

Tools and resources for accurate imputation of cattle sequence

Robert Schnabel
BIF Genetic Prediction
Kansas City, MO
12/19/2023

Acknowledgements

Troy Rowan: Imputation

Jenna Kalleberg Ridge: DeepVariant, Cue

Jacob Rissman: CNV

Hunter McConnell: Ancestral alleles

2016 NIH **Male Fertility** Sutovsky (PI), Taylor & Schnabel

- Linking Fertility-Associated Gene Polymorphisms to Aberrant Sperm Phenotypes

2018 USDA/NIFA **Bee** Elsie (PI) & Schnabel

- Identifying Genomic Regulatory Variants Associated with Resistance Traits in Honey Bee

2019 NIH **SCGE** Wells (PI), Prather, Safranski, Green & Schnabel

- Swine Somatic Cell Genome Editing Center

2020 USDA/NIFA **Cow eQTL** Schnabel (PI) & Decker

- Identification of Expression QTL Associated With Feed Efficiency in Beef Cattle

2020 USDA/NIFA **Heifer Fertility** Decker (PI) & Schnabel

- Genomics of puberty and fertility in heifers focusing on functional variants

2020 USDA/NIFA **SV/Pangenome** Brown (PI), Schnabel & Monsour

- Tools and resources for cattle pangenomics

2021 USDA/NIFA **Pig Imputation** Huang (PI), Schnabel, Steibel & Gondro

- FACT: SWIM - a cyber-enabled swine genome imputation framework and publicly accessible server for nucleotide resolution genetic mapping

2023 USDA/NIFA **GIAB** Schnabel, Murdoch, Kalleberg-Ridge

- PARTNERSHIP: Development of Genomic Reference Materials for Cattle

- “We don’t need no stinking genes!”
Curt Van Tassell, PAG 2008
- “Genotypes are actually phenotypes.” [paraphrased]
Mark Thallman, sometime around 2001
- “I shall try not to use statistics as a drunken man uses lamp-posts, for support rather than for illumination; and I shall try not to let my pen stray too far from the tethers of sanity of things seen...”
Andrew Lang, ~1937







Imputation

Rowan et al. *Genet Sel Evol* (2019) 51:77
<https://doi.org/10.1186/s12711-019-0519-x>

RESEARCH ARTICLE

Open Access

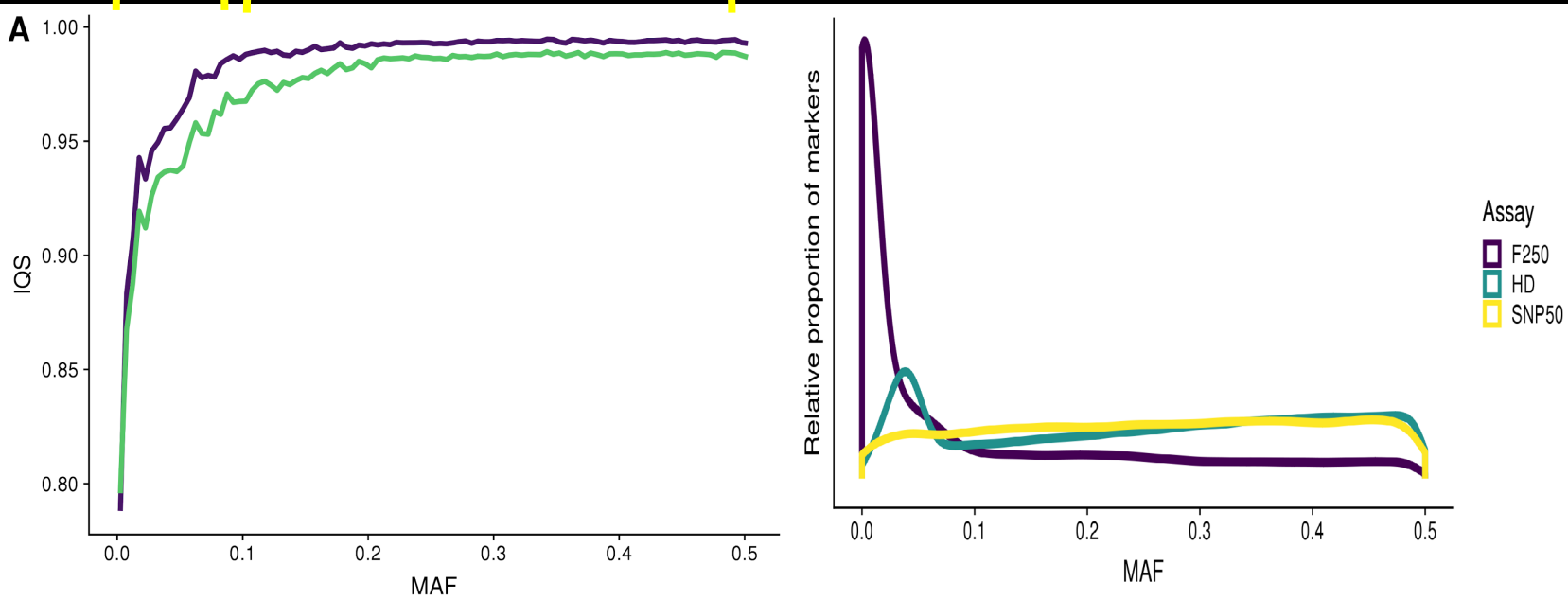
A multi-breed reference panel and additional rare variants maximize imputation accuracy in cattle

Troy N. Rowan¹ , Jesse L. Hoff¹ , Tamar E. Crum¹ , Jeremy F. Taylor¹ , Robert D. Schnabel^{1,2*} 
and Jared E. Decker^{1,2*} 



These are hard
but the VAST
majority

These are easy
but the **minority**

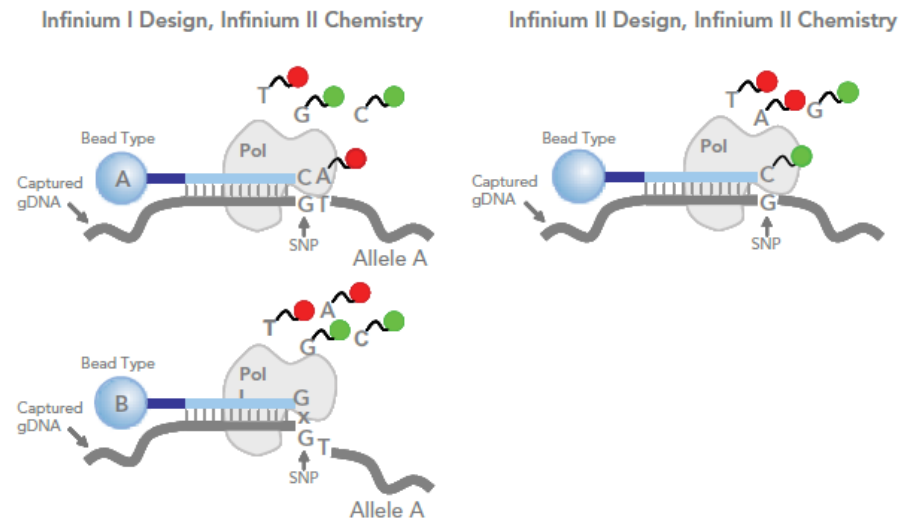


Outline

- **Sources of error**
 - **Multiallelic Chip/Sequence**
 - **Private alleles** (not really errors but misconceptions)
 - **Genotype Recall and Precision**
- **Better variants**
 - **GATK variant calling & VQSR**
 - **Deep Variant**
- **Better phasing**

Estimation of an unknown true genotype

Figure 2: Assay Design for Iselect Whole-Genome Genotyping

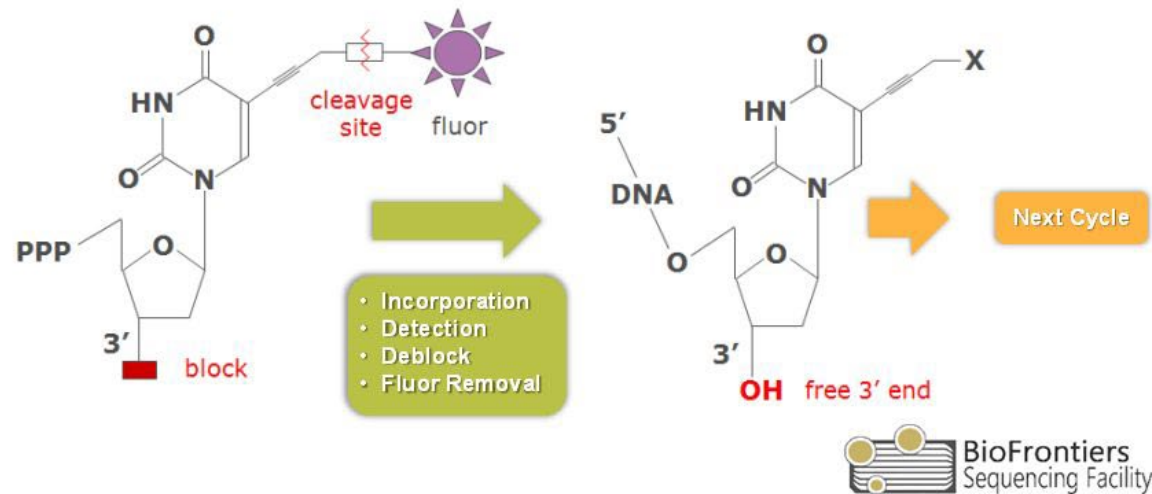


Infinium II Assay chemistry is used for all iSelect Custom Infinium Assays with either Infinium I or Infinium II probe design. Infinium I probe design includes two bead types per SNP locus. Infinium II probe design requires just one bead type per locus. Infinium I probes end at the queried SNP base. Infinium II probes end at the base preceding the queried SNP. This method allows unlimited access for SNP interrogation.

~99.5% accuracy/reproducibility
800K markers = 4,000 errors PER sample

Illumina sequencing

- Based on reversible terminator chemistry
- Sequencing by synthesis (SBS)
 - All 4 fluorescently labeled bases present



Q30 = 99.9% accuracy
20X coverage = ~60 Gb = ~60M single base errors

Multiallelic Chip

- 1000 Bulls Run9 TAUIND Tranche90.0-99.0 (excluding DoNotAnalyze)

Class	Count		SNP50	SNP50V3	HD	GGPF250
All multiallelic	6,336,169	N	52,781	52,988	758,410	205,015
PASS	2,476,198	Multiallelic	1,964	1,910	19,853	6,044
PASS Non-Major allele >1%	2,051,428	Multiallelic%	3.72%	3.60%	2.62%	2.95%
PASS Minor allele >1%	286,115	maf>0.005	587	556	2,612	963
		maf>0.01	472	447	1,804	704
		maf>0.05	185	175	621	206
		maf>0.005	1.11%	1.05%	0.34%	0.47%
		maf>0.01	0.89%	0.84%	0.24%	0.34%
		maf>0.05	0.35%	0.33%	0.08%	0.10%
		maf>0.005	0.0056%	0.0052%	0.0017%	0.0023%
		maf>0.01	0.0089%	0.0084%	0.0024%	0.0034%
		maf>0.05	0.0175%	0.0165%	0.0041%	0.0050%

Private alleles:

Are breeds as distinct as we think?

Ros-Freixedes et al.
Genetics Selection Evolution (2022) 54:39
<https://doi.org/10.1186/s12711-022-00732-8>



RESEARCH ARTICLE

Open Access



Rare and population-specific functional variation across pig lines

Roger Ros-Freixedes^{1,2*} , Bruno D. Valente³, Ching-Yi Chen³, William O. Herring³, Gregor Gorjanc¹, John M. Hickey¹ and Martin Johnsson^{1,4}

“While genome-wide association studies that involve more than one breed typically find multiple breed-specific associations, based on our results it seems unlikely that breed-specific associations arise from the low prevalence variants. Instead, **breed-specific associations depend on the effect of the differences in allele frequencies, linkage disequilibrium structure, and other genetic background features on the power** to detect the effect of prevalent variants across populations.”

