Genomics using the Assembly of the Mink Genome

B. Guldbrandtsen, Z. Cai, G. Sahana, T.M. Villumsen, T. Asp, B. Thomsen, M.S. Lund
Dept. of Molecular Biology and Genetics, Research Center Foulum, Aarhus University, 8830 Tjele, Denmark
Bernt.Guldbrandtsen@mbg.au.dk (Corresponding author)

Summary

The American Mink's (*Neovison vison*) genome has recently been sequenced. This opens numerous avenues of research both for studying the basic genetics and physiology of the mink as well as genetic improvement in mink. Using genotyping-by-sequencing (GBS) generated marker data for 2,352 Danish farm mink runs of homozygosity (ROH) were detect in mink genomes. Detectable ROH made up on average 1.7% of the genome indicating the presence of at most a moderate level of genomic inbreeding. The fraction of genome regions found in ROH varied. Ten percent of the included regions were never found in ROH. Other parts were in ROH in up to about 1/6 of the individuals. Some haplotypes must have achieved high frequencies during the relatively short time since domestication of the American mink. However, there was no evidence of haplotypes nearing fixation. Selection is therefore likely to be ongoing. The ability to detect ROH in the mink genome also demonstrates the general reliability of the new mink genome assembly.

Keywords: american mink, run of homozygosity, genome, selection, genomic inbreeding

Introduction

The genome of the American mink (*Neovison vison*) has been sequenced and assembled. The current assembly (Cai *et al., 2017*) was created using deep paired-end shotgun sequencing. In addition mate-pair sequence data with longer inserts were generated. The combined coverage of the genome corresponds to 295 fold. Assembly was done using ALLPATHS-LG (Gnerre *et al.*, 2011). The current assembly is composed of 7,175 scaffolds. The N50 is 6.3 Mb. The length of the assembled genome is 2.4 Gb. The quality of the assembly was investigated by aligning short and medium length libraries to the assembly. Also previously published BAC libraries (Anistorei *et al.*, 2011; Benkel *et al.*, 2012) were aligned to the genome to that they were collinear with genome assembly. Of the assembled genome 29% is made up of repetitive sequences. Out of these long interspersed nuclear elements and short interspersed elements make up the largest proportions, 15 and 7%, respectively.

The closest relatives whose genomes have previously been assembled are the domestic dog (*Canis familiaris*; Lindblad-Toh *et al.*, 2005) and the ferret (*Mustela putorius furo*; Peng *et al.*, 2014). The dog genome assembly is of high quality, while the assembly of ferret genome is roughly comparable to the current assembly of the mink genome. The mink and ferret genome assemblies were aligned to the dog genome. For both, about 1.5 Gb of the total genome could be aligned to the dog genome assembly. A number of chromosomal

rearrangements (compared to the dog genome) were shared between the mink and ferret genome assemblies.

One way to examine the impact of drift and selection is by detecting extended homozygous regions, so-called Runs of Homozygosity (ROH) in the genome. ROH are useful for several purposes. They allow estimation of genomic levels of inbreeding. Concentrations of ROH in certain regions may be indicators of selection affecting the genomic compositions at these positions. In this study we combine the new mink genome assembly and GBS data to investigate some genomic consequences of the domestication of the American mink.

Materials and Methods

Marker information was generated by genotype-by-sequencing for 2,352 individual mink housed at the research facility in Foulum, Denmark. The animals are individuals originating from the research farm, individuals purchased from an outside source as well as individuals that are crossbred between the Foulum animals and animals from the outside source.

Genomic DNA was extracted and digested with two restriction enzymes, PstI (rare cutter) and MspI (frequent cutter). Resulting fragments were subjected single short read sequencing to an average depth of 10 fold per site. Resulting data were aligned using bwa (Li and Durbin, 2010). Variants were called using Genotype Analysis Toolkit's HaplotypeCaller (McKenna *et al.*, 2010).

Initial filtering was performed using Genome Analysis Toolkit's SelectVariants for bi-allelic sites with mapping quality > 30, at least 50% with a genotype quality of at least 30 and depth less than 100. However, this left a large number of markers with only one of the homozygotes observed. These were removed by filtering for allelic imbalance. Thus to produce the final dataset filtering was done using the pipeline *vcffilter -s -f "ABHet > 0.35 & ABHet < 0.65 & ABHom > 0.85" in.vcf > out.vcf.*

For ROH markers in scaffolds with at least 100 markers were extracted. Runs of homozygosity were detected using plink v. 1.9 (Chang *et al.*, 2015). Parameters were default values except that at least 5 SNP were required and a minimum length of 1 Mb.

