Imputation of Single-Nucleotide Polymorphisms in Inbred Mice Using Local Phylogeny

Jeremy R. Wang,* Fernando Pardo-Manuel de Villena,† Heather A. Lawson,‡ James M. Cheverud,‡ Gary A. Churchill,§ and Leonard McMillan*,[1]

Departments of *Computer Science and, †Genetics, Lineberger Comprehensive Cancer Center, Carolina Center for Genome Science, University of North Carolina, Chapel Hill, North Carolina 27599, ‡Department of Anatomy and Neurobiology, Washington University School of Medicine, St. Louis, Missouri 63110, and, and §The Jackson Laboratory, Bar Harbor, Maine 04609

ABSTRACT We present full-genome genotype imputations for 100 classical laboratory mouse strains, using a novel method. Using genotypes at 549,683 SNP loci obtained with the Mouse Diversity Array, we partitioned the genome of 100 mouse strains into 40,647 intervals that exhibit no evidence of historical recombination. For each of these intervals we inferred a local phylogenetic tree. We combined these data with 12 million loci with sequence variations recently discovered by whole-genome sequencing in a common subset of 12 classical laboratory strains. For each phylogenetic tree we identified strains sharing a leaf node with one or more of the sequenced strains. We then imputed high- and medium-confidence genotypes for each of 88 nonsequenced genomes. Among inbred strains, we imputed 92% of SNPs genome-wide, with 71% in high-confidence regions. Our method produced 977 million new genotypes with an estimated per-SNP error rate of 0.083% in high-confidence regions and 0.37% genome-wide. Our analysis identified which of the 88 nonsequenced strains would be the most informative for improving full-genome imputation, as well as which additional strain sequences will reveal more new genetic variants. Imputed sequences and quality scores can be downloaded and visualized online.

AMONG the many advantages of inbred strains in genetic studies is that each strain needs to be genotyped only once, and that information can be reused in many experiments. Moreover, as more genotype data become available for a given inbred strain, the analysis can be updated. This cycle can continue until, ultimately, all inbred strains are fully sequenced. In the meantime, there is a need to leverage the handful of inbred strains that have been sequenced using robust imputation methods to maximize the value of existing data. High-quality imputed sequence has many potential applications including identification of functional variants and the creation of accurate scaffolds for the analysis of next-generation RNAseq and bisulfite sequencing data. Until affordable deep sequencing becomes a reality, a balanced approach that combines targeted sequencing with accurate imputation offers the best of both worlds: high-quality genomic data today at little additional cost.

A recent sequencing effort by the Wellcome Trust/Sanger Institute has made available dense genome sequences for a set of 17 inbred mouse strains, including 13 common laboratory strains, 3 wild-derived mouse strains from different subspecies of *Mus musculus*, and a single strain from a different species, *M. spretus* (Keane *et al.* 2011). This set of samples is expected to capture much of the variation found in common laboratory mouse strains and, therefore, provides a foundation for sequence imputation. A complementary resource is the recent release of Mouse Diversity Array (MDA) genotypes from 162 mouse strains (Yang *et al.* 2011). MDA is a high-density DNA microarray designed to assay diversity among commonly used laboratory mice (Yang *et al.* 2009). The density of SNP genotypes available on the MDA exceeds the density of recombination events accumulated over the development of the classical inbred strains and as such the MDA SNPs can provide a framework for imputation of the underlying whole-genome sequence.

Imputation can be used to increase the effective resolution of a lower-density SNP panel to match that of a higher-

doi: 10.1534/genetics.111.132381
Manuscript received July 10, 2011; accepted for publication October 20, 2011
Supporting information is available online at http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.111.132381/-/DC1.
[1]Corresponding author: Department of Computer Science, 319 Sitterson Hall, University of North Carolina, Campus Box 3175, Chapel Hill, NC 27599. E-mail: mcmillan@cs.unc.edu

density panel when there is a subset of samples common to both sets. Previous imputation methods use variations of a hidden Markov model (HMM) to infer sequence similarities and likely transitions between haplotypes. These methods employ probabilistic models based on local sequence similarity to infer the state of missing genotypes. Missing genotypes arise from two sources. No-calls (N's) can indicate either technical noise or an unexpected sequence variant such as a nearby SNP or an indel that interferes with probe hybridization. A second, and more extensive, source of missing genotypes is due to differences in the density of marker sets between platforms.

There have been two recent imputation efforts in the laboratory mouse (Szatkiewicz *et al.* 2008; Kirby *et al.* 2010). Szatkiewicz and co-workers imputed genotypes at 7.9 million loci by combining low-density genotypes from 51 classical and wild-derived inbred mouse strains with high-density SNP discovery data obtained on a subset of 16 inbred strains (Frazer *et al.* 2007; Yang *et al.* 2007). The authors imputed each locus consecutively across the genome, using an HMM to predict the most likely genotype among the possible alleles. Using this locus-by-locus method, they reported a 10.4% error rate over the entire genome and 4.4% error in high-confidence regions. High-confidence regions are defined by high posterior probability and cover 71% of the genome.

Kirby and co-workers imputed genotypes in 94 classical and wild-derived laboratory strains for 8.27 million SNP loci reported in the National Institute of Environmental Health Sciences (NIEHS)/Perlegen set (Frazer *et al.* 2007), using expectation-maximized integrative imputation (EMINIM) (Kang *et al.* 2010), a different HMM method that predicts genotypes by estimating haplotype blocks from the smaller set of samples with high-density genotypes. The hidden states in their model correspond to the 16 NIEHS/Perlegen strains or a 17th unknown state. Their method models recombination between haplotype blocks rather than transitions between SNPs. The authors imputed 657 million genotypes with a reported error rate of 2.4% over the entire genome and 0.27% in regions with high confidence based on posterior probability in the HMM.

These two methods do not explicitly take advantage of the local phylogenetic relationships present in classical inbred strains. This shortcoming is particularly significant given the strong population structure and the limited amount of haplotype diversity present in classical laboratory strains (Yang *et al.* 2011). Our approach estimates both haplotype blocks and the relatedness between them in the form of a local phylogenetic tree. In contrast to previous methods, our haplotype blocks and trees are inferred from a larger set of genotypes. This has the advantage that the larger set of samples can capture haplotype diversity that is not sampled in the smaller high-density set. The success of this approach requires that the SNP density in the larger sample set is sufficient to detect haplotype blocks, which is the case for the MDA genotypes (Yang *et al.* 2011). Moreover, the trees provide a measure of difference between haplotypes that is consistent with their evolutionary history.

Here we report the combined use of the MDA and Wellcome Trust/Sanger Institute resources to impute the genotypes of 88 common inbred strains (Supporting information, Figure S1). Our approach takes advantage of the local phylogenetic relationships among inbred strains to determine the confidence of local imputation and is highly accurate over most of the genome of common strains. On the basis of this imputation we discuss strategies for future sequencing and SNP discovery in the laboratory mice and the efficient use of this resource for association studies. Imputed genotypes and imputation confidence are provided in Table S1 and use of these data should cite this article as a reference. They are also publicly available at http://www.csbio.unc.edu/imputation/.

