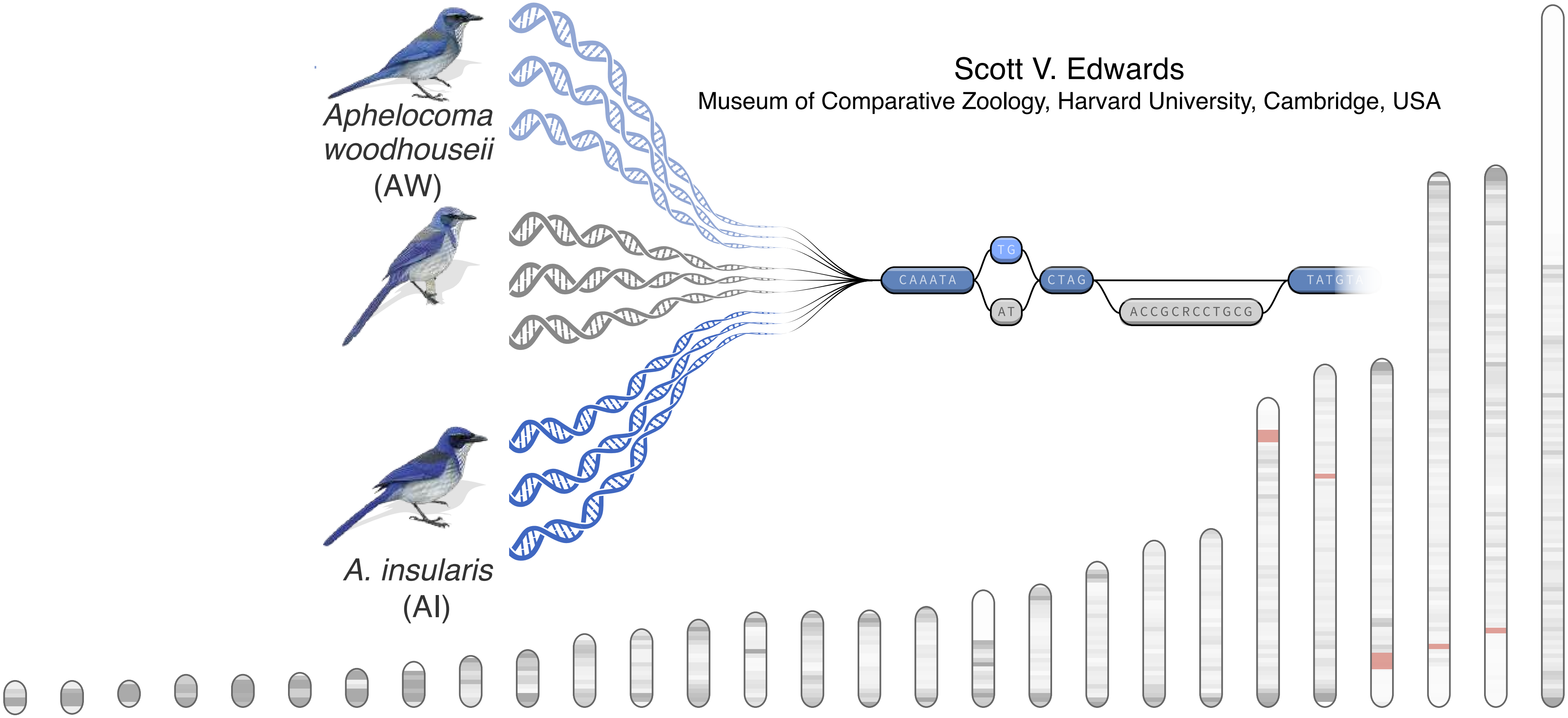
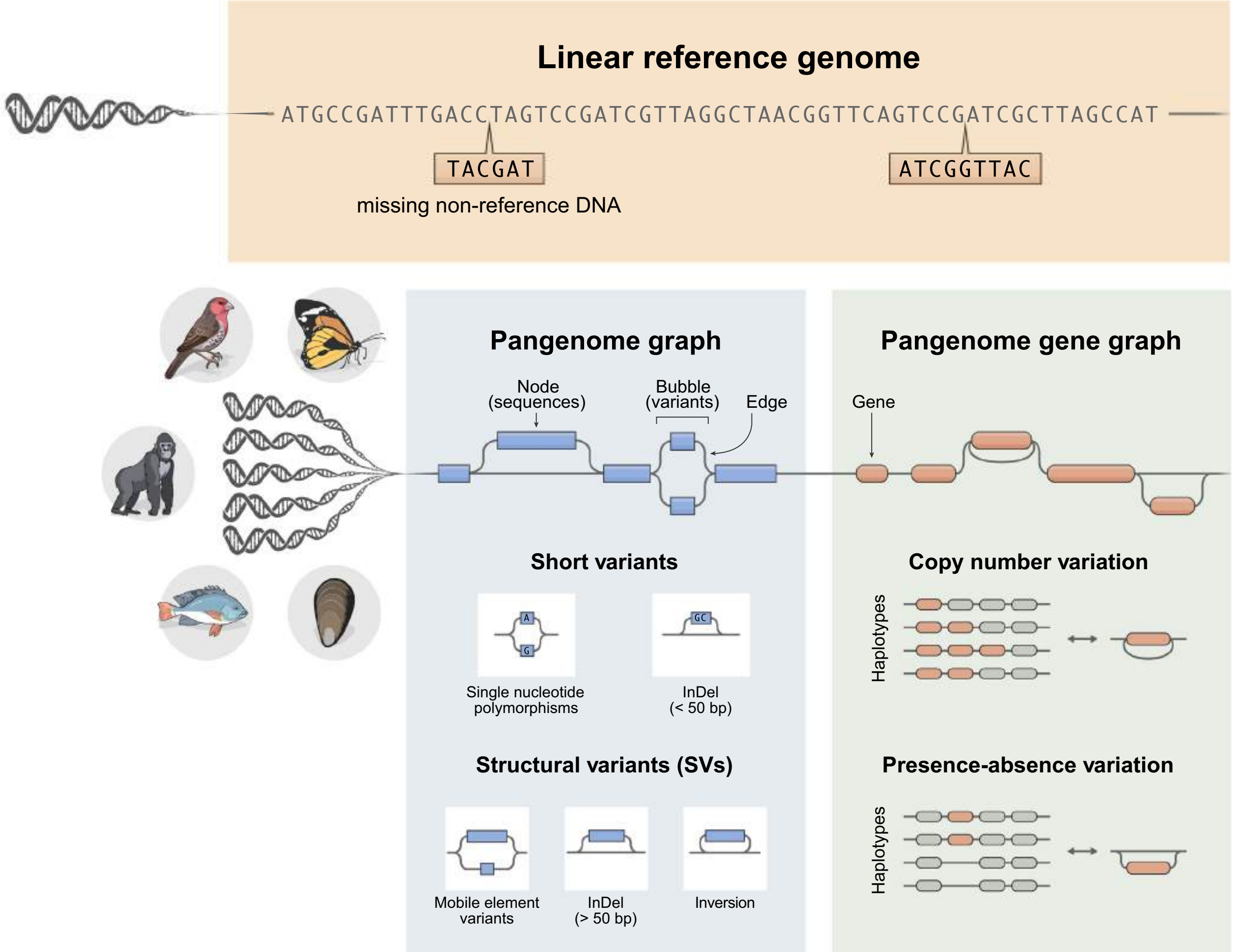
# Pangenomes: new tools for studying ecology and evolution

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



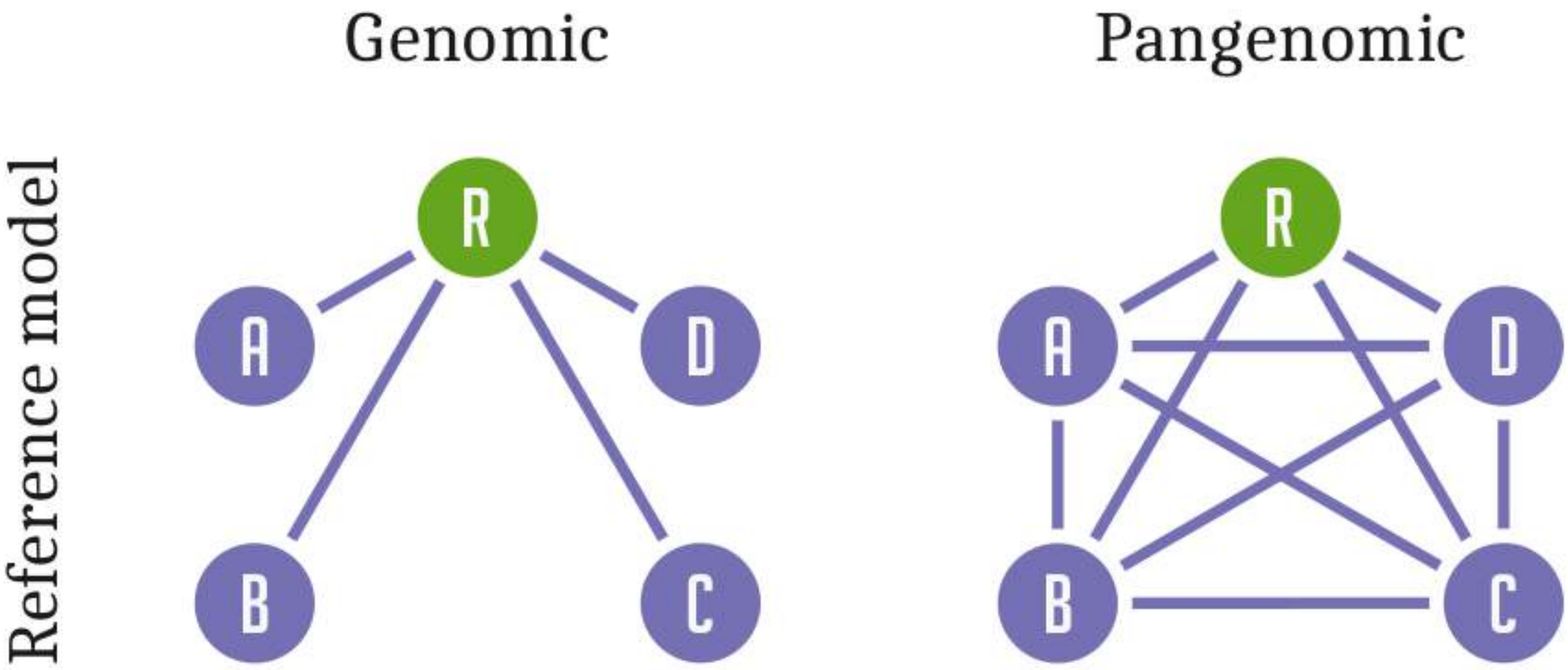
# Pangenomes: moving beyond reference-based genomics



