bioRxiv preprint doi: https://doi.org/10.1101/112979; this version posted March 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Development of a high-density, 2M SNP genotyping array and 670k SNP imputation array for the domestic horse

Robert J. Schaefer[1], Mikkel Schubert[2], Ernest Bailey[3], Danika L. Bannasch[4], Eric Barrey[5], Gila Kahila Bar-Gal[6], Gottfried Brem[7], Samantha A. Brooks[8], Ottmar Distl[9], Ruedi Fries[10], Carrie J. Finno[4], Vinzenz Gerber[11], Bianca Haase[12], Vidhya Jagannathan[13], Ted Kalbfleisch[14], Tosso Leeb[13], Gabriella Lindgren[15], Maria Susana Lopes[16], Nuria Mach[17], Artur da Câmara Machado[16], James N. MacLeod[3], Annette McCoy[18], Julia Metzger[9], Cecilia Penedo[19], Sagi Polani[6], Stefan Rieder[20], Imke Tammen[12], Jens Tetens[21,22] , Georg Thaller[21], Andrea Verini-Supplizi[23], Claire M. Wade[12], Barbara Wallner[7], Ludovic Orlando[2,24] , James R. Mickelson[25], Molly E. McCue[1*]

1. Department of Veterinary Population Medicine, College of Veterinary Medicine, University of Minnesota, St. Paul, Minnesota, USA
2. Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, Copenhagen, Denmark
3. Maxwell H. Gluck Equine Research Center, Department of Veterinary Science, University of Kentucky, Lexington, Kentucky, USA
4. School of Veterinary Medicine, University of California-Davis, Davis, CA, 95616
5. Unité de Génétique Animale et Biologie Intégrative- UMR1313, INRA, Université Paris-Saclay, AgroParisTech, 78350 Jouy-en-Josas, France
6. The Robert H. Smith Faculty of Agriculture, Food and Environment, The Koret School of Veterinary Medicine, The Hebrew University, Rehovot 76100, Israel
7. Institute of Animal Breeding and Genetics, Department of Biomedical Sciences, University of Veterinary Medicine Vienna, Vienna, Austria
8. Department of Animal Science, University of Florida, Gainesville, FL, USA
9. Institute for Animal Breeding and Genetics, University of Veterinary Medicine, Hannover, Germany
10. Lehrstuhl für Tierzucht der Technischen Universität München, Liesel-Beckmann-Strasse 1, 85354 Freising, Germany
11. Swiss Institute of Equine Medicine, Department of Clinical Veterinary Medicine, Vetsuisse Faculty, University of Bern, and Agroscope, Länggassstrasse 124, 3001 Bern, Switzerland
12. School of Life and Environmental Sciences, Faculty of Veterinary Science, University of Sydney, Regimental Drive, B19-301 RMC Gunn, Sydney, NSW 2006, Australia
13. Institute of Genetics, University of Bern, Bern 3001, Switzerland
14. Department of Biochemistry and Molecular Biology, School of Medicine, University of Louisville, Louisville KY 40202
15. Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Uppsala, Sweden
16. Biotechnology Centre of Azores, University of Azores, Angra do heroísmo, Portugal
17. Animal Genetics and Integrative Biology unit (GABI), INRA, AgroParis Tech, University of Paris-Saclay, 78350 Jouy-en-Josas, France
18. Department of Veterinary Clinical Medicine, College of Veterinary Medicine, University of Illinois at Urbana-Champaign, IL, 61802
19. Veterinary Genetics Laboratory, University of California Davis, Davis, California, United States of America
20. Agroscope, Swiss National Stud Farm, Avenches 1580, Switzerland
21. Institute of Animal Breeding and Husbandry, Christian-Albrechts-University Kiel, Hermann-Rodewald-Strasse 6, 24098, Kiel, Germany
22. Department of Animal Sciences, Functional Breeding Group, Georg-August University Göttingen. Burckhardtweg 2, 37077 Göttingen, Germany
23. Department of Veterinary Medicine - Sport Horse Research Centre, University of Perugia, Perugia, Italy
24. Laboratoire d'Anthropobiologie Moléculaire et d'Imagerie de Synthèse, CNRS UMR 5288, Université de Toulouse, Université Paul Sabatier, Toulouse, France 31000
25. Department of Veterinary and Biomedical Sciences, College of Veterinary Medicine, University of Minnesota, St. Paul, Minnesota, USA

* Corresponding Author

bioRxiv preprint doi: https://doi.org/10.1101/112979; this version posted March 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Abstract

Background

To date, genome-scale analyses in the domestic horse have been limited by suboptimal single nucleotide polymorphism (SNP) density and uneven genomic coverage of the current SNP genotyping arrays. The recent availability of whole genome sequences has created the opportunity to develop a next generation, high-density equine SNP array.

Results

Using whole genome sequence from 153 individuals representing 24 distinct breeds collated by the equine genomics community, we cataloged over 23 million *de novo* discovered genetic variants. Leveraging genotype data from individuals with both whole genome sequence, and genotypes from lower-density, legacy SNP arrays, a subset of ~5 million high-quality, high-density array candidate SNPs were selected based on breed representation and uniform spacing across the genome. Considering probe design recommendations from a commercial vendor (Affymetrix) a set of ~2 million SNPs were selected for a next-generation high-density SNP chip (MNEc2M). Genotype data were generated using the MNEc2M array from a cohort of 332 horses from 20 breeds and a lower-density array, consisting of ~670 thousand SNPs (MNEc670k), was designed for genotype imputation.

Conclusions

Here, we document the steps taken to design both the MNEc2M and MNEc670k arrays, report genomic and technical properties of these genotyping platforms, and demonstrate the imputation capabilities of these tools for the domestic horse.

Keywords:

Equine Genomics; Whole genome sequence; SNP-tagging; SNP chip; Variant recalibration; SNP discovery; SNP informativeness; SNP Validation; Linkage disequilibrium;

bioRxiv preprint doi: https://doi.org/10.1101/112979; this version posted March 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Background

Soon after the horse reference genome from Twilight, a female Thoroughbred, was sequenced using Sanger technology [1], a genotyping array (Illumina EquineSNP50 BeadChip) was developed to enable whole genome mapping using ~50k single nucleotide polymorphism (SNP) markers from low-coverage Sanger sequence of 7 horses representing 7 different breeds [2]. Shortly thereafter, a slightly higher density array (Illumina Equine SNP70 BeadChip) with ~65k informative SNP markers was developed. These widely used, now legacy, SNP arrays (with a combined ~74k unique probes, [3]), have successfully enabled genetic studies examining domestication and selection [4,5], disease and performance trait mapping [6–15], and population structure and dynamics [2,16,17]. However, extensive population structure and the low extent of linkage disequilibrium (LD) existing in many horse breeds severely limits conventional mapping approaches with the relatively low SNP density on the current SNP array [17]. Since the initial Sanger shotgun sequencing of Twilight, whole genome sequence (WGS) has been generated for hundreds of horses [15,18–25], prompting the development of a new higher-density genotyping array for the horse.

Here we describe the steps taken, including careful and extensive variant filtering, to create both a 2 million (2M) SNP array and a 670 thousand (670k) SNP array from over 23 million variants discovered from whole genome sequence of 153 horses representing 24 breeds (See Table S1). Genotypes from the 2 million probe genotyping array (MNEc2M) in a cohort of 332 horses, from 20 actively-researched or economically-important breeds that represent known genetic diversity in the domestic horse [17], were used to select SNPs for inclusion on the commercially-available 670k probe array (MNEc670k). The MNEc670k array was designed for accurate genotype imputation up to the higher density, 2M SNP set present on the MNEc2M array. We report summary statistics, broken down by breed, for both arrays as well as preliminary results on genotype imputation performance from the MNEc670k array to the SNP density on the MNEc2M array.

bioRxiv preprint doi: https://doi.org/10.1101/112979; this version posted March 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Results

Variant discovery from whole genome sequence

Whole genome, 100 base pair (bp), paired-end Illumina HiSeq reads were generated for 153 horses (including Twilight), representing 24 distinct breeds, at a depth between 1.7X and 64X, with a median depth of 13X (Table S1). Read mapping was performed using the PALEOMIX genome mapping protocol [26] to efficiently process samples in parallel and to assess individual sample quality. Each sample was mapped to the EquCab2.0 reference genome, which had been extended with an additional 7,850 *de novo* assembled scaffolds (See Methods), to produce a total of nearly 48 billion unique reads being aligned to the nuclear genome (See Table S1). Variants were identified by extending the PALEOMIX framework to identify SNPs using two variant callers (see Methods). To maximize efficiency of variant calling in individual breeds, and to minimize bias due to variable sequencing depth of coverage, individuals were broken up into 16 variant calling groups (see Table S1, Variant Calling Group columns) by estimated depth of coverage and breed. Variants were called using permissive parameters in both the GATK UnifiedGenotyper [27] as well as SAMtools 'mpileup' utilities [28]. Approximately 23 million potential SNPs were called by both GATK and SAMtools. These 23 million SNPs were kept for further analysis and validation (See Table 1; Supp. File 1).

Precision of GATK QUAL scores for variants identified by both callers

We evaluated the performance of GATK QUAL scores by comparing genotypes generated from WGS to genotypes generated from the legacy Equine 54K SNP chip in 23 horses at two different sequencing depths. Variants detected on chromosome 1 (ECA1) by both WGS and 54k SNP chip were ranked by decreasing QUAL score. The proportion of genotype calls that agreed between 54K and WGS (i.e., precision) was compared based on ranked QUAL score (i.e., recall) [29]. For each ranked variant we evaluated the precision at that point, which is the cumulative proportion of genotypes that agree between the two genotyping methods (Fig. 1). We note that this approach evaluates the concordance between the two

bioRxiv preprint doi: https://doi.org/10.1101/112979; this version posted March 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

genotyping technologies and is thus unable to assess the accuracies of the underlying genotypes themselves. However, this approach does take into account the imbalance of false positives/negatives between genotypes called by WGS and by the 54K SNP chip [30].

Considering 100 percent of variants detected by both technologies, genotypes had an overall precision of ~99% for 12X-called variants and 94-95% for 6X-called variants (See Fig. 1, x-axis). This is consistent with results from a similar study that compared *de novo* variant genotypes to array based genotypes in the Franches-Montagnes horse breed [25]. However, many SNPs with very high QUAL scores (e.g., SNPs within the top 10 percent; Fig. 1) had disagreeing genotype calls between WGS and 54K SNP chip. Additionally, when considering between 80-100% of variants ranked by QUAL score (Fig. 1), in both 6X MOR and STD comparisons, the proportion of matching genotypes increases, indicating that there are many variants with low QUAL scores and high precision. These results indicated that QUAL scores alone did not adequately rank variants, and that additional metrics were necessary to improve the reliability of SNPs ultimately chosen for the higher density genotyping arrays.

Identifying gold standard reference set for variant recalibration

In addition to QUAL type quality scores, GATK outputs additional metrics such as depth, quality of depth, Fisher strand position, map quality rank sum, and read position rank sum for each variant based on read statistics (See Methods for details of these metrics). These values were used to train linear mixed models, using the GATK VariantRecalibrator [31], to assign a composite quality score (VQSLOD) to help detect type I and II errors.

Training these models required a "gold standard" reference set of known genotypes across multiple individuals. Lacking this resource in the domestic horse, we defined three high confidence, putative "gold standard" datasets with which to train the VariantRecalibrator: 1) SNPs on the legacy SNP50 chip; 2) WGS variants which were seen in four or more (4+) calling groups (Table S1); and 3) WGS variants that were in the top one percent of QUAL scores. Models were trained on features from "gold standard" variants present

bioRxiv preprint doi: https://doi.org/10.1101/112979; this version posted March 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

on chromosomes 2-32 and used to calculate VQSLOD scores for SNPs on chromosome 1. To assess the performance of VSQLOD scores generated by each training set, scores from each group were compared to each other in addition to QUAL scores using horses genotyped on the 54K SNP Chip.

SNPs on chromosome 1 (ECA1) were ranked either by decreasing QUAL or by decreasing VQSLOD score generated from each of the three gold standard training groups. Fig. 2 shows the proportion of matching genotypes between WGS and SNP50 platforms for each of the four groups. With the exception of 6X Morgans, SNPs with high VQSLOD scores agreed more often with genotypes called on the 54K SNP chip compared to SNPs with high QUAL scores alone. For example, the top 10% highest scoring VQSLOD scores had a higher concordance rate than QUAL scores in all cases except for MOR6X, where only two of the recalibrated scores marginally outperformed QUAL scores (See Fig. 2; see discussion). Additionally, VQSLOD variants did not drop below the overall discordance rate, showing an overall better ranking than QUAL scores alone (red curve, Fig. 2).

While an exact differentiation between the gold-standard training groups remains unclear (See Discussion), recalibrated VQSLOD scores were better able to differentiate high scoring, false positives than QUAL scores alone, making them a more informative metric for selecting the final SNPs for use on the high-density commercial array. VQSLOD scores were calculated for all 23 million variants, in all remaining horses, using WGS variants called in the 4+ breed calling groups (described above) as the training set (candidate 23M SNP set). Scored variants were then filtered in several steps (described below) to select the subset of SNPs to be included on the new higher density arrays.

Preliminary 5M SNP selection for genotyping probe design

We analyzed the 23M candidate SNP set with the goal of generating a target set of ~5 million high-confidence SNPs that would be compatible with array design. Following criteria provided by Affymetrix, 780 SNPs were filtered because they fell within repetitive regions; 9,342,733 SNPs were filtered because they were within 20 base pairs (bp) of another variant; and 4,858,273 SNPs were filtered for having a minor

bioRxiv preprint doi: https://doi.org/10.1101/112979; this version posted March 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

allele frequency below 1% in the 153 WGS horses. After filtering, approximately 10 million SNPs remained for further design selection (hereafter referred to as the 10M SNP Set, Table 1). Of these ~10 million array-compatible candidates, 48,485 SNPs (65%) overlapped with the sites available from the ~74k legacy SNP sets (although all legacy SNP sites were included in MNEc2M/670k array design). Supp. Fig. 1 shows the alternative allele frequency distribution of the 23M *de novo* and 10 million design candidate SNPs (10M SNP set). The mean allele frequency is below 0.1, though there is a long tail of SNPs with a high alternative allele frequency. Interestingly, there are 445,421 SNPs (23M set) and 46,455 SNPs (10M set) where the alternative allele frequency is 1.0. This is likely due to errors in the EquCab 2.0 reference sequence (Sanger reads from Twilight) as Twilight's whole genome Illumina sequence data (collectively at more than 25X), and whole genome sequences of all other horses, contained the alternate allele. These SNPs were excluded from array design.

To ensure backwards compatibility, we started with the ~74k SNPs that were present on the two legacy (54K/65K) arrays. To fill the remaining target of ~5 million probe design candidates, we added SNPs from the 10M SNP set based on even genome distribution, informativeness among the variant calling groups, and linkage disequilibrium (LD). LD was calculated for all pairs of SNPs within 10kb of each other throughout the genome. SNPs that were in high LD ($r^2 > 0.90$) were filtered based on whether or not they were present in draft or pony variant calling groups to control for the fewer number of samples used in variant discovery. Previous SNP designs [2] show an underrepresentation of informative SNPs from these groups, which leads to poor mapping resolution in draft and pony breed groups (See Table S2). SNPs discovered in both draft and ponies were prioritized over those discovered in one group and over those absent in both. If SNPs were discovered in an equal number of priority groups, VQSLOD scores were used to break ties.

