Examining the Contribution of Recent Gene Retrocopies to Disease and Phenotypic Diversity in Dogs

By

KEVIN BATCHER
DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Integrative Genetics and Genomics

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

____
Danika Bannasch, Chair

____
Megan Dennis

____
Peter J. Dickinson

____
Anita Oberbauer

Committee in Charge

2022

ABSTRACT

Gene retrocopies are commonly found in mammalian genomes where they are typically presumed to be nonfunctional pseudogenes. However, active LINE-1 is capable of inserting new gene retrocopies, which can be functional via direct expression or through altering the expression patterns of nearby genes at the insertion site. Challenging their designation as pseudogenes, recent studies have implicated retrocopies in cancer and other disorders. Presented here is an in investigation into the recent, polymorphic retrocopies in canids. First, the effects of two canine *FGF4* retrocopies were investigated. While two separate *FGF4* retrocopies are both associated with two distinct forms of disproportionate dwarfism, the *FGF4* retrocopy on CFA12 alone was shown to be associated with calcification of the intervertebral disc and increased susceptibility to intervertebral disc disease, a phenotype known as chondrodystrophy. While multiple additional *FGF4* retrocopy insertions located elsewhere in the genome have been discovered in canids with unknown phenotypic associations, the complete landscape of retrocopies in the species was unclear. Thus, a method for the identification of recent, polymorphic retrocopies from whole genome sequencing data was developed and applied to canids. A large number of gene retrocopies insertions were identified, showing that gene retrotransposition events are a common occurrence in the canid genome and not unique to the *FGF4* gene. Target-site duplications, which are a characteristic of LINE-1 mediated retrotransposition, were identified at the insertion sites of the retrocopies, conforming they were a consequence of LINE-1 activity. In addition to the expressed and functional *FGF4* retrocopies, many other canine retrocopies were also shown to be expressed and under selection, alluding towards possible functions. A polymorphic *SNN* retrocopy associated with red coat color in Poodles was identified and characterized, where it was shown to alter the expression pattern of a nearby gene, *GPR22*. Lastly, after the identification of numerous polymorphic canine retrocopies, a similar analysis using the same methodology was performed in equine genomes for comparison. Equids have fewer gene retrocopy insertions on average then canids, indicating that LINE-1 is more active within canids. These findings highlight the contribution of LINE-1 mediated retrotransposition events to the genomic and phenotypic diversity of dogs.

Table of Contents

Table of Figures

Table of Tables

CHAPTER 1: Introduction

Transposable elements, often colloquially referred to as “jumping genes”, were famously first discovered by renowned geneticist Barbara McClintock based on her genetic analyses in maize [1]. However, the significance of transposable elements was not well understood at the time; for decades, transposable elements were considered “junk DNA” of no biological or evolutionary significance [2]. The initial draft sequence of the human genome revealed that approximately 45% of the assembly comprised transposable elements, with other mammalian genomes following similar patterns [3,4]. While it is now well understood that transposable elements have played an important role in the structure and function of genomes, the only autonomous transposable element that remains active in most mammals is the long interspersed nuclear element-1 (LINE-1) [5].

LINE-1 is a retrotransposon that replicates through an RNA intermediate via the function of the two LINE-1 encoded proteins, ORF1p and ORF2p in a process called target prime reverse transcription [6]. One of the hallmarks of a LINE-1 mediated insertion is the duplication of short (7-20bp) strands of genomic DNA flanking the insertion site, called target site duplications (TSD) [7]. While the LINE-1 encoded proteins ORF1p and ORF2p preferentially produce copies of LINE-1 mRNA in a *cis* fashion, LINE-1 proteins are also capable of producing copies of other cellular RNAs [8]. Specifically, active LINE-1 produces copies of the non-autonomous short interspersed nuclear element (SINE) as well as messenger RNA, resulting in processed pseudogenes (or gene retrocopies) [9]. This dissertation focuses on the identification and characterization of recently inserted and polymorphic gene retrocopies with functional capacity.

Gene retrocopies, which come from processed mRNA, lack the promotor region and introns present at the parent gene, and also contain a polyadenylated tail at the 3’ end. Most retrocopies present in the human and other reference assemblies are the result of ancestral retrotransposition events, and therefore, often accumulate various missense and nonsense mutations leading to their designation as “processed pseudogenes” [10-12]. However, much like how transposable elements are no longer considered “junk DNA”, the term “pseudogene” itself has become somewhat of a misnomer, as many examples of functional so-called pseudogenes have been identified [13,14]. Likewise, numerous examples of expressed and fully functional gene retrocopies have been characterized [15,16]. Gene retrocopies can function through various means, such as direct expression and translation, alteration of expression patterns of nearby insertion site genes, or through the formation of novel chimeric transcripts [16,17]. Gene retrocopies have also been implicated in cancer and other neurodegenerative, mental, and cardiovascular disorders [16,18-20]. Although gene retrotranspositions were thought to be rare genomic events, with an estimated 21 gene retrotranspositions every million years in the human lineage [15], these estimates were based on the retrocopies present in the reference genome assembly. LINE-1 is still active and capable of inserting newer, polymorphic gene retrocopies, referred to as retrotransposed gene copy number variants (retroCNVs), which may be absent from genome assemblies [21].

RetroCNVs are large structural variants that are difficult to identify from short read sequencing data using typical variant calling algorithms [22]. While the retroCNV parent genes can be identified based on the presence of intron-less copies, identification of an insertion site is considered the gold standard for calling a retroCNV [19]. Insertion site detection can be difficult because retroCNVs themselves are repetitive in nature and can insert into a repetitive region of the genome. Additionally, some highly transcribed parent genes, such as genes encoding ribosomal proteins, tend to produce many retrocopies [23-25]. Early analyses of human retroCNVs using low coverage modern human whole-genome sequencing (WGS) data from the 1000 Genomes Project dataset identified around 200 retroCNVs [21,23,26]. When higher coverage datasets were analyzed, 1,542 retroCNVs were discovered in 17,795 human genomes, indicating that gene retrotranspositions are not as rare as previously thought [27]. A recent analysis of multiple long-read human genome assemblies also confirmed a higher rate of gene retrocopy acquisition in humans than previously described [28]. However, the number of active LINE-1 varies by species; for example, humans have around 150 full length, intact copies of LINE-1, while horses have 72, dogs have 264, and mice around 2,800 [29]. Species with more total retrotransposition-active LINE-1s, or alternatively a small number of highly active LINE-1s, may acquire more gene retrocopies. Recently, a large number of retroCNV were identified in natural populations of mice correlating to their large number of active LINE-1 [30]. These murine retroCNV were also shown to be under negative selection, suggestive of deleterious effects. However, there is little information about retroCNV outside of a small number of model organisms. The goal of this dissertation is a comprehensive analysis of retroCNVs in dogs.

The domestic dog has gone through two recent population bottlenecks, providing a unique opportunity for genetic studies. First, domestication, which occurred somewhere between 11-16k years ago, was estimated to have resulted in a 16-fold reduction in population size [31]. A more recent bottleneck occurred during the formation of the modern dog breeds over the past few hundred years [32]. Most modern breeds derive from a small number of founder individuals with strong artificial selection for desired phenotypic traits. The population bottlenecks have resulted in the accumulation of deleterious variants in dogs [32,33], evidenced by the higher frequencies of disease and anatomical abnormalities within modern dog breeds [34,35]. These unique characteristics allow for the efficient mapping of genetic traits in dogs [36]. Additionally, a high prevalence of SINE and LINE dimorphisms in the canine genome are indicative of highly active LINE-1, making dog an interesting model organism for the study of retrotransposable elements [37,38].

In 2009, a form of disproportionate dwarfism called chondrodysplasia was mapped to canine chromosome (CFA) 18 and narrowed to an expressed *FGF4* retrocopy insertion [39]. A separate functional *FGF4* retrocopy insertion on CFA12 was found to be associated with another form of disproportionate dwarfism in dogs called chondrodystrophy [40]. Unlike chondrodysplasia, however, chondrodystrophy is a well-defined phenotype in dogs. In addition to disproportionate dwarfism, the chondrodystrophic breeds experience premature degeneration of the nucleus pulposus at a young age as shown by histopathological examination of the intervertebral discs [41,42]. This premature degeneration predisposes chondrodystrophic dogs to Hansen type 1 intervertebral disc disease (IVDD), typified by the acute extrusion of degenerate and often calcified nucleus pulposus into the vertebral canal [41]. Type 1 IVDD can result in severe pain and a loss of neurological function, which often requires surgery to remove the herniated disc material and restore function [43].

Both *FGF4* retroCNVs on CFA12 and CFA18 are common across many dog breeds, including several breeds with both copies, such as the Dachshund, which also has the highest prevalence of IVDD [44]. Misusage of the terms chondrodystrophy and chondrodysplasia, in addition to the presence of two distinct *FGF4* retrocopies, have led to confusion within the scientific literature on IVDD in dogs. Because of this, there is a need to clarify the roles that these two *FGF4* retroCNVs play in susceptibility to IVDD. Chapter 2 [45] is an examination into the combined effects of the two distinct *FGF4* retroCNVs on both intervertebral disc calcification and IVDD surgery. Our analysis of the two *FGF4* retrocopies—referred to as *12-FGF4RG* and *18-FGF4RG* to distinguish the copies on CFA12 and CFA18, respectively—found that *12-FGF4RG* alone was the primary driver of disc calcification and age at time of IVDD surgery. However, multiple individuals received surgery for what appeared to be Hansen Type 1 IVDD at a young age that did not have any copies of *12-FGF4RG,* indicating that other genetic risk factors likely play a role. In particular, we hypothesized that additional undiscovered *FGF4* retroCNVs present in dogs may play a role in disease. To test this hypothesis, in Chapter 3 [46], we queried canine genomes for *FGF4* retrocopy insertions using diverse molecular techniques and genomic data analysis. We identified multiple additional *FGF4* retroCNV, including two copies on CFA13, and, as such, a new method for naming the retrocopies was adopted: *FGF4L1*, for the first copy discovered on CFA18, *FGF4L2* for the copy on CFA12, and so on. While some naming conventions for retrocopies designate them as pseudogenes using the letter “P”, the letter “L” for “like” was instead used in order to avoid any functional inferences. While multiple additional *FGF4* retroCNVs were identified, no clear phenotypic associations were found with IVDD or limb abnormalities. Despite this, our results highlight the need for a more comprehensive approach for the identification of polymorphic canine retrocopies.

Chapter 4 [47] describes a novel approach for retroCNV discovery that we developed and applied to a large dataset of canine WGS data, resulting in the discovery of previously unknown retrocopy insertions. Transcriptome analysis of a subset of the discovered retroCNVs indicated expression in testes tissue for ~12% present in a set of golden retrievers, including multiple chimeric transcripts. While the phenotypic effects of the retroCNV remains unclear, the distribution of many retroCNVs shows clear breed biases, indicating possible positive selection during breed formation, much like the *FGF4* retroCNVs. Finally, I describe a new example of a functional *SNN* retroCNV in dogs associated with red coat color in Chapter 5 [48]. Appendix 1 contains a similar investigation into the landscape of retroCNVs in equids that was initiated to compare the rate of retroCNV accumulation across species. Our work makes clear the high rate of retroCNV accumulation in dogs versus other similarly well characterized species (e.g., humans and mouse) attributable to increased LINE-1 activity in the germline. This is important, considering the likely understudied contributions of canine retroCNVs to phenotypic diversity and heritable disorders observed in the domestic dog.

CHAPTER 1 REFERENCES

1. McClintock, B. The origin and behavior of mutable loci in maize. *Proceedings of the National Academy of Sciences* **1950**, *36*, 344-355.
2. Biémont, C. A brief history of the status of transposable elements: from junk DNA to major players in evolution. *Genetics* **2010**, *186*, 1085-1093.
3. Olsen, U.D.J.G.I.H.T.B.E.P.P.R.P.W.S.S.T.D.N.C.J.-F.; 9, R.G.S.C.S.Y.F.A.H.M.Y.T.T.A.I.T.K.C.W.H.T.Y.; Genoscope; 10, C.U.-W.J.H.R.S.W.A.F.B.P.B.T.P.E.R.C.W.P.; Department of Genome Analysis, I.o.M.B.R.A.P.M.N.G.T.S.R.A.; 11, G.S.C.S.D.R.D.-S.L.R.M.W.K.L.H.M.D.J.; 15, B.G.I.H.G.C.Y.H.Y.J.W.J.H.G.G.J. Initial sequencing and analysis of the human genome. *nature* **2001**, *409*, 860-921.
4. Lindblad-Toh, K.; Wade, C.M.; Mikkelsen, T.S.; Karlsson, E.K.; Jaffe, D.B.; Kamal, M.; Clamp, M.; Chang, J.L.; Kulbokas, E.J.; Zody, M.C. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **2005**, *438*, 803-819.
5. Beck, C.R.; Garcia-Perez, J.L.; Badge, R.M.; Moran, J.V. LINE-1 elements in structural variation and disease. *Annual review of genomics and human genetics* **2011**, *12*, 187.
6. Luan, D.D.; Korman, M.H.; Jakubczak, J.L.; Eickbush, T.H. Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. *Cell* **1993**, *72*, 595-605.
7. Kazazian, H.H.; Wong, C.; Youssoufian, H.; Scott, A.F.; Phillips, D.G.; Antonarakis, S.E. Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man. *Nature* **1988**, *332*, 164-166.
8. Wei, W.; Gilbert, N.; Ooi, S.L.; Lawler, J.F.; Ostertag, E.M.; Kazazian, H.H.; Boeke, J.D.; Moran, J.V. Human L1 retrotransposition: cispreference versus trans complementation. *Molecular and cellular biology* **2001**, *21*, 1429-1439.
9. Esnault, C.; Maestre, J.; Heidmann, T. Human LINE retrotransposons generate processed pseudogenes. *Nature genetics* **2000**, *24*, 363-367.
10. Vinckenbosch, N.; Dupanloup, I.; Kaessmann, H. Evolutionary fate of retroposed gene copies in the human genome. *Proceedings of the National Academy of Sciences of the United States of America* **2006**, *103*, 3220-3225, doi:10.1073/pnas.0511307103.
11. Rosikiewicz, W.; Kabza, M.; Kosiński, J.G.; Ciomborowska-Basheer, J.; Kubiak, M.R.; Makałowska, I. RetrogeneDB–a database of plant and animal retrocopies. *Database* **2017**, *2017*.
12. Kabza, M.; Ciomborowska, J.; Makałowska, I. RetrogeneDB—a database of animal retrogenes. *Molecular biology and evolution* **2014**, *31*, 1646-1648.
13. Chen, X.; Wan, L.; Wang, W.; Xi, W.-J.; Yang, A.-G.; Wang, T. Re-recognition of pseudogenes: From molecular to clinical applications. *Theranostics* **2020**, *10*, 1479.
14. Troskie, R.-L.; Jafrani, Y.; Mercer, T.R.; Ewing, A.D.; Faulkner, G.J.; Cheetham, S.W. Long-read cDNA sequencing identifies functional pseudogenes in the human transcriptome. *Genome biology* **2021**, *22*, 1-15.
15. Casola, C.; Betrán, E. The genomic impact of gene retrocopies: what have we learned from comparative genomics, population genomics, and transcriptomic analyses? *Genome biology and evolution* **2017**, *9*, 1351-1373.

16. Ciomborowska-Basheer, J.; Staszak, K.; Kubiak, M.R.; Makałowska, I. Not So Dead Genes—Retrocopies as Regulators of Their Disease-Related Progenitors and Hosts. *Cells* **2021**, *10*, 912.

17. Kubiak, M.R.; Makałowska, I. Protein-coding genes' retrocopies and their functions. *Viruses* **2017**, *9*, 80.

18. Staszak, K.; Makałowska, I. Cancer, retrogenes, and evolution. *Life* **2021**, *11*, 72.

19. Richardson, S.R.; Salvador-Palomeque, C.; Faulkner, G.J. Diversity through duplication: Whole-genome sequencing reveals novel gene retrocopies in the human population. *Bioessays* **2014**, *36*, 475-481.

20. Bim, L.V.; Navarro, F.C.; Valente, F.O.; Lima-Junior, J.V.; Delcelo, R.; Dias-da-Silva, M.R.; Maciel, R.; Galante, P.A.; Cerutti, J.M. Retroposed copies of RET gene: a somatically acquired event in medullary thyroid carcinoma. *BMC medical genomics* **2019**, *12*, 1-13.

21. Schrider, D.R.; Navarro, F.C.; Galante, P.A.; Parmigiani, R.B.; Camargo, A.A.; Hahn, M.W.; de Souza, S.J. Gene copy-number polymorphism caused by retrotransposition in humans. *PLoS genetics* **2013**, *9*, e1003242.

22. Tattini, L.; D'Aurizio, R.; Magi, A. Detection of genomic structural variants from next-generation sequencing data. *Frontiers in bioengineering and biotechnology* **2015**, *3*, 92.

23. Ewing, A.D.; Ballinger, T.J.; Earl, D.; Harris, C.C.; Ding, L.; Wilson, R.K.; Haussler, D. Retrotransposition of gene transcripts leads to structural variation in mammalian genomes. *Genome biology* **2013**, *14*, R22.

24. Gonçalves, I.; Duret, L.; Mouchiroud, D. Nature and structure of human genes that generate retropseudogenes. *Genome research* **2000**, *10*, 672-678.

25. Zhang, Z.; Harrison, P.M.; Liu, Y.; Gerstein, M. Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome. *Genome research* **2003**, *13*, 2541-2558.

26. Abyzov, A.; Iskow, R.; Gokcumen, O.; Radke, D.W.; Balasubramanian, S.; Pei, B.; Habegger, L.; Lee, C.; Gerstein, M.; Consortium, G.P. Analysis of variable retroduplications in human populations suggests coupling of retrotransposition to cell division. *Genome research* **2013**, *23*, 2042-2052.

27. Abel, H.J.; Larson, D.E.; Chiang, C.; Das, I.; Kanchi, K.L.; Layer, R.M.; Neale, B.M.; Salerno, W.J.; Reeves, C.; Buyske, S. Mapping and characterization of structural variation in 17,795 deeply sequenced human genomes. *bioRxiv* **2018**, 508515.

28. Feng, X.; Li, H. Higher Rates of Processed Pseudogene Acquisition in Humans and Three Great Apes Revealed by Long-Read Assemblies. *Molecular Biology and Evolution* **2021**, *38*, 2958-2966.

29. Penzkofer, T.; Jäger, M.; Figlerowicz, M.; Badge, R.; Mundlos, S.; Robinson, P.N.; Zemojtel, T. L1Base 2: more retrotransposition-active LINE-1s, more mammalian genomes. *Nucleic acids research* **2016**, gkw925.

30. Zhang, W.; Xie, C.; Ullrich, K.; Zhang, Y.E.; Tautz, D. The mutational load in natural populations is significantly affected by high primary rates of retroposition. *Proceedings of the National Academy of Sciences* **2021**, *118*.

31. Freedman, A.H.; Gronau, I.; Schweizer, R.M.; Ortega-Del Vecchyo, D.; Han, E.; Silva, P.M.; Galaverni, M.; Fan, Z.; Marx, P.; Lorente-Galdos, B. Genome sequencing highlights the dynamic early history of dogs. *PLoS genetics* **2014**, *10*, e1004016.

32. Ostrander, E.A.; Wayne, R.K.; Freedman, A.H.; Davis, B.W. Demographic history, selection and functional diversity of the canine genome. *Nature Reviews Genetics* **2017**, *18*, 705-720.

33. Marsden, C.D.; Ortega-Del Vecchyo, D.; O'Brien, D.P.; Taylor, J.F.; Ramirez, O.; Vilà, C.; Marques-Bonet, T.; Schnabel, R.D.; Wayne, R.K.; Lohmueller, K.E. Bottlenecks and selective sweeps during domestication have increased deleterious genetic variation in dogs. *Proceedings of the National Academy of Sciences* **2016**, *113*, 152-157.

34. Bellumori, T.P.; Famula, T.R.; Bannasch, D.L.; Belanger, J.M.; Oberbauer, A.M. Prevalence of inherited disorders among mixed-breed and purebred dogs: 27,254 cases (1995–2010). *Journal of the American Veterinary Medical Association* **2013**, *242*, 1549-1555.

35. Schoenebeck, J.J.; Ostrander, E.A. Insights into morphology and disease from the dog genome project. *Annual review of cell and developmental biology* **2014**, *30*, 535.

36. Karlsson, E.K.; Baranowska, I.; Wade, C.M.; Salmon Hillbertz, N.H.; Zody, M.C.; Anderson, N.; Biagi, T.M.; Patterson, N.; Pielberg, G.R.; Kulbokas, E.J. Efficient mapping of mendelian traits in dogs through genome-wide association. *Nature genetics* **2007**, *39*, 1321-1328.

37. Wang, W.; Kirkness, E.F. Short interspersed elements (SINEs) are a major source of canine genomic diversity. *Genome Research* **2005**, *15*, 1798-1808.

38. Halo, J.V.; Pendleton, A.L.; Shen, F.; Doucet, A.J.; Derrien, T.; Hitte, C.; Kirby, L.E.; Myers, B.; Sliwerska, E.; Emery, S. Long-read assembly of a Great Dane genome highlights the contribution of GC-rich sequence and mobile elements to canine genomes. *Proceedings of the National Academy of Sciences* **2021**, *118*.

39. Parker, H.G.; VonHoldt, B.M.; Quignon, P.; Margulies, E.H.; Shao, S.; Mosher, D.S.; Spady, T.C.; Elkahloun, A.; Cargill, M.; Jones, P.G. An expressed fgf4 retrogene is associated with breed-defining chondrodysplasia in domestic dogs. *Science* **2009**, *325*, 995-998.

40. Brown, E.A.; Dickinson, P.J.; Mansour, T.; Sturges, B.K.; Aguilar, M.; Young, A.E.; Korff, C.; Lind, J.; Ettinger, C.L.; Varon, S. FGF4 retrogene on CFA12 is responsible for chondrodystrophy and intervertebral disc disease in dogs. *Proceedings of the National Academy of Sciences* **2017**, *114*, 11476-11481.

41. Hansen, H.-J. A pathologic-anatomical study on disc degeneration in dog: With special reference to the so-called enchondrosis intervertebralis. *Acta Orthopaedica Scandinavica* **1952**, *23*, 1-130.

42. Braund, K.; Ghosh, P.; Taylor, T.; Larsen, L. Morphological studies of the canine intervertebral disc. The assignment of the beagle to the achondroplastic classification. *Research in veterinary science* **1975**, *19*, 167-172.

43. Jeffery, N.D.; Levine, J.M.; Olby, N.J.; Stein, V.M. Intervertebral disk degeneration in dogs: consequences, diagnosis, treatment, and future directions. *J Vet Intern Med* **2013**, *27*, 1318-1333, doi:10.1111/jvim.12183.

44. Bergknut, N.; Egenvall, A.; Hagman, R.; Gustås, P.; Hazewinkel, H.A.; Meij, B.P.; Lagerstedt, A.-S. Incidence of intervertebral disk degeneration–related diseases and associated mortality rates in dogs. *Journal of the American Veterinary Medical Association* **2012**, *240*, 1300-1309.

45. Batcher, K.; Dickinson, P.; Giuffrida, M.; Sturges, B.; Vernau, K.; Knipe, M.; Rasouliha, S.H.; Drögemüller, C.; Leeb, T.; Maciejczyk, K. Phenotypic Effects of FGF4 Retrogenes on Intervertebral Disc Disease in Dogs. *Genes* **2019**, *10*, 435.
46. Batcher, K.; Dickinson, P.; Maciejczyk, K.; Brzeski, K.; Rasouliha, S.H.; Letko, A.; Drögemüller, C.; Leeb, T.; Bannasch, D. Multiple FGF4 retrocopies recently derived within canids. *Genes* **2020**, *11*, 839.
47. Batcher, K.; Varney, S.; York, D.; Blacksmith, M.; Kidd, J.M.; Rebhun, R.; Dickinson, P.; Bannasch, D. Recent, full-length gene retrocopies are common in canids. *Genome Research* **2022**.
48. Batcher, K.; Varney, S.; Affolter, V.K.; Friedenberg, S.G.; Bannasch, D. An SNN retrocopy insertion upstream of GPR22 is associated with dark red coat color in Poodles. *G3 Genes| Genomes| Genetics* **2022**.

CHAPTER 2: Phenotypic Effects of *FGF4* Retrogenes on Intervertebral Disc Disease in Dogs

Kevin Batcher [1], Peter Dickinson [2], Michelle Giuffrida [2], Beverly Sturges [2], Karen Vernau [2], Marguerite Knipe [2], Sheida Hadji Rasouliha [3], Cord Drögemüller [3], Tosso Leeb [3], Kimberly Maciejczyk [1], Chris Jenkins [4], Cathryn Mellersh [4], Danika Bannasch [1]*

[1]Department of Population Health and Reproduction, University of California-Davis, USA

[2]Department of Surgical and Radiological Sciences, University of California-Davis, USA

[3]Institute of Genetics, University of Bern, 3001 Bern, Switzerland

[4]Kennel Club Genetics Centre, Animal Health Trust, Kentford, Newmarket, Suffolk CB8 7UU, UK

*Correspondence: dlbannasch@ucdavis.edu

Genes 2019, 10(6), 435; https://doi.org/10.3390/genes10060435

Received: 8 April 2019; Accepted: 5 June 2019; Published: 7 June 2019

AUTHOR CONTRIBUTIONS

Conceptualization, D.B. and P.D.; Data curation, **K.B.**, K.M., D.B., and P.D.; Formal analysis, **K.B.** and M.G.; Funding acquisition, D.B.; Investigation, B.S., C.M., C.J., D.B., **K.B.**, K.M., K.V., M.K., P.D., S.H.R., and T.L.; Project administration, D.B.; Resources, C.D., C.M., D.B., and T.L.; Supervision, D.B.; Visualization, **K.B.**; Writing—original draft, D.B., **K.B.**, and P.D.; Writing—review & editing, D.B., **K.B.**, and P.D.

ABSTRACT

Two *FGF4* retrogenes on chromosomes 12 (12-*FGF4*RG) and 18 (18-*FGF4*RG) contribute to short-limbed phenotypes in dogs. 12-*FGF4*RG has also been associated with intervertebral disc disease (IVDD). Both of these retrogenes were found to be widespread among dog breeds with allele frequencies ranging from 0.02 to 1; however, their additive contribution to disease is unknown. Surgical cases of IVDD (n = 569) were evaluated for age of onset, disc calcification, and genotypes for the *FGF4* retrogenes. Multivariable linear regression analysis identified the presence of one or two copies of 12-*FGF4*RG associated with significantly younger age at first surgery in a dominant manner. 18-*FGF4*RG had only a minor effect in dogs with one copy. Multivariable logistic regression showed that 12-*FGF4*RG had an additive effect on radiographic disc calcification, while 18-*FGF4*RG had no effect. Multivariable logistic regression using mixed breed cases and controls identified only 12-*FGF4*RG as highly associated with disc herniation in a dominant manner (Odds Ratio, OR, 18.42, 95% Confidence Interval (CI) 7.44 to 50.26; $p < 0.001$). The relative risk for disc surgery associated with 12-*FGF4*RG varied from 5.5 to 15.1 within segregating breeds and mixed breeds. The *FGF4* retrogene on CFA12 acts in a dominant manner to decrease the age of onset and increase the overall risk of disc disease in dogs. Other modifiers of risk may be present within certain breeds, including the *FGF4* retrogene on CFA18.

INTRODUCTION

The domestic dog exhibits a profound degree of phenotypic diversity in size. Two particular conditions affecting size, referred to as chondrodystrophy and chondrodysplasia, are characterized by shortened limbs and are common across many dog breeds [1,2]. The causes for both chondrodystrophy and chondrodysplasia were identified as two separate fibroblast growth factor 4 (*FGF4*) retrogenes on chromosome 12 (12-*FGF4*RG) and chromosome 18 (18-*FGF4*RG), respectively [1,2]. The most severe form of disproportionate dwarfism is seen in breeds that carry both *FGF4* retrogenes, such as Dachshunds, Basset Hounds, and Corgis [1,3]. FGF signaling is involved in early embryonal development [4,5], and appropriate levels of *FGF4* are necessary for normal limb formation [6]. Higher levels of *FGF4* transcripts were seen in dogs with either of the 12-*FGF4*RG or 18-*FGF4*RG insertions, leading to the conclusions that the *FGF4* retrogenes are expressed and that the short-limb phenotype is associated with overexpression during development [1,2]. Similarly in humans, achondroplasia, the most common form of dwarfism, is caused by the gain of function variants in the fibroblast growth factor receptor 3 that results in increased signaling [7,8].

In addition to shortened limbs, chondrodystrophic breeds are also characterized by chondroid metaplasia of the nucleus pulposus leading to premature degeneration and calcification of the intervertebral discs [9,10,11]. This degeneration predisposes chondrodystrophic dogs to intervertebral disc disease (IVDD), a debilitating disorder associated with protrusion or extrusion of intervertebral disc components into the vertebral canal, resulting in pain and/or neurological dysfunction [12]. Disc calcification may be visualized radiographically, and chondrodystrophic dogs with increased numbers of radiographically visible calcified discs have been shown to be at higher risk for clinical IVDD [13,14,15]. Calcification of the nucleus pulposus has also been described in older, non-chondrodystrophic dogs in the late stages of degeneration; however, this process generally occurs at an earlier age in the chondrodystrophic dog breeds [10,16,17].

While all breeds can be affected by IVDD, chondrodystrophic breeds are at particularly high risk [9,10,12]. Hansen classified the IVDD that occurs in the chondrodystrophic dog breeds as Type I, typified by acute extrusion of degenerate, often calcified nucleus pulposus through degenerate annulus fibrosis into the vertebral canal [10,18]. Hansen Type II IVDD generally occurs at a later age in larger breed (non-chondrodystrophic) dogs and more commonly involves chronic protrusion of degenerative disc material. Historically, type II IVDD has been reported to involve fibrous rather than chondroid disc degeneration. However, despite some specific differences in macro and microscopic pathology and in disease progression, Hansen's original work and more recent studies have shown that chondroid metaplasia is a common underlying pathological process in both chondrodystrophic and non-chondrodystrophic breeds [9,19].

Current consensus supports the use of decompressive surgery to remove the disc material impinging on the spinal cord in dogs severely affected by IVDD [12,17], although the cost of surgery can be prohibitive for many owners. Dogs susceptible to IVDD may also suffer multiple disc herniations at different locations throughout their lifetimes [20]. While chondrodystrophic dog breeds with 12-*FGF4*RG alone, such as the French Bulldog and Beagle, are at high risk for IVDD, the chondrodysplastic breeds with 18-*FGF4*RG alone, such as the Scottish Terrier and the West Highland White Terrier, are not considered at high risk [1,10]. However, in dogs with both *FGF4* retrogenes, the contribution of 18-*FGF4*RG to disc degeneration and thus IVDD is unknown. It is also unclear whether or not the *FGF4* retrogenes act in a completely dominant manner or whether any additive effect exists. Since many breeds are homozygous for the *FGF4* retrogenes, the determination of the relative risk for intervertebral disc herniation is challenging. Segregating breeds and mixed breed dogs provide an opportunity to evaluate the risk of herniation in the presence of the retrogenes.

A broad analysis of allele frequency across dog breeds was performed for both *FGF4* retrogenes, identifying breeds that segregate or are fixed for one or both retrogenes. A referral hospital DNA database was utilized to obtain information from dogs that had received decompressive surgery for IVDD, and prospective samples were collected for two years from surgical cases to obtain a large, across breed sample of 569 dogs that had IVDD defined by surgery. These samples were genotyped for both 12-*FGF4*RG and 18-*FGF4*RG. Breed, weight, sex, age at time of first surgery, and the presence of calcified discs at the time of surgery were used to determine the contribution of 12-*FGF4*RG and 18-*FGF4*RG to disease phenotype using linear and logistic regression. A separate logistic regression was performed in mixed breed dogs to determine characteristics contributing to IVDD surgery itself, and a relative risk for 12-*FGF4*RG was calculated in segregating breeds.

MATERIALS AND METHODS

Samples

Unused blood samples were collected from the University of California (UC) Davis Veterinary Medical Teaching Hospital (VMTH) hematology laboratory from 5 November 1999 to 1 February 2016 irrespective of IVDD diagnosis in a random fashion and entered in a repository. After 1 February 2016, the Bannasch laboratory began actively soliciting blood samples from IVDD cases seen at the VMTH. For the purpose of this study, samples that were collected prior to 1 February 2016 were marked as retrospective, while samples collected after 1 February 2016 were marked as prospective. Medical records for all samples in the DNA repository were queried from the VMTH database for evidence of decompressive surgery, and all surgical cases were included in the study. Samples collected from the VMTH were obtained under UC Davis IACUC protocols 12693, 15356, 18561, and 20356.

Medical information retrieved for each case included breed, sex, date of birth, weight, date of procedure, surgical procedure performed, anatomical location of surgery, a summary of findings from any medical imaging performed (Myelography, Computed Tomography, CT, Magnetic resonance Imaging, MRI, or Radiography), and a description of disc material from the surgery report. All non-disc related cases were excluded based on findings indicated in the surgery reports. All breed information was owner reported when presenting at the VMTH. All Miniature and Toy Poodles were treated as one breed, and all Dachshund varieties were treated as one breed; however, based on weight, all but two Dachshunds were miniature in size. The total numbers of dogs from the affected breeds were counted from the DNA repository in order to determine the relative breed representation of the surgical cases in the repository.

Breed-specific allele frequencies for the *FGF4* retrogenes were determined using 2333 samples from the UC Davis DNA repository irrespective of IVDD diagnosis. Additional samples for breed-specific allele frequencies were acquired from the Vetsuisse Biobank, University of Bern DNA repository without health information. Additional samples for the Dachshund breed in the United Kingdom (UK) were genotyped from a collection from the Animal Health Trust.

Phenotyping and Genotyping

Cases were classified into one of three categories based on information obtained from medical imaging and surgery reports. If the disc material removed at the time of surgery was described as calcified, mineralized, chondroid, or in other such similar terms, the case was classified as ’Group A’. The case was also classified as Group A if there was evidence of intervertebral disc calcification on the radiograph or CT report. If the disc material was described as annular or fibrous and there was no evidence of intervertebral disc calcification in any imaging, or if the disc material was described as hydrated or liquid, the case was classified as ’Group B’. Finally, if there was insufficient information to make any determination as to the nature of the IVDD due to a missing or nondescript surgery report and/or no vertebral column radiographs or CT imaging available, the case was classified as ’unknown’. Cases were also separately categorized based on the presence of calcified discs visible on radiograph irrespective of surgery report. DNA was extracted from whole blood samples using a Gentra Puregene DNA extraction kit (Qiagen, Valencia, CA, USA). Genotyping for 12-*FGF4*RG and 18-*FGF4*RG was performed using a PCR based assay as previously described [1] or through commercially available genotyping at the UC Davis Veterinary Genetics Laboratory.

Statistical Analysis

Descriptive statistics including interquartile ranges, median and 95% Confidence intervals for weight and age at surgery were obtained using GraphPad Prism 7.03 for Windows (GraphPad Software, La Jolla, CA, www.graphpad.com). The Mann-Whitney U test was used to compare weights across genotype statuses and between Groups A and B. A Chi-Squared test was used to test for significant differences in allele frequencies between Groups A and B. These analyses were also performed using GraphPad Prism 7.03 for Windows.

Multivariable linear regression analysis was used to identify characteristics associated with age at time of surgery. In order to evaluate breed contribution, the three most frequent breed representatives were used and compared to et al. (Table 1). Age was the dependent variable, and sex, reproductive status (spayed or neutered vs intact), body weight, breed (French bulldog, Dachshund, mixed breed, other pure breed), 12-*FGF4*RG status (zero copies, one copy, two copies) and 18-*FGF4*RG status (zero copies, one copy, two copies) were independent variables. Reference categories for categorical variables were mixed breed for breed and zero copies for retrogene variables. Weight was centered at population mean body weight (13.0 kg). Difference column indicates the estimated difference in age at surgery for each additional 5 kg of body weight. Univariable analyses were performed first and all independent variables with Wald $p < 0.2$ were tested for inclusion in the multivariable model. A backward elimination approach to model building was used, with variables retained in the final model if $p < 0.05$, or if they were identified as confounding variables (defined as >15% difference in coefficients). Interactions between the main effects were tested. Results were reported as differences in mean ages at surgery and surrounding 95% confidence intervals (CI).

Multivariable logistic regression was used to identify characteristics associated with the presence of disc calcification as determined by the medical record. Breeds were defined as previously. The presence of at least one calcified disc was the dependent variable and sex, age at surgery, body weight, reproductive status, breed (French bulldog, Dachshund, mixed breed, other pure breed), 12-*FGF4*RG status (zero copies, one copy, two copies), and 18-*FGF4*RG status (zero copies, one copy, two copies) were independent variables. Reference categories for categorical variables were mixed breed for breed and zero copies for retrogene variables. Weight was centered at population mean body weight (13.0 kg), and the OR represents each additional 5 kg of body weight. Univariable analyses were performed first and all independent variables with Wald $p < 0.2$ were tested for inclusion in the multivariable model. Both retrogene variables were forced into the model and other variables were retained using a backward elimination, with variables retained in the final model if p < 0.05, or if they were identified as confounding variables (defined as >15% difference in coefficients). Interactions between the main effects were tested. Results were reported as odds ratios (OR) and surrounding 95% CIs. Regression analyses were performed using Stata statistical software (StataCorp. 2015. Stata Statistical Software: Release 14. College Station, TX, USA.

A separate logistic regression among mixed breed dogs to determine characteristics associated with surgery for IVDD was also performed. All mixed breed dogs that had been collected retrospectively were 10 years or older at the time of their last visit and were free of any IVDD diagnosis were used as controls, while retrospectively collected mixed breed dogs that had received decompressive surgery for IVDD were used as cases.

Breed-specific relative risks associated with 12-*FGF4*RG were calculated for breeds that segregated the retrogene (allele frequency <0.5 and >0.05) and had a large enough sample size (number of surgeries >4). Only cases and controls that were sampled retrospectively with respect to IVDD (prior to 1 February 2016) were included. Dogs that were at least 10-years-old with no diagnosis of IVDD as of their last hospital visit were used as controls. 12-*FGF4*RG was treated as a dominant allele for relative risk calculations, where cases and controls with one or two copies of 12-*FGF4*RG were treated equally. Relative risk calculations were conducted in GraphPad Prism.

RESULTS

Allele Frequency

To identify breeds that segregate the *FGF4* retrogenes and determine allele frequencies, 3223 additional dogs from 75 different breeds were genotyped (Supplementary Table S1). 12-*FGF4*RG was identified in at least one dog from 40 different breeds, while 18-*FGF4*RG was found in 32 different breeds. At least one dog from 23 of the 75 breeds tested had both 12-*FGF4*RG and 18-*FGF4*RG.

Description and Genotype of Disc Decompressive Surgical Cases

In order to dissect *FGF4* retrogene contribution to disc herniation, both retrospective and prospective cases requiring disc decompressive surgery were evaluated. Six hundred and twelve unique cases were identified that had undergone spinal cord decompressive surgery and had DNA available for genotyping. Forty-three cases were identified as non-disc related (16 neoplasia, nine infectious or inflammatory, eight vertebral anomalies, four trauma, and six miscellaneous) and were excluded from further analysis. The final dataset contained 569 surgical cases, 272 of which were collected retrospectively (prior to 11 February 2016), and 297 that were collected prospectively (after 11 February 2016). Of the 569 cases, 56 (9.82%) had received two or more decompressive surgeries at the VMTH. The dataset included dogs from 61 different breeds, as well as 127 mixed breed dogs (Table 1). There were 257 (45.3%) neutered males, 227 (40.0%) spayed females, 68 (12.0%) intact males, and 16 (2.8%) intact females. The median age at time of surgery for all cases was 6.4 years (interquartile range, IQR, 4.6–8.8). The median body weight was 8.6 kg (IQR 6.1–14.0). Breed IVDD surgery prevalence among cases collected retrospectively is also presented in Table 1.

All cases were genotyped for 12-*FGF4*RG and 18-*FGF4*RG: 75.2% of the cases had either one or two copies of 12-*FGF4*RG (allele frequency 0.636), and 57.3% had either one or two copies of 18-*FGF4*RG (allele frequency 0.509). Segregation of 12-*FGF4*RG was found in most breeds where it was identified, although several had an extremely high allele frequency, specifically the Dachshund, Beagle, French Bulldog, Pembroke Welsh Corgi, Basset Hound, and Spaniel breeds (Table 1). Several breeds were identified without 12-*FGF4*RG, including Miniature Pinschers, Doberman Pinschers, Rottweilers, and Pomeranians.

Breed	Retrospective Surgery Cases	Percent of Total Surgeries	Total in Repository	Surgery Prevalence in Repository	Prospective Surgery cases	12-*FGF4*RG Frequency	18-*FGF4*RG Frequency	Median Age at Surgery (years)
Dachshund	86	31.62%	221	38.91%	62	0.99	0.99	6.5
Bulldog, French	20	7.35%	81	24.69%	40	0.94	0.01	3.7
Miniature Pinscher	6	2.21%	29	20.69%	0	0.00	0.00	10.3
Pekingese	3	1.10%	17	17.65%	1	0.50	0.88	6.1
Basset Hound	5	1.84%	36	13.89%	1	0.83	1.00	5.5
Beagle	9	3.31%	65	13.85%	8	1.00	0.00	7.9
Welsh Corgi, Pembroke	6	2.21%	54	11.11%	6	0.92	1.00	7.0
Maltese	5	1.84%	65	7.69%	4	0.39	1.00	6.3
Shih Tzu	7	2.57%	92	7.61%	11	0.56	0.92	6.9
Bichon Frise	4	1.47%	59	6.78%	4	0.50	0.75	8.2
Chihuahua	9	3.31%	136	6.62%	16	0.48	0.70	6.0
Pit Bull Terrier	6	2.21%	119	5.04%	5	0.14	0.00	8.0
Cocker Spaniel, American	3	1.10%	61	4.92%	1	1.00	0.00	7.0
Doberman Pinscher	3	1.10%	70	4.29%	6	0.00	0.00	7.8
Rottweiler	4	1.47%	107	3.74%	1	0.00	0.00	5.7
Mixed Breed	46	16.91%	1316	3.50%	81	0.56	0.44	5.5
German Shepherd	5	1.84%	214	2.34%	5	0.05	0.00	6.9
Other	33	12.13%	1430	2.31%	40	0.25	0.23	7.7
Labrador Retriever	12	4.41%	568	2.11%	5	0.03	0.00	8.5
Total	272		4740		297	0.636	0.509	6.4

Table 1 Descriptive statistics for dogs surgically treated for intervertebral disc disease (IVDD). Any breed with fewer than three retrospective surgery cases was included in 'Other'. Dogs from 61 different breeds and 127 mixed breed dogs were defined. The total number of dogs from each breed in the DNA repository is included, and the list is sorted by the breed prevalence of surgical cases.

The *FGF4* retrogenes have been implicated in leg length and height in dogs. Although height data were not collected for these animals, weight is routinely collected on all hospitalized cases and can be used to estimate overall size of the dog. Figure 1 shows the weight and genotype status of cases within each breed. Based on weight, there were 146 miniature Dachshunds and two standard Dachshunds included. A wide distribution of weight and genotype status was observed in the mixed breed dogs (Figure 1). Overall, dogs with one copy of 12-*FGF4*RG (median 8.6 kg, IQR 6.0–12.0) or two copies of 12-*FGF4*RG (median 7.8 kg, IQR 5.9–11.0) weighed significantly less than dogs with zero copies of 12-*FGF4*RG (median 25.4 kg, IQR 7.4–37.0; p < 0.001). There was no significant difference in weight between dogs with one or two copies of 12-*FGF4*RG (p = 0.143). Dogs with one copy of 18-*FGF4*RG (median 6.9 kg, IQR 5.5–8.8) or two copies of 18-*FGF4*RG (median 6.6 kg, IQR 5.4–8.6) weighed significantly less than dogs with zero copies of 18-*FGF4*RG (median 14.3 kg, IQR 10–30.0; p < 0.001). There was no significant difference in weight between dogs with one or two copies of 18-*FGF4*RG (p = 0.234).

This analysis was performed within the mixed breed dog population since purebreds have strong selection for specific sizes. Mixed breed dogs with one copy of 12-*FGF4*RG (median 9.1 kg, IQR 6.6–14.3) or two copies of 12-*FGF4*RG (median 7.1 kg, IQR 5.5–8.8) weighed significantly less than dogs with zero copies of 12-*FGF4*RG (median 22.6 kg, IQR 6.6–32.8; p < 0.001, p = 0.005). Mixed breed dogs with two copies of 12-*FGF4*RG also weighed significantly less than mixed breed dogs with only one copy (p = 0.001). Mixed breed dogs with one copy of 18-*FGF4*RG (median 7.6 kg, IQR 6.1–10) or two copies of 18-*FGF4*RG (median 7 kg, IQR 5.5–9.4) weighed significantly less than dogs with zero copies of 18-*FGF4*RG (median 16 kg, IQR 9–30.4;

$p < 0.001$). There was no significant difference in weight between dogs with one or two copies of 18-*FGF4*RG ($p = 0.159$).

Surgeries were primarily performed at the thoracolumbar region, with 413 (72.6%) surgeries performed between T9 and L6, 143 (25.1%) between C1 and T1 and 13 (2.3%) at the lumbosacral junction. Anatomical localization of surgical procedure as a percentage of all surgeries performed for each genotype category is shown in Supplementary Figure S1. The age at first surgery for breeds with more than four surgical cases compared with 12-*FGF4*RG and 18-*FGF4*RG genotypes is presented in Figure 2.