Materials and Methods

MDA genotype data

All genotype and haplotype data as well as the phylogenetic trees have been reported previously (Yang *et al.* 2011). This study is based on local phylogenetic trees for 100 classical laboratory strains (Figure S1) based on genotypes from MDA. MDA is an Affymetrix-based 6.5M probe platform with >600,000 SNP markers uniformly spaced across the nonrepetitive regions of the mouse genome (Yang *et al.* 2009). We used the subset of 549,683 high-quality markers that were genotyped in Yang *et al.* (2011). We also identified additional alleles in these markers including residual heterozygosity, deletions and other copy-number variation, and variable-intensity oligonucleotides (VINOs) (Yang *et al.* 2011; G. A. Churchill and F. Pardo-Manuel de Villena, unpublished data). VINOs and deletions were incorporated into haplotype and tree estimations by treating them as additional marker loci with binary alleles (*i.e.*, with and without VINO or with and without deletion) at positions coincident with the probe where they were detected (Yang *et al.* 2011).

C57BL/6J reference genome

The Wellcome Trust/Sanger Institute defines SNPs relative to the reference mouse sequence (Waterston *et al.* 2002) and uses the NCBI Genome Reference Consortium's build 37 (MGSCv37) (Church *et al.* 2009). The reference genome is derived from C57BL/6J and, thus, we included this strain as a 12th high-density sequence along with the 11 Sanger sequenced strains for which we have MDA genotypes (Figure S1).

Wellcome Trust/Sanger Institute genotype data

Our imputation incorporates the set of high-confidence SNPs recently discovered during the Wellcome Trust/Sanger Institute's sequencing of 17 inbred strains (Keane *et al.* 2011) (which we refer to henceforth as the Sanger set). Keane *et al.* (2011) identified >65 million SNPs, but most of these represent private alleles in wild-derived mouse

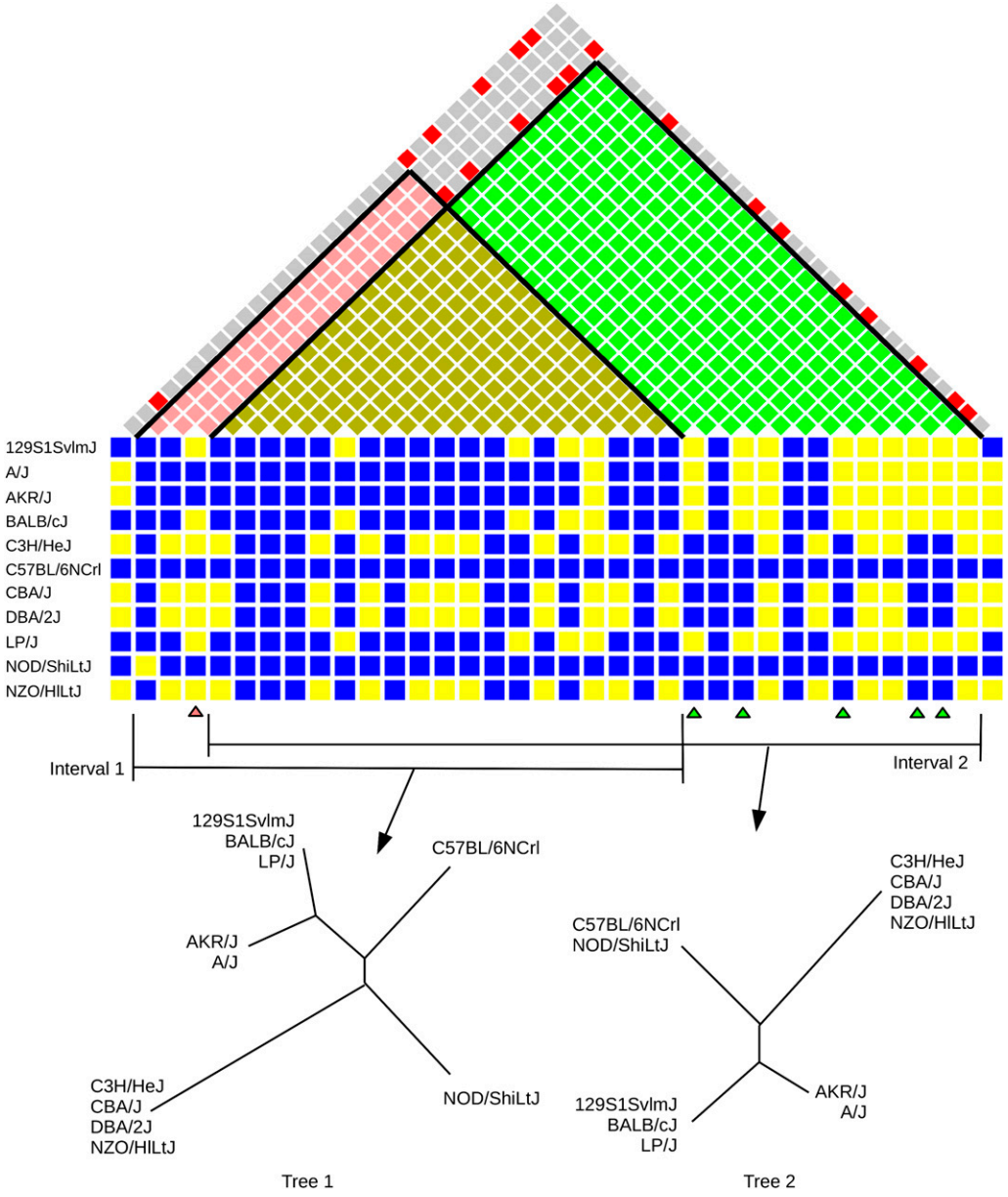

Figure 1 Identification of recombination intervals and phylogenetic tree construction. Two compatible intervals and associated phylogenetic trees (73.7–73.8 Mb on chromosome 19) are shown. The bottom blue and yellow matrix represents the panel of SNPs (columns) in this region where blue indicates the reference allele and yellow indicates the alternate allele. The top matrix represents the results of the pairwise four-gamete test where red indicates a violation (all four gametes are present in the corresponding SNP pair). The left interval (interval 1) is shown in pink and the right interval (interval 2) in green. Below the SNP matrix, the SNPs involved in violations of the four-gamete rule between these two intervals are indicated by pink and green arrowheads. The intersection of each SNP indicated with a pink arrowhead with the SNP indicated by the green arrowhead is red, indicating these SNP pairs violate the four-gamete rule.

strains. We excluded these SNPs and SNPs marked as low confidence or heterozygous in an inbred strain. Overall, 82% of the 65 million SNP loci do not vary within the subset of 12 classical strains in common between the Sanger and the MDA sets plus the reference genome (FigureS1). The strains in this common set are 129S1/SvlmJ, A/J, AKR/J, BALB/cJ, C3H/HeJ, C57BL/6N, C57BL/6J, CBA/J, DBA/2J, LP/J, NOD/ShiLtJ, and NZO/HlLtJ. We imputed the remaining 12,054,616 SNP loci in the 88 classical strains for which only MDA genotypes are available where they were medium or high confidence. We did not include wild-derived strains in our imputation (see *Discussion*).