After applying these SNP candidate criteria, 5,443,950 SNPs (5M set, see Table 1) were kept, of which 2,199,467 SNPs occurred in pony calling groups and 2,782,917 SNPs occurred in draft calling groups. A

bioRxiv preprint doi: https://doi.org/10.1101/112979; this version posted March 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

total of 1,695,347 SNPs occurred in both ponies and drafts. Flanking sequences (35 bp upstream and downstream) for these 5M filtered SNPs were submitted to Affymetrix for SNP probe design analysis.

SNP selection for the high density MNEc2M SNP array

SNP conversion recommendations for the 5M SNPs were provided from Affymetrix in both forward and reverse strand directions. SNPs were assigned to one of four groups based on decreasing probability of successful probe design: 'recommended', 'neutral', 'not recommended', and 'not possible'. To achieve an even distribution of ~2 million SNPs, the equine reference genome (~2.7 Gb) was divided into approximately 54,000 50 kb windows and a target of 37 SNPs per window was established. SNPs within each window were chosen for inclusion in the MNEc2M SNP set using a greedy algorithm. Briefly, the ~74,000 SNPs on previous generation arrays, as well as SNPs within the equine major histocompatibility complex (MHC) region (ECA20:28.7-33.6Mb), were given VIP status and were automatically included in both forward and reverse strand directions. SNPs were added to windows until the target of 37 SNPs was met. If a window had more than 37 candidates, SNPs were selected based on by Affymetrix recommendation group (See Methods for details). In total, given the above criteria, 2,001,826 high quality SNPs (2M SNP set) were chosen and submitted for probe design to comprise the MNEc2M genotyping array (Supp. File 2).

SNP selection for the MNEc670k array

A cohort of 347 horses were genotyped on the MNEc2M high-density array using DNA isolated from blood or hair roots (see Methods). The ~2M SNP genotypes were split across three different physical arrays. Genotypes were called using Affymetrix Power Tools and sample quality control was assessed according to Affymetrix best practices (See Methods). For each sample, the three physical arrays were assessed separately for quality control. In total, 320 samples passed in all three arrays while 27 samples had one or more arrays that did not pass quality control. Of the 27 samples that failed, 7 samples had two passing arrays, 5 samples had one passing array, and 15 DNA samples failed to meet genotyping quality control

bioRxiv preprint doi: https://doi.org/10.1101/112979; this version posted March 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

metrics on all three physical arrays. Failed arrays were removed from the analysis, however if a sample had at least one array, genotypes from those arrays were retained. In total, viable genotypes were produced for 332 horses.

Of the failed arrays, 25 had DNA isolated from hair roots and 2 had DNA isolated from blood. Hair root DNA had a lower average DNA concentration (Pico Green) when re-hydrated (2.7 ± 2.9 ng/μL) than did DNA samples isolated from blood (43.1 ± 55.5 ng/μL). To determine if sample origin (blood versus hair roots) or DNA concentration, or both, were associated with failure to pass genotyping quality control metrics, a logistic regression for sample success on DNA concentration and blood/hair status was performed. (Supp. Figure 5). Both DNA source ($p \leq 4.05e\text{-}04$) and DNA concentration ($p \leq 2.77e\text{-}10$) significantly influenced the probability of samples to produce quality genotypes (see Discussion).

After quality control steps for samples, genotypes were called for all 2,001,826 SNPs. Genotypes were assessed for clustering quality using the *Metrics.R* script provided by Affymetrix (see Methods). In total, 92.2% of SNPs on the MNEc2M array passed quality control producing a set of 1,846,988 high-quality SNPs (1.8M; see Table 1) genotyped on the 332 horses remaining in the analysis (see Methods).

SNPs exclusive to a single WGS variant calling group were validated (polymorphic), on average, at a rate of 80%. However, validation rates were much higher in SNPs discovered in multiple calling groups (>96%) (See Table S3). Genotypes produced from the 332 horses with passing genotypes on the MNEc2M array ("SNP" genotypes) were combined with genotypes discovered from the 153 whole genome sequence horses ("WGS" genotypes) to create a dataset containing 1.8M genotypes for 485 horses ("WGS+SNP" genotypes). These variants were analyzed to select a subset of SNPs for inclusion on the MNEc670k array, with the intent that this array would be designed for genotype imputation to the full 2M SNP set.

To maximize the information content of the MNEc670k array SNPs, multi-marker r^2 statistics were calculated on the 1.8M high-quality candidate SNPs to identify 'tagging SNPs' that allowed for efficient reconstruction of genomic haplotypes in the 485 horses both within and across breeds. Tagging SNPs that

bioRxiv preprint doi: https://doi.org/10.1101/112979; this version posted March 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

reconstructed haplotypes across all 485 horses (inter-breed tag SNPs) were identified using the software FastTagger [32]. In total, 355,903 tagging SNPs were needed to reconstruct haplotypes present across all the horses in the cohort with a minor allele frequency (MAF) > 0.01 and tag-SNP $r^2 > 0.99$. These SNPs were included on the MNEc670k imputation array (see Table S4; Inclusion criteria: Inter).

Haplotype tagging SNPs were also examined at the breed-specific level. Horses were split into 15 *tagging breed groups* based on the minimal sample size necessary to perform SNP tagging (see Table S5). Tagging SNPs were identified separately in each tagging breed group using FastTagger, then a subset of population specific tag SNPs were identified using the software Multi-Pop-TagSelect [33]. Combined, a total of 1,754,075 SNPs were needed to reconstruct fine-level, breed-specific haplotypes in all of these breed groups (MAF > 0.10 and tag SNP $r^2 > 0.90$; See Supp. File 3). Breeds varied in the number of tag SNPs needed to reconstruct haplotypes. Table 2 shows the number of tagging SNPs required to reconstruct haplotypes in each of the tagging breed groups. The Ponies, Draft and Quarter Horses (tagging breed groups) required the most tagging SNPs, each requiring over 350,000 SNPs to reconstruct breed-specific haplotypes, while Thoroughbreds, Icelandic and Lusitano tagging breed groups each required less than 150,000 tag SNPs, to reconstruct haplotypes.

Tagging SNPs that were informative in 5 or more tagging breed groups were included on the MNEc670k array (n=206,822; see Table S4; inclusion criteria: Intra). An additional 13,993 SNPs that tagged haplotypes in the breeds requiring the largest number of tag SNPs were also included in the array (inclusion criteria: Diverse). Additionally, 7,394 SNPs were included due to their location within the equine MHC region (inclusion criteria: MHC), 16,398 SNPs were included to increase SNP density in 12,104 (24.3%) 50 kb genomic windows to at least 8 SNPs (inclusion criteria: Density), and 70,295 SNPs were retained for backwards compatibility with legacy arrays (inclusion group: VIP). Collectively, 670,805 SNPs were included on the MNEc670k commercial array.

bioRxiv preprint doi: https://doi.org/10.1101/112979; this version posted March 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Imputation accuracy from the MNEc670k SNP set to the MNEc2M SNP

Genotype imputation accuracy from the MNEc670k array to the MNEc2M array was quantified using the 485 horse reference population. For each of the tagging breed groups a subset of 1/3 random individuals were masked down to the MNEc670k SNP set. Genotypes were then imputed to the MNEc2M SNP using the reference population after removing the individuals being imputed (See Methods). Imputation accuracy was measured as total genotype concordance across imputed individuals (correctly inferred genotypes/total number of imputed genotypes; see Methods for details).

Genotype imputation accuracy from the MNEc670k SNP set to the MNEc2M SNP set ranged between 96.6 and 99.4% in the 15 breeds tested (see Table 3). Tagging breed groups with over 99% mean imputation accuracy included Arabians, Belgians, Lusitano, Maremanno, Pony and Thoroughbred. No tagging groups were below 98% with the exception of the Draft group. Random sampling for the imputed Draft group individuals included the only two Percheron samples (M1542 and M1545), which underperformed (92.6% accuracy) compared to the other Draft samples in the imputed group (98.5%), which performed similarly to the other heavy horse breeds (e.g. Belgian).

SNP properties of the MNEc2M and MNEc670k arrays

SNP informativeness and Inter-SNP distance

Inter-SNP distance was calculated for the MNEc2M, as well as the MNEc670k, SNP sets at different levels of MAF (see Fig. 3 and Table 4) to assess the distribution of SNPs across the genome. On average, 1,250 and 3,756 bp, respectively, separated variants on the two arrays. Informativeness, defined as the number of SNPs with at least one heterozygote, was calculated for the same MAF cut-offs (Table 4). Inter-SNP distance, as well as informativeness, were broken down by breed for both the MNEc2M SNP set (See Supp. Fig. 2) and MNEc670k SNP set (See Supp. Fig. 3).

bioRxiv preprint doi: https://doi.org/10.1101/112979; this version posted March 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Alternate Allele Frequency

Frequency of alternate (non-reference) SNP alleles were calculated on both arrays using the full 485 sample dataset (WGS+SNP), genotypes derived from whole genome sequence only (WGS Only), and from samples genotyped on the 2M test array described above (SNP Only; see Table S5). Kernel density estimations (KDE) of the alternate allele frequency of SNPs on the MNEc2M array showed a mean frequency between 0.20 and 0.28 (See Fig. 4 and Table 5). In general, regardless of genotyping source, alternative allele frequencies shared similar distributions. Frequency distributions exhibit long tails indicating substantial numbers of samples with non-reference alleles. Alternate allele (ALT) distributions also exhibit a lower median frequency (red line) than mean frequency (red bar), a common property of right skewed distributions [34]. Alternate allele frequency distribution was also broken down by breed for both the MNEC2M array (Fig. 5) and the MNEc670k array (Supp. Fig 4) using genotypes derived from the WGS+SNP sample set. Allele frequency is balanced across breeds, though there were minute differences. For example, the median MAF for Thoroughbreds was 3-12% lower than all other breeds, though this is not unexpected given the reference genome is a Thoroughbred. Despite minor differences, all breeds had long tails and similar distributions of allele frequencies indicating a balanced and representative SNP selection for GWAS.

Linkage disequilibrium decay by breed

Linkage disequilibrium was measured using genotype r^2 between all 1.8M SNPs within 1Mb of one another within and across breeds. LD across breeds (i.e., the WGS+SNP sample sets) (see Fig 6, SNP and WGS curves) decayed faster than LD within any given breed (remaining curves). Within-breed calculations demonstrated that Quarter Horses and the Pony breeds had the lowest LD between SNPs at long distances, decaying to below 0.10 at 1Mb, while Thoroughbreds had the highest LD at all distances considered.

bioRxiv preprint doi: https://doi.org/10.1101/112979; this version posted March 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Discussion

Our goal was to provide a high-quality, standardized SNP array designed for imputation to overcome limitations in SNP density that under-power many genome mapping projects in the horse. To do this, we utilized whole genome sequencing data from 156 horses representing actively-researched and economically-important breeds that were collaboratively collected by 17 laboratories within the equine genetics community. These data, in turn, enabled a large scale, whole genome sequence mapping pipeline with *de-novo* variant calling resulting in a starting set of over 23 million variants. Variant filtering was based on genome coverage, breed representation, linkage disequilibrium, and feasibility of array probe design. This resulted in the successful design of a fully backwards compatible, high-density, 2 million SNP genotyping array (MNEc2M). Using a test cohort of 485 horses, we identified "tag" SNPs that reconstructed haplotypes both across diverse breeds and within breeds, then selected a subset of ~670k SNPs for design of the commercially available MNEc670k array. The use of tagging SNPs in this commercial array ensures its utility as an imputation tool up to a SNP density of at least 2 million.

Several successful GWAS in other agricultural animal species have been reported using this same Affymetrix® Axiom® HD genotyping array technology [35–37]. High-density ~670k genotyping arrays were used in domestic cattle to identify structural variation [38,39]. Other studies have leveraged this technology to discover and refine loci associated with production traits [40–42]. Based on reports from other domestic animals, coupled with preliminary analyses performed here, we anticipate similar performance and increased power for map-based studies to soon follow in the horse.

Assembly correction of the equine genome

Initial filtering for array compatible SNPs immediately reduced the pool from ~23 million SNPs that were discovered from *de novo* variant calling to 10 million SNPs compatible with array design. These purely technical constraints, such as removing all variants that were in close proximity to one another, substantially reduced the amount of variants considered here. While we thoroughly characterized the SNPs that were

bioRxiv preprint doi: https://doi.org/10.1101/112979; this version posted March 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

ultimately included on the genotyping array, the entire 23M SNP set has been submitted to dbSNP and additional analyses are being performed to further enhance knowledge of equine genetic diversity at the WGS level. This includes investigation of the > 400,000 sites where all WGS alleles differed from the reference allele determined by Sanger sequencing of Twilight. If the discrepancies in these data are verified, this information will be useful for further annotation and error correcting of the current equine genome reference assembly. Here, we focused on variants that were compatible with array design, however, we anticipate that improved genome annotation and finer scale haplotype maps will result from this larger SNP set.

Determining highly precise, gold standard SNP sets

Quality control and evaluation of conversion rates on the previous arrays have shown that many of the SNPs assayed on the legacy arrays are non-informative or not polymorphic in many breeds (See Table S2). These SNPs could truly occur at low frequencies within those populations, could have been singleton mutations in the individuals included in the legacy SNP discovery panel, or may have been false positive sites where no true genomic variation occurs. Compared to the SNP discovery panel used in the legacy arrays, the cohort of 153 horses used here provided a valuable opportunity to identify a precise set of SNPs that would be informative across many breeds.

While variant discovery from WGS had substantially higher overall depth of coverage than the SNP50 discovery effort, many breeds still had a modest number of individuals or relatively low average depth of coverage (Table S1). Still, genotype comparisons at variant sites from WGS to legacy arrays showed an overall high congruency rate (94% at 6X and 99% at 12X; Fig. 1), however, many SNPs with high QUAL score did not agree between WGS and legacy arrays. Since we were unable to determine the source of error in genotype mismatches between the WGS discovered variants and the legacy SNPs, especially those with high QUAL scores, we used additional information such as how many variant calling groups a SNP was discovered in as well as sequencing read level information to determine the criteria for including WGS discovered SNPs on the MNEc2M and MNEc670k arrays.

bioRxiv preprint doi: https://doi.org/10.1101/112979; this version posted March 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Distinguishing marginally higher quality variants, even at overall validation rates of over 94% becomes important when selecting a small fraction of variants to be included in a commercial array. Selecting a final subset (670k) containing only 6% of the initial ~10 million array-compatible variants posed an opportunity to control for overall false discovery rate during SNP selection. Variant recalibration allowed successful identification of variants with high QUAL scores that were less discordant between WGS and SNP50 genotypes. In the absence of high-confidence training SNP sets (i.e., dense HapMap data), we defined several different "gold standard" SNP sets for variant recalibration. While variant recalibration performed better than QUAL scores alone, there was not a single training set that clearly out-performed the others (Fig. 2). However, using recalibrated VQSLOD scores coupled with a focus on identifying high precision SNPs, did result in a high conversion rate of probes (92.3%) on the 2M test array. As more WGS data are generated in more horses allowing for further comparisons between SNPs genotyped using different technologies, and SNPs discovered here can be further validated, the ideal SNP training set for variant recalibration will become more formalized.

Genotyping success in blood versus hair root DNA

DNA samples genotyped on the MNEc2M array were primarily derived from blood, but also came from hair roots. Both DNA sources had failed samples, however, a substantially higher fraction of hair root samples produced poor genotyping rates. Hair root DNA also tended to have lower DNA concentrations when samples were re-hydrated. To maximize information, we chose to genotype all submitted samples, even though samples from hair roots did not meet minimal DNA quantity guidelines specified by Affymetrix, though samples were dropped if they did not meet genotyping quality control thresholds.

