Understanding dog breed copy number differences in the framework of gray wolf copy number variation

Aitor Serres Armero

TESI DOCTORAL UPF / ANY 2020

Dr. Tomàs Marquès i Bonet

Dr. David de Juan Sopeña

DEPARTAMENT DE CIÈNCIES EXPERIMENTALS I DE LA SALUT

A mis padres y familia

Acknowledgements

For a lack of time and the extreme weariness with which I am writing this section, I will keep my acknowledgements very standard. I apologise to the many people who will be left out, if you don't find your name here it does not mean that you are any less appreciated!

With regards to the development of this work, thank you **Tomàs** and **David** for making everything possible. Thank you **Tomàs** for giving me the opportunity to work with you during these four years and for helping me grow both scientifically and personally. Thank you **David** for your advice and your ideas on the project, specially regarding its technicalities and specifics. Last but in no way least, thank you **Inna** for helping me during the first years of my PhD and teaching me all I needed to know about tackling any scientific challenges, no matter how difficult they might seem.

Thanks to the Marquès-Bonet lab for their support, friendship and care during all these years. Of course, thank you **Clàudia, Marc, Paula, Irene** and **Laura** for being directly involved in the writing, correcting and editing of this thesis. Without your help, I still would not even know how the singular and plural of CNV work. I don't think I could have wished for a better company during all these years, and of course your scientific and personal merit go without saying! I won't ever find people so determined to playing cards during our lunch breaks, but it makes work much more fun doesn't it?

To many past and present colleagues: **Lukas, Marina (Brasó), Raquel,**

Esther, Diego, Marina (Álvarez), Txema, Jéssica, Alejandro, Xavi (both of you!), Marco, Juan, Guillem, Ignasi, Manu, Jonas, Fátima, Tom, Nerea, Martin, Sojung, Joe etc. I wish I had the time to write more... perhaps in another PhD! And of course to all the IBE people I'm unwillfully omitting.

Of course thank you **Manolo** and **Marina** for being the founders of Sardenya 383. Even if you would much rather be living in the hotlands of Seville, I'm so happy that we decided to share the flat! I'm sure I'll miss the *Carnavales* when washing the dishes and I'm quite certain that I'm not going to keep watching the Peaky Blinders unless it is during our series nights (too much Star Trek to catch up on, you know). **Dani**, **Marta** and **Lucky** it's been awesome living together during these years without any quarrelling and in a super cool vibe!

Thank you **Luis, Toni, Pablo** and **Sandra** for letting me into your smoking breaks, even if I'm not much good for the smoking part. Other than that, I'm still waiting for the famed *Cenita Guapa* to happen, you know.

To all the **friends at Corbera,** even if I have not been going there so often lately, to the all the **Xipxoperos**, and to all the colleagues from Munich!

Abstract

The study of structural variation complements and synergizes with the study of sequence variation to unravel the intricacies of phenotypic variation. Dogs are the most phenotypically variable domesticated species existing today despite their remarkably low nucleotide diversity. As such, the systematized study of copy number variation in an extensive panel of over 100 dog breeds has the potential to unravel a fraction of the bases of phenotypic diversity which remain unexplored. This study finds an excess of structural variants in dogs compared to the expectation given their genetic history, which can potentially account for some of their morphometric, anatomical and pathological variance. Indeed, trait mapping finds over 90 copy number variants associated with more than 10 phenotypes, some of which were previously unknown or uncharacterized. Moreover, there is a correlation between low effect, associated copy number variants and other relevant genomic features such as the expression patterns of long non-coding RNA or the presence of long-range chromatin contacts. Our characterization of copy number variation in dogs has generated a wealth of hypotheses for further functional testing and validation.

L'estudi de la variació estructural complementa l'estudi de la variació de seqüència per revelar les complexitats de la variació fenotípica. Els gossos són l'espècie domesticada amb més variació fenotípica que existeix avui en dia, malgrat la seva baixa diversitat nucleotídica. Com a tal, estudiar sistemàticament la variació del nombre de còpies en un panell compost de més de 100 races de gossos, permet descobrir algunes de les bases d'aquesta diversitat fenotípica. En la recerca presentada en

aquesta tesi, hem descrit un excés de variants estructurals en gossos si ho comparem amb el que s'esperaria segons la seva història genètica. Aquest fet podria explicar part de la seva variació morfomètrica, anatòmica i patològica. Per altra banda, l'estudi d'associació fenotípica que hem realitzat troba més de 90 variants de nombre de còpia associades a més de 10 fenotips, alguns dels quals desconeguts fins al moment. A més a més, hem trobat una correlació entre variants de nombre de còpia de baix efecte estadístic i altres variants genòmiques rellevants, com ara els patrons d'expressió d'ARN llargs no codificants o la presència de contactes de cromatina. El nostre estudi ha generat una gran quantitat d'hipòtesis que poden donar lloc a validacions funcionals posteriors.

Preface

The first attempts to quantize the number of copies of a determined genomic segment predate the discovery of the structure of the DNA molecule itself. Barbara McClintock's discovery of transposable elements in 1930 and similar contemporary findings on chromosomal rearrangements by Theodosius Dobzhansky, among others, pioneered the study of structural variation. Since then, the relevance of structural variation has been increasingly acknowledged in the field of genetics as a substantial cause for disease, physiological differences, and even speciation.

Dogs have mystified humans since time immemorial, and their presence in ancient culture is evidenced in religion, archeological remains, art and written texts. There has been strong evidence of their presence in certain human societies throughout the whole neolithic and even further in the past. Dogs were cherished watchers and shepherds during ancient greek and roman times and appreciated hunting assistants in the middle ages. The Victorian era saw their rise in popularity as companions and entertainers, as a consequence, intensive breeding programs were started which have persisted until the present. Way more recently, genetics have allowed us to study dogs in much more detail. Besides recapitulating part of the history of dog domestication, over 170 dog diseases and more than 40 traits have been found to have a genetic cause.

To date, copy number variation in dog whole-genomes is still an understudied subject: most copy number variation studies in dogs have been performed using array technologies. As such, there is still the

potential to discover new structural variants, especially of lower size, which could be associated with breed differences and disease. In this work, we present the first whole-genome copy number variation assessment of over 500 canine genomes belonging to over 100 dog breeds with the aim to discover new functional structural variants and assess copy number differences not only across dog breeds but also between other extant canid relatives.

Abbreviations

aCGH: array comparative genomic hybridization
ANN: artificial neural network
bp: base-pair
CBS: circular binary segmentation
CN: copy number
CNV: copy number variation/variant
DP: Dirichlet processes
FISH: fluorescence *in situ* hybridization
FT: fourier transform
GC: gene conversion
GMM: general mixture models
GWAS: genome wide association study
HMM: hidden Markov model
NAHR: non-allelic homologous recombination
NGS: next-generation sequencing
qPCR: quantitative polymerase chain reaction
RD: read depth
SNP: single nucleotide polymorphism
SNV: single nucleotide variation/variant
SSF: scale-space filtering
SV: structural variation/variant
WGAC: whole-genome assembly comparison
WGS: whole-genome sequencing

Table of Contents

1. INTRODUCTION

1.1. About genetics and genomics

Ever since the chemical structure of the DNA molecule was discovered in the mid-1950s (Watson and Crick 1953), reading and understanding it have become the core objectives of modern genetics. Even if most of the theoretical and mathematical groundwork in the field had been established several years earlier (Fisher 1937; Castle 1903; Wright 1950), bringing together these theories and the chemical and physical properties of DNA has proven a challenging task. The realization that eukaryotic DNA is an extremely long and complex molecule, together with the advent of computer science, resulted in a paradigm shift where genetics could no longer be studied theoretically from the perspective of single genes, phenotypes or loci. It became clear that true understanding of the intricacies and nuances of organismal diversity could only be achieved with genome-wide comparisons. In a certain sense, genetics and genomics have been building up towards the ultimate goal of whole-genome comparisons, a goal which only now we are starting to reach. For that purpose, since the late 1970s, there has been a co-evolution of DNA reading tools -chemical sequencing- and DNA understanding tools -statistical and computational-.

In little more than 50 years, many milestones have brought us closer to whole-genome comparisons: the development of polymerase chain reaction (PCR) in the mid 80s (Saiki et al. 1985), together with electrophoresis-based, low-throughput sequencing methods (Sanger and

Coulson 1975), made it possible to compare single, localized DNA sequences such as genes or regulatory elements. The knowledge of particular DNA sequences then enabled the design of variation arrays, which could quickly and conveniently assess predetermined genetic differences over many individuals. Based on the slow but sturdy collection of available methods, the human genome project was launched in the early 1990s with the aim to completely assemble the first whole animal genome (International Human Genome Sequencing Consortium 2004). Later, in the early 2000s, whole-genome (shotgun) sequencing (WGS) technologies bridged over the gap between array technologies and low throughput sequencing by collecting data over whole genomes in many individuals (Voelkerding, Dames, and Durtschi 2009). WGS made it possible to scan a genome for both known and unknown variation but, not least importantly, it facilitated the process of whole-genome assembly.

Parallelly, the computational bases of massive sequence analysis were laid down between the late 1960s and the 1990s. During that time, not only sequence comparison tools were developed, but also complex sequence analysis techniques based on Dynamic Bayes Networks, Dirichlet Processes or Artificial Neural Networks (Murphy 2002). Only now, after massive advances in hardware and computation, are these models being revisited and implemented in the context of bioinformatics.

Since the late 2010s, long read sequencing technologies (Rhoads and Au 2015) offer the opportunity to bring us one step closer to whole-genome comparisons. The ability to produce reads long enough to span most

kinds of variation, although seemingly unexciting, enables the direct observation of genomic variants instead of having to statistically infer their presence. This greatly facilitates the comparison of any genomic variants across similar genomes while also improving the ability to discover orthologous regions in distant species. Complementary to long reads, a promising computational advance is the transition from linear to graph-based reference genomes. Although conceptually more complicated, graph-based genomes can concisely store population information about all kinds of genomic variants and have been hypothesized to be a better platform for whole-genome comparison operations. The prospect of merging graph theory with state-of-the-art population genetics algorithms can prove really fruitful in terms of speed, efficiency and computation feasibility in extremely large datasets (Rakocevic et al. 2019).

1.2. Variation

Chemical variations in the DNA molecule and its surroundings are the driving forces of evolution and change. Variation has made it possible for organisms to adapt and survive for as long as 3-4 billion years, and it is to be thanked for all biodiversity found today. However, we are still far from understanding all the implications and forms of existing variation. We have learned to use genomic variation to predict a number of severe diseases and physical traits, normally those that have the simplest genetic bases, in various organisms (Shastry 2002; Karlsson and Lindblad-Toh 2008). But the biology of complex variation, the intricate interactions between different kinds of variation and implications of variation during

the life of an organism are still fields of active study (Zhu et al. 2018).

Classifying variation is useful but at the same time can be challenging: genomic variation can range from a simple change in a single DNA position to a complete copy of the whole genome (Dehal and Boore 2005). It is even possible to artificially combine the DNA of different species with roughly the same amount of chromosomes and produce viable offspring (Sipiczki 2018). With all this, many classifications can be proposed, here, two main categories of genomic variation will be presented: single nucleotide variation (SNV[1]) and copy number variation (CNV[2]) including repeats. Special emphasis will be placed on CNV since it is the main subject of this work.

It must be noted that variation affecting the genome structure and sequence is just a fraction of all the variation carried by organisms. Other forms of variation acting on the DNA molecule -epigenomic variation- such as methylation, histone modifications or even variation in the compaction and folding of genomic material have a deep impact on transcription regulation and phenotype. However, an in-depth review of these kinds of variation and their interactions with sequence variation is out of the scope of this work, and they will only be introduced in the context of their interactions with copy number variation.

---

[1] SNV will be used as an acronym for the general concept of single nucleotide variation and SNVs will be used as a concretion standing for specific single nucleotide variants. Sometimes the two are interchangeable.

[2] SNV will be used as an acronym for the general concept of single nucleotide variation and SNVs will be used as a concretion standing for specific single nucleotide variants. Sometimes the two are interchangeable.

1.2.1. Single Nucleotide Variation

Single nucleotide variation, also commonly known as point mutations or substitutions, involves the switching of a nucleotide for another at a single base-pair position. Removing or adding a single nucleotide at one or up to a few tens of positions -indels- tends to be considered SNV as well, although the maximum length of a SNV event and the minimum length of a CNV event are completely arbitrary. SNVs can appear as a result of replication errors, when the DNA is being copied, or as a result of DNA damage by extrinsic factors (Cooper 2000). Only SNVs affecting sexual cells -sperm cells and oocytes- will be passed onto the next generation and thus will play a role in evolution in the strict sense. Any other SNVs may affect the survival of a certain individual but will not be affecting the fitness or traits of forthcoming generations.

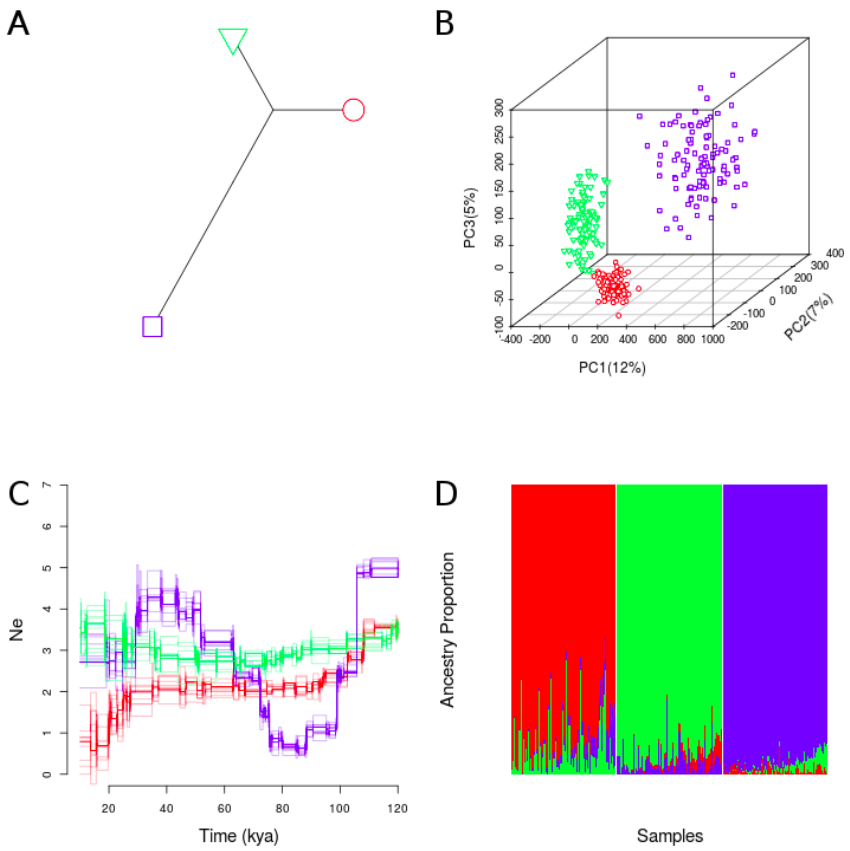

Figure 1: Standard population genetics analyses. Data generated at random for three fictional populations (green, red and blue) to fit each depiction coherently. **A)** Phylogenetic tree for population clustering based on genetic distance. **B)** Principal Component Analysis for population clustering based on SNP distribution and variance (similar to **A**). **C)** Pairwise sequentially Markovian Coalescent for population size estimation through time based on time to coalescence of different genomic segments. **D)** ADMIXTURE for ancestry sharing estimation (optimized number of components K=3).

In itself, single nucleotide variation is pretty limited: if variants were locked into a given chromosome and had no way to get shuffled into others, the amount of diversity which could be generated in a population after accounting for selective pressures would be scarce (Muller 1964).

This is the case for some bacteria and a few eukaryotes, which compensate for it with larger population sizes or low DNA replication fidelity. However, most animal and plant species can undergo recombination (San Filippo, Sung, and Klein 2008), which is the basis of sexual reproduction. Recombination can swap blocks of loci between homologous chromosomes, factorially increasing the amounts of variation which can be amassed even on small populations. Interestingly, the same mechanisms which are responsible for recombination have been proven to be one of the possible causes for copy number variation (L. Chen et al. 2014).

Single nucleotide variation has been the most thoroughly studied over the past five decades, it is also the most prevalent kind of variation, vastly outnumbering all others in terms of sheer number of events (Sudmant, Rausch, et al. 2015). As such, most of the current molecular genetic analysis techniques are tuned in to pick it up among other kinds of variation, even if its implications are not always as relevant (Rafajlović et al. 2014). Nowadays, when possible, it is oftentimes easier to scan for CNV or epigenomic variation by looking at the patterns of its surrounding SNV rather than actually scanning for the specific type of event itself (H. Zhang et al. 2014). Furthermore, given the density of SNV, it is also possible to infer missing variation patterns by looking at the surrounding SNV, since mutations are rarely inherited independently to their closest counterparts (Scheet and Stephens 2006).

Even if many diseases have been linked to SNVs, perhaps their most useful property is their ability to recapitulate the phylogenetic history of

different species. They are the optimal tool to study differences across multiple organisms, since they have been demonstrated to accumulate at a relatively predictable pace and, contrary to other kinds of variation, most of them have been hypothesized to evolve in a nearly neutral manner (Felsenstein 1987).

Single nucleotide variants that can be observed frequently enough to infer population histories or dynamics from goes by the name of single nucleotide polymorphisms (SNPs). SNPs are the basic working unit of population genetics, and through them many interesting population parameters can be assessed such as divergence times, past population sizes or migration events. The estimation of all of these parameters finds its roots in statistical treatment of the number of differences within or across different organisms and populations. For example, given similar population sizes in the ancestor, the divergence time between two populations or species will be proportional to the number of accumulated substitutions (Zuckerkandl and Pauling 1965). Similarly, the population size of a species is related to how long ago a common ancestor can be found for any gene or locus; which itself is related to intra-species diversity (Kingman 1982). In other words, if a common ancestor for most loci in a genome can be found relatively recently, a very reasonable explanation would be that the population is very small and closely related. It follows that the opposite, finding a late common ancestor for many loci or not finding one at all, should mean that the population size is really large (Figure 1). Genetic distances, understood as the pairwise number of genetic changes between two individuals, can be used to build phylogenies, and the layout of these changes in the genome has been

shown to even recapitulate geohistorical patterns (Rendine, Piazza, and Cavalli-Sforza 1986).

Nowadays, SNV is the most studied kind of variation in all organisms. SNV represents a great gateway to study new organisms and place them within the diversity and biology of their extant -or extinct- relatives.

1.2.2. Structural Variation

Structural Variation (SV) takes up a sizable portion of the genome in terms of space, but not necessarily in terms of the number of observable events. The concept of structural variation includes any kind of gain, loss or rearrangement of genetic material at any scale longer than a few nucleotides. Events of up to a few Mbp in length tend to be classified as copy number variants, while larger events are known as chromosomal disorders (Spielmann, Lupiáñez, and Mundlos 2018). However, chromosomal disorders will not be formally addressed in this work because they are rarely found in healthy animal samples.

Copy number variation, as its name intuitively implies, refers to any kind of event which alters the number of copies of a genome locus, which is canonically two in most animals -one per autosome- (Sharp et al. 2005). Submicroscopic rearrangements of genetic material, also known as inversions, should not be referred to as CNVs in a strict sense, since they do not imply any change in the number of copies of a locus, but they have been so often studied together with CNV that for the sake of conciseness they will be reviewed in this section.

CNV is a population-based term: for a locus to be copy number variable, there need to be at least two individuals in a population with a different number of copies of it; otherwise, that event would not be strictly "variable" but fixed. However, many CNV genotyping tools are not able to confidently tell fixed and variable loci apart and therefore fixed, non-diploid loci have often been studied together with CNV. Loci that are known to be duplicated but are not necessarily variable are referred to as segmental duplications[3], a denomination which is only dependant on there being more than two copies of a locus in a single individual. The term continuum of genomic variation
has been coined to refer to all possible copy number events (Conrad and Hurles 2007), ranging from small insertions or deletions to microscopic events.

Depending on how duplications and deletions are distributed in the genome, and whether they have a pair in their homologous chromosome or they are isolated, there might be restrictions in the ways that they can segregate. Based on the maximum and minimum number of copies observed in a population, it should be possible to know if there is any restriction to CNV segregation (Handsaker et al. 2015). If only two modes of segregation are possible in a population, CNVs are termed biallelic, otherwise, they are called multiallelic.

---

[3] Normally, segmental duplications are additionally defined according to a minimum length and identity between duplicates. But these two features are completely arbitrary and may change across studies.

Duplications with a higher number of copies are more likely to be multiallelic and, in fact, it should be theoretically rare to observe biallelic CNV in loci with more than 6 copies in humans (Handsaker et al. 2015). Biallelic duplications are considered to be more suitable as phylogenetic markers than their counterparts (Sudmant, Mallick, et al. 2015).

Classically, CNV has referred to complex sequence events only, meaning that any sort of repeats or transposable elements, which contain extremely low entropy sequences, were usually excluded from the classification. Mostly, repeats have eluded CNV classification for pragmatic reasons, since the methods to call CNV do not cope well with the methods to call repeats and vice versa (Gymrek et al. 2017). However, given the etymology of the concept, there is no intuitive reason why repeats should not be considered CNVs as they are indeed segments of the genome with different number of copies in different individuals.

Historically, special attention has been payed to genic CNVs -events involving partial or whole genes-, since they are major generators of new functions. SNVs alone can slightly modify or disrupt the activity and efficiency of single-copy, existing genes, but they will very rarely change their function completely. CNVs on the other hand, have the potential to shuffle new domains into existing genes, swap regulatory elements or create “backups” of existing genes which can either remain
the same as a safeguard for possible damage or freely mutate into new functions depending on the selective pressures acting upon them (Ohno 1970).

1.2.2.1. Types of CNV

There are mainly three types of CNV: duplications, deletions, and inversions -although inversions do not necessarily involve a change in the number of copies- (Figure 2).

- Duplications:

Duplications, also named insertions or expansions, are arguably the most complex of all structural variants since they comprise a potentially unbound number of genotypes, that is, any number of copies of a locus greater than the euploid is theoretically possible. Additionally, duplications do not necessarily spread across the genome in fixed blocks. On the contrary, it is easy to find partial duplicates or deletions of already duplicated loci generating artifactually misleading copy number signatures (Dennis and Eichler 2016). The history of these sequences is close to unresolvable, especially if they are collapsed in the assembly (Section 1.2.2.3), but that does not mean that accurate segmentation and copy number (CN) estimation cannot be done for such loci.

Interestingly, duplication events tend to happen in the vicinity of their original copy more often than expected by chance (She 2006). That does not mean that a duplication cannot occur in long-range distances or even in another chromosome, it is just a matter of chance and the origination mechanism of the duplication (She 2006). Thus, duplications can be classified in: (I) tandem duplications, if they occur right after the original copy, (II) intrachromosomal duplications, if they occur in the same chromosome as the original copy or (III) interchromosomal duplications if they occur in a different chromosome as the original copy. Tandem

duplications are inherently intrachromosomal but they happen frequently enough that they have received their own denomination. It must be noted that this classification of duplications can only be made if all -or at least a few- copies of the duplication are resolved in the assembly. If all copies of the duplication are collapsed in the assembly -meaning that only one copy is present in the reference, causing reads coming from all copies to be mapped to that region- or confined to unassembled scaffolds, this classification will not be possible. Importantly, the definition of an "original copy" is not trivial and can be perturbed by mechanisms such as gene conversion (Section 1.2.2.2).

- Deletions:

Deletions are the most straightforward kind of copy number variation, as they segregate and evolve in a very similar way to SNV. In diploid organisms, only three genotypes can be observed, corresponding to a total deletion of both copies of a locus, a partial deletion of one of the copies and the absence of the deletion. Genic and regulatory element deletions typically have a predictable outcome, especially total deletions, where no genic product can be created at all. So much so that artificially induced deletions are usually the go-to method for the analysis of gene functions and interactions (Santiago et al. 2008).

Deletions are the most numerous of all CNV events, especially if long indels are considered in the definition. Because of their bigger number of occurrences and their relatively simple method of inheritance, neutrally evolving deletions can potentially be good markers for phylogenetic reconstruction and population genetics studies. However, deletions have

been reported to suffer the biggest impact of GC content bias, and long-read technologies have assessed that some putative deletions are in fact artifacts of outlier GC regions (Duan et al. 2019), even after performing the necessary corrections. Indeed, deletions are the CNV type which is most susceptible to technical biases, since gaps, masked sequences and poor mappability regions will all appear as deletions.

- Inversions:

Inversions have been considered to be the rarest of all CNV events and also the most difficult to study since they cannot be detected through read depth (RD) changes. Long-read technologies might be the most suitable for inversion genotyping since reads up to 10 kbp are expected to span most inversion events. Already at their inception, long reads indicate that inversions might be more prevalent than initially thought: in the whole-genome comparison of great apes, most newly discovered SV events, relative to how many were previously known, corresponded to inversions (Kronenberg et al. 2018).

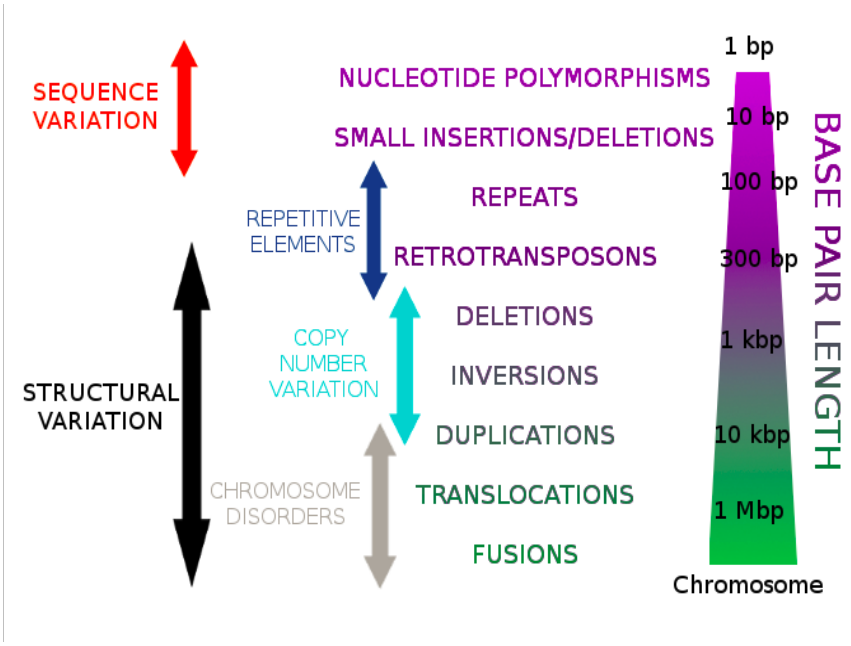

Figure 2: Classification of CNV. On the right is a gradient with an orientative size of each genomic variant. On the left is the most common classification according to the literature. Note that other works could propose different classifications.

1.2.2.2. Mechanisms of origination

Copy number variation tends to happen in unstable regions of the genome such as previous duplications, highly self-similar regions such as repeats, and retrotransposon insertion sites. All these regions are prone to both sustain more DNA breakage and to induce replication flaws, which can result in the intermixing of loci. The most studied CNV origination mechanism is Non-Allelic Homologous Recombination (NAHR) which, as evoked by the name, involves swapping two non-homologous -but similar- loci during DNA recombination (in a meiosis or a strand repair). Depending on how the NAHR took place (Figure 3) and how it is resolved, a duplication, deletion or inversion may be originated (Hastings et al. 2009; Gu, Zhang, and Lupski 2008).

There are a few other mechanisms that can originate CNVs: most prominently, active retrotransposons -viral DNA sequences inserted in a host genome- can copy themselves into another part of the genome, sometimes carrying part of the host's DNA with them. This mechanism will most likely give rise to duplications (Cordaux and Batzer 2009). Additionally, replication stalling, polymerase slippage or template switching caused by open-strand DNA secondary structures can induce the creation of generally small CNVs (Voineagu et al. 2008). It is not trivial to determine which mechanism caused a specific CNV, which makes this an active and relevant field of study nowadays (Ma et al. 2017; Thomas et al. 2019). Moreover, the prevalence of different CNV origination mechanisms across species and different evolutionary periods of a population may vary significantly (Kim et al. 2008; Tomas Marques-Bonet et al. 2009).

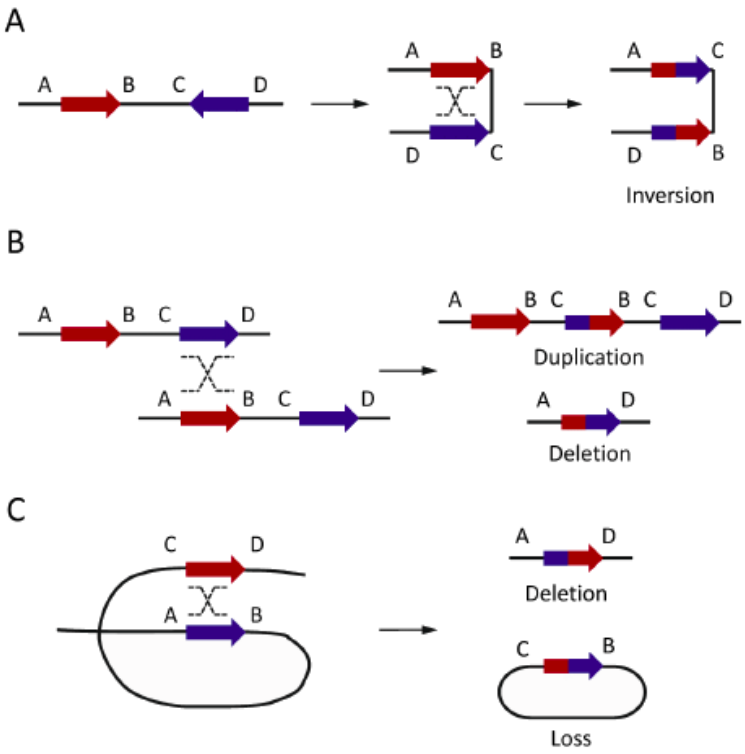

Figure 3: Consequences of NAHR. Arrows represent the direction of homologous sequences. The letters in the sequence determine the order at which loci will appear before (left) and after (right) NAHR. **A)** Intra-chromatid NAHR between inverse direction homologs generates inversions. **B)** Inter-chromatid NAHR between same direction homologs can generate both duplications and deletions. **C)** Intra-chromatid NAHR between same direction homologs generates deletions. Adapted from (Chen et al. 2014).

Finally, the impact of gene conversion (GC) must also be briefly addressed in relation to copy number variation. GC occurs when a previously duplicated genomic sequence is overwritten by any of its paralog regions during DNA synthesis (either recombination or damage repair). Briefly, GC is a byproduct of regular recombination where the resulting DNA heteroduplexes are repaired into a single, concerted haplotype (J.-M. Chen et al. 2007). Gene conversion can potentially tamper with the theoretical background surrounding duplications since it

contradicts some logical assumptions about the rate of accumulation of differences in duplications -i.e the most different copy should most likely be the oldest-. GC has since long been theorized to play a crucial role in the evolution of some gene families (Ohta 1984; Liao 1999).

1.2.2.3. Methods for CNV discovery

Classical methods

Even if most classical molecular methods to detect CNV have dropped in popularity in CNV assessment analyses over the past few years, they remain the most reliable and target-specific options to date. Here, only quantitative polymerase chain reaction (qPCR), fluorescence *in situ* hybridization (FISH) and array comparative genomic hybridization (aCGH) -the three classically most used methods- will be briefly reviewed, but it must be noted that a few other methods with more specific applications and purposes have been developed over the years (Cantsilieris, Baird, and White 2013). The methods reviewed here require a previous partial or total knowledge of the sequences to be assayed. Completely blind determination of CNV loci was generally done using restriction fragment length polymorphisms (Saiki et al. 1985).

- *qPCR*: qPCR (Higuchi et al. 1992) is based on a titered, simultaneous amplification of two DNA sequences, one with known copy number used as the baseline and another which is to be determined. If the same amount of the two DNA sequences is loaded, the ratio of amplification between the probed sequence and the baseline will result in the absolute CN of the probed

sequence. This method is tailored for the CN determination of one or a very small number of sequences, but it is still widely used for a reliable and quick assessment of CNV in a few individuals or loci.

- *FISH*: fluorescence *in situ* hybridization (Langer-Safer, Levine, and Ward 1982) is perhaps the most conceptually straightforward method of CNV determination. It consists on fluorescently labeling a predetermined DNA sequence inside a structurally integral cell and looking at it in the microscope. Of course, the labeling process is complex and the fluorescent signal must be amplified for it to be detectable. FISH normally works best at assessing larger CNV and chromosomal disorders.

- *aCGH*: much like SNP arrays, aCGH was the prevalent technique for massive CNV determination before WGS became widespread (Kallioniemi et al. 1992), and is still being used for established, commercial and biomedical purposes. aCGH consists on a relative measurement of CN between two genomes -generally a control and a test which are stained with different dyes-. Similar to SNP arrays, aCGH consists of an arrangement of thousands of fixed, locatable, different probes in a faceted surface. Each of the thousands of different probes is itself redundant in its own facet so that multiple annealings to the same sequence can happen. The stained genomes anneal to the probes and the dye intensity is measured for both genomes at each facet. If the per-facet dye intensities are significantly different after normalization, it means

that a CNV has been located and, similar to qPCR, the ratio of the two intensities should be indicative of the CNV difference. As a side note, SNP arrays can also be used for CNV determination in a single individual if one allele is found to emit significantly more light than the other.

Next-generation sequencing

As of today, short-read, high-throughput technologies are still the most prevalently used of all sequencing technologies. Their greatest values rely on their cost-effectiveness and their ability to simultaneously uncover both known and novel genomic variation, making them the optimal solution to catalog variants across as many individuals from closely related species as needed (Goodwin, McPherson, and McCombie 2016). Unfortunately, short-read technologies are slightly worse suited to perform distant species comparisons and to catalog different kinds of variation, such as repeats, complex regions or haplotype structures, especially if few samples are available. Long-read sequencing platforms might be able to cover this gap if larger throughputs are ever achieved, since they can much more easily resolve complex variation at the cost of a slightly higher error rate. Nevertheless, short reads have been successfully used to uncover repeats, structural variants and haplotypic data (Willems et al. 2017; Abyzov et al. 2011; Browning and Browning 2007).

NGS offers an unprecedented opportunity to genotype and discover new genomic variants. Different features of NGS data mapped to a reference

can be used to infer the presence of CNV (Kosugi et al. 2019; Pirooznia, Goes, and Zandi 2015). Generally, at least one of the following features is assessed and then a statistical treatment is applied to filter out technical artifacts and make the data manageable. Based on the mapping features of NGS reads, these methods are defined (Figure 4 and Supplementary Table 1):

- *Assembly based CNV discovery:* Assembly based CNV discovery, also known as whole-genome assembly comparison (WGAC) (J. A. Bailey et al. 2001), is becoming increasingly feasible with the advent of third-generation sequencing technologies. WGAC is the most consistent, straightforward and informative structural variation calling method available to date, but it requires a *de novo* assembly of the genome of interest and it is dependant on its quality. WGAC is based on re-aligning a fully assembled genome against itself to look for high identity regions. If similar regions were not collapsed during the assembly process, then WGAC will output the locations of all possible copies of a genomic region (Jeffrey A. Bailey et al. 2004). Before 2018, genomes of a certain species were seldom assembled more than once or twice in order to create a reference and then all resequencing data was mapped against the said reference. The process of *de novo* assembly was costly and often unreliable, so WGAC would largely be computed only for a few reference genomes. The first attempts to move away from single reference genomes and construct individual assemblies for each sample have proven useful in discovering new structural variation (Kronenberg et al. 2018).

- *Insert size CNV discovery:* Insert sizes are specific features from paired-end sequencing data, the most used mode of sequencing nowadays. In paired-end sequencing mode, each DNA fragment in a DNA library is read twice, once from each end, but the whole fragment is not completely sequenced. Normally, the DNA is fragmented in segments long enough that their middle portion, the insert, remains unsequenced. The distribution of inserts lengths -sizes- is therefore known from the sequencing process.

  Insert size CNV discovery consists on comparing the known insert sizes to the actual distance between the mapped read pairs. If a pair of reads maps closer together in the reference than expected given their insert size, it means that the insert could not be found in the reference i.e. there is a gain of genetic material in the assayed sample. Conversely, if the two reads map further apart than expected from their insert size, it means that the reference genome has more genetic material between the two reads than the sample and therefore the sample has a deletion (K. Chen et al. 2009).

  Paired-end sequencing especially excels at finding inversions because of the special orientation properties of read pairs: the paired-end sequencing process ensures that each read pair will be sequenced in inward, opposing directions. Since most mapping software record the relative orientation of each read pair, systematic instances where multiple read pairs map in the same

direction -instead than opposing directions- in a confined genomic region are strong indicators of the presence of an inversion breakpoint.

- *Split mapping CNV discovery:* Split mapping CNV discovery seeks to find the breakpoints of duplications and deletions by exploring the properties of the mapped reads. If a read spans the breakpoint of a duplication or deletion, only a portion of it should map to the reference and the rest should either map to a different location or not map at all. If this pattern is consistently seen in multiple reads mapping to the same region, a certain confidence can be assigned to the SV breakpoint call (Layer et al. 2014).

- *Read depth:* Read depth CNV discovery uses the amount of reads mapped to a certain region to infer the number of copies in such region of the genome. Contrary to assembly-based CNV discovery methods, read depth assumes that most copy number variable regions will be represented only once in the genome i.e. most copies will be collapsed into a single region in the assembly process. Enrichment or depletion of mapped reads in relation to the average genomic depth is, therefore, a good indicator of CNV presence (Abyzov et al. 2011). Furthermore, a discretization of read depth fold enrichment should indicate the absolute number of copies of a certain region in the genome. Based on mapping, two approaches exist to CNV read depth discovery: using all possible positions of all reads in the genome -extensive mapping-, and using only a subset of the best alignments. Although both

methods should be able to infer absolute CN, extensive mapping is more robust to poor quality assemblies and artifacts at the cost of being computationally more expensive (Alkan et al. 2009).

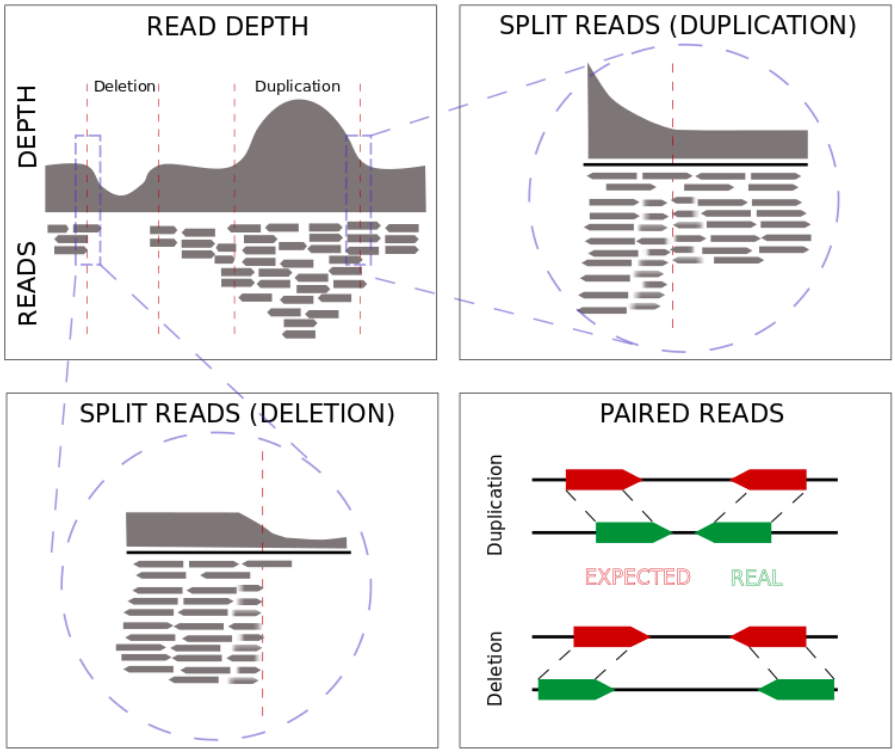

Figure 4: Cartoon of CNV calling methods. Methods explained in paragraphs above.

Statistical treatment

In practice, all the methods introduced above require different sorts of post-processing in order to be analyzed. Statistically speaking these methods could be classified into two main categories: discretization and segmentation. Some of the most used algorithms nowadays, normally taking care of both are: general mixture models, Dirichlet processes,

scale-space filtering, circular binary segmentation, hidden Markov models, and artificial neural networks.

- *General mixture models (GMM):* GMM are probably the simplest of the above. In short, they assume that the feature observations were drawn from a finite number of different distributions and try to estimate the best fitting parameters for said distributions, as well as the fraction of the data that each distribution takes up (Everitt and Hand 1981). Simply put, general mixture models are a generalization of single distribution fitting to multiple distributions, however, there is rarely an analytical, maximum likelihood solution to this fit and therefore optimization needs to be applied. GMMs require previous knowledge of the number and types of distributions that need to be fitted.

- *Dirichlet processes (DP):* DP can serve a similar classification purpose to GMMs and can be applied to very similar problems. However, Dirichlet processes do not require a pre-specification of the number or type of distributions to be fitted. Instead, they assume a preferential attachment background and a categorical distribution output which are interdependent and regulated by a "thinning" parameter (Ferguson 1973). This parameter will simultaneously regulate how many categories are created and how many observations each category is expected to explain. Dirichlet processes can be nested into other Dirichlet processes deriving into more complex models such as Latent Dirichlet allocation. Interestingly, some of the realizations of Dirichlet processes, such

as the stick-breaking process and, especially, the Pólya urn scheme bear a great resemblance to a simplified model for the duplication of alleles (Mimori et al. 2015).