LG/J and SM/J genotypes

Whole-genome sequencing for the LG/J (~20× haploid coverage) and the SM/J (~14× haploid coverage) strains was completed by the Washington University School of Medicine Genome Sequencing and Analysis Center, using Illumina sequencing in two steps as described in Mardis *et al.* (2009) and in Ding *et al.* (2010). Illumina reads from DNA extracted from the livers of a single LG/J female and a single SM/J female were aligned to the July 2007 assembly NCBI build 37 reference genome, using Mapping and Assembly with Quality (Li *et al.* 2008). SNPs for each strain were called using SamTools (Li *et al.* 2009), requiring a minimum of three reads and a SNP quality score ≥20 (H. A. Lawson, I. Nikolskiy, S. Chen, M. McLellan, J. Fay, E. Mardis, and J. M. Cheverud, unpublished data). For chromosomes 14 and 15, 305,114 SNPs were identified between LG/J and the reference and 422,879 SNPs were identified between SM/J and the reference. The LG/J and SM/J SNPs have been submitted to dbSNP (Sherry *et al.* 2001) under the

handle "Cheverud". For our validation method, we excluded SNPs for which a single allele could not be resolved, leaving 292,051 for LG/J and 416,589 for SM/J, relative to the reference genome.

Imputation

Our imputation method uses an algorithm based on the four-gamete rule (Hudson and Kaplan 1985) to define haplotype blocks that are consistent with a local perfect phylogeny. As described in Wang *et al.* (2010), we compare all pairs of SNPs and identify a minimal set of maximum-size contiguous intervals covering the genome in which no pair of SNPs violates the four-gamete rule (Figure 1). We constructed 40,647 intervals covering the entire genome. The median genomic size is 71 kb and covers 12 SNPs. Intervals had an average and median of five unique haplotypes. For each interval, we construct the local phylogenetic tree as described previously (Yang *et al.* 2011). Briefly, we computed a pairwise genotype similarity score among strains as the proportion of the matching variants in each interval. We identify shared haplotypes by connecting pairs of strains with similarity score >0.99. We constructed phylogenetic trees by connecting these haplotypes (leaves in the tree), using neighbor joining over the mean pairwise similarity of strains in each leaf. The general workflow of our imputation method is shown in Figure 2.

We compared the local phylogenetic structure in our MDA genotypes to the strain distribution patterns (SDPs) in the Sanger set. We validated that our local phylogenies matched the SDPs for the appropriate Sanger SNPs. In addition, we compared our haplotype clusters to the sequence similarity in the Sanger set and showed that our local phylogenies reflect the local sequence differences with the Sanger set. This validation prior to imputation supported our assumption that MDA SNPs were sufficient to define representative haplotype blocks.

We imputed each sample for which we have only MDA data by filling in genotypes of Sanger set samples in regions where they share a haplotype that is identical-by-descent (IBD) as indicated by a shared leaf in the local phylogenetic tree. We assign confidences to each imputed strain over each interval according to whether the imputed strain shares a haplotype block with a Sanger set sample. Figure 3 shows an example of imputation and the correspondence between phylogenetic trees and imputation confidence. High confidence was assigned to haplotypes in an interval for which there are one or more concordant Sanger sequences. Where multiple samples from the Sanger set share a haplotype, but exhibit segregating alleles (*i.e.*, evidence that our tree leaf could be further subdivided), we assign medium confidence. In medium-confidence cases, we resolve the allele at each locus independently by further subdividing the leaves using the haplotype structure in neighboring intervals until there is a consensus among remaining shared-haplotype samples. This method captures nearby haplotype structure where intervals are too small to fully differentiate samples.

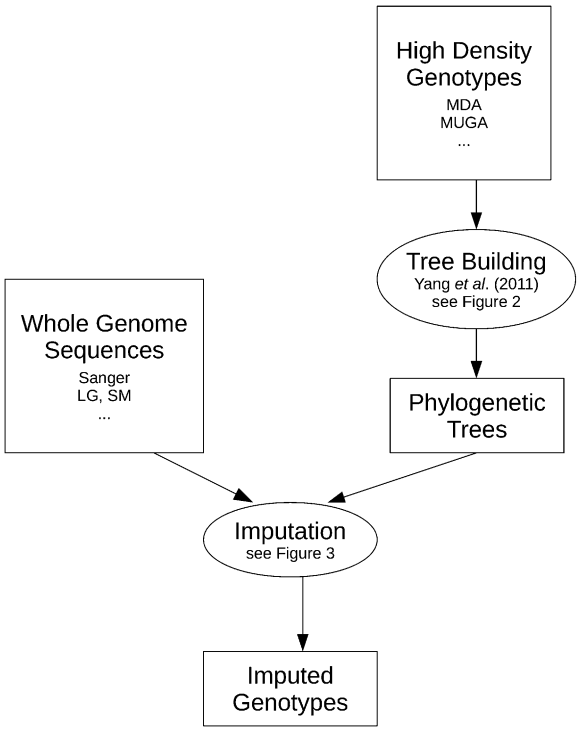

Figure 2 A diagram of the workflow showing the flow of data through our imputation method. Trees are constructed from the MDA genotype data (Figure 1). These trees are then used to inform the imputation using high-density sequence data (Figure 3).

There are many intervals where classical inbred strains do not share a haplotype with a strain from the Sanger set; thus no high-density imputation source is available. These intervals are assigned low confidence for strains not sharing a haplotype. Our approach provides no satisfactory imputation in these regions and our leave-one-out analysis indicates that they cannot be imputed with accuracy substantially better than 50%. As a result, these regions are assigned N. Since maximal compatible intervals can overlap, two intervals might cover a SNP. In this case, the interval with higher confidence is used. If the two intervals have the same confidence, the union of strains with a shared haplotype in each interval is used.

All genotypes and confidence scores are available in Table S1. Use of these data should cite this article as a reference.

