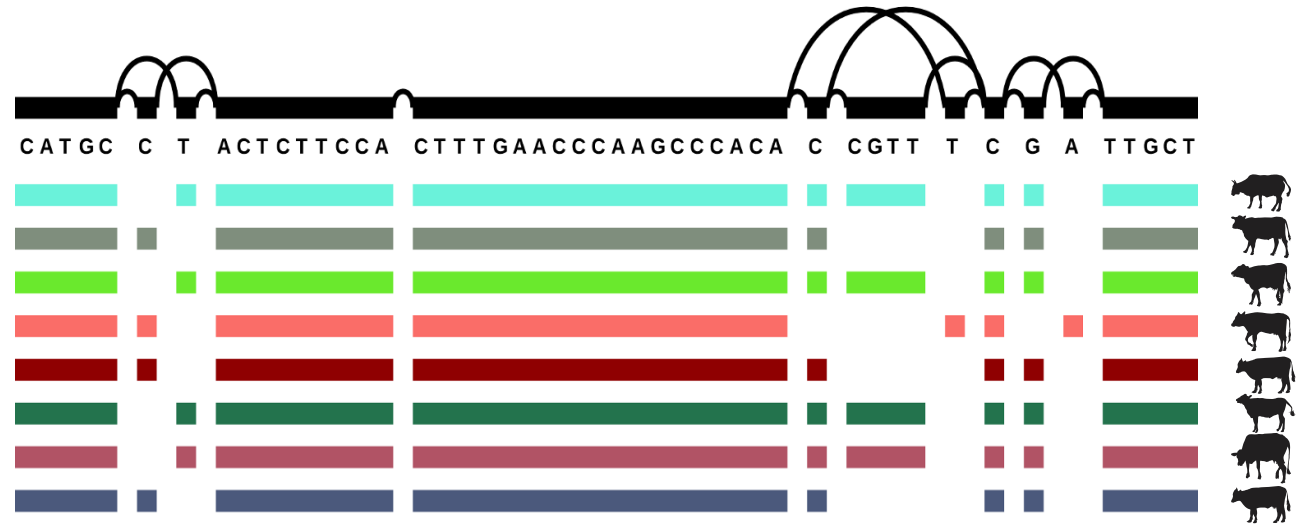


# Pangenomes: A new era of genomics

Temitayo A. Olagunju

University of Idaho

tolagunju@uidaho.edu



*Opportunities and obstacles to enhancing beef cattle evaluation with sequence data*

12th Genetic Prediction Workshop – 2023

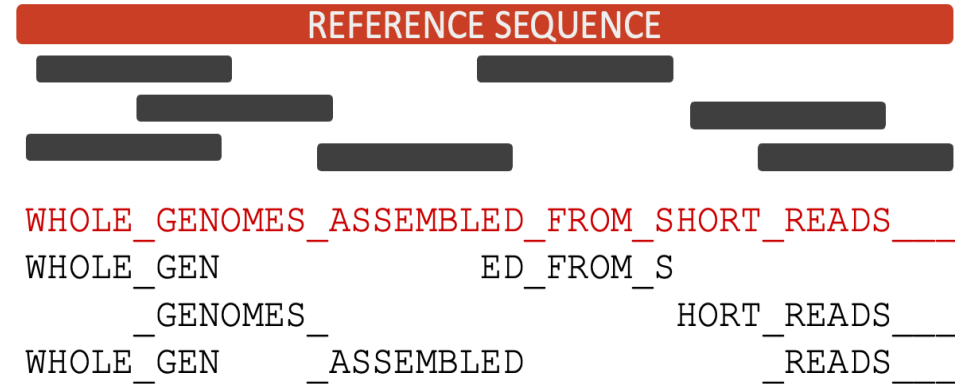
Dec 18-20

# Presentation outline

- Reference genomes
- The limitations of a single reference genome?
- Beyond a single reference - Pangenomes
- Are pangenomes useful?
- Bovine Pangenome (BP) project
- Early lessons from the BP project
- Value of a pangenome to a producer

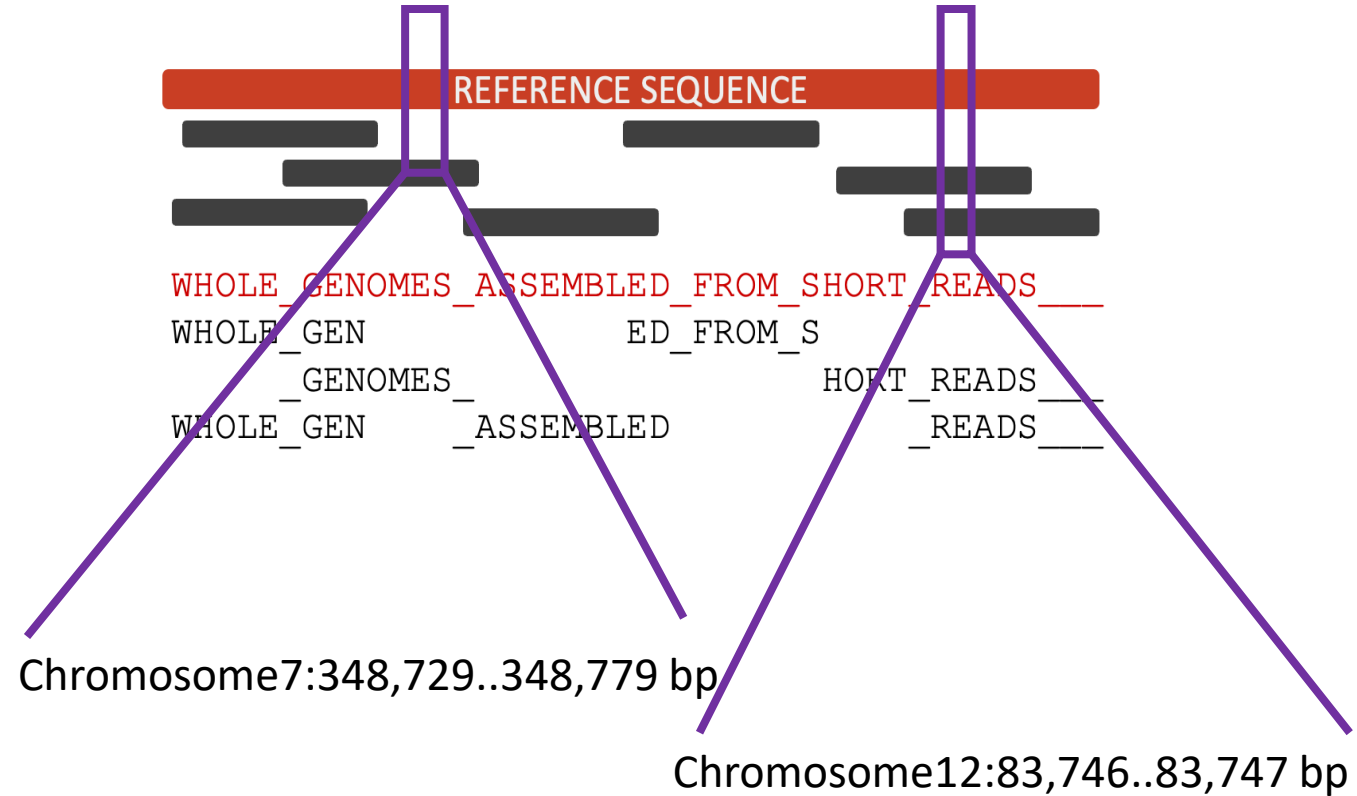
# Reference genomes

- Reference genomes
  - Short reads
- Guide for genome assembly
- Coordinate system for analysis