Trends in Ecology & Evolution

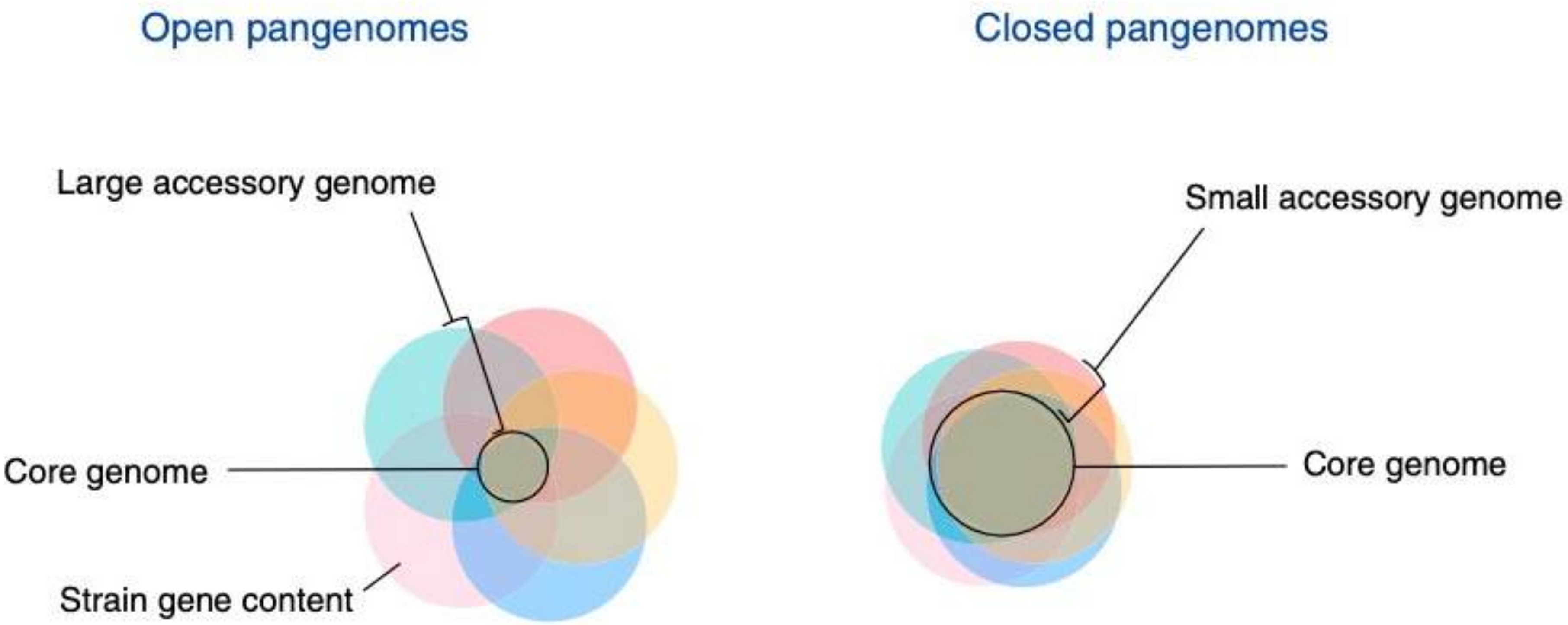
Fang and Edwards 2025 *TREE*

# Reference-free genomics



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

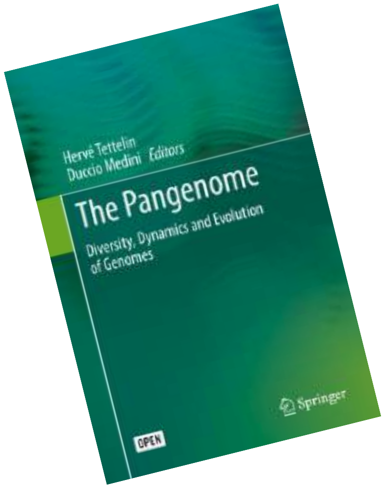
# Open and closed pangenomes



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)*



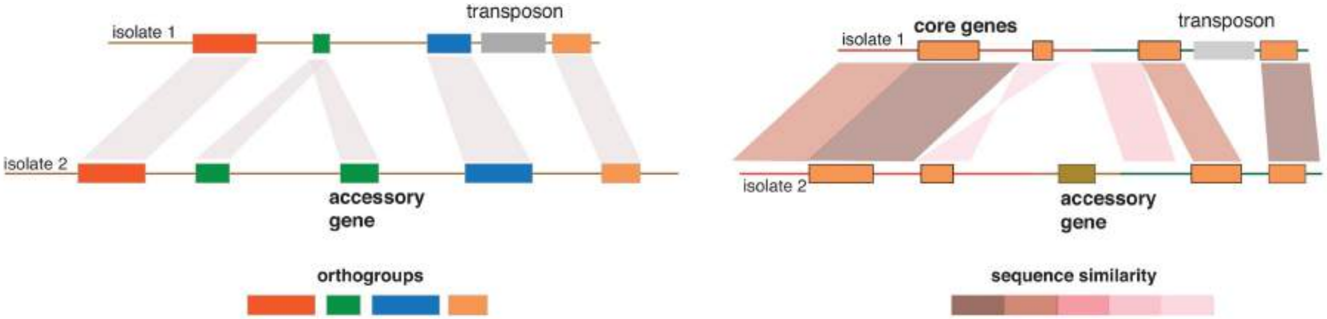
Hervé Tettelin
Duccio Medini *Editors*

The Pangenome

Diversity, Dynamics and Evolution of Genomes

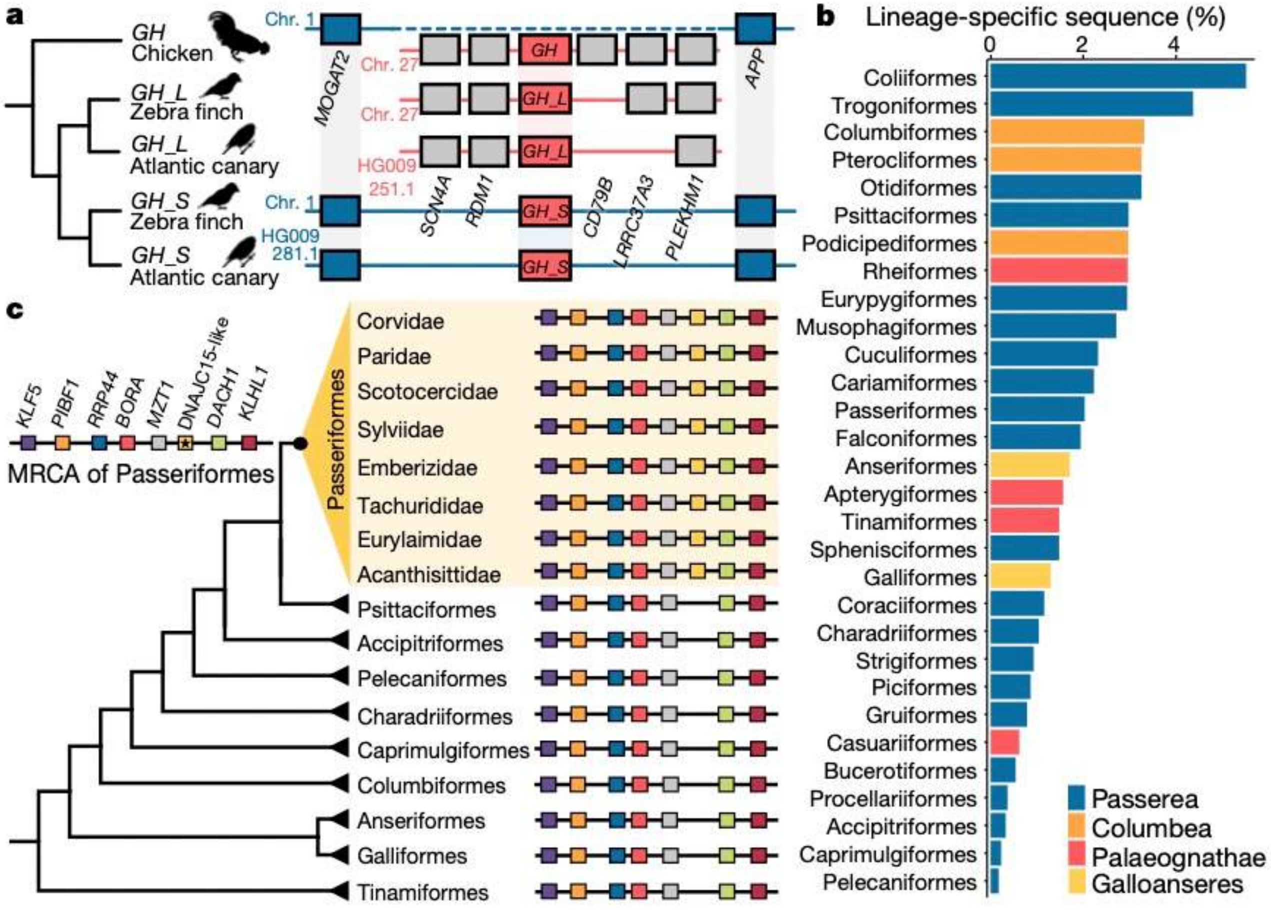
Springer

OPEN



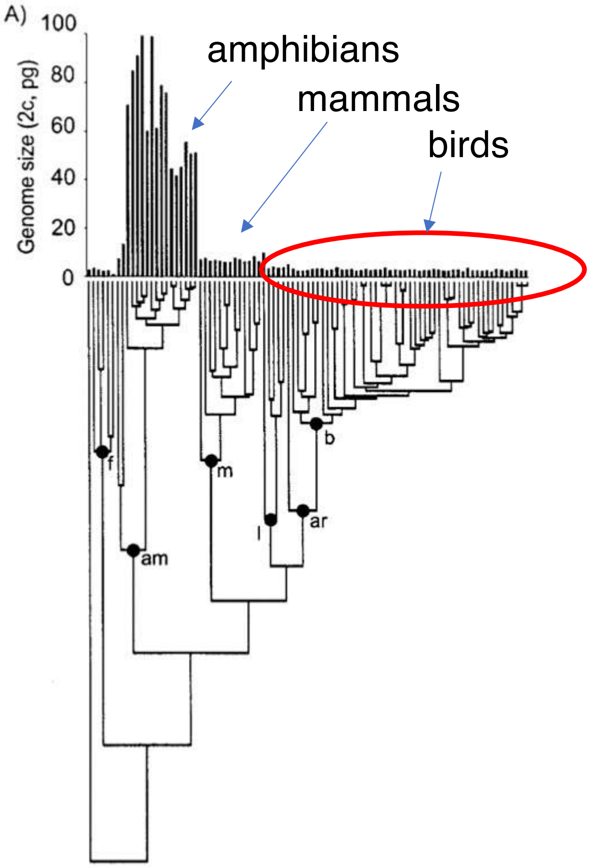
https://pathogen-genomics.org/research/

# Pangenome approach to comparative genomics

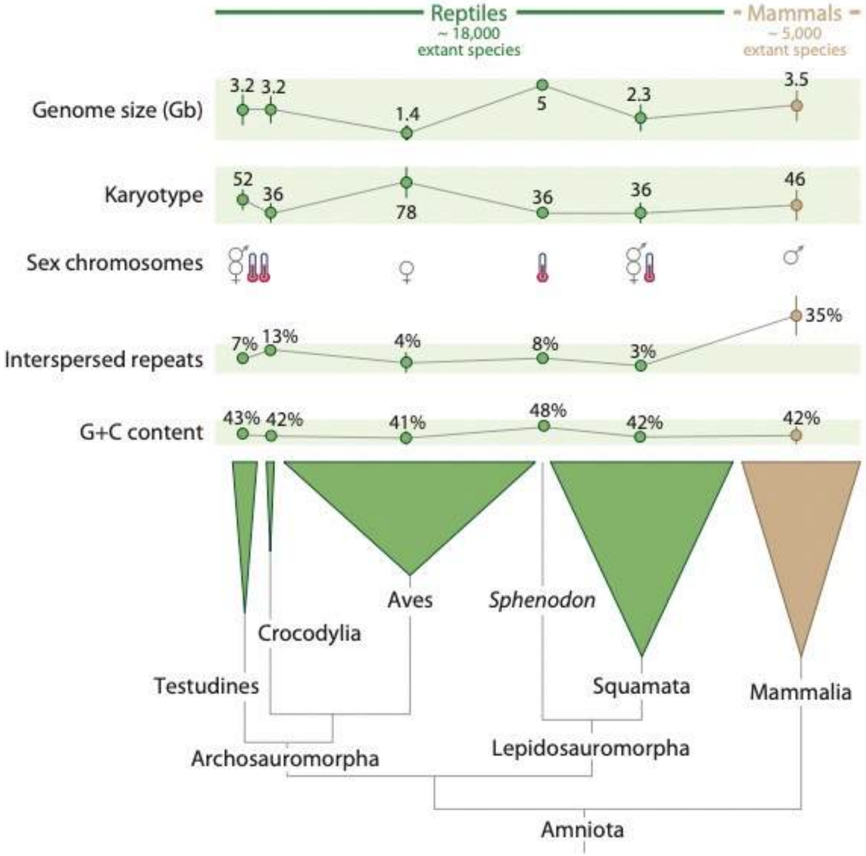


Feng et al. 2020. *Nature* 587:252-257.

# Birds have small, streamlined genomes

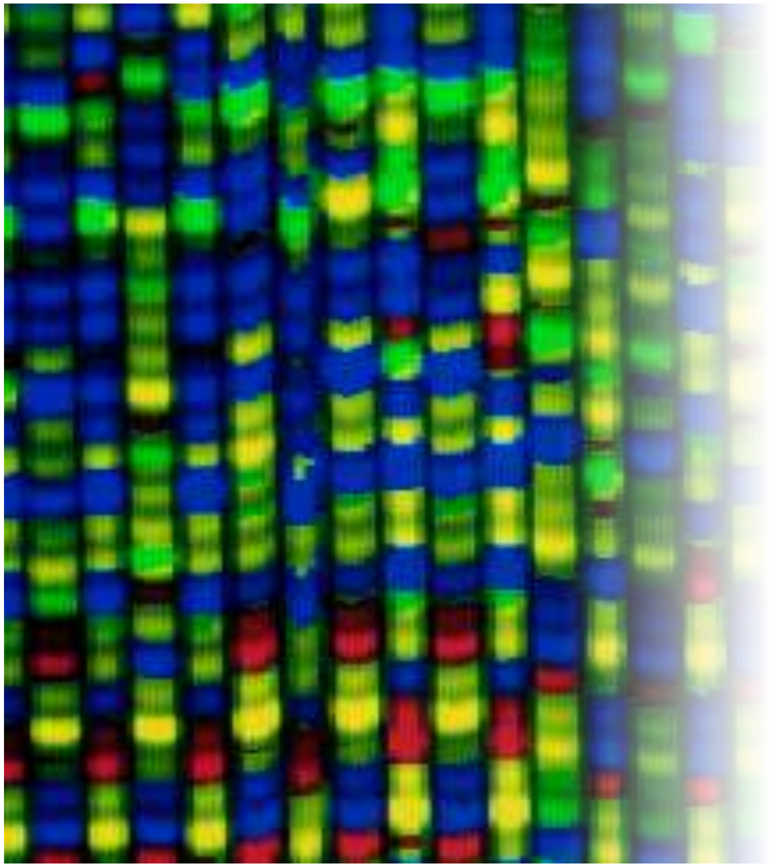


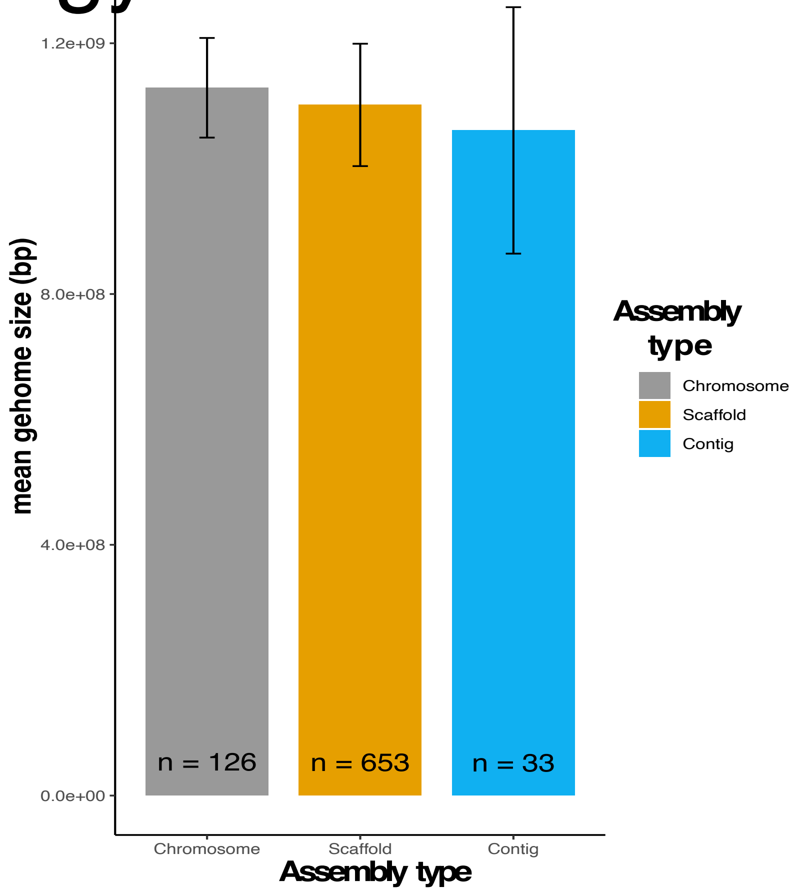
Waltari & Edwards. 2002. *Am. Nat.*



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

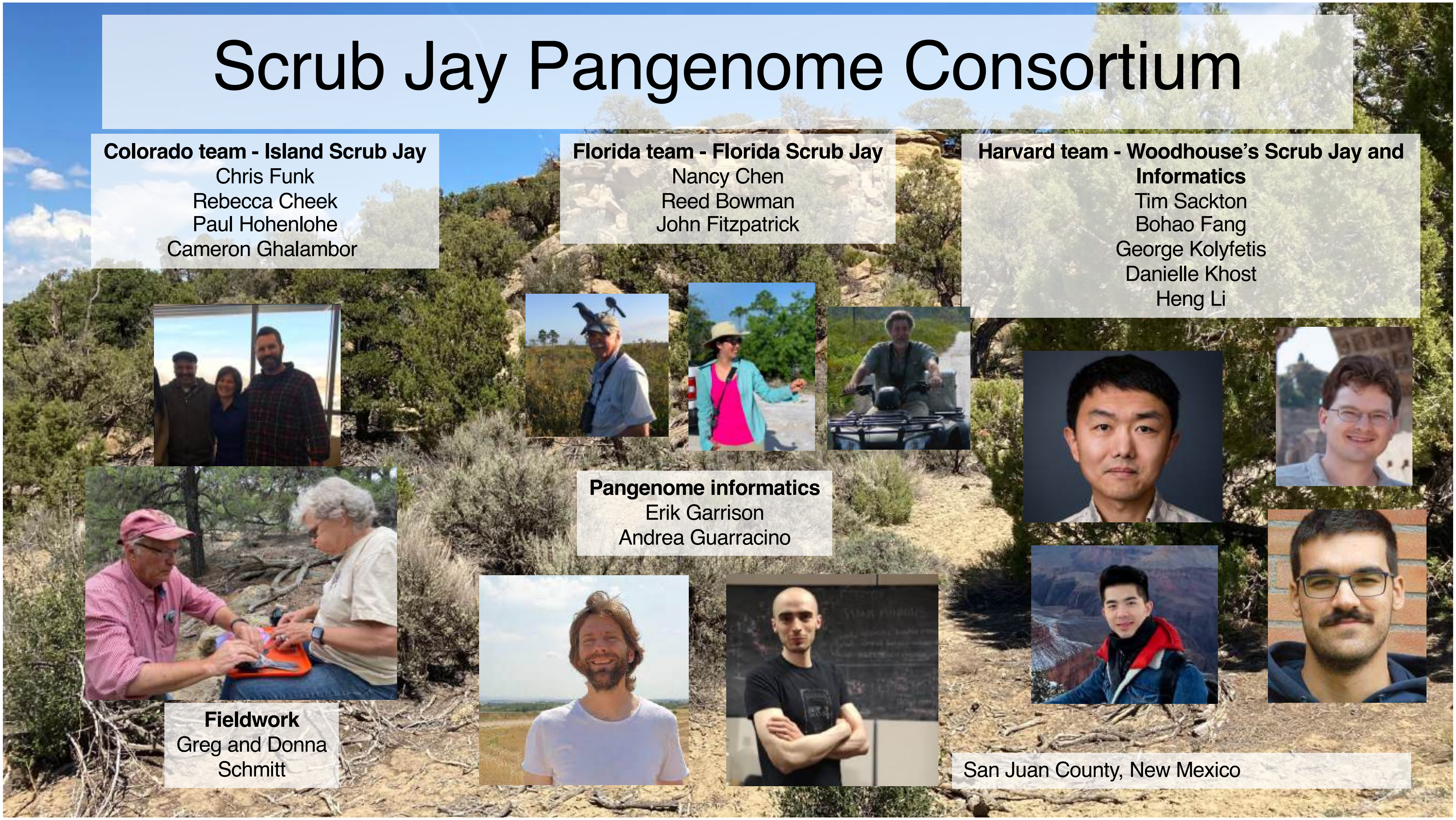
# Avian genomes are growing with each new technology





Data from NCBI, accessed 13 Nov. 2021

# Scrub Jay Pangenome Consortium



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

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

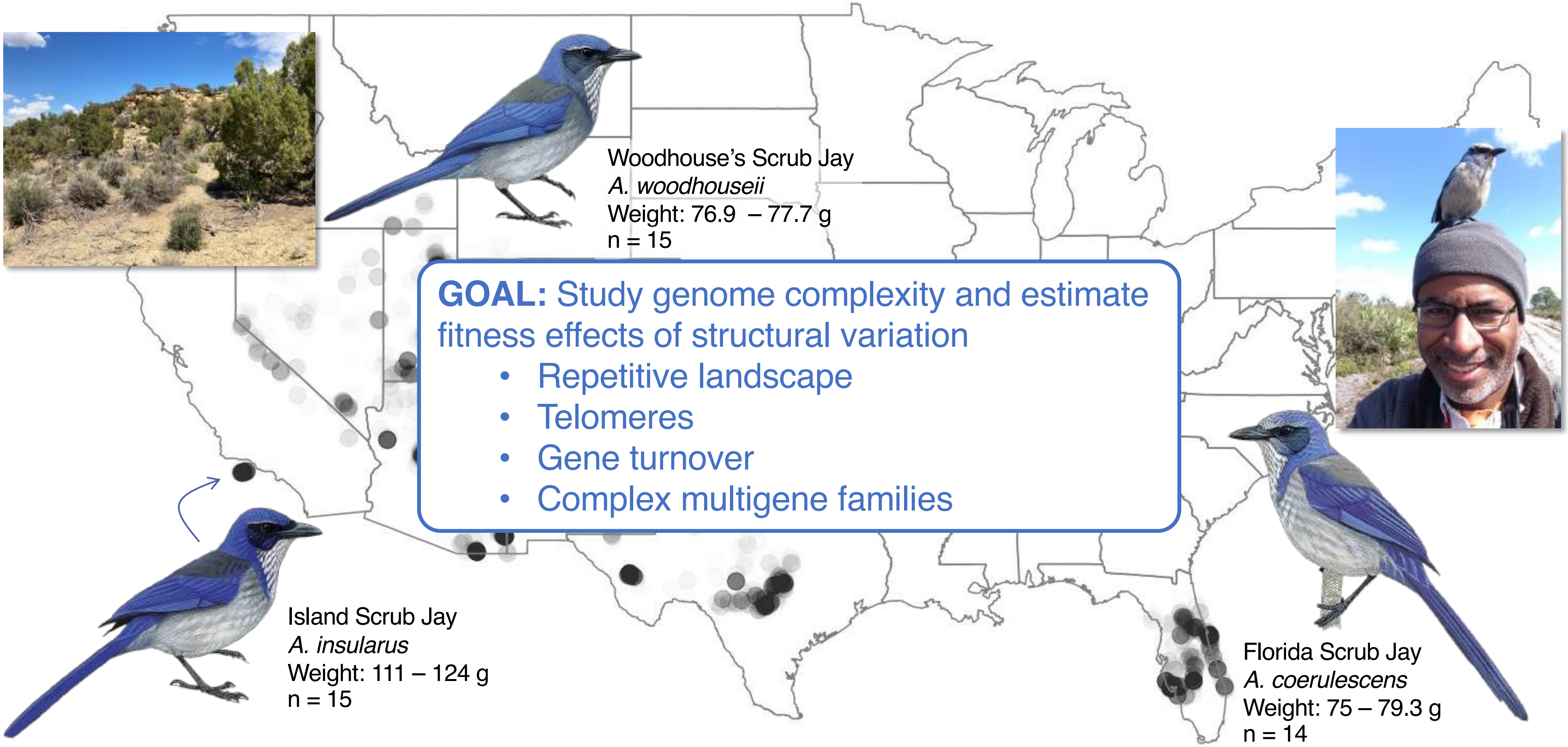
**Harvard team - Woodhouse's Scrub Jay and Informatics**
Tim Sackton
Bohao Fang
George Kolyfetis
Danielle Khost
Heng Li

