# Whole-Genome Sequencing of a Canine Family Trio Reveals a FAM83G Variant Associated with Hereditary Footpad Hyperkeratosis

Shumaila Sayyab [1,2*], Agnese Viluma [1*], Kerstin Bergvall [3], Emma Brunberg [4], Vidhya Jagannathan [5], Tosso Leeb [5], Göran Andersson [1§#], Tomas F. Bergström [1§#]

[1]Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences (SLU), Box 7023, 750 07 Uppsala, Sweden

[2]Research Center for Modeling and Simulation (RCMS), National University of Sciences and Technology (NUST), Sector H-12, Islamabad, Pakistan

[3]Department of Clinical Sciences, Swedish University of Agricultural Sciences (SLU), Box 7054, Uppsala, Sweden

[4]Norwegian Centre for Organic Agriculture, Gunnars veg 6, NO-6630 Tingvoll, Norway
[5]Institute of Genetics, University of Bern, 3001, Bern, Switzerland

*These authors contributed equally to this work

[§]Corresponding authors

#GA and TB share senior authorship

Email addresses:

TB: tomas.bergstrom@slu.se

GA: goran.andersson@slu.se

AV: agnese.viluma@slu.se

SS: shumaila@rcms.nust.edu.pk

KB: kerstin.bergvall@slu.se

EB: Emma.Brunberg@norsok.no

VJ: vidhya.jagannathan@vetsuisse.unibe.ch

TL: tosso.leeb@vetsuisse.unibe.ch

## Abstract

Over 250 Mendelian traits and disorders, caused by rare alleles have been mapped in the canine genome. Although each disease is rare in the dog as a species, they are collectively common and have major impact on canine health. With SNP-based genotyping arrays, genome-wide association studies (GWAS) have proven to be a powerful method to map the genomic region of interest when 10-20 cases and 10-20 controls are available. However, to identify the genetic variant in associated regions, fine-mapping and targeted re-sequencing is required. Here we present a new approach using whole-genome sequencing (WGS) of a family trio without prior GWAS. As a proof-of-concept, we chose an autosomal recessive disease known as hereditary footpad hyperkeratosis (HFH) in Kromfohrländer dogs. To our knowledge, this is the first time this family trio WGS-approach, has successfully been used to identify a genetic variant that perfectly segregates with a canine disorder.

The sequencing of three Kromfohrländer dogs from a family trio (an affected offspring and both its healthy parents) resulted in an average genome coverage of 9.2X per individual. After applying stringent filtering criteria for candidate causative coding variants, 527 single nucleotide variants (SNVs) and 15 indels were found to be homozygous in the affected offspring and heterozygous in the parents. Using the computer software packages ANNOVAR and SIFT to functionally annotate coding sequence differences and to predict their functional effect, resulted in seven candidate variants located in six different genes. Of these, only *FAM83G:*c155G>C (p.R52P) was found to be concordant in eight additional cases and 16 healthy Kromfohrländer dogs.

## Authors Summary

To identify genetic changes responsible for inherited diseases, the domestic dog is an attractive model. A successful approach to identify mutations associated with disease has been to perform case-control studies based on genome-wide association studies (GWAS) using large-scale SNP-array genotyping, followed by statistical association analysis. For autosomal recessive diseases caused by a single gene mutation, a sample collection of 10-20 affected dogs and equal number of healthy dogs is often needed. To recruit sufficient samples for rare diseases is difficult and time consuming and the associated region is typically very large, containing 0.5-1 million base pairs that needs to be re-sequenced for identification of disease-associated genetic variants. Here we used a different approach based on whole-genome sequencing of three dogs, a family trio where the offspring was diagnosed with Hereditary Footpad Hyperkeratosis (HFH) and the two healthy parents. HFH is a monogenic disease, and we presumed that both parents were heterozygous and that the affected offspring was homozygous for the associated mutation. A mutation in the affected individual was detected that result in a change of an amino acid in the *FAM83G* gene that was predicted to negatively influence function of the encoded protein. Perfect genetic concordance between homozygosity for this mutation and HFH was confirmed in a larger set of dogs affected by the disease.

## Introduction

In recent years, whole-genome sequencing (WGS) and whole-exome sequencing (WES) of family trios has emerged as a powerful approach to identify mutations

associated with inherited human diseases. This is the result of a rapid technological development of next generation sequencing (NGS) methods and subsequently a significant reduction in cost for sequencing an individual's genome. Both WGS and WES, generally refers to re-sequencing of an individual genome or exome and the reads are aligned to an appropriate reference genome sequence rather than being assembled *de novo*. An attractive aspect of NGS is the unbiased and large-scale detection of genetic variation, including single base pair substitutions, insertion/deletions and large structural variation resulting in inherited diseases and disorders. The majority of human studies have relied on WES and different sequence capture and enrichment techniques. In the first proof-of-concept study, WES was used to identify *MYH3* as the disease-causing gene in four unrelated individuals affected by a rare dominantly inherited disorder, the Freeman-Sheldon Syndrome (Ng et al. 2009). A large number of genes for human Mendelian diseases have now been identified using WES (for review see: Bamshad et al. 2011; Boycott et al. 2013; Chong et al. 2015; Shen et al. 2015). In contrast, WES methodologies have not been extensively used in canine Mendelian disease genetic research.

