De novo mutations in canine evolution and disease

Tracy Chew

BAnVetBioSci (Hons 1)

Faculty of Science

University of Sydney

Australia

A thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

2019

Acknowledgements

I would first like to thank my wonderful supervisors Prof. Claire Wade and Dr. Bianca Waud. I must also include Dr. Cali Willet, who was in the Wade laboratory when I started honours with Claire as an undergraduate. I am ever so grateful for your inspirational passion and knowledge in the fields of bioinformatics and animal genetics research. All three of you have taught me many valuable skills and provided me with opportunities to gain professional experience throughout my undergraduate and postgraduate candidatures. I am especially grateful for the confidence you had in me. I achieved many goals that were unimaginable to me with your ongoing support and encouragement.

I received a lot of intellectual, moral and emotional support from many postgraduate students within veterinary science. These people include past and present postgraduate students in the Wade laboratory, Jessica Gurr, Brandon Velie, Bobbie Cansdale, Mitchell O'Brien, Georgie Samaha and Diane van Rooy. I would also like to mention other veterinary science postgraduate students, Carol Lee, Annie Pan, Sally Mortlock, Theresa Li and Pamela Soh. I am so grateful to have gone through our PhD journeys together. Your bright minds, upbeat personalities and love for animals have made my journey so much more fun and memorable. I can't wait to see and hear about all your successes in the future. The people mentioned here are reflections of the broader veterinary science community here at the University of Sydney, which I am privileged to be a part of.

To our internal and external collaborators, including the University of Sydney Veterinary Teaching Hospital, Prof. Hannes Lohi and Maria Kaukonen, it would not have been possible to carry out the work in this thesis without your help and generosity in helping me collect and obtain samples. This of course includes all animals and their owners for providing these samples. I also can't forget to thank the University staff at the Sydney Informatics Hub and the ICT department, who work tirelessly to provide and service the facilities used to produce the work included in this thesis.

I would not have been able to pursue this degree without the financial support that I received as the recipient of the Australian Postgraduate Award and the Neil and Allie Lesue Scholarship. I only hope that I have honoured Neil and Allie Lesue and this country by contributing to excellent research within animal science.

Finally, I would like to thank my closest friends, biological family, whisky family and my partner Nick. Your emotional support and belief in me have kept me pursuing my goals. Thank you for all the laughter, board game nights and shenanigans, you have helped me keep my sanity. I really appreciate all of you for listening to me ramble on about genomics, especially if you didn't know what I was talking about. I hope I've taught you something about science.

I dedicate this thesis to my furry children Evie and the late Oreo, who passed away in 2014. Your love for life drew me to learn about your species.

Declaration

I declare that this thesis is my own work and has not been submitted in any form for another degree or diploma at any university of institution of tertiary education. Information derived from the published or unpublished work of others has been acknowledged in the text and reference lists that are provided within each chapter. The work contained in Chapter 5 – The Genetics of Severe Haemophilia A in the Australian Kelpie was initially analysed as part of the Honours component of my undergraduate degree. I have since completely reanalysed the data for this work using larger study cohorts, more advanced methodologies and I have completely rewritten the manuscript before including this work into this thesis.

Tracy Chew

7st January 2019

Table of Contents

List of Figures

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 4.1

Chapter 4.2

Chapter 5

List of Tables

Chapter 1

Chapter 2

Chapter 4.2

Chapter 5

List of Manuscripts and Conference proceedings

This thesis contains published manuscripts, manuscripts which are currently in submission for publication and research that was presented at a range of Faculty, national and international conferences, as listed below.

2013 Poster: Developing a genetic test for Haemophilia A in Australian Kelpies. *Genetics Society of AustralAsia Conference*, Sydney, Australia, 14th-17th July.

Oral presentation: Whole Genome Sequences and Detection of *De Novo* Mutations in Parent to Offspring Trios of The Domestic Dog. *Faculty of Veterinary Science Annual Postgraduate Conference,* Camden, Australia, 6th-7th November.

2014 Poster: Detection of *De Novo* Mutations in Parent To Offspring Trios in Whole Genome Sequences of the Domestic Dog. *Genetics Society of AustralAsia Conference,* Sydney, Australia, 6th-9th July.

Oral presentation & poster: *De Novo* Mutations in Dogs. *Faculty of Veterinary Science Annual Postgraduate Conference,* Sydney, Australia, 5th-6th November.

2015 Poster: An Evolutionarily New, Deleterious Mutation Causes Bardet Biedl Syndrome in the Hungarian Puli. *Lorne Genome Conference,* Lorne, Australia, 15th-17th February.

Oral presentation: An Evolutionarily Recent, Deleterious Mutation Causes Bardet Biedl Syndrome in the Hungarian Puli. *The 8th International Conference on Advances in Canine and Feline Genomics and Inherited Diseases.* Cambridge, England. 22nd-26th June.

Oral presentation: Discovery of a Deleterious Mutation in the Hungarian Puli That Causes Disease Similar to Bardet Biedl Syndrome in Humans. *Boden Conference*, Adelaide, Australia, 8th-10th July.

Oral presentation: Bardet Biedl Syndrome in Hungarian Puli. *Faculty of Veterinary Science Annual Postgraduate Conference,* Camden, 28th-29th October.

2016 Oral presentation & poster: Understanding How Canine Disease and Evolution Transpires Through the Analysis of *De Novo* Mutations. *Faculty of Veterinary Science Annual Postgraduate Conference,* Sydney, Australia, 9-10th November.

2017 Published manuscript: Chew, T., B. Haase, C.E. Willet, and C.M. Wade, 2017 Exclusion of known progressive retinal atrophy genes for blindness in the Hungarian Puli. Anim. Genet. DOI: 10/1111/age.12553

Published manuscript: Chew, T., B. Haase, R. Bathgate, C. E. Willet, M. K. Kaukonen *et al.*, 2017 A Coding Variant in the Gene Bardet-Biedl Syndrome 4 (BBS4) Is Associated with a Novel Form of Canine Progressive Retinal Atrophy. G3 (Bethesda). 4: g3.117.043109.

2018 In submission: Chew, T., B. Haase, C.M. Wade, 2018 Direct estimate of the *de novo* mutation rate in the domestic dog.

In submission: Chew , T., C. E. Willet, B. Haase, C. M. Wade, 2018 A performance comparison of popular single nucleotide variant detection methodologies applied to low coverage whole genome sequencing data.

Authorship Attribution Statement

Chapter 4.1 of this thesis is published as Chew, T., B. Haase, C.E. Willet, and C.M. Wade, 2017 Exclusion of known progressive retinal atrophy genes for blindness in the Hungarian Puli. Anim. Genet. DOI: 10/1111/age.12553.

I co-designed this study with Prof. Claire Wade and performed the experiments under her supervision. I performed the analysis, wrote the manuscript and developed the arguments that were included in the draft manuscript. Critical revisions were made by myself, Dr. Bianca Haase, Dr. Cali Willet and Prof. Claire Wade.

Chapter 4.2 of this thesis is published as Chew, T., B. Haase, R. Bathgate, C. E. Willet, M. K. Kaukonen *et al.*, 2017 A Coding Variant in the Gene Bardet-Biedl Syndrome 4 (BBS4) Is Associated with a Novel Form of Canine Progressive Retinal Atrophy. G3 (Bethesda). 4: g3.117.043109.

I co-designed this study with Prof. Claire Wade and Assoc. Prof. Roslyn Bathgate. The bioinformatics and sequencing analysis was carried out by myself under the supervision of Prof. Claire Wade. The fertility analysis was jointly performed by myself and Prof. Roslyn Bathgate. I developed the ideas and arguments for the manuscript and wrote the drafts. Critical revisions to the manuscript were made by Prof. Claire Wade, Dr. Bianca Haase, Assoc. Prof. Roslyn Bathgate, Dr. Cali E. Willet and Prof. Hannes Lohi.

Abbreviations

ANKC	Australian National Kennel Council
bp	Base pair
BWA	Burrows-Wheeler Aligner
^{o}C	Degrees Celsius
cDNA	Complementary DNA
CDS	Coding exonic sequence
CHISQ	Chi-squared
CpG	5' – Cytosine – phosphate – Guanine – 3'
CNV	Copy number variant
dbSNP	Database of single nucleotide polymorphisms
BBS	Bardet-Biedl Sydnrome
BED	Browser extendable format
DNA	Deoxyribonucleic acid
DSBs	Double stranded breaks
EDTA	Ethylenediaminetetraacetic acid
EXO	Exonulease
FTA	Flinders Technology Associates
GATK	Genome Analysis Tool-Kit
Gb	Gigabase
GC	Guanine-cytosine
GWAS	Genome wide association study
G_1 phase	Gap 1 phase
G_2 phase	Gap 2 phase
HapMap	Haplotype Map
IBD	Identity by descent
ID	Identifier
Indel	Insertion-deletion
Kb	Kilobase
Labrador	Labrador Retriever

LD	Linkage disequilibrium
M phase	Mitotic phase
Mb	Megabase
NCBI	National Centre for Biotechnology Information
NGS	Next generation sequencing
OMIA	Online Mendelian Inheritance in Animals
OMIM	Online Mendelian Inheritance in Man
PacBio	Pacific Biosciences
PCR	Polymerase chain reaction
Pol	Polymerase
PRA	Progressive retinal atrophy
RP	Retinitis pigmentosa
S phase	Synthesis phase
SAM	Sequence Alignment Map
Sec	seconds
SIFT	Sorting Intolerant From Tolerant
SMRT	Single molecule real time
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variation
Ti	Transition
Tv	Transversion
U	Enzyme unit
UCSC	University of California, Santa Cruz
USCF	University of Sydney, *Canis Familiaris*
UTR	Untranslated region
UV	Ultraviolet
VEP	Variant Effect Predictor
WGS	Whole genome sequencing
X	Fold of sequence coverage
yr	year
µL	Microlitres

Abstract

The domestic dog is an evolutionarily unique animal and has a special niche within genomics research. Since their domestication from the grey wolf, dogs have become one of the most phenotypically diverse living land animals. Man's desire to create individuals with specialised morphological and behavioural traits has led to the development of over 400 recognised breeds. Dogs share a significant number of inherited disease phenotypes with humans and are regarded as valuable animal models for understanding evolution and disease. New mutations are the ultimate source of new phenotypic diversity and evolutionary change. They can also cause rare spontaneous genetic disorders and collectively, they make a significant contribution to disease burden in managed populations. To comprehensively understand the mechanisms of evolution and disease, discovering the rates of occurrence, type, and patterns of distribution of *de novo* mutations across the genome is essential. Until recently, the characteristics of *de novo* mutations could be inferred only using indirect or biased methods. With recent technological advancements, it is now possible to directly observe *de novo* mutations that occur in a single generation directly through parent-offspring sequencing studies. Whole genome sequencing provides the opportunity for genomic variants associated with rare diseases caused by spontaneous mutations to be identified directly. We are on the brink of the capacity to utilize these technologies more fully in the field of personal medicine. In this thesis, *de novo* germline mutations affecting the evolution and occurrence of disease in the dog are identified and characterised. The inspiration for this work stemmed from the extraordinary phenotypic diversity in the species and its close relationship to people.

Chapter 1. Literature Review

1.1. Introduction

All genetic variation that drives evolution or contributes to disease once arose from a new DNA mutation. Characterising the rate of mutation and types of mutations that occur helps us to understand the mechanisms of evolutionary processes and disease. The identification of *de novo* variants and the methods for doing so have several practical applications. Patients with rare diseases could potentially achieve a rapid genetic diagnosis. Currently in Australia, an estimated 7% of rare disease patients do not receive a diagnosis at all and 30% received a delayed diagnosis of five years or more (Molster *et al.* 2016; Zurynski *et al.* 2017). Incorrect or delayed diagnosis can also lead to the administration of inappropriate and potentially harmful treatments, as well as incur additional emotional and financial burdens to affected families (Zurynski *et al.* 2017). Delays in obtaining genetic diagnoses for rare diseases in animals has strong potential for negative economic and ecological impacts, especially for species with a short optimal breeding age and short lifespans. Rapid genetic diagnoses are required for the quick development of accurate genetic tests. Mutation rates also have an application in research. With ancestral DNA sequence, the mutation rate can be used as a molecular clock to estimate the timing of species divergence. This was previously heavily debated in dog domestication research (Axelsson *et al.* 2013; Wang, Zhai, *et al.* 2013; Callaway 2013; Freedman *et al.* 2014). Mutation rates are also commonly used as a prior probability to obtain more accurate calling of *de novo* variants in many variant calling algorithms (Francioli *et al.* 2017).

With the recent advancements and increased accessibility in obtaining next generation sequencing (NGS) data, direct observation of new mutations is now possible through parent-offspring sequencing. Sequencing of whole genomes using NGS technologies is superior to traditional methods of *de novo* variant characterisation. Before NGS existed, new mutation events could only be indirectly inferred or observed in small proportions of large vertebrate genomes. We begin the first chapter of this thesis with a review of the current understanding of how *de novo* mutations are formed, their impacts on fitness,

methods of their detection and what is currently known about *de novo* mutation activity in animal species. Our purpose is to elucidate how *de novo* mutations impact evolution and diversity within a single species. As our interest is in mutations that persist in a species, this work primarily concentrates on germline mutations.

1.2. Causes of *de novo* mutations

Many new mutations are caused by the imperfect process of the division and proliferation of cells. The process of growth and differentiation of cells, commonly referred to as the cell cycle, involves four coordinated phases in eukaryotes: the gap 1 (G_1); synthesis (S); gap 2 (G_2); and the mitotic (M) phase (Figure 1.1). G_1, S and G_2 are collectively known as interphase and occur 95% for the duration of the cell cycle. In the G_1 phase, the cell is metabolically active. Cytosolic contents and organelles grow and replicate. In the S phase, DNA replication occurs. In G_2, the cell checks for errors in DNA replication that may have occurred in the previous phase and attempts to repair errors that are detected. The cell will continue to grow and synthesize proteins that are required in the next mitotic phase. The mitotic phase, consisting of four sub-phases (prophase, metaphase, anaphase and telophase), followed by cytokinesis, is where the cell proceeds to divide to form two daughter cells (Cooper 2000). The process described here describes the growth and replication of somatic cells. The growth and replication of germline cells is slightly different, as daughter cells are required to contain half the number of chromosomes (n) as somatic cells (2n). Gametes will undergo the prophase, metaphase, anaphase, telophase and cell division rounds twice (meiosis I and meiosis II), to produce four daughter cells. Importantly, parental chromosomes within gametes undergo homologous recombination during the first prophase.

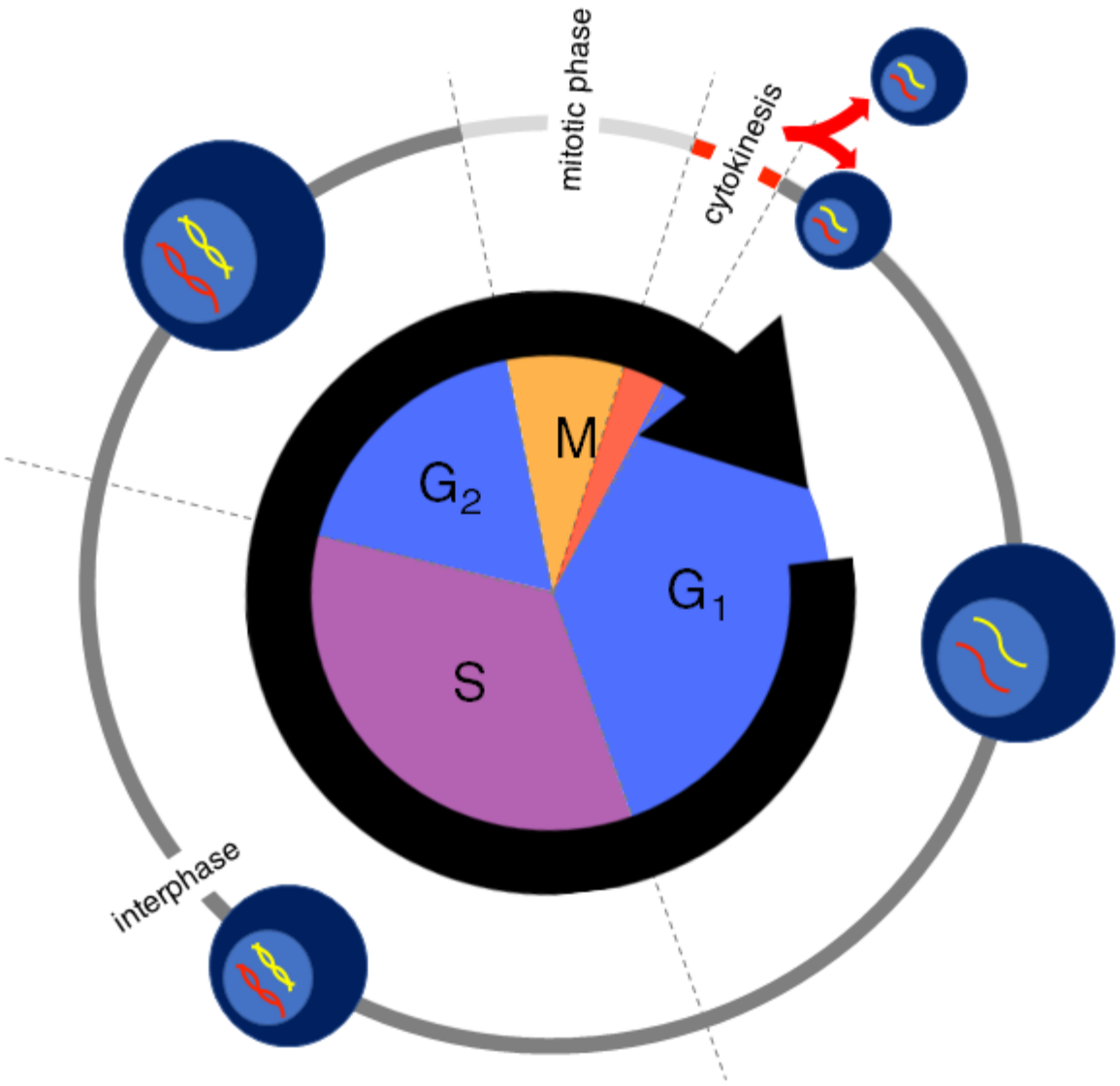

Figure 1.1. The eukaryotic cell cycle that embodies the growth and division of cells.

The cell cycle starts at the G_1 phase. DNA replication occurs in the S phase. The S phase is followed by more cellular growth. In the next G_2 phase, the cell checks for errors in that could have occurred during DNA replication. The cell then proceeds to divide in the M phase and two daughter cells are created following cytokinesis at the end of the cell cycle. Author's own artwork.

Mutations arise from errors that occur during DNA replication during the S phase of the cell cycle (Figure 1.2). In eukaryotes, the predominant DNA polymerases ε and δ replicate the leading and lagging DNA strand respectively and with high fidelity (Korona et al. 2011). The reported error rates for the replication process range between one mistake per 104 base-pairs (bp) to one per 105 bp in vitro. However the rate at which mutations which are permanently incorporated into daughter cells is much lower because most of the errors that occur are recognised and corrected by proofreading exonucleases present within DNA polymerases ε and δ (Kunkel 2009; Korona et al. 2011; Acuna-Hidalgo et al. 2016). Other accessory proteins such as the single strand binding protein also enhance the accuracy of DNA replication by DNA polymerases (Yang 2000). Errors which are missed by proofreading subunits can be corrected during the mismatch repair pathway. Mismatch repair pathway proteins (including MLH1, MLH3, MSH2, MSH3, MSH6, PMS1 and PMS2 in humans) excise the DNA containing the incorrectly incorporated nucleotide that is recognised by the proteins' ability to distinguish the newly synthesized daughter strand from the parental strand (Preston *et al.* 2010; Seshagiri 2013). Mismatch repair proteins are highly conserved across prokaryotes and eukaryotes (Yang 2000). DNA polymerase and DNA ligase then replace and seal in the correct nucleotides to the newly replicated strand. Cells that contain mutations that are not repaired will undergo DNA damage induced apoptosis if the mutation is lethal, or will be sustained in the daughter cell and its subsequent descendant cells (Preston *et al.* 2010).

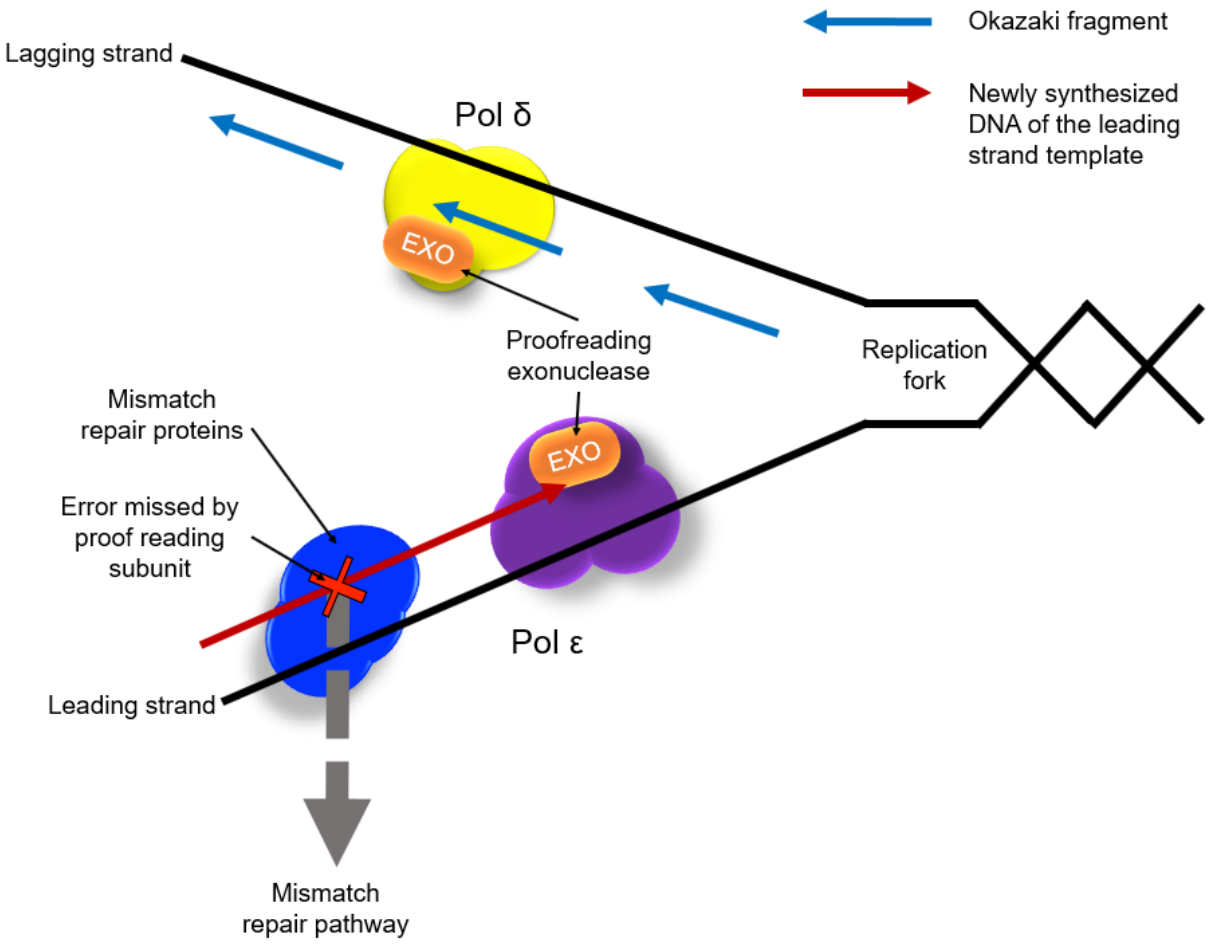

Figure 1.2. DNA replication that occurs during the S phase of the cell cycle. DNA polymerases (Pol) ε and δ replicate the leading and lagging strand respectively with an error rate of one per 104 to 105 nucleotides in eukaryotes. Both polymerases contain proofreading exonucleases (EXO) which ensure that an identical nucleotide to the leading or lagging template strand is incorporated into the newly synthesized strand. Errors which bypass the proofreading subunit can be correct in the mismatch repair pathway (this can also occur on the replication of the lagging DNA strand but is not represented in the figure). Author's own artwork.

Errors in DNA replication have the potential to generate different types of mutations. The relative frequencies of their occurrences and the relative ease of their repair influences the observable error rate for each type of mutation. The incorrect incorporation of a single nucleotide base leads to nucleotide substitutions. Among these, transitions (purine-purine involving A and G nucleotides; or pyrimidine-pyrimidine involving C and T nucleotides) occur the more frequently than transversions (purine-pyrimidine, or vice versa) in all species studied to date (Gojobori *et al.* 1982; Hershberg and Petrov 2010; Smeds *et al.* 2016). Each of the four nucleotides can obtain spontaneous, reversible rearrangements of their molecular bonds. Such rearrangements create a new form of the original nucleotide (termed a tautomer). Transition and transversion mutations can arise through tautomeric shifts. A tautomeric nucleotide sometimes pairs with a different nucleotide than the standard nucleotide that the originating nucleotide bonds with. For instance, the standard A (amino) form pairs with T, but its non-standard imino form A' pairs with the C nucleotide. If this error is not corrected during DNA replication, a transition mutation in the newly synthesized DNA strand results (Griffiths *et al.* 2000).

In the newly synthesized strand, changes are observed at C or G nucleotides more frequently than alterations at A and T bases, especially within the hyper-mutable, methylated cytosine base regions in CpG dinucleotide islands (Cooper and Youssoufian 1988). The reasons for the increased mutability are not yet clear, but it is postulated that the ease or difficulty of separation of the paired nucleotides contributes to easier repair. C and G nucleotides have a strong three hydrogen bond connection, making dissociation and repair more difficult in GC rich regions (Ségurel *et al.* 2014). A and T nucleotides are more easily separated with only two hydrogen bonds connecting these nucleotides, allowing easier repair in AT rich regions. Other causes of single nucleotide or multinucleotide substitutions include incomplete repair of the newly replicated DNA strand (Acuna-Hidalgo *et al.* 2016).

DNA polymerases can add or fail to incorporate occasional nucleotides due to misalignments to the template strand, causing small insertion-deletion (indel) errors. Larger indels involving double stranded breaks (DSBs) in DNA and larger chromosomal

segments of >1,000 bp in size are often termed structural variants (Scherer *et al.* 2007) and are most often caused by homologous recombination, non-allelic homologous recombination and replication-based mechanisms (Gu *et al.* 2008; Yang *et al.* 2013). Depending on the type of mutation, structural variants can be further classified as a copy number variant (CNV), inversion, translocation or segmental duplication. DSBs can be repaired by either the homologous recombination or non-homologous end joining pathways (Lieber 2010).

Genetic context is a major determinant of frequency of mutation. Regions of low complexity such as minisatellite and microsatellite regions have higher mutation rates than complex regions of unique DNA (Baer *et al.* 2007). Homopolymer regions are similarly hyper-mutable as they are prone to replication slippage. Replication slippage occurs when there is a misalignment in the template and newly synthesized DNA, causing expansion or contraction of the homopolymer. In eukaryotes, single nucleotide changes occur more frequently in DNA that is in close proximity to indel mutations or recombination sites (Lercher and Hurst 2002; Tian *et al.* 2008; Duret and Arndt 2008).

At the M phase of the cell cycle, aneuploidies can occur when chromosomes do not correctly segregate into their respective daughter cells (Figure 1.1). The accuracy of chromosomal segregation is dependent on the structural integrity of spindle microtubules and their ability to adequately attach onto the chromosome through a structure called the kinetochore (Compton 2011). Aneuploidies often have severe effects on cell survival and apoptosis of the affected daughter cells is usually initiated. Occasionally some cells survive and go on to have profound phenotypic effects. Aneuploidies are frequently recognised in cancerous cells and other diseases which will be reviewed later in this chapter.

Spontaneous DNA mutations caused by factors independent of the cell cycle, especially those involving DSBs, can be caused by mutagens of endogenous (retrotransposons, oxygen-free radicals, by products of metabolism) or exogenous (viruses, UV radiation, DNA-reactive chemicals commonly found in tobacco products) sources. The main mechanisms to repair these DNA lesions are through the base excision repair and

nucleotide excision repair systems (Lindahl 1999). If spontaneously caused DNA mutations are not repaired before the next round of DNA replication, they become permanently fixed into newly created daughter cells.

1.3. Somatic and germline mutations

The timing of occurrence, type and location of the cell containing a *de novo* mutation influences the possible effect that the mutation has on the individual or the population. Germline mutations occur in the gametes of an individual's parents and are therefore heritable and exposed to evolutionary processes. Somatic mutations occur in all other cells of the body and are accumulated post-fertilization throughout an individual's life (Figure 1.3). Somatic mutations are self-limiting as they are not heritable, but can have profound effects on an individual if the mutation occurs early in development (e.g. postzygotically), or if it induces oncogenesis (Li *et al.* 2014). More proliferative cell types such as those in the intestinal epithelial tissue are expected to harbour more new mutations than cells that are less proliferative, such as cells of the heart tissue (Shendure and Akey 2015). Researchers have observed that the somatic mutation rate is almost twice as high as that of the germline mutation rate for humans and mice (Milholland *et al.* 2017). The differential mutation rate highlights the importance of preserving the genome in the germ cells and that DNA repair in the soma is much less efficient.

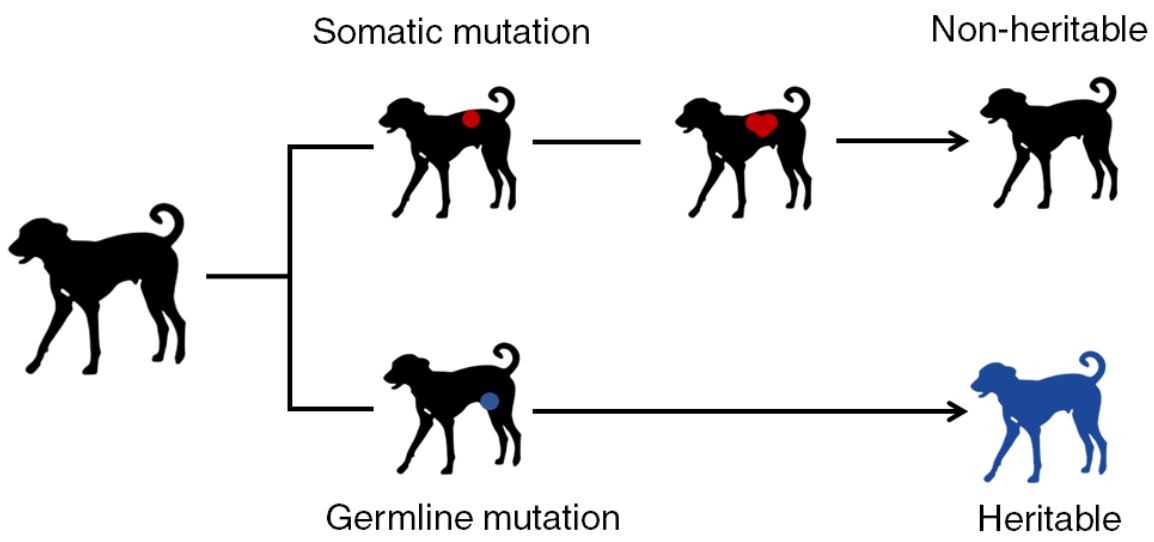

Figure 1.3. Somatic and germline mutations.
Somatic mutations are not heritable, unlike germline mutations which can be transmitted to some or all progeny. Somatic mutations can only occur in somatic tissue. Proliferation of a cell containing a somatic mutation leads to a population of cells containing the mutation that occurred in the original cell. Germline mutations occur in cells that are destined to become a sex cell (i.e. sperm or egg). Gametes containing a specific germline mutation that are fertilized result in progeny with the mutation present in all cells of their body. Author's own artwork.

1.4. *De novo* mutation detection methods

As new mutations are relatively rare especially in eukaryotes, accurately identifying and characterising *de novo* events in the whole genome has remained a challenging task (Kondrashov and Kondrashov 2010). Our ability to detect *de novo* mutations is limited by our ability to observe the DNA sequences for whole genomes, especially in eukaryotic genomes with chromosomes that are megabases in total size. With the achievement of several technological advancements in DNA sequencing, estimates have become more accurate over time and error profiles of different types of mutations has been observed at a much higher resolution than previously possible.

1.4.1. Traditional *de novo* germline mutation rate detection methods

The first few estimates of *de novo* germline mutation rates were made before DNA sequencing was even possible. Large chromosomal abnormalities, in particular aneuploidies, were easily detectable under the microscope in the early days of cytogenetic research. The path to detecting single nucleotide variation (SNV), small indels and sub-microscopic *de novo* variants such as CNVs required the development of more sophisticated methodologies and technologies. We will regard as traditional *de novo* mutation detection methods as those methods that were developed before high throughput sequencing technologies existed (current methods are later outlined in section 1.4.2).

The first methods for estimating mutation rates were based on observations of spontaneously occurring phenotypes that were caused by *de novo* mutations in functional coding DNA, such as Mendelian diseases in people or lethal mutations in laboratory animals (Danforth 1923; Haldane 1935; Keightley *et al.* 1998; Kondrashov 2002). Rates of incidence of the new phenotypes in the population were used to indirectly infer a mutation rate for that species.

Once the first DNA sequencing and amplification methods were developed, researchers could directly observe *de novo* mutations, i.e. Maxam-Gilbert and Sanger sequencing in the 1970s (Maxam and Gilbert 1977; Sanger *et al.* 1977). These sequencing and DNA amplification methods allowed scientists to obtain DNA sequences that were thousands of nucleotides in length, enabling the scientists to directly interrogate small genomes (e.g. some viruses), or small regions of larger genomes from other organisms (e.g. humans, dogs, mice).

A description of methods that have historically been used to estimate *de novo* mutation rates and their limitations are summarised in Table 1.1. Some major limitations are common to all traditional techniques. For example, given the extreme rarity of *de novo* mutation events in eukaryotes and lack of feasibility to interrogate multiple individual genomes (in humans, the current agreed estimates of the per nucleotide mutation rate is 1-3 x 10^{-8} per generation, which is equivalent to 30 – 90 nucleotides in the three

gigabase (Gb) human genome) (Conrad *et al.* 2011; Kong *et al.* 2012; Campbell and Eichler 2013), traditional methods could not provide a high resolution, accurate estimation and an investigation into the characteristics of new mutations in large eukaryotic genomes. In particular, it was not possible to observe or measure the rates: for each mutation type (single nucleotides, indels, CNVs, aneuploidies); across a variety of species of interest; in different genomic contexts; and in different physiological and environmental conditions, including in natural contexts.

Table 1.1. Methods for estimating *de novo* mutation rates in the pre-high throughput sequencing era and the potential associated biases

Measurement and methods taken to estimate rates	Potential biases and limitations	References
Incidence rates of spontaneously occurring phenotypes present in natural populations (e.g. spontaneous Mendelian diseases in people). With the Mendelian inheritance pattern, incidence, fitness effect, causal locus and its sequence length, an estimated mutation rate can be calculated. This relies on an assumption that variant is under a mutation-selection balance.	Not all mutations cause a phenotypic change leading to underestimates of the mutation rate. Deleterious mutations associated with disease may be present in mutational hotspots.	(Danforth 1923; Haldane 1935; Deng and Lynch 1996; Kondrashov 1998, 2002; Nachman 2004)
Using inbred populations to systematically measure spontaneously occurring phenotypes (e.g. mutations causing with lethal consequences)	Requires inbred lines with short generation times, up to a thousand generations are often required to make an observation. Therefore this method is not feasible in many large animals. As many generations may be required to obtain an observation of a new phenotype, this can be extremely laborious. Inbred lines may not represent true natural populations.	(Muller 1928; Keightley 1994)

Using inbred populations to directly identify *de novo* mutations that occur in a few loci by DNA sequencing or polymerase chain reaction	Only possible for sites with unusually high mutation rates (e.g. mitochondria, microsatellites), otherwise sequencing would become too expensive as large sample sizes would be necessary to make an observation.	(May *et al.* 1996; Denver *et al.* 2000)
Identifying polymorphisms at neutral sites (e.g. synonymous mutations, DNA sequencing of orthologous sequences between species). Site must be neutrally evolving so that it is proportional to the mutation rate. Timing of species divergence must be known.	Difficult to ascertain whether a site is truly neutral.	(Sueoka 1961; Kimura 1968; Kondrashov and Crow 1993; Drake *et al.* 1998; Nachman and Crowell 2000)
Applying artificial mutagens such as ethyl methanesulfonate (EMS) to introduce spontaneous mutations. EMS-induced mutations have been used to study the phenotypic effects and rate of true spontaneous mutations.	Not ethical or feasible in animals with longer generation times. Does not provide a true representation of naturally spontaneous mutations.	(Mukai 1970; Keightley *et al.* 1998)

1.4.2. High throughput sequencing technologies for detecting *de novo* germline mutations

Direct observation of *de novo* germline mutations in whole genomes would enable the accurate estimation of the *de novo* mutation rate and characterisation of their genome wide distribution for each mutation type. With multiple individual whole genomes from related individuals, many questions about *de novo* mutations can be answered. Before sequencing genomes from individuals of a family became a possibility, sequencing technologies had to become more affordable and higher in throughput and resources that complement these technologies had to be developed.

One of the most important resources in modern genomics research is the reference genome. The reference genome is a representation of a species' DNA, where nucleotides are organised linearly by physical position along the lengths of each chromosome. Researchers studying a variety of experimental questions can then use the reference genome to develop tools or describe DNA of interest in subsequent re-sequencing projects (e.g. physical positions of *de novo* mutations) in a consistent and reproducible manner. Annotations to the reference genome, including the physical position of various features such as genes, regulatory DNA and genomic context would enable a deeper understanding of how and why *de novo* mutations are formed. The first draft genome made available was the human genome in 2001 and was developed from a pool of four unique individuals (Lander *et al.* 2001; Venter *et al.* 2001). Sequencing was carried out using Sanger based technologies and the project had an estimated cost of up to $1 billion US dollars (USD). For the first time, researchers were able to characterise different features of the human genome, including its length, the number of genes and their organisation, GC content and relative rate of recombination across the genome (Lynch *et al.* 2016).

To enable utilization of the reference genome in subsequent re-sequencing projects, it was evident that major technological advancements had to be made to reduce the cost of whole genome sequencing to under $1,000 USD per individual. Reducing the cost of providing whole genome re-sequencing would make population-level studies, personalised medicine and research in other species possible (Schloss 2008; Reuter *et al.* 2015). The National Human Genome Research Institute initiated a $70 million USD scheme to make high throughput, NGS possible in the subsequent 10 years. The resources provided through this scheme resulted in the development of a variety of high throughput sequencing platforms (Reuter *et al.* 2015; Ambardar *et al.* 2016).

Since the human genome was sequenced, the genomes of other multicellular model organisms were sequenced in rapid succession including the mouse, rat, chimpanzee and the dog (Waterston *et al.* 2002; Gibbs *et al.* 2004; Mikkelsen *et al.* 2005; Lindblad-Toh *et al.* 2005). NGS platforms developed accelerated genomics research and today

there are 35,197 publicly available reference genomes, including 1,331 animal genomes (http://www.ncbi.nlm.nih.gov).

Popular NGS platforms are based on Sanger sequencing technology and involve four main wet laboratory steps, each with slight variations in chemistry depending on the sequencing platform (e.g. from 454, Illumina, Ion Torrent companies, developed between 2004-2010) (Ambardar *et al.* 2016; Mardis 2017). The four steps include nucleic acid isolation, library preparation, template amplification and sequencing by fluorescence detection (Ambardar *et al.* 2016). Creating DNA libraries involves random fragmentation of the DNA strand to create shorter DNA templates and the attachment of "adaptors" to template ends. The adaptors create stable priming sites for the ends of diverse DNA sequences and are key to enable PCR amplification and sequencing on the NGS platform (Timmerman 2015; Ambardar *et al.* 2016). Clusters of clonally PCR amplified DNA templates enhances the detectable fluorescent signal that is produced from the extension of one nucleotide during sequencing, as technologies are not yet sensitive enough to detect fluorescence using only one DNA molecule. Sequencing is carried out from either end of the DNA fragment, resulting in the production of single sequences (often termed 'reads'), or from both ends of the DNA fragment, resulting in paired-end reads on either side of a DNA fragment (the middle un-sequenced portion is commonly termed the 'insert sequence') (
Figure 1.4). Significantly, paired-end reads are on opposite strands facing the centre of the insert sequence, and library construction can filter DNA fragments so that paired reads are separated by a limited distance on the DNA which designates the "read-length". Depending on the sequencing platform, reads can vary in length (e.g. 36-300 bp, single or paired-end reads are available for Illumina platforms, 200-400 bp single reads for Ion Torrent platforms).

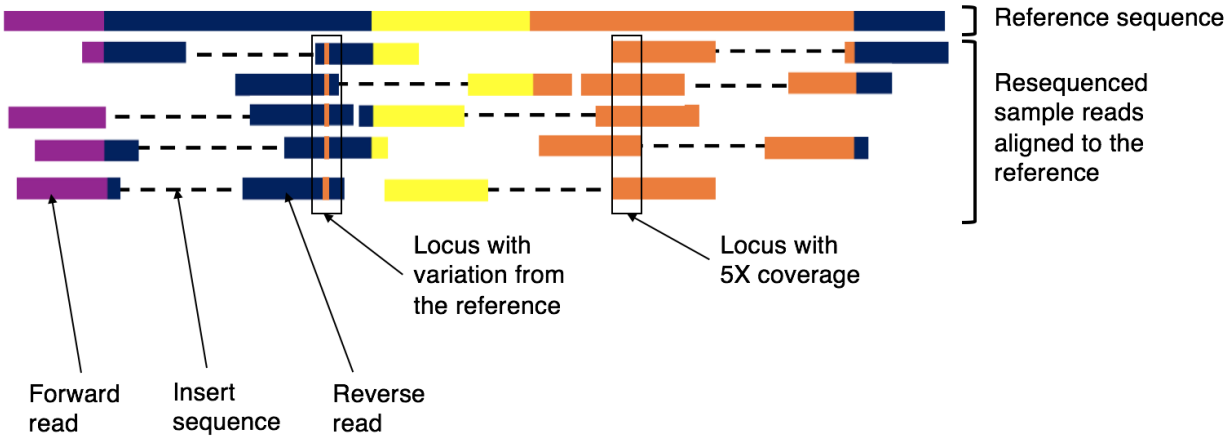

Figure 1.4. Representation of NGS data that has been aligned to a reference genome.

NGS data typically consists of short paired-end reads that are sequenced on opposite strands of the original DNA template. Paired-end reads typically contain a non-sequenced insert sequence. Coverage indicates the number of reads that have aligned to a locus in the reference genome. Variant callers detect loci with reads containing variation from the reference allele. Author's own artwork.

When individual genomes are sequenced using NGS technologies, the short reads generated must be processed using bioinformatics tools in order to achieve biologically relevant observations. First, reads must be arranged into their natural biological order. This is done by computationally aligning or 'mapping' reads to the reference genome (Figure 1.4). Commonly used mapping algorithms place reads into the most likely physical position in the reference genome, by comparing similarity between the read and all portions of the reference. Nucleotides in the optimally aligned sequences that differ relative to the reference genome can next be identified using variant calling programs.

To characterise germline *de novo* variants in eukaryotic species, studies most recently employ NGS in parent-offspring trio genomes or transcriptomes (Michaelson *et al.* 2012; Sayyab *et al.* 2016; Francioli *et al.* 2017). A “trio” consists of two parents and a progeny (often a disease “proband”). When a high-quality variant is detected at specific locus in the offspring that is not present in either parent, a deviation from Mendelian law suggests the presence of a *de novo* mutation (Goldmann *et al.* 2016; Wong *et al.* 2016) (Figure 1.5). *De novo* variant detection using this technique has been employed in humans, mice, chimpanzees and birds to date (Venn *et al.* 2014; Uchimura *et al.* 2015; Smeds *et al.* 2016).

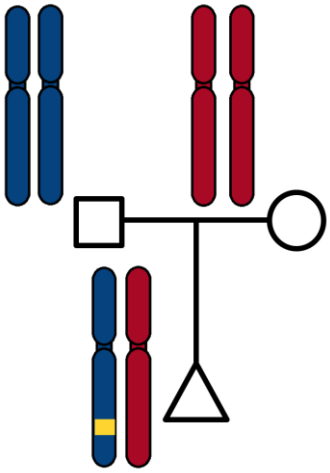

Figure 1.5. A parent-offspring trio pedigree and a representation of a germline de novo mutation.

One paternal chromosome (the father is represented on the pedigree as a square) shown in blue and one maternal chromosome (the mother is represented on the pedigree as a circle) shown in red is inherited by the offspring (represented on the pedigree as a triangle). The paternal chromosome in the offspring contains variation (shown in yellow) that deviates Mendelian inheritance laws, suggesting that this is a de novo mutation. Author’s own artwork.

Although NGS technologies have advanced genomic research drastically, sequencing, mapping and variant calling remain error prone and are limited to variant types that can be called accurately using common workflows. The sheer size and complexity of whole vertebrate genomes (~3 gigabases in humans, 48% of which are composed of repetitive sequences) means that such limitations are common (Mardis 2017). Each NGS platform is associated with its own specific propensity to particular error rates and profiles (Ambardar *et al.* 2016; Goodwin *et al.* 2016). For example, platforms that use PCR amplification are subject to PCR errors and the difficulty in re-sequencing GC rich DNA (Reuter *et al.* 2015). Read alignment to the reference genome is based on read nucleotide identity to the reference genome causing mapping bias towards reference alleles and underrepresentation of alternative alleles, especially in highly polymorphic regions such as in human leukocyte antigen and other immunity genes (Degner *et al.* 2009; Brandt *et al.* 2015). The process of alignment and variant calling assumes that the reference genome is a true representative of the studied species. Most reference genomes are "incomplete" with many containing gaps, un-localised contigs that have not been placed on their residing chromosome and technical artefacts (The Genome Reference Consortium, https://www.ncbi.nlm.nih.gov/grc/). Also, as the reference genome is developed from one or a relatively small number of unique individuals, any 'novel' DNA that is present in the subject individual but not the reference sequence may be missed due to the described mapping bias (Rosenfeld *et al.* 2012).

To alleviate issues associated with short read NGS platforms such as sequencing errors and alignment artefacts, genomes must be sequenced to a high level of redundancy (coverage of ~30X) to achieve the high specificity that is required to characterise *de novo* mutations (Francioli *et al.* 2017). Coverage of >30X can achieve relatively high sensitivity and accuracy of calling SNVs (Cheng *et al.* 2014). Because higher coverage is associated with higher sequencing costs, some laboratories opt for exome NGS to detect *de novo* variants, especially those associated with disease (Poultney *et al.* 2013; Francioli *et al.* 2017). Variant types other than SNVs, in particular longer indels and CNVs less than 100,000 bp, are more difficult to detect and genotype to a comparable degree of accuracy as genotyping SNVs using short read NGS. As a consequence, the

characteristics and rates of large *de novo* indels and CNVs have not been comprehensively studied in species other than humans (Ghoneim *et al.* 2014).

Depending upon insert sizes, paired-end sequencing on short read NGS platforms can improve the level at which indels and SNVs can be resolved. Long read sequencing platforms, which were commercially available from 2010-2014, have been designed to outperform short read NGS in accurately calling larger variants such as indels and CNVs. Long read platforms include the Oxford nanopore (minION) and single molecule real time (SMRT) sequencing by Pacific Biosciences (PacBio). The two dominating platforms produce 2,000 and 40,000 bp length reads respectively (Ambardar *et al.* 2016). Both platforms are PCR-free, with the advantage of less bias when sequencing GC rich content and all long read platforms enable improved mappability of reads due the increase in alignment confidence associated with read length (Reuter *et al.* 2015). With these benefits, long read platforms have been shown to identify 85% of novel indels and CNVs of ~500 bp that were not detected by other methods (Chaisson *et al.* 2014). Despite these benefits, long read platforms are associated with higher sequencing error rates (11 - 38.2%, predominantly composed of indel and homopolymer errors) than short read NGS platforms (0.11 – 0.28%) (Minoche *et al.* 2011). To benefit from both long and short read platforms, researchers have suggested combining both technologies in a single experiment (Weirather *et al.* 2017). However, as long read NGS are more expensive per base and are lower throughput than short read technologies, costs still limit their wide-scale use (Reuter *et al.* 2015; Ambardar *et al.* 2016).

1.4.3. Microarray based technologies for detecting *de novo* CNVs

Microarray based technologies have been successfully used to detect CNVs larger than 100,000 bp in length, especially those arising from *de novo* events (Carter 2007; Egan *et al.* 2007; Sebat *et al.* 2007; Lupski 2007; Itsara *et al.* 2010; Alvarez and Akey 2012; Elizabeth Locke *et al.* 2015). There are many types of microarray based technologies (for a review of each see (Carter 2007)) but each works using the same principle. Microarrays are developed to target multiple, evenly spread sites or markers across the lengths of reference chromosomes. Regions where individuals differ in DNA copy

number to the reference can be identified by the relative intensity of signal that is emitted from hybridized probes in the region of the variant. Because the technology relies on linkage disequilibrium for targets to represent surrounding loci, higher density microarrays such as those available for human and mice are able to provide a higher resolution of the genome. Reliability of these markers diminishes for CNVs less than 100,000 bp and for these variants NGS platforms are still preferred despite their limitations (Willet *et al.* 2013; Campbell and Eichler 2013; Poultney *et al.* 2013). Additionally, CNVs detected by microarrays cannot be physically placed without additional sequence interrogation such as through NGS.

1.4.4. Detection of somatic mutations

The *de novo* mutation detection methods that have been discussed so far relate to germline mutations and not somatic mutations, which are more difficult to identify. Somatic mutations are unique to a single cell and its descendant cells, unlike germline mutations which can be represented by all cells and tissue types in the body. Due to the relative rarity of each somatic mutation existing in an individual, obtaining high quantities of DNA to represent these mutations sufficiently is the biggest challenge in detecting somatic mutations. Many of the developed protocols used to identify somatic variants increase template number, either by careful sampling or through specialized library preparation methods.

One common purpose for somatic mutation detection in humans and dogs is in cancer studies to identify putative disease causing or risk variants (Watson *et al.* 2013; Gardner *et al.* 2016). In cancer studies, somatic mutations are typically identified by employing whole genome or whole exome sequencing on DNA obtained from tumour and normal patient-matched tissue samples (Lawrence *et al.* 2013; Watson *et al.* 2013; Alioto *et al.* 2015). A mutation is determined as somatic if it was not identified as a germline variant that was present in the cells of normal tissue. As cancers typically consist of many 'sub-clones', each with their own unique set of somatic mutations, paired-end, deep coverage (~100X) sequencing is required to detect somatic mutations from technical artefacts (Alioto *et al.* 2015; Hsu *et al.* 2017). Furthermore, as the technique of

sequencing tumour-normal pairs requires a population of tumour-affected cells, this limits its use in a clinical setting for the early detection of cancer or for non-tumourous cancers (e.g. blood cancers).

A method for the detection of somatic mutations in a cell without the need for descendant cells to increase DNA template number is single cell sequencing. Single cell sequencing was first conducted on mammalian cells for the whole genome (scDNA-seq) in 2011 and transcriptome (scRNA-seq) in 2009 (Tang *et al.* 2009; Navin and Hicks 2011). Single cell sequencing has since also been applied to metagenomics and epigenomics. The technology employs similar processes to standard 'bulk' sequencing, with some additional steps. Cells need to be isolated (e.g. through microfluidics), whole genomes are amplified to obtain enough starting template quantities and additional barcoding of DNA fragments is required in library preparation. Barcodes are later used to identify the sequence's cell of origin during downstream processing (Wang and Navin 2015). Although the technology has improved significantly in the last 10 years, technical errors may be introduced during the amplification step and such errors remain a major source of false positives in this technology (Wang and Navin 2015). Despite this, single cell sequencing technologies provide opportunity to study other biologically relevant somatic mutations other than cancer, such as neuronal mutations and somatic mutations associated with aging (Lodato *et al.* 2015; Enge *et al.* 2017). With further development, the technologies show potential for use in early clinical diagnosis of cancers caused by somatic mutations (Navin and Hicks 2011).

1.5. The effects of *de novo* mutations

All *de novo* mutations can be classed as having an advantageous, neutral or deleterious consequence to an individual's fitness. Fitness can be defined as the ability for an individual to survive and reproduce in the environment in which they reside in (Crow 2000). Identifying the distributions of fitness effects for germline mutations aids in better understanding the dynamics between the occurrence of new genetic variation and the fitness of a population (Keightley and Eyre-Walker 2007). In general, advantageous mutations are rare but over time, contribute to ongoing adaptive evolution and

speciation (Eyre-Walker and Keightley 2007; Keightley 2012). Mutations that are highly deleterious undergo purifying selection and do not persist in populations for long periods of time. This is particularly relevant for sporadically occurring CNVs and aneuploidies, which may affect larger portions of the genome (Acuna-Hidalgo *et al.* 2016). Mutations that are mildly advantageous or deleterious are under lower selective pressure and persist in populations longer. Particular attention has been paid by the research community to the accumulation of mildly deleterious alleles, which are thought to collectively contribute to common neurodevelopmental diseases such as intellectual disability, autism spectrum disorders and schizophrenia (Vissers *et al.* 2010; O'Roak *et al.* 2011; Veltman and Brunner 2012; Poultney *et al.* 2013).

1.5.1. New mutations and the evolution of canine phenotypes

The evolutionary process that allowed the rapid phenotypic evolution of the domestic dog from the grey wolf is of interest because of the amount of phenotypic diversity that has been developed in a relatively short period of time. The event of canine domestication presents a valuable model for understanding the process and relationship that influences gene variation and phenotypes as species evolve. Genetic and paleontological evidence suggests that canine domestication occurred ~15,000 - 33,000 years ago, however most of the 400 modern dog breeds were only developed in the last couple of centuries (Vilà *et al.* 1997; Savolainen *et al.* 2002; Germonpré *et al.* 2009a; Axelsson *et al.* 2013; Dreger *et al.* 2016). The phenotypic diversity is thought to be derived from genetic diversity that was already present in the wolf, however the rate and contribution from *de novo* mutations remains elusive (Wayne and Ostrander 1999). *De novo* mutations in *KIT* have been identified as a cause of white spotting in subpopulations of German Shepherd and spotted Weimaraner dogs (Gerding *et al.* 2013; Wong *et al.* 2013). The majority of other new mutations that have been reported contribute to diseases including ichthyosis, bleeding disorders and progressive retinal atrophy (Brooks 1999; Vilboux *et al.* 2008; Kropatsch *et al.* 2016; Bauer *et al.* 2017). *De novo* mutations that result in observable or measurable phenotypes such as coat colour or disease are easier to detect. However, as not all *de novo* mutations have a strong impact on visible phenotypes, many are not likely to be detected without NGS efforts.

