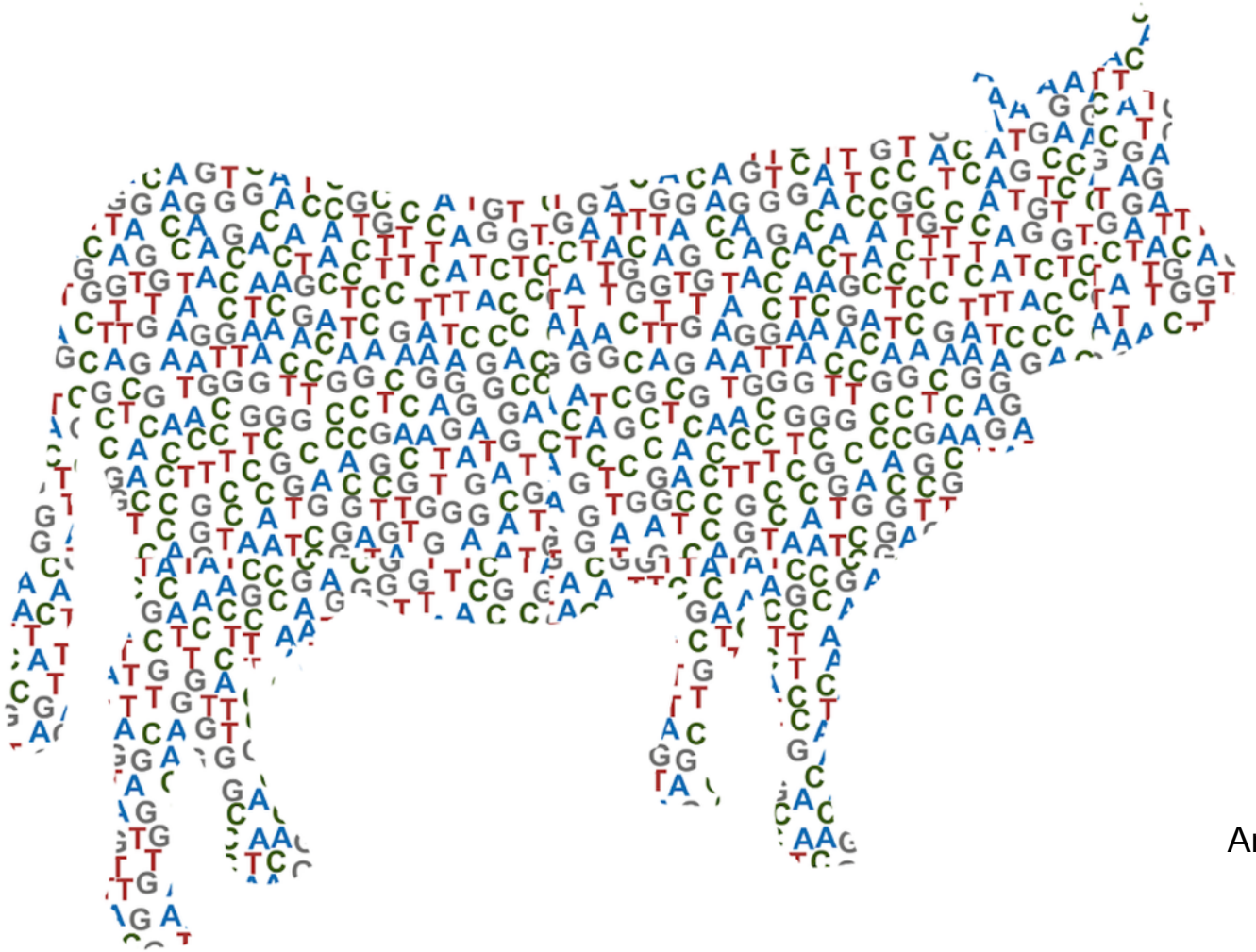
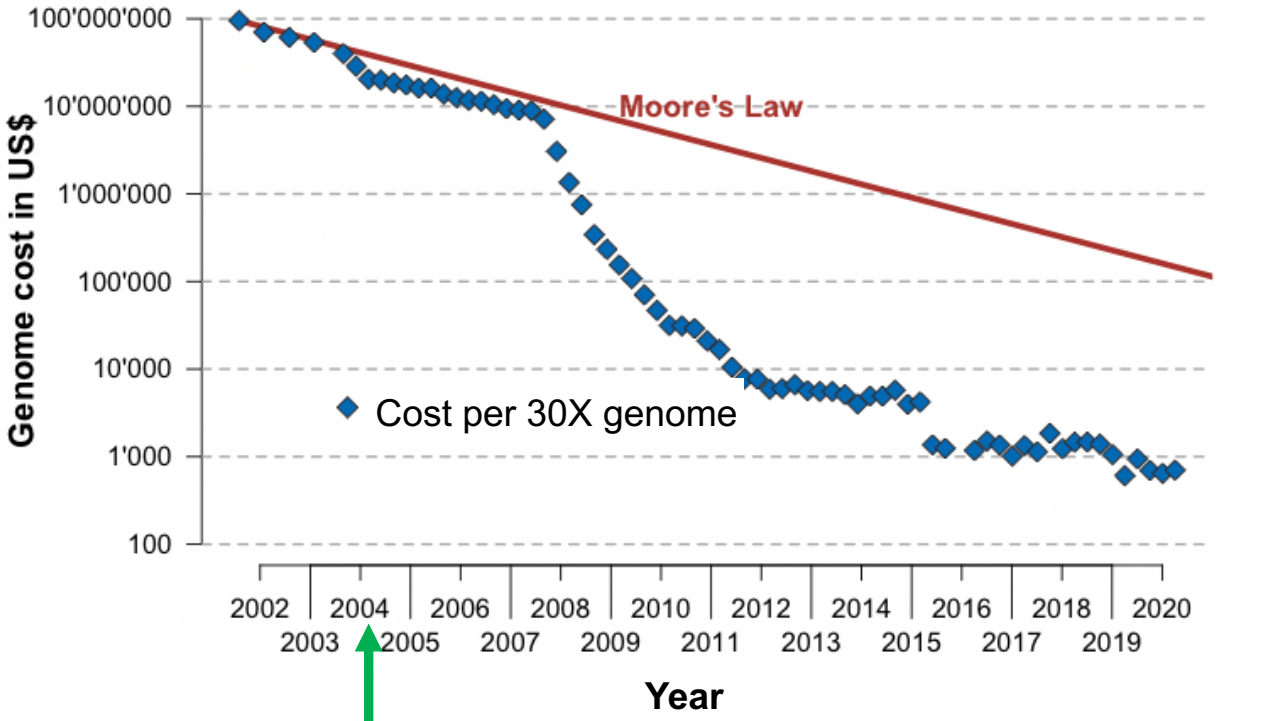


Hubert Pausch
ETH Zürich
Animal Genomics | Institute of Agricultural Sciences

Towards establishing the bovine pangenome from multiple reference-quality assemblies

Zürich / Bern, 16 March 2021

Thousands of cattle genomes have been sequenced …



https://www.genome.gov/sequencingcosts/

Btau2.0

ARTICLES

nature genetics

Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle

Hans D Daetwyler[1–3], Aurélien Capitan[4,5], Hubert Pausch[6], Paul Stothard[7], Rianne van Binsbergen[8], Rasmus F Brøndum[9], Xiaoping Liao[7], Anis Djari[10], Sabrina C Rodriguez[4], Cécile Grohs[4], Diane Esquerré[11], Olivier Bouchez[11], Marie-Noëlle Rossignol[12], Christophe Klopp[10], Dominique Rocha[4], Sébastien Fritz[5], André Eggen[4], Phil J Bowman[1,3], David Coote[1,3], Amanda J Chamberlain[1,3], Charlotte Anderson[1], Curt P VanTassell[13], Ina Hulsegge[8], Mike E Goddard[1,3,14], Bernt Guldbrandtsen[9], Mogens S Lund[9], Roel F Veerkamp[8], Didier A Boichard[4], Ruedi Fries[6] & Ben J Hayes[1–3]

LETTERS
https://doi.org/10.1038/s41588-018-0056-5
nature genetics

Meta-analysis of genome-wide association studies for cattle stature identifies common genes that regulate body size in mammals

