Title: Rapid, Reference-Free Human Genotype Imputation with Denoising Autoencoders

Authors: Raquel Dias[1,2], Doug Evans[1,2], Shang-Fu Chen[1,2], Kai-Yu Chen[1,2], Leslie Chan[1,2], Ali Torkamani[1,2,*]

[1] Scripps Research Translational Institute, Scripps Research, La Jolla, CA, 92037, USA

[2] Department of Integrative Structural and Computational Biology, Scripps Research, La Jolla, CA, 92037, USA

* **Corresponding author:**

Ali Torkamani, Ph.D.

3344 North Torrey Pines Court, Suite 300

La Jolla, CA 92037

Phone: 858-784-2082

atorkama@scripps.edu

Abstract

Genotype imputation is a foundational tool for population genetics. Standard statistical imputation approaches rely on the co-location of large whole-genome sequencing-based reference panels, powerful computing environments, and potentially sensitive genetic study data. This results in computational resource and privacy-risk barriers to access to cutting-edge imputation techniques. Moreover, the accuracy of current statistical approaches is known to degrade in regions of low and complex linkage disequilibrium.

Artificial neural network-based imputation approaches may overcome these limitations by encoding complex genotype relationships in easily portable inference models. Here we demonstrate an autoencoder-based approach for genotype imputation, using a large, commonly-used reference panel, and spanning the entirety of human chromosome 22. Our autoencoder-based genotype imputation strategy achieved superior imputation accuracy across the allele-frequency spectrum and across genomes of diverse ancestry, while delivering at least 4-fold faster inference run time relative to standard imputation tools.

Introduction

The human genome is inherited in large blocks from parental genomes, generated through a DNA-sequence-dependent shuffling process called recombination. The non-random nature of recombination breakpoints producing these genomic blocks results in correlative genotype relationships across genetic variants, known as linkage disequilibrium. Thus, genotypes for a small subset (1% – 10%) of observed common genetic variants can be used to infer the genotype status of unobserved but known genetic variation sites across the genome (on the order of ~1M of >10M sites) (Li et al., 2009; Marchini and Howie, 2010). This process, called genotype imputation, allows for the generation of nearly the full complement of known common genetic variation at a fraction of the cost of direct genotyping or sequencing. Given the massive scale of genotyping required for genome-wide association studies or implementation of genetically-informed population health initiatives, genotype imputation is an essential approach in population genetics.

Standard approaches to genotype imputation utilize Hidden Markov Models (HMM) (Browning et al., 2018; Das et al., 2016a; Rubinacci et al., 2020) distributed alongside large WGS-based reference panels (Browning and Browning, 2016). In general terms, these imputation algorithms use genetic variants shared between to-be-imputed genomes and the reference panel and apply Hidden Markov Models (HMM) to impute the missing genotypes per sample (Das et al., 2018). Genotyped variants are the observed states of the HMM, whereas the to-be-imputed genetic variants present in the reference panel are the hidden states. The HMM parameter function depends on recombination rates, mutation rates, and/or genotype error rates that must be fit by Markov Chain Monte Carlo Algorithm (MCMC) or an expectation-maximization algorithm. Thus, HMM-based imputation is a computationally intensive process, requiring access to both high-performance computing environments and large, privacy-sensitive, WGS reference panels (Kowalski et al., 2019). Often, investigators outside of large consortia will resort to submitting genotype data to imputation servers (Das et al., 2016a), resulting in privacy and scalability concerns (Sarkar et al., 2021).

Recently, artificial neural networks, especially autoencoders, have attracted attention in functional genomics for their ability to fill-in missing data from genomic assays with significant dropout events, like single-cell RNAseq and ChIP-seq (Arisdakessian et al., 2019; Koh et al., 2017; Lal et al., 2021). Autoencoders are neural networks tasked with the problem of simply reconstructing the original input data, with constraints applied to the network architecture or transformations applied to the input data in order to achieve a desired goal like dimensionality reduction or compression, and de-noising or de-masking (Abouzid et al., 2019; Liu et al., 2020; Voulodimos et al., 2018), stochastic noise or masking is used to modify or remove data inputs, training the autoencoder to reconstruct the original uncorrupted data from corrupted inputs (Tian et al., 2020). These autoencoder characteristics are well-suited for genotype imputation and may address some of the limitations of HMM-based imputation by eliminating the need for dissemination of reference panels and allowing the capture of non-linear relationships in genomic regions with complex linkage disequilibrium structures. Some attempts at genotype imputation using neural networks have been previously reported, though for specific genomic contexts (Naito et al., 2021) at genotype masking levels (5% – 20%) not applicable in typical real-world population genetics scenarios (Chen and Shi, 2019; Islam et al., 2021; Kojima et al., 2020; Sun and Kardia, 2008).

Here we present a generalized approach to unphased human genotype imputation using sparse, denoising autoencoders capable of highly accurate genotype imputation at genotype masking levels (98+%) appropriate for array-based genotyping and low-pass sequencing-based population genetics initiatives. We describe the initial training and implementation of autoencoders spanning all of human chromosome 22, achieving equivalent to superior accuracy relative to modern HMM-based methods, and dramatically improving computational efficiency at deployment without the need to distribute reference panels.

Materials and Methods

Overview

Sparse, de-noising autoencoders spanning all bi-allelic SNPs observed in the Haplotype Reference Consortium were developed and optimized. Each bi-allelic SNP was encoded as two binary input nodes, representing the presence or absence of each allele (**Figure 1A, 1D**). This encoding allows for the straightforward extension to multi-allelic architectures and non-binary allele presence probabilities. A data augmentation approach using modeled recombination events and offspring formation coupled with random masking at an escalating rate drove our autoencoder training strategy (**Figure 1B**). Because of the extreme skew of the allele frequency distribution for rarely present alleles (Auton et al., 2015), a focal-loss-based approach was essential to genotype imputation performance. The basic architecture of the template fully-connected autoencoder before optimization to each genomic segment is depicted in **Figure 1C**. Individual autoencoders were designed to span genomic segments with boundaries defined by computationally identified recombination hotspots (**Figure 1E**). The starting point for model hyperparameters were randomly selected from a grid of possible combinations and were further tuned from a battery of features describing the complexity of the linkage-disequilibrium structure of each genomic segment.

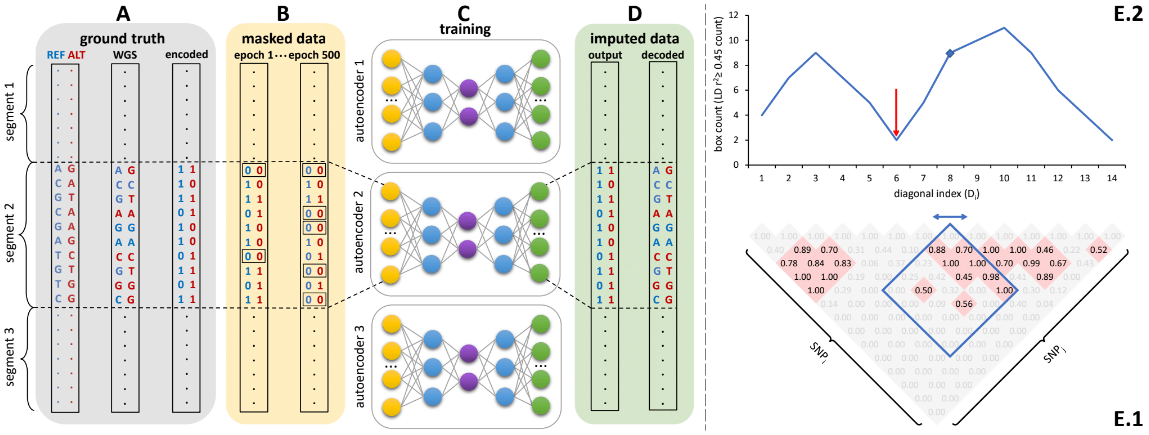


Figure 1. Schematic overview of the autoencoder training workflow. A) Ground truth whole genome sequencing data is encoded as binary values representing the presence (1) or absence

(0) of the reference allele (blue) and alternative allele (red). **B)** Variant masking (setting both alleles as absent, represented by 0) corrupts data inputs at a gradually increasing masking rate. Example masked variants are outlined. C) Fully-connected autoencoders spanning segments defined as shown in panel **E**, are then trained to reconstruct the original uncorrupted data from corrupted inputs; **D)** the reconstructed outputs (imputed data) are compared to the ground truth states for loss calculation and are decoded back to genotypes. **E)** Tiling of autoencoders across the genome is achieved by **E.1)** calculating a *n x n* matrix of pairwise SNP correlations, thresholding them at 0.45 (selected values are shown in red background, excluded values in gray), **E.2)** quantifying the overall local LD strength centered at each SNP by computing their local correlation box counts and splitting the genome into approximately independent segments by identifying local minima (recombination hotspots). The red arrow illustrates minima between strong LD regions.

Genotype Encoding

Genotypes for all bi-allelic SNPs were converted to binary values representing the presence (1) or absence (0) of the reference allele A and alternative allele B, respectively, as shown in **Equation 1**.

$$x_i = \begin{Bmatrix} if(\mathrm{G}_i = [A,A]): & x_i = [1,0] \\ if(\mathrm{G}_i = [A,B]): & x_i = [1,1] \\ if(\mathrm{G}_i = [B,A]): & x_i = [1,1] \\ if(\mathrm{G}_i = [B,B]): & x_i = [0,1] \\ if(\mathrm{G}_i = [null]): & x_i = [0,0] \end{Bmatrix} \qquad (1)$$

Where *x* is a vector containing the two allele presence input nodes to the autoencoder and their encoded allele presence values derived from the original genotype, *G,* of variant *i*. The output nodes of the autoencoder, regardless of activation function, are similarly rescaled to 0 - 1. The scaled outputs can also be regarded as probabilities and can be combined for the calculation of alternative allele dosage and/or genotype probabilities. This representation maintains the interdependencies among classes, is extensible to other classes of genetic variation, and allows for the use of probabilistic loss functions.

Training Data, Masking, and Data Augmentation

Training Data. Whole-genome sequence data from the Haplotype Reference Consortium (HRC) was used for training and as the reference panel for comparison to HMM-based imputation (McCarthy et al., 2016). The dataset consists of 27,165 samples and 39,235,157 biallelic SNPs generated using whole-genome sequence data from 20 studies of predominantly European ancestry (HRC Release 1.1): 83.92% European, 2.33% East Asian, 1.63% Native American, 2.17% South Asian, 2.96% African, and 6.99% admixed ancestry individuals. Genetic ancestry was determined using continental population classification from the 1000 Genomes Phase3 v5 (1000G) reference panel and a 95% cutoff using Admixture software (Alexander et al., 2009). Genotype imputation autoencoders were trained for all 510,442 unique SNPs observed in HRC on human chromosome 22.

Validation and Testing Data. A balanced (50%:50% European and African genetic ancestry) subset of 796 whole genome sequences from the Atherosclerosis Risk in Communities cohort (ARIC) (Mou et al., 2018), was used for model validation and selection. The Wellderly (Erikson et al., 2016), Human Genome Diversity Panel (HGDP) (Cann, 2002), and Multi-Ethnic Study of Atherosclerosis (MESA) (Bild, 2002) cohorts were used for model testing. The Wellderly cohort consisted of 961 whole genomes of predominantly European genetic ancestry. HGDP consisted of 929 individuals across multiple ancestries: 11.84% European, 14.64% East Asian, 6.57% Native American, 10.98% African, and 55.97% admixed. MESA consisted of 5,370 whole genomes across multiple ancestries: 27.62% European, 11.25% East Asian, 4.99% Native American, 5.53% African, and 50.61% admixed.

GRCh38 mapped cohorts (HGDP and MESA) were converted to hg19 using Picard v2.25 (“Picard toolkit,” 2019). All other datasets were originally mapped and called against hg19. Multi-allelic SNPs, SNPS with >10% missingness, and SNPs not observed in HRC were removed with bcftools v1.10.2 (Danecek et al., 2021). Mock genotype array data was generated from these WGS cohorts by restricting genotypes to those present on commonly used genotyping arrays (Affymetrix 6.0, UKB Axiom, and Omni 1.5M). For chromosome 22, intersection with HRC and this array-like masking respectively resulted in: 9,025, 10,615, and 14,453 out of 306,812 SNPs observed in ARIC; 8,630, 10,325, and 12,969 out of 195,148 SNPs observed in the Wellderly; 10,176, 11,086, and 14,693 out of 341,819 SNPs observed in HGDP; 9,237, 10,428, and 13,677 out of 445,839 SNPs observed in MESA.

Data Augmentation. We employed two strategies for data augmentation – random variant masking and simulating further recombination with offspring formation. During training, random masking of input genotypes was performed at escalating rates, starting with a relatively low masking rate (80% of variants) that is gradually incremented in subsequent training rounds until up to only 5 variants remain unmasked per autoencoder. Masked variants are encoded as the *null* case in **Equation 1.** During finetuning we used sim1000G (Dimitromanolakis et al., 2019) to simulate of offspring formation using the default genetic map and HRC genomes as parents. A total of 30,000 offspring genomes were generated and merged with the original HRC dataset, for a total of 57,165 genomes.

Loss Function

In order to account for the overwhelming abundance of rare variants, the accuracy of allele presence reconstruction was scored using an adapted version of focal loss (*FL*) [32], shown in **Equation 2**.

$$FL = -\alpha_t(1 - p_t)^{\gamma}\,[x_t \log(p_t) + (1 - x_t)\log(1 - p_t)] \qquad (2)$$

Where the classic cross entropy (shown as binary log loss in brackets) of the truth class (x_t) predicted probability (p_t) is weighted by the class imbalance factor α_t and a modulating factor (1 - p_t)$^{\gamma}$. The modulating factor is the standard focal loss factor with hyperparameter, γ, which amplifies the focal loss effect by down-weighting the contributions of well-classified alleles to the overall loss (especially abundant reference alleles for rare variant sites). α_t is an additional balancing hyperparameter set to the truth class frequency.

This base focal loss function is further penalized and regularized to encourage simple and sparse models in terms of edge-weight and hidden layer activation complexity. These additional penalties result in our final loss function as shown in **Equation 3**.

$$SFL = -\alpha_t(1-p_t)^{\gamma}\left[x_t \log(p_t) + (1-x_t)\log(1-p_t)\right] + \beta S_{(\rho||\hat{\rho})} + \lambda_1 L1 + \lambda_2 L2 \quad (3)$$

Where *L*1 and *L*2 are the standard *L*1 and *L*2 norms of the autoencoder weight matrix, with their contributions mediated by the hyperparameters λ_1 and λ_2. S is a sparsity penalty, with its contribution mediated by the hyperparameter β, which penalizes deviation from a target hidden node activation set by the hyperparameter (ρ) vs the observed mean activation $\hat{\rho}$ over a training batch *j* summed over total batches *n*, as shown in Equation 4:

$$S_{(\rho||\hat{\rho})} = \sum_{j=1}^{n} \rho * log\left(\frac{\rho}{\hat{\rho}_j}\right) + (1-\rho) * log\left(\frac{1-\rho}{1-\hat{\rho}_j}\right) \quad (4)$$

<u>Genome Tiling</u>

All model training tasks were distributed across a diversified set of NVIDIA graphical processing units (GPUs) with different video memory limits: 5x Titan Vs (12GB), 8x A100s (40GB), 60x V100s (32GB). Given computational complexity and GPU memory limitations, individual autoencoders were designed to span approximately independent genomic segments with boundaries defined by computationally identified recombination hotspots (**Figure 1E**). These segments were defined using an adaptation of the LDetect algorithm [33]. First, we calculated a *n x n* matrix of pairwise SNP correlations using all common genetic variation (≥5% minor allele frequency) from HRC. Correlation values were thresholded at 0.45. For each SNP, we calculated a box count of all pairwise SNP correlations spanning 500 common SNPs upstream and downstream of the index SNP. This moving box count quantifies the overall local LD strength centered at each SNP. Local minima in this moving box count were used to split the genome into approximately independent genomic segments of two types – large segments of high LD interlaced with short segments of weak LD corresponding to recombination hotspot regions. Individual autoencoders were designed to span the entirety of a single high LD segment plus its adjacent upstream and downstream weak LD regions. Thus, adjacent autoencoders overlap at their weak LD ends. If an independent genomic segment exceeded the threshold number of SNPs amenable to deep learning given GPU memory limitations, internal local minima within the high LD regions were used to split the genomic segments further to a maximum of 6000 SNPs per autoencoder. Any remaining genomic segments still exceeding 6000 SNPs were further split into 6000 SNP segments with large overlaps of 2500 SNPs given the high degree of informative LD split across these regions. This tiling process resulted in 256 genomic segments: 188 independent LD segments, 32 high LD segments resulting from internal local minima splits, and 36 segments further split due to GPU memory limitations.

Hyperparameter Initialization and Grid Search

We first used a random grid search approach to define initial hyperparameter combinations producing generally accurate genotype imputation results. The hyperparameters and their potential starting values are listed in **Table 1**. This coarse-grain grid search was performed on all genomic segments of chromosome 22 (256 genomic segments), each tested with 100 randomly selected hyperparameter combinations per genomic segment, with a batch size of 256 samples, training for 500 epochs without any stop criteria, and validating on an independent dataset (ARIC). To evaluate the performance of each hyperparameter combination, we calculated the average coefficient of determination (r-squared) comparing the predicted and observed alternative allele dosages per variant. Concordance and F1-score were also calculated to screen for anomalies but were not ultimately used for model selection.

Table1. Description and values of hyperparameters tested in grid search.

Hyperparameter description	Tested values (coarse-grid search)
λ_1 for L1 regularization	[1e-3, 1e-4, 1e-5, 1e-6, 1e-1, 1e-2, 1e-7, 1e-8]
λ_2 for L2 regularization	[1e-3, 1e-4, 1e-5, 1e-6, 1e-1, 1e-2, 1e-7, 1e-8]
Sparsity scaling factor (β)	[0, 0.001, 0.01, 0.05, 1, 5, 10]
Target average hidden layer activation (ρ)	[0.001, 0.004, 0.007, 0.01, 0.04, 0.07, 0.1, 0.4, 0.7, 1.0]
Activation function type	['sigmoid', 'tanh', 'relu', ‘softplus’]
Learning rate	[0.000001, 0.00001, 0.0001, 0.001, 0.01, 0.1, 1, 10, 100]
Amplifying factor for focal loss (γ)	[0, 0.5, 1, 2, 3, 5]
Optimizer type	["Adam", "RMS Propagation", "Gradient Descent"]
Loss type	["Binary Cross Entropy", "Custom Focal Loss"]
Number of hidden layers	[1, 2, 4, 6, 8]
Hidden layer size ratio	[0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1]
Learning rate decay ratio	[0.0, 0.25, 0.5, 0.75, 0.95, 0.99, 0.999, 0.9999]

Table 1. λ_1: scaling factor for Least Absolute Shrinkage and Selection Operator (LASSO or L1) regularization; λ_2: scaling factor for Ridge (L2) regularization; β: scaling factor for sparsity penalty described in equation (4); ρ: target hidden layer activation described in equation (4); Activation function type: defines how the output of a hidden neuron will be computed given a set of inputs; Learning rate: step size at each learning iteration while moving toward the minimum of the loss function; γ: amplifying factor for focal loss described in equation (3); Optimizer type: algorithms utilized to minimize the loss function and update the model weights in backpropagation [34]; Loss type: algorithms utilized to calculate the model error (equation (2)); Number of hidden layers: how many layers of artificial neurons to be implemented between input layer and output layer; Hidden layer size ratio: scaling factor to resize the next hidden layer with reference to the size of its previous layer; Learning rate decay ratio: scaling factor for updating the learning rate value on every 500 epochs.

Hyperparameter Tuning

In order to avoid local optimal solutions and reduce the hyperparameter search space, we developed an ensemble-based machine learning approach (Extreme Gradient Boosting - XGBoost) to predict the expected performance (r-squared) of each hyperparameter combination per genomic segment using the results of the coarse-grid search and predictive features calculated for each genomic segment. These features include the number of variants, average recombination rate and average pairwise Pearson correlation across all SNPs, proportion of rare and common variants across multiple minor allele frequency (MAF) bins, number of principal components necessary to explain at least 90% of variance, and the total variance explained by the first 2 principal components. The observed accuracies of the coarse-grid search, numbering 25,600 training inputs, were used to predict the accuracy of 500,000 new hyperparameter combinations selected from **Table 1** without training. All categorical predictors (activation function name, optimizer type, loss function type) were one-hot encoded. The model was implemented using XGBoost package v1.4.1 in Python v3.8.3 with 10-fold cross-validation and default settings.

We then ranked all hyperparameter combinations by their predicted performance and selected the top 10 candidates per genomic segment along with the single best initially tested hyperparameter combination per genomic segments for further consideration. All other hyperparameter combinations were discarded. Genomic segments with sub-optimal performance relative to Minimac were subjected to tuning with simulated offspring formation. For tuning, the maximum number of epochs was increased (35,000) with automatic stop criteria: if there is no improvement in average loss value of the current masking/training cycle versus the previous one, the training is interrupted, otherwise training continues until the maximum epoch limit is reached. Each masking/training cycle consisted of 500 epochs. Final hyperparameter selection was based on performance on the validation dataset (ARIC).

Performance Testing and Comparisons

Performance was compared to Minimac4 (Das et al., 2016b), Beagle5 (Browning et al., 2018), and Impute5 (Rubinacci et al., 2020) using default parameters. Population level reconstruction accuracy is quantified by measuring r-squared across multiple strata of data: per genomic segment, at whole chromosome level, and stratified across multiple minor allele frequency bins: [0.001-0.005), [0.005-0.01), [0.01-0.05), [0.05-0.1), [0.1-0.2), [0.2-0.3), [0.3-0.4), [0.4-0.5). While r-squared is our primary comparison metric, sample-level and population-level model performance is also evaluated with concordance and the F1-score. Wilcoxon rank-sum testing was used assess the significance of accuracy differences observed. Spearman correlations were used to evaluate the relationships between genomic segment features and observed imputation accuracy differences. Standard errors for per variant imputation accuracy r-squared is equal or less than 0.001 where not specified. Performance is reported only for the independent test datasets (Wellderly, MESA, and HGDP).

