RESEARCH ARTICLE

Figures and figure supplements

Rapid, Reference-Free human genotype imputation with denoising autoencoders

Raquel Dias *et al*

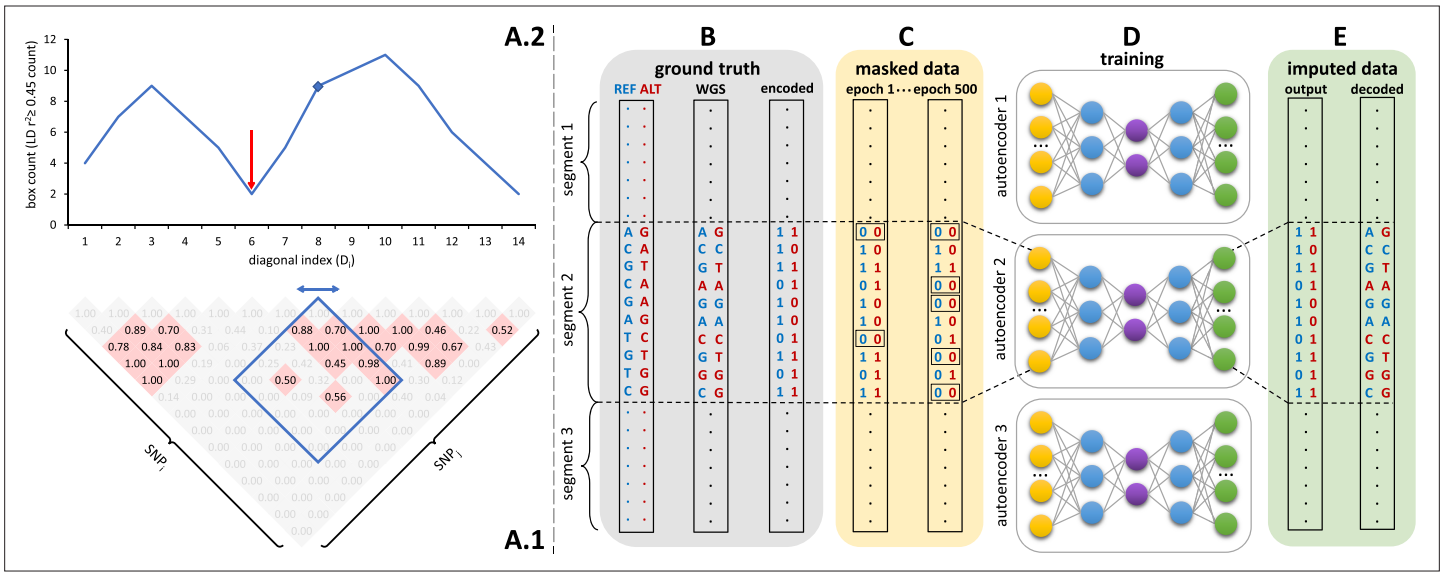

Figure 1. Schematic overview of the autoencoder training workflow. (**A**) Tiling of autoencoders across the genome is achieved by (**A.1**) calculating a *n x n* matrix of pairwise SNP correlations, thresholding them at 0.45 (selected values are shown in red background, excluded values in gray), (**A.2**) 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. For reducing computational complexity, we calculated the correlations in a fixed sliding box size of 500x500 common variants (MAF ≥ 0.5%). Thus, the memory utilization for calculating correlations will be the same regardless of genomic density. (**B**) 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). (**C**) Variant masking (setting both alleles as absent, represented by 0, corrupts data inputs at a gradually increasing masking rate). Example masked variants are outlined. (**D**) Fully-connected autoencoders spanning segments defined as shown in panel (**A**), are then trained to reconstruct the original uncorrupted data from corrupted inputs; (**E**) the reconstructed outputs (imputed data) are compared to the ground truth states for loss calculation and are decoded back to genotypes.

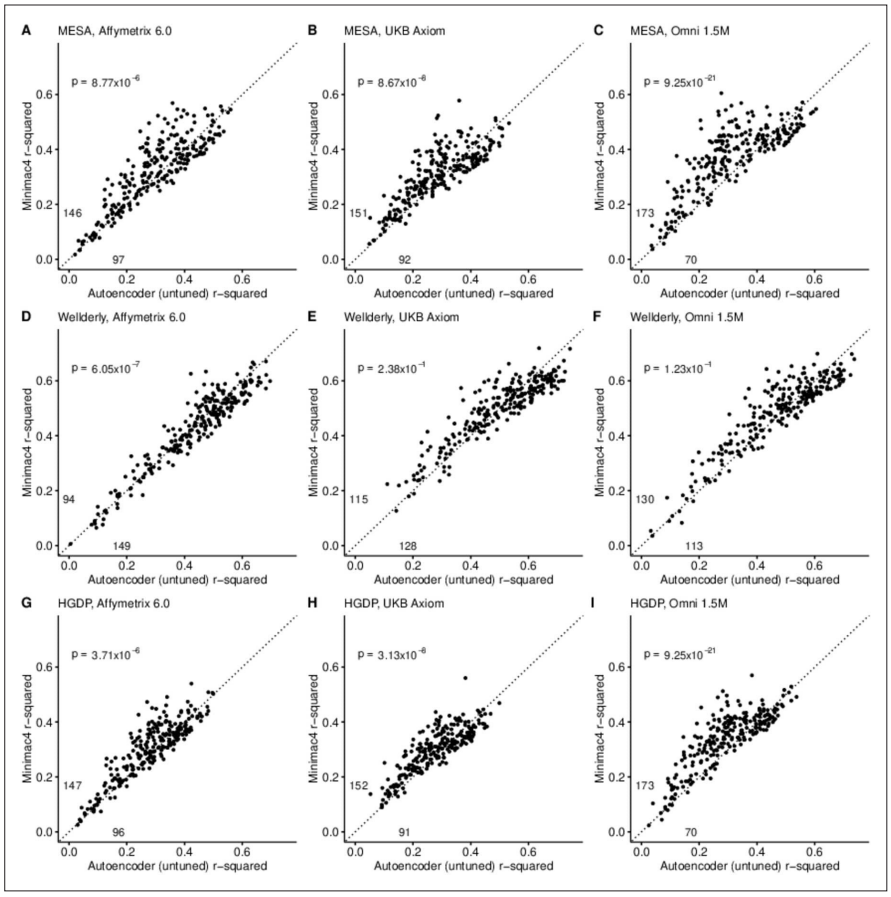

Figure 2. HMM-based (y-axis) versus autoencoder-based (x-axis) imputation accuracy prior to tuning. Minimac4 and untuned autoencoders were tested across three independent datasets–- MESA (top), Wellderly (**middle**), and HGDP (**bottom**) and across three genotyping array platforms–- Affymetrix 6.0 (**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.

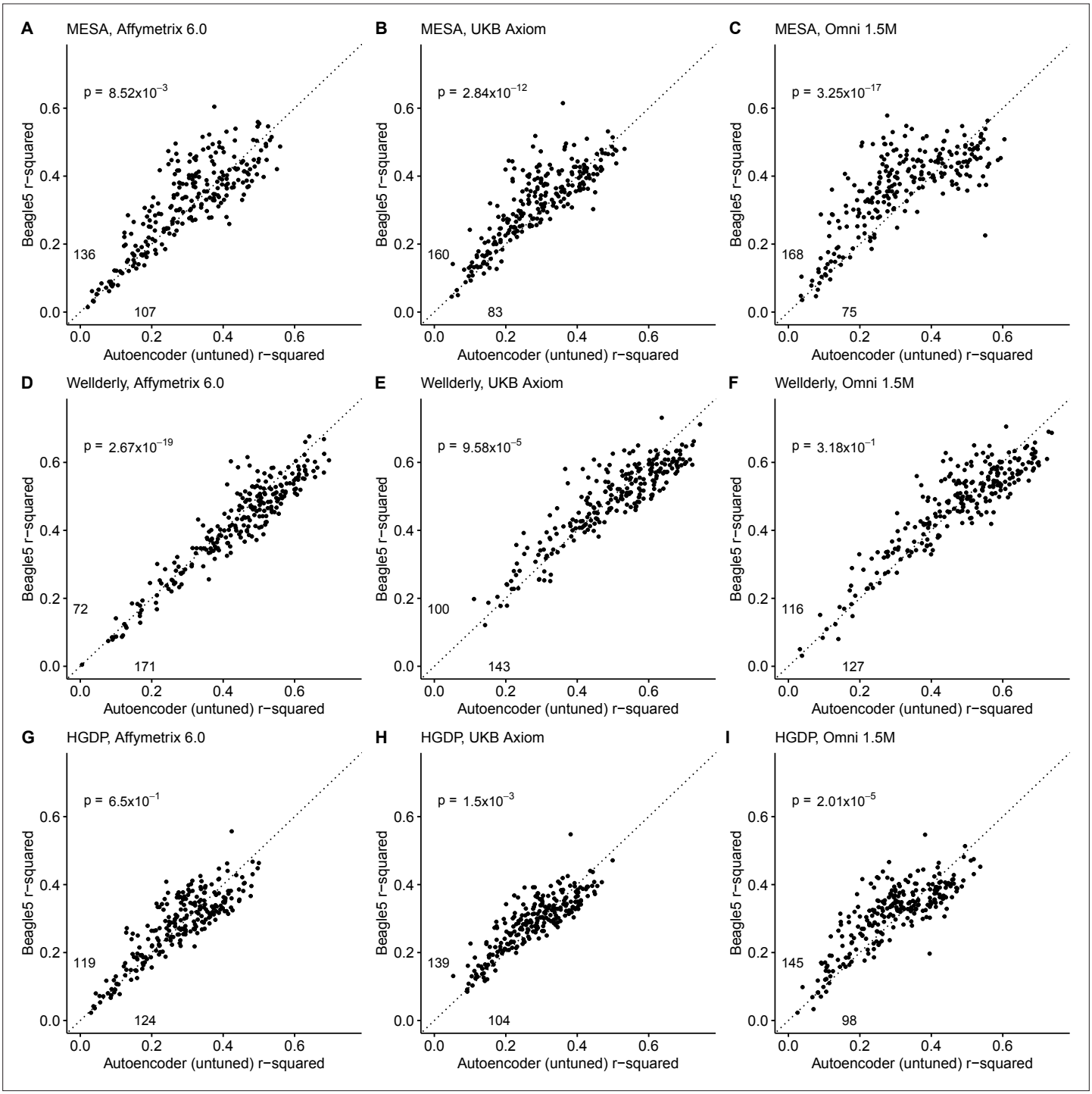

Figure 2—figure supplement 1. Beagle5 (y-axis) versus autoencoder-based (x-axis) imputation accuracy prior to tuning. Beagle5 and untuned autoencoders were tested across three independent datasets - MESA (top), Wellderly (middle), and HGDP (bottom) and across three genotyping array platforms - Affymetrix 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.

