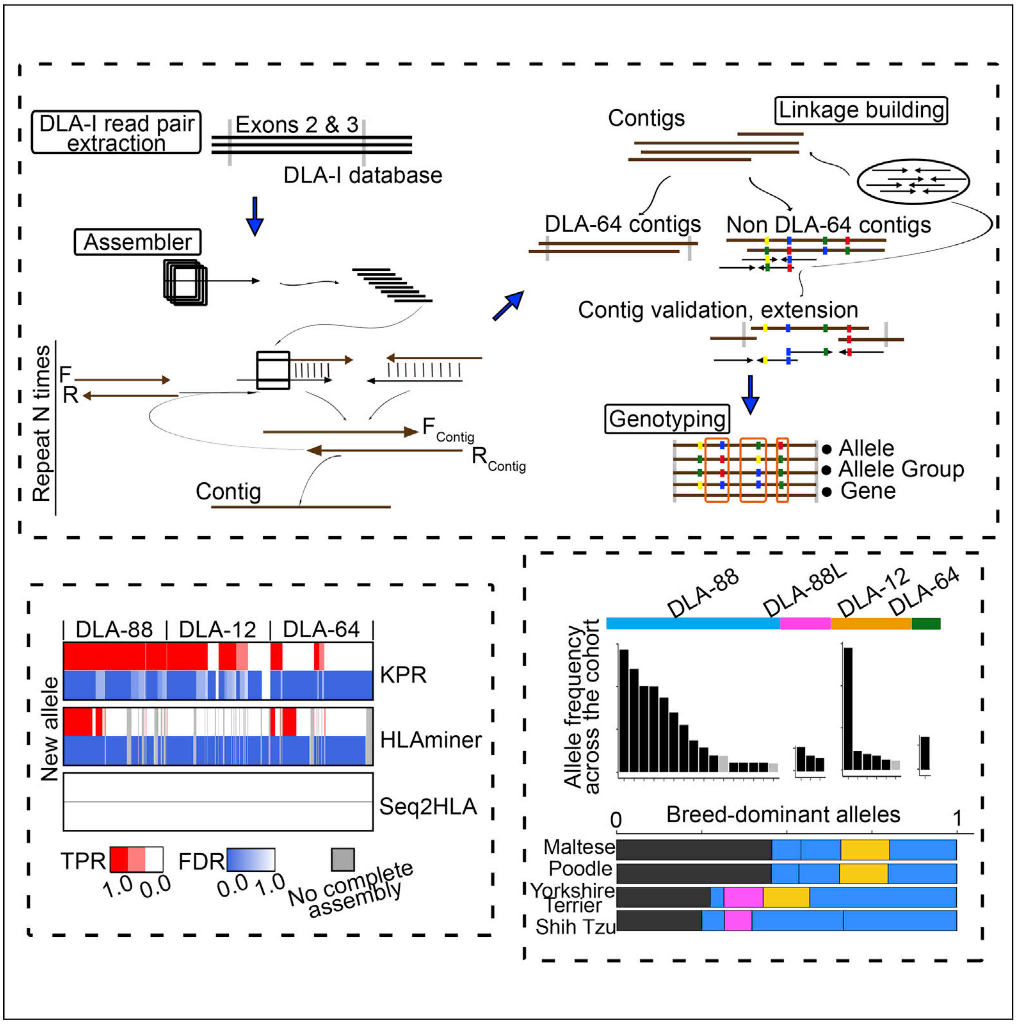
iScience

CellPress
OPEN ACCESS

Article

A Kmer-based paired-end read *de novo* assembler and genotyper for canine MHC class I genotyping



Yuan Feng, Paul R. Hess, Stephen M. Tompkins, William H. Hildebrand, Shaying Zhao

szhao@uga.edu

Highlights

KPR software genotypes DLA-I alleles and estimates expression using RNA-seq data

KPR performs *de novo* assembly and outperforms other tools in typing new alleles

Typing 152 dogs uncovers 33 putative new alleles and dominant alleles in 4 breeds

DLA-12 and DLA-88L alleles cluster with DLA-88 alleles

Feng et al., iScience *26*, 105996
February 17, 2023 © 2023 The Author(s).
https://doi.org/10.1016/j.isci.2023.105996

iScience

CellPress
OPEN ACCESS

Article

A Kmer-based paired-end read *de novo* assembler and genotyper for canine MHC class I genotyping

Yuan Feng,[1] Paul R. Hess,[2] Stephen M. Tompkins,[3] William H. Hildebrand,[4] and Shaying Zhao[1,5,*]

SUMMARY

The major histocompatibility complex class I (MHC-I) genes are highly polymorphic. MHC-I genotyping is required for determining the peptide epitopes available to an individual's T-cell repertoire. Current genotyping software tools do not work for the dog, due to very limited known canine alleles. To address this, we developed a Kmer-based paired-end read (KPR) *de novo* assembler and genotyper, which assemble paired-end RNA-seq reads from MHC-I regions into contigs, and then genotype each contig and estimate its expression level. KPR tools outperform other popular software examined in typing new alleles. We used KPR tools to successfully genotype152 dogs from a published dataset. The study discovers 33 putative new alleles, finds dominant alleles in 4 dog breeds, and builds allele diversity and expression landscapes among the 152 dogs. Our software meets a significant need in biomedical research.

INTRODUCTION

The dog serves as an important translational model of cancer, infectious disease, obesity, neurological and other disorders in humans. The canine model has the great potential to effectively bridge a current gap between preclinical models and human clinical trials,[1–5] accelerating drug discovery. This is because, unlike traditional models such as cell lines or rodent models, canine diseases arise spontaneously in animals that share the same environment as humans and have intact immune systems, thereby recapitulating the essence of human diseases. Indeed, a significant molecular homology has been noted between various canine cancer types/subtypes and their human counterparts.[6–13] However, the effective use of the canine model is at present significantly hampered by the lack of essential resource, including software tools for major histocompatibility complex class I (MHC-I) genotyping.

A human individual expresses 3–6 alleles encoded by three classical MHC-I genes HLA-A, HLA-B, and HLA-C (human MHC-I will be referred to as HLA-I hereafter). The syntenic canine MHC-I region encodes genes including DLA-88, DLA-12, and DLA-64 (canine MHC-I will be referred to as DLA-I hereafter). However, only DLA-88, being the most diversified and with the highest expression level, is considered a classical MHC-I gene.[14–17] The status of DLA-12 and DLA-64 is currently unclear,[17] due to limited data. The DLA-12 locus is reported to encodes not only DLA-12 alleles, but also DLA-88 like (DLA-88L) alleles.[17,18]

The three MHC-I genes share >90% sequence identity overall. Moreover, their exons 2 and 3, which encode the antigen-binding pocket, are among the most polymorphic regions in the genome. MHC-I genotyping (accurate determination of the ~550bp sequence of exons 2 and 3) is necessary to call alleles, which, in turn, are needed in predicting the peptide ligandome presented by an individual's collective classical MHC-I molecules to the CD8$^+$ T-cell repertoire. Thus, MHC-I genotyping is important for research on cancer (e.g., tumor-specific neoantigen discovery),[19,20] infectious disease, and other health issues.

MHC-I genotyping is traditionally achieved by PCR cloning-based Sanger sequencing.[14–17] While considered the gold standard, this method is time-consuming, labor-intensive, and low-throughput. As numerous individuals have undergone deep sequencing, many algorithms have been developed for MHC-I genotyping using these massive genome-wide data, including RNA-seq.[21–30] The challenge is that these sequence reads, generated by Illumina or similar next–generation sequencing (NGS) technologies, are often short (usually <150bp). This limitation, coupled with the high sequence homology among the three MHC-I genes and their highly polymorphic nature among individuals, complicates genotyping by this methodology.

[1]Department of Biochemistry and Molecular Biology, Institute of Bioinformatics, University of Georgia, Athens, GA 30602, USA

[2]Department of Clinical Sciences, North Carolina State University, College of Veterinary Medicine, Raleigh, NC 27607, USA

[3]Center for Vaccines and Immunology, University of Georgia, UGA, Athens, GA 30602, USA

[4]Department of Microbiology and Immunology, University of Oklahoma Health Sciences Center, Oklahoma City, OK 73104, USA

[5]Lead contact

*Correspondence: szhao@uga.edu

https://doi.org/10.1016/j.isci.2023.105996

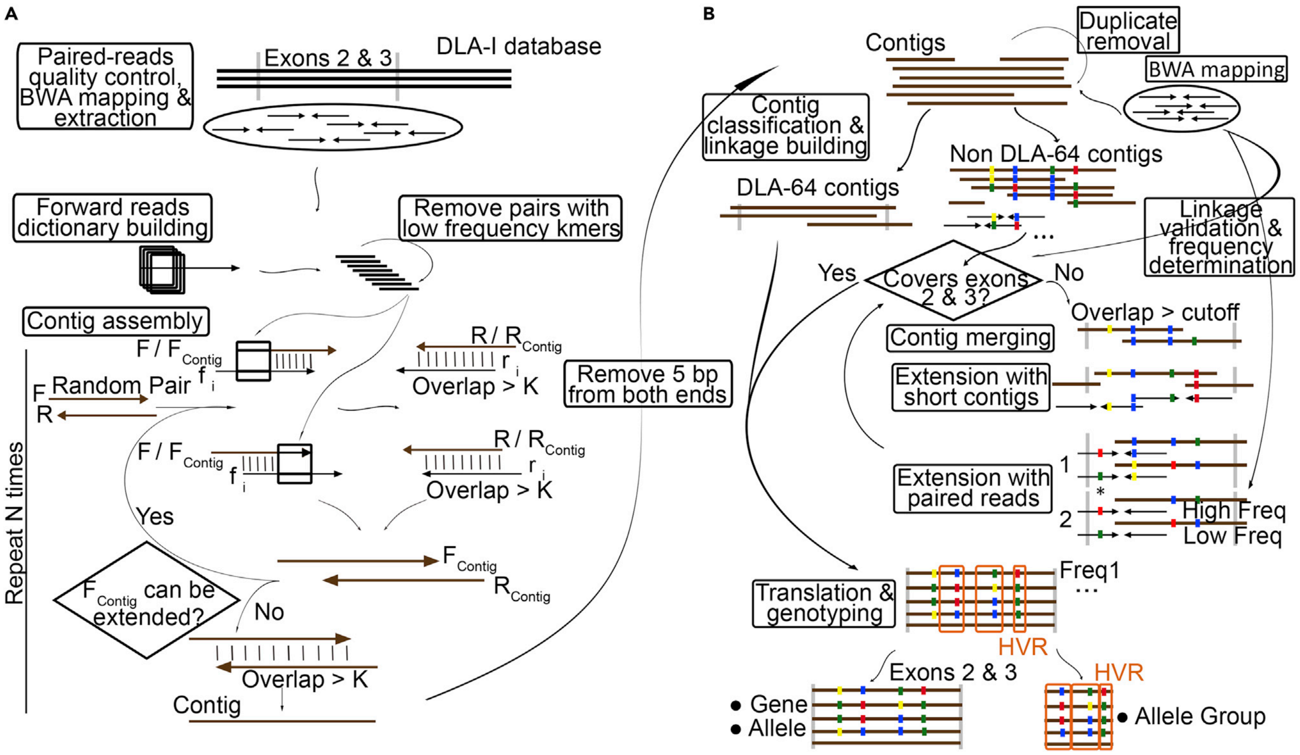


Figure 1. Kmer-based paired-end read (KPR) *de novo* assembler and genotyper conduct DLA-I genotyping using paired-end RNA-seq data

(A) Our KPR assembler assembles the highly polymorphic region, the entirety of exons 2 and 3, of DLA-I alleles of an individual *de novo*, using paired-end RNA-seq reads (see STAR Methods). Contigs are represented by bars, while paired-end reads are represented by paired-arrows facing each other. HVR: hypervariable region.

(B) Our genotyper genotypes each assembled contig with the entire exon 2 and 3 sequence. For non-DLA-64 contigs, genotyping is done after variant linkage building, validation and if needed, extension (see STAR Methods) are performed. Sequence variants are indicated by colored dots.

