



## The wolf reference genome sequence (*Canis lupus lupus*) and its implications for *Canis* spp. population genomics

Gopalakrishnan, Shyam; Samaniego Castruita, Jose Alfredo; Sinding, Mikkel Holger Strander; Kuderna, Lukas F. K.; Räikkönen, Jannikke; Petersen, Bent; Sicheritz-Pontén, Thomas; Larson, Greger; Orlando, Ludovic Antoine Alexandre; Marques-Bonet, Tomas

*Total number of authors:*  
13

*Published in:*  
B M C Genomics

*Link to article, DOI:*  
[10.1186/s12864-017-3883-3](https://doi.org/10.1186/s12864-017-3883-3)

*Publication date:*  
2017

*Document Version*  
Publisher's PDF, also known as Version of record

[Link back to DTU Orbit](#)

*Citation (APA):*  
Gopalakrishnan, S., Samaniego Castruita, J. A., Sinding, M. H. S., Kuderna, L. F. K., Räikkönen, J., Petersen, B., Sicheritz-Pontén, T., Larson, G., Orlando, L. A. A., Marques-Bonet, T., Hansen, A. J., Dalén, L., & Gilbert, M. T. P. (2017). The wolf reference genome sequence (*Canis lupus lupus*) and its implications for *Canis* spp. population genomics. *B M C Genomics*, 18, Article 495. <https://doi.org/10.1186/s12864-017-3883-3>

---

### General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal


If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

RESEARCH ARTICLE

Open Access



# The wolf reference genome sequence (*Canis lupus lupus*) and its implications for *Canis spp.* population genomics

Shyam Gopalakrishnan<sup>1</sup> , Jose A. Samaniego Castruita<sup>1</sup>, Mikkel-Holger S. Sinding<sup>1,2</sup>, Lukas F. K. Kuderna<sup>3,4</sup>, Jannikke Räikkönen<sup>5</sup>, Bent Petersen<sup>6</sup>, Thomas Sicheritz-Ponten<sup>6</sup>, Greger Larson<sup>7</sup>, Ludovic Orlando<sup>1</sup>, Tomas Marques-Bonet<sup>3,4,8</sup>, Anders J. Hansen<sup>1</sup>, Love Dalén<sup>9</sup> and M. Thomas P. Gilbert<sup>1,10,11\*</sup>

## Abstract

**Background:** An increasing number of studies are addressing the evolutionary genomics of dog domestication, principally through resequencing dog, wolf and related canid genomes. There is, however, only one de novo assembled canid genome currently available against which to map such data - that of a boxer dog (*Canis lupus familiaris*). We generated the first de novo wolf genome (*Canis lupus lupus*) as an additional choice of reference, and explored what implications may arise when previously published dog and wolf resequencing data are remapped to this reference.

**Results:** Reassuringly, we find that regardless of the reference genome choice, most evolutionary genomic analyses yield qualitatively similar results, including those exploring the structure between the wolves and dogs using admixture and principal component analysis. However, we do observe differences in the genomic coverage of re-mapped samples, the number of variants discovered, and heterozygosity estimates of the samples.

**Conclusion:** In conclusion, the choice of reference is dictated by the aims of the study being undertaken; if the study focuses on the differences between the different dog breeds or the fine structure among dogs, then using the boxer reference genome is appropriate, but if the aim of the study is to look at the variation within wolves and their relationships to dogs, then there are clear benefits to using the de novo assembled wolf reference genome.

**Keywords:** *Canis lupus*, Evolutionary genomics, Genome, Wolf

## Background

In light of the ever-decreasing cost of high-throughput DNA sequencing, it is now possible to undertake large-scale genomic studies at not only the population level, e.g. [1, 2], but also the population paleogenomic level, e.g. [3–8]. While these datasets are being exploited across a growing range of applied questions, a number of research groups are beginning to also focus on how to interpret and treat this data in a way that minimizes biases, and thus yields robust inferences from the data.

Several human population genomic datasets have noted the existence of biases that arise when mapping

the resequenced genomes of diverse individuals to a reference genome based on a single individual. Alignment against a single reference genome can lead to different samples appearing more similar to the reference genome, and underestimating the variation present in samples that come from a different population or species than the reference genome [9–11]. New mapping techniques are being developed to overcome these biases by allowing mapping to multiple genomes [12]. These methods rely on a high number of sequenced and de novo assembled samples, or a catalogue of polymorphisms for all the populations in the study. For species other than humans, such resources are scarce. Ultimately, these biases imply that thorough annotation of all variation in a genomics data set requires every individual to be represented by a de novo assembly [13–15]. Though this ideal is not feasible for a variety of

\* Correspondence: [tgilbert@snm.ku.dk](mailto:tgilbert@snm.ku.dk)

<sup>1</sup>Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, Øster Voldgade 5-7, 1350 Copenhagen, Denmark

<sup>10</sup>Trace and Environmental DNA Laboratory, Department of Environment and Agriculture, Curtin University, Perth, Western Australia, Australia  
Full list of author information is available at the end of the article



economic reasons, there is a need to broaden the pool of reference genomes to ensure that we can minimize the effects of these biases on downstream analyses.

A research discipline where population genomics is rapidly making significant contributions is the study of domestication – a topic that has long held academic interest due to both its applied relevance and its broad general public appeal. Genomic and paleogenomic resources have previously been used to address major questions in domestication, including deciphering the population structure and admixture patterns in modern and wild lineages [16–18], discovering structure among ancient pre-domestic lineages [6, 19–22], and estimating levels of introgression from wild lineages into domesticated stocks [17, 23], applied to a multitude of species, such as maize [6, 16, 22], silkworms [24], chickens [25–27], and pig [28, 29].

Although these analyses can offer powerful insights into the domestication process, they come with their own sets of challenges. While the major challenge is the need to account for genetic diversity that has been lost as a result of full or partial extinctions of original wild lineages, mapping biases arising from experimental design, such as choice of reference genome, also pose a hurdle to robust analyses. At least one domestication related study has demonstrated that these effects can be considerable. In Orlando and colleagues' [19] study of the genomic sequences of six horses (one from a pre-domestication Pleistocene sample), they showed how a variety of analyses such as D statistics, population divergence and heterozygosity estimation, led to different results when their resequenced genomes were mapped to the EquCab2.0 [30] reference genome, and a de novo assembly of the donkey genome. They attribute many of these biases to differences in how closely related the samples are to the horse reference genome. This problem is exacerbated in studies that include ancient, pre-domestication samples since the reference genomes are predominantly constructed using modern samples. Another difference in the reference genomes that might lead to different results in downstream analyses, is the technology used to generate the reference genome. Many older reference genomes were generated using Sanger sequencing while the newer reference genomes and resequenced genomes in studies have been generated using Illumina short read sequencing technology. Although the underlying causes for the biases remain unresolved, one powerful approach is to perform the analyses using several different closely related reference genomes, thus accounting for biases introduced by the mapping procedures and ensuring that the results are consistent across the choice of reference genomes.

