nature communications

Article

https://doi.org/10.1038/s41467-025-57756-z

SVLearn: a dual-reference machine learning approach enables accurate cross-species genotyping of structural variants

Received: 3 September 2024

Accepted: 4 March 2025

Published online: 11 March 2025

Check for updates

Qimeng Yang[1,7], Jianfeng Sun[2,7], Xinyu Wang[1], Jiong Wang[1], Quanzhong Liu[3], Jinlong Ru[4], Xin Zhang[3], Sizhe Wang[3], Ran Hao[3], Peipei Bian[1], Xuelei Dai[1,5], Mian Gong[1,6], Zhuangbiao Zhang[1], Ao Wang[1], Fengting Bai[1], Ran Li[1], Yudong Cai[1] ✉ & Yu Jiang[1] ✉

Structural variations (SVs) are diverse forms of genetic alterations and drive a wide range of human diseases. Accurately genotyping SVs, particularly occurring at repetitive genomic regions, from short-read sequencing data remains challenging. Here, we introduce SVLearn, a machine-learning approach for genotyping bi-allelic SVs. It exploits a dual-reference strategy to engineer a curated set of genomic, alignment, and genotyping features based on a reference genome in concert with an allele-based alternative genome. Using 38,613 human-derived SVs, we show that SVLearn significantly outperforms four state-of-the-art tools, with precision improvements of up to 15.61% for insertions and 13.75% for deletions in repetitive regions. On two additional sets of 121,435 cattle SVs and 113,042 sheep SVs, SVLearn demonstrates a strong generalizability to cross-species genotype SVs with a weighted genotype concordance score of up to 90%. Notably, SVLearn enables accurate genotyping of SVs at low sequencing coverage, which is comparable to the accuracy at 30× coverage. Our studies suggest that SVLearn can accelerate the understanding of associations between the genome-scale, high-quality genotyped SVs and diseases across multiple species.

Structural variations (SVs) are ubiquitously present in genomes and have been associated with various biological traits[1–3] and human diseases[4–6]. In recent years, the advent of long-read sequencing technologies has significantly enhanced the ability to detect SVs[7–9]. Given the high cost and scarcity of long-read sequencing data, SVs in larger populations, especially at the sequence-resolved level, have still primarily been genotyped from short-read sequencing data[10–14]. Yet, there is a long-standing challenge in correctly calling SVs from short reads for several reasons, such as their insufficient coverage of genomic regions and the limited information to resolve complex rearrangements[15,16]. To address this, several studies have experimented with both short-read and long-read sequencing technologies for accurately calling SVs[17–19]. In these cases, long reads are first used to generate accurate SV sets or graph pangenomes, while a large cohort of short reads are then applied to genotype known variant sets. As such, long-read-derived SV sets, which are able to be genotyped

[1]Key Laboratory of Animal Genetics, Breeding and Reproduction of Shaanxi Province, College of Animal Science and Technology, Northwest A&F University, Yangling Shaanxi, China. [2]Botnar Research Centre, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK. [3]College of Information Engineering, Northwest A&F University, Yangling Shaanxi, China. [4]Institute of Virology, Helmholtz Centre Munich - German Research Centre for Environmental Health, Neuherberg, Germany. [5]Yazhouwan National Laboratory, Sanya Hainan, China. [6]State Key Laboratory of Animal Biotech Breeding, Institute of Animal Science, Chinese Academy of Agricultural Sciences (CAAS), Beijing, China. [7]These authors contributed equally: Qimeng Yang, Jianfeng Sun. ✉e-mail: yudong.cai@nwafu.edu.cn; yu.jiang@nwafu.edu.cn

directly from short reads, have been accumulated rapidly across various species[20–22]. These resulting genotyping data can be seamlessly integrated into the downstream analysis workflows of genome-wide association study (GWAS)[23], selection[2,22], and introgression[18,21].

Computational strategies for identifying genotypes of SVs begin by constructing a graph to which short reads are aligned using genomic sequences of both reference and alternative alleles[24,25]. These graph-based methods have been shown effective in elevating genotyping accuracy in that biases towards the reference allele linked to a single linear genome are vastly reduced[26]. With a remarkable increase in the availability of computational algorithms and the volume of SV data, several established approaches, such as Paragraph[27] and GraphTyper2[28], have opted to omit the elimination of reference allele biases, but instead, directly project reads (aligned to SV loci in a linear reference genome) onto localized SV graphs at the cost of computational efficiency. However, the lack of information about those unmapped short reads also makes genotyping some certain types of SVs challenging[25]. Some recently released methods, such as Giraffe[26] and PanGenie[29], gain better genotyping accuracy at a noticeably increasing computational cost by mapping raw sequenced reads to graph-based reference pangenome or using k-mers that rely on haplotype-resolved graph pangenome for known variant genotyping. Nevertheless, the graph-based strategies compromise between genotyping accuracies and computational costs. Yet another computational effort, called Reference flow[30], has spearheaded read mapping in a graph-free fashion, which improves both the quantity and quality of mapped reads by leveraging a concatenated reference genome from a hybrid of populations. The alignment accuracy and bias avoidance bear a close resemblance to those by graph aligners yet are achieved at a much faster speed and a much lower memory cost.

Here, we present SVLearn, a machine-learning-based genotyper, for accurately genotyping SVs from a large population of samples sequenced with short reads. It leverages a dual-reference strategy to boost the abundance of reads at SV loci by building a reference genome and an alternative genome containing alternative allele sequences of known biallelic insertions/deletions. The genotyping results have proven the advantages of using these aligned reads and the features extracted from them. A variety of SVLearn models are optimized with stratified k-fold cross-validation, hyper-parameter fine-tuning, and the leave-one-out strategy based on genome, alignment, and genotyping features. Our genome-relevant features are extracted based on 2 × 150 bp short reads while genotypes are sourced from PacBio HiFi long reads. Compared with other existing tools, SVLearn demonstrates a remarkable improvement in genotyping both our internally crafted SVs and externally expert-curated SVs (Genome-in-a-Bottle consortium[31]) from human individuals. Additionally, this tool demonstrated the ideal generalizability of genotyping SVs from two livestock animals, cattle and sheep.

Results

Overview of SVLearn

SVLearn is a machine learning-based tool for genotyping SVs derived from short reads. Using the reference (REF) genome as the backbone, an alternative (ALT) genome is generated by replacing the reference allele sequence with the alternative allele sequence at each known biallelic SV locus in the VCF file (Fig. 1 and Supplementary Fig. 1). Different from the REF genome, the ALT genome contains insertions sequences but omits deletions sequences. The dual-reference genome is built for optimizing alignment outcomes and extracting 10 SV features (Supplementary Table 4). Short reads are separately mapped to the REF and ALT genomes. The resulting REF BAM and ALT BAM files are utilized to extract 8 alignment features about breakpoint coverage and read depths at each SV locus and 6 genotyping-relevant features by running the Paragraph tool. The SV genotypes are derived from PacBio HiFi reads of 15 humans, 15 cattle, and 15 sheep individuals from which training and validation individuals are split according to a ratio of 14: 1. SVLearn models were trained with 18 features (excluding Paragraph features) and 24 features (including Paragraph features), respectively. Detailed information about database mugging, feature extraction, and modelling training, can be found in section Methods. Moreover, SVLearn takes advantage of tandem repeat features and multi-coverage reads to amplify the SV genotyping accuracy and works as a powerful, multifaceted tool in a broad range of scenarios.

Profiles of SV genotypes

We collected 15 human individuals from the Human Pangenome Reference Consortium (HPRC), each with over 30× coverage of PacBio HiFi reads (31.19–41.70×) and Illumina short reads (29.97-35.06×, Supplementary Table 1). We aligned the PacBio HiFi reads of 15 individuals to the reference genome GRCh38, and then called SVs across all autosomes and sex chromosomes (Box 1). After two filtering steps, we retained 38,613 bi-allelic variations (17,007 deletions and 21,606 insertions) as Human SV Set (Fig. 2a). Genotypes in HG002 served as the validation set (Val-label), while genotypes in the remaining 14 individuals served as the training labels (Train-label). The homozygous reference (0/0), heterozygous (0/1), and homozygous alternate (1/1) genotypes were distributed over the 15 individuals in an average ratio of approximately 0.583:0.234:0.183, with similar distributions for deletions and insertions (Fig. 2b).

The ALT genome of the Human SV Set was generated using the REF genome GRCh38. In the REF and ALT genomes, we annotated the repeat class of SVs and categorized them into ten types (Fig. 2c, Supplementary Table 5). We found that the variable number of tandem repeats (VNTR) was most abundantly observed among all SVs (~37.81%). In addition, SVs were enriched for those classified as the short interspersed nuclear element (SINE) and the long interspersed nuclear element (LINE), which fall within ~320 bp and ~6 kb in size, respectively.

To reduce alignment bias, we determined the genomic loci of short reads by combining decisions from mapping them to both REF and ALT genomes. This can also be conducive to improving mapping quality in terms of insertion regions within the ALT genome. Intriguingly, we found that using both genomes for mapping led to approximately a 3-fold increase in the average number of mapped reads in insertion regions across the 15 individuals compared to that if the REF genome is solely used (Fig. 2d). Overall, the average number of reads mapped to SV loci increased by 45.56%. In short, the use of both genomes significantly reinforces the discovery of short reads available for SV genotyping.

