Downloaded from genome.cshlp.org on July 16, 2026 . Published by Cold Spring Harbor Laboratory Press

Overcoming limitations to customize DeepVariant for domesticated animals with TrioTrain

Jenna Kalleberg, Jacob Rissman and Robert D Schnabel

Genome Res. published online June 2, 2025
Access the most recent version at doi:10.1101/gr.279542.124

P<P Published online June 2, 2025 in advance of the print journal.

Accepted Manuscript Peer-reviewed and accepted for publication but not copyedited or typeset; accepted manuscript is likely to differ from the final, published version.

Open Access Freely available online through the *Genome Research* Open Access option.

Creative Commons License This manuscript is Open Access.This article, published in *Genome Research*, is available under a Creative Commons License (Attribution 4.0 International license), as described at http://creativecommons.org/licenses/by/4.0/.

Email Alerting Service Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or **click here.**

To subscribe to *Genome Research* go to:
https://genome.cshlp.org/subscriptions

Published by Cold Spring Harbor Laboratory Press

Downloaded from genome.cshlp.org on July 16, 2026 . Published by Cold Spring Harbor Laboratory Press

Overcoming limitations to customize DeepVariant for domesticated animals with TrioTrain

Jenna Kalleberg[1], Jacob Rissman[1], Robert D. Schnabel[1,2]

[1] University of Missouri, Division of Animal Sciences, Columbia, MO, 65201 USA

[2] University of Missouri, Genetics Area Program, Columbia, MO, 65201 USA

Corresponding author: schnabelr@missouri.edu

Downloaded from genome.cshlp.org on July 16, 2026 . Published by Cold Spring Harbor Laboratory Press

ABSTRACT

Generating high-quality variant callsets across diverse species remains challenging as most bioinformatics tools default to assumptions based on human genomes. DeepVariant (DV) excels without joint genotyping while offering fewer implementation barriers. However, the growing appeal of a "universal" algorithm has magnified the unknown impacts when used with non-human species. We use bovine genomes to assess the limits of using human-genome-trained variant callers, including the allele frequency channel (DV-AF) and joint-caller DeepTrio (DT). Our novel approach, TrioTrain, automates extending DV for diploid species lacking Genome-in-a-Bottle (GIAB) resources, using a region shuffling approach to mitigate barriers for SLURM-based clusters. Imperfect animal truth labels are curated to remove Mendelian discordant sites before training DV to genotype the offspring correctly. With TrioTrain, we use cattle, yak, and bison trios to create the first multi-species-trained DV-AF checkpoint. Although incomplete bovine truth sets constrain recall within challenging repetitive regions, we observe a mean SNV F1 score >0.990 across new checkpoints during GIAB benchmarking. With HG002, a bovine-trained checkpoint (28) decreased the Mendelian Inheritance Error (MIE) rate by a factor of two compared to the default (DV). Checkpoint 28 has a mean MIE rate of 0.03 percent in three bovine interspecies cross genomes. These results illustrate that a multi-species, trio-based training strategy reduces inheritance errors during single-sample variant calling. While exclusively training with human genomes deters transferring deep-learning-based variant calling to new species, we use the diverse ancestry within bovids to illustrate the need for advanced tools designed for comparative genomics.

Downloaded from genome.cshlp.org on July 16, 2026 . Published by Cold Spring Harbor Laboratory Press

INTRODUCTION

Machine learning applications in genomics are not new (Benzer 1959; Camin and Sokal 1965; Beyer et al. 1974). Nevertheless, the type and scale of available data continue to change what is possible through genomics-specific, deep-learning (DL) tools. Foundational, large-scale sequencing consortiums for human genomes often prompt similar work in other species. Several global efforts within the bovine genomics community run parallel to those for humans, including the Thousand Bull Genomes (1kBulls), Bovine Pangenome, and Ruminant Telomere-to-Telomere (RT2T) consortia (The 1000 Genomes Project Consortium et al. 2010; Jarvis et al. 2022; Liao et al. 2023; Kaminow et al. 2022; Nurk et al. 2022; Hayes and Daetwyler 2019; Zhou et al. 2022a; Smith et al. 2023; Kalbfleisch et al. 2024). Unfortunately, the development of species-agnostic tools has not kept pace with the exponential growth of sequenced biodiversity. Despite widespread variation in ploidy, chromosome number, or kinship, many bioinformatic tools expect the human genome by default. However, optimism surrounds deep-learning technologies and their potential translation into non-human species (Webb 2018). Context matters when evaluating the performance of any statistical inference tool. When trained exclusively with human genomes, new data from other species can affect precision (Sundaram et al. 2018). The burden of customizing these leading-edge tools often deters application with animal genomes. Comparative genomics researchers instead rely on human-based default parameters that may be inappropriate for other species (Genereux et al. 2020).

Human-centric tools include the most popular variant calling software, the GATK Best Practices (McKenna et al. 2010). In non-human genomics, alterations to the Best Practices are common, but evaluation approaches between studies are rarely standardized. Similar pervasive inconsistencies in human genomics motivated the Genome-in-a-Bottle (GIAB) consortium to create the first “highly accurate set of genotypes across a genome” using a multiplatform, ensemble curation approach (Zook et al. 2014). After the benefits of trio-based variant curation during assembly polishing were demonstrated in humans and cattle (Koren et al. 2018), the human GIAB resources were expanded to include complete trios (Zook et al. 2019). These resources remain the authority for assessing variant discovery methods (Olson et al. 2023; Chin et al. 2020; Wagner et al. 2022). Identifying genetic variation depends upon three components: the reference genome, the sequencing technologies, and the algorithms applied (Bickhart et al. 2020; Atkinson et al. 2023). A standardized ground truth from GIAB enables objective decisions when comparing

Downloaded from genome.cshlp.org on July 16, 2026 . Published by Cold Spring Harbor Laboratory Press

approaches for human genomics research (Olson et al. 2022a). For example, recent work found that variant discovery software can significantly impact genotype quality compared to the alignment method (Barbitoff et al. 2022). Benchmarking competitions revealed the limits of heuristics-based variant discovery methods like GATK; in contrast, deep-learning-based methods can identify the unknown (Sapoval et al. 2022).

For human genomics research, GIAB provides standards for communicating optimal method selection (Olson et al. 2022b). However, their authoritative catalog of genomic variation has become the default source of training data for deep learning in genomics, such as DeepVariant (DV). As the first deep-learning-based variant caller to achieve a small yet significant improvement over GATK, DeepVariant's innovation stems from approaching variant calling as an image classification problem (Poplin et al. 2018a). After converting pileup alignments into multi-layer tensors or channels, the Convolution Neural Network (CNN) uses the context surrounding a putative variant to distinguish noise from truth; relevant context is encoded within a checkpoint for genotyping. DeepVariant's specialty is providing checkpoints that control the known and unknown systematic bias for a wide range of sequencing technologies, with several notable extensions. Recent work used the 1000 Genomes Project cohort to supplement DV with an allele frequency channel to create a population-aware variant caller (DV-AF) (The 1000 Genomes Project Consortium. 2010; Chen et al. 2023). DV and DV-AF avoid GATK's scalability limitations by achieving comparable accuracy without requiring joint genotyping (Poplin et al. 2018b; Chen et al. 2023). As a joint caller, DeepTrio (DT) uses separate parental and offspring checkpoints to identify variants in related duo or trio samples (Kolesnikov et al. 2021). Furthermore, the haplotype-aware PEPPER-Margin-DeepVariant pipeline is a common approach for small variant detection with PacBio's long-read sequencing (Shafin et al. 2021). Despite relying heavily on the human GIAB reference materials for training, subsequent research has repeatedly demonstrated DeepVariant's superiority (Supernat et al. 2018; Zhao et al. 2020; Lin et al. 2022; Olson et al. 2022a; Barbitoff et al. 2022). DeepVariant's success inspired the rapid development of new methods for variant discovery (Sahraeian et al. 2019; Luo et al. 2020; Ip et al. 2020; Ramachandran et al. 2021; Khazeeva et al. 2022; Su et al. 2022; Popic et al. 2023; Ahsan et al. 2023). Although DL technologies saturate the leading edge of genomics, most overlook the challenges facing non-human genomics, limiting their relevance across biological diversity.

Downloaded from genome.cshlp.org on July 16, 2026 . Published by Cold Spring Harbor Laboratory Press

The revolutionizing impact of deep-learning-based approaches is attributed to publicly accessible genomic data and ground-truth variant callsets (Olson et al. 2023). Unfortunately, these resources only exist for human genomes, stagnating tool development for other species. Early research applied DV in mice and rice to demonstrate the variant caller's translational capabilities (Day and Poplin 2019; Poplin et al. 2018a). However, differences in variant density and distributions of nucleotide content impact DeepVariant's ability to perform equally across species (Yun et al. 2018; Sapoval et al. 2022). Custom DV checkpoints are necessary to adjust for distributional changes relative to prior experience with the human genome. Until now, their development has been driven by substantial divergence between the training and the target species, such as mosquitoes and Kākāpō parrots (Yun et al. 2018; Guhlin et al. 2023). Despite starting with the human-based DV, these resulting checkpoints are considered 'species-specific' because customization across taxonomy may inhibit generalization with humans. As the 'default' genome, models exclusively trained with human genomes are implicitly trusted in other species – the customization for Kākāpō parrots relied on truth labels created with DV (v0.9) (Guhlin et al. 2023). Recent work in cattle also relied on defaults intended for human genomes for both DV and GATK (Lloret-Villas et al. 2021, 2023). They argue DeepVariant is superior to GATK for bovine short-read sequencing (SRS); however, their conclusions were based on a prevalent limitation within animal genomics: expecting the GATK Best Practices to be optimal across species. Due to sample size (N~500), their approach with GATK required manually selecting site-level thresholds (hard-filtering). Despite these nuances, DV has seen a rapid expansion in comparative genomics, including domesticated and endangered animals, plants, and pathogenic organisms (Leonard et al. 2023; Stenløkk et al. 2022; Flack et al. 2023; Secomandi et al. 2023; Guhlin et al. 2023; Osipowski et al. 2020; Ruperao et al. 2023; Gunderson et al. 2020; Jesudoss Chelladurai et al. 2023). Researchers studying a wide range of species are drawn to DeepVariant because of substantially fewer implementation barriers than GATK.

Here, we address the non-trivial technical barriers of customizing DV for species lacking GIAB-quality truth callsets using TrioTrain, which enables automatic training extension with new data on a SLURM-based computing cluster. Existing trio-based joint-calling companion software (dv-trio, DT) requires samples with known pedigree during training and genotyping (Ip et al. 2020; Kolesnikov et al. 2021). In contrast, the curriculum within TrioTrain indirectly encodes inheritance expectations by training

Downloaded from genome.cshlp.org on July 16, 2026 . Published by Cold Spring Harbor Laboratory Press

simultaneously with two genomes that happen to be parent-offspring duos to produce a single-sample variant caller. We release the current version of our customized, trio-based, multi-species-trained DV-AF checkpoint, which reduces inheritance errors in human and bovine genomes. Our research illustrates the limitations of using the default, human-genome-trained DeepVariant in other species. Although we demonstrate how to repurpose existing non-human genomics data for deep learning, we emphasize that maximizing the quality of models built with TrioTrain will require expanding the number of organisms with comprehensive and accurate variant benchmarks.

RESULTS

Initial variant quality indicated a training gap.

As mosquito and Kākāpō genomes are appreciably more diverged from humans than cattle, we initially expected re-training to be unnecessary. For our preliminary evaluation, we used one bovine genome (SRA BioSample: SAMN1094070) to compare single-sample calling with DeepVariant (v1.0) against an existing bovine callset from multi-sample calling with GATK (v3.8). These data were created during the 1000 Bull Genomes (1kBulls) project, a global consortium that standardized methods for population-based analyses with cattle (Hayes and Daetwyler 2019). Three runs were produced after the release of the current reference genome, ARS-UCD1.2 (Rosen et al. 2020), with Run7 in 2019 (N=3818), Run8 in 2020 (N=4931), and Run9 in 2021 (N=6191). Over time, detection power improved with the increasing sample size, resulting in “the largest sequence repository of bovine genomic variation” (Dorji et al. 2024). Unlike the typical hard filtering in animal genomics, the sample size of the 1kBulls cohort enabled Variant Quality Score Recalibration (VQSR) to reduce systematic error. Although some of these data are publicly available, they are not analogous to the GIAB variant benchmarks.

In animal genomics, method evaluation typically relies on indirect quality metrics due to slight variations in context and software across studies. For example, we initially assessed the two variant calling methods using existing data produced during 1kBulls Run8 (2020). For GATK (v3.8), a single-sample VCF was extracted from the larger Run8 cohort. For DeepVariant (v1.0), a single-sample VCF was generated by providing the same BAM file as GATK. Then, the two approaches were compared using BCFtools isec (v1.15) (Danecek et al. 2021), constrained to Chromosome 29. We inferred variant quality using the

Downloaded from genome.cshlp.org on July 16, 2026 . Published by Cold Spring Harbor Laboratory Press

transition to transversion (Ti:Tv) ratio, where values below the expected genome-wide ratio of 2.0–2.3 are interpreted as lower quality (Crysnanto et al. 2019). We found variants unique to the population-scale, VQSR-filtered GATK callset achieved a marginally higher Ti:Tv ratio (+0.14 points; **Figure S1**), suggesting variants from DeepVariant were lower quality relative to the 1kBulls Run8 cohort. Based on these findings, we began developing TrioTrain to explore if DV could benefit from further training with bovine genomes.

Recently, a small cohort of dairy cattle (N=50) was used to compare callsets resulting from GATK (v4.2) and DV (v1.2) (Lloret-Villas et al. 2023). While they also observed more unique variants with DeepVariant than with GATK, they reported a slight increase in Ti:Tv with DV (+0.04 points) compared to GATK after manually selecting site-level filtering thresholds (hard filtering) (**Table S1**). Given the contradictory results across studies, we repeated our evaluation between GATK (v3.8) and DeepVariant (v1.4), using Ti:Tv to measure variant quality indirectly. For this analysis, we identified a subset of taurine cattle from the 1kBulls cohort with publicly available short-read sequencing data (N=13, **Table S1**). We performed single-sample calling with DeepVariant to compare against joint-calling with GATK with VQSR filtering (UMAGv1 callset, see Methods). Relative to our initial evaluation, mean Ti:Tv increased for concordant variants (+0.26 points). While unique variant quality with DV increased (+0.20 points), the mean genome-wide Ti:Tv ratio was slightly higher (+0.06 points) with UMAGv1 GATK-based callset. These contradictory results demonstrate the limitations of relying on indirect quality metrics for species lacking variant benchmarks.

Label inconsistency limits model quality

The challenge for training species-specific models is maximizing the quality of the labels used as truth. Here, we infer truth using SRS data and GATK, which introduces a small amount of noise that is indistinguishable from genuine variation. The initial build of TrioTrain relied on truth labels from the 1kBulls Run8 callset (**Note S1**). After adapting training to our SLURM-based HPC cluster, we assessed these existing data as potential truth labels by filtering to increase specificity. During these preliminary analyses, we reviewed plots for model loss – an indicator of prediction error that training seeks to minimize (**Figure S2**). Despite applying more strict filters (GQ>30), the loss hovered consistently around 0.2 instead of approaching zero. Subsequent testing discovered an order-of-operations flaw during filtering, creating

Downloaded from genome.cshlp.org on July 16, 2026 . Published by Cold Spring Harbor Laboratory Press

inconsistencies between the truth VCF and the PopVCF. These results indicated overfitting and potentially noisy labels. Based on these findings, we excluded the 1kBulls Run8 callset as potential training labels; however, subsequent analysis improved upon these existing standards for population-scale variant calling in cattle using GATK.