With regards to the need for multiple reference genomes, while a number of genomics studies have

recently been published that relate to the relationship between dogs and wolves, the sequence data from genome resequencing studies [21, 31–34] has either been mapped to the only currently available reference genome, that of the Boxer dog (CanFam3.1) [35], or compared to data drawn from SNP (Single Nucleotide Polymorphism) chip arrays developed to target variation in dog genomes [36, 37]. The results of such studies show that dogs are monophyletic with respect to wolves, and indicate the existence of a deep split between the modern wolf and dog lineages, and a deep split within the dogs as well [21].

There are still several questions regarding wolf and dog phylogeny, population history and domestication that remain unanswered. Although the results of these studies are largely consistent, there are some inconsistencies in the findings regarding the location and the time of the domestication event [21, 36, 38, 39]. It has also been suggested that the population of wolves that are ancestral to the modern dogs may be extinct [21, 32, 34].

It is possible that one explanation for discrepancies between studies is that important structural variation in the wolf genome is missed or misplaced by mapping to a dog reference, or targeting SNPs developed for dog variation. To test this hypothesis, we de novo generated the first wolf reference genome, then remapped the genomic datasets previously published by Wang et al., Freedman et al. and Zhang et al. [31–33]. We subsequently re-analysed the published and remapped data in the context of divergence, admixture and systematics, in order to explore whether any reference genome-specific biases occur.

## Results

### De novo reference genome assembly

In order to construct a de novo reference genome using a wolf, we generated a combination of 5–8 kilobases and 3 kilobases mate pair libraries, as well as 650 basepair and 180 basepair insert libraries. These were sequenced with 101 basepair paired end reads using 5 lanes of a Illumina HiSeq 2500, where one lane was allocated to the multiplexed mate-pair libraries, one lane to the 650 basepair insert library and the remaining three lanes were allocated for the 180 basepair insert libraries. Overall, this generated a 30× coverage of the genome. The de novo reference genome was assembled using the ALLPATHS-LG assembler [40]. The final assembly consisted of 8747 scaffolds, of which 8569 scaffolds were longer than 1 kilobase. The longest scaffold was 12.88 megabases. The scaffold N50 of the assembly is 1.56 megabases and the scaffold N80 of the assembly was 512 kilobases., the contig N50 of the assembly was 94 kilobases and the contig N80 of the assembly was 34 kilobases. The total length of the assembly was 2.34 gigabases, while the scaffolds longer

than 1 kilobase covered more than 99.99% of the assembly.

### Landscape of common repeats

To compare abundances of repetitive elements between the wolf assembly and canFam3, we sought to detect common interspersed repeats in both of them. We identify 902 megabases of repetitive elements along the wolf assembly, corresponding to 39.8% of the non-gapped assembly. We detect a similar, albeit slightly higher amount of repeats in canFam3 (1009 megabases, or 42.1% of the non-gapped assembly). When stratifying repetitive elements by their respective superfamilies, we observe similar abundances in the wolf and the dog assembly (see Additional file 1: Figure S2), with the exception of satellite sequences, a family of repetitive elements most commonly found in the telomeric and centromeric regions of the chromosomes. To investigate the patterns underlying the differences in repeat annotations, we calculated the evolutionary distance of each annotation to its consensus sequence. Overall, the divergence landscapes are very similar, however, we observe a depletion of young and highly identical long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs) insertions in the wolf assembly, most likely as an artifact of sequencing and assembly strategy (see Additional file 1: Figure S3).

### Mapping, coverage statistics

Since the choice of reference genome directly affects the mapping process, we compared the efficiency of mapping previously published short reads to the reference genome when using one of the two genomes used in this study, viz., the dog reference genome [35] and the de novo assembled wolf reference genome. We compared the proportion of uniquely mapped reads for each sample and the depth of coverage across the genome. As shown in Table S1 (Additional file 1: Table S1), we find that the samples that come from the same sub-species as the reference genome, i.e. dogs when using the dog reference genome and wolves when using the wolf reference genome, have a higher proportion of reads that map uniquely to the genome. As a result, they also have a slightly higher coverage across the genome. Note that we do not find a large difference in coverages or proportions of reads that map uniquely, and the effect is consistent across all samples.

### PCA

We performed a principal components analysis (PCA) to identify the major axes of variation in the genotype data. Fig. 1 shows the results of the PCA using data mapped to either the reference dog or the de novo wolf genome assembly. For this analysis, we used only common

variants with minor allele frequency greater than 0.05. Irrespective of the reference genome used for the alignment, the first two principal components separate dogs from wolves. The proportions of variance explained by the first and second principal components are also very similar across the choice of the two reference genomes (see Fig. 1). Changing the missingness or allele frequency threshold leads to qualitatively similar results (Additional file 1: Figure S4).

### Heterozygosity

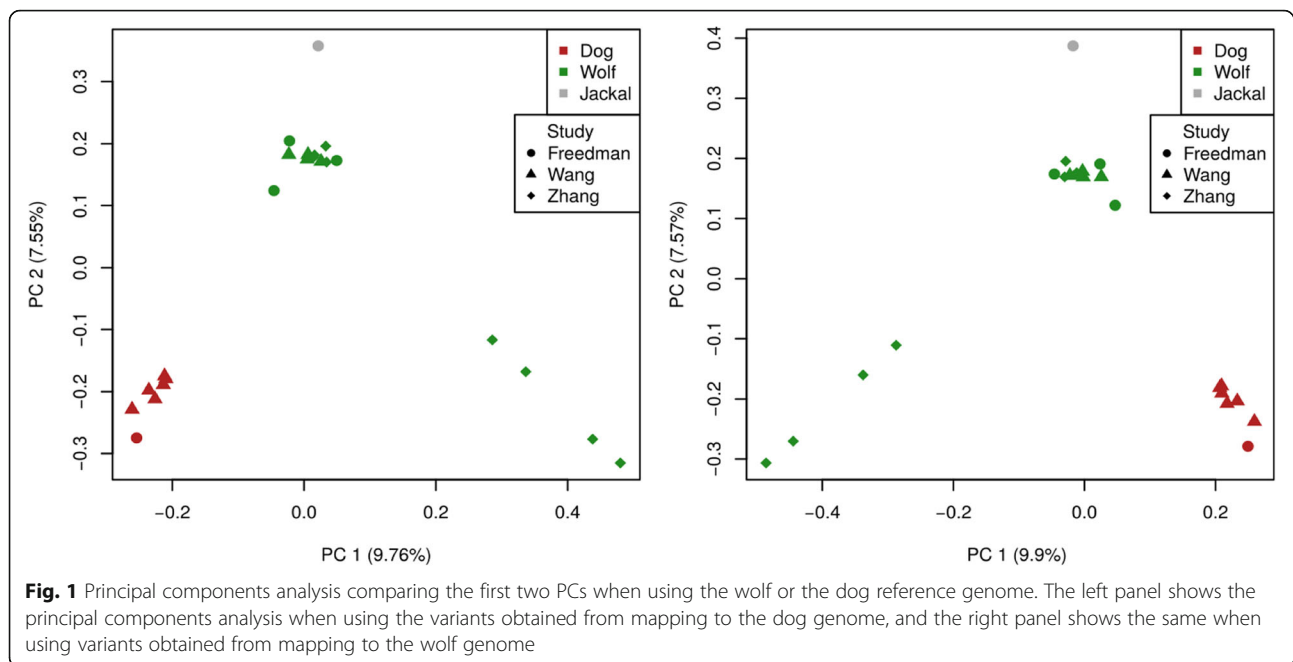
We compared the estimates of per-sample heterozygosity using alignments to the two different reference genomes. Table S1 (Additional file 1: Table S1) shows that the estimated heterozygosity of the samples depends upon the reference genomes used for mapping. The heterozygosity estimates for dogs are consistently higher by up to 10% when using the dog reference genome compared to the de novo wolf genome assembly.

