Appendix S1

Supplemental Text

Background

Two genotyping arrays have been designed for the horse. The first (Illumina Equine SNP50) was designed with 60,000 markers, of which 54,602 successfully genotyped in the commercial array. The second array (Illumina Equine SNP70) was designed to increase genome-wide marker density and to fill in coverage gaps identified in the SNP50 chip. Although all of the successful markers from the SNP50 were incorporated into the 74,000 markers designed for the SNP70 chip (65,157 successful), only 45,703 are actually genotyped with this platform. Thus, while 19,454 markers were added in the SNP70 array when compared to the SNP50, 8,899 markers remain unique to the older platform (**Figure S3**). To maximize the number of markers available for analysis when combining data obtained from the two arrays, imputation must therefore be carried out in "both directions" (i.e. from the SNP50 to the SNP70 and also from the SNP70 to the SNP50)

Methods

Data were retrieved from 248 horses of three breeds (Quarter Horse [QH], *n* = 143; Standardbred [STB], *n* = 72; Thoroughbred [TB], *n* = 33) genotyped on the Illumina Equine SNP70 beadchip (65,157 markers). These data were "masked" down to the list of 45,703 markers shared by the SNP70 and SNP50 chips, and subsequently imputed back to the complete marker set using BEAGLE (Browning & Browning, 2007). The imputed genotypes were then compared to the known genotypes at each location to determine imputation accuracy. Genotypes that were missing in the original data were excluded from analysis. On the basis of preliminary data (**Table S2**), five chromosomes were chosen for validation of imputation accuracy: equine chromosomes (ECA) 1, 6, 15, 26, and X. These were considered representative of all of the chromosomes as they reflected a range of both imputation success and chromosome size. Between 29% and 35% of the markers on each chromosome were imputed. Twenty scenarios were constructed, varying the imputed population size (range 5-30 individuals), imputed population breed (QH, STB, or TB), reference population size (range 20-100), and/or reference population make-up (breed-matched to the imputed population, or made up of an equal mix of all three breeds ["mixed" population]) (**Table S1**).

To confirm imputation accuracy for the 8,899 markers that are present on the SNP50 and not the SNP70 array, genotype masking and subsequent imputation was carried out for five chromosomes, as above, in thirty QH genotyped on the Illumina Equine SNP50 beadchip (54,602 markers). Between 11% and 37% of the markers on each chromosome were imputed (**Table S3**). Based on the public availability of genotyping data from a large number of horses of diverse breeds (www.animalgenome.org/repository/pub/UMN2012.1130/ [McCue et al., 2012; Petersen et al., 2013]) and the success of a large mixed breed population in the SNP70 scenarios (above), imputation accuracy in this QH population was confirmed in a scenario using a reference population comprised of 280 horses of thirteen diverse breeds (Thoroughbred, *n* = 44; Andalusian, *n* = 19; Arabian, *n* = 23; Belgian, *n* = 22; Franches-Montagnes, *n* = 20; French Trotter, *n* = 17; Hanoverian, *n* = 19; Icelandic, *n* = 17; Mongolian, *n* = 21; Norwegian Fjord, *n* = 21; Saddlebred, *n* = 21; Standardbred, *n* = 19; Swiss Warmblood, *n* = 17).

BEAGLE requires three input files for each chromosome to be imputed: a genotypes file (.bgl) for both the test population and the reference population, and a marker file, which includes the marker name, chromosomal position, and list of possible alleles at each locus. The marker file was generated by modifying the allele frequency output (--freq) from PLINK (Purcell et al., 2007). The genotypes files were converted from PLINK .map/.ped format to .bgl format using the phasing pipeline utility associated with GERMLINE (Gusev et al., 2009). BEAGLE was implemented using the default settings for unphased unrelated data.

To maximize the impact of imputation in a real dataset, horses genotyped on each platform (SNP50 or SNP70) should be alternately used as the reference and imputed populations such that each individual has actual genotypes from one platform and imputed genotypes from the non-overlapping markers from the other platform. To complete the data analysis pipeline for these real data, BEAGLE phased output files are converted back to PLINK .ped format using custom shell script (available at https://github.com/schae234/Beagle2Ped) with the phasing pipeline utility (above). Accompanying .map files must then be generated from the ordered list of markers in the phased BEAGLE output for the imputed population. Converted imputed files are subsequently merged with the original genotype data using PLINK (--merge). Merged imputed files can then be utilized for any number of analyses. The complete pipeline is illustrated in **Figure S1**.