1.5.2. *De novo* mutations and disease

Epidemiological studies have revealed sporadically occurring heritable diseases in both people and animal populations, with risk factors such as parental age increasing the likelihood of disease (Veltman and Brunner 2012). When there is no prior family history of a disorder expressed in a proband, researchers have recognised that causative genes are likely to be located in genomic regions that are more prone to mutations than others (these regions are often termed 'mutational hotspots') (Kong *et al.* 2012; Acuna-Hidalgo *et al.* 2016). Several disorders where new mutations are prevalent in their respective causative genes include Duchenne muscular dystrophy, haemophilia A and B, retinal atrophies and Huntington's disease (Haldane 1946; Myers *et al.* 1993; Grimm *et al.* 2012; Kropatsch *et al.* 2016). Various types of causative mutations have been identified at these loci, from simple single nucleotide mutations, CNVs, to deletions and inversions caused by non-allelic homologous recombination (Myers *et al.* 1993; Rossetti *et al.* 2011; Grimm *et al.* 2012).

Before the advent of whole genome sequencing, patients with the aforementioned diseases were unlikely to be diagnosed within their lifetime since the responsible *de novo* mutations are usually unique to an individual. If the effect on fitness is great, affected individuals are unlikely to propagate the mutation and this impacts the ability to conduct family or population-based mapping studies. Collectively, patients with spontaneous disease contribute importantly to overall disease prevalence. For example, of all reported Mendelian phenotypes (~5,129), ~32% have no reported underlying gene (OMIM, 2018). Such figures are roughly similar across domestic animal species including the dog (~23%), cat (~35%), bovine (~41%) and pig (~58%, OMIA, 2018). Many unmapped traits are believed to be caused by new mutations (Chong *et al.* 2015). Whole genome and exome trio sequencing studies are regarded as an effective method of diagnosing sporadic genetic disorders and are expected to become common in clinical practice in the foreseeable future (Zhu *et al.* 2015; Francioli *et al.* 2017; Cummings *et al.* 2017). Already, these techniques have been used to successfully map traits including common human diseases such as autism spectrum disorders and

schizophrenia and have been successfully used in animal disease studies (Sayyab *et al.* 2016; Chew, Haase, Bathgate, *et al.* 2017).

Aneuploidy is most frequently documented in humans as unassisted survival is severely impaired (Munné *et al.* 2004, 2016). Virtually all aneuploidies result from *de novo* events, as the effects of such mutations on fitness are so severe that individuals with these disorders are unable to reproduce. As with other CNVs, disease severity is impacted by a gene dosage effect caused by extra or missing chromosomes. The most common aneuploidy is trisomy 21 (Down syndrome), with a prevalence of one in 800 births (de Graaf *et al.* 2015). Other autosomal aneuploidies include trisomy 13 (Patau syndrome) and trisomy 18 (Edwards syndrome). Sex chromosome aneuploidies are also prevalent in human populations. They can be in the form of monosomy (Turner syndrome - X0) or trisomy (Jacob's syndrome -XYY; Klinefelter syndrome – XXY; and Triple X syndrome - XXX). Other forms of polysomies exist but are much more rare (other forms of Klinefelter syndrome - XXYY, XXXY, XXXXY; Tetrasomy X – XXXX and Penta X syndrome - XXXXX) (Visootsak and Graham 2006).

In addition to the diseases caused by germline mutations, somatic mutations that are acquired throughout an individual's life can become pathogenic and cause disease. Reported diseases include mutations that occur in the embryo (e.g. Proteus syndrome), or later in life (e.g. neurofibromatosis and McCune-Albright Syndrome) (Erickson 2003; Poduri et al. 2013). The most notorious and prevalent group of diseases caused by somatic mutations is cancer. Cancers can occur when disruptive mutations are acquired in proto-oncogenes, tumour suppressor genes or genes involved in DNA repair. These genes are responsible for normal cellular identity, differentiation and growth. When these normal cellular processes are disrupted, cells become abnormal and can have uncontrollable growth. The uncontrolled growth leads to formation of tumours, which is characteristic of many cancer-types (e.g. breast, lung, lymphoma) except for some blood cancers (e.g. leukemia). Disease can occur if the tumour is malignant and affect the ability of the organ or tissue it is residing in to function normally. In some cases, cells of the primary cancer can metastasize and form new tumours in other parts of the body.

Cancers are complex diseases and are genetically heterogeneous across individuals and even within the tumour cells of a single patient. Their heterogeneity stems from the stochastic nature and accumulation of somatic mutations. Most somatic mutations present in surviving cells are either neutral or mildly deleterious. The cells containing mildly deleterious mutations can clonally expand and harmful mutations can further accumulate in cancer driver genes. For this reason, age is a major risk factor for the development of many cancers (Risques and Kennedy 2018). Once cells become cancerous, tumours can develop and clonally expand into more aggressive forms. In the past decade, researchers have employed NGS technologies to determine the evolutionary trajectories of cancer to identify major genetic aberrations and the molecular interactions between cancer driving genes (Youn and Simon 2011; Krzywinski 2016; Peterson and Kovyrshina 2017). This knowledge can ultimately be used in a clinical setting such as use of identified predictive or prognostic biomarkers to enhance accuracy of diagnosis and effectiveness of personalised treatment plans.

1.6. Rates and distribution patterns of new mutations within and across species

Despite the challenges in identifying *de novo* mutations as previously described, it is evident that mutation rates vary across species, within species, within families and even across chromosomes of an individual (Ellegren *et al.* 2003; Conrad *et al.* 2011; Hodgkinson and Eyre-Walker 2011). In the current section, we describe characteristics of germline mutations across species only. Direct estimates of mammalian per base mutation rates fall around 10^{-8}, however rates in other species can vary at an order of 1,000 fold to this value (Lynch *et al.* 2016). Mutations are non-random and are influenced by different genomic contexts. Genome length and sequence constitution are unique to each species and this partially contributes to the differences observed in per species mutation rates. Variation in mutation rates also suggest that there are differences in the efficiency of DNA replication and repair across organisms (Lynch *et al.* 2016).

Although variation in rates exists, mutational patterns are shared among species. For instance, compared to non-GC rich contexts, mutations in CpG dinucleotide islands are

reported to occur 30 times more frequently in great apes, 15 times more in other mammals and 10 times more in birds (Keightley *et al.* 2011; Hodgkinson and Eyre-Walker 2011; Smeds *et al.* 2016). Apart from GC contexts, the mutation rate is also influenced by the adjacent nucleotides by two to threefold for reasons that are not completely understood (Hwang and Green 2004). Local disruptions to DNA such as recombination sites and spontaneously occurring indels can regionally influence mutation rates. In eukaryotes and some bacteria, SNP mutations are more frequent within ~50-300 bp of an indel (Tian *et al.* 2008; Zhu *et al.* 2009; Hollister *et al.* 2010). Similarly, recombination hotspots have been found to coincide with substitution mutation hotspots (Duret and Arndt 2008).

Parent of origin and age of conception has been identified as major factors that influence mutation rates. One of the first researchers to acknowledge gender differences was Haldane in 1946, who noted that the haemophilia gene was more mutagenic in men than in women (Haldane 1946). With modern technologies, Kong et al. 2012 later confirmed this, estimating that two additional mutations are transmitted to the offspring per year with increasing age of conception of the father (Kong *et al.* 2012). Whilst a similar trend is observed for mutations of maternal origin, the rate is much lower at 0.24 new mutations per additional year of the mother's age (Goldmann *et al.* 2016). This male bias is also observed in chimpanzees but at an even higher rate, with an estimated three mutations per year of the father's age (Venn *et al.* 2014). This mutational bias could reflect the reduction in the capability of DNA replication and repair during cell division as an individual ages. In females, oogenesis begins during foetal development where all a woman's primary oocytes are formed and arrested at prophase I. Further division is reinitiated at puberty when a woman begins her menstrual cycle and continues until she reaches menopause. On the other hand, the entire process of spermatogenesis in men starts at puberty and continually occurs until the death of the individual (Rahbari et al. 2016).

1.7. Aims of this thesis

In this thesis, we use modern techniques and technologies to directly observe *de novo* mutations in the dog. New mutations are very rare events; hence, methods used to identify them require high sensitivity and specificity. Unlike typical parent-offspring sequencing studies in humans which obtain sequencing depths of ~30X, we utilize sequencing datasets with lower average coverage (less than 15X). For this reason, we first compare popular SNP calling programs and pipelines to obtain the most suitable method applicable to datasets used. The results of this study enabled us to develop an optimised pipeline to obtain direct estimates of the per base mutation rate in the dog and categorise their distribution throughout the canine genome to enhance our understanding of canine evolution. Lastly, we studied two spontaneously occurring genetic diseases in the dog, aiming to map spontaneous deleterious mutations in two breeds and demonstrate that techniques used could be enforced for clinical diagnosis.

1.8. References

Acuna-Hidalgo, R., J. A. Veltman, and A. Hoischen, 2016 New insights into the generation and role of de novo mutations in health and disease. Genome Biol. 17: 1–19.

Alioto, T. S., I. Buchhalter, S. Derdak, B. Hutter, M. D. Eldridge et al., 2015 A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. Nat. Commun. 6: 10001.

Alvarez, C. E., and J. M. Akey, 2012 Copy number variation in the domestic dog. Mamm. Genome 23: 144–163.

Ambardar, S., R. Gupta, D. Trakroo, R. Lal, and J. Vakhlu, 2016 High Throughput Sequencing: An Overview of Sequencing Chemistry. Indian J. Microbiol. 56: 394–404.

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: 360–364.

Baer, C. F., M. M. Miyamoto, and D. R. Denver, 2007 Mutation rate variation in multicellular eukaryotes: Causes and consequences. Nat. Rev. Genet. 8: 619–631.

Bauer, A., D. P. Waluk, A. Galichet, K. Timm, V. Jagannathan et al., 2017 A de novo variant in the ASPRV1 gene in a dog with ichthyosis. PLoS Genet. 13: e1006651.

Brandt, D. Y. C., V. R. C. Aguiar, B. D. Bitarello, K. Nunes, J. Goudet et al., 2015 Mapping Bias Overestimates Reference Allele Frequencies at the HLA Genes in the 1000 Genomes Project Phase I Data. G3 (Bethesda). 5: 931–941.

Brooks, M., 1999 A review of canine inherited bleeding disorders: biochemical and molecular strategies for disease characterization and carrier detection. J. Hered. 90: 112–118.

Callaway, E., 2013 Dog genetics spur scientific spat. Nature 498: 282–283.

Campbell, C. D., and E. E. Eichler, 2013 Properties and rates of germline mutations in humans. Trends Genet. 29: 575–584.

Carter, N. P., 2007 Methods and strategies for analyzing copy number variation using DNA microarrays. Nat. Genet. 39: S16–S21.

Chaisson, M. J. P., J. Huddleston, M. Y. Dennis, P. H. Sudmant, M. Malig et al., 2014 Resolving the complexity of the human genome using single-molecule sequencing. Nature 517: 608-611.

Cheng, A., Y. Teo, and R. Ong, 2014 Assessing single nucleotide variant detection and genotype calling on whole-genome sequenced individuals. Bioinformatics 30: 1707–1713.

Chew, T., B. Haase, R. Bathgate, C. E. Willet, M. K. Kaukonen et al., 2017 A Coding Variant in the Gene Bardet-Biedl Syndrome 4 (BBS4) Is Associated with a Novel Form of Canine Progressive Retinal Atrophy. G3 (Bethesda). 7: 2327–2335.

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: 199–215.

Compton, D. A., 2011 Mechanisms of aneuploidy. Curr. Opin. Cell Biol. 23: 109–113.

Conrad, D. F., J. E. M. Keebler, M. A. DePristo, S. J. Lindsay, Y. Zhang et al., 2011 Variation in genome-wide mutation rates within and between human families. Nat. Genet. 43: 712–714.

Cooper, G., 2000 *The Cell: A Molecular Approach. 2nd edition*. Sinauer Associates, Sunderland, Massachusetts.

Cooper, D. N., and H. Youssoufian, 1988 The CpG dinucleotide and human genetic disease. Hum Genet 78: 151–155.

Crow, J. F., 2000 The origins, patterns and implications of human spontaneous mutation. Nat. Rev. Genet. 1: 40–47.

Cummings, B. B., J. L. Marshall, T. Tukiainen, M. Lek, S. Donkervoort et al., 2017 Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. Sci. Transl. Med. 9: 1-25.

Danforth, C., 1923 The frequency of mutation and the incidence of hereditary traits in man. Eugen. Genet. Fam. Sci. Pap. 2nd Int. Congr. Eugen. 1: 120–128.

Degner, J. F., J. C. Marioni, A. A. Pai, J. K. Pickrell, E. Nkadori et al., 2009 Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. Bioinformatics 25: 3207–3212.

Deng, H. W., and M. Lynch, 1996 Estimation of Deleterious-Mutation Parameters in Natural Populations. Genetics 144: 349–960.

Denver, D. R., K. Morris, M. Lynch, L. L. Vassilieva, and W. K. Thomas, 2000 High direct estimate of the mutation rate in the mitochondrial genome of Caenorhabditis elegans. Science 289: 2342–2344.

Drake, J. W., B. Charlesworth, D. Charlesworth, and J. F. Crow, 1998 Rates of spontaneous mutation. Genetics 148: 1667–1686.

Dreger, D. L., B. W. Davis, R. Cocco, S. Sechi, A. Di Cerbo et al., 2016 Commonalities in Development of Pure Breeds and Population Isolates Revealed in the Genome of the Sardinian Fonni's Dog. Genetics 204: 737–755.

Duret, L., and P. F. Arndt, 2008 The Impact of Recombination on Nucleotide Substitutions in the Human Genome. PLoS Genet. 4: e1000071.

Egan, C. M., S. Sridhar, M. Wigler, and I. M. Hall, 2007 Recurrent DNA copy number variation in the laboratory mouse. Nat. Genet. 39: 1384–1389.

Elizabeth Locke, M. O., M. Milojevic, S. T. Eitutis, N. Patel, A. E. Wishart et al., 2015 Genomic copy number variation in Mus musculus. BMC Genomics 16: 497.

Ellegren, H., N. G. Smith, and M. T. Webster, 2003 Mutation rate variation in the mammalian genome. Curr. Opin. Genet. Dev. 13: 562–568.

Enge, M., H. E. Arda, M. Mignardi, J. Beausang, R. Bottino et al., 2017 Single-Cell Analysis of Human Pancreas Reveals Transcriptional Signatures of Aging and Somatic Mutation Patterns. Cell 171: 321–330.

Erickson, R. P., 2003 Somatic gene mutation and human disease other than cancer. Mutat. Res. 543: 125–136.

Eyre-Walker, A., and P. D. Keightley, 2007 The distribution of fitness effects of new mutations. Nat. Rev. Genet. 8: 610–618.

Francioli, L. C., M. Cretu-Stancu, K. V Garimella, M. Fromer, W. P. Kloosterman et al., 2017 A framework for the detection of de novo mutations in family-based sequencing data. Eur. J. Hum. Genet. 25: 227–233.

Freedman, A. H., I. Gronau, R. M. Schweizer, D. Ortega-Del Vecchyo, E. Han et al., 2014 Genome Sequencing Highlights the Dynamic Early History of Dogs. PLoS Genet. 10: e1004016.

Gardner, H. L., J. M. Fenger, and C. A. London, 2016 Dogs as a Model for Cancer. Annu. Rev. Anim. Biosci. 4: 199–222.

Gerding, W. M., D. A. Akkad, and J. T. Epplen, 2013 Spotted Weimaraner dog due to de novo KIT mutation. Anim. Genet. 44: 605–606.

Germonpré, M., M. V. Sablin, R. E. Stevens, R. E. M. Hedges, M. Hofreiter et al., 2009 Fossil dogs and wolves from Palaeolithic sites in Belgium, the Ukraine and Russia: osteometry, ancient DNA and stable isotopes. J. Archaeol. Sci. 36: 473–490.

Ghoneim, D. H., J. R. Myers, E. Tuttle, and A. R. Paciorkowski, 2014 Comparison of insertion/deletion calling algorithms on human next-generation sequencing data. BMC Res. Notes 7: 864.

Gibbs, R. A., G. M. Weinstock, M. L. Metzker, D. M. Muzny, E. J. Sodergren et al., 2004 Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature 428: 493–521.

Gojobori, T., W.-H. Li, and D. Graur, 1982 Patterns of Nucleotide Substitution in Pseudogenes and Functional Genes. J Mol Evol 18: 360–369.

Goldmann, J. M., W. S. W. Wong, M. Pinelli, T. Farrah, D. Bodian et al., 2016 Parent-of-origin-specific signatures of de novo mutations. Nat. Genet. 48: 935–939.

Goodwin, S., J. D. McPherson, and W. R. McCombie, 2016 Coming of age: ten years of next-generation sequencing technologies. Nat. Rev. Genet. 17: 333–351.

de Graaf, G., F. Buckley, and B. G. Skotko, 2015 Estimates of the live births, natural losses, and elective terminations with Down syndrome in the United States. Am. J. Med. Genet. Part A 167: 756–767.

Griffiths, A., J. Miller, D. Suzuki, T. Lewontin, and W. Gelbart, 2000 *An Introduction to Genetic Analysis*. W. H. Freeman, Ed. New York.

Grimm, T., W. Kress, G. Meng, and C. R. Müller, 2012 Risk assessment and genetic counseling in families with Duchenne muscular dystrophy. Acta Myol. 31: 179–183.

Gu, W., F. Zhang, and J. R. Lupski, 2008 Mechanisms for human genomic rearrangements. BMC Pathogenetics 1: 4.

Haldane, J. B. S., 1935 The rate of spontaneous mutation of a human gene. J. Genet. 31: 317–326.

Haldane, J. B. S., 1946 The Mutation Rate of the Gene for Haemophilia, and Its Segregation Ratios in Males and Females. Ann. Eugen. 13: 262–271.

Hershberg, R., and D. A. Petrov, 2010 Evidence That Mutation Is Universally Biased towards AT in Bacteria. PLoS Genet. 6: e1001115.

Hodgkinson, A., and A. Eyre-Walker, 2011 Variation in the mutation rate across mammalian genomes. Nat. Rev. Genet. 12: 756–766.

Hollister, J. D., J. Ross-Ibarra, and B. S. Gaut, 2010 Indel-Associated Mutation Rate Varies with Mating System in Flowering Plants. Mol. Biol. Evol. 27: 409–416.

Hsu, Y. C., Y. T. Hsiao, T. Y. Kao, J. G. Chang, and G. S. Shieh, 2017 Detection of Somatic Mutations in Exome Sequencing of Tumor-only Samples. Sci. Rep. 7: 15959.

Hwang, D. G., and P. Green, 2004 Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. Proc. Natl. Acad. Sci. U. S. A. 101: 13994–14001.

Itsara, A., H. Wu, J. D. Smith, D. A. Nickerson, I. Romieu et al., 2010 De novo rates and selection of large copy number variation. Genome Res. 20: 1469–1481.

Keightley, P. D., 1994 The distribution of mutation effects on viability in Drosophila melanogaster. Genetics 138: 1315–1322.

Keightley, P. D., 2012 Rates and fitness consequences of new mutations in humans. Genetics 190: 295–304.

Keightley, P. D., L. Eöry, D. L. Halligan, and M. Kirkpatrick, 2011 Inference of mutation parameters and selective constraint in mammalian coding sequences by approximate Bayesian computation. Genetics 187: 1153–1161.

Keightley, P. D., and A. Eyre-Walker, 2007 Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. Genetics 177: 2251–2261.

Keightley, P. D., and O. Ohnishi, 1998 EMS-Induced Polygenic Mutation Rates for Nine Quantitative Characters in Drosophila melanogaster. Genetics 148: 753–766.

Kimura, M., 1968 Evolutionary rate at the molecular level. Nature 217: 624–626.

Kondrashov, A. S., 2002 Direct estimates of human per nucleotide mutation rates at 20 loci causing mendelian diseases. Hum. Mutat. 21: 12–27.

Kondrashov, A. S., 1998 Measuring spontaneous deleterious mutation process. Genetica 102: 183–197.

Kondrashov, A. S., and J. F. Crow, 1993 A molecular approach to estimating the human deleterious mutation rate. Hum. Mutat. 2: 229–234.

Kondrashov, F. A., and A. S. Kondrashov, 2010 Measurements of spontaneous rates of mutations in the recent past and the near future. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 365: 1169–1176.

Kong, A., M. L. Frigge, G. Masson, S. Besenbacher, P. Sulem et al., 2012 Rate of de novo mutations and the importance of father's age to disease risk. Nature 488: 471–475.

Korona, D. A., K. G. Lecompte, and Z. F. Pursell, 2011 The high fidelity and unique error signature of human DNA polymerase ε. Nucleic Acids Res. 39: 1763–1773.

Kropatsch, R., D. A. Akkad, M. Frank, C. Rosenhagen, J. Altmüller et al., 2016 A large deletion in RPGR causes XLPRA in Weimaraner dogs. BMC Canine Genet. Epidemiol. 3: 7.

Krzywinski, M., 2016 Molecular Cell Perspective Visualizing Clonal Evolution in Cancer. Mol. Cell 62: 652–656.

Kunkel, T. A., 2009 Evolving views of DNA replication (in)fidelity. Cold Spring Harb. Symp. Quant. Biol. 74: 91–101.

Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody et al., 2001 Initial sequencing and analysis of the human genome. Nature 409: 860–921.

Lawrence, M. S., P. Stojanov, P. Polak, G. V. Kryukov, K. Cibulskis et al., 2013 Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature 499: 214–218.

Lercher, M. J., and L. D. Hurst, 2002 Human SNP variability and mutation rate are higher in regions of high recombination. Trends Genet. 18: 337–340.

Li, R., A. Montpetit, M. Rousseau, S. Y. M. Wu, C. M. T. Greenwood et al., 2014 Somatic point mutations occurring early in development: a monozygotic twin study. J. Med. Genet. 51: 28–34.

Lieber, M. R., 2010 The Mechanism of Double-Strand DNA Break Repair by the Nonhomologous DNA End-Joining Pathway. Annu. Rev. Biochem. 79: 181–211.

Lindahl, T., 1999 Quality Control by DNA Repair. Science 286: 1897–1905.

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: 803–819.

Lodato, M. A., M. B. Woodworth, S. Lee, G. D. Evrony, B. K. Mehta et al., 2015 Somatic mutation in single human neurons tracks developmental and transcriptional history. Science 350: 94–98.

Lupski, J. R., 2007 Genomic rearrangements and sporadic disease. Nat. Genet. 39: S43–S47.

Lynch, M., M. S. Ackerman, J.-F. Gout, H. Long, W. Sung et al., 2016 Genetic drift, selection and the evolution of the mutation rate. Nat. Rev. Genet. 17: 704–714.

Mardis, E. R., 2017 DNA sequencing technologies: 2006–2016. Nat. Protoc. 12: 213–218.

Maxam, A. M., and W. Gilbert, 1977 A new method for sequencing DNA (DNA chenistry/dimethyl sulfate cleavage/hydrazine/piperidine). Biochemistry 74: 560–564.

May, C. A., A. J. Jeffreys, and J. A. Armour, 1996 Mutation rate heterogeneity and the generation of allele diversity at the human minisatellite MS205 (D16S309). Hum. Mol. Genet. 5: 1823–1833.

Michaelson, J. J., Y. Shi, M. Gujral, H. Zheng, D. Malhotra et al., 2012 Whole-Genome Sequencing in Autism Identifies Hot Spots for De Novo Germline Mutation. Cell 151: 1431–1442.

Mikkelsen, T. S., L. W. Hillier, E. E. Eichler, M. C. Zody, D. B. Jaffe et al., 2005 Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437: 69–87.

Milholland, B., X. Dong, L. Zhang, X. Hao, Y. Suh et al., 2017 Differences between germline and somatic mutation rates in humans and mice. Nat. Commun. 8: 1-8.

Minoche, A. E., J. C. Dohm, and H. Himmelbauer, 2011 Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. Genome Biol. 12: R112.

Molster, C., D. Urwin, L. Di Pietro, M. Fookes, D. Petrie et al., 2016 Survey of healthcare experiences of Australian adults living with rare diseases. Orphanet J. Rare Dis. 11: 30.

Mukai, T., 1970 Viability mutations induced by ethyl methanesulfonate in Drosophila melanogaster. Genetics 65: 335–348.

Muller, H. J., 1928 The Measurement of Gene Mutation Rate in Drosophila, Its High Variability, and Its Dependence Upon Temperature. Genetics 13: 279–357.

Munné, S., M. Bahçe, M. Sandalinas, T. Escudero, C. Márquez et al., 2004 Differences in chromosome susceptibility to aneuploidy and survival to first trimester. Reprod. Biomed. Online 8: 81–90.

Munné, S., J. Grifo, and D. Wells, 2016 Mosaicism: “survival of the fittest” versus “no embryo left behind”. Fertil. Steril. 105: 1146–1149.

Myers, R. H., M. E. Macdonald, W. J. Koroshetz, M. P. Duyao, C. M. Ambrose et al., 1993 De novo expansion of a (CAG)n repeat in sporadic Huntington’s disease. Nat. Genet. 5: 168–173.

Nachman, M. W., and S. L. Crowell, 2000 Estimate of the mutation rate per nucleotide in humans. Genetics 156: 297–304.

Nachman, M. W., 2004 Haldane and the first estimates of the human mutation rate. J. Genet. 31: 235–244.

Navin, N., and J. Hicks, 2011 Future medical applications of single-cell sequencing in cancer. Genome Med. 3: 31.

O'Roak, B. J., P. Deriziotis, C. Lee, L. Vives, J. J. Schwartz et al., 2011 Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. Nat. Genet. 43: 585–589.

Peterson, L. E., and T. Kovyrshina, 2017 Progression inference for somatic mutations in cancer. Heliyon 3: e00277.

Poduri, A., G. D. Evrony, X. Cai, and C. A. Walsh, 2013 Somatic mutation, genomic variation, and neurological disease. Science 341: 1237758.

Poultney, C. S., A. P. Goldberg, E. Drapeau, Y. Kou, H. Harony-Nicolas et al., 2013 Identification of small exonic CNV from whole-exome sequence data and application to autism spectrum disorder. Am. J. Hum. Genet. 93: 607–619.

Preston, B. D., T. M. Albertson, and A. J. Herr, 2010 DNA replication fidelity and cancer. Semin. Cancer Biol. 20: 281–293.

Rahbari, R., A. Wuster, S. J. Lindsay, R. J. Hardwick, L. B. Alexandrov et al., 2016 Timing, rates and spectra of human germline mutation. Nat. Genet. 48: 126–133.

Reuter, J. A., D. V. Spacek, and M. P. Snyder, 2015 High-Throughput Sequencing Technologies. Mol. Cell 58: 586–597.

Risques, R. A., and S. R. Kennedy, 2018 Aging and the rise of somatic cancer-associated mutations in normal tissues. PLoS Genet. 14: e1007108.

Rosenfeld, J. A., C. E. Mason, and T. M. Smith, 2012 Limitations of the Human Reference Genome for Personalized Genomics. PLoS One 7: e40294.

Rossetti, L. C., C. P. Radic, M. M. Abelleyro, I. B. Larripa, and C. D. De Brasi, 2011 Eighteen years of molecular genotyping the hemophilia inversion hotspot: from southern blot to inverse shifting-PCR. Int. J. Mol. Sci. 12: 7271–7285.

Sanger, F., S. Nicklen, and A. R. Coulson, 1977 DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. U. S. A. 74: 5463–5467.

Savolainen, P., Y. Zhang, J. Luo, J. Lundeberg, and T. Leitner, 2002 Genetic Evidence for an East Asian Origin of Domestic Dogs. Science 298: 1610–1613.

Sayyab, S., A. Viluma, K. Bergvall, E. Brunberg, V. Jagannathan et al., 2016 Whole-Genome Sequencing of a Canine Family Trio Reveals a FAM83G Variant Associated with Hereditary Footpad Hyperkeratosis. G3 (Bethesda) 6: 521-527.

Scherer, S. W., C. Lee, E. Birney, D. M. Altshuler, E. E. Eichler et al., 2007 Challenges and standards in integrating surveys of structural variation. Nat. Genet. 39: S7–S15.

Schloss, J. A., 2008 How to get genomes at one ten-thousandth the cost. Nat. Biotechnol. 26: 1113–1115

Sebat, J., B. Lakshmi, D. Malhotra, J. Troge, C. Lese-Martin et al., 2007 Strong association of de novo copy number mutations with autism. Science 316: 445–449.

Ségurel, L., M. J. Wyman, and M. Przeworski, 2014 Determinants of Mutation Rate Variation in the Human Germline. Annu. Rev. Genomics Hum. Genet 15: 47–70.

Seshagiri, S., 2013 The burden of faulty proofreading in colon cancer. Nat. Genet. 45: 121–122.

Shendure, J., and J. M. Akey, 2015 The origins, determinants, and consequences of human mutations. Science 349: 1478–83.

Smeds, L., A. Qvarnström, and H. Ellegren, 2016 Direct estimate of the rate of germline mutation in a bird. Genome Res. 26: 1211–1218.

Sueoka, N., 1961 Correlation between Base Composition of Deoxyribonucleic Acid and Amino Acid Composition of Protein. Proc. Natl. Acad. Sci. U. S. A. 47: 1141–1149.

Tang, F., C. Barbacioru, Y. Wang, E. Nordman, C. Lee et al., 2009 mRNA-Seq whole-transcriptome analysis of a single cell. Nat. Methods 6: 377–382.

Tian, D., Q. Wang, P. Zhang, H. Araki, S. Yang et al., 2008 Single-nucleotide mutation rate increases close to insertions/deletions in eukaryotes. Nature 455: 105–108.

Timmerman, L., 2015 *DNA Sequencing Market Will Exceed $20 Billion, Says Illumina CEO Jay Flatley.* [online] Forbes. Available at: https://www.forbes.com/sites/luke timmerman/2015/04/29/qa-with-jay-flatley-ceo-of-illumina-the-genomics-company-pursuing-a-20b-market/#3b8e8b6a42e7 [Accessed 21 Nov. 2018].

Uchimura, A., M. Higuchi, Y. Minakuchi, M. Ohno, A. Toyoda et al., 2015 Germline mutation rates and the long-term phenotypic effects of mutation accumulation in wild-type laboratory mice and mutator mice. Genome Res. 25: 1–10.

Veltman, J. a, and H. G. Brunner, 2012 De novo mutations in human genetic disease. Nat. Rev. Genet. 13: 565–575.

Venn, O., I. Turner, I. Mathieson, N. de Groot, R. Bontrop et al., 2014 Strong male bias drives germline mutation in chimpanzees. Science 344: 1272–1275.

Venter, J. C., M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural et al., 2001 The sequence of the human genome. Science 291: 1304–1351.

Vilà, C., P. Savolainen, J. E. Maldonado, I. R. Amorim, J. E. Rice et al., 1997 Multiple and Ancient Origins of the Domestic Dog. Science 2761687: 1687–1689.

Vilboux, T., G. Chaudieu, P. Jeannin, D. Delattre, B. Hedan et al., 2008 Progressive retinal atrophy in the Border Collie: a new XLPRA. BMC Vet. Res. 4: 10.

Visootsak, J., and J. M. Graham, 2006 Klinefelter syndrome and other sex chromosomal aneuploidies. Orphanet J. Rare Dis. 1: 42.

Vissers, L. E. L. M., J. de Ligt, C. Gilissen, I. Janssen, M. Steehouwer et al., 2010 A de novo paradigm for mental retardation. Nat. Genet. 42: 1109–1112.

Wang, Y., and N. E. Navin, 2015 Advances and Applications of Single-Cell Sequencing Technologies. Mol. Cell 58: 598–609.

Wang, G., W. Zhai, H. Yang, R. Fan, X. Cao et al., 2013 The genomics of selection in dogs and the parallel evolution between dogs and humans. Nat. Commun. 4: 1860.

Waterston, R. H., K. Lindblad-Toh, E. Birney, J. Rogers, J. F. Abril et al., 2002 Initial sequencing and comparative analysis of the mouse genome. Nature 420: 520–562.

Watson, I. R., K. Takahashi, P. A. Futreal, and L. Chin, 2013 Emerging patterns of somatic mutations in cancer. Nat. Rev. Genet. 14: 703–718.

Wayne, R. K., and E. A. Ostrander, 1999 Origin, genetic diversity, and genome structure of the domestic dog. Bioessays. 21: 247–257.

Weirather, J. L., M. de Cesare, Y. Wang, P. Piazza, V. Sebastiano et al., 2017 Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. F1000Res. 6: 100.

Willet, C. E., L. Bunbury-Cruickshank, D. Van Rooy, G. Child, M. R. Shariflou et al., 2013 Empirical assessment of competitive hybridization and noise in ultra high density canine tiling arrays. BMC Bioinformatics 14: 1.

Wong, A. K., A. L. Ruhe, K. R. Robertson, E. R. Loew, D. C. Williams et al., 2013 A de novo mutation in KIT causes white spotting in a subpopulation of German Shepherd dogs. Anim. Genet. 44: 305–310.

Wong, W. S. W., B. D. Solomon, D. L. Bodian, P. Kothiyal, G. Eley et al., 2016 New observations on maternal age effect on germline de novo mutations. Nat. Commun. 7: 10486.

Yang, W., 2000 Structure and function of mismatch repair proteins. Mutat. Res. 460: 245–256.

Yang, L., L. J. Luquette, N. Gehlenborg, R. Xi, P. S. Haseley et al., 2013 Diverse mechanisms of somatic structural variations in human cancer genomes. Cell 153: 919–929.

Youn, A., and R. Simon, 2011 Identifying cancer driver genes in tumor genome sequencing studies. Bioinformatics 27: 175–181.

Zhu, X., S. Petrovski, P. Xie, E. K. Ruzzo, Y.-F. Lu et al., 2015 Whole-exome sequencing in undiagnosed genetic diseases: interpreting 119 trios. Genet. Med. 17: 774–781.

Zhu, L., Q. Wang, P. Tang, H. Araki, and D. Tian, 2009 Genomewide Association between Insertions/Deletions and the Nucleotide Diversity in Bacteria. Mol. Biol. Evol. 26: 2353–2361.

Zurynski, Y., M. Deverell, T. Dalkeith, S. Johnson, J. Christodoulou et al., 2017 Australian children living with rare diseases: experiences of diagnosis and perceived consequences of diagnostic delays. Orphanet J. Rare Dis. 12: 68.

Chapter 2. A performance comparison of popular single nucleotide variant detection methodologies applied to low coverage whole genome sequencing data

2.1. Abstract

Next generation sequencing platforms have become essential tools for understanding DNA in a wide range of contexts. Their success heavily relies on the accuracy, sensitivity and specificity of methods used to discern differences between the reference genome and genomes under investigation. Here we compare the relative performances of five popular single nucleotide variant callers with and without their associated recommended hard filtering criteria. We compare: FreeBayes; the Genome Analysis Tool-kit's Haplotype Caller and Unified Genotyper; SAMtools; and VarScan. We tailor this comparison to suit smaller projects with modest sample numbers (n = 10) and coverage (~10X) to fill a current gap in the literature. Other comparison studies are generally applicable only to larger projects in model species, where there is access to large amounts of sequencing data and curated callsets for base and variant quality score recalibration. We estimated the accuracy, sensitivity and specificity of each pipeline according to the genotype concordance rate and number with the "truth" dataset for 10 canine samples. The truth dataset was defined as genotypes obtained from the CanineHD BeadChip array. Whole genome sequencing data was performed on the Illumina HiSeq2000 or HiSeq2500 platform as 100-101 base pair, paired end reads to an average sample coverage of 10.3X. With the exception of GATK Haplotype Caller, applying recommended hard filters did not improve the performance of genotyping concordance at the tested levels of minimum coverage. The default VarScan pipeline with no additional filters applied (VarScan uses SAMtools mpileup, without base alignment quality computation) generally outperformed other callers in terms of accuracy, sensitivity and specificity. The results of this study demonstrate that hard filtering of variant calls from low-powered genome studies can impair accuracy, sensitivity and specificity of callsets and provides some benchmark performance metrics

on a range of low coverage levels. These can be applied to future studies to aid optimal variant detection.

2.2. Introduction

Next-generation sequencing (NGS) technologies have provided scientists with unprecedented access to understanding DNA, one of the fundamental molecules of life. The range of applications has facilitated many discoveries that were made in research fields as diverse as ecology, agriculture, evolution, population diversity and human health (Schuster 2007; Mardis and Salzburg 2008; Ekblom and Galindo 2011). As sequencing technologies improve and the cost to sequence each nucleotide decreases, NGS is beginning to emerge from its role as a pure research methodology to become a common practice strategy for use in personalized medicine (O'Rawe *et al.* 2013; Hwang *et al.* 2015).

The successful use of NGS in research and in practical applications relies heavily on our ability to accurately detect, categorize and genotype true biological variants of interest in genome data. This is a complex feat as sequencing errors, alignment artefacts and other sources of error can be indistinguishable from true biological variation. Each NGS sequencing platform is associated with an expected error rate and are prone to specific known types of errors. For example, Illumina's short read sequencing technologies (36 – 250 base pairs (bp)) have an overall estimated accuracy of >99.5% (Bentley *et al.* 2008) and produce more substitution (0.11 – 0.28%) than insertion-deletion (indel) errors (3.2×10^{-6} – 2.5×10^{-5} %) (Minoche *et al.* 2011). Platforms which produce longer reads are better at resolving larger structural changes but are usually associated with higher rates of error. IonTorrent platforms (ThermoFisher Scientific) produce read lengths of 400 bp, are more prone to indel errors and have difficulty in accurately sequencing homopolymers larger than 6 – 8 bp long (Loman *et al.* 2012; Forgetta *et al.* 2013). Newer, very long read sequencing platforms such as those offered by PacBio (>10kb) and Oxford Nanopore Technologies, also have a tendency towards indel errors and have overall error rates of up to 15% (Carneiro *et al.* 2012).

To avoid the inclusion of these erroneous variants in data analyses, many variant calling algorithms have been developed. Most existing variant calling pipelines use statistical inference to determine the likelihood of a true biological variant existing at any one site (reviewed in Nielsen *et al.* 2011). Depending on the sequencing platform used and the type of variants interrogated (germline or somatic and mutation type), several different quality score types are considered. Variant callers focussing on single nucleotide polymorphism (SNP) and indels typically account base quality, mapping quality and the local assembly quality metrics to determine the most likely genotype and then to provide an associated quality score (a "genotype likelihood"). Popular structural variant calling programs may assess split reads, read pair mapping span, read pair relative orientation and relative read depth (Tattini *et al.* 2015). While structural variation remains more challenging to accurately genotype using current technologies than SNPs and indels, it is generally accepted that we have not yet perfected small variant calling and that no single approach will work best across all datasets.

Contemporary algorithms further improve the accuracies of calls within individuals by incorporating population-level data. For example, a multi-sample calling mode can be employed, where each locus is assessed across many samples simultaneously to develop a better expectation of whether the site is truly biologically polymorphic. Allele frequencies, genotype frequencies and patterns of linkage disequilibrium (LD) obtained from the observation of multiple samples can improve the confidence of a true biological variant at any given site. LD can be used to impute missing data and infer genotypes, improving calling sensitivity (Nielsen *et al.* 2011; Wang, Lu, *et al.* 2013). For well-curated species such as human and mice, prior information can be obtained from public datasets such as dbSNP (https://www.ncbi.nlm.nih.gov/SNP/) and HapMap (http://www.hapmap.org/). When a set of known variants is not available, which is often the case in non-model organisms, high quality variants may be computed from the data at hand. An initial round of variant calling creates a callset that can be used to 'teach' the calling algorithm the quality score profiles of poor and good quality variants specific to the data in hand, enabling recalibration of genotype likelihood scores (McKenna *et al.* 2010).

The extent to which multi-sample calling improves calling sensitivity over single-sample calling depends primarily on the number of samples and average coverage per sample. One study observed that single-sample calling yielded higher calling sensitivity than multi-sample calling in samples with low sequencing depths (5X)(Cheng *et al.* 2014). This result was independent of the minor allele frequency of the variant in the studied population. At low sequencing depths of 5X, multi-sample calling provided a significant improvement in sensitivity (~20%) only for low frequency and rare variant loci. This improvement required an additional 1,092 samples obtained from the 1,000 Genomes Project. In Cheng *et al.* (2014), the algorithm sensitivity to call variants always improves as the average coverage increases. Despite this, many researchers still opt for sequencing more samples at lower coverage (less than 10X) as this is believed to provide superior power for detecting common population variants relative to sequencing fewer samples at higher depths (Le and Durbin 2011; Sims *et al.* 2014; Gilly *et al.* 2017).

Although there are many variant caller comparison studies suited to model species with large datasets (for examples, see Liu *et al.* 2013; Cheng *et al.* 2014; Cornish and Guda 2014; Pirooznia *et al.* 2014), there is limited information on the performance of variant calling pipelines that are tailored to smaller studies where prior observation of variants beyond the data of a single sequenced individual cannot be obtained and used for quality recalibration. For many laboratories, obtaining NGS data on multiple samples is not economically feasible and known variant data may not be available, especially for non-model species. In these situations, the strategy used to remove sequencing errors often relies on hard filtering raw data. Hard filtering is defined as setting a threshold (usually arbitrary) for a specific parameter of the data and variants which do not meet this value are removed from further analysis. Commonly used hard filtering parameters include base quality, mapping quality, coverage and strand bias (Van der Auwera *et al.* 2013; Koboldt *et al.* 2013; Garrison 2015; Willet, Haase, *et al.* 2015b). Without sequence redundancy in low coverage data to appropriately represent sites that are more problematic to sequence such as GC rich and heterodimer prone fragments, hard filtering approaches may be too stringent and lead to false negative calls.

When hard filtering is applied, selecting appropriate threshold values is crucial to the success of whole genome sequencing (WGS) analysis in smaller studies or those using non-model species of interest. The needs will be affected by the type of analysis (for example, mapping analysis versus mutation detection analyses). The development of a validated pipeline that is well suited to a given dataset is time consuming and expensive, as additional methods of sequencing are needed to validate calls. For this reason, many researchers opt to employ recommended hard filtering cut-off values to define high quality variants when using popular variant calling programs (hard filtering recommendations are described in Methods and Supplementary Table S2).

Our goal is to compare five popular variant callers: FreeBayes (Garrison and Marth 2012); GATK Unified Genotyper (GATK UG) (McKenna *et al.* 2010); GATK Haplotype Caller (GATK HC) (McKenna *et al.* 2010); SAMtools (Li *et al.* 2009) and VarScan (Koboldt *et al.* 2013) and to observe the relative performance of pipelines after variants are filtered using recommended hard filtering criteria (Van der Auwera *et al.* 2013; Koboldt *et al.* 2013; Garrison 2015; Willet, Haase, *et al.* 2015b). The variant callers selected are tailored for short read data, such as those provided by Illumina platforms. Illumina currently has the largest market share in NGS platforms (Timmerman 2015). We apply these callers in single-sample mode to observe the sensitivity and specificity achieved when cut-off criteria are applied (hard filtering pipelines). We also observe the performance of callers without hard filtering (raw pipelines). Samples were 10 canine samples that had been subjected to WGS on popular Illumina HiSeq platforms offering genomes with a range or mean coverage from low to moderate (6 – 16X). To measure the relative calling quality of the algorithms, we assessed the concordance between genotyping calls made by the pipelines with genotypes called at 173,650 SNP markers using results on the same individuals from the CanineHD BeadChip array commercially provided by Neogen. Using these results as guidelines, researchers working with small (low coverage and number of samples) genome sequencing studies can select and adjust pipelines to suit their project goals.

2.3. Methods

2.3.1. Samples

Ten dogs (*Canis lupus familiaris*) that included three Australian Cattle Dogs, four Miniature Schnauzers and three Hungarian Puli were used in this study. These data formulate four unique parent-offspring trios. EDTA-stabilized whole blood or tissue was collected from the Australian Cattle Dogs and Miniature Schnauzers (See Table S1 for sample information). Genomic DNA was extracted using the illustra Nucleon BACC 2 kit using the manufacturer's recommended protocols (GE Healthcare). Hungarian Puli and two Miniature Schnauzer genotyping array and WGS data were obtained from previous studies (Willet, Makara, *et al.* 2015; Chew, Haase, Willet, *et al.* 2017). This study was conducted with approval from the Animal Ethics Committee at the University of Sydney (approval number N00/9–2009/3/5109 and N00/10-2012/3/5837 2015/902). See Table S1 for sample information.

2.3.2. Genotyping array data and the truth dataset

Samples were genotyped at 173,650 SNP loci on the CanineHD BeadChip array (Illumina) by GeneSeek (Lincoln, NE). Identity by descent proportions were obtained using PLINK (Purcell *et al.* 2007) and these calculations were used to confirm the pedigree relationships stated by registry data (Australian National Kennel Council) among each parent-offspring trio. SNPs genotyped on this array platform were used as the 'truth dataset' in this study. NCBI's remapping service (https://www.ncbi.nlm.nih.gov/genome/tools/remap) was used to convert CanFam 2.0 to CanFam 3.1 array coordinates to make comparison with NGS genotypes consistent. To ensure that only accurately genotyped SNPs were considered, markers that did not adhere to Mendelian inheritance laws were excluded from the analysis.

2.3.3. Next-generation sequencing

WGS data was generated on the Illumina HiSeq2000 (n = 8) or the Illumina HiSeq2500 (n = 2) by the Ramaciotti Centre, University of New South Wales, Kensington. Libraries

were prepared with the Illumina TruSeq kit. Each sample was barcoded and sequenced as 100 or 101 base-pair, paired-end reads on either one half or one full lane of the sequencing platform. See Table S1 for sample information.

The Burrows-Wheeler Alignment mem (BWA-mem) tool outperforms other popular short read aligners and is recommended for pairing with multiple variant calling programs including those used in this study (Li and Durbin 2009; Van der Auwera *et al.* 2013; Cornish and Guda 2014; Layer *et al.* 2014; Faust and Hall 2014). Here we use BWA-mem to align raw reads as pairs to the CanFam 3.1 reference genome for each sample using default parameters (Hoeppner *et al.* 2014). Polymerase chain reaction (PCR) duplicates were marked using Picard (http://picard.sourceforge.net). Local realignment around indels was performed using GATK (McKenna *et al.* 2010; DePristo *et al.* 2011).

2.3.4. Variant Calling and Hard Filtering Criteria

For each of the callers considered (FreeBayes, GATK HC, GATK UG, SAMtools and VarScan), we used recommended criteria (Figure 2.1) in the single sample mode to call variants and obtain genotypes (raw pipeline, caller-R). Next, we applied recommended hard filtering criteria to obtain high quality SNP genotypes (caller-F, Figure 2.1 and Table 2.1). Supplementary Table S2 provides a description of the parameters used in variant calling. For all pipelines, we defined indels or loci that were not bi-allelic as not called, as these are not assayed on the 'truth' platform used. Due to the stochastic nature of locus coverage in WGS experiments, we assessed genotype calls at a range of minimum base coverage thresholds. Raw base coverage at marker loci were obtained using SAMtools bedcov. Eleven different minimum coverage levels were used, ranging from zero coverage to 20X with an increment of 2X.

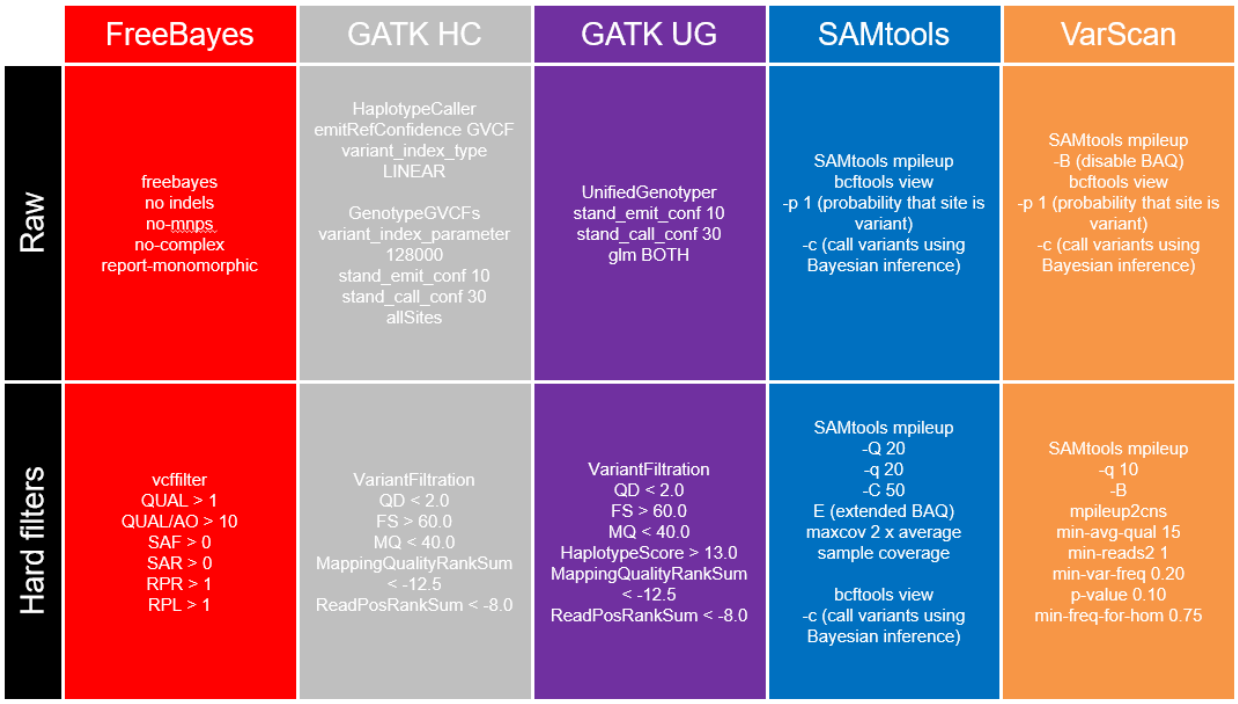

Figure 2.1. Representation of the ten variant calling pipelines used in this study. Five variant callers were used (FreeBayes, GATK HC, GATK UG, SAMtools, VarScan). The row labelled “Raw” indicates the options used for raw variant calling for each variant calling program, before hard-filters were applied. The row labelled “Hard filters” include additional hard-filtering steps performed for each variant calling program. For VarScan, we initially included all loci covered by at least one read and performed minimum coverage cut-off post variant calling. For detailed explanations of the filtering parameters, see Supplementary Table S2 and the associated software documentation.

Table 2.1. Variant callers and recommended hard filtering criteria used in this study.

Default parameters for each variant caller were used in the raw pipelines.

Program	Variant caller	Version	Hard filtering recommendation
GATK	Haplotype Caller	3.6.0	(Van der Auwera *et al.* 2013)
GATK	Unified Genotyper	3.6.0	(Van der Auwera *et al.* 2013)
SAMtools	mpileup	0.1.19	(Willet, Haase, *et al.* 2015b)
FreeBayes	FreeBayes	1.0.2-33	(Garrison 2015)
VarScan*	SAMtools mpileup mpileup2cns	2.3.9	(Koboldt *et al.* 2013)

* VarScan depends on the input from SAMtools' variant caller mpileup, without probabilistic realignment for the computation of base alignment quality (BAQ). The raw VarScan pipeline output is based on this and does not include the use of mpileup2cns.

2.3.5. Refinement of the truth dataset

We used genotypes called by each of the 10 pipelines to further refine the truth dataset. Loci that exhibited no genotype concordance across all 10 individuals and five variant callers at any one locus were removed from the truth dataset. This method aids in removing additional markers that were affected by CanFam 2.0 and CanFam 3.1 reference assembly orientation differences as well as markers that were incorrectly genotyped on the array.

2.3.6. Comparison metrics

We performed a paired, two-tailed t-test to determine whether the total concordance rates (2.1) were significantly different amongst the 10 pipelines tested regardless of minimum coverage requirement set. We then compared total concordance rates (%) amongst the 10 pipelines and across the 11 different coverage cut-off levels. For each of the variant callers, we compared their raw pipeline to their corresponding pipeline including hard filters to determine if filtering improved genotype concordance. To estimate the sensitivity, specificity and accuracy of each variant calling pipeline, we compared the total number of loci where the genotype called by the pipeline agreed with the truth dataset (concordant) and where the genotype called differed from the truth dataset (discordant). We calculated the standard deviation of the total number of concordant loci at each minimum coverage requirement level as a measure of variance between pipelines. Genotyping rates (including concordant and discordant genotype calls) are in supplementary Table S3. To determine if there are genotyping biases, we compared these concordance metrics for homozygous (2.2) and heterozygous (2.3) array genotypes separately.

$$\boldsymbol{Total\ concordance\ rate} = \frac{\sum_{n=1}^{10} \boldsymbol{Number\ of\ concordant\ genotypes\ called\ by\ the\ pipeline}}{\sum_{n=1}^{10} \boldsymbol{Number\ of\ loci\ in\ the\ truth\ dataset}}\ \boldsymbol{x\ 100} \qquad (2.1)$$

$$\boldsymbol{Homozygous\ concordance} = \frac{\sum_{n=1}^{10} \boldsymbol{Number\ of\ homozygous\ concordant\ genotypes\ called\ by\ the\ pipeline}}{\sum_{n=1}^{10} \boldsymbol{Number\ of\ loci\ with\ homozygous\ genotypes\ in\ the\ truth\ dataset}} \qquad (2.2)$$

$$\boldsymbol{Heterozygous\ concordance} = \frac{\sum_{n=1}^{10} \boldsymbol{Number\ of\ heterozygous\ concordant\ genotypes\ called\ by\ the\ pipeline}}{\sum_{n=1}^{10} \boldsymbol{Number\ of\ loci\ with\ \ heterozygous\ genotypes\ in\ the\ truth\ dataset}} \qquad (2.3)$$

2.4. Results

2.4.1. Truth and whole genome sequencing variant dataset

A total of 171,672 unique markers were considered in each of the 10 pipelines and 10 individuals after removing 796 SNPs which: did not conform to Mendelian inheritance; 40 that could not be converted to the CanFam 3.1 reference assembly; and a further 712 markers that had no genotype concordance with any of the 10 pipelines and 10 individuals. Loci which were genotyped on both WGS data and the CanineHD BeadChip array differed depending on the individual, variant caller, pipeline and coverage.