To determine if sample origin (blood versus hair roots), DNA quantity, or both, were associated with failure to pass genotyping quality control metrics, a logistic regression for sample success on DNA concentration (Pico Green) and blood/hair status was performed (Supp. Figure 5). It is clear that higher DNA concentrations increased the probability of genotyping success. However, while blood samples were more

bioRxiv preprint doi: https://doi.org/10.1101/112979; this version posted March 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

likely to produce passing samples regardless of DNA concentration, at adequate concentrations, hair root samples were also highly likely to produce passing genotypes.

We also noted variation in genotyping quality of specific SNPs based on tissue origin. During array design, SNPs were not disqualified that potentially performed better in blood versus hair root, e.g. 'PolyHighResolution' in one and not the other (Table S4). While it is difficult to determine if certain probes perform better using DNA isolated from one tissue versus the other, in the samples tested here, DNA isolated from blood was preferred over hair root when DNA concentration is low or questionable.

Setup for precision imputation

Although the legacy equine arrays were not designed for imputation, several previous studies have demonstrated their utility in imputing genotypes, which, if performed on a larger scale across many breeds, could greatly improve the chances of success in horse genome mapping studies. A study in Standardbreds, Quarter Horses and Thoroughbreds showed high fidelity imputation from legacy arrays (54k and 65k) to a higher marker density (74k) using a reference population of 248 horses [3]. Another study found that genotype imputation in Thoroughbreds was feasible from a very low density (1-3K markers) to a legacy SNP set (70K), although it was impacted by the minor allele frequency of the SNPs being imputed, as well as potentially complex LD structure [43].

We masked variants genotyped on the MNEc2M array down to the MNEc670k SNP set in a test cohort of 485 horses and found a high overall imputation accuracy (96-99%) up to the 2M density, across several different breeds of horses (See Table 3). While our imputation scenarios focused on imputing to the MNEc2M SNP set, imputation to higher SNP densities (>2M) is feasible. A recent study using 44 Franches-Montagnes and Warmbloods imputed SNPs from the legacy SNP50 genotypes to nearly 13M variants with 95% accuracy [44].

As sample size continues to plague the characterization of complex equine phenotypes, despite falling genotyping costs, the development of this affordable, backwards compatible, 670K imputation array will

bioRxiv preprint doi: https://doi.org/10.1101/112979; this version posted March 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

allow the inclusion of already genotyped individuals in higher-powered analysis, providing an economical trade-off for studies that must choose between more markers or more samples. We expect that imputation will complement future GWAS studies and mapping studies designed based on haplotype structure. In the future, improvements in imputation software as well as development of equine-specific imputation protocols will further increase accuracy of genotype imputation. Furthermore, as additional WGS are integrated into reference populations, imputation performance will continue to improve.

Conclusions

Here we report, through a community effort, the leveraging of WGS data from 153 individuals with a 13X median coverage to achieve new variant discovery in the domestic horse. The use of WGS enabled us to generate orders of magnitude more data than previous technologies allowed. Empirical SNP properties were used to produce composite scores for each SNP based on machine learning approaches, allowing better detection of false positive variant calls that could not be achieved using generic QUAL scores alone. Thus, from a starting set of over 23 million SNPs, we identified a set of ~5 million SNPs which were considered for array design. With probe design recommendations from Affymetrix, we further filtered this list to 2M SNPs (MNEc2M), which adequately represented the breeds used in variant discovery, were evenly spaced across the genome, and had the highest chance of conversion in the array implementation.

Using a test cohort of 332 horses, we used the MNEc2M SNP set to identify haplotype tagging SNPs both within and across breeds. Choices made in 2M array design, particularly with regard to variant filtering, resulted in greater than 92% of the SNPs on the 2M array (MNEc2M) returning high-quality genotypes. Filtering further, we designed the MNEc670k genotyping array to contain SNPs which allowed for accurate imputation. Together these genotyping platforms represent the next generation of genomic array technologies for the domestic horse.

bioRxiv preprint doi: https://doi.org/10.1101/112979; this version posted March 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Methods

Whole-genome sequencing and mapping pipeline

Paired-end Illumina Hi-Seq 100 base pair whole genome sequences were generated for 153 horses representing 24 distinct breeds (Table S1). Raw reads for each individual were mapped using the PALEOMIX pipeline [26] and aligned to an extended EquCab version 2.0 genome (see below). Specifically, reads for each individual were processed separately by sequencing lane and filtered for quality control using the program Adapter Removal [45], which removed PCR adapters, trimmed low quality base-pairs, and collapsed overlapping paired-end reads into a single high-quality read, and the resulting reads were filtered by length. Passing reads were mapped to each reference file using BWA [46]. Paired-end reads for which the mate was filtered were mapped in single-end mode. PCR duplicates were detected and removed, the resulting bam-files were merged, and reads were realigned around detected indels.

Extended EquCab 2.0 reference genome

An extended version of the EquCab2 reference genome was used which included the 31 autosomal chromosomes, ECAX, and equine chromosome unknown 1 (ChrUn1) [1], together with an additional 7,850 contigs designated as ChrUn2 generated from unmapped Twilight genomic DNA reads which were de novo assembled using the Velvet assembler [47]. These additional contigs were required to meet any of the following criteria: 1) longer than 1,000 base pairs with no BLAST alignment (bit score > 99) to the human, canine, or bovine genomes; 2) longer than 1,000 base pair and either a single BLAST alignment to a human, canine, or bovine chromosome or multiple alignments that mapped to a single chromosome; 3) between 500 and 999 base pairs and a BLAST alignment to a single human, canine, or bovine chromosome with a bit score >499; 4) between 500 and 999 bp with alignment to a single chromosome where the total coverage of the aligned region included more than 80% of the coding length of an existing human, canine, or bovine annotated protein-coding gene.

bioRxiv preprint doi: https://doi.org/10.1101/112979; this version posted March 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

PALEOMIX *vcf pipeline* implementation and python source code

Programs within PALEOMIX are abstracted as nodes within the program to allow files to be generated within a temporary directory and only to be moved to the final directory upon successful completion. In addition to nodes for the read alignment mapping programs, additional nodes were created to run variant calling programs implemented by GATK [27] and SAMtools [28]. Additional PALEOMIX nodes were created to perform variant recalibration as well as to assess precision versus recall for Morgan and Standardbred breed groups (Fig. 1 and Fig. 2). Source code for the extended PALEOMIX nodes are available at https://github.com/schae234/pypeline.

Variant calling and validation PALEOMIX nodes

The PALEOMIX computational framework was further extended to process alignments and produce variant calls. Individuals were split into 16 different calling groups based on both breed and sequencing depth in order to minimize biases due to coverage and population stratification. Variant call files (.vcf) were produced for each group using both GATK Unified Genotyper with" –stand_call_conf" set to 30, "—stand_emit_conf" set to 10, and "–dcov" set to 200. Variants called by GATK were retained if they were independently called by SAMtools "mpileup" with default calling parameters. This allowed possible false negatives to be assessed downstream using other quality metrics. A total of ~23M variants called independently by both GATK and SAMtools were retained for further analysis.

Variant recalibration (VQSLOD scores)

Both GATK UnifiedGenotyper as well as SAMtools assign generic quality scores (QUAL) to each discovered variant, which is the posterior probability that a true variant exists given the pileup of reads at a given locus using base pair quality and expected allelic distribution of samples. Using GATK, additional metrics were generated for each variant including coverage (DP), quality of depth (QD), Fisher strand bias (FS), mapping quality rank sum (MQRankSum) and read position rank sum (ReadPosRankSum). Definitions from the GATK manual for each metric are below.

bioRxiv preprint doi: https://doi.org/10.1101/112979; this version posted March 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Coverage (DP) – Total, unfiltered depth of coverage.

Quality by Depth (QD) – Variant confidence (from QUAL field) divided by depth of non-reference samples.

Fisher Strand Bias (FS) – Measure of strand bias, i.e. the variation seen on only the forward or reverse strand.

Mapping Quality Rank Sum (MQRankSum) – The rank sum test for mapping qualities.

Read Position Rank Sum (ReadPosRankSum) – The rank sum test for the distance of the variant from the end of reads.

These parameters were used to train linear mixed models using GATK VariantRecalibrator which produces a composite variant quality metric called a VQSLOD score which can be summarized as using the formula VQSLOD ~ DP + QD + FS + MQRankSum + ReadPosRankSum. We defined three "gold standard" datasets: 1) SNPs previously genotyped on the 54K SNP chip, 2) variants called in four or more calling groups, 3) and variants within the 99th percentile of QUAL scores. Models were trained on chromosomes 2-31 and validated on chromosome 1 in both Morgan and Standardbred breed groups as a subset of each breed group had whole genome sequence data at target depths of 6X (12 Morgans and 6 Standardbreds) and 12X (6 Morgans and 4 Standardbreds) target coverage (actual coverage MOR6X: 5.23-6.96X, mean=6.11X; STD6X: 4.61-5.89X, mean=5.29X; MOR12X: 10.42-15.14X, mean=12.96X; STD12X: 11.21-11.99X, mean=11.63X). These "gold standard" training SNPs were used to assess the precision of different quality metrics of *de novo* variant calling.

Preliminary 5M high-density SNP selection

A preliminary list of approximately 5 million SNPs was prepared for probe design (conversion) recommendation by Affymetrix. VQSLOD scores were generated for all ~23 million initial bi-allelic candidate SNPs. SNPs were filtered out if they 1) fell within repetitive regions designated by RepeatMasker

bioRxiv preprint doi: https://doi.org/10.1101/112979; this version posted March 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

3.3.0 [48] (780 SNPs); 2) fell within 20bp of another variant (9,342,733 SNPs); and 3) had fewer than 2 observed instances of minor allele (4,858,273). Of the remaining 11,435,936 SNPS, further filtering was applied to obtain equal coverage throughout the genome. Pairwise LD was calculated between SNPs within 10kb windows. If SNPs within a window had an r^2 value of over 0.90, priority was given to SNPs that were in called in Draft or Pony groups. If a SNP was in an equal number of priority groups, VQSLOD scores were used to break ties. After all filtering criteria were applied 5,443,950 SNPs remained.

High-density 2M SNP selection

From the ~5 million candidate SNPs, Affymetrix provided four classes of recommendations based on the probability the SNP would be convertible through probe design ('recommended', neutral, 'non-recommended', 'not possible'). A recommended SNP has: a probability of conversion > 0.60, no interfering polymorphisms within 24 bases, and unique flanking sequence. Non-recommended SNPs have: either duplicate flanking sequence, a probability of conversion < 0.40, interfering polymorphisms within 21 bases, or more than 2 interfering polymorphisms within 24 bases. A 'not possible' designation is given to probes which probes cannot be created, and neutral recommendations cover all other cases. Furthermore, SNPs with alleles of A/T or C/G required two probes to differentiate between allele states and were tagged as allele-specific SNPs. Recommendations were generated for both forward and reverse strands based on the above criteria. Groups of SNPs were ranked within 50kb windows and probe design criteria were chosen using a greedy algorithm until 37 SNPs were chosen per window using the following criteria: 1) VIP SNPs previously designed on 54/65K chip (regardless of recommendation or strand), 2) SNPs for known Mendelian traits such as coat color (regardless of recommendation or strand), 3) SNPs designated as 'recommended' by Affymetrix and designable with one probe, 4) SNPs with a 'neutral' recommendation from Affymetrix designable with one probe, 5) any SNP within the equine MHC region, 6) SNPs requiring multiple probes. If a SNP was equally designable in forward and reverse strand, the forward strand was arbitrarily chosen. Using these criteria, 2,001,826 SNPs were chosen for array design.

bioRxiv preprint doi: https://doi.org/10.1101/112979; this version posted March 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

2M SNP test array and sample quality control

Using probes designed according to the above criteria, 347 horses from 20 breeds were genotyped on the 2M array using DNA isolated from blood using the Gentra PureGene Blood kit. DNA from hair roots was isolated using the Gentra Puregene Blood Kit with a modified version of the Gentra Puregene Mouse Tail purification protocol. The amount of Proteinase K was increased to 20 µL, isopropanol to 650 µL, and ethanol to 500 µL. The DNA hydration solution was reduced to 20-30 µL depending on size of DNA pellet. Quality control metrics were calculated on arrays grouped by tissue type (blood vs hair) as well as array batches, and arrays were dropped for various reasons at multiple QC steps. "DishQC" evaluates call rates between A/T probes and C/G probes of non-polymorphic sites to assess background probe contrast and was calculated using Affymetrix Power Tools. Samples with DQC scores below 0.60 were removed from further evaluation. Passing arrays were then genotyped using approximately 20,000 high-confidence probes provided by Affymetrix in their R1 release package (See http://www.affymetrix.com/estore/catalog/prod1010002/AFFY/Axiom%26%23174%3B+Equine+Genotyping+Array#1_1). If the number of arrays per tissue/array group was above 96 samples, generic priors were used for probe-set genotyping; otherwise, SNP specific priors were used (provided in the R1 package released by Affymetrix). Sample arrays with 20K call rates below 0.97 were removed from further analysis.

Remaining samples (n=347) were genotyped on all probes using a similar approach. Generic or SNP specific priors (provided in R1 package) were used for groups containing more or less than 96 samples, respectively, and samples with call rates below 0.97 were removed from the analysis. Successfully genotypes samples were then grouped by genotyping plates to check for batch effects. Plates with pass rates below 95% in blood or 93% in hair were dropped.

Up to 4 probes were used to genotype each SNP across samples passing quality control criteria. Probe performance was calculated using the `*Metrics.R*` script provided by Affymetrix which assessed several quality criteria. Twelve criteria were used to assign each SNP into 6 categories representing probe conversion types. High-quality probes fell in one of three categories in descending order of quality:

bioRxiv preprint doi: https://doi.org/10.1101/112979; this version posted March 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

'PolyHighResolution' SNPs had good cluster resolution and at least two examples of the minor allele, 'MonoHighResolution' SNPs had good SNP clustering but less than two samples with the minor allele, and 'NoMinorHom' had good cluster resolution but no samples with homozygous, minor alleles. Poor quality SNPs were qualified as having off target variation (OTV), a call rate below threshold, or a combination of poor performing cluster properties. SNPs with multiple probes were assigned a best probe based on the above high-quality conversion types with the added constraint that SNPs with no minor homozygote did not result in extreme Hardy-Weinberg values ($p \leq 10e\text{-}5$; Chi-Squared Test). 1,846,988 SNPs passed the above quality control metrics.

670K commercial array SNP selection

Genotype information derived from the horses on the 2M test array (n=332) were combined with SNPs in horses called by whole genome sequencing (n=153) resulting in a total of 485 horses genotyped at 1,846,988 SNPs. Tag SNPs were calculated using FastTagger [32] using two different tagging scenarios. Tagging SNPs informative across populations (inter-population) were identified by running FastTagger on the entire dataset. Parameters provided to FastTagger identified inter-population tag-SNPs at a resolution of down to 0.01 minor allele frequency and up to 0.99 r^2.

To identify SNPs tagging population specific haplotypes (intra-population), samples were split into 15 tagging breed groups based on available sample size: Land Race, Arabian, Belgian, Draft, Franches-Montagne, Icelandic, Lusitano, Maremanno, Morgan, Pony, Quarter Horse, Standardbred, Thoroughbred, Trotter, and Warmblood (See Table S5). FastTagger was run with parameters to differentiate tag-SNPs within each population at a resolution of 0.10 minor allele frequency and 0.90 r^2. Representative SNPs from each population specific set of tag SNPs were collapsed using the program `Multi-Pop-TagSelect` which assesses overlap between sets of tag SNPs and identifies the subset of SNPs which near-minimally spans the set of populations [33]. Intra-breed specific SNPs were kept as long as they were tagged haplotypes in five or more breed groups.

bioRxiv preprint doi: https://doi.org/10.1101/112979; this version posted March 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Quarter Horses, Ponies, Morgans, and Standardbreds needed a much higher number of intra-breed specific SNPs to tag haplotypes, indicating they were the most diverse breeds. Additional tag SNPs (n=13,993) were included in the final commercial array if they tagged haplotypes in three or more of these diverse breed groups.