Since the publication of the canine reference genome in 2005 (Lindblad-Toh et al. 2005), the most common approach for identifying canine genetic risk factors has been to use SNP-based genotyping and genome-wide association studies (GWAS). The first proof-of-principle study used a two-tiered GWAS approach with two breeds where the haplotype associated with white coat colour was identified (Karlsson et al. 2007). Since then, case-control studies based on SNP genotyping and GWAS in a single breed has been used extensively and frequently resulted in the successful identification of genes associated with traits and diseases. For Mendelian traits, these

studies have typically involved SNP-based genotyping of 10-20 unrelated cases and equal number of healthy controls. With the extensive degree of linkage disequilibrium (LD) generally observed within dog breeds, these studies often result in an associated region spanning between 0.5-1 Mb that requires fine-mapping with additional SNPs and targeted resequencing to identify a candidate mutation (for review see: Karlsson et al. 2007; Karlsson and Lindblad-Toh 2008; Parker et al. 2010)). The recent update of the canine reference genome (CanFam3.1) and the development of more dense SNP-arrays such as the 170K CanineHD BeadChip (Illumina), have further increased the efficacy of this experimental design (Hoeppner et al. 2014; Vaysse et al. 2011). It is however not unusual that the associated region spans several megabases (e.g. Downs et al. 2014; Vernau et al. 2013).

In recent years, several canine disease-associated mutations have been identified with a combination of GWAS and WGS (Frischknecht et al. 2013; Gerber et al. 2015; Jagannathan et al. 2013; Kyostila et al. 2015; Owczarek-Lipska et al. 2013; Wolf et al. 2015). With this approach, GWAS performed on only a handful of individuals is used for defining an associated region in the genome followed by WGS of one affected dog. The critical interval of the affected dog's genome sequence is then compared to non-affected dogs to find candidate mutations for the disease. In rare diseases where only a few affected and unaffected individuals are available from a small family pedigree, successful identification of candidate mutations has also been achieved by a combination of SNP-based linkage analysis and WGS (Kyostila et al. 2015; Steffen et al. 2015; Willet et al. 2015). The power of WGS in candidate gene studies is further illustrated by two studies where the genomes of single individuals affected by a neuronal ceroid lipofuscinosis were sequenced and the data were mined for candidate

mutations (Gilliam et al. 2015; Guo et al. 2014). This led to the successful identifications of a frame shift mutation in the *CLN5* gene in Golden retriever (Gilliam et al. 2015) and a nonsense mutation in the *CLN8* gene in a mixed breed dog with Australian shepherd ancestry (Guo et al. 2014).

In a situation where no obvious candidate genes are known, an alternative approach is to use family-based WGS for gene discovery of rare variants associated with Mendelian diseases and traits. In comparison to WES, analysis of the complete genome eliminates biases associated with capture-based enrichment technologies and allows for the detection of non-coding variants that may be associated with a trait. In addition, the WGS data will remain useful even when the annotation of the canine reference genome is further improved. The Ion Proton™ System (Thermo Fisher Scientific/Life Technologies) has previously been evaluated for WGS of dogs (Viluma et al. 2015) and in the present study we applied the technology to perform WGS of a family trio as a straightforward approach to discover genetic variants associated with an autosomal recessive disease termed Hereditary Footpad Hyperkeratosis (HFH) that is segregating in the Kromfohrländer breed. This allowed the identification of a missense variant in *FAM83G* (c155G>C, p.R52P) associated with HFH.

During the progress of our study, Drögemüller and colleagues independently identified the same missense variant using a combination of a two-tiered GWAS in two breeds (42 Kromfohrländer dogs, 13 cases and 29 controls and 31 Irish Terriers, 10 cases and 21 controls) and re-sequencing of a single affected Kromfohrländer individual (Drogemuller et al. 2014). Combined, the use of different canine cohorts and experimental designs strengthens the conclusion that the genetic variant in

*FAM83G* is associated with HFH. It further demonstrates the efficacy of the trio-based WGS approach and serves as a proof-of-concept for detecting genetic variants associated with development of canine monogenic diseases.

# Results

### Next Generation Sequencing

Sequence libraries were constructed from genomic DNA prepared from a family trio consisting of two healthy parents and their HFH-affected offspring. The sequence reactions from each library were loaded on two Ion PI™ chips, yielding an average of 158 million reads per dog of which 98% were aligned to the reference genome (CanFam3.1). The average read length was 138 bp and on average 21.8 Gb sequence was generated for each dog (Table S1). Alignment of the reads to the canine reference genome resulted in a genomic coverage of 9.2X per individual dog with around 96% of the genome and 91% of the exome covered with at least one read.
In total, 3 726 772 SNVs and 2 474 309 indels were detected using GATK v.2.7 Unified Genotyper tool (Boycott et al. 2013) of which 3 449 902 SNVs and 198 165 indels remained after standard hard filtering.

### Functional Annotation of Sequence Variants

Functional annotation of the detected sequence variants was performed. In total, 23 185 SNVs were found to be exonic and of those substitutions, 9 858 were missense, 184 were nonsense and 13 143 were silent (Figure 1). Four hundred and ninety-one

indels were located in exons. Next, a conditional filtering was applied based on a pattern of inheritance where both parents were heterozygous for a substitution and the affected offspring was homozygous. This resulted in the detection of 527 SNVs that fulfilled this conditional inheritance pattern. Of those, 217 were found to be missense, 2 nonsense and 308 were silent substitutions. We identified fifteen indels that followed the conditional inheritance pattern and all but one was predicted to result in a frameshift of the open reading frame. One hundred and sixty-five missense substitutions, both nonsense substitutions and all 14 indels were found to be known common variants, thus we were left with 52 missense substitutions for functional prediction analysis using SIFT (Sim et al. 2012).