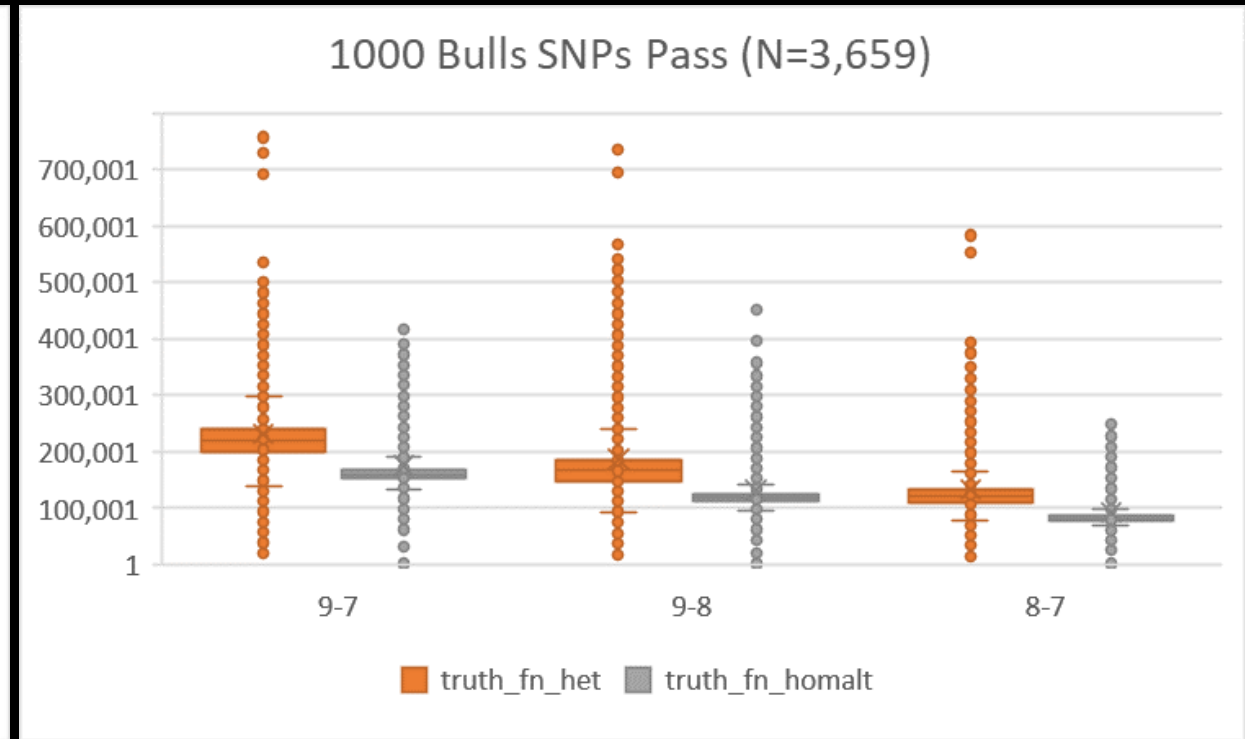
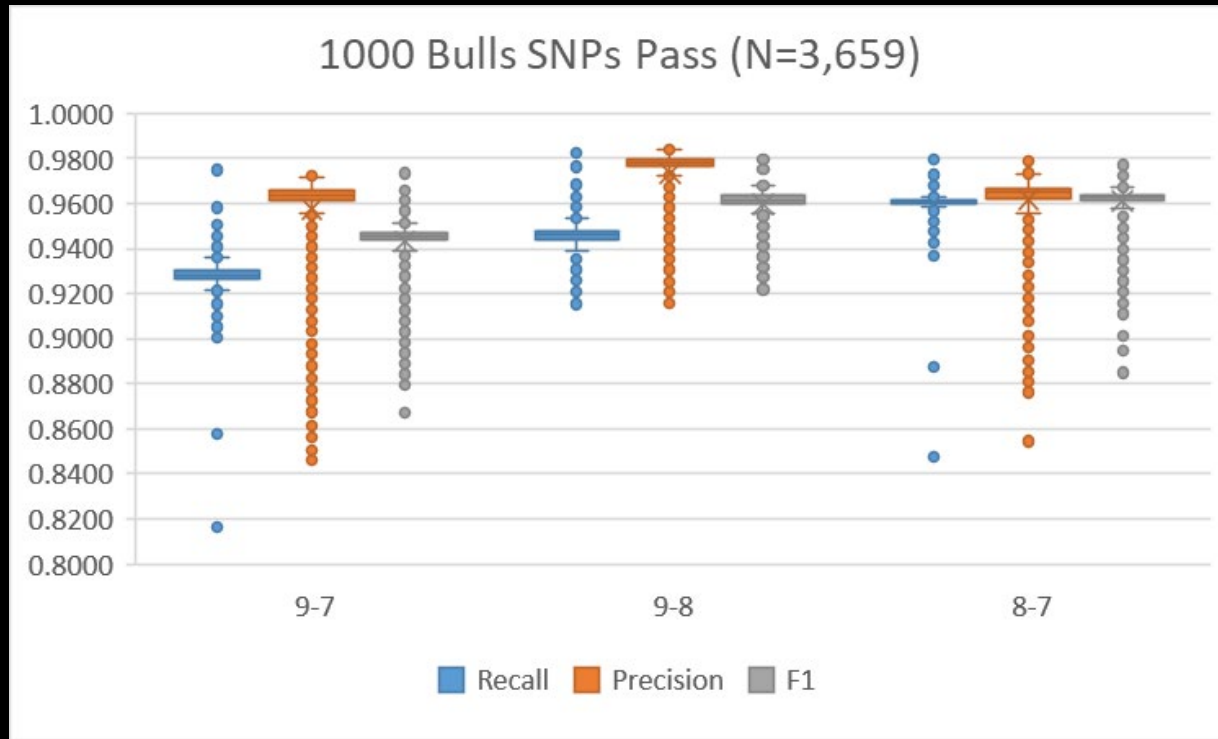
Private alleles

	ANG	BSW	HOL	JER	NRC
Chr24 Filter	N=265	N=262	N=930	N=179	N=345
Total Variants	443,640	454,414	496,121	377,398	459,060
Private Biallelic AC>1	8,987	19,470	11,098	4,570	10,869
Private AF>0.05	3,355	8,364	1,376	2,515	3,003
Private Biallelic AC>1	2.03%	4.28%	2.24%	1.21%	2.37%
Private AF>0.05	0.76%	1.84%	0.28%	0.67%	0.65%

- **Number of private alleles proportional to genetic distance**
- **As you increase N within breed or N across breeds the number of private alleles decreases**

Genotype Recall & Precision

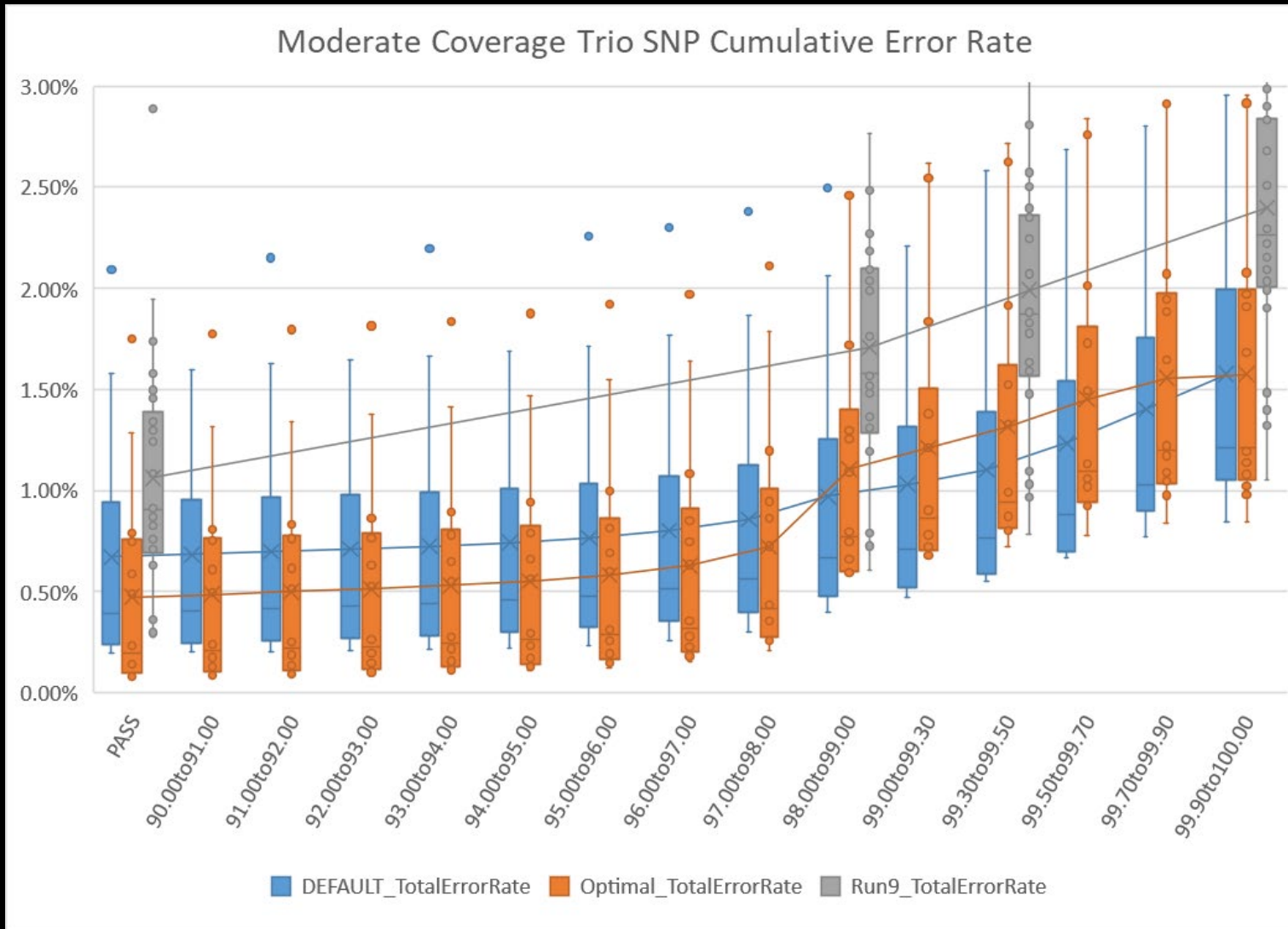
- 3,659 samples in all three Runs, compare calls between runs
- Run9 considered “TRUTH”
- **Recall** = $TP / (TP + FN)$
 - How many variants did you miss in the previous run
- **Precision** $TP / (TP + FP)$
 - How many variants in previous run were not real variants



~1.5M CPU hours (a LOT of data not show)

Mendelian error rate

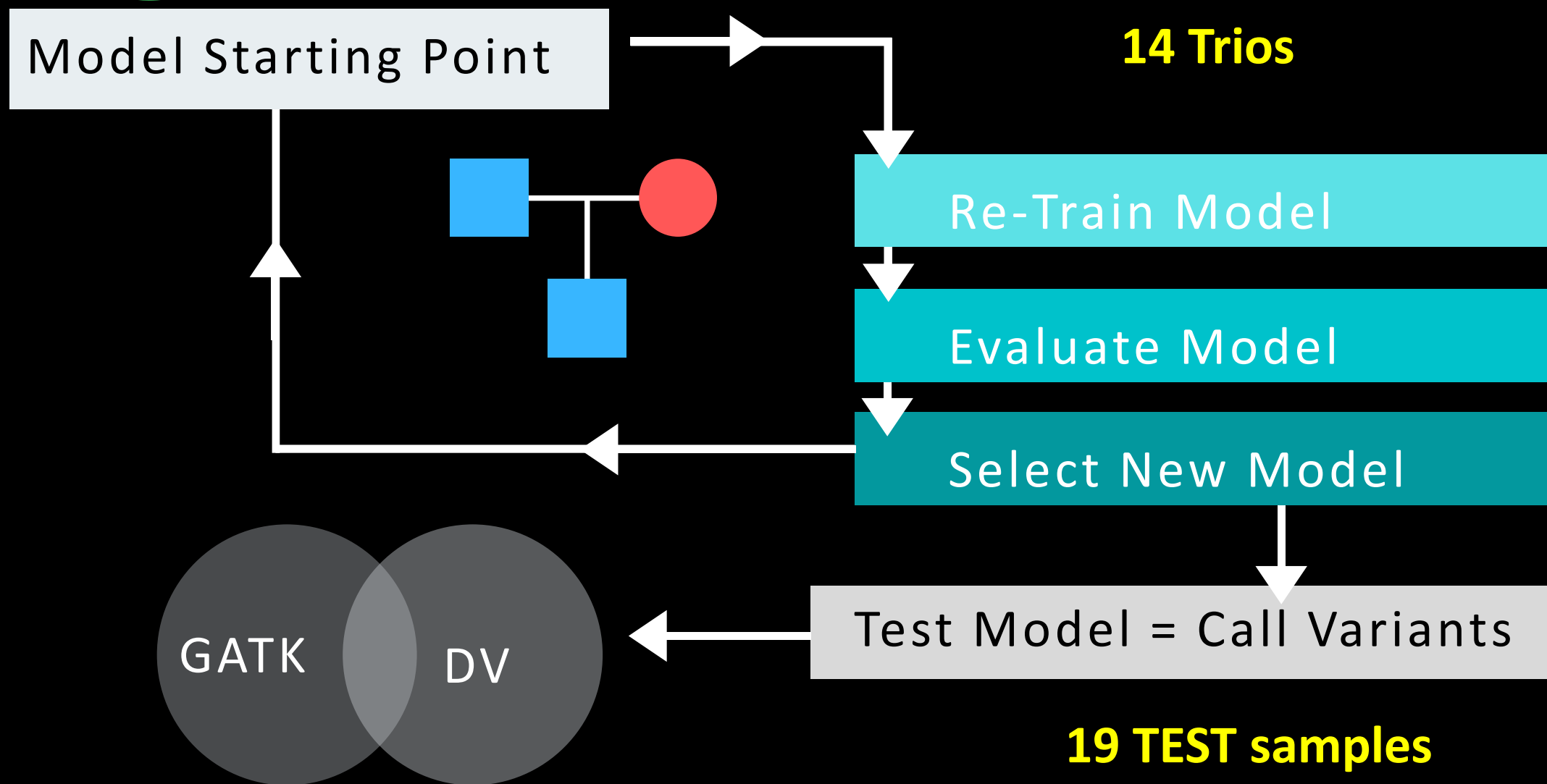
Optimal vs Default vs Run9



Can we produce even
better Genotypes?