**Pangenome informatics**
Erik Garrison
Andrea Guarracino

**Fieldwork**
Greg and Donna Schmitt

San Juan County, New Mexico

# Three Scrub Jay (*Aphelocoma*) species in pangenome project



Woodhouse’s Scrub Jay
*A. woodhouseii*
Weight: 76.9 – 77.7 g
n = 15

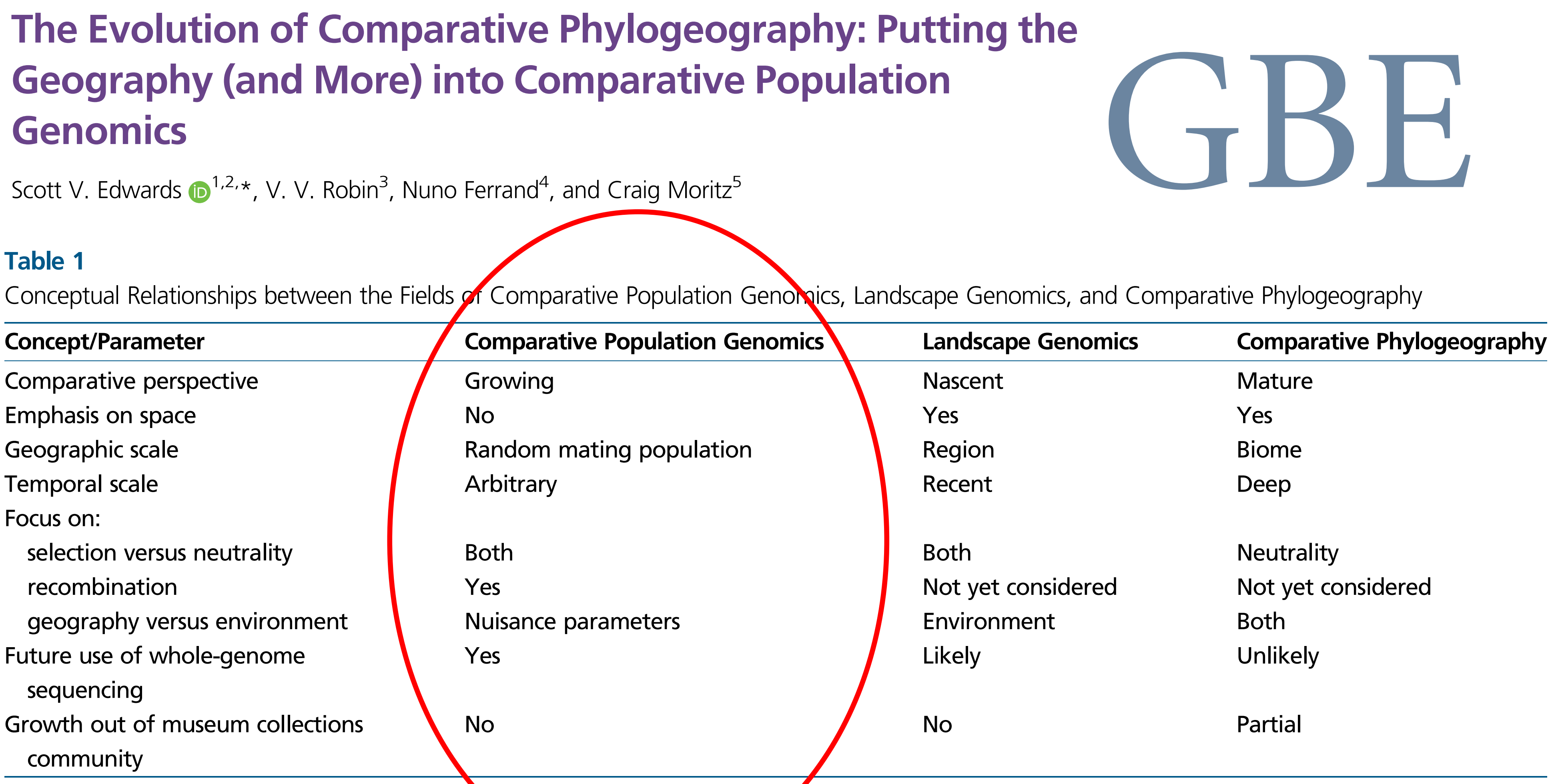
Island Scrub Jay
*A. insularus*
Weight: 111 – 124 g
n = 15

Florida Scrub Jay
*A. coerulescens*
Weight: 75 – 79.3 g
n = 14

**GOAL:** Study genome complexity and estimate fitness effects of structural variation

- Repetitive landscape
- Telomeres
- Gene turnover
- Complex multigene families

Datapoints from gbif.org
Images from [https://birdsoftheworld.org](https://birdsoftheworld.org)



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

GBE

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

**Table 1**

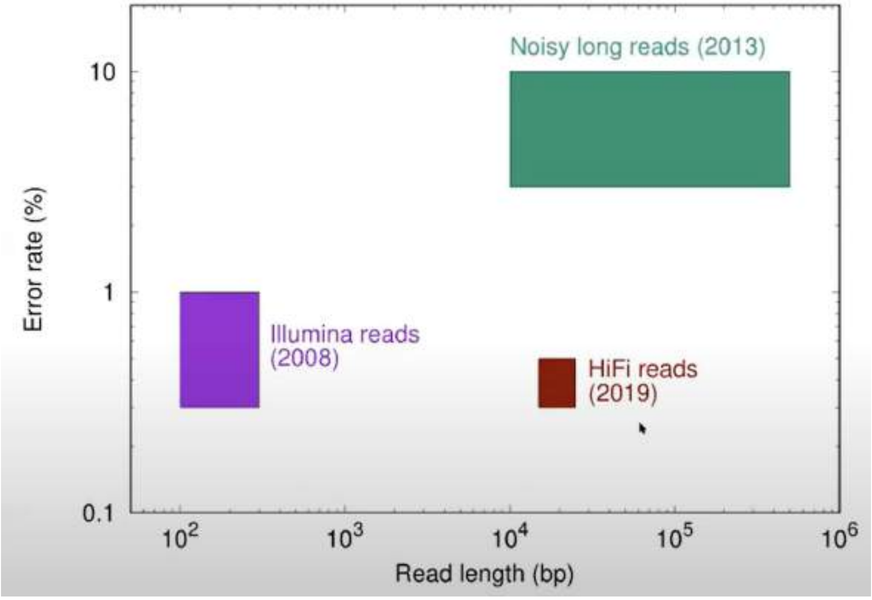
Conceptual Relationships between the Fields of Comparative Population Genomics, Landscape Genomics, and Comparative Phylogeography

| 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

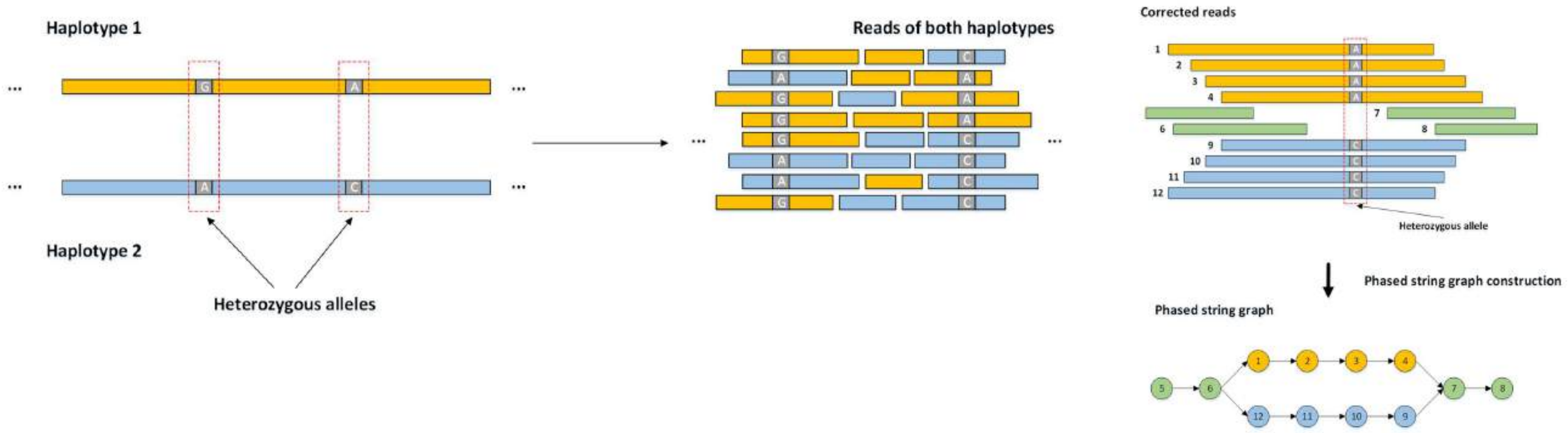
# PacBio HiFi reads are long <u>and</u> accurate

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



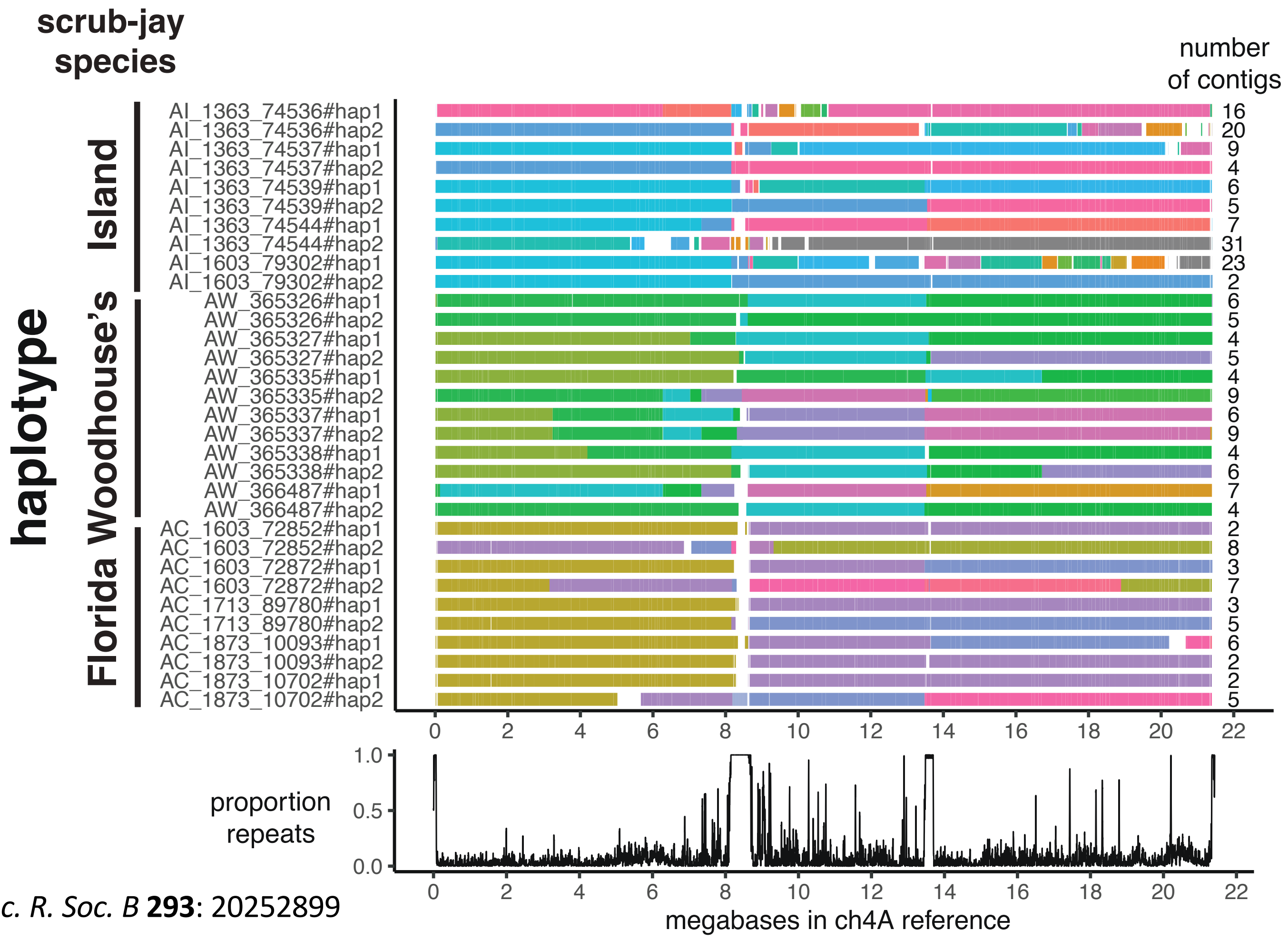
Coutesy Haoyu Cheng, Dana Farber Cancer Institute

# Hifiasm – a HiFi accurate read assembler that resolves haplotypes



Coutesy Haoyu Cheng, Dana Farber Cancer Institute

# Partially phased haplotypes span megabases



Edwards & Li. 2026 *Proc. R. Soc. B* **293**: 20252899

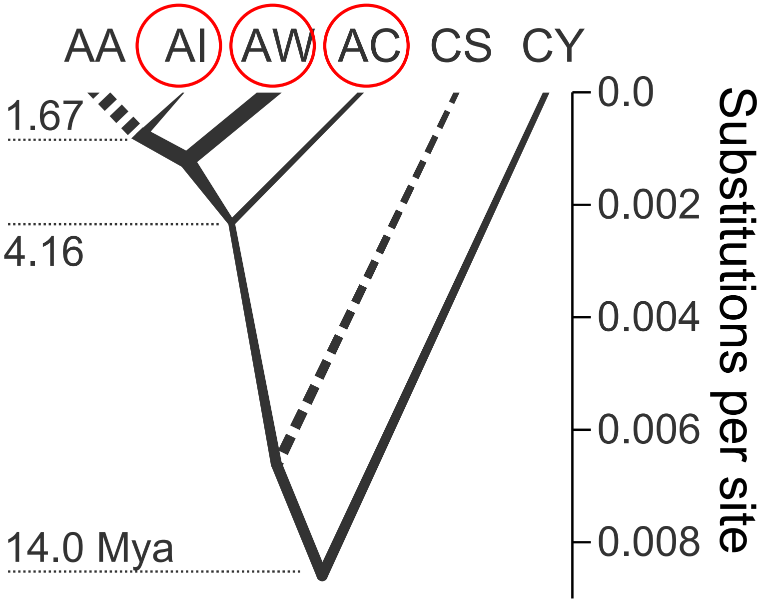
# Sampling of scrub-jay species and outgroups

**Core species**

**AI** = *A. insularis* – 30 haplotypes
**AW** = *A. woodhouseii* – 30 haplotypes
**AC** = *A. coerulescens* – 28 haplotypes

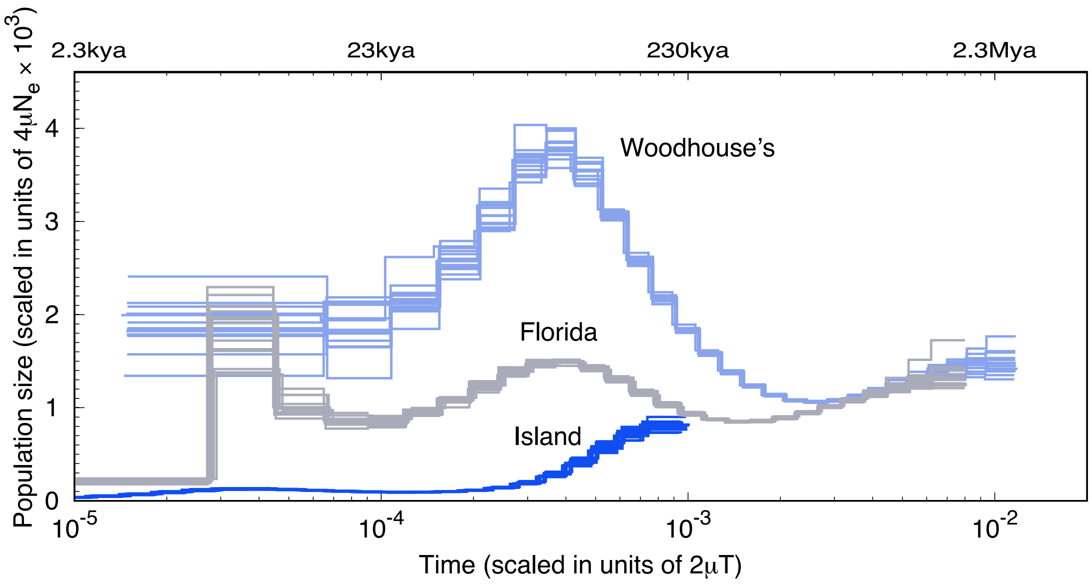
**Outgroups and additional species**

**AA** = *A. californica* – 2 haplotypes
**CS** = *C. stelleri* – 2 haplotypes
**CY** = *Cyanocorax yucatanicus* – 2 haplotypes