We used the MESA cohort for inference runtime comparisons. Runtime was determined using the average and standard error of three imputation replicates. Two hardware configurations were used for the tests: 1) a low-end environment: 16-core Intel Xeon CPU (E5-2640 v2 2.00GHz), 250GB RAM, and one GPU (NVIDIA GTX 1080); 2) a high-end environment: 24-Core AMD CPU (EPYC 7352 2.3GHz), 250GB RAM, using one NVIDIA A100 GPU. We report computation time only, input/output (I/O) reading/writing times are excluded as separately optimized functions.

Data availability

The data that support the findings of this study are available from dbGAP and European Genome-phenome Archive (EGA), but restrictions apply to the availability of these data, which were used under ethics approval for the current study, and so are not openly available to the public. The computational pipeline for autoencoder training and validation is available at https://github.com/TorkamaniLab/Imputation_Autoencoder/tree/master/autoencoder_tuning_pipeline. The python script for calculating imputation accuracy is available at https://github.com/TorkamaniLab/imputation_accuracy_calculator.

Results

Untuned Performance and Model Optimization.

A preliminary comparison of the best performing autoencoder per genomic segment vs HMM-based imputation was made after the initial grid (Minimac4: **Figure 2**, Beagle5 and Eagle5: **Supplemental Figures S1-S2**). Untuned autoencoder performance was equivalent or inferior to all tested HMM-based methods except when tested on the European ancestry-rich Wellderly dataset when masked using the Affymetrix 6.0 and UKB Axiom marker sets, but not Omni 1.5M markers. HMM-based imputation was consistently superior across the more ancestrally diverse test datasets (MESA and HGDP) (two proportion test, $p \leq 8.77\text{x}10^{-6}$). Overall, when performance across genomic segments, test datasets, and test array marker sets was combined, the autoencoders exhibited an average r-squared per variant of 0.352±0.008 in reconstruction of WGS ground truth genotypes versus an average r-squared per variant of 0.374±0.007, 0.364±0.007, and 0.357±0.007 for HMM-based imputation methods (Minimac4, Beagle5, and Impute5, respectively) (**Table 2**). This difference was statistically significant only relative to Minimac4 (Minimac4: Wilcoxon rank-sum test p=0.037, Beagle5 and Eagle5: p≥0.66).

In order to understand the relationship between genomic segment features, hyperparameter values, and imputation performance, we calculated predictive features (see ***Methods***) for each genomic segment and determined their Spearman correlation with the differences in r-squared observed for the autoencoder vs Minimac4 (**Supplemental Figure S3**). We observed that the autoencoder had superior performance when applied to the genomic segments with the most complex LD structures: those with larger numbers of observed unique haplotypes, unique diplotypes, and heterozygosity, as well as high average MAF, and low average pairwise Pearson correlation across all SNPs (average LD) (Spearman correlation ($\rho \geq 0.22$, $p \leq 9.8\text{x}10^{-04}$). Similarly, we quantified genomic segment complexity by the proportion of variance explained by the first two principal components as well as the number of principal components needed to explain at least 90% of the variance of HRC genotypes from each genomic segment.

Concordantly, superior autoencoder performance was associated with a low proportion explained by the first two components and positively correlated with the number of components required to explained 90% of variance (Spearman ρ ≥ 0.22, p ≤ 8.3x10^{-04}). These observations informed our tuning strategy.

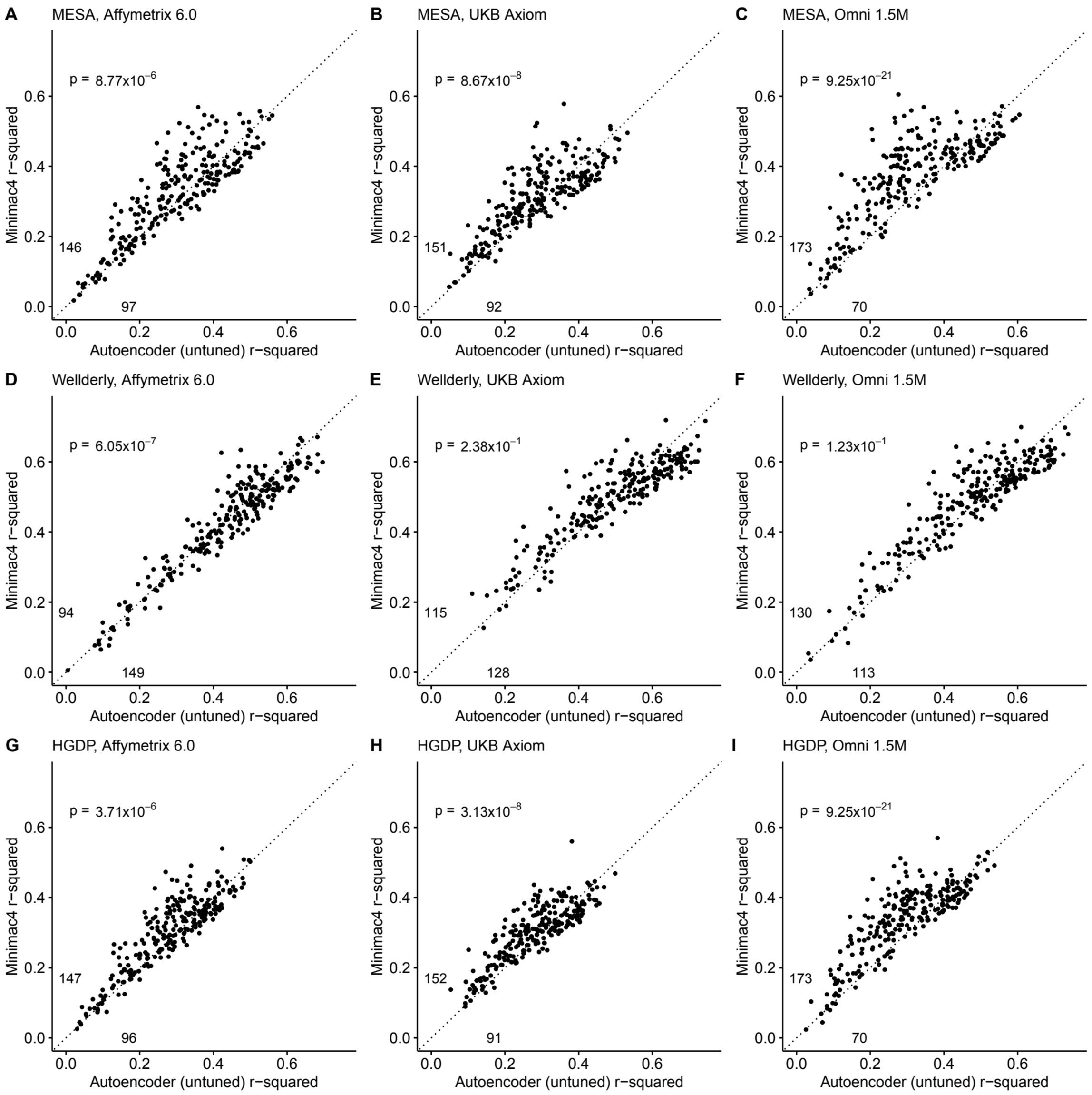


Figure 2. HMM-based (y-axis) versus autoencoder-based (x-axis) imputation accuracy prior to tuning. Minimac4 and untuned autoencoders were tested across three independent datasets - MESA (**top**), Wellderly (**middle**), and HGDP (**bottom**) - and across three genotyping array

platforms - Affymetrix 6.0 (**left**), UKB Axiom (**middle**), Omni1.5M (**right**). Each data point represents the imputation accuracy (average r-squared per variant) for an individual genomic segment relative to its WGS-based ground truth. The numerical values presented on the left side and below the identity line (dashed line) indicate the number of genomic segments in which Minimac4 outperformed the untuned autoencoder (left of identity line) and the number of genomic segments in which the untuned autoencoder surpassed Minimac4 (below the identity line). Statistical significance was assessed through two-proportion Z-test p-values.

Table 2. Performance comparisons between untuned autoencoder (AE) and HMM-based imputation tools (Minimac4, Beagle5, and Impute5).

	MESA	Wellderly	HGDP	Affymetrix 6.0	UKB Axiom	Omni 1.5M	Combined
AE (untuned)	0.303±0.008	0.470±0.009	0.285±0.006	0.339±0.008	0.356±0.007	0.362±0.008	0.352±0.008
Minimac4	0.337±0.007*	0.471±0.008	0.314±0.006**	0.352±0.008	0.370±0.006	0.400±0.007**	0.374±0.007*
Beagle5	0.336±0.007*	0.460±0.008	0.296±0.005	0.342±0.007	0.367±0.006	0.384±0.007*	0.364±0.007
Impute5	0.326±0.007*	0.458±0.008	0.289±0.006	0.336±0.008	0.354±0.006	0.383±0.008*	0.358±0.007

Table 2. Average r-squared per variant was extracted from each genomic segment of chromosome 22. We applied Wilcoxon rank-sum tests to compare the HMM-based tools to the reference untuned autoencoder (AE). * represents p-values ≤ 0.05, ** indicates p-values ≤ 0.001, and *** indicates p-values ≤ 0.0001.

We then used the genomic features significantly correlated with imputation performance to predict the performance of and select the hyperparameter values to advance to fine-tuning. An ensemble model inference approach was able to predict the genomic segment-specific performance of hyperparameter combinations with high accuracy (**Supplemental Figure S4**, mean r-squared = 0.935±0.002 of predicted vs observed autoencoders accuracies via 10-fold cross validation). The top 10 best performing hyperparameter combinations were advanced to fine-tuning (**Table 3**).

Autoencoder tuning with simulated offspring formation was then executed as described in ***Methods***.

Table 3. Top 10 best performing hyperparameter combinations that advanced to fine-tuning.

λ_1	λ_2	β	ρ	Activation	Learn rate	γ	Optimizer	Loss type	Hidden layers	Size ratio	Decay
0.1	0	0.01	0.01	tanh	$1.0*10^{-4}$	0	adam	CE	4	1	0.95
0.1	0	1	0.5	sigmoid	$1.0*10^{-4}$	1	adam	CE	2	0.9	0.95
0.1	0	5	0.5	sigmoid	$1.0*10^{-1}$	4	adam	CE	2	0.5	0
0.1	0	1	0.005	relu	$1.0*10^{-1}$	4	adam	FL	6	1	0.25
0.1	0	5	0.01	relu	$1.0*10^{-5}$	5	adam	FL	4	1	0.95
0.1	0	0.01	0.1	leakyrelu	$1.0*10^{-5}$	0	adam	FL	8	0.9	0.95
0.1	0	1	0.01	tanh	$1.0*10^{-4}$	0	adam	CE	6	1	0.95
0	$1.0*10^{-8}$	0.001	0.05	relu	$1.0*10^{-5}$	4	adam	CE	8	0.6	0.95
0.1	0	0	0.01	relu	$1.0*10^{-1}$	5	adam	FL	8	0.9	0
0.1	0	0.01	0.01	tanh	$1.0*10^{-3}$	5	adam	CE	2	1	0.95

Table 3. See ***Methods*** and **Table 1** for a detailed description of the hyperparameters.

<u>Tuned Performance.</u>

After tuning, autoencoder performance surpassed HMM-based imputation performance across all imputation methods, independent test datasets, and genotyping array marker sets. At a minimum, autoencoders surpassed HMM-based imputation performance in >62% of chromosome 22 genomic segments (two proportion test $p=1.02x10^{-11}$) (Minimac4: **Figure 3**, Beagle5 and Eagle5: **Supplemental Figures S5-S6**). Overall, the optimized autoencoders exhibited superior performance with an average r-squared of 0.395±0.007 vs 0.374±0.007 for Minimac4 (Wilcoxon rank sum test p=0.007), 0.364±0.007 for Beagle5 (Wilcoxon rank sum test $p=1.53*10^{-4}$), and 0.358±0.007 for Impute5 (Wilcoxon rank sum test $p=2.01*10^{-5}$) (**Table 4**). This superiority was robust to the marker sets tested, with the mean r-squared per genomic segment for autoencoders being 0.373±0.008, 0.399±0.007, and 0.414±0.008 versus 0.352±0.008, 0.370±0.006, and 0.400±0.007 for Minimac4 using Affymetrix 6.0, UKB Axiom, and Omni 1.5M marker sets

(Wilcoxon rank-sums test p-value=0.029, 1.99*10^{-4}, and 0.087, respectively). Detailed comparisons to Beagle5 and Eagle5 are presented in **Supplemental Figures S5-S6**.

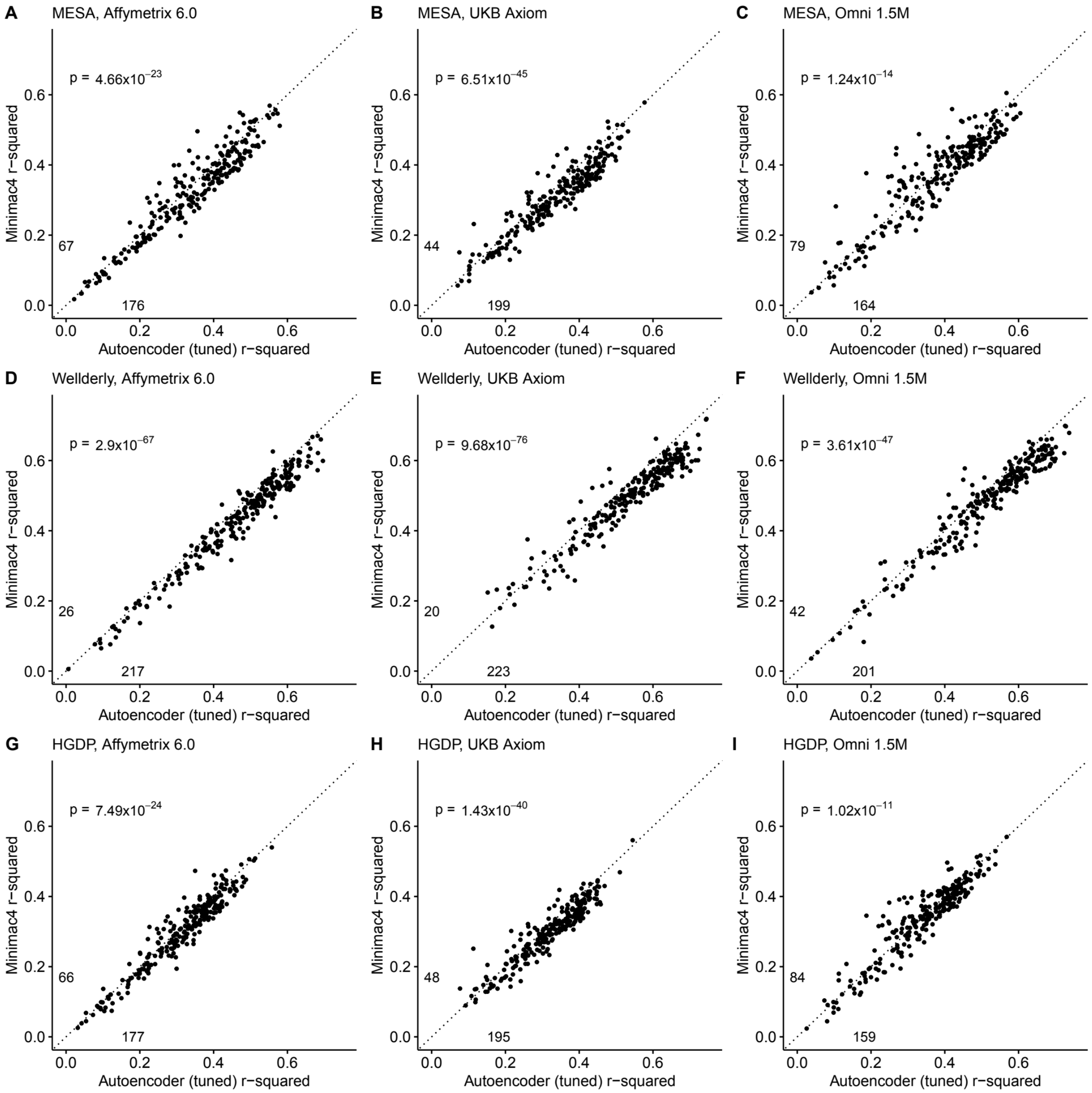


Figure 3. HMM-based (y-axis) versus autoencoder-based (axis) imputation accuracy after tuning. Minimac4 and tuned autoencoders were validated across three independent datasets - MESA (**top**), Wellderly (**middle**), and HGDP (**bottom**) - and across three genotyping array platforms - Affymetrix 6.0 (**left**), UKB Axiom (**middle**), Omni1.5M (**right**). Each data point represents the imputation accuracy (average r-squared per variant) for an individual genomic

segment relative to its WGS-based ground truth. The numerical values presented on the left side and below the identity line (dashed line) indicate the number of genomic segments in which Minimac4 outperformed the untuned autoencoder (left of identity line) and the number of genomic segments in which the untuned autoencoder surpassed Minimac4 (below the identity line). Statistical significance was assessed through two-proportion Z-test p-values.

Table 4. Performance comparisons between tuned autoencoder (AE) and HMM-based imputation tools (Minimac4, Beagle5, and Impute5).

	MESA	Wellderly	HGDP	Affymetrix 6.0	UKB Axiom	Omni 1.5M	Combined
AE (tuned)	0.355±0.007	0.505±0.008	0.327±0.006	0.373±0.008	0.399±0.007	0.414±0.008	0.396±0.007
AE (untuned)	0.303±0.008***	0.470±0.009*	0.285±0.006***	0.339±0.008*	0.356±0.007***	0.362±0.008***	0.352±0.008***
Minimac4	0.337±0.007*	0.471±0.008**	0.314±0.006	0.352±0.008*	0.370±0.006**	0.400±0.007	0.374±0.007*
Beagle5	0.336±0.007*	0.460±0.008***	0.296±0.005***	0.342±0.007**	0.367±0.006***	0.384±0.007**	0.364±0.007**
Impute5	0.326±0.007*	0.458±0.008***	0.289±0.006***	0.336±0.008**	0.354±0.006***	0.383±0.008**	0.358±0.007***

Table 4. Average r-squared per variant was extracted from each genomic segment of chromosome 22. We applied Wilcoxon rank-sum tests to compare the HMM-based tools to the reference untuned autoencoder (AE). * represents p-values ≤ 0.05, ** indicates p-values ≤ 0.001, and *** indicates p-values ≤ 0.0001.

Tuning improved performance of the autoencoders across all genomic segments, generally improving the superiority of autoencoders relative to HMM-based approaches in genomic segments with complex haplotype structures while equalizing performance relative to HMM-based approaches in genomic segments with more simple LD structures (as described in *Methods*, by the number of unique haplotypes: **Supplemental Figure S7**, diplotypes: **Supplemental Figure S8**, average pairwise LD: **Supplemental Figure S9**, proportion variance explained: **Supplemental Figure S10**). Concordantly, genomic segments with higher recombination rates exhibited the largest degree of improvement with tuning (**Supplemental Figure S11**). Use of the augmented reference panel did not improve HMM-based imputation, having no influence on Minimac4 performance (original overall r-squared of 0.374±0.007 versus 0.363±0.007 after augmentation, Wilcoxon rank-sum test p=0.0917), and significantly degrading performance of Beagle5 and Impute5 (original r-squared of 0.364±0.007 and 0.358±0.007 versus 0.349±0.006 and 0.324±0.007 after augmentation, p=0.026 and $p=1.26*10^{-4}$ respectively). Summary statistics for these comparisons are available in **Supplemental Table S1**.

Overall Chromosome 22 Imputation Accuracy.

After merging the results from all genomic segments, the whole chromosome accuracy of autoencoder-based imputation remained superior to all HMM-based imputation tools, across all independent test datasets, and all genotyping array marker sets (Wilcoxon rank-sums test $p \leq 5.55x10^{-67}$). The autoencoder's mean r-squared per variant ranged from 0.363 for HGDP to 0.605 for the Wellderly vs 0.340 to 0.557 for Minimac4, 0.326 to 0.549 for Beagle5, and 0.314 to 0.547 for Eagle5, respectively. Detailed comparisons are presented in in **Table 5** and **Supplemental Table S2**.

Further, when imputation accuracy is stratified by MAF bins, the autoencoders maintain superiority across all MAF bins by nearly all test dataset and genotyping array marker sets (**Figure 4**, and **Supplemental Table S3**). Concordantly, autoencoder imputation accuracy is similarly superior when measured with F1-scores (**Supplemental Figure S12**) and concordance (**Supplemental Figure S13**), though these metrics are less sensitive at capturing differences in rare variant imputation accuracy.