We began by examining genotyping consistency between the three 1kBulls genotyping runs (2019–2021) by selecting samples included in all three joint-calling runs (N=3659; mean coverage = 13.7×). From each cohort VCF, a sample's PASS genotypes were extracted into a single-sample VCF, enabling a pairwise comparison using established methods (Krusche et al. 2019). To our knowledge, this is the most extensive comparison with these data, requiring >1.5M CPU hours (**Note S2**). In theory, the genotypes for a sample should be identical (F1 score >0.990); instead, we observed lower recall and precision across the 1kBulls runs than expected (mean F1 scores between 0.95-0.97) (**Table S2**). These findings reveal variability between genotypes derived from the same underlying sequence data, affecting all downstream analyses that assume accurate and consistent genotypes. Given these limitations, we improved upon the 1kBulls variant calling protocol by iteratively altering parameters during VQSR, resulting in a new variant callset (UMAGv1). Animal genomics typically settles for performing VQSR once or hard-filtering, using parameters intended for human genomes (Lloret-Villas et al. 2023). Instead, we tested a range of values for each tunable parameter, resulting in approximately 300–1,000 VQSR runs. From these analyses, we selected VQSR parameter values that produced fewer violations of Mendelian inheritance within bovine trios while increasing the total number of PASS variants. We then compared our optimal parameters against the defaults and the 1kBulls Run9 (2021). The new UMAGv1 callset contains more PASS variants and 2× fewer Mendelian Inheritance Errors (MIEs) (**Supplemental Methods: Figure S30**). While our research demonstrates current boundaries for producing higher-quality variant callsets in animal genomics, we provide further evidence that the default values intended for human genomes introduce error when used in other organisms. Relative to the 1kBulls Run9, the quality improvements within the UMAGv1 callset enabled us to reconsider extending DV with bovine genomes.

Downloaded from genome.cshlp.org on July 16, 2026 . Published by Cold Spring Harbor Laboratory Press

Building TrioTrain to design a bovine-specific curriculum

A lack of high-quality benchmarking variants has thwarted most pursuits for non-human DV models. Previous extensions of DV focused on nonmammalian diploid organisms while altering the training curriculum (comparative summary in **Table S3**) (Yun et al. 2018; Guhlin et al. 2023). For example, in the mosquito-specific approach, the parental genomes were never given to DeepVariant but instead used to curate inherited variants within the offspring (Yun et al. 2018). Our approach modifies the mosquito-specific strategy based on pedigree structure, genome length, and sample size. For cattle, violations of Mendelian inheritance patterns were attributed to systematic error during SRS since de novo variants are comparably rare (Harland et al. 2017; Lee et al. 2023). Based on these expectations, our trio-based variant curation excluded sites violating Mendelian inheritance expectations within a complete bovine trio. Relative to mosquitoes, we hypothesized that the larger sample size of reported pedigrees across bovids could potentially offset imperfect truth labels. However, agricultural genomics historically prioritized larger sample sizes with lower coverage. Since DV-AF had improved accuracy in lower coverage human SRS samples, we include population allele frequencies from the UMAGv1 callset when creating new models with TrioTrain (Chen et al. 2023).

Using the reported pedigree from public SRS data, we identified known families within the UMAGv1 cohort, resulting in 15 complete trios from multiple bovine species (**Table 1**). With these trios, we built 30 bovine-trained DV-AF checkpoints sequentially over five phases, exploring potential strategies for customizing DV. Our trio-based experimental design structures the training phases into biological questions to provide insight into DV's behavior. For example, as with prior nonmammalian approaches, our training strategy during the first two phases focuses on creating a 'cattle-specific' model, using data typical for domesticated animal genomics. Training begins with six cattle trios representing multiple unique cattle breeds, while the second phase adds breed-specific replicates. Next, we investigate if DV's species-specific experience could be extended for comparative genomics since accuracy in humans previously increased with diversity (Chen et al. 2023). The third phase uses two bison trios with the same father to determine whether a multi-species, 'bovine-specific' model was possible. The final two phases maximize truth-label confidence, demonstrating the value of a potential cattle-specific, GIAB-quality resource. The boundaries of quality are rapidly developing as the Q100 project works towards a perfect, diploid, telomere-to-telomere

Downloaded from genome.cshlp.org on July 16, 2026 . Published by Cold Spring Harbor Laboratory Press

(T2T) genome assembly from a GIAB sample (HG002) (Phillippy and Hansen 2023; Rautiainen et al. 2023). The potential for minimal per-base error for both haplotypes is shifting GIAB's previous read-reference variant benchmarks toward assembly-reference benchmarks. For cattle, we relied on the ongoing Bovine Pangenome project, which uses trio-binning to produce haplotype-resolved alternative reference assemblies. Multiple platforms were used to sequence three F1 crosses with extant bovine lineages, including the taurus and indicus subspecies, Angus-Brahman (Low et al. 2020), and two interspecific crosses, Highlander-Yak and Simmental-Bison (Rice et al. 2020; Heaton et al. 2021; Oppenheimer et al. 2021). During our analysis, these reconstructed haploid parental assemblies represented the highest quality assembly-based truth set for cattle (**Supplemental Methods: Table S11**). However, the increased heterozygosity in these hybrid offspring challenges DeepVariant's capabilities with genotype class imbalance (**Table S4**). Therefore, phase 4 strategically uses these data after exhausting standard cattle and bison trios. Given the prohibitive curation effort required to build the original GIAB resources, we considered the potential of using synthetic truth labels in animal genomics. For phase 5, we used synthetic SRS reads created for the Angus × Brahman hybrid offspring (Methods). As publicly reported pedigree records will restrict other animal species, we contrast each phase of our multi-trio approach against three single-trio replicates. The first represents a pragmatic scenario of repurposing a trio sequenced only with SRS (checkpoint 2). Lastly, we perform two parallel single-trio experiments, resulting in four additional bovine-trained DV-AF checkpoints. These represent an alternative strategy of training with the highest-quality bovine truth labels immediately, effectively skipping the first 20 iterations to only extend DeepVariant with the original SRS data from a hybrid trio, either the Angus-Brahman or the Yak-Highlander (checkpoints 2B and 2C, respectively).

With TrioTrain, DeepVariant is given an entire trio after two sequential iterations, one for each parent (Methods). Within a trio, both iterations use the same labeled examples from the child to select an optimal checkpoint that distinguishes inherited variants from noise. TrioTrain automatically splits each trio into parent-offspring duos – the parental genome provides new context during training, which is continuously evaluated using the offspring, also known as tuning (**Figure 1A**). When provided a set of genomes previously withheld from training or evaluation, the pipeline will automatically estimate generalization, known as testing. These bovine genomes (N=19; **Figure 1B**) include real SRS from the

Downloaded from genome.cshlp.org on July 16, 2026 . Published by Cold Spring Harbor Laboratory Press

offspring of the three F1 crosses, three corresponding offspring replicates with synthetic SRS, along with 13 non-pedigreed taurine cattle that remain independent because truth labels are withheld from DeepVariant (**Table S5**). Bovine data was selected to be representative of publicly available SRS, with similar mean coverage (**Figure 1C**). TrioTrain is designed to enable training with any diploid vertebrate species and begins by initializing weights or warm-starting from the existing human-trained DV checkpoint (**Figure 1D**). Bovine truth sets used for either training, evaluation, and testing were constrained to the autosomes and X Chromosome, then restricted with a sample-specific callable regions BED file generated by GATK (Methods). As such, the UMAGv1 bovine truth sets are more conservative than similar files provided by GIAB. The number of steps during a training iteration varies based on the example variants produced for a parent; within each trio, all training examples are used once to prevent overfitting. Tuning with the offspring continuously calculates stratified performance metrics. Pedigree information is not explicitly provided to DeepVariant; instead, the checkpoint that performs best with the offspring is selected as the warm-starting point for the next iteration. Then, the entire re-training process is iteratively repeated until all trios have been exhausted.

Factors that influence training behavior

When used appropriately, deep-learning-based methods are superior because they can account for the numerous subtle factors that alter distributions within the data. However, the nature of DL typically obscures which features are relevant for accurate prediction; previous work suggests the importance of demographic history, variant distribution and density, and genome complexity (Zou et al. 2019; Whalen et al. 2021; Musolf et al. 2022). With TrioTrain, the metrics from training and evaluation indicate if the parental truth labels helped DV accurately predict the offspring's genotypes. However, our ability to estimate training performance depends on the definition of truth. False Positives (FP) include any variant discovered by DeepVariant that GATK misses. Variants DV ignores within the UMAGv1 truth labels are considered False Negatives (FN). True Positives (TP) are an exact match between DeepVariant and GATK. These definitions are essential caveats as FP errors may be novel variants missing from the GATK-based truth labels. We measure performance using the harmonic mean of precision and recall, known as F1 score (TP / TP + ½ [FP + FN]), where a value of one indicates perfect prediction. Since model selection relies on F1 score, we

Downloaded from genome.cshlp.org on July 16, 2026 . Published by Cold Spring Harbor Laboratory Press

compared the successively created bovine DV-AF checkpoints stratified by variant type (**Figure 2A**) and genotype class (**Figure 2B**). Performance assessment of TrioTrain starts within a trio (x-axis: odd–even pairs). For example, in Figure 2A, comparing two checkpoints (27–28) evaluated using the same offspring reveals that maternal examples (28) slightly improved prediction within the offspring relative to only paternal examples (27). Alternatively, comparing performance between trios by reviewing patterns across several checkpoint pairs (i.e., 3–4 vs. 11–12) helps reveal bovine genome characteristics that alter DeepVariant's behavior, indicating a distribution change relative to the GIAB trios (**Table S6**).

Building the mosquito-specific DV revealed that the default checkpoint struggled with higher variant density; notably, cattle have approximately 40 percent more nucleotide diversity and approximately 1.5× more small variants than humans despite similar genome lengths (Gibbs et al. 2009). Other factors, such as the ratios for variant type and genotype class, reflect subtle nuances learned by DeepVariant. Essentially, DV anticipates a certain level of imbalanced classification from the human GIAB samples (**Table S6**). With balanced data, the ratio between predicted classes approaches one. In contrast, training with GIAB truth sets gave DV over 6× more SNV examples (mean SNV:indel = 6.58). Cattle are expected to have similar distributions (SNV:indel = 7.05) (Crysnanto et al. 2019), which is reflected during the first 18 iterations of TrioTrain (mean SNV:indel = 7.54, **Tables S6, S7**). However, the diversity of the bovine trios in subsequent iterations provides substantially more SNVs (mean SNV:indel = 9.02) relative to DV's expectations. Training performance over multiple bovine genomes reflects how these data deviate from the established human-genome priors.

Multiple factors contribute to the priors learned by DeepVariant. Heterozygous variant confidence decreases with coverage (<30×) because fewer reads at a locus can introduce allele sampling bias. With SRS, indels are notoriously challenging because mapping accuracy declines for reads that cannot span the variant. By relying on a single sequencing platform, some bovine indels detected by DV will be incorrectly labeled as errors (FPs) simply because they are missing from the bovine truth set. Despite these caveats, overall performance across all iterations is robust, with a median optimal checkpoint F1 Score of 0.9865 (**Table S8)**. The training performance observed during phases 1 and 2 demonstrates that high-quality variant callsets derived from existing, population-scale SRS can be curated into 'silver-standard' training data. However, this required exhaustively improving joint-calling accuracy before training; therefore, we

Downloaded from genome.cshlp.org on July 16, 2026 . Published by Cold Spring Harbor Laboratory Press

caution that our findings represent an upper limit threshold rather than typical performance with non-human SRS data. Across the 30 bovine DV-AF checkpoints, the F1 score for SNVs approaching one (**Figure 2A**) indicates that DV successfully uses the parent genome to inform offspring variant calling (median SNV F1 score = 0.991), with one iteration (29) deviating from the others. Phase 5 begins with replicating the parental genome in the previous phase (23); however, the offspring used for model tuning were inspired by pseudo-diploid (SynDip) benchmarking strategies in humans (Li et al. 2018). Synthetic SRS reads were sampled from the reconstructed parental haplotype assemblies, resulting in a higher per-base error rate than the offspring's Illumina sequencing data. Since we do not intend to call variants in synthetic data, we halted training with other synthetic offspring samples.

Investigating why some bovine genomes underperform during training provides more insight into the human training curriculum. Previous work found that including the allele frequency channel (DV-AF) stabilized performance in lower-coverage human genomes (Chen et al. 2023). Since allele frequencies from the UMAGv1 cohort were included during training, our bovine-trained DV-AF should handle lower coverage genomes relatively well. As expected, regardless of parental coverage being above or below 30×, we observe robust offspring F1 score (**Figure S4E**), with most checkpoints achieving high recall (fewer FNs) during training. Performance with indels was lower than SNVs, which is expected given how the bovine truth labels were curated (Methods). Although variable coverage has minimal impact on offspring SNV precision (**Figure S4C**), six lower-coverage parents produce a cluster of checkpoints with decreased training indel precision (**Figure S4D**).

However, interpreting change during training from a discrete attribute such as coverage overlooks DeepVariant's ability to learn complex and hidden interactions. The cluster of checkpoints indicates that training underperforms specifically with parents from the same breed as the reference genome (HE = Hereford) and lower coverage (<20×) (Rosen et al. 2020). Unlike the composite human reference genome (GRCh38), the cattle reference assembly represents a single, highly inbred cow named Dominette. This reference captures most variation between Hereford individuals, typically resulting in fewer homozygous alternate (HomAlt) genotypes relative to cattle from other breeds. While the raw variant count of HomAlts for Hereford cattle is similar to the GIAB genomes (mean HomAlt/genome: humans ~1.60M; HE cattle ~1.68M), the species difference is more pronounced for Hets (mean Hets/genome: humans ~2.3M; HE

Downloaded from genome.cshlp.org on July 16, 2026 . Published by Cold Spring Harbor Laboratory Press

cattle ~3.4M). Relative to other cattle breeds used during training, the truth labels for the HE offspring contain proportionally more Hets (**Table S7**), making it imperative for DV to experience similar data during training. Gravitating towards 'breed-specific' or 'species-specific' explanations may be intuitive; however, DeepVariant remains unaware of those labels. Instead, these differences become evident to DV through novel characteristics within the provided training callsets.