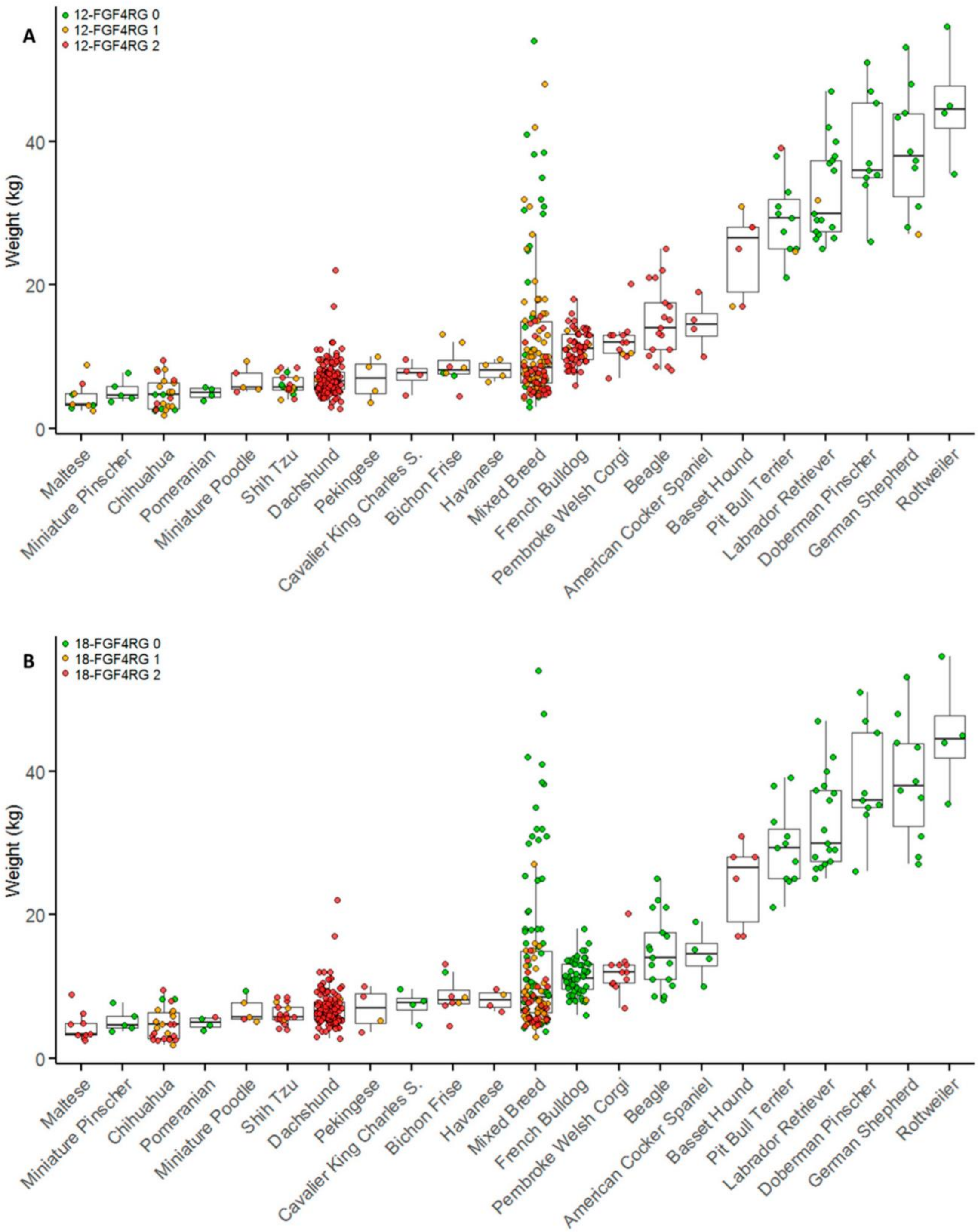
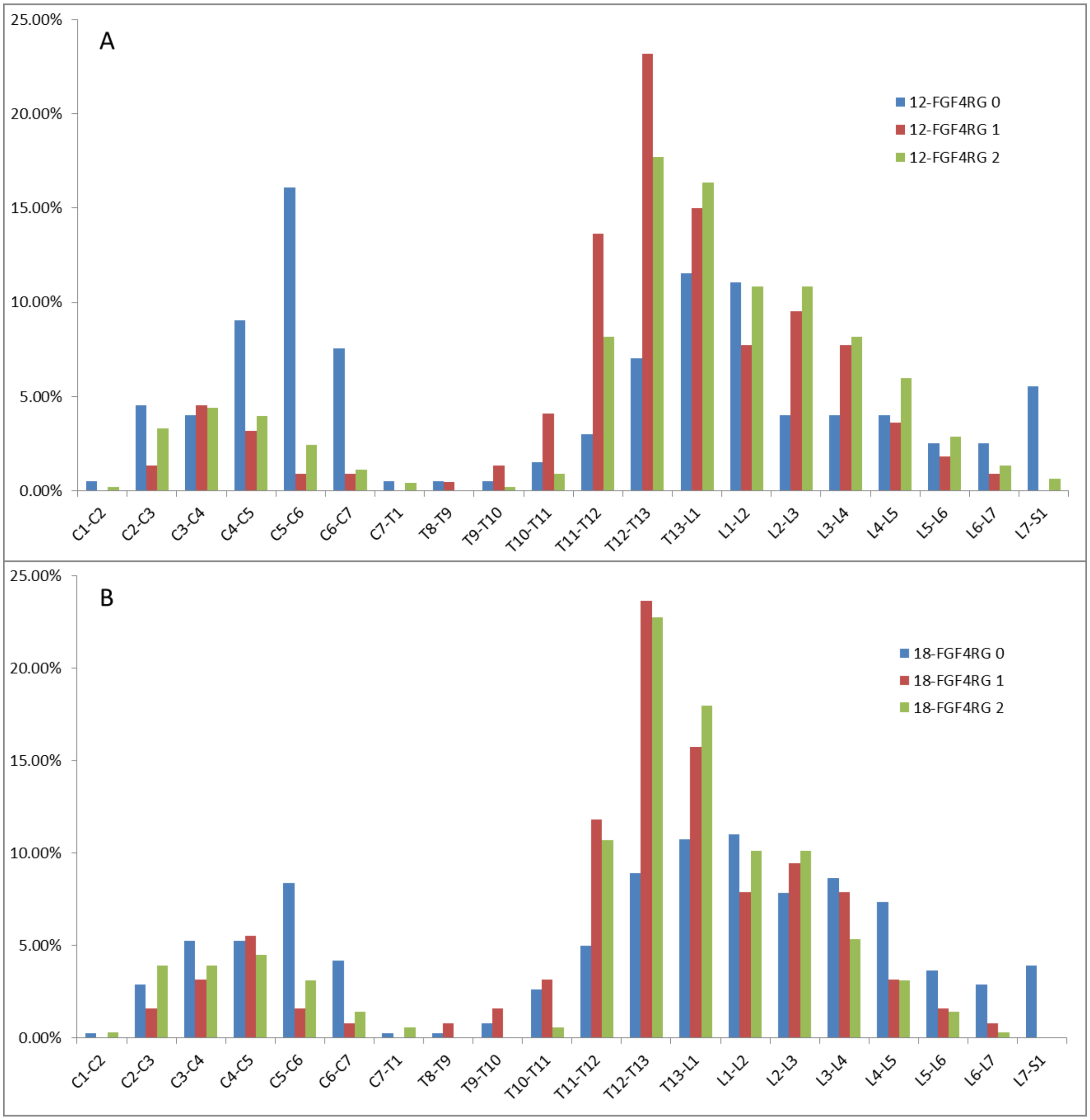


Figure 1 Breed and genotype distribution of surgical IVDD cases by body weight. Breeds with fewer than four cases are not included in this figure. Breeds are plotted in order of ascending median weights and colored by 12-*FGF4*RG genotype (A) and 18-*FGF4*RG genotype (B) n = 510). Red indicates two copies of each retrogene, orange indicates one copy, and green indicates zero copies.



Supplemental Figure S1 Anatomical localization of surgical procedures in dogs

Linear Regression of Age at Surgery

Univariable and multivariable linear regression were performed to determine the effect of each retrogene on age at time of surgery. On multivariable linear regression, 12-*FGF4*RG status and breed category had the largest effect on age at surgery, followed by weight and 18-*FGF4*RG (Table 2). Dogs with one copy of 12-*FGF4*RG (Table 2; p < 0.001) or with two copies of 12-*FGF4*RG (Table 2; p < 0.001) were significantly younger at the time of surgery when compared to dogs with zero copies of 12-*FGF4*RG. There was no significant difference in age at surgery between dogs with one versus two copies of 12-*FGF4*RG (mean difference −4.6 months, 95% CI −12.7 to 3.5; p = 0.266). Adjusting for other variables, French Bulldogs were significantly younger at the time of surgery than mixed breeds (Table 2; p < 0.001), Dachshunds (mean difference −37.3 months, 95% CI −50.3 to −24.4; p < 0.001), and other pure breeds (mean difference −30.1 months, 95% CI −41.8 to −18.3; p < 0.001). Dachshunds were significantly older at surgery when compared to mixed breed dogs (Table 2; p = 0.014). Dogs with one copy of 18-*FGF4*RG (Table 2; p = 0.014) were significantly younger at the time of surgery when compared to dogs with zero copies of 18-*FGF4*RG. There was no significant difference in age at surgery between dogs with zero versus two copies of 18-*FGF4*RG (Table 2; p = 0.239) or between dogs with one versus two copies of 18-*FGF4*RG (mean difference 7.6 months, 95%CI −2.29 to 17.5; p = 0.132).

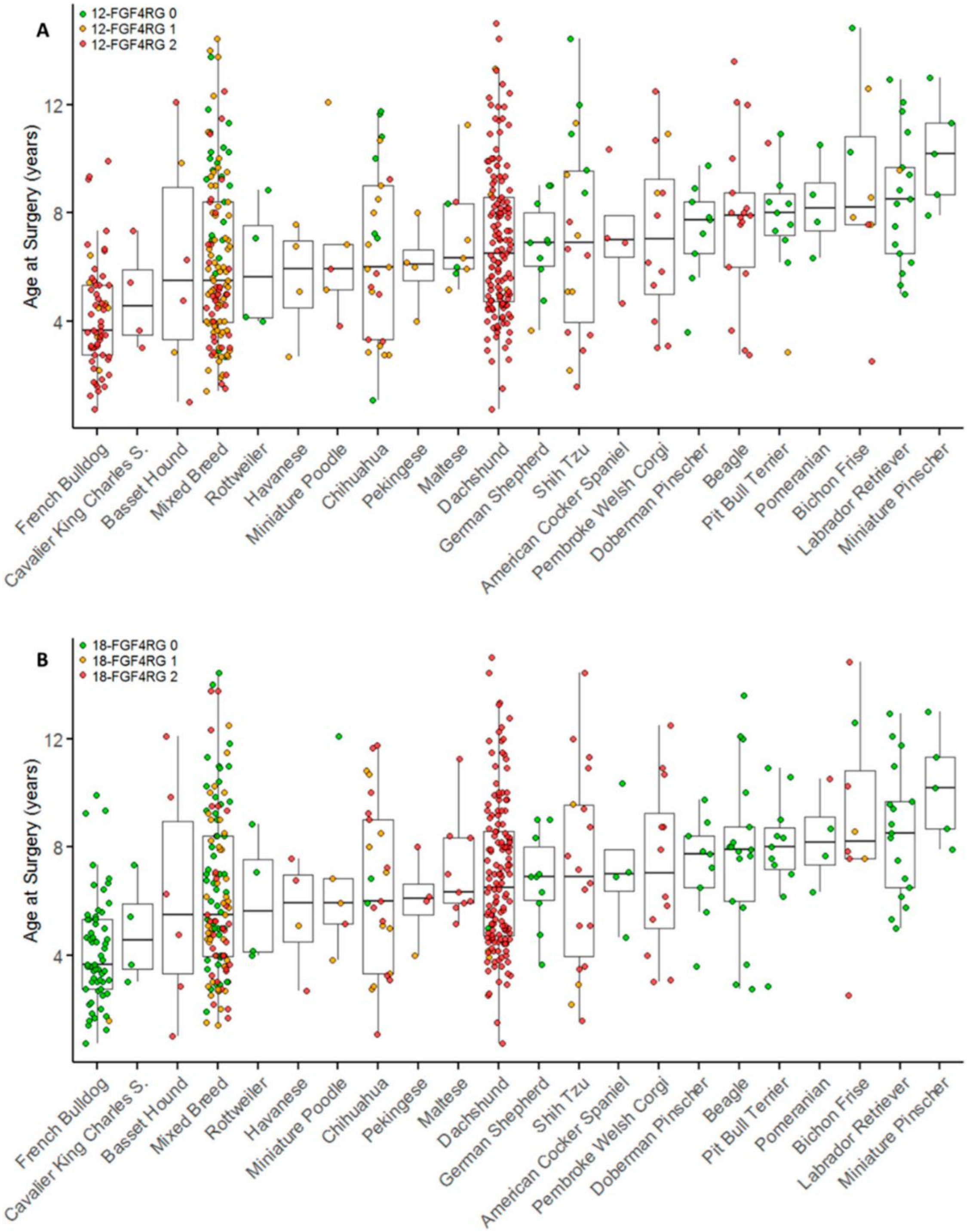


Figure 2 Breed and genotype distribution of surgical IVDD cases by age at surgery. Breeds are plotted in order of ascending median age at surgery. Individuals are colored by 12-*FGF4*RG genotype (**A**) and 18-*FGF4*RG genotype (**B**). Breeds with fewer than four cases are not included (n = 510). Red indicates two copies of each retrogene, orange indicates one copy, and green indicates zero copies.

	Age at IVDD surgery		Univariable (unadjusted) linear regression			Multivariable (adjusted) linear regression		
	Mean	SE	Difference[1]	95% CI	P	Difference[1]	95% CI	P
12-*FGF4*RG								
0 copies	102.6	2.7	Reference	-	<0.001	Reference	-	<0.001
1 copy	73.5	3.0	-29.1	-37.1 to -21.0		-26.6	-35.9 to -17.2	
2 copies	73.0	2.0	-29.5	-36.4 to -22.7		-31.2	-40.5 to -21.8	
Breed								
Mixed breed	74.7	3.2	Reference	-	<0.001	Reference	-	<0.001
Dachshund	82.8	2.7	8.1	0.0 to 16.2		12.9	2.6 to 23.3	
French Bulldog	49.4	3.1	-25.3	-35.8 to -14.8		-24.4	-36.7 to -12.1	
Other purebred	90.1	2.3	15.4	8.0 to 22.7		5.7	-2.1 to 13.4	
Body Weight (5 kg)	80.4	1.5	1.6	0.3 to 2.9	0.017	-2.2	-3.8 to -0.6	0.008
18-*FGF4*RG								
0 copies	82.1	2.4	Reference	-	0.020	Reference	-	0.049
1 copy	69.4	3.9	-12.7	-22.1 to -3.2		-13.1	-23.5 to -2.6	
2 copies	82.1	2.2	0.0	-6.3 to 6.3		-5.5	-14.7 to 3.7	
Sex								
Female	78.0	2.4	Reference	-	0.158	-	-	-
Male	82.3	1.9	4.3	-1.7 to 10.3				
Reproductive								
Status	82.1	3.8	Reference	-	0.639	-	-	-
Intact	80.1	1.6	-2.0	-10.4 to 6.4				
Spayed/Neutered								

Table 2 Linear regression identifying characteristics associated with age at time of surgery. Standard Error (SE); Confidence Interval (CI). [1] Difference column indicates the estimated difference in age (in months) at surgery compared with the reference level. Any positive value indicates later age at surgery, and any negative value indicates earlier age at surgery, versus reference level.

Evaluation of Disc Calcification

Surgical cases were then categorized based on evidence of disc calcification into Group A and Group B. Table 3 shows genotyping results and characteristics for dogs based on group classification. Dogs in Group A (calcified disc material on radiograph or at surgery) were smaller (median 8.1 kg, IQR 6.1–12.8)($p < 0.001$, effect size 14 kg, 95% CI 8.7 to 19.3) and younger (median 5.5 years, IQR 3.9–8.0)($p < 0.001$, effect size 2.7 years, 95% CI 1.8 to 3.6) at the time of surgery compared to the Group B (non-calcified disc material) dogs (median 25 kg, IQR 8.5–38)(median 9 years, IQR 6.4–11). 12–*FGF4*RG was more common in Group A than in Group B; the allele frequency of 12-*FGF4*RG was 0.765 in Group A and 0.149 in Group B ($\chi^2 = 149.9$, $p < 0.001$). 18-*FGF4*RG was also more common in Group A than in Group B with an allele frequency of 0.587 and 0.160 ($\chi^2 = 61.6$, $p < 0.001$). Forty-six cases classified as Group A had zero copies of 12-*FGF4*RG. The breeds represented in this category included Labrador Retriever, Doberman Pinscher, German Shepherd, Pit Bull Terrier, Rottweiler, and Pomeranian, as well as several individual cases from other breeds and mixed breed dogs.

Diagnosis	Count	Median Weight (kg)	Median Age at Surgery (Years)	12-*FGF4*RG Frequency	18-*FGF4*RG Frequency
Group A	378	8.1	5.5	0.765	0.587
Group B	47	25.0	9.0	0.149	0.160

Table 3 IVDD phenotypes among surgical cases. Dogs included in Group A had evidence of calcified intervertebral discs either radiographically or at the time of surgery, while dogs categorized as Group B had no evidence of disc calcification on radiograph or at time of surgery.

Within Group A, dogs with zero copies of 12-*FGF4*RG were significantly older at time of surgery than Group A dogs with one copy of 12-*FGF4*RG (mean difference 22.7 months, 95% CI 10.4 to 35.0; p < 0.001) or two copies of 12-*FGF4*RG (mean difference 20.2 months, 95% CI 9.5 to 30.9; p < 0.001). There was no significant difference in age at surgery between Group A dogs with one versus two copies of 12-*FGF4*RG (p = 0.372). One hundred and thirty Group A dogs had zero copies of 18-*FGF4*RG. There was no significant difference in age at time of surgery between Group A dogs with zero copies of 18-*FGF4*RG and Group A dogs with one copy of 18-*FGF4*RG (p = 0.0.240) or two copies of 18-*FGF4*RG (p = 0.202). However, Group A dogs with one copy 18-*FGF4*RG were significantly younger at time of surgery than group A dogs with two copies of 18-*FGF4*RG (mean difference 11.8 months, 95% CI −22.3 to −1.3; p = 0.026).

Within Group A, dogs with zero copies of 12-*FGF4*RG were significantly older at time of surgery than Group A dogs with one copy of 12-*FGF4*RG (mean difference 22.7 months, 95% CI 10.4 to 35.0; p < 0.001) or two copies of 12-*FGF4*RG (mean difference 20.2 months, 95% CI 9.5 to 30.9; p < 0.001). There was no significant difference in age at surgery between Group A dogs with one versus two copies of 12-*FGF4*RG (p = 0.372). One hundred and thirty Group A dogs had zero copies of 18-*FGF4*RG. There was no significant difference in age at time of surgery between Group A dogs with zero copies of 18-*FGF4*RG and Group A dogs with one copy of 18-*FGF4*RG (p = 0.0.240) or two copies of 18-*FGF4*RG (p = 0.202). However, Group A dogs with one copy 18-*FGF4*RG were significantly younger at time of surgery than group A dogs with two copies of 18-*FGF4*RG (mean difference 11.8 months, 95% CI −22.3 to −1.3; p = 0.026).

Radiographic screenings for calcified discs are used by some breeding programs in an attempt to reduce the incidence of IVDD [21]. Therefore, we compared the characteristics of cases with and without calcified discs defined by radiography. Four hundred and twenty-three of the surgical cases had radiographic reports available for review. Calcified discs were observed significantly (p < 0.001) more often in dogs with two copies of 12-*FGF4*RG (84.8%; 190/224) than in dogs with one copy (63.8%; 51/80) or zero copies (18.5%; 22/119) of 12-*FGF4*RG. Univariable and multivariable logistic regression was performed in order to determine what characteristics had an effect on disc calcification (Table 4). The main contributor to disc calcification was the presence of 12-*FGF4*RG. The odds of disc calcification increased with increasing number of copies of 12-*FGF4*RG. When compared to dogs with one copy of 12-*FGF4*RG, dogs with two copies had 2.5 times greater odds of disc calcification (OR 2.46, 95% CI 1.21 to 5.03; p = 0.013). Other significant contributors to disc calcification were age and breed. The rate of having at least one calcified disc differed significantly (p < 0.001) by breed. At least one calcified disc was observed in 105/116 (90.5%) Dachshunds, 36/51 (70.6%) French Bulldogs, 51/82 (60.2%) mixed breeds, and 71/174 (40.8%) other pure breeds. Dachshunds had significantly higher odds of disc calcification versus other pure breeds (OR 2.53, 95% CI 1.01 to 6.36; p = 0.048) and versus French bulldogs (OR 5.03, 95% CI 1.54 to 16.46; p = 0.007). While copy number of 18-*FGF4*RG was significant on univariable analysis, on multivariable analysis adjusting for other variables there was no significant difference in odds of disc calcification when comparing dogs with zero copies of 18-*FGF4*RG to dogs with one or two copies (Table 4), or when comparing dogs with one versus two copies (OR 1.56, 95% CI 0.58 to 4.20; p = 0.376).

Relative Risk

Since a large number of mixed breed dogs were available from the retrospective collection that had disc decompressive surgery (N = 46) and suitable controls were available (N

= 460: mixed breed dogs aged 10 years or older without history of disc herniation), univariable and multivariable logistic regression was performed to determine which factors are associated with risk of decompressive surgery for IVDD. On univariable logistic regression, 12-*FGF4*RG, 18-*FGF4*RG, and weight were significantly associated with IVDD surgery (Table 5). However, under multivariable analysis, 12-*FGF4*RG status was the only contributor significantly associated with IVDD surgery. There was no significant difference in odds of IVDD surgery when comparing dogs with one or two copies of 12-*FGF4*RG (OR 2.40, 95% CI 0.89 to 6.46; p = 0.083).

Because 12-*FGF4*RG was found to be the main contributor to both age at IVDD surgery and disc calcification among surgical cases and was the only factor significantly associated with IVDD surgery in mixed breed dogs, a breed specific relative risk associated with the presence of 12-*FGF4*RG was then calculated. The high allele frequency of 12-*FGF4*RG among chondrodystrophic breeds such as Dachshunds and French Bulldogs precluded the calculation of a relative risk, and therefore, only breeds that segregated 12-*FGF4*RG (allele frequency < 0.5) and mixed breed dogs were included (Table 6). Only samples that had been collected randomly with respect to IVDD were used for relative risk calculations. Mixed breeds with 12-*FGF4*RG had the highest associated relative risk for IVDD at 15.1.

	Univariable logistic regression			Multivariable logistic regression		
	OR	**95%CI**	**P**	**OR**	**95%CI**	**P**
12-*FGF4*RG						
0 copies	Reference	-	<0.001	Reference	-	<0.001
1 copy	7.75	4.04 to 14.84		6.02	2.75 to 13.18	
2 copies	24.64	13.67 to 44.42		14.82	6.46 to 34.04	
Age at IVDD, year	0.81	0.75 to 0.87	<0.001	0.88	0.80 to 0.95	0.003
Breed						
Mixed breed	Reference	-	<0.001	Reference	-	0.035
Dachshund	5.80	2.70 to 12.47		1.82	0.61 to 5.39	
French Bulldog	1.46	0.69 to 3.09		0.36	0.13 to 1.02	
Other purebred	0.42	0.24 to 0.72		0.72	0.34 to 1.51	
18-*FGF4*RG						
0 copies	Reference	-	<0.001	Reference	-	0.641
1 copy	1.88	0.94 to .374		0.66	0.26 to 1.70	
2 copies	4.37	2.80 to 6.83		1.04	0.49 to 2.20	
Body weight (5 kg)	0.78	0.70 to 0.84	<0.001	-	-	-
Male sex	0.99	0.67 to 1.47	0.960	-	-	-
Spayed or neutered	2.10	1.23 to 3.56	0.006	-	-	-

Table 4 Logistic regression analysis of factors associated with radiographically calcified discs. Results of univariable and multivariable logistic regression analysis of factors associated with having at least one radiographically calcified disc at the time of IVDD surgery in 423 dogs. Odds Ratio (OR).

	Univariable logistic regression			Multivariable logistic regression		
	OR	**95%CI**	**P**	**OR**	**95%CI**	**P**
12-*FGF4*RG						
0 copies	Reference	-	<0.001	Reference	-	<0.001
1 copy	18.38	8.51 to 43.26		18.4	7.44 to 50.26	
2 copies	43.11	15.10 to 129.8		44.2	12.92 to 163.3	
18-*FGF4*RG						
0 copies	Reference	-	<0.001	Reference	-	0.079
1 copy	4.9	2.36 to 10.19		3.1	1.14 to 8.34	
2 copies	7.7	3.25 to 17.46		2.3	0.77 to 6.90	
Body weight (5 kg)	0.7	0.59 to 0.81	<0.001	1.2	0.92 to 1.44	0.22
Male sex	1.33	0.73 to 2.48	0.355	-	-	-

Table 5 Logistic regression analysis of factors associated with IVDD surgery in 506 mixed breed dogs (46 surgical cases and 460 non-surgical). All non-surgical cases were at least 10 years of age at their last visit to the Veterinary Medical Teaching Hospital (VMTH) and had no diagnosis of IVDD.

Breed	Total Dogs	12-*FGF4*RG allele frequency	Relative Risk	95% CI	P Value
Shih Tzu	52	0.25	10.3	1.8 – 62.1	0.005
Bichon Frise	39	0.18	10.0	1.7 – 60.1	0.011
Mixed breed	508	0.10	15.1	7.6 – 29.9	<0.0001
Chihuahua	60	0.10	5.5	1.7 – 18.2	0.008

Table 6 Relative risk associated with 12-*FGF4*RG for IVDD. Dogs included in the IVDD categories were at least 10 years of age as of their last visit to the VMTH. Breed allele frequencies information is taken from dogs with no diagnosis of IVDD (supplemental table 1).

DISCUSSION

Both 12-*FGF4*RG and 18-*FGF4*RG are common across many dog breeds, as they were found in at least one dog in 53% and 43%, respectively, of the 75 breeds tested. Among the dogs treated surgically for IVDD, dogs that carried at least one copy of 12-*FGF4*RG were significantly younger, smaller and more likely to have radiographically calcified discs than the dogs without 12-*FGF4*RG. 12-*FGF4*RG was the only retrogene with a statistically significant effect on disc calcification in an additive manner. Age and breed also had modest effects on disc calcification. Under multivariable logistic regression, 12-*FGF4*RG was the only factor contributing to IVDD surgery in mixed breed dogs. The relative risk for 12-*FGF4*RG varied among segregating breeds, with mixed breed dogs carrying 12-*FGF4*RG the most at risk for IVDD.

The 12-*FGF4*RG has been described in association with the chondrodystrophic phenotype [1], as previously defined clinically and pathologically by Hansen, Braund, and others [10,11]. Consistent with these historical data, presence of the 12-*FGF4*RG in this large data set was found to be significantly associated with four major clinico-pathological phenotypes associated with chondrodystrophy-associated IVDD; specifically, small size, anatomical location of IVDD, early age of onset, and presence of calcification. IVDD in 12-*FGF4*RG carrying dogs was mostly located in the thoracolumbar region, while only 14.2% of the extruded discs localized to the cervical region. Hansen observed that only 15% of the extruded discs in the chondrodystrophic breeds were cervical, and other studies have also shown that thoracolumbar herniation is most common in the chondrodystrophic breeds [18,22,23]. In contrast, among the dogs with zero copies of 12-*FGF4*RG, 42.2% of the extruded discs localized to the cervical region. The caudal cervical region, in particular, was the most affected, although thoracolumbar and lumbosacral IVDD was also present among the dogs with zero copies of 12-*FGF4*RG, results which are consistent with previous studies in non-chondrodystrophic breeds [18,23,24].

While age at time of surgery across all breeds was significantly lower for 12-*FGF4*RG dogs, significant differences were present within specific chondrodystrophic breeds suggestive of additional factors, either genetic or environmental, that may contribute to overall disease presentation. As has been described previously, French Bulldogs had a mean age at time of surgery significantly lower than other breeds, at 4.1 years [25]. Alterations in WNT pathway signaling have been consistently implicated in aging and degeneration of the intervertebral disc [26,27,28]. Down regulation of WNT signaling has been described specifically in chondrodystrophic dog degenerate nucleus pulposus [27], and a downregulating frameshift variant in the WNT pathway gene Dishevelled 2 (DVL2) was also recently identified and associated with screw tail and brachycephaly in Bulldogs, French Bulldogs, and Boston terriers [29]. Bulldogs and Boston Terriers rarely carry the 12-*FGF4*RG and are rarely reported with IVDD; however, it is interesting to speculate that the significantly earlier onset of IVDD in French Bulldogs carrying the 12-*FGF4*RG may be related to additional perturbation of WNT signaling exacerbating *FGF4* retrogene-related pathology.

Dachshunds, the breed with the highest prevalence of disc disease in this study and elsewhere, surprisingly had a significantly older age of onset than mixed breeds. While fewer individuals were identified from other breeds, there is a trend that breeds with a high allele frequency of 12-*FGF4*RG have a later median age of onset of disc herniation. It is possible that breeds with high allele frequencies of 12-*FGF4*RG, such as Beagles and Dachshunds, have undergone additional selection with younger onset affected animals being more likely to be excluded from the breeding pool. It is also possible that owners treat their dogs differently since they are aware of the risk of disc herniation within these breeds. Within breed selection for protective effects could also explain why mixed breed dogs suffer the greatest relative risk for IVDD associated with 12-*FGF4*RG, as they would not benefit from any protective alleles.

The original classification of type I and type II IVDD made by Hansen was done based on histopathological examinations of intervertebral discs and signalment. For this study, in the absence of histopathology, when descriptions of calcified or mineralized disc material were available from the surgery or radiographic reports, the cases were classified as Group A or Group B. Under this classification system, the majority (87%) of dogs in Group A had at least one copy of 12-*FGF4*RG, consistent with a dominant mode of inheritance. Interestingly, 46 out of the 378 cases classified as Group A had no copies of 12-*FGF4*RG. Previous studies have shown that Hansen type I IVDD can occur in non-chondrodystrophic breeds, and, more recently, it has also been shown that the histopathological progression of disc degeneration is similar between the chondrodystrophic and non-chondrodystrophic breeds [17,19,22]. Therefore, it is reasonable to expect that some older non-chondrodystrophic dogs could present with clinical IVDD resembling that seen in the chondrodystrophic breeds. In this study, the Group A dogs with zero copies of 12-*FGF4*RG were on average 22 months older than the Group A dogs with one or two copies of 12-*FGF4*RG. It is unclear whether non-chondrodystrophic dogs presenting with either calcified intervertebral discs or chronic disc protrusions are reflections of a spectrum of presentations within a common phenotype, similar to heterogeneity that may be seen even within chondrodystrophic breeds, or whether additional genetic factors may be present in non-chondrodystrophic dogs, resulting in histopathologically similar end points of disc degeneration with a later onset of disease.

Although the current and previous data support 12-*FGF4*RG as the chondrodystrophy locus, many chondrodystrophic breeds expressing the 12-*FGF4*RG also carry the 18-*FGF4*RG, possibly reflecting prolonged breeding selection for a short limbed phenotype through two different loci (Supplementary Table S1; Table 1). The effect of 18-*FGF4*RG was modest and only significant in one copy and not two. The number of dogs in this category (heterozygous for 18-*FGF4*RG) was low (N = 73) compared to the number of homozygous animals (zero copies 18-*FGF4*RG N = 243; two copies 18-*FGF4*RG N = 253). These differences in allele frequency are reflective of the allele frequency results within breeds reported here and elsewhere [2]. The fact that two copies of 18-*FGF4*RG does not significantly reduce the age of onset of IVDD argues that the significant effect seen in this dataset for one copy is less likely to be biologically significant. Logistic regression using calcification as an outcome also found that the addition of 18-*FGF4*RG did not improve the regression equation compared to 12-*FGF4*RG alone. Multivariable logistic regression also showed that 18-*FGF4*RG did not significantly contribute to IVDD surgery in mixed breed dogs, where 12-*FGF4*RG alone explained the outcome. While this study does not rule out that 18-*FGF4*RG is contributing to the IVDD disease phenotype in minor ways, such as a younger age of onset (13 months) in dogs with one copy of 18-*FGF4*RG, the effect of 12-*FGF4*RG was found to be far greater on all aspects of the disease.

Progression of the nucleus pulposus from normal to a radiographically visible degenerate and mineralized pathology appears to be under the influence of copy number of 12-*FGF4*RG since there is an additive effect. This is in contrast to the effect on age of onset of herniation, which is the same between dogs with one or two copies of 12-*FGF4*RG. In previous work evaluating height in the Nova Scotia Duck Tolling Retriever, the effect of 12-*FGF4*RG was also additive. Beagle neonatal puppies with two copies of 12-*FGF4*RG were previously shown to have 20-fold higher expression of *FGF4* in the intervertebral disc compared to Cane Corso neonatal puppies with no *FGF4* retrogenes [1]. It is possible that continued high expression in the disc in adults is contributing to the rate of mineralization in an additive fashion but that the presence of degeneration is enough to predispose dogs to herniation.

The mechanisms underlying the differential phenotypes associated with the *FGF4* retrogenes remain to be elucidated. The relatively minimal effects of 18-*FGF4*RG on disc disease compared to that of 12-*FGF4*RG could be due to differences in expression patterns between the two *FGF4* retrogenes. Although retrogenes are often regarded as non-expressing pseudogenes due to the frequent lack of defined regulatory elements [30], the 5′ end of the *FGF4* retrogenes contains a highly conserved CpG island which is predicted to function as a promotor [31,32]. Both *FGF4* retrogenes have been shown to be transcriptionally active, although associated clinical phenotypes appear to be different [1,2]. It was previously theorized that expression of 12-*FGF4*RG in the intervertebral disc is based on the chromosomal environment in which it was inserted, since all nearby genes were shown to be expressed in the intervertebral disc [1]. Temporal and tissue-specific expression profiles of the two *FGF4* retrogenes (with associated different clinical phenotypes) may, therefore, be more dependent on the genomic context at the different insertion sites.

Dachshunds are the most commonly affected breed with IVDD, and most of the demographic and breed selection related studies have been conducted relating to the various

Dachshund breed varieties [13,24,33,34]. Dachshunds, similarly, made up the largest portion of this study population, accounting for 31.6% of all retrospectively collected surgical cases, while only making up 4.7% of the DNA repository. While Dachshund variety information was not defined in this study, weights indicated that only two of the 148 surgical cases were Standard Dachshunds, while 146 were miniatures. The allele frequency for 12-*FGF4*RG was high within the breed; however, there was variation depending on where the samples were collected (0.98 in USA/UK samples and 0.94 in Swiss samples), indicating that some populations may be less homozygous than others. It is possible that some varieties of Dachshund segregate 12-*FGF4*RG more than others, and studies on IVDD in Dachshunds outside the USA suggest that this may be true: A previous genetic analysis of disc calcification in Wirehaired Standard Dachshunds registered to the Danish Kennel Club identified an associated region on chromosome 12 near the 12-*FGF4*RG, indicating that the population studied likely segregated 12-*FGF4*RG [35], and wire-haired varieties also appear to be less often clinically affected by IVDD [18,21,33,36].

Several studies have defined incidence of calcification in Dachshunds, the relationship between calcification and risk of clinical disc disease, and the heritability of disc calcification, providing a body of data to try and inform selective breeding to reduce IVDD incidence in the breed [15,21,37]. Interestingly, recent pilot data (Proschowsky and Fredholm, Gravhunden 1-2018 pp 12-13, Magazine for members of the Danish Dachshund Club) evaluating genotype of 12-*FGF4*RG and calcification scores in Dachshunds from Denmark showed an OR of 6.1 for high calcification scores (K6–K15) associated with either one or two copies of 12-*FGF4*RG allele within the wire haired variety. It is possible that data from historical calcification and heritability studies, particularly when segregating varieties were included, may have been a reflection of 12-*FGF4*RG allele frequency within the heterogeneous Dachshund populations studied [15,21,37].

Determining whether disease incidence as opposed to age of onset increases with two versus one copy of the retrogene is important information for the purpose of IVDD breed eradication strategies. Circumstantial evidence may support an increase in disease incidence with two copies given the association between historical calcification scoring (at a defined age) and disease incidence in one breed of dog [13,14], and the correlation of radiographically calcified discs and 12-*FGF4*RG allele frequency in this study. However, age at surgery in this study was highly variable, and data relating to radiographic presence of calcification should be interpreted in this context, since it has been shown that the number of radiographically calcified discs declines with age [21,38]. Selection against higher numbers of disc calcifications through radiographic screening programs in Dachshunds has been implemented in some countries as a way of reducing the incidence of IVDD [14,38], although progress has been limited to date [15]. This may reflect the inherent sensitivity and specificity issues associated with using disc calcification scoring, and its application over potentially heterogeneous 12-*FGF4*RG populations of Dachshund varieties. Among the Dachshund surgical cases in this study, 9.4% had no radiographic evidence of disc calcification. These results are similar to previous retrospective studies in Dachshunds and Pekingese that found that 13% and 17% of cases with disc extrusions had no radiographic calcification [39,40], likely reflecting low sensitivity (0.3–0.6) compared to histopathological assessment, as well as limitations of calcification as the sole marker for “clinically” relevant pathology [41,42].

CONCLUSIONS

Here, we report that 12-*FGF4*RG is both associated with intervertebral disc calcification and with age at time of surgery for IVDD across all affected breeds. The presence of 12-*FGF4*RG increases the risk for disc herniation 5.5–15.1-fold over the background risk in segregating and mixed breeds. Our findings suggest that breeding priorities should be for dogs with fewer copies of 12-*FGF4*RG, so that the allele frequency can be reduced. In breeds with lower allele frequencies of 12-*FGF4*RG, selection against the allele should reduce the incidence of disc disease. Even among breeds with high allele frequencies, genetic screening may be desirable to identify dogs with only one copy of 12-*FGF4*RG so that dogs with zero copies may eventually be bred, significantly improving the overall health of affected breeds.

SUPPLEMENTARY MATERIALS

The following are available online at https://www.mdpi.com/2073-4425/10/6/435/s1, Figure S1: Anatomical localization of surgical procedures in dogs, Table S1: Breed list with genotypes for 12-*FGF4*RG and 18-*FGF4*RG.

CHAPTER 2 REFERENCES

1. Brown, E.A.; Dickinson, P.J.; Mansour, T.; Sturges, B.K.; Aguilar, M.; Young, A.E.; Korff, C.; Lind, J.; Ettinger, C.L.; Varon, S., et al. *FGF4* retrogene on CFA12 is responsible for chondrodystrophy and intervertebral disc disease in dogs. *Proceedings of the National Academy of Sciences of the United States of America* **2017**, *114*, 11476-11481, doi:10.1073/pnas.1709082114.
2. Parker, H.G.; VonHoldt, B.M.; Quignon, P.; Margulies, E.H.; Shao, S.; Mosher, D.S.; Spady, T.C.; Elkahloun, A.; Cargill, M.; Jones, P.G., et al. An expressed *FGF4* retrogene is associated with breed-defining chondrodysplasia in domestic dogs. *Science (New York, N.Y.)* **2009**, *325*, 995-998, doi:10.1126/science.1173275.
3. Sutter, N.B.; Mosher, D.S.; Gray, M.M.; Ostrander, E.A. Morphometrics within dog breeds are highly reproducible and dispute Rensch’s rule. *Mammalian genome : official*

journal of the International Mammalian Genome Society **2008**, *19*, 713-723, doi:10.1007/s00335-008-9153-6.

4. Boulet, A.M.; Capecchi, M.R. Signaling by *FGF4* and FGF8 is required for axial elongation of the mouse embryo. *Developmental biology* **2012**, *371*, 235-245, doi:10.1016/j.ydbio.2012.08.017.
5. Niswander, L.; Martin, G.R. Fgf-4 expression during gastrulation, myogenesis, limb and tooth development in the mouse. *Development (Cambridge, England)* **1992**, *114*, 755-768.
6. Lu, P.; Minowada, G.; Martin, G.R. Increasing <em>*FGF4*</em> expression in the mouse limb bud causes polysyndactyly and rescues the skeletal defects that result from loss of <em>Fgf8</em> function. *Development (Cambridge, England)* **2006**, *133*, 33-42, doi:10.1242/dev.02172.
7. Horton, W.A.; Hall, J.G.; Hecht, J.T. Achondroplasia. *The Lancet* **2007**, *370*, 162-172.
8. Foldynova-Trantirkova, S.; Wilcox, W.R.; Krejci, P. Sixteen years and counting: the current understanding of fibroblast growth factor receptor 3 (FGFR3) signaling in skeletal dysplasias. *Human mutation* **2012**, *33*, 29-41, doi:10.1002/humu.21636.
9. Smolders, L.A.; Bergknut, N.; Grinwis, G.C.; Hagman, R.; Lagerstedt, A.S.; Hazewinkel, H.A.; Tryfonidou, M.A.; Meij, B.P. Intervertebral disc degeneration in the dog. Part 2: chondrodystrophic and non-chondrodystrophic breeds. *Vet J* **2013**, *195*, 292-299, doi:10.1016/j.tvjl.2012.10.011.
10. Hansen, H.J. A pathologic-anatomical study on disc degeneration in dog, with special reference to the so-called enchondrosis intervertebralis. *Acta orthopaedica Scandinavica. Supplementum* **1952**, *11*, 1-117.
11. Braund, K.; Ghosh, P.; Taylor, T.; Larsen, L. Morphological studies of the canine intervertebral disc. The assignment of the beagle to the achondroplastic classification. *Research in veterinary science* **1975**, *19*, 167-172.
12. Jeffery, N.D.; Levine, J.M.; Olby, N.J.; Stein, V.M. Intervertebral disk degeneration in dogs: consequences, diagnosis, treatment, and future directions. *J Vet Intern Med* **2013**, *27*, 1318-1333, doi:10.1111/jvim.12183.
13. Lappalainen, A.K.; Vaittinen, E.; Junnila, J.; Laitinen-Vapaavuori, O. Intervertebral disc disease in Dachshunds radiographically screened for intervertebral disc calcifications. *Acta Veterinaria Scandinavica* **2014**, *56*, 89, doi:10.1186/s13028-014-0089-4.
14. Jensen, V.F.; Beck, S.; Christensen, K.A.; Arnbjerg, J. Quantification of the association between intervertebral disk calcification and disk herniation in Dachshunds. *Journal of the American Veterinary Medical Association* **2008**, *233*, 1090-1095, doi:10.2460/javma.233.7.1090.
15. Lappalainen, A.K.; Maki, K.; Laitinen-Vapaavuori, O. Estimate of heritability and genetic trend of intervertebral disc calcification in Dachshunds in Finland. Acta Vet. Scand. **2015**, 57, 78.
16. Goggin, J.E.; Li, A.S.; Franti, C.E. Canine intervertebral disk disease: characterization by age, sex, breed, and anatomic site of involvement. *American journal of veterinary research* **1970**, *31*, 1687-1692.

17. Cherrone, K.L.; Dewey, C.W.; Coates, J.R.; Bergman, R.L. A retrospective comparison of cervical intervertebral disk disease in nonchondrodystrophic large dogs versus small dogs. *J Am Anim Hosp Assoc* **2004**, *40*, 316-320, doi:10.5326/0400316.
18. Hansen, H.-J. A pathologic-anatomical interpretation of disc degeneration in dogs. *Acta Orthop Scand* **1951**, *20*, 280-293.
19. Hansen, T.; Smolders, L.A.; Tryfonidou, M.A.; Meij, B.P.; Vernooij, J.C.M.; Bergknut, N.; Grinwis, G.C.M. The Myth of Fibroid Degeneration in the Canine Intervertebral Disc: A Histopathological Comparison of Intervertebral Disc Degeneration in Chondrodystrophic and Nonchondrodystrophic Dogs. *Veterinary pathology* **2017**, *54*, 945-952, doi:10.1177/0300985817726834.
20. Mayhew, P.D.; McLear, R.C.; Ziemer, L.S.; Culp, W.T.; Russell, K.N.; Shofer, F.S.; Kapatkin, A.S.; Smith, G.K. Risk factors for recurrence of clinical signs associated with thoracolumbar intervertebral disk herniation in dogs: 229 cases (1994-2000). *Journal of the American Veterinary Medical Association* **2004**, *225*, 1231-1236.
21. Jensen, V.F.; Christensen, K.A. Inheritance of Disc Calcification in the Dachshund. J. Vet. Med. Ser. A **2001**, 47, 331–340.
22. Kranenburg, H.J.; Grinwis, G.C.; Bergknut, N.; Gahrmann, N.; Voorhout, G.; Hazewinkel, H.A.; Meij, B.P. Intervertebral disc disease in dogs - part 2: comparison of clinical, magnetic resonance imaging, and histological findings in 74 surgically treated dogs. *Veterinary journal (London, England : 1997)* **2013**, *195*, 164-171, doi:10.1016/j.tvjl.2012.06.001.
23. Bergknut, N.; Smolders, L.A.; Grinwis, G.C.; Hagman, R.; Lagerstedt, A.S.; Hazewinkel, H.A.; Tryfonidou, M.A.; Meij, B.P. Intervertebral disc degeneration in the dog. Part 1: Anatomy and physiology of the intervertebral disc and characteristics of intervertebral disc degeneration. *Vet J* **2013**, *195*, 282-291, doi:10.1016/j.tvjl.2012.10.024.
24. Bergknut, N.; Egenvall, A.; Hagman, R.; Gustås, P.; Hazewinkel, H.A.; Meij, B.; Lagerstedt, A. Incidence And Mortality Of Diseases Related To Intervertebral Disc Degeneration In A Population Of Over 600,000 Dogs. *Journal of Veterinary Internal Medicine* **2012**, *26*, 847.
25. Mayousse, V.; Desquilbet, L.; Jeandel, A.; Blot, S. Prevalence of neurological disorders in French bulldog: a retrospective study of 343 cases (2002-2016). *BMC Vet Res* **2017**, *13*, 212, doi:10.1186/s12917-017-1132-2.
26. Hiyama, A.; Sakai, D.; Risbud, M.V.; Tanaka, M.; Arai, F.; Abe, K.; Mochida, J. Enhancement of intervertebral disc cell senescence by WNT/beta-catenin signaling-induced matrix metalloproteinase expression. *Arthritis and rheumatism* **2010**, *62*, 3036-3047, doi:10.1002/art.27599.
27. Smolders, L.A.; Meij, B.P.; Onis, D.; Riemers, F.M.; Bergknut, N.; Wubbolts, R.; Grinwis, G.C.; Houweling, M.; Groot Koerkamp, M.J.; van Leenen, D., et al. Gene expression profiling of early intervertebral disc degeneration reveals a down-regulation of canonical Wnt signaling and caveolin-1 expression: implications for development of regenerative strategies. *Arthritis research & therapy* **2013**, *15*, R23, doi:10.1186/ar4157.
28. Dahia, C.L.; Mahoney, E.J.; Durrani, A.A.; Wylie, C. Intercellular signaling pathways active during intervertebral disc growth, differentiation, and aging. *Spine (Phila Pa 1976)* **2009**, *34*, 456-462, doi:10.1097/BRS.0b013e3181913e98.