The results from the SIFT analysis suggested that 17 out of the 52 missense substitutions were deleterious (SIFT score <0.05). Single missense substitutions were found in 11 genes and three genes had two missense substitutions resulting in a total of fourteen candidate genes to be further evaluated for their potential association with HFH. Using IGV and manual inspection, nine of the 17 missense substitutions were eliminated due to genotype errors or insufficient coverage. The remaining eight non-synonymous substitutions were located in the following functional genes: *TDRD6, USP4, TACC2, DLGAP2, GRAPL* (two substitutions), *FAM83G* and *PDILT*. These were further considered as candidate causative coding variants associated with development of HFH.

**Validation of Candidate Causative Coding Variants**

To validate the NGS data from the family trio, the same three individuals were genotyped using Sanger sequencing of the eight candidate causative coding variants (Figure 2A, Table 1). All variants except the missense substitution present in the *TACC2* gene was confirmed to be heterozygous in the parents and homozygous in the affected offspring. Thus, leaving six candidate genes with potential causative coding variants to be further investigated.

| **Gene:** | ***GRAPL*** | ***GRAPL*** | ***FAM83G*** | ***PDILT*** | ***DLGAP2*** | ***TDRD6*** | ***USP4*** |
|---|---|---|---|---|---|---|---|
| **Position** | chr5: 41014230 | chr5: 41014231 | chr5: 41055619 | chr6: 24951743 | chr37: 30832376 | chr12: 14853811 | chr20: 39951727 |
| **Affected Offspring** | T/T | T/T | C/C | T/T | T/T | T/T | T/T |
| **Healthy Sire** | G/T | C/T | C/G | G/T | C/T | C/T | G/T |
| **Healthy Dam** | G/T | C/T | C/G | G/T | C/T | C/T | G/T |

**Table 1:** *Validation of the genotypes obtained by WGS of the family trio. Genotyping was made by Sanger sequencing of the seven candidate causative variants located in six different genes: GRAPL, FAM83G, PDILT, DLGAP2, TDRD6, and USP4. The genes and their respective chromosomal positions are indicated above. Genotypes obtained from the affected offspring, the healthy sire and the healthy dam are indicated.*

To further explore the concordance of the predicted deleterious variants in *GRAPL*, *FAM83G*, *PDILT*, *DLGAP2*, *TDRD6*, and *USP4*, an additional eight HFH-affected and 16 healthy Kromfohrländer dogs were genotyped using Sanger sequencing. All dogs affected by HFH were found to be homozygous for the variants in the *GRAPL* and *FAM83G* genes. In contrast, the missense substitutions found in the other four

genes were non-concordant in affected and healthy dogs (Table 2) and could consequently be excluded as being causative for HFH.

| Differences in: | GRAPL | GRAPL | FAM83G | PDILT | DLGAP2 | TDRD6 | USP4 |
|---|---|---|---|---|---|---|---|
| Position | chr5:41014230 | chr5:41014231 | chr5:41055619 | chr6:24951743 | chr37:30832376 | chr12:14853811 | chr20:39951727 |
| Case 1 | T/T | T/T | C/C | G/T | C/T | C/C | T/T |
| Case 2 | T/T | T/T | C/C | G/T | C/C | C/C | G/G |
| Case 3 | T/T | T/T | C/C | G/T | C/T | C/T | G/G |
| Case 4 | T/T | T/T | C/C | T/T | C/C | T/T | G/G |
| Case 5 | T/T | T/T | C/C | G/T | C/T | C/T | T/T |
| Case 6 | T/T | T/T | C/C | G/G | C/T | T/T | T/T |
| Case 7 | T/T | T/T | C/C | G/G | C/T | C/T | G/G |
| Case 8 | T/T | T/T | C/C | G/G | C/T | C/T | G/T |
| Control 1 | G/G | C/C | G/G | G/T | T/T | C/T | T/T |
| Control 2 | G/G | C/C | G/G | G/G | T/T | C/T | G/G |
| Control 3 | G/G | C/C | G/G | G/T | T/T | C/T | G/G |
| Control 4 | G/G | C/C | G/G | G/G | C/T | C/C | G/G |
| Control 5 | G/T | C/T | C/G | G/T | C/T | C/T | G/G |
| Control 6 | G/T | C/T | G/G | G/T | C/T | C/T | G/T |
| Control 7 | G/G | C/C | G/G | G/G | C/T | C/C | G/G |
| Control 8 | G/G | C/C | G/G | G/G | C/T | C/C | G/G |
| Control 9 | G/T | C/T | C/G | G/G | T/T | C/C | G/T |
| Control 10 | G/G | C/C | G/G | G/G | C/T | C/C | G/T |
| Control 11 | G/T | C/T | C/G | G/T | C/C | C/T | G/G |
| Control 12 | G/G | C/C | G/G | G/T | C/T | C/T | G/G |
| Control 13 | G/G | C/C | G/G | G/T | C/C | C/T | G/G |
| Control 14 | G/G | C/C | G/G | G/G | C/T | C/C | G/G |
| Control 15 | G/G | C/C | G/G | G/T | C/T | C/T | G/G |
| Control 16 | G/G | C/C | G/G | G/T | C/C | C/T | G/T |
| CanFam3 | G/G | C/C | G/G | G/G | C/C | C/C | G/G |

**Table 2**: *Genotyping by Sanger sequencing of the seven candidate causative variants located in six different genes to validate the concordance among cases. The GRAPL, FAM83G, PDILT, DLGAP2, TDRD6, and USP4 genes and their respective chromosomal positions are indicated above. The eight additional affected individuals (case 1-8), the 16 healthy controls (control 1-16) and the genotype in the CanFam3.1 reference genome sequence are indicated to the left.*

**Description of Disease-Associated Variants**

The missense variant in the *FAM83G* gene (Ensembl gene ID: ENSCAFG00000018245), was a G to C transversion at position chr5:41,055,619 (CanFam3.1) that was predicted to cause a non-synonymous substitution in the first exon of *FAM83G.* This missense variant (c.155G>C; p. R52P) substitutes a positively charged arginine to a neutral proline (Figure 2B). In a comparison with available genome sequence data from 29 placental mammals (Lindblad-Toh et al. 2011), it was found that the arginine residue at p.52 was completely conserved. A prediction analysis using HOPE (Venselaar et al. 2010) suggested that the substitution is damaging to the FAM83G protein.