- *Scale-space filtering (SSF):* SSF consists on convolving the data with a smoothing kernel multiple times using different smoothing constants -binwidths-. Effectively, this is analogous to calculating a density graph on discrete data, which is routinely done by many data analysis softwares (Ramsay and Scott 1993). The rate of change of the smoothed data under different kernels can be used as an indicative of whether there is a significant change in the unsmoothed data itself. Calculating the rate of change of the smoothed data instead of the raw data is especially useful to process very noisy signals. The multiple smoothed signals can be then compiled into segmented and discretized output (Witkin 1987). SSF has been extensively used in image processing and telecommunications.

- *Circular binary segmentation (CBS):* CBS is one of the multiple realizations of the sliding window style of algorithms. It consists on sequentially analyzing groups of observations -windows- and formally testing whether there are differences between contiguous groups. Contrary to canonical GMMs or DPs, CBS accounts for the order in which observations appear, thus rejecting the assumption that observations are sequentially independent. CBS is quickly implementable and applicable but it cannot emit absolute values, instead, it can only locate where changes within a

sequence occur (Olshen et al. 2004). A variety of tests can be applied to assay whether a CN change has taken place.

- *Hidden Markov models (HMM):* HMM are the quintessential probability framework for working with sequentially dependent data (Baum and Petrie 1966). In essence, they are a generalization of GMMs for ordered data, where observations are being assigned a label -state- according to a finite number of distributions but the sequence of the labels is not random. The greatest mathematical property of Markóv chains is that, even if they model the probability of entire sequences of observations, they need only be conditioned on the hidden state of the previous observation. This property emerges from some of the core axioms of conditional probability statistics and becomes extremely useful in the computational treatment of data since the overall probability of a sequence can be derived from reading it only once. We develop an HMM based method for CN determination described in Supplementary Figure 1 and in Section 3.1. [Methods].

- *Artificial neural networks (ANN):* ANN bring together the properties of Bayesian networks and Markóv Chains, thus being able to model almost any kind of data, be it sequential, hierarchical or a combination of either (McCulloch and Pitts 1990). Such is the extent of ANNs that they have been named Universal function approximators, as they have been hypothesized to be able to approximate any possible continuous function in real space, way beyond the scope of statistical

discrimination. Furthermore, contrary to canonical HMMs, ANNs can account for possible higher-order interactions in the data -when an observation is conditioned not only by those surrounding it but also by others further away-. With all their potential, neural networks require extensive training sets and can quickly become computationally intensive. Additionally, ANNs might be slightly overcomplicated for the sole purpose of CNV calling from a single NGS feature. Most of the single features explained in this section have been extensively studied and can be parametrized without the need for ANN learning. Instead, ANNs could be more efficient at combining the data from different features to increase CNV calling accuracy. However, ANNs need to be trained on an extensive true set and, to date, very few genomes have been fully and accurately CNV-resolved.

1.2.2.4. Relevance of CNV

Typically, mammalian genomes contain between 50 and 200 Mbp of CNV, which amounts to ~2.5-10% of the genome. In terms of the number of events in extensive sample panels, most studies, especially those before the early 2010s, discarded all CNV smaller than 5-10 kb and would typically discover 200-700 duplications and up to 2,000 deletions in the whole population (Ghosh et al. 2014; Paudel et al. 2013; Bickhart et al. 2012). Extending the CNV definition to >500 bp results in calling almost one order of magnitude more events of both kinds. Lowering the threshold for detection below typical read length (<100 bp) can lead to calling over 100,000 CNV events, but read depth based software is not tuned to work at such low resolutions, and all calls should be refined with

other methods. The rough estimates provided here are probably an underestimation of the real number of CNV loci in a genome. Reference biases have been shown to potentially cause a drop-out of up to ~50 CNV Mb in top quality genome assemblies such as human (Sherman et al. 2019), and private CNV variation is likely to be neglected by default. Still, compared to the total number of SNP within a WGS mammalian sample panel, which rarely exceeds 30 million, that means that the fraction of the genome affected by CNV is easily about 10 times higher than that affected by SNPs.

Generally speaking, exons tend to be depleted in deletions in most humans, and genic deletions tend to appear at lower frequencies (Sudmant, Mallick, et al. 2015). Duplications have been hypothesized to be enriched in genes in a few species such as the human or dog (Sudmant, Mallick, et al. 2015), although that could be a by-product of specific population histories. The conclusions of these analyses should be interpreted cautiously since some gene families such as olfactory receptors, late cornified envelope proteins or immunoglobulin light chains tend to be heavily duplicated in some animals. Differential gene annotation qualities or failure to control for these overrepresented genes could lead to contradictory results. CNVs are generally not randomly distributed across genomes: globally, certain chromosomes tend to harbor unusual amounts SV, as is the case of the human chromosome 19 (Grimwood et al. 2004), the chicken chromosome 15 (Yan et al. 2014), or the mouse chromosomes 6 and 7 (Grimwood et al. 2004; Morgan et al. 2017). Additionally, duplication events tend to cluster together at the moment of their occurrence (Jeffrey A. Bailey et al. 2002).

Nevertheless, most studies agree that CNV calls either tag or intersect different gene and functional element annotations than those typically affected by SNV calls. Particularly, this has been reported in genome-wide association studies (GWAS) and cross-population studies in humans and other domesticated animals (Sudmant, Mallick, et al. 2015; Zhou et al. 2018; Brahmachary et al. 2014). This is not surprising, since most SNV-focused studies tend to exclude complex regions and avoid calling SNPs within and near CNV-like loci (Prado-Martinez et al. 2013; Mallick et al. 2016). Similarly, complex regions have been shown to be heavily underrepresented in assembly alignments (Siepel et al. 2005). Additionally, all the copies of a duplication that are collapsed within a single assembly region are inherently untaggable by SNPs. Although unsurprising, this SNP-CNV orthogonality is extremely relevant, since it might possibly account for a part of the unexplained variation which haunts many SNP-based association studies. Indeed many anatomical traits and diseases have been linked to CNV in a number of species. In dogs, about 10% of the commercially assayed phenotypes correspond to CNV, and a few breeds defining traits are caused by copy number variation.

Although tangent to the scope of this thesis, CNVs are of the utmost importance in the field of cancer genetics. Cancerous cells have for decades been known to accumulate more CNVs than healthy cells as their repair mechanisms shut down, making them an instrumental tool for the study and diagnosis of cancer. Particularly, different types of cancer have been shown to accumulate CNVs differently according to their tissue and

stage of development (Ni et al. 2013; Stephens et al. 2012), suggesting that CNV accumulation might be a good indicator of the initial causes of cancer, which could assist medical personnel in the choice of treatment and prognosis. Of course, the software and analysis techniques to profile cancer CNV are oriented towards comparison with healthy tissues and recognition of isolated events, which often diverges from the goal of evolutionary genomics. However, it must be noted that generally speaking, the cancer CNV discovery tools are more broadly applied and more regularly maintained than the ones used in evolutionary genomics (Zare et al. 2017).

CNV in humans

CNV has been extensively and repeatedly studied in humans. Different consortia such as the 1000 Genomes Project (The 1000 Genomes Project Consortium 2015), the Coriell repository or the Simons Genome Diversity Project (Mallick et al. 2016) have performed CNV calls in large cohorts of samples. CNV has been shown to recapitulate the major geocultural distribution of human populations, although with a worse resolution than SNPs (Sudmant, Mallick, et al. 2015). Some of the main findings of these studies are a better correlation between deletions and SNPs (Figure 5) and a similar load of CNVs between African and non-African populations, which is striking given larger population size and overall higher nucleotide diversity of African populations (Sudmant, Rausch, et al. 2015; Sharp et al. 2005).

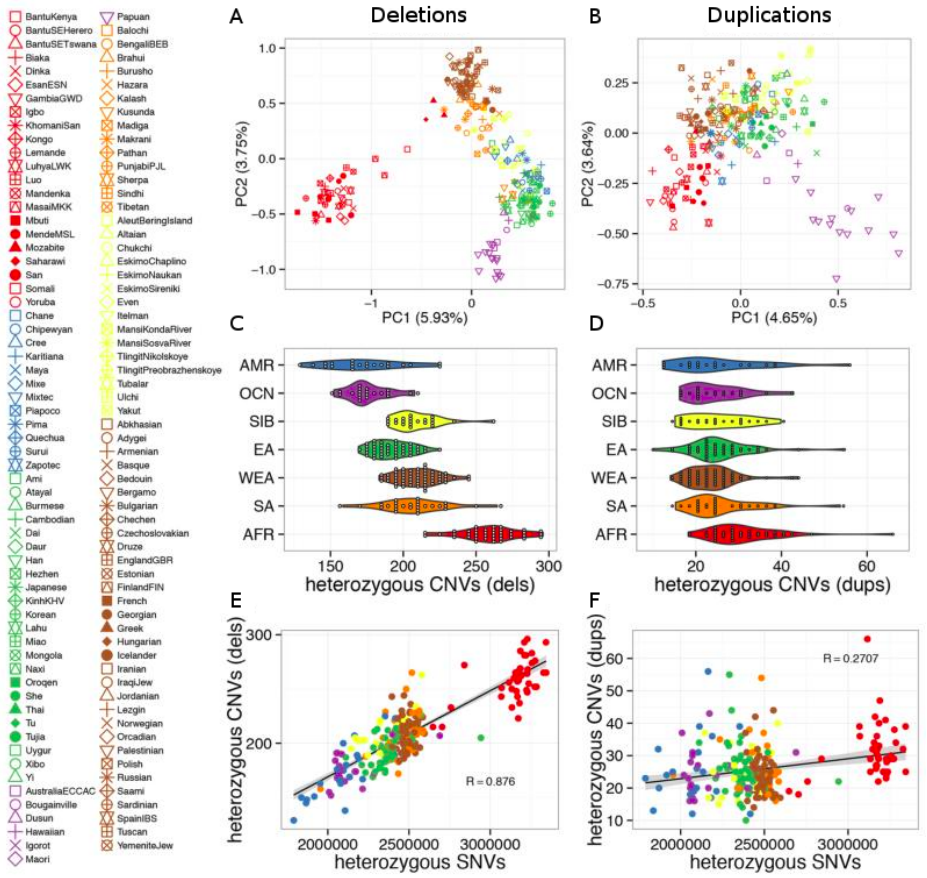

Figure 5: CNV in human populations using deletions (left) and duplications (right). A-B) PCA clustering of human populations. **C-D)** Heterozygosity of biallelic CNV. **E-F)** Correlation between biallelic CNV and SNP heterozygosity. Adapted from (Sudmant et al. 2015).

Additionally, CNVs have been associated with many disease-driving genes in a vast array of disease types, however, it is important to highlight the difference in the number and rate of validation of disease associated CNVs versus SNVs. A quick survey on 1,323 random entries of the Infevers medical database (Touitou et al. 2004), which recopilates data on curated human autoinflammatory disease driving variants, shows that only 9.4% of them are putatively caused by CNVs. Similarly, 43.2%

of the nucleotide changes reported in the database are experimentally validated, while only 19.3% CNVs are. Of course, this estimation is biased and likely to vary between diseases, but it still hints at two consolidated facts: disease-linked CNVs are more often lethal, and therefore less prone to let the patient survive until developing a clinical condition, and CNVs are heavily underrepresented and under-validated in medical studies. Some of the most relevant human diseases caused, at least in some cases, by CNVs are Willams' syndrome, DiGeorge's syndrome, Gaucher disease , primary immunodeficiency and colitis (F. Zhang et al. 2009; Moens et al. 2014; Afzali et al. 2017). Additionally, CNVs have also been associated with neurodegenerative diseases such as Alzheimer's, multiple sclerosis or Parkinson's disease (F. Zhang et al. 2009).

CNV in Great Apes

The evolution of copy number variation in the human lineage needs to be traced back, at least, to the ancestor of orangutans at the base of the great ape tree. Gene expansions and contractions have the potential to account for the pronounced phenotypic differences across human and non-human primates, which can hardly be exclusively explained by the relatively small amount of nucleotide differences. Indeed, the landscape of structural variation in great apes is enriched in lineage-specific CNVs and there is an acceleration in the rate of emergence of novel duplications in the human-chimpanzee ancestor (T. Marques-Bonet and Eichler 2009). Deletions can both reconstruct the great ape phylogeny and correctly recapitulate the subspecies tree of the *Pan*, *Gorilla* and *Pongo* genera

(Sudmant et al. 2013). Interestingly, CNVs in the *PRDM* gene family have been found in the *Pan* genus, which have the potential to alter the recombination map and nucleotide composition of the closest species to humans. A particularly interesting SV-driven trait in great apes is the absence or presence of penile spines (Reno et al. 2013). It seems that the ancestral state in primates was the presence of penile spines, but they have been lost or attenuated in some great ape lineages through different mechanisms: while chimpanzees and bonobos still preserve the trait, humans have lost it through the deletion of an enhancer (McLean et al. 2011) and gorillas may have attenuated the phenotype through an inversion (Kronenberg et al. 2018). Also interesting is the absence of the retrotransposon PtERV1 in humans and orangutans when it is supposed to be basal to all African primates (Gifford et al. 2008).

CNV in domesticated animals

Animals with a huge impact on human economy have usually been subject to CNV analyses. Species like the chicken, horse, or cow have been extensively CNV-profiled over the last two decades (X. Wang and Byers 2014; Ghosh et al. 2014; Keel, Lindholm-Perry, and Snelling 2016). It is interesting to point out that studies involving these species tend to have a narrower scope than those performed in model animals, as they normally aim to uncover the biology of very specific traits of interest instead of performing a global assessment of the species' CNV landscape. As such, most studies preferentially use existing, affordable and standardized technologies such as commercial SNP arrays and aCGH to infer CNV (Gorla et al. 2017; Jia et al. 2013; Doan et al. 2012;

Metzger et al. 2013; Hou et al. 2012; Jiang et al. 2013), at the cost of a potentially lower resolution and a lower genome coverage. Interestingly, some domesticated species such as the pig or the horse do not have a curated, assembly-based, structural variation map in any repository, even if they have had recent assembly updates over the past few years.

CNV in non-domesticated animal species

Non-domesticated animal species that are not used as medical or research models have historically rarely been studied outside the scope of conservation and phylogeny assessment. Since, CNVs are not instrumental for either of those purposes, partial- and whole-genome assembly endeavors for basic nucleotide diversity calculations are usually prioritized to the assessment of other kinds of variation. A quick survey of the UCSC database shows that out of about 120 animal species with public assemblies, only 13 have a curated structural variation annotation track. Of course, that does not mean that CNV analyses cannot, or have not, been performed in those species; a search in variation databases such as dbVar will find particular CNVs for most species. But this highlights the fact that many assemblies are ill-suited for CNV discovery endeavors, and that there is a noticeable lack of global CNV discovery analyses.

However, the advent and convenience of NGS and the development of better assembly strategies have promoted the creation of the Genome 10K initiative (Scientists and Genome 10K Community of Scientists 2009), which is aimed at studying variation in vertebrates. Through this project, a number of high quality mammalian and avian genome

assemblies are being produced which are planned to be used, among other purposes, for CNV assessment. Together with the advent of affordable long-read technologies, this might lead to a more global, high-resolution structural variation panel in a great number of vertebrate species.

1.3. Dogs

Nowadays, dogs play a substantial yet inconspicuous role in human welfare, economy, and society. From the point of view of medicine and genetics, dogs are subjected to similar selective pressures and environments as humans, which makes them the perfect model to study disease and phenotype. The genetic history of artificial selection bottlenecks has provided dogs with a simplified genetic architecture which, additionally, is more receptive to slightly deleterious variants (Cruz, Vilà, and Webster 2008). As such, the analysis of highly complex and polygenic traits in humans such as behavior, intelligence, and body mass is generally reduced to a handful of variants with medium effects in dogs (Plassais et al. 2019; MacLeant et al., n.d.). Dogs have also proven to be great models to naturally assess the phenotypic repercussions of transgenerational gene alterations which are hardly tolerated in other species (Freedman, Lohmueller, and Wayne 2016).

Dog health generates a very large proportion of the veterinary industry revenue, which is estimated to easily gross over 50 billion dollars worldwide yearly and employ hundreds of thousands of people in the US only (American Veterinary Medical Association 01 Jan, 2016). In

addition to that, the industries of dog nutrition and wellbeing are thought to reach similar or even higher figures. Less obvious but still significant are the revenues of dog racing and dog-based entertainment, which recent surveys report to amount to nearly 1 billion pounds in the UK alone (Lange 2019), but could potentially reach much higher figures worldwide.

A few niche business markets involving dogs have been going on for over 200 years: dog breeding and training, although overall much less profitable, sustain the dog show-business and gambling industries, and report huge benefits to particular owners and kennels. A purebred dog litter fitting breed standards with a certified pedigree will surely sell for over 10 thousand US dollars, but there are numerous precedents of its actual values reaching higher orders of magnitude.

With all this, dog genetic testing is quickly becoming a rising business opportunity. About a dozen companies worldwide offer ancestry and disease propensity determination in dogs and generate a net average yearly revenue of 2 to 10 million US dollars each. Although well below the grossing of the veterinary industry, it is interesting to note that their revenues are a close second to those of human ancestry determination where much fewer companies generate a revenue of about 200 million US dollars altogether[4].

[4] All estimates and values were extracted from hoovers ("Company Search | Company Information | Hoovers Company Profiles - D&B Hoovers - Companies & Details - Hoovers.com" n.d.)

Overall, it is very difficult to quantify the impact of dogs in human society. Nowadays, they fulfill a key role of companionship, assistance, and protection, which can hardly be found in any other domesticated species. Dogs are so interwoven in the human lifestyle and routine that their contribution might be considered invaluable by many.

1.3.1. Domestication

The dog is, by far, the oldest domesticated species known to date and has coexisted with humans for the largest part of our societal modern life. It is known that dogs were domesticated from gray wolves, but the actual gray wolf subspecies which gave rise to modern dogs is still undetermined. Even if many gray wolf populations are currently dwindling, gray wolves have been known to inhabit Europe, Asia, America (Fan et al. 2016) during the Pleistocene, and many other genetically similar species can be found in Africa (Gopalakrishnan et al. 2018). However, none of the current extant gray wolf subspecies seem to be sufficiently closely related to dogs to be their closest ancestor, the reason being that most ancestry signatures in the current dog breeds have been overridden by strong artificial selection and inbreeding, and any excess similarity can be better explained by recent introgression.

The matter of dog domestication is so complex that one of the current hypotheses proposes that dogs could have been the descendants of an extinct, Eurasian Pleistocene gray wolf lineage. A few key ancient canine samples which could shed light to this matter have been found recently: a bone and a perfectly preserved head of two different Pleistocene wolves dating back from 30 to 40 thousand years before the present were found

in Siberia in 2015 (Skoglund et al. 2015) and 2019. Additionally, permafrost preserved gray wolf cubs from even further in the past were discovered in Canada in 2016. All these ancient wolf samples could either help pinpoint or belong to the gray wolf population from which dogs descend. Additionally, samples from ancient dogs from 5,000 to 16,000 years in the past have been found in Germany, Siberia, Ireland, and Scandinavia (Botigué et al. 2017; Pitulko and Kasparov 2017). These samples should be free from inbreeding and artificial selection signals, and their similarity to current gray wolf subspecies could be truly informative of the origin of dog domestication. With all these new sources of genetic and archaeological information, the prevalent hypothesis based on genetics that dogs were domesticated between 10 and 15 thousand years ago is being pushed back further into the past (Skoglund et al. 2015). As such, for now, it is safer to say that dogs were domesticated in Eurasia sometime between the last glacial maximum and 40,000 years before the present.

Dog domestication has been a hotly debated topic, not only because of the difficulty to trace back the date and origin place of the event, but also because contrary to the case of farm animals, where the purpose of the domestication event is clearly confined to sustenance or transportation, the intent of dog domestication is yet to be fully unraveled. Clearly dogs have been used for protection, hunting assistance and even transportation or entertainment over different ages and by different civilizations, but the initial intent of the domestication can mostly only be theorized about. Nowadays, the most prevalent hypothesis is that dog domestication might have been a byproduct of the expansion of anatomically modern humans

into Eurasia. This might have forced gray wolves to scavenge for food near human settlements, making them tamer and more receptive to training. Eventually, humans might have benefited from dog presence for hunting and protection, which seem to be the most heritable traits in dog behavior (MacLean et al. 2019).

Figure 6: Pictures of 20th-century and current dog breeds. A) Bulldog. **B)** Basset Hound. **C)** Bull Terrier. **D)** Dachshund. **E)** German Shepherd. **F)** Boxer. **G)** Airedale Terrier. **H)** Pug. **I)** Shetland Sheepdog. **J)** St. Bernard. All images borrowed from (https://www.vintag.es/2019/04/then-and-now-dog-breeds.html).

1.3.2. Dog phylogeny

The phylogenetic history of dogs is as rich and interesting as it might be deceiving. The founder wolf population from which dogs were domesticated has been theorized to be way below the 1,000 individuals, which would mean that all current dog genomic variation stems from at most a few hundred individuals[5] (Niskanen et al. 2013). However, most of the common dog breeds existing in the present are not direct descendants of this first domestication bottleneck (Karlsson and Lindblad-Toh 2008). The only current breeds which are directly related to that bottleneck event are Arctic and Asian breeds -such as the Siberian husky and Chow Chow-, and to a lesser extent, toy Asian breeds -such as the Shih Tzu or the Pekingese- (Freedman et al. 2014). However, all these breeds have experienced introgression events with current dog breeds, which has diluted their ancestral component. Most current common European dog breeds -such German Shepherds, Rottweilers or Retrievers among many others- are descendants of a number of breed-specific, secondary bottlenecks with the intent of trait selection which took place between 100 and 300 years ago (Parker et al. 2017). Some of these secondary domestication events gave rise to more than one current breed, therefore, breeds that share a common secondary domestication origin are said to belong to the same breed clade. These recent, pure-bred genealogies have been maintained up to the present date at the cost of heavy inbreeding.

---

[5] These estimates might change in the light of the most recent ancient sample findings (Section 1.3.1).

Interestingly, there are written records of the date and protocol of creation of most recent European breeds, so there is a true positive set of data to match to genetically reconstructed breed genealogies. Indeed, genetic and historical data recapitulate dog phylogeny considerably better than plain morphological and occupational classifications (Parker et al. 2017) (Figure 7).

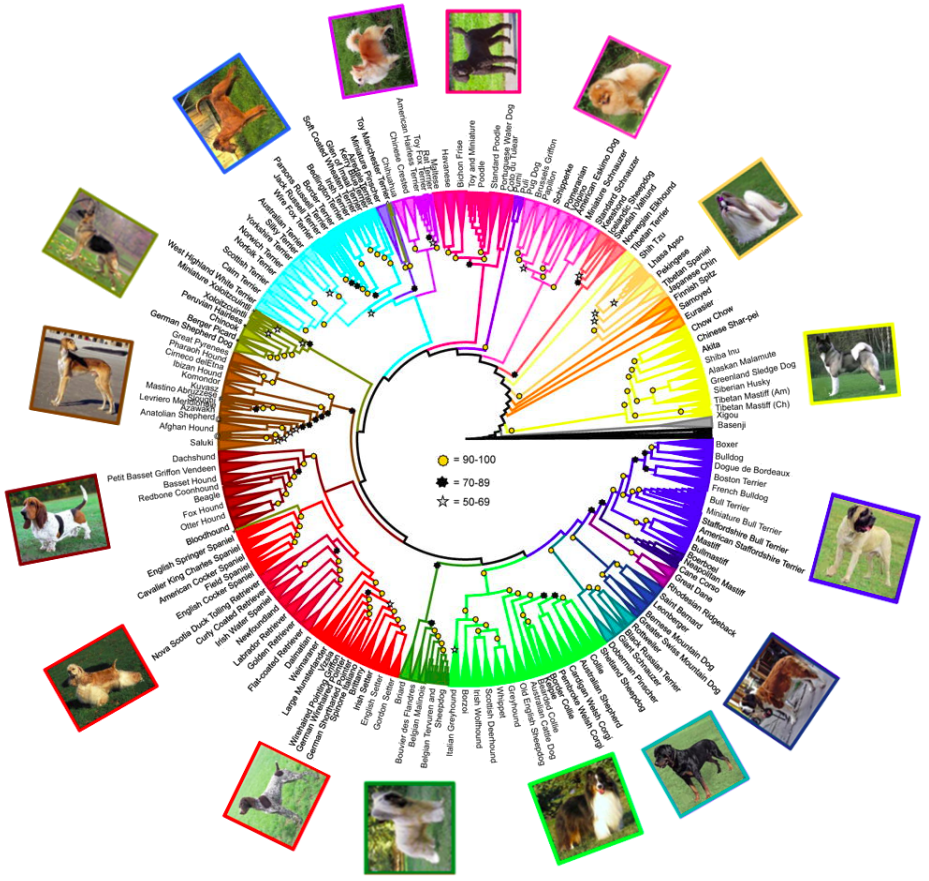

Figure 7: Dog breed phylogeny. Breed clade representatives depicted around the tree. Adapted from (Parker et al. 2017).

This specific process of breed origination has resulted in a highly hierarchical classification of dog genetic differences, where dogs from the same breed are expected to share up to 25% of their total genetic material -which is a similar proportion to that of second-degree relatives- (Parker et al. 2017). Next come breed clades, which share from 2.5 to 5 percent of their DNA. Finally, breeds from different breed clades are expected to share less than 1.5 percent of their genomes. With these numbers in mind, it is easy to see how dogs are a great model for genome-phenome assays since intra-breed comparisons should reveal environmental contribution to trait variance while inter-breed comparisons should reveal the genetic bases of traits.

1.3.3. Dog phenotypes

A consequence of the isogenic background of dog breeds has been the creation of breed stereotypes in dog morphometrics (Jones et al. 2008). Breed stereotypes -or standards- are anatomical measurements such as limb length, ear shape or body size from which purebred dogs rarely deviate. Breed standards are highly sought after in dog shows and pageants, but they have deeper implications than that: dog phenotypes can be accurately inferred based on their breed so long as they are purebred. This has been used in many dog GWAS studies where no phenotypes were measured for the genotyped samples, but instead they were imputed from their breed identity.

Extensive phenotypic studies have been done in dogs over the past few decades by means of GWAS and trait mapping. Different studies have pinpointed the genetic bases of several physical traits such as coat

coloring, ear and tail shape and other morphometrics (Jones et al. 2008; Vaysse et al. 2011; Hayward et al. 2016; Plassais et al. 2019). Many other assays have discovered over 170 nucleotide changes related to dog disease propensity, penetrance and morbidity (Chase et al. 2009; Shearin and Ostrander 2010; Mellersh 2014; Baker et al. 2017; Karlsson and Lindblad-Toh 2008) (Supplementary Table 2). Lately, even 15 key dog behavioral traits have been found to be remarkably heritable and explainable by less than 150 loci (MacLean et al. 2019).

1.3.4. Dog CNV

Copy number variation has been assessed in dogs via multiple methods, most prominently aCGH and SNP arrays (W.-K. Chen et al. 2009; Nicholas et al. 2009, 2011; Berglund et al. 2012), but also through WGS (Serres-Armero et al. 2017; G.-D. Wang et al. 2019). One of the most recognized examples of CNV in canids is the presence or absence of an amylase gene expansion, which is correlated with a better ability to digest and process starch. Wild canids rarely present this gene duplication and the cases where they do have been attributed to dog introgression (Freedman et al. 2014). On the other hand, all dog breeds display this expansion to a bigger or lesser extent, meaning that dog ancestors must have either carried it as standing variation or that it appeared after the first dog domestication bottleneck (Ollivier et al. 2016). Perhaps even more interesting is the correlation between the absolute number of copies of the gene and expected starch consumption, where sled dog breeds living in the arctic, which are rarely fed starch, present fewer copies of the gene -putatively a more ancestral genotype- while dogs living elsewhere have more copies (Arendt et al. 2016). Interestingly, a similar

correlation has been found between low starch and high starch consumption in human populations (Perry et al. 2007). Other than amylase, a few other genes have been found to have different copy number in dogs and wild canids such as *MAGI2* and *PDE4D* (W.-K. Chen et al. 2009; Ramirez et al. 2014).

CNVs have been found to play a role in breed origination and identity. A classical example is the breed-defining formation of a dorsal hair ridge in Rhodesian and Thai Ridgeback dogs, which is caused by a CNV (Salmon Hillbertz et al. 2007). Similarly, a duplication is the main cause of blue eye coloring in Siberian Huskies (Deane-Coe et al. 2018) and CNVs tend to be associated with breed-specific coat features such as saddle tan (Dayna L. Dreger et al. 2013), agouti patterning (D. L. Dreger and Schmutz 2011) or plain hairlessness (Drögemüller et al. 2008). Some genes with a less apparent phenotypic implication have been discovered to be differentiated across dog breeds (Nicholas et al. 2011), and could potentially be implicated in differences in olfaction, growth and cognitive abilities.

2. OBJECTIVES

- Development of a statistical framework to determine copy number differences using read depth data.
- Global assessment of copy number variation in dogs and wolves using whole-genome sequencing.
- Recapitulation and discovery of copy number variation differences across canids.
- Genome-wide association of copy number variants to phenotypes using breed stereotypes.
- Recapitulation of dog breed phylogeny using copy number variation.

3. RESULTS

3.1.

Serres-Armero, Aitor, Inna S. Povolotskaya, Javier Quilez, Oscar Ramirez, Gabriel Santpere, Lukas F. K. Kuderna, Jessica Hernandez-Rodriguez, et al. 2017. Similar Genomic Proportions of Copy Number Variation within Gray Wolves and Modern Dog Breeds Inferred from Whole Genome Sequencing. *BMC Genomics* 18 (1): 977.

Serres-Armero et al. BMC Genomics (2017) 18:977
DOI 10.1186/s12864-017-4318-x

BMC Genomics

RESEARCH ARTICLE Open Access

Similar genomic proportions of copy number variation within gray wolves and modern dog breeds inferred from whole genome sequencing

Aitor Serres-Armero[1†], Inna S. Povolotskaya[1†], Javier Quilez[1,2†], Oscar Ramirez[1,3], Gabriel Santpere[1,4], Lukas F. K. Kuderna[1], Jessica Hernandez-Rodriguez[1], Marcos Fernandez-Callejo[2], Daniel Gomez-Sanchez[1], Adam H. Freedman[5], Zhenxin Fan[6], John Novembre[5], Arcadi Navarro[1,2,7], Adam Boyko[8], Robert Wayne[5], Carles Vilà[9], Belen Lorente-Galdos[1,4*] and Tomas Marques-Bonet[1,2,7*]

Abstract

Background: Whole genome re-sequencing data from dogs and wolves are now commonly used to study how natural and artificial selection have shaped the patterns of genetic diversity. Single nucleotide polymorphisms, microsatellites and variants in mitochondrial DNA have been interrogated for links to specific phenotypes or signals of domestication. However, copy number variation (CNV), despite its increasingly recognized importance as a contributor to phenotypic diversity, has not been extensively explored in canids.

Results: Here, we develop a new accurate probabilistic framework to create fine-scale genomic maps of segmental duplications (SDs), compare patterns of CNV across groups and investigate their role in the evolution of the domestic dog by using information from 34 canine genomes. Our analyses show that duplicated regions are enriched in genes and hence likely possess functional importance. We identify 86 loci with large CNV differences between dogs and wolves, enriched in genes responsible for sensory perception, immune response, metabolic processes, etc. In striking contrast to the observed loss of nucleotide diversity in domestic dogs following the population bottlenecks that occurred during domestication and breed creation, we find a similar proportion of CNV loci in dogs and wolves, suggesting that other dynamics are acting to particularly select for CNVs with potentially functional impacts.

Conclusions: This work is the first comparison of genome wide CNV patterns in domestic and wild canids using whole-genome sequencing data and our findings contribute to study the impact of novel kinds of genetic changes on the evolution of the domestic dog.

Keywords: Copy number variation, Dog genomics, Evolution, Domestication

* Correspondence: belen.lorente@gmail.com; tomas.marques@upf.edu
†Equal contributors
[1]IBE, Institut de Biologia Evolutiva (Universitat Pompeu Fabra/CSIC), Ciencies Experimentals i de la Salut, 08003 Barcelona, Spain
Full list of author information is available at the end of the article

Background

The dog (*Canis familiaris*) was domesticated from the gray wolf (*C. lupus*) [1–4] more than 10,000 years ago, although when and where domestication happened as well as the role of humans in the process have been focus of intense debate [5–10]. Beginning several hundred years ago, modern dog breeds were established as isolated gene pools, in parallel with strong artificial selection for specific physical and behavioral phenotypes favored by humans. A large number of dog breeds have been developed since then, which has resulted in a broad variety of traits and exceptional phenotypic variation [11].

Detecting and understanding the footprint that domestication left in the canine genome is an area of active research. To this end, genetic variation in dogs and wolves has been extensively studied using single nucleotide polymorphisms (SNPs) and microsatellites [12–16]. These studies have shown that nucleotide diversity is between 1.5 and 2 times lower in dogs than in wolves as a result of a 9 to 16-fold reduction in the effective population size associated with dog domestication [4, 17, 18]. Selective breeding further led to reduction in variation, longer linkage disequilibrium (LD) blocks and a lower number of haplotypes among purebred dogs compared to wolves and "village dogs", which have not gone through the breeding process [15, 17, 19–22]. This reduction in diversity is striking in the light of the great phenotypic variation observed in modern dog breeds [12]. Several studies have focused on the identification of functional variants responsible for phenotypic changes associated with domestication [23] or contributing to phenotypic variation of the modern dog breeds [11, 19, 24–29].

Although CNV contributes to phenotypic differences and genetic diseases [28, 30–32], structural variation in multiple canine genomes has not been thoroughly interrogated yet genome-wide. Absolute copy number (CN) values in short genomic windows can be predicted computationally from whole genome sequencing experiments [33–40] and this approach has been used to study CNV patterns in many species. A number of studies have investigated CNV in dogs and wolves using experimental approaches, namely array comparative genomic hybridizations (aCGH) [30, 41–45] and intensity data from SNP genotyping arrays [46]. However, these techniques are limited to relatively low CN regions [47], produce CN values relative to the CN in the reference individual [48], have strong limitation in size of the detectable structural variants [49, 50] and only the parts of the genome in which probes have been placed can be interrogated [47].

In the present study, we aimed to investigate CNV regions in dogs and wolves. However, the analysis of the genome-wide patterns of segregating CNV across a set of individuals is a challenging task and requires precise estimates of the absolute CN of each CNV locus for each of the individual genomes. The accuracy of all the existing methods for absolute CN inference decreases rapidly as CN increases, and thus, nearly all of the studies of CNV diversity up to date are limited to biallelic loci with segregating alleles CN_1 and CN_2 per haplotype [40, 51, 52]. In addition, methods based on read depth only produce point estimates and do not provide confidence intervals, which are extremely important to distinguish between true CN variability and increased technical noise (especially for higher CN values) [53]. This is an important caveat considering that, as reported in humans, population differentiation in loci with a high number of copies might be an important contributor to phenotypic differences [40, 54, 55]. Here, we designed a new probabilistic framework of the read depth based approach for accurate absolute CN inference and CNV detection, which enabled us to perform a comprehensive genome-wide analysis of the patterns and dynamics of CNV loci across the entire range of CNs in a set of 34 canid genomes.

Results

We analyzed a set of 34 sequenced individuals at a mean initial coverage of 16.8X [4, 56, 57]. Our dataset included 12 dogs (*C. familiaris*), 16 gray wolves (*C. lupus*), 2 red wolves (*C. rufus*), 3 coyotes (*C. latrans*) and 1 golden jackal (*C. aureus*) (Table 1) from diverse populations and breeds across Europe, America and Asia [57].

We generated individual genome-wide fine-scale CN profiles using a previously published method [33]. Further, we developed and applied a new probabilistic approach, which allowed us to overcome some of the limitations of the previous methods by estimating probabilities for each CN and broaden the analysis to include loci of high CN.

Validation

We validated our computational predictions with the available aCGH data [43] for 14 of the samples that are common in both studies (Table 1). We compared "digital" log2ratios between the reference individual ("bxr") and each of the other samples included in the aCGH study [43], which showed a high correlation with the aCGH log2-ratios (mean correlation coefficient $R = 0.77 \pm 0.06$, Additional file 1: Table S1). Additionally, $95.4 \pm 3.3\%$ of windows with sample specific CN gains relative to the reference individual (Boxer) have passed the validation threshold (See METHODS and Additional file 1: Table S1). Boxer specific duplications had a lower validation rate ($69.3 \pm 6.8\%$), most likely as a result of sequencing biases specific to this sample (Additional file 1: Figure S1).

Genomic duplications

Duplicated genomic regions spanned 114.05 Mbps (43.44 Mbps in autosomal chromosomes and 70.69 Mbps in unplaced scaffolds) or about 5% of the size of dog autosomes. Dogs have 111.82 Mbps of duplicated

Table 1 Samples and sequencing coverage

Species	Sample	Abbreviation	HMM function	Raw coverage	Effective coverage	aCGH data	Dataset	Diversity analysis
Dog	Chinese indigenous dog	DogCI2	Training	9.83	–	No	Wang et al.	No
Dog	Dingo	din	Analysis	7.09	5.1	No	Freedman et al.	Yes
Dog	Basenji	mba	Analysis	11.8	8.49	Yes	Freedman et al.	Yes
Dog	Kerry Blue Terrier	ali	Analysis	21.28	15.32	No	Fan et al.	Yes
Dog	Boxer	bxr	Analysis	31.27	22.29	No	Fan et al.	Yes
Dog	English cocker	cec	Analysis	11.81	8.5	No	Fan et al.	Yes
Dog	Labrador retriever	dlr	Analysis	12.6	9.07	No	Fan et al.	Yes
Dog	Chinese crest	jcc	Analysis	19.04	13.71	No	Fan et al.	Yes
Dog	Standard poodle	osp	Analysis	12.91	9.29	No	Fan et al.	Yes
Dog	Belgium Malanois	DogBM	Analysis	10.11	7.57	No	Wang et al.	Yes
Dog	German shepherd	DogGS	Analysis	9.56	5.61	No	Wang et al.	Yes
Dog	Tibetan Mastiff	DogTM	Analysis	10.37	5.8	No	Wang et al.	Yes
Gray wolf	Wolf Russia	GW3	Training	11.1	–	No	Wang et al.	No
Gray wolf	Wolf China	chw	Analysis	17.94	12.91	Yes	Freedman et al.	Yes
Gray wolf	Wolf Croatia	crw	Analysis	9.73	6.94	No	Freedman et al.	Yes
Gray wolf	Israeli wolf	isw	Analysis	7.37	5.26	No	Freedman et al.	Yes
Gray wolf	Wolf Great Lakes	glw	Analysis	26.8	19.3	Yes	Fan et al.	Yes
Gray wolf	Wolf India	inw	Analysis	27.42	19.74	Yes	Fan et al.	Yes
Gray wolf	Wolf Iran	irw	Analysis	30.15	21.71	Yes	Fan et al.	Yes
Gray wolf	Wolf Italy	ita	Analysis	7.59	6.07	Yes	Fan et al.	Yes
Gray wolf	Wolf Mexico	mxa	Analysis	25.64	18.46	Yes	Fan et al.	Yes
Gray wolf	Wolf Mexico	mxb	Analysis	7.08	5.66	No	Fan et al.	No
Gray wolf	Wolf Portugal	ptw	Analysis	28.46	20.49	Yes	Fan et al.	Yes
Gray wolf	Wolf Spain	spw	Analysis	28.88	20.79	Yes	Fan et al.	Yes
Gray wolf	Wolf Yellowstone	ysa	Analysis	28.21	20.31	Yes	Fan et al.	Yes
Gray wolf	Wolf Yellowstone	ysb	Analysis	18.82	13.55	Yes	Fan et al.	No
Gray wolf	Wolf Yellowstone	ysc	Analysis	8.44	6.75	Yes	Fan et al.	No
Gray wolf	Wolf China	GW4	Analysis	9.61	6.75	No	Wang et al.	No
Coyote	Coyote California	cac	Training	26.87	19.35	No	Fan et al.	No
Coyote	Coyote Alabama	alc	Analysis	7.69	5.54	No	Fan et al.	No
Coyote	Coyote Midwest	mwc	Analysis	9.11	6.56	No	Fan et al.	No
Jackal	Golden Jackal Kenya	jaa	Analysis	27.47	19.78	Yes	Freedman et al.	No
Red wolf	Red wolf	rwa	Analysis	30.28	21.8	No	Fan et al.	No
Red wolf	Red wolf	rwb	Analysis	7.72	6.17	No	Fan et al.	No

Sequences were retrieved from previously published work from Fan et al. [57], Freedman et al. [4] and Wang et al. [56]. The raw coverage is calculated from the total number of reads before mapping and referred to the 2,413,045,422 bps of the prepared version of CanFam3.1. The effective coverage is calculated after removing poor-quality sequencing lanes and read ends. For 14 samples aCGH data from Ramirez et al. [43] were available. Coyote, jackal and red wolf samples were combined as a single group for the analyses

sequence, gray wolves 111.46 Mbps and related canids 109.74 Mbps. We found, that 79% of the genomic duplications were present in all the individuals (89.72 Mbps in total, 24.03 Mbps in chromosomes and 65.70 Mbps in unplaced scaffolds), 93.04 Mbps (~83%) were present in all the dogs and 95.53 Mbp (~86%) in all the wolves (Fig. 1a). Dogs and gray wolves showed the same average amount of duplicated sequence per individual (104.21 ± 1.89 and 105.54 ± 0.71 Mbps, respectively, Fig. 1b) and 38.22 Mbps were duplicated in at least one individual from each subspecies excluding unassembled scaffolds (Fig. 1c). The average length of duplicated segments did not depend on the sample coverage (Additional file 1: Figure S2).