Aniek C. Bouwman[1], Hans D. Daetwyler[2,3], Amanda J. Chamberlain[2], Carla Hurtado Ponce[2,4], Mehdi Sargolzaei[5,6], Flavio S. Schenkel[5], Goutam Sahana[7], Armelle Govignon-Gion[8], Simon Boitard[9], Marlies Dolezal[10], Hubert Pausch[2,11,12], Rasmus F. Brøndum[7], Phil J. Bowman[2], Bo Thomsen[9], Bernt Guldbrandtsen[7], Mogens S. Lund[7], Bertrand Servin[13], Dorian J. Garrick[14], James Reecy[14], Johanna Vilkki[15], Alessandro Bagnato[16], Min Wang[2,3], Jesse L. Hoff[17], Robert D. Schnabel[17], Jeremy F. Taylor[17], Anna A. E. Vinkhuyzen[18,19], Frank Panitz[9], Christian Bendixen[9], Lars-Erik Holm[9], Birgit Gredler[20], Chris Hozé[8,21], Mekki Boussaha[8], Marie-Pierre Sanchez[8], Dominique Rocha[8], Aurelien Capitan[8,21], Thierry Tribout[8], Anne Barbat[8], Pascal Croiseau[8], Cord Drögemüller[22], Vidhya Jagannathan[22], Christy Vander Jagt[2], John J. Crowley[23], Anna Bieber[24], Deirdre C. Purfield[25], Donagh P. Berry[25], Reiner Emmerling[26], Kay-Uwe Götz[26], Mirjam Frischknecht[20], Ingolf Russ[27], Johann Sölkner[28], Curtis P. Van Tassell[29], Ruedi Fries[11], Paul Stothard[30], Roel F. Veerkamp[1], Didier Boichard[8], Mike E. Goddard[2,4] and Ben J. Hayes[2,31*]

… and aligned to ARS-UCD1.2

- First assembly of the bovine ***reference sequence*** (2004)
  - Hereford cow «*Dominette*»
  - Cost over 50 million US$

The Genome Sequence of Taurine Cattle: A Window to Ruminant Biology and Evolution

The Bovine Genome Sequencing and Analysis Consortium,* Christine G. Elsik,[1] Ross L. Tellam,[2] Kim C. Worley[3]

To understand the biology and evolution of ruminants, the cattle genome was sequenced to about sevenfold coverage. The cattle genome contains a minimum of 22,000 genes, with a core set of 14,345 orthologs shared among seven mammalian species of which 1217 are absent or undetected in noneutherian (marsupial or monotreme) genomes. Cattle-specific evolutionary breakpoint regions in chromosomes have a higher density of segmental duplications, enrichment of repetitive elements, and species-specific variations in genes associated with lactation and immune responsiveness. Genes involved in metabolism are generally highly conserved, although five metabolic genes are deleted or extensively diverged from their human orthologs. The cattle genome sequence thus provides a resource for understanding mammalian evolution and accelerating livestock genetic improvement for milk and meat production.

Domesticated cattle (*Bos taurus* and *Bos taurus indicus*) provide a significant source of nutrition and livelihood to nearly 6.6 billion humans. Cattle belong to a clade phylogenetically distant from humans and rodents, the Cetartiodactyl order of eutherian mammals, which

2009 VOL 324 SCIENCE www.sciencemag.org

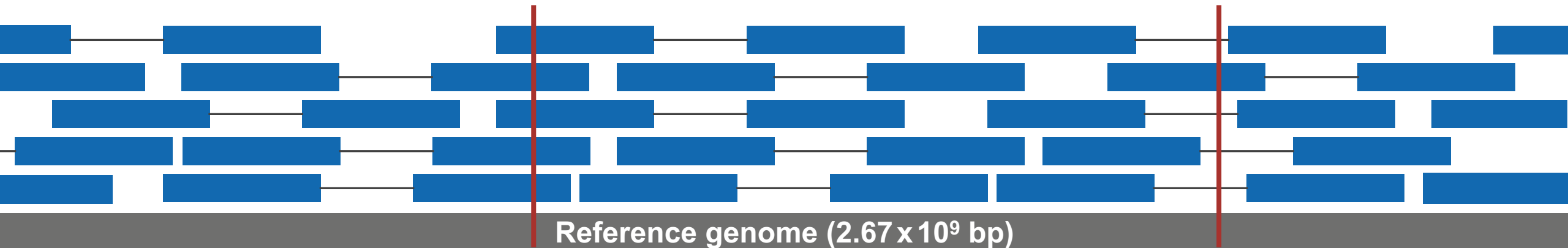
Source: Michael MacNeil, USDA

Typically, more than 100 million sequence read pairs (2 x 150 bases) for each individual corresponding to >10-fold genome coverage

```
GCCCCATGTAGCACAGCCTCTGGAGGCTCTCCACCACTGAGGAAATGTAGCGTCTCTTGTAGAGAAGCTGCTGCCGGATGGTCCGGGGGGAGAAAGCGGATC
GTGAAGAACCTGCCTCTGATGAAACTGCCTGTACGGAGTCTCGAAGTTCAAAGGCGGATGGGGAGGTTGAGACCATCAAACCGAGGCGCTCTGAGAGCCTG
TCACTGATGACCTTGTTGAGCCGATCGTCGTCCGCCTCGATGCCCACGCTGTCCAGGATCTTCTTGATGTCCTTGGCGCTAGGAGAAGTATTGCCCCCGAG
.
.
.
CTGAAAGTTTCCACATGTGGAATATAGATACAACATTGAACAAAATATGTGGCCTCCCATGTACATTGGTTACCTATGTACAAGTATCCTATACACCAGTA
CTTTGGAAAATCCATGTTCTCCACCCGTGTTCTGTGTCTGCCTGTCTGTCAGCACAACTGTGACTCTGAGAACGCAGACTGCGTGGCCTGTGCCGTCAAGG
GATCGGAAGAGCACACGTCTGAACTCCAGTCACCCGTCCCGATCTCGTATGCCGTCTTCTGCTTGAAAAAACAAAAAGAGAGTCTTGCAGGCATCGCCCTG
```