Whole genome sequencing on the Illumina HiSeq2000 or HiSeq2500 produced an average of 273 million reads per sample, with 99.13% of these successfully mapping to the CanFam 3.1 reference genome. This corresponds to an average mapped coverage of 10.3X.

2.4.2. Comparison of genotype concordance rates of the 10 variant calling pipelines to truth dataset

We found that the VarScan-R pipeline generally had significantly better genotype concordance than all other pipelines studied ($P_{T\text{-}TEST} < 0.05$, Table 2.2). The VarScan-R pipeline uses SAMtools mpileup without BAQ and achieved concordance rates of 98.37 – 99.67%. The GATK UG-R pipeline achieve similar levels of genotype concordance at higher levels of minimum coverage requirement (10X, see Table S4 for percent concordances for all pipelines and minimum coverage requirements). SAMtools-F outperformed the other pipelines at 20X minimum coverage requirement. The FreeBayes-F and VarScan-F pipelines underperformed significantly in comparison to the other eight pipelines tested ($P_{T\text{-}TEST} < 0.05$), especially when lower minimum coverage requirements were set (Table S4). We also observed the effect on genotype concordance when the minimum coverage requirement increased (Figure 2.2 and Table S4). For all pipelines, genotype concordance rates improved as the minimum coverage requirement increased for both raw and filtered variants, except at 20X where improvement was only seen for FreeBayes-F, SAMtools-F and VarScan-F.

Table 2.2. *P*-values from paired, two-tailed t tests on average genotype concordance rates of 10 different pipelines using five different variant callers with and without hard filtering (FreeBayes, GATK HC, GATK UG, SAMtools and VarScan) compared against genotypes obtained using the CanineHD BeadChip array.

	FreeBayes Raw	FreeBayes Filtered	GATK HC Raw	GATK HC Filtered	GATK UG Raw	GATK UG Filtered	SAMtools Raw	SAMtools Filtered	VarScan Raw	VarScan Filtered
FreeBayes Raw		0.001	0.061	0.334	0.018	0.242	0.042	0.173	0.016	0.002
FreeBayes Filtered			0.001	0.001	0.001	0.001	0.002	0.001	0.001	0.001
GATK HC Raw				0.182	0.003	0.001	0.017	0.041	0.020	0.002
GATK HC Filtered					0.003	0.117	0.024	0.258	0.011	0.002
GATK UG Raw						0.001	0.215	0.011	0.040	0.001
GATK UG Filtered							0.017	0.569	0.008	0.002
SAMtools Raw								0.041	0.001	0.002
SAMtools Filtered									0.020	0.001
VarScan Raw										0.002

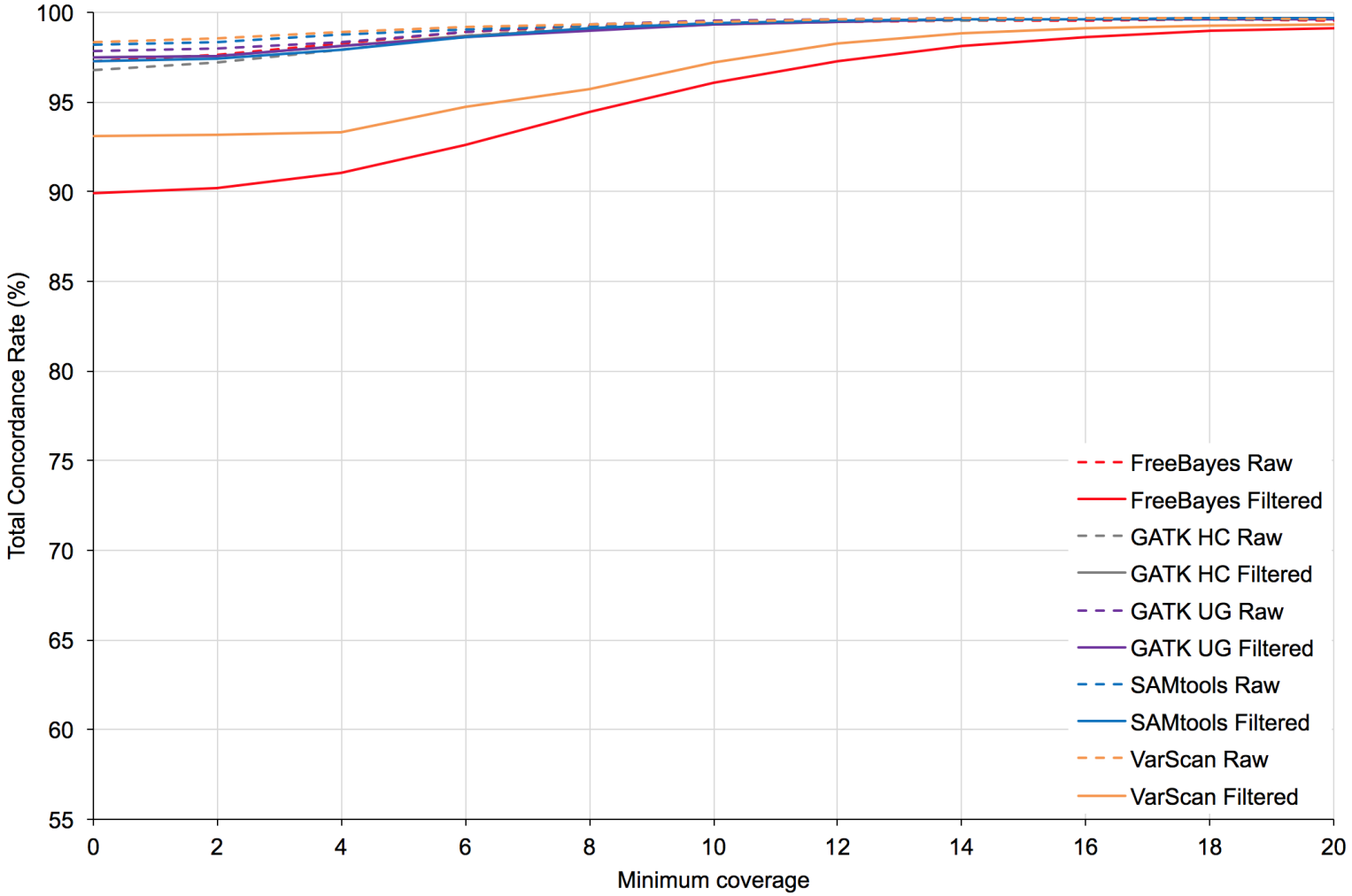

Figure 2.2. Percent concordance of all genotypes (homozygous and heterozygous) called by 10 different pipelines using five different variant callers with and without hard filtering (FreeBayes, GATK HC, GATK UG, SAMtools and VarScan) compared against genotypes obtained using the CanineHD BeadChip array.

2.4.3. Comparison of genotype concordance rates between raw pipelines and corresponding pipelines that include hard filters

For each variant caller, we compared their raw pipeline to their corresponding pipeline including hard filters applied to observe the effect of hard filters to genotype concordance. In general, applying hard filters to variants using recommended criteria did not improve genotype concordance rates. The exceptions where applying filters did improve genotype concordance occurred for GATK HC-F for low (0 – 2X) and higher minimum coverage and SAMtools-F for higher levels (14 – 20X) of minimum coverage requirements (Figure 2.2 and Table S4).

2.4.4. Total concordance and discordance and standard deviation of genotypes called by each of the pipelines to the truth dataset

To estimate the sensitivity and specificity of each pipeline, we calculated the number of concordant (Table 2.3) and discordant genotypes (Table 2.4) to the genotypes of the truth dataset. The VarScan-R pipeline called the most number of concordant genotypes at low minimum coverage requirements (0 – 8X). At moderate to higher levels of minimum coverage requirements (10 – 20X), GATK UG-R and VarScan-R both called the highest number of total concordant genotypes. However, as standard deviation decreased as the level of minimum coverage requirement increased, the number of concordant genotypes was similar between most pipelines (Table 2.3).

A similar trend was observed for the number and standard deviation of discordant genotypes across the 10 pipelines and minimum coverage requirement levels. The VarScan-R pipeline had the lowest number of discordant genotypes at lower minimum coverage requirement levels (0 – 8X). At moderate to higher minimum coverage requirement levels (10 – 20X), GATK UG-R, SAMtools-F and VarScan-R had the lowest number of discordant genotypes. Variation amongst the pipelines was minimal at higher levels of minimum coverage requirement (Table 2.4). Like the total number of concordant genotypes, standard deviation and minimum coverage requirement levels had an inverse relationship.

Table 2.3. Total numbers and standard deviation of concordant genotypes called by 10 different pipelines using five variant callers with and without hard filtering (FreeBayes, GATK HC, GATK UG, SAMtools and VarScan) compared against genotypes obtained using the CanineHD BeadChip array.

Shaded cells designate the pipeline with the highest number of concordant genotypes at each minimum coverage threshold. Performance was generally similar for most pipelines at minimum coverage >10X.

Minimum Coverage	FreeBayes Raw	FreeBayes Filtered	GATK HC Raw	GATK HC Filtered	GATK UG Raw	GATK UG Filtered	SAMtools Raw	SAMtools Filtered	VarScan Raw	VarScan Filtered	Standard Deviation
0	1,639,706	1,499,652	1,631,264	1,634,521	1,644,833	1,635,701	1,654,915	1,628,495	1,657,501	1,535,511	53,505
2	1,632,138	1,492,141	1,624,304	1,626,840	1,637,095	1,628,128	1,644,661	1,621,297	1,647,137	1,533,125	52,209
4	1,573,448	1,441,869	1,567,531	1,567,118	1,574,148	1,568,019	1,581,352	1,561,117	1,583,428	1,490,281	46,588
6	1,416,130	1,312,268	1,412,521	1,409,977	1,415,718	1,410,022	1,418,551	1,406,174	1,419,994	1,355,233	35,422
8	1,170,929	1,103,386	1,169,450	1,166,525	1,171,527	1,166,425	1,171,091	1,164,005	1,172,146	1,128,791	23,253
10	900,319	862,846	900,136	898,190	901,473	898,862	900,741	895,419	901,428	880,652	12,587
12	655,057	636,238	654,913	654,074	655,791	654,765	655,423	650,619	655,824	646,683	6,228
14	457,623	448,570	457,330	457,114	457,946	457,525	457,737	453,275	457,964	454,174	3,065
16	307,387	303,290	307,305	307,269	307,678	307,448	307,568	303,545	307,672	305,879	1,708
18	196,889	195,142	196,759	196,776	196,991	196,845	196,922	193,420	196,980	196,189	1,157
20	119,587	118,887	119,506	119,514	119,649	119,550	119,605	117,202	119,628	119,244	752

Table 2.4. Total number and standard deviation discordant genotypes called by 10 different pipelines using five variant callers with and without hard filtering (FreeBayes, GATK HC, GATK UG, SAMtools and VarScan) compared against genotypes obtained using the CanineHD BeadChip array.

Shaded cells designate the pipeline with the lowest number of discordant genotypes at each minimum coverage threshold. Performance was generally similar for most pipelines at minimum coverage >10X.

Minimum Coverage	FreeBayes Raw	FreeBayes Filtered	GATK HC Raw	GATK HC Filtered	GATK UG Raw	GATK UG Filtered	SAMtools Raw	SAMtools Filtered	VarScan Raw	VarScan Filtered	Standard Deviation
0	45,230	167,798	53,795	42,542	35,692	42,314	30,842	45,345	27,396	113,494	44,812
2	38,941	161,463	46,702	40,244	33,478	40,071	27,154	42,885	23,907	112,709	44,414
4	27,193	142,012	32,822	30,036	26,242	30,197	19,801	33,223	17,179	106,102	42,073
6	15,465	104,695	18,613	18,707	15,707	19,587	13,345	19,600	11,571	75,431	31,856
8	8,703	64,875	9,667	10,858	7,985	11,756	8,692	10,215	7,453	50,350	20,649
10	5,634	35,435	5,332	6,167	4,403	6,253	5,272	5,511	4,494	24,955	10,766
12	3,415	17,737	3,154	3,397	2,627	3,290	3,078	3,149	2,623	11,547	5,091
14	1,932	8,624	1,921	1,867	1,583	1,831	1,832	1,805	1,575	5,227	2,311
16	1,303	4,304	1,162	1,069	992	1,117	1,120	1,084	995	2,703	1,083
18	765	2,074	754	676	653	734	725	648	648	1,407	468
20	506	1,040	496	448	436	481	475	390	435	815	208

2.4.5. Homozygous verse heterozygous concordance

To identify genotyping biases, we separately explored the rate of homozygous and heterozygous genotypes called by each of the 10 pipelines that were concordant with the truth genotypes derived from the array platform. For all variant calling pipelines and levels of coverage, homozygous concordance rates were higher than heterozygous concordance rates (Figure 2.3 and Figure 2.4). Heterozygous genotype concordance rates were more heavily influenced by the level of minimum coverage requirement. The difference in the levels of concordance between homozygous and heterozygous genotypes decreased as the level of coverage increased.

The highest rates of homozygous concordance were achieved with the VarScan-F at lower to moderate levels of minimum coverage (0 – 12X, 99.86 – 99.85%) and the SAMtools-F pipeline at higher levels of minimum coverage (14 – 20X, 99.85 – 99.85%, Figure 2.3). Table S5 contains percentages of homozygous concordance for all 10 pipelines and minimum coverage requirement levels. Heterozygous concordance rates were highest using the VarScan-R (0 – 8X, 16 – 18X) and GATK UG-R (10 – 14X and 20X, Figure 2.4). Table S6 contains percentages of heterozygous concordance for all 10 pipelines and minimum coverage requirement levels.

We compared homozygous and heterozygous concordance rates for each variant caller between the raw and corresponding pipeline that includes hard filters applied. With a few exceptions, applying hard filters to variants improved homozygous concordance rates (Table S5). Applying hard filters generally worsened heterozygous concordance rates, except for GATK HC at most minimum coverage requirement levels (Figure 2.4 and Table S6).

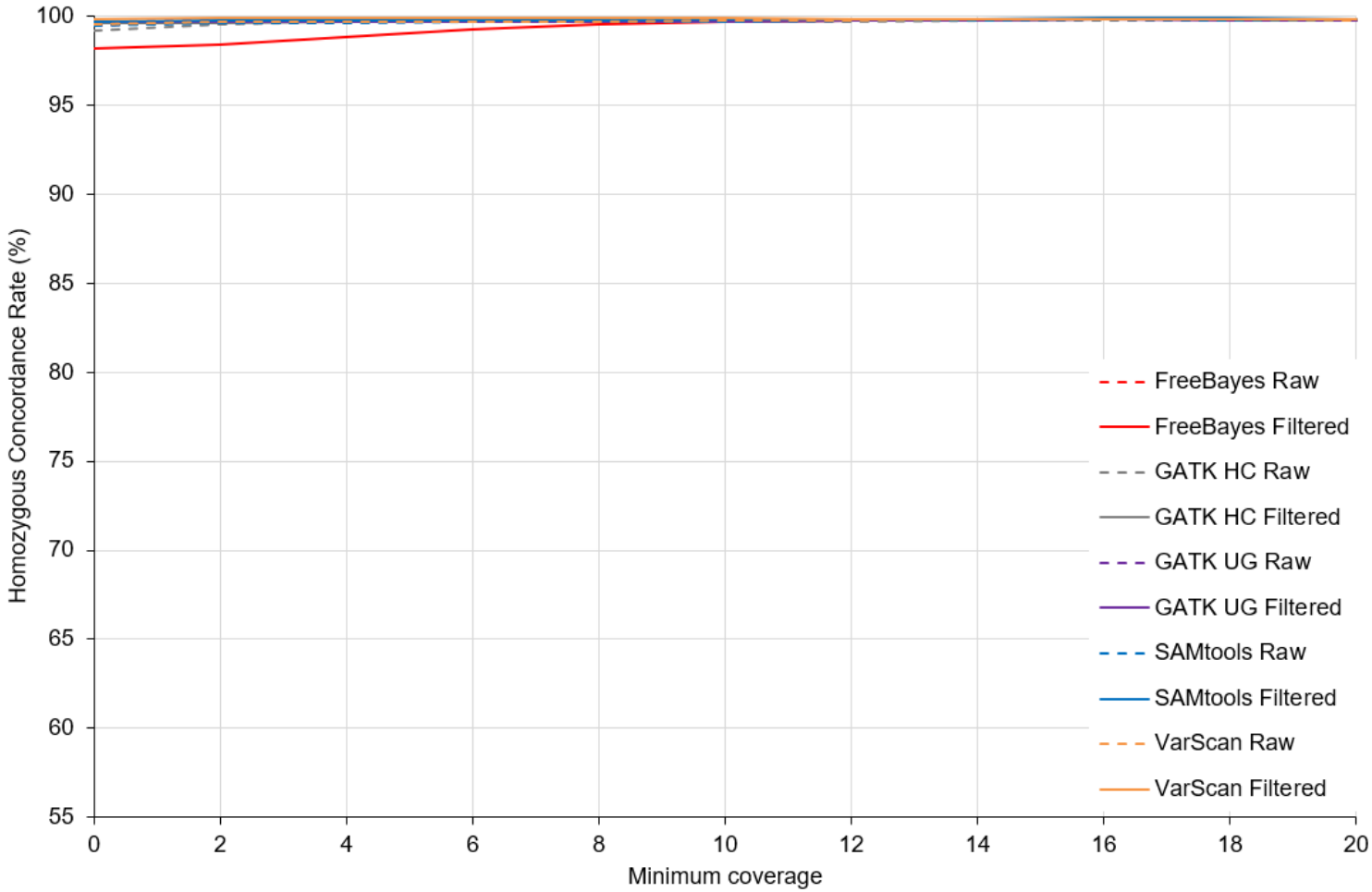

Figure 2.3 Percentage concordance of homozygous genotypes called by raw and filtered pipelines using five different variant callers (FreeBayes, GATK HC, GATK UG, SAMtools and VarScan) compared against genotypes obtained using the CanineHD BeadChip array.

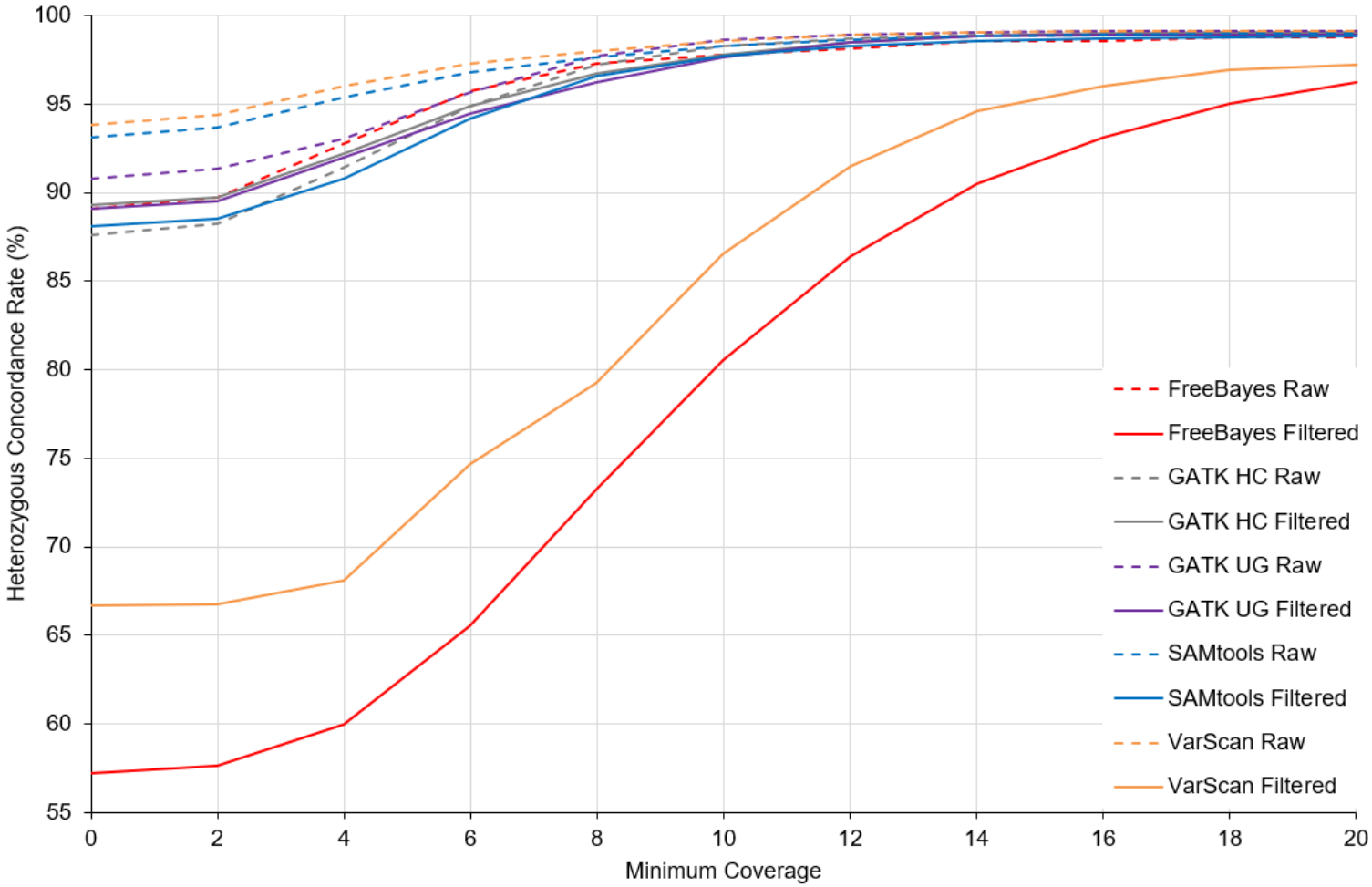

Figure 2.4. Percentage concordance of heterozygous genotypes called by raw and filtered pipelines using five different variant callers (FreeBayes, GATK HC, GATK UG, SAMtools and VarScan) compared against genotypes obtained using the CanineHD BeadChip array.

To compare the differences in sensitivity between homozygous and heterozygous genotypes that were concordant to the truth dataset, we observed the number of homozygous and heterozygous concordant genotypes separately (Table 2.5 and Table 2.6 respectively). In general, FreeBayes-R and VarScan-F had the highest number of homozygous concordant genotypes. We observed that the GATK UG-R and VarScan-R pipelines had the highest number of heterozygous concordant genotypes.

Table 2.5. Total numbers of concordant homozygous genotypes called by raw and filtered pipelines using five variant callers (FreeBayes, GATK HC, GATK UG, SAMtools and VarScan) compared against genotypes obtained using the CanineHD BeadChip array.

Shaded cells designate the pipeline with the highest number of homozygous concordant genotypes at each minimum coverage threshold.

Minimum Coverage	FreeBayes Raw	FreeBayes Filtered	GATK HC Raw	GATK HC Filtered	GATK UG Raw	GATK UG Filtered	SAMtools Raw	SAMtools Filtered	VarScan Raw	VarScan Filtered
0	309,700	191,374	304,445	308,026	315,055	307,296	323,898	304,472	326,176	223,433
2	309,508	191,263	304,444	307,629	315,055	306,937	323,417	304,358	325,688	223,285
4	306,105	190,094	301,543	302,458	307,177	301,756	314,906	298,632	316,919	223,274
6	280,008	183,809	277,551	276,178	279,931	274,900	283,171	274,504	284,540	218,290
8	230,751	166,197	230,337	228,265	231,674	227,109	231,486	227,991	232,427	187,896
10	174,202	137,671	174,946	173,591	175,667	173,440	175,029	173,174	175,603	154,137
12	124,073	105,558	124,672	124,179	125,060	124,276	124,710	123,270	125,008	115,609
14	85,062	76,115	85,277	85,173	85,508	85,217	85,336	84,134	85,499	81,657
16	56,266	52,211	56,443	56,429	56,585	56,435	56,505	55,496	56,591	54,802
18	35,650	33,930	35,675	35,697	35,770	35,680	35,722	34,918	35,775	34,969
20	21,505	20,821	21,524	21,539	21,584	21,522	21,553	21,027	21,582	21,169

Table 2.6. Total numbers of concordant heterozygous genotypes called by raw and filtered pipelines using five different variant callers (FreeBayes, GATK HC, GATK UG, SAMtools and VarScan) compared against genotypes obtained using the CanineHD BeadChip array.

Shaded cells designate the pipeline with the highest number of heterozygous concordant genotypes at each minimum coverage threshold.

Minimum Coverage	FreeBayes Raw	FreeBayes Filtered	GATK HC Raw	GATK HC Filtered	GATK UG Raw	GATK UG Filtered	SAMtools Raw	SAMtools Filtered	VarScan Raw	VarScan Filtered
0	309,700	191,374	304,445	308,026	315,055	307,296	323,898	304,472	326,176	223,433
2	309,508	191,263	304,444	307,629	315,055	306,937	323,417	304,358	325,688	223,285
4	306,105	190,094	301,543	302,458	307,177	301,756	314,906	298,632	316,919	223,274
6	280,008	183,809	277,551	276,178	279,931	274,900	283,171	274,504	284,540	218,290
8	230,751	166,197	230,337	228,265	231,674	227,109	231,486	227,991	232,427	187,896
10	174,202	137,671	174,946	173,591	175,667	173,440	175,029	173,174	175,603	154,137
12	124,073	105,558	124,672	124,179	125,060	124,276	124,710	123,270	125,008	115,609
14	85,062	76,115	85,277	85,173	85,508	85,217	85,336	84,134	85,499	81,657
16	56,266	52,211	56,443	56,429	56,585	56,435	56,505	55,496	56,591	54,802
18	35,650	33,930	35,675	35,697	35,770	35,680	35,722	34,918	35,775	34,969
20	21,505	20,821	21,524	21,539	21,584	21,522	21,553	21,027	21,582	21,169

We compared the difference in specificity between homozygous and heterozygous genotypes that were concordant to the truth dataset by observing the total number of discordant calls for homozygous and heterozygous genotypes separately (Table 2.7 and Table 2.8) . For homozygous concordant genotypes, VarScan-F had the least number for lower to moderate levels of coverage (0 – 12X), whilst SAMtools-F had the least number for higher levels of coverage (14 – 20X). For heterozygous concordant genotypes, VarScan-R had the lowest number for lower levels of minimum coverage requirement levels (0 – 8X). VarScan-R and GATK UG-R had the lowest number of discordant heterozygous concordant genotypes at higher levels of coverage (10 – 20X).

Table 2.7. Total numbers of discordant homozygous genotypes called by raw and filtered pipelines using five variant callers (FreeBayes, GATK HC, GATK UG, SAMtools and VarScan) compared against genotypes obtained using the CanineHD BeadChip array.

Shaded cells designate pipeline with the lowest number of discordant homozygous genotypes at each minimum coverage threshold.

Minimum Coverage	FreeBayes Raw	FreeBayes Filtered	GATK HC Raw	GATK HC Filtered	GATK UG Raw	GATK UG Filtered	SAMtools Raw	SAMtools Filtered	VarScan Raw	VarScan Filtered
0	7,215	24,751	10,565	5,688	3,816	4,728	6,861	4,088	5,857	1,903
2	3,428	20,997	6,190	4,965	3,581	4,077	5,403	3,429	4,573	1,553
4	3,228	15,227	4,405	4,347	3,369	3,784	4,536	2,894	4,027	1,510
6	2,889	8,247	3,739	3,848	3,067	3,468	3,881	2,575	3,526	1,407
8	2,317	4,113	3,028	3,088	2,530	2,843	3,014	2,085	2,742	1,195
10	1,671	2,102	2,262	2,280	1,910	2,113	2,128	1,504	1,933	986
12	1,070	1,132	1,527	1,497	1,272	1,382	1,362	960	1,213	776
14	670	659	966	876	768	812	840	554	749	587
16	461	442	559	476	469	493	513	339	477	412
18	318	307	373	321	326	347	347	218	325	286
20	230	222	266	235	239	247	245	148	236	207

Table 2.8. Total numbers of discordant heterozygous genotypes called by raw and filtered pipelines using five different variant callers (FreeBayes, GATK HC, GATK UG, SAMtools and VarScan) compared against genotypes obtained using the CanineHD BeadChip array.

Shaded cells designate pipeline with the lowest number of discordant heterozygous genotypes at each minimum coverage threshold.

Minimum Coverage	FreeBayes Raw	FreeBayes Filtered	GATK HC Raw	GATK HC Filtered	GATK UG Raw	GATK UG Filtered	SAMtools Raw	SAMtools Filtered	VarScan Raw	VarScan Filtered
0	38,015	143,047	43,230	36,854	31,876	37,586	23,981	41,257	21,539	111,591
2	35,513	140,466	40,512	35,279	29,897	35,994	21,751	39,456	19,334	111,156
4	23,965	126,785	28,417	25,689	22,873	26,413	15,265	30,329	13,152	104,592
6	12,576	96,448	14,874	14,859	12,640	16,119	9,464	17,025	8,045	74,024
8	6,386	60,762	6,639	7,770	5,455	8,913	5,678	8,130	4,711	49,155
10	3,963	33,333	3,070	3,887	2,493	4,140	3,144	4,007	2,561	23,969
12	2,345	16,605	1,627	1,900	1,355	1,908	1,716	2,189	1,410	10,771
14	1,262	7,965	955	991	815	1,019	992	1,251	826	4,640
16	842	3,862	603	593	523	624	607	745	518	2,291
18	447	1,767	381	355	327	387	378	430	323	1,121
20	276	818	230	213	197	234	230	242	199	608

2.5. Discussion

Selecting the best variant calling algorithm and parameters to classify true biological variants from sequencing errors is notoriously difficult. Many variant calling comparison studies that cater towards projects in model species with large datasets and high levels of average coverage (~30X) have been performed for calling SNPs in Illumina NGS data (Bauer 2011; Liu *et al.* 2013; Cheng *et al.* 2014; Cornish and Guda 2014; Pirooznia *et al.* 2014). In this study, we aimed to compare 10 variant calling pipelines that include five variant callers (FreeBayes, GATK HC, GATK UG, SAMtools and VarScan) with and without recommended hard filters applied. We applied the pipelines to data that is representative of a non-model species and smaller study with a lower coverage dataset (~10X). The metrics of performance provided in this study including the genotype concordance rate, total number of concordant and discordant genotypes to estimate sensitivity and specificity can be used as a guide to determine the optimal variant calling pipeline for other small studies with similar datasets.

The VarScan-R pipeline was generally the most accurate as measured by total genotype concordance to the truth dataset ($P_{T\text{-}TEST} < 0.05$, Figure 2.2), achieving concordance rates of 98.37 – 99.37% across the minimum coverage requirements tested (0 – 20X). The VarScan variant caller is identical to the SAMtools pipeline, both using SAMtools' mpileup except that BAQ computation is disabled for VarScan. As the VarScan authors observe (Koboldt *et al.* 2013), we also found that BAQ is too stringent by comparing VarScan-R and SAMtools-R. VarScan-R was generally the most sensitive and specific caller estimated by the highest number of total concordant and lowest number of total discordant genotypes. The superior performance of VarScan-R compared to other nine pipelines is evidently due to its performance compared to the other pipelines at lower minimum coverage requirement levels (less than 10X). At low levels of coverage, the standard deviation of the total number of concordant and discordant genotypes was relatively high. Studies with low average sample coverages should consider the VarScan-R pipeline as it outperformed all other 9 pipelines in genotype concordance, estimated sensitivity and specificity.

At minimum coverage requirement of 10X and over, genotype concordance rates become similar and at 20X each pipeline is within 0.1% of each other, except for VarScan-F and FreeBayes-F which were substantially lower (Table S4). Standard deviation in total concordance and discordance also continually decreases and the difference between the 10 tested pipelines became minimal (Table 2.3 and Table 2.4). GATK UG-R and SAMtools-F had better or similar total accuracy, sensitivity and specificity than VarScan-R at minimum coverage requirement of 10X and higher (Figure 2.2, Table 2.3 and Table 2.4). As the truth dataset comprised of only commonly occurring SNP loci, there is a potential bias where a called variant is more likely to be true, inflating genotype concordances across all pipelines and coverage levels. Subsequent studies should include known rare and *de novo* variants to reduce this source of bias.

Minimum coverage requirement levels had a high impact on the accuracy, sensitivity and specificity of the caller, as has been previously described (Cheng *et al.* 2014). The variance across the 10 pipelines decreased as minimum coverage requirement increased and most pipelines performed quite similarly at higher minimum coverage levels (Table 2.3 and Table 2.4). Besides the variant calling pipeline, the minimum coverage requirement level should be carefully considered depending on the average coverage of the samples and project goals as coverage had the greatest impact on calling sensitivity and specificity (Table S3 contains genotyping rates for each minimum coverage requirement level for each of the 10 pipelines).

Applying the recommended hard filtering criteria to variants generally did not improve the accuracy of genotype concordance to the genotypes of the truth dataset in this study (Figure 2.2). The only exceptions include GATK HC and SAMtools where filtering did improve genotype concordance at some levels of low and high minimum coverage requirement (Figure 2.2 and Table S4). In Figure 2.2, the difference in genotype concordance rate between the raw and pipeline with hard filters applied becomes smaller as minimum coverage requirement increases. We suspect that many of the

developed hard filtering pipelines were developed for samples with higher coverage, where metrics that are used in filtering can be calculated more accurately.

Obtaining high genotyping accuracy (>99.99%) is extremely difficult for relatively low coverage data. As other studies have observed (Sims *et al.* 2014; Willet, Haase, *et al.* 2015a; De Summa *et al.* 2017), we found a higher rate of homozygous than heterozygous concordance to array genotypes, especially when hard filters are applied. Heterozygous genotyping heavily influences the total concordance rate and is dependent on the minimum coverage cut-off value used. Apart from the FreeBayes-F and VarScan-F pipelines, heterozygous concordance rates drastically improve at ≥10X, and become comparable to homozygous concordance rates at ≥12X (the difference between homozygous and heterozygous concordance rates is 0.93 – 1.7%, depending on the variant caller at this coverage level, Figure 2.4). Bias towards homozygous genotypes is evidently caused by low average sample coverage. At low coverage (less than 10X), distinguishing sequencing errors from true alternative variants becomes difficult without the additional support for the alternative allele of multiple reads and we observed that applying hard filters for all five variant calling pipelines was too stringent.

Researchers with low coverage, short read, whole genome sequencing data should select tools and variant calling filtering parameters based on the desired sensitivity and specificity that is appropriate for the research question. For example, when the research question is to identify genetic variants associated with disease, higher sensitivity is more desirable than higher specificity, within reason. From this study, this is achieved by sequencing samples with at least an average depth of 10X to ensure that high genotyping rates are achieved. Applying no additional hard filters generally increases the number of non-reference alleles that are captured. On the contrary, when genotyping accuracy is desired, including hard filters with SAMtools or VarScan can reduce the number of false positive genotypes called.

This study provides reference metrics that can be used to tailor recommended hard filtering pipelines towards specific project goals. Its use would be suitable for projects with small sample sizes and WGS depth (~10X) that wish to call SNPs from Illumina

NGS data. For low coverage data, hard filtering generally reduces sensitivity to detect SNPs, particularly at heterozygous loci. The most value is achieved for samples with a minimum average coverage of 10X per sample, as sensitivities and specificities drastically improve up until this level, where improvement with each additional coverage level begins to plateau.

2.6. References

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. Bioinform. 43: 1-33.

Bauer, D. C., 2011 *Variant calling comparison CASAVA1.8 and GATK*. [online] Nature Precedings. Available at: http://dx.doi.org/10.1038/npre.2011.6107.1 [Accessed 21 Nov. 2018].

Bentley, D. R., S. Balasubramanian, H. P. Swerdlow, G. P. Smith, J. Milton et al., 2008 Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456: 53–59.

Carneiro, M. O., C. Russ, M. G. Ross, S. B. Gabriel, C. Nusbaum et al., 2012 Pacific biosciences sequencing technology for genotyping and variation discovery in human data. BMC Genomics 13: 375.

Cheng, A. Y., Y. Y. Teo, and R. T. H. Ong, 2014 Assessing single nucleotide variant detection and genotype calling on whole-genome sequenced individuals. Bioinformatics 30: 1707–1713.

Chew, T., B. Haase, R. Bathgate, C. E. Willet, M. K. Kaukonen et al., 2017 A Coding Variant in the Gene Bardet-Biedl Syndrome 4 (BBS4) Is Associated with a Novel Form of Canine Progressive Retinal Atrophy. G3 (Bethesda). 7: 2327–2335.

Cornish, A., and C. Guda, 2014 A Comparison of Variant Calling Pipelines Using Genome in a Bottle as a Reference. Biomed Res. Int. 2015: 456479.

DePristo, M. A., E. Banks, R. Poplin, K. V Garimella, J. R. Maguire et al., 2011 A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. 43: 491–498.

Ekblom, R., and J. Galindo, 2011 Applications of next generation sequencing in molecular ecology of non-model organisms. Heredity (Edinb). 107: 1–15.

Faust, G. G., and I. M. Hall, 2014 SAMBLASTER: fast duplicate marking and structural variant read extraction. Bioinformatics 30: 2503–2505.

Forgetta, V., G. Leveque, J. Dias, D. Grove, R. Lyons et al., 2013 Sequencing of the Dutch elm disease fungus genome using the Roche/454 GS-FLX titanium system in a comparison of multiple genomics core facilities. J. Biomol. Tech. 24: 39–49.

Garrison, E., 2015 Freebayes in Depth : Model , Filtering , and Walk-Through. Presented at the University of Cambridge, May 2015. Available at: https://wiki.uiowa.edu/download/attachments/145192256/erik%20garrison%20-%20iowa%20talk%202.pdf?api=v2.

Garrison, E., and G. Marth, 2012 Haplotype-based variant detection from short-read sequencing. https://arxiv.org/abs/1207.3907v2.

Gilly, A., K. Kuchenbaecker, L. Southam, D. Suveges, R. Moore et al., 2017 Very low depth whole genome sequencing in complex trait association studies. bioRxiv 169789.

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

Hwang, S., E. Kim, I. Lee, and E. M. Marcotte, 2015 Systematic comparison of variant calling pipelines using gold standard personal exome variants. Nat. Publ. Gr. 5: 17875.

Koboldt, D. C., D. E. Larson, and R. K. Wilson, 2013 Using VarScan 2 for Germline Variant Calling and Somatic Mutation Detection. Curr. Protoc. Bioinforma. 44: 15.4.1-17.

Layer, R. M., C. Chiang, A. R. Quinlan, and I. M. Hall, 2014 LUMPY: a probabilistic framework for structural variant discovery. Genome Biol. 15: R84.

Le, S. Q., and R. Durbin, 2011 SNP detection and genotyping from low-coverage sequencing data on multiple diploid samples. Genome Res. 21: 952–960.

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

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

Liu, X., S. Han, Z. Wang, J. Gelernter, B.-Z. Yang et al., 2013 Variant Callers for Next-Generation Sequencing Data: A Comparison Study. PLoS One 8: e75619.

Loman, N. J., R. V Misra, T. J. Dallman, C. Constantinidou, S. E. Gharbia et al., 2012 Performance comparison of benchtop high-throughput sequencing platforms. Nat. Biotechnol. 30: 434–439.

Mardis, E. R., and S. Salzburg, 2008 The impact of next-generation sequencing technology on genetics. Trends Genet. 24: 133–141.

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: 1297–1303.

Minoche, A. E., J. C. Dohm, and H. Himmelbauer, 2011 Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. Genome Biol. 12: R112.

Nielsen, R., J. S. Paul, A. Albrechtsen, and Y. S. Song, 2011 Genotype and SNP calling from next-generation sequencing data. Nat. Rev. Genet. 12: 443–451.

O'Rawe, J., T. Jiang, G. Sun, Y. Wu, W. Wang et al., 2013 Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. Genome Med. 5: 28.

Pirooznia, M., M. Kramer, J. Parla, F. S. Goes, J. B. Potash et al., 2014 Validation and assessment of variant calling pipelines for next-generation sequencing. Hum. Genomics 8: 14.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira et al., 2007 PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81: 559–575.

Schuster, S. C., 2007 Next-generation sequencing transforms today's biology. Nat. Methods 5: 16–18.

Sims, D., I. Sudbery, N. E. Ilott, A. Heger, and C. P. Ponting, 2014 Sequencing depth and coverage: key considerations in genomic analyses. Nat. Rev. Genet. 15: 121–132.

De Summa, S., G. Malerba, R. Pinto, A. Mori, V. Mijatovic et al., 2017 GATK hard filtering: tunable parameters to improve variant calling for next generation sequencing targeted gene panel data. BMC Bioinformatics 18: 119.

Tattini, L., R. D'Aurizio, and A. Magi, 2015 Detection of Genomic Structural Variants from Next-Generation Sequencing Data. Front. Bioeng. Biotechnol. 3: 92.

Timmerman, L., 2015 *DNA Sequencing Market Will Exceed $20 Billion, Says Illumina CEO Jay Flatley.* [online] Forbes. Available at: https://www.forbes.com/sites/luke timmerman/2015/04/29/qa-with-jay-flatley-ceo-of-illumina-the-genomics-company-pursuing-a-20b-market/#3b8e8b6a42e7 [Accessed 21 Nov. 2018].

Wang, Y., J. Lu, J. Yu, R. A. Gibbs, and F. Yu, 2013 An integrative variant analysis pipeline for accurate genotype/haplotype inference in population NGS data. Genome Res. 23: 833–842.

Willet, C. E., B. Haase, M. A. Charleston, and C. M. Wade, 2015a Simple, rapid and accurate genotyping-by-sequencing from aligned whole genomes with ArrayMaker. Bioinformatics 31: 599–601.

Willet, C. E., B. Haase, M. A. Charleston, and C. M. Wade, 2015b Simple, rapid and accurate genotyping-by-sequencing from aligned whole genomes with ArrayMaker. Bioinformatics 31: 599–601.

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: e0117055.

Chapter 3. Direct estimate of the *de novo* mutation rate in the domestic dog

3.1. Abstract

All genetic variation that drives evolution and contributes to disease once arose from a spontaneously occurring new DNA mutation. In this chapter, we characterise the rate and distribution of autosomal germline mutations in one of the most phenotypically diverse species, the domestic dog. By characterising *de novo* mutations, their contributions to canine health and evolution can be better understood. There are currently over 400 recognised dog breeds, many of which were created only in the last couple of centuries. Through parent-offspring whole genome sequencing, we estimate the probability of *de novo* mutation to be 3.9×10^{-9} per nucleotide per generation. This corresponds to 81 – 112 new nucleotide mutations in each individual canine genome that is 2.4×10^{9} nucleotides in size. The observed transition to transversion ratio in the canine is 2.3 units, like other vertebrate species. The rate of *de novo* mutations per generation is slightly higher in the dog than the rate of all other studied species including humans, mice, chimpanzees and birds. We theorize that the elevated *de novo* mutation rate may have contributed to the rapid phenotypic diversification of the domestic dog.

3.2. Introduction

The dog (*Canis lupus familiaris*) is believed to be the first animal to be domesticated and today there are over 400 recognised breeds that were developed for a variety of social and economic purposes (Karlsson and Lindblad-Toh 2008; Axelsson *et al.* 2013). The sole ancestor of the domestic dog is the grey wolf (*Canis lupus*) (Vilà *et al.* 1997). The timing, location and process of dog domestication has been heavily debated and many studies have been performed to understand how the species evolved to become one of the most phenotypically diverse living land animals (Vonholdt *et al.* 2010; Boyko *et al.* 2010; Larson *et al.* 2012; Callaway 2013; Axelsson *et al.* 2013; Freedman *et al.*

2014). Whilst many of the previous studies have provided extensive insight into understanding how rapid canine evolution occurred from the perspective of existing ancestral variation and intense artificial selection, no study has yet considered the contribution of *de novo* mutations to canine evolution.

The earliest archaeological evidence of dog domestication include the discovery of ~32,000 to 36,000 year old dog-like fossil remains in Siberia, Belgium and the Czech Republic (Germonpré *et al.* 2009b; Ovodov *et al.* 2011; Germonpré *et al.* 2015). However, it is uncertain whether the fossils represent domestic dogs, animals from failed attempts at domestication, or simply rare, morphologically unique extant wolves (Freedman *et al.* 2014). Dog fossils found at burial sites in Israel and Germany are regarded as more indicative of domestication because their burial reflects their importance in human civilization at the time. Buried canine fossils are dated to be between 11,500 to 16,000 years old (Davis and Valla 1978; Boyko 2011). Studies of wolf and dog mitochondrial DNA variation have previously suggested that dogs were domesticated over 100,000 years ago (Vilà *et al.* 1997; Wayne and Ostrander 1999). Later genetic studies on genomic SNP variation indicate that it is more probable that dogs were domesticated from populations of wolves of either Middle Eastern or Southeast Asian origin only 10,000 years ago (Pang *et al.* 2009; Vonholdt *et al.* 2010). Such genetic studies rely on assumptions of the number of founding events and levels of admixture between wolves, but the true values of these cannot be known for certain. It is likely that the process of domestication was long and complex, involving multiple ancestral populations and multiple back crossing events with wolves.

Although an agreement on the origins of the dog has not yet been reached, it is evident through observed patterns of linkage disequilibrium (LD) that there were two significant bottlenecking events in dog evolutionary history (Lindblad-Toh *et al.* 2005; Boyko 2011). The first bottleneck reflects the initial domestication of dogs from wolves. The creation of modern dog breeds was brought about in the second bottlenecking event, involving intense artificial selection and breeding within closed populations (Lindblad-Toh *et al.* 2005). The second bottlenecking event has only occurred in the last few centuries and has resulted in more than 400 breeds that are recognised worldwide today (Dreger *et al.*

2016). Numerous specialized breeds were formed to suit a specific purpose such as for guarding, herding, retrieving, hunting and racing. Many other breeds were developed for aesthetic and behavioural traits suited for companionship.

The two significant bottlenecking events have led to unique patterns of LD in the dog. LD extending several megabases can be found when analysing dogs of a single breed. Dogs across multiple breeds share a much shorter range of LD that only extends tens of kilobases (Lindblad-Toh *et al.* 2005; Stern *et al.* 2013; Friedenberg and Meurs 2016). This unique genetic architecture of the dog has led scientists to recognise the advantages of mapping traits more efficiently in this species, as fewer individuals and genetic markers are required compared to other species (Karlsson *et al.* 2007). Genome wide association studies (GWAS) have led to the successful mapping of many genes underlying a variety of canine phenotypes, some of which are shared across breeds through introgressive breeding. Gene variants or haplotypes in *FGF4* for Dwarfism (chondrodysplasia), *THBS2* for short-snouts (brachycephaly) and *MSRB3* for floppy ears are examples of successful mapping of traits shared across multiple breeds through GWAS (Parker *et al.* 2009; Bannasch *et al.* 2010; Boyko *et al.* 2010; Boyko 2011).

Despite these successes, there is an unexpectedly large number of Mendelian traits that have no reported underlying causal variant (~23% of reported Mendelian traits have no known underlying causal variant; OMIA, 2018). Similar figures are observed in humans (32%; OMIM, 2018), despite humans being one of the most comprehensively studied species. Many of these unmapped variants are thought to be *de novo* mutations which are not in LD with common genetic markers and hence cannot be found through GWAS (Chong *et al.* 2015). *De novo* mutations have been implicated in the spontaneous occurrence of several canine phenotypes including visible traits such as white spotting in subpopulations of German Shepherd dogs and Weimaraners (Gerding *et al.* 2013; Wong *et al.* 2013); and spontaneously occurring diseases such as ichthyosis, bleeding disorders and progressive retinal atrophy (Brooks 1999; Vilboux *et al.* 2008; Kropatsch *et al.* 2016; Bauer *et al.* 2017). It is likely that many other *de novo* variants cause more subtle influences on phenotype and thus remain undetected.

To understand the contributions of *de novo* mutations to the processes of canine evolution and disease, we aim to characterise the germline *de novo* mutation rate and their distribution in the autosomes of the domestic dog. We identify single nucleotide variant (SNV) *de novo* mutations, estimate the per base mutation rate per generation, estimate the transition (Ti) to transversion (Tv) rates, and characterise new mutations according to the genomic feature in which they reside in. We do this through Illumina whole genome sequencing of five unique parent-offspring trios, comprised of 12 individuals of three purebred dog breeds.

3.3. Materials and Methods

3.3.1. Samples

Twelve individuals that form five unique father-mother-offspring trios were used in this study (see Table S1 for sample information). Samples included dogs from purebred Australian Cattle Dog, Labrador and Miniature Schnauzer families. Some of the offspring are full or half siblings and have one or both parents in common (Figure 3.1).

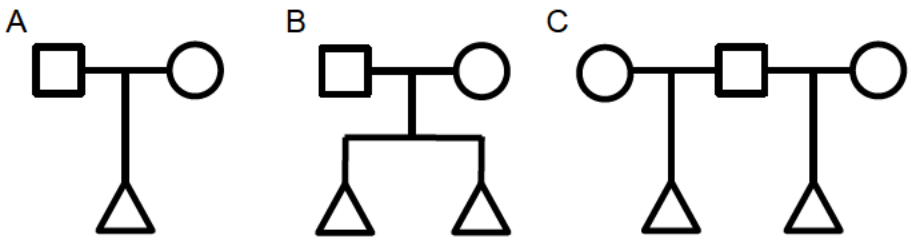

Figure 3.1. Parent-offspring trio configurations.
Squares represent fathers, circles represent mothers and triangles represent offspring. (A) A simple father-mother-offspring trio configuration formed by the Australian Cattle Dog individuals used in this study. (B) Two full siblings with both parents in common, forming two unique parent-offspring trios. The four Miniature Schnauzer samples represents this configuration. (C) Two half siblings with one parent in common. The five Labrador samples are of this configuration representing

two unique parent-offspring trios. Each unique parent-offspring trio can be referred to by the identification number (ID) of the child: USCF134, USCF136, USCF1014, USCF1119 and USCF1294. See Table S1 for pedigree information.

Genomic DNA was extracted from EDTA-stabilized whole blood obtained from the 12 samples using the phenol-chloroform method or the illustra Nucleon BACC2 kit following the manufacturer's protocol (GE Healthcare). This research was conducted with the consent of the animal's owners and with animal ethics approval granted by the Animal Ethics Committee at the University of Sydney (approval number N00/9–2009/3/5109 and N00/10-2012/3/5837 2015/902).

3.3.2. Whole genome sequencing

Whole genome sequencing was performed for each sample using the Illumina HiSeq 2000 (Illumina, San Diego, CA) by the Ramaciotti Centre at the University of New South Wales, Kensington. Libraries were prepared with the Illumina PCR-free TruSeq kit according to the vendor's instructions. Each sample was sequenced as 100-101 base pair (bp), paired-end reads using either half or a full lane of the flow cell.

All bioinformatics analysis was performed on the University of Sydney's High-Performance Computing Cluster (Artemis). Raw reads were aligned to the canine reference genome (CanFam 3.1) as pairs using the Burrows-Wheeler Alignment (BWA)-mem algorithm version 0.7.15 with default parameters (Li and Durbin 2009). Polymerase chain reaction (PCR) duplicates were marked with samblaster, version 0.1.22 (Faust and Hall 2014). Local realignment around insertion-deletions (indels) was performed using the Genome Analysis Toolkit (GATK), version 3.6.0 (McKenna *et al.* 2010; DePristo *et al.* 2011). The number of mapped and unmapped reads was obtained using SAMtools idxstats.

3.3.3. Variant calling and genotyping

To ensure a high level of variant calling accuracy, we obtained sites where genotypes were concordant between two popular variant callers: GATK (version 3.6.0) and SAMtools (version 1.6) (Li *et al.* 2009; McKenna *et al.* 2010). Both of these callers are consistently found to be the best amongst other popular callers at accurately genotyping SNVs in Illumina data when used in conjunction with BWA-mem as the aligner (Cheng *et al.* 2014; Cornish and Guda 2014; Hwang *et al.* 2015). Raw variants at all sites were first called with GATK's Haplotype Caller (HC) (McKenna *et al.* 2010; Van der Auwera *et al.* 2013). The minimum phred-scaled emission and calling confidence threshold was set at 50, which is higher than the recommended values of 10 and 30 respectively (Van der Auwera *et al.* 2013). We chose a higher calling confidence to obtain highly confident genotype calling accuracy. GATK HC raw SNPs were excluded using GATK's VariantFiltration tool if Quality Depth < 2.0, Fisher Strand > 60.0, Mapping Quality < 40.0, HaplotypeScore > 13.0, MappingQualityRankSum < -12.5 and ReadPosRankSum < -8.0, as previously recommended (Van der Auwera *et al.* 2013). SAMtools mpileup and bcftools (version 1.6) was used to call and genotype SNPs, excluding bases and reads with base quality < 20 and mapping quality < 20. Only properly paired reads were considered. As recommended by SAMtools, a coefficient of 50 was applied to reduce the effect of reads with excessive mismatches. Using vcffilter (version 1.0.0), we further filtered SAMtools mpileup SNP calls and excluded sites with QUAL < 50 and MQ < 40.

A single, high quality set of genotypes for each locus and individual were obtained for sites if genotypes were concordant between both filtered GATK HC and SAMtools callsets. This was obtained using bcftools isec with default parameters. Sites were retained if coverage was greater than or equal to 10, and less than or equal to two times the average coverage of the individual. The maximum coverage is applied to avoid regions with duplications as previously recommended (Willet, Haase, *et al.* 2015b). We filtered each locus by coverage using vcflib's vcffilter (version 1.6).

