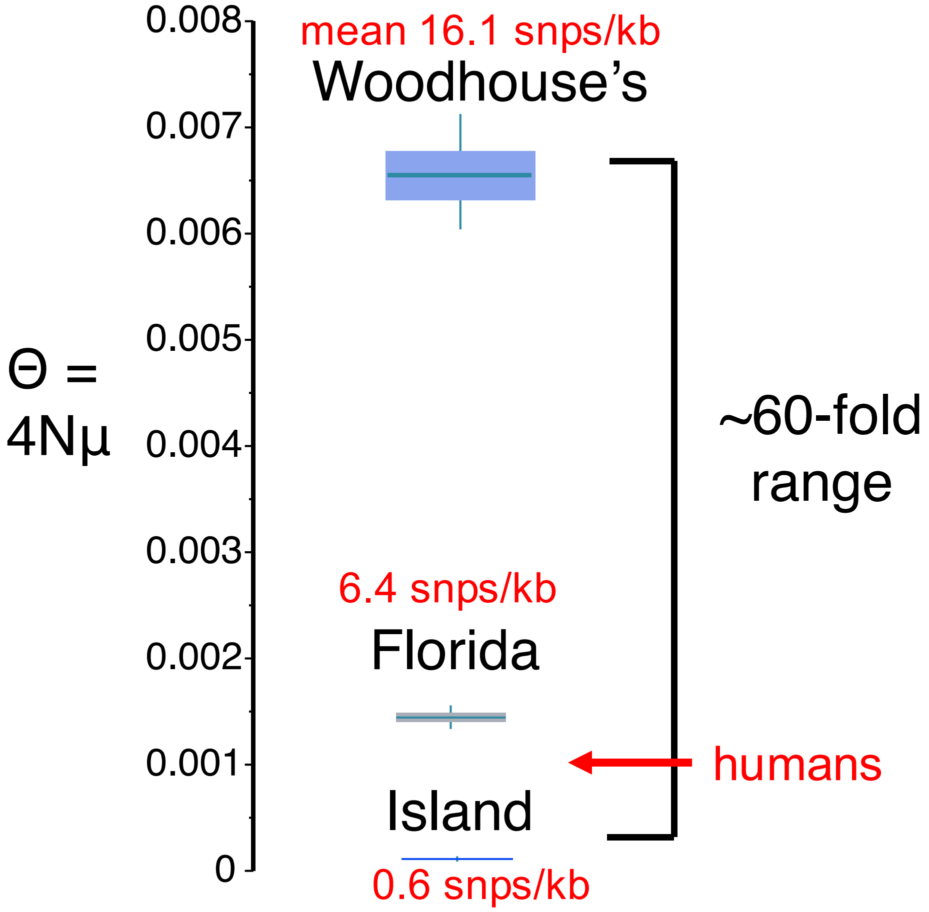
PSMC

SNP data from genome-wide dipcall regions



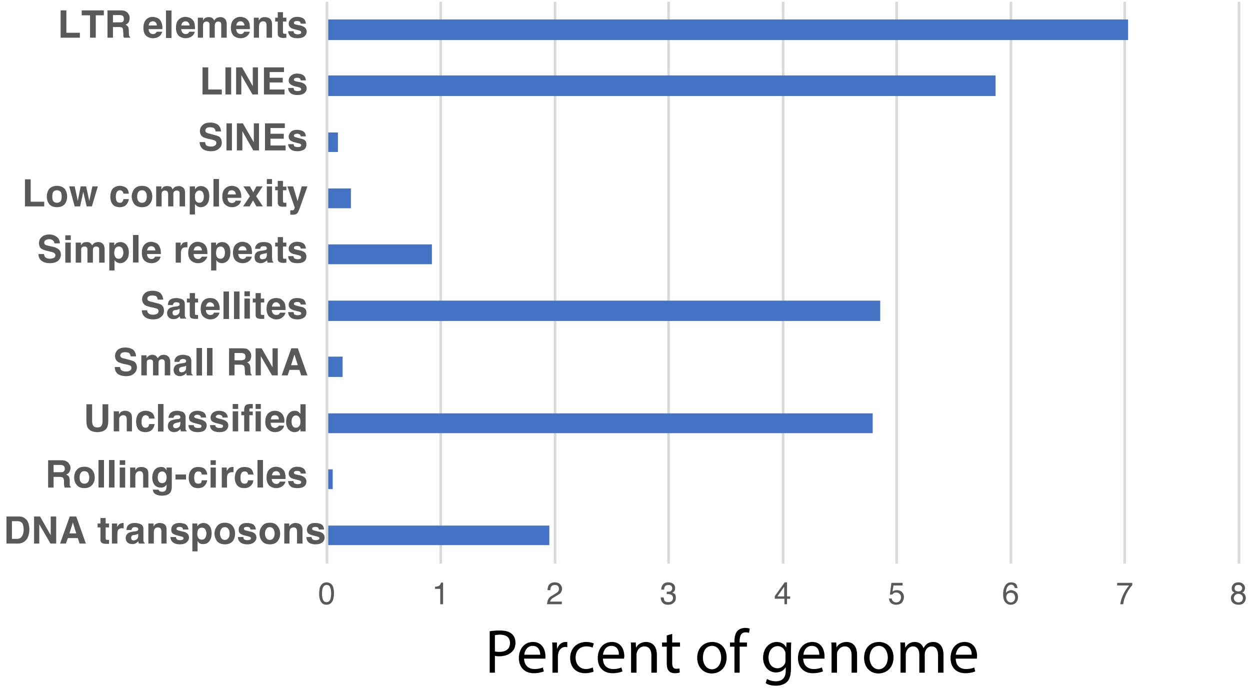
bpp

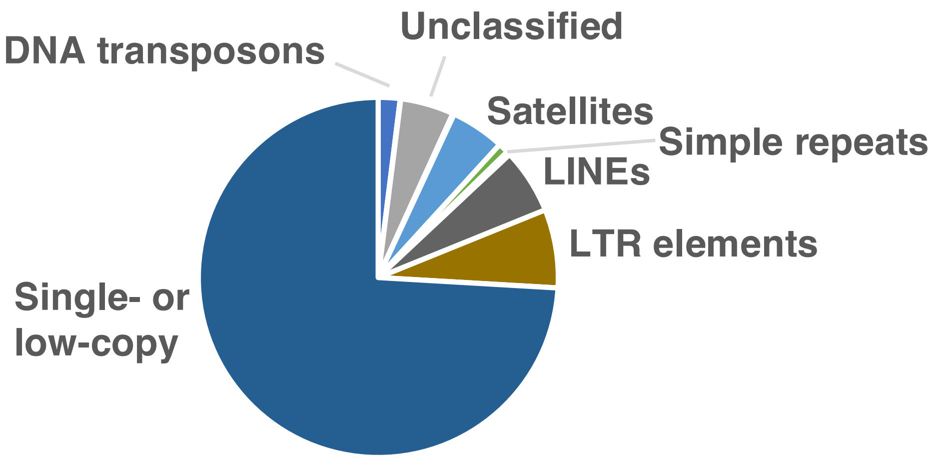
~2500 1-kb dipcall regions



Li & Durbin. 2011. *Nature* 2011, 475:493-496; Flouri et al. 2020. *Mol. Biol. Evol.* 37: 1211-1223

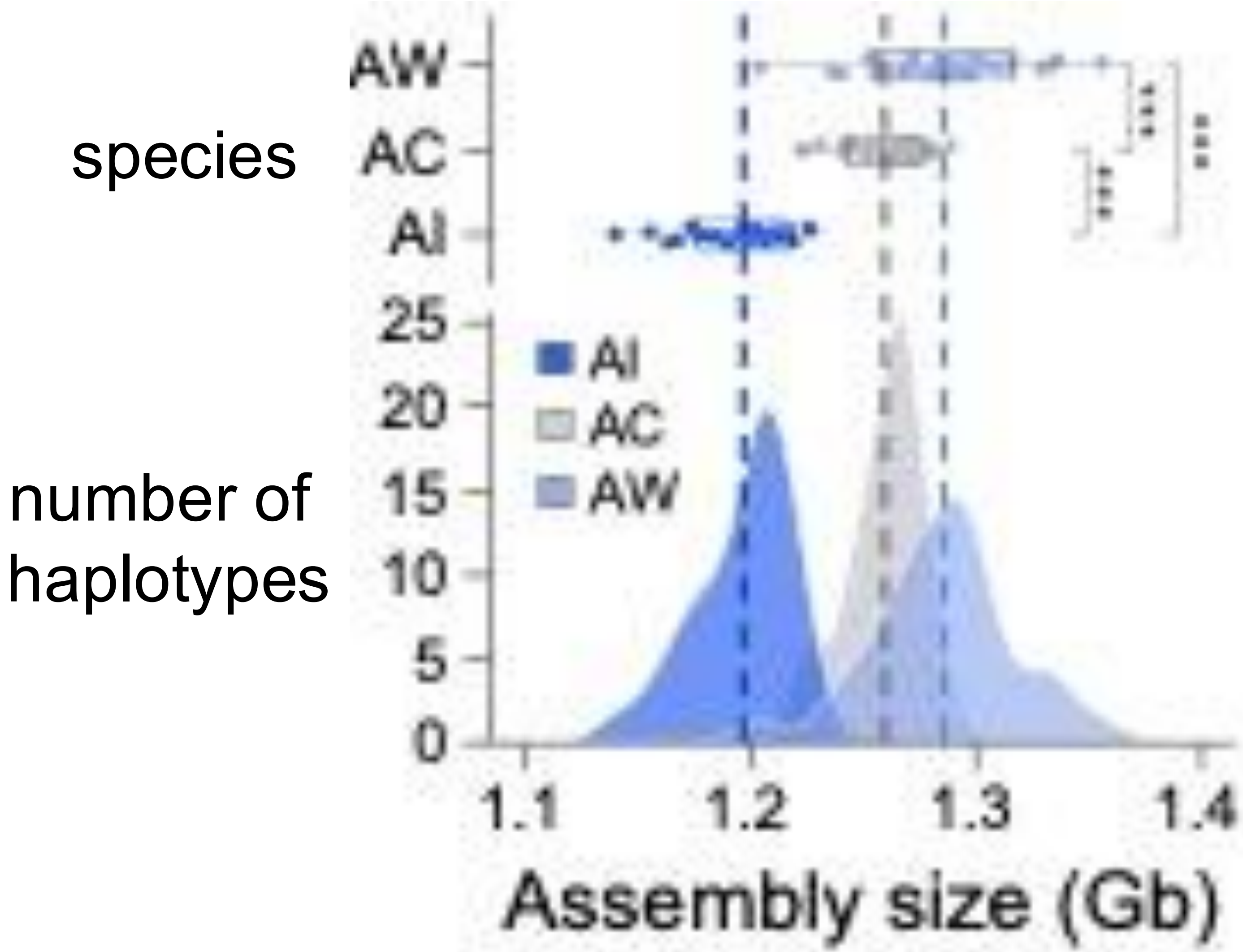
RepeatMasker analysis suggests over 25% repeats and transposable elements





Assemblies of Island Scrub Jays are ~100 Mb smaller than Woodhouse's Scrub Jay

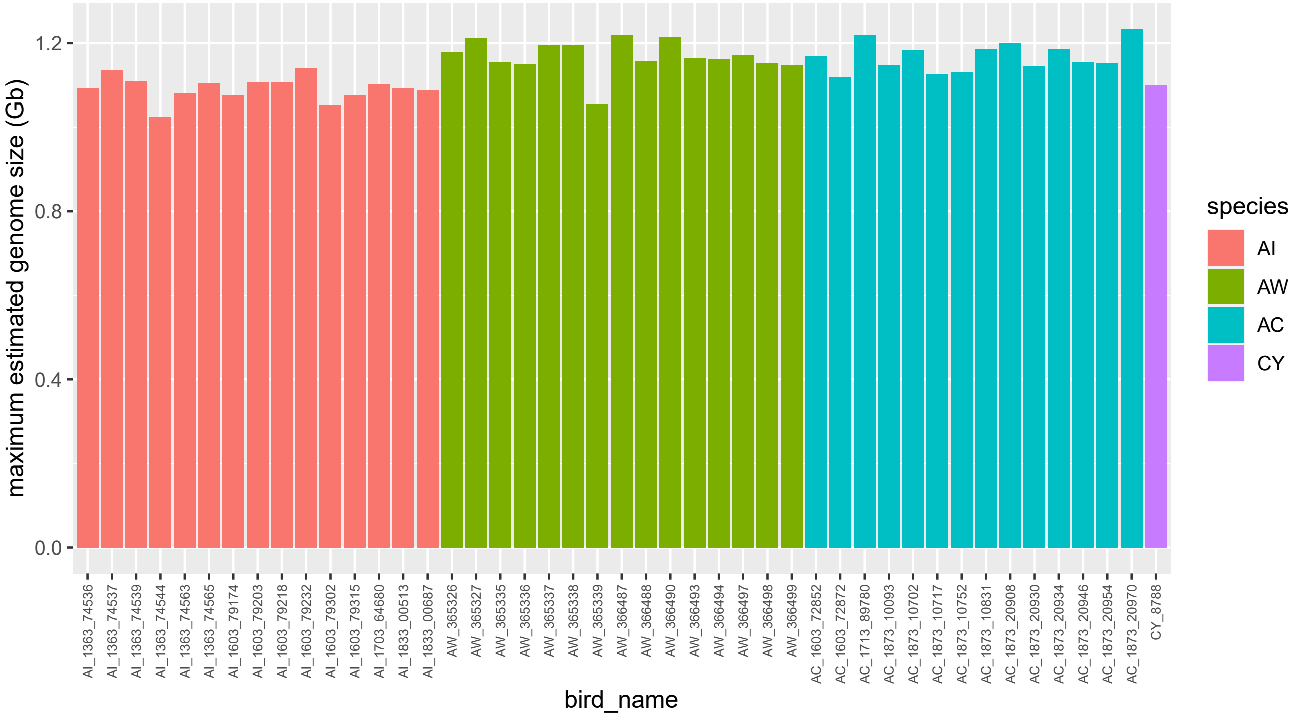
N = 30, 30 and 28 haplotypes (AI, AW, AC)