Training DV with bovine genomes shifts several distributions of the predicted classes. For example, the Het:HomAlt ratio is influenced by depth of coverage and variant density. Deviating from the priors of the human GIAB samples (mean Het:HomAlt = 1.46) during training is one way for DV to gain new experience. For example, aligning bison genomes to the cattle reference, on average, leads to nearly 9× more HomAlt genotypes in the truth set relative to previous iterations with cattle (mean Het:HomAlt = 0.21; **Figure S3E, Table S6**). The lower training recall (**Figure S5F**) and the corresponding decrease in offspring F1 score (**Figure S5B**) for the minor class (HomAlts) during phase 3 are expected because the bison truth labels contain proportionally more HomAlts. A similar performance shift occurs when training with the interspecific crosses during phase 4; the hybrid offspring provide nearly 6× more Het examples than all previous offspring (**Table S6**). DV must adapt to the increased heterozygosity in the hybrid offspring relative to the parents (**Figure 2C**). Increasing the examples from the major class (Hets) (**Figure S3D**) leads to a corresponding recall decrease (**Figure S5E**). These shifts in training performance for specific classes reflect the bovine training data diverging from the human-based priors, a necessary step to build new experience. Our results indicate that truth variants from related species can provide additional context for accurate variant discovery.

Assessing generalization across species

Recommending a TrioTrain checkpoint for variant calling requires confirmation of accurate genotype predictions outside the training data, known as model testing. We assessed the 30 new bovine-trained DV-AF checkpoints created during sequential iterations with TrioTrain against existing versions (DV and DV-AF). These checkpoints were first tested using a multi-breed set of taurine cattle genomes (N=13) that were not used previously during either training or tuning (**Table S5, top**). Unlike training, the truth genotypes for these samples are withheld from DeepVariant but do not exclude Mendelian discordancies

Downloaded from genome.cshlp.org on July 16, 2026 . Published by Cold Spring Harbor Laboratory Press

due to incomplete pedigree or missing parental genotypes. Applying the existing human-trained DV to cattle genomes does not severely decrease testing recall (**Figure S6**), indicating that priors from the human genome are informative for genotyping cattle. However, a sudden decline in testing recall occurs when using Angus paternal examples to predict variation in the Angus-Simmental hybrid cross offspring. During this iteration (23), DeepVariant recalls Hets (**Figure S6C**) at the cost of missing SNVs, indels, and HomAlts. With our trio-based training approach, testing recall recovers after the next iteration (24) provides maternal examples. The limitations of the UMAGv1 truth labels become apparent when comparing precision and recall within a particular variant type. Testing precision increases slightly between iterations 1 and 28 for indels (**Figure S6B**). Since fewer indel examples are provided per training iteration relative to SNVs, more iterations are required to accumulate the same volume of examples. However, the corresponding decrease in testing recall for the same class indicates that the bovine truth set is overly conservative (fewer indels).

As expected, single-sample variant calling with existing human-trained checkpoints introduces variants relative to the bovine GATK-based truth labels (UMAGv1), resulting in low testing precision (**Figure S6**) and F1 score (**Figure S7**). Further exploration of these putative false positives focused on a single genome – the same individual as the cattle reference (SAMN03145444). Variants missing from the bovine truth set correctly identified by the human-trained DV include Het SNVs within homopolymers (**Figure S8**) or BovB repetitive regions (**Figure S9**). Including the UMAGv1 allele frequencies reduced HomAlt errors in low-coverage regions (**Figure S10**) as intended: extending DV with a single cattle genome (iteration 1) results in a 4.88-fold reduction in these HomAlt errors. Although some FPs remain with an altered genotype (16 percent), a manual review of HomAlt SNVs introduced by the default checkpoint that were resolved after the first TrioTrain iteration confirmed that training was necessary. Heterozygous variant density likely contributed to some actual genotyping errors made by DeepVariant (**Figures S11, S12**). Adjusting the heterozygosity priors learned from humans by training in cattle also improves genotyping error from DV (**Figures S13, S14**).

The relative similarity between bovine and human genomes enables comparison across taxonomies unreported by prior nonmammalian extensions of DV (Yun et al. 2018; Guhlin et al. 2023). Given the potential ambiguity with the non-pedigreed bovine testing genomes, they were supplemented with the three real and three synthetic bovine offspring, along with the six human genomes (**Table S5,**

Downloaded from genome.cshlp.org on July 16, 2026 . Published by Cold Spring Harbor Laboratory Press

bottom). Although the truth labels for these six bovine offspring represent the highest-confidence truth sets within the UMAGv1 cohort, we can only describe changes to generalization relative to GATK. In contrast, the GIAB trios enable definitive benchmarking as previously described (Krusche et al. 2019). Six checkpoints built with TrioTrain were selected to simplify comparison between species and across TrioTrain phases (bold vertical lines **Figure 2**). Performance with these multi-trio checkpoints (**Figure 3**) is contrasted against two additional single-trio checkpoints (2B, 2C; **Figure S16**), along with existing versions of the joint-caller (DT=DeepTrio) and the two single-sample, human-trained versions (DV and DV-AF). Given that the UMAGv1 truth set is expected to miss some indels, models built with these data are expected to do the same in bovine (**Figure 3A, 3B**) and human genomes (**Figure 3C, 3D**). All eight bovine-trained checkpoints generalize well in both species for all other classifications relative to indels (**Figure S15**). However, out of the eight bovine DV-AF checkpoints, checkpoint 28 (model.ckpt-282383) maximized SNV F1 score in bovine (max = 0.997239; N=19) and human (max = 0.992415; N=6) samples.

During benchmarking with the GIAB trios, the best-performing TrioTrain checkpoint expectedly underperformed relative to those exclusively trained with those samples (**Figures 4A, B**). Genome-wide performance with the Ashkenazi Jewish (AJ) offspring HG002 confirmed that a multi-trio approach outperforms all single-trio checkpoints (**Figure S16**). The slight decrease during human benchmarks is primarily due to training with a different definition of truth. For example, bovine truth labels were constrained to a sample-specific callable region BED file (Methods). While the current versions of the GIAB truth sets include complex and challenging regions, the bovine truth sets are constrained to variation detected confidently with SRS, limiting the amount of novel, difficult examples. Challenging or repetitive regions are not excluded since they are not characterized to the same extent as in humans. As such, the bovine-trained checkpoints have lower recall in complex and repetitive regions, regardless of species. Outside known segmental duplication regions in HG002, the bovine-trained checkpoint (28) is nearly perfect in SNVs (**Figure 4C**) and substantially improved for indels (**Figure 4D**). Lower performance within segmental duplications (**Figures 4E, 4F**) indicates that training with bovine genomes alters DeepVariant's expectations of copy number variation in addition to the heterozygosity priors. Without a definitive ground truth specific to bovine genomes, selecting the optimal checkpoint required further analyses.

Downloaded from genome.cshlp.org on July 16, 2026 . Published by Cold Spring Harbor Laboratory Press

Fewer Mendelian Inheritance Errors without joint genotyping

Given the imperfect bovine truth labels, we used an additional metric, the Mendelian Inheritance Error (MIE) rate, to assess performance across species, inferring that a lower rate indicates improvement. Although DeepVariant was not directly given pedigree information, our training approach is designed to reduce the number of Mendelian discordant genotypes during variant calling. For these analyses, we used the GIAB Ashkenazi Jewish (AJ) and Han Chinese (HC) trios, along with three bovine F1-hybrids (Methods). Across the remaining eight checkpoints, bovine-trained DV-AF (28) resulted in proportionally fewer Mendelian discordant SNVs in all three F1-hybrid bovine trios, even compared to DeepTrio (**Figure 5A**). Joint calling with DeepTrio (DT) typically outperforms other versions of DeepVariant (DV); however, the wall time to produce results varies significantly, depending on sample coverage and variant content. DeepVariant produces results in as little as six hours when embarrassingly parallelized, while DeepTrio takes at least two days to achieve the same results. TrioTrain models reduce the MIE rate for individual samples without joint calling, retaining inherited variants in less time.

As with bovine trios, checkpoint 28 achieves the lowest Mendelian inheritance error rate in the AJ offspring (HG002). However, the new checkpoint calls approximately 31k fewer indels than the comparable human version (DV-AF), illustrating the trade-off of using our bovine-trained model in human genomes: under-calling indels. Given the curation process of SRS-based bovine truth labels, the observed limitation is expected. Notably, performance with the HC offspring (HG005) behaves inconsistently after re-training with cattle – the MIE rate nearly doubles with the first two phases of TrioTrain (**Figure 5B**). However, performance recovers over the entire training period, with the bovine DV-AF checkpoint (28) achieving the same Mendelian discordancy rate as DV while producing 86k more variants. While the existing human-trained models identify more variants in bovine samples, they also introduce more variants that violate Mendelian inheritance expectations. As the number of calls made in the same individual varies between checkpoints, we sorted the AJ trio VCFs by genotype quality (GQ) in the offspring (HG002). The resulting MIE calibration curve illustrates that TrioTrain substantially reduces inheritance errors relative to existing human-trained single-sample checkpoints, more closely resembling DeepTrio (**Figure 5C**). Based on these findings, we selected checkpoint 28 (model.ckpt-282383, see Data Access) as the best-performing DV-AF built with TrioTrain.

Downloaded from genome.cshlp.org on July 16, 2026 . Published by Cold Spring Harbor Laboratory Press

During a manual review of Mendelian inheritance errors in HG002, we contrast discordant genotypes from the default version (DV) that were concordant with the bovine-trained checkpoint (28). Most MIEs were indels (DV: 87%, 28: 67%), which are difficult to resolve due to representation differences between individuals. Discordant SNVs predominantly fell outside the benchmark regions for HG002 (DV: ~94%, 28: ~91%), which were excluded from manual review. We randomly selected a subset of discordant SNVs with known genotypes for the trio. With these, the human-trained DV typically introduced discordant HomAlts that were labeled as concordant Hets with the bovine-trained DV-AF checkpoint. Relative to HG002's GIAB truth calls, genuine errors with checkpoint 28 occurred because the context provided by short-read sequencing alone contradicted the GIAB truth genotypes (**Figures S17-S19**). The discordant variants produced due to genotyping errors from DV were typically not made by the new checkpoint 28 (**Figures S20-S22**). The genuine discordant FPs with DV that were subsequently corrected by our checkpoint were commonly found in HG003 (Father) – the only GIAB sample consistently withheld during prior training – indicating that training in bovine samples provided relevant context for genotyping humans. Further review of known de novo variants in the GIAB offspring focused on biallelic Mendelian discordant SNV (**Table S12**). Of the four checkpoints compared (DT, DV, DV-AF, 28), DeepTrio (DT) produced more True Positive genotypes, but all expected Mendelian discordancies became concordant. However, sites that remained discordant with the other checkpoints occurred due to inconsistent genotypes rather than correctly inferring de novo variation. During these analyses, a more recent version of DeepTrio (v1.6) introduced a feature designed to preserve de novo variant recall by increasing the importance of known variants during training. Without LRS for other species, this feature could potentially improve genotyping accuracy for the single-sample variant callers given validated de novo variation (Porubsky et al. 2025).

Recommendations for applying TrioTrain in other species

Our results demonstrate that an optimal variant caller for human genomes may not be appropriate in all mammals. However, using the TrioTrain pipeline for species-specific customization of a DeepVariant checkpoint requires substantial upfront investment in maximizing callset quality. Avoiding missteps during truth set curation requires appreciating how every prior step affects the data (Whalen et al. 2021). As such, model builders should identify any limitations with their current genotyping approach in samples

Downloaded from genome.cshlp.org on July 16, 2026 . Published by Cold Spring Harbor Laboratory Press

representative of the intended cohort. First, compare summary distributions (i.e., depth of coverage, variant density, SNV:indel, Het:HomAlt, Ti:Tv) relative to all prior data used to train a particular DeepVariant checkpoint (**Tables S6, S7**) as unexpected deviations may indicate re-training is necessary. Even if high-quality training data or trios are unavailable, we encourage standardized reporting of these variant calling metrics when using human-based defaults in other species. Comparative data would help justify developing more diverse benchmarking resources and quantify progress over time.

Next, consider the impacts of selecting training samples based on the trios available, focusing on sample quality rather than quantity. Since the multi-trio approach shows incremental improvement over a single-trio approach (**Figure S16**), multiple trios would be ideal but not mandatory. We expect fewer than 14 trios could produce a generalized species-specific model with strategic training sample selection. In domesticated animal species, we recommend using breeds that differ from the reference genome, prioritizing diverse training samples over replicates within the same breed. Our results suggest selecting unrelated trios rather than using shared pedigree, such as half-sibling offspring. Pragmatically, a minimum of two trios would allow one for training, with the second remaining independent for testing. While single-sample variant calling with the optimal bovine-trained checkpoint (28) requires known population allele frequencies, these data are not mandatory to pursue extending the default DeepVariant with TrioTrain.

Before pursuing training, identify potential strategies for maximizing genotyping confidence while avoiding overly strict filtering. Currently, alternative variant calling methods remain essential to build independent species-specific truth labels despite the burden of optimization. Although the human-trained versions of DeepVariant can discover potential novel variants, using these loci as truth variants requires tedious manual inspection. Most novel variants identified by DV within cattle trios were ambiguous, where truth was not easily identifiable when visualizing aligned short-reads in IGV (v2.16) (Robinson et al. 2011). Resolving these difficulties would benefit from long-read sequencing, but unfortunately, these data do not exist for many of the cattle trios from historical data (phases 1 and 2). We acknowledge that reliance on SRS data alone to build bovine truth sets impacts training accuracy with indels and highly repetitive regions, restricting generalization with human genomes. These limitations may require strategically adding the human GIAB trios between species-specific iterations until the characterization of these genomic regions in other species improves.

Downloaded from genome.cshlp.org on July 16, 2026 . Published by Cold Spring Harbor Laboratory Press

Fortunately, efforts to improve non-human assembly quality are continuously expanding the number of samples sequenced with multiple platforms, simultaneously creating new sources of training data. However, we recommend estimating the per-base error rate before creating synthetic training data from any assembly. Using synthetic reads during the last bovine training phase increased the SNV MIE rate in all trios compared to the prior phase (**Figure 5A, B**). Given the reported Quality Values (QV) (**Table S11**), we do not consider synthetic data derived from these three existing bovine assemblies to be a viable source of training data. Regardless of species, prioritizing assembly quality in multiple individuals is essential for increasing the diversity of assembly-based truth labels.

DISCUSSION

The remaining challenges within genome biology require technologies designed for the continuously expanding diversity of sequencing data. Variant discovery is the first of many hurdles encountered during comparative genomics research, resulting in a trade-off between quality and time. The difficulty of optimizing methods for new species leads many to rely on the defaults intended for the human genome. Studying critically endangered species adds urgency to the decision, as recording population diversity is essential to monitor decline and prioritize intervention. However, the growing prevalence of deep-learning-based methods introduces unknown effects of implicitly trusting a sophisticated model, as illustrated by a previously customized model for the endangered Kākāpō parrots (Guhlin et al. 2023). Prior to re-training, the small population size (<200) impacted GATK's accuracy, resulting in a higher rate of Mendelian inheritance errors compared to DeepVariant (v0.9) (3.38% and 1.28%, respectively). However, using DV to create the Kākāpō truth labels leads to circular training that may unintentionally retain bias towards the human genome. The SNV Mendelian inheritance error rate reported for Kākāpō is 2.6× higher than our benchmarks with the default DeepVariant (v1.4) with the GIAB samples (mean: 0.49%). Despite potential bias from the human genome, the Kākāpō DV callset was curated by altering Mendelian discordant sites in the offspring to match the parental genotypes. There are always limitations during experiment design, particularly in novel or endangered species with inherently small sample sizes, but our research with TrioTrain quantifies their magnitude.