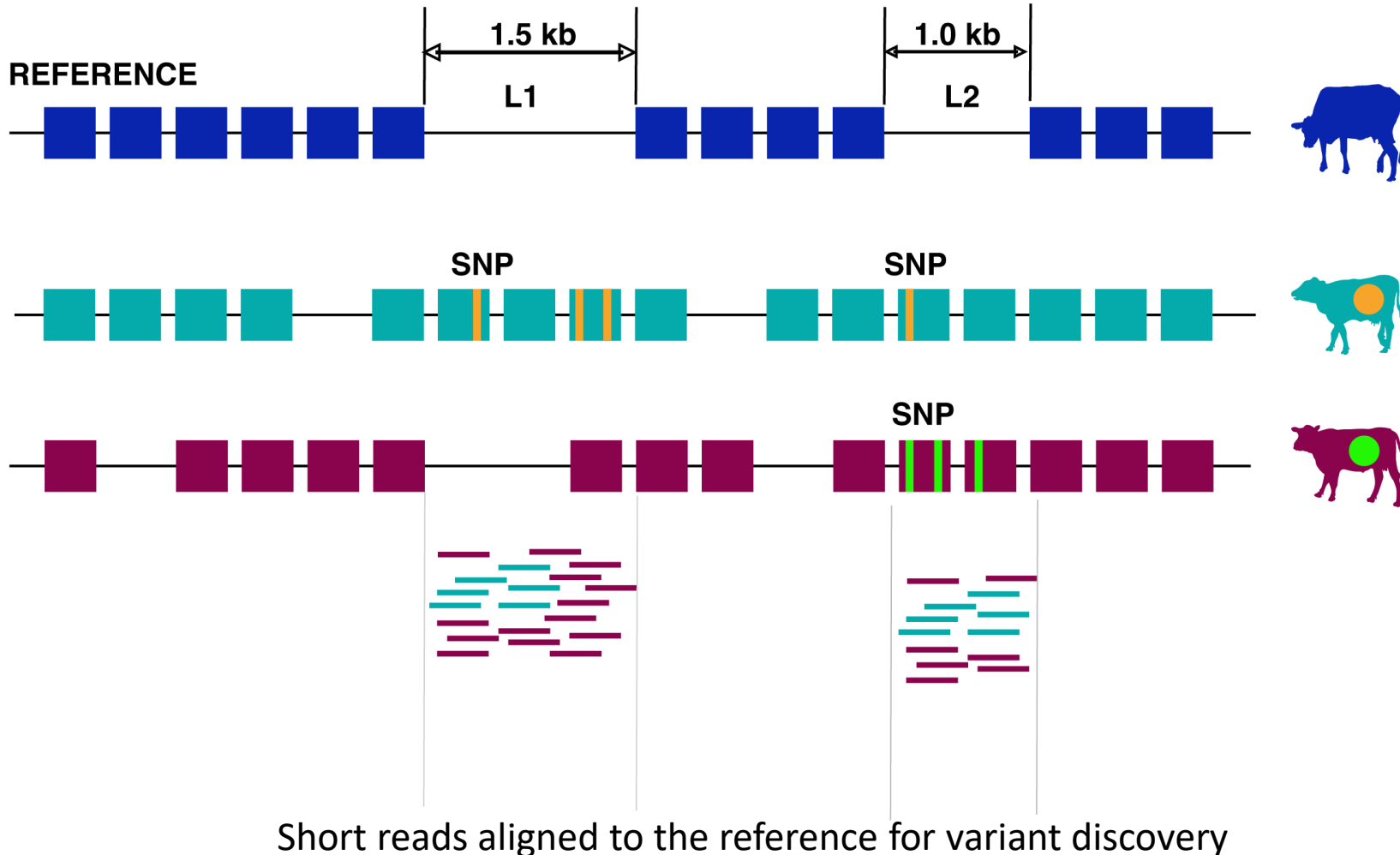


# Reference genomes

- Reference genomes
  - Short reads
- Guide for genome assembly
- Coordinate system for analysis



# Reference genomes - limitations



- **Reference bias**
- Missed true variants in individuals due to absence of region on reference
- Erroneous variants due to reads aligning to wrong places

# Beyond a single reference - Pangenomes

- **Pan-genome**
  - Origin of pangenomes
- 8 strains exhibited variation
- 2,160,262 bp (2,175 genes)
  - ~1400x – cattle genome

## Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for the microbial "pan-genome"

Hervé Tettelin<sup>a,b</sup>, Vega Massignani<sup>b,c</sup>, Michael J. Cieslewicz<sup>b,d,e</sup>, Claudio Donati<sup>c</sup>, Duccio Medini<sup>c</sup>, Naomi L. Ward<sup>a,f</sup>, Samuel V. Angiuoli<sup>g</sup>, Jonathan Crabtree<sup>g</sup>, Amanda L. Jones<sup>g</sup>, A. Scott Durkin<sup>g</sup>, Robert T. DeBoy<sup>g</sup>, Tanja M. Davidsen<sup>g</sup>, Marirosa Mora<sup>c</sup>, Maria Scarselli<sup>c</sup>, Immaculada Margarit y Ros<sup>c</sup>, Jeremy D. Peterson<sup>g</sup>, Christopher R. Hauser<sup>g</sup>

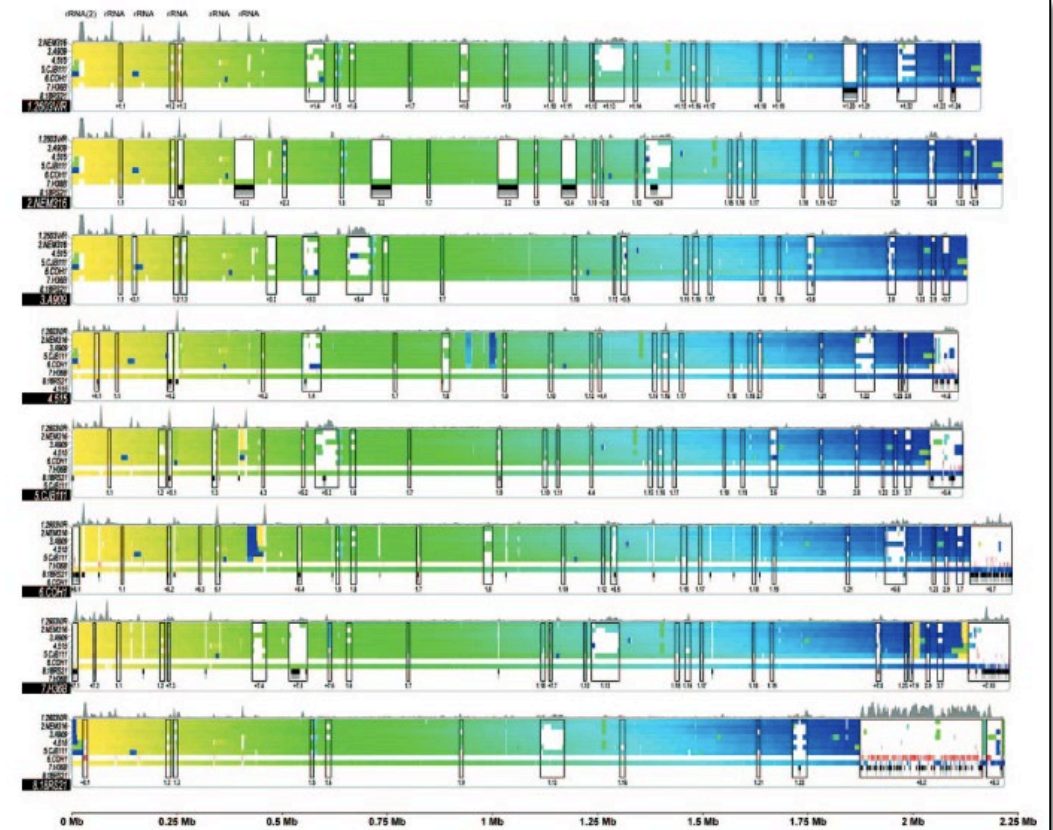
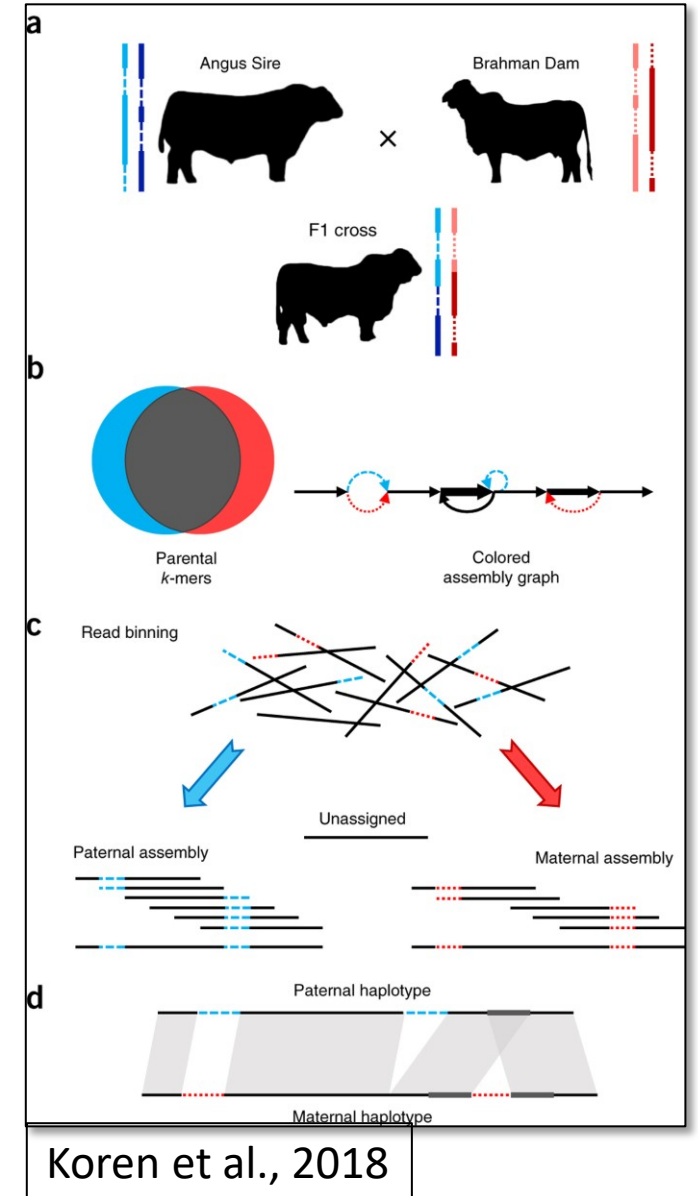