Results

In the end, 28,336 markers were available for further analysis in scaffolds covering 2.26 Gb. Of these, 18,926 were in 73 scaffolds with at least 100 markers (range 100 – 839) in total spanning 767 Mb. All these scaffolds were at least 1 Mb long (range 1-40 Mb).

A total of 13,034 ROH were detected or 5.54 per individual (median 5, range 0-24). Given that the retained scaffolds cover 1/3 of the assembled genome this corresponds to a average of about 16 ROH per individual. The average length of genome in ROH was 13.4 Mb (median 12 Mb, range 0-97.6 Mb) corresponding to an average genomic inbreeding coefficient of 13.4/767=1.7% (median 1.5%, range 0-12.6%). On average each SNP was part of an ROH in 77.2 (median 75, range 0-395) out of 2,352 individuals corresponding to an

average genomic inbreeding rate of 3.3%. In total, 2,262 SNP were never observed to be part of an ROH.

The number of individuals with an ROH at a position was plotted against position in the scaffold. This was done for the scaffolds with the highest maximal number of ROH. The results are shown for scaffold 140 in Figure 1 and for scaffold 114 in Figure 2. Of the 73 scaffolds retained, 7 had at least one SNP which was in an ROH in at least 200 individuals (out of 2,352), while 43 in at least 100 individuals.

Discussion

ROH allow estimation of levels of genomic inbreeding. Compared to other species of farm animals, estimates of genomic inbreeding for the American mink population investigated are low. Estimates of inbreeding based on the fraction of markers in ROH and the fraction of the genome in ROH differ; presumably due to uneven marker densities. ROH detection is more efficient in regions with many SNP, so the proportion of SNP in ROH will tend to overestimate genomic inbreeding levels. Due to that many of the scaffolds only are a few Mb long, the requirement for 1 Mb to declare an ROH precludes many ROH spanning endpoints of scaffolds from being detected. The genomic inbreeding level based on the proportion of ROH, 1.7%, is therefore likely to be an underestimate of the true value.

Ten percent of the scaffolds contain at least one SNP which is in an ROH in about 8% of the population. This shows that certain haplotypes across the genome have become common. This must represent a combination of the effects of the domestication bottleneck, limited subsequent effective population size as well as natural and artificial selection. The mink has only been kept in farm systems for a limited time, in Denmark since the 1930ies (Hammershøj, 2004). Therefore selection due to domestication processes is likely still ongoing. This is supported by that no ROH peaks were anywhere close to fixation. Inbreeding tends to generate ROHs uniformly across the whole genome, so these peaks are not due to general inbreeding phenomena. About 10% of the SNPs in scaffolds containing more than 100 Mb, are never observed to be part of an ROH. Had the ROHs been an expression of general inbreeding in the population, the distribution of numbers in ROH would have been more uniform. Thus the concentration of ROH around a few regions suggests effects specific to certain chromosome regions – that is, selection.

The right shoulder in the ROH peak in scaffold 114 at 3.5 Mb coincides with the location of a genomic region associated with hair length identified through GWAS (*unpubl.*). The associated region is close the mink RADIL gene. In Zebrafish (*Danio rerio*) this gene is associated with neural crest development and pigmentation. No traits examined by GWAS so far showed association with markers on scaffold 140.

The ability to detect ROH in the mink genome using GBS-generated markers mapped to the new assembly demonstrates that the assembly of scaffolds of the mink genome is reliable. Assembly errors, e.g., matching unrelated pieces or over assembly of similar, but distinct genome segments would destroy the ability to detect ROH. The fact that ROH are detectable in substantial numbers in large parts of the genome demonstrates the reliability of the assembly.

Taking advantage of GBS data and the new genome assembly we have been able to detect the genomic impact of domestication, drift and selection in the genome of farmed American mink. Genomic diversity remains high, and selection processes for the most part seem not yet to have reached completion.