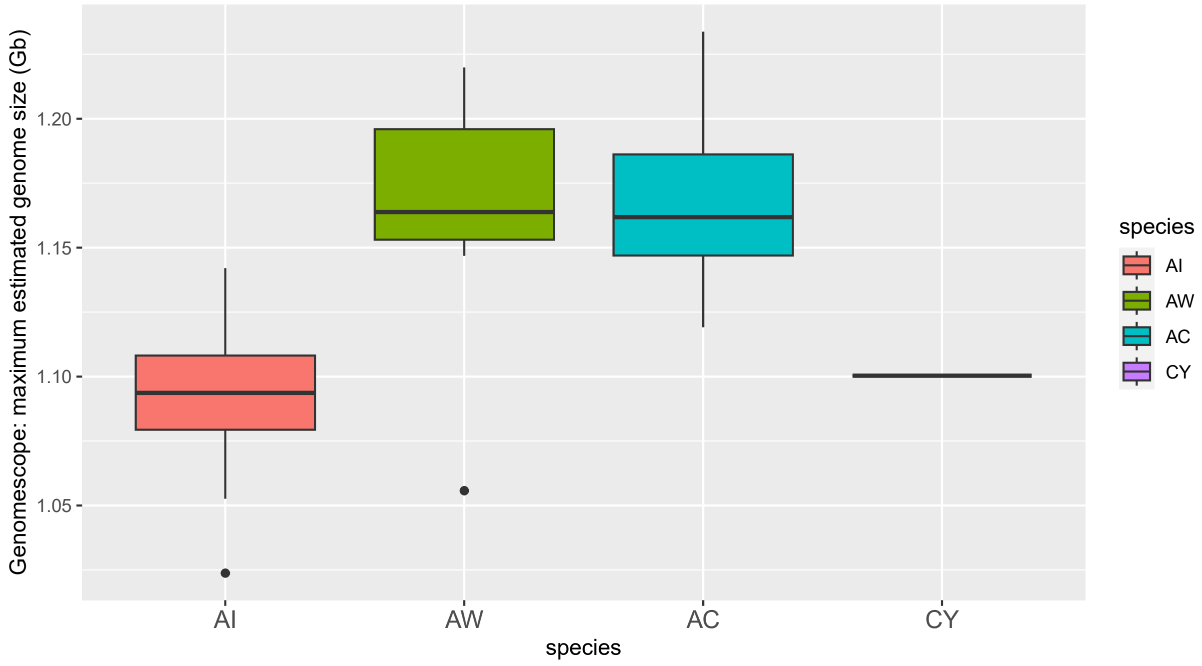


---- mean within species

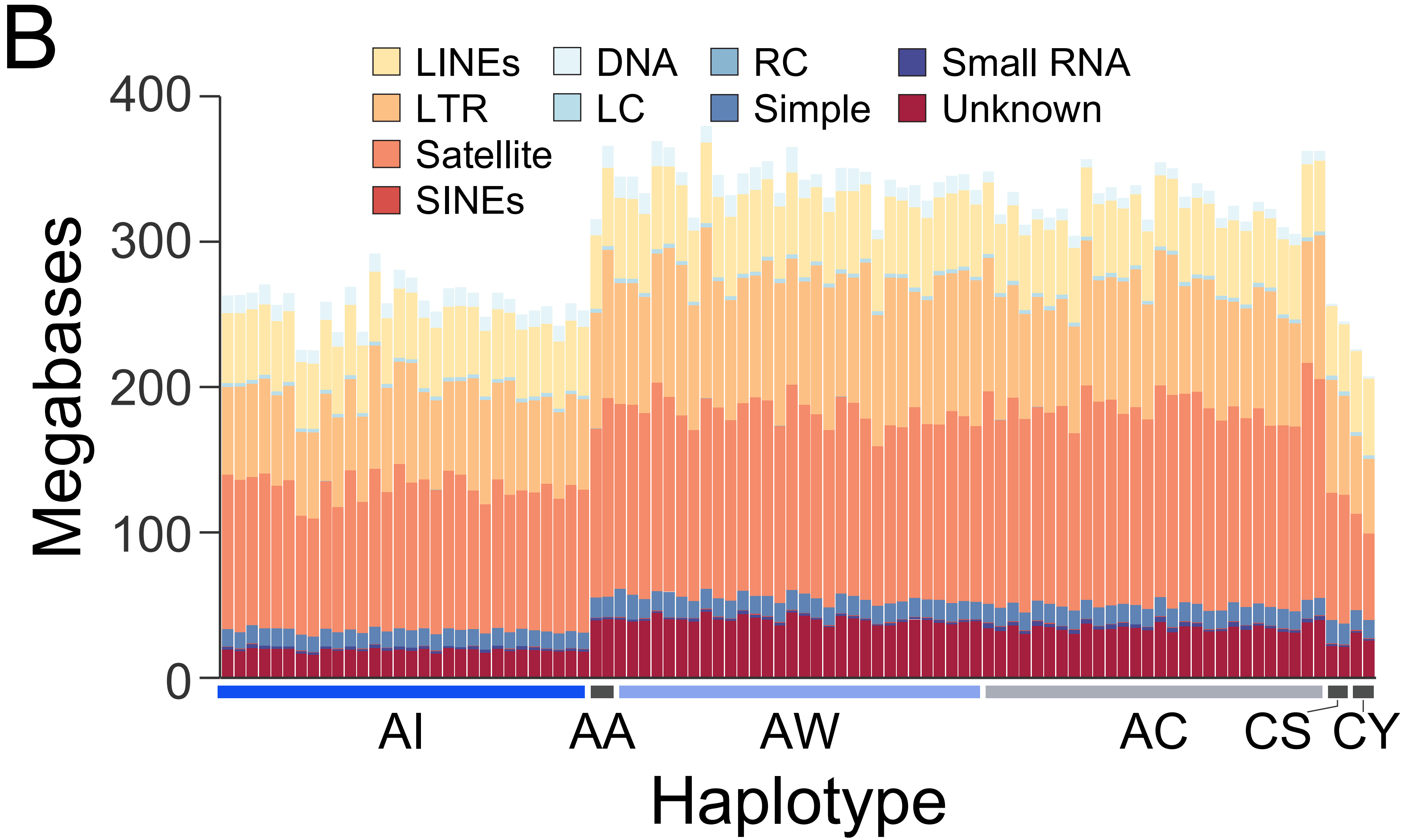
sum of contigs of primary assembly from hifiasm