Fig. 1. Whole genome alignment of GBS strains. The eight genomes are compared to each other by using COG (41) and NUCMER analyses (see Materials and

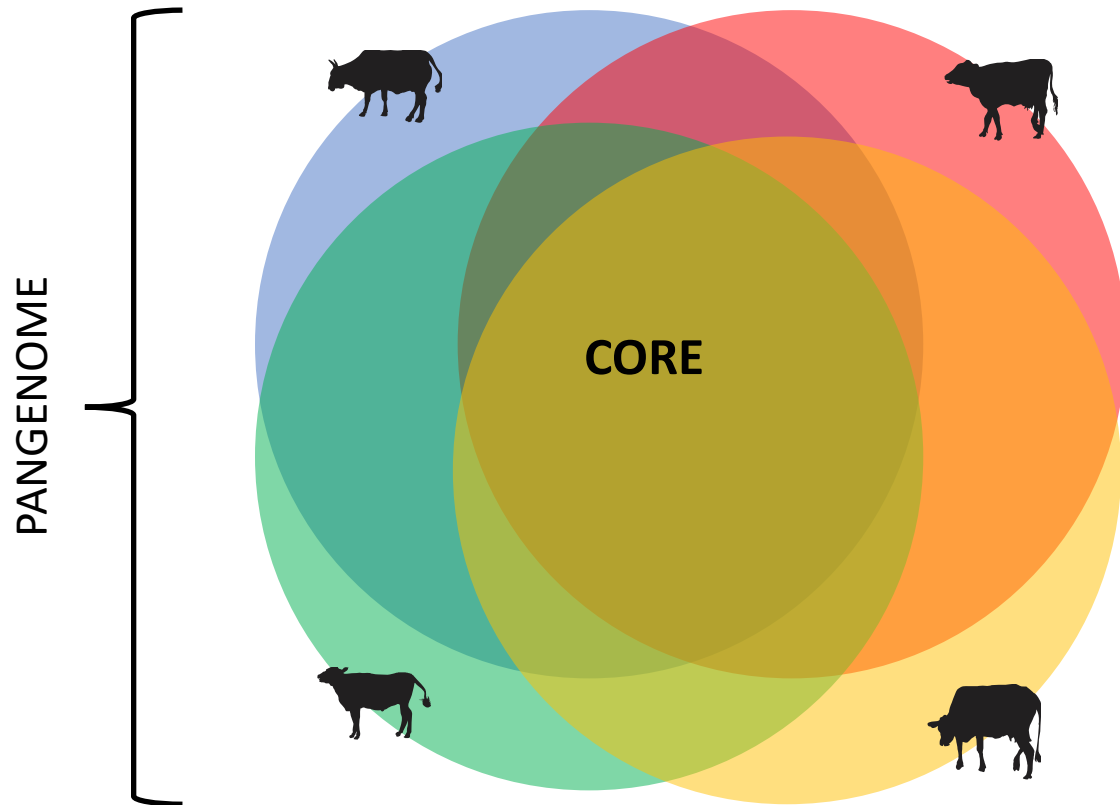
# Beyond a single reference - Pangenomes

- Advancement in sequencing technologies provided an opportunity for *de novo* assemblies
  - HiFi reads – 15-20,000 bp @ 99.9% accuracy
  - UL reads – 100,000 bp length
  - Trio-binning – **haplotype** phasing
  - Hi-C technology – **haplotype** phasing
- Variants calling is only as good as the **query genome** and **the reference**



# Pangenomes

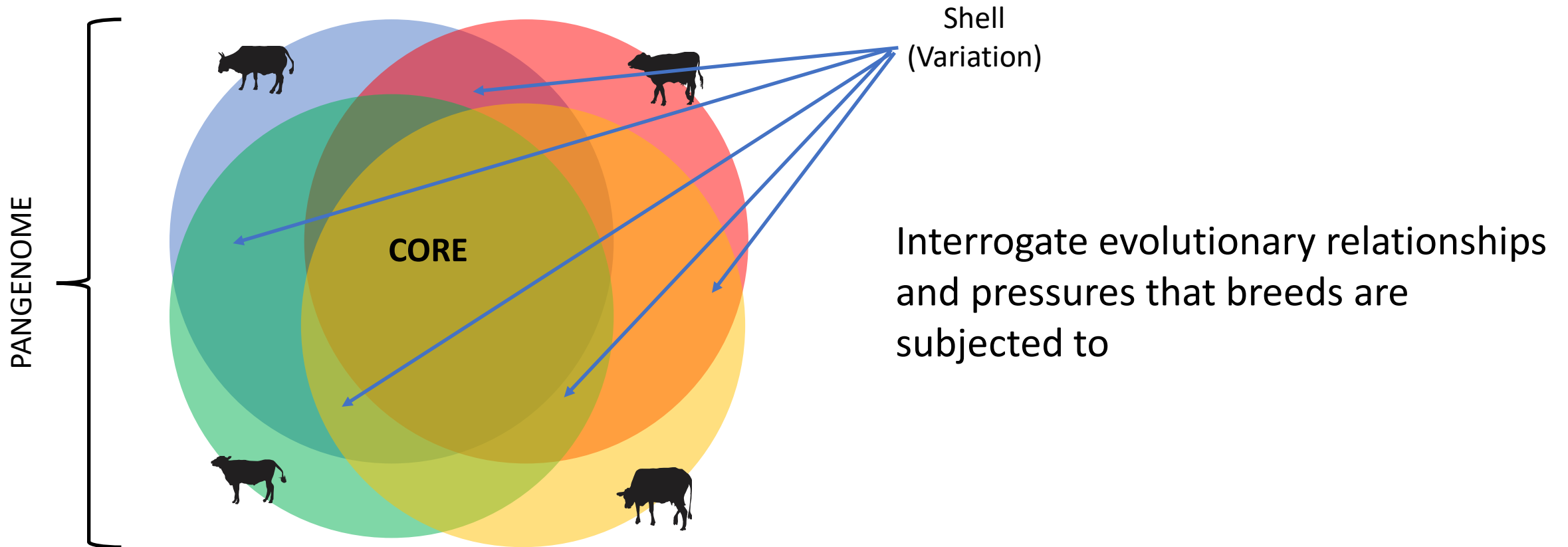
- The genome as a set + other components
  - The shell comprises intersections between multiple individuals