Validation

To assess the accuracy of our imputation, we used a leave-one-out approach and compared our results directly to the Sanger sequence data. In the leave-one-out method, we removed 1 of the 12 Sanger strains and imputed SNPs using only the remaining 11 strains in the Sanger set. This method has the advantage that it allows us to consider the entire genome when determining accuracy. In addition, we performed external validation using sequence data for

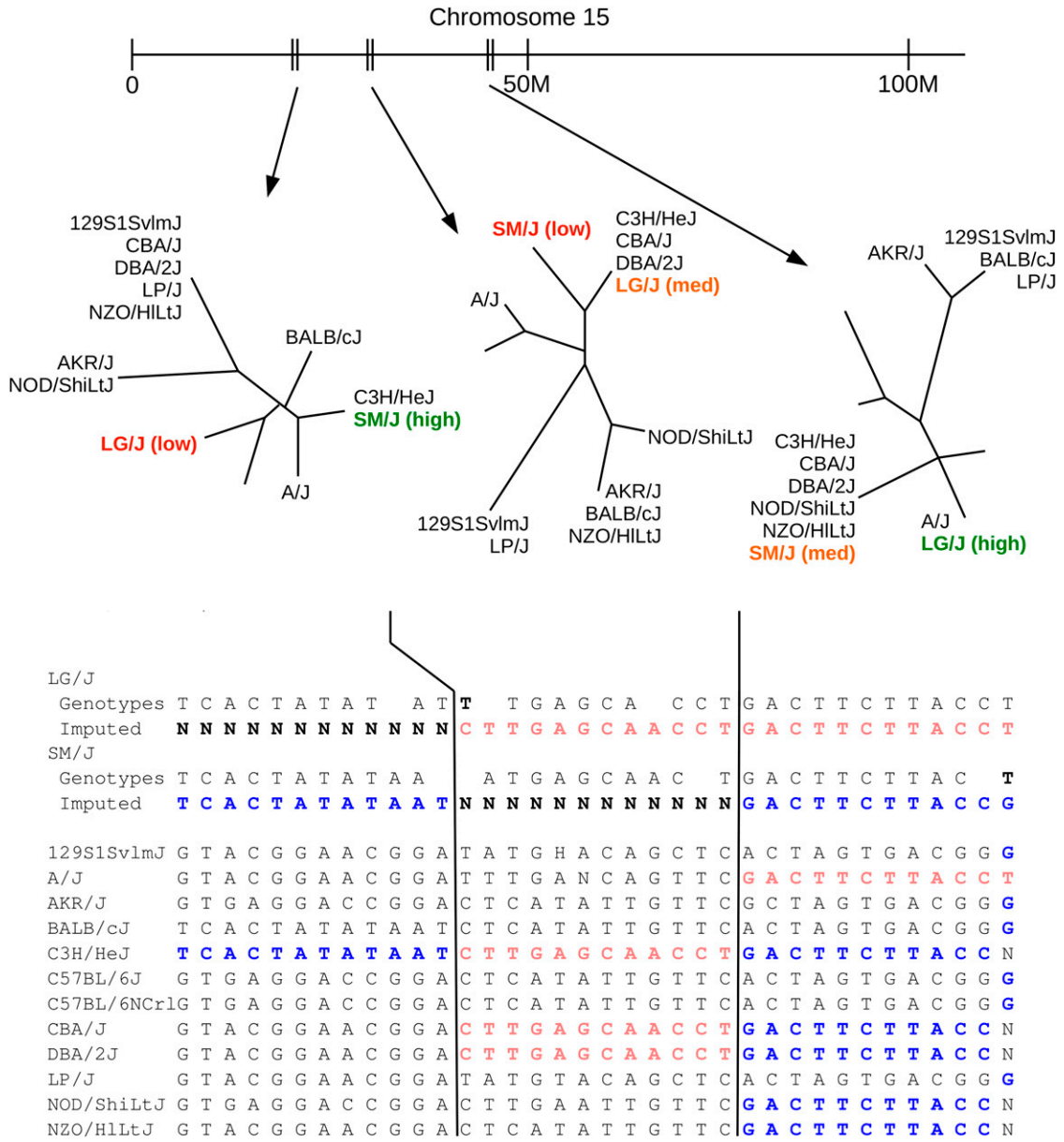

Figure 3 A representative example of our imputation method. Three trees on chromosome 15 (22.8–23.0 Mb, 31.3–31.4 Mb, and 45.1–45.5 Mb) that exhibit all three levels of confidence for LG/J and SM/J are shown. At the bottom, a sampling of SNPs in these regions is shown with the alleles contributing to the imputed sequences highlighted. Allele shown in blue contribute to SM/J, and those in pink contribute to LG/J.

chromosomes 14 and 15 obtained from strains LG/J and SM/J. We determined imputation accuracy by comparing genotypes over the intersection of our imputed SNP set (Sanger set SNPs) and the SNP sets in the external validation sequences.

Results and Discussion

We imputed 12 million SNPs over 88 samples for a total of 977 million new genotypes (Table S1). On average, we imputed 70.76% of the genotypes with high confidence and 21.36% with medium confidence. The remaining 7.88% were low confidence (Table 1). There is a wide range of variation in the fraction of SNPs imputed with different confidence among these 88 strains (Table 1). The greatest fraction of high-confidence SNPs is observed in substrains derived from a common inbred ancestor that differ only at loci harboring new mutations (97.00%) (Yang *et al.* 2011). These are followed by the 129T2/SvEmsJ and TSJ/LeJ strains that have very large fractions of their genome imputed at high confidence (94.21% and 93.29%, respectively). On the other hand, KK/HlJ and TALLYHO/JngJ have the smallest fraction of genome imputed at high confidence (39.52% and 42.10%, respectively). However, it is the CE/J strain that has the greatest fraction of the genome with low confidence (23.48%).

Table 1 Fraction of SNPs at different levels of confidence in 88 classical inbred strains