29. Mansour, T.A.; Lucot, K.; Konopelski, S.E.; Dickinson, P.J.; Sturges, B.K.; Vernau, K.L.; Choi, S.; Stern, J.A.; Thomasy, S.M.; Ho, H.H., et al. Whole genome variant association across 100 dogs identifies a frame shift mutation in DISHEVELLED 2 which contributes to Robinow-like syndrome in Bulldogs and related screw tail dog breeds. *PLoS Genet* **2018**, *14*, e1007850, doi:10.1371/journal.pgen.1007850.
30. Kaessmann, H.; Vinckenbosch, N.; Long, M. RNA-based gene duplication: mechanistic and evolutionary insights. *Nature reviews. Genetics* **2009**, *10*, 19-31, doi:10.1038/nrg2487.
31. Carelli, F.N.; Hayakawa, T.; Go, Y.; Imai, H.; Warnefors, M.; Kaessmann, H. The life history of retrocopies illuminates the evolution of new mammalian genes. *Genome Res* **2016**, *26*, 301-314.
32. Antequera, F. Structure, function and evolution of CpG island promoters. *Cellular and Molecular Life Sciences CMLS* **2003**, *60*, 1647-1658, doi:10.1007/s00018-003-3088-6.
33. Simpson, S.T. Intervertebral disc disease. *The Veterinary clinics of North America. Small animal practice* **1992**, *22*, 889-897.
34. Packer, R.M.; Seath, I.J.; O'Neill, D.G.; De Decker, S.; Volk, H.A. DachsLife 2015: an investigation of lifestyle associations with the risk of intervertebral disc disease in Dachshunds. *Canine genetics and epidemiology* **2016**, *3*, 8, doi:10.1186/s40575-016-0039-8.
35. Mogensen, M.S.; Karlskov-Mortensen, P.; Proschowsky, H.F.; Lingaas, F.; Lappalainen, A.; Lohi, H.; Jensen, V.F.; Fredholm, M. Genome-Wide Association Study in Dachshund: Identification of a Major Locus Affecting Intervertebral Disc Calcification. *Journal of Heredity* **2011**, *102*, S81-S86, doi:10.1093/jhered/esr021.
36. Packer, R.M.; Hendricks, A.; Volk, H.A.; Shihab, N.K.; Burn, C.C. How long and low can you go? Effect of conformation on the risk of thoracolumbar intervertebral disc extrusion in domestic dogs. PLoS ONE **2013**, 8, e69650.
37. Stigen, O.; Christensen, K. Calcification of intervertebral discs in the dachshund: an estimation of heritability. *Acta Vet Scand* **1993**, *34*, 357-361.
38. Stigen, Ø. Calcification of intervertebral discs in the dachshund: A radiographic study of 115 dogs at 1 and 5 years of age. Acta Vet. Scand. **1996**, 37, 229–237.
39. Rohdin, C.; Jeserevic, J.; Viitmaa, R.; Cizinauskas, S. Prevalence of radiographic detectable intervertebral disc calcifications in Dachshunds surgically treated for disc extrusion. *Acta Vet Scand* **2010**, *52*, 24, doi:10.1186/1751-0147-52-24.
40. Chai, O.; Harrosh, T.; Bdolah-Avram, T.; Mazaki-Tovi, M.; Shamir, M.H. Characteristics of and risk factors for intervertebral disk extrusions in Pekingese. *Journal of the American Veterinary Medical Association* **2018**, *252*, 846-851, doi:10.2460/javma.252.7.846.
41. Stigen, O.; Kolbjornsen, O. Calcification of intervertebral discs in the dachshund: a radiographic and histopathologic study of 20 dogs. *Acta Vet Scand* **2007**, *49*, 39, doi:10.1186/1751-0147-49-39.
42. Stigen, Ø.; Ciasca, T.; Kolbjørnsen, Ø. Calcification of extruded intervertebral discs in dachshunds: a radiographic, computed tomographic and histopathological study of 25 cases. *Acta Veterinaria Scandinavica* **2019**, *61*, 13, doi:10.1186/s13028-019-0448-2.

CHAPTER 3: Multiple *FGF4* Retrocopies Recently Derived within Canids

Kevin Batcher [1], Peter Dickinson [2], Kimberly Maciejczyk [1], Kristin Brzeski [3], Sheida Hadji Rasouliha [4], Anna Letko [4], Cord Drögemüller [4], Tosso Leeb [4], and Danika Bannasch [1,*]

[1]Department of Population Health and Reproduction, University of California-Davis, USA

[2]Department of Surgical and Radiological Sciences, University of California-Davis, USA

[3]College of Forest Resources and Environmental Science, Michigan Technological University, Houghton, MI 49931, USA;

[4]Institute of Genetics, Vetsuisse Faculty, University of Bern, 3012 Bern, Switzerland;

[*]Correspondence: dlbannasch@ucdavis.edu

Genes 2020, 11(8), 839; https://doi.org/10.3390/genes11080839

Received: 30 June 2020; Accepted: 21 July 2020; Published: 23 July 2020

AUTHOR CONTRIBUTIONS

Conceptualization, D.B.; Data curation, **K.B.**, P.D., K.M., and D.B.; Formal analysis, **K.B.**; Funding acquisition, D.B.; Investigation, **K.B.**, K.M., S.H.R., A.L., C.D., T.L., and D.B.; Methodology, **K.B.** and D.B.; Project administration, D.B.; Resources, K.B. (Kristin Brzeski) and D.B.; Supervision, D.B.; Visualization, **K.B.**; Writing—original draft, **K.B.** and D.B.; Writing—review & editing, **K.B.**, P.D., T.L., and D.B.

ABSTRACT

Two transcribed retrocopies of the fibroblast growth factor 4 (*FGF4*) gene have previously been described in the domestic dog. An *FGF4* retrocopy on chr18 is associated with disproportionate dwarfism, while an *FGF4* retrocopy on chr12 is associated with both disproportionate dwarfism and intervertebral disc disease (IVDD). In this study, whole-genome sequencing data were queried to identify other *FGF4* retrocopies that could be contributing to phenotypic diversity in canids. Additionally, dogs with surgically confirmed IVDD were assayed for novel *FGF4* retrocopies. Five additional and distinct *FGF4* retrocopies were identified in canids including a copy unique to red wolves (Canis rufus). The *FGF4* retrocopies identified in domestic dogs were identical to domestic dog *FGF4* haplotypes, which are distinct from modern wolf *FGF4* haplotypes, indicating that these retrotransposition events likely occurred after domestication. The identification of multiple, full length *FGF4* retrocopies with open reading frames in canids indicates that gene retrotransposition events occur much more frequently than previously thought and provide a mechanism for continued genetic and phenotypic diversity in canids.

INTRODUCTION

Gene retrocopies, often previously referred to as processed pseudogenes, are formed through the mRNA-mediated gene duplication of cellular gene transcripts [1]. In mammals, this process is mediated by long interspersed nuclear elements 1 (L1) acting in trans [2,3]. L1s are the only autonomous, retrotransposable elements still active today in mammals, and while over 100 active, full-length L1s have been identified in humans, dogs have more than 200 active L1s [4]. L1 insertion is accomplished through target primed reverse transcription, a process that results in duplication of genomic DNA at the insertion site, called a target site duplication (TSD) [5]. Because gene retrocopies are formed from processed mRNA, they also lack introns and contain a polyA tail, features that distinguish them from their parental gene. Although retrocopy insertions can occur anywhere in the genome, the L1 machinery shows a preference for the TTAAAA consensus sequence as an insertion site [3,6].

Retrocopies are more likely to come from highly expressed genes [7], with some genes having over a dozen retrocopies [8]. Most of the retrocopies present in any given reference genome arose millions of years ago and have since acquired numerous sequence variants differentiating them from their parent genes [8]. While many retrocopies have also lost their open reading frame (ORF), L1 is still actively producing retrocopies with intact ORFs in mammalian genomes. Hundreds of recent, polymorphic gene retrocopies have been reported in humans and mice [9,10,11]. Notably, polymorphic retrocopies were more common in mice than in humans, consistent with mice having more active L1s [9]. A recent survey of retrocopies in the canfam3 reference genome has identified over 3000 retrocopies, 476 of which were intact [12], and several gene retrocopies have also been identified on the canine Y chromosome [13]. However, it is still unclear how many recent, polymorphic retrocopies are in canids that are not present in the canfam3 reference genome.

Two expressed, polymorphic fibroblast growth factor 4 gene (*FGF4*) retrocopies have been described previously in dogs on chr18 [14] and chr12 [15], referred to as *FGF4*L1 (CFA18) and *FGF4*L2 (CFA12) in this study. Both *FGF4*L1 and *FGF4*L2 are associated with forms of disproportionate dwarfism that are common across many popular dog breeds, and there is evidence that these genes have been under selection owing their strong phenotypic effects [16,17]. *FGF4*L2 has also been associated with canine chondrodystrophy, a disorder characterized by premature degeneration of the intervertebral discs, which predisposes affected dogs to intervertebral disc herniation [18]. However, chondroid disc degeneration can also be seen in dogs without *FGF4*L2, indicating the possibility of alternative risk loci for the disorder [19].

Because two recent, functional *FGF4* retrocopies had already been described in dogs, we hypothesized that more *FGF4* retrocopies could be segregated across dog breeds, which may contribute to limb morphology and/or disc disease. Previous *FGF4* retrocopies were identified following genome-wide associations for disproportionate dwarfism. In the current study, two approaches were utilized to identify additional *FGF4* retrocopies in dogs. First, discordant read mapping of paired-end Illumina reads from publicly available whole-genome sequence data was used to identify additional polymorphic *FGF4* retrocopies in canid genomes that would not be identified by common variant calling techniques. The second approach was to perform exon to exon polymerase chain reaction (PCR) to identify the presence of an intron-less retrocopy, followed by inverse PCR to identify the site of insertion. Five additional *FGF4* retrocopies were then identified, sequenced, and characterized.

MATERIALS AND METHODS

FGF4 Retrocopy Discovery in Whole-Genome Sequence Data

Data from six BioProjects (PRJNA448733, PRJEB16012, PRJNA288568, PRJNA377155, PRJEB20635, and PRJEB32865) were utilized for this approach [20,21,22,23,24,25]. This included 1125 individuals from 160 different breeds, as well as 101 indigenous dogs, 141 wolves, and 3 coyotes (Supplemental Table S1). The canine reference genome, CanFam3, does not contain any full length *FGF4* retrocopies, and thus all reads coming from *FGF4* retrocopies are aligned to the parental *FGF4* gene locus. To identify any such novel *FGF4* retrocopies, aligned paired end

Illumina sequence data in the region surrounding the *FGF4* gene (CanFam3 chr18:48,412,000–48,418,000) were downloaded from the Sequence Read Archive and analyzed. Sequencing files were viewed in Integrative Genomics Viewer [26]. Discordant paired end reads mapping from exon to exon (Supplemental Figure S1 shown in red) are indicative of the presence of an *FGF4* gene retrocopy somewhere in the genome as retrocopies lack introns, while discordant paired end reads, wherein one mate maps to the *FGF4* gene locus and the other mate maps to another region of the genome, are indicative of the putative insertion site for an *FGF4* retrocopy (Supplemental Figure S1 shown in teal). The presence of both forms of discordant reads was used as an indication of an *FGF4* retrocopy insertion.

***FGF4* Retrocopy Discovery in Clinical Cases**

Whole-genome sequence data were not available for any of the individuals treated by surgical decompression for presumed IVDD. Therefore, a molecular approach was developed to test for novel *FGF4* retrocopies in DNA samples. A total of 164 surgical cases that were previously shown to have 0 copies of *FGF4*L1 and *FGF4*L2 were used for novel *FGF4* retrocopy discovery [19]. The presence of *FGF4* retrocopies was tested by amplifying the region between exon 1 and exon 3 of *FGF4* (Supplemental Table S2). The identification of a reduced size, intron-less product indicates the presence of an *FGF4* retrocopy (Supplemental Figure S2). When an individual tested positive for an *FGF4* retrocopy and negative for the two known *FGF4* retrocopy insertions, inverse PCR [27] was then performed to identify the insertion site of the *FGF4* retrocopy. For inverse PCR, 1 µg of genomic DNA was digested with the MboI restriction enzyme according to the manufacturer's instructions (New England Biolabs, Ipswich, MA, USA), and fragments were then circularized by ligation at final concentrations varying between 1 and

10 ng/µL using T4 DNA ligase according to the manufacturer's instructions for sticky end ligation (New England Biolabs, Ipswich, MA, USA). A set of inverted primers were designed that amplified circular DNA fragments containing the 5' end of the *FGF4* retrocopy insertion (Supplemental Table S2). PCR was then performed using LongAmp Taq DNA polymerase according to the manufacturer's instructions (New England Biolabs, Ipswich, MA, USA). PCR products were visualized by gel electrophoresis and isolated for Sanger sequencing using a QIAquick Gel Extraction Kit (Qiagen, Valencia, CA, USA). All PCR primers were designed using primer3 (http://bioinfo.ut.ee/primer3/) [28].

Sequencing and Comparitive Analysis of *FGF4* Retrocopies

All canine DNA samples used for retrocopy sequencing and subsequent population genotyping of the *FGF4* retrocopies came from the Bannasch Canine Repository and were obtained under UC Davis Animal Care and Use Committee protocol 18,561 [19] (Supplemental Table S3). Red wolf tissue samples for DNA extraction were obtained with the approval of the United States Fish and Wildlife Services. PCR primers were designed to flank the insertion sites of *FGF4* retrocopies identified via discordant paired end reads or inverse PCR (Supplemental Table S1). Entire retrocopy insertions were then amplified through PCR using LongAmp Taq DNA polymerase according to the manufacturer's instructions (New England Biolabs, Ipswich, MA, USA). The full sequence of each retrocopy was obtained through Sanger sequencing using a series of internal primers (Supplemental Table S2). Variants in the parental *FGF4* gene were observed in a dataset of 722 canids to determine which single-nucleotide variants (SNVs) were unique to *FGF4* retrocopies [20].

Conservation at Insertion Sites

Evolutionarily conserved elements (ECR) near the *FGF4* retrocopy insertion sites were defined using the 4-Way Multiz Alignment & Conservation track for CanFam2 on the UCSC genome browser [29], which shows a measure of evolutionary conservation between dog, human, mouse, and rat genomes using Multiz alignment [30].

Population Genotyping

Breeds were selected for population genotyping based on the breeds in which they were identified, excluding breeds where whole-genome sequencing data indicated they did not contain any *FGF4* retrocopies. PCR assays utilizing three primers per assay were designed for each *FGF4* retrocopy for population genotyping, as previously described [15]. In each assay, a shared internal primer at the 3' end of the *FGF4* retrocopy produces a different size amplicon when the retrocopy is present (Supplemental Table S1).

Height Measurements

Height was measured in selected cases to determine if *FGF4* retrocopies had any effect on height. All height measurements were performed by the same individual using a standard wicket (height measuring device for dogs). Multivariable linear regression was performed in R studio using the generalized linear model function with sex and *FGF4* genotype included to identify any association with height.

RESULTS

FGF4 Retrocopy Discovery from Whole-Genome Sequence Data

In addition to the two known *FGF4* retrocopies, *FGF4*L1 and *FGF4*L2, evidence for four additional *FGF4* retrocopies in canids was observed in the whole-genome sequence dataset (Table 1). The novel *FGF4* retrocopies include a copy on CFA27 (*FGF4*L3) seen in three Nova Scotia Duck Tolling Retrievers (NSDTR); a copy on CFA22 (*FGF4*L4) seen in two Norwich Terriers; a copy on CFA13 (*FGF4*L5) seen in a Belgian Malinois and a Dutch Shepherd; and a copy on CFA36 (*FGF4*L6) seen in two red wolves. Sequence read archive (SRA) accession numbers for these individuals are available in Supplemental Table S4.

Discordant reads were also observed at the 3′ end of the *FGF4* gene locus aligning to a partial *FGF4* retrocopy insertion in the CanFam3 reference genome at chr7:68,372,263–68,373,442. To confirm whether this was a real *FGF4* retrocopy fragment or a mistake in the reference assembly, primers were designed flanking the insertion and the region was amplified in six Boxers. Five were heterozygous for the CFA7 partial *FGF4* retrocopy insertion, and Sanger sequencing confirmed the sequence matched the reference genome. Because this retrocopy only contains the 3′ UTR of the gene and has no ORF, this retrocopy fragment was not considered for further analysis.

***FGF4* Retrocopy Discovery in Dogs Treated for Disc Disease**

A surgically treated population of 164 individuals that had neither *FGF4*L1 nor *FGF4*L2 was then tested for the presence of any *FGF4* retrocopy using an exon–exon PCR assay. Four of these individuals tested positive for the presence of an *FGF4* retrocopy. These samples were first tested for the other newly discovered *FGF4* retrocopies. One sample, a Shetland Sheepdog, was heterozygous for *FGF4*L5. The medical history of this individual indicates that it received a hemilaminectomy to treat a mass that was not disc-related.

Name	Location	Sequence at Insertion Site	Strand	G/C	ECR	Method
***FGF4*L1**	Chr18:20,443,703–20,443,735	ACCATGAAAT**AAGTCAGACAGAG**AAAGACAAGT	+	36.4	2	GWAS[14]
***FGF4*L2**	Chr12:33,710,158–33,710,188	ATTCCTATTC**AAGTGCTTTGA**ACTCTTCAAA	+	32.3	1	GWAS[15]
***FGF4*L3**	Chr27:24,834,102–24,834,135	TGAGAATACT**CAGGGACCATTTCT**ATTGACTTTT	-	35.3	0	DRM
***FGF4*L4**	Chr22:47,761,852–47,761,888	TGTCTTTGTC**AAGAATATTCTGGTTGT**GAGTAATAGA	+	32.4	2	DRM
***FGF4*L5**	Chr13:28,020,009–28,020,044	GCAGTTTCTT**AAAACTTAGAGGAACA**AAGTAGCTTG	+	36.1	6	DRM
***FGF4*L6**	Chr36:11,456,175–11,456,208	AAAGCATTAA**TTACCAAAGTACTA**TTTCATAACT	+	23.5	1	DRM
***FGF4*L7**	Chr13:25,020,597–25,020,632	GAATCGTGTT**TAAGAAGGGGTGGTAT**GACTTGCCCT	-	47.2	3	Inverse PCR

Table 1 Genomic sequence at fibroblast growth factor 4 gene (*FGF4*) retrocopy insertion sites in canids. Target site duplications (TSDs) are in bold and underlined. Ten bases upstream and downstream from the TSD are included, as well as the strand orientation of the retrocopy, G/C content of the region, and evolutionarily conserved elements (ECR) within 2.5 kb of the insertion site. *FGF4* retrocopies were identified by GWAS, discordant read mapping (DRM), and inverse PCR.

The three remaining dogs were all Pit Bull Terrier mixes that had received hemilaminectomies for IVDD at relatively young ages (age at time of surgery of 3, 5, and 8 years), and none of the newly discovered or previously defined *FGF4* retrocopies were present in these individuals, indicating they contained a novel *FGF4* retrocopy. Inverse PCR was then performed to discover the insertion site of the novel *FGF4* retrocopies in these individuals, which was on CFA13 (*FGF4*L7) at approximately CFA13:25,020,600. The three dogs were all heterozygous for *FGF4*L7, and Sanger sequencing revealed that *FGF4*L7 is a full length *FGF4* retrocopy.

Comparative Analysis of *FGF4* Retrocopies

Novel *FGF4* retrocopies were confirmed through PCR amplification and sequencing. The genomic location for the *FGF4* retrocopies, their TSD, and genomic sequence surrounding the TSD are shown in Table 1. Exact TSD length varied from 11 to 17 bases, with a median of 15 bp. The loosely conserved L1 consensus insertion site sequence (TTAAAA) was only observed at the *FGF4*L5 insertion site. Insertion sites for 6/7 of the *FGF4* retrocopies had a low G/C content compared with the Canfam3 average of 41.3% (Table 1). All *FGF4* retrocopies inserted into intergenic regions of the genome. Both *FGF4*L1 and *FGF4*L3 inserted into a LINE element, while *FGF4*L4 inserted into a long terminal repeat (LTR). The number of evolutionarily conserved regions within 2.5 kb of the insertion sites is also reported in Table 1.

Comparison of each *FGF4* retrocopy to the parental *FGF4* sequence showed that each novel copy has a fully conserved ORF (Figure 1). The 5’ UTR of *FGF4*L3 is truncated by 112 bp compared with the other retrocopies, and the 3’ UTR in both *FGF4*L1 and *FGF4*L4 is truncated by 530 bp and 83 bp. No single-nucleotide variants (SNVs) were identified in either the ORF or the 5′ UTR of any of the retrocopies. However, six SNVs were identified in the 3′ UTR that differed from the reference genome *FGF4* gene sequence. Analysis of a whole-genome sequencing variant calling dataset from 722 canids indicated that these SNVs are also present at the parental *FGF4* gene (Supplemental Table S5). Therefore, no SNV specific to any of the dog *FGF4* retrocopies was identified. Rather, the differences between *FGF4* retrocopies are owing to different haplotypes of the parental *FGF4* gene from which the retrocopies formed. Notably, the 3′ end of the parental *FGF4* gene in wolves contains several SNV not identified in any domestic dogs (Supplemental Table S5).

The red wolf *FGF4*L6 3′ UTR sequence contained two single nucleotide indels not observed in any domestic dog *FGF4* sequences: a deletion (CFA18:48,415,685delA) and an insertion (CFA18:48,416,575_48,416,576insA). These indels were not identified in any canids other than the two red wolves in a whole-genome sequencing variant calling dataset, which included 46 gray wolves. The parental *FGF4* locus was sequenced in seven red wolves to determine whether these variants also exist in the parental gene in red wolves or if they are unique to the retrocopy. While three individuals were heterozygous for the CFA18:48,416,575T>TA insertion at the parental *FGF4* gene, CFA18:48,415,685CA C was not identified in any of the parental *FGF4* sequences, indicating this variant may have occurred after retrotransposition and may be unique to *FGF4*L6.

Population Genotyping of Novel *FGF4* Retrocopies

A targeted population genotyping approach based on the breeds in which *FGF4* retrocopies were identified was utilized to determine allele frequencies of the *FGF4* retrocopies. A complete list of *FGF4* retrocopy genotyping results is available in Supplemental Table S3. *FGF4*L3 was only observed in the NSDTR breed in the whole-genome sequencing dataset, and was thus tested in 100 randomly selected NSDTR. The allele frequency of *FGF4*L3 was 8.5% in the NSDTR.

*FGF4*L4 had an allele frequency of 16.7% in Norwich Terriers (n = 30). Further testing for *FGF4*L4 in related terrier breeds identified this retrocopy in Norfolk Terriers (n = 10, allele frequency 30%), Border Terriers (n = 32, allele frequency 71.9%), and Skye Terriers (n = 10, allele frequency 5%). Given the previous association of *FGF4* retrogenes with skeletal dysplasia, *FGF4*L4 genotype was also compared to height at the withers in 24 Border Terriers using multiple linear regression. The regression analysis identified no significant association between *FGF4*L4 and height in Border Terriers (p = 0.877, n = 24), although only one homozygous wild type individual was included (Supplemental Figure S3).

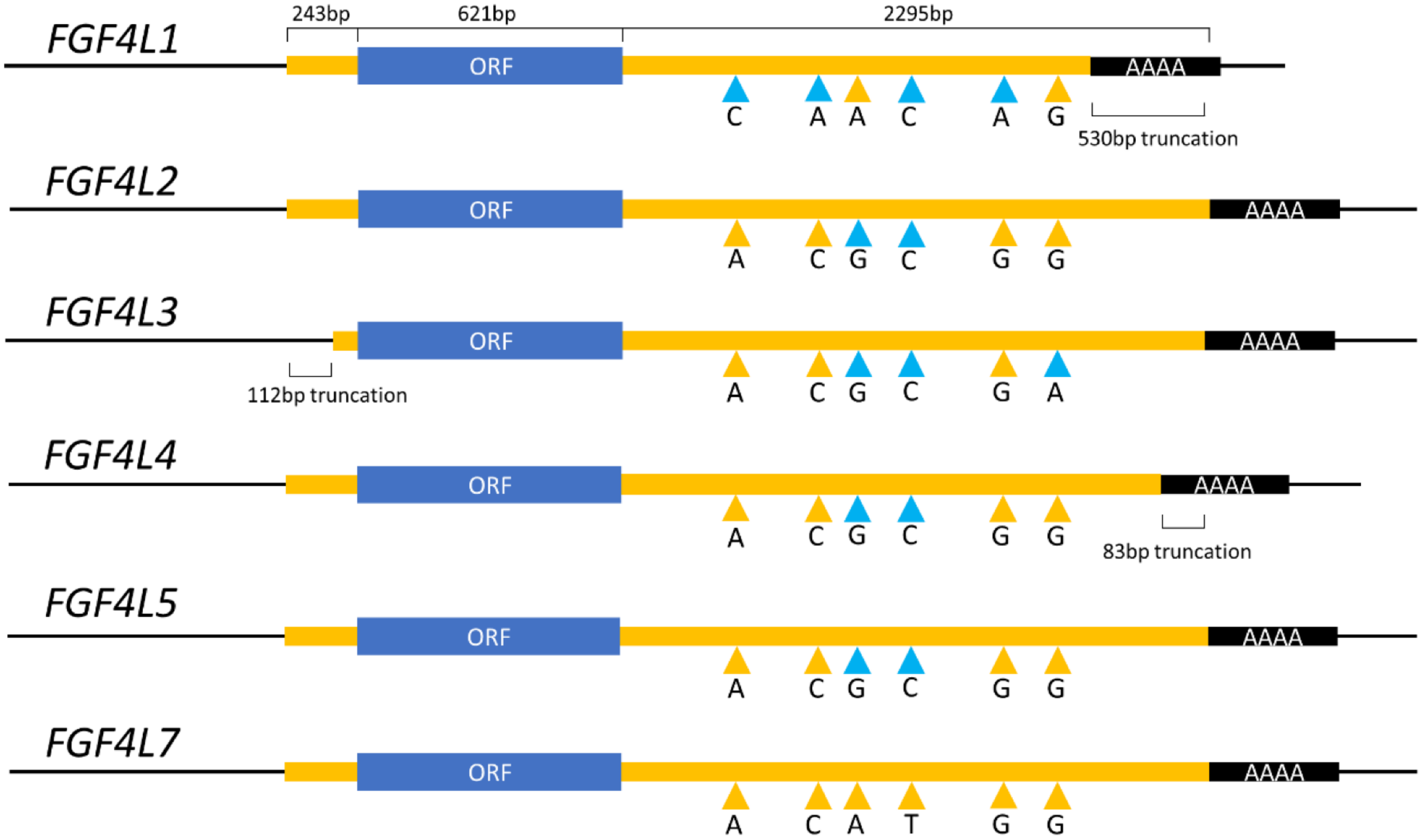


Figure 1 Comparison of the six full length *FGF4* retrocopies identified in domestic dogs. From left to right, the letters with colored arrowheads represent variants within the 3’ UTR of the *FGF4* gene at genomic locations CFA18:48,415,400A>C; CFA18:48,415,405C>A; CFA18:48,415,585A>G; CFA18:48,415,608T>C; CFA18:48,415,661G>A; and CFA18:48,416,537G>A. SNV colored in blue represent non-reference alleles. ORF, open reading frame.

*FGF4*L5 was not identified in any other Shetland Sheepdogs (n = 58) or Belgian Malinois (n = 14). Australian Shepherds (n = 19) and Anatolian Shepherd dogs (n = 5) also tested negative for *FGF4*L5. No additional Dutch Shepherd samples were available for population genotyping of *FGF4*L5 in the breed. *FGF4*L6 was tested in 14 red wolf samples, 5 of which were heterozygous (allele frequency 15.6%).

Pit Bull Terriers and Pit Terrier Mixes were then tested for *FGF4*L7 (n = 201), and all tested negative for the retrocopy. Because *FGF4*L7 was identified in dogs treated for IVDD and could be contributing to the disorder, all mixed breed dogs from the Bannasch Canine Repository that had been treated surgically for IVDD were also tested for *FGF4*L7 (n = 55), all of which tested negative. However, two discordant reads mapping to the *FGF4*L7 were subsequently identified in the whole-genome sequence data of a single Chinese village dog (SRR7107669). Several breeds developed in Asia were then tested for *FGF4*L7, including Chow Chow (n = 22), Pugs (n = 9), Pekingese (n = 8), and Tibetan terriers (n = 6), none of which tested positive. However, *FGF4*L7 was identified in Chinese Shar-Pei (n = 22, allele frequency 34.1%).

DISCUSSION

Multiple recently transposed *FGF4* retrocopies exist in canids in addition to the previously identified *FGF4*L1 and *FGF4*L2. Novel retrocopies appear to be breed or breed group specific, contain intact ORFs, and have not accrued mutations that differentiate them from parental *FGF4* gene haplotypes. The *FGF4* retrocopies were retrotransposed from *FGF4* genes with distinct haplotypes, indicating that the same copy has not been retrotransposed multiple times. It is unclear whether any of these novel copies are expressed retrogenes, or in what tissue types they could be expressed. *FGF4*L7 was identified in three dogs treated surgically for IVDD, however, the significance relative to IVDD is unknown. The majority of IVDD surgical cases examined in this study that were not explained by *FGF4*L2 were found to have no *FGF4* retrocopies, indicating that there are risk factors other than *FGF4* retrocopies that predispose dogs to IVDD.

Evidence for the expression of both *FGF4*L1 and *FGF4*L2 has indicated that the *FGF4* retrocopies are capable of expression [14,15]. The 5′ end of the *FGF4* gene is GC-rich and contains many evolutionarily conserved transcription factor binding sites that were previously hypothesized to be conducive towards expression of the retrocopies [15]. Thus, the 5′ end truncation of the *FGF4*L3 retrocopy likely affects expression. It has also been reported that the expression of retrocopies is highly dependent on the genomic environment of the insertion sites [31]. Both *FGF4*L1 and *FGF4*L2 have inserted into regions containing nearby evolutionarily conserved elements (ECRs). Similarly, ECRs at all but one of the *FGF4* retrocopy insertion sites may be conducive towards expression. The different genomic context at the insertion sites for *FGF4*L1 and *FGF4*L2 could also explain the different phenotypes between the copies. If expressed, the novel *FGF4* retrocopies may show unique expression profiles, resulting in phenotypic associations other than height and IVDD. *FGF4* is involved in several cellular processes including cell growth, tissue repair, tumor growth and invasion, and is also a well-known proto-oncogene [32,33].

Although *FGF4*L2 has been shown to have a major association with IVDD [15,19,34], clinically significant IVDD has been reported in dogs lacking the *FGF4*L2 retrogene, implicating alternate causative factors [19]. Additional *FGF4* retrogenes are logical candidates for these *FGF4*L2 negative IVDD cases, however the additional *FGF4* retrocopies identified in this study do not appear to provide a compelling explanation for this group of dogs owing to the limited frequency of the retrogenes in affected animals. Although *FGF4*L7 was identified in three dogs treated surgically for IVDD, it was not seen in any other breeds in the surgically treated data set, and the breed with the highest identified allele frequency (Shar-Pei; 0.341) is not known to be among the breeds highly predisposed to IVDD [35]. Similarly, clinical IVDD is uncommon in Border Terriers and Norfolk Terriers, which had the highest allele frequency of the *FGF4*L4 retrogene. Interestingly, *FGF4*L7 inserted 5 Mb downstream from the HAS2 gene, a gene that has been implicated in the Shar-Pei wrinkled skin phenotype as well as Familial Shar-Pei Fever [36]. Strong selection in this region of the genome in Shar Peis could explain the high allele frequency of *FGF4*L7 in the breed.

While *FGF4*L4 was not found to be associated with height in Border Terriers, the majority of Border Terriers tested had either one or two copies of *FGF4*L4, and only one individual with 0 copies was identified. If the retrocopy has a dominant effect on height in the breed, more homozygous wild type individuals will need to be measured to determine any effect. *FGF4*L4 was also found at low allele frequencies in other related terrier breeds, including the Skye, Norwich, and Norfolk Terriers, and may have originated in a common progenitor to the terrier breed group. As dog breeds are known to be highly inbred [37,38], a high allele frequency alone does not indicate selection, as it could be the result of random genetic drift followed by decreasing genetic diversity, as characterizes purebred dogs. Interestingly, *FGF4*L1 is also very common in Norwich and Norfolk Terriers [14], and the Skye Terriers used in this study were homozygous for both *FGF4*L1 and *FGF4*L2, making them the first breed to be identified with three *FGF4* retrocopies.

All the *FGF4* retrocopies in canids appear to have been very recently retrotransposed with no new mutations differentiating them from the parental *FGF4* gene. Even the red wolf *FGF4* retrocopy, *FGF4*L6, is nearly identical to the red wolf specific *FGF4* haplotype. Dating the *FGF4* retrocopy insertions is difficult owing to their short length (3.2 kbp) and sequence identity to the parental gene sequence; however, the *FGF4* retrocopies are identical to canine-specific *FGF4* gene haplotypes, which are distinct from modern wolf *FGF4* haplotypes (Supplemental Table S5). This could indicate that the dog retrocopies occurred after domestication. Recently inserted, fully intact retrocopies such as the *FGF4* retrocopies are very uncommon in reference genomes. Studies have found that less than 18% of the retrocopies in the human reference have a fully intact ORF, while only 1% of retrocopies share greater than 99% of their DNA sequence with their parental gene [31,39]. However, these studies have focused on analyzing reference genomes, which miss polymorphic retrocopies that are more likely to be recent, such as the *FGF4* retrocopies in canids, which are not found in CanFam3. It is possible that some unique aspects of the *FGF4* gene increase its rate of L1 mediated retrotransposition. A search for *FGF4* through a database of all retrocopies identified in over 40 mammalian reference genomes reveals that a squirrel (Ictidomys tridecemlineatus) and a hedgehog (Echinops telfairi) also have *FGF4* retrocopies, although they are only 61.2% and 90.6% identical to the parental genes, indicating they are not recent [8], but it is unknown whether other species have polymorphic *FGF4* retrocopies not found in their reference genomes. Another possibility is that L1 mediated gene retrotransposition in general is occurring more frequently in canids. If this was the case, recent, polymorphic retrocopies may be more common in canids in a greater number of genes than just *FGF4*.

While next generation sequencing allows for the detection of polymorphic gene retrocopies, they often go unidentified or misidentified by common variant calling methods [40]. However, more careful analysis of discordant Illumina paired-end reads has shown they are more common than previously thought [41,42]. As with *FGF4*L1 and *FGF4*L2, retrocopies of other genes may have phenotypic consequences. As such, the possibility of retrocopy insertions should be considered when scanning critical intervals for disease trait associations. Recently inserted gene retrocopies can result in overexpression of the parental gene product, resulting in gain of function, which could be deleterious [43]. In this study, whole-genome sequence data were successfully used to identify novel, polymorphic retrocopies of the *FGF4* gene; a similar approach could be generalized to all genes to identify other polymorphic gene retrocopies in canids. Similar to the *FGF4* retrocopies, other polymorphic retrocopies may play an important role in both breed health and phenotypic variation across dogs.

SUPPLEMENTARY MATERIALS

The following are available online at https://www.mdpi.com/2073-4425/11/8/839/s1, Figure S1: Retrocopy discovery from whole-genome sequencing data; Figure S2: *FGF4* genotyping assay; Figure S3: Height in Border Terriers; Table S1: Breed list of whole-genome sequence files analyzed; Table S2: *FGF4* retrocopy primers; Table S3: Breed list and *FGF4* retrocopy genotype results; Table S4: SRA accession numbers for canids with novel *FGF4* retrocopies; Table S5: SNV at the *FGF4* gene locus in 722 canids.

ACKNOWLEDGEMENTS

We would like to thank all researchers who made their genome sequences publicly available, which allowed this study to be possible. We would also like to thank the United States Fish and Wildlife Services for granting access to red wolf samples for DNA extraction.

CHAPTER 3 REFERENCES

1. Casola, C.; Betrán, E. The genomic impact of gene retrocopies: What have we learned from comparative genomics, population genomics, and transcriptomic analyses? Genome Biol. Evol. 2017, 9, 1351–1373.
2. Ostertag, E.M.; Kazazian, H.H., Jr. Biology of mammalian L1 retrotransposons. Annu. Rev. Genet. 2001, 35, 501–538.
3. Jurka, J. Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. Proc. Natl. Acad. Sci. USA 1997, 94, 1872–1877.
4. Penzkofer, T.; Jäger, M.; Figlerowicz, M.; Badge, R.; Mundlos, S.; Robinson, P.N.; Zemojtel, T. L1Base 2: More retrotransposition-active LINE-1s, more mammalian genomes. Nucleic Acids Res. 2017, 45, D68–D73.
5. Luan, D.D.; Korman, M.H.; Jakubczak, J.L.; Eickbush, T.H. Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: A mechanism for non-LTR retrotransposition. Cell 1993, 72, 595–605.
6. Noll, A.; Raabe, C.A.; Churakov, G.; Brosius, J.; Schmitz, J. Ancient traces of tailless retropseudogenes in therian genomes. Genome Biol. Evol. 2015, 7, 889–900.
7. Zhang, Z.; Harrison, P.M.; Liu, Y.; Gerstein, M. Millions of years of evolution preserved: A comprehensive catalog of the processed pseudogenes in the human genome. Genome Res. 2003, 13, 2541–2558.
8. Rosikiewicz, W.; Kabza, M.; Kosiński, J.G.; Ciomborowska-Basheer, J.; Kubiak, M.R.; Makałowska, I. RetrogeneDB–a database of plant and animal retrocopies. Database 2017, 2017.
9. Ewing, A.D.; Ballinger, T.J.; Earl, D.; Harris, C.C.; Ding, L.; Wilson, R.K.; Haussler, D. Retrotransposition of gene transcripts leads to structural variation in mammalian genomes. Genome Biol. 2013, 14, R22.
10. Abyzov, A.; Iskow, R.; Gokcumen, O.; Radke, D.W.; Balasubramanian, S.; Pei, B.; Habegger, L.; Lee, C.; Gerstein, M.; Consortium, G.P. Analysis of variable retroduplications in human populations suggests coupling of retrotransposition to cell division. Genome Res. 2013, 23, 2042–2052.
11. Schrider, D.R.; Navarro, F.C.; Galante, P.A.; Parmigiani, R.B.; Camargo, A.A.; Hahn, M.W.; de Souza, S.J. Gene copy-number polymorphism caused by retrotransposition in humans. PLoS Genet. 2013, 9.
12. Gao, X.; Li, Y.; Adetula, A.A.; Wu, Y.; Chen, H. Analysis of new retrogenes provides insight into dog adaptive evolution. Ecol. Evol. 2019, 9, 11185–11197.

13. Tsai, K.L.; Evans, J.M.; Noorai, R.E.; Starr-Moss, A.N.; Clark, L.A. Novel Y chromosome retrocopies in canids revealed through a genome-wide association study for sex. Genes 2019, 10, 320.

14. Parker, H.G.; VonHoldt, B.M.; Quignon, P.; Margulies, E.H.; Shao, S.; Mosher, D.S.; Spady, T.C.; Elkahloun, A.; Cargill, M.; Jones, P.G. An expressed fgf4 retrogene is associated with breed-defining chondrodysplasia in domestic dogs. Science 2009, 325, 995–998.

15. Brown, E.A.; Dickinson, P.J.; Mansour, T.; Sturges, B.K.; Aguilar, M.; Young, A.E.; Korff, C.; Lind, J.; Ettinger, C.L.; Varon, S. FGF4 retrogene on CFA12 is responsible for chondrodystrophy and intervertebral disc disease in dogs. Proc. Natl. Acad. Sci. USA 2017, 114, 11476–11481.

16. Akey, J.M.; Ruhe, A.L.; Akey, D.T.; Wong, A.K.; Connelly, C.F.; Madeoy, J.; Nicholas, T.J.; Neff, M.W. Tracking footprints of artificial selection in the dog genome. Proc. Natl. Acad. Sci. USA 2010, 107, 1160–1165.

17. Vaysse, A.; Ratnakumar, A.; Derrien, T.; Axelsson, E.; Pielberg, G.R.; Sigurdsson, S.; Fall, T.; Seppälä, E.H.; Hansen, M.S.; Lawley, C.T. Identification of genomic regions associated with phenotypic variation between dog breeds using selection mapping. PLoS Genet. 2011, 7.

18. Hansen, H.-J. A pathologic-anatomical study on disc degeneration in dog: With special reference to the so-called enchondrosis intervertebralis. Acta Orthop. Scand. 1952, 23, 1–130.

19. Batcher, K.; Dickinson, P.; Giuffrida, M.; Sturges, B.; Vernau, K.; Knipe, M.; Rasouliha, S.H.; Drögemüller, C.; Leeb, T.; Maciejczyk, K. Phenotypic effects of FGF4 retrogenes on intervertebral disc disease in dogs. Genes 2019, 10, 435.

20. Plassais, J.; Kim, J.; Davis, B.W.; Karyadi, D.M.; Hogan, A.N.; Harris, A.C.; Decker, B.; Parker, H.G.; Ostrander, E.A. Whole genome sequencing of canids reveals genomic regions under selection and variants influencing morphology. Nat. Commun. 2019, 10, 1–14.

21. Bauer, A.; Waluk, D.P.; Galichet, A.; Timm, K.; Jagannathan, V.; Sayar, B.S.; Wiener, D.J.; Dietschi, E.; Müller, E.J.; Roosje, P. A de novo variant in the ASPRV1 gene in a dog with ichthyosis. PLoS Genet. 2017, 13, e1006651.

22. Decker, B.; Davis, B.W.; Rimbault, M.; Long, A.H.; Karlins, E.; Jagannathan, V.; Reiman, R.; Parker, H.G.; Drögemüller, C.; Corneveaux, J.J. Comparison against 186 canid whole-genome sequences reveals survival strategies of an ancient clonally transmissible canine tumor. Genome Res. 2015, 25, 1646–1655.

23. Lucot, K.L.; Dickinson, P.J.; Finno, C.J.; Mansour, T.A.; Letko, A.; Minor, K.M.; Mickelson, J.R.; Drögemüller, C.; Brown, C.T.; Bannasch, D.L. A missense mutation in the vacuolar protein sorting 11 (VPS11) gene is associated with neuroaxonal dystrophy in Rottweiler dogs. G3 Genes Genomes Genet. 2018, 8, 2773–2780.

24. Kardos, M.; Åkesson, M.; Fountain, T.; Flagstad, Ø.; Liberg, O.; Olason, P.; Sand, H.; Wabakken, P.; Wikenros, C.; Ellegren, H. Genomic consequences of intensive inbreeding in an isolated wolf population. Nat. Ecol. Evol. 2018, 2, 124–131.

25. Jagannathan, V.; Drögemüller, C.; Leeb, T.; Consortium, D.B.V.D.; Aguirre, G.; André, C.; Bannasch, D.; Becker, D.; Davis, B.; Ekenstedt, K. A comprehensive biomedical variant

catalogue based on whole genome sequences of 582 dogs and eight wolves. Anim. Genet. 2019, 50, 695–704.