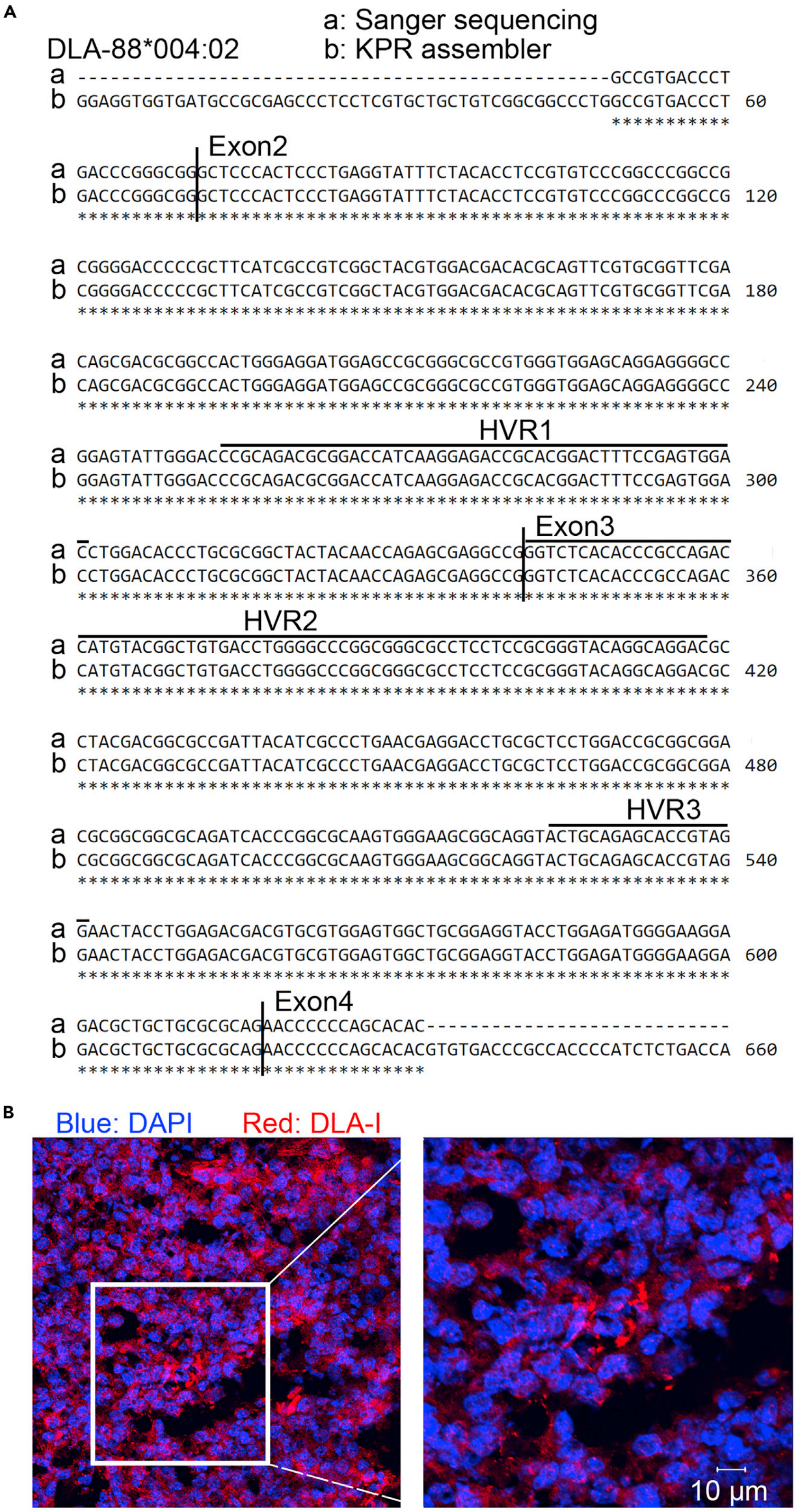
In humans, the challenge is addressed by taking advantage of the vast number of known HLA-I alleles,[21–30] about 23,694 in the IMGT/HLA database[31] as of March 2022. With the assumption that many HLA-I alleles existing in the human population have already been defined, many current HLA-I genotyping tools focus on accurately mapping the NGS reads to the known allele reference and then genotyping the individual by identifying known alleles with the most unambiguously placed NGS reads. A few tools perform assembly, which either aims to extend the short reads to achieve more accurate mapping to known alleles,[24] or is reference-based but not *de novo* assembly.[25] Hence, these assembly performing tools still heavily rely on the catalog of known alleles.

In dogs, known DLA-I alleles are very limited, with 150 alleles for DLA-88, 8 alleles for DLA-88L, 20 alleles for DLA-12, and 7 alleles for DLA-64 (185 alleles in total).[14–18] Hence, current RNA-seq based MHC-I genotyping tools,[21–30] whose accuracy heavily depends on the comprehensiveness of known alleles, do not work for the dog. To address this deficiency, we are developing software tools that perform *de novo* assembly[32] of complete contigs for genotyping. Our software outperforms widely used genotyping tools in typing new alleles.

RESULTS

Kmer-based paired-end read *de novo* assembler and genotyper conduct DLA-I genotyping using paired-end RNA-seq data

Our KPR assembler assembles the highly polymorphic region, the entirety of exons 2 and 3, of DLA-I alleles of an individual *de novo*, using paired-end RNA-seq reads that are mapped to a reference consisting of 196 known DLA-I alleles (Figure 1A). Our genotyper then genotypes each complete contig assembled, and estimates the expression level of each allele, after contig validation and if needed, contig extension (Figure 1B).



```
a -------------------------------------------------GCCGTGACCCT
b GGAGGTGGTGATGCCGCGAGCCCTCCTCGTGCTGCTGTCGGCGGCCCTGGCCGTGACCCT  60
                                                   ***********

a GACCCGGGCGGGCTCCCACTCCCTGAGGTATTTCTACACCTCCGTGTCCCGGCCCGGCCG
b GACCCGGGCGGGCTCCCACTCCCTGAGGTATTTCTACACCTCCGTGTCCCGGCCCGGCCG  120
  ************************************************************

a CGGGGACCCCCGCTTCATCGCCGTCGGCTACGTGGACGACACGCAGTTCGTGCGGTTCGA
b CGGGGACCCCCGCTTCATCGCCGTCGGCTACGTGGACGACACGCAGTTCGTGCGGTTCGA  180
  ************************************************************

a CAGCGACGCGGCCACTGGGAGGATGGAGCCGCGGGCGCCGTGGGTGGAGCAGGAGGGGCC
b CAGCGACGCGGCCACTGGGAGGATGGAGCCGCGGGCGCCGTGGGTGGAGCAGGAGGGGCC  240
  ************************************************************

a GGAGTATTGGGACCCGCAGACGCGGACCATCAAGGAGACCGCACGGACTTTCCGAGTGGA
b GGAGTATTGGGACCCGCAGACGCGGACCATCAAGGAGACCGCACGGACTTTCCGAGTGGA  300
  ************************************************************

a CCTGGACACCCTGCGCGGCTACTACAACCAGAGCGAGGCCGGGTCTCACACCCGCCAGAC
b CCTGGACACCCTGCGCGGCTACTACAACCAGAGCGAGGCCGGGTCTCACACCCGCCAGAC  360
  ************************************************************

a CATGTACGGCTGTGACCTGGGGCCCGGCGGGCGCCTCCTCCGCGGGTACAGGCAGGACGC
b CATGTACGGCTGTGACCTGGGGCCCGGCGGGCGCCTCCTCCGCGGGTACAGGCAGGACGC  420
  ************************************************************

a CTACGACGGCGCCGATTACATCGCCCTGAACGAGGACCTGCGCTCCTGGACCGCGGCGGA
b CTACGACGGCGCCGATTACATCGCCCTGAACGAGGACCTGCGCTCCTGGACCGCGGCGGA  480
  ************************************************************

a CGCGGCGGCGCAGATCACCCGGCGCAAGTGGGAAGCGGCAGGTACTGCAGAGCACCGTAG
b CGCGGCGGCGCAGATCACCCGGCGCAAGTGGGAAGCGGCAGGTACTGCAGAGCACCGTAG  540
  ************************************************************

a GAACTACCTGGAGACGACGTGCGTGGAGTGGCTGCGGAGGTACCTGGAGATGGGGAAGGA
b GAACTACCTGGAGACGACGTGCGTGGAGTGGCTGCGGAGGTACCTGGAGATGGGGAAGGA  600
  ************************************************************

a GACGCTGCTGCGCGCAGAACCCCCCAGCACAC------------------------------
b GACGCTGCTGCGCGCAGAACCCCCCAGCACACGTGTGACCCGCCACCCCATCTCTGACCA  660
  ******************************
```

Figure 2. Contigs assembled by the KPR software are validated by Sanger sequencing

(A) Sanger sequence is identical to the contig assembled with RNA-seq reads of the same dog by the KPR software. HVRs are drawn as reported previously.[16,34]

(B) Representative confocal images indicate that the MHC-I protein complex is expressed in the same tissue of the dog. See Table S1.

Note that paired-end reads are both derived from the same allele, and our core algorithm relies on paired-end sequencing for contig assembly, verification, and extension (Figure 1). Thus, our KPR tools do not rely on the completeness of the known allele reference as heavily as existing genotyping tools and can type new alleles. The detailed algorithms are described in the STAR Methods section.

The contigs generated by the Kmer-based paired-end read assembler are validated by Sanger sequencing

We applied the KPR assembler (Figure 1A) to the canine RNA-seq data generated by our group,[7,33] which consists of 27–87 million paired-end reads of 76 x 76bp per sample. We performed PCR cloning-based Sanger sequencing for DLA-88 by sequencing 5–10 clones per sample for 7 samples. The experiment validated the DLA-88 assemblies at the nucleotide level. For example, both KPR and Sanger sequencing identified a single allele (DLA-88*004:02) in one tumor (Figure 2A), two alleles (DLA-88*012:01 and DLA-88*032:01) in MDCKII cells,[33] and 4 alleles (DLA-88*501:01, DLA-88*501:02, DLA-88*004:02; DLA-88*004:New1) in tumor 2577T (Table S1). In tumor LilyT, all 7 Sanger sequenced clones support the highly expressed allele (DLA-88*004:02) genotyped by KPR (Table S1). In tumor T76, the only KPR-genotyped allele (DLA-88*051:01) is validated by three Sanger sequenced clones (Table S1).

To more quantitatively assess the validation, we performed paired T and Wilcoxon tests on DLA-88 alleles of the 7 samples genotyped by the KPR software and Sanger sequencing (Table S1). Both tests indicate no significant differences between KPR and Sanger sequencing in either the alleles or their expression levels (p values range from 0.33 to 0.78; see Table S1).

We also performed immunostaining experiments to show the expression of the MHC-I protein complex by these cells (Figure 2B). These results support that our KPR tools are effective.

Kmer-based paired-end read running parameters are optimized via simulation

Three parameters are required to run the KPR tools, including the Kmer length (K), the number of assembly runs (N), and the read pair depth (PD) for potential misassembly identification (Figure 1).

To evaluate these parameters, we performed simulations based on the largest RNA-seq study published so far for the dog.[35,36] The dataset consists of 25–77 million RNA-seq read pairs of 101 x 101bp per sample for 222 tumor and/or normal samples from 158 dogs (Table S2).

We first conducted comprehensive data quality control (QC), examining the sequencing amount and quality, as well as using germline mutations to validate the tumor-normal pairing accuracy and breed data accuracy as described[6] (Figures S1A–S1J). After excluding one sample of low quality and correcting three samples from two dogs with misassigned breeds (Table S2), all other samples passed the QC measures.

We first chose three dogs with paired normal and tumor samples (six in total) from this cohort (Table S2). The six samples represent the majority of the dogs of the cohort in DLA-I read total amount and distribution, as well as the sequencing error rate (Table S2; Figures S2A–S2C).

For allele simulation, we randomly chose 1 or 2 alleles per DLA-I gene from the known allele database. For each allele, we simulated its sequence read-pairs in a sample chosen above using the actual read length, insert size, and expression level of the corresponding DLA-I gene, as well as the overall sequencing error rate of the sample (see STAR Methods). Then, we replaced the actual DLA-I read pairs with the simulated ones. We made 10 allele combinations (ACs) per sample, by varying DLA-88 allele type, canonical or 50X (which encodes an extra amino acid residue), and homozygosity or heterozygosity of DLA-88 and DLA-12 (DLA-64 is not considered because it is more distant, separately genotyped as shown in Figure 1B, and has only 3 known alleles) (Table S2). We randomly sampled each AC three times. These yield a total of $6\ (samples) \times 10\ (ACs) \times 3\ (repeats) = 180$ simulated samples.

PD is a parameter for ''variant linkage validation,'' a function implemented in KPR to use paired-end RNA-seq reads to identify misassemblies for contig validation (Figure 1B). Briefly, variant linkages are built via multiple sequence alignment of contigs assembled for a sample. If a linkage is also found in one or more RNA-seq read pairs mapped to the contig, then the linkage is validated. This however needs a certain amount of RNA-seq read pair coverage over the linkage, i.e., PD. In KPR, we define PD as the maximum read pairs that are mapped to any two polymorphic sites among true positive alleles of a sample when none of the read pairs supports the variant linkage of a true positive allele. If the total read pair coverage is >PD and yet no any read pair supports the linkage, the contig will be classified as misassembly.

To estimate PD, we genotyped the 180 simulated samples at different K and N values. As shown in Figure 3A, PD stays roughly at 15 for known DLA-88 alleles at K and N values when false negatives are minimized. For known DLA-12 alleles, PD is around 38 (Figure S3A). We thus set $PD = 15$ for DLA-88 and $PD = 38$ for DLA-12 (no PD cutoff is used for DLA-64 in our KPR software; see Figures 1 and S3B) for further studies.

For all studies later in discussion, we used two values, true positive rate (TPR) and false discovery rate (FDR) to evaluate the genotyping results. TPR and FDR were calculated using either allele number (treating individual alleles as 1 or 0), or allele expression (using the expression levels of individual allele) (see STAR Methods).

To evaluate N, we set $K = 50bp$ and varied N from 1 to 30,000 to genotype the 180 simulated samples. The genotyping results of DLA-88 are stabilized when N is > 100, especially after variant linkage validation with paired-end reads (Figure 3B). For allele number, the mean TPR ranges from 0.98 to 0.99 while the mean FDR ranges from 0.11 to 0.12 (Figure 3B; Table S2). For allele expression, the mean TPR ranges from 0.87 to 0.88 and the mean FDR is at approximately 0.2 (Figure 3B; Table S2).

To optimize K, we set $N = 1,000$ and genotyped the 180 simulated samples by changing K from 20mer (unique in a mammalian genome) to 80mer (as the read length is 101bp). For DLA-88, the mean TPR ranges from 0.98 to 1 for allele number and 0.83 to 0.88 for allele expression when K changes from 20mer to 50mer, and decreases afterward (Figure 3C; Table S2). Variant linkage validation with paired-end reads increases expression TPRs and decreases FDRs, especially when K is small (Figure 3C; Table S2). Considering that both TPR and FDR are at the acceptable level (mean TPR $= 0.99$ and mean FDR $= 0.11$ for allele number, while mean TPR $= 0.88$ and mean FDR $= 0.19$ for allele expression; see Table S2), we chose 50bp as the optimal K for running KPR on these samples for DLA-88.

As the choice of K is clearly related to the sequence read length, we performed simulations by varying the read length from 50bp to 150bp. We genotyped 180 simulated samples at each read length with different K values (Figures S4A–S4C; Table S2). The results indicate that setting K to half of the read length generally yields better genotyping results (Figure S5; Table S2). The analysis also indicates that genotyping results improve with longer read length (Figure 3D; Table S2).

