Short Communication

Comprehensive characterization of horse genome variation by whole genome sequencing of 88 horses

Vidhya Jagannathan[1], Vinzenz Gerber[2], Stefan Rieder[3], Jens Tetens[4], Georg Thaller[5], Cord Drögemüller[1], Tosso Leeb[1]

[1] Institute of Genetics, Vetsuisse Faculty, University of Bern, 3001 Bern, Switzerland

[2] Department of Clinical Veterinary Medicine, Swiss Institute of Equine Medicine, Vetsuisse Faculty, University of Bern, and Agroscope, Länggassstrasse 124, 3012, Bern, Switzerland.

[3] Agroscope, Swiss National Stud Farm, 1580, Avenches, Switzerland

[4] Department of Animal Sciences, Functional Breeding Group, Georg-August University Göttingen, Burckhardtweg 2, 37077, Göttingen, Germany.

[5] Institute of Animal Breeding and Husbandry, Christian-Albrechts University Kiel, Hermann-Rodewald-Strasse 6, 24098, Kiel, Germany.

Running title: Whole genome sequence analysis of 88 horses

Address for correspondence

Vidhya Jagannathan
Institute of Genetics
Vetsuisse Faculty
University of Bern
Bremgartenstrasse 109a
3001 Bern
Switzerland

Phone: +41-31-6312325
E-mail: Vidhya.Jagannathan@vetsuisse.unibe.ch

Summary

Whole genome sequencing studies are vital to gain a thorough understanding of genomic variation. Here we summarize the results of a whole genome sequencing study comprising 88 horses and ponies from diverse breeds at 19.1x average coverage. The paired end reads were mapped to the current EquCab3.0 horse reference genome assembly and we identified approximately 23.5 million SNVs and 2.3 million short indel variants. Our dataset includes at least 7 million variants that were not previously reported. On average each individual horse genome carried ~5.7 million single nucleotide and 0.8 million small indel variants with respect to the reference genome assembly. The variants were functionally annotated. We provide two examples for potentially deleterious recessive alleles that were identified in heterozygous state in individual genome sequences. Appropriate management of such deleterious recessive alleles in horse breeding programs should help to improve fertility and reduce the prevalence of heritable diseases. This comprehensive dataset has been made publicly available and will represent a valuable resource for future horse genetic studies and supports the goal of accelerating the rates of genetic gain in domestic horse.

Keywords: genetic diversity, WGS, database, variant, Equus caballus, bioinformatics

Genetic variation comprising single nucleotide variants (SNVs), small insertion/deletion variants (indels) and structural variants (SVs) controls the heritable part of phenotypic diversity in animals and humans. SNVs are the most abundant genetic variation followed by indels and SVs. Protein changing variants include missense SNVs, but also nonsense SNVs, indels or SVs in protein coding regions, splice site variants and other types of variants. Protein changing variants often cause diseases or other phenotypic traits (Xue et al. 2012; Andersson, 2016). Whole genome sequencing data of various animal and plant species support the hypothesis that domestication commonly resulted in reduced overall genetic variation and an increased proportion of weakly deleterious variants segregating in populations of domesticated species (Makino et al. 2018). The advance of sequencing technologies has enabled large scale sequencing of individual genomes. Cataloging the features of the variants in these genomes helps to obtain a complete understanding of the vast pattern of genomic variation. The availability of large numbers of individual genomes and their genetic variants also accelerates the search for causal variants in medical genetics as these data allow to quickly rule out common, functionally neutral variants (Das et al. 2015; Broeckx et al. 2017).

Similar to other important domestic animal species, such as the dog, pig, cattle, sheep and chicken, the domestic horse is considered a viable large animal model for genetic research (Andersson 2016). In the era of mammalian genome projects, the horse was chosen as representative of the order perissodactyla and the first high quality draft genome assembly became available in 2005 (Wade et al. 2009). Since then, 50 k, 70 k, and 670 k SNV genotyping arrays for the domestic horses were developed by the Equine Genome Diversity Consortium (Schaefer et al. 2017). A new updated and greatly improved genome reference assembly, EquCab3.0, was released in April 2018 (Kalbfleisch et al. 2018).

