Supplementary Figures

Cross-species and tissue imputation of species-level DNA methylation samples across mammalian species.

Emily Maciejewski, Steve Horvath, Jason Ernst

Contents

Supplementary Figures

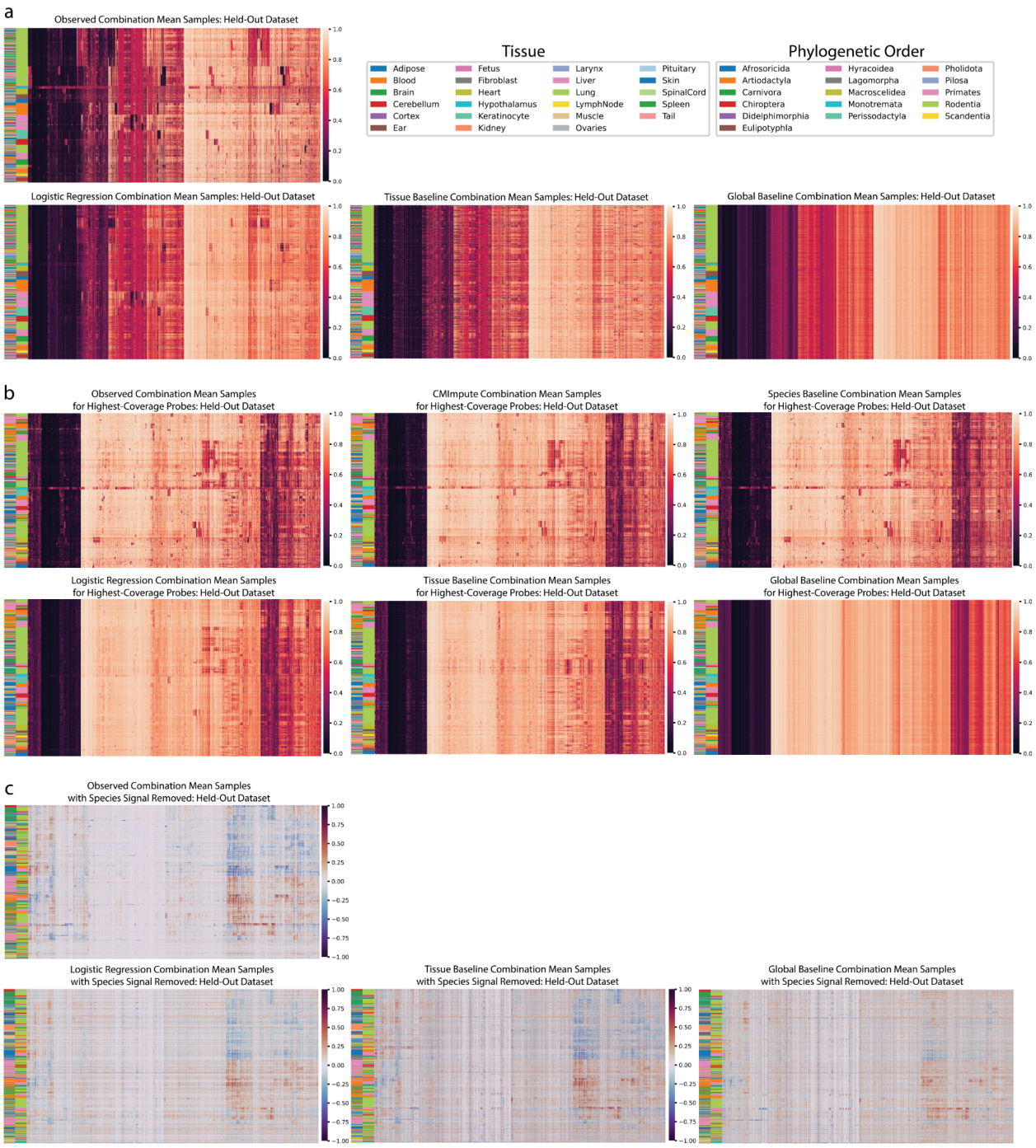
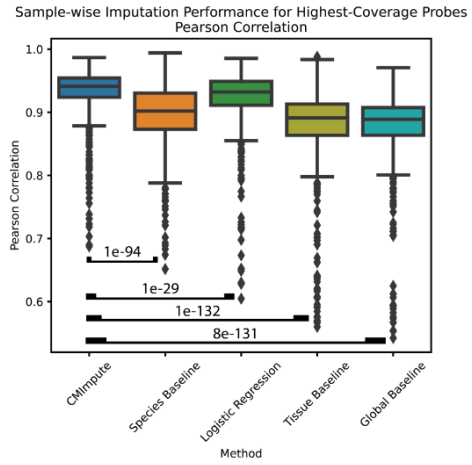


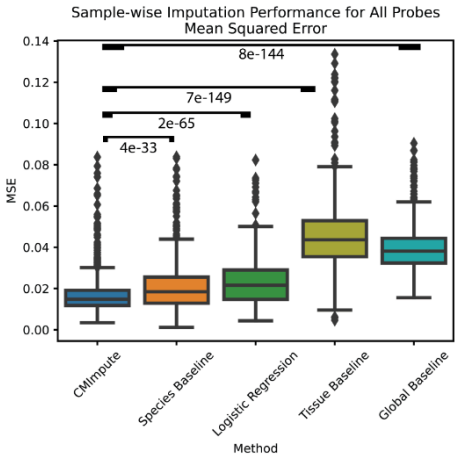
Figure S1. Imputed species-tissue combination mean sample visualizations. a) Similar heatmaps as in Fig. 2a-c of predictions of held-out datasets' methylation probe values shown here for logistic regression, tissue baseline, and global baseline (bottom row; left to right). Observed methylation probe values included again for comparison (top row). Each row within a

heatmap is a species-tissue combination mean sample and each column is a methylation probe. Samples and probes were ordered based on hierarchical clustering followed by optimal leaf ordering. Color bars on the left indicate the phylogenetic order (inner) and tissue (outer) corresponding to the samples. Legends corresponding to the color bars can be found in the top right of the figure. Color scale representing methylation values from 0 to 1 on the right. **b)** Heatmaps of methylation probe values for the observed data held-out during cross-validation and CMImpute and baseline predictions restricted to the highest coverage probes. Samples and probes are ordered and labeled similarly to a. **c)** Similar heatmaps as in Fig. 2d-f of predicted datasets with the species signal removed shown here to highlight the differentially methylated tissue regions for logistic regression, tissue baseline, and global baseline. Observed dataset with species signal removed included again for comparison. Color scale representing methylation delta values from -1 to 1 on the right.

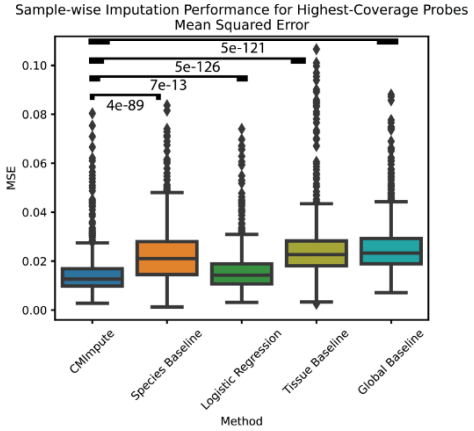
a


Sample-wise Imputation Performance for Highest-Coverage Probes
Pearson Correlation
Pearson Correlation
1e-94
1e-29
1e-132
8e-131
CMImpute
Species Baseline
Logistic Regression
Tissue Baseline
Global Baseline
Method

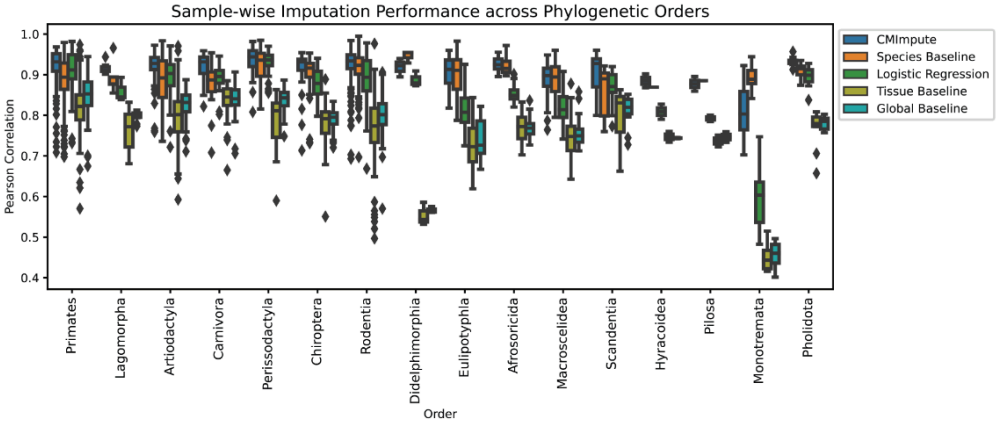
b


Sample-wise Imputation Performance for All Probes
Mean Squared Error
MSE
8e-144
7e-149
2e-65
4e-33
CMImpute
Species Baseline
Logistic Regression
Tissue Baseline
Global Baseline
Method

c


Sample-wise Imputation Performance for Highest-Coverage Probes
Mean Squared Error
MSE
5e-121
5e-126
7e-13
4e-89
CMImpute
Species Baseline
Logistic Regression
Tissue Baseline
Global Baseline
Method

d


Sample-wise Imputation Performance across Phylogenetic Orders
Pearson Correlation
CMImpute
Species Baseline
Logistic Regression
Tissue Baseline
Global Baseline
Primates
Lagomorpha
Artiodactyla
Carnivora
Perissodactyla
Chiroptera
Rodentia
Didelphimorphia
Eulipotyphla
Afrosoricida
Macroscelidea
Scandentia
Hyracoidea
Pilosa
Monotremata
Pholidota
Order

e

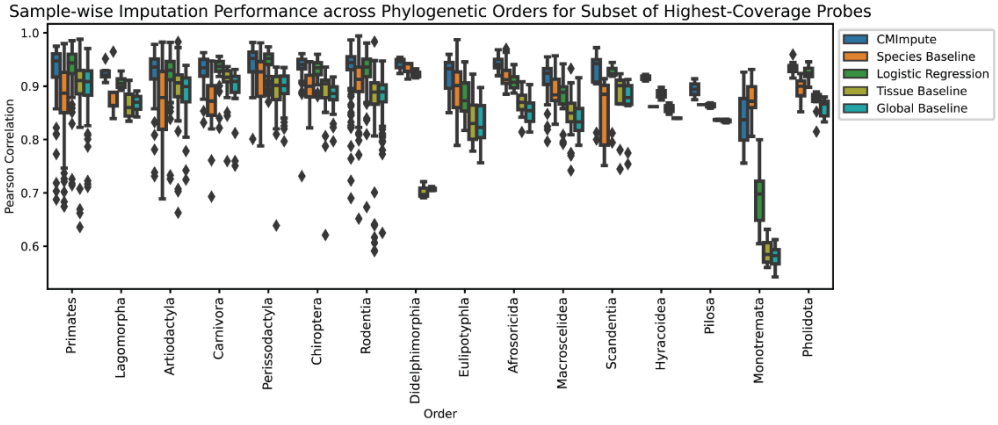

Sample-wise Imputation Performance across Phylogenetic Orders for Subset of Highest-Coverage Probes
Pearson Correlation
CMImpute
Species Baseline
Logistic Regression
Tissue Baseline
Global Baseline
Primates
Lagomorpha
Artiodactyla
Carnivora
Perissodactyla
Chiroptera
Rodentia
Didelphimorphia
Eulipotyphla
Afrosoricida
Macroscelidea
Scandentia
Hyracoidea
Pilosa
Monotremata
Pholidota
Order

Figure S2. Sample-wise performance distributions. a) Box-plot showing distribution of sample-wise Pearson correlation of imputed species-tissue combination mean samples with held-out observed values when restricted to highest-coverage methylation probes for CMImpute and all baselines. Baselines labeled by Wilcoxon signed-rank test p-value comparing CMImpute's sample-wise Pearson correlation and each baseline's sample-wise Pearson correlation for each imputed combination ([CMImpute, Species Baseline], [CMImpute, Logistic Regression], [CMImpute, Tissue Baseline], [CMImpute, Global Baseline]). **b-c)** Sample-wise MSE of imputed combination mean samples based on **b)** all probes and **c)** highest-coverage probes only. **d-e)** Sample-wise Pearson correlation of imputed combination mean samples based on phylogenetic order of the imputed species for **d)** all probes and **e)** the subset of highest-coverage probes.