5-fold coverage

1-fold coverage

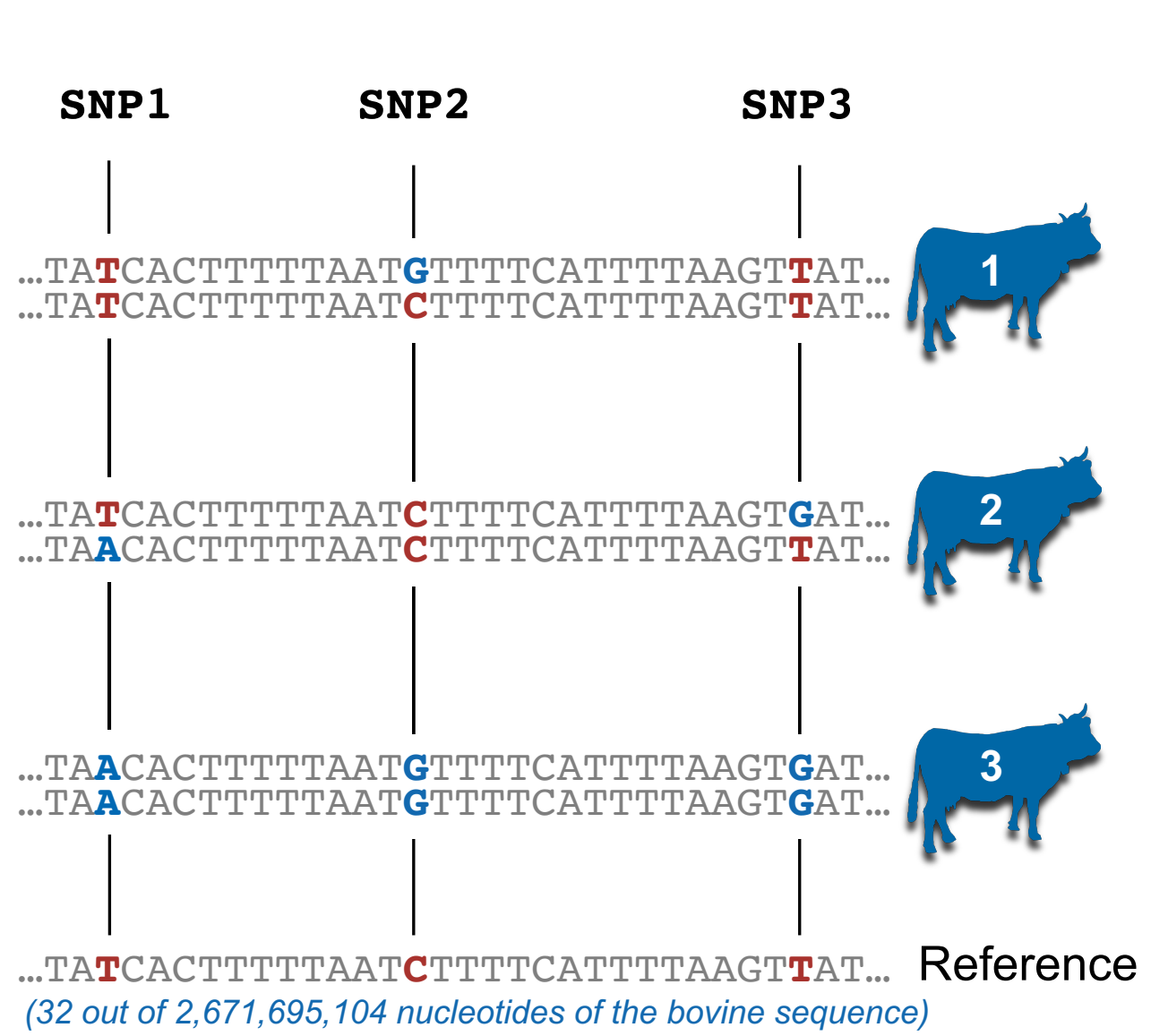


Genetic variation in molecules and matrices

Millions of SNPs

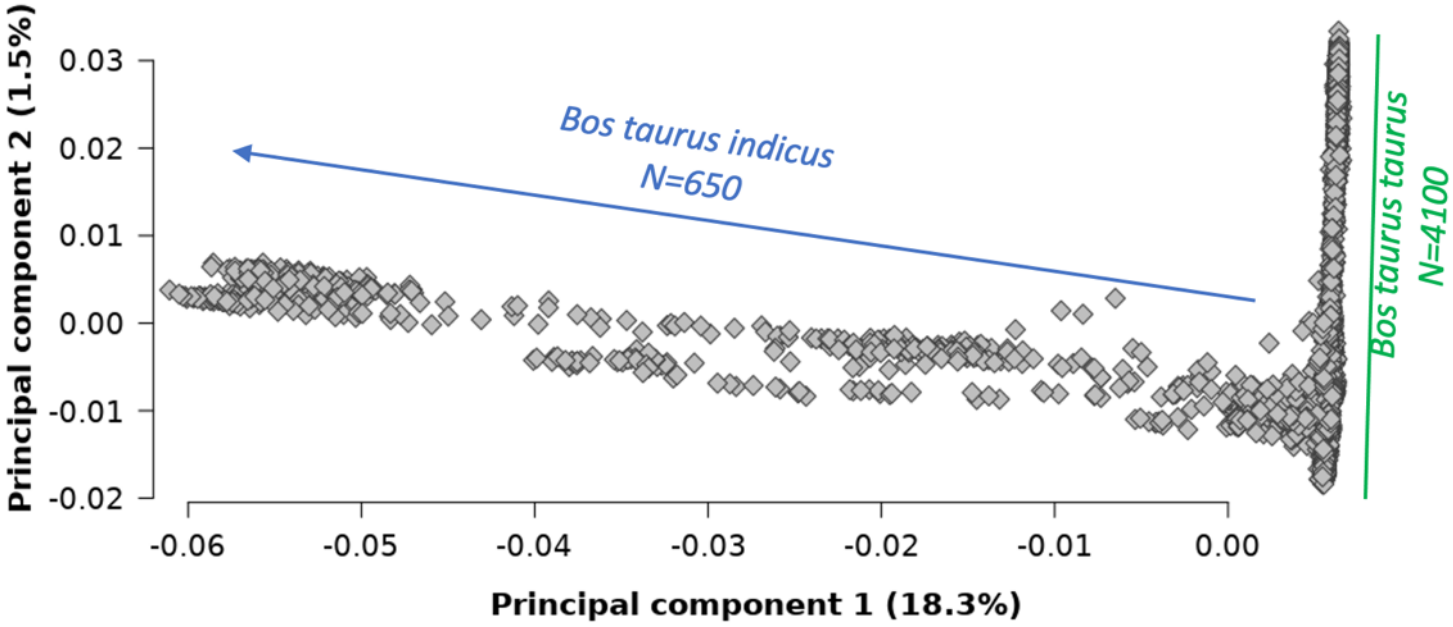
Thousands of animals

	SNP1 # of T's	SNP2 # of C's	SNP3 # of T's
1	2	1	2
2	1	2	1
3	0	0	0

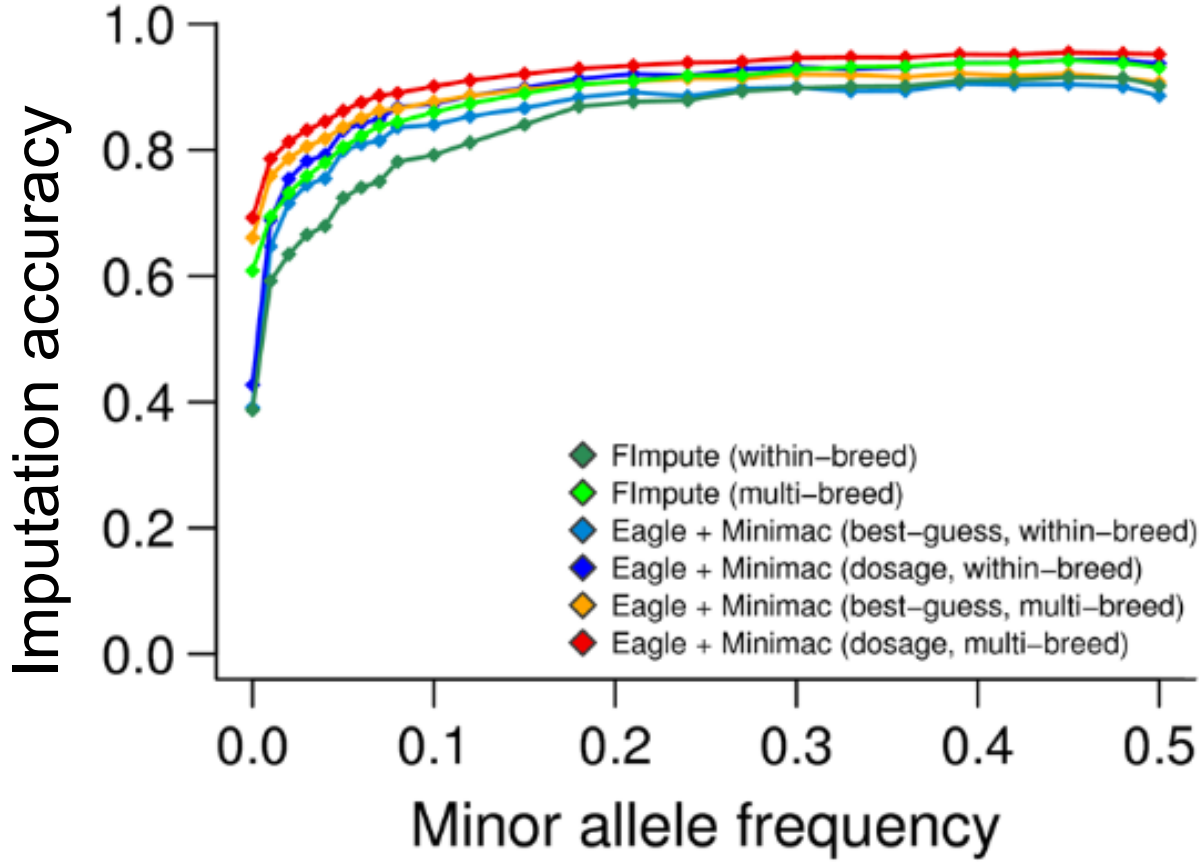


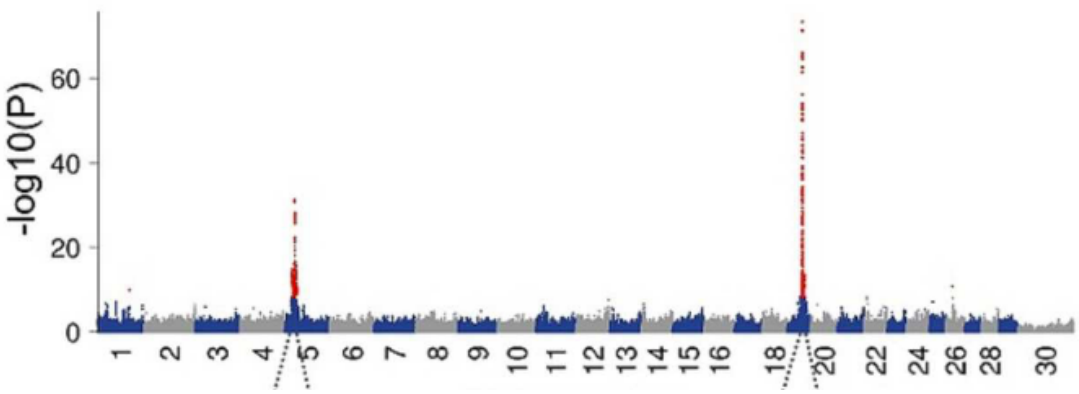
Reference-guided variant discovery revealed >140 million polymorphic sites in >4700 cattle

