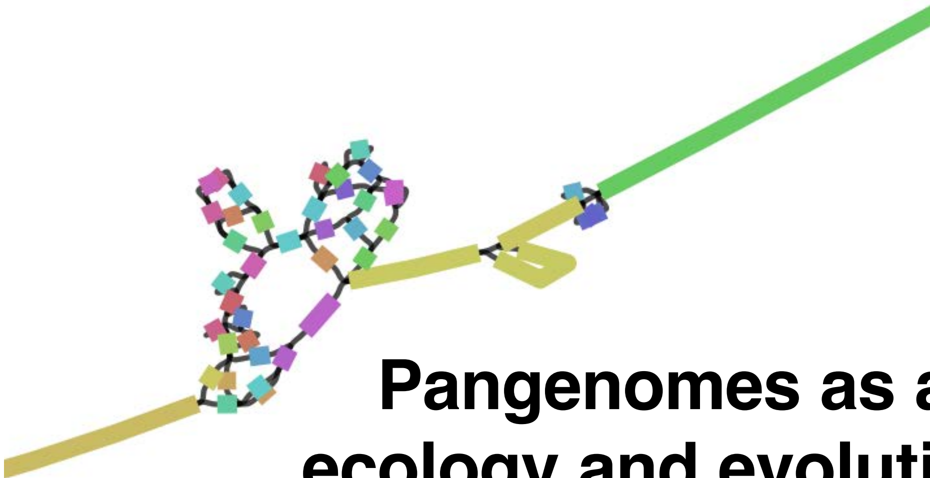
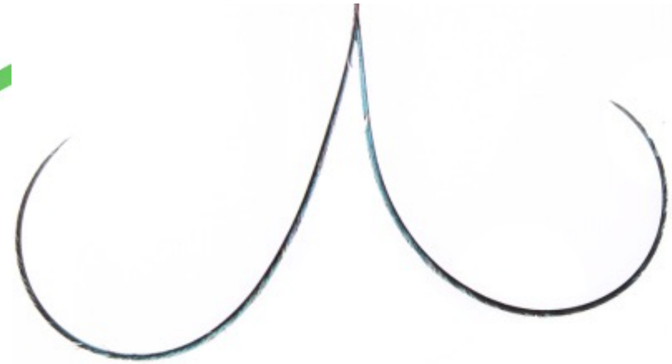
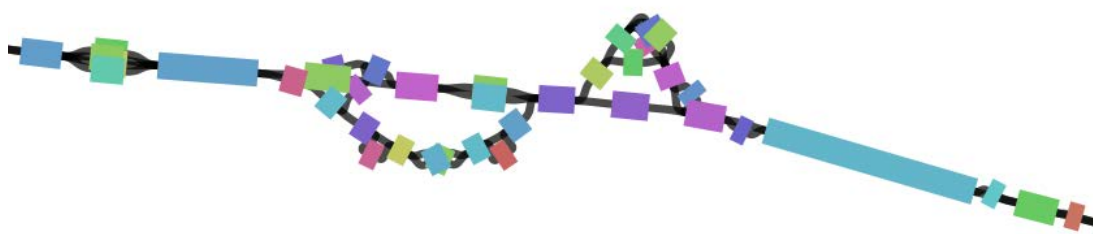



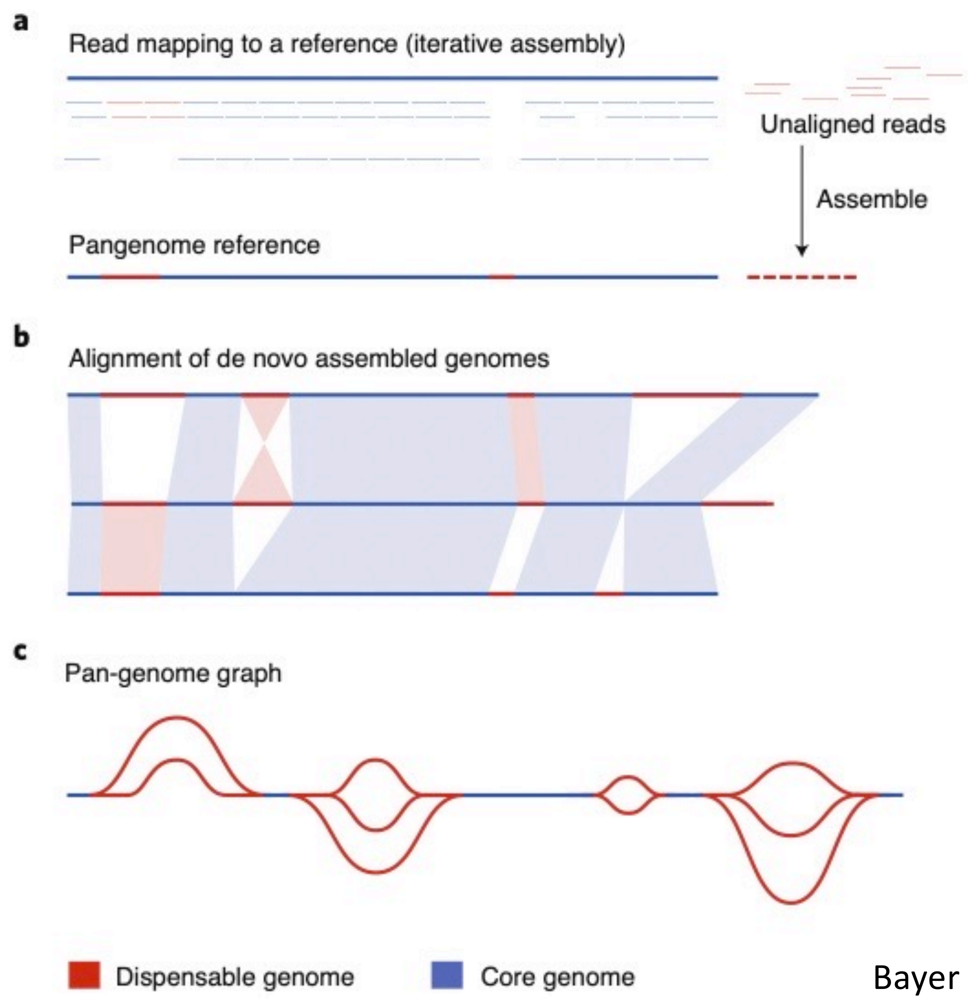

Pangenomes as a new tool for studying ecology and evolution of natural populations

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



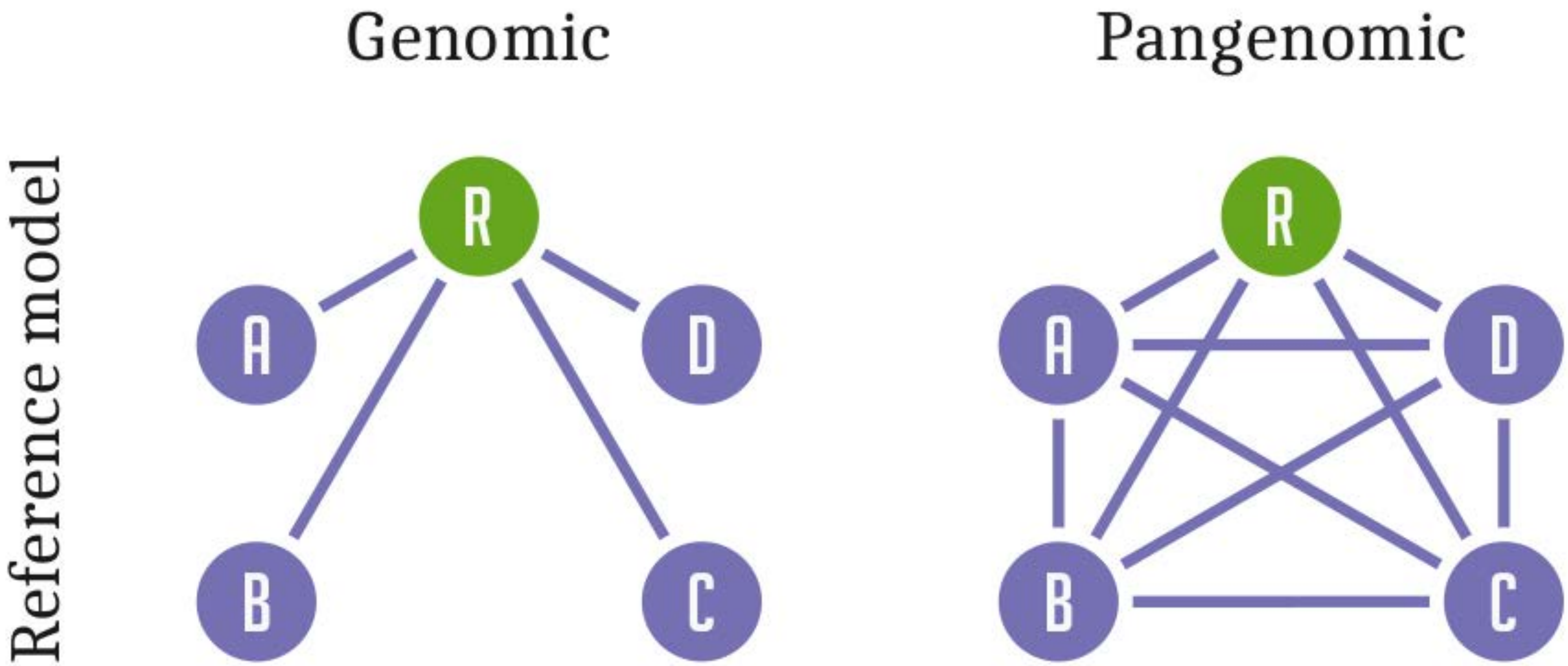


Pangenomes: moving beyond reference-based genomics



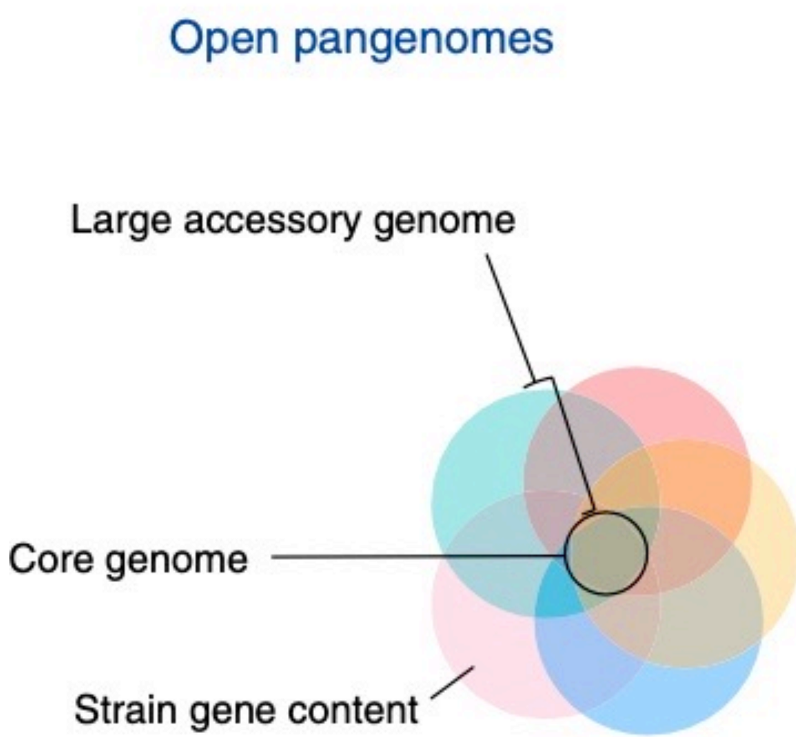
Bayer et al. 2020. *Nature Plants* 6: 914-920.

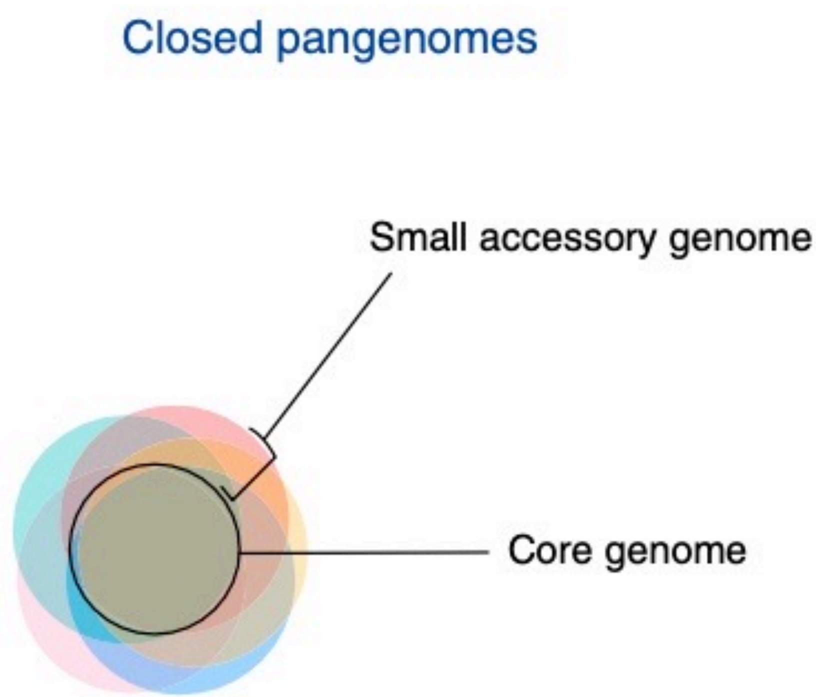
Reference-free genomics



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

Open and closed pangenomes

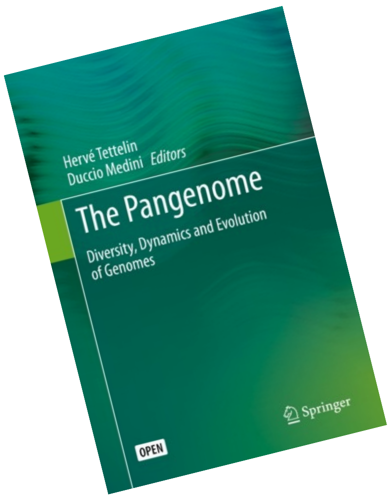


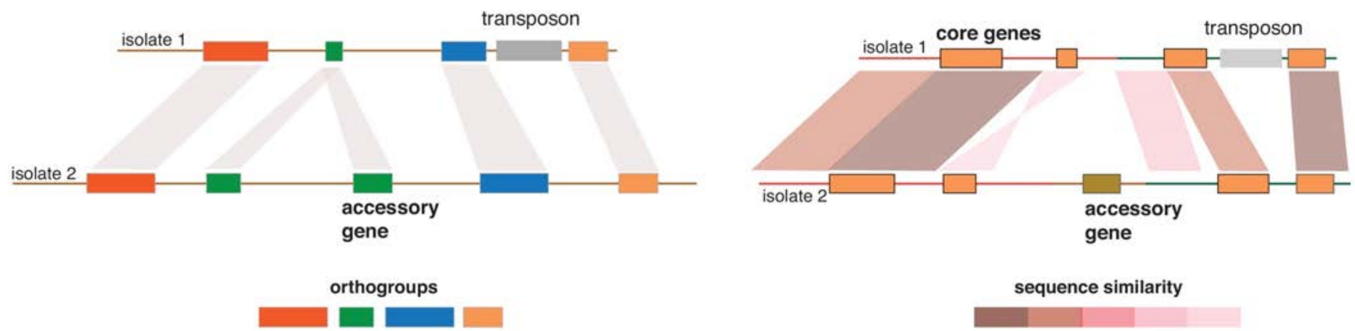


Brokhurst et al. 2019. *Curr. Biol.*

The eukaryotic pangenome

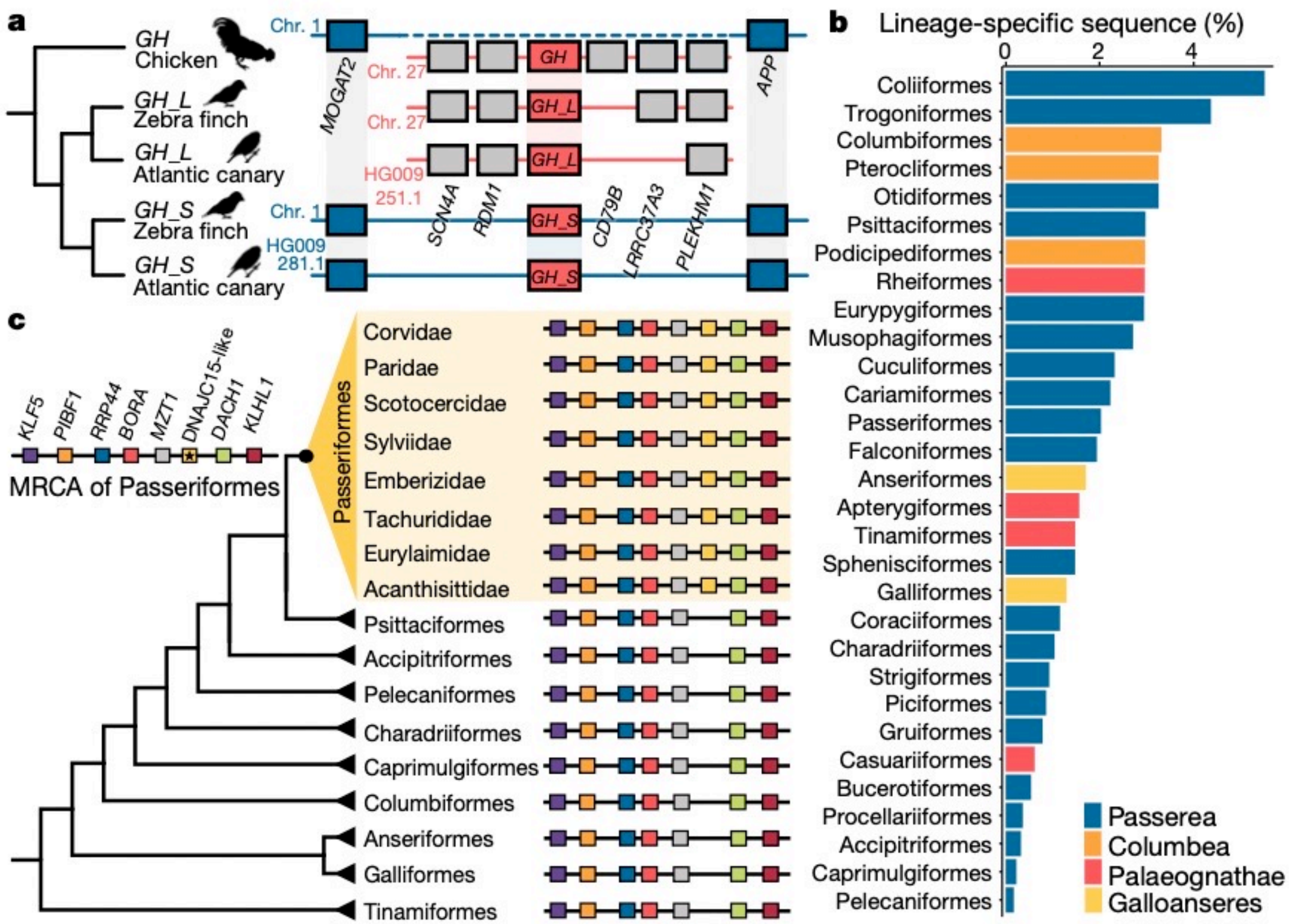
- “The existence of pangenomes in eukaryotes is debated…Pangenome studies in eukaryotes are challenging due to their more complex genome and architectures and a lack of replete genome-level sampling” (Brockhurst et al. 2019. *Current Biology)*





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

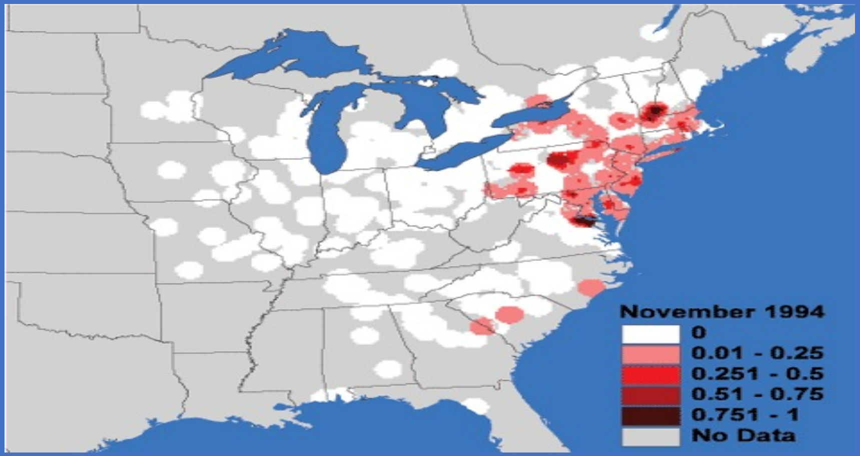
Pangenome approach to comparative genomics