Strain	HC %	MC %	LC %
129P1/ReJ	92.54	4.05	3.41
129P3/J	90.13	6.35	3.52
129S6	90.97	6.90	2.13
129T2/SvEmsJ	94.21	4.29	1.51
129X1/SvJ	90.14	6.38	3.48
A/WySnJ	97.26	1.41	1.33
AEJ/GnLeJ	76.11	18.14	5.76
AEJ/GnRk	75.57	18.57	5.86
ALR/LtJ	53.46	35.68	10.86
ALS/LtJ	57.47	32.78	9.75
BALB/cByJ	97.63	1.27	1.11
BDP/J	58.81	26.78	14.41
BPH/2J	79.92	15.79	4.29
BPL/1J	70.77	23.24	6.00
BPN/3J	68.72	25.08	6.20
BTBRT+tf/J	71.04	21.71	7.26
BUB/BnJ	56.14	32.43	11.42
BXSB/MpJ	88.38	9.36	2.26
C3HeB/FeJ	92.39	6.07	1.54
C57BL/10J	85.72	12.44	1.85
C57BL/10ScNJ	85.63	12.53	1.85
C57BL/10ScSnJ	85.66	12.50	1.84
C57BL/6NCrl	98.98	1.01	0.02
C57BL/6NTac	98.98	1.01	0.02
C57BLKS/J	70.87	21.04	8.09
C57BR/cdJ	65.74	25.67	8.59
C57L/J	70.84	21.06	8.10
C58/J	67.74	23.85	8.42
CBA/CaJ	85.47	11.39	3.14
CE/J	46.04	30.48	23.48
CHMU/LeJ	79.16	17.97	2.88
DBA/1J	84.31	12.50	3.19
DBA/1LacJ	84.70	12.09	3.21
DBA/2DeJ	96.76	1.48	1.76
DBA/2HaSmnJ	76.99	18.95	4.06
DDK	46.12	37.57	16.31
DDY/JclSidSeyfrkJ	42.14	41.76	16.11
DLS/LeJ	80.43	15.43	4.13
EL/SuzSeyfrkJ	43.96	41.46	14.58
FVB/NJ	49.52	34.92	15.56
HPG/BmJ	88.64	9.12	2.24
I/LnJ	51.49	32.43	16.08
JE/LeJ	77.22	17.42	5.36
KK/HlJ	39.52	42.42	18.06
LG/J	60.78	27.65	11.57
LT/SvEiJ	78.25	15.70	6.05
MRL/MpJ	61.72	28.72	9.55
NON/LtJ	50.87	36.24	12.89
NONcNZO10/LtJ	46.36	40.08	13.56
NONcNZO5/LtJ	56.95	32.28	10.77
NOR/LtJ	84.40	12.19	3.42
NU/J	54.68	32.14	13.18
NZB/BINJ	58.76	29.49	11.75
NZL/LtJ	81.44	13.62	4.93
NZM2410/J	49.12	36.00	14.88
NZW/LacJ	49.09	35.58	15.33
P/J	58.39	27.11	14.50
PL/J	60.00	30.83	9.18
PN/nBSWUmaDJ	46.78	37.32	15.90
RF/J	66.97	25.02	8.01
RHJ/LeJ	82.57	14.57	2.86
RIIIS/J	45.13	37.46	17.41
RSV/LeJ	87.86	10.43	1.71
SB/LeJ	84.83	12.08	3.09
SEA/GnJ	72.31	20.11	7.58
SEC/1GnLeJ	82.88	11.96	5.17
SEC/1ReJ	83.13	11.72	5.15
SH1/LeJ	83.49	13.85	2.66
SI/ColTyrp1bDnahc11iv/J	75.42	19.96	4.62
SJL/Bm	53.84	32.52	13.65
SJL/J	53.89	32.49	13.62
SM/J	51.79	29.97	18.24
SSL/LeJ	82.21	13.46	4.34
ST/bJ	53.69	33.28	13.04
STX/Le	89.96	8.69	1.35
SWR/J	47.74	33.68	18.58
TALLYHO/JngJ	42.10	42.16	15.75
TKDU/DnJ	85.40	10.63	3.97
TSJ/LeJ	93.29	5.29	1.42
YBR/EiJ	47.36	37.29	15.35
ZRDCT Rax+/ChUmdJ	65.05	23.42	11.54
IBWSP2	67.83	25.23	6.94
IBWSR2	65.80	26.70	7.50
ICOLD2	70.84	22.58	6.59
IHOT1	70.31	24.01	5.68
IHOT2	69.29	24.49	6.22
ILS	76.18	16.99	6.83
ISS	74.25	20.01	5.74
Average	**70.76**	**21.36**	**7.88**
Standard deviation	**16.58**	**11.46**	**5.49**

Using a leave-one-out method, we can estimate our imputation accuracy using the Sanger set. In these samples, 1.05% of genotypes could not be imputed because the haplotype structure is unknown for some strains in regions due to deletion or copy-number variation (Yang *et al.* 2011). An additional 2.25% of genotypes are uncalled (N) in the Sanger set to which we compared the leave-one-out imputed genotypes and, therefore, cannot be verified. On average, we imputed 73.8 ± 19.1% of the remaining genotypes with high confidence and 18.0 ± 11.8% with medium confidence (Table 2). High-confidence genotypes had an average error rate of 0.083 ± 0.019% and the error rate for medium-confidence genotypes was 1.57 ± 0.75%. The remaining 8.2 ± 7.5% were low confidence. In these regions, methods based on the haplotype similarity and phylogeny trees performed no better than a consensus sequence among all strains in the Sanger set because there is no single representative haplotype to use for imputation. The consensus genotypes had an error rate of 44.5 ± 10.1% compared to the Sanger genotypes. Since our accuracy in low-confidence regions is little better than chance, we do not impute these genotypes and indicate them by N's.

In addition to leave-one-out analysis, we performed validation with the chromosomes 14 and 15 sequence data for LG/J and SM/J, two strains that are not included in the Sanger set. We compared imputed genotypes on chromosomes 14 and 15, containing 1,316,845 imputed Sanger

Table 2 Fraction of the genome imputed and error in leave-one-out imputation

Strain	HC %	HC error %	MC %	MC error %	LC %	LC error %
129S1SvlmJ	76.83	0.07	15.35	1.74	7.82	43.75
A/J	81.26	0.09	13.47	2.30	5.27	44.98
AKR/J	58.50	0.11	29.98	1.85	11.52	44.18
BALB/cJ	82.07	0.08	13.39	1.86	4.55	37.81
C3H/HeJ	85.29	0.09	11.92	0.82	2.79	53.76
C57BL/6J	97.05	0.08	2.89	0.54	0.06	33.87
C57BL/6NCrl	97.21	0.04	2.77	0.13	0.02	19.94
CBA/J	81.91	0.08	13.56	1.52	4.54	52.73
DBA/2J	66.46	0.10	22.54	1.60	11.00	50.67
LP/J	79.05	0.07	14.32	1.76	6.63	55.25
NOD/ShiLtJ	42.00	0.10	38.60	1.84	19.40	45.85
NZO/HlLtJ	38.04	0.10	37.12	2.83	24.84	51.20
Average	**73.81**	**0.08**	**17.99**	**1.57**	**8.20**	**44.50**
SD	**18.32**	**0.018**	**11.30**	**0.71**	**7.20**	**9.64**

HC, high confidence; MC, medium confidence; LC, low confidence.

Table 3 Fraction of the genome imputed and error in external validation of imputed genotypes

Strain	HC %	HC error %	MC %	MC error %
LG/J	65.58	0.08	25.20	4.43
SM/J	67.90	0.07	21.25	3.14
Average	**66.74**	**0.075**	**23.23**	**3.79**
SD	**1.16**	**0.005**	**1.98**	**0.65**

SNPs, with high-density genotypes for LG/J and SM/J containing 292,051 and 416,589 SNPs, respectively. For SM/J, there were 362,362 SNPs in common with our imputed genotypes. Of these, 67.90% were imputed with high confidence and had an error rate of 0.07% (Table 3). An additional 21.25% were imputed with medium confidence at an error rate of 3.14%. The remaining were low confidence. Similar results were found for LG/J (Table 3). We could increase the number of markers available for validation if we included all loci in our imputed genotypes and assume that these markers have the reference allele in our validation sequences where a SNP is not present. We do not do this because the SNP density is much lower in the validation sets than in our imputed sequences, so there are likely many unreported SNPs in LG/J and SM/J at the resolution of our imputed sequence. While LG/J has 292,051 SNPs and SM/J has 416,589 SNPs in the validation genotypes, any pair of two strains in the Sanger SNP sets has, on average, 908,493 SNPs across chromosomes 14 and 15.

We achieved very low error rates in regions of high and medium confidence with both validation approaches. Our imputation method provides several improvements over existing techniques and resources. Our methods outperform previously published imputation methods in regions in which we can impute with high confidence. In addition, we identify regions that cannot be accurately imputed with our samples because they do not share common haplotypes based on local phylogeny. Furthermore, our analysis can inform the selection of strains for which full genomic sequence would substantially improve the ability to confidently impute other strains and the optimal strains for maximal SNP discovery.

