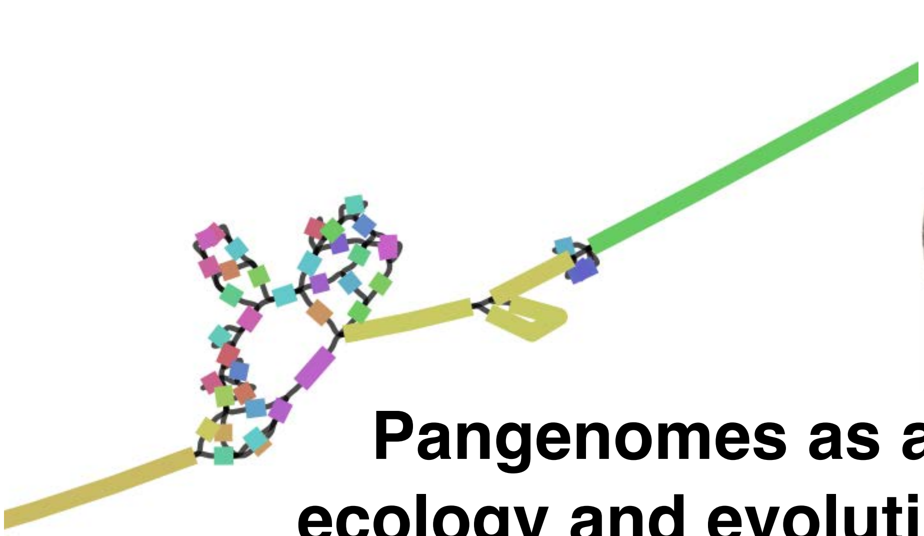
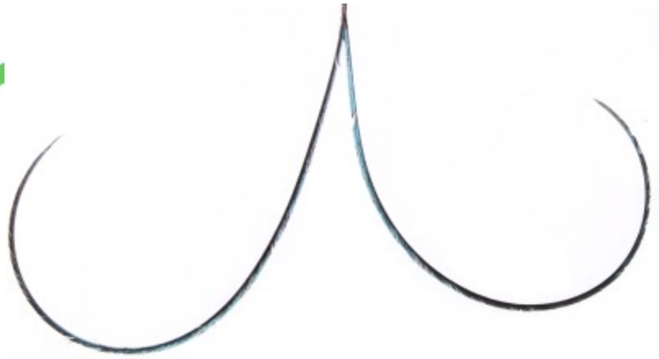
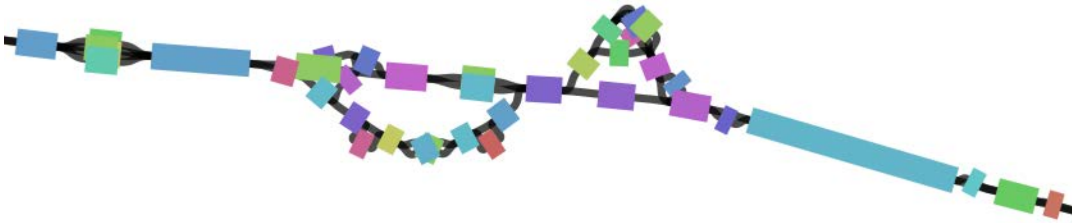
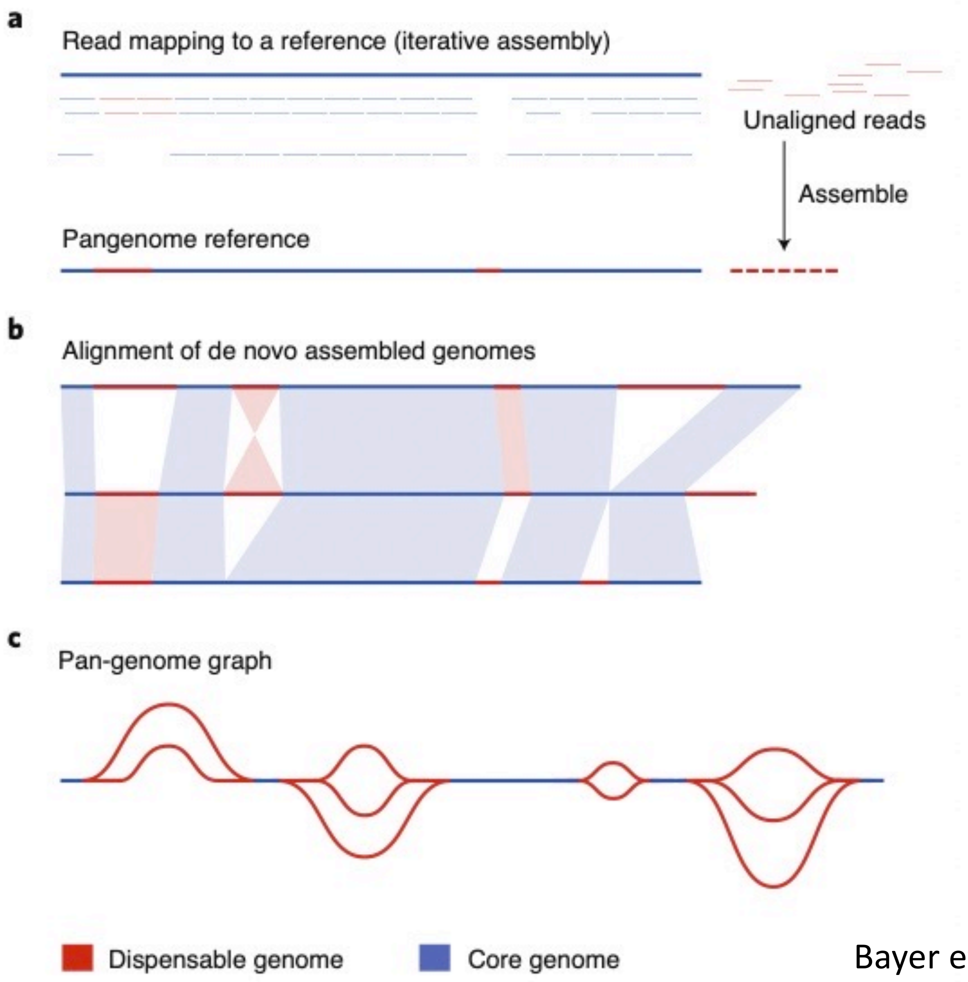



Pangenomes as a new tool for studying ecology and evolution of natural populations

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

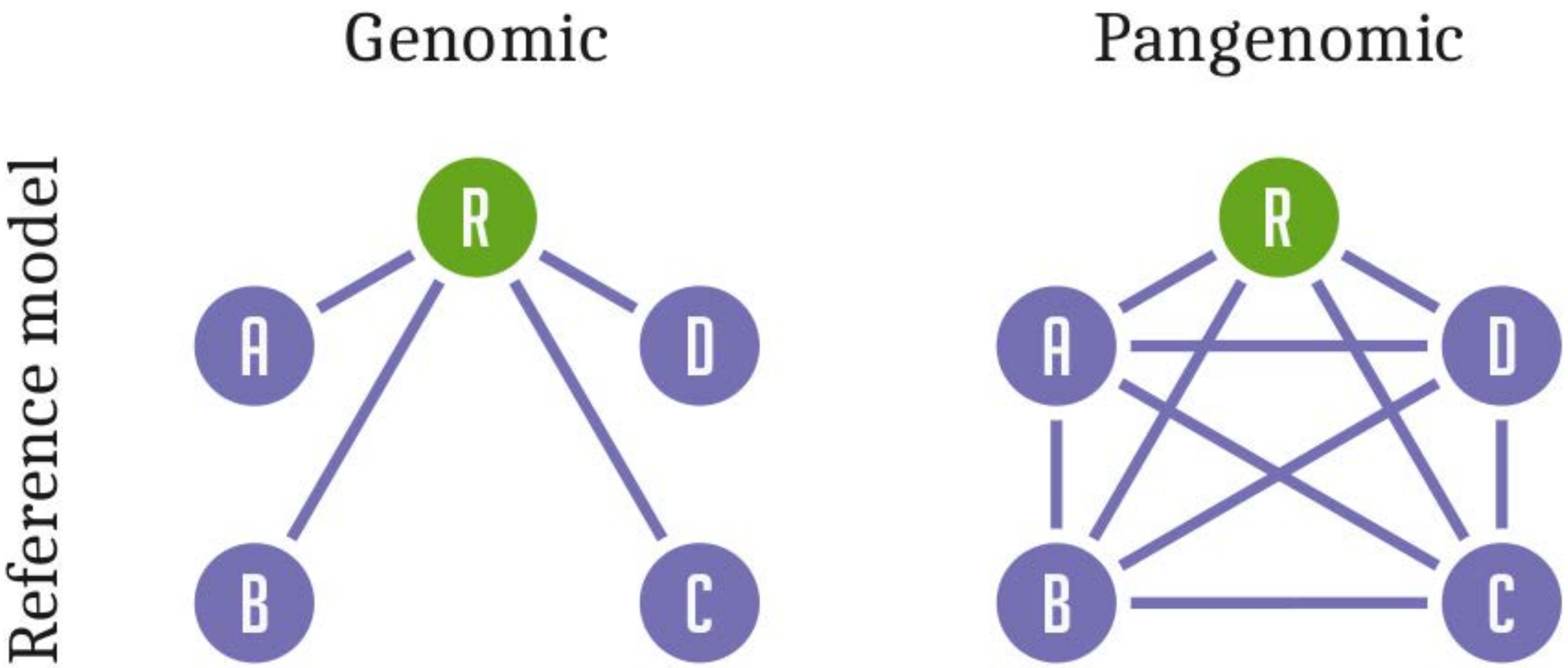


Pangenomes: moving beyond reference-based genomics



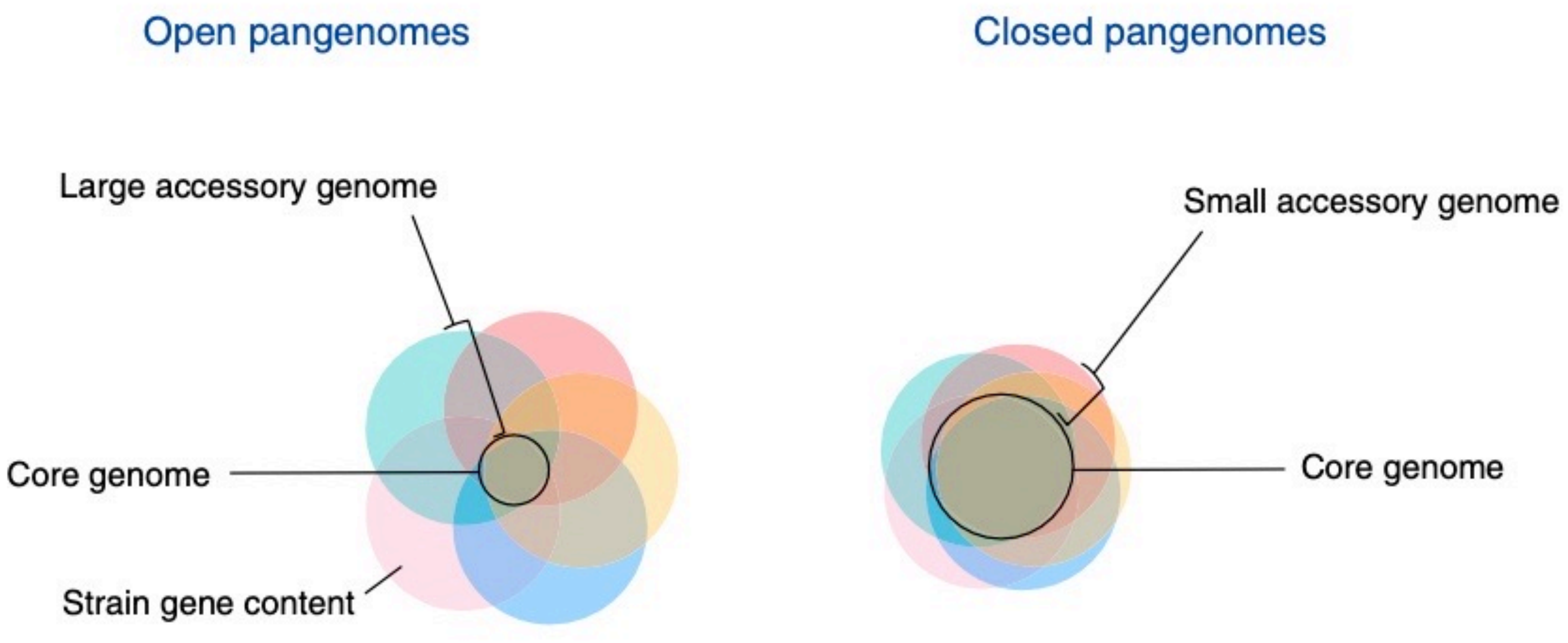
Bayer et al. 2020. *Nature Plants* 6: 914-920.

Reference-free genomics



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

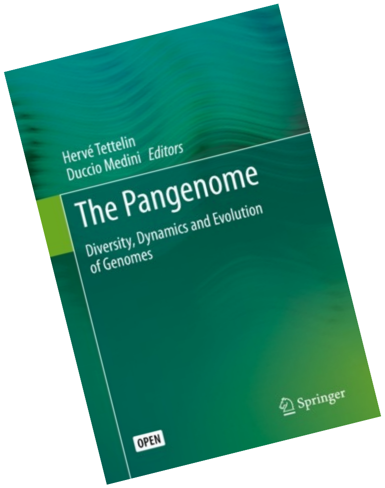
Open and closed pangenomes

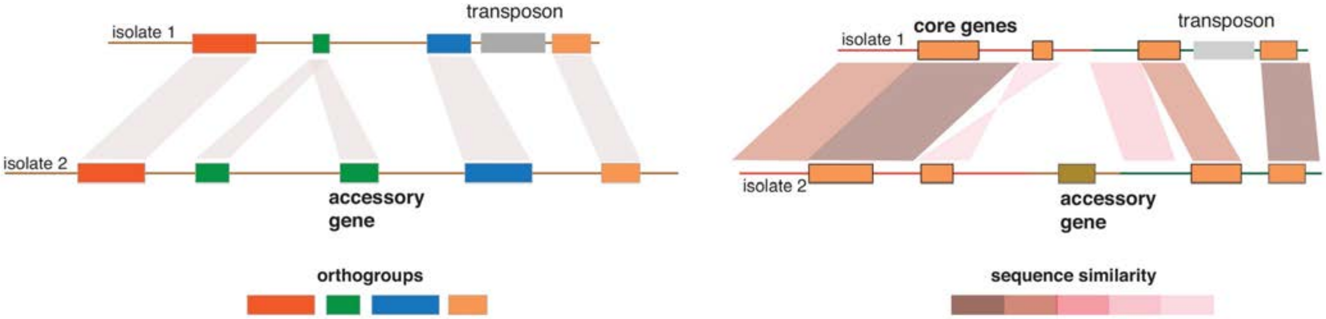


Brokhurst et al. 2019. *Curr. Biol.*

The eukaryotic pangenome

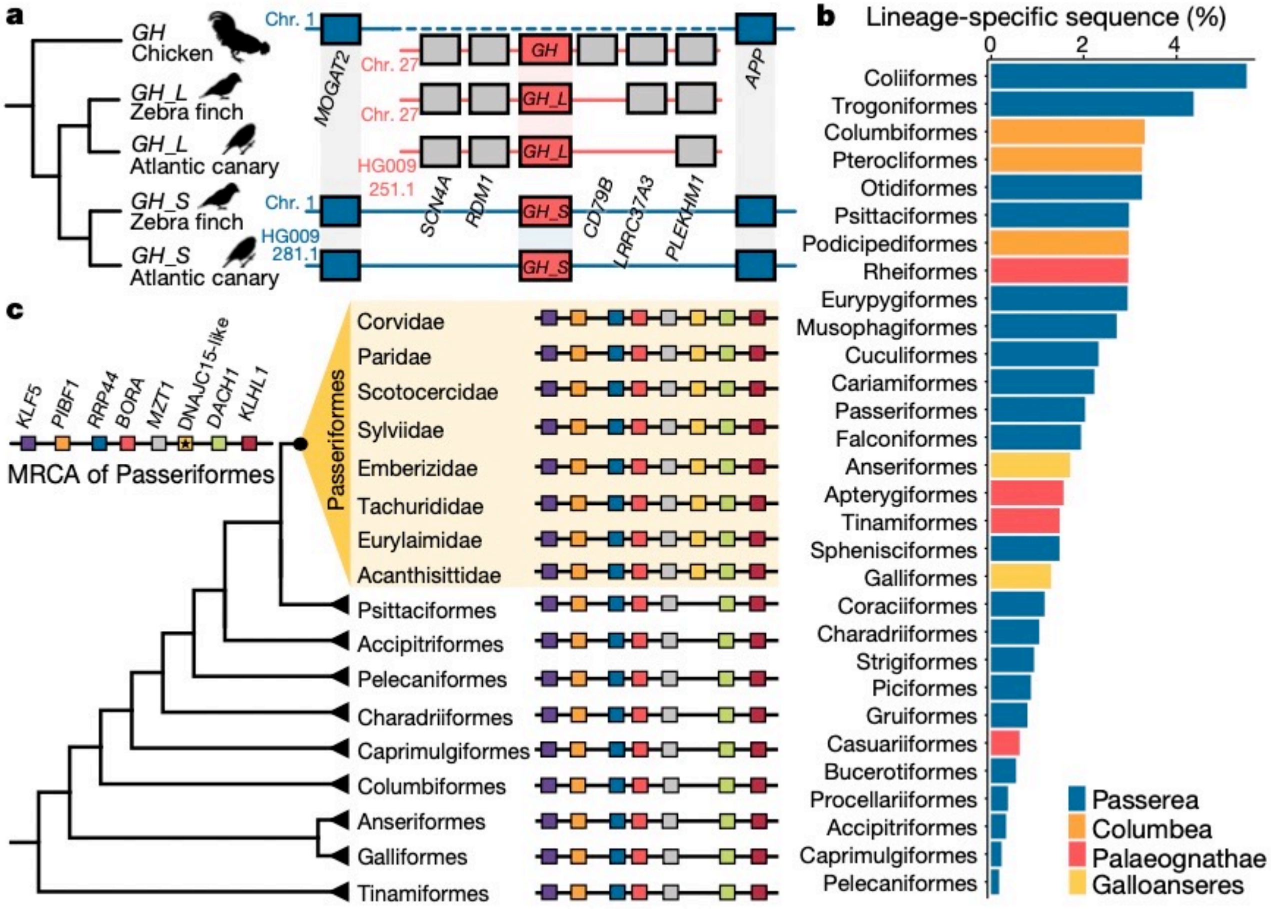
- “The existence of pangenomes in eukaryotes is debated…Pangenome studies in eukaryotes are challenging due to their more complex genome and architectures and a lack of replete genome-level sampling” (Brockhurst et al. 2019. *Current Biology)*





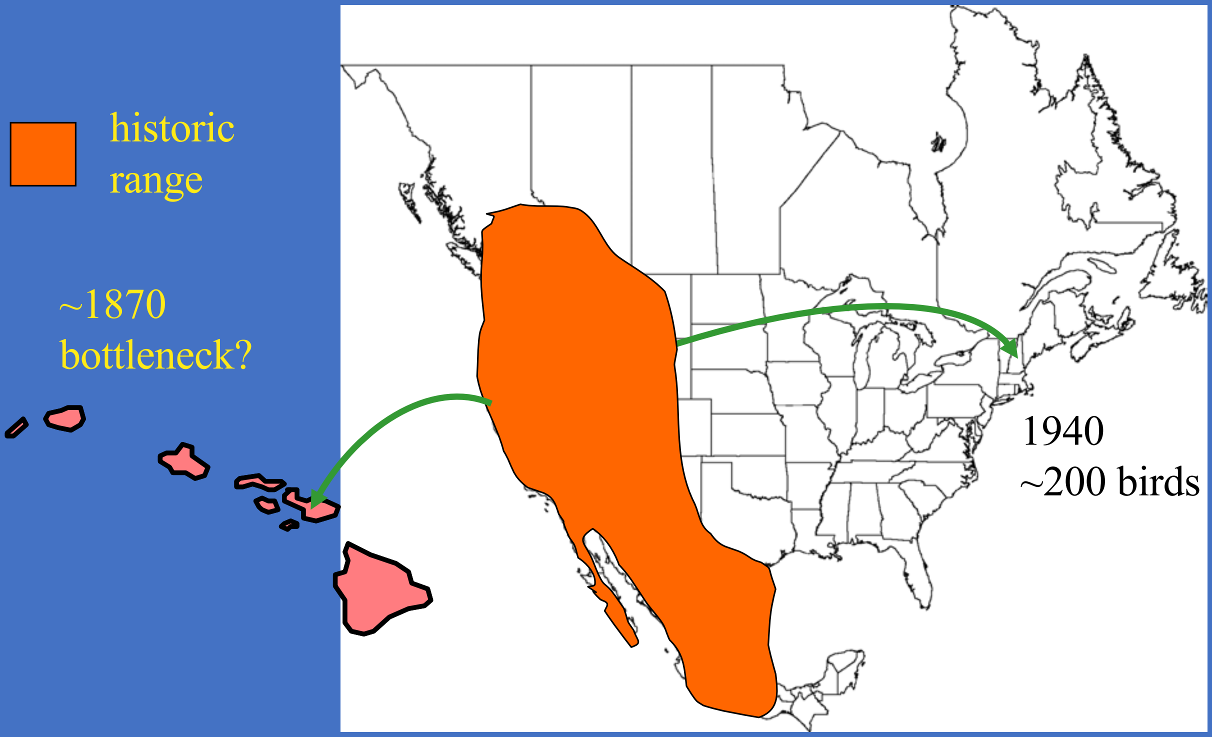
https://pathogen-genomics.org/research/

Pangenome approach to comparative genomics

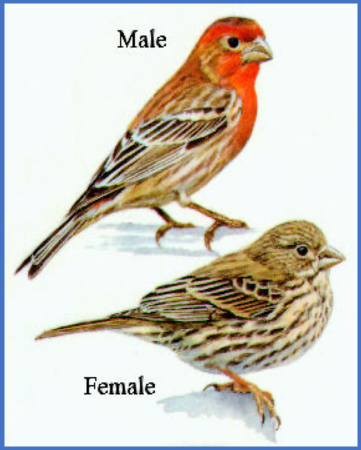


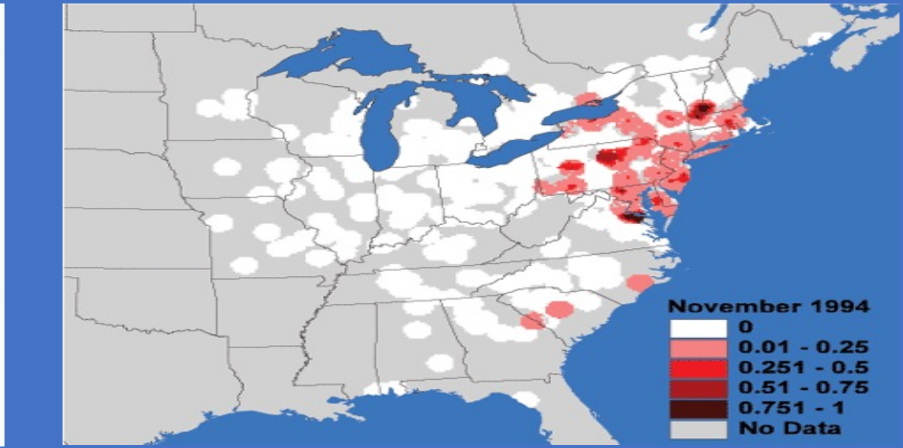
Feng et al. 2020. *Nature* 587:252-257.

Recent history of House Finch populations



Rapid spread of *Mycoplasma* in House Finch populations

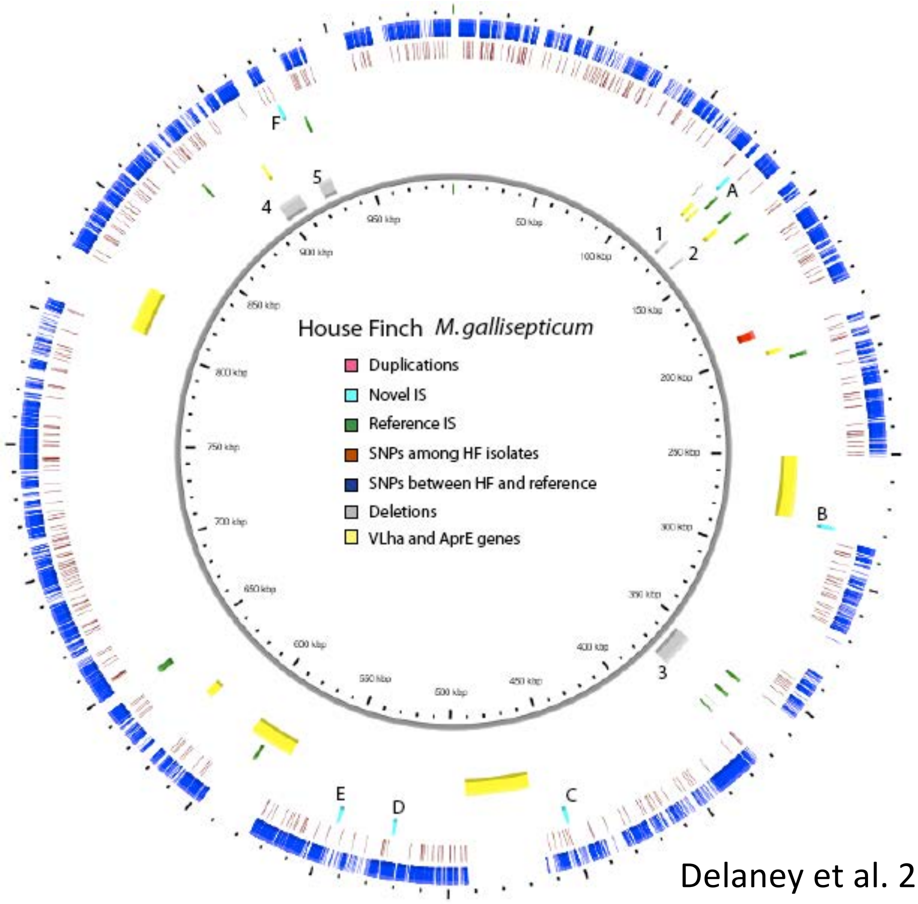




Courtesy Cornell Lab of Ornithology

- *Mycoplasma* is transmitted horizontally, often at bird feeders
- Expanded throughout the eastern US in just five years
- Has now crossed the Rockies and is spreading south through California and the southwest.