### Population size

We additionally used the pairwise sequentially markovian coalescent (PSMC) [41] to explore the effect of reference genome on the estimated population size history of the populations that the resequenced individuals were obtained from. Figure 2 shows the reconstructed population size history for a subset of the samples in our study. The comparison of the population sizes shows that the estimates obtained are largely consistent. For the dogs in this study, the population size trajectories estimated using the two different reference genomes coincide beyond 10kya. However, the effective population sizes for the wolves are a bit lower when using the wolf reference genome, compared to the same when using the dog reference genome. We observed reference genome specific differences in the recent histories, which can be attributed to the difference in the rare/private variants discovered in the two species when using the different reference genomes. If the primary effect of changing the reference genome is in the number of rare variants discovered, the effect on analyses such as PSMC will be greatest in the recent population size estimates. As PSMC does not have the power to estimate these parameters well, the effect of this bias is not expected to be high in this analysis.

### Phylogeny

We used *RAxML* and *ExaML* [42, 43] to estimate the phylogenetic relationships between samples using the variants identified by aligning to the wolf or the dog reference genome. Since our analysis only uses variant sites, we accounted for the ascertainment scheme of the variants using the ascertained version of the GTRGAMMA model of sequence evolution. As shown in



Additional file 1: Figure S1, the overall topology of the resulting phylogenies differ depending on the choice of the reference genome. Specifically, when using the dog reference genome the dogs and wolves are reciprocally monophyletic. While using the de novo assembled wolf reference genome, the dogs were monophyletic with respect to the wolves but the wolves were not monophyletic with respect to the dogs. Note that the support values for these nodes that differ between the two topologies have very low bootstrap support values. Additionally, using a neighbour joining approach to estimate the phylogenetic relationships led to qualitatively similar results (data not shown).

### Admixture

We estimated the ancestry proportions in the 23 samples using ngsAdmix [44]. When using two ancestry components for estimating admixture proportions, dogs and wolves are split into two different clusters for both choices of reference genome. In both cases, all the wolves, except for the high altitude wolves from the Zhang study [33], show up to 20% of the estimated dog ancestral component (Fig. 3). Increasing the number of estimated ancestral components from two to three leads to similar results, with the dogs and the wolves being separated into two clusters. Additionally, the wolves split into two clusters where the high altitude wolves are separated from the rest of the wolves. Further, the contribution of the estimated dog ancestry components in the wolves becomes negligible.

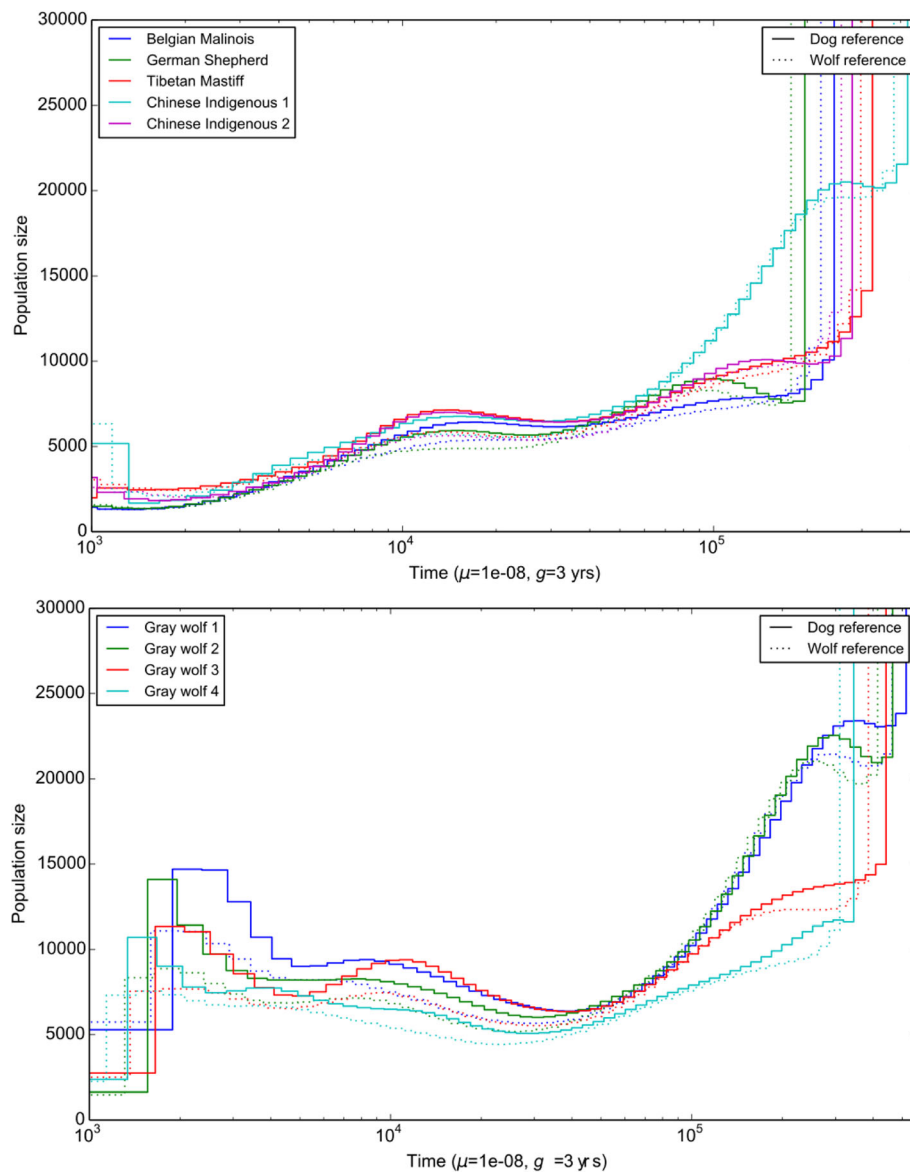
When estimating admixture with four ancestry clusters, the choice of the reference genome has an impact

on the qualitative outcome of the admixture analyses. When using the de novo wolf reference genome, the newly added ancestry component separates the golden jackal (*Canis aureus*) from the other samples, whereas using the boxer dog reference genome reveals additional structure in the wolves, with the golden jackal assigned to one of the clusters containing the wolves. When estimating a higher number of ancestry components, the additional ancestry components explain variance in dogs if the dog reference genome was used and conversely, the use of the de novo wolf reference genome leads to additional structure in the wolves.