Kirby *et al.* (2010) imputed 657 million genotypes over 94 strains consisting of 65 classical and 13 wild-derived low-density sequenced strains and 12 classical and 4 wild-derived high-density sequenced strains. They imputed the 78 low-density sequences (121,433 SNPs) with the high-density NIEHS/Perlegen data set (8.27 million SNPs) (Frazer *et al.* 2007) and missing genotypes in the 16 high-density sequences. In addition to including wild-derived strains in the sample set, 64% of SNPs in the full Perlegen data set include private alleles seen only in wild-derived strains. Szatkiewicz *et al.* (2008) imputed 269 million genotypes using a cleaned subset of 7.9 million NIEHS/Perlegen SNPs over 51 strains including 39 classical and 12 wild derived. These previous imputation efforts attempted to impute a mixture of classical laboratory strains and wild-derived strains including SNPs with private alleles in wild-derived samples. Including wild-derived strains contributes many SNPs that are nonvarying among classical strains and results in inflated estimates of accuracy because many of the imputed variants are actually constant within the classical population. Here we exclude wild-derived strains because they exhibit few variants seen among classical strains and many variants not seen in any classical strain. We have imputed significantly more SNPs that segregate among classical inbred strains. In addition, our estimated error rate in high-confidence regions is 0.083% compared to 0.27% and 4.4% reported in previous studies. Our error rate in high-confidence regions is in line with sequencing error and the rate of recurrent mutations at highly mutable sites (homoplasy). These improvements are due, in part, to the use of a more complete set of sequence data and the fact that our lower-density set, at >500,000 markers, is considerably denser than in previous studies. Furthermore, the MDA platform, designed specifically to highlight the diversity within our sample set, is better equipped to identify appropriate imputation genotypes than the sparser ~135,000-marker Broad SNP set (Kirby *et al.* 2010) used in previous efforts.

Due to the different SNP and strain sets used for imputation in our work compared with previous imputation methods, it is useful to analyze differences against a common set of sequence data. We identified 174,891 SNPs common

Table 4 Performance of imputation methods based on fraction of SNPs imputed at high confidence (HC) and fraction of error in chromosomes 14 and 15

	This study	Kirby *et al.* (2010)	Szatkiewicz *et al.* (2007)
LG/J HC %	60.91	65.88	73.07
LG/J HC error %	0.06	3.75	6.57
SM/J HC %	64.95	57.26	72.39
SM/J HC error %	0.03	1.60	5.76

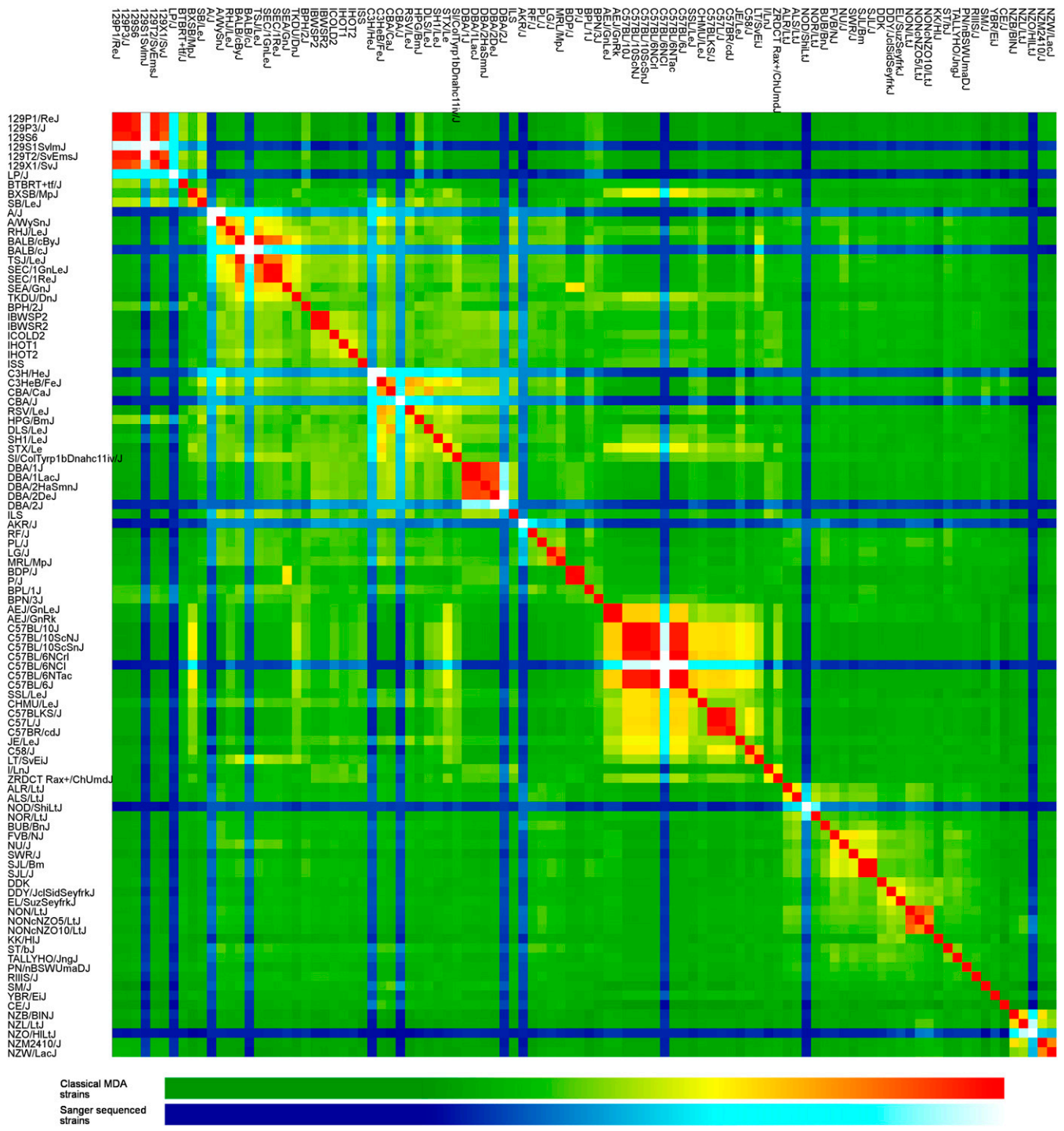

Figure 4 Frequency of leaf/haplotype sharing within the set of 100 Mouse Diversity Array genotyped samples is shown as a heat map. Green to red intensity colors indicate similarity among only classical MDA strains. Blue to white colors indicate similarity with and between strains in the Sanger set.

to our imputed genotypes, previous imputation results, and our LG/J and SM/J validation genotypes on chromosomes 14 and 15 against which we can directly compare (Table 4). In SM/J, our method imputed 64.95% of SNPs with high confidence, Kirby *et al.* (2010) imputed 57.26% with high confidence (they define this as posterior probability >0.98), and Szatkiewicz *et al.* (2008) imputed 72.39% with high confidence (posterior probability >0.9). Using our validation genotypes as the ground truth, our method achieved a per-SNP error rate of 0.03% while previous methods achieved error rates of 1.60% and 5.76%, respectively (Table 4).

