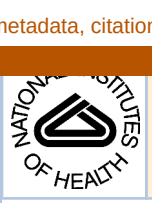
View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Carolina Digital Repository



NIH Public Access
Author Manuscript
Mamm Genome. Author manuscript; available in PMC 2009 August 12.

Published in final edited form as:
Mamm Genome. 2008 March ; 19(3): 199–208. doi:10.1007/s00335-008-9098-9.

An imputed genotype resource for the laboratory mouse

Jin P. Szatkiewicz[1], **Glen L. Beane**[1], **Yueming Ding**[1], **Lucie Hutchins**[1], **Fernando Pardo-Manuel de Villena**[2], and **Gary A. Churchill**[1]

[1]The Jackson Laboratory, Bar Harbor, Maine 04609, USA

[2]Department of Genetics, Carolina Center for Genome Sciences and Lineberger Comprehensive Cancer Center, University of North Carolina, Chapel Hill, Chapel Hill, North Carolina 27599, USA

Abstract

We have created a high-density SNP resource encompassing 7.87 million polymorphic loci across 49 inbred mouse strains of the laboratory mouse by combining data available from public databases and training a hidden Markov model to impute missing genotypes in the combined data. The strong linkage disequilibrium found in dense sets of SNP markers in the laboratory mouse provides the basis for accurate imputation. Using genotypes from eight independent SNP resources, we empirically validated the quality of the imputed genotypes and demonstrate that they are highly reliable for most inbred strains. The imputed SNP resource will be useful for studies of natural variation and complex traits. It will facilitate association study designs by providing high density SNP genotypes for large numbers of mouse strains. We anticipate that this resource will continue to evolve as new genotype data become available for laboratory mouse strains. The data are available for bulk download or query at http://cgd.jax.org/.

Keywords

mouse; SNP; hidden Markov model; missing data

INTRODUCTION

The laboratory mouse owes much of its popularity as a model organism in biomedical research to the existence of a large collection of inbred strains that represent an immortal population of genetic clones derived by repeated brother sister mating (Lyon et al. 1996). Because mice from each strain are genetically identical it is possible to collect and combine biological data over time and space leading to a depth of phenotype characterization rarely achieved in other mammalian systems (Bogue 2003). Furthermore, the existence of a definite set of genetic

Corresponding author: Gary A. Churchill, The Jackson Laboratory, Bar Harbor, Maine 04609, USA, **Email:**gary.churchill@jax.org, Phone: 207-288-6189, Fax: 207-288-6847.

WEB SITE REFERENCES

http://www.broad.mit.edu/~claire/MouseHapMap/
http://mouse.perlegen.com/
http://www.well.ox.ac.uk/mouse/INBREDS/
http://www.sanger.ac.uk/modelorgs/mouse.shtml/
http://www.ensemble.org/
http://www.ncbi.nlm.nih.gov/SNP/
http://snp.gnf.org/
http://phenome.jax.org/
http://stt.gsc.riken.jp/msm/
http://mousesnp.roche.com/
http://cgd.jax.org/

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

differences among inbred strains allows scientists to explore the effect of genetic diversity on almost any phenotype of interest (Wade and Daly 2005). These studies require an accurate description of the level and distribution of genetic variation present among the hundreds of existing inbred strains. This is a challenging problem because the diversity between strains varies from extremely low levels found among sister substrains to very high levels found among strains derived from different species and subspecies (Petkov et al. 2004; Ideraabdullah et al. 2004; Yang et al. 2007).

Inbred strains can be classified into classical and wild-derived strains according to whether they were derived in the 20th century from a small set of founders known as “fancy” mice or derived from mice captured from natural populations more recently. Common wild-derived strains include representatives from two species *Mus spretus* and *M. musculus*, thought to have diverged almost 2 million years ago (Guenet and Bonhomme 2003). There is well over 1% sequence divergence between these species resulting in one SNP every 75bp (Ideraabdullah et al. 2004). Within the *M. musculus* species there are four subspecies, *M. m. domesticus, M. m. castaneus. M. m. musculus* and *M. m. molossinus* from which inbred strains have been derived. These subspecies are thought to have diverged 750,000 years ago (Guenet and Bonhomme 2003). There is roughly 1% divergence among subspecies corresponding to one SNP every 150bp (Ideraabdullah et al. 2004). Recent analysis of high density genotype data demonstrates that many wild-derived strain genomes carry regions of intersubspecific introgression (“contamination”) from a different subspecies (Yang et al. 2007). This analysis also confirms that classical strains are derived from multiple subspecies but that the contribution of *M. m. domesticus* represents over 90% of the genome in most of these strains. These data are critical to interpret the results of any mouse experiment in the proper evolutionary context and may have profound implications for our understanding of basic biological processes such as divergence, selection and speciation (Payseur and Hoekstra 2005; Mott 2007).

Since the genomic sequence of the C57BL/6 strain was reported (Waterston et al. 2002) much effort has been focused on the discovery and characterization of single nucleotide polymorphisms (SNPs) in inbred strains (Wade et al. 2002; Wiltshire et al. 2003; Yalcin et al. 2004; Pletcher et al. 2004; Frazer et al. 2004). Early SNP discovery projects carried out resequencing in a limited number of classical strains (Mural et al. 2002). More recently, the NIEHS used a hybridization based strategy to discover ~8.3 million SNPs in a survey of 15 inbred mouse strains, including four wild-derived strains representing the major subspecies of *M. musculus* (Frazer et al. 2007). In parallel to these SNP discovery efforts the Broad Institute of Harvard and MIT carried out genotyping of ~138,000 known SNPs on 49 inbred mouse strains (Wade and Daly 2005) and the Wellcome-CTC genotyped 499 inbred strains and outbred stocks at a lower SNP density with only ~13,370 SNPs (Shifman et al. 2006). Additional SNP resources are listed in Table 1.