3.3.4. Direct estimate of the per base mutation rate in dogs

To estimate the per base mutation rate in dogs, sites passing filters in variant calling were further filtered by genotype. Sites where both parents were homozygous for the reference allele and where the child was either homozygous reference (non-*de novo* site) or heterozygous (*de novo* site) were obtained using vcflib's vcffilter tool. We term these sites the total number of observable sites passing all filtering requirements used in this study. After visual inspection of potential *de novo* sites using SAMtools tview (Li *et al.* 2009), we noticed that some parents contained poor quality alternative bases despite being called as homozygous reference. As these are more likely to represent non-*de novo* sites, we manually excluded these from further analysis.

In each trio, we defined the per base mutation rate to be:

$$\frac{Number\ of\ \ observed\ de\ novo\ SNP\ sites\ passing\ filters}{Total\ number\ of\ observed\ de\ novo\ and\ non-denovo\ sites\ passing\ filters}$$

De novo mutation events were categorised as either Ti or Tv events and a transition:tranversion (Ti:Tv) rate was calculated.

3.3.5. Characterising *de novo* mutations

We characterised sites that passed all quality filters according to their occurrence within any of seven local genomic features: coding exonic sequence (CDS), CpG island, intergenic, intronic, conserved, 3' untranslated region (3' UTR) and the 5' untranslated region (5' UTR). Using UCSC's Table Browser (https://genome.ucsc.edu/), we obtained CDS, intronic, 3' UTR and 5' UTR regions using the refGene and xenoRefGene tracks in BED file format. Intergenic regions were defined as regions in the reference genome that were not already defined as CDS, intronic, 3' UTR and 5' UTR and were obtained using a custom perl script. Conserved regions of the genome were defined as regions with a phastCons score of > 0.5, calculated using reference genomes of 33 placental mammals. PhastCons scores that were calculated relative to the human genome (GRCh37/hg19) were obtained from UCSC (Pollard *et al.* 2010). Positions were

converted to the CanFam 3.1 reference genome using UCSC's LiftOver tool. We determined whether the per base mutation rate was significantly different between each feature by performing a paired, two tailed t-test.

3.4. Results

3.4.1. Whole genome sequencing

Sequencing on the Illumina HiSeq 2000 platform for the 12 samples used in this study produced between 175,095,946 – 469,840,272 reads, 98.2 – 99.6% of which were aligned to the CanFam 3.1 reference genome. This is equivalent to an average raw coverage of 6.6 – 17.9X per individual (Table S2). The average raw coverage per unique parent-offspring trio ranged between 8.4 – 13.5X.

3.4.2. Variant calling and per base mutation rate estimates

The number of observable loci that passed all filtering criteria used in this study ranged between 64,397,375 – 1,010,866,409 bp for the five trios observed which corresponds to 2.9 – 45.9% of the canine reference autosomes (~2.2 gigabases, Table S3 contains the number of observable loci per trio, per autosome). The number of *de novo* mutations detected ranged from 3 – 51 nucleotide variants for each offspring in the five unique parent-offspring trio (Table S4 contains physical position and genotypes for parent-offspring trios at observed *de novo* sites). There was one identical *de novo* mutation identified between full siblings USCF134 and USCF136 at chromosome 18 position 28,418,247 (CanFam 3.1).

The per-base mutation rate was estimated to be 3.9×10^{-8} (95% confidence interval 3.5 – 4.4×10^{-8}) per meiosis (Table 3.1). This is equivalent to 81 – 112 nucleotide mutations in the canine genome (2.4 gigabases in size). The average Ti:Tv rate was estimated to be 2.3 (95% confidence interval 1.3 – 3.3). The trio that included USCF1119 as the child was excluded from Ti:Tv analyses as no transversions were identified.

Table 3.1. Per base mutation, transition and transversion rate estimates for the domestic dog in five unique parent-offspring trios.

The trio including the offspring USCF1119 was not included in the average transition and transversion rate as no transversions were detected.

Offspring identifier in Trio observed	Per base mutation rate estimate per generation	Ti	Tv	Ti:Tv
USCF134	4.3×10^{-8}	12	8	1.5
USCF136	3.9×10^{-8}	6	4	1.5
USCF1014	3.4×10^{-8}	22	7	3.1
USCF1119	4.7×10^{-8}	3	0	NA
USCF1294	3.5×10^{-8}	26	9	2.9
Average	3.9×10^{-8}	14	6	2.3
Minimum	3.4×10^{-8}	3	0	1.5
Maximum	4.7×10^{-8}	26	9	3.1
Standard Deviation	4.9×10^{-9}	8.9	3.3	0.75

Of the 97 *de novo* mutations collectively observed in the five parent-offspring trios, 71.1% were transition mutations (Figure 3.2). Transition mutations constitute mutations between either A and G nucleotides representing 33.0% of those observed in our data or mutations between C and T nucleotides that represented 38.1% of observed

occurrences of transitions in our data. The remaining 28.9% of all *de novo* mutation events comprised transversion mutations. Transversions include mutations between A and C nucleotides (9.3%), A and T (9.3%), C and G (1.0%) and G and T (9.3%) mutations.

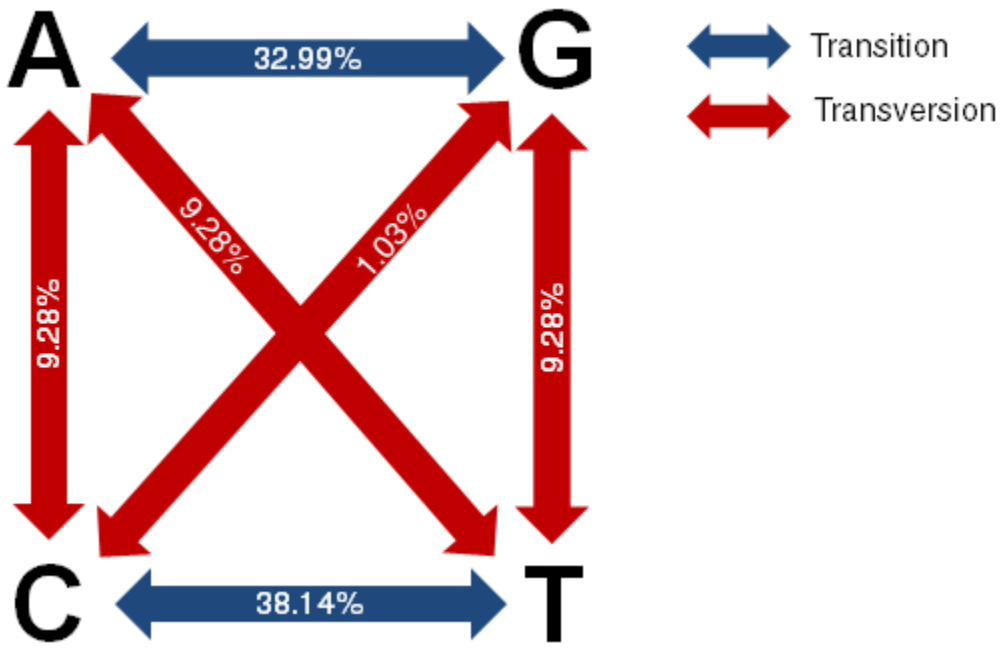

Figure 3.2. Percentage of transition and transversion mutations observed in four parent-offspring trios.

3.4.3. Characteristics of observed *de novo* mutations

Observed *de novo* and non-*de novo* mutation events were categorised as representing any of seven genomic features (see previous description) and the per-base *de novo* mutation rate was calculated for each genomic feature category (Table 3.2). Total number of event observations for each genomic feature can be obtained in Table S5.

Table 3.2. Per base mutation rate estimates (x 10^{-8}) within coding, CpG islands, intergenic, intronic, conserved, 3' UTR and 5' UTR features in dogs using five unique parent-offspring samples.

Offspring identifier in Trio	Coding	CpG Island	Intergenic	Intronic	Conserved	3' UTR	5' UTR
USCF134	0	0	5.2	4.1	8.4	0	0
USCF136	0	0	5.6	2.9	8.4	0	0
USCF1014	1.3	9.1	3.2	3.2	2.3	0	0
USCF1119	0	0	3.2	3.9	0	0	0
USCF1294	1.1	0	4.4	2.2	2.0	0	4.4
Average	4.8	1.8	4.3	3.3	4.2	0	8.7

A paired, two tailed t-test was performed to determine if the per base mutation rate was significantly higher or lower between the seven genomic features observed. The total number of loci observed for each feature can be found in Table S5. The 3' UTR feature had significantly less mutations per base than both intergenic and intronic features ($P < 0.05$).

Table 3.3. P-values obtained from paired, two tailed t-tests performed between seven genomic features to determine if the per base mutation rate was significantly different between each feature in the dog.

	Coding	CpG Island	Intergenic	Intronic	Conserved	3' UTR	5' UTR
Coding		0.458	0.894	0.657	0.900	0.179	0.627
CpG Island			0.494	0.458	0.900	0.374	0.691
Intergenic				0.178	0.948	**0.001**	0.639
Intronic					0.608	**0.001**	0.575
Conserved						0.074	0.657
3' UTR							0.374

3.5. Discussion

The result of this study was a per-base, per-generation germline mutation rate in dogs of 3.9×10^{-8} (95% confidence interval $3.5 – 4.4 \times 10^{-8}$), which is slightly higher than rates estimated for other vertebrates including humans, chimpanzees, laboratory mice and birds (Table 3.4) (Campbell and Eichler 2013; Venn *et al.* 2014; Uchimura *et al.* 2015; Smeds *et al.* 2016; Narasimhan *et al.* 2017). The practical consequence is an expectation of approximately 81 – 112 *de novo* nucleotide changes (also called "private mutations") in each individual genome. The number of transitions outnumber the number of transversions (Ti:Tv) by 2.3 fold, (95% confidence interval 1.3 – 3.3). This

figure is similar to estimated Ti:Tv rates of other vertebrate species (Table 3.4) (Campbell and Eichler 2013; Venn *et al.* 2014; Uchimura *et al.* 2015; Smeds *et al.* 2016; Narasimhan *et al.* 2017).

Table 3.4. Relative predicted *de novo* mutation rate estimates for dogs, humans, mice, chimpanzees and birds.

Human, mice, chimpanzee and bird figures were obtained from several recent studies (Campbell and Eichler 2013; Venn *et al.* 2014; Uchimura *et al.* 2015; Smeds *et al.* 2016; Narasimhan *et al.* 2017).

	Dogs	Humans	Mice	Chimpanzees	Birds
Per base per generation mutation rate	3.93×10^{-8}	$1 – 3 \times 10^{-8}$	5.4×10^{-9}	1.2×10^{-8}	4.6×10^{-9}
Ti: Tv	2.3	2.2	2.1	2.2	2.7

It is plausible that elevated mutation rate in the dog in addition to relatively large litter sizes (mean litter size is 5.4 for purebred dogs) (Sverdrup Borge *et al.* 2011) and shorter generation times in comparison to other studied species and domesticated animals may have facilitated more rapid phenotypic diversification of the dog. We observed a common *de novo* mutation at chromosome 18, position 28,418,247 (CanFam 3.1) in both siblings (USCF134, USCF136) within a nuclear family (Table S1). This could demonstrate that *de novo* mutation events that occur in early stages of spermatogenesis or oogenesis in the parents can propagate to more offspring at a time, resulting in more rapid dissemination of new genetic variation.

Studies in *de novo* mutation rates in humans and chimpanzees have suggested that *de novo* mutations occur more frequently in the offspring as the age at conception of the parents increases, especially of the father (Kong *et al.* 2012; Venn *et al.* 2014). Sperm

are created through multiple rounds of spermatogenesis throughout a mature male's life, unlike in females where oogenesis begins during foetal development. The process of DNA replication in aging individuals is thought to foster more DNA replication mistakes, and hence result in more observable *de novo* mutations in the offspring. Age of conception information was unavailable for the current study. Trios including a wide range of ages of parents at the time their offspring was conceived should be included in future studies to determine parent of origin effects on the mutation rate as a function of age. Additionally, as different breeds of dog are known to have different average lifespans, future studies should be repeated on a per breed bases, and subsequently, an assessment of variance of the mutation rate across breeds can be determined. This may suggest how much the mutation rate could influence the rate of evolution in some breeds compared to others.

Differences in the average mutation rate that may be observed across species is likely to be influenced by technical nuances. This includes sequencing technologies and bioinformatics pipelines employed to process the raw sequencing data. The studies discussed here applied a whole genome approach using Illumina sequencing technologies (Venn *et al.* 2014; Uchimura *et al.* 2015; Smeds *et al.* 2016; Narasimhan *et al.* 2017). However, the sequencing depth varied from approximately 10 to 40X coverage. A coverage of at least 10X is deemed to be sufficient for providing accurate variant calling, however slight improvements are still evident with increasing coverage (Cheng *et al.* 2014). Variant calling accuracy is also variable across variant calling algorithms and quality filtering parameters enforced (for performance metrics see Cheng *et al.* 2014). In order to minimize the effects technical biases when assessing mutation rate differences between species, future studies should ensure that sequencing technologies and bioinformatics pipelines used are consistent for all samples of all species studied. Future studies should also consider a *de novo* assembly rather than a reference genome mapping approach to mitigate potential biases from differences in the quality of reference genomes across species. Reference genomes that are more complete and representative of a population enable improved mappability of raw sequencing reads as mapping is dependent on sequence similarity (Degner *et al.*

2009; Brandt *et al.* 2015). This would therefore affect all downstream processing, including variant calling and *de novo* mutation calling accuracy.

Improved estimates of average mutation rates per species can benefit several practical applications that ultize per base mutation rates. For instance, the timing of species divergence is calculated by using the mutation rate as a molecular clock in the study of ancestral DNA sequences. We provide an alternative means for scientists to estimate dog divergence from the grey wolf and this might inform debate on dog domestication. Better estimates of *de novo* mutation rate and transition to transversion ratios are expected to improve future *de novo* detection studies via the more accurate application of priors in computational prediction algorithms (Ramu *et al.* 2013; Francioli *et al.* 2017). *De novo* mutations are never in LD with genetic markers and are notoriously difficult to map through GWAS. Hence, accurate identification and calling of *de novo* mutations is essential for clinical diagnosis for patients with spontaneously occurring genetic diseases.

To better understand the potential effects on phenotype and evolution arising from the of *de novo* mutations that we identified, we categorised the observed variants into any of seven genomic features: protein-coding, CpG island, intergenic, intronic, conserved, 3' UTR and the 5' UTR. We did not detect any significant difference in the mutation rate between each feature, other than mutations in 3' UTR regions are significantly less common compared with those occurring in intergenic and intronic features ($P_{T\text{-}TEST} < 0.05$, Table 3.3). We observed no difference in the per-base mutation rate in CpG dinucleotide islands compared with other genomic features. This is an unexpected finding as methylated CpG islands have previously reported to be more mutagenic (Cooper and Youssoufian 1988). In humans, the per-base mutation rate has been observed to be 10 – 18 times higher in CpG dinucleotides compared with non-CpG dinucleotides (Kondrashov 2002; Lynch 2010; Kong *et al.* 2012; Narasimhan *et al.* 2017). Higher mutation rates in CpG islands have also been observed in other animals including apes (30 times higher), mammals (15 times higher) and birds (10 times higher) (Keightley *et al.* 2011; Hodgkinson and Eyre-Walker 2011; Smeds *et al.* 2016).

The disparity in relative CpG island mutation rates in dogs compared to other mammals are most likely caused by technical limitations of this study. Next generation sequencers such as Illumina platforms rely on PCR amplification techniques that require specific optimisation in order to successfully sequence GC rich DNA (Reuter *et al.* 2015). The sequencing difficulty was evident in this study, with only ~5.5 million high-quality observed nucleotides occurring in CpG islands considered compared with 37,355,082 that exists in the canine reference genome (Table S5). The rate of C to G transversion (1.0% of all mutations) is also likely to have been affected by this sequencing bias (Figure 3.2). To better characterise relative occurrences of *de novo* mutations across a variety of genomic features, future studies would benefit from the inclusion of parent-offspring trios sequenced at higher coverage. The reliable identification of *de novo* mutations across the whole genome requires a high level of variant calling and genotyping accuracy. Such accuracy and coverage can be achieved through via higher levels of sequencing depth (>30X) (Francioli *et al.* 2017). However, the relatively high costs of sequencing currently impede on the opportunities to perform *de novo* mutation characterisations in non-model species.

To the author's knowledge, we are first to describe and directly observe *de novo* mutation rates in the domestic dog using a genome-wide strategy. The estimated rates reveal an elevated *de novo* mutation rate for canines in comparison to other studied species to date, indicating a possible mechanism for the rapid generation of phenotypic diversity in the evolution of the domestic dog from the grey wolf. We have provided metrics in relation to *de novo* mutation rates in the dog that might be used as a molecular clock. This will enable scientists to better elucidate the timing of dog domestication. Metrics that we have provided can also be used as better priors for use in variant calling algorithms for more accurate genotyping of *de novo* mutations.

3.6. 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: 360–364.

Bannasch, D., A. Young, J. Myers, K. Truvé, P. Dickinson et al., 2010 Localization of Canine Brachycephaly Using an Across Breed Mapping Approach. PLoS One 5: e9632.

Bauer, A., D. P. Waluk, A. Galichet, K. Timm, V. Jagannathan et al., 2017 A de novo variant in the ASPRV1 gene in a dog with ichthyosis. PLoS Genet. 13: e1006651.

Boyko, A. R., 2011 The domestic dog: man's best friend in the genomic era. Genome Biol. 12: 216.

Boyko, A. R., P. Quignon, L. Li, J. J. Schoenebeck, J. D. Degenhardt et al., 2010 A Simple Genetic Architecture Underlies Morphological Variation in Dogs. PLoS Biol. 8: e1000451.

Brandt, D. Y. C., V. R. C. Aguiar, B. D. Bitarello, K. Nunes, J. Goudet et al., 2015 Mapping Bias Overestimates Reference Allele Frequencies at the HLA Genes in the 1000 Genomes Project Phase I Data. G3 (Bethesda). 5: 931–941.

Brooks, M., 1999 A review of canine inherited bleeding disorders: biochemical and molecular strategies for disease characterization and carrier detection. J. Hered. 90: 112–118.

Callaway, E., 2013 Dog genetics spur scientific spat. Nature 498: 282–283.

Campbell, C. D., and E. E. Eichler, 2013 Properties and rates of germline mutations in humans. Trends Genet. 29: 575–584.

Cheng, A., Y. Teo, and R. Ong, 2014 Assessing single nucleotide variant detection and genotype calling on whole-genome sequenced individuals. Bioinformatics 30: 1707–1713.

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: 199–215.

Cooper, D. N., and H. Youssoufian, 1988 The CpG dinucleotide and human genetic disease. Hum Genet 78: 151–155.

Cornish, A., and C. Guda, 2014 A Comparison of Variant Calling Pipelines Using Genome in a Bottle as a Reference. Biomed Res. Int. 2015: 456479.

Davis, J. M., and F. R. Valla, 1978 Evidence for domestication of the dog 12,000 years ago in the Natufian of Israel. Nature 276: 608–610.

Degner, J. F., J. C. Marioni, A. A. Pai, J. K. Pickrell, E. Nkadori et al., 2009 Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. Bioinformatics 25: 3207–3212.

DePristo, M. A., E. Banks, R. Poplin, K. V Garimella, J. R. Maguire et al., 2011 A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. 43: 491–498.

Dreger, D. L., B. W. Davis, R. Cocco, S. Sechi, A. Di Cerbo et al., 2016 Commonalities in Development of Pure Breeds and Population Isolates Revealed in the Genome of the Sardinian Fonni's Dog. Genetics 204: 737–755.

Faust, G. G., and I. M. Hall, 2014 SAMBLASTER: fast duplicate marking and structural variant read extraction. Bioinformatics 30: 2503–2505.

Francioli, L. C., M. Cretu-Stancu, K. V Garimella, M. Fromer, W. P. Kloosterman et al., 2017 A framework for the detection of de novo mutations in family-based sequencing data. Eur. J. Hum. Genet. 25: 227–233.

Freedman, A. H., I. Gronau, R. M. Schweizer, D. Ortega-Del Vecchyo, E. Han et al., 2014 Genome Sequencing Highlights the Dynamic Early History of Dogs. PLoS Genet. 10: e1004016.

Friedenberg, S. G., and K. M. Meurs, 2016 Genotype imputation in the domestic dog. Mamm. Genome 27: 485–494.

Gerding, W. M., D. A. Akkad, and J. T. Epplen, 2013 Spotted Weimaraner dog due to de novo KIT mutation. Anim. Genet. 44: 605–606.

Germonpré, M., M. Lázničková-Galetová, R. J. Losey, J. Räikkönen, and M. V. Sablin, 2015 Large canids at the Gravettian Předmostí site, the Czech Republic: The mandible. Quat. Int. 359–360: 261–279.

Germonpré, M., M. V. Sablin, R. E. Stevens, R. E. M. Hedges, M. Hofreiter et al., 2009 Fossil dogs and wolves from Palaeolithic sites in Belgium, the Ukraine and Russia: osteometry, ancient DNA and stable isotopes. J. Archaeol. Sci. 36: 473–490.

Hodgkinson, A., and A. Eyre-Walker, 2011 Variation in the mutation rate across mammalian genomes. Nat. Rev. Genet. 12: 756–766.

Hwang, S., E. Kim, I. Lee, and E. M. Marcotte, 2015 Systematic comparison of variant calling pipelines using gold standard personal exome variants. Nat. Publ. Gr. 5: 17875.

Karlsson, E. K., I. Baranowska, C. M. Wade, N. H. C. Salmon Hillbertz, M. C. Zody et al., 2007 Efficient mapping of mendelian traits in dogs through genome-wide association. Nat. Genet. 39: 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: 713–725.

Keightley, P. D., L. Eöry, D. L. Halligan, and M. Kirkpatrick, 2011 Inference of mutation parameters and selective constraint in mammalian coding sequences by approximate Bayesian computation. Genetics 187: 1153–1161.

Kondrashov, A. S., 2002 Direct estimates of human per nucleotide mutation rates at 20 loci causing mendelian diseases. Hum. Mutat. 21: 12–27.

Kong, A., M. L. Frigge, G. Masson, S. Besenbacher, P. Sulem et al., 2012 Rate of de novo mutations and the importance of father's age to disease risk. Nature 488: 471–475.

Kropatsch, R., D. A. Akkad, M. Frank, C. Rosenhagen, J. Altmüller et al., 2016 A large deletion in RPGR causes XLPRA in Weimaraner dogs. Canine Genet. Epidemiol. 3: 7.

Larson, G., E. K. Karlsson, A. Perri, M. T. Webster, S. Y. W. Ho et al., 2012 Rethinking dog domestication by integrating genetics, archeology, and biogeography. Proc. Natl. Acad. Sci. 109: 8878–8883.

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

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

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: 803–819.

Lynch, M., 2010 Rate, molecular spectrum, and consequences of human mutation. Proc. Natl. Acad. Sci. U. S. A. 107: 961–968.

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: 1297–1303.

Narasimhan, V. M., R. Rahbari, A. Scally, A. Wuster, D. Mason et al., 2017 Estimating the human mutation rate from autozygous segments reveals population differences in human mutational processes. Nat. Commun. 8: 303.

Ovodov, N. D., S. J. Crockford, Y. V. Kuzmin, T. F. G. Higham, G. W. L. Hodgins et al., 2011 A 33,000-Year-Old Incipient Dog from the Altai Mountains of Siberia: Evidence of

the Earliest Domestication Disrupted by the Last Glacial Maximum (A. Stepanova, Ed.). PLoS One 6: e22821.

Pang, J.-F., C. Kluetsch, X.-J. Zou, A. Zhang, L.-Y. Luo et al., 2009 mtDNA data indicate a single origin for dogs south of Yangtze River, less than 16,300 years ago, from numerous wolves. Mol. Biol. Evol. 26: 2849–2864.

Parker, H. G., B. M. VonHoldt, P. Quignon, E. H. Margulies, S. Shao et al., 2009 An expressed fgf4 retrogene is associated with breed-defining chondrodysplasia in domestic dogs. Science 325: 995–998.

Pollard, K. S., M. J. Hubisz, K. R. Rosenbloom, and A. Siepel, 2010 Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. 20: 110–121.

Ramu, A., M. J. Noordam, R. S. Schwartz, A. Wuster, M. E. Hurles et al., 2013 DeNovoGear: de novo indel and point mutation discovery and phasing. Nat. Methods 10: 985–987.

Reuter, J. A., D. V. Spacek, and M. P. Snyder, 2015 High-Throughput Sequencing Technologies. Mol. Cell 58: 586–597.

Smeds, L., A. Qvarnström, and H. Ellegren, 2016 Direct estimate of the rate of germline mutation in a bird. Genome Res. 26: 1211–1218.

Stern, J. A., S. N. White, and K. M. Meurs, 2013 Extent of linkage disequilibrium in large-breed dogs: chromosomal and breed variation. Mamm. Genome 24: 409–415.

Sverdrup Borge, K., R. Tønnessen, A. Nødtvedt, and A. Indrebø, 2011 Litter size at birth in purebred dogs—A retrospective study of 224 breeds. Theriogenology. 15: 911-919.

Uchimura, A., M. Higuchi, Y. Minakuchi, M. Ohno, A. Toyoda et al., 2015 Germline mutation rates and the long-term phenotypic effects of mutation accumulation in wild-type laboratory mice and mutator mice. Genome Res. 25: 1–10.

Venn, O., I. Turner, I. Mathieson, N. de Groot, R. Bontrop et al., 2014 Strong male bias drives germline mutation in chimpanzees. Science. 344: 1272–1275.

Vilà, C., P. Savolainen, J. E. Maldonado, I. R. Amorim, J. E. Rice et al., 1997 Multiple and Ancient Origins of the Domestic Dog. Science. 276: 1687–1689.

Vilboux, T., G. Chaudieu, P. Jeannin, D. Delattre, B. Hedan et al., 2008 Progressive retinal atrophy in the Border Collie: a new XLPRA. BMC Vet. Res. 4: 10.

Vonholdt, B. M., J. P. Pollinger, K. E. Lohmueller, E. Han, H. G. Parker et al., 2010 Genome-wide SNP and haplotype analyses reveal a rich history underlying dog domestication. Nature 464: 898–902.

Wayne, R. K., and E. A. Ostrander, 1999 Origin, genetic diversity, and genome structure of the domestic dog. 21: 247–257.

Willet, C. E., B. Haase, M. A. Charleston, and C. M. Wade, 2015 Simple, rapid and accurate genotyping-by-sequencing from aligned whole genomes with ArrayMaker. Bioinformatics 31: 599–601.

Wong, A. K., A. L. Ruhe, K. R. Robertson, E. R. Loew, D. C. Williams et al., 2013 A de novo mutation in KIT causes white spotting in a subpopulation of German Shepherd dogs. Anim. Genet. 44: 305–310.

Chapter 4. The Genetics of Progressive Retinal Atrophy in the Hungarian Puli

4.1. Synopsis - Exclusion of known progressive retinal atrophy genes for blindness in the Hungarian Puli

In this chapter, we begin to explore the contribution of *de novo* mutations in canine disease and demonstrate how next generation sequencing data can be used to study low frequency variants that are associated with disease. Low frequency variants, such as *de novo* mutations that are involved in rare disease, are difficult to detect with traditional genome wide association analyses. In addition, studies on rare diseases are challenged with having a limited number of case samples. The research in chapter 4 is focussed on progressive retinal atrophy in the Hungarian Puli. In section 4.1, we present published research which reports that this form of disease is potentially novel, by performing comprehensive testing of reported canine progressive retinal atrophy genes. The supplementary materials associated with the original publication have been included in section 4.1.1 to provide the reader with greater context for this chapter.

ANIMAL GENETICS Immunogenetics, Molecular Genetics and Functional Genomics

BRIEF NOTE

doi: 10.1111/age.12553

Exclusion of known progressive retinal atrophy genes for blindness in the Hungarian Puli

Tracy Chew, Bianca Haase, Cali E. Willet and Claire M. Wade

School of Life and Environmental Sciences, Faculty of Veterinary Science, University of Sydney, Camperdown, NSW 2145, Australia

Accepted for publication 23 January 2017

Description: Progressive retinal atrophy (PRA) is a common cause of blindness in many pure and mixed breed dogs (*Canis lupus familiaris*). The typical onset of PRA begins with gradual night vision loss followed by day vision loss due to the death of rod and cone photoreceptors, respectively.[1] There are currently no mutations or genes reported to be causative or associated with PRA in the Hungarian Puli. In this study, we use an extensive list of 53 known PRA genes to screen for putative causal variants in this breed of dog.

Samples: Two half sibling Hungarian Puli dogs (USCF516, USCF519) were diagnosed with PRA at the age of two years by registered specialists in veterinary ophthalmology. Diagnosis was based on ophthalmologic changes observed including vascular attenuation, hyper reflectivity and reduced myelination in the optic nerve head. The dams (USCF347, USCF524) and sire (USCF525) were assessed as clear of PRA. Nine other dogs from the same kennel had normal vision. The pedigree can be found in Fig. S1.

Most identified PRA mutations are inherited in an autosomal recessive manner, except for two that are X linked and one other that has dominant inheritance.[2] The pattern of inheritance for PRA in the Hungarian Puli family is consistent with the autosomal recessive form of inheritance.

Candidate gene screening: We screened for functional variants that followed an autosomal recessive pattern of inheritance in 53 candidate genes (Table S1) in Illumina paired end, whole genome sequencing data of a father mother pro band trio (USCF525, USCF347, USCF516) and one additional half sibling case (USCF519) (sequencing data have been deposited in NCBI's Sequence Read Archive under BioProject accession no. PRJNA344694). Target loci were identified using two affected (USCF516, USCF519) and 12 non PRA affected relatives genotyped at 172 938 SNP markers distributed across the canine genome using the

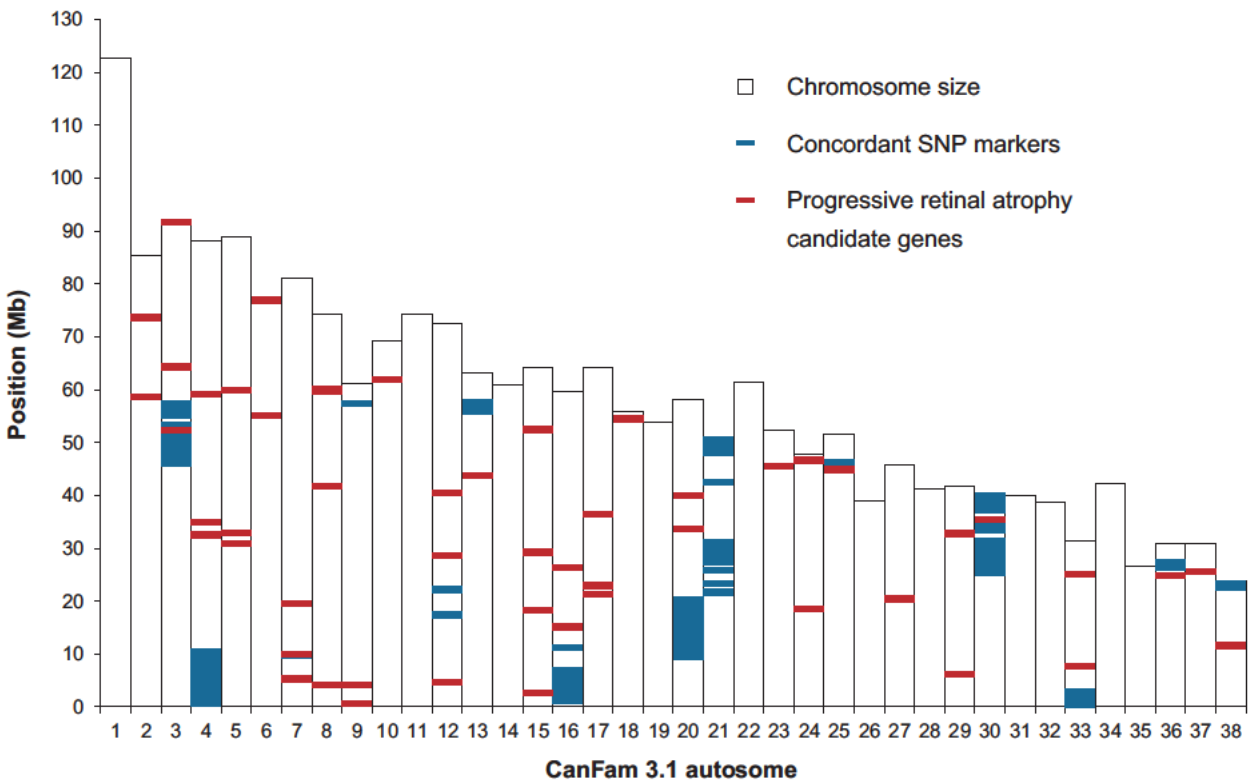

Figure 1 Location of 53 candidate genes and 364 SNP markers that are concordant with autosomal recessive inheritance on the CanFam 3.1 autosomes. Blue indicates the location of SNP markers on the CanineHD BeadChip array that are concordant with the expected inheritance for the two cases and 12 controls including three parents genotyped. Red indicates the locations of the 53 candidate genes. Two genes, *RLPB1* at chr3:52 260 877 52 278 803 and *NR2E3* at chr30:35 378 421 35 381 822 were the only genes that co located with concordant loci (chr3:45 592 262 57 397 690 with 33 concordant markers and chr30:25 254 123 39 976 525 with 103 concordant markers).

CanineHD BeadChip (Illumina) (genotyping array data of all 14 individuals used in this study have been deposited in NCBI's Gene Expression Omnibus under the accession number GSE87642). Coding exons and untranslated regions of genes that co located with loci conforming to an autosomal recessive inheritance pattern (Table S2) were then Sanger sequenced for the two affected dogs. A full description of materials and methods can be found in Appendix S1.

Conclusions: Exhaustive screening of 53 candidate loci in a Hungarian Puli family segregating blindness identified no coding variants for the phenotype of interest. Two candidate genes in loci concordant with recessive inheritance were identified (*RLBP1* position chr3:52 260 877 52 278 803 and *NR2E3* chr30:35 378 421 35 381 822, CanFam 3.1; Fig. 1). Sanger sequencing of exons and untranslated regions revealed no variation to the reference genome. An additional eight recessively inherited SNPs in other regions of the genome were found in *PDE6A, RD3, PRCD* and *MERTK* using whole genome sequencing data; however these were either intronic or non coding variants and are not likely to cause disease (Table S3). This study provides the basis for mapping and further screening of potentially novel canine PRA genes followed by testing in a wider sample cohort.

References

1 Parry H.B. (1953) *Br J Ophthalmol* **37**, 487 502.
2 Miyadera K. *et al.* (2012) *Mamm Genome* **23**, 40 61.

Correspondence: C. M. Wade (claire.wade@sydney.edu.au)

Supporting information

Additional supporting information may be found online in the supporting information tab for this article:
Appendix S1 Materials and methods.
Figure S1 Pedigree of Hungarian Puli segregating progressive retinal atrophy.
Table S1 A list of the PRA candidate genes screened.
Table S2 PCR primer sequences.
Table S3 Putative variants identified from screening 53 candidate genes in parent proband and an affected half sibling case.

4.1.1. Supplementary materials for section 4.1

Materials and Methods

Samples

Two affected half sibling Hungarian Puli dogs (USCF516, USCF519) and 12 other individuals including their parents (USCF347, USCF524, USCF525) from the same kennel were used in this study (Figure S1). USCF516 and USCF519 were diagnosed with progressive retinal atrophy (PRA) by registered specialists in veterinary ophthalmology based on vascular attenuation in the eye, hyper-reflectivity and reduced myelination in the optic nerve head. The parents also underwent testing and were confirmed to be PRA clear. The remaining 9 dogs who were over 5 years of age had normal vision. Dogs with PRA reach complete blindness at 6 months – 4 years, hence we considered these 9 dogs to be PRA clear. EDTA-stabilized blood was collected from all 14 dogs and genomic DNA was extracted using the illustra Nucleon BACC2 kit (GE Healthcare). This study was carried out with the consent of the dog's owners' and with Animal Ethics approval granted by the Animal Ethics Committee at the University of Sydney (N00/9–2009/3/5109).

Candidate gene selection

A comprehensive list of 53 candidate genes was screened for PRA causative variants. Candidates were selected from a toolset that was developed to allow rapid screening of dog families with PRA (Winkler *et al.* 2016). Genes include PRA associated genes identified in multiple dog breeds and genes associated with analogous disease in humans. Additional genes were selected from another PRA-screening study involving multiple dog breeds and a review (Downs *et al.* 2014; Miyadera *et al.* 2012). Exhaustive screening for candidate mutations was performed with two methods – putative variant detection in in whole genome sequencing data and Sanger sequencing exons of genes that resided in loci concordant with autosomal recessive inheritance. The second method ensures that candidate genes within putative loci are completely and accurately

sequenced, especially where there is no – low coverage in whole genome sequencing data.

Whole genome sequencing and putative mutation detection

One father-mother-proband trio (USCF525, USCF347, USCF516) and the additional half sibling case (USCF519) were whole genome sequenced by the Ramaciotti Centre, University of New South Wales, Kensington. Library preparation was performed with the Illumina TruSeq DNA PCR-free kit. The four samples were barcoded and sequenced on two lanes as 101 base paired-end reads on the Illumina HiSeq 2000 (Illumina, San Diego, CA).

For each sample, reads were aligned as pairs using the Burrows-Wheeler Alignment tool with default parameters (Li & Durbin 2009). Polymerase chain reaction (PCR) duplicates were marked using Picard (http://picard.sourceforge.net). Local realignment around insertion-deletions (INDELs) was performed using GATK (McKenna *et al.* 2010; DePristo *et al.* 2011). Raw SNP and INDEL variants were called in the 53 candidate genes (Table S1) using Unified Genotyper provided by GATK (McKenna *et al.* 2010). The VariantFiltration tool was used to filter for high quality variants using recommended hard filtering parameters for small datasets (Van der Auwera *et al.* 2013). SNPs were removed if Quality Depth < 2.0, Fisher Strand > 60.0, Mapping Quality < 40.0, HaplotypeScore > 13.0, MappingQualityRankSum < -12.5 and ReadPosRankSum < -8.0. INDELs were removed if Quality Depth < 2.0, Fisher Strand > 200.00 and ReadPosRankSum < -20.0.

High quality variants that conformed to an autosomal recessive inheritance pattern were retained. Concordant variants were annotated with Variant Effect Predictor (VEP) provided by Ensembl (McLaren *et al.* 2010). Known, common SNPs listed in dbSNP were removed and remaining functional coding variants were considered for testing in a wider study cohort.

Identification of regions concordant with recessive inheritance

The two case (USCF516, USCF519) and 12 control dogs including the PRA clear parents (USCF347, USCF524, USCF525) were genotyped at 172,938 SNP markers using the CanineHD BeadChip array (Illumina, San Diego, CA) by GeneSeek (Lincoln, NE). Markers that were genotyped as homozygous for the minor allele for the cases only were regarded as target loci (concordant). Candidate genes and concordant loci were charted onto the concordance map (Figure 1).

Sanger sequencing of candidate genes in associated loci

USCF516 and USCF519 were screened for putative functional variants in coding exons of candidate genes that resided within concordant loci using PCR and Sanger sequencing. Primers were designed in Primer3 (Rozen & Skaletsky 2000) to amplify *RLBP1* and *NR2E3* exons (primer sequences, melting temperatures and product sizes can be found in Table S2). PCR was carried out using the AmpliTaq Gold 360 Master Mix (Applied Biosystems) in a 20 µL reaction volume. Thermocycling conditions were carried out as follows: denaturation at 95 oC for 15 min; 35 cycles of 95 oC for 30 sec, annealing at melting temperatures (Tm) according to Table S2 for 30 sec, 72 oC for 45 sec; and lastly a final elongation step at 72 oC for 10 min. PCR products were purified by dispensing 7 µL of PCR product into 3 µL of a master mix containing 10x shrimp alkaline phosphatase (SAP) buffer, 1U SAP, 1U Exo I and water. The 10 µL reaction volume was placed into a thermocycler to allow enzymatic activity for 30 min at 37oC, followed by a deactivation period of 15 min at 80oC. Sanger sequencing of purified PCR products was carried out by the Australian Genome Research Facility at Westmead in accordance with the vendor's instructions. Sequences were assessed for any variants alternative to the reference genome.

References

DePristo M.A., Banks E., Poplin R., Garimella K.V., Maguire J.R., Hartl C., Philippakis A.A., del Angel G., Rivas M.A., Hanna M., McKenna A., Fennell T.J., Kernytsky A.M., Sivachenko A.Y., Cibulskis K., Gabriel S.B., Altshuler D. & Daly M.J. (2011) A

framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics* 43, 491–8.

Downs L.M., Hitti R., Pregnolato S. & Mellersh C.S. (2014) Genetic screening for PRA-associated mutations in multiple dog breeds shows that PRA is heterogeneous within and between breeds. *Veterinary Ophthalmology* 17, 126–30.

Li H. & Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–60.

McKenna A., Hanna M., Banks E., Sivachenko A., Cibulskis K., Kernytsky A., Garimella K., Altshuler D., Gabriel S., Daly M. & DePristo M.A. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* 20, 1297–303.

McLaren W., Pritchard B., Rios D., Chen Y., Flicek P. & Cunningham F. (2010) Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* 26, 2069–70.

Miyadera K., Acland G. M. & Aguirre G. D. (2012) Genetic and phenotypic variations of inherited retinal diseases in dogs: The power of within- and across-breed studies. *Mammalian Genome* 23, 40–61.

Rozen S. & Skaletsky H. (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods in Molecular Biology* 132, 365–86.

Van der Auwera G.A., Carneiro M.O., Hartl C., Poplin R., del Angel G., Levy-Moonshine A., Jordan T., Shakir K., Roazen D., Thibault J., Banks E., Garimella K.V., Altshuler D., Gabriel S., DePristo M.A. (2013) From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Current Protocols in Bioinformatics* 43, 11.10.1-11.10.33.

Winkler P. A., Davis J. A., Petersen-Jones S. M., Venta P. J. & Bartoe, J. T. (2016) A tool set to allow rapid screening of dog families with PRA for association with candidate genes. *Veterinary Ophthalmology* 20, 372-376.

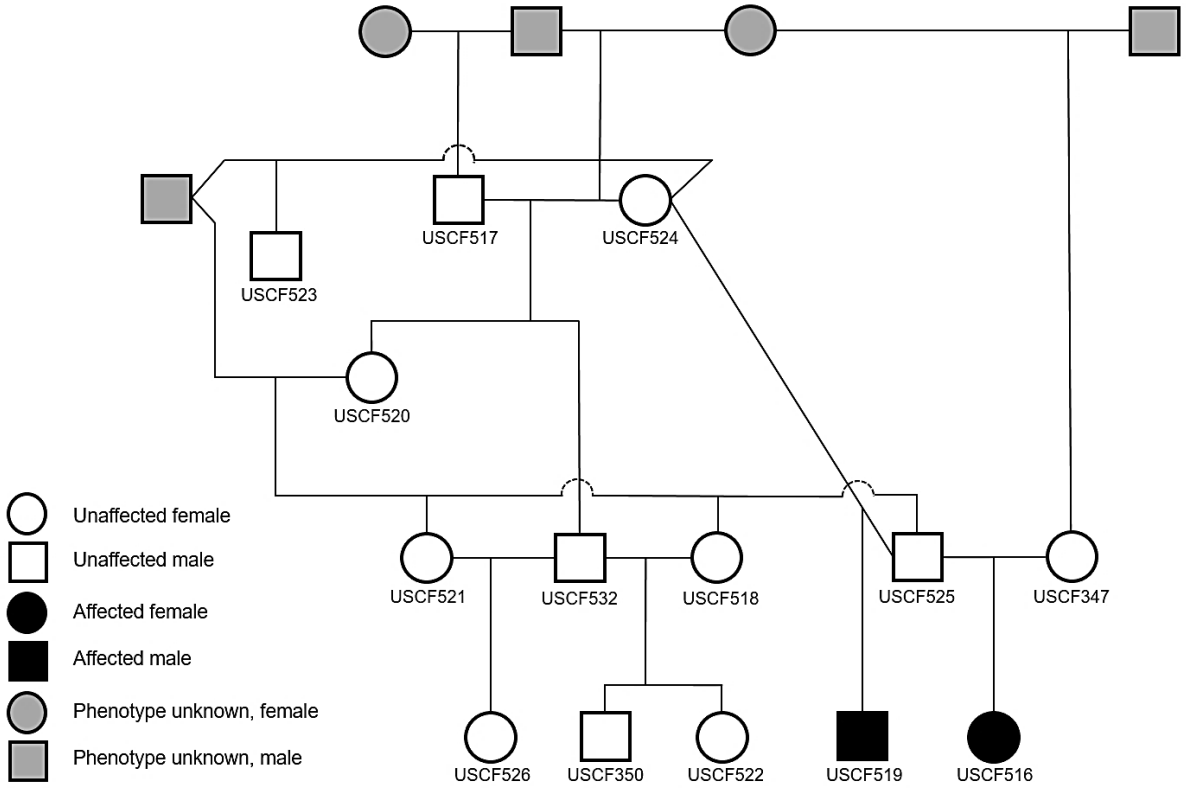

Figure S1. Pedigree of Hungarian Puli dogs segregating progressive retinal atrophy. Two dogs are affected with progressive retinal atrophy (USCF516, USCF519). Their parents (USCF347, USCF524, USCF525) and 9 dogs in the pedigree have normal vision. Dogs with individual identifiers were used in this study.

Table S1. A list of the 53 PRA candidate genes screened. Candidates include PRA genes causative or associated with PRA in other purebred dogs and genes that cause analogous autosomal recessive disease in humans.

Gene	CanFam3.1 Position	Reference
CNGB1	chr2:58574552-58640412	(Winkler *et al.* 2016)
DHDDS	chr2:73593302-73608757	(Winkler *et al.* 2016)
RLBP1	chr3:52261271-52269489	(Winkler *et al.* 2016)
PROM1	chr3:64260671-64360950	(Winkler *et al.* 2016)
PDE6B	chr3:91746571-91775372	(Winkler *et al.* 2016; Downs *et al.* 2014)
RGR	chr4:32492211-32495738	(Winkler *et al.* 2016)
RBP3	chr4:34972797-34983018	(Winkler *et al.* 2016)
PDE6A	chr4:59103965-59163857	(Winkler *et al.* 2016; Downs *et al.* 2014)
AIPL1	chr5:30828619-30834894	(Winkler *et al.* 2016)
GUCY2D	chr5:32844033-32859263	(Winkler *et al.* 2016)
NPHP4	chr5:59819237-59935037	(Winkler *et al.* 2016; Downs *et al.* 2014)
ABCA4	chr6:55058361-55253309	(Winkler *et al.* 2016)
RPE65	chr6:76887399-76911133	(Winkler *et al.* 2016; Downs *et al.* 2014)
CRB1	chr7:5277947-5419381	(Winkler *et al.* 2016)
RD3	chr7:9874740-9887791	(Winkler *et al.* 2016)
C1ORF36	chr7:9875590-9875862	(Winkler *et al.* 2016)
PDC	chr7:19498785-19514226	(Downs *et al.* 2014)
NRL	chr8:4086435-4091100	(Winkler *et al.* 2016)
RDH12	chr8:41686714-41689770	(Winkler *et al.* 2016)
SPATA7	chr8:59658291-59697320	(Winkler *et al.* 2016)
TTC8	chr8:60077187-60108376	(Winkler *et al.* 2016; Downs *et al.* 2014)
PDE6G	chr9:527987-528889	(Winkler *et al.* 2016)
PRCD	chr9:4185466-4188777	(Winkler *et al.* 2016; Downs *et al.* 2014)
FAM161A	chr10:61812850-61839706	(Winkler *et al.* 2016; Downs *et al.* 2014)
TULP1	chr12:4633093-4639684	(Winkler *et al.* 2016)

EYS	chr12:28547134-28708579	(Winkler *et al.* 2016)
C6ORF152	chr12:40445676-40489047	(Winkler *et al.* 2016)
CNGA1	chr13:43831161-43864273	(Winkler *et al.* 2016)
COL9A2	chr15:2647620-2661552	(Miyadera *et al.* 2012)
RPGRIP1	chr15:18331912-18385143	(Winkler *et al.* 2016; Downs *et al.* 2014)
CEP290	chr15:29194929-29281291	(Winkler *et al.* 2016)
LRAT	chr15:52401373-52401765	(Winkler *et al.* 2016)
SLC4A3	chr16:15117725-15126206	(Winkler *et al.* 2016; Downs *et al.* 2014)
ADAM9	chr16:26413196-26551132	(Winkler *et al.* 2016; Downs *et al.* 2014)
ZNF513	chr17:21332278-21335631	(Winkler *et al.* 2016)
C2ORF71	chr17:22899286-22910537	(Winkler *et al.* 2016; Downs *et al.* 2014)
MERTK	chr17:36336858-36445789	(Winkler *et al.* 2016)
BEST1	chr18:54468844-54480311	(Vilboux *et al.* 2008)

References

Downs L. M., Hitti, R., Pregnolato, S. & Mellersh, C. S. (2014) Genetic screening for PRA-associated mutations in multiple dog breeds shows that PRA is heterogeneous within and between breeds. *Veterinary Ophthalmology* **17,** 126–30.

Miyadera, K., Acland, G. M. & Aguirre, G. D. (2012) Genetic and phenotypic variations of inherited retinal diseases in dogs: The power of within- and across-breed studies. *Mammalian Genome* **23,** 40–61.

Vilboux, T., Chaudieu, G., Jeannin, P., Delattre, D., Hedan, B., Bourgain, C., Queney, G., Galibert, F., Thomas, A. & André, C. (2008) Progressive retinal atrophy in the Border Collie: a new XLPRA. *BMC Veterinary Research* **4,** 10.

Winkler, P. A., Davis, J. A., Petersen-Jones, S. M., Venta, P. J. & Bartoe, J. T. (2016) A tool set to allow rapid screening of dog families with PRA for association with candidate genes. *Veterinary Ophthalmology* 20*,* 372-376.

Table S2. PCR primer sequences.

CanFam3.1 Position		Forward primer (5'-3')	Reverse Primer (5'-3')	Tm (ºC)	Product length (bp)
chr3:52,260,877-52,278,803	1	TTGGTAGTAAAGCTGAGGTCATTG	TGGCCCTATCTCTCCATTTG	60	373
	2	GGATGGCCCCTAGAATAAGC	TTCCCAAAGTGTAGCCCAAG	60	866
	3	GGATGGCCCCTAGAATAAGC	TTCCCAAAGTGTAGCCCAAG	60	866
	4	CAATCCATGTTTCGGGTAGG	GGAAGTGGAGGCTATTGTCG	60	645
	5	GACCCACACCTCACTTCCAC	TGCGTATCCTGCTCAGTCAC	60	460
	6	AAGGTGTAGGCAGGTTCAAGTC	TTTCACCAGTCCCTTATTGTTG	59	744
	7	CCACACACAAGTCCTAAACCTC	CTCCTAGTGGGCTATCCTTTG	58	758
	8	CCACACACAAGTCCTAAACCTC	CTCCTAGTGGGCTATCCTTTG	58	758
chr30:35,378,421-35,381,822	1	CCCAGGCATCTAGGACCAG	TAGATGCTGGATTCGTGCTG	60	829
	2	CCCAGGCATCTAGGACCAG	TAGATGCTGGATTCGTGCTG	60	829
	3	CCCAGGCATCTAGGACCAG	TAGATGCTGGATTCGTGCTG	60	829
	4	CTCACCCACAAAAATCATGC	TGGAACTGCTAGGTCACAGG	59	522

Table S3. Putative variants identified from screening 53 candidate genes in parent-proband and an affected half sibling case.

Parent Genotype[1]	Case Genotype[1]	Gene	Position[2]	Type	Consequence[3]
G A	G G	*PDE6A*	4:59105480	Intronic	Modifier
G A	A A	*PDE6A*	4:59139522	Intronic	Modifier
C T	T T	*RD3*	7:9886063	Intronic, non-coding transcript	Modifier
C T	T T	*PRCD*	9:4187887	Intronic	Modifier
G C	A A	*PRCD*	9:4188050	Intronic	Modifier
A G	G G	*MERTK*	17:36360580	Intronic	Modifier
G A	A A	*MERTK*	17:36361700	Intronic	Modifier
C T	T T	*MERTK*	17:36371425	Intronic	Modifier

[1]High quality genotypes were called using Unified Genotyper provided by GATK and recommended hard filtering parameters (McKenna *et al.* 2010; Van der Auwera *et al*. 2013). [2]CanFam 3.1 positions. [3]Variant consequences on protein function or expression was predicted by Ensembl's Variant Effect Predictor (McLaren *et al.* 2010).

References

McKenna A., Hanna M., Banks E., Sivachenko A., Cibulskis K., Kernytsky A., Garimella K., Altshuler D., Gabriel S., Daly M. & DePristo M.A. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20,** 1297–303.

Van der Auwera G.A., Carneiro M.O., Hartl C., Poplin R., del Angel G., Levy-Moonshine A., Jordan T., Shakir K., Roazen D., Thibault J., Banks E., Garimella K.V., Altshuler D., Gabriel S., DePristo M.A. (2013) From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Current Protocols in Bioinformatics* **43,** 11.10.1-11.10.33.

McLaren W., Pritchard B., Rios D., Chen Y., Flicek P. & Cunningham F. (2010) Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26,** 2069–70.

4.2. Synopsis - A Coding Variant in the Gene Bardet-Biedl Syndrome 4 (*BBS4)* Is Associated with a Novel Form of Canine Progressive Retinal Atrophy

With the indication that a potentially novel form of progressive retinal atrophy was affecting the case Hungarian Puli as identified in section 4.1, in section 4.2, we present a published original research article that describes our methods for identifying a putative variant for canine progressive retinal atrophy. This method involved genotyping array and whole genome sequencing of parent-offspring trio samples. We executed this method and identified a highly associated nonsense SNP in *BBS4* (c.58.A > T, P_{CHISQ} = $3.43e^{14}$, n = 103). *BBS4* is a novel canine progressive atrophy gene. In humans, this gene is involved with Bardet-Biedl Syndrome, a ciliopathy that can cause other disease phenotypes including obesity and infertility. In this paper, we also provide evidence that the identified mutation in canine *BBS4* may cause syndromic disease as we observe similar phenotypes in the cases. *BBS4* is the second Bardet-Biedl Syndrome gene that has been linked to canine progressive retinal atrophy.