SNP50 imputation results and comments

Results for imputation for the SNP50 chip are presented in **Table S3**. The average imputation success across all chromosomes was 94.2% (range for individual horses 87.4%-98.8%). When compared to results for the same size imputed population (*n* = 30) in the SNP70 scenarios, imputation success in this SNP50 scenario was somewhat lower for ECA1, higher for ECA6, and about the same for ECA 15, 26 and X. The mean R^2 across all chromosomes, reflecting confidence in the imputed genotype calls, was 0.725 (range 0.680-0.795). This is lower than was found in the SNP70 scenario with a large mixed breed population (mean 0.76). However, in that scenario, one-third of the horses in the reference population were of the same breed as the imputed population (QH), while in the SNP50 scenario, there were no QH in the reference population. This supports findings reported in the main text that a reference population that is breed-matched to the imputed population gives better results than a mixed reference population. Although the results cannot truly be directly compared because they looked at performance of imputation in different arrays, it is of note that nearly tripling the size of the reference population (from 100 to 280 individuals) did not result in a marked increase in imputation success. This reflects findings reported in human data, in which increasing reference population sizes over a threshold gave diminishing returns for improvement in imputation, except for very low frequency polymorphisms (Howie et al., 2011; Li et al., 2011).

References

Browning S.R., & Browning, G. L. (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *American Journal of Human Genetics* 81, 1084-1097.

Gusev A., Lowe J.K., Stoffel M., Daly M.J., Altshuler D., Breslow J.L., Friedman J.M. Pe'er I. (2009) Whole population, genome-wide mapping of hidden relatedness. *Genome Research* 19, 318-326.

Howie B., Marchini J., Stephens M. (2011) Genotype imputation with thousands of genomes. *Genes Genomes Genetics* 1, 457-470.

Li L., Li Y., Browning S.R., Browning B.L. Slater A.J., Kong X., Aponte J.L., Mooser V.E., Chissoe S.L., Whittaker J.C., Nelson M.R., Ehm M.G. (2011) Performance of genotype imputation for rare variants identified in exons and flanking regions of genes. *PLoS ONE* 6:e24945. doi:10.1371/journal.pone.0024945.

McCue M.E., Bannasch D.L., Petersen J.L., Bailey E., Binns M.M., Distl O., Guérin G., Hasegawa T., Hill E.W., Leeb T., Lindgren G., Penedo M.C.T., Røed K.H., Ryder O.A., Swinburne J.E., Tozaki T., Valberg S.J., Vaudin M., Lindblad-Toh K., Wade C.M., Mickelson J.R. (2012) A high density SNP array for the domestic horse and extant Perissodactyla: utility for association mapping, genetic diversity and phylogeny studies. *PLoS Genetics* 8:e1002451. doi:10.1371/journal.pgen.1002451.

Petersen J.L., Mickelson J.R., Cothran E.G., Andersson L.S., Axelsson J., Bailey E., Bannasch D., Binns M.M., Borges A.S., Brama P., de Câmara Machado A., Distl O., Felicetti M., Fox-Clipsham L., Graves K.T., Guérin G., Haase B., Hasegawa T., Hemmann K., Hill E.W., Leeb T., Lindgren G., Lohi H., Lopes M.S., McGivney B.A., Mikko S., Orr N., Penedo M.C.T., Piercy R.J., Raekallio M., Rieder S., Røed K.H., Silvestrelli M., Swinburne J., Tozaki T., Vaudin M., Wade C.M., McCue M.E. (2013) Genetic diversity in the modern horse illustrated from genome-wide SNP data. *PLoS One* 8:e54997. doi:10.1371/journal.pone.0054997.

Purcell S., Neale B., Todd-Brown K., Thomas, L., Ferreira M.A.R., Bender D., Maller J., Sklar P., de Bakker P.I.W., Daly M.J., Sham P.C. (2007) PLINK: a tool set for whole-genome association and population-based linkage analysis. *American Journal of Human Genetics* 81:559-575.