Estimates of max genome size from Genomescope using k-mers

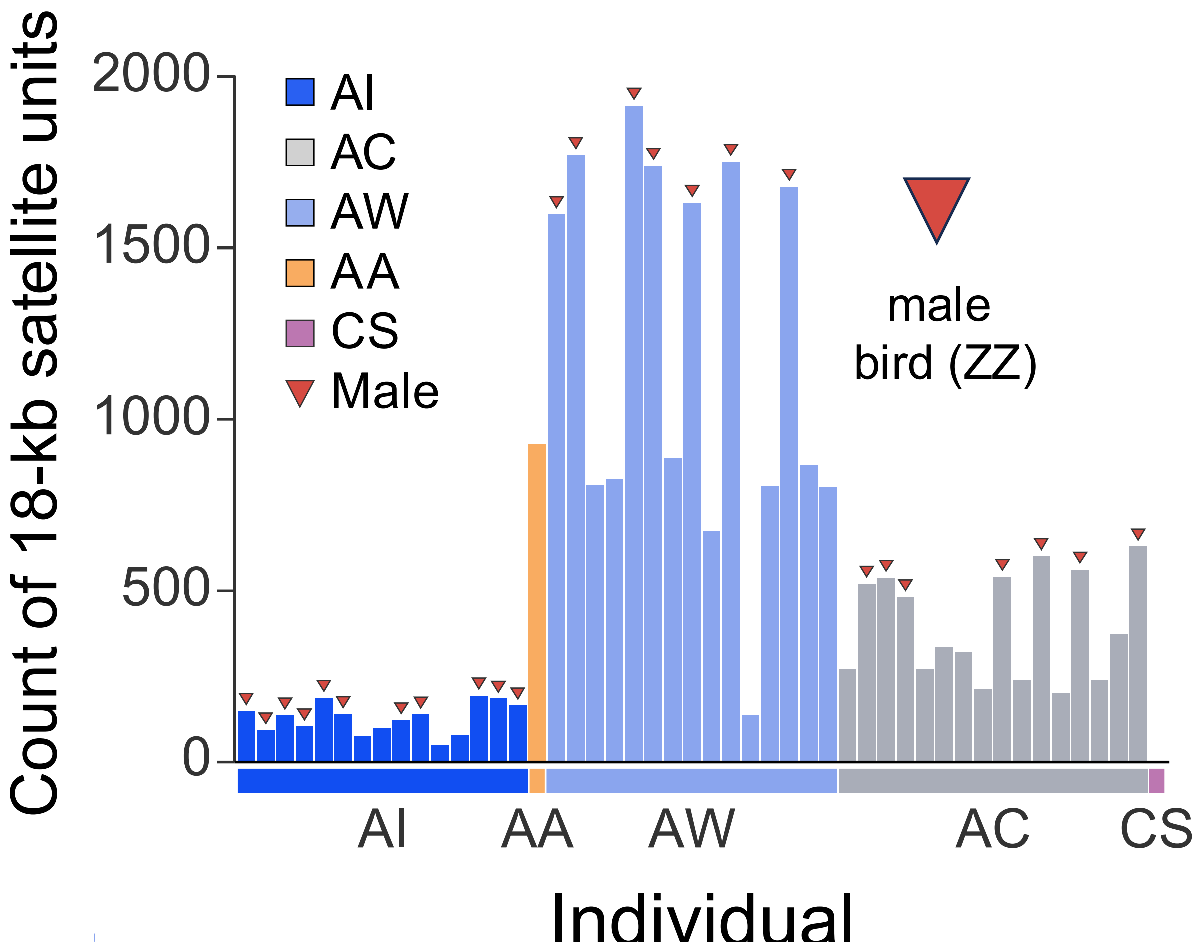


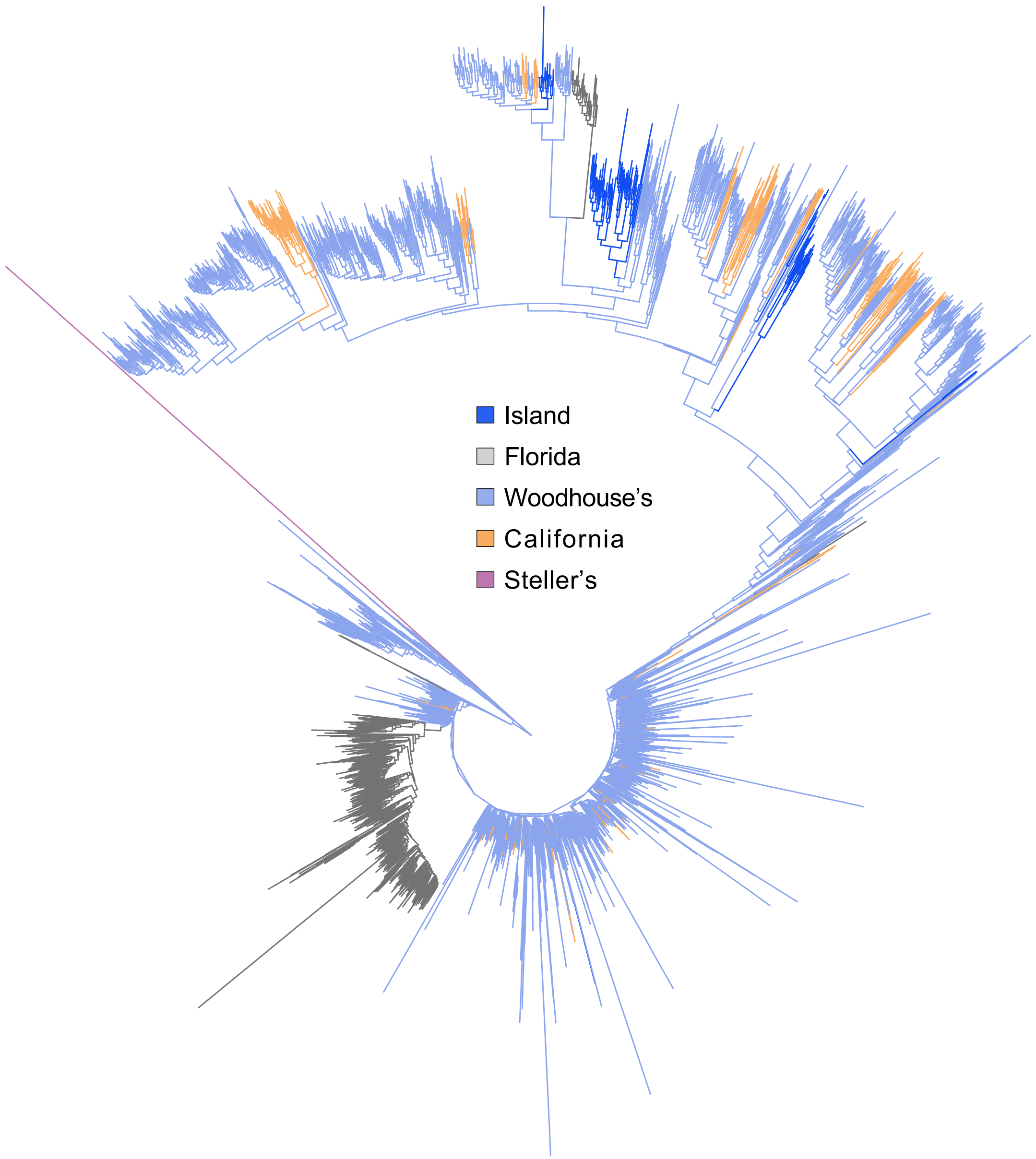


Interspecific variation in repeat content…



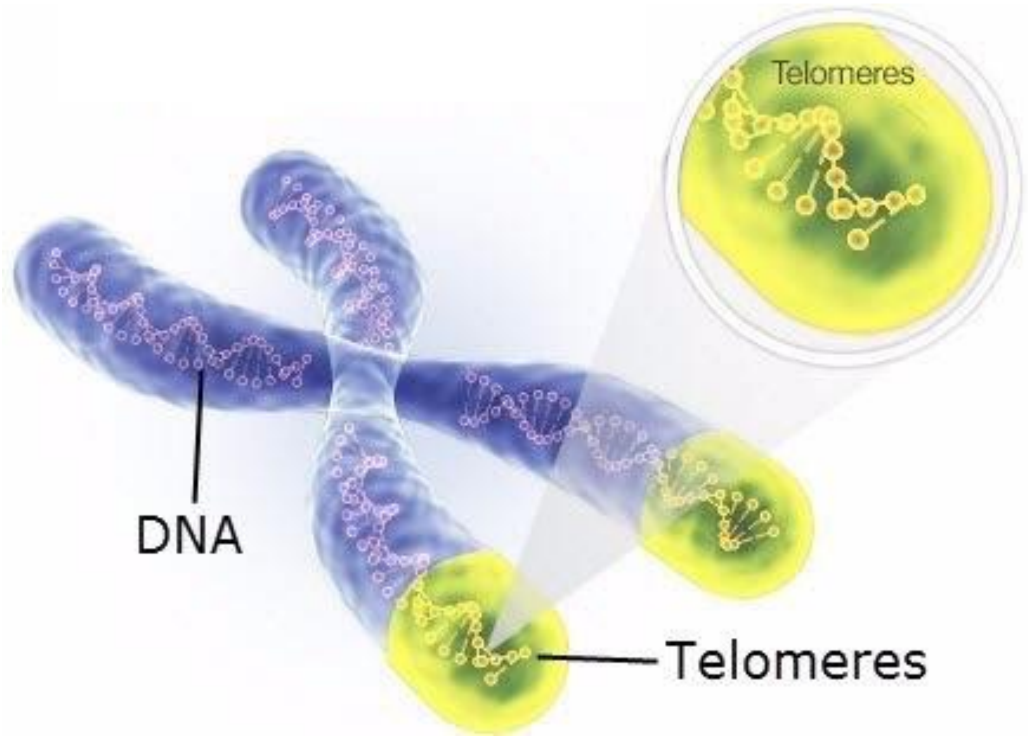
…and satellites





IQ-TREE v. 1.6.12, 3500 18-kb satellite sequences

Telomeres – barometers of age and stress in birds

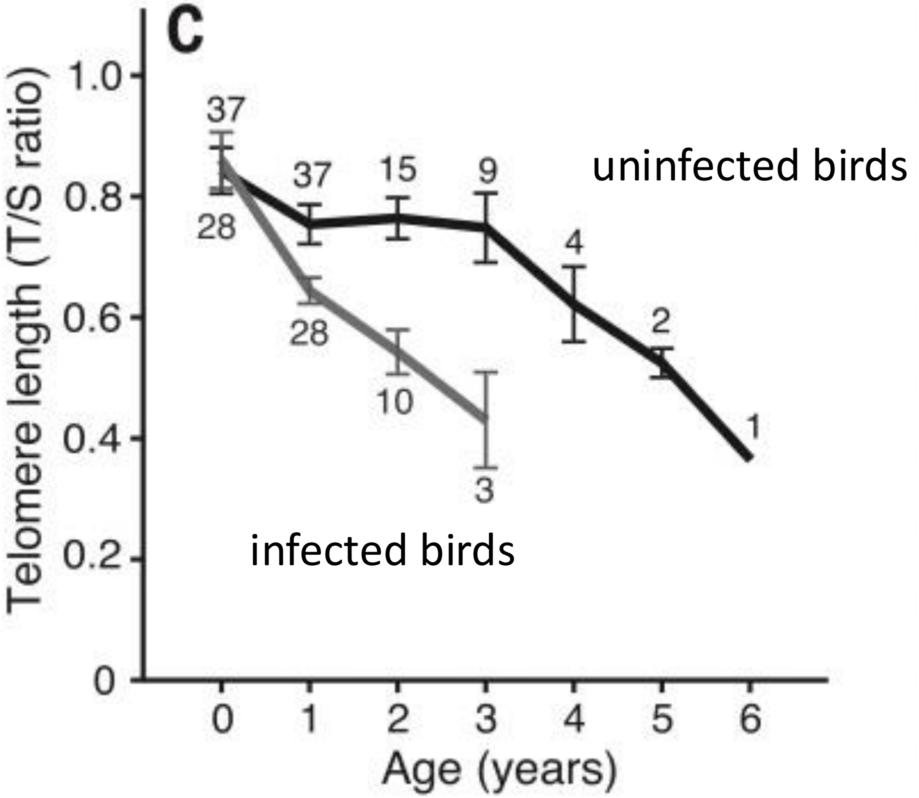


https://medibalans.com/telomere/

RESEARCH | *REPORTS*

CHRONIC INFECTION

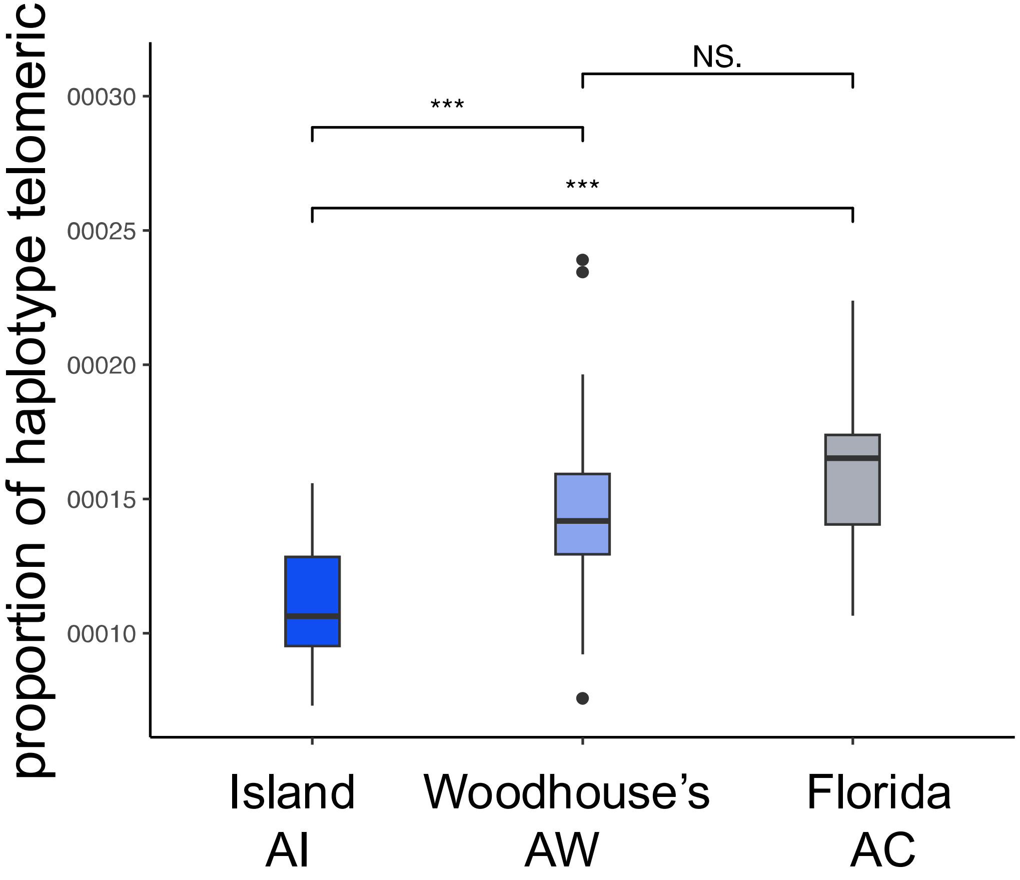
Hidden costs of infection: Chronic malaria accelerates telomere degradation and senescence in wild birds

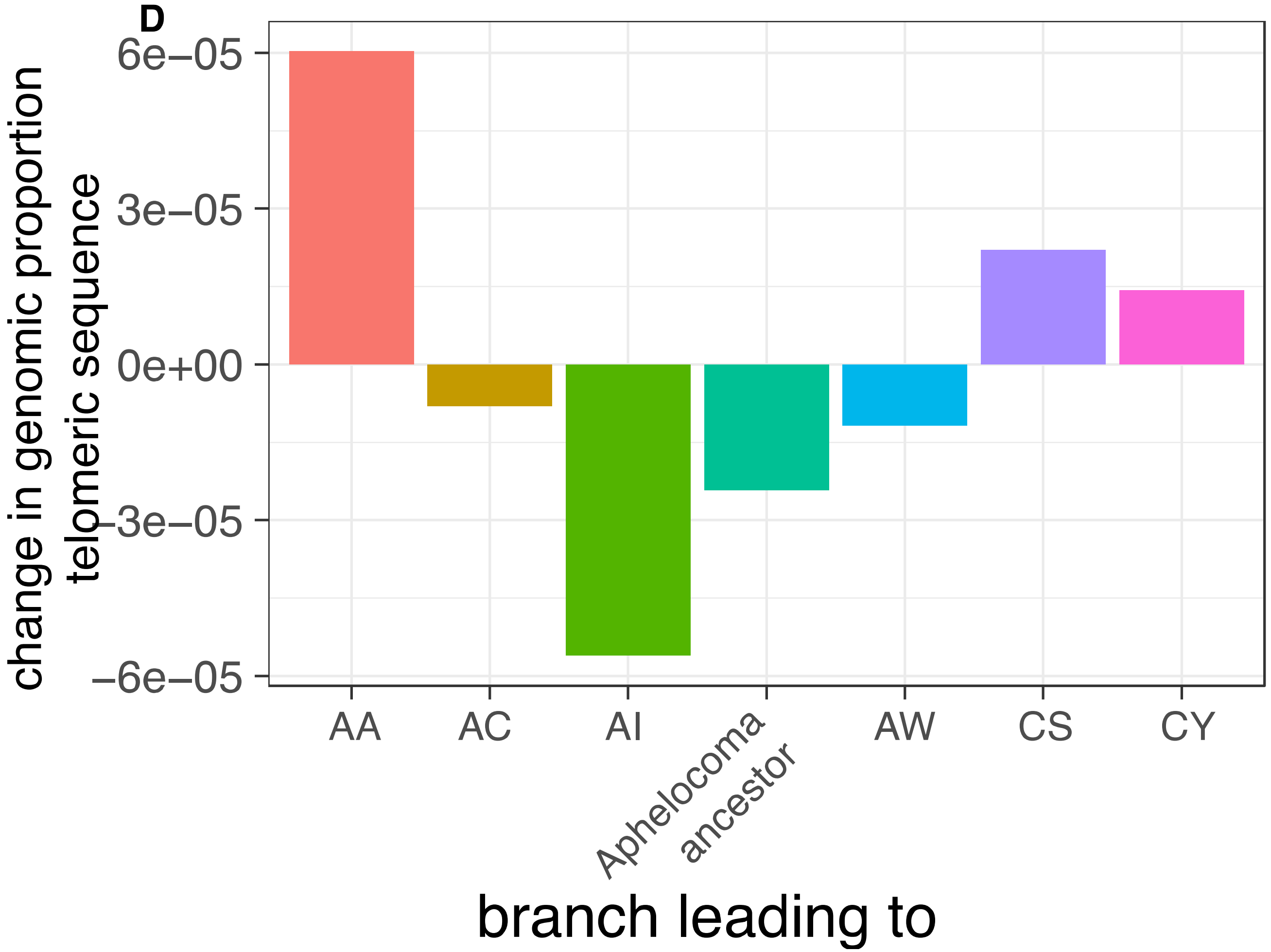


Ashgar et al. 2015. Science 347:436-438

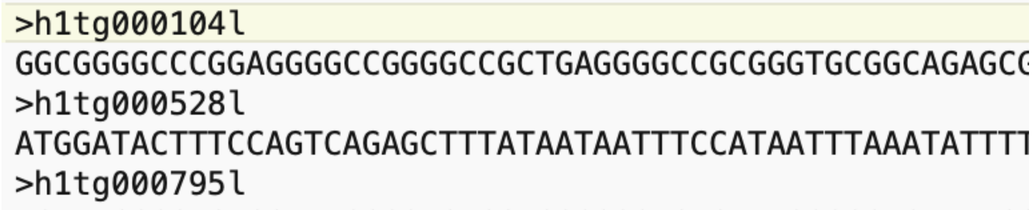
Telomere abundances influenced by population size

Species with smaller N_e are predicted to have shorter telomeres:
Brown et al. 2024. *Genome Biol. Evol*. evae111