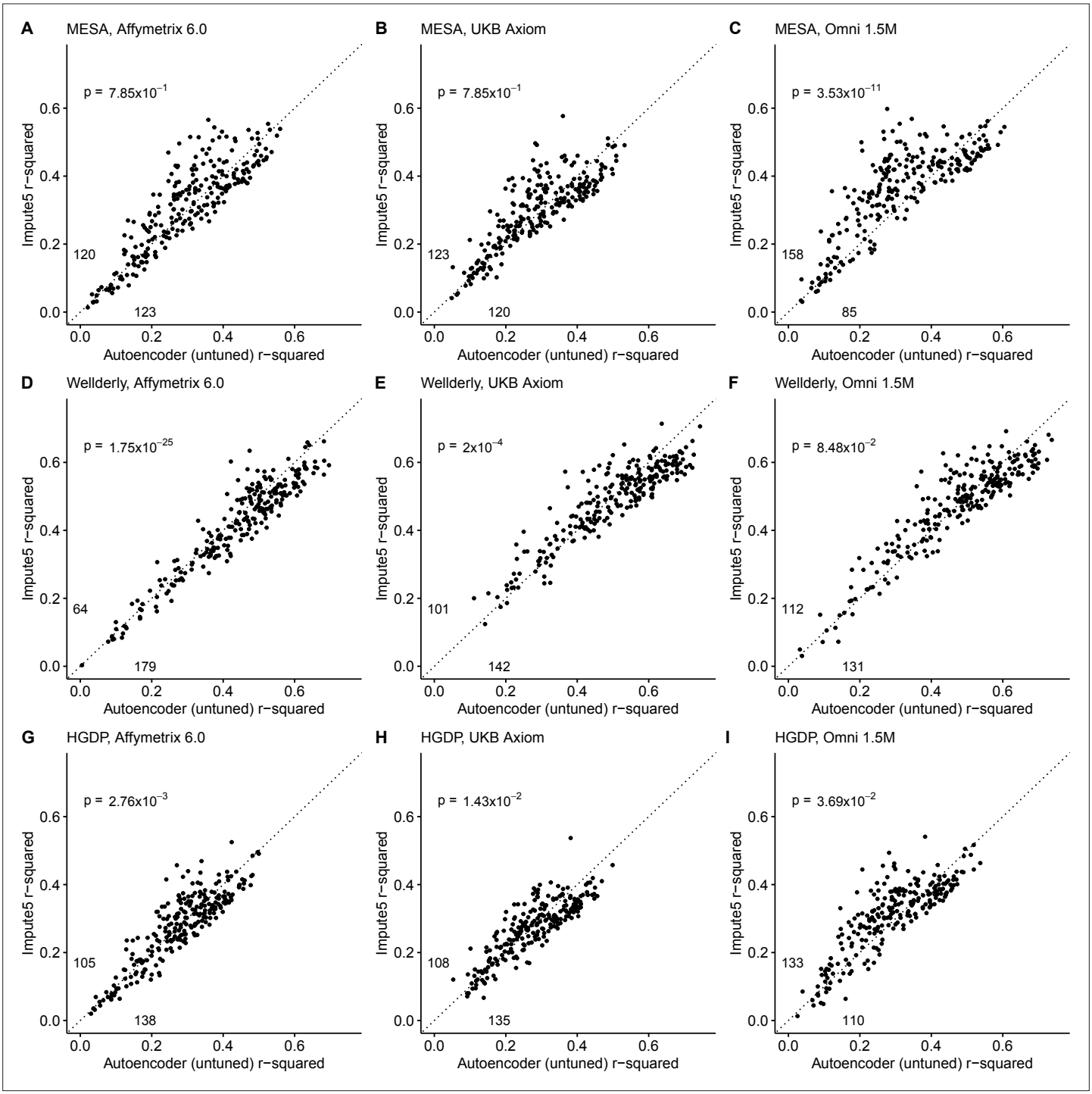

Figure 2—figure supplement 2. Impute5 (y-axis) versus autoencoder-based (x-axis) imputation accuracy prior to tuning. Impute5 and untuned autoencoders were tested across three independent datasets - MESA (top), Wellderly (middle), and HGDP (bottom) and across three genotyping array platforms - Affymetrix 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.

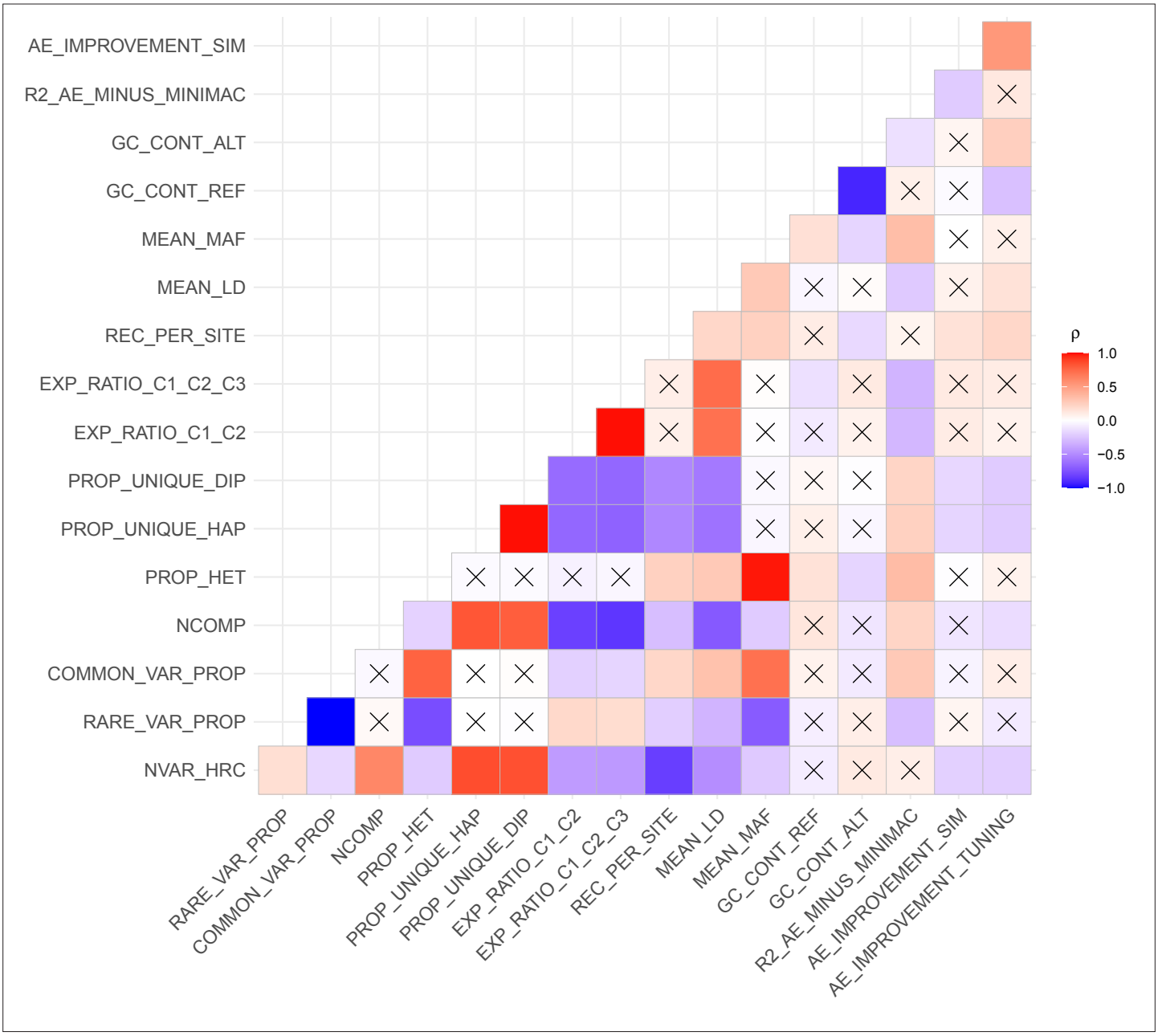

Figure 2—figure supplement 3. Relationship between genomic segment features and autoencoder performance. Spearman correlations (ρ) between genomic segment features and autoencoder performance metrics are presented. An *"X"* denotes Spearman correlations that are not statistically significant (p>0.05). 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).