Deep Variant: TrioTrain



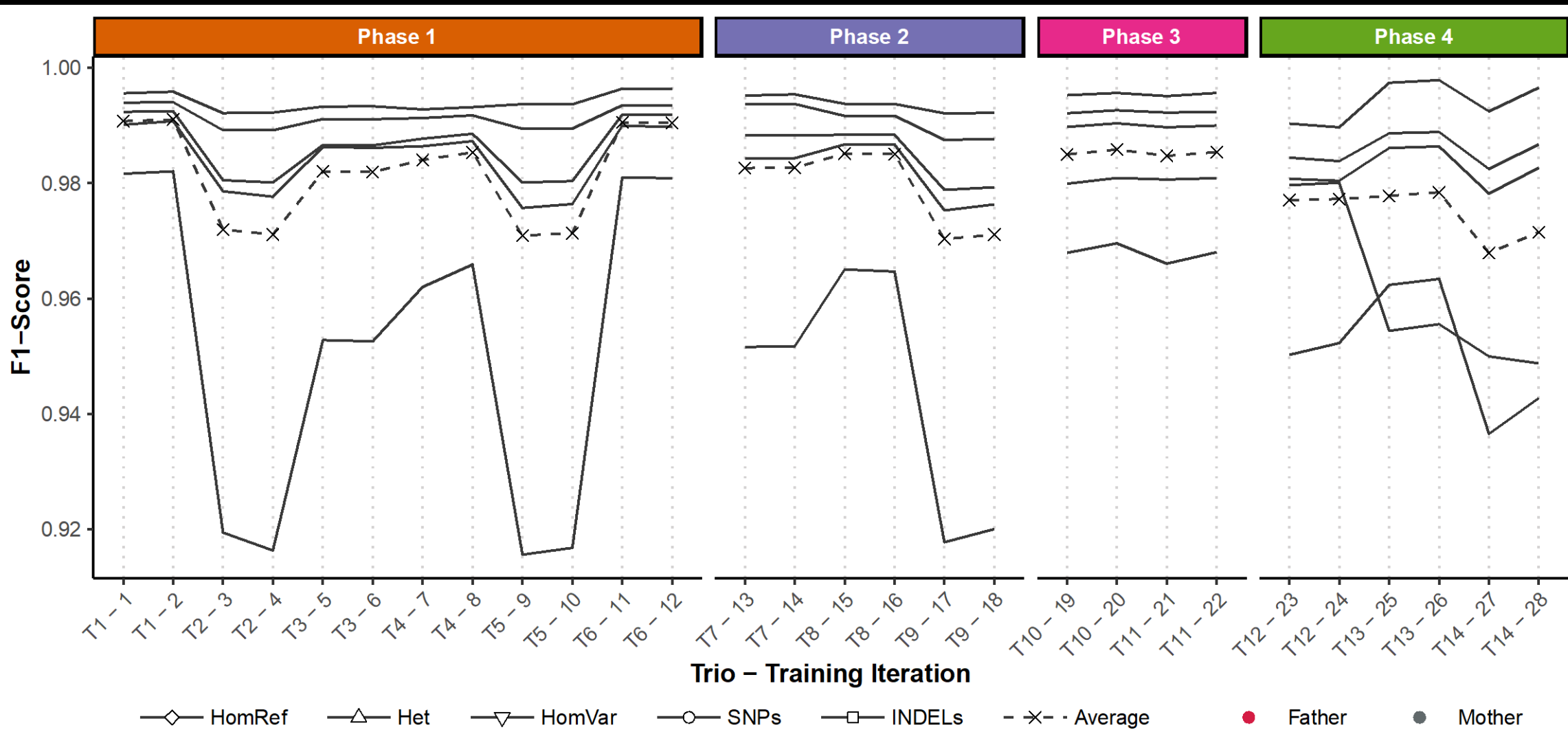


Figure 1.3) Model training performance across phases

For each trio, the first iteration begins with the father by giving labeled examples from CHR 1 – 29, X. Pedigree information is not explicitly provided to the model; instead, the checkpoint that achieves the maximum F1-Score in the offspring's labeled examples is chosen as the starting point for the next iteration.

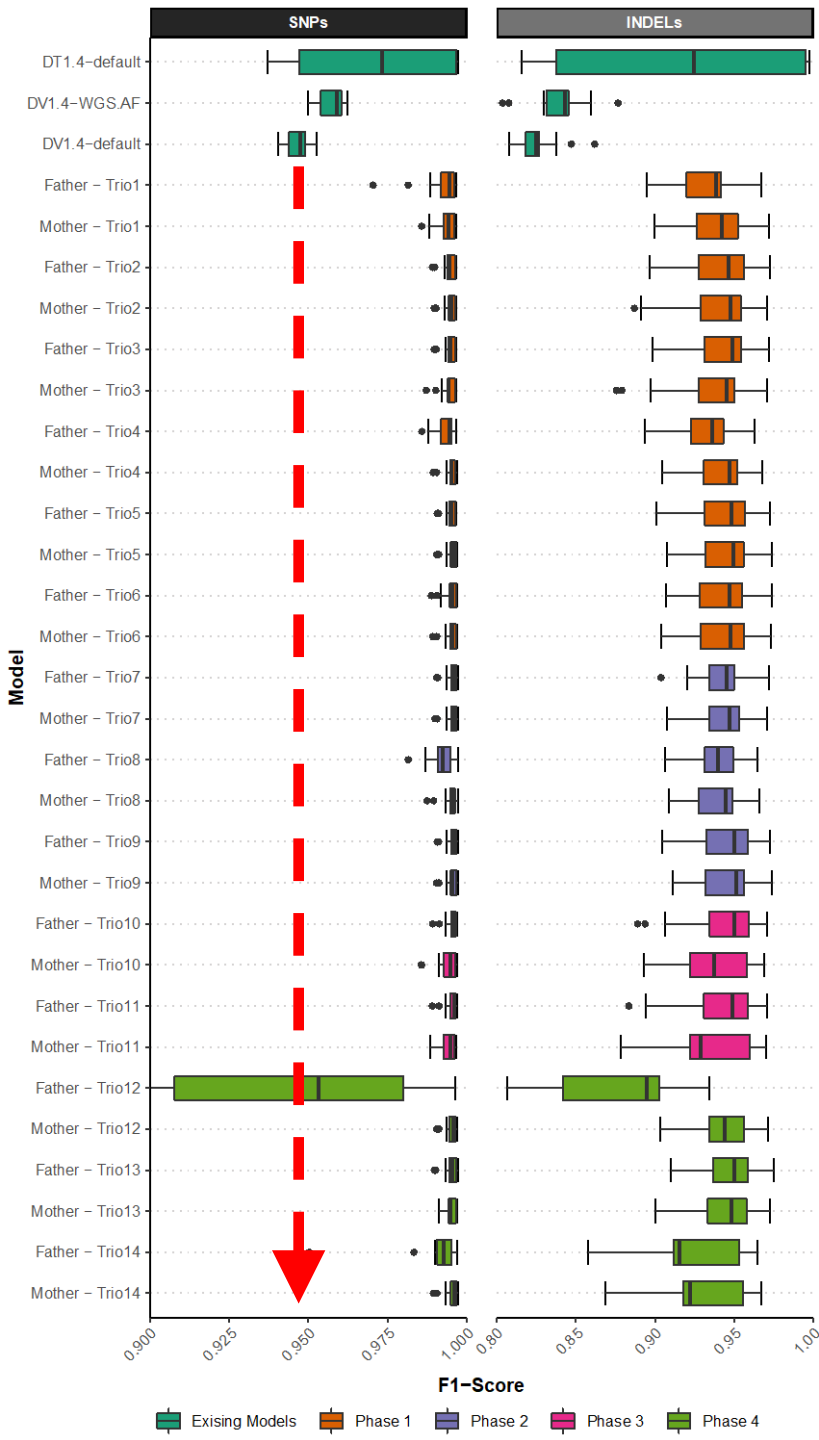


Figure 1.4) Comparing variants from test genomes.

Each box-and-whisker represents the F1-score with an independent set of bovine samples previously unseen by the model

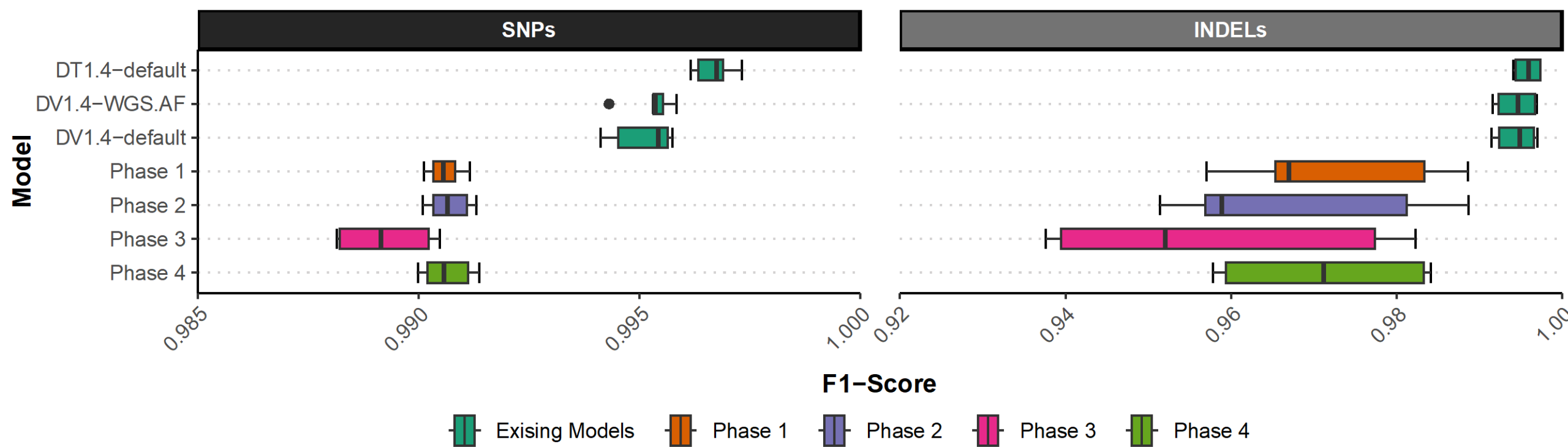


Figure 1.5) Comparing variants from *human* genomes.

Each box-and-whisker represents the F1-score with the GIAB human samples ($n = 6$). We compared the variants produced by each model against their respective GIAB v4.2.1 benchmark sets using hap.py.