### Discussion

Previous studies have speculated that the choice of reference genome has wide ranging effects, especially on the identification of population structure and the timing of demographic events in studies using multiple related species. This problem is expected to be exacerbated when the reference genome is closer to some species in the study than others. Given that there is currently a considerable amount of effort being applied to the sequencing and analysis of dog and wolf genomes, we decided to both explore the impact of the phenomenon in general, and specifically explore whether it holds implications for the results of several relevant previously published dog and wolf genome studies. In this regard, because the time of divergence between dogs and wolves is relatively recent (a conservative estimate of the divergence time is around 35,000 years ago [31, 34]) and the genetic divergence between the extant wolves and modern dogs is low, we did not, a priori, expect the choice of





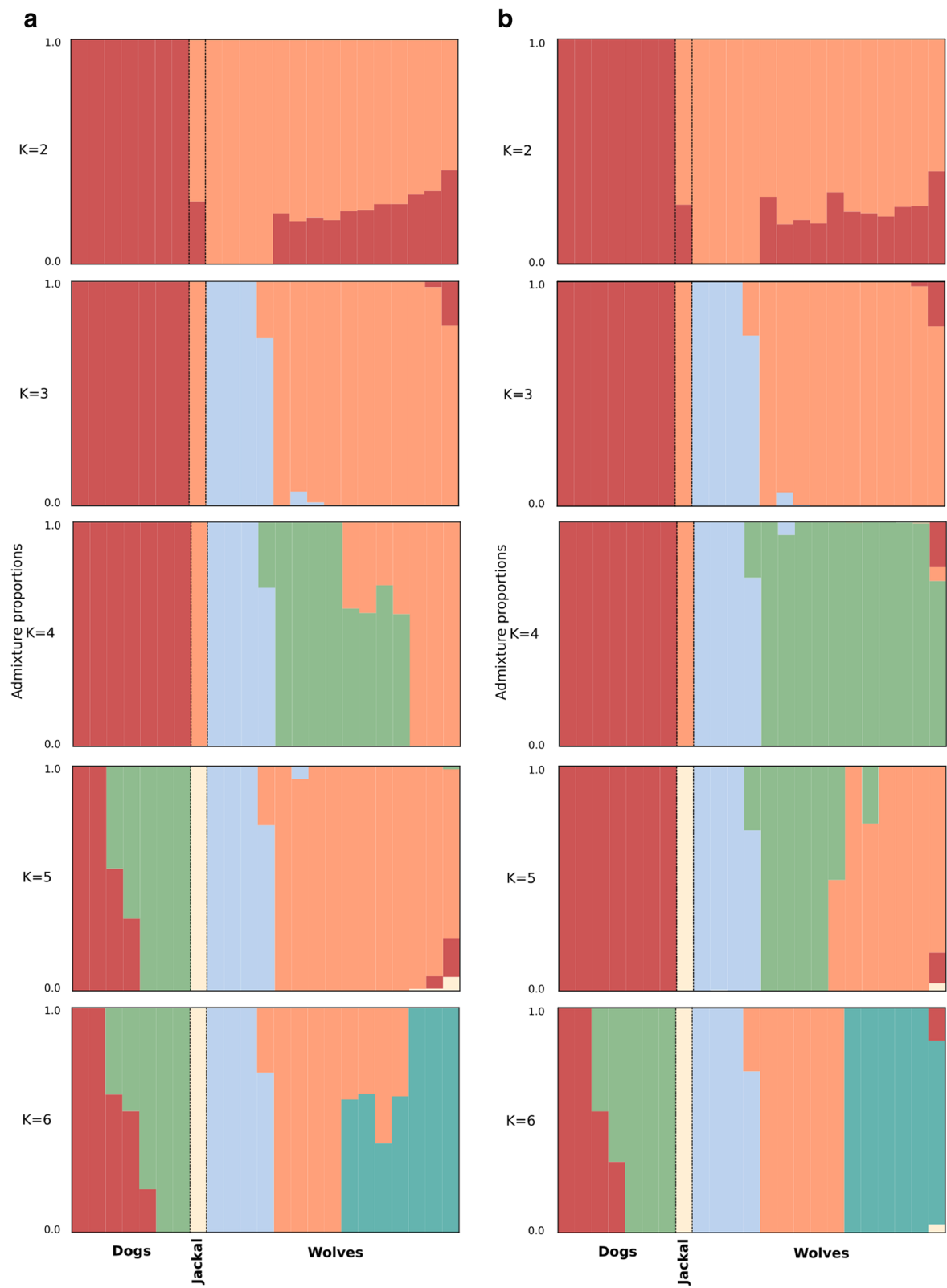
**Fig. 2** Effective population sizes estimated using PSMC. The left panel shows the effective population sizes for dogs in the Wang dataset, estimated using the data mapped to both the reference genomes. The right panel shows the population size estimates for the wolves in the Wang dataset, when using the data mapped to both the reference genomes

the reference genome to have a big impact on the qualitative inferences in the standard population genetics analyses. Overall, our findings bear this expectation out - the analyses that are primarily driven by common variation, such as principal components analysis and admixture analysis with low number of clusters result in very similar findings across the two reference genomes.

Nevertheless, since these two species are genetically very similar, the rare and/or private variation is informative for the differences between the two species. Regarding these variants, the choice of reference genome is clearly more important than for the common variants. As shown in both the table of heterozygosity

(Additional file 1: Table S1) and the results from admixture analyses with higher number of estimated ancestry clusters (Fig. 3), the rare variation in the two datasets can lead to qualitatively different results. This is especially evident in the admixture analyses with four or more clusters, where the structure that is revealed is dependent on the choice of the reference genome. Using the data aligned to the dog genome results in earlier identification of structure in dogs, and vice versa.

One main concern when interpreting these results is the differences in the quality of the two reference assemblies. Clearly, the dog reference genome is in a much more mature state than our de novo assembly of the



**Fig. 3** Admixture plot showing the estimated ancestry components. The plots on the left, panel **a**, show the estimates obtained when the dog reference genome is used, whereas the figures on the right, panel **b**, were obtained using the wolf reference genome

wolf reference genome. This difference in quality could lead to biases in the analyses, especially analyses that require large continuous regions with variant calls, e.g., effective population size estimation using PSMC as well as characterization of inbreeding levels using runs of homozygosity. Although the effective population size estimates are consistent for the two reference genomes, the difference in quality of assembly could result in different estimates in the most recent time periods, where the methods are typically underpowered.

