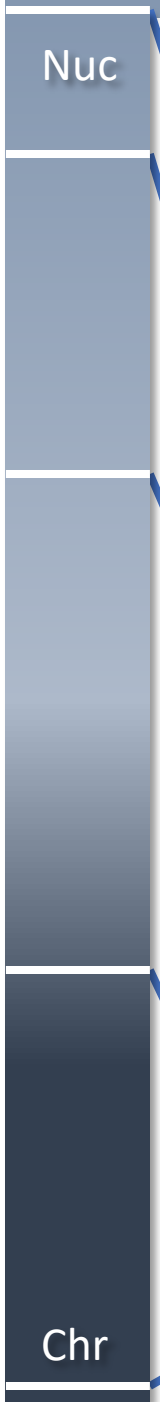
Genotyping complex structural variants using a Chicken Pangenome reference

Presentation by Derek M Bickhart, Ed. S. Rice, Laurent A.F. Frantz, Wes C. Warren

Original concept and manuscript: Rice et al. 2023. BMC Biology. https://doi.org/10.1186/s12915-023-01758-0



Genetic variants ordered by scale

A sense of scale (and individual frequency)

- SNP -- Single nucleotide polymorphisms (~ 5,000,000 on average)
- INDEL – Insertions/Deletions (~600,000 on average)
- Mobile Elements – SINE, LINE Transposition (???)
- Genomic structural variation (25,000 on average (?))
  - Large-scale Insertions/Deletions [Copy Number Variation: CNV]
  - Segmental Duplications
  - Inversions, Translocations, Fusions.

Animal phenotypes are caused by structural variants (SV)

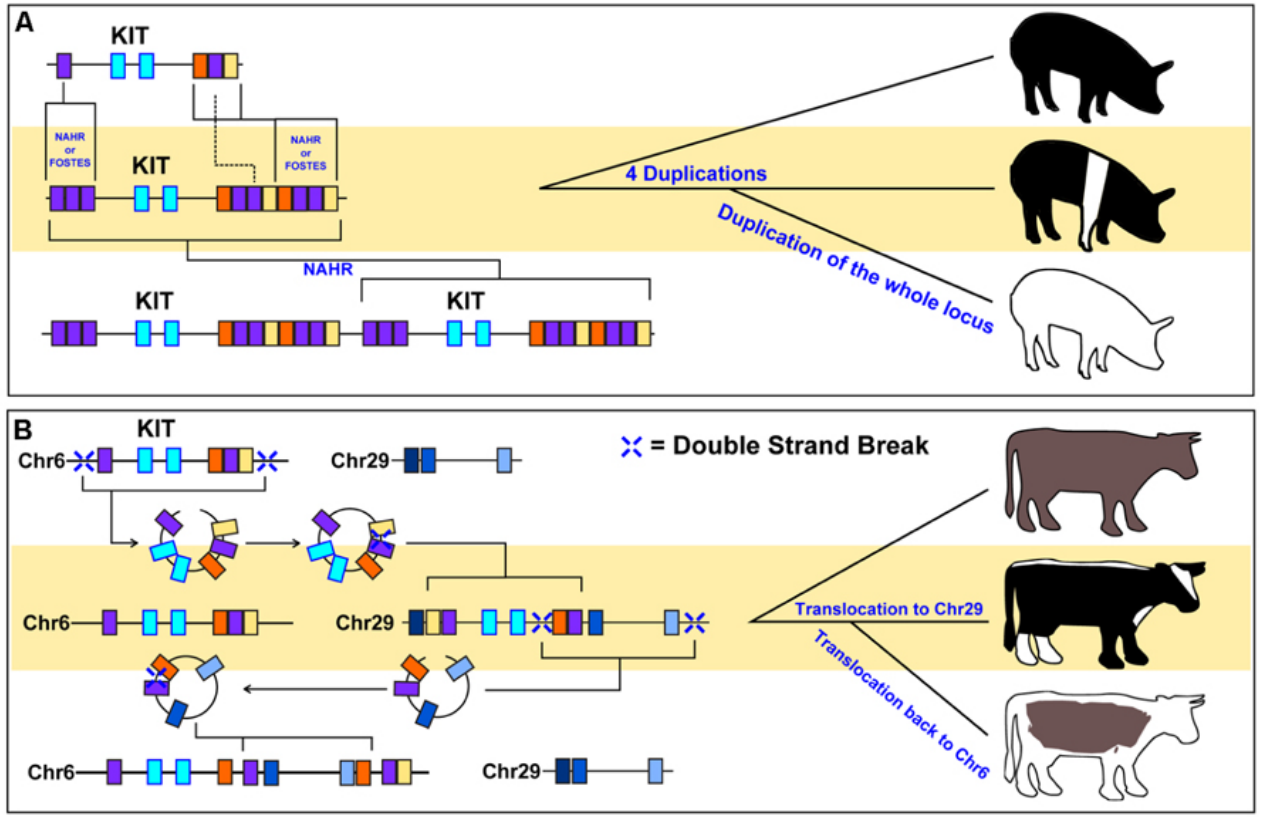
Known positive phenotypes caused by SV

- Top: Belted-pig, and dominant white
- Bottom: Color-sidedness in cattle

Deletions tend to be very harmful

- HH5 deletion causes embryonic death in cattle
- Deletion of the *fanci* gene causes early death in cattle

Many more examples we have not uncovered



Problem: Structural variant detection is prone to errors

- DNA sequence data is still the best detection method
- Many landmark studies acknowledge issues with false discovery rate (FDR)
  - Short-read SV and CNV studies: ~15-30% FDR
  - Long-read SV and CNV studies: ~1-11% FDR
- What are the sources of errors?
  - Unknown structurally variant regions
  - Non-reference, novel DNA sequence
  - Repetitive DNA regions

Problem: the best solution to detect SVs is to use longer reads

Longer reads resolve errors

Span repetitive regions

Can contain many smaller SVs (complexity)

Could resolve inversions

Longer reads are more costly

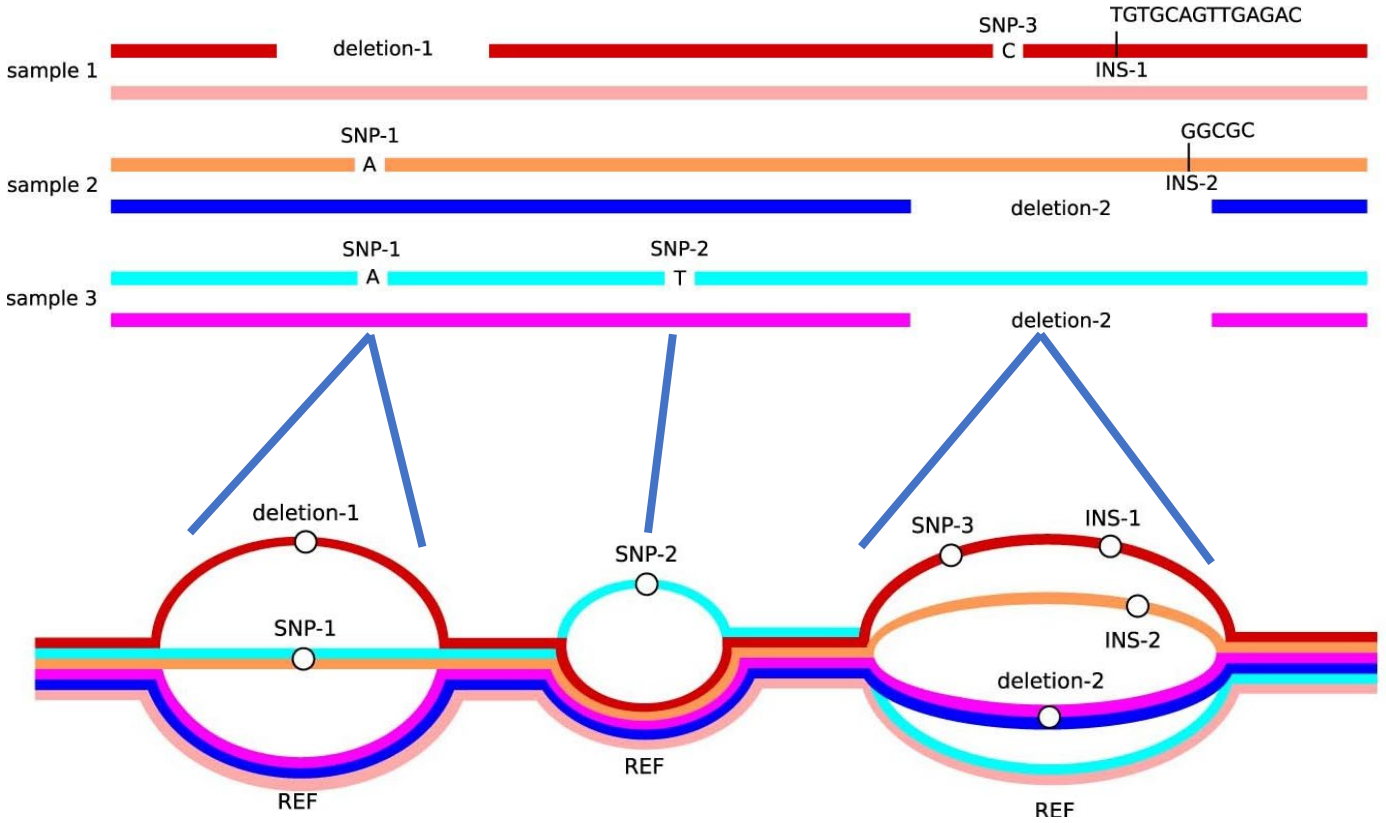
Approximately 20 to 30 euros per Gbp

Short reads can be 5 euros per Gbp

About half an order of magnitude cheaper!