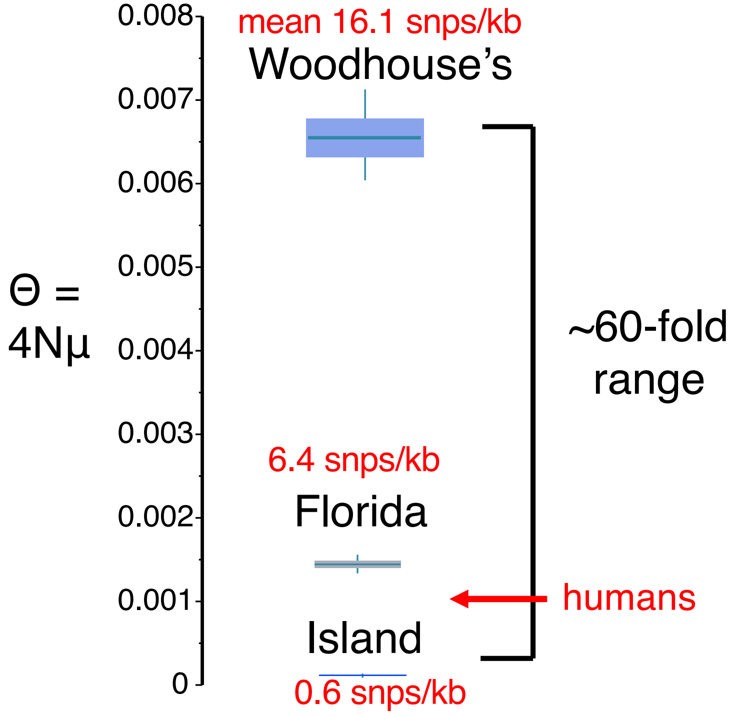
~2500 dipcall loci fitted to tree with bpp
Flouri et al. 2020. *Mol. Biol. Evol.* 37: 1211-1223

# 60-fold range in effective population size across species



**PSMC**

SNP data from genome-wide dipcall regions

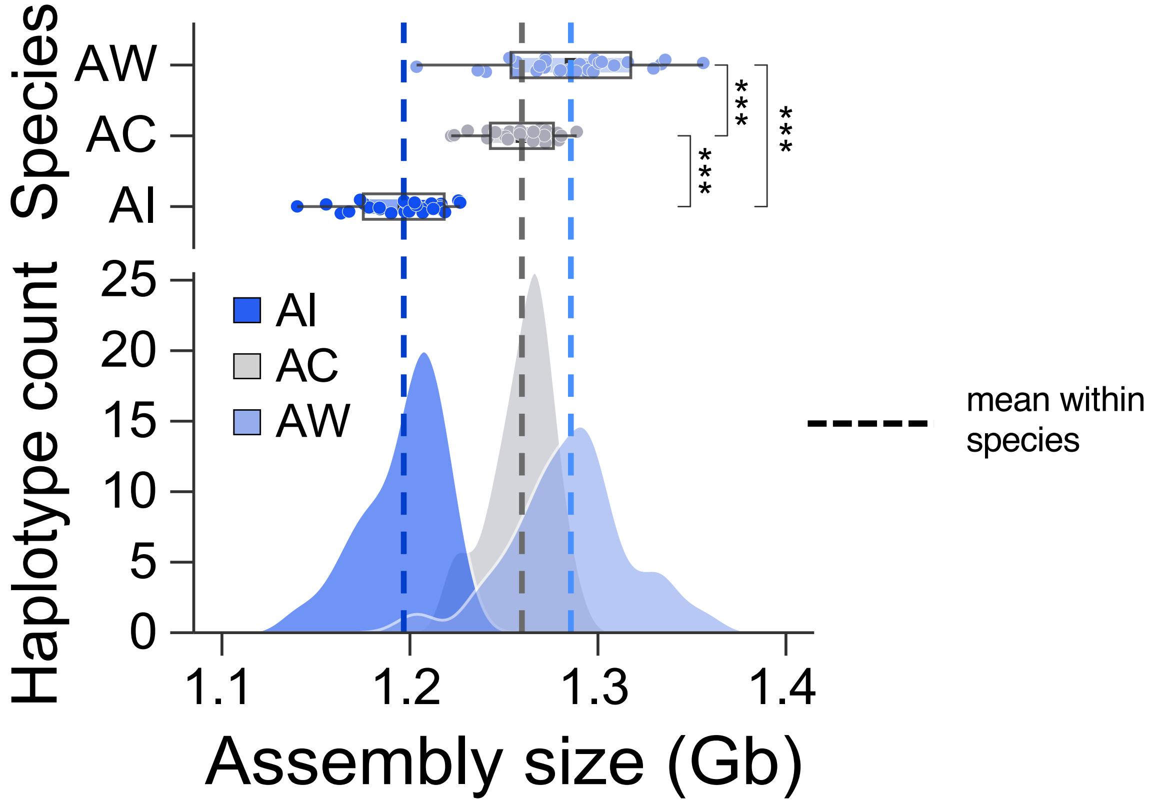


**bpp**

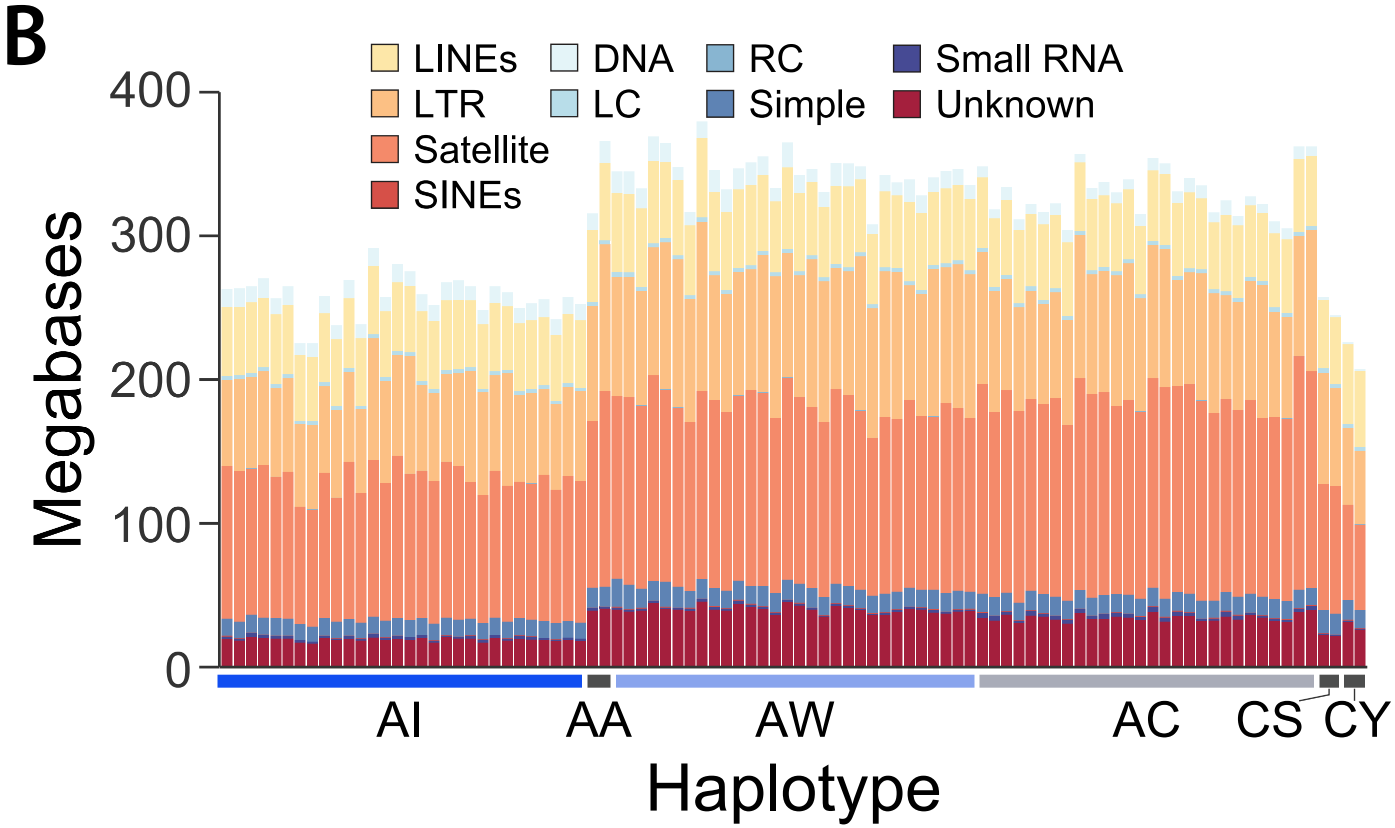
~2500 1-kb dipcall regions

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

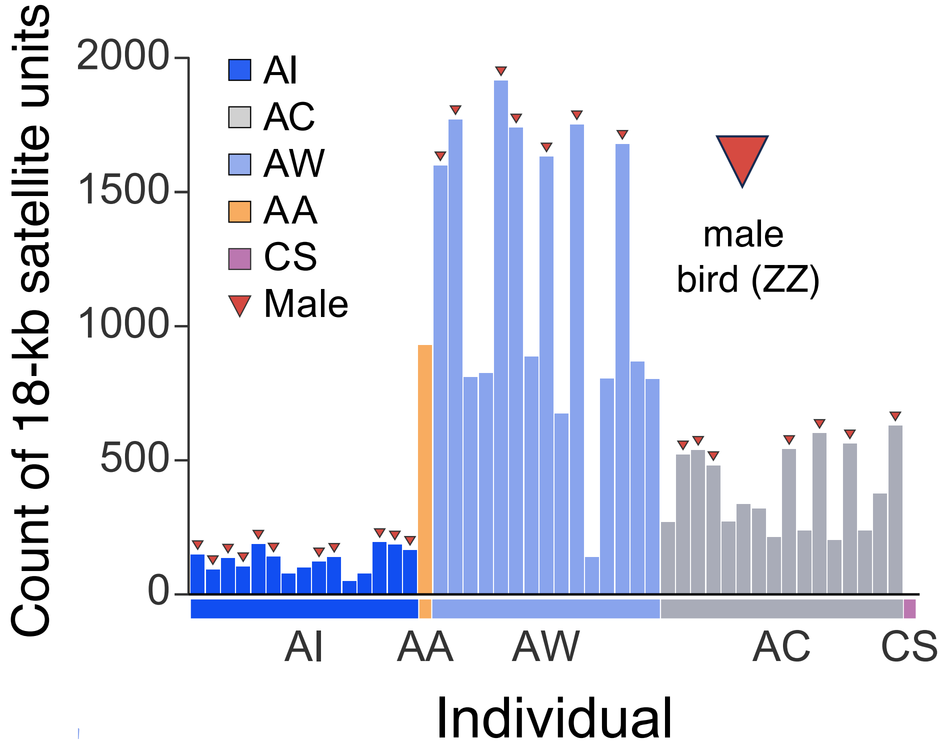
# Assemblies of Island Scrub-Jays are ~80 Mb smaller than other species