The effect of the choice of the reference genome seems to be limited to analyses that rely on low frequency and private variants. When comparing the effect from mapping against wolf and dog reference genomes, we found the largest effect in the higher order structure identified in the wolves or dogs when estimating ancestry components. At lower number of ancestry components, the choice of reference genome had no effect on the identification of clusters.

In this study, neither of the two reference genomes used were equally distant from the wolves and dogs samples analysed. Ideally, one could use the genome of a relatively close outgroup – the golden jackal in our case – to ensure that there are no biases introduced due to the choice of the reference genome. Although this would avoid the pitfalls of choosing a reference genome that is more close to some of the samples than others, it may not be feasible in many cases, e.g. due to the relatively high economic and computational costs of generating outgroup genomes, or the absence of an appropriate outgroup. Since the reference genomes for most studies tend to not be equally distant from all samples, it is important to account for the biases while interpreting the findings from population and phylogenetics analyses.

## Conclusions

We have generated the first de novo assembled wolf reference genome, which will be a useful resource for future studies exploring the genomic structure and relationship between dogs, wolves and other canids. Since the two species that are the focus of this paper are so closely related, the effect of the reference genome was minimal on many of the downstream analyses such as PCA and estimating the phylogeny of the samples. However, some analyses like admixture showed the effects of the reference genome at higher number of clusters. Since the use of the wolf reference genome results in identification of population structure that is hidden when using the dog reference genome, we recommend the use of the de novo wolf reference genome for any studies where the focus is on identifying the relationships between wolves and dogs or teasing apart the relationship between the various wolves of the world.

## Methods

### De novo reference genome assembly

We used a muscle sample from a Swedish wolf to construct (Additional file 1: Figure S5) our de novo reference genome. The sample (specimen ID: NRM201105024) was obtained from the Environmental Specimen Bank at the Swedish Museum of Natural History, and originated from a male yearling shot during a licensed hunt in Värmland, Sweden in January 2011. The individual (Grimsö ID: D-11-21) was born in the Jangen 5 territory [45]. The pedigree-based inbreeding level (F) for the offspring born in this territory has been estimated to be  $F = 0.30$  [46].

In order to generate a de novo wolf reference genome assembly, we created libraries with different insert sizes, viz., one 5–8 kb mate pair library, one 3 kb mate pair library, and 650 bp and 180 bp insert libraries. In all, these libraries were sequenced using 5 lanes of HiSeq 2500. Using the short reads generated from these libraries, the de novo reference genome was assembled at NGI Stockholm using the ALLPATHS-LG [40] assembler. Different assemblers were tested before choosing the genome assembled using the ALLPATHS-LG assembler, based on statistics of the assembly, such as the number of scaffolds, the N50, N80 and total assembly length.

### Repeat identification

To identify common genomic interspersed repeats in the wolf assembly, we ran RepeatMasker [47] (version 4.0.6) with RMBLAST (version 2.2.27+) as engine. We used the dog-specific repeat libraries derived from the latest available Repbase database (version 20,160,829, available at [www.girinst.org](http://www.girinst.org)). To put our results into context, we also identified interspersed repeats with the exact same approach in the latest dog reference genome assembly canFam3 [35], as published annotations might differ slightly in parameter settings or engine.

### Resequencing data used in comparisons

To quantify the effects of the choice of the reference genome on the downstream analyses, we used publicly available datasets that contained whole genome sequences for canids. We used the raw short reads from the sequenced wolves and dogs from Freedman et al., Wang et al. and Zhang et al. [31–33]. From Freedman et al., we obtained the short reads for the 6 canids that were whole-genome sequenced as part of their study, viz., a dingo, an Israeli wolf, a Croatian wolf, a Chinese wolf, a basenji and a golden jackal. Of these samples, we did not use the basenji due to data corruption issues. From the Wang et al. study, we downloaded the short reads for four gray wolves, three Chinese indigenous dogs, two European dog breeds – the German shepherd and the Belgian Malinois – and a Tibetan mastiff. We also obtained the shotgun short reads from the whole genomes of 8 Chinese



wolves that were sequenced as part of the Zhang et al. study.

The details of the samples included in our study, including the sequencing depth of coverage, and the source from which the data were obtained are given in Table S1 (Additional file 1: Table S1).

### Data processing

Since the different datasets that were used in this study were obtained from various different sources, we built a custom processing pipeline to ensure that all the data were processed using the same tools and were subject to the same quality control and filtering. We built our pipeline on the Paleomix pipeline developed by Schubert et al. [48]. The various parts of our processing pipeline are detailed below.

### Mapping

Each sample used in this study was mapped against both the dog reference genome (canFam3.1) [35] and the de-novo assembled wolf reference genome. We used the Paleomix pipeline to map the short reads from the samples to both the genomes. Specifically, we used bwa-0.7.10 (the aln algorithm) [49] to map the reads to the genome. After the initial alignment step, we discarded any reads that did not map uniquely to the reference genome. Using only the uniquely mapped reads, we used GATK [50] to perform an indel realignment step to account for increased error rates in reads whose ends span an indel. As there are no available curated sets of indels for the dog or the wolf populations, we did not use an external database of indels for indel realignment.

Since the aim of this paper is to compare the choice of reference genome, mapping to the reference genome is a critical step in the bioinformatics processing. Any biases or errors introduced as part of the mapping process will propagate to the downstream analysis resulting in incorrect inferences. In order to ensure that we did not introduce any such biases, we used exactly the same settings while mapping to the wolf or the dog reference genomes.

### Genotyping

After mapping the reads to the reference genome, we called genotypes for all sites in the genome. Since the samples in this study consist of different populations and species, each sample was processed independently. For each sample, the positions in the genome that were covered by at least 5 reads were genotyped. At each such site, the genotype was called using *samtools-0.1.19* [49, 51]. We used a minimum genotype quality threshold of 30 to filter out low quality genotypes from each sample.

### Variant identification and quality filtering

Using the initial set of genotypes for all samples at all sequenced sites, we identified variant sites in the study

sample. We did not use a multi-sample variant caller since we have a heterogeneous set of samples. Since we did not use a multi-sample variant caller, the variants identified will consist of a lot of false positives. We used multiple different filters to exclude false positives and get a final set of analysis ready variants. Similar to the Freedman et al. study, we used different sets of filters for different analyses. The filtering criteria are detailed below.

### Genotype and variant quality

We marked all genotypes that had a phred scaled genotype quality of less than 30 as missing genotypes. Further, we also excluded variants that had a variant quality of less than 20.

### Depth of coverage

For each sample, we excluded sites that had an abnormally low or high coverage compared to the rest of the genome. We removed any sites that did not have at least 5 reads covering that site, since the uncertainty in the genotype call is high when it is based on a low number of reads. In addition, we also excluded any sites that had more than twice the average genome-wide coverage. The rationale behind discarding sites with high coverage is that these sites have a high coverage either due to mapping artifacts or the presence of homologs in the genome.