G3 Genes | Genomes | Genetics

INVESTIGATION

A Coding Variant in the Gene Bardet-Biedl Syndrome 4 (*BBS4*) Is Associated with a Novel Form of Canine Progressive Retinal Atrophy

Tracy Chew,*,[1] Bianca Haase,[†] Roslyn Bathgate,[†] Cali E. Willet,[‡] Maria K. Kaukonen,[§,,††] Lisa J. Mascord,* Hannes T. Lohi,[§,**,††] and Claire M. Wade*,[1]**

*School of Life and Environmental Sciences, [†]Sydney School of Veterinary Science, Faculty of Science, and [‡]Sydney Informatics Hub, Core Research Facilities, University of Sydney, 2006, Australia, [§]Department of Veterinary Biosciences, **Research Programs Unit, Molecular Neurology, and [††]Folkhälsan Institute of Genetics, University of Helsinki, 00014, Finland

ORCID IDs: 0000 0001 9529 7705 (T.C.); 0000 0001 8449 1502 (C.E.W.); 0000 0003 1087 5532 (H.T.L.); 0000 0003 3413 4771 (C.M.W.)

ABSTRACT Progressive retinal atrophy is a common cause of blindness in the dog and affects >100 breeds. It is characterized by gradual vision loss that occurs due to the degeneration of photoreceptor cells in the retina. Similar to the human counterpart retinitis pigmentosa, the canine disorder is clinically and genetically heterogeneous and the underlying cause remains unknown for many cases. We use a positional candidate gene approach to identify putative variants in the Hungarian Puli breed using genotyping data of 14 family based samples (CanineHD BeadChip array, Illumina) and whole genome sequencing data of two proband and two parental samples (Illumina HiSeq 2000). A single nonsense SNP in exon 2 of *BBS4* (c.58A > T, p.Lys20*) was identified following filtering of high quality variants. This allele is highly associated (P_{CHISQ} = 3.425*e* ^{14}, *n* = 103) and segregates perfectly with progressive retinal atrophy in the Hungarian Puli. In humans, *BBS4* is known to cause Bardet Biedl syndrome which includes a retinitis pigmentosa phenotype. From the observed coding change we expect that no functional BBS4 can be produced in the affected dogs. We identified canine phenotypes comparable with *Bbs4* null mice including obesity and spermatozoa flagella defects. Knockout mice fail to form spermatozoa flagella. In the affected Hungarian Puli spermatozoa flagella are present, however a large proportion of sperm are morphologically abnormal and <5% are motile. This suggests that BBS4 contributes to flagella motility but not formation in the dog. Our results suggest a promising opportunity for studying Bardet Biedl syndrome in a large animal model.

KEYWORDS
Hungarian Puli
whole genome sequencing
blindness
obesity
infertility

Progressive retinal atrophy (PRA) (OMIA #000830 9615) is the most common cause of hereditary blindness in the domestic dog (*Canis lupus familiaris*), affecting >100 pure breeds (Whitley *et al.* 1995). It is clinically and genetically heterogeneous and encompasses several forms of disease which vary by etiology, rate of progression, and age of onset (Downs *et al.* 2014a). The typical characteristics are gradual night, followed by day vision loss due to the degeneration of rod and cone photoreceptors, and this degeneration continues until the affected animal is completely blind (Parry 1953). Ophthalmic features that become apparent as the retina deteriorates include tapetal hyper reflectivity, vascular attenuation, pigmentary changes, and atrophy of the optic nerve head (Parry 1953; Clements *et al.* 1996; Petersen Jones 1998).

PRA is recognized as the veterinary equivalent of retinitis pigmentosa (RP) in humans due to the clinical and genetic similarities between the disorders (Petersen Jones 1998; Cideciyan *et al.* 2005; Zangerl *et al.* 2006; Downs *et al.* 2011). RP is a common cause of blindness in humans and affects ~1 in 4000 people (Hamel 2006). There are very limited treatment options for both PRA and RP at present (Hamel 2006). For this reason, the dog has become a valuable large animal

doi: https://doi.org/10.1534/g3.117.043109
Manuscript received March 19, 2017; accepted for publication May 15, 2017; published Early Online May 22, 2017.

Supplemental material is available online at www.g3journal.org/lookup/suppl/doi:10.1534/g3.117.043109/-/DC1.
[1]Corresponding authors: Faculty of Veterinary Science, University of Sydney, RMC Gunn Bldg., B19-301 Regimental Cres., Camperdown, NSW 2006, Australia. E-mail: tracy.chew@sydney.edu.au; and claire.wade@sydney.edu.au

model for retinal degeneration, in particular, for testing the efficacy of novel therapeutics such as gene therapy (Pearce Kelling *et al.* 2001; Acland *et al.* 2001; Narfström *et al.* 2003; Cideciyan *et al.* 2005; Beltran *et al.* 2012; Pichard *et al.* 2016). As of 2016, 256 retinal disease associated genes were identified for humans (https://sph.uth.edu/retnet/). Some of these genes cause nonsyndromic RP, while others contribute to syndromic disorders such as Bardet Biedl syndrome (BBS) (Hamel 2006).

Currently, retinal dystrophies in 58 domestic dog breeds have been linked to at least 25 mutations in 19 different genes (Miyadera *et al.* 2012; Downs *et al.* 2014b). Canine PRA is typically inherited in an autosomal recessive pattern, although two forms that are X linked (Vilboux *et al.* 2008) and one that has dominant inheritance have been reported (Kijas *et al.* 2002, 2003). Many of these discoveries in the canine were made using candidate gene studies, linkage mapping and genome wide association studies (GWAS) followed with fine mapping (Acland *et al.* 1999; Goldstein *et al.* 2006; Kukekova *et al.* 2009; Downs *et al.* 2014b). This success has been facilitated by the unique breeding structure of dogs. Intense artificial selection, genetic drift, and strong founder effects have resulted in stretches of linkage disequilibrium (LD) that can persist for several Mb within breeds, but only tens of kb across breeds (Lindblad Toh *et al.* 2005). This species population structure has allowed for the successful mapping of Mendelian traits with fewer markers and subjects compared to human gene mapping studies: as few as 10 unrelated cases and 10 controls (Karlsson *et al.* 2007; Frischknecht *et al.* 2013; Jagannathan *et al.* 2013; Willet *et al.* 2015; Gerber *et al.* 2015; Wolf *et al.* 2015). Such methods are accepted to work extremely well for mapping monogenic traits that segregate within a single breed.

Despite this achievement, there are still many forms of PRA in several breeds of dog that have yet to be genetically characterized. Traits with underlying genetic heterogeneity and a late onset are notoriously difficult to map using linkage or GWAS methods (Hirschhorn and Daly 2005; Korte and Farlow 2013). Although PRA is collectively common, individually, specific forms are relatively rare and it may take many generations until an adequately sized cohort of unrelated case samples are collected. The genetic heterogeneity of PRA can complicate the results of linkage mapping and GWAS, as different causative variants and genes can be responsible for an identical phenotype. In addition, both linkage and GWAS rely on markers to be in LD and segregate with the disease gene, making it difficult to detect rare or *de novo* variants (Hirschhorn and Daly 2005).

Since the advent of whole genome sequencing (WGS) and whole exome sequencing technologies, the discovery of causal variants for rare or genetically heterogeneous diseases has become more rapid with fewer case samples necessary for success. One study design of note that has been used in human and more recently in canine studies is the sequencing of parent proband trios (Zhu *et al.* 2015; Sayyab *et al.* 2016). As this method provides the chance for earlier diagnosis than previously possible, this gives patients the opportunity to access more personalized treatment options (Farwell *et al.* 2015; Zhu *et al.* 2015; Sawyer *et al.* 2016).

In a preliminary study, extensive screening of 53 genes associated with autosomal recessive PRA or RP revealed no putative variants that could be associated with PRA in the Hungarian Puli breed (Chew *et al.* 2017). Here, we combine genotyping array data and WGS data of a parent proband trio with an additional half sibling case to identify a potentially novel canine PRA gene. We successfully identify a highly associated mutation in exon 2 of *BBS4* (c.58A > T, $P_{CHISQ} = 3.425e^{-14}$, $n = 103$) that segregates perfectly with the disease phenotype. This mutation encodes a premature stop codon which is expected to result in complete loss of function of the BBS4 protein. The association of *BBS4* with canine PRA is a novel finding and presents the first description of an associated variant for PRA in the Hungarian Puli.

MATERIALS AND METHODS

Samples

This study involved 255 dogs (*C. lupus familiaris*) that comprised 103 Hungarian Puli and 152 Hungarian Pumi samples. This sample cohort included 14 Hungarian Pulis segregating PRA in an autosomal recessive pattern from a previous study (Chew *et al.* 2017). Three affected Hungarian Pulis (USCF516, USCF519, and USCF1311) were diagnosed with PRA at the age of 2 yr by registered specialists in veterinary ophthalmology. Diagnosis was based on observed ophthalmologic changes including vascular attenuation, hyper reflectivity, and reduced myelination in the optic nerve head. The parents (USCF347, USCF524, and USCF525) were similarly tested and confirmed as PRA clear. The remaining dogs were >3 yr of age and had normal vision as reported by their owners or veterinarians. Hungarian Pumis are a very closely related breed to the Hungarian Pulis and have been considered as a unique breed only since the 1920s, so were considered as a compatible cohort for this study.

Biological samples from the 255 dogs were collected either as EDTA stabilized whole blood or buccal cells using noninvasive swabs (DNA Genotek) or indicating Whatman FTA Cards (GE Healthcare). Genomic DNA was isolated from whole blood using the illustra Nucleon BACC2 kit (GE Healthcare) or from buccal cells on swabs using the Performagene Kit. For samples collected on an FTA card, DNA on discs was purified according to the manufacturer's guidelines.

We ensured that recommendations from the Australian Code for the Care and Use of Animals for Scientific Purposes were strictly followed throughout the study. Animal ethics approval was granted to conduct this research by the Animal Ethics Committee at the University of Sydney (approval number N00/9 2009/3/5109, September 24, 2009) and the State Provincial Office of Southern Finland (ESAVI/6054/04.10.03/2012).

Genotyping array data

Genotyping array data of 14 Hungarian Puli and WGS data of a parent proband trio and one additional half sibling case (USCF347, USCF516, USCF519, and USCF525) were obtained from the preliminary study (Chew *et al.* 2017). Genotyping was performed on the CanineHD BeadChip array (Illumina, San Diego, CA) by GeneSeek (Lincoln, NE). WGS was performed as 101 bp, paired end reads on the Illumina HiSeq 2000 by the Ramaciotti Centre, University of New South Wales, Kensington. The Illumina TruSeq DNA polymerase chain reaction (PCR) free kit was used to prepare the libraries. The four samples were barcoded and sequenced on two lanes of the sequencing machine. For additional information on sample and data collection, refer to the supplementary information in Chew *et al.* (2017). Sample information for this study can be found in Supplemental Material, File S1.

Candidate gene selection

Comprehensive screening of 53 PRA loci in the Hungarian Puli family revealed no obvious functional variants for the phenotype of interest (Chew *et al.* 2017). To identify novel candidates, regions concordant with a recessive inheritance pattern were identified using two case (USCF516 and USCF519) and 12 control dogs that were genotyped at 172,938 SNP markers on the CanineHD array. The control dogs included three PRA clear parents (USCF347, USCF524, and USCF525). Only markers that were genotyped as homozygous for the minor allele in cases,

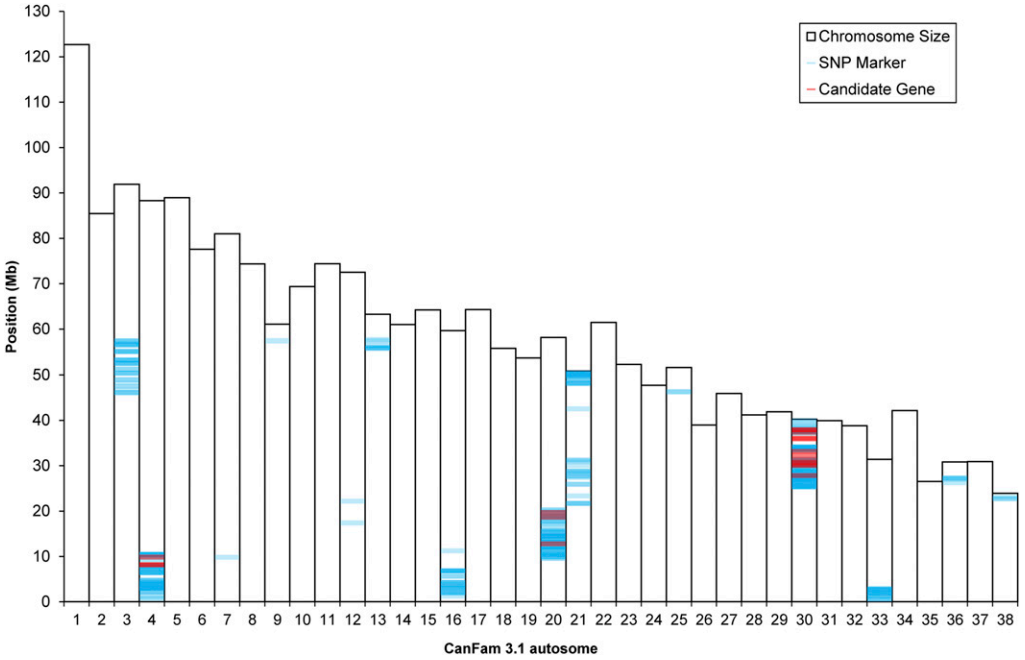

Figure 1 Positions of SNP array markers that segregate with the PRA phenotype and candidate genes are identified. Concordant markers are indicated in blue. Color opacity describes the density of concordant markers with darker hues corresponding with higher concordant marker density. Candidate genes are depicted in red. The locus with the highest frequency and density of markers is chr30: 25,254,123 39,976,525, with 103 markers and 12 candidate genes residing on the region. Following this is chr4: 556,510 10,473,708 with 61 markers and three candidate genes and chr20: 9,562,689 20,226,838 with 60 markers and three candidate genes.

heterozygous in the parents, and heterozygous or homozygous for the reference allele in the remaining nine control dogs were regarded as target loci (Microsoft Excel 2010).

Candidate genes were selected from the region with the highest frequency and density of concordant SNPs. LD in purebred dogs can span several Mb long (Lindblad Toh *et al.* 2005), thus we considered markers within 5 Mb to be in a single haplotype block. Using the corresponding syntenic positional region in the mouse reference genome (mouse genome assembly GRCm38, January 2012 build, the Genome Reference Consortium), we restricted our analysis to genes with a known phenotypic connection to vision using the Mouse Genome Browser (http://jbrowse.informatics.jax.org/). Any genes within the identified regions that were not already assessed in the preliminary PRA gene screening study (Chew *et al.* 2017) were chosen as positional candidate genes and considered for further analysis.

Whole-genome sequence processing and putative mutation detection

Next generation sequencing data from two cases (USCF516 and USCF519) and two parental controls (USCF347 and USCF525) were aligned to CanFam 3.1 (Hoeppner *et al.* 2014). Reads were aligned as pairs using the Burrows Wheeler Alignment tool with default parameters (Li and Durbin 2009). PCR duplicates were marked using Picard (http://broadinstitute.github.io/picard/). Local realignment around insertion deletions (indels) was performed using the Genome Analysis Tool Kit (GATK) (McKenna *et al.* 2010; DePristo *et al.* 2011).

High quality variants were called for all four individuals simultaneously over 12 candidate genes that were selected from the locus with the highest density of SNPs concordant with autosomal recessive inheritance. Raw variants were first called using HaplotypeCaller provided by GATK (Van der Auwera *et al.* 2013; McKenna *et al.* 2010). SNPs were then removed if Quality Depth <2.0, Fisher Strand >60.0, Mapping Quality <40.0, HaplotypeScore >13.0, MappingQualityRankSum $<$ 12.5, and ReadPosRankSum $<$ 8.0. Indels were removed if Quality Depth <2.0, Fisher Strand >200.00, and ReadPosRankSum $<$ 20.0.

The remaining high quality SNPs and indels were annotated using Variant Effect Predictor provided by Ensembl (McLaren *et al.* 2010). Known population variants obtained from publically available data were not considered as candidates (Lindblad Toh *et al.* 2005; Vaysse *et al.* 2011; Axelsson *et al.* 2013). Exonic variants were manually evaluated for genotype quality and conformation to the expected inheritance pattern using SAMtools tview (Li *et al.* 2009) and the UCSC Genome Browser. Remaining variants which were predicted by SIFT (Sim *et al.* 2012) to be deleterious (<0.05) were then considered for genotype validation and segregation analysis in the wider population by Sanger sequencing.

Variant validation and segregation analysis

The pedigree relationships among the 14 array genotyped individuals for which registered (Australian National Kennel Council) pedigree data were available were tested through identity by descent proportions calculated using PLINK (Purcell *et al.* 2007).

To confirm that the identified mutation was not a sequencing error and that the variant was concordant with the Mendelian expectation of the disorder phenotype, we genotyped 103 Hungarian Puli and 152 Hungarian Pumi for the candidate causative mutation c 58A $>$ T in *BBS4* using PCR and Sanger sequencing.

Forward (5′ GTTAGCAAGATACATGGTGTTGC 3′) and reverse (5′ GACTATTACTGCTTTCCCCAAAA 3′) primers were designed with Primer3 (Rozen and Skaletsky 2000) to amplify a 225 bp product flanking the candidate mutation. PCR was carried out using the AmpliTaq Gold 360 Master Mix (Applied Biosystems) in a 20 μl reaction volume. Following denaturation at 95° for 15 min, samples underwent amplification for 35 cycles at 95° for 30 sec, 55° for 30 sec, 72° for 45 sec, followed by a final elongation step at 72° for 10 min. For the purification of each sample, 7 μl of PCR product was dispensed into 3 μl of master mix containing 10× shrimp alkaline phosphatase (SAP) buffer, 1 U SAP, 1 U Exo I, and water. Enzymatic activity was enabled for 30 min at 37° and was then deactivated during 15 min at 80°. Sanger sequencing of purified PCR products was carried out by the Australian Genome Research Facility at Westmead in accordance with the vendor's instructions.

Table 1 Number of SNP and indel variants detected after applying standard hard filtering criteria

Filtering Criteria	SNP	Indel
High quality variants in candidate regions	2726	912
Not a common in canine population	1918	900
Exonic variants (total)	44	4
Synonymous	27	
Missense	16	
Nonsense	1	
In frame insertion		1
In frame deletion		2
Multiple nucleotide polymorphism		1
Manual check for quality and inheritance pattern	3	0
Predicted as deleterious by SIFT (<0.05)	1	0

Assessment of Bbs4 / mouse phenotypes in the dog

In addition to retinal degeneration, previous studies with *Bbs4* null mice demonstrated that the protein is implicated in obesity and infertility caused by a failure to form spermatozoa flagella (Mykytyn *et al.* 2004; Aksanov *et al.* 2014). Veterinarians who assessed the three affected Hungarian Puli anecdotally described these individuals as obese. A fertility assessment was performed for the sole intact male (USCF519; USCF347 is female and USCF1311 was desexed) by an animal reproduction specialist at the University of Sydney. Semen characteristics were compared with previously reported data for healthy dogs as a breed matched control was not able to be obtained (Schaer 2009). The sperm rich fraction of semen was collected by digital stimulation into a polypropylene test tube. Semen volume and color were noted immediately. Spermatozoa were observed using phase contrast microscopy at 100× magnification, and motility was subjectively determined. An aliquot of semen was smeared onto a slide for morphology assessment under oil at 1000× magnification, using previously described criteria (Feldman and Nelson 1987). Sperm count was determined by use of a hemocytometer.

Data availability

File S1 contains a list of accession numbers for available sample data. Genotyping array data were deposited in NCBI's Gene Expression Omnibus under the accession number GSE87642. WGS data for four Hungarian Puli (BAM files) can be obtained from NCBI's Sequence Read Archive under BioProject accession number PRJNA344694.

RESULTS

Target loci and candidate genes

Of the 172,938 SNP markers that were genotyped on the CanineHD BeadChip array for two cases and 12 controls, 363 markers segregated with PRA. Chromosome 30 (chr30) position 25.3 40.0 Mb demonstrated the highest density of concordant SNPs with 103 markers (Figure 1). This region is syntenic to mouse chromosome 9, 55.5 96.3 Mb (GRCm38/mm10 Assembly). The mouse phenome browser indicated that in this region 13 genes involved in vision have been identified, of which 12 are not currently known to be implicated in canine PRA. Chr4 position 0 5 10.5 Mb had the second most ($n = 61$) number of concordant markers. This region is syntenic to mouse chr13, 9.5 14.5 Mb, and chromosome 8, 122.7 127.7 Mb. Followed by this region is chr20 position 9.5 20.3 Mb

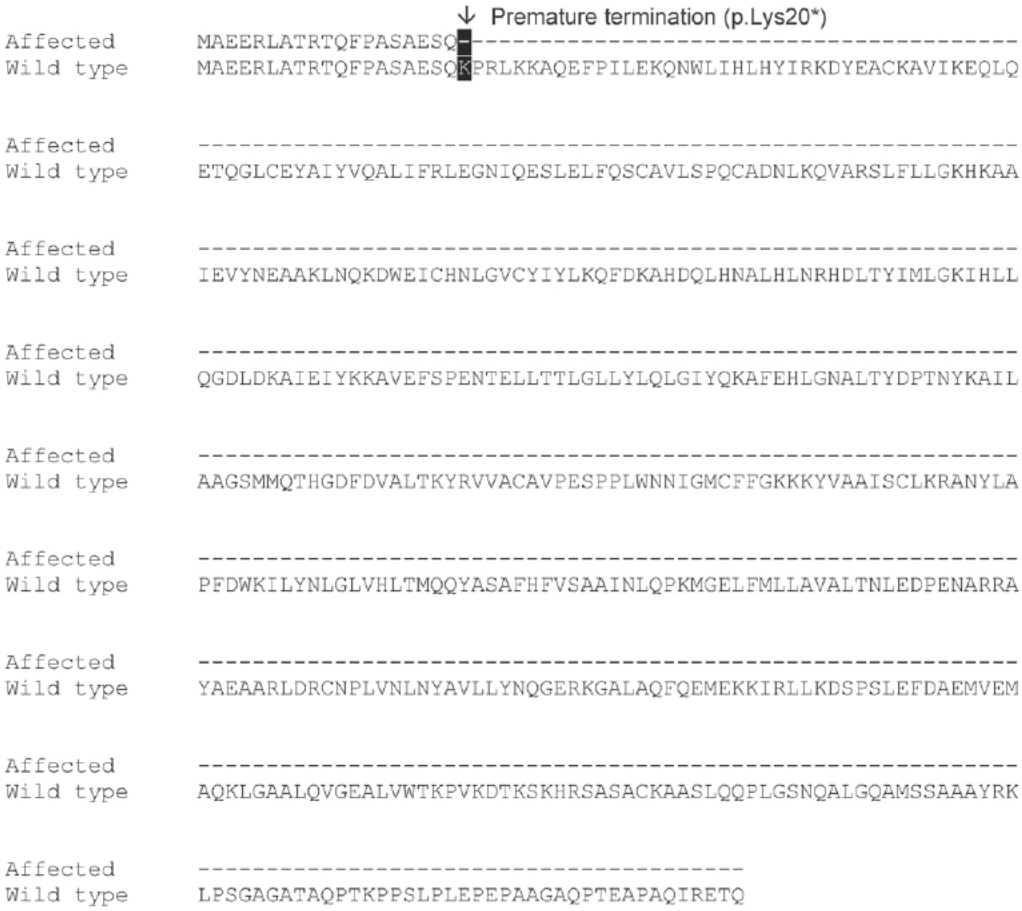

Figure 2 BBS4 protein sequence alignment of affected dogs containing the c.58A > T SNP and of the wild type protein. The SNP in affected dogs results in a premature stop codon (p.Lys20*). Hyphens (-) refer to missing amino acids in the affected dogs relative to the wild type protein.

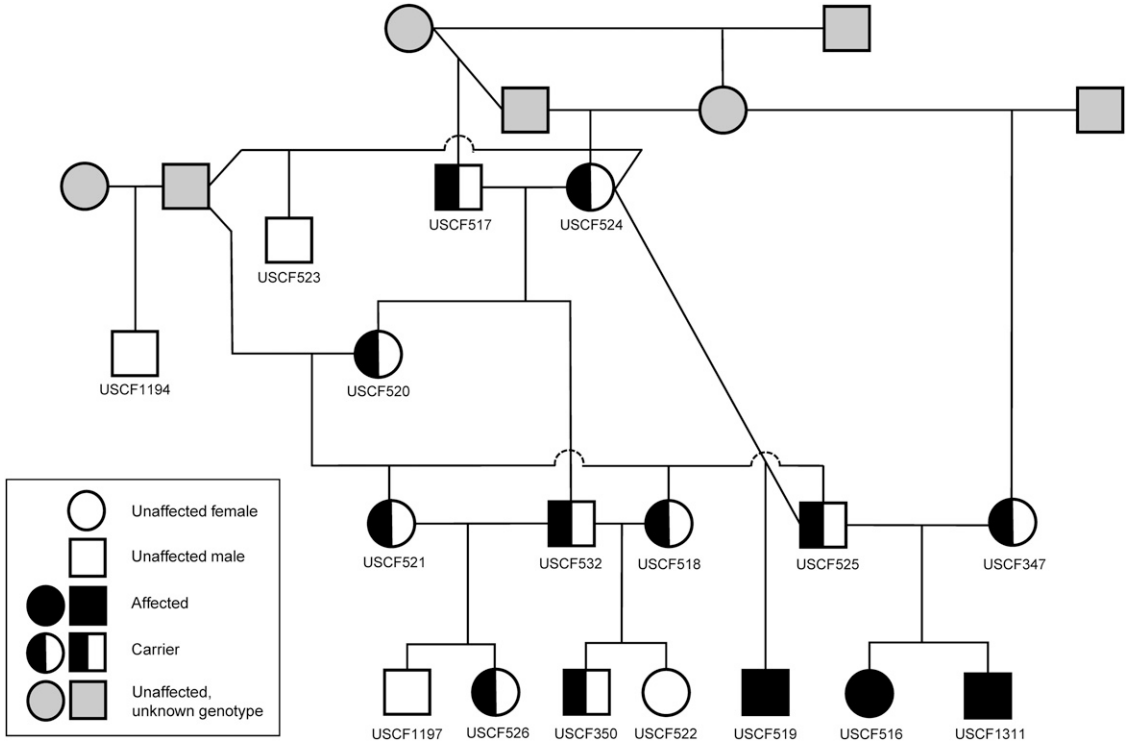

Figure 3 Segregation of the *BBS4* SNP (c.58A > T, p.Lys20*) in the Hungarian Puli family. DNA samples were available for all individuals with an identifier (n 17). PRA is consistent with an autosomal recessive form in this family. Genotypes confirmed through Sanger sequencing repre sented by unfilled (homozygous wild type *A*/*A*), filled (homozygous mutant *T*/*T*), or half filled (heterozygous *A*/*T*) circles (females) or squares (males) support this mode of inheritance.

with 60 concordant markers. This is syntenic to mouse chr6, 100.0 110.0 Mb. The mouse phenome browser revealed three candidate genes on each of the chr4 and chr20 regions. A total of 18 genes were selected as positional candidates in the current study (Table S1).

WGS and variant detection

Sequencing on the Illumina HiSeq 2000 produced an average of 171 million raw reads per dog. Of these reads, 99.3% were successfully mapped to the CanFam 3.1 reference genome, resulting in an average mapped coverage of 6.9× per individual.

In the 18 selected candidate genes, 2726 high quality SNPs were detected, 1918 of which are not currently known population variants (Table 1; Lindblad Toh *et al.* 2005; Vaysse *et al.* 2011; Axelsson *et al.* 2013). Of the 44 exonic SNPs, there were 27 synonymous, 16 missense, and one nonsense SNP. Two of the missense SNPs, one at *MEGF11*, chr30: 30,251,670 and the other at *STRA6*, chr30: 37,344,538, followed the expected inheritance pattern. Both were predicted by SIFT to be tolerated ($P = 1$) and therefore were not considered for further analysis. The single nonsense SNP detected occurred at *BBS4*, chr30: 36,063,748 and followed the expected inheritance pattern. This was predicted to be a deleterious mutation and was considered for validation and segregation analysis.

A total of 912 indels were detected, 900 of which are not currently known population variants (Table 1). Four of these were exonic, and by manual inspection none of these followed the expected inheritance pattern and so were not considered for further analysis.

Validation and segregation of putative nonsense variant in BBS4

A single, putative functional coding variant that passed all hard filtering criteria was identified. The variant results in a stop gained mutation in *BBS4* and is predicted to be deleterious. We manually completed the annotation of *BBS4* in the CanFam 3.1 reference genome as exon 1 was evidently missing (refer to File S2 for a full description of the methods used). The complete canine BBS4 protein can be accessed through Genbank (accession KX290494). In the complete *BBS4* gene, the puta tive mutation results in a premature stop codon (p Lys20*) as a result of a c.58A > T SNP in exon 2 (Figure 2).

The 103 Hungarian Puli included the three affected animals and 14 others with normal vision from the same kennel (Figure 3). Pedigree relationships for the 14 individuals for which genotyping array data were available were confirmed through identity by descent estimations (Table S2). Through Sanger sequencing, we observed that all three affected dogs (USCF516, USCF519, and USCF1311) were homozygous for the variant allele (*T*/*T*), all three obligate carrier parents were het erozygous (*A*/*T*), and the remaining unaffected Hungarian Puli were either heterozygous or homozygous for the wild type allele (*A*/*A*, Figure 4). All Hungarian Pumi were homozygous for the wild type allele. Genotypes for each individual in the study can be found in File S1.

An association of $P_{CHISQ} = 3.425e^{-14}$ between the c.58A > T SNP in *BBS4* to the disease phenotype was found for all validated Hungarian Puli genotypes ($n = 103$). When including validated Hungarian Pumi genotypes, the association is $P_{CHISQ} = 3.252e^{-34}$ ($n = 255$). The geno types are perfectly consistent with an autosomal recessive pattern of inheritance for the 17 Hungarian Puli individuals with pedigree infor mation, which supports the expected segregation pattern for PRA in this breed (Figure 3).

Assessment of Bbs4 / mouse phenotypes in the dog

The intact affected male Hungarian Puli ($n = 1$) was found to be subfertile. Semen analysis indicated normal sperm concentration but

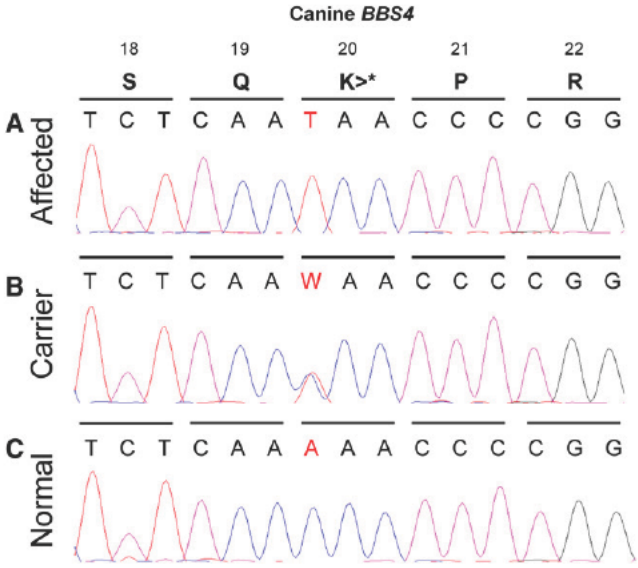

Figure 4 Sanger sequencing of a PCR fragment containing the c.58A > T SNP at position chr30: 36,063,748 on CanFam 3.1 in exon 2 of *BBS4*. The numbers above the amino acid code (S serine, Q glycine, K lysine, P proline, R arginine) denote their position in the protein sequence. Nucleotides (W A or T) are arranged in the 5′ to 3′ direction. (A) All affected Hungarian Puli cases (USCF516, USCF519, and USCF1311) have the *T/T* genotype which results in a nonsense mutation (p.Lys20*). (B) Carrier individuals including all parents (USCF347, USCF524, and USCF525) have the *A/T* genotype. (C) Unaffected individuals have the *A/A* genotype. This is identical to the canine reference sequence (CanFam 3.1 build).

a low total sperm count (13.65e^6). A high proportion (78%) of sperm had abnormal morphology, predominantly as a consequence of spermatozoa tail defects (74%, Table 2). There were <5% of sperm with normal motility.

Table 2 Semen analysis report of affected Hungarian Puli

	Normal Dog (Schaer 2009)	Affected Hungarian Puli
Normal morphology (%)	≥80	22
Abnormal morphology (%)		
Head defects		18
Midpiece defects		2
Tail defects		74
Normal motility (%)	≥70	<5
Concentration (number per ml)	4 400e^6	6.5e^6
Total sperm	100 3000e^6	13.65e^6
Volume (ml)	0.4 40	2.1
Color	Cloudy white	Transparent

Normal canine semen characteristics were obtained from Schaer 2009.

DISCUSSION

In this study, we identify a putative functional variant that is highly associated with the PRA disease phenotype in the Hungarian Puli breed. The variant occurs in a novel canine PRA gene. A preliminary study suggested that it was likely that a novel gene was causing disease in this family because no obvious functional variants were identified in the exons or promoters of any of 53 previously described PRA genes (Chew *et al.* 2017). Using genotyping array data and WGS data of a parent proband trio (USCF525, USCF347, and USCF516) and an additional half sibling case (USCF519), we identify a nonsense SNP (p.Lys20*) in exon 2 of *BBS4* that is significantly associated with disease (P_{CHISQ} = 3.425e^{-14}, n = 103). The associated SNP perfectly segregates in an autosomal recessive mode of inheritance. The mutation results in truncation at the N terminal of the translated BBS4 protein, reducing a 520 amino acid protein down to a 19 amino acid peptide. We predict that nonsense mediated decay of *BBS4* messenger RNA would hinder the expression of functional BBS4 protein (Popp and Maquat 2013). In humans and mice, *BBS4* is associated with the syndromic disease, BBS. We also provide some evidence that this form of PRA in the dog is part of a syndromic disease. There are now two BBS genes implied in canine PRA (*BBS4* and *TTC8*; Downs *et al.* 2014b). As BBS has not been previously reported in the dog, future PRA cases should be monitored for BBS phenotypes and gene mutations as they may provide a potential canine model for human disease.

BBS4 is one of eight evolutionarily conserved proteins that together form a multi protein complex referred to as the BBSome (Nachury *et al.* 2007; Loktev *et al.* 2008). This complex localizes to primary cilia, a small hair like organelle that is present on almost all vertebrate cells. Cilia play a vital role in many developmental pathways that occur during vertebrate embryogenesis enabling correct organ differentiation and spatial organization within the body. Primary cilia mediate multiple cell signaling activities in nondividing cells, responding to both mechanical and chemosensory stimuli in multiple body systems as they contain tissue specific sensory receptors (Singla and Reiter 2006; Goetz and Anderson 2010). The ubiquity of primary cilia and the concurrent differences in characteristics that they possess depending on their residing cell type give ciliopathies their clinical heterogeneity. Presumably, a dysfunctional protein that normally localizes to cilia of only one cell type will result in nonsyndromic disease, while proteins essential to cilia on multiple cell types such as those involved in its maintenance will result in syndromic disease.

The formation and maintenance of cilia are highly dependent on the bidirectional (anterograde and retrograde) movement of nonmembrane bound particles between the cell body and the tip of the cilia via its axonemal microtubules. The mechanism for this is referred to as intraflagellar transport (IFT) (Rosenbaum and Witman 2002). While the BBSome is not directly required for cilia formation, it is essential for the trafficking and organization of IFT complexes and hence has an indirect role in ciliary maintenance (Wei *et al.* 2012). Disruption in any of the BBSome genes (among 11 others that are not part of the BBSome) can cause failure of this mechanism, resulting in the rare ciliopathy, BBS (Suspitsin and Imyanitov 2016). The degree of importance of each BBS protein and their effect on the ability for the BBSome to carry out its ciliary functions within the various cell types remains elusive.

Studies of human BBS type 4 (OMIM #615982) and *Bbs4* null mice show that functions of the canine BBS4 protein are consistent with these theories. Structurally normal primary and motile cilia were observed in knockout mice, suggesting that *BBS4* is not required for the formation of cilia (Mykytyn *et al.* 2004). All affected individuals including the dogs used in this study experience retinal degeneration, despite having normal vision at a very young age (Iannaccone *et al.* 1999, 2005; Riise *et al.* 2002; Mykytyn *et al.* 2004; Li *et al.* 2014). This suggests that cilia are correctly formed, however IFT of newly synthesized proteins in the inner segment to the outer segment of photoreceptor cells is compromised, as the only route between the two is through connecting cilia (Marszalek *et al.* 2000; Mykytyn *et al.* 2004). These proteins are essential to photoreceptor maintenance and without these, the photoreceptor cells undergo apoptosis.

BBS is recognized as a syndromic disease, however in the dog, the disease may appear as nonsyndromic PRA. Like canine PRA, BBS is typically inherited in an autosomal recessive manner, except for one report of triallelic inheritance (Katsanis *et al.* 2001; Forsythe and Beales 2013). In human BBS type 4, symptoms that are observed in addition to RP include obesity, hypogenitalism, polydactyly, mental retardation, renal anomalies, and decreased olfaction (Iannaccone *et al.* 1999, 2005; Riise *et al.* 2002; Li *et al.* 2014; Aksanov *et al.* 2014). The severity and frequency of occurrence of each of these symptoms is variable like for all types of BBS, and clinical diagnosis is based on the presence of three to four primary and two secondary symptoms (Forsythe and Beales 2013). The difference in the underlying genetic mutation for reports of BBS type 4 is likely to contribute to this heterogeneity.

The affected Hungarian Puli in this study were predicted to have no functional BBS4, so we compared their phenotypes to those observed in *Bbs4* null mice. In these mice, obesity and a complete lack of sperma tozoa flagella were observed in addition to retinal degeneration (Mykytyn *et al.* 2004). In the dog, we observed all of these phenotypes but found that canine spermatozoa flagella were not as severely affected as those in the mouse. We observed 22% of sperm with normal mor phology in the dog; however, a large proportion of abnormal sperm had defective flagella (74%) and a very small proportion were motile (<5%; Table 2). This suggests that *BBS4* is only of moderate importance to flagella formation but is necessary for providing motility in the dog. More canine samples are required to confirm this.

The difficulty with differentiating nonsyndromic and syndromic disease in companion animals such as the dog is that many of the concurrent symptoms may not be diagnosed or recognized. Obesity is common with 26 43% of pure and mixed breed dogs classed as overweight in an Australian survey (McGreevy *et al.* 2005). As it is widely recognized as a nutritional disease, many people would un derestimate the genetic component of this phenotype. Further, in Australia many companion animals are desexed prior to maturity, limiting the opportunity to recognize fertility deficits. Other symp toms such as learning or developmental delay and decreased olfac tion may be difficult to assess in animals. For these reasons, we recommend that all human BBS genes might be considered as potential candidate genes for cases of canine PRA with unknown genetic causation. Further studies are required to confirm that *BBS4* causes syndromic disease in the dog and this should be mon itored as it may potentially be a useful large animal model for human BBS.

ACKNOWLEDGMENTS

We thank the owners and their pets for providing these samples and the veterinarians who phenotyped the Hungarian Puli. We acknowl edge Vidhya Jagannathan and Ranja Eklund with great appreciation for their technical assistance in providing control sample data. We also thank the Sydney Informatics Hub for providing access to the Artemis High Performance Computing system at the University of Sydney.

Author contributions: T.C , B H., and C.M.W. conceived and designed the experiments. Sample collection and preparation was done by T.C., B.H., M.K.K., H.T.L., and C.M.W. R.B. performed the fertility assessment. T.C. and C.M.W. performed the whole genome geno typing and resequencing analysis. T.C. and C.E.W. performed bioinformatic analysis of whole genome sequence data. L.J.M. provided intellectual insight into cilia biology and development of the discussion. T.C. wrote the article with the input and approval of all coauthors.

LITERATURE CITED

Acland, G. M., K. Ray, C. S. Mellersh, A. A. Langston, J. Rine *et al.*, 1999 A novel retinal degeneration locus identified by linkage and comparative mapping of canine early retinal degeneration. Genomics 59: 134 142.

Acland, G. M., G. D. Aguirre, J. Ray, Q. Zhang, T. S. Aleman *et al.*, 2001 Gene therapy restores vision in a canine model of childhood blindness. Nat. Genet. 28: 92 95.

Aksanov, O., P. Green, and R. Z. Birk, 2014 BBS4 directly affects prolifer ation and differentiation of adipocytes. Cell. Mol. Life Sci. 71: 3381 3392.

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: 360 364.

Beltran, W. A., A. V. Cideciyan, A. S. Lewin, S. Iwabe, H. Khanna *et al.*, 2012 Gene therapy rescues photoreceptor blindness in dogs and paves the way for treating human X linked retinitis pigmentosa. Proc. Natl. Acad. Sci. USA 109: 2132 2137.

Chew, T., B. Haase, C. E. Willet, and C. M. Wade, 2017 Exclusion of known progressive retinal atrophy genes for blindness in the Hungarian Puli. Anim. Genet. DOI: 10.1111/age.12553

Cideciyan, A. V., S. G. Jacobson, T. S. Aleman, D. Gu, S. E. Pearce Kelling *et al.*, 2005 *In vivo* dynamics of retinal injury and repair in the *rho dopsin* mutant dog model of human retinitis pigmentosa. Proc. Natl. Acad. Sci. USA 102: 5233 5238.

Clements, P. J., D. R. Sargan, D. J. Gould, and S. M. Petersen Jones, 1996 Recent advances in understanding the spectrum of canine gener alised progressive retinal atrophy. J. Small Anim. Pract. 37: 155 162.

DePristo, M. A., E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire *et al.*, 2011 A framework for variation discovery and genotyping using next generation DNA sequencing data. Nat. Genet. 43: 491 498.

Downs, L. M., B. Wallin Håkansson, M. Boursnell, S. Marklund, Å. Hedhammar *et al.*, 2011 A frameshift mutation in golden retriever dogs with pro gressive retinal atrophy endorses *SLC4A3* as a candidate gene for human retinal degenerations. PLoS One 6: e21452.

Downs, L. M., R. Hitti, S. Pregnolato, and C. S. Mellersh, 2014a Genetic screening for PRA associated mutations in multiple dog breeds shows that PRA is heterogeneous within and between breeds. Vet. Ophthalmol. 17: 126 130.

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

Farwell, K. D., L. Shahmirzadi, D. El Khechen, Z. Powis, E. C. Chao *et al.*, 2015 Enhanced utility of family centered diagnostic exome sequencing with inheritance model based analysis: results from 500 unselected fam ilies with undiagnosed genetic conditions. Genet. Med. 17: 578 586.

Feldman, E. C., and R. W. Nelson, 1987 *Canine and Feline Endocrinology and Reproduction.* W. B. Saunders, Philadelphia, PA.

Forsythe, E., and P. L. Beales, 2013 Bardet Biedl syndrome. Eur. J. Hum. Genet. 21: 8 13.

Frischknecht, M., H. Niehof Oellers, V. Jagannathan, M. Owczarek Lipska, C. Drögemüller *et al.*, 2013 A *COL11A2* mutation in Labrador Re trievers with mild disproportionate dwarfism. PLoS One 8: e60149.

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

Goetz, S. C., and K. V. Anderson, 2010 The primary cilium: a signalling centre during vertebrate development. Nat. Rev. Genet. 11: 331 344.

Goldstein, O., B. Zangerl, S. Pearce Kelling, D. J. Sidjanin, J. W. Kijas *et al.*, 2006 Linkage disequilibrium mapping in domestic dog breeds narrows the progressive rod cone degeneration interval and identifies ancestral disease transmitting chromosome. Genomics 88: 541 550.

Hamel, C., 2006 Retinitis pigmentosa. Orphanet J. Rare Dis. 1: 40.

Hirschhorn, J. N., and M. J. Daly, 2005 Genome wide association studies for common diseases and complex traits. Nat. Rev. Genet. 6: 95 108.

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: e91172.

Iannaccone, A., B. Falsini, N. Haider, G. Del Porto, E. M. Stone et al., 1999 Phenotypic characteristics associated with the BBS4 locus, pp. 187 199 in *Retinal Degenerative Diseases and Experimental Therapy*, edited by Hollyfield, J. G., M. M. La Vail, and R. E. Anderson. Plenum Press, New York.

Iannaccone, A., K. Mykytyn, A. M. Persico, C. C. Searby, A. Baldi *et al.*, 2005 Clinical evidence of decreased olfaction in Bardet Biedl syndrome caused by a deletion in the *BBS4* gene. Am. J. Med. Genet. A. 132A: 343 346.

Jagannathan, V., J. Bannoehr, P. Plattet, R. Hauswirth, C. Drögemüller *et al.*, 2013 A mutation in the *SUV39H2* gene in Labrador Retrievers with hereditary nasal parakeratosis (HNPK) provides insights into the epige netics of keratinocyte differentiation. PLoS Genet. 9: 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: 1321 1328.

Katsanis, N., S. J. Ansley, J. L. Badano, E. R. Eichers, R. A. Lewis *et al.*, 2001 Triallelic inheritance in Bardet Biedl syndrome, a Mendelian re cessive disorder. Science 293: 2256 2259.

Kijas, J. W., A. V. Cideciyan, T. S. Aleman, M. J. Pianta, S. E. Pearce Kelling *et al.*, 2002 Naturally occurring rhodopsin mutation in the dog causes retinal dysfunction and degeneration mimicking human dominant reti nitis pigmentosa. Proc. Natl. Acad. Sci. USA 99: 6328 6333.

Kijas, J. W., B. J. Miller, S. E. Pearce Kelling, G. D. Aguirre, and G. M. Acland, 2003 Canine models of ocular disease: outcross breedings de fine a dominant disorder present in the English mastiff and bull mastiff dog breeds. J. Hered. 94: 27 30.

Korte, A., and A. Farlow, 2013 The advantages and limitations of trait analysis with GWAS: a review. Plant Methods 9: 29.

Kukekova, A. V., O. Goldstein, J. L. Johnson, M. A. Richardson, S. E. Pearce Kelling *et al.*, 2009 Canine *RD3* mutation establishes rod cone dysplasia type 2 (*rcd2*) as ortholog of human and murine *rd3*. Mamm. Genome 20: 109 123.

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

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan *et al.*, 2009 The sequence alignment/map format and SAMtools. Bioinformatics 25: 2078 2079.

Li, Q., Y. Zhang, L. Jia, and X. Peng, 2014 A novel nonsense mutation in BBS4 gene identified in a Chinese family with Bardet Biedl syndrome. Chin. Med. J. (Engl.) 127: 4190 4196.

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: 803 819.

Loktev, A. V., Q. Zhang, J. S. Beck, C. C. Searby, T. E. Scheetz *et al.*, 2008 A BBSome subunit links ciliogenesis, microtubule stability, and acetylation. Dev. Cell 15: 854 865.

Marszalek, J. R., X. Liu, E. A. Roberts, D. Chui, J. D. Marth *et al.*, 2000 Genetic evidence for selective transport of opsin and arrestin by kinesin II in mammalian photoreceptors. Cell 102: 175 187.

McGreevy, P. D., P. C. Thomson, C. Pride, A. Fawcett, T. Grassi *et al.*, 2005 Prevalence of obesity in dogs examined by Australian veterinary practices and the risk factors involved. Vet. Rec. 156: 695 702.

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis *et al.*, 2010 The genome analysis toolkit: a MapReduce framework for ana lyzing next generation DNA sequencing data. Genome Res. 20: 1297 1303.

McLaren, W., B. Pritchard, D. Rios, Y. Chen, P. Flicek *et al.*, 2010 Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics 26: 2069 2070.

Miyadera, K., G. M. Acland, and G. D. Aguirre, 2012 Genetic and pheno typic variations of inherited retinal diseases in dogs: the power of within and across breed studies. Mamm. Genome 23: 40 61.

Mykytyn, K., R. F. Mullins, M. Andrews, A. P. Chiang, R. E. Swiderski *et al.*, 2004 Bardet Biedl syndrome type 4 (BBS4) null mice implicate Bbs4 in flagella formation but not global cilia assembly. Proc. Natl. Acad. Sci. USA 101: 8664 8669.

Nachury, M. V., A. V. Loktev, Q. Zhang, C. J. Westlake, J. Peränen *et al.*, 2007 A core complex of BBS proteins cooperates with the GTPase Rab8 to promote ciliary membrane biogenesis. Cell 129: 1201 1213.

Narfström, K., M. L. Katz, R. Bragadottir, M. Seeliger, A. Boulanger *et al.*, 2003 Functional and structural recovery of the retina after gene therapy in the RPE65 null mutation dog. Invest. Ophthalmol. Vis. Sci. 44: 1663 1672.

Parry, H. B., 1953 Degenerations of the dog retina. II. Generalized progressive atrophy of hereditary origin. Br. J. Ophthalmol. 37: 487 502.

Pearce Kelling, S. E., T. S. Aleman, A. Nickle, A. M. Laties, and G. D. Aguirre, 2001 Calcium channel blocker D *cis* diltiazem does not slow retinal degeneration in the *PDE6B* mutant *rcd1* canine model of retinitis pig mentosa. Mol. Vis. 7: 42 47.

Petersen Jones, S. M., 1998 A review of research to elucidate the causes of the generalized progressive retinal atrophies. Vet. J. 155: 5 18.

Pichard, V., N. Provost, A. Mendes Madeira, L. Libeau, P. Hulin *et al.*, 2016 AAV mediated gene therapy halts retinal degeneration in PDE6β deficient dogs. Mol. Ther. 24: 867 876.

Popp, M. W., and L. E. Maquat, 2013 Organizing principles of mam malian nonsense mediated mRNA decay. Annu. Rev. Genet. 47: 139 165.

Purcell, S., B. Neale, K. Todd Brown, L. Thomas, M. A. Ferreira *et al.*, 2007 PLINK: a tool set for whole genome association and population based linkage analyses. Am. J. Hum. Genet. 81: 559 575.

Riise, R., K. Tornqvist, A. F. Wright, K. Mykytyn, and V. C. Sheffield, 2002 The phenotype in Norwegian patients with Bardet Biedl syn drome with mutations in the BBS4 gene. Arch. Ophthalmol. 120: 1364 1367.

Rosenbaum, J. L., and G. B. Witman, 2002 Intraflagellar transport. Nat. Rev. Mol. Cell Biol. 3: 813 825.

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

Sawyer, S. L., T. Hartley, D. A. Dyment, C. L. Beaulieu, J. Schwartzentruber *et al.*, 2016 Utility of whole exome sequencing for those near the end of the diagnostic odyssey: time to address gaps in care. Clin. Genet. 89: 275 284.

Sayyab, S., A. Viluma, K. Bergvall, E. Brunberg, V. Jagannathan *et al.*, 2016 Whole genome sequencing of a canine family trio reveals a *FAM83G* variant associated with hereditary footpad hyperkeratosis. G3 (Bethesda) 6: 521 527.

Schaer, M., 2009 *Clinical Medicine of the Dog and Cat*. CRC Press, Boca Raton, FL.

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: W452 W457.

Singla, V., and J. F. Reiter, 2006 The primary cilium as the cell's antenna: signaling at a sensory organelle. Science 313: 629 633.

Suspitsin, E. N., and E. N. Imyanitov, 2016 Bardet Biedl syndrome. Mol. Syndromol. 7: 62 71.

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 ge nome analysis toolkit best practices pipeline. Curr. Protoc. Bioinformatics 11: 11.10.1 11.10.33.

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

Vilboux, T., G. Chaudieu, P. Jeannin, D. Delattre, B. Hedan *et al.*, 2008 Progressive retinal atrophy in the Border Collie: a new XLPRA. BMC Vet. Res. 4: 10.

Wei, Q., Y. Zhang, Y. Li, Q. Zhang, K. Ling *et al.*, 2012 The BBSome controls IFT assembly and turnaround in cilia. Nat. Cell Biol. 14: 950 957.

Whitley, R. D., S. A. McLaughlin, and B. C. Gilger, 1995 Update on eye disorders among purebred dogs. Vet. Med. 90: 574 592.

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: 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: e1005059.

Zangerl, B., O. Goldstein, A. R. Philp, S. J. Lindauer, S. E. Pearce Kelling *et al.*, 2006 Identical mutation in a novel retinal gene causes progressive rod cone degeneration in dogs and retinitis pigmentosa in humans. Genomics 88: 551 563.

Zhu, X., S. Petrovski, P. Xie, E. K. Ruzzo, Y. F. Lu *et al.*, 2015 Whole exome sequencing in undiagnosed genetic diseases: interpreting 119 trios. Genet. Med. 17: 774 781.

Communicating editor: D. L. Bannasch

Chapter 5. The Genetics of Severe Haemophilia A in the Australian Kelpie

5.1. Abstract

Haemophilia A is a bleeding disorder caused by a reduced activity of factor VIII (FVIII) that is required in the coagulation cascade to create blood clots. Disease is associated with a wide range of mutation types in the FVIII gene that either have X-linked inheritance or have occurred sporadically. In this study, we report the occurrence of haemophilia A in two purebred Australian Kelpie pups. The affected dogs had a FVIII coagulation activity of <1.5 %, which is considered to be severe by the Animal Health Diagnostic Centre, Cornell University. To our knowledge, there was no family history of this disease in the affected kennel. Using genotyping array data from the CanineHD BeadChip array, we inferred haplotypes in the FVIII loci. The apparently healthy maternal grandsire contained the same apparent haplotypes as the affected dogs, suggesting a putative *de novo* mutation in the FVIII gene. Using 100 base pair, paired-end Illumina sequencing reads from one affected and 11 unrelated control dogs, we called putative SNP, indel and structural variants in the FVIII gene and 13 other bleeding disorder loci using GATK's haplotype caller, SVtyper and LUMPY. A total 37 intronic SNPs unique to the affected dog were identified, but without functional data, their effect could not be confirmed. Comprehensive screening of the FVIII gene in whole genome sequence data and through Sanger sequencing revealed no candidate exonic mutations. The affected dogs were also clear of a commonly reported inversion mutation affecting intron 22, which was tested by long range PCR. Low mutation detection success rates are common in rare disease research, and this emphasizes the need for the development of more statistically powerful and effective methodologies to provide sufferers with a rapid diagnosis.