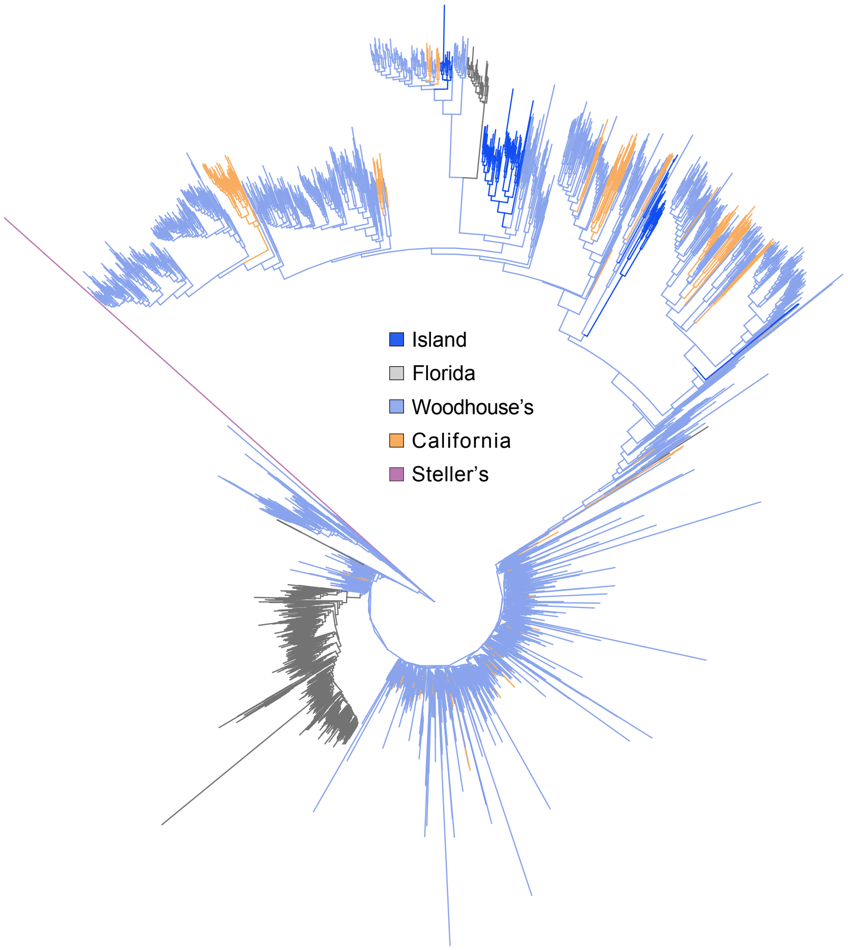


# Interspecific variation in repeat content…



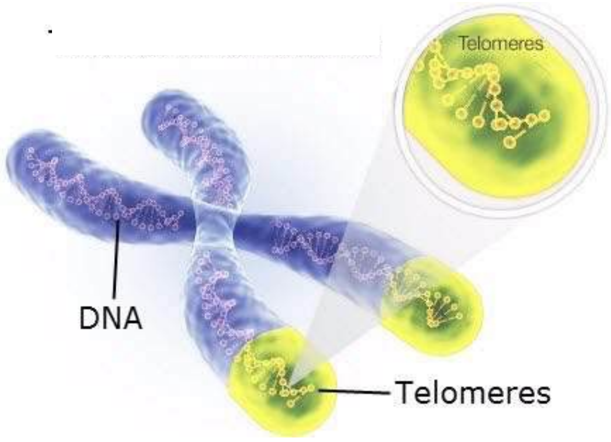
# …and satellites

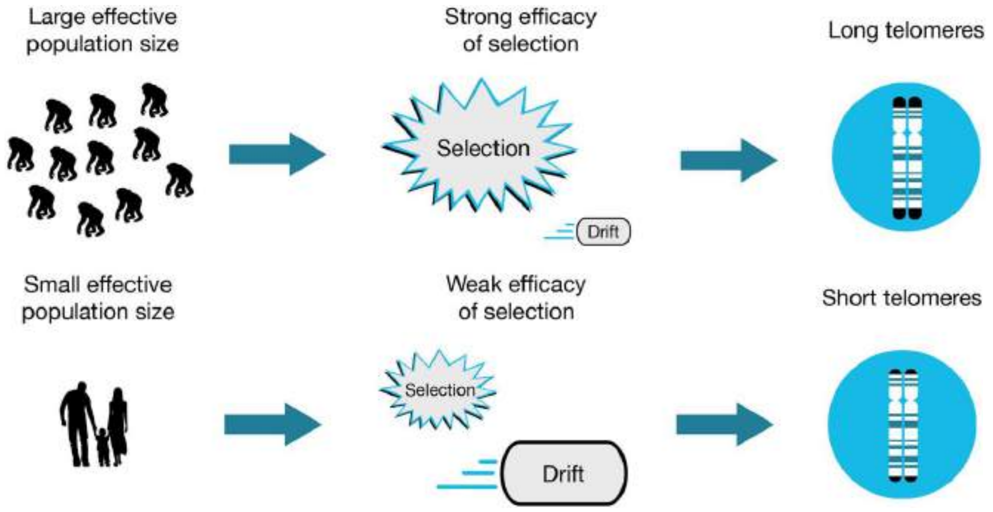




IQ-TREE v. 1.6.12, 3500 18-kb satellite sequences

# Telomeres predicted to be shorter in species with small $N_e$




Large effective population size

Strong efficacy of selection

Long telomeres

Small effective population size

Weak efficacy of selection

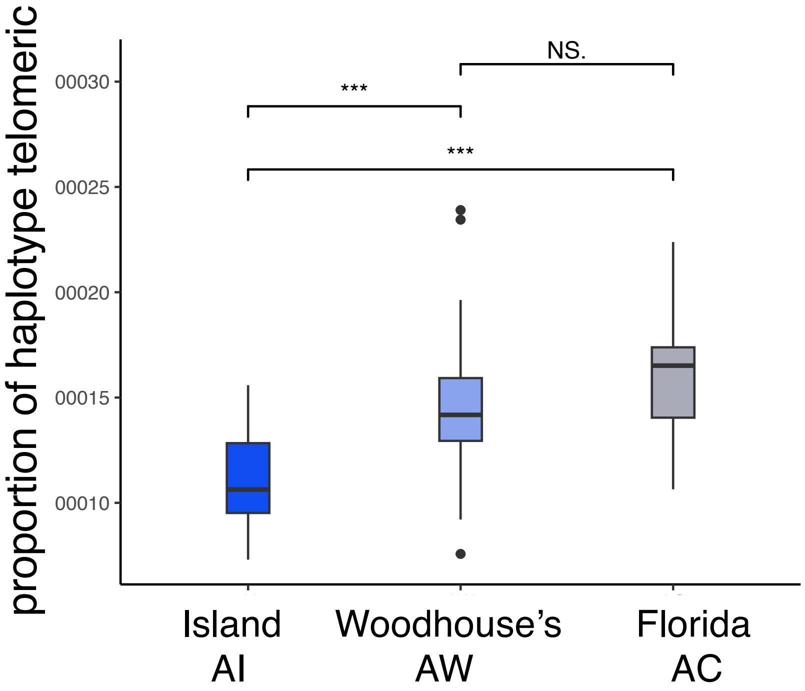
Short telomeres

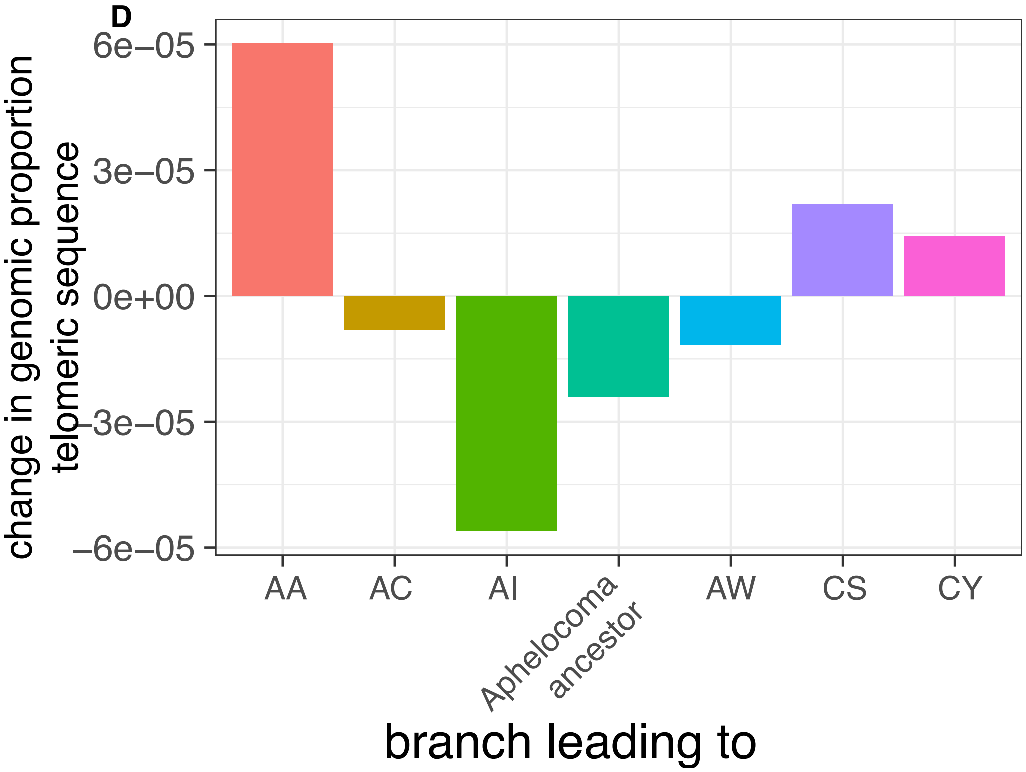
Brown LM, Elbon MC, Bharadwaj A, Damle G, Lachance J. 2024. *Genome Biol Evol*: 16.

https://medibalans.com/telomere/

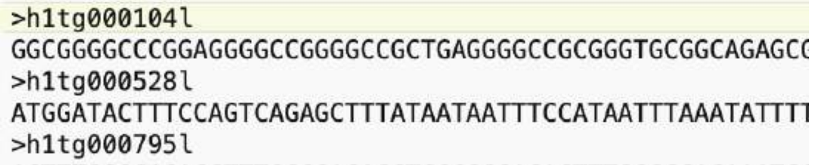
# Telomere abundances influenced by population size

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
GGCGGGGCCCGGAGGGGCCGGGGCCGCTGAGGGGCCGCGGGTGCGGCAGAGCC
>h1tg000528l
ATGGATACTTTCCAGTCAGAGCTTTATAATAATTTCCATAATTTAAATATTT
>h1tg000795l
ACTTTGGGGACACCTTTGGGGACACCTCGGGGGACACTTTGGGCCACAAATC
```

unaligned fasta files

→

**minigraph**

**odgi**

build genome graph

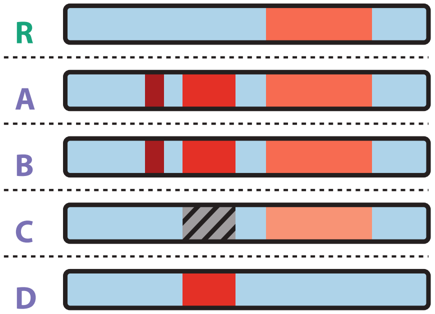
→

**odgi viz**

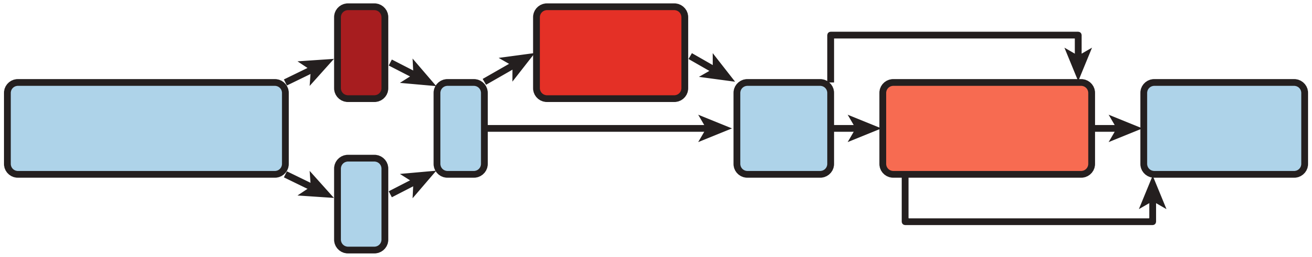
**bandage**

visualize genome graph

→



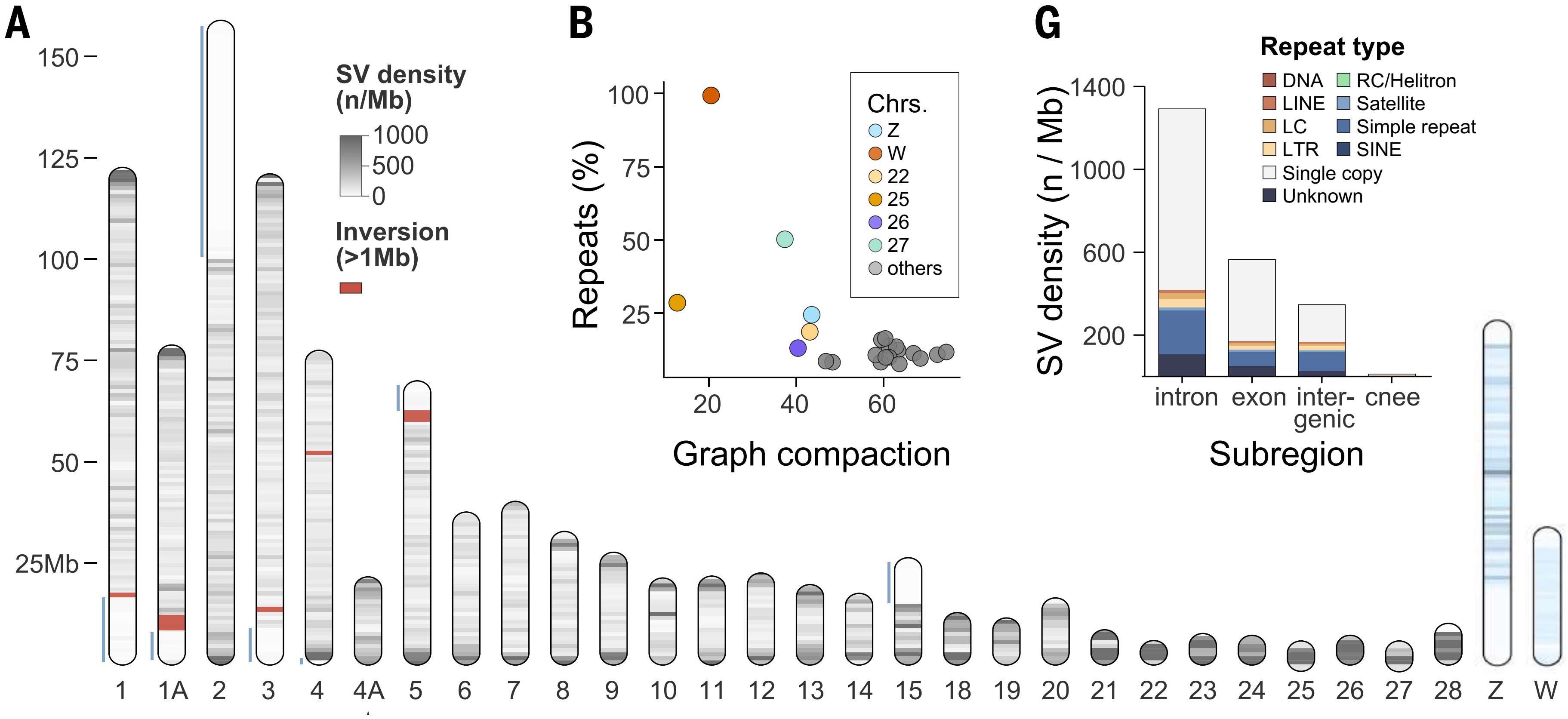
**Multiple sequence alignment**



**Bidirected genome graph**

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

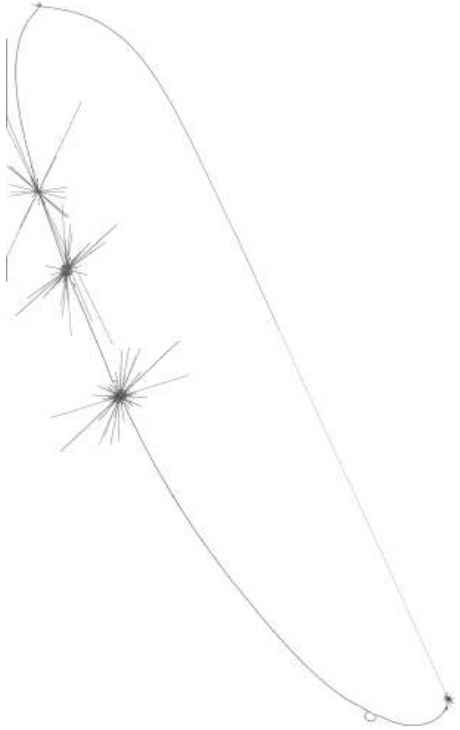
# Distribution of genome-wide structural variants



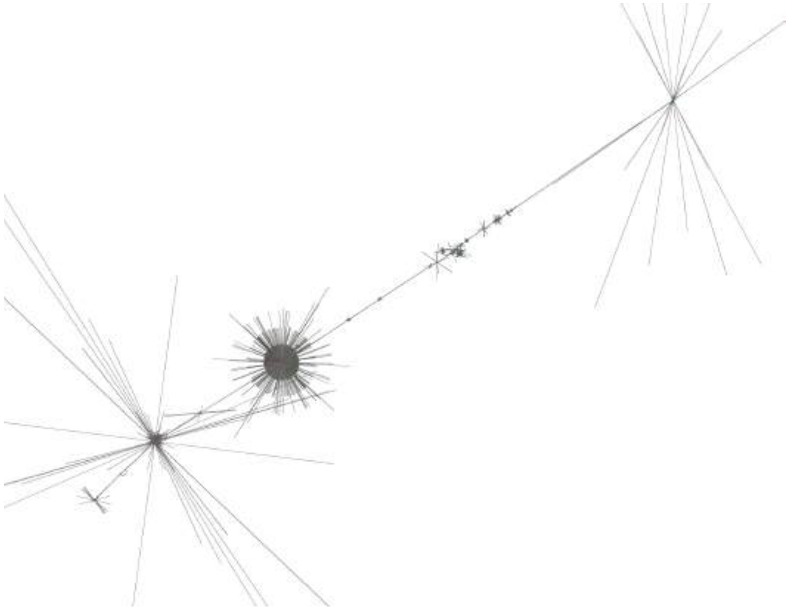
# 2D pangenome graph visualizations – PGGB/Odgi

Chr 18 – 12 Mb

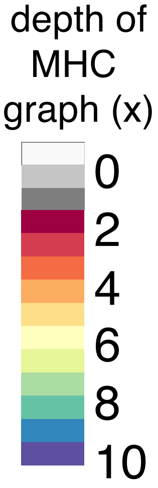
Chr 1 – 160 Mb
‘telomere kiss’



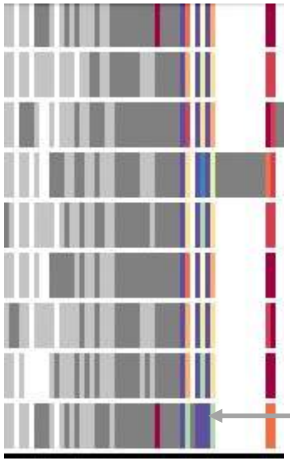
Chr 23 – 8 Mb



# Variation in depth of a pangenome graph



<u>medium</u>
(self-)node-depth regions
depth = ~ n haplotypes

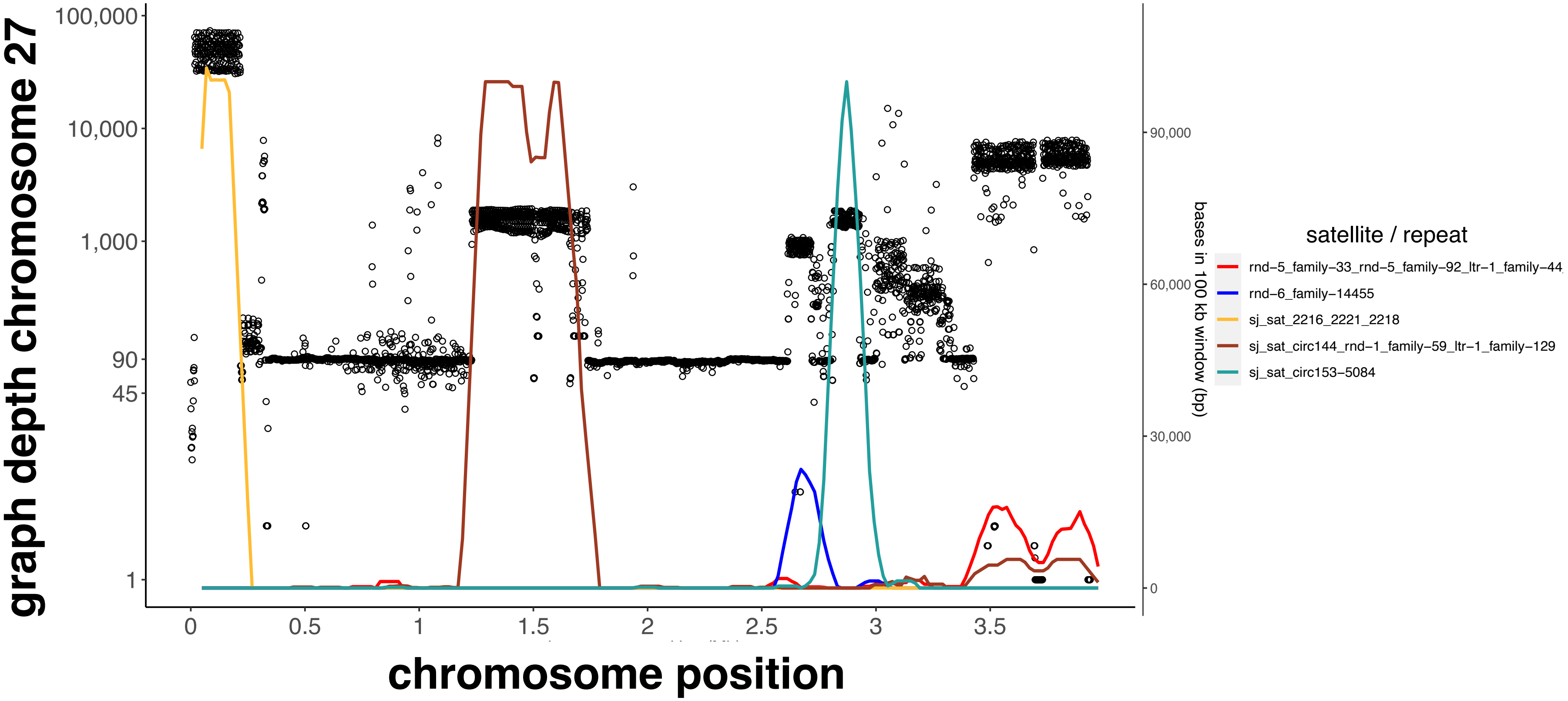