House Finch *Mycoplasma* genome ~1 Mb



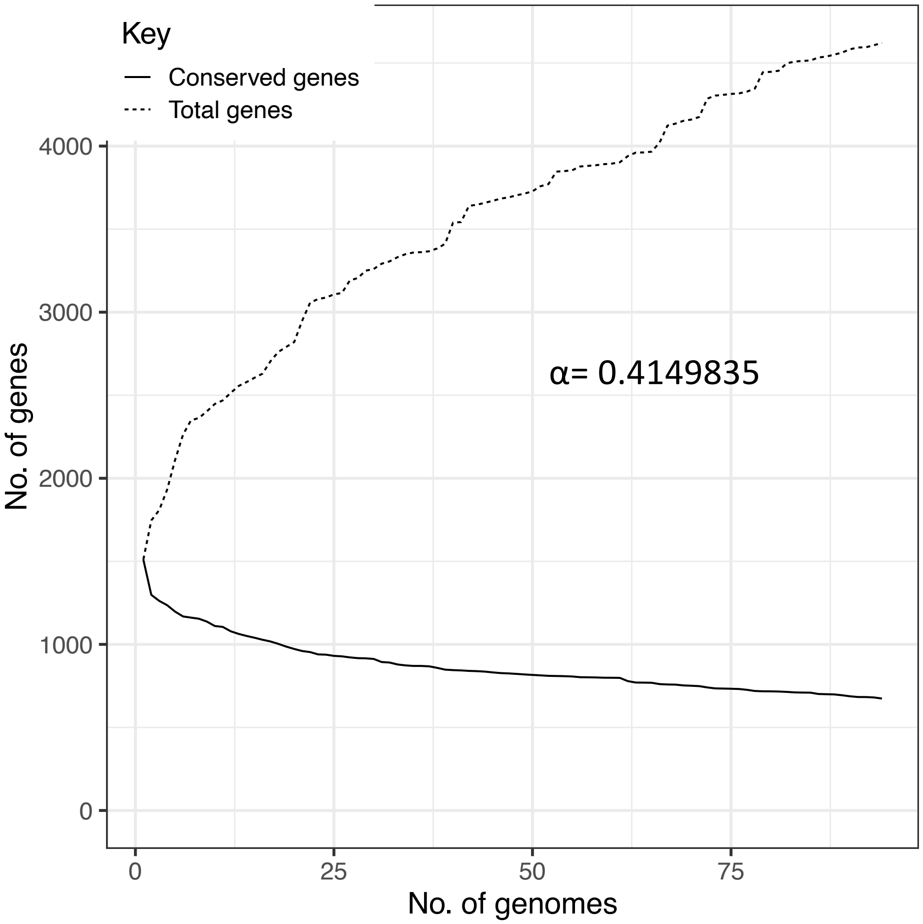
Analyzed 81 Mycoplasma strains from chicken, turkey and house finch, available on NCBI

Added 12 new House Finch Mycoplasma strains, sequenced with PacBio

Used

Delaney et al. 2012. *PLoS Genetics*

Pangenome of *Mycoplasma gallisepticum*



The size of the pan-genome was determined using 10,000 permutations by microPan

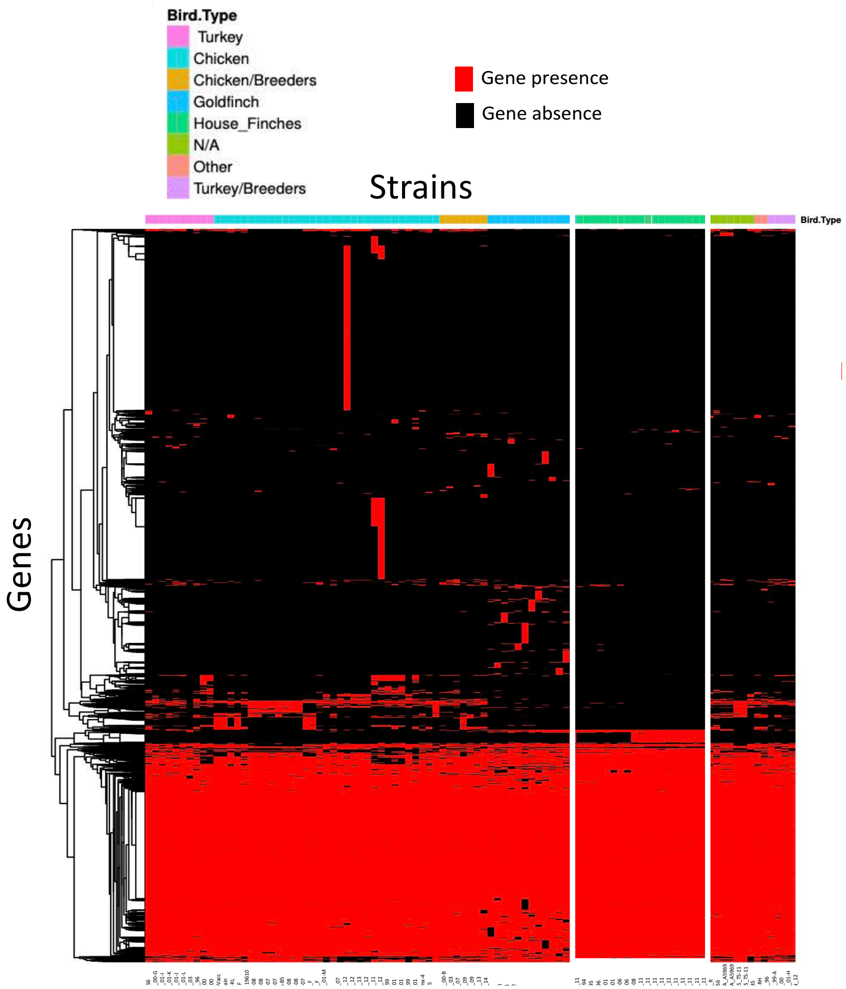
Feature	Info	Number of genes	Percentage
Core genes	(99% <= strains <= 100%)	674	14.586
Soft core genes	(95% <= strains < 99%)	464	10.041
Shell genes	(15% <= strains < 95%)	412	8.916
Cloud genes	(0% <= strains < 15%)	3071	66.457
SGF	one copy in all strains	141	3.051
SGF	without recombination signals	117	2.532
Total genes	(0% <= strains <= 100%)	4621	100

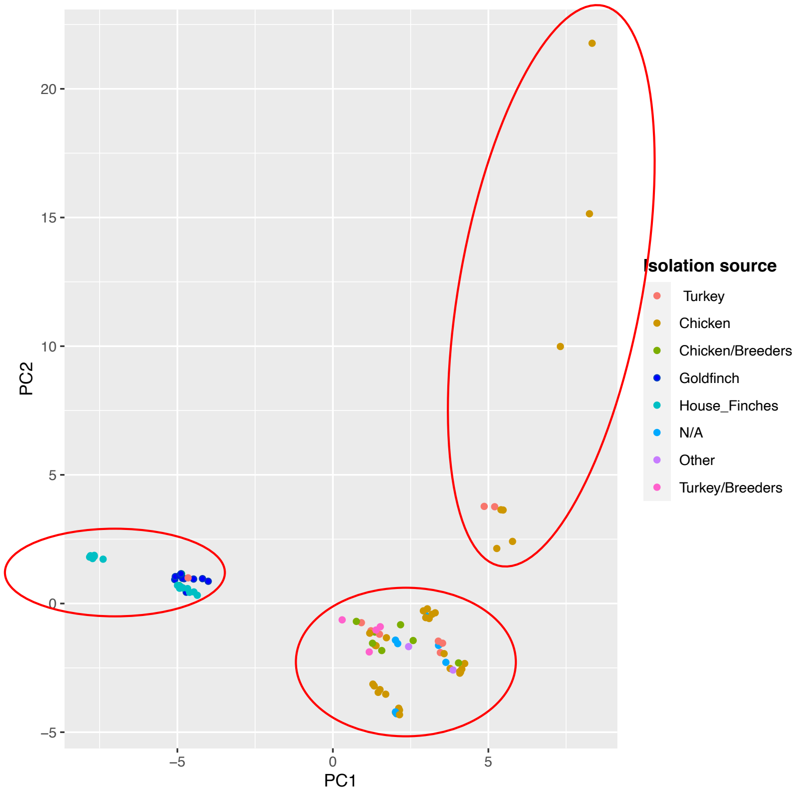
Alpha value: the number of gene clusters we would see if we collected *all* genomes of the species

New data: Determine the alpha value using MicroPan

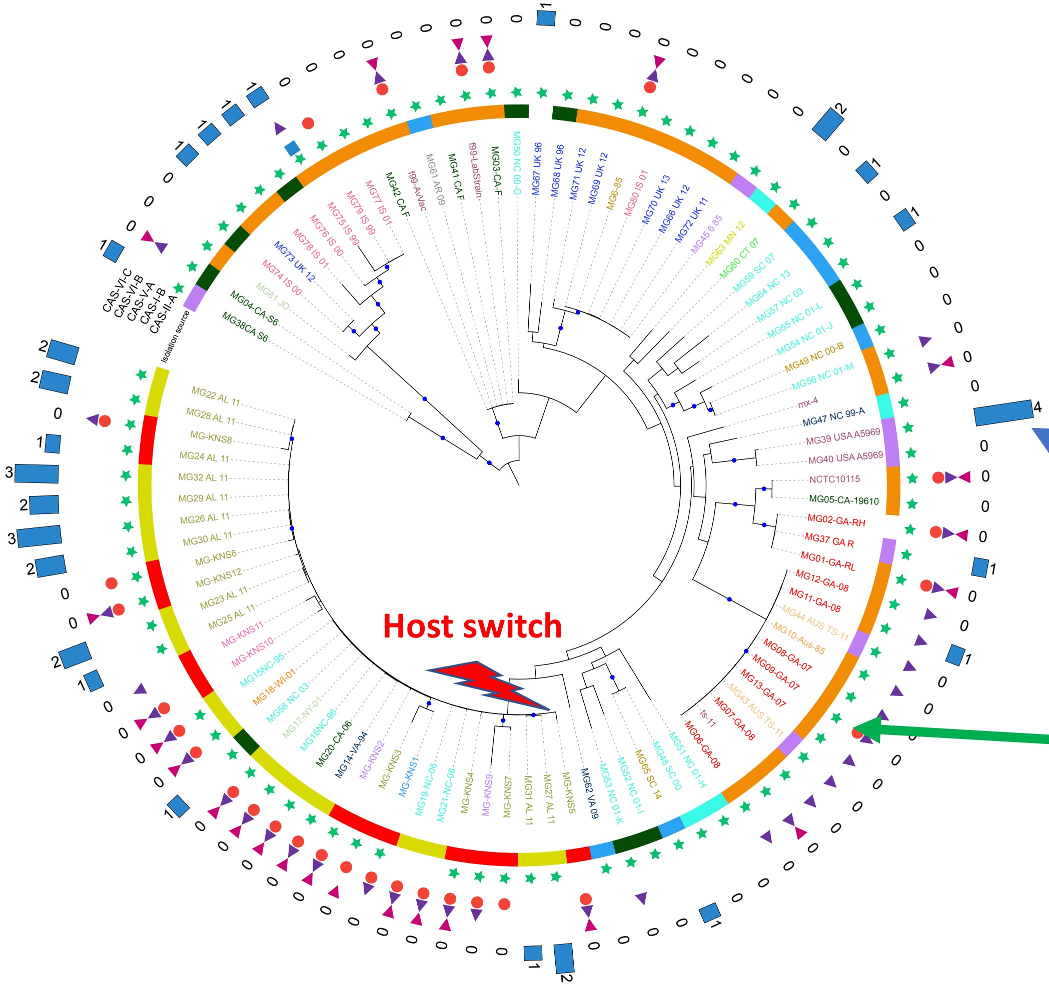
*the pan-genome is closed if the estimated alpha is above 1.0

Mycoplasma pangenome gene repertoire is highly strain-specific





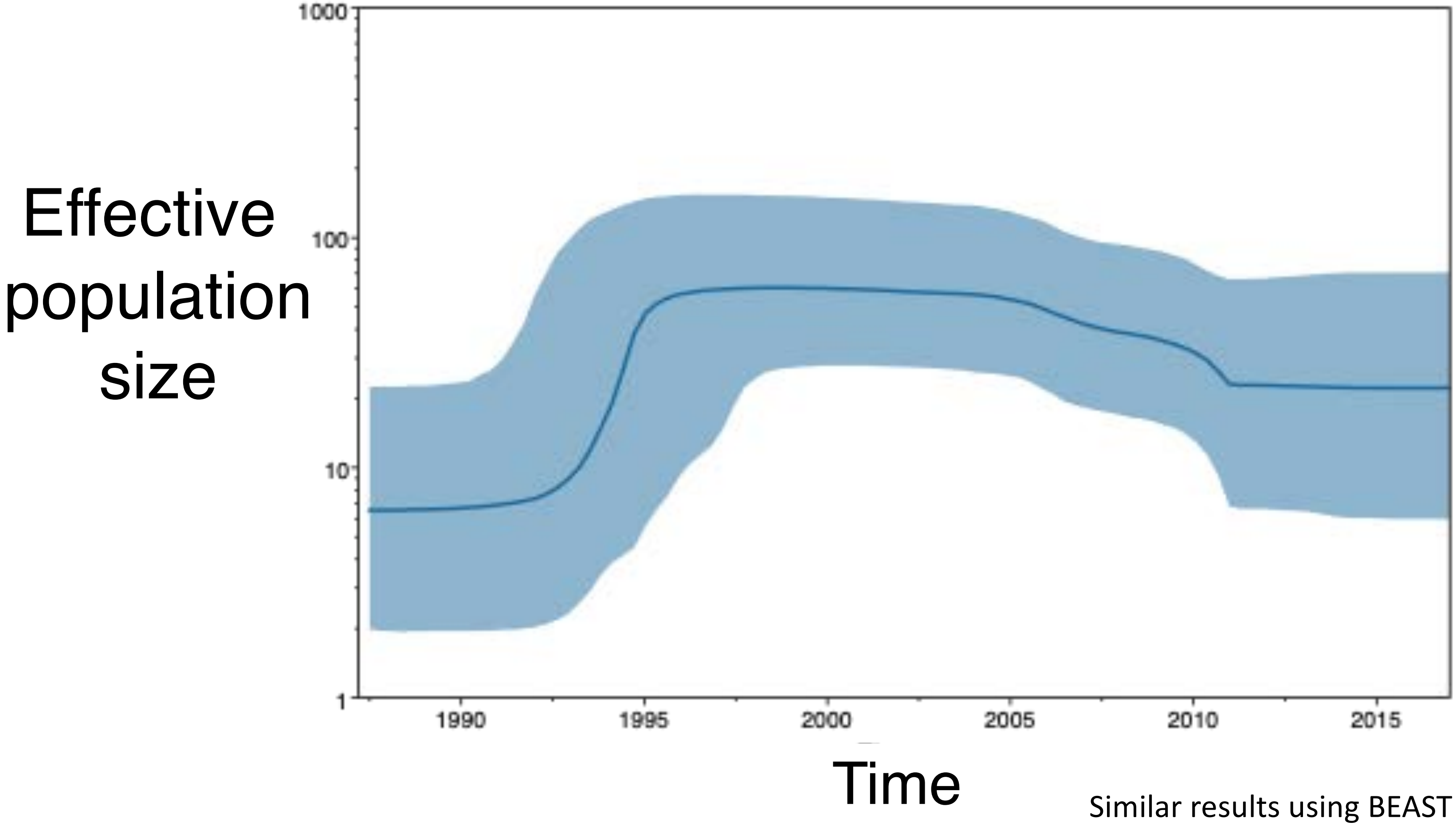
House Finch *Mycoplasma* strains have distinct CRISPR and prophage landscapes



Prophages integrated into Mycoplasma genome

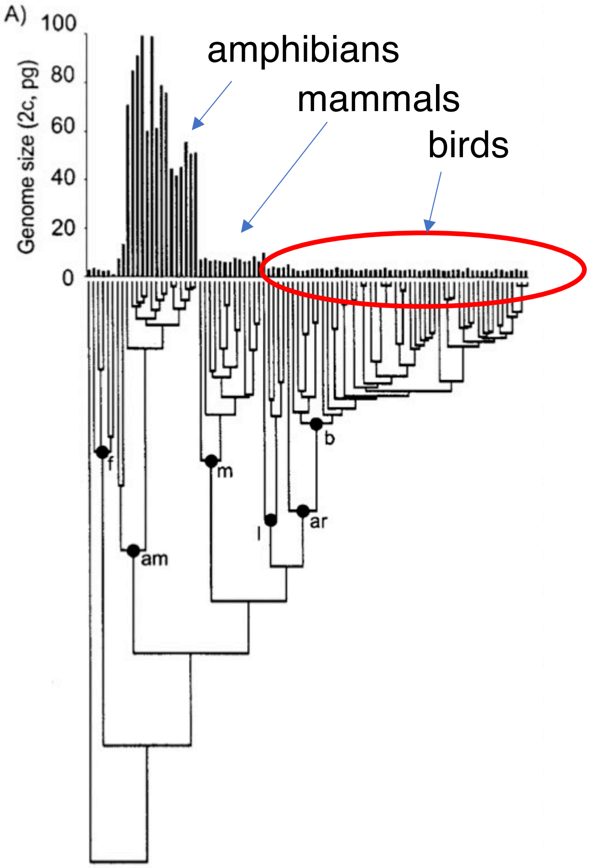
CRISPR landscape

Mycoplasma epizootic likely began ~2 years before first detection

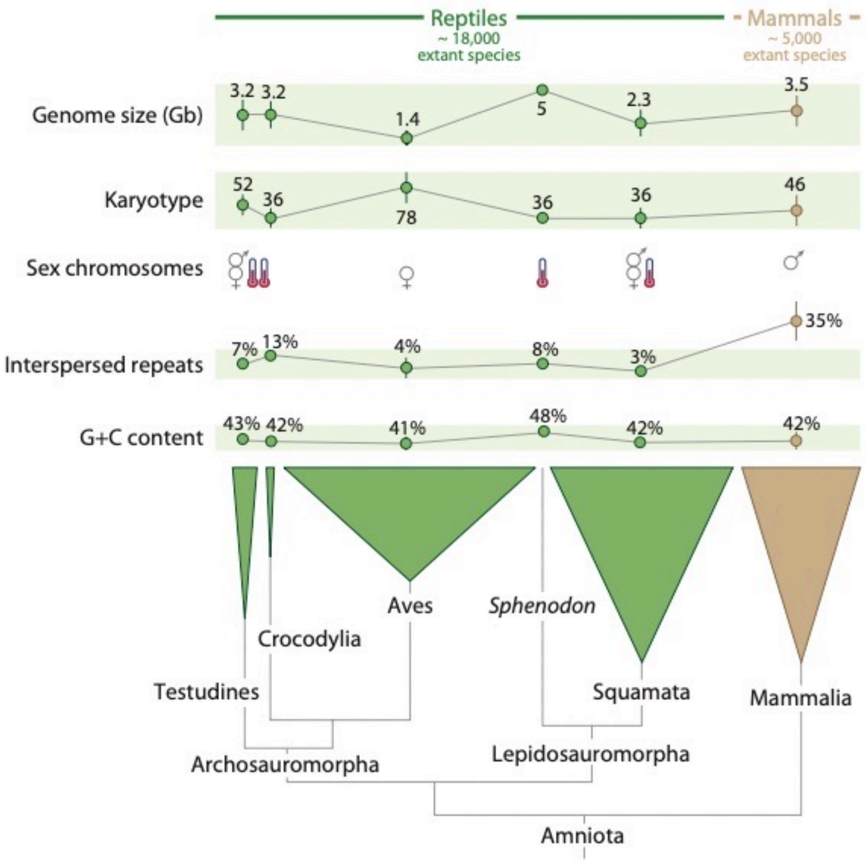


Similar results using BEAST and Stairway plot

Birds have small, streamlined genomes

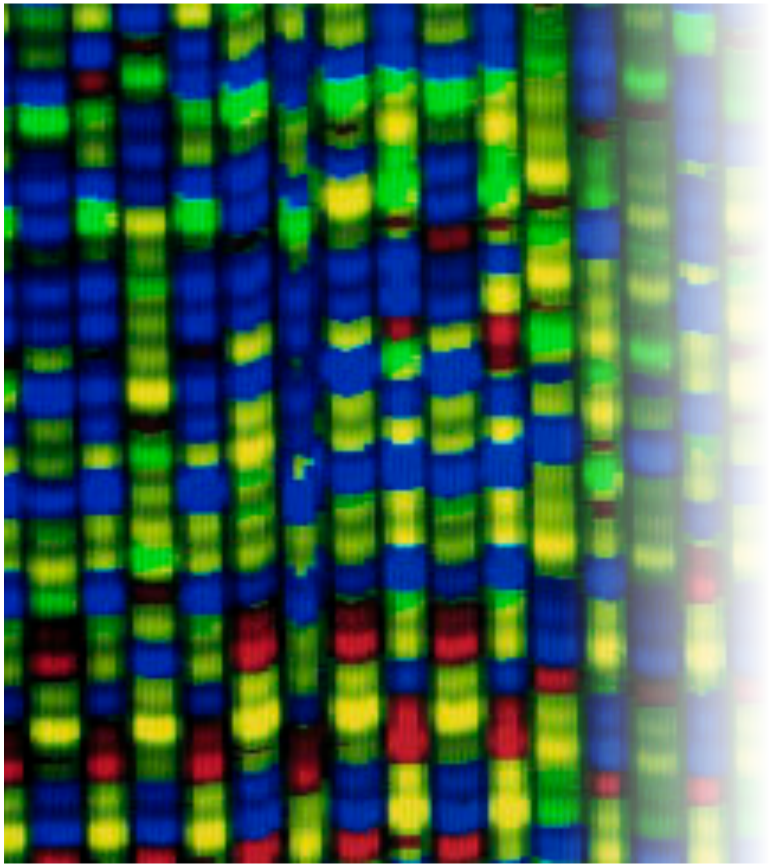


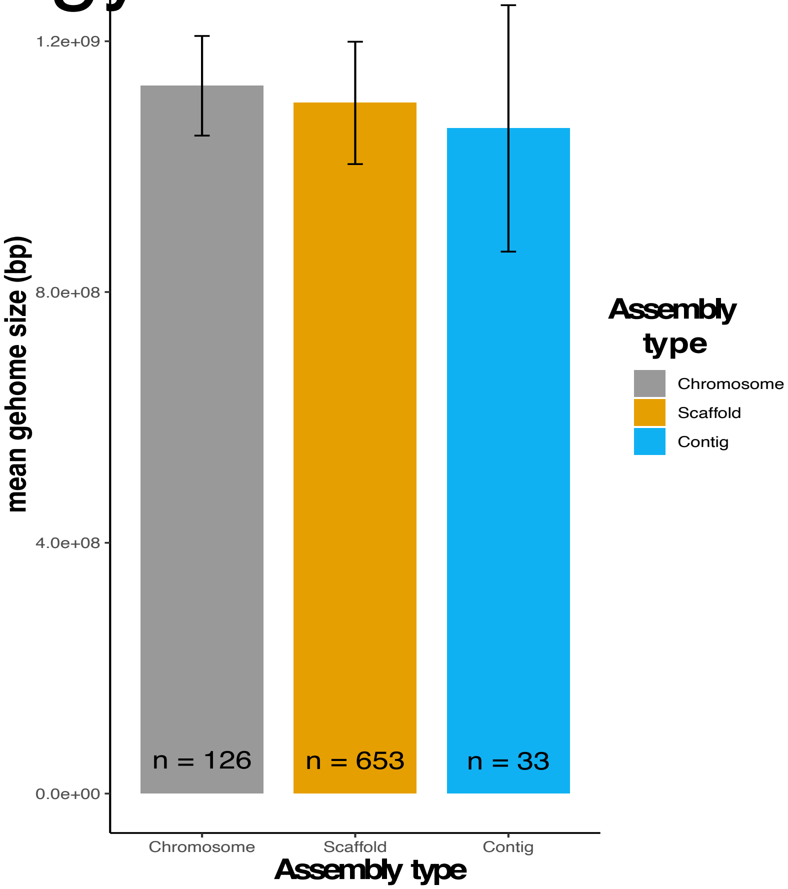
Waltari & Edwards. 2002. *Am. Nat.*



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

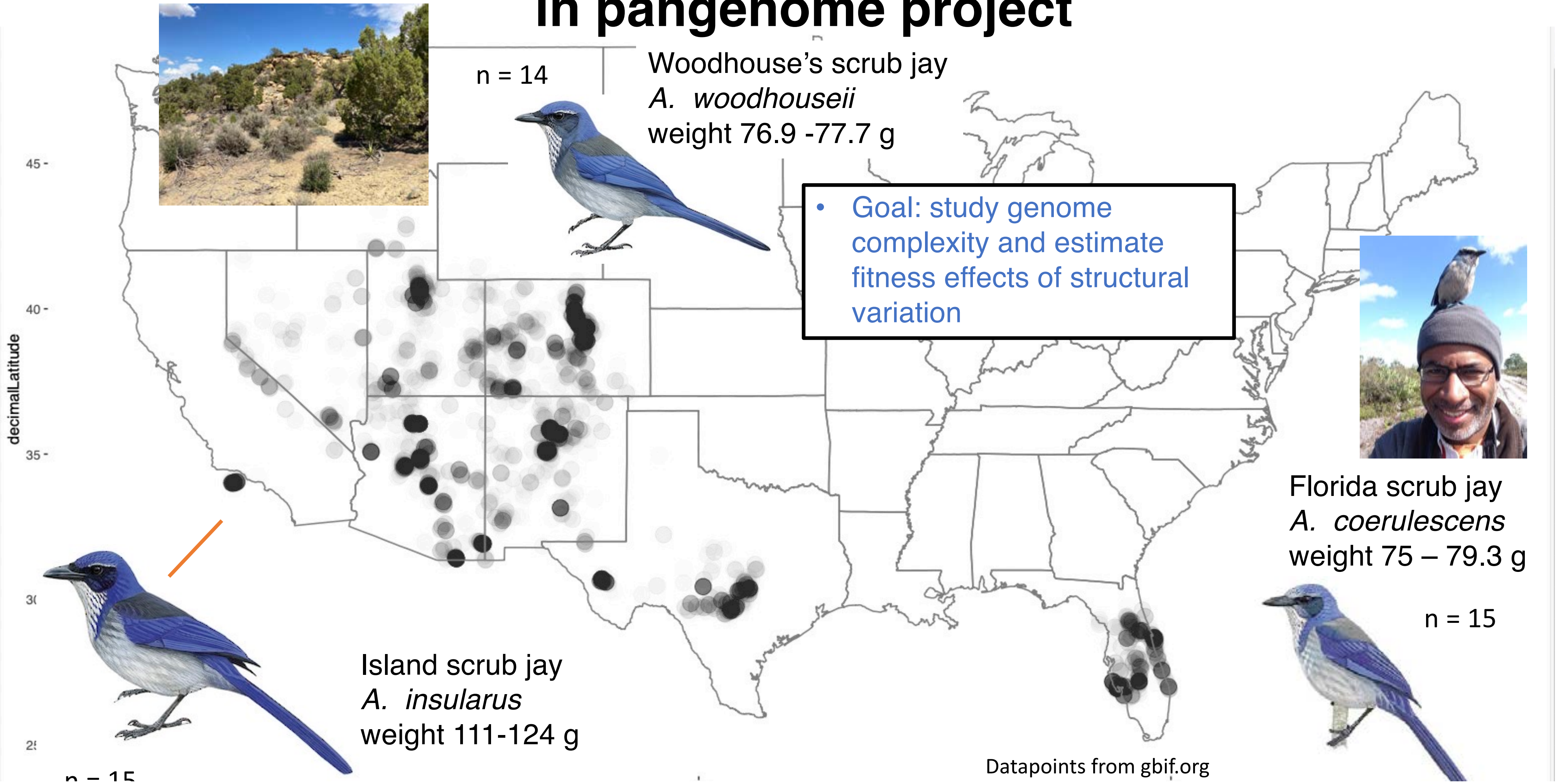
Avian genomes are growing with each new technology