In the follow we refer to the NIEHS data as high density (>7 million genotyped SNPs). Medium density genotypes are similar in magnitude to the Broad set (>100,000 genotyped SNPs) and low density genotypes are similar in magnitude to Wellcome-CTC set (>10,000 genotyped SNPs). Much of biomedical research involves inbred strains for which the description of the diversity is based on low to medium density SNP panels (Liao et al. 2004; Pletcher et al. 2004; Cervino et al. 2005; Shifman et al. 2006; McClurg et al. 2007; Payseur and Place 2007). Linkage disequilibrium (LD) among classical inbred strains is extensive (Wade et al. 2002; Petkov et al. 2005), suggesting that we could leverage the NIEHS data to impute genotypes at high density in a larger set of inbred mouse strains. Achieving this goal should immediately empower hundreds of laboratories to narrow quantitative trait loci, help design the next generation of experiments in mammalian genetics and provide invaluable support in the field of comparative and evolutionary genomics for the study of biological processes such

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

as recombination, mutation and selection (Dipetrillo et al. 2005; Siebert and Schadt 2007; Roberts et al. 2007).

We propose a method to impute genotypes at high density in strains for which only medium or low density genotype data are available. We apply this method to create a resource of SNP genotypes at ~7.9 million loci across 49 inbred strains by combining existing public databases and imputing missing genotypes. The quality of the imputed genotypes is quantified and empirically validated. We find that the imputed genotypes are most reliable for classical strains that have at least medium density genotyping data available. The accuracy of imputed genotypes is somewhat lower in wild-derived strains. We provide a confidence score that can be used to identify those imputed genotypes that are most reliable.

MATERIALS AND METHODS

Data preparation

Prior to combining databases, a multi-step quality control procedure was applied to the original NIEHS (http://mouse.perlegen.com/mouse/download.html, July 2006 release) and Broad (http://www.broad.mit.edu/~claire/MouseHapMap, February 2006 release) SNPs. First, we eliminated SNPs whose reported physical locations are impossible. We compared the genotypes and the 100-mer flanking sequences of all C57BL/6 SNPs in the data to the published mouse genome sequence NCBI build 36. This process remapped the NCBI build 33 Broad data to NCBI build 36 coordinates and removed those SNPs that revealed any discrepancy. A total of 75 Broad and 2,925 NIEHS SNPs were excluded. We then identified SNPs that are present in duplicated regions. We have previously observed that these SNPs can have very high false positive rates due to the detection of paralogous variation at other sites (Yang et al. 2007; unpublished data). We used BLAT (Kent 2002) under the most sensitive parameter settings to map all 25-mers centered in each SNP and defined a duplication as a SNP for which the 25-mer map to multiple genomic locations with less than three mismatches. We then used a sliding window to search for clustered duplications. Whenever two duplications were found out of four consecutive SNPs, the duplicated SNPs and any intervening SNPs were removed. A total of 448,999 SNPs (~5.4%) were removed. Of the remaining NIEHS SNPs, 15,068 were reported to have same genomic location. We kept one copy of each when all genotypes were fully consistent; otherwise, they were removed. For Broad data, we removed 2762 SNPs (~2%) that mapped to the duplicated locations in the NIEHS data. SNPs that mapped to identical genomic locations (redundant SNPs) were combined. During the process of combining databases, strand orientation adjustment was done whenever necessary. Conflicting genotypes were recoded as missing data. When more than half of the strains had discordant genotypes, the SNP locus was excluded.

Genotype Imputation

We use a hidden Markov model (HMM) with left to right architecture (Figure 1) to impute the missing genotypes. In this model, there are six hidden states (H = 6) representing different haplotypes at each SNP. State transitions proceed from one SNP (columns in figure 1) to the next according to a Markov process. The haplotypes of a strain can be viewed as a path through the model visiting one state per SNP locus, from the first SNP to the last on a given chromosome. Given a trained model and the genotypes of a strain, the path decoding problem is solved by Viterbi's algorithm (Viterbi 1967). In Figure 1, the Viterbi paths are shown as colored lines. Strains with identical path through this region are grouped, but in general each strain will have its own unique path through the haplotype states. States have a probabilistic output, representing the observed genotype. Missing genotypes are imputed as the allele that is most likely to be emitted by the states along the Viterbi path. The most probable genotype for each state is indicated in the Figure 1. For every genotype, imputed and experimental, the

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

posterior probability under the trained HMM serves as a confidence score, which is computed as the product of the posterior probability of the inferred haplotype state and genotype probability given the state.

Training the HMM involves estimation of a large number of free parameters. Parameter estimation is accomplished using the Expectation-Maximization (E-M) method as elaborated in Churchill (1989). The convergence criterion for the E-M algorithm is set as 10^{-6} change in the log-likelihood. Initial values for the EM algorithm are sampled at least 10 times and the training run that achieves the highest likelihood is chosen.

Despite the vast amount of data (millions of genotypes) in the training sets, this is a data poor problem. At any given SNP we have genotypes for only a small to moderate number of strains, distinct haplotypes may not be equally represented, and the information available in adjacent SNPs decays more or less rapidly depending on marker density and the extent of local linkage disequilibrium. The number of parameters in the HMM is large and it grows linearly with the number of SNPs; therefore the prior distributions can be influential and should be chosen carefully to obtain the best results.

Transition and emission probabilities are assumed to follow Dirichlet prior distributions. For state transitions, the prior density is biased towards the transitions between the same haplotype, with probability $1 - \lambda$, and is equally distributed among the other (H-1) haplotypes. A prior that favors small values of λ will encourage the use of more information from adjacent SNPs. Emission probabilities are assumed to follow a uniform prior distribution for the two possible alleles. The Dirichlet pseudocount method is used to combine prior information with maximum likelihood estimates (Durbin et al. 1998).

A series of computational experiments was carried out to optimize the predictive accuracy of the HMM. We varied both the number of haplotypes and the prior parameters and assessed the accuracy of imputation by randomly masking portions of the genotype data. When H is too small, accuracy declines but we saw little improvement for these data when H is greater than 5 or 6. In genomic regions with fewer than six distinct haplotypes, a subset of the states will typically have small marginal probabilities and are effectively unused. Based on these studies, we chose $H=6$ and a prior mean transition rate of 0.01 as optimal values for imputation in the merged NIEHS-Broad data (Figure 2).

RESULTS

Combining large scale SNP panels

In order to create the imputed genotype resource, we first merged the SNP genotypes in the NIEHS and Broad data. We refer to these data as the *merged set*. The relationship of the merged set to other SNP sets used in this study is summarized in Figure 2.