26. Robinson, J.T.; Thorvaldsdóttir, H.; Winckler, W.; Guttman, M.; Lander, E.S.; Getz, G.; Mesirov, J.P. Integrative genomics viewer. Nat. Biotechnol. 2011, 29, 24–26.
27. Ochman, H.; Gerber, A.S.; Hartl, D.L. Genetic applications of an inverse polymerase chain reaction. Genetics 1988, 120, 621–623. [Google Scholar]
28. Untergasser, A.; Cutcutache, I.; Koressaar, T.; Ye, J.; Faircloth, B.C.; Remm, M.; Rozen, S.G. Primer3—new capabilities and interfaces. Nucleic Acids Res. 2012, 40, e115.
29. Kent, W.J.; Sugnet, C.W.; Furey, T.S.; Roskin, K.M.; Pringle, T.H.; Zahler, A.M.; Haussler, D. The human genome browser at UCSC. Genome Res. 2002, 12, 996–1006.
30. Blanchette, M.; Kent, W.J.; Riemer, C.; Elnitski, L.; Smit, A.F.; Roskin, K.M.; Baertsch, R.; Rosenbloom, K.; Clawson, H.; Green, E.D. Aligning multiple genomic sequences with the threaded blockset aligner. Genome Res. 2004, 14, 708–715.
31. Carelli, F.N.; Hayakawa, T.; Go, Y.; Imai, H.; Warnefors, M.; Kaessmann, H. The life history of retrocopies illuminates the evolution of new mammalian genes. Genome Res. 2016, 26, 301–314.
32. Parish, A.; Schwaederle, M.; Daniels, G.; Piccioni, D.; Fanta, P.; Schwab, R.; Shimabukuro, K.; Parker, B.; Helsten, T.; Kurzrock, R. Fibroblast growth factor family aberrations in cancers: Clinical and molecular characteristics. Cell Cycle 2015, 14, 2121–2128.
33. Huebner, K.; Ferrari, A.; Delli, B.P.; Croce, C.; Basilico, C. The FGF-related oncogene, K-FGF, maps to human chromosome region 11q13, possibly near int-2. Oncogene Res. 1988, 3, 263.
34. Murphy, B.G.; Dickinson, P.; Marcellin-Little, D.J.; Batcher, K.; Raverty, S.; Bannasch, D. Pathologic Features of the Intervertebral Disc in Young Nova Scotia Duck Tolling Retrievers Confirms Chondrodystrophy Degenerative Phenotype Associated With Genotype. Vet. Pathol. 2019, 56, 895–902.
35. Bergknut, N.; Egenvall, A.; Hagman, R.; Gustås, P.; Hazewinkel, H.A.; Meij, B.P.; Lagerstedt, A.-S. Incidence of intervertebral disk degeneration–related diseases and associated mortality rates in dogs. J. Am. Vet. Med. Assoc. 2012, 240, 1300–1309.
36. Olsson, M.; Meadows, J.R.; Truve, K.; Pielberg, G.R.; Puppo, F.; Mauceli, E.; Quilez, J.; Tonomura, N.; Zanna, G.; Docampo, M.J. A novel unstable duplication upstream of HAS2 predisposes to a breed-defining skin phenotype and a periodic fever syndrome in Chinese Shar-Pei dogs. PLoS Genet 2011, 7, e1001332.
37. Wade, C.M. Inbreeding and genetic diversity in dogs: Results from DNA analysis. Vet. J. 2011, 189, 183–188.
38. Leroy, G. Genetic diversity, inbreeding and breeding practices in dogs: Results from pedigree analyses. Vet. J. 2011, 189, 177–182.
39. Marques, A.C.; Dupanloup, I.; Vinckenbosch, N.; Reymond, A.; Kaessmann, H. Emergence of young human genes after a burst of retroposition in primates. PLoS Biol. 2005, 3.
40. Chatron, N.; Cassinari, K.; Quenez, O.; Baert-Desurmont, S.; Bardel, C.; Buisine, M.P.; Calpena, E.; Capri, Y.; Corominas Galbany, J.; Diguet, F. Identification of mobile retrocopies during genetic testing: Consequences for routine diagnosis. Hum. Mutat. 2019, 40, 1993–2000.

41. Abel, H.J.; Larson, D.E.; Regier, A.A.; Chiang, C.; Das, I.; Kanchi, K.L.; Layer, R.M.; Neale, B.M.; Salerno, W.J.; Reeves, C.; et al. Mapping and characterization of structural variation in 17,795 human genomes. Nature 2020, 583, 83–89.

42. Zhang, Y.; Li, S.; Abyzov, A.; Gerstein, M.B. Landscape and variation of novel retroduplications in 26 human populations. PLoS Comput. Biol. 2017, 13, e1005567.

43. Kubiak, M.R.; Makałowska, I. Protein-coding genes' retrocopies and their functions. Viruses 2017, 9, 80.

CHAPTER 4: Recent, full-length gene retrocopies are common in canids

Kevin Batcher[1], Scarlett Varney[1], Daniel York[2], Matthew Blacksmith[3], Jeffrey M. Kidd[3, 4], Robert Rebhun[2], Peter Dickinson[2], and Danika Bannasch[1]*

[1] Department of Population Health and Reproduction, University of California, Davis, CA, USA

[2] Department of Surgical and Radiological Sciences, University of California, Davis, CA, USA

[3] Department of Human Genetics, University of Michigan Medical School, Ann Arbor, MI, USA

[4] Department of Computational Medicine and Bioinformatics, University of Michigan Medical School, Ann Arbor, MI, USA

*Correspondence: dlbannasch@ucdavis.edu

Genome Res. 2022, 32; https://doi.org/10.1101/gr.276828.122

Received: April 8, 2022; Accepted: July 19, 2022; Published in Advance: August 12, 2022

AUTHOR CONTRIBUTIONS

Conceptualization, D.B., **K.B.**, and J.K.; Data curation, **K.B.**; Formal analysis, **K.B.**, M.B., and J.K.; Funding acquisition, D.B. and J.K.; Investigation, **K.B.**, M.B., and S.V.; Project administration, D.B.; Resources, D.B., **K.B.**, M.B., and J.K.; Supervision, D.B.; Visualization, **K.B.**; Writing—original draft, D.B. and **K.B.**; Writing—review & editing, D.B., **K.B.**, P.D., R.R., and D.Y.

ABSTRACT

Gene retrocopies arise from the reverse transcription and insertion into the genome of processed mRNA transcripts. Although many retrocopies have acquired mutations that render them functionally inactive, most mammals retain active LINE-1 sequences capable of producing new retrocopies. New retrocopies, referred to as retro copy number variants (retroCNVs), may not be identified by standard variant calling techniques in high-throughput sequencing data. While multiple functional *FGF4* retroCNVs have been associated with skeletal dysplasias in dogs, the full landscape of canid retroCNVs has not been characterized. Here, retroCNV discovery was performed on a whole-genome sequencing dataset of 293 canids from 76 breeds. We identified retroCNV parent genes via the presence of mRNA specific 30-mers, and then identified retroCNV insertion sites through discordant read analysis. In total, we resolved insertion sites for 1911 retroCNVs from 1179 parent genes, 1236 of which appeared identical to their parent genes. Dogs had on average 54.1 total retroCNVs and 1.4 private retroCNVs. We found evidence of expression in testes for 12% (14/113) of the retroCNVs identified in 6 Golden Retrievers, including four chimeric transcripts, and 97 retroCNVs also had significantly elevated F_{ST} across dog breeds, possibly indicating selection. We applied our approach to a subset of human genomes and detected an average of 4.2 retroCNVs per sample, highlighting a 13-fold relative increase of retroCNV frequency in dogs. Particularly in canids, retroCNVs are a largely unexplored source of genetic variation which can contribute to genome plasticity and which should be considered when investigating traits and diseases.

INTRODUCTION

Gene retrotransposition occurs when mRNA is reverse transcribed into DNA and inserted back into the genome, resulting in an intron-less copy of a gene referred to as a retrocopy or a processed pseudogene. This process is carried out in mammals by long interspersed nuclear element 1 (LINE-1 or L1) proteins acting in trans on cellular mRNA [1-3]. LINE-1 mediated retrotransposition results in the duplication of short (10-20bp) segments of genomic DNA flanking the insertion, referred to as a target site duplications (TSD). Thousands of retrocopies have been identified in mammalian reference assembles, although the exact number varies by species and by annotation method applied [4,5]. Most of these reference retrocopies are the consequence of ancestral retrotransposition events, evidenced by the accumulation of mutations that differentiate the retrocopy sequence from that of the parent gene [5]. While these ancestral retrocopies tend to be fixed in a species [6-8], most mammalian genomes contain active LINE-1s capable of producing novel retrocopy insertions [9]. These more recent retrocopy insertions may not be fixed in a species, resulting in gene copy number variation between individuals, referred to as retro copy number variants (retroCNVs) [1,6,7,10,11]. While some of the retrocopies present in a reference genome assembly may be polymorphic in a species and thus reference retroCNVs, there are also non-reference retroCNVs which are not found in the assembly itself [6].

RetroCNVs are a type of complex structural variant that require specialized techniques for identification within whole genome sequencing (WGS) data [11]. When Illumina paired-end reads are aligned to a reference assembly, any reference retroCNV that is absent in an individual will appear as a deletion relative to the assembly. However, when an individual has a non-reference retroCNV, the reads coming from that retroCNV will align to the parent gene. Because the retrocopy lacks introns, discordant reads are observed aligning only to the exons of the parent gene. Discordant reads are also found at the 3' and 5' end of the gene mapping to the insertion site of the retroCNV. These two features can be used to identify non-reference retroCNVs from WGS data, with the "gold standard" in retroCNV discovery requiring identification of the parent gene and characterization of the insertion site [1,12]. Estimates of gene retrotransposition rates have varied, although recent analyses using high coverage WGS data have identified 1663 retroCNV parent genes in populations of mice [12] and 503 in human populations [13], indicating that gene retrotransposition is a common occurrence. There is limited information about retroCNVs in dogs [14-16], however, short interspersed nuclear element (SINE) insertions, which are also mobilized in trans by LINE-1 encoded proteins in a manner similar to gene retrocopies, have been shown to be highly dimorphic in dogs [17,18].

While retrocopies have historically been referred to as processed pseudogenes and presumed to be nonfunctional, evidence has accumulated for retrocopy expression and functionality [12,19,20]. It has been argued that retroCNVs are likely to be deleterious based on negative selection in natural populations of mice [12], and consistent with this hypothesis, retroCNV in humans have been shown to be involved in cancer as well as neurodegenerative, mental or cardiovascular disorders [21]. In dogs, two recently inserted and expressed *FGF4* retrocopies are associated with dominant skeletal dysplasias [22,23]. Several additional *FGF4* retrocopies with no known phenotypic associations were also discovered, indicating that retroCNV formation may also be a common occurrence in dogs [16]. Artificial breed selection by humans could also increase the allele frequency of functional retroCNVs, as appears to have been the case with the *FGF4* retroCNVs which are common in many breeds [24]. As such, analysis of the full landscape of retroCNVs in dogs could lead to interesting insights into retrocopy biology and may additionally help in identifying causative variants for phenotypic associations in dogs. The goal of this current study was to characterize the landscape of retroCNV in dogs by performing retroCNV discovery on a diverse dataset of canids and by further analyzing the retroCNV for evidence of function.

RESULTS

RetroCNV Discovery

We used high coverage WGS data from 293 canids (median coverage of 25.6×) aligned to the CanFam3.1 reference genome for our retroCNV discovery dataset (Supplemental Table S1). Our approach to retroCNV discovery was to first identify mRNA specific 30-mers, which are present in spliced gene sequences but absent from the CanFam3.1 reference assembly (Figure 1A). These sequences are only found in genomic DNA when a non-reference retroCNV is present (Figure 1B). No mRNA specific 30-mers were identified from single exon genes (N=1574) or genes with recent retrocopies already present in the reference assembly (N=75). We identified mRNA specific 30-mers for 18,192 protein coding genes and 10,807 long non-coding RNAs which were then used for retroCNV parent gene discovery. In total, 1870 putative retroCNV parent genes were identified in the 293 canid dataset based on the presence of these mRNA specific 30-mers.

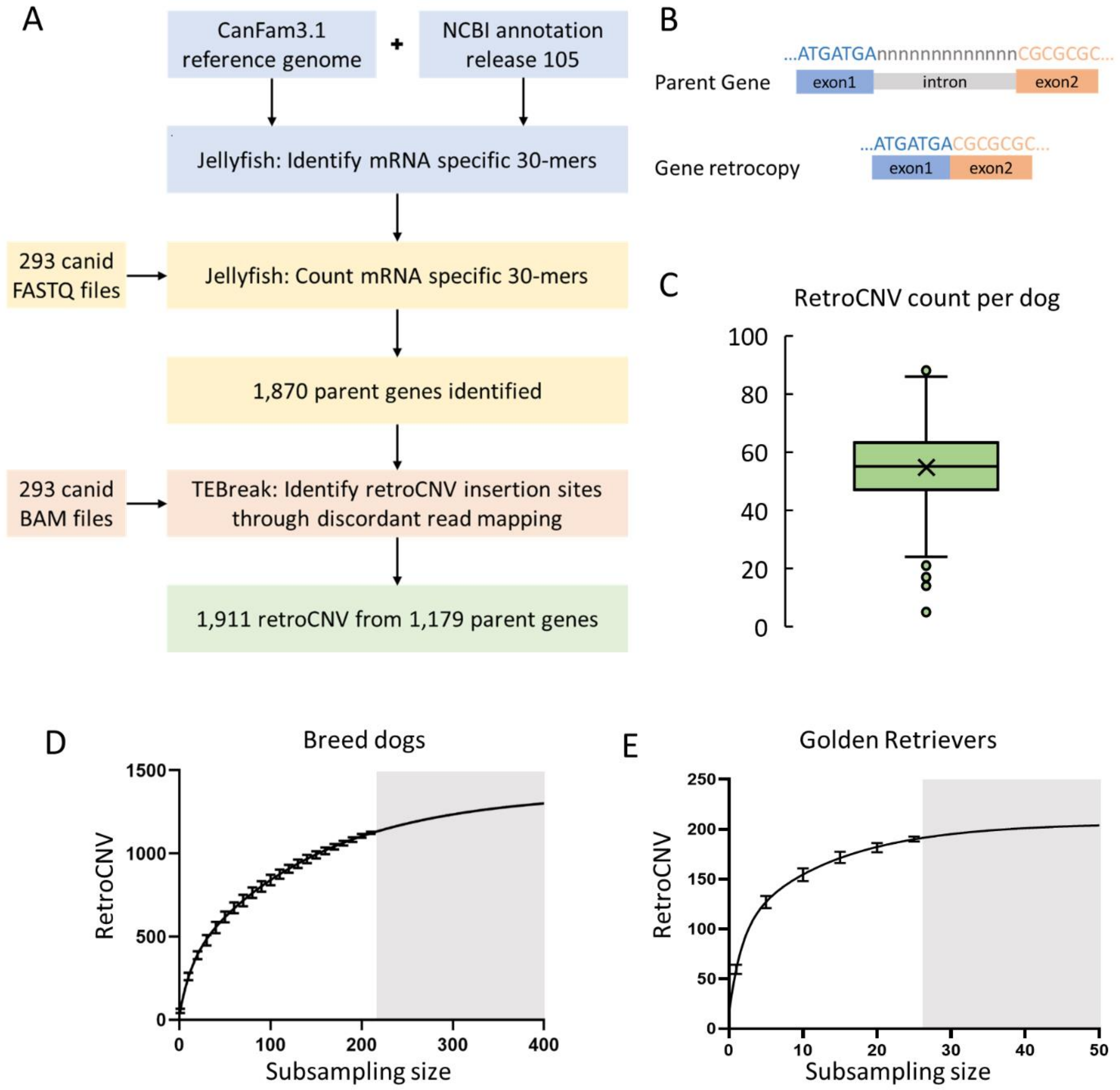


Figure 1 RetroCNV discovery. (A) Discovery pipeline. (B) mRNA specific sequences are formed due to intron removal; these sequences are used to identify non-reference retroCNV parent genes in FASTQ reads from genomic DNA sequence. (C) Total non-reference retroCNV count seen in individual domestic dogs (average 54.1). (D) Number of detected retroCNV insertions with increasing resampling sample sizes in breed dogs. Subsample sizes were selected from 1 to 210, with a step size of 10, and 100 replicates within each subsample. Bars represent standard deviation of the replicates at each subsample size, and the gray area shows the prediction of increasing the number of dogs beyond the number used in this study. (E) Number of detected retroCNV insertions with increasing resampling sample sizes in Golden Retrievers. Subsample sizes were selected from 1 to 25, with a step size of 5 and 100 replicates within each subsample. Bars represent standard deviation of the replicates at each subsample size.

In order to resolve the insertion site of non-reference retroCNVs (Supplemental Figure S1), we analyzed discordant paired-end reads. From the 1870 parent genes identified as having putative retroCNVs, insertion sites were resolved for 1911 total non-reference retroCNVs coming from 1179 parent genes (Figure 2A). Many (808/1911, 42.3%) insertion sites were located within the introns of other protein coding genes (Supplemental Table S2; Supplemental Figure S2). Of the 1179 retroCNV parent genes, 1150 were protein coding while 29 were lncRNA. Four of the previously identified *FGF4* retroCNVs were successfully identified through our approach (Supplemental Table S2). The TSD was resolved for 1676 (87.7%) of the retroCNVs, and the median TSD length was 16bp. No insertion sites could be identified for retroCNVs derived from 691 of the putative retroCNV parent genes; however, 125 of these parent genes had discordant reads mapping between exons, which may indicate the retroCNV inserted into repetitive or unresolved regions of the CanFam3.1 reference genome (Supplemental Table S3). Further inspection revealed that 21 of the putative retroCNV parent genes had discordant reads mapping to satellite DNA, suggesting insertion of the retroCNV into a telomeric or centromeric region. Additionally, four parent genes (*MITF*, *NME7*, *TXNDC12*, and *PPP2CB*) that had discordant reads in all of the males and none of the females were identified, indicating that the retroCNVs were likely on the Y Chromosome, which is not represented in the CanFam3.1 assembly.

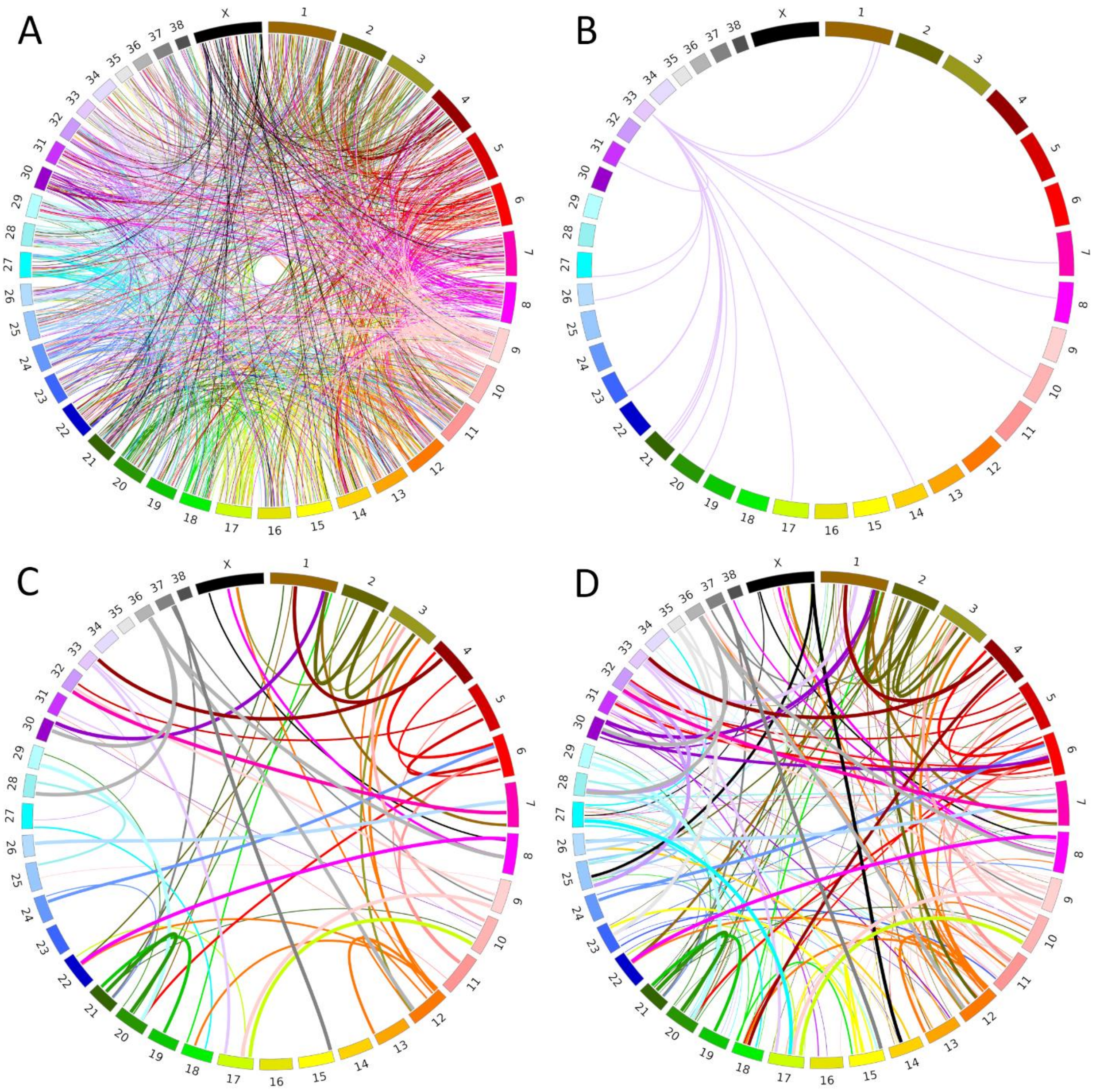


Figure 2 Circos plots highlighting the location of retroCNV parent genes and their insertion sites. Links are colored based on the chromosome of the parent gene. (A) All recent retroCNVs identified in canids with no retroCNV specific variants. (B) All retroCNVs of the *GAP43* parent gene. (C) All retroCNV present in a single Golden Retriever (SRR7107792). (D) All retroCNV present in all (N=26) Golden Retrievers. In figures C-D, the thickness of the link represents how common the retroCNV is within the Golden Retriever dataset.

We also identified retrocopies which are present in the CanFam3.1 reference genome but missing in at least one of the 293 canids, which we refer to as reference retroCNVs. We identified 58 reference retroCNVs which were confirmed through visual analysis of aligned WGS data (Supplemental Table S4). Several reference retroCNVs were highly prevalent in dogs (>=70%) and rare in wolves (<=10%), including the reference retroCNV *AKR1B1* (retro_cfam_63) which has previously been associated with domestication [25] as well as retroCNVs of *MGST3* (retro_cfam_35) and *RPL27A* (retro_cfam_145).

We focused all subsequent analysis on the 1911 non-reference retroCNV with resolved insertion sites. A full matrix matching individual canids with retroCNVs is available in Supplemental Table S5. While 880 of the retroCNV parent genes only had one retroCNV insertion site identified, 338 retroCNV parent genes had multiple insertion sites identified, including genes such as *GAPDH*, which had 21 retroCNV insertions (Supplemental Figure S3; Supplemental Table S2). Most of the retroCNV parent genes (900/1179, 76.3%) had no retrocopies present in the CanFam3.1 reference assembly. Additionally, 231 retroCNV parent genes had no known retrocopies in any of the mammalian reference genome assemblies, including *GAP43*, which had 18 retroCNV insertion sites identified in canids (Figure 2B; Supplemental Table S2).

Table 1 highlights some aspects of the non-reference retroCNVs by population. Dogs with an assigned breed, which we refer to as breed dogs, had 54.1 non-reference retroCNVs on average (95% CI 52.5-55.7) (Figure 1C). Within the 227 breed dogs, there were 325 private (unique to an individual dog) retroCNVs, or 1.4 per breed dog on average (95% CI 1.2-1.7), while the 43 free ranging dogs had 214 private retroCNVs and 5.0 each on average (95% CI 4.0-6.0).

Most non-reference retroCNVs were identified in a small number of dogs and at a low allele frequency in the entire population (Supplemental Figure S4). There were also 689 retroCNVs exclusive to breed dogs, 247 retroCNVs exclusive to the 18 wild canids, and 197 retroCNVs exclusive to the three African wild dogs (Table 1). We also observed that individuals which were sequenced at higher depth tended to have more retroCNV insertion sites resolved (Supplemental Figure S5). While samples at greater than 30× coverage had an average of 57.9 retroCNVs (95% CI 56.2-59.6), samples between 10× and 30× coverage had 54.8 (95% CI 52.3-57.3) and samples less than 10× coverage had 40 (95% CI 35.2 to 44.8) on average.

Population	Total retroCNV	Average retroCNV	Private retroCNV	Exclusive retroCNV
Breed dogs (N=227)	1165	54.1	325	689
Free ranging dogs (N=43)	705	56.6	214	254
Dingoes (N=3)	85	55.3	9	13
Wolves (N=10)	354	71.5	144	185
Coyotes (N=5)	104	33.8	31	49
African wild dogs (N=3)	214	173.0	44	197

Table 1 Population summary of non-reference retroCNVs. Private refers to retroCNVs unique to a specific individual, while exclusive refers to retroCNVs unique to a specific population.

RetroCNV specific gene variants

To estimate how recently the retroCNVs inserted, we identified variants which occurred in the retroCNV after insertion through the analysis of variants at the parent gene locus. Most retroCNV had not acquired any new variants after insertion, as retroCNV specific variants were only identified in 153/1390 (11.0%) of the retroCNVs analyzed, and only 8/1390 (0.6%) retroCNV had high impact mutations (Supplemental Table S6). Highlighting their more recent origin, only 92/1212 (7.6%) of the dog exclusive retroCNV had any retroCNV specific variants, while 61/178 (34.3%) of the retroCNV which were shared across canids had retroCNV specific variants.

Resampling

We expected that more retroCNVs may be detected in even larger datasets. To test this, we performed random resampling of individual breed dogs. While we identified 1140 retroCNVs in the 227 breed dogs, extrapolation from random resampling indicated that we would expect to identify over 1300 retroCNVs in a dataset of 400 breed dogs (Figure 1D). We similarly performed resampling within Golden Retrievers and observed that smaller sample sizes (N=26) can sufficiently capture the majority of retroCNVs within a single breed (Figure 1E; Figure 2C,D). Resampling within the free ranging dogs indicated that a large number of retroCNVs remain to be discovered, highlighting the heterogeneous nature of the free ranging dogs (Supplemental Figure S6).

RetroCNV Validation

We expected that some of the non-reference retroCNVs, while absent from the CanFam3.1 reference genome assembly, would be present in the alternative canid genome assemblies. This would confirm them as true retroCNVs and validate our discovery method. To test this, we first applied our retroCNV discovery pipeline to Illumina WGS data generated from four canids that have been used to create alternate canid reference assemblies, and identified 248 non-reference retroCNVs (Table 2). We then directly confirmed the presence of 173 of the 248 (69.8%) retroCNVs within their respective assemblies (Supplemental Table S7). We also identified the insertion site for 6 retroCNVs which had not been resolved through discordant read mapping (Supplemental Table S7). Most of the retroCNVs were full length with respect to the parent genes, with 153/179 (85.5%) containing the entire parental gene coding sequence.

Genome assembly involves the collapsing of heterozygous haplotypes, a process which could have resulted in heterozygous retroCNVs being excluded from the assembly sequence [26]. Therefore, we also analyzed PacBio long read data from each sample aligned to the CanFam3.1 assembly for evidence of retrocopy insertions (Supplemental Data S1). We found evidence for 236/248 (95.1%) of the retroCNVs within the long read data (Supplemental Table S8). A poly(A) tail 10bp or larger was observed in 211/236 (89.4%) of the retroCNVs, with median length of 29bp. A TSD was identified for 218/236 (92.4%) of the retroCNVs, with median length of 14bp. Overall, 244/248 (98.4%) of the retroCNVs were validated either within their respective genome assembly or in the long read data (Table 2).

Assembly	Predicted retroCNV	RetroCNV in assembly	RetroCNV in assembly or PacBio
UMICH_Zoey_3.1	49	33	47
UU_Cfam_GSD_1.0	59	41	59
Canfam_GSD	60	41	59
CanLup_DDS	80	58	79
Total	248	173	244

Table 2 Analysis of RetroCNVs in alternative canid genome assemblies.

We also developed PCR assays for nine of the predicted retroCNVs and validated them through Sanger sequencing (Figure 3A). In individuals positive for the retroCNV, we observed a poly(A) tail at one end of the insertion site and the 5' gene sequence of the retroCNV parent gene on other end of the insertion site (Figure 3C), which matched the expectation from the discordant reads and confirmed the presence of a retrocopy. For each retroCNV sequenced, the TSD was identified as the duplicated genomic sequence present at both ends of the insertion (Figure 3C, Supplemental Table S9). The TSD identified through Sanger sequencing for these nine retroCNVs matched the predicted TSD from the discordant reads. We additionally identified dogs lacking the retroCNVs (Figure 3B), confirming the retrocopies as polymorphic insertions. The list of retroCNVs validated through PCR and Sanger sequencing and the associated breeds is available in Supplemental Table S9.

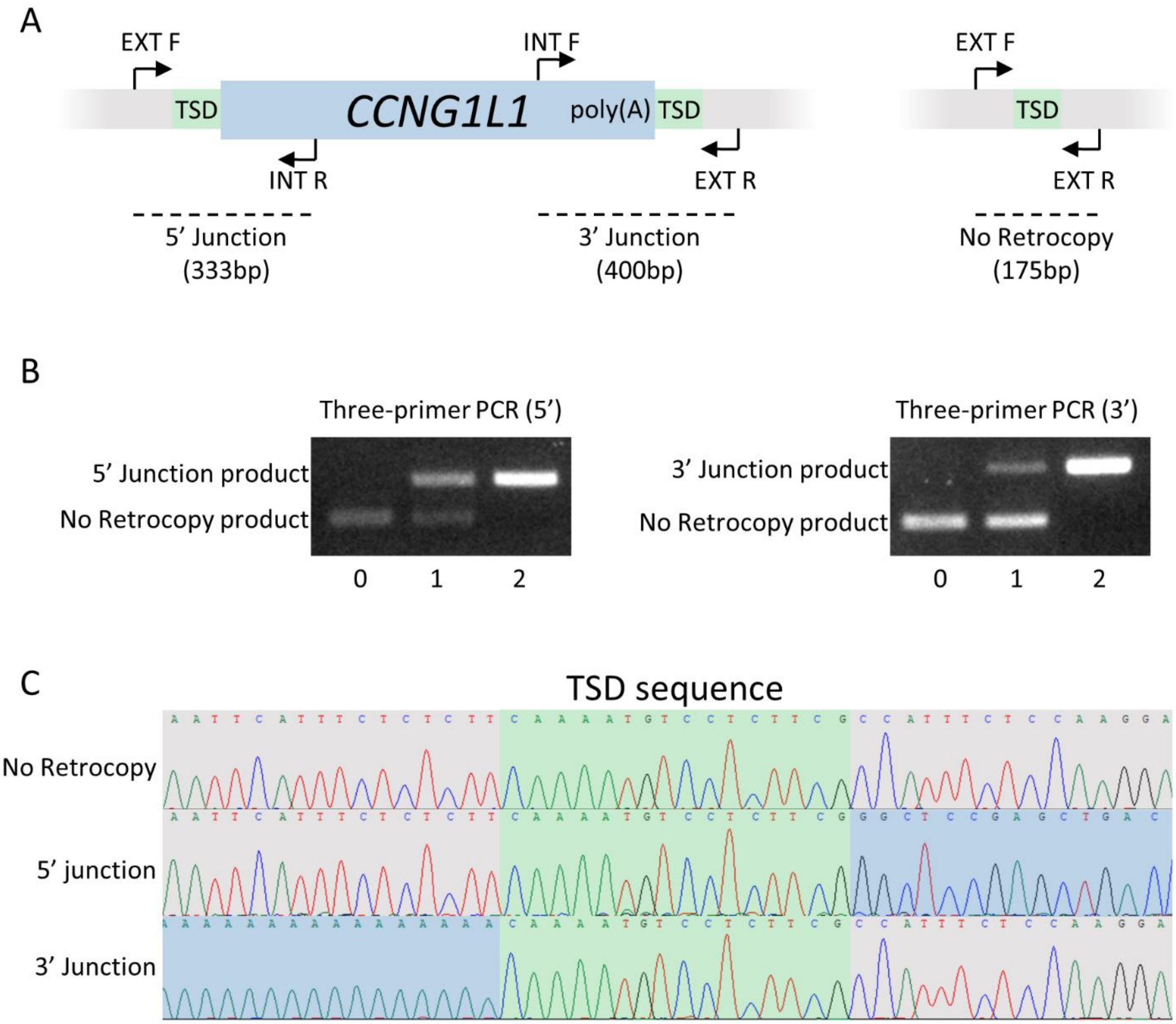


Figure 3 RetroCNV validation. (A) Primer design for retroCNV genotyping, with *CCNG1L1* as an example. When *CCNG1L1* is present, the EXT_F and INT_R primers produce a 333bp product at the 5' junction and the INT_F and EXT_R primers produce a 400bp product at the 3' junction. When the retrocopy is absent, the EXT_F and EXT_R primers produce a 175bp product. (B) Three-primer PCR results for *CCNG1L1* at the 5' and 3' junctions for individuals with 0, 1, and 2 copies of *CCNG1L1.* The two external primers EXT_F and EXT_R are included in both reactions as well as one of the internal primers, INT_F (5' Junction) or INT_R (3' Junction). (C) Sanger sequencing results for the *CCNG1L1* retroCNV. The TSD is identified as the genomic sequence from the insertion site which is present at both the 5' and 3' ends of the retroCNV.

RetroCNV selection and expression

We calculated a fixation index (F_{ST}) for the retroCNV across breed clades [27] and identified 97 retroCNVs which had significantly elevated F_{ST} (see methods). This included the two previously identified *FGF4* retroCNVs (Table 3; full list and clade distribution Supplemental Table S10, S11).

RetroCNV	F_{ST}	Insertion site	Total dogs	Insertion site gene
NDUFAF4L1	0.484	Chr2:6242423-6242444	5	*LOC111091106*
FGF4L2	0.439	Chr12:33710166-33710178	16	-
RHEBL1	0.430	Chr9:27506566-27506581	4	*CA10*
RPS2	0.430	Chr8:5138945-5139948	155	-
S100PL4	0.413	Chr34:34324611-34324631	13	-
PREPL1	0.403	Chr8:7565459-7565475	5	-
ARHGAP5L1	0.399	Chr5:18750413-18750430	9	-
RPS16L1	0.399	Chr3:26720828-26720840	9	-
ARPC1BL1	0.398	Chr14:34916675-34916690	5	-
NAA20L1	0.395	Chr18:34191916-34191933	9	*KIAA1549L*
RESTL1	0.388	Chr12:36658275-36658284	13	*COL12A1*
NAP1L1L2	0.365	Chr17:29923085-29923100	7	*LOC102154187*
NAP1L1L3	0.358	Chr18:49079285-49079285	11	*IGHMBP2*
C16orf87	0.350	Chr22:35188887-35190448	113	*C22H16orf87*
FAM133BL4	0.348	Chr29:31980926-31981012	8	*CA2*
RPL10L1	0.348	Chr3:32009305-32009323	8	*NIPA1*
ST13	0.342	Chr1:55086310-55087949	107	*UNC93A*
LSM2L1	0.341	Chr22:35508014-35508030	13	-
FGF4L1	0.337	Chr18:20443708-20443726	15	-
EIF4BL3	0.336	Chr18:16670414-16670429	14	*RELN*

Table 3 RetroCNVs with the highest F_{ST} between breed clades.

To determine if the retroCNVs showed evidence of expression, we first performed WGS and retroCNV discovery in six Golden Retrievers, and then RNA-seq using testes from the same 6 Golden Retrievers. There were 113 total non-reference retroCNVs identified in the six individuals. RetroCNV specific variants were present in 24 of the 113 retroCNVs, allowing for the distinction between parent and retroCNV derived transcripts and confirmed the expression for two retroCNV. Among the retroCNVs which had inserted within the introns of another gene, chimeric reads between retroCNV parent gene and a gene at the insertion site were observed in 4/42. This included the *COILL2* retroCNV, which is inserted within the 5' UTR of *LOC100686934*, producing a novel chimeric transcript in two of the six Golden Retrievers (Supplemental Figure S7). However, five of the 42 insertion site genes were not sufficiently expressed in testes (<1 transcript per million in all samples) to allow for the evaluation of chimera formation with the retroCNVs. Additionally, discordant reads mapping to the parent gene loci were observed at 11 of the insertion sites. Overall, at least one form of evidence for expression was observed for 12.4% (14/113) of the retroCNVs in 6 Golden Retriever testes (Supplemental Table S12). The expressed retroCNV *FARSBL1* was present in 60% of all dogs and only a single wild canid.

RetroCNV discovery in humans

To determine the rate of retroCNV occurrence in another species for comparison, we performed retroCNV discovery in 78 individuals from 26 populations using The 1000 Genomes Project Consortium phase 3 high coverage dataset [28,The 1000 Genomes Project Consortium 29]. We resolved insertion sites for 46 non-reference retroCNVs from 44 parent genes in the 78 samples (Supplemental Table S13). Of the 44 retroCNV parent genes, 40 have been previously identified in human datasets, while 34 of the 46 retroCNV insertion sites had been previously identified. A full analysis of which human retroCNVs have been identified in previous studies is available in Supplemental Table S14. Individuals in this dataset had 4.2 retroCNVs on average (95% CI 3.9-4.6).

Rate of retroCNV formation in canids

We observed a total of 214 non-reference retroCNVs in the three African wild dogs (*Lycaon pictus*), which had 173 retroCNVs each on average. Most of these retroCNVs were private to the African wild dogs, and only 17 retroCNVs were shared between the African wild dogs and any other canid, indicating that most of the retroCNVs identified in either species inserted after the species had diverged. Similarly, we identify 194 retroCNVs that are exclusive to grey wolves and 1010 retroCNVs exclusive to breed dogs. Genetic analyses have indicated that domestication in dogs occurred around 25,000 years ago, while breed formation largely occurred within the last 200 years [30]. If the 1010 breed dog specific retroCNVs inserted after domestication, we can estimate the rate of retroCNV accumulation at approximately four per 100 years. Alternatively, within our dataset of Golden Retrievers (N=26), we identify 10 retroCNVs that are exclusive to the breed and not found in any other breed dogs or canids. Since the Golden Retriever breed was formed roughly 200 years ago, 10 Golden Retriever exclusive retroCNV would indicate a similar rate of retroCNV accumulation at five per 100 years.

DISCUSSION

Previous analyses of canid retrocopies have focused on those retrocopies present in the CanFam3.1 reference assembly, which was produced from a single dog [5,14,31]. In this study we characterized a rich landscape of retroCNVs in canids, consistent with an active LINE-1. By applying a novel approach to retroCNV discovery on a diverse dataset of 293 canids, we identified 1911 retroCNVs, most of which had inserted recently. Domestic dogs have 54.1 non-reference retroCNVs each on average, and, as many of the retroCNV are private to a single individual, we also expect to find additional retroCNVs in larger discovery datasets. We observed many retroCNV that appear under selection, and that even within a single tissue type, 12% of the retroCNVs were expressed or forming novel chimeric transcripts with nearby genes, indicating that some of the retroCNVs may have functional and phenotypic consequences in canids.

Our approach to retroCNV discovery successfully resolved a large number of retroCNV in canids. In humans, previous studies using low-coverage WGS datasets have underestimated the rate of retrocopy insertion [1]. When long-read assemblies were analyzed, the estimated rate of retroCNV formation in humans was increased from 39 events per 939 individuals to 40 events per 22 individuals, or 4.1 per individual on average [26]. Similar to the analysis by Feng and Li, we found that humans had on average 4.2 retroCNV insertions using our discovery method, which identified 46 total retroCNV insertions in 78 human genomes. While most of the retroCNV parent genes have been reported in previous analyses of human genomes, 8 of the insertion sites have not been previously identified [13]. Previous analysis of canine SINEs, which are also mobilized via LINE-1 proteins acting in trans, highlighted a rate of SINE insertions 10 to

100-fold higher than that observed in humans [18]. While phenotypic associations with retroCNV are rare, SINE and LINE-1 insertions are a significant source of phenotypic variability in dogs, being responsible for 10% of the phenotype associated variants identified to date (Online Mendelian Inheritance in Animals, OMIA. Sydney School of Veterinary Science, 3/29/2022. World Wide Web URL: https://omia.org/). Whereas in humans, transposable elements are responsible for only 0.27% of all disease mutations [32]. In a recent analysis of the CanFam3.1 and UMICH_Zoey_3.1 reference assemblies, 16,221 dimorphic SINE and 1,121 dimorphic LINE-1 were identified, which represented a 17-fold increase in SINE differences and an 8-fold increase in LINE-1 differences compared to the number found in humans [17]. Domestic dogs, at 54.1 retroCNVs on average, also have a 13-fold increase in retroCNVs relative to humans, data which is consistent with either a highly active or a highly promiscuous LINE-1 in dogs. We also identified 231 retroCNV parent genes that have no known retrocopies in any other mammalian species, which might indicate that canine LINE-1 proteins are less selective. Among these retroCNV parent genes was *GAP43,* which had 18 retroCNV insertions and which may have implications in cognitive function in dogs [33].

In this study, we only considered retroCNVs with identifiable insertion sites as valid, although we provide evidence for the presence of 125 additional retroCNVs with unresolved insertion sites. Our method of retroCNV parent gene discovery also cannot identify single exon genes or genes with recent retrocopies present in the CanFam3.1 reference assembly. Still, we estimated the rate of retroCNV formation in domestic dogs at around four retroCNVs per 100 years, which is still an underestimate as our dataset of 228 breed dogs does not capture every retroCNV in the entire population. A recent analysis of retroCNVs in mice estimated their rate of retroCNV formation at two per 100 years [12]. This would indicate that domestic dogs have a rate of retroCNV formation even 2x greater than mice, which are known to have a large number of active LINE-1s [9] and have been shown to have 5-fold as many retroCNVs as humans [7]. In natural populations of mice, however, retroCNVs were also shown to be under negative selection due to deleterious effects, where retroCNVs are quickly removed from the population [12]. Also, in human populations, retroCNV insertions are not found in evolutionarily conserved regions, which indicates that highly deleterious retroCNVs are under negative selection [13]. We observed that many retroCNVs appear to be under positive selection in dogs, with significantly elevated F_{ST} values. Like the *FGF4* retrogenes, other retroCNVs may be under positive selection by breeders due to their phenotypic effects [24], although they may also be neutral variants whose frequencies differ due to the dynamics of breed formation and low genetic diversity within breeds [34].

While gene retrocopies were historically considered nonfunctional pseudogenes, more recently it has been recognized that retrocopies, which are a form of structural variant, are often functional through a variety of mechanisms which have been explored in recent reviews [20,21]. We found evidence for expression in 14 out of 113 retroCNVs in testes tissue from six Golden Retrievers, including four novel chimeric transcripts. However, this is likely an underestimate due to our use of a single tissue type for transcriptional assessment. Additionally, many of the retroCNVs were indistinguishable from their parent genes and thus cannot be effectively queried for evidence of expression through RNA-seq analysis alone. Comparison of overall gene expression in larger RNA-seq datasets including individuals with and without specific retroCNVs may be required to determine expression of retroCNVs which are identical to their parent genes. When such analyses were performed in mice, differences in retroCNV parent gene expression were found between individuals with and without the respective retroCNV, including many cases where overall expression was significantly reduced in the individuals with the retroCNV [12]. Our data confirms that the *AKR1B1* reference retroCNV, which has previously been shown to be expressed and associated with dog domestication, is common in dogs while rare in wolves [25]. Similarly, the *MGST3* and *RPL27A* reference retroCNVs and the expressed *FARSBL1* retroCNV are both common across dogs while rare in wild canids and may play roles in domestication. *AKR1B1*, *MGST3*, and *FARSB* are also involved in metabolism [35,36], while the *RPL27A* retrogene is inserted within *NCOA3*, a coactivator of transcription involved in thyroid function [37].

The four Y Chromosome retroCNVs were present in all male canids and thus not true retroCNV, and their sequence deviation from their parent genes of origin indicates that they are likely ancestral. Two of the Y Chromosome retroCNVs had been previously identified through the identification of autosomal variants which were actually sex-linked and due to the retroCNV insertion [38]. In this study, we also identified variants at the retroCNV parent gene loci which are likely attributable to non-reference retroCNVs. These variants would normally be attributed to variation in the parent genes, potentially hindering any analyses looking for causative mutations; it’s noteworthy that some variants identified in the coding sequences of genes may in fact be attributable to non-reference retroCNVs elsewhere in the genome.

Domesticated dog breeds have undergone artificial selection, which has led to extreme phenotypic diversity between breeds as well as breed predispositions to many heritable disorders [34,39-42]. In particular, many dog breeds are at a substantially higher risk for specific cancers when compared to human populations [43]. We have shown that dogs have many retroCNVs, consistent with a highly active LINE-1, and we also observed that some retroCNVs are under selection or also capable of expression or insertional mutagenesis through the formation of novel chimeric transcripts with nearby genes. These functional retroCNVs are likely a contributing factor to the phenotypic diversity seen in canids and are also strong candidates for disease associations in dogs, including susceptibility to cancers. We hope this list of retroCNV insertion sites will be a useful resource for the canine research community and that further assessment of the retroCNVs for evidence of function will provide insight into the genetics of phenotypic traits under selection in dogs.

METHODS

Data Selection

Illumina sequencing data aligned to the CanFam3.1 reference were downloaded from the sequence read archive [31]. The dataset included 227 dogs from 76 different breeds (referred to as 'breed dogs'), 43 free ranging dogs (marked as either 'village' or 'indigenous' by the data provider), three dingoes, ten grey wolves, two red wolves, five coyotes and three African wild dogs. A full list of samples used in this study and their accession numbers is available (Supplemental Table S1). Ten samples were removed from the analysis due to a low number of retroCNV insertion sites being resolved, possibly due to the quality of the sequencing data.

Parent gene discovery using mRNA specific 30-mers

The nucleotide sequence of a gene retrocopy resembles the processed mRNA transcript of its parent gene, with unique nucleotide sequences formed at the exon-exon junctions. These unique nucleotide sequences are only observed in genomic DNA when a retrocopy insertion is present, and thus can be used to identify the parent genes of retroCNVs from WGS data. We used Gffread [44] to obtain the spliced gene sequences for each gene transcript found in the NCBI CanFam3.1 annotation release 105). For each spliced gene sequence, we created a set of mRNA specific 30-mers that are absent from the CanFam3.1 reference assembly using Jellyfish count [45]. This identified all unique 30-mer sequences within the transcriptome that are absent from CanFam3.1 due to intron removal, giving a maximum of 29 mRNA specific 30-mers per exon-exon junction. We attributed mRNA specific 30-mers from alternatively spliced gene transcripts to their respective parent gene. To reduce false positives due to sequencing errors, we filtered out 30-mers with an edit distance of two substitutions from the reference genome using mrsFAST [46]. We removed any gene which had <5 mRNA specific 30-mers from further analysis, which included 1,574 single exon genes and 75 genes with recent retrocopies in the CanFam3.1 assembly. In total, 5,884,280 30-mers from 30,792 genes (median per gene: 106) were retained for retrocopy parent gene discovery. WGS data were then queried for the presence of the mRNA specific 30-mers using Jellyfish. Genes which had at least five mRNA specific 30-mers and at least 10% of the total 30-mers for that gene identified were considered as putative retroCNV parent genes for further analysis.