Can we improve the accuracy of SV detection with short-read sequencing?

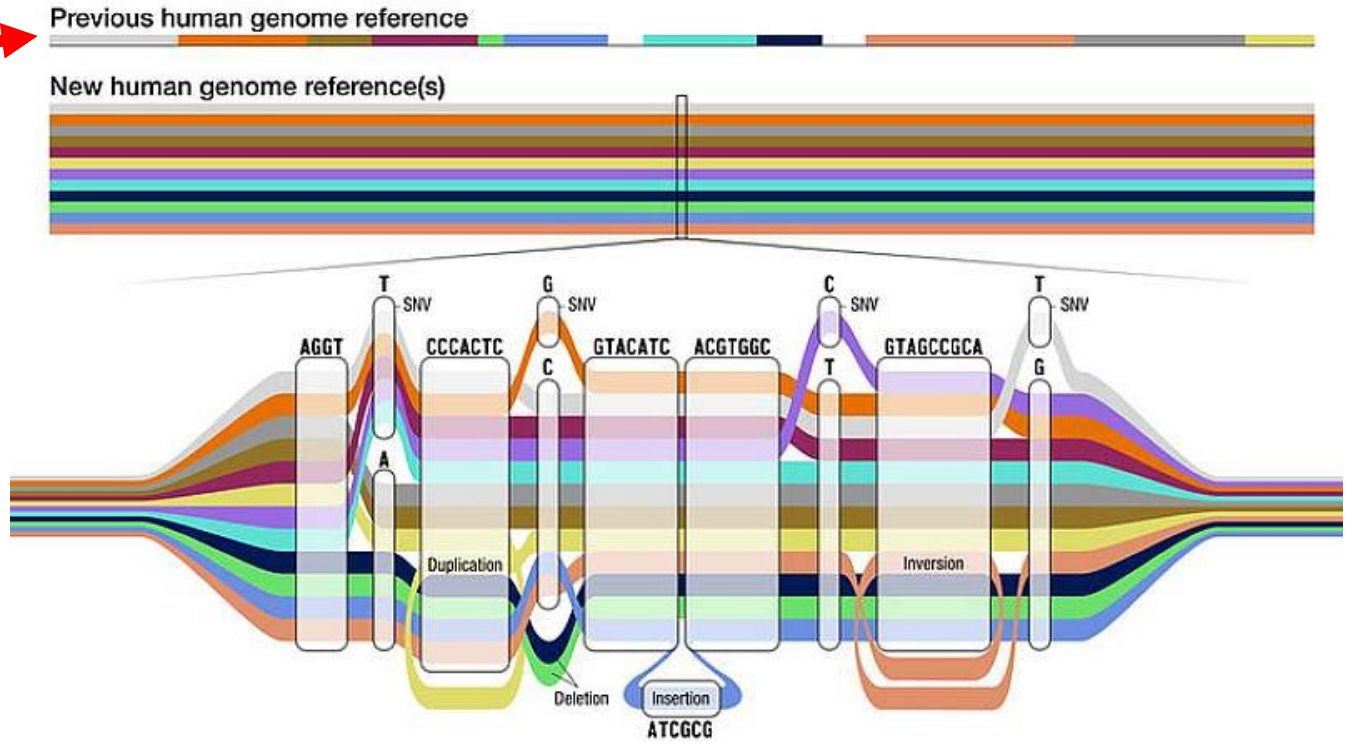
- If variant sites are known, genotyping is easier
  - Known problem regions for realignment
  - Remapping reads to correct locations
- Several “variant-aware” mappers available for short reads
- Hypothesis: better mapping accuracy for short reads will improve SV detection



From Ebler et al. 2022. Nat Gen.

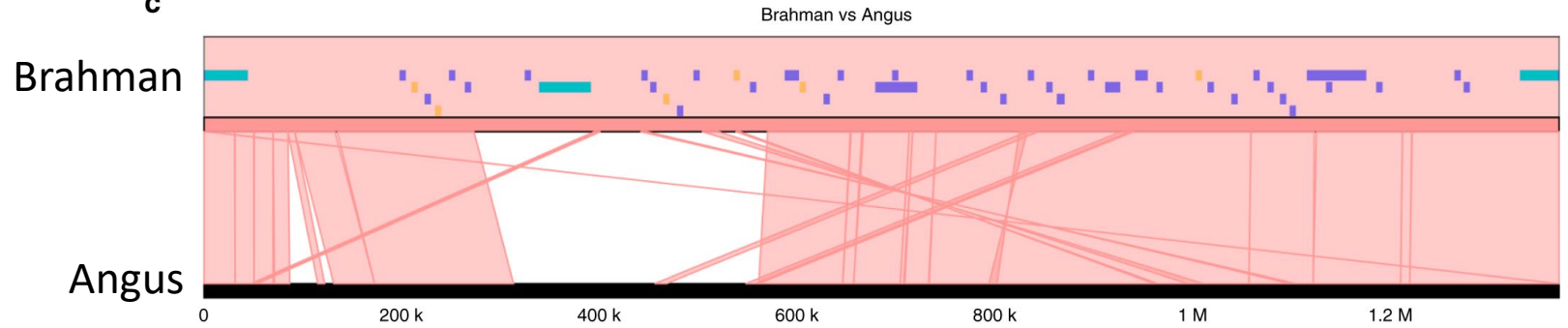
Pangenome resources: using graphs to encode variation

- Pangenome: Originated in field of Microbiology
- Linear references only represent one allelic state
- A Pangenome (graph format) can represent many alleles
- Individual variation is represented by traversing the graph



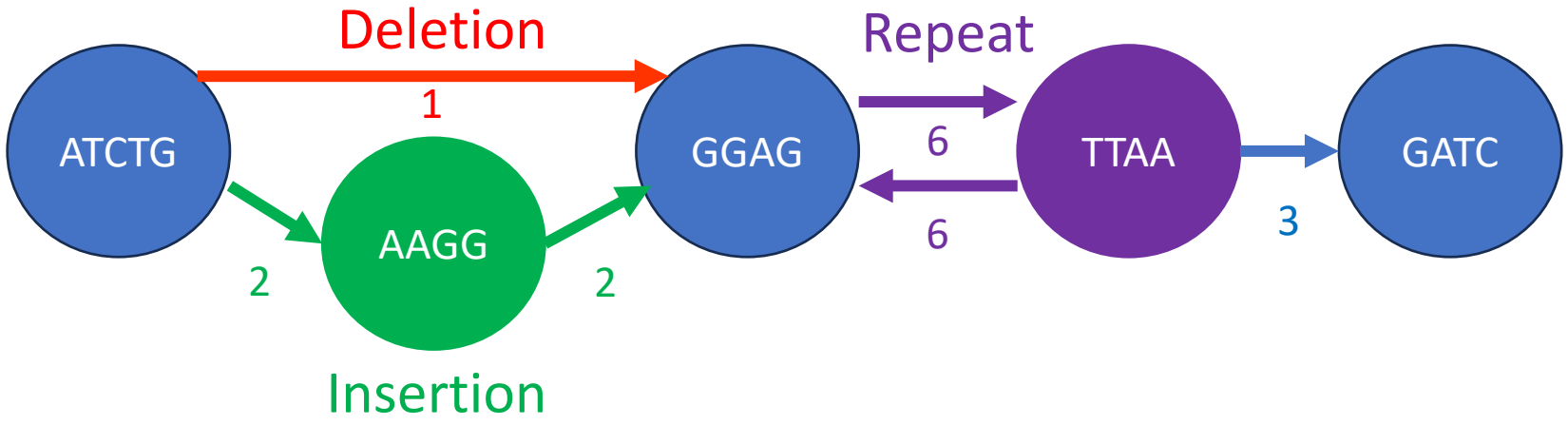
Linear references compared to graphs

- Conflicts and coordinates among versions – makes it difficult to interpret



From Low et al. 2020. Nat comm.