Our imputation method highlights an important feature of the imputed MDA sample set. Our method of haplotype

Table 5 Strains contributing the most high-confidence (HC) genotypes unrepresented in the Sanger set

Strain	HC genotypes	% increase
SWR/J	12,873,012	1.21
SJL/Bm	12,806,101	1.21
SJL/J	12,699,148	1.20
BDP/J	12,592,551	1.19
P/J	12,497,039	1.18
DDY/JclSidSeyfrkJ	12,002,104	1.13
FVB/NJ	11,985,984	1.13
DDK	11,892,009	1.12
KK/HlJ	11,546,347	1.09
I/LnJ	11,497,933	1.08

identification and assignment is based on the notion that classical laboratory mice are derived from a small set of recent common ancestors. To impute missing genotypes, we identify intervals with no evidence of ancestral recombination, in which shared haplotype ancestry can be assumed (Yang *et al.* 2011), and identify sequenced strains that share these haplotypes with strains that are to be imputed. In some cases, no evidence exists of a shared haplotype with a sequenced strain; these we consider low confidence (Figure 3). Since these intervals are not derived from a haplotype common to one of our high-density sequences, no method can produce accurate genotypes given the data on hand. This feature allows us to suggest a method for improving our imputation power by identifying those sequences that share haplotypes unrepresented in our high-density genotypes.

We have identified strains that would provide the greatest improvement in imputation accuracy as those that share the greatest number and size of unrepresented haplotype blocks with the greatest number of other strains. These intervals are currently identified as low confidence. If we had whole-genome sequence for even one sample with the shared haplotype in these intervals (Figure 4), we could impute these genotypes with high confidence. In other words, the discriminating function can be described as the greatest total number of genotypes changed from low confidence to high confidence by introducing a new fully sequenced sample into these haplotypes (Table 5). The strain that would contribute the greatest number of new high-confidence genotypes is SWR/J, which would contribute an additional 12,973,012 high- and medium-confidence genotypes, reducing overall low confidence regions from 7.88% to 6.65%.

Imputation, by its nature, cannot increase the number of SNPs since all imputed genotypes are derived from existing sequence. However, using our local phylogenies, we can identify samples that would contribute the greatest discovery of additional sequence variation. While high-confidence imputation power is related to the haplotype group membership (Figure 4) in the local phylogenetic structure, the level of sequence variation can be inferred from the edge length in the local phylogenetic trees. The edge lengths in our local phylogenetic trees are derived from the sequence differences in each compatible interval. The longer the edge, the further a leaf/haplotype is from the high-density samples and the greater sequence variation we expect to see in the unrepresented sequence. This is unlike the metric for imputation power since the haplotype frequency and sharing with other samples are not relevant. The sample that would contribute the greatest additional sequence variation is KK/HlJ. This strain has the most unrepresented sequence variation, with an average sequence variation of 5.55% from the nearest Sanger sequenced sample across the genome. The top 10 candidates and their sequence variation are shown in Table 6.

We deliberately omitted wild-derived strains from our imputation since wild-derived strains do not share a recent history with the classical laboratory strains (Beck *et al.* 2000). We do not have sufficient marker density to catalog the variants among wild-derived strains and they are likely so widely divergent that a local phylogeny cannot be constructed. However, using wild-derived strains sequenced by the Sanger Institute to impute classical inbred strains could potentially improve our current design in regions of contamination in the wild-derived strains [*i.e.*, CAST/EiJ and PWK/PhJ (Yang *et al.* 2011)] or in putative regions in which a few classical strains share a haplotype that was rare or absent in fancy mice. We conclude that this could be worth doing but is unlikely to have significant impact and is not consistent with our tree-based approach. Imputing wild-derived strains would be of little value as only a single wild-derived strain from each species or subspecies would be used to impute SNPs of all additional strains in the same taxon. Our recent work (Yang *et al.* 2011) demonstrates that there is far more sequence variation among wild-derived than classical mice, making the imputation an exercise in futility.

Our imputation model can be further fine-tuned to better identify the appropriate intervals over which we assign a local phylogenetic structure. The optimal haplotype blocks should be small enough to accurately represent only a single indivisible haplotype but large enough to capture all appropriate variation in these haplotypes. As more samples are incorporated into the haplotype derivation model, it will be especially important to accurately represent the structure and how it relates to high-density sequenced samples.

Table 6 Strains with the highest sequence variation unrepresented in the Sanger set

Strain	% unrepresented variation
KK/HlJ	5.55
NZM2410/J	3.47
EL/SuzSeyfrkJ	3.30
DDK	3.09
DDY/JclSidSeyfrkJ	3.08
BDP/J	2.79
P/J	2.77
SWR/J	2.57
SJL/Bm	2.22
SJL/J	2.22

Imputed genotypes are evolving resources [Szatkiewicz *et al.* 2008 (http://cgd.jax.org/datasets/popgen/imputed.shtml); Kirby *et al.* 2010 (http://mouse.cs.ucla.edu/mousehapmap/)]. As the sequencing pace increases for mouse, the need for a codified and coherent resource will be even more important. Using our method of imputation, additional full-genome sequences can be easily incorporated and will further improve the imputation accuracy (Figure 3). We plan to include LG/J and SM/J as soon as their full-genome sequence is released. Since our model explicitly takes advantage of local phylogeny and haplotype structure and accounts for multiple instances of a single derived haplotype, we can incorporate multiple and varying sequences representing a single sample or strain. This will help consolidate and derive a consensus from possibly discordant sources. Our phylogeny modeling can be extended to include classical strains genotyped with the MDA, such as the upcoming Collaborative Cross strains (Collaborative Cross Consortium 2012), to provide an even larger and more diverse resource of imputed genotypes. The method we present here is not limited only to the laboratory mouse, but could be extended to any organism for which inbred populations exist, such as rats and dogs, as well as many plant species. Imputation in any species will be most effective when inbred populations are derived from a common and relatively small set of ancestral populations.

Acknowledgments

This work was supported by National Institute of General Medical Sciences Centers of Excellence in Systems Biology program, grant GM-076468. Sequencing of LG/J and SM/J was supported by National Institutes of Health grant R01 AR053224 and National Institute of Diabetes and Digestive and Kidney Diseases grant DK75112.

Literature Cited

Beck, J. A., S. Lloyd, M. Hafezparast, M. Lennon-Pierce, J. T. Eppig *et al.*, 2000 Genealogies of mouse inbred strains. Nat. Genet. 24: 23–25.

Church, D. M., L. Goodstadt, L. W. Hillier, M. C. Zody, S. Goldstein *et al.*, 2009 Lineage-specific biology revealed by a finished genome assembly of the mouse. PLoS Biol. 7: e1000112.

Collaborative Cross Consortium, 2012 The genome architecture of the Collaborative Cross mouse genetic reference population. Genetics 190: 389–401.

Ding, L., M. J. Ellis, S. Li, D. E. Larson, K. Chen *et al.*, 2010 Genome remodelling in a basal-like breast cancer metastasis and xenograft. Nature 464: 999–1005.