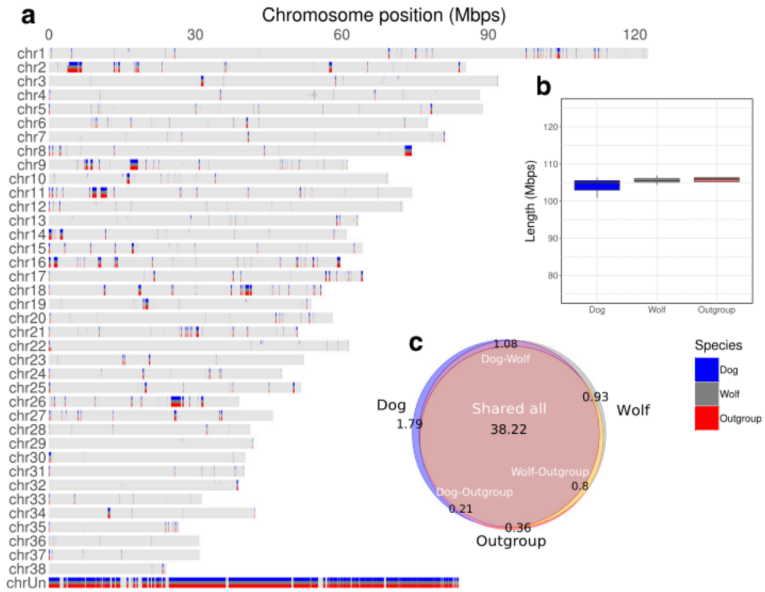

Fig. 1 Landscape of canine segmental duplications. **a** Genome-wide map of canine SDs. Autosomes are represented by horizontal bars, and each mark represent a duplicated region identified in at least one sample of the group indicated. **b** Total length of genomic duplications identified per subspecies. **c** Venn diagram showing intersection of duplicated regions identified in dogs, gray wolves and canines in the outgroup (chrUn excluded)

We found that the set of genomic duplications detected in the 34 canine samples overlapped with 433 genes annotated in the CanFam3.1 dog genome assembly, an overlap significantly higher than the random expectation (randomization p-value = 0.0023) (Additional file 1: Figure S3). Moreover, we found a significant enrichment of duplicated genes involved in detection of chemical stimulus and G-protein coupled receptor signaling pathways, both with p-val < 10^{-30} (Additional file 1: Table S2). These two pathways are closely associated with the perception and transduction of smell and other sensory functions. We also detected a significant enrichment in the pathways of immunoglobulin production and phagocytosis recognition with a p-value of ~10^{-6}. Many essential genes were duplicated in all of our samples, including major cytoskeleton components, a number of ribosomal genes/proteins, mitochondrion maintenance and ubiquitination enzymes or DNA repair mechanisms among many others.

We further looked at the private duplications, present in one subspecies and not in the other. We restricted subsequent analysis to include only 11 dogs and 11 wolves from distinct populations (Table 1, Additional file 1: Table S3), consequently these differences do not result from different sample sizes of dogs and gray wolves (see METHODS). The number of duplications that were unique to dogs (3.67 Mbps or ~3.29% of dog duplications) was substantially greater than for gray wolves (2.19 Mbps or ~1.97% of gray wolf duplications) and they mainly corresponded to events in single individuals (Additional file 1: Figure S4) and none of the private duplications was shared by more than 7 individuals. These private duplications were also significantly enriched in genes for both dogs (randomization p-value = 0.0075) and wolves (randomization p-value = 0.003) with genes involved in iron homeostasis and elastin catabolism overrepresented in dogs, and genes involved in arginine transport overrepresented in wolves (Additional file 1: Table S4).

Genomic proportion of CNV

Our CN calls allowed us to identify windows with segregating CN alleles within populations. We assessed whether the proportion of the genome classified as CNV was reduced in the dog lineage relative to the gray wolf, as has happened for nucleotide diversity (Fig. 2a) reflecting domestication and breed creation bottlenecks. As an overall measure of the fraction of the genome with segregating CNV in either subspecies, we used the number of 1-Kbps windows for which at least two individuals presented non-overlapping CN intervals (further referred to as variable windows specifically or CN variability globally) divided by the total number of 1-Kbp windows called inside duplications, taken as the most likely substrate for CNVs [58, 59]. In striking contrast to the 1.6-fold reduction in single nucleotide diversity in our dataset of dogs (in accordance with estimates of 1.5 to 2-fold reduction reported previously [4, 17, 18], see Additional

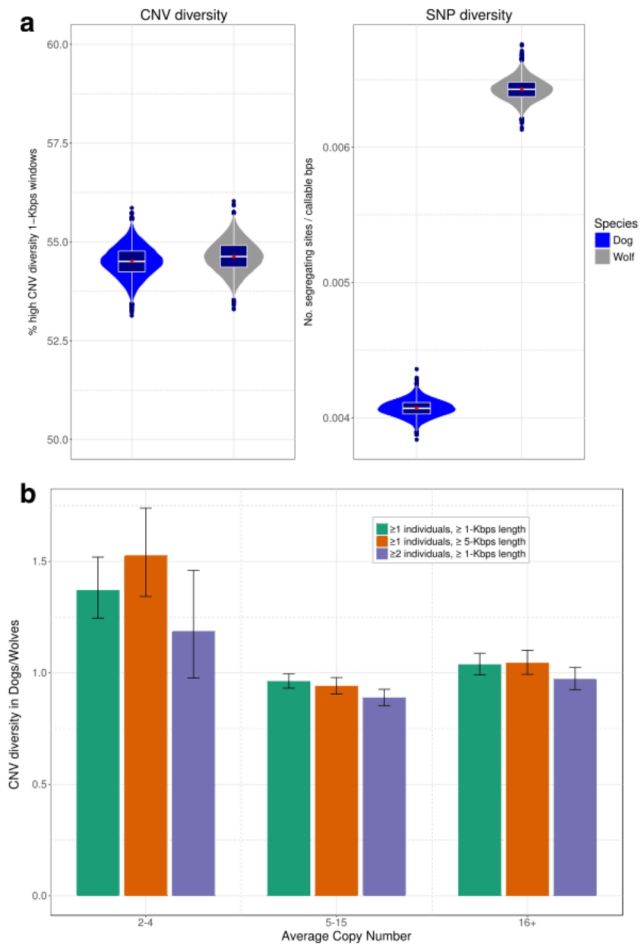

Fig. 2 Proportion of genomic CNV and SNP diversity in dogs and gray wolves. **a** Boxplots indicate the observed values of overall genomic proportion of CNVs and SNP diversity in dogs and gray wolves. Violin plots correspond to 5000 bootstrap values. **b** Ratios of the level of CNVs in dogs to the level of CNVs in wolves for low (2–4), medium (5–15) and high (16+) copy number categories. In green all the regions are taken into account, in orange short regions (less than 5 consecutive windows) of variable CN are filtered and in purple singletons are filtered

file 1: Tables S4 and S5), we found similar proportion of the duplicated genome space with CNVs in the two canine subspecies (54.5% and 54.6% variable windows per total number of duplicated windows in dogs and wolves respectively) (Fig. 2a, Additional file 1: Table S6). Among all variable windows in dogs, 78.8% are also variable in wolves, while for wolves this proportion is slightly higher (80.9%). Most of these regions represent in principle, variability originated before the lineage split, whereas those regions not shared (21.2% and 19.1% respectively), represent subspecies-specific variability, which could potentially contribute to functional differences between the two subspecies (Additional file 1: Figure S5). Alternatively, these regions may represent independent inheritance of CNVs from a common ancestor.

We sought then to eliminate the possibility that artifacts of our CN calling algorithm might influence our estimates. Given the known differences in accuracy of depth of coverage methods for different CN magnitudes, we first divided all genomic duplications into three categories according to their corresponding CN. Specifically, these categories included duplications of low CN (mean CN across all windows in all duplicated individuals between 2 and 4), medium CN (mean CN between 4 and 15) and high CN (mean CN larger than 15). We then calculated variability levels for each of the categories separately (Additional file 1:

Table S6). Surprisingly, the proportion of CNV windows within genomic duplications with low CN (2–4), is even higher in dogs than in wolves (28% and 20% respectively). In this category we assessed the quality of the calls of variable windows for each pair of samples with a two-way aCGH comparison. We required that the absolute value: $\log_2\frac{CN_{Sample1}}{CN_{Sample2}} = aCGH_{Sample1} - aCGH_{Sample2}$ exceeds the cut-off of $aCGH_{CUTOFF} = \pm 3^{*}\sigma_{aCGH(CR)}$ (see METHODS for details) for all the windows with predicted CN differences between the two samples, when one of the samples was not predicted to be duplicated. We thus validated 89% of windows per sample for relative losses and 88% per sample for relative gains (median values, Additional file 1: Table S7).

To further investigate if our measure of CN variability is affected by singletons, we repeated the analysis requiring a minimum of two individuals to be called with a different CN. Even so, dogs and wolves presented similar genomic proportions of CNVs and the value in the low CN category is still slightly higher for dogs (Fig. 2b, Additional file 1: Table S6). Finally, we tested whether the similar levels of genomic variation are not driven by hyper variable duplication breakpoints [60] and are not a result of inaccurate calls of short variable regions. To do so we required for CN regions to be comprised of a minimum of 5 consecutive windows which are identified as variant within the population, and still found overall similar genomic proportion of CNVs comparing dogs and wolves (Fig. 2b, Additional file 1: Table S6).

Variable duplicated genomic segments, defined as 1-Kbps windows for which there were at least two individuals with non-overlapping CN intervals, are enriched in genes in the low and medium CN categories for both lineages (dogs: $p_{CN2-4} = 0.018$ and $p_{CN5-15} = 0.023$; wolves: and $p_{CN2-4} = 0.014$ and $p_{CN5-15} = 0.053$) and many of these genes are involved in both innate immunity (6 genes related to phagocytosis recognition) and adaptive immunity (15 genes involved in immunoglobulin production and MHC maturation). A striking enrichment was found in the pathway of DNA recombination and the most significant signal belonged again to olfactory receptor activity (Additional file 1: Table S8).

We further looked for genes which show a high degree of CN differentiation between the two subspecies based on the V_{ST} statistic. We recover a number of genic CNVs previously reported to be associated with the dog specific phenotypes. Among these genes is the paralogue to the canine alpha-2B-amylase gene (*AMY2B*), which catalyzes the first step in the digestion of dietary starch and glycogen (Fig. 3a and Additional file 1: Table S9). Another case of CN expansion in dogs is a 150-Kbps duplication in chromosome 24 [16, 42]. This duplication spans three members of the signal-regulatory protein (SIRP) gene family, which mediate immune-cell regulation [61] (Fig. 3b and Additional file 1: Table S9). Similarly, the *CBR1* gene (Fig. 3c), coding for a carbonyl reductase enzyme involved in the degradation of both environmental and biologically synthesized quinones, lies within a region duplicated in most samples with some dog samples having a higher number of copies (Additional file 1: Table S9).

Discussion

Inferring absolute CN values from sequencing read depth and determining gains in the number of copies is not trivial. Among computational methods, read depth based approaches are the most accurate [62]. Here, we develop an accurate probabilistic expansion of the sequencing read depth based method to call CN genome wide and use this method to produce fine-scale maps of genomic duplications and CNV regions in dogs, gray wolves and the more basal coyote and golden jackal lineages. The novelty of our approach relies in the population-wide Bayesian probabilistic method to CN estimation, hence allowing us to reliably compare CN values across groups of genomes.

The CNV and duplication maps that we present in this study greatly improve on the landscape of structural variation in the canine genome. We analyzed a set of dogs from different breeds, but we additionally included a wide range of gray wolf samples from a broad geographic distribution and several individuals from other wild canine species. All samples were previously sequenced [4, 56, 57] by next generation sequencing and we utilized a computational read depth approach to estimate fine-scale CN for each individual.

The main objective of our study was to investigate whether the proportion of the genome with CNV regions is reduced in dog compared to gray wolf genomes, which would indicate a reduction of CN polymorphisms similar to the expectations based on SNP diversity and the inferred bottlenecks. To do that, precise estimates of the absolute CN of each CNV locus for each of the individual genomes and a probability associated to them are required. We applied an HMM prediction, local multi-sample re-genotyping and created accurate interval estimates of absolute CNs. We further validated our calling method with an available experimental dataset and found that the accuracy of the method is comparable and in some cases slightly superior to the accuracy of previous methods for copy gain predictions. Accuracy of the calls varies across samples but is not dependent on coverage depth (Additional file 1: Figure S6), as expected since the uncertainties associated with coverage are taken into account by HMM predictions and multi-sample re-genotyping.

Our fine-scale duplication maps indicate that dog genomes present similar genomic proportion of CNV compared to those of gray wolves (Fig. 2a). Nucleotide diversity in dogs compared to their canine ancestors has

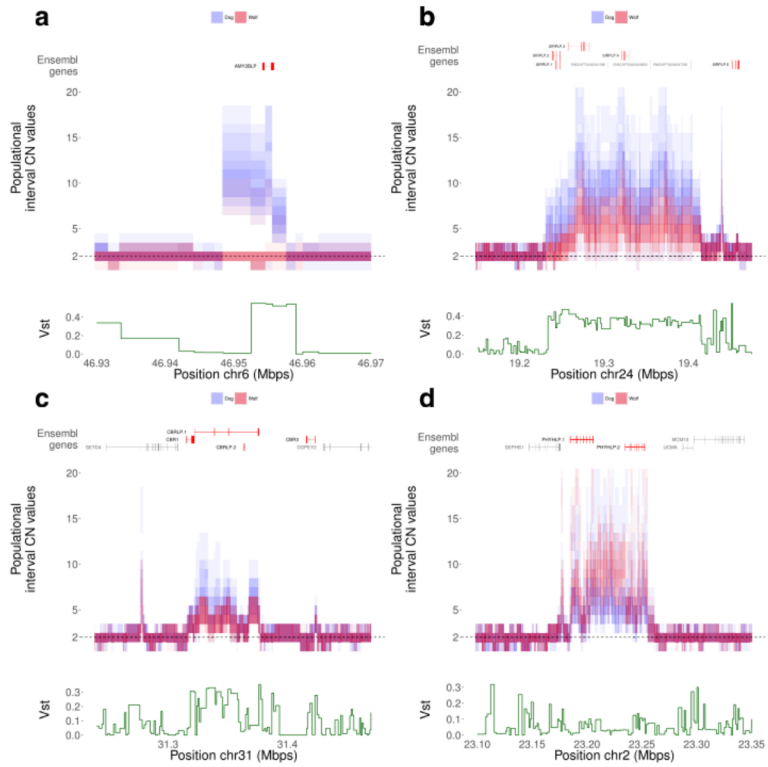

Fig. 3 Genes in high V_{ST} between dogs and wolves. (**a**) AMY2B (**b**) SIRP (**c**) CBR1 (**d**) PHYH

been reduced genome-wide, as reported previously [4, 17, 18] and supported by our data in an extended set of samples (Fig. 2a, Additional file 1: Table S5 and Additional file 1: Table S6). This reduction in nucleotide diversity has been attributed to the population bottlenecks and inbreeding that dogs have suffered as a result of domestication and the creation of dog breeds [13, 20, 63]. With regard to duplications, we observe that ~80% of the CNV sites are mostly shared by both subspecies. It is notable that still ~20% of these genomic CNV regions are not shared, and they might then contribute to the phenotypic plasticity observed in modern dog breeds or represent different sampling of CNV regions from a common ancestor.

We explicitly addressed potential biases that could affect the calculation of the proportion of CNVs. First, the proportions are maintained when considering a high confidence subset of regions, for which at least two individuals are called with a different CN. As for each individual we require the cumulative probability of the CN interval to reach at least 0.99, the probability that two distinct individuals would be called incorrectly is lower than 10^{-4}. Second, accuracy of duplication calls increases with the length of the duplicated region [33] and the same is true with the accuracy for the calls of variability. After exclusion of all short variable segments, the resemblance between CN variation levels in the two subspecies is still maintained and the relative genomic CNV proportion in the low CN category even increases in dogs (Fig. 2b).

Finally, perhaps the greatest challenge in our estimates of the genomic proportion of variable duplications, is the fact that the total length of duplications represented in the dog assembly that are unique to dog samples might be either collapsed or misrepresented. To determine the extent of this problem in our estimates, we used the duplications identified in the genome with the whole-genome assembly comparison (WGAC) [42] to count each duplication detected in each subspecies. After correcting for duplications annotated in CanFam2, we found a slight increase in the proportion of CNVs observed in the dog compared to the wolf, although the final magnitudes were reduced 15–20% in both

subspecies (Additional file 1: Figure S7). It is worth mentioning that our approach is based on counting the proportion of variable windows or equivalently the proportion of total length of variable duplications in the entire duplicated space and therefore it is just an estimation of the actual duplication units. For a more accurate assessment, better resolution of duplication events and breakpoints is required, which could be achieved by whole genome reconstruction based on long read sequencing technologies.

Altogether, similar levels of CNV load in dogs and wolves are extremely unlikely to be explained by an artifact or a bias alone. A key question is then why CN variability is not as reduced in dogs compared to single-nucleotide variation. Below, we considered each of the forces driving mutation-selection-migration equilibrium separately.

There are two scenarios in which selection might increase CNV levels in dogs above those expected given their demographic history. First, the maintenance of relatively high CNV levels in dogs is consistent with diversifying selection among different canine populations if regions of CNV are strongly functional. However, if that is the case, selected functional variants should show high frequency in breeds sharing a trait under selection and be at low frequency or absent in other dogs, resulting in a high overall proportion of genomic CNV. Although, this idea is difficult to test with the current dataset due to the limited number of samples per dog breed, data with aCGH suggest that most of the CNV found in dogs are not shared within breeds but across individuals of different breeds [30]. However, this data does not eliminate across breed variability in high CNs, which would not be detected given the lower dynamic range for such values in aCGH. Alternatively, domestication has relaxed selective pressure on dogs [64] and the consequences of this relaxation can be seen in differences in coding sequence variation [65]. Then, if CNV is generally slightly deleterious, the reduced efficiency of natural selection in small populations during the domestication bottleneck might affect CNVs differently than general SNP diversity especially if the distribution of selective effects is biased toward a greater frequency of neutral or nearly neutral variants in CNVs.

The CNV mutational landscape might also be altered in the canine lineage. Notably, the recombination hotspot gene *PRDM9* gene was pseudogenized in the dog genome. This gene is involved in recombination and novel CNV formation in primate and rodent lineages [66, 67]. Its absence in the dog genome might imply different conditions for CNV formation in the canine lineage. The genomes of closely related domestic cat, panda and ferret all carry a functional copy of *PRDM9*. Interestingly, a region with *RPA3*, one of the genes which binds and stabilizes single-stranded DNA during DNA replication and plays a role in double-strand break repair via homologous recombination, is duplicated in all canid genomes in our study and is variable in dogs. Given 80% of CNVs are shared between the two subspecies, many of them likely originated before the two lineages split, but it could also indicate recurrent duplication events happening at hotspots. However, great uncertainty exists about the overall mutation rate of SNPs [22, 68] and CNV in canines and even less is known about the variation of this rate between dogs and wild canids.

Finally, a reduction in our estimates of CNV relative to SNP diversity also could have been accomplished by reducing the number of genotypes [69, 70] that are segregating in dogs. CNV loci carry on average more alleles than SNP loci, which normally carry just two [71–73]. Although the dynamics of the loss of the number of alleles might be similar between two types of variation, the levels of variability in case of CN will be affected less [35, 73]. The number of alleles per loci is higher for high CN regions [72] and thus, even with a significant reduction of the number of alleles per locus, the level of variability of those high CN loci will not be reduced to the same degree. This effect might underlie the dynamics of our median and high CN categories. Remarkably, duplications with relatively low mean CN are consistently more variable in dogs than in wolves. These low CN duplications are significantly enriched in genes and some have subspecies specific variants, suggesting to a certain extent they might be novel and contribute to functional changes that have occurred after the lineages split.

Regardless of the proposed scenarios, some of the CNV loci with a high degree of variability in dogs or wolves, and specifically gene expansions in CN in the dog lineage, might affect phenotypic differences between subspecies given that 20% are unique to one of these subspecies. A good example is the unique amplification of the amylase gene CN in all dogs, as opposed to the single-copy number in almost all gray wolves, which has been linked to a starch-rich diet in dogs [23] (Fig. 3a). Another example is a highly variable tandem duplication of the *PHYH* gene [42], which in humans is linked to Refsum disease, with multiple epiphyseal dysplasia among variable features [74] (Fig. 3d). In addition, homozygous *PHYH* knockout mice exhibit slightly reduced tibia length [75]. We also detected a remarkable enrichment in the levels of SDs and CNVs in the pathways of immunoglobulin production and phagocytosis recognition, as a CNV region comprising the cluster of *SIRP* genes (Fig. 3b), which are involved in the adaptive immune system [61]. The levels of immunoglobulin A have been shown to vary greatly across dog breeds [76] but, to our knowledge, copy number has never been

studied as a possible cause for this variation. An example of natural and artificial selection acting in opposite directions might be the widespread duplication upstream of the *KITLG* gene, which is linked to the increased risk for squamous cell carcinoma in black standard poodles [77]. *KITLG* locus has been shown to be under strong selective pressure in dogs [19] and a number of other species [78–80]. Interestingly, in humans and stickleback fish this locus is associated with variation in skin pigmentation [81, 82] and therefore possibly also plays a role in coat color and patterning in dogs. Thus, high frequency of this duplication might be explained by artificial selection favoring coat color traits preferred by humans, despite its negative impact on overall fitness.

Conclusions

We present the first genome-wide assessment of CNV landscape in canids based on CN maps generated from high-coverage whole genome sequencing data. The novelty of this study resides in its focus on structural genome variation, which has not been as extensively explored as single-nucleotide variation in canids [4, 17, 19–22]. Additionally, we present a novel method for the application to the whole-genome sequencing read depth data to predict absolute genomic CN under a probabilistic framework. We find that the proportion of genome-wide CNVs in dogs and wolves has been maintained at similar levels in contrast to the decline of nucleotide variation seen in dogs. This result could reflect diversifying selection among dog breeds and populations if CNV are generally functional as with *AMY2B* [43]. The enrichment of genes in CNV regions further supports this assertion. Furthermore, we identify genes with divergent CN variation in dogs and gray wolves, which might have contributed to phenotypic and behavioral differences between the two subspecies. Determining the functional importance of CNV and amount of dog breed specific CNVs should be a focus of future studies.

Methods

Samples and sequencing data

We use sequence data from a panel of 22 canids including 6 dogs, 13 wolves and 3 coyotes sequenced previously [57]. Further, we included the genomes for another 12 canids recently published [4, 56], provided that they had a raw coverage greater than 5X (see below). Altogether, our final dataset comprised 12 dogs, 16 gray wolves, 2 red wolves, 3 coyotes and 1 golden jackal (Table 1) at a mean initial coverage of 16.8X [4, 56, 57]. Each dog sample was from a different so-called modern dog breed with the exception of the Dingo, Basenji and Chinese indigenous dog, which are typically regarded as old lineages. The wolves were sampled from a broad geographic distribution and included a family trio (male, female and offspring) from Yellowstone. For the subsequent analyses we considered the red wolves, coyotes and the jackal samples as a single group (referred to as "outgroup").

Pipeline for calling copy number from sequencing data

We extended a read depth based approach for detection of SDs with a HMM for CN prediction from raw sequencing read depth and incorporated it to the pipeline for calling CN and CNV regions genome-wide (see Fig. 4 for the pipeline overview). To create raw, continuous, genome-wide CN predictions we applied a previously described [33] approach, which consists of the following steps:

(i) Masking of over-represented kmers in the assembly. In addition to the repeats already masked in the UCSC Genome Browser [83] version of CanFam3.1 with RepeatMasker [84] and Tandem Repeat Finder [85], we sought to identify and mask potential hidden repeats in CanFam3.1. In order to do so, chromosomes were partitioned into 36-bps kmers (with adjacent kmers overlapping 5 bps) and the resulting kmers were mapped against CanFam3.1 using mrsFast [86]. Then we masked positions in the assembly mapped by kmers with more than 20 placements in the genome, resulting in 6,910,707 bps additionally masked compared to the original CanFam3.1.

(ii) Mapping 36-bps reads against the assembly. Illumina reads from each individual were split into 36-bps portions (positions 10–45 and 46–81 of the original reads in order to exclude the lower-quality ends of the reads) and mapped to the prepared version of CanFam3.1 using mrFast [86].

(iii) Read depth calculation in 1-Kbps non-overlapping windows of non-repetitive sequence. To avoid edge problems with masked regions, which would underestimate the CN, the 36 bps flanking the masked regions were masked as well (referred to as 36-bps padding onwards). We then calculated the read depth in 1-Kbps non-overlapping windows of non-masked sequence.

(iv) GC-corrected absolute CN estimation from read depth. Read depth values in 1-Kbps windows were then corrected for GC bias using a set of diploid control regions. Control regions were defined as a set of diploid windows totally included in autosomal regions that had not been reported as a CNV in previous studies [4, 41, 42, 87]. These studies were based on CanFam2 and we lifted over the final set of control regions to CanFam3.1. Finally, we removed gaps (plus a 36-bps padding at the start and end of the gap) from the control regions in CanFam3.1. Altogether, this resulted in 21,260 control regions (mean = 94.9 Kbps) with a total size of 2,017,239,131 bps (83.7% of CanFam3.1) and the majority being smaller than 500 Kbps. Of the total 1,151,822 1-Kbps windows 998,077 (86.7%) and 153,745 (13.3%) were control and non-control, respectively.

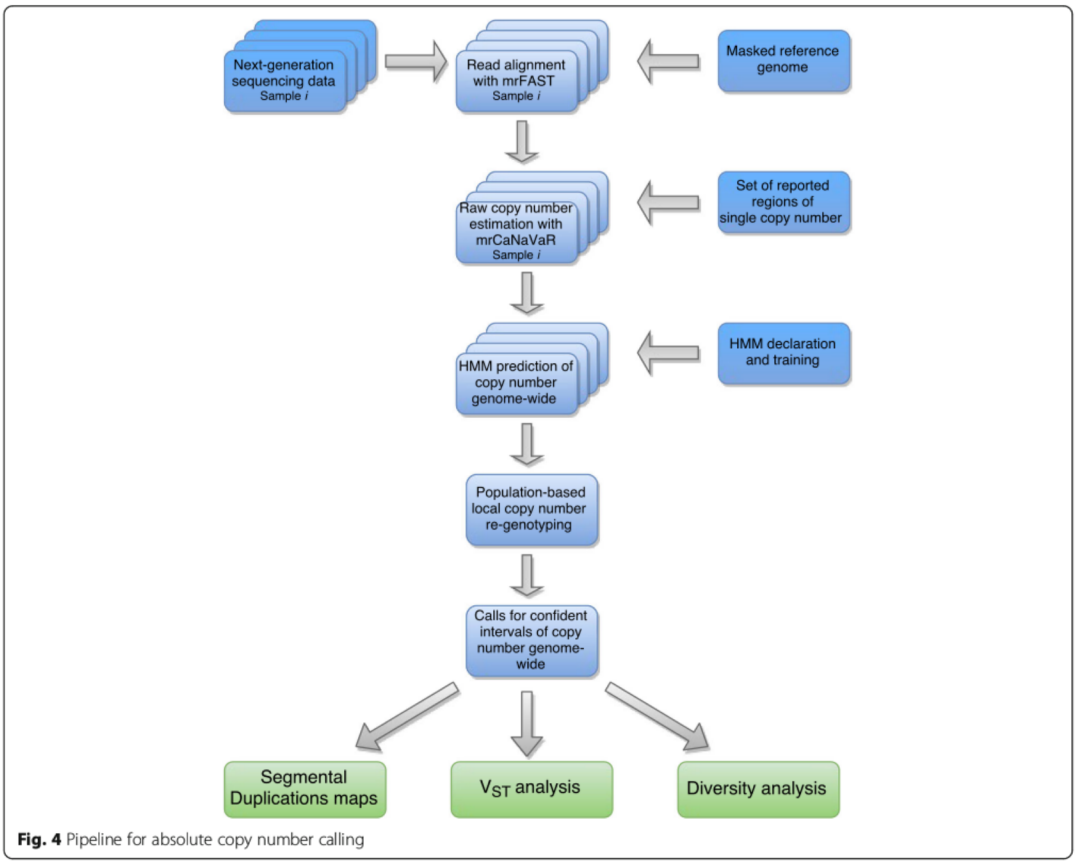

Fig. 4 Pipeline for absolute copy number calling

(v) Raw CN estimation. Finally, we determined CN in the 1-Kbps windows of non-repetitive sequence as the read depth of each window divided the mean read depth in the control regions set.

We noted that this setup is equivalent to a Hidden Markov Process, where hidden states correspond to the true integral CN of the genomic region, emission symbols correspond to CN estimates based on read depth, emission probabilities are drawn from the corresponding read depth distributions and transitions between states correspond to genomic changes of the CN in the adjunct genomic regions. We thus applied this HMM approach to estimate the probabilities of each CN for 1 Kbp windows. We declared HMM states as a set of all possible CN for low CN *(from CN = 0 to CN = 20)* and their corresponding emission distributions as normal distributions with the mean corresponding to the CN, $\mu_N = N$, and standard deviation derived from standard deviation in control regions *(CR, CN = 2)* as $\sigma_N = \sqrt{0.5N} \times \sigma_{CR}$. For high CNs we declared states as an interval of CNs *(CN21–100, CN101–1000)* and modeled their emission distribution as a mixture of normal distributions with weights proportional to the estimated frequencies of each CN. We trained the transition matrix of this HMM with the Baum–Welch algorithm coded in the Pomegranate Python package [88] until convergence. We trained the HMM separately for dogs and wolves, using continuous CN predictions from read depth for one of the samples as observations (see Table 1). We excluded samples used for training from further analysis.

As samples differ greatly in coverage, which in turn leads to differences in standard deviations in the control regions, we redefined the HMM for each individual separately, so the emission distributions would resemble sample specific standard deviation of the read depth in regions of $CN = 2$. The transition matrix is, on the contrary, subspecies specific and does not depend on the sequencing quality. We predict probabilities of each of the declared states at each point by using the forward-reverse algorithm coded in the Pomegranate Python Package. For each individual we thus predict probabilities of each CN at each 1-

Kbp window. Individual read depth based predictions of CN are very noisy and in order to improve them we additionally performed local population-based re-genotyping. For a particular observation of read depth derived raw CN $cn = x$, we use Bayes' theorem to estimate the probability to draw this value from each of the distributions corresponding to CN states:

$$p(CN = N|cn \in [x + dx]) = \frac{p(cn \in [x + dx]|CN = N) \times p(CN = N)}{p(cn \in [x + dx])}$$

$$= \frac{PDF(cn, N, \sqrt{0.5N}\sigma_{CR}) \times dx \times p(N)}{\sum_{CN} PDF(cn, N, \sqrt{0.5N}\sigma_{CR}) \times dx} = \frac{PDF(cn, N, \sqrt{0.5N}\sigma_{CR})}{\sum_{CN} PDF(cn, N, \sqrt{0.5N}\sigma_{CR})} \times p(N);$$

where $p(N)$ is the expected probability to observe $CN = N$ in the data, the only variable which could be tuned locally. For every 1-Kbp window and each possible state of $CN = N$, we calculated its average probability across all the individuals in 5 consecutive windows, centered at the window of interest, and used this mean probability as a prior for the expected probability $p(N)$ of $CN = N$ in the data.

For a fraction of 1-Kbp windows (~2.5% inside duplications, ~51% genome wide) we can call the underlying CN with high confidence $(p > 0.99)$ as a unique integer value. But for complex regions of high CN which are variable across individuals, the probability of each CN is low $(p < 0.99)$. For such windows we consider confidence intervals of the underlying CN. To do so, for each window, we order CN states according to their probability after population based local re-genotyping, and add them to the interval one by one, until their cumulative probability reaches $p = 0.99$ threshold. We further call underlying CN of the window to belong to this interval. We thus could assess if for a particular window any two individuals have the same CN (if we confidently call them with the exact value), different (if we confidently call them to belong to non-overlapping CN intervals), or unresolved (if we call the individuals to belong to overlapping intervals).

We defined duplicated regions as regions of the genome, which harbor at least 5 consecutive windows, which we confidently call as CN > = 3 in at least one of the individual canine genomes. The collection of all such regions we call duplication track, and perform all further analyses only for windows which belong to this track.

aCGH data and validation of the method

For 14 of the samples (1 dog, 1 jackal and 11 gray wolves and 1 red wolf) in which we predicted fine scale confidence CN values, aCGH data assays were available [43] (Table 1). This aCGH chip contains 598,733 probes which target, with a higher density, previously reported regions in the canine genome harboring structural variation [87]. In this study a Boxer sample was used as a reference in the array and we sequenced the same individual in the present study (bxr). Because the aCGH data was based on CanFam2 we generated the 1-Kbps CN predictions based on this version of the dog genome reference assembly and called confidence CN intervals for these 14 samples in the described fashion.

We performed quality control of aCGH experiments by assessing density function of aCGH probes for each individual (Additional file 1: Figure S8). The standard deviation for sample ysc was 2.5 times higher than for the rest of the samples, and we thus excluded ysc from subsequent aCGH validation analysis. We than calculated a threshold to separate true aCGH signals corresponding to gains and losses from diploid noise. To do so, we defined true CN = 2 windows as the intersection between regions which were previously experimentally identified as diploid [4, 41, 42, 87] and the regions which we confidently called as CN = 2 (probability greater than 0.99). As the aCGH chip was designed to target duplications and CNV regions previously reported in the canine genome, genome-wide 1-Kbp windows may be not covered uniformly with aCGH probes or covered at all, so we restricted our analysis only to the windows which harbor at least 2 different aCGH probes. We plotted the distribution of median aCGH signals for Boxer sample in these subset of windows ($n = 1452$), and used a cutoff for aCGH signal $CUTOFF = aCGH_{MEAN}(CN = 2) \pm 3 * aCGH_{SD}(CN = 2) = \pm 0.20$ to discriminate between true gains and losses from false ones.

To validate our calls, we assessed if the difference in the CN which we predict computationally is confirmed by aCGH values. For each individual separately, we detected windows inside SDs, which we computationally predicted to be of a different CN than the reference Boxer sample. This difference could be a duplication compared to Boxer, if the sample CN is predicted to be higher than in Boxer, or a deletion compared to Boxer, if sample CN is lower than Boxer's. We assessed the accuracy in detecting duplications and deletions separately, and calculated it as percentage of windows, which we predict to be CN different from Boxer, which have median aCGH above or below the $CUTOFF = \pm 0.2$ respectively.

Diversity analysis

Our probabilistic method has enabled us to analyze for the first time the fraction of CNV genome-wide and compare it to SNP diversity. To avoid sample sizes biases between dogs ($n = 11$) and gray wolves ($n = 17$), we matched the number of individuals from either subspecies by selecting a subset of 11 gray wolves based on various criteria (Additional file 1: Table S3); the selection of samples also ensured that only one gray wolf from each population was used.

SNP calling and overall SNP diversity

After mapping sequencing reads to the canine genome with BWA [89], we used the CallableLoci tool of GATK [90], with default parameters, to determine areas of the genome that could be considered callable in each of the samples used in the analysis of CNV. We then defined the "callable genome" as the intersection of the callable regions across all the individuals. In addition, we subtracted from the callable genome the X chromosome and mitochondrial DNA, those regions that were masked in the version of the dog genome assembly used here (see above) and 1-Kbps windows with CN exceeding the sample-specific cutoff in at least one sample (Additional file 1: Table S5). After indel realignment we used the UnifiedGenotyper and VariantFiltration tools of GATK [90], with filtering parameters suggested when Variant Quality Score Recalibration (VQSR) is not available [91], to call SNP variants in the total of 11 dogs and 11 gray wolves used in the analysis of the genomic CNV proportion. For this analysis, however, we only retained those variants within the final callable genome (Additional file 1: Table S5). We then split SNPs into those seen in either dog or gray wolf samples and calculated, as a measure of overall SNP diversity, the number of segregating sites in either subspecies divided by the number of bps in the final callable, allowing for zero or two missing alleles (Additional file 1: Table S5). We also calculated the number of segregating sites per bps of callable using the subset of 8 dogs and 8 gray wolves with raw coverage >7X. We observed that the callable genome was greatly reduced by including those samples with a lower raw coverage (Additional file 1: Table S5). We therefore also performed the SNP calling and calculated SNP diversity in the subset of 8 dogs and 8 gray wolves with sequencing raw coverage >7X (see Table 1). We generated bootstrap values for the observed overall SNP diversity as follows: (i) partition the callable genome into intervals of 1 Mbps (I); (ii) random sampling with replacement of I intervals and re-calculated the number of segregating sites divided by the length of the callable genome.

Genomic fraction of CNVs

Within dogs and gray wolves separately, we identify CNV windows as windows for which there are at least two individuals with non-overlapping predicted CN intervals. We measure variability within subspecies as percentage of variable windows among all the windows inside duplicated regions. In either subspecies we obtained an overall measure of CN variability as follows: (i) subset 1-Kbps windows which lie inside duplicated regions of a given subspecies (N); (ii) from those subset 1-Kbp windows which are variable in a given subspecies; (iii) generated bootstrap values by randomly sampling with replacement N windows and re-calculating CN variability, for a total of 5000 times.

To assess the patterns of variable CN across different CN values, we divided all the duplications into the CN bins. To each 1-Kbp window we assign a value, which is average of median points of CN intervals across individuals within subspecies. We further created bins of absolute CNs in such a way, that each bin contains at least 5% of the total number of duplicated windows: low CN (mean CN = 2–4), medium CN (mean CN = 5–15) and high CN (mean CN > 15). We classified all the windows to the bins and assessed the proportion of variable windows in each of them separately for dogs and wolves. To control for the high levels of noise in individual CN predictions we assessed variability for regions comprising at least 5 consecutive variable windows. As a separate control, we excluded singletons from the variability calls and required at least 2 individuals to belong to each of the non-overlapping CN intervals (Fig. 2, Additional file 1: Table S6).

Genes overlapping with genomic duplications and enrichment analysis

We downloaded the 29,884 Ensembl gene models available for CanFam3.1 from the UCSC Genome Browser [83]. Additionally, we considered as of higher confidence those transcripts, 26,748 genes (89.51%), comprising at least one exon present in the xenoRef set of positions syntenic to exons in other species (n = 2,381,071), which was downloaded from the UCSC Genome Browser [83]. These transcripts were converted back to the gene coordinates and only the total of N = 20,328 genes in autosomes were considered for further analysis. For the gene enrichment tests we only selected genes which were entirely covered by duplications. We estimated the gene enrichment associated p-values by the bootstrap. We performed 10,000 repetitions of shuffling duplications coordinates, while keeping their true size and avoiding placing smaller duplications (<100 Kb) on gaps in order to generate an empirical distribution of the expected overlap between genes and SDs. The empirical p-value of the true observed value was calculated by dividing the rank of the true observation by the total number of permutations. The enrichment analysis was performed using the elimination algorithm of the TopGO R package [92], which scores GO terms hierarchically and subtracts specific, significant terms from the more global ones to avoid an overrepresentation of the latter. This conditions the results of the recursive tests on the topology of the gene ontology tree and reduces the effect of multiple testing to a level where no further conventional correction is required [93]. Instead, we refined our result set with the browser tool REVIGO [94], which implements semantic search algorithms in order to merge closely related GO terms and extract the most significant relations between them.

Analysis of CNV differentiation between dogs and gray wolves

In every 1-Kbps window we used CN predictions in dog and gray wolf samples to calculate the V_{ST} statistic [51] between the two subspecies. The V_{ST} statistic is a variation of the F_{ST} [95] to measure between-populations differentiation in CNV regions:$(V_T - V_S)/V_T$ where V_T is the variance in the CN midpoints of all subspecies together, and V_S is the weighted average of the variance in CN midpoints for each subspecies separately. For consistency with the analysis of CNV and SNP diversity we calculated V_{ST} values between the same subsets of 11 dogs and 11 gray wolves (Additional file 1: Table S3). We looked for genes with median $V_{ST} > 0.15$ between dogs and wolves, which corresponds to the windows with the top 10% of V_{ST} values. We focused on the genes with more than 3 copies in dogs while less than 3 copies in wolves (Additional file 1: Table S9).

Additional file

Additional file 1: Collection of all supplementary figures and tables. (DOCX 721 kb)

Acknowledgements
We thank Dorina Twigg, Stefan Sirakov, and Jeffrey M. Kidd for their valuable contribution to the analysis and processing of our data.

Funding
TMB is supported by MINECO BFU2014-55090-P (FEDER), U01 MH106874 grant, Howard Hughes International Early Career, Obra Social "La Caixa" and Secretaria d'Universitats i Recerca del Departament d'Economia i Coneixement de la Generalitat de Catalunya. ISP is supported with a Juan de la Cierva - Formación FJCI-2015-24,275 fellowship. JHR is supported by the Spanish Ministry of Education under FPI grant (BES-2013-064333). BLG is supported with a Beatriu de Pinós (BP-DGR 2014) fellowship.

Availability of data and materials
The WGS datasets analysed during the current study are available in the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) database under the accessions SRA068869 (Wang et al. [56]), SRP044399 (Fan et al. [57]) and PRJNA274504 (Freedman et al. [4]) [https://www.ncbi.nlm.nih.gov/sra/?term=SRA068869, https://www.ncbi.nlm.nih.gov/sra/?term=SRP044399, https://www.ncbi.nlm.nih.gov/sra/?term=PRJNA274504].