Data from NCBI, accessed 13 Nov. 2021

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

GBE

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

Table 1

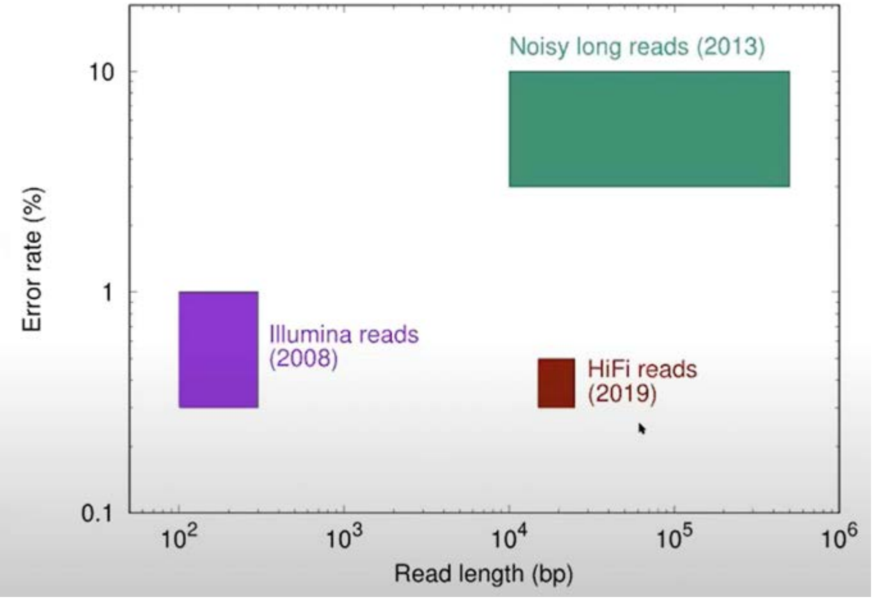
Conceptual Relationships between the Fields of Comparative Population Genomics, Landscape Genomics, and Comparative Phylogeography

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

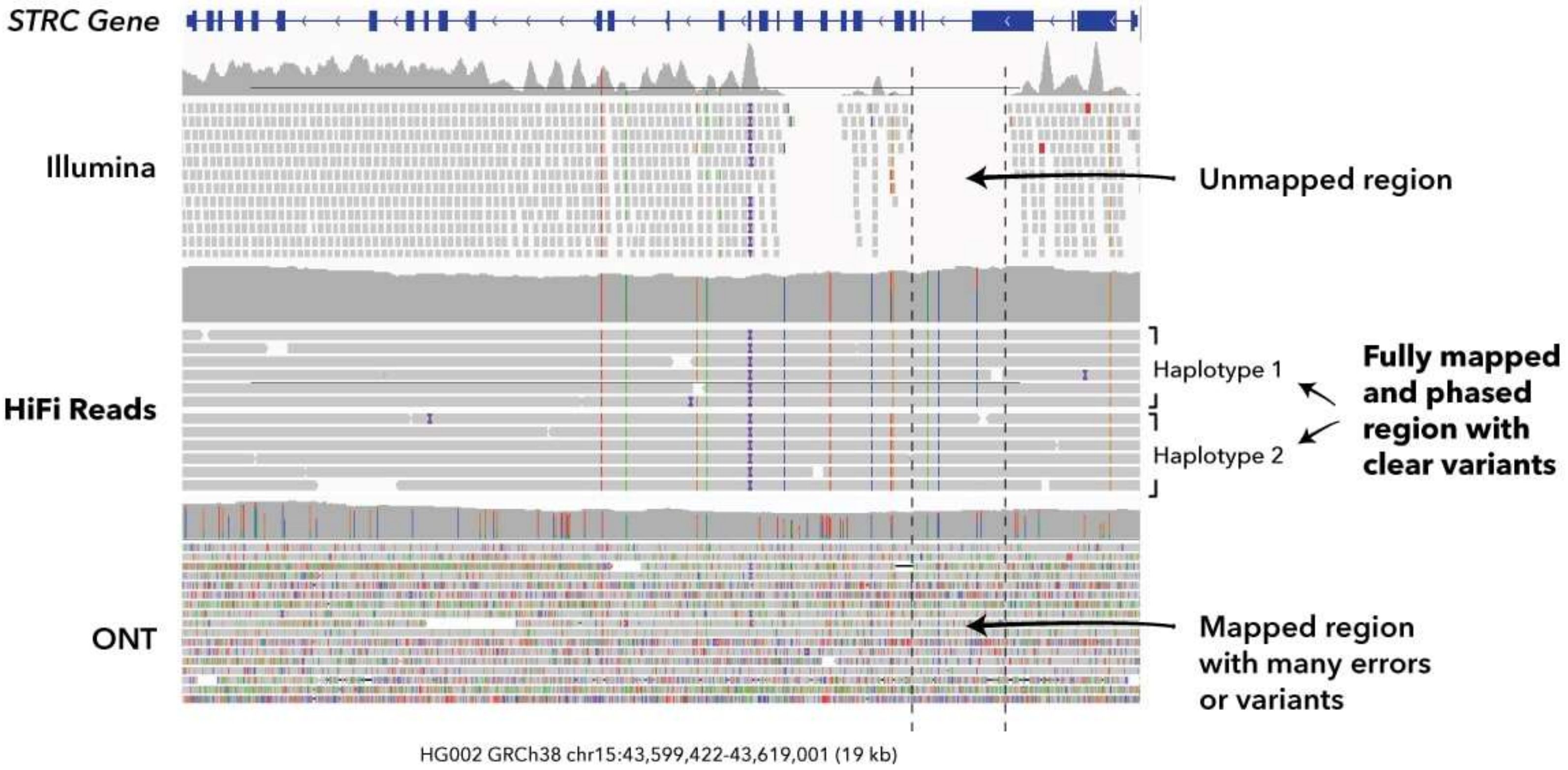
PacBio HiFi reads are long <u>and</u> accurate

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

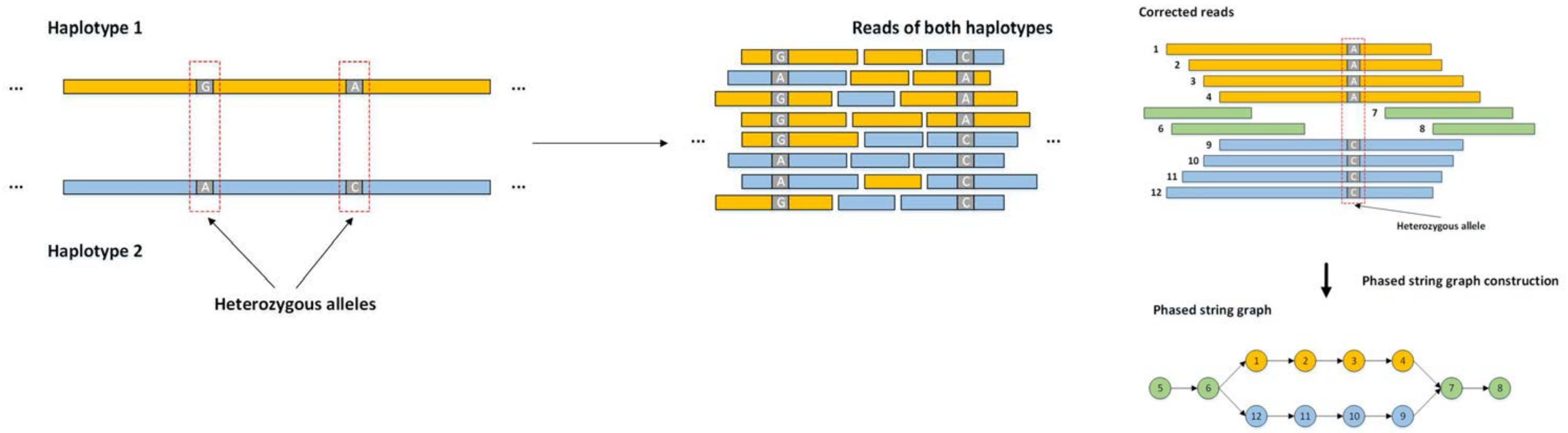


Coutesy Haoyu Cheng, Dana Farber Cancer Institute

PacBio HiFi reads are long and accurate



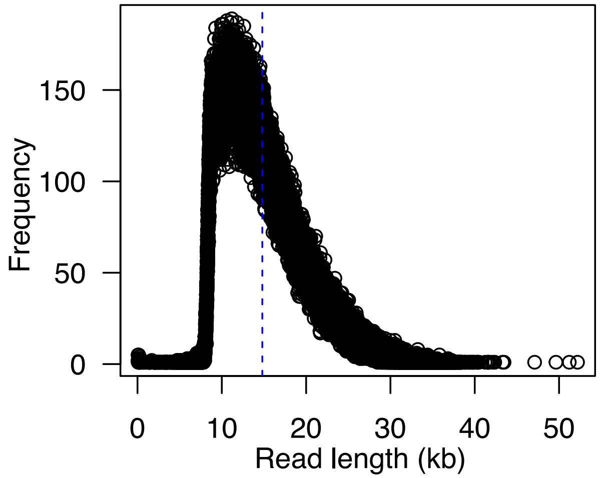
Hifiasm – a HiFi accurate read assembler that resolves haplotypes



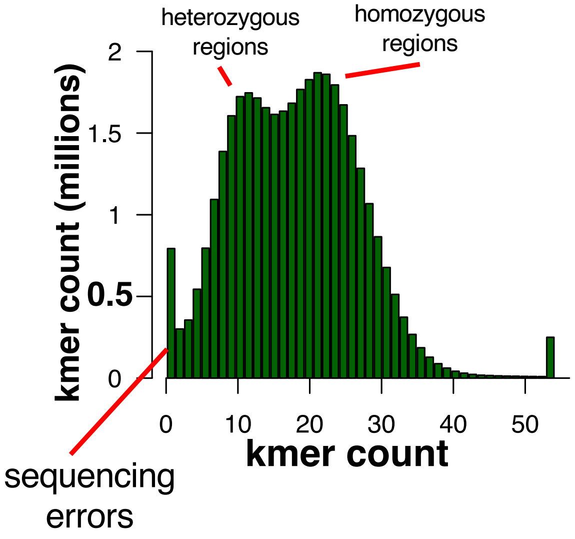
Coutesy Haoyu Cheng, Dana Farber Cancer Institute

Scrub-jay PacBio HiFi data characteristics

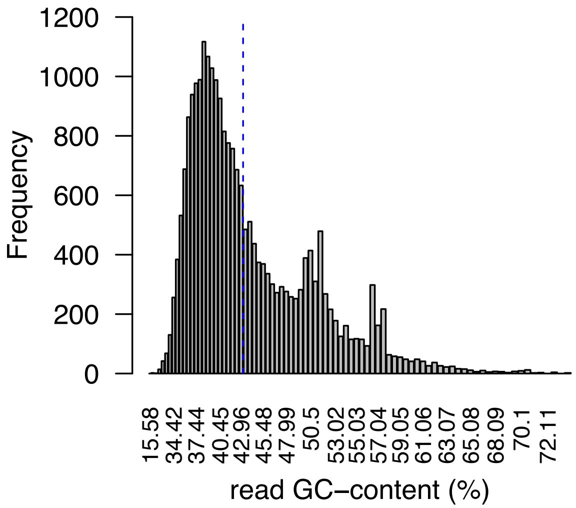
Average read length 14.8 kb



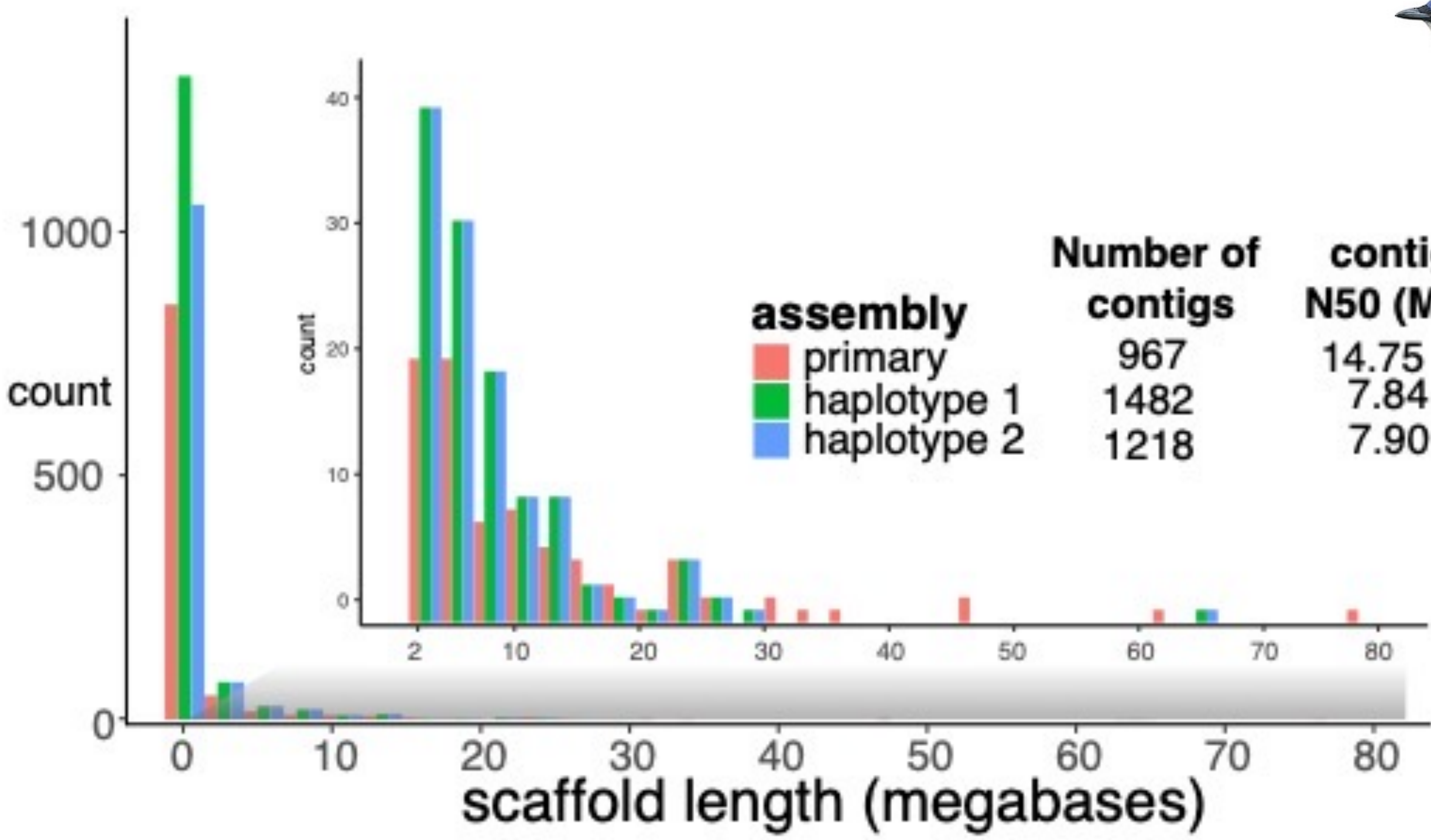
Average coverage ~40X



GC-content



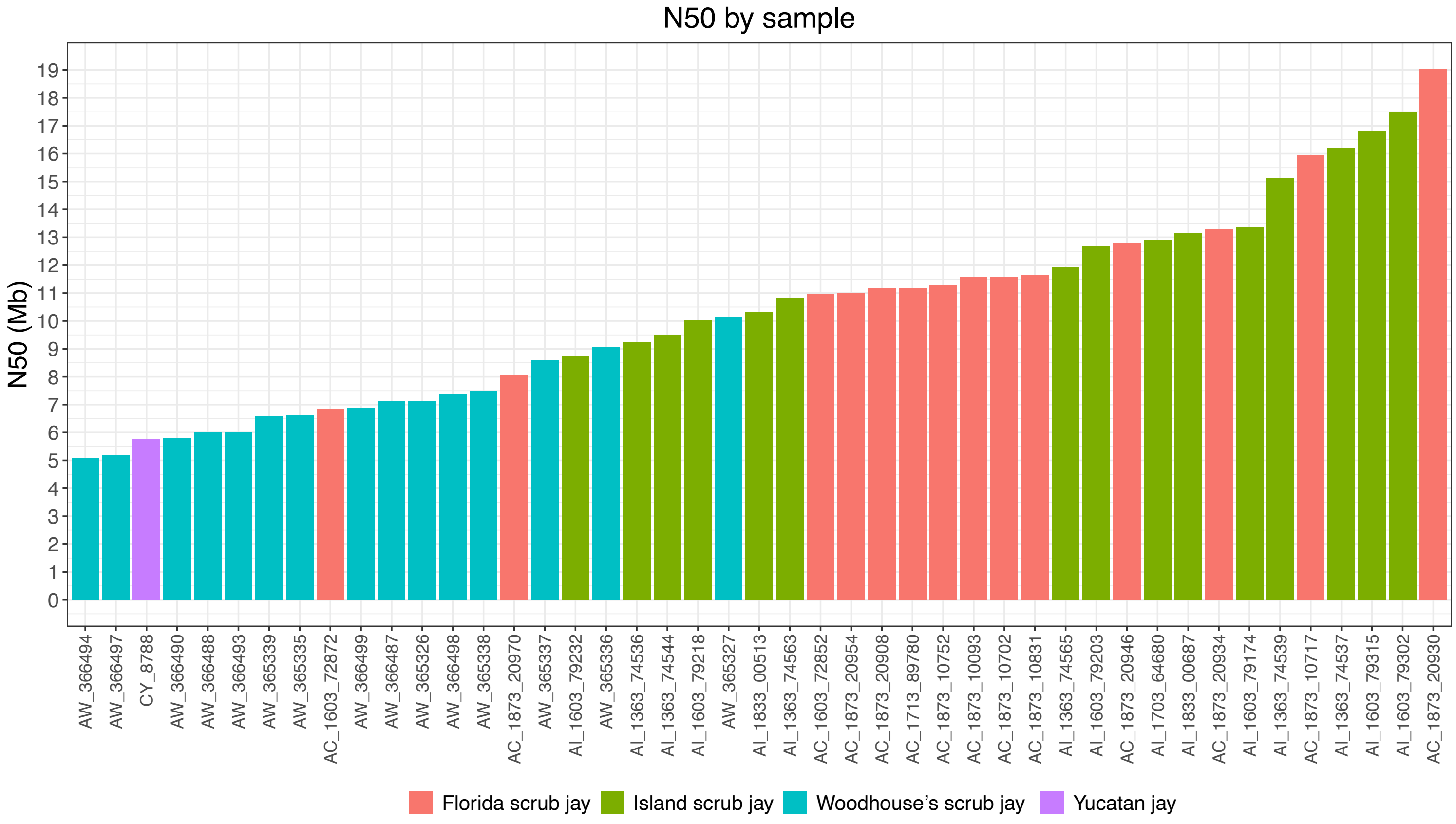
Genome assembly with hifiasm yields ~1.3 Gb primary and haplotype assemblies



assembly	Number of contigs	contig N50 (Mb)
🟥 primary	967	14.75
🟩 haplotype 1	1482	7.84
🟦 haplotype 2	1218	7.90

A. woodhouseii

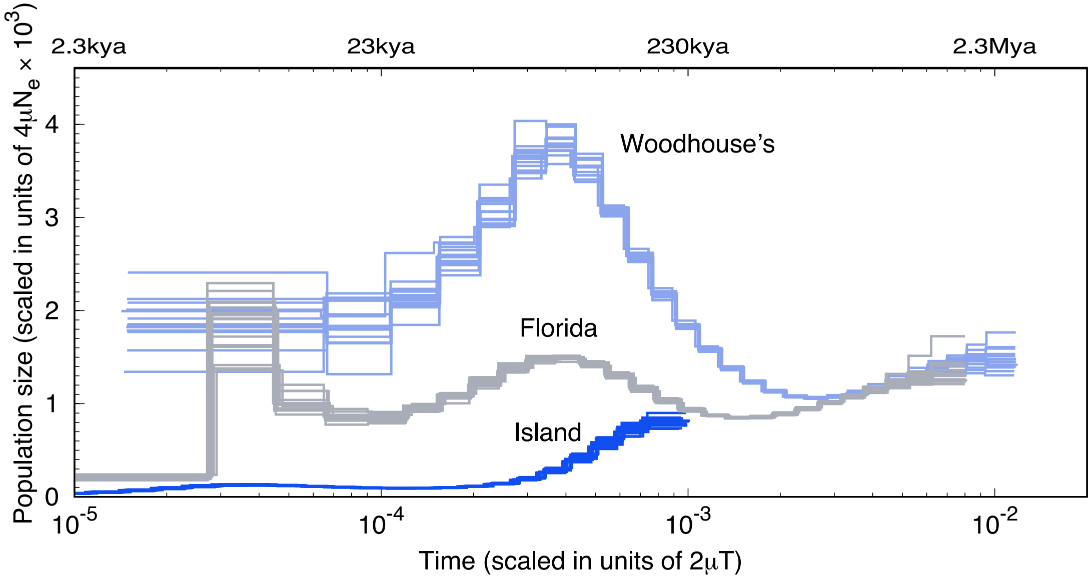
Quality of assemblies varies by species/tissue type



60-fold range in effective population size across species

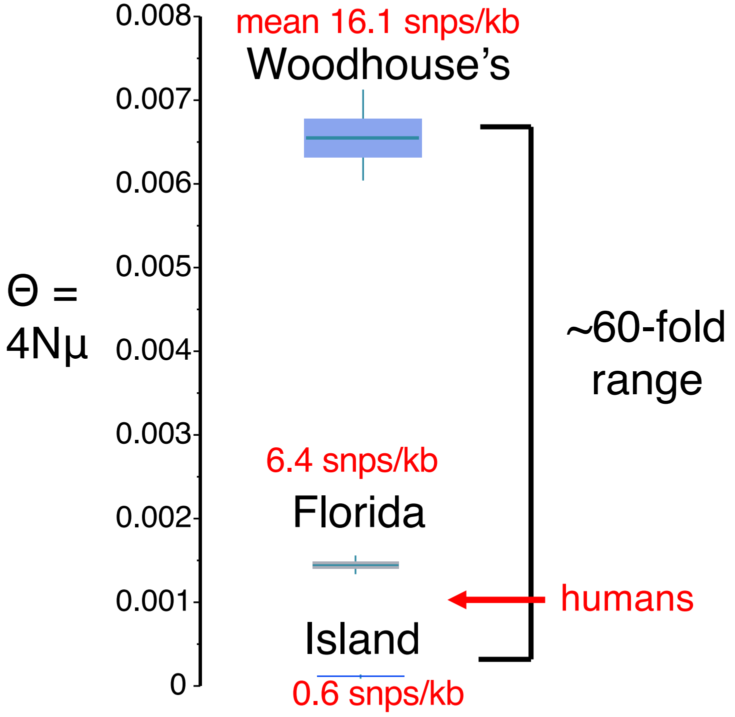
PSMC

SNP data from genome-wide dipcall regions



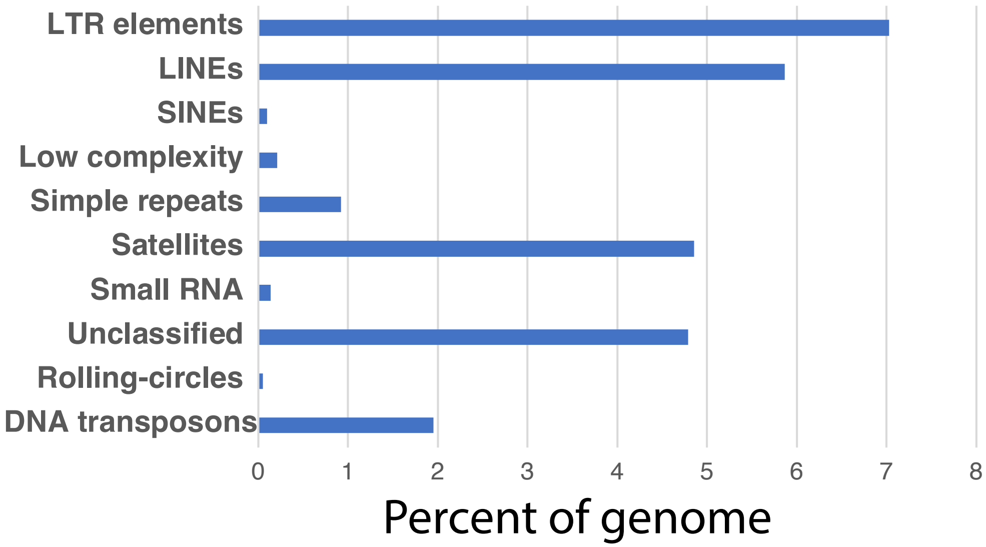
bpp

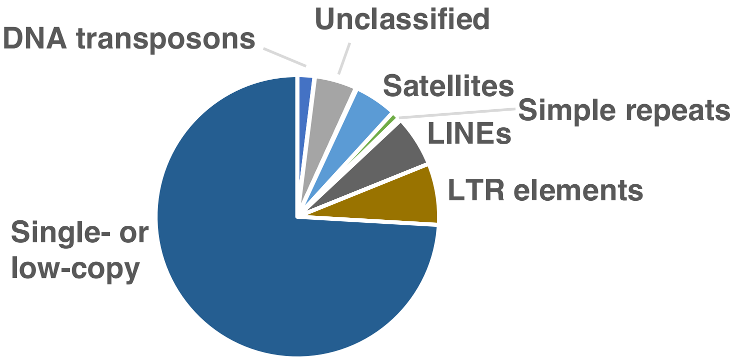
~2500 1-kb dipcall regions



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

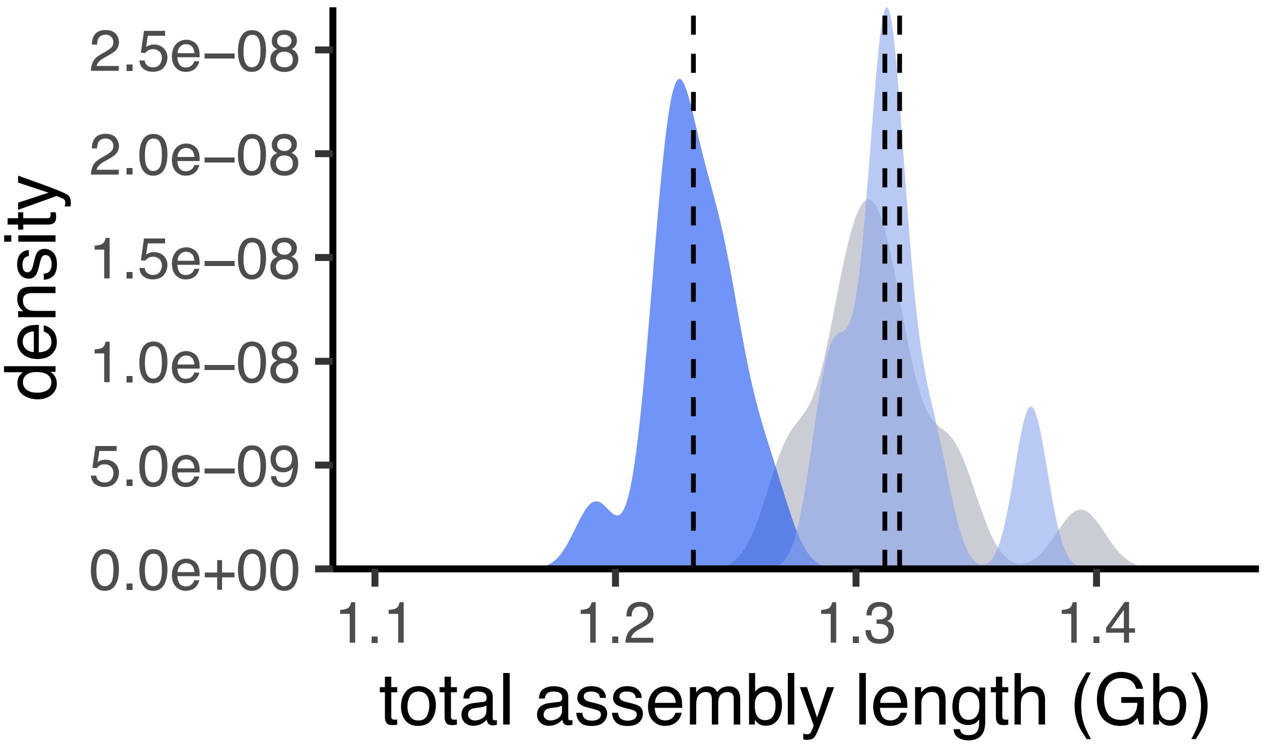
RepeatMasker analysis suggests over 25% repeats and transposable elements