```
GGGGATGATTTTGGGAGGTTTGGGGTTGATTTAAAGACATTTTGGGATATTT
TGAGGGGATTTGGGGACATTTTGAGGGGTTTGGGGACACTTTGGGGACAATT
TTTGGGAGGTTTGGGGCTGATTTTGAGAGGTTTGGATCCAACTTTGGAGCAT
TTTGAGGGGGTTTGGGGCCACTTTGGGGCCACTTTGGTGGCCTCAGTGCCCC
TTGGGGACACCTGGAGGGGTCTGGGGACAACTTGGGGTGCTTGGGAACACTT
TAAGGATATTTTGGGACATTTTGAGGGGTTTGGGGACACTTTGGGGACAATT
```

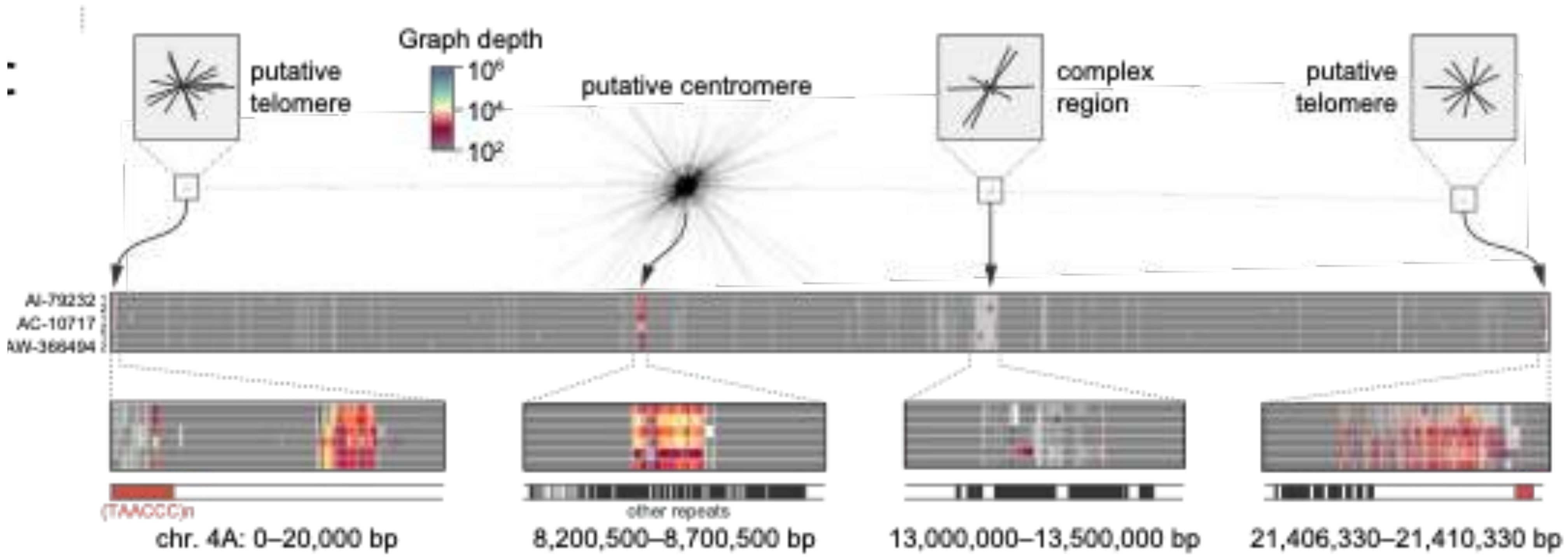
<u>low</u> node-depth
regions – SNPs
or small
indels

<u>high</u> node-depth regions –
large structural variants, satellites

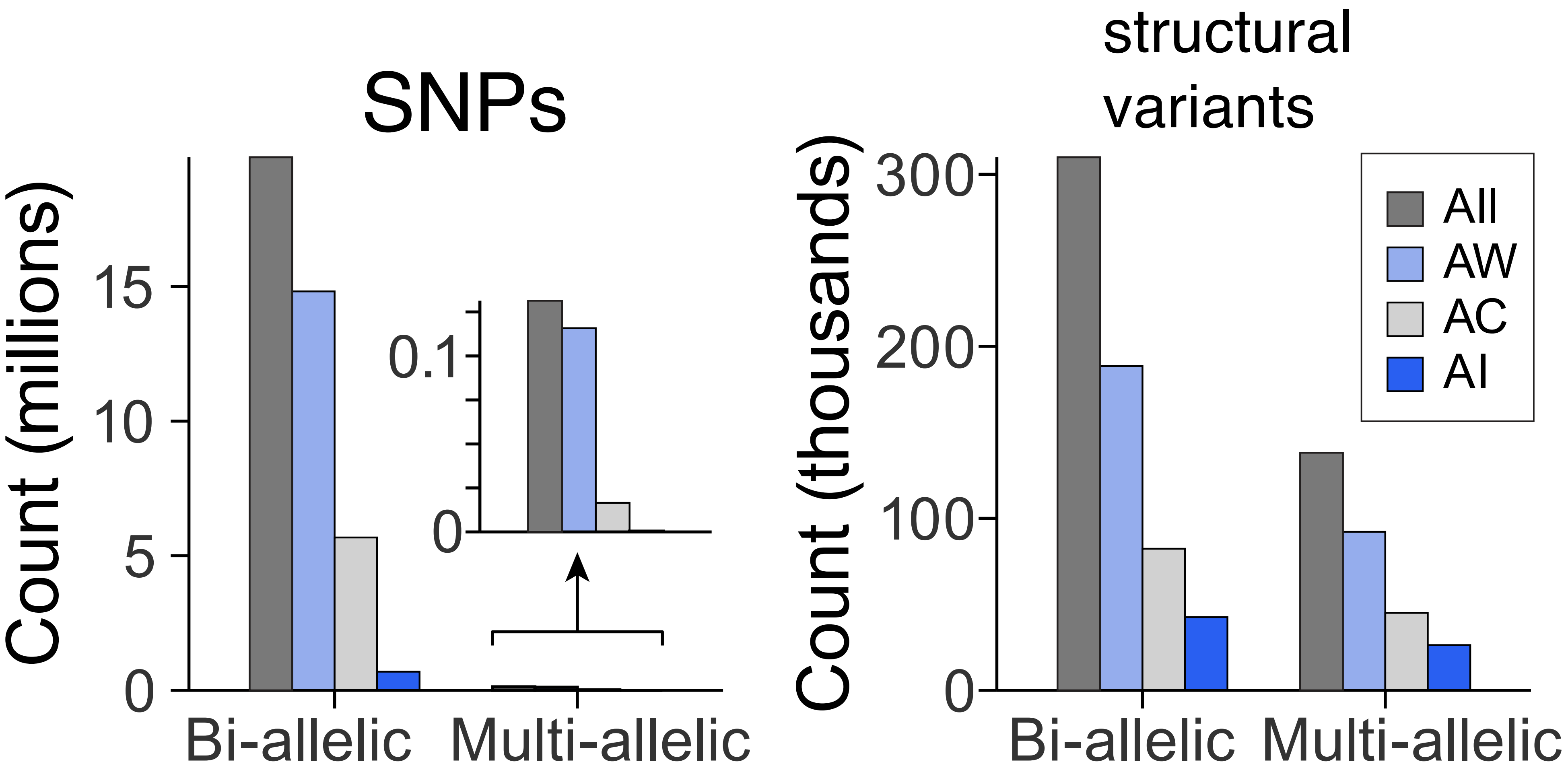
# Graph depth of microchromosome 27 correlates with LTRs and satellites



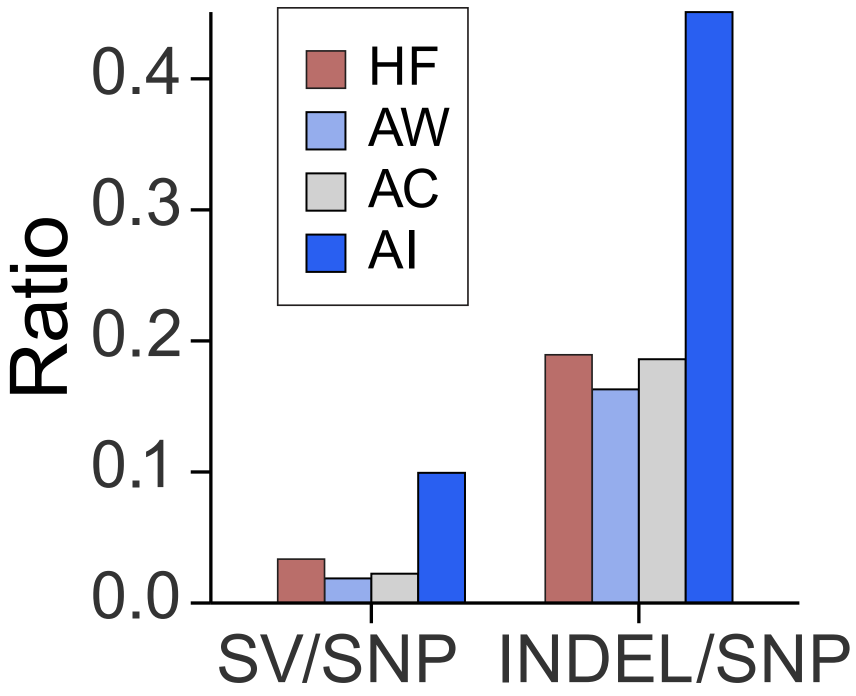
# Pangenome depth reveals chromosomal features

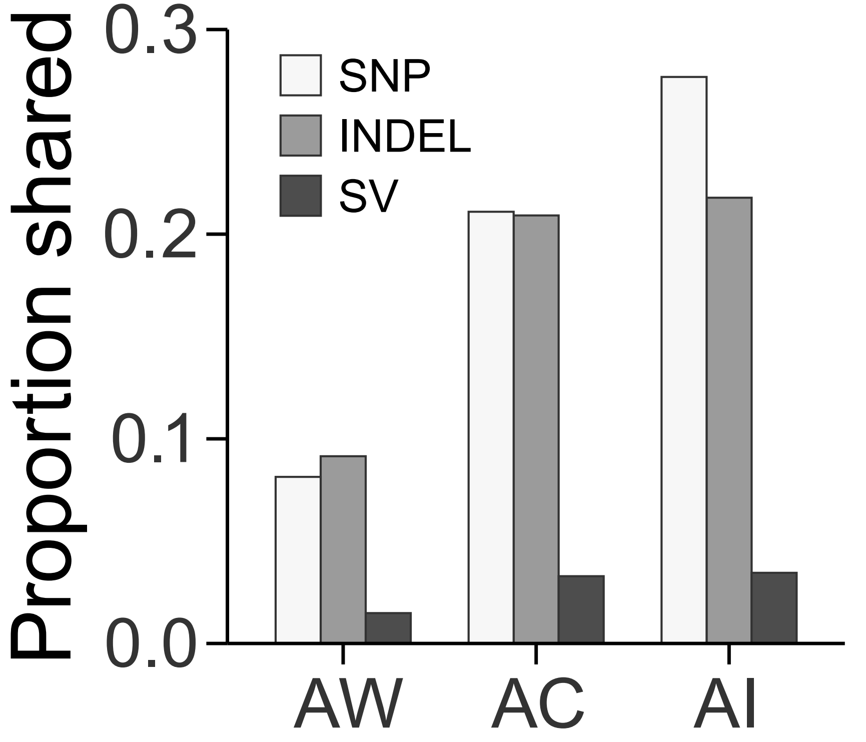


**Number of structural variants scales with population size**

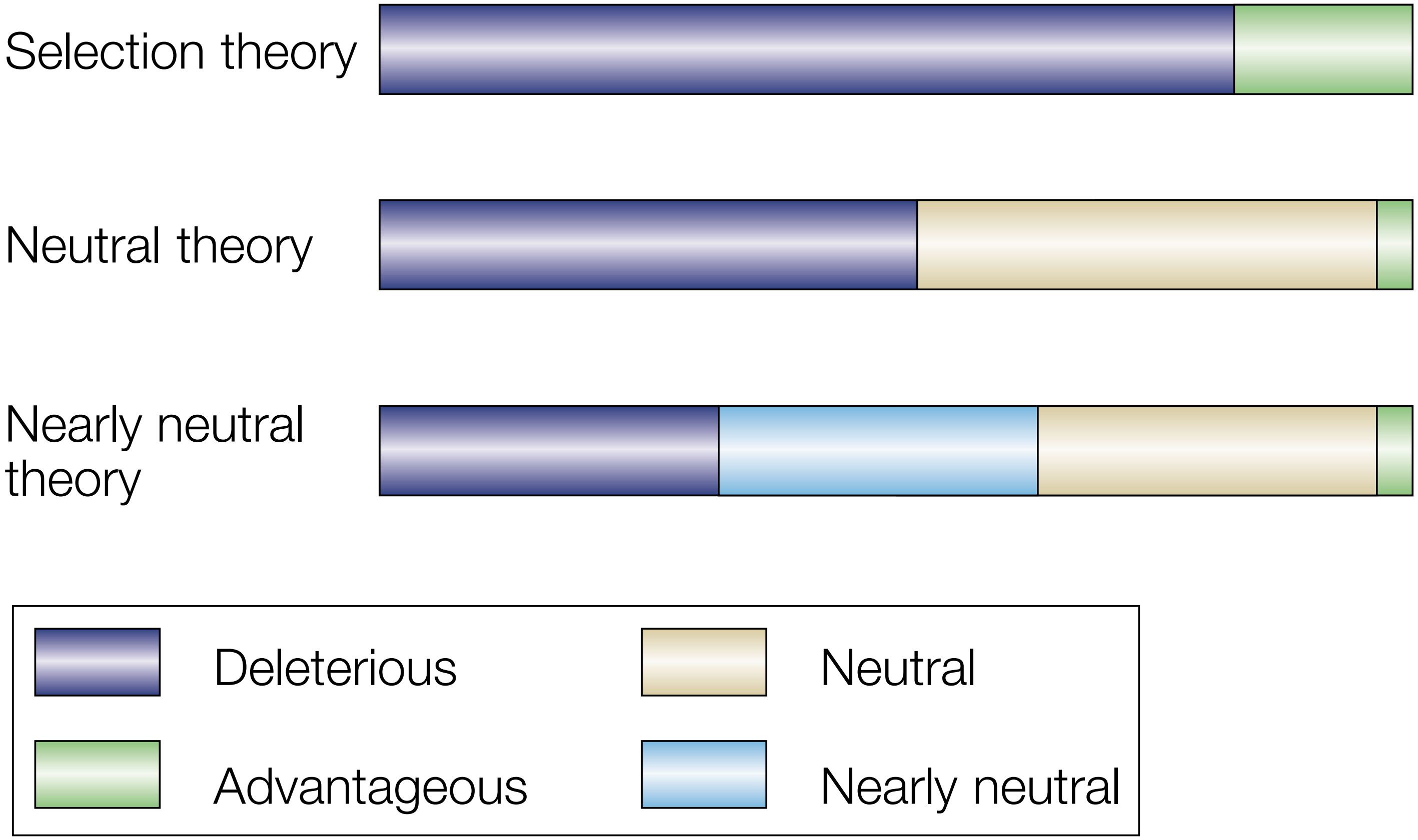


# Indirect evidence for an interaction between selection and genetic drift for structural variants


ratio of SVs to SNPs shifts
in small population (AI)

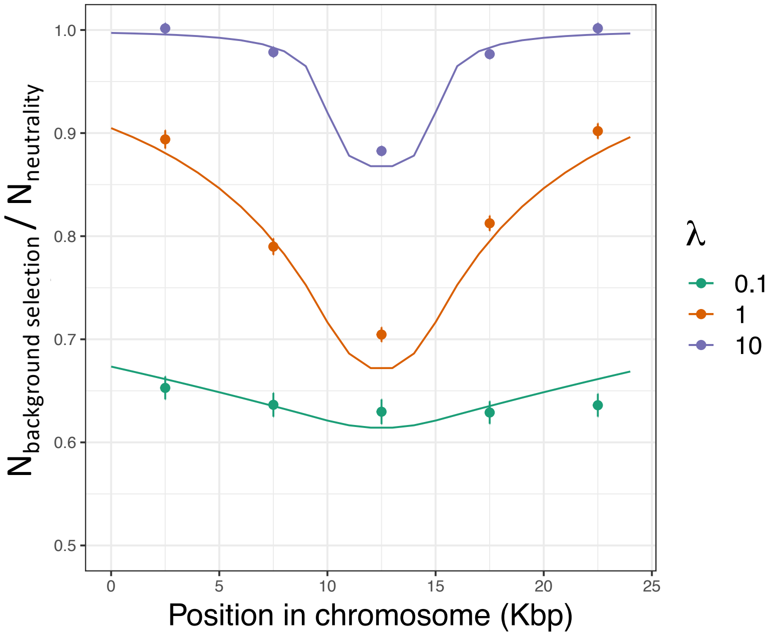

proportion of variants shared between species
differs for SVs and SNPs

# Nearly neutral molecular evolution



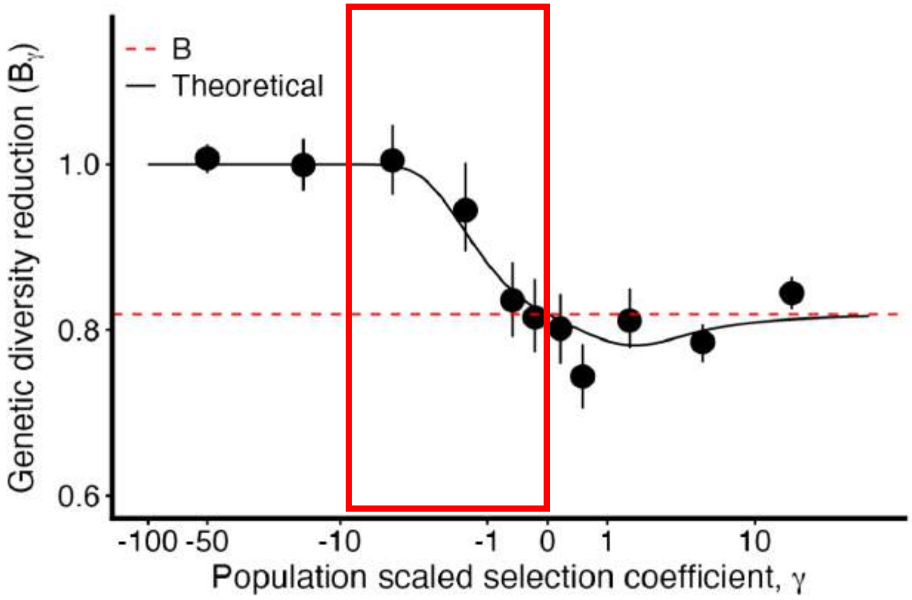
Bromham and Penny. 2003. *Nature Reviews Genetics*

# Weakly deleterious mutations reduce nearby genetic diversity



λ = change in recombination rate

Booker et al. 2022. *Proc. R. Soc. Lond. B*. 289:20220782