The affected dogs were also homozygous for two substitutions adjacent to each other at positions chr5:41,014,230-41,014,231 in the *GRAPL* gene (Ensembl gene ID: ENSCAFG00000029170). However, based on comparison to other species, the Ensembl annotation of exon 1 of the *GRAPL* gene appears to be incorrect. Only in dogs, this part of the gene is annotated as being protein coding. Furthermore, we analysed NGS sequence data from 49 additional dog genomes derived from dogs unaffected by HFH. The genotype distribution from this analysis revealed that nine dogs, one Kromfohrländer case and eight dogs from other breeds (Beagle, Entlebucher mountain dog, Eurasier, Siberian husky, Lagotto Romagnolo, and Sloughi) were homozygous for the mutant allele (TT/TT). In addition nine dogs were heterozygous (GC/TT). The majority (31 dogs) of the other dogs were homozygous for the GC variant (data not shown).

# Discussion

GWAS have for almost a decade been the main strategy for identifying disease-associated alleles in canine Mendelian disorders. In the present proof-of-concept study, we show that WGS of a canine family trio can successfully be used to identify a candidate mutation for HFH, which is an autosomal recessive disease in Kromfohrländer dogs.

Today, the goal of "the $1000 genome" is close to being achieved (Hayden 2014) and NGS has already transformed human genetic research (Kilpinen and Barrett 2013). In studies of human monogenic diseases in small pedigrees, candidate mutations have been identified with WGS or WES (e.g. Glazov et al. 2011; Harakalova et al. 2012; Ng et al. 2010; Ng et al. 2009; Roach et al. 2010). A striking example of a family trio-based WES approach was the identification of a non-synonymous mutation in the *ABCC9* gene that is causative of the Cantú syndrome, a dominant disorder characterized by cardiac defects, congenital hypertrichosis (abnormal hairiness) and distinctive facial features. The missense mutation was detected by exome sequencing of one affected child and both his unaffected parents (Harakalova et al. 2012). Similarly, applying WGS on a family quartet where both offspring were affected by two autosomal recessive disorders (Miller syndrome and primary ciliary dyskinesia) and the parents were unaffected resulted in the identification of four candidate genes for these two Mendelian disorders (Roach et al. 2010).

HFH also known as digital hyperkeratosis (DH) is an orthokeratotic palmoplantar hyperkeratosis with autosomal recessive inheritance, first described clinically in Irish

Terriers (Binder et al. 2000), and is also affecting the Kromfohrländer breed. Here we show that a missense variant in *FAM83G* (c155G>C, p.R52P) is associated with HFH in Kromfohrländer dogs. The *FAM83G* gene encodes a protein with only limited knowledge concerning its function. However, FAM83G was recently identified as a novel SMAD1 interactor that was shown to be a modulator for bone morphogenetic protein (BMP)-dependent signalling (Vogt et al. 2014). The amino acid residue arginine (R) at position 52 is highly conserved among 29 placental mammals (Lindblad-Toh et al. 2011) and the substitution p.R52P in FAM83G is likely to be damaging. However, to conclusively define the biological effect of this non-synonymous substitution functional experiments are needed. The identification of *FAM83G* (c155G>C, p.R52P) as a likely candidate for HFH is in agreement with results published by Drögemüller and colleagues based on a GWAS of 42 Kromfohrländer dogs (13 cases and 29 controls) and 31 Irish Terriers (10 cases and 21 controls) followed by deep WGS of one affected dog (Drogemuller et al. 2014). In that study, a 611 kb haplotype was identified that was shared by all the affected dogs. WGS of one affected Kromfohrländer dog that was compared to 46 genomes from healthy dogs of different breeds revealed the causative allele at the *FAM83G* locus.

In the present study we undertook an alternative approach to identify the genetic variant associated with HFH in a population of Swedish Kromfohrländer dogs. Without prior GWAS and directly using WGS of a single family trio (an affected offspring and both its healthy parents) we confirmed that the *FAM83G*:c155G>C (p.R52P) variant is associated with HFH. To our knowledge, this is the first time WGS on a canine family trio has successfully been used to identify a disease-associated allele.

## Conclusions

We conclude that the missense variant *FAM83G*:c155G>C (p.R52P) is associated with HFH in Kromfohrländer dogs. Our result independently confirms a previous report using GWAS of a case-control population followed by WGS of a single affected dog and show that WGS of a family trio with two healthy carrier parents and one affected offspring is an efficient strategy for the identification of rare alleles associated with canine Mendelian disorders. This approach circumvents the challenge of sampling sufficient number of cases and controls needed to achieve statistical power in GWAS.

## Materials and Methods

### Canine Subjects

Kromfohrländer dogs were recruited via the breed association and breeders of Kromfohrländer dogs. A boarded veterinary dermatologist clinically examined the dogs. Pedigrees from all dogs were collected. Dogs with a history of hyperkeratosis affecting all four paw pads since juvenility and with clinical lesions compatible of the disease, with hyperkeratotic, firm and cracking pads were diagnosed as cases, whereas dogs with clinically normal paw pads were included as healthy controls. Biopsies, when taken from cases, revealed histopathological changes typical for paw pad hyperkeratosis.