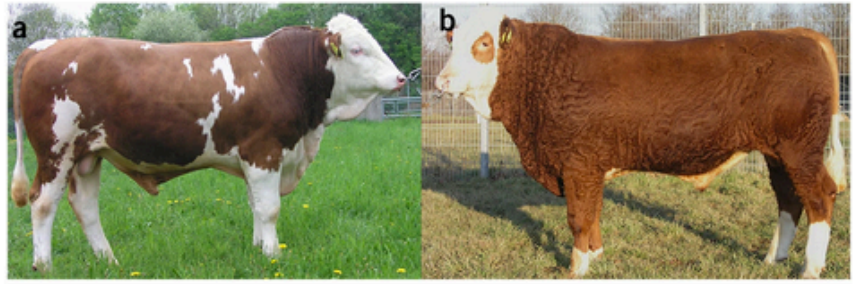
(1000 Bull Genomes Project)

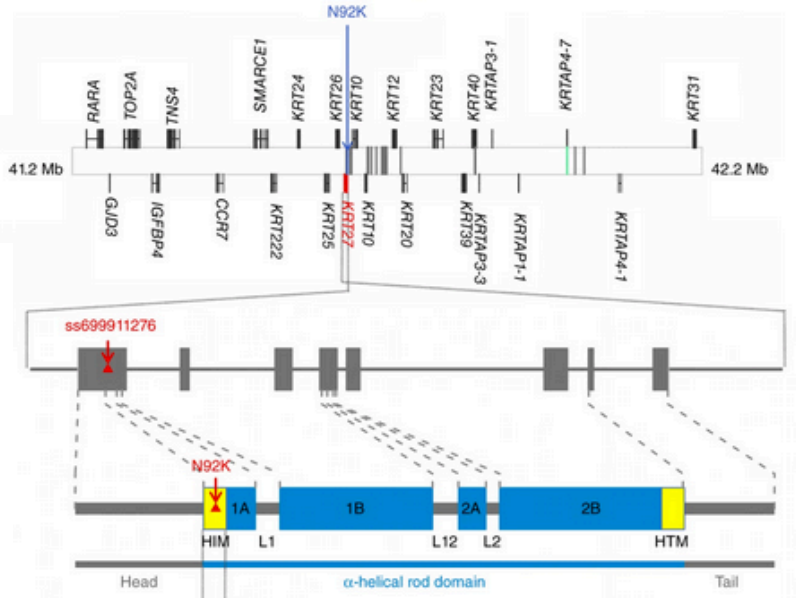


This catalogue contains causal variants









Daetwyler et al. ***Nature Genetics*** 2014
Pausch et al. ***Genet Sel Evol*** 2017

Reference-guided variant discovery is biased

- DNA sequencing reads that contain only alleles that match corresponding reference nucleotides are more likely to align correctly than DNA fragments that also contain non-reference alleles
- Reads originating from DNA fragments that are highly diverged from corresponding reference nucleotides will either obtain low alignment scores, or align at incorrect locations, or remain un-mapped
- Reads originating from DNA fragments that are missing in the reference genome remain unmapped. Reference-guided variant discovery is blind to these sites.

Reference allele bias

Paten et al. ***Genome Res*** 2017;27:665–76
Pritt et al. ***Genome Biol*** 2018;19:220
Sherman et al. ***Nat Genet*** 2019;51:30–35

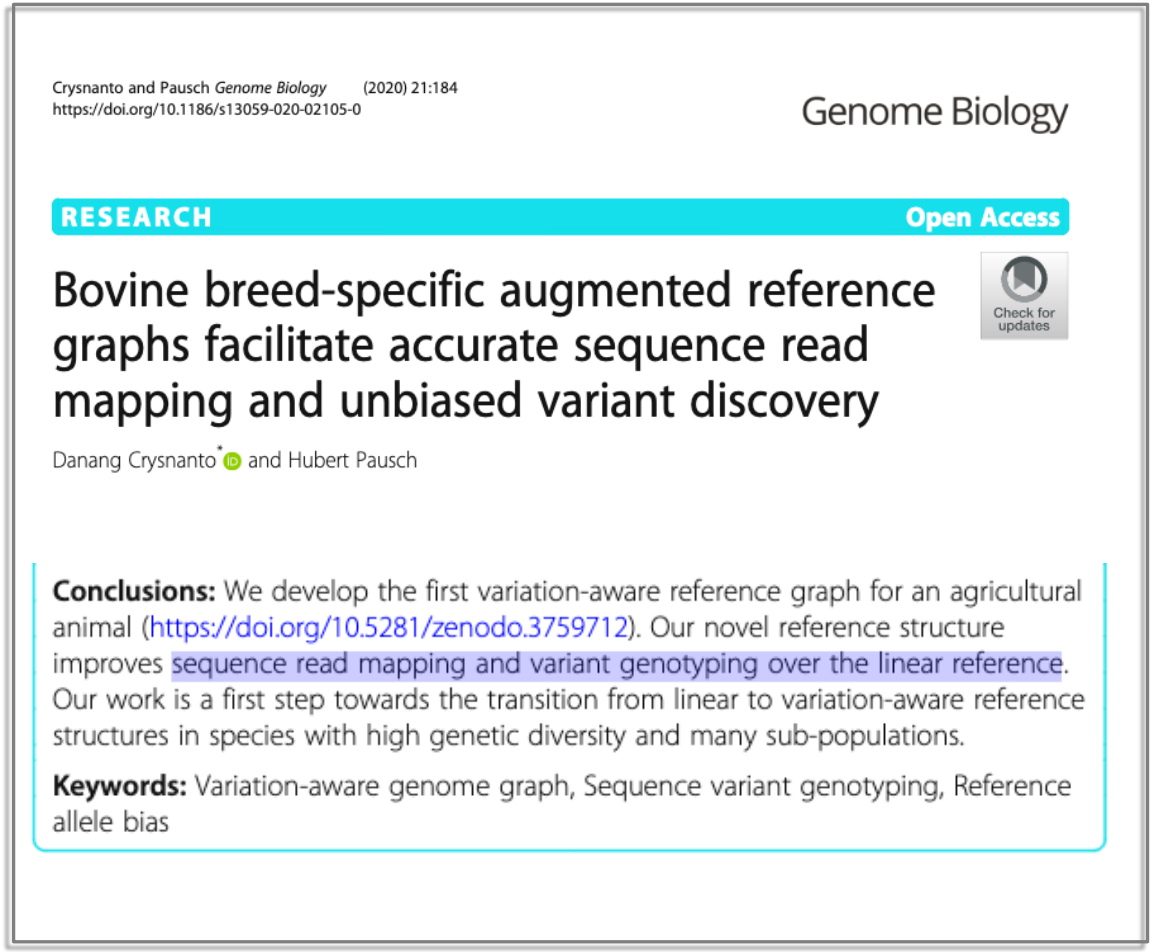
Strategies to mitigate reference allele bias

- *Personalized reference genomes*
  - Easy to implement
  - Standard genomic analysis tools work well
  - Coordinates remain unchanged
- *Graph-based references*
  - Relatively easy to implement
  - Standard genomic analysis tools work with workarounds
  - Variant prioritization is crucial
  - Coordinates remain unchanged

van de Geijn et al. **Nat Methods** 2015;12:1061–1063
Garrison et al. **Nat Biotechnol** 2018;36:875–81