The NIEHS data (http://mouse.perlegen.com/mouse/download.html, July 2006 release), include 109 million genotypes on ~8.3 million SNPs spanning the 19 autosomes, the X and Y chromosomes, and the mitochondrial genome, for 11 classical and four wild-derived strains. The genotypes were generated by Perlegen Sciences using high density oligonucleotide arrays. Approximately 54% of the NIEHS SNPs have missing genotypes for one or more of the 15 strains summing to 12% incomplete genotypes. The frequency of missing data is higher in the three wild-derived strains of non-*domesticus* origin (CAST, MOLF and PWD). We removed 5.5% of SNPs from the initial NIEHS set due to potential problems (see Materials and Methods and Table S1). The 7,804,762 remaining SNPs spanning the autosomes and the X chromosome were used in this study.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

The Broad data (http://www.broad.mit.edu/~claire/MouseHapMap, February 2006 release) include over 6 million genotypes for 138,793 SNPs distributed at ~20kb intervals across the autosomes and the X chromosome, for 38 classical and 11 wild-derived strains representing four subspecies of *M. musculus* and two representatives of *M. spretus*. Genotypes were generated using custom Affymetrix SNP array technology. Approximately 68% of the SNPs have missing genotypes for one or more strains summing to 8% incomplete genotypes. The frequency of incomplete genotypes is higher for wild-derived strains with the exception of the two wild-derived *M. m. domesticus* strains, WSB and PERA. After mapping to NCBI build 36 coordinates and data preparation (see Materials and Methods), 135,846 SNPs were retained for this study.

The 15 NIEHS strains are common to both datasets and the remaining 33 strains are unique to the Broad data. Genotypes from the reference C57BL/6 genome sequence are included in the merged data. In total there are 116 million genotypes for 7,870,134 SNPs spanning the autosomes and the X chromosome of 49 strains.

In the process of merging datasets of dramatically different marker densities we assigned all of the unavailable genotypes to be missing. Thus there are two types of missing genotypes in the merged set. Experimental missing data are due to failure of a genotyping assay. Missing data created as a result of merging the data have, for the most part, not been directly assayed. In the merged set, 7,734,384 of the loci are assayed only in NIEHS data, 65,468 only in the Broad data and 70,282 loci have been assayed in both. There are a total of 14,078,485 experimental missing genotypes, 255,234,672 missing genotypes created by merging these data, and 6,318 missing values due to conflicts (see Methods). The frequency of incomplete genotypes is 98.3% for the 33 strains unique to the Broad set and ranges from 10% to 16% for the 15 strains common to both Broad and NIEHS data.

The imputed genotypes

We implemented a hidden Markov model (HMM) with a left to right architecture (Figure 1) for the primary purpose of genotype imputation and for the secondary purpose of haplotype identification. The architecture is similar to those described in Kimmel and Shamir (2005), and in Scheet and Stephens (2006). The number of hidden states at each SNP locus and the prior distribution of the model parameters of the HMM were optimized for genome-wide imputation accuracy (see Materials and Methods). All 49 strains in the merged set were used to train the model (see Materials and Methods). The posterior probability of each imputed genotype under the trained model provides a confidence score. A total of 269,319,475 missing genotypes were imputed in the merged set (*merged-imputed set* in Figure 2), of which 14.9%, 15.0%, and 70.1%, fall into the low, medium and high confidence score bins of (0,0.6), (0.6,0.9) and (0.9,1), respectively (Table S2).

In order to assess the quality of the imputed genotypes, we assembled a *validation set* of genotypes reported in eight SNP resources developed independently from the NIEHS and Broad sets (Figure 2, Table 1, Table S3). A total of 969,457 imputed genotypes could be validated using these resources, of which 15.4%, 13.7%, and 70.8%, fall into the low medium and high confidence score bins. The number of validated genotypes varies substantially among inbred strains (Table 2) and seven strains (129S4/SvJae, DDK/Pas, MAI/Pas, O20, Qsi5, ST/bJ and SEG/Pas) have no genotypes available for validation.

We compared imputed genotypes to genotypes in the validation set and conservatively assume that discordant genotypes represent imputation errors. The overall imputation error rate based on comparison with the validation set is 0.104 (Table 2). Error rates vary substantially among strains and to a lesser degree across chromosomes (Table S2). Strain specific error rates are lower for classical strains than for wild-derived strains. Furthermore, error rates vary for the

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

two different types of missing data (Table 2). For imputed genotypes with high confidence scores, the error rate is 0.044 (Table 2). Among validated genotypes with high confidence scores wild-derived strains have higher error rates than classical strains. The NIEHS strains have higher error rates because most of the missing genotypes are experimental.

As a consequence of high levels of divergence between the classical and wild-derived strains, 59% of SNPs in the merged set are private to the wild-derived strains (i.e., the genotypes of those SNPs are constant within classical strains). We estimated error rates stratified by the status of a SNP being constant or polymorphic within the classical strains and found that they were essentially identical (Table S4). We note that, for strains A, DBA/2 and 129S1, the error rates within the constant SNPs (26% of the total validated SNPs) were elevated compared to the unstratified version (Table 2). These strains have a large number of validation genotypes available (Mural et al. 2002). This interesting pattern and the higher experimental error rates in the NIEHS strains suggest that gene conversion may be responsible for a large fraction of the imputation error.

To test how wild-derived strains impact the imputation accuracy among the classical strains, we retrained the HMM on a subset of the merged data including only the 38 classical strains and estimated the imputation error rates by comparison with the validation data. The impact on error rates varies across chromosomes and is most evident in chromosomes where there is a substantial contribution to the classical strains from *M. m. musculus* (Yang et al. 2007). The higher errors observed overall suggest that the inclusion of the wild-derived strains improves the imputation of missing genotypes in regions where some classical inbred strains are not of *M. m. domesticus* origin, without negatively impacting the rest.

Based on this empirical validation, we conclude that the quality of the imputed genotypes improves as information from more strains is utilized in training the HMM, and that the imputation of missing genotypes is highly reliable for most strains.