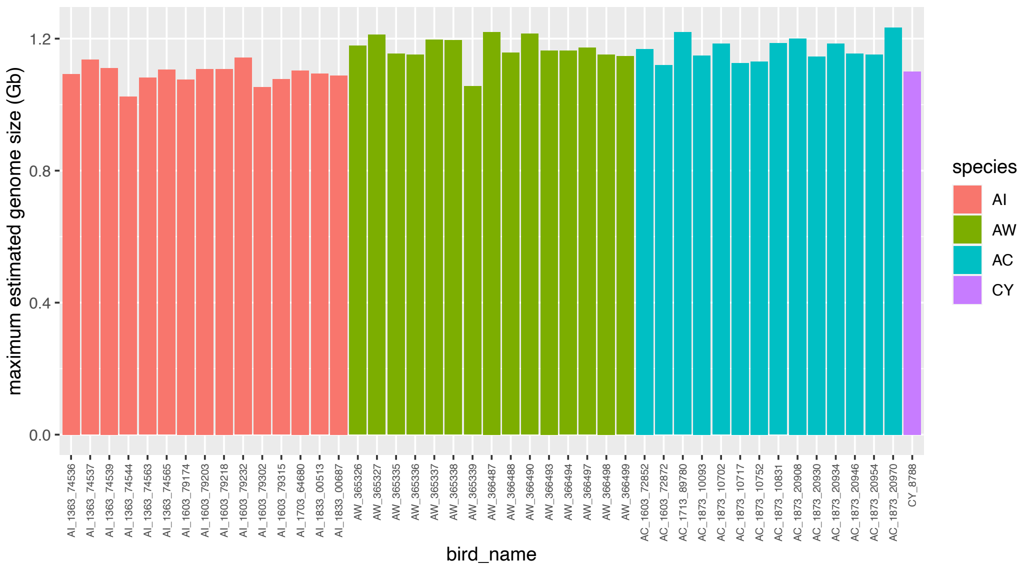
Assemblies of Island Scrub Jays are ~100 Mb smaller than Woodhouse's Scrub Jay

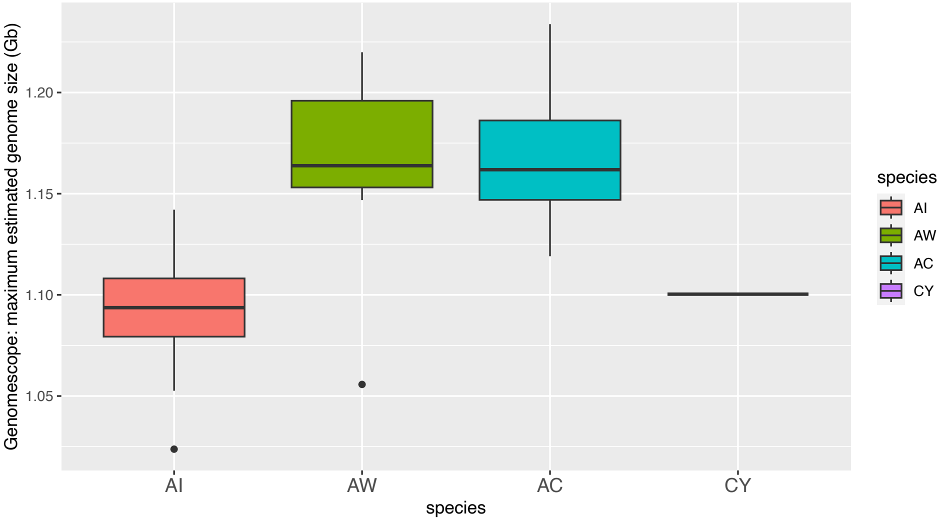
N = 30, 30 and 28 haplotypes (AI, AW, AC)



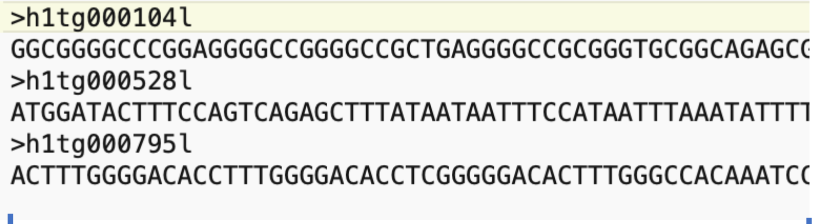
sum of contigs of primary assembly from hifiasm

Estimates of max genome size from Genomescope using k-mers





Pangenome graphs capture structural variation within species



```
>h1tg000104l
GGCGGGGCCCGGAGGGGCCGGGGCCGCTGAGGGGCCGCGGGTGCGGCAGAGCG
>h1tg000528l
ATGGATACTTTCCAGTCAGAGCTTTATAATAATTTCCATAATTTAAATATTTT
>h1tg000795l
ACTTTGGGGACACCTTTGGGGACACCTCGGGGGACACTTTGGGCCACAAATCC
```

unaligned fasta files

→

minigraph

odgi

build genome graph

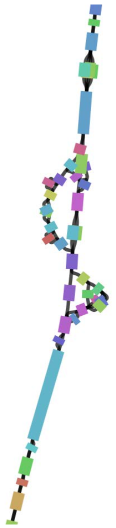
→

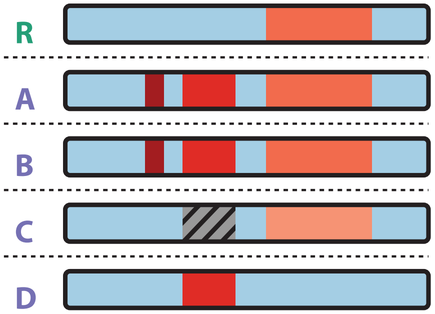
odgi viz

bandage

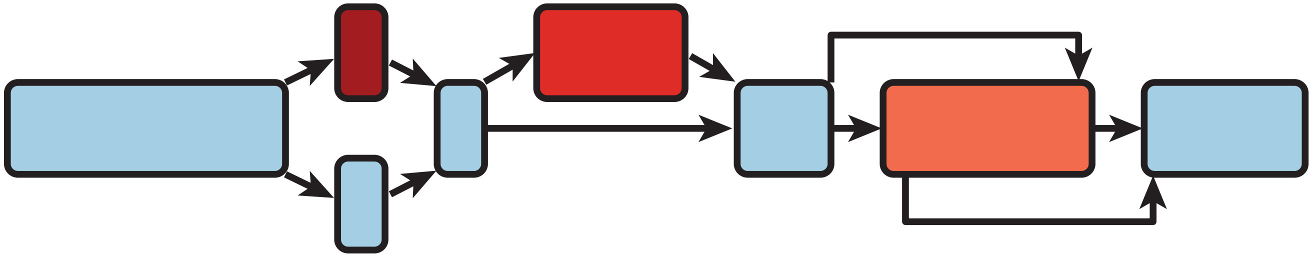
visualize genome graph

→





Multiple sequence alignment

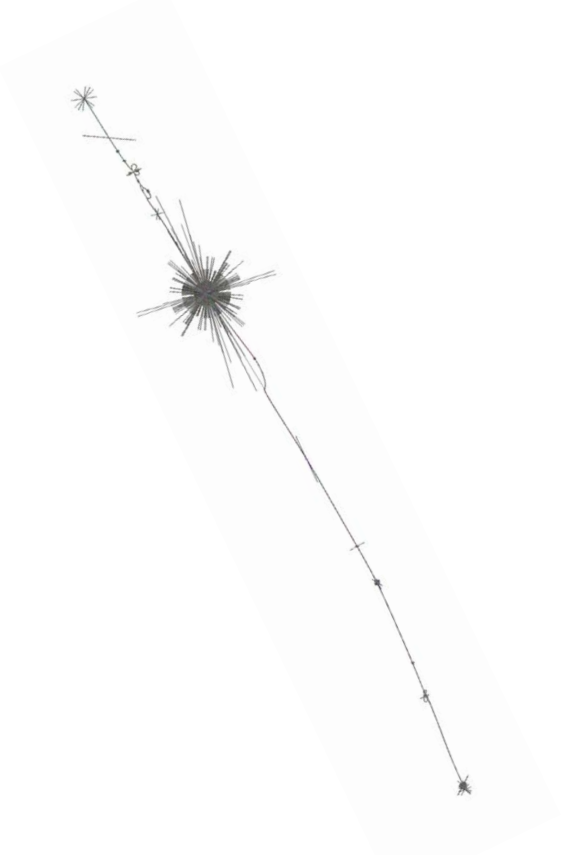


Bidirected genome graph

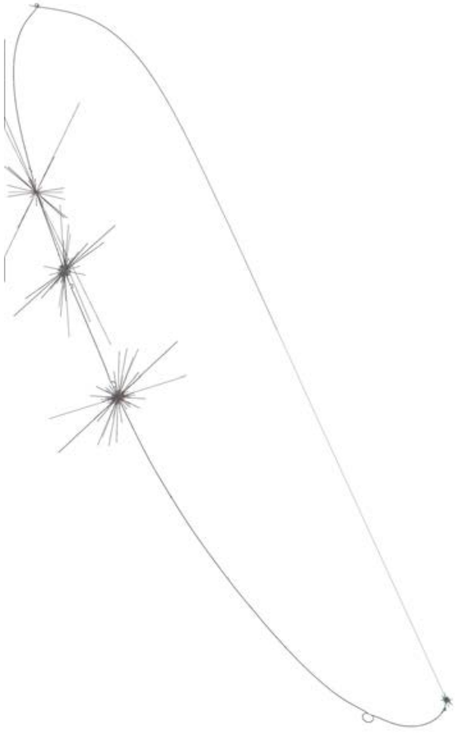
Eizenga et al. 2021. *Ann. Rev. Genomics Hum. Genetics*

2D pangenome graph visualizations – PGGB/Odgi

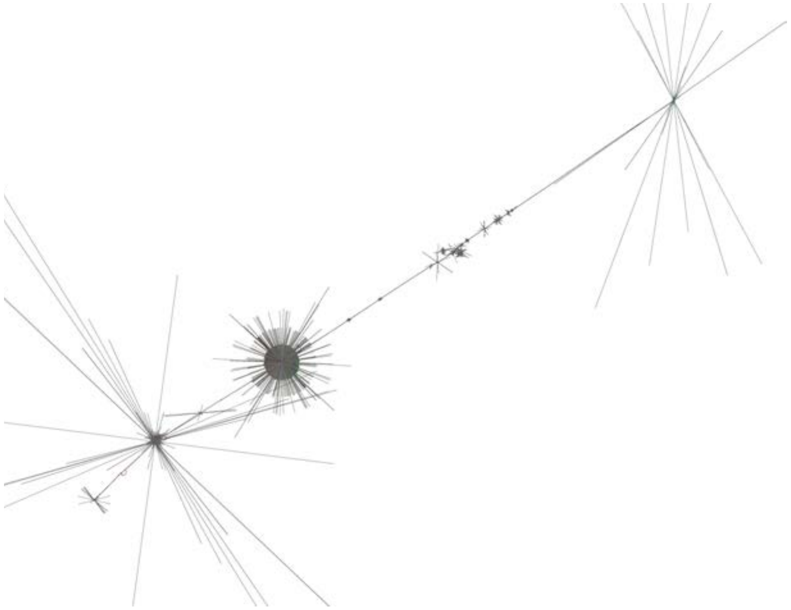
Chr 18 – 12 Mb



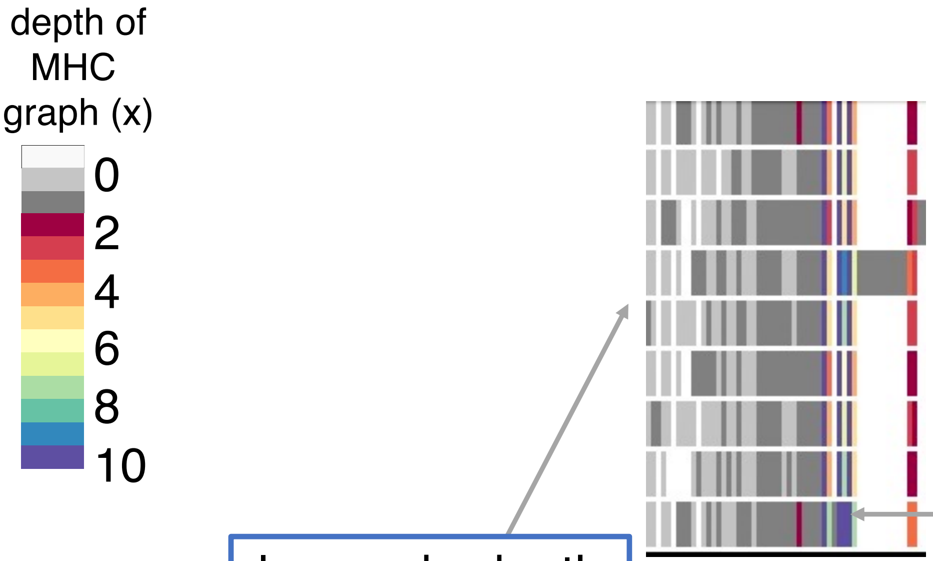
Chr 1 – 160 Mb
‘telomere kiss’



Chr 23 – 8 Mb



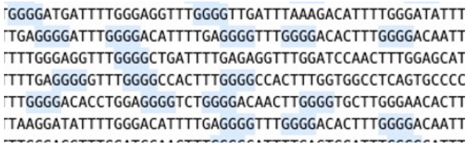
Variation in depth of a pangenome graph



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

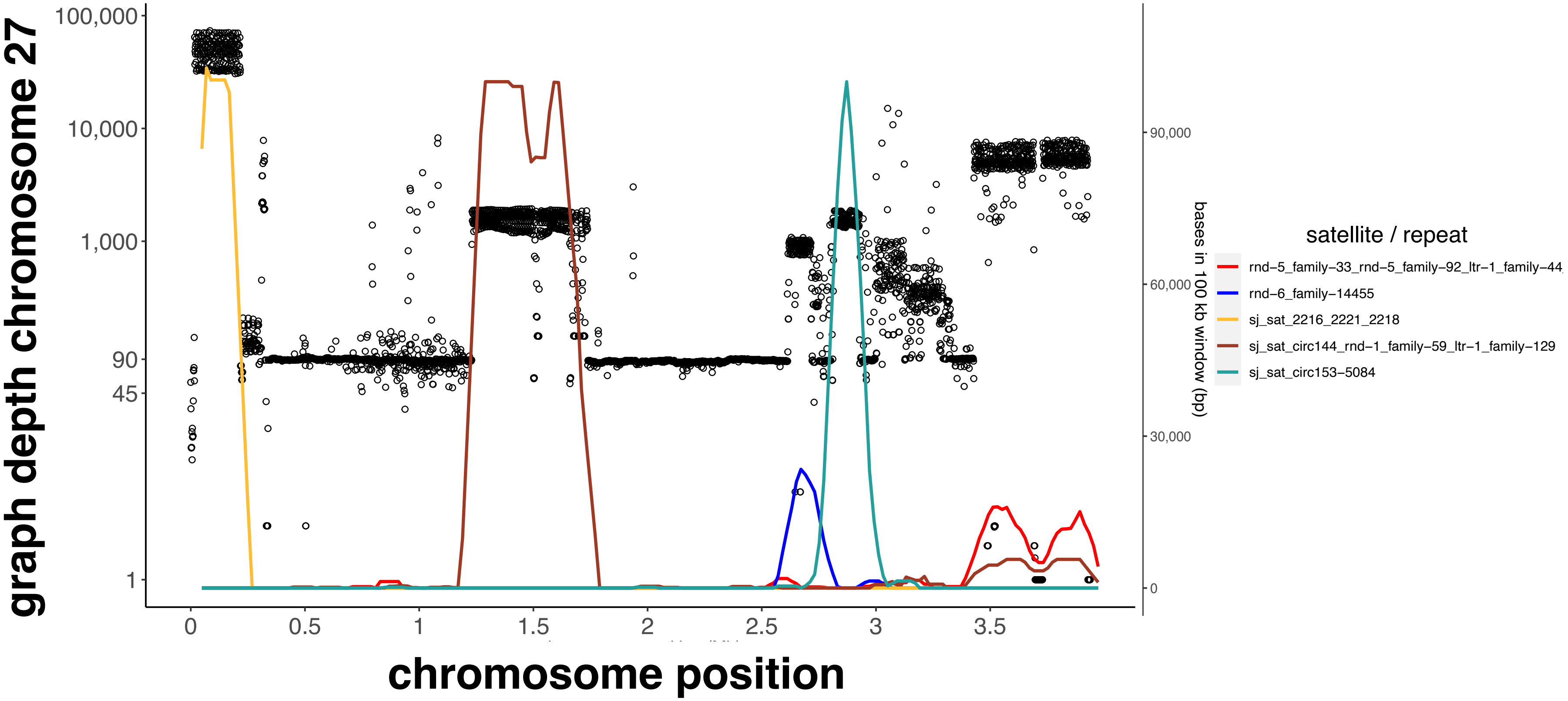
<u>low</u> node-depth regions – SNPs or small indels