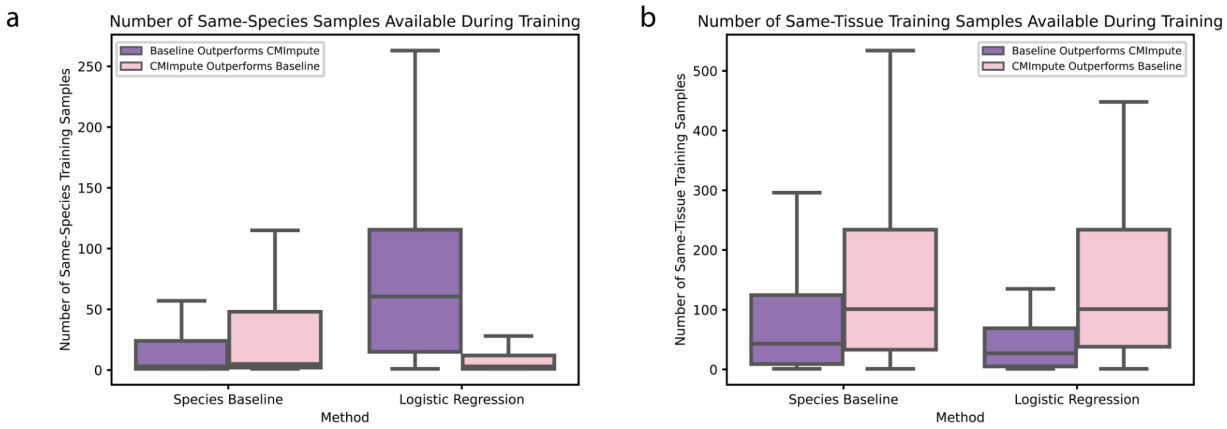


Figure S3. Analysis of subset of samples where species baseline and logistic regression outperform CMImpute and vice versa. a-b) The number of **a)** individual same-species samples and **b)** individual same-tissue samples available during model training for imputed combination mean samples where CMImpute outperforms the species baseline and logistic regression and vice versa. The imputed combination mean samples where the baseline outperforms CMImpute are the 32% of samples (291 out of 907 combination mean samples) below the black line in the top left plot of Fig. 3b for the species baseline and the 22% (200 out of 907 combination mean samples) of samples below the black line in the top right plot of Fig. 3b for logistic regression.

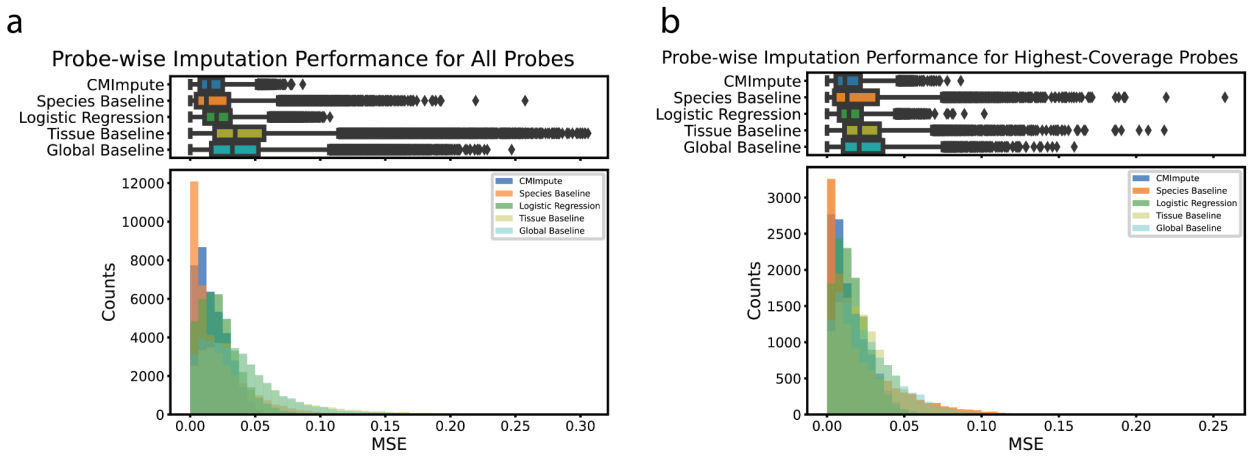


Figure S4. Probe-wise mean squared error distributions. a) Distributions of probe-wise MSE with held-out observed values based on all probes for CMImpute and all baselines. Plots are formatted in the way as Fig. 4a-b. The top boxplots show the distribution of probe-wise correlations with held-out observed values. The bottom boxplots show the number of imputed combination mean samples across 50 performance bins. Legend for both the boxplot and histogram shown in histogram plot. CMImpute yields the best mean probe-wise MSE of 0.017 compared to the species baseline's 0.021, logistic regression's 0.023, the tissue baseline's 0.045, and the global baseline's 0.039. Corresponding plots for subsets of higher variance probes can be found in Figure S10d-f. **b)** Same as a) except restricted to highest-coverage probes. CMImpute yields the best mean probe-wise MSE of 0.014 compared to the species baseline's 0.023, logistic regression's 0.016, the tissue baseline's 0.025, and the global baseline's 0.026. Corresponding plots for subsets of higher variance probes can be found in Figure S10a-c.

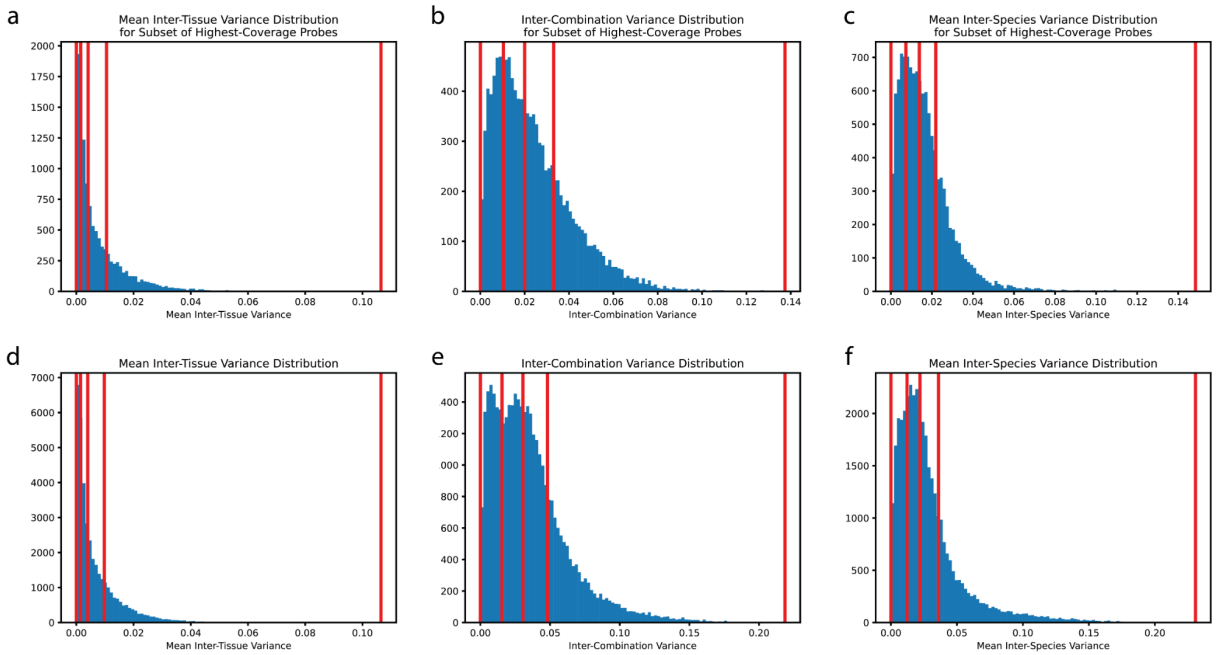


Figure S5. Probe-wise variance distributions with quartiles. a-f) Histogram showing the distribution of probe-wise variances in the observed data. Each histogram has five red vertical lines denoting the boundaries of the quartiles of that variance metric. **a-c)** Distributions showing the **a)** mean inter-tissue, **b)** inter-combination, and **c)** mean inter-species variances across the subset of highest-coverage probes. **d-f)** Corresponding variance distribution plots for all probes.

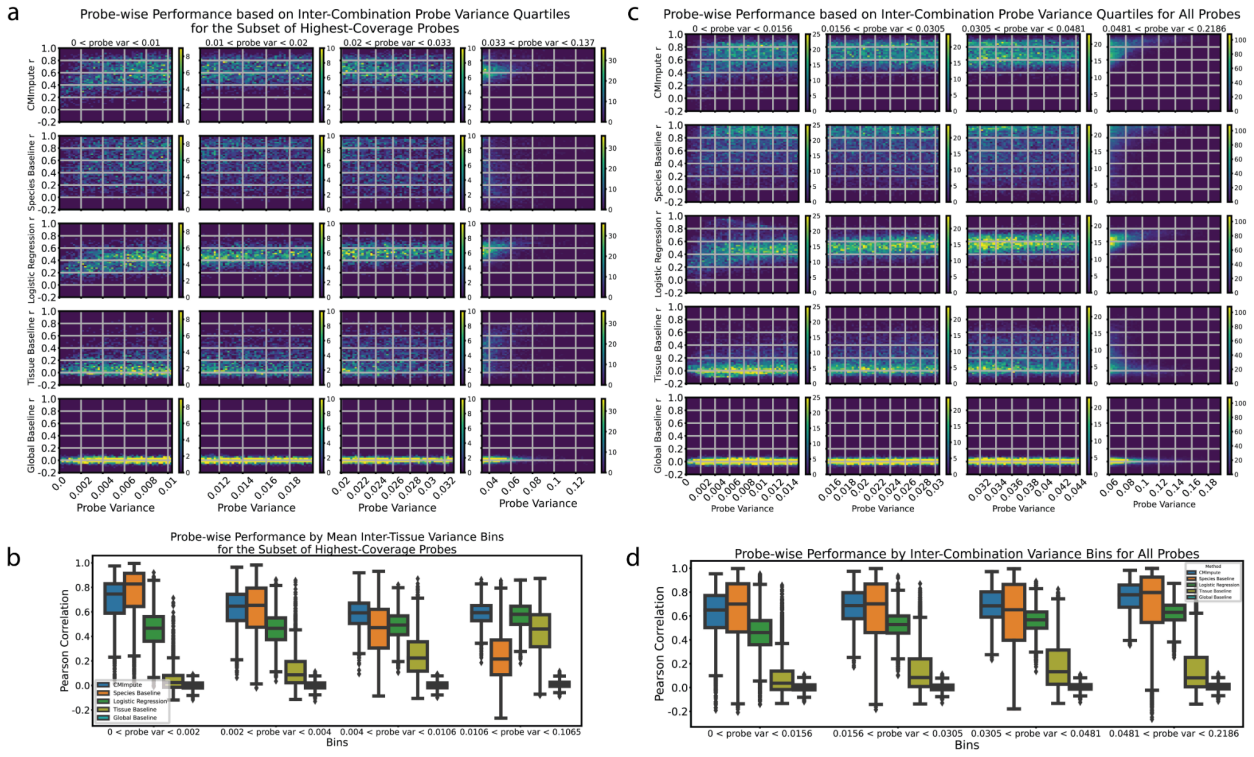


Figure S6. Probe-wise Pearson correlation relative to mean inter-combination probe variance. a) 2-d histograms comparing mean inter-combination variance with probe-wise performance when considering the subset of highest-coverage probes with CMImpute, species baseline, logistic regression, tissue baseline, and global baseline (displayed from top to bottom) probe-wise Pearson correlation. Heatmaps are formatted in the same way as Fig. 4c. Each row contains four heatmaps corresponding to variance quartiles. Within each variance quartile, the heatmap shows the number of probes within a probe variance bin along the x-axis and a probe-wise Pearson correlation bin along the y-axis split into 50 bins along each axis. **b)** Boxplots of probe-wise Pearson correlation with held-out observed values for highest-coverage probes in each mean inter-combination variance quartile. Each variance quartile represented in the boxplots correspond to the variance quartile in the 2-d histograms from a. **c)** Same as a) but when considering all probes. **d)** Same as b) but when considering all probes.