- Pangenome Graph: A data structure that represents DNA sequence as "Nodes" and the connections between them as "Edges"

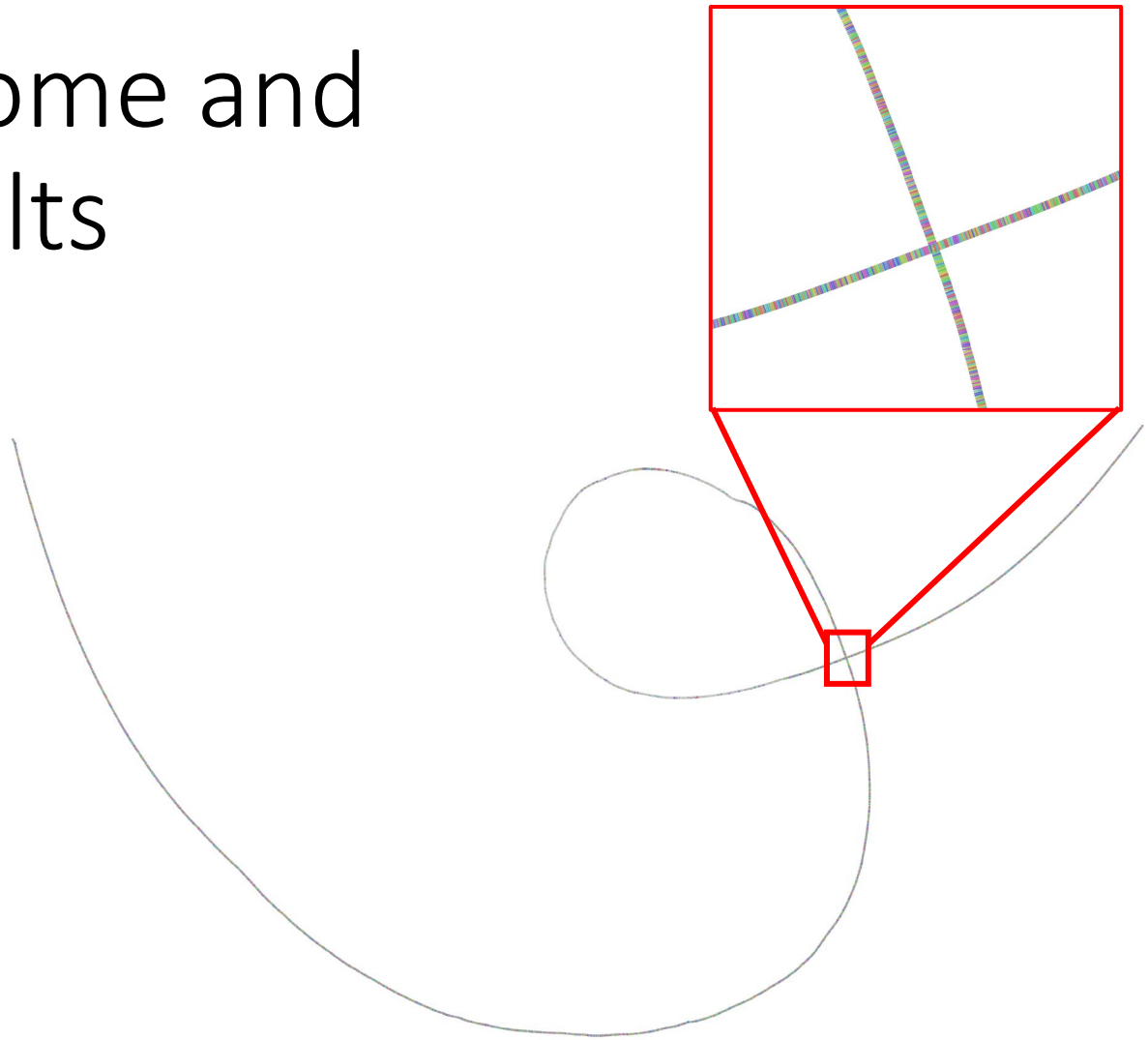


Visualizing the pangenome and understanding the results

- Pangenome graphs are less human-readable

- Visualizations of graph structure
  - Bandage
  - ODGI

- Surjection
  - Decomposing graph structure to fit a linear framework
  - Strategy used by VG
  - Used here to compare to linear methods



Subgraph of K locus
chrZ: 11159196-11400464

Comprehensive chicken pangenome resource

Research article | Open access | Published: 22 November 2023

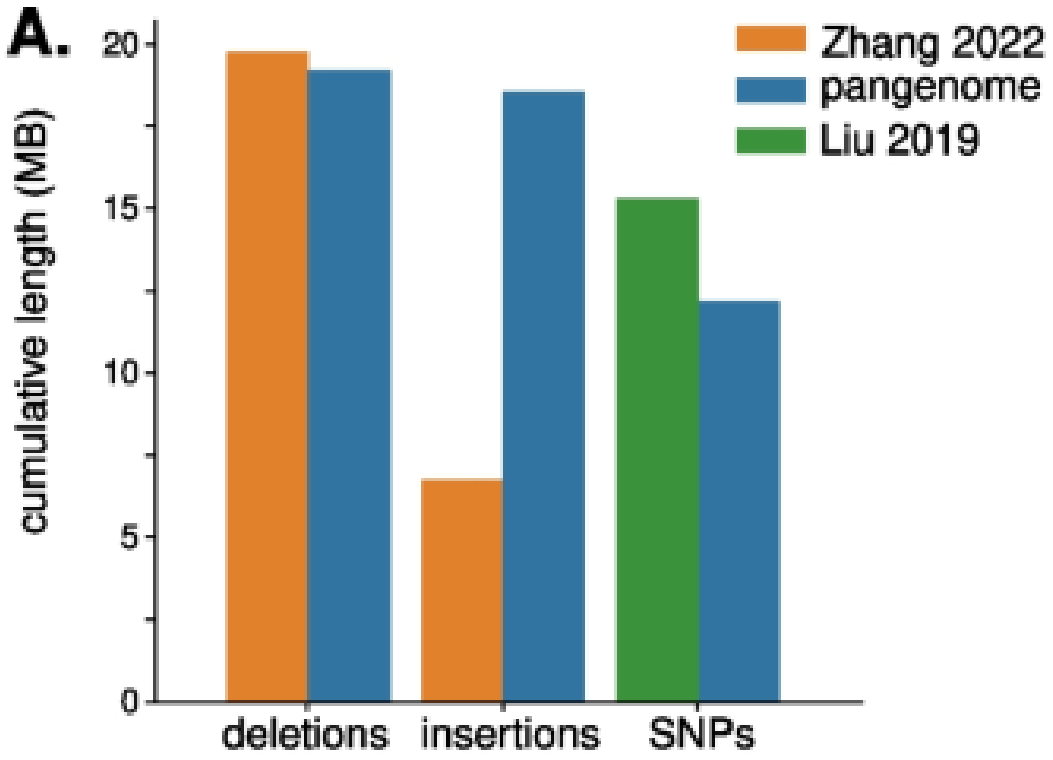
A pangenome graph reference of 30 chicken genomes allows genotyping of large and complex structural variants

Edward S. Rice, Antton Alberdi, James Alfieri, Giridhar Athrey, Jennifer R. Balacco, Philippe Bardou, Heath Blackmon, Mathieu Charles, Hans H. Cheng, Olivier Fedrigo, Steven R. Fiddaman, Giulio Formenti, Laurent A. F. Frantz, M. Thomas P. Gilbert, Cari J. Hearn, Erich D. Jarvis, Christophe Klopp, Sofia Marcos, Andrew S. Mason, Deborah Velez-Irizarry, Luohao Xu & Wesley C. Warren

BMC Biology **21**, Article number: 267 (2023) | Cite this article

- Assembly-based Pangenome Graph
- Composed of 30 chicken assemblies
  - T2T chicken assembly
  - Broiler chicken assemblies
  - More Broiler than Layer
- Rice et al. 2023. BMC Biology