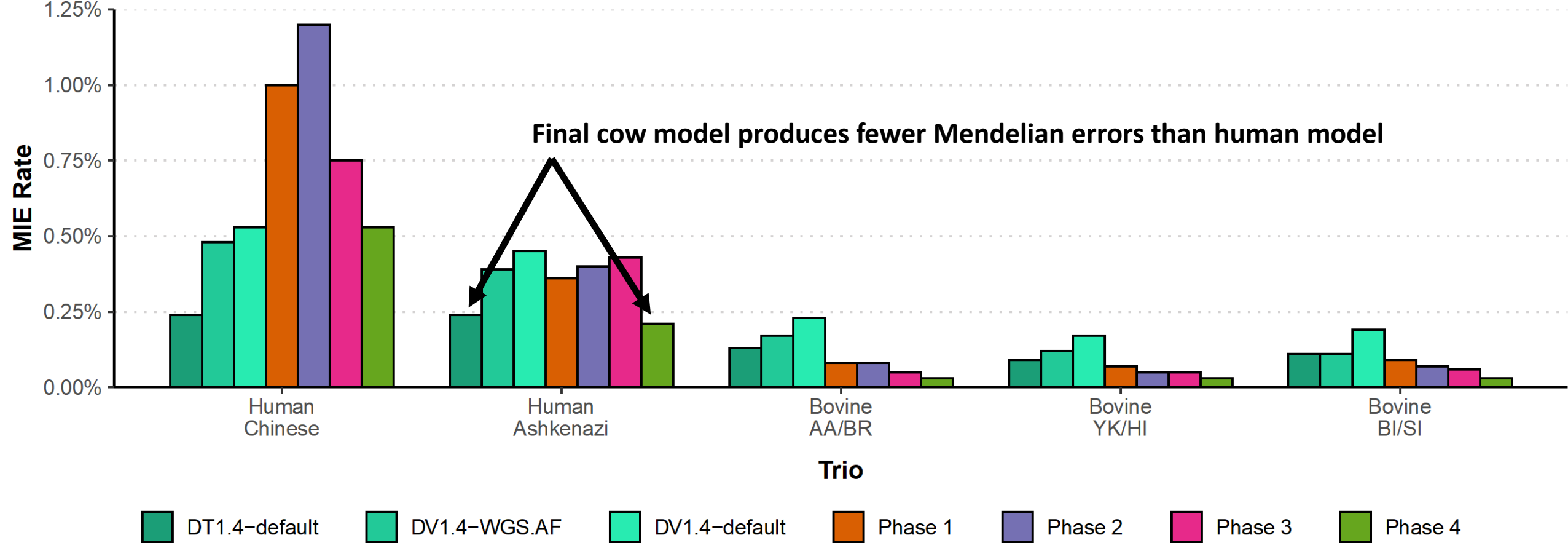


Figure 1.6) Inheritance error rate in human and bovine trios.

Mendelian Inheritance Errors (MIE) were identified in PASS variants in the autosomes and X chromosome for two GIAB human trios and six bovine hybrid-cross trios.

Conclusion: We need better truth sets for cows!

PARTNERSHIP: DEVELOPMENT OF GENOMIC REFERENCE MATERIALS FOR CATTLE

GIAB-Ag

The objectives of this proposal are to:

- 1. Identify an optimal set of individuals** from those available in the Bovine Pangenome Consortium and PanEpigenome projects to develop cattle Reference Materials.
2. Generate primary and immortalized **cell lines** for the Reference Material samples to enable distribution to the community for future use.
3. Aggregate and generate sequence data for Reference Material samples to **produce a definitive truth set** for variant calls (SNP, INDEL, SV) to serve as the authoritative benchmark resources for the community.
- 4. Develop best practices guidelines** and standard procedures to support the genomics community's use of the generated Reference Materials and provide a **roadmap for replicating our research in other species.**

<https://cris.nifa.usda.gov/cgi-bin/starfinder/0?path=fastlink1.txt&id=anon&pass=&search=R=98382&format=WEBLINK>

**What about structural
variants?**

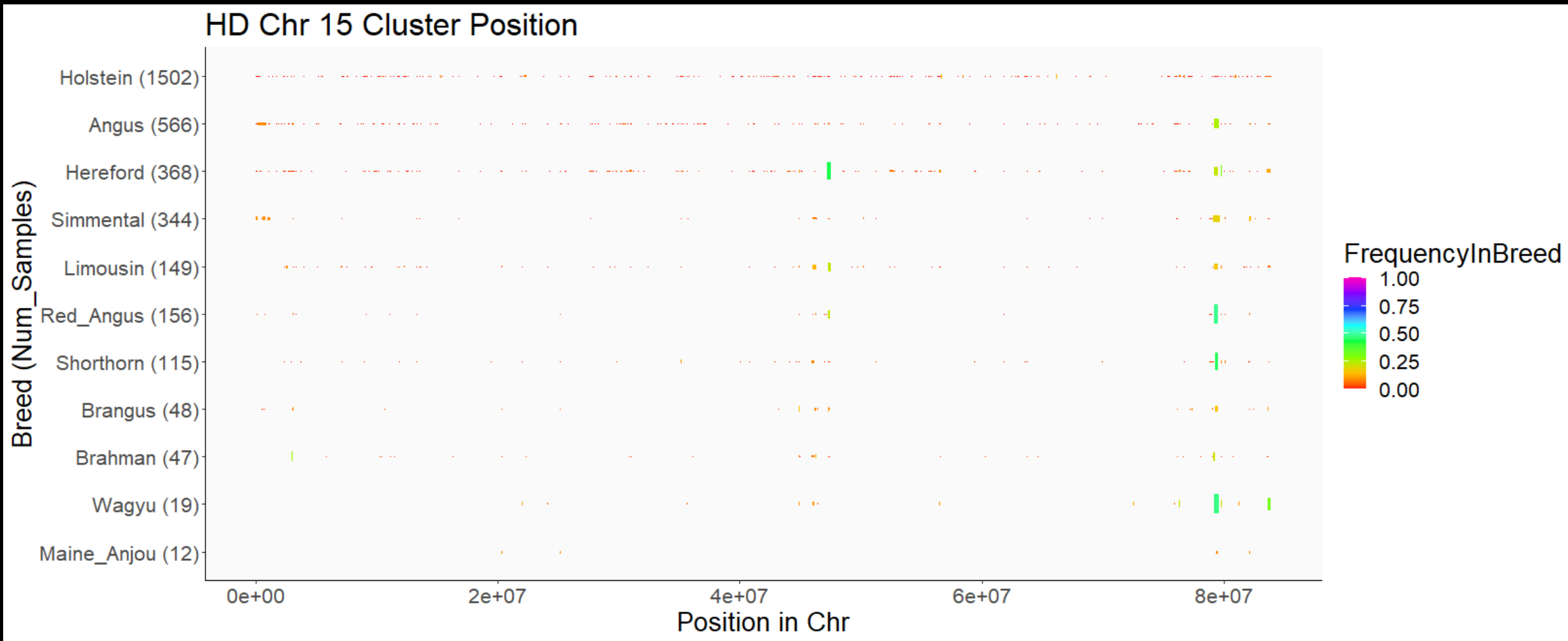
Mapping **Copy Number Variants** Across The Cattle Genome

Jacob Rissman

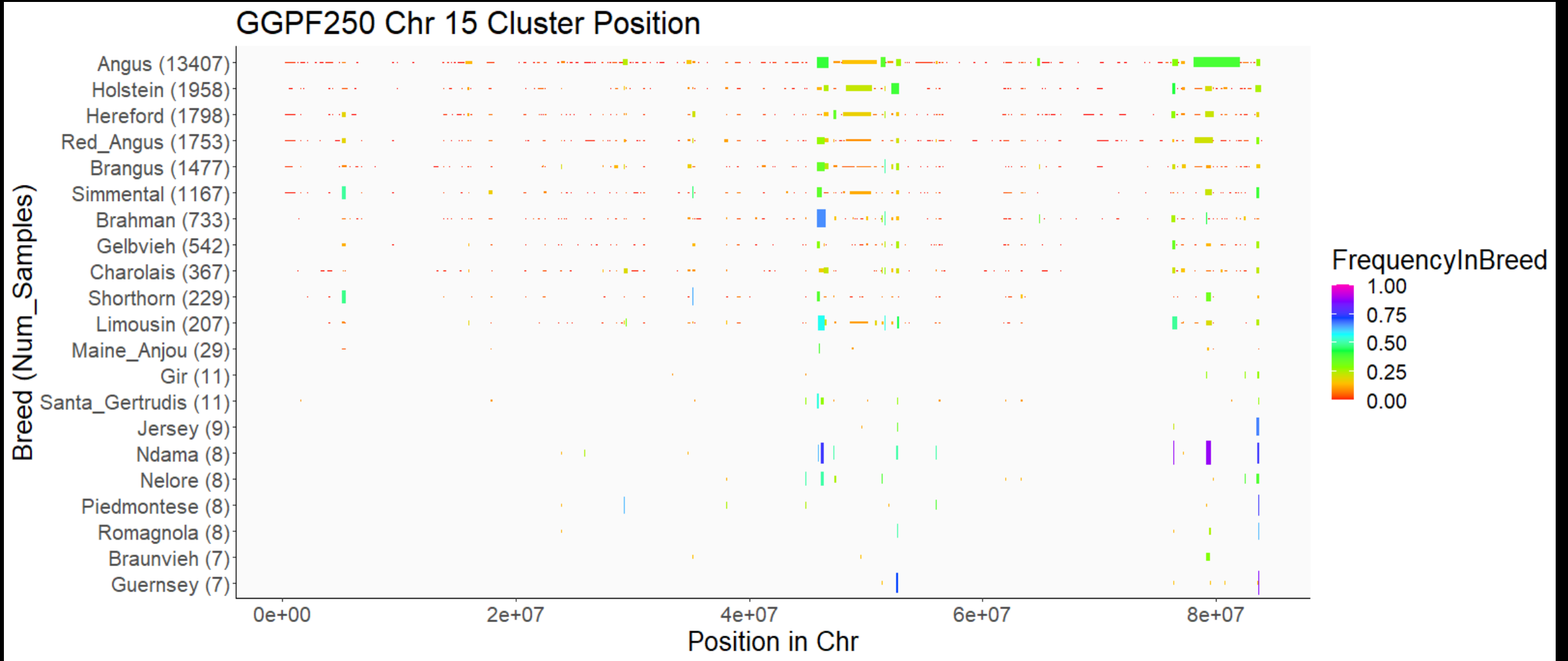
- ~61,000 SNP Chip Samples
- 89 Different Breeds
- 652 Trios
 - BOVG50V1 (19 trios)
 - SNP50 (282 trios)
 - GGPF250 (351 trios)

# Samples	Assay
5,203	BOVG50V1 (50k)
18,788	SNP50 (SNP50_B) (50k)
882	SNP50V3 (SNP50_C) (50k)
3,347	GGP100V1 (100k)
32,252	GGPF250 (250k)
3,741	HD (777k)
1,241	GGPF250 and SNP50
1,605	HD and GGPF250
366	HD and SNP50