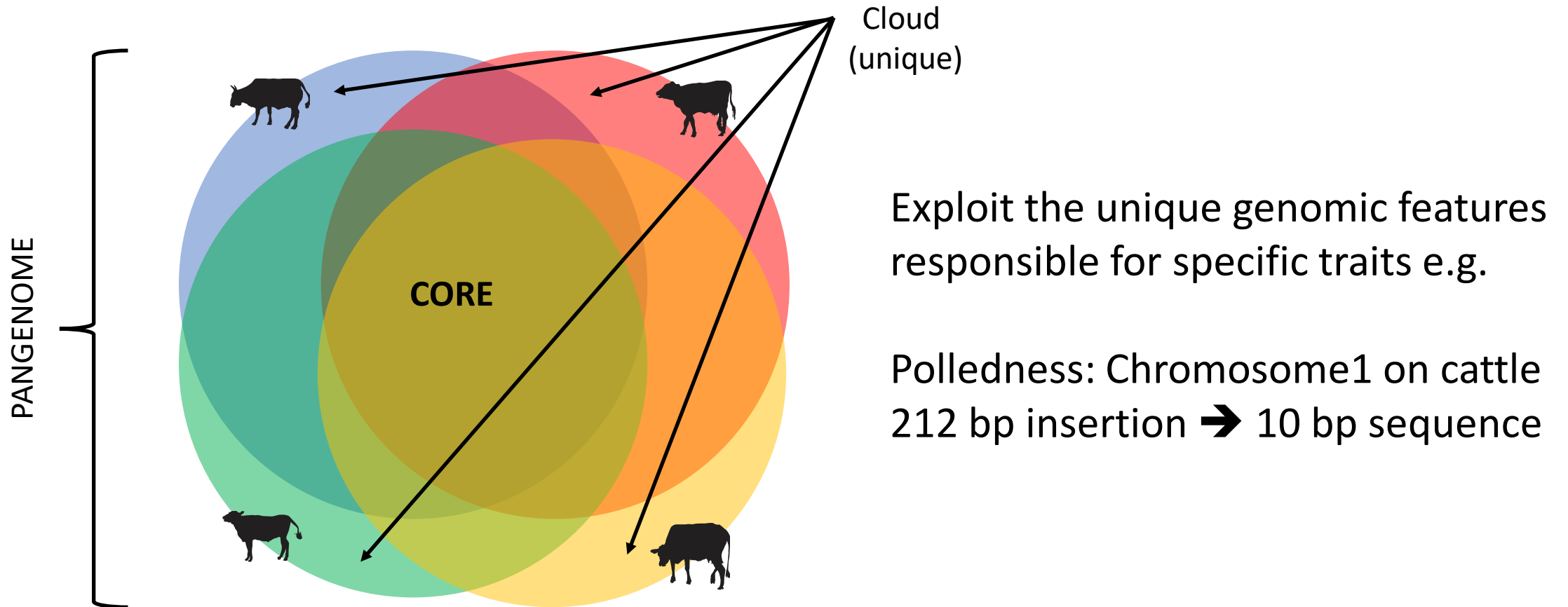
# Pangenomes

- The genome as a set + other components
  - The shell comprises intersections between multiple individuals



# Pangenomes

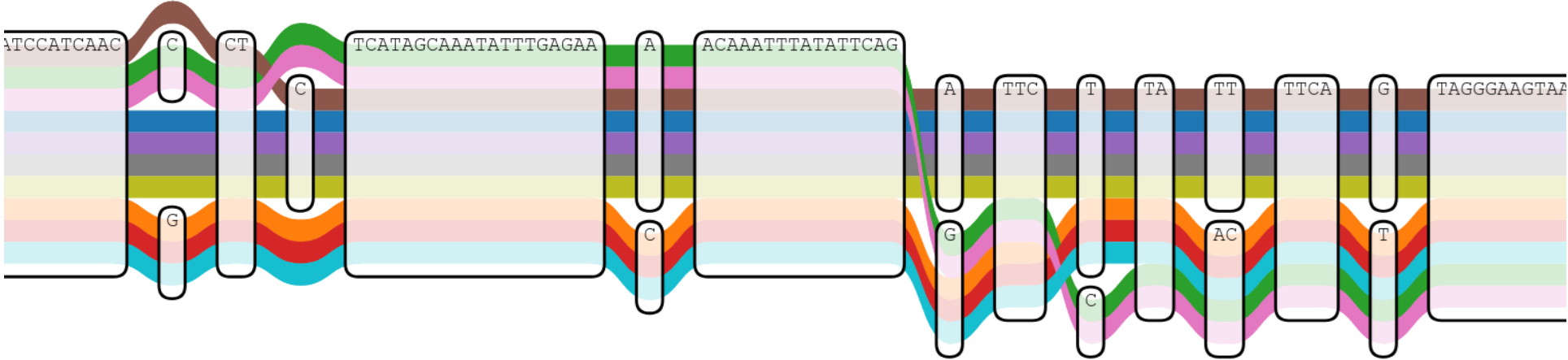
- The genome as a set + other components
  - Cloud comprising only unique genome



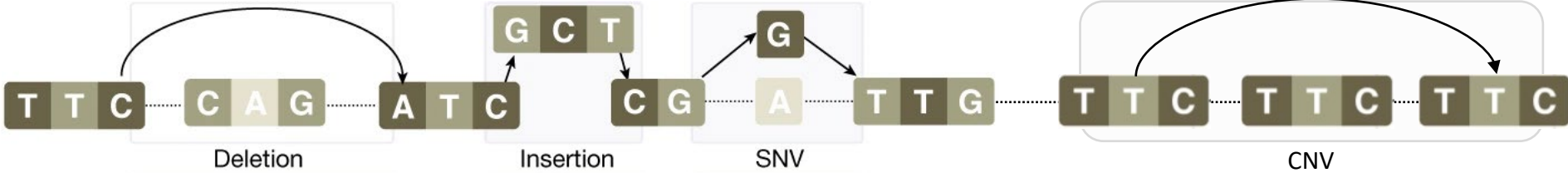
# Pangenomes

- Representing a pangenome
  - Variation graph (VG)