A *Bos taurus* reference graph

- ARS-UCD1.2 used as backbone
- Augmented with variants (SNPs and Indels) that were prioritized based on allele frequency
  - Breed-specific genome graphs
  - Pan-genome graphs
- Read mapping and variant genotyping accuracy compared between graphs and the linear reference sequence



Crysnanto and Pausch *Genome Biology* (2020) 21:184
https://doi.org/10.1186/s13059-020-02105-0

Genome Biology

RESEARCH Open Access

Bovine breed-specific augmented reference graphs facilitate accurate sequence read mapping and unbiased variant discovery

Danang Crysnanto* and Hubert Pausch

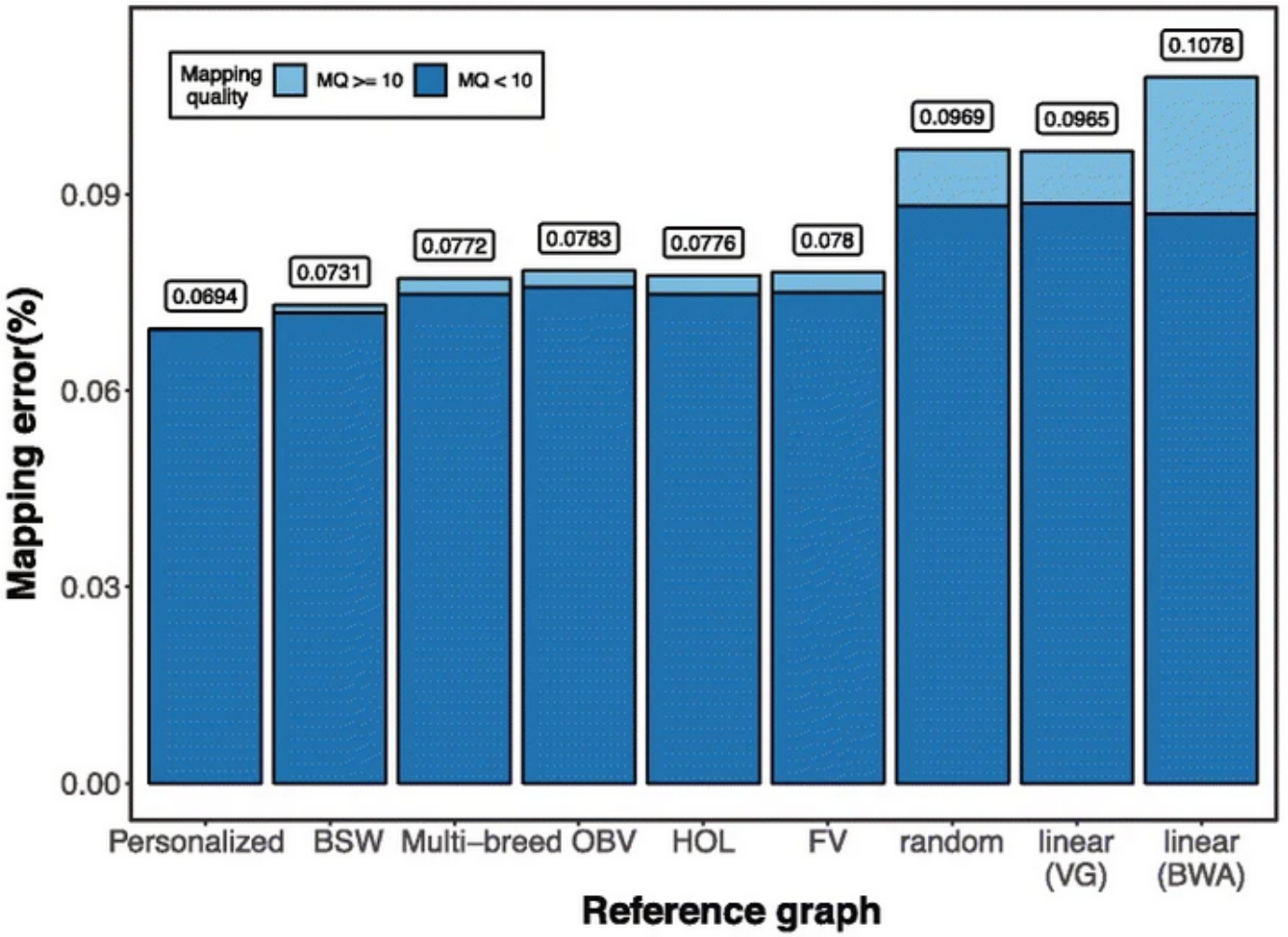
Conclusions: We develop the first variation-aware reference graph for an agricultural animal (https://doi.org/10.5281/zenodo.3759712). Our novel reference structure improves sequence read mapping and variant genotyping over the linear reference. Our work is a first step towards the transition from linear to variation-aware reference structures in species with high genetic diversity and many sub-populations.

Keywords: Variation-aware genome graph, Sequence variant genotyping, Reference allele bias

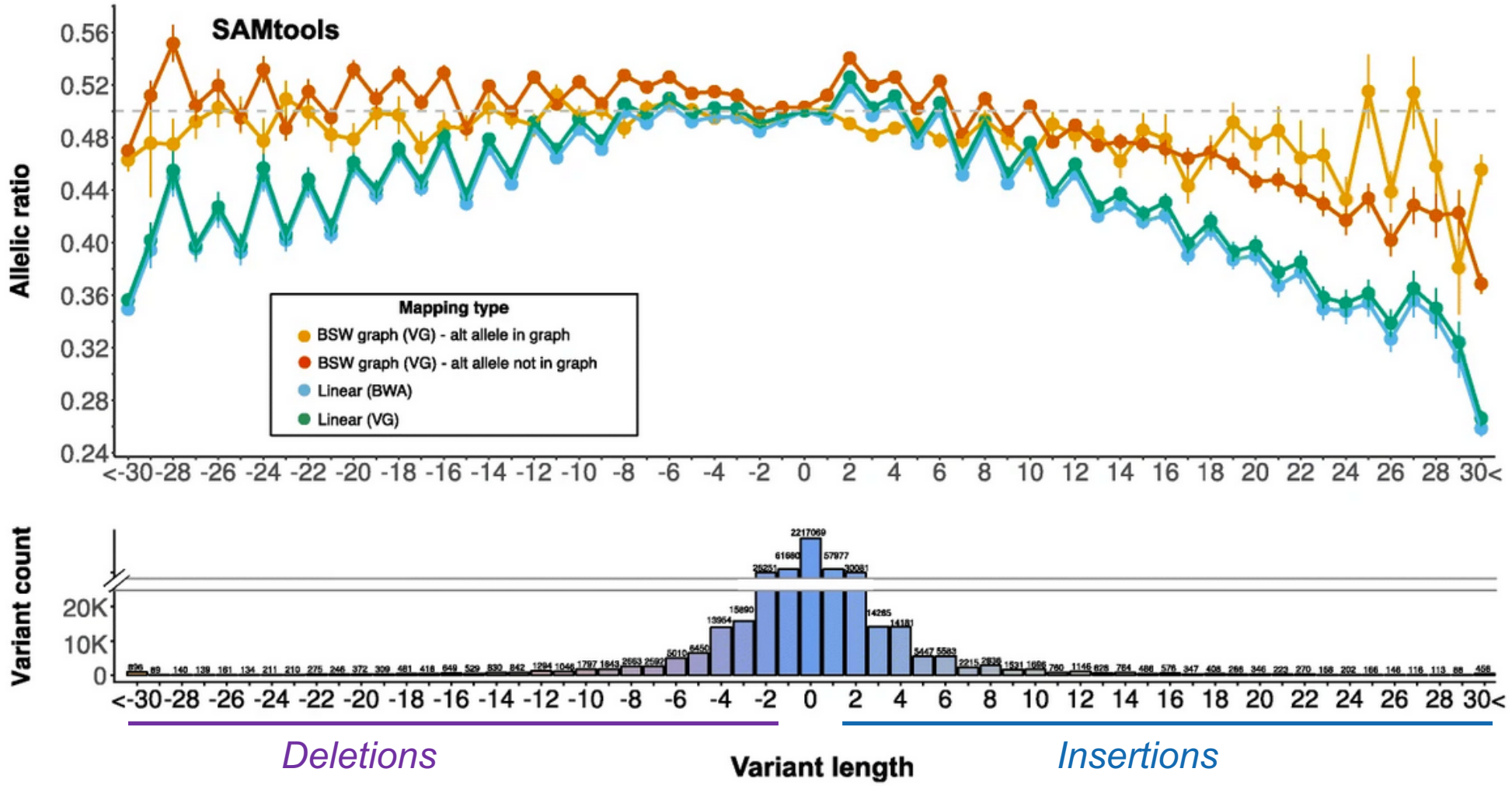
https://github.com/danangcrysnanto/bovine-graphs-mapping