DLA-12 and DLA-64 genotyping are less effective, as mean TPR ranges from 0.2 to 0.5 for DLA-64 and mean FDR ranges from 0.17 to 0.42 for DLA-12 (Figures S3A and S3B; Table S2). However, $K = 50bp$ for read length of 101bp and $N > 100$ are still appropriate parameters to use, as they yield higher TPRs and lower FDRs compared to many other conditions (Figures S3A and S3B; Table S2).

We repeated the analysis with another RNA-seq dataset, consisting of 50–70 million 51 x 51bp read pairs per sample from blood specimens of 39 dogs, of which 21 dogs have immune-mediated hemolytic anemia and 18 dogs are unaffected.[37] All samples passed our QC measures (Figures S6A–S6E). We chose six samples that represent the entire cohort (Table S2) for simulation. The simulation analysis indicates that the optimized running parameters are $N \geq 1,000$, K being half of the read length, and $PD = 12$ for DLA-88 (Figure S7A), largely agreeing with those from the mammary dataset (Figure 3). However, for DLA-12 and DLA-64, the optimized N is much larger, with $N \geq 10,000$ for the blood dataset (Figure S7B) versus $N \geq 1,000$ for the mammary dataset (Figures S3A and S3B). No significant difference was found for the other parameters, as $K = 20$ (roughly half of the read length) works well for all three DLA-I genes and $PD = 38$ works well for DLA-12 (Figures S7A and S7B). Using the optimized parameters, we achieved a similar genotyping accuracy for DLA-88, but a significantly higher accuracy for DLA-64, for the blood samples compared to the mammary samples (Figures 3, S3, S4, and S7). One reason may be that the blood samples have significantly more DLA-I read pairs per sample, but approximately the same DLA-64/DLA-I ratios, compared to the

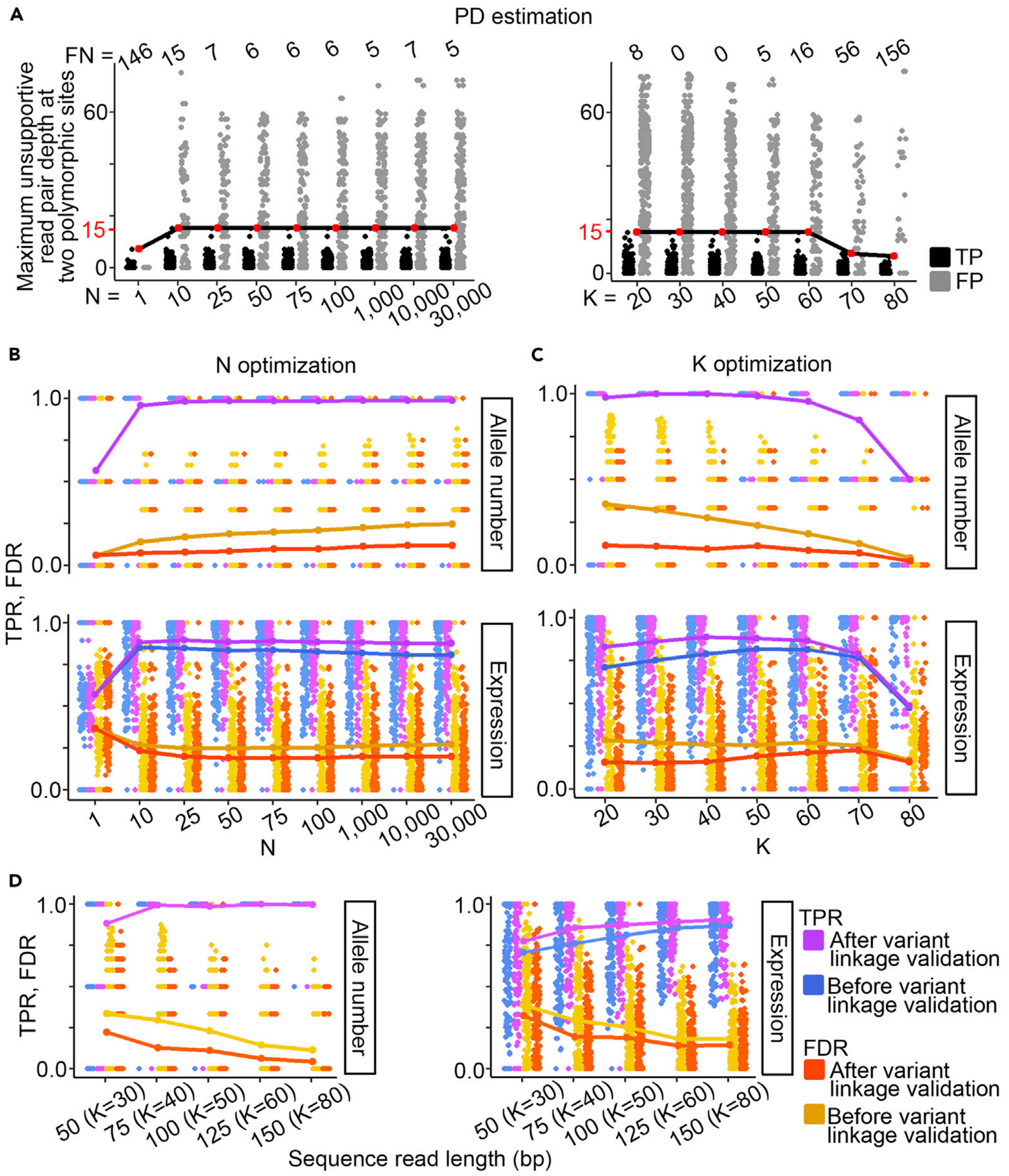


Figure 3. KPR running parameters are optimized for DLA-88 genotyping via simulation

(A) Optimization of two-polymorphic site read pair depth (PD). A total of 180 simulated samples (see STAR Methods) were genotyped by the KPR software with each value of the assembly runs (N) or the Kmer length (K) specified by the X axis. The Y axis indicates the maximum number of read pairs that do not support any variant linkage of known alleles in a simulated sample. Red dots in black line show the optimized PD at each N or K. The numbers at the top indicate the total false negatives (FN) in 180 samples. TP: true positives. FP: false positives.

(B and C) Optimization of N and K. Plotted are distributions of true positive rate (TPR) and false discovery rate (FDR) in each of the 180 simulated samples at a specified N (B) or K (C) value. The genotyping was done with N = 1,000 for B and K = 50 for C. TPR and FDR were calculated using allele numbers (top) or estimated expression levels (bottom) before and after paired-end validation. Each dot represents the TPR or FDR of a sample, while the line indicates the mean FDR or TPR of the 180 samples at each N or K value.

(D) K varies with sequence read length. A total 180 simulated samples were genotyped at each read length with the specified K value. Images are presented as B and C. See Figures S1–S8, and Table S2.

mammary samples (Figure S8A). Thus, a larger N is required for DLA-64 assembly, and the increased amount of DLA-64 read pairs, coupled with lower sequencing error rates (Figure S8B), improves the genotyping accuracy.

Other factors influence genotyping accuracy

Beside read length (Figure 3D), we wanted to know if the KPR genotyping accuracy is influenced by AC, sequencing error rate (E), DLA-I read pair amount (D), and DLA-I read distribution evenness along exons 2 and 3 (RD). We hence performed simulations with optimized parameters of $PD = 15$ for DLA-88 and $PD = 38$ for DLA-12, $K = 50bp$ (as read length is 101bp), and $N = 1,000$ for the mammary dataset (Figures 3, S3, and S4). Moreover, when evaluating one factor (AC, E, D, or RD), we kept all other factors optimized to reduce confounding.

To assess the effect of AC on genotyping, we used the 10 ACs made previously (Figure 3), and randomly sampled alleles from the known allele pool 30 times per AC. We then simulated 300 samples, using the most optimized E (0.3%), D (180,435 read pairs), and RD (0.45) values across the mammary cohort (Tables S2 and S3). The genotyping results reveal no significant difference in either TPR or FDR among the 10 ACs for DLA-88 (Figure 4A), as well as for DLA-64 (Figure S9; Table S3). For DLA-12, however, ACs of heterozygous alleles have significantly lower TPRs and higher FDRs, compared to ACs of homozygous alleles (Figure S9A; Table S3). Thus, the accuracy of KPR tools may drop when genotyping samples with heterozygous DLA-12 alleles. As such, we used only the 5 ACs with homozygous DLA-12 alleles to evaluate other factors later in discussion.

To assess the effect of E on genotyping, we kept D and RD optimized as stated above, and varied E from 0 to 1.6% ($0.3\% < E < 0.8\%$ for the cohort; see Figure S2A and Table S2). We simulated and genotyped 50 samples (10 samples per AC) with each E (350 samples in total). For DLA-88, the mean TPR is stabilized at 1.0 for allele number and 0.9 for allele expression, while the mean FDR decreases to 0.08 to 0.02 for allele number and 0.13 to 0.09 for allele expression, when $E < 0.4\%$ (Figure 4B). Interestingly, nearly the same conclusion is reached for DLA-12 and DLA-64 (Figures S9A and S9B).

To assess the effect of D on genotyping, we kept E and RD optimized, and varied D from 100 to 180,000 read pairs ($4,485 \leq D \leq 180,435$ for the cohort; see Figure S2B and Table S2). With these, we simulated and genotyped 50 samples per D (400 samples in total). For DLA-88, the mean TPR reaches 1.0 for allele number and 0.8 to 0.9 for allele expression when $D \geq 3,000$, and largely stabilizes afterward. Meanwhile, the mean FDR ranges from 0.05 to 0.13 for allele number and 0.11 to 0.2for allele expression (Figure 4C). The same mean TPR and FDR are reached when $D \geq 10,000$ for DLA-12 and $D \geq 30,000$ for DLA-64 (Figures S9A and S9B).

To assess the effect of RD on genotyping, we chose 5 RD values that represent the entire cohort (Table S2). We simulated and genotyped 50 samples per RD (250 samples in total) by keeping E and D optimized. The results reveal no significant difference in either TPR or FDR among different RDs (Figures 4D, S9A, and S9B; Table S3), indicating that RD may not influence genotyping when E and D are optimized. To test this possibility, we repeated the analysis by reducing D to 10,000 and 3,000 read pairs. Increasing RD then indeed improves the genotyping accuracy (Figures 4D, S9A, and S9B; Table S3).

Kmer-based paired-end read software outperforms other genotyping tools examined in typing new alleles

We compared the KPR software to well-benchmarked and recommended genotyping tools that are RNA-seq based, with or without assembly.[21,22] For such tools that perform assembly, one is HLAminer,[24] which conducts *de novo* assembly, like the KPR assembler. Another is Kourami,[25] which conducts the reference-based assembly but at the time of writing only works for human alleles, and hence, is not included in the comparison. For tools that do not perform assembly, most rely on read-mapping for genotyping (see Introduction), and we chose Seq2HLA[23] as the representative.

To compare the software tools, we performed two sets of simulation, one with known alleles and the other with new alleles, which are represented by newly assembled allele candidates described in the next section (Table S4). A total of 600 samples per set were simulated, using the same 6 samples shown in Figure 3 and considering allele type and combination. Note that neither HLAminer nor Seq2HLA reports allele expression levels. Thus, only TPR and FDR for allele number were used to evaluate the genotyping results of each software.

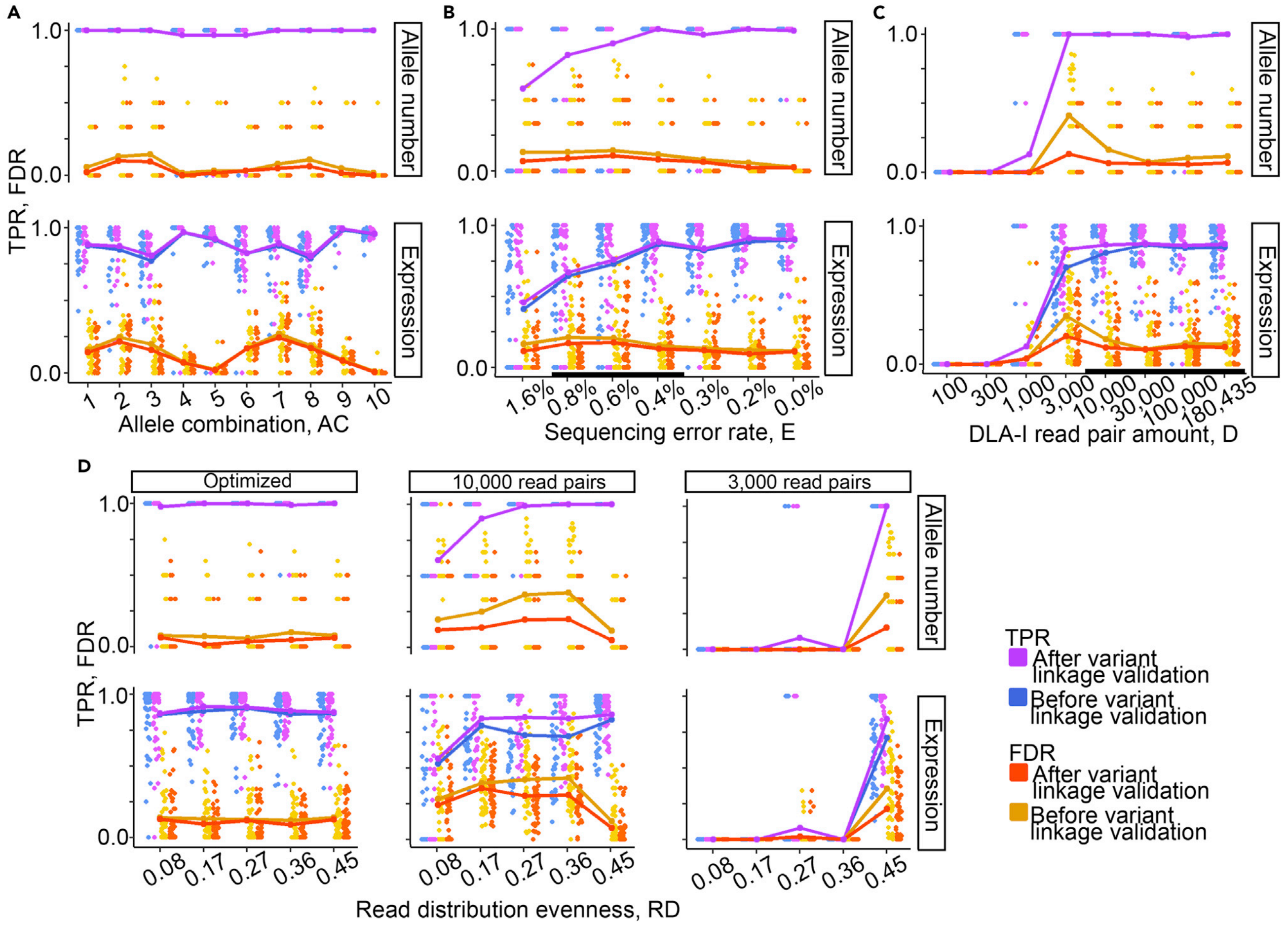


Figure 4. The influence of allele combination (AC), sequencing error rate (E), DLA-I read amount (D) and distribution (RD) on DLA-88 genotyping is evaluated via simulation

(A) AC evaluation. Genotyping results are shown for 300 samples, simulated with the 10 ACs (30 samples per AC) indicated and with E, D, and RD optimized (see STAR Methods). TPR and FDR are presented as in Figure 3.