Representation



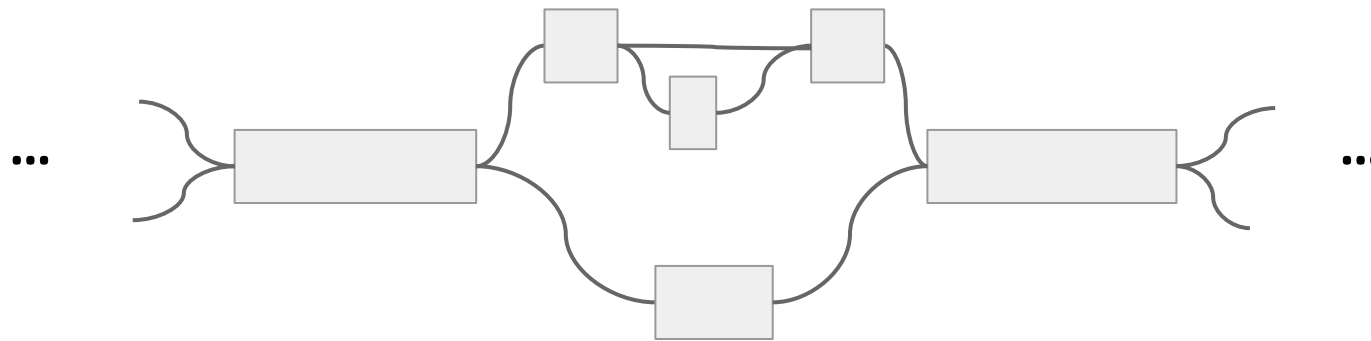
SVs on a VG



Adapted from Erik Garrison

# Pangenomes

- **Variation graph**
  - Data structure that captures the variation

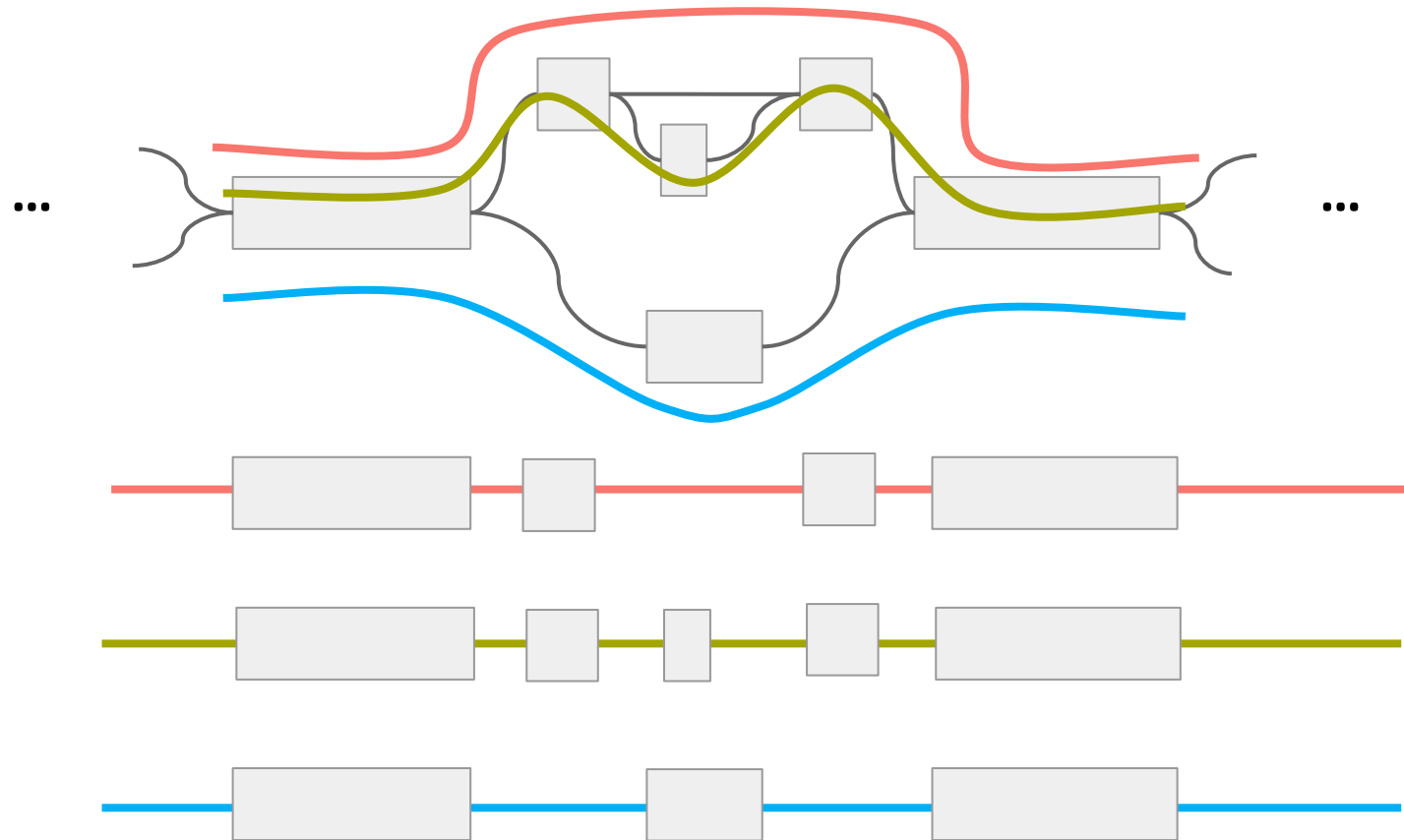


Nodes are genome segments

Edges connect the segments

# Pangenomes

- **Variation graph**
  - Data structure that captures the variation



Nodes are genome segments

Edges connect the segments

Genome-1

Genome-2

Genome-3

# Pangenomes - requirements

- Constructing pangenomes
  - A highly accurate and complete genome assembly is required to reduce false positives and negatives – the role of T2T chromosome-level assemblies
    - ASM errors will lead to false SVs discovery
    - Reference-agnostic ASM QC metrics
- The pangenome becomes a new reference:
  - Inferences based on this “neo-reference” can only be good as the reference

# Pangenomes make a difference



## Novel functional sequences uncovered through a bovine multiassembly graph

Danang Crysanto<sup>a,1</sup> , Alexander S. Leonard<sup>a</sup> , Zih-Hua Fang<sup>a</sup> , and Hubert Pausch<sup>a</sup> 

<sup>a</sup>Animal Genomics, Eidgenössische Technische Hochschule (ETH) Zürich, 8315 Zürich, Switzerland

Edited by Harris A. Lewin, University of California, Davis, CA, and approved April 2, 2021 (received for review January 18, 2021)

Many genomic analyses start by aligning sequencing reads to a linear reference genome. However, linear reference genomes are imperfect, lacking millions of bases of unknown relevance and are unable to reflect the genetic diversity of populations. This makes reference-guided methods susceptible to reference-allele bias. To

in sequences that are not present in the reference genome (11). Recent estimates suggest that millions of bases are missing in mammalian reference genomes (12, 13), indicating a high potential for bias.

Efforts to mitigate reference-allele bias and increase the genetic

We show that the nonreference sequences contain transcripts that are differentially expressed as well as polymorphic sites that segregate within and between breeds of cattle.

### Significance

Most sequence variant analyses rely on a linear reference genome that is assumed to lack millions of bases that occur in the genomes of other individuals. To quantify the extent and functional relevance of such missing bases, we integrate six genome assemblies from cattle and related species into a pangenome. This allows us to uncover more than 70 million bases that are not included in the *Bos taurus* reference genome. Through complementary bioinformatics, genomics, and transcriptomics methods, we discover putative genes from nonreference sequences that are differentially expressed and thousands of polymorphic sites that were unused so far. Our work provides a computational framework, broadly applicable to many species, to make a so-far neglected source of genomic variation amenable to genetic investigations.


Author contributions: D.C. and H.P. designed research; D.C., A.S.L., Z.-H.F., and H.P. performed research; D.C., A.S.L., and H.P. analyzed data; and D.C., A.S.L., and H.P. wrote the paper.

More than **69 million** non-reference bases added to the pangenome

# Pangenomes make a difference



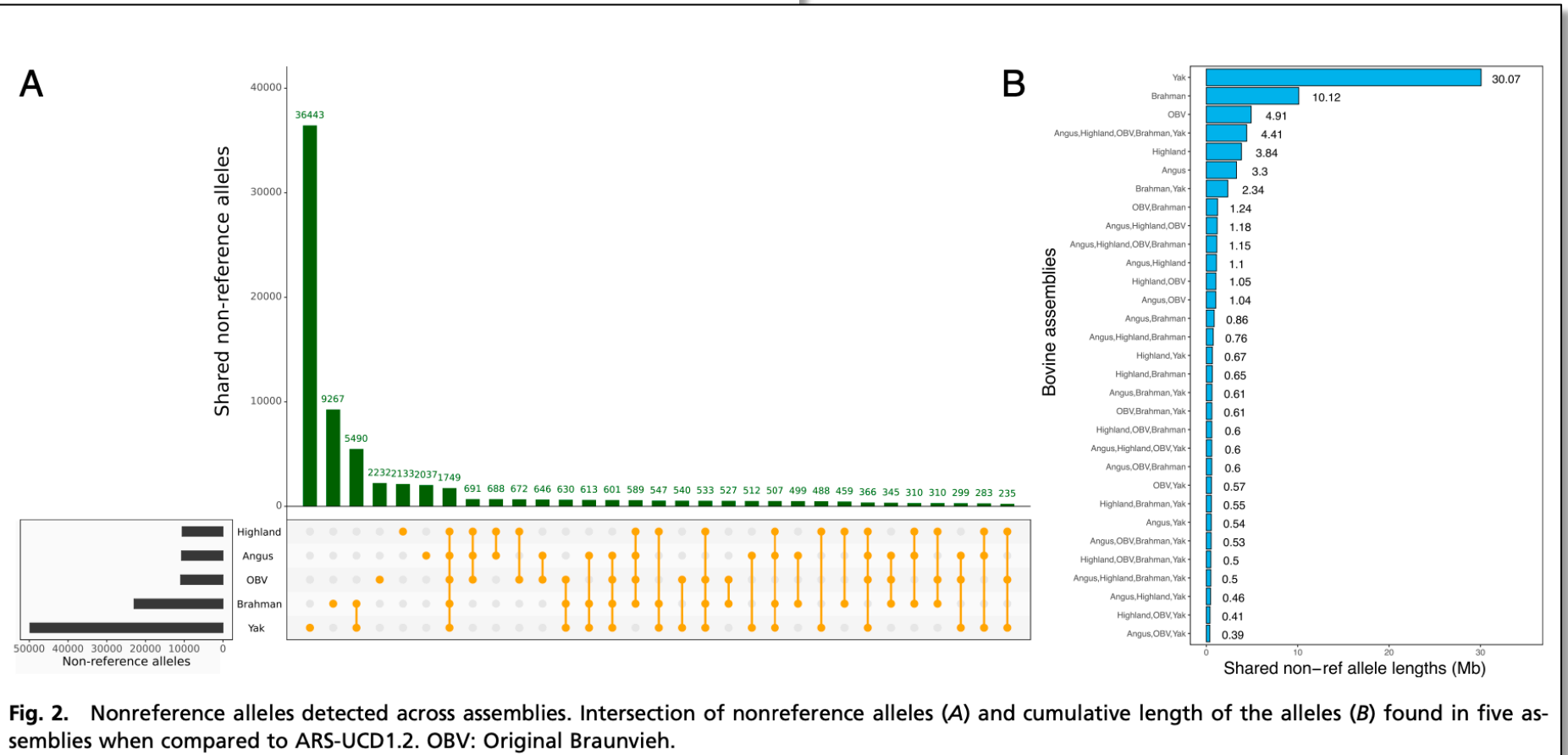
## Novel functional sequences uncovered through a bovine multiassembly graph

Danang Crysanto<sup>a,1</sup> , Alexander S.