5.2. Introduction

Haemophilia A is recognized as one of the most common and severe bleeding disorders affecting a range of animals including both dogs (*Canis lupus familiaris*) and people

(Brooks 1999; Graw *et al.* 2005). It has an X-linked recessive mode of inheritance and occurs in ~1 in 10,000 live male births worldwide. Disease occurs when damaging mutations in the FVIII gene cause coagulation factor VIII (FVIII) to be dysfunctional, limiting an individual's ability to propagate the intrinsic coagulation pathway that is necessary to control bleeding (Tantawy 2010). In addition to excessive bleeding, clinical signs include increased risk of haematoma and spontaneous haemorrhaging in joints and muscles. The severity of disease may be classified as mild, moderate or severe depending on endogenous circulating FVIII levels (< 1%, 2-5% and 5-20% respectively) (Brooks 1999; Bolton-Maggs and Pasi 2003). Disease symptoms can be managed by the periodic transfusion of plasma-derived of recombinant FVIII and lifestyle changes to minimize the risk of bleeding. Due to the short half-lives of the proteins and the development of neutralizing antibodies in patients, treatment success is limited and the search for better treatment options is ongoing. Dogs with severe disease are often euthanized before reaching the age of one year due to recurrent and high risk of fatal haemorrhaging.

The FVIII gene is a large and complex gene that includes 26 exons spanning over 186 kilobases (kb) of DNA at the end of the long arm of the X chromosome in both humans and dogs. It coincides with a mutational hotspot and for 30% of human cases, spontaneous disease occurs in individuals with no prior family history of haemophilia A. In humans, over 2,015 distinct causative mutations have been identified and each is linked to a specific type of clinical severity (http://www.factorviii-db.org/) (Graw *et al.* 2005; Repessé *et al.* 2007; Tantawy 2010). The most commonly occurring mutation (~45% of severe human cases) is an inversion causing a breakpoint at intron 22 within the FVIII gene. Intra-chromosomal recombination between a 9.5 kb region within intron 22 (termed *int22h1*) and two highly homologous regions distal to the FVIII locus and towards the telomere on the X chromosome (*int22h1 and int22h3*) can cause this inversion event spontaneously. A wide range of other causal variant types including other inversions, insertions, deletions, nonsense and missense mutations have been reported.

Haemophilia A has been documented in several pure and mixed breed dogs (OMIA #000437-9615). There is a breed predisposition for the clinical severity of haemophilia A and for the tendency for disease to occur sporadically (Brooks 1999; Brooks *et al.* 2008; Dunning *et al.* 2009). Unlike human haemophilia A, the genetic characterisation of disease is lacking in the veterinary literature and so genetic testing for the detection of carrier individuals is not available in many breeds (Mischke *et al.* 2011). Amongst the cases that have been characterised, a mutation that resembles the intron 22 inversion in humans was evident in two separate dog colonies. The colony housed at Queen's University established from affected Miniature Schnauzers and obligate carrier Schnauzer-Brittany Spaniel females were found with an abnormal FVIII transcript that contained a novel sequence element following exons 1-22 (Hough *et al.* 2002). Researchers similarly described an aberrant FVIII transcript in a colony of Irish Setters from the University of North Carolina in Chapel Hill (Lozier *et al.* 2002). Several other nonsense and missense mutations in breeds including German Shepherds, Boxers and an Old English Sheepdog have been found to associate with severe haemophilia A (Mischke *et al.* 2011; Christopherson *et al.* 2014; Lozier *et al.* 2016).

Mice, pigs and dogs are popular animal models for testing novel therapies for haemophilia A (and B, caused by dysfunctional coagulation factor IX) (Yen *et al.* 2016). Besides their use in preclinical trials for FVIII infusion safety testing, dogs are popular models for a variety of gene therapy technologies to treat and ultimately cure haemophilia in humans because of the high homology of the FVIII gene and immune systems between the two species (Yen *et al.* 2016). In the last 20 years, researchers have used adeno-associated viral vectors and are challenged with achieving prolonged, high FVIII expression levels, whilst reducing vector toxicity and curbing patient response of inhibiting antibodies that block the function of infused FVIII (Ward and Walsh 2017). More recently, CRISPR-Cas9 mediated genome editing strategies have been used to achieve mutation correction without off target side effects in mice, showing the potential for pre-clinical testing in a larger model such as the dog (Ohmori *et al.* 2017).

This study reports the first case of HA in the Australian Kelpie breed. Two male littermates were diagnosed with severe disease through FVIII coagulation assays

conducted by the Animal Health Diagnostic Centre, Cornell University. Here we explore the FVIII gene and genetic cause of disease in this family. We utilize CanineHD BeadChip genotyping array data from the two affected, male pups and 10 of their unaffected relatives to infer haplotypes at the FVIII loci. Whole genome sequencing (WGS) data from one affected and 11 unrelated Australian Kelpie dogs obtained from the Illumina HiSeq 2000 or 2500 were used to call variants in FVIII and 13 other bleeding candidate loci. We also used PCR and Sanger sequencing to confirm that the affected dogs were clear of obvious exonic mutations and the alleged intron 22 inversion. We identified 37 putative intronic SNPs which should be further investigated for functional importance.

5.3. Methods

5.3.1. Animals

Two male, Australian Kelpie littermates (USCF305 and USCF311) were brought to the University Veterinary Teaching Hospital at the University of Sydney. The pups were presenting with symptoms consistent with severe coagulopathy. One individual died and the second was subsequently euthanized. The two affected and 21 control dogs of the same breed were selected for this study. The control cohort included 10 samples from the same extended pedigree as the affected individuals (refer to the section 5.4.1 for pedigree information). The diagnosis of haemophilia A in the affected dogs and nine of their unaffected relatives was confirmed by factor VIII and/or factor IX coagulation assays carried out by the Animal Health Diagnostic Centre, Cornell University. FIX tests are commonly performed with FVIII because reduced activity in this factor is commonly seen in haemophilia A patients and it is used to test the affection status of an even rarer bleeding disorder, haemophilia B. All other individuals exhibited no disease phenotypes or were unrelated, and so were assumed healthy. To the author's knowledge, haemophilia A has not been diagnosed in this family prior. HA has an X-linked recessive inheritance pattern which supports the segregation of disease in this family.

Genomic DNA was extracted from EDTA stabilized whole blood samples from each individual using the illustra Nucleon BACC2 kit using the manufacturer's recommended protocol (GE Healthcare). Recommendations from the Australian Code for the Care and Use of Animals for Scientific Purposes were strictly adhered to throughout this study. Research was conducted with animal ethics approval, granted by the Animal Ethics Committee at the University of Sydney (approval number N00/9–2009/3/5109, September 24, 2009).

5.3.2. Genotyping array and whole genome sequencing data

Genotyping array data for two affected and 10 unaffected relatives were obtained from the CanineHD BeadChip array at 173, 650 SNPs (Illumina, San Diego, CA) by Neogen (Lincoln). The unaffected dogs include the dam of the affected animals. The data was used to infer FVIII haplotypes by observing sample genotypes its residing region on chromosome X (122,897,137 – 123,043,373). Only markers with a minor allele frequency (MAF) > 0.1 were considered.

Next generation sequencing data was performed for one affected Kelpie (USCF305) on the Illumina HiSeq 2000 as 100 base pair, paired end reads on a single lane of the sequencing platform at the Ramaciotti Centre (University of New South Wales, Kensington). The quality and quantity of DNA for the other affected male was insufficient for whole genome sequencing. Whole genome sequencing data similarly sequenced on the Illumina HiSeq 2000 or 2500 for 11 Australian Working Kelpies were obtained from an unrelated study (Arnott *et al.* 2015; Pan *et al.* 2017).

Raw reads were aligned as pairs to the CanFam 3.1 reference sequence using the Burrows-wheeler Alignment (BWA-MEM) tool (version 0.7.15) with default parameters (Li and Durbin 2009). PCR duplicates were marked with SAMBLASTER (version 0.1.22) (Faust and Hall 2014). Local realignment was performed around insertion-deletions with the Genome Analysis Tool Kit (GATK version 3.6.0) (McKenna *et al.* 2010; DePristo *et al.* 2011). Base quality scores were recalibrated with GATK using known variants that were downloaded from Ensembl's dbSNP database.

5.3.3. Screening for putative variants in known bleeding disorder loci

A reduction in coagulation FVIII can be influenced by proteins other than FVIII including FIX and von Willebrand factor. For this reason, we called SNP and small indel variants in WGS data in a total of 14 genes known to be associated with a bleeding tendency phenotype (Table S1). GATK's HaplotypeCaller was used to call these variants following best practice recommendations (Van der Auwera *et al.* 2013). Low quality SNPs defined by Quality Depth < 2.0, Fisher Strand > 60.0, Mapping Quality < 40.0, HaplotypeScore > 13.0, MappingQualityRankSum < -12.5 or ReadPosRankSum < -8.0 were removed. Similarly, low quality indels defined by Quality Depth < 2.0, Fisher Strand > 200.0 and ReadPosRankSum < -20.0 were removed. We also called structural variants by using LUMPY (version 0.2.11) and genotyped these calls with SVtyper (version 0.0.2) with the default settings applied (Layer *et al.* 2014; Chiang *et al.* 2015). Common population variants were not considered as candidates (Lindblad-Toh *et al.* 2005; Vaysse *et al.* 2011; Axelsson *et al.* 2013). The remaining variants were annotated with Ensembl's Variant Effect Predictor (McLaren *et al.* 2010). Alleles which segregated in an autosomal or X-linked recessive manner, or which were predicted to have a high impact by VEP were considered for further validation.

5.3.4. Screening the *FVIII* gene

To ensure that no variants were missed, we manually screened for variants in the 26 exons of the FVII gene using WGS data of affected individual USCF305 using SAMtools tview. For exons with low to no coverage, we performed polymerase chain reaction (PCR) and Sanger sequencing for two affected (USCF305 and USCF311) and two control dogs (USCF316, USCF1290). Primer 3 (Rozen and Skaletsky 2000) was used to design forward and reverse primers that captured both exons 23 and 24 in a 896 bp product (5'-ATGTCTGTGCGGATTCTTCC -3' and 5'-TTGTACCCTGTCTGCACCTG-3' respectively) and exon 26 in a 844 bp product (5'-GTCACTGCACAGAGGACGTG-3' and 5'-TGGGTTCGACGTGATGAAG-3', respectively).

PCR was performed using the AmpliTaq Gold 360 Master Mix (Applied Biosystems) in a 10 µL reaction volume. Solutions underwent PCR cycling conditions using the Veriti 96 Fast Thermal Cycler (Applied Biosystems) as follows: denaturation at 95°C for 15 min; amplification for 35 cycles at 95 °C for 30 sec, 55 °C for 30 sec, 72 °C for 45 secs; a final elongation step at 72 °C for 10 min. PCR products were purified in a 10 µL reaction volume. This contained 7 µL of PCR product and a 3 µL master mix containing 1 U Exo I, 1 U shrimp alkaline phosphatase (SAP) and 10x SAP buffer. Sanger sequencing was performed at the Australian Genome Research Facility (Westmead) according to the vendor's instructions.

5.3.5. Analysis of a putative inversion mutation at intron 22 of *FVIII*

In humans, the most prevalent causative mutation of severe HA is an inversion causing a break of the FVIII gene within intron 22, occurring in ~45% of cases. Two independent canine studies have alluded to a similar mutational event causing severe HA in their dog colonies. We used LUMPY to call structural variants in WGS Kelpies. We also extracted improperly paired reads using SAMtools from intron 22 of *FVIII*.

To test for a putative inversion with a breakpoint at this location, we designed a long-range PCR test using Primer 3 to design primers to capture intron 22. We decided to capture the entire intron 22 which spans 16,081 bp as the exact location of the putative breakpoint is unknown. We designed two separate PCR tests with overlapping fragments as we were unable to amply the 16 kb fragment in a single PCR. The test was performed for the two cases (USCF305, USCF311) with the SequalPrep Long PCR kit with dNTPs (Thermo Fisher Scientific). The first PCR contained forward (5'-GTAATGGGTTGGGTGCAAAC-3') and reverse (5'- AAGGAGCCAATGACAAATGG-3') primers that captured an 11,031 bp fragment. The subsequent PCR contained forward (5- TGTCATTGGCTCCTTTATAGCTC-3) and reverse (5'- TCTCCAGCCTCTACGTGTC TC-3') primers that captured a 5,256 bp fragment. Each PCR was performed in a 10 µL reaction volume. Solutions underwent PCR cycling conditions using the Veriti 96 Fast Thermal Cycler (Applied Biosystems) using the recommended cycling conditions provided by the SequalPrep Long PCR kit.

5.4. Results

5.4.1. *FVIII* assessment in the affected family

Coagulation activity assays for FVIII and FIX for the samples obtained from the affected pedigree are shown in Figure 5.1. Affected individual USCF305 and USCF311 have a coagulation FVIII activity of <1.5 % and 1.5% respectively. A FVIII activity of <2% is defined by the Animal Health Diagnostic Centre, Cornell University as an individual with severe HA. FIX coagulation could not be obtained for USCF311 and four other relatives due to poor sample quality. USCF305 had the lowest FIX activity in the family (29%). Regional haplotypes at FVIII on chromosome X indicate that the unaffected grandparent contains the same haplotype as the affected littermates (Figure 5.1).

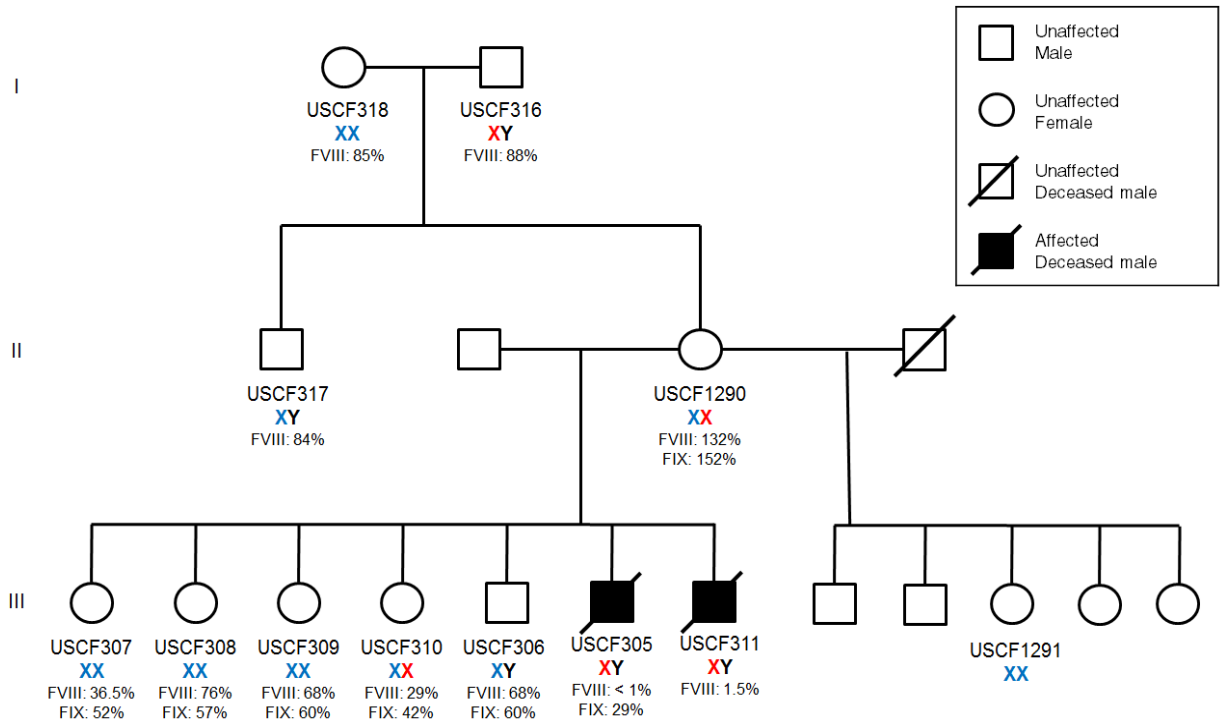

Figure 5.1 Pedigree of the Australian Kelpie family segregating haemophilia A. Individual identifiers are displayed under the individuals that were included in this study. Severe disease was diagnosed in USCF305 and USCF311 based on the level of coagulation FVIII and FIX activity that were measured by the Animal Health Diagnostic Centre, Cornell University. Reference levels for FVIII and FIX coagulation are 50 – 200% and 50 – 150% respectively. Haplotypes represented by genotyping data obtained from the Canine HD BeadChip array are depicted by the colour of the X under each individual.

5.4.2. Detection of variants in bleeding disorder loci

Variants were called and annotated in 14 clotting factor loci for the single case (USCF305) and 11 unrelated control Australian Kelpie dogs. Using GATK's HaplotypeCaller, 2,282 and 668 raw SNPs and indels were called respectively. There were no structural variants called by LUMPY in any of the candidate genes. After filtering for quality and for variants which are not known to be common in the population,

1,580 SNPs and 646 indels remained. None of the indels conformed to the expected mode of inheritance with at least one control animal containing an identical genotype as the case animal at the putative loci. There were 37 SNPs which followed the expected inheritance pattern. All SNPs were intronic (Table 5.1).

Table 5.1. Variants detected in 14 bleeding tendency candidate genes using whole genome sequencing data of one case and 11 control Australian Kelpies.
Each locus was genotyped for each of the 12 animals. Alternative alleles detected were only present in affected dog USCF305 and were not found in the control dogs. Positions and reference alleles are relative to the CanFam 3.1 reference genome.

Chromosome	Position	Gene	Intron	Reference Allele	Alternative Allele
9	9,212,395	*GPIIIa*	2	C	T
9	9,213,121	*GPIIIa*	2	A	G
9	9,214,462	*GPIIIa*	2	C	G
9	9,214,942	*GPIIIa*	2	G	A
9	9,215,281	*GPIIIa*	1	C	A
9	9,216,703	*GPIIIa*	1	G	A
9	9,216,981	*GPIIIa*	1	C	A
9	9,217,017	*GPIIIa*	1	G	A
9	9,217,129	*GPIIIa*	1	C	T
9	9,224,785	*GPIIIa*	1	A	T
9	9,226,881	*GPIIIa*	1	G	A

9	19,053,999	*ITGA2B*	10	A	G
9	19,054,961	*ITGA2B*	12	G	T
9	19,058,091	*ITGA2B*	16	T	C
18	42,783,087	*F2*	13	G	A
18	42,783,320	*F2*	12	T	C
18	42,783,362	*F2*	12	G	A
18	42,783,978	*F2*	12	A	T
18	42,784,501	*F2*	12	T	C
18	42,784,518	*F2*	12	A	G
18	42,785,453	*F2*	12	T	C
18	42,785,747	*F2*	12	T	C
18	42,785,748	*F2*	12	G	A
18	42,785,756	*F2*	12	G	A
18	42,786,022	*F2*	12	T	A
18	42,786,584	*F2*	12	A	G
18	42,788,637	*F2*	12	A	G
18	42,788,747	*F2*	12	C	A
18	42,788,894	*F2*	12	T	C
18	42,791,141	*F2*	12	C	T

18	42,791,283	*F2*	12	A	G
22	60,576,451	*F7*	2	C	T
22	60,591,358	*F10*	4	C	G
22	60,591,363	*F10*	4	G	A
22	60,594,634	*F10*	6	C	A
X	109,526,498	*F9*	6	A	T
X	122,916,938	*F8*	22	G	T

5.4.3. Screening the *FVIII* gene for novel and known mutations

We performed SNP, indel and structural variant calling in WGS data and Sanger sequencing of exons which were not sufficiently covered (exons 23, 24 and 26). In addition to the intron variant detected previously (Table 5.1), we found four exonic SNP variants in USCF305. This included one missense mutation in exon 14 and three synonymous SNPs in exons 1 and 15 (Table S2). These alleles were genotyped as homozygous alternative in USCF305 and one or more of the control dogs and so are not likely to be causative for haemophilia A. An additional missense SNP (G > T) was detected in the WGS of the affected dog in exon 23 (chrX: 122,907,870) but confirmed to be homozygous for the reference allele through Sanger sequencing. No structural variants were detected in any of the candidate gene loci.

We manually detected seven improperly paired reads in intron 22 using SAMtools tview (Table S3). All reads in intron 22 were in the forward orientation. Five of the reverse read mates mapped to a location distal to the *FVIII* gene, suggesting a possible inversion event similar to the prevalent intron 22 inversion seen in humans. Two separate, long range PCR tests that were designed to overlapping fragments that

together span the entire intron 22 were performed. The expected fragment sizes (~11kb and 5kb) were observed for both affected and unaffected Australian Kelpies, confirming that the improperly paired reads were caused by an alignment artefact and are not of a biological cause.

5.5. Discussion

Haemophilia A is considered the most common bleeding disorder loci in both humans and dogs. Although it affects a variety of pure and mixed breeds, the underlying genetic cause of disease is unknown for many cases and thus genetic testing is unavailable for many breeds. Of the genetic mutations that are known, similarities to the human disorder are apparent. In both species, a variety of inherited and spontaneously occurring mutations in the *FVIII* gene can either be associated with mild, moderate or severe disease. The most commonly reported mutation in humans involves a large fragment of chromosome X at the telomere, which is inverted causing disruption in the *FVIII* gene with a breakpoint occurring in intron 22. A similar event has also been observed in dogs with severe haemophilia A.

Here we report occurrence of severe haemophilia A in purebred Australian Kelpies. The two affected dogs (USCF305, USCF311) presented with classic clinical signs of haemophilia A, including low FVIII activity (<1.5%) in addition to reduced FIX activity in USCF305. Only males were affected in the litter, which conforms to the X-linked recessive mode of inheritance of haemophilia A. We screened for mutations in the *FVIII* gene. To ensure that we tested for a comprehensive list of putative loci, 13 other genes that have a known association with a bleeding tendency phenotype in humans were included in the analysis and hence we also considered the possibility of autosomal recessive inheritance.

Using WGS data of affected Kelpie USCF305 and 11 unrelated control dogs, we detected 37 intronic SNP variants in 7 of the selected candidate genes that fit the expected mode of inheritance. These were predicted to have a modifier (one with low impact in *ITGA2B* at CanFam 3.1 chr9: 19,058,091) effect on the corresponding

proteins and includes one intronic mutation in *FVIII* (CanFam 3.1, chrX: 122,916,938). Whilst intronic mutations have no obvious impact on protein function, several point mutations and a deletion in intron 22 have been associated with severe haemophilia A in people (http://www.factorviii-db.org/). As we had not collected transcriptome samples and the affected dogs were deceased, we were unable to confirm the effect of putative variants on protein function. No indel or structural variants that were concordant to the expected mode of inheritance were detected.

As haemophilia A is known to be caused by defects in *FVIII*, we performed comprehensive screening at this locus. Using 173, 650 SNP genotyping array data obtained from the CanineHD BeadChip, we inferred *FVIII* haplotypes in two cases and in 10 apparently healthy relatives. This revealed that the maternal grandsire (USCF316) had the same apparent haplotype as the affected pups despite being healthy. If *FVIII* was the causal locus, this suggests that the mutation occurred sporadically, either in the grandsire's sperm or mother's oocyte (USCF1290). Presuming the latter situation, the mother would not appear to be a carrier in direct genetic testing and we took this into consideration throughout this study.

The FVIII gene was manually screened at each of the 26 exons and Sanger sequencing was performed for exons that were insufficiently covered in WGS data to ensure that all coding sequence was assessed. Two non-synonymous SNPs were found in exons 14 and 23, however were also present in other healthy WGS samples and did not conform to the expected mode of inheritance. We made the decision to test for a mutation similar to the most prevalent intron 22 inversion observed in humans, especially because a similar event in severely affected Miniature Schnauzers and Irish Setters was found (Hough *et al.* 2002; Lozier *et al.* 2002). Improperly paired reads in intron 22 of USCF305 provided some evidence for this mutation. The mates of five of these forward reads that mapped to intron 22 mapped 387,826 - 387,858 bp telomeric to the *FVIII* gene (Table S3). Interestingly, there is also a SNP variant following the expected mode of inheritance in this intron (Table 5.1). Without transcript data, we used long range PCR to test for a putative breakpoint within intron 22, however found no evidence for an inversion event as fragments of the expected size were amplified in affected dogs.

The detection of rare variants that are causative of disease is extremely challenging and mutation detection studies are often underpowered due to low case numbers. With canine genomes, regions that are highly associated with disease can be identified with ~10 case and ~10 control samples for simple Mendelian traits (Karlsson *et al.* 2007). Whilst this is easily achieved for common traits, it is often unachievable for rare traits without colony creation, a time consuming and expensive task with obvious animal welfare implications. The challenge in identifying candidate variants causative for diseases is exacerbated by several technical limitations associated with current sequencing technologies and resources available. Sequencing depth, raw read mappability to the reference genome and completeness of annotation vary across the genome (Sims *et al.* 2014). For example, GC-rich regions which are characteristic at transcription start sites of protein coding genes are prone to low sequencing depth and hence, are likely to be less represented than other genomic contexts by whole genome sequencing data.

For human medicine, where rapid diagnosis and personalized treatment for people with rare diseases is more pertinent, researchers are using WGS and whole exome sequencing from parent-offspring trios as a powerful approach to map causative variants in these scenarios (Zhu *et al.* 2015). This approach has more recently been applied in two independent canine studies with success (Sayyab *et al.* 2016; Chew, Haase, Bathgate, *et al.* 2017) and should be considered in the experimental design of future rare disease mapping studies. The emergence and reducing costs of next generation sequencing technologies have enabled these successes, however, the overall the diagnostic rate for rare diseases still remains relatively low at 25 – 50% (Yang *et al.* 2014; Ankala *et al.* 2015; Taylor *et al.* 2015; Chong *et al.* 2015; Cummings *et al.* 2017). For rare diseases where mutations are sporadically occurring, the issue of achieving statistical power in mutation detection studies is exacerbated because the disease may not appear to have Mendelian inheritance.

Although each rare disease affects a small percentage of individuals in a population, when all types of rare diseases are considered collectively, they are a common problem. The challenges presented in this study which are common in rare disease

research highlight the need for the development of mapping strategies and the careful consideration in the experimental design of future research in this field.

5.6. References

Ankala, A., C. da Silva, F. Gualandi, A. Ferlini, L. J. H. Bean et al., 2015 A comprehensive genomic approach for neuromuscular diseases gives a high diagnostic yield. Ann. Neurol. 77: 206–214.

Arnott, E. R., L. Peek, J. B. Early, A. Y. H. Pan, B. Haase et al., 2015 Strong selection for behavioural resilience in Australian stock working dogs identified by selective sweep analysis. Canine Genet. Epidemiol. 2: 6.

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. Bioinform. 43: 1-33.

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: 360–364.

Bolton-Maggs, P. H. B., and K. J. Pasi, 2003 Haemophilias A and B. Lancet 361: 1801–1809.

Brooks, M., 1999 A review of canine inherited bleeding disorders: biochemical and molecular strategies for disease characterization and carrier detection. J. Hered. 90: 112–118.

Brooks, M. B., R. MacNguyen, R. Hall, R. Gupta, and J. G. Booth, 2008 Indirect carrier detection of canine haemophilia A using factor VIII microsatellite markers. Anim. Genet. 39: 278–283.

Chew, T., B. Haase, R. Bathgate, C. E. Willet, M. K. Kaukonen et al., 2017 A Coding Variant in the Gene Bardet-Biedl Syndrome 4 (BBS4) Is Associated with a Novel Form of Canine Progressive Retinal Atrophy. G3 (Bethesda). 7: 2327–2335.

Chiang, C., R. M. Layer, G. G. Faust, M. R. Lindberg, D. B. Rose et al., 2015 SpeedSeq: ultra-fast personal genome analysis and interpretation. Nat. Methods 12: 966–968.

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: 199–215.

Christopherson, P. W., L. M. Bacek, K. B. King, and M. K. Boudreaux, 2014 Two novel missense mutations associated with hemophilia A in a family of Boxers, and a German Shepherd dog. Vet. Clin. Pathol. 43: 312–316.

Cummings, B. B., J. L. Marshall, T. Tukiainen, M. Lek, S. Donkervoort et al., 2017 Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. Sci. Transl. Med. 9: 1-25.

DePristo, M. A., E. Banks, R. Poplin, K. V Garimella, J. R. Maguire et al., 2011 A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. 43: 491–498.

Dunning, M. D., G. F. Averis, H. Pattinson, M. Targett, S. Cade et al., 2009 Haemophilia A (factor VIII deficiency) in a litter of Weimaraners. J. Small Anim. Pract. 50: 357–359.

Faust, G. G., and I. M. Hall, 2014 SAMBLASTER: fast duplicate marking and structural variant read extraction. Bioinformatics 30: 2503–2505.

Graw, J., H.-H. Brackmann, J. Oldenburg, R. Schneppenheim, M. Spannagl et al., 2005 Haemophilia A: from mutation analysis to new therapies. Nat. Rev. Genet. 6: 488–501.

Hough, C., S. Kamisue, C. Cameron, C. Notley, S. Tinlin et al., 2002 Aberrant splicing and premature termination of transcription of the FVIII gene as a cause of severe canine hemophilia A: similarities with the intron 22 inversion mutation in human hemophilia. Thromb. Haemost. 87: 659–665.

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

Layer, R. M., C. Chiang, A. R. Quinlan, and I. M. Hall, 2014 LUMPY: a probabilistic framework for structural variant discovery. Genome Biol. 15: R84.

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

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: 803–819.

Lozier, J. N., A. Dutra, E. Pak, N. Zhou, Z. Zheng et al., 2002 The Chapel Hill hemophilia A dog colony exhibits a factor VIII gene inversion. Proc. Natl. Acad. Sci. U. S. A. 99: 12991–12996.

Lozier, J. N., M. T. Kloos, E. P. Merricks, N. Lemoine, M. H. Whitford et al., 2016 Severe Hemophilia A in a Male Old English Sheep Dog with a C -> T Transition that Created a Premature Stop Codon in Factor VIII. Comp. Med. 66: 405–411.

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: 1297–1303.

McLaren, W., B. Pritchard, D. Rios, Y. Chen, P. Flicek et al., 2010 Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics 26: 2069–2070.

Mischke, R., C. Wilhelm, a Czwalinna, M. Varvenne, K. Narten et al., 2011 Canine haemophilia A caused by a mutation leading to a stop codon. Vet. Rec. 169: 496b.

Ohmori, T., Y. Nagao, H. Mizukami, A. Sakata, S. Muramatsu et al., 2017 CRISPR/Cas9-mediated genome editing via postnatal administration of AAV vector cures haemophilia B mice. Sci. Rep. 7: 4159.

Pan, A. Y. H., C. M. Wade, R. M. Taylor, and P. Williamson, 2017 Exclusion of known gene loci for cerebellar abiotrophy in the Australian Working Kelpie. Anim. Genet. 48: 730–732.

Repessé, Y., M. Slaoui, D. Ferrandiz, P. Gautier, C. Costa et al., 2007 Factor VIII (FVIII) gene mutations in 120 patients with hemophilia A: detection of 26 novel mutations and correlation with FVIII inhibitor development. J. Thromb. Haemost. 5: 1469–1476.

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

Sayyab, S., A. Viluma, K. Bergvall, E. Brunberg, V. Jagannathan et al., 2016 Whole-Genome Sequencing of a Canine Family Trio Reveals a FAM83G Variant Associated with Hereditary Footpad Hyperkeratosis. G3 Genes, Genomes, Genet. 6: 521-527.

Sims, D., I. Sudbery, N. E. Ilott, A. Heger, and C. P. Ponting, 2014 Sequencing depth and coverage: key considerations in genomic analyses. Nat. Rev. Genet. 15: 121–132.

Tantawy, A. a. G., 2010 Molecular genetics of hemophilia A: Clinical perspectives. Egypt. J. Med. Hum. Genet. 11: 105–114.

Taylor, J. C., H. C. Martin, S. Lise, J. Broxholme, J.-B. Cazier et al., 2015 Factors influencing success of clinical genome sequencing across a broad spectrum of disorders. Nat. Genet. 47: 717–726.

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: e1002316.

Ward, P., and C. E. Walsh, 2017 Expert Review of Hematology Current and future prospects for hemophilia gene therapy Current and future prospects for hemophilia gene therapy. Expert Rev. Hematol. 9: 649–659.

Yang, Y., D. M. Muzny, F. Xia, Z. Niu, R. Person et al., 2014 Molecular Findings Among Patients Referred for Clinical Whole-Exome Sequencing. JAMA 312: 1870.

Yen, C. T., M. N. Fan, Y. L. Yang, S. C. Chou, I. S. Yu et al., 2016 Current animal models of hemophilia: the state of the art. Thromb. J. 14: 22.

Zhu, X., S. Petrovski, P. Xie, E. K. Ruzzo, Y.-F. Lu et al., 2015 Whole-exome sequencing in undiagnosed genetic diseases: interpreting 119 trios. Genet. Med. 17: 774–781.

Chapter 6. General Discussion and Conclusions

The accumulation of new spontaneously occurring variation in the genome is known to drive evolution and contribute to disease. In this thesis, we explored the role of *de novo* germline mutations in the evolution of and diseases occurring in the domestic dog (*Canis lupus familiaris*). We used next generation sequencing (NGS) technologies to interrogate the whole genomes of canine samples to: estimate the *de novo* mutation rate in dogs (chapter 3); identify a putative mutation in a potentially novel canine progressive retinal atrophy gene that is associated with blindness in the Hungarian Puli (chapter 4); and explore the genetics of spontaneously occurring severe haemophilia A in Australian Kelpies (chapter 5). Whilst NGS platforms are extremely powerful in their high throughput and in the abundance of data that they can generate, important biological gene variants can be missed due to technical limitations of the technology and bioinformatics methodology employed to the data. Because of this, we first carried out a performance comparison of single nucleotide variant (SNV) detection methodologies that was applied to our specific data type to ensure that we utilized optimal bioinformatics pipelines in the subsequent chapters (chapter 2). In this chapter, we discuss the outcomes and knowledge gained from each experimental chapter separately.

6.1. Conclusions from chapter 2

A preliminary study prompted us to initiate the study in chapter 2 where we compared the performance of 10 SNV calling pipelines using five popular variant callers on the data used in this thesis. In the preliminary study, we tested several pipelines to determine the optimal bioinformatics methodology that would allow us to achieve an accurate estimation of the *de novo* per base mutation rate in dogs. Achieving a high level of accuracy requires extremely high calling specificity so that true *de novo* mutation loci are detected without the inclusion of other variants, including population variation, sequencing errors and variants caused by other technical artefacts. Yet, increasing specificity is often at the cost of sensitivity, and without sufficient sensitivity, *de novo* variants may not be captured due to the rarity of new mutation events. To

obtain an adequate balance between sensitivity, specificity and accuracy, we systematically compared 10 SNV calling pipelines in chapter 2. The metrics provided can be used to formulate optimal variant calling pipelines tailored for all other types of studies that utilize similar datasets, specifically, studies with Illumina NGS sequencing data, relatively small sample sizes and average sample coverage (~10X).

In chapter 2 we compared five popular SNV callers: FreeBayes (Garrison and Marth 2012); the Genome Analysis Tool-kit's Haplotype Caller (GATK HC) and Unified Genotyper (GATK UG) (McKenna *et al.* 2010); SAMtools (Li *et al.* 2009); and VarScan (Koboldt *et al.* 2013). We ran each variant caller without any additional quality filtering (raw pipeline), and then applied recommended hard filtering (filtered pipeline) where genotypes are considered as 'not called' if they do not meet certain quality metric criteria (Van der Auwera *et al.* 2013; Koboldt *et al.* 2013; Garrison 2015; Willet *et al.* 2015). As many other studies have observed (Yu *et al.* 2013; Cheng *et al.* 2014; De Summa *et al.* 2015; Willet, Haase, *et al.* 2015a), we found that the level of minimum coverage requirement parameter had a major impact on genotyping accuracy rates, estimated sensitivity and estimated specificity of each variant caller. The differences in the measured metrics were greatest between the pipelines with no minimum coverage requirement. The raw VarScan pipeline outperformed the other nine pipelines at minimum coverage requirement levels less than 10X in this study. As the minimum coverage requirement increased, the pipelines performed more similarly in accuracy, sensitivity and specificity, except for two underperforming pipelines (FreeBayes and VarScan with filters applied). There was no clear overall outperforming pipeline at minimum coverage requirement levels of over 10X. There is a common agreement that genotypes can be called with a sufficient level of confidence at sites with at least 10X coverage (Koboldt *et al.* 2013).

Applying hard filters that were recommended by other studies (Van der Auwera *et al.* 2013; Koboldt *et al.* 2013; Garrison 2015; Willet *et al.* 2015) in variant calling pipelines generally did not improve genotyping accuracy for the dataset used in this study. The genotyping accuracy was greatly affected at loci that had low coverage. As the minimum coverage requirements of the algorithm were increased, hard filtering became

more effective and the differences between the accuracies of the raw and pipelines including hard filters became minimal. However, we observed only two variant callers (GATK HC and SAMtools) that had slightly improved genotyping accuracy at higher levels of minimum coverage requirement (over 14X).

We found that achieving higher sensitivity always costs in calling specificity, and vice versa. An optimal variant calling pipeline should be unique to each specific project depending on the nature of the sample data and the project goals. For example, a project with the goals of identifying a causative genetic variant for a phenotype may prioritise sensitivity over specificity, within reason, where there is a feasible number of putative variants to further investigate. As previously described, we used the results of chapter 2 to obtain an optimal pipeline for estimating the canine per generation *de novo* mutation rate, a primary aim of the research conducted in chapter 3.

The additional benefit of sequencing at higher depths of over 10X for accurate SNV calling drastically reduces with increasing depth of coverage. Realistically, achieving high levels of per sample coverage requires continued decline in sequencing costs. For the same cost, many researchers will still opt to sequence more samples at a reduced coverage, as opposed to fewer samples at higher coverage (Le and Durbin 2011; Sims et al. 2014; Gilly et al. 2017). The largest sequencing company, Illumina, achieved its promise to decrease the cost of sequencing a single human genome for ~$1,000 USD at 30X depth of coverage with the release of the HiSeq X Ten platform in 2014. Illumina has since promised to reduce this cost to just $100, however, Schwarze et al. has found little evidence for this cost reduction in whole exome and genome studies conducted from 2013 - 2017 (Schwarze *et al.* 2018). The deceleration of cost reduction may be an intended business choice rather than restriction by technical limitations, as Illumina already owns the largest market share in sequencing platforms.

Despite the ever-increasing affordability of short read NGS, these technologies are still limited by their ability to resolve other types of variants accurately. For this reason, we were restricted to analysing SNVs in chapter 2 and 3 in this thesis. This limitation provides an opportunity for companies such as Pacific Biosciences and Oxford

Nanopore that produce longer read platforms to compete in the current market. Resolving GC-rich regions, repetitive, insertion-deletion (indel) and copy number variants (CNVs) greater than 500 bp are more successful with long read sequencing technologies (Chaisson *et al.* 2014; Reuter *et al.* 2015; Pollard *et al.* 2018). In addition, *de novo* assemblies with long reads provide more complete and accurate coverage of whole genomes than *de novo* assemblies using short reads (Pollard *et al.* 2018). *De novo* assembly methods organise reads without the need for a reference genome. Biases and restrictions associated with using a reference genome (see section 1.4.2) can thus be avoided. Omitting cost, the current high per base error rates of long read (11 – 38.2%) compared to short read (0.11 – 0.28%) sequencing limit the potential for long read platforms to completely replace short read platforms (Minoche *et al.* 2011; Pollard *et al.* 2018).

In the foreseeable future, more complete interrogation of whole genomes by accurate identification of all variant types and sequence contexts will be possible by combining short and long read sequencing data. Efficiency needs to be improved for both technologies, especially long read sequencing, which is still unobtainable for many researchers. With access to both short and long read sequencing data and subsequently, more accurate genotyping of all variant types, applications and research questions that utilise sequencing technologies can be broadened.

6.2. Conclusions from chapter 3

New DNA mutations are the primary source of genetic diversity and enable evolution to occur. In only a few hundred years, hundreds of dog breeds that specialise in a variety of morphological, physiological and behavioural traits have been created. To improve our understanding of canine evolution, we characterised the germline *de novo* mutation rates and variant distributions throughout the canine genome in chapter 3. Through direct observation of *de novo* mutations by whole genome sequencing of parent-offspring trios, we were able to estimate the per-base, per-generation mutation rate to be 3.9×10^{-8} (95% confidence interval $3.5 – 4.4 \times 10^{-8}$). In the canine genome which is 2.4 gigabases in size, this is equivalent to 81 – 112 *de novo* nucleotide changes in each

individual genome per meiosis. Transitions outnumber transversions by 2.3 (95% confidence interval 1.3 – 3.3), which is similar to the transition to transversion ratio in other mammals including humans (2.2), mice (2.1) and chimpanzees (2.2) (Campbell and Eichler 2013; Venn *et al.* 2014; Uchimura *et al.* 2015; Narasimhan *et al.* 2017).

Our estimate of the per base mutation rate in dogs is slightly higher than the reported estimates for other species including humans, mice, chimpanzees and birds by 1 – 4 x 10^{-9} nucleotide changes per generation (Campbell and Eichler 2013; Venn *et al.* 2014; Uchimura *et al.* 2015; Smeds *et al.* 2016; Narasimhan *et al.* 2017). This elevated rate coupled with relatively large litter sizes (5.4 puppies on average for purebred dogs) and shorter generation times in comparison to other studied species may have facilitated rapid phenotypic diversification of the dog.

To understand the possible effects of *de novo* mutations on phenotype in the dog, we categorised observed mutations into seven genomic features based on their physical position in the annotated reference genome. The categories included protein-coding, CpG island, intergenic, intronic, conserved, 3' untranslated region (UTR) and the 5' UTR. We did not find any significant bias towards mutation in any particular feature, except that mutations are significantly less likely in the 3' UTR compared to intronic and intergenic regions ($P_{T\text{-}TEST} < 0.05$). This is unlike other species, who have found that CpG dinucleotides are highly hypermutable compared to other genomic contexts (10 – 30 times, depending on the animal) (Kondrashov 2002; Lynch 2010; Keightley *et al.* 2011; Hodgkinson and Eyre-Walker 2011; Kong *et al.* 2012; Smeds *et al.* 2016; Narasimhan *et al.* 2017).

The accuracy in the estimated per base mutation rate for dogs derived in chapter 3 is likely to be impacted by technical quirks that are unique to our dataset. Sequencing difficulty of GC rich contexts in Illumina NGS data, relatively low sample sequencing coverage and strict variant calling filtering criteria resulted in a low observation of GC contexts for this study. Due to our limited ability to observe GC rich contexts, there is a possible underestimate in the average per base mutation rate in dogs. We expect there to be an underestimate because the mutation rate in other animal species in CpG

islands is reported to be 10-30 times higher compared with non-CpG sites (Kondrashov 2002; Lynch 2010; Kong *et al.* 2012; Narasimhan *et al.* 2017). To overcome these technical limitations, a greater a sequencing depth of over 30X has previously been recommended to obtain the high specificity and sensitivity required for accurate SNV calling, as well as appropriate representation of all genomic sequence contexts (Cheng *et al.* 2014; Francioli *et al.* 2017). With additional utilization of long read sequencing technologies, characteristics other types of *de novo* variants including indels and CNVs can be characterised accurately in a range of genomic contexts. As discussed at the end of section 6.1, the opportunity for sequencing samples at a higher depth or to obtain long read sequencing data is limited by cost.

6.3. Conclusions from chapter 4

De novo mutations that are associated with disease are notoriously difficult to detect through common mapping methods such as genome wide association analyses (Lee *et al.* 2014). New variants are not in linkage disequilibrium to genetic markers that are typically used to identify genetic variants associated with phenotypes. In addition, spontaneously occurring diseases caused by new mutations are often rare and limited to a very small number of individuals within families. In chapter 4, we studied spontaneously occurring progressive retinal atrophy (PRA) in three related Hungarian Puli dogs. Although PRA is typically an autosomal recessive disorder in other breeds, there was no prior history of PRA in the subject Hungarian Puli family.

In the first part of this chapter, we performed screening of a comprehensive list of reported canine PRA genes using NGS data (Miyadera *et al.* 2012; Downs, Hitti, *et al.* 2014; Winkler *et al.* 2016). We found no coding variants in 53 candidate loci for the phenotype of interest. To ensure that we captured all possible coding variants in likely candidate genes, we used Sanger sequencing to sequence the exons that were not sufficiently covered in NGS data of any of the 53 genes that were segregating in an autosomal recessive pattern. These genes were identified using genotyping array markers that followed the expected inheritance pattern and co-located with candidate

genes. With no potential damaging variants identified and no family history of PRA, we considered the possibility that PRA in this family of Hungarian Puli was novel.

We further assessed positional candidate genes in loci segregating in an autosomal recessive pattern and identified a single nonsense SNP in exon 2 of *BBS4* that was significantly associated with the disorder (c.58.A > T, $P_{CHISQ} = 3.43e^{14,}$ n = 103). Dysfunctional *BBS4* is known to cause Bardet-Biedl Syndrome in people, a ciliopathy which is characterised by many phenotypes including retinitis pigmentosa, obesity and infertility (Katsanis *et al.* 2002; Mykytyn *et al.* 2004; Wei *et al.* 2012). We also found evidence that the identified nonsense SNP could cause syndromic disease, as affected Puli were anecdotally obese, and the sole intact male was confirmed to be infertile primarily due to morphologically abnormal flagella. This is the first report of *BBS4* and its involvement in canine PRA.

Since the emergence of NGS technologies, researchers have recognised and demonstrated the potential for these platforms to identify low frequency, rare and *de novo* variants, particularly through sequencing of parent-offspring samples (Buermans and den Dunnen 2014; Zhu *et al.* 2015; Francioli *et al.* 2017; Sayyab *et al.* 2016). We applied this technique of parent-offspring whole genome sequencing here to successfully identify a rare SNP that is strongly associated with a potentially novel form of *BBS4*, whilst contributing to research in canine PRA. Fast identification of spontaneous disease-causing variants in turn allows for DNA screening protocols to commence earlier. Subsequently, the identified *BBS4* nonsense SNP associated with PRA can be eliminated from Hungarian Puli dogs more rapidly.

The manuscript presented in section 4.2 is the second report implicating a Bardet-Biedl Syndrome gene with canine PRA. Besides *BBS4*, *TTC8* was found to be associated with PRA in Golden Retriever Dogs (Downs, Wallin-Håkansson, *et al.* 2014). Our study and Downs et al. were opportunistic and our ability to obtain data was restricted by access and need to maintain the welfare of the pet dogs involved. Opportunistic studies enable scientists to understand diseases and how they arise in a natural context.

Researchers are also able to provide pet owners with a diagnosis and tests to manage the disease, which are particularly pertinent to animal breeders.

Despite the benefits of opportunistic research, they lack a controlled environment and perpetual research to further understand the disease and explore treatment possibilities is limited by access to the samples. As we experienced, Downs et al. had limited opportunities to thoroughly explore other known Bardet-Biedl phenotypes found in people with the disease, or observed in BBS knockout gene mouse models (Nishimura *et al.* 2004; Iannaccone *et al.* 2005; Benzinou *et al.* 2006; Swiderski *et al.* 2007; Aksanov *et al.* 2014). Establishing a canine colony for research could confirm whether the *BBS4* or *TTC8* are associated with syndromic diseases, as in humans and mice. Colonies could also present as large animal models for the human disease counterpart. However, with obvious animal welfare consequences associated with maintaining a research colony, the benefit of this research needs to be carefully evaluated.

6.4. Conclusion to chapter 5

In chapter 5, we investigated the genetic basis of severe haemophilia A that was presented in two Australian Kelpie purebred littermates. Haemophilia A is a rare disease that is characterised by uncontrollable bleeding and has been reported in several species including dogs and people (Brooks 1999; Graw *et al.* 2005). In people, over 2,015 distinct, causative mutations for haemophilia A have been identified in the factor VIII gene (Graw *et al.* 2005; Repessé *et al.* 2007; Tantawy 2010). The gene encodes for coagulation factor VIII (FVIII) and is necessary for the successful operation of the coagulation cascade. Disease is generally inherited in an X-linked recessive pattern, however, a third of all human haemophilia A cases occur sporadically (Crow 2000; Graw *et al.* 2005). Sporadic haemophilia A has also been reported in several dog breeds (Brooks 1999). The present study is the first to our knowledge to report haemophilia A in the Australian Kelpie breed and in a sporadic form, as there was no prior family history of bleeding tendencies in the family.

As Haldane was first to realise, the mutation rate is higher in male germ cells than in female germ cells (Haldane 1935, 1946; Crow 2000; Nachman 2004). He made this hypothesis because sporadic haemophilia A was often associated with an apparently heterozygous, rather than a homozygous normal mother (Haldane 1935). Through genotyping array data of the affected Australian Kelpie family, we also observed this pattern (Figure 5.1). The maternal grandfather of the two cases had the apparently affected haplotype, despite being in the clinically normal range for coagulation assays for FVIII clotting. This suggests that a spontaneously occurring mutation occurred in the grandfather's germ cell.

In chapter 5, we screened for putative mutations in FVIII as well as 13 other bleeding disorder loci in NGS of one case and 11 unrelated Australian Kelpie control samples. Whilst we did not detect any damaging variants in the candidate genes selected including the FVIII gene, 37 intronic SNP variants in 7 of the candidate genes were found. As we described in the conclusions for chapter 4, research in sporadically occurring or rare diseases are challenging. In chapter 5, we were tested with low sample numbers and limited access to sample type. As we found in chapter 4, the value in creating a canine colony to provide the required samples to sufficiently model the haemophilia A in dogs should be evaluated. Although it was not possible to access additional case samples, access to cDNA would reveal the functional impact of the putative variants identified. Whole genome sequences or transcriptomes of the parents and the additional case would greatly enhance the power to detect damaging variants. Potential use of NGS technologies in a clinical setting for obtaining a genetic diagnosis for spontaneously occurring diseases can also be faced with similar challenges described.

6.5. Final remarks

In this thesis, we used NGS technologies to assess the contribution of germline *de novo* mutations in canine evolution and disease. We systematically compared popular, recommended variant calling pipelines to provide benchmark performance metrics that can be used as a guide to develop an optimal pipeline that is tailored to a specific study

depending on their priorities for accuracy, sensitivity and specificity. We used these results to develop our own pipeline that enabled us to directly observe characteristics of *de novo* mutations and to estimate the per base, per generation germline mutation rate in the domestic dog. Finally, we contribute to research in two canine rare diseases that were potentially caused by *de novo* mutations and the outcomes of the research highlighted potential applications and challenges with using NGS in clinical diagnosis for spontaneously occurring diseases.

6.6. References

Aksanov, O., P. Green, and R. Z. Birk, 2014 BBS4 directly affects proliferation and differentiation of adipocytes. Cell. Mol. Life Sci. 71: 3381–3392.

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. Bioinform. 43: 1-33.

Benzinou, M., A. Walley, S. Lobbens, M.-A. Charles, B. Jouret et al., 2006 Bardet-Biedl syndrome gene variants are associated with both childhood and adult common obesity in French Caucasians. Diabetes 55: 2876–2882.

Brooks, M., 1999 A review of canine inherited bleeding disorders: biochemical and molecular strategies for disease characterization and carrier detection. J. Hered. 90: 112–118.

Buermans, H. P. J., and J. T. den Dunnen, 2014 Next generation sequencing technology: Advances and applications. Biochim. Biophys. Acta - Mol. Basis Dis. 1842: 1932–1941.

Campbell, C. D., and E. E. Eichler, 2013 Properties and rates of germline mutations in humans. Trends Genet. 29: 575–584.

Chaisson, M. J. P., J. Huddleston, M. Y. Dennis, P. H. Sudmant, M. Malig et al., 2015 Resolving the complexity of the human genome using single-molecule sequencing. Nature 517: 608–611.

Cheng, A. Y., Y. Y. Teo, and R. T. H. Ong, 2014 Assessing single nucleotide variant detection and genotype calling on whole-genome sequenced individuals. Bioinformatics 30: 1707–1713.

Crow, J. F., 2000 The origins, patterns and implications of human spontaneous mutation. Nat. Rev. Genet. 1: 40–47.

Downs, L. M., R. Hitti, S. Pregnolato, and C. S. Mellersh, 2014 Genetic screening for PRA-associated mutations in multiple dog breeds shows that PRA is heterogeneous within and between breeds. Vet. Ophthalmol. 17: 126–130.

Downs, L. M., B. Wallin-Håkansson, T. Bergström, 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.

Francioli, L. C., M. Cretu-Stancu, K. V Garimella, M. Fromer, W. P. Kloosterman et al., 2017 A framework for the detection of de novo mutations in family-based sequencing data. Eur. J. Hum. Genet. 25: 227–233.

Francioli, L. C., M. Cretu-Stancu, K. V Garimella, M. Fromer, W. P. Kloosterman et al., 2016 A framework for the detection of de novo mutations in family-based sequencing data Genome of the Netherlands Consortium. Eur. J. Hum. Genet. 25147: 227–233.

Garrison, E., 2015 Freebayes in Depth: Model, Filtering, and Walk-Through. Presented at the University of Cambridge, May 2015. Available at:https://wiki.uiowa.edu/download/attachments/145192256/erik%20garrison%20-%20iowa%20talk%202.pdf?api=v2.

Garrison, E., and G. Marth, 2012 Haplotype-based variant detection from short-read sequencing. https://arxiv.org/abs/1207.3907v2.

Graw, J., H. H. Brackmann, J. Oldenburg, R. Schneppenheim, M. Spannagl et al., 2005 Haemophilia A: from mutation analysis to new therapies. Nat. Rev. Genet. 6: 488–501.

Haldane, J. B. S., 1935 The rate of spontaneous mutation of a human gene. J. Genet. 31: 317–326.

Haldane, J. B. S., 1946 The Mutation Rate of the Gene for Haemophilia, and Its Segregation Ratios in Males and Females. Ann. Eugen. 13: 262–271.

Hodgkinson, A., and A. Eyre-Walker, 2011 Variation in the mutation rate across mammalian genomes. Nat. Rev. Genet. 12: 756–766.