In this study we utilized whole genome sequence data from 88 horses of genetically diverse breeds that had been sequenced with illumina paired-end sequencing technology to 19-fold coverage on average (range 5.3-43.5). We called and catalogued ~26 million sequence variants with respect to the new EquCab3.0 assembly (Table 1, Table S1). The variants have

been submitted to the European Variant Databases (EVA) and are available under the project accession number PRJEB28306. The detailed methodology is described in File S1.

Table 1: Summary information on the WGS data.

Breeds	Horses	Aligned bases (Gb)[1]	Coverage depth[1]	Genome covered[1]	SNVs (total)	Indels (total)
25	88	46.72	19.1x	99.1%	23,559,582	2,396,022

[1]Average per individual.

One high quality SNV was called every 105 bp on average. The transition to transversion ratio of SNVs was 1.97 and the heterozygote to homozygote ratio for SNVs was 2.16. On average, each horse genome contained 1,814,886 homozygous and 3,922,449 heterozygous SNVs with respect to the reference genome.

The indels in the 88 horses consisted of 1,282,573 short insertions (range 1–396 bp) and 1,591,971 short deletions (range 1–317 bp), with an average of 395,882 insertions and 426,380 deletions in an individual genome. The estimated heterozygote to homozygote ratio was 1.73 for short indels. On average, one indel occurred every 867 bp.

A total of 22,874,328 variants (87.5 %) were shared between at least two horses, and only 3,254,386 (12.5%) were private to individual horses (Figure S1A). The minimum and maximum numbers of private variants including SNVs and indels in a single genome were 4,499 and 151,107, respectively (Figure S1B).

The ~26 million variants were annotated with SnpEff (Cingolani et al. 2012). SnpEff annotates a single variant with more than one effect depending on the number of isoforms of a gene or overlapping genes at the genomic location of the variant. The SnpEff annotation contained more than 400,000 protein-changing effects (moderate and high impact variants) (Table 2).

Table 2: Summary statistics of identified variants and their predicted effects.

Category	SNVs	Indels	Total
Number of variants	23,559,582	2,396,022	25,955,604
Number of SnpEff variants[1]	23,761,421	2,874,544	26,635,965
Number of effects	61,877,764	7,961,562	69,839,326
Modifier	60,832,667 (98.3%)	7,914,882 (99.4%)	68,747,549 (98.3%)
Low	675,266 (1.1%)	14,563 (0.2%)	689,829 (1.1%)
Moderate	360,867 (0.5%)	10,232 (0.1%)	371,099 (0.5%)
High	8,964 (0.01%)	21,885 (0.3%)	30,849 (0.04%)

[1]SnpEff counts some variants multiple times, if they are located in overlapping genes.

The known *Equus caballus* SNVs and indels from dbSNP version 51, which still refers to EquCab2 and is now available from the European Variant Database, were remapped to EquCab3.0 using the remap.pl program of NCBI (https://www.ncbi.nlm.nih.gov/genome/tools/remap/). The VCF file downloaded from Ensembl consisted of 21,546,500 variants of which 21,325,592 (99%) could be successfully remapped to EquCab3.0. Among the mapped variants we found that 2,123,397 were marked REF_EDIT in the info tag of the VCF file meaning that the reference allele in the EquCab3.0 assembly had changed with respect to EquCab2. As these do no longer represent valid variants in the VCF file, we could only compare the remaining 19,202,195 dbSNP variants with the variants of our own dataset. A total of 16,407,561 (85%) of these remaining variants were contained in our dataset. On the other hand, our dataset contained 9,548,043 variants that are not contained in the dbSNP 51 dataset after remapping to EquCab3.0. Thus, even, if we assume that these 9.5 million variants also contain all of the ~2.1 million dbSNP REF_EDIT variants

that were lost during the re-mapping process to EquCab3.0, our dataset still contains at least 7 million previously unknown equine variants.