Genotype imputation between the MNEc670k and MNEc2M arrays

A reference population of the 485 horses with genotypes at ~2M SNPs (generated either by the MNEc2M array or whole-genome sequence) was used as a reference population for imputation. Horses from each different breed were used to test the imputation accuracy of the array. A random set of 1/3 of individuals were removed from each breed group and masked down to the MNEc670k SNP set as well as the SNP65 SNP set. Genotypes were then imputed using the remaining (non-masked) individuals as a reference population. Imputation was performed using Beagle 4.0 [49] using default parameters. Genotype concordance was calculated using VCFtools (0.1.15) [50].

Abbreviations

SNP: Single Nucleotide Polymorphism; MNEc2M: the 2 million SNP array developed here; MNEc670k: the 670 thousand SNP array developed here; WGS: whole genome sequencing; bp: base pair; GATK: genome analysis tool kit; QUAL: generic quality score output by GATK and SAMtools; MOR: the Morgan horse breed; STD: the Standardbred horse breed; VQSLOD: variant quality score log-odds; Gb: giga-bases; MHC: major histocompatibility complex; ECA: Equus caballus chromosome; VIP: very important probe; MAF: minor allele frequency; ALT: alternate allele; KDE: kernel density estimation; LD: linkage disequilibrium; ChrUn1: Equus calballus unknown chromosome;

bioRxiv preprint doi: https://doi.org/10.1101/112979; this version posted March 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Declarations

Acknowledgements

These genomic tools are the result of a collaborative effort that involved a number of laboratories within the equine genetics community, many of which provided support or insight at various steps for this community resource. We would like to acknowledge Chad Dow from Affymetrix for the support he provided while designing the MNEc2M array.

Author Contributions

Experimental concept and design: MEM, JRM, LO, RJS, MS; Sample collection and data contribution: EB, DLB, EB, GKB, GB, SAB, OD, RF, CJF, VG, BH, VJ, TK, TL, GL, MSL, NM, ACM, JNM, AM, JM, CP, SP, SR, IT, JT, GT, AVS, CMW, BW, LO, JRM, MEM; Data analysis and interpretation: RJS, MS, MEM; Computational support: RJS, MS; Manuscript writing and figures: RJS, JRM, MEM; Manuscript review: All authors read and approved the final manuscript.

Competing Interests

The authors declare that they have no competing interests.

Animal Welfare

DNA samples were previously collected with approval from the Animal Care and Use Committees at the respective institutions.

Funding

This work was supported by:

- USDA NIFA project 2012-67015-19432 and Minnesota Agricultural Experiment Station Multistate project MIN-62-090.

bioRxiv preprint doi: https://doi.org/10.1101/112979; this version posted March 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

- The National Animal Genome Project (NRSP8) through the equine genome coordinator: USDA-NRSP8 (2013-2018) horse-technical-committee coordinator funds.
- The Danish Council for Independent Research, Natural Sciences (Grant 4002-00152B); the Danish National Research Foundation (Grant DNRF94); Initiative d'Excellence Chaires d'attractivité, Université de Toulouse (OURASI), and; the European Research Council (ERC-CoG-2015-681605);
- The Bavarian Ministry State Ministry for Food and Agriculture, and Forestry (A/13/39) for providing funding for generating whole genome sequence.
- The Laboratory of Molecular Evolution, The Koret School of Veterinary Medicine, The Hebrew University of Jerusalem, Israel) for contributing pure-bred Arabian whole-genomes on behalf of The Israel Science Foundation (ISF) grant #1365/10
- The Swedish Research Council Formas (221-2013-1661) and the Swedish Research Council VR (621-2012-4666)

Data Availability

Whole genome sequences are available in the following NCBI BioProjects: PRJEB14779, PRJNA273402, and PRJEB10098. Additional sequences have are restricted in availability due to pre-existing material transfer agreements and can be requested by contacting the contributing investigator in Table S1. Genotypes for horses on the MNec2M array will be released upon publication. Genome positions for all 23 million discovered SNPs have been submitted to dbSNP.

bioRxiv preprint doi: https://doi.org/10.1101/112979; this version posted March 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Figures

Fig. 1

Comparing high QUAL genotypes called de-novo to the SNP50 array in Morgans and Standardbreds

SNPs on chromosome (ECA1) 1 with genotype calls from both WGS and SNP50 array were ranked by decreasing QUAL score for Morgans (MOR) at 6X (3,759 SNPs; 12 individuals) and 12X (3,751 SNPs; 6 individuals) coverage, and Standardbreds (STD) at 6X (4,422 SNPs; 5 individuals) and 12X (4,380 SNPs; 3 individuals) coverage. The cumulative proportion of genotypes that agree between platforms was compared based on ranking *de-novo* variants by QUAL score. Variants with missing data on either platform were excluded.

Fig. 2

Comparing QUAL ranked SNPs to VQSLOD ranked SNPs

VQSLOD scores were calculated from three different "gold standard" reference groups in Morgans (MOR) and Standardbreds (STD) using GATK VariantRecalibrator. Compared to QUAL scores (red line), high VQSLOD scored variants (top 10% variants by score) have a lower number of mismatched genotypes across the SNP50 BeadChip and variants discovered *de novo*.

Fig. 3

MNEc2M and MNEc670K Inter-SNP Distance

Distance between SNPs on each array was calculated using various minor allele frequency (MAF) cutoffs. Considering all available SNPs genotyped on the MNEc2M and MNEc670k arrays, on average, 1,250 and 3,756 bp separate markers (See Table 4 for average inter-SNP distances at the various MAF). Median values (red lines) and mean values (red boxes) were calculated at each MAF cutoff.

bioRxiv preprint doi: https://doi.org/10.1101/112979; this version posted March 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Fig. 4

MNEc2M and MNEc670k Alternate Allele Frequency

Genotypes from horses on the 2M test array (SNP Only; n=332) as well as whole genome sequence (WGS Only; n=153) were combined (WGS+SNP; n=485) to estimate alternate allele frequency of the SNPs represented on the (**A**) 2M and (**B**) 670K arrays. Fig. 4 shows kernel density estimated (KDE) distributions of alternate allele (ALT) frequency in each sample group using variants that are on each array. Boxplots show ALT frequency distribution median (red line), mean (red square) as well as variance (See Table 5 for values).

Fig. 5

MNEc2M Breed Specific Alternate Allele Frequency

Alternate allele frequencies from variants present on the MNEc2M chip were split by breed group. Samples (WGS+SNP) were split into 15 tagging breed groups (See Table S5). Breed groups with asterisk (*) indicate a combination of studbook breeds. Boxplots show median (red line) and mean (red box) of the alternate allele frequency distribution.

Fig 6

Linkage Disequilibrium decay within and across breeds

Pairwise r^2 was calculated between each SNP within 1Mb having a minor allele frequency greater than 0.05. LD curves are broken down by breed as well as for samples derived from the all-breed WGS as well as from all-breed SNP cohorts. Breeds are ordered (TBLR) in the legend by their r^2 values at 400kb.

bioRxiv preprint doi: https://doi.org/10.1101/112979; this version posted March 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Tables

Table 1

SNP Sets used at various steps in array design

Variants discovered from whole genome sequencing were filtered at various steps for quality control or using array design criteria. Six distinct sets of variants ranging from the initial ~23M high-quality, variants discovered from WGS to the 670k variants available in the commercial genotyping array are described throughout this manuscript.

Table 2

Number of breed specific tagging SNPs

Number of tagging SNPs required to reconstruct haplotypes in each breed (MAF > 0.10 and 0.90 r^2, See Methods).

Table 3

Imputation accuracy of the MNEc670k SNP genotyping array

Breed specific imputation accuracy (mean +/- s.e.m.) of genotypes from MNEc670k to MNEc2M SNP sets. In each tagging breed group, 1/3 of samples genotypes were masked to lower density SNP sets and removed from the reference population of 485 horses. Imputation was performed using Beagle 4.0 and concordance was determined with VCFtools.

Table 4

MNEc2M and MNEc670k Inter-SNP distance at various minor allele frequency cutoffs

Inter-SNP distance was calculated between SNPs informative at minor allele cutoffs greater than 0, 0.01, 0.03, 0.05 and 0.10. The number of SNPs included at this MAF cutoff is included. Distance and

bioRxiv preprint doi: https://doi.org/10.1101/112979; this version posted March 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

informativeness was re-calculated on both MNEc2M and MNEc670k arrays which were further broken down by tagging breed group (See Table S5).

Table 5

MNEc2M and MNEc670k variant mean and median alternate allele frequency

Average (mean and median) values for MNEc2M and MNEc670k arrays broken down by genotype information available from WGS, CHIP or WGS+CHIP.

Supplementary Files

Table S1

Whole genome sequencing (WGS) samples

Whole genome sequencing samples with horse identification and read depth. 153 Individuals representing 24 breeds (Twilight is included as 4 entries). Table includes horse identifier, breed, contributing laboratory and coverage statistics for each individual in both the nuclear and mitochondrial genomes. Variant calling groups indicate which individuals were grouped together during variant discovery.

Table S2

SNP50 breed informativeness

Percent of legacy SNP50 BeadChip SNPs that are informative (MAF $\geq$ 0.05) in light, draft and pony breeds broken down by discovery breed described by McCue et al. [1,2].

bioRxiv preprint doi: https://doi.org/10.1101/112979; this version posted March 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Table S3

SNP validation by discovery by variant calling group

SNPs discovered from whole genome sequence were validated by assessing the minor allele frequency in individuals genotyped on the MNEc2M test array.

Table S4

MNEc670k SNP information

Contains information on MNEc670k SNPs. Includes Affymetrix probe ID, genetic coordinates, MNEc Identifier, tissue origin indicating high quality genotyping in either hair root or blood, Affymetrix assigned genotyping cluster resolution, and inclusion criteria:

Intra – SNP tags intra-breed haplotype

Inter – SNP tags inter-breed haplotype

MHC – SNP within Equine Major Histocompatibility Complex region

Diverse – SNP tags haplotype in 3 or more of diverse breeds that need many tagging SNPs (Quarter Horse, Pony, Morgan, Standardbred; See Table S4)

Density – SNP included to target at least 8 SNPs per 50kb genomic window (targeting uniform genomic coverage)

Table S5

Sample breed and tagging breed groups

Samples (n=485) from either whole genome sequence (WGS) or from the 2M test array (SNP) were assigned to one of 15 tagging breed groups based on their reported breed.

bioRxiv preprint doi: https://doi.org/10.1101/112979; this version posted March 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Supp. File 1

Position of variants discovered from WGS data

File contains chromosome, bp position, distance from previous discovered SNP and alternate allele frequency for the 23 million SNP set.

Supp. File 2

MNEc670k SNP information

Contains information on MNEc670k SNPs. Includes Affymetrix probe ID, genetic coordinates, MNEc Identifier, tissue origin indicating high quality genotyping in either hair root or blood, Affymetrix assigned genotyping cluster resolution, and inclusion criteria:

Intra – SNP tags intra-breed haplotype

Inter – SNP tags inter-breed haplotype

MHC – SNP within Equine MHC region

Diverse – SNP tags haplotype in 3 or more of diverse breeds that need many tagging SNPs (Quarter Horse, Pony, Morgan, Standardbred; See Table S4)

Density – SNP included to target at least 8 SNPs per 50kb genomic window (targeting uniform genomic coverage)

Supp. File 3

Breed specific tagging SNPs

File containing informative tagging SNPs broken down by breed.

bioRxiv preprint doi: https://doi.org/10.1101/112979; this version posted March 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Supplementary Figures

Supp. Fig. 1

Alternate allele frequency for 23M and 10M SNP sets.

Histograms show the alternative allele frequency for the ~23 million SNPs discovered by WGS and the ~10 million SNPs that were compatible with array design (see Table 1). A high number of SNPs were observed at 100% alternate allele frequency in a sample cohort that included deep sequencing (see Table S1) of Twilight (the horse used for the reference genome) indicating likely Sanger sequence errors.

Supp. Fig. 2

MNEc2M Inter-SNP distance and informativeness by breed.

Distance between SNPs on the MNEc2M array were calculated by breed group (See Table S5) at various minor allele frequency cutoffs (0, 0.01, 0.03, 0.05, 0.10). Breed groups with asterisk (*) indicate a combination of studbook breeds.

Supp. Fig. 3

MNEc670k Inter-SNP distance and informativeness by breed

Distance between SNPs on the MNEc670k array were calculated by breed group (Table S5) at various minor allele frequency cutoffs (0, 0.01, 0.03, 0.05, and 0.10). Breed groups with asterisk (*) indicate a combination of studbook breeds.

bioRxiv preprint doi: https://doi.org/10.1101/112979; this version posted March 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Supp. Fig 4

MNEc670k Breed Specific Alternate Allele Frequency

Alternate allele frequencies from variants present on the MNEc670k chip were split by breed group. Samples (WGS+SNP) were split into 15 tagging breed groups (See Table S5). Breed groups with asterisk (*) indicate a combination of studbook breeds.

Supp. Figure 5

Logistic regression of DNA Sample success

A logistic regression was fit, predicting the probability of a sample passing quality control using DNA concentration (Pico Green) and sample source as independent variables.