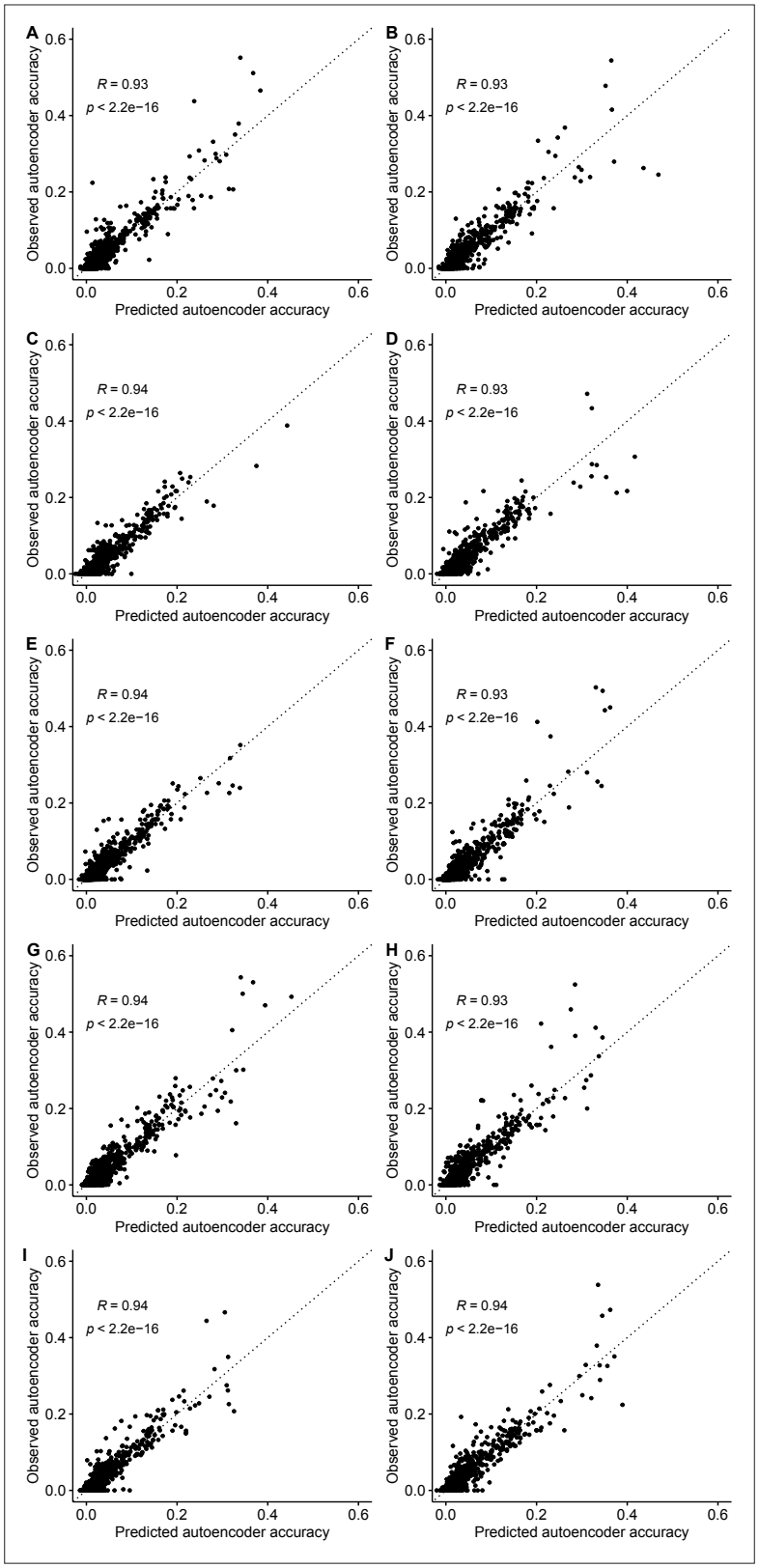

Figure 2—figure supplement 4. Projecting autoencoder performance from hyperparameters and genomic features. We developed an ensemble-based machine learning approach (Extreme Gradient Boosting - XGBoost) to predict the expected performance (r-squared) of each hyperparameter combination per genomic segment using the results of the coarse-grid search and predictive features calculated for each genomic segment (see

Figure 2—figure supplement 4 continued on next page

Figure 2—figure supplement 4 continued

Materials and 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 in the ARIC validation dataset.

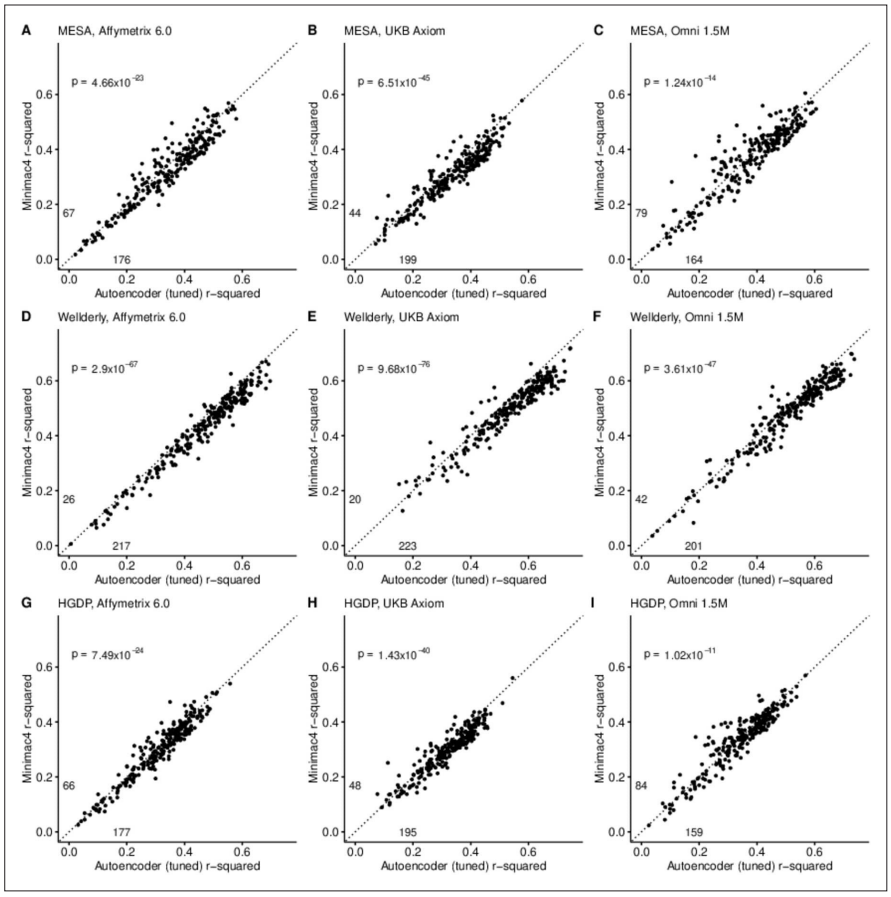

Figure 3. HMM-based (y-axis) versus autoencoder-based (axis) imputation accuracy after tuning. Minimac4 and tuned autoencoders were validated across three independent datasets–- MESA (top), Wellderly (**middle**), and HGDP (**bottom**) and across three genotyping array platforms–- Affymetrix 6.0 (**left**), 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.