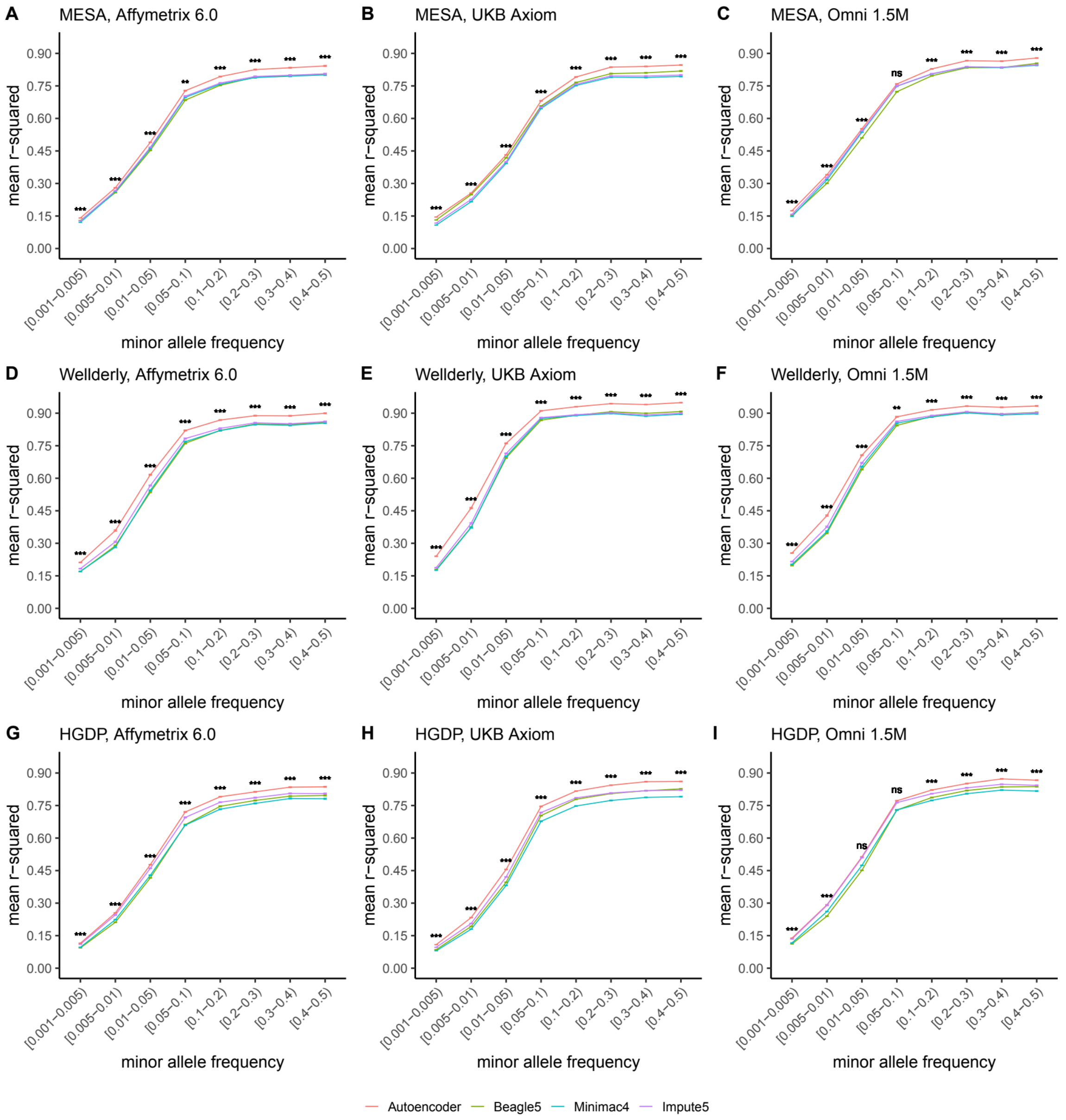


Figure 4. HMM-based versus autoencoder-based imputation accuracy across MAF bins. Autoencoder-based (**red**) and HMM-based (Minimac4 (**blue**), Beagle5 (**green**), and Impute5 (**purple**)) imputation accuracy was validated across three independent datasets - MESA (**top**), Wellderly (**middle**), and HGDP (**bottom**) - and across three genotyping array platforms - Affymetrix 6.0 (**left**), UKB Axiom (**middle**), Omni1.5M (**right**). Each data point represents the imputation accuracy (average r-squared per variant) relative to WGS-based ground truth across MAF bins.

Error bars represent standard errors. We applied Wilcoxon rank-sum tests to compare the HMM-based tools to the tuned autoencoder (AE). * represents p-values ≤ 0.05, ** indicates p-values ≤ 0.001, and *** indicates p-values ≤ 0.0001, ns represents non-significant p-values.

Table 5. Whole chromosome level comparisons between autoencoder (AE) and HMM-based imputation tools (Minimac4, Beagle5, and Impute5).

	MESA			Wellderly			HGDP		
	Affymetrix 6.0	UKB Axiom	Omni 1.5M	Affymetrix 6.0	UKB Axiom	Omni 1.5M	Affymetrix 6.0	UKB Axiom	Omni 1.5M
AE (tuned)	0.410	0.395	0.452	0.537	0.605	0.586	0.363	0.364	0.392
Minimac4	0.390***	0.364***	0.436***	0.500***	0.557***	0.551***	0.350***	0.340***	0.385***
Beagle5	0.383***	0.379***	0.420***	0.484***	0.549***	0.534***	0.326***	0.328***	0.353***
Impute5	0.384***	0.356***	0.429***	0.485***	0.547***	0.539***	0.328***	0.314***	0.359***

Table 5. Average r-squared per variant was extracted at whole chromosome level. We applied Wilcoxon rank-sum tests to compare the HMM-based tools to the reference tuned autoencoder (AE). * represents p-values ≤ 0.05, ** indicates p-values ≤ 0.001, and *** indicates p-values ≤ 0.0001. Standard errors that are equal or less than 0.001 are not shown.

Ancestry-Specific Chromosome 22 Imputation Accuracy.

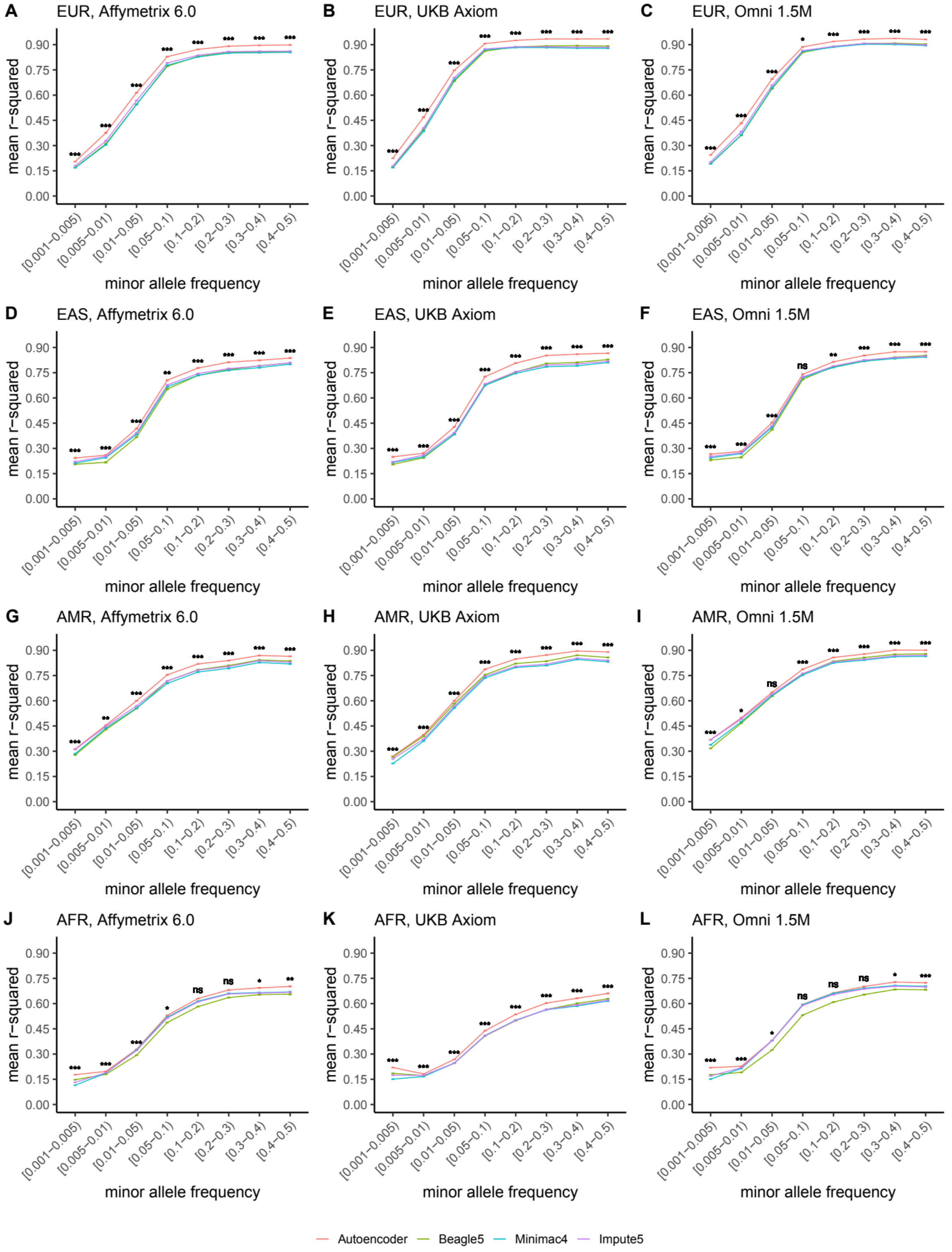
Finally, we evaluated ancestry-specific imputation accuracy. As before, overall autoencoder-based imputation maintains superiority across all continental populations present in MESA (**Figure 5**, Wilcoxon rank-sums test p=5.39x10^{-19}). The autoencoders' mean r-squared ranged from 0.357 for African ancestry to 0.614 for East Asian ancestry vs 0.328 to 0.593 for Minimac4, 0.330 to 0.544 for Beagle5, and 0.324 to 0.586 for Impute5, respectively. Note, East Asian ancestry exhibits a slightly higher overall imputation accuracy relative to European ancestry due to improved rare variant imputation. Autoencoder superiority replicates when HGDP is split into continental populations (**Supplemental Figure S14**).

Further stratification of ancestry-specific imputation accuracy results by MAF continues to support autoencoder superiority across all ancestries, MAF bins, and nearly all test datasets, and genotyping array marker sets (**Figure 5, Supplemental Figure S14**). Minimum and maximum accuracies across MAF by ancestry bins ranged between 0.177 to 0.937 for the autoencoder, 0.132 to 0.907 for Minimac4, 0.147 to 0.909 for Beagle5, and 0.115 to 0.903 for Impute5, with a maximum standard error of ±0.004.

Thus, autoencoder performance was superior across all variant allele frequencies and ancestries with the primary source of superiority arising from hard to impute regions with complex LD structures.

<u>Inference Speed.</u>

Inference runtimes for the autoencoder vs HMM-based methods were compared in a low-end and high-end computational environment as described in *Methods*. In the low-end environment, the autoencoder's inference time is at least ~4X faster than all HMM-based inference times (summing all inference times from all genomic segments of chromosome 22, the inference time for the autoencoder was $2.4\pm1.1*10^{-3}$ seconds versus 1,754±3.2, 583.3±0.01, and $8.4\pm4.3*10^{-3}$ seconds for Minimac4, Beagle5, and Impute5, respectively (**Figure 6A**)). In the high-end environment, this difference narrows to a ~3X advantage of the autoencoder vs HMM-based methods ($2.1\pm8.0*10^{-4}$ versus 374.3±1.2, 414.3±0.01, and $6.1\pm2.1*10^{-4}$ seconds for Minimac4, Beagle5, and Impute5, respectively (**Figure 6B**). These unoptimized results indicate that autoencoder-based imputation can be executed rapidly, without a reference cohort, and without the need for a high-end server or high-performance computing (HPC) infrastructure.


A
EUR, Affymetrix 6.0
*** *** *** *** *** *** *** ***
B
EUR, UKB Axiom
*** *** *** *** *** *** *** ***
C
EUR, Omni 1.5M
*** *** *** * *** *** *** ***
D
EAS, Affymetrix 6.0
*** *** *** ** *** *** *** ***
E
EAS, UKB Axiom
*** *** *** *** *** *** *** ***
F
EAS, Omni 1.5M
*** *** *** ns ** *** *** ***
G
AMR, Affymetrix 6.0
*** ** *** *** *** *** *** ***
H
AMR, UKB Axiom
*** *** *** *** *** *** *** ***
I
AMR, Omni 1.5M
*** * ns *** *** *** *** ***
J
AFR, Affymetrix 6.0
*** *** *** * ns ns * **
K
AFR, UKB Axiom
*** *** *** *** *** *** *** ***
L
AFR, Omni 1.5M
*** *** * ns ns ns * ***
mean r−squared
0.00
0.15
0.30
0.45
0.60
0.75
0.90
[0.001−0.005)
[0.005−0.01)
[0.01−0.05)
[0.05−0.1)
[0.1−0.2)
[0.2−0.3)
[0.3−0.4)
[0.4−0.5)
minor allele frequency
Autoencoder
Beagle5
Minimac4
Impute5

Impute5 (**purple**)) imputation accuracy was validated across individuals of diverse ancestry from MESA cohort (EUR: European (**top**); EAS: East Asian (**2nd row**); AMR: Native American (**3rd row**); AFR: African (**bottom**)) and multiple genotype array platforms (Affymetrix 6.0 (**left**), UKB Axiom (**middle**), Omni1.5M (**right**)). Each data point represents the imputation accuracy (average r-squared per variant) relative to WGS-based ground truth across MAF bins. Error bars represent standard errors. We applied Wilcoxon rank-sum tests to compare the HMM-based tools to the tuned autoencoder (AE). * represents p-values ≤ 0.05, ** indicates p-values ≤ 0.001, and *** indicates p-values ≤ 0.0001, ns represents non-significant p-values.

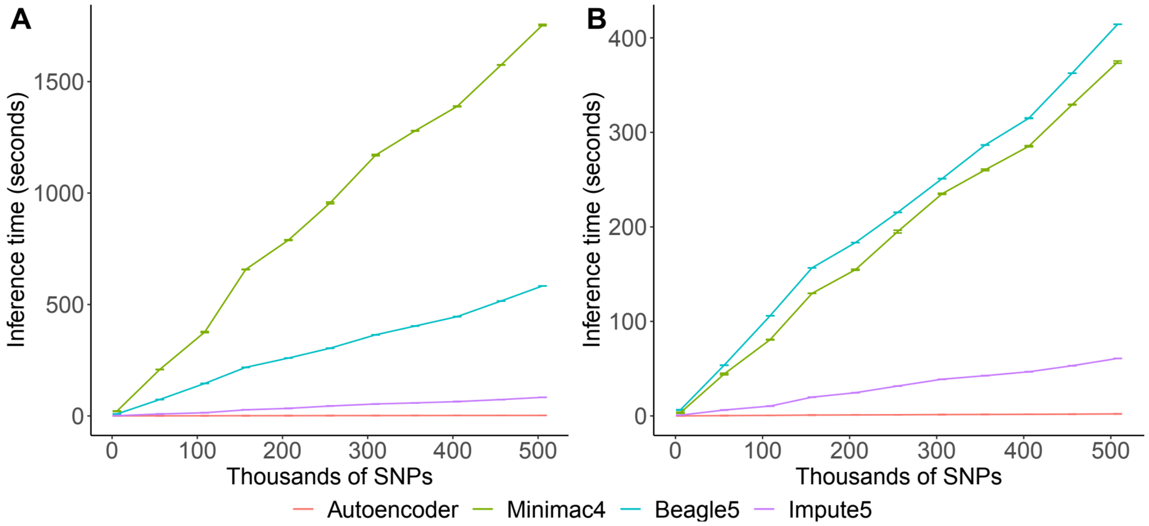


Figure 6. HMM-based versus autoencoder-based inference runtimes. We plot the average time and standard error of three imputation replicates. Two hardware configurations were used for the tests: A) a low-end environment: 16-core Intel Xeon CPU (E5-2640 v2 2.00GHz), 250GB RAM, and one GPU (NVIDIA GTX 1080); B) a high-end environment: 24-Core AMD CPU (EPYC 7352 2.3GHz), 250GB RAM, using one NVIDIA A100 GPU.

Discussion

Artificial neural network-based data mining techniques are revolutionizing biomedical informatics and analytics(Dias and Torkamani, 2019; Jumper et al., 2021). Here, we have demonstrated the potential for these techniques to execute a fundamental analytical task in population genetics, genotype imputation, producing superior results in a computational efficient and portable framework. The trained autoencoders can be transferred easily, and execute their functions rapidly, even in modest computing environments, obviating the need to transfer private genotype data to external imputation servers or services. Furthermore, our fully trained autoencoders robustly surpass the performance of all modern HMM-based imputation approaches across all tested independent datasets, genotyping array marker sets, minor allele frequency spectra, and diverse ancestry groups. This superiority was most apparent in genomic regions with low LD and/or high complexity in their linkage disequilibrium structure.

Superior imputation accuracy is expected to improve GWAS power, enable more complete coverage in meta-analyses, and improve causal variant identification through fine-mapping. Moreover, superior imputation accuracy in low LD regions may enable the more accurate interrogation of specific classes of genes under a greater degree of selective pressure and involved in environmental sensing. For example, promoter regions of genes associated with inflammatory immune responses, response to pathogens, environmental sensing, and neurophysiological processes (including sensory perception genes) are often located in regions of low LD (Dias and Torkamani, 2019; Frazer et al., 2007). These known disease-associated biological processes that are critical to interrogate accurately in GWAS. Thus, the autoencoder-based imputation approach both improves statistical power and biological coverage of individual GWAS’ and downstream meta-analyses.

HMM-based imputation tools depend on large reference panels or datasets to impute a single genome whereas pre-trained autoencoder models eliminate that dependency. However, further development is required to actualize this approach in practice for broad adoption. Autoencoders must be pre-trained and validated across all segments of the human genome. Here we performed training only for chromosome 22. Autoencoder training is computationally intensive, shifting the computational burden to model trainers, and driving performance gains for end-users. As a result, inference time scales only with the number of variants to be imputed, whereas HMM-based inference time depends on both reference panel and the number of variants to be imputed. This allows for autoencoder-based imputation to extend to millions of genomes but introduces some challenges in the continuous re-training and fine-tuning of the pre-trained models as larger reference panels are made available. In addition, our current encoding approach lacks phasing information, which leads to substantial improvements in imputation accuracy. Future models will need to address the need for phasing and continuous fine-tuning of models for application to modern, ever-growing, genomic datasets.

Ideas and Speculation

After expanding this approach across the whole genome, our work will provide a more efficient genotype imputation platform on whole genome scale and thus beneficial for genome association studies and clinical applications in precision medicine. In addition to the speed, cost and accuracy benefits, our proposed approach can potentially improve automation for downstream analyses. The autoencoder naturally generates a hidden encoding with latent features representative of the original data. This latent representation of the original data acts as an automatic feature extraction and dimensionality reduction technique for downstream tasks such as genetic risk prediction. Moreover, the autoencoder-based imputation approach only requires a reference panel during training – only the neural network needs to be distributed for implementation. Thus, the neural network is portable and avoids privacy issues associated with standard statistical imputation. This privacy-preserving feature will allow developers to deploy real-time data-driven algorithms on personal devices (edge computing). These new features will expand the clinical applications of genomic imputation, as well as its role in preventive healthcare.

Acknowledgments

This work is supported by KL2TR002552 to RD, by R01HG010881 to AT as well as grants U24TR002306 and UL1TR002550. We would like to thank J.C. Ducom and Lisa Dong from the Scripps High Performance Computing center, as well as Fernanda Foertter, Johnny Israeli, Ohad Mosafi, and Joyjit Daw from NVIDIA for their technical support and collaboration in this project. A portion of this research was conducted using a startup account at the Summit supercomputer from Oak Ridge National Laboratory (ORNL).

Competing interests

The authors declare no competing interests.