### Distance to neighboring variants

The presence of indels can cause the identification of false positive variants due to mapping artifacts around the indel. Similarly, a cluster of SNPs close to each other is an indicator of mapping artifacts. To filter out these false variants arising from mapping artifacts, we filter out variants that within 5 base pairs of another SNP or indel. Further, we used a lower quality score threshold for identification of neighboring variants, i.e., variants with a quality score between 10 and 20 were included in the pool of variants when filtering for distance to other variants. This ensures that we filter out any SNPs that are close to other variants, even if the neighboring variants do not pass our quality filters.

### Triallelic single nucleotide polymorphisms (SNPs)

We identified triallelic SNPs across the 23 samples. We used a genotype quality threshold of 30 and a variant quality threshold of 20 to identify such variants. We excluded all tri-allelic sites from downstream analyses.

### Minor allele frequency

We used multiple different minor allele frequency (maf) thresholds to prune our data depending on the analysis. For all the analyses performed in this study, we excluded singletons i.e. variants with only one copy of the rare allele in the sample. For PCA and admixture analyses,

we used two different maf thresholds (0.05–0.2) to obtain different datasets on which we repeated the analysis to explore the effect of low frequency variants on the analysis.

### Missingness

We excluded variant sites that had a high proportion of samples with missing genotype calls. These sites were excluded for analysis that required called genotypes. Missingness threshold dependent upon the analysis that are performed. For each of these analysis, the missingness threshold is indicated in the relevant section.

### Principal components analysis

As a first step to assessing the effect of reference genomes on the outcome of the data, we performed a principal components analysis (PCA) on the 23 samples. From the genotypes obtained for the samples after aligning to the dog reference genome, we identified variants in the combined set of samples. We excluded any variants which had a minor allele frequency less than 0.05. We also discarded variant sites with greater than 5% missingness. We used three additional filters to prune our dataset: triallelic SNPs, distance to nearest variant and depth of coverage. Using this filtered dataset, we performed PCA using the ngsCovar program available as part of the ngsTools suite of tools [52]. We repeated the analysis with different levels of missingness (10%, 20%) and minor allele frequency thresholds (0.05, 0.1 and 0.2) to check the robustness of our findings. We performed an identical analysis using the alignments obtained from mapping to the de novo wolf genome.

### Heterozygosity

To compute the heterozygosity per sample, we used the per sample genotype calls and excluded sites with a genotype quality less than 20 and a variant quality less than 30. We also discarded sites that were within 5 basepairs of another variant (indel or SNP) with a variant quality greater than 10. Using this filtered set of variants, we used *plink* [53] to compute the heterozygosity for each sample.

### Population size (PSMC)

We used pairwise sequentially markovian coalescent (PSMC) [41] approach to compute the population size history of the samples in our dataset. For each sample, we used the genotypes obtained from the alignments to the dog or the de novo wolf reference genome, to obtain a consensus fasta sequence across the entire genome. We filtered out sites with genotype quality lower than 20 or site quality lower than 30. Further, we used the depth filter to exclude sites with abnormally low or high number of reads covering it. Finally, we excluded variant

sites that were less than 5 basepairs away from another variant site with quality score greater than 10. We generated the PSMC input file using a window size of 100 base pairs. During this process, we marked all windows with more than 80 unknown/missing bases as missing. Using this input file, we ran *psmc* by dividing time into 64 bins, and used the pattern of “1\*6 + 58\*1” to estimate 59 independent population size parameters.

### Phylogenetic analysis

We constructed the phylogeny of all the samples using the variants identified by mapping the reads either to the de novo wolf reference assembly or the publicly available dog reference genome. For this analysis, we filtered out variants that were closer than 5 bp from another variant. In addition, we also excluded variants with quality scores lower than 30 and genotypes with quality scores less than 20.

We used the paleomix pipeline to obtain the phylogeny from these variants. As part of the *paleomix* pipeline, we used *RAxML* and *ExaML* [42, 43] to estimate the phylogeny of these samples. Since we do not have the annotations of genes for the de novo wolf reference genome, we used 5 megabases of sequences - 100 randomly selected regions, each 5 kb long - to estimate the phylogeny. We generated the consensus sequences for each of the samples using *samttools-0.1.19* [49, 51]. Since all the samples are mapped against the same genome and the indels are discarded, the multiple alignment of these regions were readily obtained by matching the genomic positions of the regions across samples. Using *RAxML* and *ExaML*, we generated the phylogeny of all the samples. We used 5 random starting points to generate 5 replicates from the data. We used 100 bootstrap runs to obtain the support for the nodes in the tree.

### Admixture

We performed an admixture analysis to identify structure and admixture in the samples. From the variants identified by combining the genotypes from all the samples, we excluded sites with qualities lower than 30, missingness greater than 25% and minor allele frequencies less than 5%. In addition, we also marked all genotypes with qualities lower than 20 as missing.

We used *ngsAdmix* [44] on this filtered dataset to obtain the admixture proportions in the samples. We ran *ngsAdmix* for different numbers of clusters, ranging from 2 to 6. For each value of the number of clusters, we ran the analysis 10 times and chose the run that gave us the best likelihood at convergence. Similar to our other analyses, we performed the same analysis for data obtained from mapping to the dog reference genome and the data from mapping to the de novo wolf reference genome.

## Additional file

**Additional file 1: Figure S1.** Phylogenies. The left panel shows the phylogenetic tree of all the samples, estimated from reads that are mapped to the boxer dog reference genome, while the right panel shows the phylogenetic tree estimated from the data after mapping reads to the *de novo* assembled wolf reference genome. **Figure S2.** Distribution of repeat elements. Total amount of bases in different repeat classes across the two reference genomes. **Figure S3.** Comparison of the divergence of the different repeat elements from their consensus sequence. The top panel shows the total number of bases against the divergence from the consensus sequence in each repeat family when using the *de novo* wolf reference genome for alignment. The bottom panel shows the same figures when using the boxer reference genome. **Figure S4.** Principal Components Analysis (PCA). Panels A and B show the first four principal components of the genotypes when using the *de novo* wolf reference assembly. For making these PCA plots, we used a missingness cutoff of 0.9 and a minor allele frequency cutoff of 0.2. Panels C and D show the first four principal components of the genotypes when using the boxer reference genome while using the same filtering thresholds. **Figure S5.** Picture of the skull of the Swedish wolf sample used for reference genome assembly. **Table S1.** Coverage and heterozygosity estimates. The coverage and heterozygosity are shown for each sample included in the study. For each animal, the higher estimate of coverage are bolded. (PDF 1652 kb)

## Abbreviations

LINE: Long interspersed nuclear elements; PCA: Principal components analysis; PSMC: Pairwise sequentially markovian coalescent; SINE: Short interspersed nuclear elements; SNP: Single nucleotide polymorphism

## Acknowledgements

The authors would like to acknowledge support from Science for Life Laboratory, the National Genomics Infrastructure (NGI) in Sweden and its genome assembly team, the Knut and Alice Wallenberg Foundation, UPPMAX and DTU-Computerome for providing assistance with massively parallel DNA sequencing and computational infrastructure. We further thank Guo-dong Wang and John Novembre and Bob Wayne for providing the raw sequencing data from their published genomic studies on dogs and wolves.