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

Recent history of House Finch populations

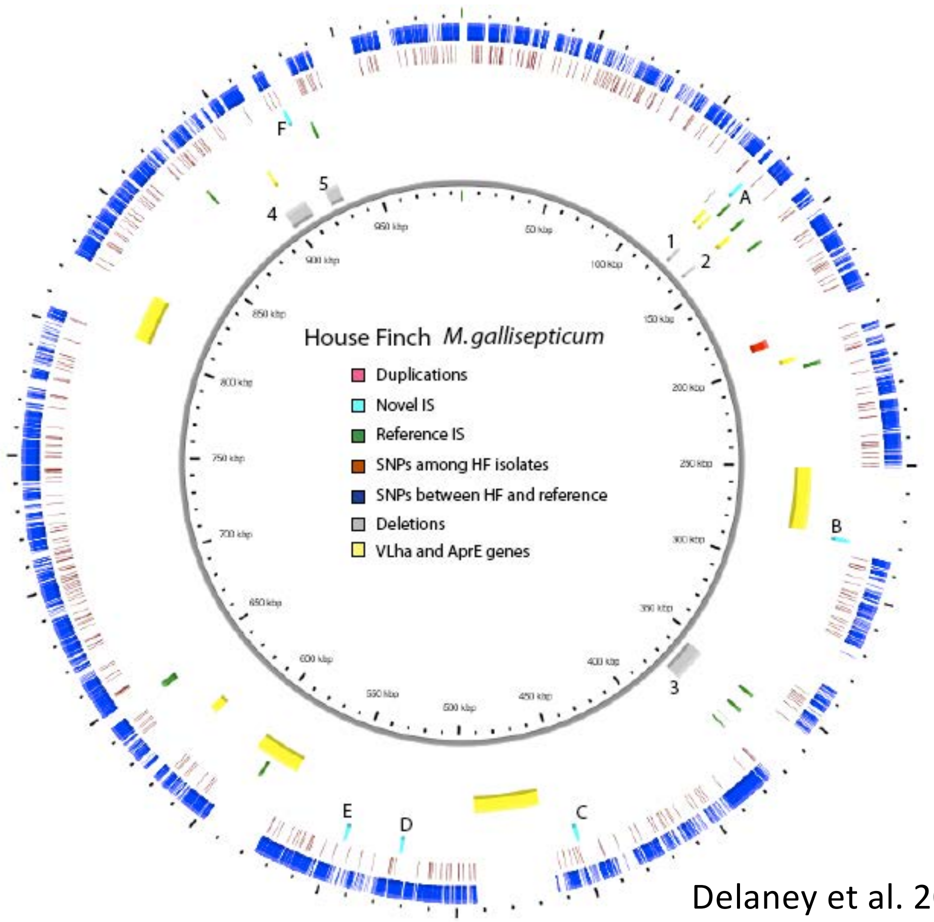
Rapid spread of *Mycoplasma* in House Finch populations



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



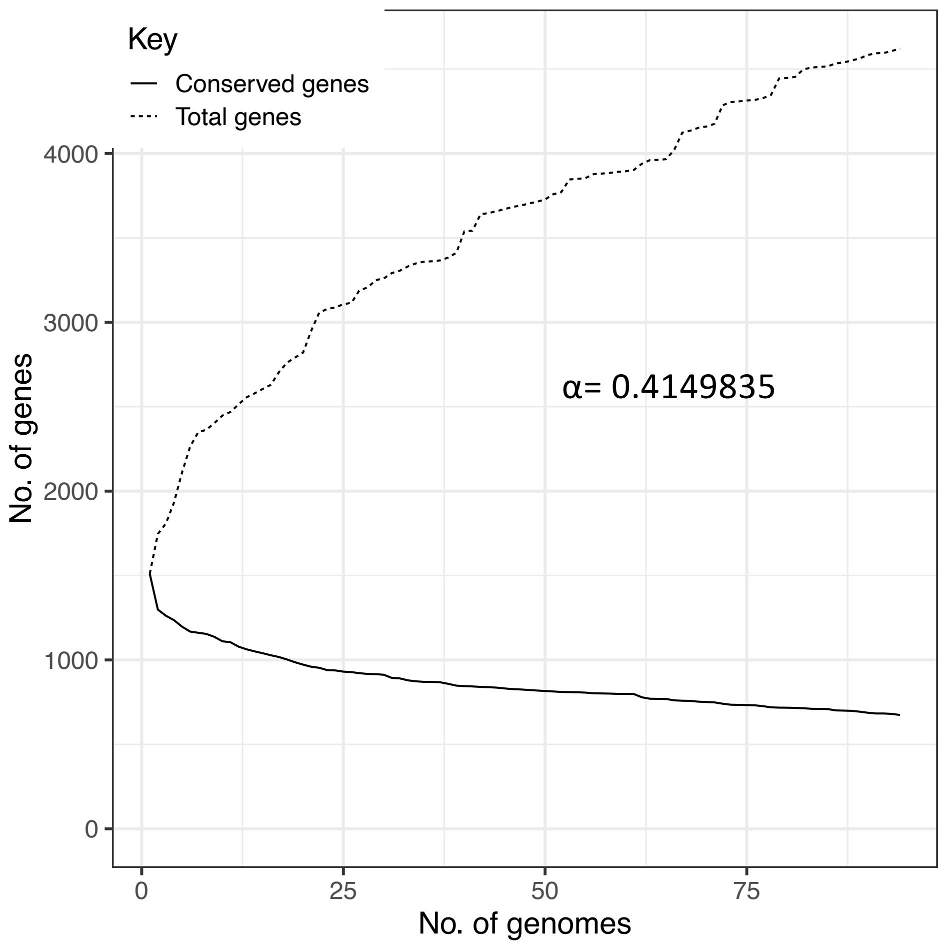
Analyzed 81 Mycoplasma strains from chicken, turkey and house finch, available on NCBI

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

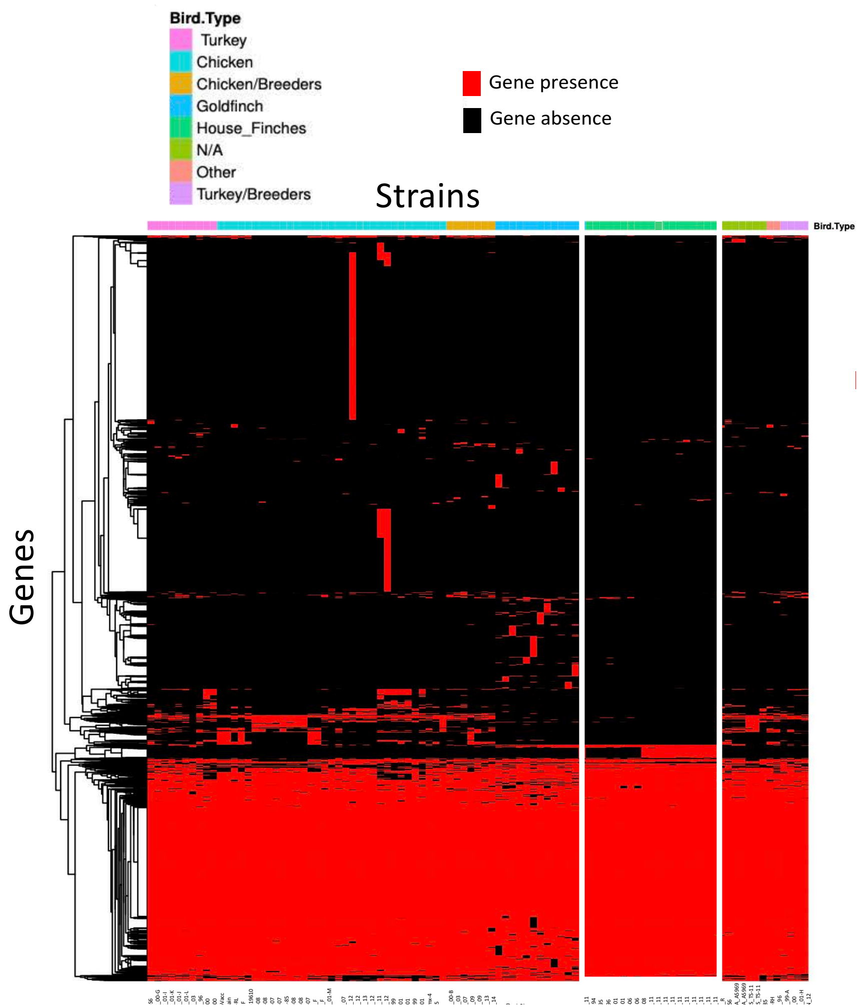
Feature	Info	Number of genes	Percentage
Core genes	(99% <= strains <= 100%)	674	14.586
Soft core genes	(95% <= strains < 99%)	464	10.041
Shell genes	(15% <= strains < 95%)	412	8.916
Cloud genes	(0% <= strains < 15%)	3071	66.457
SGF	one copy in all strains	141	3.051
SGF	without recombination signals	117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

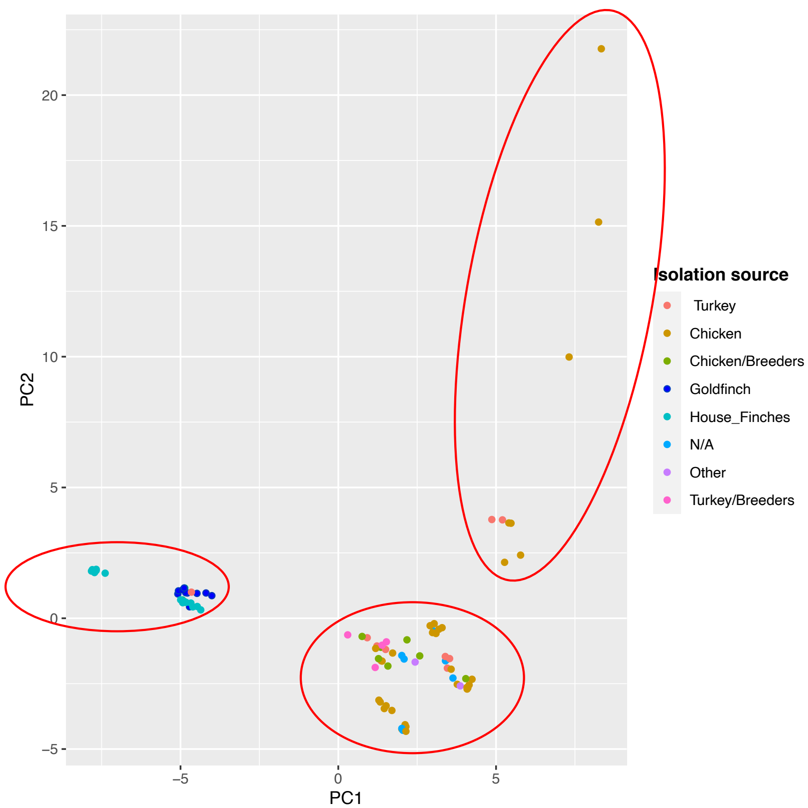
Alpha value: the number of gene clusters we would see if we collected *all* genomes of the species

New data: Determine the alpha value using MicroPan

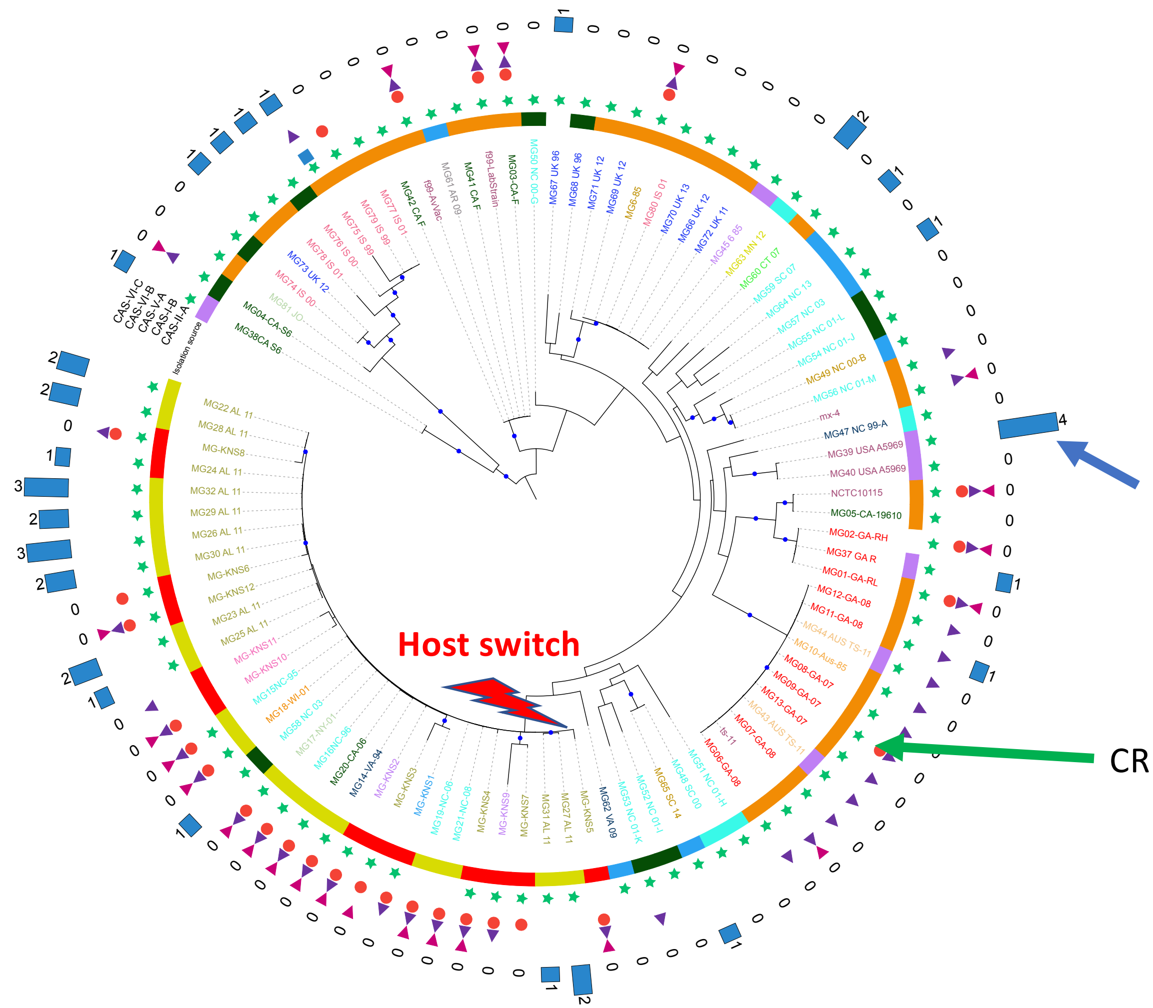
*the pan-genome is closed if the estimated alpha is above 1.0

Mycoplasma pangenome gene repertoire is highly strain-specific





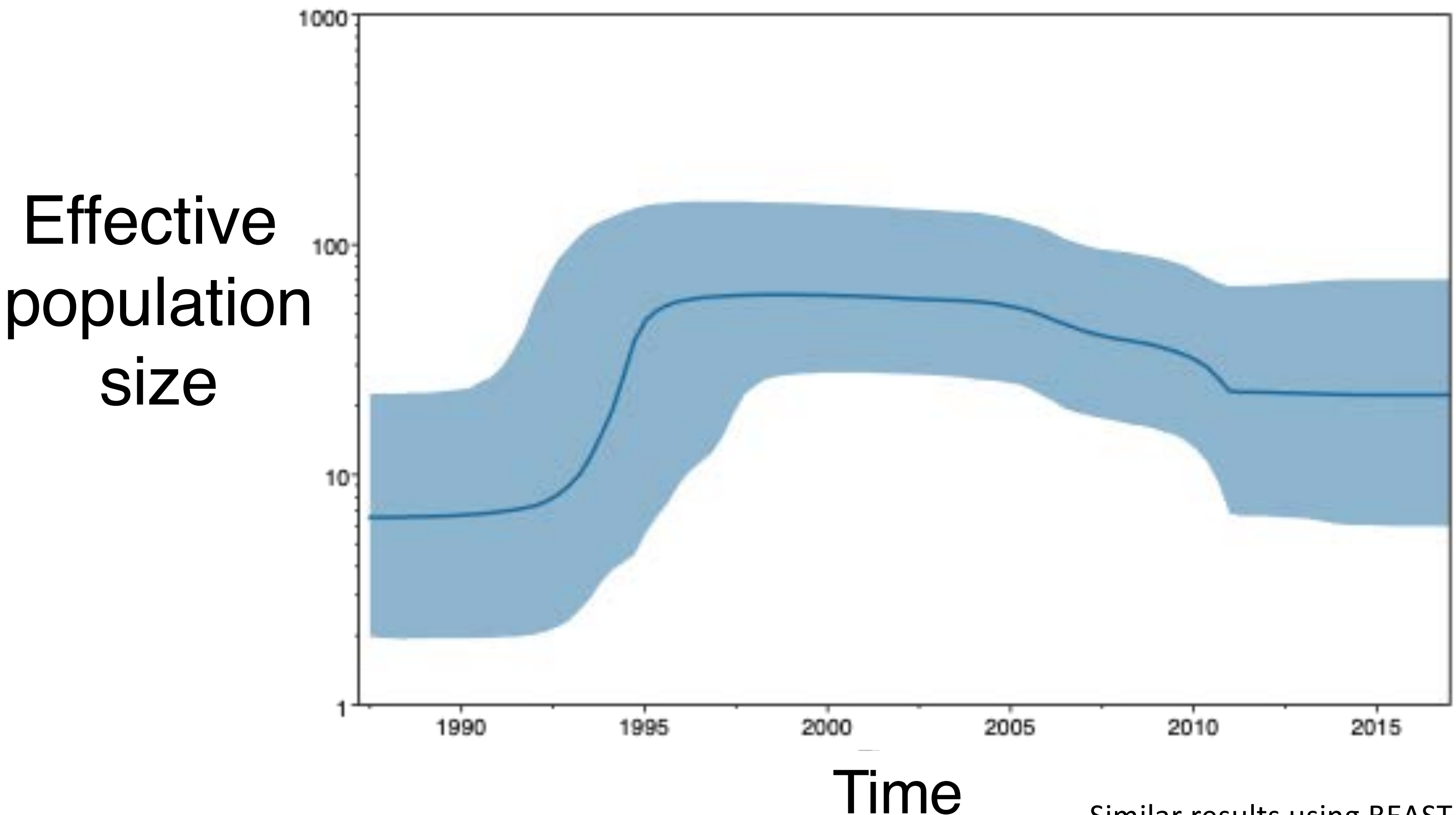
House Finch *Mycoplasma* strains have distinct CRISPR and prophage landscapes