Pangenome graphs capture structural variation within species



```
>h1tg000104l
GGCGGGGCCCGGAGGGGCCGGGGCCGCTGAGGGGCCGCGGGTGCGGCAGAGC
>h1tg000528l
ATGGATACTTTCCAGTCAGAGCTTTATAATAATTTCCATAATTTAAATATTTT
>h1tg000795l
ACTTTGGGGACACCTTTGGGGACACCTCGGGGGACACTTTGGGCCACAAATC
```

unaligned fasta files

minigraph

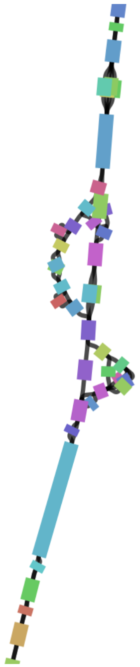
odgi

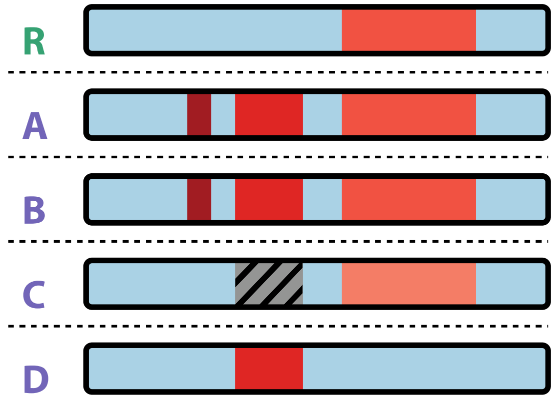
build genome graph

odgi viz

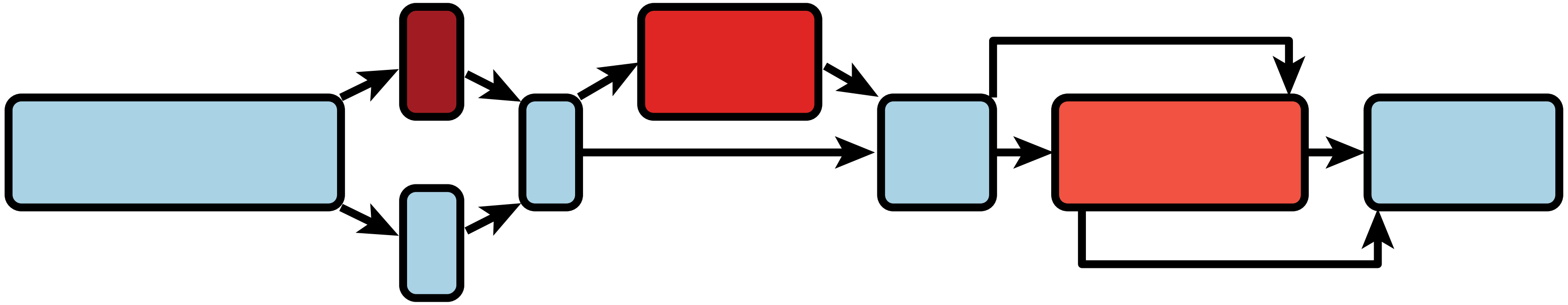
bandage

visualize genome graph





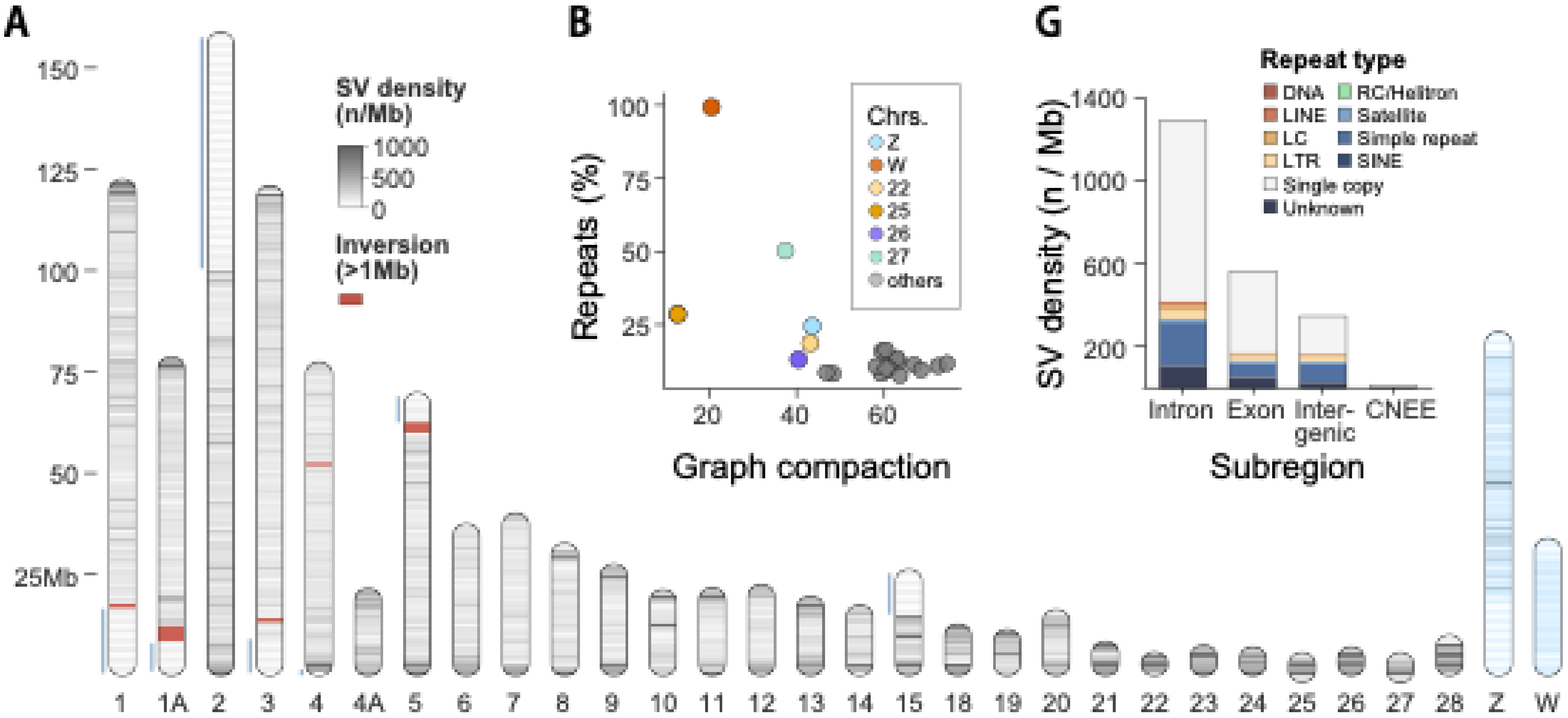
Multiple sequence alignment



Bidirected genome graph

Eizenga et al. 2021. *Ann. Rev. Genomics Hum. Genetics*

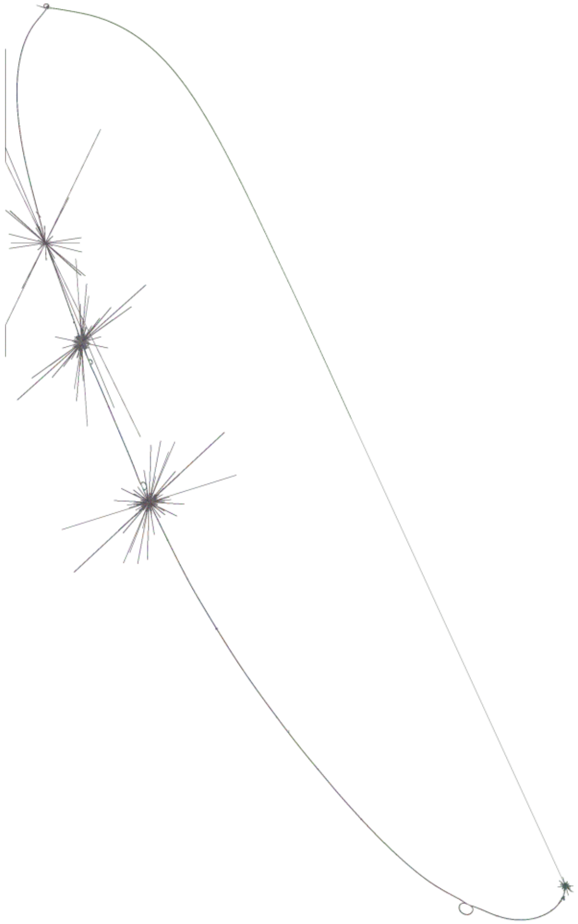
Distribution of genome-wide structural variants



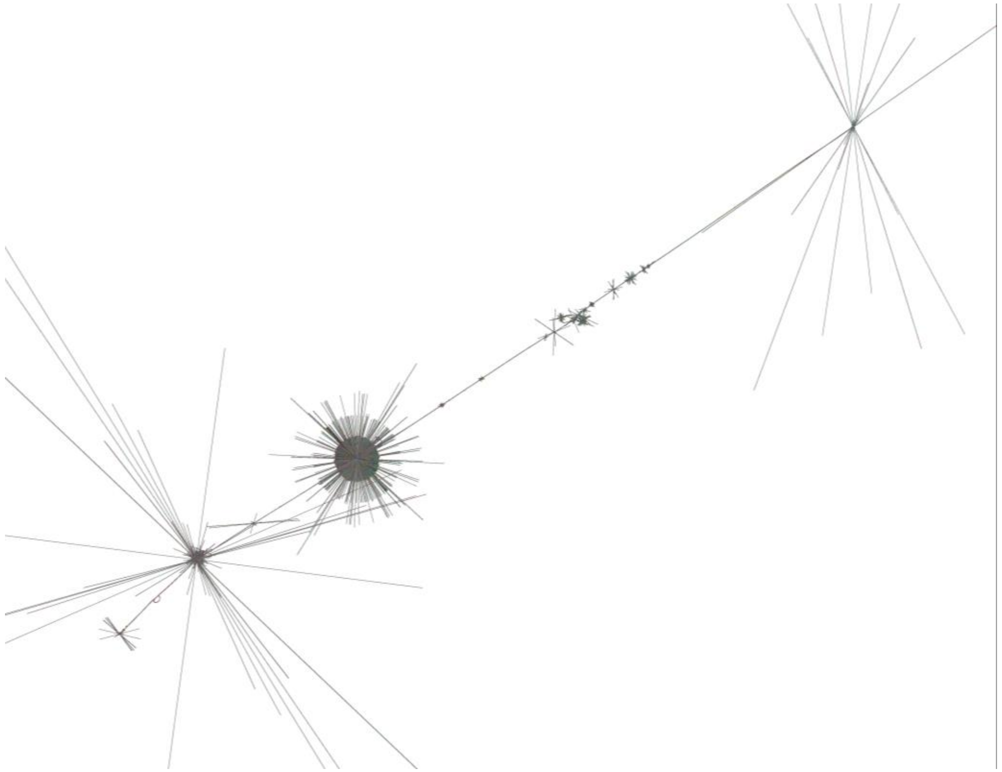
2D pangenome graph visualizations – PGGB/Odgi

Chr 18 – 12 Mb

Chr 1 – 160 Mb
‘telomere kiss’



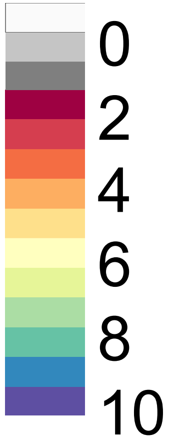
Chr 23 – 8 Mb

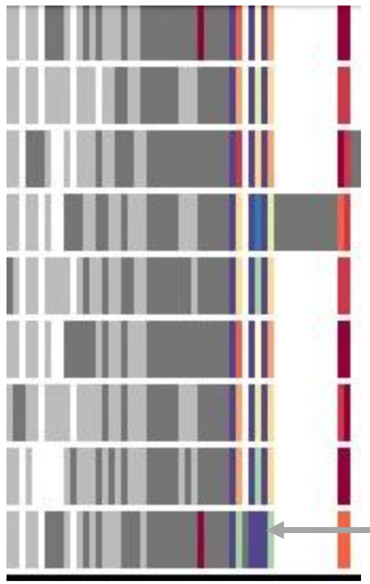


Variation in depth of a pangenome graph

medium
(self-)node-depth regions
depth = ~ n haplotypes

depth of
MHC
graph (x)



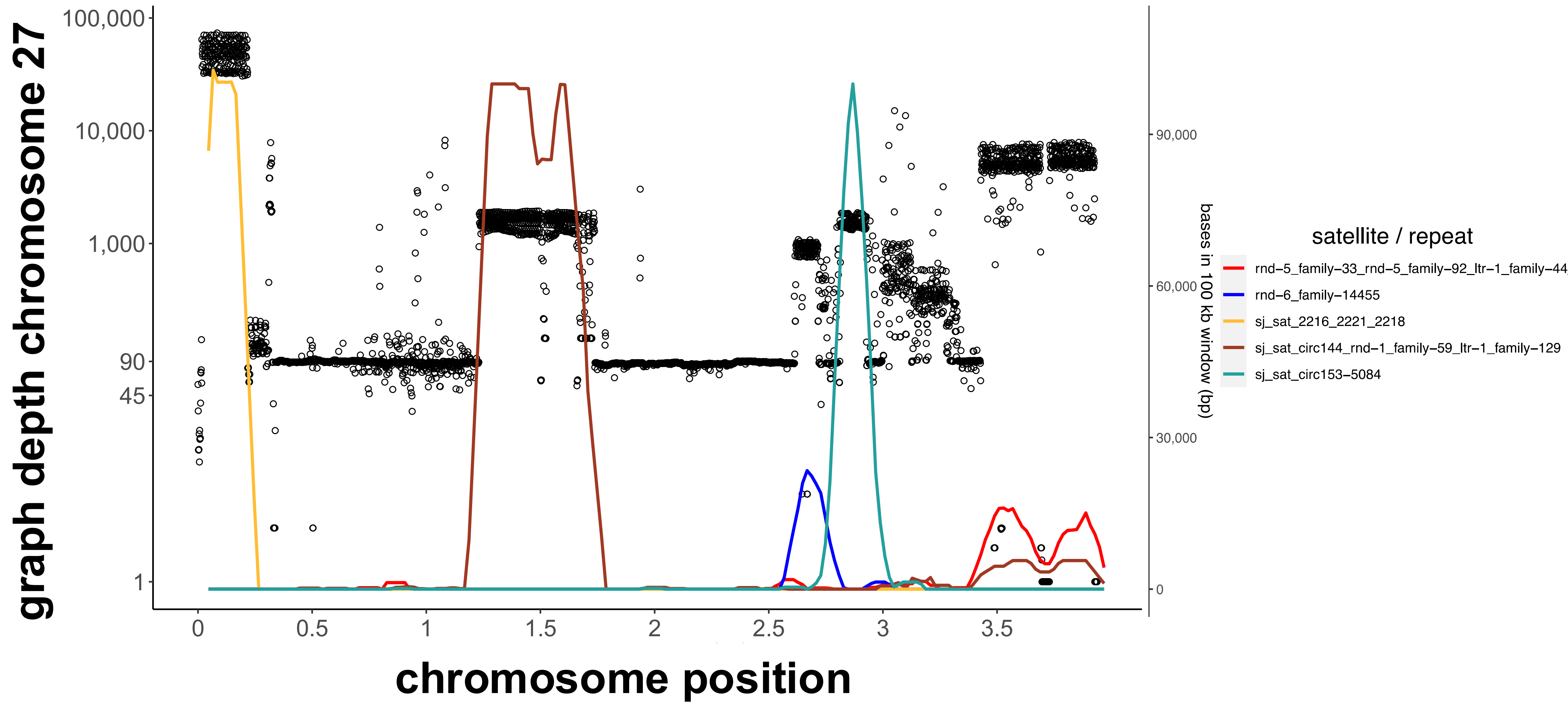


```
TGGGGATGATTTTGGGAGGTTTGGGGTTGATTTAAAGACATTTTGGGATATTT
TTGAGGGGATTTGGGGACATTTTGAGGGGTTTGGGGACACTTTGGGGACAATT
TTTTGGGAGGTTTGGGGCTGATTTTGAGAGGTTTGGATCCAACTTTGGAGCAT
TTTTGAGGGGGTTTGGGGCCACTTTGGGGCCACTTTGGTGGCCTCAGTGCCCC
TTTGGGGACACCTGGAGGGGTCTGGGGACAACTTGGGGTGCTTGGGAACACTT
TTAAGGATATTTTGGGACATTTTGAGGGGTTTGGGGACACTTTGGGGACAATT
```