Downloaded from genome.cshlp.org on July 16, 2026 . Published by Cold Spring Harbor Laboratory Press

Regardless of the method, relying on defaults intended for human genomes has consequences. Parallel research using the default human-genome-trained DeepVariant (v1.5) to polish cattle assemblies reported double the SNV error rate (1.06%) relative to our benchmarks with humans (Leonard et al. 2024). For context, prior to removing Mendelian discordant sites, the VQSR-optimized GATK callsets for within-species trios (phases 1 through 3) have a minimum MIE rate of 1.14% (mean = 2.40%, max = 3.92%). The overlapping error rate between VQSR-optimized GATK and human-trained DV illustrates the nuances of method selection in non-human genomics. Although a lack of genomic benchmarks in cattle obstructs optimal method selection, releasing these data (VQSR-optimized UMAGv1) represents the first step towards a GIAB-quality ground truth in bovids and other agricultural species. Although creating a cattle-specific variant benchmark is beyond the scope of this manuscript, we are in the early stages of developing a GIAB-quality variant benchmark in a new cattle trio, starting with developing cell lines to ensure the long-term value of the resource.

As these resources continue to mature, potential limitations must be readily transparent. For example, previous work using human cohorts illustrated the pathogenic impacts of VQSR due to the omission of rare and de novo variants (AlDubayan et al. 2020). By excluding all discordant sites from the UMAGv1 trio VCFs, we compromise unknown de novo variation to address systematic errors during variant calling, which has limitations. For example, the predicted genotype for the offspring can be correct yet classified as discordant due to parental genotyping errors. Systematic errors could lead to a concordant site with incorrect genotypes in all three individuals. Filtering out discordant variants will also exclude some genuine indels represented differently between individuals. However, we emphasize that the UMAGv1 cohort callset exceeds typical variant quality compared to the community standards in animal genomics (1kBulls). Short of producing species-specific GIAB benchmarks, these data serve as a current upper limit of variant quality achievable for any agricultural species. While referred to as “truth labels,” the genotypes used for training must be considered preliminary. They cannot represent all possible true variants; instead, they are regions we can reliably and accurately genotype in cattle with SRS data. As such, the truth labels exclude highly repetitive and GC-rich regions due to shorter read lengths and platform-specific limitations. Given these factors, the bovine-trained checkpoint is more conservative than the human-trained default in the GIAB samples. Unfortunately, many non-human species lack extensive characterization of complex or

Downloaded from genome.cshlp.org on July 16, 2026 . Published by Cold Spring Harbor Laboratory Press

repetitive genome regions; without these resources, customization of DV in non-human species will continue to struggle with these areas.

Despite these caveats, our bovine-trained model reduces the Mendelian error rate in bovine trios by an order of magnitude compared to the default DeepVariant checkpoint, confirming that TrioTrain enables DeepVariant to select model weights to identify inherited variants. We release this initial version to bridge the divide between GATK and DV for cattle genomics while minimizing the potential of making incorrect biological inferences in cattle. For example, previous work found that DeepVariant increased the proportion of HomAlt genotypes, which increased the amount of high-impact functional consequences predicted in cattle (Lloret-Villas et al. 2023). Until more comprehensive cattle-specific variant benchmarks compel re-training, the bovine-trained checkpoint (28) reduces implementation barriers compared to GATK while improving upon the current standards (1kBulls). The established tools, resources, and population-scale SRS data for cattle often guide research strategies in other agricultural species, so the initial success with bovine genomes provides encouraging evidence about replicating our work. Although we anticipate that checkpoint 28 will produce high-confidence callsets of common variation in other domesticated animals, future work must evaluate how well our checkpoint generalizes in new species relative to those exclusively trained with human genomes and evaluate the impact of warm-starting with a revised human DV-AF model (v1.4+) that combines both channels (The 1000 Genomes Project Consortium et al. 2010; Chen et al. 2023). For now, those working in other species must remain skeptical of any downstream consequences and pursue validation. We caution against inferring that the performance of any variant caller will replicate consistently in all other species, including the model accompanying this manuscript.

Prioritizing human genomes has enabled many advances in specialized deep-learning methods, but comparative genomics research depends on high-quality variant callsets from multiple species. Previous work focused on building species-specific models that require continuous re-training every time the human model is improved. Developing separate models for each species is cost-prohibitive across the exponentially growing species list with sequencing data. However, TrioTrain establishes an iterative, trio-based training approach for customizing single-sample variant calling checkpoints to any diploid species with the necessary input data. The pipeline focuses on reducing the computational burdens required to

Downloaded from genome.cshlp.org on July 16, 2026 . Published by Cold Spring Harbor Laboratory Press

customize DeepVariant. With TrioTrain, we provide the first DeepVariant model trained in multiple mammalian species for comparative genomics research.

Our results demonstrate that historical sequencing data can adjust DeepVariant to the distributional shifts that arise when aligning multiple related species against the same reference genome. Given that training with a bison trio improves training performance in homozygous alternate variants, our work suggests previously unknown benefits of training within genera. Future research must determine if the human-trained model would benefit from re-training with other hominid species. Exploring DeepVariant's behavior during training with high-quality data from more organisms may reveal previously hidden unknowns in genomics. For example, learning behavior across more species could reveal new insight based on the phylogenic boundaries that achieve optimal generalization. With similar high-quality variant benchmarks, DeepVariant could feasibly become a 'golden model' for comparative genomics. Compared to the human-specific DeepVariant, a theoretical species-agnostic variant caller would be generalizable across broader phylogenic classifications but could also struggle with rare variant discovery in humans. In building TrioTrain, we provide a tool for exploring training across species to illustrate how the behavior of deep-learning models enables investigating the persistent challenges in genomics. Future innovation requires looking beyond the human genome; our work provides an initial template for creating diverse, pedigree-based benchmarking resources for other species.

METHODS

Raw sequencing data

The University of Missouri Animal Genomics (UMAG) group maintains a repository containing the publicly available genomes for several agricultural species, including bovines. As new samples are published, the paired-end, short-read Illumina WGS data is obtained from the NCBI's Sequence Read Archive (SRA). We identify bovine samples based on SRA metadata for taxon ID for multiple species (**Note S3**). Where available, we also collect reported sex, breed, and pedigree information. Samples are aggregated based on BioSample ID and assigned an internal, integer-based UMAG lab ID unique to a single animal. We refer to the bovine cohort as the UMAGv1 cohort (**Note S4**). At the time of analysis, the cohort included 5,612 individuals with complete metadata, spanning modern and ancient genomes from

Downloaded from genome.cshlp.org on July 16, 2026 . Published by Cold Spring Harbor Laboratory Press

related bovine species: *Bos taurus taurus* (Taurus cattle), *Bos taurus indicus* (Indicine cattle), *Bison bison* (American Bison), *Bos mutus* (Yak), *Bos grunniens* (Yak), *Bos frontalis* (Gayal), *Bos javanicus* (Banteng), and *Bos gaurus* (Gaur).

Whole genome sequence alignment

After downloading the raw FASTQ files from SRA, the data are pre-processed (**Notes S5 – S8**). Briefly, low-quality bases and adapters were trimmed from the reads with Trimmomatic (v0.38 or v0.39) (Bolger et al. 2014) and subsequently aligned to the reference genome, ARS-UCD1.2 with Btau5.0.1 Y Chromosome (Rosen et al. 2020) using either BWA-MEM (v0.7.17) (Li 2013) or BWA-MEM2 (2.2.1) (Md et al. 2019). The resulting sequence alignment map (SAM) output is sorted, indexed, and compressed into binary alignment map (BAM) files. Duplicate reads are marked with Picard MarkDuplicates (v2.18.19 or v2.26.10) (Broad Institute 2019). Our internal pipeline automatically performs indel realignment and base quality score recalibration (BQSR) and produces callable region BED files (**Notes S9 and S10, Data Access**). Note that DeepVariant does not require the realignment and recalibration of raw data. Coverage and other summary metrics were calculated using SAMtools (v1.9) (Danecek et al. 2021) (**Note S11**). The average coverage for the entire cohort before sample QC was 12.77× (minimum = 0.07×, maximum = 182.78×).

Initial variant calling

The processed sequence data was used for germline variant calling for single nucleotide variants (SNVs) and short inserts and deletions (indels, <50bp). We refer to the resulting VCF as the UMAGv1 callset. Genotypes for the UMAGv1 cohort were prepared for all 5,612 bovine samples, including outgroup species, using an internal workflow that expands upon the established 1kBulls GATK FASTQ to GVCF guidelines (Hayes and Daetwyler 2019). The 1kBulls protocol uses GATK HaplotypeCaller (v3.8) (Poplin et al. 2018b) to generate single sample callsets, which are merged before joint genotyping (**Notes S12 and S13**). However, using bovine trios, our internal workflow is further optimized (**Figures S23 – S32; Note S14 – S17**). In brief, significant improvement was made during Variant Quality Score Recalibration (VQSR) as parameters were extensively tuned from defaults intended for human genomes.

Downloaded from genome.cshlp.org on July 16, 2026 . Published by Cold Spring Harbor Laboratory Press

Sample selection

A subset of individuals from the final UMAGv1 callset were selected as the high-quality bovine genotypes used for re-training DeepVariant. We used mean sequencing coverage to select candidate samples from the UMAGv1 cohort with moderate sequencing depth by excluding samples with a mean coverage of <10×. Related samples were identified using reported pedigree metadata. Candidates for training samples were constrained to parent-offspring trios by excluding samples missing pedigree data or samples with sequencing data for only one parent. We identified 14 unique trios, resulting in 40 unique variant datasets with a mean coverage of 25.99× (minimum = 10.66×, maximum =50.83×). These 14 trios consist of sires (N=12), dams (N=14), and their offspring (N=14) across multiple cattle breeds and crossbreeds (**Note S18, Table S4**). In addition to the trios, we identified other genomes we could use to represent the final data given to the deep-learning model. These genomes were used to monitor performance changes over the iterative re-training process and assess the generalizability of each model created with TrioTrain. We selected 13 samples from multiple cattle breeds with a mean coverage of 29.63× (minimum = 13.54×; maximum = 60.00×). These testing genomes included representatives from 15 cattle breeds, with eight additional breeds not used during training (**Note S19, Table S5**).

Bovine truth set curation

From the UMAGv1 cohort dataset, we extracted variants for all three samples within a trio into family VCF files (**Note S20**). Briefly, within each family VCF, we removed any Mendelian discordant genotypes. For example, at one locus, if one parent is A/B and the other parent is B/B, but the child has a genotype A/A, that position was dropped from all three individuals. We then split the family VCF into single-sample VCFs. Due to a lack of parental sequencing data, Mendelian discordant variants could not be removed for the non-trio samples used during testing. The single-sample VCFs used for training, tuning, and testing were then filtered to retain PASS filter sites only, where PASS was set to tranche level 99.7. Lastly, to meet the specifications required by DeepVariant for labels, the final VCFs exclude all homozygous reference genotypes.

Downloaded from genome.cshlp.org on July 16, 2026 . Published by Cold Spring Harbor Laboratory Press

Synthetic bovine offspring

Synthetic diploid (Q) approaches have been proposed as alternatives to GIAB-quality benchmarking data (Li et al. 2018). These efforts inspired us to consider if ongoing efforts to produce multiple high-quality bovine assemblies could be used to generate high-quality synthetic benchmark variants for cattle (Smith et al. 2023; Zhou et al. 2022b). To compare both strategies of truth label curation in non-human species, we used the haploid parental assemblies to create synthetic reads for the offspring in all three trios (**Note S21**). Briefly, we created synthetic reads from each parental assembly using NEAT (v3.2) (Stephens et al. 2016), introducing no sequencing error, then merged these synthetic reads, resulting in three more offspring replicates. As such, each new trio consists of the parents' original SRS data, but the offspring is substituted with the synthetic SRS. These F1-hybrid offspring have unique identifiers, separate truth labels, and a different noise signature relative to the non-synthetic samples. The three unique F1-hybrid offspring and their respective synthetic replicates (N=6) have a mean coverage of 26.7× (minimum coverage = 14.46×; maximum coverage = 50.01×).

Shuffling labeled examples

First, TrioTrain (v0.8) calculates the number of regions based on the variants within the Truth VCF input for a genome (**Notes S22 and S23**). These regions' BED files are passed to DeepVariant through the `--regions` flag. We designed region shuffling to batch the complete genome to create representative samples of the total examples. These region-specific steps are submitted across compute nodes as independent SLURM jobs. Crucially, a region-specific tfrecord file contains examples produced from across the genome. Next, within-region shuffling is performed, and then all tfrecord "shards" are concatenated into a single file. Finally, the resulting labeled and shuffled examples across all regions are fed to the DeepVariant model in a randomized order. Our Regions Shuffling process successfully enables the DeepVariant Beam shuffling pipeline on a SLURM-based HPC cluster. Rather than rely on a different resource manager, our pipeline creates embarrassingly parallel shuffling jobs. Our design avoids extensive system-administrator support to create a Spark cluster on top of the existing SLURM-controlled cluster. Instead, our approach uses Beam's Direct Runner to shuffle examples. Direct Runner is constrained to the

Downloaded from genome.cshlp.org on July 16, 2026 . Published by Cold Spring Harbor Laboratory Press

local memory within a single compute node and is typically applied for pipeline debugging. However, SLURM effectively controls multiple independent Direct Runner jobs simultaneously.

Initializing model weights

Due to the wide variability in coverage levels within the complete UMAGv1 cohort, we opted to add the allele frequency channel during our re-training. Before re-training beings, an existing model checkpoint from human genomes is used to initialize model weights, or 'warm-starting.' At the time of analysis, the only available WGS-specific model with DeepVariant (v1.4) added a new channel for insert size compared to previous model versions (v1.1). Since the additional channel was not yet incorporated into the existing human-trained DV-AF, we used the new DV checkpoint as the training starting point. TrioTrain enabled building new checkpoints that encode for both insert size and population-level allele frequencies (**Note S24**). We created the PopVCF using the UMAGv1 callset by dropping genotypes and retaining only variant allele frequencies for the joint genotypes across the entire cohort (N=5,612).

Re-training

Running TrioTrain requires organizing metadata and file paths into a single metadata CSV file (**Note S21**). The number of examples created and used varies for each parent (**Table S6**). For re-training the network, we used a learning rate of 0.005, a batch size of 32, and defined the training period as one epoch that spans all labeled candidate variants, known as examples, for a training genome. Performance metrics in the offspring (**Table S7**) from the selected best checkpoint for all 34 iterations were stratified by variant type and genotype class (**Table S8**).

Summarizing performance

We obtained the 1000 Genomes Project's population VCF and the GRCh38 reference genome from Google Genomics. We then acquired the GIAB benchmarking files (v4.2.1), corresponding sample-specific callsets, and callable region files from the National Institute for Standards and Technology (NIST) FTP site (**Note S25**). We used two trios: (1) Ashkenazi trio with HG002_NA24385_son, HG003_NA24149_father, HG004_NA24143_mother; (2) Han Chinese trio with HG005_NA24631_son,

Downloaded from genome.cshlp.org on July 16, 2026 . Published by Cold Spring Harbor Laboratory Press

HG006_NA24694_father, HG007_NA24695_mother. Benchmarking parameters were identical to those used to calculate bovine performance (**Note S26**). The only difference is switching to the human reference genome, truth labels, callable regions, and the corresponding model checkpoints. The resulting VCFs were then passed to hap.py (v0.3.12) using the appropriate reference genome and truth file. Performance metrics, such as recall, precision, and F1 score, were stratified by variant type (SNV, indel) and genotype class (HomRef, Het, HomAlt). Raw metric values were collected and plotted using a custom Python (v3.12) script. Summary metrics from a subset of checkpoints are available in bovine trios (**Table S9**) and human GIAB samples (**Table S10**).