Authors' contributions
TMB, BLG, ISP designed the study, analyses and the method; ASA, ISP, JQ performed most of the analyses; MFC, DGS performed variant calling and SNP diversity analysis; OR and JHR performed the experimental analyses; BLG, LFKK contributed to the analyses; OR, GS, AN contributed to the design of the analyses; ISP, JQ, ASA, TMB, BLG wrote the manuscript; AHF, ZF, JN, AB, CV, RW collected samples; all authors read and approved the manuscript.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no conflict of interest. The aCGH dataset analysed during the current study is available in the Gene Expression Omnibus database under the accession GSE58195 (Ramirez et al. [43]) [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE58195].

Publisher's Note

Author details
[1]IBE, Institut de Biologia Evolutiva (Universitat Pompeu Fabra/CSIC), Ciencies Experimentals i de la Salut, 08003 Barcelona, Spain. [2]CNAG-CRG, Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology (BIST), Barcelona, Spain. [3]Vetgenomics, 08193 Barcelona, Spain. [4]Department of Neuroscience, Yale School of Medicine, New Haven, CT, USA. [5]UCLA, Department of Ecology and Evolutionary Biology, Los Angeles, CA 90095, USA. [6]Key Laboratory of Bioresources and Ecoenvironment (Ministry of Education), College of Life Sciences, Sichuan University, Chengdu 610064, People's Republic of China. [7]Institucio Catalana de Recerca i Estudis Avançats (ICREA), 08010 Barcelona, Catalonia, Spain. [8]Cornell University, Department of Biological Statistics and Computational Biology, New York, NY 14853, USA. [9]Estación Biológica de Doñana EBD-CSIC, Department of Integrative Ecology, 41092 Sevilla, Spain.

Received: 10 April 2017 Accepted: 17 November 2017
Published online: 19 December 2017

References
1. Vilà C, Savolainen P E. Maldonado J R. Amorim I E. Rice J L. Honeycutt R, et al. Multiple and Ancient Origins of the Domestic Dog. 1997. Science. doi:10.1126/science.276.5319.1687.
2. Lindblad-Toh K. Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature. 2005;438(7069):803–19.
3. Germonpré M, Sablin MV, Stevens RE, Hedges REM, Hofreiter M, Stiller M, et al. Fossil dogs and wolves from Palaeolithic sites in Belgium, the Ukraine and Russia: osteometry, ancient DNA and stable isotopes. J Archaeol Sci. 2009;36:473–90.
4. Freedman AH, Gronau I, Schweizer RM, Ortega-Del Vecchyo D, Han EE, Silva PM, et al. Genome sequencing highlights the dynamic early history of dogs. PLoS Genet. 2014;10:e1004016. doi:10.1371/journal.pgen.1004016.
5. Boyko AR. The domestic dog: man's best friend in the genomic era. Genome Biol. 2011;12:216. doi:10.1186/gb-2011-12-2-216.
6. Larson G, Bradley DG. How much is that in dog years? The advent of canine population genomics. PLoS Genet. 2014;10:e1004093. doi:10.1371/journal.pgen.1004093.
7. Skoglund P. Estimation of population divergence times from non-overlapping genomic sequences: examples from dogs and wolves. Mol Biol Evol. 2011;28(4):1505–17.
8. Thalmann O, Shapiro B, Cui P, Schuenemann VJ, Sawyer SK, Greenfield DL, et al. Complete mitochondrial genomes of ancient Canids suggest a European origin of domestic dogs. Science (80-). 2013;2013:342. http://science.sciencemag.org/content/342/6160/871.full. Accessed 14 Sep 2016
9. Frantz LAF, Mullin VE, Pionnier-Capitan M, Lebrasseur O, Ollivier M, Perri A, et al. Genomic and archaeological evidence suggest a dual origin of domestic dogs. Science (80-). 2016;2016:352.
10. Wang L, Ma Y-P, Zhou Q-J, Zhang Y-P, Savolaimen P, Wang G-D. The geographical distribution of grey wolves (Canis Lupus) in China: a systematic review. Zool Res. 2016;37:315–26. 10.13918/j.issn.2095-8137.2016.6.315.
11. Vaysse A, Ratnakumar A, Derrien T, Axelsson E, Rosengren Pielberg G, Sigurdsson S, et al. Identification of genomic regions associated with phenotypic variation between dog breeds using selection mapping. PLoS Genet. 2011;7:e1002316. doi:10.1371/journal.pgen.1002316.
12. Irion DN, Schaffer AL, Famula TR, Eggleston ML, Hughes SS, Pedersen NC. Analysis of genetic variation in 28 dog breed populations with 100 microsatellite markers. J Hered. 2003;94:81–7. doi:10.1093/JHERED/ESG004.
13. Ostrander EA, Wayne RK. The canine genome. Genome Res. 2005;15:1706–16. doi:10.1101/gr.3736605.
14. Gundry RL, Allard MW, Moretti TR, Honeycutt RL, Wilson MR, Monson KL, et al. Mitochondrial DNA analysis of the domestic dog: control region variation within and among breeds. J Forensic Sci. 2007;52:562–72. doi:10.1111/j.1556-4029.2007.00425.x.
15. Shannon LM, Boyko RH, Castelhano M, Corey E, Hayward JJ, McLean C, et al. Genetic structure in village dogs reveals a central Asian

domestication origin. Proc Natl Acad Sci U S A. 2015;112:13639–44. doi:10.1073/pnas.1516215112.
16. Decker B, Davis BW, Rimbault M, Long AH, Karlins E, Jagannathan V, et al. Comparison against 186 canid whole-genome sequences reveals survival strategies of an ancient clonally transmissible canine tumor. Genome Res. 2015;25:1646–55. doi:10.1101/gr.190314.115.
17. Gray MM, Granka JM, Bustamante CD, Sutter NB, Boyko AR, Zhu L, et al. Linkage disequilibrium and demographic history of wild and domestic Canids. Genetics. 2009;181:1493–505. doi:10.1534/genetics.108.098830.
18. Pang J-F, Kluetsch C, Zou X-J, Zhang A-B, Luo L-Y, Angleby H, et al. mtDNA data indicate a single origin for dogs south of Yangtze River, less than 16,300 years ago, from numerous wolves. Mol Biol Evol. 2009;26:2849–64. doi:10.1093/molbev/msp195.
19. Boyko AR, Quignon P, Li L, Schoenebeck JJ, Degenhardt JD, Lohmueller KE, et al. A simple genetic architecture underlies morphological variation in dogs. PLoS Biol. 2010;8:e1000451. doi:10.1371/journal.pbio.1000451.
20. VonHoldt BM, Pollinger JP, Lohmueller KE, Han E, Parker HG, Quignon P, et al. Genome-wide SNP and haplotype analyses reveal a rich history underlying dog domestication. Nature. 2010;464:898–902. doi:10.1038/nature08837.
21. Karlsson EK, Baranowska I, Wade CM, Salmon Hillbertz NHC, Zody MC, Anderson N, et al. Efficient mapping of mendelian traits in dogs through genome-wide association. Nat Genet. 2007;39:1321–8. doi:10.1038/ng.2007.10.
22. Auton A, Rui Li Y, Kidd J, Oliveira K, Nadel J, Holloway JK, et al. Genetic recombination is targeted towards gene promoter regions in dogs. PLoS Genet. 2013;9:1–2.
23. Axelsson E, Ratnakumar A, Arendt M-L, Maqbool K, Webster MT, Perloski M, et al. The genomic signature of dog domestication reveals adaptation to a starch-rich diet. Nature. 2013;495:360–4. doi:10.1038/nature11837.
24. Bannasch D, Young A, Myers J, Truvé K, Dickinson P, Gregg J, et al. Localization of canine Brachycephaly using an across breed mapping approach. PLoS One. 2010;5:e9632. doi:10.1371/journal.pone.0009632.
25. Cadieu E, Neff MW, Quignon P, Walsh K, Chase K, Parker HG, et al. Coat variation in the domestic dog is governed by variants in three genes. Science. 2009;326:150–3. doi:10.1126/science.1177808.
26. Olsson M, Meadows JRS, Truvé K, Rosengren Pielberg G, Puppo F, Mauceli E, et al. A novel unstable duplication upstream of HAS2 predisposes to a breed-defining skin phenotype and a periodic fever syndrome in Chinese Shar-Pei dogs. PLoS Genet. 2011;7:e1001332. doi:10.1371/journal.pgen.1001332.
27. Quilez J, Short AD, Martínez V, Kennedy LJ, Ollier W, Sanchez A, et al. A selective sweep of >8 Mb on chromosome 26 in the boxer genome. BMC Genomics. 2011;12:339. doi:10.1186/1471-2164-12-339.
28. Salmon Hillbertz NHC, Isaksson M, Karlsson EK, Hellmén E, Pielberg GR, Savolainen P, et al. Duplication of FGF3, FGF4, FGF19 and ORAOV1 causes hair ridge and predisposition to dermoid sinus in ridgeback dogs. Nat Genet. 2007;39:1318–20. doi:10.1038/ng.2007.4.
29. Schoenebeck JJ, Hutchinson SA, Byers A, Beale HC, Carrington B, Faden DL, et al. Variation of BMP3 contributes to dog breed skull diversity. PLoS Genet. 2012;8:e1002849. doi:10.1371/journal.pgen.1002849.
30. Berglund J, Nevalainen EM, Molin A-M, Perloski M, André C, Zody MCM, et al. Novel origins of copy number variation in the dog genome. Genome Biol. 2012;13:R73. doi:10.1186/gb-2012-13-8-r73.
31. Coe BP, Witherspoon K, Rosenfeld JA, van Bon BWM, Vulto-van Silfhout AT, Bosco P, et al. Refining analyses of copy number variation identifies specific genes associated with developmental delay. Nat Genet. 2014;46:1063–71. doi:10.1038/ng.3092.
32. Fanciulli M, Norsworthy PJ, Petretto E, Dong R, Harper L, Kamesh L, et al. FCGR3B copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. Nat Genet. 2007;39:721–3. doi:10.1038/ng2046.
33. Alkan C, Kidd JM, Marques-Bonet T, Aksay G, Antonacci F, Hormozdiari F, et al. Personalized copy number and segmental duplication maps using next-generation sequencing. Nat Genet. 2009;41:1061–7. doi:10.1038/ng.437.
34. Ghosh S, Qu Z, Das PJ, Fang E, Juras R, Cothran EG, et al. Copy number variation in the horse genome. PLoS Genet. 2014;10:e1004712. doi:10.1371/journal.pgen.1004712.
35. Chain FJJ, Feulner PGD, Panchal M, Eizaguirre C, Samonte IE, Kalbe M, et al. Extensive copy-number variation of young genes across stickleback populations. PLoS Genet. 2014;10:e1004830. doi:10.1371/journal.pgen.1004830.
36. Yi G, Qu L, Liu J, Yan Y, Xu G, Yang N. Genome-wide patterns of copy number variation in the diversified chicken genomes using next-generation sequencing. BMC Genomics. 2014;15:962. doi:10.1186/1471-2164-15-962.
37. Jiang J, Wang J, Wang H, Zhang Y, Kang H, Feng X, et al. Global copy number analyses by next generation sequencing provide insight into pig genome variation. BMC Genomics. 2014;15:593. doi:10.1186/1471-2164-15-593.
38. Sudmant PH, Kitzman JO, Antonacci F, Alkan C, Malig M, Tsalenko A, et al. Diversity of human copy number variation and multicopy genes. Science (80-). 2010;330:641–6. doi:10.1126/science.1197005.
39. Sudmant PH, Huddleston J, Catacchio CR, Malig M, Hillier LW, Baker C, et al. Evolution and diversity of copy number variation in the great ape lineage. Genome Res. 2013;23:1373–82. doi:10.1101/gr.158543.113.
40. Sudmant PH, Mallick S, Nelson BJ, Hormozdiari F, Krumm N, Huddleston J, et al. Global diversity, population stratification, and selection of human copy-number variation. Science (80-). 2015;349: aab3761. doi:10.1126/science.aab3761.
41. Chen W-K, Swartz JD, Rush LJ, Alvarez CE. Mapping DNA structural variation in dogs. Genome Res. 2008;19:500–9. doi:10.1101/gr.083741.108.
42. Nicholas TJ, Cheng Z, Ventura M, Mealey K, Eichler EE, Akey JM. The genomic architecture of segmental duplications and associated copy number variants in dogs. Genome Res. 2009;19:491–9. doi:10.1101/gr.084715.108.
43. Ramirez O, Olalde I, Berglund J, Lorente-Galdos B, Hernandez-Rodriguez J, Quilez J, et al. Analysis of structural diversity in wolf-like canids reveals post-domestication variants. BMC Genomics. 2014;15: 465. doi:10.1186/1471-2164-15-465.
44. Poorman K, Borst L, Moroff S, Roy S, Labelle P, Motsinger-Reif A, et al. Comparative cytogenetic characterization of primary canine melanocytic lesions using array CGH and fluorescence in situ hybridization. Chromosom Res. 2015;23:171–86. doi:10.1007/s10577-014-9444-6.
45. Rossi E, Radi O, De Lorenzi L, Vetro A, Groppetti D, Bigliardi E, et al. Sox9 duplications are a relevant cause of Sry-negative XX sex reversal dogs. PLoS One. 2014;9:e101244. doi:10.1371/journal.pone.0101244.
46. Molin A-M, Berglund J, Webster MT, Lindblad-Toh K. Genome-wide copy number variant discovery in dogs using the CanineHD genotyping array. BMC Genomics. 2014;15:210. doi:10.1186/1471-2164-15-210.
47. Coe BP, Ylstra B, Carvalho B, Meijer GA, MacAulay C, Lam WL. Resolving the resolution of array CGH. Genomics. 2007;89:647–53.
48. Sharp AJ, Itsara A, Cheng Z, Alkan C, Schwartz S, Eichler EE. Optimal design of oligonucleotide microarrays for measurement of DNA copy-number. Hum Mol Genet. 2007;16:2770–9. doi:10.1093/hmg/ddm234.
49. Cooper GM, Zerr T, Kidd JM, Eichler EE, Nickerson DA. Systematic assessment of copy number variant detection via genome-wide SNP genotyping. Nat Genet. 2008;40:1199–203. doi:10.1038/ng.236.
50. Winchester L, Yau C, Ragoussis J. Comparing CNV detection methods for SNP arrays. Brief Funct Genomic Proteomic. 2009;8:353–66. doi:10.1093/bfgp/elp017.
51. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, et al. Global variation in copy number in the human genome. Nature. 2006;444:444–54. doi:10.1038/nature05329.
52. Lou H, Li S, Yang Y, Kang L, Zhang X, Jin W, et al. A map of copy number variations in Chinese populations. PLoS One. 2011;6:e27341. doi:10.1371/journal.pone.0027341.
53. Zhao M, Wang Q, Wang Q, Jia P, Zhao Z. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. BMC Bioinformatics. 2013;14(Suppl 11):S1. doi:10.1186/1471-2105-14-S11-S1.
54. Girirajan S, Campbell CD, Eichler EE. Human copy number variation and complex genetic disease. Annu Rev Genet. 2011;45:203–26. doi:10.1146/annurev-genet-102209-163544.
55. Yang Y, Chung EK, YL W, Savelli SL, Nagaraja HN, Zhou B, et al. Gene copy-number variation and associated polymorphisms of complement component C4 in human systemic lupus erythematosus (SLE): low copy number is a risk factor for and high copy number is a protective factor against SLE susceptibility in European America. Am J Hum Genet. 2007;80:1037–54.
56. Wang G, Zhai W, Yang H, Fan R, Cao X, Zhong L, et al. The genomics of selection in dogs and the parallel evolution between dogs and humans. Nat Commun. 2013;4:1860. doi:10.1038/ncomms2814.
57. Fan Z, Silva P, Gronau I, Wang S, Armero AS, Schweizer RM, et al. Worldwide patterns of genomic variation and admixture in gray wolves. Genome Res. 2016;26:163–73. doi:10.1101/gr.197517.115.
58. Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, et al. Mapping and sequencing of structural variation from eight human genomes. Nature. 2008;453:56–64. doi:10.1038/nature06862.
59. Armengol L, Villatoro S, González JR, Pantano L, García-Aragonés M, Rabionet R, et al. Identification of copy number variants defining

genomic differences among major human groups. PLoS One. 2009;4: e7230. doi:10.1371/journal.pone.0007230.
60. Marques-Bonet T, Kidd JM, Ventura M, Graves TA, Cheng Z, Hillier LW, et al. A burst of segmental duplications in the genome of the African great ape ancestor. Nature. 2009;457:877–81. doi:10.1038/nature07744.
61. Barclay AN, Brown MH. The SIRP family of receptors and immune regulation. Nat Rev Immunol. 2006;6:457–64. doi:10.1038/nri1859.
62. Pirooznia M, Goes FS, Zandi PP. Whole-genome CNV analysis: advances in computational approaches. Front Genet. 2015;6:138. doi:10.3389/fgene.2015.00138.
63. Larson G, Karlsson EK, Perri A, Webster MT, Ho SYW, Peters J, et al. Rethinking dog domestication by integrating genetics, archeology, and biogeography. Proc Natl Acad Sci U S A. 2012;109:8878–83. doi:10.1073/pnas.1203005109.
64. Cruz F, Vila C, Webster MT. The legacy of domestication: accumulation of deleterious mutations in the dog genome. Mol Biol Evol. 2008;25:2331–6. doi:10.1093/molbev/msn177.
65. Marsden CD, Ortega-Del Vecchyo D, O'Brien DP, Taylor JF, Ramirez O, Vilà C, et al. Bottlenecks and selective sweeps during domestication have increased deleterious genetic variation in dogs. Proc Natl Acad Sci U S A. 2016;113:152–7. doi:10.1073/pnas.1512501113.
66. Carvalho CMB, Lupski JR. Mechanisms underlying structural variant formation in genomic disorders. Nat Rev Genet. 2016;17:224–38. doi:10.1038/nrg.2015.25.
67. Miller DE, Hawley RS. Tetrad analysis in the mouse. Nat Genet. 2014;46: 1045–6. doi:10.1038/ng.3104.
68. Freedman AH, Schweizer RM, Ortega-Del Vecchyo D, Han E, Davis BW, Gronau I, et al. Demographically-based evaluation of genomic regions under selection in domestic dogs. PLoS Genet. 2016;12:e1005851. doi:10.1371/journal.pgen.1005851.
69. Allendorf FW. Genetic drift and the loss of alleles versus heterozygosity. Zoo Biol. 1986;5:181–90. doi:10.1002/zoo.1430050212.
70. Maruyama T, Fuerstt PA. POPULATION BOTTLENECKS AND NONEQUILIBRIUM MODELS IN POPULATION GENETICS. 11. NUMBER O F ALLELES IN a SMALL POPULATION T H a T WAS FORMED BY a RECENT BOTTLENECK HE fate of genes in a population that experiences a sudden reduction in. Genetics. 1985;111:675–89. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1202664/pdf/675.pdf. Accessed 24 Mar 2017.
71. Hodgkinson A, Eyre-Walker A. Human triallelic sites: evidence for a new mutational mechanism? Genetics. 2010;184:233–41. doi:10.1534/genetics.109.110510.
72. Handsaker RE, Van Doren V, Berman JR, Genovese G, Kashin S, Boettger LM, et al. Large multiallelic copy number variations in humans. Nat Genet. 2015; 47:296–303. doi:10.1038/ng.3200.
73. Duvaux L, Geissmann Q, Gharbi K, Zhou J-J, Ferrari J, Smadja CM, et al. Dynamics of copy number variation in host races of the pea aphid. Mol Biol Evol. 2015;32:63–80.
74. Skjeldal OH, Stokke O, Refsum S, Norseth J, Petit H. Clinical and biochemical heterogeneity in conditions with phytanic acid accumulation. J Neurol Sci. 1987;77:87–96. http://www.ncbi.nlm.nih.gov/pubmed/2433405. Accessed 4 Oct 2017.
75. Skarnes WC, Rosen B, West AP, Koutsourakis M, Bushell W, Iyer V, et al. A conditional knockout resource for the genome-wide study of mouse gene function. Nature. 2011;474:337–42. doi:10.1038/nature10163.
76. Olsson M, Frankowiack M, Tengvall K, Roosje P, Fall T, Ivansson E, et al. The dog as a genetic model for immunoglobulin a (IgA) deficiency: identification of several breeds with low serum IgA concentrations. Vet Immunol Immunopathol. 2014;160:255–9. doi:10.1016/j.vetimm.2014.05.010.
77. Karyadi DM, Karlins E, Decker B, vonHoldt BM, Carpintero-Ramirez G, Parker HG, et al. A copy number variant at the KITLG locus likely confers risk for canine Squamous cell carcinoma of the digit. PLoS Genet. 2013;9:e1003409. doi:10.1371/journal.pgen.1003409.
78. Lao O, de Gruijter JM, van Duijn K, Navarro A, Kayser M. Signatures of positive selection in genes associated with human skin pigmentation as revealed from analyses of single nucleotide polymorphisms. Ann Hum Genet. 2007;71:354–69. doi:10.1111/j.1469-1809.2006.00341.x.
79. Metzger J, Karwath M, Tonda R, Beltran S, Águeda L, Gut M, et al. Runs of homozygosity reveal signatures of positive selection for reproduction traits in breed and non-breed horses. BMC Genomics. 2015;16:764. doi:10.1186/s12864-015-1977-3.
80. Gutierrez-Gil B, Arranz JJ, Wiener P. An interpretive review of selective sweep studies in Bos Taurus cattle populations: identification of unique and shared selection signals across breeds. Front Genet. 2015;6:167. doi:10.3389/fgene.2015.00167.
81. Miller CT, Beleza S, Pollen AA, Schluter D, Kittles RA, Shriver MD, et al. Cis-regulatory changes in kit Ligand expression and parallel evolution of pigmentation in sticklebacks and humans. Cell. 2007;131:1179–89. doi:10.1016/j.cell.2007.10.055.
82. Mengel-From J, Wong TH, Morling N, Rees JL, Jackson IJ. Genetic determinants of hair and eye colours in the Scottish and Danish populations. BMC Genet. 2009;10:88. doi:10.1186/1471-2156-10-88.
83. Raney BJ, Dreszer TR, Barber GP, Clawson H, Fujita PA, Wang T, et al. Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC genome browser. Bioinformatics. 2014;30:1003–5. doi:10.1093/bioinformatics/btt637.
84. Smit A. The origin of interspersed repeats in the human genome Arian FA Smit. Curr Opin Genet Dev. 1996;6:743–8.
85. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 1999;27:573–80. http://www.ncbi.nlm.nih.gov/pubmed/9862982. Accessed 24 Feb 2017.
86. Hach F, Hormozdiari F, Alkan C, Hormozdiari F, Birol I, Eichler EE, et al. mrsFAST: a cache-oblivious algorithm for short-read mapping. Nat Methods. 2010;7:576–7. doi:10.1038/nmeth0810-576.
87. Nicholas TJ, Baker C, Eichler EE, Akey JM. A high-resolution integrated map of copy number polymorphisms within and between breeds of the modern domesticated dog. BMC Genomics. 2011;12:414. doi:10.1186/1471-2164-12-414.
88. Schreiber J. Pomegranate. 2014. https://github.com/jmschrei/pomegranate.
89. Li H, Durbin R. Fast and accurate short read alignment with burrows-wheeler transform. Bioinformatics. 2009;25:1754–60. doi:10.1093/bioinformatics/btp324.
90. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43:491–8. doi:10.1038/ng.806.
91. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, et al. From FastQ data to high confidence variant calls: the genome analysis toolkit best practices pipeline. Curr Protoc Bioinforma. 2013;43:11.10.1–33. doi:10.1002/0471250953.bi1110s43.
92. Alexa A. Rahnenfuhrer J. TopGO. 2016; https://bioconductor.org/packages/release/bioc/html/topGO.html.
93. Alexa A, Joorg Rahnenfuhrer. TopGO Manual, page 19, section 6.2. 2017. https://bioconductor.org/packages/release/bioc/vignettes/topGO/inst/doc/topGO.pdf.
94. Supek F, Bošnjak M, Škunca N, Šmuc T. REVIGO summarizes and visualizes long lists of gene ontology terms. PLoS One. 2011;6:e21800. doi:10.1371/journal.pone.0021800.
95. Weir BS, Cockerham CC. Estimating F-statistics for the analysis of population structure. Evolution (N Y). 1984;38:1358. doi:10.2307/2408641.

3.2.

Serres-Armero Aitor, David Juan, Brian W. Davis, Inna S. Povolotskaya, Carlos Morcillo-Suarez, Jocelyn Plassais, Elaine A. Ostrander and and Tomas Marques-Bonet. Dog breed variation in genomic copy number underlies complex and novel phenotype associations. (*In preparation*).

Dog breed variation in genomic copy number underlies complex and novel phenotype associations.

Aitor Serres-Armero[1], David Juan[1], Brian W. Davis[2,3], Inna S. Povolotskaya, Carlos Morcillo-Suarez[1], Jocelyn Plassais[2], Elaine A. Ostrander[2§] and Tomas Marques-Bonet[1,4,5§]

[1] IBE, Institut de Biologia Evolutiva (Universitat Pompeu Fabra/CSIC), Ciencies Experimentals i de la Salut, Barcelona, 08003, Spain

[2]Cancer Genetics and Comparative Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, 20892, USA

[3] Department of Veterinary Integrative Biosciences, College of Veterinary Medicine, Texas A&M University, College Station, TX 77843

[4] CNAG-CRG, Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology (BIST), Barcelona, Spain

[5] Institucio Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Catalonia 08010, Spain

[§] Corresponding authors

Keywords: Copy number variation, dog genomics, dog breeds, morphometrics, disease

Email addresses:

ASA: aitor.serres@upf.edu

DJ: david.juan@upf.edu

BWD: bdavis@cvm.tamu.edu

ISP: ipovolotskaya@gmail.com

CMS: carlos.morcillo@upf.edu

JP: jocelyn.plassais@nih.gov

EAO: eostrand@mail.nih.gov

TMB: tomas.marques@upf.edu

Abstract

The extreme phenotypic diversity, history of artificial selection and socioeconomic value make domestic dog breeds a compelling subject for genomic research. Copy number variation is known to account for a significant part of inter-individual genomic diversity. However, the relevance of structural variation in the definition of breed-specific phenotypic and disease-related traits remains underexplored. We have generated whole genome copy number variation (CNV) maps for more than 300 canids. Our dataset extends the canine structural variation landscape to more than 100 dog breeds, including novel variants which cannot be assessed with classical technologies. We have taken advantage of this dataset to perform the first CNV-GWAS in canids. Here we report 96 loci displaying differences in copy number across breeds statistically associated with a previously compiled set of breed-specific morphometrics and disease susceptibilities. Integration with external information highlights a subset of specific loci and genes with promising functional relevance to explain trait variabilities among dog breeds. ..

Introduction

Dogs have been the subject of intense study over many decades (Vilà et al. 1999; Ostrander and Wayne 2005; Freedman et al. 2014), providing invaluable insight into human history, disease and evolution (Ní Leathlobhair et al. 2018; Wang et al. 2018; Coelho et al. 2018). Much has been learned about canine phylogenetics and association to traits through traditional approaches, originally using micro-satellite genotyping, single nucleotide polymorphisms (SNP) and finally genotyping using whole genome sequencing (WGS) data (Irion 2003; Gundry et al. 2007; Plassais et al. 2019). Unfortunately, there is still a lack of genome-wide analyses of small indels and other genomic variants across dog breeds, a notable exception when compared to humans and other model organisms (Yalcin et al. 2011; Brown et al. 2012; Sudmant et al. 2015b). The canine genome assembly and annotation have had very minor updates since 2011 (Kim et al. 1998; Lindblad-Toh et al. 2005; Wucher et al. 2017), and few extensive transcriptomic, epigenomic or chromatin conformation experiments have been performed using multiple dog breeds (Hoeppner et al. 2014; Vietri Rudan et al. 2015). Copy number variation has been previously studied in dogs and wolves to elucidate specific phenotypes (Arendt et al. 2014; Waldo and Diaz 2015; Deane-Coe et al. 2018) However, most of the literature has focused on the comparison of dogs and wolves using array-based technologies (Berglund et al. 2012; Schoenebeck et al. 2012). Previous studies were designed to genotype known CNVs and did not seek to build a repository of new variation. Additionally, most CNV related studies only aimed to

qualitatively find segmentally duplicated regions but were unable to emit absolute CN genotypes (Quilez et al. 2012; Molin et al. 2014). This is especially remarkablein the light of recent findings (Serres-Armero et al. 2017)proposing that dogs have a comparable amount of copy number variable loci in dogs to wild canids which have not undergone domestication.

Currently, there exist about 400 dog breeds, 193 breeds registered by the American Kennel Club and 360 by the Fédération Cynologique Internationale. Dogs were initially domesticated from gray wolves 13,000-30,000 years ago (Freedman et al. 2014) with rapid diversification of breeds occurring within the past few hundred years. Several breed classifications have been proposed based on breed occupation, morphology or history (American Kennel Club 2007; Wucher et al. 2017). The most recent genetic analysis encompassing nearly 200 breeds suggests a monophyletic origin of most modern breeds and provides data regarding breed origins and timing (Parker et al. 2017). There, clusters of genetically similar breed groups are found, which sometimes resemble the occupational and historical classifications. We adopt the term breed clades, or macro-groups, to refer to these genetically similar clusters.

The recent and intense artificial selective pressure exerted on dogs has led to pronounced inter-breed phenotypic differences while preserving intra-breed homogeneity (cite). This makes dogs from the same breed much more likely to not only share morphometric traits, but also behavioral patterns and disease propensities (cite). Such is the level of similarity within dog breeds that anatomical standards, also referred to as stereotypes, have been created for most of the existing breeds. Purebred dogs tend to adhere so tightly to these standards that phenotype inference based on breed stereotypes is a relatively common and fruitful practice (cite). This scenario is especially exploitable by genotype-phenotype mapping assays such as genome wide association studies (GWAS), since genetic differences across breeds are more likely to explain trait variation (cite). A number of GWAS studies in dogs have shed light into the biology of both morphometric features and veterinary conditions (cite). Indeed GWAS in canines has been unexpectedly efficient in unraveling the genetic bases of apparently complex traits such as body size or behavior (cite), which remain elusive even in humans. However, all these analyses have been performed using a subset of indicative single nucleotide markers or more recently WGS, but other forms of genomic variation have rarely been studied in such a systematic manner.

Here we present a fine-scale copy number assessment of > 400 canid samples, analyzed at the whole genome level. We examine > 145 individual breeds, as well as non-breed dogs including village dogs, dingoes, captive New Guinea singing dogs and wild canids such as wolves. We use this dataset to recreate the current dog phylogeny using genome-wide copy number differences. Moreover, we test for canid and breed associations in the first CNV-based GWAS performed in dogs to date.

Results

We created a fine-scale CNV map for a panel of 263 purebred dog genomes, 59 village dogs from diverse locations, and 17 grey and Tibetan wolves. All the samples were previously sequenced at low to medium coverage (methods, Supplemental Data S1,2). We report 26,991 autosomal CNV events larger than 1 kb in at least one sample. We find over 95% concordance between the structural variants generated in this study and those that we previously reported (Serres-Armero et al. 2017). We inferred the sample phenotypes based on breed standards from the American Kennel Club (AKC), the Fédération Cynologique Internationale (FCI) and the Orthopedic Foundation for Animals (OFA) among others (methods, Supplemental Data S3) and referred to purebred dogs for the majority of our analyses.

Copy Number Statistics of modern dogs, village dogs and wolves

Overall, we report a total of 221.32 Mb of copy number change across all samples, amounting to approximately 9.66% of the whole dog genome. Of these regions, 169.96 Mb are duplications and 51.36 Mb deletions relative to the reference genome. We detected a total of 6,657 autosomal duplication events with an average size of 25.53 kb and a median spacing of 119.44 kb. A total of 20,334 deletion events with an average size of 2.36 kb and a median spacing of 48.21 kb are also observed. Neither duplications nor deletions seem to be randomly distributed throughout the genome (coordinate randomization test with p-value <0.001), as has been extensively described previously in other species (Li et al. 2009; Upadhyay et al. 2017). We observe 50.73 Mb, amounting to 22.9% of the global CNV map, primarily duplications, encompassing 360 whole gene annotations with CNV status in more than 320 samples. This implies that for the entirety of the panel these are largely copy number variable. Additionally, we detect 129.86 Mb in structural variants which partially overlap with over 7,200 gene annotations. We only find a total of 14.43 singleton Mb in our panel, meaning that most of the CNV loci observed are present in at least two individuals (Supplemental Fig. S1, methods). A total of 193.9 Mb or 87.6% of all CNVs are segregating in the population, i.e. there are at least two individuals with different CN genotypes for said loci., These segregating CNVs are divided in 143.8 Mb present in duplications and 50.1 Mb in deletions,

We assessed how much of the current breed phylogeny, constructed using 170,000 SNPs (Parker et al. 2017) can be recapitulated using the duplications and deletions reported here (Supplemental Fig. S2,3). Altogether, we are able to separate breeds resulting from the first domestication bottleneck (i.e. Arctic and Asian spitz, ancient sighthounds...) (Freedman et al. 2014), but we do not achieve a fully monophyletic separation of breeds derived in the eighteenth century and after, even when accounting for possible described admixture and inbreeding effects (Parker et al. 2017). As the correlation of

CNV with geography and genealogy is not as clear as in single nucleotide variation (SNV), we expect the confounding effects of population stratification to be reduced accordingly (Zhang et al. 2008; Price et al. 2010). As such, we found no private structural variant for any breed group; that is, we did not identify any copy number variable region present in all samples from a certain breed group which was simultaneously single copy in all others.

Comparative of modern dogs, village dogs and wolves

We did not observe a significant reduction in the number of CNV sites in purebred dogs when compared to wolves (Serres-Armero et al. 2017) (Supplemental Fig. S4). This is in stark contrast to the SNV decline reported in numerous domesticated animals (Freedman et al. 2014; Makino et al. 2018). Surprisingly, village dogs show a slightly smaller number of CNV sites compared to dogs and wolves, which might suggest that CNVs have been artificially maintained in domestic dogs through selective breeding (Serres-Armero et al. 2017).

We applied the pairwise Vst statistic (Redon et al. 2006) to test for highly differentiated regions overlapping genes across dog and wolf pairs. We identify a number of stratified CNVs that in total equal 11 Mb. Not surprisingly, some of these have been extensively reported multiple times before, such as *AMY2B* or *MAGI2* (Chen et al. 2009; Arendt et al. 2014). We also found some novel gene-overlapping CNVs (Supplemental Table S1). Of particular interest is *IRS1* (Fig. 1A), a gene involved in insulin resistance in humans, which is present in one copy in many wolves (Morgane Ollivier, Anne Tresset, Fabiola Bastian, Laetitia Lagoutte, Erik Axelsson, Maja-Louise Arendt, Adrian Bălăşescu, Marjan Marshour, Mikhail V. Sablin, Laure Salanova, Jean-Denis Vigne, Christophe Hitte, Catherine Hänni 2016). Moreover, we find an unexpectedly large proportion of CN differentiated genes involved in fatty acid metabolism (GO p-value < 10E-3), some of which have been previously reported. This enrichment is potentially attributable to lifestyle differences between the three groups (Björnerfeldt et al. 2006; Li et al. 2014). We also report differences in *HBB1* (Fig. 1B) (CN 2 in wolf) and *SLIT2* (Fig. 1C) (single copy in wolf), which have been associated with adaptation and development (Hu 1999; Bigham 2016). Domestic and village dogs and wolves can also be discriminated via principal component analysis or by sheer pairwise euclidean distance (Fig. 1D).

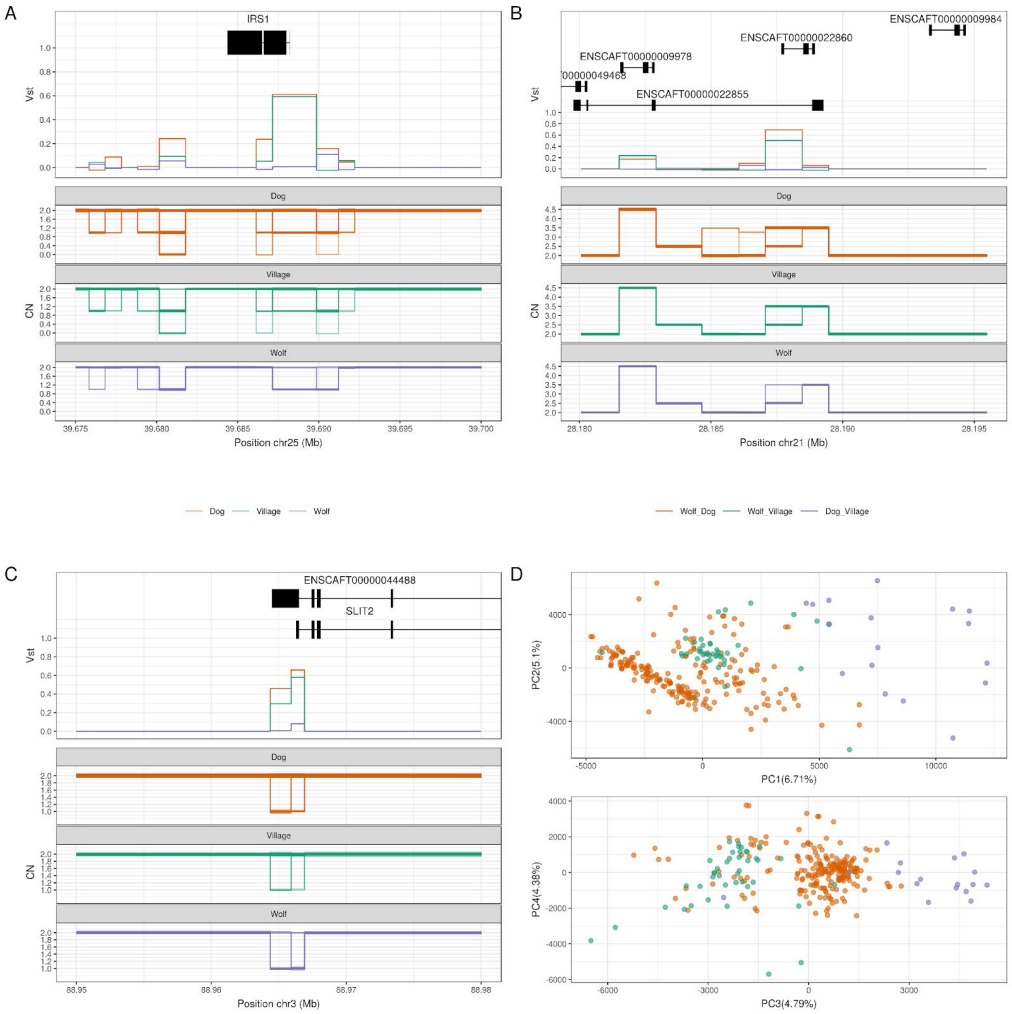

Figure 1. (A) Copy number based principal component analysis of dogs (green), village dogs (red) and wolves (blue). (**B-D)** Depictions of the copy number values for highly differentiated loci HBB1, IRS1 and SLIT1. Discretized copy number calls for all samples are depicted in step plots with a small jitter. Blue: wolves; Green: village dogs; Orange: modern dogs. Top panel: copy number window values; Bottom panel: Vst values for the same genomic windows.

GWAS

To date, a great number of traditional genome-wide association analyses have been performed in domestic dogs (Vaysse et al. 2011; Hayward et al. 2016; Plassais et al. 2019). However, given the lack of global maps of genome-wide CNV analyses, absolute copy number has never been globally assessed for trait associations. Here, we used over 20 phenotypes to uncover associated CNVs. In order to assess different association trends, we have implemented and compared four generalized traditional association tests, both discrete and continuous (see Methods). Nevertheless, it is worth keeping in mind that copy number inference remains a field of active investigation (Layer et al. 2014; Trost et al. 2018) and some of its genetic bases remain elusive (Hastings et al. 2009; Muñoz-Amatriaín et al. 2013).

As discussed previously, population stratification in copy number is expected to be small (Supplemental Fig. S2). Even so, we decided to control for population stratification by using the first four principal components of the data structure (Wu et al. 2011) as covariates in regression analyses or by partitioning the data into substrata when applying categorical tests (Agresti 2002). We found this approach to be overly conservative at times, so we resorted to its application only when we detected an excess of significant p-values after controlling for inflation (Tsepilov et al. 2013). Our GWAS associations were enriched in intergenic regions and non-coding genes such as lncRNA. Therefore, we decided to use orthogonal data sources (Hoeppner et al. 2014; Vietri Rudan et al. 2015; Sudharsan et al. 2018; Le Béguec et al. 2018) (methods) in order to validate and explore the relevance of the most interesting results. We used GWAS, Hi-C, transcription and conservation data to annotate and contextualize our association signals, especially those showing lower significance.