Frazer, K. A., E. Eskin, H. M. Kang, M. A. Bogue, D. A. Hinds *et al.*, 2007 A sequence-based variation map of 8.27 million SNPs in inbred mouse strain. Nature 448: 1050–1053.

Hudson, R. R., and N. L. Kaplan, 1985 Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics 111: 147–164.

Kang, H. M., N. A. Zaitlen, and E. Eskin, 2010 EMINIM: an adaptive and memory-efficient algorithm for genotype imputation. J. Comput. Biol. 17: 547–560.

Keane, T. M., L. Goodstadt, P. Danecek, M. A. White, K. Wong *et al.*, 2011 Mouse genomic variation and its effect on phenotypes and gene regulation. Nature 477: 289–294.

Kirby, A., H. M. Kang, C. M. Wade, C. Cotsapas, E. Kostem *et al.*, 2010 Fine mapping in 94 inbred mouse strains using a high-density haplotype resource. Genetics 185: 1081–1095.

Li, H., J. Ruan, and R. Durbin, 2008 Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. 18: 1851–1858.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan *et al.*, 2009 The sequence alignment/map format and SAMtools. Bioinformatics 25: 2078–2079.

Mardis, E. R., L. Ding, D. J. Dooling, D. E. Larson, M. D. McLellan *et al.*, 2009 Recurring mutations found by sequencing an acute myeloid leukemia genome. N. Engl. J. Med. 361: 1058–1066.

Sherry, S. T., M. H. Ward, M. Kholodov, J. Baker, L. Phan *et al.*, 2001 dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 29: 308–311.

Szatkiewicz, J. P., G. L. Beane, Y. Ding, L. Hutchins, F. Pardo-Manuel de Villena *et al.*. 2008 An imputed genotype resource for the laboratory mouse. Mamm. Genome 19: 199–208.

Wang, J., K. J. Moore, Q. Zhang, F. Pardo-Manuel de Villena, W. Wang *et al.*. 2010 Genome-wide compatible SNP intervals and their properties. Proceedings of the First ACM International Conference on Bioinformatics and Computational Biology. Association for Computing Machinery, Niagara Falls, NY.

Waterston, R. H., K. Lindblad-Toh, E. Birney, J. Rogers, J. F. Abril *et al.*, 2002 Initial sequencing and comparative analysis of the mouse genome. Nature 420: 520–562.

Yang, H., T. A. Bell, G. A. Churchill, and F. Pardo-Manuel de Villena, 2007 On the subspecific origin of the laboratory mouse. Nat. Genet. 39: 1100–1107.

Yang, H., Y. Ding, L. N. Hutchins, J. Szatkiewicz, T. A. Bell *et al.*, 2009 A customized and versatile high-density genotyping array for the mouse. Nat. Methods 6: 663–666.

Yang, H., J. Wang, J. Didion, R. J. Buus, T. A. Bell *et al.*, 2011 Subspecific origin and haplotype diversity in the laboratory mouse. Nat. Genet. 43(648): 655.

Edited by Lauren M. McIntyre, Dirk-Jan de Koning, and 4 dedicated Associate Editors

GENETICS

Supporting Information
http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.111.132381/-/DC1

Imputation of Single-Nucleotide Polymorphisms in Inbred Mice Using Local Phylogeny

Jeremy R. Wang, Fernando Pardo-Manuel de Villena, Heather A. Lawson, James M. Cheverud, Gary A. Churchill, and Leonard McMillan

DOI: 10.1534/genetics.111.132381

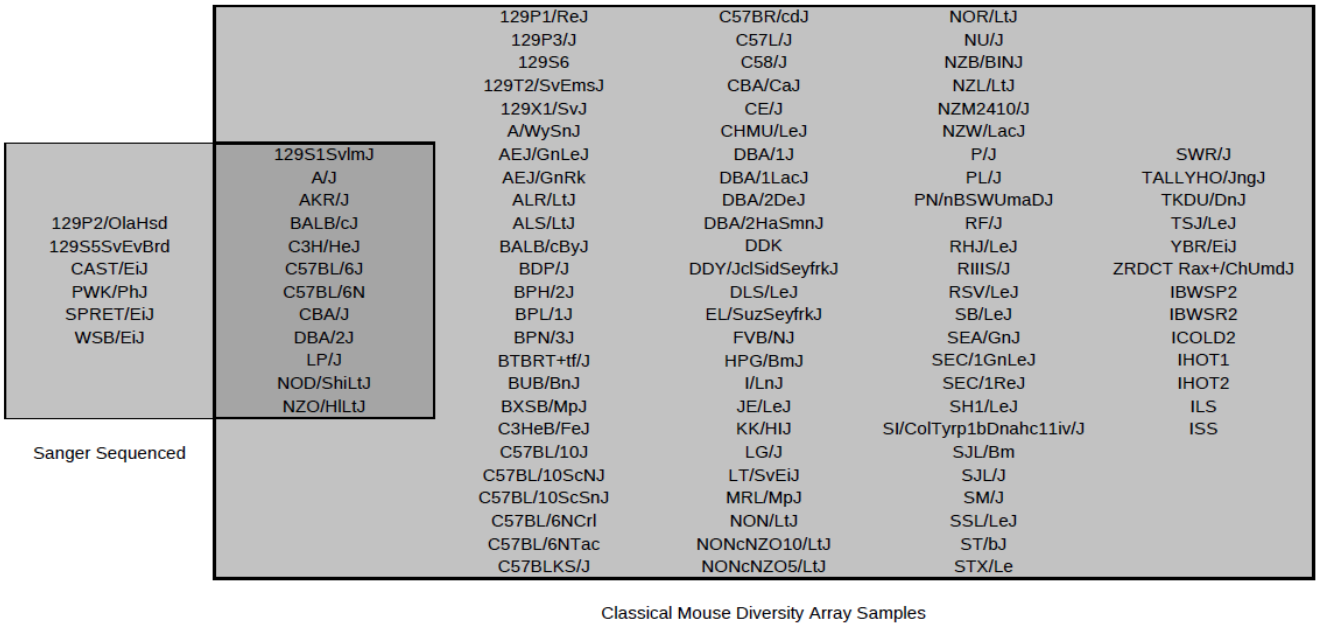

Figure S1 The set of mouse strains genotyped at medium-density on the Mouse Diversity Array are shown on the right. The strains in the Sanger set are shown on the left. The overlapping set we used as the source of high-density genotypes in our imputation.

Table S1 Imputed genotypes and confidence scores for 88 classical laboratory mouse strains.

Table S1 is available at http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.111.132381/-/DC1 for download as a compressed file.

Positions are based on NCBI build 37. Alleles are listed for each strain, followed by the confidence value, 0, 1, or 2 indicating low, medium, and high confidence, respectively. N indicates no call in the case of low confidence or unknown genotypes in the imputing sequence. Use of these data should cite this publication as a reference. These data can also be downloaded at http://www.csbio.unc.edu/imputation/imputed_88_20111013.zip.