Works Cited

1. 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 [Internet]. 2009 [cited 2015 Jan 5];326:865–7. Available from: http://www.sciencemag.org/content/326/5954/865.short

2. McCue ME, Bannasch DL, Petersen JL, Gurr J, Bailey E, Binns MM, et al. A high density SNP array for the domestic horse and extant Perissodactyla: utility for association mapping, genetic diversity, and phylogeny studies. Georges M, editor. PLoS Genet. [Internet]. Public Library of Science; 2012 [cited 2012 Jul 30];8:e1002451. Available from: http://dx.plos.org/10.1371/journal.pgen.1002451

3. McCoy AM, McCue ME. Validation of imputation between equine genotyping arrays. Anim. Genet. 2014;45:153.

4. 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. Proc. Natl. Acad. Sci. [Internet]. National

bioRxiv preprint doi: https://doi.org/10.1101/112979; this version posted March 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Academy of Sciences; 2014 [cited 2014 Dec 16];111:201416991. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25512547

5. Jónsson H, Schubert M, Seguin-Orlando A, Ginolhac A, Petersen L, Fumagalli M, et al. Speciation with gene flow in equids despite extensive chromosomal plasticity. Proc. Natl. Acad. Sci. U. S. A. [Internet]. National Academy of Sciences; 2014 [cited 2016 Sep 7];111:18655–60. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25453089

6. McCoy AM, Beeson SK, Splan RK, Lykkjen S, Ralston SL, Mickelson JR, et al. Identification and validation of risk loci for osteochondrosis in standardbreds. BMC Genomics [Internet]. BioMed Central; 2016 [cited 2016 Sep 7];17:41. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26753841

7. McQueen CM, Dindot S V, Foster MJ, Cohen ND. Genetic Susceptibility to Rhodococcus equi. J. Vet. Intern. Med. [Internet]. Wiley-Blackwell; 2015 [cited 2016 Sep 7];29:1648–59. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26340305

8. Hauswirth R, Haase B, Blatter M, Brooks SA, Burger D, Drögemüller C, et al. Mutations in MITF and PAX3 cause “splashed white” and other white spotting phenotypes in horses. Barsh GS, editor. PLoS Genet. [Internet]. Public Library of Science; 2012 [cited 2016 Sep 7];8:e1002653. Available from: http://dx.plos.org/10.1371/journal.pgen.1002653

9. Hill EW, McGivney BA, Gu J, Whiston R, Machugh DE. A genome-wide SNP-association study confirms a sequence variant (g.66493737C>T) in the equine myostatin (MSTN) gene as the most powerful predictor of optimum racing distance for Thoroughbred racehorses. BMC Genomics [Internet]. BioMed Central; 2010 [cited 2016 Sep 7];11:552. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20932346

10. Lykkjen S, Dolvik NI, McCue ME, Rendahl AK, Mickelson JR, Roed KH. Genome-wide association analysis of osteochondrosis of the tibiotarsal joint in Norwegian Standardbred trotters. Anim. Genet.

bioRxiv preprint doi: https://doi.org/10.1101/112979; this version posted March 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

[Internet]. 2010 [cited 2016 Sep 7];41 Suppl 2:111–20. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21070284

11. Raudsepp T, McCue ME, Das PJ, Dobson L, Vishnoi M, Fritz KL, et al. Genome-Wide Association Study Implicates Testis-Sperm Specific FKBP6 as a Susceptibility Locus for Impaired Acrosome Reaction in Stallions. Barsh GS, editor. PLoS Genet. [Internet]. Public Library of Science; 2012 [cited 2016 Jun 7];8:e1003139. Available from: http://dx.plos.org/10.1371/journal.pgen.1003139

12. Lykkjen S, Dolvik NI, McCue ME, Rendahl AK, Mickelson JR, Røed KH. Equine developmental orthopaedic diseases--a genome-wide association study of first phalanx plantar osteochondral fragments in Standardbred trotters. Anim. Genet. [Internet]. 2013 [cited 2016 Sep 7];44:766–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23742657

13. Signer-Hasler H, Flury C, Haase B, Burger D, Simianer H, Leeb T, et al. A genome-wide association study reveals loci influencing height and other conformation traits in horses. Weedon MN, editor. PLoS One [Internet]. Public Library of Science; 2012 [cited 2016 Sep 7];7:e37282. Available from: http://dx.plos.org/10.1371/journal.pone.0037282

14. Corbin LJ, Blott SC, Swinburne JE, Sibbons C, Fox-Clipsham LY, Helwegen M, et al. A genome-wide association study of osteochondritis dissecans in the Thoroughbred. Mamm. Genome [Internet]. Springer-Verlag; 2012 [cited 2016 Sep 7];23:294–303. Available from: http://link.springer.com/10.1007/s00335-011-9363-1

15. Finno CJ, Stevens C, Young A, Affolter V, Joshi NA. SERPINB11 Frameshift Variant Associated with Novel Hoof Specific Phenotype in Connemara Ponies. 2015;1–17.

16. Kader A, Li Y, Dong K, Irwin DM, Zhao Q, He X, et al. Population Variation Reveals Independent Selection toward Small Body Size in Chinese Debao Pony. Genome Biol. Evol. [Internet]. Oxford University Press; 2016 [cited 2016 Sep 7];8:42–50. Available from:

bioRxiv preprint doi: https://doi.org/10.1101/112979; this version posted March 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

http://www.ncbi.nlm.nih.gov/pubmed/26637467

17. Petersen JL, Mickelson JR, Cothran EG, Andersson LS, Axelsson J, Bailey E, et al. Genetic diversity in the modern horse illustrated from genome-wide SNP data. PLoS One [Internet]. Public Library of Science; 2013 [cited 2014 Oct 20];8:e54997. Available from: http://www.plosone.org/article/info%25253Adoi%25252F10.1371%25252Fjournal.pone.0054997#pone-0054997-g004

18. Schubert M, Ermini L, Der Sarkissian C, Jónsson H, Ginolhac A, Schaefer R, et al. Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX. Nat. Protoc. [Internet]. 2014;9:1056–82. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24722405

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. Nature [Internet]. 2013;499:74–8. Available from: http://dx.doi.org/10.1038/nature12323%5Cn10.1038/nature12323%5Cnhttp://www.nature.com/nature/journal/v499/n7456/abs/nature12323.html#supplementary-information

20. Sarkissian C Der, Ermini L, Schubert M, Yang MA, Librado P, Fumagalli M, et al. Przewalski ' s Horse. 2016;25:2577–83.

21. Librado P, Der Sarkissian C, Ermini L, Schubert M, Jónsson H, Albrechtsen A, et al. Tracking the origins of Yakutian horses and the genetic basis for their fast adaptation to subarctic environments. Proc. Natl. Acad. Sci. [Internet]. 2015;112:201513696. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4687531&tool=pmcentrez&rendertype=abstract

22. Metzger J, Tonda R, Beltran S, Agueda L, Gut M, Distl O. Next generation sequencing gives an

bioRxiv preprint doi: https://doi.org/10.1101/112979; this version posted March 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

insight into the characteristics of highly selected breeds versus non-breed horses in the course of domestication. BMC Genomics [Internet]. 2014;15:562. Available from: http://www.scopus.com/inward/record.url?eid=2-s2.0-84904420952&partnerID=tZOtx3y1

23. Doan R, Cohen ND, Sawyer J, Ghaffari N, Johnson CD, Dindot SV. Whole-Genome sequencing and genetic variant analysis of a quarter Horse mare. BMC Genomics [Internet]. 2012;13:78. Available from: http://www.biomedcentral.com/1471-2164/13/78

24. Andersson LS, Larhammar M, Memic F, Wootz H, Schwochow D, Rubin C-J, et al. Mutations in DMRT3 affect locomotion in horses and spinal circuit function in mice. Nature [Internet]. 2012;488:642–6. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3523687&tool=pmcentrez&rendertype=abstract

25. Frischknecht M, Neuditschko M, Jagannathan V, Drögemüller C, Tetens J, Thaller G, et al. Imputation of sequence level genotypes in the Franches-Montagnes horse breed. Genet. Sel. Evol. [Internet]. 2014;46:63. Available from: http://www.gsejournal.org/content/46/1/63

26. Schubert M, Ermini L, Sarkissian C Der, Jónsson H, Ginolhac A, Schaefer R, et al. Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX. Nat. Protoc. [Internet]. Nature Publishing Group; 2014 [cited 2016 Jun 7];9:1056–82. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24722405

27. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. [Internet]. 2010 [cited 2016 Sep 13];20:1297–303. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20644199

28. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and

bioRxiv preprint doi: https://doi.org/10.1101/112979; this version posted March 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

population genetical parameter estimation from sequencing data. Bioinformatics [Internet]. 2011 [cited 2016 Sep 13];27:2987–93. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21903627

29. Raghavan V, Bollmann P, Jung GS. A critical investigation of recall and precision as measures of retrieval system performance. ACM Trans. Inf. Syst. 1989;7:205–29.

30. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One. 2015;10:1–21.

31. DePristo MA, Banks E, Poplin R, Garimella K V, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. [Internet]. Nature Research; 2011 [cited 2016 Oct 26];43:491–8. Available from: http://www.nature.com/doifinder/10.1038/ng.806

32. Liu G, Wang Y, Wong L. FastTagger: an efficient algorithm for genome-wide tag SNP selection using multi-marker linkage disequilibrium. BMC Bioinformatics [Internet]. 2010 [cited 2014 Sep 11];11:66. Available from: http://www.biomedcentral.com/1471-2105/11/66

33. Howie BN, Carlson CS, Rieder MJ, Nickerson D a. Efficient selection of tagging single-nucleotide polymorphisms in multiple populations. Hum. Genet. 2006;120:58–68.

34. von Hippel PT. Mean, Median, and Skew: Correcting a Textbook Rule. J. Stat. Educ. American Statistical Association. 732 North Washington Street, Alexandria, VA 22314. Tel: 703-684-1221; Tel: 888-231-3473; Fax: 703-684-2037; e-mail: asainfo@amstat.org; Web site: http://www.amstat.org/publications/jse; 2005;13.

35. Kranis A, Gheyas AA, Boschiero C, Turner F, Yu L, Smith S, et al. Development of a high density 600K SNP genotyping array for chicken. BMC Genomics [Internet]. 2013;14:59. Available from: http://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-14-59

bioRxiv preprint doi: https://doi.org/10.1101/112979; this version posted March 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

36. Groenen MAM. Development of a high-density Axiom® porcine genotyping array to meet research and commercial needs [Internet]. Plant Anim. Genome XXIII Conf. San Diego, CA: Plant & Animal Genome XXIII Conference; 2015. Available from: https://pag.confex.com/pag/xxiii/webprogram/Paper17949.html

37. Rincon G, Weber KL, Eenennaam a L Van, Golden BL, Medrano JF. Hot topic: performance of bovine high-density genotyping platforms in Holsteins and Jerseys. J. Dairy Sci. [Internet]. Elsevier; 2011;94:6116–21. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22118099

38. Salomón-Torres R, González-Vizcarra VM, Medina-Basulto GE, Montaño-Gómez MF, Mahadevan P, Yaurima-Basaldúa VH, et al. Genome-wide identification of copy number variations in Holstein cattle from Baja California, Mexico, using high-density SNP genotyping arrays. Genet. Mol. Res. [Internet]. 2015;14:11848–59. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26436509

39. Salomon-Torres R, Villa-Angulo R, Villa-Angulo C. Analysis of copy number variations in Mexican Holstein cattle using axiom genome-wide Bos 1 array. Genomics Data [Internet]. The Authors; 2016;7:97–100. Available from: http://dx.doi.org/10.1016/j.gdata.2015.12.007

40. Romé H, Varenne A, Hérault F, Chapuis H, Alleno C, Dehais P, et al. GWAS analyses reveal QTL in egg layers that differ in response to diet differences. Genet. Sel. Evol. [Internet]. BioMed Central; 2015;47:83. Available from: http://www.gsejournal.org/content/47/1/83

41. Lu D, Akanno EC, Crowley JJ, Schenkel F, Li H, De Pauw M, et al. Accuracy of genomic predictions for feed efficiency traits of beef cattle using 50K and imputed HD genotypes. J. Anim. Sci. 2016;94:1342–53.

42. Li G, Li D, Yang N, Qu L, Hou Z, Zheng J, et al. A genome-wide association study identifies novel single nucleotide polymorphisms associated with dermal shank pigmentation in chickens. Poult. Sci. [Internet]. 2014;93:2983–7. Available from:

bioRxiv preprint doi: https://doi.org/10.1101/112979; this version posted March 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

http://ps.oxfordjournals.org/content/93/12/2983%5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/25260525

43. Corbin LJ, Kranis A, Blott SC, Swinburne JE, Vaudin M, Bishop SC, et al. The utility of low-density genotyping for imputation in the Thoroughbred horse. Genet. Sel. Evol. [Internet]. BioMed Central; 2014 [cited 2016 Sep 7];46:9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24495673

44. Frischknecht M, Neuditschko M, Jagannathan V, Drögemüller C, Tetens J, Thaller G, et al. Imputation of sequence level genotypes in the Franches-Montagnes horse breed. Genet. Sel. Evol. 2014;46:63.

45. Lindgreen S. AdapterRemoval: easy cleaning of next-generation sequencing reads. BMC Res. Notes [Internet]. 2012 [cited 2015 Sep 4];5:337. Available from: http://www.biomedcentral.com/1756-0500/5/337

46. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics [Internet]. 2009 [cited 2014 Jul 9];25:1754–60. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2705234&tool=pmcentrez&rendertype=abstract

47. Zerbino DR, Birney E. Velvet : Algorithms for de novo short read assembly using de Bruijn graphs. 2008;821–9.

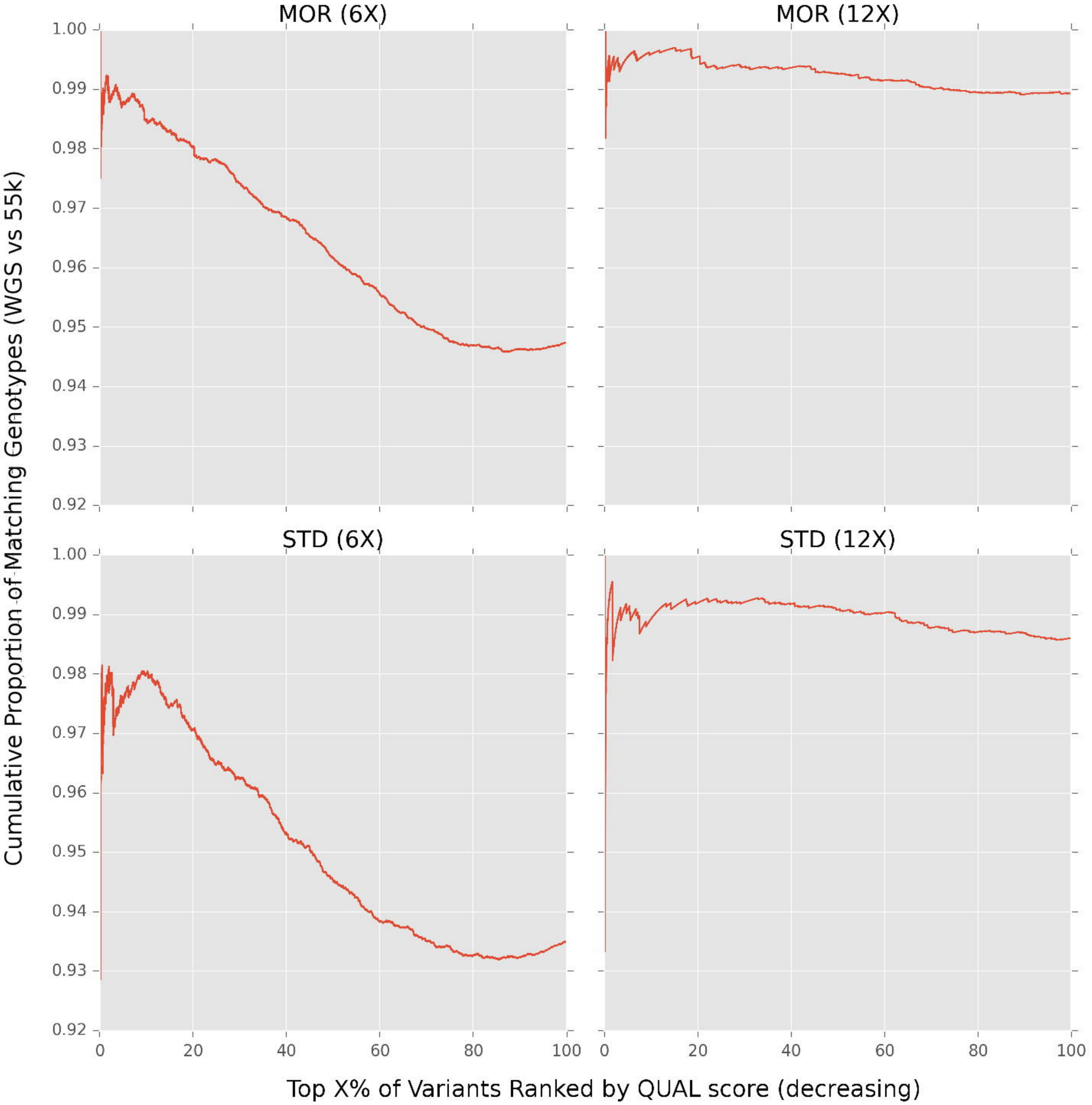
48. Smit A, Hubley R, Green P. RepeatMasker Open-4.0 [Internet]. Available from: http://repeatmasker.org