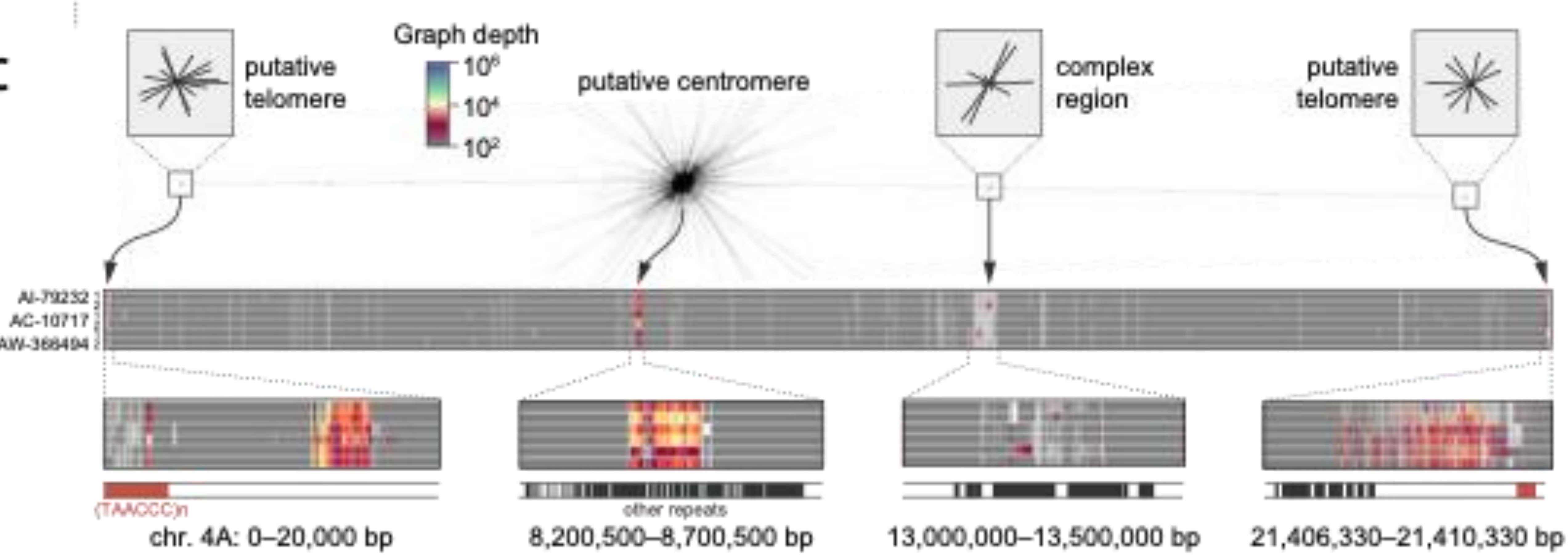
low node-depth regions – SNPs or small indels

high node-depth regions – large structural variants, satellites

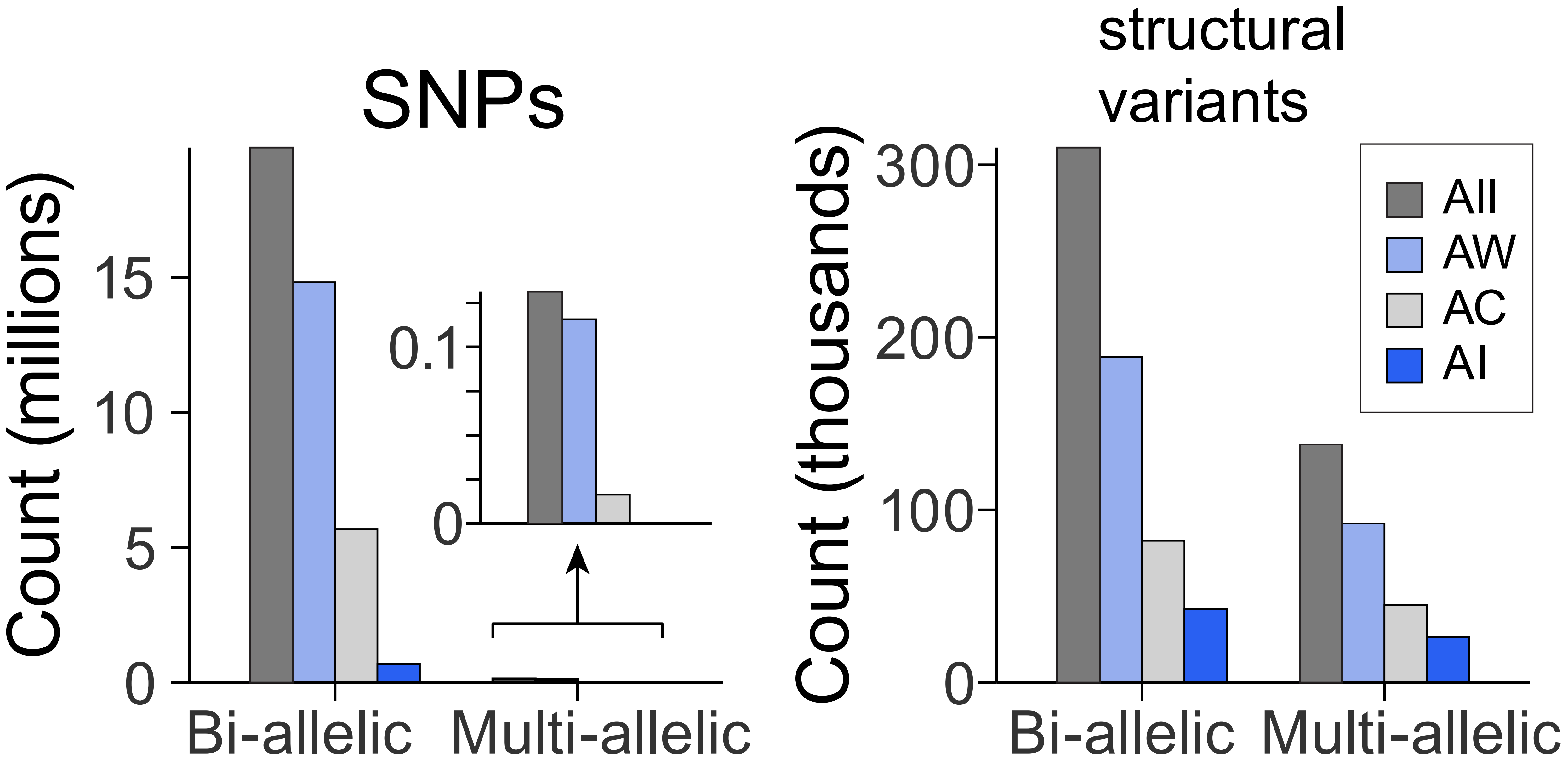
Graph depth of microchromosome 27 correlates with LTRs and satellites



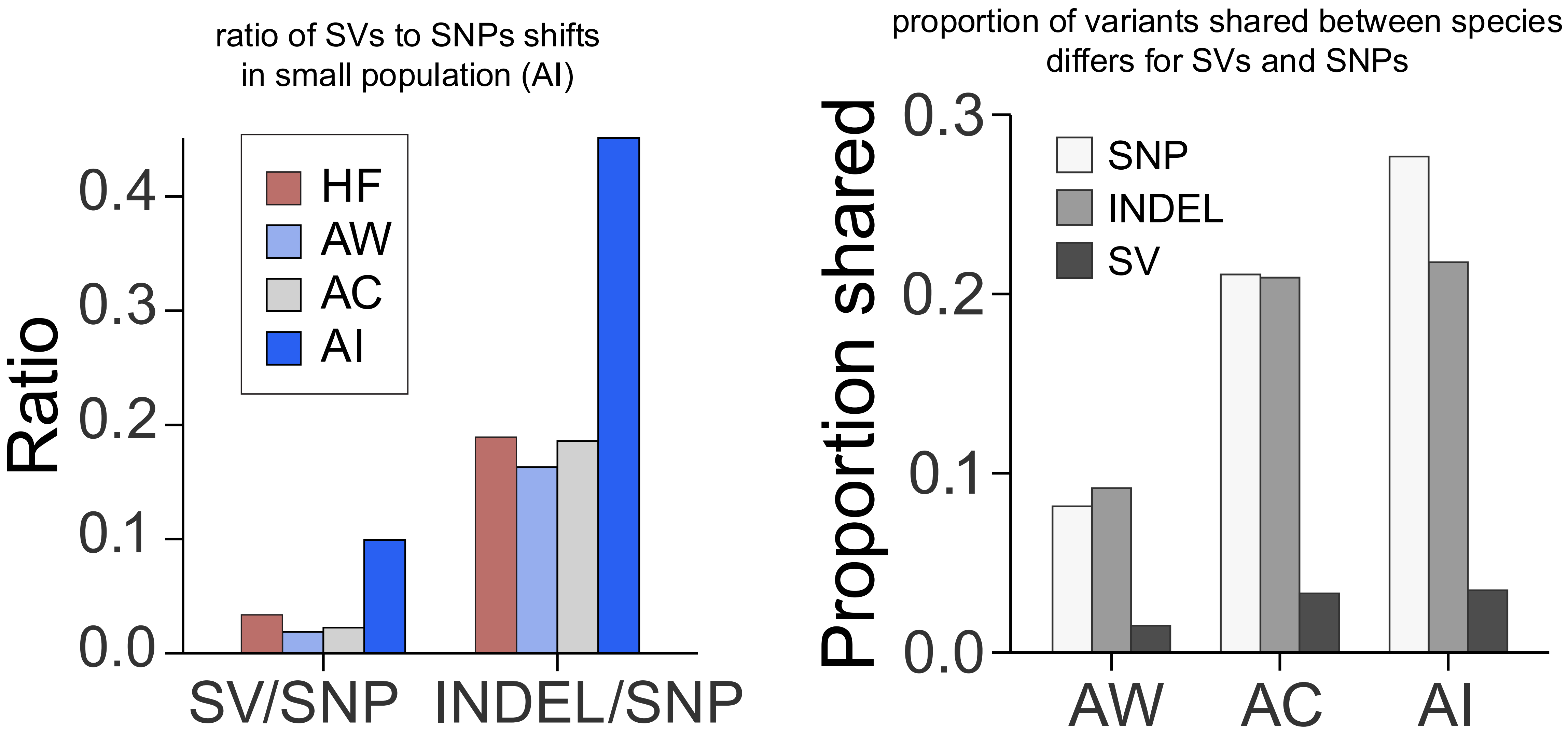
Pangenome depth reveals chromosomal features



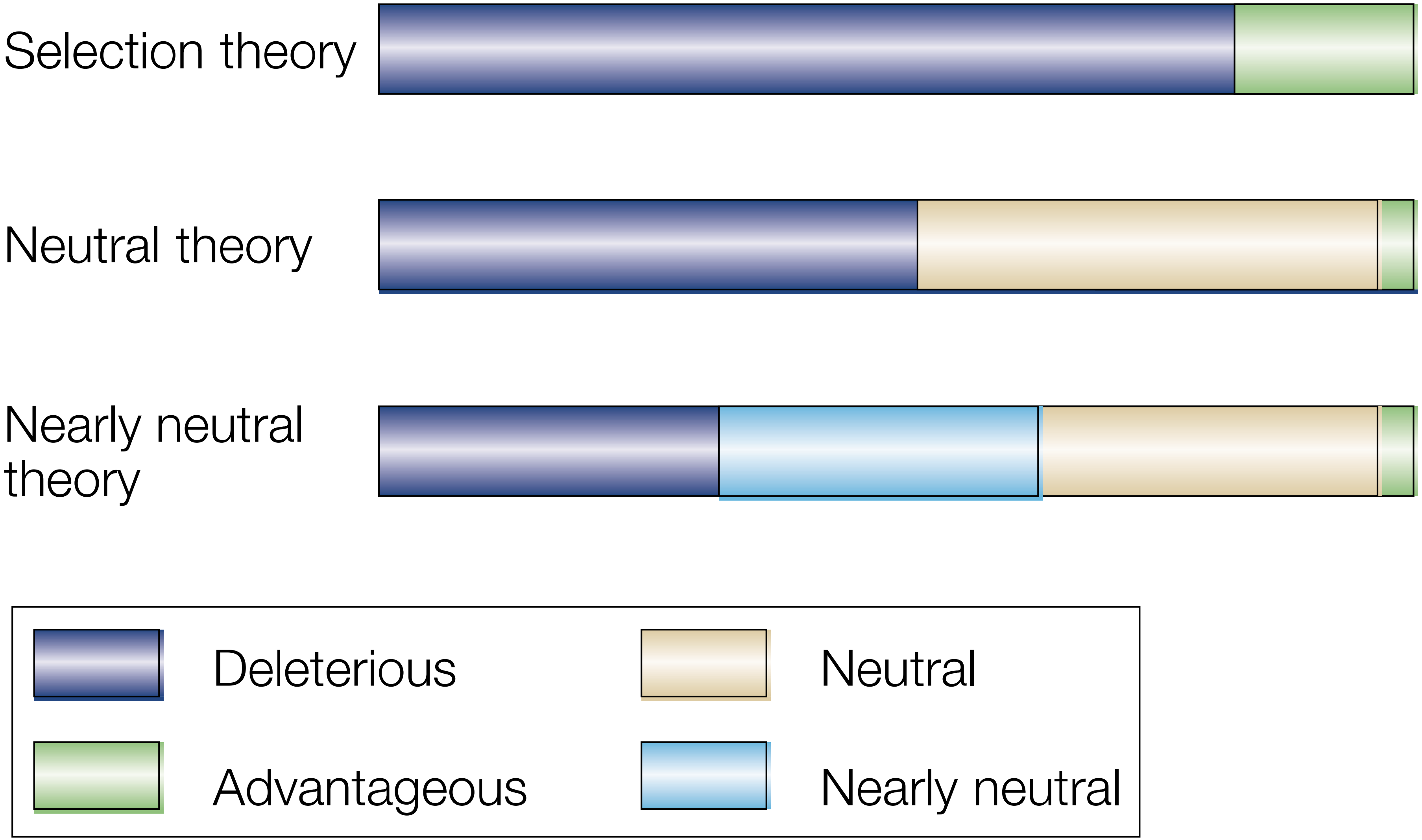
Number of structural variants scales with population size