Validating performance Using Mendelian discordancy

Without a benchmarking callset for bovine genomes, we cannot directly compare the F1 score across species. However, "most methods used to identify and genotype genetic variants from [high-throughput] sequencing data ignore the relationships between samples, resulting in significant Mendelian errors, false positives, and negatives" (Cleary et al. 2014). Therefore, we used the Mendelian Inheritance Error (MIE) rate to infer relative performance changes between species. We used Real Time Genomics (RTG) software, rtg-tools mendelian (v3.12.1), to calculate the MIE rate (Cleary et al. 2015). For this analysis, we again used the same models from GIAB benchmarking: five terminal checkpoints from the TrioTrain phases and the three existing DeepVariant models (v1.4: default WGS, WGS.AF, DeepTrio). However, we used human and bovine trios, specifically, the two GIAB trios and the three F1-hybrid trios described previously. Each sample was re-run using DeepVariant separately, except with DeepTrio (DT), a joint caller; all models created per-sample VCFs as output (**Note S27**).

Datasets

A subset of the data for the two most recent genotyping runs produced by the 1000 Bull Genome (1KBulls) project are available at the European Nucleotide Archive (ENA; https://www.ebi.ac.uk/ena/browser/home) under accession numbers PRJEB42783 and PRJEB56689. For bovine samples, all raw sequencing data used in this study were obtained from the NCBI Sequence Read Archive (SRA; https://www.ncbi.nlm.nih.gov/sra) under accession numbers provided in the Supplemental Material (training and evaluation: **Table S4**, testing: **Table S5**). The Supplemental Material for human

Downloaded from genome.cshlp.org on July 16, 2026 . Published by Cold Spring Harbor Laboratory Press

GIAB benchmarking samples also provide details about how the files were obtained for this study (**Note S25**).

Software Availability

Copies of the source code for the TrioTrain pipeline v0.8, as well as custom scripts used in this study, are provided as Supplemental Code. Further details about the source code, installation instructions, documentation, and tutorials are available at GitHub (https://github.com/jkalleberg/DV-TrioTrain/releases/tag/v0.8). Source code used for sequence data processing (**Notes S4 – S11**), using GATK (**Notes S12 – S17**), and producing input files for TrioTrain (**Note 20**) is provided as Supplemental Material.

Data access

The variant data generated in this study have been submitted to the European Nucleotide Archive (ENA; https://www.ebi.ac.uk/ena/browser/home) under accession number PRJEB86883. Additional data including pedigree and breed labels (metadata.csv), the per-sample callable region files (CallableRegions.tar.gz), UMAGv1 cohort population allele frequency file (UMAG1.POP.FREQ.vcf.gz), reference genome file (ReferenceGenome.tar.gz), and the final selected TrioTrain checkpoint (28; ModelCheckpoint.tar.gz) can be found at Zenodo (https://doi.org/10.5281/zenodo.15482484).

COMPETING INTEREST STATEMENT

The authors declare that they have no competing interests.

ACKNOWLEDGMENTS

The research reported here was partly supported by NIH NLM with award number T32LM012410 and USDA NIFA with award numbers 2020-67015-31675, 2023-67015-39261, and HATCH MO HAAS0001. We thank the MU Animal Genomics group for their feedback and discussions on this work. The computation for this work was performed on the high-performance computing infrastructure operated by

Downloaded from genome.cshlp.org on July 16, 2026 . Published by Cold Spring Harbor Laboratory Press

Research Support Solutions in the Division of IT at the University of Missouri, Columbia, MO. DOI: 10.32469/10355/97710.

AUTHOR CONTRIBUTIONS

RDS and JK conceived the study. RDS performed the initial variant calling with GATK, optimized the VQSR parameters to produce the final UMAGv1 callset, and curated the truth labels. JR developed code to identify callable regions. JK built the TrioTrain pipeline with input from RDS. JK performed all re-training iterations with bovine genomes and performed the subsequent analyses for model benchmarking, comparison, and evaluation. JK and RDS wrote the manuscript. All authors have read and approved the final manuscript.

REFERENCES

Ahsan MU, Liu Q, Perdomo JE, Fang L, Wang K. 2023. A survey of algorithms for the detection of genomic structural variants from long-read sequencing data. *Nat Methods* 20: 1143–1158.

AlDubayan SH, Conway JR, Camp SY, Witkowski L, Kofman E, Reardon B, Han S, Moore N, Elmarakeby H, Salari K, et al. 2020. Detection of pathogenic variants with germline genetic testing using deep learning vs standard methods in patients with prostate cancer and melanoma. *JAMA - Journal of the American Medical Association* 324: 1957–1969.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* 215: 403–410. https://linkinghub.elsevier.com/retrieve/pii/S0022283605803602.

Atkinson EG, Artomov M, Loboda AA, Rehm HL, MacArthur DG, Karczewski KJ, Neale BM, Daly MJ. 2023. Discordant calls across genotype discovery approaches elucidate variants with systematic errors. *Genome Res* 33: 999–1005.

Barbitoff YA, Abasov R, Tvorogova VE, Glotov AS, Predeus A V. 2022. Systematic benchmark of state-of-the-art variant calling pipelines identifies major factors affecting accuracy of coding sequence variant discovery. *BMC Genomics* 23: 1–17. https://doi.org/10.1186/s12864-022-08365-3.

Benzer S. 1959. On the Topology of the Genetic Fine Structure. *Proceedings of the National Academy of Sciences* 45: 1607–1620.

Beyer WA, Stein ML, Smith TF, Ulam SM. 1974. A molecular sequence metric and evolutionary trees. *Math Biosci* 19: 9–25. https://linkinghub.elsevier.com/retrieve/pii/0025556474900285.

Bickhart DM, McClure JC, Schnabel RD, Rosen BD, Medrano JF, Smith TPL. 2020. Symposium review: Advances in sequencing technology herald a new frontier in cattle genomics and genome-enabled selection. *J Dairy Sci* 103: 5278–5290. http://dx.doi.org/10.3168/jds.2019-17693.

Downloaded from genome.cshlp.org on July 16, 2026 . Published by Cold Spring Harbor Laboratory Press

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* 30: 2114–2120.

Broad Institute. 2019. Picard Toolkit. *GitHub*. https://broadinstitute.github.io/picard (Accessed November 24, 2023).

Camin JH, Sokal RR. 1965. A Method for Deducing Branching Sequences in Phylogeny. *Evolution (N Y)* 19: 311.

Chen NC, Kolesnikov A, Goel S, Yun T, Chang PC, Carroll A. 2023. Improving variant calling using population data and deep learning. *BMC Bioinformatics* 24.

Chin CS, Wagner J, Zeng Q, Garrison E, Garg S, Fungtammasan A, Rautiainen M, Aganezov S, Kirsche M, Zarate S, et al. 2020. A diploid assembly-based benchmark for variants in the major histocompatibility complex. *Nat Commun* 11: 1–9. http://dx.doi.org/10.1038/s41467-020-18564-9.

Cleary JG, Braithwaite R, Gaastra K, Hilbush BS, Inglis S, Irvine SA, Jackson A, Littin R, Nohzadeh-Malakshah S, Rathod M, et al. 2014. Joint variant and de novo mutation identification on pedigrees from high-throughput sequencing data. *Journal of Computational Biology* 21: 405–419.

Cleary JG, Braithwaite R, Gaastra K, Hilbush BS, Inglis S, Irvine SA, Jackson A, Littin R, Rathod M, Ware D, et al. 2015. Comparing Variant Call Files for Performance Benchmarking of Next-Generation Sequencing Variant Calling Pipelines. *bioRxiv* 1–6. https://doi.org/10.1101/023754.

Clift B, Haussler D, McConnell R, Schneider TD, Stormo GD. 1986. Sequence landscapes. *Nucleic Acids Res* 14: 141–158.

Crysnanto D, Wurmser C, Pausch H. 2019. Accurate sequence variant genotyping in cattle using variation-aware genome graphs. *Genetics Selection Evolution* 51.

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM. 2021. Twelve years of SAMtools and BCFtools. *Gigascience* 10.

Day A, Poplin R. 2019. Analyzing 3024 rice genomes characterized by DeepVariant. *Google Cloud Blog*. https://cloud.google.com/blog/products/data-analytics/analyzing-3024-rice-genomes-characterized-by-deepvariant (Accessed November 24, 2023).

Dorji J, Reverter A, Alexandre PA, Chamberlain AJ, Vander-Jagt CJ, Kijas J, Porto-Neto LR. 2024. Ancestral alleles defined for 70 million cattle variants using a population-based likelihood ratio test. *Genetics Selection Evolution* 56.

Flack N, Drown M, Walls C, Pratte J, McLain A, Faulk C. 2023. Chromosome-level, nanopore-only genome and allele-specific DNA methylation of Pallas's cat, *Otocolobus manul*. *NAR Genom Bioinform* 5. https://academic.oup.com/nargab/article/doi/10.1093/nargab/lqad033/7103190.

Genereux DP, Serres A, Armstrong J, Johnson J, Marinescu VD, Murén E, Juan D, Bejerano G, Casewell NR, Chemnick LG, et al. 2020. A comparative genomics multitool for scientific discovery and conservation. *Nature* 587: 240–245.

Gibbs RA, Taylor JF, Van Tassell CP, Barendse W, Eversole KA, Gill CA, Green RD, Hamernik DL, Kappes SM, Lien S, et al. 2009. Genome-Wide Survey of SNP Variation Uncovers the Genetic Structure of Cattle Breeds. *Science (1979)* 324: 528–532. https://www.science.org/doi/10.1126/science.1167936.

Downloaded from genome.cshlp.org on July 16, 2026 . Published by Cold Spring Harbor Laboratory Press

Guhlin J, Le Lec MF, Wold J, Koot E, Winter D, Biggs PJ, Galla SJ, Urban L, Foster Y, Cox MP, et al. 2023. Species-wide genomics of kākāpō provides tools to accelerate recovery. *Nat Ecol Evol*.

Gunderson EL, Vogel I, Chappell L, Bulman CA, Lim KC, Luo M, Whitman JD, Franklin C, Choi YJ, Lefoulon E, et al. 2020. The endosymbiont Wolbachia rebounds following antibiotic treatment. *PLoS Pathog* 16.

Harland C, Charlier C, Karim L, Cambisano N, Deckers M, Mni M, Mullaart E, Coppieters W, Georges M. 2017. Frequency of mosaicism points towards mutation-prone early cleavage cell divisions in cattle. *bioRxiv*. http://biorxiv.org/lookup/doi/10.1101/079863.

Hayes BJ, Daetwyler HD. 2019. 1000 Bull Genomes Project to Map Simple and Complex Genetic Traits in Cattle: Applications and Outcomes. *Annu Rev Anim Biosci* 7: 89–102. https://www.annualreviews.org/doi/10.1146/annurev-animal-020518-115024 (Accessed August 8, 2022).

Heaton MP, Smith TPL, Bickhart DM, Vander Ley BL, Kuehn LA, Oppenheimer J, Shafer WR, Schuetze FT, Stroud B, Mcclure JC, et al. 2021. A Reference Genome Assembly of Simmental Cattle, Bos taurus taurus. *Journal of Heredity* 112: 184–191.

Ip EKK, Hadinata C, Ho JWK, Giannoulatou E. 2020. dv-trio: A family-based variant calling pipeline using DeepVariant. *Bioinformatics* 36: 3549–3551.

Jarvis ED, Formenti G, Rhie A, Guarracino A, Yang C, Wood J, Tracey A, Thibaud-Nissen F, Vollger MR, Porubsky D, et al. 2022. Semi-automated assembly of high-quality diploid human reference genomes. *Nature* 611: 519–531.

Jesudoss Chelladurai JRJ, Abraham A, Quintana TA, Ritchie D, Smith V. 2023. Comparative Genomic Analysis and Species Delimitation: A Case for Two Species in the Zoonotic Cestode Dipylidium caninum. *Pathogens* 12.

Kalbfleisch TS, McKay SD, Murdoch BM, Adelson DL, Almansa-Villa D, Becker G, Beckett LM, Benítez-Galeano MJ, Biase F, Casey T, et al. 2024. The Ruminant Telomere-to-Telomere (RT2T) Consortium. *Nat Genet*.

Kaminow B, Ballouz S, Gillis J, Dobin A. 2022. Pan-human consensus genome significantly improves the accuracy of RNA-seq analyses. *Genome Res* 32: 738–750.

Khazeeva G, Sablauskas K, van der Sanden B, Steyaert W, Kwint M, Rots D, Hinne M, van Gerven M, Yntema H, Vissers L, et al. 2022. DeNovoCNN: a deep learning approach to de novo variant calling in next generation sequencing data. *Nucleic Acids Res* 50: e97.

Kolesnikov A, Goel S, Nattestad M, Yun T, Baid G, Yang H, McLean CY, Chang P-C, Carroll A. 2021. DeepTrio: variant calling in families using deep learning. *bioRxiv* 1–16.

Koren S, Rhie A, Walenz BP, Dilthey AT, Bickhart DM, Kingan SB, Hiendleder S, Williams JL, Smith TPL, Phillippy AM. 2018. De novo assembly of haplotype-resolved genomes with trio binning. *Nat Biotechnol* 36: 1174–1182. https://www.nature.com/articles/nbt.4277.

Krusche P, Trigg L, Boutros PC, Mason CE, De La Vega FM, Moore BL, Gonzalez-Porta M, Eberle MA, Tezak Z, Lababidi S, et al. 2019. Best practices for benchmarking germline small-variant calls in human genomes. *Nat Biotechnol* 37: 555–560. http://dx.doi.org/10.1038/s41587-019-0054-x.

Lee YL, Bouwman AC, Harland C, Bosse M, Moreira GCM, Veerkamp RF, Mullaart E, Cambisano N, Groenen MAM, Karim L, et al. 2023. The rate of de novo structural variation is increased in in vitro–produced offspring and preferentially affects the paternal genome. *Genome Res* 33: 1455–1464.

Downloaded from genome.cshlp.org on July 16, 2026 . Published by Cold Spring Harbor Laboratory Press

Leonard AS, Crysnanto D, Mapel XM, Bhati M, Pausch H. 2023. Graph construction method impacts variation representation and analyses in a bovine super-pangenome. *Genome Biol* 24.

Leonard AS, Mapel XM, Pausch H. 2024. Pangenome genotyped structural variation improves molecular phenotype mapping in cattle. *Genome Res* gr.278267.123. http://genome.cshlp.org/lookup/doi/10.1101/gr.278267.123.

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv*. http://arxiv.org/abs/1303.3997 (Accessed February 4, 2024).

Li H, Bloom JM, Farjoun Y, Fleharty M, Gauthier L, Neale B, MacArthur D. 2018. A synthetic-diploid benchmark for accurate variant-calling evaluation. *Nat Methods* 15: 595–597. http://dx.doi.org/10.1038/s41592-018-0054-7.