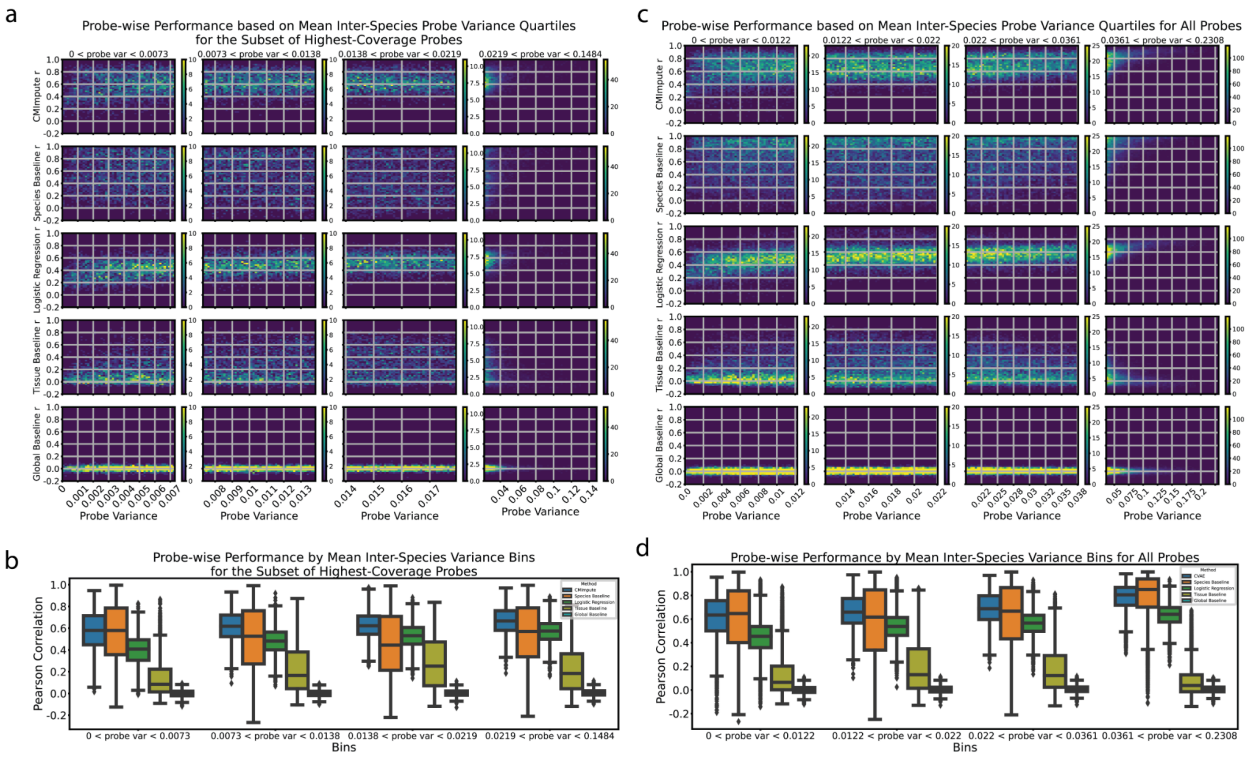


Figure S7. Probe-wise Pearson correlation relative to mean inter-species probe variance. a) 2-d histograms comparing mean inter-species variance with probe-wise performance when considering the subset of highest-coverage probes with CMImpute, species baseline, logistic regression, tissue baseline, and global baseline (displayed from top to bottom) probe-wise Pearson correlation. Heatmaps are formatted in the same way as Fig. 4c. Each row contains four heatmaps corresponding to variance quartiles. Within each variance quartile, the heatmap shows the number of probes within a probe variance bin along the x-axis and a probe-wise Pearson correlation bin along the y-axis split into 50 bins along each axis. **b)** Boxplots of probe-wise Pearson correlation with held-out observed values for highest-coverage probes in each mean inter-species variance quartile. Each variance quartile represented in the boxplots correspond to the variance quartile in the 2-d histograms from a. **c)** Same as a) but when considering all probes. **d)** Same as b) but when considering all probes.

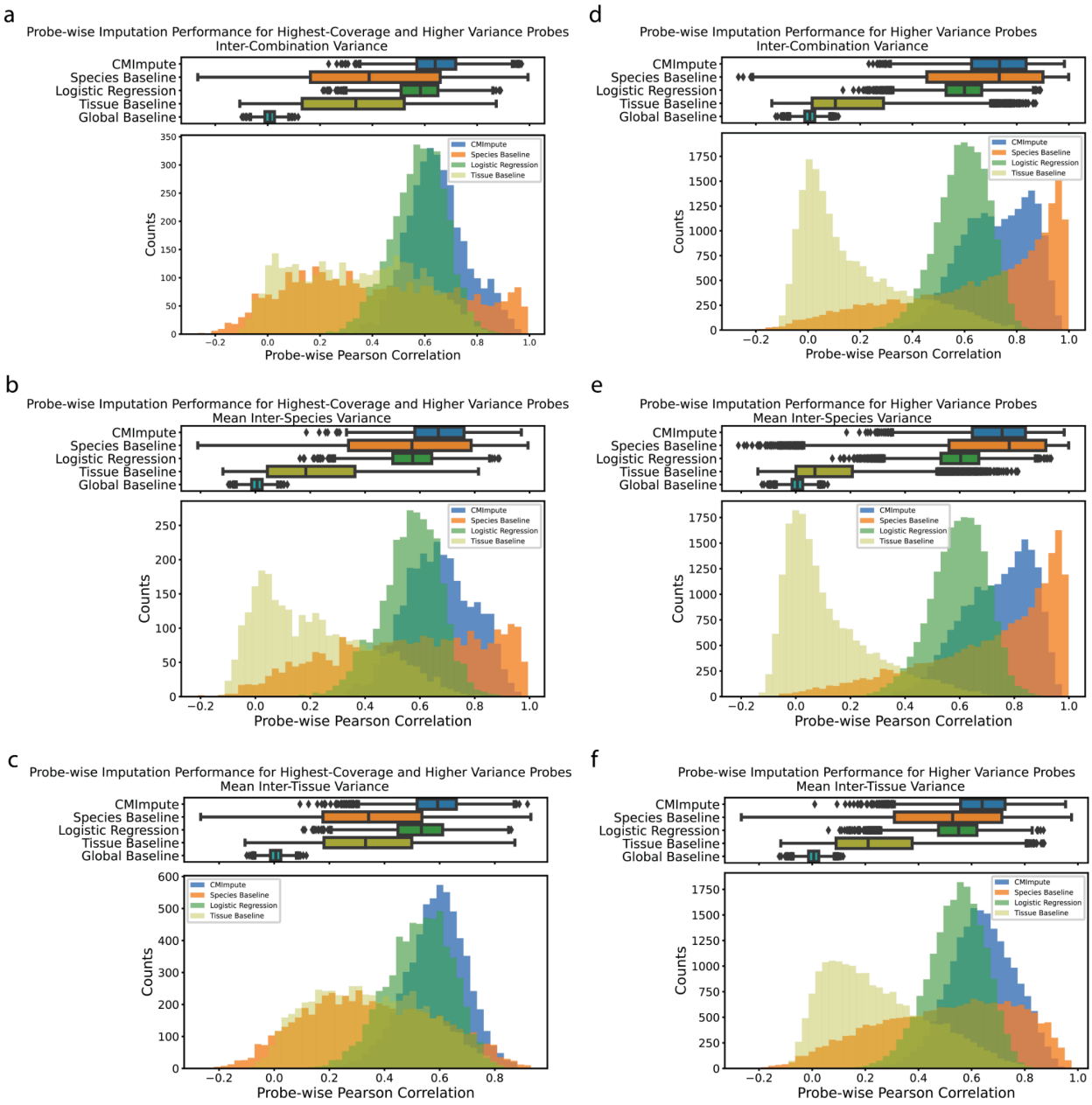


Figure S8. Probe-wise Pearson correlation distributions for higher variance and highest-coverage probes. Distributions of probe-wise Pearson correlations with held-out observed values for **a)** 3,390 probes that are among the set of the highest-coverage probes and have an inter-combination variance > 0.031 (median inter-combination variance among the set of highest-coverage probes), **b)** 2,883 probes that are among the set of the highest-coverage and have a mean inter-species variance > 0.022 (median mean inter-species variance among the set of

highest-coverage probes), **c)** 5,997 probes that are among the set of the highest-coverage probes and have a mean inter-tissue variance > 0.004 (median mean inter-tissue variance among the set of highest-coverage probes), **d)** 18,746 probes with an inter-combination variance > 0.031 (median inter-combination variance), **e)** 18,746 probes with a mean inter-species variance > 0.022 (median mean inter-species variance), and **f)** 18,746 probes with a mean inter-tissue variance > 0.004 (median mean inter-tissue variance). The top boxplots show the distribution of probe-wise correlations with held-out observed values. The bottom histograms show the number of imputed combination mean samples across 50 performance bins. Legend for both boxplots and histograms shown in histogram plot and indicate which method (CMImpute and all baselines) are being considered. As the global baseline predictions do not vary within a fold, the probe-wise performance is not meaningful and this is not included in the histograms.

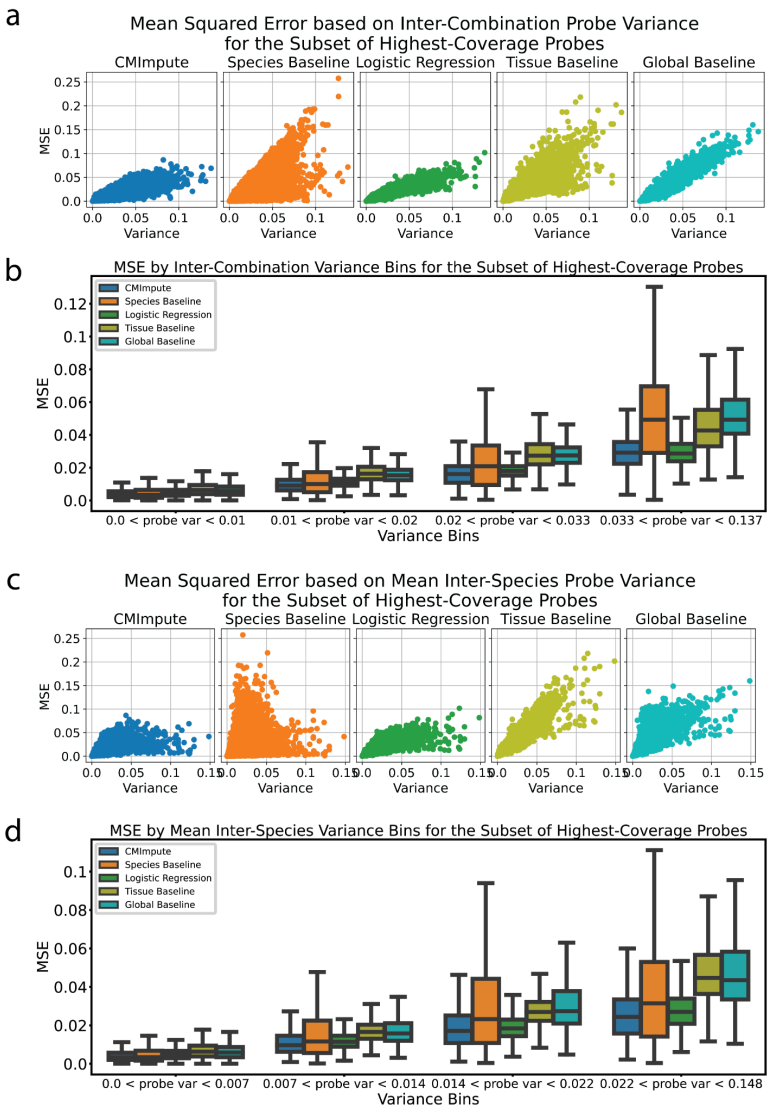


Figure S9. Probe-wise mean squared error based on inter-combination and mean inter-species variance. a) Scatterplot showing the relationship between probe-wise MSE and mean inter-combination variance restricting to the highest-coverage probes in the same format as Fig. 4d. The y-axis is the probe-wise MSE with held-out observed values for CMImpute, species baseline, logistic regression, tissue baseline, or global baseline (left to right). The x-axis is the probe variance. Each dot corresponds to a single imputed probe. **b)** Boxplot of the probe-wise MSE with held-out observed data across mean inter-combination probe variance bins similar to Fig. 4f. **c)** Scatter plot similar to a) for the mean inter-species variance. **d)** Boxplot similar to b)

for the mean inter-species variance.