In August 2018, the Online Mendelian Inheritance in Animals (OMIA) database listed 83 records of likely causal variants for Mendelian traits in horse (http://omia.org). We identified 20 of them in our dataset and provide updated EquCab3.0 coordinates for them (Table S2).

Our set of annotated equine variants will facilitate the search for functional variants that cause phenotypic variation. To highlight just a few examples, among the variants with SnpEff high impact predictions we identified a nonsense variant in the *PALB2* gene encoding the partner and localizer of BRCA2 (XM_005598843.3:c.1402C>T; XP_005598900.2:p.(Arg468Ter)). The mutant allele containing a premature stop codon was present in heterozygous state in two of the sequenced Warmblood horses and we did not observe any homozygous mutant horses. It probably represents a severely deleterious recessive allele as PALB2 has been shown to be essential for mesoderm development. *Palb2*$^{-/-}$ knockout mice present with embryonic lethality and do not survive past E9.5 of development (Rantakari et al. 2010).

Another potentially deleterious recessive allele may be caused by a nonsense variant in the *PLEKHM1* gene encoding the pleckstrin homology domain-containing protein, family M, member 1 (XM_014739613.2:c.202C>T, p. XP_014595099.1:p.(Arg68Ter)). This variant was discovered in heterozygous state in another Warmblood horse. Human patients and rats that carry inactivating variants on both *PLEKHM1* alleles develop osteopetrosis due to diminished osteoclast function (OMIM #611497; van Wiesenbeeck et al. 2007).

Thus, similar to humans and many other domestic species, horses can also be expected to carry a small number of severely deleterious recessive alleles in their genomes (Das et al. 2015; Charlier et al. 2016). Such deleterious recessive alleles may rapidly increase in frequency due to the breeding practices involving the heavy use of only very few breeding animals (e.g. the popular sires). This can then result in unintentional carrier x carrier matings and homozygous embryos that are lost during gestation or develop into foals with severe hereditary diseases. Our variant catalog provides a means for the early identification of such

problematic alleles and facilitates their monitoring in breeding programs. This should help to reduce the prevalence of hereditary diseases and improve fertility.

The accuracy of algorithms for variant calling and assigning correct genotypes to sequenced individuals is far from perfect (Kim et al. 2017). We estimate, that there are tens of thousands of false-positive variant calls per genome in our dataset. Unfortunately, artifacts in whole-genome variant calling are widely known using current genome sequencing technology (Li 2014; Zook et al. 2014). In our dataset we can see that there is a small discrepancy between the SNV and indel heterozygote to homozygote ratio (2.16 and 1.73 respectively). This discrepancy could be due to a major portion of heterozygous SNVs being calling errors (Li 2014) or heterozygous indels were missed due to insufficient read coverage and were called erroneously as homozygous (Levy at al. 2007). Large SVs can have a profound functional effect, but they cannot be reliably called from illumina short read data with existing methods (Cretu Stancu et al. 2017; Huddleston et al. 2017). The quality of the horse genome annotation also still needs improvement. During our search for potentially deleterious recessive alleles we noticed that more than 50% of the “high impact” predictions by SnpEff were incorrect due to incorrect annotation of equine transcripts and their protein coding regions.

In conclusion, we provide a comprehensive annotated set of equine variants that have been mapped to the EquCab3.0 reference assembly. This set of variants comprises a large fraction of the existing genome diversity in horses. It includes many functionally important variants and should facilitate the search for causative variants underlying new rare Mendelian traits as it allows to identify common and most likely functionally neutral genetic variants.

Acknowledgements

The authors would like to thank Nathalie Besuchet Schmutz, Muriel Fragnière, and Sabrina Schenk for expert technical assistance. We also acknowledge the Next Generation Sequencing Platform of the University of Bern for performing the high-throughput sequencing experiments, and the Interfaculty Bioinformatics Unit of the University of Bern for providing high performance computing infrastructure.

References

Andersson L. (2016) Domestic animals as models for biomedical research. Upsala Journal of Medical Sciences 121, 1–11.