Indirect evidence for an interaction between selection and genetic drift for structural variants

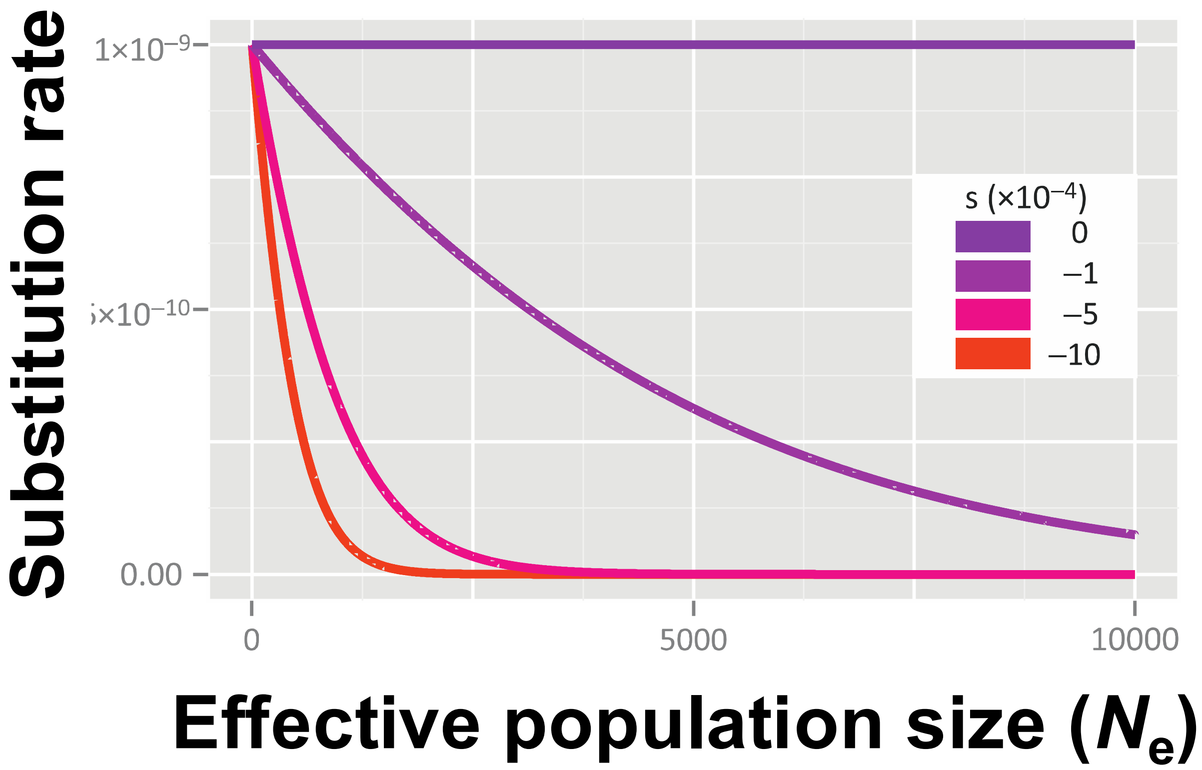


Nearly neutral molecular evolution

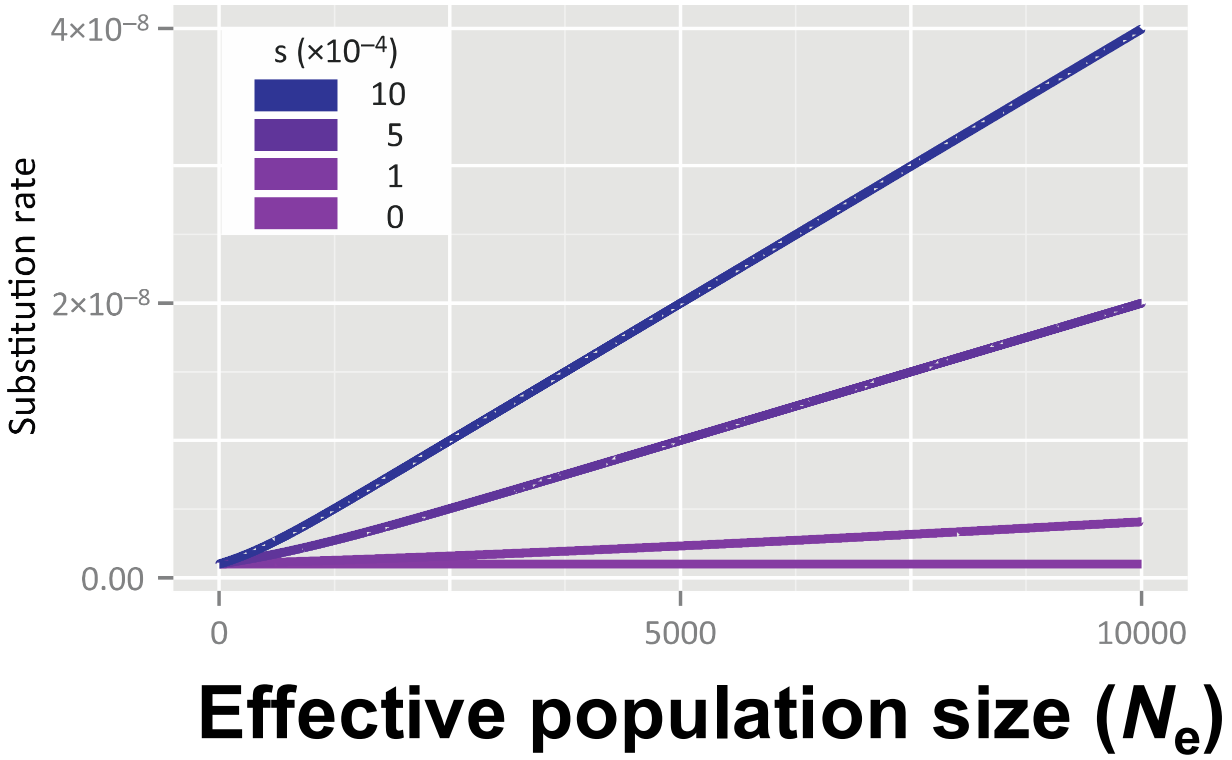


Bromham and Penny. 2003. *Nature Reviews Genetics*

Population size and rate of molecular evolution



deleterious mutations



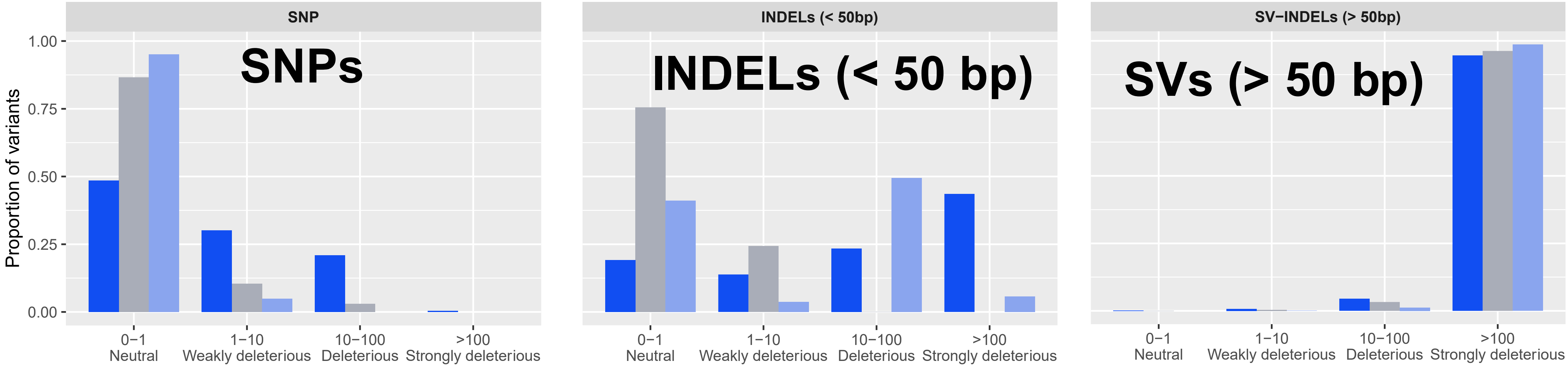
advantageous mutations

Lanfear et al. 2014. *Trends in Ecology and Evolution*

SVs are on average more deleterious than SNPs

Tomoko Ohta

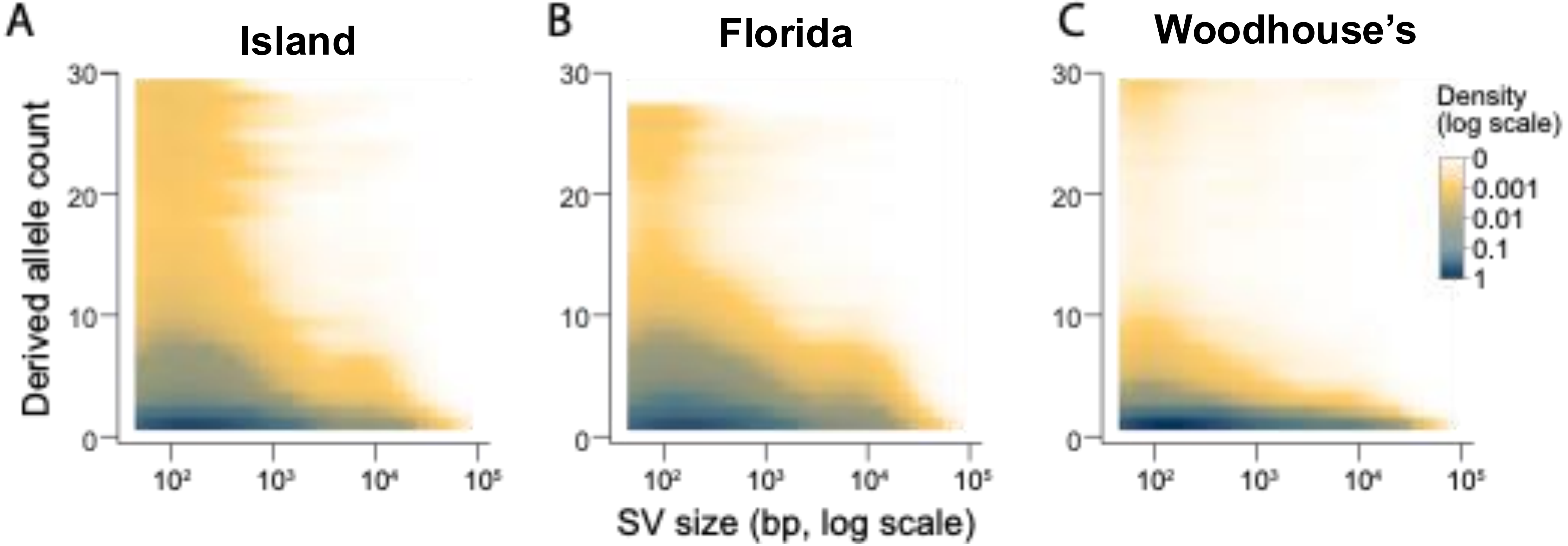
■ **Florida**
■ **Woodhouse's**
■ **Island**



$-N_eS$ $-N_eS$ $-N_eS$

Barton HJ, Zeng K: **New Methods for Inferring the Distribution of Fitness Effects for INDELs and SNPs**. *Mol Biol Evol* 2018, **35**:1536-1546.