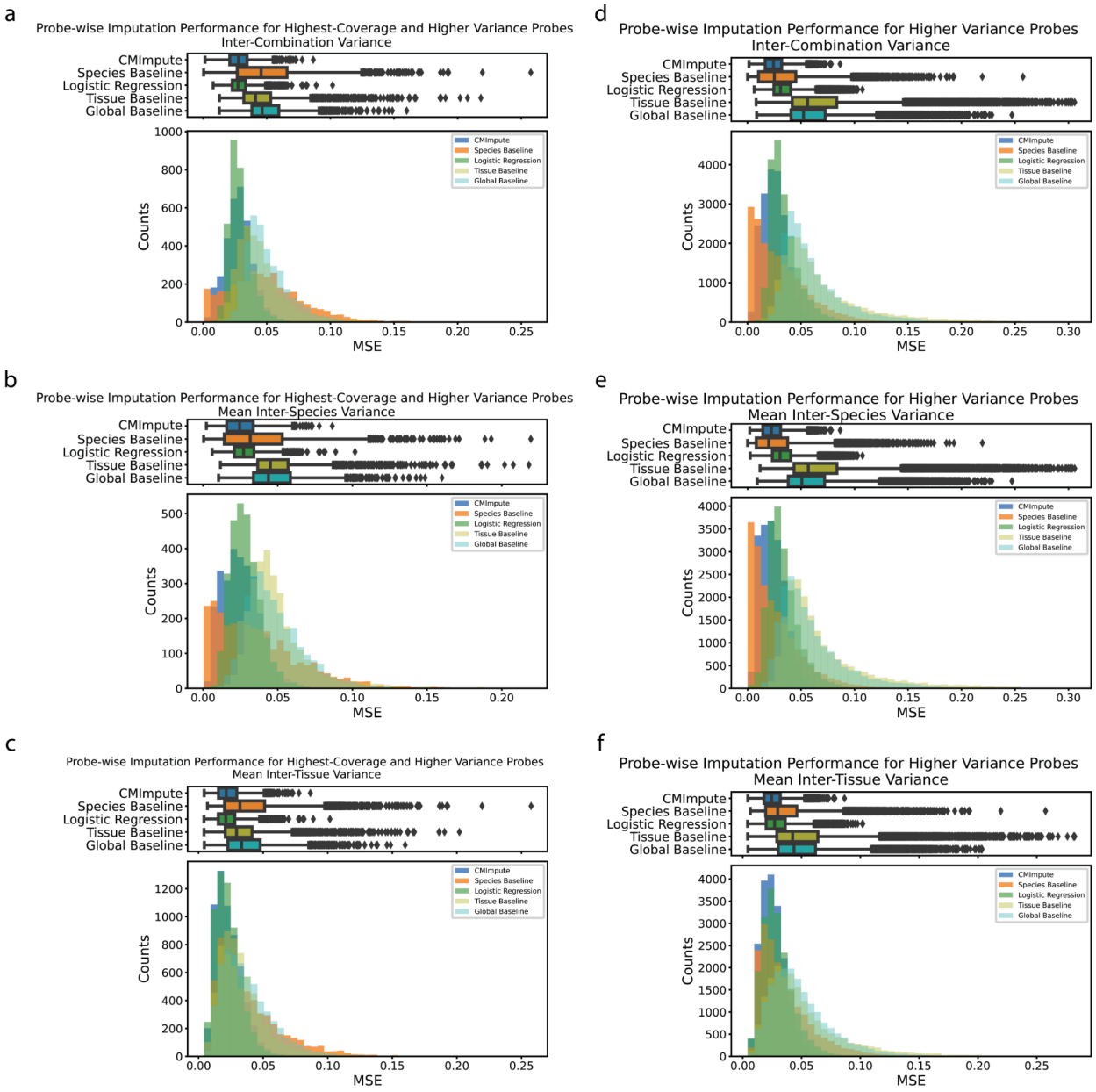


Figure S10. Probe-wise MSE distributions for higher variance probes and highest-coverage probes. Analogous to Pearson correlation distributions in Figure S8. Distributions of probe-wise MSE with held-out observed values for **a)** 3,390 probes that are among the set of the highest-coverage probes and have an inter-combination variance > 0.031 (median inter-combination variance among the set of highest-coverage probes), **b)** 2,883 probes that are

among the set of the highest-coverage and have a mean inter-species variance > 0.022 (median mean inter-species variance among the set of highest-coverage probes), **c)** 5,997 probes that are among the set of the highest-coverage probes and have a mean inter-tissue variance > 0.004 (median mean inter-tissue variance among the set of highest-coverage probes), **d)** 18,746 probes with an inter-combination variance > 0.031 (median inter-combination variance), **e)** 18,746 probes with a mean inter-species variance > 0.022 (median mean inter-species variance), and **f)** 18,746 probes with a mean inter-tissue variance > 0.004 (median mean inter-tissue variance). The top boxplots show the distribution of probe-wise correlations with held-out observed values. The bottom histograms show the number of imputed combination mean samples across 50 performance bins. Legend for both boxplots and histograms shown in histogram plot and indicate which method (CMImpute and all baselines) are being considered. As the global baseline predictions do not vary within a fold, the probe-wise performance is not meaningful and this is not included in the histograms.

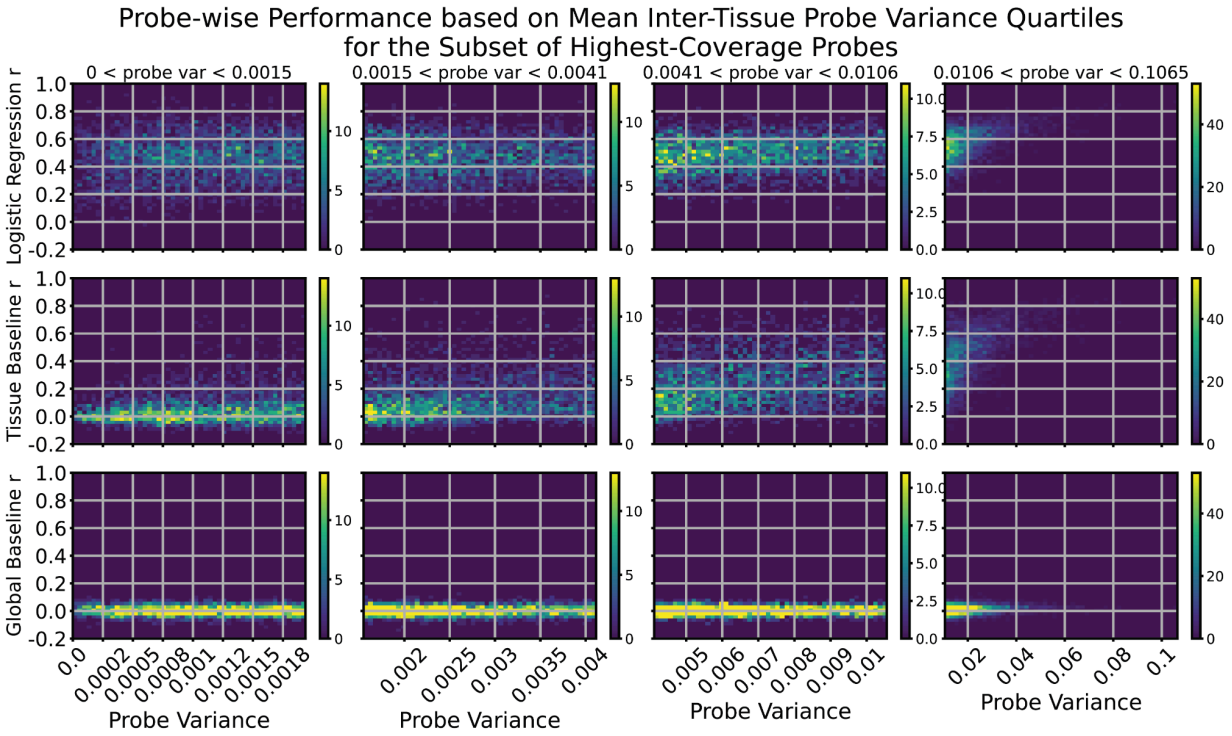


Figure S11. Probe-wise Pearson correlation relative to mean inter-tissue probe variance for baselines. Similar 2-d histogram as Fig. 4c comparing mean inter-tissue variance with probe-wise Pearson correlation shown here for logistic regression (top row), tissue baseline (middle row), and global baseline (bottom row) probe performance for the subset of highest-coverage probes.

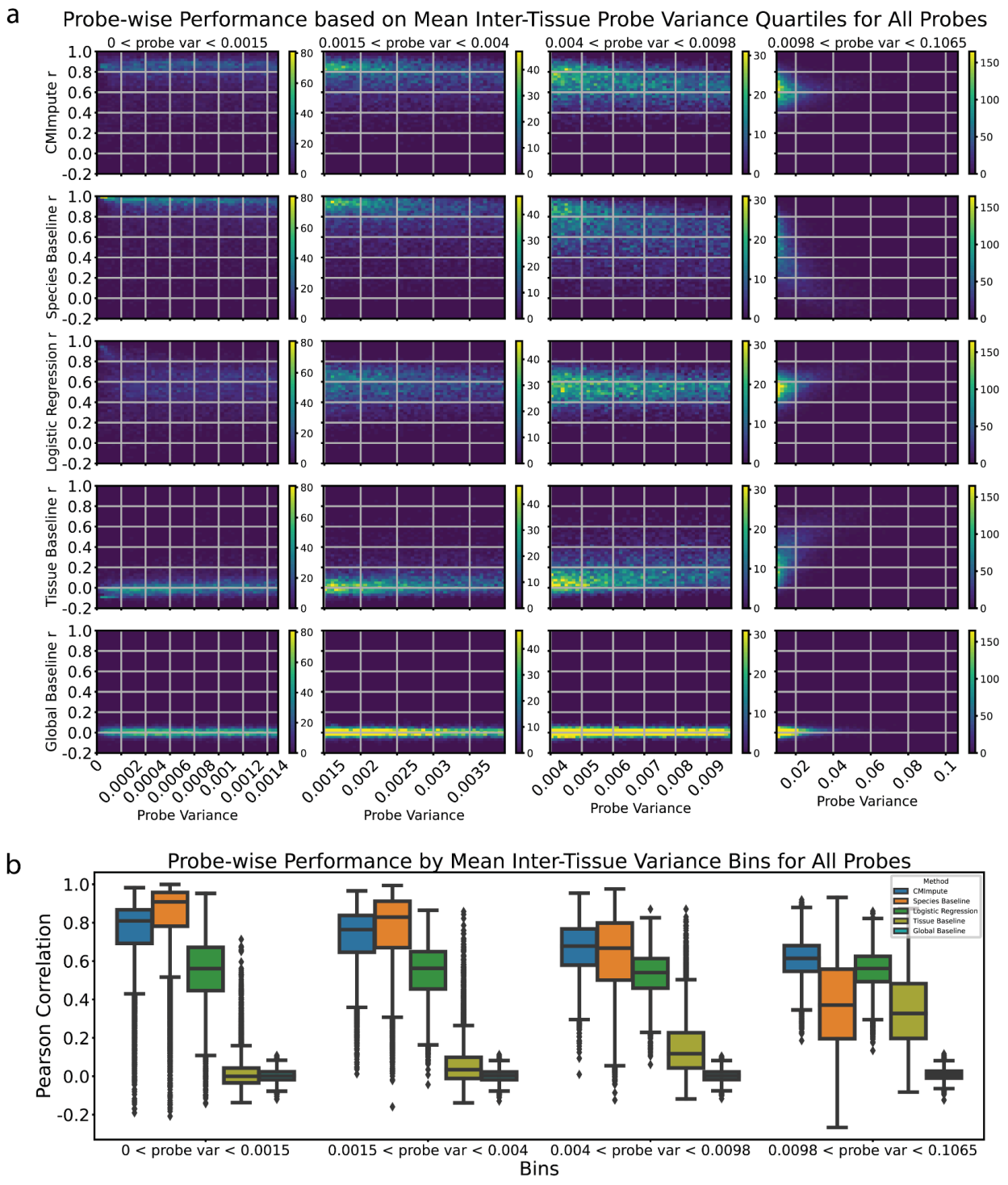


Figure S12. Probe-wise Pearson correlation relative to mean inter-tissue probe variance for all probes. a) 2-d histograms comparing mean inter-tissue variance with CMImpute, species baseline, logistic regression, tissue baseline, and global baseline (displayed from top to bottom)

probe-wise Pearson correlation when considering all probes. Heatmaps are formatted in the same way as Fig. 4c. Each row contains four heatmaps corresponding to variance quartiles. Within each variance quartile, the heatmap shows the number of probes within a probe variance bin along the x-axis and a probe-wise Pearson correlation bin along the y-axis split into 50 bins along each axis. **b)** Boxplots of probe-wise Pearson correlation with held-out observed values for probes in each mean inter-tissue variance quartile when considering all probes. Each variance quartile represented in the boxplots corresponds to the variance quartile in the 2-d histograms from a).