RetroCNV insertion site discovery through discordant read analysis

Gene retrocopies are derived from processed mRNA transcripts and thus lack introns. When Illumina paired end reads containing retroCNV sequences are aligned to a reference genome, they align to the parent gene locus, resulting in discordant read pairs which align only to the exons of the parent gene. Discordant read pairs can also be observed at the 5’ and 3’ ends of the parent gene, wherein one read aligns at the parent gene loci, while the other read aligns at the insertion site elsewhere in the genome. Discordant reads can thus be used to verify the presence of a RetroCNV as well as identify the insertion site. We performed discordant read analysis on aligned WGS data using TEBreak [47]. The “--disc_only” option was used to obtain a list of discordant read clusters of at least four reads (“--min_disc_reads 4”) mapping from a putative parent gene to elsewhere in the genome via “--disco_target”. Putative insertion sites for retroCNVs were visually confirmed in Integrative Genomics Viewer (IGV) [48] (Supplemental Figure S8). A retrocopy insertion site was considered valid if discordant reads were observed mapping to the same genomic locus from both the 3’ and 5’ end of the parent gene, or if discordant reads were found mapping from either the 3’ or 5’ as well as exon-exon discordant reads at the parent gene. Any discordant reads mapping from parent genes to known CanFam3.1 reference retrocopy insertion sites were ignored, as reference retroCNVs were analyzed separately. The TSD sequence was identified as the overlap between forward and reverse discordant reads at the insertion site. The 5’ and 3’ junction sequences for the retroCNV insertions were resolved using TEBreak and are available in Supplemental Table S15. Visual representations of retroCNV insertion sites were produced using Circos [49]. As has been proposed by Cheetham et al. (2020), rather than “pseudogene”, we chose a term which does not make functional inferences for the retrocopies: “like”, e.g. retrocopies of the *FGF4* gene were labeled *FGF4L1*, *FGF4L2*, etc.

CanFam3.1 reference assembly retroCNVs

We also examined retrocopies present in the CanFam3.1 assembly for evidence of being retroCNVs; an individual lacking a CanFam3.1 reference retrocopy would appear to have a deletion at those loci when aligned to CanFam3.1. A list of reference retrocopy locations was downloaded from RetrogeneDB [5]. The dataset of 293 canids previously aligned to CanFam3.1 was analyzed using DELLY with default settings to identify structural variants within 1kb of reference retrocopies [50]. All deletions were confirmed visually in IGV. Additionally, aligned sequence data from a gray wolf and a coyote were visually analyzed in IGV at all recent canine RetrogeneDB retrocopy loci (>95% identity with parent gene) to identify any retroCNVs which may have gone undetected by DELLY.

RetroCNV specific variant identification

Sequence variants between a retroCNV and its parent gene sequence are either due to germline variants present within the parent gene, or new polymorphisms unique to the retroCNV which occurred after insertion, which we refer to as retroCNV specific variants. As all retroCNV derived reads align to the parent gene loci, we analyzed variants at the parent gene loci in order to identify retroCNV specific variants. We first identified variants at the retroCNV parent gene loci using BCFtools mpileup [51]. We then compare variant allele frequencies between individuals positive or negative for each retroCNV, and variants that only appeared in individuals with the retroCNV were considered unique to the retroCNV. RetroCNVs that were unique to wild canids were excluded from this analysis as the wild canids contained many unique variants that could not be easily differentiated between variation within the parent genes or the retroCNV sequences. RetroCNV which had multiple insertions from the same parent gene were also excluded. Variant effect prediction was performed using the UCSC Genome Browser Variant Annotation Integrator tool [52].

RetroCNV validation

We performed retroCNV discovery using Illumina data aligned to CanFam3.1 on four individuals which were previously used to generate additional dog genome assemblies: UMICH_Zoey_3.1 [17], UU_Cfam_GSD_1.0 [53], Canfam_GSD [54], and CanLup_DDS [55]. We then assessed the presence of the retroCNVs within their respective assemblies using BLAST [56]. PacBio data was also examined for evidence of the retroCNV insertions. Individual long read FASTQ files were aligned to the CanFam3.1 reference with minimap2 version 2.17 [57]. Alignment files were sorted, merged, and indexed with SAMtools version 1.5 [51]. The predicted retroCNV insertion sites +/- 100bp were analyzed using a modified version of a pipeline designed to detect LINE-1 insertions in long read sequenced genomes. The pipeline extracts the raw long reads which align to a locus of interest and uses a combination of Canu and wtdbg2 to assemble the reads into contigs [58,59]. The contigs are then polished using Racon [60], aligned to the reference using minimap2 version 2.20 and put in orientation with the reference. Precise breakpoints were identified using with AGE [61]. We developed three primer PCR assays for retroCNVs using Primer3 software [62], with forward and reverse primers flanking the insertion site and internal primers at the 5’ or 3’ ends of the parent gene. A panel of 10 dogs from a breed identified as carrying each retroCNV were selected at random from a the Bannasch Lab DNA Repository for testing [24]. A list of the primers used in this study and their expected product sizes is available in Supplemental Table S16. Sanger sequencing was performed on an

Applied Biosystems 3500 Genetic Analyzer using a Big Dye Terminator Sequencing Kit (Applied Biosystems, Burlington, ON, Canada).

Population analysis

For F_{ST} calculations, dog breeds were placed into the multi-breed clades identified by Parker et al. [27]. Only clades containing individuals from at least three breeds were included in this analysis. Additionally, only three Golden Retrievers were selected at random to include in the retriever clade. F_{ST} between clades was calculated as described by Zhang et al. [13], including the calculation of a null distribution from 1,000 fake population sets generated through shuffling individual labels for significance estimates. RetroCNVs for which 1,000 fake population sets never produced an equal or higher F_{ST} than the real population were considered significant.

WGS and RNA-seq

Adult Golden Retriever testes were obtained from routine castration procedures. Tissue samples were flash frozen in liquid nitrogen and stored at -80˚C. Genomic DNA was extracted using a Gentra Puregene DNA extraction kit (Qiagen, Valencia, CA, USA), and RNA was extracted using an RNeasy Fibrous Tissue Mini Kit (Qiagen, Valencia, CA, USA). Library preparation and NovaSeq S4 Illumina paired end sequencing were performed at the UC Davis Genome Center. Reads were aligned to the CanFam3.1 reference assembly using minimap2 [57]. PCR duplicate reads were removed and the aligned files were sorted and indexed using SAMtools [51]. Evidence for chimeric transcripts in the RNA-seq dataset was found through visual analysis of the retroCNV insertion sites and nearby genes in IGV. Due to the small sample size and heterogeneity of retroCNVs between individuals, evidence of expression was determined visually through examination of the insert site, the 5' UTR of the parent gene, and insertion site genes for chimeric transcripts. For retroCNVs which contained any retroCNV specific variants, the parent gene loci were examined for evidence of the retroCNV specific variant which would indicate expression of the retroCNV. A minimum of two discordant reads was used to consider a retroCNV expressed.

Human comparative analysis

We performed retroCNV discovery in a subset of individuals from The 1000 Genomes Project Consortium high coverage phase 3 dataset [The 1000 Genomes Project Consortium 29,63]. We selected 3 individuals from each of 26 human populations at random for this analysis (supplemental table 8) and used the GRCh38 annotation release 109.20210514 for 30-mer construction. We then performed retroCNV discovery in the same manner as was performed on the canid dataset, and compared the retroCNVs identified by our approach to those identified in a previous study that used the same individuals [13]. We compared the retroCNVs identified in this study to those retroCNV identified in the same 78 individuals by Zhang et al as well as retroCNV identified in four other studies which used different datasets [6,7,10,26].

DATA ACCESS

The WGS and RNA-seq data generated in this study have been submitted to the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/) under accession PRJNA776905. The source code is available as supplemental material (Supplemental_Code.zip). The retroCNV insertion sites in bigBed format is available in the supplemental material.

ACKNOWLEDGEMENTS

This work was supported in part by the Lodric Maddox Graduate Fellowship, Maxine Adler Endowed Chair funds, and the Center for Companion Animal Health. Additional financial support was provided by the Students Training in Advanced Research (STAR) Program through a UC Davis School of Veterinary Medicine Endowment Fund. MB and JMK were supported by the National Institute of General Medical Sciences of the National Institutes of Health under award numbers R01GM140135 and T32GM007544. We thank Laura Kirby for helpful discussions related to the use of long read data to detect retrogene insertions. We thank Guide Dogs for the Blind Inc. San Rafael CA for the samples.

CHAPTER 4 REFERENCES

1. Richardson, S.R.; Salvador-Palomeque, C.; Faulkner, G.J. Diversity through duplication: Whole-genome sequencing reveals novel gene retrocopies in the human population. *Bioessays* **2014**, *36*, 475-481.
2. Jurka, J. Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. *Proceedings of the National Academy of Sciences* **1997**, *94*, 1872-1877.
3. Esnault, C.; Maestre, J.; Heidmann, T. Human LINE retrotransposons generate processed pseudogenes. *Nature genetics* **2000**, *24*, 363-367.
4. Frankish, A.; Diekhans, M.; Ferreira, A.-M.; Johnson, R.; Jungreis, I.; Loveland, J.; Mudge, J.M.; Sisu, C.; Wright, J.; Armstrong, J. GENCODE reference annotation for the human and mouse genomes. *Nucleic acids research* **2019**, *47*, D766-D773.
5. Rosikiewicz, W.; Kabza, M.; Kosiński, J.G.; Ciomborowska-Basheer, J.; Kubiak, M.R.; Makałowska, I. RetrogeneDB–a database of plant and animal retrocopies. *Database* **2017**, *2017*.
6. Schrider, D.R.; Navarro, F.C.; Galante, P.A.; Parmigiani, R.B.; Camargo, A.A.; Hahn, M.W.; de Souza, S.J. Gene copy-number polymorphism caused by retrotransposition in humans. *PLoS genetics* **2013**, *9*, e1003242.
7. Ewing, A.D.; Ballinger, T.J.; Earl, D.; Harris, C.C.; Ding, L.; Wilson, R.K.; Haussler, D. Retrotransposition of gene transcripts leads to structural variation in mammalian genomes. *Genome biology* **2013**, *14*, R22.
8. Kabza, M.; Kubiak, M.R.; Danek, A.; Rosikiewicz, W.; Deorowicz, S.; Polański, A.; Makałowska, I. Inter-population differences in retrogene loss and expression in humans. *PLoS genetics* **2015**, *11*, e1005579.

9. Penzkofer, T.; Jäger, M.; Figlerowicz, M.; Badge, R.; Mundlos, S.; Robinson, P.N.; Zemojtel, T. L1Base 2: more retrotransposition-active LINE-1s, more mammalian genomes. *Nucleic acids research* **2016**, gkw925.

10. Abyzov, A.; Iskow, R.; Gokcumen, O.; Radke, D.W.; Balasubramanian, S.; Pei, B.; Habegger, L.; Lee, C.; Gerstein, M.; Consortium, G.P. Analysis of variable retroduplications in human populations suggests coupling of retrotransposition to cell division. *Genome research* **2013**, *23*, 2042-2052.

11. Casola, C.; Betrán, E. The genomic impact of gene retrocopies: what have we learned from comparative genomics, population genomics, and transcriptomic analyses? *Genome biology and evolution* **2017**, *9*, 1351-1373.

12. Zhang, W.; Xie, C.; Ullrich, K.; Zhang, Y.E.; Tautz, D. The mutational load in natural populations is significantly affected by high primary rates of retroposition. *Proceedings of the National Academy of Sciences* **2021**, *118*.

13. Zhang, Y.; Li, S.; Abyzov, A.; Gerstein, M.B. Landscape and variation of novel retroduplications in 26 human populations. *PLoS computational biology* **2017**, *13*, e1005567.

14. Gao, X.; Li, Y.; Adetula, A.A.; Wu, Y.; Chen, H. Analysis of new retrogenes provides insight into dog adaptive evolution. *Ecology and Evolution* **2019**.

15. Kim, S.; Mun, S.; Kim, T.; Lee, K.-H.; Kang, K.; Cho, J.-Y.; Han, K. Transposable element-mediated structural variation analysis in dog breeds using whole-genome sequencing. *Mammalian Genome* **2019**, *30*, 289-300.

16. Batcher, K.; Dickinson, P.; Maciejczyk, K.; Brzeski, K.; Rasouliha, S.H.; Letko, A.; Drögemüller, C.; Leeb, T.; Bannasch, D. Multiple FGF4 retrocopies recently derived within canids. *Genes* **2020**, *11*, 839.

17. Halo, J.V.; Pendleton, A.L.; Shen, F.; Doucet, A.J.; Derrien, T.; Hitte, C.; Kirby, L.E.; Myers, B.; Sliwerska, E.; Emery, S. Long-read assembly of a Great Dane genome highlights the contribution of GC-rich sequence and mobile elements to canine genomes. *Proceedings of the National Academy of Sciences* **2021**, *118*.

18. Wang, W.; Kirkness, E.F. Short interspersed elements (SINEs) are a major source of canine genomic diversity. *Genome Research* **2005**, *15*, 1798-1808.

19. Troskie, R.-L.; Jafrani, Y.; Mercer, T.R.; Ewing, A.D.; Faulkner, G.J.; Cheetham, S.W. Long-read cDNA sequencing identifies functional pseudogenes in the human transcriptome. *Genome biology* **2021**, *22*, 1-15.

20. Cheetham, S.W.; Faulkner, G.J.; Dinger, M.E. Overcoming challenges and dogmas to understand the functions of pseudogenes. *Nature Reviews Genetics* **2020**, *21*, 191-201.

21. Ciomborowska-Basheer, J.; Staszak, K.; Kubiak, M.R.; Makałowska, I. Not So Dead Genes—Retrocopies as Regulators of Their Disease-Related Progenitors and Hosts. *Cells* **2021**, *10*, 912.

22. Parker, H.G.; VonHoldt, B.M.; Quignon, P.; Margulies, E.H.; Shao, S.; Mosher, D.S.; Spady, T.C.; Elkahloun, A.; Cargill, M.; Jones, P.G. An expressed fgf4 retrogene is associated with breed-defining chondrodysplasia in domestic dogs. *Science* **2009**, *325*, 995-998.

23. Brown, E.A.; Dickinson, P.J.; Mansour, T.; Sturges, B.K.; Aguilar, M.; Young, A.E.; Korff, C.; Lind, J.; Ettinger, C.L.; Varon, S. FGF4 retrogene on CFA12 is responsible for

chondrodystrophy and intervertebral disc disease in dogs. *Proceedings of the National Academy of Sciences* **2017**, *114*, 11476-11481.

24. Batcher, K.; Dickinson, P.; Giuffrida, M.; Sturges, B.; Vernau, K.; Knipe, M.; Rasouliha, S.H.; Drögemüller, C.; Leeb, T.; Maciejczyk, K. Phenotypic Effects of FGF4 Retrogenes on Intervertebral Disc Disease in Dogs. *Genes* **2019**, *10*, 435.
25. Wang, G.-D.; Shao, X.-J.; Bai, B.; Wang, J.; Wang, X.; Cao, X.; Liu, Y.-H.; Wang, X.; Yin, T.-T.; Zhang, S.-J. Structural variation during dog domestication: insights from gray wolf and dhole genomes. *National Science Review* **2019**, *6*, 110-122.
26. Feng, X.; Li, H. Higher Rates of Processed Pseudogene Acquisition in Humans and Three Great Apes Revealed by Long-Read Assemblies. *Molecular Biology and Evolution* **2021**, *38*, 2958-2966.
27. Parker, H.G.; Dreger, D.L.; Rimbault, M.; Davis, B.W.; Mullen, A.B.; Carpintero-Ramirez, G.; Ostrander, E.A. Genomic analyses reveal the influence of geographic origin, migration, and hybridization on modern dog breed development. *Cell reports* **2017**, *19*, 697-708.
28. Byrska-Bishop, M.; Evani, U.S.; Zhao, X.; Basile, A.O.; Abel, H.J.; Regier, A.A.; Corvelo, A.; Clarke, W.E.; Musunuri, R.; Nagulapalli, K. High coverage whole genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. **2021**.
29. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **2015**, *526*, 68-74.
30. Ostrander, E.A.; Wang, G.-D.; Larson, G.; Vonholdt, B.M.; Davis, B.W.; Jagannathan, V.; Hitte, C.; Wayne, R.K.; Zhang, Y.-P. Dog10K: an international sequencing effort to advance studies of canine domestication, phenotypes and health. *National science review* **2019**, *6*, 810-824.
31. Hoeppner, M.P.; Lundquist, A.; Pirun, M.; Meadows, J.R.; Zamani, N.; Johnson, J.; Sundström, G.; Cook, A.; FitzGerald, M.G.; Swofford, R. An improved canine genome and a comprehensive catalogue of coding genes and non-coding transcripts. *PloS one* **2014**, *9*, e91172.
32. Callinan, P.; Batzer, M. Retrotransposable elements and human disease. *Genome and disease* **2006**, *1*, 104-115.
33. Routtenberg, A.; Cantallops, I.; Zaffuto, S.; Serrano, P.; Namgung, U. Enhanced learning after genetic overexpression of a brain growth protein. *Proceedings of the National Academy of Sciences* **2000**, *97*, 7657-7662.
34. Bannasch, D.; Famula, T.; Donner, J.; Anderson, H.; Honkanen, L.; Batcher, K.; Safra, N.; Thomasy, S.; Rebhun, R. The effect of inbreeding, body size and morphology on health in dog breeds. *Canine Medicine and Genetics* **2021**, *8*, 12, doi:10.1186/s40575-021-00111-4.
35. Crosas, B.; Hyndman, D.J.; Gallego, O.; Martras, S.; Parés, X.; Flynn, T.G.; Farrés, J. Human aldose reductase and human small intestine aldose reductase are efficient retinal reductases: consequences for retinoid metabolism. *Biochemical Journal* **2003**, *373*, 973-979.
36. Jakobsson, P.-J.; Mancini, J.A.; Riendeau, D.; Ford-Hutchinson, A.W. Identification and characterization of a novel microsomal enzyme with glutathione-dependent transferase and peroxidase activities. *Journal of Biological Chemistry* **1997**, *272*, 22934-22939.

37. Nolan, J.; Campbell, P.J.; Brown, S.J.; Zhu, G.; Gordon, S.; Lim, E.M.; Joseph, J.; Cross, S.M.; Panicker, V.; Medland, S.E. Genome-wide analysis of thyroid function in Australian adolescents highlights SERPINA7 and NCOA3. *European Journal of Endocrinology* **2021**, *185*, 743-753.

38. Tsai, K.L.; Evans, J.M.; Noorai, R.E.; Starr-Moss, A.N.; Clark, L.A. Novel Y chromosome retrocopies in canids revealed through a genome-wide association study for sex. *Genes* **2019**, *10*, 320.

39. Wayne, R.K.; Ostrander, E.A. Lessons learned from the dog genome. *TRENDS in Genetics* **2007**, *23*, 557-567.

40. Gough, A.; Thomas, A.; O'Neill, D. *Breed predispositions to disease in dogs and cats*; John Wiley & Sons: John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA, 2018.

41. Asher, L.; Diesel, G.; Summers, J.F.; McGreevy, P.D.; Collins, L.M. Inherited defects in pedigree dogs. Part 1: disorders related to breed standards. *The Veterinary Journal* **2009**, *182*, 402-411.

42. Summers, J.F.; Diesel, G.; Asher, L.; McGreevy, P.D.; Collins, L.M. Inherited defects in pedigree dogs. Part 2: Disorders that are not related to breed standards. *The Veterinary Journal* **2010**, *183*, 39-45.

43. Schiffman, J.D.; Breen, M. Comparative oncology: what dogs and other species can teach us about humans with cancer. *Philosophical Transactions of the Royal Society B: Biological Sciences* **2015**, *370*, 20140231.

44. Pertea, G.; Pertea, M. GFF utilities: GffRead and GffCompare. *F1000Research* **2020**, *9*.

45. Marcais, G.; Kingsford, C. Jellyfish: A fast k-mer counter. *Tutorialis e Manuais* **2012**, *1*, 1-8.

46. Hach, F.; Sarrafi, I.; Hormozdiari, F.; Alkan, C.; Eichler, E.E.; Sahinalp, S.C. mrsFAST-Ultra: a compact, SNP-aware mapper for high performance sequencing applications. *Nucleic acids research* **2014**, *42*, W494-W500.

47. Carreira, P.E.; Ewing, A.D.; Li, G.; Schauer, S.N.; Upton, K.R.; Fagg, A.C.; Morell, S.; Kindlova, M.; Gerdes, P.; Richardson, S.R. Evidence for L1-associated DNA rearrangements and negligible L1 retrotransposition in glioblastoma multiforme. *Mobile DNA* **2016**, *7*, 1-14.

48. Thorvaldsdóttir, H.; Robinson, J.T.; Mesirov, J.P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in bioinformatics* **2013**, *14*, 178-192.

49. Krzywinski, M.; Schein, J.; Birol, I.; Connors, J.; Gascoyne, R.; Horsman, D.; Jones, S.J.; Marra, M.A. Circos: an information aesthetic for comparative genomics. *Genome research* **2009**, *19*, 1639-1645.

50. Rausch, T.; Zichner, T.; Schlattl, A.; Stütz, A.M.; Benes, V.; Korbel, J.O. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **2012**, *28*, i333-i339.

51. Danecek, P.; Bonfield, J.K.; Liddle, J.; Marshall, J.; Ohan, V.; Pollard, M.O.; Whitwham, A.; Keane, T.; McCarthy, S.A.; Davies, R.M. Twelve years of SAMtools and BCFtools. *Gigascience* **2021**, *10*, giab008.

52. Hinrichs, A.S.; Raney, B.J.; Speir, M.L.; Rhead, B.; Casper, J.; Karolchik, D.; Kuhn, R.M.; Rosenbloom, K.R.; Zweig, A.S.; Haussler, D. UCSC data integrator and variant annotation integrator. *Bioinformatics* **2016**, *32*, 1430-1432.
53. Wang, C.; Wallerman, O.; Arendt, M.-L.; Sundström, E.; Karlsson, Å.; Nordin, J.; Mäkeläinen, S.; Pielberg, G.R.; Hanson, J.; Ohlsson, Å. A novel canine reference genome resolves genomic architecture and uncovers transcript complexity. *Communications biology* **2021**, *4*, 1-11.
54. Field, M.A.; Rosen, B.D.; Dudchenko, O.; Chan, E.K.; Minoche, A.E.; Edwards, R.J.; Barton, K.; Lyons, R.J.; Tuipulotu, D.E.; Hayes, V.M. Canfam_GSD: De novo chromosome-length genome assembly of the German Shepherd Dog (Canis lupus familiaris) using a combination of long reads, optical mapping, and Hi-C. *GigaScience* **2020**, *9*, giaa027.
55. Field, M.A.; Yadav, S.; Dudchenko, O.; Esvaran, M.; Rosen, B.D.; Skvortsova, K.; Edwards, R.J.; Keilwagen, J.; Cochran, B.J.; Manandhar, B. The Australian dingo is an early offshoot of modern breed dogs. *Science advances* **2022**, *8*, eabm5944.
56. Madden, T. The BLAST sequence analysis tool. *The NCBI handbook* **2013**, *2*, 425-436.
57. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **2018**, *34*, 3094-3100.
58. Ruan, J.; Li, H. Fast and accurate long-read assembly with wtdbg2. *Nature methods* **2020**, *17*, 155-158.
59. Koren, S.; Walenz, B.P.; Berlin, K.; Miller, J.R.; Bergman, N.H.; Phillippy, A.M. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome research* **2017**, *27*, 722-736.
60. Vaser, R.; Sović, I.; Nagarajan, N.; Šikić, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome research* **2017**, *27*, 737-746.
61. Abyzov, A.; Gerstein, M. AGE: defining breakpoints of genomic structural variants at single-nucleotide resolution, through optimal alignments with gap excision. *Bioinformatics* **2011**, *27*, 595-603.
62. Untergasser, A.; Cutcutache, I.; Koressaar, T.; Ye, J.; Faircloth, B.C.; Remm, M.; Rozen, S.G. Primer3—new capabilities and interfaces. *Nucleic acids research* **2012**, *40*, e115-e115.
63. Sudmant, P.H.; Rausch, T.; Gardner, E.J.; Handsaker, R.E.; Abyzov, A.; Huddleston, J.; Zhang, Y.; Ye, K.; Jun, G.; Fritz, M.H.-Y. An integrated map of structural variation in 2,504 human genomes. *Nature* **2015**, *526*, 75-81.

CHAPTER 5: An *SNN* retrocopy insertion upstream of *GPR22* is associated with dark red coat color in Poodles

Kevin Batcher[1], Scarlett Varney[1], Verena K. Affolter[2], Steven G. Friedenberg[3], and Danika Bannasch[1]*

[1] Department of Population Health and Reproduction, University of California, Davis, CA, USA

[2] Department of Pathology, Microbiology, & Immunology, University of California, Davis, CA, USA

[3] Department of Veterinary Clinical Sciences, University of Minnesota, Saint Paul, MA, USA

G3 Genes|Genomes|Genetics, jkac227, https://doi.org/10.1093/g3journal/jkac227

Received: 12 July 2022; Accepted: 27 August 2022; Published: 1 September 2022

AUTHOR CONTRIBUTIONS

Conceptualization, D.B. and K.B.; Data curation, K.B. and S.V.; Formal analysis, V.A., K.B., and S.V.; Funding acquisition, D.B. and S.F.; Investigation, V.A., K.B., and S.V.; Project administration, D.B.; Resources, D.B. and S.F.; Supervision, D.B.; Visualization, V.A., K.B., and S.V.; Writing—original draft, K.B. and S.V.; Writing—review & editing, D.B., K.B., V.A., and S.V.

ABSTRACT

Pigment production and distribution is controlled through multiple genes, resulting in a wide range of coat color phenotypes in dogs. Dogs that produce only the pheomelanin pigment vary in intensity from white to deep red. The Poodle breed has a wide range of officially recognized coat colors, including the pheomelanin-based white, cream, apricot and red coat colors, which are not fully explained by the previously identified genetic variants involved in pigment intensity. Here, a genome-wide association study for pheomelanin intensity was performed in Poodles which identified an association on canine chromosome 18. Whole

genome sequencing data revealed an *SNN* retrocopy insertion (*SNNL1)* in apricot and red Poodles within the associated region on chromosome 18. While equal numbers of melanocytes were observed in all Poodle skin hair bulbs, higher melanin content was observed in the darker Poodles. Several genes involved in melanogenesis were also identified as highly overexpressed in red Poodle skin. The most differentially expressed gene however was *GPR22*, which was highly expressed in red Poodle skin while unexpressed in white Poodle skin ($\log_2$ fold-change in expression 6.1, P<0.001). *GPR22* is an orphan G-protein coupled receptor normally expressed exclusively in the brain and heart. The *SNNL1* retrocopy inserted 2.8kb upstream of *GPR22* and is likely disrupting regulation of the gene, resulting in atypical expression in the skin. Thus, we identify the *SNNL1* insertion as a candidate variant for the CFA18 pheomelanin intensity locus in red Poodles.

INTRODUCTION

In dogs, as with other mammals, coat color patterns are the result of varied production of the yellow-red pigment, pheomelanin, and the black pigment, eumelanin. While most dogs produce a mixture of both pigments, loss of function mutations in the pigment-type switching genes melanocortin 1 receptor (*MC1R*) and agouti signaling protein (*ASIP*) result in production of only one pigment type [1-3]. Among pheomelanin-based dogs, pigment intensity can vary greatly within and between breeds, from white to deep red [4]. Multiple genetic variants that modify pheomelanin pigment intensity have been identified in dogs, highlighting the complex, multigenic nature of coat color phenotypes. A missense variant in the *MFSD12* gene and a copy number variant near KITLG have both been associated with pheomelanin intensity in a variety of breeds [5,6]. An across breed analysis of pheomelanin intensity that was published while the current study was being performed identified that genetic variants at 5 loci explained 70% of pheomelanin intensity in dogs, which included the variants at *MFSD12* and *KITLG* as well as 3 novel loci on canine chromosomes (CFA) 2, CFA18 and CFA21 [7]. However, it is unclear how much each of the loci contribute to pheomelanin intensity within individual breeds.

The Poodle breed has 3 size varieties (toy, miniature, and standard) and 11 coat colors that are officially recognized by the American Kennel Club, 4 of which are pheomelanin-based: white, cream, apricot, and red (www.akc.org). While the *MFSD12* dilution variant was present in the white Poodles, it alone does not explain the range of pheomelanin intensity between the cream, apricot and red Poodles [5]. Additionally, the copy number variant near KITLG, which was associated with pigment intensity in the pheomelanin-based Nova Scotia Duck Tolling Retrievers and the eumelanin-based silver and black Poodles, was not found to be associated with pigment intensity between the pheomelanin-based white and red Poodles, indicating that additional genetic factors affecting pheomelanin intensity exist within the Poodle breed [6].

In this study, the genetics of pheomelanin intensity was analyzed within a single breed, the Poodle. A quantitative genome-wide association study (GWAS) was performed and a single associated locus on CFA18 was identified. An *SNN* gene retrocopy insertion was then identified as the most likely causative variant behind pheomelanin intensity in Poodles.

METHODS

Sample collection

Collection of all Poodle samples (N=225) was approved by the University of California, Davis Animal Care and Use Committee (protocol #18561). Breed, date of birth, sex, weight, and color were reported by the owner. Owners provided whole blood or buccal swabs from their privately owned dogs (58 white, 17 cream, 3 apricot, 6 red) in collaboration with the Poodle Club of America Foundation (Grant #A182159001), and DNA was extracted using a Gentra Puregene DNA extraction kit (Qiagen, Valencia, CA, USA). Additional Poodle DNA samples from the Bannasch lab DNA repository at UC Davis were also included in the study (67 white, 9 cream, 23 apricot, 42 red). RNA from 8 red and 9 white Poodles was extracted from neonatal canine dewclaw samples using an RNeasy Fibrous Tissue Mini Kit (Qiagen, Valencia, CA, USA).

Genome wide association

In order to perform a quantitative GWAS, Poodles were designated from 1 to 4 based on owner described coat color (also the AKC registered coat color), with white as '1', cream as '2', apricot as '3' and red as '4'. Genome-wide SNV genotyping was performed on the Illumina Canine HD BeadChip array. All dogs were confirmed homozygous for the recessive yellow 'e' allele at MC1R with the exception of two cream Poodles which were heterozygous and thus excluded from further analysis [1]. Variants with a minor allele frequency of less than 5% or less than 90% total genotyping rate were excluded using PLINK, resulting in 163,753 total variants [8]. The Bonferroni corrected genome-wide significance threshold was set at P = 3.05 x 10-7. A multidimensional scaling plot showed that standard Poodles clustered separately from the toy and miniature Poodles, highlighting population stratification in the dataset (Figure S1). To control for this, the GWAS was performed using a univariate mixed model with a standardized relatedness matrix in GEMMA v.0.97 [9]. A similar GWAS using only the miniature and toy Poodles (N=57) was also performed using GEMMA.

Variant Detection

Whole Genome Sequencing (WGS) data from standard Poodles (7 white, 1 apricot and 1 red) were aligned to UU_Cfam_GSD_1.0 [10] using BWA v0.717 and converted to BAM files using samtools v1.14, both with default parameter settings [11]. Variant calling across the critical interval was performed using bcftools mpileup [12]. Based on the GWAS results, the assumed inheritance pattern was alternate homozygotes for the red and white Poodles and heterozygous for the apricot Poodle. The WGS samples were confirmed to match this pattern at the top four GWAS SNV. Variants were tested for function using the Ensembl variant effect predictor [13] with the UU_Cfam_GSD_1.0 annotation. Missense variants were tested for function using SIFT and Polyphen-2 [14,15]. The region was also analyzed for structural variants through visual analysis of the alignment files of a single red Poodle in comparison to a single white Poodle using Integrative Genomics Viewer (IGV) [16]. To maintain consistency with the variants reported from the GWAS, all genomic locations were reported as their location in the CanFam3.1 reference.

Genotyping and Sanger Sequencing

A three primer PCR assay was developed for genotyping *SNNL1*, with a forward and reverse primer flanking the insertion site and one primer internal to the retrocopy. Internal primers were then used for sequencing the entire retrocopy. Primers were also designed for genotyping the *SLC26A4* chr18:12,910,382 C/T variant. All primers were developed using Primer3 software [17]. The primers used in this study are available in Table S1. Sanger sequencing was performed on an Applied Biosystems 3500 Genetic Analyzer using a Big Dye

Terminator Sequencing Kit (Life Technologies, Burlington, ON, Canada). Additional genotyping of *SNNL1* from WGS data was performed visually in IGV. Samples which had no reads crossing either of the breakends at the insertion site were considered homozygous for the retrocopy insertion.

Histopathological and immunohistochemical examinations

Submitted skin samples from the dewclaws of a white Poodle (*SNNL1* 0 copies), a cream Poodle (*SNNL1* 1 copy), and a red Poodle (*SNNL1* 2 copies) were used for histopathological analysis. The samples were fixed in 4% buffered formalin, bisected and embedded in paraffin. Five-micron paraffin sections were used for both histopathology and immunohistochemistry. Presence of melanin granules within matrical cells of the hair follicles as well as within hair shafts was assessed by Fontana-Masson's stain. Anti-Sox10 antibody (mouse monoclonal, Abcam Ref. ab212843), which recognizes cells of neural crest origin, was used to identify melanocytes among the matrical cells within hair bulbs of anagen hair follicles, which are actively forming a new hair shaft. For immunohistochemistry, sections were deparaffinized (xylene: 10 min 2x, followed by 100% ethanol: 1 min 3x, 95% ethanol: 1 min and 70% ethanol: 1min), followed by quenching of endogenous peroxidase (500 ul 10% sodium azide; 500 ul 30% hydrogen peroxide in 50 ml PBS; 25 min at room temperature) and three rinses in PBS. Antigen retrieval was performed by immersing slides in preheated antigen retrieval solution (1x Dako Target Retrieval Solution; stock solution S1699, pH6 at 95 to 100oC; 5 min). Slides were then cooled down to room temperature and washed three times in PBS. After exposing slides to 10% horse serum in PBS (15 min), the anti-Sox10 antibody (mouse monoclonal, Abcam Ref. ab212843) was applied at a 1:100 dilution for one hour. After three rinses in PBS the following steps were performed: 1) application of ImmPRESS HRP Horse Anti-Mouse IgG Polymer Reagent (Vector Cat.# MP-7402; 30 min), 2) thorough PBS rinses, and 3) addition of substrate (Vector, SK-4800). Development was monitored microscopically and reaction was stopped by immersing the slides in Milli-Q/distilled water. Counterstain in Gill's Hematoxylin #2 (RICCA, 3536-16; 15-30 s) was stopped by washing slides in running tap water. Slides were then cover-slipped using Shandon-Mount media (Thermo Scientific, 1900331).

RNAseq analysis

Poly(A) capture RNAseq Library preparation and NovaSeq S4 Illumina paired end sequencing were performed in three red and one white Poodle at the UC Davis Genome Center. RNAseq data was aligned to UU_Cfam_GSD_1.0 [10] using minimap v2.21 [18]. Alignment files were analyzed for evidence of chimeric transcripts using IGV. Batch 3' TagSeq library preparation and HiSeq 4000 Illumina single end sequencing were performed on eight red and nine white Poodles at the UC Davis Genome Center. TaqSeq generates a single initial library molecule per transcript which is ideal for differential gene expression analysis [19]. Unique molecular identifiers were removed from the TagSeq data using UMI-tools [20], and reads were also trimmed to remove Illumina adaptors and polyA read through using bbduk [21]. Reads were aligned to UU_Cfam_GSD_1.0 [10] using STAR v2.7.9a [22]. The UU_GSD_1.0 annotation was used to perform gene counting with htseq-count [23]. Genes with overlapping 3' UTR in the annotation resulted in reads not being counted due to ambiguity; therefore, the '--nonunique all' option was used, which counts ambiguous reads to all overlapping features. Differential gene expression was performed using Limma-Voom [24] and is reported as log2 fold-change (FC) increase in expression in the red Poodles, where a negative FC indicates higher expression in white Poodles. Genes which had fewer than 5 normalized read counts across all samples were filtered. Different individuals were used in the RNAseq and TagSeq analyses than those used for variant discovery in the WGS analysis.

RESULTS

Genome wide association for pheomelanin intensity in Poodles

To identify regions of the genome associated with pheomelanin intensity specific to Poodles, a quantitative GWAS was performed in white (N=51), cream (N=5), apricot (N=15) and red (N=8) Poodles (Figure 1A). A single locus on chromosome 18 reached genome-wide significance (Figure 1B). A Q-Q plot of expected and observed chi-squared values indicated that population stratification was successfully controlled for (λ=1.014; Figure S2). The top four associated variants, shown in table 1, were in near perfect LD. Analysis of linkage disequilibrium (LD) between the top associated SNV (chr18:16,968,786) and nearby variants revealed a large region of LD in Poodles (Figure 1C). To determine if population structure within the dataset was affecting the association, a separate GWAS using only the miniature and toy Poodles (N=57) was performed, which identified the same top four SNVs and confirmed the CFA18 association with pheomelanin intensity (Figure S3).

CanFam3.1 position	Red Allele	AF white	AF cream	AF apricot	AF red	P_{LRT}
chr18:16968786	G	0	0.200	0.500	0.722	5.11×10^{-23}
chr18:17006104	A	0	0.200	0.500	0.722	5.11×10^{-23}
chr18:12910382	T	0	0.100	0.500	0.722	6.59×10^{-23}
chr18:13022106	G	0	0.100	0.500	0.722	6.59×10^{-23}

Table 1: Top genetic markers associated with red coat color in Poodles (CanFam3.1). AF = Allele frequency of the red associated allele in each category.

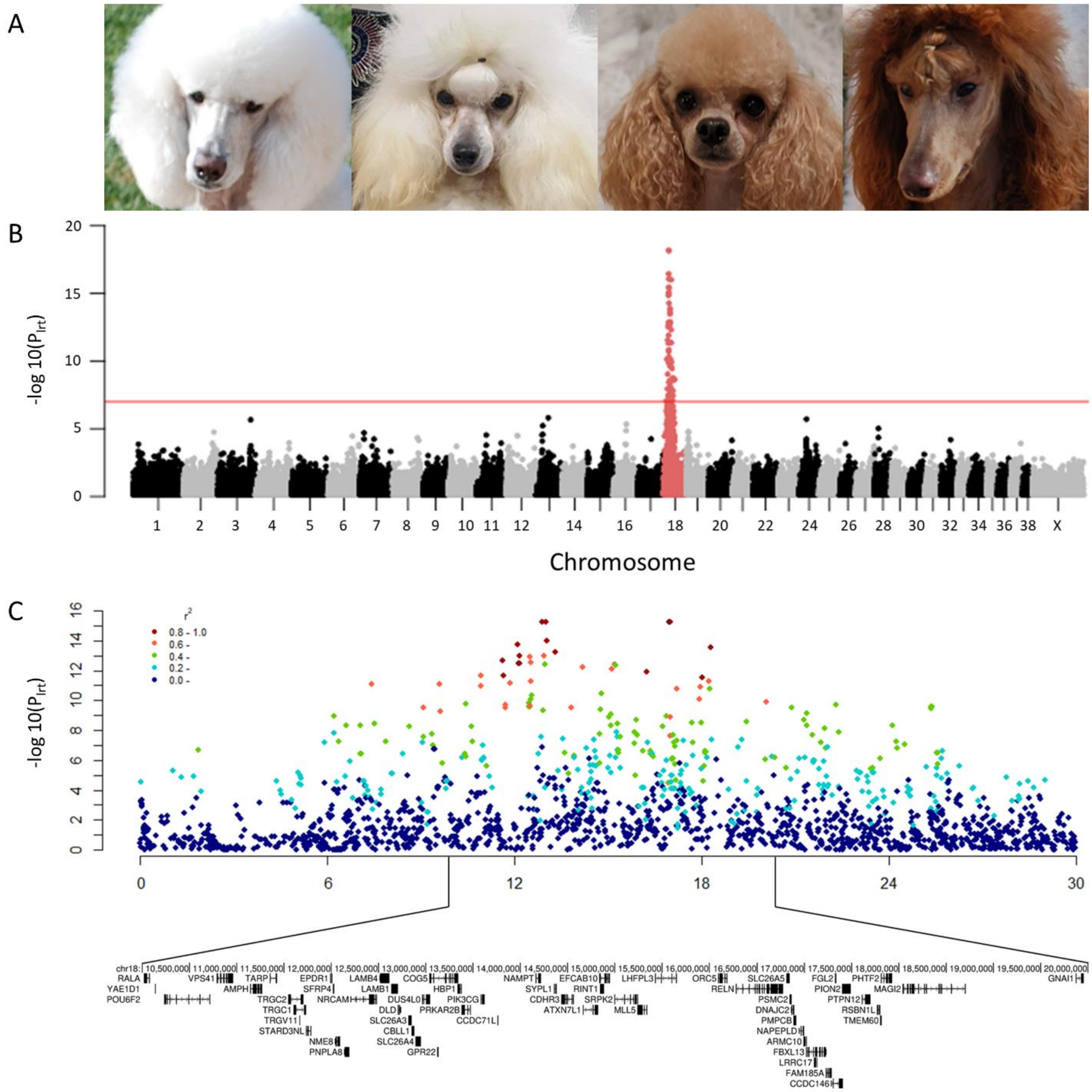


Figure 1: Genome-wide association for red coat color in Poodles. (A) Owner provided photographs of a white, cream, apricot, and red Poodle. (B) Manhattan plot showing P values for likelihood-ratio tests calculated in GEMMA (λ=1.014). Bonferroni corrected Genome-wide significance is indicated by the solid red line. (C) The genome-wide significant region on chromosome 18 with nearby SNV colored based on linkage disequilibrium (r2) with one of the top associated SNV (chr18:16,968,786).

Whole genome sequencing analysis

Variants within the interval of CanFam3.1 chr18:10-20mb were analyzed from WGS data in 9 standard Poodles (7 white, 1 apricot, 1 red). Out of 47,095 total variants identified, 5,603 segregated by phenotype, including 4,834 SNV and 762 short indels. Variant effect prediction identified missense variants in 5 genes (Table 2), including a previously reported variant in the *SLC26A4* gene (chr18:12,910,382 T>C) that was associated with pheomelanin intensity across breeds [7]. While the two missense variants in *ARMC10* and *GSAP* were predicted to be deleterious by both SIFT and Polyphen-2, none of the missense variants affected genes known to be involved in any pigment pathways. Therefore, visual analysis of the aligned sequence data was also performed to identify larger structural variants. A cluster of discordant reads was observed in the red and apricot Poodles at approximately chr18:13,134,000-13,134,500 which mapped to the Stannin (*SNN*) gene locus (chr6:31,137,750-31,147,848), highlighting a putative retrocopy insertion (Figure 2A).

CanFam3.1 Location	Gene	Amino acid	Codons	dbSNP	SIFT	Polyphen-2
chr18:12529314	*LAMB4*	G/E	gGa/gAa	rs852976135	Tolerated(0.44)	Benign(0.035)
chr18:12910382	*SLC26A4*	I/M	atA/atG	rs852750854	Deleterious(0.03)	Benign(0.014)
chr18:14468631	*CDHR3*	G/R	Ggg/Agg	rs22643100	Tolerated(0.42)	Benign(0.011)
chr18:17006104	*ARMC10*	F/C	tTt/tGt	rs853061060	Deleterious(0)	Damaging(1.0)
chr18:17469333	*GSAP*	D/N	Gat/Aat	rs850968557	Deleterious(0)	Damaging(1.0)

Table 2: Missense variants identified in red Poodles.