Imputation of genotypes in other mouse strains

The trained HMM can be used to infer high density genotypes for strains that are not included in the training set using the Viterbi algorithm (Viterbi, 1967). Strains NZO/HILtJ and PWK/PhJ are not in the merged set but are of interest because they are founder strains of the Collaborative Cross (Churchill et al. 2004). NZO is a classical inbred strain closely related to strain NZB of the merged set. PWK is a wild derived *M. musculus* strain that is most closely related to strain PWD in the merged set, although the overall sequence divergence between PWK and PWD is greater than between any pair of classical strains. We have assembled 140,269 genotypes and 132,862 genotypes for NZO and PWK, respectively, from four independent resources (Table S3; Tim Wiltshire, personal communication) and created a medium density SNP set by retaining only genotypes that correspond to Broad loci in the merged set. The remaining SNPs were used for validation (Table 3). Similarly, we created low density versions of NZO and PWK SNPs by selecting SNP loci that correspond to loci genotyped in the Wellcome-CTC study (Shifman et al, 2007). We ran the Viterbi algorithm and imputed missing genotypes for both strains at medium and low density. At medium density, NZO imputations yield a large proportion of high confidence SNPs; but at low density, confidence scores shift substantially downward. At medium density about one third of the imputed PWK SNPs have low confidence and at low density, nearly 80% of imputed SNPs fall into the lowest confidence category. The low confidence in the PWK imputations is likely due to the presence of *M .m. musculus* haplotypes in PWK that are not represented in the NIEHS strains. The validation accuracy of the imputed SNPs (Table 3) follows the same pattern as the confidence scores. Although the overall validation error rate of 38% for low density PWK imputations is unacceptably high, but for those SNPs that achieve the highest confidence scores (>0.9), chromosome specific error rates range from 1% to 8%. These results suggest that

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

additional strains with medium density SNP genotyping can be accurately imputed and that the best overall results will be achieved for classical strains.

The imputed genotype resource

To generate the most accurate imputation of genotypes on the 49 inbred mouse strains, we added the 969,457 SNP genotypes in the validation set to the merged set and created a *combined set* of SNP genotypes (Figure 2). We then retrained the HMM using 49 strains in the combined set and imputed the missing genotypes. The confidence score distribution of imputed genotypes from the combined set demonstrated a slight shift towards the higher confidence score bins, indicating increased accuracy with 14.1%, 14.7%, 71.2% of all imputed genotypes falling into the low, medium and high confidence score bins, respectively. In addition, we recomputed the imputation for two strains, NZO and PWK, using all of the available genotypes. These strains were not included in the training set. The complete set of imputed genotypes and confidence scores for 51 strains are available at http://cgd.jax.org/. The data can be downloaded or queried through a MYSQL database. The SNPs have been quality-checked, and cross-linked with other features, including ENSEMBL annotations, GO annotations, and MGI gene and phenotype information. We created a web interface that allows SNP retrieval filtered on features such as neighboring genes, genomic location, SNP functional implication, CpG sites, and substitution types.

DISCUSSION

We have created a data resource of experimental and imputed SNP genotypes at a density of 7,870,134 loci on 49 commonly used inbred mouse strains by combining data and imputing missing genotypes from two major public SNP collections. Our results support the hypothesis that strains with medium density genotyping can be accurately imputed to obtain high density genotypes. Confidence scores assigned to each genotype reflect the reliability of imputed genotypes and identify experimental genotypes that depart from expectations based on local LD. We have demonstrated the accuracy of imputed genotypes by comparison to experimental genotypes obtained independently.

Imputed genotypes are not a replacement for experimental measurements. We encourage investigators to use this resource as an exploratory tool, but critical conclusions based on imputed genotypes, should be validated. Low confidence imputations are more common in the wild-derived strains due the limited representation of appropriate taxa in the NIEHS strain panel. Nonetheless, the reliability of most genotypes in this resource is sufficient for high throughput analysis and hypothesis generating.

Extensive local LD, reflecting the small number of founders and the presence of admixture in laboratory mice, provides the essential structure that allows accurate imputation of missing genotypes. To achieve this, the density of SNP genotyping should be sufficient to tag most regions of local LD in the population of strains to which the imputation algorithm is being applied. Imputation may be unreliable if a novel haplotype, not present in the high density training data, is encountered. Furthermore, SNPs that were not identified in the discovery process are not represented in the imputed data resource. We have previously estimated the false negative rate in the NIEHS data to be 67% (the false negative rate in the classical strains is 43%) but the rate is significantly higher for singleton SNPs (Yang et al. 2007). Therefore, absence of a SNP in this, or any, resource does not imply genetic identity. False negatives are of concern when the missing SNPs are private to one strain or to a small group of related strains. Discovery bias will significantly impact our ability to accurately impute SNPs in inbred strains derived from diverse mouse lineages. The solution is to carry out more SNP discovery at high density. The report of the sequence of multiple *Drosophila* species highlights the importance and benefits of resequencing and SNP discovery in diverse taxa (*Drosophila* 12 Genomes

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Consortium 2007). Similarly, it will be particularly useful to identify lineage specific SNPs and to genotype additional representatives of lineages with high error rates such as *M. m. castaneus* and *M. spretus*.

During the data preparation phase of this project we intentionally removed SNPs from regions with evidence of multiple copies in the C57BL/6 genome due to higher error rates observed in these SNPs (unpublished) and there are variable repeat regions present in other strains. Repeats and copy number variation need to be included to achieve a complete understand of the landscape of genetic variation in the laboratory mouse.

Other important features of genetic variation, such as gene conversion and recurrent mutations, may be missed because they are not consistent with the local LD pattern. In fact, the patterns of errors observed in the Celera strains suggest that both processes contribute significantly to the imputation error. This situation can be solved through additional experimental determination of missing genotypes. Finally, genotyping errors in the training data present a major barrier to improving the accuracy of imputed genotypes. Identification and resolution of genotyping errors in data as extensive as these is a daunting task. Confidence scores obtained from the HMM can suggest potential genotyping errors, but not all genotyping errors can be detected in this way. New computational approaches to error detection would be beneficial.

We used an empirical approach to validate imputed genotypes. Our estimated error rates are likely to be conservative because of genotyping errors in the validation set. An alternative method to assess imputation accuracy (Roberts et al. 2007) is to randomly mask a proportion of the genotypes within the combined dataset and to compare the imputed values to the masked genotypes. We found that the random masking approach provides higher estimated accuracy compared to the empirical validation. The processes that lead to missing data are likely to be more complex than the masking model, thus we prefer the empirical estimates but acknowledge their conservative bias.