Prophages integrated into Mycoplasma genome

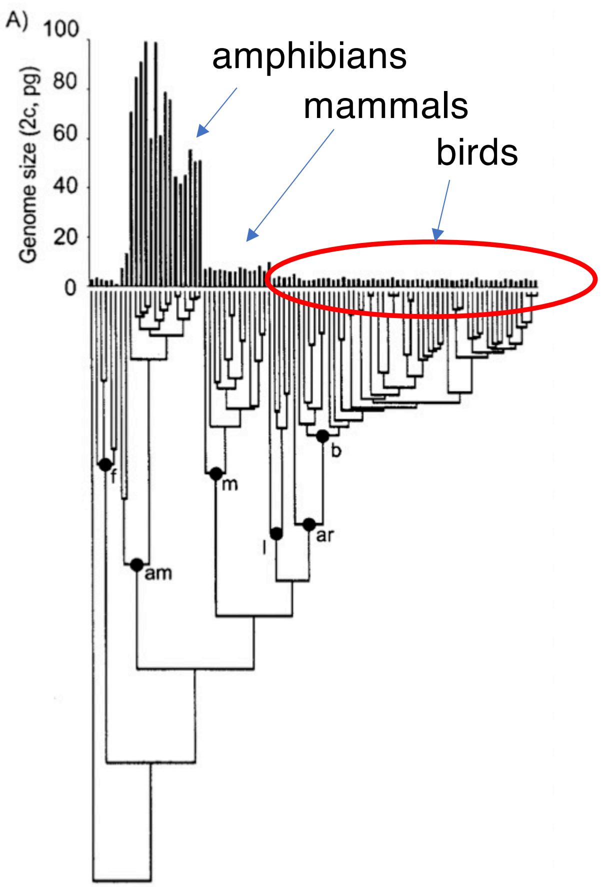
CRISPR landscape

Mycoplasma epizootic likely began ~2 years before first detection

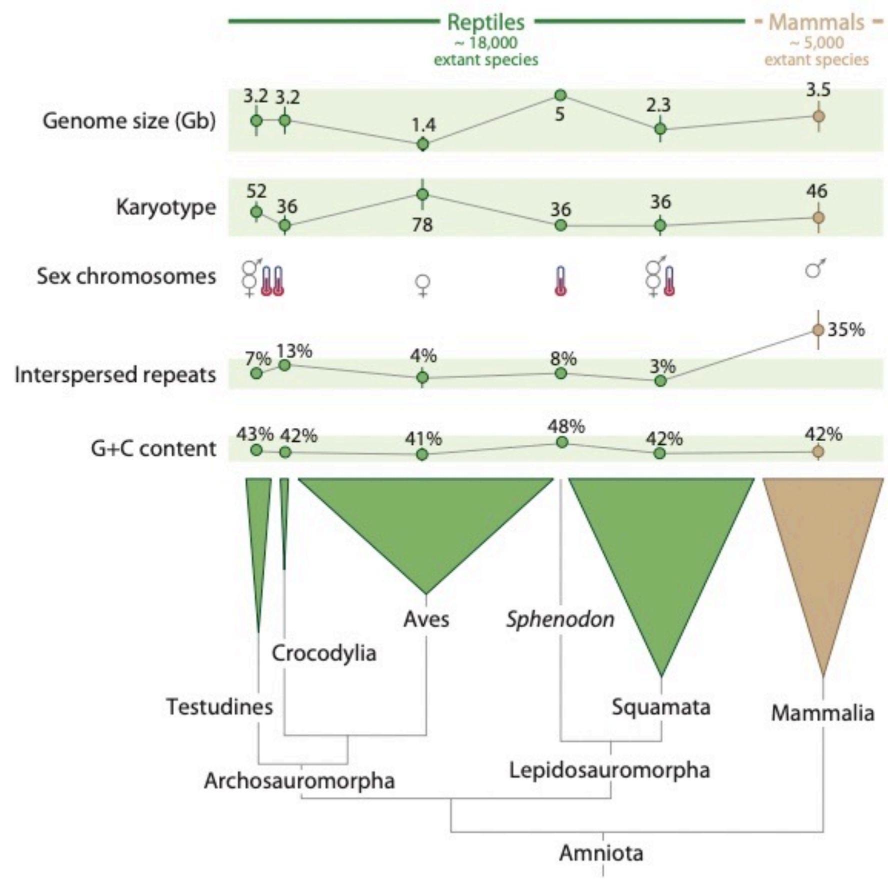


Similar results using BEAST and Stairway plot

Birds have small, streamlined genomes

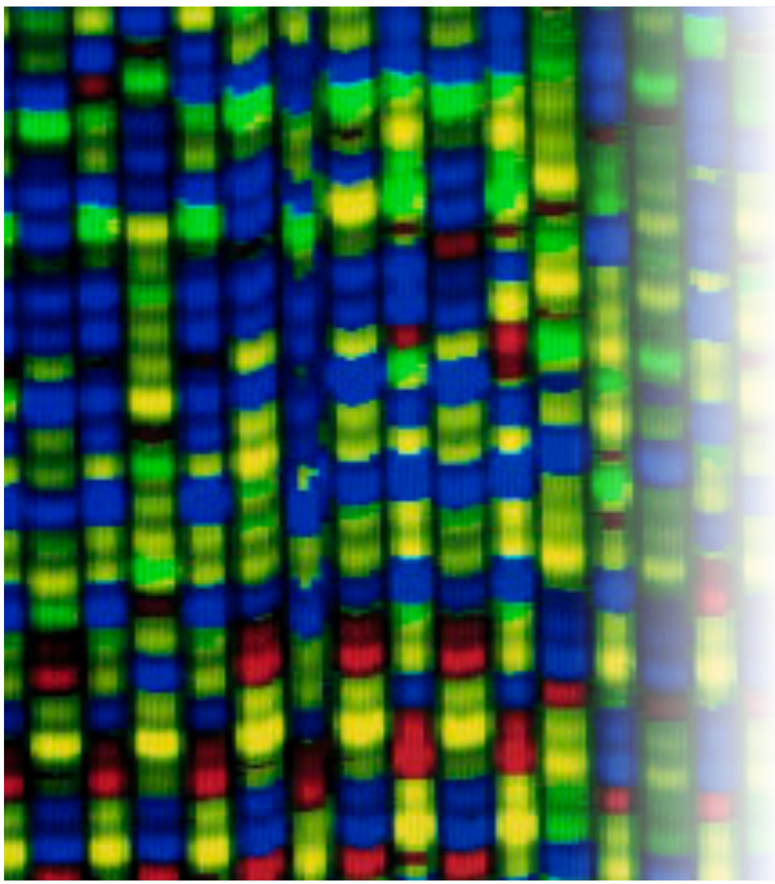


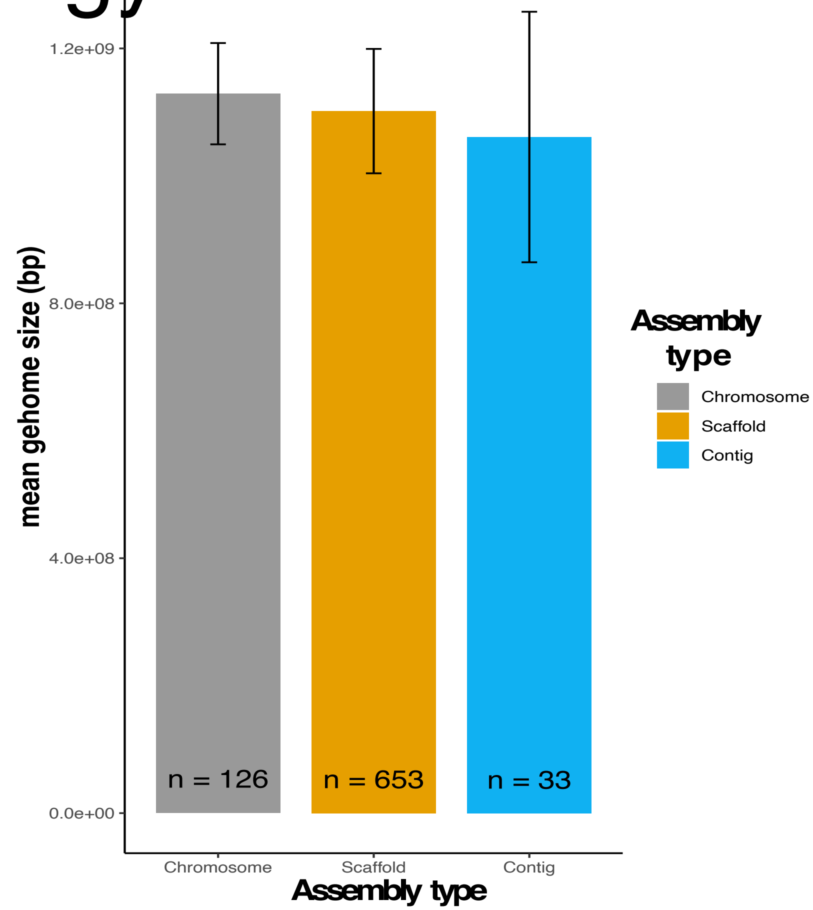
Waltari & Edwards. 2002. *Am. Nat.*



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

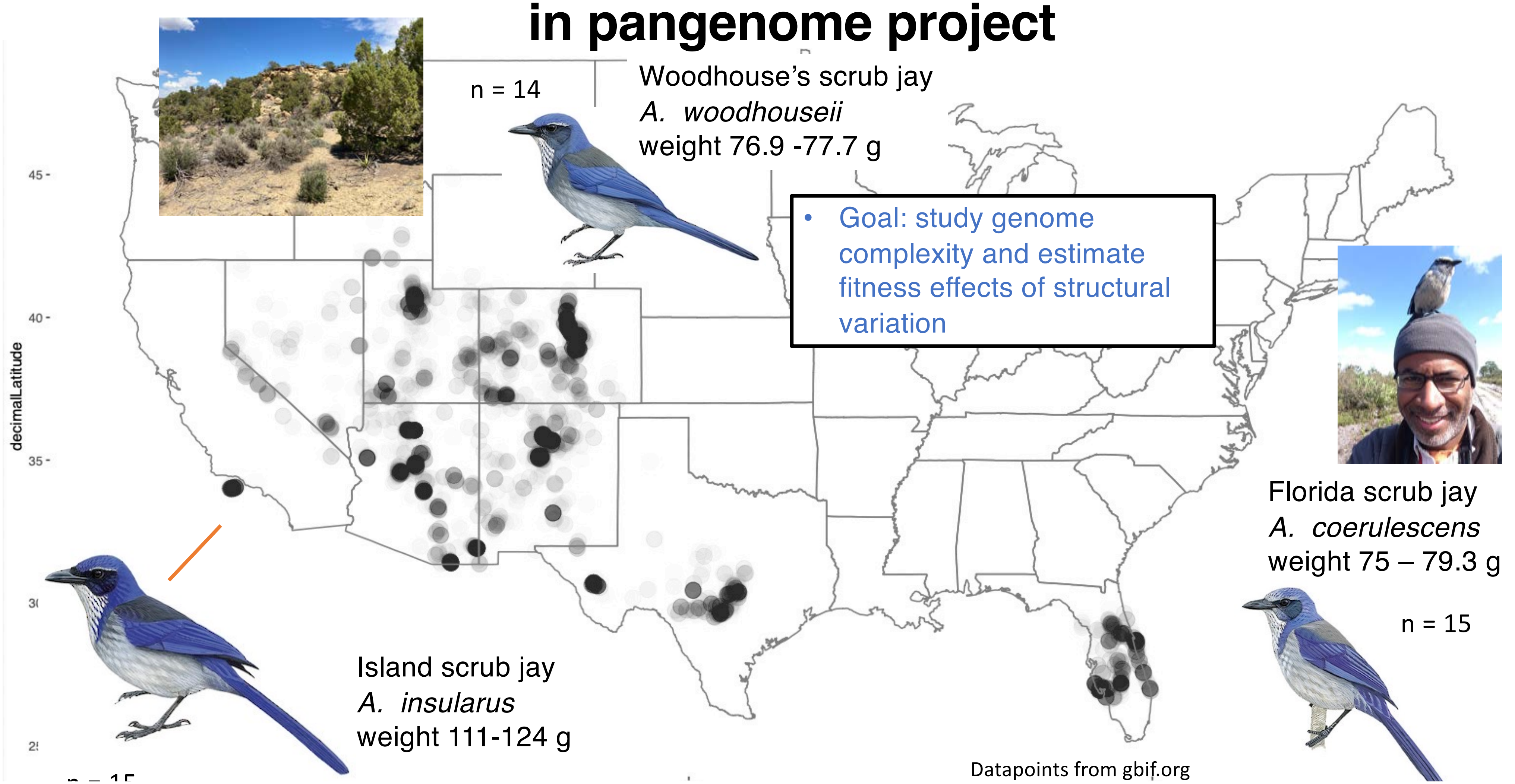
Avian genomes are growing with each new technology





Data from NCBI, accessed 13 Nov. 2021

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



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

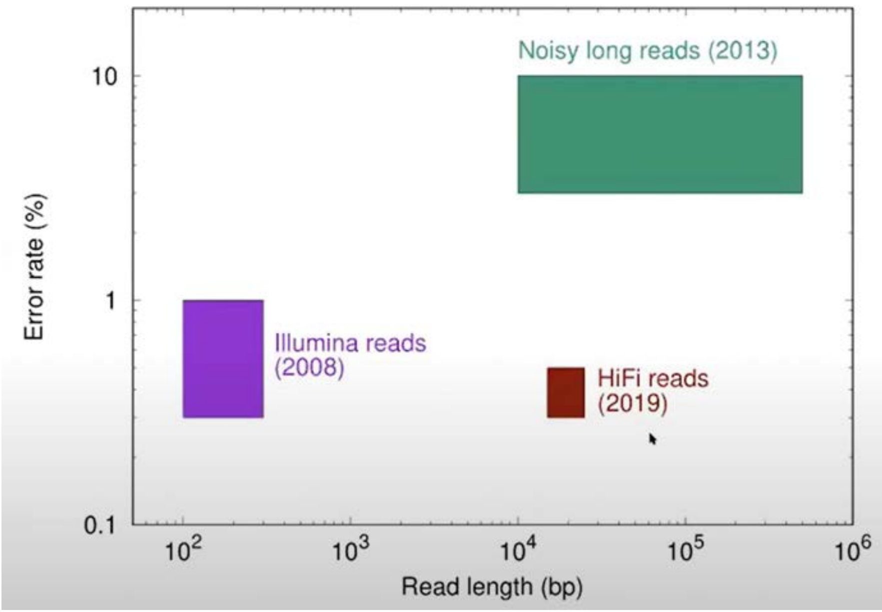
Conceptual Relationships between the Fields of Comparative Population Genomics, Landscape Genomics, and Comparative Phylogeography