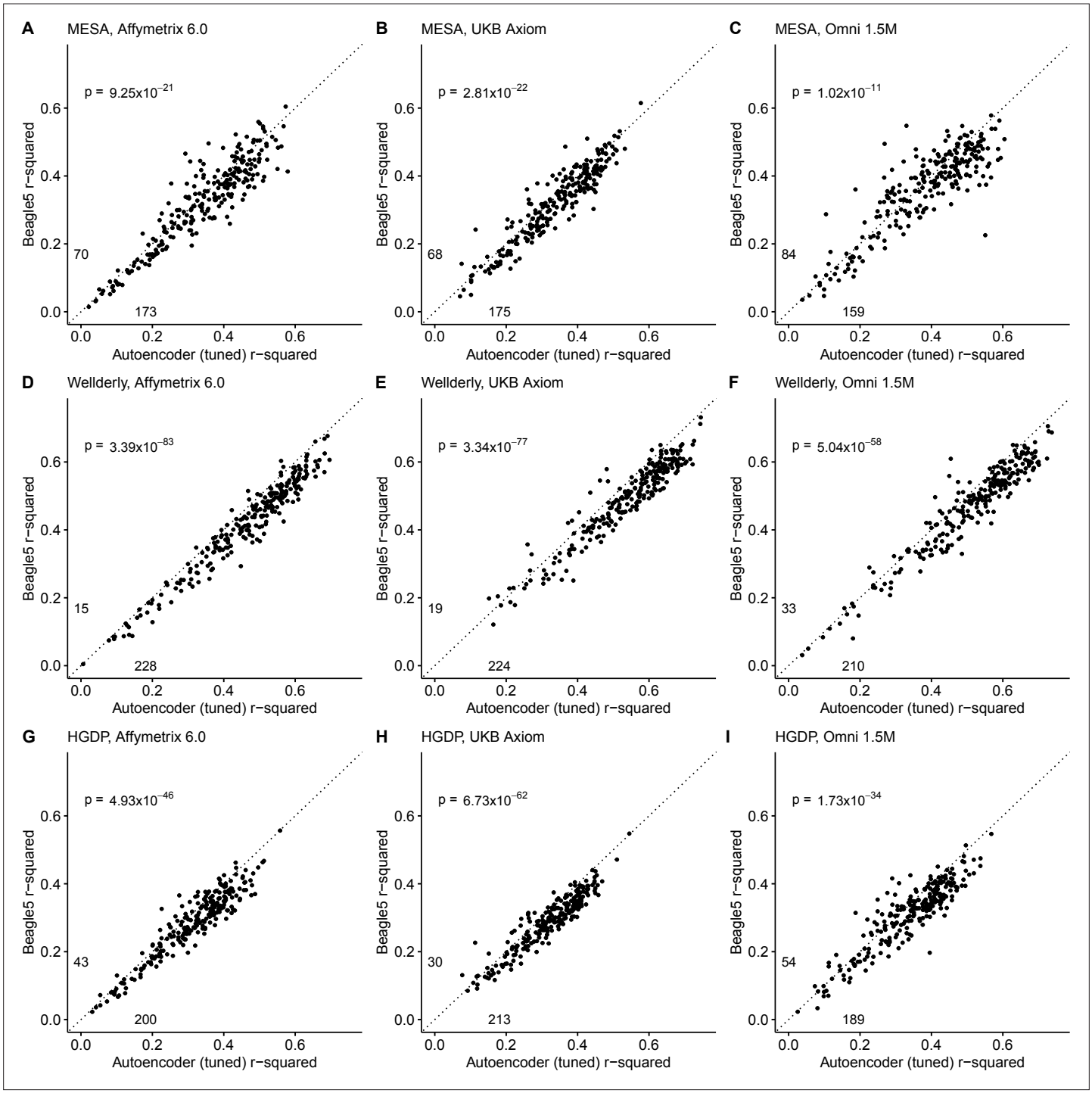

Figure 3—figure supplement 1. Beagle5 (y-axis) versus autoencoder-based (axis) imputation accuracy after tuning. Beagle5 and tuned autoencoders were validated across three independent datasets - MESA (top), Wellderly (**middle**), and HGDP (**bottom**) and across three genotyping array platforms - Affymetrix 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.

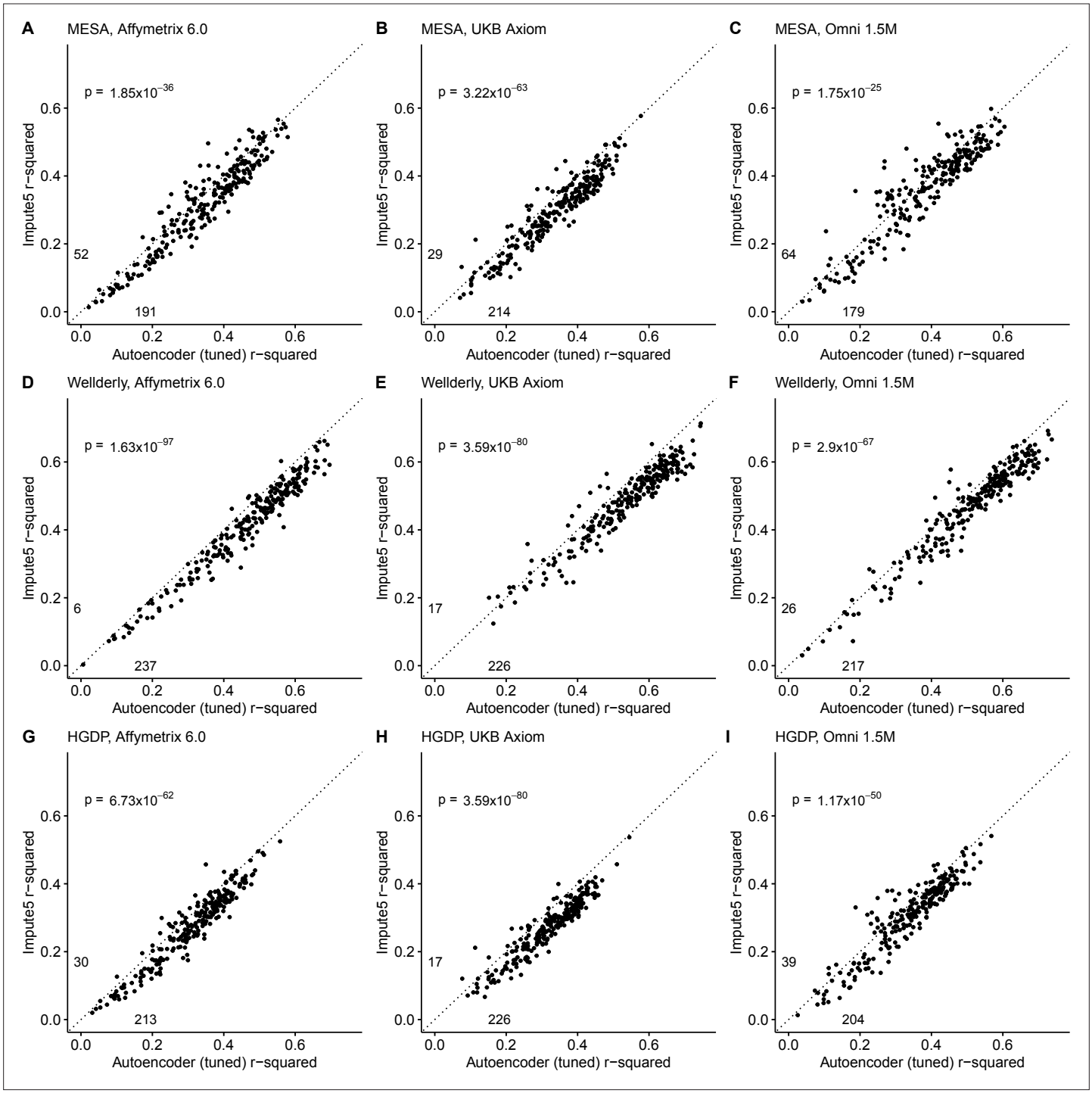

Figure 3—figure supplement 2. Impute5 (y-axis) versus autoencoder-based (axis) imputation accuracy after tuning. Impute5 and tuned autoencoders were validated across three independent datasets - MESA (**top**), Wellderly (**middle**), and HGDP (**bottom**) and across three genotyping array platforms - 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.

Figure 3—figure supplement 3. 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 1.5 M; (**D**) Wellderly Affymetrix 6.0; (**E**) Wellderly UKB Axiom; (**F**) Wellderly Omni 1.5 M; (**G**) HGDP Affymetrix 6.0; (**H**) HGDP UKB Axiom; (**I**) HGDP Omni 1.5 M.

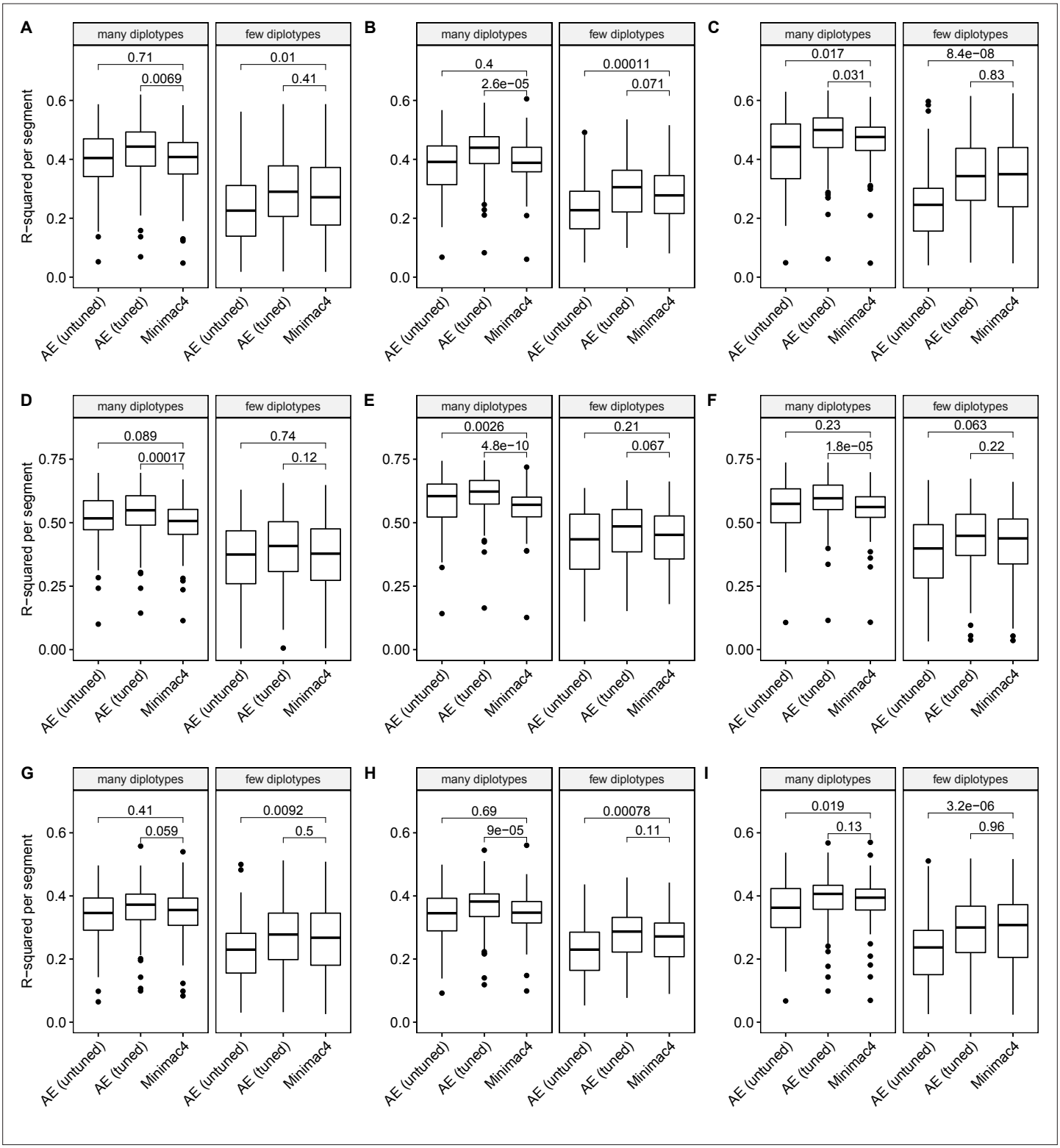

Figure 3—figure supplement 4. Imputation accuracy as a function of unique diplotype abundance. Minimac4 and tuned and untuned autoencoders (AE) were tested across three independent datasets - MESA (top), Wellderly (**middle**), and HGDP (**bottom**) - and across three genotyping array platforms - Affymetrix 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 1.5 M; (**D**) Wellderly Affymetrix 6.0; (**E**) Wellderly UKB Axiom; (**F**) Wellderly Omni 1.5 M; (**G**) HGDP Affymetrix 6.0; (**H**) HGDP UKB Axiom; (**I**) HGDP Omni 1.5 M.