Liao WW, Asri M, Ebler J, Doerr D, Haukness M, Hickey G, Lu S, Lucas JK, Monlong J, Abel HJ, et al. 2023. A draft human pangenome reference. *Nature* 617: 312–324.

Lin YL, Chang PC, Hsu C, Hung MZ, Chien YH, Hwu WL, Lai FP, Lee NC. 2022. Comparison of GATK and DeepVariant by trio sequencing. *Sci Rep* 12.

Lloret-Villas A, Bhati M, Kadri NK, Fries R, Pausch H. 2021. Investigating the impact of reference assembly choice on genomic analyses in a cattle breed. *BMC Genomics* 22.

Lloret-Villas A, Pausch H, Leonard AS. 2023. The size and composition of haplotype reference panels impact the accuracy of imputation from low-pass sequencing in cattle. *Genetics Selection Evolution* 55.

Low WY, Tearle R, Liu R, Koren S, Rhie A, Bickhart DM, Rosen BD, Kronenberg ZN, Kingan SB, Tseng E, et al. 2020. Haplotype-resolved genomes provide insights into structural variation and gene content in Angus and Brahman cattle. *Nat Commun* 11.

Luo R, Wong C-L, Wong Y-S, Tang C-I, Liu C-M, Leung C-M, Lam T-W. 2020. Exploring the limit of using a deep neural network on pileup data for germline variant calling. *Nat Mach Intell* 2: 220–227. https://www.nature.com/articles/s42256-020-0167-4.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. 2010. The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 20: 1297–1303.

Md V, Misra S, Li H, Aluru S. 2019. Efficient architecture-aware acceleration of BWA-MEM for multicore systems. In *Proceedings - 2019 IEEE 33rd International Parallel and Distributed Processing Symposium, IPDPS 2019*, pp. 314–324, Institute of Electrical and Electronics Engineers Inc.

Musolf AM, Holzinger ER, Malley JD, Bailey-Wilson JE. 2022. What makes a good prediction? Feature importance and beginning to open the black box of machine learning in genetics. *Hum Genet* 141: 1515–1528.

Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze A V, Mikheenko A, Vollger MR, Altemose N, Uralsky L, Gershman A, et al. 2022. *The complete sequence of a human genome*. https://www.science.org.

Olson ND, Wagner J, Dwarshuis N, Miga KH, Sedlazeck FJ, Salit M, Zook JM. 2023. Variant calling and benchmarking in an era of complete human genome sequences. *Nat Rev Genet* 24: 464–483.

Olson ND, Wagner J, McDaniel J, Stephens SH, Westreich ST, Prasanna AG, Johanson E, Boja E, Maier EJ, Serang O, et al. 2022a. PrecisionFDA Truth Challenge V2: Calling variants

Downloaded from genome.cshlp.org on July 16, 2026 . Published by Cold Spring Harbor Laboratory Press

from short and long reads in difficult-to-map regions. *Cell Genomics* 2: 100129. https://linkinghub.elsevier.com/retrieve/pii/S2666979X22000581.

Olson ND, Wagner J, McDaniel J, Stephens SH, Westreich ST, Prasanna AG, Johanson E, Boja E, Maier EJ, Serang O, et al. 2022b. PrecisionFDA Truth Challenge V2: Calling variants from short and long reads in difficult-to-map regions. *Cell Genomics* 2: 100129. https://www.sciencedirect.com/science/article/pii/S2666979X22000581.

Oppenheimer J, Rosen BD, Heaton MP, Vander Ley BL, Shafer WR, Schuetze FT, Stroud B, Kuehn LA, Mcclure JC, Barfield JP, et al. 2021. A Reference Genome Assembly of American Bison, Bison bison bison. *Journal of Heredity* 112: 174–183.

Osipowski P, Pawełkowicz M, Wojcieszek M, Skarzyńska A, Przybecki Z, Pląder W. 2020. A high-quality cucumber genome assembly enhances computational comparative genomics. *Molecular Genetics and Genomics* 295: 177–193.

Phillippy A, Hansen N. 2023. The Q100 Project. https://genomeinformatics.github.io/HG002v1/ (Accessed October 9, 2024).

Popic V, Rohlicek C, Cunial F, Hajirasouliha I, Meleshko D, Garimella K, Maheshwari A. 2023. Cue: a deep-learning framework for structural variant discovery and genotyping. *Nat Methods* 20: 559–568.

Poplin R, Chang PC, Alexander D, Schwartz S, Colthurst T, Ku A, Newburger D, Dijamco J, Nguyen N, Afshar PT, et al. 2018a. A universal snp and small-indel variant caller using deep neural networks. *Nat Biotechnol* 36: 983. https://www.nature.com/articles/s41587-021-00861-3.

Poplin R, Ruano-Rubio V, Depristo MA, Fennell TJ, Carneiro MO, Van Der Auwera GA, Kling DE, Gauthier LD, Levy-Moonshine A, Roazen D, et al. 2018b. Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv* 1–22. https://doi.org/10.1101/201178.

Porubsky D, Dashnow H, Sasani TA, Logsdon GA, Hallast P, Noyes MD, Kronenberg ZN, Mokveld T, Koundinya N, Nolan C, et al. 2025. Human de novo mutation rates from a four-generation pedigree reference. *Nature*. https://www.nature.com/articles/s41586-025-08922-2.

Ramachandran A, Lumetta SS, Klee EW, Chen D. 2021. HELLO: improved neural network architectures and methodologies for small variant calling. *BMC Bioinformatics* 22: 1–31.

Rautiainen M, Nurk S, Walenz BP, Logsdon GA, Porubsky D, Rhie A, Eichler EE, Phillippy AM, Koren S. 2023. Telomere-to-telomere assembly of diploid chromosomes with Verkko. *Nat Biotechnol* 41: 1474–1482.

Rice ES, Koren S, Rhie A, Heaton MP, Kalbfleisch TS, Hardy T, Hackett PH, Bickhart DM, Rosen BD, Ley B Vander, et al. 2020. Continuous chromosome-scale haplotypes assembled from a single interspecies F1 hybrid of yak and cattle. *Gigascience* 9.

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. *Nat Biotechnol* 29: 24–26.

Rosen BD, Bickhart DM, Schnabel RD, Koren S, Elsik CG, Tseng E, Rowan TN, Low WY, Zimin A, Couldrey C, et al. 2020. De novo assembly of the cattle reference genome with single-molecule sequencing. *Gigascience* 9.

Ruperao P, Gandham P, Odeny DA, Mayes S, Selvanayagam S, Thirunavukkarasu N, Das RR, Srikanda M, Gandhi H, Habyarimana E, et al. 2023. Exploring the sorghum race level diversity utilizing 272 sorghum accessions genomic resources. *Front Plant Sci* 14.

Downloaded from genome.cshlp.org on July 16, 2026 . Published by Cold Spring Harbor Laboratory Press

Sahraeian SME, Liu R, Lau B, Podesta K, Mohiyuddin M, Lam HYK. 2019. Deep convolutional neural networks for accurate somatic mutation detection. *Nat Commun* 10.

Sapoval N, Aghazadeh A, Nute MG, Antunes DA, Balaji A, Baraniuk R, Barberan CJ, Dannenfelser R, Dun C, Edrisi M, et al. 2022. Current progress and open challenges for applying deep learning across the biosciences. *Nat Commun* 13.

Secomandi S, Gallo GR, Sozzoni M, Iannucci A, Galati E, Abueg L, Balacco J, Caprioli M, Chow W, Ciofi C, et al. 2023. A chromosome-level reference genome and pangenome for barn swallow population genomics. *Cell Rep* 42.

Shafin K, Pesout T, Chang PC, Nattestad M, Kolesnikov A, Goel S, Baid G, Kolmogorov M, Eizenga JM, Miga KH, et al. 2021. Haplotype-aware variant calling with PEPPER-Margin-DeepVariant enables high accuracy in nanopore long-reads. *Nat Methods* 18: 1322–1332.

Smith TPL, Bickhart DM, Boichard D, Chamberlain AJ, Djikeng A, Jiang Y, Low WY, Pausch H, Demyda-Peyrás S, Prendergast J, et al. 2023. The Bovine Pangenome Consortium: democratizing production and accessibility of genome assemblies for global cattle breeds and other bovine species. *Genome Biol* 24.

Stenløkk K, Saitou M, Rud-Johansen L, Nome T, Moser M, Árnyasi M, Kent M, Barson NJ, Lien S. 2022. The emergence of supergenes from inversions in Atlantic salmon. *Philosophical Transactions of the Royal Society B: Biological Sciences* 377.

Stephens ZD, Hudson ME, Mainzer LS, Taschuk M, Weber MR, Iyer RK. 2016. Simulating next-generation sequencing datasets from empirical mutation and sequencing models. *PLoS One* 11.

Su J, Zheng Z, Ahmed SS, Lam T-W, Luo R. 2022. Clair3-trio: high-performance Nanopore long-read variant calling in family trios with trio-to-trio deep neural networks. *Brief Bioinform* 23: 1–31. https://academic.oup.com/bib/article/doi/10.1093/bib/bbac301/6645484.

Sundaram L, Gao H, Padigepati SR, McRae JF, Li Y, Kosmicki JA, Fritzilas N, Hakenberg J, Dutta A, Shon J, et al. 2018. Predicting the clinical impact of human mutation with deep neural networks. *Nat Genet* 50: 1161–1170.

Supernat A, Vidarsson OV, Steen VM, Stokowy T. 2018. Comparison of three variant callers for human whole genome sequencing. *Sci Rep* 8.

The 1000 Genomes Project Consortium, Yang H, Zhang X, Zheng H, Ambrogio L, Bloom T, Cibulskis K, Fennell TJ, Jaffe DB, Shefler E, et al. 2010. A map of human genome variation from population-scale sequencing. *Nature* 467: 1061–1073. https://www.nature.com/articles/nature09534 (Accessed December 19, 2024).

Wagner J, Olson ND, Harris L, Khan Z, Farek J, Mahmoud M, Stankovic A, Kovacevic V, Yoo B, Miller N, et al. 2022. Benchmarking challenging small variants with linked and long reads. *Cell Genomics* 2: 100128. https://doi.org/10.1016/j.xgen.2022.100128.

Webb S. 2018. Deep learning for biology. *Nature* 554: 555–557.

Whalen S, Schreiber J, Noble WS, Pollard KS. 2021. Navigating the pitfalls of applying machine learning in genomics. *Nat Rev Genet* 23: 169–181. https://doi.org/10.1038/s41576-021-00434-9.

Yun T, McLean C, Chang P-C, Carroll A. 2018. Improved non-human variant calling using species-specific DeepVariant models. *DeepVariant Blog*. https://google.github.io/deepvariant/posts/2018-12-05-improved-non-human-variant-calling-using-species-specific-deepvariant-models/.

Downloaded from genome.cshlp.org on July 16, 2026 . Published by Cold Spring Harbor Laboratory Press

Zhao S, Agafonov O, Azab A, Stokowy T, Hovig E. 2020. Accuracy and efficiency of germline variant calling pipelines for human genome data. *Sci Rep* 10.

Zhou Y, Yang L, Han X, Han J, Hu Y, Li F, Xia H, Peng L, Boschiero C, Rosen BD, et al. 2022a. Assembly of a pangenome for global cattle reveals missing sequences and novel structural variations, providing new insights into their diversity and evolutionary history. *Genome Res* 32: 1585–1601. https://doi.org/10.1101/gr.276550.122.

Zhou Y, Yang L, Han X, Han J, Hu Y, Li F, Xia H, Peng L, Boschiero C, Rosen BD, et al. 2022b. Assembly of a pangenome for global cattle reveals missing sequences and novel structural variations, providing new insights into their diversity and evolutionary history. *Genome Res* 32: 1585–1601.

Zook JM, Chapman B, Wang J, Mittelman D, Hofmann O, Hide W, Salit M. 2014. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat Biotechnol* 32: 246–251.

Zook JM, McDaniel J, Olson ND, Wagner J, Parikh H, Heaton H, Irvine SA, Trigg L, Truty R, McLean CY, et al. 2019. An open resource for accurately benchmarking small variant and reference calls. *Nat Biotechnol* 37: 561–566.

Zou J, Huss M, Abid A, Mohammadi P, Torkamani A, Telenti A. 2019. A primer on deep learning in genomics. *Nat Genet* 51: 12–18. http://dx.doi.org/10.1038/s41588-018-0295-5.

Downloaded from genome.cshlp.org on July 16, 2026 . Published by Cold Spring Harbor Laboratory Press

Table 1. Bovine genomes used during successive TrioTrain iterations. Individual genomes have a distinct 'truth' VCF. Extending DeepVariant with bovine genomes was achieved over five phases with 30 TrioTrain iterations, representing 14 family trios and one synthetic hybrid offspring. Additional details are provided in **Table S4.**

		Training: Parents				***Evaluation: Offspring***			
	Trio Name	*SRA Accession*	*Sex*	*Breed Code*	*Mean Cov.*	*SRA Accession*	*Sex*	*Breed Code*	*Mean Cov.*
Training Phase 1									
1	Trio1	SAMN10940502	M	AA	35.96	SAMN10940538	F	AA	23.64
2		SAMN10940537	F	AA	23.10				
3	Trio2	SAMN10598569	M	HE	17.47	SAMN10598563	F	HE	14.45
4		SAMN10598566	F	HE	12.92				
5	Trio3	SAMEA4644756	M	BS	15.84	SAMEA4644754	F	BS	16.78
6		SAMEA4644755	F	BS	15.14				
7	Trio4	SAMN15780082	M	HO	35.23	SAMN15779741	M	HO	23.66
8		SAMN15779919	F	HO	46.68				
9	Trio5	SAMN10598570	M	HE	13.51	SAMN10598562	F	HE	15.89
10		SAMN10598567	F	HE	15.08				
11	Trio6	SAMEA5159887	M	TG	21.63	SAMEA5159888	F	TG	20.41
12		SAMEA5159889	F	TG	19.77				
Training Phase 2									
13	Trio7	SAMN15779644	M	HJ	39.35	SAMN15779593	F	HJ	50.83
14		SAMN15779569	F	HJ	44.75				
15	Trio8	SAMN15779606	M	HJ	30.81	SAMN15779717	M	HJ	27.65
16		SAMN15779596	F	HJ	39.31				
17	Trio9	SAMN10598569	M	HE	17.47	SAMN10598564	M	HE	14.79
18		SAMN10598565	F	HE	14.62				
Training Phase 3									
19	Trio10	SAMN05788493	M	BI	22.69	SAMN13655887	F	BI	17.95
20		SAMN13655886	F	BI	23.72				
21	Trio11	SAMN05788493	M	BI	22.69	SAMN10940703	M	BI	24.06
22		SAMN10940702	F	BI	23.78				
Training Phase 4									
23	Trio12	SAMN08473804	M	AA	36.99	SAMN08473802	M	F1X	50.01
24		SAMN08473803	F	BR	43.32				
25	Trio13	SAMN12153485	M	HI	34.69	SAMN12153487	F	F1X	17.92
26		SAMN12153486	F	YK	10.66				
27	Trio14	SAMN16823422	M	BI	27.44	SAMN16780309	M	F1X	14.46
28		SAMN16825967	F	SI	35.42				
Training Phase 5									
29	Trio15	SAMN08473804	M	BI	36.99	SYNTHETIC	M	F1X	26.20
30		SAMN08473803	F	SI	43.32				

Breed Code Abbreviations: AA = Angus, HE = Hereford, BS = Brown Swiss, HO = Holstein, TG = Tyrolean Grey, HJ = Holstein-Jersey, BI = Bison, BR = Brahman, HI = Scottish Highlander, YK = Yak, SI = Simmental, F1X = F1 Hybrid Cross

Downloaded from genome.cshlp.org on July 16, 2026 . Published by Cold Spring Harbor Laboratory Press

FIGURE LEGENDS

Figure 1. Before extending DeepVariant, the available data was assigned to different purposes: training (**A**) or testing (**B**). The x-axis provides thresholds for genome-wide mean coverage, while the y-axis provides sample counts. **A.** We selected samples from the UMAGv1 cohort with complete pedigree information (family trios). With TrioTrain, DV simultaneously uses labeled examples from two genomes: a parental genome provides new information, while learning is evaluated based on genotype predictions in the offspring. **B.** Most samples used for model testing (13 / 19) lacked a complete pedigree but served as an independent test set because their GATK-based truth labels were withheld from DeepVariant during training. **C**. The Kernel Density Estimate (KDE) plot allows for comparing the two distributions after normalizing for sample size. **D.** Workflow diagram of TrioTrain's iterative re-training process. All re-training begins with weights from an existing checkpoint (warm-starting); our experiments began with DV (v1.4) without the allele frequency channel. After initialization, all bovine-trained checkpoints include the Allele Frequency channel (DV-AF), encoding the PopVCF from the UMAGv1 cohort. Two iterations are performed for each trio, one for each parent. An iteration covers all labeled examples within a bovine parent-child duo constrained to sample-specific callable regions from Chr 1-29, X. Each iteration produces a candidate checkpoint that is then assessed using a set of test genomes (N=19). The resulting variant calls are compared against sample-specific, GATK-derived, truth labels to monitor training behavior over time.