Whole blood was collected from 9 cases and 18 healthy Kromfohrländer dogs into EDTA tubes. Genomic DNA was extracted from peripheral blood leukocytes, using 1 ml blood on a QIAsymphony SP instrument using the QIAsymphony DSP DNA Kit (Qiagen, Hilden, Germany).

### Ion Proton Sequencing

One µg of gDNA was fragmented using the Covaris S2 instrument (Covaris inc.) and library preparation was performed using the Ion Xpress™ Plus Fragment Library Kit for AB Library Builder™ System (Thermo Fisher Scientific/Life Technologies) followed by 5 cycles of amplification. Emulsion PCR was done on the Ion OneTouch™ 2 system with Ion PI™ Template OT2 200 Kit v2 chemistry (Thermo Fisher Scientific/Life Technologies). Enrichment was conducted using the Ion OneTouch™ ES (Thermo Fisher Scientific/Life Technologies). Samples were loaded on two Ion PI™ chip Kit v2 and sequenced on the Ion Proton™ System using Ion PI™ Sequencing 200 Kit v2 chemistry (200 bp read length, Thermo Fisher Scientific/Life Technologies).

### Next-Generation Sequence Analysis

Reads were aligned to the canine reference sequence assembly CanFam3.1 (Hoeppner et al. 2014) using TMAP, an implementation of BWA, SSAHA and Super-maximal Exact Matching (Li and Durbin 2010, 2009; Ning et al. 2001; Li 2012) included in the TorrentSuit 3.6 (Thermo Fisher Scientific/Life Technologies) software with default settings. Following best practice (GATK forum) guidelines, for each raw binary alignment file duplicated reads were detected and removed using the software Picard

v.1.69 tool MarkDuplicates Picard v.1.69 tool [http://picard.sourceforge.net](Van der Auwera et al. 2013). We further realigned reads in regions of potential insertions or deletions (indels) with the GATK tool (McKenna et al. 2010) Indel Realignment and performed Base Quality Recalibration with the covariates read group, quality score, context size (3bp) and cycle. Publically available genetic variation (SNPs and indels) from the Ensembl variation database (CanFam3.1, dog release 75) was used as "true positives" in base quality score recalibration and variant calling Ensembl Variation Release 77 (Canis lupus familiaris) [ftp://ftp.ensembl.org/pub/release-77/variation/vcf/canis_familiaris/] .

SNVs and indels were called simultaneously on the entire family trio using SAMtools (Li et al. 2009) and the GATK (McKenna et al. 2010) tool UnifiedGenotyper. Raw variant calls were filtered with VariantFiltration tool, using the standard hard filtering parameters (Van der Auwera et al. 2013): with (Mapping quality) MQ <40 and (Quality by depth) QD < 2.0 for both SNVs and indels, (Read Position Rank Sum test) ReadPosRankSum < -8.0 for SNVs and ReadPosRankSum < -20.0 for indels, (Fisher Strand) FS >60.0 for SNVs and FS >200.0 for indels.

**Detection of candidate mutations**

The SNVs and indels were classified using gene predictions from Ensembl build version 75 with ANNOVAR (Wang et al. 2010). A custom Perl script was used to extract variants with the conditional filtering using the following criteria: autosomal recessive pattern of inheritance *i.e*. parents heterozygous and the affected offspring homozygous for the variant allele. We used publicly available genetic variation in dogs (Axelsson et al. 2013) Ensembl Variation Release 77 (Canis lupus familiaris)

[ftp://ftp.ensembl.org/pub/release-77/variation/vcf/canis_familiaris/] and our custom SNV data set derived from genome sequences of other dog breeds to extract novel/unknown candidates (or - eliminate already known variants). To evaluate identified missense and nonsense mutations SIFT (Sim et al. 2012)was used. The coverage over identified positions was then manually inspected using IGV (Robinson et al. 2011; Thorvaldsdottir et al. 2013).

**Sanger Sequencing**

Amplification primers for seven missense mutations predicted as deleterious by SIFT were designed using the online version (http://biotools.umassmed.edu/bioapps/primer3_www.cgi) of Primer3 (Rozen and Skaletsky 2000). The primers were constructed with M13 forward or reverse tails (Table S3). Amplification and sequencing was done using the BigDye® Direct Cycle Sequencing Kit (Thermo Fisher Scientific/Life Technologies) on a ProFlex™ PCR System (Applied Biosystems) and GeneAmp® PCR System 9700 (Applied Biosystems). Four ng of genomic DNA was used in a 10 µl amplification reaction with an annealing temperature of 60 °C in accordance with the manufacturers' protocol. Cycle Sequencing was performed with both the forward (5′ -TGTAAAACGACGGCCAGT- 3′) and reverse (5′ -CAGGAAACAGCTATGACC-3′) sequencing primer on an ABI 3500XL DNA Analyser (Applied Biosystems). The sequences were then analysed using CodonCode Aligner v5.0.2 (CodonCode Corporation).

**Conservation in Mammals**

For each of the candidate mutations, multiple sequence alignments at DNA and protein level were performed. For DNA level we extracted human alignments for 100 vertebrates available at the Santa Cruz Genome Browser (UCSC) for each of the corresponding dog positions. For protein alignments, we first extracted the protein sequences from the HomoloGene NCBI site for each of the gene containing the candidate substitution. Next, we used standalone version of MUSCLE (Edgar 2004) software using the default parameters for multiple sequence alignments.

## Data availability

The data sets (three BAM files and one VCF file) supporting the results of this article are available in the European Nucleotide Archive (ENA) repository, [study accession number: PRJEB12301, http://www.ebi.ac.uk/ena/data/view/PRJEB12301

## Authors' contributions

GA, KB and TB conceived and designed the study. KB and EB collected field material and KB diagnosed the subjects. AV, SS, GA, TB performed experiments and analysed the data. VJ and TL contributed genome sequence data of control dogs. All authors approved the final manuscript.