References

Abouzid H, Chakkor O, Reyes OG, Ventura S. 2019. Signal speech reconstruction and noise removal using convolutional denoising audioencoders with neural deep learning. *Analog Integrated Circuits and Signal Processing* **100**:501–512. doi:10.1007/s10470-019-01446-6

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research* **19**:1655–1664. doi:10.1101/gr.094052.109

Arisdakessian C, Poirion O, Yunits B, Zhu X, Garmire LX. 2019. DeepImpute: An accurate, fast, and scalable deep neural network method to impute single-cell RNA-seq data. *Genome Biology* **20**:211. doi:10.1186/s13059-019-1837-6

Auton A, Abecasis GR, Altshuler DM, Durbin RM, Bentley DR, Chakravarti A, Clark AG, Donnelly P, Eichler EE, Flicek P, Gabriel SB, Gibbs RA, Green ED, Hurles ME, Knoppers BM, Korbel JO, Lander ES, Lee C, Lehrach H, Mardis ER, Marth GT, McVean GA, Nickerson DA, Schmidt JP, Sherry ST, Wang J, Wilson RK, Boerwinkle E, Doddapaneni H, Han Y, Korchina V, Kovar C, Lee S, Muzny D, Reid JG, Zhu Y, Chang Y, Feng Q, Fang X, Guo X, Jian M, Jiang H, Jin X, Lan T, Li G, Li J, Li Yingrui, Liu S, Liu Xiao, Lu Y, Ma X, Tang M, Wang B, Wang G, Wu H, Wu R, Xu X, Yin Y, Zhang D, Zhang W, Zhao J, Zhao M, Zheng X, Gupta N, Gharani N, Toji LH, Gerry NP, Resch AM, Barker J, Clarke L, Gil L, Hunt SE, Kelman G, Kulesha E, Leinonen R, McLaren WM, Radhakrishnan R, Roa A, Smirnov D, Smith RE, Streeter I, Thormann A, Toneva I, Vaughan B, Zheng-Bradley X, Grocock R, Humphray S, James T, Kingsbury Z, Sudbrak R, Albrecht MW, Amstislavskiy VS, Borodina TA, Lienhard M, Mertes F, Sultan M, Timmermann B, Yaspo ML, Fulton L, Ananiev V, Belaia Z, Beloslyudtsev D, Bouk N, Chen C, Church D, Cohen R, Cook C, Garner J, Hefferon T, Kimelman M, Liu C, Lopez J, Meric P, O'Sullivan C, Ostapchuk Y, Phan L, Ponomarov S, Schneider V, Shekhtman E, Sirotkin K, Slotta D, Zhang H, Balasubramaniam S, Burton J, Danecek P, Keane TM, Kolb-Kokocinski A, McCarthy S, Stalker J, Quail M, Davies CJ, Gollub J, Webster T, Wong B, Zhan Y, Campbell CL, Kong Y, Marcketta A, Yu F, Antunes L, Bainbridge M, Sabo A, Huang Z, Coin LJM, Fang L, Li Q, Li Z, Lin H, Liu B, Luo R, Shao H, Xie Y, Ye C, Yu C, Zhang F, Zheng H, Zhu H, Alkan C, Dal E, Kahveci F, Garrison EP, Kural D, Lee WP, Leong WF, Stromberg M, Ward AN, Wu J, Zhang M, Daly MJ, DePristo MA, Handsaker RE, Banks E, Bhatia G, del Angel G, Genovese G, Li H, Kashin S, McCarroll SA, Nemesh JC, Poplin RE, Yoon SC, Lihm J, Makarov V, Gottipati S, Keinan A, Rodriguez-Flores JL, Rausch T, Fritz MH, Stütz AM, Beal K, Datta A, Herrero J, Ritchie GRS, Zerbino D, Sabeti PC, Shlyakhter I, Schaffner SF, Vitti J, Cooper DN, Ball E v., Stenson PD, Barnes B, Bauer M, Cheetham RK, Cox A, Eberle M, Kahn S, Murray L, Peden J, Shaw R, Kenny EE, Batzer MA, Konkel MK, Walker JA, MacArthur DG, Lek M, Herwig R, Ding L, Koboldt DC, Larson D, Ye Kai, Gravel S, Swaroop A, Chew E, Lappalainen T, Erlich Y, Gymrek M, Willems TF, Simpson JT, Shriver MD, Rosenfeld JA, Bustamante CD, Montgomery SB, de La Vega FM, Byrnes JK, Carroll AW, DeGorter MK, Lacroute P, Maples BK, Martin AR, Moreno-Estrada A, Shringarpure SS, Zakharia F, Halperin E, Baran Y, Cerveira E, Hwang J, Malhotra A, Plewczynski D, Radew K, Romanovitch M, Zhang C, Hyland FCL, Craig DW, Christoforides A, Homer N, Izatt T, Kurdoglu AA, Sinari SA, Squire K, Xiao C, Sebat J, Antaki D, Gujral M, Noor A, Ye Kenny, Burchard EG, Hernandez RD, Gignoux CR, Haussler D, Katzman SJ, Kent WJ, Howie B, Ruiz-Linares A, Dermitzakis ET, Devine SE, Kang HM, Kidd JM, Blackwell T, Caron S, Chen W, Emery S, Fritsche L, Fuchsberger C, Jun G, Li B, Lyons R, Scheller C, Sidore C, Song S, Sliwerska E, Taliun D, Tan A, Welch R, Wing MK, Zhan X, Awadalla P, Hodgkinson A, Li Yun, Shi X, Quitadamo A, Lunter G, Marchini JL, Myers S, Churchhouse C, Delaneau O, Gupta-Hinch A, Kretzschmar W, Iqbal Z, Mathieson I, Menelaou A, Rimmer A, Xifara DK, Oleksyk TK, Fu Yunxin, Liu Xiaoming, Xiong M, Jorde L, Witherspoon D, Xing J, Browning BL, Browning SR, Hormozdiari F, Sudmant PH, Khurana E, Tyler-Smith C, Albers CA, Ayub Q, Chen Y, Colonna V, Jostins L, Walter K, Xue Y, Gerstein MB, Abyzov A, Balasubramanian S, Chen J, Clarke D, Fu Yao, Harmanci AO, Jin M, Lee D, Liu J, Mu XJ, Zhang J, Zhang Yan, Hartl C, Shakir K, Degenhardt J, Meiers S, Raeder B, Casale FP, Stegle O, Lameijer EW, Hall I, Bafna V, Michaelson J, Gardner EJ, Mills RE, Dayama G, Chen K, Fan X, Chong Z, Chen T, Chaisson MJ, Huddleston J, Malig M, Nelson BJ, Parrish NF, Blackburne B, Lindsay SJ, Ning Z, Zhang Yujun, Lam H, Sisu C, Challis D, Evani US, Lu J, Nagaswamy U, Yu J, Li W, Habegger L, Yu H, Cunningham F, Dunham I, Lage K, Jespersen JB, Horn H, Kim D, Desalle R, Narechania A, Sayres MAW, Mendez FL, Poznik GD, Underhill PA, Mittelman D, Banerjee R, Cerezo M, Fitzgerald TW, Louzada S, Massaia A, Yang F, Kalra D, Hale W, Dan X, Barnes KC, Beiswanger C, Cai H, Cao H, Henn B, Jones D, Kaye JS, Kent A, Kerasidou A, Mathias R, Ossorio PN, Parker M, Rotimi CN, Royal CD, Sandoval K, Su Y, Tian Z, Tishkoff S, Via M, Wang Y, Yang H, Yang L, Zhu J, Bodmer W, Bedoya G, Cai Z, Gao Y, Chu J, Peltonen L, Garcia-Montero A, Orfao A, Dutil J, Martinez-Cruzado JC, Mathias RA, Hennis A, Watson H, McKenzie C, Qadri F, LaRocque R, Deng X, Asogun D, Folarin O, Happi C, Omoniwa O, Stremlau M, Tariyal R, Jallow M, Joof FS, Corrah T, Rockett K, Kwiatkowski D, Kooner J, Hien TT, Dunstan SJ, ThuyHang N, Fonnie R, Garry R, Kanneh L, Moses L, Schieffelin J, Grant DS, Gallo C, Poletti G, Saleheen D, Rasheed A, Brooks LD, Felsenfeld AL, McEwen JE, Vaydylevich Y, Duncanson A, Dunn M, Schloss JA. 2015. A global reference for human genetic variation. *Nature*. doi:10.1038/nature15393

Berisa T, Pickrell JK. 2016. Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics* **32**:283–285. doi:10.1093/bioinformatics/btv546

Bild DE. 2002. Multi-Ethnic Study of Atherosclerosis: Objectives and Design. *American Journal of Epidemiology* **156**:871–881. doi:10.1093/aje/kwf113

Browning BL, Browning SR. 2016. Genotype Imputation with Millions of Reference Samples. *American Journal of Human Genetics* **98**:116–126. doi:10.1016/j.ajhg.2015.11.020

Browning BL, Zhou Y, Browning SR. 2018. A One-Penny Imputed Genome from Next-Generation Reference Panels. *American Journal of Human Genetics* **103**:338–348. doi:10.1016/j.ajhg.2018.07.015

Cann HM. 2002. A Human Genome Diversity Cell Line Panel. *Science* **296**:261b–2262. doi:10.1126/science.296.5566.261b

Chen J, Shi X. 2019. Sparse Convolutional Denoising Autoencoders for Genotype Imputation. *Genes* **10**:652. doi:10.3390/genes10090652

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, Li H. 2021. Twelve years of SAMtools and BCFtools. *GigaScience* **10**:1–4. doi:10.1093/gigascience/giab008

Das S, Abecasis GR, Browning BL. 2018. Genotype Imputation from Large Reference Panels. *Annual Review of Genomics and Human Genetics* **19**:73–96. doi:10.1146/annurev-genom-083117-021602

Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, Vrieze SI, Chew EY, Levy S, McGue M, Schlessinger D, Stambolian D, Loh PR, Iacono WG, Swaroop A, Scott LJ, Cucca F, Kronenberg F, Boehnke M, Abecasis GR, Fuchsberger C. 2016a. Next-generation genotype imputation service and methods. *Nature Genetics* **48**:1284–1287. doi:10.1038/ng.3656

Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, Vrieze SI, Chew EY, Levy S, McGue M, Schlessinger D, Stambolian D, Loh PR, Iacono WG, Swaroop A, Scott LJ, Cucca F, Kronenberg F, Boehnke M, Abecasis GR, Fuchsberger C. 2016b. Next-generation genotype imputation service and methods. *Nature Genetics* **48**:1284–1287. doi:10.1038/ng.3656

Dias R, Torkamani A. 2019. Artificial intelligence in clinical and genomic diagnostics. *Genome Medicine*. doi:10.1186/s13073-019-0689-8

Dimitromanolakis A, Xu J, Krol A, Briollais L. 2019. sim1000G: A user-friendly genetic variant simulator in R for unrelated individuals and family-based designs. *BMC Bioinformatics* **20**:26. doi:10.1186/s12859-019-2611-1

Erikson GA, Bodian DL, Rueda M, Molparia B, Scott ER, Scott-Van Zeeland AA, Topol SE, Wineinger NE, Niederhuber JE, Topol EJ, Torkamani A. 2016. Whole-Genome Sequencing of a Healthy Aging Cohort. *Cell* **165**:1002–1011. doi:10.1016/j.cell.2016.03.022

Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, Pasternak S, Wheeler DA, Willis TD, Yu F, Yang H, Zeng C, Gao Y, Hu H, Hu W, Li C, Lin W, Liu S, Pan H, Tang X, Wang J, Wang W, Yu J, Zhang B, Zhang Q, Zhao Hongbin, Zhao Hui, Zhou J, Gabriel SB, Barry R, Blumenstiel B, Camargo A, Defelice M, Faggart M, Goyette M, Gupta S, Moore J, Nguyen H, Onofrio RC, Parkin M, Roy J, Stahl E, Winchester E, Ziaugra L, Altshuler D, Shen Yan, Yao Z, Huang W, Chu X, He Y, Jin L, Liu Y, Shen Yayun, Sun W, Wang Haifeng, Wang Yi, Wang Ying, Xiong X, Xu L, Waye MMY, Tsui SKW, Xue H, Wong JTF, Galver LM, Fan JB, Gunderson K, Murray SS, Oliphant AR, Chee MS, Montpetit A, Chagnon F, Ferretti V, Leboeuf M, Olivier JF, Phillips MS, Roumy S, Sallée C, Verner A, Hudson TJ, Kwok PY, Cai D, Koboldt DC, Miller RD, Pawlikowska L, Taillon-Miller P, Xiao M, Tsui LC, Mak W, You QS, Tam PKH, Nakamura Y, Kawaguchi T, Kitamoto T, Morizono T, Nagashima A, Ohnishi Y, Sekine A, Tanaka T, Tsunoda T, Deloukas P, Bird CP, Delgado M, Dermitzakis ET, Gwilliam R, Hunt S, Morrison J, Powell D, Stranger BE, Whittaker P, Bentley DR, Daly MJ, de Bakker PIW, Barrett J, Chretien YR, Maller J, McCarroll S, Patterson N, Pe'Er I, Price A, Purcell S, Richter DJ, Sabeti P, Saxena R, Schaffner SF, Sham PC, Varilly P, Stein LD, Krishnan L, Smith AV, Tello-Ruiz MK, Thorisson GA, Chakravarti A, Chen PE, Cutler DJ, Kashuk CS, Lin S, Abecasis GR, Guan W, Li Y, Munro HM, Qin ZS, Thomas DJ, McVean G, Auton A, Bottolo L, Cardin N, Eyheramendy S, Freeman C, Marchini J, Myers S, Spencer C, Stephens M, Donnelly P, Cardon LR, Clarke G, Evans DM, Morris AP, Weir BS, Johnson TA, Mullikin JC, Sherry ST, Feolo M, Skol A, Zhang H, Matsuda I, Fukushima Y, MacEr DR, Suda E, Rotimi CN, Adebamowo CA, Ajayi I, Aniagwu T, Marshall PA, Nkwodimmah C, Royal CDM, Leppert MF, Dixon M, Peiffer A, Qiu R, Kent A, Kato K, Niikawa N, Adewole IF, Knoppers BM, Foster MW, Clayton EW, Watkin J, Muzny D, Nazareth L, Sodergren E, Weinstock GM, Yakub I, Birren BW, Wilson RK, Fulton LL, Rogers J, Burton J, Carter NP, Clee CM, Griffiths M, Jones MC, McLay K, Plumb RW, Ross MT, Sims SK, Willey DL, Chen Z, Han H, Kang L, Godbout M, Wallenburg JC, L'Archevêque P, Bellemare G, Saeki K, Wang Hongguang, An D, Fu H, Li Q, Wang Z, Wang R, Holden AL, Brooks LD, McEwen JE, Guyer MS, Wang VO, Peterson JL, Shi M, Spiegel J, Sung LM, Zacharia LF, Collins FS, Kennedy K, Jamieson R, Stewart J. 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**:851–861. doi:10.1038/nature06258

Islam T, Kim CH, Iwata H, Shimono H, Kimura A, Zaw H, Raghavan C, Leung H, Singh RK. 2021. A Deep Learning Method to Impute Missing Values and Compress Genome-ide Polymorphism Data in Rice In: Lorenz R, Fred ALN, Gamboa H, editors. Proceedings of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies, {BIOSTEC} 2021, Volume 3: BIOINFORMATICS, Online Streaming, February 11-13, 2021. SCITEPRESS. pp. 101–109.

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature* 1–7. doi:10.1038/s41586-021-03819-2

Koh PW, Pierson E, Kundaje A. 2017. Denoising genome-wide histone ChIP-seq with convolutional neural networks. *Bioinformatics* **33**:i225–i233. doi:10.1093/bioinformatics/btx243

Kojima K, Tadaka S, Katsuoka F, Tamiya G, Yamamoto M, Kinoshita K. 2020. A genotype imputation method for de-identified haplotype reference information by using recurrent neural network. *PLoS Computational Biology* **16**:e1008207. doi:10.1371/journal.pcbi.1008207

Kowalski MH, Qian H, Hou Z, Rosen JD, Tapia AL, Shan Y, Jain D, Argos M, Arnett DK, Avery C, Barnes KC, Becker LC, Bien SA, Bis JC, Blangero J, Boerwinkle E, Bowden DW, Buyske S, Cai J, Cho MH, Choi SH, Choquet H, Adrienne Cupples L, Cushman M, Daya M, de Vries PS, Ellinor PT, Faraday N, Fornage M, Gabriel S, Ganesh SK, Graff M, Gupta N, He J, Heckbert SR, Hidalgo B, Hodonsky CJ, Irvin MR, Johnson AD, Jorgenson E, Kaplan R, Kardia SLR, Kelly TN, Kooperberg C, Lasky-Su JA, Loos RJF, Lubitz SA, Mathias RA, McHugh CP, Montgomery C, Moon JY, Morrison AC, Palmer ND, Pankratz N, Papanicolaou GJ, Peralta JM, Peyser PA, Rich SS, Rotter JI, Silverman EK, Smith JA, Smith NL, Taylor KD, Thornton TA, Tiwari HK, Tracy RP, Wang T, Weiss ST, Weng LC, Wiggins KL, Wilson JG, Yanek LR, Zöllner S, North KE, Auer PL, Raffield LM, Reiner AP, Li Y. 2019. Use of >100,000 NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium whole genome sequences improves imputation quality and detection of rare variant associations in admixed African and Hispanic/Latino populations. *PLoS Genetics* **15**:e1008500. doi:10.1371/journal.pgen.1008500

Lal A, Chiang ZD, Yakovenko N, Duarte FM, Israeli J, Buenrostro JD. 2021. Deep learning-based enhancement of epigenomics data with AtacWorks. *Nature Communications* **12**. doi:10.1038/s41467-021-21765-5

Li Y, Willer C, Sanna S, Abecasis G. 2009. Genotype Imputation. *Annual Review of Genomics and Human Genetics* **10**:387–406. doi:10.1146/annurev.genom.9.081307.164242

Lin T-Y, Goyal P, Girshick R, He K, Dollár P. 2017. Focal Loss for Dense Object Detection.

Liu Y, Gu J, Goyal N, Li X, Edunov S, Ghazvininejad M, Lewis M, Zettlemoyer L. 2020. Multilingual Denoising Pre-training for Neural Machine Translation. *Transactions of the Association for Computational Linguistics* **8**:726–742. doi:10.1162/tacl_a_00343

Marchini J, Howie B. 2010. Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*. doi:10.1038/nrg2796

McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, Kang HM, Fuchsberger C, Danecek P, Sharp K, Luo Y, Sidore C, Kwong A, Timpson N, Koskinen S, Vrieze S, Scott LJ, Zhang H, Mahajan A, Veldink J, Peters U, Pato C, van Duijn CM, Gillies CE, Gandin I, Mezzavilla M, Gilly A, Cocca M, Traglia M, Angius A, Barrett JC, Boomsma D, Branham K, Breen G, Brummett CM, Busonero F, Campbell H, Chan A, Chen S, Chew E, Collins FS, Corbin LJ, Smith GD, Dedoussis G, Dorr M, Farmaki AE, Ferrucci L, Forer L, Fraser RM, Gabriel S, Levy S, Groop L, Harrison T, Hattersley A, Holmen OL, Hveem K, Kretzler M, Lee JC, McGue M, Meitinger T, Melzer D, Min JL, Mohlke KL, Vincent JB, Nauck M, Nickerson D, Palotie A, Pato M, Pirastu N, McInnis M, Richards JB, Sala C, Salomaa V, Schlessinger D, Schoenherr S, Slagboom PE, Small K, Spector T, Stambolian D, Tuke M, Tuomilehto J, van den Berg LH, van Rheenen W, Volker U, Wijmenga C, Toniolo D, Zeggini E, Gasparini P, Sampson MG, Wilson JF, Frayling T, de Bakker PIW, Swertz MA, McCarroll S, Kooperberg C, Dekker A, Altshuler D, Willer C, Iacono W, Ripatti S, Soranzo N, Walter K, Swaroop A, Cucca F, Anderson CA, Myers RM, Boehnke M, McCarthy MI, Durbin R, Abecasis G, Marchini J. 2016. A reference panel of 64,976 haplotypes for genotype imputation. *Nature Genetics* **48**:1279–1283. doi:10.1038/ng.3643

Mou L, Norby FL, Chen LY, O'Neal WT, Lewis TT, Loehr LR, Soliman EZ, Alonso A. 2018. Lifetime Risk of Atrial Fibrillation by Race and Socioeconomic Status: ARIC Study (Atherosclerosis Risk in Communities). *Circulation: Arrhythmia and Electrophysiology* **11**. doi:10.1161/CIRCEP.118.006350

Naito T, Suzuki K, Hirata J, Kamatani Y, Matsuda K, Toda T, Okada Y. 2021. A deep learning method for HLA imputation and trans-ethnic MHC fine-mapping of type 1 diabetes. *Nature Communications* **12**:1–14. doi:10.1038/s41467-021-21975-x

Okewu E, Adewole P, Sennaike O. 2019. Experimental Comparison of Stochastic Optimizers in Deep LearningLecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Springer Verlag. pp. 704–715. doi:10.1007/978-3-030-24308-1_55

Picard toolkit. 2019.

Rubinacci S, Delaneau O, Marchini J. 2020. Genotype imputation using the Positional Burrows Wheeler Transform. *PLOS Genetics* **16**:e1009049. doi:10.1371/journal.pgen.1009049

Sarkar E, Chielle E, Gürsoy G, Mazonka O, Gerstein M, Maniatakos M. 2021. Fast and scalable private genotype imputation using machine learning and partially homomorphic encryption. *IEEE Access*.

Sun Y v., Kardia SLR. 2008. Imputing missing genotypic data of single-nucleotide polymorphisms using neural networks. *European Journal of Human Genetics* **16**:487–495. doi:10.1038/sj.ejhg.5201988

Tian C, Fei L, Zheng W, Xu Y, Zuo W, Lin CW. 2020. Deep learning on image denoising: An overview. *Neural Networks*. doi:10.1016/j.neunet.2020.07.025

Voulodimos A, Doulamis N, Doulamis A, Protopapadakis E. 2018. Deep Learning for Computer Vision: A Brief Review. *Computational Intelligence and Neuroscience*. doi:10.1155/2018/7068349

Supplemental Figures

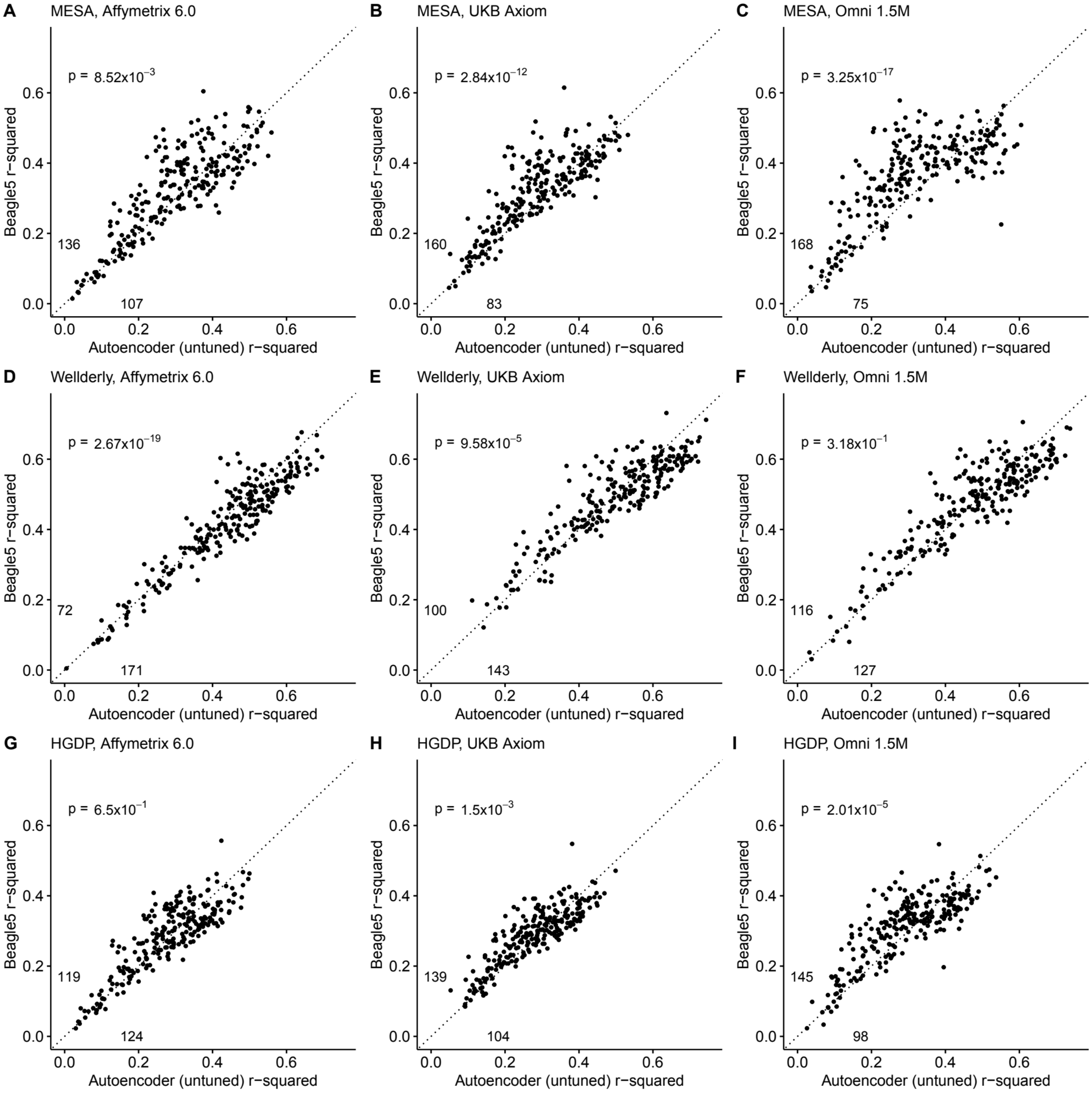


Fig. S1. Beagle5 (y-axis) versus autoencoder-based (x-axis) imputation accuracy prior to tuning. Beagle5 and untuned autoencoders were tested across three independent datasets - MESA (**top**), Wellderly (**middle**), and HGDP (**bottom**) - and across three genotyping array platforms - Affymetrix 6.0 (**left**), UKB Axiom (**middle**), Omni1.5M (**right**). Each data point

represents the imputation accuracy (average r-squared per variant) for an individual genomic segment relative to its WGS-based ground truth. The numerical values presented on the left side and below the identity line (dashed line) indicate the number of genomic segments in which Beagle5 outperformed the untuned autoencoder (left of identity line) and the number of genomic segments in which the untuned autoencoder surpassed Beagle5 (below the identity line). Statistical significance was assessed through two-proportion Z-test p-values.