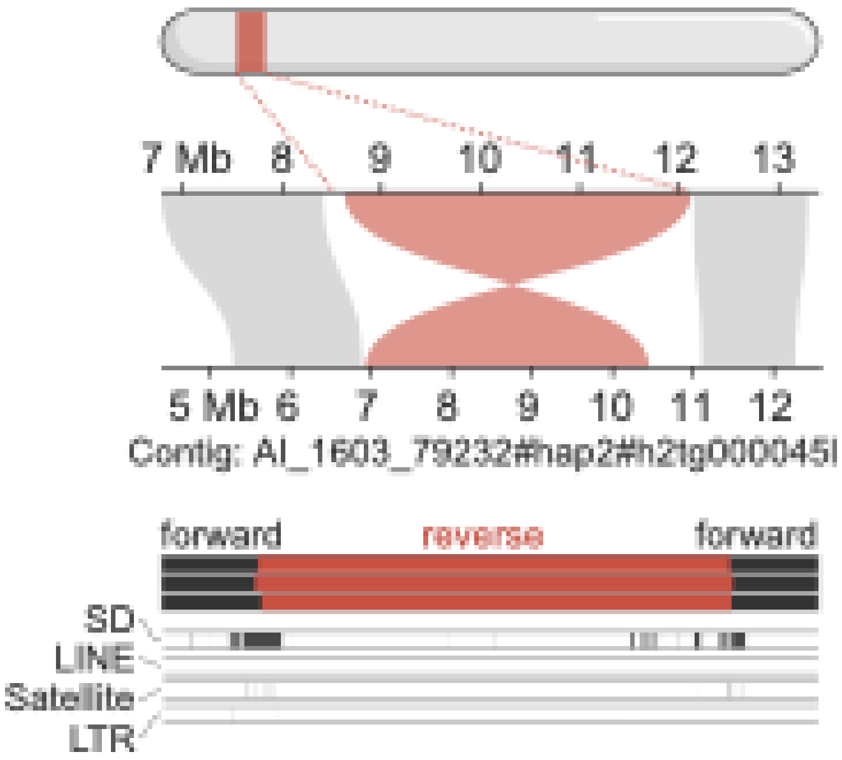
Longer SVs rise to higher frequencies in island population



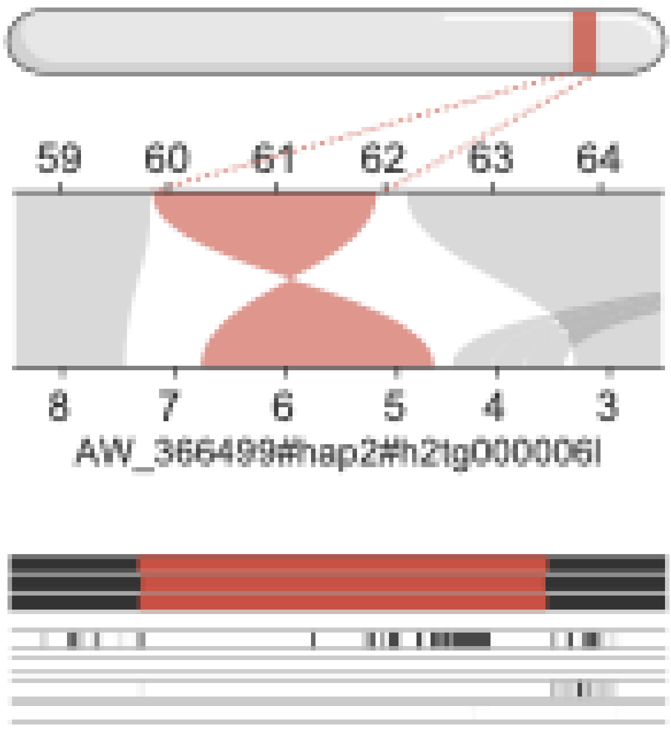
Density of ~485,000 structural variants extracted from PGGB pangenome graph

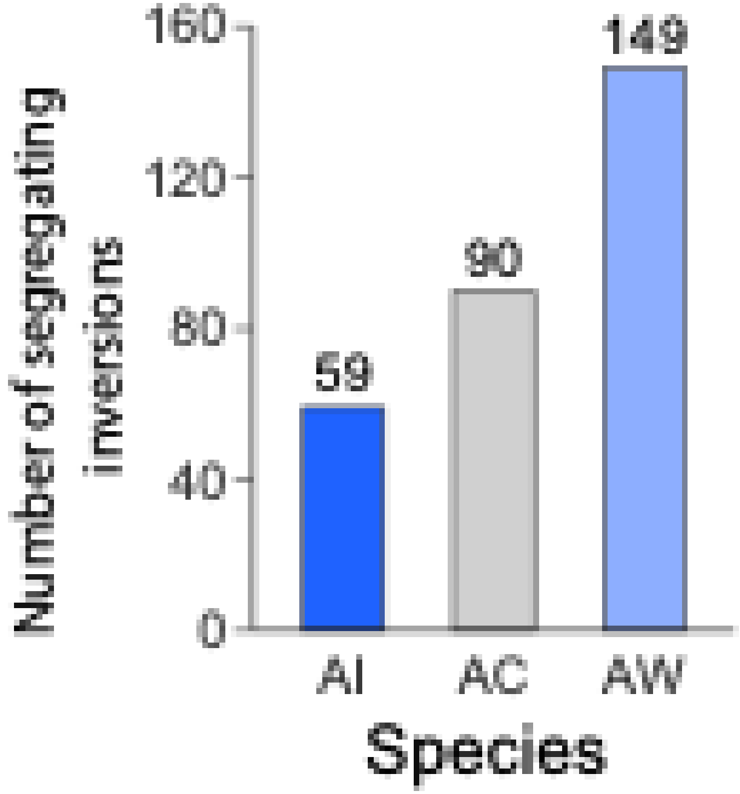
Inversions are common and track population size

Chr. 1A

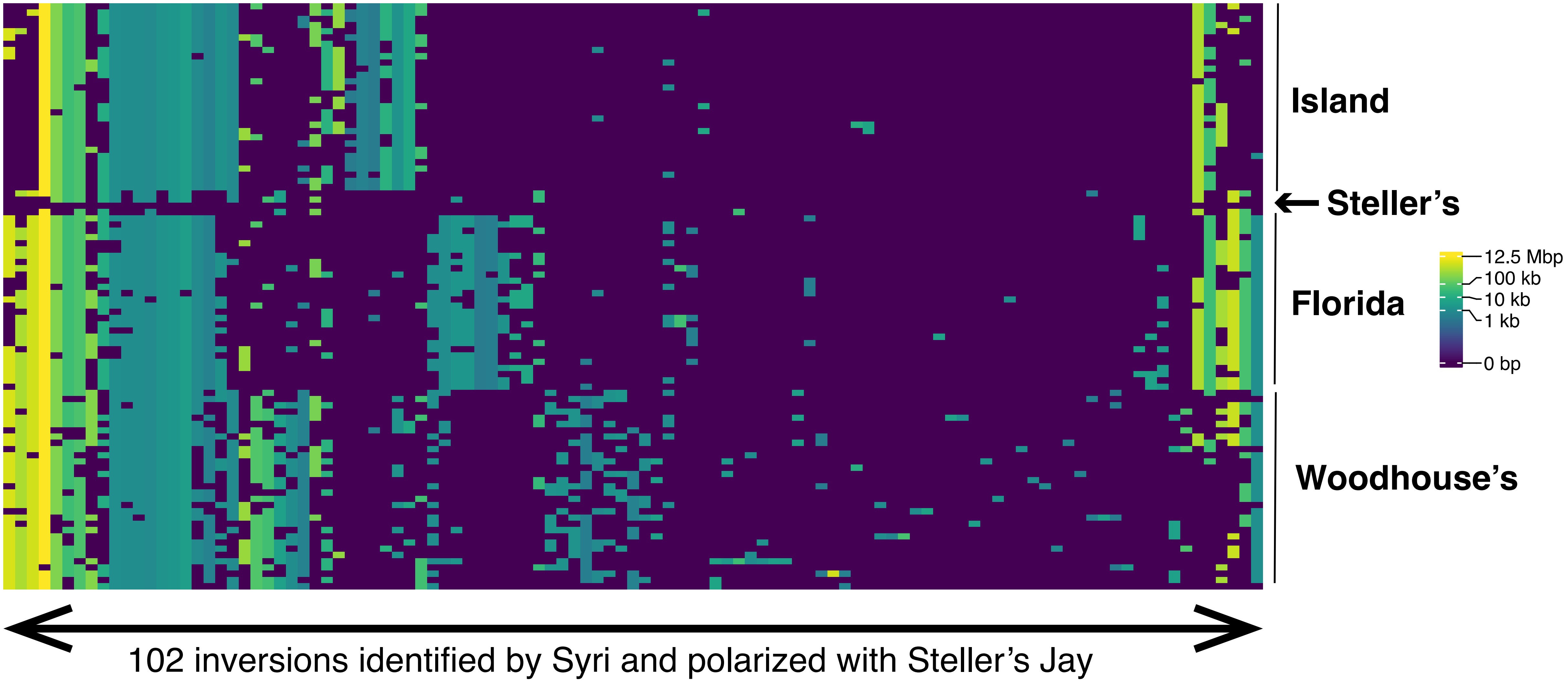


Chr. 5

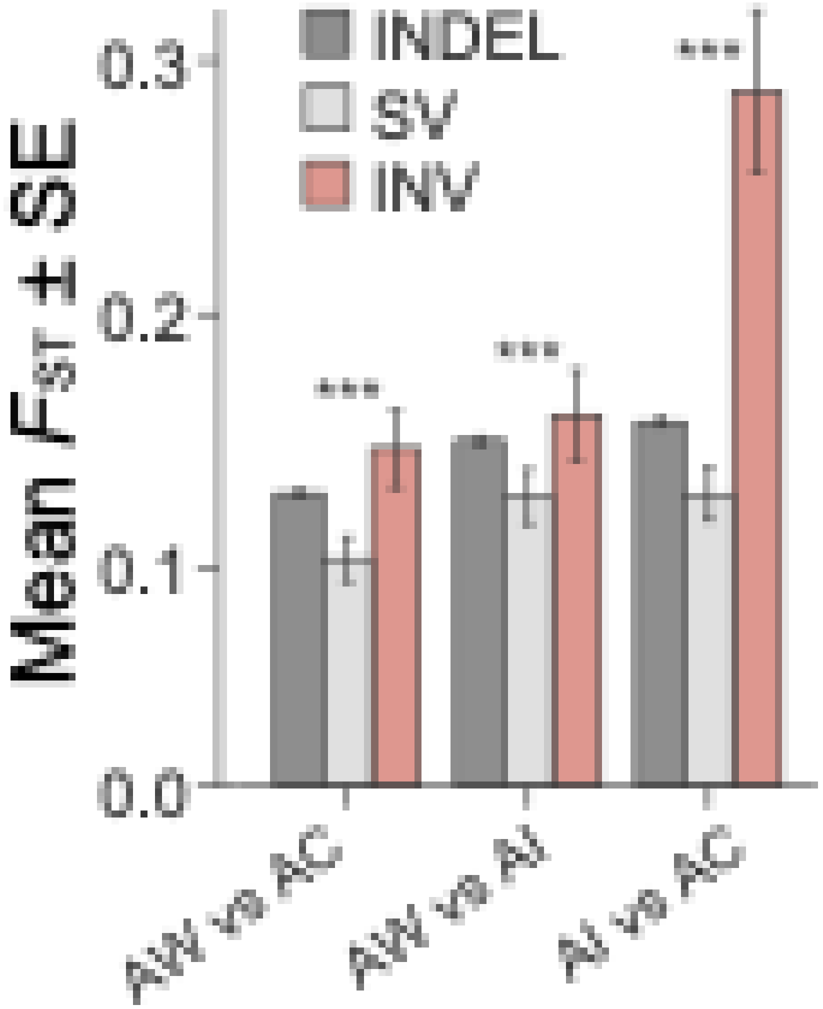




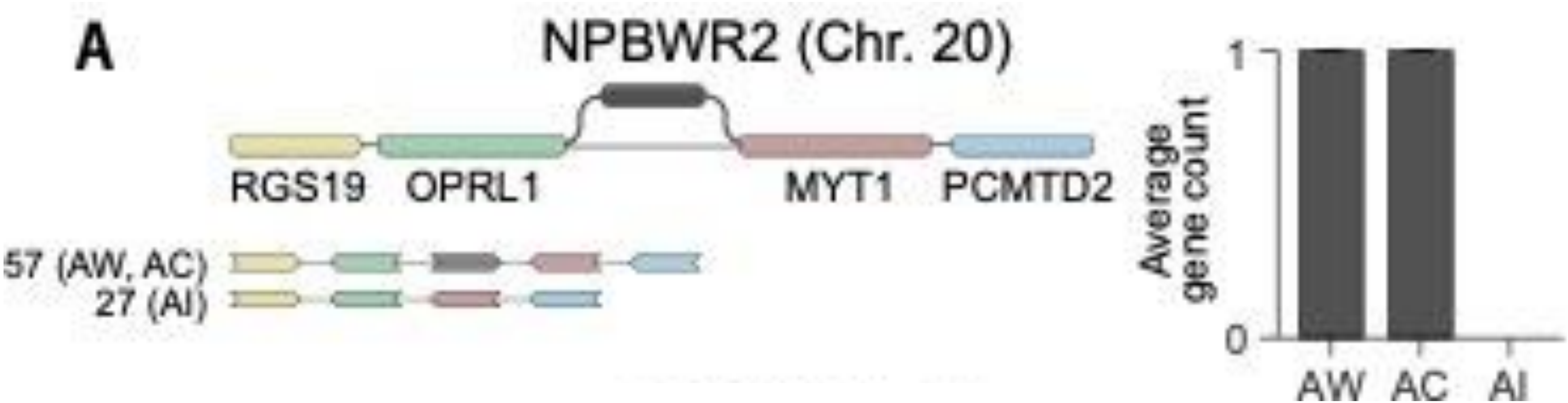
Numerous inversions identified by pangenome and reference-based methods