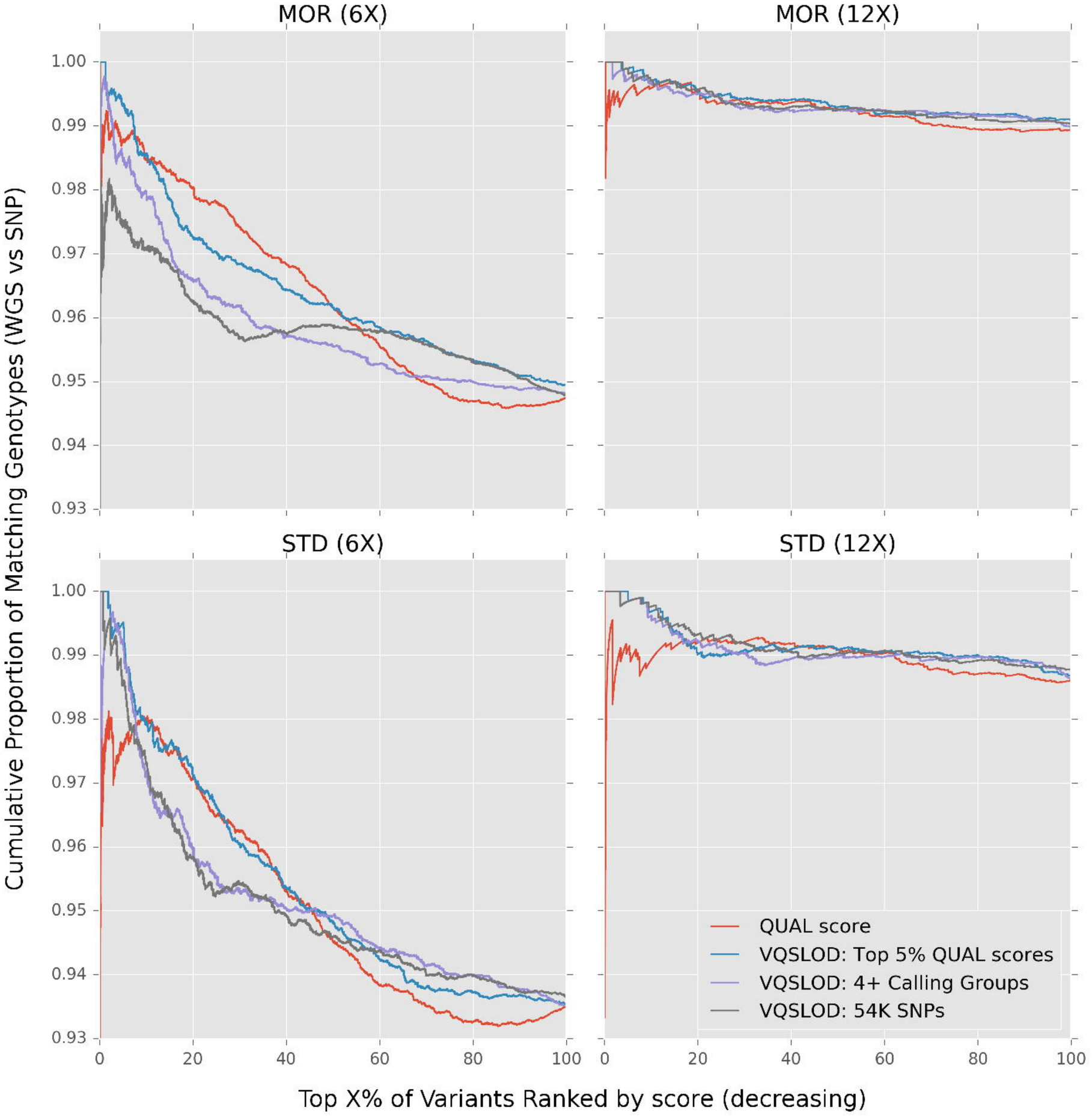
49. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am. J. Hum. Genet. 2007;81:1084–97.

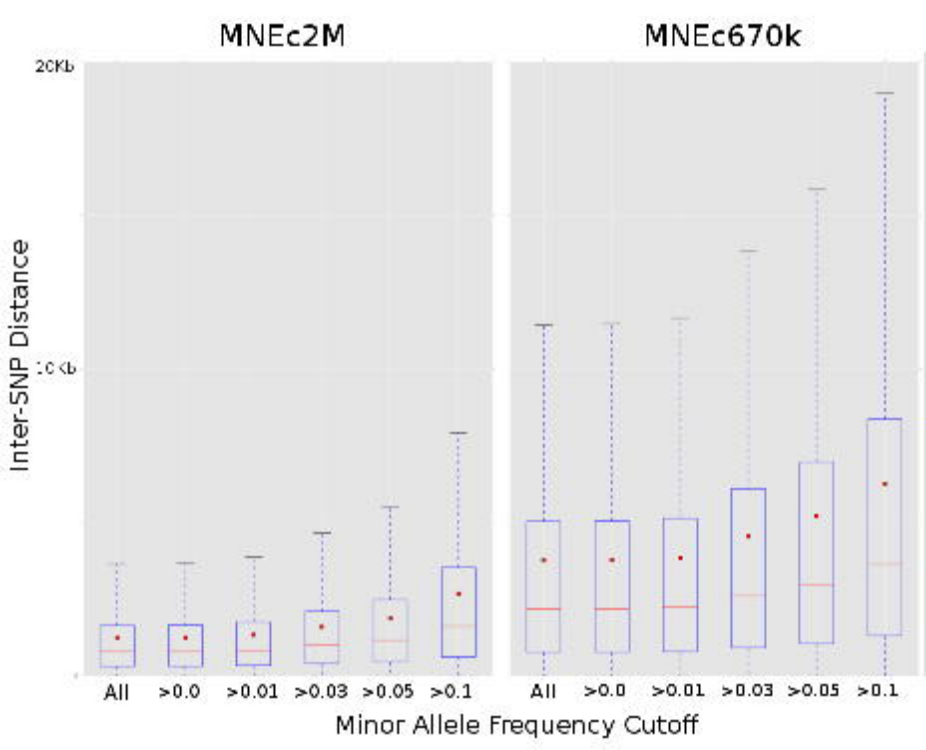
50. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format

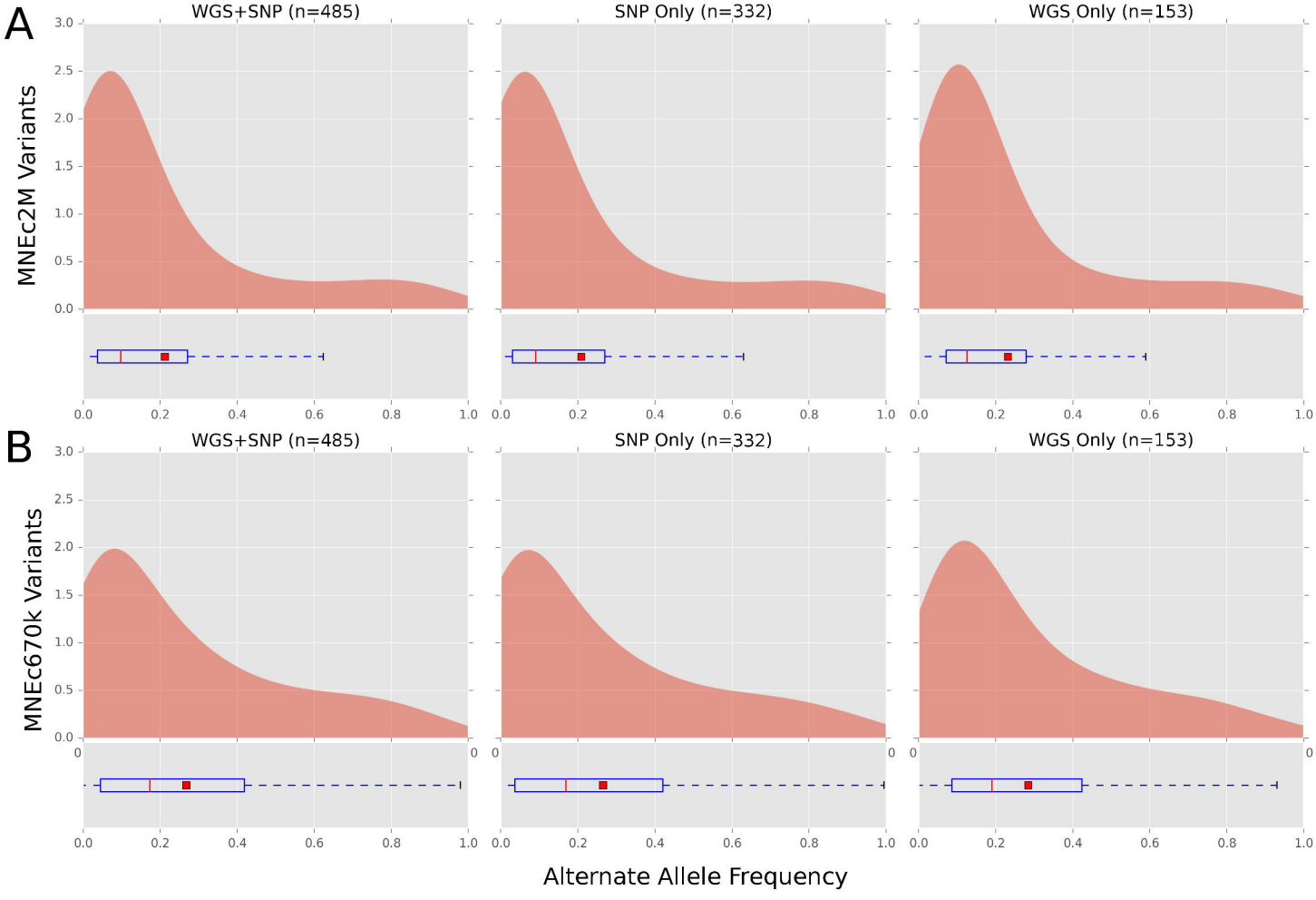
bioRxiv preprint doi: https://doi.org/10.1101/112979; this version posted March 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

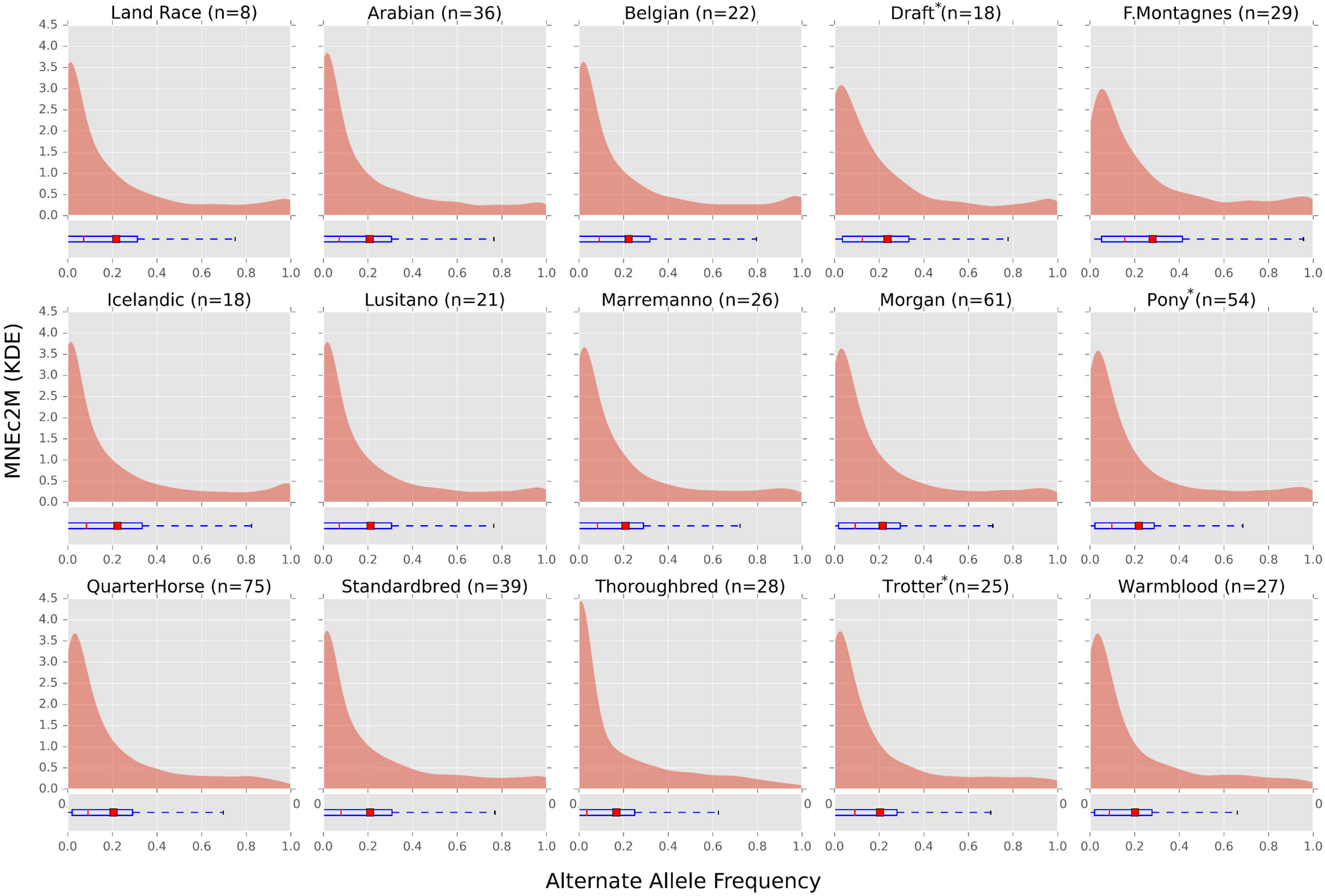
and VCFtools. Bioinformatics. 2011;27:2156–8.