(B) E evaluation. Genotyping results are shown for 350 samples, simulated with 7 E values (50 samples per E) indicated and with AC, D, and RD optimized (see STAR Methods). The thick bar of the X axis indicates the actual E range of the RNA-seq data of the cohort.

(C) D evaluation. Genotyping results are shown for 400 samples, simulated with the 8 D values (50 samples per D) indicated and with AC, E and RD optimized (see STAR Methods). The thick bar of the X axis indicates the actual D range for the RNA-seq data of the cohort.

(D) RD evaluation. Genotyping results are shown for 250 samples, simulated with the 5 RD values (50 samples per RD) indicated and with all AC, E and D optimized (left), or only AC and E optimized and D = 10, 000 (middle) or D = 3, 000 (right) (see STAR Methods). See Figure S9 and Tables S2 and S3.

In simulations with known alleles, KPR performs significantly better than the assembler of HLAminer (only the assembly results of HLAminer were used for this comparison). KPR is comparable to Seq2HLA in DLA-88 and DLA-12 genotyping. For DLA-88, the mean TPR is 0.99 for KPR, 0.50 for HLAminer, and 0.84 for Seq2HLA, while the corresponding mean FDR is 0.08, 0.07, and 0.36 respectively (Figure 5A; Table S4). For DLA-12, the mean TPR is 0.59 for KPR, 0.02 for HLAminer, and 0.57 for Seq2HLA, while the corresponding FDR is 0.20, 0.01, and 0.63 respectively (Figure 5A). However, for DLA-64, KPR performs worse than Seq2HLA, with a mean TPR of 0.23 versus 0.90, and a mean FDR of 0.02 versus 0.25 respectively (Figure 5A).

In simulations with new alleles, KPR performs significantly better than HLAminer and Seq2HLA (Figure 5B). As expected, Seq2HLA is unable to genotype new alleles, while KPR types new alleles with similar accuracy as known alleles (Figure 5; Table S4). KPR achieves a mean TPR of 0.99 and a mean FDR of 0.15 for DLA-88, compared to 0.42 and 0.02 for HLAminer (Figure 5B). KPR obtains a mean TPR of 0.62 and a mean FDR of 0.28 for DLA-12, also better than HLAminer (Figure 5B). For DLA-64, KPR and HLAminer are comparable, with 0.19 and 0.22 for TPR and 0.03 and 0.02 for FDR respectively (Figure 5B).

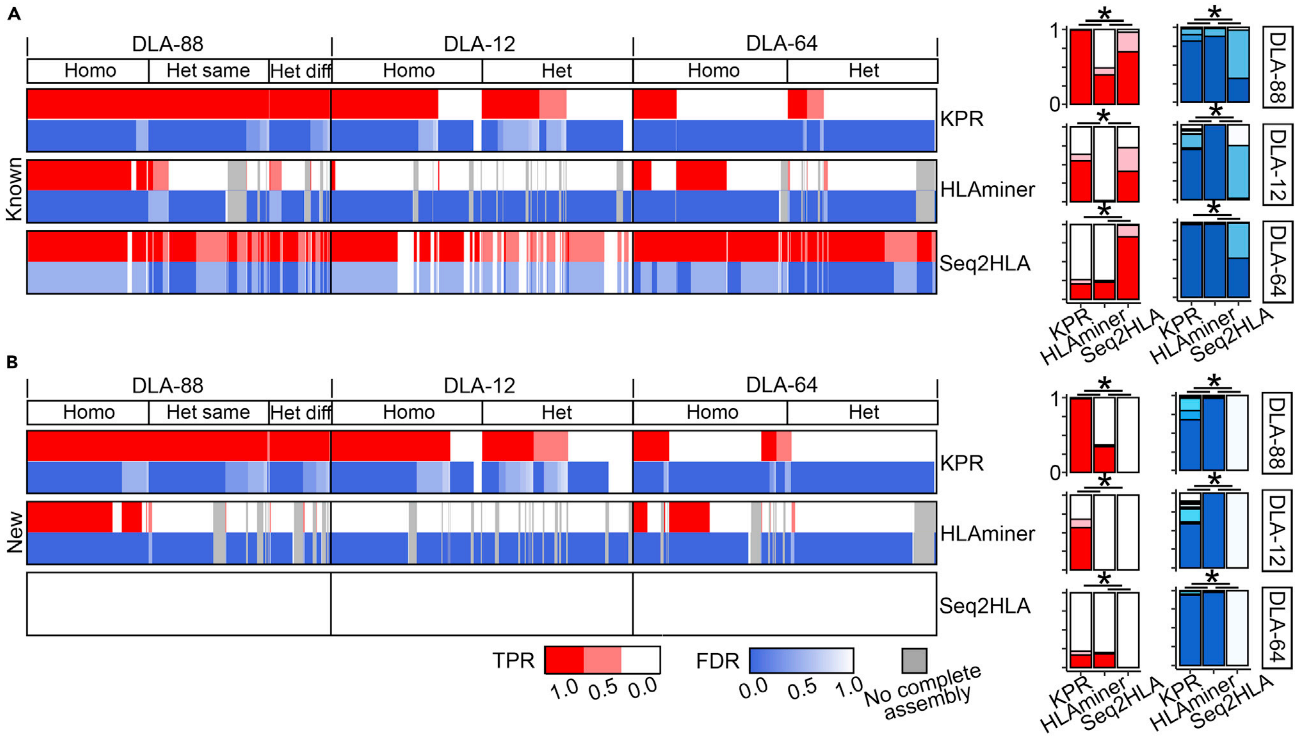


Figure 5. KPR software outperforms the HLAminer assembly tool and Seq2HLA in new allele typing

(A) Genotyping with known alleles. Left heatmaps indicate TPR (red) and FDR (blue) in each of the 600 simulated samples (see STAR Methods). Columns represent simulated samples, grouped based on the gene and AC, and then ordered by TPR and FDR. Right bar plots summarize the genotyping results of all samples, with * indicating a significant ($p < 0.05$) difference via Wilcoxon rank sum tests between the software tools compared. Homo: homozygous; Het: heterozygous; Het same: heterozygous with same type of alleles (DLA-88 or DLA-88L) only; Het Diff: with both DLA-88 and DLA-88L alleles.

(B) Genotyping with new alleles. Images are presented as described in A. See Table S4.

A cohort of 157 dogs was genotyped by Kmer-based paired-end read software

We used KPR to genotype all 157 mammary tumors and 64 matching normal samples from 157 dogs of the cohort[35,36] passed QC measures (Figures S1A–S1J; Table S2) and used in our simulations described above. We applied KPR to these 221 samples using the optimized parameters (Figure 3), by setting $K = 50\text{bp}, N = 1,000$, and $PD = 15$ for DLA-88 and $PD = 38$ for DLA-12. We excluded genotyping results of 48 misassemblies identified by KPR. For DLA-12, we only kept the results of three new allele candidates that are recurrent and highly expressed (Table S5) to reduce false positives, as our simulation indicates a lower genotyping accuracy for DLA-12 (Figure S3A). A total of 208 samples (94%) were then genotyped, including 146 tumor and 62 normal samples from 152 dogs, yielding 45 known alleles and 60 new allele candidates (Table S5). The remaining 13 samples failed genotyping, as no complete contigs of exons 2 and 3 were assembled. These 13 samples have a significantly lower amount of DLA-I read pairs (Figure S10A), a critical factor that affects the KPR genotyping efficiency based on our simulation analysis (Figures 4C and S9).

Most of the genotyping results are validated

We assessed the genotyping accuracy. Among 848 genotyping results, 730 (86%) are with 45 known alleles (Table S5), and should be accurate. The remaining are with 60 new allele candidates. Among the new allele candidates, 6 have their variant linkages fully validated with paired-end reads (Figure 1B), 24 are recurrently found among ≥2 samples, and 10 have an expression fraction of ≥50% within a sample (Table S5). These subsets of new allele candidates are less likely to be artifacts arising from sequencing or assembly errors; their genotyping results (84 total) hence have a higher probability to be accurate.

To further validate the genotyping results, we used the whole exome sequencing (WES) data that are also published for these dogs,[35,36] via two approaches. First, WES read pairs were mapped to the RNA-seq read-assembled contigs, which consist of 45 known and 60 new allele candidates. For each WES sample and each DLA-I gene, the top two contigs with the most WES reads concordantly and identically mapped were identified. If the WES sample and the two contigs are from the same dog, then validation is achieved. This analysis validates 726 (86%) of 848 total genotyping results, of which 684 are with 31 known alleles and 42 are with 20 new allele candidates (Figures S10B and S10C; Table S5). Second, we conducted assembly with WES reads for exon 2 and exon 3 separately. If either contig is ≥ 90 amino acid long (exon 2 encodes 90 amino acids; exon 3 encodes 92 or 93 amino acids) and completely matches an RNA-seq contig from the same dog, then the genotyping result is validated. This analysis validates 173 genotyping results, of which 162 are with 21 known alleles and 11 are with 8 new allele candidates (Figure S10C; Table S5).

Lastly, by manually examining 13 samples with >6 alleles genotyped, we identified 7 new allele candidates that are misassembled from specific allele combinations. For instance, when a sample carries both DLA-88*501:01 and DLA-12*001:01, chimeric assemblies DLA-12*14:01 and DLA-88*501:New3 are likely to be generated, due to a long (>270bp) identical region amid two polymorphic sites between DLA-88*501:01 and DLA-12*001:01 (Table S5). To alert the users of potential misassembly, the KPR output files have flagged any new alle candidates that share >270bp identical sequence with another allele (known or new) in the sample.

Combining all validation analyses described above, 742 of the 848 (88%) genotyping results are either validated or considered unlikely to be false. Moreover, the analyses yield 33 new allele candidates that are unlikely to be artifacts (referred to as putative new alleles hereafter), with 20 for DLA-88, 7 for DLA-88L, 2 for DLA-12, and 4 for DLA-64 (Table S5).

DLA-88 has the highest expression and diversity

We investigated the 742 validated genotyping results of 208 samples (146 tumor and 62 normal) from 152 dogs. Among the three DLA-I loci, the DLA-88 locus is the most expressed and diversified (Figures 6A–6C). A total of 28 known alleles and 20 putative new alleles of DLA-88 were identified in 199 samples, with a relative allele expression range of 3-100% and a median of 35% (Figures 6A and 6B; Table S5). The DLA-64 locus is the least diversified and expressed, with one known and four putative new alleles found in 97 samples and having a relative allele expression range of 0.5–33.3% and a median of 5% (Figures 6A and 6B; Table S5). The DLA-12/DLA-88L locus has intermediate allele diversity and expression levels. For DLA-12, a total of 12 known alleles and two putative new alleles were identified in 166 samples with a relative expression range of 0.5-100% and a median of 11% (Figures 6A and 6B; Table S5). For DLA-88L, four known alleles and 7 putative new alleles were genotyped in 45 samples, with a relative expression range of 2-59% and a median of 23% (Figures 6A and 6B; Table S5). These diversity and expression level differences are clearly seen in the cumulative relative allele expression level normalized within the population (Figures 6C and S11A). The observations are consistent with literature reports.[14–17]

As expected, many concordant alleles between tumor and normal samples were found for the 56 dogs with both samples genotyped. Specifically, among 207 genotyping results from the normal samples and 206 genotyping results from the tumor samples, 166 (81%) agree (Table S5). Our statistical analysis reveals no significant difference between tumor and normal samples in either the alleles or their expression levels ($0.6 \leq p \leq 1$; see Table S5). Individual dog-wise, 55 of 56 total animals have at least one allele, and 3 alleles on average, shared between the two samples of each animal, with 15 dogs having the agreement at 100% (Figure S11B; Table S5). The disagreement found in some of the dogs may indicate that their tumor and normal samples express the same alleles at different levels, due to different cell composition and/or epigenetic or other alterations in the tumor.[38,39] As a result, an allele might fail assembly and genotyping only in the tumor (or normal) sample due to a lower expression level.

Our study reveals breed-dominant alleles (Figure 6D), consistent with published work.[15,17] Specifically, we identified prevalent alleles in 4 breeds with a sample size of ≥10 dogs (Figures 6D and S11C). Examples include DLA-88*004:02 in Yorkshire Terrier, as well as DLA-88*014:01 and DLA-88*035:01 in Shih Tzu. The dominant alleles in Shih Tzu match published studies,[17] and differ more from those of the other three breeds, consistent with their places of origin.