The imputed genotype resource must be viewed dynamically because the coverage of strains, the number of SNP loci, and the accuracy of imputation will improve with additional data from ongoing genotyping projects. In conclusion, this study reports a method for accurate imputation of missing genotypes and its use in generating a dense map of the genetic variation in the mouse genome. Our results support the proposal by Frazer and coworkers (2007) that such a resource could and must be generated.

Acknowledgments

This work was supported by the US National Institutes of General Medical Sciences as part of the Center of Excellence in Systems Biology (1P50 GM076468). We thank Tim Wiltshire for sharing genotyping data prior to its publication, Jesse Hammer and Susan Moxley for graphics assistance.

References

Abe K, Noguchi H, Tagawa K, Yuzuriha M, Toyoda A, et al. Contribution of Asian mouse subspecies Mus musculus molossinus to genomic constitution of strain C57BL/6J, as defined by BAC-end sequence-SNP analysis. Genome Res 2004;14:2439–2447. [PubMed: 15574823]

Bogue MA. Mouse Phenome Project: understanding human biology through mouse genetics and genomics. J Appl Physiol 2003;95:1335–1337. [PubMed: 12970372]

Cervino AC, Li G, Edwards S, Zhu J, Laurie C, et al. Integrating QTL and high-density SNP analyses in mice to identify Insig2 as a susceptibility gene for plasma cholesterol levels. Genomics 2005;86:505–517. [PubMed: 16126366]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Churchill GA, Airey DC, Allayee H, Angel JM, Attie AD, et al. The Collaborative Cross, a community resource for the genetic analysis of complex traits. Nat Genet 2004;36:1133–1137. [PubMed: 15514660]

Churchill GA. Stochastic models for heterogeneous DNA sequences. Bulletin of Mathematical Biology 1989;51:79–94. [PubMed: 2706403]

DiPetrillo K, Wang X, Stylianou L, Pagien B. Bioinformatics toolbox for narrowing rodent quantitative trait loci. Trends Genet 2005;21:684–692.

Durbin, R.; Eddy, SR.; Krogh, A.; Mitchison, G. Biological sequence analysis. Cambridge University Press; Cambridge, UK: 1998.

Drosophila 12 genomes consortium. Evolution of genes and genomes on the Drosophila phylogeny. Nature 2007;450:203–218. [PubMed: 17994087]

Frazer KA, Wade CM, Hinds DA, Patil N, Cox DR, et al. Segmental phylogenetic relationships of inbred mouse strains revealed by fine-scale analysis of sequence variation across 4.6 Mb of mouse genome. Genome Res 2004;14:1493–1500. [PubMed: 15289472]

Frazer KA, Eskin E, Kang HM, Bogue MA, Hinds DA, et al. A sequence-based variation map of 8.27 million SNPs in inbred mouse strain. Nature 2007;448:1050–1053. [PubMed: 17660834]

Guenet JL, Bohomme F. Wild mice: an ever-increasing contribution to a popular mammalian model. Trends Genet 2003;19:24–31. [PubMed: 12493245]

Ideraabdullah FY, de la Casa-Esperon E, Bell TA, Detwiler DA, Magnuson T, et al. Genetic and haplotype diversity among wild derived mouse inbred strains. Genome Res 2004;14:1880–1887. [PubMed: 15466288]

Kent WJ. BLAT--the BLAST-like alignment tool. Genome Res 2002;12:656–664. [PubMed: 11932250]

Kimmel G, Shamir R. A block-free hidden Markov model for genotypes and its application to disease association. J Comput Biol 2005;12:1243. [PubMed: 16379532]

Liao G, Wang J, Guo J, Allard J, Cheng J, et al. In silico genetics: identification of a functional element regulating H2-Ealpha gene expression. Science 2004;306:690–695. [PubMed: 15499019]

Lyon, MF.; Rastan, S.; Brown, SDM., editors. Genetic variants and strains of the laboratory mouse. Vol. 3. Oxford Univeristy Press; Oxford, UK: 1996.

McClurg P, Janes J, Wu C, Delano DL, Walker JR, et al. Genomewide association analysis in diverse inbred mice: power and population structure. Genetics 2007;176:675–683. [PubMed: 17409088]

Mott R. A haplotype map for the laboratory mouse. Nat Genet 2007;39:1054–1056. [PubMed: 17728771]

Mural RJ, Adams MD, Myers EW, Smith HO, Miklos GL, et al. A comparison of whole-genome shotgun-derived mouse chromosome 16 and the human genome. Science 2002;296:1661–1671. [PubMed: 12040188]

Payseur BA, Hoekstra HE. Signatures of reproductive isolation in patterns of single nucleotide diversity across inbred strains of mice. Genetics 2005;171:1905–1016. [PubMed: 16143616]

Payseur BA, Place M. Prospects for association mapping in classical inbred mouse strains. Genetics 2007;175:1999–2008. [PubMed: 17277361]

Pletcher MT, McClurg P, Batalov S, Su AI, Barnes SW, et al. Use of a dense single nucleotide polymorphism map for in silico mapping in the mouse. PLoS Biol 2004;2:2159–2169.

Petkov PM, Graber JH, Churchill GA, DiPetrillo K, King BL, et al. Evidence of a large-scale functional organization of mammalian chromosomes. PLoS Genet 2005;1:e33. [PubMed: 16163395]

Petkov PM, Ding Y, Cassell MA, Zhang W, Wagner G, et al. An efficient SNP system for mouse genome scanning and elucidating strain relationships. Genome Res 2004;14:1806–1811. [PubMed: 15342563]

Roberts A, McMillan L, Wang W, Parker J, Rusyn I, et al. Inferring missing genotypes in large SNP panels using fast nearest-neighbor searches over sliding windows. Bioinformatics 2007;23:i401. [PubMed: 17646323]

Roberts A, Pardo-Manuel de Villena F, Wang W, McMillan L, Threadgill DW. The polymorphism architecture of mouse genetic resources elucidated using genome-wide resequencing data: implications for QTL discovery and systems genetics. Mamm Genome 2007;18:473–481. [PubMed: 17674098]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Siebert SK, Schadt EE. Moving toward a system genetics view of disease. Mamm Genome 2007;18:389–401. [PubMed: 17653589]

Scheet P, Stephens M. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet 2006;78:129.