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

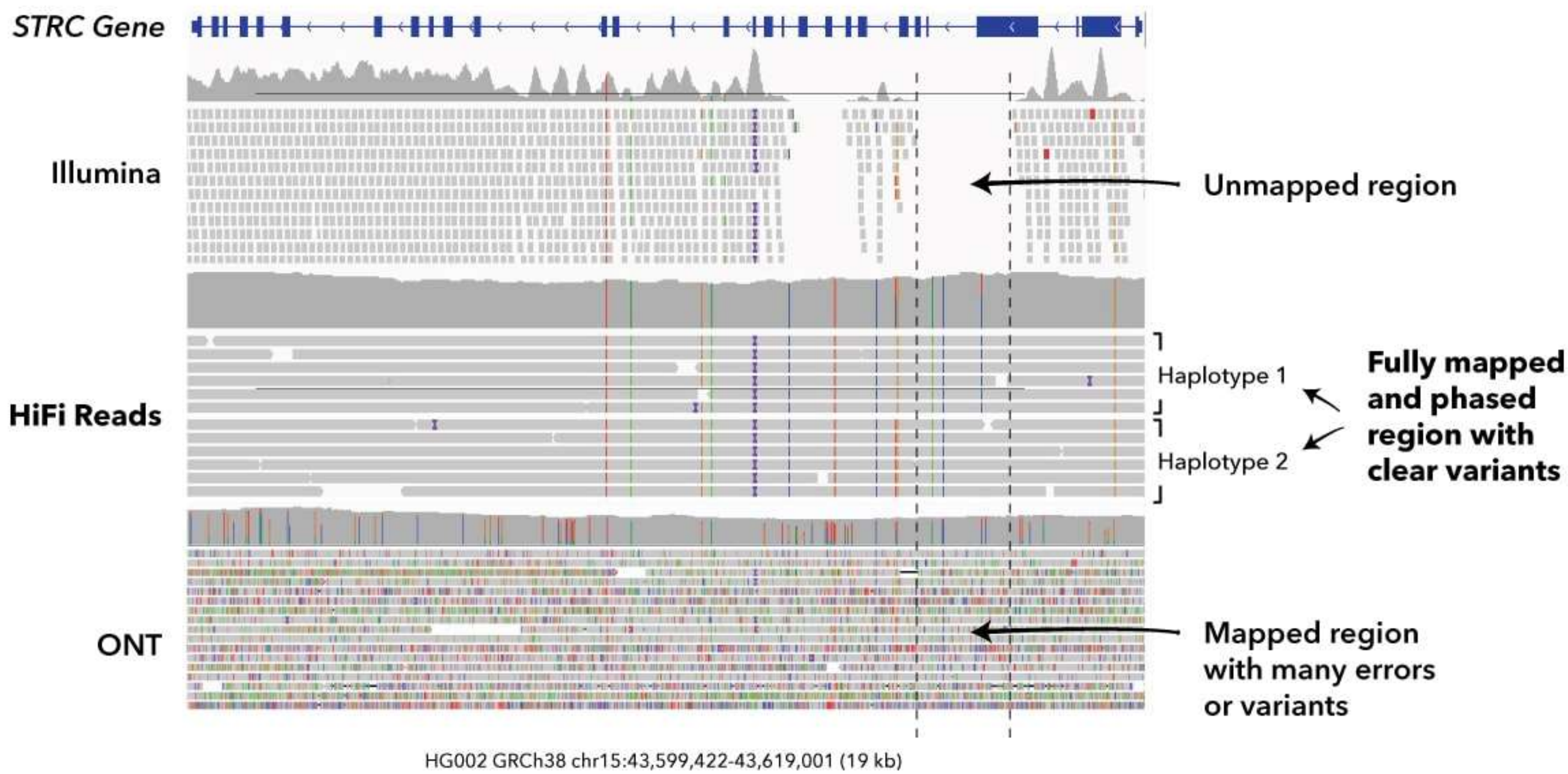
PacBio HiFi reads are long and accurate

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

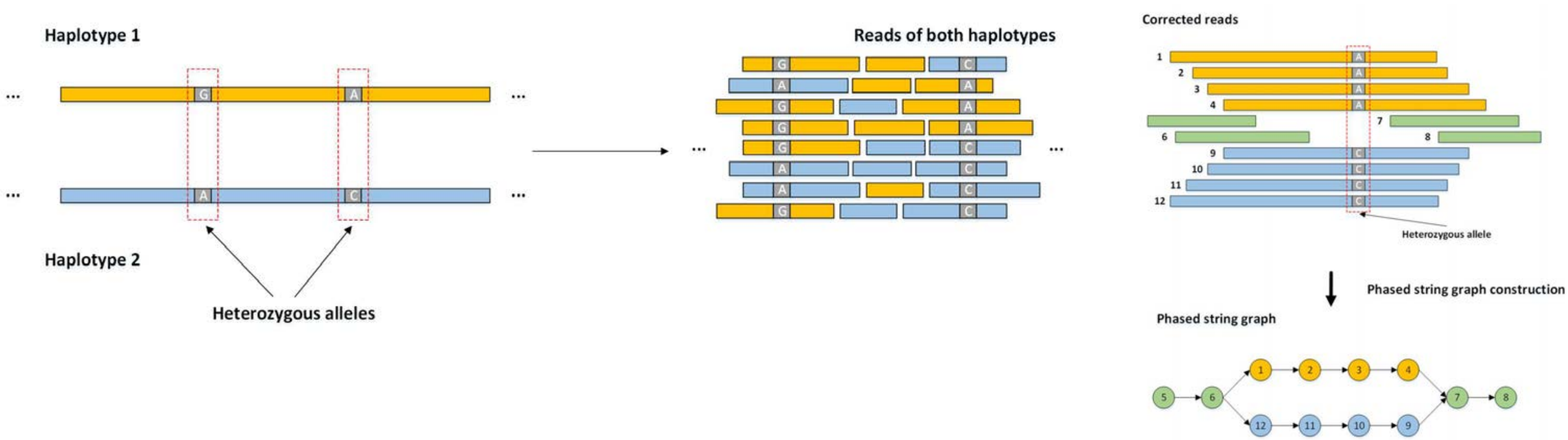


Coutesy Haoyu Cheng, Dana Farber Cancer Institute

PacBio HiFi reads are long and accurate

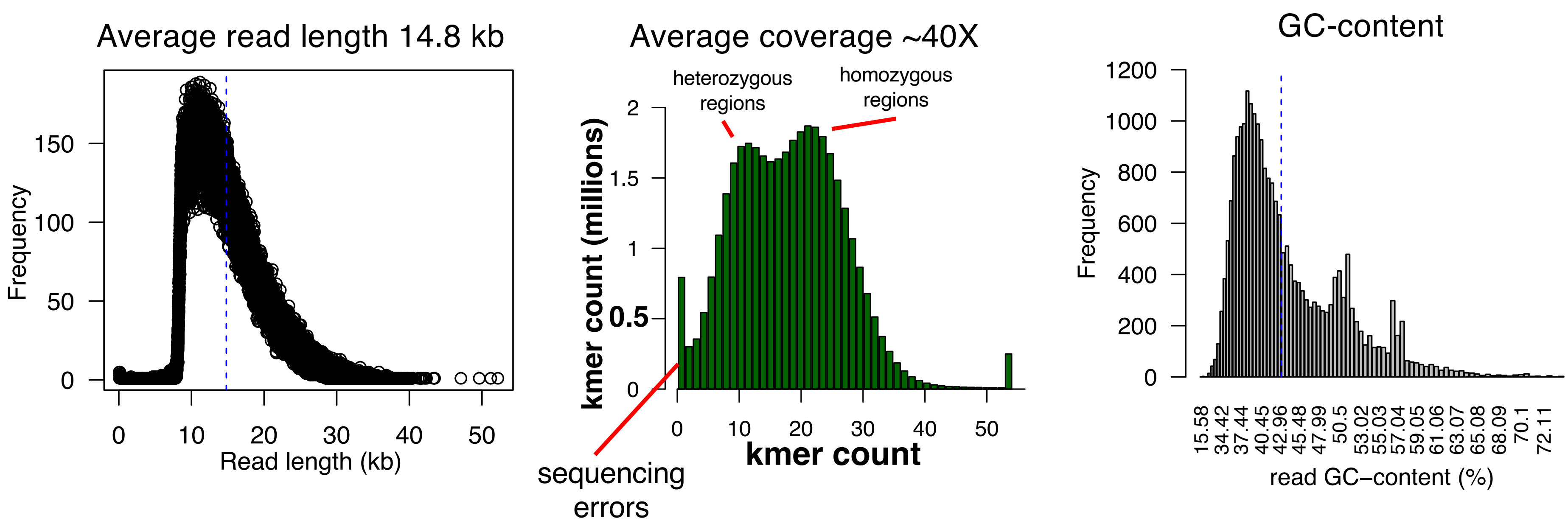


Hifiasm – a HiFi accurate read assembler that resolves haplotypes

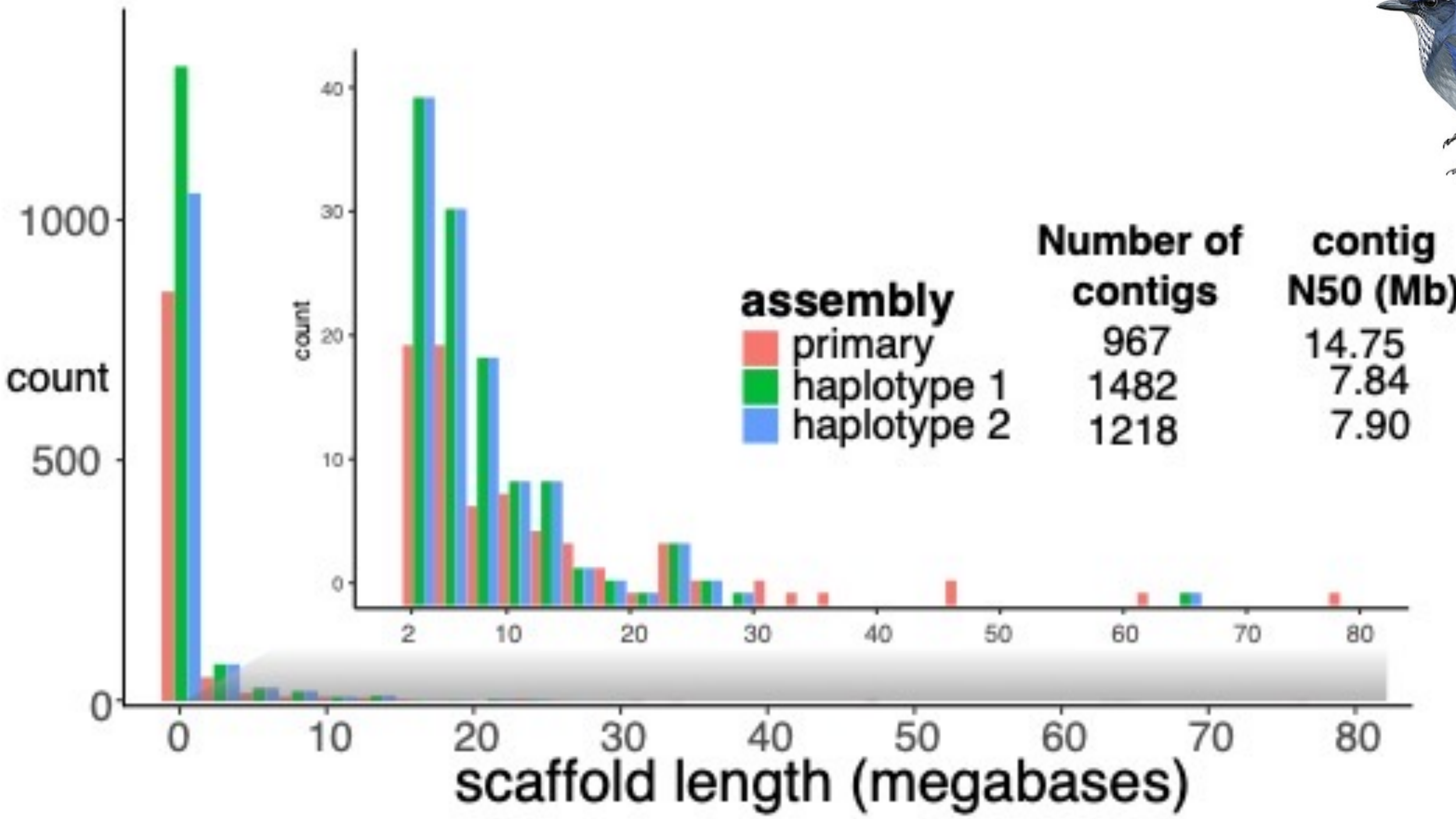


Coutesy Haoyu Cheng, Dana Farber Cancer Institute

Scrub-jay PacBio HiFi data characteristics

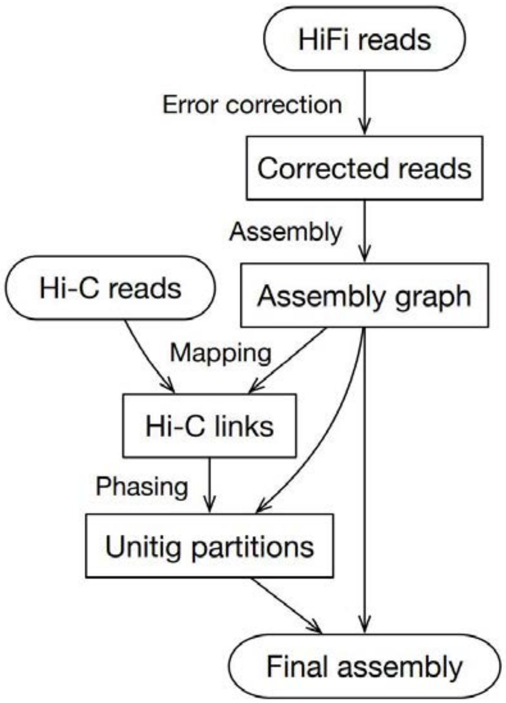


Genome assembly with hifiasm yields ~1.3 Gb primary and haplotype assemblies

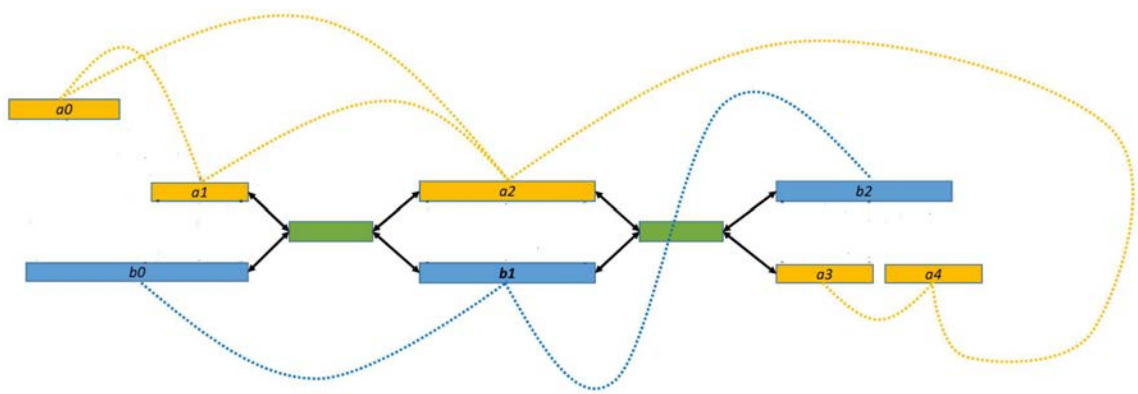


A. woodhouseii

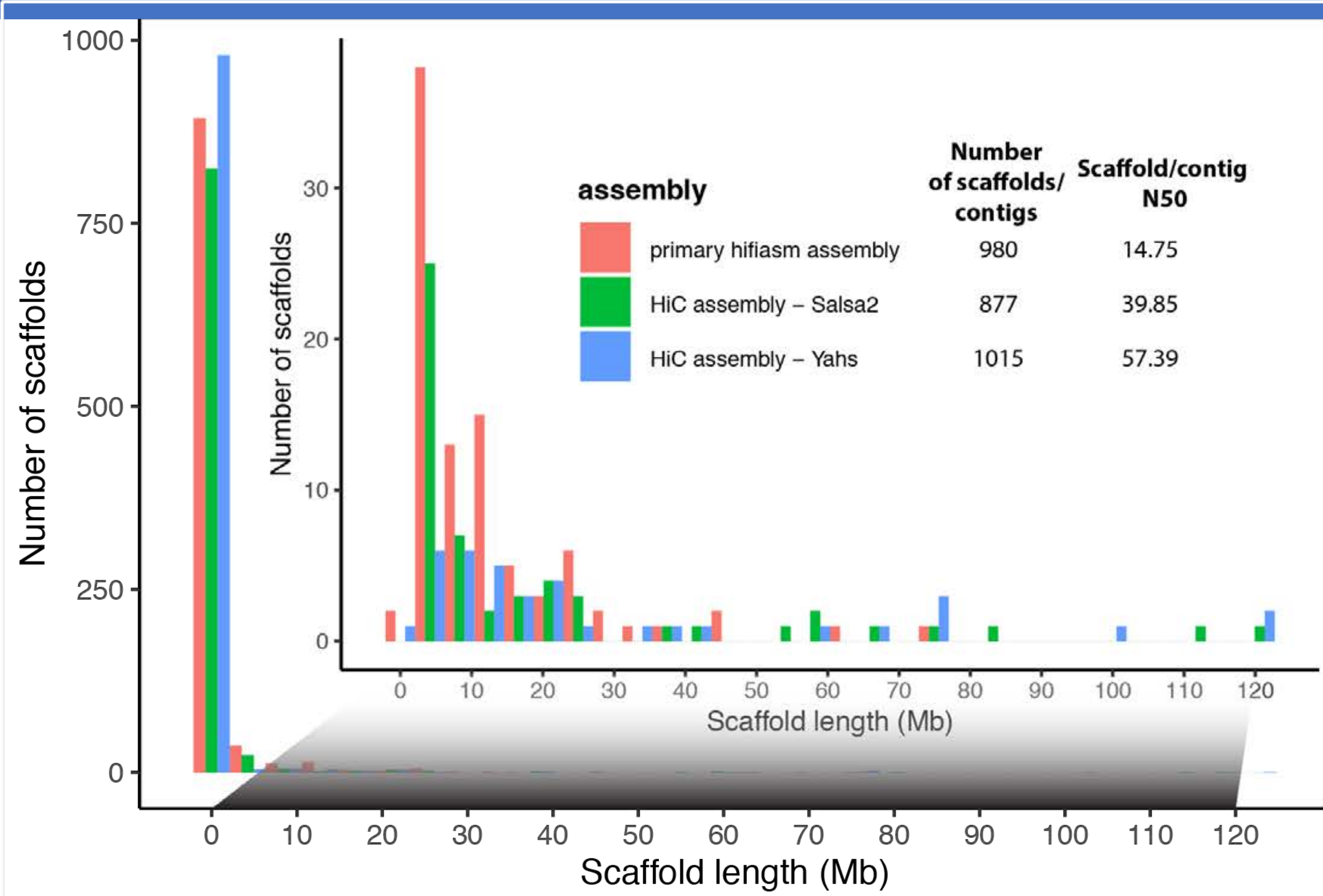
Hifiasm – improved assemblies using HiC



- Procedure:
  - Identify heterozygous unitigs by coverage
  - Build index by unique *k*-mers from heterozygous unitigs
  - Align Hi-C reads using unique *k*-mers

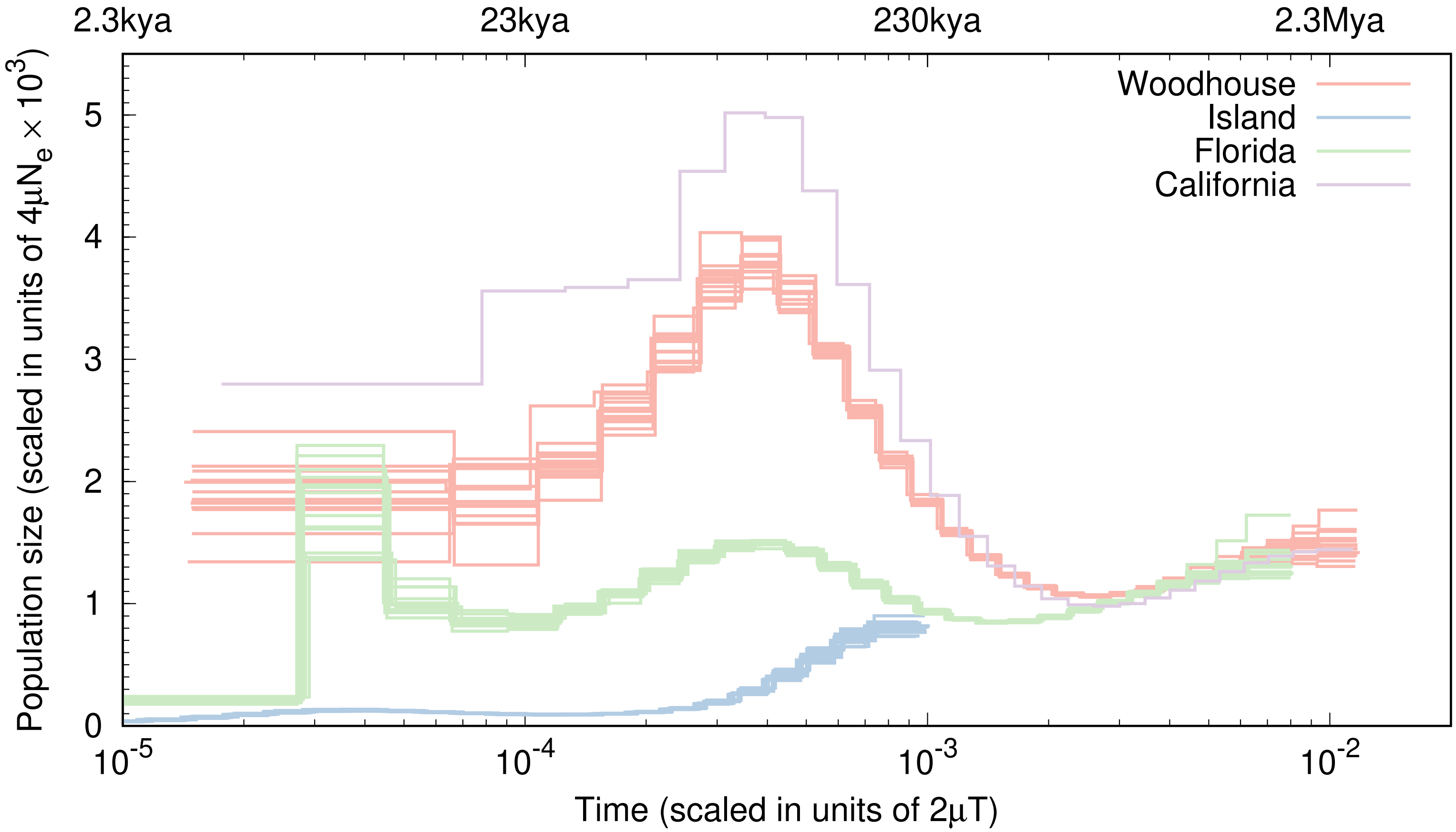


HiC greatly improves contiguity of scrub jay assemblies

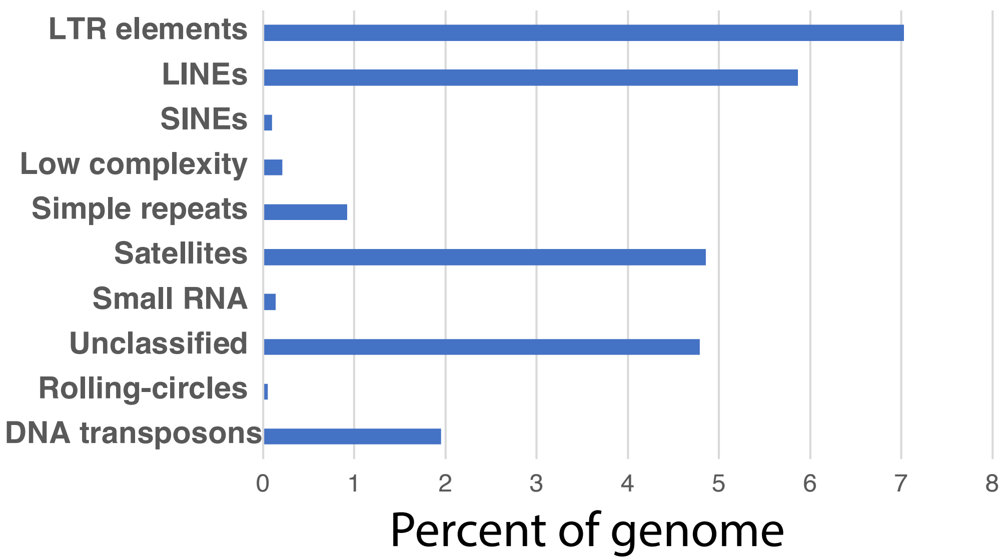


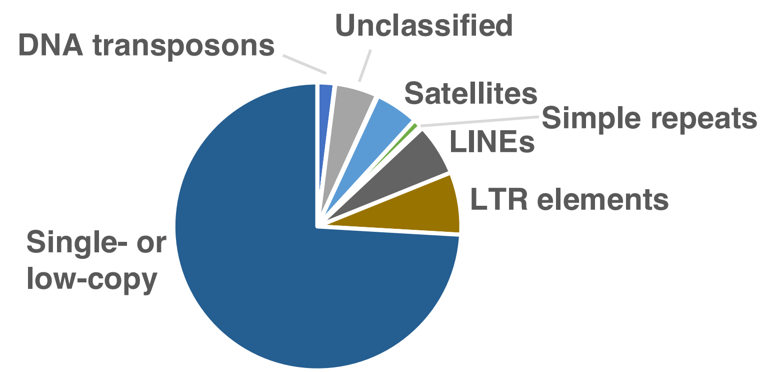
PSMC analysis confirms variation in effective population size through time

Pairwise Sequentially Markovian Coalescent (Li & Durbin 2011. Nature)

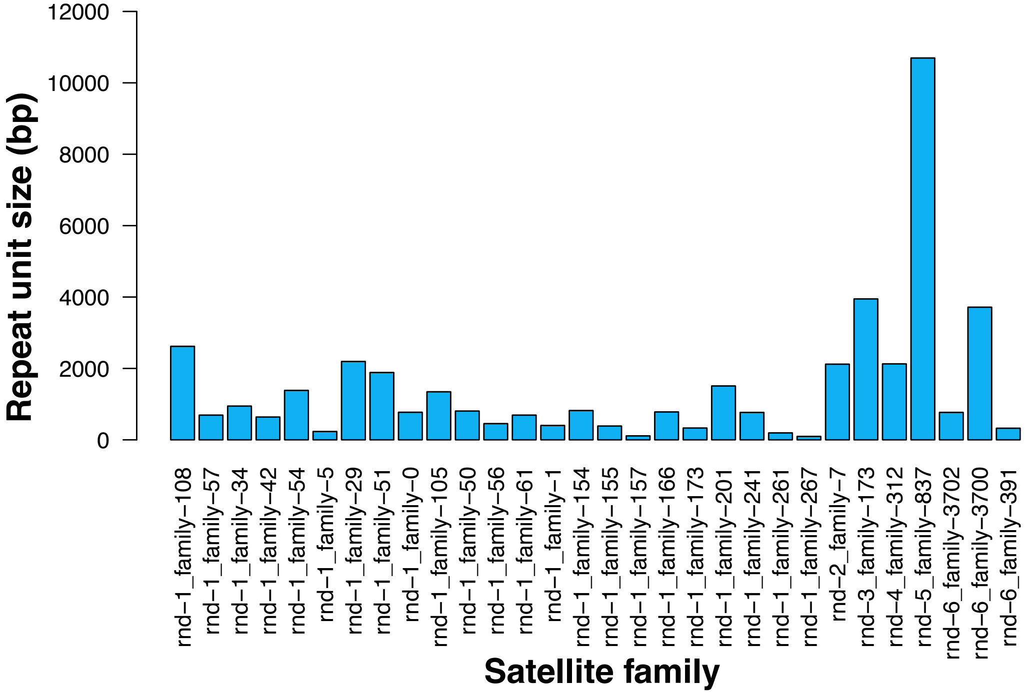


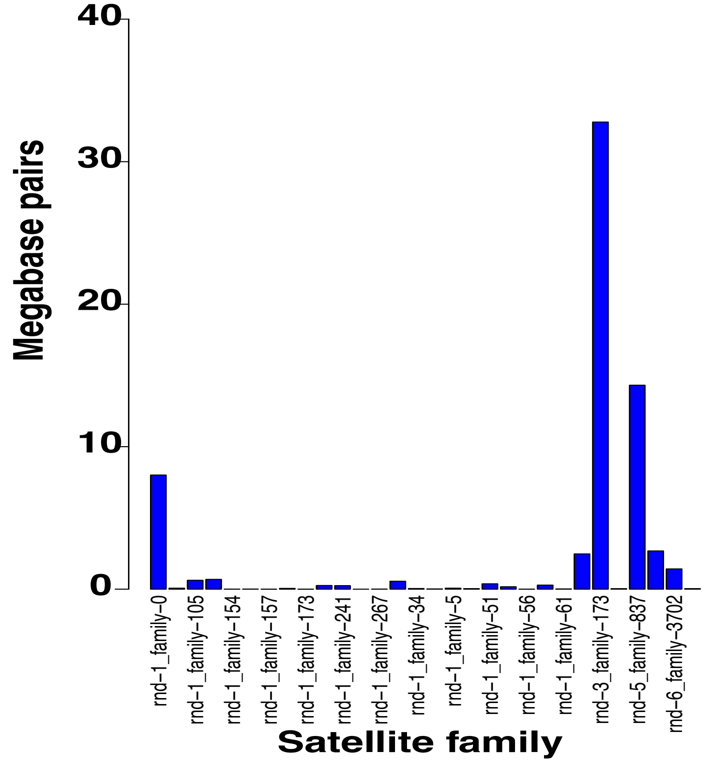
RepeatMasker analysis suggests over 25% repeats and transposable elements



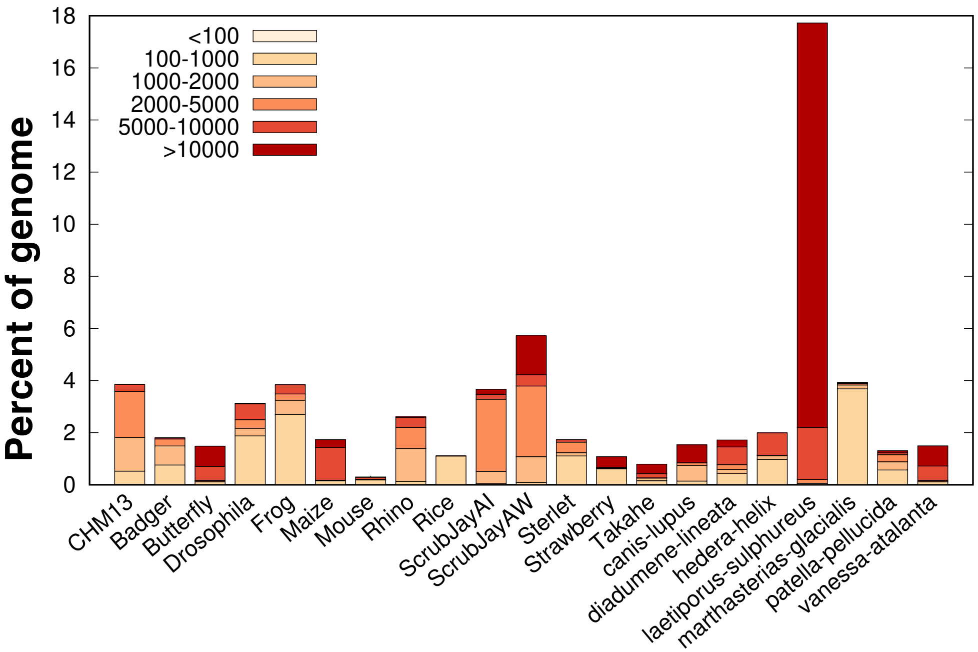


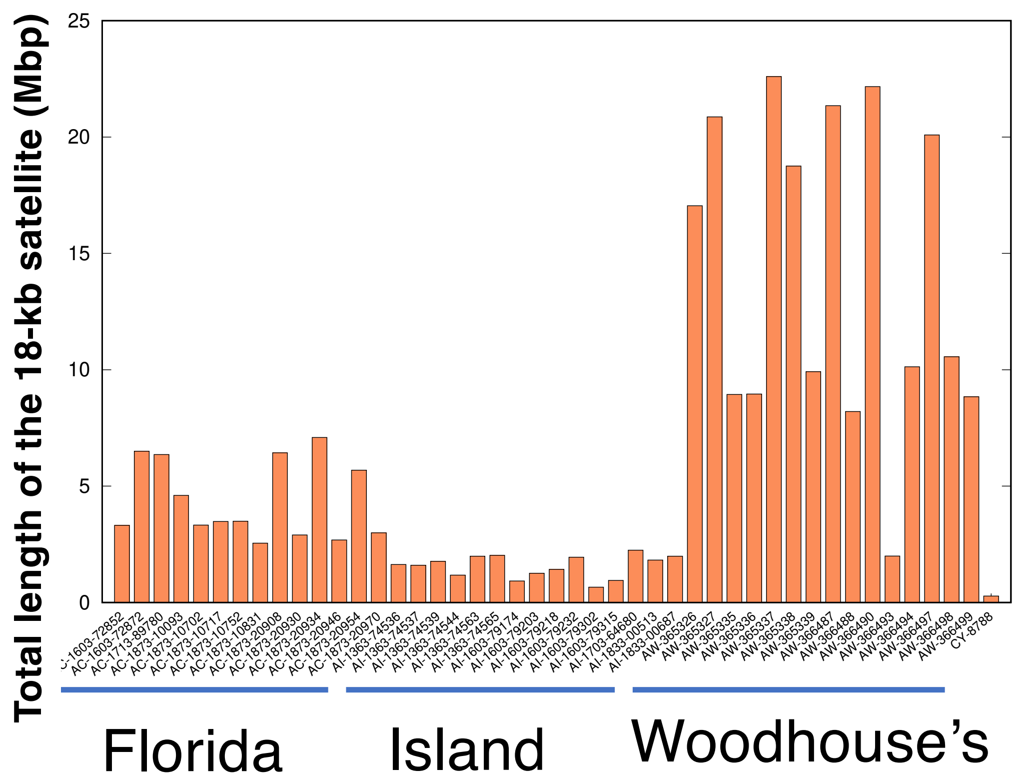
Satellites are long and prevalent in scrub jay genomes