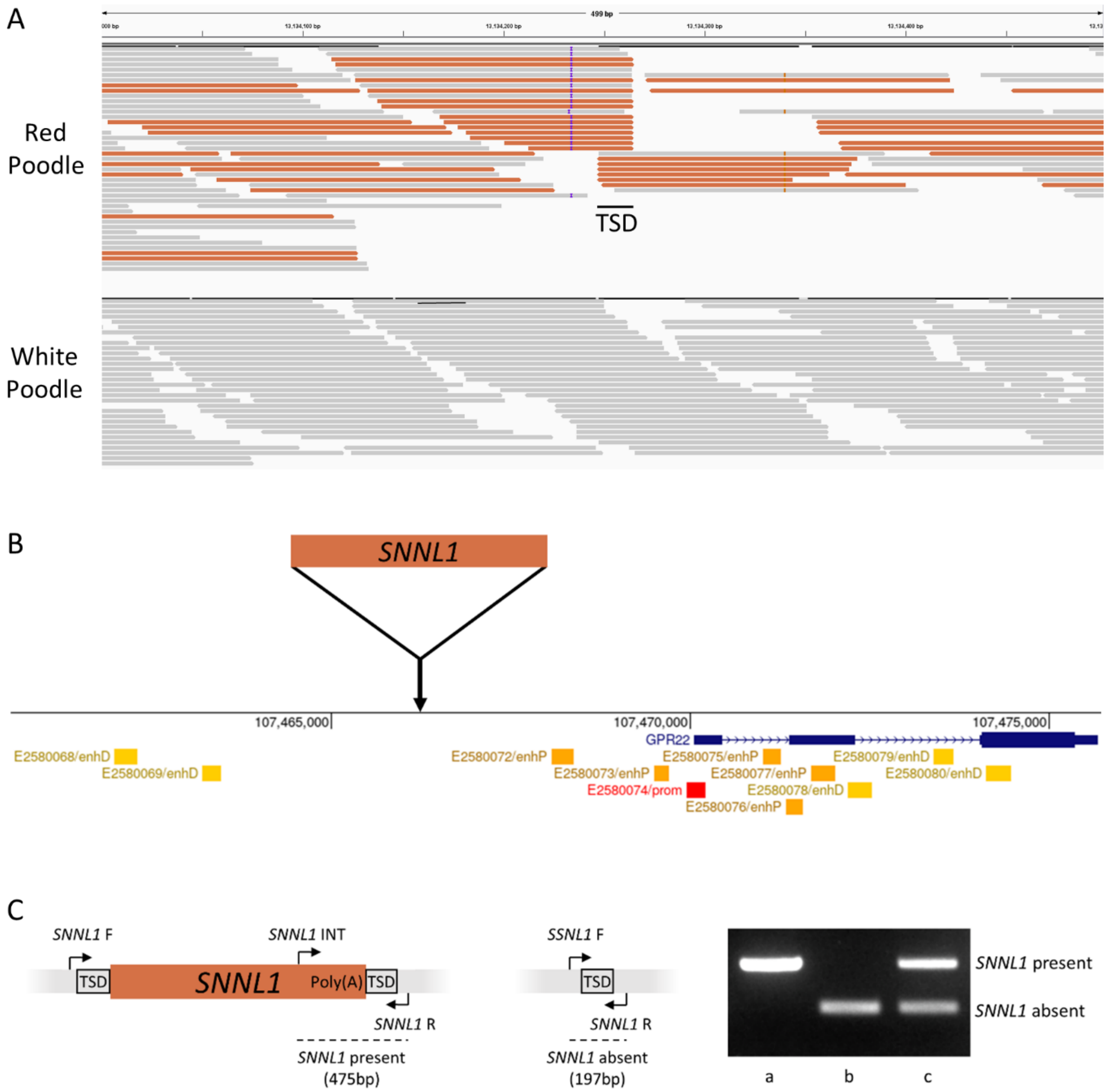


Figure 2: An *SNN* retrocopy identified in red Poodles. (A) A discordant read cluster observed in IGV indicating the presence of an *SNN* retrocopy insertion in a red Poodle (top track) which was absent from white Poodles (bottom track). (B) The location of the *SNNL1* retrocopy insertion within the syntenic region in humans. (C) Two external primers flanking the *SNNL1* insertion and one internal primer are used to genotype *SNNL1*. When *SNNL1* is present, a 475bp product is observed (a), while a 197bp product is observed when *SNNL1* is absent (b). Heterozygous individuals have both bands (c).

SNN retrocopy analysis

The putative *SNN* retrocopy was investigated using primers flanking the insertion site to PCR amplify the region in a red Poodle. Sanger sequencing confirmed the insertion as a full length *SNN* retrocopy (File S1), referred to here as *SNNL1*. *SNNL1* is inserted within the intron of COG5 and 2.8kbp upstream of and in the same orientation as *GPR22*. The *SNNL1* retrocopy sequence contains two SNV in the 3' UTR (chr6:31,139,403 C>A and chr6:31,140,045 G>A), but is otherwise identical to the parent gene sequence. *SNNL1* has a 3' poly (A) tail approximately 27bp in length, and a 17bp target site duplication (TGTGAAATACTGAAGTT) was also observed flanking the insertion, putting the exact insertion location at chr18:13,134,248-13,134,264. The syntenic region in humans for *SNNL1* was viewed to determine its location relative to regulatory elements. *SNNL1* inserted 2.8kb upstream of *GPR22*, nearby multiple predicted *GPR22* enhancers (Figure 2B).

Genotyping SNNL1 and the SLC26A4 missense variant

A three primer PCR genotyping assay was developed for *SNNL1* (Figure 2C). The retrocopy was then genotyped in a larger dataset of white, cream, apricot and red Poodles to test the association with coat color (N=224). *SNNL1* copy number was highly predictive of red coat color in the breed (adjusted R2=0.840, P=2.17x10-90) (Table 3). All (N=125) white Poodles had 0 copies of *SNNL1*, and all red Poodles (N=48) had at least one copy of *SNNL1*, with 38/48 of them having 2 copies. Most (19/25) apricot Poodles had 1 copy of *SNNL1*. The allele frequencies were 0.096 in cream, 0.500 in apricot, and 0.896 in red Poodles, indicating an additive effect on pheomelanin intensity. The nearby missense variant in *SLC26A4* (chr18:12,910,382 T>C) was also genotyped in the same set of dogs to access LD in the region, and the ‘C’ allele and the *SNNL1* insertion were found to be in complete LD in the white, apricot and red Poodles, however, 1 cream Poodle was identified with 0 copies of *SNNL1* that was heterozygous for the *SLC26A4* variant.

The linkage between the chr18:12,910,382 T>C variant and *SNNL1* was further assessed in a publicly available WGS dataset [25]. The ‘C’ allele was observed in Tibetan Mastiffs, Chow Chows, village dogs, and a Xoloitzcuintli, Qingchuan and Chongqing dog (Table S2). *SNNL1* was also genotyped in these same breeds through visual analysis of the aligned whole genome sequencing data, and while *SNNL1* was in strong LD with chr18:12,910,382 T>C, 8 village dogs and 1 Tibetan Mastiff were identified that have the SNV but do not appear to have *SNNL1*, indicating that linkage between the two is incomplete (Table S2). Although Tibetan Mastiffs, Chow Chows, Xoloitzcuintli, Qingchuan and Chongqing dogs all have deep red pheomelanin segregating within the breeds, we did not have access to phenotype data for the dogs from the WGS to confirm any associations.

Coat Color	***SNNL1* Copy number**			**Total**	**Allele frequency**
	0	**1**	**2**		
White	125	0	0	125	0.000
Cream	21	5	0	26	0.096
Apricot	3	19	3	25	0.500
Red	0	10	38	48	0.896

Table 3: *SNNL1* copy number in white, cream, apricot and red Poodles.

Histopathological and immunohistochemical examinations

Histopathological analysis was performed in skin tissue from a white (0 copies *SNNL1*), cream (1 copy *SNNL1*), and red (2 copies *SNNL1*) Poodle. Expression of Sox10, identifying melanocytes within the hair bulb, was observed in all Poodles irrespective of coat color or *SNNL1* copy number (Figure 3 a,d,g). However, the white Poodle with 0 copies of *SNNL1* lacked melanin within the hair bulbs and hair shaft cuticle (Figure 3 b,c). Some melanin was observed in a cream Poodle with one copy of *SNNL1* (Figure 3 e,f), but melanin was most prominent in the red Poodles with two copies of *SNNL1* (Figure 3 h,i). The equivalent melanocytes and differential melanin indicated that the red coat color was occurring due to an increase in pigment synthesis.

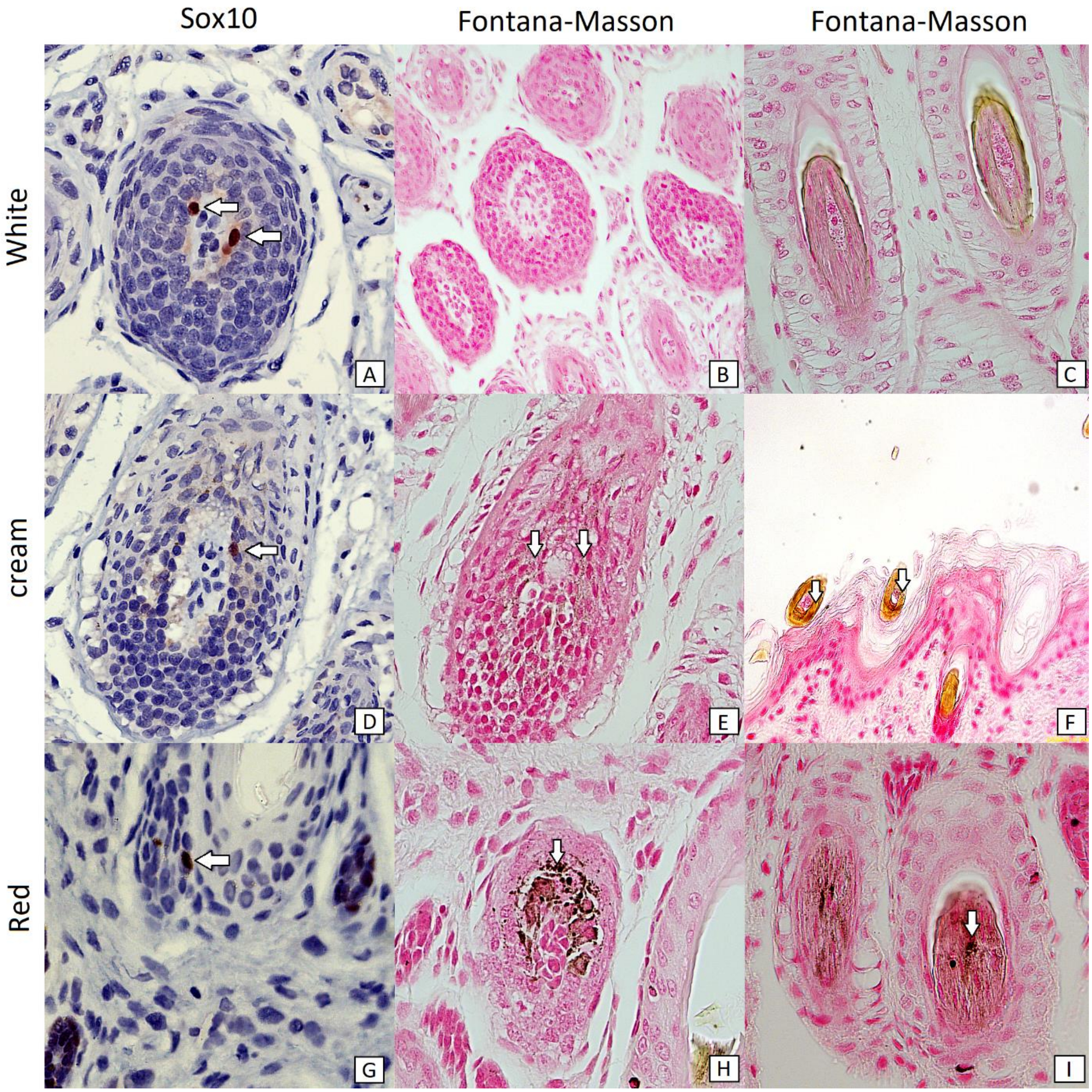


Figure 3: Presence of melanocytes within hair bulbs and melanin granules within matrical cells and hair shafts. Melanocytes were identified by the presence of SOX10, and melanin was identified by Fontana-Mason staining. White Poodle with zero copies of *SNNL1*: despite the presence of melanocytes (A) in the hair bulb, no pigment is noted in the matrical cells of the hair bulb (B) or the cuticle of hair shafts (C). Cream Poodle with one copy of *SNNL1*: in addition to melanocytes (D) in the hair bulb, there is fine melanin dusting of matrical cells of the hair bulb (E). The hair shaft cuticle contains melanin granules (F). Red Poodle with two copies of *SNNL1*: in addition to melanocytes (G) in the hair bulb, there is marked presence of melanin granules in matrical cells of the hair bulb with melanin pigment (H) as well as in hair shaft cuticle (I).

Gene expression analysis in Poodle skin

Gene expression was analyzed in red and white Poodle skin. Poly(A) capture RNAseq was first performed in 3 red Poodles and 1 white Poodle to determine if *SNNL1* was forming novel chimeric transcripts with the nearby genes *GPR22* and COG5, and no novel chimeric transcripts were observed in any samples. Overall differences in expression were then analyzed in 8 red and 9 white Poodle skin samples using TagSeq (Table S3). Among the most highly overexpressed genes in the red Poodles were several genes involved in melanogenesis, including TYR, PMEL, MLANA, SLC24A5, and MC1R (Figure 4). Notably, among genes involved in the production of eumelanin, TYRP1 was not expressed in either red or white poodle skin, and no changes in expression were observed for DCT. The most differentially expressed gene in red Poodle skin was *GPR22*, which was unexpressed in the white Poodles and highly expressed in the red Poodles (FC 6.1; adjusted P=0.00042). *SNN* also had small but significantly increased expression in the red Poodle skin (FC 0.49; adjusted P=0.0315), as did *COG5* (FC 0.41; adjusted P=0.0433). Among the genes with missense variants in the red Poodles, neither *SLC26A4* nor CDHR3 had sufficient expression in either the red or white Poodle skin to allow for differential expression analysis. However, while low levels of differential expression were observed in *LAMB4* (FC 1.2; adjusted P= 0.0024) and GSAP (FC -1.1; adjusted P=0.0006). *GPR22*, with a FC of 6.1, was the only gene within the chr18:10-20mb interval that had greater than 2 FC in expression in the red Poodles.

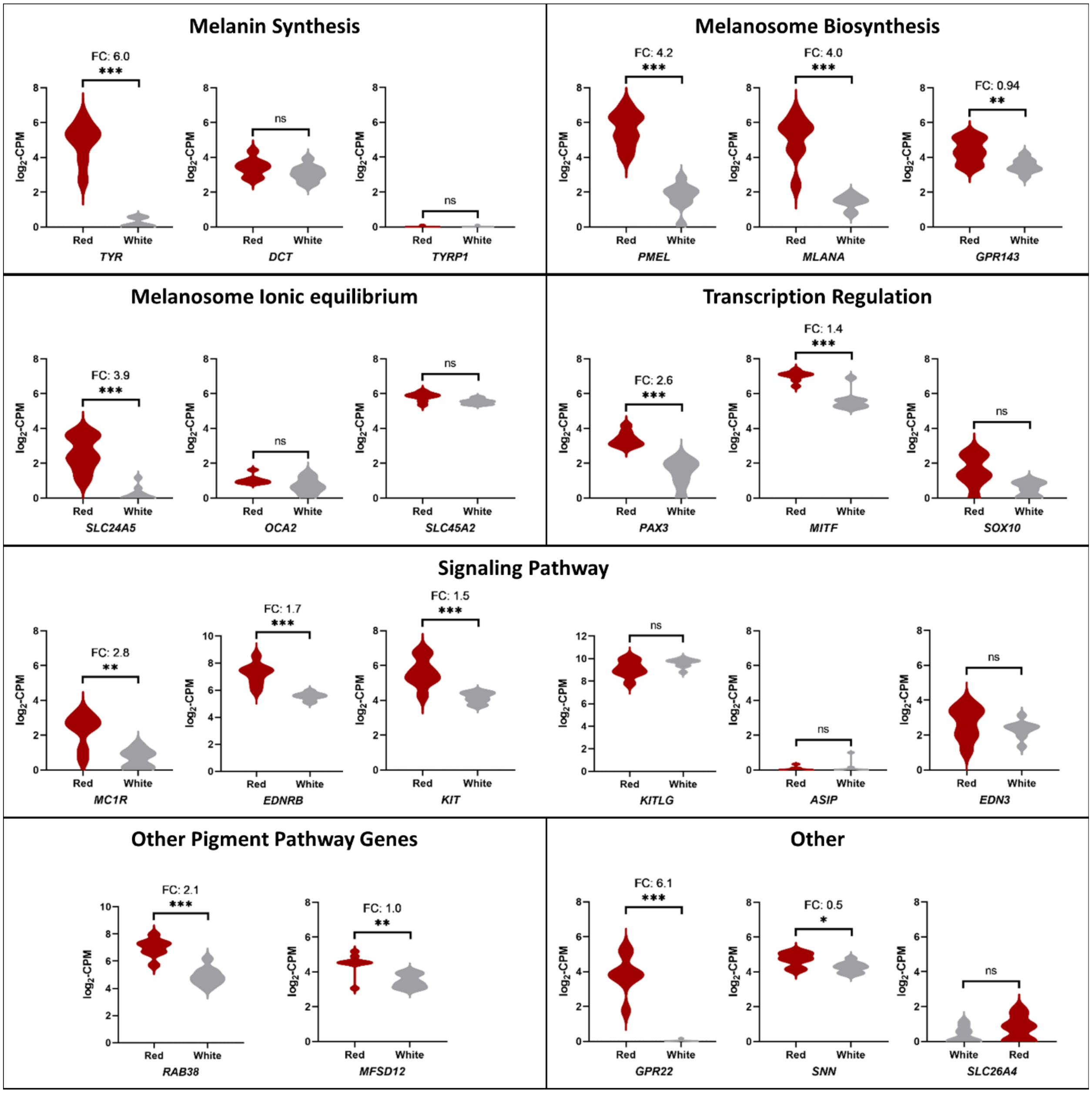


Figure 4: Differential gene expression analysis in Poodle skin tissue. Various genes involved in pigment production were overexpressed in the red Poodles (N=8) compared to the white Poodles (N=9). *GPR22* and *SNN* are also enriched in red Poodle skin.

DISCUSSION

Here we report the discovery of a recent *SNN* gene retrocopy insertion (*SNNL1*) which is 2.8kb upstream from *GPR22* and is strongly associated with pheomelanin intensity in Poodles. *SNNL1* was part of a large LD block in Poodles which was identified through quantitative GWAS of pheomelanin colors in Poodles. Missense variants in 5 genes were identified in red Poodles within this region, including a variant in *SLC26A4* which had previously been reported as a candidate for pheomelanin intensity across breeds. However, *SLC26A4* was not expressed in Poodle skin, whereas *GPR22* was identified as the most differentially expressed gene between red and white Poodles. *SNNL1*, which has inserted just upstream of *GPR22*, appears to interfere with regulation of the gene, resulting in ectopic expression of *GPR22* in skin tissue. This study implicates *GPR22* as being involved in the pigment production pathway, and the misregulation of *GPR22* via the insertion of *SNNL1* is likely the causal genetic influence behind red coat color in the Poodle breed.

Retrocopy insertions are a type of large structural variant which can be expressed directly, form chimeric transcripts with nearby genes, or otherwise interrupt the typical expression patterns of nearby genes [26]. The *SNN* retrocopy insertion, *SNNL1*, is a full length copy of the parent gene. Additionally, it has no coding sequence variants differentiating it from the parent gene sequence, indicating that it is likely a recent insertion. *SNN* codes for Stannin, a metal ion binding mitochondrial membrane protein which may be involved in response to toxic substance as well as cell growth and apoptosis [27,28]. No other *SNN* retrocopies are present in an across species database of reference genome retrocopies, indicating that *SNN* is not a commonly retrotransposed gene [29]. Interestingly, overall expression of *SNN* was also increased in the red Poodles, which may indicate that *SNNL1* is capable of expression. Notably, multiple recent FGF4 retrocopy insertions have also been reported in dogs, several of which are expressed and involved in skeletal dysplasias [30-32]. In red and apricot Poodles, the insertion of *SNNL1* immediately upstream of *GPR22* is likely affecting regulation of the *GPR22* gene, resulting in its atypical expression in skin tissue. Structural variants, which include retrocopy insertions, have been identified as a major source of gene expression differences which often affect multiple nearby genes [33]. *SNNL1* is inserted within an intron of *COG5*, and while no chimeric reads were observed between the genes, a small increase in expression was observed for COG5 in the red Poodle skin which may also be a consequence of the retrocopy.

While several genes involved in pheomelanin intensity have been identified across dog breeds, the single-breed GWAS presented in this study only identified the CFA18 locus as significant within the Poodle breed. One of the top SNVs was a missense variant in *SLC26A4* which has previously been associated with pheomelanin intensity across dog breeds, where it was hypothesized to be the causative variant [7]. The researchers found that the CFA18 locus explained a relatively small percent of the total variance in pheomelanin across dog breeds (adjusted R2=0.047), whereas loci on CFA2 and CFA20 explained over 50% of the variance across breeds. However, when looking within a single breed, the Poodle, the CFA18 locus, identified herein as *SNNL1*, explained the majority of the variance between white, cream, apricot and red Poodles (adjusted R2=0.840). Notably, 70 out of 73 apricot or red poodles had at least 1 copy of *SNNL1*, while none of the 125 white Poodles tested had any copies of *SNNL1*. Only 19% of the cream poodles had 1 copy of *SNNL1*, while the rest had 0 copies. Likely other genes involved in pheomelanin intensity, such as the MFSD12 dilution variant, explain the differences between white and cream coat colors [5]. Analysis of the *SLC26A4* missense variant and the *SNNL1* insertion in a larger whole genome sequencing dataset revealed that they are rare across breeds and were only observed in East Asian dog breeds and village dogs, such as Tibetan Mastiffs, Chow Chows, and, notable for their rich pheomelanin appearance, Qingchuan and Chongqing dogs. It's possible that the CFA18 locus explains a relatively small proportion of the variance in pheomelanin intensity across breeds due to this breed exclusivity. Whereas, within breeds, *SNNL1* may actually mask the effects of other genes involved in pheomelanin intensity. While the CNV upstream of *KITLG* was associated with pheomelanin intensity in Nova Scotia Duck Tolling Retrievers, the *KITLG* CNV was not significantly associated with pheomelanin intensity in Poodles, possibly due to the effects of *SNNL1* within the breed [34].

Pigment production is canonically regulated through the MC1R-tmAC-MITF pathway, which induces changes in expression of pigment genes such as *TYR*, *PMEL*, *MLANA*, *SLC24A5*, *TYRP1* and *DCT* [35,36]. The MC1R-tmAC-MITF pathway uses the second messenger molecule cyclic adenosine monophosphate (cAMP), and loss of function mutations in the MC1R gene lead to impairments in downstream cAMP signaling, resulting in impaired eumelanogenesis and the phenotype known as recessive yellow in dogs [1]. In addition to cAMPs role as a second messenger molecule in the MC1R-tmAC-MITF pathway, tyrosinase itself is also affected by cAMP; reduction in cAMP within the melanosome results in a higher melanosomal pH, leading to greater tyrosinase activity and increased melanogenesis [37]. The sex hormones estrogen and progesterone have been found to regulate melanin synthesis through the alteration of cAMP signaling, showing that external factors which affect cAMP concentrations can have a downstream effect on melanogenesis [38]. While the pheomelanin-based Poodles used in this study were homozygous for the recessive yellow mutation in *MC1R*, several pigment genes were still observed to be highly overexpressed in red Poodle skin, including *TYR*, *PMEL* and *MLANA*, indicating upregulation of the melanogenesis pathway in red Poodles. Histopathological analysis in Poodle skin found that melanocytes were present within the hair bulbs of all Poodles, however melanin granules were present in high amounts in the red Poodles, consistent with an upregulation in pigment production in red Poodles. Notably, however, the main drivers behind eumelanogenesis, *TYRP1* and *DCT* [39], were not overexpressed in the red Poodles, suggesting specific upregulation in pheomelanogenesis. While *SLC26A4* was not expressed in either red or white Poodle skin, differential expression was observed in several other genes within the red Poodle associated region on chr18:10-20mb. However, the degree of differential expression observed in *GPR22* was much higher than any other genes, and was more comparable to the differential expression observed in genes involved in the pigment production pathway. Most notably, a near identical increase in expression was observed for *GPR22* and TYR (FC 6.05 and 5.96), which were also the top two most differentially expressed genes between red and white Poodle skin.

GPR22 is an orphan G-coupled protein receptor with a highly restrictive expression pattern in the heart and brain [40]. While *GPR22* knockout mice are viable and grossly indistinguishable from wild-type mice, they may be more susceptible to functional cardiac decompensation following aortic banding [40]. Deregulation of *GPR22* within the zebrafish embryo lead to defects in left-right patterning and resulted in abnormal cilia structure and length, indicating a possible developmental role for *GPR22* [41]. *GPR22* has also been implicated in osteoarthritis in humans through GWAS [42,43]. While *GPR22* was absent from healthy cartilage, it was found expressed in damaged cartilage, and overexpression of *GPR22* was also shown to accelerate chondrocyte hypertrophy [44,45]. While the *GPR22* ligand is unknown, overexpression of *GPR22* in HEK-293 cells identified that the GPR22 protein signals through the G inhibitory pathway, resulting in inhibition of adenylyl cyclase and a reduction in cAMP [40]. The second messenger cAMP regulates numerous functions in melanocytes, and the abnormal expression of *GPR22* in the melanocytes of red Poodles may have effects on pigmentation through this G inhibitory pathway.

GWAS in dogs often succeed at identifying genomic regions, yet due to the extensive LD in dog breeds, it can be difficult to pinpoint causative mutations. In this study, several nearby missense variants within the pheomelanin associated region on chr18 could not be ruled out by segregation analysis alone. While the expression analysis in the skin was able to rule out the *SLC26A4* and CDHR3 missense variants, further analysis including quantitative phenotyping in other breeds with the *SNNL1* insertion such as Chow Chows and Tibetan Mastiffs may be required to rule out the other missense variants. Still, the differential expression analysis in Poodle skin highlighted the *SNNL1* insertion and its effects on the expression of *GPR22* as the likely causal mutation behind the red coat color phenotype in Poodles. Further analysis may also quantify the effects of *SNNL1* on eumelanin based coat patterns, as the upregulation of numerous genes involved in melanogenesis observed in the red Poodles might indicate an effect on eumelanin as well. The identification of another recent, functional retrocopy insertion further serves to highlight the often complex nature behind genomic associations, and also shows that novel retrotransposition events continue to contribute to genomic and phenotypic diversity in dogs.

DATA ACCESS

All sequencing data has been uploaded to the Sequence Read Archive under the BioProject accession PRJNA830895. A description of these files is available in Table S4.

ACKNOWLEDGEMENTS

We thank the Poodle owners for providing samples of their dogs. We also thank Cindy Paul, Lisa Ritson and Barbara Hoopes for photographs of their Poodles.

CHAPTER 5 REFERENCES

1. Newton, J.; Wilkie, A.L.; He, L.; Jordan, S.A.; Metallinos, D.L.; Holmes, N.G.; Jackson, I.J.; Barsh, G.S. Melanocortin 1 receptor variation in the domestic dog. Mammalian Genome 2000, 11, 24-30.
2. Kerns, J.A.; Newton, J.; Berryere, T.G.; Rubin, E.M.; Cheng, J.-F.; Schmutz, S.M.; Barsh, G.S. Characterization of the dog Agouti gene and a nonagoutimutation in German Shepherd Dogs. Mammalian Genome 2004, 15, 798-808.
3. Berryere, T.G.; Kerns, J.A.; Barsh, G.S.; Schmutz, S.M. Association of an Agouti allele with fawn or sable coat color in domestic dogs. Mammalian Genome 2005, 16, 262-272.
4. Sponenberg, D.; Rothschild, M.F. Genetics of coat colour and hair texture. The genetics of the dog 2001, 61-85.
5. Hédan, B.; Cadieu, E.; Botherel, N.; Dufaure de Citres, C.; Letko, A.; Rimbault, M.; Drögemüller, C.; Jagannathan, V.; Derrien, T.; Schmutz, S. Identification of a missense variant in MFSD12 involved in dilution of Phaeomelanin leading to white or cream coat color in dogs. Genes 2019, 10, 386.
6. Weich, K.; Affolter, V.; York, D.; Rebhun, R.; Grahn, R.; Kallenberg, A.; Bannasch, D. Pigment intensity in dogs is associated with a copy number variant upstream of KITLG. Genes 2020, 11, 75.
7. Slavney, A.J.; Kawakami, T.; Jensen, M.K.; Nelson, T.C.; Sams, A.J.; Boyko, A.R. Five genetic variants explain over 70% of hair coat pheomelanin intensity variation in purebred and mixed breed domestic dogs. PloS one 2021, 16, e0250579.
8. Purcell, S.; Neale, B.; Todd-Brown, K.; Thomas, L.; Ferreira, M.A.; Bender, D.; Maller, J.; Sklar, P.; De Bakker, P.I.; Daly, M.J. PLINK: a tool set for whole-genome association and population-based linkage analyses. The American journal of human genetics 2007, 81, 559-575.
9. Zhou, X.; Stephens, M. Genome-wide efficient mixed-model analysis for association studies. Nature genetics 2012, 44, 821-824.
10. Wang, C.; Wallerman, O.; Arendt, M.-L.; Sundström, E.; Karlsson, Å.; Nordin, J.; Mäkeläinen, S.; Pielberg, G.R.; Hanson, J.; Ohlsson, Å. A novel canine reference genome resolves genomic architecture and uncovers transcript complexity. Communications biology 2021, 4, 1-11.

11. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv:1303.3997 2013.
12. Danecek, P.; Bonfield, J.K.; Liddle, J.; Marshall, J.; Ohan, V.; Pollard, M.O.; Whitwham, A.; Keane, T.; McCarthy, S.A.; Davies, R.M. Twelve years of SAMtools and BCFtools. Gigascience 2021, 10, giab008.
13. McLaren, W.; Gil, L.; Hunt, S.E.; Riat, H.S.; Ritchie, G.R.; Thormann, A.; Flicek, P.; Cunningham, F. The ensembl variant effect predictor. Genome biology 2016, 17, 1-14.
14. Adzhubei, I.A.; Schmidt, S.; Peshkin, L.; Ramensky, V.E.; Gerasimova, A.; Bork, P.; Kondrashov, A.S.; Sunyaev, S.R. A method and server for predicting damaging missense mutations. Nature methods 2010, 7, 248-249.
15. Ng, P.C.; Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. Nucleic acids research 2003, 31, 3812-3814.
16. Robinson, J.T.; Thorvaldsdóttir, H.; Winckler, W.; Guttman, M.; Lander, E.S.; Getz, G.; Mesirov, J.P. Integrative genomics viewer. Nature biotechnology 2011, 29, 24-26.
17. Untergasser, A.; Cutcutache, I.; Koressaar, T.; Ye, J.; Faircloth, B.C.; Remm, M.; Rozen, S.G. Primer3—new capabilities and interfaces. Nucleic acids research 2012, 40, e115-e115.
18. Li, H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 2018, 34, 3094-3100.
19. Meyer, E.; Aglyamova, G.; Matz, M. Profiling gene expression responses of coral larvae (Acropora millepora) to elevated temperature and settlement inducers using a novel RNA-Seq procedure. Molecular ecology 2011, 20, 3599-3616.
20. Smith, T.; Heger, A.; Sudbery, I. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. Genome research 2017, 27, 491-499.
21. Bushnell, B. BBMap: a fast, accurate, splice-aware aligner; Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States): 2014.
22. Dobin, A.; Davis, C.A.; Schlesinger, F.; Drenkow, J.; Zaleski, C.; Jha, S.; Batut, P.; Chaisson, M.; Gingeras, T.R. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 2013, 29, 15-21.
23. Anders, S.; Pyl, P.T.; Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. bioinformatics 2015, 31, 166-169.
24. Law, C.W.; Chen, Y.; Shi, W.; Smyth, G.K. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. Genome biology 2014, 15, 1-17.
25. Plassais, J.; Kim, J.; Davis, B.W.; Karyadi, D.M.; Hogan, A.N.; Harris, A.C.; Decker, B.; Parker, H.G.; Ostrander, E.A. Whole genome sequencing of canids reveals genomic regions under selection and variants influencing morphology. Nature communications 2019, 10, 1-14.
26. Kubiak, M.R.; Makałowska, I. Protein-coding genes' retrocopies and their functions. Viruses 2017, 9, 80.
27. Buck-Koehntop, B.A.; Mascioni, A.; Buffy, J.J.; Veglia, G. Structure, dynamics, and membrane topology of stannin: a mediator of neuronal cell apoptosis induced by trimethyltin chloride. Journal of Molecular Biology 2005, 354, 652-665.

28. Billingsley, M.; Yun, J.; Reese, B.; Davidson, C.; Buck-Koehntop, B.; Veglia, G. Functional and structural properties of stannin: roles in cellular growth, selective toxicity, and mitochondrial responses to injury. Journal of cellular biochemistry 2006, 98, 243-250.

29. Rosikiewicz, W.; Kabza, M.; Kosiński, J.G.; Ciomborowska-Basheer, J.; Kubiak, M.R.; Makałowska, I. RetrogeneDB–a database of plant and animal retrocopies. Database 2017, 2017.

30. Parker, H.G.; VonHoldt, B.M.; Quignon, P.; Margulies, E.H.; Shao, S.; Mosher, D.S.; Spady, T.C.; Elkahloun, A.; Cargill, M.; Jones, P.G. An expressed fgf4 retrogene is associated with breed-defining chondrodysplasia in domestic dogs. Science 2009, 325, 995-998.

31. Brown, E.A.; Dickinson, P.J.; Mansour, T.; Sturges, B.K.; Aguilar, M.; Young, A.E.; Korff, C.; Lind, J.; Ettinger, C.L.; Varon, S. FGF4 retrogene on CFA12 is responsible for chondrodystrophy and intervertebral disc disease in dogs. Proceedings of the National Academy of Sciences 2017, 114, 11476-11481.

32. Batcher, K.; Dickinson, P.; Maciejczyk, K.; Brzeski, K.; Rasouliha, S.H.; Letko, A.; Drögemüller, C.; Leeb, T.; Bannasch, D. Multiple FGF4 retrocopies recently derived within canids. Genes 2020, 11, 839.

33. Scott, A.J.; Chiang, C.; Hall, I.M. Structural variants are a major source of gene expression differences in humans and often affect multiple nearby genes. Genome research 2021, 31, 2249-2257.

34. Bannasch, D.L.; Affolter, V.K.; York, D.; Rebhun, R.B.; Grahn, R.A.; Weich, K.M.; Kallenberg, A. Correction: Weich, K., et al. Pigment Intensity in Dogs Is Associated with a Copy Number Variant Upstream of KITLG. Genes 2020, 11, 75. Genes 2021, 12, 357.

35. Bang, J.; Zippin, J.H. Cyclic adenosine monophosphate (cAMP) signaling in melanocyte pigmentation and melanomagenesis. Pigment Cell & Melanoma Research 2021, 34, 28-43.

36. Kawakami, A.; Fisher, D.E. The master role of microphthalmia-associated transcription factor in melanocyte and melanoma biology. Laboratory investigation 2017, 97, 649-656.

37. Zhou, D.; Ota, K.; Nardin, C.; Feldman, M.; Widman, A.; Wind, O.; Simon, A.; Reilly, M.; Levin, L.R.; Buck, J. Mammalian pigmentation is regulated by a distinct cAMP-dependent mechanism that controls melanosome pH. Science signaling 2018, 11, eaau7987.

38. Natale, C.A.; Duperret, E.K.; Zhang, J.; Sadeghi, R.; Dahal, A.; O'Brien, K.T.; Cookson, R.; Winkler, J.D.; Ridky, T.W. Sex steroids regulate skin pigmentation through nonclassical membrane-bound receptors. Elife 2016, 5, e15104.

39. Slominski, A.; Tobin, D.J.; Shibahara, S.; Wortsman, J. Melanin pigmentation in mammalian skin and its hormonal regulation. Physiological reviews 2004, 84, 1155-1228.

40. Adams, J.W.; Wang, J.; Davis, J.R.; Liaw, C.; Gaidarov, I.; Gatlin, J.; Dalton, N.D.; Gu, Y.; Ross Jr, J.; Behan, D. Myocardial expression, signaling, and function of *GPR22*: a protective role for an orphan G protein-coupled receptor. American Journal of Physiology-Heart and Circulatory Physiology 2008, 295, H509-H521.

41. Verleyen, D.; Luyten, F.P.; Tylzanowski, P. Orphan G-protein coupled receptor 22 (*GPR22*) regulates cilia length and structure in the zebrafish Kupffer's vesicle. PLoS One 2014, 9, e110484.

42. Kerkhof, H.J.; Lories, R.J.; Meulenbelt, I.; Jonsdottir, I.; Valdes, A.M.; Arp, P.; Ingvarsson, T.; Jhamai, M.; Jonsson, H.; Stolk, L. A genome-wide association study identifies an osteoarthritis susceptibility locus on chromosome 7q22. Arthritis & Rheumatism: Official Journal of the American College of Rheumatology 2010, 62, 499-510.
43. Evangelou, E.; Valdes, A.M.; Kerkhof, H.J.; Styrkarsdottir, U.; Zhu, Y.; Meulenbelt, I.; Lories, R.J.; Karassa, F.B.; Tylzanowski, P.; Bos, S.D. Meta-analysis of genome-wide association studies confirms a susceptibility locus for knee osteoarthritis on chromosome 7q22. Annals of the rheumatic diseases 2011, 70, 349-355.
44. Guns, L.; Monteagudo, S.; Calebiro, D.; Lohse, M.; Cailotto, F.; Lories, R. Increased *GPR22* activation triggers osteoarthritis. Annals of the Rheumatic Diseases 2018, 77, A5.
45. Guns, L.-A.; Calebiro, D.; Lohse, M.; Lories, R.; Cailotto, F. A cholecystokinin receptor antagonist inhibits chondrocyte hypertrophy and protein kinase a activity decrease induced by activation of *GPR22*. Osteoarthritis and Cartilage 2016, 24, S178-S179.

CHAPTER 6: Conclusion

The domestic dog, represented by over 350 distinct breeds (www.akc.org), is the most phenotypically diverse domesticated species, with two orders of magnitude differences in size between the smallest and largest breeds and an equally impressive degree of variation in behavioral and other physiological attributes [1-3]. Breed dogs also suffer from a predisposition to numerous heritable disorders [4-7], making them a useful translational model for human health. Many of the large effect size genetic variants underlying breed differences in morphological traits have been identified [8-12], although smaller effect size variants affecting morphology within breeds have yet to be discovered [13]. Dogs are noteworthy for having two distinct, functional *FGF4* retroCNVs that both have large effects on skeletal morphology [14,15]. In addition to the *FGF4* retroCNVs, other transposable elements including LINE-1 and SINE insertions make up a substantial portion (around 10%) of the phenotype associated variants identified in dogs (www.omia.org). Transposable elements have contributed substantially to the evolution of the canine genome, and they continue to play an active role to this day [16-19]. This dissertation research, focused on examining the landscape of retroCNV in dogs, has broader implications into the biology of LINE-1 activity in canids.

The disproportionate dwarfism phenotype in dogs called 'chondrodysplasia' was associated with an *FGF4* retroCNV on CFA18 [15]. The CFA18 *FGF4* retroCNV is common to many breeds, including some breeds which have also been characterized as chondrodystrophic, such as the Dachshund, leading many to incorrectly believe that this was the chondrodystrophy gene. However, many of the breeds that are positive for the CFA18 *FGF4* retroCNV, including several terrier breeds, are not canonically considered chondrodystrophic, nor do they appear to be at increased risk for IVDD [20,21]. The second *FGF4* retroCNV on CFA12, however, is present within all the canonically described chondrodystrophic breeds, such as the French bulldog and Beagle, two breeds which lack the CFA18 *FGF4* retroCNV [14,20,22]. However, to best advise breeders, it is necessary to determine the genetic contribution to disc disease of the two *FGF4* retroCNVs. Through the analysis of the *FGF4* retroCNVs in hundreds of IVDD surgical cases in Chapter 2, it was shown that the CFA12 *FGF4* retroCNV alone was associated with increased risk of disc calcification and herniation leading to surgery. Additionally, the analysis of all other canine *FGF4* retroCNV in Chapter 3 did not reveal additional common *FGF4* retroCNV as contributing to IVDD in dogs. These findings indicate that breeders can substantially decrease the incidence of IVDD by reducing the prevalence of the CFA12 *FGF4* retroCNV alone, keeping the CFA18 *FGF4* retroCNV for the disproportionate dwarfism phenotype that is preferrable in many popular breeds.

Chapter 4 contains a broader analysis of retroCNV in dogs; prior to this research, the landscape of retroCNV in dogs outside of the *FGF4* retrocopies was incomplete. The identification of a large number of retroCNVs across canids, most of which are full length and many functional, indicates that retroCNVs are a potential source for the phenotypic variability observed in dogs. Human genomes contained 13-fold fewer retroCNV on average compared to the canids, showing how relatively frequent retroCNV insertions are in the canid genome. An additional analysis in equid genomes for comparison is described in the Appendix, again finding fewer retroCNV than in the canids, and uncovering an ancient retroCNV of the *LCORL* gene that appears important to equine evolution. While the canine *FGF4* retroCNVs were discovered due to their strong phenotypic association with easily observable morphological traits, it is unclear what phenotypes could be associated with the other canine or equine retroCNVs. The availability of an extensive list of retroCNVs should, however, be a useful resource to the research community performing association analyses. As an example, Chapter 5 characterizes an *SNN* retroCNV associated with red coat color in Poodles following a GWAS. While a nearby gene, *SLC26A4*, had initially been identified by other researchers as a top candidate for pheomelanin intensity [23], it was shown through RNAseq analysis of red and white Poodle skin that the *SNN* retroCNV was the likely causative mutation due to its effects on the expression of *GPR22,* a gene nearby to its insertion site. Although the *SNN* retroCNV was only identified through the analysis of WGS from red Poodles after the analysis described in Chapter 4 was performed, which contained no red Poodles, it shows that the availability of a comprehensive list of canine retroCNVs is a useful resource for following up on genomic associations.

Interestingly, while dogs have less than twice as many full length LINE-1s as humans do [24], they have approximately 13-fold as many retroCNVs. In previous studies, dogs also had 8-fold and 17-fold as many dimorphic LINE-1s and SINEs compared to humans [17,18]. Horses also had an increased prevalence of retroCNV compared to humans, which would not be predicted based on the fewer number of full length LINE-1s in horses [24]. This suggests that the total number of full-length LINE-1 in a species may not necessarily correlate with overall LINE-1 activity, as some of the canine LINE-1 appear highly active. LINE-1 is both epigenetically and post-transcriptionally suppressed through various mechanisms, so across-species variance in LINE-1 activity may reflect changes to or defects in LINE-1 suppression [25,26]. One study has suggested that the recent accumulation of SINE insertions in dogs may be a consequence of a weak piRNA response, a mechanism used to suppress transposable elements [27]. It is also unclear to what extent species' LINE-1 activity in the germline could correlate with activity in somatic cells. For example, there are tissue-type specific mechanisms for LINE-1 suppression, particularly in the male germline, that could vary between species [28-31].

While the focus of this dissertation research was on retroCNVs acquired in the germline, the large number identified in canids indicate that somatic retrotransposition events are also likely to be a common occurrence in dogs. In humans, somatically acquired LINEs and SINEs as well as gene retrocopies have all been implicated in various disorders including cancers [32-35]. Meanwhile, breed dogs have a high predisposition to many disorders, and are at substantially higher risk for specific cancers than humans [4,36]. The germline retroCNVs identified in Chapter 4 may play a role in this increased susceptibility to disease in modern dog breeds, which was proven to be the case for the *FGF4* retroCNV causing chondrodystrophy and increased susceptibility to IVDD. However, somatic retrotransposon activity, including somatically acquired retroCNV, could also be a concern in dogs. Although the technology is still nascent, single cell whole genome sequencing has recently opened the door for the discovery of somatic retrotransposition events, which are surprisingly prevalent in primates [37-40]. In dogs, which have much higher LINE-1 activity than humans, the somatic accumulation of retrotransposition events could contribute to cancer susceptibility. An interesting observation in cancer biology, known as Peto's paradox, is that if every cell has a chance of becoming cancerous, then larger, longer-lived species should be at increased risk of developing cancers compared to smaller, shorter-lived species, and yet, no correlation exists between body size and cancer risk across species [41]. One possible explanation for Peto's paradox is that longer-lived species have a lower somatic mutation rate, resulting in fewer tumors [42]. Recently, somatic mutation rates across 16 mammalian species were indeed confirmed to be negatively correlated with life span [43]. Similarly, the rate of somatic retrotransposon activity within a species might correlate with both life span and cancer prevalence. While smaller dog breeds do live longer and are at lower risk of certain cancers than larger breeds, purebred dogs as a whole are at higher risk of cancer than humans [44,45]. It may be necessary to quantify canine LINE-1 activity in somatic tissues, particularly in those breeds and tissue types that have a high prevalence of specific cancers, to fully understand the contribution of LINE-1 mediated retrotransposition to cancer.

Finally, it is interesting to speculate why LINE-1 is highly active in the domestic dog. If it were the case that LINE-1 mediated retrotranspositions were a contributing factor to the increased prevalence of disease and cancer in dogs, one might expect individuals to have compensatory mechanisms to suppress LINE-1 activity for higher fitness. Transposable elements are recognized as mutagenic, and are involved in an ongoing “arms race” between them and their host genomes that seek to suppress their activity [46,47]. Evolutionary pressure should favor the suppression of LINE-1 activity, resulting in less retrotransposition, and yet LINE-1 remains highly active within the canine genome. In response to a similar question regarding evolutionary pressure acting to reduce mutation rates, the geneticist A.H. Sturtevant once stated: "It would evidently be fatal for a species, in the long run, if its mutation rate fell to zero, for adjustment to changing conditions would then not long remain possible" [48]. Similarly, a reduction in LINE-1 activity could be predicted to result in a reduction in a species ability to adjust to changing conditions. While somatic changes to the genome such as point mutations and retroposition events are harmful to the individual, leading to cancers and aging, some level of LINE-1 activity may be beneficial for the species as a whole, allowing for greater adaptability to environmental changes and the ability to take better advantage of new niches. The domestic dog is arguably one of the most phenotypically diverse mammalian species and was able to rapidly adapt and thrive during and following domestication. One of the models for domestication of canids proposes that, unlike with other species which had domestication forced upon them, such as plants and ungulates, dogs underwent self-domestication, where they had to adapt to humans. Part of the reason dogs have been so successful could be due to an adaptive advantage provided by increased LINE-1 activity.