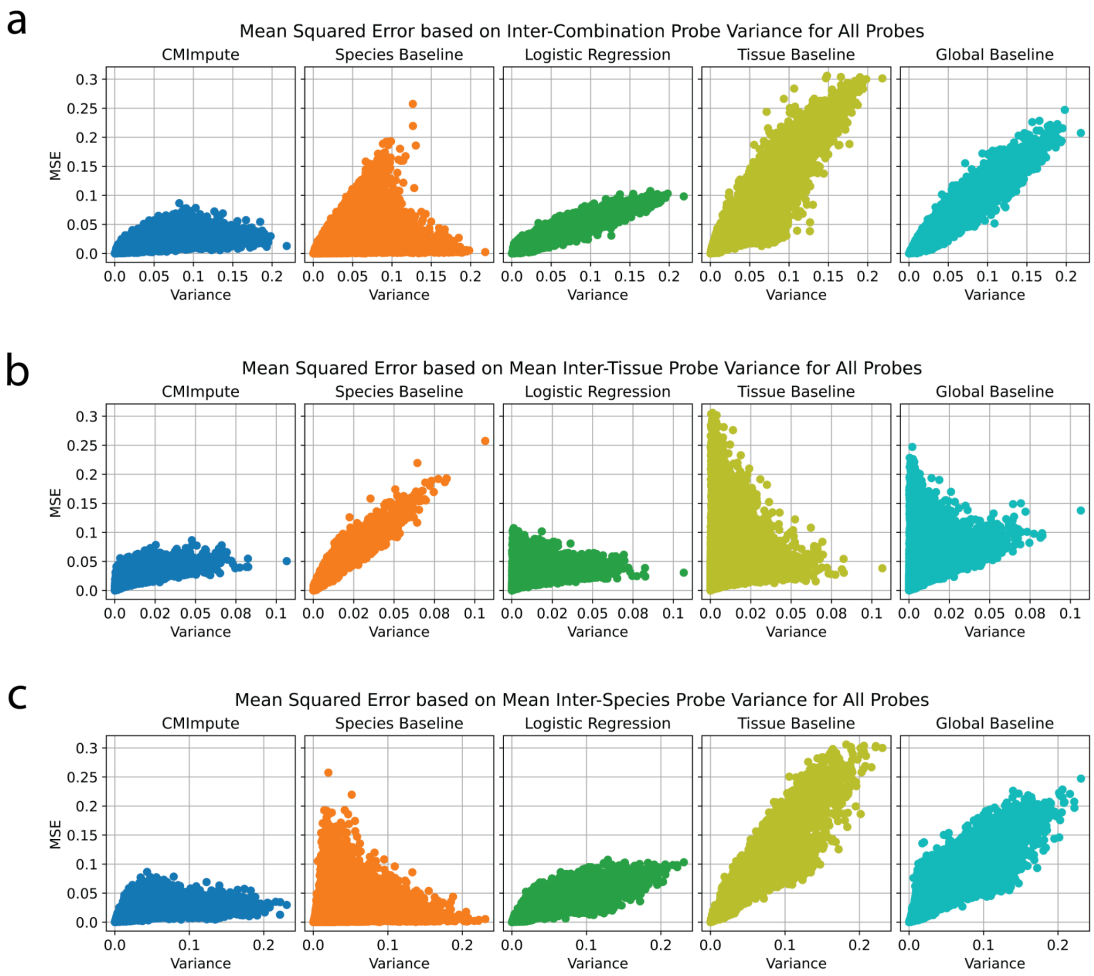


Figure S13. Probe-wise MSE relative to probe variance for all probes. Scatterplots showing the relationship between probe-wise MSE and **a)** inter-combination, **b)** mean inter-tissue, and **c)** mean inter-species variance when considering all probes in the same format as Fig. 4d. The y-axis is the probe-wise MSE with held-out observed values for CMImpute, species baseline, logistic regression, tissue baseline, or global baseline (left to right). The x-axis is the probe variance. Each dot corresponds to a single imputed probe.

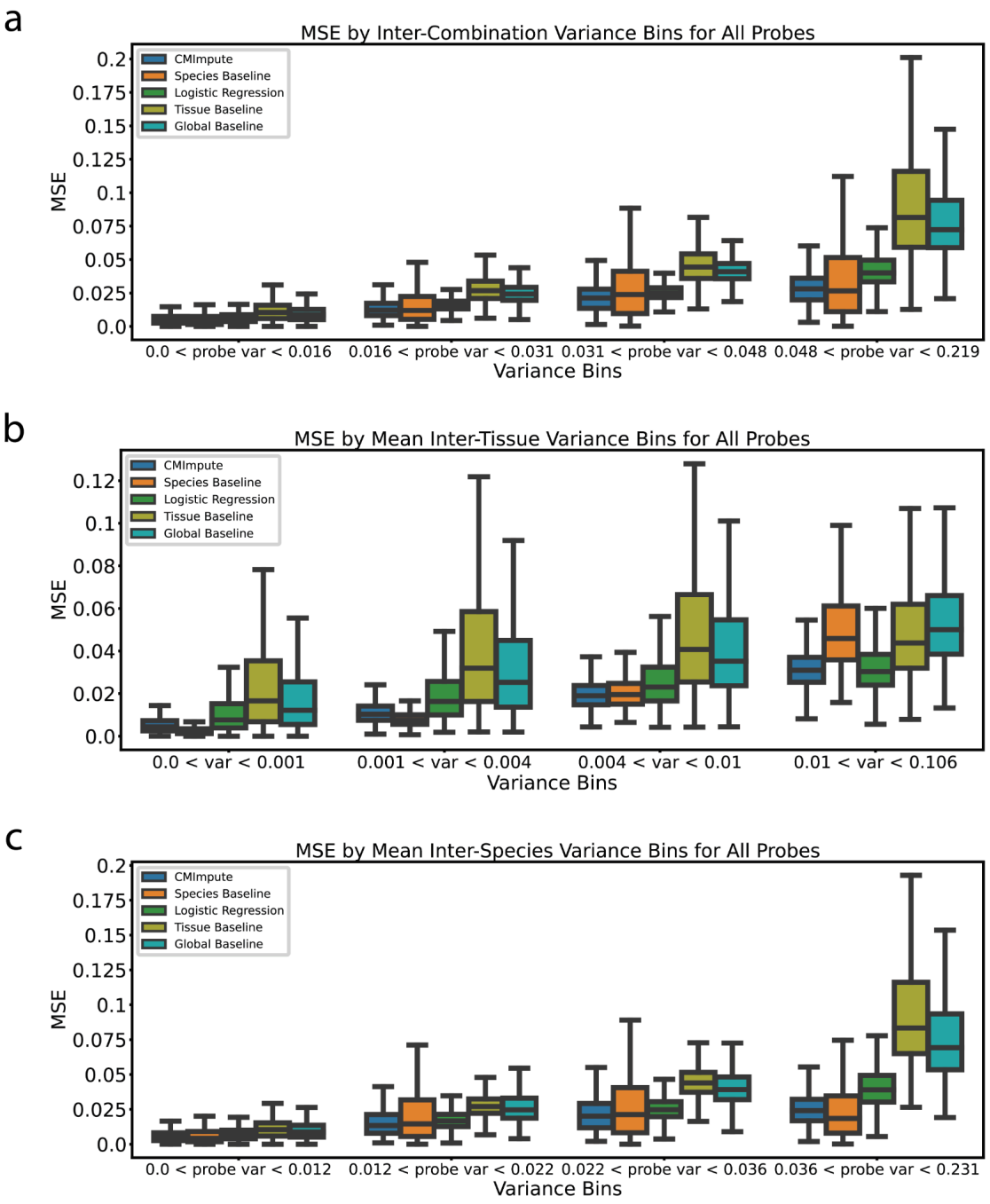
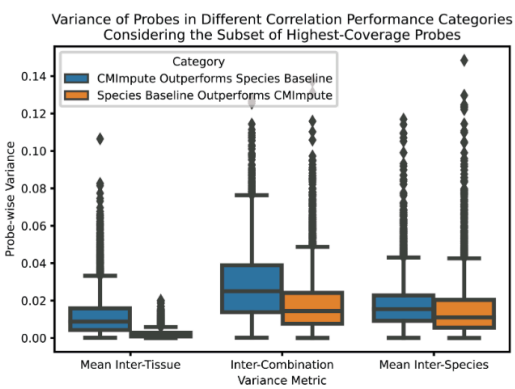
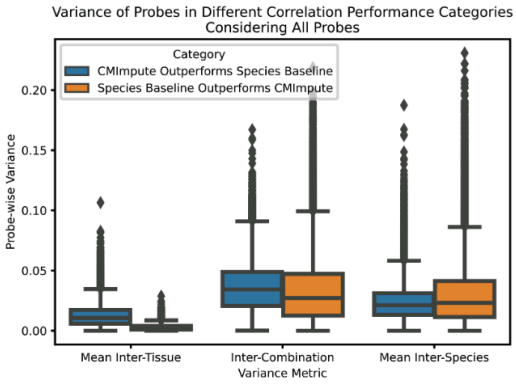


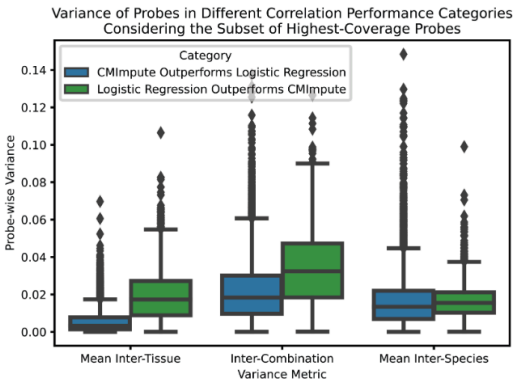
Figure S14. Mean probe-wise MSE across probe variance bins for all probes. Boxplots of the probe-wise MSE with held-out observed data across probe variance bins similar to Fig. 4f. Considering **a)** inter-combination, **b)** mean inter-tissue, and **c)** mean inter-species variance across all probes.

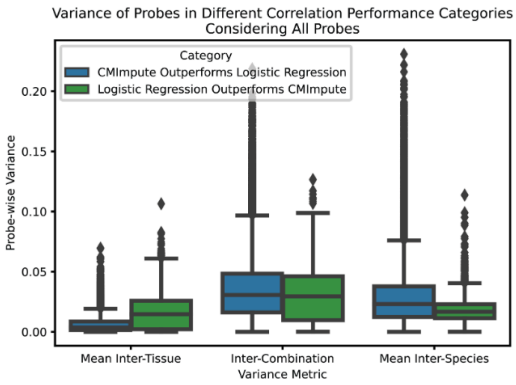
a


Variance of Probes in Different Correlation Performance Categories
Considering the Subset of Highest-Coverage Probes
Category
CMImpute Outperforms Species Baseline
Species Baseline Outperforms CMImpute
Probe-wise Variance
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
Mean Inter-Tissue
Inter-Combination
Mean Inter-Species
Variance Metric

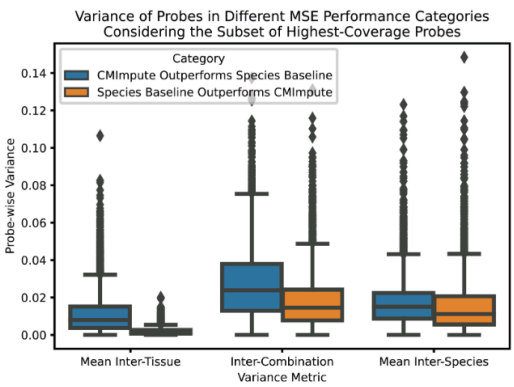

Variance of Probes in Different Correlation Performance Categories
Considering All Probes
Category
CMImpute Outperforms Species Baseline
Species Baseline Outperforms CMImpute
Probe-wise Variance
0.00
0.05
0.10
0.15
0.20
Mean Inter-Tissue
Inter-Combination
Mean Inter-Species
Variance Metric