Figure 3—figure supplement 5. 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 1.5 M; (**D**) Wellderly Affymetrix 6.0; (**E**) Wellderly UKB Axiom; (**F**) Wellderly Omni 1.5 M; (**G**) HGDP Affymetrix 6.0; (**H**) HGDP UKB Axiom; (**I**) HGDP Omni 1.5 M.

Figure 3—figure supplement 6. 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 1.5 M; (**D**) Wellderly Affymetrix 6.0; (**E**) Wellderly UKB Axiom; (**F**) Wellderly Omni 1.5 M; (**G**) HGDP Affymetrix 6.0; (**H**) HGDP UKB Axiom; (**I**) HGDP Omni 1.5 M.

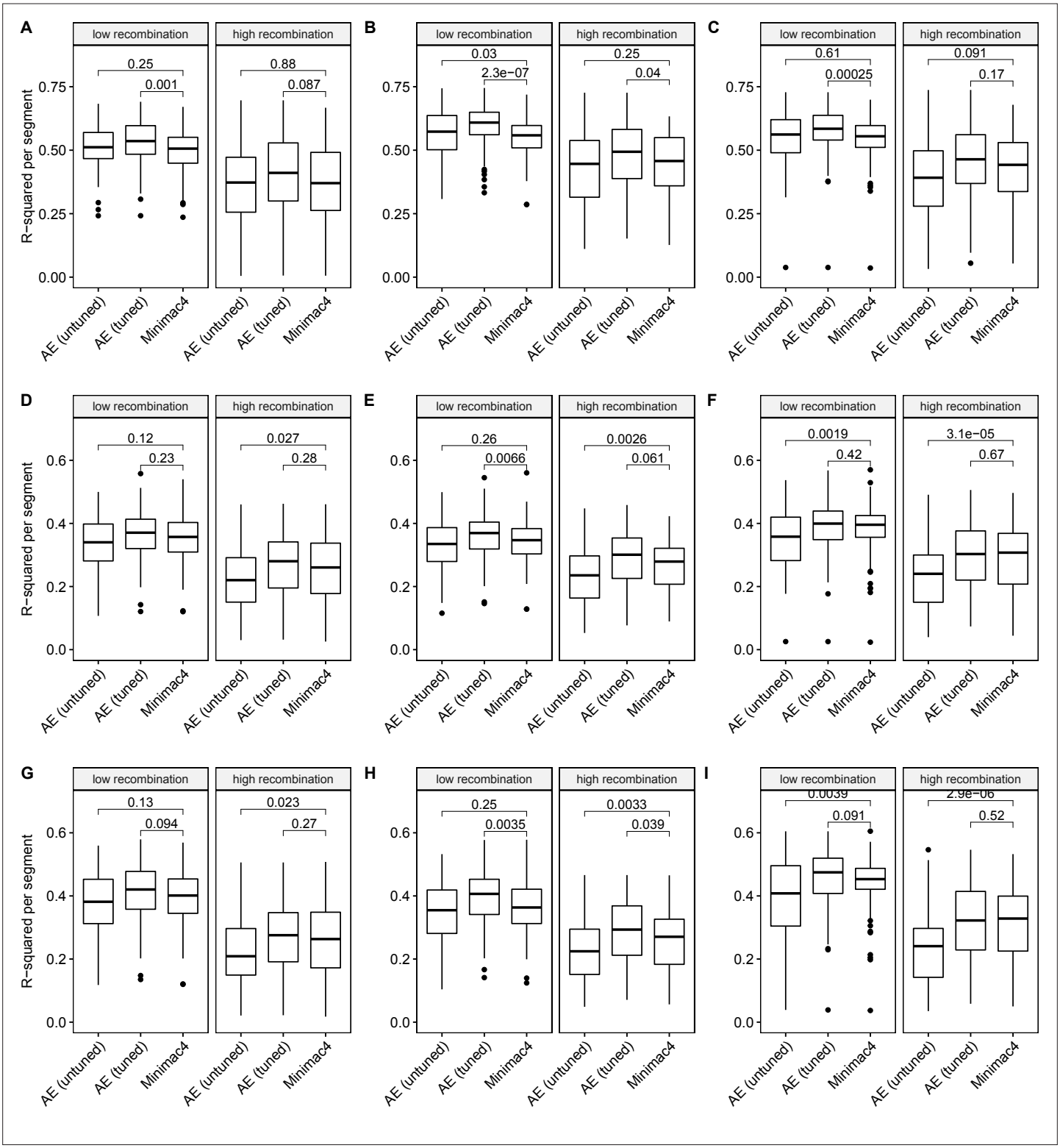

Figure 3—figure supplement 7. Imputation accuracy as a function of recombination rate. Minimac4 and tuned and untuned autoencoders (AE) were tested across three independent datasets - MESA (**top**), Wellderly (**middle**), and HGDP (**bottom**) - and across three genotyping array platforms - Affymetrix 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 1.5 M; (**D**) Wellderly Affymetrix 6.0; (**E**) Wellderly UKB Axiom; (**F**) Wellderly Omni 1.5 M; (**G**) HGDP Affymetrix 6.0; (**H**) HGDP UKB Axiom; (**I**) HGDP Omni 1.5 M.

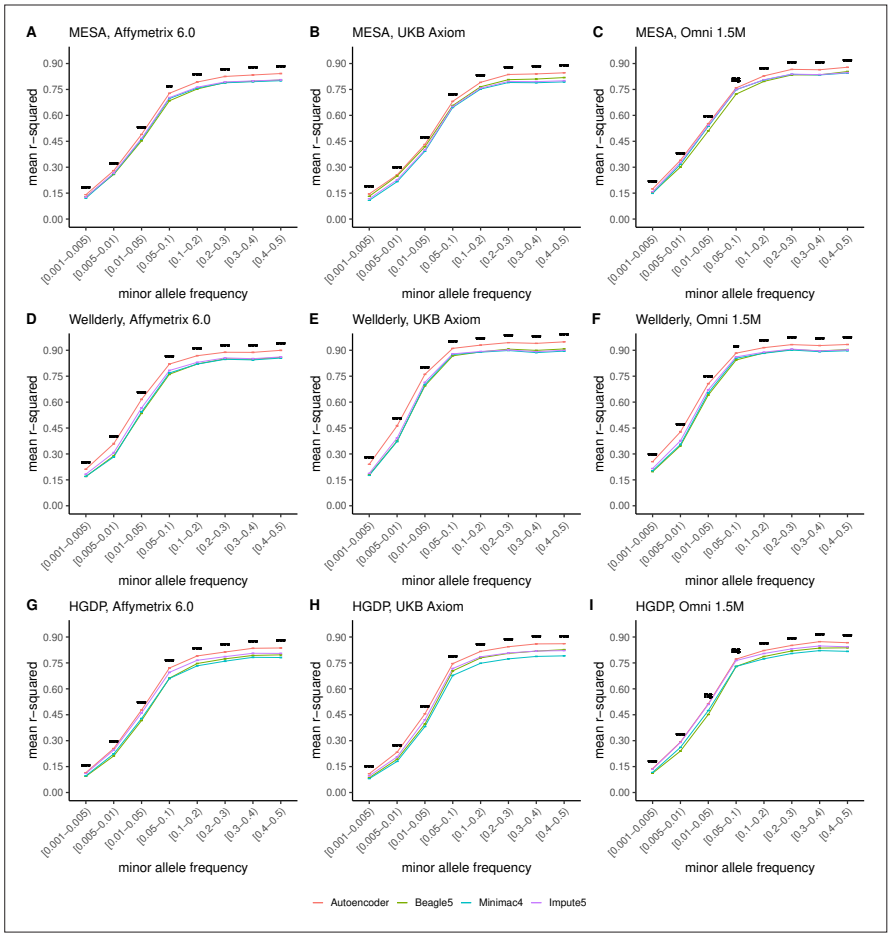

Figure 4. HMM-based versus autoencoder-based imputation accuracy across MAF bins. Autoencoder-based (**red**) and HMM-based (Minimac4 (**blue**), Beagle5 (**green**), and Impute5 (**purple**)) imputation accuracy was validated across three independent datasets–- MESA (**top**), Wellderly (**middle**), 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.

