Using the new ARS-UCD1 reference map in US genomic predictions

Dan Null[1]*, Derek Bickhart[2], Paul VanRaden[1], John Cole[1], Jeff O'Connell[3], and Ben Rosen[1]

[1] Animal Genomics and Improvement Laboratory, ARS, USDA, Beltsville, MD, USA

[2] Cell Wall Biology and Utilization Research , ARS, USDA, Madison, WI, USA

[3] University of Maryland School of Medicine, Baltimore, MD, USA

Email: daniel.null@ars.usda.gov

Web site: https://aipl.arsusda.gov/

USDA

Topics

- **Strategies to liftover SNP locations from one map to another**

- **Compare previous UMD3 map to new ARS-UCD1 map**

  - **Imputation results from 5 breeds (HOL, JER, BSW, RDC, GUE)**

  - **Editing SNPs from potentially mismapped regions**

  - **Inheritance and consistency of haplotypes across generations**

- **US implementation of the ARS-UCD1 map**

- **Further research with HD or sequence data**

USDA

Liftover: Converting locations to a different map

- **Each chromosome may be longer or shorter than previous assembly**

  – **Autosomes range from -4.3% shorter (chr12) to 0.6% longer (chr26)**

  – **X chromosome 6.6% longer than UMD3, with more distinct PAR**

- **Many insertions, deletions, inversions, translocations, SNPs vs. UMD3**

  – **Official liftover tool developed by the UCSC Genome Browser (scaffolds)**

  – **Use probe from manifest or flanking sequence, align to new map**

  – **Simulate paired end reads from old map, align to new map (very fast)**

  – **Use partial matches within a single chromosome**

USDA

Liftover of SNP locations

- **Probes used for genotyping SNPs on arrays**

  - **A 50-base sequence extended to the right or left of SNP, such as**

  - **[C/T]AGTCAGCTCTGTGGCCTGGGCAGGTTCCCAGGATATTCCAGCCAGAC**

  - **Where was this located on the old map?**

  - **Where is this located on the new map?**

- **File of new SNP locations provided by Bob Schnabel (U Missouri)**

  - **https://www.animalgenome.org/repository/cattle/UMC_bovine_coordinates/**

  - **Liftover of other SNPs not on manifests was done at AGIL**

USDA

SNPs now located on different chromosomes


Whole Genome
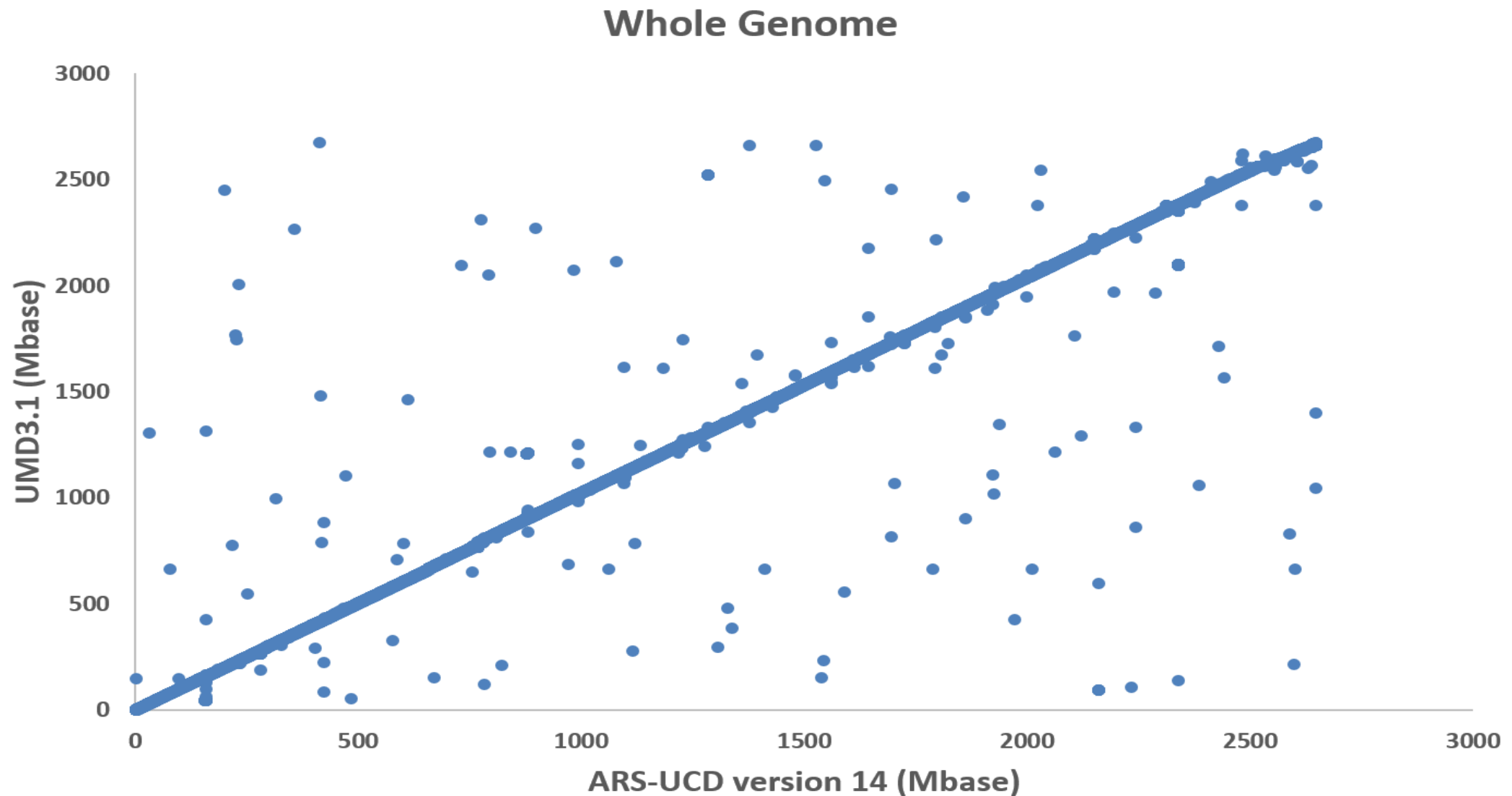
Switching from UMD3 to ARS-UCD1 map