Selection of best-performing models and identification of informative features

Random Forest was demonstrated as best-performing among six machine learning algorithms upon completion of the whole training process (Supplementary Fig. 3). Then, we set out to optimize the best-performing model with the leave-one-out strategy (Supplementary Figs. 4-5), suggesting that the performance of SV genotyping is reliable and not significantly influenced by any specific test individual. Through an importance analysis of features, we found that in the final models (Human 18 Feature Model and Human 24 Feature Model), several Alignment features are top-ranked in terms of the importance of correctly genotyping SVs (Fig. 2e, f and Supplementary Fig. 6). Additionally, two Paragraph features, Ref_GT and Alt_GT, which were indicative of the preliminary SV genotyping, contributed most to the performance of Human 24 Feature Model. While the contribution of most of the SV features seems to be limited, the length-associated features (e.g., SV length and TR length) stuck out as useful for genotyping. In addition, we performed a cumulative feature ablation analysis to assess the contribution of our handcrafted SV features to the performance of SV genotyping in tandem repeat (TR) regions. The

https://doi.org/10.1038/s41467-025-57756-z

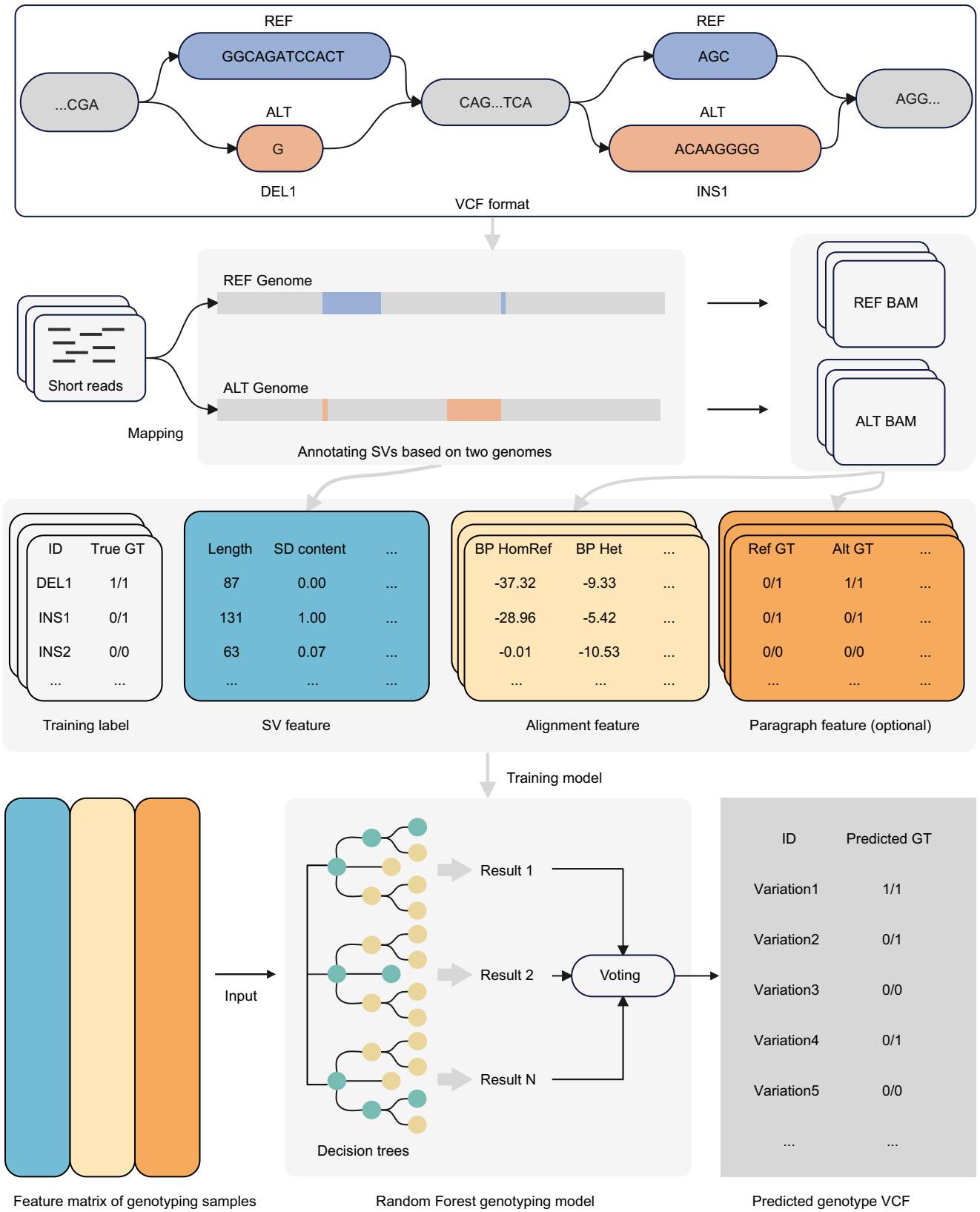


Fig. 1 | Workflow of SVLearn. Based on a known SV set, an alternative (ALT) genome was constructed relative to the reference (REF) genome. Short reads were mapped to REF and ALT genomes to generate REF BAM and ALT BAM files, respectively. SV features were extracted from the two genomes. Alignment features and Paragraph features (optional) were extracted for each SV from the REF BAM and ALT BAM files. The true genotype (GT) of each SV is taken as the label used for training. The model then takes the feature matrix as input and outputs the predicted genotypes of SVs used.

results show an average decrease of 7.55% in weighted genotype concordance (wGC)[29] values within TR regions, compared to an average decrease of 1.77% in non-TR regions (Supplementary Fig. 7), highlighting the critical role of these features in improving the accuracy of SV genotyping in genomic repeat regions.

Evaluation of SV genotyping performance

We compared the genotyping performance of SVLearn with Paragraph[27], BayesTyper[32], GraphTyper2[28], and SVTyper[33] across multiple coverage levels (30×, 20×, 10×, and 5×) using evaluation metrics (Supplementary Fig. 2), including precision, recall, genotype rates, F1 scores, and weighted genotype concordance (wGC). To comprehensively examine the ability of SV genotyping, we derived SVs from a variety of trustworthy resources (Box 1), including PacBio HiFi long reads, haplotype-resolved genomes, as well as datasets from the Genome-in-a-Bottle Consortium (GIAB)[31] and the Human Genome Structural Variation Consortium (HGSVC)[34].

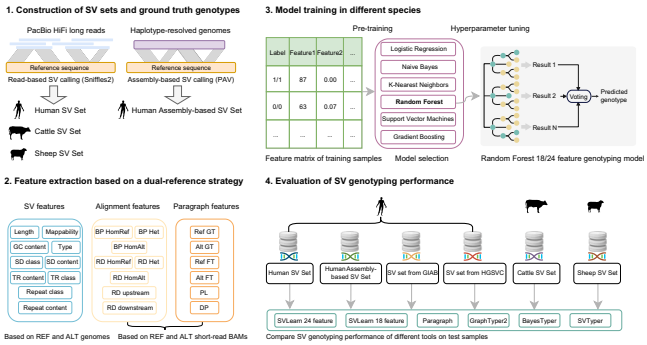
Overall, SVLearn shows a much more pronounced improvement than other tools for genotyping long-read-derived insertions and deletions in the Human SV Set. (Fig. 3a, b, Supplementary Figs. 8 and 14). The best-performing variant model of SVLearn, Human 24 Feature Model, gains wGC of 85.37% at 30× coverage, which is approximately 3% better than the second best model, Human 18 Feature Model. Both models demonstrate superior SV genotyping performance compared to other existing tools (Fig. 3a). For example, they outperform Paragraph by 9.35% and 6.85% in terms of wGC. In addition, SVLearn is superior to other tools in terms of precision, recall, and F1 scores. For instance, the two SVLearn models achieve F1 scores of 0.7922 and 0.7678, whereas all the other tools only have F1 scores of below 0.7 (Fig. 3b). Moreover, we noticed that BayesTyper, one of the representative external tools, while robust in wGC at 30× coverage, performs poorly in light of the genotyping rate (45.27%) compared to the rest of genotyping tools (> 95%). Our results also show that SVLearn leads to varying degrees of improvement with respect to SV lengths ranging from 50 to 5000 bp (Supplementary Fig. 15).

Our results show that, SVLearn gains a slightly better performance in genotyping SVs derived from haplotype-resolved genomes (termed assembly-based) than those obtained by mapping long reads to the GRCh38 reference genome (termed read-based) (Supplementary Fig. 9). The average difference between the two types of SVs is relatively minor, with both F1-scores and wGC values differing by only 1% to 2%. Additionally, Human 24 Feature Model and Human 18 Feature Model, trained using the Human SV Set (read-based), show highly consistent genotyping performance when applied to both assembly-based and read-based SV sets (Fig. 3a, b and Supplementary Fig. 10), suggesting their reliable generalizability.

To further validate the generalization capability of SV genotyping tools, we collected the HG002_SVs_Tier1_v0.6_plus dataset from GIAB, consisting of a total of 16,451 expert-curated SVs, which are independent of the Human SV Set for the HG002 individual. Detailed data processing procedures can be found in section Methods. As seen in Fig. 3c, Human 24 Feature Model of SVLearn achieves the best wGC, reaching 81.09% and 80.33% at 30× and 20× coverage, respectively. Also, Human 18 Feature Model exhibits strong genotyping performance, with wGC of 78.11% and 77.3% at the same coverage levels. However, at 10× and 5× coverage, this model is surpassed by both Human 24 Feature Model and Paragraph by a large margin, as

BOX 1

Overview of the protocol for constructing SVLearn



This box outlines a comprehensive approach for constructing and evaluating SV genotyping models across different species. The workflow is divided into four key steps. **1)** SV sets and ground truth genotypes are generated using PacBio HiFi long reads for 15 individuals per species (human, cattle, sheep). For human, an additional assembly-based SV set and their corresponding ground truth genotypes are generated based on haplotype-resolved genomes. **2)** For each SV set, an alternative (ALT) genome is generated relative to the corresponding reference (REF) genome that is primarily used to characterize the surroundings of each SV in a fine-grained manner. Short reads are then mapped to both REF and ALT genomes, generating BAM files per each. For each SV, ten SV features are extracted directly from the REF and ALT genomes. Eight alignment features and six paragraph features are derived from the aforementioned BAM files. **3)** For each species, SVs derived from 14 individuals are used for training, while those from the remaining one are used for testing. To evaluate the contribution of paragraph features to SV genotyping performance, two feature sets (i.e., including and excluding paragraph features) are formed to train random forest models, respectively, resulting in SVLearn. **4)** The performance between SVLearn and other genotyping tools is evaluated on a held-out test sample for each species. A total of six SV sets were used to comprehensively evaluate the SV genotyping ability of the tools, including four SV sets generated in this study, as well as two external SV sets from the Genome-in-a-Bottle Consortium (GIAB) and the Human Genome Structural Variation Consortium (HGSVC). The human, cattle, and sheep silhouettes were collected from PhyloPic (https://www.phylopic.org) and created by Malio Kodis, T. Michael Keesey, and Mozillian, respectively. The silhouettes are made freely accessible according to the CC0 1.0 Universal Public Domain Dedication.