Over half a dozen loci had already been identified as major drivers of body size variation in canines, the major gene contributors being *IGF1, IGF1R, STC2* and *GHR* (cite), however only *SMAD2* had been previously related to CNV (cite). Our CNV analysis reproduces a two of the previously reported GWAS body size associations -chr26:12796099-13004170 (Fig. 2A,B, Fig. 3A) (Hayward et al. 2016; Plassais et al. 2019) and *SMAD2* (Fig. 2A,D, Fig. 3B) (Rimbault et al. 2013) (Supplemental Table S2)- which may be taken as a proof of concept of the global approach. We report an associated duplication (chr26:12,739,546-12,754,676) which harbors a CpG island 20 kb upstream the *MED13L* gene. Interestingly, we detect a well-supported Hi-C interaction between this duplication and the region containing the *TBX3* gene located almost one Mb downstream (Fig. 4A). *TBX3* has been reported to cause short stature in humans and is also a major contributor to height in horses (Kader et al. 2015). *SMAD2* is tagged by a 9.9 kb deletion located 24 kb downstream from the protein-coding region (chr7:43,794,129) which engulfs a CpG island on the gene tail and has previously been hypothesized to be an enhancer (Rimbault et al. 2013). Additionally, the gene *FGF4* (Fig. 2A,B, Fig. 3C) (Brown et al. 2017) displays a strong association signal with height at the withers due to the

documented retrogene copy (Parker et al. 2009) which confers chondrodysplasia to certain breeds. Each of the three genes contributes significantly to the trait and only one copy number variant seems to be sufficient for a dog to have small size (Fig. 3D). The three genes do not present evident statistical interactions and remain significant in the absence of the others.

We report an interesting, potentially functional finding in a ~26 kb deletion (chr6:15,642,612-15,668,739) containing homeobox gene *UNCX* and many CpG islands (Fig. 2E,F), which is associated with tail to body ratio. *UNCX* has been reported to play an important role in tail formation during mice development (Chalamalasetty et al. 2014).

Disease propensities were also tested for CNV association genome wide following the same procedure. Importantly, the lack of individual phenotypes could potentially be more detrimental to the power to detect associations in this assay, since penetrance and morbidity were not accounted for. This manifests as an overall p-value deflation over most assayed pathoses, especially thyroid and cardiac conditions. Despite these caveats, we find an association for generalized-progressive retinal atrophy susceptibility in a duplication covering more than ten exons of the *DMBT1* gene (chr28:32,220,591-32,260,415) (Fig. 2G,H). Interestingly, a CNV in the same gene has been hypothesized to cause macular degeneration in humans (Polley et al. 2016), but to our knowledge, this is the first report for a similar eye condition in dogs. Here, we also report an intronic deletion in the osteoclast activating gene *PTPRε* (Chiusaroli et al. 2004) associated with patellar luxation propensity (Supplemental Table S2). This deletion notably shortens the space between the exons 8 and 9 from approximately two kb to less than one, a phenomenon which may potentially affect the function of the associated gene (Chen et al. 2015; Rigau et al. 2019).

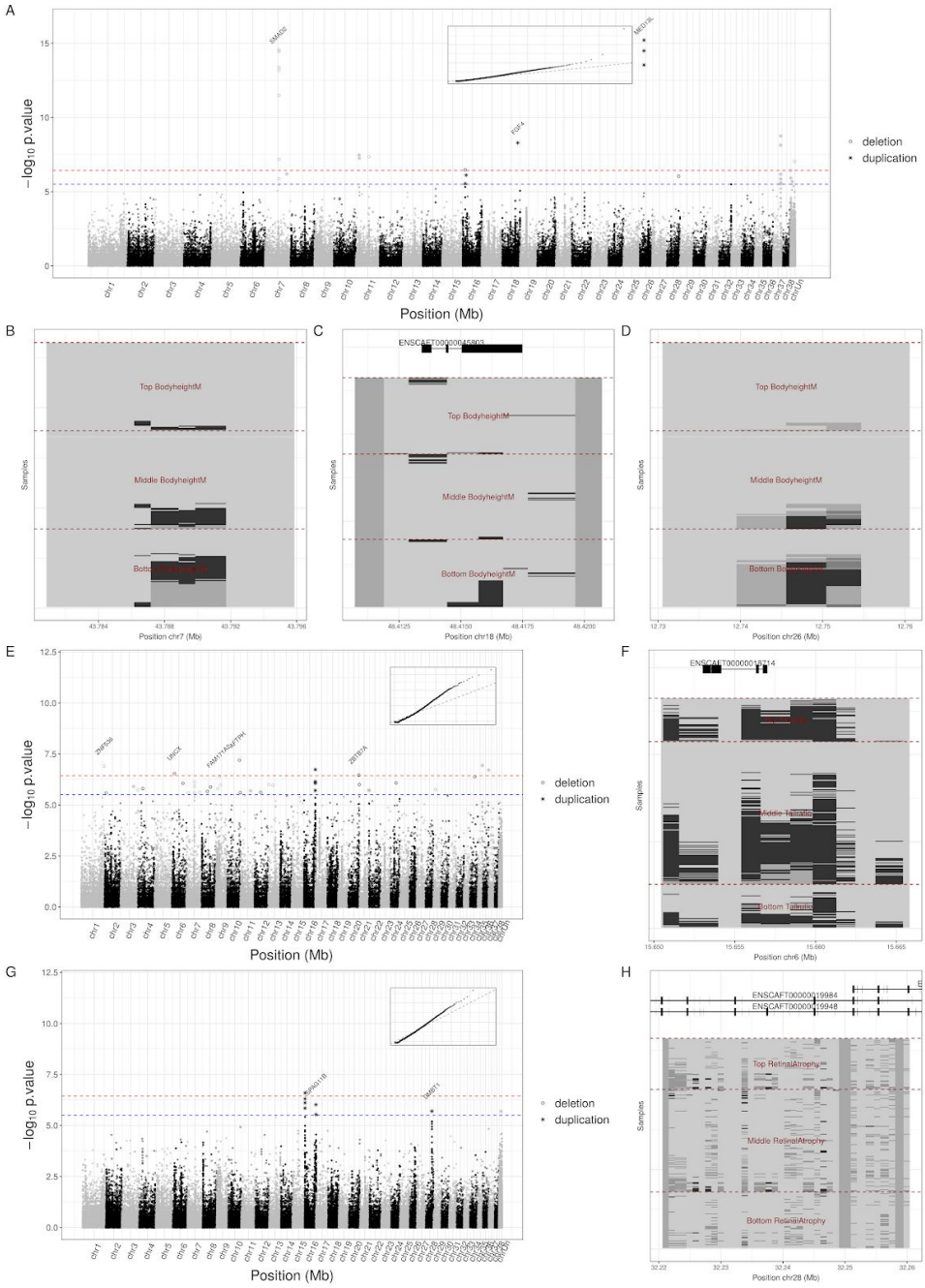

Figure 2. (A, E, G) Manhattan plots of the copy number GWAS for breed standard height, retinal atrophy susceptibility and tail to body ratio respectively (Jones et al. 2008). Red line: Bonferroni correction. Blue line: one order of magnitude below Bonferroni correction. P-values were calculated using the generalized CMH test and corrected for inflation. **(B-D, F, H)** Close-up of the relevant regions for each of the traits respectively (CFA26, *FGF4* and *SMAD2* for height, *DMBT1* for retinal atrophy and UNCX for tail size). Each sample corresponds to a line along the y-axis and is ordered according to the trait in question. CN windows for each sample are colored according to their normalized distance to the median CN in the region. The x axis shows the genomic position of each window.

We also detect some interesting associations which are in close proximity to genes of interest, such as *RXRA* and *DECR2* for aging and *CORIN* with respect to hair length. *RXRA* has been shown to play a role in reversing brain aging processes in mice (Natrajan et al. 2015). *DECR2* is involved in fatty acid metabolism and could be potentially involved in aspects of aging in mice (Miinalainen et al. 2009). The *CORIN* protease has been reported to influence hair follicle development in mice and could be associated with hair coloring and shaping (Enshell-Seijffers et al. 2008). A point mutation in CORIN is causative for the snow white tiger pelage color phenotype (Xu et al. 2017).

We report an interesting case where two genes (*NLRP13* and *NLRP8*, involved in innate immune response) seemingly unrelated to the herding phenotype show moderately high association signals for it (Fig. 5A). This is association results from German Shepherds and Rottweilers (notoriously used for livestock herding in Germany) (van der Borg Elisabeth A. M. Graat BonneBeerda 2017) having different copy numbers for those genes than most other breeds. Even if immunity is out of the scope of this paper due to the lack of breed-wide phenotypes, these two breeds have been reported to contract more auto-inflammatory diseases than others (Day 1999; Wiberg et al. 2000; Jokinen et al. 2011), mimicking an extensively reported process with the human *NLRP* family variation (Amin et al. 2017).

GWAS annotation

We gathered data on alternative genomic variants aside from copy number variation to test whether our secondary-threshold GWAS associations followed any discernible patterns. Concordance between multiple non-coding or intergenic regions and their previous, independent annotations could both serve as a validation and potentially point to polygenic effects.

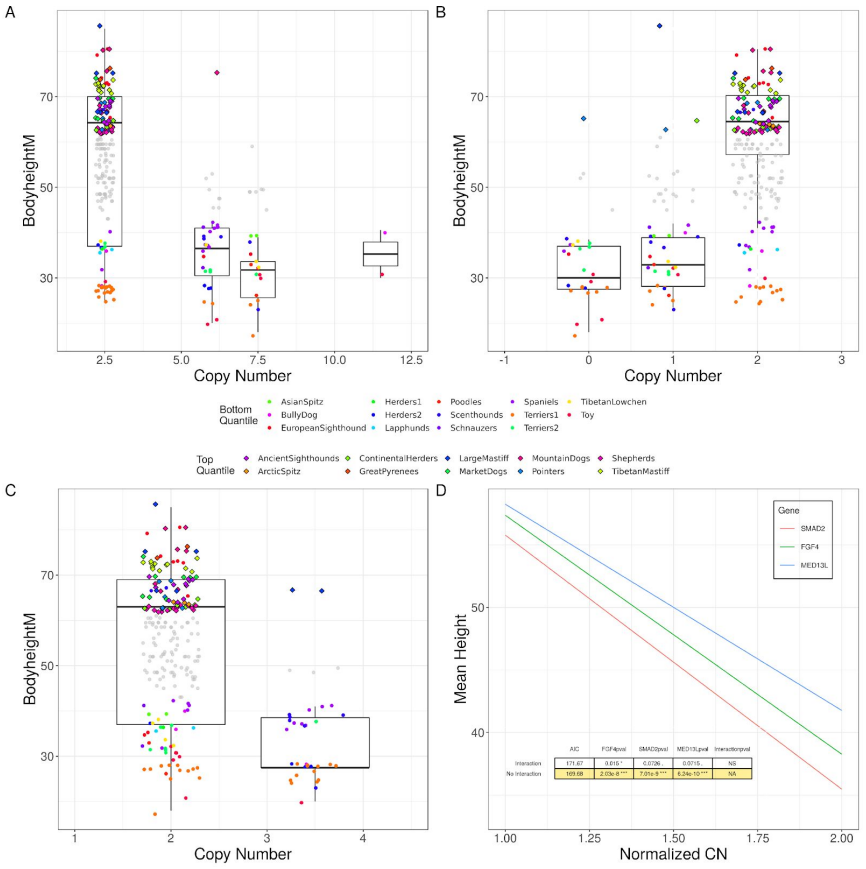

Figure 3. (A-C) Boxplot dissection of the three main body mass/height associations SMAD2, FGF4 and CFA26 respectively. Breeds are colored by trait percentile. **(D)** Interaction plot for the three aforementioned genes. Within the plot is a table showing that the model without interaction is preferred.

We cross-referenced our CN GWAS signals with one preceding WGS-GWAS study for similar traits (Plassais et al. 2019). For each reported SNV association, we assessed whether the closest CNV signals had increased p-values (methods). Even if we were able to identify this trend in some sparse cases, most prominently deletions, the majority of our associations were poorly tagged by SNPs (mean distance to the closest Illumina CanineHD SNP >37 kb) (Sudmant et al. 2015a). Of interest, one of our most significant GWAS hits, SMAD2, segregates together with a previously reported SNP at frequencies 0.6 ± 0.29 depending on the breed (Jones et al. 2008; Chase et al. 2009). Similarly, we observe smaller p-values around a polymorphism related to tail curl near the *CHSY3* gene (Plassais et al. 2019) (Supplemental Fig. S6).

In order to assess whether our intergenic and intronic association hits could point to unannotated regulatory regions, we studied the enrichment in conserved motifs. For this, we intersected the 75

way GERP score ensembl annotation (Hunt et al. 2018) with our significant calls (methods). We found no significant increase in the number of conserved and associated copy number variants compared to the global background of all non-genic structural variation (Supplemental Table S3). This means that associated copy number events behave the same as the rest of copy number events in terms of sequence conservation. Indeed, there seems to be an overall depletion in highly conserved motifs in our whole structural variation space, most likely due to the poor alignment properties of complex regions when constructing the conservation scores.

A substantial part of our CNV associations either overlapped or was in close proximity to long non-coding RNAs. Therefore we assessed the concordance between the lncRNAs tissue expression patterns and their trait association as indicative of a non-spurious distribution of minor association results. We used the dog lncRNA database (Le Béguec et al. 2018) to annotate the GWAS signals within a 10 kb range of a lncRNA based on the tissue where the lncRNA is most abundant. We compared empirical GWAS-lncRNA contingency table against an independent distribution of both features (methods). Remarkably, we found that some traits are enriched in their respective, concordant lncRNA tissues (e.g. brain lncRNA expression for intelligence traits) while retaining the expected counts in all others. Particular examples of this are testis expression for a litter size association, muscle, blood, and heart for racing and adrenal gland for temperament among others (Supplemental Fig. S7). While these results should be interpreted cautiously due to the low contingency table counts, the consistency between traits with no major protein-coding associations encourages further exploration of complex traits such as intelligence or temperament.

Finally, we also assessed whether any associated region aside from the aforementioned *TBX3* displayed any distal Hi-C contacts (Fig. 4A). We verified all contacts reported here using ChIP-seq data for dog CTCF motifs (Schmidt et al. 2012), assessing whether each end of the contact contains at least one CTCF motif in inward opposing directions (methods). We find about seven interesting, well-supported associated interactions in our dataset. Most prominently, an association signal for hair length in a largely unannotated genomic region (chr9:16,780,483-16,782,227) interacts with the *MAP2K6* gene located almost one Mb away (Fig. 4B), which has been associated with hypertrichosis in humans (Clark et al. 2016). Also interesting is an association signal for intelligence with a moderately low p-value (p-value < 10E-6) (chr4:64,513,062-64,514,795) which significantly interacts with the last exons of the *HCN1* gene (Fig. 4C). *HCN1* has been extensively studied in humans for its involvement in attention, memory and fluid intelligence (Thuault et al. 2013). Finally, we report a recurrent, low p-value (p-value < 10E-6) association signal (chr7:43,788,068-43,804,257) for many traits correlated to body mass (e.g. litter size, lifespan, body height...). The interaction involves the multipurpose microtubule processing *KATNAL2* which does not seem to have any clear relation with the traits in question (Supplemental Table S4). Long-range interactions involving lncRNAs were also

accounted for in the corresponding analysis.

Breed Vst

We then set out to study possible differences in copy number arising from breed differentiation, even if they do not involve any measurable phenotypic effect. Despite the reportedlack of global phylogenetic gene expansion regarding modern dogs, highly differentiated loci might still correspond to genomic signatures that were swept along in the process of breed derivation (Parker et al. 2017)

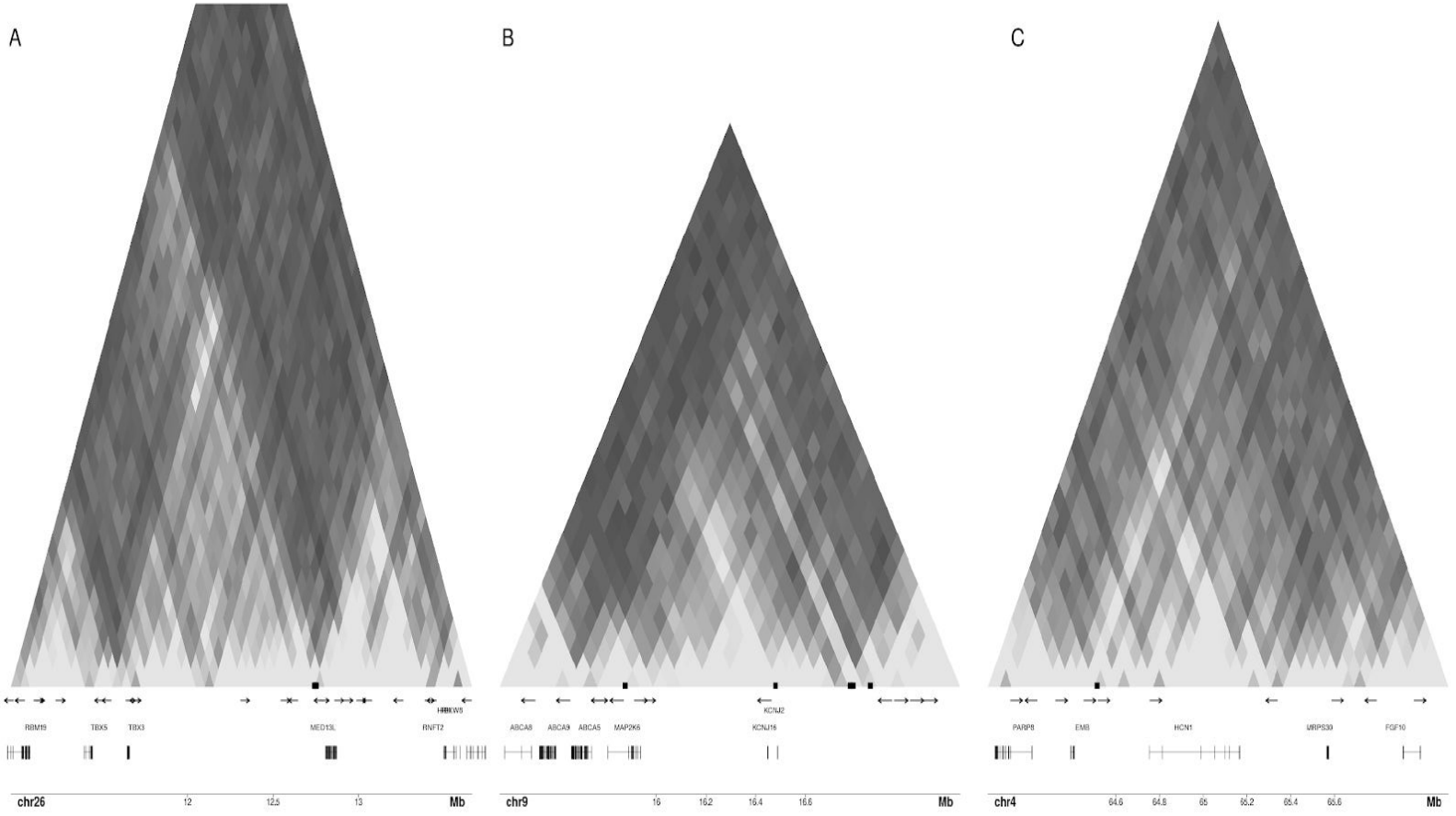

Figure 4. (A-C) Genomic interaction plots for our significant interaction hits on chromosomes 26, 4 and 9. The black squares on the interaction plot mark the position the GWAS hits. The arrows mark the position and directionality of the most significant CTCF motifs of the area.

We applied the pairwise Vst statistic (Redon et al. 2006) (methods, Supplemental Table S4) to all pairs of breed macro-groups consisting of more than six individuals (methods). Overall, we found some very differentiated regions in certain breed groups (mainly Tibetan Mastiffs, Arctic Spitz, Shepherds, Ancient Sighthounds, and Scenthounds). Most if these regions map to gene poor regions. However they often contain one or more members of extensive protein families (olfactory receptors, solute carriers, late cornified envelope proteins, ...). Of special mention is an intronic deletion of ~35 kb (chr4:17,945,894-17,986,191) in *CTNNA3* intron 3 (Fig. 5B). A deletion of this size in a near ~75 kbp intron of a big gene could theoretically have an impact in expression levels, transcription times or splicing activity (Rigau et al. 2019). Terriers and retrievers possess fewer copies of this gene than most other breeds. About eight out of the 16 sampled West Highland White terriers carry a possibly homozygous deletion of this intron chunk. In humans, variation in this gene has been associated with congenital progressive macular dystrophy, which is a common condition of terriers (Supplemental Data S1). We also find a homozygous deletion (chr27:25,735,696-25,844,995) in many German shepherds which encompasses two *SLC7A* orthologs (Fotiadis et al. 2013) (Fig. 5C).

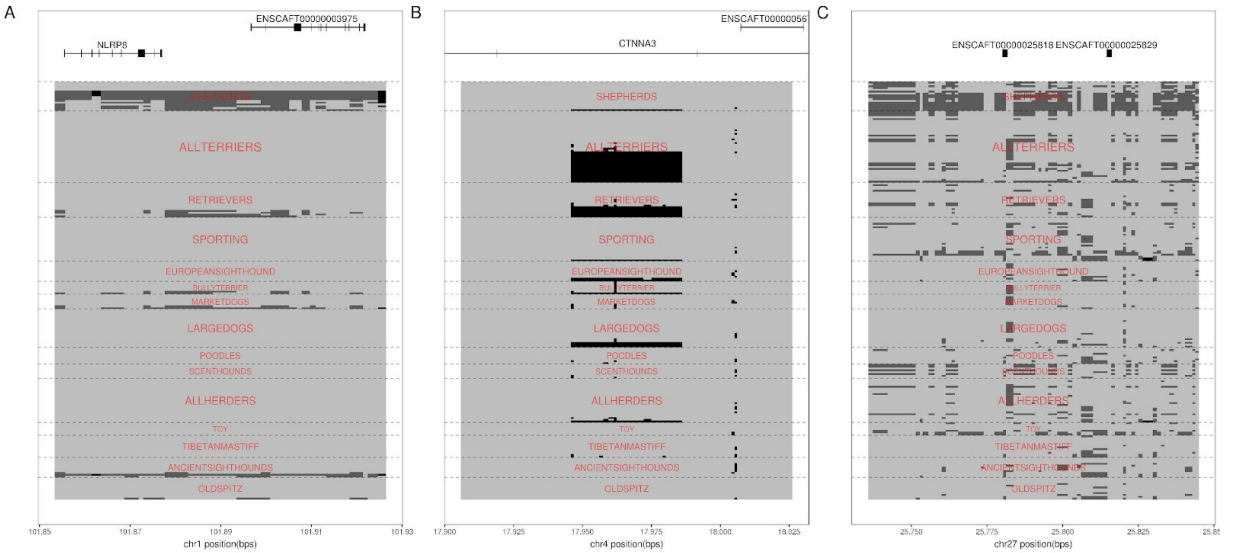

Figure 5. (A-C) Representation of high inter-breed Vst regions for the *NLRP8-13* genes, *CTNNA3* and *SLC7A* respectively. Normalized copy number is represented by color where the modal CN is the lightest color and deviations from the mode (either deletions or duplications) are colored darker. Each row represents a sample ordered by breed.

Discussion

Genome-wide association analyses with absolute copy number as the testable variable are not common in the literature, especially outside of the field of disease (Wellcome Trust Case Control Consortium et al. 2010), and less so in domesticated animals. In part, this is due to the increased technical difficulties of working with CNVs (Zhao et al. 2013; Sudmant et al. 2015a) compared to SNPs, and a more complex genetic scenario arising to their birth and death (Sudmant et al. 2015a; Xu et al. 2016) These hindrances ultimately result in a lack of specific tools, validation methods and workflows. However, copy number events have been correlated to distinctive traits in a vast array of organisms (Hegele 2007; Karyadi et al. 2013; Chain et al. 2014; Upadhyay et al. 2017). Moreover, copy number is known to explain a complex different landscape of variation which is often poorly tagged by SNPs (Sudmant et al. 2015a). However, few structural variants and their phenotypic effects have already been explored in dogs (Alvarez and Akey 2012) In fact, until this work, a comprehensive and cohesive whole genome approach had not yet been attempted in this organism.

Through our approach, we are able to discover new associations between phenotypes of interest and copy number variants within or adjacent to excellent candidate genes. The replication of previous findings in CNV associations, especially with height and body mass, has been used as a proof of concept for this global approach (Rimbault et al. 2013). We find some novel structural variants harboring whole or partial genes and regulatory elements. We hypothesize that these variant are not only associated with the trait but also play a role in the actual phenotype. In spite of the supportive concurrence of orthogonal evidence in these associations, , additional functional studies will beneeded to further validate such hypotheses. We also use orthogonal genomic variants to contextualize and ascertain many of our secondary threshold p-value associations, and discover many potentially interesting trends which might have been otherwise overlooked. We, therefore, highlight the importance of using non-coding variants, particularly when working with potentially complex traits with many subtle underlying associations. The accumulation of concordant features within our associations encourages further investigation, even if the false discovery rate associated with a breed-to-sample GWAS phenotype inference is potentially high (Brzyski et al. 2017).

We demonstrate how copy number can be used to separate domestic dogs from their closest extant relatives. We further propose some differentiated genomic regions between domestic dogs, village dogs, and wolves which could be associated with dietary and metabolic disparities. Although we do not have the power to reconstruct the whole phylogeny of modern dog breeds, we are able to separate a small part of the old Spitz dogs and Tibetan Mastiffs from modern breeds (Shubkina et al. 2012; Skoglund et al. 2015). It must, however, be noted that the most recent attempts at dog phylogeny reconstruction using SNV arrays featured three times the number of informative markers of this study

and almost five times the number of samples (Parker et al. 2017). It has been well established in human populations that copy number can seldom match the resolution of SNV when it comes to phylogeny reconstruction (Sudmant et al. 2015a) and doing so would involve a massive increase in sample size.

Additionally, we find a few isolated, highly differentiated structural variants across dog breeds, often corresponding to highly variable genome segments and multicopy gene families. Even though extreme variation could be part of the reason why we are able to identify these regions, the absolute copy numbers of many gene-containing regions offers a wealth of hypotheses for further testing.

Altogether, we reemphasize the importance of copy number variation in physiology and morphometrics (Jones et al. 2008; Plassais et al. 2019). Whole genome copy number studies provide an interesting, orthogonal source of information to the more traditional genomic assays and could potentially prove useful to unravel hidden diversity in many understudied organisms.

Materials and methods

Samples

We analyzed a panel of 431 canid samples containing purebred dogs, free ranging (village) dogs and wolves. Four wolves (Wolf34,WOLF6116,Wolf23,Wolf18) were used to train our Hidden Markov Model transition matrix (explained below) and discarded from the final panel. After quality control (described below), we kept a total of 263 dog genomes, 59 village dogs and 17 wolves. Our purebred dog samples classify into more than 130 breeds, which altogether can be divided into more than 30 breed macro-groups (Supplemental Data S1). We used the breed status of each sample to infer its phenotype.

Phenotypes

We composed a database of anatomical, behavioral and disease records for each dog breed in our panel to be used for our association studies. When available, the information was retrieved from the FCI (http://www.fci.be/en/Nomenclature/) or the AKC (https://www.akc.org/dog-breeds/) databases. Temperament and intelligence data were available from the ATTS (http://atts.org/breed-statistics/) database and (Coren 1994) respectively. Disease data was exclusively extracted from the OFA database (https://www.ofa.org/diseases/breed-statistics#detail). Data for purebred dog litter size was obtained from (Borge et al. 2011).

Extra morphometrics as well as confirmation for our measures were extracted from the publication by (Jones et al. 2008).

Copy number calling

Sample pre-processing:

Our initial collection of sample sequencing formats was coerced into fastq format using the appropriate tools (biobambam, qseq2fastq, fastq dump) and all sequencing qualities were regularized

to the standard phred 33 encoding. Adapters were trimmed with TrimGalore (Martin 2011), using paired end data when possible and restricting the output length to a minimum of 36 base pairs. The trimmed sequencing reads were then further split into 78mers to facilitate the mapping process.

Reference assembly preparation:
In order to use an exhaustive mapper and further perform the necessary read depth calculations, the CanFam3.1 assembly was prepared following these steps:

1. Standard repeat masking: masking of the corresponding genome wide tandem repeat finder annotations (Haeussler et al. 2019).
2. Assembly kmer masking: in order to identify potentially hidden repeats, the assembly was split into 36mers with a 5 bp overlap and re-mapped against itself using GEM (Marco-Sola et al. 2012) at 6% divergence with a 10% edit distance. Kmers mapping to more than 20 positions were additionally masked. This version of the assembly was indexed (bwa, GEM, samtools) and used for all subsequent sample mappings.
3. Padding and assembly windowing: all the masked locations described in steps 1 and 2 were extended for a length of 78 bps on each side. This attempts to correct for the common effect of read depth deflation around masked loci. Next, the assembly was partitioned into 1000 bp windows of non-masked sequence as described in (Alkan et al. 2009). The resulting 1kb genomic window coordinates were used for copy number estimation and are theoretically comparable across samples due to the common reference.

Mapping and read depth post-processing:
The pre-processed samples were aligned against the masked CanFam3.1 reference using the GEM exhaustive mapper at 6% divergence and 10% edit distance. The resulting files were processed with mrCanavar (Alkan et al. 2009), a tool for absolute copy number prediction based on read depth normalization which performs GC correction and discriminates between CN 2 (aka control or diploid regions) and potentially duplicated windows.

Quality control:
We assessed three main parameters to decide which samples to include in our experiments:

- First, we formally tested that the distribution of our control regions did not differ much from a Gaussian centered at diploid CN using the kolmogorov-Smirnov test. Extreme deviations from a bell shape or distribution mean shifts could be a product of faulty normalization and might lead to excessive CN miscallings.
- Second, we imposed a hard threshold (0.45) on the dispersion of the diploid regions, as this will later be used to model our HMM emissions and too high a deviation might lead to poor quality genotypes.
- Third, we checked for independence and homoscedasticity in the control region variance. That is we assessed that the variance in the CRs was not locally correlated using Pearson's coefficient.

Copy number genotyping and smoothing

We sought to discretize our copy number estimations to enhance comparability and produce a more biologically relevant CN measure. In order to do so, we prepared a similar setup to the one described

in (Serres-Armero et al. 2017).
We implemented a Hidden Markov Model where the observed read depth (emissions) is linked to a certain integer CN value (hidden state) via a Gaussian distribution. The HMM potentially accounts for spatial CN correlation (transitions) and scatter (emissions) at once.
Briefly, we declare a set of hidden states ranging from 0 to 20 (plus Gaussian mixtures of states with CN above 20) with variance proportional to the empirical diploid dispersion and the hidden CN. We train our transition matrix using the Baum-Welch algorithm in the python pomegranate (https://github.com/jmschrei/pomegranate) package and then predict the forward-backward probability of each state for each 1kb window in every sample.
Additionally, we update our CN distributions by using predicted probabilities of all our samples together. We define a sliding window range of five windows with four window overlap and calculate the expected probability of each state within it. The expected local probabilities of each state are then used as priors to apply Bayes' rule on the third 1kb window within the range for each sample. Finally, instead of just reporting the maximum likelihood CN estimate, we emit the range of CN states whose cumulative posterior distribution sums up to 0.95.

$$p(cn \in [x+dx]) = \frac{p(cn \in [x+dx] \mid CN = N)\, p(CN = N)}{p(cn \in [x+dx])} = \frac{PDF(cn, N, \sigma_{CR}\sqrt{0.5\,CR})\, p(N)}{\sum_{CN} PDF(cn, N, \sigma_{CR}\sqrt{0.5\,N})}$$

Structural variation calling

We defined any windows where at least one individual had a CN range above (and not overlapping) CN 2 as duplications. Similarly we defined all windows with a CN range below (and not overlapping) CN 2 as deletions.
Most of our analyses were restricted to the duplication/deletion space defined here.

Copy number classification and deletion re-calling

Working with ranged CN genotypes can sometimes make it difficult to find natural sample clusters or do genotype classification. Therefore, we implemented a recursive algorithm where for each duplicated 1 kb window we define the set of the most distant non-overlaping CN interval(s) compared to the modal CN. We then assign the rest of the ranges to any of the intervals based on overlap and repeat the process until all intervals have been optimally classified, with the option to create new non-overlaping intervals.
Additionally, we attempted to emit definite, unranged genotypes for our deletion space (defined via HMM) by refitting the empirical observations with a Gaussian Mixture Model. We used the R mixtools package (Benaglia 2009) to fit the mixture weights of a model with fixed means 0, 1, 2 and variances sd(2)/2,sd(2)/2, sd(2). We then used the expected probabilities of each component in all samples to update the individual probabilities on each site using Bayes' rule and kept the most likely genotype. The successful application of this re-genotyping strategy relies on the fact that deletions are believed to be shorter and scattered along the genome, so the spatial component of an HMM is not necessarily needed.

Vst analyses

We apply an in-house implementation of the pairwise Vst statistic (Redon et al. 2006) for each genotyped 1kb window in the dog genome to all breed groups containing 6 or

more individuals . Much like Fst, Vst compares the statistical variance of copy number values within each breed to that of both breeds taken together. We exclude diploid regions reported in our HMM method, as they are assumed to have constant CN.
As we detected that small sample sizes can bias the genomic Vst distribution behavior. We performed 1000 subsamples of all breed group comparisons to 6 individuals and kept the median value for each window and comparison.

$$VST(B_1, B_2) = 1 - \frac{len(B_1)Var(B_1) + len(B_2)Var(B_2)}{len([B_1, B_2])Var([B_1, B_2])}$$

GWAS

Categorical analysis:
We applied an in-house implementation of the generalized Cochran–Mantel–Haenszel (CMH) test by Richard Landis (Landis et al. 1979) as explained in Alan Agresti's 2002 statistical handbook (Agresti 2002). This generalization allows for stratification of data into subpopulations as well as for the ordinalization of phenotypes and copy number.
We split the phenotype data into the top 70 and bottom 30 percentiles (two groups). We further classified copy number into categories as described in the previous corresponding section. Finally, we accounted for possible population stratification by dividing the data into three similarly sized substrata based on the breed tree proposed by Parker et al 2017 (Parker et al. 2017).
The test was repeated for each window in the genome and bonferroni-corrected based on the number of non-diploid windows.

Van elteren test:
The van elteren test is a version of the wilcoxon rank sum test for stratified data coded in the R sanon package (Kawaguchi and Koch 2015) which we applied to all CN-mapped genes. CN was left as a continuous variable while the phenotypes and the subpopulation strata were defined as described above.

Phylogeny

Tree construction:
All euclidean distance matrices were calculated directly from the CN values using the stats R package. The distance matrices were then used to construct phylogenetic trees with the ape (Paradis et al. 2004) R package.

Tree comparisons:
In the occasions when trees containing different samples, breeds and metrics had to be compared, we extracted the common tree topologies by projecting the different distance matrices against the column space of their respective indicator matrices (where each ordered column signals which samples belong to a common breed). The column values of the resulting compound, repetitive matrices were collapsed by breed and propagated across the diagonal to create a symmetric, synthetic distance matrix which retains the topological properties of the original matrix.
The resulting distance matrices were thus ordered, filtered and comparable under common scaling conditions. In our case, we applied simple correlation and 2-norm comparisons.

$$B(B^T B^{-1})^{-1} B^T D$$

$$\forall\, sample \ni [1,2,\ldots I],\ \forall\, breed \ni [1,2,\ldots J] \quad b_{ij} := \begin{cases} 1 & breed \ni sample \\ 0 & breed \notin sample \end{cases}$$

$$d_{ij}^2 = \| CN_{samp\,i} - CN_{samp\,j} \|_2^2$$

Vst tree:

We attempted to construct a tree using the structural variants which differentiate dog breeds the most, under the pretext that those variants might better recapitulate the previously described SNP topologies. In order to do so, only CN windows with Vst values above 0.3 in more than 8 comparisons were retained.

PCA

Principal component analyses were performed using the prcomp function from the R stats package. In order to prevent sample size biases, a common pca basis was created using a random balanced subset of all breed macro-groups. All other samples were projected into this common basis by applying the centering, scaling and rotation matrices outputted by prcomp.

Haplotype sharing tree

In order to test if the possible excessive haplotype sharing across seemingly unrelated breeds was affecting our tree topologies, we attempted to remove these potentially confounding loci.

We subtracted the positions of the pairwise shared haplotype locations in (Parker et al. 2017) from our deletion space and recalculated the sample distances based on the remaining deletions, correcting for the amount of subtracted positions.

We also designed a similar setup where instead of only removing the haplotypes shared in one pairwise breed comparison, we removed the haplotypes for all pairwise comparisons which shared a breed in common. The resulting topologies were compared as described above.

GWAS comparisons and validations

All genome arithmetics were performed using the bedtools suite (Quinlan and Hall 2010) enforcing the necessary parameters. In broad terms, window-based association signals were mapped to their respective structural variants and then intersected with the corresponding annotation files.

lncRNA:

We proposed the independent joint distribution of all copy number lncRNA tissues and the proportion of association signals across traits as the null hypothesis to test for deviations in the associated copy number variant (ptissue $\otimes$ passociation). We assume multinomial distributions over the association table and report excessive cell counts in terms of standard deviations. lncRNA data was downloaded from (Le Béguec et al. 2018).

Conservation scores:

Highly conserved regions were defined by binning GERP scores according to their 95th quantile value

(~3). All non-exonic structural variants were used as a background to test whether non-exonic associated variants were enriched in highly conserved elements. We tested the null hypothesis of variable independence using Fisher's test (variables association & conservation). GERP scores were downloaded from (Hunt et al. 2018).

Hi-C:
We computed the ratio of main contact read support (region against itself) and every other region involving this contact in CNV regions. We set the threshold at the 95th quantile of the distribution (~.25) to call significant contacts. Hi-C data was downloaded from (Vietri Rudan et al. 2015).

ChIP-seq:
All putative significant contacts were verified by assessing both that they contained at least one CTCF motif on each side (Vietri Rudan et al. 2015) and that the CTCF motifs were correctly oriented i.e facing each other. The ChIP-seq data was downloaded from (Schmidt et al. 2012) and lifted over form the CanFam2 genome build to CanFam3.1. We re-annotated the CTCF orientation for the relevant loci using the dog-specific CTCF position weight matrix (https://www.ebi.ac.uk/research/flicek/publications/FOG03) and the software PWMTools (Ambrosini et al. 2018).

Leading SNP:
We gathered all structural variation GWAS values within 1 Mb surrounding the leading SNP GWAS signals proposed by (Plassais et al. 2019). Next, for each leading SNP, we binned the structural variation data into equally sized blocks and assessed if the block containing the leading SNP was significantly enriched in GWAS hits in comparison with the rest.

SNP genotyping

We used FreeBayes (Garrison E, Marth G, preprint) to detect single nucleotide polymorphisms in our data. Provided that the SNP calls performed with extensive mappings on short length reads are expected to be noisier than usual, for each sample, we calculated the 33rd read depth percentile and the 83rd genotype quality percentile. We kept only SNPs within the range of RD33+-RD33/10 and above GQ83 for analysis, as they seemed to empirically correct the allele balance distributions for most of our samples. Variant gathering, filtering and tabulating for further analyses was carried out with custom scripts.

Acknowledgements

We thank EAO, Adam Boyko, Robert Wayne and many other contributors of dog sequencing data for making the samples publically available. We acknowledge the role of the Orthopedic Foundation for Animals, the American Kennel Club and the Fédération Cynologique Internationale for the public availability of their data regarding dog breeds.

References:

Agresti A. 2002. Categorical Data Analysis. *Wiley Series in Probability and Statistics*. http://dx.doi.org/10.1002/0471249688.

Alkan C, Kidd JM, Marques-Bonet T, Aksay G, Antonacci F, Hormozdiari F, Kitzman JO, Baker C,

Malig M, Mutlu O, et al. 2009. Personalized copy number and segmental duplication maps using next-generation sequencing. *Nat Genet* **41**: 1061–1067.

Alvarez CE, Akey JM. 2012. Copy number variation in the domestic dog. *Mamm Genome* **23**: 144–163.

Ambrosini G, Groux R, Bucher P. 2018. PWMScan: a fast tool for scanning entire genomes with a position-specific weight matrix. *Bioinformatics* **34**: 2483–2484.

American Kennel Club. 2007. *The Complete Dog Book: 20th Edition*. Ballantine Books.

Amin J, Boche D, Rakic S. 2017. What do we know about the inflammasome in humans? *Brain Pathol* **27**: 192–204.

Arendt M, Fall T, Lindblad-Toh K, Axelsson E. 2014. Amylase activity is associated with AMY2B copy numbers in dog: implications for dog domestication, diet and diabetes. *Anim Genet* **45**: 716–722.

Benaglia T. 2009. *Mixtools: An R Package for Analyzing Finite Mixture Models*.

Berglund J, Nevalainen EM, Molin A-M, Perloski M, André C, Zody MC, Sharpe T, Hitte C, Lindblad-Toh K, Lohi H, et al. 2012. Novel origins of copy number variation in the dog genome. *Genome Biol* **13**: R73.

Bigham AW. 2016. Genetics of human origin and evolution: high-altitude adaptations. *Curr Opin Genet Dev* **41**: 8–13.

Björnerfeldt S, Webster MT, Vilà C. 2006. Relaxation of selective constraint on dog mitochondrial DNA following domestication. *Genome Res* **16**: 990–994.

Borge KS, Tønnessen R, Nødtvedt A, Indrebø A. 2011. Litter size at birth in purebred dogs—A retrospective study of 224 breeds. *Theriogenology* **75**: 911–919. http://dx.doi.org/10.1016/j.theriogenology.2010.10.034.

Brown EA, Dickinson PJ, Mansour T, Sturges BK, Aguilar M, Young AE, Korff C, Lind J, Ettinger CL, Varon S, et al. 2017. FGF4 retrogene on CFA12 is responsible for chondrodystrophy and intervertebral disc disease in dogs. *Proc Natl Acad Sci U S A* **114**: 11476–11481.

Brown KH, Dobrinski KP, Lee AS, Gokcumen O, Mills RE, Shi X, Chong WWS, Chen JYH, Yoo P, David S, et al. 2012. Extensive genetic diversity and substructuring among zebrafish strains revealed through copy number variant analysis. *Proc Natl Acad Sci U S A* **109**: 529–534.