<sup>a</sup>Animal Genomics, Eidgenössische Technische Hochschule

Edited by Harris A. Lewin, University of California

Many genomic analyses start by aligning linear reference genome. However, linear imperfect, lacking millions of bases of unknown, unable to reflect the genetic diversity of pangenomes, reference-guided methods susceptible to





# Pangenomes make a difference



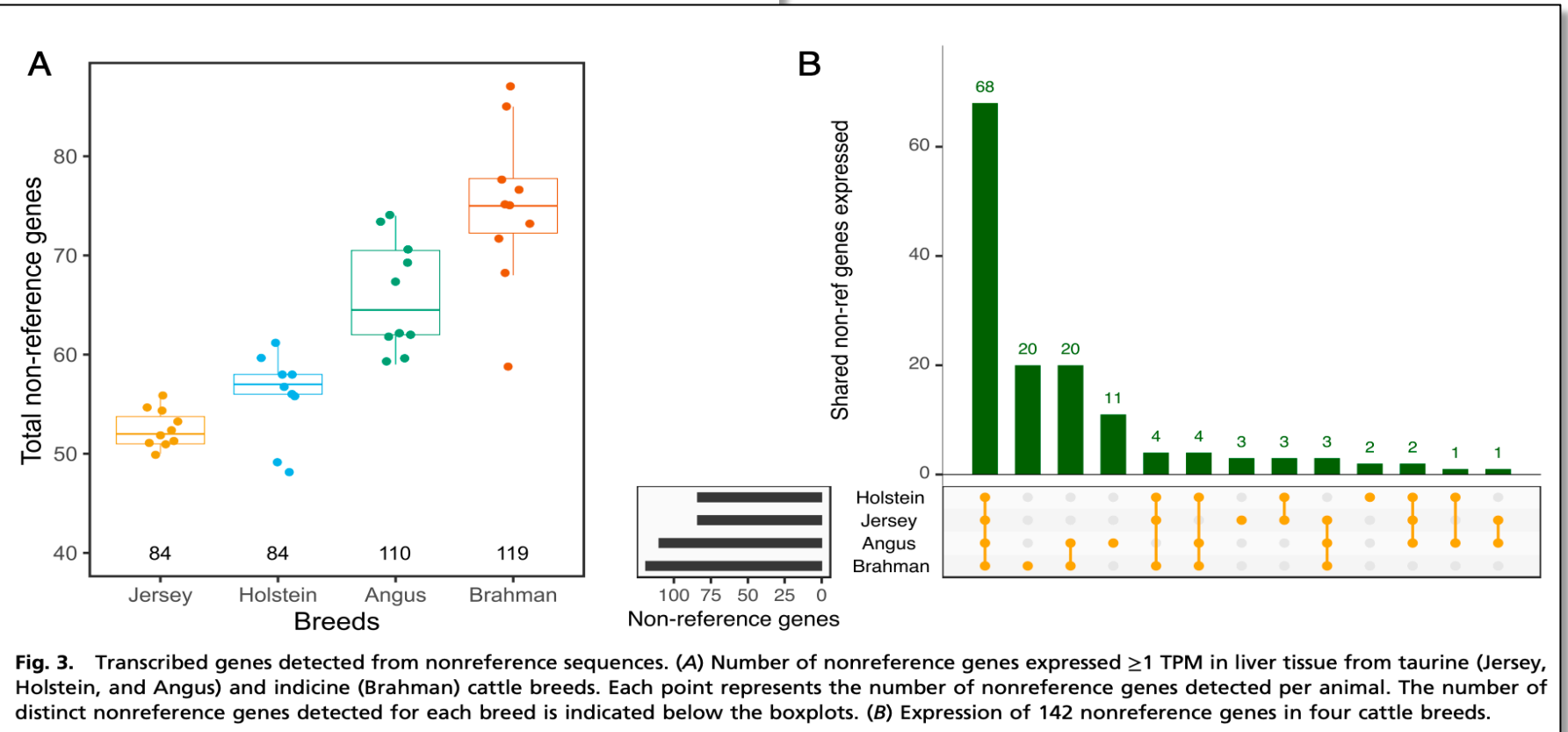
## Novel functional sequences uncovered through a bovine multiassembly graph

Danang Crysanto<sup>a,1</sup> , Alexander S. Leon

<sup>a</sup>Animal Genomics, Eidgenössische Technische Hochschule

Edited by Harris A. Lewin, University of California, Davis

Many genomic analyses start by aligning sequences to a linear reference genome. However, linear references are imperfect, lacking millions of bases of unknown sequence, and are unable to reflect the genetic diversity of populations. Reference-guided methods are susceptible to reference bias.



**Fig. 3.** Transcribed genes detected from nonreference sequences. (A) Number of nonreference genes expressed  $\geq 1$  TPM in liver tissue from taurine (Jersey, Holstein, and Angus) and indicine (Brahman) cattle breeds. Each point represents the number of nonreference genes detected per animal. The number of distinct nonreference genes detected for each breed is indicated below the boxplots. (B) Expression of 142 nonreference genes in four cattle breeds.




**1,431** non-reference genes, 885 expressed

# Pangenomes make a difference

nature > nature communications > articles > article

Article | [Open access](#) | [Published: 31 May 2022](#)

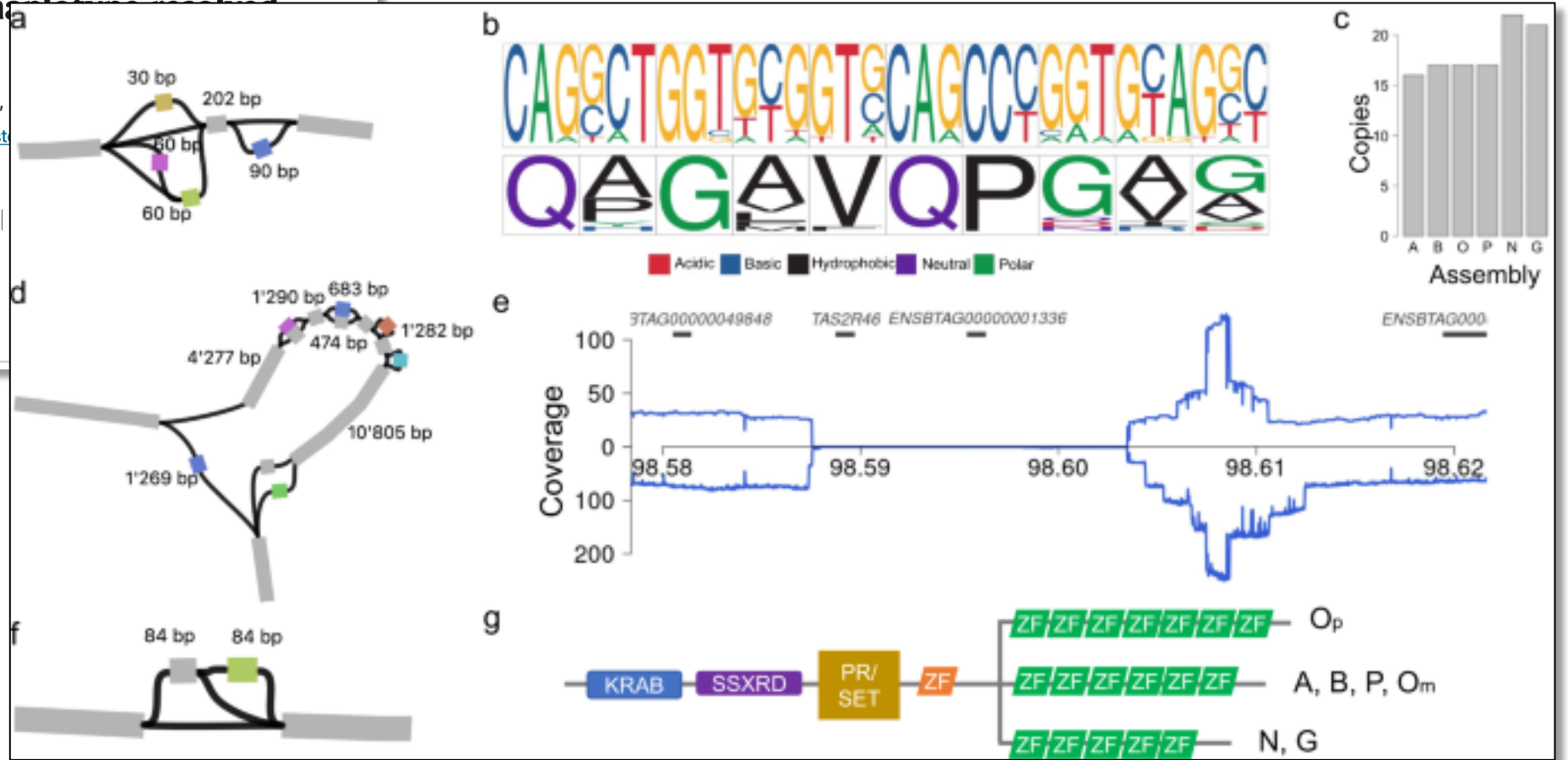
## Structural variant-based pangenome construction has low sensitivity to variability of haplotypes in bovine assemblies