https://doi.org/10.1038/s41467-025-57756-z

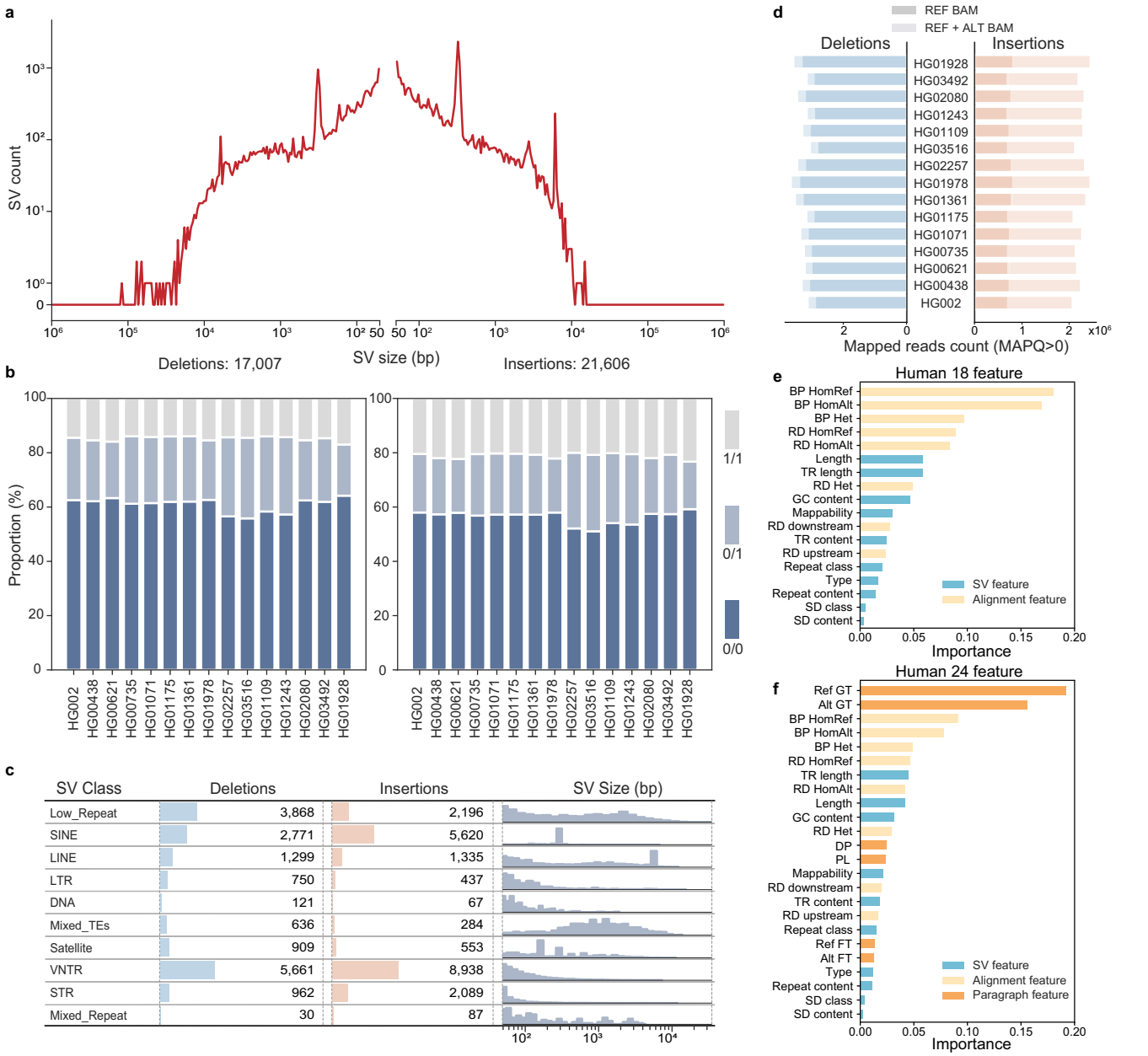


Fig. 2 | Statistics and analysis of Human SV Set and features. a Length distribution of deletions and insertions in Human SV Set. **b** Proportion of three genotypes for deletions and insertions across 15 human individuals. **c** Classification of SVs, with four columns from left to right representing the class of SVs, the count of deletions, the count of insertions, and the length distribution of SVs. **d** Count of mapped reads at SV loci from 15 human individuals using the REF BAM and the combined REF and ALT BAMs. **e** and **f** Feature importance of the Human 18 Feature Model and Human 24 Feature Model, respectively.

evidenced by significantly decreased wGC from 69.87% at 10× to 59.38% at 5×.

Accurately genotyping SVs—which are called from a large number of individuals—in an individual has proved challenging due to the intrinsic nature of the SVs, such as multiple overlapped SVs situated in a genomic region[10]. Our results indicate that genotyping accuracy progressively decreases as the SV set size increases (Supplementary Fig. 16). However, compared to other methods, SVLearn demonstrates a relatively smaller decrease in F1 score and exhibits greater overall stability. For instance, in the 30× data from HG002, as the total number of SVs increases from 45 k to 164 k, the F1-score of SVLearn 24-feature models decreases moderately from 0.6019 to 0.4301. In contrast, Paragraph, which initially performs similarly to SVLearn with an F1-score of 0.5917, experiences a sharp drop to 0.3027. Compared to Paragraph, SVLearn exclusively incorporates multiple SV and alignment features in its training process, potentially enhancing its capability to accurately genotype complex SVs.

Performance improvement in genotyping SVs in tandem repeat regions

We then performed a stratified analysis of genotyping SVs from the Human SV Set within and beyond regions of tandem repeats (TRs) since we observed substantially high proportions of our collected SVs associated with TRs (68.81% of deletions and 84.93% of insertions according to Tandem Repeats Finder, Supplementary Table 6). Our results show that SVLearn significantly improves SV genotyping in TR regions (Supplementary Fig. 8). The precision of genotyping deletions and insertions in TR regions using Human 24 Feature Model was at

https://doi.org/10.1038/s41467-025-57756-z

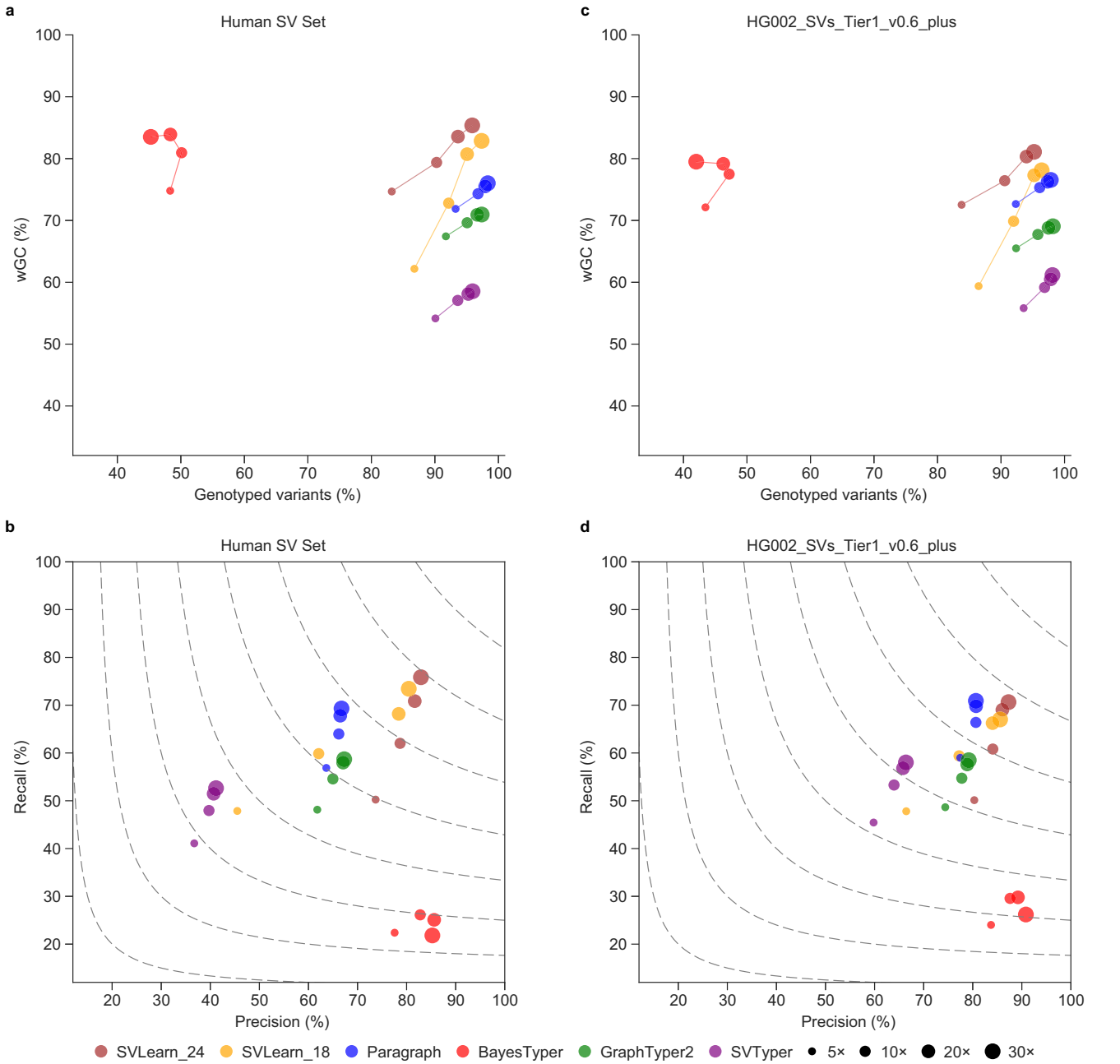


Fig. 3 | Comparison of genotyping tools. a Weighted genotype concordance (wGC) and genotyping rate for Human SV Set. **b** Precision and recall for Human SV Set, with contour lines in the figure representing F1 scores ranging from 0.2 to 0.9. **c** Weighted genotype concordance (wGC) and genotyping rate for HG002_SVs_Tier1_v0.6_plus SV dataset. **d** Precision and recall for HG002_SVs_Tier1_v0.6_plus SV dataset. The performance of Paragraph, BayesTyper, GraphTyper2, SVTyper, SVLearn_18 (Human 18 Feature Model), and SVLearn_24 (Human 24 Feature Model) was evaluated with the HG002 individual in different coverage settings.

least 13.75% and 15.61% higher than that of all the other tools at 30× coverage, except for BayesTyper. Additionally, SVLearn models demonstrated the best recall for genotyping all types of SVs (i.e., deletions inside and outside TRs and insertions inside and outside TRs) across all benchmarked tools, regardless of whether these SVs are associated with TRs.

Impact of sequencing coverage on model performance

Subsequently, we investigated the impact of different sequencing coverage levels on genotyping performance. To this end, we first down-sampled the raw reads of the 14 individuals from 30× coverage to 20×, 10×, and 5× coverage, re-generated training data with 18 and 24 respective features, and performed re-training procedures, leading to the following models: Human 18 Feature 20× Model, Human 18 Feature 10× Model, Human 18 Feature 5× Model, Human 24 Feature 20× Model, Human 24 Feature 10× Model, Human 24 Feature 5× Model. These models were then validated using Human SV Set of HG002 at the four coverage levels. Our results show that the genotyping performance decreases progressively at the inter-coverage level as the coverage decreases, suggesting that an inadequate cohort of genetic material has a detrimental impact on accurate genotyping of SVs (Fig. 4). For example, using 18 features, the best-performing model at 30× coverage, Human 18 Feature 30× model, achieves wGC of over 82%. Yet, this metric is slightly above 80% for the best-performing model (Human 18 Feature 5× model) at 5× coverage. At the intra-coverage level, the best performance peaks at the model that is specifically trained at its

corresponding sequencing coverage. For example, it is observed that at 5× coverage, the Human 18 Feature 5× model achieved a much better wGC value (80.23%) than the Human 18 Feature 30× model (62.18%) (Fig. 4a), in line with F1 score improving from 0.4666 to 0.6787 (Supplementary Fig. 11). In addition, SVLearn models overall show a better genotyping result than other state-of-the-art tools at different coverage levels. For instance, the Human 18 Feature 10× Model and Human 24 Feature 10× Model achieve wGC values of 82.88% and 84.08%, respectively, while Paragraph gains only 74.32% at 10× coverage (Supplementary Fig. 12).