Brzyski D, Peterson CB, Sobczyk P, Candès EJ, Bogdan M, Sabatti C. 2017. Controlling the Rate of GWAS False Discoveries. *Genetics* **205**: 61–75.

Chain FJJ, Feulner PGD, Panchal M, Eizaguirre C, Samonte IE, Kalbe M, Lenz TL, Stoll M, Bornberg-Bauer E, Milinski M, et al. 2014. Extensive copy-number variation of young genes across stickleback populations. *PLoS Genet* **10**: e1004830.

Chalamalasetty RB, Garriock RJ, Dunty WC Jr, Kennedy MW, Jailwala P, Si H, Yamaguchi TP. 2014. Mesogenin 1 is a master regulator of paraxial presomitic mesoderm differentiation. *Development* **141**: 4285–4297.

Chase K, Jones P, Martin A, Ostrander EA, Lark KG. 2009. Genetic mapping of fixed phenotypes: disease frequency as a breed characteristic. *J Hered* **100 Suppl 1**: S37–41.

Chen W-K, Swartz JD, Rush LJ, Alvarez CE. 2009. Mapping DNA structural variation in dogs. *Genome Res* **19**: 500–509.

Chen W, Liu Y, Li H, Chang S, Shu D, Zhang H, Chen F, Xie Q. 2015. Intronic deletions of tva receptor gene decrease the susceptibility to infection by avian sarcoma and leukosis virus subgroup A. *Sci Rep* **5**: 9900.

Chiusaroli R, Knobler H, Luxenburg C, Sanjay A, Granot-Attas S, Tiran Z, Miyazaki T, Harmelin A, Baron R, Elson A. 2004. Tyrosine phosphatase epsilon is a positive regulator of osteoclast function in vitro and in vivo. *Mol Biol Cell* **15**: 234–244.

Clark J-ABJ, Whalen D, Marshall HD. 2016. Genomic analysis of gum disease and hypertrichosis in foxes. *Genet Mol Res* **15**. http://dx.doi.org/10.4238/gmr.15025363.

Coelho LP, Kultima JR, Costea PI, Fournier C, Pan Y, Czarnecki-Maulden G, Hayward MR, Forslund SK, Schmidt TSB, Descombes P, et al. 2018. Similarity of the dog and human gut microbiomes in gene content and response to diet. *Microbiome* **6**: 72.

Coren S. 1994. *The intelligence of dogs: canine consciousness and capabilities*.

Day MJ. 1999. Possible immunodeficiency in related rottweiler dogs. *J Small Anim Pract* **40**: 561–568.

Deane-Coe PE, Chu ET, Slavney A, Boyko AR, Sams AJ. 2018. Direct-to-consumer DNA testing of 6,000 dogs reveals 98.6-kb duplication associated with blue eyes and heterochromia in Siberian Huskies. *PLoS Genet* **14**: e1007648.

Enshell-Seijffers D, Lindon C, Morgan BA. 2008. The serine protease Corin is a novel modifier of the Agouti pathway. *Development* **135**: 217–225.

Fotiadis D, Kanai Y, Palacín M. 2013. The SLC3 and SLC7 families of amino acid transporters. *Mol Aspects Med* **34**: 139–158.

Freedman AH, Gronau I, Schweizer RM, Ortega-Del Vecchyo D, Han E, Silva PM, Galaverni M, Fan Z, Marx P, Lorente-Galdos B, et al. 2014. Genome sequencing highlights the dynamic early history of dogs. *PLoS Genet* **10**: e1004016.

Gundry RL, Allard MW, Moretti TR, Honeycutt RL, Wilson MR, Monson KL, Foran DR. 2007. Mitochondrial DNA analysis of the domestic dog: control region variation within and among breeds. *J Forensic Sci* **52**: 562–572.

Haeussler M, Zweig AS, Tyner C, Speir ML, Rosenbloom KR, Raney BJ, Lee CM, Lee BT, Hinrichs AS, Gonzalez JN, et al. 2019. The UCSC Genome Browser database: 2019 update. *Nucleic Acids Res* **47**: D853–D858.

Hastings PJ, Lupski JR, Rosenberg SM, Ira G. 2009. Mechanisms of change in gene copy number. *Nat Rev Genet* **10**: 551–564.

Hayward JJ, Castelhano MG, Oliveira KC, Corey E, Balkman C, Baxter TL, Casal ML, Center SA, Fang M, Garrison SJ, et al. 2016. Complex disease and phenotype mapping in the domestic dog. *Nat Commun* **7**: 10460.

Hegele RA. 2007. Copy-number variations and human disease. *Am J Hum Genet* **81**: 414–5; author reply 415.

Hoeppner MP, Lundquist A, Pirun M, Meadows JRS, Zamani N, Johnson J, Sundström G, Cook A, FitzGerald MG, Swofford R, et al. 2014. An improved canine genome and a comprehensive catalogue of coding genes and non-coding transcripts. *PLoS One* **9**: e91172.

Hu H. 1999. Chemorepulsion of neuronal migration by Slit2 in the developing mammalian forebrain. *Neuron* **23**: 703–711.

Hunt SE, McLaren W, Gil L, Thormann A, Schuilenburg H, Sheppard D, Parton A, Armean IM, Trevanion SJ, Flicek P, et al. 2018. Ensembl variation resources. *Database* **2018**. http://dx.doi.org/10.1093/database/bay119.

Irion DN. 2003. Analysis of Genetic Variation in 28 Dog Breed Populations With 100 Microsatellite Markers. *J Hered* **94**: 81–87.

Jokinen P, Rusanen EM, Kennedy LJ, Lohi H. 2011. MHC class II risk haplotype associated with canine chronic superficial keratitis in German Shepherd dogs. *Vet Immunol Immunopathol* **140**: 37–41.

Jones P, Chase K, Martin A, Davern P, Ostrander EA, Lark KG. 2008. Single-nucleotide-polymorphism-based association mapping of dog stereotypes. *Genetics* **179**: 1033–1044.

Kader A, Li Y, Dong K, Irwin DM, Zhao Q, He X, Liu J, Pu Y, Gorkhali NA, Liu X, et al. 2015. Population Variation Reveals Independent Selection toward Small Body Size in Chinese Debao Pony. *Genome Biol Evol* **8**: 42–50.

Karyadi DM, Karlins E, Decker B, vonHoldt BM, Carpintero-Ramirez G, Parker HG, Wayne RK, Ostrander EA. 2013. A copy number variant at the KITLG locus likely confers risk for canine squamous cell carcinoma of the digit. *PLoS Genet* **9**: e1003409.

Kawaguchi A, Koch GG. 2015. sanon: AnRPackage for Stratified Analysis with Nonparametric Covariable Adjustment. *Journal of Statistical Software* **67**. http://dx.doi.org/10.18637/jss.v067.i09.

Kim KS, Lee SE, Jeong HW, Ha JH. 1998. The complete nucleotide sequence of the domestic dog (Canis familiaris) mitochondrial genome. *Mol Phylogenet Evol* **10**: 210–220.

Landis JR, Cooper MM, Kennedy T, Koch GG. 1979. A computer program for testing average partial association in three-way contingency tables (PARCAT). *Comput Programs Biomed* **9**: 223–246.

Layer RM, Chiang C, Quinlan AR, Hall IM. 2014. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol* **15**: R84.

Le Béguec C, Wucher V, Lagoutte L, Cadieu E, Botherel N, Hédan B, De Brito C, Guillory A-S, André C, Derrien T, et al. 2018. Characterisation and functional predictions of canine long non-coding RNAs. *Sci Rep* **8**: 13444.

Li J, Yang T, Wang L, Yan H, Zhang Y, Guo Y, Pan F, Zhang Z, Peng Y, Zhou Q, et al. 2009. Whole Genome Distribution and Ethnic Differentiation of Copy Number Variation in Caucasian and Asian Populations. *PLoS One* **4**: e7958.

Lindblad-Toh K, Wade CM, Mikkelsen TS, Karlsson EK, Jaffe DB, Kamal M, Clamp M, Chang JL, Kulbokas EJ 3rd, Zody MC, et al. 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438**: 803–819.

Li Y, Wang G-D, Wang M-S, Irwin DM, Wu D-D, Zhang Y-P. 2014. Domestication of the dog from the wolf was promoted by enhanced excitatory synaptic plasticity: a hypothesis. *Genome Biol Evol* **6**: 3115–3121.

Makino T, Rubin C-J, Carneiro M, Axelsson E, Andersson L, Webster MT. 2018. Elevated Proportions of Deleterious Genetic Variation in Domestic Animals and Plants. *Genome Biol Evol* **10**: 276–290.

Marco-Sola S, Sammeth M, Guigó R, Ribeca P. 2012. The GEM mapper: fast, accurate and versatile alignment by filtration. *Nat Methods* **9**: 1185–1188.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**: 10. http://dx.doi.org/10.14806/ej.17.1.200.

Miinalainen IJ, Schmitz W, Huotari A, Autio KJ, Soininen R, Ver Loren van Themaat E, Baes M, Herzig K-H, Conzelmann E, Hiltunen JK. 2009. Mitochondrial 2,4-dienoyl-CoA reductase deficiency in mice results in severe hypoglycemia with stress intolerance and unimpaired ketogenesis. *PLoS Genet* **5**: e1000543.

Molin A-M, Berglund J, Webster MT, Lindblad-Toh K. 2014. Genome-wide copy number variant discovery in dogs using the CanineHD genotyping array. *BMC Genomics* **15**: 210.

Morgane Ollivier, Anne Tresset, Fabiola Bastian, Laetitia Lagoutte, Erik Axelsson, Maja-Louise Arendt, Adrian Bălăşescu, Marjan Marshour, Mikhail V. Sablin, Laure Salanova, Jean-Denis Vigne, Christophe Hitte, Catherine Hänni. 2016. Amy2B copy number variation reveals starch diet adaptations in ancient European dogs. *https://royalsocietypublishing.org/doi/101098/rsos160449*. https://doi.org/10.1098/rsos.160449 (Accessed April 26, 2019).

Muñoz-Amatriaín M, Eichten SR, Wicker T, Richmond TA, Mascher M, Steuernagel B, Scholz U, Ariyadasa R, Spannagl M, Nussbaumer T, et al. 2013. Distribution, functional impact, and origin mechanisms of copy number variation in the barley genome. *Genome Biol* **14**: R58.

Natrajan MS, de la Fuente AG, Crawford AH, Linehan E, Nuñez V, Johnson KR, Wu T, Fitzgerald DC, Ricote M, Bielekova B, et al. 2015. Retinoid X receptor activation reverses age-related deficiencies in myelin debris phagocytosis and remyelination. *Brain* **138**: 3581–3597.

Ní Leathlobhair M, Perri AR, Irving-Pease EK, Witt KE, Linderholm A, Haile J, Lebrasseur O, Ameen C, Blick J, Boyko AR, et al. 2018. The evolutionary history of dogs in the Americas. *Science* **361**: 81–85.

Ostrander EA, Wayne RK. 2005. The canine genome. *Genome Res* **15**: 1706–1716.

Paradis E, Claude J, Strimmer K. 2004. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20**: 289–290. http://dx.doi.org/10.1093/bioinformatics/btg412.

Parker HG, Dreger DL, Rimbault M, Davis BW, Mullen AB, Carpintero-Ramirez G, Ostrander EA. 2017. Genomic Analyses Reveal the Influence of Geographic Origin, Migration, and Hybridization on Modern Dog Breed Development. *Cell Rep* **19**: 697–708.

Parker HG, VonHoldt BM, Quignon P, Margulies EH, Shao S, Mosher DS, Spady TC, Elkahloun A, Cargill M, Jones PG, et al. 2009. An expressed fgf4 retrogene is associated with breed-defining chondrodysplasia in domestic dogs. *Science* **325**: 995–998.

Plassais J, Kim J, Davis BW, Karyadi DM, Hogan AN, Harris AC, Decker B, Parker HG, Ostrander EA. 2019. Whole genome sequencing of canids reveals genomic regions under selection and

variants influencing morphology. *Nat Commun* **10**: 1489.

Polley S, Cipriani V, Khan JC, Shahid H, Moore AT, Yates JRW, Hollox EJ. 2016. Analysis of copy number variation at DMBT1 and age-related macular degeneration. *BMC Med Genet* **17**: 44.

Price AL, Zaitlen NA, Reich D, Patterson N. 2010. New approaches to population stratification in genome-wide association studies. *Nat Rev Genet* **11**: 459–463.

Quilez J, Martínez V, Woolliams JA, Sanchez A, Pong-Wong R, Kennedy LJ, Quinnell RJ, Ollier WER, Roura X, Ferrer L, et al. 2012. Genetic Control of Canine Leishmaniasis: Genome-Wide Association Study and Genomic Selection Analysis. *PLoS One* **7**: e35349.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–842.

Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, et al. 2006. Global variation in copy number in the human genome. *Nature* **444**: 444–454.

Rigau M, Juan D, Valencia A, Rico D. 2019. Intronic CNVs and gene expression variation in human populations. *PLoS Genet* **15**: e1007902.

Rimbault M, Beale HC, Schoenebeck JJ, Hoopes BC, Allen JJ, Kilroy-Glynn P, Wayne RK, Sutter NB, Ostrander EA. 2013. Derived variants at six genes explain nearly half of size reduction in dog breeds. *Genome Res* **23**: 1985–1995.

Schmidt D, Schwalie PC, Wilson MD, Ballester B, Gonçalves A, Kutter C, Brown GD, Marshall A, Flicek P, Odom DT. 2012. Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell* **148**: 335–348.

Schoenebeck JJ, Hutchinson SA, Byers A, Beale HC, Carrington B, Faden DL, Rimbault M, Decker B, Kidd JM, Sood R, et al. 2012. Variation of BMP3 Contributes to Dog Breed Skull Diversity. *PLoS Genet* **8**: e1002849.

Serres-Armero A, Povolotskaya IS, Quilez J, Ramirez O, Santpere G, Kuderna LFK, Hernandez-Rodriguez J, Fernandez-Callejo M, Gomez-Sanchez D, Freedman AH, et al. 2017. Similar genomic proportions of copy number variation within gray wolves and modern dog breeds inferred from whole genome sequencing. *BMC Genomics* **18**: 977.

Shubkina AV, Severtsov AS, Chepeleva KV. 2012. Factors influencing the hunting success of the predator: A model with sighthounds. *Biology Bulletin* **39**: 65–76. http://dx.doi.org/10.1134/s1062359012010074.

Skoglund P, Ersmark E, Palkopoulou E, Dalén L. 2015. Ancient wolf genome reveals an early divergence of domestic dog ancestors and admixture into high-latitude breeds. *Curr Biol* **25**: 1515–1519.

Sudharsan R, Beiting DP, Aguirre GD, Beltran WA. 2018. Author Correction: Involvement of Innate Immune System in Late Stages of Inherited Photoreceptor Degeneration. *Sci Rep* **8**: 17041.

Sudmant PH, Mallick S, Nelson BJ, Hormozdiari F, Krumm N, Huddleston J, Coe BP, Baker C, Nordenfelt S, Bamshad M, et al. 2015a. Global diversity, population stratification, and selection of human copy-number variation. *Science* **349**: aab3761.

Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Fritz MH-Y, et al. 2015b. An integrated map of structural variation in 2,504 human genomes. *Nature* **526**: 75–81.

Thuault SJ, Malleret G, Constantinople CM, Nicholls R, Chen I, Zhu J, Panteleyev A, Vronskaya S, Nolan MF, Bruno R, et al. 2013. Prefrontal cortex HCN1 channels enable intrinsic persistent neural firing and executive memory function. *J Neurosci* **33**: 13583–13599.

Trost B, Walker S, Wang Z, Thiruvahindrapuram B, MacDonald JR, Sung WWL, Pereira SL, Whitney J, Chan AJS, Pellecchia G, et al. 2018. A Comprehensive Workflow for Read Depth-Based Identification of Copy-Number Variation from Whole-Genome Sequence Data. *Am J Hum Genet* **102**: 142–155.

Tsepilov YA, Ried JS, Strauch K, Grallert H, van Duijn CM, Axenovich TI, Aulchenko YS. 2013. Development and Application of Genomic Control Methods for Genome-Wide Association Studies Using Non-Additive Models. *PLoS One* **8**: e81431.

Upadhyay M, da Silva VH, Megens H-J, Visker MHPW, Ajmone-Marsan P, Bâlteanu VA, Dunner S,

Garcia JF, Ginja C, Kantanen J, et al. 2017. Distribution and Functionality of Copy Number Variation across European Cattle Populations. *Front Genet* **8**: 108.

van der Borg Elisabeth A. M. Graat BonneBeerda JAM. 2017. Behavioural testing based breeding policy reduces the prevalence of fear and aggression related behaviour in Rottweilers. https://www.sciencedirect.com/science/article/pii/S0168159117301818?via%3Dihub (Accessed April 26, 2019).

Vaysse A, Ratnakumar A, Derrien T, Axelsson E, Pielberg GR, Sigurdsson S, Fall T, Seppälä EH, Hansen MST, Lawley CT, et al. 2011. Identification of Genomic Regions Associated with Phenotypic Variation between Dog Breeds using Selection Mapping. *PLoS Genetics* **7**: e1002316. http://dx.doi.org/10.1371/journal.pgen.1002316.

Vietri Rudan M, Barrington C, Henderson S, Ernst C, Odom DT, Tanay A, Hadjur S. 2015. Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture. *Cell Rep* **10**: 1297–1309.

Vilà C, Maldonado JE, Wayne RK. 1999. Phylogenetic relationships, evolution, and genetic diversity of the domestic dog. *J Hered* **90**: 71–77.

Waldo JT, Diaz KS. 2015. Development and validation of a diagnostic test for Ridge allele copy number in Rhodesian Ridgeback dogs. *Canine Genet Epidemiol* **2**: 2.

Wang X, Zhou B-W, Yin T-T, Chen F-L, Esmailizadeh A, Turner MM, Poyarkov AD, Savolainen P, Wang G-D, Fu Q, et al. 2018. Canine transmissible venereal tumor genome reveals ancient introgression from coyotes to arctic sled dogs. *Evolutionary Biology*.

Wellcome Trust Case Control Consortium, Craddock N, Hurles ME, Cardin N, Pearson RD, Plagnol V, Robson S, Vukcevic D, Barnes C, Conrad DF, et al. 2010. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature* **464**: 713–720.

Wiberg ME, Saari SA, Westermarck E, Meri S. 2000. Cellular and humoral immune responses in atrophic lymphocytic pancreatitis in German shepherd dogs and rough-coated collies. *Vet Immunol Immunopathol* **76**: 103–115.

Wu C, DeWan A, Hoh J, Wang Z. 2011. A comparison of association methods correcting for population stratification in case-control studies. *Ann Hum Genet* **75**: 418–427.

Wucher V, Legeai F, Hédan B, Rizk G, Lagoutte L, Leeb T, Jagannathan V, Cadieu E, David A, Lohi H, et al. 2017. FEELnc: a tool for long non-coding RNA annotation and its application to the dog transcriptome. *Nucleic Acids Res* **45**: e57.

Xu L, Hou Y, Bickhart DM, Zhou Y, Hay EHA, Song J, Sonstegard TS, Van Tassell CP, Liu GE. 2016. Population-genetic properties of differentiated copy number variations in cattle. *Sci Rep* **6**: 23161.

Xu X, Dong G-X, Schmidt-Küntzel A, Zhang X-L, Zhuang Y, Fang R, Sun X, Hu X-S, Zhang T-Y, Yang H-D, et al. 2017. The genetics of tiger pelage color variations. *Cell Res* **27**: 954–957.

Yalcin B, Wong K, Agam A, Goodson M, Keane TM, Gan X, Nellåker C, Goodstadt L, Nicod J, Bhomra A, et al. 2011. Sequence-based characterization of structural variation in the mouse genome. *Nature* **477**: 326–329.

Zhang F, Wang Y, Deng H-W. 2008. Comparison of Population-Based Association Study Methods Correcting for Population Stratification. *PLoS One* **3**: e3392.

Zhao M, Wang Q, Wang Q, Jia P, Zhao Z. 2013. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinformatics* **14 Suppl 11**: S1.

4. DISCUSSION

This work has aimed to study the dynamics and contribution of copy number variation to an extensive panel of dog and wolf genomes at both a global and fine scale. The initial assessment of duplication differences between dogs, wolves and other wild canids revealed the surprising presence of gene-enriched, low copy number loci private to dogs which contributed to maintain a similar overall number of CNV loci across all species. The resilience of duplications to the harsh population bottlenecks of dog domestication is remarkable since other kinds of genomic variation have been reported to have experienced a significant decline (Serres-Armero et al. 2017; Freedman et al. 2014). This result was consequently followed up with an assessment of whether dog-specific CNVs are equally distributed across all dog breeds and if they correlate with dog phenotypes. We report significant cases where breed-specific phenotypic variance can be explained by CNV events and open the way for more specific functional assays. Overall, a substantial contribution of CNV to dog phenotype is observed, especially considering the lower rate of occurrence of CNV in comparison to SNV events. However, the aggregated effect of known CNV to morphometrics remains second to that of SNV.

The ability of CNVs to recapitulate the canine phylogeny and the cladistics breed creation was also assessed here. CNVs can be used to classify dogs and wolves, but they do not seem to completely adhere to the accepted phylogenetic tree built using high-density genetic markers and historical records. Similar difficulties to accurately reproduce the

most recent dog breed phylogeny can be seen in previous studies performed with a smaller set of markers or fewer individuals (Vonholdt et al. 2010).

4.1. Similar genomic proportions of copy number variation within gray wolves and modern dog breeds inferred from whole-genome sequencing

Since the creation of the first dog assembly in the early 2000s (Lindblad-Toh et al. 2005), few global whole-genome sequencing CNV studies have been performed in this organism. Most novel CNV discovery since then has been performed using tiling arrays (Nicholas et al. 2011, 2009), which can potentially neglect small CNVs and are generally less exhaustive than WGS. WGS screening of canine structural variation in our study showed high concordance with both previous aCGH and WGAC. Additionally, a substantial sharing of over 80% of all segmental duplications was observed across dogs, wolves, coyotes, and jackals, concordant with a basal loss of *PRDM9* to the Canidae family which would have had a great impact in CNV origination mechanisms (Paigen and Petkov 2018). An important remark is that all canid sequences were mapped against the dog reference, which diverged from the most distant sample present in the panel -a jackal- between 3 and 4 million years ago (Lindblad-Toh et al. 2005). Although this is common practice, and much earlier diverging species have been mapped to a single genome, even if specific references were available (Sudmant et al. 2013), important CNV dropouts could be resulting from this practice (Sherman et al. 2019). However, since most of our analyses revolved around dogs and wolves,

whose divergence time does not surpass 50,000 years, our claims should be robust to this bias. It must be noted that the publication of the wolf genome assembly (Gopalakrishnan et al. 2017) coincided with the review period of this work, but replication of these results using the wolf as a reference would add substantial evidence to its claims. The comparison of genomic proportions of different repeat families between the dog and wolf assemblies, which is theorized to correlate well with CNV (Brahmachary et al. 2014), shows that dogs systematically present a bigger or equal proportion of each of the categories than wolves and therefore backs our results.

Similar proportions of genomic CNV were found between dogs and wolves, both in terms of the number of CNV loci and in terms of the amount of variability within those loci. This study contains a balanced set of dogs, comprising recently originated and ancient breeds, and a similarly balanced set of gray wolves from all over Eurasia and America and therefore the sampling is unlikely to play a role in this effect. On the contrary, the worldwide diversity of gray wolves should outweigh that of bottlenecked dogs, as observed in the almost doubled SNP diversity of wolves. As an additional precaution, the number of samples was chosen to be equal in both subspecies and the sequencing qualities were assessed to be comparable across all individuals to minimize technical bias. The CNV proportions remained the same for all CN ranges -low, middle and high- and were invariable to bootstrapping and singleton effects. Perhaps even more interestingly, these CNV loci, specifically duplications within the low copy number range, were enriched in genes and private to dogs. Since low CN instances are projected to lose variability much more easily

and are also expected to be younger, this preservation of variability could be an indication of artificial selection in favor of potentially functional CNV loci. The opposite argument could be made as well: if CNVs are expected to be deleterious and purged from free-ranging populations, the relaxation in selective pressures in domesticates could facilitate their accumulation. Finally, naturally evolving CNVs could be more resilient to the loss of diversity just because of their bigger number of segregating alleles. Some evidence against this last hypothesis is presented by finding constitutively low genomic Vst values and by the previously reported excess of low CN loci in dogs.

Lastly, this global approach made it possible to recapitulate most previous relevant CNV findings (Reiter, Jagoda, and Capellini 2016; Nicholas et al. 2009) and propose two mostly unexplored CNV-differentiated loci: the *SIRP* gene cluster in chr24:19,200,000-19,400,000 and the *CBR1* gene.

4.2. Dog breed variation in genomic copy number underlies complex and novel phenotype associations

If dogs have maintained a significant amount of potentially functional copy number variation, the exploration of how it is distributed across dog breeds and how it might impact dog physiology should ensue. Even if a few specific CNVs have been associated with dog traits in the past, to date, no formal CNV-GWAS has been performed in dogs. Besides discovering new interesting CNV loci, finding associations with crucial

domestication traits such as behavior, snout morphology, tail length or body size could contribute new evidence that CNVs were indeed selected for during breed origination.

CNV-GWAS replicates most of the previously known CNV related trait associations and finds a few novel interesting genic CNVs such as *UNCX* associated with tail length and *DMBT1* and *CTNNA3* with eye disorders. Remarkably, a poorly characterized, low effect SNP variant which had been loosely associated to body size in previous studies (Hayward et al. 2016) is close to a significantly associated duplication with high effect in our study -comparable to *SMAD2*, a well-characterized deletion which is present in many small size breeds-. This duplication is located 40 kbp upstream of the *MED13L* gene, in a block of mammalian synteny containing genes *TBX5*, *TBX3* and *MED13L* and, additionally, is found inside a Hi-C contact region between genes *TBX3* and *MED13L*. Of note, *TBX3* has been associated with the size reduction of Debao ponys (Kader et al. 2015), and haploinsufficiency in this block of synteny has been associated with syndromes which potentially involve size reduction in humans (van Weerd et al. 2014). Similarly, a few CNVs associated with hair length are found in significant Hi-C interactions with the *MAP2K6* gene, which is known to be a cause for hypertrichosis in humans.

Interestingly, most of the associated structural variants in our study overlap non-coding, functional elements such as lncRNA or CpG islands. We report a significant concordance between the tissue-specific expression of these genomic variants and the phenotype they are associated with. For example, the lncRNA CNVs associated with

intelligence are primarily expressed in the brain, while those associated with litter size are mostly expressed in the testes. These cases of phenotype concordance with tissue expression point to complex, oligogenic effects. However, since they account for a smaller variance than most SNPs found in previous GWAS, they can hardly be the only causative effect acting on the phenotype. On the contrary, these CNVs could potentially be acting on phenotype modulation and regulation, but functional analyses should be conducted in order to disentangle the molecular mechanisms in which they interact with the phenotype.

Future prospects concerning this project would involve using non-domestic dogs and wolves to assess the ancestral state of the associated CNV loci. This could provide more insight into whether the associated CNVs are private to dogs or shared with their closely related ancestors. More importantly, an enrichment analysis of private dog associated CNVs against dog-wolf shared associated CNV could provide new evidence on whether the domestication process favored the accumulation of functional copy number variation in dogs. Principal component analysis of CN genotypes shows a separation between domestic and non-domestic dogs, which at least indicates that there are some differentiated CNVs between the two groups. However, the comparison is not trivial, since village dog samples show generally lower genotype qualities, smaller read depths and have been shown to potentially display more deletion due to artifacts. Hence, increasing the number and quality of samples from this group would be helpful to assess the role of CNVs in dog domestication.

4.3. Methodological considerations

This work features the largest panel of whole-genome sequencing data used to estimate CNV in purebred dogs. During the whole study, over 500 canine genomes, belonging to more than 100 pure breeds, feral dogs and wild canids, were scanned for CNV. All these genomes were gathered from publicly available resources and lack any metadata or phenotypes, their only available information being breed identity. As such, there is an inherent heterogeneity in read depths, read lengths, sample origins, modes of sequencing and even year of production which may affect sample-specific CNV genotype qualities. Even if CNV quality was controlled for, these effects could potentially correlate with the recently reported deletion guanine-cytosine biases reported in other dog CNV analyses (Kidd 17 Sep, 2018 - 20 Sep, 2018) which are also observable in some of our samples. Even if this effect is not strong enough to sway properly corrected association analyses, it might be hampering our ability to successfully reconstruct breed phylogenies.

4.3.1. Advantages and limitations of Hidden Markov models for the inference of CN

The development of a probabilistic framework to distinguish CN across different individuals was instrumental to confidently explore the hypothesis that the proportion of CNV loci is comparable between wild canids and domestic dogs. However, it is important to analyze its methodological implications, caveats and possible extensions.

Most importantly, the idea of using a HMM only makes sense if the read

depth observations are spatially correlated. Since we are *a priori* partitioning the genome for normalization and correction purposes, the resolution of this *a priori* segmentation will indirectly influence the predictive abilities of the model. In other words, if the genome is segmented into windows large enough so that most of the structural variants are expected to fit into one or just a few, a first-order HMM will no longer be the optimal choice for an accurate probabilistic prediction. Moreover, there is no certainty that the HMM parameters are invulnerable to segmentation changes. If different window resolutions are to be assessed, the model parameters need to be re-estimated for each possible attempt.

Another important remark needs to be made on the sparsity of CNV instances. Many “incomplete” vertebrate assemblies such as the horse or the dog tend to have most CNV and other complex regions aggregate into non-chromosomal scaffolds. This undermines the representation of CNV regions within chromosomal scaffolds and alters the biological order of CNV transitions. Besides affecting parameter estimation, this phenomenon is likely to produce cumulative probability value underflow and very low estimated transition probabilities, which are difficult to process computationally. Implementing partial hidden Markov chains -breaking down the sequence of read depth observations into smaller, independent subsequences-, and applying sparse data methods (Bicego, Cristani, and Murino 2007) for parameter estimation is likely to enhance computational performance and aid prediction accuracy.

On the subject of *a priori* genome partitioning, another essential

consideration is related to the rigidity of the window creation process. Since the genome is segmented into fixed-length windows, it is very likely that many windows will partially overlap both CNV and non-CNV regions, effectively containing a part of each. This will produce aberrant read depth measures which can potentially sway the absolute CN values of short structural variants and hinder the HMM probability calculations.

The issues explained above should be alleviated by picking smaller genome partition sizes or just by completely avoiding genome partitions. Genome partitioning was established by early heuristic CNV calling software such as CNVnator (Abyzov et al. 2011) or mrCaNaVar (Alkan et al. 2009) as a means to ensure that the program could run efficiently and quickly in any computer. This is not required in our current setup, an HMM can take care of genomic segmentation in single samples at a quasi-base-pair resolution without the use of parallel computing or heavy virtual memory requirements, especially if partial HMM chains are implemented. New issues would be expected to arise from a more fine-grained approach, particularly, differences in sample quality and coverage are much more noticeable at almost single base-pair levels. These could make parameter estimation and sample regularization much more complex and unwieldy. Nevertheless, the probabilistic nature of the HMM together with posterior populational Bayes re-genotyping could be expected to partially tackle this problem. Another heuristic, easily implementable, approaches to the fixed window length problem could involve performing multiple genome segmentations with different offsets, or redefining window boundaries for windows with unexpectedly uninformative posteriors.

Altogether, much like many other omic problems nowadays, CNV estimation must strike a balance between resolution and accuracy. Low-resolution CNV calls will not be able to fully benefit from the advantages of an HMM and will neglect small SV, but they can be easily compared across samples and the differences that are found will be reliable. High-resolution CNV calls will be subjected to more noise and false positives, but an HMM should be able to quantify this uncertainty. Additionally, other features besides read depth, such as allele balance or variant density, might be integrable with read depth methods to increase robustness and accuracy at smaller base-pair resolution.

4.3.2. Methodological extensions

Read depth is one of the most robust indicators of copy number change in single samples, however, it can show great cross-sample variance and specific QC and sequence biases. The consideration of semi-orthogonal factors into the CN calculation might thereby alleviate some of these issues while introducing very little extra computational burdens.

Nowadays, allele balances (and therefore, variant densities[6]) are almost trivially calculated from mapped data. Allele balances (AB), also called variant allele fractions, are the proportion of reads supporting a variant in a specific locus for a certain sample. In a canonical heterozygous diploid position, the fraction of reads supporting either allele should theoretically

[6] Variants do not need to be formally called using a genotyper. Only the presence reads supporting alternatives alleles in the mapped file should be assessed.

be around 50%, while in a homozygous position no reads supporting two different alleles should ever be observed. Allele balances and read depths can be calculated in roughly 1.7 times[7] the time it would take to calculate read depth alone. However, since the raw calculation needs to be coupled with more intensive operations such as normalization, the time differential of the two calculations should be reduced even further.

Read depth methods are based on the assumption that unresolved genomic sequences will be mapped to their most identical counterpart which is represented in the assembly in proportion to their absolute number. As such, all variation from these unresolved regions will be projected into the mappings in similar proportions, increasing the density of variant occurrence. Moreover, the fraction of reads containing a certain allele should decrease in accordance with the region's CN, or at least be concordant with any integer fraction of said CN. All these features can be either independently or jointly used for absolute CN prediction and to assess CN concordance across samples. Indeed, some variant calling software already take these factors into consideration. Specifically, the latest incarnations of Haplotypecaller, one of the most used callers nowadays, avoid confidently calling variation on hypervariable regions with abnormal allele balances (Wu et al. 2017).

[7] Estimates based on 10 replicates on different files using the samtools suite 1.3 in a 2.6 GHz CPU.

For variation to be used to aid read depth CN calculation, a compound probability measure can be calculated from the required features, which could additionally be arbitrarily down-weighted by the expected noise-to-signal ratio of each feature. This is equivalent to an HMM with multivariate observations and a single hidden state, however, more sophisticated relations and dependencies between the features could be built into a Bayesian network.

4.3.2.1. Modelling allele balances

As introduced above, the number of copies of a long collapsed assembly segment at a relatively high coverage should, in theory, be determinable based on the fractions of the alleles found in it. For example a region with CN=7 -reads from 7 different regions map to it-, would be expected to contain variants at fractions 1/7, 2/7, 3/7, etc. These fractions are not trivially independent from read depth, since the denominator of the fraction corresponds to the RD itself. However statistical independence can be claimed assuming Lukacs's proportion-sum independence theorem (Lukacs 1955) if read depth is high enough (between 10 and 15X) that allele count distributions are expected to be bell-shaped.

If each possible fraction is expected to appear mostly at random -i.e it is similarly likely to observe alleles at fraction 1/7 or at fraction 4/7 for a CN=7 region-, then the CN can be predicted using the Fourier Transform (FT). Variance in the AB fraction values will be observed as frequency phase shifts and can be quantified using the imaginary part of the transform (Figure 8). The FT real values can be normalized into

probabilities and independently included in the compound model.

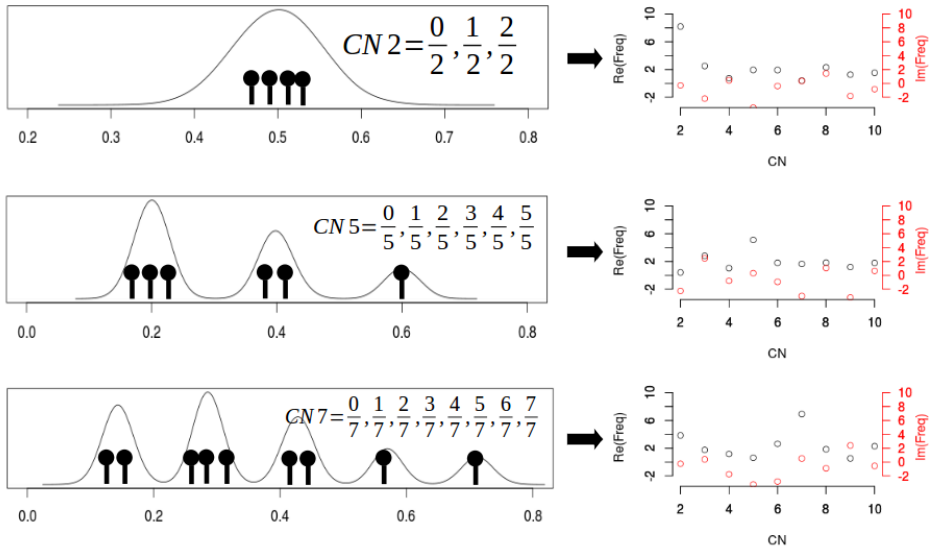

Figure 8: Cartoon of the allele balance modelling using the Fourier Transform. Right: Examples of possible allele balance densities which could be found in a CN 2,5 and 7 locus respectively. Note that a minimum number of measurements will be needed in order to reliably perform the transform. **Left:** Simulations of FT outputs considering 30 polymorphic sites which can belong to any possible fraction with a N(0,0.05) jitter on the AB measurements. Note how the real part of the transform (black) becomes higher while the imaginary part (red) remains close to 0 for the "real" CN.

However, allele fractions are not necessarily expected to appear at random. Quite the opposite, when a diploid, heterozygous locus duplicates, only one of the two alleles will be copied. If the same locus happens to duplicate again, it is twice as likely that the initially duplicated allele will duplicate again because it is represented twice as much. This effect scales with CN, making it more likely to find small (or big) AB fractions and less likely to find medium AB fractions. Furthermore, if the variation was not present before the first duplication happened, it will be introduced as the smallest possible fraction, which is

to say that new mutations can only appear in one of the copies at a time and will produce low AB.

Preferential attachment processes, also known as *rich get richer* processes, comprise models where the probability of a certain event happening increases in proportion with its frequency (Figure 9). In other words, the most likely events will grow even more likely as the process moves forward. This process greatly resembles the evolution of duplicated alleles described above and can be modeled with a distribution named beta-binomial. The multivariate version of a beta-binomial distribution, named Dirichlet-multinomial distribution, generalizes into a preferential attachment process called a Dirichlet process, which could be analogous to a process of duplication with mutation. However, there is a small difference between the Dirichlet process and the duplication model: the probability of introducing new variation in a Dirichlet process increases or decreases according to preferential attachment, while in a duplication model it could be considered constant. The constant introduction of variation, together with the fact that the number of variation categories is initially bounded should make it possible to transform the Dirichlet process into a Dirichlet-multinomial distribution, which can potentially be treated analytically much more easily than the Dirichlet process version of the model.

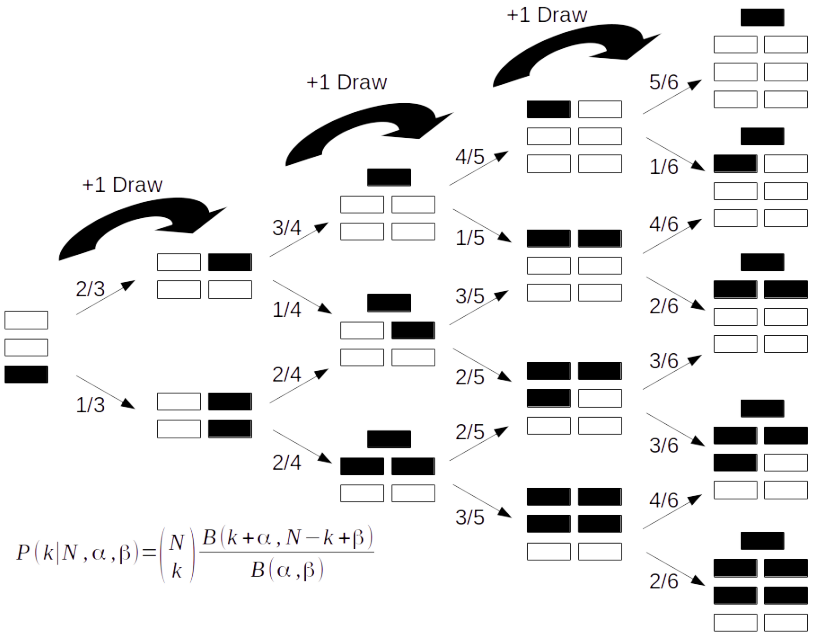

Figure 9: Cartoon example of a Beta-Binomial distribution. An analogy could be traced to the evolution of an locus with two alleles (white and black) after 4 duplications without the ability to mutate. The formula on the bottom left is the density function. α: initial number of white "alleles" (2). β: initial number of black "alleles" (1). N: total final number of "alleles" after duplication (7). k: final number of white "alleles" whose probability is being assessed.

Much like the negative binomial distribution, the negative Dirichlet-multinomial distribution measures the probability of the number of trials needed to reach a certain category configuration in a preferential attachment process. Since each trial in the analogous Dirichlet-Multinomial process proportionally corresponds to a CN increase, the most likely number of trials should be proportional to the true CN of the region. Note that, since the input to the model are allele counts and not fractions, this model is not independent of read depth, but rather distributes the depth into allele categories in a preferential attachment manner. This is a way to integrate RD and AB into a single probability

measure, which can be directly applied to our current HMM schema.

4.3.2.2. Modeling Variant densities

Variant densities can be modeled either by counting the number of variants in arbitrarily long genome segments or by storing the genomic distances from each assayed variable to the next. The most easily implementable model considering the current setup is plain variant counting, since the genome is already segmented into fixed-length windows. The variant counts in each window could be naively assumed to be Poisson-distributed with potentially variable rates and therefore, quantized rate changes between windows -which would correspond to copy number changes- could be assigned a likelihood using Poisson densities.