- **Better assembly of Dominette DNA using longer reads**

- **UMD3: Zimin et al. 2009 Genome Biology 10:R42**

- **ARS-UCD1: Rosen et al. 2018 WCGALP, vol. Molecular Genetics 3, p. 802**

  – **https://www.ncbi.nlm.nih.gov/assembly/GCA_002263795.2/**

  – **Annotation (gene structure) released by NCBI in May 2018**

  – **Also used for 1000 Bull Genomes Project Run7 in 2019**

- **Edited versions used by AGIL after mis-mapped SNPs removed**

USDA

PAR-X comparison
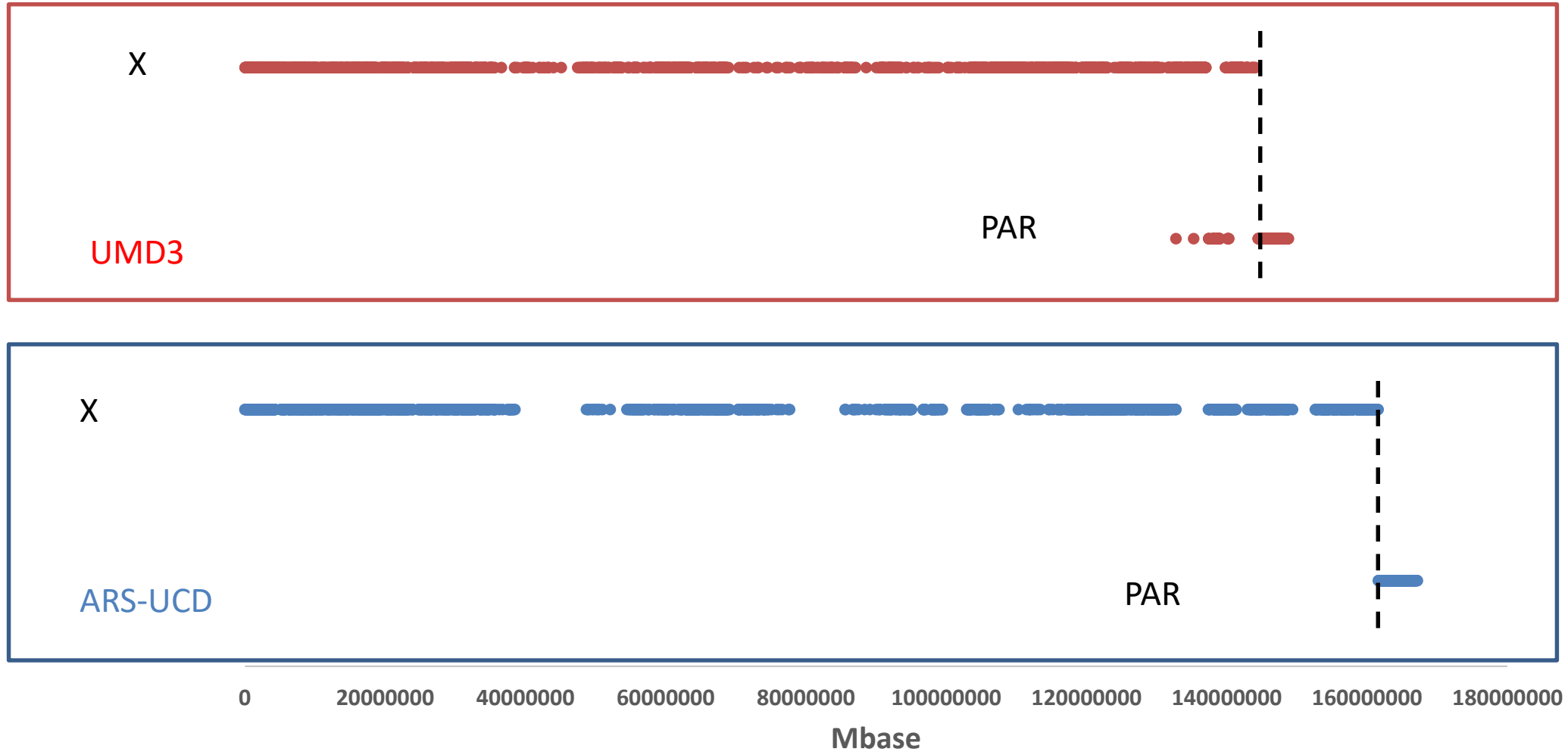
Imputation tests

- **Locations for the 60,367 usable SNPs were converted to ARS-UCD1**

- **Genotypes were imputed to 60K from 30 chips of differing density**

- **Genotypes from 5 breeds were imputed separately:**

  - **1,748,453 Holsteins (HO)**

  - **215,800 Jerseys (JE)**

  - **32,724 Brown Swiss (BS)**

  - **4,834 Ayrshires (AY)**

  - **3,517 Guernseys (GU)**

USDA

60K Imputation results – 1: Inheritance rate

Haplotypes with parent-progeny noninheritance (%)			
Breed	UMD3	ARS-UCD1	
	edited	original	edited
HOL	1.6	1.3	1.1
JER	4.3	3.9	3.8
BSW	1.5	1.3	1.2
RDC	1.8	1.5	1.4
GUE	1.6	1.4	1.3

Maximum non-inheritance within any haplotype (%)			
Breed	UMD3	ARS-UCD1	
	edited	original	edited
HOL	13	31	12
JER	25	35	18
BSW	17	21	10
RDC	16	27	10
GUE	18	16	9