Comparisons Across Breeds

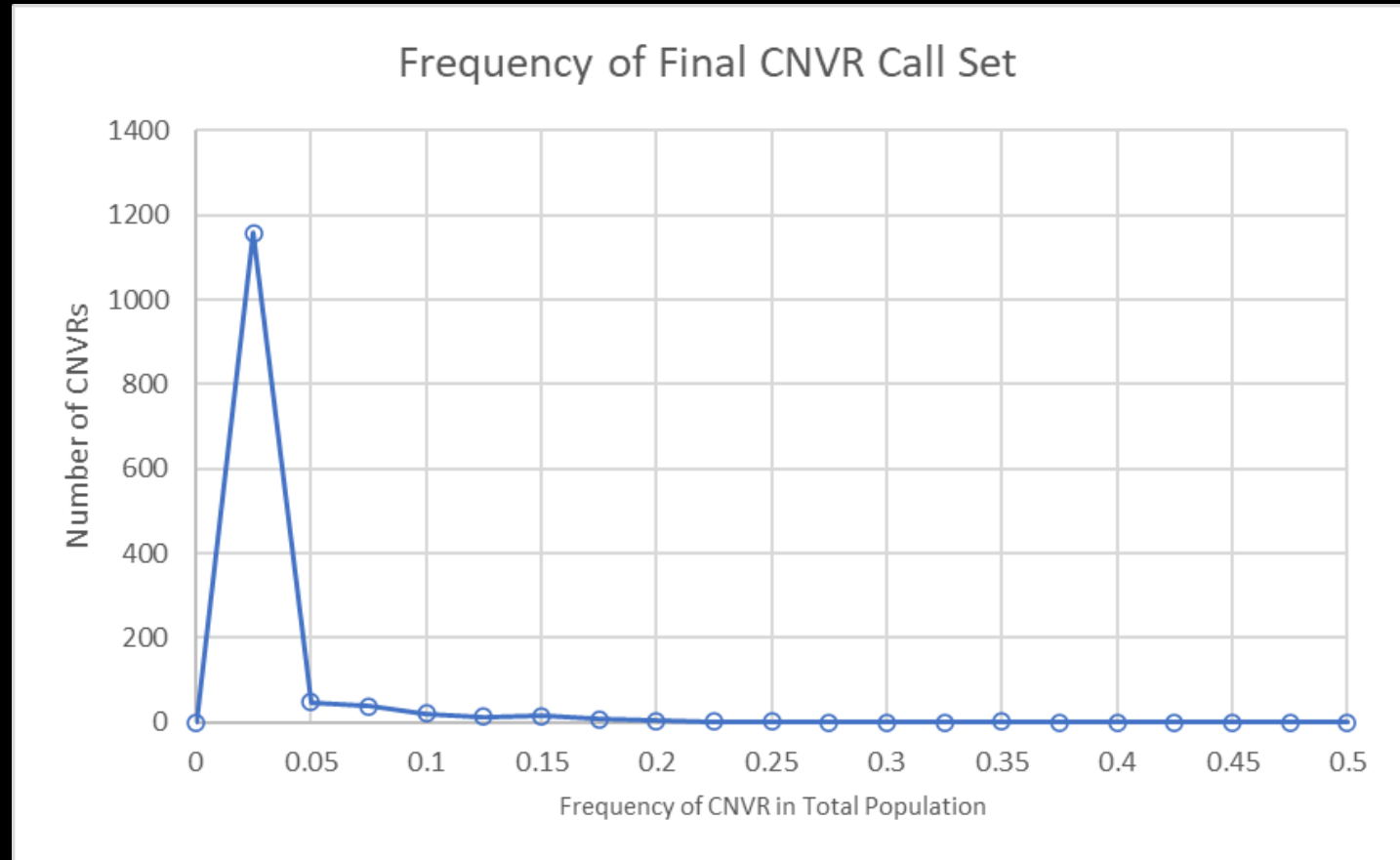


Comparisons Across Breeds



Confident Final CNVR Call Set

1,316 CNVRs created
8.51% of the genome



POPULATION SCALE CHARACTERIZATION OF STRUCTURAL VARIANTS

Jenna Kalleberg-Ridge

Cue: a deep learning framework for SV calling and genotyping

<https://github.com/PopicLab/cue>

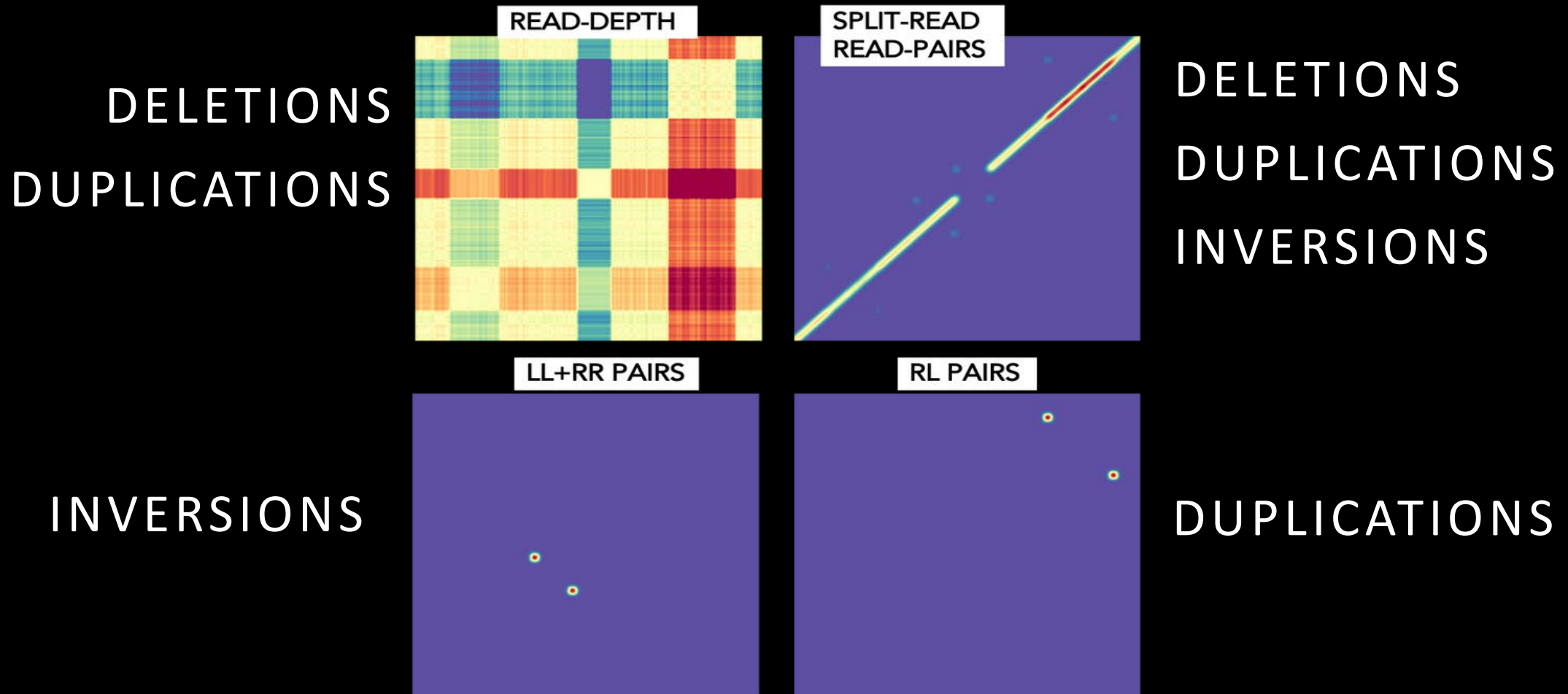


Table 2.1) Stratified SV counts. Using summary metrics from the per-sample VCFs produced with the UMAGv1 cohort, we stratified by SV type and genotype class to observe how different QC thresholds changed SV calls.

	TYPE				CLASS		
	DEL	INV	DUP	IDUP	HET	HOMALT	TOTAL
All Samples ^[1]							
Count	1,996,215	316,497	172,806	9,431	2,029,540	465,409	2,494,949
% Total	80.01%	12.69%	6.93%	0.38%	81.35%	18.65%	
Mean / Genome	298.34	47.44	25.90	2.17	303.32	69.56	374
Median / Genome	310	36	25	2	312	57	
Excluding Outliers ^[2]							
Count	1,991,646	314,814	172,374	9,356	2,024,380	463,810	2,315,816
% Total	80.04%	12.65%	6.93%	0.38%	81.36%	18.64%	
Mean / Genome	298.15	47.255179	25.87	2.16	303.05	69.43	326
Median / Genome	310	36	25	2	312	57	
Samples with Average Coverage $\geq 7.5x$ ^[3]							
Count	1,763,271	290,873	154,255	8,740	1,787,070	430,069	2,217,139
% Total	79.53%	13.12%	6.96%	0.39%	80.60%	19.40%	
Mean / Genome	351.74	58.05	39.54	2.88	412.50	132.23	452
Median / Genome	349	47	30	2	354	73	
Samples with Average Coverage $\geq 15x$ ^[4]							
Count	804,518	172,607	76,787	5,039	801,893	257,058	1,058,951
% Total	75.97%	16.30%	7.25%	0.48%	75.73%	24.27%	
Mean / Genome	413.85	88.84	30.80	2.28	412.50	132.23	536
Median / Genome	409	78	40	3	410	123	

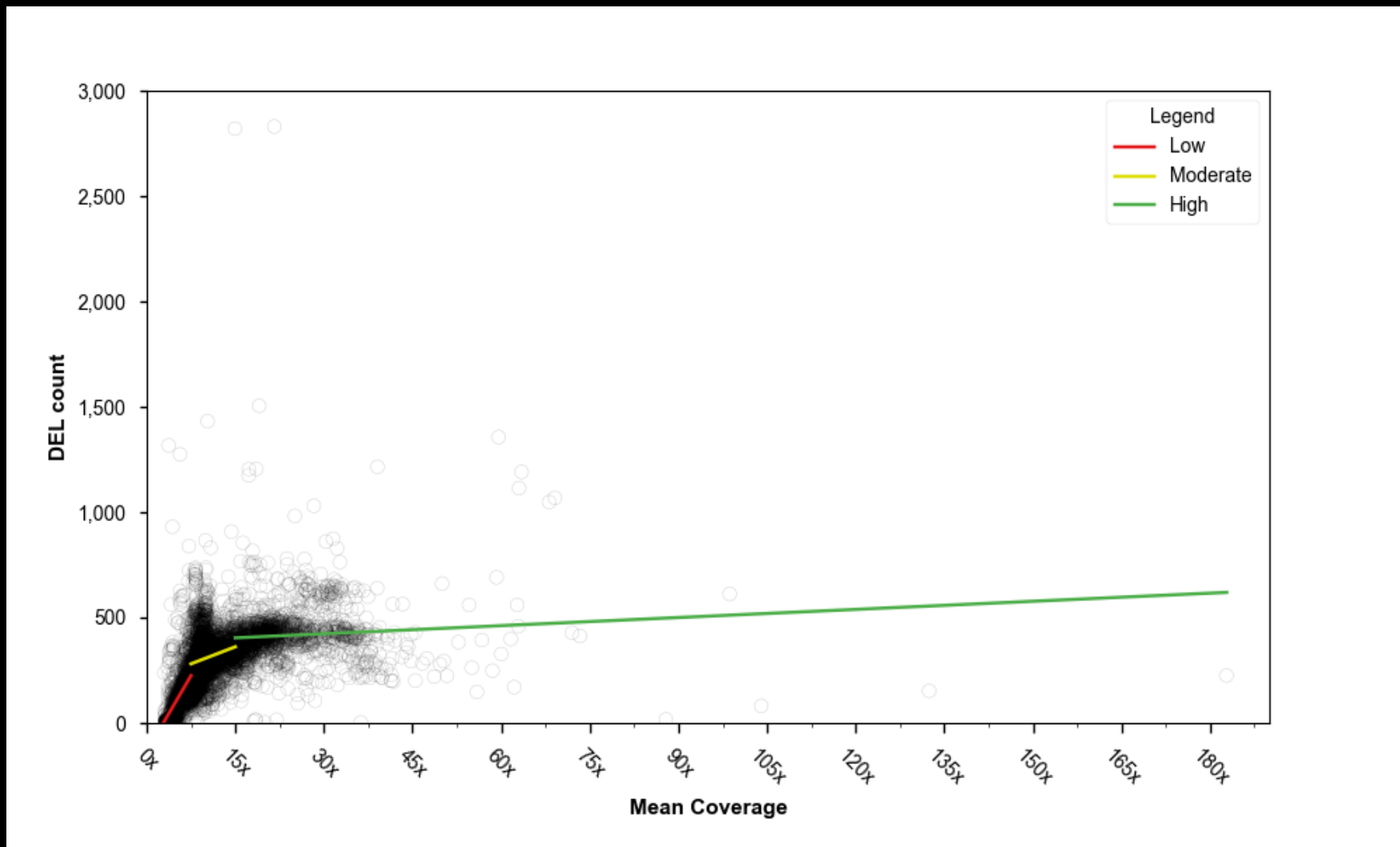


Figure 2.1) Sample coverage versus DEL counts. After excluding outliers, each black circle represents the number of deletions for each sample within the UMAGv1 cohort. Linear regression trend lines were created for three different coverage categories. Red for low coverage (<7.5x), yellow for moderate coverage (≥7.5x – <15x), and green for high coverage (≥15x) [respective Pearson's $r = 0.51, 0.18, 0.09$]. **Overall, counts plateau around 400 DEL per genome, with expected variation between individuals.**

Estimation of ancestral alleles

Hunter McConnell

Chrom	Pos	REF	ALT	AF_Banteng	AF_Bison	AF_Gaur	AF_Yak
25	735	G	A	0	0	0	0
25	18924	C	T	0.2786	0.554	0.5592	0.583
25	31877	A	G	0.98	0.914	0.956	0.224