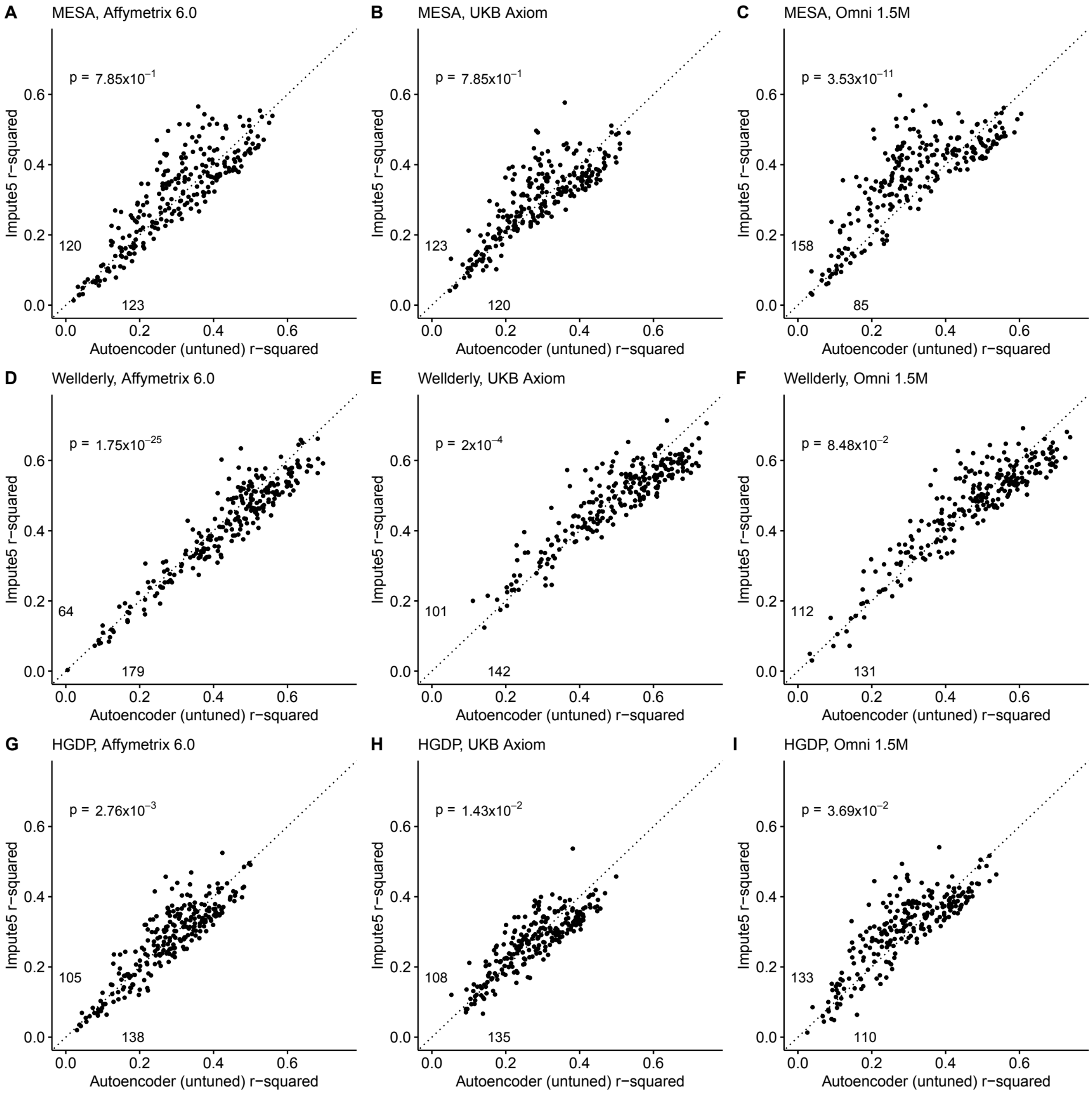


Fig. S2. Impute5 (y-axis) versus autoencoder-based (x-axis) imputation accuracy prior to tuning. Impute5 and untuned autoencoders were tested across three independent datasets - MESA (**top**), Wellderly (**middle**), and HGDP (**bottom**) - and across three genotyping array platforms - Affymetrix 6.0 (**left**), UKB Axiom (**middle**), Omni1.5M (**right**). Each data point represents the imputation accuracy (average r-squared per variant) for an individual genomic segment relative to its WGS-based ground truth. The numerical values presented on the left side and below the identity line

(dashed line) indicate the number of genomic segments in which Impute5 outperformed the untuned autoencoder (left of identity line) and the number of genomic segments in which the untuned autoencoder surpassed Impute5 (below the identity line). Statistical significance was assessed through two-proportion Z-test p-values.

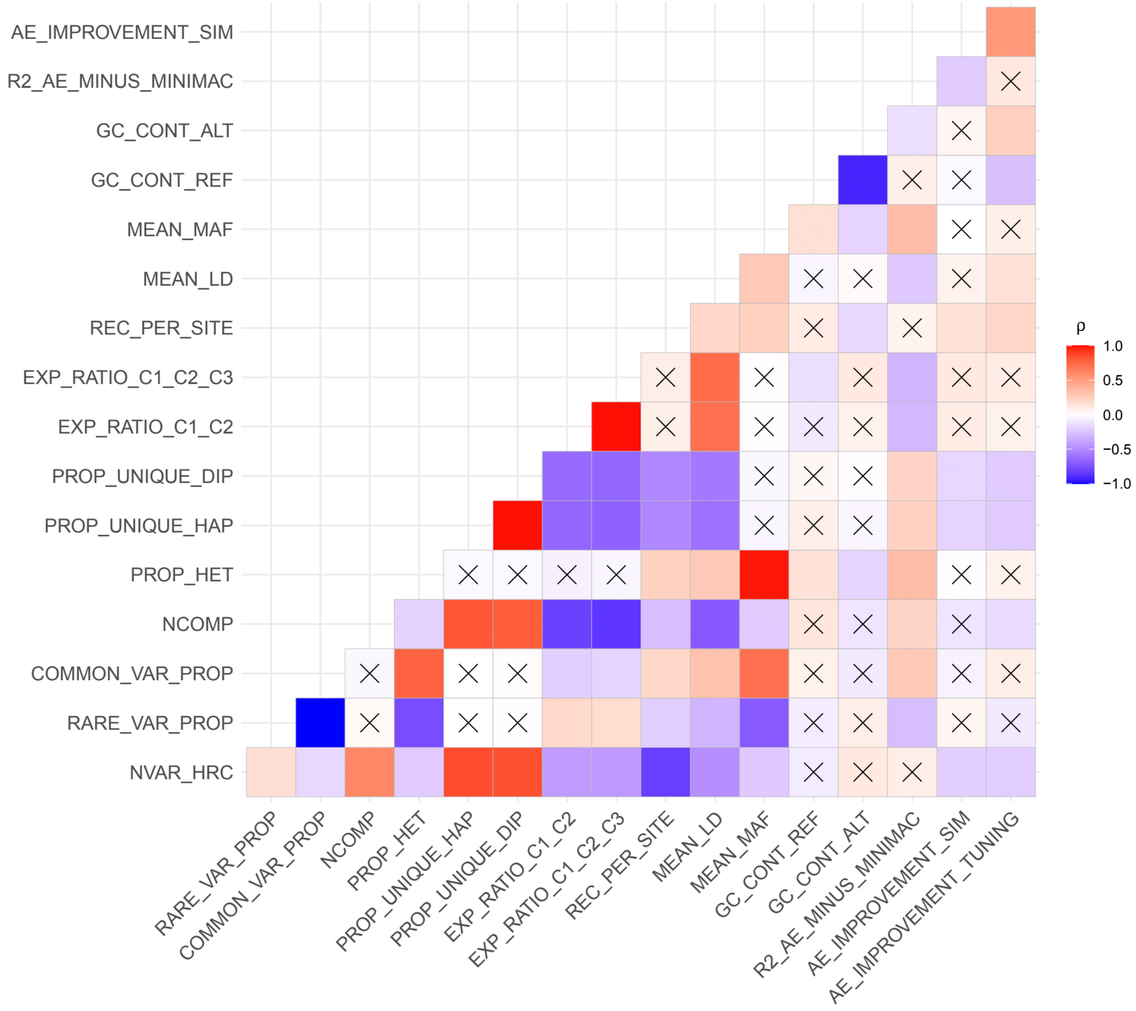
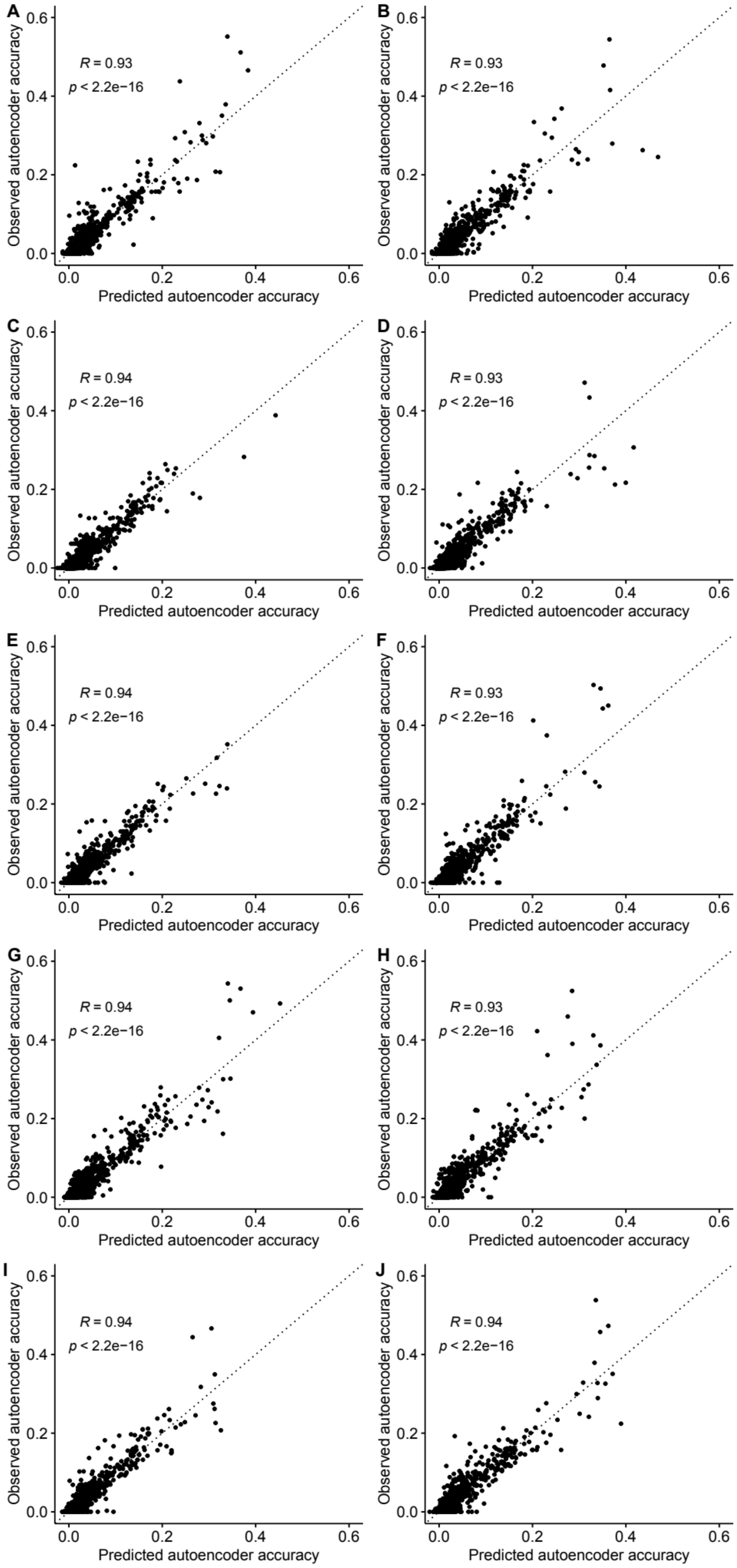


Fig. S3. Relationship between genomic segment features and autoencoder performance. Spearman correlations (ρ) between genomic segment features and autoencoder performance metrics are presented. An *"X"* denotes Spearman correlations that are not statistically significant (p>0.05). The performance metrics include the mean validation accuracy of Minimac4 and autoencoder (R2_AE_MINUS_MINIMAC), the autoencoder's improvement in accuracy observed after offspring formation (AE_IMPROVEMENT_SIM) and the autoencoder's improvement in accuracy after fine tuning of hyperparameters (AE_IMPROVEMENT_TUNING). The genomic features include the total number of variants per genomic segment in HRC (NVAR_HRC), proportion of rare variants at MAF≤0.5% threshold (RARE_VAR_PROP), proportion of common

variants at MAF>0.5% threshold (COMMON_VAR_PROP), number of components needed to explain at least 90% of variance after running Principal Component Analysis (NCOMP), proportion of heterozygous genotypes (PROP_HET), proportion of unique haplotypes (PROP_UNIQUE_HAP) and diplotypes (PROP_UNIQUE_DIP), sum of ratios of explained variance from first two (EXP_RATIO_C1_C2) and three (EXP_RATIO_C1_C2_C3) components from Principal Component Analysis, recombination per variant per variant (REC_PER_SITE), mean pairwise correlation across all variants in each genomic segment (MEAN_LD), mean MAF (MEAN_MAF), GC content of reference alleles (GC_CONT_REF), GC content of alternate alleles (GC_CONT_ALT).



1 **Fig. S4. Projecting autoencoder performance from hyperparameters and genomic features.**

We developed an ensemble-based machine learning approach (Extreme Gradient Boosting - XGBoost) to predict the expected performance (r-squared) of each hyperparameter combination per genomic segment using the results of the coarse-grid search and predictive features calculated for each genomic segment (see Methods). We plot the observed accuracy of trained autoencoders versus the accuracy predicted by the XGBoost model after 10-fold cross-validation. Each subplot shows one iteration of the 10-fold validation process and its respective Pearson correlation between the predicted and observed accuracy values.

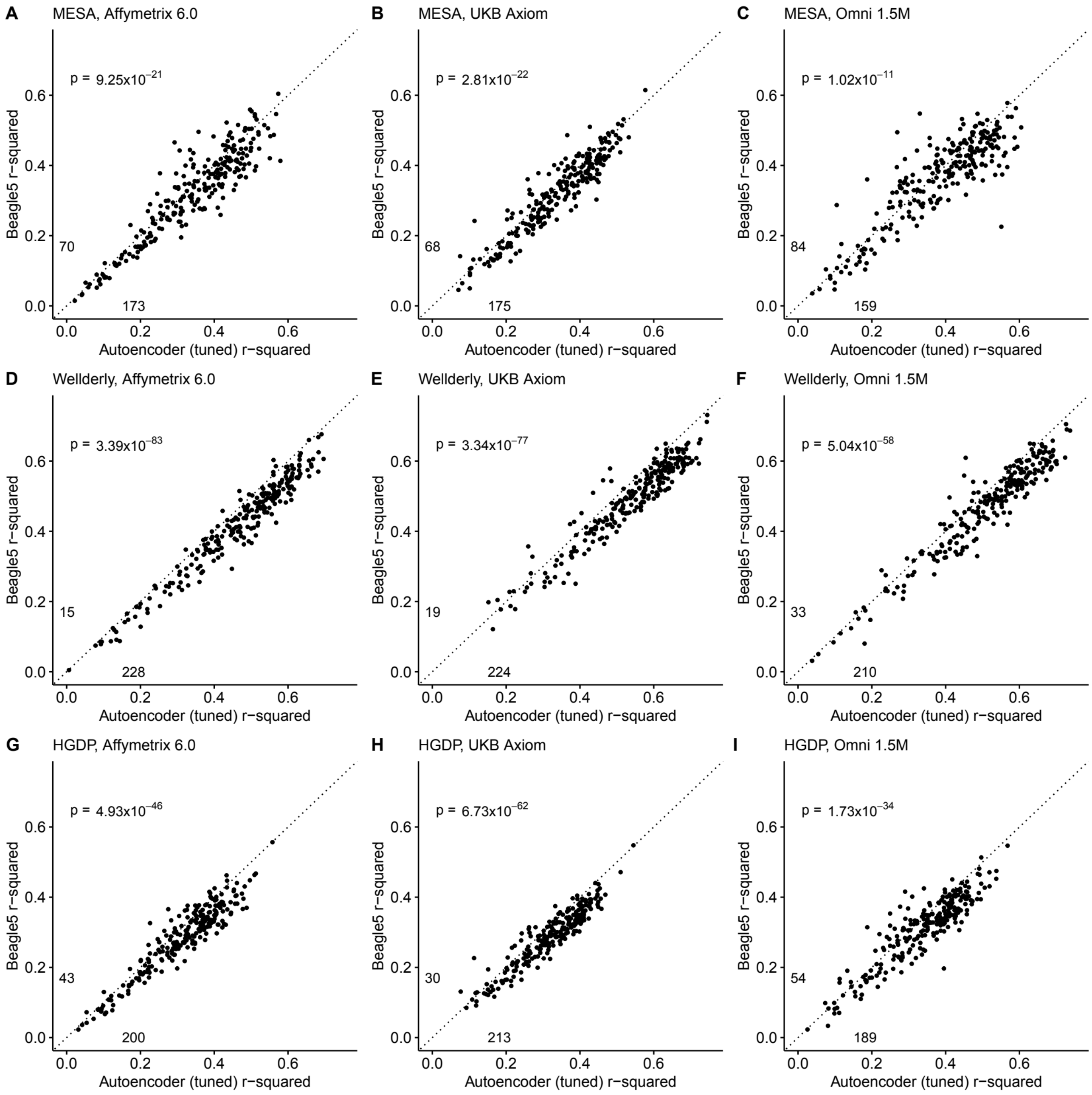


Fig. S5. Beagle5 (y-axis) versus autoencoder-based (axis) imputation accuracy after tuning. Beagle5 and tuned autoencoders were validated across three independent datasets - MESA (**top**), Wellderly (**middle**), and HGDP (**bottom**) - and across three genotyping array platforms - Affymetrix 6.0 (**left**), UKB Axiom (**middle**), Omni1.5M (**right**). Each data point represents the imputation accuracy (average r-squared per variant) for an individual genomic segment relative to its WGS-based ground truth. The numerical values presented on the left side and below the identity line (dashed line) indicate the number of genomic segments in which Beagle5 outperformed the untuned autoencoder (left of identity line) and the number of genomic segments in which the untuned autoencoder surpassed Beagle5 (below the identity line). Statistical significance was assessed through two-proportion Z-test p-values.