USDA

60K Imputation results – 2: Maximum haplotypes

Maximum number of haplotypes per segment			
Breed	**UMD3**	**ARS-UCD1**	
	edited	original	edited
HOL	47,987	81,316	36,690
JER	39,628	33,034	29,732
BSW	11,602	12,846	9,447
RDC	2,030	2,512	1,606
GUE	1,970	1,657	1,427

- **Haplotype counts and inheritance improved for all breeds**

- **Much better on the X chromosome and its pseudoautosomal region**

- **Further edits used HD data to identify smaller regions potentially mismapped**

USDA

Imputation for published gene tests / haplotypes

- **The new map more cleanly separated truly lethal haplotypes from false candidate haplotypes.**

- **Gene tests were added within previously published haplotypes:**

  - **Added HH0, HH3, HH4, Red, Black/Red, Dominant Red, BH2, AH1, AH2.**

  - **HH1, JH1, BLAD, CVM, DUMPs, Mulefoot, SDM, SMA, and Weaver were already included in previous 60K list. Some of these were progeny tests.**

- **The added gene tests tended to identify more carrier animals.**

- **For most haplotypes >99% of animals maintained the same status.**

USDA

Implementation steps

- **Converted UMD3 locations to ARS-UCD1 SNP locations**

  - **Detected a few SNPs with different names, locations actually were same**

  - **Detected a few SNPs with different locations, but names were the same**

- **Increased usable SNP list from 60K to 80K (Wiggans ADSA talk)**

- **Imputed genotypes for all breeds and crossbreds**

- **Checked published haplotypes to ensure proper inheritance**

- **Obtained prior allele effect estimates for all breeds, traits, and BBR**

- **Implemented during the December 2018 full release**

USDA

Published gene tests / haplotypes - HOL

Test	Same	Changed to:		Freq
	(%)	Carrier	Non-C	(%)
HH0	99.8	0.17	0.05	3.2
HH1	99.7	0.24	0.05	2.6
HH2	99.3	0.71	0.02	2.6
HH3	99.5	0.46	0.03	4.6
HH4	99.9	0.02	0.01	0.5
HH5	98.2	1.69	0.15	6.2
HH6	New			0.5

Test	Same	Changed to:		Freq
	(%)	Carrier	Non-C	(%)
BLAD	99.9	0.03	0.05	0.2
CVM	99.1	0.68	0.14	1.9
DUMPS	100.0	0.00	0.00	<0.1
MuleFt	99.9	0.02	0.01	0.1
Polled	99.2	0.57	0.18	3.8
Red	97.1	1.33	1.53	9.4
Blk/Red	99.7	0.21	0.08	1.2
DomRed	99.9	0.03	0.00	0.2
HCD	99.1	0.26	0.61	5.9

USDA

Published gene tests / haplotypes – Other breeds

Test	Same	Changed to:		Freq
	(%)	Carrier	Non-C	(%)
BSW:				
BH2	99.4	0.42	0.20	13.3
SDM	98.7	0.92	0.37	3.0
SMA	99.3	0.66	0.06	4.0
Polled	99.6	0.07	0.31	2.5
Weaver	99.9	0.07	0.06	1.2

Test	Same	Changed to:		Freq
	(%)	Carrier	Non-C	(%)
JER:				
JH1	98.4	1.27	0.29	18.4
JHP	99.0	0.17	0.79	4.1
RDC:				
AH1	98.2	1.55	0.23	22.2
AH2	99.0	0.69	0.35	21.0

USDA

HD Imputation results – 1: Inheritance rate

Haplotypes with parent-progeny noninheritance (%)		
	ARS-UCD1	
Breed	60K edits	HD edits
HOL	3.0	2.9
JER	3.9	4.9
BSW	2.3	2.2
RDC	2.4	2.4
GUE	1.5	1.4

Maximum non-inheritance within any haplotype (%)		
	ARS-UCD1	
Breed	60K edits	HD edits
HOL	48	18
JER	47	22
BSW	30	17
RDC	37	18
GUE	27	13