The Pangenome Composition compares favorably with other variation maps



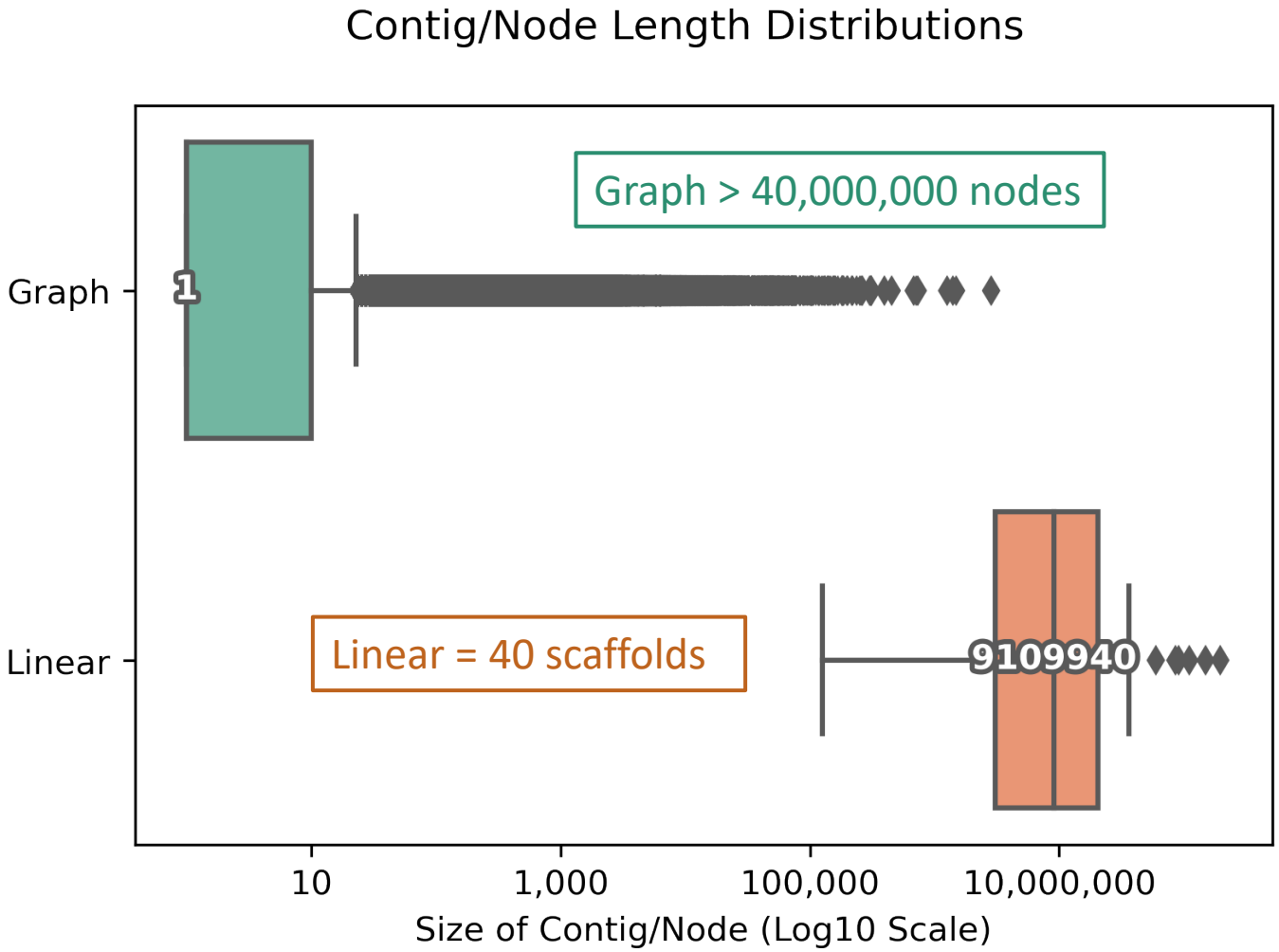
Comparisons to other Chicken studies

Zhang 2022: long-read SV survey

Liu 2019: Chinese chicken SNP survey

Excellent representation of diversity and variation maps

Chicken pangenome graph properties visualized

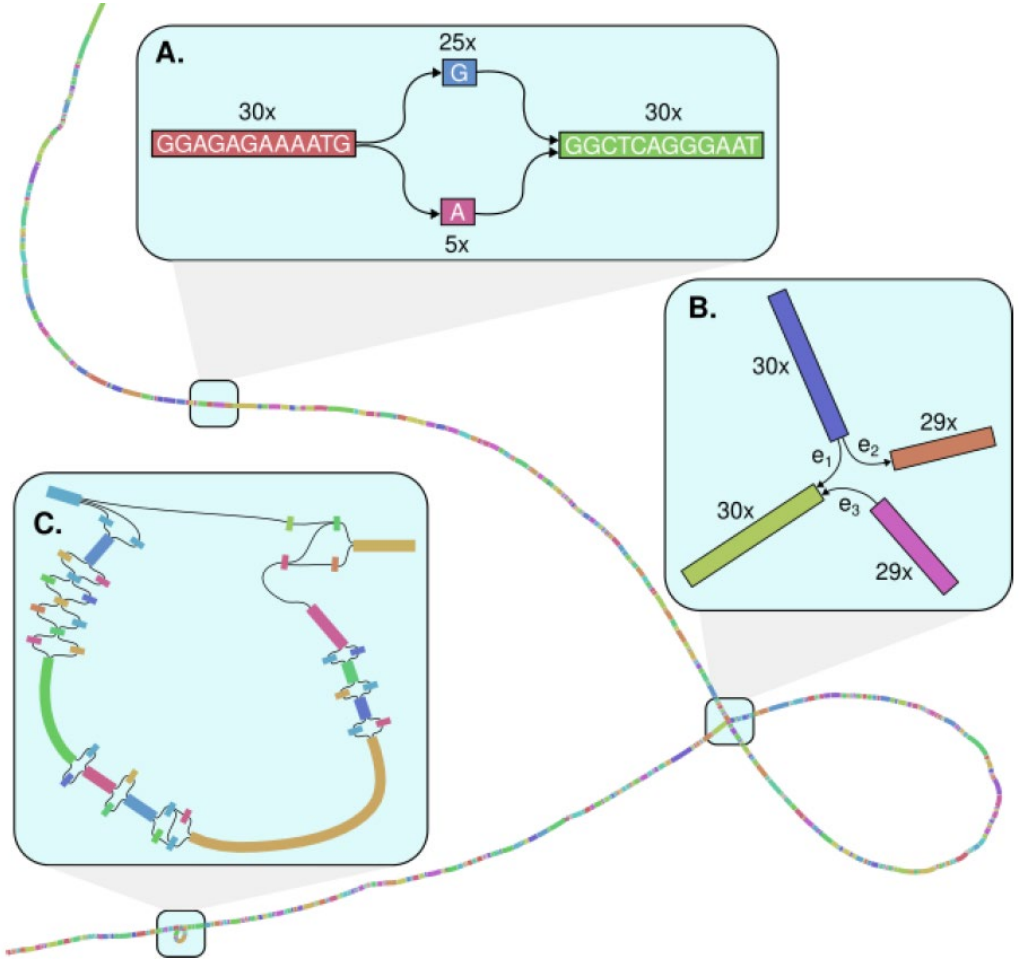


- Pangenome complexity is difficult to visualize
- Sheer size impossible to load in viewing tools
- Pangenome nodes:
  - Majority are SNPs among assemblies
  - Linear reference (bGalGal1b) is mostly contained in chromosome scaffolds