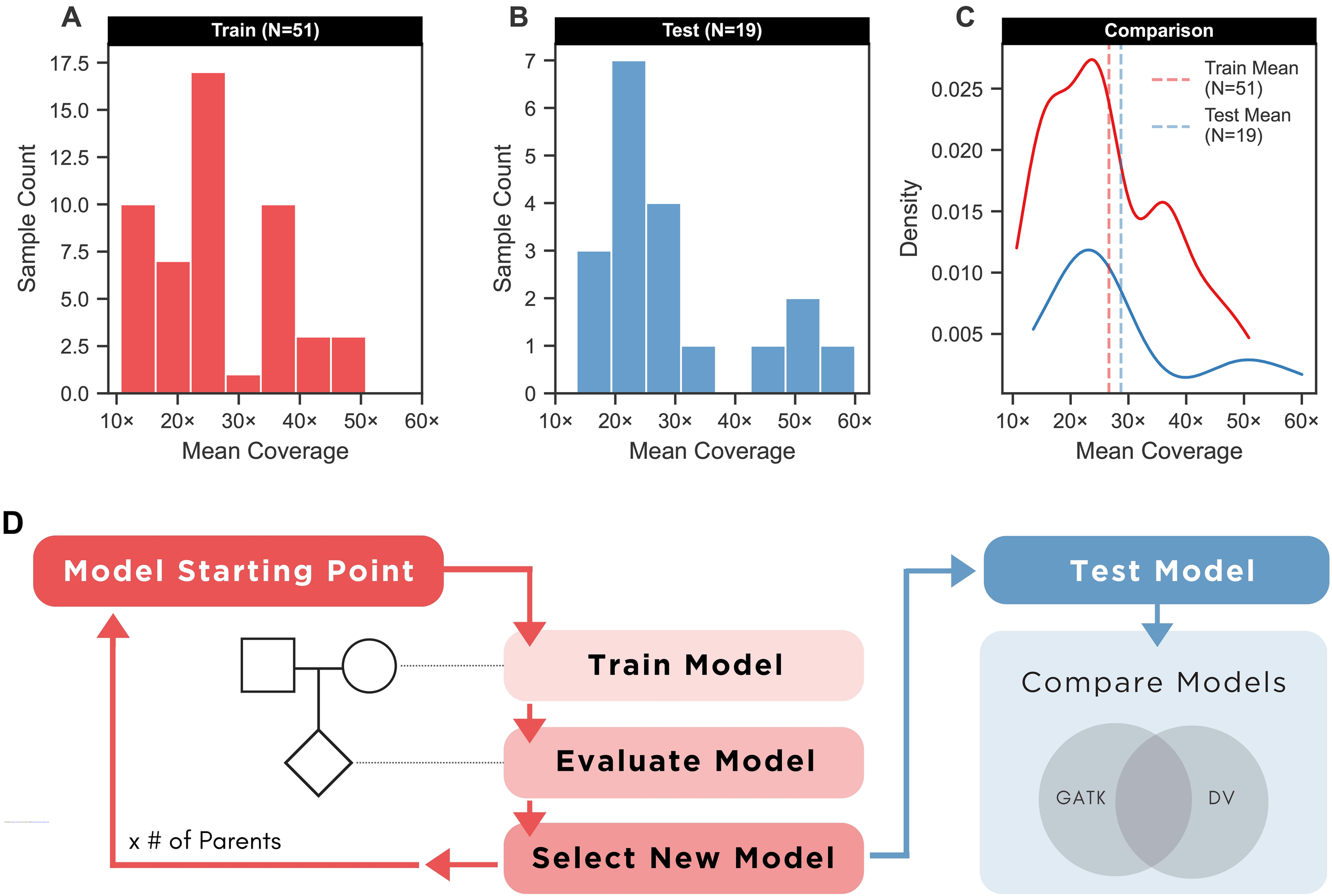
Figure 2. Training performance across successive iterations with bovine trios reveals insights into DeepVariant's behavior. The darker vertical lines highlight the checkpoints (2, 12, 18, 22, 28, 30) used in subsequent analyses to compare across phases. TrioTrain enables structuring the bovine inputs to match inheritance expectations – variation is transmitted from parents to offspring. Pedigree data is not explicitly given to DeepVariant; instead, parental genotypes are used to inform predictions in the offspring. The number of checkpoints (x-axis) represents the number of parental genomes given to DV (odd = paternal; even = maternal). Two iterations are required to cover both parents. For example, iterations 1 and 2 represent a complete bovine trio where truth labels from a single offspring were used to evaluate the optimal training stopping point (**Table 1**). In the first two panels, the y-axis represents the maximum F1 score achieved in the offspring; comparing across iterations reveals if DV found the parental truth labels informative. Each line represents stratified performance by different variant classifications. **A. Variant Type (SNVs, indels):** As expected, training struggles with indels due to imbalanced classification, as the bovine truth labels contain more SNV relative to those from humans. **B. Genotype Class (HomAlts, Hets).** Between iterations 18 and 19, we observe a shift in the best-performing genotype class for the offspring due to distributional changes to heterozygosity. **C. Truth label contents contribute to training performance.** Phase 3 uses truth labels from Bison genomes aligned to the cattle reference; a sudden increase in HomAlts (**Figure S3)** is reflected by the decreased Het:HomAlt ratio. Counterintuitively, increasing the number of HomAlt examples given to DeepVariant contributes to a decrease in F1 score for that genotype class during training.

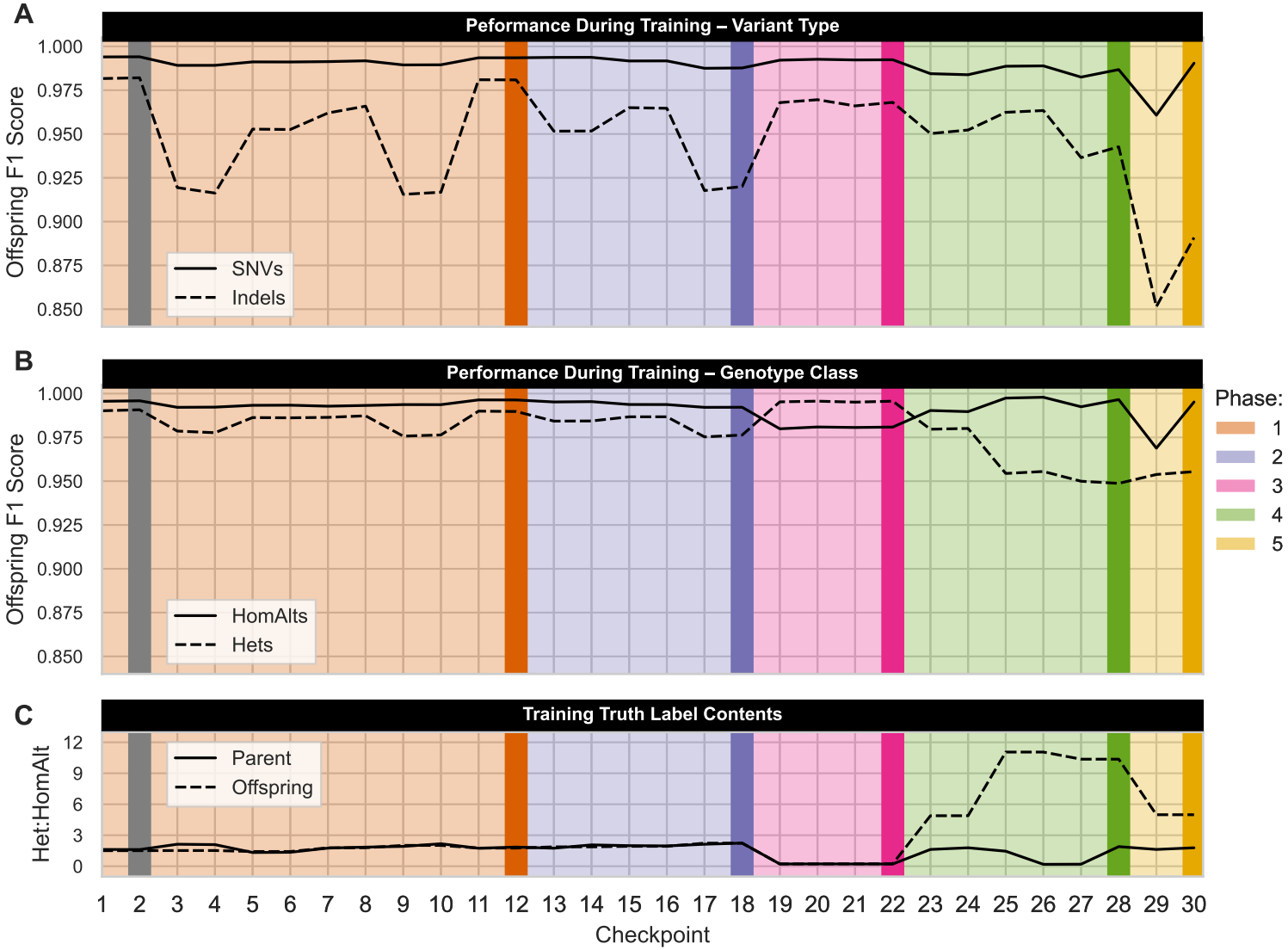
Figure 3. Comparing variant type generalization across species. Three human-trained versions of DeepVariant (DV, DV-AF, DT) are contrasted against eight bovine-trained DV-AF checkpoints created with TrioTrain. Each box-and-whisker represents the distribution of F1 scores observed in bovine testing genomes (top; N=19, except for DT, where N=3 offspring) or the human GIAB trios (bottom; N=6). As expected, exclusively training with the human GIAB samples achieves a higher F1 score with these samples. Although bovine-trained checkpoints ignore genuine indels in humans, the high-quality SNV variants within the UMAGv1 callset enable SNV generalization across species. However, bovine genomes (top) results quantify genotyping changes relative to the GATK-based truth (UMAGv1). See Figure S15 for results stratified by genotype class.

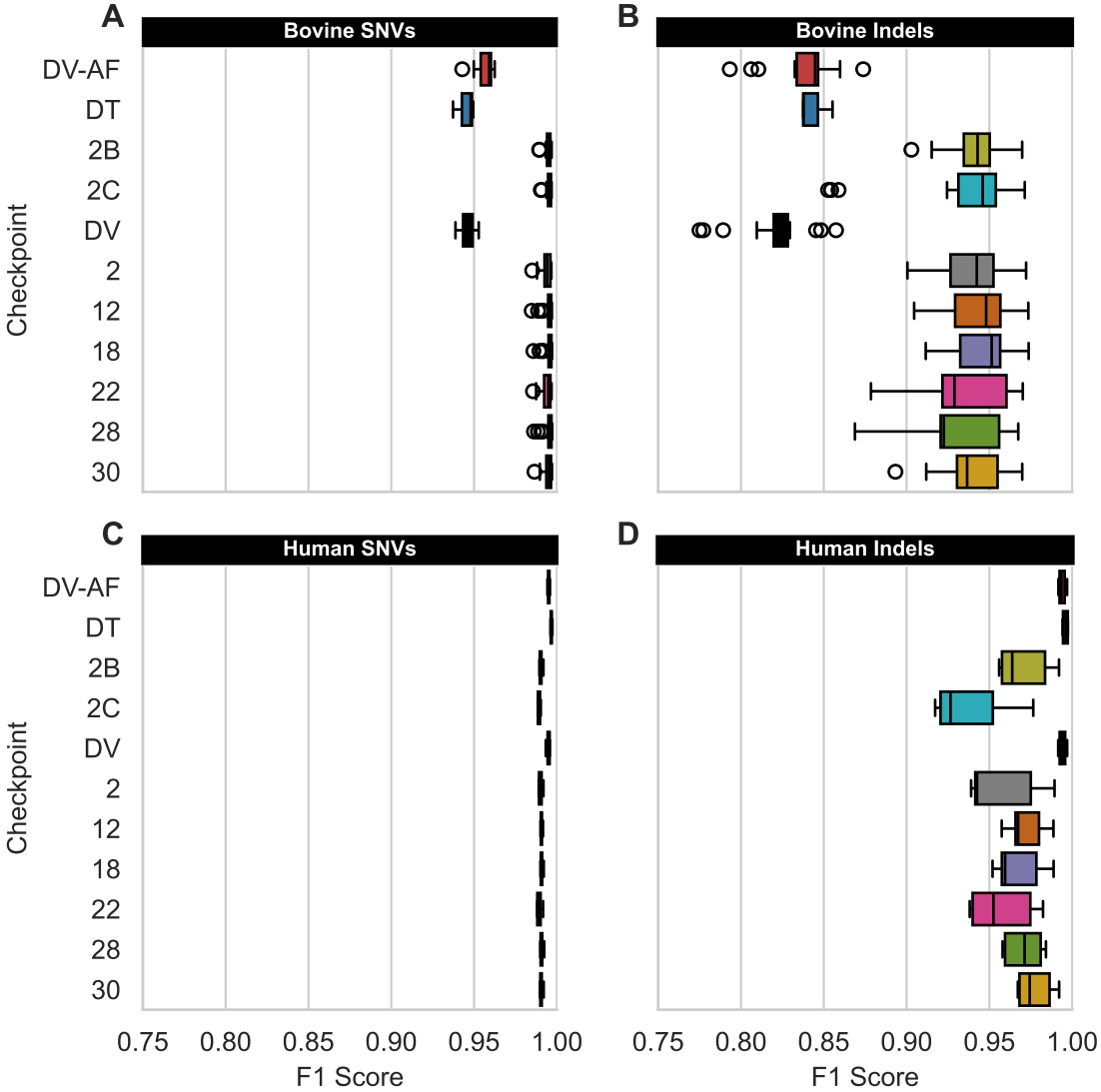
Downloaded from genome.cshlp.org on July 16, 2026 . Published by Cold Spring Harbor Laboratory Press