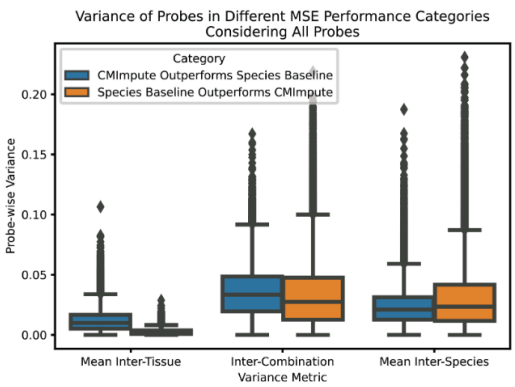
b


Variance of Probes in Different Correlation Performance Categories
Considering the Subset of Highest-Coverage Probes
Category
CMImpute Outperforms Logistic Regression
Logistic Regression Outperforms CMImpute
Probe-wise Variance
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
Mean Inter-Tissue
Inter-Combination
Mean Inter-Species
Variance Metric

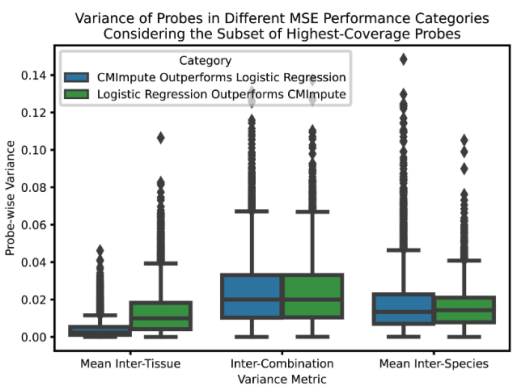

Variance of Probes in Different Correlation Performance Categories
Considering All Probes
Category
CMImpute Outperforms Logistic Regression
Logistic Regression Outperforms CMImpute
Probe-wise Variance
0.00
0.05
0.10
0.15
0.20
Mean Inter-Tissue
Inter-Combination
Mean Inter-Species
Variance Metric

c


Variance of Probes in Different MSE Performance Categories
Considering the Subset of Highest-Coverage Probes
Category
CMImpute Outperforms Species Baseline
Species Baseline Outperforms CMImpute
Probe-wise Variance
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
Mean Inter-Tissue
Inter-Combination
Mean Inter-Species
Variance Metric


Variance of Probes in Different MSE Performance Categories
Considering All Probes
Category
CMImpute Outperforms Species Baseline
Species Baseline Outperforms CMImpute
Probe-wise Variance
0.00
0.05
0.10
0.15
0.20
Mean Inter-Tissue
Inter-Combination
Mean Inter-Species
Variance Metric

d


Variance of Probes in Different MSE Performance Categories
Considering the Subset of Highest-Coverage Probes
Category
CMImpute Outperforms Logistic Regression
Logistic Regression Outperforms CMImpute
Probe-wise Variance
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
Mean Inter-Tissue
Inter-Combination
Mean Inter-Species
Variance Metric

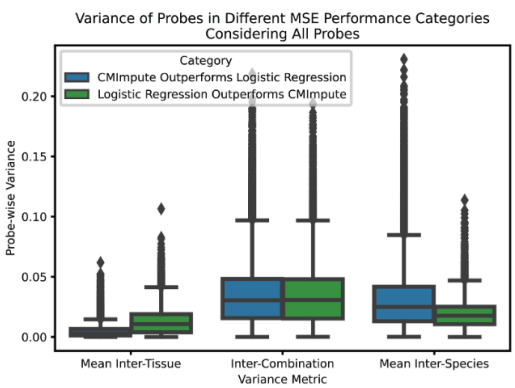

Variance of Probes in Different MSE Performance Categories
Considering All Probes
Category
CMImpute Outperforms Logistic Regression
Logistic Regression Outperforms CMImpute
Probe-wise Variance
0.00
0.05
0.10
0.15
0.20
Mean Inter-Tissue
Inter-Combination
Mean Inter-Species
Variance Metric

Figure S15. Comparison of subset of probes where species baseline and logistic regression outperform CMImpute and vice versa. a-d) The probe-wise variance of probes in the subset of highest-coverage (left) and overall probes (right) for mean inter-tissue, inter-combination, and mean inter-species variance in the subsets of probes where **a)** CMImpute yields a higher probe-wise correlation than the species baseline, **b)** CMImpute yields a higher probe-wise correlation than logistic regression, **c)** CMImpute yields a lower MSE than the species baseline, and **d)** CMImpute yields a lower MSE than logistic regression and vice versa.

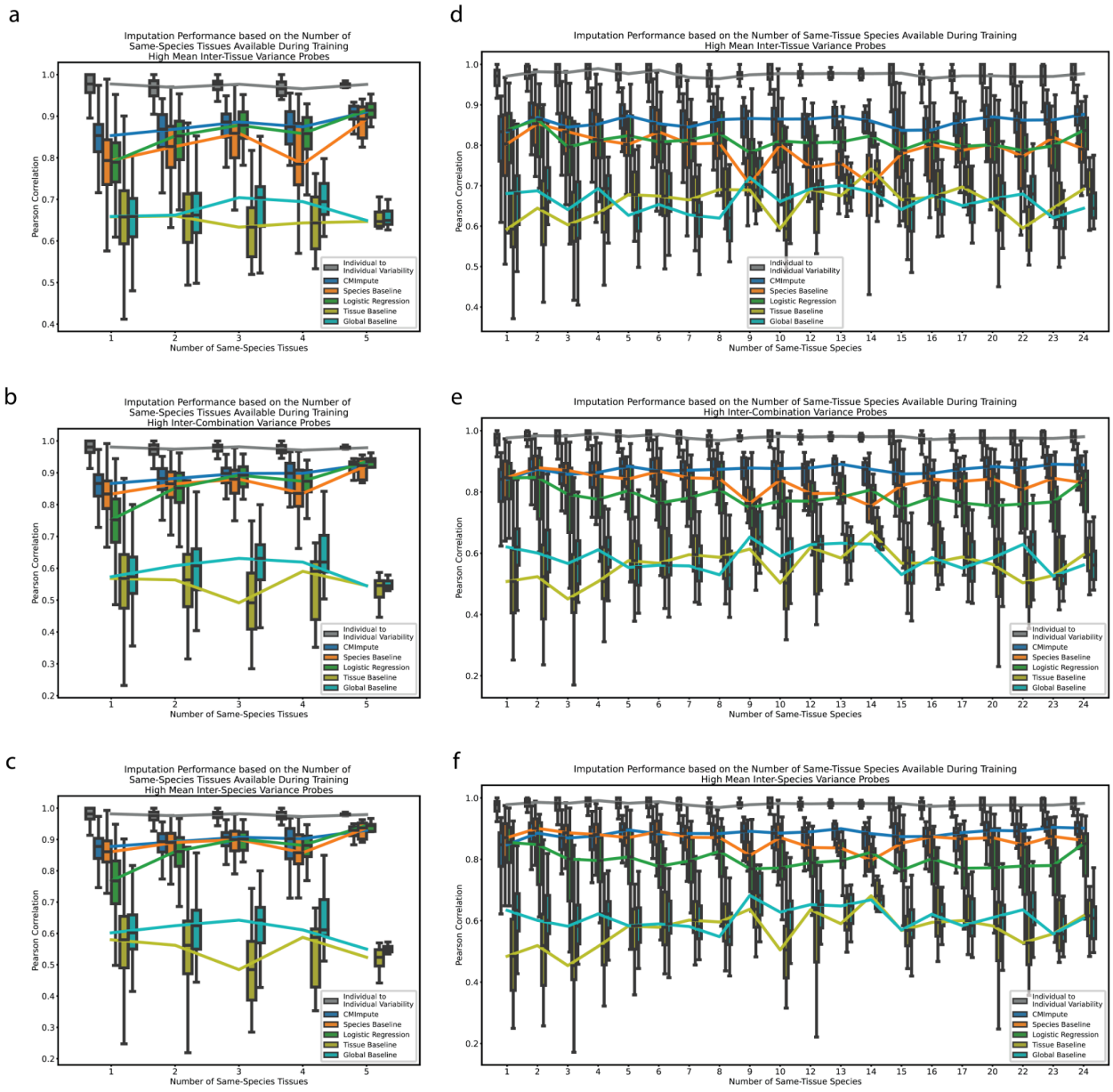


Figure S16. Impact of the amount of available training data on CMImpute and baseline performance across higher variance probes. a-c) Sample-wise Pearson correlation across a subset of **a)** 18,746 probes with a mean inter-tissue variance > 0.004 (median mean inter-tissue variance), **b)** 18,746 probes with an inter-combination variance > 0.031 (median inter-combination variance), and **c)** 18,746 probes with a mean inter-species variance > 0.022 (median mean inter-species variance) based on the number of tissue types available in the target

species during training. **d-f)** Sample-wise Pearson correlation across the corresponding set of higher variance probes in a-c based on the number of species available during training in the target tissue.

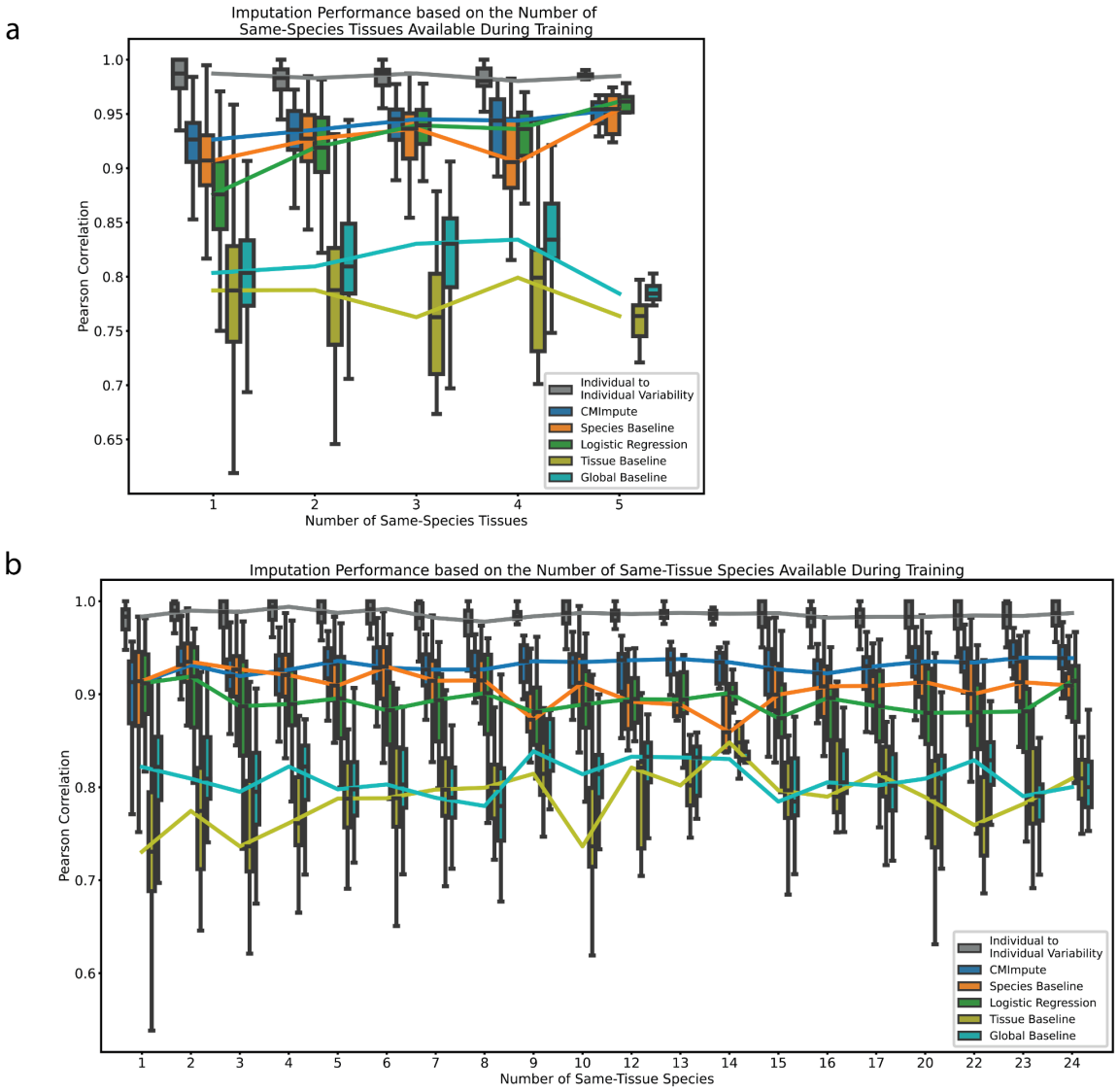


Figure S17. Impact of the amount of available training data on CMImpute and baseline performance. a) Sample-wise Pearson correlation based on the number of tissue types available in the target species during training. The box plot shows the distribution of Pearson correlation for each number of tissue types. Line connects the median correlations for an imputation method across all tissue type counts. Individual to individual variability represents the average pairwise correlation of observed data between individuals of the same species and tissue type for each combination. **b)** Sample-wise Pearson correlation based on the number of species available during training in the target tissue.