Chicken pangenome graph representation

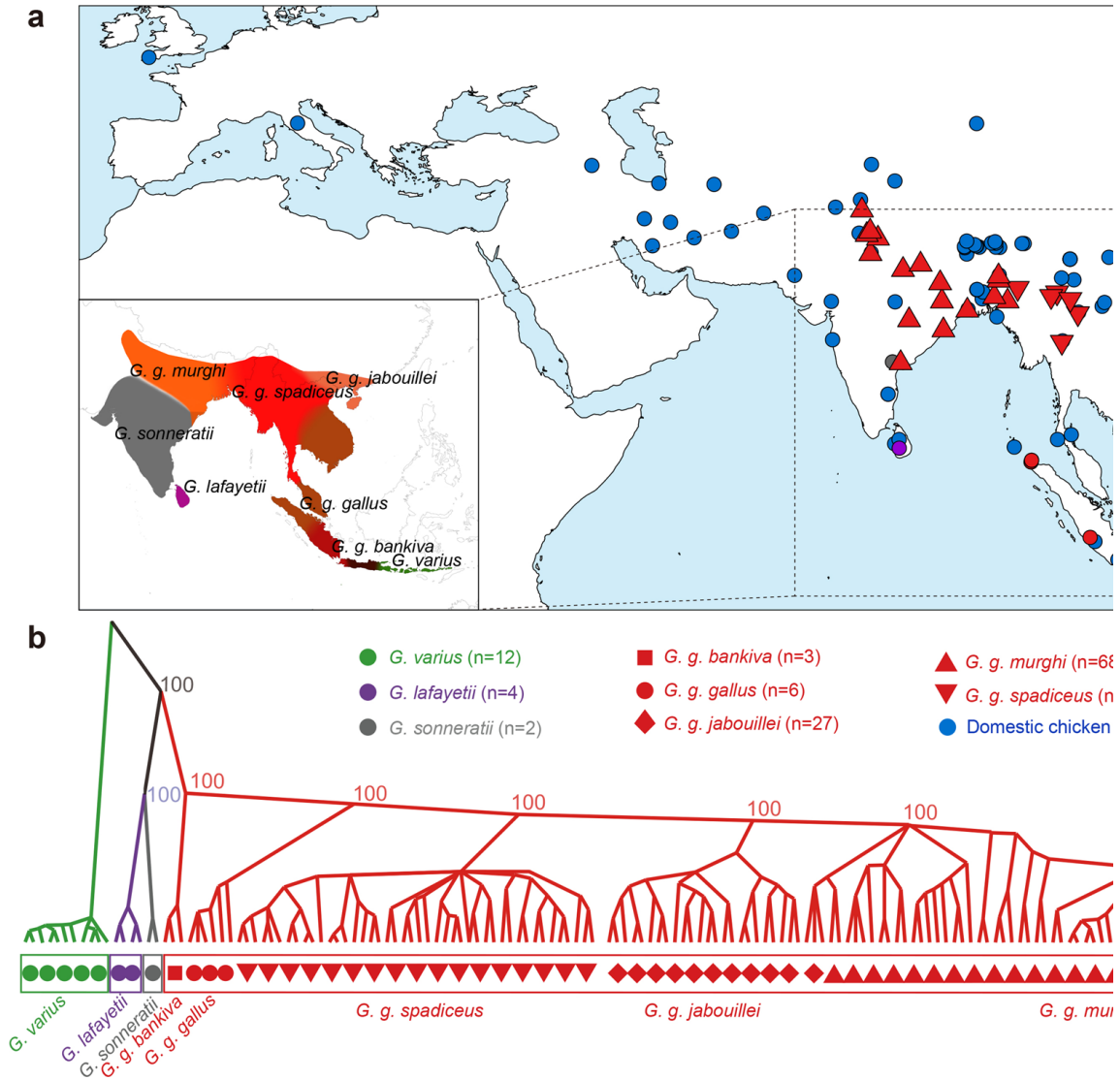
- Complex structural variants can be visualized and incorporated

- Structure of the *IGLLI* gene
  - Haplotype with a deletion (B) missing from reference
  - Structural variant and allele structure is more complex (C)



Population survey of 863 chicken WGS datasets

- Available on SRA and high quality
- Consist of several chicken (sub)species
  - *Gallus varius*
  - *Gallus gallus jabouillei*
  - *Gallus gallus bankiva*
  - *Gallus gallus murgha*
  - *Gallus gallus spadiceus*
- Selected first 98 samples from SRA list (> 10X coverage) for alignment