## Acknowledgements

The authors would like to acknowledge support of the National Genomics Infrastructure (NGI) / Uppsala Genome Center and UPPMAX for providing assistance in massive parallel sequencing and computational infrastructure. Work performed at NGI / Uppsala Genome Center has been funded by RFI/VR and Science for Life Laboratory, Sweden. This study was funded by generous support by FORMAS, The Swedish Kennel Club and AGRIA. The authors would also like to thank the veterinary clinicians diagnosing and collecting samples. We would also like to acknowledge the support of dedicated dog owners whom allowed their dogs to take part in this study.

# Figures

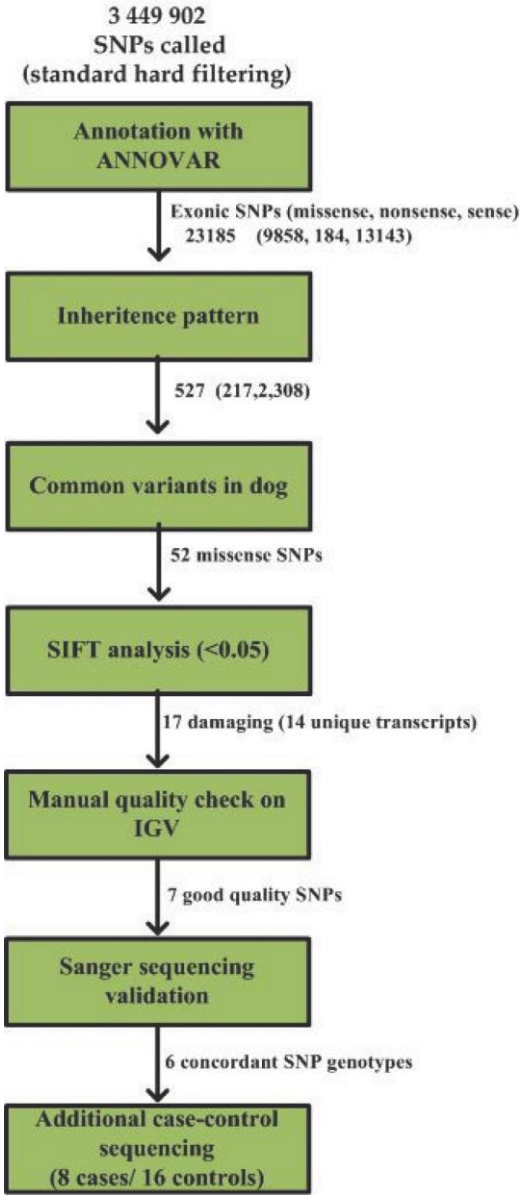


**Figure 1.** Schematic representation of the filtering pipeline used to evaluate genetic variants

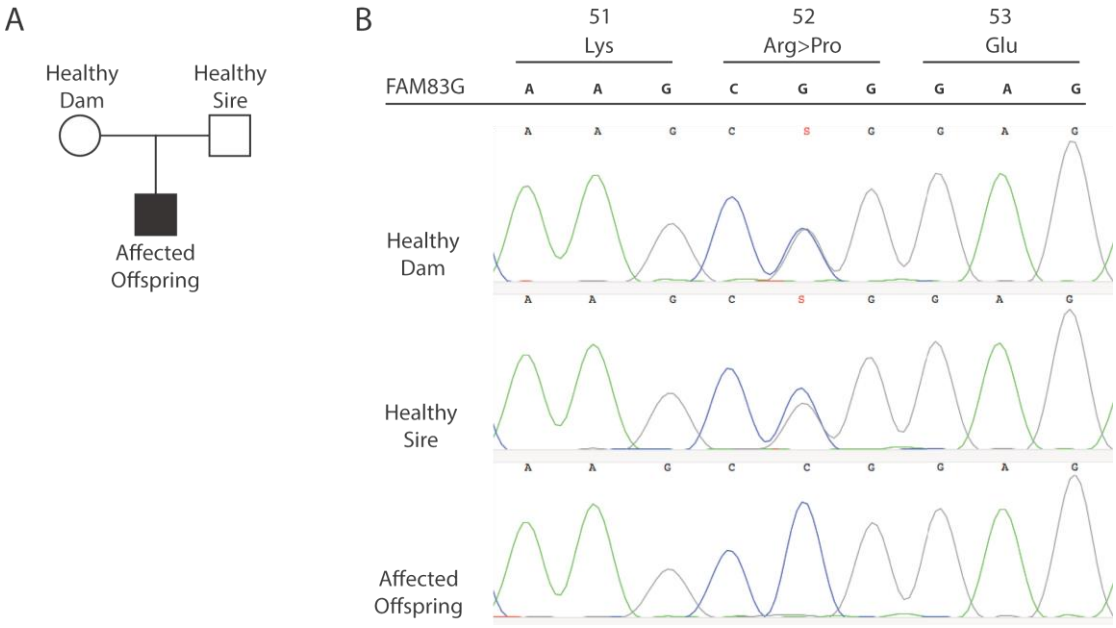


**Figure 2.** The family trio used for WGS and the validation of the missense variant at *FAM83G* (c155G>C, p.R52P) by Sanger sequencing. **A.** The genome of Kromfohrländer dogs from a family trio consisting of an offspring affected by HFH and both its healthy parents was sequenced. **B**. Both parents were heterozygous G/C (S) at position 155 of the *FAM83G* gene and the affected offspring was homozygous C/C resulting in a change at amino acid position 52 from an arginine to proline. The nucleotide sequence of codons 51-53 from the canine genome reference sequence is shown at the top.