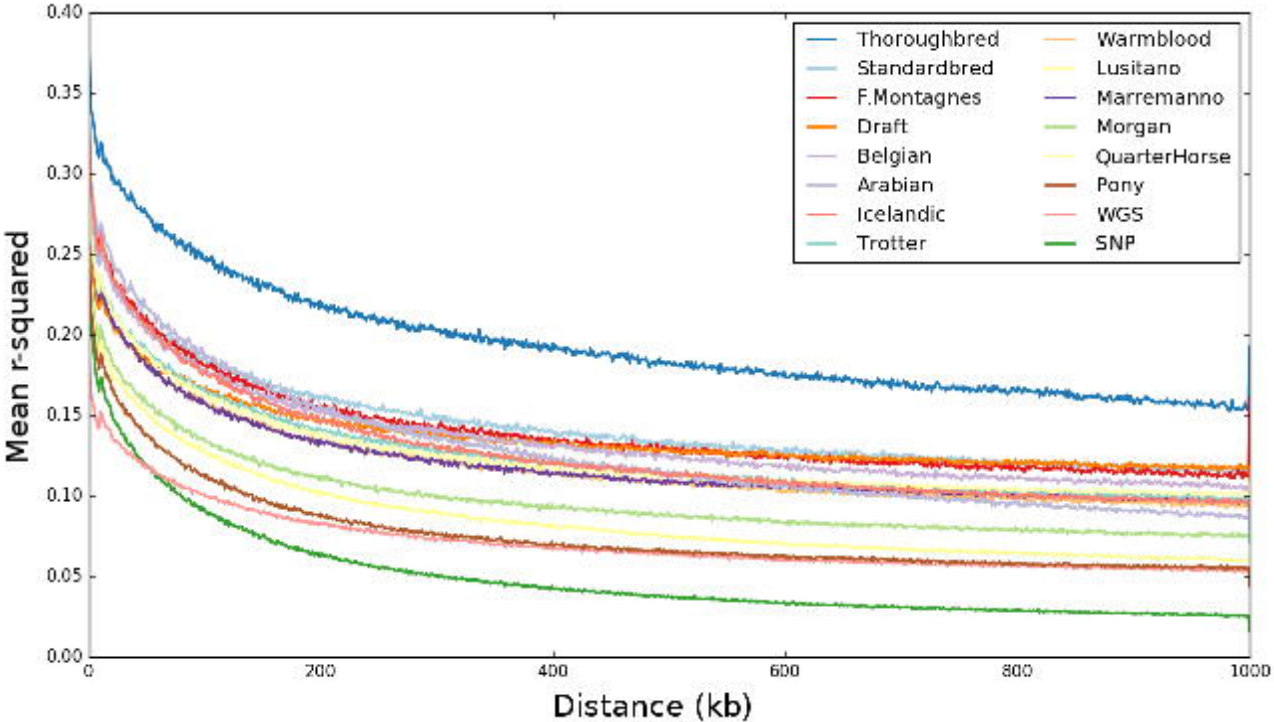

MOR (6X)
MOR (12X)
STD (6X)
STD (12X)
Cumulative Proportion of Matching Genotypes (WGS vs 55k)
Top X% of Variants Ranked by QUAL score (decreasing)
1.00
0.99
0.98
0.97
0.96
0.95
0.94
0.93
0.92
0
20
40
60
80
100

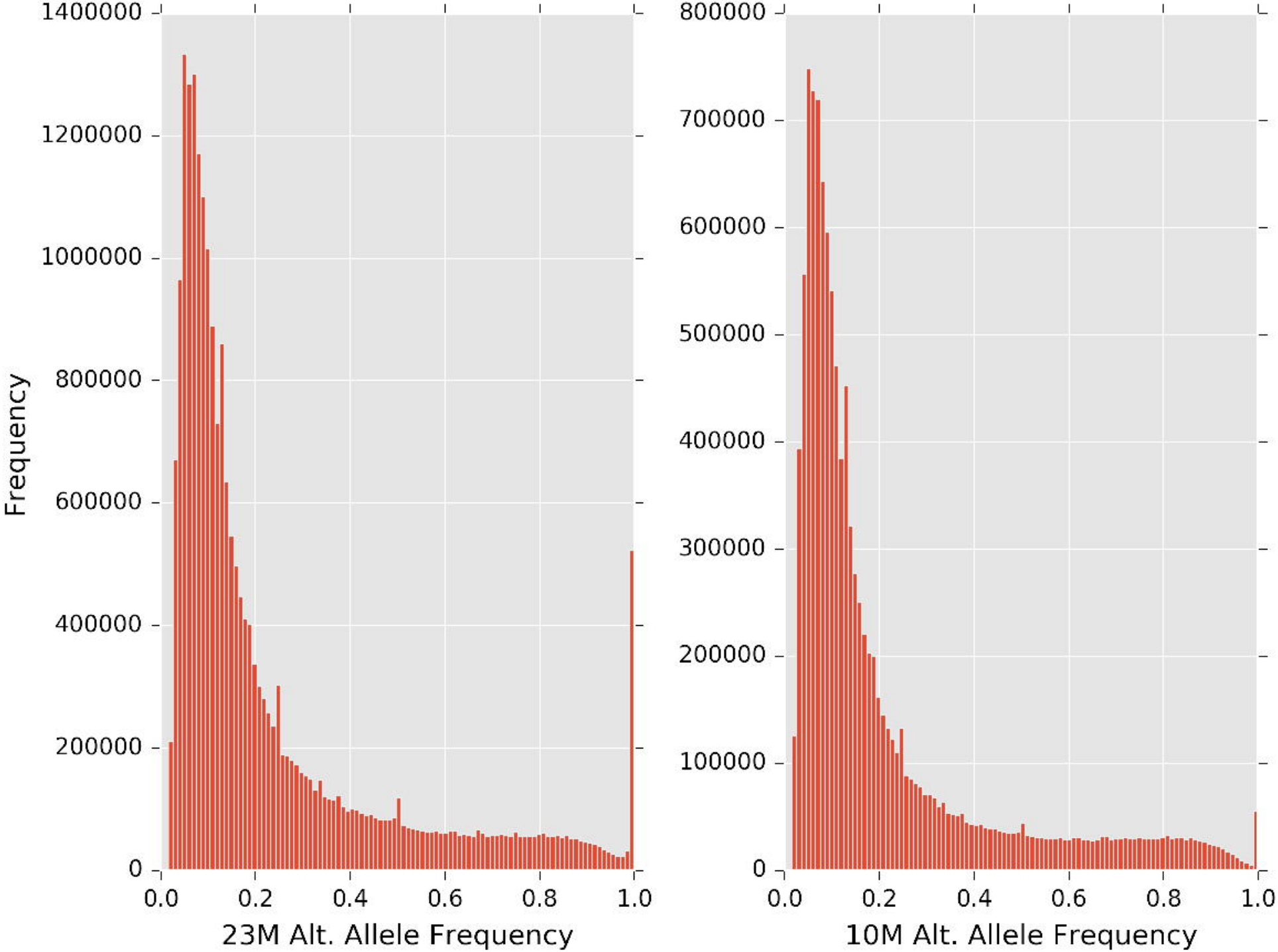

MOR (6X)
MOR (12X)
STD (6X)
STD (12X)
Cumulative Proportion of Matching Genotypes (WGS vs SNP)
1.00
0.99
0.98
0.97
0.96
0.95
0.94
0.93
0
20
40
60
80
100
Top X% of Variants Ranked by score (decreasing)
QUAL score
VQSLOD: Top 5% QUAL scores
VQSLOD: 4+ Calling Groups
VQSLOD: 54K SNPs

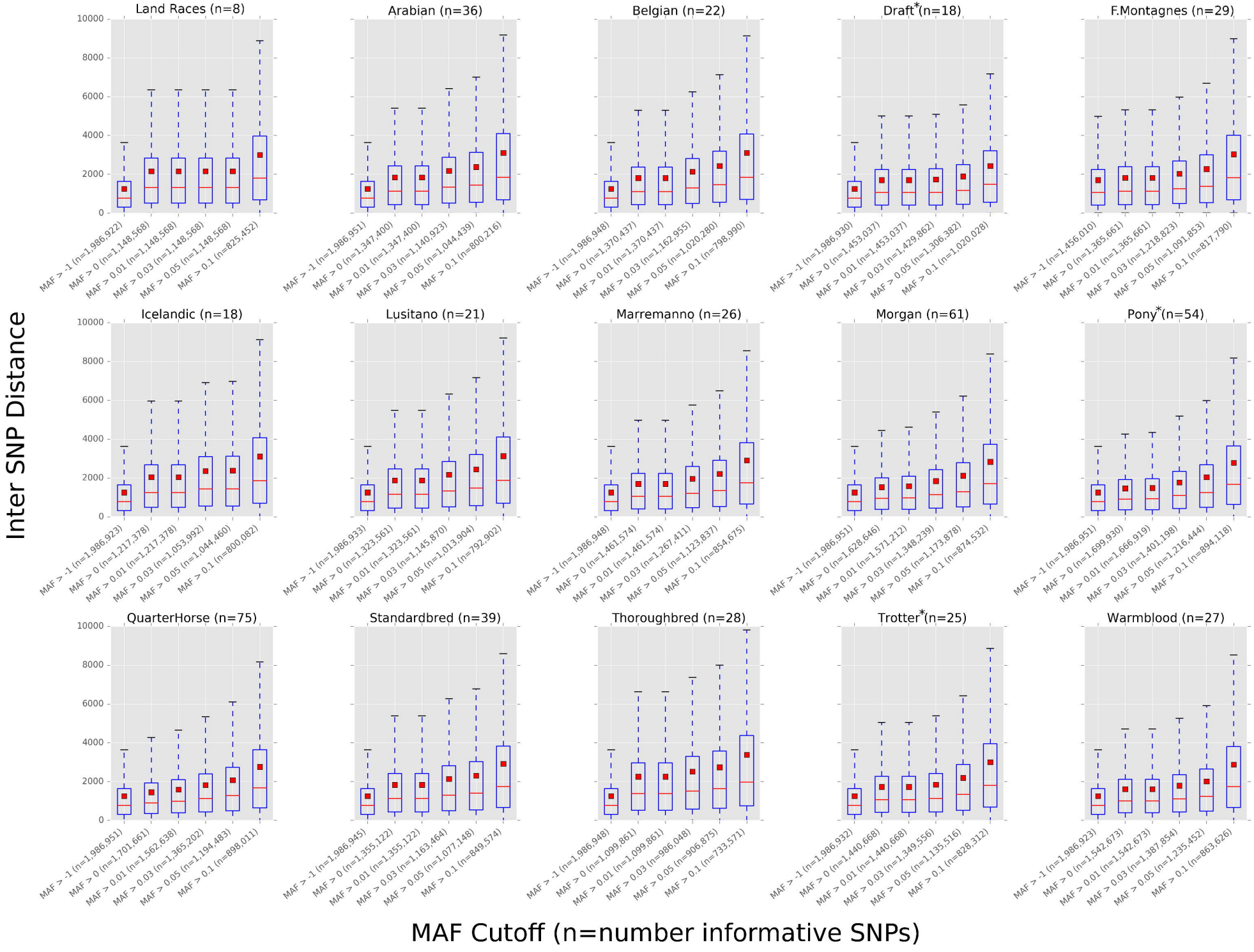

MNEc2M
MNEc670k
20Kb
10Kb
Inter-SNP Distance
All
>0.0
>0.01
>0.03
>0.05
>0.1
All
>0.0
>0.01
>0.03
>0.05
>0.1
Minor Allele Frequency Cutoff

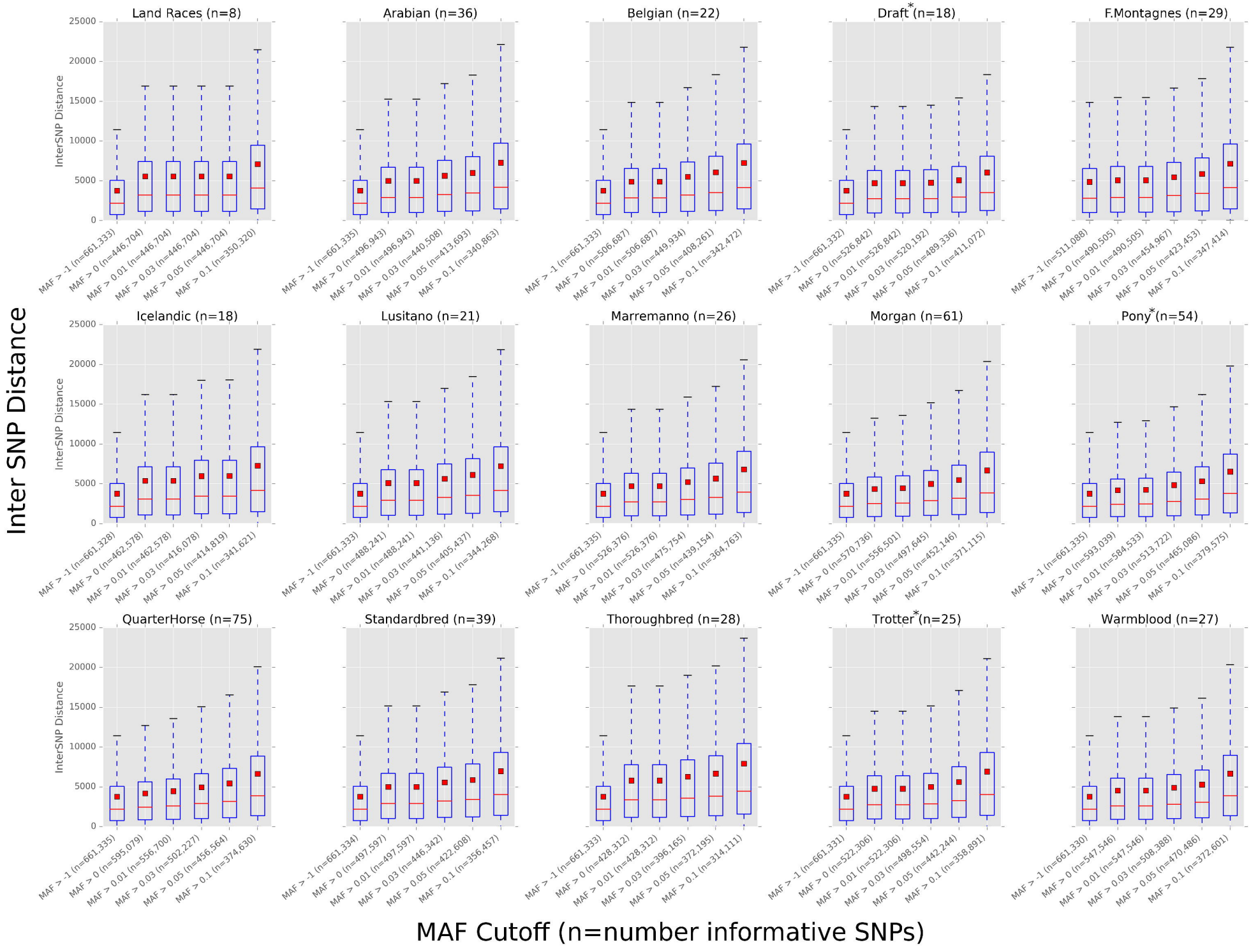

A
WGS+SNP (n=485)
SNP Only (n=332)
WGS Only (n=153)
MNEc2M Variants
B
WGS+SNP (n=485)
SNP Only (n=332)
WGS Only (n=153)
MNEc670k Variants
Alternate Allele Frequency

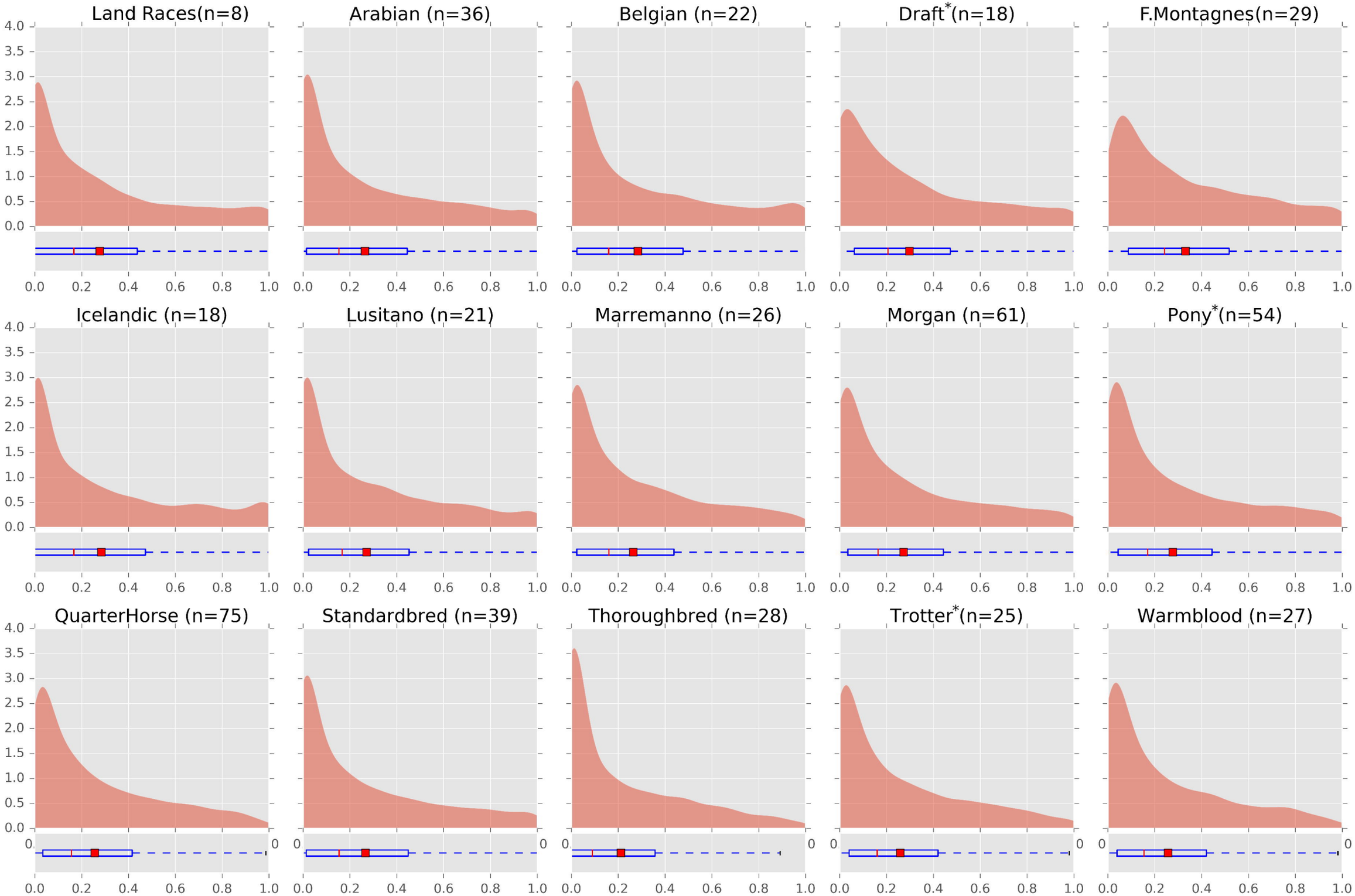

Land Race (n=8)
Arabian (n=36)
Belgian (n=22)
Draft*(n=18)
F.Montagnes (n=29)
Icelandic (n=18)
Lusitano (n=21)
Marremanno (n=26)
Morgan (n=61)
Pony*(n=54)
QuarterHorse (n=75)
Standardbred (n=39)
Thoroughbred (n=28)
Trotter*(n=25)
Warmblood (n=27)
MNEc2M (KDE)
Alternate Allele Frequency