Shifman S, Bell JT, Copley RR, Taylor MS, Williams RW, et al. A high-resolution single nucleotide polymorphism genetic map of the mouse genome. PLoS Biol 2006;4:e395. [PubMed: 17105354]

Viterbi AJ. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE Trans Information Theory 1967;13:260–269.

Wade CM, Kulbokas EJ 3rd, Kirby AW, Zody MC, Mullikin JC, et al. The mosaic structure of variation in the laboratory mouse genome. Nature 2002;420:574–578. [PubMed: 12466852]

Wade CM, Daly MJ. Genetic variation in laboratory mice. Nat Genet 2005;37:1175–1180. [PubMed: 16254563]

Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, et al. Initial sequencing and comparative analysis of the mouse genome. Nature 2002;420:520–562. [PubMed: 12466850]

Wiltshire T, Pletcher MT, Batalov S, Barnes SW, Tarantino LM, et al. Genome-wide single-nucleotide polymorphism analysis defines haplotype patterns in mouse. Proc Natl Acad Sci USA 2003;100:3380–3385. [PubMed: 12612341]

Yalcin B, Fullerton J, Miller S, Keays DA, Brady S, et al. Unexpected complexity in the haplotypes of commonly used inbred strains of laboratory mice. Proc Natl Acad Sci USA 2004;101:9734–9739. [PubMed: 15210992]

Yang H, Bell TA, Churchill GA, Pardo-Manuel de Villena F. On the subspecific origin of the laboratory mouse. Nat Genet 2007;39:1100–1107. [PubMed: 17660819]

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

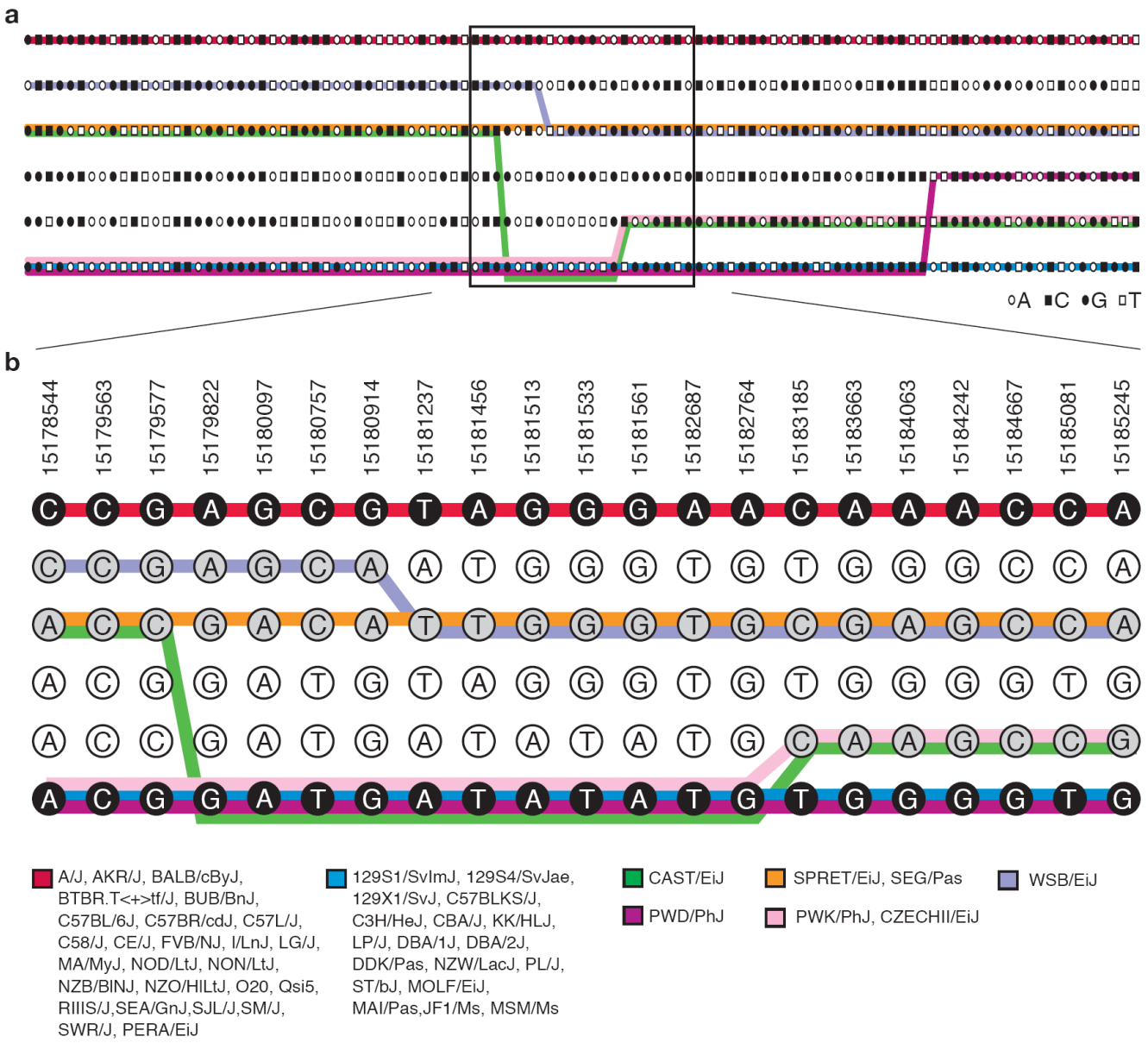


Figure 1.