# Supporting Information

**Table S1**: Summary alignment statistics.

Proportion of aligned reads and average depth for each chip and each individual is given.

**Table S2**: Primers used for Sanger sequencing.

Primers for Sanger sequencing of seven candidate damaging missense mutations. For both forward and reverse primer, the sequence is given in 5’ to 3’ direction.

# References


**Axelsson, E., A. Ratnakumar, M.L. Arendt, K. Maqbool, M.T. Webster *et al.*, 2013 The genomic signature of dog domestication reveals adaptation to a starch-rich diet. *Nature* 495 (7441):360-364.**

**Bamshad, M.J., S.B. Ng, A.W. Bigham, H.K. Tabor, M.J. Emond *et al.*, 2011 Exome sequencing as a tool for Mendelian disease gene discovery. *Nat Rev Genet* 12 (11):745-755.**

**Binder, H., S. Arnold, C. Schelling, M. Suter, and P. Wild, 2000 Palmoplantar hyperkeratosis in Irish terriers: evidence of autosomal recessive inheritance. *J Small Anim Pract* 41 (2):52-55.**

**Boycott, K.M., M.R. Vanstone, D.E. Bulman, and A.E. MacKenzie, 2013 Rare-disease genetics in the era of next-generation sequencing: discovery to translation. *Nat Rev Genet* 14 (10):681-691.**

**Chong, J.X., K.J. Buckingham, S.N. Jhangiani, C. Boehm, N. Sobreira *et al.*, 2015 The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. *Am J Hum Genet* 97 (2):199-215.**

**Downs, L.M., B. Wallin-Hakansson, T. Bergstrom, and C.S. Mellersh, 2014 A novel mutation in TTC8 is associated with progressive retinal atrophy in the golden retriever. *Canine Genet Epidemiol* 1:4.**

**Drogemuller, M., V. Jagannathan, D. Becker, C. Drogemuller, C. Schelling *et al.*, 2014 A mutation in the FAM83G gene in dogs with hereditary footpad hyperkeratosis (HFH). *PLoS Genet* 10 (5):e1004370.**

**Edgar, R.C., 2004 MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32 (5):1792-1797.**

**Frischknecht, M., H. Niehof-Oellers, V. Jagannathan, M. Owczarek-Lipska, C. Drogemuller *et al.*, 2013 A COL11A2 mutation in Labrador retrievers with mild disproportionate dwarfism. *PLoS One* 8 (3):e60149.**

**Gerber, M., A. Fischer, V. Jagannathan, M. Drogemuller, C. Drogemuller *et al.*, 2015 A deletion in the VLDLR gene in Eurasier dogs with cerebellar hypoplasia resembling a Dandy-Walker-like malformation (DWLM). *PLoS One* 10 (2):e0108917.**

**Gilliam, D., A. Kolicheski, G.S. Johnson, T. Mhlanga-Mutangadura, J.F. Taylor *et al.*, 2015 Golden Retriever dogs with neuronal ceroid lipofuscinosis**

**have a two-base-pair deletion and frameshift in CLN5. *Mol Genet Metab* 115 (2-3):101-109.**

**Glazov, E.A., A. Zankl, M. Donskoi, T.J. Kenna, G.P. Thomas *et al.*, 2011 Whole-exome re-sequencing in a family quartet identifies POP1 mutations as the cause of a novel skeletal dysplasia. *PLoS Genet* 7 (3):e1002027.**

**Guo, J., G.S. Johnson, H.A. Brown, M.L. Provencher, R.C. da Costa *et al.*, 2014 A CLN8 nonsense mutation in the whole genome sequence of a mixed breed dog with neuronal ceroid lipofuscinosis and Australian Shepherd ancestry. *Mol Genet Metab* 112 (4):302-309.**

**Harakalova, M., J.J. van Harssel, P.A. Terhal, S. van Lieshout, K. Duran *et al.*, 2012 Dominant missense mutations in ABCC9 cause Cantu syndrome. *Nat Genet* 44 (7):793-796.**

**Hayden, E.C., 2014 Technology: The $1,000 genome. *Nature* 507 (7492):294-295.**

**Hoeppner, M.P., A. Lundquist, M. Pirun, J.R. Meadows, N. Zamani *et al.*, 2014 An improved canine genome and a comprehensive catalogue of coding genes and non-coding transcripts. *PLoS One* 9 (3):e91172.**

**Jagannathan, V., J. Bannoehr, P. Plattet, R. Hauswirth, C. Drogemuller *et al.*, 2013 A mutation in the SUV39H2 gene in Labrador Retrievers with hereditary nasal parakeratosis (HNPK) provides insights into the epigenetics of keratinocyte differentiation. *PLoS Genet* 9 (10):e1003848.**

**Karlsson, E.K., I. Baranowska, C.M. Wade, N.H. Salmon Hillbertz, M.C. Zody *et al.*, 2007 Efficient mapping of mendelian traits in dogs through genome-wide association. *Nat Genet* 39 (11):1321-1328.**

**Karlsson, E.K., and K. Lindblad-Toh, 2008 Leader of the pack: gene mapping in dogs and other model organisms. *Nat Rev Genet* 9 (9):713-725.**

**Kilpinen, H., and J.C. Barrett, 2013 How next-generation sequencing is transforming complex disease genetics. *Trends Genet* 29 (1):23-30.**

**Kyostila, K., P. Syrja, V. Jagannathan, G. Chandrasekar, T.S. Jokinen *et al.*, 2015 A missense change in the ATG4D gene links aberrant autophagy to**