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



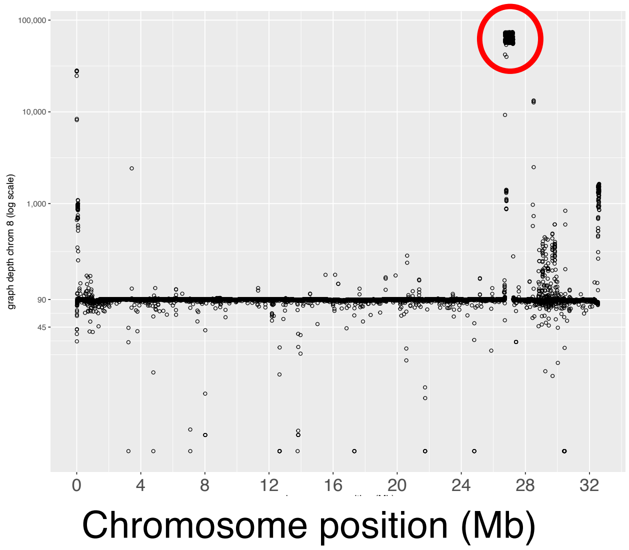
```
TGGGGATGATTTTGGGAGGTTTGGGGTTGATTTAAAGACATTTTGGGATATTT
TTGAGGGGATTTGGGGACATTTTGAGGGGTTTGGGGACACTTTGGGGACAATT
TTTTGGGAGGTTTGGGGCTGATTTTGAGAGGTTTGGATCCAACTTTGGAGCAT
TTTTGAGGGGGTTTGGGGCCACTTTGGGGCCACTTTGGTGGCCTCAGTGCCCC
TTTGGGGACACCTGGAGGGGTCTGGGGACAACTTGGGGTGCTTGGGAACACTT
TTAAGGATATTTTGGGACATTTTGAGGGGTTTGGGGACACTTTGGGGACAATT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

Graph depth of microchromosome 27 correlates with LTRs and satellites

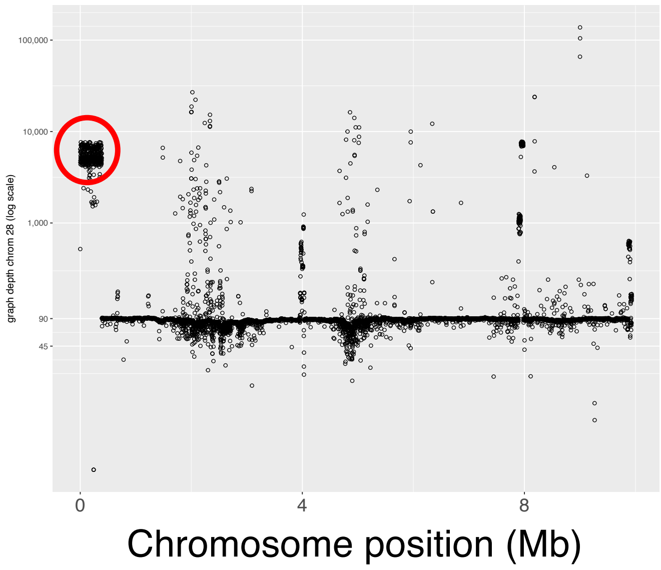


Further examples of graph depth scans

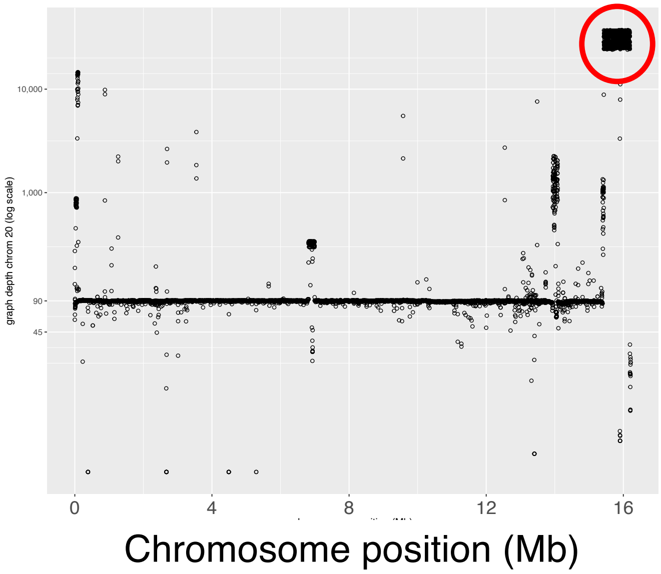
Chr 8



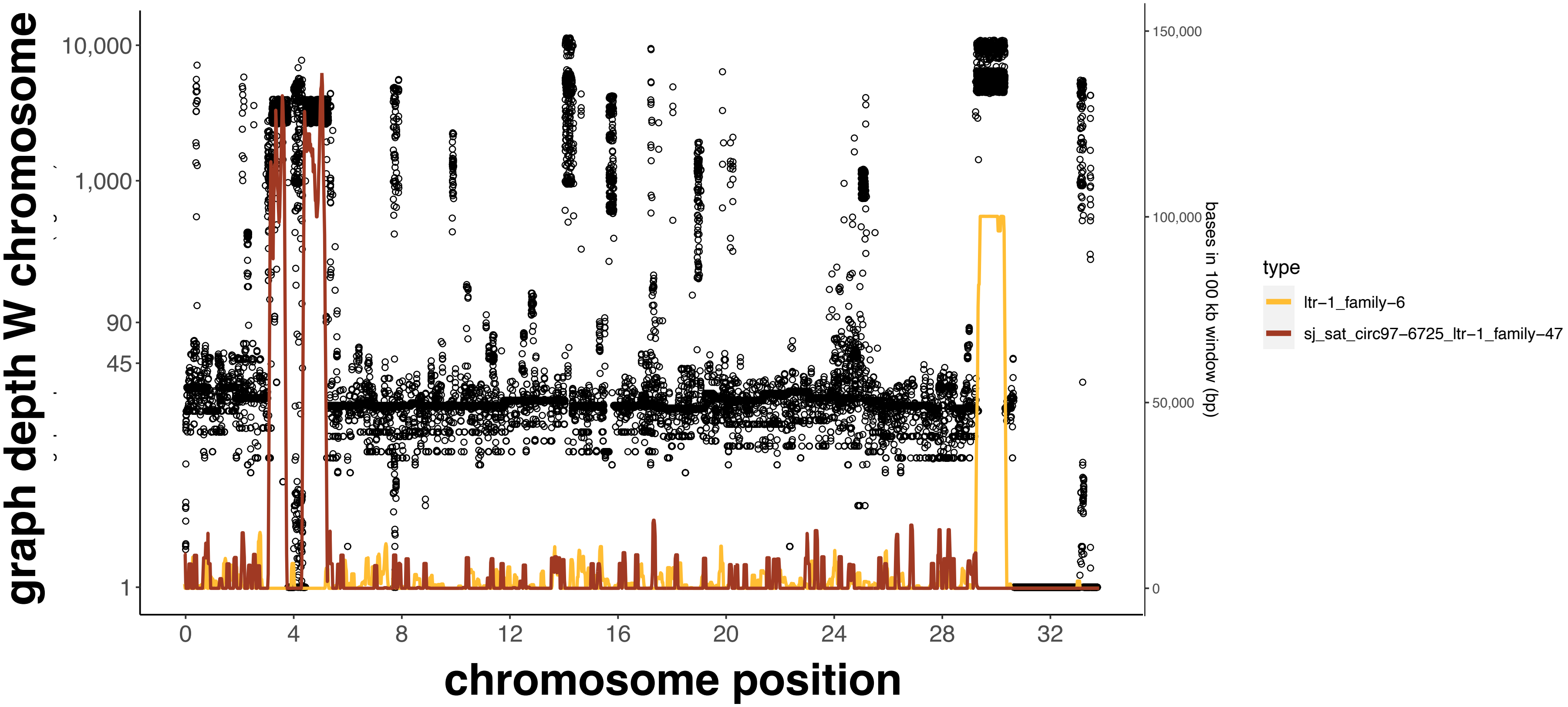
Chr 1A



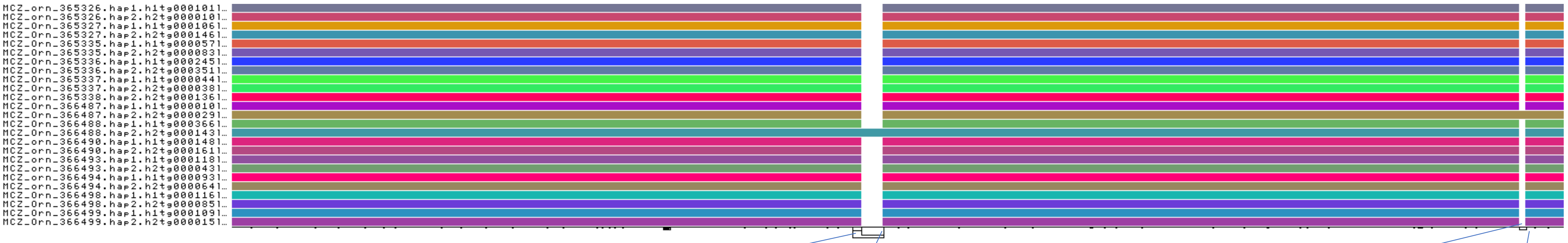
Chr 20



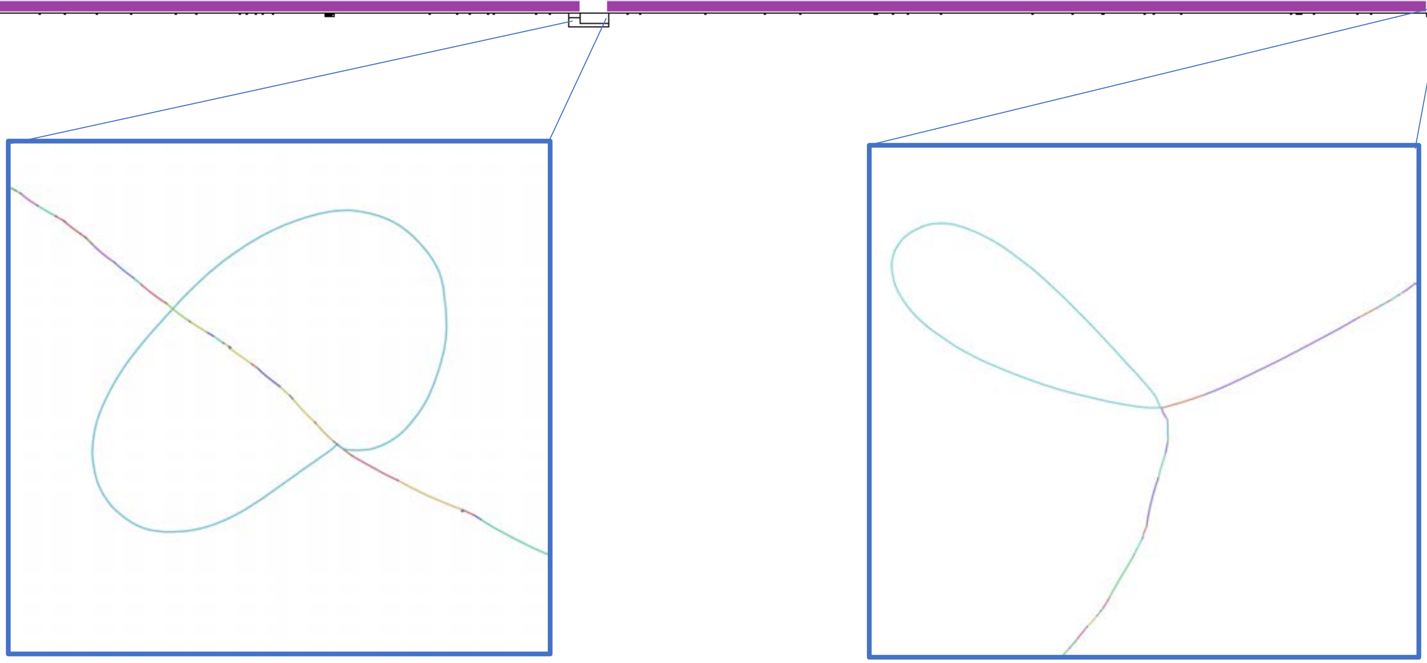
Graph depth of W chromosome correlates with LTRs and satellites



Genomic stability of 400-kb hox1a region in Western Scrub Jays



Pangenome graphs
generated with odgi
and visualized with Bandage

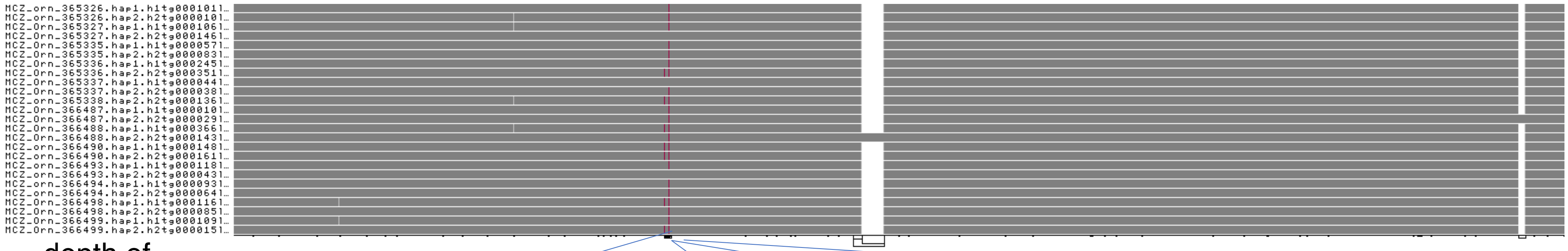


7.5 kb polymorphic indel

2.5 kb polymorphic indel

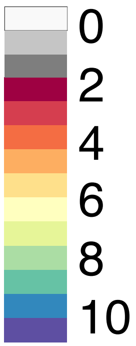
Guarracino et al. 2021.
Bioinformatics, in press.
Wick et al. 2015.
Bioinformatics 31:3350.

Smaller regions of complexity in hox1a region

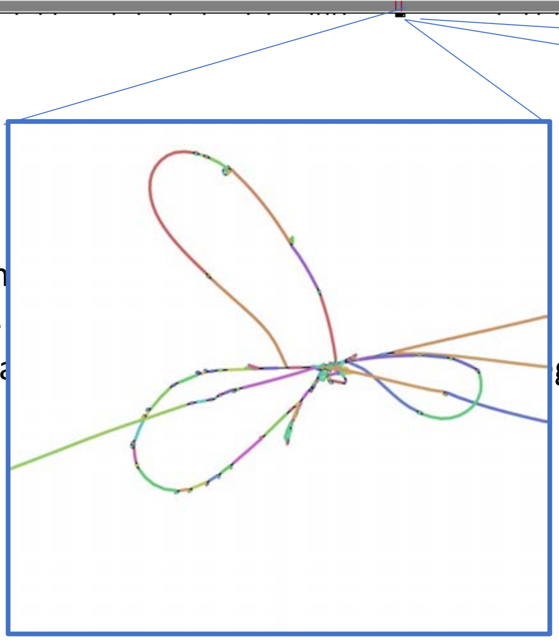


MCZ_orn_365326.hap1.h1tg0001011…
MCZ_orn_365326.hap2.h2tg0000101…
MCZ_Orn_365327.hap1.h1tg0001061…
MCZ_Orn_365327.hap2.h2tg0001461…
MCZ_Orn_365335.hap1.h1tg0000571…
MCZ_Orn_365335.hap2.h2tg0000831…
MCZ_Orn_365336.hap1.h1tg0002451…
MCZ_Orn_365336.hap2.h2tg0003511…
MCZ_Orn_365337.hap1.h1tg0000441…
MCZ_Orn_365337.hap2.h2tg0000381…
MCZ_orn_365338.hap2.h2tg0001361…
MCZ_Orn_366487.hap1.h1tg0000101…
MCZ_Orn_366487.hap2.h2tg0000291…
MCZ_Orn_366488.hap1.h1tg0003661…
MCZ_Orn_366488.hap2.h2tg0001431…
MCZ_orn_366490.hap1.h1tg0001481…
MCZ_orn_366490.hap2.h2tg0001611…
MCZ_orn_366493.hap1.h1tg0001181…
MCZ_orn_366493.hap2.h2tg0000431…
MCZ_orn_366494.hap1.h1tg0000931…
MCZ_orn_366494.hap2.h2tg0000641…
MCZ_Orn_366498.hap1.h1tg0001161…
MCZ_Orn_366498.hap2.h2tg0000851…
MCZ_Orn_366499.hap1.h1tg0001091…
MCZ_Orn_366499.hap2.h2tg0000151…

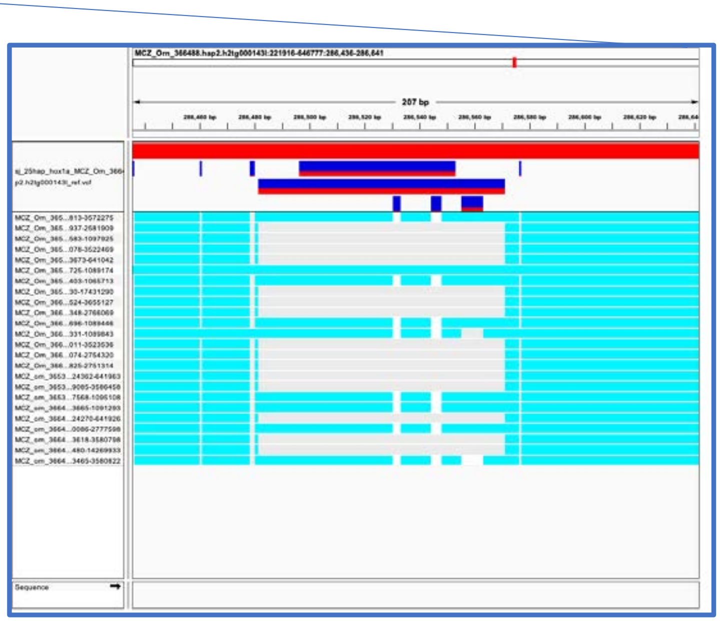
depth of
hox1a region
graph (x)



White region
Gray regions
Red regions a



gions



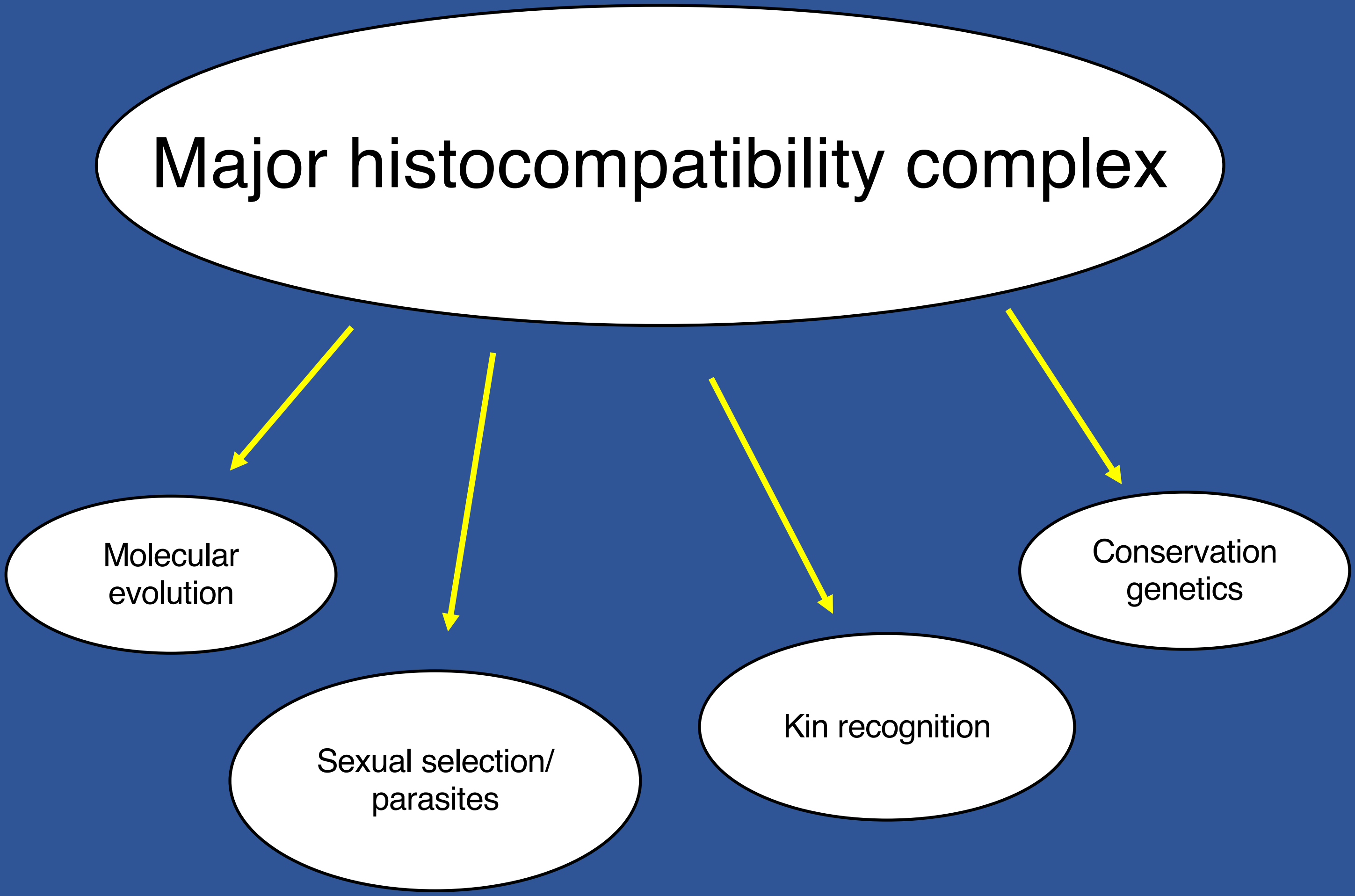
MCZ_Orn_366488.hap2.h2tg000143l:221916-646777:286,436-286,641

207 bp

286,460 bp 286,480 bp 286,500 bp 286,520 bp 286,540 bp 286,560 bp 286,580 bp 286,600 bp 286,620 bp 286,64

sj_25hap_hox1a_MCZ_Orn_366
p2.h2tg000143l_ref.vcf

MCZ_Orn_365...813-3572275
MCZ_Orn_365...937-2681909
MCZ_Orn_365...583-1097925
MCZ_Orn_365...078-3522469
MCZ_Orn_365...3673-641042
MCZ_Orn_365...725-1089174
MCZ_Orn_365...403-1065713
MCZ_Orn_365...30-17431290
MCZ_Orn_366...524-3655127
MCZ_Orn_366...348-2766069
MCZ_Orn_366...696-1089446
MCZ_Orn_366...331-1089843
MCZ_Orn_366...011-3523536
MCZ_Orn_366...074-2754320
MCZ_Orn_366...825-2751314
MCZ_orn_3653...24362-641963
MCZ_orn_3653...9065-3586458
MCZ_orn_3653...7568-1096108
MCZ_orn_3664...3665-1091293
MCZ_orn_3664...24270-641926
MCZ_orn_3664...0086-2777598
MCZ_orn_3664...3618-3580798
MCZ_orn_3664...480-14269933
MCZ_orn_3664...3465-3580822



Major histocompatibility complex

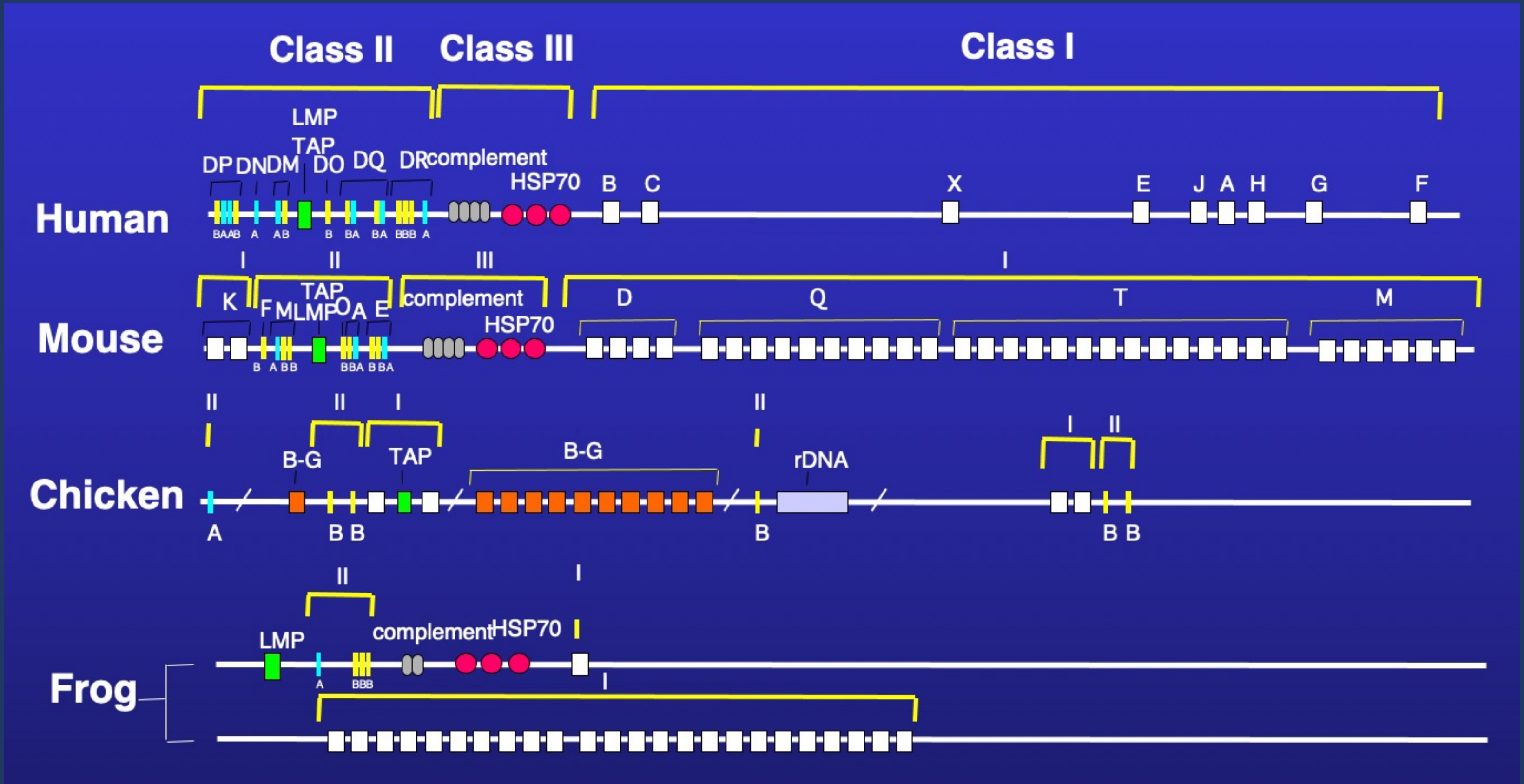
Molecular evolution

Sexual selection/ parasites

Kin recognition

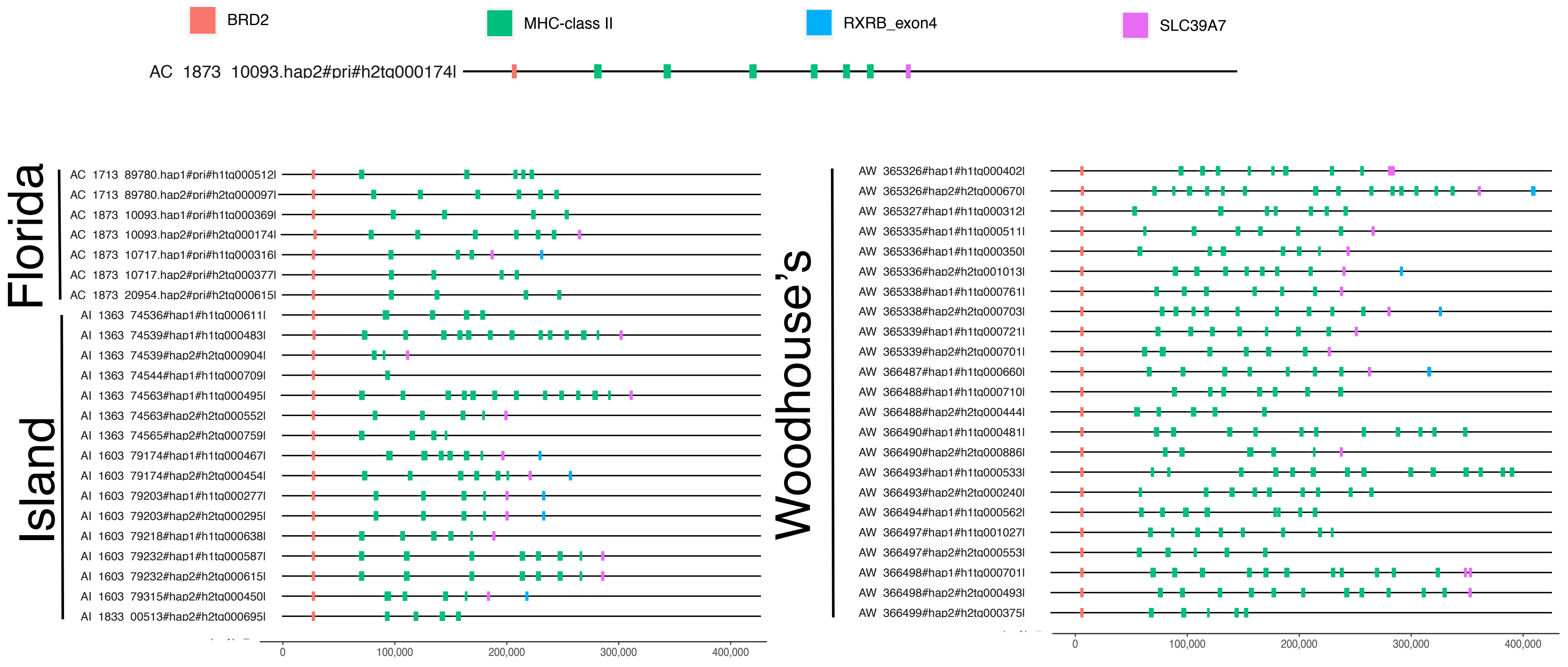
Conservation genetics

The chicken MHC is small (~99 kb) and compact

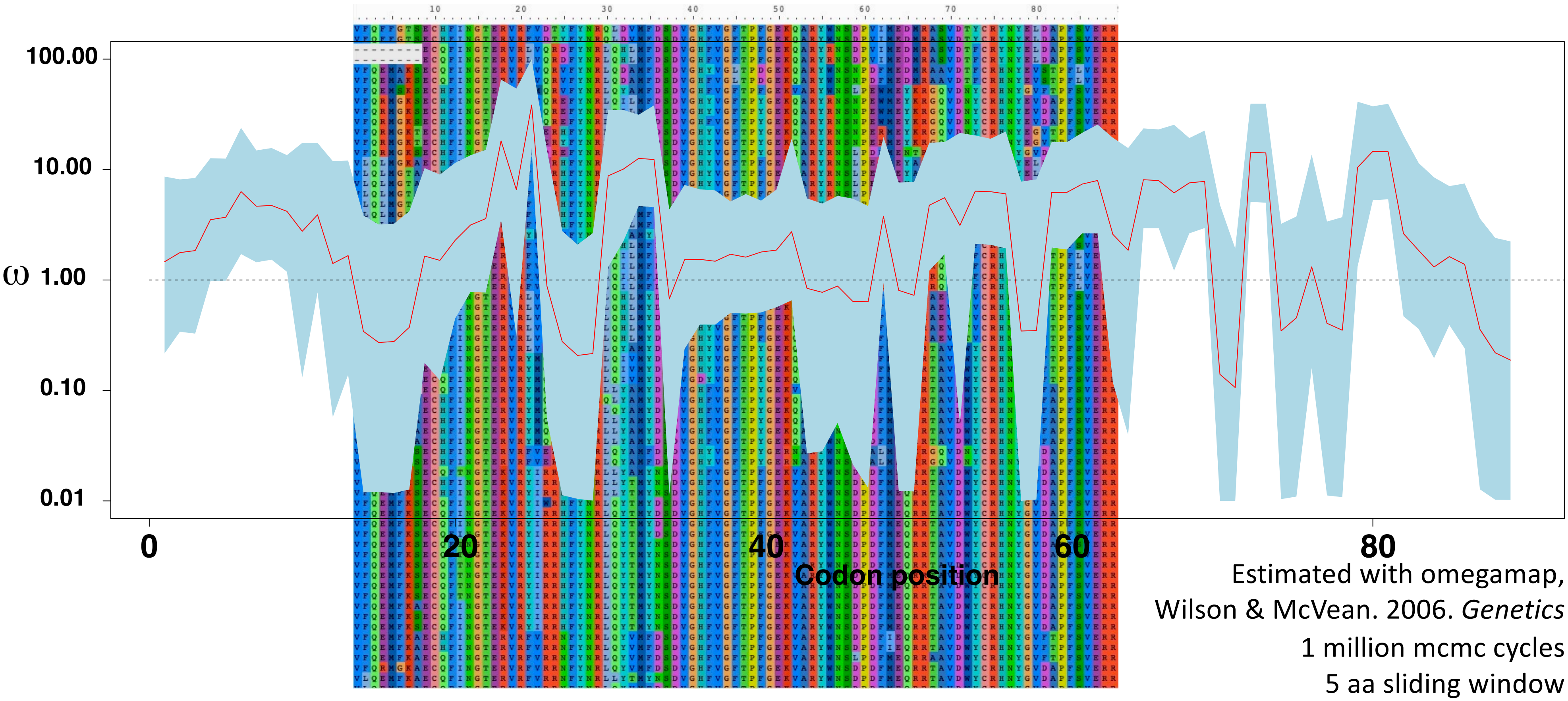


(not to scale)

Extraordinary haplotype diversity in MHC class II region in Scrub Jays

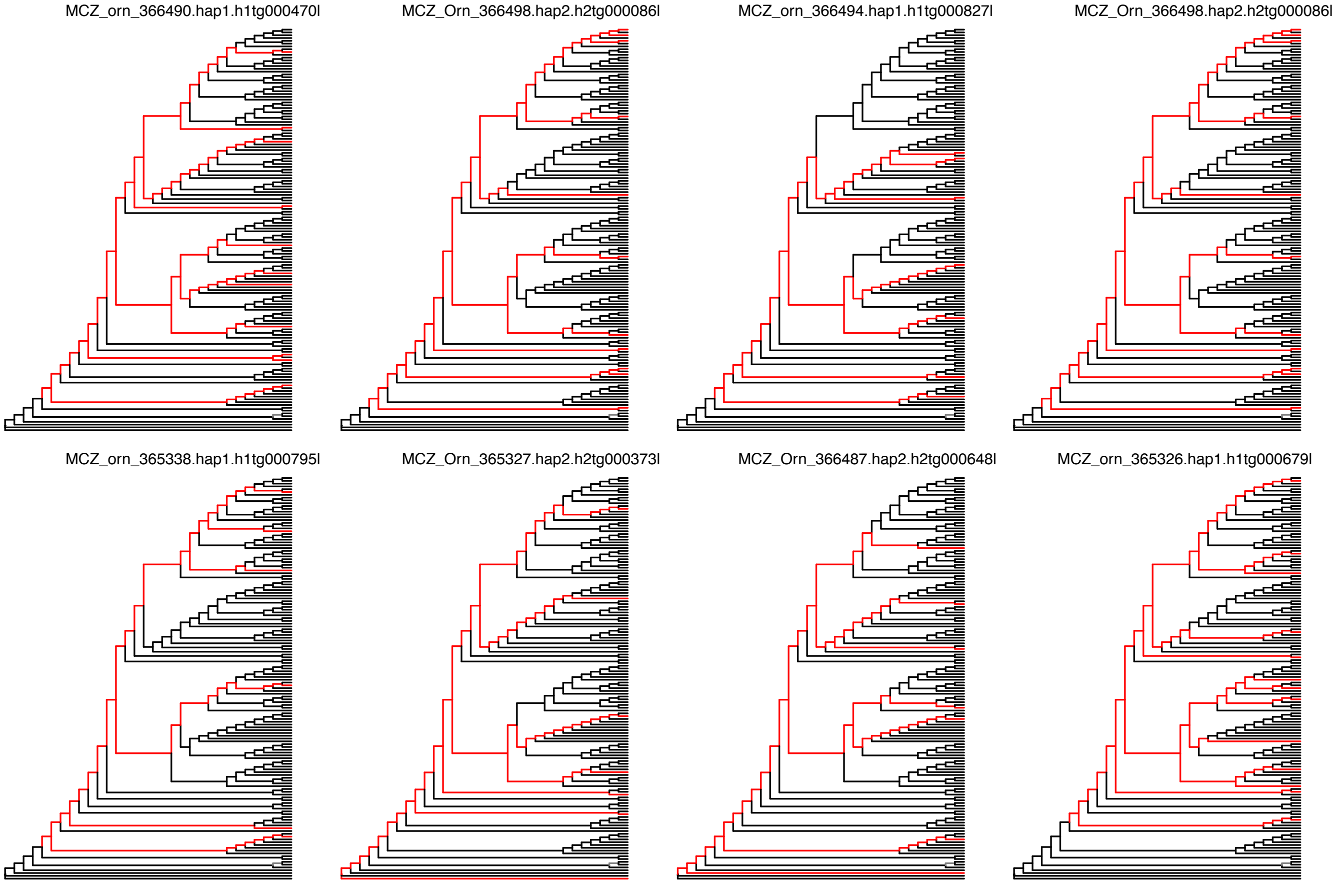


Mhc class II peptide-binding region shows solid evidence of balancing selection



Estimated with omegamap,
Wilson & McVean. 2006. *Genetics*
1 million mcmc cycles
5 aa sliding window

Mhc class II peptide binding regions are phylogenetically diverse on individual haplotypes


MCZ_orn_366490.hap1.h1tg000470l

MCZ_Orn_366498.hap2.h2tg000086l

MCZ_orn_366494.hap1.h1tg000827l

MCZ_Orn_366498.hap2.h2tg000086l

MCZ_orn_365338.hap1.h1tg000795l

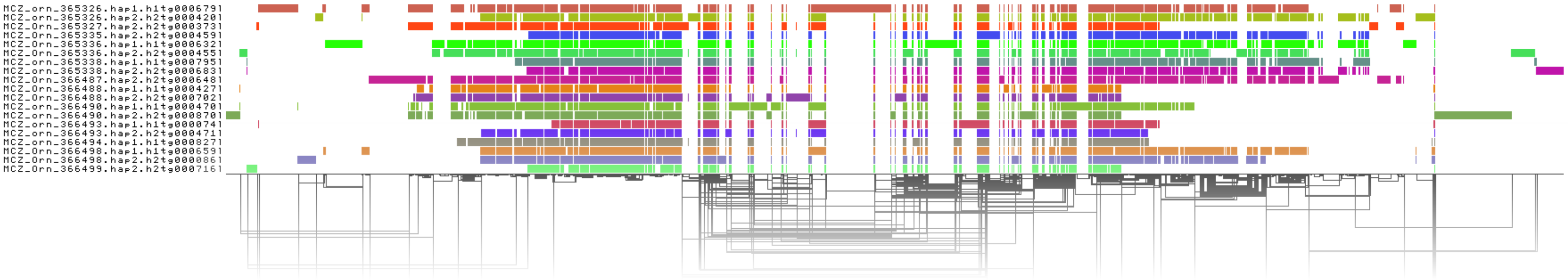
MCZ_Orn_365327.hap2.h2tg000373l

MCZ_Orn_366487.hap2.h2tg000648l

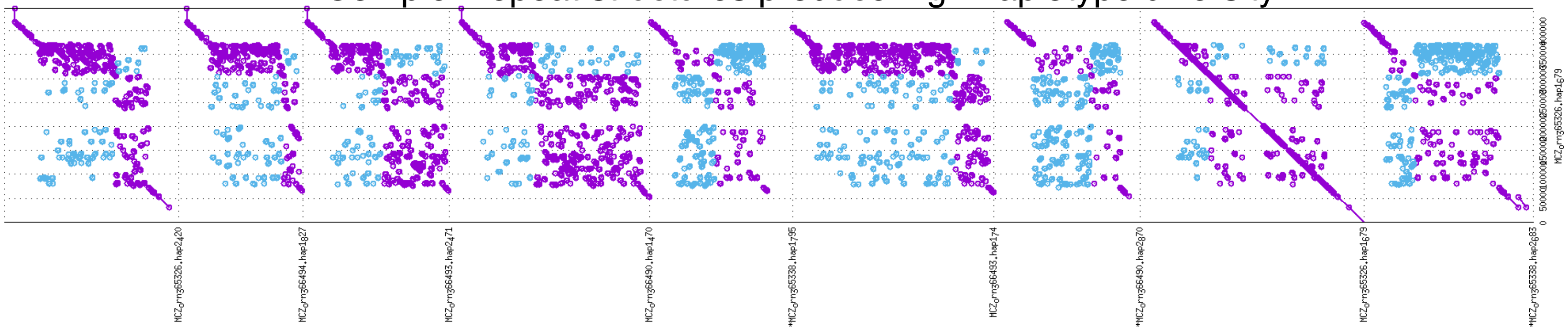
MCZ_orn_365326.hap1.h1tg000679l

🟥 Phylogenetic paths of
Mhc exon2 alleles
on individual haplotypes