Iannaccone, A., K. Mykytyn, A. M. Persico, C. C. Searby, A. Baldi et al., 2005 Clinical evidence of decreased olfaction in Bardet-Biedl syndrome caused by a deletion in the BBS4 gene. Am. J. Med. Genet. A 132A: 343–346.

Katsanis, N., E. R. Eichers, S. J. Ansley, R. A. Lewis, H. Kayserili et al., 2002 BBS4 is a minor contributor to Bardet-Biedl syndrome and may also participate in triallelic inheritance. Am. J. Hum. Genet. 71: 22–29.

Keightley, P. D., L. Eöry, D. L. Halligan, and M. Kirkpatrick, 2011 Inference of mutation parameters and selective constraint in mammalian coding sequences by approximate Bayesian computation. Genetics 187: 1153–1161.

Koboldt, D. C., D. E. Larson, and R. K. Wilson, 2013 Using VarScan 2 for Germline Variant Calling and Somatic Mutation Detection. Curr. Protoc. Bioinforma. 44: 15.4.1-17.

Kondrashov, A. S., 2002 Direct estimates of human per nucleotide mutation rates at 20 loci causing mendelian diseases. Hum. Mutat. 21: 12–27.

Kong, A., M. L. Frigge, G. Masson, S. Besenbacher, P. Sulem et al., 2012 Rate of de novo mutations and the importance of father’s age to disease risk. Nature 488: 471–5.

Lee, S., G. R. Abecasis, M. Boehnke, and X. Lin, 2014 Rare-Variant Association Analysis: Study Designs and Statistical Tests. Am. J. Hum. Genet. 95: 5–23.

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

Lynch, M., 2010 Rate, molecular spectrum, and consequences of human mutation. Proc. Natl. Acad. Sci. U. S. A. 107: 961–968.

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: 1297–1303.

Minoche, A. E., J. C. Dohm, and H. Himmelbauer, 2011 Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems. Genome Biol. 12: R112.

Miyadera, K., G. M. Acland, and G. D. Aguirre, 2012 Genetic and phenotypic variations of inherited retinal diseases in dogs: The power of within- and across-breed studies. Mamm. Genome 23: 40–61.

Mykytyn, K., R. F. Mullins, M. Andrews, A. P. Chiang, R. E. Swiderski et al., 2004 Bardet-Biedl syndrome type 4 (BBS4)-null mice implicate Bbs4 in flagella formation but not global cilia assembly. Proc. Natl. Acad. Sci. U. S. A. 101: 8664–8669.

Nachman, M. W., 2004 Haldane and the first estimates of the human mutation rate. J. Genet. 31: 235–244.

Narasimhan, V. M., R. Rahbari, A. Scally, A. Wuster, D. Mason et al., 2017 Estimating the human mutation rate from autozygous segments reveals population differences in human mutational processes. Nat. Commun. 8: 303.

Nishimura, D. Y., M. Fath, R. F. Mullins, C. Searby, M. Andrews et al., 2004 Bbs2-null mice have neurosensory deficits, a defect in social dominance, and retinopathy associated with mislocalization of rhodopsin. Proc. Natl. Acad. Sci. U. S. A. 101: 16588–16593.

Pollard, M. O., D. Gurdasani, A. J. Mentzer, T. Porter, and M. S. Sandhu, 2018 Long reads: their purpose and place. Hum. Mol. Genet. 27: R234–R241.

Repessé, Y., M. Slaoui, D. Ferrandiz, P. Gautier, C. Costa et al., 2007 Factor VIII (FVIII) gene mutations in 120 patients with hemophilia A: detection of 26 novel mutations and correlation with FVIII inhibitor development. J. Thromb. Haemost. 5: 1469–1476.

Reuter, J. A., D. V Spacek, and M. P. Snyder, 2015 High-throughput sequencing technologies. Mol. Cell 58: 586–597.

Sayyab, S., A. Viluma, K. Bergvall, E. Brunberg, V. Jagannathan et al., 2016 Whole-Genome Sequencing of a Canine Family Trio Reveals a FAM83G Variant Associated with Hereditary Footpad Hyperkeratosis. G3 (Bethesda) 6: 521-527.

Schwarze, K., J. Buchanan, J. C. Taylor, and S. Wordsworth, 2018 Are whole-exome and whole-genome sequencing approaches cost-effective? A systematic review of the literature. Genet. Med. 20: 1122–1130.

Smeds, L., A. Qvarnström, and H. Ellegren, 2016 Direct estimate of the rate of germline mutation in a bird. Genome Res. 26: 1211–1218.

De Summa, S., G. Malerba, R. Pinto, A. Mori, V. Mijatovic et al., 2015 GATK hard filtering: tunable parameters to improve variant calling for next generation sequencing targeted gene panel data. BMC Bioinformatics. 18: 119.

Swiderski, R. E., D. Y. Nishimura, R. F. Mullins, M. a Olvera, J. L. Ross et al., 2007 Gene expression analysis of photoreceptor cell loss in bbs4-knockout mice reveals an early stress gene response and photoreceptor cell damage. Invest. Ophthalmol. Vis. Sci. 48: 3329–3340.

Tantawy, A. a. G., 2010 Molecular genetics of hemophilia A: Clinical perspectives. Egypt. J. Med. Hum. Genet. 11: 105–114.

Uchimura, A., M. Higuchi, Y. Minakuchi, M. Ohno, A. Toyoda et al., 2015 Germline mutation rates and the long-term phenotypic effects of mutation accumulation in wild-type laboratory mice and mutator mice. Genome Res. 25: 1–10.

Venn, O., I. Turner, I. Mathieson, N. de Groot, R. Bontrop et al., 2014 Strong male bias drives germline mutation in chimpanzees. Science 344: 1272–1275.

Wei, Q., Y. Zhang, Y. Li, Q. Zhang, K. Ling et al., 2012 The BBSome controls IFT assembly and turnaround in cilia. Nat. Cell Biol. 14: 950–957.

Willet, C. E., B. Haase, M. A. Charleston, and C. M. Wade, 2015 Simple, rapid and accurate genotyping-by-sequencing from aligned whole genomes with ArrayMaker. Bioinformatics 31: 599–601.

Winkler, P. A., J. A. Davis, S. M. Petersen-Jones, P. J. Venta, and J. T. Bartoe, 2016 A tool set to allow rapid screening of dog families with PRA for association with candidate genes. Vet. Ophthalmol. 20: 372-376.

Yu, X., S. Sun, F. Collins, L. Brooks, A. Chakravarti et al., 2013 Comparing a few SNP calling algorithms using low-coverage sequencing data. BMC Bioinformatics 14: 274.

Zhu, X., S. Petrovski, P. Xie, E. K. Ruzzo, Y. F. Lu et al., 2015 Whole-exome sequencing in undiagnosed genetic diseases: interpreting 119 trios. Genet. Med. 17: 774–781.

Appendices

Appendix I: Supplementary material for chapter 2

Table S1. Sample information

Breed	Identifier	Genotyping array	Sample type	Sequencing platform	Library	Sequencing lanes	Read length	Number of raw reads	Reads mapped to CanFam 3.1 (%)	Raw average coverage
Australian Cattle Dog	USCF1292	Illumina CanineHD BeadChip	Whole blood	Illumina HiSeq 2000	TruSeq (PCR-free)	1/2 x 2	100	326,493,791	98.4	12.5
Australian Cattle Dog	USCF1293	Illumina CanineHD BeadChip	Whole blood	Illumina HiSeq 2000	TruSeq (PCR-free)	1/2 x 1	101	348,600,508	99.12	13.3
Australian Cattle Dog	USCF1294	Illumina CanineHD BeadChip	Whole blood	Illumina HiSeq 2000	TruSeq (PCR-free)	1/2 x 1	101	381,441,309	99.0	14.5

Miniature Schnauzer	USCF138	Illumina CanineHD BeadChip	Whole blood	Illumina HiSeq 2500	TruSeq (PCR-free)	1/2 x 2	100	273,023,711	98.8	10.2
Miniature Schnauzer	USCF301	Illumina CanineHD BeadChip	Whole blood	Illumina HiSeq 2500	TruSeq (PCR-free)	1/2 x 2	100	422,811,078	99.0	15.9
Miniature Schnauzer	USCF134	Illumina CanineHD BeadChip	Tissue	Illumina HiSeq 2000	TruSeq	1/2 x 1	101	229,564,555	99.6	8.7
Miniature Schnauzer	USCF136	Illumina CanineHD BeadChip	Tissue	Illumina HiSeq 2000	TruSeq	1/2 x 1	101	192,777,994	99.6	7.2
Hungarian Puli	USCF525	Illumina CanineHD BeadChip	Whole blood	Illumina HiSeq 2000	TruSeq (PCR-free)	1/2 x 1	101	197,791,391	99.3	7.5
Hungarian Puli	USCF347	Illumina CanineHD BeadChip	Whole blood	Illumina HiSeq 2000	TruSeq (PCR-free)	1/2 x 1	101	167,359,624	99.3	6.4
Hungarian Puli	USCF516	Illumina CanineHD BeadChip	Whole blood	Illumina HiSeq 2000	TruSeq (PCR-free)	1/2 x 1	101	191,605,356	99.3	7.3

Table S2. Description of filtering parameters used in raw and filtered pipelines for five variant calling software

Software	Pipeline	Filter	Description
FreeBayes	Raw	No indels	Ignore insertion-deletions
		No mnps	Ignore multi-nucleotide polymorphisms
		No-complex	Ignore complex events (composites of other classes of allele types)
		Report monomorphic	Report loci which are non-variant to the reference allele
	Filtered	QUAL > 1	Include sites in the output only if the reported quality of the site is greater than 1. QUAL is the Phred-scaled probability that the variant reported in the ALT field of the VCF file exists in the sequencing data
		QUAL/AO > 10	AO is to observation count of the alternate allele (depth). Include sites if QUAL/AO is greater than 10.
		SAF > 0	Include sites if the alternate allele is present on more than 0 sites on the forward sequencing reads
		SAR > 0	Include sites if the alternate allele is present on more than 0 sites on the reverse sequencing reads

		RPR > 1	Include sites if the reads placed right (number of reads supporting the alternate is balanced to the 3' end) is greater than 1
		RPL > 1	Include sites if the reads placed left (number of reads supporting the alternate is balanced to the 5' end) is greater than 1
GATK HC	Raw	emitRefConfidence GVCF	Emit reference confidence scores
		variant_index_type LINEAR	Type of variant indexing to use
		variant_index_parameter 128000	Variant to pass to the VCF/BCF IndexCreator
		stand_emit_conf 10	Include sites in the output if the emission confidence threshold (Phred-scaled) that the site is possibly variant is greater than 10
		stand_call_conf 30	Include sites if the calling confidence (Phred-scaled) is greateer than 30
		allSites	Report loci which are non-variant to the reference allele
	Filtered	QD > 2.0	Exclude sites if quality of depth (quality score normalized by the allele depth, AD) is greater than 2.0
		FS > 60.0	Exclude sites if the Fisher's exact test (Phred-scaled) is greater than 60.0 to remove sites with evidence of strand bias

		MQ < 40.0	Exclude sites if mapping quality (Phred-scaled) is less than 40.0
		MappingQualityRankSum < -12.5	Exclude sites if the mapping qualities of reads supporting the reference and alternate is less than -12.5. An ideal value is 0, which indicates no difference in quality between alleles
		ReadPosRankSum < -8.0	ReadPosRankSum measures bias in the position of the alleles in the sequencing reads. Exclude sites where ReadPosRankSum is less than -8.0
GATK UG	Raw	stand_emit_conf 10	Include sites in the output if the emission confidence threshold (Phred-scaled) that the site is possibly variant is greater than 10
		stand_call_conf 30	Include sites if the calling confidence (Phred-scaled) is greateer than 30
		glm BOTH	Perform genotype likelihoods calculation for both SNP and indel
	Filtered	QD > 2.0	Exclude sites if quality of depth (quality score normalized by the allele depth, AD) is greater than 2.0
		FS > 60.0	Exclude sites if the Fisher's exact test (Phred-scaled) is greater than 60.0 to remove sites with evidence of strand bias

		MQ < 40.0	Exclude sites if mapping quality (Phred-scaled) is less than 40.0
		HaplotypeScore > 13.0	HaplotypeScore measures evidence of regions with poor quality alignments and is based on the expectation that the sample is diploid. Exclude sites with a HaplotypeScore greater than 13.0
		MappingQualityRankSum < 12.5	Exclude sites if the mapping qualities of reads supporting the reference and alternate is less than -12.5. An ideal value is 0, which indicates no difference in quality between alleles
		ReadPosRankSum < -8.0	Exclude sites if the mapping qualities of reads supporting the reference and alternate is less than -12.5. An ideal value is 0, which indicates no difference in quality between alleles
SAMtools	Raw	p 1	p is the probability that the site is variant. Include sites if p is less than or equal to 1
		c	Use Bayesian inference in variant calling
	Filtered	Q 20	Exclude bases with base quality less than 20
		q 20	Exclude reads with mapping quality less than 20
		C 50	Recommended by SAMtools if mapping quality is overestimated

			for reads containing excessive mismatches
		E	Perform extended base alignment quality calculation (probability of a read being mis-aligned)
		Maximum coverage 2 x average sample coverage	Exclude if the maximum coverage at the loci is greater than 2 times the average coverage in the sample
		c	Use Bayesian inference in variant calling
VarScan	Raw	B	Disable base alignment quality calculation
		p 1	p is the probability that the site is variant. Include sites if p is less than or equal to 1
		c	Use Bayesian inference in variant calling
	Filtered	q 10	Exclude reads with mapping quality less than 10
		B	Disable base alignment quality calculation
		min-avg-qual 15	Include base if quality (Phred-scaled) is greater than 15
		min-reads2 1	Include if minimum number of reads supporting the variant allele is greater than 1
		min-var-freq 0.20	Include variants if the variant allele frequency is greater than 0.20 of the total reads present at the site

	p-value 0.10	Set p-value threshold of 0.10 (Fisher's exact test) which a variant call is deemed significant
	min-freq-for-hom 0.75	Minimum variant allele frequency above which a variant will be called homozygous in a given sample

Table S3. Genotyping rates (n = 10) across 11 minimum coverage requirement levels obtained from raw and filtered pipelines using five different variant callers (FreeBayes, GATK HC, GATK UG, SAMtools and VarScan) compared against genotypes obtained using the CanineHD BeadChip array.

Variant Caller	Filtering type	Minimum Coverage	Number of genotypes called from WGS	Number not called from WGS	Genotyping rate (%)
FreeBayes	Raw	0	1,715,475	1,245	99.927
FreeBayes	Raw	2	1,701,023	15,697	99.086
FreeBayes	Raw	4	1,629,333	87,387	94.910
FreeBayes	Raw	6	1,457,691	259,029	84.911
FreeBayes	Raw	8	1,202,033	514,687	70.019
FreeBayes	Raw	10	924,196	792,524	53.835
FreeBayes	Raw	12	672,794	1,043,926	39.191
FreeBayes	Raw	14	470,366	1,246,354	27.399
FreeBayes	Raw	16	316,495	1,400,225	18.436
FreeBayes	Raw	18	203,059	1,513,661	11.828
FreeBayes	Raw	20	123,641	1,593,079	7.2021
FreeBayes	Filtered	0	1,697,474	19,246	98.879
FreeBayes	Filtered	2	1,683,039	33,681	98.038
FreeBayes	Filtered	4	1,612,095	104,625	93.906
FreeBayes	Filtered	6	1,442,670	274,050	84.036
FreeBayes	Filtered	8	1,190,351	526,369	69.339
FreeBayes	Filtered	10	916,316	800,404	53.376
FreeBayes	Filtered	12	668,160	1,048,560	38.921
FreeBayes	Filtered	14	467,913	1,248,807	27.256
FreeBayes	Filtered	16	315,338	1,401,382	18.369

FreeBayes	Filtered	18	202,584	1,514,136	11.801
FreeBayes	Filtered	20	123,451	1,593,269	7.191
GATK HC	Raw	0	1,715,726	994	99.942
GATK HC	Raw	2	1,700,986	15,734	99.083
GATK HC	Raw	4	1,629,061	87,659	94.894
GATK HC	Raw	6	1,457,226	259,494	84.884
GATK HC	Raw	8	1,201,509	515,211	69.989
GATK HC	Raw	10	923,690	793,030	53.806
GATK HC	Raw	12	672,373	1,044,347	39.166
GATK HC	Raw	14	470,047	1,246,673	27.381
GATK HC	Raw	16	316,261	1,400,459	18.422
GATK HC	Raw	18	202,908	1,513,812	11.820
GATK HC	Raw	20	123,542	1,593,178	7.196
GATK HC	Filtered	0	1,706,809	9,911	99.423
GATK HC	Filtered	2	1,696,622	20,098	98.829
GATK HC	Filtered	4	1,625,493	91,227	94.686
GATK HC	Filtered	6	1,454,488	262,232	84.725
GATK HC	Filtered	8	1,199,534	517,186	69.874
GATK HC	Filtered	10	922,396	794,324	53.730
GATK HC	Filtered	12	671,639	1,045,081	39.123
GATK HC	Filtered	14	469,676	1,247,044	27.359
GATK HC	Filtered	16	316,055	1,400,665	18.410
GATK HC	Filtered	18	202,795	1,513,925	11.812
GATK HC	Filtered	20	123,462	1,593,258	7.192
GATK UG	Raw	0	1,710,535	6,185	99.640
GATK UG	Raw	2	1,700,394	16,326	99.049
GATK UG	Raw	4	1,628,997	87,723	94.890
GATK UG	Raw	6	1,457,458	259,262	84.898
GATK UG	Raw	8	1,201,866	514,854	70.009
GATK UG	Raw	10	924,084	792,636	53.828
GATK UG	Raw	12	672,719	1,044,001	39.186

GATK UG	Raw	14	470,324	1,246,396	27.397
GATK UG	Raw	16	316,462	1,400,258	18.434
GATK UG	Raw	18	203,040	1,513,680	11.827
GATK UG	Raw	20	123,628	1,593,092	7.201
GATK UG	Filtered	0	1,707,817	8,903	99.481
GATK UG	Filtered	2	1,697,822	18,898	98.899
GATK UG	Filtered	4	1,626,644	90,076	94.753
GATK UG	Filtered	6	1,455,494	261,226	84.783
GATK UG	Filtered	8	1,200,413	516,307	69.925
GATK UG	Filtered	10	923,232	793,488	53.779
GATK UG	Filtered	12	672,280	1,044,440	39.161
GATK UG	Filtered	14	470,086	1,246,634	27.383
GATK UG	Filtered	16	316,304	1,400,416	18.425
GATK UG	Filtered	18	202,935	1,513,785	11.821
GATK UG	Filtered	20	123,541	1,593,179	7.196
SAMtools	Raw	0	1,716,385	335	99.980
SAMtools	Raw	2	1,701,837	14,883	99.133
SAMtools	Raw	4	1,629,904	86,816	94.943
SAMtools	Raw	6	1,458,027	258,693	84.931
SAMtools	Raw	8	1,202,210	514,510	70.029
SAMtools	Raw	10	924,265	792,455	53.839
SAMtools	Raw	12	672,825	1,043,895	39.192
SAMtools	Raw	14	470,374	1,246,346	27.400
SAMtools	Raw	16	316,483	1,400,237	18.435
SAMtools	Raw	18	203,038	1,513,682	11.827
SAMtools	Raw	20	123,612	1,593,108	7.200
SAMtools	Filtered	0	1,703,114	13,606	99.207
SAMtools	Filtered	2	1,693,281	23,439	98.635
SAMtools	Filtered	4	1,622,273	94,447	94.498
SAMtools	Filtered	6	1,451,191	265,529	84.533
SAMtools	Filtered	8	1,196,001	520,719	69.668

SAMtools	Filtered	10	918,599	798,121	53.509
SAMtools	Filtered	12	667,564	1,049,156	38.886
SAMtools	Filtered	14	465,388	1,251,332	27.109
SAMtools	Filtered	16	311,959	1,404,761	18.172
SAMtools	Filtered	18	199,040	1,517,680	11.594
SAMtools	Filtered	20	120,764	1,595,956	7.035
VarScan	Raw	0	1,715,452	1,268	99.926
VarScan	Raw	2	1,701,004	15,716	99.085
VarScan	Raw	4	1,629,313	87,407	94.908
VarScan	Raw	6	1,457,667	259,053	84.910
VarScan	Raw	8	1,202,002	514,718	70.017
VarScan	Raw	10	924,160	792,560	53.833
VarScan	Raw	12	672,760	1,043,960	39.189
VarScan	Raw	14	470,339	1,246,381	27.398
VarScan	Raw	16	316,460	1,400,260	18.434
VarScan	Raw	18	203,018	1,513,702	11.826
VarScan	Raw	20	123,595	1,593,125	7.199
VarScan	Filtered	0	1,678,687	38,033	97.785
VarScan	Filtered	2	1,675,183	41,537	97.580
VarScan	Filtered	4	1,624,844	91,876	94.648
VarScan	Filtered	6	1,456,627	260,093	84.849
VarScan	Filtered	8	1,201,446	515,274	69.985
VarScan	Filtered	10	923,780	792,940	53.811
VarScan	Filtered	12	672,508	1,044,212	39.174
VarScan	Filtered	14	470,180	1,246,540	27.388
VarScan	Filtered	16	316,362	1,400,358	18.428
VarScan	Filtered	18	202,987	1,513,733	11.824
VarScan	Filtered	20	123,596	1,593,124	7.200

Table S4. Percent concordance of all genotypes (homozygous and heterozygous) called by raw and filtered pipelines using five different variant callers (FreeBayes, GATK HC, GATK UG, SAMtools and VarScan) compared against genotypes obtained using the CanineHD BeadChip array.

Minimum Coverage	FreeBayes Raw	FreeBayes Filtered	GATK HC Raw	GATK HC Filtered	GATK UG Raw	GATK UG Filtered	SAMtools Raw	SAMtools Filtered	VarScan Raw	VarScan Filtered
0	97.316	89.937	96.808	97.463	97.876	97.478	98.170	97.291	98.374	93.117
2	97.670	90.236	97.205	97.586	97.996	97.598	98.376	97.423	98.569	93.152
4	98.301	91.034	97.949	98.119	98.360	98.111	98.763	97.916	98.927	93.354
6	98.920	92.611	98.699	98.691	98.903	98.630	99.068	98.625	99.192	94.728
8	99.262	94.447	99.180	99.078	99.323	99.002	99.263	99.130	99.368	95.730
10	99.378	96.055	99.411	99.318	99.514	99.309	99.418	99.388	99.504	97.244
12	99.481	97.288	99.521	99.483	99.601	99.500	99.533	99.518	99.602	98.246
14	99.580	98.114	99.582	99.593	99.656	99.601	99.601	99.603	99.657	98.862
16	99.578	98.601	99.623	99.653	99.679	99.638	99.637	99.644	99.677	99.124
18	99.613	98.948	99.619	99.658	99.670	99.629	99.633	99.666	99.672	99.288
20	99.579	99.133	99.587	99.627	99.637	99.599	99.604	99.668	99.638	99.321

Table S5. Percentage concordance of homozygous genotypes called by raw and filtered pipelines using five different variant callers (FreeBayes, GATK HC, GATK UG, SAMtools and VarScan) compared against genotypes obtained using the CanineHD BeadChip array.

Minimum Coverage	FreeBayes Raw	FreeBayes Filtered	GATK HC Raw	GATK HC Filtered	GATK UG Raw	GATK UG Filtered	SAMtools Raw	SAMtools Filtered	VarScan Raw	VarScan Filtered
0	99.460	98.143	99.210	99.573	99.714	99.645	99.487	99.692	99.562	99.855
2	99.741	98.412	99.533	99.625	99.730	99.692	99.593	99.740	99.655	99.882
4	99.746	98.798	99.653	99.657	99.735	99.702	99.643	99.771	99.683	99.881
6	99.746	99.274	99.672	99.662	99.731	99.695	99.659	99.773	99.690	99.876
8	99.754	99.563	99.679	99.672	99.732	99.698	99.680	99.778	99.709	99.873
10	99.770	99.711	99.689	99.686	99.738	99.710	99.708	99.792	99.734	99.864
12	99.799	99.787	99.713	99.718	99.761	99.740	99.744	99.818	99.772	99.854
14	99.820	99.823	99.741	99.765	99.794	99.782	99.775	99.850	99.799	99.843
16	99.817	99.824	99.778	99.811	99.814	99.804	99.796	99.864	99.810	99.836
18	99.803	99.810	99.769	99.801	99.798	99.785	99.785	99.863	99.799	99.823
20	99.766	99.774	99.729	99.761	99.757	99.749	99.751	99.846	99.760	99.789

Table S6. Percentage concordance of heterozygous genotypes called by raw and filtered pipelines using five different variant callers (FreeBayes, GATK HC, GATK UG, SAMtools and VarScan) compared against genotypes obtained using the CanineHD BeadChip array.

Minimum Coverage	FreeBayes Raw	FreeBayes Filtered	GATK HC Raw	GATK HC Filtered	GATK UG Raw	GATK UG Filtered	SAMtools Raw	SAMtools Filtered	VarScan Raw	VarScan Filtered
0	89.067	57.225	87.566	89.314	90.812	89.102	93.107	88.067	93.806	66.692
2	89.707	57.656	88.256	89.712	91.333	89.504	93.698	88.524	94.396	66.764
4	92.739	59.989	91.388	92.172	93.070	91.951	95.377	90.780	96.015	68.099
6	95.702	65.586	94.914	94.894	95.680	94.461	96.766	94.160	97.250	74.677
8	97.307	73.228	97.198	96.708	97.700	96.224	97.606	96.557	98.013	79.264
10	97.776	80.507	98.275	97.810	98.601	97.669	98.235	97.738	98.563	86.542
12	98.145	86.408	98.712	98.493	98.928	98.488	98.643	98.255	98.885	91.477
14	98.538	90.527	98.893	98.850	99.056	98.818	98.851	98.535	99.043	94.623
16	98.526	93.113	98.943	98.960	99.084	98.906	98.937	98.675	99.093	95.987
18	98.762	95.050	98.943	99.015	99.094	98.927	98.953	98.784	99.105	96.894
20	98.733	96.220	98.943	99.021	99.096	98.924	98.944	98.862	99.086	97.208

Appendix II: Supplementary material for chapter 3

Table S1. Sample and pedigree information for the parent-offspring trios used in the study

Identification Number	Breed	Trio identification number (same ... Father, Mother or child in trio	Sequencing centre	Sequencing platform	Library	Read length
USCF1292	Cattle Dog	USCF1294	Father	Ramaciotti	Illumina HiSeq 2000	PCR-free TruSeq 550bp insert	100
USCF1293	Cattle Dog	USCF1294	Mother	Ramaciotti	Illumina HiSeq 2000	PCR-free TruSeq 550bp insert	101
USCF1294	Cattle Dog	USCF1294	Child	Ramaciotti	Illumina HiSeq 2000	PCR-free TruSeq 550bp insert	101
USCF1225	Labrador	USCF1119; USCF1014	Father	Ramaciotti	Illumina HiSeq 2000	PCR-free TruSeq 550bp insert	101
USCF1224	Labrador	USCF1119	Mother	Ramaciotti	Illumina HiSeq 2000	PCR-free TruSeq 550bp insert	100
USCF1222	Labrador	USCF1014	Mother	Ramaciotti	Illumina HiSeq 2000	PCR-free TruSeq 550bp insert	100

USCF1119	Labrador	USCF1119	Child	Ramaciotti	Illumina HiSeq 2000	PCR-free TruSeq 550bp insert	101
USCF1014	Labrador	USCF1014	Child	Ramaciotti	Illumina HiSeq 2000	PCR-free TruSeq 550bp insert	101
USCF138	Miniature Schnauzer	USCF136; USCF134	Father	Ramaciotti	Illumina HiSeq 2000	PCR-free TruSeq 350bp insert	100
USCF301	Miniature Schnauzer	USCF136; USCF134	Mother	Ramaciotti	Illumina HiSeq 2000	PCR-free TruSeq 350bp insert	100
USCF136	Miniature Schnauzer	USCF136	Child	Ramaciotti	Illumina HiSeq 2000	PCR-free TruSeq 350bp insert	101
USCF134	Miniature Schnauzer	USCF134	Child	Ramaciotti	Illumina HiSeq 2000	PCR-free TruSeq 350bp insert	101

Table S2. Average coverage in parent-offspring trio samples

Child identifier	Father identifier	Mother identifier	Raw average coverage (Child)	Raw average coverage (Father)	Raw average coverage (Mother)	Raw average coverage in trio
USCF1294	USCF1292	USCF1293	14.5	12.5	13.3	13.4
USCF1119	USCF1225	USCF1224	7.5	11.0	6.6	8.4
USCF1014	USCF1225	USCF1222	11.7	11.0	17.9	13.5
USCF136	USCF138	USCF301	7.2	10.2	15.9	11.1
USCF134	USCF138	USCF301	8.7	10.2	15.9	11.6

Table S3. Observed sites per parent-offspring trio, per chromosome

Chromosome	Chromosome size in CanFam 3.1	Offspring identifier in trio observed	Sites considered	*De novo* sites observed (total)	Transitions (AG)	Transition (CT)	Transversion (AC)	Transversion (AT)	Transversion (CG)	Transversion (GT)
1	122,678,785	USCF134	25,430,142	0	0	0	0	0	0	0
2	85,426,708	USCF134	16,165,018	0	0	0	0	0	0	0
3	91,889,043	USCF134	20,019,354	0	0	0	0	0	0	0
4	88,276,631	USCF134	19,528,607	1	1	0	0	0	0	0
5	88,915,250	USCF134	14,645,917	0	0	0	0	0	0	0
6	77,573,801	USCF134	14,642,397	0	0	0	0	0	0	0
7	80,974,532	USCF134	17,849,351	0	0	0	0	0	0	0
8	74,330,416	USCF134	16,469,421	3	0	1	2	0	0	0
9	61,074,082	USCF134	9,013,245	0	0	0	0	0	0	0
10	69,331,447	USCF134	13,681,867	0	0	0	0	0	0	0
11	74,389,097	USCF134	16,093,419	1	0	1	0	0	0	0
12	72,498,081	USCF134	17,542,020	0	0	0	0	0	0	0
13	63,241,923	USCF134	14,309,623	1	1	0	0	0	0	0
14	60,966,679	USCF134	15,416,789	1	0	0	0	1	0	0
15	64,190,966	USCF134	14,592,372	1	0	0	1	0	0	0
16	59,632,846	USCF134	12,063,722	0	0	0	0	0	0	0
17	64,289,059	USCF134	13,030,745	2	0	0	0	0	0	2

18	55,844,845	USCF134	10,017,989	2	0	2	0	0	0	0
19	53,741,614	USCF134	12,722,779	0	0	0	0	0	0	0
20	58,134,056	USCF134	10,101,522	0	0	0	0	0	0	0
21	50,858,623	USCF134	10,929,873	0	0	0	0	0	0	0
22	61,439,934	USCF134	15,367,588	1	1	0	0	0	0	0
23	52,294,480	USCF134	12,190,912	1	1	0	0	0	0	0
24	47,698,779	USCF134	7,772,509	0	0	0	0	0	0	0
25	51,628,933	USCF134	10,939,621	0	0	0	0	0	0	0
26	38,964,690	USCF134	5,051,270	1	0	0	1	0	0	0
27	45,876,710	USCF134	10,410,574	1	0	1	0	0	0	0
28	41,182,112	USCF134	7,502,532	0	0	0	0	0	0	0
29	41,845,238	USCF134	10,364,260	0	0	0	0	0	0	0
30	40,214,260	USCF134	8,816,006	0	0	0	0	0	0	0
31	39,895,921	USCF134	8,839,281	0	0	0	0	0	0	0
32	38,810,281	USCF134	10,639,933	0	0	0	0	0	0	0
33	31,377,067	USCF134	7,771,882	2	2	0	0	0	0	0
34	42,124,431	USCF134	9,013,408	0	0	0	0	0	0	0
35	26,524,999	USCF134	5,301,285	1	0	1	0	0	0	0
36	30,810,995	USCF134	7,804,227	1	0	0	0	1	0	0
37	30,902,991	USCF134	7,113,522	0	0	0	0	0	0	0
38	23,914,537	USCF134	4,832,706	0	0	0	0	0	0	0
1	122,678,785	USCF136	13,865,057	0	0	0	0	0	0	0
2	85,426,708	USCF136	8,414,400	0	0	0	0	0	0	0
3	91,889,043	USCF136	11,446,442	0	0	0	0	0	0	0
4	88,276,631	USCF136	11,040,943	0	0	0	0	0	0	0

5	88,915,250	USCF136	7,771,766	0	0	0	0	0	0	0
6	77,573,801	USCF136	7,976,768	1	0	1	0	0	0	0
7	80,974,532	USCF136	9,724,290	0	0	0	0	0	0	0
8	74,330,416	USCF136	9,248,297	1	0	0	0	0	1	0
9	61,074,082	USCF136	4,513,086	0	0	0	0	0	0	0
10	69,331,447	USCF136	7,229,263	0	0	0	0	0	0	0
11	74,389,097	USCF136	9,050,430	1	0	0	0	0	0	1
12	72,498,081	USCF136	10,157,185	0	0	0	0	0	0	0
13	63,241,923	USCF136	8,287,198	0	0	0	0	0	0	0
14	60,966,679	USCF136	8,556,469	1	1	0	0	0	0	0
15	64,190,966	USCF136	8,209,955	0	0	0	0	0	0	0
16	59,632,846	USCF136	6,900,144	0	0	0	0	0	0	0
17	64,289,059	USCF136	7,160,794	0	0	0	0	0	0	0
18	55,844,845	USCF136	5,777,671	1	0	1	0	0	0	0
19	53,741,614	USCF136	7,707,702	0	0	0	0	0	0	0
20	58,134,056	USCF136	5,304,180	0	0	0	0	0	0	0
21	50,858,623	USCF136	6,428,309	0	0	0	0	0	0	0
22	61,439,934	USCF136	9,329,742	0	0	0	0	0	0	0
23	52,294,480	USCF136	6,646,976	0	0	0	0	0	0	0
24	47,698,779	USCF136	4,040,944	0	0	0	0	0	0	0
25	51,628,933	USCF136	6,020,721	0	0	0	0	0	0	0
26	38,964,690	USCF136	2,711,785	0	0	0	0	0	0	0
27	45,876,710	USCF136	5,726,948	1	0	1	0	0	0	0
28	41,182,112	USCF136	3,927,984	0	0	0	0	0	0	0
29	41,845,238	USCF136	6,104,239	1	0	1	0	0	0	0

30	40,214,260	USCF136	4,704,568	0	0	0	0	0	0	0
31	39,895,921	USCF136	5,416,406	2	0	1	0	1	0	0
32	38,810,281	USCF136	6,240,165	0	0	0	0	0	0	0
33	31,377,067	USCF136	4,251,515	1	0	0	0	1	0	0
34	42,124,431	USCF136	5,183,011	0	0	0	0	0	0	0
35	26,524,999	USCF136	2,818,856	0	0	0	0	0	0	0
36	30,810,995	USCF136	4,351,416	0	0	0	0	0	0	0
37	30,902,991	USCF136	3,964,706	0	0	0	0	0	0	0
38	23,914,537	USCF136	2,831,219	0	0	0	0	0	0	0
1	122,678,785	USCF1014	46,730,472	2	2	0	0	0	0	0
2	85,426,708	USCF1014	28,656,218	0	0	0	0	0	0	0
3	91,889,043	USCF1014	37,970,148	0	0	0	0	0	0	0
4	88,276,631	USCF1014	36,191,043	1	1	0	0	0	0	0
5	88,915,250	USCF1014	27,094,983	3	1	2	0	0	0	0
6	77,573,801	USCF1014	27,289,875	2	1	1	0	0	0	0
7	80,974,532	USCF1014	32,077,117	2	0	1	1	0	0	0
8	74,330,416	USCF1014	30,048,531	1	0	1	0	0	0	0
9	61,074,082	USCF1014	15,941,552	0	0	0	0	0	0	0
10	69,331,447	USCF1014	24,742,921	1	1	0	0	0	0	0
11	74,389,097	USCF1014	29,758,111	0	0	0	0	0	0	0
12	72,498,081	USCF1014	32,700,207	1	0	1	0	0	0	0
13	63,241,923	USCF1014	26,974,267	0	0	0	0	0	0	0
14	60,966,679	USCF1014	27,814,599	0	0	0	0	0	0	0
15	64,190,966	USCF1014	26,889,028	1	0	0	0	0	0	1
16	59,632,846	USCF1014	23,169,967	1	0	1	0	0	0	0

17	64,289,059	USCF1014	23,627,586	1	0	1	0	0	0	0
18	55,844,845	USCF1014	19,467,098	0	0	0	0	0	0	0
19	53,741,614	USCF1014	24,765,288	1	1	0	0	0	0	0
20	58,134,056	USCF1014	18,276,715	1	0	1	0	0	0	0
21	50,858,623	USCF1014	20,571,130	0	0	0	0	0	0	0
22	61,439,934	USCF1014	29,814,498	0	0	0	0	0	0	0
23	52,294,480	USCF1014	21,958,424	2	1	1	0	0	0	0
24	47,698,779	USCF1014	14,257,328	1	0	0	0	1	0	0
25	51,628,933	USCF1014	19,940,398	0	0	0	0	0	0	0
26	38,964,690	USCF1014	9,902,023	0	0	0	0	0	0	0
27	45,876,710	USCF1014	18,822,187	0	0	0	0	0	0	0
28	41,182,112	USCF1014	13,678,576	0	0	0	0	0	0	0
29	41,845,238	USCF1014	19,747,443	1	0	0	0	0	0	1
30	40,214,260	USCF1014	15,467,753	0	0	0	0	0	0	0
31	39,895,921	USCF1014	17,438,733	0	0	0	0	0	0	0
32	38,810,281	USCF1014	19,629,246	0	0	0	0	0	0	0
33	31,377,067	USCF1014	14,102,693	0	0	0	0	0	0	0
34	42,124,431	USCF1014	17,378,887	1	1	0	0	0	0	0
35	26,524,999	USCF1014	9,633,840	2	0	0	0	1	0	1
36	30,810,995	USCF1014	14,524,417	2	1	0	0	0	0	1
37	30,902,991	USCF1014	13,076,859	2	0	2	0	0	0	0
38	23,914,537	USCF1014	9,369,603	0	0	0	0	0	0	0
1	122,678,785	USCF1119	3,305,006	0	0	0	0	0	0	0
2	85,426,708	USCF1119	1,882,224	0	0	0	0	0	0	0
3	91,889,043	USCF1119	2,908,986	0	0	0	0	0	0	0

4	88,276,631	USCF1119	2,763,173	0	0	0	0	0	0	0
5	88,915,250	USCF1119	1,613,746	0	0	0	0	0	0	0
6	77,573,801	USCF1119	1,903,218	0	0	0	0	0	0	0
7	80,974,532	USCF1119	2,340,767	0	0	0	0	0	0	0
8	74,330,416	USCF1119	2,305,060	0	0	0	0	0	0	0
9	61,074,082	USCF1119	925,742	0	0	0	0	0	0	0
10	69,331,447	USCF1119	1,698,459	0	0	0	0	0	0	0
11	74,389,097	USCF1119	2,225,884	0	0	0	0	0	0	0
12	72,498,081	USCF1119	2,730,763	0	0	0	0	0	0	0
13	63,241,923	USCF1119	2,165,272	0	0	0	0	0	0	0
14	60,966,679	USCF1119	2,264,187	0	0	0	0	0	0	0
15	64,190,966	USCF1119	2,105,469	0	0	0	0	0	0	0
16	59,632,846	USCF1119	1,720,829	1	1	0	0	0	0	0
17	64,289,059	USCF1119	1,636,973	0	0	0	0	0	0	0
18	55,844,845	USCF1119	1,401,134	0	0	0	0	0	0	0
19	53,741,614	USCF1119	2,054,344	1	1	0	0	0	0	0
20	58,134,056	USCF1119	1,179,438	0	0	0	0	0	0	0
21	50,858,623	USCF1119	1,572,824	0	0	0	0	0	0	0
22	61,439,934	USCF1119	2,654,496	1	0	1	0	0	0	0
23	52,294,480	USCF1119	1,663,253	0	0	0	0	0	0	0
24	47,698,779	USCF1119	834,802	0	0	0	0	0	0	0
25	51,628,933	USCF1119	1,479,466	0	0	0	0	0	0	0
26	38,964,690	USCF1119	508,932	0	0	0	0	0	0	0
27	45,876,710	USCF1119	1,448,255	0	0	0	0	0	0	0
28	41,182,112	USCF1119	867,561	0	0	0	0	0	0	0

29	41,845,238	USCF1119	1,663,548	0	0	0	0	0	0	0
30	40,214,260	USCF1119	1,134,303	0	0	0	0	0	0	0
31	39,895,921	USCF1119	1,566,164	0	0	0	0	0	0	0
32	38,810,281	USCF1119	1,765,144	0	0	0	0	0	0	0
33	31,377,067	USCF1119	1,152,217	0	0	0	0	0	0	0
34	42,124,431	USCF1119	1,332,050	0	0	0	0	0	0	0
35	26,524,999	USCF1119	631,670	0	0	0	0	0	0	0
36	30,810,995	USCF1119	1,189,373	0	0	0	0	0	0	0
37	30,902,991	USCF1119	1,035,278	0	0	0	0	0	0	0
38	23,914,537	USCF1119	767,365	0	0	0	0	0	0	0
1	122,678,785	USCF1294	55,273,848	1	1	0	0	0	0	0
2	85,426,708	USCF1294	33,962,454	0	0	0	0	0	0	0
3	91,889,043	USCF1294	44,808,613	0	0	0	0	0	0	0
4	88,276,631	USCF1294	42,405,666	3	1	1	0	0	0	1
5	88,915,250	USCF1294	31,540,124	2	0	1	0	1	0	0
6	77,573,801	USCF1294	31,510,673	2	0	1	1	0	0	0
7	80,974,532	USCF1294	37,673,467	1	0	1	0	0	0	0
8	74,330,416	USCF1294	35,374,163	1	1	0	0	0	0	0
9	61,074,082	USCF1294	18,425,316	4	2	1	1	0	0	0
10	69,331,447	USCF1294	29,111,937	0	0	0	0	0	0	0
11	74,389,097	USCF1294	34,929,857	0	0	0	0	0	0	0
12	72,498,081	USCF1294	38,591,544	2	2	0	0	0	0	0
13	63,241,923	USCF1294	31,436,333	0	0	0	0	0	0	0
14	60,966,679	USCF1294	32,686,722	2	0	2	0	0	0	0
15	64,190,966	USCF1294	31,934,038	2	1	1	0	0	0	0

16	59,632,846	USCF1294	27,445,376	0	0	0	0	0	0	0
17	64,289,059	USCF1294	27,828,461	0	0	0	0	0	0	0
18	55,844,845	USCF1294	22,735,961	1	0	0	1	0	0	0
19	53,741,614	USCF1294	29,306,186	1	0	0	0	1	0	0
20	58,134,056	USCF1294	21,537,088	3	1	2	0	0	0	0
21	50,858,623	USCF1294	24,198,736	0	0	0	0	0	0	0
22	61,439,934	USCF1294	34,847,974	1	0	0	1	0	0	0
23	52,294,480	USCF1294	26,112,421	0	0	0	0	0	0	0
24	47,698,779	USCF1294	16,567,521	0	0	0	0	0	0	0
25	51,628,933	USCF1294	23,480,203	1	0	0	0	1	0	0
26	38,964,690	USCF1294	11,670,827	0	0	0	0	0	0	0
27	45,876,710	USCF1294	22,444,721	0	0	0	0	0	0	0
28	41,182,112	USCF1294	15,860,271	0	0	0	0	0	0	0
29	41,845,238	USCF1294	23,393,326	1	0	1	0	0	0	0
30	40,214,260	USCF1294	18,288,054	0	0	0	0	0	0	0
31	39,895,921	USCF1294	20,339,076	3	1	1	0	0	0	1
32	38,810,281	USCF1294	23,068,897	0	0	0	0	0	0	0
33	31,377,067	USCF1294	16,566,263	1	1	0	0	0	0	0
34	42,124,431	USCF1294	20,477,674	2	1	1	0	0	0	0
35	26,524,999	USCF1294	11,616,663	0	0	0	0	0	0	0
36	30,810,995	USCF1294	17,128,788	1	1	0	0	0	0	0
37	30,902,991	USCF1294	15,402,868	0	0	0	0	0	0	0
38	23,914,537	USCF1294	10,884,299	0	0	0	0	0	0	0

Table S4. Physical position on CanFam 3.1 and genotypes for *de novo* mutations observed in parent-offspring trios

Chromosome	Position	Sire genotype	Dam genotype	Offspring genotype	Offspring identifier in Trio
1	3,779,247	G G	G G	A G	USCF1294
1	121,739,924	G G	G G	A G	USCF1014
1	121,739,926	A A	A A	G A	USCF1014
4	8,941,678	C C	C C	T C	USCF1294
4	20,422,856	G G	G G	T G	USCF1294
4	24,592,772	G G	G G	A G	USCF134
4	49,638,228	G G	G G	A G	USCF1294
4	70,281,398	G G	G G	A G	USCF1014
5	21,562,488	T T	T T	C T	USCF1014
5	22,537,046	T T	T T	A T	USCF1294
5	22,998,572	C C	C C	T C	USCF1014
5	46,644,004	A A	A A	G A	USCF1014
5	85,195,818	T T	T T	C T	USCF1294
6	9,814,311	A A	A A	G A	USCF1014
6	28,518,209	C C	C C	T C	USCF136
6	32,980,209	C C	C C	T C	USCF1014
6	40,729,015	C C	C C	T C	USCF1294
6	45,422,851	A A	A A	C A	USCF1294

7	8,365,994	T T	T T	C T	USCF1294
7	16,247,159	C C	C C	T C	USCF1014
7	33,746,227	A A	A A	C A	USCF1014
8	20,425,502	G G	G G	C G	USCF136
8	22,350,581	C C	C C	T C	USCF1014
8	55,853,926	A A	A A	G A	USCF1294
8	73,607,623	C C	C C	A C	USCF134
8	73,719,433	C C	C C	T C	USCF134
8	73,765,592	A A	A A	C A	USCF134
9	7,983,435	C C	C C	A C	USCF1294
9	9,424,569	G G	G G	A G	USCF1294
9	26,817,438	G G	G G	A G	USCF1294
9	28,733,303	T T	T T	C T	USCF1294
10	37,953,116	A A	A A	G A	USCF1014
11	33,950,398	T T	T T	C T	USCF134
11	35,963,522	G G	G G	T G	USCF136
12	26,613,923	A A	A A	G A	USCF1294
12	44,019,294	A A	A A	G A	USCF1294
12	55,060,734	T T	T T	C T	USCF1014
13	21,462,229	G G	G G	A G	USCF134
14	25,558,020	G G	G G	A G	USCF136
14	26,858,614	A A	A A	T A	USCF134
14	58,324,036	C C	C C	T C	USCF1294
14	59,269,467	T T	T T	C T	USCF1294

15	26,825,863	C C	C C	A C	USCF134
15	32,458,744	T T	T T	C T	USCF1294
15	51,809,687	A A	A A	G A	USCF1294
15	51,904,204	G G	G G	T G	USCF1014
16	25,737,306	A A	A A	G A	USCF1119
16	54,487,862	C C	C C	T C	USCF1014
17	9,819,454	T T	T T	G T	USCF134
17	32,261,540	T T	T T	C T	USCF1014
17	41,353,018	T T	T T	G T	USCF134
18	28,418,247	T T	T T	C T	USCF134
18	28,418,247	T T	T T	C T	USCF136
18	29,365,717	C C	C C	T C	USCF134
18	55,700,372	A A	A A	C A	USCF1294
19	16,343,711	T T	T T	A T	USCF1294
19	20,163,470	G G	G G	A G	USCF1014
19	37,757,686	G G	G G	A G	USCF1119
20	3,790,590	T T	T T	C T	USCF1294
20	5,526,459	T T	T T	C T	USCF1294
20	41,440,801	C C	C C	T C	USCF1014
20	57,437,942	G G	G G	A G	USCF1294
22	13,103,933	G G	G G	A G	USCF134
22	37,991,629	C C	C C	A C	USCF1294
22	41,906,946	T T	T T	C T	USCF1119
23	6,329,495	T T	T T	C T	USCF1014

23	9,749,895	A A	A A	G A	USCF134
23	49,416,247	G G	G G	A G	USCF1014
24	843,131	A A	A A	T A	USCF1014
25	32,245,185	A A	A A	T A	USCF1294
26	36,568,797	A A	A A	C A	USCF134
27	14,937,023	C C	C C	T C	USCF134
27	17,145,045	T T	T T	C T	USCF136
29	1,948,843	C C	C C	T C	USCF1294
29	6,803,696	T T	T T	G T	USCF1014
29	35,609,029	C C	C C	T C	USCF136
31	8,781,152	A A	A A	T A	USCF136
31	12,579,393	G G	G G	A G	USCF1294
31	13,978,308	C C	C C	T C	USCF136
31	16,302,436	T T	T T	C T	USCF1294
31	20,782,899	T T	T T	G T	USCF1294
33	6,036,956	G G	G G	A G	USCF1294
33	8,560,784	A A	A A	T A	USCF136
33	10,917,719	A A	A A	G A	USCF134
33	10,917,725	G G	G G	A G	USCF134
34	9,822,177	G G	G G	A G	USCF1294
34	37,099,031	T T	T T	C T	USCF1294
34	39,504,460	A A	A A	G A	USCF1014
35	4,004,809	A A	A A	T A	USCF1014
35	5,316,122	T T	T T	G T	USCF1014

35	10,820,922	C C	C C	T C	USCF134
36	4,491,107	T T	T T	A T	USCF134
36	17,100,470	G G	G G	A G	USCF1014
36	28,660,864	G G	G G	T G	USCF1014
36	30,792,870	G G	G G	A G	USCF1294
37	14,223,810	T T	T T	C T	USCF1014
37	20,717,756	T T	T T	C T	USCF1014

Table S5. Number of sites passing quality filters for each parent-offspring trio in seven genomic features observed including coding sequence (cds), CpG island (cpg), intergenic, intronic, conserved (phastCons33), 3' untranslated region (utr3) and 5' untranslated region (utr5)

Chromosome	Feature	Non-*de novo* observed in USCF134 trio	*De novo* observed in USCF134	Non-*de novo* observed in USCF136 trio	*De novo* observed in USCF136	Non-*de novo* observed in USCF1014 trio	*De novo* observed in USCF1014	Non-*de novo* observed in USCF1119 trio	*De novo* observed in USCF1119	Non-*de novo* observed in USCF1294 trio	*De novo* observed in USCF1294	Total non-*de novo* observed	Total *de novo* observed
1	cds	197,848	0	100,598	0	455,176	0	23,760	0	503,783	0	1,281,165	0
1	cpg	5,122	0	972	0	190,546	2	3,862	0	137,304	0	337,806	2
1	intergenic	11,444,649	0	6,531,414	0	21,810,056	0	1,612,715	0	25,754,093	0	67,152,927	0
1	intron	10,877,576	0	5,730,122	0	19,267,383	0	1,343,739	0	22,799,433	0	60,018,253	0
1	phastCons33	1,131,539	0	560,633	0	2,098,615	0	146,136	0	2,423,006	0	6,359,929	0
1	utr3	198,801	0	100,840	0	366,218	0	25,827	0	414,975	0	1,106,661	0
1	utr5	53,841	0	26,162	0	112,444	0	7,024	0	131,434	0	330,905	0
2	cds	145,788	0	62,765	0	300,131	0	16,132	0	349,901	0	874,717	0
2	cpg	3,603	0	567	0	179,587	0	3,410	0	102,474	0	289,641	0
2	intergenic	6,788,105	0	3,675,063	0	12,385,484	0	796,244	0	14,728,973	1	38,373,869	1
2	intron	7,336,713	0	3,751,349	0	12,681,210	0	864,921	0	14,995,021	0	39,629,214	0
2	phastCons33	850,004	0	388,148	0	1,498,100	0	98,265	0	1,737,769	0	4,572,286	0
2	utr3	151,047	0	68,075	0	253,156	0	17,850	0	283,790	0	773,918	0

2	utr5	46,826	0	20,751	0	87,837	0	5,134	0	105,186	0	265,734	0
3	cds	127,029	0	59,633	0	261,991	0	14,128	0	297,804	0	760,585	0
3	cpg	2,426	0	667	0	105,083	0	2,269	0	65,481	0	175,926	0
3	intergenic	11,189,427	0	6,738,147	0	21,877,611	0	1,748,341	0	25,899,472	0	67,452,998	0
3	intron	7,236,481	0	3,857,468	0	13,075,545	0	943,555	0	15,393,182	0	40,506,231	0
3	phastCons33	996,189	0	514,823	0	1,846,133	0	141,841	0	2,149,834	0	5,648,820	0
3	utr3	124,712	0	58,890	0	223,334	0	16,577	0	256,405	0	679,918	0
3	utr5	33,611	0	14,635	0	63,814	0	3,187	0	73,773	0	189,020	0
4	cds	136,704	0	61,552	0	276,120	0	14,305	0	325,558	0	814,239	0
4	cpg	2,690	0	287	0	87,079	0	1,241	0	56,534	0	147,831	0
4	intergenic	9,792,679	0	5,901,086	0	18,807,740	0	1,522,316	0	22,026,302	2	58,050,123	2
4	intron	7,570,701	1	3,980,926	0	13,421,841	1	955,756	0	15,747,985	1	41,677,209	3
4	phastCons33	1,011,362	0	516,511	0	1,848,738	0	138,914	0	2,158,043	0	5,673,568	0
4	utr3	147,689	0	66,596	0	240,524	0	16,077	0	278,425	0	749,311	0
4	utr5	38,853	0	18,512	0	71,664	0	4,420	0	88,539	0	221,988	0
5	cds	125,882	0	57,861	0	301,801	0	12,677	0	321,400	0	819,621	0
5	cpg	4,060	0	222	0	122,450	0	1,962	0	76,587	0	205,281	0
5	intergenic	5,667,754	0	3,216,298	0	10,967,760	2	674,613	0	12,795,729	2	33,322,154	4
5	intron	6,757,305	0	3,423,377	0	12,068,108	1	691,638	0	13,988,847	0	36,929,275	1
5	phastCons33	818,482	0	392,855	0	1,574,054	0	94,426	0	1,777,430	0	4,657,247	0
5	utr3	127,985	0	54,847	0	225,141	0	13,271	0	245,087	0	666,331	0
5	utr5	30,149	0	12,588	0	65,469	0	2,557	0	70,575	0	181,338	0
6	cds	141,102	0	65,967	0	302,112	0	16,933	0	329,856	1	855,970	1
6	cpg	2,703	0	333	0	109,396	0	1,752	0	79,570	0	193,754	0
6	intergenic	5,997,716	0	3,466,145	1	11,592,417	1	836,422	0	13,323,945	0	35,216,645	2
6	intron	6,616,754	0	3,490,542	0	12,084,920	1	831,436	0	13,939,348	1	36,963,000	2
6	phastCons33	866,628	0	445,639	0	1,633,563	0	116,738	0	1,859,393	1	4,921,961	1
6	utr3	132,890	0	57,705	0	220,958	0	14,907	0	246,798	0	673,258	0
6	utr5	42,631	0	17,506	0	80,915	0	4,952	0	93,598	1	239,602	1
7	cds	171,518	0	79,939	0	341,704	0	18,532	0	394,493	0	1,006,186	0