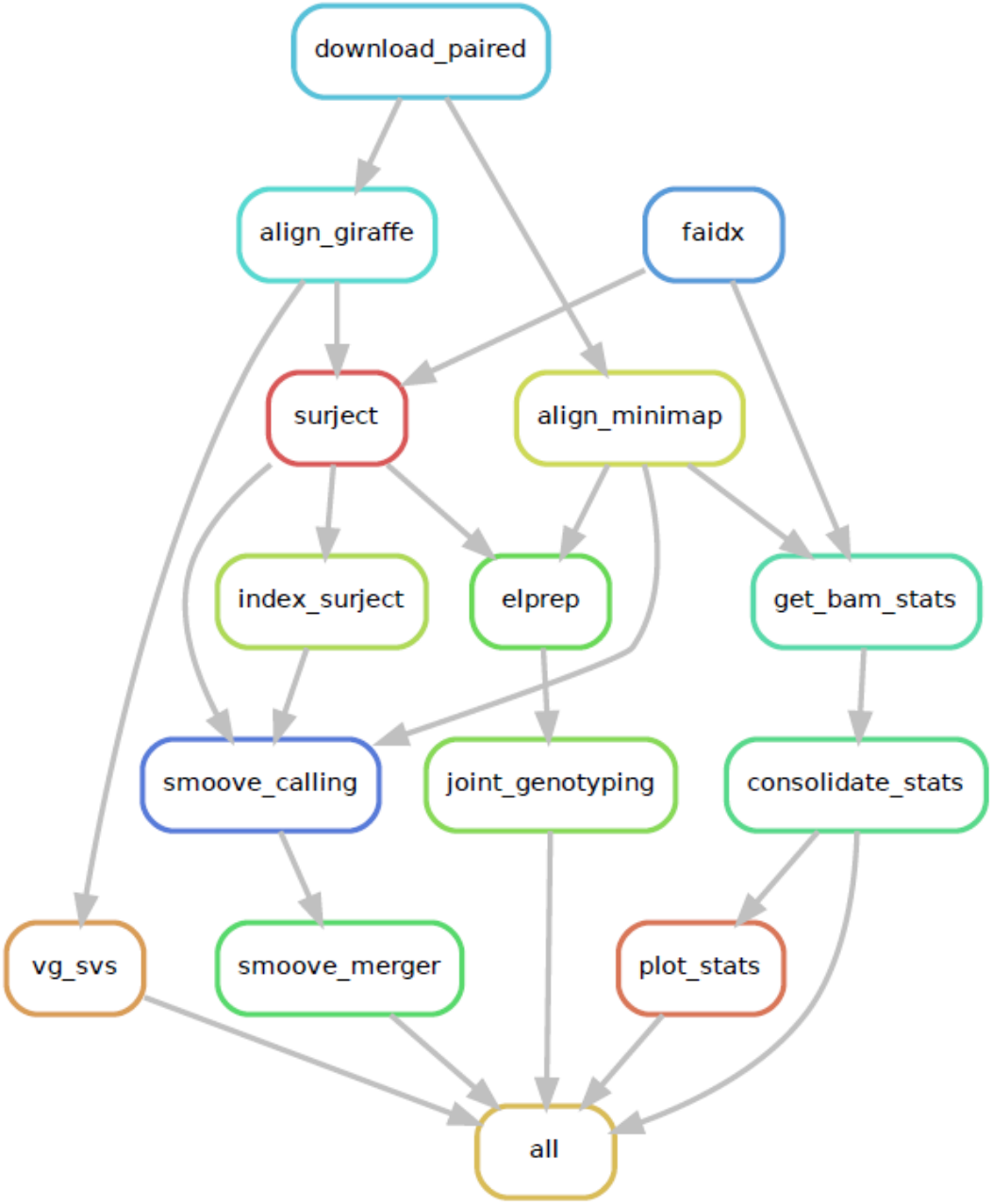


From Wang et al. 2020. Cell Research

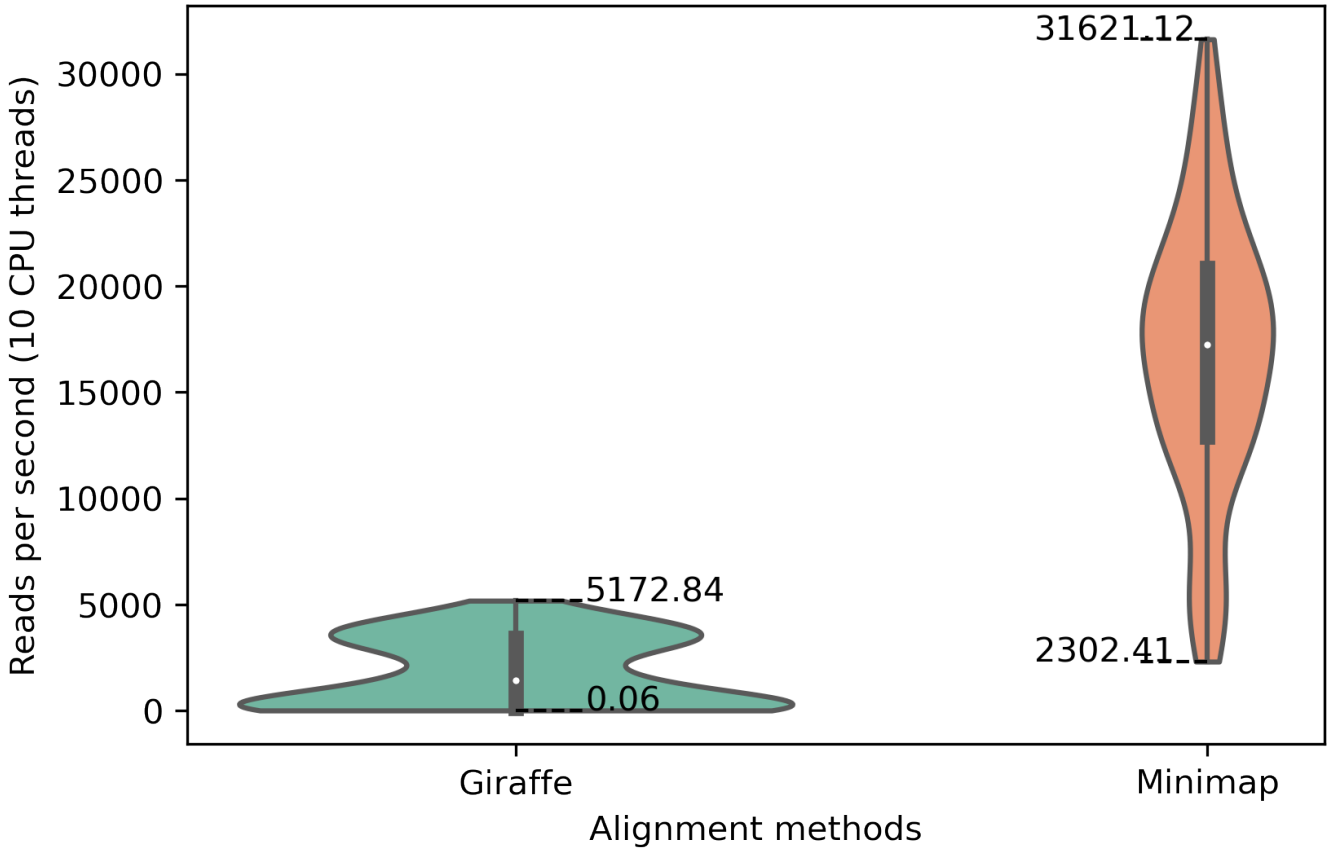
Methodology to discuss and replicate the results

- **The results of the study**
  - Aligned datasets: 98 chickens
  - Mapping comparisons between linear and graph
  - Known SV typing (K locus)
  - *De novo* SV calling

- **Methodology**
  - Snakemake workflow (Right)
  - Surjection to bGalGal1b coords
  - **Minimap** to bGalGal1b (linear)
  - **Giraffe** (vg) to pangenome (graph)



Graph alignment resource efficiency is still not fully optimized



Read mapping rates were higher for linear references

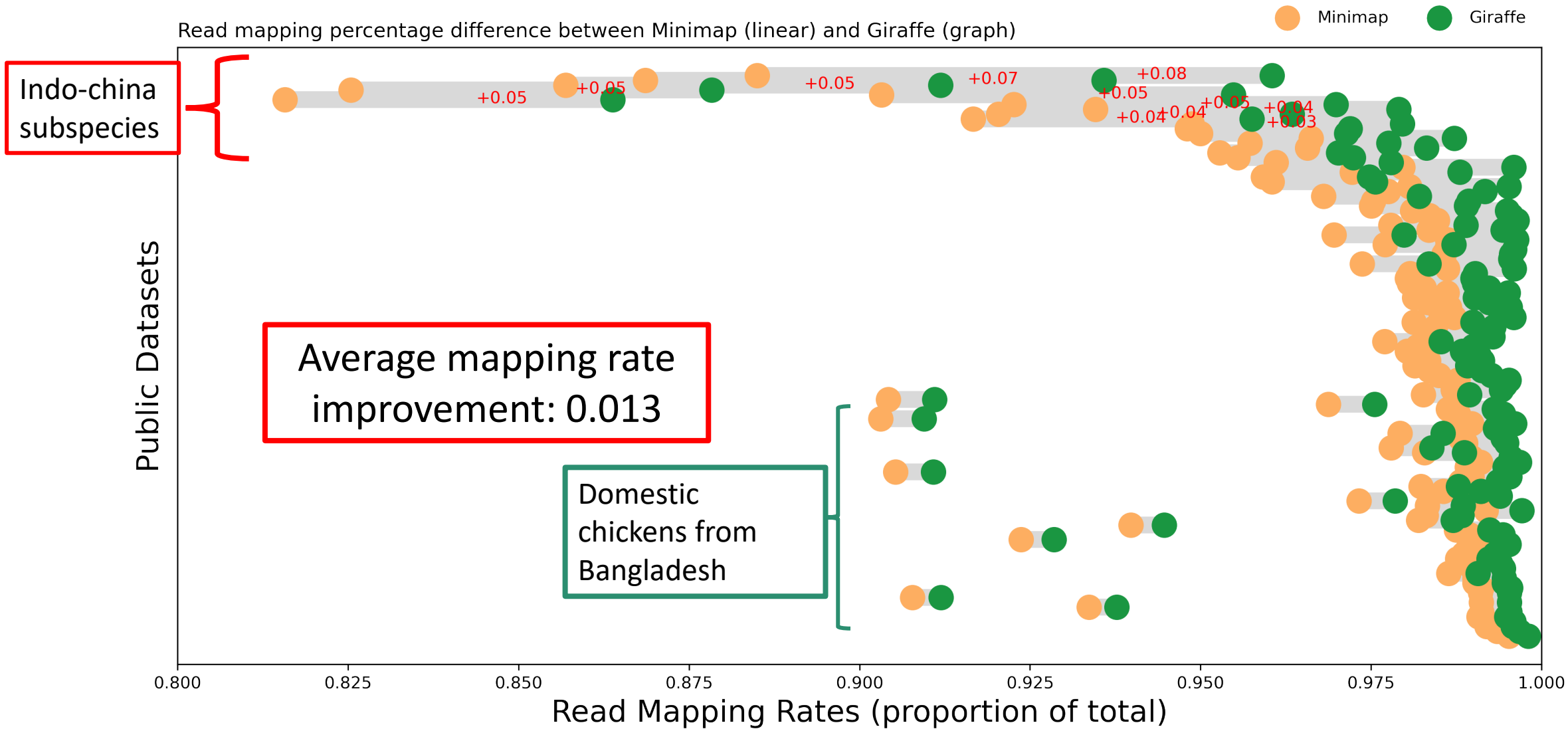
- On average: 6-fold faster
- Giraffe stalled with mapping reads to repetitive regions

Memory usage

- Giraffe: 26 Gb (avg)
- Minimap: ~9 Gb (avg)

Reason: maturity of tools for linear alignment

Mapping ratio improvements in every sample

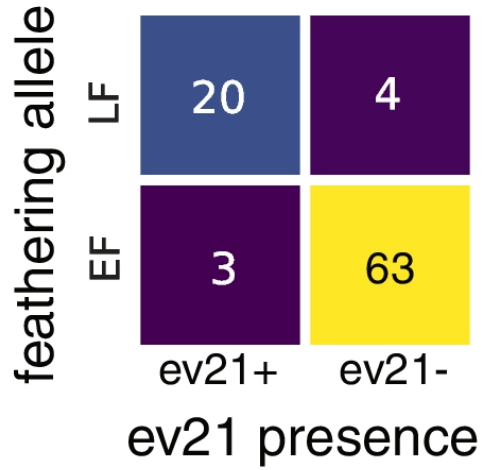


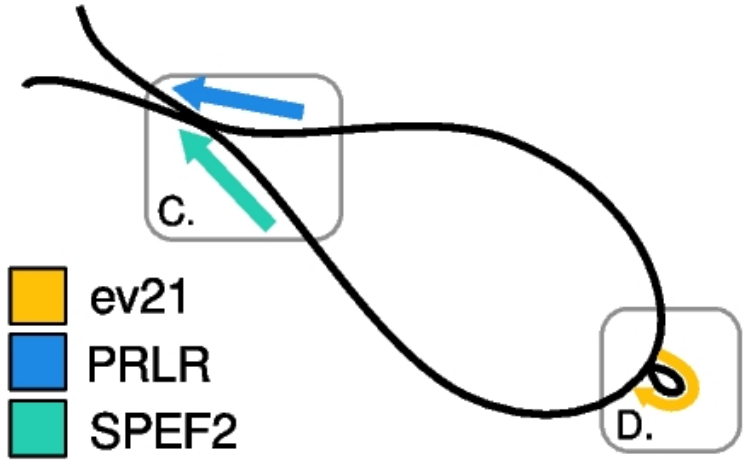
K locus EF and LF genotyping

- Two alleles:
  - Early Feathering (EF)
  - Late Feathering (LF)