We also genotyped the 39 dogs with blood RNA-seq data,[37] using the optimized parameters shown in Figure S7. The analysis yields the same major findings as the mammary study (Figures 6, S11A, and S12). First, DLA-88 is the most diversified and expressed, compared to DLA-12 and 64, among these 39 dogs

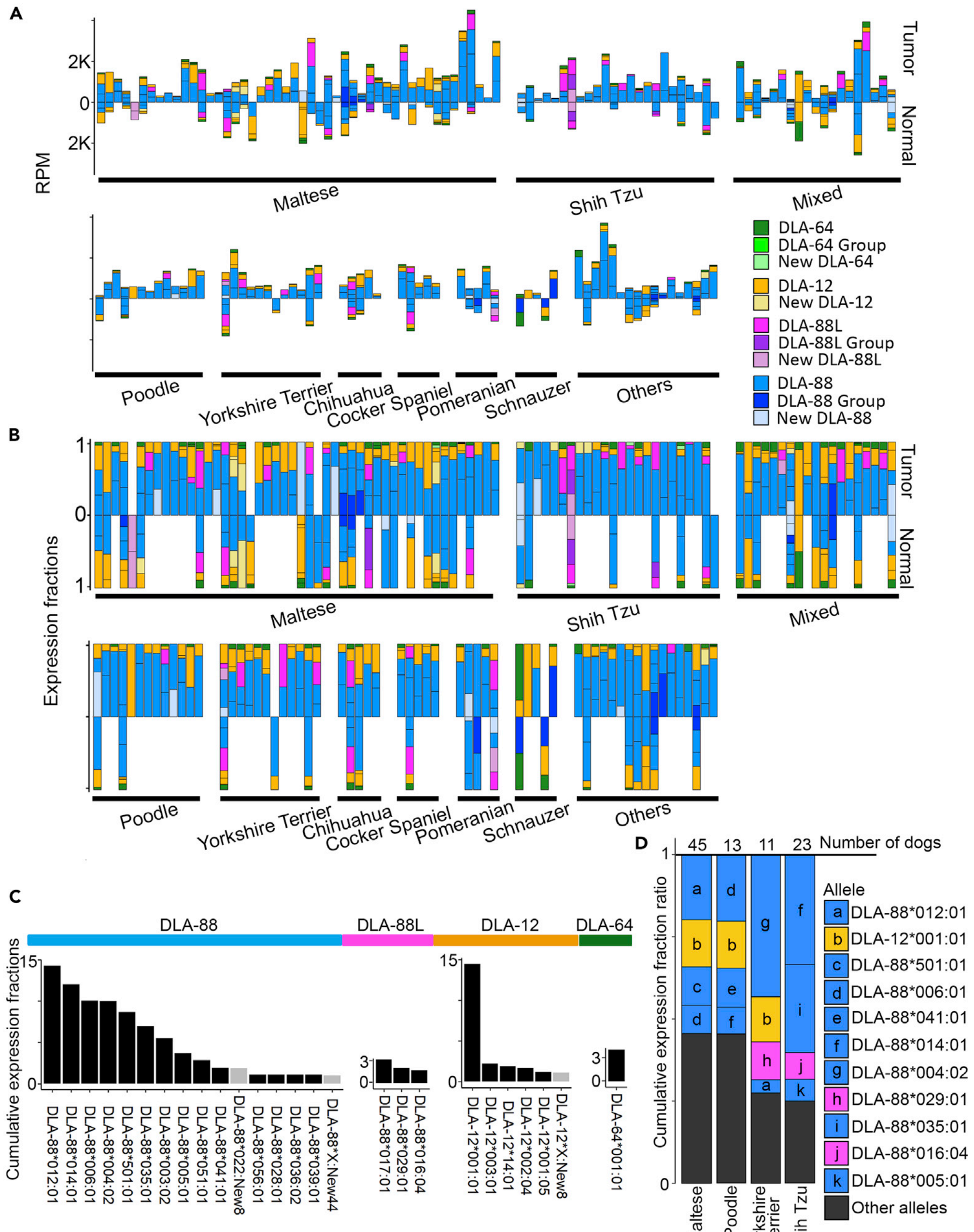


Figure 6. KPR tools genotyped 152 dogs from the largest canine RNA-seq study published so far[35,36]

(A and B) Genotyping results of the tumor (top) and normal (bottom) samples of the 152 dogs grouped by breeds. A dog is represented by single or paired vertical bars. The lines inside each bar separate individual alleles, with the height indicating the allele expression in reads per million (RPM) (A) or expression fractions within the animal (B). DLA-88/88L/12/64: known alleles. DLA-88/88L/12/64 group: new allele candidates with allele group assigned. DLA-88/88L/12/64 new: new allele candidates with no allele group assigned.

(C) Distribution of allele cumulative expression fractions within the 146 tumor samples. Alleles with cumulative expression fractions ≥1.0 are shown. Known alleles are shown as black bars, while new alleles are as gray bars.

(D) Breed-dominant alleles in the tumor samples of 4 pure breeds with each having ≥10 dogs. Top 4 alleles or alleles with cumulative expression fraction ratio reaching >50% within a breed are specified. See Figures S10–S12 and Table S5.

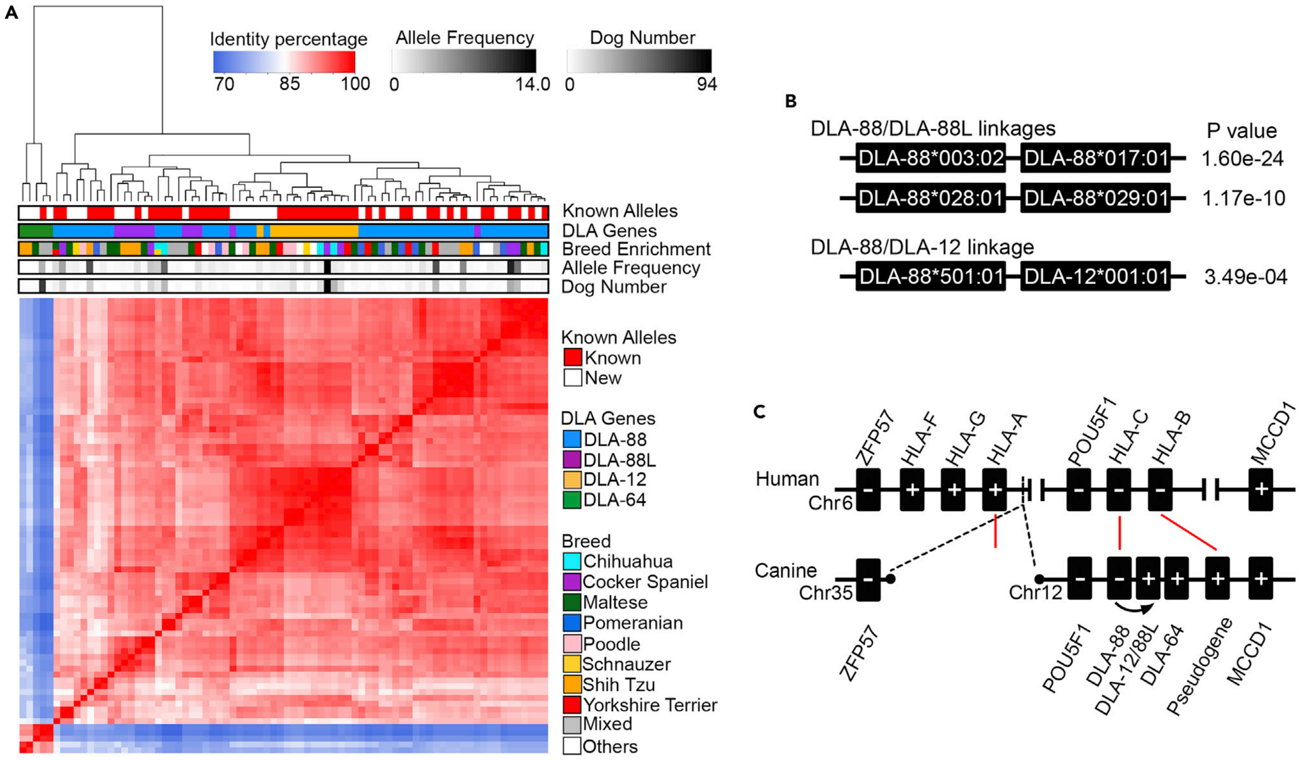


Figure 7. DLA-12 and DLA-88L alleles cluster with DLA-88 alleles

(A) The clustering of 45 known alleles and 33 putative new alleles identified in the 152 dogs, based on amino acid sequence identities. Breed enrichment is represented by the fraction of dogs within a breed that carry the allele. "Allele frequency" is the cumulated allele expression fraction, while "Dog number" specifies the number of dogs having the allele, within the cohort.

(B) Examples of DLA-88/DLA-88L linkages and a putative DLA-88/DLA-12 linkage indicated by allele co-occurrence identified by Fisher's exact tests, with p values shown. The two DLA-88/DLA-88L linkages shown have been reported previously.[17]

(C) Proposed evolution of the DLA-I genes. Red lines link the likely corresponding HLA-I and DLA-I genes. Dash lines indicate the breakage of the HLA-I locus in the canine genome. "-" and "+" specify the gene orientation. See Figures S13–S15 and Table S6.

(Figures S12A and S12B; Table S5). Furthermore, the most prevalent alleles among these dogs (Figures S12C and S12D) also match those in the mammary cohort (Figures 6C and S11D). DLA-88L alleles were found in Shih Tzu as expected, and also in Collie interestingly (Table S5).

DLA-12 and DLA-88L alleles cluster with DLA-88 alleles

To examine the allele diversity, we clustered the 45 known alleles and 33 putative new alleles genotyped in the cohort, using their encoded amino acid sequences (Table S5). As expected, the DLA-64 alleles are outliers (Figure 7A). Importantly, the DLA-88 alleles are clearly divided into two clusters (Figures 7A, S13, and S14). One cluster has fewer alleles and is more distant (Figure 7A).[40] The other cluster is much larger and can be further divided into 2 subclusters, which are named DLA-88/DLA-88L and DLA-88/DLA-12 because of the enrichment in DLA-88L and DLA-12 alleles respectively (Figures 7A and S6A–S6C; Table S6). The same conclusions are reached when including all 160 known alleles or just using the 160 known alleles in the analysis (Figures S13 and S14).

We observed possible "pioneers," which are the few highly prevalent alleles in each DLA-I cluster (Figure 7A). Examples are DLA-88*014:01 in the distant DLA-88 cluster, as well as DLA-88*012:01 and DLA-12*001:01 in the DLA-88/DLA-12 cluster (Table S6). This points out a potential "founder" effect in forming these allele clusters.

The results indicate that DLA-88L and DLA-12 are distinct allele groups. Within each group, DLA-88L alleles are more divergent, while DLA-12 alleles are more conserved (Figure 7A). Interestingly, certain DLA-12 and DLA-88L alleles are distributed differently among breeds. For example, DLA-12*001:01, found in 96 animals and representing the most prevalent allele, is significantly depleted in Shih Tzu. However, DLA-88*016:04 (previously known as DLA-88*L[17]), a DLA-88L allele, is found only in Shih Tzu (Figure S15A; Table S6). DLA-88L and DLA-12 are proposed to be both encoded by the DLA-12 locus.[15,17] Indeed, our analysis reveals significant co-occurrence of specific DLA-88 and DLA-88L allelic combinations (e.g., DLA-88*028:01 and DLA-88*029:01; DLA-88*003:02 and DLA-88*017:01), as well as of DLA-88 and DLA-12 alleles (e.g., DLA-88*501:01 and DLA-12*001:01) (Figures 7B and S15B; Table S6). These results are consistent with the reported DLA-88/DLA-88L and DLA-88/DLA-12 haplotypes.[17,18]

To better understand the relationship between DLA-88, -88L, and 12 alleles, we examined the dog-human synteny at the MHC-I loci with genomic sequences. The canine counterpart of HLA-A is possibly lost,[41] silenced or not assembled in the current CanFam3.1 genome after the ancestral region near the HLA-A locus broke in the canine genome, forming the end of canine chromosome 35 (Figure 7C). The canine orthologous site of HLA-B encodes only a pseudogene (Figure 7C); thus, the canine counterpart of HLA-B may be also destroyed. While the DLA-88 locus is an unambiguous ortholog of the HLA-C locus (Figure 7C), the DLA-12 and DLA-64 loci lack clear human orthologs. Because of the high sequence homology of DLA-12/88L to DLA-88 (Figure 7A), it is possible that the original DLA-12 locus was created via duplication of DLA-88 (Figure 7C).

DISCUSSION

Kmer-based paired-end read *de novo* assembler and genotyper

NGS-based methodology has become a focus for human HLA-I genotyping in recent years, and numerous software tools have been published.[21–30,42–46] Unfortunately, these tools, which are built upon the >23,694 human alleles, do not work for the dog, due to very limited alleles defined (185 total currently). The development of additional software tools, such as the KPR *de novo* assembler and genotyper reported here, could help to remedy this situation, enhancing the usage of the canine model.

Because our core algorithm relies on paired-end sequencing (note that paired-end reads are both derived from the same allele), the accuracy of our tools is not as heavily dependent on known alleles, as are existing methods.[21–30] Indeed, Sanger sequencing validation, simulation analysis, and concordance with published results[17] all indicate that our tools are effective. Our software outperforms popular published genotyping tools (with or without assembly) in typing new alleles.

Our KPR software however needs improvement. Factors including the amount of DLA-I reads (determined by the allele expression levels) influence the KPR genotyping accuracy. For lowly expressed alleles, the KPR assembler may be unable to assemble complete contigs that contain the entirety of exons 2 and 3, failing genotyping. We plan to test if the implementation of dynamic programming, such as the Floyd-Warshall algorithm, into the assembly process will yield improvement. We have successfully implemented dynamic programming in our copy number software SEG.[47] Second, KPR is able to estimate the allele expression level but the accuracy needs to be improved. KPR currently uses individual variant sites for expression estimation, and we plan to implement a haplotype-based strategy for improvement. Third, our current tools are designed for RNA-seq, but not for whole genome sequencing (WGS) and WES data,[42–46] which contain intronic sequences. Importantly, unlike RNA-seq, WGS and WES lack sequence reads that directly link exons 2 and 3. Thus, new functionality is needed for our tools to work for WGS and WES.