## Funding

We acknowledge the following for funding our research; Carlsbergfondet grant CF14-0995 and Marie Curie grant 655732-Wherewolf to SG, Danish National Research Foundation grant DNRF94, Lundbeckfonden grant R52-5062 and ERC Consolidator Grant (681396-Extinction Genomics) (to MTPG) and the Swedish Research Council (to LD). LFKK is supported by an FPI fellowship associated to BFU2014-55090-P (FEDER). TMB is supported by MINECO BFU2014-55090-P (FEDER), Fundacio Zoo Barcelona and Secretaria d'Universitat i Recerca del Departament d'Economia i Coneixement de la Generalitat de Catalunya. GL was supported by a European Research Council grant (ERC-2013-StG-337574-UN-DEAD) and Natural Environmental Research Council grants (NE/K005243/1 and NE/K003259/1).

## Availability of data and materials

The genome resequencing data used in the manuscript are available from the original publications and are indicated as such in the manuscript. The final assembly of the wolf reference genome can be downloaded from [https://sid.erda.dk/wsgi-bin/lis.py?share\\_id=f1ppDgUPQG](https://sid.erda.dk/wsgi-bin/lis.py?share_id=f1ppDgUPQG).

## Authors' contributions

SG and JS performed the bioinformatics and the comparative analyses. MS performed the lab work for the reference sample. TG, LO, AH and LD conceived the project. LK and TMB performed the copy number variation and repeat family analyses. JR and LD were involved the selection of the reference sample and sequencing to generate the assembly. BP and TSP provided computational infrastructure and expertise for the bioinformatics analyses. LO, AH and GL were involved in the discussion and interpretation of the results. SG, MS, LK, and TG wrote the majority of the manuscript. All authors were involved in the writing and editing of the manuscript.

## Ethics approval and consent to participate

Not applicable.

## Consent for publication

Not applicable.

## Competing interests

The authors declare that they have no competing interests.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## Author details

<sup>1</sup>Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, Øster Voldgade 5-7, 1350 Copenhagen, Denmark. <sup>2</sup>Natural History Museum, University of Oslo, N-0318 Oslo, Norway. <sup>3</sup>Institute of Evolutionary Biology (UPF-CSIC), PRBB, Dr. Aiguader 88, 08003 Barcelona, Spain. <sup>4</sup>CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Baldori i Reixac 4, 08028 Barcelona, Spain. <sup>5</sup>Department of Environmental Research and Monitoring, Swedish Museum of Natural History, Box 50007, 10405 Stockholm, Sweden. <sup>6</sup>Department of Bio and Health Informatics, Technical University of Denmark, 2800 Kongens Lyngby, Denmark. <sup>7</sup>Palaeogenomics & Bio-Archaeology Research Network, Research Laboratory for Archaeology and the History of Art, University of Oxford, OX1 3QY, Oxford, UK. <sup>8</sup>Catalan Institution of Research and Advanced Studies (ICREA), Passeig de Lluís Companys, 23, 08010 Barcelona, Spain. <sup>9</sup>Department of Bioinformatics and Genetics, Swedish Museum of Natural History, Box 50007, 10405 Stockholm, Sweden. <sup>10</sup>Trace and Environmental DNA Laboratory, Department of Environment and Agriculture, Curtin University, Perth, Western Australia, Australia. <sup>11</sup>NTNU University Museum, Norwegian University of Science and Technology, Trondheim, Norway.

Received: 12 December 2016 Accepted: 20 June 2017

Published online: 29 June 2017

## References

- Ekblom R, Galindo J. Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity*. 2011;107:1–15.
- Koboldt DC, Steinberg KM, Larson DE, Wilson RK, Mardis ER. The next-generation sequencing revolution and its impact on genomics. *Cell*. 2013;155:27–38.
- Allentoft ME, Sikora M, Sjögren K-G, Rasmussen S, Rasmussen M, Stenderup J, et al. Population genomics of bronze age Eurasia. *Nature*. 2015;522:167–72.
- Orlando L, Gilbert MTP, Willerslev E. Reconstructing ancient genomes and epigenomes. *Nat Publ Group*. 2015;16:395–408.
- Raghavan M, DeGiorgio M, Albrechtsen A, Moltke I, Skoglund P, Korneliusen TS, et al. The genetic prehistory of the new world Arctic. *Sci Am Assoc Adv Sci*. 2014;345:1255832.
- da Fonseca RR, Smith BD, Wales N, Cappellini E, Skoglund P, Fumagalli M, et al. The origin and evolution of maize in the Southwestern United States. *Nat Plants Nat Publishing Group*. 2015;1:14003.
- Skoglund P, Malmström H, Raghavan M, Stora J, Hall P, Willerslev E, et al. Origins and genetic legacy of Neolithic farmers and hunter-gatherers in Europe. *Sci Am Assoc Adv Sci*. 2012;336:466–9.
- Paper AL. Of AATAA at TE of T. Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature*. *Nat Publ Group*. 2015;513:409–13.
- Sousa V, Hey J. Understanding the origin of species with genome-scale data: modelling gene flow. *Nat Rev Genet*. 2013;14:404–14.
- Degner JF, Marioni JC, Pai AA, Pickrell JK, Nkadori E, Gilad Y, et al. Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics*. 2009;25:3207–12.
- Brandt DY, Aguiar VRC, Bitarello BD, Nunes K, Goudet J, Meyer D. Mapping bias overestimates reference allele frequencies at the HLA genes in the 1000 genomes project phase I data. *G3*. 2015;5:931–41.
- Huang L, Popic V, Batzoglu S. Short read alignment with populations of genomes. *Bioinformatics*. 2013;29:361–70.
- Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, et al. *De novo* assembly of human genomes with massively parallel short read sequencing. *Genome Res*. 2010;20:265–72.
- Li Y, Zheng H, Luo R, Wu H, Zhu H, Li R, et al. Structural variation in two human genomes mapped at single-nucleotide resolution by whole genome *de novo* assembly. *Nat Biotechnol*. 2011;29:723–30.