[Alexander S. Leonard](#) , [Danang Crysanto](#), [Zih-Hua Fang](#), [Carolina Herrera](#), [Heinrich Bollwein](#), [Derek M. Bickhart](#), [Kristin D. Rosen](#)  & [Hubert Pausch](#) 

*Nature Communications* **13**, Article number: 3012 (2022)

6839 Accesses | 14 Citations | 15 Altmetric | [Metrics](#)

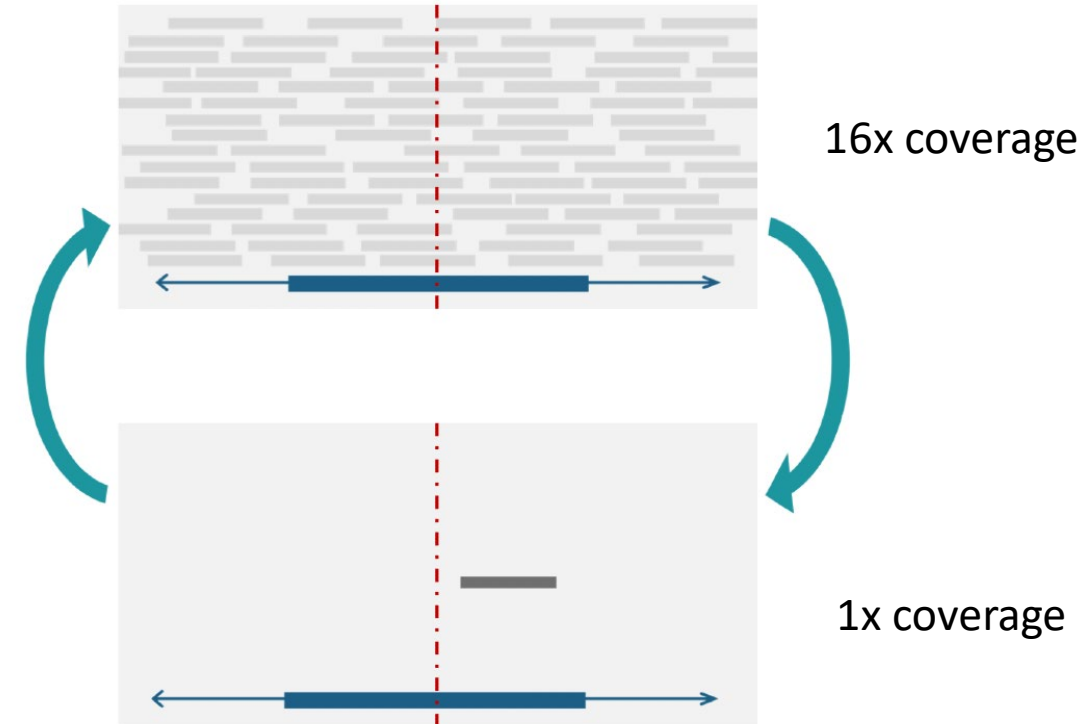
### Abstract



Pangenome reveals structural variation on some coding genes

# Pangenomes make a difference

- With a high-quality pangenome reference, more accurate variants detection can be made with low sequencing coverage.
  - Shallow / low-pass WG sequencing
  - Cheaper than deep WG sequencing

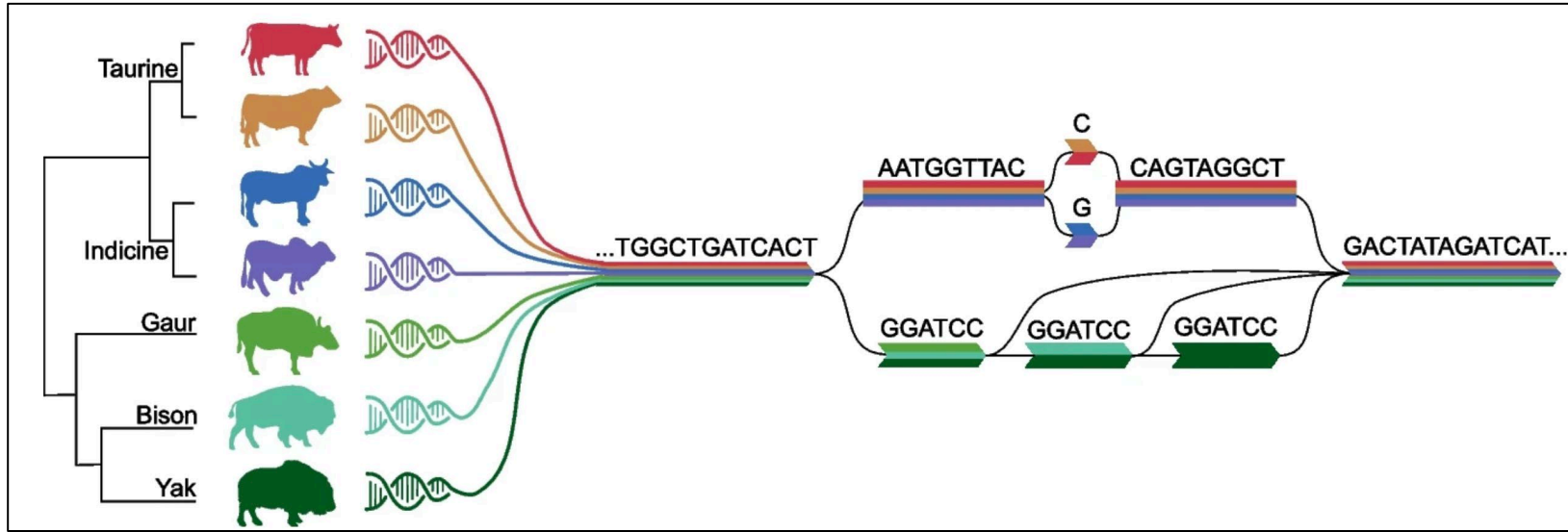


<https://www.cancer.gov/ccg/blog/2019/low-coverage-seq>

# Pangenomes: Value to the producer

- Pangenomes enable exploration of the influence of other SVs besides SNPs on livestock traits
- Better genome prediction
- Exploit the genetic diversity in other breeds that could be helpful to a specific breed