Our software is effective in DLA-88 genotyping, but less effective in DLA-12 and DLA-64 genotyping. DLA-12 and DLA-64 are expressed less abundantly, compared to DLA-88. Moreover, DLA-12 has a high sequence identity to DLA-88, making it harder to separate DLA-12 reads from DLA-88 reads. DLA-12 alleles are highly identical and two variant sites are often too distant to be validated by sequence read-pairs. All of these contribute to the lower genotyping accuracy of DLA-12 by our software. The DLA-64 genotyping issue is simpler. DLA-64 is more distant from DLA-88 and DLA-12, and its genotyping accuracy can be improved by increasing the sequencing amount. Lastly, while parameters optimized for one dataset work well for another dataset in our simulation study for DLA-88 genotyping, dataset-specific parameter optimization is required to achieve effective DLA-64 genotyping.

Our software is capable of genotyping new alleles with the same accuracy as known alleles. However, for the new allele candidates discovered by our software, experimental validation such as Sanger sequencing is still needed.

Our KPR software can be easily modified to genotype numerous other species that, like the dog, have very few alleles defined. Note that the KPR assembler is species-independent, and only the KPR genotyper needs modifications.

DLA-I allele diversity landscape

Consistent with published work,[14–17] our genotyping study of a cohort of 152 dogs[35] clearly indicates that the DLA-88 locus encodes a classical MHC-I gene, being highly diversified and expressed, while DLA-64 is the opposite. The DLA-12 locus is however less clear.

A previous study reports that the DLA-12 locus encodes DLA-12 alleles in 80% of dogs and DLA-88L alleles in 20% of dogs examined.[17] This is supported by our finding of significant co-occurrence of certain DLA-88 and DLA-88L alleles, as well as certain DLA-88 and DLA-12 alleles, in the same animals. Our sequence analysis reveals that DLA-12 and DLA-88L alleles are closer to some DLA-88 alleles, compared to among more distant DLA-88 alleles themselves. Unlike the DLA-88 locus, we have not found a human orthologue for the DLA-12 locus. These findings support the notion that the original DLA-12 locus may be duplicated from the DLA-88 locus. A recent study reports that DLA-88L arose via interlocus and intralocus gene conversion between DLA-88 and DLA-12.[18] As such, DLA-12 and DLA-88L alleles should perhaps be classified as subgroups of DLA-88, which further increases the DLA-88 diversity.

As DLA-88L is highly homologous to DLA-88 and new DLA-88L alleles are yet to be identified, it is possible that some DLA-88L alleles may be mistakenly classified as DLA-88. This may explain why >2 alleles of DLA-88 are genotyped in some animals by us and others.[15]

Kmer-based paired-end read software can lead to more biological insights

Over 1,000 dogs have publicly available paired-end RNA-seq data, and more dogs are on the way. Applying our KPR software to these rapidly accumulating data will assess the diversity landscape of DLA-I alleles, and discover dominant alleles within and across breeds. This knowledge can in turn be used to address important questions, including how the MHC-I genotype shapes the tumor mutational landscape. By investigating 9,176 human cancer patients, a team reports that recurrent mutations are more likely to locate in peptides that are poorly presented by the HLA-I molecules across patients, and that the HLA-I genotype is associated with the emergence of specific oncogenic mutations.[48] The authors conclude that the HLA-I genotype restricts the oncogenic mutational landscape and could be used in predicting personal cancer susceptibility.[48] Once a large number of cancerous and healthy dogs are genotyped, we can conduct similar research to determine if the DLA-I genotype also shapes the mutational landscape in dogs.[6,49] Importantly, investigating the association between breed-dominant alleles and prevalent oncogenic mutations may contribute to the understanding of the breed-predisposition to cancer.[50] For example, Maltese, Yorkshire terrier, and poodle are among breeds reported to be susceptible to mammary tumors.[50] Are dominate alleles in these predisposed breeds (Figure 6D) poorly presenting the prevalent oncogenic mutations detected in mammary tumors, such as PIK3CA H1047R,[6,49] compared to other breeds? These will deepen our understanding of tumorigenesis and cancer susceptibility.

Limitations of the study

First, our KPR software performs *de novo* assembly of the entire exon 2 and 3 sequences with RNA-seq data. It hence cannot genotype DLA-I alleles that are not expressed or expressed at a low level. Second, for new allele candidates discovered by our software, validation (e.g., Sanger sequencing) is needed. This is especially so for new allele candidates of DLA-12 due to the higher probability of misassembly. Third, the KPR software assembles and genotypes alleles individually, and hence cannot determine the allele haplotype. For samples where >2 DLA-88 alleles are genotyped, more studies are needed to determine if and which DLA-88 alleles are encoded by the DLA-12 locus. The confirmed alleles should be reclassified as DLA-88L alleles.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2023.105996.

ACKNOWLEDGMENTS

We thank Ms. Hannah Lorelle Cavanagh, Dr. Xinfu Zhang, and Ms. Jin Qian for conducting the Sanger sequencing and Dr. Tianfang Wang for conducting the IHC experiment; Mr. Joshua Watson and Mr. Maxwell Colonna for editing the article; the Georgia Advanced Computing Resource Center, University of Georgia (GACRC) for supporting this work; and the authors who published canine NGS data used in this work. This work is funded by NCI R01 CA252713 and R01 CA182093.

AUTHOR CONTRIBUTIONS

Conceptualization, Y.F. and S.Z.; methodology, Y.F. and S.Z.; investigation, Y.F. and S.Z.; writing, S.Z., Y.F., P.R.H., S.M.T., and W.H.H.; visualization, Y.F.; resources, Y.F. and S.Z.; supervision, S.Z. and S.M.T.; funding acquisition, S.Z. and W.H.H.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: August 5, 2022
Revised: November 28, 2022
Accepted: January 11, 2023
Published: February 17, 2023

REFERENCES

1. Somarelli, J.A., Boddy, A.M., Gardner, H.L., DeWitt, S.B., Tuohy, J., Megquier, K., Sheth, M.U., Hsu, S.D., Thorne, J.L., London, C.A., and Eward, W.C. (2020). Improving cancer drug discovery by studying cancer across the tree of Life. Mol. Biol. Evol. *37*, 11–17. https://doi.org/10.1093/molbev/msz254.

2. London, C.A., Acquaviva, J., Smith, D.L., Sequeira, M., Ogawa, L.S., Gardner, H.L., Bernabe, L.F., Bear, M.D., Bechtel, S.A., and Proia, D.A. (2018). Consecutive day hsp90 inhibitor administration improves efficacy in murine models of kit-driven malignancies and canine mast cell tumors. Clin. Cancer Res. *24*, 6396–6407. https://doi.org/10.1158/1078-0432.CCR-18-0703.

3. Regan, D.P., Chow, L., Das, S., Haines, L., Palmer, E., Kurihara, J.N., Coy, J.W., Mathias, A., Thamm, D.H., Gustafson, D.L., and Dow, S.W. (2022). Losartan blocks osteosarcoma-elicited monocyte recruitment, and combined with the kinase inhibitor toceranib, exerts significant clinical benefit in canine metastatic osteosarcoma. Clin. Cancer Res. *28*, 662–676. https://doi.org/10.1158/1078-0432.CCR-21-2105.

4. Boyko, A.R. (2011). The domestic dog: man's best friend in the genomic era. Genome Biol. *12*, 216. https://doi.org/10.1186/gb-2011-12-2-216.

5. Dow, S. (2019). A role for dogs in advancing cancer immunotherapy research. Front. Immunol. *10*, 2935. https://doi.org/10.3389/fimmu.2019.02935.

6. Alsaihati, B.A., Ho, K.L., Watson, J., Feng, Y., Wang, T., Dobbin, K.K., and Zhao, S. (2021). Canine tumor mutational burden is correlated with TP53 mutation across tumor types and breeds. Nat. Commun. *12*, 4670. https://doi.org/10.1038/s41467-021-24836-9.

7. Wang, J., Wang, T., Sun, Y., Feng, Y., Kisseberth, W.C., Henry, C.J., Mok, I., Lana, S.E., Dobbin, K., Northrup, N., et al. (2018). Proliferative and invasive colorectal tumors in pet dogs provide unique insights into human colorectal cancer. Cancers *10*, 330. https://doi.org/10.3390/cancers10090330.

8. Wang, J., Wang, T., Bishop, M.A., Edwards, J.F., Yin, H., Dalton, S., Bryan, L.K., and Zhao, S. (2018). Collaborating genomic, transcriptomic and microbiomic alterations lead to canine extreme intestinal polyposis. Oncotarget *9*, 29162–29179. https://doi.org/10.18632/oncotarget.25646.

9. Liu, D., Xiong, H., Ellis, A.E., Northrup, N.C., Dobbin, K.K., Shin, D.M., and Zhao, S. (2015). Canine spontaneous head and neck squamous cell carcinomas represent their human counterparts at the molecular level. PLoS Genet. *11*, e1005277. https://doi.org/10.1371/journal.pgen.1005277.

10. Liu, D., Xiong, H., Ellis, A.E., Northrup, N.C., Rodriguez, C.O., Jr., O'Regan, R.M., Dalton, S., and Zhao, S. (2014). Molecular homology and difference between spontaneous canine mammary cancer and human breast cancer. Cancer Res. https://doi.org/10.1158/0008-5472.

11. Tang, J., Li, Y., Lyon, K., Camps, J., Dalton, S., Ried, T., and Zhao, S. (2014). Cancer driver-passenger distinction via sporadic human and dog cancer comparison: a proof-of-principle study with colorectal cancer. Oncogene *33*, 814–822. https://doi.org/10.1038/onc.2013.17.

12. Tang, J., Le, S., Sun, L., Yan, X., Zhang, M., Macleod, J., Leroy, B., Northrup, N., Ellis, A., Yeatman, T.J., et al. (2010). Copy number abnormalities in sporadic canine colorectal cancers. Genome Res. *20*, 341–350. https://doi.org/10.1101/gr.092726.109.

13. Wong, K., van der Weyden, L., Schott, C.R., Foote, A., Constantino-Casas, F., Smith, S., Dobson, J.M., Murchison, E.P., Wu, H., Yeh, I., et al. (2019). Cross-species genomic landscape comparison of human mucosal melanoma with canine oral and equine melanoma. Nat. Commun. *10*, 353. https://doi.org/10.1038/s41467-018-08081-1.

14. Graumann, M.B., DeRose, S.A., Ostrander, E.A., and Storb, R. (1998). Polymorphism analysis of four canine MHC class I genes. Tissue Antigens *51*, 374–381.

15. Ross, P., Buntzman, A.S., Vincent, B.G., Grover, E.N., Gojanovich, G.S., Collins, E.J., Frelinger, J.A., and Hess, P.R. (2012). Allelic diversity at the DLA-88 locus in golden retriever and boxer breeds is limited. Tissue Antigens *80*, 175–183. https://doi.org/10.1111/j.1399-0039.2012.01889.x.

16. Kennedy, L.J., Angles, J.M., Barnes, A., Carter, S.D., Francino, O., Gerlach, J.A., Happ, G.M., Ollier, W.E., Thomson, W., and Wagner, J.L. (2001). Nomenclature for factors of the dog major histocompatibility system (DLA), 2000: second report of the ISAG DLA nomenclature committee. Tissue Antigens *58*, 55–70.

17. Miyamae, J., Suzuki, S., Katakura, F., Uno, S., Tanaka, M., Okano, M., Matsumoto, T., Kulski, J.K., Moritomo, T., and Shiina, T. (2018). Identification of novel polymorphisms and two distinct haplotype structures in dog leukocyte antigen class I genes: DLA-88, DLA-12 and DLA-64. Immunogenetics *70*, 237–255. https://doi.org/10.1007/s00251-017-1031-5.

18. Miyamae, J., Okano, M., Nishiya, K., Katakura, F., Kulski, J.K., Moritomo, T., and Shiina, T. (2022). Haplotype structures and polymorphisms of dog leukocyte antigen (DLA) class I loci shaped by intralocus and interlocus recombination events. Immunogenetics *74*, 245–259. https://doi.org/10.1007/s00251-021-01234-5.

19. Chang, T.C., Carter, R.A., Li, Y., Li, Y., Wang, H., Edmonson, M.N., Chen, X., Arnold, P., Geiger, T.L., Wu, G., et al. (2017). The neoepitope landscape in pediatric cancers. Genome Med. *9*, 78. https://doi.org/10.1186/s13073-017-0468-3.

20. Kroemer, G., Galassi, C., Zitvogel, L., and Galluzzi, L. (2022). Immunogenic cell stress and death. Nat. Immunol. *23*, 487–500. https://doi.org/10.1038/s41590-022-01132-2.

21. Richters, M.M., Xia, H., Campbell, K.M., Gillanders, W.E., Griffith, O.L., and Griffith, M. (2019). Best practices for bioinformatic characterization of neoantigens for clinical utility. Genome Med. *11*, 56. https://doi.org/10.1186/s13073-019-0666-2.

22. Bauer, D.C., Zadoorian, A., Wilson, L.O.W.; Melbourne Genomics Health Alliance, and Thorne, N.P. (2018). Evaluation of computational programs to predict HLA genotypes from genomic sequencing data. Brief. Bioinform. *19*, 179–187. https://doi.org/10.1093/bib/bbw097.

23. Boegel, S., Löwer, M., Schäfer, M., Bukur, T., de Graaf, J., Boisguérin, V., Türeci, O., Diken, M., Castle, J.C., and Sahin, U. (2012). HLA typing from RNA-Seq sequence reads. Genome Med. *4*, 102. https://doi.org/10.1186/gm403.

24. Warren, R.L., Choe, G., Freeman, D.J., Castellarin, M., Munro, S., Moore, R., and Holt, R.A. (2012). Derivation of HLA types from shotgun sequence datasets. Genome Med. *4*, 95. https://doi.org/10.1186/gm396.