**a neurodegenerative vacuolar storage disease. *PLoS Genet* 11 (4):e1005169.**

**Li, H., 2012 Exploring single-sample SNP and INDEL calling with whole-genome de novo assembly. *Bioinformatics* 28 (14):1838-1844.**

**Li, H., and R. Durbin, 2009 Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25 (14):1754-1760.**

**Li, H., and R. Durbin, 2010 Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 26 (5):589-595.**

**Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan *et al.*, 2009 The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25 (16):2078-2079.**

**Lindblad-Toh, K., M. Garber, O. Zuk, M.F. Lin, B.J. Parker *et al.*, 2011 A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* 478 (7370):476-482.**

**Lindblad-Toh, K., C.M. Wade, T.S. Mikkelsen, E.K. Karlsson, D.B. Jaffe *et al.*, 2005 Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* 438 (7069):803-819.**

**McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis *et al.*, 2010 The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 20 (9):1297-1303.**

**Ng, S.B., K.J. Buckingham, C. Lee, A.W. Bigham, H.K. Tabor *et al.*, 2010 Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet* 42 (1):30-35.**

**Ng, S.B., E.H. Turner, P.D. Robertson, S.D. Flygare, A.W. Bigham *et al.*, 2009 Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* 461 (7261):272-276.**

**Ning, Z., A.J. Cox, and J.C. Mullikin, 2001 SSAHA: a fast search method for large DNA databases. *Genome Res* 11 (10):1725-1729.**

**Owczarek-Lipska, M., V. Jagannathan, C. Drogemuller, S. Lutz, B. Glanemann *et al.*, 2013 A frameshift mutation in the cubilin gene (CUBN) in Border**

**Collies with Imerslund-Grasbeck syndrome (selective cobalamin malabsorption). *PLoS One* 8 (4):e61144.**

**Parker, H.G., A.L. Shearin, and E.A. Ostrander, 2010 Man's best friend becomes biology's best in show: genome analyses in the domestic dog. *Annu Rev Genet* 44:309-336.**

**Roach, J.C., G. Glusman, A.F. Smit, C.D. Huff, R. Hubley *et al.*, 2010 Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* 328 (5978):636-639.**

**Robinson, J.T., H. Thorvaldsdottir, W. Winckler, M. Guttman, E.S. Lander *et al.*, 2011 Integrative genomics viewer. *Nat Biotechnol* 29 (1):24-26.**

**Rozen, S., and H. Skaletsky, 2000 Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol* 132:365-386.**

**Shen, T., A. Lee, C. Shen, and C.J. Lin, 2015 The long tail and rare disease research: the impact of next-generation sequencing for rare Mendelian disorders. *Genet Res (Camb)* 97:e15.**

**Sim, N.L., P. Kumar, J. Hu, S. Henikoff, G. Schneider *et al.*, 2012 SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res* 40 (Web Server issue):W452-457.**

**Steffen, F., T. Bilzer, J. Brands, L. Golini, V. Jagannathan *et al.*, 2015 A Nonsense Variant in COL6A1 in Landseer Dogs with Muscular Dystrophy. *G3 (Bethesda)*.**

**Thorvaldsdottir, H., J.T. Robinson, and J.P. Mesirov, 2013 Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* 14 (2):178-192.**

**Van der Auwera, G.A., M.O. Carneiro, C. Hartl, R. Poplin, G. Del Angel *et al.*, 2013 From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics* 11 (1110):11 10 11-11 10 33.**

**Vaysse, A., A. Ratnakumar, T. Derrien, E. Axelsson, G. Rosengren Pielberg *et al.*, 2011 Identification of genomic regions associated with phenotypic variation between dog breeds using selection mapping. *PLoS Genet* 7 (10):e1002316.**

**Venselaar, H., T.A. Te Beek, R.K. Kuipers, M.L. Hekkelman, and G. Vriend, 2010 Protein structure analysis of mutations causing inheritable diseases. An e-Science approach with life scientist friendly interfaces. *BMC Bioinformatics* 11:548.**

**Vernau, K.M., J.A. Runstadler, E.A. Brown, J.M. Cameron, H.J. Huson *et al.*, 2013 Genome-wide association analysis identifies a mutation in the**

**thiamine transporter 2 (SLC19A3) gene associated with Alaskan Husky encephalopathy. *PLoS One* 8 (3):e57195.**

**Viluma, A., S. Sayyab, S. Mikko, G. Andersson, and T.F. Bergstrom, 2015 Evaluation of whole-genome sequencing of four Chinese crested dogs for variant detection using the ion proton system. *Canine Genet Epidemiol* 2:16.**

**Vogt, J., K.S. Dingwell, L. Herhaus, R. Gourlay, T. Macartney *et al.*, 2014 Protein associated with SMAD1 (PAWS1/FAM83G) is a substrate for type I bone morphogenetic protein receptors and modulates bone morphogenetic protein signalling. *Open Biol* 4:130210.**

**Wang, K., M. Li, and H. Hakonarson, 2010 ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 38 (16):e164.**

**Willet, C.E., M. Makara, G. Reppas, G. Tsoukalas, R. Malik *et al.*, 2015 Canine disorder mirrors human disease: exonic deletion in HES7 causes autosomal recessive spondylocostal dysostosis in miniature Schnauzer dogs. *PLoS One* 10 (2):e0117055.**

**Wolf, Z.T., H.A. Brand, J.R. Shaffer, E.J. Leslie, B. Arzi *et al.*, 2015 Genome-wide association studies in dogs and humans identify ADAMTS20 as a risk variant for cleft lip and palate. *PLoS Genet* 11 (3):e1005059.**