CHAPTER 6 REFERENCES

1. Drake, A.G.; Klingenberg, C.P. Large-scale diversification of skull shape in domestic dogs: disparity and modularity. *The American Naturalist* **2010**, *175*, 289-301.
2. Wayne, R.K.; Vonholdt, B.M. Evolutionary genomics of dog domestication. *Mammalian Genome* **2012**, *23*, 3-18.
3. Wilcox, B.; Walkowicz, C. *Atlas of dog breeds of the world. New rev*; 1989.
4. Gough, A.; Thomas, A.; O'Neill, D. *Breed predispositions to disease in dogs and cats*; John Wiley & Sons: John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA, 2018.
5. Summers, J.F.; Diesel, G.; Asher, L.; McGreevy, P.D.; Collins, L.M. Inherited defects in pedigree dogs. Part 2: Disorders that are not related to breed standards. *The Veterinary Journal* **2010**, *183*, 39-45.
6. Asher, L.; Diesel, G.; Summers, J.F.; McGreevy, P.D.; Collins, L.M. Inherited defects in pedigree dogs. Part 1: disorders related to breed standards. *The Veterinary Journal* **2009**, *182*, 402-411.
7. Bannasch, D.; Famula, T.; Donner, J.; Anderson, H.; Honkanen, L.; Batcher, K.; Safra, N.; Thomasy, S.; Rebhun, R. The effect of inbreeding, body size and morphology on health in dog breeds. *Canine Medicine and Genetics* **2021**, *8*, 12, doi:10.1186/s40575-021-00111-4.
8. Plassais, J.; Kim, J.; Davis, B.W.; Karyadi, D.M.; Hogan, A.N.; Harris, A.C.; Decker, B.; Parker, H.G.; Ostrander, E.A. Whole genome sequencing of canids reveals genomic regions under selection and variants influencing morphology. *Nature communications* **2019**, *10*, 1-14.
9. Boyko, A.R.; Quignon, P.; Li, L.; Schoenebeck, J.J.; Degenhardt, J.D.; Lohmueller, K.E.; Zhao, K.; Brisbin, A.; Parker, H.G.; Cargill, M. A simple genetic architecture underlies morphological variation in dogs. *PLoS biology* **2010**, *8*, e1000451.
10. Mansour, T.A.; Lucot, K.; Konopelski, S.E.; Dickinson, P.J.; Sturges, B.K.; Vernau, K.L.; Choi, S.; Stern, J.A.; Thomasy, S.M.; Döring, S. Whole genome variant association across 100 dogs identifies a frame shift mutation in DISHEVELLED 2 which contributes to Robinow-like syndrome in Bulldogs and related screw tail dog breeds. *PLoS genetics* **2018**, *14*, e1007850.
11. Marchant, T.W.; Johnson, E.J.; McTeir, L.; Johnson, C.I.; Gow, A.; Liuti, T.; Kuehn, D.; Svenson, K.; Bermingham, M.L.; Drögemüller, M. Canine brachycephaly is associated with a retrotransposon-mediated missplicing of SMOC2. *Current Biology* **2017**, *27*, 1573-1584. e1576.
12. Schoenebeck, J.J.; Hutchinson, S.A.; Byers, A.; Beale, H.C.; Carrington, B.; Faden, D.L.; Rimbault, M.; Decker, B.; Kidd, J.M.; Sood, R. Variation of BMP3 contributes to dog breed skull diversity. *PLoS genetics* **2012**, *8*, e1002849.
13. Bannasch, D.L.; Baes, C.F.; Leeb, T. Genetic variants affecting skeletal morphology in domestic dogs. *Trends in genetics* **2020**, *36*, 598-609.
14. Brown, E.A.; Dickinson, P.J.; Mansour, T.; Sturges, B.K.; Aguilar, M.; Young, A.E.; Korff, C.; Lind, J.; Ettinger, C.L.; Varon, S. FGF4 retrogene on CFA12 is responsible for chondrodystrophy and intervertebral disc disease in dogs. *Proceedings of the National Academy of Sciences* **2017**, *114*, 11476-11481.

15. Parker, H.G.; VonHoldt, B.M.; Quignon, P.; Margulies, E.H.; Shao, S.; Mosher, D.S.; Spady, T.C.; Elkahloun, A.; Cargill, M.; Jones, P.G. An expressed fgf4 retrogene is associated with breed-defining chondrodysplasia in domestic dogs. *Science* **2009**, *325*, 995-998.
16. Biémont, C. A brief history of the status of transposable elements: from junk DNA to major players in evolution. *Genetics* **2010**, *186*, 1085-1093.
17. Halo, J.V.; Pendleton, A.L.; Shen, F.; Doucet, A.J.; Derrien, T.; Hitte, C.; Kirby, L.E.; Myers, B.; Sliwerska, E.; Emery, S. Long-read assembly of a Great Dane genome highlights the contribution of GC-rich sequence and mobile elements to canine genomes. *Proceedings of the National Academy of Sciences* **2021**, *118*.
18. Wang, W.; Kirkness, E.F. Short interspersed elements (SINEs) are a major source of canine genomic diversity. *Genome Research* **2005**, *15*, 1798-1808.
19. Lindblad-Toh, K.; Wade, C.M.; Mikkelsen, T.S.; Karlsson, E.K.; Jaffe, D.B.; Kamal, M.; Clamp, M.; Chang, J.L.; Kulbokas, E.J.; Zody, M.C. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **2005**, *438*, 803-819.
20. Hansen, H.-J. A pathologic-anatomical study on disc degeneration in dog: With special reference to the so-called enchondrosis intervertebralis. *Acta Orthopaedica Scandinavica* **1952**, *23*, 1-130.
21. Bellumori, T.P.; Famula, T.R.; Bannasch, D.L.; Belanger, J.M.; Oberbauer, A.M. Prevalence of inherited disorders among mixed-breed and purebred dogs: 27,254 cases (1995–2010). *Journal of the American Veterinary Medical Association* **2013**, *242*, 1549-1555.
22. Braund, K.; Ghosh, P.; Taylor, T.; Larsen, L. Morphological studies of the canine intervertebral disc. The assignment of the beagle to the achondroplastic classification. *Research in veterinary science* **1975**, *19*, 167-172.
23. Slavney, A.J.; Kawakami, T.; Jensen, M.K.; Nelson, T.C.; Sams, A.J.; Boyko, A.R. Five genetic variants explain over 70% of hair coat pheomelanin intensity variation in purebred and mixed breed domestic dogs. *PloS one* **2021**, *16*, e0250579.
24. Penzkofer, T.; Jäger, M.; Figlerowicz, M.; Badge, R.; Mundlos, S.; Robinson, P.N.; Zemojtel, T. L1Base 2: more retrotransposition-active LINE-1s, more mammalian genomes. *Nucleic acids research* **2016**, gkw925.
25. Pizarro, J.G.; Cristofari, G. Post-transcriptional control of LINE-1 retrotransposition by cellular host factors in somatic cells. *Frontiers in cell and developmental biology* **2016**, *4*, 14.
26. Slotkin, R.K.; Martienssen, R. Transposable elements and the epigenetic regulation of the genome. *Nature reviews genetics* **2007**, *8*, 272-285.
27. Vandewege, M.W.; Platt, R.N.; Ray, D.A.; Hoffmann, F.G. Transposable element targeting by piRNAs in Laurasiatherians with distinct transposable element histories. *Genome biology and evolution* **2016**, *8*, 1327-1337.
28. Kuramochi-Miyagawa, S.; Watanabe, T.; Gotoh, K.; Totoki, Y.; Toyoda, A.; Ikawa, M.; Asada, N.; Kojima, K.; Yamaguchi, Y.; Ijiri, T.W. DNA methylation of retrotransposon genes is regulated by Piwi family members MILI and MIWI2 in murine fetal testes. *Genes & development* **2008**, *22*, 908-917.

29. Goodier, J.L.; Pereira, G.C.; Cheung, L.E.; Rose, R.J.; Kazazian Jr, H.H. The broad-spectrum antiviral protein ZAP restricts human retrotransposition. *PLoS genetics* **2015**, *11*, e1005252.
30. Tan, K.; Kim, M.E.; Song, H.-W.; Skarbrevik, D.; Babajanian, E.; Bedrosian, T.A.; Gage, F.H.; Wilkinson, M.F. The Rhox gene cluster suppresses germline LINE1 transposition. *Proceedings of the National Academy of Sciences* **2021**, *118*, e2024785118.
31. Di Giacomo, M.; Comazzetto, S.; Saini, H.; De Fazio, S.; Carrieri, C.; Morgan, M.; Vasiliauskaite, L.; Benes, V.; Enright, A.J.; O'Carroll, D. Multiple epigenetic mechanisms and the piRNA pathway enforce LINE1 silencing during adult spermatogenesis. *Molecular cell* **2013**, *50*, 601-608.
32. Bim, L.V.; Navarro, F.C.; Valente, F.O.; Lima-Junior, J.V.; Delcelo, R.; Dias-da-Silva, M.R.; Maciel, R.; Galante, P.A.; Cerutti, J.M. Retroposed copies of RET gene: a somatically acquired event in medullary thyroid carcinoma. *BMC medical genomics* **2019**, *12*, 1-13.
33. Rodriguez-Martin, B.; Alvarez, E.G.; Baez-Ortega, A.; Zamora, J.; Supek, F.; Demeulemeester, J.; Santamarina, M.; Ju, Y.S.; Temes, J.; Garcia-Souto, D. Pan-cancer analysis of whole genomes identifies driver rearrangements promoted by LINE-1 retrotransposition. *Nature genetics* **2020**, *52*, 306-319.
34. Xiao-Jie, L.; Hui-Ying, X.; Qi, X.; Jiang, X.; Shi-Jie, M. LINE-1 in cancer: multifaceted functions and potential clinical implications. *Genetics in Medicine* **2016**, *18*, 431-439.
35. Ciomborowska-Basheer, J.; Staszak, K.; Kubiak, M.R.; Makałowska, I. Not So Dead Genes—Retrocopies as Regulators of Their Disease-Related Progenitors and Hosts. *Cells* **2021**, *10*, 912.
36. Schiffman, J.D.; Breen, M. Comparative oncology: what dogs and other species can teach us about humans with cancer. *Philosophical Transactions of the Royal Society B: Biological Sciences* **2015**, *370*, 20140231.
37. Billon, V.; Sanchez-Luque, F.J.; Rasmussen, J.; Bodea, G.O.; Gerhardt, D.J.; Gerdes, P.; Cheetham, S.W.; Schauer, S.N.; Ajjikuttira, P.; Meyer, T.J. Somatic retrotransposition in the developing rhesus macaque brain. *Genome research* **2022**, *32*, 1298-1314.
38. Erwin, J.A.; Paquola, A.; Singer, T.; Gallina, I.; Novotny, M.; Quayle, C.; Bedrosian, T.A.; Alves, F.I.; Butcher, C.R.; Herdy, J.R. L1-associated genomic regions are deleted in somatic cells of the healthy human brain. *Nature neuroscience* **2016**, *19*, 1583-1591.
39. Evrony, G.D.; Lee, E.; Mehta, B.K.; Benjamini, Y.; Johnson, R.M.; Cai, X.; Yang, L.; Haseley, P.; Lehmann, H.S.; Park, P.J. Cell lineage analysis in human brain using endogenous retroelements. *Neuron* **2015**, *85*, 49-59.
40. Nam, C.H.; Youk, J.; Kim, J.Y.; Lim, J.; Park, J.W.; Oh, S.A.; Lee, H.J.; Park, J.W.; Jeong, S.-Y.; Lee, D.-S. Extensive mosaicism by somatic L1 retrotransposition in normal human cells. *bioRxiv* **2022**.
41. Peto, R.; Roe, F.; Lee, P.; Levy, L.; Clack, J. Cancer and ageing in mice and men. *British journal of cancer* **1975**, *32*, 411-426.
42. Caulin, A.F.; Maley, C.C. Peto's Paradox: evolution's prescription for cancer prevention. *Trends in ecology & evolution* **2011**, *26*, 175-182.
43. Cagan, A.; Baez-Ortega, A.; Brzozowska, N.; Abascal, F.; Coorens, T.H.; Sanders, M.A.; Lawson, A.R.; Harvey, L.M.; Bhosle, S.; Jones, D. Somatic mutation rates scale with lifespan across mammals. *Nature* **2022**, *604*, 517-524.

44. Rowell, J.L.; McCarthy, D.O.; Alvarez, C.E. Dog models of naturally occurring cancer. *Trends in molecular medicine* **2011**, *17*, 380-388.

45. Michell, A. Longevit of British breeds of dog and its relationships with-sex, size, cardiovascular variables and disease. *Veterinary Record* **1999**, *145*, 625-629.

46. Aravin, A.A.; Hannon, G.J.; Brennecke, J. The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race. *science* **2007**, *318*, 761-764.

47. Saito, K.; Siomi, M.C. Small RNA-mediated quiescence of transposable elements in animals. *Developmental cell* **2010**, *19*, 687-697.

48. Sturtevant, A. Essays on evolution. I. On the effects of selection on mutation rate. *The Quarterly Review of Biology* **1937**, *12*, 464-467.

APPENDIX: A large number of segmentally duplicated *LCORL* retrocopies in equids

Kevin Batcher[1], Scarlett Varney[1], Vidhya Jagannathan[2], Tosso Leeb[2], and Danika Bannasch[1]*

[1] Department of Population Health and Reproduction, University of California, Davis, CA, USA

[2] Institute of Genetics, Vetsuisse Faculty, University of Bern, 3001 Bern, Switzerland

*Correspondence: dlbannasch@ucdavis.edu

ABSTRACT

Although gene retrocopies are often presumed to be nonfunctional pseudogenes, evidence of functional retrocopies has grown over time. Additionally, LINE-1 is still active and capable of inserting new retrocopies, resulting in retro-copy number variants (retroCNVs) between individuals. Here, retroCNV discovery was performed on a whole genome sequencing dataset of 78 horses and 8 wild equids. In total, 437 retrocopy insertions from 353 parent genes were identified. Each horse had 22.5 retroCNV on average. Additionally, 32 retrocopies present in the EquCab3.0 reference assembly were found to be absent from all wild equids. Only 5 retroCNV were shared between at least one horse and one wild equid, indicating that the majority of retroCNV inserted after the species diverged. Among the shared retroCNV, a large number of segmentally duplicated *LCORL* retrocopies were identified in all equids, with horses and zebras having between 17 and 35 copies. The *LCORL* retrocopy was estimated to have inserted 18 MYA (17-19 95% CI) and was absent from other members of the family Perissodactyl (Tapirs and Rhinoceroses). The *LCORL* retrocopies were highly expressed and accounted for the majority of overall *LCORL* expression in all tissue types except testes in both horses and donkeys. Evolutionary conservation of the *LCORL* retrocopy segmental duplication in

the Equidae family and the ancient timeline for *LCORL* retrotransposition indicate an essential role for this structural variant in a distinguishing feature of their biology. The dramatic increase in body size and reduction in digit number during equid evolution are predicted to have occurred after the retrotransposition of a gene known to affect body size across mammals.

INTRODUCTION

Long interspersed nuclear element 1 (LINE-1) is the only autonomous transposable element still active in mammalian genomes (*1*). LINE-1 encodes two functional proteins: an mRNA binding protein and a combined reverse transcriptase endonuclease. These LINE-1 proteins function to reverse transcribe and insert mRNA copies of LINE-1 into the genome in a process called target-site primed reverse transcription (*2, 3*). One hallmark of LINE-1 mediated retrotransposition is the duplication of short segments of genomic DNA at the insertion site, called target site duplications (TSD)(*4*). While LINE-1 proteins preferentially act on LINE-1 mRNA, LINE-1 proteins can also create new genomic copies of short interspersed nuclear element (SINE) as well as copies of cellular mRNA, which are referred to as retrocopies or processed pseudogenes (*5, 6*). Because retrocopies are derived from processed mRNA, they have a poly(A) tail and lack the introns and regulatory elements present at the parent gene, aspects which distinguish retrocopies from their parent genes (*7*). Most mammalian reference assemblies have thousands of retrocopies, many of which no longer code for functional proteins (*8*). Whereas these ancestral retrocopies tend to be fixed within species, LINE-1 is still active and capable of inserting new retrocopies (*9, 10*). These recently inserted retrocopies vary between individuals, resulting in what have been referred to as retrocopy number variants (retroCNVs) (*11, 12*).

While retroCNVs are not typically identified from whole genome sequencing (WGS) data by most common variant calling programs, techniques which take advantage of the differences between the retrocopy insertion and the parent gene of origin can be used to identify retroCNV, with the "gold standard" of retroCNV discovery requiring identification of the retroCNV parent gene and characterization of the insertion site. (*2, 13*). Estimates of gene retrotransposition rates have varied due to the difficulty in identifying retroCNVs and the varying number of active LINE-1 between species (*10*). However, an analysis of multiple high quality human genome assemblies indicate retroCNV are more common than previously believed (*14*). Studies using high coverage WGS data identified 1663 retroCNV parent genes in mice (*13*) and 503 in human populations (*15*), while 1911 retroCNV insertions from 1179 parent genes were recently identified in canids (*16*). A recent study in Thoroughbred horses found 62 retroCNV parent genes, although the insertion sites were not identified (*17*).

Retrocopies are often presumed to be nonfunctional and designated as pseudogenes, yet there is a growing body of evidence indicating that retrocopies are commonly expressed and functional (*13, 18, 19*). Expressed retrocopies have been implicated in various diseases including cancer and neurodegenerative disorders in humans (*20*). Even older retrocopies that no longer code for functional proteins can act as regulatory long non-coding RNAs (lncRNA) which alter the expression or translation of the parent gene (*20*). For example, elephants have around 20 transcriptionally active, segmentally duplicated *TP53* retrocopies which, despite being truncated and no longer coding for a fully functional TP53 proteins, are thought to play a role in DNA response to damage and cancer resistance (*21, 22*). Segmental or tandem duplications of genes or retrocopies of genes are a common evolutionary mechanism across species (*23*). While most retrocopies present within reference assemblies are ancestral and have accumulated nonsense mutations (*8*), retroCNVs, due to their recent insertion, typically still code for functional proteins (*16*). In dogs, multiple fully functional and expressed *FGF4* retroCNVs have been identified based on their association with skeletal dysplasias (*24, 25*). Recent retrocopy insertions may also have functional effects on the expression of nearby genes or form chimeric transcripts when inserted within the intron of other genes (*11, 12*). While the retroCNVs identified in natural populations of wild mice were found to be under strong negative selection, likely due to deleterious effects (*13*), significant population differentiation by breed was observed for many of the retroCNV in dogs, possibly as a consequence of artificial selection by breeders (*16*).

Because retroCNV have a propensity to be harmful, there is a need to characterize the repertoire of retroCNV in domesticated species. In this study, a recently developed method to identify retroCNV was used to characterize the landscape of retroCNV in 86 equids. While most retroCNV were not shared between the domestic horses and other equids, many *LCORL* retrocopies were identified in all equids mapping to the same locus, highlighting a segmental duplication containing an *LCORL* retrocopy which is absent from the reference assembly. Retrocopy specific variants were identified and used to evaluate *LCORL* expression in horse and donkey RNAseq datasets, and estimate the age of the retrotransposition event.

RESULTS

RetroCNV discovery in equids

RetroCNV discovery was performed using Illumina paired-end WGS data from 86 equids aligned to EquCab3.0 (Supplemental Table S1). The median coverage across all samples was 21.2x. An initial analysis using mRNA specific 30-mers identified 693 total putative retroCNV parent genes across all samples. RetroCNV insertion site discovery was then performed using discordant read analysis focused on these putative retroCNV parent genes. In total, 437 retroCNV insertions were resolved from 353 parent genes (Figure 1). Many of the retroCNV were inserted within the introns of other protein coding genes (146/437; 33.4%) (Supplemental Table S2). The TSD sequence was identified for 316 of the retroCNV insertions based on overlapping discordant reads. The TSD sequences ranged from 10 to 31 bp with a median length of 17. A genotyping matrix of the retroCNV in each sample is available in Supplemental Table S2. Primers were designed for eight retroCNVs, and a set of thoroughbreds were tested to validate the insertion sites. Sanger sequencing of the PCR products confirmed the retrocopy insertions and their TSD sequences (Supplemental Table S3).

The number of total non-reference retroCNV in each individual horse ranged from 8 to 40 with an average of 22.5 retroCNV (Table 1; 95% CI 21.3-23.7). Meanwhile, the 8 wild equids averaged 61.1 retroCNV (95% CI 55.0-67.2). Horses had on average less than 1 private retroCNV (average 0.60; 95% CI 0.40-0.81), while the wild equids, which only included one individual from each species, had on average 15.6 private retroCNV (95% CI 9.9 -21.4). Nearly half of the total non-reference retroCNV (218/437, 49.9%) were exclusive to the 8 wild equids. The horse retroCNV were also analyzed for unique variants that differentiate them from their parental gene sequence. Most (157/213, 73.7%) of the retroCNVs in horses did not contain any unique variants, and only one high impact variant was detected, indicating that most retroCNV still likely code for functional proteins. The list of retroCNV specific variants and their predicted effects is available in Supplemental Table S4.

All Horses (N=78)	219	22.5	38	191
Przewalski Horse (N=3)	50	29	9	23
Wild equids (N=8)	223	61.1	125	218

Table 1: Summary of non-reference retroCNV in equids. Average retroCNV is the average number of total retroCNV carried by an individual animal. Private retroCNV are unique to an individual, and exclusive retroCNV are unique to the population.

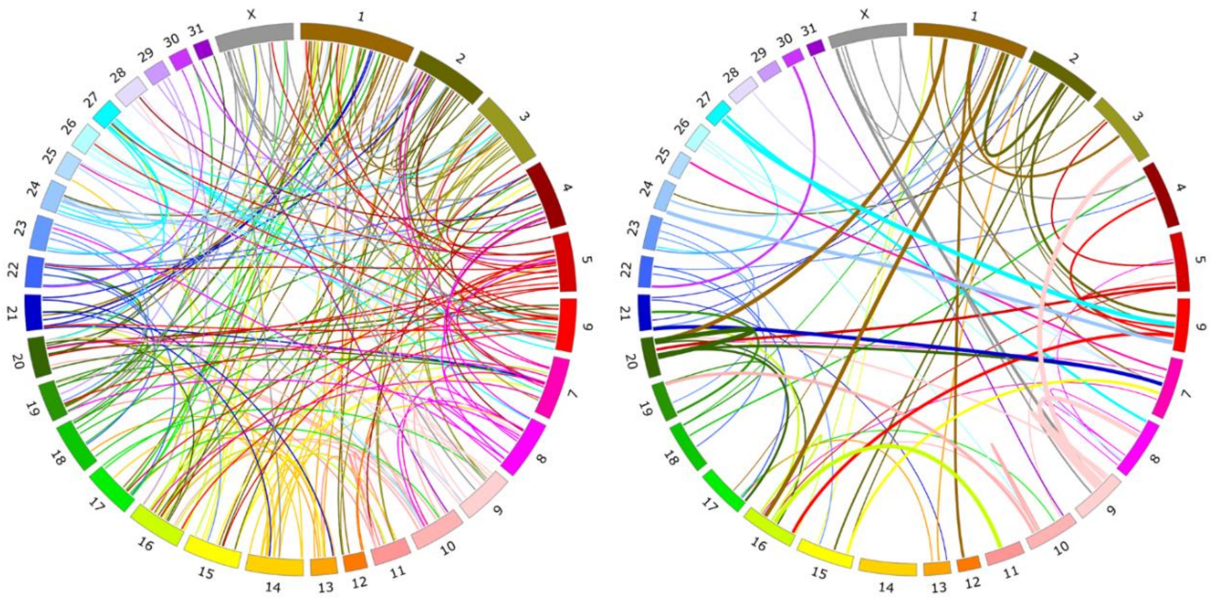


Figure 1: Circos plots of equine retroCNVs. Links are colored based on the chromosome of parent gene. Left: all retroCNV identified in 86 equids (N=437). Right: All the retroCNV in 22 Thoroughbred horses (N=81). Thickness of the band indicates how common the retroCNV was in thoroughbreds.

While discordant read analysis failed to identify an insertion site for 340 of the 693 putative retroCNV parent genes, exon-exon discordant reads were observed for only 51 of those 340 putative parent genes, indicating that retroCNVs with unresolved insertion sites were present for those parent genes (Supplemental Table S5). Four of these putative parent genes

(*ATP6V0C*, *EIF3CL*, *LOC100066257*, *TIRAP*) were observed to have discordant reads present only in the male samples, possibly indicating that the retroCNV was on the Y chromosome, which is absent from the EquCab3.0 reference assembly. To test this, an equine Y chromosome assembly was separately analyzed for these four retroCNV, three of which (*ATP6V0C*, *EIF3CL*, and *LOC100066257*) were identified.

The program sideRETRO was also used to identify retroCNV in the same dataset. A total of 385 retroCNV candidates were identified by sideRETRO under default settings, 196 of which were confirmed to be retrocopy insertions through visual analysis, and 176 of which overlapped with the initial analysis (Supplemental Table S6).

Retrocopies present in the EquCab3.0 assembly itself were also analyzed to determine if they were retroCNV. In this analysis, any retrocopy absent from an individual was considered a retroCNV. A total of 32 reference retroCNV were identified (Supplemental Table S7). All the 32 reference retroCNV were absent from all the wild equids, making them horse-specific retroCNV; additionally, 21 of them were present in every horse sample. The reference retroCNVs shared 98.3% sequence identity with their parent gene sequences on average, and also tended to be the full length of the parental protein coding sequence (97.4% on average).

RetroCNV parent gene analysis

Some of the retroCNV parent genes were identified as genes which commonly form retrocopies, including multiple ribosomal protein genes and *GAPDH*. The 353 retroCNV parent genes were highly enriched for biological processes of translation and peptide synthesis, and the molecular functions related to the ribosome and RNA binding (Table 2).

GO biological process	Fold Enrichment	P value	FDR
translation	5.79	3.49E-15	5.01E-11
peptide biosynthetic process	5.53	1.14E-14	8.17E-11
peptide metabolic process	4.59	6.79E-13	3.25E-09
cellular macromolecule biosynthetic process	3.68	2.21E-12	7.93E-09
GO molecular function	**Fold Enrichment**	**P value**	**FDR**
structural constituent of ribosome	6.63	1.44E-11	6.41E-08
RNA binding	2.96	1.12E-10	2.49E-07
structural molecule activity	3.11	1.41E-07	6.95E-05
mRNA binding	4.33	2.23E-07	9.01E-05

Table 2: GO analysis of equine retroCNV parent genes.

LCORL retrocopy in Equids

Only 5 retroCNV were shared between any wild equids and horses. Among them, an *LCORL* retrocopy was identified which was present in all 86 samples, indicating that it was common to all equids and was not present in the EquCab3.0 assembly. Analysis of the EquCab2.0 assembly indicated that an *LCORL* retrocopy was present on a chromosomal contig which was absent from the EquCab3.0 assembly. Discordant reads identified the insertion site as chr9:30194359-30194380 in EquCab3.0, within the intron of *MRPL15*. However, an abnormally large number of discordant reads were observed for the *LCORL* retrocopy; each equid had on average 750 discordant reads (95% CI 654-847) mapping between the *LCORL* parent gene and the insertion site on chromosome 9, which indicated there were many *LCORL* retrocopies at this location (Figure 2A; Supplemental Figure S1). A previous study had also indicated the presence of multiple *LCORL* retrocopies on chromosome 9 using fluorescence in situ hybridization (26). Visual analysis of the *LCORL* retrocopy insertion site revealed similar increased coverage nearby the insertion, indicating that the retrocopy was part of a larger segmental duplication (Supplemental Figure S1).

An analysis based on increased coverage per individual indicated that there were between 17 and 35 copies of the segmental duplication carrying the *LCORL* retrocopy in horses (Supplemental Table S8). The number of *LCORL* retrocopies appears to vary between individuals and species; members of the Subgenus *Asinus* had the least number of copies on average at 13.7 (95% CI 10.3-16.6), while domestic horses have an average of 25.1 (95% CI 24.4-25.9). Interestingly the Przewalski horses were estimated to have 33.2 copies of *LCORL* on average (95% CI 30.0-36.4). When the UCLA_HA_Equagga_1.0 zebra genome assembly was examined for evidence of the *LCORL* retrocopy, six *LCORL* retrocopy fragments were identified on unplaced-scaffolds, indicating that the region had not been properly assembled. The ASM1607732v2 donkey genome assembly, however, contains 10 full length *LCORL* retrocopies (Supplemental Table S9). Curiously, only one of the *LCORL* retrocopies was on donkey chr12, which is the equivalent of horse chr9; the other nine *LCORL* retrocopies were assembled on donkey chr3.

Because the *LCORL* gene locus has been associated with height in horses (*27, 28*) as well as other mammals (*29-33*), the *LCORL* retrocopies were further examined for evidence of function. The *LCORL* gene contains multiple different isoforms formed through alternative splicing. The retrocopy is derived from a transcript which lacks exon 3 and which includes the first of two possible terminal exons (Figure 2A). This specific transcript has been observed in mice as *LCORL* transcript variant x12, where a second start codon in exon 2 is utilized to produce a 512 amino acid product (NCBI: XM_030254341.2). The assembled retrocopy sequences present in the EquCab2.0 and ASM1607732v2 assemblies were all predicted to result in a truncated protein product between 43 and 74 amino acids in length when the first start codon in exon 1 is used (Supplemental Data 1). The predicted protein product of the retrocopies using the secondary start codon in exon 2 is only 8 amino acids in length (Supplemental Data 2). Based on visual inspection of the *LCORL* retrocopy, it had many SNV indicating that it was an older retrogene insertion (Figure 2B). Variant calling was performed across the *LCORL* parent gene locus to identify variants specific to the retrocopies. Numerous variants were detected that were unique to the *LCORL* retrocopy sequence and which differentiate it from the parent gene, including multiple frameshift mutations, a 33bp deletion in exon 1 corresponding to the loss of 11 alanine codons, and a nonsense mutation in exon 4 (Supplemental Table S10). The *LCORL* retrocopy specific variants also differed by species; while 15 variants appeared fixed in all *LCORL* retrocopies in all Equid species, 24 variants appeared unique to and fixed within only the horse *LCORL* retrocopies, and 6 variants appeared in all wild equid *LCORL* retrocopies which were absent from horses. The horse retrocopies also contain a 76bp deletion in exon 7 (chr3:107,550,405-107,550,481) and a SINE insertion in the 3’ UTR (chr3:107,553,402) which are not present in any of the wild equid retrocopies.

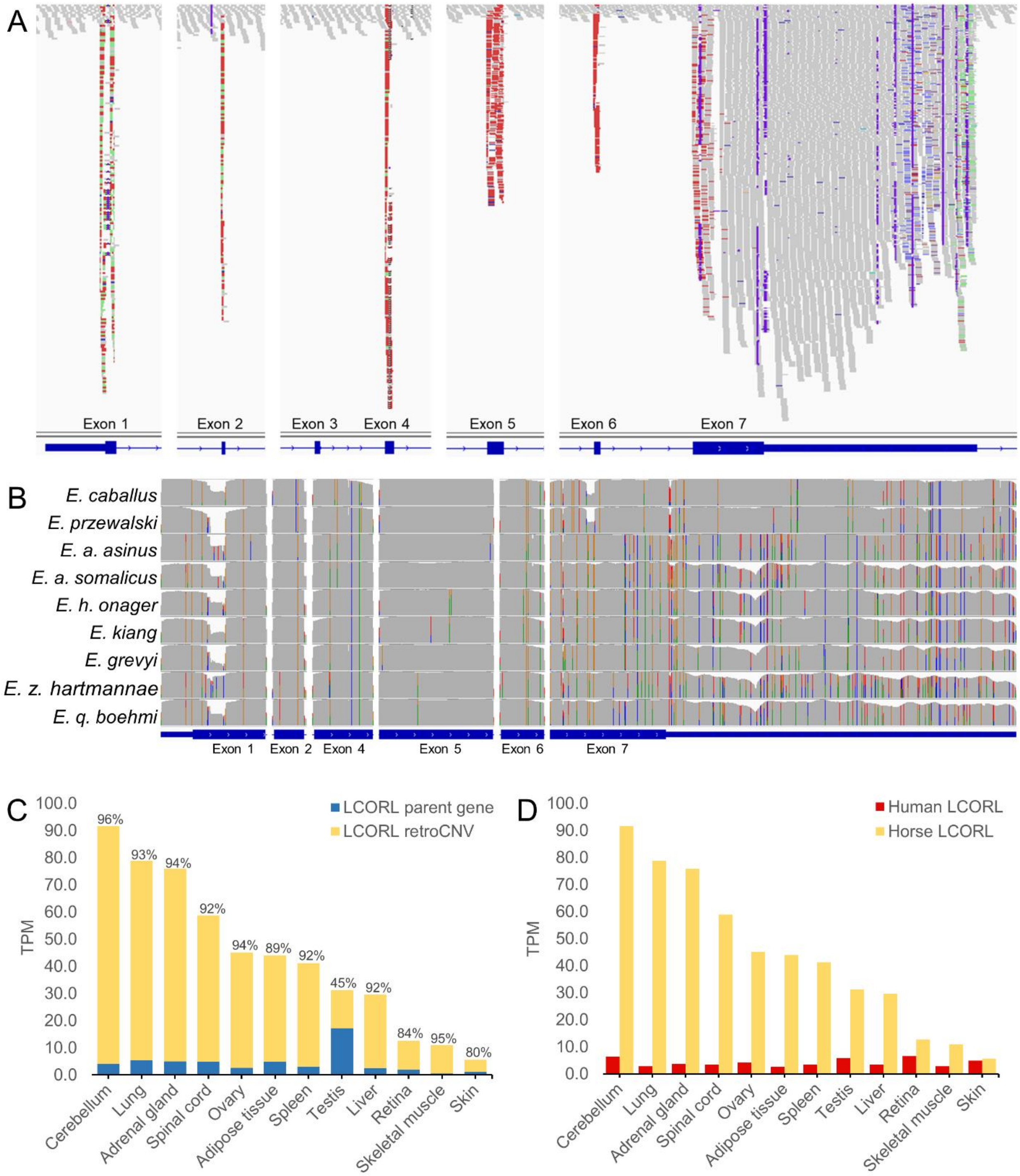


Figure 2: Large number of expressed *LCORL* retrocopies in equids. A) Increased coverage over the exons of the *LCORL* gene indicate the presence of several *LCORL* retrocopies in equids. Notably, the third exon is missing from the retrocopy transcript. B) Visual representation of the variation across the *LCORL* gene highlighting the sequence differentiation within the *LCORL* retrocopies between species. Each colored line represents a polymorphism within the

retrocopy sequence. Not all variants appear fixed across all retrocopies. C) *LCORL* expression analysis in horses based on the SNVs identified between the retrocopy and the parent gene indicated that the *LCORL* retrocopies comprised the majority of overall *LCORL* transcripts in all tissue types except testis. D) Overall *LCORL* expression was also greatly increased in horses relative to humans.

LCORL expression analysis

RNAseq data from the Functional Annotation of Equine Genome (FAANG) dataset (*34*) were then analyzed using the *LCORL* retrocopy specific variants to differentiate the parental gene transcripts from the retrocopy transcripts. The majority of *LCORL* transcripts were determined to be derived from the retrocopies (Figure 2C; Supplemental Figure S2). This same trend was observed in a donkey RNAseq dataset, where the majority of *LCORL* expression could be attributed to the retrocopy (Supplemental Figure S3). Overall expression of *LCORL* was also increased in all horse tissues relative to humans (Figure 2D), with an average fold increase in TPM across 12 tissues of 11.8 (95% CI 7.4-16.3). The increase in overall expression of *LCORL* in horses was due entirely to expression of the retrocopies (Supplemental Figure S4). *MRPL15*, which is also included within the segmentally duplicated region in equids, was also overexpressed in horses compared to humans (Supplemental Figure S5; average fold-increase in expression 12.1, 95% CI 8.1-16.1). Tissue specific trends in expression of *MRPL15* and *LCORL* in donkeys and horses are available in Supplemental Figure S6. Long read RNAseq data was also analyzed to characterize the full length *LCORL* retrocopy transcript. The *LCORL* retrocopy often appears to form a chimeric transcript with nearby *MRPL15* (Supplemental Figure S7). Additionally, antisense transcripts of the *LCORL* retrocopy were observed at the insertion site.

Phylogenetic analysis of the LCORL retrocopy

Phylogenetic analysis using the *LCORL* retrocopy variants replicate known relationships between the species and that the *LCORL* retrocopy insertion predates equid speciation (Figure 3A). Interestingly, rearrangements of the segmental duplication containing the *LCORL* retrocopy appear to have resulted in gene conversions: several variants were identified in both the parent gene and retrocopy sequence in one species while absent from other species' parent or retrocopy sequences. No evidence of *LCORL* retrocopies was found in genome assemblies of two other extant members of the order Perissodactyla, the South American tapir (*Tapirus terrestris*) and the white rhinoceros (*Ceratotherium simum*). Additionally, no *LCORL* retrocopies were present within the WGS data of two tapirs (*Tapirus terrestris* and *Tapirus indicus*) and two rhinoceroses (*Rhinoceros unicornis* and *Cerototherium simum*), indicating that the *LCORL* retrocopies were specific to the Equidae family (Figure 3B).

Since polymorphism exists within the duplicated copies of the *LCORL* retrocopy at varying allele frequencies, the single haplotype retroCNV sequences present within the EquCab2.0, ASM1607732v2 and UCLA_HA_Equagga_1.0 references assemblies were compared to their parent gene sequences in order to estimate how old the *LCORL* retrocopy is (Supplemental Table S11). Estimates ranged from 12.2 to 21.7 million years ago (MYA), with an average age estimate of 18.0 MYA (95% CI 17.0-19.0 MYA). The age of the retrotransposition placed this event much deeper than the extant Equus genus and lead us to explore the timing relative to evolutionary changes that occurred within the Equidae family. The estimate of the time of the *LCORL* retrocopy insertion was compared to the increase in body weight and decrease in digit number during equid evolution (Figure 4).

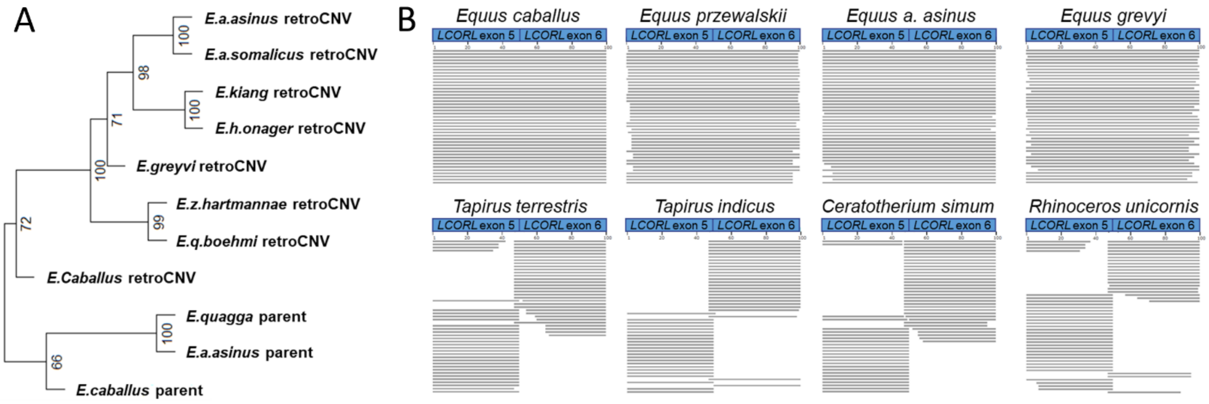


Figure 3: Across species analysis of *LCORL* retrocopies. A) Variants within the *LCORL* retrocopies as compared to the parent *LCORL* genes are used to reconstruct the equine phylogeny, indicating that the retrocopy inserted prior to Equus divergence. Approximately unbiased p-values indicating the strength of the clustering are shown at each node. *LCORL* parent gene sequences were obtained from three recent equine genome assemblies: EquCab3.0 (*E. caballus*), UCLA_HA_Equagga_1.0 (*E. quagga*) and ASM1607732v2 (*E. asinus*). B) WGS reads spanning an *LCORL* exon-exon junction are observed in all Equids (top) while absent from tapirs and rhinoceroses.

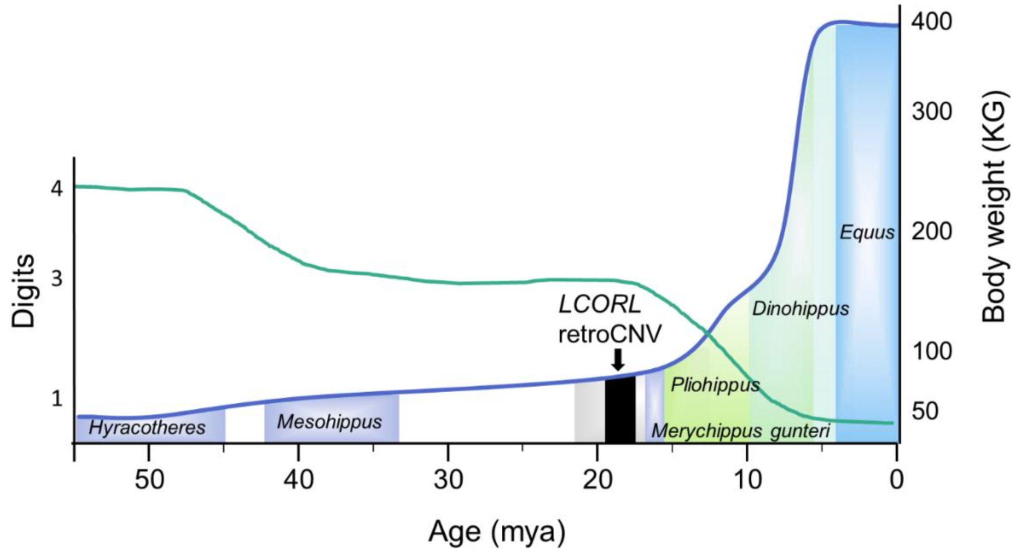


Figure 4: Insertion of the *LCORL* retrocopy within the timeline of equid evolution. The body weight (blue line) and digit number (green line) are shown relative to the evolutionary history of the equid. Additional changes that occurred not shown are changes in tooth anatomy. There is overlap of different equid species which is shown by transparency of the bars representing the different species. The black bar represents the average age estimate with 95% confidence interval of the *LCORL* retrocopy insertion. The grey bar shows the complete range of insertion times.

DISCUSSION

In this study, a recently developed approach to retroCNV discovery was successfully applied to equine WGS data to resolve the insertion sites for 437 retroCNVs. Horses had on average 22.5 retroCNV each, and 219 retroCNV were specific to horses. While the wild equids had a larger number of retroCNV on average (61.1), this is likely due to the use of a horse reference assembly for discovery. When retrocopies present in the EquCab3.0 reference assembly itself were also analyzed, 32 retrocopies were found to be absent from all wild equids, identifying them as horse specific retroCNV. Of particular interest was the identification of a large number of *LCORL* retrocopies in all equids. The *LCORL* retrocopies were also shown to be highly expressed in all tissue types analyzed in both horse and donkey RNAseq data and is part of a segmental duplication whose copy number was estimated to vary between species and individuals. Based on the number of SNV between the *LCORL* parent gene and the retrocopy, the retroCNV insertion, common to all extant equids, was dated to 18 MYA. Both the age of the *LCORL* retrogene and the conservation of a large number of expressed *LCORL* retrocopies in all extant equids provides evidence of function for the *LCORL* retrocopies in equid evolution.

A recent study which used the same method to identify retroCNVs in human and dog genomes found on average 4.2 retroCNVs in humans, while dogs had on average 54.1 retroCNVs (*16*). In dogs, the increased rate in gene retrotransposition relative to humans was correlated with increased LINE-1 and SINE dimorphism rates, possibly indicative of higher LINE-1 activity in general (*35, 36*). The domestic horse, with 22.5 retroCNVs on average, appears to have an intermediary rate of gene retrotransposition between that of humans and dogs. An analysis of the horse genome identified only 72 full length, intact LINE-1 sequences, which is fewer than both humans, which have 142, and dogs, which have 264 (*10*), indicating that while horses have fewer total LINE-1, LINE-1 appears to be more actively capable of gene retrotransposition in horses than in humans. As retroCNV are only one aspect of overall LINE-1 activity, a proportional increase in frequency for LINE-1 and SINE polymorphisms could be expected in equine genomes relative to humans. However, phenotypic associations with transposable element insertions are relatively rare in horses, being responsible for less than 1% (1/103) of the total phenotype associated variants identified in horses, compared to around 10% in dogs (Online Mendelian Inheritance in Animals, OMIA. Sydney School of Veterinary Science, 10/18/2022. World Wide Web URL: https://omia.org/).

RetroCNV, which are a form of large structural variant, can be difficult to identify from Illumina short read datasets. The retroCNVs identified in low coverage human genomes have varied based on the approach used (*2, 9, 12, 37*). Some approaches to retroCNV discovery also focus on retroCNV parental gene analysis (*15*), but the resolution of the retroCNV insertion site is important for downstream functional analyses. Additionally, multiple insertions can exist for a single parent gene; in this study, 437 total insertions were identified from 353 parent genes. While a previous analysis using multiple long-read human genome assemblies appeared to give the most accurate results for retroCNV, this quality of data is not yet available for most species (*14*). The approach to retroCNV employed in the current study was also used in canids, where it was tested using long-read assemblies and found to be highly accurate (*16*). SideRETRO, a recently developed tool for retroCNV discovery (*38*), was also applied to the same dataset used in this study, and 177 of the same retroCNV were confirmed as well as 20 unique retroCNV. However, sideRETRO failed to identify 260 of retroCNVs which were identified using discordant reads, possibly due to issues in the annotation file or the parameters used. While sideRETRO is a useful, easy to use tool, manual curation of the output file was still required to remove false positives caused by the presence of other complex structural variants in the genome. Due to the complexity of structural variant analyses such as retroCNV discovery, any technique applied requires validation. A previous study focused on Thoroughbreds found 62 retroCNV parent genes which were PCR validated (*17*); the current study confirmed the presence of 54 of those retroCNV, including the identification of the insertion sites.