HMM architecture. Each SNP locus is modeled using six hidden states representing haplotypes. Each state is labeled by the nucleotide that is most likely to be observed in that haplotype. Colored lines represent the most probable haplotypes (Viterbi paths) for strains shown at the bottom. (a) A 40kb interval on chromosome 14 spanning 105 SNPs. (b) Detailed view of a 6.7 kb segment. Shading of the state nodes corresponds to the marginal state probability with darker color indicating haplotypes that are well represented in the training data.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

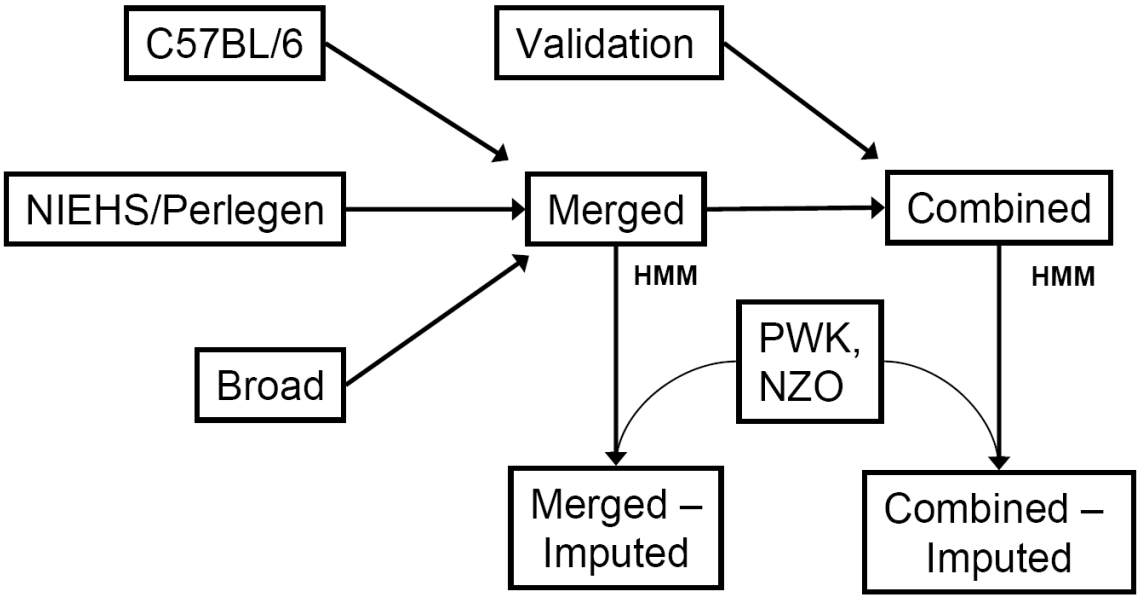


Figure 2.
Relationships among SNP data sources. The sequence of C57BL/6J provides reference genotypes for all SNPs in this study. The NIEHS data on 15 strains and Broad data on 48 strains were combined to create a merged set of experimental SNPs. The HMM was trained on the merged set and used to create the merged-imputed set. The validation set of experimental SNPs was assembled and curated from sources listed in Supplemental Table S3 and compared to the merged-imputed set in our validation study. The validation set was then combined with the merged SNPs to create the combined set of experimental SNPs. The HMM was retrained on the combined set to generate the combined-imputed set. SNP data from strains PWK and NZO were threaded through both of the trained HMMs. The results from threading with the merged-imputed model were used in our validation study. The PWK and NZO SNP data were not used in the training of either HMM.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Table 1

A summary of SNP resources. For each SNP resource, the name, the number of SNPs, the number of the strains reported, the genotyping technology used, and a literature citation are shown.

Resource	No. SNP	No. Strain	Genotyping technology	Web site	Reference
NIEHS	8,272,574	16	Perlegen Science oligonucleotide arrays	mouse.perlegen.com	Frazer et al. (2007)
Broad	138,608	49	Affymetrix SNP array	www.broad.mit.edu/~claire/MouseHapMap	Wade and Daly (2005)
Celera	2,093,327	5	Whole genome shotgun assembly	www.celera.com	Mural et al. (2002)
GNF	9,594	48	PCR and DNA sequencing	snp.gnf.org	Wiltsire et al. (2003)
TJL	1,638	144	Amplifluor technology	aretha.jax.org	Petkov et al. (2004)
Japanese MSM	210,160	1	BAC-end sequence analysis	stt.gsc.riken.jp/msm/	Abe et al. (2004)
Wellcome-CTC	13,348	499	Illumina SNP array	www.well.ox.ac.uk/mouse/INBREDS	Shifman et al. (2006)
Sanger	748,723	7	whole gneome shotgun, clone based sequencing	www.ensembl.org; www.sanger.ac.uk	Waterston et al. (2002)
db-SNP	11,814	7	Central repository	www.ncbi.nlm.nih.gov/SNP	
Rosetta/Merck	12,473	62	Illumina	www.rii.com	Cervino et al. (2005)
Roche	214,706	20	Sequencing	mousesnp.roche.com/	Liao et al (2004)

Mamm Genome. Author manuscript; available in PMC 2009 August 12.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Table 2

Estimated strain specific error rates in the merged-imputed set. The table provides (1) the total number of imputed missing genotypes, the percentage of imputed genotypes that were missing due to merging datasets; (2) the total number of validated imputed genotypes, the error rate of validated imputed genotypes, the error rates for imputed genotypes that were experimental missing data and for imputed genotypes that were missing due to merging datasets; (3) the number and error rates of validated imputed genotypes with confidence scores greater than 0.9. The table is divided to indicate classical (top) versus wild derived (bottom) strains. NIEHS strains are listed first within these categories.