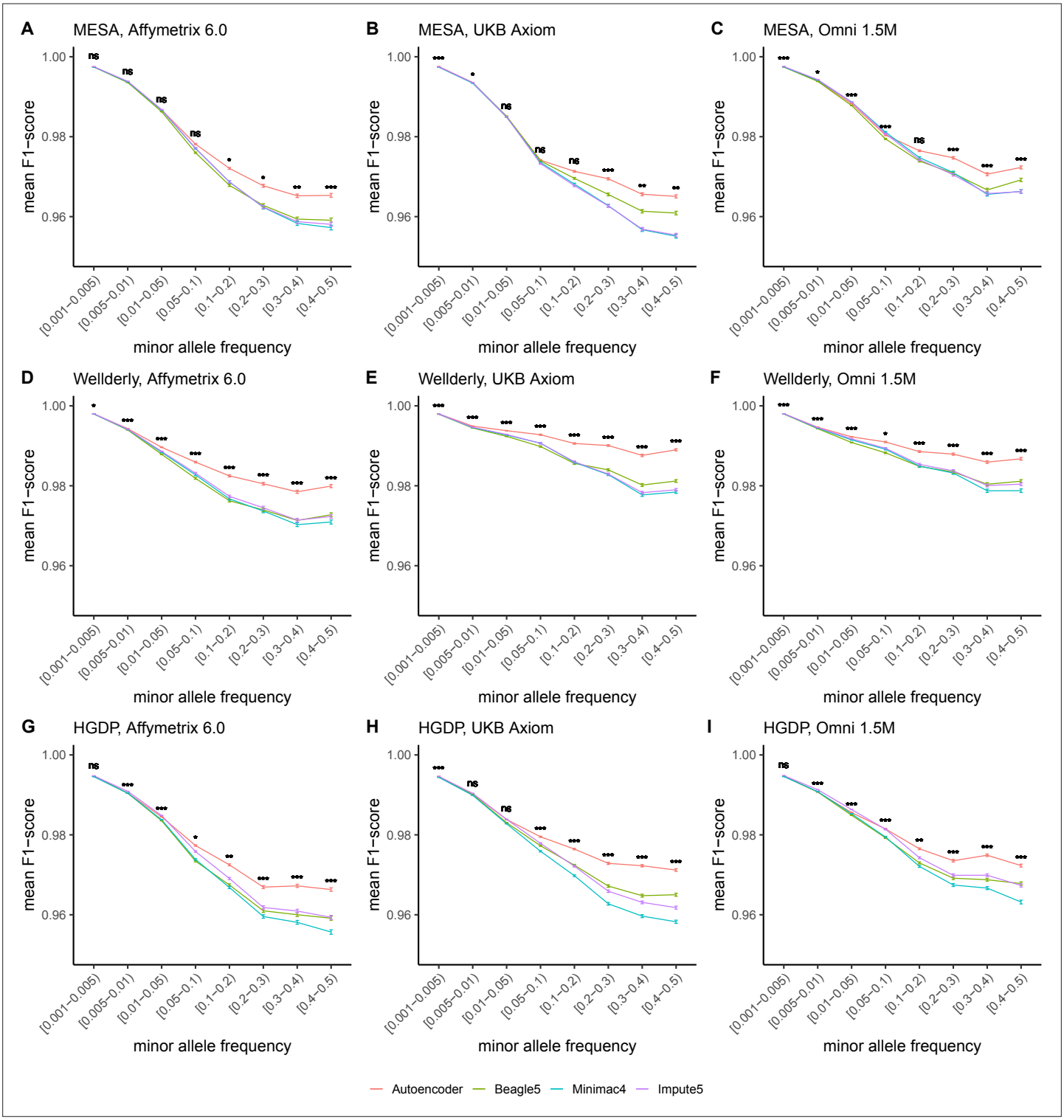

Figure 4—figure supplement 1. HMM-based versus autoencoder-based imputation accuracy across MAF bins (F1 score). Autoencoder-based (**red**) and HMM-based (Minimac4 (**blue**), Beagle5 (**green**), and Impute5 (**purple**)) imputation accuracy was validated across three independent datasets - MESA (**top**), 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. Please note that F1 scores are high for rare variations given the high degree of class imbalance, most alternative alleles are not present for rare variants, leading to high accuracy in the negative class. R-squared depicted in ***Figure 4*** provides a more accurate picture of balanced class accuracy.

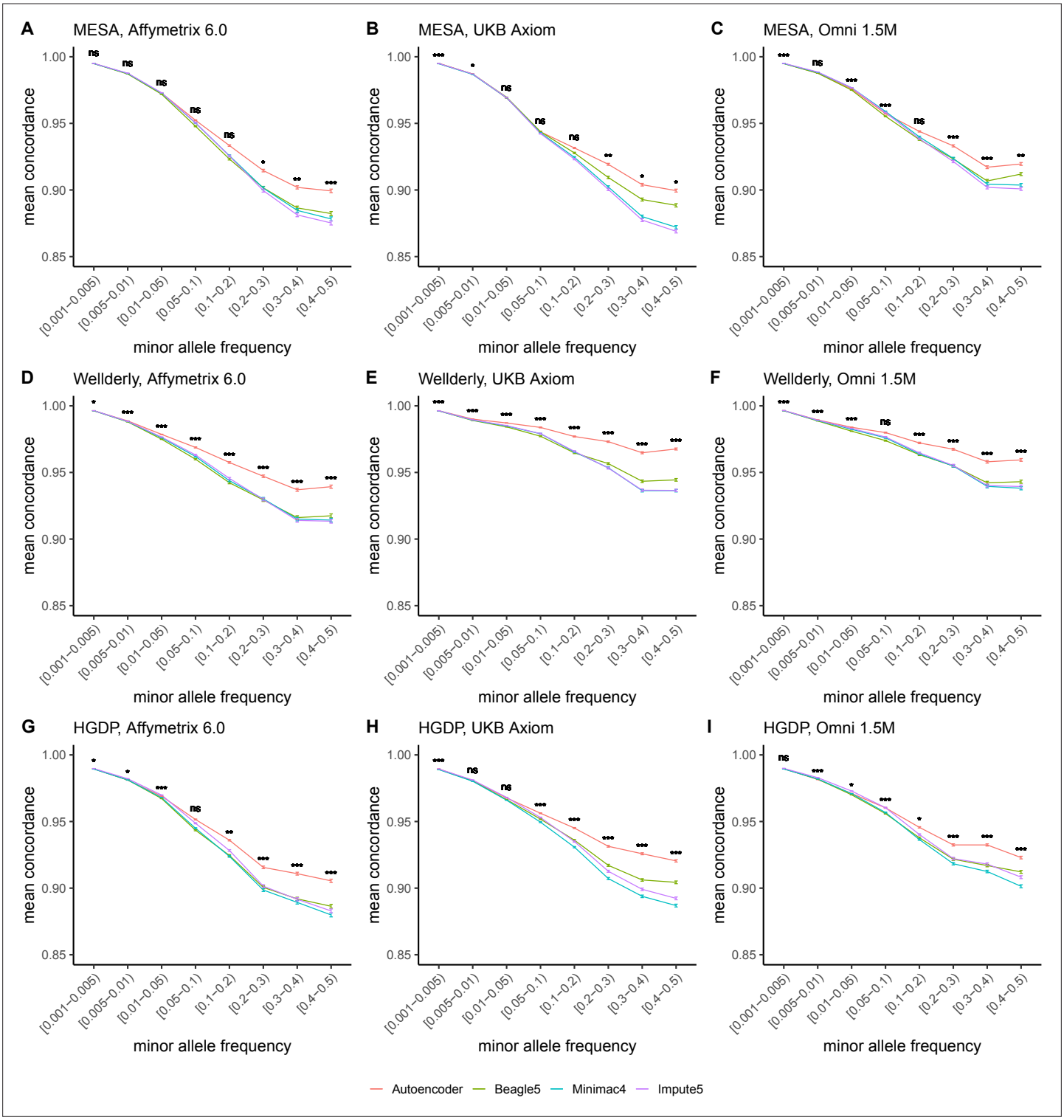

Figure 4—figure supplement 2. HMM-based versus autoencoder-based imputation accuracy across MAF bins (concordance). Autoencoder-based (**red**) and HMM-based (Minimac4 (**blue**), Beagle5 (**green**), and Impute5 (**purple**)) imputation accuracy was validated across three independent datasets - MESA (**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. Please note that F1 scores are high for rare variations given the high degree of class imbalance, most alternative alleles are not present for rare variants, leading to high accuracy in the negative class. R-squared depicted in ***Figure 4*** provides a more accurate picture of balanced class accuracy.