- LF allele
  - Duplication of portion of SPEF2 and PRLR genes
  - Represented on graph

- Insertion of EV21 detected – independent of feathering allele





K locus genotyping from surjection

Carriers in the test dataset:

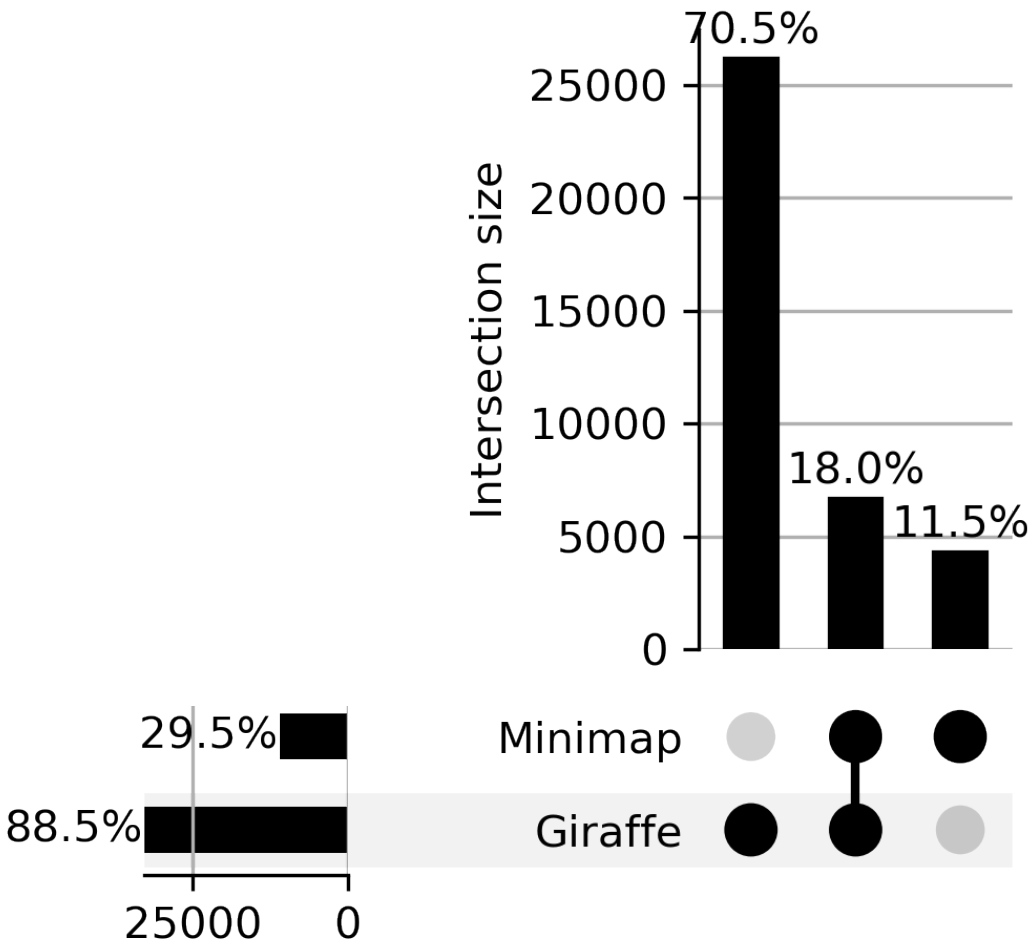
- 18 carriers for LF allele
- Read depth signal from CNVNATOR/JARMS
- 100% concordance Giraffe and Minimap mappings

Only one ev21 carrier (EF allele)

Graph based genotyping:

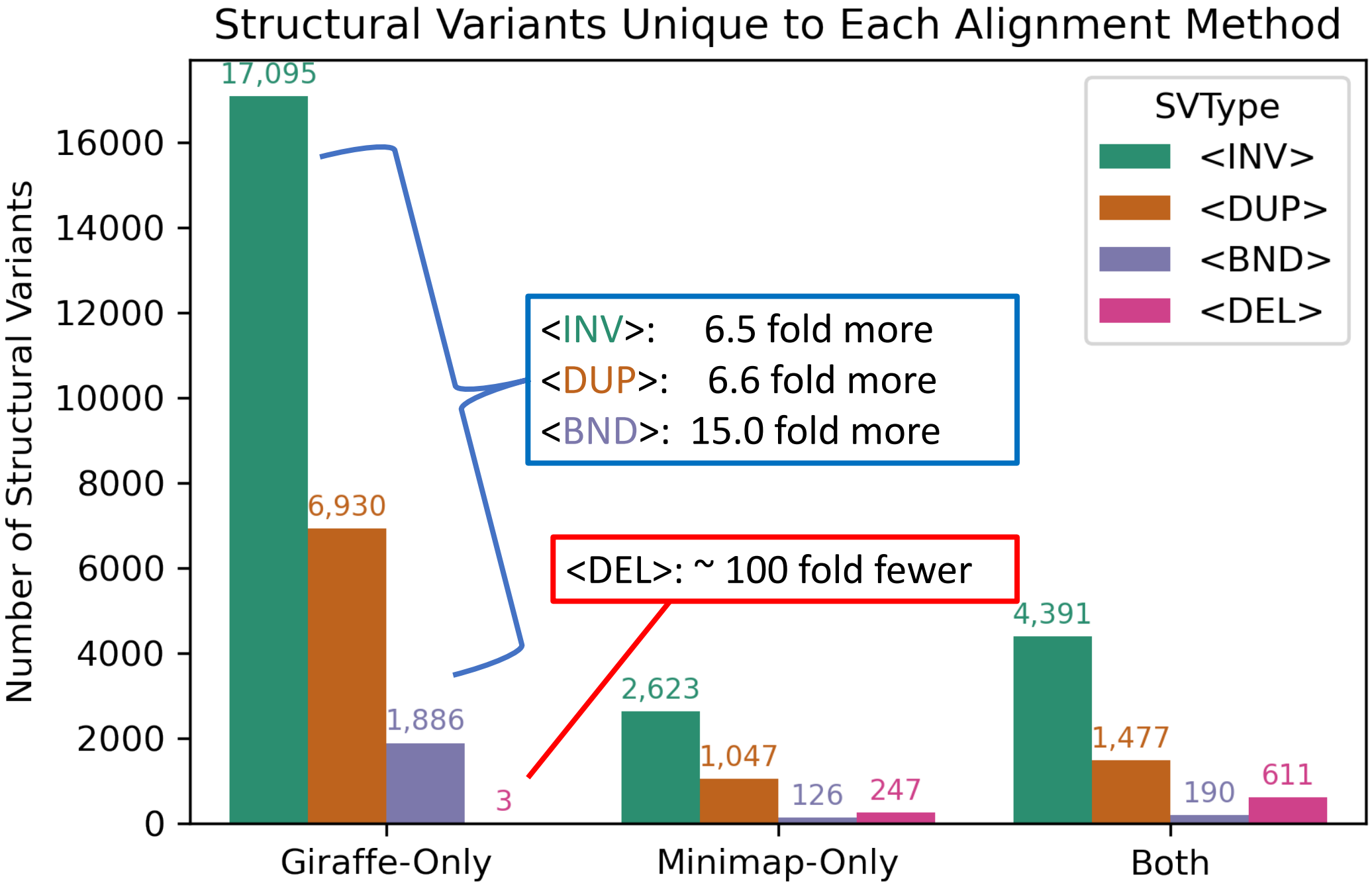
- Difficult to resolve the signal
- Read-depth CNV calling still difficult to incorporate

De novo SV calling from surjected graph results in more calls than a linear reference