Acknowledgements

This research was supported by the Center for Genomic Selection in Animals and Plants (GenSAP) funded by Innovation Fund Denmark (grant 0603-00519B).

List of References

Anistoroaei, R., ten Hallers, B., Nefedov, M., Christensen, K. & de Jong, P., 2011. Construction of an American mink bacterial artificial chromosome (BAC) library and sequencing candidate genes important for the fur industry. BMC Genomics **12**, 354

Benkel, B. F. *et al*., 2012. A comparative, BAC end sequence enabled map of the genome of the American mink (*Neovison vison*). Genes & Genomics **34**, 83-91.

Cai, Z. *et al*. The draft reference genome of the American mink (*Neovison vison*) opens new opportunities of genomic research in mink, 2017. Scientific Reports **7**:14564.

Chang, C.C., Chow, C.C., Tellier, L.C.A.M., Vattikuti, S., Purcell, S.M., & Lee, J.J., 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience **4**: 7.

Comuzzie, A.G., Cole, S.A., Laston, S.L., Voruganti, V.S., Haack, K., *et al*., 2012. Novel Genetic Loci Identified for the Pathophysiology of Childhood Obesity in the Hispanic Population. PLoS ONE **7**(12): e51954. doi:10.1371/journal.pone.0051954

Gnerre, S. *et al*., 2011. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc Natl Acad Sci USA **108**: 1513-1518

Hammershøj, M., 2004. Population ecology of free-ranging American mink *Mustela vison* in Denmark, PhD thesis, National Environmental Research Institute, Denmark.

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

Lindblad-Toh, K. *et al.*, 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature **438**, 803-819.

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

Peng, X. *et al.*, 2014. The draft genome sequence of the ferret (*Mustela putorius furo*) facilitates study of human respiratory disease. Nat Biotechnol **32**, 1250-1255