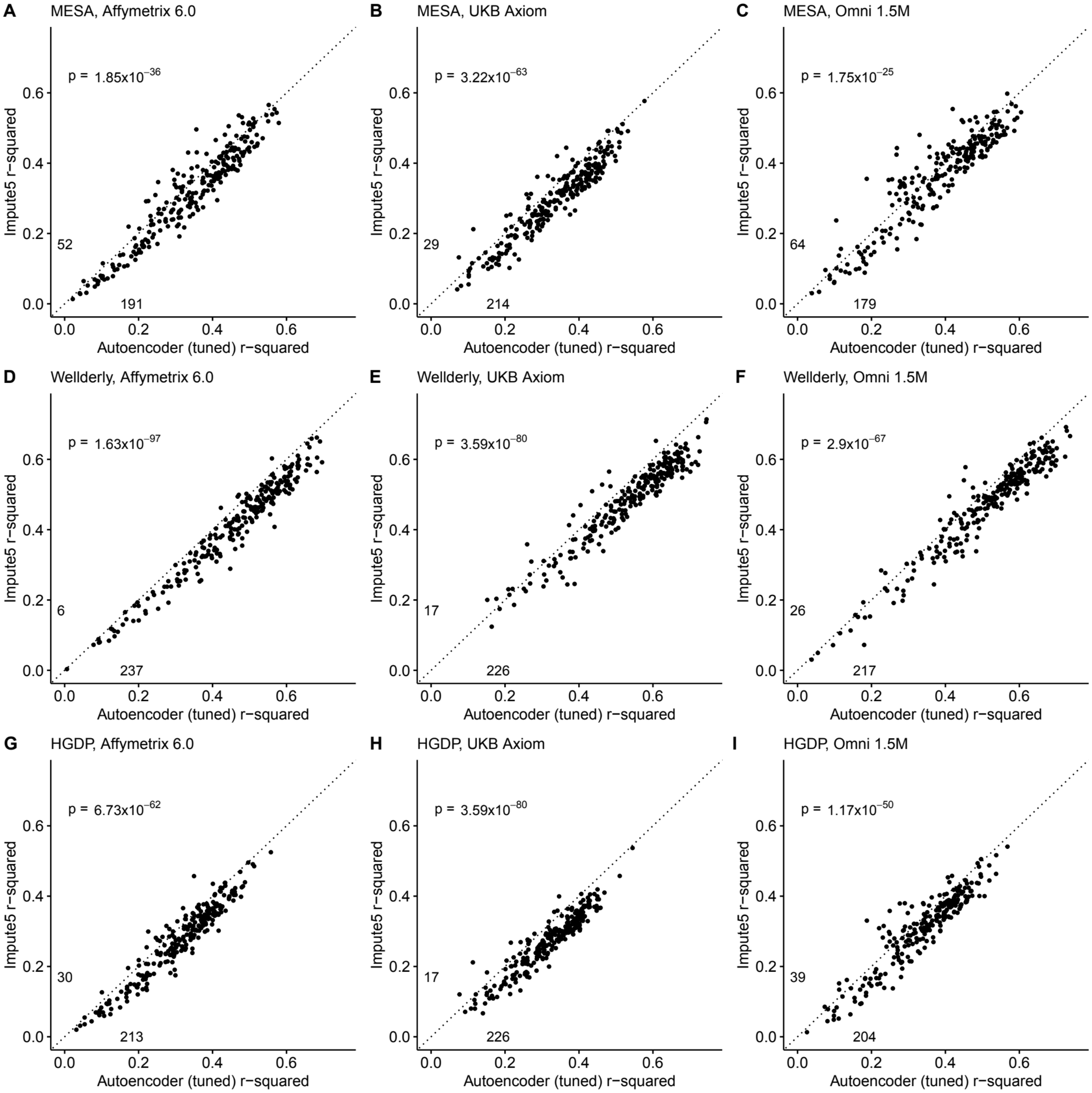


Fig. S6. Impute5 (y-axis) versus autoencoder-based (axis) imputation accuracy after tuning. Impute5 and tuned autoencoders were validated across three independent datasets - MESA (**top**), Wellderly (**middle**), and HGDP (**bottom**) - and across three genotyping array platforms - Affymetrix 6.0 (**left**), UKB Axiom (**middle**), Omni1.5M (**right**). Each data point represents the imputation accuracy (average r-squared per variant) for an individual genomic segment relative to its WGS-based ground truth. The numerical values presented on the left side and below the identity line (dashed line) indicate the number of genomic segments in which Impute5 outperformed the untuned autoencoder (left of identity line) and the number of genomic segments in which the untuned autoencoder surpassed Impute5 (below the identity line). Statistical significance was assessed through two-proportion Z-test p-values.

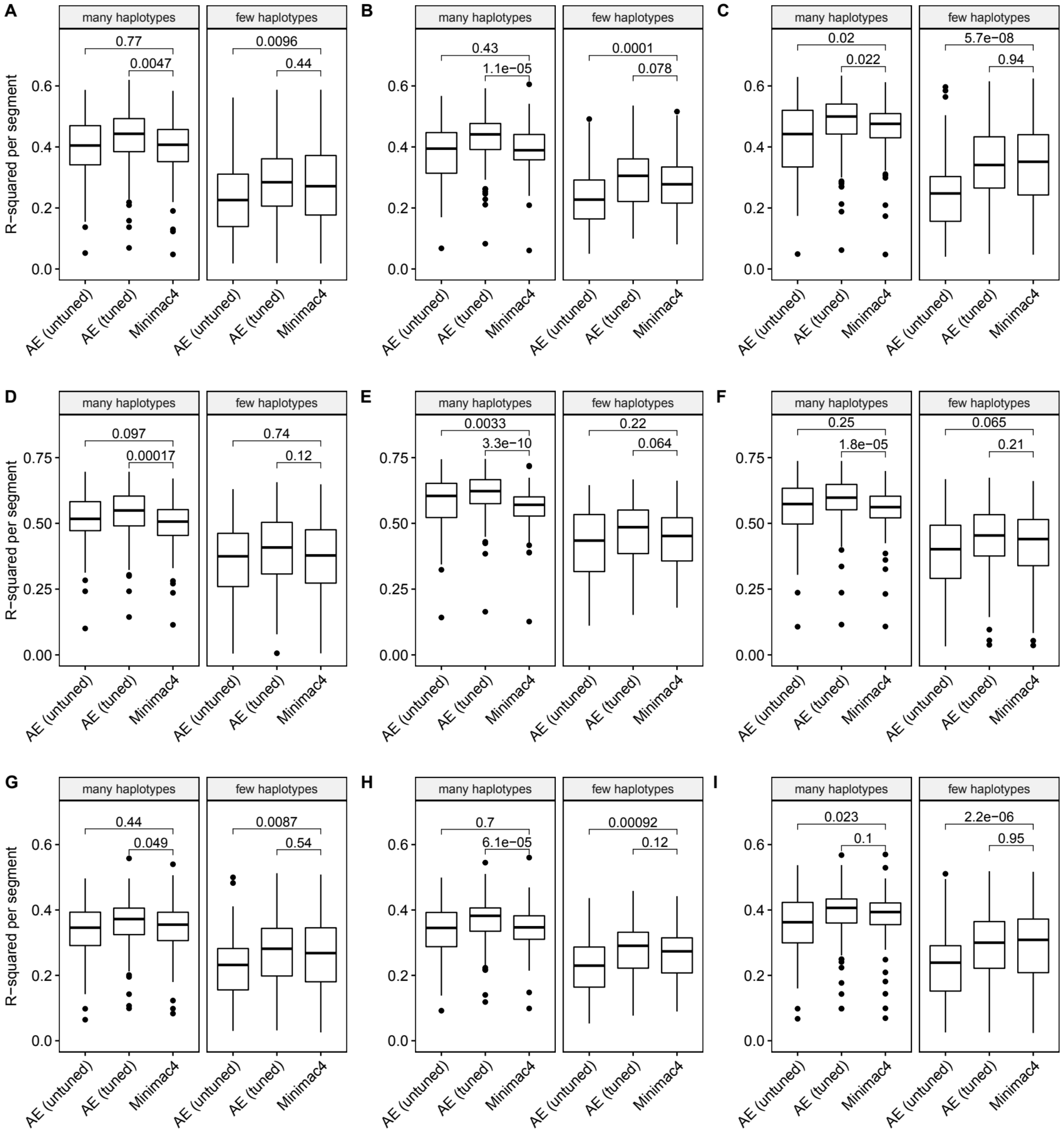


Fig. S7. Imputation accuracy as a function of unique haplotype abundance. Minimac4 and tuned and untuned autoencoders (AE) were tested across three independent datasets - MESA (**top**), Wellderly (**middle**), and HGDP (**bottom**) - and across three genotyping array platforms - Affymetrix 6.0 (**left**), UKB Axiom (**middle**), Omni1.5M (**right**). “Many” vs “Few” haplotypes are

defined by splitting genomic segments into those with greater than vs less than the median number of unique haplotypes per genomic segment. We applied Wilcoxon rank-sum tests to compare the untuned and tuned autoencoder to Minimac4. The validation datasets consist of: A) MESA Affymetrix 6.0; B) MESA UKB Axiom; C) MESA Omni 1.5M; D) Wellderly Affymetrix 6.0; E) Wellderly UKB Axiom; F) Wellderly Omni 1.5M; G) HGDP Affymetrix 6.0; H) HGDP UKB Axiom; I) HGDP Omni 1.5M.

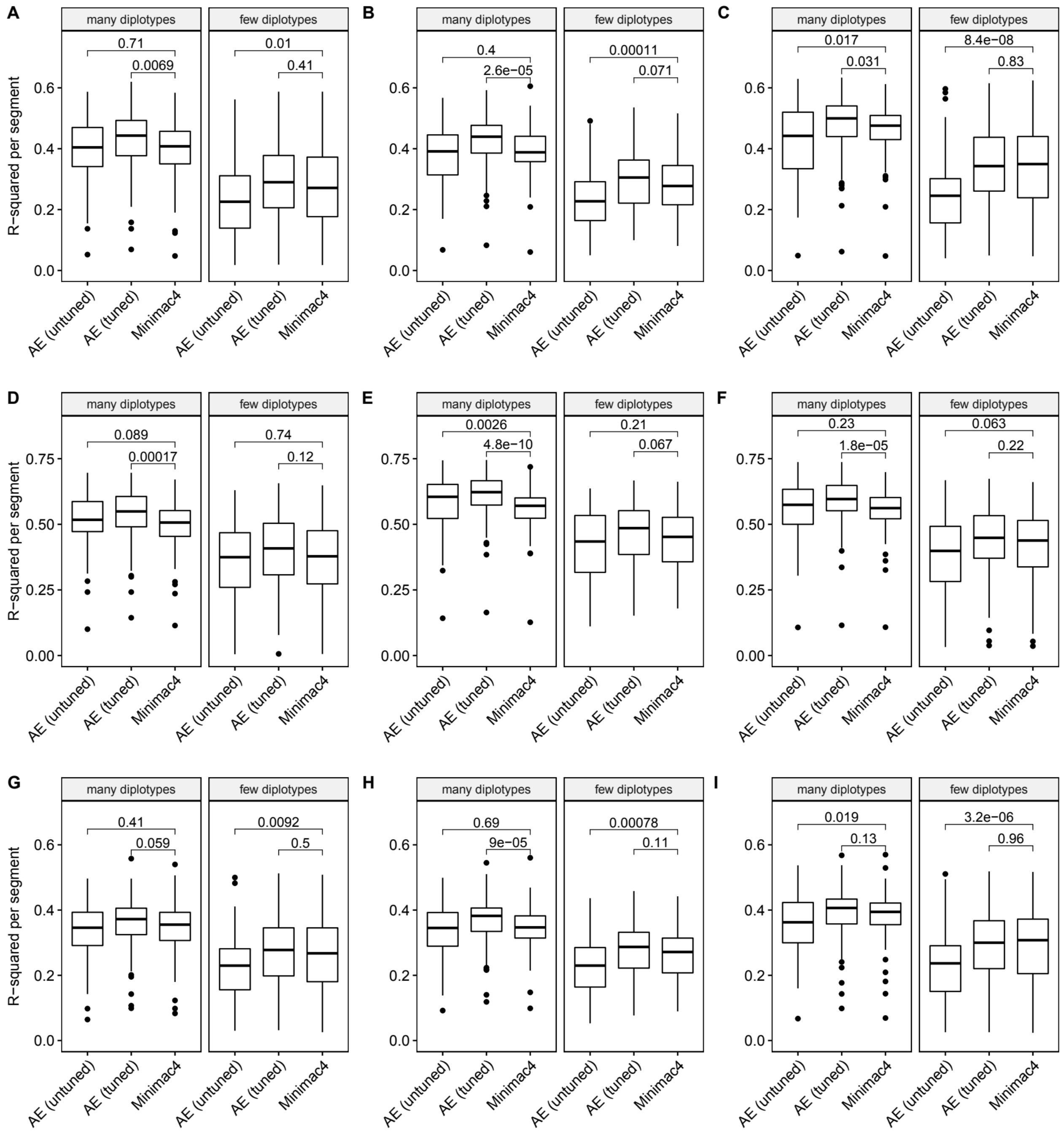


Fig. S8. Imputation accuracy as a function of unique diplotype abundance. Minimac4 and tuned and untuned autoencoders (AE) were tested across three independent datasets - MESA (**top**), Wellderly (**middle**), and HGDP (**bottom**) - and across three genotyping array platforms - Affymetrix 6.0 (**left**), UKB Axiom (**middle**), Omni1.5M (**right**). “Many” vs “Few” diplotypes are

defined by splitting genomic segments into those with greater than vs less than the median number of unique diplotypes per genomic segment.. We applied Wilcoxon rank-sum tests to compare the untuned and tuned autoencoder to Minimac4. The validation datasets consist of: A) MESA Affymetrix 6.0; B) MESA UKB Axiom; C) MESA Omni 1.5M; D) Wellderly Affymetrix 6.0; E) Wellderly UKB Axiom; F) Wellderly Omni 1.5M; G) HGDP Affymetrix 6.0; H) HGDP UKB Axiom; I) HGDP Omni 1.5M.

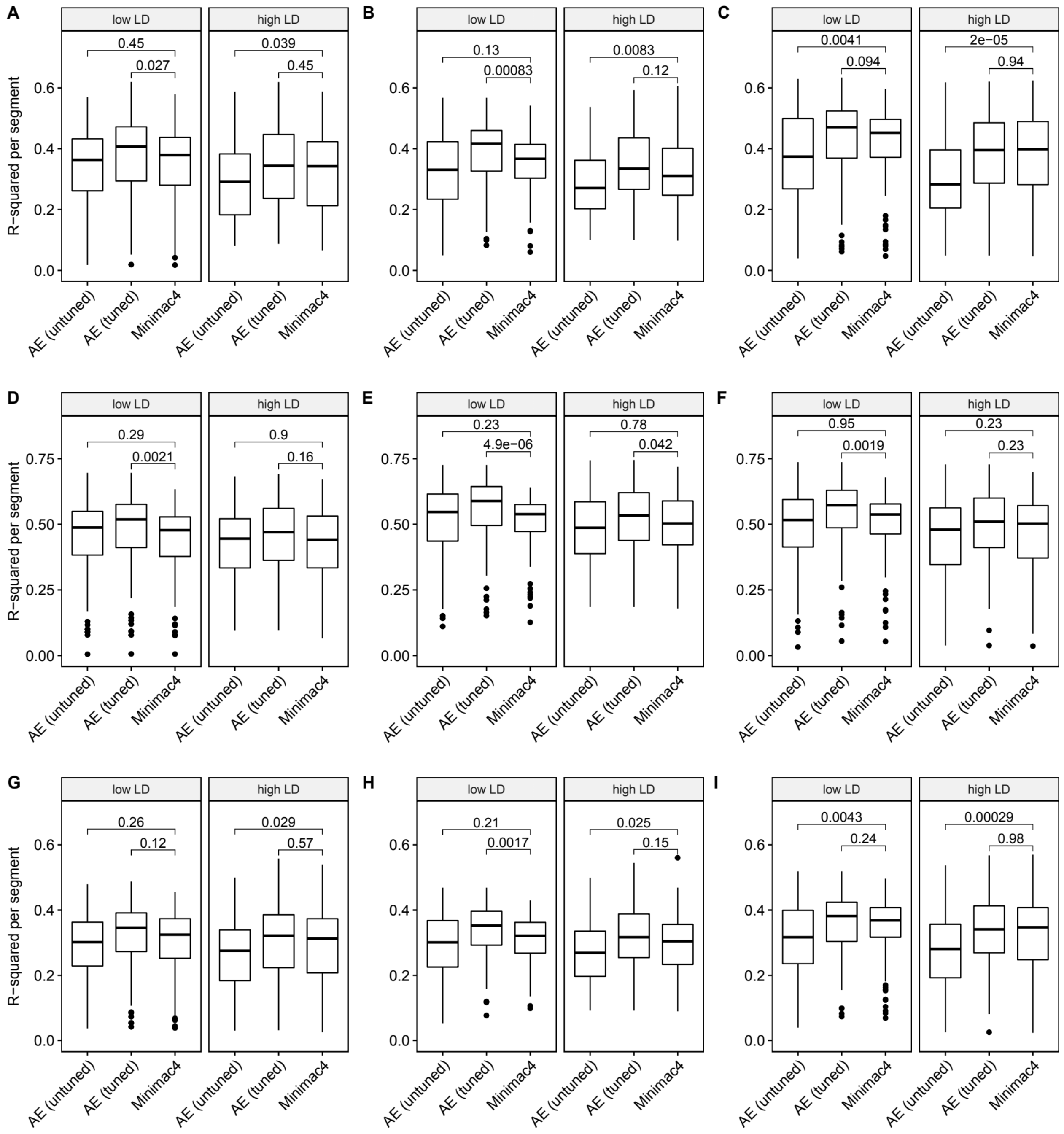


Fig. S9. Imputation accuracy as a function of linkage disequilibrium (LD). Minimac4 and tuned and untuned autoencoders (AE) were tested across three independent datasets - MESA (**top**), Wellderly (**middle**), and HGDP (**bottom**) - and across three genotyping array platforms - Affymetrix 6.0 (**left**), UKB Axiom (**middle**), Omni1.5M (**right**). “High” vs “Low” LD is defined by splitting

genomic segments into those with greater than vs less than the average pairwise LD strength per genomic segment. We applied Wilcoxon rank-sum tests to compare the untuned and tuned autoencoder to Minimac4. The validation datasets consist of: A) MESA Affymetrix 6.0; B) MESA UKB Axiom; C) MESA Omni 1.5M; D) Wellderly Affymetrix 6.0; E) Wellderly UKB Axiom; F) Wellderly Omni 1.5M; G) HGDP Affymetrix 6.0; H) HGDP UKB Axiom; I) HGDP Omni 1.5M.

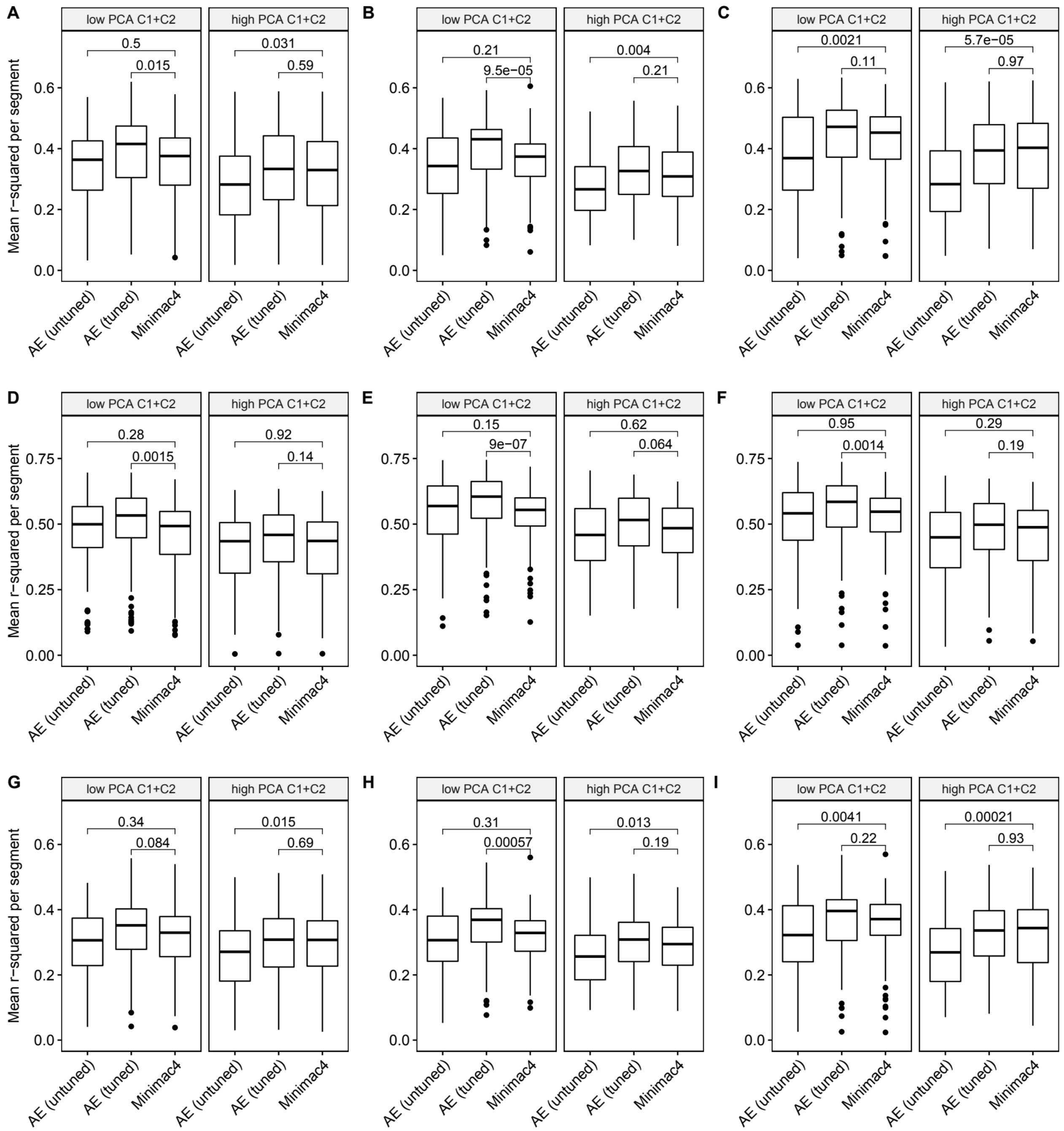


Fig. S10. Imputation accuracy as a function of data complexity. Minimac4 and tuned and untuned autoencoders (AE) were tested across three independent datasets - MESA (**top**), Wellderly (**middle**), and HGDP (**bottom**) - and across three genotyping array platforms - Affymetrix 6.0 (**left**), UKB Axiom (**middle**), Omni1.5M (**right**). "High" vs "Low" data complexity is defined by

splitting genomic segments into those with greater than vs less than the median proportion of variance explained by first two components of Principal Component Analysis per genomic segment (PCA C1+C2). We applied Wilcoxon rank-sum tests to compare the untuned and tuned autoencoder to Minimac4. The validation datasets consist of: A) MESA Affymetrix 6.0; B) MESA UKB Axiom; C) MESA Omni 1.5M; D) Wellderly Affymetrix 6.0; E) Wellderly UKB Axiom; F) Wellderly Omni 1.5M; G) HGDP Affymetrix 6.0; H) HGDP UKB Axiom; I) HGDP Omni 1.5M.