Moreover, it is observed that the fluctuation of these evaluation metrics between different SVLearn models turns to subside at each coverage more clearly if the 6 Paragraph features are added. Human 24 Feature 30× Model genotypes SVs with wGC of 74.68% and an F1 score of 0.5974 at 5× coverage. Notably, wGC and F1 score increased to 81.73% and 0.6781, respectively, by Human 24 Feature 5× Model (Fig. 4b and Supplementary Fig. 11). These findings suggest that coverage-specific training (i.e., matching the training coverage levels to the test coverage levels) significantly improves SVLearn's genotyping performance at different coverage levels (Supplementary Fig. 12).

Propensity for classifying SV genotypes

To fathom the propensity for classifying SV genotypes, we calculated confusion matrices to gain the proportions of correctly and incorrectly classified genotypes. Overall, our results show a significant improvement of using SVLearn to genotype SVs across all three genotypes (0/0, 0/1, 1/1) at all coverage levels compared with BayesTyper, GraphTyper2, Paragraph, and SVTyper (Supplementary Fig. 13).

While all tools encounter challenges in accurately classifying heterozygous SVs (0/1), often misclassifying them as 0/0, SVLearn demonstrates improved performance in addressing this issue. For instance, at 30× coverage of the HG002 individual, SVLearn_24 (24-feature SVLearn models) achieved a genotyping accuracy of 75.34% for 0/1, significantly outperforming all other tools. SVLearn misclassified 0/1 as 0/0 at a rate of only 16.4%, whereas other tools exhibited misclassification rates exceeding 20%. At 5× coverage, the genotyping accuracy for 0/1 by Paragraph dropped significantly from 70.44% to 51.14%, whereas SVLearn demonstrated a more stable performance (from 75.34% to 67.22%). We conjecture that the dual-reference strategy enables more reads to be mapped to the genome, thus leading to the generation of more informative features to handle the classification of different SV genotypes.

In the meantime, we found that SVLearn_24 exhibited minimal genotype flipping at all coverage levels. For instance, SVLearn_24 achieved 85.75% accuracy for 1/1 in the HG002 individual at 30× coverage (Supplementary Fig. 13). Among the misclassified 1/1 SVs, 9.6% were labeled as 0/1, while only 4.65% were completely misclassified as 0/0. This represents the lowest full-flip rate among all the tools tested.

Assessment in cattle and sheep

We also built a series of models specialized for another two critical livestock species: cattle and sheep. We collected PacBio HiFi reads

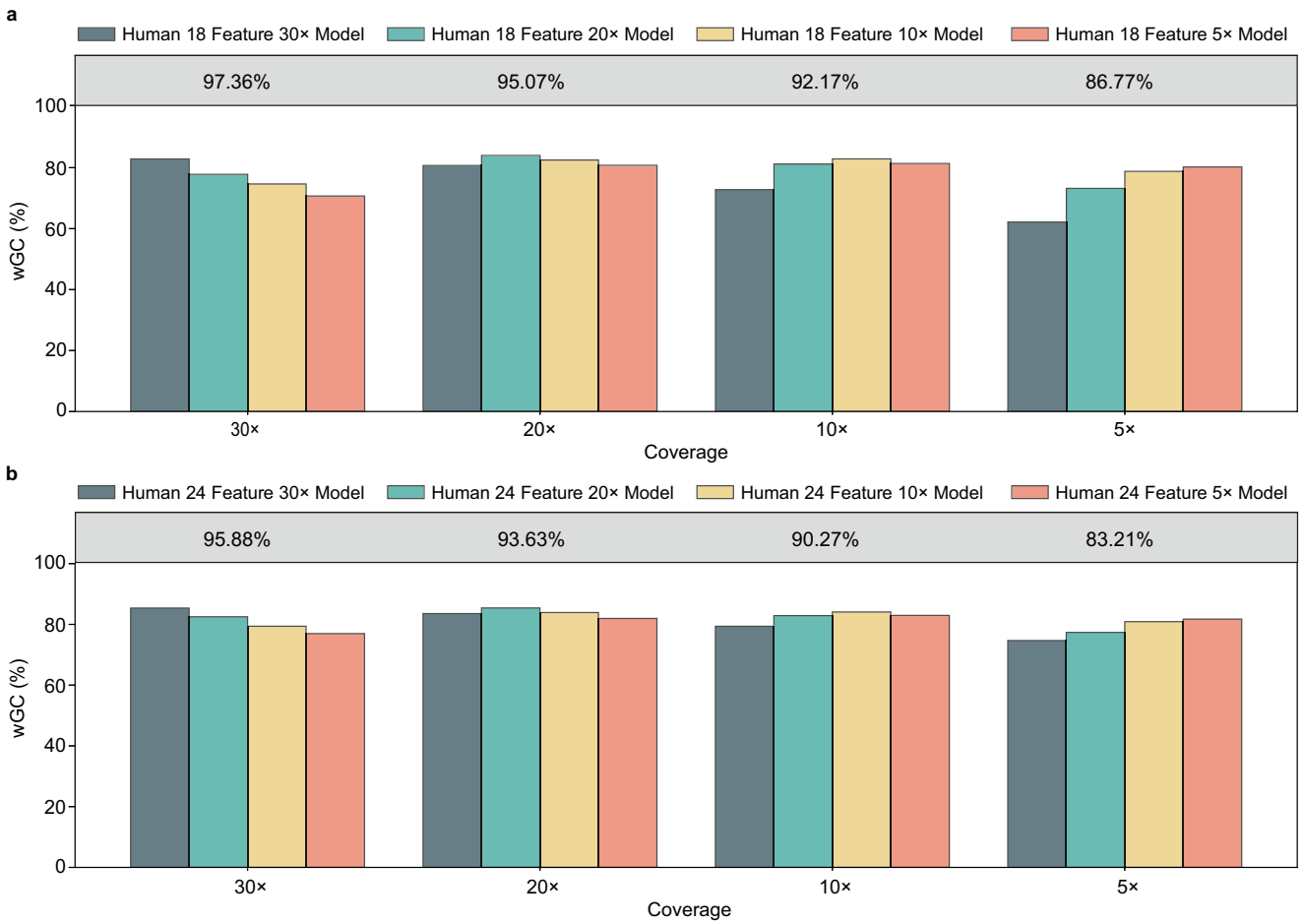


Fig. 4 | Comparison of models trained at multiple coverage levels. a Weighted genotype concordance (wGC) of genotyping SVs using 18 features. **b** Weighted genotype concordance (wGC) of genotyping SVs using 24 features. The x-axis represents the different short-read coverages of the test individual (HG002). The percentages in the shaded area at the top of figures represent the genotyping rate at different coverages of the test individual. The bars in different colors represent models trained with different coverages of the training individuals.

https://doi.org/10.1038/s41467-025-57756-z

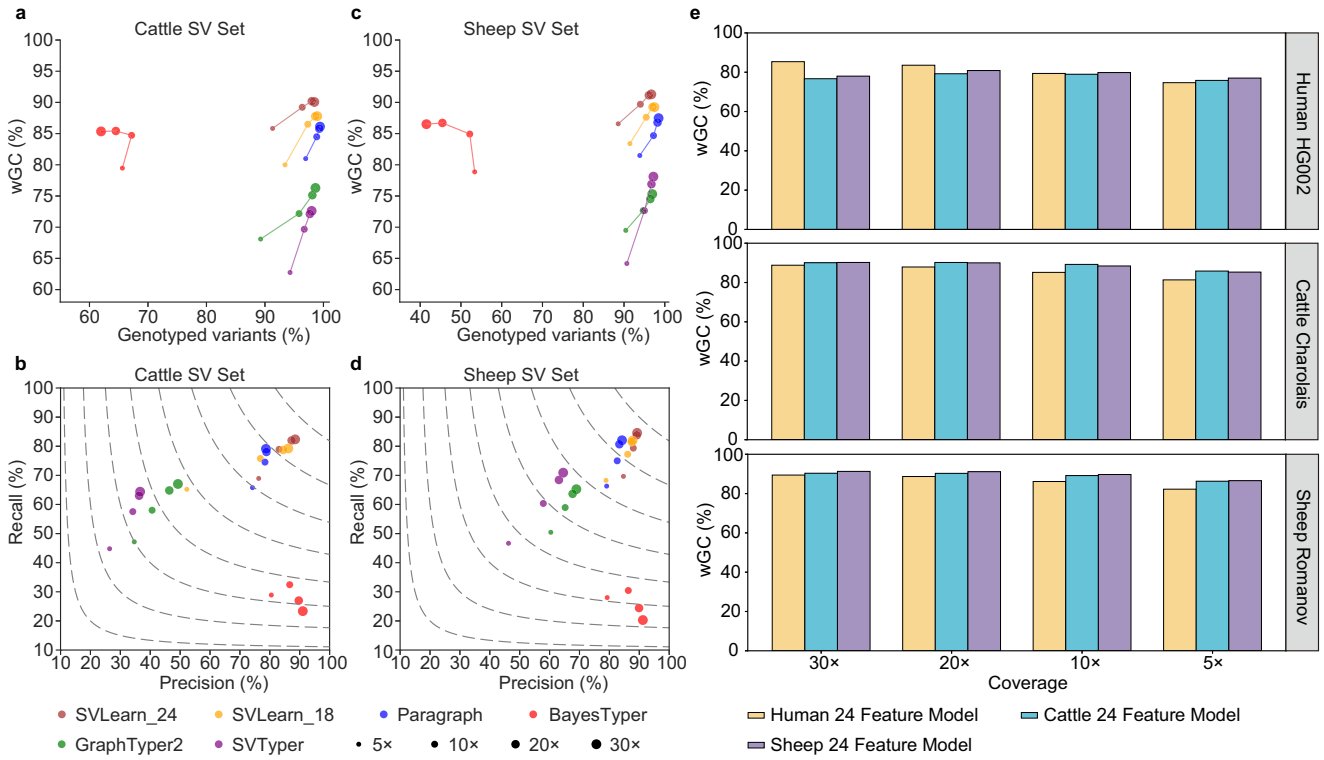


Fig. 5 | SVLearn validation in cattle and sheep data. a Weighted genotype concordance (wGC) and genotyping rate for Cattle SV Set. **b** Precision and recall for Cattle SV Set, with contour lines in the figure representing F1 scores ranging from 0.2 to 0.9. The performance of various tools, including SVLearn with Cattle 18 Feature Model (SVLearn_18) and Cattle 24 Feature Model (SVLearn_24), as well as other tools, was evaluated with the Charolais individual in different coverage settings. **c** Weighted genotype concordance (wGC) and genotyping rate for Sheep SV Set. **d** Precision and recall for Sheep SV Set, with contour lines in the figure representing F1 scores ranging from 0.2 to 0.9. The performance of various tools, including SVLearn with Sheep 18 Feature Model (SVLearn_18) and Sheep 24 Feature Model (SVLearn_24), as well as other tools, was evaluated with the Romanov individual in different coverage settings. **e** Weighted genotype concordance (wGC) of genotyping SVs across species, with human, cattle, and sheep SVLearn models each evaluated on test samples from all three species.