Visualization of MHC class II region in 22 haplotypes of Woodhouse's scrub-jays with odgi


MCZ_orn_365326.hap1.h1tg000679l
MCZ_orn_365326.hap2.h2tg000420l
MCZ_Orn_365327.hap2.h2tg000373l
MCZ_Orn_365335.hap2.h2tg000459l
MCZ_Orn_365336.hap1.h1tg000632l
MCZ_Orn_365336.hap2.h2tg000455l
MCZ_orn_365338.hap1.h1tg000795l
MCZ_orn_365338.hap2.h2tg000683l
MCZ_Orn_366487.hap2.h2tg000648l
MCZ_Orn_366488.hap1.h1tg000427l
MCZ_Orn_366488.hap2.h2tg000702l
MCZ_orn_366490.hap1.h1tg000470l
MCZ_orn_366490.hap2.h2tg000870l
MCZ_orn_366493.hap1.h1tg000074l
MCZ_orn_366493.hap2.h2tg000471l
MCZ_orn_366494.hap1.h1tg000827l
MCZ_Orn_366498.hap1.h1tg000659l
MCZ_Orn_366498.hap2.h2tg000086l
MCZ_Orn_366499.hap2.h2tg000716l

Complex repeat structures produce high haplotype diversity

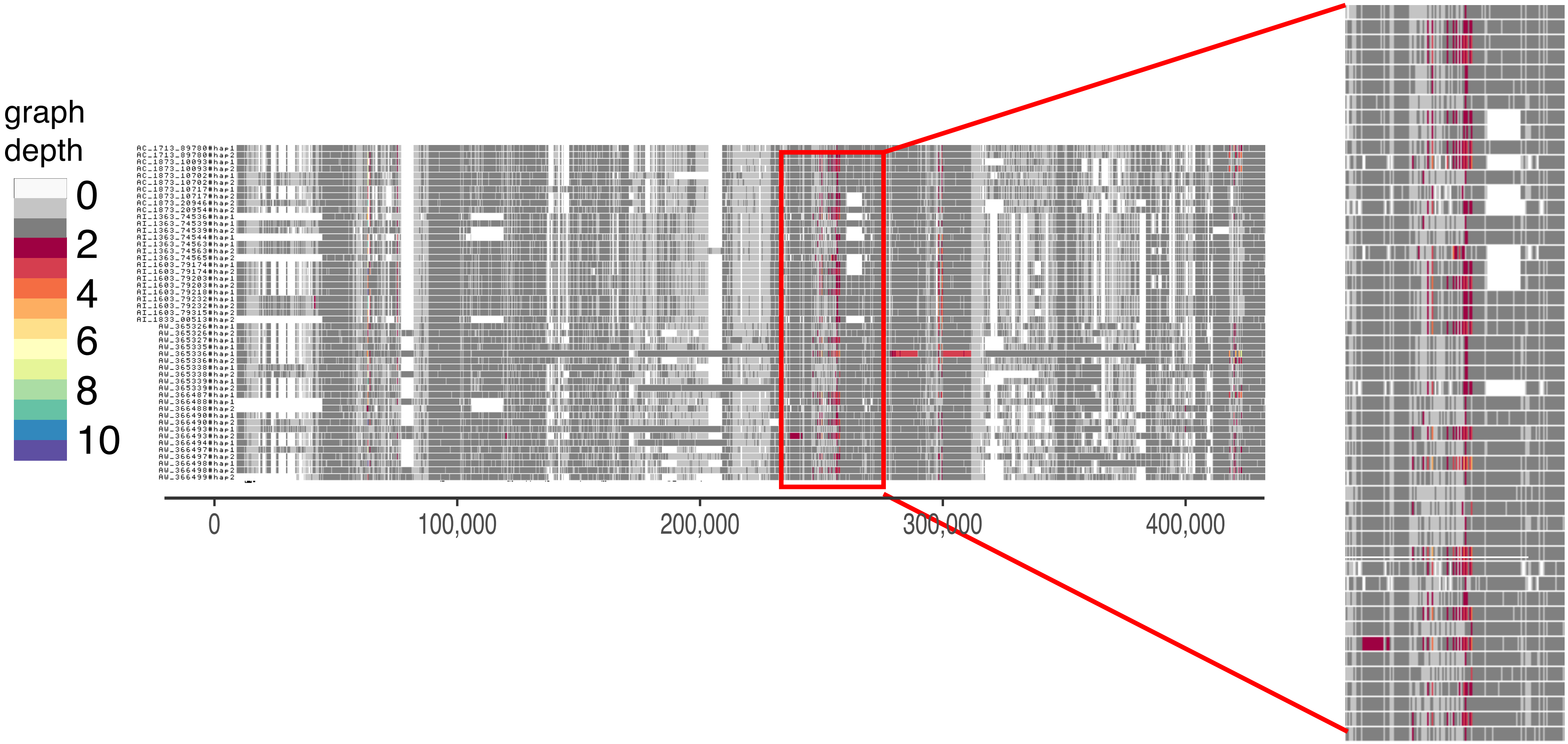


made with odgi and pangenome graph builder pipeline
Guarracino et al. 2021. *Bioinformatics*, in press.

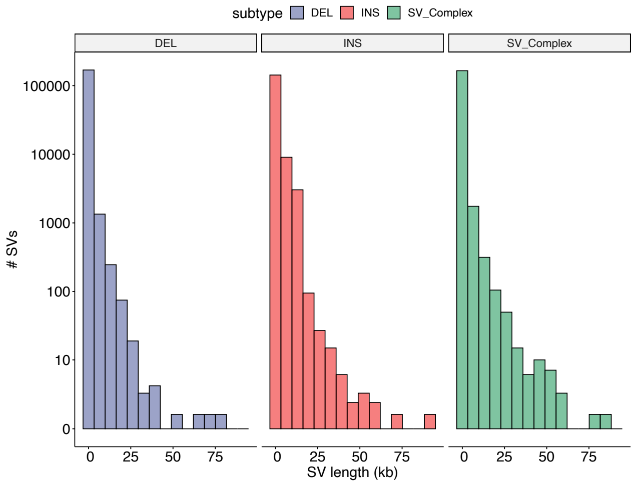
Example satellites in MHC class II region of Woodhouse’s scrub jay

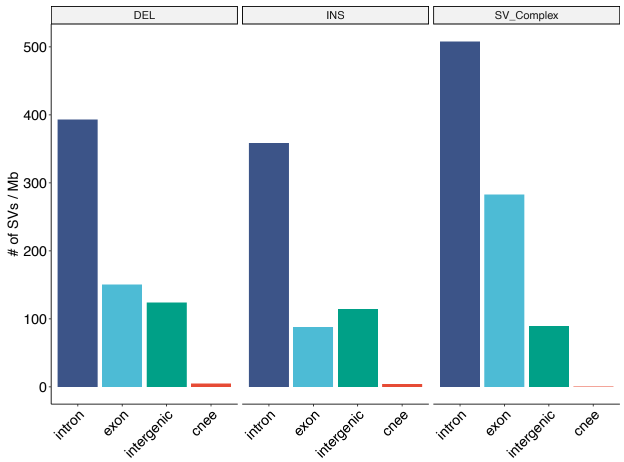
```
TTCTGTGTGGTTTTGGGGTGGTTTTAGGTGGTTTTGGGGAGGTTTTAGGGGGGTTTGGGGAGGTTTTAGGGGGTGTTGGGGAGGTTTTAGGGGGATTTTGGGGGGCTTGGGGAGACTTTGGGGTCACTGGGACACCTCGGGGACAGTTTTGGGGCGGGTTTGAGGCGGTTTTGGGGCTGTTTGAG                              GGGGT
GGTTTTAGGGCGGTATTGGGGTGGGTTTAGGGGGTGTTGGGGGGGTTTTAGGGGGGTTTGGGGGGCCTTGGGGACACTTTGGGGGCACTGGGACACCTTGGGGACAGTTTTGGGGCGGGTTTGAGGCGGTTTTGGGGCTATTTGGGGCAATTTTGGGGCAATTTCTGTGTGGTTTTGGGGCGGTT                              AGGTT
TTAGGGGGTGTTGGGGAGGTTTTAGGGGGTTTTGGGGGGCTTAGGGACACTTTGGGGGCACTGGGACAGCTCGGGGACAGCTTTAGGGTGGTTTGGGGTGGGTTTGGGGCGGTTTTGGGGCTGTTTGGGGCGGTTTTGGGGCATTTTGGGGCAATTTTGGGGCTTTTTGGGGTGGTTTTAGGGCG                              GGGGT
TTTGGGGGGCCTTGGGGACACTTTGGGGGCACTGGGACACCTTGGGGACAGTTTTAGGGTGGGTTTGGGGCGGTTTTGGGGCTATTTGGGGCAATTTTGGGGCAATTTCTGTGTGGTTTTGGGGCGGTTTTAGGGGGAGTTCCGGGCAATTTCTGCGCGGGTTTGGGGAGGTTTTAGGGGGTGTT                              TTTGG
GGTCACTGGGACACCTCGGGGACAGTTTTGGGGCGGGTTTGAGGCGGTTTTGGGGCTGTTTGAGGCAATTTTGGGGCATTTTGGGGCAGTTTTGGGGCGGTTTTAGGGGGAGTTCCGGGCAATTTCTGCGCAGTTTTGGGGAGGTTTTGGGGGTGTTGGGGAGGTTTTAGGGGGGTTTGGGGAAA                              GCACT
GGGACACCTCGGGGACAGTTTTGGGGCGGTTTTGGGGCAATTTCTGTGCAGTTTTGGGGCGGTTTTAGGGGGGGTTTGGGGCAATTTCTGCGCGGTTTTGGGGTGGTTTGGGGTGGTTTTGGGGGGCTTGGGGCCACTTTGGGGGCACTGGGACAGCTCGGGGACAGTTTTAGGGCGGGTTTGGG                              TGCGC
GGTTTCGGGGAGGTTTTAGGGGGGTGTTGGGGGGGTTAGGGGGGTTTTGGGGGGTTTGGGGACACTTTGGGGGCACTGGGACACCTCGGGGACAGTTTTAGGGTGGGTTTGGGGCGGTTTTGGGGCAATTTTGGGGCTTTTTGGGGTGGTTTGGGGCAATTTCTGCGCGGTTTCGGGGAGGTTTT                              CTTGG
GGACACTTTGGGGGCACTGGGACAGCTCGGGGACAGTTTTAGGGCGGGTTTGAGGCGGTTTTGGGGCTATTTGGGGCTTTTTGGGGCAATTTCTGTGTGGTTTTGGGGTGGTTTTAGGGGGTGTTCGGGAGGTTTTCGGGGGGTTTGGGGTGGTTTTAGGGGGATTTTGGGGGGCTTGGGGAGAC                              GGGTT
TGAGGCGGTTTTGGGGCTGTTTGAGGCAATTTTGGGGCATTTTGGGGCTGTTTTGGGGCAGTTTTGGGGTGGTTTTAGGGGGAGTTTGGGGCAATTTCTGCGCGGTGTTGGGGAGGTTTTAGGGGGGTTTTGGGGACCTTGGGGACACTTTGGGGTCACTGGGACACCTCGGGGACCGTTTTAGG                              GGGGC
AATTTCGGGGAGTTTTGGGGTGATTTGGGGCGGTTTTAGGGTGGTTTTAGGGGGGTTTGGGGAGGTTTTAGGGGGGCTTGGGGGCCTTGGGGGACACTTTGGGGGCACTGGGACACCTCGGGGACAGTTTTAGGATTGTTTTGGCAATTTTGGGGCTGTTTGGGGCAATTTTGGGGCAGTTTTGGGGTGGTTTGGGGCAATTTTTGGGCGGTTTGGGGCAGTTCGGGG
CCATTTGGGGCAATTTCTGTGCAGTTTTGGGGGGGTTTGGGGTGGTTTTGGGGGGCTTGGGGACACTTTGGGGGCACTGGGACACCTTGGGGACATTTTTAGGGTGGTTTGGGGTGGTTTTGGGGCTGTTTGGGGCAATTTTGGGGCTTTTTGGGGTGGTTTGGGGCAATTTCTGTGCGGTTTGGGGGTTTGGGGCGGTTTAGGGGGGTTTGGGGAGGTTTTGGGGTGG
TTTTAGGGGGTTTGGGGTGGTTTTGGGGGGCCTTAGGGACACTTTGGGGGCACTGGGACACCTTGGGGACAGTTTGGGACAATTTTGGGGCAATTTCTGGGTGGTTTTGGGGTGGTTTTAGGGGGGATTGGGGTGGTTTAGGGTGGTTTGGGGCGATTTTAGGGGGGTCTGGGGCGGTTTTGGGGGTTTTTGGGGTGGTTTTGGGGGGCTTGGGGTCACTTTGGGGACA
CTGGGACAACTTGGGGACACTTTTGGGGGGAGCTCGGGGCTTTTGGGGTGATTTGGGGCGGTTTTGGGGTGGTTTGGGGTGGTTTTGGGGTGGTTTTAGGAGGTTTTGAAGGCCTTTTGGGGTATTTTTAAGGGATTTGGGGACATTTAGGGACATCCGTGCCCCCTCCTTGTGTCCCCACAGCCCCCCACGAGCACCACGAGGGCCACGGCCACTCCCACCACGGGCA
CGGTGAGAGCCCAGGGCACCCTGGGGACATTTTGGGGACGATTTTGGGGGGTTTGGGGCCAATTTTGGGGCATTTTGAGGGGATTTGGGGACGTTTTTAGGGGGTTTGGGGACACCTGGAGGGGTCTGGGGACACCTTGGGGTGGGTGGGGACACTTTGGGGACAATTTTGGGAGGTTTGGGGCCAAATTAGGACATTTTGAGGGGATTTGGGGACATTTTGAGGGGGT
TTGGGGACATTTTGGGGACGATTTGGGGGGATTTGGGGACGATTTGGGGGGATTTGGGGCCAATTTGGGGCATTTTGAGGGGATTTGGGGACATTTTGAGGGGTTTGGGGACGATTTTGGGAGGTTTGGGGTTGATTTTGGGGCATTTTTAAGGGTTGTGGGGCCACCTGGAGATGGTTGGGGACACTCAGGGACATCAGGGCCACTTTGGGGATGATTTGGGGAGGTT