Linear mapping is outperformed by graph-based mapping

Breed-specific graph reduced mapping errors by 30% compared to current gold standard (*BWA mem*)



Graph-based variant genotyping is unbiased



Options to mitigate reference allele bias

- *Personalized reference genomes*
  - Easy to implement
  - Standard genomic analysis tools work well
  - Coordinates remain unchanged
- *Graph-based references*
  - Relatively easy to implement
  - Standard genomic analysis tools work with workarounds
  - Variant prioritization is crucial
  - Coordinates remain unchanged
- *Multi-assembly graphs*
  - Full spectrum of sequence diversity accessible
  - New paradigm that requires novel tools (*minigraph*, *progressive cactus, pggb*)

Largely blind to structural variations

Li et al. **Genome Biol** 2020;21:265; Armstrong et al. **Nature** 2020:587;246-251; https://github.com/pangenome/pggb

The assembly of reference-quality assemblies is possible *at scale*

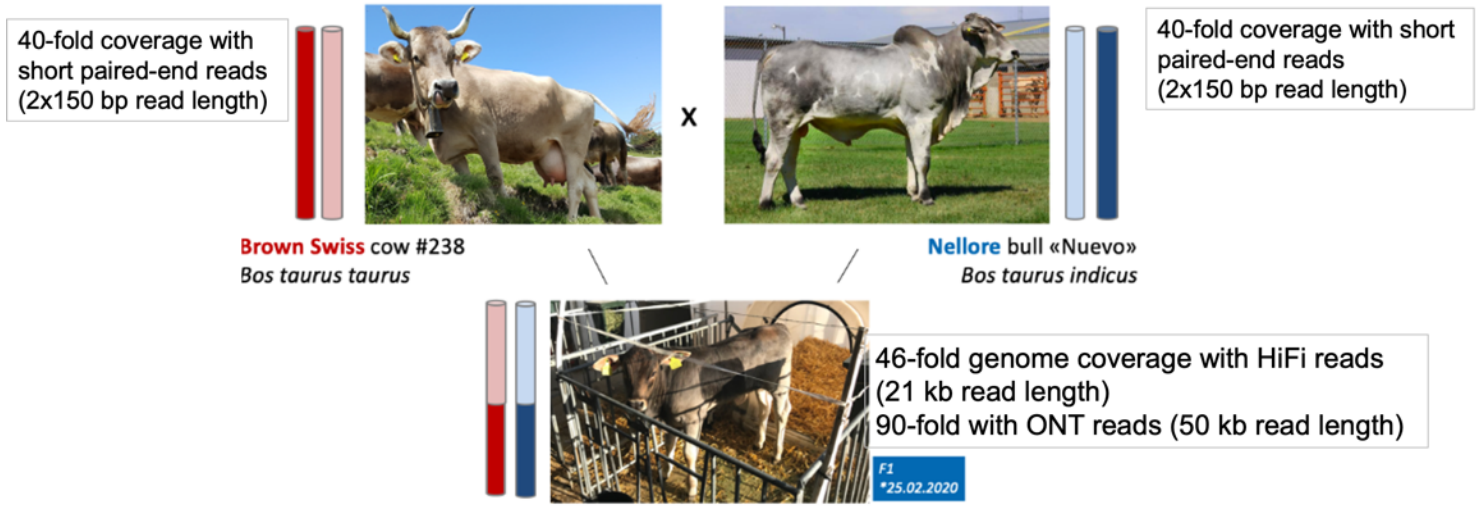
- Long-read sequencing is affordable
  - ONT, PacBio
- Sophisticated assembly algorithms have been developed
  - Trio-Binning (*Canu*, *hifiasm*, *GraphUnzip*)
  - Noisy reads (*Shasta*, *Raven*, *Ratatosk*)
- Long-read sequencing is highly accurate
  - PacBio Circular Consensus Sequencing (a.k.a. «HiFi-sequencing»)
  - PEPPER + Margin + DeepVariant for variant calling from ONT reads

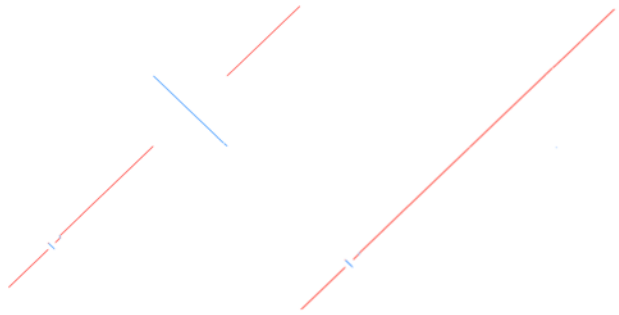
Koren et al. ***Nat Biotechnol*** 2018; Wenger et al. ***Nat Biotechnol*** 2019; Cheng et al. ***Nat Method*** 2021; Shafin et al. **BioRxiv**, 2021

Bovine HiFi-assemblies

AgroVet Strickhof

		contiguity	*correctness*	*completeness*		
Animal	**Size (gb)**	**Contigs**	**NG50 (mb)**	**k-QV**	**BUSCO (% complete)**	**Assembler**
Original Braunvieh (primary)	3.17	765	91	50.9	96.1	Hifiasm
Brown Swiss (F1 haplotype)	3.07	1036	86.6	46.2	95.9	Hifiasm
Nellore (F1 haplotype)	2.95	1177	94.7	45.6	93.3	Hifiasm

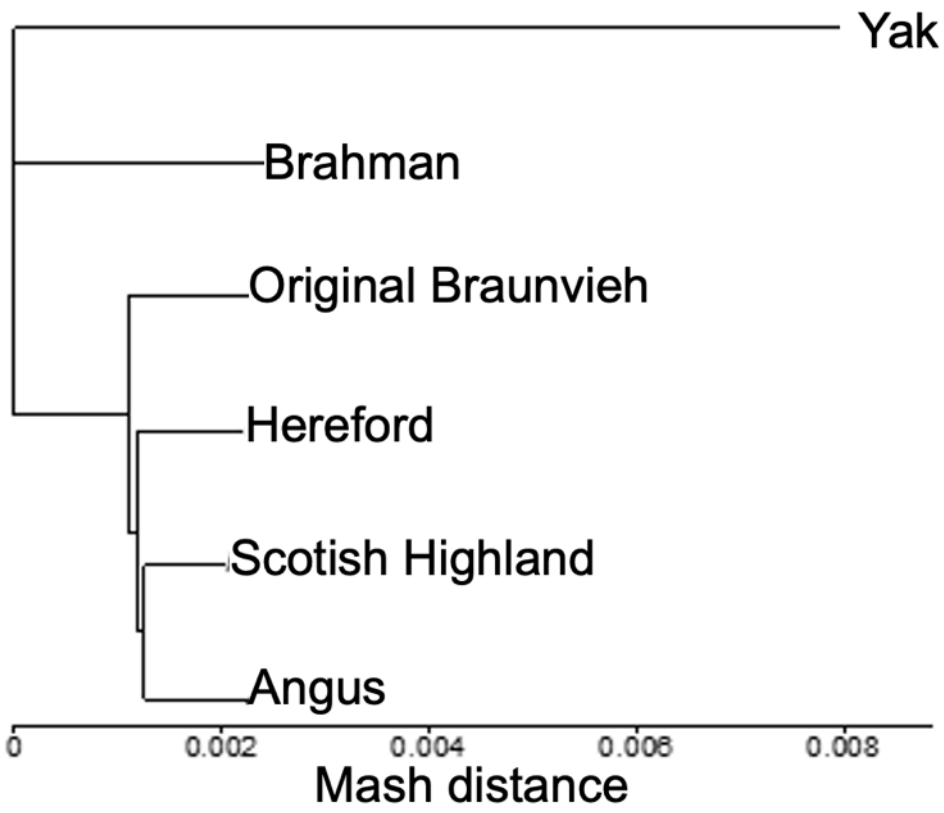




https://github.com/AnimalGenomicsETH/bovine-assembly

Establishing a bovine pangenome

(*Bos taurus taurus, Bos taurus indicus, Bos grunniens*)



Novel functional sequences uncovered through a bovine multi-assembly graph
bioRxiv https://doi.org/10.1101/2021.01.08.425845

The pangenome uncovers novel functional sequences



Cold Spring Harbor Laboratory

bioRxiv

THE PREPRINT SERVER FOR BIOLOGY

HOME | ABOUT | SUBMIT | NEWS & NOTES | ALERTS / RSS | CHANNELS

Search

Advanced Search

bioRxiv is receiving many new papers on coronavirus SARS-CoV-2. A reminder: these are preliminary reports that have not been peer-reviewed. They should not be regarded as conclusive, guide clinical practice/health-related behavior, or be reported in news media as established information.