USDA

Check SNP correlation problems using heat maps

End of chromosome 29

Before HD edits
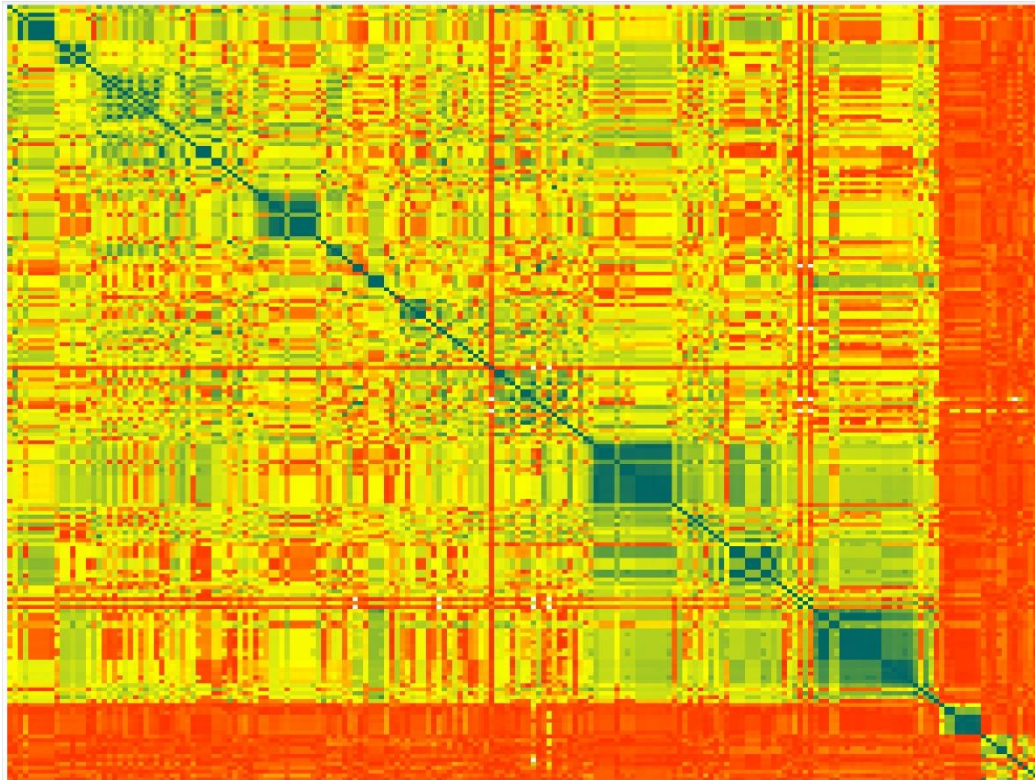
After HD edits