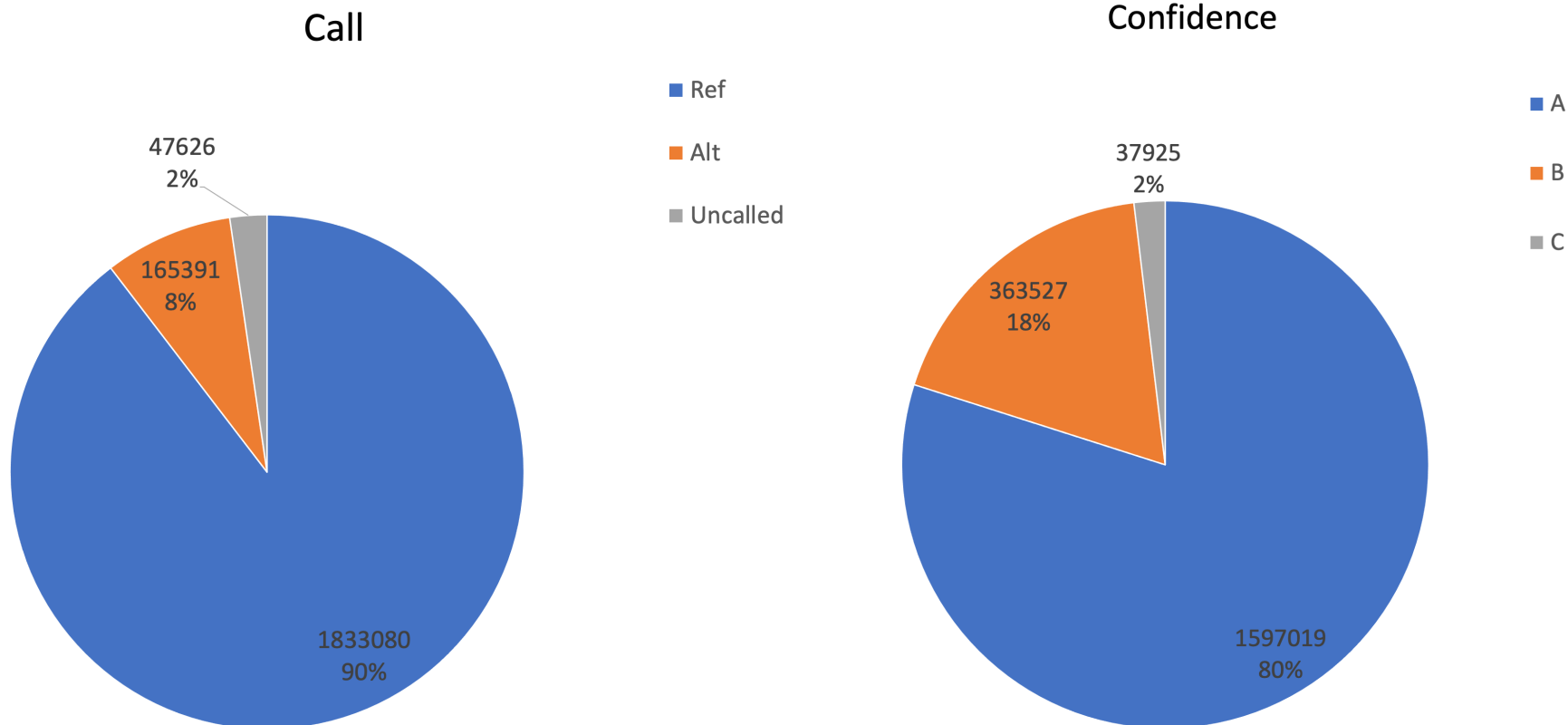
Chrom	Pos	REF	ALT	AF_Bin_Banteng	AF_Bin_Bison	AF_Bin_Gaur	AF_Bin_Yak
25	735	G	A	1	1	1	1
25	18924	C	T	2	3	3	3
25	31877	A	G	5	5	5	2

- 198 Bovid genomes
- Chromosome 25
- 2,046,097 Called Variants

$$175,931,726 \times 0.8 = 141,603,390$$

3.5x Previous Attempts

Dominette is a good representation of bovids



Phasing

Shapeit5 Lots of parameters, scales sub-linearly

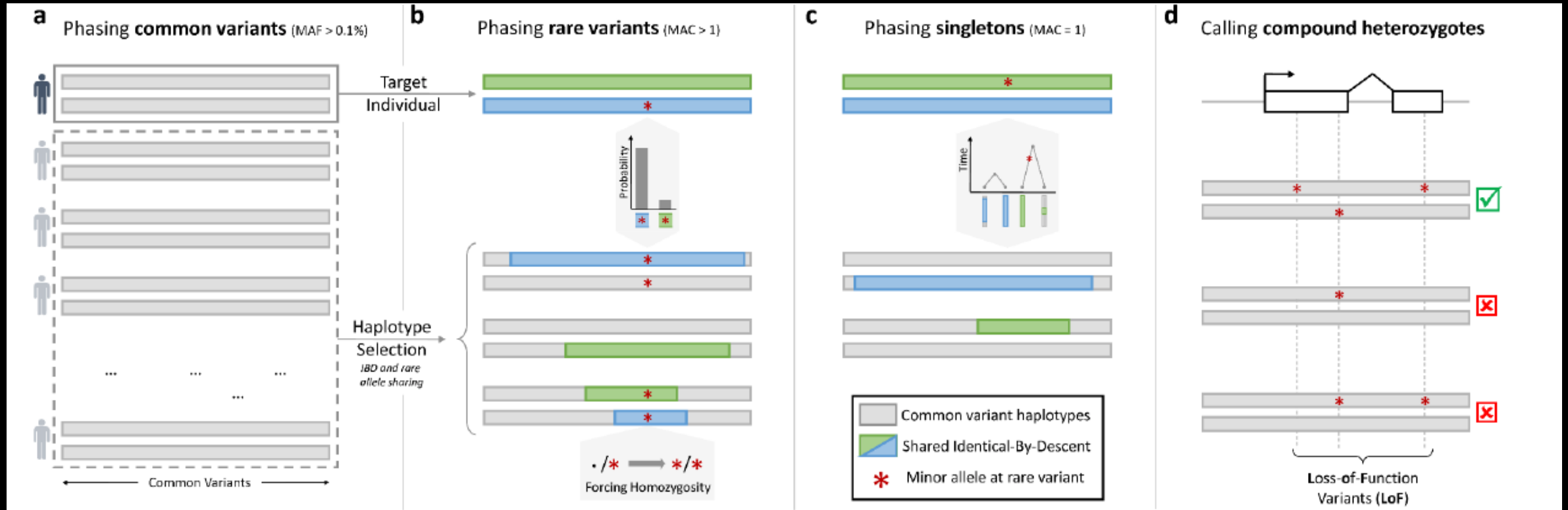
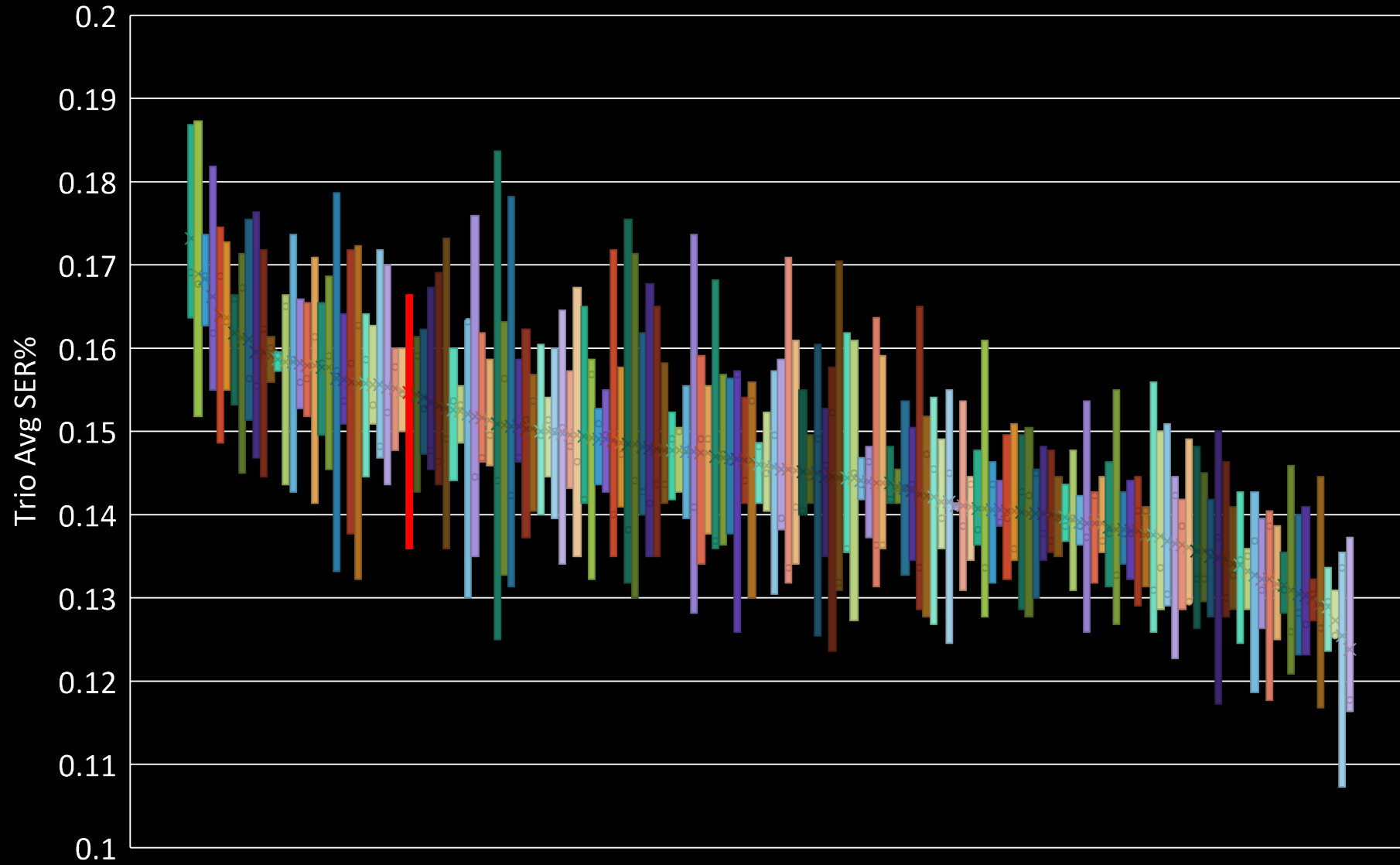


Figure 1: Rationale of SHAPEIT5. From left to right. (a) All samples are phased at common variants ($MAF \geq 0.1\%$). (b) Phasing of a given rare variant onto the haplotypes at common variants. Conditioning haplotypes used in the estimation share long matches with the target (in green and blue) and are not monomorphic at the rare variant. (c) Singleton phasing by assigning the new allele on the target haplotype with the shortest match. (d) Compound heterozygous event mapping based on the rare variant phasing (a-c).

