Title

Canine reference genome accuracy impacts variant calling: Lessons learned from investigating embryonic lethal variants.

Journal

Animal Genetics, 53(5)

Authors

Kinsey, Nathan
Belanger, Janelle
Oberbauer, Anita

Publication Date

2022-10-01

DOI

10.1111/age.13241

Peer reviewed

HHS Public Access

Author manuscript

Anim Genet. Author manuscript; available in PMC 2023 October 01.

Published in final edited form as:

Anim Genet. 2022 October ; 53(5): 706–708. doi:10.1111/age.13241.

Canine reference genome accuracy impacts variant calling: Lessons learned from investigating embryonic lethal variants

Nathan A. Kinsey,

Janelle M. Belanger,

Dog Biomedical Variant Database Consortium[†],

Anita M. Oberbauer

Department of Animal Science, University of California, Davis, USA

Abstract

Deficient homozygosity of a variant maintained in a population suggests that the variant may be embryonic lethal. We examined whole genome sequence data from 675 canids to investigate for variants with missing homozygosity and high predicted impact. Our analysis identified 45 variants, in 32 genes. However, further scrutiny of the sequence reads revealed that all but one of these variants were artifacts of the variant calling process when using CanFam3.1, a widely utilized canine reference genome. We demonstrate that the use of multiple, newer reference genomes could reduce artifacts and lead to more accurate variant identification.

Keywords

bioinformatics; canis lupus familiaris; genomics; lethal recessive; sequence mapping; whole genome sequencing

Variants with considerably less frequent homozygosity than expected under Hardy–Weinberg equilibrium (HWE) have been suggested to be embryonic lethal or to cause developmental disorders in many species, including canids (Georges et al., 2019). Using missing homozygosity to annotate variants, we characterized novel and previously reported variants for potential impairment of canine development; we then examined the reliability of the mapping used to determine missing homozygosity.

Whole genome sequence data from 675 canids mapped to CanFam3.1 from the Dog Biomedical Variant Database Consortium were analyzed using ENSEMBL variant effect predictor (version 101) for variant impact prediction and sorting intolerant from tolerant

Correspondence: Nathan A. Kinsey, Department of Animal Science, University of California, Davis, CA 95616, USA. natkinsey@ucdavis.edu.

[†]Affiliations, funding and email addresses for DBVDC members are listed in Appendix S1.

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

(SIFT) scoring to predict missense variant impact (Howe et al., 2021; Jagannathan et al., 2019; McLaren et al., 2016; Vaser et al., 2016).

Variants marked as 'high-impact' or having a SIFT prediction score of 0.0–0.05 ('not tolerated') were considered deleterious. Allelic frequencies and HWE exact test statistics were determined using PLINK2.3A (Chang et al., 2015; Wigginton et al., 2005). A minimum minor allele frequency of 5% and a maximum heterozygote frequency of 25% were set, using the maximum frequencies reported in cattle for recessive lethal variants as a guideline (Upperman et al., 2019). Variants with HWE p-values > 0.05 following a correction to control the false discovery rate were discarded; variants with more than two homozygous cases in the dataset were also discarded (Benjamini & Yekutieli, 2001). This pipeline yielded 45 variants suspected to be embryonic lethal or causing developmental disorders owing to their missing homozygosity and predicted deleterious impact; the 45 variants are located within 32 genes (Table S1). The variants were further investigated using a separate cohort of 39 dogs whose whole genome sequence reads were mapped to CanFam3.1. Integrative Genomics Viewer and BLAST was used to manually assess sequence read quality in variant regions for those 39 dogs and on selected sequence reads against the five available dog reference genomes on NCBI (annotation release 105) to check for mismapping (Altschul et al., 1990; Robinson et al., 2017).

Of the 45 predicted deleterious variants deficient in homozygosity, only a single variant appeared to be valid and to have a developmental impact. The variant is a frameshift in *ENSCAFG00000043059* (CFA 1), a zinc finger gene with widespread embryonic expression, making it a promising candidate for further investigation (Megquier et al., 2019; NCBI, 2021). The remaining identified variants appear to be false. Mismapping of sequence reads occurred for two main reasons. First, CanFam3.1 is based on a female boxer, and thus sequence reads from the Y chromosome are incorrectly mapped to autosomes (Tsai et al., 2019). For example, the suspected deleterious variant ascribed to *CASP6*, a gene found on chromosome CFA 32, is actually associated with sequence reads that most accurately align to an intergenic region found on CFA Y (Table S1). Secondly, CanFam3.1 possesses over 20 000 gaps with approximately 20% of them occurring within genes (Wang et al., 2021). Sequence reads that cover these gaps either erroneously map within the correct gene or map to an entirely incorrect gene. In either case, the variant would appear to be in a heterozygous state and would never be called homozygous, which could then be interpreted as missing homozygosity and attributed to causing embryonic lethality or developmental disorders (Jagannathan et al., 2019). The high percentage of candidate genes indicated by false variants owing to assembly issues is a major obstacle to studies focusing on loci with high heterozygosity. Several other dog genome assemblies (Table S2) have been recently reported that might mitigate the shortcomings of a single alignment (i.e. CanFam3.1), including ROS_Cfam_1.0, UNSW_CanFamBas_1.0, Basenji_breed-1.1, Dog10K_Boxer_Tasha, UU_Cfam_GSD_1.0 and UMICH_Zoey_3.1 (Edwards et al., 2021; Halo et al., 2021; Jagannathan et al., 2021; The Roslin Institute, 2020; Wang et al., 2021). The present findings demonstrate the importance of confirming putative variants across multiple assemblies, as a single reference genome may be insufficient for accurately identifying variants.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Funding information

Catherine André and Christophe Hitte were supported by the French Cani-DNA CRB (http://dog-genetics.genouest.org), which is part of the CRB-Anim infrastructure, ANR-11-INBS-0003. Kari J. Ekenstedt was supported by the Office of the Director, National Institutes of Health under award number K01-OD027051. Eva Furrow was supported by the Office of the Director, National Institutes of Health under award number K01-OD019912. Tosso Leeb was supported by the Albert-Heim Foundation (project no. 105). Kim M. Summers receives core support from the Mater Foundation, Brisbane, Australia.