(20× coverage) and 2 × 150 bp short reads of 15 cattle and 15 sheep individuals from previously published studies (Supplementary Tables 1 and 2). Following the same procedures for generation of the Human SV Set, we identified a total number of 121,435 SVs from the cattle individuals (dubbed Cattle SV Set, Supplementary Fig. 17) and 113,042 SVs from the sheep individuals (dubbed Sheep SV Set, Supplementary Fig. 18). We extracted Charolais from the cattle individuals and Romanov from the sheep individuals for validation and left the rest of the individuals for training. Similarly, after feature extraction and training, we obtained Cattle 18 Feature Model, Cattle 24 Feature Model, Sheep 18 Feature model, and Sheep 24 Feature Model. The model training and selection processes can be found in Supplementary Figs. 3–6. In line with the human results, the SVLearn models achieve the best genotyping performance compared to other tools on cattle SVs (Fig. 5a, b) and sheep SVs (Fig. 5c, d). In particular, Cattle 24 Feature Model yields wGC of 90.07% at 30× coverage and F1 score of 0.853 for Charolais, while Sheep 24 Feature Model gives wGC of 91.29% and F1 score of 0.8687 for Romanov. Besides, SVLearn improves the SV genotyping performance by approximately 4%-6% in both metrics for both species compared to the start-of-the-art tool, Paragraph. Furthermore, SVLearn significantly improved the precision of genotyping SVs in TR regions (Supplementary Figs. 19-20).

Examination of cross-species genotyping performance

Lastly, we set out to examine the cross-species genotyping ability of SVLearn. To this end, we tested the best-performing model using 24 features, which was trained specifically for one of the three species, on the SV sets of the remaining two species each time. It can be observed from Fig. 5e that the gain or loss in genotyping performance is not sensitive to the change in the types of species. We also find that the model trained specifically on human-derived SVs performs worse at low levels of coverage compared to the other two species. This possibly arises from the constrained availability of human SV samples for training and the sequencing coverage exclusive to a certain level. The number of human SVs that we were able to collect for training is nearly three times lower than that of either cattle or sheep SVs, leaving the model for humans comparatively insufficiently trained. In addition, there is a substantial difference in sequencing coverage levels between these species. In our case, the sequencing coverage of short reads is ~32× for humans versus ~17× for cattle and ~19× for sheep (Supplementary Table 3). We speculate that the genomic features extracted from reads at high coverage potentially avail Human 24 Feature Model capable of genotyping SVs at similar levels of coverage, and consequently, it performs better at 30× and 20× coverage. The model for humans might be relatively underfitted at low coverage, especially 5×, compared to the models for cattle and sheep. At 10× coverage, there is only a minimal difference in wGC values (79.38%, 78.97%, and 79.83%) and F1 scores (0.6936, 0.6875, and 0.6999) between the human, cattle, and sheep models for genotyping the HG002, respectively (Fig. 5e, Supplementary Fig. 21).

Discussion

SVLearn presented in this study has been demonstrated as a practical approach for accurate genotyping of SVs. The addition of an alternative genome to the reference genome using alternative allele sequences has proven to be of great avail to further boost the performance of SV genotyping. Compared to using only the reference genome, the number of short reads mapped to SV loci in the reference and alternative genomes increased by 45.56% (Fig. 2d). To our knowledge, this scheme has not been implemented and tested by

previously released tools (e.g., Paragraph and GraphTyper2). Accurate genotyping of insertions by previous tools is typically found incredibly challenging and its performance is usually significantly inferior to that for deletions. However, our strategy results in a similar, strong ability of genotyping SVs in insertion regions to that in deletion regions (Supplementary Figs. 14, 15).

Features used for training SVLearn integrate multisource information from genomes, alignments, and genotyping statistics. Through the feature importance analysis and using the leave-one-out strategy, the inclusion of each class of these features proves necessary to progressively improve the genotyping performance. The addition of repeat-related features imparts SVLearn with a notable advantage over other tools in accurately genotyping SVs in repeat regions (Supplementary Fig. 8). The sensitivity of SVLearn models to coverage primarily hinges on the diversity of coverage in our training SVs and the robustness of its selected features. The 18-feature model relatively lacks the global smoothing benefit gained from Paragraph, rendering it more vulnerable to coverage-induced noise. By tapping into the genotyping output of Paragraph, our 24-feature model remains more stable under moderate or low coverage.

Our exploration reveals that the genotyping ability of SVLearn is largely affected by the coverage of short reads at both intra- and inter-species levels. The highest genotyping performance is achieved by the model that is specifically trained at that coverage. At the inter-species level, we show that the performance of SVLearn does not significantly rest upon the species from which training SVs are derived, suggesting a robust generalizability of SVLearn produced especially from cattle and sheep individuals. Even without relying on the haplotype-resolved graph pangenome, the performance of SV genotyping has been improved markedly by SVLearn models trained only with sequence-resolved SVs. This paves the way for rapidly studying the spectrum of large-scale SVs in human and other species populations. We have streamlined the training process in the SVLearn package for mass-producing cross-species models.

Despite the aforementioned advantages, it also has some limitations, such as high computational consumption in handling short reads due to two rounds of mapping to dual references (Supplementary Table 9). Due to the dual-reference strategy, the current version of SVLearn only supports biallelic SV genotyping rather than duplications and inversions. We anticipate that genotyping those SVs beyond the current study can be accelerated by our streamlined processes of feature extraction and model training. Additionally, the information about the local linkage single nucleotide polymorphisms (SNPs) has been successfully used in PanGenie[29] to increase the number of typable SVs and significantly improve the genotyping rate (nearly 100%). This strategy could potentially improve the future versions of SVLearn.

Methods

Sequencing data collection

Sequencing data used throughout this study were sourced from publicly available databases. A total of 45 individuals were collected, of which 15 human individuals were obtained from the Human Pangenome Reference Consortium[35], 15 sheep individuals were derived from a sheep pangenome study[22], and 15 cattle individuals were obtained from five different studies (Charolais[36], Holstein[37], NxB and OxO[38], Yunling[39], the remaining ten individuals[21]). Each individual includes PacBio HiFi long reads and 2×150 bp paired-end short reads (Supplementary Table 1).

Construction of ground truth sets for SV genotypes

Minimap2[40] (version 2.26) was used to align the PacBio HiFi long reads of each human individual to the reference genome GRCh38 (https://ftp.ncbi.nlm.nih.gov/1000genomes/ftp/technical/reference/GRCh38_reference_genome/GRCh38_full_analysis_set_plus_decoy_hla.fa). Subsequently, we used Sniffles2[41] (version 2.2) in multi-sample mode with parameter --minsupport 4 to derive SVs and their genotypes across 15 human individuals. We retained only the insertions and deletions within the range between 50 bp and 1 Mb on the chromosomes (autosomes and XY). To curate high-quality data, we removed any two SVs that were separated by less than 300 bp. This arises from the fact that closely arranged SVs are often characterized intricately and thus difficult to detect by computational methods[17,35,42]. SVs with missing genotypes in any individual were also removed, resulting in the Human SV Set containing genotypes for all 15 individuals. We applied the same procedure to obtain SVs and their genotypes for 15 cattle and 15 sheep individuals using PacBio HiFi long reads, generating the Cattle SV Set and Sheep SV Set, respectively. The reference genome ARS-UCD1.2 (GCF_002263795.1) was used for mapping cattle individuals, while ARS-UI_Ramb_v2.0 (GCF_016772045.1) was used for mapping sheep individuals.

For genotyping performance comparison, we additionally derived SVs and their genotypes based on haplotype-resolved genomes of 15 human individuals. We downloaded haplotype-resolved genomes of the 15 human individuals from the Human Pangenome Reference Consortium (HPRC)[35]. First, we excluded potentially contaminated contigs based on a previously published list (https://github.com/human-pangenomics/HPP_Year1_Assemblies/blob/main/genbank_changes/y1_genbank_remaining_potential_contamination.txt). Next, we called SVs and their genotypes using PAV[17] (version 2.3.4) with default parameters, which were then merged using Jasmine[43] (version 1.1.5). The parameters for Jasmine are provided in the Supplementary Methods. Quality control was subsequently performed using the same criteria applied to the Human SV Set. After removing those SVs overlapped with genomic gap regions, we were left with a total of 39,746 SVs and the genotypes (referred to as the Human Assembly-based SV Set).

Generation of ALT genome

An ALT sequence from the ATL genome was constructed by applying alterations induced by SVs to its corresponding REF sequence from the REF genome. A genomic segment can be overwhelmed by multiple SV events detected from different individuals, such as those 2 small-sized deletions overlapped with 1 large-sized deletion (Supplementary Fig. 1). These overlapping SVs can create multi-allelic bubbles[35,44], thereby undermining the fidelity of SVs as biallelic. To avoid this, we devised two strategies, threshold-based and pseudo-contig-based sequence assemblies, to generate ALT sequences.

Using the threshold-based strategy, SVs occurring in a REF sequence are classified as either overlapping or non-overlapping. An SV is considered as non-overlapping if there exists a distance of at least 300 bp between the SV and each of its adjacent SVs. Finally, only non-overlapped SVs are utilized to construct each ALT sequence. However, this strategy often leads to a significant decrease in the number of SVs retained for assembling the ALT genome. For instance, in dataset *"variants_GRCh38_sv_insdel_alt_HGSVC2024v1.0"*, we were left with 51,110 SVs, less than one third of the raw SVs.

To maximize the retention of SVs, we employed the pseudo-contig-based strategy (Supplementary Fig. 1b), which builds a primary ALT sequence and several pseudo-contigs based on a single REF sequence containing overlapped SVs. The primary sequence was built by replacing the reference sequence at the leftmost overlapped SV (such as DEL2) with its ALT allele. Then, for each of the other overlapped SVs, a pseudo-contig (such as DEL3 and DEL4) for a deletion was built by concatenating 150 bp of flanking sequences on both the upstream and downstream sides of the deletion together, while a pseudo-contig for an insertion (such as INS3) was built by truncating the insertion and incorporating 150 bp of flanking sequences on each side. This approach effectively prevents the overwriting of previously replaced segments on the main chromosome and minimizes disruptions to read alignment.