The GO enrichment analysis of the retroCNV parent genes showed that many were derived from genes encoding ribosomal proteins, which are known to be common retrocopy forming genes (*39*). Retrocopies tend to derive from highly expressed genes, including genes encoding ribosomal proteins (*9, 39, 40*). However, the retroCNV identified in this study may function through a variety of mechanisms, including alteration of nearby gene expression (*19, 20*). In natural populations of mice, many retroCNV were shown to significantly alter the overall expression of the parent gene, and the retroCNV were also under negative selection, likely due to deleterious effects (*13*). Still, many retroCNV appear to have been selected towards fixation in horses; 21 retrocopies were identified in the EquCab3.0 genome assembly that were homozygous in all horses yet absent from all wild equids.

More intriguing still is the presence of many *LCORL* retrocopies across all equids. The *LCORL* retrocopies, which are part of a larger duplication that is poorly assembled in most available equine genome assemblies, were only identified through our discovery pipeline due to their absence from the EquCab3.0 assembly. In humans, 7% of the human genome consists of segmental duplications and, in the past, they were also poorly resolved in genome assemblies

(*41*). The *LCORL* retrocopies were previously mapped to horse chromosome 9 using fluorescence in situ hybridization, where it was also noted that very strong signals indicated the presence of multiple copies (*26*). Discordant read mapping confirmed the *LCORL* retrocopies were on chromosome 9. While multiple *LCORL* retrocopies were present in the Dezhou Donkey assembly (ASM1607732v2), they appeared on different chromosomes, indicating that the *LCORL* retrocopy had translocated between species or that there may be an error in the assembly. Increased coverage over the segmental duplication containing the *LCORL* retrocopies confirmed the presence of many copies, with horses having 25 on average, which warranted further investigation of the retroCNV.

The *LCORL* retrocopies are of particular interest due to the nature of the gene; the *LCORL*/*NCAPG* locus has been associated with body size across many mammalian species, including humans and horses, with identification of the causative mutation often difficult (*26-29, 31, 33, 42*). While *LCORL* encodes a putative transcription factor, the function of *LCORL* is not well understood (*43*). Alternative splicing of the *LCORL* gene results in multiple transcript variants as well as multiple vastly different protein isoforms (*31*). How *LCORL* mediates changes in body size is still unknown, making it difficult to speculate how transcription of the retrocopy might influence morphology. The equid *LCORL* retrocopies have accumulated numerous mutations, including frameshift mutations and a predicted nonsense mutation in exon 4, resulting in a truncated protein product. However, the *LCORL* retrocopies were shown to be highly expressed in horse and donkey tissues, where the retrocopies made up the vast majority of overall *LCORL* expression. Even if they no longer code for a functional protein, the *LCORL* retrocopy transcripts, which can form chimeric transcripts with the nearby *MRPL15* gene, may still have an effect on expression or translation of the parent gene (*19, 20*). The presence of these highly expressed *LCORL* retrocopies may have also interfered with functional analyses into the *LCORL* parent gene (*27, 44*). Additionally, the large copy number of the *MRPL15* gene itself may also have functional consequences in equids. *MRPL15* is a mitochondrial ribosomal protein that plays a role in protein synthesis within the mitochondria which has been identified as a biomarker for cancers (*45-47*). Overall expression of *MRPL15* was increased in horse and donkey tissues relative to humans, likely because of the high copy number of the gene. While the known involvement of *LCORL* body size is intriguing, *MRPL15* could be contributing to the function of the segmental duplication or the chimeric reads between the two genes could be driving the selection and conservation of this segmental duplication.

The *LCORL* retrocopies are absent from rhinoceros and tapir genomes, the two other members of Perissodactyla, making the retrocopies unique to the equid family, the only living monodactyly (*48*). While the *LCORL* retrocopies are present in all equid species, the retrocopy sequences have begun to diverge; separate variants have become fixed across all retrocopies differentially between the species. This is possible evidence of gene conversion due to the effects of concerted evolution, a process which results in greater sequence similarity of repetitive elements within than among species (*49*). This, in addition to the across-species perseveration of large copy numbers of the *LCORL* retrocopies, imply a functional role. Notably, a similar process of retrocopy insertion followed by segmental duplication has resulted in a large number of partial *TP53* gene retrocopies in elephants (21). Although many of the *TP53* retrocopies are truncated and no longer functional, they have been under positive selection due to a predicted role in cancer resistance among elephants, although this conclusion has been brought into question recently (*50*). Another example of retrocopies leading to expanded gene copy number is for the APOBEC3 antiviral proteins in primates (*51*). It's possible that the *LCORL* retrocopies play a similarly important role in the evolutionary history of the equids.

The *LCORL* retrotransposition was estimated to occur 18 million years ago based on sequence divergence between the parent genes and the retrocopy sequences available in assemblies from horse, donkey, and zebra. The retrotransposition event happened first, followed by segmental duplication events and additional subsequent rearrangements within the extant Equids. In addition to preservation of high copy number of the segmental duplication, there has been high expression levels maintained across the extant equids a time frame of almost 5 million years. The large size and complex nature of the segmental duplication make it challenging to completely resolve and as a result the functionality of this highly transcribed locus is difficult to test. Without DNA sequence data from other extinct equids, it is also difficult to verify the timing of the retrotransposition event or the expansion in copy number. The complete landscape of segmental duplications as well as other evolutionary changes during equid evolution when identified may provide additional insight into the importance of this locus, however the timing of the retrotransposition of *LCORL* predating the major increase in body size and digit reduction in equids is compelling.

METHODS

RetroCNV parent gene discovery

A previously described method utilizing mRNA specific 30-mers to identify putative retroCNV parent genes was adopted for the EquCab3.0 reference (*16*). Briefly, spliced gene sequences for every gene transcript in the NCBI EquCab3.0 annotation release 103 were obtained using Gffread (*52*). mRNA specific 30-mers which are absent from the EquCab3.0 assembly were then identified for each gene sequence using Jellyfish (*53*). To prevent false positives due to sequencing errors, all 30-mers with an edit distance of 2 substitutions from the EquCab3.0 reference assembly were filtered using mrsFAST (*54*). In total, 26,849 genes were used for retroCNV discovery using 5,672,946 unique 30-mers (median 30-mers per gene: 131). RetroCNV discovery was then performed using high coverage Illumina paired end whole genome sequencing (WGS) fastq files downloaded from the Sequence Read Archive or European Bioinformatics Institute (*55*). The dataset included 75 breed horses, 3 Przewalski horses and 8 wild Equids (Supplemental Table S1). Fastq files were queried for the presence of the mRNA specific 30-mers using Jellyfish, and any genes which had greater than 5 mRNA specific 30-mers and at least 10% of the total 30-mers for that gene identified were considered as a putative retroCNV parent genes.

RetroCNV insertion site analysis

RetroCNV insertion sites were identified using a previously developed pipeline (*16*). Briefly, WGS fastq files were aligned to the EquCab3.0 reference assembly (*56*) using Minimap2 v2.24 with the preset '-ax sr' for Illumina paired end reads (*57*). Aligned data were sorted and duplicate reads were removed using samtools (*58*). TEBreak was then used to obtain discordant read clusters at putative retroCNV parent gene loci (*59*). All retroCNV insertion sites were visually confirmed in Integrative Genomics Viewer (IGV) (*60*). The TEBreak 5' and 3' junction sequences for the retroCNV are available in Supplemental Table S12. To validate a set of retroCNV insertion sites and TSD, three primer genotyping PCR assays were designed as previously described (Supplemental Table S13). Thoroughbred horses from a DNA repository maintained by the Bannasch lab were then selected at random for genotyping, and Sanger sequencing was performed on an Applied Biosystems 3500 Genetic Analyzer using a Big Dye Terminator Sequencing Kit (Life Technologies, Burlington, ON, Canada). The horse Y chromosome assembly (*61*) was analyzed for evidence of RetroCNV that had been predicted to be on the Y chromosome based on sex bias. The retroCNV parent gene sequence was used to query the Y chromosome for the retrocopy using BLAST (*62*). SideRETRO was then used with default settings on the same dataset in a separate analysis (*38*).

Reference assembly retrocopy analysis

Gene retrocopies present in RetrogeneDB (*8*) were analyzed to identify retroCNV within the EquCab3.0 reference assembly. Retrocopy locations were first converted to EquCab3.0 using the NCBI remap tool, and then WGS data were analyzed for deletions at these locations using Delly (*63*). Deletions which confirmed the retrocopy as a retroCNV were manually confirmed visually using IGV. All recent (>95% identity, N=88) retrocopies from RetrogeneDB were also visually analyzed in the 8 wild equids using IGV to manually determine if they were retroCNV.

RetroCNV specific variant analysis

Single nucleotide variants (SNV) at retroCNV parent gene loci were identified from the WGS dataset using bcftools mpileup (*58*). For each gene, SNVs which were present only in individuals who were positive for a retroCNV of that gene were attributed to the retroCNV, while SNV which were present in individuals who lacked the retroCNV were considered to be variants within the parental gene sequence itself. RetroCNV which were unique to wild equids were excluded from this analysis.

LCORL retroCNV copy number

In order to estimate the total copy number of the *LCORL* retrocopies in each sample, the average coverage over a portion of the segmental duplication (Equcab3 Chr9:30186485-30216186) was first calculated using the samtools depth (*58*). This was then divided by the average genome-wide coverage which was estimated using samtools idxstats (*58*).

LCORL expression analysis

A list of *LCORL* retrocopy specific variants was obtained from WGS data using bcftools mpileup at the *LCORL* parent gene locus (*58*). The *LCORL* retrocopy specific variants were analyzed in IGV, and only variants which appeared fixed across all retrocopies were used to create a masked *LCORL* transcript (Supplemental Data 3). Illumina paired-end RNA-seq data were downloaded from the Functional Annotation of Animal Genomes (FAANG) archive (*34*) and aligned to the masked *LCORL* sequence using HISAT2 (*64*). SNPsplit was then used to perform allele specific analysis of *LCORL* to differentiate the parental gene transcripts from the retroCNV transcripts (*65*). RNA-seq data were also aligned to the full EquCab3.0 reference assembly using HISAT2 and transcripts per million (TPM) for *LCORL* and *MRPL15* were counted using Kallisto (*66*). TPM were then averaged across all sample of the same tissue type. The same analysis was also performed using Illumina paired-end RNA-seq data from a Dezhou Donkey, BioProject accession PRJNA431818 (*67*) using a separate masked *LCORL* sequence file using the wild equid *LCORL* retrocopy specific variants (Supplemental Data 4). Normalized expression data in humans for *MRPL15* and *LCORL* was obtained from the Human Protein Atlas website (proteinatlas.org) (*68*). Expression of the *LCORL* retrocopies was also analyzed in long-read RNAseq data from the FAANG initiative (*69*). Circular consensus sequence files were downloaded and aligned to EquCab3.0 using minimap2 v.2.24 using the preset '-ax splice:hq' for spliced Iso-seq data. Aligned files were then visually analyzed in IGV (*60*).

Phylogenetic analysis of the LCORL retrocopies

Two recent, chromosome level wild equid genome assemblies were queried for evidence of the *LCORL* retrocopy via BLAST using the *LCORL* cDNA sequence: The Dezhou donkey (*Equus asinus*) assembly ASM1607732v2 (*67*), and the plains zebra (*Equus quagga)* assembly UCLA_HA_Equagga_1.0 (*70*). The *LCORL* retrocopy sequences were extracted from the assemblies, aligned to the respective parent gene sequence using MUSCLE (*71*), and the number of SNPs between the retrocopies and parent gene were counted using SNP-sites (*72*). The retroCNV were then aged using an estimated mutation rate in horses of 7.24 x 10^{-9} mutations/site/generation, with a generation time of 8 years (*73, 74*). The Southern white rhinoceros (*Ceratotherium simum*) and South American tapir (*Tapirus terrestris*) genome assemblies (*75*) were also searched for evidence of the *LCORL* retrocopies with BLAST using the *LCORL* cDNA sequence. To further confirm that no *LCORL* retrocopies were present in other Perissodactylas, WGS data from two species of rhinoceros (DRR308100, SRR20637451) and two species of tapir (SRR11097180, SRR13167957) were also queried via BLAST using a spliced *LCORL* sequence between exon 5 and exon 6 (Supplemental data 5). Phylogenetic analysis of the *LCORL* parent genes and retrocopies was performed using the previously identified retrocopy specific variants (supplemental table S10) and the R package pvclust with method.dist set to Euclidian (*76*). Information on weight, age, and digit status for the Equidae were obtained from previous publications (*74, 77-80*).

DATA ACCESS

The retroCNV insertion sites in bigBed format and a track hub for the UCSC Genome Browser are available at GitHub: https://github.com/klbatcher/retroCNV_insertions.

ACKNOWLEDGEMENTS

This work was supported in part by the Lodric Maddox Graduate Fellowship and Maxine Adler Endowed Chair funds. The unpublished genome assemblies of *Ceratotherium simum* and *Tapirus terrestris* were used with permission from the DNA Zoo Consortium (dnazoo.org).

APPENDIX REFERENCES

1. E. M. Ostertag, H. H. Kazazian Jr, Biology of mamalian L1 retrotransposons. *Annual review of genetics* **35**, 501 (2001).
2. S. R. Richardson, C. Salvador-Palomeque, G. J. Faulkner, Diversity through duplication: Whole-genome sequencing reveals novel gene retrocopies in the human population. *Bioessays* **36**, 475-481 (2014).
3. D. D. Luan, M. H. Korman, J. L. Jakubczak, T. H. Eickbush, Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. *Cell* **72**, 595-605 (1993).
4. J. Jurka, Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. *Proceedings of the National Academy of Sciences* **94**, 1872-1877 (1997).
5. C. Esnault, J. Maestre, T. Heidmann, Human LINE retrotransposons generate processed pseudogenes. *Nature genetics* **24**, 363-367 (2000).
6. W. Wei *et al.*, Human L1 retrotransposition: cispreference versus trans complementation. *Molecular and cellular biology* **21**, 1429-1439 (2001).
7. E. F. Vanin, Processed pseudogenes: characteristics and evolution. *Annual review of genetics* **19**, 253-272 (1985).
8. W. Rosikiewicz *et al.*, RetrogeneDB–a database of plant and animal retrocopies. *Database* **2017**, (2017).
9. A. D. Ewing *et al.*, Retrotransposition of gene transcripts leads to structural variation in mammalian genomes. *Genome biology* **14**, R22 (2013).

10. T. Penzkofer *et al.*, L1Base 2: more retrotransposition-active LINE-1s, more mammalian genomes. *Nucleic acids research*, gkw925 (2016).
11. C. Casola, E. Betrán, The genomic impact of gene retrocopies: what have we learned from comparative genomics, population genomics, and transcriptomic analyses? *Genome biology and evolution* **9**, 1351-1373 (2017).
12. D. R. Schrider *et al.*, Gene copy-number polymorphism caused by retrotransposition in humans. *PLoS genetics* **9**, e1003242 (2013).
13. W. Zhang, C. Xie, K. Ullrich, Y. E. Zhang, D. Tautz, The mutational load in natural populations is significantly affected by high primary rates of retroposition. *Proceedings of the National Academy of Sciences* **118**, (2021).
14. X. Feng, H. Li, Higher Rates of Processed Pseudogene Acquisition in Humans and Three Great Apes Revealed by Long-Read Assemblies. *Molecular Biology and Evolution* **38**, 2958-2966 (2021).
15. Y. Zhang, S. Li, A. Abyzov, M. B. Gerstein, Landscape and variation of novel retroduplications in 26 human populations. *PLoS computational biology* **13**, e1005567 (2017).
16. K. Batcher *et al.*, Recent, full-length gene retrocopies are common in canids. *Genome Research*, (2022).
17. T. Tozaki *et al.*, Identification of processed pseudogenes in the genome of Thoroughbred horses: Possibility of gene-doping detection considering the presence of pseudogenes. *Animal Genetics* **53**, 183-192 (2022).
18. R.-L. Troskie *et al.*, Long-read cDNA sequencing identifies functional pseudogenes in the human transcriptome. *Genome biology* **22**, 1-15 (2021).
19. S. W. Cheetham, G. J. Faulkner, M. E. Dinger, Overcoming challenges and dogmas to understand the functions of pseudogenes. *Nature Reviews Genetics* **21**, 191-201 (2020).
20. J. Ciomborowska-Basheer, K. Staszak, M. R. Kubiak, I. Makałowska, Not So Dead Genes—Retrocopies as Regulators of Their Disease-Related Progenitors and Hosts. *Cells* **10**, 912 (2021).
21. L. M. Abegglen *et al.*, Potential mechanisms for cancer resistance in elephants and comparative cellular response to DNA damage in humans. *Jama* **314**, 1850-1860 (2015).
22. M. Sulak *et al.*, TP53 copy number expansion is associated with the evolution of increased body size and an enhanced DNA damage response in elephants. *elife* **5**, e11994 (2016).
23. E. Kuzmin, J. S. Taylor, C. Boone, Retention of duplicated genes in evolution. *Trends in Genetics*, (2021).
24. H. G. Parker *et al.*, An expressed fgf4 retrogene is associated with breed-defining chondrodysplasia in domestic dogs. *Science* **325**, 995-998 (2009).
25. E. A. Brown *et al.*, FGF4 retrogene on CFA12 is responsible for chondrodystrophy and intervertebral disc disease in dogs. *Proceedings of the National Academy of Sciences* **114**, 11476-11481 (2017).
26. E. A. Staiger *et al.*, Skeletal variation in Tennessee Walking Horses maps to the *LCORL*/NCAPG gene region. *Physiological genomics* **48**, 325-335 (2016).
27. J. Metzger, R. Schrimpf, U. Philipp, O. Distl, Expression levels of *LCORL* are associated with body size in horses. *PloS one* **8**, e56497 (2013).

28. J. Tetens, P. Widmann, C. Kühn, G. Thaller, A genome-wide association study indicates *LCORL*/NCAPG as a candidate locus for withers height in German Warmblood horses. *Animal genetics* **44**, 467-471 (2013).
29. N. Soranzo *et al.*, Meta-analysis of genome-wide scans for human adult stature identifies novel Loci and associations with measures of skeletal frame size. *PLoS genetics* **5**, e1000445 (2009).
30. A. K. Lindholm-Perry *et al.*, Adipose and muscle tissue gene expression of two genes (NCAPG and *LCORL*) located in a chromosomal region associated with cattle feed intake and gain. *PloS one* **8**, e80882 (2013).
31. R. Saif *et al.*, The *LCORL* locus is under selection in large-sized Pakistani goat breeds. *Genes* **11**, 168 (2020).
32. A. Takasuga, PLAG1 and NCAPG-*LCORL* in livestock. *Animal Science Journal* **87**, 159-167 (2016).
33. J. Plassais *et al.*, Whole genome sequencing of canids reveals genomic regions under selection and variants influencing morphology. *Nature communications* **10**, 1-14 (2019).
34. E. N. Burns *et al.*, Generation of an equine biobank to be used for Functional Annotation of Animal Genomes project. *Animal genetics* **49**, 564-570 (2018).
35. J. V. Halo *et al.*, Long-read assembly of a Great Dane genome highlights the contribution of GC-rich sequence and mobile elements to canine genomes. *Proceedings of the National Academy of Sciences* **118**, (2021).
36. W. Wang, E. F. Kirkness, Short interspersed elements (SINEs) are a major source of canine genomic diversity. *Genome Research* **15**, 1798-1808 (2005).
37. A. Abyzov *et al.*, Analysis of variable retroduplications in human populations suggests coupling of retrotransposition to cell division. *Genome research* **23**, 2042-2052 (2013).
38. T. L. Miller, F. Orpinelli Rego, J. L. L. Buzzo, P. A. Galante, sideRETRO: a pipeline for identifying somatic and polymorphic insertions of processed pseudogenes or retrocopies. *Bioinformatics* **37**, 419-421 (2021).
39. I. Gonçalves, L. Duret, D. Mouchiroud, Nature and structure of human genes that generate retropseudogenes. *Genome research* **10**, 672-678 (2000).
40. Z. Zhang, P. M. Harrison, Y. Liu, M. Gerstein, Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome. *Genome research* **13**, 2541-2558 (2003).
41. M. R. Vollger *et al.*, Segmental duplications and their variation in a complete human genome. *Science* **376**, eabj6965 (2022).
42. S. Makvandi-Nejad *et al.*, Four loci explain 83% of size variation in the horse. *PLoS One* **7**, e39929 (2012).
43. T. Kunieda, J.-M. Park, H. Takeuchi, T. Kubo, Identification and characterization of Mlr1, 2: two mouse homologues of Mblk-1, a transcription factor from the honeybee brain. *FEBS letters* **535**, 61-65 (2003).
44. K. Srikanth *et al.*, Comprehensive genome and transcriptome analyses reveal genetic relationship, selection signature, and transcriptome landscape of small-sized Korean native Jeju horse. *Scientific reports* **9**, 1-16 (2019).

45. F. Sotgia, M. Fiorillo, M. P. Lisanti, Mitochondrial markers predict recurrence, metastasis and tamoxifen-resistance in breast cancer patients: Early detection of treatment failure with companion diagnostics. *Oncotarget* **8**, 68730 (2017).
46. F. Deng, L. Shen, H. Wang, L. Zhang, Classify multicategory outcome in patients with lung adenocarcinoma using clinical, transcriptomic and clinico-transcriptomic data: machine learning versus multinomial models. *American journal of cancer research* **10**, 4624 (2020).
47. Y. Zeng *et al.*, Prognostic value and related regulatory networks of *MRPL15* in Non-Small-cell lung cancer. *Frontiers in oncology*, 1479 (2021).
48. B. K. McHorse, A. A. Biewener, S. E. Pierce, The evolution of a single toe in horses: causes, consequences, and the way forward. *Integrative and Comparative Biology* **59**, 638-655 (2019).
49. J. F. Elder Jr, B. J. Turner, Concerted evolution of repetitive DNA sequences in eukaryotes. *The Quarterly review of biology* **70**, 297-320 (1995).
50. L. Nunney, Cancer suppression and the evolution of multiple retrogene copies of TP53 in elephants: A re-evaluation. *Evolutionary Applications* **15**, 891-901 (2022).
51. L. Yang, M. Emerman, H. S. Malik, R. N. M. Jnr, Retrocopying expands the functional repertoire of APOBEC3 antiviral proteins in primates. *Elife* **9**, e58436 (2020).
52. G. Pertea, M. Pertea, GFF utilities: GffRead and GffCompare. *F1000Research* **9**, (2020).
53. G. Marcais, C. Kingsford, Jellyfish: A fast k-mer counter. *Tutorialis e Manuais* **1**, 1-8 (2012).
54. F. Hach *et al.*, mrsFAST-Ultra: a compact, SNP-aware mapper for high performance sequencing applications. *Nucleic acids research* **42**, W494-W500 (2014).
55. R. Leinonen, H. Sugawara, M. Shumway, I. N. S. D. Collaboration, The sequence read archive. *Nucleic acids research* **39**, D19-D21 (2010).
56. T. S. Kalbfleisch *et al.*, Improved reference genome for the domestic horse increases assembly contiguity and composition. *Communications biology* **1**, 1-8 (2018).
57. H. Li, Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094-3100 (2018).
58. P. Danecek *et al.*, Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).
59. P. E. Carreira *et al.*, Evidence for L1-associated DNA rearrangements and negligible L1 retrotransposition in glioblastoma multiforme. *Mobile DNA* **7**, 1-14 (2016).
60. H. Thorvaldsdóttir, J. T. Robinson, J. P. Mesirov, Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in bioinformatics* **14**, 178-192 (2013).
61. J. E. Janečka *et al.*, Horse Y chromosome assembly displays unique evolutionary features and putative stallion fertility genes. *Nature communications* **9**, 1-15 (2018).
62. T. Madden, The BLAST sequence analysis tool. *The NCBI handbook* **2**, 425-436 (2013).
63. T. Rausch *et al.*, DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333-i339 (2012).
64. D. Kim, J. M. Paggi, C. Park, C. Bennett, S. L. Salzberg, Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature biotechnology* **37**, 907-915 (2019).

65. F. Krueger, S. R. Andrews, SNPsplit: Allele-specific splitting of alignments between genomes with known SNP genotypes. *F1000Research* **5**, (2016).

66. N. L. Bray, H. Pimentel, P. Melsted, L. Pachter, Near-optimal probabilistic RNA-seq quantification. *Nature biotechnology* **34**, 525-527 (2016).

67. C. Wang *et al.*, Donkey genomes provide new insights into domestication and selection for coat color. *Nature communications* **11**, 1-15 (2020).

68. M. Uhlén *et al.*, Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).

69. S. Peng *et al.*, Long-read RNA Sequencing Improves the Annotation of the Equine Transcriptome. *bioRxiv*, (2022).

70. E. Cappelletti *et al.*, Robertsonian fusion and centromere repositioning contributed to the formation of satellite-free centromeres during the evolution of zebras. *Molecular Biology and Evolution*, (2022).

71. R. C. Edgar, MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research* **32**, 1792-1797 (2004).

72. A. J. Page *et al.*, SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *biorxiv*, 038190 (2016).

73. A. O. Vershinina *et al.*, Ancient horse genomes reveal the timing and extent of dispersals across the Bering Land Bridge. *Molecular Ecology* **30**, 6144-6161 (2021).

74. L. Orlando *et al.*, Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse. *Nature* **499**, 74-78 (2013).

75. O. Dudchenko *et al.*, De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92-95 (2017).

76. R. Suzuki, H. Shimodaira, Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* **22**, 1540-1542 (2006).

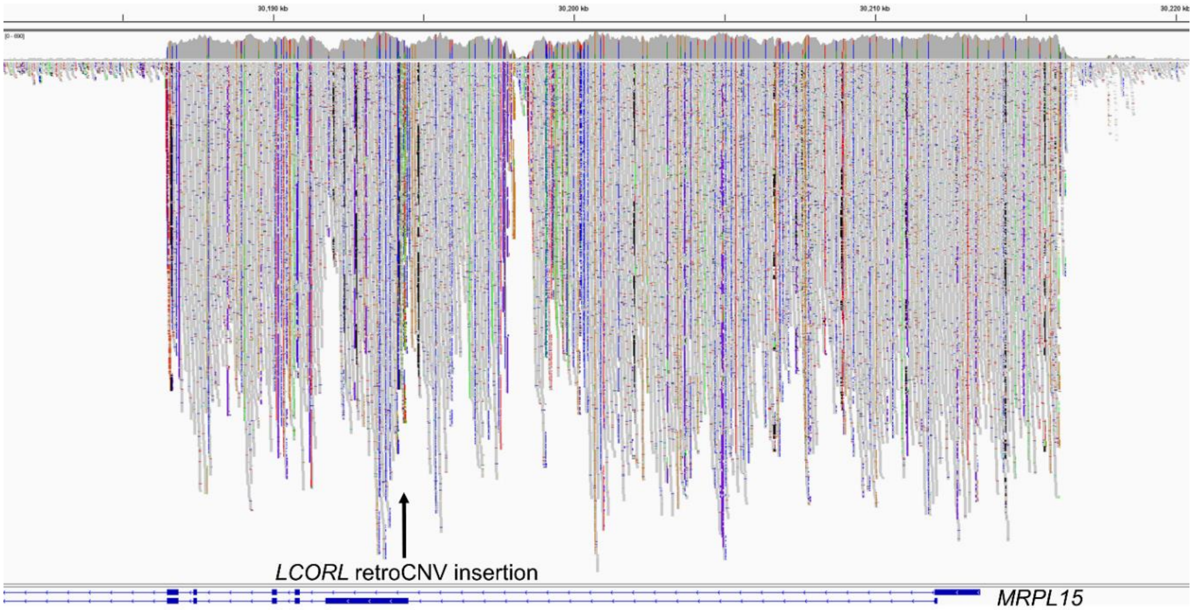
77. B. J. MacFadden, Fossil horses from “Eohippus”(Hyracotherium) to Equus: scaling, Cope's Law, and the evolution of body size. *Paleobiology* **12**, 355-369 (1986).

78. C. M. Janis, R. L. Bernor, The evolution of equid monodactyly: a review including a new hypothesis. *Frontiers in Ecology and Evolution* **7**, 119 (2019).

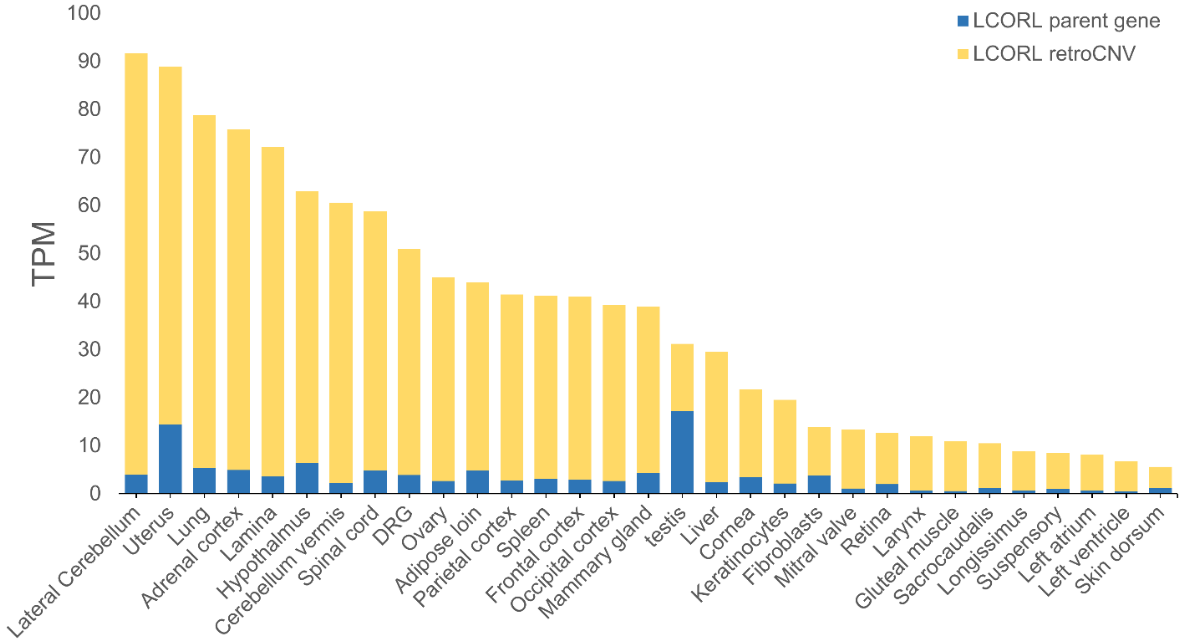
79. J. L. Cantalapiedra, J. L. Prado, M. Hernández Fernández, M. T. Alberdi, Decoupled ecomorphological evolution and diversification in Neogene-Quaternary horses. *Science* **355**, 627-630 (2017).

80. H. Jónsson *et al.*, Speciation with gene flow in equids despite extensive chromosomal plasticity. *Proceedings of the National Academy of Sciences* **111**, 18655-18660 (2014).

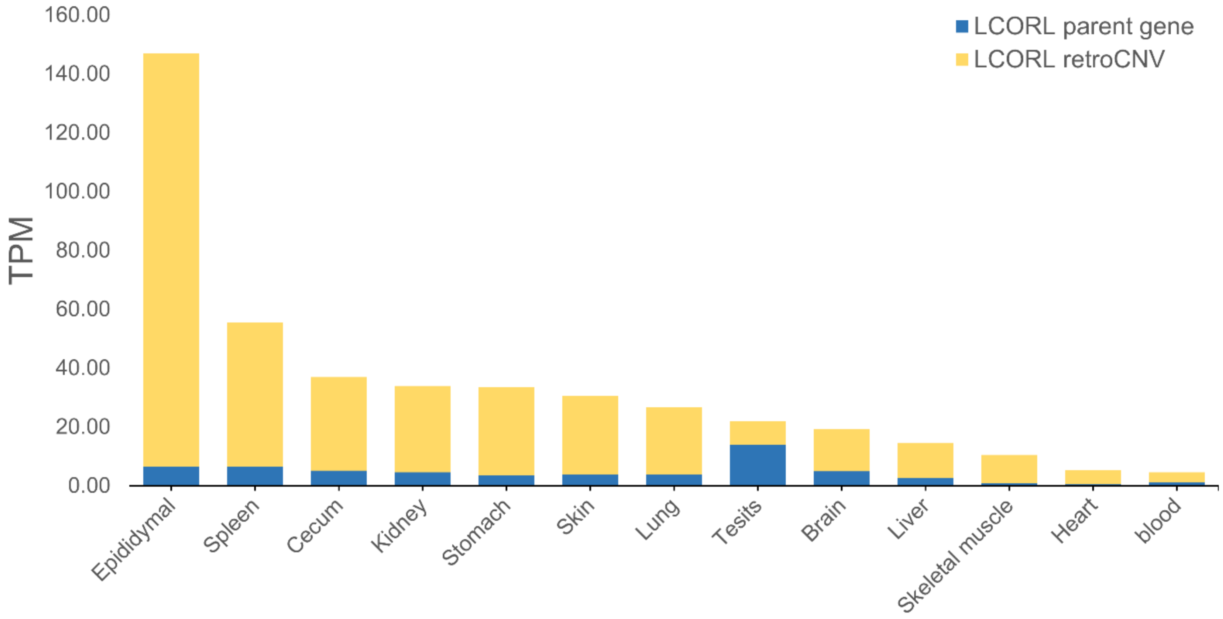
SUPPLEMENTAL FIGURES



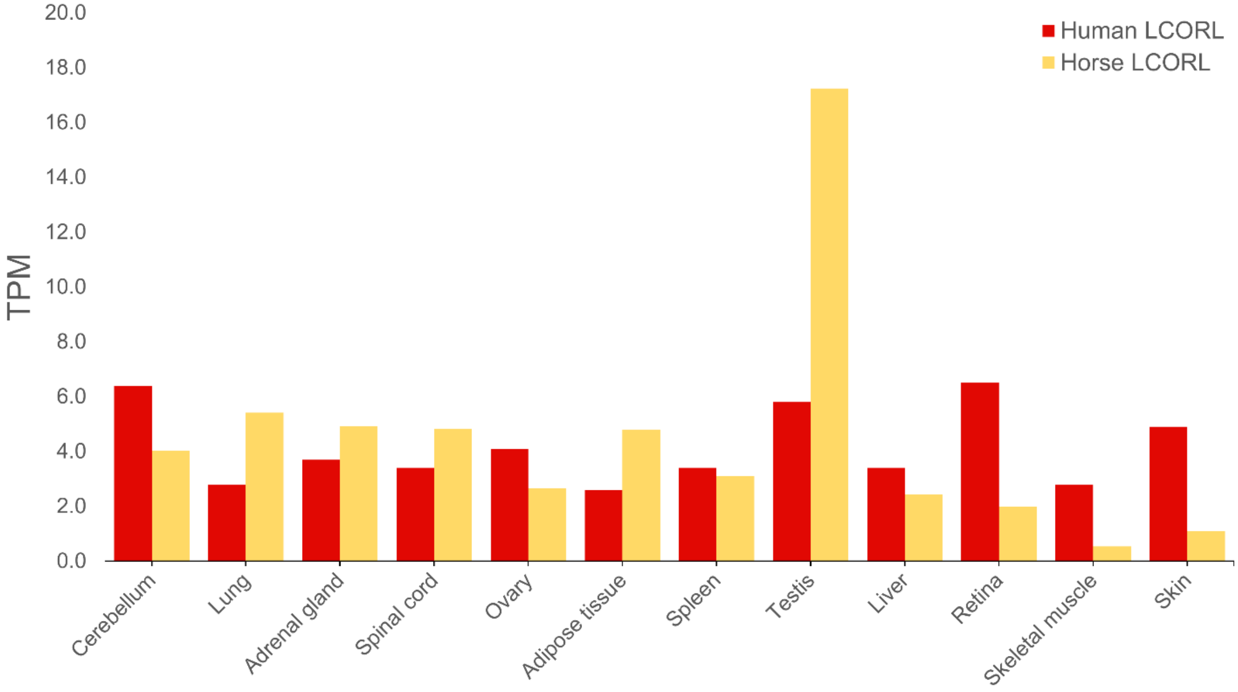
Supplemental Figure S1: Large number of expressed *LCORL* retrocopies in equids. A segmental duplication encompassing the *MRPL15* gene on chromosome 9 contains the *LCORL* retrocopy insertion at chr9:30194359-30194380.



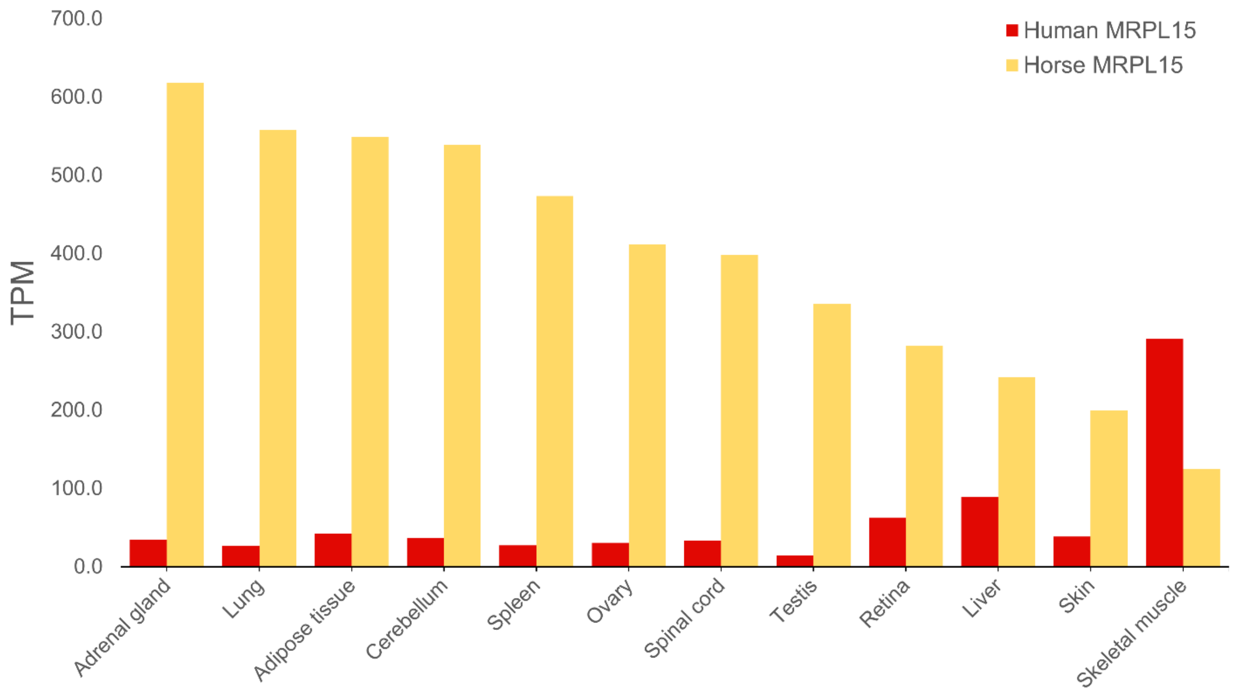
Supplemental Figure S2: *LCORL* expression analysis in horse tissue. Percentages of total *LCORL* transcripts derived from the retroCNV is shown for each tissue type. The *LCORL* retrocopy comprised the majority of overall *LCORL* transcripts in all tissue types except testis.



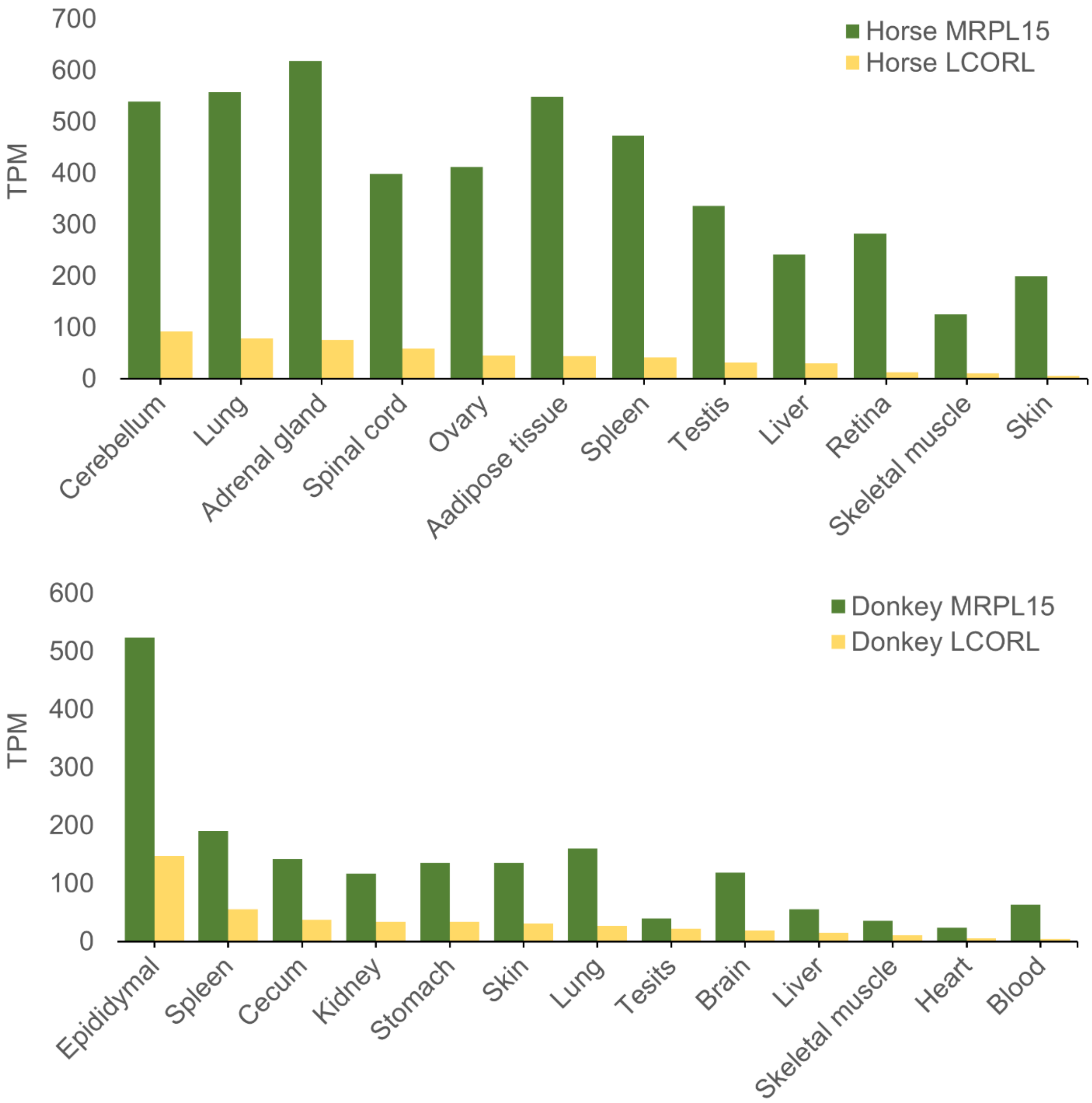
Supplemental Figure S3: *LCORL* retrocopies are highly expressed in donkey tissue. The *LCORL* retrocopy comprised the majority of overall *LCORL* transcripts in all tissue types except testis.



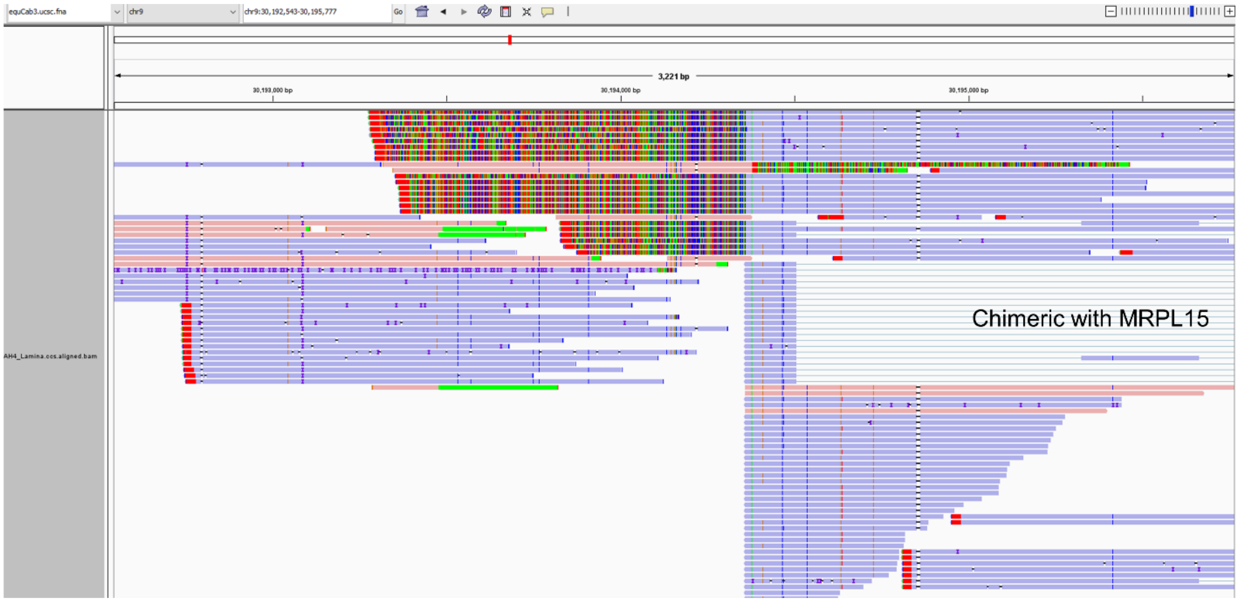
Supplemental Figure S4: Expression of the *LCORL* parent gene is comparable to expression in humans across tissue types.



Supplemental Figure S5: Increased expression of the segmentally duplicated *MRPL15* gene in horses compared to humans.



Supplemental Figure S6: Overall expression of *LCORL* and *MRPL15* in horses and donkeys.



Supplemental Figure S7: Long read RNAseq data viewed at the *LCORL* retrocopy insertion site. Chimeric reads were observed between the *LCORL* retrocopy and the nearby *MRPL15* gene. Antisense reads highlighted in red were also observed at the insertion site.