Abundance of an 18-kb unit repeat satellite varies strongly among species





Pangenome graphs capture structural variation within species

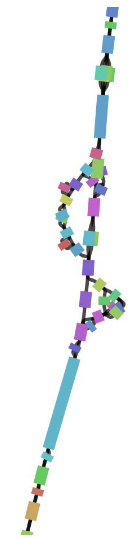
```
>h1tg000104l
GGCGGGGCCCGGAGGGGCCGGGGCCGCTGAGGGGCCGCGGGTGCGGCAGAGCC
>h1tg000528l
ATGGATACTTTCCAGTCAGAGCTTTATAATAATTTCCATAATTTAAATATTTT
>h1tg000795l
ACTTTGGGGACACCTTTGGGGACACCTCGGGGGACACTTTGGGCCACAAATCC
```

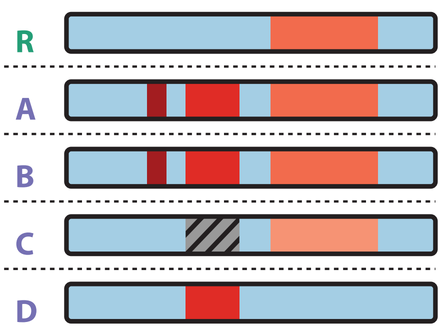
unaligned fasta files

→ **minigraph** **odgi** — build genome graph

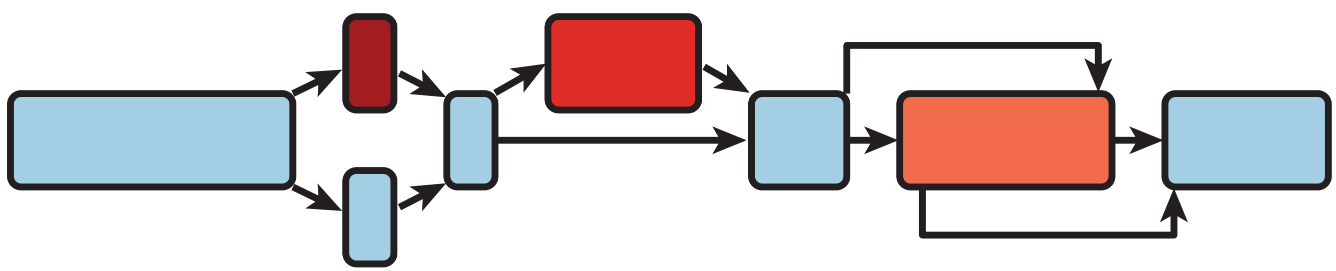
→ **odgi viz** **bandage** — visualize genome graph

→





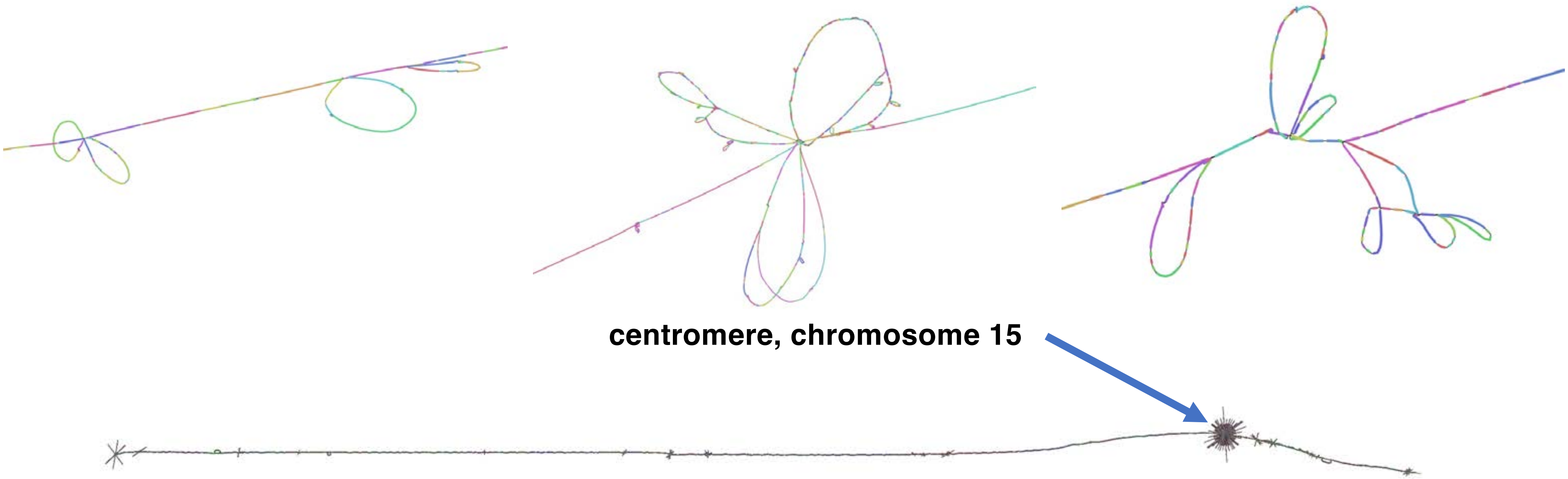
Multiple sequence alignment



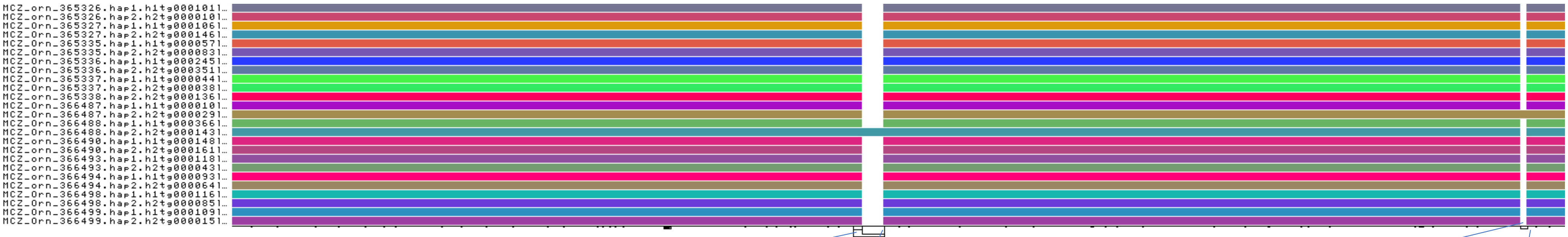
Bidirected genome graph

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

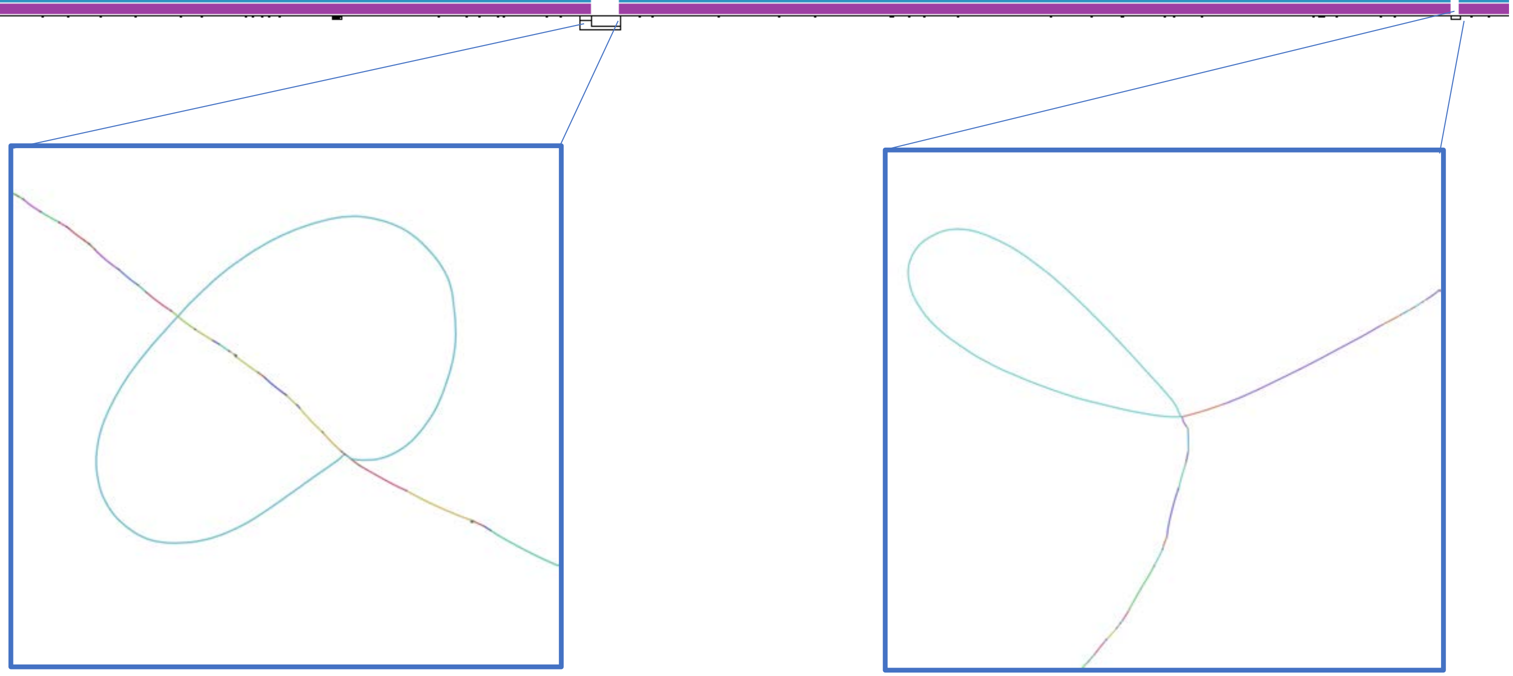
Pangenome graphs of haplotype variation in Scrub Jays



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



Pangenome graphs generated with odgi and visualized with Bandage

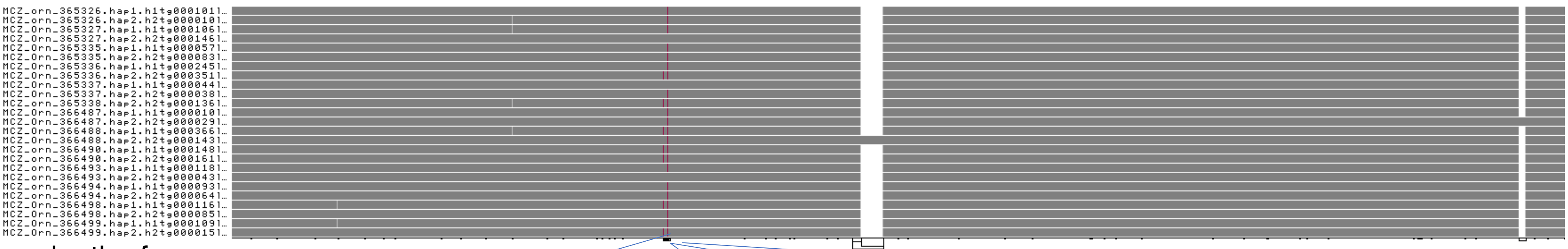


7.5 kb polymorphic indel

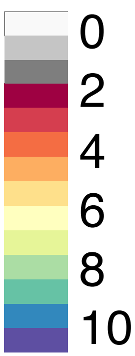
2.5 kb polymorphic indel

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

Smaller regions of complexity in hox1a region



depth of *hox1a* region graph (x)