New Results

Comment on this paper

Previous

Next

Novel functional sequences uncovered through a bovine multi-assembly graph

Posted January 08, 2021.

Danang Crysnanto, Alexander S. Leonard, Zih-Hua Fang, Hubert Pausch

doi: https://doi.org/10.1101/2021.01.08.425845

This article is a preprint and has not been certified by peer review [what does this mean?].

Download PDF

Email

Supplementary Material

Share

Citation Tools

XML

Abstract | Full Text | Info/History | Metrics | Preview PDF

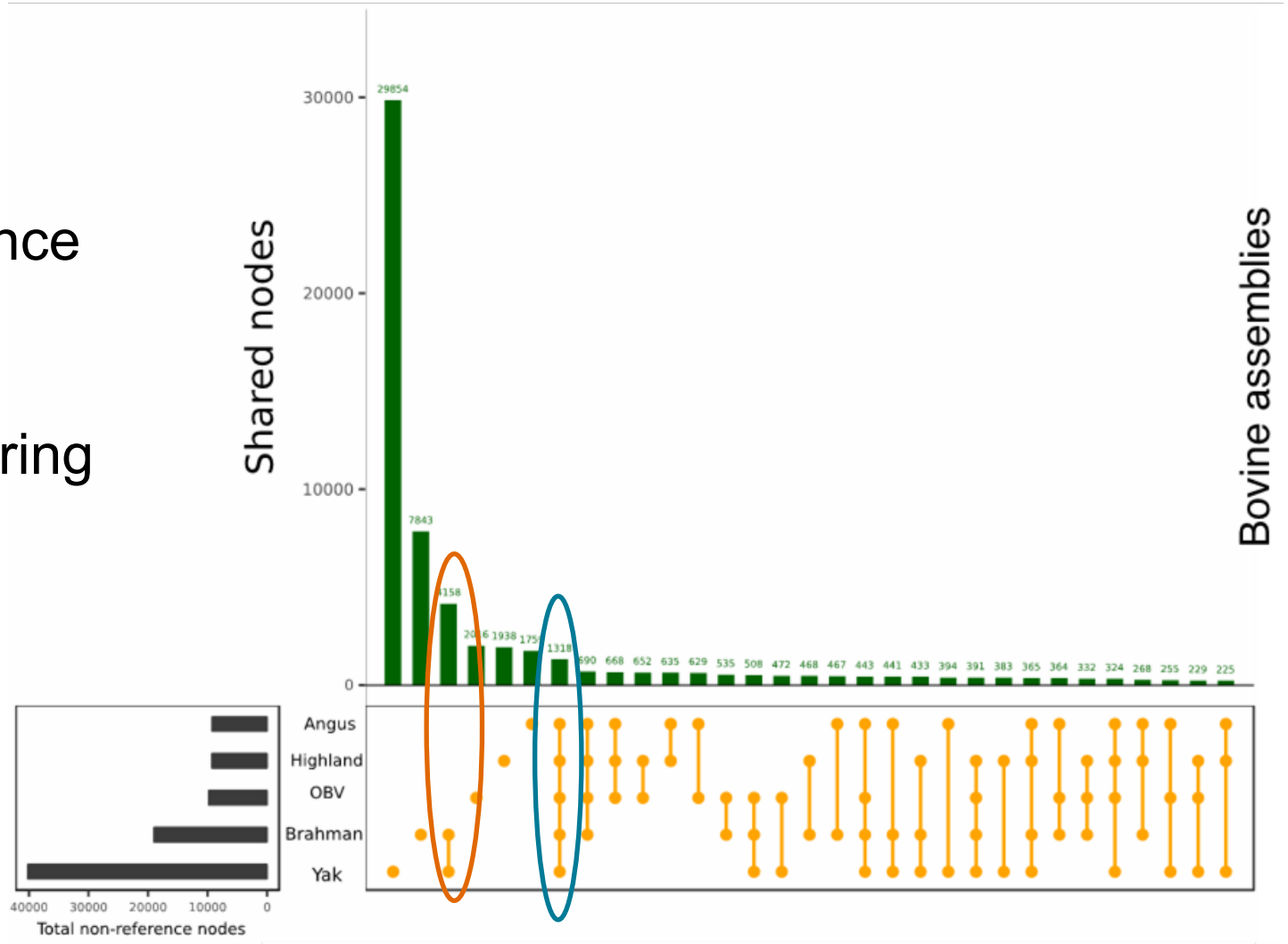
Tweet

Like 0

https://github.com/AnimalGenomicsETH/bovine-graphs

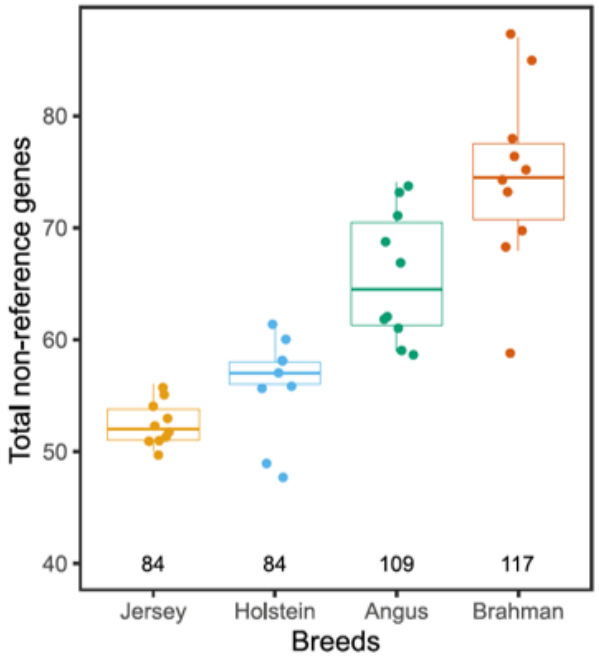
Structural variants (*non-reference nodes*) of the multi-assembly graph contain 70 million bases

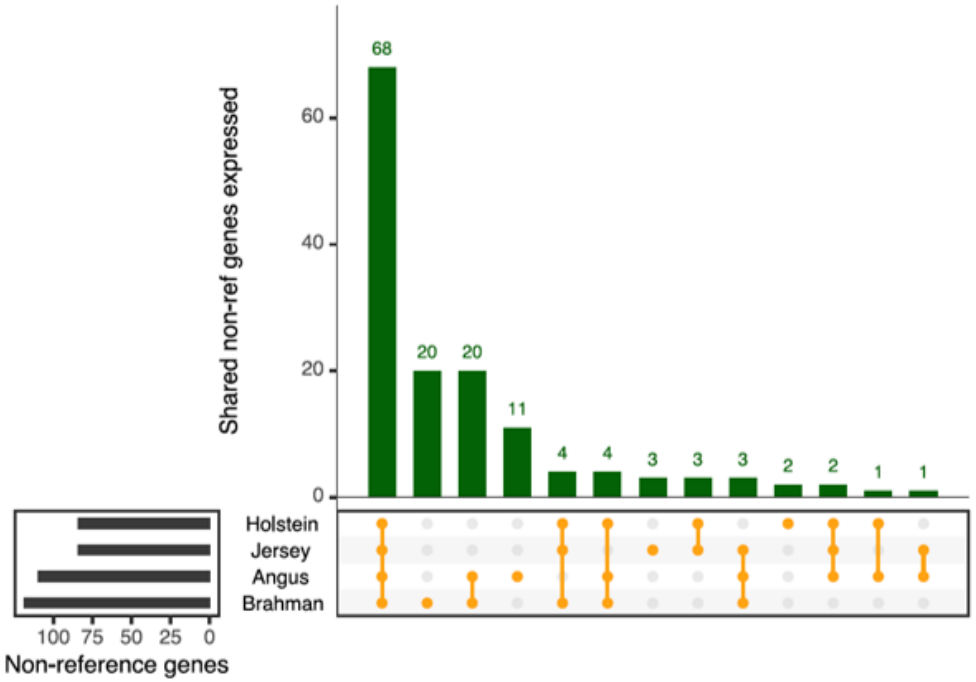
- Identify nonreference sequence
  - Evolutionary relationships?
  - Reference bias?
- Highlights areas worth exploring