25. Lee, H., and Kingsford, C. (2018). Kourami: graph-guided assembly for novel human leukocyte antigen allele discovery. Genome Biol. *19*, 16. https://doi.org/10.1186/s13059-018-1388-2.

26. Buchkovich, M.L., Brown, C.C., Robasky, K., Chai, S., Westfall, S., Vincent, B.G., Weimer, E.T., and Powers, J.G. (2017). HLAProfiler utilizes k-mer profiles to improve HLA calling accuracy for rare and common alleles in RNA-seq data. Genome Med. *9*, 86. https://doi.org/10.1186/s13073-017-0473-6.

27. Kim, H.J., and Pourmand, N. (2013). HLA haplotyping from RNA-seq data using hierarchical read weighting. PLoS One *8*, e67885. https://doi.org/10.1371/journal.pone.0067885.

28. Orenbuch, R., Filip, I., Comito, D., Shaman, J., Pe'er, I., and Rabadan, R. (2020). arcasHLA: high-resolution HLA typing from RNAseq. Bioinformatics *36*, 33–40. https://doi.org/10.1093/bioinformatics/btz474.

29. Szolek, A., Schubert, B., Mohr, C., Sturm, M., Feldhahn, M., and Kohlbacher, O. (2014). OptiType: precision HLA typing from next-generation sequencing data. Bioinformatics *30*, 3310–3316. https://doi.org/10.1093/bioinformatics/btu548.

30. Bai, Y., Ni, M., Cooper, B., Wei, Y., and Fury, W. (2014). Inference of high resolution HLA types using genome-wide RNA or DNA sequencing reads. BMC Genom. *15*, 325. https://doi.org/10.1186/1471-2164-15-325.

31. Robinson, J., Halliwell, J.A., Hayhurst, J.D., Flicek, P., Parham, P., and Marsh, S.G.E. (2015). The IPD and IMGT/HLA database: allele variant databases. Nucleic Acids Res. *43*, D423–D431. https://doi.org/10.1093/nar/gku1161.

32. Tian, L., Li, Y., Edmonson, M.N., Zhou, X., Newman, S., McLeod, C., Thrasher, A., Liu, Y., Tang, B., Rusch, M.C., et al. (2020). CICERO: a versatile method for detecting complex and diverse driver fusions using cancer RNA sequencing data. Genome Biol. *21*, 126. https://doi.org/10.1186/s13059-020-02043-x.

33. Wang, T., Kwon, S.H., Peng, X., Urdy, S., Lu, Z., Schmitz, R.J., Dalton, S., Mostov, K.E., and Zhao, S. (2020). A Qualitative change in the transcriptome occurs after the first cell cycle and coincides with lumen establishment during MDCKII Cystogenesis. iScience *23*, 101629. https://doi.org/10.1016/j.isci.2020.101629.

34. Kennedy, L.J., Altet, L., Angles, J.M., Barnes, A., Carter, S.D., Francino, O., Gerlach, J.A., Happ, G.M., Ollier, W.E., Polvi, A., et al. (1999). Nomenclature for factors of the dog major histocompatibility system (DLA), 1998. First report of the ISAG DLA Nomenclature Committee. International Society for Animals Genetics. Tissue Antigens *54*, 312–321. https://doi.org/10.1034/j.1399-0039.1999.540319.x.

35. Kim, T.M., Yang, I.S., Seung, B.J., Lee, S., Kim, D., Ha, Y.J., Seo, M.K., Kim, K.K., Kim, H.S., Cheong, J.H., et al. (2020). Cross-species oncogenic signatures of breast cancer in canine mammary tumors. Nat. Commun. *11*, 3616. https://doi.org/10.1038/s41467-020-17458-0.

36. Kim, K.K., Seung, B.J., Kim, D., Park, H.M., Lee, S., Song, D.W., Lee, G., Cheong, J.H., Nam, H., Sur, J.H., and Kim, S. (2019). Whole-exome and whole-transcriptome sequencing of canine mammary gland tumors. Sci. Data *6*, 147. https://doi.org/10.1038/s41597-019-0149-8.

37. Borchert, C., Herman, A., Roth, M., Brooks, A.C., and Friedenberg, S.G. (2020). RNA sequencing of whole blood in dogs with primary immune-mediated hemolytic anemia (IMHA) reveals novel insights into disease pathogenesis. PLoS One *15*, e0240975. https://doi.org/10.1371/journal.pone.0240975.

38. Campoli, M., and Ferrone, S. (2008). HLA antigen changes in malignant cells: epigenetic mechanisms and biologic significance. Oncogene *27*, 5869–5885. https://doi.org/10.1038/onc.2008.273.

39. Garcia-Lora, A., Algarra, I., and Garrido, F. (2003). MHC class I antigens, immune surveillance, and tumor immune escape.

J. Cell. Physiol. *195*, 346–355. https://doi.org/10.1002/jcp.10290.

40. Parker, H.G., Dreger, D.L., Rimbault, M., Davis, B.W., Mullen, A.B., Carpintero-Ramirez, G., and Ostrander, E.A. (2017). Genomic analyses reveal the influence of geographic origin, migration, and hybridization on modern dog breed development. Cell Rep. *19*, 697–708. https://doi.org/10.1016/j.celrep.2017.03.079.

41. Yuhki, N., Beck, T., Stephens, R., Neelam, B., and O'Brien, S.J. (2007). Comparative genomic structure of human, dog, and cat MHC: HLA, DLA, and FLA. J. Hered. *98*, 390–399. https://doi.org/10.1093/jhered/esm056.

42. Shukla, S.A., Rooney, M.S., Rajasagi, M., Tiao, G., Dixon, P.M., Lawrence, M.S., Stevens, J., Lane, W.J., Dellagatta, J.L., Steelman, S., et al. (2015). Comprehensive analysis of cancer-associated somatic mutations in class I HLA genes. Nat. Biotechnol. *33*, 1152–1158. https://doi.org/10.1038/nbt.3344.

43. Xie, C., Yeo, Z.X., Wong, M., Piper, J., Long, T., Kirkness, E.F., Biggs, W.H., Bloom, K., Spellman, S., Vierra-Green, C., et al. (2017). Fast and accurate HLA typing from short-read next-generation sequence data with xHLA. Proc. Natl. Acad. Sci. USA *114*, 8059–8064. https://doi.org/10.1073/pnas.1707945114.

44. Kiyotani, K., Mai, T.H., and Nakamura, Y. (2017). Comparison of exome-based HLA class I genotyping tools: identification of platform-specific genotyping errors. J. Hum. Genet. *62*, 397–405. https://doi.org/10.1038/jhg.2016.141.

45. Huang, Y., Yang, J., Ying, D., Zhang, Y., Shotelersuk, V., Hirankarn, N., Sham, P.C., Lau, Y.L., and Yang, W. (2015). HLAreporter: a tool for HLA typing from next generation sequencing data. Genome Med. *7*, 25. https://doi.org/10.1186/s13073-015-0145-3.

46. Liu, C., Yang, X., Duffy, B., Mohanakumar, T., Mitra, R.D., Zody, M.C., and Pfeifer, J.D. (2013). ATHLATES: accurate typing of human leukocyte antigen through exome sequencing. Nucleic Acids Res. *41*, e142. https://doi.org/10.1093/nar/gkt481.

47. Zhang, M., Liu, D., Tang, J., Feng, Y., Wang, T., Dobbin, K.K., Schliekelman, P., and Zhao, S. (2018). Seg - a software Program for finding somatic copy number alterations in whole genome sequencing data of cancer. Comput. Struct. Biotechnol. J. *16*, 335–341.

48. Marty, R., Kaabinejadian, S., Rossell, D., Slifker, M.J., van de Haar, J., Engin, H.B., de Prisco, N., Ideker, T., Hildebrand, W.H., Font-Burgada, J., and Carter, H. (2017). MHC-I genotype restricts the oncogenic mutational landscape. Cell *171*, 1272–1283.e15. https://doi.org/10.1016/j.cell.2017.09.050.

49. Rodrigues, L., Watson, J., Feng, Y., Lewis, B., Harvey, G., Post, G., Megquier, K., Lambert, L., Miller, A., Lopes, C., and Zhao, S. (2021). Shared hotspot mutations in spontaneously arising cancers position dog as an unparalleled comparative model for precision therapeutics. Preprint at bioRxiv. https://doi.org/10.1101/2021.10.22.465469.

50. Dobson, J.M. (2013). Breed-predispositions to cancer in pedigree dogs. ISRN Vet. Sci. *2013*, 941275. https://doi.org/10.1155/2013/941275.

51. Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120. https://doi.org/10.1093/bioinformatics/btu170.

52. Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics *26*, 589–595. https://doi.org/10.1093/bioinformatics/btp698.

53. Maccari, G., Robinson, J., Ballingall, K., Guethlein, L.A., Grimholt, U., Kaufman, J., Ho, C.S., de Groot, N.G., Flicek, P., Bontrop, R.E., et al. (2017). IPD-MHC 2.0: an improved inter-species database for the study of the major histocompatibility complex. Nucleic Acids Res. *45*, D860–D864. https://doi.org/10.1093/nar/gkw1050.

54. Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods *17*, 261–272. https://doi.org/10.1038/s41592-019-0686-2.

55. Ewels, P., Magnusson, M., Lundin, S., and Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics *32*, 3047–3048. https://doi.org/10.1093/bioinformatics/btw354.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Dog monoclonal anti-MHC-I DG-H58A	Monoclonal Antibody Center; Washington State Univ.	Cat# DG-BOV2001; RRID: AB_2722788
Alexa fluor® 594 Goat anti-mouse IgG	Jackson ImmunoResearch	Code: 115-585-006; RRID: AB_2338872
Biological samples		
Canine tumor samples	Collected from veterinary hospitals with owner informed consent.	N/A
MDCKII, RNA samples	Wang, et al. 2020[33]	https://pubmed.ncbi.nlm.nih.gov/33089114/
Chemicals, peptides, and recombinant proteins		
Fetal Bovine Serum	Atlanta Biological	Cat# S11550
Critical commercial assays		
AllPrep DNA/RNAMini Kit	Qiagen	Cat# 80204
iScript cDNA Synthesis Kit	Bio-Rad	Cat# 1708880
Phusion High-Fidelity DNA Polymerase	NEB	Cat# M0530S
Deposited data		
Canine mammary tumor RNA-seq data	The SRA database	https://www.ncbi.nlm.nih.gov/sra/?term=PRJNA489087
Canine mammary tumor whole exome sequencing (WES) data	The SRA database	https://www.ncbi.nlm.nih.gov/bioproject/PRJNA489159/
Canine whole blood RNA-seq data	The SRA database	https://www.ncbi.nlm.nih.gov/bioproject?term=PRJNA629466
Other canine RNA-seq data (from our lab)	The SRA database Wang et al., 2018[7,8] Wang et al., 2020[33]	https://www.ncbi.nlm.nih.gov/bioproject?term=PRJNA418842 https://www.ncbi.nlm.nih.gov/bioproject/PRJNA532904/
Canine MHC-I reference alleles, with official allele names	The IPD-MHC database	https://www.ebi.ac.uk/ipd/mhc/group/DLA/
Canine MHC-I reference alleles, with no official allele names	GenBank	LC130502.1-LC130527.1; LC171419,1-LC171438.1; KX583613.1, KX583617.1, KX583623.1, KX583624.1, KX583625.1, KX583628.1
Canine reference genome, CanFam3.1		https://www.ncbi.nlm.nih.gov/data-hub/genome/GCF_000002285.3/
Oligonucleotides		
Primers for DLA-88 exons 2 and 3 cloning CGGAGATGGAGGTGGTGA” (forward) “GGTGGCGGGTCACACG” (reverse)	This paper	N/A
Software and algorithms		
KPR *de novo* assembler and genotyper	This paper	https://github.com/ZhaoS-Lab/KPR.git
BWA 0.7.17		http://bio-bwa.sourceforge.net/
SAMtools 1.9		http://samtools.sourceforge.net/
Python 2.7.14		https://www.python.org/downloads/release/python-2714/
Trimmomatic 0.39	Bolger et al., 2014[51]	https://github.com/usadellab/Trimmomatic
Clustal-Omega 1.2.4		http://www.clustal.org/omega/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Shaying Zhao (szhao@uga.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- Data: All data used in the study, including RNA-seq data, are publicly available, as listed in the key resources table.
- Code: The software package of KPR *de novo* assembler and genotyper is available at Github at https://github.com/ZhaoS-Lab/KPR.git. It is written in Python (version 2.7) and runs on the Unix/Linux platform. Other software tools including SAMtools (version 1.9), Trimmomatic (version 0.39), BWA (version 0.7.17), and Clustal-Omega (version 1.2.4) are needed to execute the pipeline. These tools are publicly available as listed in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Canine cDNA samples

Canine cDNA samples used in this study (listed in Table S1) are described in our previous publications[7,8,10,33]. Briefly, one cDNA sample was extracted from cells of Madin-Darby canine kidney II (MDCKII),[33] a widely used cell line in research.[33] Three cDNA samples (T76, 414642T, and 435726T) were extracted from mammary tumors of adult female dog patients,[10] with one being a 14-year-old Staffordshire Terrier (T76) and the other two with exact age and breed unknown. The remaining three cDNA samples (N14-77-T5-E1, 2577T, and LilyT) were extracted from intestinal tumors of three canine patients,[7,8] including a 9-year-old neutered male Golden Retriever-mix (N14-77-T5-E1), an 11-year-old intact male English Cocker Spaniel (2577T), and a 9-year-old spayed female English Cocker Spaniel (LilyT). Tumors samples were collected from client-owned dogs that develop the disease spontaneously, under the guidelines of the Institutional Animal Care and Use Committee for use of residual diagnostic specimens and with owner informed consent. The research received the ethical approval from the Institutional Animal Care and Use Committee of the University of Georgia.

METHOD DETAILS

Kmer-based paired-end read (KPR) *de novo* assembler and genotyper

The KPR assembler assembles DLA-I alleles *de novo*, as illustrated Figure 1A and below.