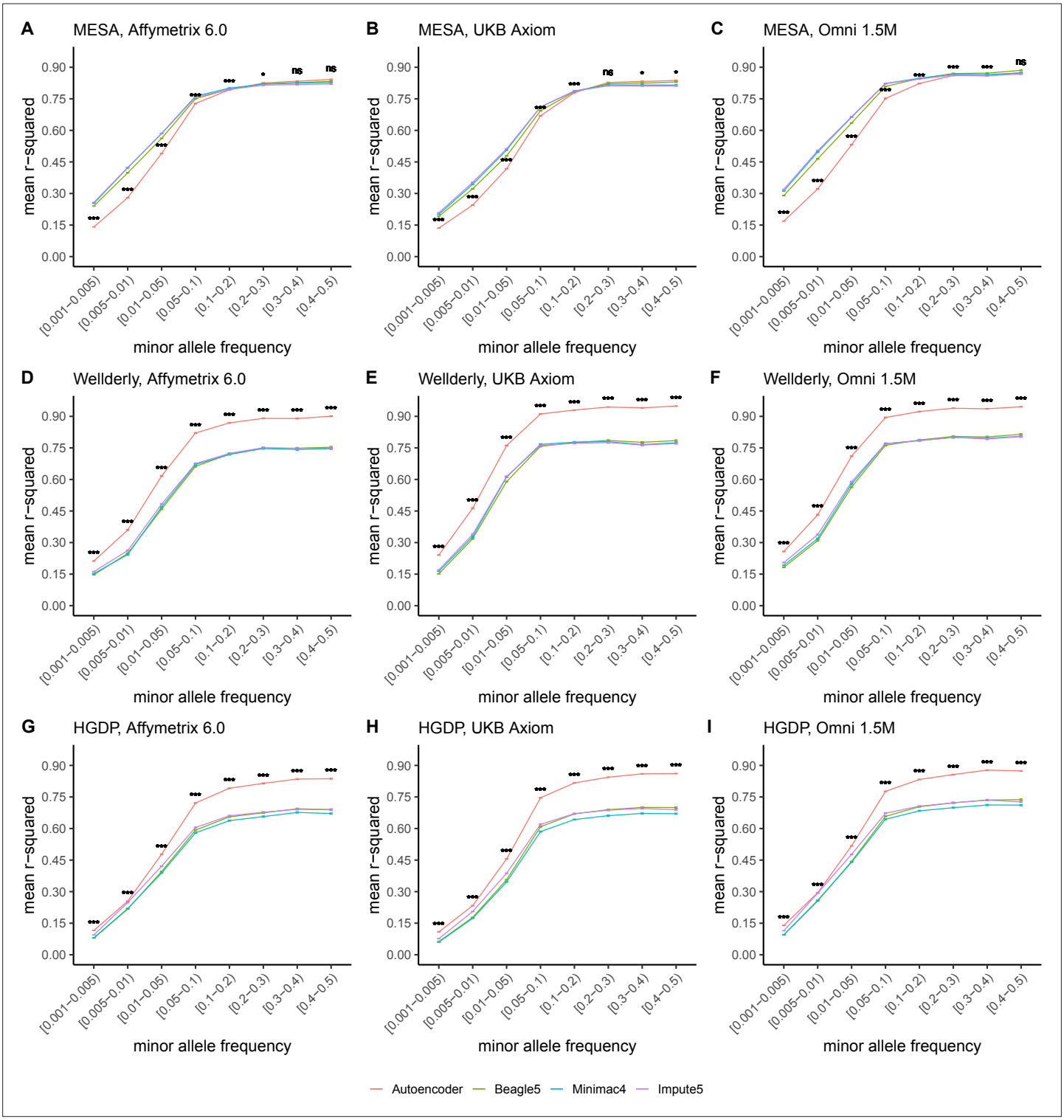

Figure 4—figure supplement 3. TOPMed cohort HMM-based imputation versus HRC cohort autoencoder-based imputation accuracy across MAF bins. Autoencoder-based imputation using the HRC reference panel (**red**) was compared to HMM-based (Minimac4 (**blue**), Beagle5 (**green**), and Impute5 (**purple**)) imputation accuracy using the upgraded TOPMed cohort. Accuracy was determined across three datasets–- MESA (**top** – not independent), Wellderly (**middle** - independent), and HGDP (**bottom** - independent) 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.

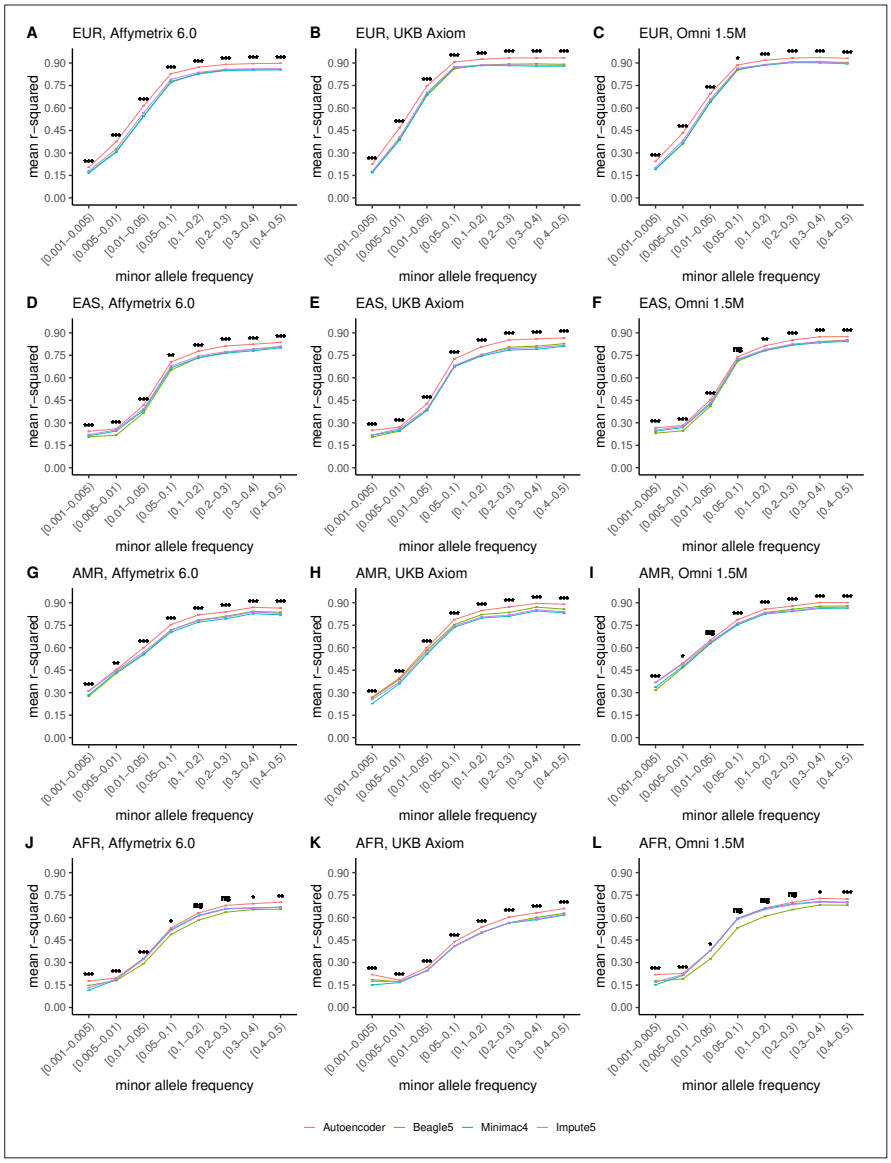

Figure 5. HMM-based versus autoencoder-based imputation accuracy across ancestry groups. Autoencoder-based (**red**) and HMM-based (Minimac4 (**blue**), Beagle5 (**green**), and Impute5 (**purple**)) imputation accuracy was validated across individuals of diverse ancestry from MESA cohort (EUR: European (**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.

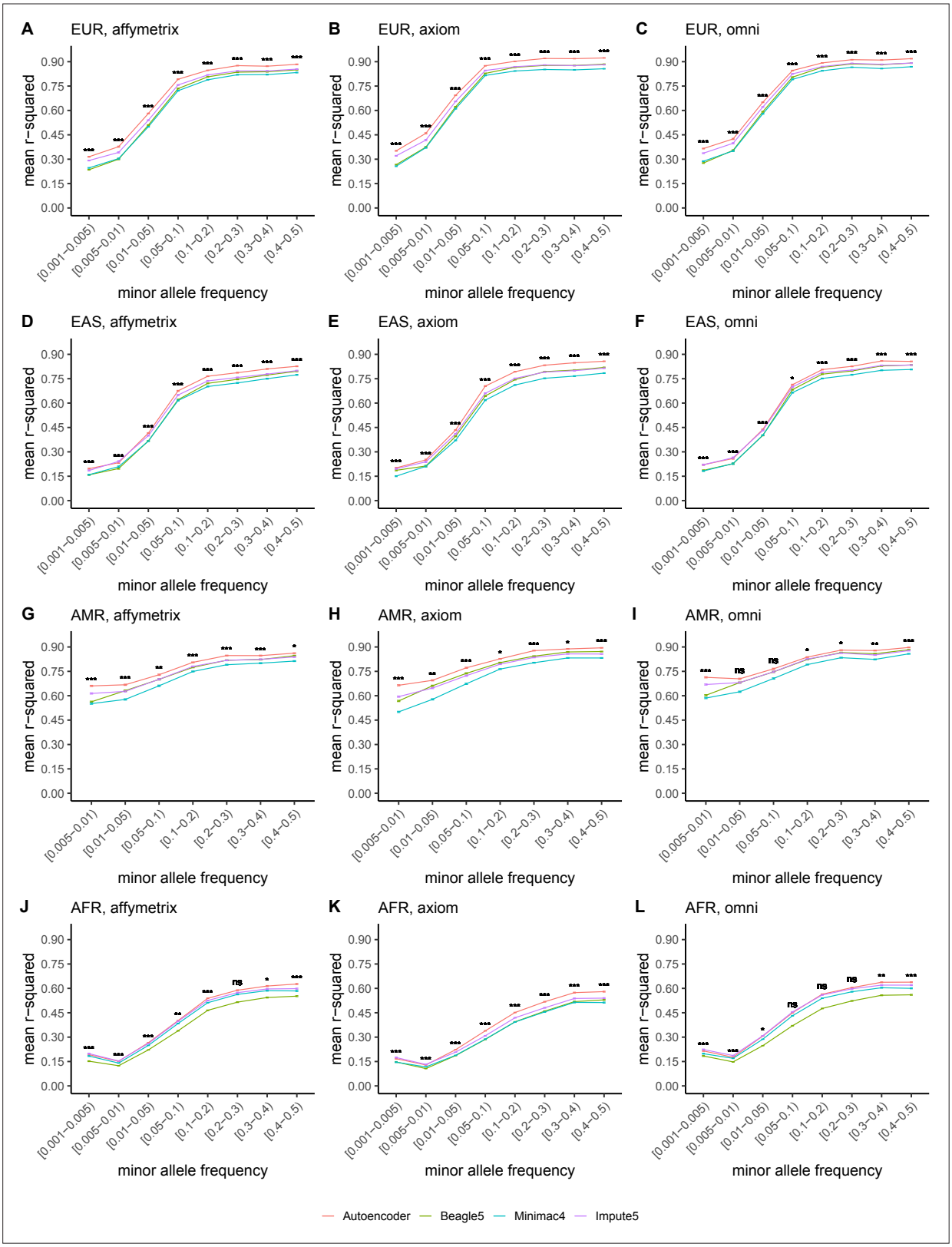

Figure 5—figure supplement 1. HMM-based versus autoencoder-based imputation accuracy across ancestry groups. Autoencoder-based (**red**) and HMM-based (Minimac4 (**blue**), Beagle5 (**green**), and Impute5 (**purple**)) imputation accuracy was validated across individuals of diverse ancestry from HGDP 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

Figure 5—figure supplement 1 continued on next page

Figure 5—figure supplement 1 continued

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.