Finally, genomic positions of the ALT sequences were recorded. Using the ALT sequence and positional information, BED and VCF files of SVs were then produced.

Short reads mapping

Short reads were mapped to the REF and ALT genomes using BWA-MEM2[45] (version 2.2.1) and the mapped reads in the BAM files were sorted using SAMtools[46] (version 1.17). PCR duplicates were removed using Sambamba[47] (version 1.0.1) by setting parameter markdup to -r, yielding BAM files for the two genomes, respectively. The CRAM files aligned to GRCh38 were downloaded for 14 of the 15 human individuals, except for HG002. Subsequently, SAMtools was used to convert the CRAM to BAM. Bazam[48] (version 1.0.1) was employed to extract short reads from the BAM files and remap them to the ALT genome. Mosdepth[49] (version 0.3.6) was utilized to calculate the coverage of both short and long reads with parameters --fast-mode --no-per-base --by 300. We analyzed short reads mapped to SV sites in reference (REF) and alternative (ALT) BAM files for 15 human individuals under 30× coverage. During the counting process, we excluded secondary alignments, PCR duplicates, and reads with a mapping quality of zero, as these reads are generally unsuitable for SV genotyping. To ensure data uniqueness, each read was counted only once when calculating the number of mapped reads in the REF and ALT BAM files.

SV feature

We built ten SV features based on the REF and ALT genomes (Box 1). Using the REF genome, we first extracted three features, SV type (deletion or insertion), SV length, and GC content. The length of the longest allele sequence at an SV locus was regarded as the SV length. The GC content of the SV was calculated based on the longest allele sequence.

Since the longest allele sequences for deletions or insertions are determined exclusive to the use of the REF or ALT genome, the remaining seven features related to deletions and insertions were calculated with different genomes.

Next, interspersed repeats were annotated using RepeatMasker[50] (version 4.1.5). NCBI/RMBLAST (version 2.14.0) was used to search against the Dfam database[51] (version 3.7) and the RepBase database (release 20181026). Tandem repeats were identified using Tandem Repeats Finder[52] (version 4.09) with parameters 2 7 7 80 10 50 500 -h -f -d. For each SV, the repeat classes and contents were finally determined with the annotations from both RepeatMasker and Tandem Repeats Finder. There are a total of ten repeat classes: short interspersed nuclear element (SINE), long interspersed nuclear element (LINE), long terminal repeat (LTR), DNA Transposons, Mixed_TEs, Satellite, variable number of tandem repeat (VNTR), short tandem repeat (STR), Mixed_repeat, and Low_Repeat. An SV was assigned SINE, LINE, LTR, DNA, Satellite, VNTR, or STR if the sequence of the repeat covers over 80% of the SV sequence. An SV was classified as a Mixed_TE if the combined length of SINE, LINE, LTR, and DNA Transposons exceeded 80% of the length of the SV sequence. An SV was classified as Mixed_repeat if the combined length of all the repeat sequences exceeded 80% of the length of the SV sequence. Otherwise, the SV was classified as Low_Repeat. The repeat content was calculated by the percentage of all the repeat sequences in the SV sequence.

To characterize the negative impact of tandem repeats (TRs) on SV genotyping, we used their two compositional features, namely, the length and content of TRs at an SV locus, and calculated them from the output of Tandem Repeats Finder. We only considered the segments of TRs overlapped with the SV sequence. Similarly, the TR repeat content was measured using the percentage of all the TRs in the SV sequence.

For the annotation of segmental duplications (SDs) in the masked genomes, BISER[53] (version 1.4) was used. Subsequently, the regions of SDs in the REF and ALT genomes were picked based on the following criteria[54]: 1) a minimum of 90% gap-compressed identity, 2) a maximum of 50% gapped sequences in the alignment, 3) at least 1 kbp of aligned sequence, and 4) a maximum of 70% satellite sequences determined by RepeatMasker. An SV was assigned an SD class if over 80% or more than 200 bp of its sequence was within SD regions and a non-SD class, otherwise. The SD content was determined as the percentage of the SV sequence in SD regions.

GenMap[55] (version 1.3.0) was used to calculate genome mappability for REF and ALT genomes with parameters -K 50 -E 1 -fl. The median mappability index of the deletion and insertion region of an SV in the REF and ALT genomes was seen as the final mappability.

Alignment feature

We next extracted eight alignment-based features (Alignment features) for each SV from REF and ALT BAM files. An alignment feature can further be categorized into a BreakPoint or ReadDepth class according to whether the information of breakpoints or read depths is required.

Each SV site has three possible genotypes G: homozygous reference HR (0/0), heterozygous HT (0/1), and homozygous alternate HA (1/1), which we refer to as G_{xy}. Read mapping statistics at SV breakpoints were used to calculate genotype likelihoods via a Bayesian classification method similar to SVTyper[33]. $S(G_{xy})$ represents the prior probability of observing an alternate read in a single trial given any genotype G_{xy} at a locus. These priors were set to 0.001, 0.5, and 0.9 for HR, HT, and HA, respectively, as suggested by SVTyper (version 0.7.1).

In the REF BAM, reads spanning SV breakpoints were treated to associate with the reference allele, while split-reads were considered supportive of the alternative allele. There is the other way around for the ALT BAM that reads spanning SV breakpoints were treated to associate with the alternative allele, while split-reads were considered supportive of the reference allele. Specifically, reads spanning SV breakpoints (NM<3) were indicative of the current allele, while split-reads at SV breakpoints (± 3 bp) were supportive of the other allele. Reads with a mapping quality of zero were discarded. In cases where the allele type of a read became inconclusive, we further calculated an alignment score (Ali) to settle the conflict, such that

$$Ali = M - 2 \times NM \tag{1}$$

Here, M stands for the match base number. The allele type with the highest Ali score was retained. Moreover, a read was discarded if both allele types had the same Ali score. The number of reads mapped to SV sites for different genotypes is subjected to a binomial distribution[33], $B(N_R + N_A, S(G_{xy}))$.

$$P(N_R, N_A|G_{xy}) = \frac{(N_R + N_A)!}{N_R! \cdot N_A!} \cdot S(G_{xy})^{N_A} \cdot (1 - S(G_{xy}))^{N_R} \tag{2}$$

Here, N_R and N_A denote the counts of reads supporting the reference and alternative alleles, respectively. Assuming that the prior probabilities of the three genotypes $P(G_{xy})$ are known to be 1/3, their conditional probabilities upon the number of the two allele types $P(HR|N_R, N_A)$, $P(HT|N_R, N_A)$, and $P(HA|N_R, N_A)$ can be calculated using Bayes' theorem, serving as three Breakpoint features: BP_HOMREF_likelihood, BP_HET_likelihood, and BP_HOMALT_likelihood.

$$P(G_{xy}|N_R, N_A) = \frac{P(N_R, N_A|G_{xy}) \cdot P(G_{xy})}{P(N_R, N_A)} = \frac{P(N_R, N_A|G_{xy}) \cdot P(G_{xy})}{\Sigma_{G_{xy} \in G} P(N_R, N_A|G_{xy}) \cdot P(G_{xy})} \tag{3}$$

We exploited the information about the read depth (RD) to further delineate the quality of sequenced reads at SV loci. We constructed a global read depth (RD_Global) feature and two local read depth features (RD_SVd and RD_SVi) according to whether SV loci are

concerned with deletions or insertions. Initially, genome-wide sequencing depths for each 300 bp window were calculated using Mosdepth. After GC correction following CNVcaller[56], we treated the median of the window-placed sequencing depths across the whole genome as RD_Global.

Next, we constructed RD_SVd upon correction for GC bias and removal of reads with mapping quality of zero and base quality of below 20, which was given by the following mapping f

$$f(s)=\begin{cases} d(R_s), & R_s \subseteq \mathrm{REF} \\ \max(0, RD_{Global} - d(R_s)), & otherwise \end{cases} \quad (4)$$

Here, d is a function used to calculate the median sequencing depth at deletion locus s. R_s is a set of reads at the locus mapped to the REF BAM. This feature can reflect the depths of reads supportive of the reference and alternative alleles, respectively[28]. Similarly, to calculate RD_Svi at insertion locus, we constructed a mapping g based on the ALT BAM

$$g(s)=\begin{cases} d(R_s), & R_s \subseteq ALT \\ \max(0, RD_{Global} - d(R_s)), & otherwise \end{cases} \quad (5)$$

The conditional probabilities of the three genotypes were calculated as akin to the BreakPoint feature. The resulting ReadDepth features were denoted as RD_HOMREF_likelihood, RD_HET_likelihood, and RD_HOMALT_likelihood. Prior probabilities of genotypes HR, HT, and HA for observing alternate read ($S(G_{xy})$) were set to 0.0625, 0.5, and 0.99, respectively, as suggested by GraphTyper2[28]. Note that there is a slight change for the last value as a result of optimization of our in silico experiments.

To enhance the ability to genotype SVs within tandem repeat regions of the genome, we sought to characterize SVs using their upstream and downstream sequences, leading to two features

$$RD_{upstream} = \frac{RD_{up}^{REF}}{RD_{up}^{REF} + RD_{up}^{ALT}} \quad (6)$$

$$RD_{downstream} = \frac{RD_{down}^{REF}}{RD_{down}^{REF} + RD_{down}^{ALT}} \quad (7)$$

Here, RD_{up}^{REF} and RD_{down}^{REF} are RDs for each SV locus in the 1 kb regions upstream and downstream calculated using the REF BAM, respectively. RD_{up}^{ALT} and RD_{down}^{ALT} are RDs for each SV locus in the 1 kb regions upstream and downstream calculated using the ALT BAM, respectively. These two features indicate the changes in the coverage of upstream and downstream sequences in the presence or absence of an SV sequence.