TGGGGTCCATTTAAGGCCATTTTGGGACATTTTGGGACGTTTTGAGGGGATTTGGGGCATTTTGAGGGGATTTGGGGCCATTTTGAGGGCTTTGGGGCCACTTTGGGGCCACTTTGGTGGCCTCAGTGCCCCCTTTTGTCCCTGCAGGTCCCCACGAGGGCCACGGCCACTCCCACCACGGGCACAGTGAGAGCCCAGGGCACCCTGGGGACATTTTGGGGACGATTTT
GGGGGGGTTGAGACCAATTTTGGGGTATTTGGAGGGGATTTGGGGCATTTTGAGGGGATTTGGGGACGTTTTTAGGGGGTTTGGGGACACCTGGAGGGGTCTGGGGACACCTTGGGCTGGGTGGGGACACTTTGGGGACGATTTGGGGAGGTTTGGGGTTGATTTAAGGGCATTTTGGGACATTTTGAGGGGTTTTGGGAACACTTTGGGGACGATTTGGGGAGGTTTG
GGGCCAATTTTGGGTCCATTTTGAGGGGATTTGGGGCATTTTGAGGGGATTTGGAGACACCTGGAGGGGTCTGGGGACACCTTGGGGTGGTTGGGGACACTTTGGGGACGATTTTGGGAGTTGTGGGGCCAATTTTGGGGCATTTTGAGAGGATTTGGGGCTGATTTTGGGAGGTTTGGATCCAACTTTGGGGCATTTTGAGGGGATTTGGGGCATTTTTAGGGGGTTT
GGGGATGATTTTGGGAGGTTTGGGGCCAATTTTGGGGACATTTTGAGGGGATTTGGGGACATTTTGAGGGGATTTGGGGACATTTTGAGGGGTTTGGGGACAATTTTGGGAGGTTTGGGGTTGATTTAAGGACATTTTGGGGCTTTTTGAGGGGATTTGGGGACATTTTGGGGGGGTTTGGGGATACCTGGAGGGGTCTGGGGACACCTTGGGGTGGTTTGGGACATTT
TGGGGATGATTTTGGGAGGTTTGGGGTTGATTTAAAGACATTTTGGGATATTTTGAGGGGATTTGGGGACGTTTCGGGACATTTCGAGGAGATTTTTGGGACATTTTGAGGAGTTGTGGGGCCACCTGGAGATGGTTGGGGACACTTAGGGACATCAGGGCCACTTTGGGGATGATTTGGGGAGGTTTGGGGTCGATTTAAGGACATTTTTGGGACATTTTGGGACGTT
TTGAGGGGATTTGGGGACATTTTGAGGGGTTTGGGGACACTTTGGGGACAATTTTGGGAGGTTTGAGACCAATTTTAGGGCATTTTGAGGGGATTTGGGGACATTTTGAGGGGATTTGGAGACACCTGGAGGGGTCTGGGGACAACTTGGGCTGGTTGGGGACACTTTGGGGATGATTTGGGGACGTTTGGGTTTCATTTTGGAGCATTTTGAGAGGATTTGGGGACCA
TTTTGGGAGGTTTGGGGCTGATTTTGAGAGGTTTGGATCCAACTTTGGAGCATTTTGAGGGGATTTGGGGACATTTTGAGGGGTTTGGGGATGATTTTGGGAGGTTTGGGACCAATTTTGGGGCATTTTGAGGGGATTTGGGAACATTTTGAGGGGTTTGGGGATGATTTTGGGAGGTTTAGGGTCGATTTAAGGCCATTTTGGGGCATTTTGAGGGGATTTGGGGCCA
TTTTGAGGGGGTTTGGGGCCACTTTGGGGCCACTTTGGTGGCCTCAGTGCCCCCTTTTGTCCCTGCAGATCCCCACGAGGGCCACGGCCACTCCCACCACGGGCACGGTGAGAGCCCAGGGCACCCTGGGGACATTTTGGGGACGATTTTGGGGGGTTTGGGGCCAATTTTGGGGCATTTTGAGGGGATTTGGGGCATTTTGAGGGGATTTGGGGACGTTTTTAGGGGG
TTTGGGGACACCTGGAGGGGTCTGGGGACAACTTGGGGTGCTTGGGAACACTTTGGGGACGATTTTGGGAGGTTTGGGGCCAATTTTGGGTCATTTTGAGGGGATTTGGGGCATTTTGAGGGGATTTGGGGACGTTTTTAGGGGGTTTGGGGACACCTGGAGGGGTCTGGGGACACCTTGGGGTGGTTGGGAACACTTTGGGGACGATTTGGGGAGGTTTGGGGTTGAT
TTAAGGATATTTTGGGACATTTTGAGGGGTTTGGGGACACTTTGGGGACAATTTTGGGAGGTTTGGGGCCAATTTTGGGTCCATTTTGAGGGGATTTGGGGCATTTTGAGGGGTTTGGAGACACCTGGAGGGGTCTGGGGACACCTTGGGGTGGTTGGGGACACTTTGGGGACGATTTTGGGAGTTGTGGGGCCAATTTTGGGGCATTTTGAGAGGATTTGGGAACGAT
TTTGGGAGGTTTGGATCCAACTTTGGGGCATTTTGAGTGGATTTGGGGCATTTTTAGGGGGTTTGGGGATGATTTTGGGAGGTTTGGGGCCAATTTTGGGGCATTTTGCGGGGATTTGGGGACATTTTGAGGGGTTTGGGGGCCACCTGGATATGGTTGGGGACACTCAGGGACATCAGGGCCACTTTGGGGACGATTTGGGGAGGTTTGGGGTCGATTTAAGGCCATT
TTGGGAGGTTTGGGGACATTTTGAGGGGATTTGGGGCCATTTTGAGGGCTTTGGGGCCACTTTGGGGCCACTTTGGTGGCCTCAGTGCCCCCTTTTGTCCCTGCAGATCCCCACGAGGGCCACGGCCACTCCCACCACGGGCACGGTGAGAGCCCAGGGCACCCTGGGGACATTTTGGGGACGATTTTGGGGGGTTTGGGGCCAATTTTGGGGCATTTTGAGGGGATTT
GGGGCTTTTTGAGGGGATTTGGGGACGTTTTTAGGGGGTTTGGGGACACCTGGAGGAGTCTGGGGACACCTTGGGGTGGTTGGGAACACTTTGGGGACGATTTGGGGAGGTTTGGGGTTGATTGAAGGGCATTTTGGGACATTTTGAGGGGTTTGGGGCCACTTTGGGGACAATTTTGGGAGGTTTGGGGCCAATTTTGGGGCCATTTTGAGGGGATTTGGGGCATTTT
TAGGGGGTTTAGAGACACCTGGAGGGGTCTGGGGACACCTTGGGCTGGGTGGGGACACTTTGGGGACAATTTTGGGAGGTGTGGGGCCAAATTAGGACATTTTGAGGGGATTTGGGGACATTTTGGGGATGATTTGGGGGGATTTGGGGCCAATTTTGGGACATTTTGAGGGGATTTGGGGACATTTTGAGGGATTTGGGGACGATTT
TGGGAGGTTTGGGGTCGATTTAAGGCCATTTTGAAGGGTTTTGGGGCCACCTGGAGATGGTTGAGGACACTTAGGGACATCAGGGCCACTTTGGGGATGATTTGGGGAGGTTTGGGGTCCATTTAAGGCCATTTTGAGAGGTTTTGGGACATTTTGAGGGGATTTGGGGCCATTTTGAGGGGATTTGGGGCCATTTTGAGGGGGTTTGGGGCCACTTTGGGGCCACTTT
GGTGGCCTCAGTGCCCCCTTTTGTCCCTGCAGGTCCCCACGAGGGCCACGGCCACTCCCACCACGGGCACGGTGAGAGCCCAGGGCACCCTGGGGACGTTTTGGGGACGATTTTGGGGGGTTTGGGGCCAATTTTGGGGCATTTTGAGGGGATTTGGGGACGTTTTTAGGGGGTTTGGGGACACCTGGAGGGGTCTGGGGACACCTTGGGGTGGTGGGGACACTTTGG
GGACAATTTTGGGAGGTTTGGGGCCAAATTAGGACATTTTGAGGGGATTTGGGGACATTTTGAGGGGGTTTGGGGACATTTTGGGGACGATTTGGGGGGATTTGGGGACGATTTGGGGGGATTTGGGGCCAATTTGGGGCATTTTGAGGGGATTTGGGGACATTTTGAGGGGTTTGGGGACGATTTTGGGAGGTTTGGGGTCGATTTAAGGCCATTTTGAAGGGTTTTG
GGGCCACCTGGAGATGGTTGGGGACACTCAGGGACATCAGGGCCACTTTGGGGACGATTTGGGGAGGTTTGGGGTCGATTTAAGGCCATTTTGTGACATTTTGGGACGTTTTGAGGGGATTTGGGGCATTTTGAGGGGATTTGGGGCCATTTTGAGGGCTTTGGGGCCACTTTGGGGCCACTTTGGTGGCCTCAGTGCCCCCTTTTGTCCCTGCAGGTCCCCACGAGGG
CCACGGCCACTCCCACCACGGGCACGGTGAGAGCCCAGGGCACCCTGGGGACATTTTGGGGACGATTTGGGGGGGTTTGGGGCCAATTTTGGGTCATTTTGAGGGGATTTGGGGCATTTTGAGGGGATTTGGGGACGTTTTTAGGGGGTTTGGGGACACCTGGAGGAGTCTGGGGACACCTTGGGGTGGTTGGGGACCCTTTGGGGATGATTTTGGGAGGTTTGGGGTT