Phasing

Shapeit5

Replicates



Default: 0.155%
Best: 0.124%
Diff: 0.031%
%Diff: 20%

Phasing

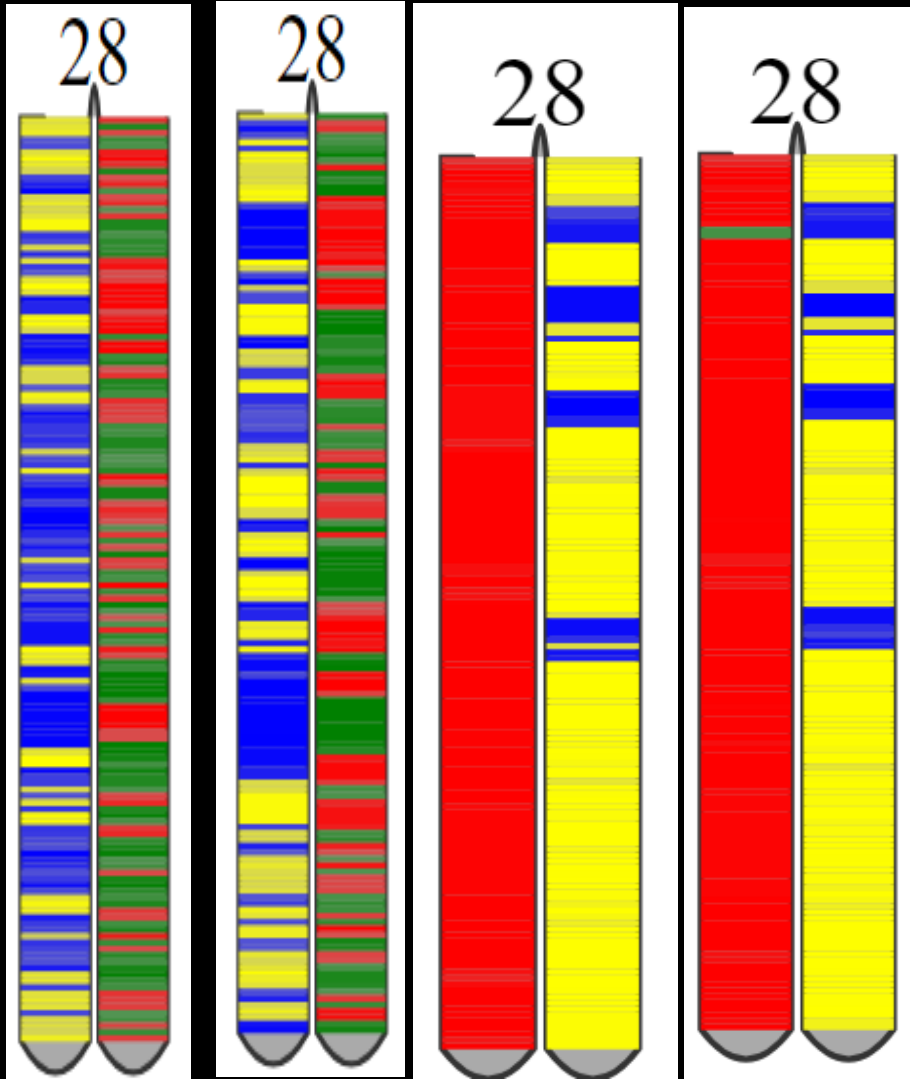
Sequence Phasing

- 6,137 genomes
- 212,942,459 variants
- N_e 300 & 30,000
- Recombination Map, LD Map, No Map
- 5 replicates
- With pedigree & without pedigree
- **Each chromosome phased 60 times**

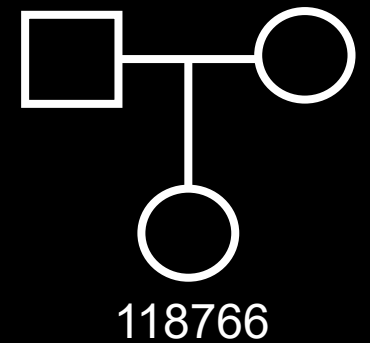
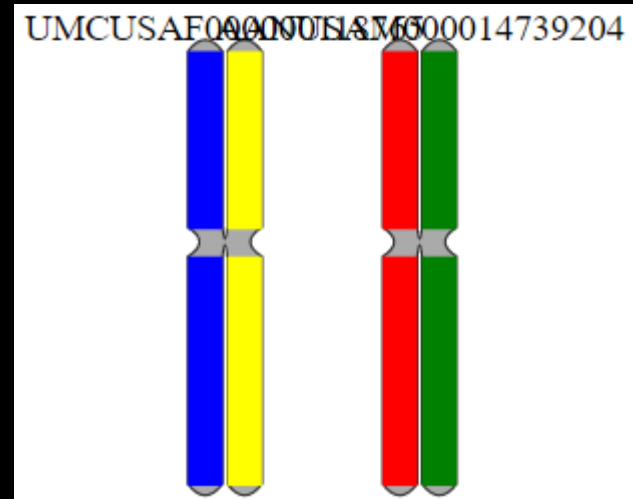
118766 UMCUSA000000118766

- A. 1kbulls Run8 Beagle phased
- B. 1kbulls Run8 Eagle phased
- C. UMAG2 Shapeit5 Ne 300
- D. UMAG2 Shapeit5 Ne 30,000

Phasing



Version	nVariants	nMendelian Errors	nSwitch	Switch%
A	62,418	1019	609	0.98
B	56,541	670	405	0.72
C	51,318	7	0	0
D	51,321	5	0	0



assay	num_loci	num_samples
ANGGS	48,434	21,982
BOVG50V1	45,606	112,664
BOVGGPV3K	2,837	175,647
BOVLDC	6,819	289,009
BOVLDV2A	7,822	273,092
BOVMD	50,135	71,881
DAIRYULDB	4,047	65,021
GGP100V1	92,442	25,556
GGP90KT	75,970	40,379
GGP9K	8,564	143,532
GGPF250	203,920	41,016
GGPHDV3	137,390	35,198
GGPIND35	34,984	22,076
GGPLDV1	6,778	289,253
GGPLDV3	25,658	71,617
GGPLDV4	29,138	68,889
GGPRANULD	28,908	126,867
GGPSIMULD	29,161	121,517
HD_GGPF250	924,261	22,031
HD	756,704	10,592
IDBV3	50,282	14,703
IND90KH	73,609	22,117
SNP50	52,642	73,165
WBSV1	49,388	25,827
ZLD2	17,373	21,999
ZLD4	19,657	21,995
ZLD5	33,494	21,994
ZMD2	60,106	21,978
ZOETIS1	50,062	24,600

SNP-Chip Phasing

- 29 assay
- # Samples 10,529 – 289,253
- # Loci 2,837 – 924,261
- DepthCommon 4, 8, 16, 32
- McMcIter 15, 25, 45
- N_e 300 & 30,000
- Recombination Map, LD Map, No Map
- 5 replicates
- With pedigree & without pedigree
- **Each chromosome phased 240 times**

Summary

- Sources of error
 - Multiallelic Chip/Sequence
 - **Probably not a huge issue but is a source of error**
 - Private alleles
 - **Cows are cows, sample size is far more important**
 - Genotype Recall and Precision
 - **Increased sample size = “new” variants in “old” samples**
- Better variants
 - GATK variant calling & VQSR
 - **Simply picking appropriate values produces more and better data**
 - Deep Variant
 - **Limited by truth set, better genotypes soon**
 - Structural Variants
 - **A LOT of additional variation we are ignoring, mostly rare**
- Better phasing
 - **Better reference panels will produce better imputation**







Imputation

Rowan et al. *Genet Sel Evol* (2019) 51:77
<https://doi.org/10.1186/s12711-019-0519-x>

RESEARCH ARTICLE

Open Access

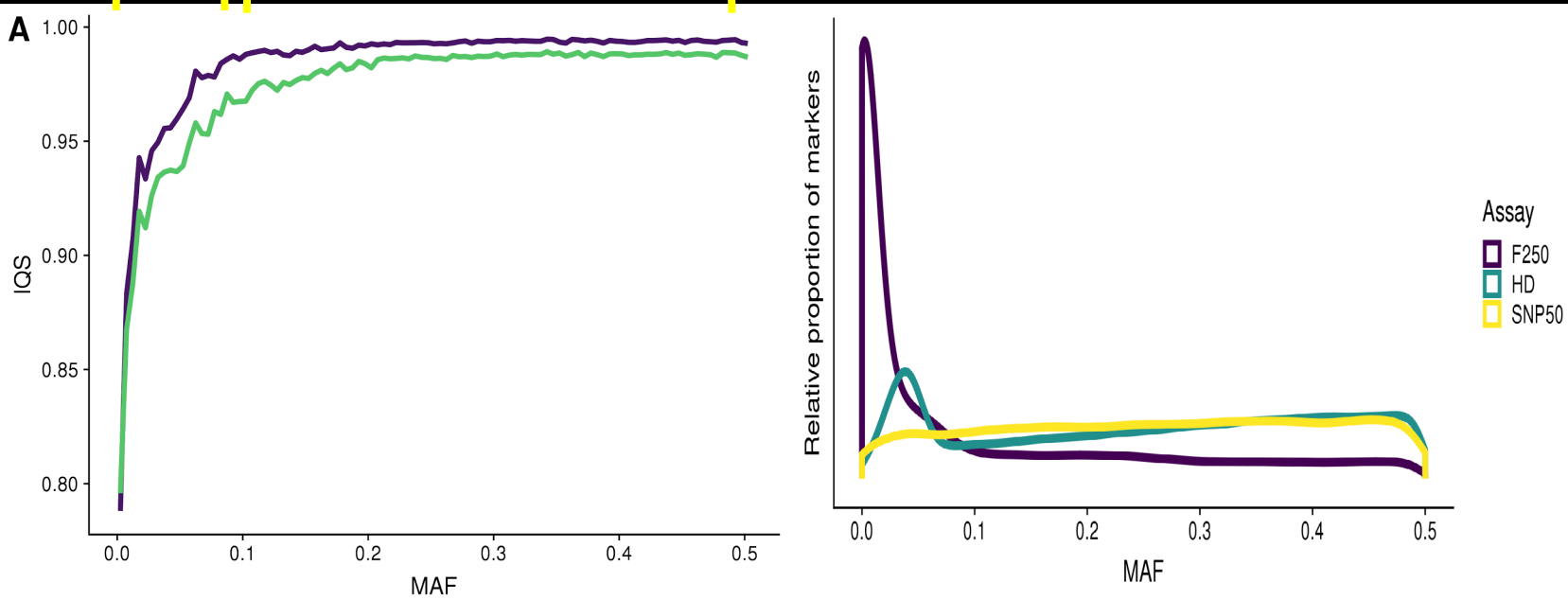
A multi-breed reference panel and additional rare variants maximize imputation accuracy in cattle

Troy N. Rowan¹ , Jesse L. Hoff¹ , Tamar E. Crum¹ , Jeremy F. Taylor¹ , Robert D. Schnabel^{1,2*} 
and Jared E. Decker^{1,2*} 



These are hard
but the VAST
majority

These are easy
but the **minority**



Phasing

Modeling Linkage Disequilibrium and Identifying Recombination Hotspots Using Single-Nucleotide Polymorphism Data

Na Li* and Matthew Stephens^{†,1}

*Department of Biostatistics and [†]Department of Statistics, University of Washington, Seattle, Washington 98195

Manuscript received January 30, 2003

Accepted for publication August 11, 2003

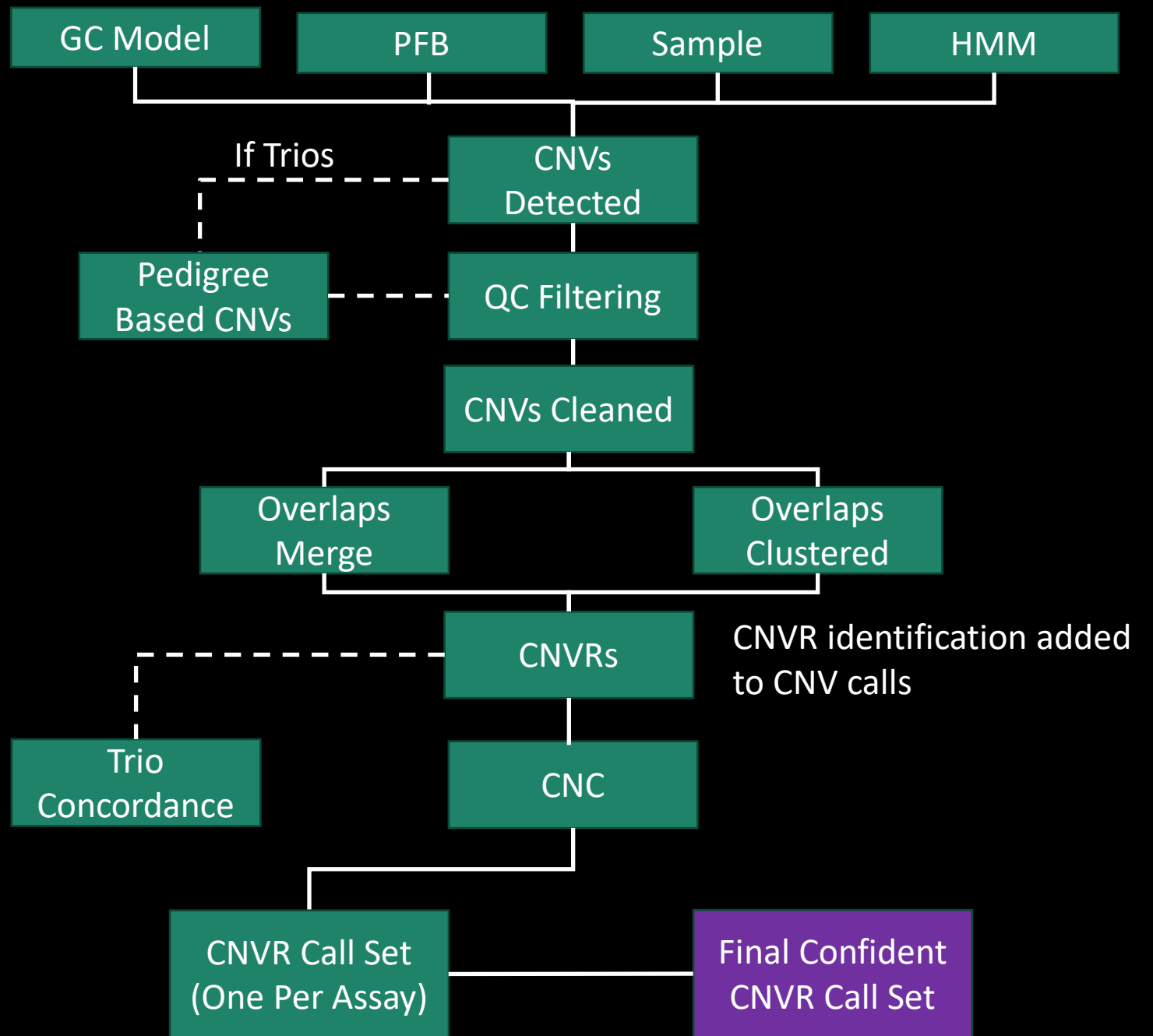
Despite the ease with which coalescent models can be simulated from, using these models for *inference* remains extremely challenging. For example, consider the problem of estimating the underlying recombination rate in a region, using data from a random population sample. It follows from coalescent theory that population samples contain information on the value of the product of the recombination rate c and the effective (diploid) population size N , but not on c and N separately. It has therefore become standard to attempt to estimate the compound parameter $\rho = 4Nc$, and several methods have been proposed. Some (*e.g.*, GRIFFITHS and MARJORAM 1996; KUHNER *et al.* 2000; NIELSEN 2000; FEARN-