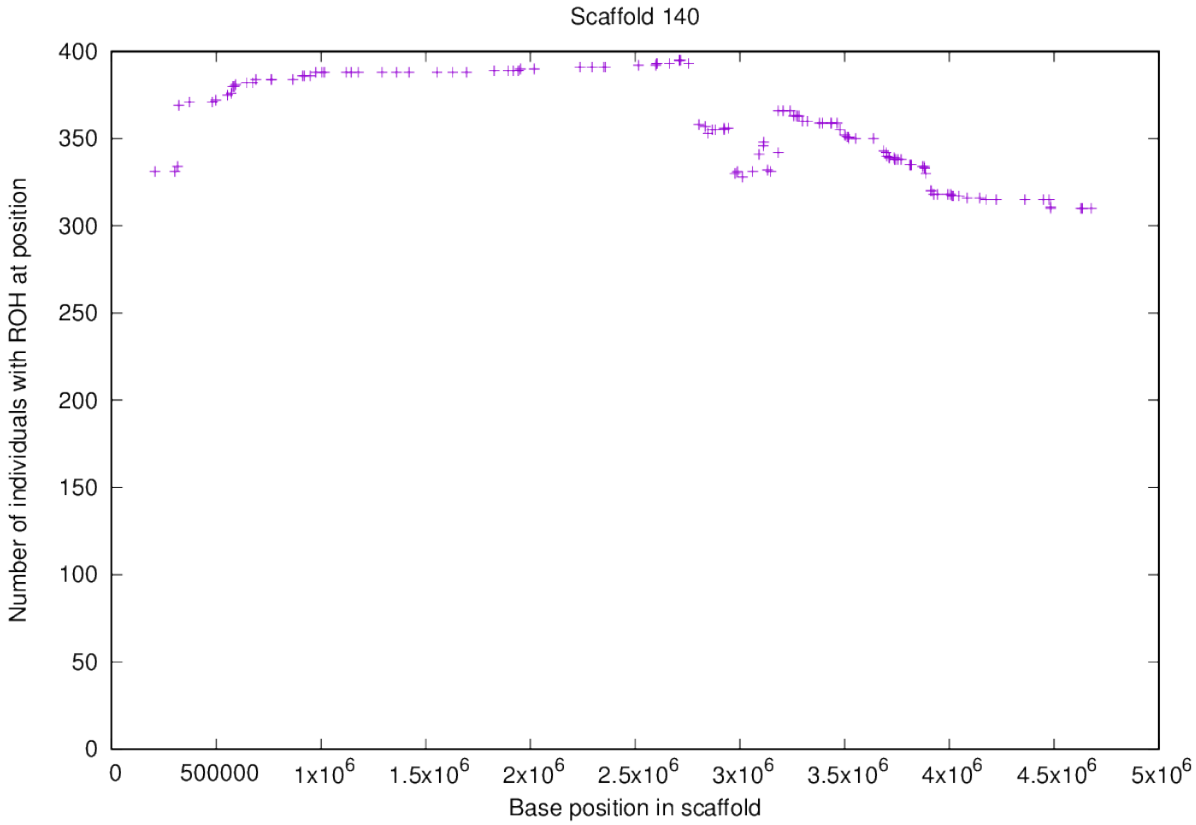

Figure 1: Number of individuals (out of 2,352) who had each position included in a detected ROH in scaffold 140.

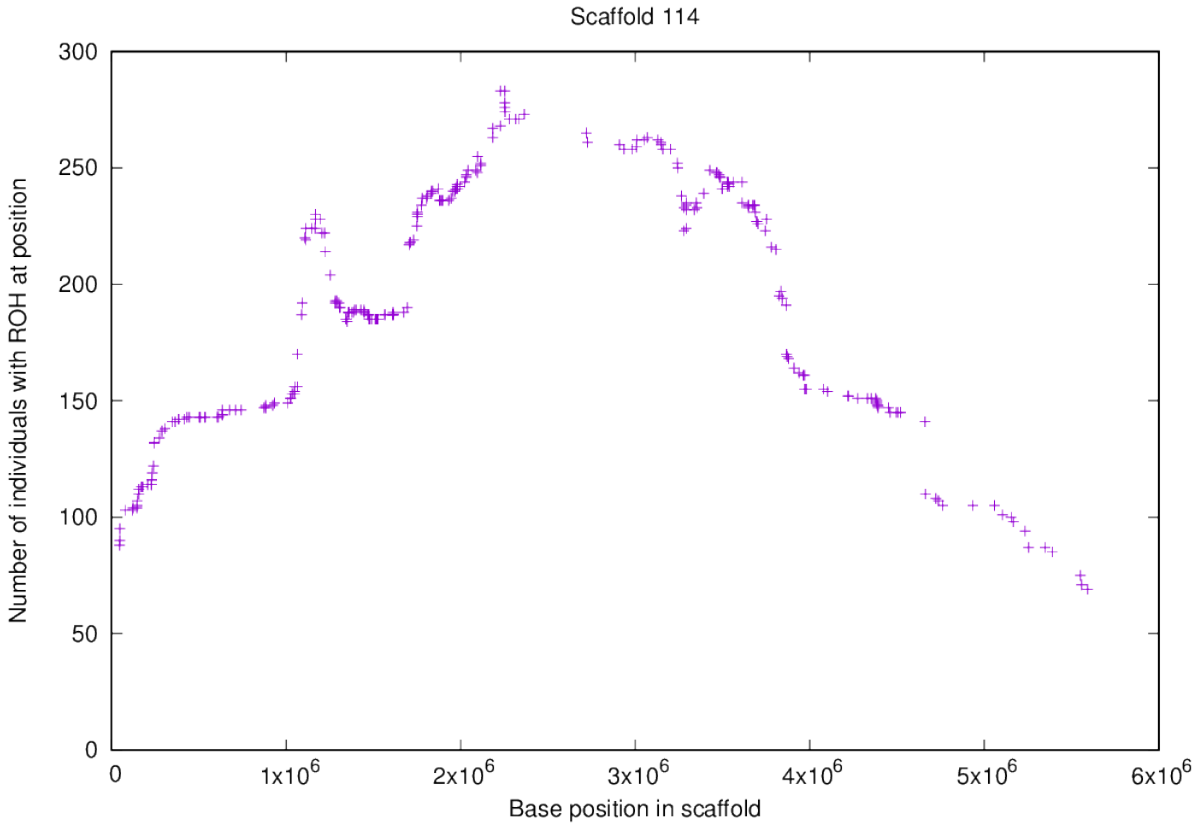

Figure 2: Number of individuals (out of 2,352) who had each position included in a detected ROH in scaffold 140.