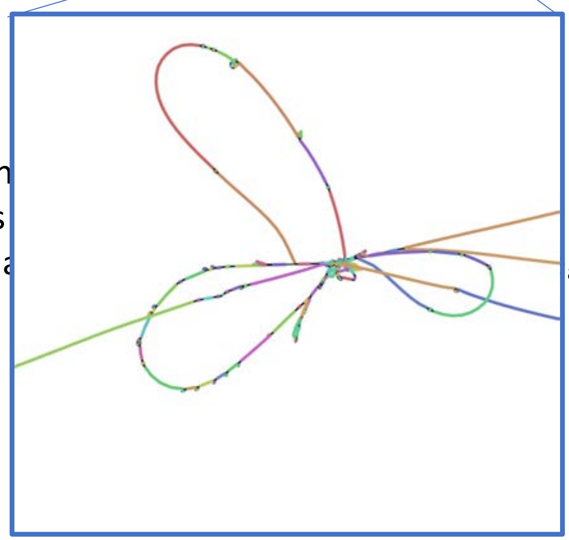


White region

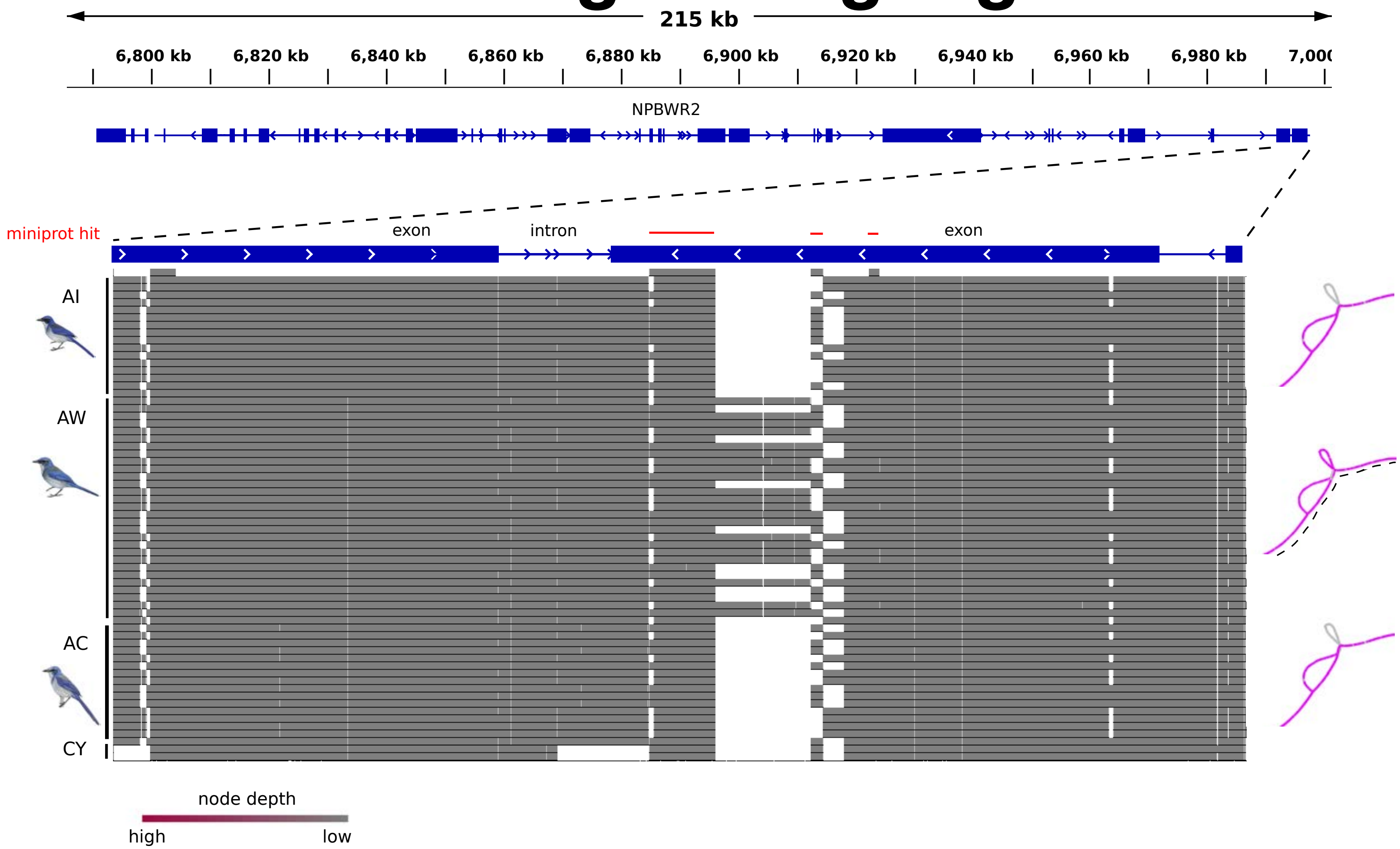
Gray regions

Red regions a

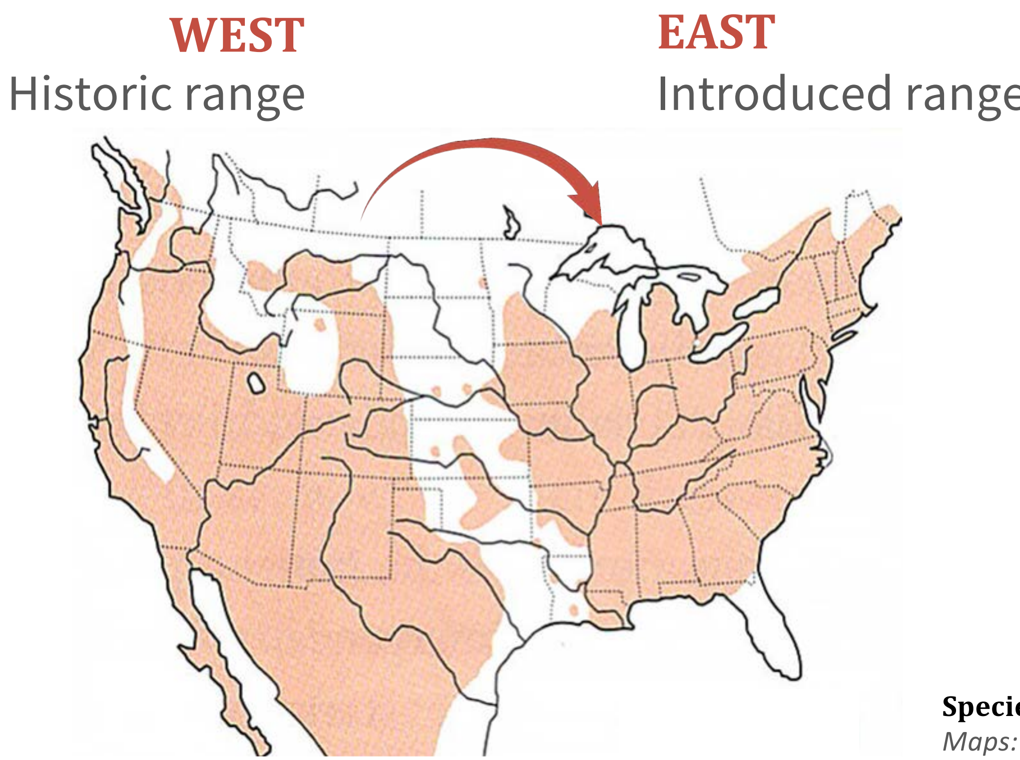
gions



Structural variation in another conservatively evolving coding region

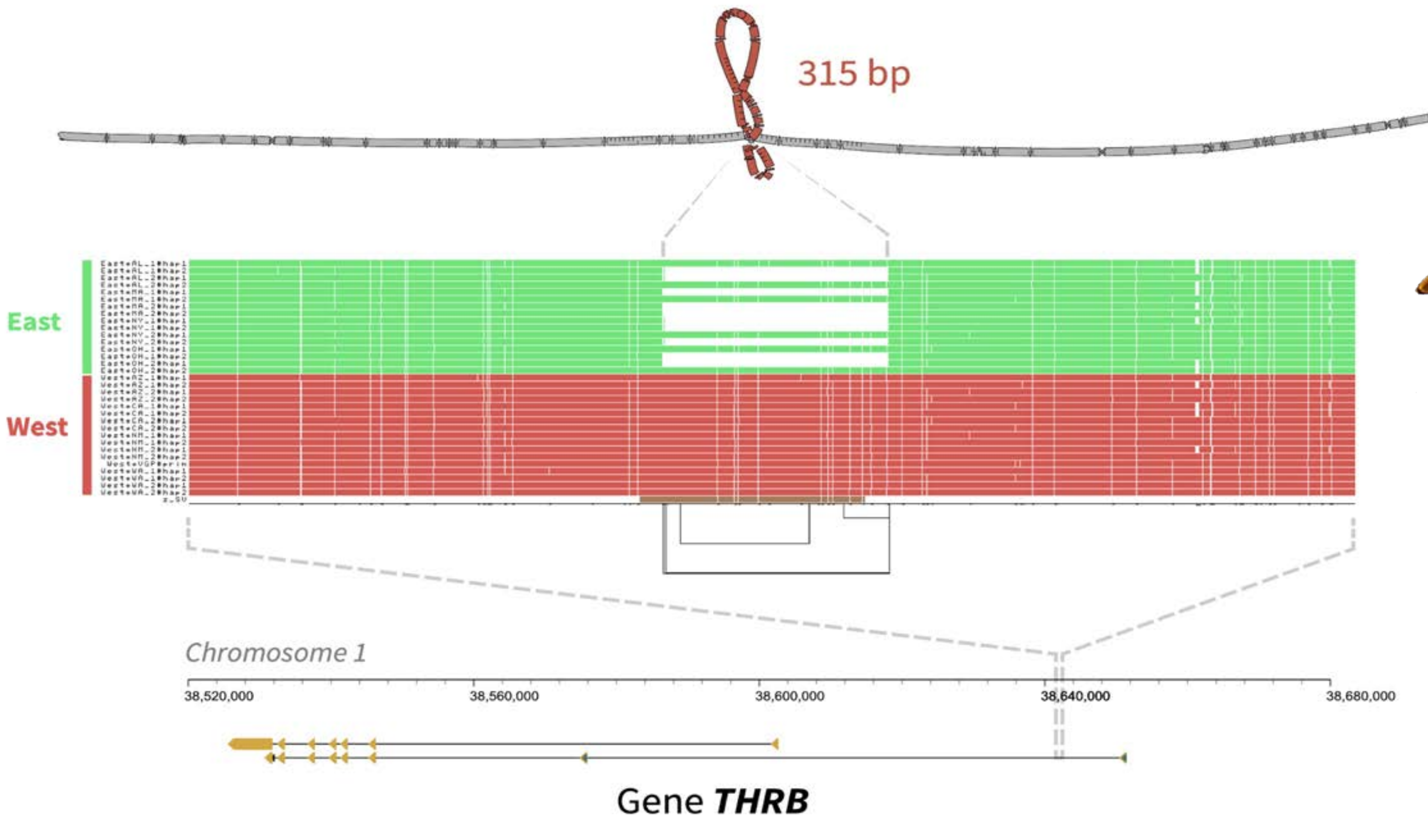


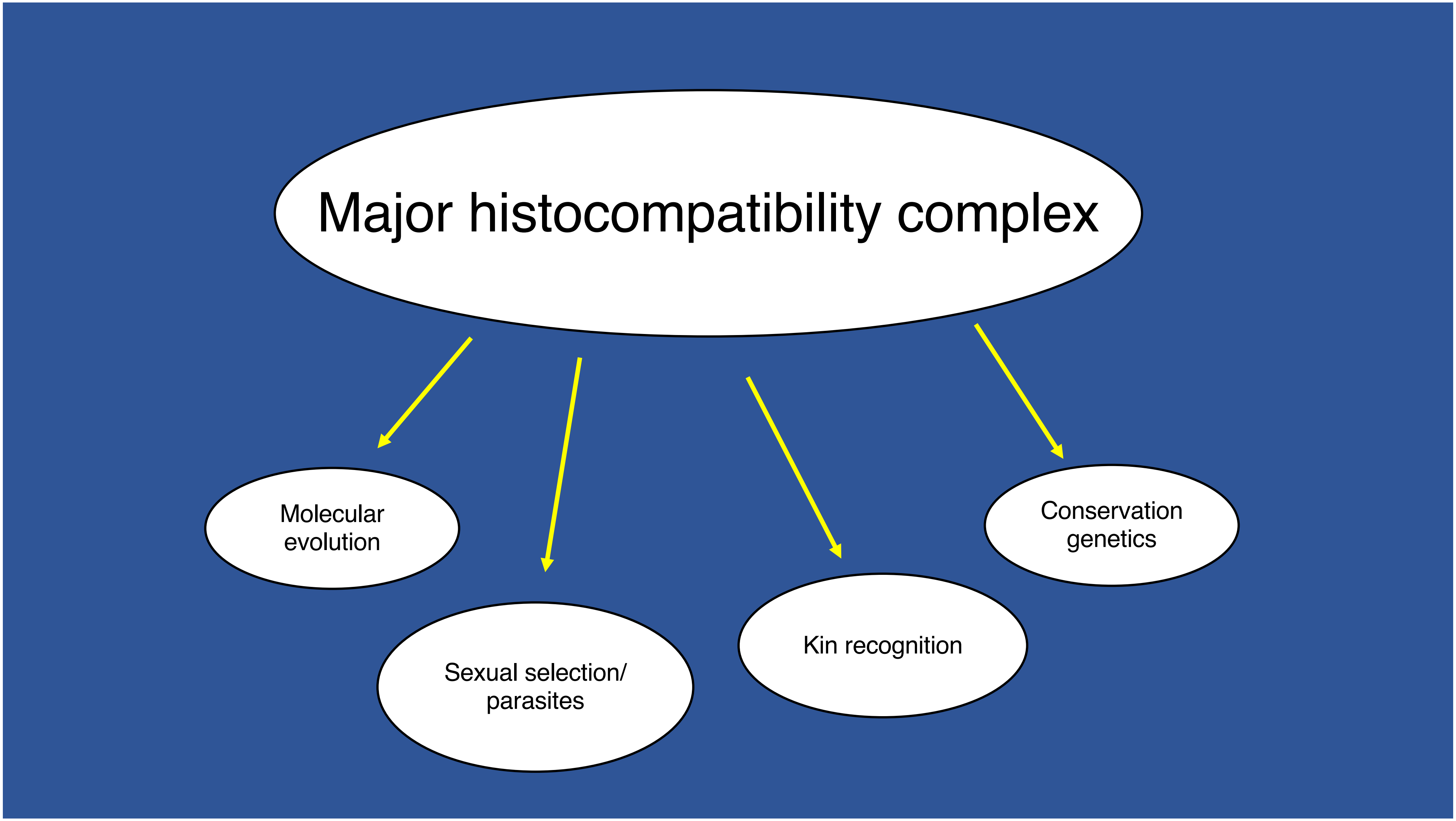
Demographic history of House Finches



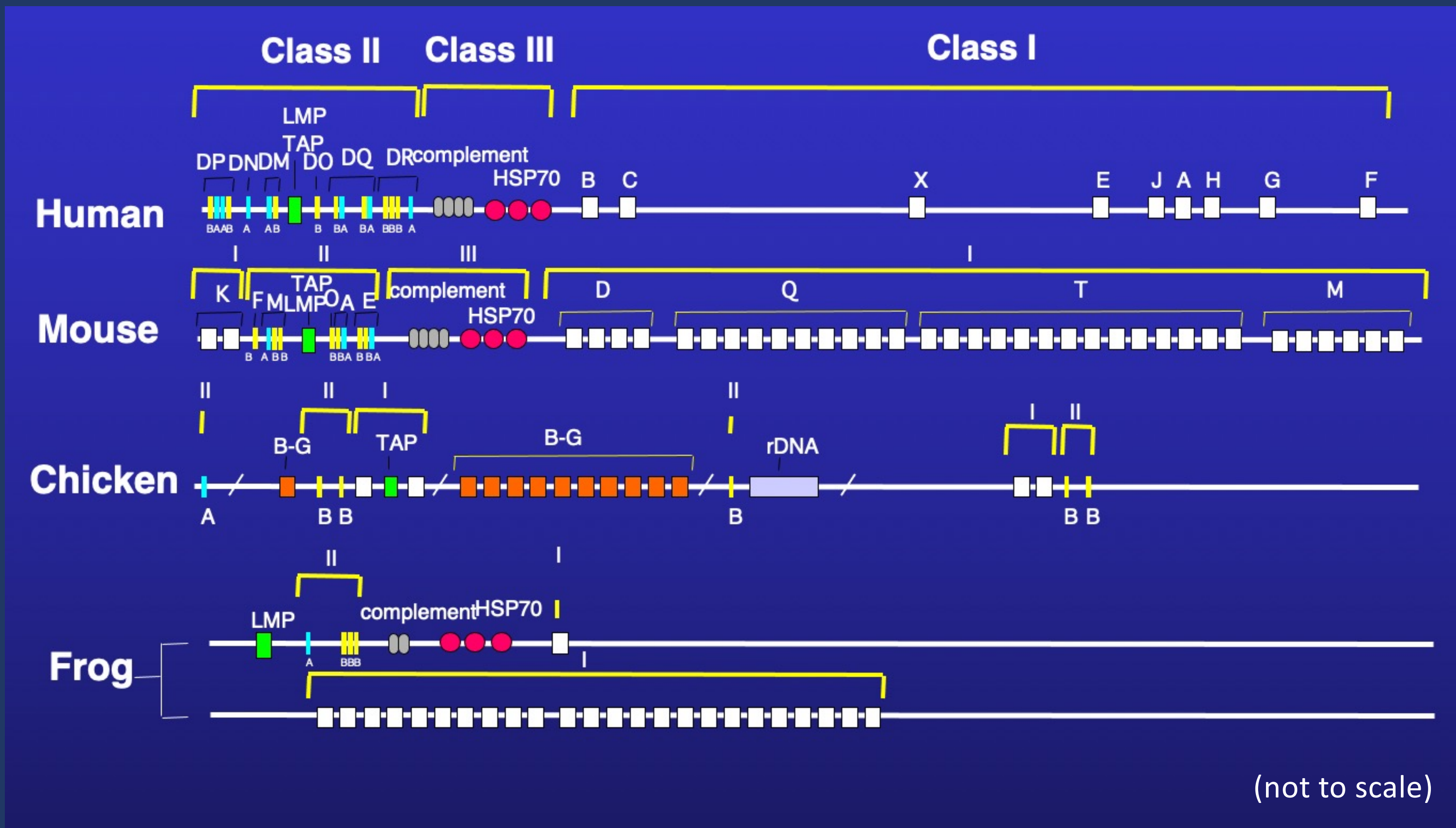
Species distribution in ~1990
Maps: Birds of North America

Structural variants in the thyroid receptor-β gene

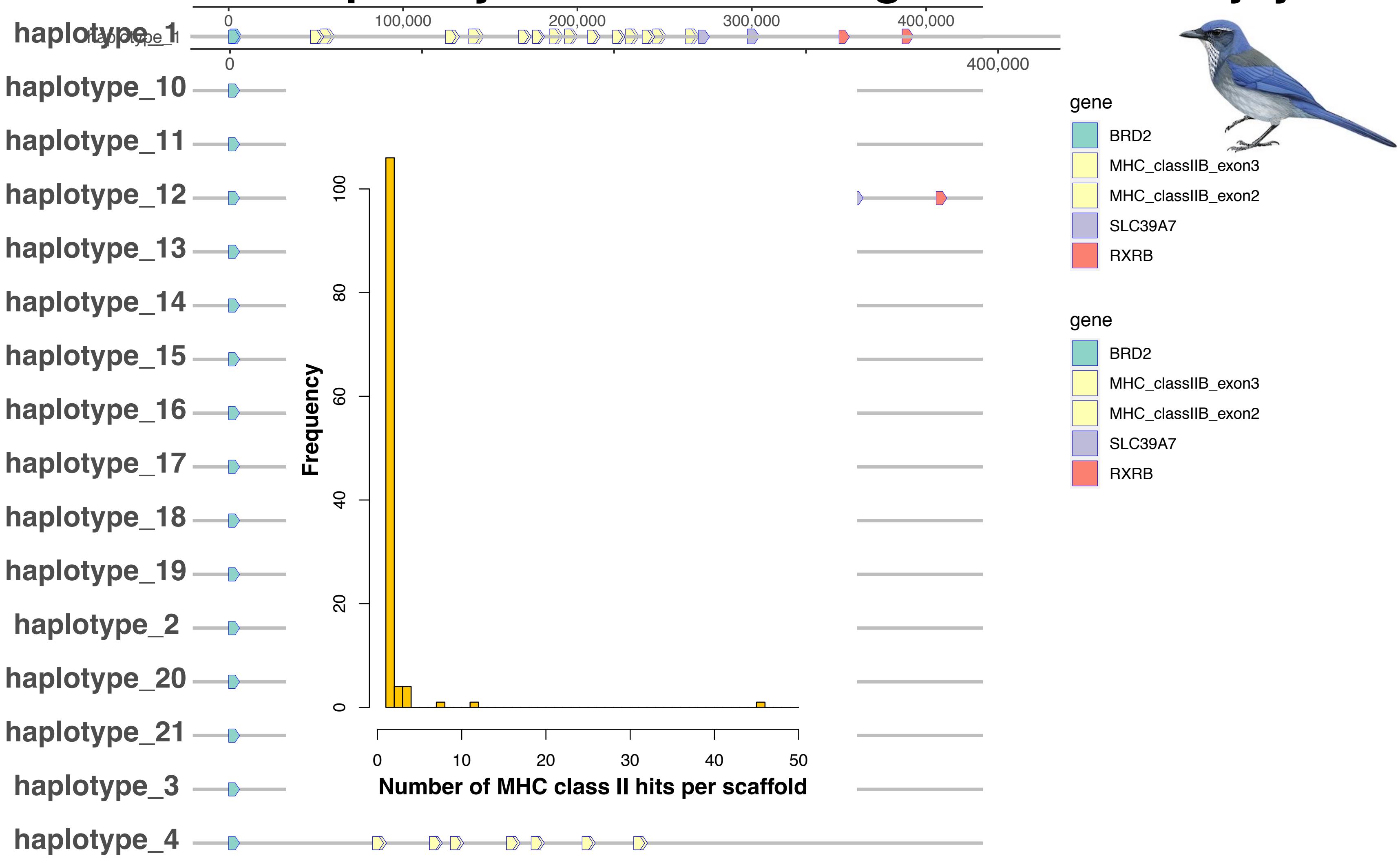



Major histocompatibility complex
Molecular evolution
Sexual selection/ parasites
Kin recognition
Conservation genetics

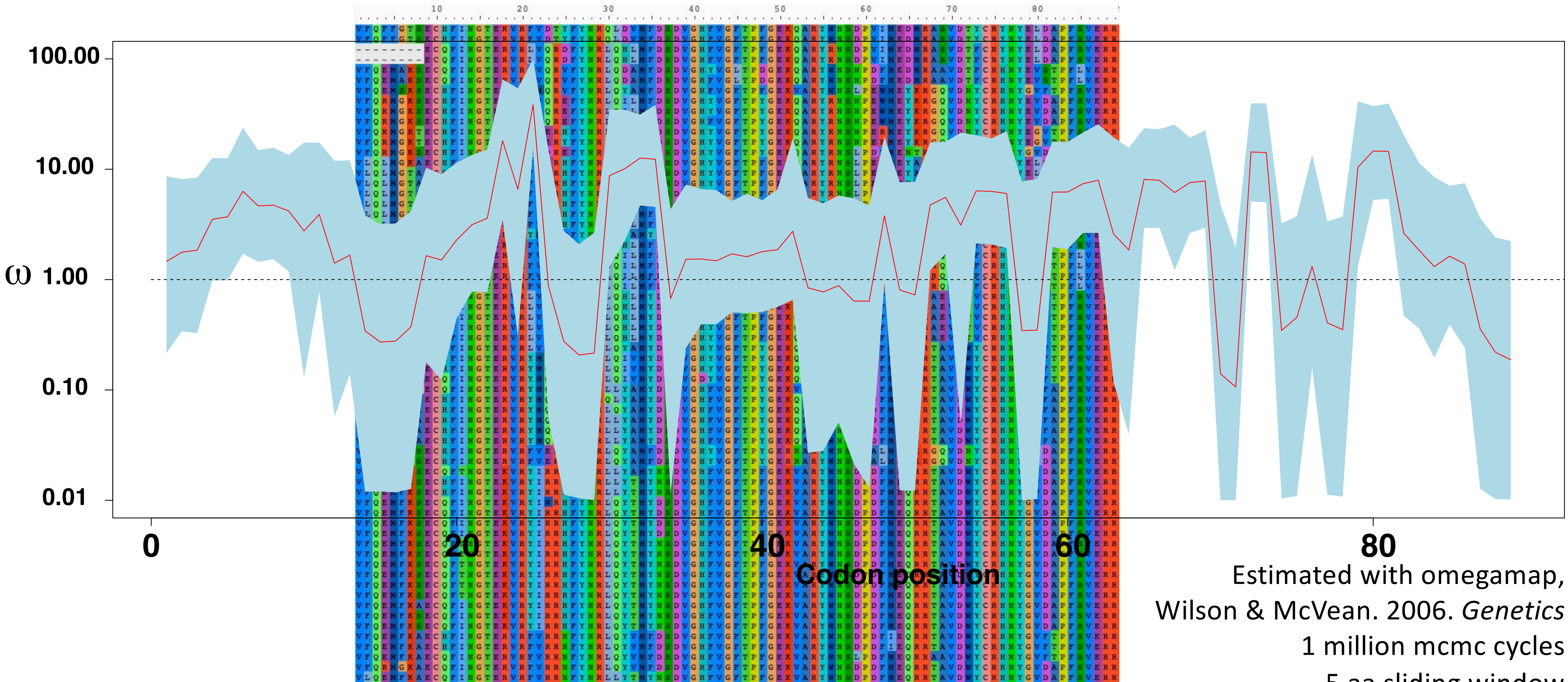
The chicken MHC is small (~99 kb) and compact



Unprecedented complexity of MHC class II genes in scrub jays

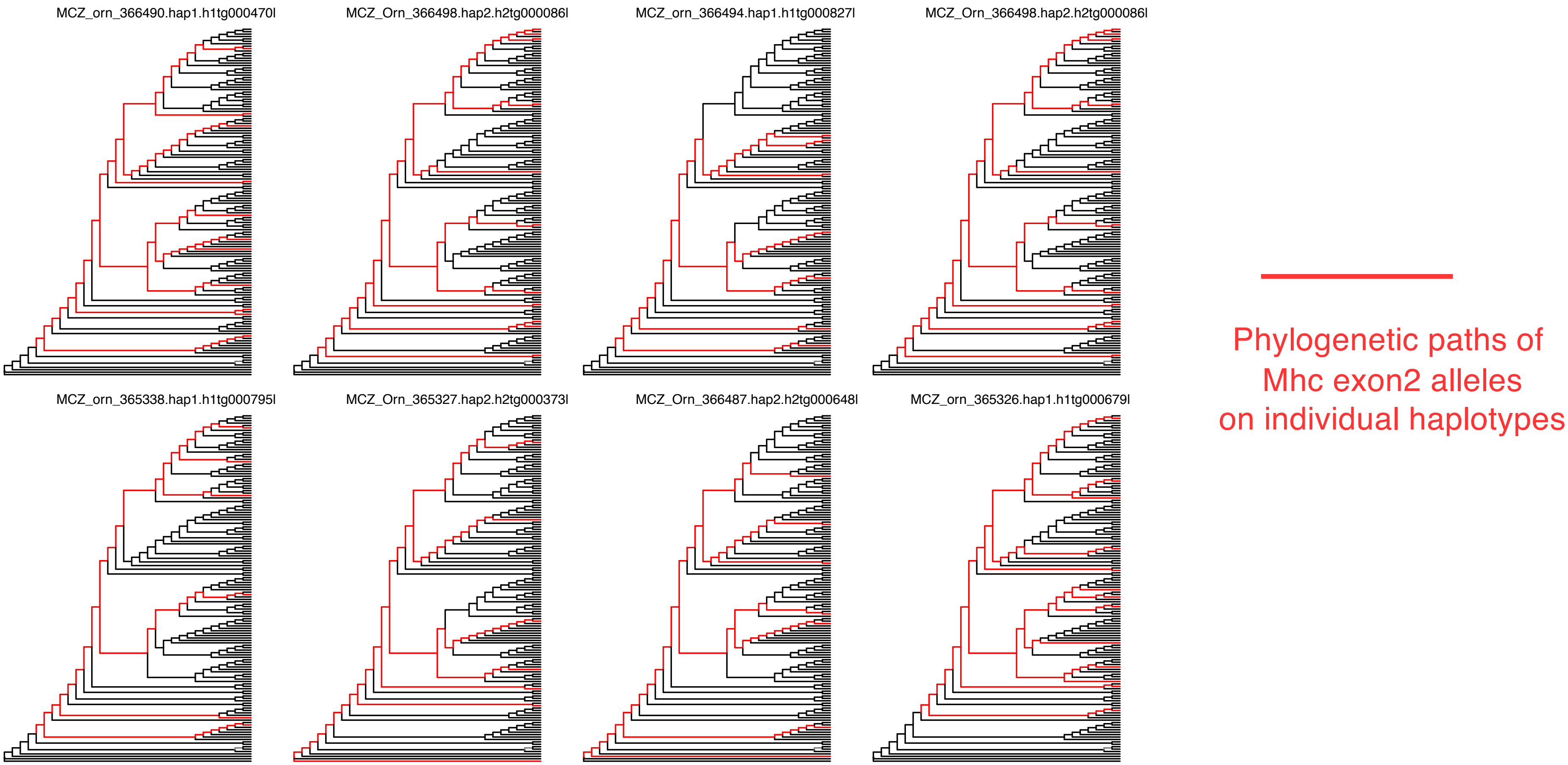


Mhc class II peptide-binding region shows solid evidence of balancing selection

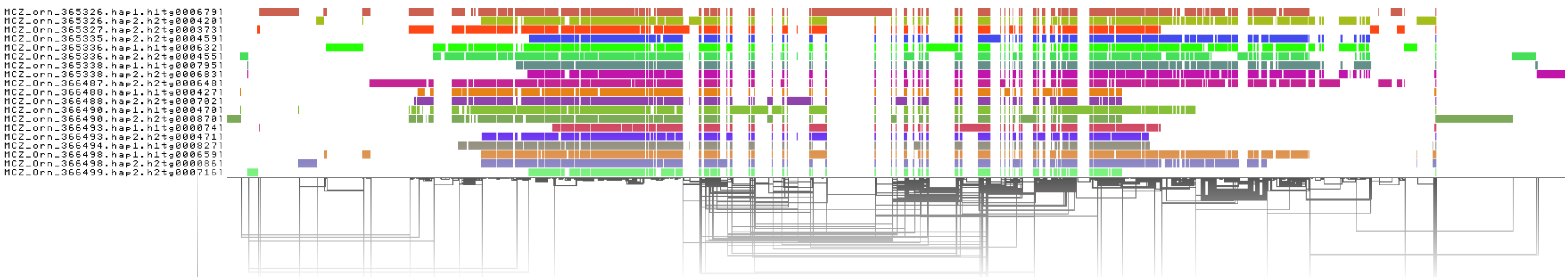


Estimated with omegamap,
Wilson & McVean. 2006. *Genetics*
1 million mcmc cycles
5 aa sliding window