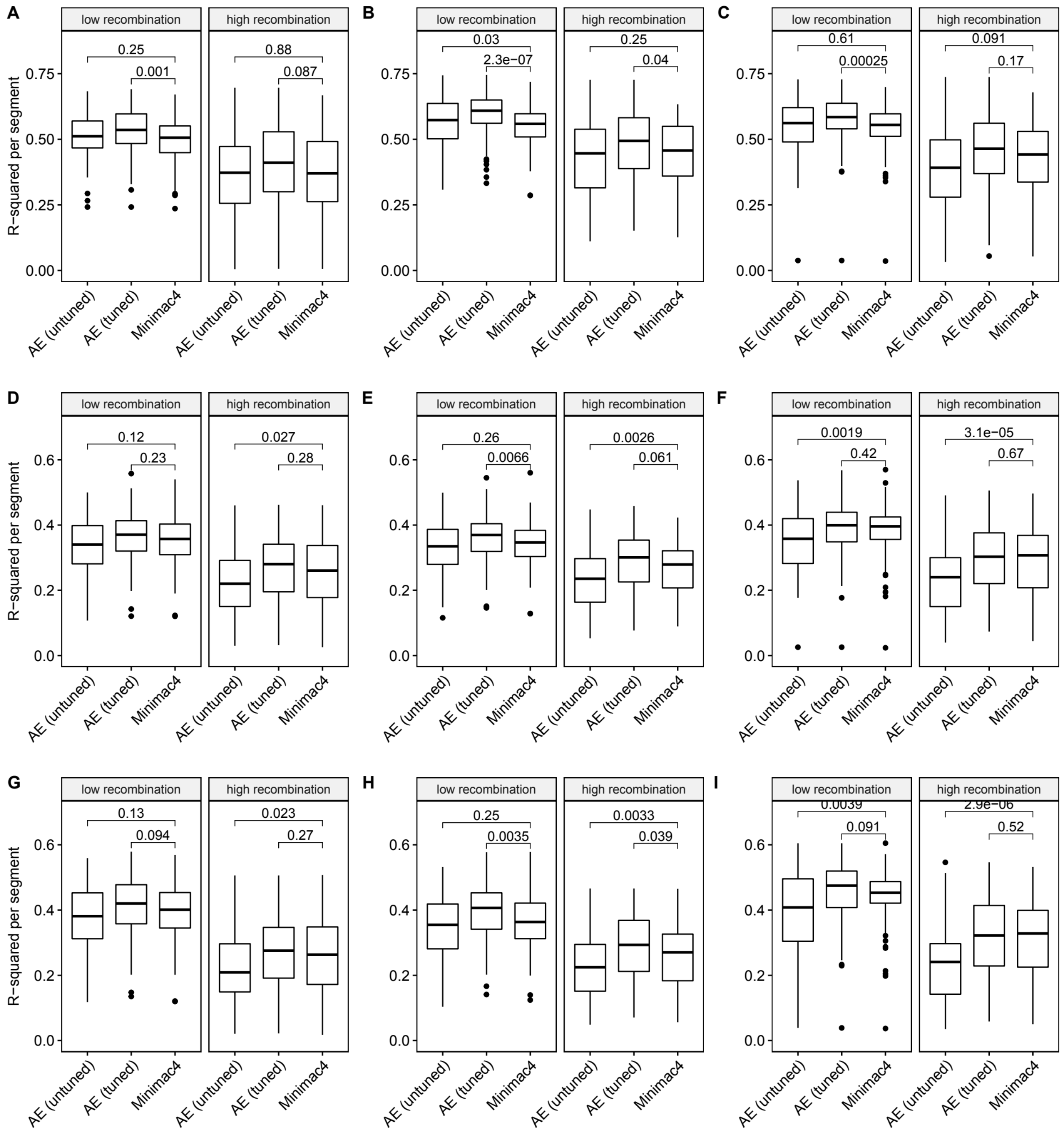


Fig. S11. Imputation accuracy as a function of recombination rate. Minimac4 and tuned and untuned autoencoders (AE) were tested across three independent datasets - MESA (**top**), Wellderly (**middle**), and HGDP (**bottom**) - and across three genotyping array platforms - Affymetrix 6.0 (**left**), UKB Axiom (**middle**), Omni1.5M (**right**). “High” vs “Low” recombination rate is defined

by splitting genomic segments in those with greater than vs less than the median recombination rate per variant per genomic segment. We applied Wilcoxon rank-sum tests to compare the untuned and tuned autoencoder to Minimac4. The validation datasets consist of: A) MESA Affymetrix 6.0; B) MESA UKB Axiom; C) MESA Omni 1.5M; D) Wellderly Affymetrix 6.0; E) Wellderly UKB Axiom; F) Wellderly Omni 1.5M; G) HGDP Affymetrix 6.0; H) HGDP UKB Axiom; I) HGDP Omni 1.5M.

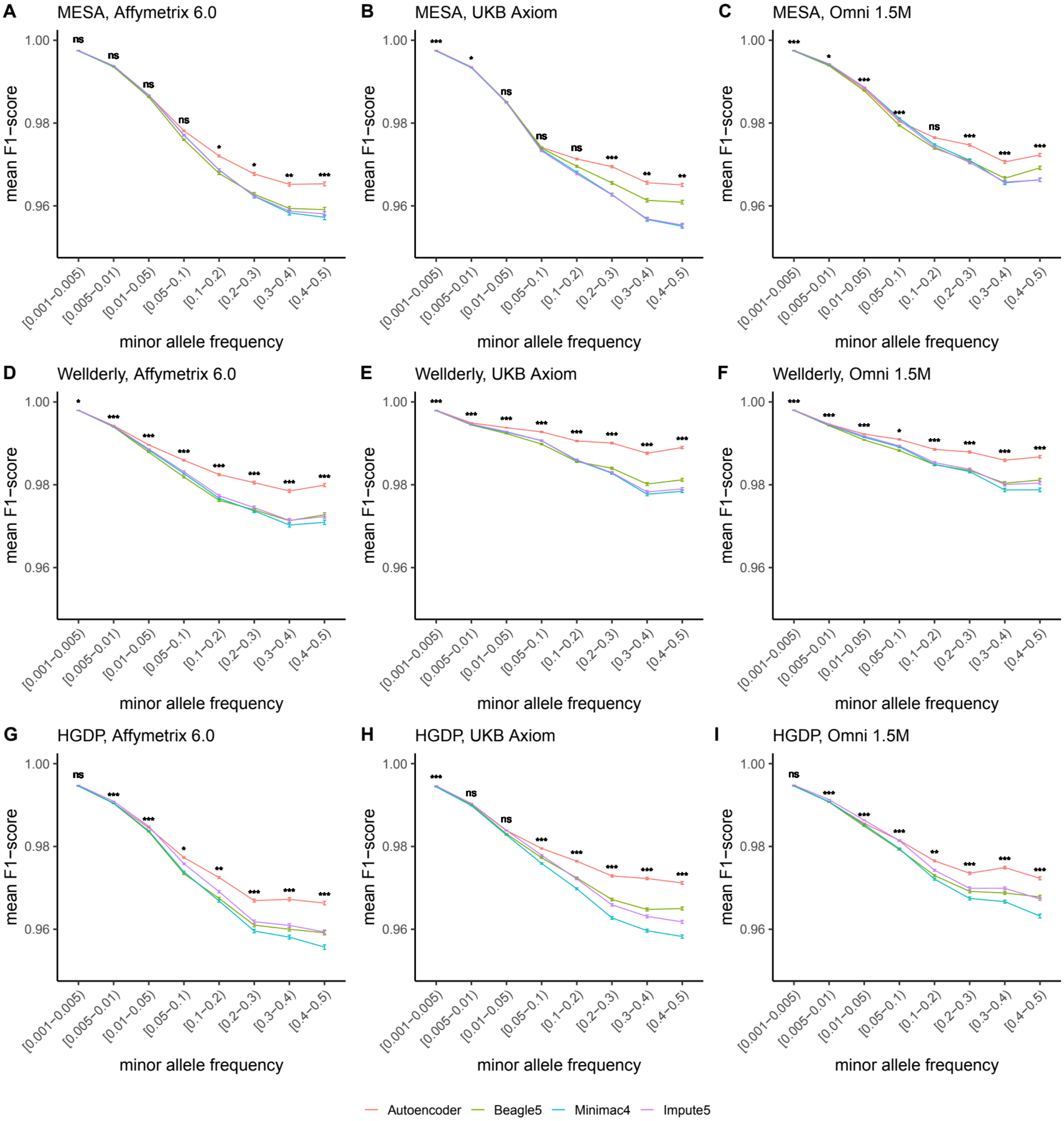


Fig. S12. HMM-based versus autoencoder-based imputation accuracy across MAF bins (F1 score). Autoencoder-based (**red**) and HMM-based (Minimac4 (**blue**), Beagle5 (**green**), and Impute5 (**purple**)) imputation accuracy was validated across three independent datasets - MESA (**top**), Wellderly (**middle**), and HGDP (**bottom**) - and across three genotyping array platforms -

Affymetrix 6.0 (**left**), UKB Axiom (**middle**), Omni1.5M (**right**). Each data point represents the imputation accuracy (mean F1-score per variant) relative to WGS-based ground truth across MAF bins. Error bars represent standard errors. We applied Wilcoxon rank-sum tests to compare the HMM-based tools to the tuned autoencoder (AE). * represents p-values ≤ 0.05, ** indicates p-values ≤ 0.001, and *** indicates p-values ≤ 0.0001, ns represents non-significant p-values.

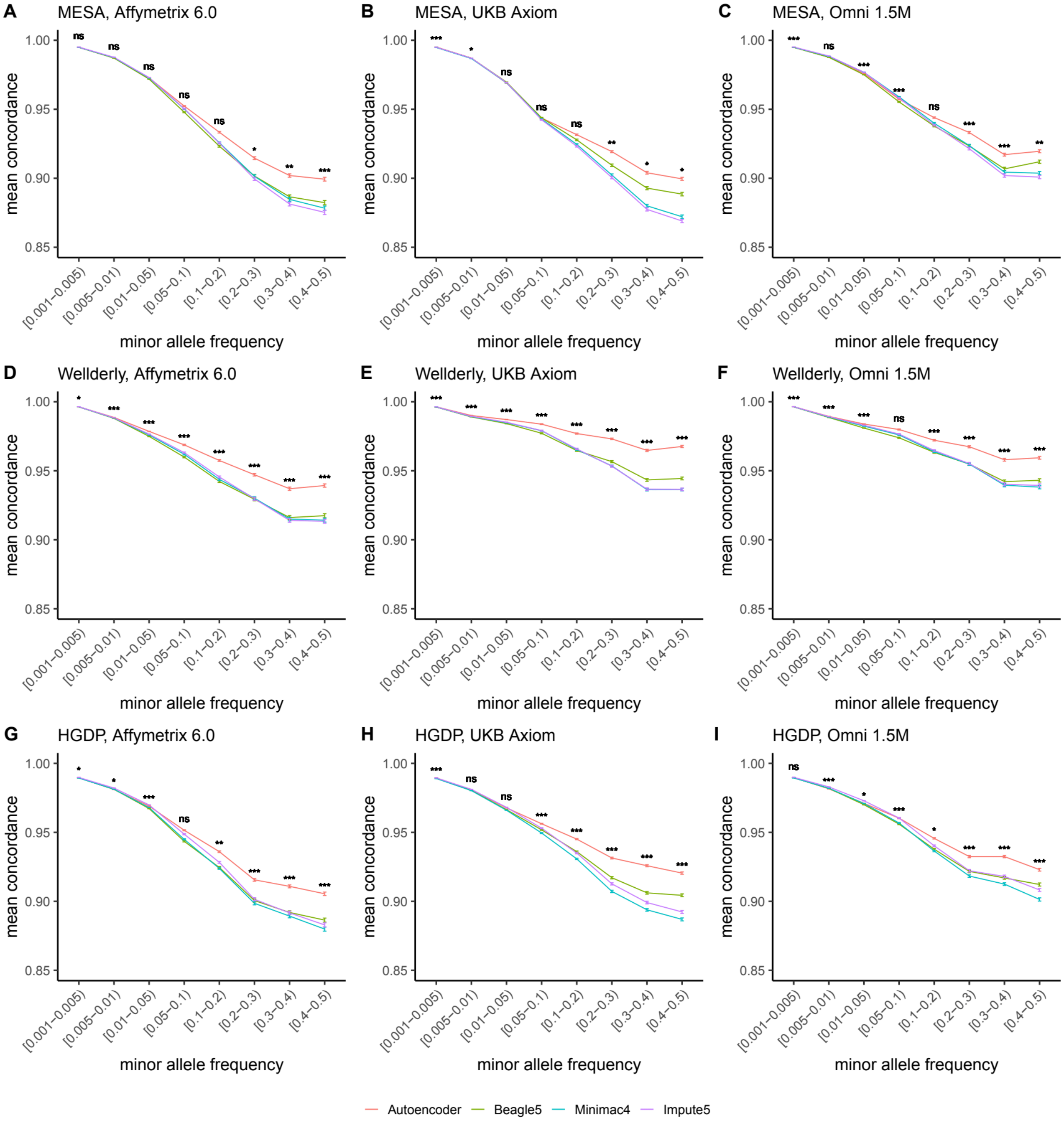


Fig. S13. HMM-based versus autoencoder-based imputation accuracy across MAF bins (concordance). Autoencoder-based (**red**) and HMM-based (Minimac4 (**blue**), Beagle5 (**green**), and Impute5 (**purple**)) imputation accuracy was validated across three independent datasets - MESA (**top**), Wellderly (**middle**), and HGDP (**bottom**) - and across three genotyping array

platforms - Affymetrix 6.0 (**left**), UKB Axiom (**middle**), Omni1.5M (**right**). Each data point represents the imputation accuracy (mean concordance per variant) relative to WGS-based ground truth across MAF bins. Error bars represent standard errors. We applied Wilcoxon rank-sum tests to compare the HMM-based tools to the tuned autoencoder (AE). * represents p-values ≤ 0.05, ** indicates p-values ≤ 0.001, and *** indicates p-values ≤ 0.0001, ns represents non-significant p-values.

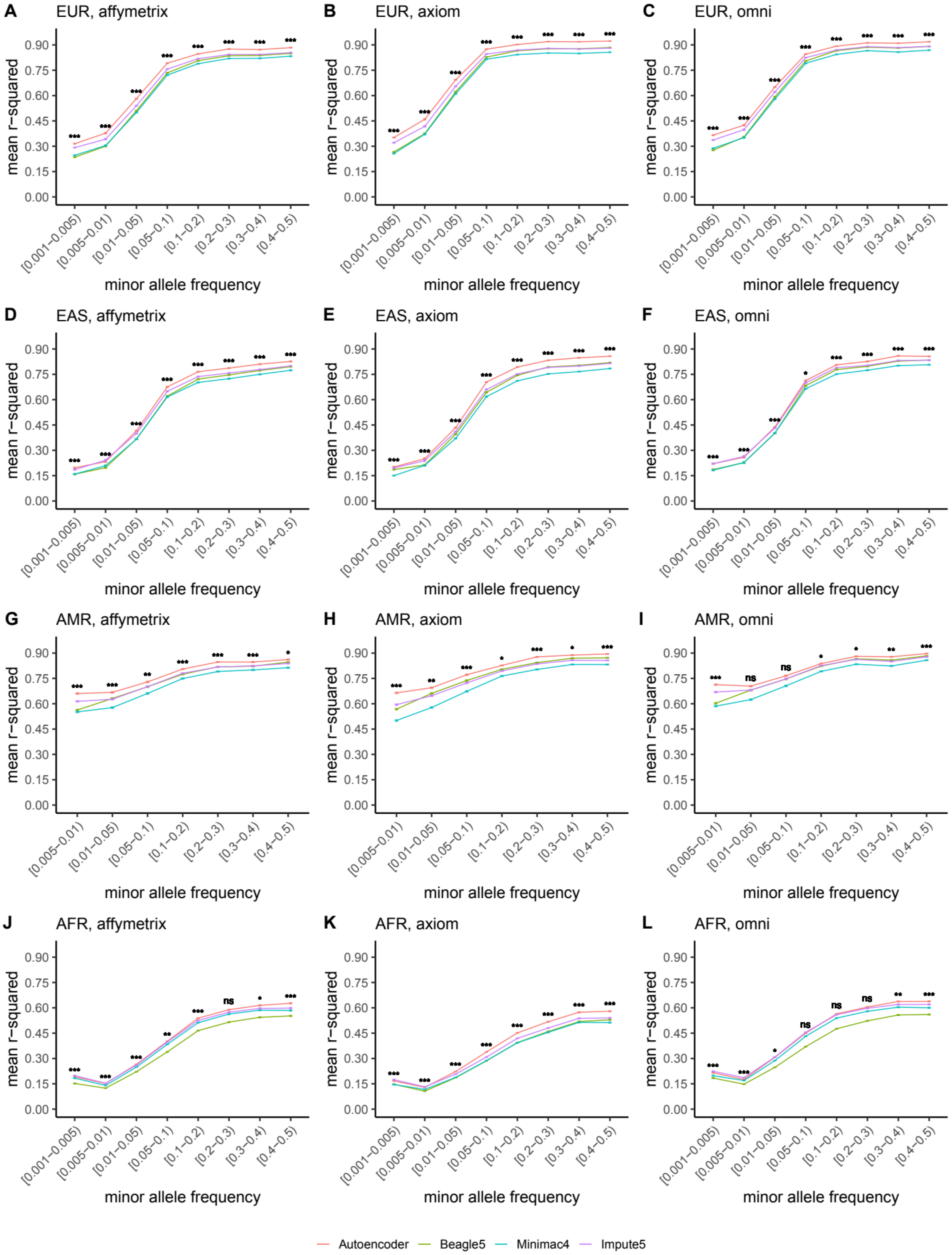

A
EUR, affymetrix
*** *** *** *** *** *** *** ***
B
EUR, axiom
*** *** *** *** *** *** *** ***
C
EUR, omni
*** *** *** *** *** *** *** ***
D
EAS, affymetrix
*** *** *** *** *** *** *** ***
E
EAS, axiom
*** *** *** *** *** *** *** ***
F
EAS, omni
*** *** *** * *** *** *** ***
G
AMR, affymetrix
*** *** ** *** *** *** *
H
AMR, axiom
*** ** *** * *** * ***
I
AMR, omni
*** ns ns * * ** ***
J
AFR, affymetrix
*** *** *** ** *** ns * ***
K
AFR, axiom
*** *** *** *** *** *** *** ***
L
AFR, omni
*** *** * ns ns ns ** ***
mean r−squared
0.00
0.15
0.30
0.45
0.60
0.75
0.90
[0.001−0.005)
[0.005−0.01)
[0.01−0.05)
[0.05−0.1)
[0.1−0.2)
[0.2−0.3)
[0.3−0.4)
[0.4−0.5)
minor allele frequency
Autoencoder
Beagle5
Minimac4
Impute5

Fig. S14. HMM-based versus autoencoder-based imputation accuracy across ancestry groups. Autoencoder-based (**red**) and HMM-based (Minimac4 (**blue**), Beagle5 (**green**), and Impute5 (**purple**)) imputation accuracy was validated across individuals of diverse ancestry from HGDP cohort (EUR: European (**top**); EAS: East Asian (**2nd row**); AMR: Native American (**3rd row**); AFR: African (**bottom**)) and multiple genotype array platforms (Affymetrix 6.0 (**left**), UKB Axiom (**middle**), Omni1.5M (**right**)). Each data point represents the imputation accuracy (average r-squared per variant) relative to WGS-based ground truth across MAF bins. Error bars represent standard errors. We applied Wilcoxon rank-sum tests to compare the HMM-based tools to the tuned autoencoder (AE). * represents p-values ≤ 0.05, ** indicates p-values ≤ 0.001, and *** indicates p-values ≤ 0.0001, ns represents non-significant p-values.

Supplemental Tables

Table S1. Performance comparisons between tuned autoencoder (AE) and HMM-based imputation tools (Minimac4, Beagle5, and Impute5) after applying data augmentation to HMM-based tools.

	MESA	Wellderly	HGDP	Affymetrix 6.0	UKB Axiom	Omni 1.5M	Combined
AE (tuned) vs Minimac4 (augmented)	1.36e-04*	3.49e-06*	1.18e-03*	6.05e-04*	6.98e-08*	1.95e-03*	3.39e-05*
AE (tuned) vs Beagle5 (augmented)	1.71e-05*	1.68e-09*	2.88e-09*	1.54e-06*	3.94e-10*	4.30e-07*	2.30e-08*
AE (tuned) vs Impute5 (augmented)	1.24e-09*	3.15e-15*	5.28e-15*	4.41e-11*	2.47e-18*	4.90e-10*	8.64e-14*
Minimac4 (original vs augmented)	4.91e-02*	2.07E-01	1.03E-01	1.74E-01	4.36e-02*	1.13E-01	9.17E-02
Beagle5 (original vs augmented)	1.21e-02*	8.21E-02	2.35e-02*	8.96E-02	6.59e-03*	5.27E-02	2.58e-02*
Impute5 (original vs augmented)	5.45e-04*	6.89e-05*	1.78e-04*	7.01e-04*	1.16e-05*	4.15e-04*	1.26e-04*
AE (tuned)	0.355±0.007	0.505±0.008	0.327±0.006	0.373±0.008	0.399±0.007	0.414±0.008	0.396±0.007
Minimac4 (augmented)	0.322±0.007	0.462±0.008	0.303±0.006	0.342±0.008	0.358±0.006	0.388±0.007	0.363±0.007
Beagle5 (augmented)	0.316±0.007	0.446±0.008	0.283±0.005	0.327±0.007	0.348±0.006	0.370±0.007	0.349±0.006
Impute5 (augmented)	0.294±0.007	0.416±0.008	0.261±0.006	0.302±0.008	0.318±0.006	0.351±0.008	0.324±0.007

We applied Wilcoxon rank-sum tests to compare the HMM-based tools to the reference tuned autoencoder (AE). * represents p-values ≤ 0.05, ** indicates p-values ≤ 0.001, and *** indicates p-values ≤ 0.0001.

Table S2. Detailed performance comparisons between tuned autoencoder (AE) and HMM-based imputation tools (Minimac4, Beagle5, and Impute5).