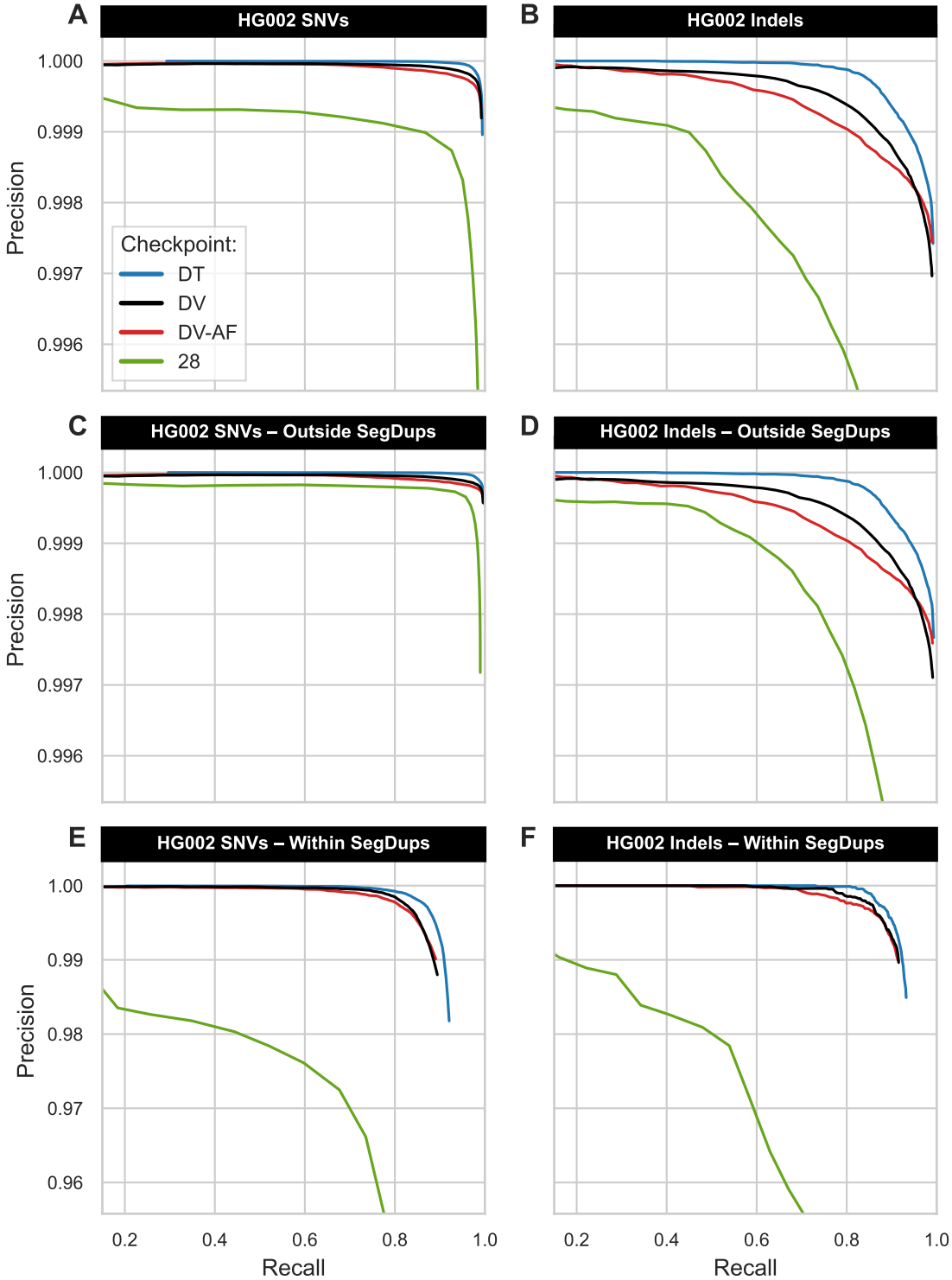
Figure 4. Benchmarking checkpoints with HG002. We used hap.py to calculate the Precision-Recall curve in a single human genome (HG002), where values above 0.5 indicate skilled prediction. We contrast three human-trained versions of DeepVariant (DV, DV-AF, DT) against a bovine-trained DV-AF checkpoint (28) that generalized well in humans and cattle. The top panels stratify genome-wide classification accuracy in SNV (**A**) or indels (**B**), controlling for the variable number of genotypes to allow direct model comparison. The middle panels use the GIAB stratifications (v3.5) to compare classification accuracy outside known segmental duplication (SegDup) regions in SNV (**C**) and indels (**D**). Note that the y-axis has a different scale for the bottom panels for stratification within SegDups in SNV (**E**) and indels (**F**). Models exclusively trained with human genomes outperform the bovine-trained model in HG002. However, the TrioTrain checkpoint's lower performance in repetitive regions is expected since bovine SegDups are not characterized to the same extent as in humans. Outside of known SegDups in HG002, the bovine-trained checkpoint created with TrioTrain is nearly perfect in SNVs (**C**). We infer that training with bovine genomes alters DeepVariant's priors for heterozygosity and copy number variation; these adjustments contribute to the marginally lower curve observed in human genome-wide precision and recall (top).

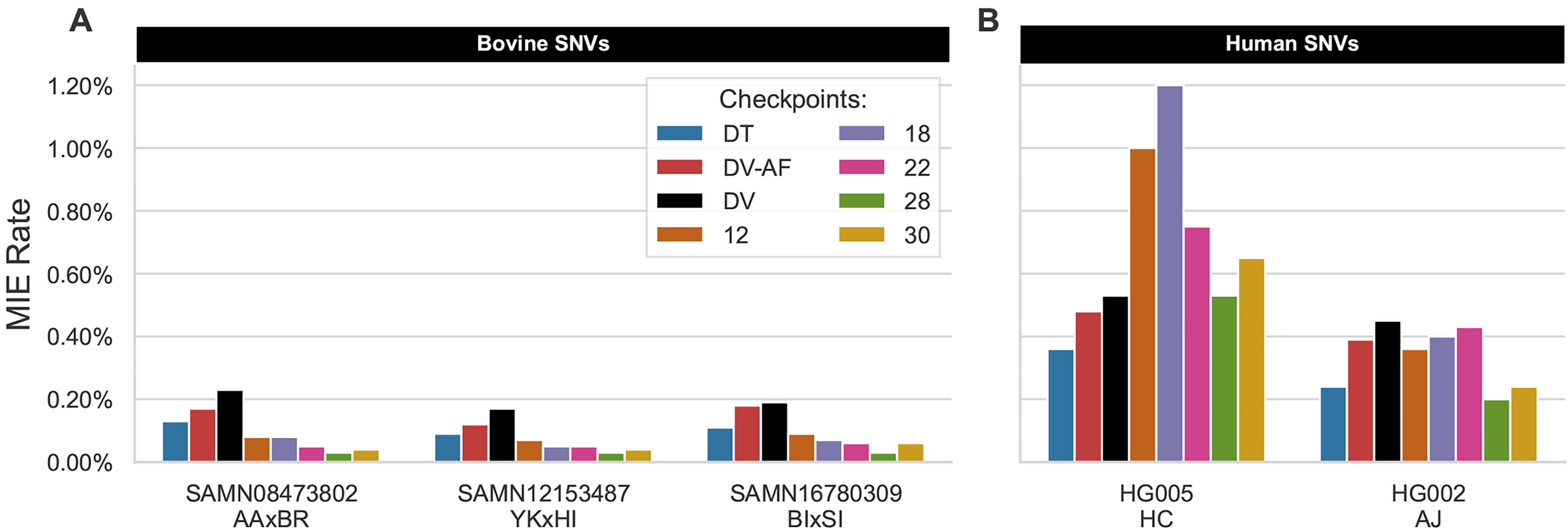
Figure 5. Assessing inheritance error rate in human and bovine trios. Mendelian Inheritance Errors (MIE) were identified in PASS variants from the autosomes and X chromosome for three bovine hybrid-cross trios (**A**) and two GIAB human trios (**B**). Each cluster of bars on the x-axis represents a distinct trio, with a bar representing the total number of discordant SNVs proportional to the total number of variants analyzed for eight distinct checkpoints. The Han Chinese (HC) trio deviates from all other trios, with the MIE rate nearly doubling for the first two TrioTrain phases. However, relative to the default checkpoint (DV), one bovine-AF checkpoint (28) achieves an identical SNV MIE rate in HG005 (0.53%) while also producing fewer discordant SNVs in the Ashkenazi (AJ) trio. Since checkpoints differ in total variants called in the sample individual, the line plot (**C**) enables a more direct comparison for HG002 only. The y-axis reports the cumulative number of errors found across four distinct versions of DeepVariant (DV, DV-AF, DT, C28). PASS variants in the trio VCF were sorted by Genotype Quality (GQ) in the offspring, where GQ score decreases as the number of variants increases from left to right on the x-axis. Compared to the human-trained single-sample variant callers, the error curve with checkpoint 28 approaches DeepTrio, confirming that our approach successfully recognizes inherited variation. While DT is a joint caller requiring a minimum of two samples with known pedigree, the bovine-trained checkpoint built with TrioTrain reduces the MIE rate in individual samples, enabling speed improvements.

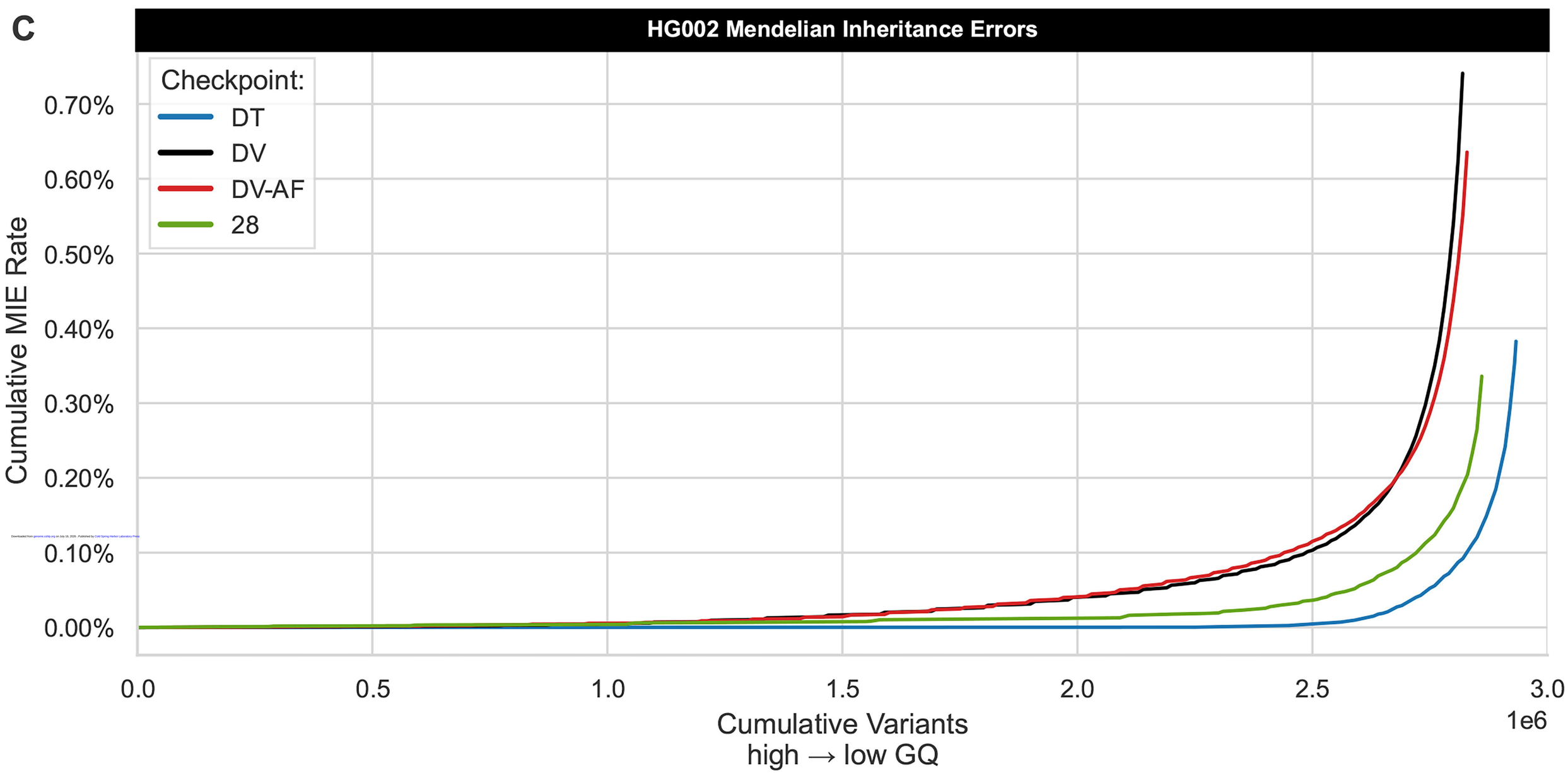

A
Train (N=51)
Sample Count
Mean Coverage
B
Test (N=19)
Sample Count
Mean Coverage
C
Comparison
Density
Train Mean (N=51)
Test Mean (N=19)
Mean Coverage
D
Model Starting Point
Train Model
Evaluate Model
Select New Model
x # of Parents
Test Model
Compare Models
GATK
DV


A
Peformance During Training – Variant Type
Offspring F1 Score
1.000
0.975
0.950
0.925
0.900
0.875
0.850
SNVs
Indels
B
Performance During Training – Genotype Class
Offspring F1 Score
1.000
0.975
0.950
0.925
0.900
0.875
0.850
HomAlts
Hets
Phase:
1
2
3
4
5
C
Training Truth Label Contents
Het:HomAlt
12
9
6
3
0
Parent
Offspring
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
Checkpoint


A
Bovine SNVs
B
Bovine Indels
C
Human SNVs
D
Human Indels
Checkpoint
DV-AF
DT
2B
2C
DV
2
12
18
22
28
30
0.75
0.80
0.85
0.90
0.95
1.00
F1 Score


A
HG002 SNVs
B
HG002 Indels
C
HG002 SNVs – Outside SegDups
D
HG002 Indels – Outside SegDups
E
HG002 SNVs – Within SegDups
F
HG002 Indels – Within SegDups
Precision
1.000
0.999
0.998
0.997
0.996
1.00
0.99
0.98
0.97
0.96
Checkpoint:
DT
DV
DV-AF
28
0.2
0.4
0.6
0.8
1.0
Recall





Downloaded from genome.cshlp.org on July 16, 2026 - Published by Cold Spring Harbor Laboratory Press