Paragraph feature

We used the Paragraph[27] tool for genotyping of SVs and obtained a total of six features. Using the REF and ALT BAMs, we generated two versions of Paragraph-predicted genotyping results for a given SV locus. The two predicted genotypes were used as the first feature and denoted as Ref_GT and Alt_GT, respectively. It should be noticed, in terms of Alt_GT, that the two genotypes 0/0 and 1/1 predicted by Paragraph were converted to 1/1 and 0/0, respectively, as deletions and insertions were treated as one another if the ALT BAM was used. Yet another three features were extracted directly from the fields in the Paragraph output files, including FORMAT/FT, FORMAT/DP, and FORMAT/PL. The FORMAT/FT field accommodates a variety of filtering strategies to classify genotypes of SVs, and denoted as the FT_{REF} and FT_{ALT} features. The FORMAT/DP field (DP_{REF} and DP_{ALT}) provides information relevant with the total sequencing depths of SVs. The FORMAT/PL field displays Phred-scaled likelihoods of genotypes of SVs. The second smallest PL values from this field were extracted as features and denoted as PL_{REF} and PL_{ALT}. In addition, we designed two features to quantify the difference between these existing measurements from the REF and ALT BAMs, which were computed by

$$DP = DP_{REF} - DP_{ALT} \quad (8)$$

$$PL = PL_{REF} - PL_{ALT} \quad (9)$$

Training machine learning models

We trained various machine learning models using two feature sets built with and without the Paragraph features, respectively.

The HG002 individual was used for testing and the remaining 14 individuals were used for training. The genotypes of SVs for all individuals were derived from the BAM of mapped PacBio HiFi reads. To ensure a high-quality dataset, we did not perform data imputation but removed those SVs with missing features instead. Discrete features were uniformly one-hot encoded. A more detailed representation of the features can be found in Supplementary Table 4. Similar preprocessing was applied to cattle and sheep training sets, with Charolais and Romanov as test individuals.

We utilized six classical machine learning algorithms to train SV genotyping models, including Logistic Regression, Naive Bayes, K-Nearest Neighbors, Random Forest, Support Vector Machines, and Gradient Boosting (Box 1). The StratifiedKFold[57] method was employed for 10-fold cross validation as stratified sampling can ensure a relatively even distribution over different classes in each fold. To reproduce our in silico experiments, we generated random seeds with a unified random state (random_state=42). The training, testing, and evaluation procedures were conducted using scikit-learn[58] (version 1.3.0).

Our training results pinpointed Random Forest as the best-performing method. To further reinforce the performance of Random Forest models, we performed a fine-tuning analysis using HalvingGridSearchCV to optimize six hyperparameters (Supplementary Table 7), resulting in six models across three species, including Human 18 Feature Model and Human 24 Feature Model, Cattle 18 Feature Model, Cattle 24 Feature Model, Sheep 18 Feature Model, and Sheep 24 Feature Model. These trained models with respect to the optimal parameters are displayed in Supplementary Table 8.

We performed 15 rounds of leave-one-out cross-validation experiments on the Human SV Set and the Human Assembly-based SV Set, respectively. In each round, SVs and their genotypes from 1 out of 15 individuals were held out as a test set, while the rest of the SVs and genotypes were used for training random forest models. After each round of model training, we used the mean decrease of impurity (MDI)[59] to assess the importance of features by using parameter feature_importances_ in the scikit-learn package. For categorical features that were one-hot encoded, we summed the importance of all dummy variables corresponding to the same original feature. Additionally, we repeated the same procedures for training models specific to cattle and sheep.

We trained coverage-specific models by down-sampling short reads of the 15 human individuals from coverage ~30× to 20×, 10×, and 5×. The Alignment features and the Paragraph features were re-generated at each coverage. Then, ten SV features were combined to train 18 feature models and 24 feature models at three different coverages. Human 18 Feature Model and Human 24 Feature Model were used as the Human 30× models. The models were validated using the HG002 individual at four different coverages. The coverage-specific models were trained on the Human SV Set.

We also performed cross-species validation to examine the genotyping performance of Human 24 Feature Model, Cattle 24 Feature

Model, the Sheep 24 Feature Model on three validation individuals (HG002, Charolais, and Romanov) across multiple sequencing coverage levels.

Finally, a cumulative feature ablation experiment was conducted to determine the contribution of SV features to genotyping SVs in repeat genomic regions. Specifically, the process iterated the model training by cumulatively ablating one SV feature on the human 24-feature dataset. Upon removal of each feature, the model performance was re-evaluated. To ensure the reliability of the results, each model was trained by 15 rounds of leave-one-out experiments in which parameters of the model were determined through stratified 10-fold cross validation.

Comparison with existing genotyping tools

The genotyping performance of SVLearn was compared with Paragraph (version 2.4a), GraphTyper2 (version 2.7.5), BayesTyper (version 1.5), and SVTyper (version 0.7.1) using the test individual at 30×, 20×, 10×, and 5× coverage. Given that SVTyper is specialized for genotyping deletions, we treated insertions as deletions in the ALT BAM and then converted them back to insertions in the REF Genome to gain their genotypes.

We conducted performance comparisons of different tools using HG002 as the test individual across SV sets obtained using four different strategies. In addition to generating the Human SV Set derived from long reads and the Human Assembly-based SV Set derived from haplotype-resolved genomes in this study, we further validated the generalization capability of SVLearn using reliable SV sets publicly available from the Genome-in-a-Bottle Consortium (GIAB) and the Human Genome Structural Variation Consortium (HGSVC).

To derive sequence-resolved SVs from GIAB, we first pulled 7281 insertions and 5464 deletions from the HG002_SVs_Tier1_v0.6[31] dataset. Different from our Human SV Set, a total of 12,745 SVs were detected through alignment of raw sequencing reads of HG002 to the GRCh37 Genome. To expand the volume of SVs, we additionally derived a set of SVs of the HG005 individual from the GIAB HG005 PacBio CCS dataset[60]. The genotype of these SVs is homozygous reference (0/0) for HG002. Upon removal of SVs that are less than 500 bp away from those in the HG002_SVs_Tier1_v0.6, we retained 3706 SVs that were all located within the benchmark intervals of HG002_SVs_Tier1_v0.6.bed. Altogether, we built a set of 16,451 SVs, dubbed HG002_SVs_Tier1_v0.6_plus, for benchmarking the generalization capability of models.

We also derived a large set of SVs from HGSVC. The most recent version of the Phase 3 SV set (variants_GRCh38_sv_insdel_alt_HGSVC2024v1.0) from HGSVC[34], containing 176,231 insertion/deletion (SV) events from 65 individuals. We first removed 1406 SVs located within genomic gap regions or whose reference sequences mismatch the genome, leaving 174,825 SVs. Next, we chose the HG002 individual of interest and excluded 10,354 SVs due to a lack of genotypic information for this individual. We selected subsets of 20,000; 40,000; 60,000; 80,000; 100,000; 120,000; and all 139,254 0/0 SVs (absent in HG002). Each subset was then combined with the 25,217 0/1 or 1/1 SVs present in HG002, resulting in seven SV sets ranging from 45 k to 164 k in size (Supplementary Fig. 16). We then re-genotyped seven SV sets using different tools. Furthermore, we employed the prepareAlt --no-filter-overlaps parameters to retain as many SVs as possible for genotyping and used coverage-specific models to maximize SVLearn's genotyping performance across different coverage levels. To validate the robustness of our approach, we included the analysis of another sample HG00514 and repeated the same data pre-processing procedures. We were left with a set of 164,749 SVs after removing 10,076 SVs with missing genotypic data from 174,825 SVs.

We also compared SVLearn with other tools using the Cattle SV Set and Sheep SV Set, where Charolais served as the test sample for cattle and Romanov served as the test sample for sheep. The detailed commands for running each genotyping tool are accessible at https://github.com/yangqimeng99/svlearn/wiki/Compare-with-other-tools.

Evaluation metrics

Unlike variant calling that identifies SVs and their genomic positions, genotyping is destined for determining the genotypes of known SVs. Thus, the performance of genotyping tools is only evaluated with predicted and known genotypes. We use the predicted and ground-truth labels to compute precision, recall, F1 score, genotype rate, and weighted genotype concordance[29] (wGC) (Supplementary Fig. 2). Particularly, wGC can balance the weights assigned to the genotyping concordance of all three genotypes, preventing the overrepresented category of genotype 0/0 from overshadowing the performance on the less frequent but biologically important 0/1 and 1/1 genotypes (Supplementary Fig. 2). The evaluation process was automated in the SVLearn package. Furthermore, to gain deeper insights into the performance nuances of each genotyping tool, we conducted a stratified analysis based on specific SV features by differentiating insertions and deletions. We performed detailed evaluations from two perspectives: SV size variations (genotyping performance in regard to SV sizes) and genomic locations (genotyping performance in regard to whether SVs are located in tandem repeat regions).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The 2 × 150 bp paired-end short reads and PacBio HiFi long reads were obtained from https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=working for 15 human individuals, from BioProject accession no. PRJNA945429 for 15 sheep individuals, from BioProject accession no. PRJEB55064 for Charolais, from https://ngdc.cncb.ac.cn/gsa/browse/CRA006888 for Holstein, from BioProject accession no. PRJEB42335 for NxB and OxO, and from BioProject accession no. PRJNA978937 for Yunling. For the rest of the cattle individuals, their sequenced reads were all obtained from BioProject accession no. PRJNA786777. The haplotype genomes of 15 human individuals were downloaded from https://github.com/human-pangenomics/HPP_Year1_Assemblies/blob/main/assembly_index/Year1_assemblies_v2_genbank.index. The GIAB HG002_SVs_Tier1_v0.6 SV benchmark set was downloaded at https://ftp.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/AshkenazimTrio/HG002_NA24385_son/NIST_SV_v0.6/. The GIAB HG005 PacBio CCS dataset was downloaded at https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/ChineseTrio/analysis/PacBio_CCS_15kb_20kb_chemistry2_12072020/HG005/HG005.hs37d5.pbsv.vcf.gz. The HGSVC Phase3 SV set was downloaded at https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC3/release/Variant_Calls/1.0/GRCh38/variants_GRCh38_sv_insdel_alt_HGSVC2024v1.0.vcf.gz. The accession number for HG00514 short reads is ERR3988781. The SV genotyping models generated in this study are available from Zenodo [https://doi.org/10.5281/zenodo.11144997][61]. The SV sets, as well as the training and validation datasets produced in this study, are available from Zenodo [https://doi.org/10.5281/zenodo.13309024][62]. Source Data are provided with this paper.

Code availability

SVLearn is available at GitHub (https://github.com/yangqimeng99/svlearn) and Zenodo [https://doi.org/10.5281/zenodo.14897730][63]. The code used for analysis in this study can also be found at GitHub (https://github.com/yangqimeng99/svlearn-paper-code) and Zenodo [https://doi.org/10.5281/zenodo.14891769][64].