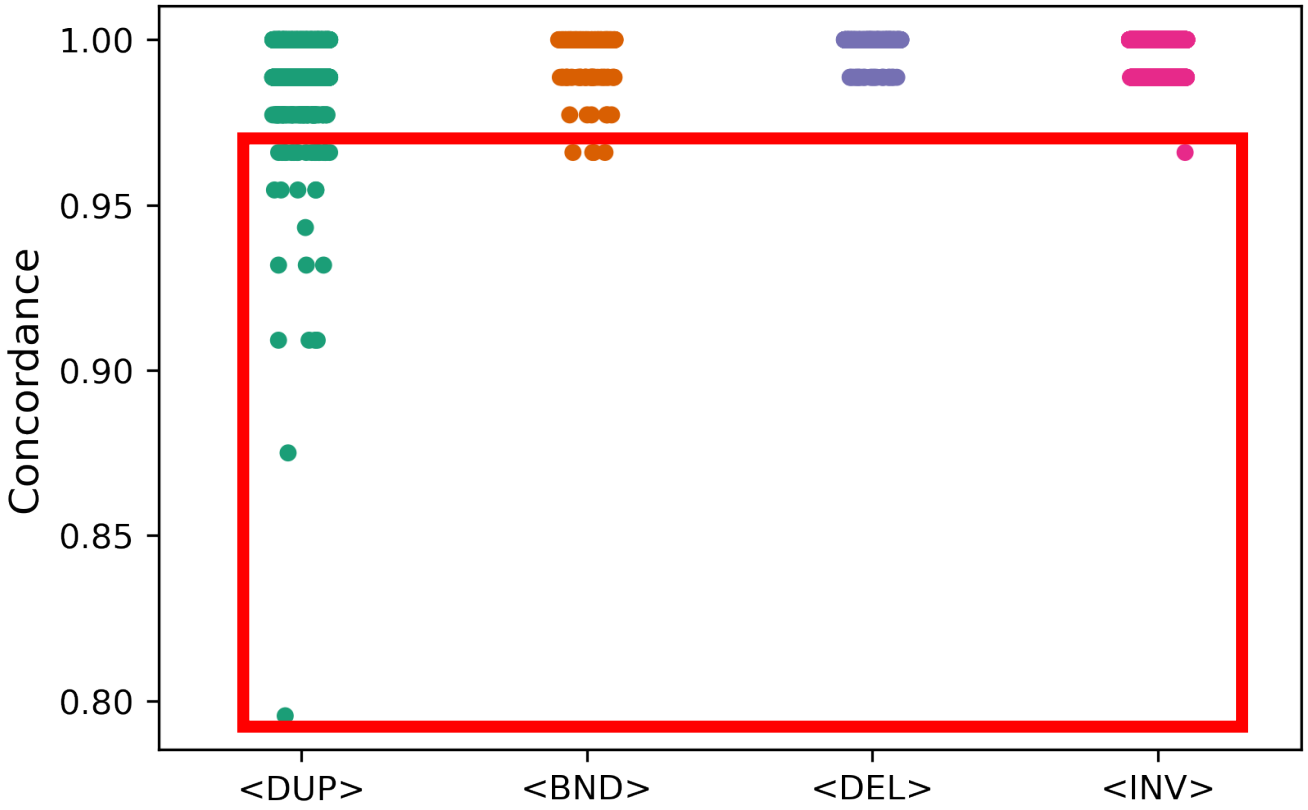
- Workflow:
  - Alignment to respective reference
  - (Giraffe-only) surjection to bGalGal1b coordinates
  - Smoove SV calling (Lumpy wrapper)
  - Combine compare (bcftools)

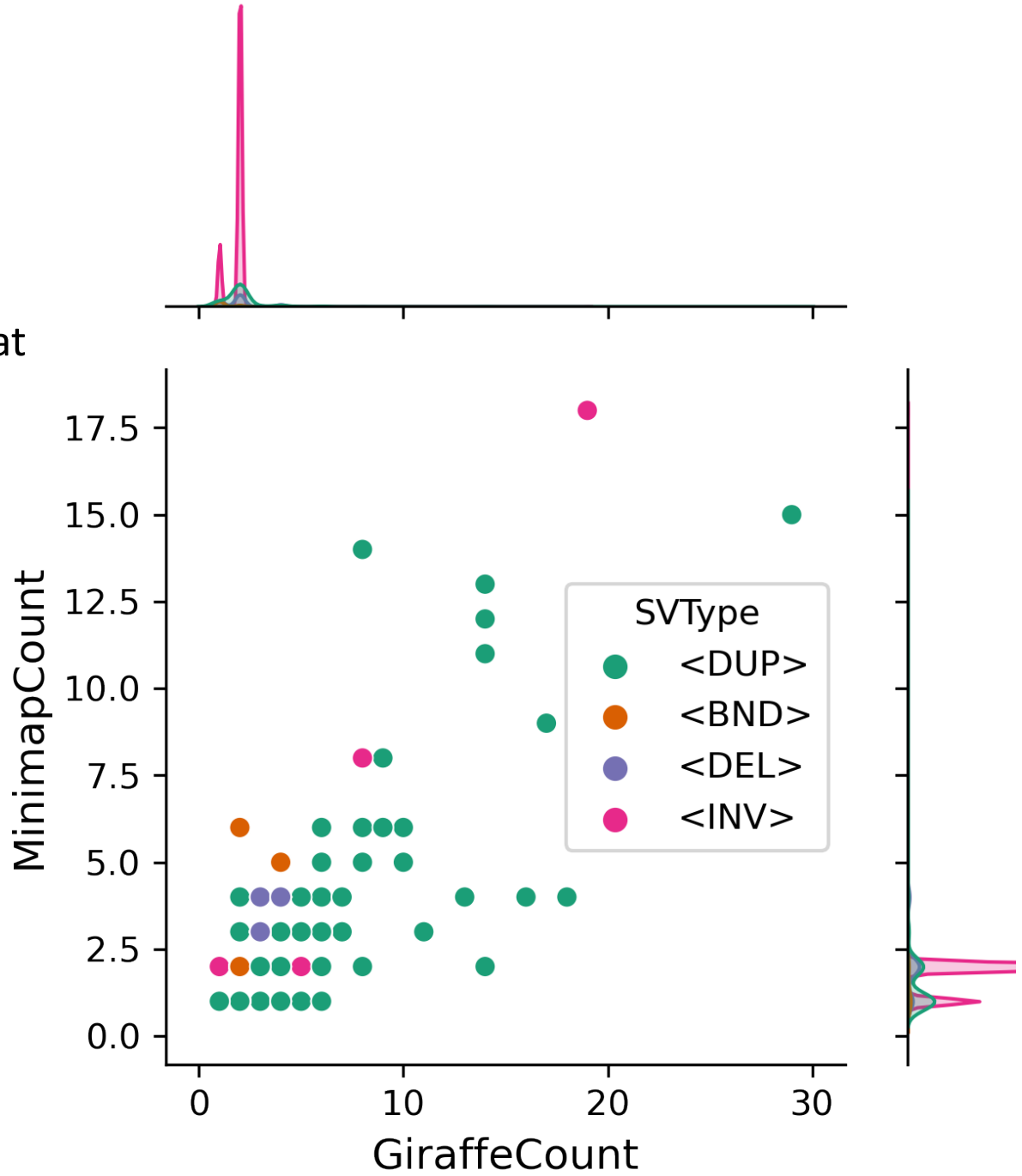
- Total calls:
  - Minimap: 10,934
  - Giraffe: 32,842


Structural Variants Unique to Each Alignment Method
Number of Structural Variants
0
2000
4000
6000
8000
10000
12000
14000
16000
SVType
<INV>
<DUP>
<BND>
<DEL>
17,095
6,930
1,886
3
2,623
1,047
126
247
4,391
1,477
190
611
<INV>: 6.5 fold more
<DUP>: 6.6 fold more
<BND>: 15.0 fold more
<DEL>: ~ 100 fold fewer
Giraffe-Only
Minimap-Only
Both

Shared Structural Variant Concordances

- Allele counts for common SV calls
- Giraffe nearly doubled frequencies
- Distinct sub-group of calls that had low concordance





Most novel graph-based calls were in heterochromatin regions

Likely mappings to repetitive regions not possible with linear reference and alignment

- Centromeric repeats
- Sub-telomeric regions

Comprise nearly all BND variants and the majority of INV

Likely neutral variants, or minor subspecies-specific differences

Conclusions

Pangenome resources improve accuracy of variant detection with cost-efficient short read datasets

Mapping rate improvements: 1.3%

SNP and INDEL calling reference bias: < 38%

SV calling rates: 10-fold more BND events

Tools for pangenome alignment and variant calling are about 4-5 years behind linear tools in terms of efficiency, but still useable!

Worthwhile investment in newer resources, but tools to interact with the graph formats need more development time

Acknowledgments

- Hendrix Genetics:
  - Katrijn Peeters
  - Bruno Perez
  - Carolien Vermeij
  - Marco Bink
  - Johan Van Arendonk
  - Many more
- The Wageningen Cluster (Anunna)
- University of Missouri
  - Ed Rice
  - Wes Warren
- Ludwig-Maximilians-Universität
  - Laurent Frantz
- NCBI NLM