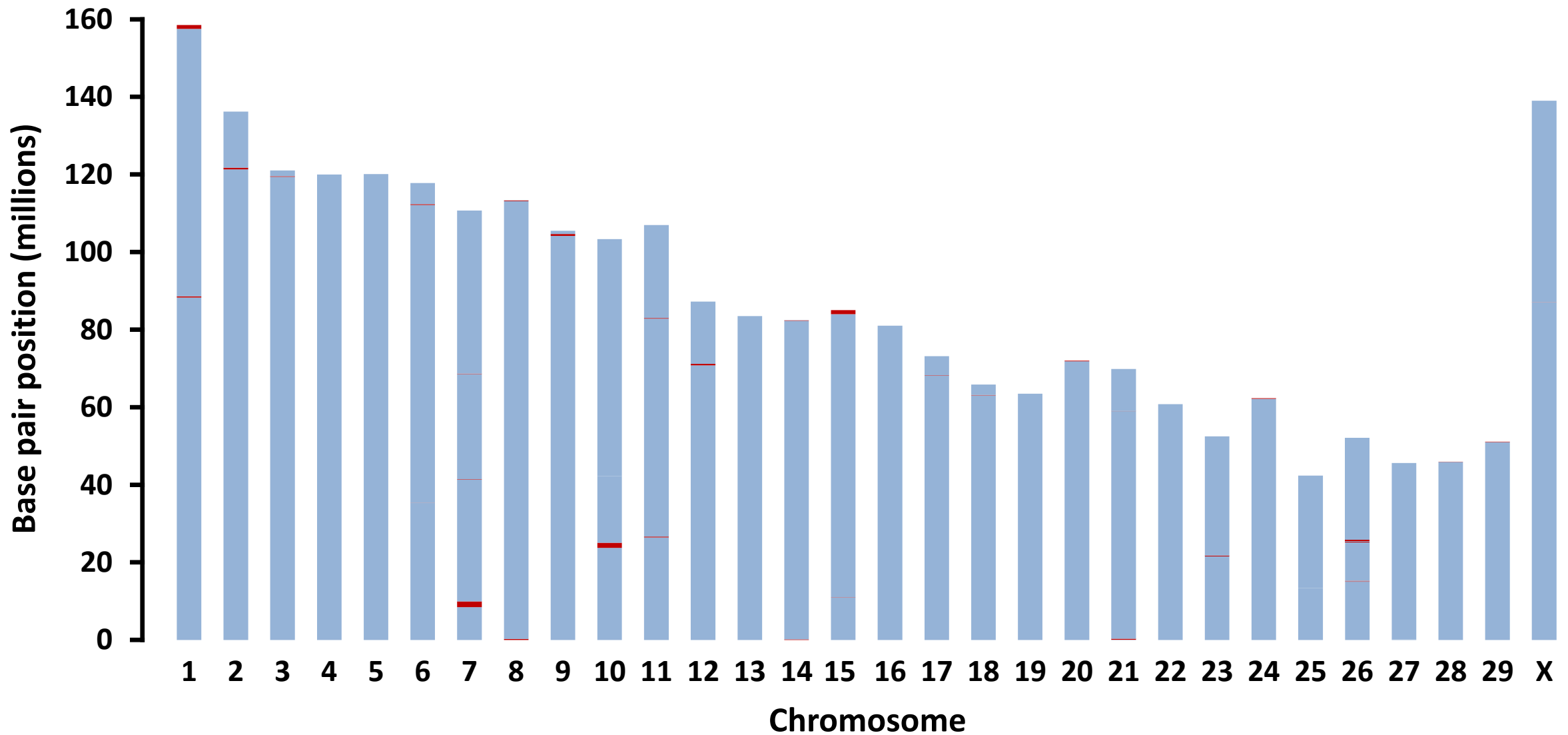37% Non-inheritance

4% Non-inheritance

USDA
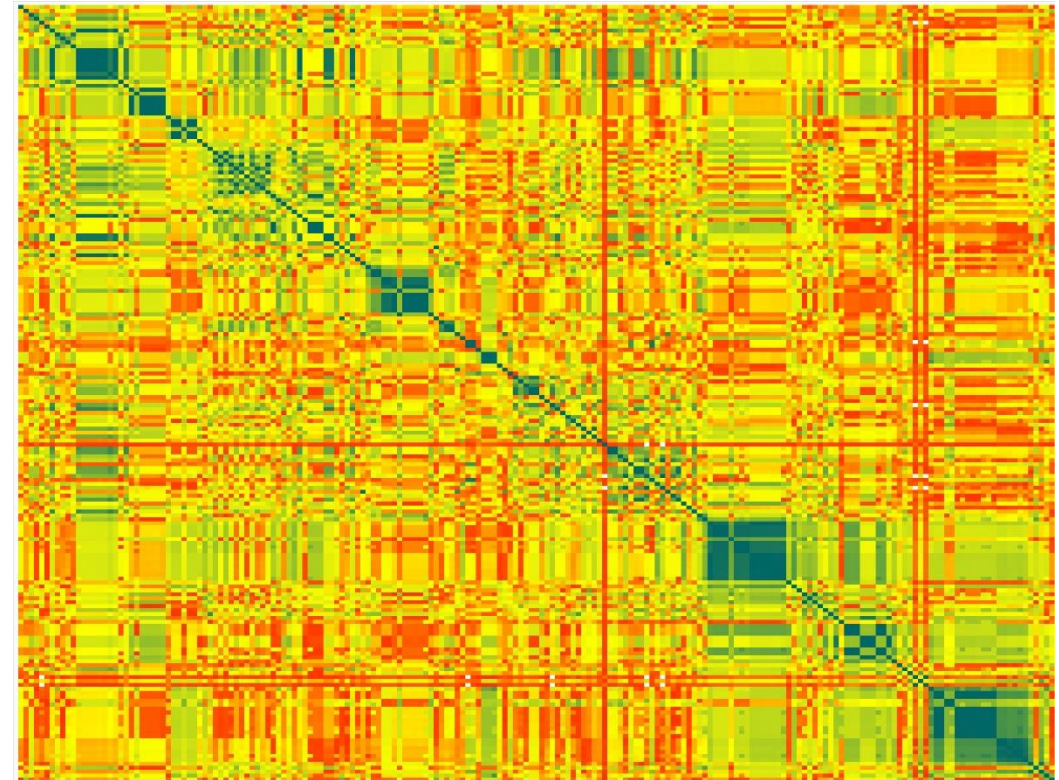
Potentially mismapped sections edited