References

1. Hollox, E. J., Zuccherato, L. W. & Tucci, S. Genome structural variation in human evolution. *Trends Genet.* **38**, 45–58 (2022).
2. Liu, X. et al. Evolutionary origin of genomic structural variations in domestic yaks. *Nat. Commun.* **14**, 5617 (2023).
3. Cai, Y. et al. Ancient Genomes Reveal the Evolutionary History and Origin of Cashmere-Producing Goats in China. *Mol. Biol. Evol.* **37**, 2099–2109 (2020).
4. Weischenfeldt, J., Symmons, O., Spitz, F. & Korbel, J. O. Phenotypic impact of genomic structural variation: insights from and for human disease. *Nat. Rev. Genet.* **14**, 125–138 (2013).
5. Cortés-Ciriano, I. et al. Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing. *Nat. Genet.* **52**, 331–341 (2020).
6. Yu, Z. et al. Genetic variation across and within individuals. *Nat. Rev. Genet.* **25**, 548–562 (2024).
7. Ho, S. S., Urban, A. E. & Mills, R. E. Structural variation in the sequencing era. *Nat. Rev. Genet.* **21**, 171–189 (2020).
8. Ahsan, M. U., Liu, Q., Perdomo, J. E., Fang, L. & Wang, K. A survey of algorithms for the detection of genomic structural variants from long-read sequencing data. *Nat. Methods* **20**, 1143–1158 (2023).
9. Marx, V. Method of the year: long-read sequencing. *Nat. Methods* **20**, 6–11 (2023).
10. Quan, C., Lu, H., Lu, Y. & Zhou, G. Population-scale genotyping of structural variation in the era of long-read sequencing. *Comput. Struct. Biotechnol.* **20**, 2639–2647 (2022).
11. Zhou, Y. et al. 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 (2022).
12. Jun, G. et al. Structural variation across 138,134 samples in the TOPMed consortium. Preprint at *bioRxiv* https://doi.org/10.1101/2023.01.25.525428 (2023).
13. Yang, L. et al. Mapping and functional characterization of structural variation in 1060 pig genomes. *Genome Biol.* **25**, 116 (2024).
14. Du, Z.-Z., He, J.-B. & Jiao, W.-B. A comprehensive benchmark of graph-based genetic variant genotyping algorithms on plant genomes for creating an accurate ensemble pipeline. *Genome Biol.* **25**, 91 (2024).
15. Huddleston, J. et al. Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Res.* **27**, 677–685 (2017).
16. Aganezov, S. et al. A complete reference genome improves analysis of human genetic variation. *Science* **376**, eabl3533 (2022).
17. Ebert, P. et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372**, eabf7117 (2021).
18. Quan, C. et al. Characterization of structural variation in Tibetans reveals new evidence of high-altitude adaptation and introgression. *Genome Biol.* **22**, 159 (2021).
19. Leonard, A. S., Mapel, X. M. & Pausch, H. Pangenome genotyped structural variation improves molecular phenotype mapping in cattle. *Genome Res* **34**, 300–309 (2024).
20. Shi, J. et al. Structural variants involved in high-altitude adaptation detected using single-molecule long-read sequencing. *Nat. Commun.* **14**, 8282 (2023).
21. Dai, X. et al. A Chinese indicine pangenome reveals a wealth of novel structural variants introgressed from other Bos species. *Genome Res.* **33**, 1284–1298 (2023).
22. Li, R. et al. A sheep pangenome reveals the spectrum of structural variations and their effects on tail phenotypes. *Genome Res.* **33**, 463–477 (2023).
23. Jin, S. et al. Structural variation (SV)-based pan-genome and GWAS reveal the impacts of SVs on the speciation and diversification of allotetraploid cottons. *Molecular Plant* **16**, 678–693 (2023).
24. Garrison, E. et al. Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nat. Biotechnol.* **36**, 875–879 (2018).
25. Hickey, G. et al. Genotyping structural variants in pangenome graphs using the vg toolkit. *Genome Biol.* **21**, 35 (2020).
26. Sirén, J. et al. Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. *Science* **374**, abg8871 (2021).
27. Chen, S. et al. Paragraph: a graph-based structural variant genotyper for short-read sequence data. *Genome Biol.* **20**, 291 (2019).
28. Eggertsson, H. P. et al. GraphTyper2 enables population-scale genotyping of structural variation using pangenome graphs. *Nat. Commun.* **10**, 5402 (2019).
29. Ebler, J. et al. Pangenome-based genome inference allows efficient and accurate genotyping across a wide spectrum of variant classes. *Nat. Genet.* **54**, 518–525 (2022).
30. Chen, N.-C., Solomon, B., Mun, T., Iyer, S. & Langmead, B. Reference flow: reducing reference bias using multiple population genomes. *Genome Biol.* **22**, 8 (2021).
31. Zook, J. M. et al. A robust benchmark for detection of germline large deletions and insertions. *Nat. Biotechnol.* **38**, 1347–1355 (2020).
32. Sibbesen, J. A., Maretty, L. & Krogh, A. Accurate genotyping across variant classes and lengths using variant graphs. *Nat. Genet.* **50**, 1054–1059 (2018).
33. Chiang, C. et al. SpeedSeq: ultra-fast personal genome analysis and interpretation. *Nat. Methods* **12**, 966–968 (2015).
34. Logsdon, G. A. et al. Complex genetic variation in nearly complete human genomes. Preprint at *bioRxiv* https://doi.org/10.1101/2024.09.24.614721 (2024).
35. Liao, W.-W. et al. A draft human pangenome reference. *Nature* **617**, 312–324 (2023).
36. Eché, C. et al. A Bos taurus sequencing methods benchmark for assembly, haplotyping, and variant calling. *Sci. Data* **10**, 369 (2023).
37. Li, T.-T. et al. De novo genome assembly depicts the immune genomic characteristics of cattle. *Nat. Commun.* **14**, 6601 (2023).
38. Leonard, A. S. et al. Structural variant-based pangenome construction has low sensitivity to variability of haplotype-resolved bovine assemblies. *Nat. Commun.* **13**, 3012 (2022).
39. Chen, J. et al. Population Structure and Genetic Diversity of Yunling Cattle Determined by Whole-Genome Resequencing. *Genes* **14**, 2141 (2023).
40. Li, H. New strategies to improve minimap2 alignment accuracy. *Bioinformatics* **37**, 4572–4574 (2021).
41. Smolka, M. et al. Detection of mosaic and population-level structural variants with Sniffles2. *Nat. Biotechnol.* **42**, 1571–1580 (2024).
42. Milia, S. et al. Taurine pangenome uncovers a segmental duplication upstream of KIT associated with depigmentation in white-headed cattle. *Genome Res.* https://doi.org/10.1101/gr.279064.124 (2024). Online ahead of print.
43. Kirsche, M. et al. Jasmine and Iris: population-scale structural variant comparison and analysis. *Nat. Methods* **20**, 408–417 (2023).
44. Crysnanto, D., Leonard, A. S., Fang, Z.-H. & Pausch, H. Novel functional sequences uncovered through a bovine multiassembly graph. *Proc. Natl. Acad. Sci. USA* **118**, e2101056118 (2021).
45. Vasimuddin, Md., Misra, S., Li, H. & Aluru, S. Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems. in *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)* 314–324 (2019). https://doi.org/10.1109/IPDPS.2019.00041.
46. Danecek, P. et al. Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008 (2021).
47. Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J. & Prins, P. Sambamba: fast processing of NGS alignment formats. *Bioinformatics* **31**, 2032–2034 (2015).
48. Sadedin, S. P. & Oshlack, A. Bazam: a rapid method for read extraction and realignment of high-throughput sequencing data. *Genome Biol.* **20**, 78 (2019).

49. Pedersen, B. S. & Quinlan, A. R. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics* **34**, 867–868 (2018).
50. Smit, A. F. A., Hubley, R. & Green, P. RepeatMasker Open-4.0. http://www.repeatmasker.org (2013).
51. Storer, J., Hubley, R., Rosen, J., Wheeler, T. J. & Smit, A. F. The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mob. DNA* **12**, 2 (2021).
52. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
53. Išerić, H., Alkan, C., Hach, F. & Numanagić, I. Fast characterization of segmental duplication structure in multiple genome assemblies. *Algorithms Mol. Biol.* **17**, 4 (2022).
54. Vollger, M. R. et al. Segmental duplications and their variation in a complete human genome. *Science* **376**, eabj6965 (2022).
55. Pockrandt, C., Alzamel, M., Iliopoulos, C. S. & Reinert, K. GenMap: ultra-fast computation of genome mappability. *Bioinformatics* **36**, 3687–3692 (2020).
56. Wang, X. et al. CNVcaller: highly efficient and widely applicable software for detecting copy number variations in large populations. *GigaScience* **6**, gix115 (2017).
57. Sharma, H., Zerbe, N., Klempert, I., Hellwich, O. & Hufnagl, P. Deep convolutional neural networks for automatic classification of gastric carcinoma using whole slide images in digital histopathology. *Comput. Med. Imag. Graph.* **61**, 2–13 (2017).
58. Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
59. Louppe, G., Wehenkel, L., Sutera, A. & Geurts, P. Understanding variable importances in forests of randomized trees. *Adv. Neural Inf. Process. Syst.* **26**, Available at: https://proceedings.neurips.cc/paper/2013/hash/e3796ae838835da0b6f6ea37bcf8bcb7-Abstract.html (2013).
60. Zook, J. M. et al. Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci. Data* **3**, 160025 (2016).
61. Yang, Q. et al. Machine learning model for SVLearn. *Zenodo* https://doi.org/10.5281/zenodo.11144997 (2024).
62. Yang, Q. et al. Dataset for SVLearn. *Zenodo* https://doi.org/10.5281/zenodo.13309024 (2024).
63. Yang, Q. SVLearn-v0.0.5. *Zenodo* https://doi.org/10.5281/zenodo.14897730 (2025).
64. Yang, Q. SVLearn-paper-code. *Zenodo* https://doi.org/10.5281/zenodo.14891769 (2025).

Acknowledgements

This work was supported by grants from the National Key R&D Program of China (2023YFD1300402 to Y.C., 2022YFF1000100 to Y.J.), National Natural Science Foundation of China (U21A20120 to Y.J.) and Shaanxi Livestock and Poultry Breeding Double-chain Fusion Key Project (2022GD-TSLD-46-0401 to Y.J.). We thank the High-Performance Computing platform of Northwest A&F University and Computing Center in Xi'an for providing computing resources. We also thank the Human Genome Structural Variation Consortium for publicly sharing their structural variation (SV) dataset, which, along with the true genotypes of the test samples, significantly aided in the evaluation of our software.

Author contributions

Y.J. and Y.C. designed and coordinated the study. Q.Y., Y.C. and J.S. developed the algorithm and software. J.S., X.W. and J.W. contributed to the assessment and analysis of the machine learning models. Q.Y. and X.W. benchmarked the tools. Q.L., J.R. and X.Z. contributed to the code testing process. S.W. and R.H. collected the sequencing data. P.B., X.D., M.G., Z.Z., A.W., F.B. and R.L. provided constructive suggestions for the software. Q.Y. and J.S. wrote a draft paper and all authors contributed edits and comments. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-57756-z.

Correspondence and requests for materials should be addressed to Yudong Cai or Yu Jiang.

Peer review information *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work. A peer review file is available.