DATA AVAILABILITY STATEMENT

The whole genome sequencing data used for the initial cohort (n = 675) is publicly available under the sample accession numbers listed in Table S3. Mapped reads have been submitted to the Sequence Read Archive and are accessible under project accession number PRJNA816174.

REFERENCES

Altschul SF, Gish W, Miller W, Myers EW & Lipman DJ (1990) Basic local alignment search tool. Journal of Molecular Biology, 215, 403–410. 10.1016/S0022-2836(05)80360-2 [PubMed: 2231712]

Benjamini Y & Yekutieli D (2001) The control of the false discovery rate in multiple testing under dependency. The Annals of Statistics, 29, 1165–1188. 10.1214/aos/1013699998

Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM & Lee JJ (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience, 4, 7. 10.1186/s13742-015-0047-8 [PubMed: 25722852]

Edwards RJ, Field MA, Ferguson JM, Dudchenko O, Keilwagen J, Rosen BD et al. (2021) Chromosome-length genome assembly and structural variations of the primal basenji dog (*Canis lupus* familiaris) genome. BMC Genomics, 22, 188. 10.1186/s12864-021-07493-6 [PubMed: 33726677]

Georges M, Charlier C & Hayes B (2019) Harnessing genomic information for livestock improvement. Nature Reviews. Genetics, 20, 135–156. 10.1038/s41576-018-0082-2

Halo JV, Pendleton AL, Shen F, Doucet AJ, Derrien T, Hitte C et al. (2021) Long-read assembly of a great Dane genome highlights the contribution of GC-rich sequence and mobile elements to canine genomes. PNAS, 118, 1–9. 10.1073/pnas.2016274118

Howe KL, Achuthan P, Allen J, Allen J, Alvarez-Jarreta J, Amode MR et al. (2021) Ensembl 2021. Nucleic Acids Research, 49, D884–D891. 10.1093/nar/gkaa942 [PubMed: 33137190]

Jagannathan V, Drögemüller C, Leeb T & Dog Biomedical Variant Database Consortium, (DBVDC). (2019) A comprehensive biomedical variant catalogue based on whole genome sequences of 582 dogs and eight wolves. Animal Genetics, 50, 695–704. 10.1111/age.12834 [PubMed: 31486122]

Jagannathan V, Hitte C, Kidd JM, Masterson P, Murphy TD, Emery S et al. (2021) Dog10k_Boxer_Tasha_1.0: a long-read assembly of the dog reference genome. Genes, 12, 847. 10.3390/genes12060847 [PubMed: 34070911]

McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A et al. (2016) The Ensembl variant effect predictor. Genome Biology, 17, 122. 10.1186/s13059-016-0974-4 [PubMed: 27268795]

Megquier K, Genereux DP, Hekman J, Swofford R, Turner-Maier J, Johnson J et al. (2019) BarkBase: epigenomic annotation of canine genomes. Genes, 10, 433. 10.3390/genes10060433 [PubMed: 31181663]

NCBI, 2021. LOC102155960 zinc finger protein 577-like [Canis lupus familiaris (dog)] – Gene – NCBI. https://www.ncbi.nlm.nih.gov/gene/102155960. Accessed September 3, 2022.

Robinson JT, Thorvaldsdóttir H, Wenger AM, Zehir A & Mesirov JP (2017) Variant review with the integrative genomics viewer. Cancer Research, 77, e31–e34. 10.1158/0008-5472.CAN-17-0337 [PubMed: 29092934]

The Roslin Institute, 2020. ROS_Cfam_1.0 – genome – assembly – NCBI National Center for Biotechnology Information https://www.ncbi.nlm.nih.gov/assembly/GCF_014441545.1/. Accessed April 01, 2022.

Tsai KL, Evans JM, Noorai RE, Starr-Moss AN & Clark LA (2019) Novel Y chromosome retrocopies in canids revealed through a genome-wide association study for sex. Genes, 10, 320. 10.3390/genes10040320 [PubMed: 31027231]

Upperman LR, Kinghorn BP, MacNeil MD & Van Eenennaam AL (2019) Management of lethal recessive alleles in beef cattle through the use of mate selection software. Genetics Selection Evolution, 51, 36. 10.1186/s12711-019-0477-3

Vaser R, Adusumalli S, Leng SN, Sikic M & Ng PC (2016) SIFT missense predictions for genomes. Nature Protocols, 11, 1–9. 10.1038/nprot.2015.123 [PubMed: 26633127]

Wang C, Wallerman O, Arendt M-L, Sundström E, Karlsson Å, Nordin J et al. (2021) A novel canine reference genome resolves genomic architecture and uncovers transcript complexity. Commun Biol, 4, 1–11. 10.1038/s42003-021-01698-x [PubMed: 33398033]

Wigginton JE, Cutler DJ & Abecasis GR (2005) A note on exact tests of hardy-weinberg equilibrium. American Journal of Human Genetics, 76, 887–893. [PubMed: 15789306]