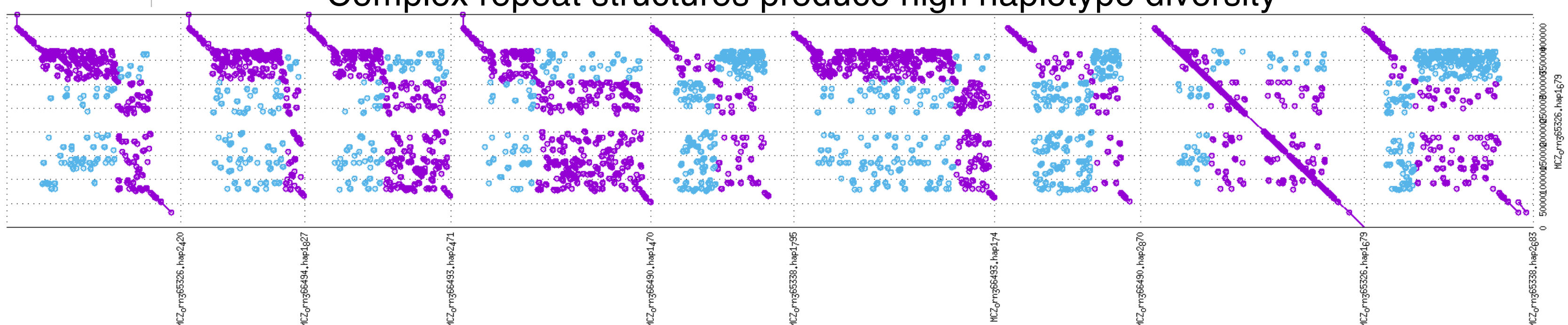
Mhc class II peptide binding regions are phylogenetically diverse on individual haplotypes



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

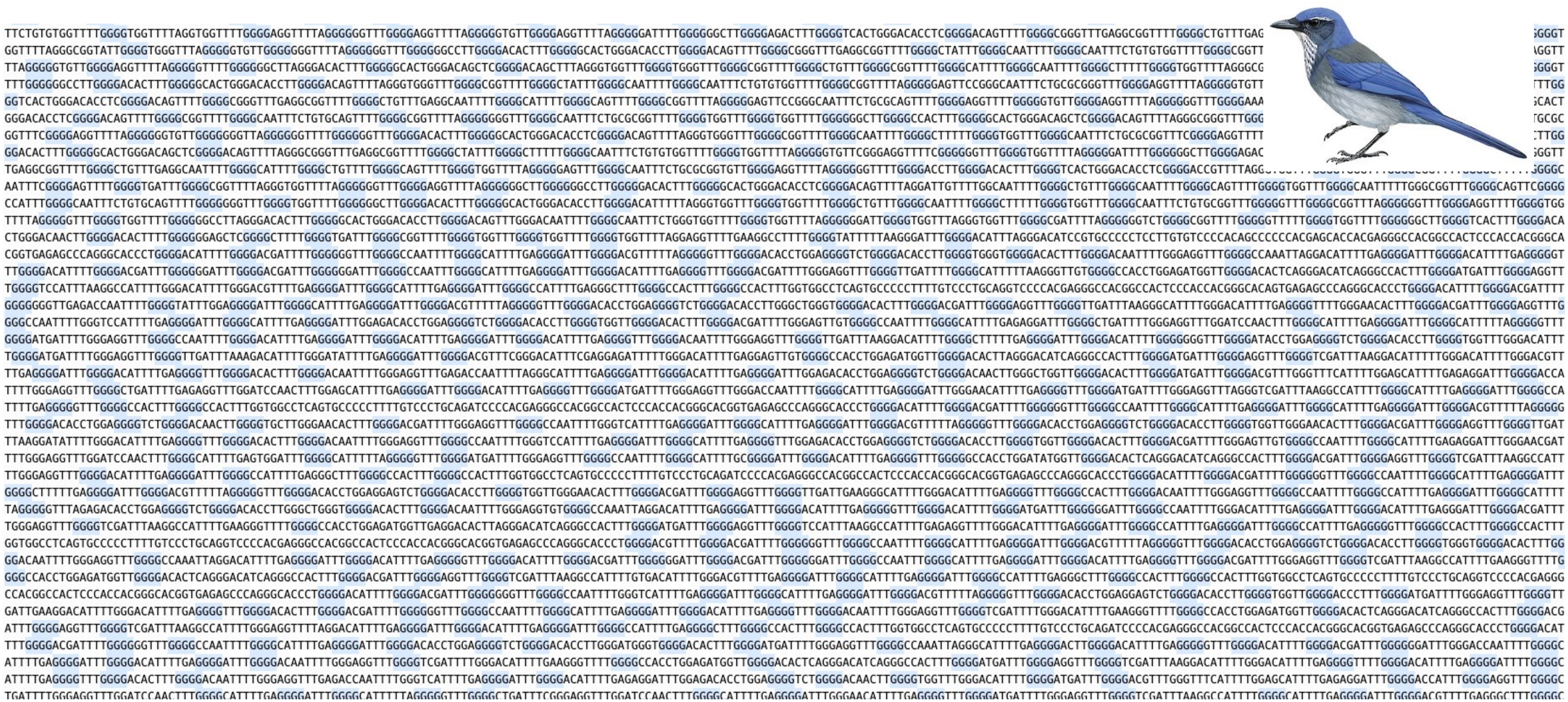


Complex repeat structures produce high haplotype diversity

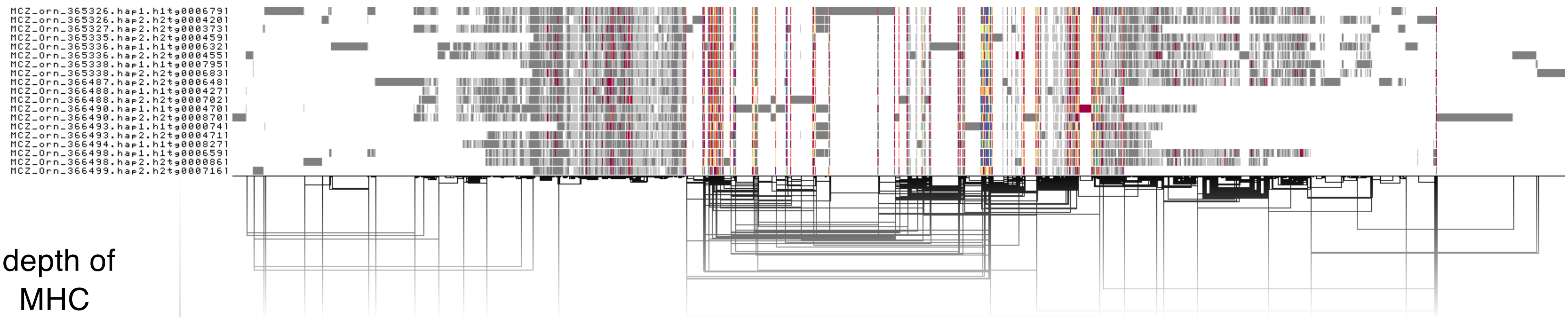


made with odgi and pangenome graph builder pipeline
Guarracino et al. 2021. *Bioinformatics*, in press.

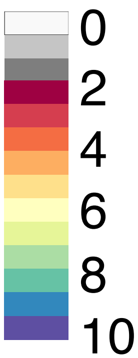
Example satellites in MHC class II region of Woodhouse's scrub jay



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



depth of MHC graph (x)

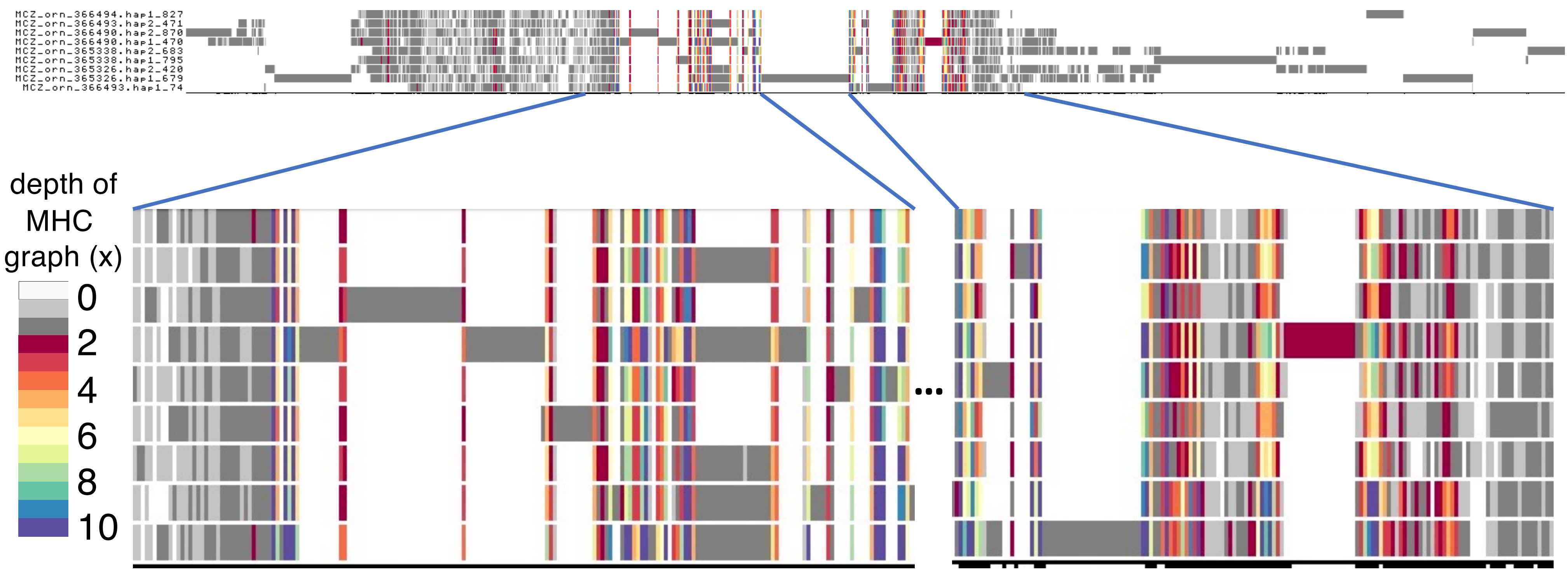


odgi visualization of pangenome graph depth based on 26 MHC-containing scrub jay haplotypes, up to ~480 kb

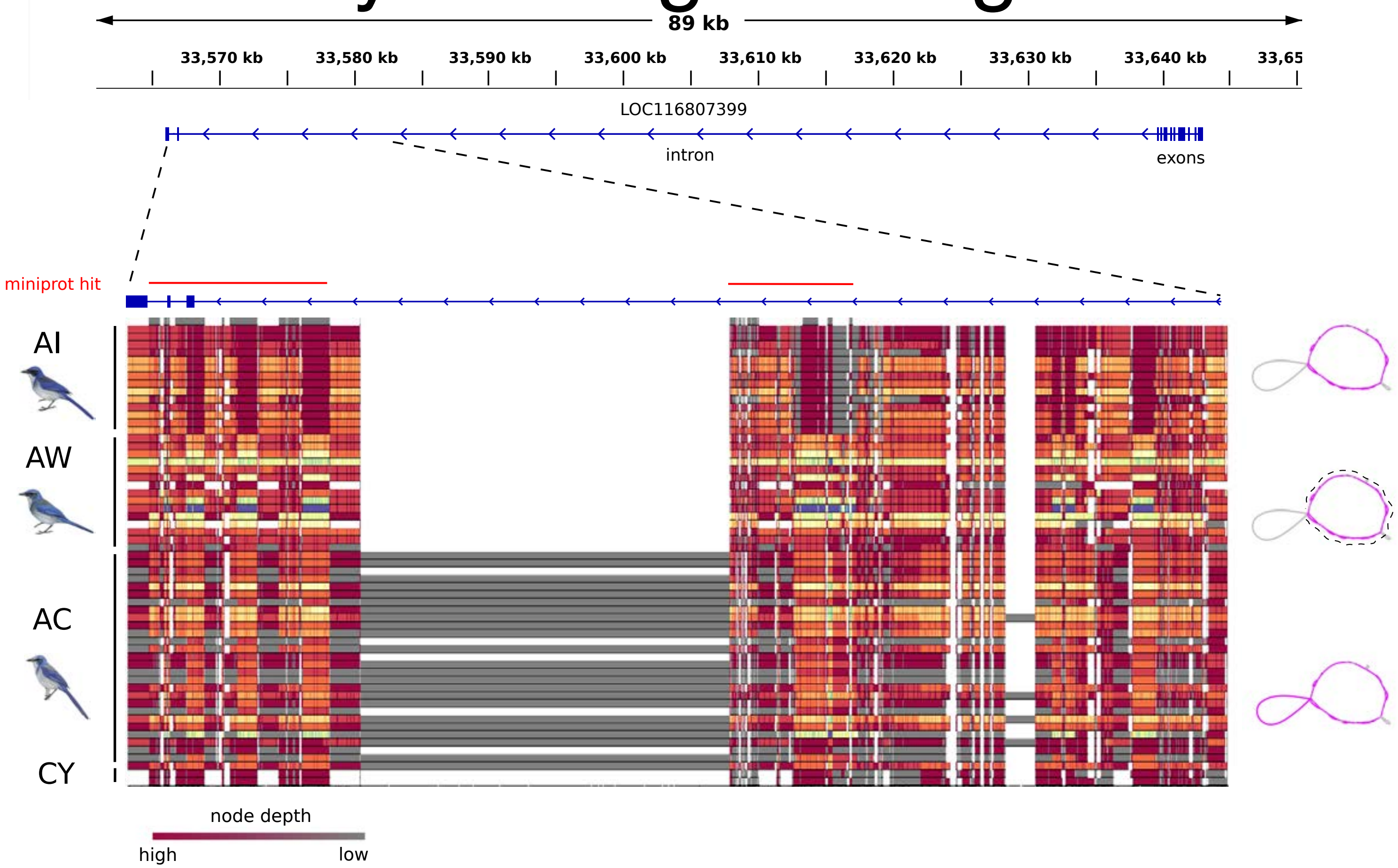
White regions are indels between haplotypes
Gray regions are SNP variation
Red regions are low-complexity and repetitive regions

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

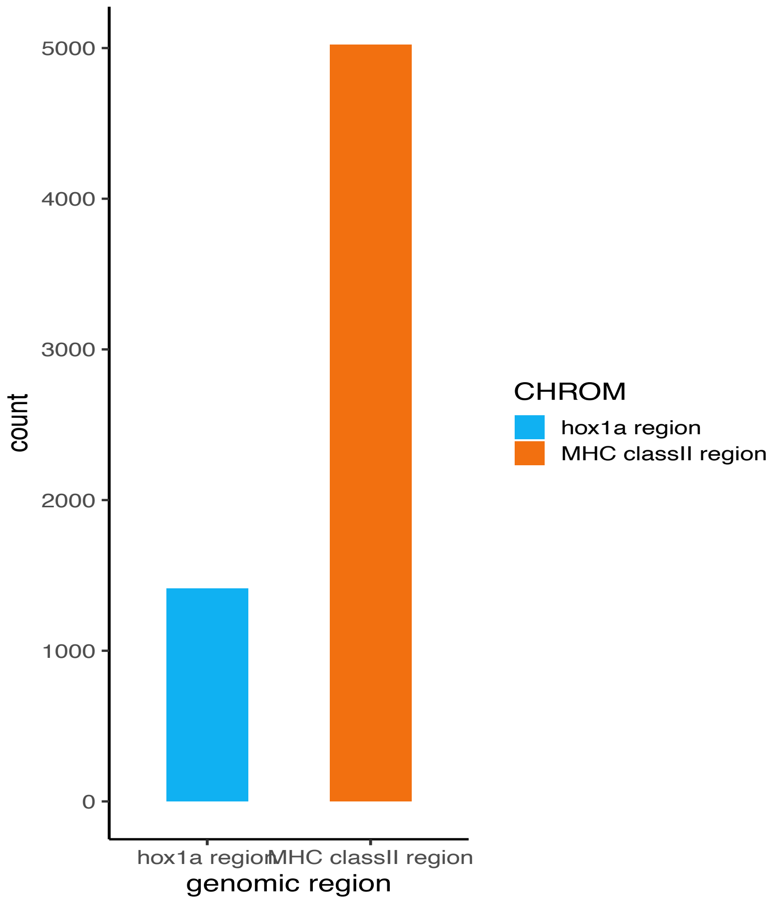
Graph depth shows single-copy regions surrounded by complex VNTRs