Novel sequences contain novel genes

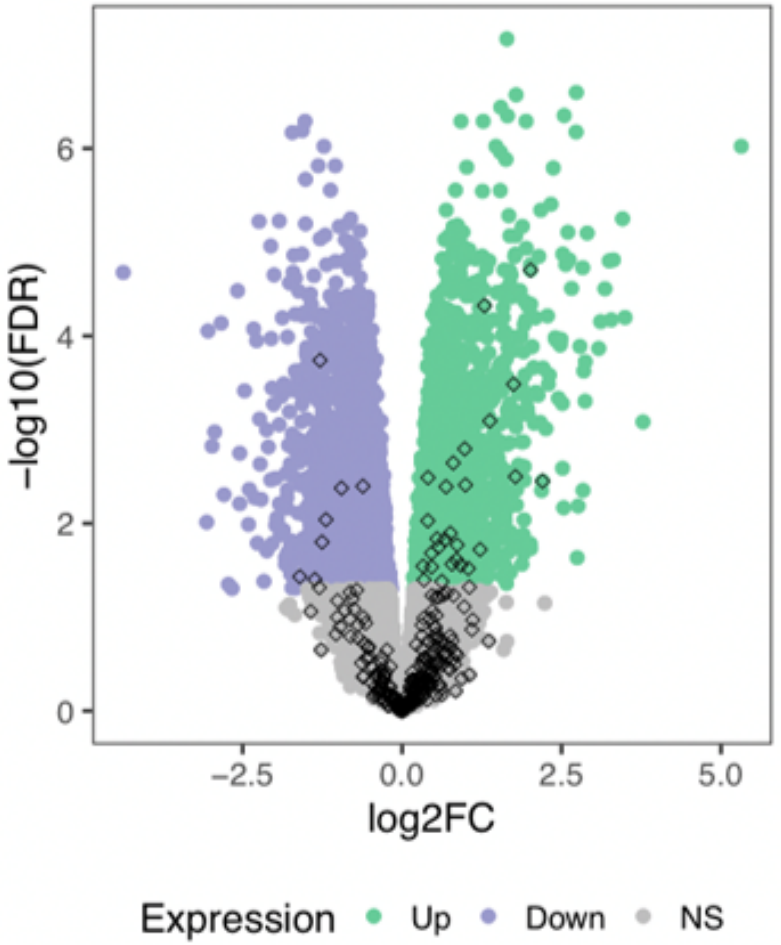
- Gene models were predicted using Augustus & BLASTP
- Liver transcriptomes from 45 cattle provide additional experimental evidence





Novel genes are differentially expressed

- We revisit an analysis from McLoughlin et al (*Frontiers in immunology* 5, 396. (2014))
  - Transcriptome data of peripheral blood leukocytes of 16 cattle
    - Eight *Mycobacterium bovis*-infected
    - Eight unaffected controls
- The novel sequences contain 36 putatively novel genes that are differentially expressed
  - The top down-regulated novel gene encodes leukocyte immunoglobulin-like receptor A5 (LILRA5) which is missing in ARS-UCD1.2



Conclusions

- Linear reference genomes are incomplete and susceptible to reference allele bias
- Haplotype-resolved assemblies are required to accurately detect alleles at structural variations
- Novel reference structures such as multi-assembly graphs make «non-reference sequences» amenable to genetic investigations

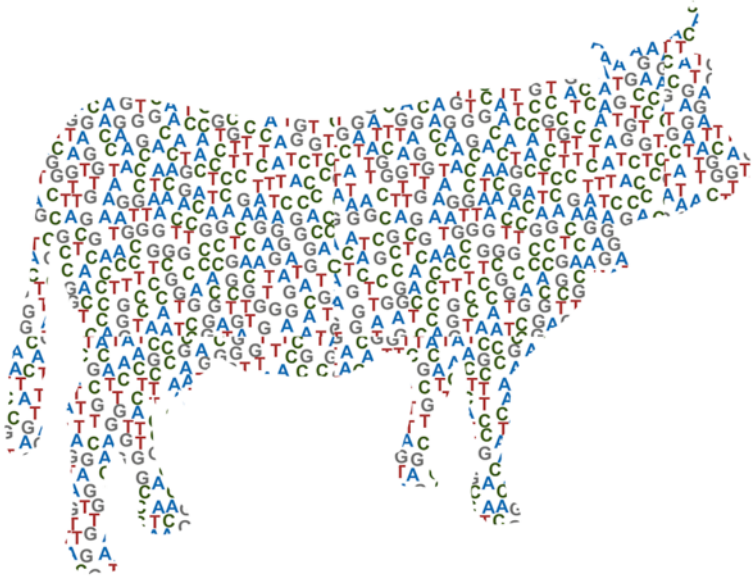
Contributors, collaborators & funding

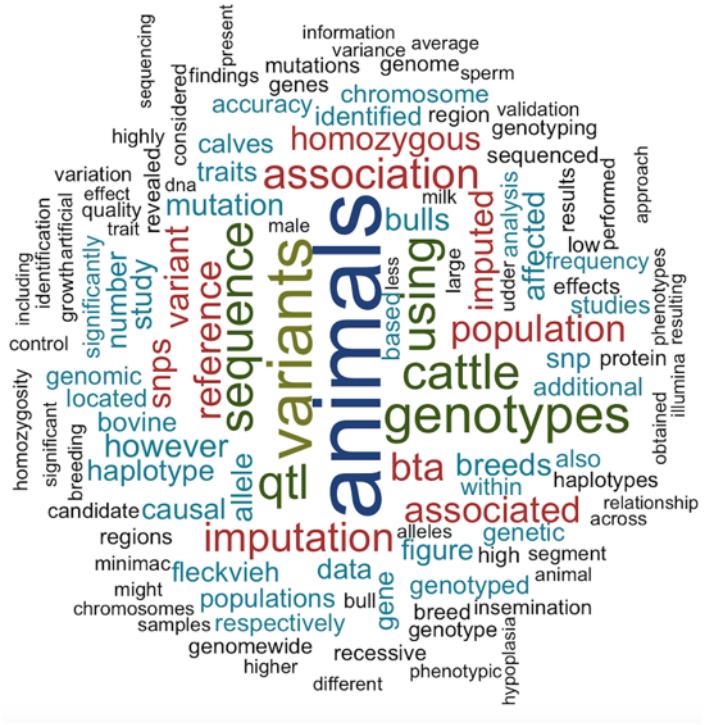
- Danang Crysnanto, Alexander Leonard, Zih-Hua Fang, ETH Animal Genomics

- Ben Hayes, Hans Daetwyler, Michael Goddard, Brisbane / Melbourne, Australia, for the *1000 Bull Genomes Consortium*
- Ben Rosen, Derek Bickhart, Tim Smith, USDA, USA, for the *Bovine Pangenome Consortium*

AgroVet Strickhof
Eine Kooperation in Bildung und Forschung: strickhof ETH zürich University of Zurich
ETH zürich
University of Zurich
Functional Genomics Center Zurich
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
Federal Office for Agriculture FOAG
FNSNF
Fonds national suisse
Schweizerischer Nationalfonds
Fondo nazionale svizzero
Swiss National Science Foundation
swissgenetics

Thank you for your attention





ETH Zürich
Hubert Pausch
Animal Genomics
Eschikon 27 | EHB E 21
CH-8315 Lindau

hubert.pausch@usys.ethz.ch
www.ag.ethz.ch