7	cpg	1,824	0	181	0	69,493	0	1,301	0	48,175	0	120,974	0
7	intergenic	7,481,137	0	4,266,470	0	13,865,276	0	1,031,148	0	16,287,797	0	42,931,828	0
7	intron	8,410,292	0	4,463,906	0	14,676,044	2	1,066,290	0	17,280,255	0	45,896,787	2
7	phastCons33	898,134	0	448,314	0	1,638,392	0	115,722	0	1,891,642	0	4,992,204	0
7	utr3	162,820	0	73,703	0	270,602	0	18,316	0	317,107	0	842,548	0
7	utr5	48,174	0	20,383	0	87,436	0	6,254	0	103,373	0	265,620	0
8	cds	144,180	0	65,907	0	288,692	0	16,420	0	334,336	0	849,535	0
8	cpg	3,179	0	712	0	92,810	0	1,490	0	65,359	0	163,550	0
8	intergenic	7,525,206	3	4,502,240	1	14,415,114	1	1,150,146	0	16,922,554	1	44,515,260	6
8	intron	6,850,652	0	3,648,936	0	11,955,113	0	898,543	0	14,087,747	0	37,440,991	0
8	phastCons33	994,126	0	500,787	0	1,803,507	0	139,007	0	2,087,607	0	5,525,034	0
8	utr3	139,504	0	63,206	0	234,954	0	18,860	0	258,005	0	714,529	0
8	utr5	38,098	0	16,053	0	71,232	0	4,721	0	85,587	0	215,691	0
9	cds	123,563	0	51,805	0	288,888	0	10,005	0	303,863	0	778,124	0
9	cpg	3,158	0	197	0	114,339	0	2,011	0	79,177	0	198,882	0
9	intergenic	2,365,854	0	1,245,066	0	4,363,431	0	281,295	0	4,899,877	2	13,155,523	2
9	intron	5,107,705	0	2,531,592	0	8,844,827	0	504,605	0	10,379,862	1	27,368,591	1
9	phastCons33	631,281	0	287,317	0	1,150,265	0	64,420	0	1,275,367	0	3,408,650	0
9	utr3	133,485	0	52,020	0	225,195	0	11,765	0	241,639	0	664,104	0
9	utr5	37,478	0	15,411	0	73,129	0	2,888	0	85,458	0	214,364	0
10	cds	128,488	0	58,026	0	255,914	0	14,044	0	287,700	0	744,172	0
10	cpg	2,273	0	866	0	148,882	0	2,049	0	85,900	0	239,970	0
10	intergenic	6,018,300	0	3,366,069	0	11,365,889	1	810,125	0	13,449,014	0	35,009,397	1
10	intron	5,858,500	0	2,955,658	0	10,081,845	0	680,145	0	11,832,780	0	31,408,928	0
10	phastCons33	805,628	0	388,736	0	1,446,202	0	102,593	0	1,685,531	0	4,428,690	0
10	utr3	132,201	0	56,532	0	208,252	0	13,318	0	242,495	0	652,798	0
10	utr5	35,749	0	14,170	0	65,446	0	3,504	0	76,651	0	195,520	0
11	cds	119,282	0	53,810	0	241,935	0	13,236	0	284,308	0	712,571	0
11	cpg	2,326	0	417	0	84,266	0	1,506	0	57,208	0	145,723	0
11	intergenic	8,171,150	0	4,849,145	0	15,621,971	0	1,222,391	0	18,320,203	0	48,184,860	0

11	intron	5,993,690	1	3,176,116	1	10,671,211	0	751,324	0	12,551,484	0	33,143,825	2
11	phastCons33	903,110	1	452,612	1	1,659,683	0	115,254	0	1,924,991	0	5,055,650	2
11	utr3	126,967	0	57,238	0	208,002	0	15,614	0	238,901	0	646,722	0
11	utr5	32,683	0	14,489	0	59,582	0	3,934	0	71,419	0	182,107	0
12	cds	136,178	0	63,496	0	271,672	1	14,672	0	325,639	0	811,657	1
12	cpg	2,494	0	470	0	60,024	0	1,401	0	51,926	0	116,315	0
12	intergenic	8,181,131	0	4,995,700	0	15,923,292	0	1,347,288	0	18,752,559	1	49,199,970	1
12	intron	7,152,297	0	3,972,398	0	12,763,068	0	1,076,169	0	15,064,486	1	40,028,418	1
12	phastCons33	878,308	0	457,486	0	1,604,856	1	127,798	0	1,885,561	0	4,954,009	1
12	utr3	134,464	0	61,670	0	223,364	0	18,304	0	250,399	0	688,201	0
12	utr5	30,960	0	11,991	0	59,250	0	3,327	0	69,943	0	175,471	0
13	cds	100,035	0	47,753	0	203,981	0	12,705	0	238,049	0	602,523	0
13	cpg	1,417	0	405	0	79,574	0	1,199	0	45,564	0	128,159	0
13	intergenic	7,570,508	0	4,610,487	0	14,696,817	0	1,182,296	0	17,226,474	0	45,286,582	0
13	intron	5,486,050	0	3,005,644	0	9,977,425	0	811,284	0	11,542,150	0	30,822,553	0
13	phastCons33	640,471	1	331,696	0	1,178,224	0	92,809	0	1,360,719	0	3,603,919	1
13	utr3	94,223	0	46,543	0	163,554	0	14,063	0	184,561	0	502,944	0
13	utr5	26,213	0	11,117	0	46,068	0	3,863	0	53,925	0	141,186	0
14	cds	103,787	0	49,897	0	207,563	0	13,389	0	240,521	0	615,157	0
14	cpg	787	0	113	0	58,613	0	1,272	0	41,477	0	102,262	0
14	intergenic	6,397,078	1	3,693,019	1	11,801,137	0	983,933	0	13,934,829	1	36,809,996	3
14	intron	6,733,905	0	3,615,825	0	11,896,289	0	951,691	0	13,970,333	1	37,168,043	1
14	phastCons33	919,772	0	467,423	0	1,609,169	0	127,420	0	1,878,266	0	5,002,050	0
14	utr3	112,618	0	55,529	0	190,582	0	17,362	0	220,309	0	596,400	0
14	utr5	30,313	0	14,236	0	58,390	0	4,414	0	67,223	0	174,576	0
15	cds	136,343	0	63,287	0	268,688	0	18,069	0	311,975	0	798,362	0
15	cpg	1,464	0	302	0	83,854	1	1,577	0	50,086	0	137,283	1
15	intergenic	6,279,628	1	3,642,650	0	11,886,997	1	935,982	0	14,208,755	1	36,954,012	3
15	intron	6,578,733	0	3,600,178	0	11,754,704	0	929,296	0	13,899,632	1	36,762,543	1
15	phastCons33	762,228	0	383,478	0	1,394,823	0	103,875	0	1,606,644	0	4,251,048	0

15	utr3	111,631	0	51,379	0	189,047	0	14,212	0	212,021	0	578,290	0
15	utr5	38,111	0	18,136	0	74,725	0	4,834	0	85,822	0	221,628	0
16	cds	82,795	0	38,208	0	179,080	0	9,982	0	208,089	0	518,154	0
16	cpg	2,306	0	1,644	0	139,129	0	9,629	0	102,493	0	255,201	0
16	intergenic	5,920,776	0	3,553,494	0	11,750,531	1	893,059	0	13,959,944	0	36,077,804	1
16	intron	4,636,724	0	2,556,022	0	8,636,017	0	628,958	0	10,189,555	0	26,647,276	0
16	phastCons33	446,205	0	229,812	0	848,592	0	61,162	0	983,198	0	2,568,969	0
16	utr3	78,920	0	35,563	0	136,253	0	9,388	0	158,366	0	418,490	0
16	utr5	23,539	0	10,923	0	46,795	0	3,385	0	54,261	0	138,903	0
17	cds	130,700	0	58,431	0	249,434	0	13,611	0	286,343	0	738,519	0
17	cpg	4,587	0	1,526	0	134,234	0	2,410	0	80,016	0	222,773	0
17	intergenic	6,361,515	2	3,760,497	0	12,201,272	1	885,981	0	14,323,472	0	37,532,737	3
17	intron	4,976,985	0	2,526,673	0	8,430,026	0	554,739	0	9,961,034	0	26,449,457	0
17	phastCons33	607,390	0	299,221	0	1,108,081	0	74,281	0	1,268,414	0	3,357,387	0
17	utr3	123,551	0	53,160	0	196,645	0	12,403	0	227,887	0	613,646	0
17	utr5	28,759	0	12,543	0	54,163	0	2,709	0	63,009	0	161,183	0
18	cds	80,253	0	38,679	0	197,265	0	9,135	0	217,073	0	542,405	0
18	cpg	2,395	0	603	0	106,505	0	1,922	0	67,205	0	178,630	0
18	intergenic	4,605,710	1	2,814,365	0	9,142,304	0	689,559	0	10,744,776	1	27,996,714	2
18	intron	4,132,886	1	2,289,807	1	7,869,023	0	561,217	0	9,136,594	0	23,989,527	2
18	phastCons33	432,564	0	213,989	0	835,848	0	55,064	0	944,109	0	2,481,574	0
18	utr3	72,385	0	32,224	0	129,239	0	7,569	0	141,404	0	382,821	0
18	utr5	17,901	0	9,038	0	40,656	0	2,032	0	46,964	0	116,591	0
19	cds	67,448	0	31,624	0	138,263	0	9,494	0	163,025	0	409,854	0
19	cpg	689	0	230	0	43,215	0	892	0	34,053	0	79,079	0
19	intergenic	7,446,859	0	4,740,407	0	14,971,762	1	1,264,666	0	17,713,137	1	46,136,831	2
19	intron	4,403,756	0	2,502,353	0	8,188,131	0	664,251	1	9,680,549	0	25,439,040	1
19	phastCons33	651,553	0	343,989	0	1,197,364	0	98,301	0	1,406,716	0	3,697,923	0
19	utr3	65,568	0	32,036	0	111,852	0	9,196	0	126,690	0	345,342	0
19	utr5	14,390	0	7,470	0	27,406	0	2,274	0	33,220	0	84,760	0

20	cds	74,998	0	34,689	0	216,490	0	7,317	0	213,416	0	546,910	0
20	cpg	5,037	0	194	0	137,933	0	1,500	0	88,087	0	232,751	0
20	intergenic	3,210,072	0	1,801,041	0	5,939,636	0	425,461	0	7,009,288	2	18,385,498	2
20	intron	5,439,258	0	2,768,029	0	9,638,693	1	599,580	0	11,322,703	0	29,768,263	1
20	phastCons33	513,560	0	248,302	0	970,889	0	65,091	0	1,097,045	0	2,894,887	0
20	utr3	74,001	0	29,658	0	138,222	0	8,498	0	153,973	0	404,352	0
20	utr5	22,278	0	9,782	0	54,486	0	2,339	0	63,409	0	152,294	0
21	cds	85,205	0	41,572	0	189,619	0	10,456	0	218,169	0	545,021	0
21	cpg	832	0	24	0	36,468	0	826	0	27,740	0	65,890	0
21	intergenic	4,639,674	0	2,885,278	0	9,183,913	0	712,456	0	10,868,425	0	28,289,746	0
21	intron	5,003,056	0	2,809,303	0	8,959,902	0	677,173	0	10,518,677	0	27,968,111	0
21	phastCons33	496,806	0	249,576	0	911,039	0	67,905	0	1,046,565	0	2,771,891	0
21	utr3	76,748	0	35,803	0	131,373	0	9,303	0	152,401	0	405,628	0
21	utr5	19,522	0	8,606	0	37,730	0	2,413	0	45,545	0	113,816	0
22	cds	74,476	0	30,970	0	143,958	0	8,742	0	168,698	0	426,844	0
22	cpg	623	0	118	0	40,988	0	1,006	0	29,707	0	72,442	0
22	intergenic	9,604,922	1	6,177,364	0	19,434,867	0	1,800,167	1	22,630,406	1	59,647,726	3
22	intron	4,435,098	0	2,411,279	0	7,920,224	0	655,257	0	9,327,909	0	24,749,767	0
22	phastCons33	779,164	0	425,916	0	1,426,537	0	126,259	0	1,646,435	0	4,404,311	0
22	utr3	67,646	0	33,549	0	115,331	0	10,337	0	136,581	0	363,444	0
22	utr5	16,380	0	6,477	0	30,239	0	2,192	0	36,487	0	91,775	0
23	cds	98,759	0	44,760	0	196,320	0	12,198	0	226,403	0	578,440	0
23	cpg	1,195	0	425	0	45,932	0	1,072	0	34,294	0	82,918	0
23	intergenic	5,294,224	0	3,028,471	0	9,967,023	0	769,859	0	11,917,179	0	30,976,756	0
23	intron	5,238,814	1	2,771,599	0	9,131,389	1	681,472	0	10,789,657	0	28,612,931	2
23	phastCons33	574,203	0	282,831	0	1,018,793	0	77,066	0	1,192,249	0	3,145,142	0
23	utr3	88,719	0	40,084	0	147,917	0	10,653	0	167,503	0	454,876	0
23	utr5	29,825	0	11,532	0	53,289	0	3,697	0	63,960	0	162,303	0
24	cds	57,917	0	25,491	0	135,712	0	5,714	0	141,261	0	366,095	0
24	cpg	2,575	0	478	0	146,167	0	1,864	0	70,097	0	221,181	0

24	intergenic	3,412,169	0	1,826,564	0	6,344,424	1	372,521	0	7,401,564	0	19,357,242	1
24	intron	3,454,047	0	1,763,376	0	6,233,232	0	368,565	0	7,216,476	0	19,035,696	0
24	phastCons33	419,142	0	198,876	0	788,684	0	45,929	0	893,733	0	2,346,364	0
24	utr3	62,165	0	27,648	0	110,632	0	5,135	0	120,999	0	326,579	0
24	utr5	20,009	0	8,181	0	39,746	0	1,844	0	44,816	0	114,596	0
25	cds	92,005	0	44,841	0	191,702	0	10,784	0	219,454	0	558,786	0
25	cpg	3,468	0	892	0	136,602	0	2,083	0	69,715	0	212,760	0
25	intergenic	4,677,041	0	2,639,526	0	8,625,822	0	640,532	0	10,189,888	0	26,772,809	0
25	intron	4,827,030	0	2,585,563	0	8,618,827	0	641,179	0	10,126,664	1	26,799,263	1
25	phastCons33	396,676	0	200,535	0	730,279	0	54,769	0	852,236	0	2,234,495	0
25	utr3	81,519	0	41,177	0	144,733	0	11,212	0	166,568	0	445,209	0
25	utr5	26,306	0	10,766	0	46,379	0	2,683	0	54,496	0	140,630	0
26	cds	47,405	0	21,450	0	121,114	0	3,751	0	131,858	0	325,578	0
26	cpg	1,691	0	408	0	72,919	0	1,047	0	48,036	0	124,101	0
26	intergenic	1,674,416	0	960,596	0	3,408,188	0	180,304	0	3,993,492	0	10,216,996	0
26	intron	2,756,964	1	1,456,235	0	5,310,029	0	276,365	0	6,275,415	0	16,075,008	1
26	phastCons33	208,746	0	94,332	0	413,615	0	18,989	0	464,646	0	1,200,328	0
26	utr3	47,431	0	17,837	0	89,520	0	4,310	0	97,223	0	256,321	0
26	utr5	17,195	0	6,210	0	34,768	0	1,710	0	39,575	0	99,458	0
27	cds	104,743	0	51,541	0	216,236	0	13,802	0	245,319	0	631,641	0
27	cpg	884	0	40	0	36,192	0	641	0	27,451	0	65,208	0
27	intergenic	3,594,595	1	2,057,998	1	6,683,512	0	516,881	0	8,058,533	0	20,911,519	2
27	intron	5,082,339	0	2,713,616	0	8,900,953	0	692,652	0	10,538,466	0	27,928,026	0
27	phastCons33	489,054	0	241,677	0	887,966	0	67,109	0	1,028,208	0	2,714,014	0
27	utr3	96,524	0	48,253	0	165,878	0	12,364	0	189,032	0	512,051	0
27	utr5	27,694	0	12,899	0	53,530	0	2,839	0	64,039	0	161,001	0
28	cds	88,603	0	37,919	0	171,640	0	9,380	0	193,164	0	500,706	0
28	cpg	2,633	0	1,107	0	156,933	0	3,106	0	74,504	0	238,283	0
28	intergenic	2,926,808	0	1,604,399	0	5,504,230	0	339,657	0	6,398,155	0	16,773,249	0
28	intron	3,543,947	0	1,799,613	0	6,240,420	0	409,251	0	7,244,241	0	19,237,472	0

28	phastCons33	462,166	0	217,212	0	832,012	0	55,776	0	958,354	0	2,525,520	0
28	utr3	93,310	0	39,760	0	147,534	0	10,945	0	173,653	0	465,202	0
28	utr5	22,007	0	10,819	0	40,129	0	2,846	0	48,658	0	124,459	0
29	cds	66,074	0	32,083	0	127,957	0	9,454	0	155,707	0	391,275	0
29	cpg	1,210	0	221	0	40,503	0	870	0	33,559	0	76,363	0
29	intergenic	5,151,337	0	3,154,182	1	10,069,205	0	859,800	0	11,894,535	1	31,129,059	2
29	intron	3,933,683	0	2,231,433	0	7,238,504	1	599,880	0	8,628,838	0	22,632,338	1
29	phastCons33	534,676	0	281,618	0	977,597	0	83,287	0	1,146,770	0	3,023,948	0
29	utr3	81,632	0	40,541	0	132,680	0	12,714	0	153,601	0	421,168	0
29	utr5	23,743	0	11,671	0	42,400	0	3,038	0	51,781	0	132,633	0
30	cds	111,046	0	54,091	0	227,513	0	12,685	0	263,305	0	668,640	0
30	cpg	1,422	0	92	0	53,643	0	802	0	31,463	0	87,422	0
30	intergenic	3,115,653	0	1,694,752	0	5,512,469	0	400,129	0	6,558,665	0	17,281,668	0
30	intron	4,535,122	0	2,420,275	0	7,921,179	0	594,846	0	9,342,342	0	24,813,764	0
30	phastCons33	572,215	0	281,053	0	1,021,863	0	75,130	0	1,174,888	0	3,125,149	0
30	utr3	113,935	0	52,884	0	184,783	0	14,947	0	208,505	0	575,054	0
30	utr5	23,947	0	10,265	0	43,026	0	2,421	0	51,617	0	131,276	0
31	cds	38,379	0	17,718	0	86,626	0	5,058	0	94,459	0	242,240	0
31	cpg	2,152	0	483	0	120,525	0	2,009	0	49,964	0	175,133	0
31	intergenic	5,029,894	0	3,200,386	2	10,071,602	0	931,905	0	11,856,994	2	31,090,781	4
31	intron	2,590,491	0	1,461,203	0	4,916,019	0	407,140	0	5,690,242	1	15,065,095	1
31	phastCons33	363,633	0	200,251	0	678,816	0	60,891	0	786,909	0	2,090,500	0
31	utr3	40,740	0	21,429	0	75,757	0	6,793	0	82,207	0	226,926	0
31	utr5	12,046	0	4,502	0	23,809	0	1,726	0	26,551	0	68,634	0
32	cds	81,143	0	37,527	0	152,978	0	10,473	0	183,567	0	465,688	0
32	cpg	525	0	81	0	13,828	0	159	0	15,005	0	29,598	0
32	intergenic	4,669,845	0	2,854,235	0	8,914,212	0	798,405	0	10,443,948	0	27,680,645	0
32	intron	4,644,106	0	2,607,193	0	8,214,541	0	743,762	0	9,684,619	0	25,894,221	0
32	phastCons33	430,284	0	225,167	0	783,583	0	68,751	0	924,277	0	2,432,062	0
32	utr3	75,514	0	38,604	0	129,427	0	10,810	0	151,053	0	405,408	0

32	utr5	17,978	0	8,381	0	30,714	0	2,652	0	37,634	0	97,359	0
33	cds	66,721	0	32,105	0	133,724	0	7,746	0	155,705	0	396,001	0
33	cpg	994	0	83	0	38,780	0	572	0	25,316	0	65,745	0
33	intergenic	3,357,287	0	1,915,663	0	6,372,765	0	542,551	0	7,454,645	1	19,642,911	1
33	intron	3,432,577	2	1,841,226	1	5,991,840	0	481,952	0	7,096,521	0	18,844,116	3
33	phastCons33	398,232	0	197,674	0	712,990	0	54,912	0	828,252	0	2,192,060	0
33	utr3	74,160	0	32,782	0	118,774	0	9,204	0	140,063	0	374,983	0
33	utr5	18,993	0	8,868	0	33,438	0	2,729	0	37,178	0	101,206	0
34	cds	59,693	0	26,374	0	116,018	0	6,413	0	131,174	0	339,672	0
34	cpg	736	0	255	0	48,101	0	952	0	37,128	0	87,172	0
34	intergenic	4,577,400	0	2,797,611	0	9,222,780	0	725,903	0	10,874,537	1	28,198,231	1
34	intron	3,597,219	0	1,939,811	0	6,541,589	1	492,889	0	7,722,784	1	20,294,292	2
34	phastCons33	432,829	0	221,168	0	795,476	0	59,965	0	929,798	0	2,439,236	0
34	utr3	64,800	0	31,478	0	114,308	0	9,768	0	130,115	0	350,469	0
34	utr5	19,824	0	8,670	0	35,826	0	2,256	0	41,782	0	108,358	0
35	cds	27,571	0	11,916	0	60,652	0	2,473	0	71,061	0	173,673	0
35	cpg	2,495	0	228	0	44,828	0	936	0	36,256	0	84,743	0
35	intergenic	2,420,964	1	1,358,419	0	4,523,574	1	309,160	0	5,440,479	0	14,052,596	2
35	intron	2,042,159	0	1,027,644	0	3,582,332	0	226,855	0	4,320,255	0	11,199,245	0
35	phastCons33	223,168	0	111,399	0	411,958	0	25,896	0	491,294	0	1,263,715	0
35	utr3	32,955	0	14,538	0	54,689	0	3,308	0	66,658	0	172,148	0
35	utr5	9,816	0	3,933	0	18,424	0	781	0	21,928	0	54,882	0
36	cds	91,359	0	41,096	0	182,396	0	11,922	0	209,672	0	536,445	0
36	cpg	1,722	0	280	0	37,837	0	881	0	27,128	0	67,848	0
36	intergenic	3,188,358	0	1,883,942	0	6,235,004	1	508,086	0	7,396,467	0	19,211,857	1
36	intron	3,538,940	1	1,892,407	0	6,281,275	1	525,559	0	7,374,475	0	19,612,656	2
36	phastCons33	568,641	0	284,881	0	1,029,813	0	82,334	0	1,192,674	0	3,158,343	0
36	utr3	74,169	0	36,548	0	124,806	0	10,959	0	143,307	0	389,789	0
36	utr5	22,015	0	9,811	0	37,906	0	2,902	0	43,601	0	116,235	0
37	cds	70,238	0	32,317	0	134,874	0	8,481	0	157,271	0	403,181	0

37	cpg	499	0	134	0	48,009	0	1,042	0	28,395	0	78,079	0
37	intergenic	3,053,606	0	1,772,628	0	5,852,293	1	451,489	0	6,919,735	0	18,049,751	1
37	intron	3,214,021	0	1,749,093	0	5,746,647	1	472,958	0	6,742,146	0	17,924,865	1
37	phastCons33	450,543	0	227,304	0	814,670	0	61,938	0	946,382	0	2,500,837	0
37	utr3	70,540	0	34,766	0	115,211	0	8,915	0	126,627	0	356,059	0
37	utr5	17,557	0	7,536	0	30,961	0	1,610	0	36,091	0	93,755	0
38	cds	26,608	0	12,143	0	62,849	0	3,701	0	69,040	0	174,341	0
38	cpg	574	0	160	0	38,633	0	546	0	16,559	0	56,472	0
38	intergenic	2,799,463	0	1,737,281	0	5,608,742	0	487,020	0	6,539,296	0	17,171,802	0
38	intron	1,507,363	0	814,638	0	2,757,038	0	209,537	0	3,178,244	0	8,466,820	0
38	phastCons33	210,709	0	106,568	0	407,038	0	32,520	0	467,350	0	1,224,185	0
38	utr3	30,081	0	12,865	0	50,415	0	3,660	0	58,872	0	155,893	0
38	utr5	9,971	0	4,495	0	19,317	0	1,548	0	22,142	0	57,473	0
All	cds	3,761,866	0	1,739,841	0	7,898,788	1	431,779	0	8,972,419	1	22,804,693	2
All	cpg	80,770	0	16,417	0	3,303,900	3	65,069	0	2,096,993	0	5,563,149	3
All	intergenic	211,602,910	11	124,918,098	7	407,000,000	13	31,640,806	1	479,000,000	21	1,254,201,072	53
All	intron	195,533,939	8	104,142,428	3	348,000,000	11	25,466,429	1	410,000,000	9	1,083,211,110	32
All	phastCons33	23,769,421	2	11,919,809	1	43,587,827	1	3,198,543	0	50,372,305	1	132,847,905	5
All	utr3	3,718,050	0	1,707,960	0	6,308,862	0	458,714	0	7,164,205	0	19,357,791	0
All	utr5	1,025,385	0	449,518	0	1,952,538	0	119,639	0	2,291,250	1	5,838,330	1

Appendix III: Supplementary material for chapter 4

Appendix for chapter 4.2

File S1. Sample information including data availability and genotypes at *BBS4* c.58A > T

BREED	ID	GENOTYPING ARRAY	ACCESSION	WHOLE GENOME SEQUENCING PLATFORM	LIBRARY	ACCESSION	BBS4 c.58A > T GENOTYPES (BY SANGER SEQUENCING)
Hungarian Puli	USCF347	Illumina CanineHD BeadChip	GSE87642	Illumina HiSeq 2000	TruSeq, PCR-free	PRJNA344694	A T
Hungarian Puli	USCF350	Illumina CanineHD BeadChip	GSE87642	N/A	N/A	N/A	A T
Hungarian Puli	USCF516	Illumina CanineHD BeadChip	GSE87642	Illumina HiSeq 2000	TruSeq, PCR-free	PRJNA344694	T T
Hungarian Puli	USCF517	Illumina CanineHD BeadChip	GSE87642	N/A	N/A	N/A	A T
Hungarian Puli	USCF518	Illumina CanineHD BeadChip	GSE87642	N/A	N/A	N/A	A T
Hungarian Puli	USCF519	Illumina CanineHD BeadChip	GSE87642	Illumina HiSeq 2000	TruSeq, PCR-free	PRJNA344694	T T
Hungarian Puli	USCF520	Illumina CanineHD BeadChip	GSE87642	N/A	N/A	N/A	A T

Hungarian Puli	USCF521	Illumina CanineHD BeadChip	GSE87642	N/A	N/A	N/A	A T
Hungarian Puli	USCF522	Illumina CanineHD BeadChip	GSE87642	N/A	N/A	N/A	A A
Hungarian Puli	USCF523	Illumina CanineHD BeadChip	GSE87642	N/A	N/A	N/A	A A
Hungarian Puli	USCF524	Illumina CanineHD BeadChip	GSE87642	N/A	N/A	N/A	A T
Hungarian Puli	USCF525	Illumina CanineHD BeadChip	GSE87642	Illumina HiSeq 2000	TruSeq, PCR-free	PRJNA344694	A T
Hungarian Puli	USCF526	Illumina CanineHD BeadChip	GSE87642	N/A	N/A	N/A	A T
Hungarian Puli	USCF532	Illumina CanineHD BeadChip	GSE87642	N/A	N/A	N/A	A T
Hungarian Puli	USCF1194	N/A	N/A	N/A	N/A	N/A	A A
Hungarian Puli	USCF1195	N/A	N/A	N/A	N/A	N/A	A A
Hungarian Puli	USCF1197	N/A	N/A	N/A	N/A	N/A	A A
Hungarian Puli	USCF1198	N/A	N/A	N/A	N/A	N/A	A A
Hungarian Puli	USCF1263	N/A	N/A	N/A	N/A	N/A	A A
Hungarian Puli	USCF1264	N/A	N/A	N/A	N/A	N/A	A A
Hungarian Puli	USCF1265	N/A	N/A	N/A	N/A	N/A	A A

Hungarian Puli	USCF1266	N/A	N/A	N/A	N/A	N/A	A A
Hungarian Puli	USCF1267	N/A	N/A	N/A	N/A	N/A	A A
Hungarian Puli	USCF1268	N/A	N/A	N/A	N/A	N/A	A A
Hungarian Puli	USCF1269	N/A	N/A	N/A	N/A	N/A	A A
Hungarian Puli	USCF1270	N/A	N/A	N/A	N/A	N/A	A T
Hungarian Puli	USCF1271	N/A	N/A	N/A	N/A	N/A	A A
Hungarian Puli	USCF1272	N/A	N/A	N/A	N/A	N/A	A A
Hungarian Puli	USCF1273	N/A	N/A	N/A	N/A	N/A	A T
Hungarian Puli	USCF1274	N/A	N/A	N/A	N/A	N/A	A A
Hungarian Puli	USCF1275	N/A	N/A	N/A	N/A	N/A	A A
Hungarian Puli	USCF1276	N/A	N/A	N/A	N/A	N/A	A T
Hungarian Puli	USCF1277	N/A	N/A	N/A	N/A	N/A	A A
Hungarian Puli	USCF1278	N/A	N/A	N/A	N/A	N/A	A A
Hungarian Puli	USCF1279	N/A	N/A	N/A	N/A	N/A	A T
Hungarian Puli	USCF1280	N/A	N/A	N/A	N/A	N/A	A A
Hungarian Puli	USCF1281	N/A	N/A	N/A	N/A	N/A	A A
Hungarian Puli	USCF1282	N/A	N/A	N/A	N/A	N/A	A A
Hungarian Puli	USCF1283	N/A	N/A	N/A	N/A	N/A	A A

Hungarian Puli	USCF1284	N/A	N/A	N/A	N/A	N/A	A A
Hungarian Puli	USCF1285	N/A	N/A	N/A	N/A	N/A	A A
Hungarian Puli	USCF1286	N/A	N/A	N/A	N/A	N/A	A A
Hungarian Puli	USCF1287	N/A	N/A	N/A	N/A	N/A	A A
Hungarian Puli	USCF1289	N/A	N/A	N/A	N/A	N/A	A A
Hungarian Puli	USCF1311	N/A	N/A	N/A	N/A	N/A	T T
Hungarian Puli	PUL001	N/A	N/A	N/A	N/A	N/A	A A
Hungarian Puli	PUL002	N/A	N/A	N/A	N/A	N/A	A A
Hungarian Puli	PUL003	N/A	N/A	N/A	N/A	N/A	A A
Hungarian Puli	PUL004	N/A	N/A	N/A	N/A	N/A	A A
Hungarian Puli	PUL005	N/A	N/A	N/A	N/A	N/A	A A
Hungarian Puli	PUL006	N/A	N/A	N/A	N/A	N/A	A A
Hungarian Puli	PUL007	N/A	N/A	N/A	N/A	N/A	A A
Hungarian Puli	PUL008	N/A	N/A	N/A	N/A	N/A	A A
Hungarian Puli	PUL009	N/A	N/A	N/A	N/A	N/A	A A
Hungarian Puli	PUL010	N/A	N/A	N/A	N/A	N/A	A A
Hungarian Puli	PUL011	N/A	N/A	N/A	N/A	N/A	A A
Hungarian Puli	PUL012	N/A	N/A	N/A	N/A	N/A	A A

Hungarian Puli	PUL013	N/A	N/A	N/A	N/A	N/A	A A
Hungarian Puli	PUL014	N/A	N/A	N/A	N/A	N/A	A A
Hungarian Puli	PUL015	N/A	N/A	N/A	N/A	N/A	A A
Hungarian Puli	PUL016	N/A	N/A	N/A	N/A	N/A	A A
Hungarian Puli	PUL017	N/A	N/A	N/A	N/A	N/A	A A
Hungarian Puli	PUL018	N/A	N/A	N/A	N/A	N/A	A A
Hungarian Puli	PUL019	N/A	N/A	N/A	N/A	N/A	A A
Hungarian Puli	PUL020	N/A	N/A	N/A	N/A	N/A	A A
Hungarian Puli	PUL021	N/A	N/A	N/A	N/A	N/A	A A
Hungarian Puli	PUL022	N/A	N/A	N/A	N/A	N/A	A A
Hungarian Puli	PUL023	N/A	N/A	N/A	N/A	N/A	A A
Hungarian Puli	PUL024	N/A	N/A	N/A	N/A	N/A	A A
Hungarian Puli	PUL025	N/A	N/A	N/A	N/A	N/A	A A
Hungarian Puli	PUL026	N/A	N/A	N/A	N/A	N/A	A A
Hungarian Puli	PUL027	N/A	N/A	N/A	N/A	N/A	A A
Hungarian Puli	PUL028	N/A	N/A	N/A	N/A	N/A	A A
Hungarian Puli	PUL029	N/A	N/A	N/A	N/A	N/A	A A
Hungarian Puli	PUL030	N/A	N/A	N/A	N/A	N/A	A A

Hungarian Puli	PUL031	N/A	N/A	N/A	N/A	N/A	A A
Hungarian Puli	PUL032	N/A	N/A	N/A	N/A	N/A	A A
Hungarian Puli	PUL033	N/A	N/A	N/A	N/A	N/A	A A
Hungarian Puli	PUL034	N/A	N/A	N/A	N/A	N/A	A A
Hungarian Puli	PUL035	N/A	N/A	N/A	N/A	N/A	A A
Hungarian Puli	PUL036	N/A	N/A	N/A	N/A	N/A	A A
Hungarian Puli	PUL037	N/A	N/A	N/A	N/A	N/A	A A
Hungarian Puli	PUL038	N/A	N/A	N/A	N/A	N/A	A A
Hungarian Puli	PUL039	N/A	N/A	N/A	N/A	N/A	A A
Hungarian Puli	PUL040	N/A	N/A	N/A	N/A	N/A	A A
Hungarian Puli	PUL041	N/A	N/A	N/A	N/A	N/A	A A
Hungarian Puli	PUL042	N/A	N/A	N/A	N/A	N/A	A A
Hungarian Puli	PUL043	N/A	N/A	N/A	N/A	N/A	A A
Hungarian Puli	PUL044	N/A	N/A	N/A	N/A	N/A	A A
Hungarian Puli	PUL045	N/A	N/A	N/A	N/A	N/A	A A
Hungarian Puli	PUL046	N/A	N/A	N/A	N/A	N/A	A A
Hungarian Puli	PUL047	N/A	N/A	N/A	N/A	N/A	A A
Hungarian Puli	PUL048	N/A	N/A	N/A	N/A	N/A	A A

Hungarian Puli	PUL049	N/A	N/A	N/A	N/A	N/A	A A
Hungarian Puli	PUL051	N/A	N/A	N/A	N/A	N/A	A A
Hungarian Puli	PUL052	N/A	N/A	N/A	N/A	N/A	A A
Hungarian Puli	PUL053	N/A	N/A	N/A	N/A	N/A	A A
Hungarian Puli	PUL054	N/A	N/A	N/A	N/A	N/A	A A
Hungarian Puli	PUL055	N/A	N/A	N/A	N/A	N/A	A A
Hungarian Puli	PUL056	N/A	N/A	N/A	N/A	N/A	A A
Hungarian Puli	PUL057	N/A	N/A	N/A	N/A	N/A	A A
Hungarian Puli	PUL058	N/A	N/A	N/A	N/A	N/A	A A
Hungarian Puli	PUL059	N/A	N/A	N/A	N/A	N/A	A A

All Hungarian Pumi used in this study (n = 152) had the ‘A A’ genotype at BBS4 c.58A > T, determined by Sanger sequencing.

File S2. Completion of the current *BBS4* annotation in CanFam3.1 reference genome

Introduction

Exon 1 of BBS4 is not annotated in the most current reference sequence (CanFam3.1). Its absence is evident by a lack of an initiation codon. When observing a multiple sequence alignment of BBS4 protein sequences from a variety of vertebate species including human, orangutan, mouse, rat, cow, cat and elephant it is clear that the dog is lacking the first part of the transcript. We hypothesized that exon 1 resided in a reference genome assembly gap ~9.7 Kb upstream to the current BBS4 annotation and adjacent to a region of high guanine-cytosine density.

Methods

Using the popular *de novo* aligner Velvet version 1.2.10 (Zerbino and Birney 2008), we attempted to resolve this gap by assembling unmapped reads and reads that partially mapped to the vicinity of the gap in chromosome 30 from four Hungarian Puli dogs. The initial attempt at assembly was unsuccessful in building a contig that completely resolved the gap. Alternatively, we performed a manual alignment using sequences from unmapped mates of reads that had aligned adjacent to the gap. A multiple sequence alignment with the assembled contig and exon 1 of human (NR_033028.4), mouse (NM_175325.3) and cat (XM_011282956.1) was performed using Clustal Omega (Sievers *et al.* 2011). The assembled contig was translated into an amino acid sequence using ExPASy's translate tool (Gasteiger *et al.* 2003). We similarly aligned the predicted canine BBS4 protein corresponding to exon 1 to human (NP_149017.2), mouse (NP_780534.1) and cat (XP_011281258.1) BBS4 proteins.

Results and Conclusions

Multiple sequence alignment of the contig produced from manual assembly revealed that putative exon 1 of *BBS4* in the dog is identical to that of the domestic cat (*Felis catus*) and differs from the human sequence by two nucleotides (Figure S1). Thus, the

complete BBS4 protein in dogs consists of 520 amino acids encoded by 1,560 base pairs of mRNA organised into 16 exons on chromosome 30 of CanFam 3.1. Protein sequences corresponding to exon 1 of *BBS4* are identical for dog and cat but differ to human and mouse by one and five amino acids respectively (Figure S2).

```
Canine Contig  AGCCAAGATGGCTGAGGAGAGGCTGGCGACGGTGAGCGCCGACCTGCCGCTCGGTGTCCC
Homo sapiens   -------ATGGCTGAGGAGAGAGTCGCGACG----------------------------
Mus musculus   -------ATGGCTGAAGTGAAGCTTGGGATG----------------------------
Felis catus    -------ATGGCTGAGGAGAGACTCGCGACG----------------------------
                      ******** * **   * * ** *
```

Figure S1. Multiple sequence alignment of a manually assembled canine contig with exon 1 of *BBS4* of human (*Homo sapiens*), mouse (*Mus musculus*) and domestic cat (*Felis catus*) nucleotide sequences. The canine contig was assembled using reads from four Hungarian Puli dogs. Reads include unmapped mates of pairs that aligned adjacent to a reference genome gap on chromosome 30, putative to the location of *BBS4* exon 1. An asterisk denotes full identity of the nucleotide across species.

```
Canis familiaris      MAEERLATRTQFPASAESQKPRLKK
Felis catus           MAEERLATRTQLPASAESQKPRLKK
Homo sapien           MAEERVATRTQFPVSTESQKPRQKK
Mus musculus          MAEVKLGMKTQVPASVESQKPRSKK
                      *** ::. :**.*.*.****** **
```

Figure S2. Multiple amino acid sequence alignment of partial canine BBS4 protein with domestic cat (*Felis catus*), human (*Homo sapien*) and mouse (*Mus musculus*) homologs corresponding to exon 1 and 2 only. The canine protein sequence corresponding to exon 1 (highlighted in grey) and exon 2 was obtained from translation of genomic sequence of a contig produced by manual *de novo* assembly of canine Illumina HiSeq 2000 reads.

The complete mRNA and amino acid sequences for canine *BBS4* have been deposited in Genbank (KX290494).

Literature Cited

Gasteiger, E., A. Gattiker, C. Hoogland, I. Ivanyi, R. D. Appel *et al.*, 2003 ExPASy: The proteomics server for in-depth protein knowledge and analysis. Nucleic Acids Res. 31(13): 3784–3788.

Sievers, F., A. Wilm, D. Dineen, T. J. Gibson, K. Karplus *et al.*, 2011 Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol. 7: 539.

Zerbino, D. R., and E. Birney, 2008 Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18(2): 821–829.

Table S1. Candidate genes for progressive retinal atrophy in the Hungarian Puli dog breed. Candidates were selected from the region with the highest density of SNP markers (chromosome 30, 25.3 – 40.0 Mb on CanFam 3.1) that were concordant to a recessive pattern of inheritance. All genes have a phenotypic connection to vision as indicated by the Mouse Genome Browser.

Gene	CanFam 3.1 Position	Ensembl Transcript ID
CPLX3	chr30: 37,888,801-37,892,561	ENSCAFT00000028507
CSK	chr30: 37,866,652-37,869,289	ENSCAFT00000028485
STRA6	chr30: 37,332,568-37,346,312	ENSCAFT00000048873
BBS4	chr30: 36,063,713-36,109,202	ENSCAFT00000028102
HEXA	chr30: 35,838,158-35,843,722	ENSCAFT00000028088

GLCE	chr30: 33,142,327-33,208,130	ENSCAFT00000046216
CLN6	chr30: 32,246,411-32,264,240	ENSCAFT00000027690
SMAD3	chr30: 31,246,313-31,360,098	ENSCAFT00000027577
MAP2K1	chr30: 30,683,192-30,760,479	ENSCAFT00000043934
MEGF11	chr30: 30,234,191-30,446,063	ENSCAFT00000027347
SLC24A1	chr30: 29,967,634-29,996,958	ENSCAFT00000027314
RAB8B	chr30: 27,784,338-27,845,901	ENSCAFT00000026890

**NR2E3 (*chr30: 35,378,421-35,381,822) was excluded as it is a known canine PRA gene. A preliminary study (Chew et al., 2017 [Animal Genetics in press]) confirms that no putative variants are present in this gene.

Literature Cited

Chew, T., B. Haase, C. E. Willet, and C. M. Wade, 2017 Exclusion of known progressive retinal atrophy genes for blindness in the Hungarian Puli. Anim. Genet. 48: 500–501.

Table S2. Relationships between 14 Hungarian Puli individuals from the same pedigree estimated through proportion of identity by descent (IBD) calculations performed using PLINK (Purcell *et al.* 2007). Relationships were obtained from pedigree records (Australian National Kennel Council).

Individual 1	Individual 2	Relationship	Proportion IBD
USCF532	USCF347	OT	0
USCF532	USCF350	PO	0.5
USCF532	USCF516	OT	0
USCF532	USCF517	PO	0.5
USCF532	USCF518	HS	0.244
USCF532	USCF519	HS	0.1755
USCF532	USCF520	FS	0.3737
USCF532	USCF522	PO	0.5
USCF532	USCF523	HS	0.2037
USCF532	USCF524	PO	0.5
USCF532	USCF525	OT	0.1626
USCF532	USCF526	PO	0.5
USCF532	USCF521	OT	0.2215
USCF347	USCF350	OT	0
USCF347	USCF516	PO	0.5
USCF347	USCF517	OT	0.0551
USCF347	USCF518	OT	0
USCF347	USCF519	OT	0.0577
USCF347	USCF520	OT	0
USCF347	USCF522	OT	0

USCF347	USCF523	OT	0.1117
USCF347	USCF524	HS	0.2835
USCF347	USCF525	OT	0
USCF347	USCF526	OT	0
USCF347	USCF521	OT	0
USCF350	USCF516	OT	0.0656
USCF350	USCF517	OT	0.3262
USCF350	USCF518	PO	0.5
USCF350	USCF519	OT	0.1257
USCF350	USCF520	OT	0.3004
USCF350	USCF522	FS	0.4293
USCF350	USCF523	OT	0.2823
USCF350	USCF524	OT	0.393
USCF350	USCF525	OT	0.2628
USCF350	USCF526	HS	0.2777
USCF350	USCF521	OT	0.2286
USCF516	USCF517	OT	0.183
USCF516	USCF518	OT	0.1856
USCF516	USCF519	HS	0.3014
USCF516	USCF520	OT	0.1565
USCF516	USCF522	OT	0
USCF516	USCF523	OT	0.1211
USCF516	USCF524	OT	0.2613
USCF516	USCF525	PO	0.5

USCF516	USCF526	OT	0.1235
USCF516	USCF521	OT	0.2273
USCF517	USCF518	OT	0.5
USCF517	USCF519	OT	0.2414
USCF517	USCF520	PO	0.5037
USCF517	USCF522	OT	0.3605
USCF517	USCF523	OT	0.2289
USCF517	USCF524	HS	0.3298
USCF517	USCF525	OT	0.3651
USCF517	USCF526	OT	0.3112
USCF517	USCF521	OT	0.3831
USCF518	USCF519	OT	0.2845
USCF518	USCF520	PO	0.5
USCF518	USCF522	PO	0.5
USCF518	USCF523	HS	0.374
USCF518	USCF524	OT	0.5
USCF518	USCF525	OT	0.3875
USCF518	USCF526	OT	0.3268
USCF518	USCF521	FS	0.5548
USCF519	USCF520	HS	0.3208
USCF519	USCF522	OT	0
USCF519	USCF523	HS	0.2537
USCF519	USCF524	PO	0.5
USCF519	USCF525	PO	0.5

USCF519	USCF526	OT	0.1335
USCF519	USCF521	HS	0.3223
USCF520	USCF522	OT	0.3634
USCF520	USCF523	HS	0.2515
USCF520	USCF524	PO	0.5049
USCF520	USCF525	PO	0.5
USCF520	USCF526	OT	0.403
USCF520	USCF521	PO	0.5
USCF522	USCF523	OT	0.2233
USCF522	USCF524	OT	0.3924
USCF522	USCF525	OT	0.2033
USCF522	USCF526	HS	0.2459
USCF522	USCF521	OT	0.2668
USCF523	USCF524	PO	0.5
USCF523	USCF525	OT	0.2419
USCF523	USCF526	OT	0.1263
USCF523	USCF521	HS	0.3722
USCF524	USCF525	OT	0.3249
USCF524	USCF526	OT	0.3451
USCF524	USCF521	OT	0.5
USCF525	USCF526	OT	0.2616
USCF525	USCF521	FS	0.4689
USCF526	USCF521	PO	0.5

Samples were genotyped on the CanineHD BeadChip array. Relationships provided by pedigree records are consistent with proportion of IBD estimations that were expected

depending on the type of relationship. Parent-offspring (PO) relationships have an expected IBD = 0.5; full-sibling (FS) relationships have an expected IBD = 0.5; half-sibling relationships have an expected IBD = 0.25. OT indicates ‘other’ relationships.

References

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira et al., 2007 PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81: 559–575.

Appendix IV: Supplementary material for chapter 5

Table S1. Candidate genes screened in canine individual USCF305 presenting with severe haemophilia A. Genes associated with a bleeding tendency phenotype in humans were selected as candidates.

Gene	CanFam 3.1 Position	Ensembl Transcript ID	Reference
F2	chr18:42,782,384 - 42,799,459	ENSCAFG00000009122	Archarya et al 2003
F7	chr22:60,572,511 - 60,582,729	ENSCAFG00000006257	Archarya et al 2003
F8	chrX:122,897,137 - 123,043,373	ENSCAFG00000019631	Archarya et al 2003
F9	chrX:109,501,341 - 109,533,798	ENSCAFG00000018998	Archarya et al 2003
F10	chr22:60,585,600 - 60,596,983	ENSCAFG00000006258	Archarya et al 2003
F11	chr11:44,466,300 - 44,487,120	ENSCAFG00000007348	Archarya et al 2003
F13A1	chr35:6,185,898 - 6,347,853	ENSCAFG00000009509	Archarya et al 2003
F13B	chr7:5,674,454 - 5,702,366	ENSCAFG00000011416	Archarya et al 2003
FGA	chr15:52,238,946 - 52,246,920	ENSCAFG00000023178	Acharya and Dimichele, 2008
FGB	chr15:52,220,662 - 52,229,692	ENSCAFG00000008424	Acharya and Dimichele, 2008
FGG	chr15:52,261,220 - 52,270,169	ENSCAFG00000008440	Acharya and Dimichele, 2008
GPIIIa	chr9:9,182,562 - 9,231,046	ENSCAFG00000013735	Nurden, 2006
ITGA2B	chr9:19,050,132 - 19,063,992	ENSCAFG00000014145	Nurden, 2006
vWF	chr27:38,834,909 - 38,972,738	ENSCAFG00000015228	Archarya et al 2003

Table S2. SNPs detected in the *F8* gene in whole genome sequencing data of one Australian Kelpie with haemophilia A (USCF305) and in 11 unrelated controls of the same breed. SNPs were genotyped as homozygous alternative in affected dog USCF305 and one or more of the 11 control dogs in whole genome sequencing data. Positions are relative to the CanFam 3.1 reference genome. Two missense and three synonymous SNPs were identified and have been previously reported in dbSNP.

Chromosome	CanFam 3.1 Position	Exon	Reference Allele	Alternative Allele	Consequence	Amino acid	Reference SNP cluster ID
X	122,938,611	15	G	A	synonymous	N	rs852651766
X	122,956,540	14	G	A	missense	P/L	rs852844707
X	122,957,205	14	C	T	synonymous	E	rs851733901
X	123,043,038	1	G	A	synonymous	D	rs852021679

Table S3. Improperly paired reads of USCF305 aligning to CanFam 3.1 in intron 22 in *FVIII*

Read ID	Forward read mapping position	Reverse read mapping position	Forward read sequence	Reverse read sequence
HWI-ST1213:110:C0MHBACXX:7:2215:19129:75378	122,916,994	123,304,842	TNAGCAACGGGGAAG CAGTCAGTAGGTAAGA AAATACAAAAGAGGCC CATCTGACACAGACTC CGCCACCAGTCCTGCG CACTCACGTGGCTGCC TGGAAG	ATGTAGGCCTGGGCAG CTTTCTTACTGTCTTAT GACAAGAATGCTTAGG AGTTACGGAATGTGACT GGTGATAGTATTTGGGT TTGGGTTTAAGAAAAAG C
HWI-ST1213:110:C0MHBACXX:7:1210:13147:51601	122,917,100	122,917,100	TTCGGAGCCCTAAAAG CCTAGTCTAACTTATTG CAACAGTGTTAGGGTGT ATCCTCCTTTGTAACTTA GCTTTTTCTGGTACAAT CTTCTCAACCGGAAAT	GATTCTGTTCATTTATAT CTCTAGAGAAATCCAAT GCTGCTCATATACCTAA CACCAGGGTTTTTGGTA ACCTCTCTATATCATCA ATGCAAGGAGTTAGA
HWI-ST1213:110:C0MHBACXX:7:1210:2901:13903	122,917,016	123,304,874	TAGGTAAGAAAATACA AAAGAGGCCCATCTGA CACAGACTCCGCCACC AGTCCTGCGCACTCA CGTGGCTGCCTGGAA GGGTCTTTCGGAGCC CTAAAAGC	TTACTGTCTTATGACAA GAATGCTTAGGAGTTAC GGAATGTGACTGGTGA TAGTATTTGGGTTTGG GTTTAAGAAAAAGCCTC CTTAGGCCTCTGGTCT NA
HWI-ST1213:110:C0MHBACXX:7:2215:19102:8424	122,917,020	123,304,874	TAAGAAAATACAAAA GAGGCCCATCTGACA CAGACTCCGCCACC AGTCCTGCGCACTCAC GTGGCTGCCTGGAAGG GTCTTTCGGAGCCCTA AAAGCCTAG	ACTGTCTTATGACAAG AATGCTTAGGAGTTACG GAATGTGACTGGTGATA GTATTTGGGTTTGGGTT TAAGAAAAAGCCTCCTT AGGCCTCTGGTCTAANT
HWI-ST1213:110:C0MHBACXX:7:1311:3150:27769	122,917,037	123,304,863	GGCCCATCTGACACA GACTCCGCCACCAGT CCTGCGCACTCACGT	TCTTACTGTCTTATGAC AAGAATGCTTAGGAGT TACGGAATGTGACTGG

			GGCTGCCTGGAAGGGT CTTTCGGAGCCCTAAA AGCCTAGTCTAACTTAT TGCAACA	TGATAGTATTTGGGTTT GGGTTTAAGAAAAAGC CTCCTTAGGCCTCTGG TCT
HWI-ST1213:110: C0MHBACXX:7:2109: 10573:9641	122,917,058	123,304,942	GCCACCAGTCCTGCG CACTCACGTGGCTGC CTGGAAGGGTCTTTCG GAGCCCTAAAAGCCT AGTCTAACTTATTGCA ACAGTGTTAGGGTGTA TCCTCCTT	AAGAAAAAGCCTCCTT AGGCCTCTGGTCTAAC TCCTTGCATTGATGATA TAGAGAGGTTACCAAA AACCCTGGTGTTAGGT ATATGAGCAGCATTGG ATTT
HWI-ST1213:110: C0MHBACXX:7:1208: 6051:65684	122,917,105	122,917,105	TCATTTATATCTCTAGA GAAATCCAATGCTGCT CATATACCTAACACCA GGGTTTTTGGTAACCT CTCTATATCATCAATG CAAGGAGTTAGACCA GAGGC	AGCCCTAAAAGCCTA GTCTAACTTATTGCAA CAGTGTTAGGGTGTAT CCTCCTTTGTAACTTA GCTTTTTCTGGTACAA TCTTCTCAACCGGAAA TGTAGG