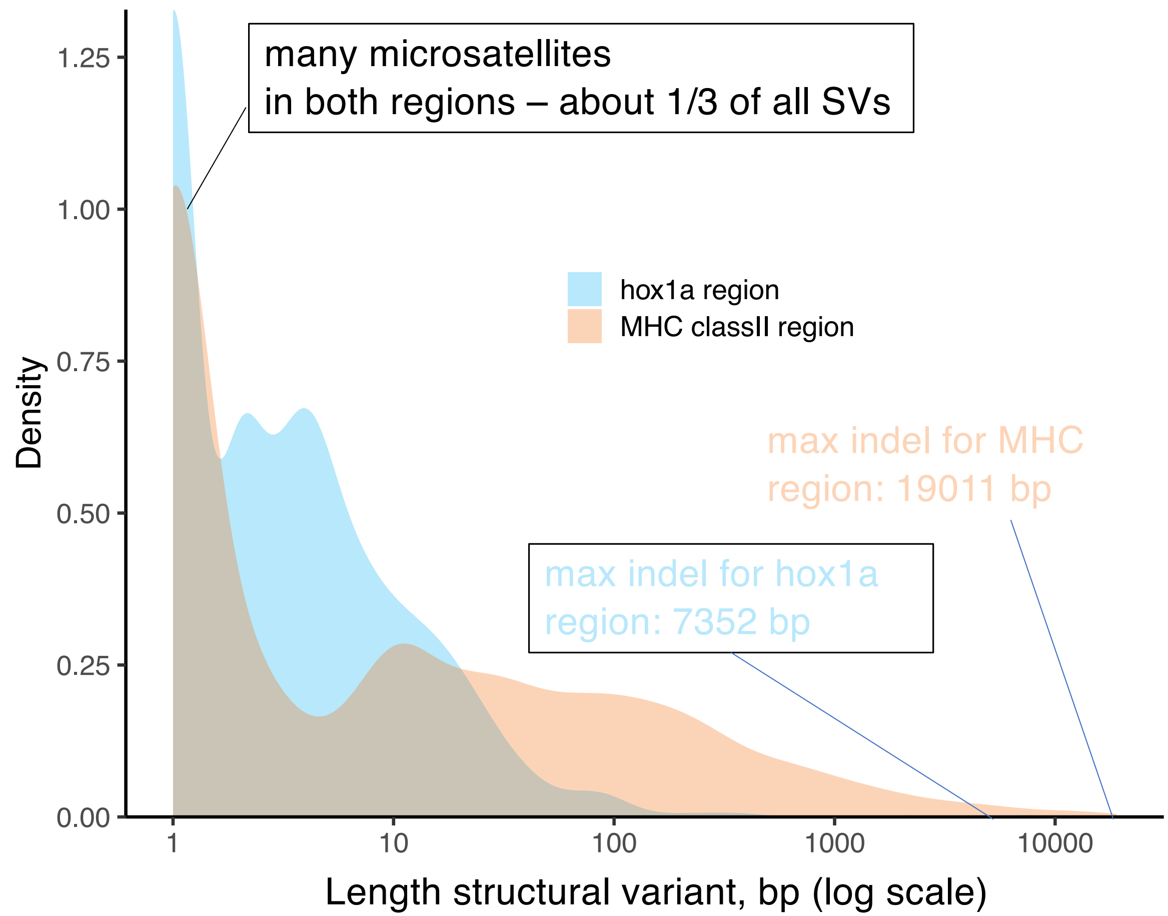


Structural variation in another dynamic genic region

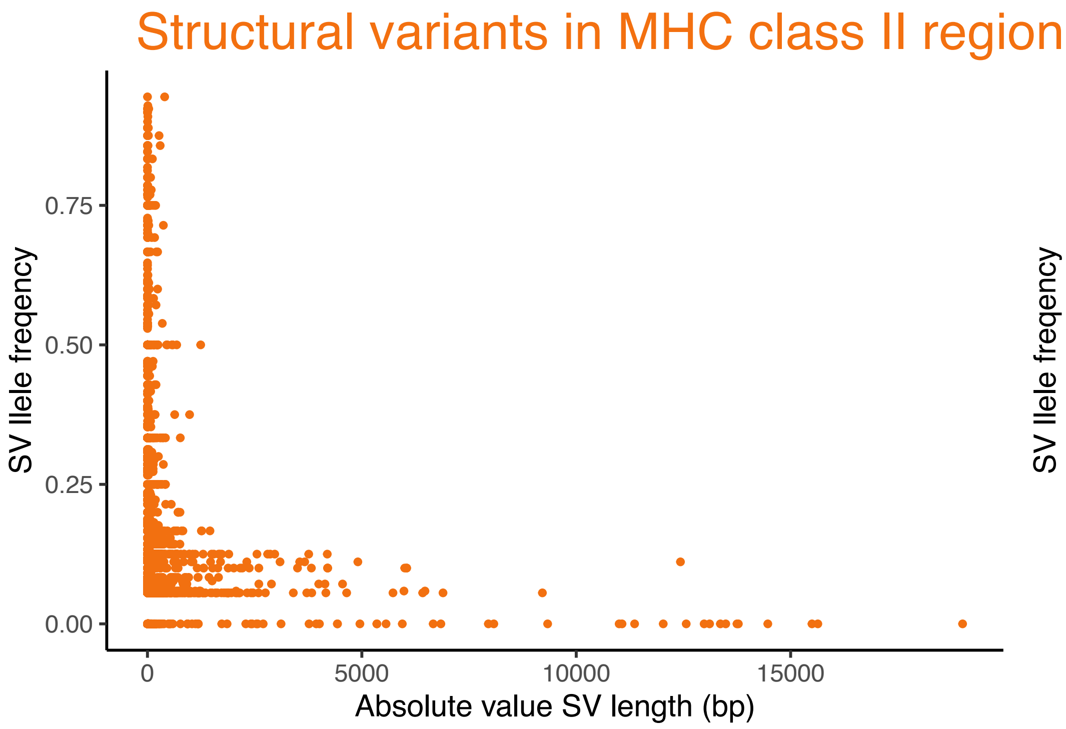


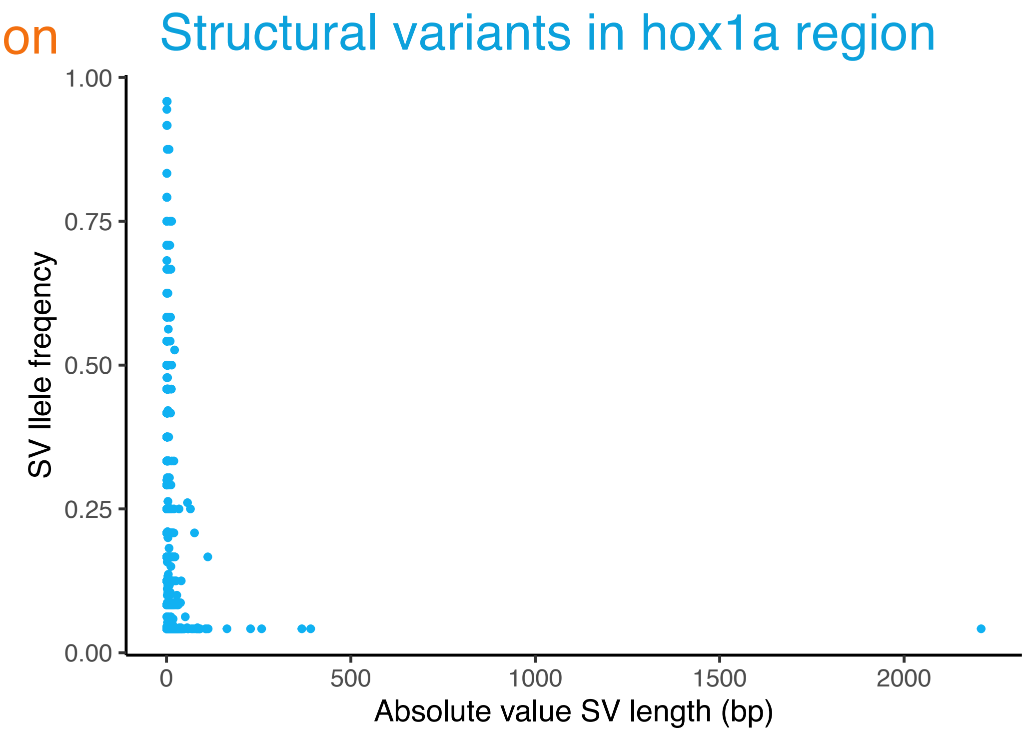
MHC region has more numerous and longer structural variants than hox1a region



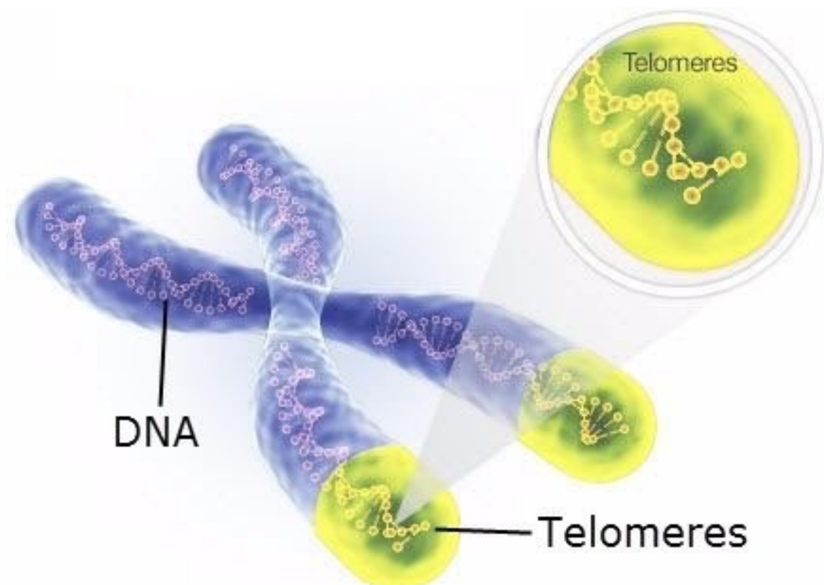


Low frequency of large SVs in both MHC and hox1a regions (~400-kb)





Telomeres – barometers of age and stress in birds

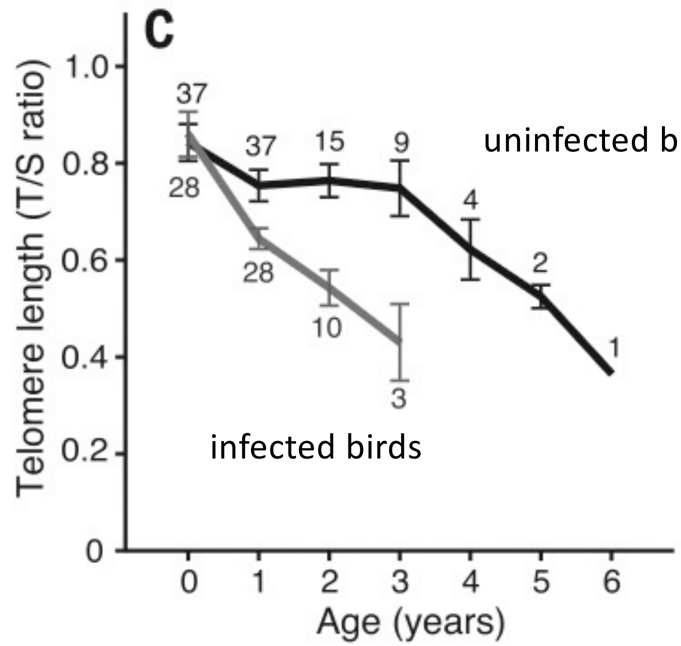


https://medibalans.com/telomere/

RESEARCH | *REPORTS*

CHRONIC INFECTION

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

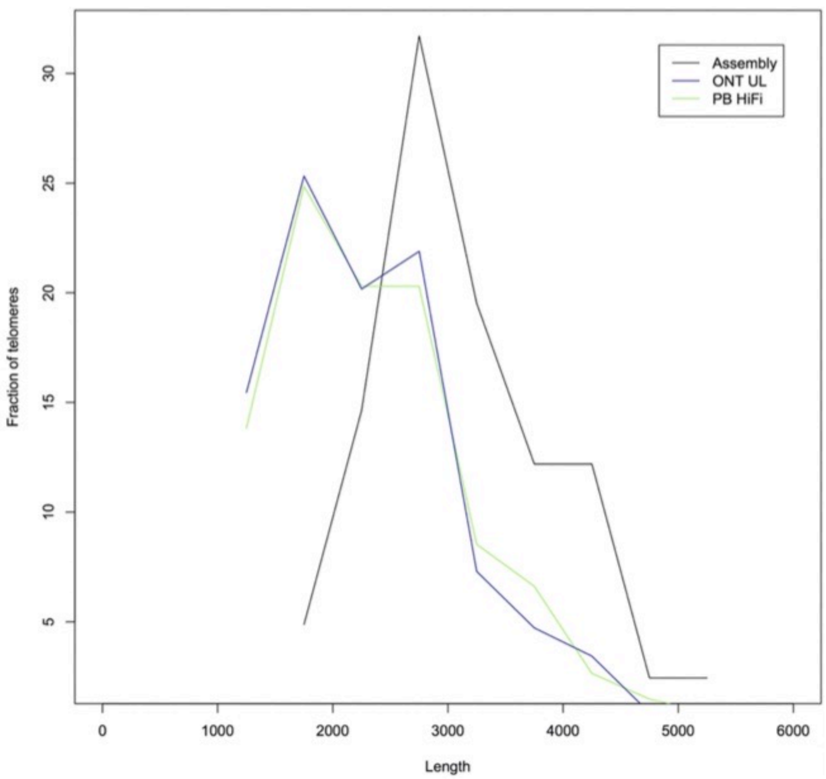


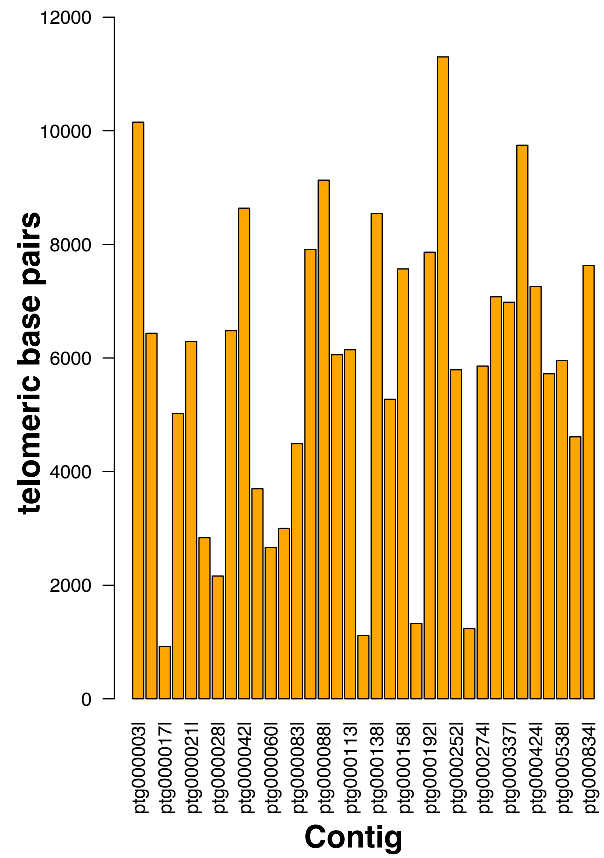
Ashgar et al. 2015. Science 347:436-438

CTCAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGG
GTTAGGGTTAGGGTTAGGGGGTTAGGGTTAGGGGGTTAGGGTTAGGGTTAGGGTTAGGGT
TAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGT
TAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGT
TAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTGGGTTAGGGCTAGGGTT
AGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTT
AGGGTTAGGGTTAGGGTTAGGGTTAGGGGGTTAGGGTTGGGTTAGGGTTAGGGTTAGGGT
TAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTGGTTAGGGTTAGGGTTAGGGTTAG
GGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAG
GGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAG
GGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTAGG
GTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGG
GTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGG
GTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTTGGGTTAGGGTTAGGGTTAGG
GTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGG
GTTAGGGTTAGGGTTTGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGG
GTTAGGGTTAGGGTTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAG
GGTTAGGGTTAGGGTTAGGGTTAGGGTTTGGGTTAGGGGGTTAGGGTTAGGGTTTGGGTT
AGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGGTTAGGGTTAGGGTTAGGGT
TAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTGGGTTAGGGTTAGGGGT
TAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGGTTAGGTTAGGGTTAGGGT
TGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTTGGGTTAGGGT
TAGGGTTAGGGTTTGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGT
TAGGGTTAGGGTTAGGGTTTGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGT
TAGGGTTAGGGTTAGGGTTGGGTTAGGGGTTGGGTTAGGGTTAGGGTTAGGGGTTAGGGT
TAGGGTTAGGGTTTGGGTTAGGGTTAGGGGTAGGGTTAGGGTTAGGGTTTGGGTTAGGGT
TAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGGTTAGGGTTAGGG
TTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGG
TTGGGTTAGGGTTAGGGTTGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTT
AGGGTTAGGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTGGGGTTAGGGT

GGGGTTAGGGTTAGGGTTAGGGTTAGGGTAGGGTTAGGGTTAGGGTTAGGGGTTAGGGTT
AGGGTTAGGGTAGGGTTAGGGGTTAGGGTTAGGGTTAGGGGTTAGGGGTTAGGGTTAGGG
TTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGG
TTGGTTAGGGTTAGGGTTAGGGTTAGGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGT
TAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGGTTAGGGGTAGGGTTAGGG
TTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGG
TTAGGGTTGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGT
TAGGGTTAGGGTTAGGGTTGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTT
AGGGTTAGGGTTAGGGTTAGGGTTGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAG
GGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGGTTAGGGTTAGGGTTAGGGTTA
GGGTTAGGTTAGGGTTAGGGTTAGGGTTAGGGTTTGGGTTAGGGTTAGGGTTAGGGTTAG
GGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAG
GGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAG
GGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAG
GGTTAGGGTTGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGG
GTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTTGG
GTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGGTTAGGGTTAGGGTTAG
GGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAG
GGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAG
GGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAG
GGTTAGGGTTAGGGTTAGGGTTAGGGTTTGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAG
GGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTGGGTTAGGGTTAGGGTTAGG
GGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAG
GGTTAGGGTTAGGGTTAGGGTTTGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAG
GGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAG
GGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTGGGTTAGGGTTAGGGTTAGG
GTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTTGG
GGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTAGGGTTAGGGTTAGG
GTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGG
GTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGG

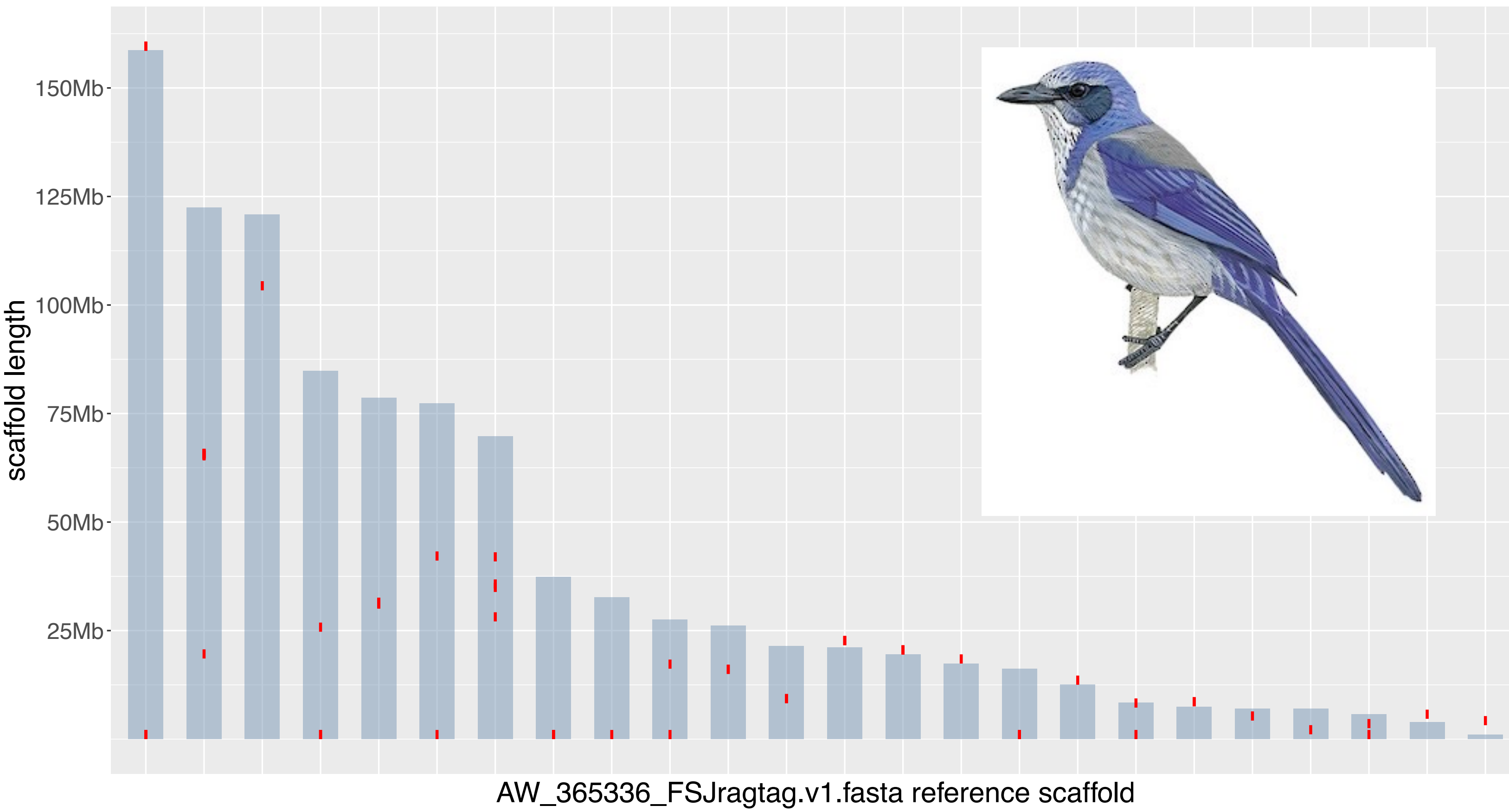
Scrub jay telomeres are usually ~3-10 kb long



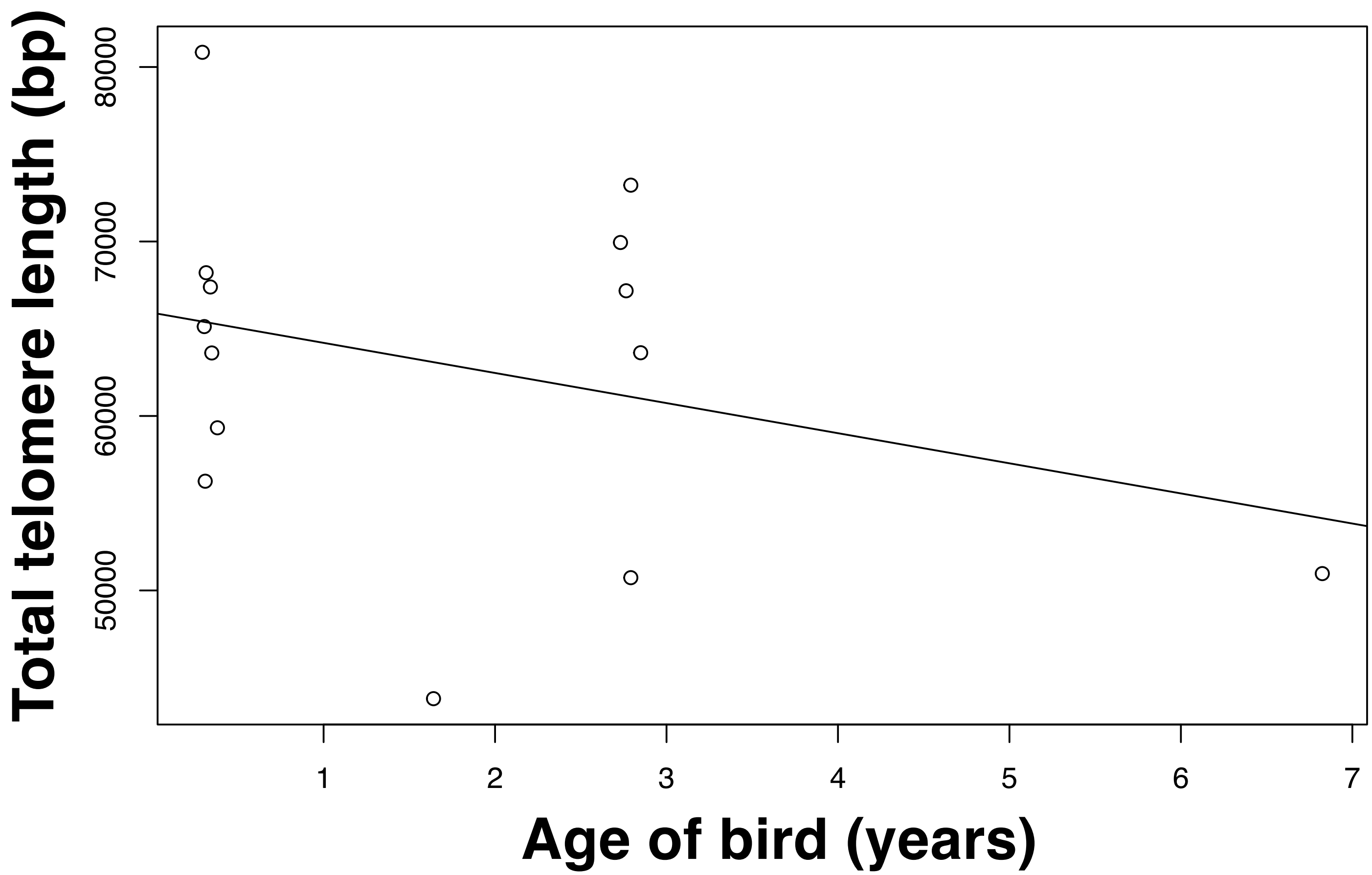


Miga et al. 2020. Nature 485:79-87.

Telomere sequences are generally found at chromosome ends



Telomere abundance declines with age in Florida birds



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

Pangenome informatics
Erik Garrison
Andrea Guarracino

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

Fieldwork
Greg and Donna