Dataset	**MESA**			**Wellderly**			**HGDP**		
array	**Affymetrix 6.0**	**UKB Axiom**	**Omni 1.5M**	**Affymetrix 6.0**	**UKB Axiom**	**Omni 1.5M**	**Affymetrix 6.0**	**UKB Axiom**	**Omni 1.5M**
AE (tuned) vs Minimac4	9.47e-185***	0.00e+00***	7.22e-89***	6.27e-209***	0.00e+00***	2.75e-198***	2.35e-151***	0.00e+00***	5.55e-67***
AE (tuned) vs Beagle5	0.00e+00***	0.00e+00***	0.00e+00***	0.00e+00***	0.00e+00***	0.00e+00***	0.00e+00***	0.00e+00***	0.00e+00***
AE (tuned) vs Impute5	0.00e+00***	0.00e+00***	5.37e-191***	0.00e+00***	0.00e+00***	0.00e+00***	0.00e+00***	0.00e+00***	0.00e+00***
Minimac4 vs Beagle5	1.73e-259***	6.06e-16***	0.00e+00***	2.68e-86***	1.65e-11***	6.87e-64***	0.00e+00***	2.62e-185***	0.00e+00***
Minimac4 vs Impute5	4.87e-38***	3.59e-48***	1.17e-22***	3.05e-74***	5.94e-15***	1.00e-25***	9.43e-261***	0.00e+00***	1.73e-251***
Beagle5 vs Impute5	1.92e-96***	2.75e-09***	1.23e-175***	3.65E-01	1.98E-01	9.53e-09***	1.22e-25***	2.61e-17***	8.36e-57***
AE (tuned)	0.410±0.001	0.395±0.001	0.452±0.001	0.537±0.001	0.605±0.001	0.586±0.001	0.363±0.001	0.364±0.001	0.392±0.001
Minimac4	0.390±0.001	0.364±0.001	0.436±0.001	0.500±0.001	0.557±0.001	0.551±0.001	0.350±0.001	0.340±0.001	0.385±0.001
Beagle5	0.383±0.001	0.379±0.001	0.420±0.001	0.484±0.001	0.549±0.001	0.534±0.001	0.326±0.001	0.328±0.001	0.353±0.001
Impute5	0.384±0.001	0.356±0.001	0.429±0.001	0.485±0.001	0.547±0.001	0.539±0.001	0.328±0.001	0.314±0.001	0.359±0.001

Validation accuracies were stratified by dataset (MESA, Wellderly, HGDP) and genotype array platform (Affymetrix 6.0, UKB Axiom, Omni 1.5M). We applied Wilcoxon rank-sum tests to compare the HMM-based tools to the reference tuned autoencoder (AE). * represents p-values ≤ 0.05, ** indicates p-values ≤ 0.001, and *** indicates p-values ≤ 0.0001.

Table S3. Detailed performance comparisons between tuned autoencoder (AE) and HMM-based imputation tools (Minimac4, Beagle5, and Impute5).

Dataset	Array	MAF	AE (tuned) vs Minimac4	AE (tuned) vs Beagle5	AE (tuned) vs Impute5	Minimac4 vs Beagle5	Minimac4 vs Impute5	Beagle5 vs Impute5	AE (tuned)	Minimac4	Beagle5	Impute5
MESA	**Affymetrix 6.0**	**[0.001-0.005)**	3.84e-306***	0.00e+00***	0.00e+00***	0.00e+00***	1.30e-122***	0.00e+00***	0.141±0.001	0.128±0.001	0.129±0.001	0.121±0.001
MESA	**Affymetrix 6.0**	**[0.005-0.01)**	1.37e-48***	3.39e-280***	8.96e-86***	3.12e-127***	1.13e-08***	1.22e-68***	0.280±0.001	0.266±0.001	0.258±0.001	0.262±0.001
MESA	**Affymetrix 6.0**	**[0.01-0.05)**	2.62e-48***	6.14e-119***	3.54e-72***	4.61e-23***	8.85e-05***	7.81e-09***	0.490±0.001	0.467±0.001	0.453±0.001	0.461±0.001
MESA	**Affymetrix 6.0**	**[0.05-0.1)**	2.11e-05***	4.44e-49***	2.36e-04**	3.26e-23***	7.11E-01	3.53e-24***	0.728±0.002	0.703±0.002	0.684±0.002	0.698±0.002
MESA	**Affymetrix 6.0**	**[0.1-0.2)**	1.17e-15***	2.04e-50***	3.36e-09***	6.41e-11***	6.50E-02	8.64e-16***	0.793±0.002	0.763±0.002	0.753±0.002	0.758±0.002
MESA	**Affymetrix 6.0**	**[0.2-0.3)**	1.02e-16***	5.61e-25***	6.87e-09***	5.09E-02	2.67e-02*	5.57e-05***	0.825±0.002	0.794±0.002	0.790±0.002	0.789±0.002
MESA	**Affymetrix 6.0**	**[0.3-0.4)**	2.41e-19***	3.04e-28***	2.92e-09***	9.05E-02	1.08e-02*	3.17e-05***	0.834±0.002	0.799±0.002	0.798±0.002	0.795±0.002
MESA	**Affymetrix 6.0**	**[0.4-0.5)**	2.46e-18***	7.67e-23***	3.85e-11***	3.80E-01	7.17E-02	9.77e-03*	0.842±0.002	0.806±0.002	0.805±0.003	0.801±0.003
MESA	**UKB Axiom**	**[0.001-0.005)**	0.00e+00***	0.00e+00***	0.00e+00***	0.00e+00***	1.04e-106***	1.10e-95***	0.145±0.001	0.117±0.001	0.132±0.001	0.108±0.001
MESA	**UKB Axiom**	**[0.005-0.01)**	7.45e-142***	7.16e-191***	4.41e-242***	2.76e-23***	8.03e-18***	1.28E-01	0.255±0.001	0.226±0.001	0.249±0.001	0.216±0.001
MESA	**UKB Axiom**	**[0.01-0.05)**	2.85e-128***	6.38e-41***	1.57e-181***	1.76e-12***	3.39e-07***	2.75e-34***	0.432±0.001	0.400±0.001	0.418±0.001	0.393±0.001
MESA	**UKB Axiom**	**[0.05-0.1)**	5.91e-21***	4.88e-09***	2.14e-23***	2.68e-03*	4.80E-01	3.13e-04**	0.681±0.002	0.652±0.002	0.657±0.002	0.646±0.002
MESA	**UKB Axiom**	**[0.1-0.2)**	1.12e-42***	2.66e-11***	5.58e-37***	9.55e-11***	5.04E-01	2.58e-08***	0.791±0.001	0.758±0.002	0.766±0.002	0.752±0.002
MESA	**UKB Axiom**	**[0.2-0.3)**	8.25e-59***	1.62e-15***	5.06e-54***	2.23e-14***	8.15E-01	1.07e-12***	0.837±0.001	0.796±0.002	0.807±0.002	0.790±0.002
MESA	**UKB Axiom**	**[0.3-0.4)**	8.13e-81***	7.00e-14***	3.82e-72***	3.98e-26***	8.34E-01	2.26e-22***	0.840±0.002	0.795±0.002	0.810±0.002	0.789±0.002
MESA	**UKB Axiom**	**[0.4-0.5)**	5.60e-82***	8.03e-12***	1.69e-79***	6.34e-30***	6.20E-01	3.13e-29***	0.846±0.002	0.800±0.002	0.819±0.002	0.794±0.002

MESA	**Omni 1.5M**	**[0.001-0.005)**	6.78e-179***	0.00e+00***	0.00e+00***	0.00e+00***	8.51e-66***	0.00e+00***	0.174±0.001	0.158±0.001	0.152±0.001	0.148±0.001
MESA	**Omni 1.5M**	**[0.005-0.01)**	1.37e-23***	7.31e-253***	5.27e-53***	6.68e-147***	5.95e-08***	1.44e-90***	0.340±0.001	0.327±0.001	0.301±0.002	0.317±0.001
MESA	**Omni 1.5M**	**[0.01-0.05)**	1.51e-05***	6.15e-118***	9.61e-14***	2.01e-77***	1.54e-03*	6.53e-53***	0.552±0.001	0.542±0.001	0.510±0.001	0.537±0.001
MESA	**Omni 1.5M**	**[0.05-0.1)**	8.96E-01	2.53e-52***	1.91E-01	2.25e-48***	2.57E-01	2.05e-54***	0.759±0.002	0.750±0.002	0.723±0.002	0.749±0.002
MESA	**Omni 1.5M**	**[0.1-0.2)**	7.72e-19***	3.14e-57***	5.03e-09***	5.40e-11***	6.88e-03*	2.65e-19***	0.828±0.001	0.806±0.002	0.797±0.002	0.805±0.002
MESA	**Omni 1.5M**	**[0.2-0.3)**	2.53e-32***	2.68e-65***	6.69e-18***	8.66e-06***	6.61e-03*	6.35e-12***	0.866±0.002	0.838±0.002	0.834±0.002	0.838±0.002
MESA	**Omni 1.5M**	**[0.3-0.4)**	4.87e-32***	3.52e-34***	5.66e-20***	8.88E-01	3.34e-02*	2.46e-02*	0.864±0.002	0.836±0.002	0.835±0.002	0.834±0.002
MESA	**Omni 1.5M**	**[0.4-0.5)**	2.05e-53***	1.11e-22***	1.92e-38***	3.17e-08***	7.47E-02	4.31e-04**	0.879±0.002	0.847±0.002	0.854±0.002	0.845±0.002
Wellderly	**Affymetrix 6.0**	**[0.001-0.005)**	0.00e+00***	0.00e+00***	0.00e+00***	0.00e+00***	0.00e+00***	5.06e-19***	0.212±0.001	0.183±0.001	0.170±0.001	0.170±0.001
Wellderly	**Affymetrix 6.0**	**[0.005-0.01)**	9.46e-58***	4.82e-140***	2.37e-167***	1.84e-26***	1.44e-39***	1.40e-02*	0.359±0.003	0.307±0.003	0.289±0.003	0.283±0.003
Wellderly	**Affymetrix 6.0**	**[0.01-0.05)**	4.01e-43***	2.50e-149***	1.02e-88***	6.52e-34***	9.91e-12***	1.85e-06***	0.616±0.002	0.566±0.002	0.536±0.002	0.544±0.002
Wellderly	**Affymetrix 6.0**	**[0.05-0.1)**	4.80e-19***	2.25e-55***	7.24e-16***	7.05e-12***	5.76E-01	7.95e-12***	0.820±0.002	0.783±0.003	0.761±0.003	0.769±0.003
Wellderly	**Affymetrix 6.0**	**[0.1-0.2)**	1.08e-46***	4.14e-53***	3.92e-34***	2.40E-01	1.23E-01	1.03e-02*	0.869±0.002	0.830±0.002	0.821±0.002	0.820±0.002
Wellderly	**Affymetrix 6.0**	**[0.2-0.3)**	7.27e-35***	4.70e-38***	1.12e-23***	5.90E-01	9.11E-02	2.47e-02*	0.889±0.002	0.856±0.002	0.850±0.002	0.848±0.002
Wellderly	**Affymetrix 6.0**	**[0.3-0.4)**	1.93e-49***	3.99e-32***	1.43e-35***	7.35e-03*	1.17E-01	3.62E-01	0.888±0.002	0.851±0.002	0.848±0.003	0.844±0.002
Wellderly	**Affymetrix 6.0**	**[0.4-0.5)**	8.77e-51***	7.23e-32***	5.15e-31***	3.34e-03*	1.11e-02*	8.04E-01	0.900±0.002	0.861±0.002	0.859±0.002	0.855±0.002
Wellderly	**UKB Axiom**	**[0.001-0.005)**	0.00e+00***	0.00e+00***	0.00e+00***	1.56e-98***	2.93e-173***	5.08e-13***	0.240±0.001	0.188±0.001	0.179±0.001	0.176±0.001
Wellderly	**UKB Axiom**	**[0.005-0.01)**	1.42e-78***	3.64e-137***	4.44e-134***	8.25e-14***	2.51e-11***	3.79E-01	0.463±0.003	0.392±0.003	0.374±0.003	0.373±0.003
Wellderly	**UKB Axiom**	**[0.01-0.05)**	3.20e-43***	4.20e-124***	1.06e-54***	5.93e-22***	3.75e-02*	1.29e-12***	0.761±0.002	0.714±0.002	0.694±0.002	0.701±0.002

Wellderly	**UKB Axiom**	**[0.05-0.1)**	1.63e-62***	1.70e-80***	1.47e-32***	6.88e-03*	7.27e-05***	1.74e-10***	0.911±0.001	0.879±0.002	0.868±0.002	0.875±0.002
Wellderly	**UKB Axiom**	**[0.1-0.2)**	1.21e-141***	4.36e-96***	4.21e-87***	1.88e-04**	2.59e-05***	5.92E-01	0.930±0.001	0.893±0.001	0.891±0.001	0.889±0.001
Wellderly	**UKB Axiom**	**[0.2-0.3)**	4.84e-219***	7.82e-121***	4.78e-154***	9.07e-15***	9.63e-04**	3.48e-05***	0.944±0.001	0.902±0.001	0.907±0.001	0.899±0.001
Wellderly	**UKB Axiom**	**[0.3-0.4)**	5.32e-257***	3.33e-135***	1.46e-196***	1.38e-21***	4.46e-02*	4.42e-12***	0.940±0.001	0.892±0.002	0.899±0.002	0.886±0.002
Wellderly	**UKB Axiom**	**[0.4-0.5)**	0.00e+00***	5.32e-156***	1.60e-244***	1.62e-34***	1.02e-02*	6.59e-19***	0.949±0.001	0.900±0.001	0.908±0.002	0.895±0.002
Wellderly	**Omni 1.5M**	**[0.001-0.005)**	0.00e+00***	0.00e+00***	0.00e+00***	2.69e-183***	1.06e-207***	7.98e-03*	0.255±0.001	0.215±0.001	0.197±0.001	0.203±0.001
Wellderly	**Omni 1.5M**	**[0.005-0.01)**	1.20e-46***	1.49e-125***	1.09e-102***	8.69e-25***	4.07e-15***	1.55e-02*	0.428±0.003	0.376±0.003	0.347±0.003	0.355±0.003
Wellderly	**Omni 1.5M**	**[0.01-0.05)**	1.67e-13***	1.07e-93***	1.95e-25***	2.72e-36***	1.68e-03*	7.71e-19***	0.706±0.002	0.671±0.002	0.641±0.002	0.653±0.002
Wellderly	**Omni 1.5M**	**[0.05-0.1)**	5.40e-07***	6.60e-59***	8.37e-04**	1.59e-26***	1.12E-01	1.08e-31***	0.883±0.002	0.862±0.002	0.844±0.002	0.854±0.002
Wellderly	**Omni 1.5M**	**[0.1-0.2)**	2.89e-42***	1.73e-55***	1.08e-22***	4.85e-02*	1.70e-03*	1.76e-06***	0.915±0.002	0.889±0.002	0.885±0.002	0.883±0.002
Wellderly	**Omni 1.5M**	**[0.2-0.3)**	2.56e-68***	6.21e-68***	9.93e-47***	8.71E-01	3.06e-02*	6.20E-02	0.933±0.001	0.907±0.002	0.904±0.002	0.901±0.002
Wellderly	**Omni 1.5M**	**[0.3-0.4)**	3.99e-89***	1.54e-58***	3.55e-68***	1.15e-04**	1.52E-01	3.57e-02*	0.927±0.002	0.896±0.002	0.897±0.002	0.892±0.002
Wellderly	**Omni 1.5M**	**[0.4-0.5)**	1.88e-101***	5.27e-60***	2.12e-68***	6.34e-07***	1.44e-02*	2.86e-02*	0.933±0.002	0.902±0.002	0.904±0.002	0.897±0.002
HGDP	**Affymetrix 6.0**	**[0.001-0.005)**	0.00e+00***	0.00e+00***	0.00e+00***	0.00e+00***	0.00e+00***	4.57e-76***	0.115±0.000	0.110±0.001	0.094±0.000	0.097±0.000
HGDP	**Affymetrix 6.0**	**[0.005-0.01)**	9.42e-37***	0.00e+00***	1.41e-180***	1.23e-172***	2.29e-59***	6.90e-31***	0.255±0.001	0.245±0.001	0.212±0.001	0.224±0.001
HGDP	**Affymetrix 6.0**	**[0.01-0.05)**	1.27e-19***	5.35e-251***	5.51e-161***	3.40e-124***	1.46e-68***	3.88e-09***	0.477±0.001	0.461±0.001	0.416±0.001	0.428±0.001
HGDP	**Affymetrix 6.0**	**[0.05-0.1)**	5.84e-07***	1.05e-90***	3.49e-67***	5.86e-46***	3.24e-32***	4.07e-02*	0.720±0.002	0.695±0.002	0.662±0.002	0.660±0.002
HGDP	**Affymetrix 6.0**	**[0.1-0.2)**	4.01e-09***	6.45e-78***	6.97e-76***	1.44e-32***	1.49e-33***	4.55E-01	0.791±0.002	0.765±0.002	0.747±0.002	0.733±0.002
HGDP	**Affymetrix 6.0**	**[0.2-0.3)**	2.52e-13***	1.54e-56***	2.43e-64***	9.52e-15***	2.08e-20***	6.08E-02	0.813±0.002	0.786±0.002	0.774±0.002	0.760±0.002

HGDP	**Affymetrix 6.0**	**[0.3-0.4)**	5.85e-16***	6.47e-64***	7.68e-72***	2.25e-15***	9.49e-21***	7.35E-02	0.835±0.002	0.806±0.002	0.793±0.002	0.782±0.002
HGDP	**Affymetrix 6.0**	**[0.4-0.5)**	2.89e-16***	7.59e-40***	9.69e-63***	7.01e-06***	3.98e-16***	7.45e-05***	0.836±0.002	0.805±0.002	0.797±0.002	0.781±0.002
HGDP	**UKB Axiom**	**[0.001-0.005)**	0.00e+00***	0.00e+00***	0.00e+00***	0.00e+00***	0.00e+00***	1.59e-03*	0.109±0.000	0.096±0.000	0.086±0.000	0.080±0.000
HGDP	**UKB Axiom**	**[0.005-0.01)**	6.42e-154***	0.00e+00***	0.00e+00***	1.60e-75***	4.56e-76***	5.59E-01	0.233±0.001	0.206±0.001	0.193±0.001	0.180±0.001
HGDP	**UKB Axiom**	**[0.01-0.05)**	1.58e-100***	7.46e-277***	0.00e+00***	7.30e-49***	2.07e-107***	1.88e-12***	0.455±0.001	0.421±0.001	0.396±0.001	0.382±0.001
HGDP	**UKB Axiom**	**[0.05-0.1)**	9.25e-18***	8.34e-48***	3.52e-139***	2.05e-08***	9.53e-57***	2.15e-24***	0.746±0.002	0.718±0.002	0.704±0.002	0.678±0.002
HGDP	**UKB Axiom**	**[0.1-0.2)**	1.69e-51***	1.48e-73***	1.93e-253***	3.65e-03*	9.55e-77***	3.82e-55***	0.817±0.001	0.785±0.001	0.779±0.001	0.748±0.002
HGDP	**UKB Axiom**	**[0.2-0.3)**	2.31e-71***	5.99e-63***	4.54e-249***	3.50E-01	2.43e-55***	2.29e-60***	0.844±0.001	0.808±0.002	0.805±0.002	0.773±0.002
HGDP	**UKB Axiom**	**[0.3-0.4)**	1.63e-104***	4.84e-81***	3.67e-302***	9.67e-03*	2.63e-52***	3.67e-70***	0.860±0.001	0.819±0.002	0.818±0.002	0.788±0.002
HGDP	**UKB Axiom**	**[0.4-0.5)**	3.74e-93***	5.24e-50***	8.92e-262***	5.86e-08***	2.88e-45***	3.80e-83***	0.861±0.001	0.821±0.002	0.827±0.002	0.791±0.002
HGDP	**Omni 1.5M**	**[0.001-0.005)**	5.46e-226***	0.00e+00***	0.00e+00***	0.00e+00***	3.78e-302***	1.17e-153***	0.139±0.001	0.135±0.001	0.112±0.001	0.117±0.001
HGDP	**Omni 1.5M**	**[0.005-0.01)**	1.04e-09***	0.00e+00***	2.39e-110***	2.55e-221***	2.71e-56***	1.73e-59***	0.292±0.001	0.290±0.001	0.240±0.001	0.261±0.001
HGDP	**Omni 1.5M**	**[0.01-0.05)**	4.61E-01	7.26e-252***	8.61e-99***	1.15e-221***	3.08e-85***	2.14e-34***	0.513±0.001	0.510±0.001	0.451±0.001	0.474±0.001
HGDP	**Omni 1.5M**	**[0.05-0.1)**	1.58E-01	5.99e-77***	2.18e-41***	5.10e-79***	2.87e-45***	2.85e-05***	0.772±0.002	0.764±0.002	0.730±0.002	0.730±0.002
HGDP	**Omni 1.5M**	**[0.1-0.2)**	1.63e-08***	3.95e-95***	7.52e-111***	1.44e-43***	6.73e-57***	5.16e-03*	0.822±0.001	0.805±0.002	0.787±0.002	0.774±0.002
HGDP	**Omni 1.5M**	**[0.2-0.3)**	2.18e-14***	5.45e-80***	1.70e-107***	2.92e-25***	5.09e-43***	7.27e-05***	0.851±0.002	0.832±0.002	0.819±0.002	0.805±0.002
HGDP	**Omni 1.5M**	**[0.3-0.4)**	1.09e-34***	1.27e-114***	5.18e-157***	7.10e-21***	4.34e-44***	4.80e-07***	0.873±0.001	0.848±0.002	0.836±0.002	0.821±0.002
HGDP	**Omni 1.5M**	**[0.4-0.5)**	2.78e-29***	8.21e-50***	4.50e-131***	1.90e-03*	2.67e-37***	1.11e-23***	0.867±0.002	0.843±0.002	0.837±0.002	0.817±0.002

Validation accuracies were stratified by dataset (MESA, Wellderly, HGDP), genotype array platform (Affymetrix 6.0, UKB Axiom, Omni 1.5M), and MAF bin. We applied Wilcoxon rank-sum tests to compare the HMM-based tools to the reference tuned autoencoder (AE). * represents p-values ≤ 0.05, ** indicates p-values ≤ 0.001, and *** indicates p-values ≤ 0.0001.