- Step 1: DLA-I read-pairs are identified by first mapping paired-end RNA-seq reads of a dog with BWA[52] (version 0.7.17), after cleaning with Trimmomatic[51] (version 0.39), to the DLA-I reference, which consists of 185 known DLA-I alleles with official allele names[16] and 11 alleles derived from GenBank[16,17,34,53] with no official allele names assigned. Then, concordantly mapped pairs with at least one read placed to exon 2 or 3 of DLA-I genes are identified, and further selected for those with each base quality at $\geq$30 to reduce sequencing error rate to $\leq$0.1%.
- Step 2: The K*mer* dictionary is built from the forward (or reverse) read of each pair selected above at an exhaustive 1bp sliding window. Low frequency Kmers are identified, and read-pairs harboring any of them are removed to further reduce sequencing errors.
- Step 3: Read pair extension starts with a randomly chosen pair with forward read *F* and reverse read *R*. It is extended by another pair with forward read f_i and reverse read r_i, if *F* and *R* share identical sequence of >K in length with f_i and r_i, respectively, at their 3′ or 5′-end (Figure 1A). This yields $F_{contig,}$ and R_{contig}, which are then looped through the same extension process until no further extension is possible.

- Step 4: Final contig is assembled from the last $F_{contig,}$ and R_{contig} from step 3, if they share identical sequence of >K at their 5′ or 3′-end.
- Step 5: Steps 3 and 4 are repeated *N* (e.g., 1,000) times such that all extension paths have been exhausted.

The KPR genotyper genotypes assembled contigs, as illustrated in Figure 1B and below

- Contig sorting and variant linkage building: After removing duplicates, assembled contigs are classified as DLA-64 contigs and non-DLA64 contigs. All full length (containing the entire sequence of exons 2 and 3) DLA-64 contigs are subjected to "DLA-I genotyping" described below. All non-DLA-64 contigs are subjected to multiple sequence alignment for consensus sequence building to identify variable sites and variant linkage groups.
- Variant linkage validation and frequency estimation with paired-end reads: To reduce false results from misassembly, variant linkage groups of non-DLA64 contigs are built via multiple sequence alignment of the contigs, and validated with paired-end reads after mapping the DLA-I read-pairs (see step 1 in the assembly section) to the contigs. If a linkage is also found in one or more RNA-seq read pairs mapped to the contig, then the linkage is validated (as the two sequences of a read pair are from the same allele). The frequency of each linkage group is estimated with read-pair coverage at its unique variant sites.
- Contig extension (if not completely covering exons 2 and 3): MHC-I genotyping requires the entire sequence of exons 2 and 3. Incomplete contigs are first extended by other incomplete contigs via overlapping variant linkage (≥75% *overlap*) or paired-end reads (Figure 1B). If still incomplete, these contigs are further extended with paired-end reads based on variants or frequency (Figure 1B).
- DLA-I genotyping: Each complete contig (with entire exons 2 and 3) is translated into amino acid sequence, and compared to the known DLA-I reference for gene and allele identification. Allele groups are assigned by examining the sequences from three HVRs, following the scheme by the International Society for Animal Genetics DLA Nomenclature Committee,[16,34] including 3-digit typing for allele group (e.g., DLA-88*034) and 5-digit typing for allele (e.g., DLA-88*034:01).
- New allele candidate naming: If matching to a known allele group based on hypervariable regions (HVRs),[16] the allele will be named by assigning to that allele group (3-digit typing, e.g., DLA-88*034), followed by ":New" and an integer (5-digit typing, e.g., DLA-88*001:New1). If not matching to any known allele group, we will first assign the allele to a specific DLA-I gene (DLA-88, 88L, 12 or 64) via sequence clustering with known alleles. Then, we will name the alleles as described above, except for using "X" to represent the allele group, e.g., DLA-88*X:New2.

Sanger sequencing

Sanger sequencing was performed following published procedures.[15,17] Briefly, exons 2 and 3 of DLA-88 were amplified from cDNA samples of canine cells and tissues, using primers "CGGAGATGGAGG TGGTGA" (forward) and "GGTGGCGGGTCACACG" (reverse) (Table S1). Then, the amplified PCR products were cloned into a vector and 5–10 clones per sample were subjected to sanger sequencing using the two primers specified above.

Immunohistochemistry (IHC) and confocal imaging

IHC with frozen sections of canine tissue samples and confocal imaging were performed following published procedures.[7–10] Briefly, H58A, a pan DLA-I monoclonal antibody purchased from the Washington State University (WSU) Monoclonal Antibody Center, was used the primary antibody. Alexa Fluor594–conjugated secondary antibody was purchased from Jackson ImmunoResearch. Images were taken with a Zeiss LSM 710 confocal microscope.

Sample simulation

Six paired tumor and normal RNA-seq samples of 3 dogs from a published study[35,36] were used for the simulation. Their identifiers in the sequence read archive (SRA) database are: 1) SRR7779554 and SRR7779476 from dog CMT-162; 2) SRR7779670 and SRR7779671 from dog CMT-785; and 3) SRR7779469 and SRR7779468 from dog CMT-149 (Table S2). Allele simulation was performed by random sampling of the DLA-I database. A total of 10 allele combinations (AC) were made to simulate scenarios including homozygosity versus heterozygosity, and DLA-88 canonical versus DLA-88*50X alleles. Allele sampling was repeated 3 times for each AC (Table S2). For PD, K, and N optimization simulation (Figure 3), a total of $6\ (\textit{samples}) \times 10\ (\textit{ACs}) \times 3\ (\textit{allele sampling repeats}) = 180\ \textit{samples}$ were simulated.

PD is defined as the maximum read pairs that are mapped to two polymorphic sites among true positive alleles of a sample when none of the read pairs supports the variant linkage of a true positive allele (Figure 3A). PD is used in variant linkage validation for potential misassembly identification (Figure 1B).

For AC, D, and E evaluation (Figure 4), all samples were simulated based on the RNA-seq data of sample SRR7779668, as well as optimized values derived from the best samples from the 158 dogs of the cohort.[35] These include 180,435 read pairs of sample SRR7779475 for D, 0.31% of SRR7779475 for E, and 0.45 of SRR7779668 for RD (Table S2). To evaluate the impact of AC (Figure 4A), random allele sampling was repeated 30 times per AC group (Table S3), resulting in a total of $10\ (\textit{ACs}) \times 30\ (\textit{allele sampling repeats}) = 300\ \textit{simulated samples}$. To minimize the influence of AC, only 5 AC groups with homozygous DLA-12 were used for E, D and RD simulation (Figures 4B–4D). RD was represented by the uniformity test statistics calculated with mapped read count at each base position of DLA-I exons 2 and 3, using the SciPy package.[54] To evaluate RD, 5 samples with an RD range representing the entire cohort were selected, including SRR7779632 with $RD = 0.08$, SRR7779628 with $RD = 0.17$, SRR7779614 with $RD = 0.27$, SRR7779679 with $RD = 0.36$, and SRR7779668 with $RD = 0.45$. A total of simulated samples of $7\ (\textit{Es}) \times 5\ (\textit{ACs}) \times 10\ (\textit{allele sampling repeats}) = 350$ for E, $8\ (\textit{Ds}) \times 5\ (\textit{ACs}) \times 10\ (\textit{allele sampling repeats}) = 400$ for D, and $5\ (\textit{RDs}) \times 5\ (\textit{ACs}) \times 10\ (\textit{allele sampling repeats}) = 250$ for RD were generated.

For software comparison (Figure 5), random allele sampling was repeated 10 times for each of the 10 ACs separately for known alleles (Figure 5A) and new alleles (Figure 5B), using the 6 samples chosen as described in Figure 3 (Table S4). New alleles are from contigs assembled by the KPR tools, as shown in Figure 6 (Table S5). A total of $6\ (\textit{samples}) \times 10\ (\textit{ACs}) \times 10\ (\textit{allele sampling repeats}) = 600\ \textit{simulated samples}$ were generated separately known and new allele genotyping.

Simulated DLA-I read pairs of each sample were generated from selected alleles, considering 3 parameters derived from the actual RNA-seq data of the sample. First, sequencing error rate was estimated from the non-DLA-I CDS regions after aligning the RNA-seq reads to the canFam3 reference genome with BWA[52] (version 0.7.17). Based on the estimated rate, sequencing errors were randomly introduced to the simulated RNA-seq reads. Second, relying on the mapping position to the reference alleles, each simulated read-pair was made to replace its corresponding actual RNA-seq read pair, having the same read length and insert size. Third, the expression levels of the 3 DLA-I genes were estimated via a unique region in exon 3, i.e., "ACCATGTAC" for DLA-88, "TGGATGTTT" for DLA-12, and "TGGACTTCG" for DLA-64. Based on this, the amount of simulated read-pairs of each DLA-I allele was determined (and evenly split among the two alleles for heterozygous allele simulation).

TPR and FDR

Using the allele number (integer) or the relative allele expression level (non-integer) to represent each positive call, TPR was calculated by $TPR = \frac{TP}{TP+FN}$ and $FDR = \frac{FP}{TP+FP}$, where TP, TN, and FP respectively represent the number of true positives, false negatives, and false positives. A relative expression level is the expression fraction of an allele in a sample by treating the combined expression of all alleles in the sample as 1.

Genotyping a published cohort

RNA-seq data of 158 tumors and 64 normal samples of 158 dogs from a published study[35,36] were downloaded from the SRA database (PRJNA489087) (Table S5), and quality controlled as follows. First, MultiQC[55] (version 1.5) was used to examine GC content and duplicate level (Figures S1B–S1D). Base quality distribution before and after Trimmomatic trimming was also examined (Figure S1A). Second, the distributions of per sample read-pair total amount, mapping rate and quality, and CDS targeting rate were examined to identify and exclude samples that fail to meet the cutoffs (Figures S1G–S1J). Third, tumor-normal sample pairing accuracy was evaluated by examining the fraction of shared germline variants between any two samples of a study[6] (Figure S1E). Forth, breed validation was performed with using breed-specific germline variants identified previously.[6]

After QC, each sample was genotyped using the KPR *de novo* assembler and genotyper, with optimized parameters of $N = 1,000$, $K = 50$, $PD = 15$ for DLA-88 and $PD = 38$ for DLA-12, following steps described previously and outlined in Figure 1.

RNA-seq genotyping validation with WES data

WES data from the same 158 dogs genotyped were downloaded from the SRA database (PRJNA489159). The sequence reads were first cleaned by Trimmomatic with default setting. Germline variants were used to validate the pairing of WES and RNA-seq samples via dog IDs (Figure S1F).

For WES read validation, WES read pairs of a sample were mapped with BWA to a database that consists of all complete contigs assembled from RNA-seq data. All concordantly and identically mapped reads, including both uniquely and repetitively mapped, were extracted. All RNA-seq contigs of each such mapped read were identified. An RNA-seq contig and the genotyping result are considered valid if the contig has the most or the second most WES reads from the same dog mapped.

For WES contig validation, WES read pairs were used to assemble exon 2 and exon 3 of DLA-I genes independently, following steps outlined in Figure 1. WES contigs were then compared with RNA-seq contigs with BLAST. An RNA-seq contig and the genotyping result are considered valid if the RNA-seq contig is matched with 100% sequence identify and an aligned length ≥ 90 amino acids, which is roughly the size of an exon, by a WES contig from the same dog.

Manual examination for false positive reduction

Manual examinations were performed to identify false positives originated from misassembly or sequencing errors as exemplified as follows. First, DLA-88*501:01 and DLA-12*001:01 share long stretch of identical sequence in the middle. The expression of both alleles in the same sample may result in chimeric assemblies that consist of the head of one allele and the tail of the other allele, e.g., DLA-12*14:01 and DLA-88*501:New3 in certain samples (Table S5). Unfortunately, the RNA-seq read pairs are often not long enough to identify these mis-assemblies. To alert the users of potential misassembly, the KPR output files have flagged any new alle candidates that share >270bp identical sequence with another allele (known or new) in the same sample. Second, DLA-88*006:New42 represents an artifact assembly caused by sequencing error, as only one read pair out of 443 total supports its linkage between positions 466 and 537. These manual examinations excluded 7 new allele candidates. The known allele, DLA-12*14:01, was also removed if the same sample was found to expressing both DLA-88*501:01 and DLA-12*001:01.

Allele clustering and bootstrapping

DLA-I alleles were clustered via Clustal-Omega multi-sequence alignment of amino acid sequences encoded by exons 2 and 3. Euclidean distance and complete agglomeration method were used in both sequence clustering and bootstrapping. Heatmaps and dendrograms were created by "gplot" in R.

Parameter optimization and genotyping with the blood RNA-seq dataset

The RNA-seq data were downloaded from the SRA database (PRJNA629466) and quality controlled as described above (Figure S6). Six samples (SRR11650119, SRR11650123, SRR11650138, SRR11650141, SRR11650147, and SRR11650153) were selected for simulations (Figure S7), as they represent the majority of the dogs of the cohort in DLA-I read pair amount and sequencing error rate (Figure S8). The parameter optimization simulations were performed as described above. Then, all 39 dogs of this cohort were genotyped by KPR using optimized parameters of $N = 30,000$, $K = 20$, $PD = 12$ for DLA-88, and $PD = 38$ for DLA-12 (Figure S12).

QUANTIFICATION AND STATISTICAL ANALYSIS

Paired T tests and Wilcoxon tests were conducted on DLA-88 alleles of the 7 samples genotyped by the KPR software and Sanger sequencing (Table S1). The tests are based on: 1) the alleles by assigning equal value to each allele genotyped in a sample; and 2) the allele expression levels by using the actual expression fraction of each allele in a sample estimated by KPR or by the number of Sanger sequencing clones, assuming all genotyped alleles in a sample summing to 1. The same statistical tests were performed on all DLA-I alleles genotyped between the tumor and normal samples from the same animals (Table S5). Software comparison (Figure 5) was conducted using Wilcoxon rank-sum test on genotyping results of simulated samples.