Inversions show enhanced differentiation between species

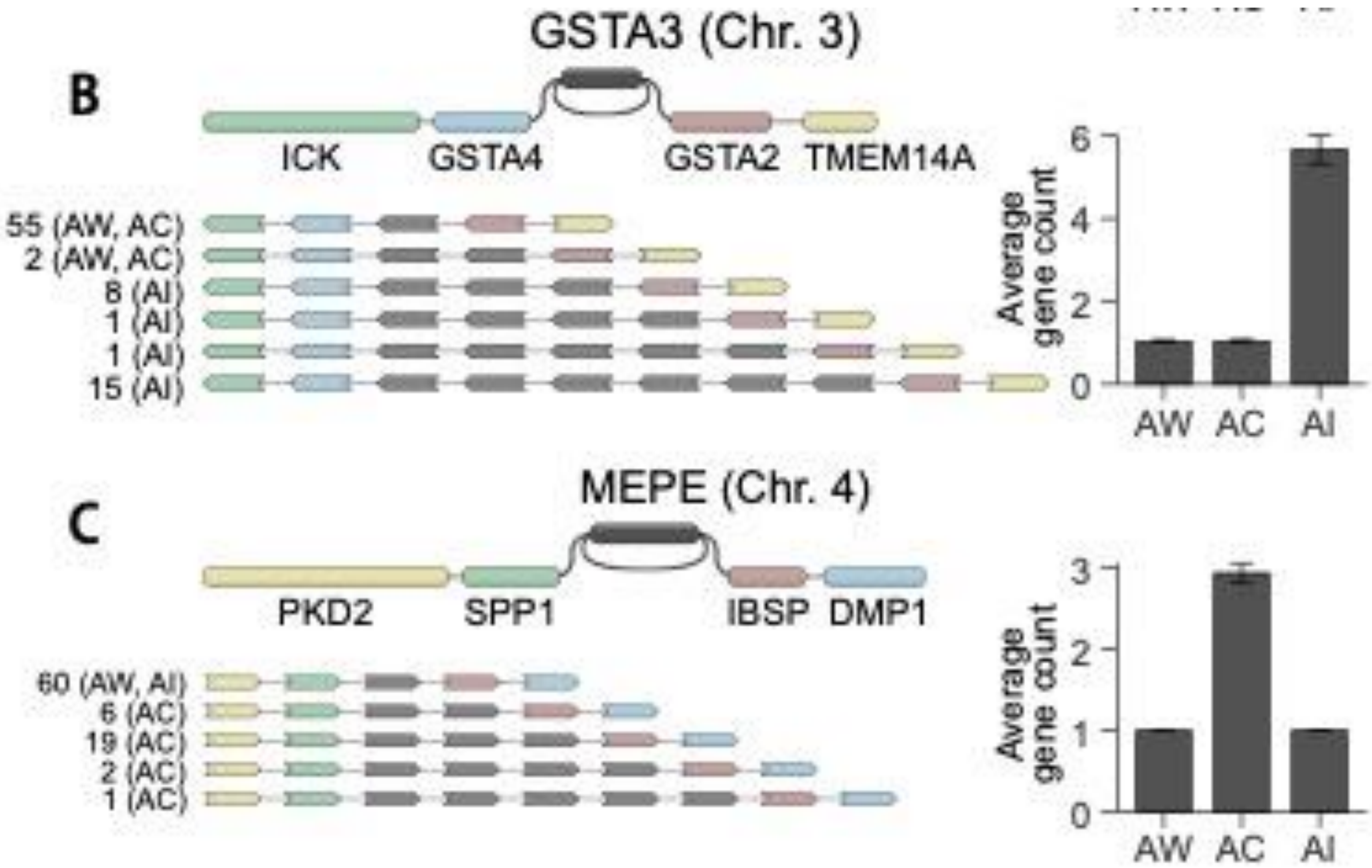


Abundant copy-number variants and gene deletions

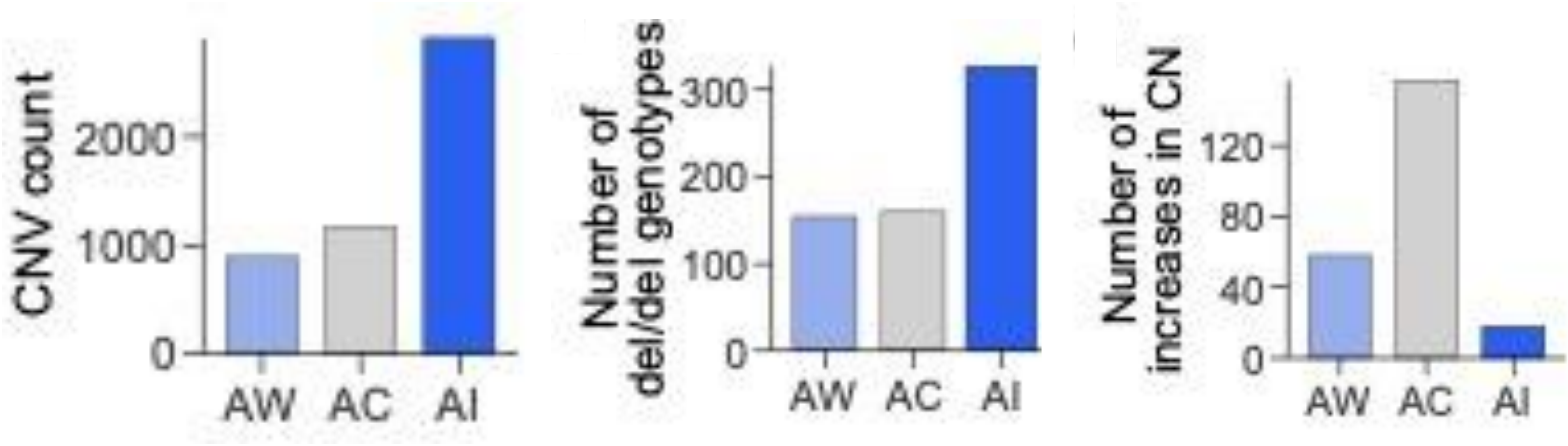


Using miniprot and pangene
Li et al. 2024. *Bioinformatics*
Li, H. 2023. *Bioinformatics*

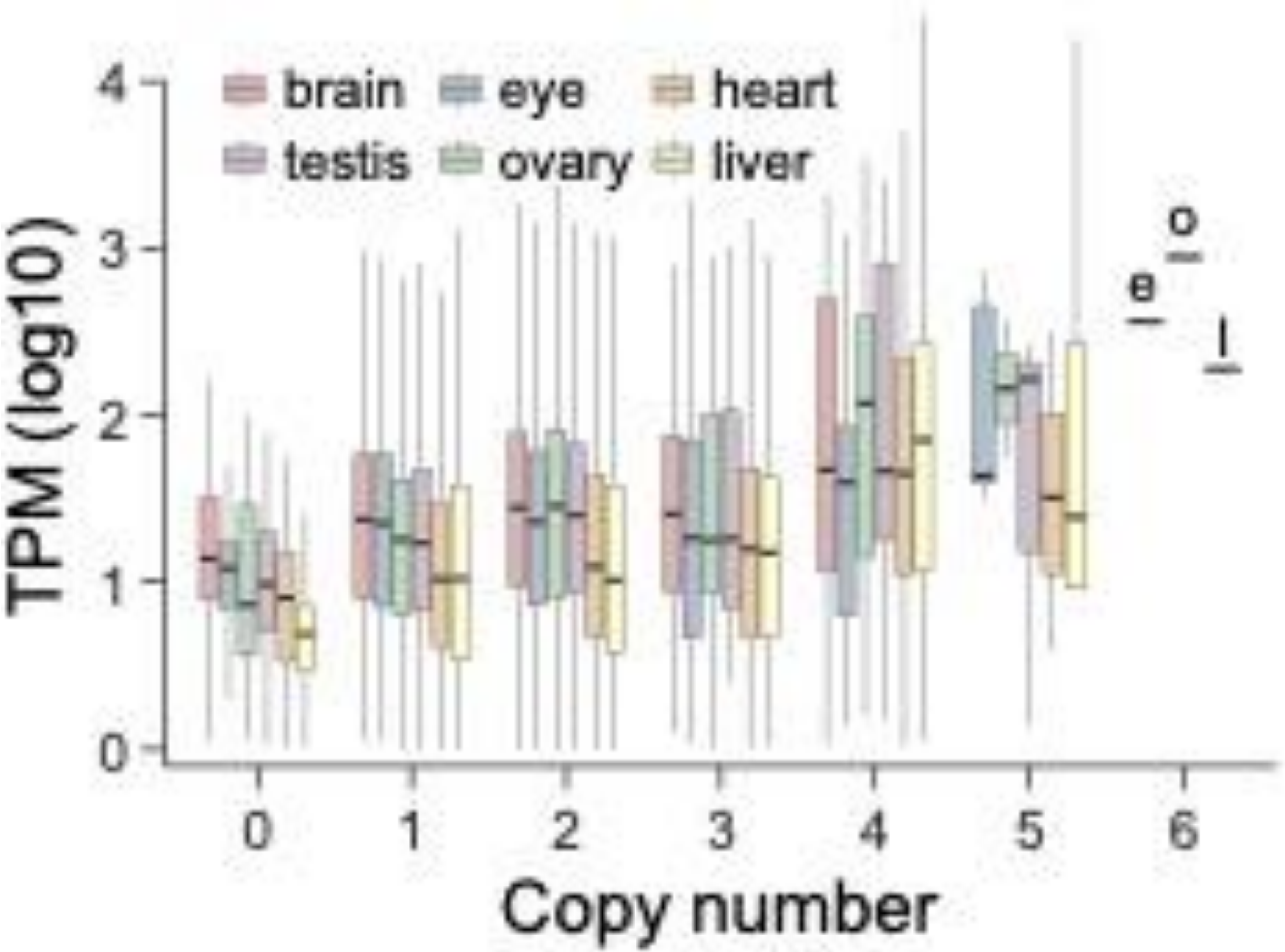
Abundant copy-number variants and gene deletions



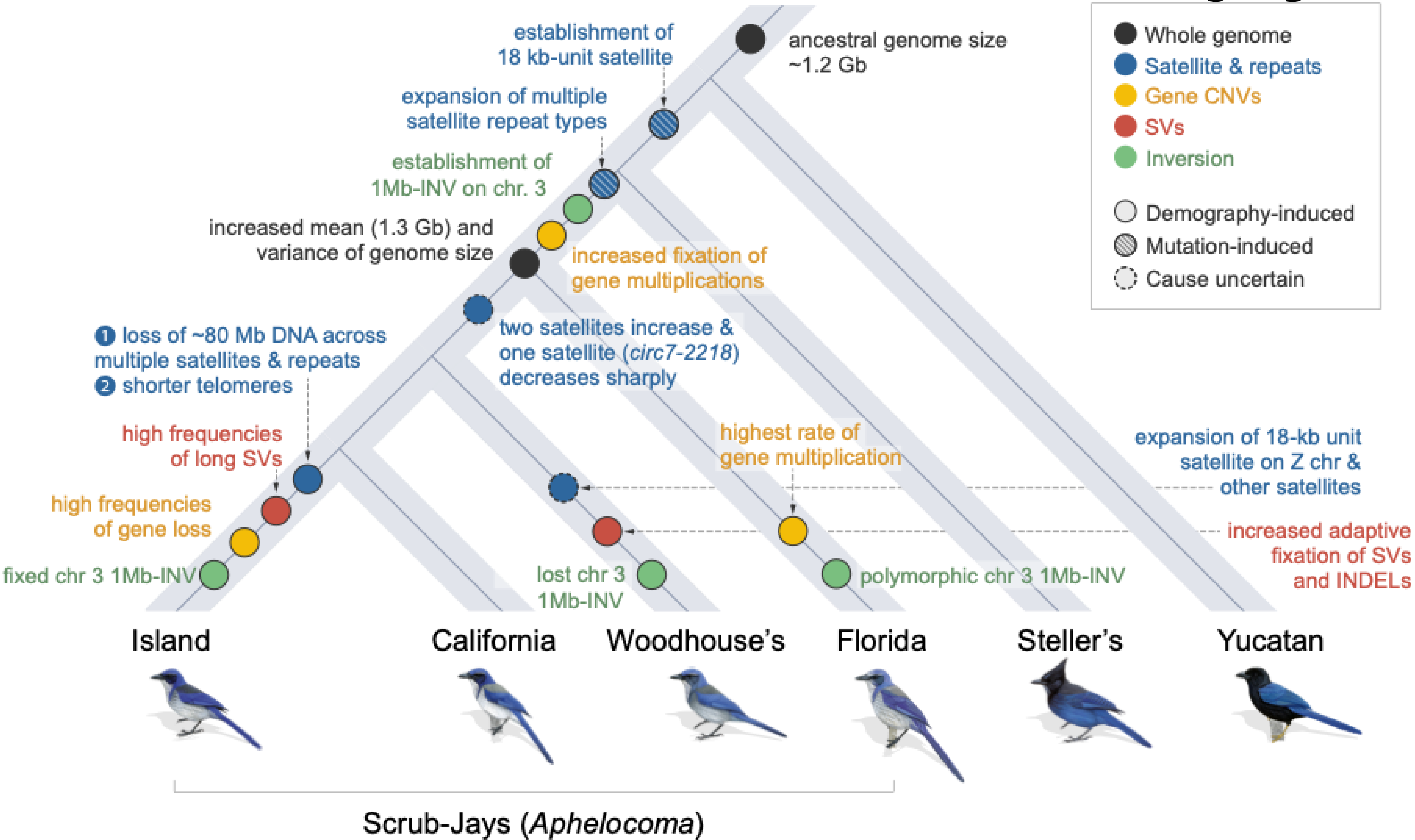
Gene copy number variants exhibit a surprising pattern



Copy number variants have functional consequences



Population size influences diversity of structural variants in scrub-jays



● Whole genome
● Satellite & repeats
● Gene CNVs
● SVs
● Inversion

○ Demography-induced
▨ Mutation-induced
◌ Cause uncertain

ancestral genome size
~1.2 Gb

establishment of
18 kb-unit satellite

expansion of multiple
satellite repeat types

establishment of
1Mb-INV on chr. 3

increased mean (1.3 Gb) and
variance of genome size

increased fixation of
gene multiplications

two satellites increase &
one satellite (*circ7-2218*)
decreases sharply

❶ loss of ~80 Mb DNA across
multiple satellites & repeats
❷ shorter telomeres

high frequencies
of long SVs

high frequencies
of gene loss

fixed chr 3 1Mb-INV

lost chr 3
1Mb-INV

highest rate of
gene multiplication

polymorphic chr 3 1Mb-INV

expansion of 18-kb unit
satellite on Z chr &
other satellites

increased adaptive
fixation of SVs
and INDELs

Island California Woodhouse's Florida Steller's Yucatan

Scrub-Jays (*Aphelocoma*)

Conclusions

- Scrub-jay genomes are repeat-rich
- Structural variants generally track population size
- Structural variants are estimated to be slightly deleterious
- Gene copy number variants look strongly deleterious
- Pangenome analysis will likely become the common standard