Potentially mismapped sections edited

Chromosome	Mismapped Region Starting Location	Ending Location	Size (Mb)	Chromosome	Mismapped Region Starting Location	Ending Location	Size (Mb)	Chromosome	Mismapped Region Starting Location	Ending Location	Size (Mb)
1	88,288,254	88,589,699	0.30	10	23,703,142	25,059,514	1.36	21	58,940,778	59,023,509	0.08
1	157,525,745	159,000,000	1.47	10	42,156,286	42,292,668	0.14	23	21,510,406	21,737,841	0.23
2	121,316,059	121,751,997	0.44	11	26,412,888	26,626,298	0.21	24	62,162,854	63,000,000	0.84
3	119,378,630	119,559,889	0.18	11	82,836,074	83,006,400	0.17	25	13,347,316	13,376,441	0.03
4	1,336,110	1,417,557	0.08	12	70,762,931	71,197,493	0.43	26	14,965,770	14,988,648	0.02
6	35,300,301	35,344,991	0.04	14	1	146,714	0.15	26	15,079,004	15,241,107	0.16
6	112,109,597	112,324,326	0.21	14	82,285,783	82,366,657	0.08	26	25,236,723	25,297,707	0.06
7	8,433,502	9,956,060	1.52	15	10,979,970	11,040,240	0.06	26	25,495,761	25,798,796	0.30
7	41,323,000	41,434,573	0.11	15	83,996,174	86,000,000	2.00	27	1	139,701	0.14
7	68,528,026	68,619,931	0.09	17	68,078,456	68,217,960	0.14	28	45,859,111	46,000,000	0.14
8	1	234,625	0.23	18	63,017,258	63,135,541	0.12	29	50,977,673	52,000,000	1.02
8	113,096,592	113,273,051	0.18	20	71,850,046	72,000,000	0.15	31	86,996,907	87,035,672	0.04
9	104,172,188	104,634,958	0.46	21	1	454,527	0.45				

USDA

Conclusions

- **The ARS-UCD map fixes many defects in the UMD3 map**

- **Mis-mapped regions of UMD3 are now usable on ARS-UCD**

- **The new map has several better properties than previous:**

  - **Marker locations, genotype imputation, haplotype inheritance, sequence alignment, and gene annotation**

- **ARS-UCD1 map used in U.S. genomic evaluations since Dec 2018**

- **ARS-UCD1 map used in 1000 Bull Genomes run7 sequence data files delivered in 2019**

USDA

Acknowledgments

- **USDA-ARS project 1265-31000-101-00, "Improving Genetic Predictions in Dairy Animals Using Phenotypic and Genomic Information" (AGIL funding)**

- **Council on Dairy Cattle Breeding (CDCB) and its industry suppliers for data**

USDA