Charlier C., Li W., Harland C., Littlejohn M., Coppieters W., Creagh F., et al. (2016) NGS-based reverse genetic screen for common embryonic lethal mutations compromising fertility in livestock. Genome Research 26, 1333-41.

Cingolani P., Platts A., Wang le L., Coon M., Nguyen T., Wang L., Land S.J., Lu X. & Ruden D.M. (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) 6, 80-92.

Cretu Stancu M, van Roosmalen MJ, Renkens I, Nieboer MM, Middelkamp S, de Ligt J, et al. (2017). Mapping and phasing of structural variation in patient genomes using nanopore sequencing. Nat Commun 8, 1326.

Das A., Panitz F., Gregersen V.R., Bendixen C. & Holm L.E. (2015). Deep sequencing of Danish Holstein dairy cattle for variant detection and insight into potential loss-of-function variants in protein coding genes. BMC Genomics 16, 1043.

Huddleston J, Chaisson MJP, Steinberg KM, Warren W, Hoekzema K, Gordon D, et al. (2017) Discovery and genotyping of structural variation from long-read haploid genome sequence data. Genome Res 27, 677-685.

Kalbfleisch T.S., Rice E.S., DePriest Jr M.S., Walenz B.P., Hestand M.S., Vermeesch J.R., et al. (2018) EquCab3, an updated reference genome for the domestic horse. bioRxiv, doi:10.1101/306928

Kim B.Y., Park J.H., Jo H.Y., Koo S.K., Park M.H. (2017) Optimized detection of insertions/deletions (INDELs) in whole-exome sequencing data. PLoS One 12, e0182272.

Levy S., Sutton G., Ng PC., Feuk L., Halpern AL., Walenz BP., Axelrod N., Huang J., et al.(2007) The diploid genome sequence of an individual human. PLoS Biol.5(10): e254.

Li H. (2014) Toward better understanding of artifacts in variant calling from high-coverage samples. Bioinformatics. 30, 2843-51.

Makino T., Rubin C.-J., Carneiro M., Axelsson E., Andersson L. & Webster M.T. (2018) Elevated proportions of deleterious genetic variation in domestic animals and plants. Genome Biology and Evolution 10, 276–90.

Rantakari P, Nikkilä J, Jokela H, Ola R, Pylkäs K, Lagerbohm H, et al. (2010) Inactivation of Palb2 gene leads to mesoderm differentiation defect and early embryonic lethality in mice. Human Molecular Genetics 19, 3021-9.

Schaefer R.J., Schubert M., Bailey E., Bannasch D.L., Barrey E., Bar-Gal G.K., et al. (2017) Developing a 670k genotyping array to tag ~2M SNPs across 24 horse breeds. BMC Genomics 18, 565.

van Wesenbeeck L., Odgren P.R., Coxon F.P., Frattini A., Moens P., Perdu B., et al. (2007) Involvement of PLEKHM1 in osteoclastic vesicular transport and osteopetrosis in incisors absent rats and humans. Journal of Clinical Investigations 117, 919-30.

Wade C.M., Giulotto E., Sigurdsson S., Zoli M., Gnerre S., Imsland F., et al. (2009) Genome sequence, comparative analysis, and population genetics of the domestic horse. Science 326, 865–7.

Xue Y., Chen Y., Ayub Q., Huang N., Ball E.V., Mort M., et al. (2012) Deleterious- and disease-allele prevalence in healthy individuals: Insights from current predictions, mutation databases, and population-scale resequencing. American Journal of Human Genetics 91, 1022–32.

Zook JM, Chapman B, Wang J, Mittelman D, Hofmann O, Hide W, Salit M. (2014) Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. Nat Biotechnol 32, 246-51.

Supplementary Material

Figure S1. Variant sharing between horses and summary of private variants for each horse.

File S1. Methodology for mapping, variant calling and effect prediction of whole genome sequence data.

Table S1. Detailed descriptive statistics and accessions of 88 horse genomes.

Table S2. OMIA variants in the dataset.