	ALL MISSING DATA IMPUTED		VALIDATED IMPUTED				VALIDATED IMPUTED CONFIDENCE >0.9			
Strains	**# missing genotype imputed**	**% missing data due to merge**	**# validated imputed**	**error for ALL missing data**	***error for experimental missing***	***error for missing due to merge***	**# validated imputed conf>0.9**	**error for ALL missing data**	***error for experimental missing***	***error for missing due to merge***
DBA/2J	887413	0	70678	0.156	*0.156*	-	50734	0.077	*0.077*	-
A/J	738284	0	66222	0.160	*0.160*	-	47485	0.078	*0.078*	-
129S1/SvImJ	860002	0	26846	0.168	*0.168*	-	19284	0.089	*0.089*	-
C3H/HeJ	921492	0	1337	0.147	*0.147*	-	944	0.058	*0.058*	-
BTBR.T<+>tf/J	845230	0	1256	0.143	*0.143*	-	926	0.054	*0.054*	-
FVB/NJ	852011	0	1212	0.152	*0.152*	-	850	0.069	*0.069*	-
KK/HLJ	845405	0	1209	0.168	*0.168*	-	816	0.082	*0.082*	-
NOD/LtJ	871145	0	1190	0.155	*0.155*	-	828	0.070	*0.070*	-
AKR/J	836285	0	1168	0.170	*0.170*	-	809	0.079	*0.079*	-
NZW/LacJ	846575	0	1160	0.147	*0.147*	-	812	0.074	*0.074*	-
BALB/cByJ	795150	0	7	0.000	*0.000*	-	7	0.000	*0.000*	-
129X1/SvJ	7740291	98.3%	485452	0.076	*0.060*	*0.076*	364089	0.030	*0.018*	*0.030*
SM/J	7753890	98.3%	10695	0.108	*0.102*	*0.109*	7108	0.048	*0.041*	*0.048*
NZB/BlNJ	7745598	98.3%	10456	0.103	*0.098*	*0.103*	7075	0.040	*0.021*	*0.040*
NON/LtJ	7747432	98.3%	10447	0.086	*0.113*	*0.086*	7081	0.027	*0.036*	*0.027*
SJL/J	7743333	98.3%	10424	0.089	*0.086*	*0.089*	7060	0.031	*0.007*	*0.032*
CBA/J	7742711	98.3%	10423	0.077	*0.039*	*0.078*	7276	0.022	*0.019*	*0.022*
CE/J	7745164	98.3%	10421	0.124	*0.132*	*0.124*	6892	0.058	*0.056*	*0.059*
BUB/BnJ	7742202	98.3%	10416	0.090	*0.067*	*0.090*	7111	0.030	*0.028*	*0.030*
LG/J	7743879	98.3%	10413	0.094	*0.123*	*0.093*	7098	0.038	*0.038*	*0.038*
SWR/J	7742958	98.3%	10407	0.097	*0.111*	*0.097*	7069	0.041	*0.034*	*0.041*
LP/J	7742766	98.3%	10387	0.093	*0.085*	*0.093*	7096	0.030	*0.030*	*0.030*
C58/J	7741022	98.3%	10379	0.098	*0.123*	*0.098*	7131	0.034	*0.024*	*0.034*
C57BR/cdJ	7742389	98.3%	10366	0.096	*0.059*	*0.097*	7195	0.034	*0.000*	*0.035*

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

	ALL MISSING DATA IMPUTED		VALIDATED IMPUTED				VALIDATED IMPUTED CONFIDENCE >0.9			
Strains	**# missing genotype imputed**	**% missing data due to merge**	**# validated imputed**	**error for ALL missing data**	***error for experimental missing***	***error for missing due to merge***	**# validated imputed conf>0.9**	**error for ALL missing data**	***error for experimental missing***	***error for missing due to merge***
I/LnJ	7743740	98.3%	10363	0.095	*0.101*	*0.095*	7092	0.039	*0.014*	*0.040*
PL/J	7741981	98.3%	10332	0.087	*0.076*	*0.088*	7092	0.029	*0.009*	*0.029*
RIIIS/J	7743061	98.3%	10331	0.104	*0.131*	*0.104*	6862	0.039	*0.040*	*0.039*
C57L/J	7743799	98.3%	7996	0.111	*0.105*	*0.111*	5266	0.033	*0.008*	*0.034*
MA/MyJ	7743694	98.3%	7973	0.108	*0.119*	*0.107*	5232	0.036	*0.042*	*0.035*
SEA/GnJ	7743682	98.3%	7971	0.098	*0.078*	*0.099*	5272	0.024	*0.000*	*0.025*
C57BLKS/J	7741314	98.3%	7958	0.084	*0.092*	*0.084*	5668	0.026	*0.026*	*0.026*
DBA/1J	7742347	98.3%	7926	0.103	*0.110*	*0.103*	5292	0.026	*0.020*	*0.027*
CAST/EiJ	1170222	0	1877	0.295	*0.295*	-	603	0.250	*0.250*	-
MOLF/EiJ	1137192	0	1764	0.190	*0.190*	-	887	0.106	*0.106*	-
WSB/EiJ	876918	0	1121	0.241	*0.241*	-	667	0.186	*0.186*	-
PWD/Ph	1211132	0	520	0.212	*0.212*	-	238	0.134	*0.134*	-
MSM/Ms	7750801	98.3%	79953	0.133	*0.165*	*0.133*	45522	0.052	*0.122*	*0.052*
PERA/EiJ	7748601	98.3%	10498	0.199	*0.196*	*0.199*	6487	0.130	*0.134*	*0.130*
SPRET/EiJ	7765710	98.3%	8909	0.219	*0.238*	*0.218*	5084	0.193	*0.175*	*0.194*
CZECHII/EiJ	7749890	98.3%	7789	0.114	*0.167*	*0.112*	4846	0.070	*0.131*	*0.067*
JF1/Ms	7751232	98.3%	3205	0.070	*0.119*	*0.068*	1966	0.033	*0.082*	*0.030*
Genome-wide	**215,077,943**		**969,457**	**0.104**	***0.158***	***0.090***	**686,852**	**0.044**	***0.078***	***0.036***

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

Table 3

Estimated error rates for imputed genotypes using SNP genotyping data at different densities. The table provides (1) the total number of genotyped SNPs; the proportion of the merged data; (2) the number of imputed genotypes, (3) the number of validated imputed genotypes and error estimate; (4) the number of validation imputed genotypes and error estimates for confidence score bins of (0,0.6), (0.6,0.9), (0.9,1).

Strain	# genotyped snp	% of total snp in merged set	Number Imputed	Number validated.	Error validated	Number validated in.confidence.score.bin			error. in.confidence.score.bin		
						(0,0.6)	(0.6,0.9)	(0.9,1)	(0,0.6)	(0.6,0.9)	(0.9,1)
NZO.medium	130,387	1.7%	7,739,747	9,882	0.092	1,387	1,822	6,673	0.327	0.149	0.028
NZO.low	8,491	0.1%	7,861,643	131,778	0.159	50,220	45,532	36,026	0.336	0.080	0.011
PWK.medium	123,709	1.6%	7,746,425	9,153	0.086	2,408	1,474	5,271	0.206	0.091	0.030
PWK.low	8,085	0.1%	7,862,049	124,777	0.378	89,148	23,609	12,020	0.481	0.161	0.032