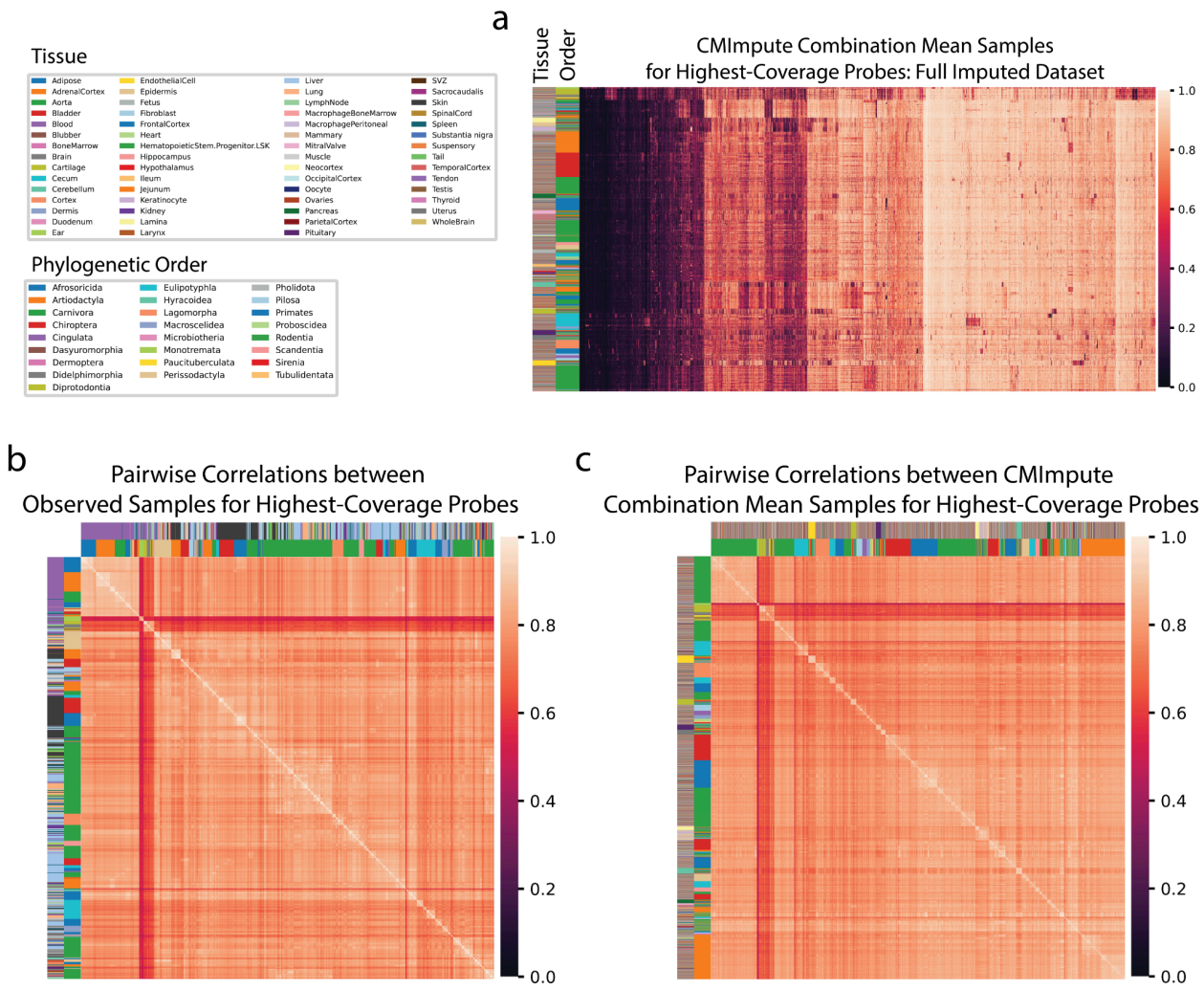


Figure S18. Visualization of highest-coverage probes for CMImpute's full imputed dataset. **a)** Heatmap of imputed dataset's methylation probe values when considering the subset of highest-coverage probes. Each row is an imputed species-tissue combination mean sample and each column is a methylation probe restricted to the subset of highest-coverage probes. Samples and probes ordered by hierarchical clustering followed by optimal leaf ordering. Samples labeled via color bar by phylogenetic order (inner) and tissue (outer). **b-c)** Similar heatmap of pairwise correlations to Fig. 6b-c for **b)** the 746 observed species-tissue combinations or **c)** 20,251 CMImpute-imputed species-tissue combinations when restricted to the subset of highest-coverage probes.

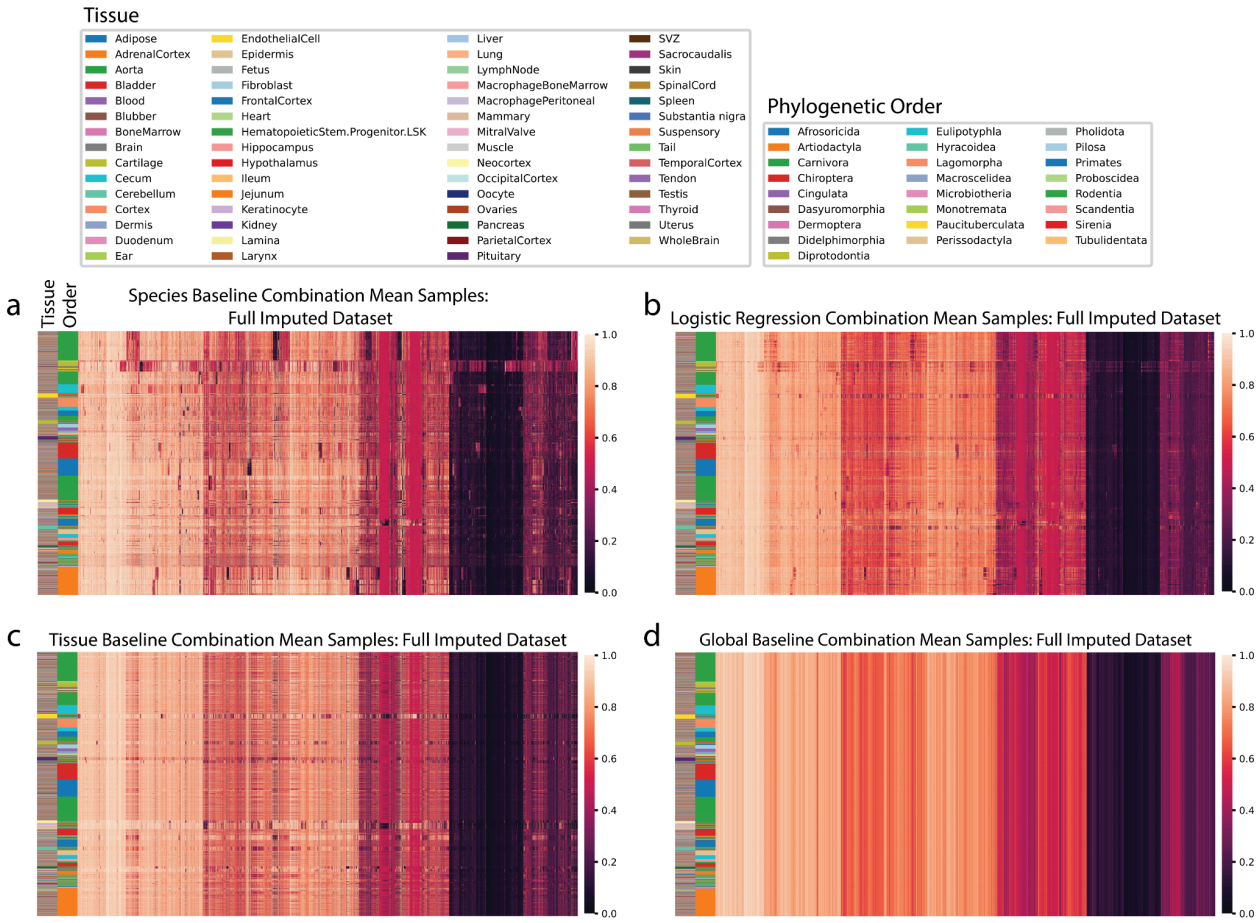


Figure S19. Visualization of baseline-imputed samples of non-observed combinations on all probes. a-d) Heatmaps of **a)** species baseline, **b)** logistic regression, **c)** tissue baseline, and **d)** global baseline-imputed datasets' methylation probe values corresponding to the CMImpute-imputed dataset displayed in Fig. 6a. Samples and probes ordered by hierarchical clustering followed by optimal leaf ordering. Color bars on the left indicate the phylogenetic order (inner) and tissue (outer) corresponding to the samples.

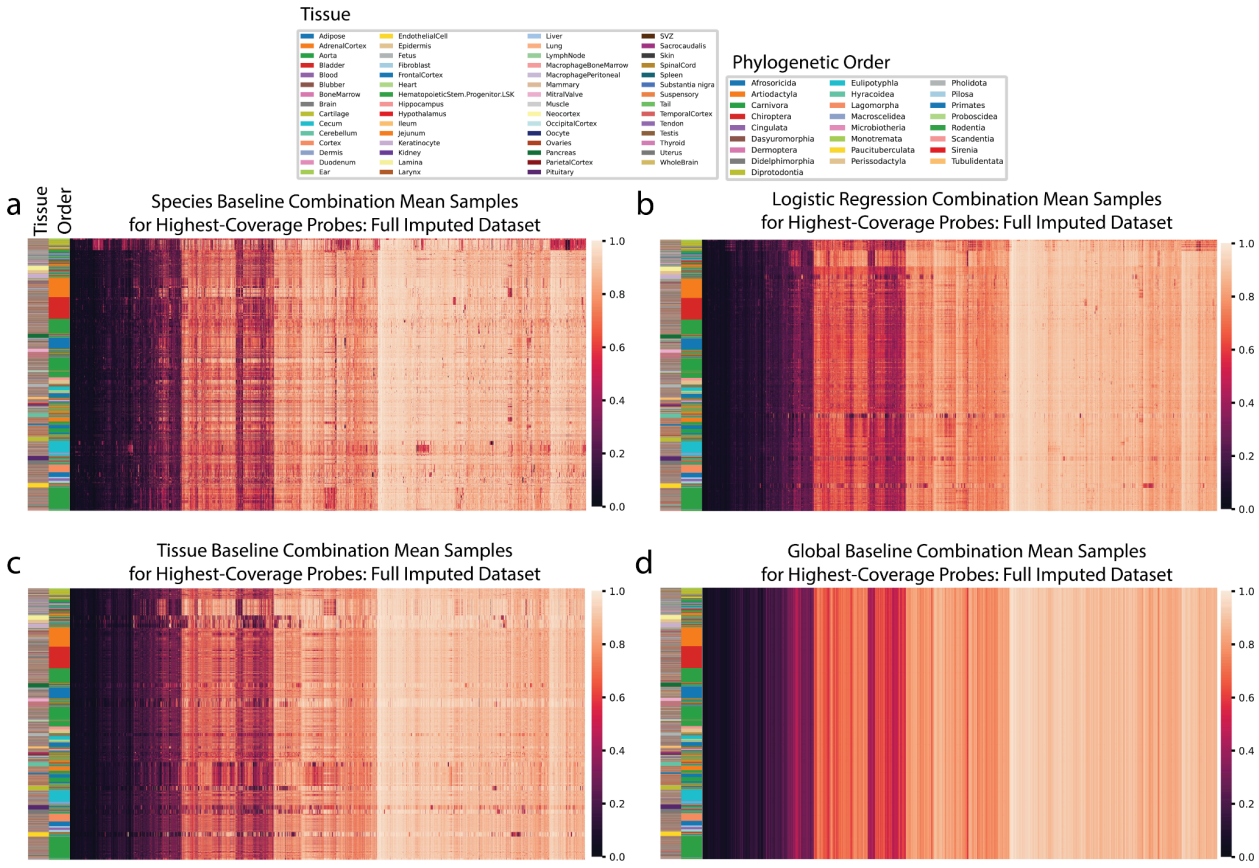


Figure S20. Visualization of baseline-imputed samples of non-observed combinations on the subset of highest-coverage probes. a-d) Heatmaps of **a)** species baseline, **b)** logistic regression, **c)** tissue baseline, and **d)** global baseline-imputed datasets' methylation probe values when considering the subset of highest-coverage probes corresponding to the CMImpute-imputed dataset displayed in Figure S18a. Samples and probes ordered via hierarchical clustering followed by optimal leaf ordering. Color bars on the left indicate the phylogenetic order (inner) and tissue (outer) corresponding to the samples.