GATTGAAGGACATTTTGGGACATTTTGAGGGGTTTGGGGACACTTTGGGGACGATTTTGGGGGGTTTGGGGCCAATTTTGGGGCATTTTGAGGGGATTTGGGGACATTTTGAGGGGTTTGGGGACAATTTTGGGAGGTTTGGGGTCGATTTTGGGACATTTTGAAGGGTTTTGGGGCCACCTGGAGATGGTTGGGGACACTCAGGGACATCAGGGCCACTTTGGGGACG
ATTTGGGGAGGTTTGGGGTCGATTTAAGGCCATTTTGGGAGGTTTTAGGACATTTTGAGGGGATTTGGGGACATTTTGAGGGGATTTGGGGCCATTTTGAGGGGCTTTGGGGCCACTTTGGGGCCACTTTGGTGGCCTCAGTGCCCCCTTTTGTCCCTGCAGATCCCCACGAGGGCCACGGCCACTCCCACCACGGGCACGGTGAGAGCCCAGGGCACCCTGGGGACAT
TTTGGGGACGATTTTGGGGGGTTTGGGGCCAATTTTGGGGCATTTTGAGGGGATTTGGGGACACCTGGAGGGGTCTGGGGACACCTTGGGATGGGTGGGGACACTTTGGGGATGATTTTGGGAGGTTTGGGGCCAAATTAGGGCATTTTGAGGGGACTTGGGGACATTTTGAGGGGGTTTGGGGACATTTTGGGGACGATTTGGGGGGATTTGGGACCAATTTTGGGGC
ATTTTGAGGGGATTTGGGGACATTTTGAGGGGATTTGGGGACAATTTTGGGAGGTTTGGGGTCGATTTTGGGACATTTTGAAGGGTTTTGGGGCCACCTGGAGATGGTTGGGGACACTCAGGGACATCAGGGCCACTTTGGGGATGATTTGGGGAGGTTTGGGGTCGATTTAAGGACATTTTGGGACATTTTGAGGGGTTTTGGGGACATTTTGAGGGGATTTTGGGGC
ATTTTGAGGGGTTTGGGGACACTTTGGGGACAATTTTGGGAGGTTTGAGACCAATTTTGGGTCATTTTGAGGGGATTTGGGGACATTTTGAGAGGATTTGGAGACACCTGGAGGGGTCTGGGGACAACTTGGGGTGGTTTGGGACATTTTGGGGATGATTTGGGGACGTTTGGGTTTCATTTTGGAGCATTTTGAGAGGATTTGGGGACCATTTGGGGAGGTTTGGGGC
TGATTTTGGGAGGTTTGGATCCAACTTTGGGGCATTTTGAGGGGATTTGGGGCATTTTTAGGGGGTTTGGGGCTGATTTCGGGAGGTTTGGATCCAACTTTGGGGCATTTTGAGGGGATTTGGGAACATTTTGAGGGGTTTGGGGATGATTTTGGGAGGTTTGGGGTCGATTTAAGGCCATTTTGGGGCATTTTGAGGGGATTTGGGGACGTTTTGAGGGCTTTGGGGC
```

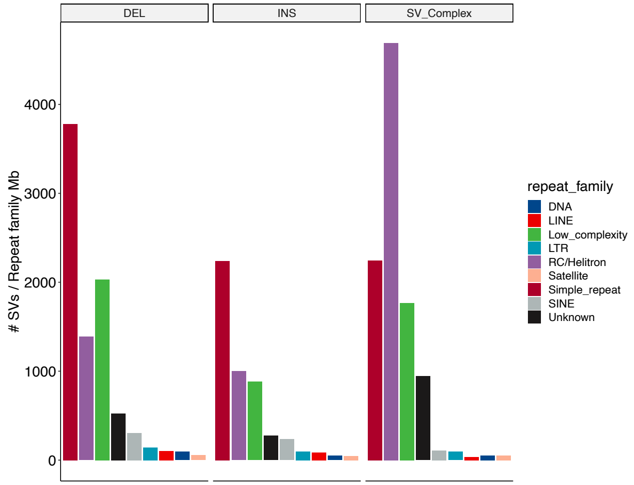
MHC graph depth shows single-copy MHC regions surrounded by complex VNTRs

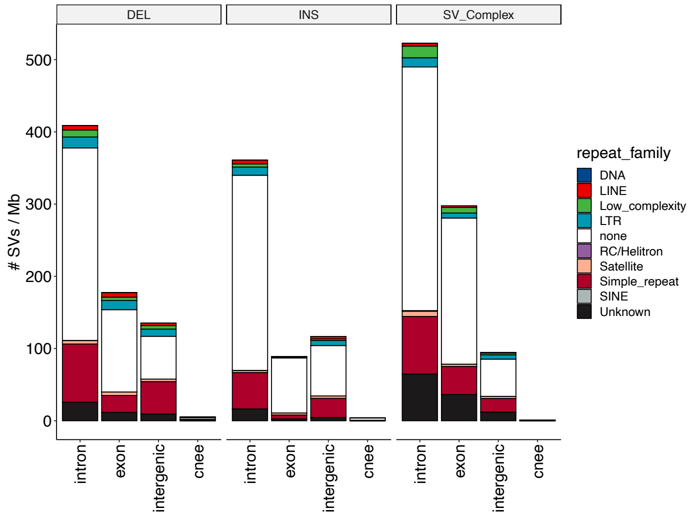


SVs from PGGB: Danielle’s plots

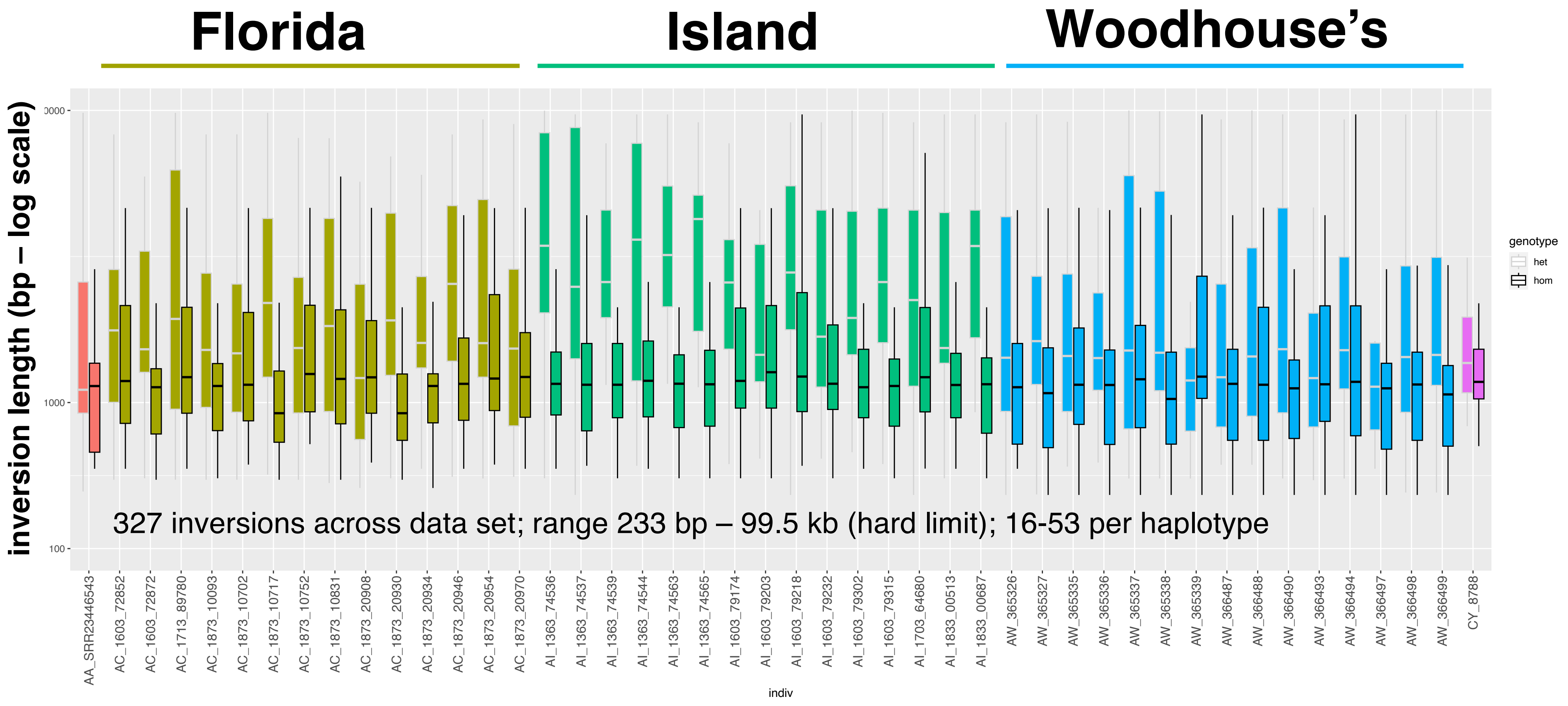








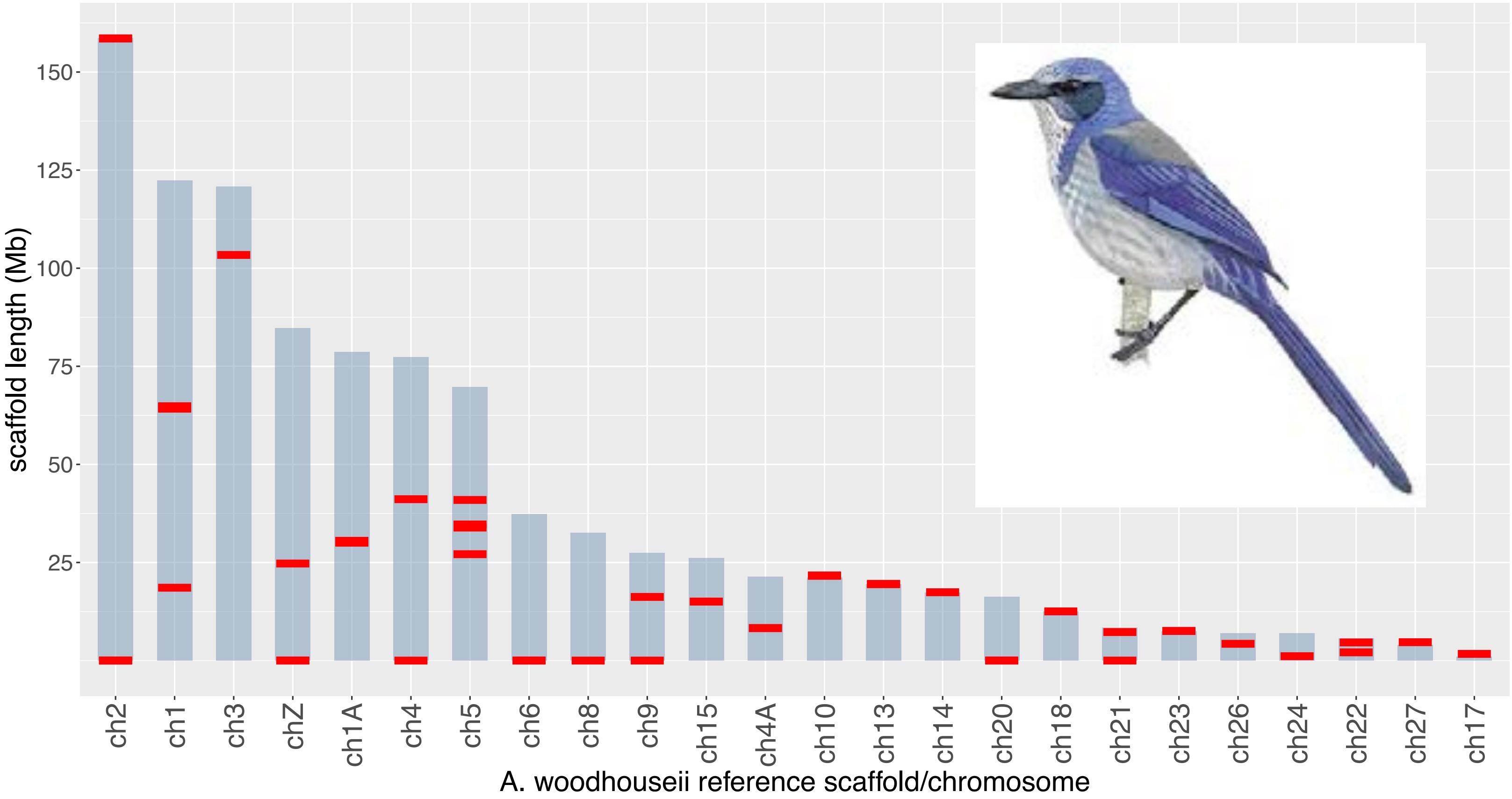
inversion lengths longer in heterozygotes



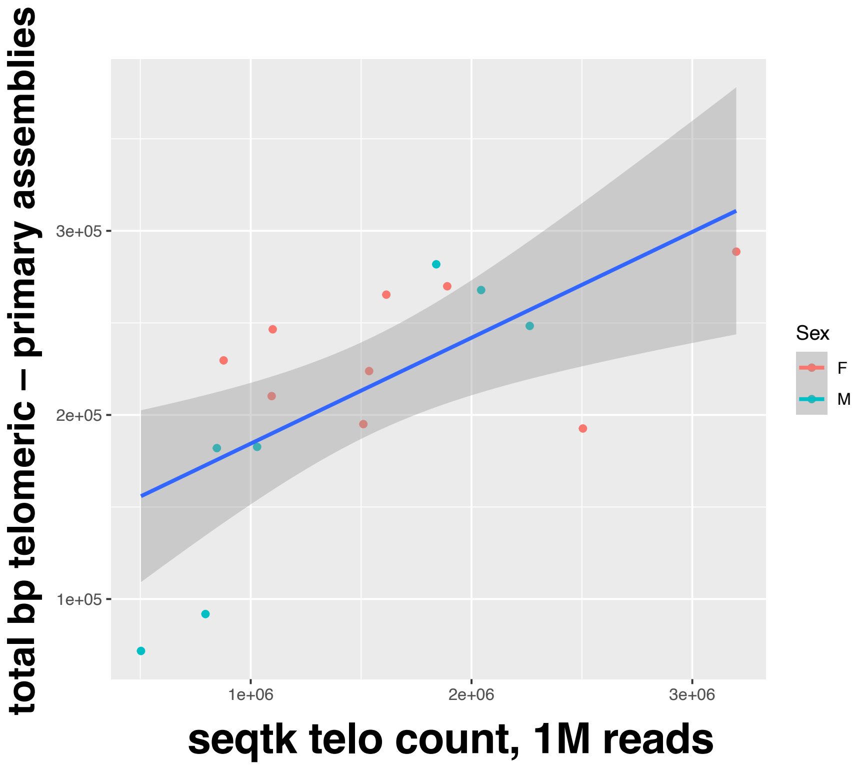
327 inversions across data set; range 233 bp – 99.5 kb (hard limit); 16-53 per haplotype

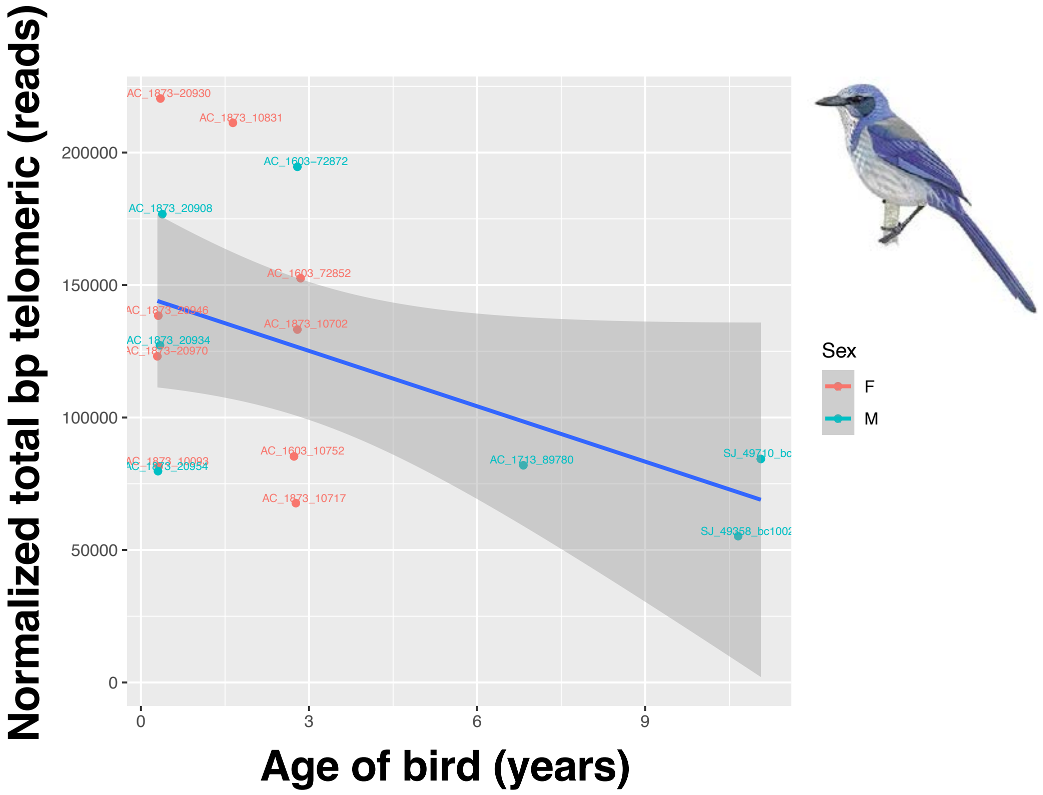
SVs counted in diploid mode with svim-asm: Heller et al. 2021. *Bioinformatics* **36:** 5519-5521.

Telomere sequences are generally found at chromosome ends



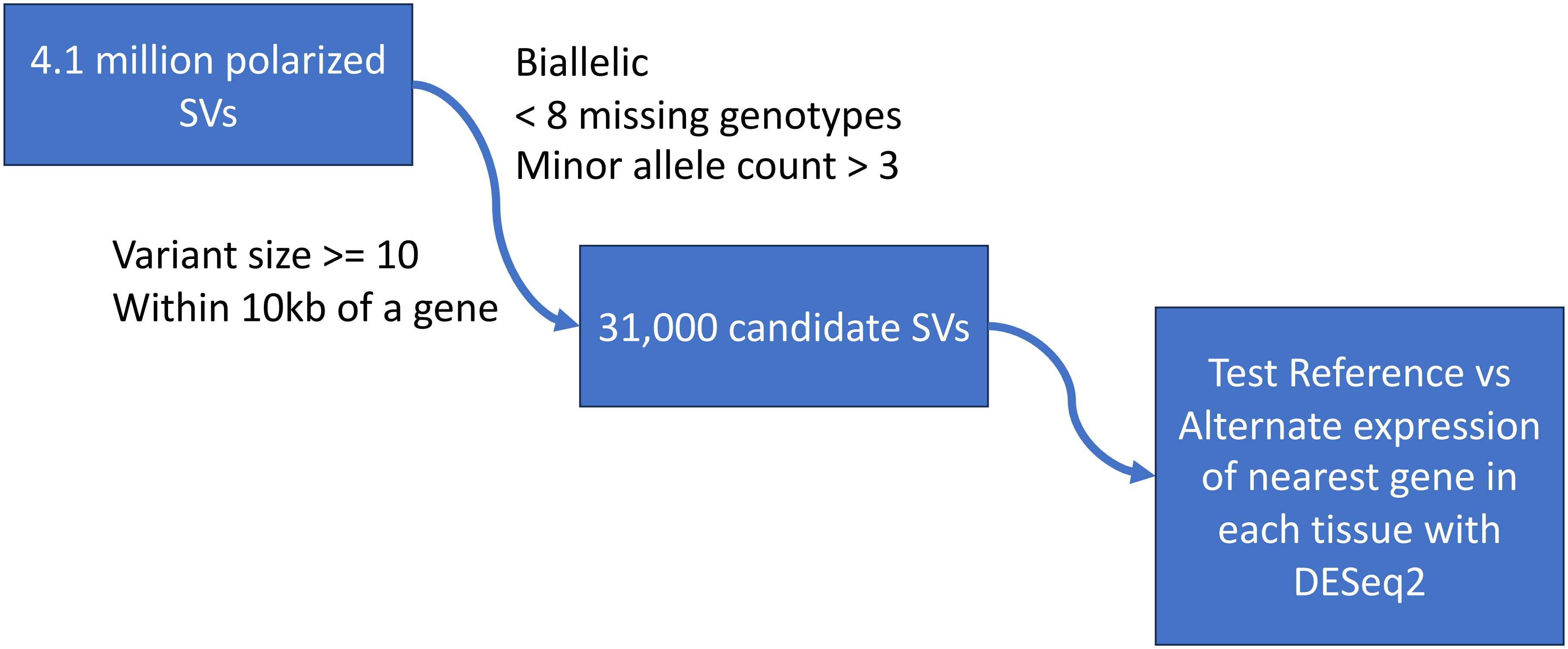
Telomere abundance declines with age in Florida birds





using seqtk telo, H. Li unpubl.

Linking structural variants to gene expression



4.1 million polarized SVs

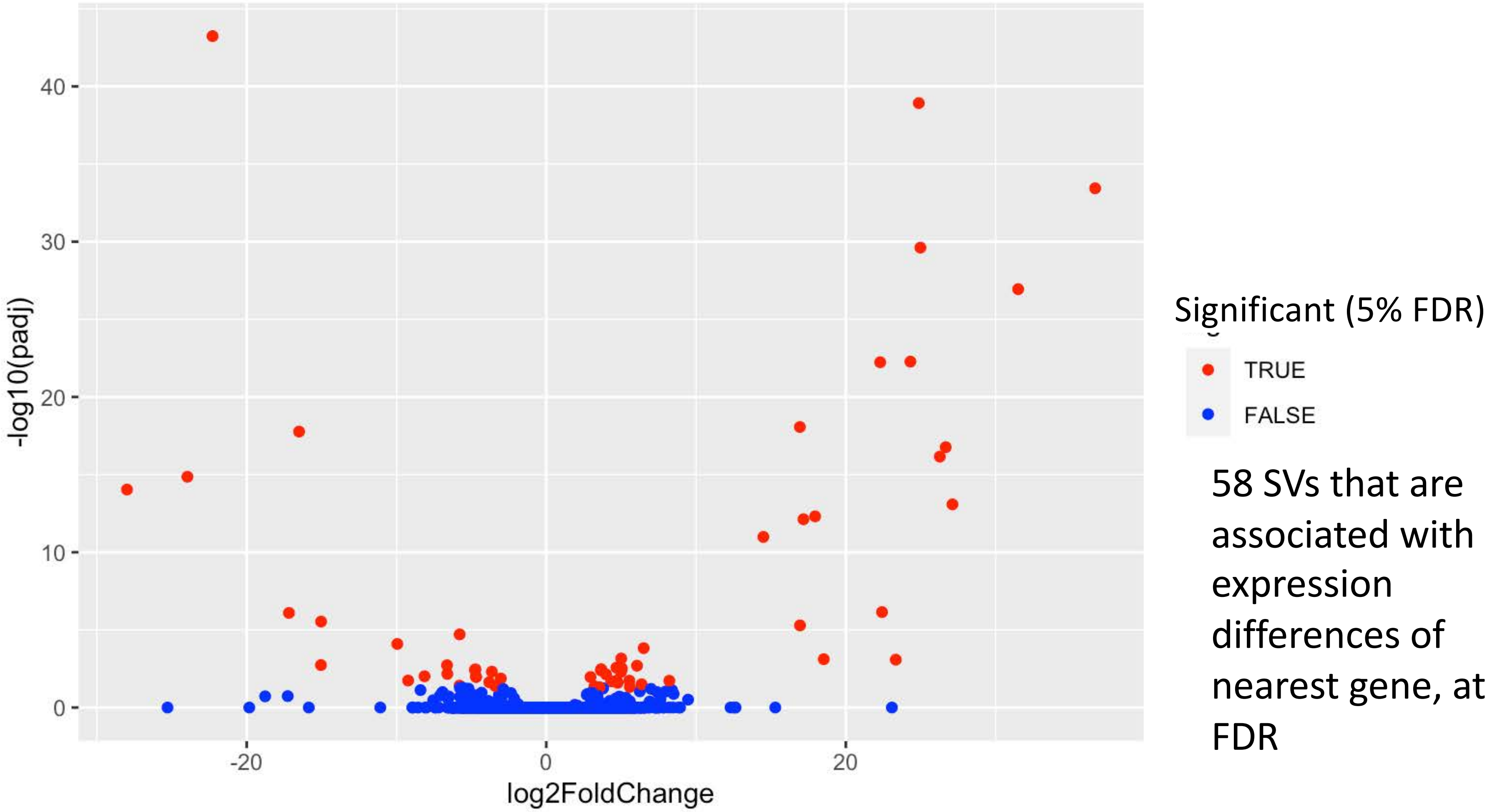
Biallelic
< 8 missing genotypes
Minor allele count > 3

Variant size >= 10
Within 10kb of a gene

31,000 candidate SVs

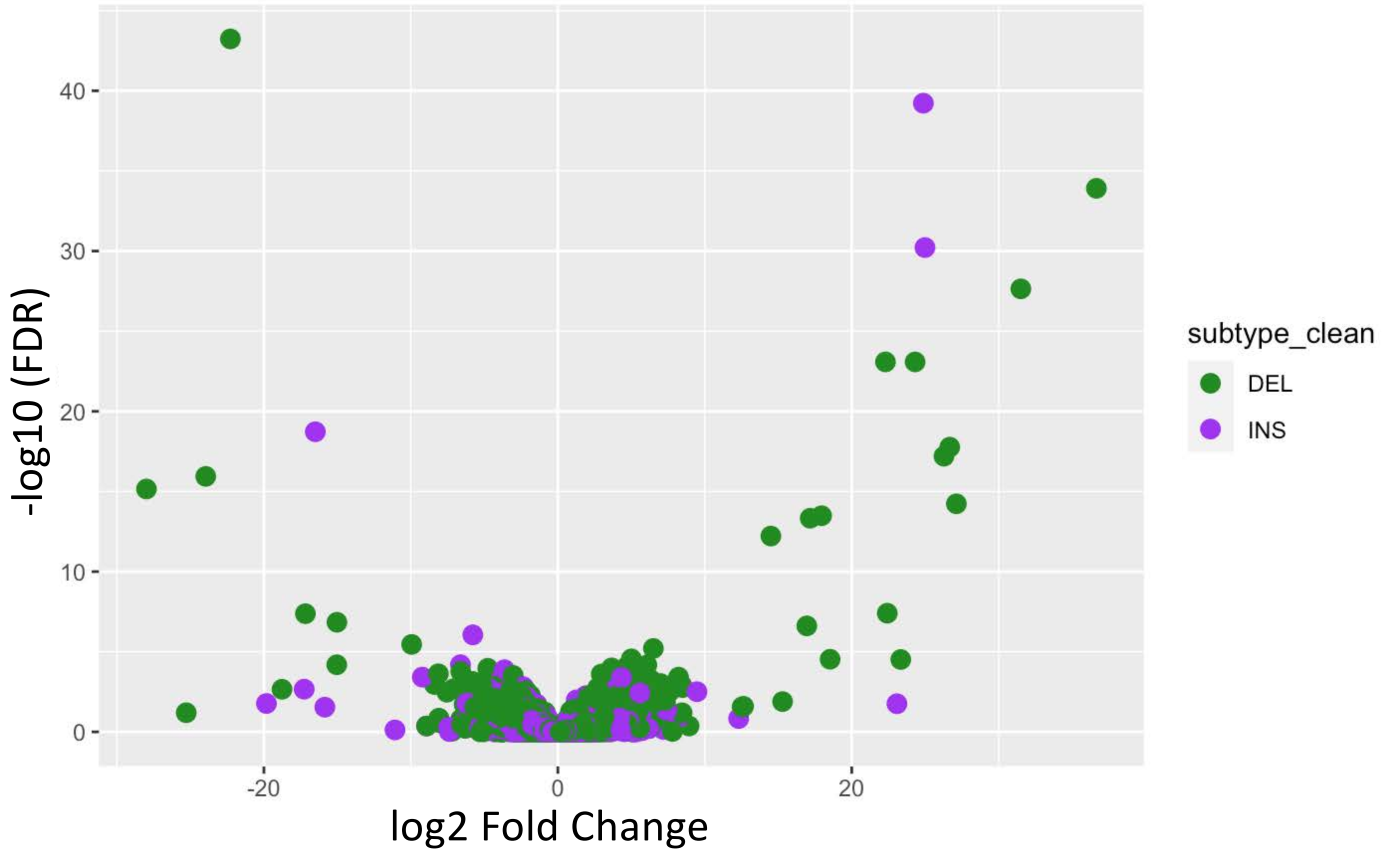
Test Reference vs Alternate expression of nearest gene in each tissue with DESeq2

Structural variants affect gene expression



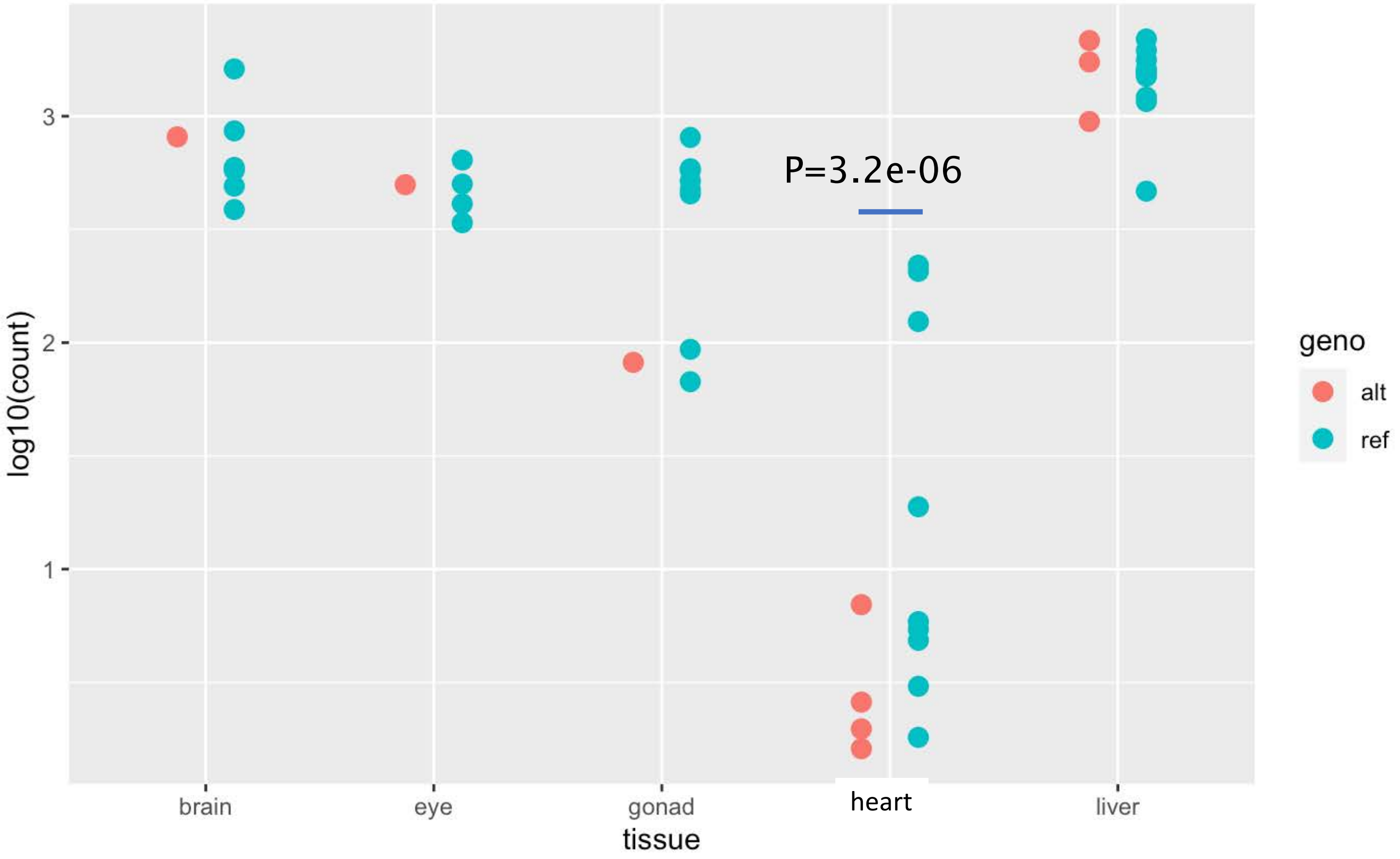
58 SVs that are associated with expression differences of nearest gene, at 5% FDR

Deletions have a bigger impact on gene expression than insertions



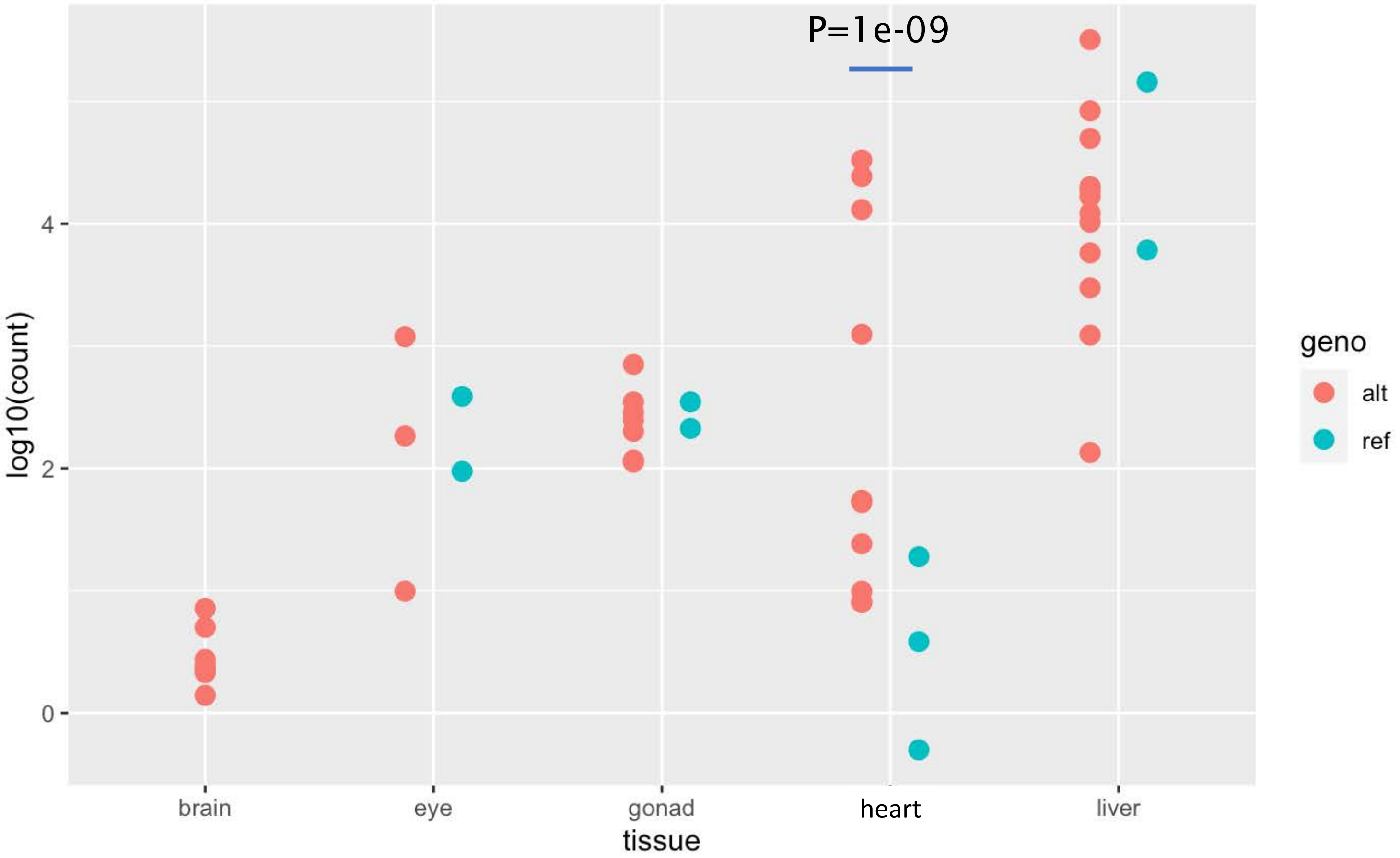
SVs influence tissue-specific gene expression

16 bp deletion in CNEE near CDH4 gene



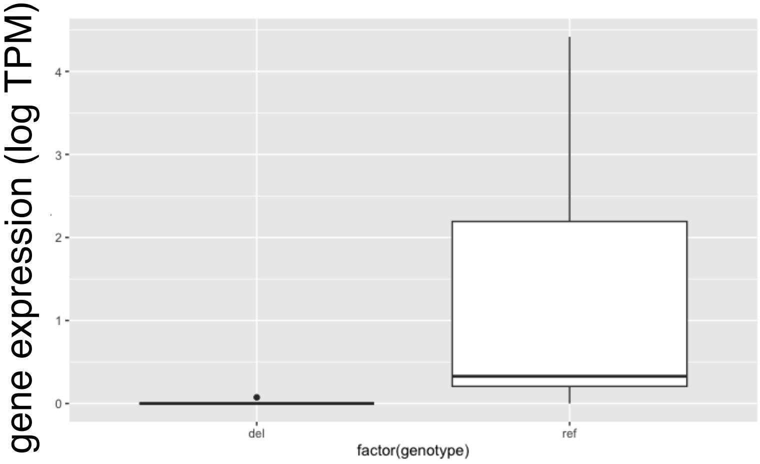
SVs influence tissue-specific gene expression

43 bp deletion near THRSPB

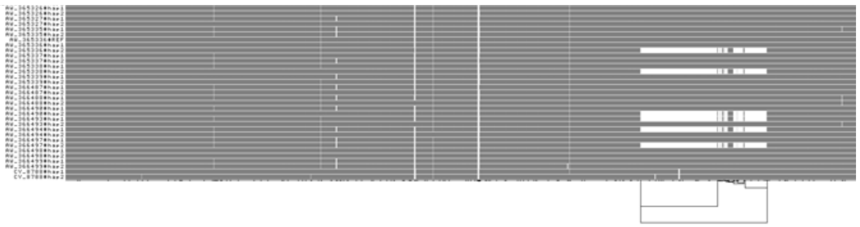


Visualizing examples of SVs influencing gene expression

gene-LOC116806707 – some AW birds with 0 and some with > 100 tpm

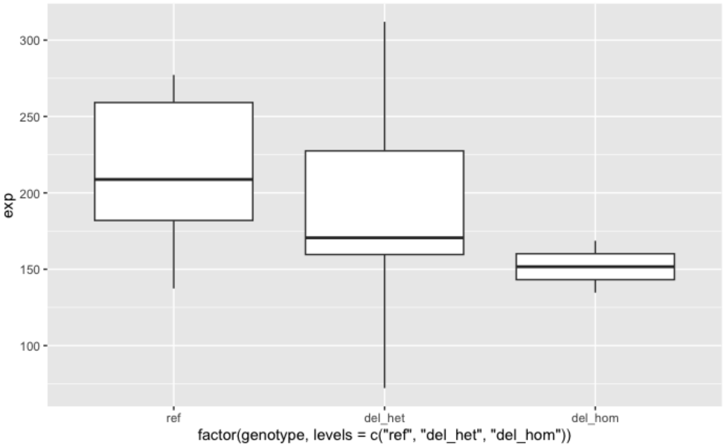


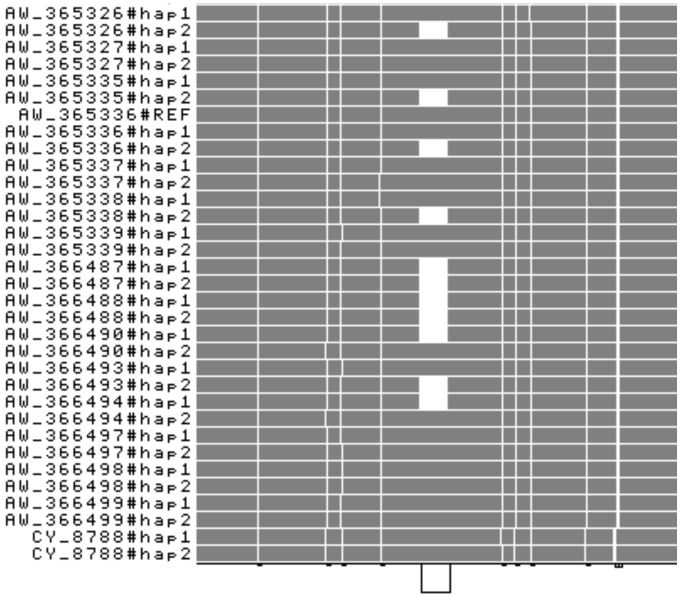
odgi viz 1D visualization of region around gene-LOC116806707



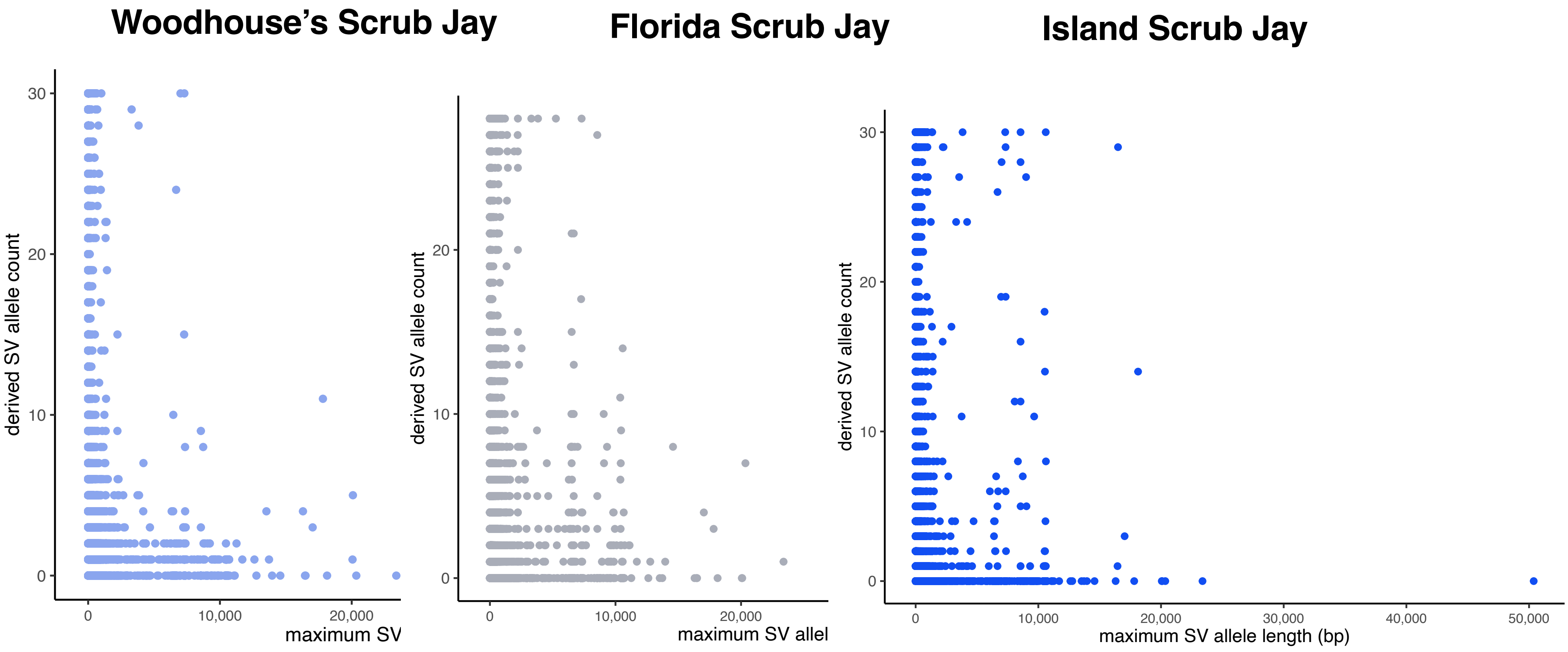
Another example of SV and gene expression

17bp deletion in a CNEE near the GTPBP2 gene





Long SVs increase in frequency in small populations



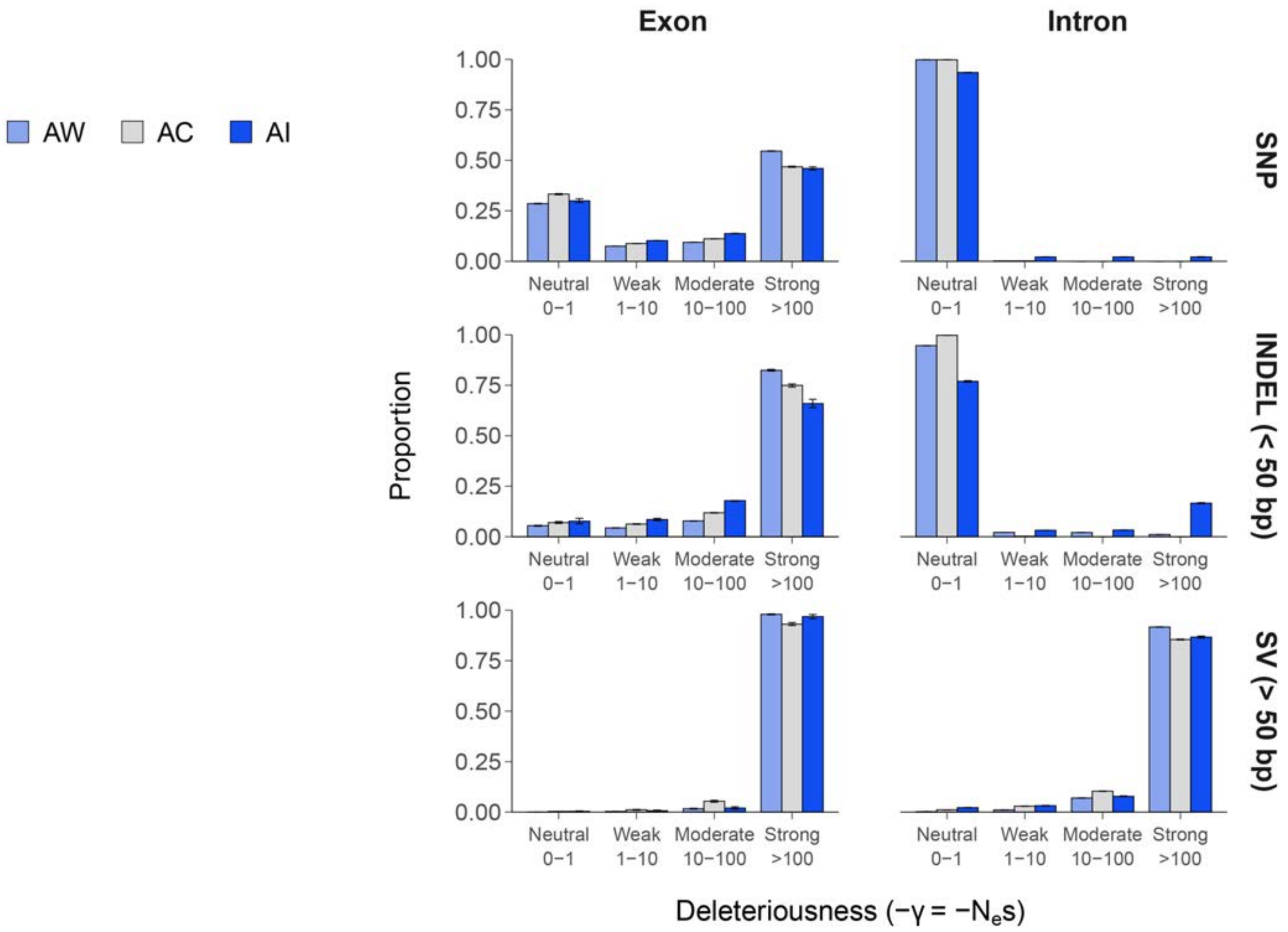
Woodhouse’s Scrub Jay

Florida Scrub Jay

Island Scrub Jay

4,032,715 SVs from PGGB vcf downsampled to 100k SVs

Distribution of fitness effects of SVs



Conclusions

- Scrub-jay genomes are repeat-rich
- The MHC class II region is much more complex than chicken and likely dispersed on multiple contigs and chromosomes
- Pangenome graph analysis illustrates dynamic and conserved regions of the scrub-jay genome
- Large structural variants appear in lower frequency than small ones
- Pangenome analysis will likely become the common standard

Acknowledgements

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
Danielle Khost
Heng Li
Bohao Fang
George Kolyfetis

Pangenome informatics
Erik Garrison
Andrea Guarracino

Fieldwork
Greg and Donna