15. Rasmussen M, Guo X, Wang Y, Lohmueller KE, Rasmussen S, Albrechtsen A, et al. An aboriginal Australian genome reveals separate human dispersals into Asia. *Sci Am Assoc Adv Sci*. 2011;334:94–8.
16. Hufford MB, Xu X, van Heerwaarden J, Rvi TPA. J. A, Chia J-M, Cartwright RA, et al. comparative population genomics of maize domestication and improvement. *Nat Genet Nat Publishing Group*. 2012;44:808–11.
17. Kopaliani N, Shkarashvili M, Gurididze Z, Qurkhuli T, Tarkhnishvili D. Gene flow between wolf and shepherd dog populations in Georgia (Caucasus). *J Hered*. 2014;105:345–53.
18. Randi E, Lucchini V. Detecting rare introgression of domestic dog genes into wild wolf (*Canis lupus*) populations by Bayesian admixture analyses of microsatellite variation. *Conserv Genet Kluwer Acad Publishers*. 2002;3:29–43.
19. Orlando L, Ginolhac A, Zhang G, Froese D, Albrechtsen A, Stiller M, et al. Recalibrating Equus evolution using the genome sequence of an early middle Pleistocene horse. *Nat Nat Publishing Group*. 2013:1–8.
20. Schubert M, Jónsson H, Chang D, Der Sarkissian C, Ermini L, Ginolhac A, et al. Prehistoric genomes reveal the genetic foundation and cost of horse domestication. *Proceedings of the National Academy of Sciences. National Acad Sci*. 2014;111:E5661–9.
21. 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*. 2016;352:1228–31.
22. Ramos-Madriral J, Smith BD, Victor Moreno-Mayar J, Gopalakrishnan S, Ross-Ibarra J, Gilbert MTP, et al. Genome Sequence of a 5,310-Year-Old Maize Cob Provides Insights into the Early Stages of Maize Domestication. *Curr Biol*. 2016;26:3195–3201.
23. Park SDE, Magee DA, McGettigan PA, Teasdale MD, Edwards CJ, Lohan AJ, et al. Genome sequencing of the extinct Eurasian wild aurochs, *Bos primigenius*, illuminates the phylogeography and evolution of cattle. *Genome Biol BioMed Central Ltd*. 2015:1–15.
24. Xia Q, Guo Y, Zhang Z, Li D, Xuan Z, Li Z, et al. Complete resequencing of 40 genomes reveals domestication events and genes in silkworm (*Bombyx*). *Sci Am Assoc Adv Sci*. 2009;326:433–6.
25. Eriksson J, Larson G, Gunnarsson U, Bed'hom B, Tixier-Boichard M, Strömstedt L, et al. Identification of the yellow skin gene reveals a hybrid origin of the domestic chicken. *PLoS Genet*. 2008;4:e1000010.
26. Kanginakudru S, Metta M, Jakati RD, Nagaraju J. Genetic evidence from Indian red jungle fowl corroborates multiple domestication of modern day chicken. *BMC Evol Biol*. 2008;8:174.
27. Xiang H, Gao J, Yu B, Zhou H, Cai D, Zhang Y, et al. Early Holocene chicken domestication in northern China. *Proc Natl Acad Sci U S A*. 2014;111:17564–9.
28. Giuffra E, Kijas JM, Amarger V, Carlborg O, Jeon JT, Andersson L. The origin of the domestic pig: independent domestication and subsequent introgression. *Genetics*. 2000;154:1785–91.
29. Paper AL. Of AATAA at TE of T. Analyses of pig genomes provide insight into porcine demography and evolution. *Nat Nat Publishing Group*. 2012;491:393–8.
30. Wade CM, Giulotto E, Sigurdsson S, Zoli M, Gnerre S, Imsland F, et al. Genome sequence, comparative analysis, and population genetics of the domestic horse. *Science*. 2009;326:865–7.
31. Wang G-D, Zhai W, Yang H-C, Fan R-X, Cao X, Zhong L, et al. The genomics of selection in dogs and the parallel evolution between dogs and humans. *Nat Commun Nat Publishing Group*. 2013;4:1860–9.
32. Freedman AH, Gronau I, Schweizer RM, Ortega-Del Vecchyo D, Han E, Silva PM, et al. Genome sequencing highlights the dynamic early history of dogs. *PLoS Genet*. 2014;10:e1004016.
33. Zhang W, Fan Z, Han E, Hou R, Zhang L, Galaverni M, et al. Hypoxia adaptations in the Grey wolf (*Canis lupus chanco*) from Qinghai-Tibet plateau. *PLoS Genet*. 2014;10:e1004466.
34. Skoglund P, Ersmark E, Palkopoulou E, Dalén L. Ancient wolf genome reveals an early divergence of domestic dog ancestors and admixture into high-latitude breeds. *Curr Biol Elsevier Ltd*. 2015:1–6.
35. Lindblad-Toh K, Wade CM, Mikkelsen TS, Karlsson EK, Jaffe DB, Kamal M, et al. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nat Nat Publishing Group*. 2005;438:803–19.
36. 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. *Nat Nat Publishing Group*. 2010;464:898–902.
37. von Holdt BM, Pollinger JP, Earl DA, Knowles JC, Boyko AR, Parker H, et al. A genome-wide perspective on the evolutionary history of enigmatic wolf-like canids. *Genome Res*. 2011;21:1294–305.
38. Savolainen P, Zhang Y-P, Luo J, Lundeberg J, Leitner T. Genetic evidence for an east Asian origin of domestic dogs. *Sci Am Assoc Adv Sci*. 2002;298:1610–3.
39. 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. *Sci Am Assoc Adv Sci*. 2013;342:871–4.
40. Gnerre S, Maccallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, et al. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc Natl Acad Sci U S A*. 2011;108:1513–8.
41. Li H, Durbin R. Inference of human population history from individual whole-genome sequences. *Nature*. 2011;475:493–6.
42. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*. 2014;30:1312–3.
43. Kozlov AM, Aberer AJ, Stamatakis A. ExaML version 3: a tool for phylogenomic analyses on supercomputers. *Bioinformatics*. 2015;31:2577–9.
44. Skotte L, Korneliusen TS, Albrechtsen A. Estimating individual admixture proportions from next generation sequencing data. *Genetics*. 2013;195:693–702.
45. Svensson L, Åkesson M. Licensjakt varg 2011: DNA-analys och inventeringsdata [Internet]. 2011. Available from: [http://pub.epsilon.slu.se/13681/7/svensson\\_s\\_akesson\\_m\\_160928.pdf](http://pub.epsilon.slu.se/13681/7/svensson_s_akesson_m_160928.pdf).
46. Åkesson, M, Svensson, L. 2015. Sammanställning av släktträdet över den skandinaviska vargstammen fram till 2014. Available from: <http://www.viltskadecenter.se/images/stories/Publikationer/sammanstallning-slaktrad-skand-varg-2014-webb.pdf>.
47. Smit AFA, Hubley R, Green P. RepeatMasker Open-3.0. 1996. <http://www.repeatmasker.org/faq.html>.
48. Schubert M, Ermini L, Der Sarkissian C, Nsson HAKJO, Ginolhac AEL, Schaefer R, et al. Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX. *Nat Protoc Nat Publishing Group*. 2014;9:1056–82.
49. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. *Bioinformatics*. 2009;25:2078–9.
50. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernysky A, et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010;20:1297–303.
51. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetic parameter estimation from sequencing data. *Bioinformatics*. 2011;27:2987–93.
52. Fumagalli M, Vieira FG, Linderth T, Nielsen R. ngsTools: methods for population genetics analyses from next-generation sequencing data. *Bioinformatics*. 2014;30:1486–7.
53. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 2007;81:559–75.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at  
[www.biomedcentral.com/submit](http://www.biomedcentral.com/submit)