However, the counting approach will become less effective the smaller the windows get, instead, the distance between variants can be used for finer-grain approaches. This will produce a continuous observation that is inversely proportional to the absolute CN, and could potentially be modeled as the waiting time of a Poisson process using exponential or geometric distributions. This second approach combined with a Markov chain would yield a continuous-time Markov chain, where changes in waiting times are linked to state changes.

4.3.2.3. Comparisons across samples

Where read depth can show great variance across the same loci in different samples, the patterns of variation -i.e. the genomic position of the variants and their allele balances- should be better indicators of

whether two individuals have identical or different CN states. If it can be proved that two individuals have relatively similar allele balances at roughly the same one to one positions, it is really probable that they will both have the same CN. The probability that two individuals have the same allele balances can be measured using statistical metrics such as the Bhattacharyya distance or the Kullback–Leibler divergence during the re-genotyping step of our pipeline (Supplementary Figure 1, Section 3.1. [Methods]).

4.4. Concluding remarks

An in-depth exploration of copy number variation in dogs and wolves has been carried out in this thesis, its main value being the use of whole genome sequencing in place of less effective techniques. Additionally, the most extensive panel of whole genome samples in the context CN discovery has been analyzed. Most previously reported CNV events have been reproduced in our dataset and a few novel and promising CNV instances have been found.

By surveying over 20,000 CNV events in the whole dataset, we confirm that the most impactful CNV events had already been characterized over two decades of work in the field of canine genomics. WGS offers a cost-effective way to genotype known CNVs and to discover potentially novel variation, however, the resolution that can be confidently assessed using WGS is still well above the 1,000 base-pairs. As such, any newly discovered variants are likely to be infrequent or else they would have been previously picked up by classical methods. Therefore, future major CNV discoveries in this organism will most likely sprout from the study

of smaller CNV events, which are too short to produce distinguishable read depth changes but too long to be entirely spanned by short reads. Third generation sequencing technologies hold a lot of promise in this regard, as they excel at genotyping structural variation within this range.

On a similar note, long read assembly updates of the dog genome might dramatically improve the CNV discovery rate in this organism. Currently, the dog assembly unknown chromosome contains over 83 Mbp of unplaced, mostly complex sequence which is hardly accessible by CNV analyses. The correct assembly of this sequence might not only enable the discovery of CNVs in their proper genomic context, but also the classification of segmental duplications which are currently collapsed in the assembly. This will lead to more robust structural variation analyses, especially in the context of population genetics and trait mapping.

This work has risen to the challenge of reliably and systematically assessing copy number variation in canine whole genomes. Through our methodological advances, we have been able to confidently determine cross-sample copy number differences, making it easier to perform classification and enhancing the predictive value of our trait associations. As such, disease-associated CNVs that could potentially contribute to eye disease propensities have been discovered. In the advent of commercial genetic testing, should these CNVs be easily assayable, they could potentially be included in more exhaustive test panels. Additionally, we have discovered novel interactions between copy number variants and other genomic variants which could be explored and validated in further studies.

Annexes

List of contributions to other publications during the span of this thesis (some lists of authors are truncated for readability purposes):

Fan, Zhenxin, Pedro Silva, Ilan Gronau, Shuoguo Wang, **Aitor Serres-Armero**, Rena M. Schweizer, Oscar Ramirez, et al. 2016. "Worldwide Patterns of Genomic Variation and Admixture in Gray Wolves." *Genome Research* 26 (2): 163–73.

Kuderna, Lukas F. K., Chad Tomlinson, Ladeana W. Hillier, Annabel Tran, Ian T. Fiddes, Joel Armstrong, Hafid Laayouni, David Gordon, John Huddleston, Raquel Garcia Perez, Inna Povolotskaya, **Aitor Serres-Armero**, …, et al. 2017. "A 3-Way Hybrid Approach to Generate a New High-Quality Chimpanzee Reference Genome (Pan_tro_3.0)." *GigaScience* 6 (11): 1–6.

Librado, Pablo, Cristina Gamba, Charleen Gaunitz, Clio Der Sarkissian, Mélanie Pruvost, Anders Albrechtsen, Antoine Fages, …, **Aitor Serres-Armero**, …, et al. 2017. "Ancient Genomic Changes Associated with Domestication of the Horse." *Science* 356 (6336): 442–45.

Gopalakrishnan, Shyam, Mikkel-Holger S. Sinding, Jazmín Ramos-Madrigal, Jonas Niemann, Jose A. Samaniego Castruita, Filipe G. Vieira, Christian Carøe, ... **Aitor Serres-Armero**, …, et al. 2018. "Interspecific Gene Flow Shaped the Evolution of the Genus Canis." *Current Biology: CB* 28 (21): 3441–49.e5.

Fages Antoine, Kristian Hanghøj, Naveed Khan, Charleen Gaunitz, Andaine Seguin-Orlando, Michela Leonardi, Christian McCrory Constantz, …, **Aitor Serres-Armero**, …et al. 2019. "Tracking Five Millennia of Horse Management with Extensive Ancient Genome Time Series." *Cell* 177 (6): 1419–35.e31.

Kuderna, Lukas F. K., Esther Lizano, Eva Julià, Jessica Gomez-Garrido, **Aitor Serres-Armero**, Martin Kuhlwilm, Regina Antoni Alandes, et al. 2019. "Selective Single Molecule Sequencing and Assembly of a Human Y Chromosome of African Origin." *Nature Communications* 10 (1): 4.

Gelabert, Pere, Marcela Sandoval-Velasco, **Aitor Serres-Armero**, Marc

de Manuel, Pere Renom, Ashot Margaryan, Toni de-Dios, et al. n.d. “Evolutionary History, Genomic Adaptation to Toxic Diet and Extinction of the Carolina Parakeet.” *SSRN Electronic Journal (Accepted in Current Biology)*. https://doi.org/10.2139/ssrn.3417467.

Bibliography

Abyzov, Alexej, Alexander E. Urban, Michael Snyder, and Mark Gerstein. 2011. “CNVnator: An Approach to Discover, Genotype, and Characterize Typical and Atypical CNVs from Family and Population Genome Sequencing.” *Genome Research* 21 (6): 974–84.

Afzali, Behdad, Juha Grönholm, Jana Vandrovcova, Charlotte O’Brien, Hong-Wei Sun, Ine Vanderleyden, Fred P. Davis, et al. 2017. “BACH2 Immunodeficiency Illustrates an Association between Super-Enhancers and Haploinsufficiency.” *Nature Immunology* 18 (7): 813–23.

Alkan, Can, Jeffrey M. Kidd, Tomas Marques-Bonet, Gozde Aksay, Francesca Antonacci, Fereydoun Hormozdiari, Jacob O. Kitzman, et al. 2009. “Personalized Copy Number and Segmental Duplication Maps Using next-Generation Sequencing.” *Nature Genetics* 41 (10): 1061–67.

American Veterinary Medical Association. 01 Jan, 2016. “2016 AVMA Report onTHE MARKET FOR VETERINARIANS.” https://www.aavmc.org/assets/site_18/files/annual %20reports/v3_econ_2016_report3_mketvet_061416.pdf.

Arendt, M., K. M. Cairns, J. W. O. Ballard, P. Savolainen, and E. Axelsson. 2016. “Diet Adaptation in Dog Reflects Spread of Prehistoric Agriculture.” *Heredity* 117 (5): 301–6.

Bailey, J. A., A. M. Yavor, H. F. Massa, B. J. Trask, and E. E. Eichler. 2001. “Segmental Duplications: Organization and Impact within the Current Human Genome Project Assembly.” *Genome Research* 11 (6): 1005–17.

Bailey, Jeffrey A., Deanna M. Church, Mario Ventura, Mariano Rocchi, and Evan E. Eichler. 2004. “Analysis of Segmental Duplications and Genome Assembly in the Mouse.” *Genome Research* 14 (5): 789–801.

Bailey, Jeffrey A., Zhiping Gu, Royden A. Clark, Knut Reinert, Rhea V. Samonte, Stuart Schwartz, Mark D. Adams, Eugene W. Myers, Peter W. Li, and Evan E. Eichler. 2002. “Recent Segmental Duplications in the Human

Genome.” *Science* 297 (5583): 1003–7.

Baker, Lauren A., Brian Kirkpatrick, Guilherme J. M. Rosa, Daniel Gianola, Bruno Valente, Julia P. Sumner, Wendy Baltzer, et al. 2017. “Genome-Wide Association Analysis in Dogs Implicates 99 Loci as Risk Variants for Anterior Cruciate Ligament Rupture.” *PloS One* 12 (4): e0173810.

Baum, Leonard E., and Ted Petrie. 1966. “Statistical Inference for Probabilistic Functions of Finite State Markov Chains.” *The Annals of Mathematical Statistics*. https://doi.org/10.1214/aoms/1177699147.

Berglund, Jonas, Elisa M. Nevalainen, Anna-Maja Molin, Michele Perloski, LUPA Consortium, Catherine André, Michael C. Zody, et al. 2012. “Novel Origins of Copy Number Variation in the Dog Genome.” *Genome Biology* 13 (8): R73.

Bicego, Manuele, Marco Cristani, and Vittorio Murino. 2007. “Sparseness Achievement in Hidden Markov Models.” *14th International Conference on Image Analysis and Processing (ICIAP 2007)*. https://doi.org/10.1109/iciap.2007.4362759.

Bickhart, Derek M., Yali Hou, Steven G. Schroeder, Can Alkan, Maria Francesca Cardone, Lakshmi K. Matukumalli, Jiuzhou Song, et al. 2012. “Copy Number Variation of Individual Cattle Genomes Using next-Generation Sequencing.” *Genome Research* 22 (4): 778–90.

Botigué, Laura R., Shiya Song, Amelie Scheu, Shyamalika Gopalan, Amanda L. Pendleton, Matthew Oetjens, Angela M. Taravella, et al. 2017. “Ancient European Dog Genomes Reveal Continuity since the Early Neolithic.” *Nature Communications* 8 (July): 16082.

Brahmachary, Manisha, Audrey Guilmatre, Javier Quilez, Dan Hasson, Christelle Borel, Peter Warburton, and Andrew J. Sharp. 2014. “Digital Genotyping of Macrosatellites and Multicopy Genes Reveals Novel Biological Functions Associated with Copy Number Variation of Large Tandem Repeats.” *PLoS Genetics* 10 (6): e1004418.

Browning, Sharon R., and Brian L. Browning. 2007. “Rapid and Accurate

Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies by Use of Localized Haplotype Clustering.” *American Journal of Human Genetics* 81 (5): 1084–97.

Cantsilieris, Stuart, Paul N. Baird, and Stefan J. White. 2013. “Molecular Methods for Genotyping Complex Copy Number Polymorphisms.” *Genomics* 101 (2): 86–93.

Castle, W. E. 1903. “The Laws of Heredity of Galton and Mendel, and Some Laws Governing Race Improvement by Selection.” *Proceedings of the American Academy of Arts and Sciences*. https://doi.org/10.2307/20021870.

Chase, Kevin, Paul Jones, Alan Martin, Elaine A. Ostrander, and Karl G. Lark. 2009. “Genetic Mapping of Fixed Phenotypes: Disease Frequency as a Breed Characteristic.” *The Journal of Heredity* 100 Suppl 1 (July): S37–41.

Chen, Jian-Min, David N. Cooper, Nadia Chuzhanova, Claude Férec, and George P. Patrinos. 2007. “Gene Conversion: Mechanisms, Evolution and Human Disease.” *Nature Reviews. Genetics* 8 (10): 762–75.

Chen, Ken, John W. Wallis, Michael D. McLellan, David E. Larson, Joelle M. Kalicki, Craig S. Pohl, Sean D. McGrath, et al. 2009. “BreakDancer: An Algorithm for High-Resolution Mapping of Genomic Structural Variation.” *Nature Methods* 6 (9): 677–81.

Chen, Lu, Weichen Zhou, Ling Zhang, and Feng Zhang. 2014. “Genome Architecture and Its Roles in Human Copy Number Variation.” *Genomics & Informatics* 12 (4): 136–44.

Chen, Wei-Kang, Joshua D. Swartz, Laura J. Rush, and Carlos E. Alvarez. 2009. “Mapping DNA Structural Variation in Dogs.” *Genome Research* 19 (3): 500–509.

“Company Search | Company Information | Hoovers Company Profiles - D&B Hoovers - Companies & Details - Hoovers.com.” n.d. Accessed October 26, 2019. http://www.hoovers.com/company-information/company-search.html?term=23andme.

Conrad, Donald F., and Matthew E. Hurles. 2007. “The Population Genetics of Structural Variation.” *Nature Genetics*. https://doi.org/10.1038/ng2042.

Cooper, Geoffrey M. 2000. *The Cell: A Molecular Approach*. Sinauer Associates.

Cordaux, Richard, and Mark A. Batzer. 2009. “The Impact of Retrotransposons on Human Genome Evolution.” *Nature Reviews. Genetics* 10 (10): 691–703.

Cruz, Fernando, Carles Vilà, and Matthew T. Webster. 2008. “The Legacy of Domestication: Accumulation of Deleterious Mutations in the Dog Genome.” *Molecular Biology and Evolution* 25 (11): 2331–36.

Deane-Coe, Petra E., Erin T. Chu, Andrea Slavney, Adam R. Boyko, and Aaron J. Sams. 2018. “Direct-to-Consumer DNA Testing of 6,000 Dogs Reveals 98.6-Kb Duplication Associated with Blue Eyes and Heterochromia in Siberian Huskies.” *PLoS Genetics* 14 (10): e1007648.

Dehal, Paramvir, and Jeffrey L. Boore. 2005. “Two Rounds of Whole Genome Duplication in the Ancestral Vertebrate.” *PLoS Biology*. https://doi.org/10.1371/journal.pbio.0030314.

Dennis, Megan Y., and Evan E. Eichler. 2016. “Human Adaptation and Evolution by Segmental Duplication.” *Current Opinion in Genetics & Development* 41 (December): 44–52.

Doan, Ryan, Noah Cohen, Jessica Harrington, Kylee Veazey, Rytis Juras, Gus Cothran, Molly E. McCue, Loren Skow, and Scott V. Dindot. 2012. “Identification of Copy Number Variants in Horses.” *Genome Research* 22 (5): 899–907.

Dreger, Dayna L., Heidi G. Parker, Elaine A. Ostrander, and Sheila M. Schmutz. 2013. “Identification of a Mutation That Is Associated with the Saddle Tan and Black-and-Tan Phenotypes in Basset Hounds and Pembroke Welsh Corgis.” *The Journal of Heredity* 104 (3): 399–406.

Dreger, D. L., and S. M. Schmutz. 2011. “A SINE Insertion Causes the Black-and-Tan and Saddle Tan Phenotypes in Domestic Dogs.” *Journal of*

Heredity. https://doi.org/10.1093/jhered/esr042.

Drögemüller, Cord, Elinor K. Karlsson, Marjo K. Hytönen, Michele Perloski, Gaudenz Dolf, Kirsi Sainio, Hannes Lohi, Kerstin Lindblad-Toh, and Tosso Leeb. 2008. “A Mutation in Hairless Dogs Implicates FOXI3 in Ectodermal Development.” *Science* 321 (5895): 1462.

Duan, Junbo, Han Liu, Lanling Zhao, Xiguo Yuan, Yu-Ping Wang, and Mingxi Wan. 2019. “Detection of False-Positive Deletions from the Database of Genomic Variants.” *BioMed Research International* 2019 (April). https://doi.org/10.1155/2019/8420547.

Everitt, B. S., and D. J. Hand. 1981. “Finite Mixture Distributions.” https://doi.org/10.1007/978-94-009-5897-5.

Fages, Antoine, Kristian Hanghøj, Naveed Khan, Charleen Gaunitz, Andaine Seguin-Orlando, Michela Leonardi, Christian McCrory Constantz, et al. 2019. “Tracking Five Millennia of Horse Management with Extensive Ancient Genome Time Series.” *Cell* 177 (6): 1419–35.e31.

Fan, Zhenxin, Pedro Silva, Ilan Gronau, Shuoguo Wang, Aitor Serres Armero, Rena M. Schweizer, Oscar Ramirez, et al. 2016. “Worldwide Patterns of Genomic Variation and Admixture in Gray Wolves.” *Genome Research* 26 (2): 163–73.

Felsenstein, Joe. 1987. “Molecular Evolutionary Genetics.” *Cell*. https://doi.org/10.1016/0092-8674(87)90630-1.

Ferguson, Thomas S. 1973. “A Bayesian Analysis of Some Nonparametric Problems.” *The Annals of Statistics*. https://doi.org/10.1214/aos/1176342360.

Fisher, R. A. 1937. “THE WAVE OF ADVANCE OF ADVANTAGEOUS GENES.” *Annals of Eugenics*. https://doi.org/10.1111/j.1469-1809.1937.tb02153.x.

Freedman, Adam H., Ilan Gronau, Rena M. Schweizer, Diego Ortega-Del Vecchyo, Eunjung Han, Pedro M. Silva, Marco Galaverni, et al. 2014. “Genome Sequencing Highlights the Dynamic Early History of Dogs.”

PLoS Genetics 10 (1): e1004016.

Freedman, Adam H., Kirk E. Lohmueller, and Robert K. Wayne. 2016. “Evolutionary History, Selective Sweeps, and Deleterious Variation in the Dog.” *Annual Review of Ecology, Evolution, and Systematics*. https://doi.org/10.1146/annurev-ecolsys-121415-032155.

Gelabert, Pere, Marcela Sandoval-Velasco, Aitor Serres, Marc de Manuel, Pere Renom, Ashot Margaryan, Toni de-Dios, et al. n.d. “Evolutionary History, Genomic Adaptation to Toxic Diet and Extinction of the Carolina Parakeet.” *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3417467.

Ghosh, Sharmila, Zhipeng Qu, Pranab J. Das, Erica Fang, Rytis Juras, E. Gus Cothran, Sue McDonell, et al. 2014. “Copy Number Variation in the Horse Genome.” *PLoS Genetics* 10 (10): e1004712.

Gifford, R. J., A. Katzourakis, M. Tristem, O. G. Pybus, M. Winters, and R. W. Shafer. 2008. “A Transitional Endogenous Lentivirus from the Genome of a Basal Primate and Implications for Lentivirus Evolution.” *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.0807873105.

Goodwin, Sara, John D. McPherson, and W. Richard McCombie. 2016. “Coming of Age: Ten Years of next-Generation Sequencing Technologies.” *Nature Reviews. Genetics* 17 (6): 333–51.

Gopalakrishnan, Shyam, Jose A. Samaniego Castruita, Mikkel-Holger S. Sinding, Lukas F. K. Kuderna, Jannikke Räikkönen, Bent Petersen, Thomas Sicheritz-Ponten, et al. 2017. “The Wolf Reference Genome Sequence (Canis Lupus Lupus) and Its Implications for Canis Spp. Population Genomics.” *BMC Genomics* 18 (1): 495.

Gopalakrishnan, Shyam, Mikkel-Holger S. Sinding, Jazmín Ramos-Madrigal, Jonas Niemann, Jose A. Samaniego Castruita, Filipe G. Vieira, Christian Carøe, et al. 2018. “Interspecific Gene Flow Shaped the Evolution of the Genus Canis.” *Current Biology: CB* 28 (21): 3441–49.e5.

Gorla, E., M. C. Cozzi, S. I. Román-Ponce, F. J. Ruiz López, V. E. Vega-

Murillo, S. Cerolini, A. Bagnato, and M. G. Strillacci. 2017. “Genomic Variability in Mexican Chicken Population Using Copy Number Variants.” *BMC Genetics* 18 (1): 61.

Grimwood, Jane, Laurie A. Gordon, Anne Olsen, Astrid Terry, Jeremy Schmutz, Jane Lamerdin, Uffe Hellsten, et al. 2004. “The DNA Sequence and Biology of Human Chromosome 19.” *Nature* 428 (6982): 529–35.

Gu, Wenli, Feng Zhang, and James R. Lupski. 2008. “Mechanisms for Human Genomic Rearrangements.” *PathoGenetics*. https://doi.org/10.1186/1755-8417-1-4.

Gymrek, Melissa, Thomas Willems, David Reich, and Yaniv Erlich. 2017. “Interpreting Short Tandem Repeat Variations in Humans Using Mutational Constraint.” *Nature Genetics* 49 (10): 1495–1501.

Handsaker, Robert E., Vanessa Van Doren, Jennifer R. Berman, Giulio Genovese, Seva Kashin, Linda M. Boettger, and Steven A. McCarroll. 2015. “Large Multiallelic Copy Number Variations in Humans.” *Nature Genetics*. https://doi.org/10.1038/ng.3200.

Hastings, P. J., James R. Lupski, Susan M. Rosenberg, and Grzegorz Ira. 2009. “Mechanisms of Change in Gene Copy Number.” *Nature Reviews. Genetics* 10 (8): 551–64.

Hayward, Jessica J., Marta G. Castelhano, Kyle C. Oliveira, Elizabeth Corey, Cheryl Balkman, Tara L. Baxter, Margret L. Casal, et al. 2016. “Complex Disease and Phenotype Mapping in the Domestic Dog.” *Nature Communications* 7 (January): 10460.

Higuchi, Russell, Gavin Dollinger, P. Sean Walsh, and Robert Griffith. 1992. “Simultaneous Amplification and Detection of Specific DNA Sequences.” *Bio/Technology*. https://doi.org/10.1038/nbt0492-413.

Hou, Yali, Derek M. Bickhart, Hoyoung Chung, Jana L. Hutchison, H. Duane Norman, Erin E. Connor, and George E. Liu. 2012. “Analysis of Copy Number Variations in Holstein Cows Identify Potential Mechanisms Contributing to Differences in Residual Feed Intake.” *Functional &*

Integrative Genomics 12 (4): 717–23.

International Human Genome Sequencing Consortium. 2004. “Finishing the Euchromatic Sequence of the Human Genome.” *Nature* 431 (7011): 931–45.

Jiang, Li, Jicai Jiang, Jie Yang, Xuan Liu, Jiying Wang, Haifei Wang, Xiangdong Ding, Jianfeng Liu, and Qin Zhang. 2013. “Genome-Wide Detection of Copy Number Variations Using High-Density SNP Genotyping Platforms in Holsteins.” *BMC Genomics*. https://doi.org/10.1186/1471-2164-14-131.

Jia, X., S. Chen, H. Zhou, D. Li, W. Liu, and N. Yang. 2013. “Copy Number Variations Identified in the Chicken Using a 60K SNP BeadChip.” *Animal Genetics*. https://doi.org/10.1111/age.12009.

Jones, Paul, Kevin Chase, Alan Martin, Pluis Davern, Elaine A. Ostrander, and Karl G. Lark. 2008. “Single-Nucleotide-Polymorphism-Based Association Mapping of Dog Stereotypes.” *Genetics* 179 (2): 1033–44.

Kader, Adiljan, Yan Li, Kunzhe Dong, David M. Irwin, Qianjun Zhao, Xiaohong He, Jianfeng Liu, et al. 2015. “Population Variation Reveals Independent Selection toward Small Body Size in Chinese Debao Pony.” *Genome Biology and Evolution* 8 (1): 42–50.

Kallioniemi, A., O. Kallioniemi, D. Sudar, D. Rutovitz, J. Gray, F. Waldman, and D. Pinkel. 1992. “Comparative Genomic Hybridization for Molecular Cytogenetic Analysis of Solid Tumors.” *Science*. https://doi.org/10.1126/science.1359641.

Karlsson, Elinor K., and Kerstin Lindblad-Toh. 2008. “Leader of the Pack: Gene Mapping in Dogs and Other Model Organisms.” *Nature Reviews. Genetics* 9 (9): 713–25.

Keel, Brittney N., Amanda K. Lindholm-Perry, and Warren M. Snelling. 2016. “Evolutionary and Functional Features of Copy Number Variation in the Cattle Genome.” *Frontiers in Genetics* 7 (November): 207.

Kidd, Jeffrey M. 17 Sep, 2018 - 20 Sep, 2018. “De Novo Assembly and

Analysis of a Canine Genome." presented at the Genome Informatics, Wellcome Genome Campus Conference Centre, Hinxton, Cambridge,UK. http://www.genetics.org.uk/events/genome-informatics/.

Kim, Philip M., Hugo Y. K. Lam, Alexander E. Urban, Jan O. Korbel, Jason Affourtit, Fabian Grubert, Xueying Chen, Sherman Weissman, Michael Snyder, and Mark B. Gerstein. 2008. "Analysis of Copy Number Variants and Segmental Duplications in the Human Genome: Evidence for a Change in the Process of Formation in Recent Evolutionary History." *Genome Research* 18 (12): 1865–74.

Kingman, J. F. C. 1982. "On the Genealogy of Large Populations." *Journal of Applied Probability*. https://doi.org/10.2307/3213548.

Kosugi, Shunichi, Yukihide Momozawa, Xiaoxi Liu, Chikashi Terao, Michiaki Kubo, and Yoichiro Kamatani. 2019. "Comprehensive Evaluation of Structural Variation Detection Algorithms for Whole Genome Sequencing." *Genome Biology* 20 (1): 117.

Kronenberg, Zev N., Ian T. Fiddes, David Gordon, Shwetha Murali, Stuart Cantsilieris, Olivia S. Meyerson, Jason G. Underwood, et al. 2018. "High-Resolution Comparative Analysis of Great Ape Genomes." *Science* 360 (6393). https://doi.org/10.1126/science.aar6343.

Kuderna, Lukas F. K., Esther Lizano, Eva Julià, Jessica Gomez-Garrido, Aitor Serres-Armero, Martin Kuhlwilm, Regina Antoni Alandes, et al. 2019. "Selective Single Molecule Sequencing and Assembly of a Human Y Chromosome of African Origin." *Nature Communications* 10 (1): 4.

Kuderna, Lukas F. K., Chad Tomlinson, Ladeana W. Hillier, Annabel Tran, Ian T. Fiddes, Joel Armstrong, Hafid Laayouni, et al. 2017. "A 3-Way Hybrid Approach to Generate a New High-Quality Chimpanzee Reference Genome (Pan_tro_3.0)." *GigaScience* 6 (11): 1–6.

Lange, David. 2019. "Betting Industry of the United Kingdom." Statista. June 14, 2019. https://www.statista.com/statistics/469762/gambling-turnover-dogs-in-great-britain-betting/.

Langer-Safer, P. R., M. Levine, and D. C. Ward. 1982. “Immunological Method for Mapping Genes on Drosophila Polytene Chromosomes.” *Proceedings of the National Academy of Sciences of the United States of America* 79 (14): 4381–85.

Layer, Ryan M., Colby Chiang, Aaron R. Quinlan, and Ira M. Hall. 2014. “LUMPY: A Probabilistic Framework for Structural Variant Discovery.” *Genome Biology* 15 (6): R84.

Liao, D. 1999. “Concerted Evolution: Molecular Mechanism and Biological Implications.” *American Journal of Human Genetics* 64 (1): 24–30.

Librado, Pablo, Cristina Gamba, Charleen Gaunitz, Clio Der Sarkissian, Mélanie Pruvost, Anders Albrechtsen, Antoine Fages, et al. 2017. “Ancient Genomic Changes Associated with Domestication of the Horse.” *Science* 356 (6336): 442–45.

Lindblad-Toh, Kerstin, Claire M. Wade, Tarjei S. Mikkelsen, Elinor K. Karlsson, David B. Jaffe, Michael Kamal, Michele Clamp, et al. 2005. “Genome Sequence, Comparative Analysis and Haplotype Structure of the Domestic Dog.” *Nature* 438 (7069): 803–19.

Lukacs, Eugene. 1955. “A Characterization of the Gamma Distribution.” *The Annals of Mathematical Statistics*. https://doi.org/10.1214/aoms/1177728549.

MacLean, Evan L., Noah Snyder-Mackler, Bridgett M. vonHoldt, and James A. Serpell. 2019. “Highly Heritable and Functionally Relevant Breed Differences in Dog Behaviour.” *Proceedings. Biological Sciences / The Royal Society* 286 (1912): 20190716.

MacLeant, Evan L., Noah Snyder-Mackler, Bridgett M. vonHoldt, and James A. Serpell. n.d. “Highly Heritable and Functionally Relevant Breed Differences in Dog Behavior.” https://doi.org/10.1101/509315.

Mallick, Swapan, Heng Li, Mark Lipson, Iain Mathieson, Melissa Gymrek, Fernando Racimo, Mengyao Zhao, et al. 2016. “The Simons Genome Diversity Project: 300 Genomes from 142 Diverse Populations.” *Nature*

538 (7624): 201–6.

Marques-Bonet, T., and E. E. Eichler. 2009. “The Evolution of Human Segmental Duplications and the Core Duplicon Hypothesis.” *Cold Spring Harbor Symposia on Quantitative Biology* 74 (August): 355–62.

Marques-Bonet, Tomas, Jeffrey M. Kidd, Mario Ventura, Tina A. Graves, Ze Cheng, Ladeana W. Hillier, Zhaoshi Jiang, et al. 2009. “A Burst of Segmental Duplications in the Genome of the African Great Ape Ancestor.” *Nature* 457 (7231): 877–81.

Ma, Ruiyu, Linbei Deng, Yan Xia, Xianda Wei, Yingxi Cao, Ruolan Guo, Rui Zhang, Jing Guo, Desheng Liang, and Lingqian Wu. 2017. “A Clear Bias in Parental Origin of de Novo Pathogenic CNVs Related to Intellectual Disability, Developmental Delay and Multiple Congenital Anomalies.” *Scientific Reports* 7 (March): 44446.

McCulloch, W. S., and W. Pitts. 1990. “A Logical Calculus of the Ideas Immanent in Nervous Activity. 1943.” *Bulletin of Mathematical Biology* 52 (1-2): 99–115; discussion 73–97.

McLean, Cory Y., Philip L. Reno, Alex A. Pollen, Abraham I. Bassan, Terence D. Capellini, Catherine Guenther, Vahan B. Indjeian, et al. 2011. “Human-Specific Loss of Regulatory DNA and the Evolution of Human-Specific Traits.” *Nature* 471 (7337): 216–19.

Mellersh, Cathryn S. 2014. “The Genetics of Eye Disorders in the Dog.” *Canine Genetics and Epidemiology*. https://doi.org/10.1186/2052-6687-1-3.

Metzger, Julia, Ute Philipp, Maria Susana Lopes, Artur da Camara Machado, Michela Felicetti, Maurizio Silvestrelli, and Ottmar Distl. 2013. “Analysis of Copy Number Variants by Three Detection Algorithms and Their Association with Body Size in Horses.” *BMC Genomics* 14 (July): 487.

Mimori, Takahiro, Naoki Nariai, Kaname Kojima, Yukuto Sato, Yosuke Kawai, Yumi Yamaguchi-Kabata, and Masao Nagasaki. 2015. “Estimating Copy Numbers of Alleles from Population-Scale High-Throughput Sequencing Data.” *BMC Bioinformatics* 16 Suppl 1 (January): S4.

Moens, Lotte N., Elin Falk-Sörqvist, A. Charlotta Asplund, Ewa Bernatowska, C. I. Edvard Smith, and Mats Nilsson. 2014. “Diagnostics of Primary Immunodeficiency Diseases: A Sequencing Capture Approach.” *PloS One* 9 (12): e114901.

Morgan, Andrew P., Daniel M. Gatti, Maya L. Najarian, Thomas M. Keane, Raymond J. Galante, Allan I. Pack, Richard Mott, Gary A. Churchill, and Fernando Pardo-Manuel de Villena. 2017. “Structural Variation Shapes the Landscape of Recombination in Mouse.” *Genetics* 206 (2): 603–19.

Muller, H. J. 1964. “THE RELATION OF RECOMBINATION TO MUTATIONAL ADVANCE.” *Mutation Research* 106 (May): 2–9.

Murphy, Kevin Patrick. 2002. *Dynamic Bayesian Networks: Representation, Inference and Learning*.

Nicholas, Thomas J., Carl Baker, Evan E. Eichler, and Joshua M. Akey. 2011. “A High-Resolution Integrated Map of Copy Number Polymorphisms within and between Breeds of the Modern Domesticated Dog.” *BMC Genomics* 12 (August): 414.

Nicholas, Thomas J., Ze Cheng, Mario Ventura, Katrina Mealey, Evan E. Eichler, and Joshua M. Akey. 2009. “The Genomic Architecture of Segmental Duplications and Associated Copy Number Variants in Dogs.” *Genome Research* 19 (3): 491–99.

Niskanen, A. K., E. Hagström, H. Lohi, M. Ruokonen, R. Esparza-Salas, J. Aspi, and P. Savolainen. 2013. “MHC Variability Supports Dog Domestication from a Large Number of Wolves: High Diversity in Asia.” *Heredity* 110 (1): 80–85.

Ni, Xiaohui, Minglei Zhuo, Zhe Su, Jianchun Duan, Yan Gao, Zhijie Wang, Chenghang Zong, et al. 2013. “Reproducible Copy Number Variation Patterns among Single Circulating Tumor Cells of Lung Cancer Patients.” *Proceedings of the National Academy of Sciences of the United States of America* 110 (52): 21083–88.

Ohno, Susumu. 1970. “Evolution by Gene Duplication.”

https://doi.org/10.1007/978-3-642-86659-3.

Ohta, T. 1984. “Population Genetics Theory of Concerted Evolution and Its Application to the Immunoglobulin V Gene Tree.” *Journal of Molecular Evolution* 20 (3-4): 274–80.

Ollivier, Morgane, Anne Tresset, Fabiola Bastian, Laetitia Lagoutte, Erik Axelsson, Maja-Louise Arendt, Adrian Bălăşescu, et al. 2016. “Amy2B Copy Number Variation Reveals Starch Diet Adaptations in Ancient European Dogs.” *Royal Society Open Science*. https://doi.org/10.1098/rsos.160449.

Olshen, A. B., E. S. Venkatraman, R. Lucito, and M. Wigler. 2004. “Circular Binary Segmentation for the Analysis of Array-Based DNA Copy Number Data.” *Biostatistics*. https://doi.org/10.1093/biostatistics/kxh008.

Paigen, Kenneth, and Petko M. Petkov. 2018. “PRDM9 and Its Role in Genetic Recombination.” *Trends in Genetics*. https://doi.org/10.1016/j.tig.2017.12.017.

Parker, Heidi G., Dayna L. Dreger, Maud Rimbault, Brian W. Davis, Alexandra B. Mullen, Gretchen Carpintero-Ramirez, and Elaine A. Ostrander. 2017. “Genomic Analyses Reveal the Influence of Geographic Origin, Migration, and Hybridization on Modern Dog Breed Development.” *Cell Reports* 19 (4): 697–708.

Paudel, Yogesh, Ole Madsen, Hendrik-Jan Megens, Laurent A. F. Frantz, Mirte Bosse, John W. M. Bastiaansen, Richard P. M. A. Crooijmans, and Martien A. M. Groenen. 2013. “Evolutionary Dynamics of Copy Number Variation in Pig Genomes in the Context of Adaptation and Domestication.” *BMC Genomics* 14 (July): 449.

Perry, George H., Nathaniel J. Dominy, Katrina G. Claw, Arthur S. Lee, Heike Fiegler, Richard Redon, John Werner, et al. 2007. “Diet and the Evolution of Human Amylase Gene Copy Number Variation.” *Nature Genetics* 39 (10): 1256–60.

Pirooznia, Mehdi, Fernando S. Goes, and Peter P. Zandi. 2015. “Whole-Genome

CNV Analysis: Advances in Computational Approaches." *Frontiers in Genetics* 6 (April): 138.

Pitulko, Vladimir, and Aleksey Kasparov. 2017. "Archaeological Dogs from the Early Holocene Zhokhov Site in the Eastern Siberian Arctic." *Journal of Archaeological Science: Reports* 13 (June): 491–515.

Plassais, Jocelyn, Jaemin Kim, Brian W. Davis, Danielle M. Karyadi, Andrew N. Hogan, Alex C. Harris, Brennan Decker, Heidi G. Parker, and Elaine A. Ostrander. 2019. "Whole Genome Sequencing of Canids Reveals Genomic Regions under Selection and Variants Influencing Morphology." *Nature Communications* 10 (1): 1489.

Prado-Martinez, Javier, Peter H. Sudmant, Jeffrey M. Kidd, Heng Li, Joanna L. Kelley, Belen Lorente-Galdos, Krishna R. Veeramah, et al. 2013. "Great Ape Genetic Diversity and Population History." *Nature* 499 (7459): 471–75.

Rafajlović, M., A. Klassmann, A. Eriksson, T. Wiehe, and B. Mehlig. 2014. "Demography-Adjusted Tests of Neutrality Based on Genome-Wide SNP Data." *Theoretical Population Biology* 95 (August): 1–12.

Rakocevic, Goran, Vladimir Semenyuk, Wan-Ping Lee, James Spencer, John Browning, Ivan J. Johnson, Vladan Arsenijevic, et al. 2019. "Fast and Accurate Genomic Analyses Using Genome Graphs." *Nature Genetics* 51 (2): 354–62.

Ramirez, Oscar, Iñigo Olalde, Jonas Berglund, Belen Lorente-Galdos, Jessica Hernandez-Rodriguez, Javier Quilez, Matthew T. Webster, et al. 2014. "Analysis of Structural Diversity in Wolf-like Canids Reveals Post-Domestication Variants." *BMC Genomics* 15 (June): 465.

Ramsay, Philip H., and David W. Scott. 1993. "Multivariate Density Estimation, Theory, Practice, and Visualization." *Technometrics*. https://doi.org/10.2307/1270280.

Reiter, Taylor, Evelyn Jagoda, and Terence D. Capellini. 2016. "Dietary Variation and Evolution of Gene Copy Number among Dog Breeds." *PloS*

One 11 (2): e0148899.

Rendine, S., A. Piazza, and L. L. Cavalli-Sforza. 1986. “Simulation and Separation by Principal Components of Multiple Demic Expansions in Europe.” *The American Naturalist*. https://doi.org/10.1086/284597.

Reno, Philip L., Cory Y. McLean, Jasmine E. Hines, Terence D. Capellini, Gill Bejerano, and David M. Kingsley. 2013. “A Penile Spine/vibrissa Enhancer Sequence Is Missing in Modern and Extinct Humans but Is Retained in Multiple Primates with Penile Spines and Sensory Vibrissae.” *PloS One* 8 (12): e84258.

Rhoads, Anthony, and Kin Fai Au. 2015. “PacBio Sequencing and Its Applications.” *Genomics, Proteomics & Bioinformatics*. https://doi.org/10.1016/j.gpb.2015.08.002.

Saiki, R. K., S. Scharf, F. Faloona, K. B. Mullis, G. T. Horn, H. A. Erlich, and N. Arnheim. 1985. “Enzymatic Amplification of Beta-Globin Genomic Sequences and Restriction Site Analysis for Diagnosis of Sickle Cell Anemia.” *Science* 230 (4732): 1350–54.

Salmon Hillbertz, Nicolette H. C., Magnus Isaksson, Elinor K. Karlsson, Eva Hellmén, Gerli Rosengren Pielberg, Peter Savolainen, Claire M. Wade, et al. 2007. “Duplication of FGF3, FGF4, FGF19 and ORAOV1 Causes Hair Ridge and Predisposition to Dermoid Sinus in Ridgeback Dogs.” *Nature Genetics* 39 (11): 1318–20.

San Filippo, Joseph, Patrick Sung, and Hannah Klein. 2008. “Mechanism of Eukaryotic Homologous Recombination.” *Annual Review of Biochemistry* 77: 229–57.

Sanger, F., and A. R. Coulson. 1975. “A Rapid Method for Determining Sequences in DNA by Primed Synthesis with DNA Polymerase.” *Journal of Molecular Biology* 94 (3): 441–48.

Santiago, Yolanda, Edmond Chan, Pei-Qi Liu, Salvatore Orlando, Lin Zhang, Fyodor D. Urnov, Michael C. Holmes, et al. 2008. “Targeted Gene Knockout in Mammalian Cells by Using Engineered Zinc-Finger

Nucleases.” *Proceedings of the National Academy of Sciences of the United States of America* 105 (15): 5809–14.

Scheet, Paul, and Matthew Stephens. 2006. “A Fast and Flexible Statistical Model for Large-Scale Population Genotype Data: Applications to Inferring Missing Genotypes and Haplotypic Phase.” *American Journal of Human Genetics* 78 (4): 629–44.

Scientists, Genome 10k Community of, and Genome 10K Community of Scientists. 2009. “Genome 10K: A Proposal to Obtain Whole-Genome Sequence for 10 000 Vertebrate Species.” *Journal of Heredity*. https://doi.org/10.1093/jhered/esp086.

Serres-Armero, Aitor, Inna S. Povolotskaya, Javier Quilez, Oscar Ramirez, Gabriel Santpere, Lukas F. K. Kuderna, Jessica Hernandez-Rodriguez, et al. 2017. “Similar Genomic Proportions of Copy Number Variation within Gray Wolves and Modern Dog Breeds Inferred from Whole Genome Sequencing.” *BMC Genomics* 18 (1): 977.

Sharp, Andrew J., Devin P. Locke, Sean D. McGrath, Ze Cheng, Jeffrey A. Bailey, Rhea U. Vallente, Lisa M. Pertz, et al. 2005. “Segmental Duplications and Copy-Number Variation in the Human Genome.” *The American Journal of Human Genetics*. https://doi.org/10.1086/431652.

Shastry, Barkur S. 2002. “SNP Alleles in Human Disease and Evolution.” *Journal of Human Genetics* 47 (11): 561–66.

Shearin, Abigail L., and Elaine A. Ostrander. 2010. “Leading the Way: Canine Models of Genomics and Disease.” *Disease Models & Mechanisms* 3 (1-2): 27–34.