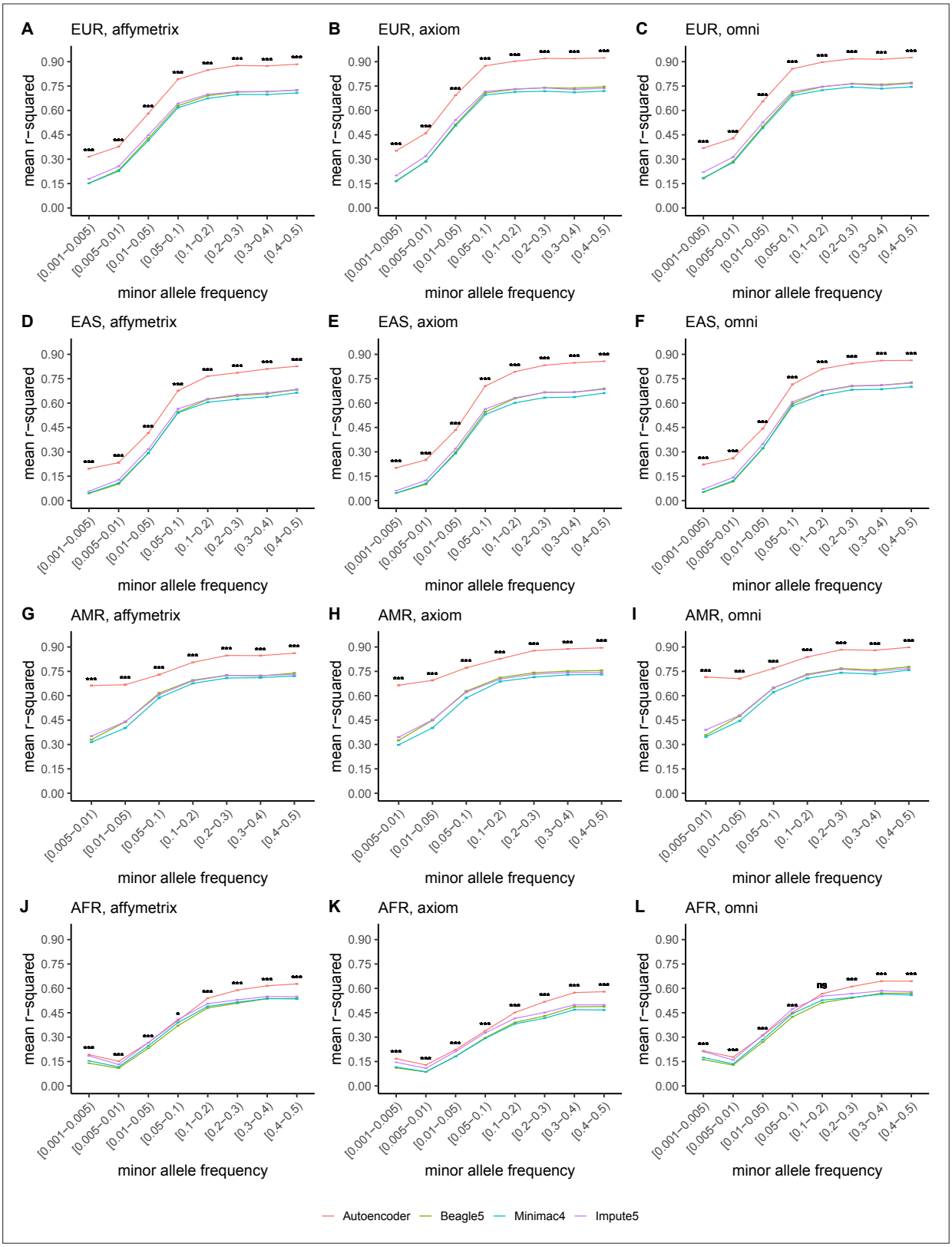

Figure 5—figure supplement 2. TOPMed cohort HMM-based versus HRC cohort autoencoder-based imputation accuracy across ancestry groups. Autoencoder-based imputation using the HRC reference panel (**red**) was compared to HMM-based (Minimac4 (**blue**), Beagle5 (**green**), and Impute5 (**purple**)) imputation using the TOPMed reference panel. Accuracy was determined across individuals of diverse ancestry from the 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

Figure 5—figure supplement 2 continued on next page

Figure 5—figure supplement 2 continued

(**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.

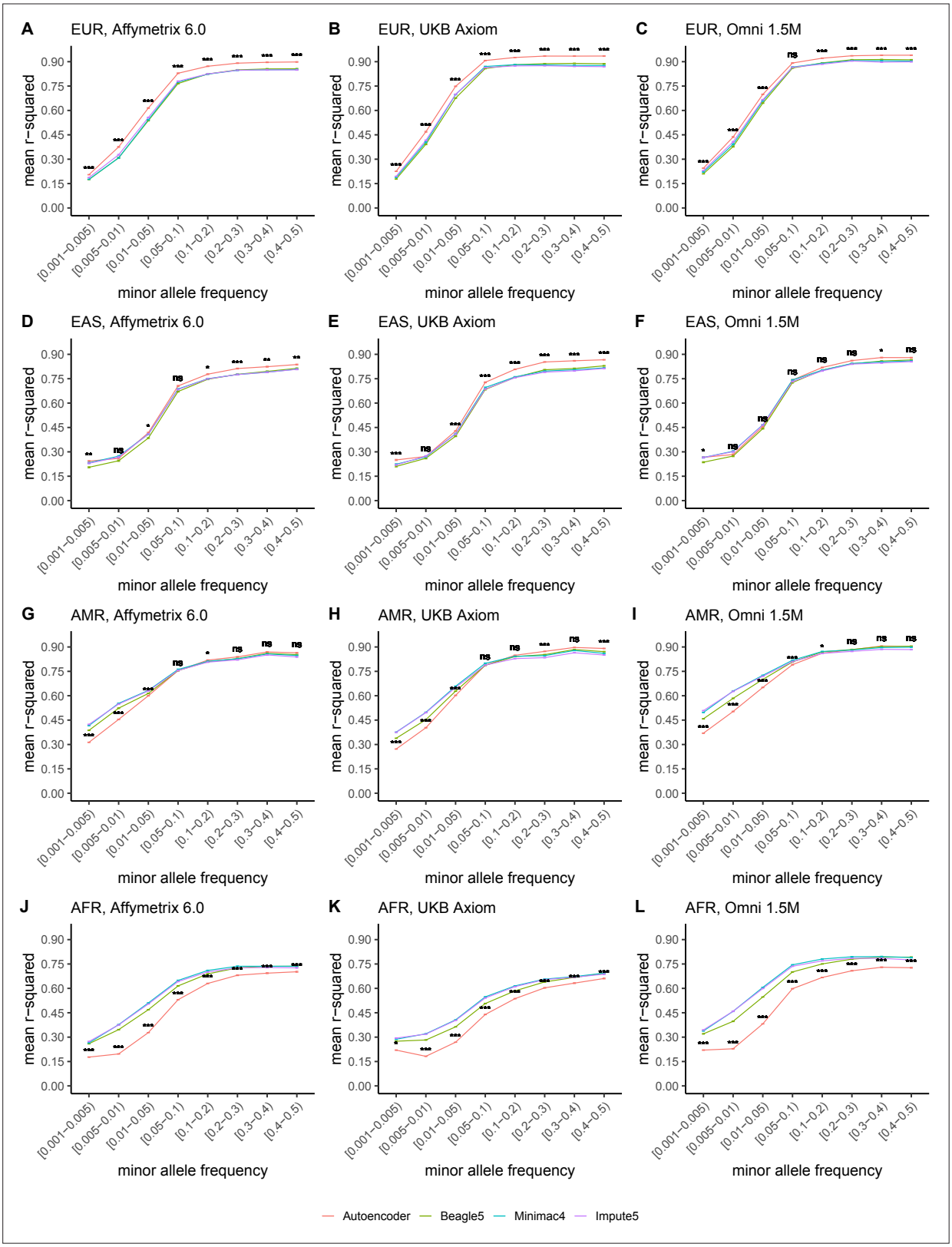

Figure 5—figure supplement 3. TOPMed cohort HMM-based versus HRC cohort autoencoder-based imputation accuracy across ancestry groups. Autoencoder-based imputation using the HRC reference panel (**red**) was compared to HMM-based (Minimac4 (**blue**), Beagle5 (**green**), and Impute5 (**purple**)) imputation using the TOPMed reference panel. Accuracy was determined across individuals of diverse ancestry from the 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

Figure 5—figure supplement 3 continued on next page

Figure 5—figure supplement 3 continued

(**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.

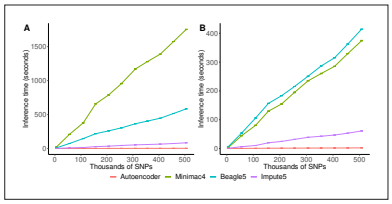

Figure 6. HMM-based versus autoencoder-based inference runtimes. We plot the average time and standard error of three imputation replicates. Two hardware configurations were used for the tests: (**A**) a low-end environment: 16-core Intel Xeon CPU (E5-2640 v2 2.00 GHz), 250 GB RAM, and one GPU (NVIDIA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