Further, in these kinds of applications, where estimation of underlying recombination rates may be of only indirect interest, the usefulness of our model will depend only on whether $\Pr(h_1, \dots, h_n | \rho)$ is a sensible distribution for h_1, \dots, h_n for *some* value of the parameters ρ , even if this ρ does not correspond precisely to the background recombination rate scaled by the effective population size. Under these circumstances our two approxima-

Confident Final CNVR Call Set

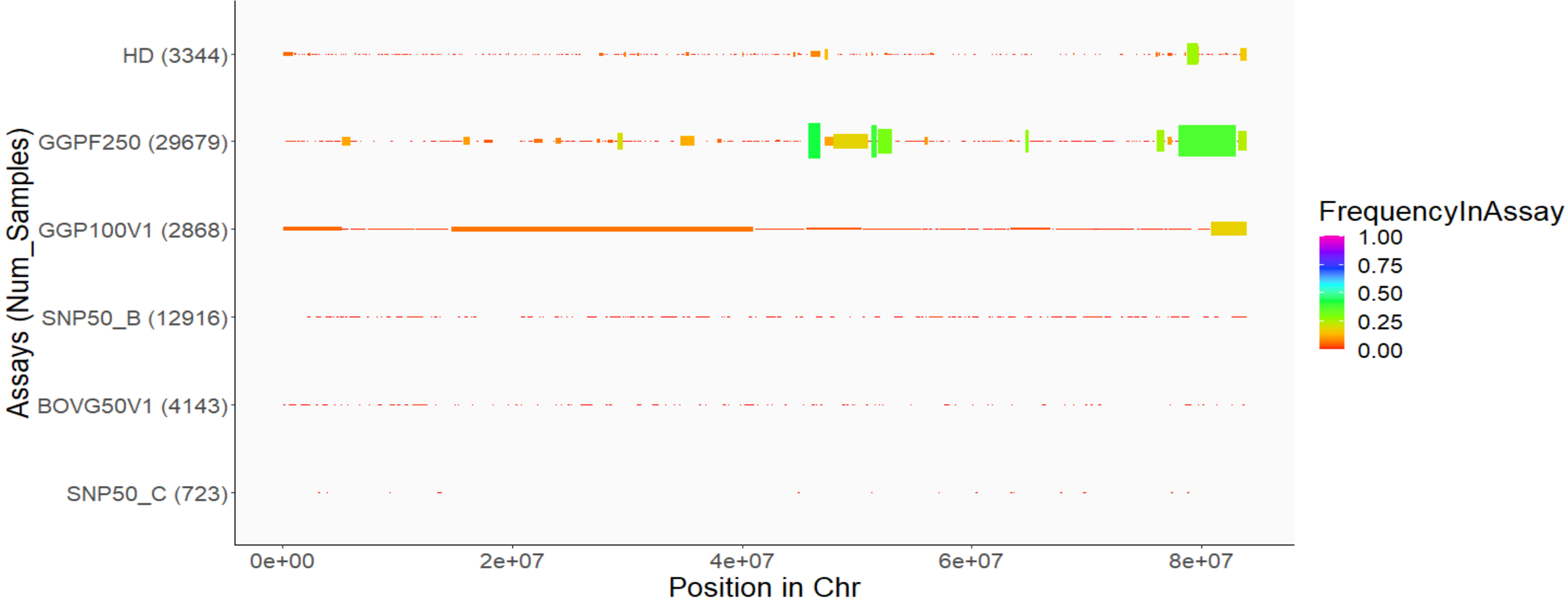
Requirements

1. CNVR must be detected in at least 2 Assays.
2. CNVR must be present in either the HD or GGPF250 Assay
3. Frequency of CNVR > 0.01% across all samples



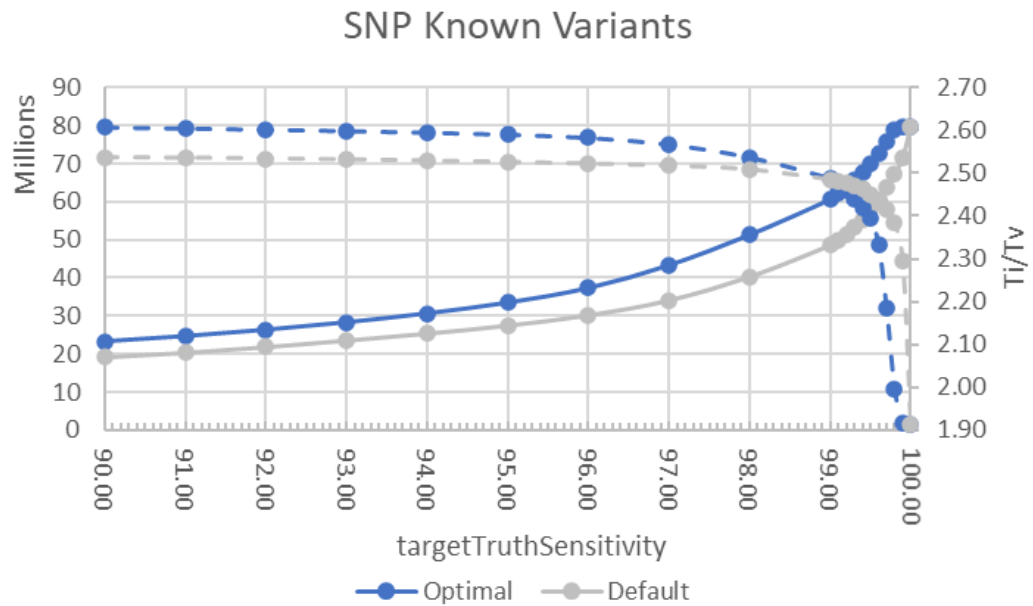
Comparisons Across Assays

Chr 15 CNVR Positions



GGPF250 Comparing Breeds Using Jaccard Similarity Coefficient

Sample Size	13407	1958	1798	1753	1477	1167	733	543	367	339	307	39	11	11	9	8	8	8	8	7	7	3	3	1
Breed	AN	HO	HFD	RAN	BG	SIM	BR	GEL	CHA	SH	LM	ANR	GIR	SGT	JER	NDAM	NEL	PIED	RMG	BRVH	GNS	BEFM	SHK	CHIA
AN	4,873	0.315	0.381	0.375	0.306	0.310	0.262	0.205	0.219	0.141	0.167	0.045	0.037	0.035	0.031	0.017	0.028	0.037	0.027	0.009	0.018	0.008	0.009	0.020
HO	945	3,670	0.404	0.365	0.439	0.373	0.344	0.352	0.343	0.260	0.302	0.106	0.088	0.087	0.082	0.043	0.066	0.094	0.071	0.021	0.046	0.019	0.023	0.052
HFD	970	556	3,872	0.411	0.391	0.418	0.342	0.349	0.329	0.256	0.282	0.089	0.069	0.068	0.064	0.033	0.050	0.074	0.055	0.016	0.036	0.015	0.018	0.042
RAN	1,005	603	652	4,093	0.373	0.428	0.326	0.303	0.277	0.234	0.248	0.080	0.064	0.064	0.058	0.030	0.048	0.067	0.051	0.015	0.033	0.014	0.017	0.038
BG	949	473	559	598	3,600	0.386	0.401	0.337	0.318	0.275	0.294	0.109	0.093	0.089	0.082	0.043	0.069	0.093	0.070	0.022	0.045	0.020	0.025	0.054
SIM	938	489	541	567	482	2,852	0.363	0.388	0.300	0.332	0.295	0.117	0.089	0.090	0.084	0.041	0.067	0.097	0.073	0.020	0.047	0.019	0.024	0.056
BR	957	483	556	594	462	467	3,742	0.335	0.282	0.277	0.289	0.115	0.105	0.098	0.089	0.044	0.079	0.101	0.077	0.015	0.049	0.023	0.028	0.059
GEL	910	398	471	520	400	378	377	2,408	0.353	0.381	0.375	0.182	0.141	0.140	0.138	0.066	0.094	0.157	0.120	0.030	0.078	0.030	0.040	0.092
CHA	938	436	514	567	442	440	429	325	2,562	0.276	0.298	0.136	0.115	0.111	0.114	0.055	0.077	0.129	0.096	0.022	0.064	0.026	0.033	0.070
SH	907	378	456	498	371	347	345	239	296	1,526	0.345	0.234	0.186	0.170	0.180	0.075	0.117	0.201	0.157	0.029	0.103	0.034	0.051	0.133
LM	909	385	467	513	385	377	362	258	311	218	2,024	0.184	0.149	0.158	0.146	0.083	0.116	0.166	0.134	0.033	0.091	0.039	0.047	0.100
ANR	902	347	442	484	344	332	313	201	251	143	170	507	0.315	0.256	0.294	0.101	0.177	0.316	0.272	0.055	0.221	0.059	0.095	0.300
GIR	902	346	443	484	342	333	309	202	249	142	169	58	403	0.327	0.413	0.150	0.286	0.449	0.382	0.039	0.291	0.112	0.176	0.314
SGT	901	344	441	482	341	331	309	200	248	142	166	59	50	529	0.383	0.242	0.288	0.406	0.391	0.036	0.316	0.131	0.198	0.269
JER	901	342	439	481	340	329	308	197	244	138	164	54	44	44	387	0.257	0.247	0.548	0.573	0.031	0.419	0.132	0.214	0.329
NDAM	901	343	440	482	340	331	309	199	246	139	162	53	44	39	35	419	0.175	0.219	0.249	0.052	0.281	0.220	0.279	0.123
NEL	903	347	444	486	344	334	310	204	252	145	168	59	48	47	45	37	528	0.238	0.257	0.030	0.237	0.125	0.188	0.203
PIED	901	343	440	482	341	330	310	198	246	140	166	58	47	47	40	41	50	364	0.528	0.033	0.381	0.113	0.191	0.310
RMG	901	342	439	481	340	329	308	197	245	137	162	52	43	41	34	32	42	38	340	0.040	0.453	0.147	0.233	0.304
BRVH	901	343	440	482	340	331	311	198	246	138	162	48	41	39	36	23	35	41	32	145	0.043	0.061	0.024	0.021
GNS	901	342	439	482	340	329	308	196	244	136	161	48	39	37	31	25	35	36	28	23	288	0.175	0.248	0.297
BEFM	901	343	440	482	340	330	307	197	245	137	160	47	37	35	31	19	31	37	28	15	20	207	0.227	0.076
SHK	901	343	440	482	340	329	307	197	244	136	161	46	37	34	31	19	31	36	27	17	20	13	218	0.152
CHIA	901	343	439	481	340	329	307	196	245	134	162	47	40	40	35	30	39	40	33	26	26	24	24	167



Tranche 99.00 Ti/Tv 2.49
Default 94,611,143
Optimal 116,026,576 (+22.6%)
“free” data ^^