# The Bovine Pangenome Consortium



Represent global cattle variation

Genome Biology


A global collaborative effort

## CORRESPONDENCE

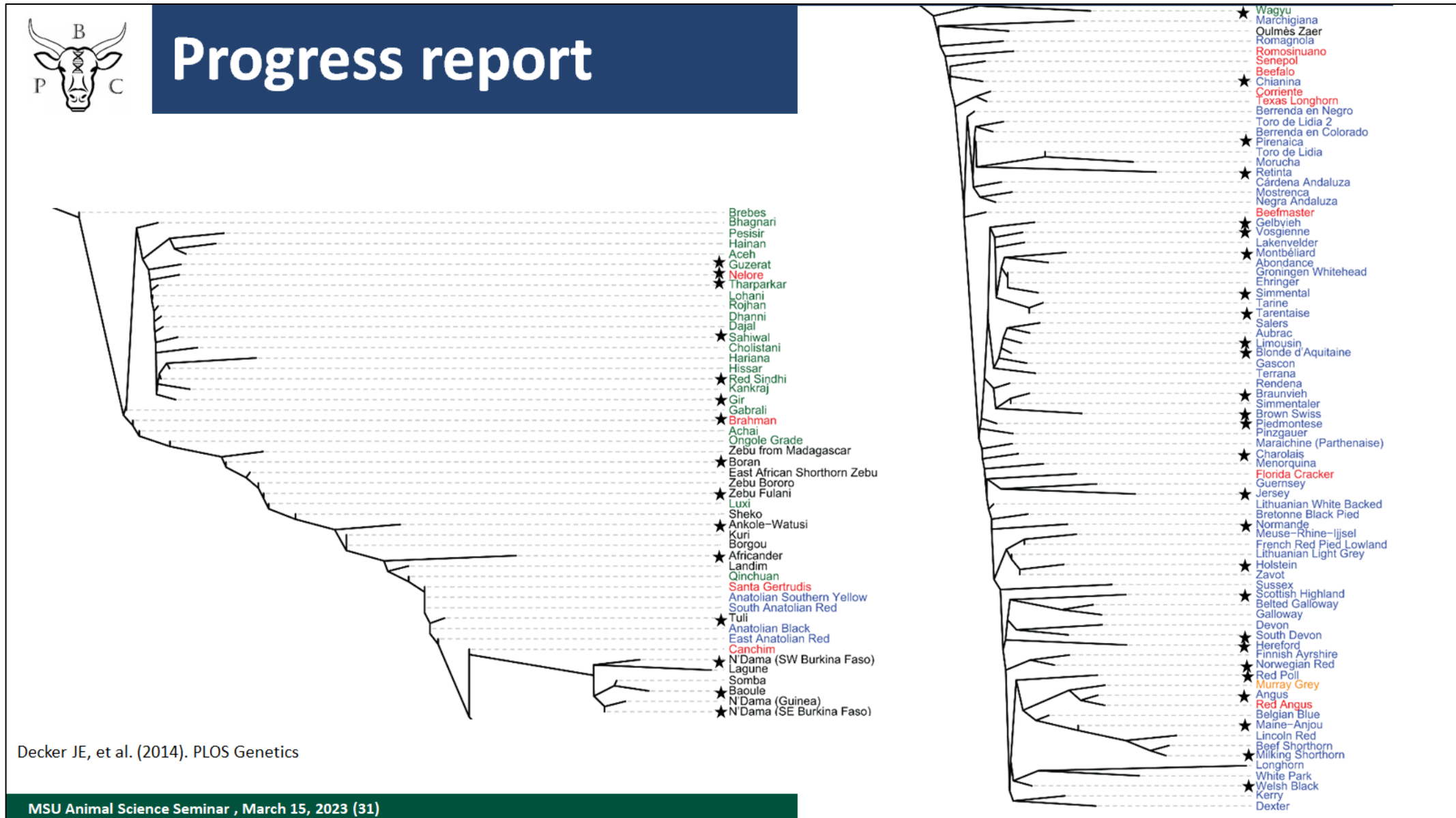
Open Access

## The Bovine Pangenome Consortium: democratizing production and accessibility of genome assemblies for global cattle breeds and other bovine species



Timothy P. L. Smith<sup>1</sup>, Derek M. Bickhart<sup>2</sup>, Didier Boichard<sup>3</sup>, Amanda J. Chamberlain<sup>4,5</sup>, Appolinaire Djikeng<sup>6,7</sup>, Yu Jiang<sup>8</sup>, Wai Y. Low<sup>9</sup>, Hubert Pausch<sup>10</sup>, Sebastian Demyda-Peyrás<sup>11,12</sup>, James Prendergast<sup>7,13</sup>, Robert D. Schnabel<sup>14</sup>, Benjamin D. Rosen<sup>15\*</sup>  and Bovine Pangenome Consortium

# The Bovine Pangenome Consortium



# The Bovine Pangenome Consortium

S/N	Cattle Breed	In progress	Number of individuals
1	Angus	y	Multiple
2	Asmo	y	1
3	Ayrshire	y	1
4	Brahman	y	1
5	Brown Swiss	y	2
6	Charolais	y	3
7	Gelbvieh	y	1
8	Guzerat	y	1
9	Holstein	y	Multiple
10	Luvattu	y	1
11	Red Wagyu	y	1
12	Retinta	y	1
13	Rubia Gallega	y	1
14	Shorthorn	y	1
15	Welsh Black	y	1
16	White Fulani	y	1
17	Whitebred	y	1
18	Yiling cattle	y	1

Samples obtained for the listed breeds

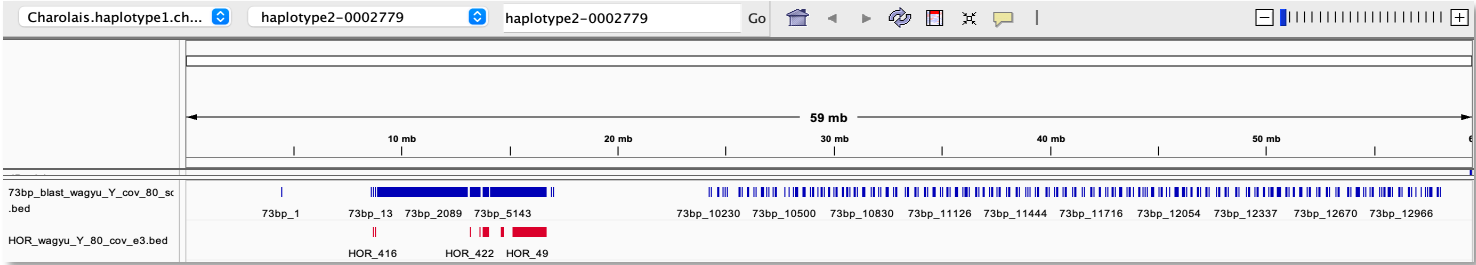
More breeds are still expected – a bottleneck

A data freeze will be implemented soon to enable analysis across breeds

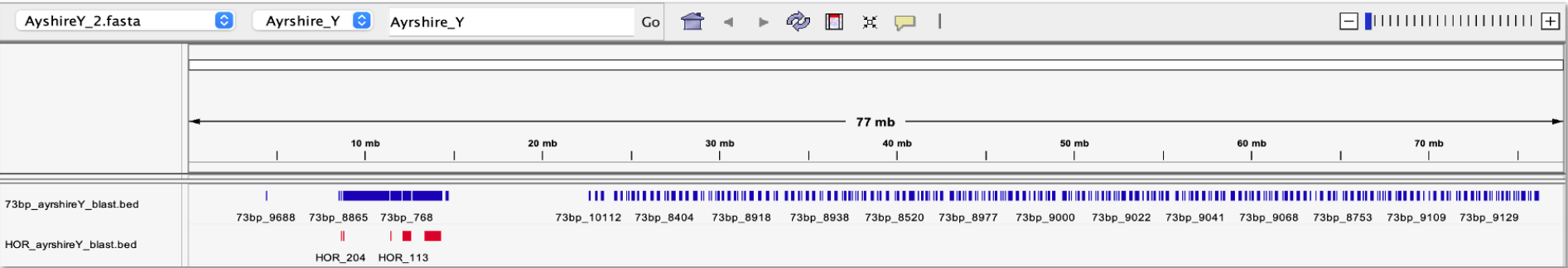
# Early lessons



Wagyu  
59.4 Mb



Ayrshire  
77.1 Mb



Chromosome Y: length difference but similar chromosome structure



# Summary

- Single reference genomes are insufficient to capture species diversity
- New sequencing and assembly technologies provide an opportunity to produce high quality *de-novo* genomes.
- A set of diverse genomes – Pangenome – required to alleviate the insufficiencies of a single reference genome.
- A pangenome provides a unique opportunity for better genome prediction from traits-linked structural variants.
- The Bovine Pangenome Consortium (BPC) currently making efforts to produce a global representation of the cattle breeds diversity.

# Acknowledgements



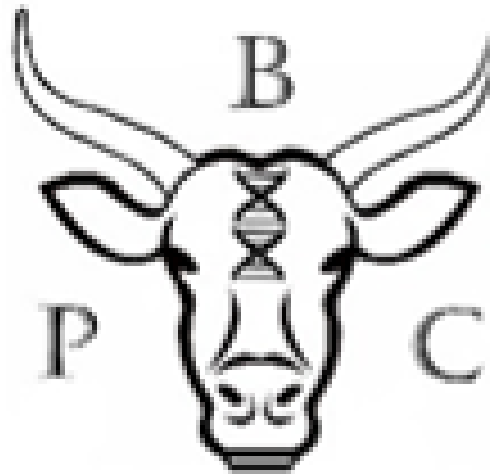
Dr. Brenda M. Murdoch



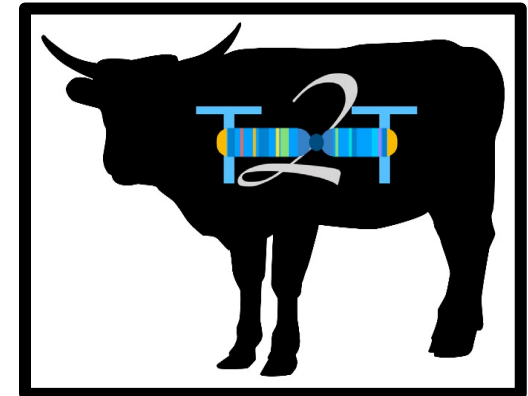
Dr. Ben D. Rosen



Dr. Timothy P.L. Smith



Bovine Pangenome consortium



Ruminant T2T consortium

THANK YOU FOR YOUR ATTENTION

THANK FOR YOUR ATTENTION

YOU FOR YOUR ATTENTION

THANK YOU FOR ATTENTION

THANK YOU FOR YOUR

THANK YOU YOUR ATTENTION