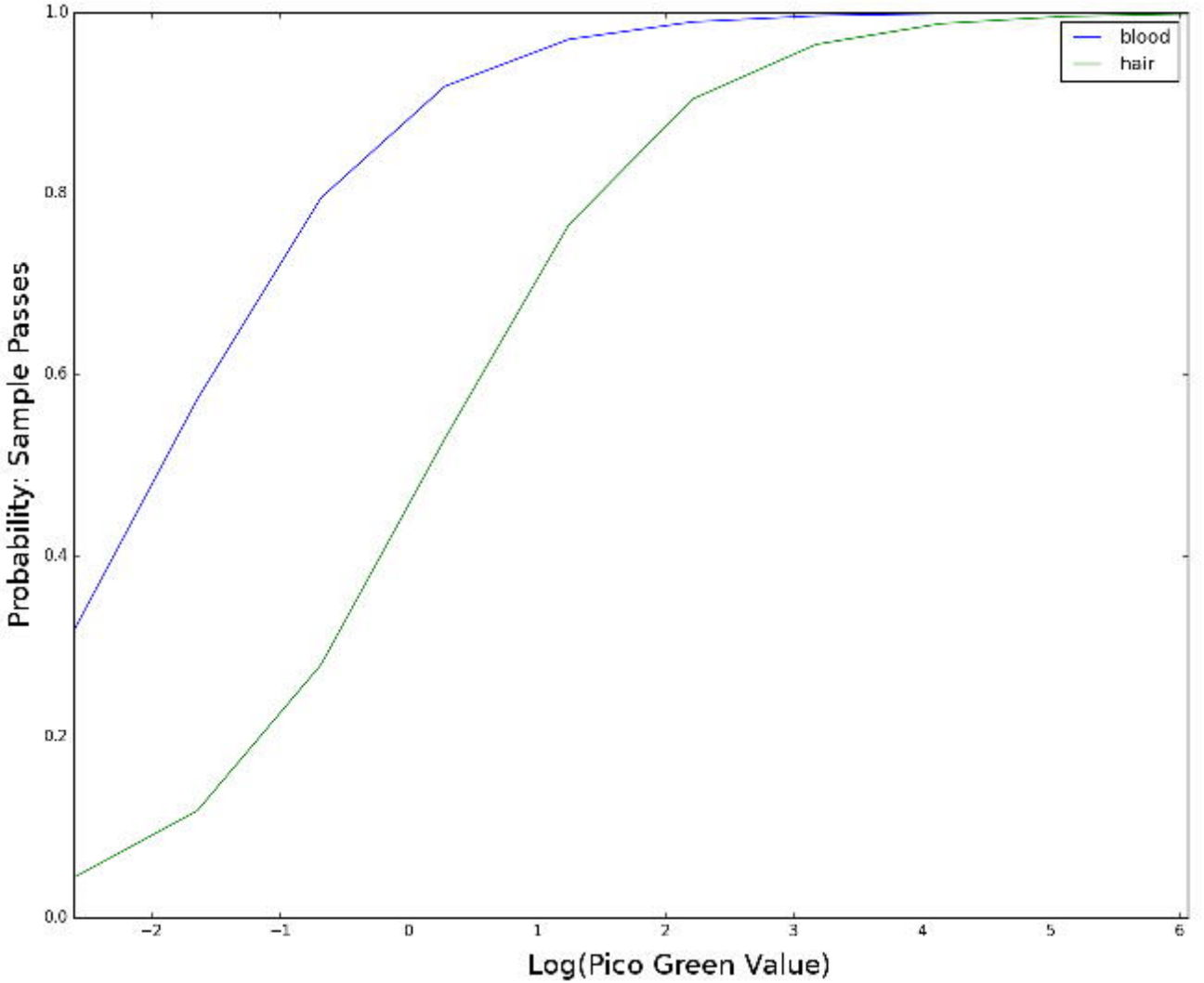

Thoroughbred
Standardbred
F.Montagnes
Draft
Belgian
Arabian
Icelandic
Trotter
Warmblood
Lusitano
Marremanno
Morgan
QuarterHorse
Pony
WGS
SNP
Mean r-squared
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0
200
400
600
800
1000
Distance (kb)


Frequency
1400000
1200000
1000000
800000
600000
400000
200000
0
0.0
0.2
0.4
0.6
0.8
1.0
23M Alt. Allele Frequency
800000
700000
600000
500000
400000
300000
200000
100000
0
0.0
0.2
0.4
0.6
0.8
1.0
10M Alt. Allele Frequency


Land Races (n=8)
MAF > -1 (n=1,986,922)
MAF > 0 (n=1,148,568)
MAF > 0.01 (n=1,148,568)
MAF > 0.03 (n=1,148,568)
MAF > 0.05 (n=1,148,568)
MAF > 0.1 (n=825,452)
Arabian (n=36)
MAF > -1 (n=1,986,951)
MAF > 0 (n=1,347,400)
MAF > 0.01 (n=1,347,400)
MAF > 0.03 (n=1,140,923)
MAF > 0.05 (n=1,044,439)
MAF > 0.1 (n=800,216)
Belgian (n=22)
MAF > -1 (n=1,986,948)
MAF > 0 (n=1,370,437)
MAF > 0.01 (n=1,370,437)
MAF > 0.03 (n=1,162,955)
MAF > 0.05 (n=1,020,280)
MAF > 0.1 (n=798,990)
Draft*(n=18)
MAF > -1 (n=1,986,930)
MAF > 0 (n=1,453,037)
MAF > 0.01 (n=1,453,037)
MAF > 0.03 (n=1,429,862)
MAF > 0.05 (n=1,306,382)
MAF > 0.1 (n=1,020,028)
F.Montagnes (n=29)
MAF > -1 (n=1,456,010)
MAF > 0 (n=1,365,661)
MAF > 0.01 (n=1,365,661)
MAF > 0.03 (n=1,218,823)
MAF > 0.05 (n=1,091,853)
MAF > 0.1 (n=817,790)
Icelandic (n=18)
MAF > -1 (n=1,986,923)
MAF > 0 (n=1,217,378)
MAF > 0.01 (n=1,217,378)
MAF > 0.03 (n=1,053,992)
MAF > 0.05 (n=1,044,460)
MAF > 0.1 (n=800,082)
Lusitano (n=21)
MAF > -1 (n=1,986,933)
MAF > 0 (n=1,323,561)
MAF > 0.01 (n=1,323,561)
MAF > 0.03 (n=1,145,870)
MAF > 0.05 (n=1,013,904)
MAF > 0.1 (n=792,902)
Marremanno (n=26)
MAF > -1 (n=1,986,948)
MAF > 0 (n=1,461,574)
MAF > 0.01 (n=1,461,574)
MAF > 0.03 (n=1,267,411)
MAF > 0.05 (n=1,123,837)
MAF > 0.1 (n=854,675)
Morgan (n=61)
MAF > -1 (n=1,986,951)
MAF > 0 (n=1,628,646)
MAF > 0.01 (n=1,571,212)
MAF > 0.03 (n=1,348,239)
MAF > 0.05 (n=1,173,878)
MAF > 0.1 (n=874,532)
Pony*(n=54)
MAF > -1 (n=1,986,951)
MAF > 0 (n=1,699,930)
MAF > 0.01 (n=1,666,919)
MAF > 0.03 (n=1,401,198)
MAF > 0.05 (n=1,216,444)
MAF > 0.1 (n=894,118)
QuarterHorse (n=75)
MAF > -1 (n=1,986,951)
MAF > 0 (n=1,701,661)
MAF > 0.01 (n=1,562,638)
MAF > 0.03 (n=1,365,202)
MAF > 0.05 (n=1,194,483)
MAF > 0.1 (n=898,011)
Standardbred (n=39)
MAF > -1 (n=1,986,945)
MAF > 0 (n=1,355,122)
MAF > 0.01 (n=1,355,122)
MAF > 0.03 (n=1,163,464)
MAF > 0.05 (n=1,077,148)
MAF > 0.1 (n=849,574)
Thoroughbred (n=28)
MAF > -1 (n=1,986,948)
MAF > 0 (n=1,099,861)
MAF > 0.01 (n=1,099,861)
MAF > 0.03 (n=986,048)
MAF > 0.05 (n=906,875)
MAF > 0.1 (n=733,571)
Trotter*(n=25)
MAF > -1 (n=1,986,932)
MAF > 0 (n=1,440,668)
MAF > 0.01 (n=1,440,668)
MAF > 0.03 (n=1,349,556)
MAF > 0.05 (n=1,135,516)
MAF > 0.1 (n=828,312)
Warmblood (n=27)
MAF > -1 (n=1,986,923)
MAF > 0 (n=1,542,673)
MAF > 0.01 (n=1,542,673)
MAF > 0.03 (n=1,387,854)
MAF > 0.05 (n=1,235,452)
MAF > 0.1 (n=863,626)
Inter SNP Distance
MAF Cutoff (n=number informative SNPs)


Land Races (n=8)
Arabian (n=36)
Belgian (n=22)
Draft* (n=18)
F.Montagnes (n=29)
Icelandic (n=18)
Lusitano (n=21)
Marremanno (n=26)
Morgan (n=61)
Pony* (n=54)
QuarterHorse (n=75)
Standardbred (n=39)
Thoroughbred (n=28)
Trotter* (n=25)
Warmblood (n=27)
InterSNP Distance
0
5000
10000
15000
20000
25000
MAF > -1 (n=661,333)
MAF > 0 (n=446,704)
MAF > 0.01 (n=446,704)
MAF > 0.03 (n=446,704)
MAF > 0.05 (n=446,704)
MAF > 0.1 (n=350,320)
MAF > -1 (n=661,335)
MAF > 0 (n=496,943)
MAF > 0.01 (n=496,943)
MAF > 0.03 (n=440,508)
MAF > 0.05 (n=413,693)
MAF > 0.1 (n=340,863)
MAF > -1 (n=661,333)
MAF > 0 (n=506,687)
MAF > 0.01 (n=506,687)
MAF > 0.03 (n=449,934)
MAF > 0.05 (n=408,261)
MAF > 0.1 (n=342,472)
MAF > -1 (n=661,332)
MAF > 0 (n=526,842)
MAF > 0.01 (n=526,842)
MAF > 0.03 (n=520,192)
MAF > 0.05 (n=489,336)
MAF > 0.1 (n=411,072)
MAF > -1 (n=511,088)
MAF > 0 (n=490,505)
MAF > 0.01 (n=490,505)
MAF > 0.03 (n=454,967)
MAF > 0.05 (n=423,453)
MAF > 0.1 (n=347,414)
MAF > -1 (n=661,328)
MAF > 0 (n=462,578)
MAF > 0.01 (n=462,578)
MAF > 0.03 (n=416,078)
MAF > 0.05 (n=414,819)
MAF > 0.1 (n=341,621)
MAF > -1 (n=661,333)
MAF > 0 (n=488,241)
MAF > 0.01 (n=488,241)
MAF > 0.03 (n=441,136)
MAF > 0.05 (n=405,437)
MAF > 0.1 (n=344,268)
MAF > -1 (n=661,335)
MAF > 0 (n=526,376)
MAF > 0.01 (n=526,376)
MAF > 0.03 (n=475,754)
MAF > 0.05 (n=439,154)
MAF > 0.1 (n=364,763)
MAF > -1 (n=661,335)
MAF > 0 (n=570,736)
MAF > 0.01 (n=556,501)
MAF > 0.03 (n=497,645)
MAF > 0.05 (n=452,146)
MAF > 0.1 (n=371,115)
MAF > -1 (n=661,335)
MAF > 0 (n=593,039)
MAF > 0.01 (n=584,533)
MAF > 0.03 (n=513,722)
MAF > 0.05 (n=465,086)
MAF > 0.1 (n=379,575)
MAF > -1 (n=661,335)
MAF > 0 (n=595,079)
MAF > 0.01 (n=556,700)
MAF > 0.03 (n=502,227)
MAF > 0.05 (n=456,564)
MAF > 0.1 (n=374,630)
MAF > -1 (n=661,334)
MAF > 0 (n=497,597)
MAF > 0.01 (n=497,597)
MAF > 0.03 (n=446,342)
MAF > 0.05 (n=422,608)
MAF > 0.1 (n=356,457)
MAF > -1 (n=661,333)
MAF > 0 (n=428,312)
MAF > 0.01 (n=428,312)
MAF > 0.03 (n=396,165)
MAF > 0.05 (n=372,195)
MAF > 0.1 (n=314,111)
MAF > -1 (n=661,331)
MAF > 0 (n=522,306)
MAF > 0.01 (n=522,306)
MAF > 0.03 (n=498,554)
MAF > 0.05 (n=442,244)
MAF > 0.1 (n=358,891)
MAF > -1 (n=661,330)
MAF > 0 (n=547,546)
MAF > 0.01 (n=547,546)
MAF > 0.03 (n=508,388)
MAF > 0.05 (n=470,486)
MAF > 0.1 (n=372,601)
Inter SNP Distance
MAF Cutoff (n=number informative SNPs)


Land Races(n=8)
Arabian (n=36)
Belgian (n=22)
Draft*(n=18)
F.Montagnes(n=29)
Icelandic (n=18)
Lusitano (n=21)
Marremanno (n=26)
Morgan (n=61)
Pony*(n=54)
QuarterHorse (n=75)
Standardbred (n=39)
Thoroughbred (n=28)
Trotter*(n=25)
Warmblood (n=27)
0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0
0.0 0.2 0.4 0.6 0.8 1.0


blood
hair
Probability: Sample Passes
Log(Pico Green Value)
1.0
0.8
0.6
0.4
0.2
0.0
−2
−1
0
1
2
3
4
5
6