Li & Berg. 2026. PNAS 123:e2513613123

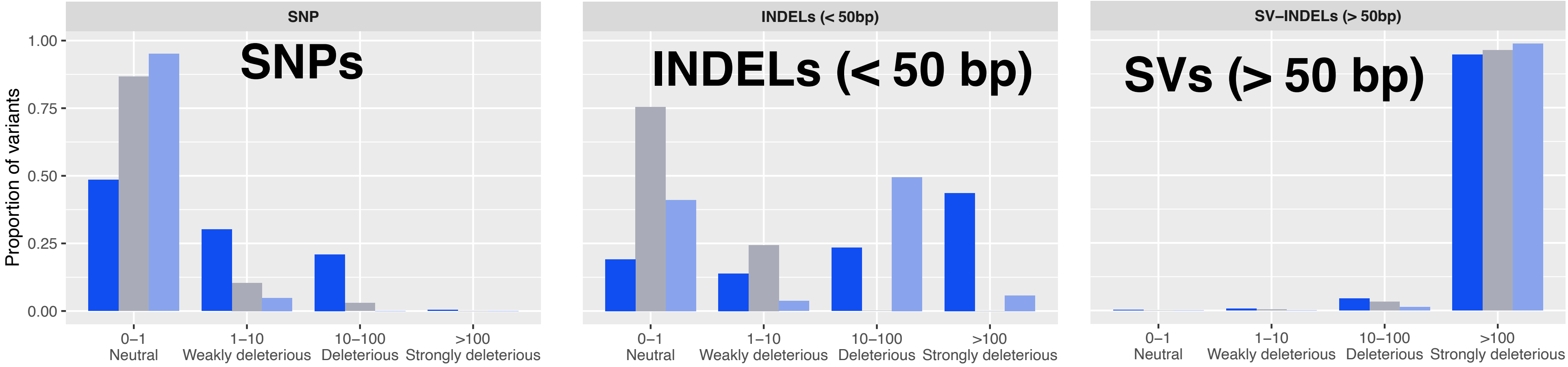
# SVs are on average more deleterious than SNPs

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.

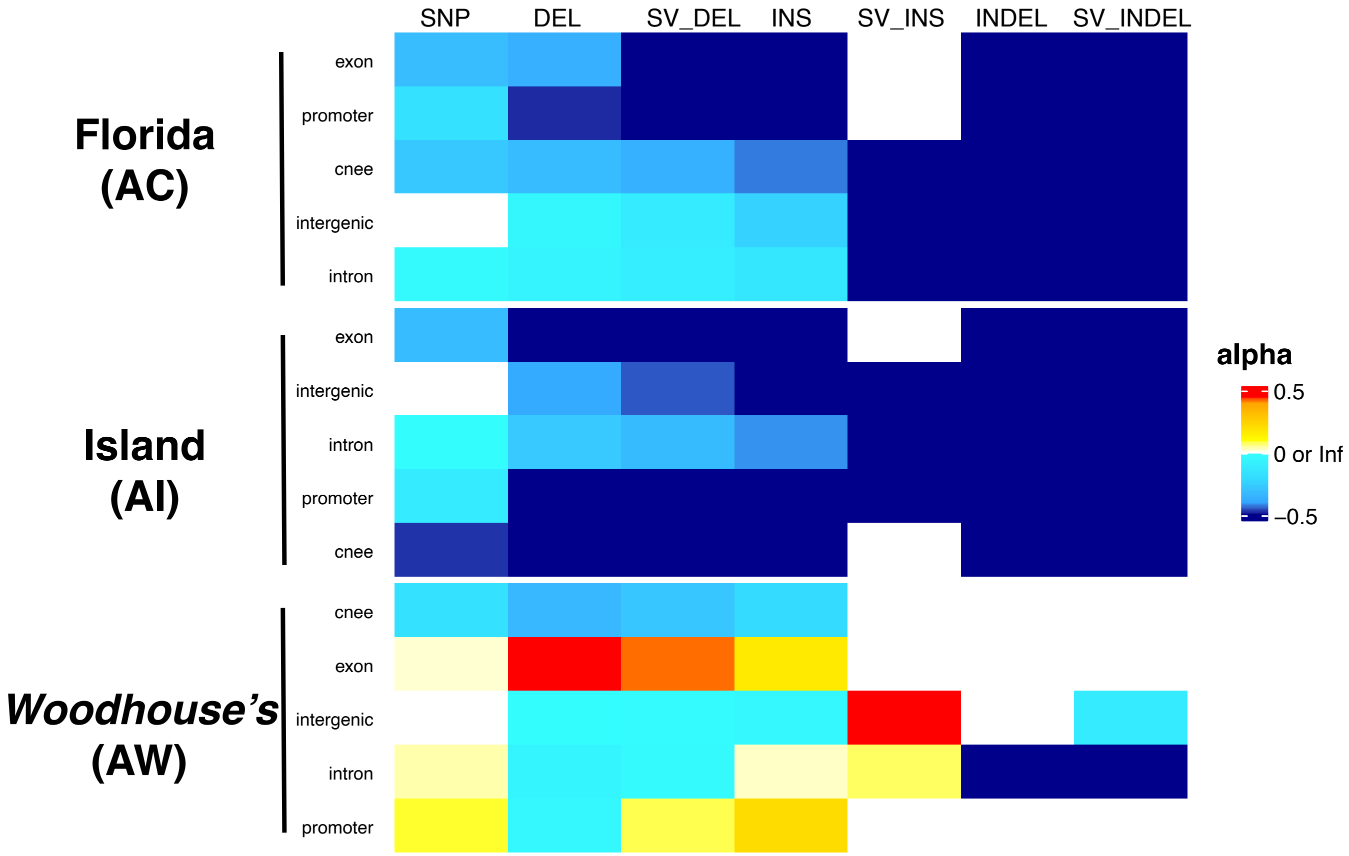
# Estimating the fraction of sites fixed adaptively

$$\alpha = 1 - \frac{D_s P_n}{D_n P_s} \qquad \mathrm{DoS} = D_n/(D_n + D_s) - P_n/(P_n + P_s).$$

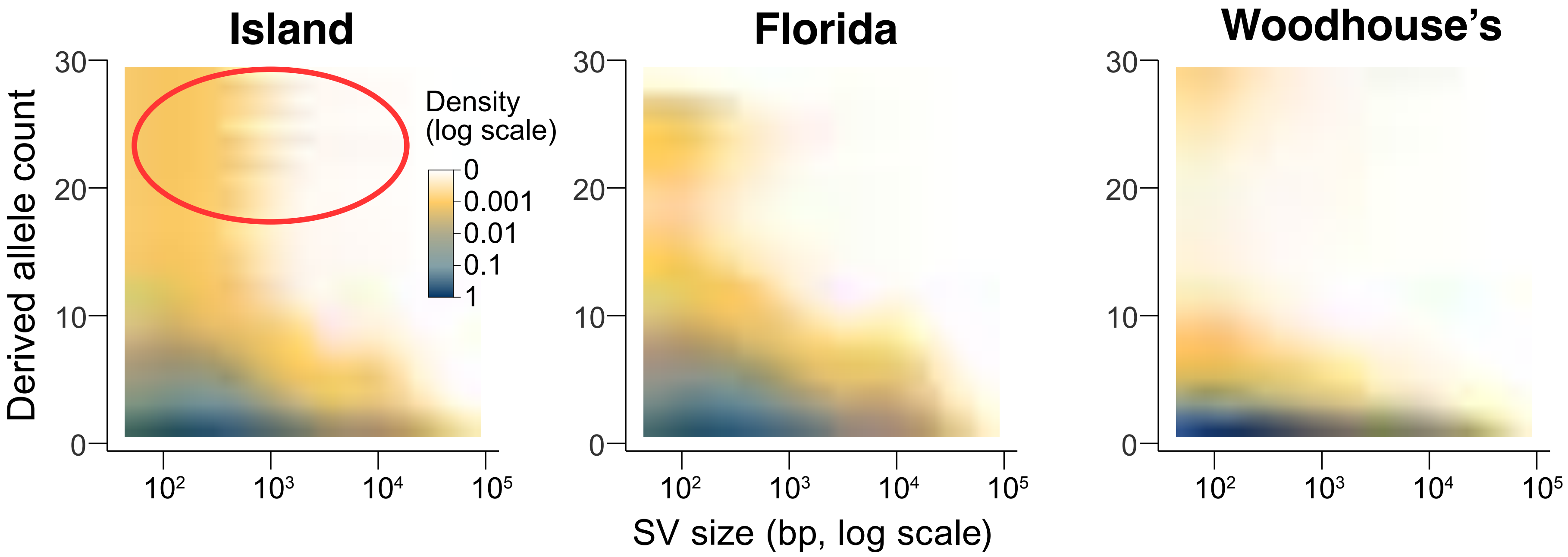
- **Intergenic SNPs** used as the neutral standard ($D_s$, $P_s$)
- **SVs** used as non-neutral test cases ($D_n$, $P_n$)

Smith NG, Eyre-Walker A. 2002. ***Nature***: 415:1022-1024.
Stoletzki N, Eyre-Walker A. 2011. ***Mol. Biol. Evol***. 28:63-70

# Evidence of adaptive fixation only in species with large population size (*A. woodhouseii*)

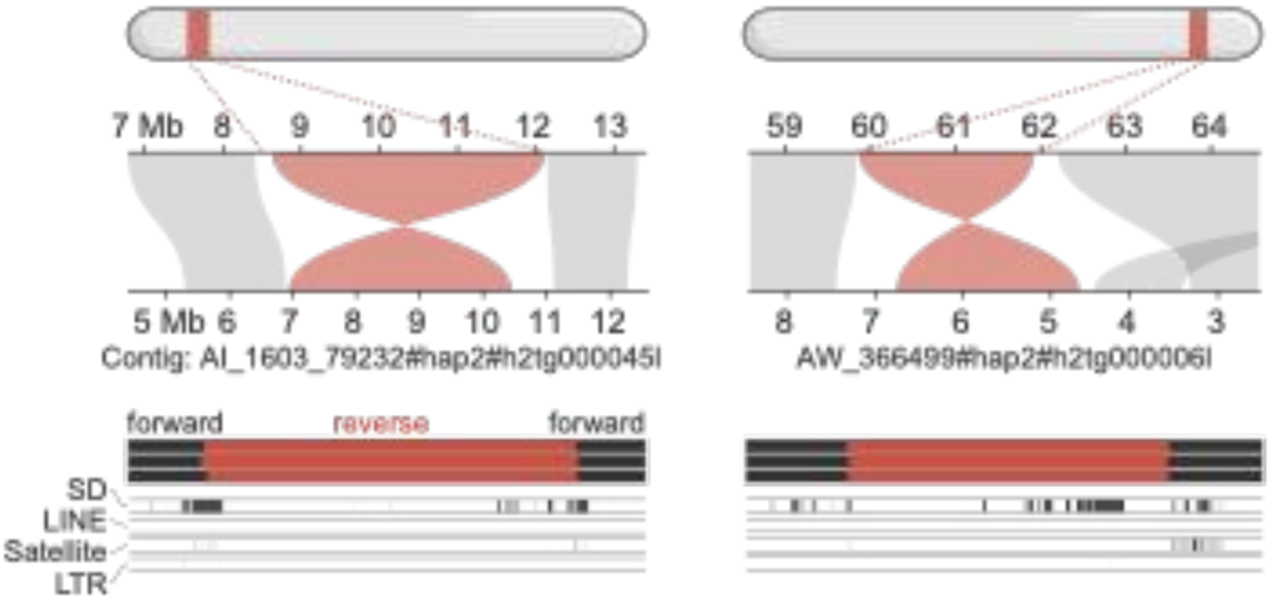


# Longer SVs rise to higher frequencies in island population



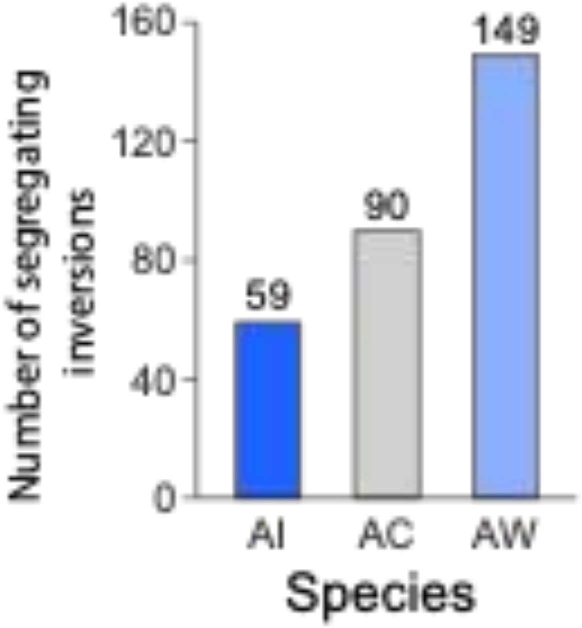
Density of ~485,000 structural variants extracted from PGGB pangenome graph

# Inversions are common and track population size

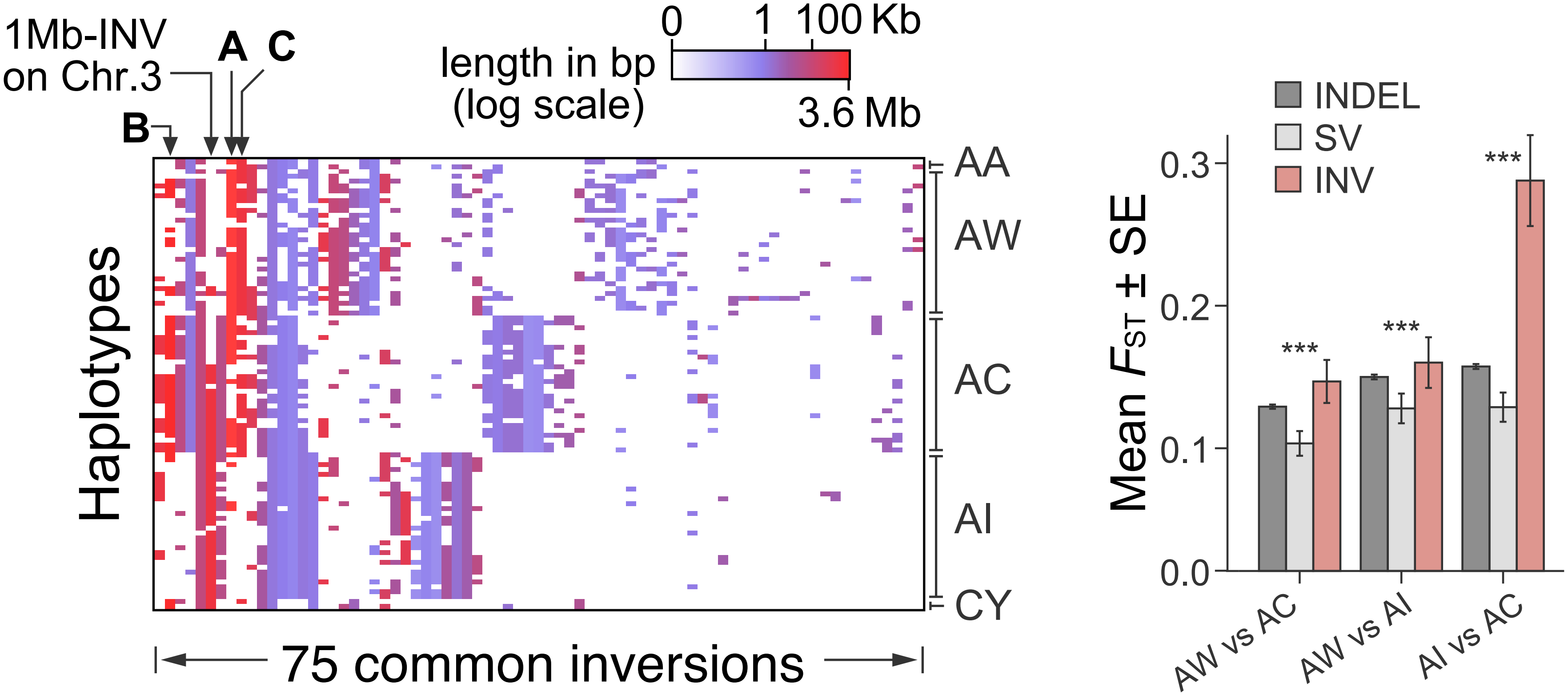


Chr. 1A

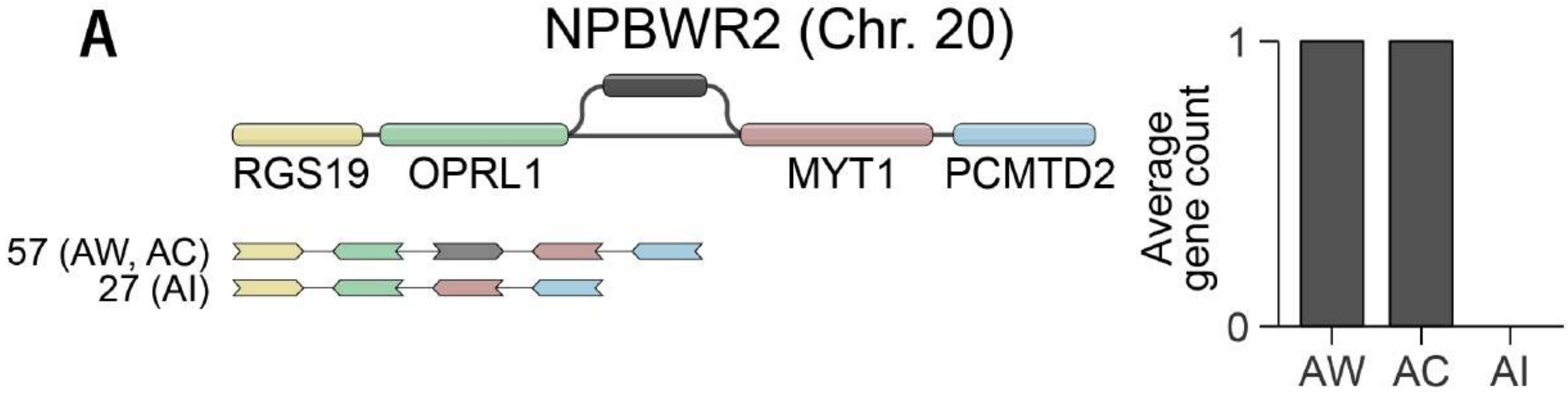
Chr. 5



# 382 inversions identified by pangenome and reference-based methods

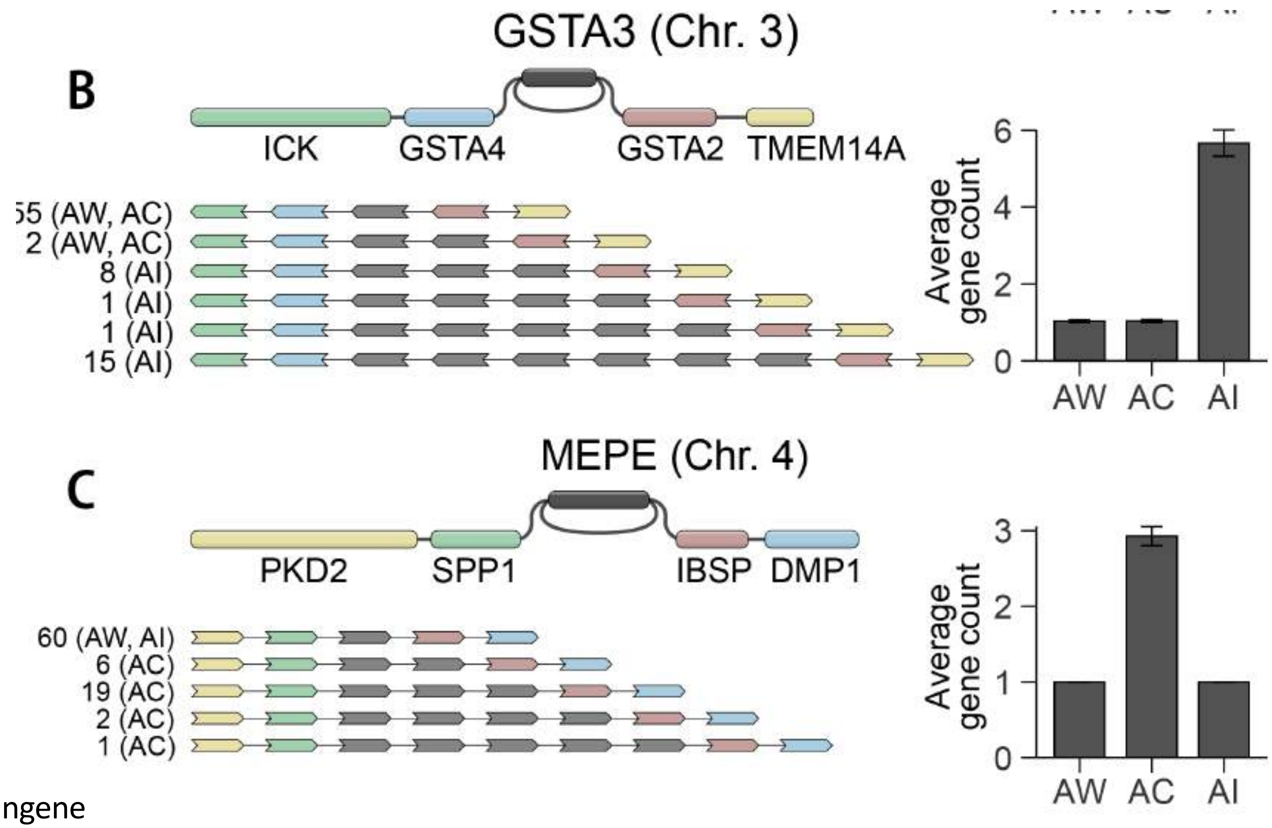


# Abundant copy-number variants and gene deletions



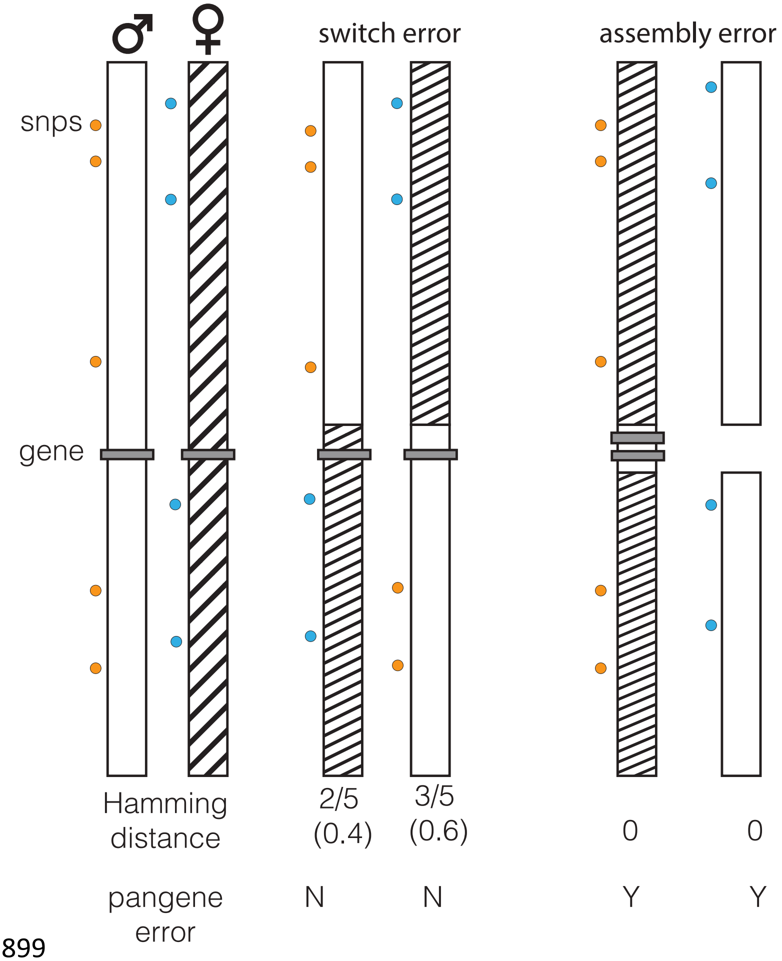
Using miniprot and pangene
Li et al. 2024. *Bioinformatics*