Sherman, Rachel M., Juliet Forman, Valentin Antonescu, Daniela Puiu, Michelle Daya, Nicholas Rafaels, Meher Preethi Boorgula, et al. 2019. “Assembly of a Pan-Genome from Deep Sequencing of 910 Humans of African Descent.” *Nature Genetics* 51 (1): 30–35.

She, X. 2006. “A Preliminary Comparative Analysis of Primate Segmental Duplications Shows Elevated Substitution Rates and a Great-Ape

Expansion of Intrachromosomal Duplications.” *Genome Research*. https://doi.org/10.1101/gr.4949406.

Siepel, Adam, Gill Bejerano, Jakob S. Pedersen, Angie S. Hinrichs, Minmei Hou, Kate Rosenbloom, Hiram Clawson, et al. 2005. “Evolutionarily Conserved Elements in Vertebrate, Insect, Worm, and Yeast Genomes.” *Genome Research* 15 (8): 1034–50.

Sipiczki, Matthias. 2018. “Interspecies Hybridisation and Genome Chimerisation in Saccharomyces: Combining of Gene Pools of Species and Its Biotechnological Perspectives.” *Frontiers in Microbiology*. https://doi.org/10.3389/fmicb.2018.03071.

Skoglund, Pontus, Erik Ersmark, Eleftheria Palkopoulou, and Love Dalén. 2015. “Ancient Wolf Genome Reveals an Early Divergence of Domestic Dog Ancestors and Admixture into High-Latitude Breeds.” *Current Biology*. https://doi.org/10.1016/j.cub.2015.04.019.

Spielmann, Malte, Darío G. Lupiáñez, and Stefan Mundlos. 2018. “Structural Variation in the 3D Genome.” *Nature Reviews. Genetics* 19 (7): 453–67.

Stephens, Philip J., Patrick S. Tarpey, Helen Davies, Peter Van Loo, Chris Greenman, David C. Wedge, Serena Nik-Zainal, et al. 2012. “The Landscape of Cancer Genes and Mutational Processes in Breast Cancer.” *Nature* 486 (7403): 400–404.

Sudmant, Peter H., John Huddleston, Claudia R. Catacchio, Maika Malig, Ladeana W. Hillier, Carl Baker, Kiana Mohajeri, et al. 2013. “Evolution and Diversity of Copy Number Variation in the Great Ape Lineage.” *Genome Research* 23 (9): 1373–82.

Sudmant, Peter H., Swapan Mallick, Bradley J. Nelson, Fereydoun Hormozdiari, Niklas Krumm, John Huddleston, Bradley P. Coe, et al. 2015. “Global Diversity, Population Stratification, and Selection of Human Copy-Number Variation.” *Science* 349 (6253): aab3761.

Sudmant, Peter H., Tobias Rausch, Eugene J. Gardner, Robert E. Handsaker, Alexej Abyzov, John Huddleston, Yan Zhang, et al. 2015. “An Integrated

Map of Structural Variation in 2,504 Human Genomes." *Nature* 526 (7571): 75–81.

The 1000 Genomes Project Consortium. 2015. "A Global Reference for Human Genetic Variation." *Nature* 526 (7571): 68–74.

Thomas, Gregg W. C., Richard J. Wang, Jelena Nguyen, R. Alan Harris, Muthuswamy Raveendran, Jeffrey Rogers, and Matthew W. Hahn. 2019. "Origins and Long-Term Patterns of Copy-Number Variation in Rhesus Macaques." *bioRxiv*. https://doi.org/10.1101/749416.

Touitou, Isabelle, Suzanne Lesage, Michael McDermott, Laurence Cuisset, Hal Hoffman, Catherine Dode, Nitza Shoham, et al. 2004. "Infevers: An Evolving Mutation Database for Auto-Inflammatory Syndromes." *Human Mutation* 24 (3): 194–98.

Vaysse, Amaury, Abhirami Ratnakumar, Thomas Derrien, Erik Axelsson, Gerli Rosengren Pielberg, Snaevar Sigurdsson, Tove Fall, et al. 2011. "Identification of Genomic Regions Associated with Phenotypic Variation between Dog Breeds Using Selection Mapping." *PLoS Genetics* 7 (10): e1002316.

Voelkerding, K. V., S. A. Dames, and J. D. Durtschi. 2009. "Next-Generation Sequencing: From Basic Research to Diagnostics." *Clinical Chemistry*. https://doi.org/10.1373/clinchem.2008.112789.

Voineagu, Irina, Vidhya Narayanan, Kirill S. Lobachev, and Sergei M. Mirkin. 2008. "Replication Stalling at Unstable Inverted Repeats: Interplay between DNA Hairpins and Fork Stabilizing Proteins." *Proceedings of the National Academy of Sciences of the United States of America* 105 (29): 9936–41.

Vonholdt, Bridgett M., John P. Pollinger, Kirk E. Lohmueller, Eunjung Han, Heidi G. Parker, Pascale Quignon, Jeremiah D. Degenhardt, et al. 2010. "Genome-Wide SNP and Haplotype Analyses Reveal a Rich History Underlying Dog Domestication." *Nature* 464 (7290): 898–902.

Wang, Guo-Dong, Xiu-Juan Shao, Bing Bai, Junlong Wang, Xiaobo Wang, Xue

Cao, Yan-Hu Liu, et al. 2019. “Structural Variation during Dog Domestication: Insights from Gray Wolf and Dhole Genomes.” *National Science Review*. https://doi.org/10.1093/nsr/nwy076.

Wang, Xiaofei, and Shannon Byers. 2014. “Copy Number Variation in Chickens: A Review and Future Prospects.” *Microarrays (Basel, Switzerland)* 3 (1): 24–38.

Watson, J. D., and F. H. C. Crick. 1953. “Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid.” *Nature* 171 (4356): 737–38.

Weerd, Jan Hendrik van, Ileana Badi, Malou van den Boogaard, Sonia Stefanovic, Harmen J. G. van de Werken, Melisa Gomez-Velazquez, Claudio Badia-Careaga, et al. 2014. “A Large Permissive Regulatory Domain Exclusively Controls Tbx3 Expression in the Cardiac Conduction System.” *Circulation Research*. https://doi.org/10.1161/circresaha.115.303591.

Willems, Thomas, Dina Zielinski, Jie Yuan, Assaf Gordon, Melissa Gymrek, and Yaniv Erlich. 2017. “Genome-Wide Profiling of Heritable and de Novo STR Variations.” *Nature Methods* 14 (6): 590–92.

Witkin, Andrew P. 1987. “SCALE-SPACE FILTERING.” *Readings in Computer Vision*. https://doi.org/10.1016/b978-0-08-051581-6.50036-2.

Wright, Sewall. 1950. “Genetical Structure of Populations.” *BMJ*. https://doi.org/10.1136/bmj.2.4669.36.

Wu, Steven H., Rachel S. Schwartz, David J. Winter, Donald F. Conrad, and Reed A. Cartwright. 2017. “Estimating Error Models for Whole Genome Sequencing Using Mixtures of Dirichlet-Multinomial Distributions.” *Bioinformatics* 33 (15): 2322–29.

Yan, Yiyuan, Guoqiang Yi, Congjiao Sun, Lujiang Qu, and Ning Yang. 2014. “Genome-Wide Characterization of Insertion and Deletion Variation in Chicken Using next Generation Sequencing.” *PloS One* 9 (8): e104652.

Zare, Fatima, Michelle Dow, Nicholas Monteleone, Abdelrahman Hosny, and Sheida Nabavi. 2017. “An Evaluation of Copy Number Variation Detection

Tools for Cancer Using Whole Exome Sequencing Data." *BMC Bioinformatics* 18 (1): 286.

Zhang, Feng, Wenli Gu, Matthew E. Hurles, and James R. Lupski. 2009. "Copy Number Variation in Human Health, Disease, and Evolution." *Annual Review of Genomics and Human Genetics*. https://doi.org/10.1146/annurev.genom.9.081307.164217.

Zhang, Huiping, Fan Wang, Henry R. Kranzler, Can Yang, Hongqin Xu, Zuoheng Wang, Hongyu Zhao, and Joel Gelernter. 2014. "Identification of Methylation Quantitative Trait Loci (mQTLs) Influencing Promoter DNA Methylation of Alcohol Dependence Risk Genes." *Human Genetics*. https://doi.org/10.1007/s00439-014-1452-2.

Zhou, Yang, Erin E. Connor, George R. Wiggans, Yongfang Lu, Robert J. Tempelman, Steven G. Schroeder, Hong Chen, and George E. Liu. 2018. "Genome-Wide Copy Number Variant Analysis Reveals Variants Associated with 10 Diverse Production Traits in Holstein Cattle." *BMC Genomics* 19 (1): 314.

Zhu, Ying, André M. M. Sousa, Tianliuyun Gao, Mario Skarica, Mingfeng Li, Gabriel Santpere, Paula Esteller-Cucala, et al. 2018. "Spatiotemporal Transcriptomic Divergence across Human and Macaque Brain Development." *Science* 362 (6420). https://doi.org/10.1126/science.aat8077.

Zuckerkandl, Emile, and Linus Pauling. 1965. "Evolutionary Divergence and Convergence in Proteins." *Evolving Genes and Proteins*. https://doi.org/10.1016/b978-1-4832-2734-4.50017-6.

Supplementary figures and tables

Supplementary Table 1: List of CNV calling software.

RP: read pairs; RD: read depth; SR: split reads; AS: assembly-based; LR: long reads; SP: span of reads; SV library: public SV libraries; RF: Random Forest; MEI: mobile element insertion; NUMT: nuclear mitochondrial genome insertion; VEI: virus element insertion.

Software	SV types	Detection	Input	Version	Author	Paper
1-2-3-SV	DEL,INS,INV,TRA	RP	bam	1(2012)	https://github.com/Vityay/1-2-3-SV	n.a.
AS-GENSENG	DEL,DUP	RD	bam	1.0.2(2015)	Wang W et al. (2015)	Allele-specific copy-number discovery from whole-genome and whole-exome sequencing
BASIL-ANISE	INS	RP,AS	bam	0.3.0(2015)	Holtgrewe M et al. (2015)	Methods for the detection and assembly of novel sequence in high-throughput sequencing data
BatVI	VEI	RP,SR	fastq	1.02(2017)	Tennakoon C et al. (2017)	BATVI: Fast, sensitive and accurate detection of virus integrations
BICseq2	DEL,DUP	RD	bam	0.2.4/0.7.2(20	Xi R et al. (2016)	Copy number analysis of whole-genome data using BIC-seq2 and its application to detection of cancer susceptibility variants
BreakDancer	DEL,INS,INV,TRA	RP	bam	1.3.6(2015)	Chen K et al. (2009)	BreakDancer: an algorithm for high-resolution mapping of genomic structural variation
BreakSeek	DEL,INS	RP,SR	bam	1.2(2015)	Zhao H et al. (2015)	BreakSeek: a breakpoint-based algorithm for full spectral range INDEL detection
BreakSeq2	DEL,INS	SV-library	bam	2.2(2015)	Lam HY et al. (2010)	Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library
Breakway	DEL,INS	RP	bam	0.7.1(2011)	Clark MJ et al. (2010)	U87MG decoded: the genomic sequence of a cytogenetically aberrant human cancer cell line
CLEVER	DEL,INS	RP	bam	2(2015)	Marschall T et al. (2012)	CLEVER: clique-enumerating variant finder
cn.MOPS	DEL,DUP	RD	bam	1.0(2012)	Klambauer et al. (2012)	cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate
cnvHiTSeq	DEL,DUP	RD,SR	bam	1.0(2012)	Bellos E et al. (2012)	cnvHiTSeq: integrative models for high-resolution copy number variation detection and genotyping using population sequencing data
CNVnator	DEL,DUP	RD	bam	0.3.2(2015)	Abyzov A et al. (2011)	CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing
CNV-seq	DEL,DUP	RD	bam	1.0(2009)	Xie et al. (2009)	CNV-seq, a new method to detect copy number variation using high-throughput sequencing
Control-FREEC	DEL,DUP	RD	bam	8.7(2016)	Boeva V et al. (2012)	Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data
CREST	DEL,INS,INV,TRA	RP,SR,AS	bam	1.0(2011)	Wang J et al. (2011)	CREST maps somatic structural variation in cancer genomes with base-pair resolution
DELLY	DEL,DUP,INV,TRA	RP,SR	bam	0.7.2(2016)	Rausch T et al. (2012)	DELLY: structural variant discovery by integrated paired-end and split-read analysis
DIGTYPER	DUP,INV	RP,SR	bam	unknown(201	Ebler J et al. (2017)	Genotyping inversions and tandem duplications
DINUMT	NUMT	RP	bam	0.0.23(2014)	Dayama G et al. (2014)	The genomic landscape of polymorphic human nuclear mitochondrial insertions
ERDS	DEL,DUP	RP,SR,RD	bam	1.1(2013)	Zhu M et al. (2012)	Using ERDS to infer copy-number variants in high-coverage genomes
FermiKit	DEL,INS	AS	fastq	0.13(2015)	Li H (2015).	FermiKit: assembly-based variant calling for Illumina resequencing data
forestSV	DEL,DUP	RP,RD,RF	bam	0.3.3(2013)	Michaelson JJ et al. (2012)	forestSV: structural variant discovery through statistical learning
GASVpro	DEL,INV	RP,RD	bam	1.2.1(2013)	Sindi SS et al. (2012)	An integrative probabilistic model for identification of structural variation in sequencing data
GenomeSTRiP	DEL	RP,RD	bam	2.00(2016)	Handsaker RE et al. (2011)	Discovery and genotyping of genome structural polymorphism by sequencing on a population scale
GIndel	DEL,DUP	RP,SR,RD	bam	1.0(2014)	Chu et al. (2014)	GINDEL: accurate genotype calling of insertions and deletions from low coverage population sequence reads
GRIDSS	DEL,DUP,INS,INV,TRA	RP,SR,AS	bam	1.5.0(2017)	Cameron DL et al. (2017)	GRIDSS: sensitive and specific genomic rearrangement detection using positional de Bruijn graph assembly
Gustaf	DEL,DUP,INS	SR	fasta	1.0(2014)	Trappe et al. (2014)	Gustaf: detecting and correctly classifying SVs in the NGS twilight zone
HGT-ID	VEI	RP,SR	bam	1.0(2018)	Baheti S et al. (2018)	HGT-ID: an efficient and sensitive workflow to detect human-viral insertion sites using next-generation sequencing data
Hydra-sv	DEL	RP,AS	bam	0.5.3(2014)	Quinlan AR et al. (2010)	Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome
iCopyDAV	DEL,DUP	RD	bam	1.0(2018)	Dharanipragada P et al. (2018)	iCopyDAV: Integrated platform for copy number variations-Detection, annotation and visualization
indelMINER	DEL	RP,SR	bam	0.10(2015)	Ratan A et al. (2015)	Identification of indels in next-generation sequencing data
inGAP-sv	DEL,DUP,INS,INV	RP,RD	sam	3.1.1(2014)	Qi J et al. (2011)	inGAP-sv: a novel scheme to identify and visualize structural variation from paired end mapping data
ITIS	MEI	RP,SR	bam	1.0(2015)	Jiang C et al. (2015)	ITIS, a bioinformatics tool for accurate identification of transposon insertion sites using next-generation sequencing data
laSV	DEL,DUP,INV,TRA	AS	fastq	1.0.2(2015)	Zhuang J et al. (2015)	Local sequence assembly reveals a high-resolution profile of somatic structural variations in 97 cancer genomes
Lumpy	DEL,DUP,INV,TRA	RP,SR,RD	bam	0.2.9(2015)	Layer RM et al. (2014)	LUMPY: a probabilistic framework for structural variant discovery
Manta	DEL,DUP,INS,INV,TRA	RP,SR,AS	bam	0.29.5(2016)	Chen X et al. (2016)	Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications
MATCHCLIP	DEL,DUP	RP,SR,RD	bam	2.0(2013)	Wu Y et al. (2013)	MATCHCLIP: locate precise breakpoints for copy number variation using CIGAR string by matching soft clipped reads
Meerkat	DEL,DUP,INS,INV,TRA	RP,SR	bam	0.185(2015)	Yang L et al. (2013)	Diverse mechanisms of somatic structural variations in human cancer genomes
MELT	MEI,NUMT	RP,SR	bam	2.0.1(2016)	Gardner EJ et al. (2017)	The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology
MetaSV	DEL,DUP,INS,INV	RP,SR,RD	bam	0.5.2(2016)	Mohiyuddin M et al. (2015)	MetaSV: an accurate and integrative structural-variant caller for next generation sequencing
MindTheGap	INS	k-mer,AS	fastq	0.6447(2014)	Rizk G et al. (2014)	MindTheGap: integrated detection and assembly of short and long insertions
Mobster	MEI,NUMT,VEI	RP,SR	bam	0.2.4.1(2018)	Thung DT et al. (2014)	Mobster: accurate detection of mobile element insertions in next generation sequencing data
mrCaNaVar	DEL,DUP	RD	sam	0.51(2013)	Alkan C et al. (2009)	Personalized copy number and segmental duplication maps using next-generation sequencing
OncoSNP-SEQ	DEL,DUP	RD	bam	2.0(2015)	Yau C (2013)	OncoSNP-SEQ: a statistical approach for the identification of somatic copy number alterations from next-generation sequencing of cancer genomes
Pamir	INS	RP,SR,AS	fastq,bam	unknown(201	Kavak P et al. (2017)	Discovery and genotyping of novel sequence insertions in many sequenced individuals
PBHoney	DEL,INS	LR(SR,SP)	fastq	15.8.24(2015)	English AC et al. (2014)	PBHoney: identifying genomic variants via long-read discordance and interrupted mapping
pbsv	DEL,INS	LR	fastq,bam	Unknown(201	https://github.com/PacificBiosciences	n.a.
PEMer	DEL,DUP,INS	RP	bam	Unknown(200	Korbel et al. (2009)	PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data
PennCNV-Seq	DEL,DUP	RD	bam	Unknown(201	de Araujo Lima L et al. (2017)	PennCNV in whole-genome sequencing data
Pindel	DEL,DUP,INS,INV,TRA	SR	bam	0.2.5(2015)	Ye K et al. (2009)	Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads

Supplementary Table 1 (continued)

PopIns	INS	RP,SR,AS	bam	unknown(201	Kehr B et al. (2016)	PopIns: population-scale detection of novel sequence insertions
PRISM	DEL,INS,INV	RP,SR	sam	1.1.6(2012)	Jiang Y et al. (2012)	PRISM: pair-read informed split-read mapping for base-pair level detection of insertion, deletion and structural variants
RAPTR-SV	DEL,DUP,INS	RP,SR	bam	0.0.15(2015)	Bickhart DM et al. (2011)	RAPTR-SV: a hybrid method for the detection of structural variants
RDXplorer	DEL,DUP	RD	bam	3.2(2011)	Yoon et al. (2009)	Sensitive and accurate detection of copy number variants using read depth of coverage
readDepth	DEL,DUP	RD	bam	0.9.8.4(2015)	Miller CA et al. (2011)	Identification of indels in next-generation sequencing data
RetroSeq	MEI	RP,SR	bam	1.41(2014)	Keane TM et al. (2013)	RetroSeq: transposable element discovery from next-generation sequencing data
SegSeq	DEL,DUP	RD	bam	1.0(2008)	Chiang DY et al. (2008)	High-resolution mapping of copy-number alterations with massively parallel sequencing
Sniffles	DEL,DUP,INS,INV,TRA	LR(SR)	bam	1.0.6(2017)	Sedlazeck FJ et al. (2018)	Accurate detection of complex structural variations using single-molecule sequencing
Socrates	DEL,INS	SR	bam	1.1(2015)	Schroder J et al. (2014)	Socrates: identification of genomic rearrangements in tumour genomes by re-aligning soft clipped reads
SoftSearch	DEL,DUP,INS,INV,TRA	RP,SR	bam	2.4(2014)	Hart SN et al. (2013)	SoftSearch: integration of multiple sequence features to identify breakpoints of structural variations
SoftSV	DEL,DUP,INV,TRA	RP,SR	bam	1.4.2(2015)	Bartenhagen C et al. (2016)	Robust and exact structural variation detection with paired-end and soft-clipped alignments: SoftSV compared with eight algorithms
SoloDel	DEL	RP,RD	bam	1.0.0(2015)	Kim J et al. (2015)	SoloDel: a probabilistic model for detecting low-frequent somatic deletions from unmatched sequencing data
Sprites	DEL	SR	bam	0.3(2016)	Zhang Z et al. (2016)	Sprites:detection of deletions from sequencing data by re-aligning split reads
SV	DEL,DUP	RP,SR,RD	bam	1.4.0(2017)	Antaki D et al. (2017)	SV2: Accurate Structural Variation Genotyping and De Novo Mutation Detection from Whole Genomes
SvABA	DEL,DUP,INS,INV	RP,SR,AS	bam	0.2.1(2017)	Wala JA et al. (2018)	SvABA: genome-wide detection of structural variants and indels by local assembly
SVDetect	DEL,DUP,INS,INV,TRA	RP	bam	0.8b(2013)	Zeitouni B et al. (2010)	SVDetect: a tool to identify genomic structural variations from paired-end and mate-pair sequencing data
SVelter	DEL,DUP,INV	RP,SR,RD	bam	1.1(2016)	Zhao X et al. (2016)	Resolving complex structural genomic rearrangements using a randomized approach
SVfinder	DEL,INS,INV,TRA	RP	bam	unknown(201	Yang R et al. (2014)	Integrated analysis of whole-genome paired-end and mate-pair sequencing data for identifying genomic structural variations in multiple myeloma
SVSeq2	DEL,INS	SR	bam	2.2(2015)	Zhang J et al. (2012)	An improved approach for accurate and efficient calling of structural variations with low-coverage sequence data
Tangram	MEI	RP,SR	bam	0.3.1(2015)	Wu J et al. (2014)	Tangram: a comprehensive toolbox for mobile element insertion detection
Tea	MEI	RP,SR	bam	0.6.2(2015)	Lee E et al. (2012)	Landscape of somatic retrotransposition in human cancers
TEMP	MEI	RP,SR	bam	1.0.5(2017)	Zhuang J et al. (2014)	TEMP: a computational method for analyzing transposable element polymorphism in populations
TIDDIT	DEL,DUP,INV,TRA	RP,SR,RD	bam	1.0.2(2017)	Eisfeldt J et al. (2017)	TIDDIT, an efficient and comprehensive structural variant caller for massive parallel sequencing data
Ulysses	DEL,DUP,INV	RP	bam	1.0(2015)	Gillet-Markowska A et al. (2015)	Ulysses:accurate detection of low-frequency structural variations in large insert-size sequencing libraries
VariationHunter	DEL,INS,INV	RP	divet	0.04(2012)	Hormozdiari F et al. (2009)	Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes
VirusFinder2	VEI	Others	bam	2.0(2015)	Wang Q et al. (2015)	VERSE: a novel approach to detect virus integration in host genomes through reference genome customization
VirusSeq	VEI	RP	fastq	Unknown(201	Chen Y et al. (2013)	VirusSeq: software to identify viruses and their integration sites using next-generation sequencing of human cancer tissue
Wham	DEL,DUP,INS,INV	RP,SR	bam	1.8(2016)	Kronenberg ZN et al. (2015)	Wham: Identifying Structural Variants of Biological Consequence

Supplementary Table 2: Small excerpt of dog diseases found online. Data gathered from OMIA (https://omia.org), (Karlsson, E, and Lindblad-Toh, K 2008) and (Jagannathan, V et al. 2019).

Phenotype	Gene	Breeds	Variant	Author
Exercise-induced metabolic myopathy	*ACADVL*	German Hunting Terrier	nonsense (stop-gain)	Lepori et al. (2018)
Amelogenesis imperfecta	*ACP4*	Akita	n.a.	Hytönen et al. (2019)
Upper airway syndrome	*ADAMTS3*	Norwich Terrier	n.a.	Marchant et al. (2019)
Primary hyperoxaluria type I (Oxalosis I)	*AGXT*	Coton de Tulear,Tibetan spaniel	missense	Johnson, J.L. et al. (2012)
Hypophosphatasia	*ALPL*	Karelian Bear Dog	n.a.	Kyöstilä et al. (2019)
Imerslund–Grasbeck disorder (cobalamin malabsorption)	*AMN*	Giant schnauzer	33 bp coding deletion	Fyfe et al. (2004)
Respiratory distress syndrome	*ANLN*	Dalmatian	nonsense (stop-gain)	Holopainen et al. (2017)
Cyclic hematopoiesis (cyclic neutropaenia)	*AP3B1*	Collie	1 bp coding insertion	Benson et al. (2003)
Progressive retinal atrophy, Basenji	*ARR*	Basenji	extension (stop-lost)	Ying, R.W. et al. (2006)
Ichthyosis	*ASPRV1*	German Shepherd Dog	missense	Bauer et al. (2017a)
Neurodegenerative vacuolar storage disease	*ATG4D*	Lagotto Romagnolo	missense	Kyöstilä et al. (2015)
Glycogen storage disease VII	*ATP-PFK*	American cocker spaniel,English Cocker Spaniel,Wachtelhund,Whippet	missense	Santos, G.A. et al. (1992)
Neuronal ceroid lipofuscinosis (CLN12)	*ATP13A2*	Australian Cattle Dog	splicing	Schmutz et al. (2019)
Spongy degeneration with cerebellar ataxia 2 (SDCA2)	*ATP1B2*	Belgian Shepherd	insertion, gross (>20)	Mauri et al. (2017a)
Renal cancer syndrome	*BHD*	German Shepherd Dog	missense	Cooley, J. et al. (2000)
Coat colour, white spotting, KIT-related	*c-KIT*	German Shepherd Dog	insertion, small (<=20)	Garbade, P. et al. (2013)
Myasthenic syndrome, congenital, owing to CHRNE	*CHRNE*	Heideterrier	insertion, small (<=20)	Herder et al. (2017)
Hyperkeratosis, epidermolytic	*CK-10*	Norfolk terrier	splicing	Jaiswal, A.K. et al. (2005)
Myotonia congenita	*CLCN1*	Labrador Retriever	insertion, small (<=20)	Quitt et al. (2018)
Neuronal ceroid lipofuscinosis, 2	*CLN2*	Dachshund	deletion, small (<=20)	Hayward, J.J. et al. (2011)
Ceroid lipofuscinosis	*CLN8*	English setter	14 bp coding deletion	Katz et al. (2005)
Neuronal ceroid lipofuscinosis	*CLN8*	Alpenländische Dachsbracke	deletion, gross (>20)	Hirz et al. (2017)
Cone degeneration	*CNGB3*	Alaskan malamute	Gene deletion	Sidjanin et al. (2002)
Cone degeneration	*CNGB3*	German short-haired pointer	Missense mutation	Sidjanin et al. (2002)
Skeletal dysplasia 2 (SD2)	*COL11A2*	Labrador Retriever	missense	Frischknecht et al. (2013)
Osteogenesis imperfecta	*COL1A2*	Lagotto Romagnolo	splicing	Letko et al. (2019a)
Alport syndrome	*COL4A5*	Samoyed	Nonsense mutation	Zheng et al. (1994)
Ehlers–Danlos syndrome	*COL5A1*	Labrador Retriever	n.a.	Bauer et al. (2019a)
Ehlers–Danlos syndrome	*COL5A1*	mixed breed	n.a.	Bauer et al. (2019a)
Muscular dystrophy, Ullrich type	*COL6A1*	Landseer	nonsense (stop-gain)	Steffen et al. (2015)
Epidermolysis bullosa (dystrophic form)	*COL7A1*	Golden retriever	Missense mutation	Baldeschi et al. (2003)
Epidermolysis bullosa, dystrophic	*COL7A1*	Central Asian Shepherd	nonsense (stop-gain)	Niskanen et al. (2017)
Intestinal cobalamin malabsorption owing to CUBN mutation	*CUBN*	Beagle	deletion, small (<=20)	Drögemüller et al. (2014b)
Intestinal cobalamin malabsorption owing to CUBN mutation	*CUBN*	Border Collie	deletion, small (<=20)	Owczarek-Lipska et al. (2013)
Epilepsy, generalized myoclonic, with photosensitivity	*DIRAS1*	Rhodesian Ridgeback	deletion, small (<=20)	Wielaender et al. (2017)
X-linked dystophin muscular dystrophy	*DMD*	Golden retriever	Splice-junction point mutation	Sharp et al. (1992)
X-linked hypohidrotic ectodermal dysplasia	*EDA*	Dachshund	splicing	Hadji Rasouliha et al. (2018)
X-linked hypohidrotic ectodermal dysplasia	*EDA*	mixed breed	deletion, small (<=20)	Waluk et al. (2016)
Amelogenesis imperfecta	*ENAM*	Parson Russell Terrier	deletion, small (<=20)	Hytönen et al. (2019)
Osteochondromatosis	*EXT2*	American Staffordshire Terrier	nonsense (stop-gain)	Friedenberg et al. (2018)
Von Willebrand disease I	*F8VWF*	Doberman Pincher	splicing	Brooks, M.B. et al. (2001)
Von Willebrand disease II	*F8VWF*	Chinese Crested Dog,Chinese Crested Dog,German Shorthair Pointer,German Shorthair Pointer, German Wirehaired Pointer,German Wirehaired Pointer	missense	Chiodo, V.A. et al. (2012)
Von Willebrand disease III	*F8VWF*	Dutch Kooiker,Scottish Terrier, Shetland Sheepdog	deletion, small (<=20)	Garosi, L. et al. (2004)
Haemophilia B (factor IX deficiency)	*F9*	Mixed breed	Missense mutation	Evans et al. (1989)
Dental hypomineralization	*FAM20C*	Border Collie	n.a.	Hytönen et al. (2016)
Hyperkeratosis, palmoplantar	*FAM83G*	Irish Terrier, Kromfohrländer	missense	Drögemüller et al. (2014a)
Renal cystadenocarcinoma and nodular dermatofibrosis	*FLCN*	German shepherd	Missense mutation	Lingaas et al. (2003)
Glycogen storage disease Ia	*G6Pase*	Maltese Terriers	missense	Kunz, E. et al. (2011)
Muscular hypertrophy (double muscling)	*GDF8*	Whippet	deletion, small (<=20)	Gong, H. et al. (2011)
Polyneuropathy (LPN2)	*GJA9*	Leonberger	deletion, small (<=20)	Becker et al. (2017)
Bernard–Soulier syndrome	*GP9*	Cocker Spaniel	nonsense (stop-gain)	Gentilini et al. (unpublished data)
Thrombasthenia	*GPIIB*	Great Pyrenees	missense	Chen, N. et al. (1967)

Supplementary Table 2 (continued)

Narcolepsy	*HCRTR2*	Dachshund	Intronic SINE insertion	Lin et al. (1999)
Mucopolysaccharidosis IIIA	*HSS*	Wire-haired dachshund	deletion, small (<=20)	Brenner, S. et al. (1998)
Severe combined immunodeficiency	*IL2RG*	Basset hound	4 bp coding deletion	Henthorn et al. (2003)
Leukocyte adhesion deficiency	*ITGB2*	Irish setter	Missense mutation	Kijas et al. (1999)
Hyperkeratosis, epidermolytic	*K10*	Norfolk terrier	splicing	Jaiswal, A.K. et al. (2005)
Spongy degeneration with cerebellar ataxia 1 (SDCA1)	*KCNJ10*	Malinois	insertion, small (<=20)	Mauri et al. (2017b)
Hyperkeratosis, epidermolytic	*Ker10*	Norfolk terrier	splicing	Jaiswal, A.K. et al. (2005)
Curly hair	*KRT71*	many	missense	Bauer et al. (2019a, 2019b) and Salmela et al. (2019)
Epidermolysis bullosa (junctional form)	*LAMA3*	German short-haired pointer	Intronic repeat insertion	Capt et al. (2005)
Glycogen storage disease VII	*M-PFK*	American cocker spaniel,English Cocker Spaniel,Wachtelhund,Whippet	missense	Santos, G.A. et al. (1992)
Cream coat colour	*MC1R*	Australian Cattle Dog	deletion, small (<=20)	Dürig et al. (2018)
Drug sensitivity (Ivermectin)	*MDR1*	Collie	4 bp coding deletion	Mealey et al. (2001)
Invermectin sensitivity	*MDR1*	Australian Shepherd,Border Collie, Collie,German Shepherd Dog, Longhaired whippet,McNab shepherd, mixed breed,Old English Sheepdog, Shetland Sheepdog,Silken windhound,Waller,White Swiss shepherd	insertion, small (<=20)	Hässig, M. et al. (2005)
Phaeomelanin dilution	*MFSD12*	Many	n.a.	Hédan et al. (2019)
Neuronal ceroid lipofuscinosis	*MFSD8*	Chihuahua	deletion, small (<=20)	Karli et al. (2016)
Lethal acrodermatitis	*MKLN1*	Bull Terrier and Miniature Bull Terrier	splicing	Bauer et al. (2018a)
Coat colour dilution	*MLPH*	Chow Chow, Sloughi, Thai Ridgeback	splicing	Bauer et al. (2018b)
Menkes disease	*MNK*	Labrador Retriever	missense	Gower, S. et al. (1996)
Fanconi syndrome	*MTMR15*	Basenji,Irish Wolfhound	deletion, gross (>20)	Baumgärtner, W. et al. (2004)
Copper toxicosis	*MURR1*	Bedlington terrier	One exon deleted	van De Sluis et al. (2002)
Wilson disease, COMMD1 type	*MURR1*	Bedlington Terrier	deletion, gross (>20)	Agerholm, J.S. et al. (2002)
Leukoencephalomyelopathy	*NAPEPLD*	Great Dane and Rottweiler	insertion, small (<=20)	Minor et al. (2018)
Progressive retinal atrophy	*NECAP1*	Giant Schnauzer	n.a.	Hitti et al. (2019)
Epilepsy (Lafora type)	*NHLRC1*	Minature wire-haired dachshund	Tandem 12-bp repeat expansion	Bradbury et al. (2005)
Myoclonus epilepsy of Lafora	*NHLRC1*	Chihuahua	repeat variation	Barrientos et al. (2019)
Primary ciliary dyskinesia	*NME5*	Alaskan Malamute	n.a.	Anderegg et al. (2019)
Primary hereditary cataract	*NOL3*	Australian Shepherd,Australian Shepherd,Boston Terrier,Boston Terrier,Staffordshire Bull Terrier, Staffordshire Bull Terrier	deletion, small (<=20)	Coulson, N.R. et al. (2009)
CHILD-like syndrome	*NSDHL*	Labrador Retriever	deletion, gross (>20)	Bauer et al. (2017b)
Oculocutaneous albinism II	*OCA2*	German Spitz	splicing	Caduff et al. (2017a)
Goniodysgenesis	*OLFML3*	Border Collie	n.a.	Pugh et al. (2019)
Invermectin sensitivity	*p-gp*	Australian Shepherd,Border Collie, Collie,German Shepherd Dog, Longhaired whippet,McNab shepherd, mixed breed,Old English Sheepdog, Shetland Sheepdog,Silken windhound,Waller,White Swiss shepherd	insertion, small (<=20)	Hässig, M. et al. (2005)
ADP response impaired; Postoperative hemorrhage	*P2Y12*	Greater Swiss Mountain	deletion, small (<=20)	Esquerré, D. et al. (2011)
Histiocytosis, malignant	*P53*	Bernese Mountain dog,Flat-coated retriever,Golden Retriever,Rottweiler	insertion, small (<=20)	Permi, P. et al. (2012)
Rod-cone dysplasia 1a	*PDBS*	Sloughi	insertion, small (<=20)	Hu, X. et al. (2012)
Cone-rod dystrophy 1	*PDBS*	American Staffordshire terrier	deletion, small (<=20)	Hu, X. et al. (2012)
Rod-cone dysplasia 1	*PDBS*	Irish Setter	nonsense (stop-gain)	Muir, W.W. et al. (1975)
Rod–cone dysplasia 3	*PDE6A*	Cardigan Welsh corgi	1 bp coding deletion	Petersen-Jones et al. (1999)
Rod–cone dysplasia 1	*PDE6B*	Irish setter	Nonsense mutation	Suber et al. (1993)
Rod-cone dysplasia 3	*PDEA*	Cardigan Welsh Corgi	deletion, small (<=20)	Hu, X. et al. (2012)
Rod-cone dysplasia 1	*PDEB*	Irish Setter	nonsense (stop-gain)	Muir, W.W. et al. (1975)
Rod-cone dysplasia 1a	*PDEB*	Sloughi	insertion, small (<=20)	Hu, X. et al. (2012)
Cone-rod dystrophy 1	*PDEB*	American Staffordshire terrier	deletion, small (<=20)	Hu, X. et al. (2012)
Glycogen storage disease VII	*PFK-M*	American cocker spaniel,English Cocker Spaniel,Wachtelhund,Whippet	missense	Santos, G.A. et al. (1992)
Dilated cardiomyopathy	*PLN*	Welsh Springer Spaniel	n.a.	Yost et al. (2019)
Shaking puppy (generalized tremor)	*PLP1*	English Springer spaniel	Missense mutation	Nadon et al. (1990)
Coat colour, merle	*Pmel17*	Shetland Sheepdog	insertion, gross (>20)	Niu, L. et al. (1982)
Pyruvate dehydrogenase deficiency	*PPM2C*	Clumber Spaniel,Sussex Spaniel	nonsense (stop-gain)	Wang, T. et al. (2000)
Photoreceptor dysplasia	*PPT1*	Miniature Schnauzer	n.a.	Murgiano et al. (2019)
Centronuclear myopathy	*PTPLA*	Labrador retriever	SINE insertion in exon	Pele et al. (2005)

Supplementary Table 2 (continued)

Type II fiber deficiency; Autosomal recessive muscular dystrophy; Hereditary myopathy of Labrador retrievers (HMLR)	*PTPLA*	Labrador Retriever	insertion, gross (>20)	Jia, B.Y. et al. (2003)
Deafness	*PTPRQ*	Doberman Pinscher	insertion, small (<=20)	Guevar et al. (2018)
Polyneuropathy, ocular abnormalities and neuronal vacuolation	*RAB3GAP1*	Alaskan Husky	deletion, small (<=20)	Wiedmer et al. (2015)
Cone-rod dystrophy 1	*RCD-1*	American Staffordshire terrier	deletion, small (<=20)	Hu, X. et al. (2012)
Rod-cone dysplasia 1	*RCD-1*	Irish Setter	nonsense (stop-gain)	Muir, W.W. et al. (1975)
Rod-cone dysplasia 1a	*RCD-1*	Sloughi	insertion, small (<=20)	Hu, X. et al. (2012)
Autosomal dominant PRA	*Rho*	English mastiff	Missense mutation	Kijas et al. (2002)
Autosomal dominant PRA	*RHO1*	Bull Mastiff,English Mastiff	missense	Lohi, H. et al. (2009)
Congenital night blindness	*RPE65*	Briard	4 bp coding deletion	Veske et al. (1999)
Progressive retinal atrophy, Basenji	*sag1*	Basenji	extension (stop-lost)	Ying, R.W. et al. (2006)
Van den Ende–Gupta syndrome	*SCARF2*	Wirehaired Fox Terrier	deletion, small (<=20)	Hytönen et al. (2016)
Spinocerebellar ataxia	*SCN8A*	Alpenländische Dachsbracke	n.a.	Letko et al. (2019b)
Hypotrichosis, recessive	*SGK3*	Scottish Deerhound	deletion, small (<=20)	Hytönen & Lohi (2019)
Coat colour, merle	*SILV*	Shetland Sheepdog	insertion, gross (>20)	Niu, L. et al. (1982)
Eye malformation, congenital	*SIX6*	Golden Retriever	n.a.	Hug et al. (2019)
Leigh-like subacute necrotising encephalopathy	*SLC19A3*	Yorkshire Terrier	insertion, small (<=20)	Drögemüller et al. (2019)
Craniomandibular osteopathy	*SLC37A2*	West Highland White Terrier, Scottish Terrier, Cairn Terrier	splicing	Hytönen et al. (2016)
Coat colour, albinism, oculocutaneous type IV	*SLC45A2*	Bull Mastiff	deletion, gross (>20)	Caduff et al. (2017b)
Brachycephaly	*SMOC2*	Many	insertion, gross (>20)	Marchant et al. (2017)
Nasal parakeratosis	*SUV39H2*	Greyhound	missense	Bauer et al. (2018c)
Nasal parakeratosis	*SUV39H2*	Labrador Retriever	deletion, small (<=20)	Jagannathan et al. (2013)
Neuroaxonal dystrophy	*TECPR2*	Spanish Water Dog	missense	Hahn et al. (2015)
Cerebellar hypoplasia	*VLDLR*	Eurasier	deletion, small (<=20)	Gerber et al. (2015)
Canine multifocal retinopathy	*VMD2*	Coton de Tulear	missense	Mahla, R. et al. (2012)
Canine multifocal retinopathy	*VMD2*	Lapponian Herder	deletion, small (<=20)	Mahla, R. et al. (2012)
Canine multifocal retinopathy	*VMD2*	American Bulldog,Australian Shepherd,Boerboel,Bull Mastiff,Dogue de Bordeaux,English Bulldog,English Mastiff,Great Pyrenees,Italian Cane Corso,Perro de Presa Canario	nonsense (stop-gain)	Mahla, R. et al. (2012)
Von Willebrand disease type II	*VWF*	German pointers	Missense mutation	Kramer et al. (2004)
Von Willebrand disease type III	*VWF*	Scottish terrier	1 bp coding deletion	Venta et al. (2000)

Supplementary Figure 1: Cartoon of our read depth pipeline. The mapped RD is normalized to the average RD and corrected for GC content. The window values are discretized using integer, CN-valued gaussian mixtures as HMM hidden states under a Baum-Welch trained model. The 95% highest posterior density intervals are emitted and bulk-corrected using all available samples in 5-window intervals. For a more detailed description see the methods on Section 3.1.