Li, H. 2023. *Bioinformatics*

# Abundant copy-number variants and gene deletions



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

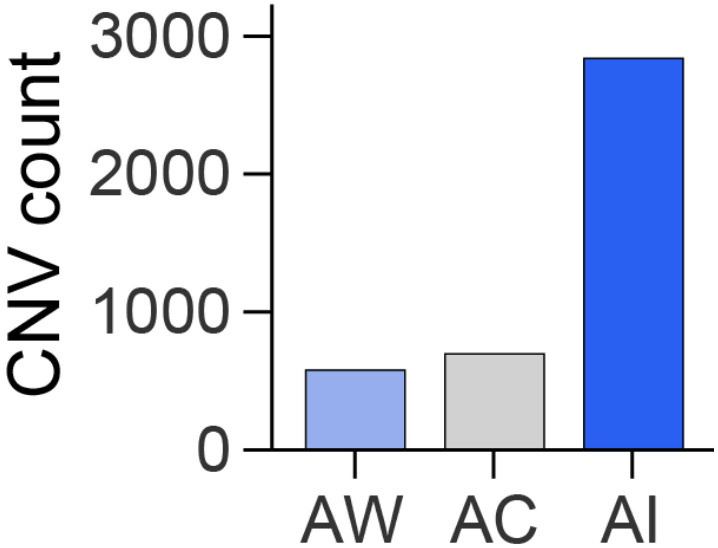
# Assembly errors can mimic gene deletions



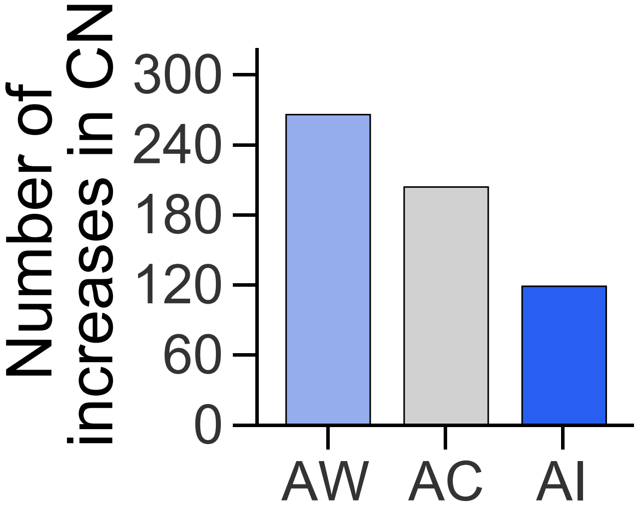
Edwards & Li. 2026 *Proc. R. Soc. B* **293**: 20252899

# Gene copy number variants exhibit a surprising pattern

gene deletion is most common type of CNV
and most common in Island Scrub-Jay

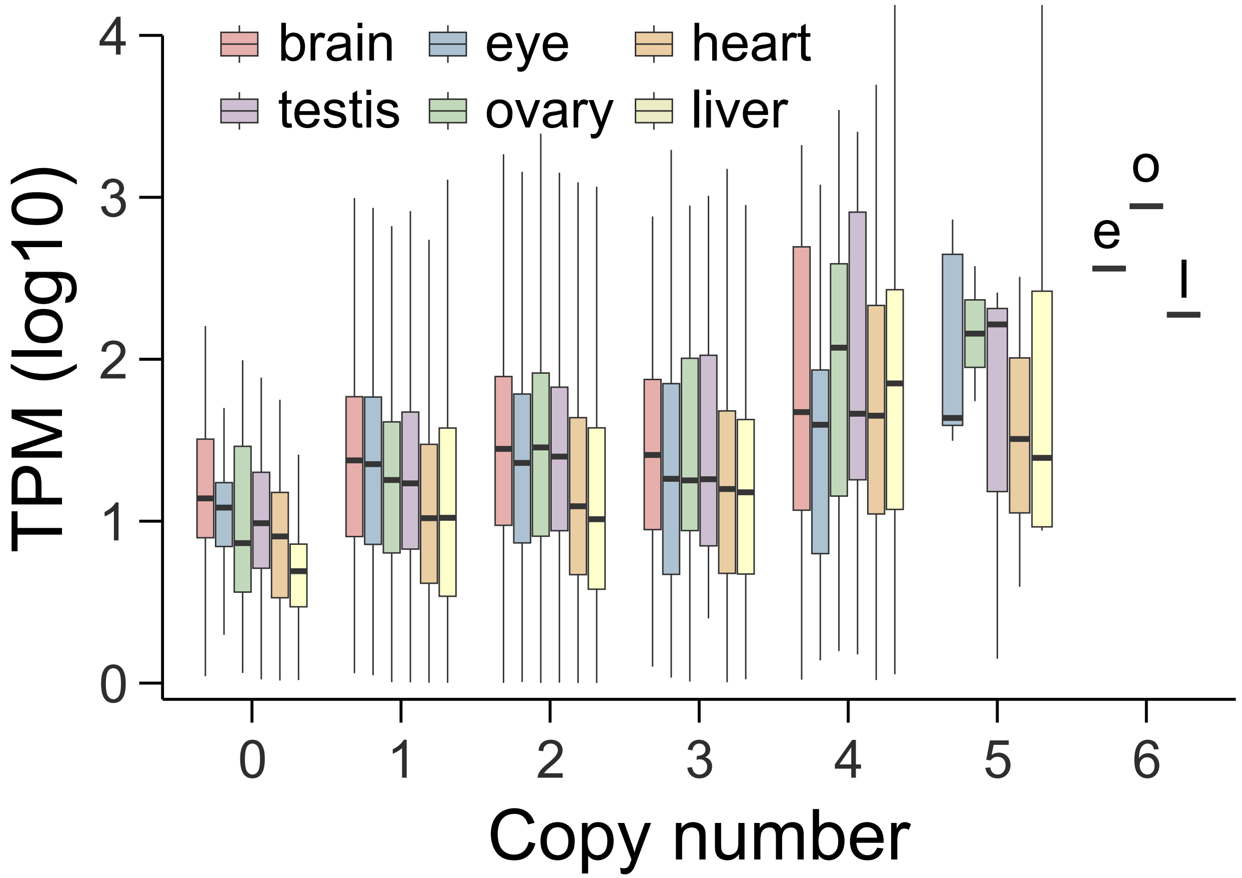


Gene duplications are rarest in Island Scrub-Jay

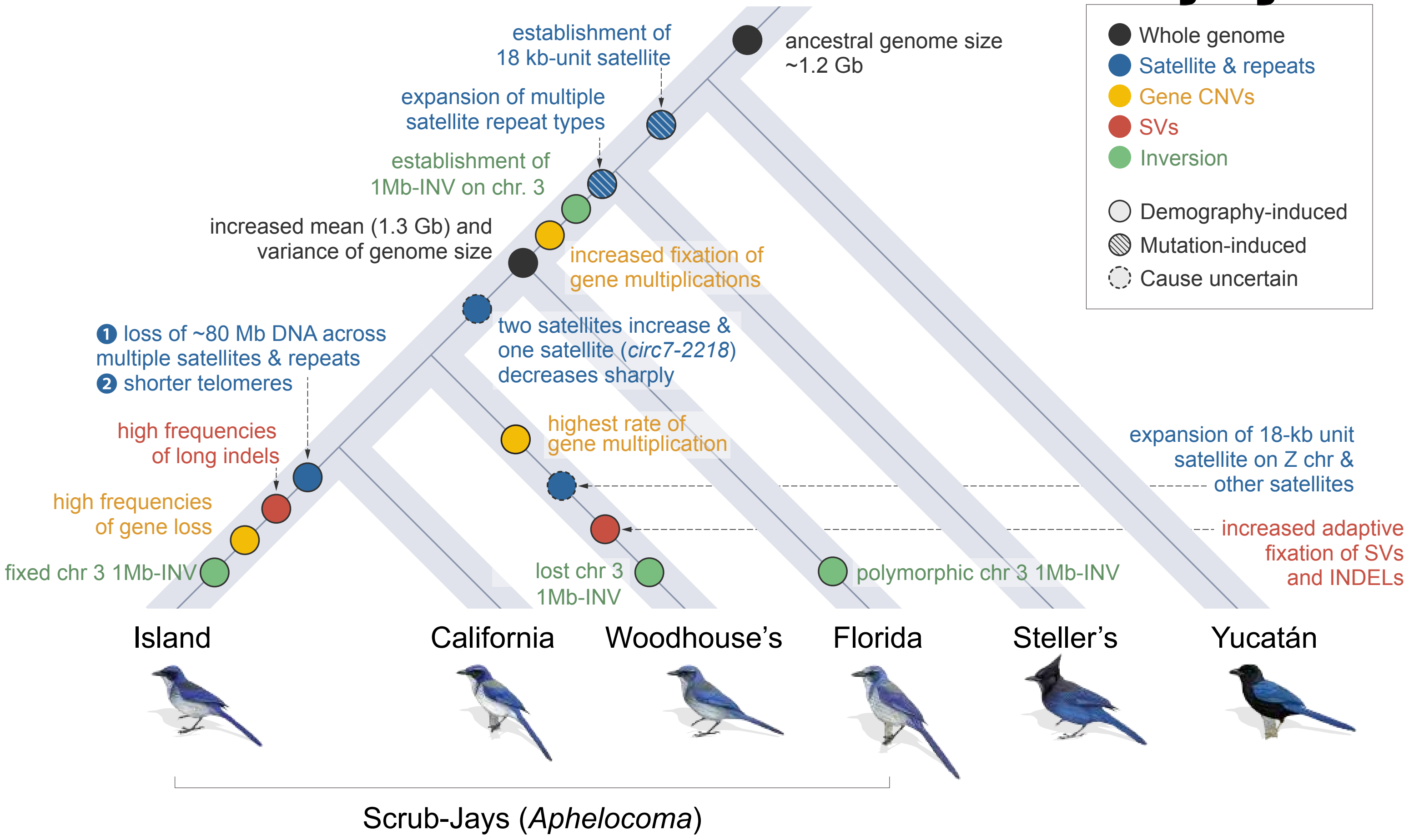


Using miniprot and pangene
Li et al. 2024. *Bioinformatics*
Li, H. 2023. *Bioinformatics*

# Copy number variants influence gene expression



# 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 indels

high frequencies of gene loss

fixed chr 3 1Mb-INV

highest rate of gene multiplication

lost chr 3 1Mb-INV

polymorphic chr 3 1Mb-INV

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

increased adaptive fixation of SVs and INDELs

Scrub-Jays (*Aphelocoma*)

# Population size influences diversity of structural variants in scrub-jays

- Long-read pangenomes reveal increased complexity of bird genomes
- Telomeres, but not genome size, follow predictions of $N_e$
- Structural variants generally track population size and are estimated to be slightly deleterious
- Gene copy number variants behave as if strongly deleterious