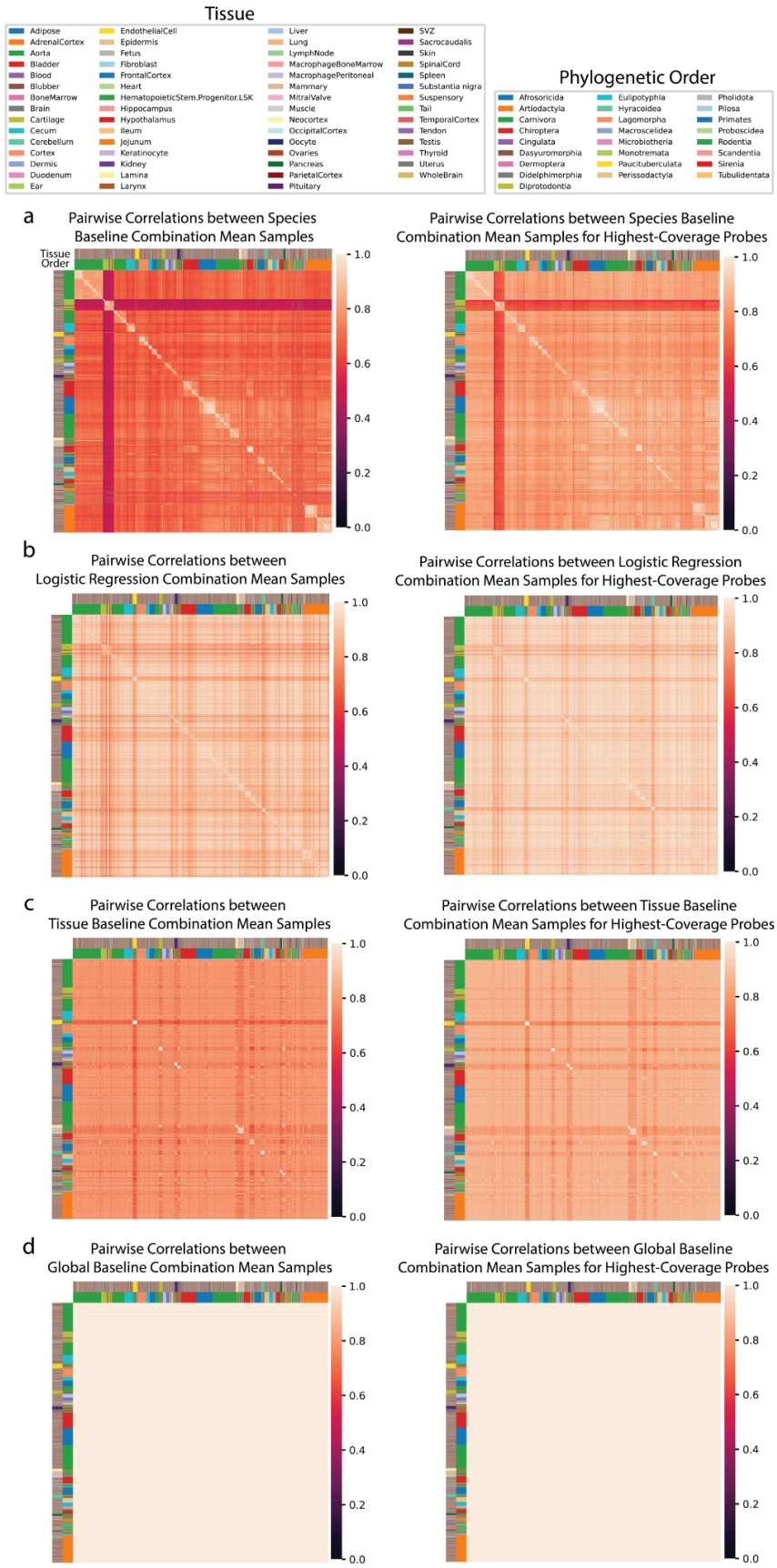

Tissue
Adipose
AdrenalCortex
Aorta
Bladder
Blood
Blubber
BoneMarrow
Brain
Cartilage
Cecum
Cerebellum
Cortex
Dermis
Duodenum
Ear
EndothelialCell
Epidermis
Fetus
Fibroblast
FrontalCortex
Heart
HematopoieticStem.Progenitor.LSK
Hippocampus
Hypothalamus
Ileum
Jejunum
Keratinocyte
Kidney
Lamina
Larynx
Liver
Lung
LymphNode
MacrophageBoneMarrow
MacrophagePeritoneal
Mammary
MitralValve
Muscle
Neocortex
OccipitalCortex
Oocyte
Ovaries
Pancreas
ParietalCortex
Pituitary
SVZ
Sacrocaudalis
Skin
SpinalCord
Spleen
Substantia nigra
Suspensory
Tail
TemporalCortex
Tendon
Testis
Thyroid
Uterus
WholeBrain
Phylogenetic Order
Afrosoricida
Artiodactyla
Carnivora
Chiroptera
Cingulata
Dasyuromorphia
Dermoptera
Didelphimorphia
Diprotodontia
Eulipotyphla
Hyracoidea
Lagomorpha
Macroscelidea
Microbiotheria
Monotremata
Paucituberculata
Perissodactyla
Pholidota
Pilosa
Primates
Proboscidea
Rodentia
Scandentia
Sirenia
Tubulidentata
a
Pairwise Correlations between Species Baseline Combination Mean Samples
Tissue
Order
Pairwise Correlations between Species Baseline Combination Mean Samples for Highest-Coverage Probes
b
Pairwise Correlations between Logistic Regression Combination Mean Samples
Pairwise Correlations between Logistic Regression Combination Mean Samples for Highest-Coverage Probes
c
Pairwise Correlations between Tissue Baseline Combination Mean Samples
Pairwise Correlations between Tissue Baseline Combination Mean Samples for Highest-Coverage Probes
d
Pairwise Correlations between Global Baseline Combination Mean Samples
Pairwise Correlations between Global Baseline Combination Mean Samples for Highest-Coverage Probes
1.0
0.8
0.6
0.4
0.2
0.0

Figure S21. Visualization of species and tissue signals via pairwise correlations for baselines. a-d) Similar heatmaps as in Fig. 6c and Figure S18c here showing pairwise correlations between 20,251 **a)** species baseline, **b)** logistic regression, **c)** tissue baseline, and **d)** global baseline-imputed combination mean samples when considering all probes (left) and the subset of highest-coverage probes (right). Samples are ordered based on hierarchical clustering followed by optimal leaf ordering of the full methylation samples from each dataset. Color bars on the left indicate the phylogenetic order (inner) and tissue (outer) corresponding to the samples.

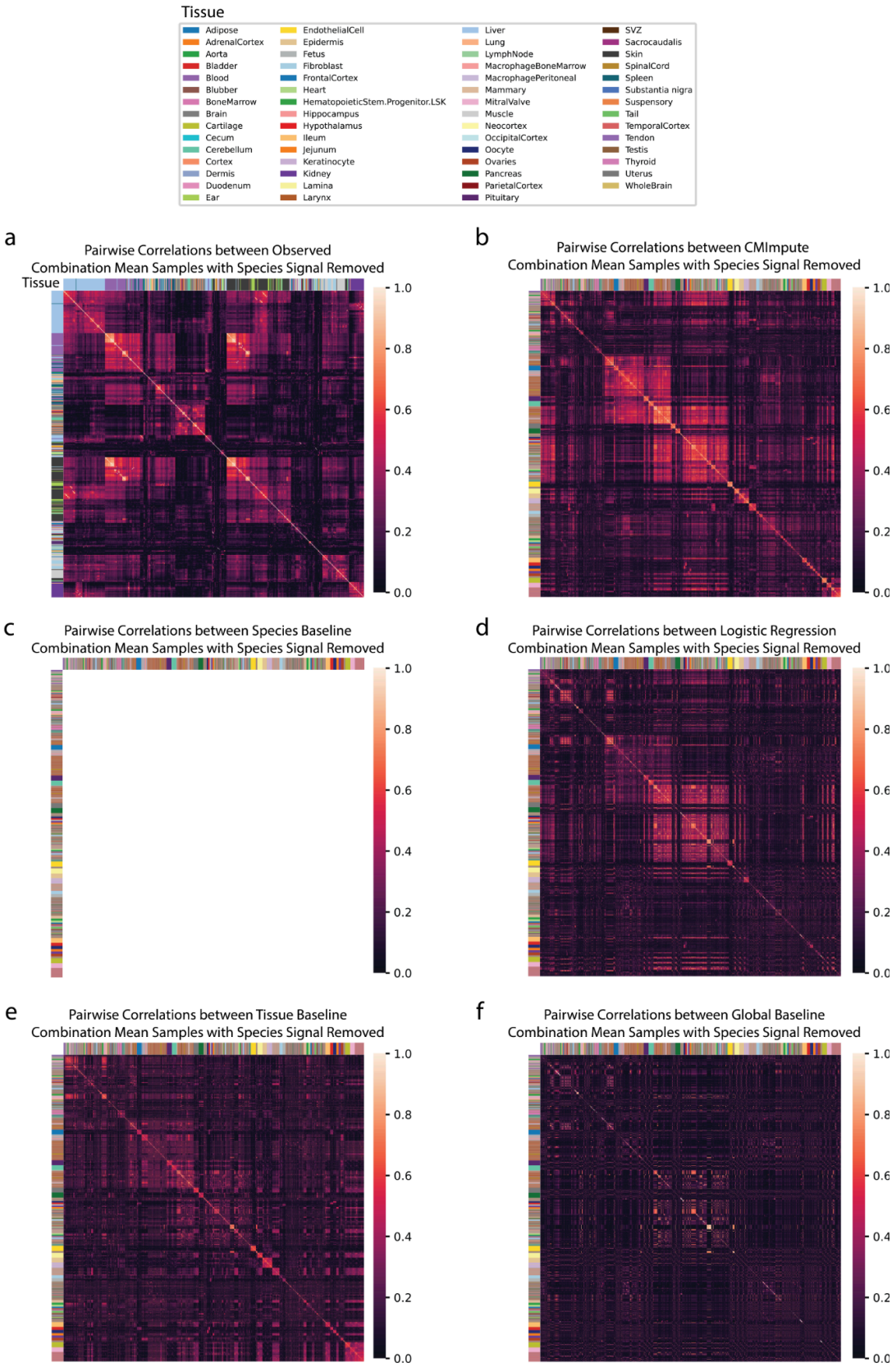

Tissue
Adipose
AdrenalCortex
Aorta
Bladder
Blood
Blubber
BoneMarrow
Brain
Cartilage
Cecum
Cerebellum
Cortex
Dermis
Duodenum
Ear
EndothelialCell
Epidermis
Fetus
Fibroblast
FrontalCortex
Heart
HematopoieticStem.Progenitor.LSK
Hippocampus
Hypothalamus
Ileum
Jejunum
Keratinocyte
Kidney
Lamina
Larynx
Liver
Lung
LymphNode
MacrophageBoneMarrow
MacrophagePeritoneal
Mammary
MitralValve
Muscle
Neocortex
OccipitalCortex
Oocyte
Ovaries
Pancreas
ParietalCortex
Pituitary
SVZ
Sacrocaudalis
Skin
SpinalCord
Spleen
Substantia nigra
Suspensory
Tail
TemporalCortex
Tendon
Testis
Thyroid
Uterus
WholeBrain
a
Pairwise Correlations between Observed
Combination Mean Samples with Species Signal Removed
Tissue
1.0
0.8
0.6
0.4
0.2
0.0
b
Pairwise Correlations between CMImpute
Combination Mean Samples with Species Signal Removed
c
Pairwise Correlations between Species Baseline
Combination Mean Samples with Species Signal Removed
d
Pairwise Correlations between Logistic Regression
Combination Mean Samples with Species Signal Removed
e
Pairwise Correlations between Tissue Baseline
Combination Mean Samples with Species Signal Removed
f
Pairwise Correlations between Global Baseline
Combination Mean Samples with Species Signal Removed

Figure S22. Visualization of tissue signal in imputed samples of non-observed combinations. a) Heatmap of pairwise correlations between 746 observed combination mean samples, with at least two tissue types available, with the species signal removed. White space represents undefined Pearson correlation values because of no variation in the samples. **b-f)** Heatmaps of pairwise correlations between 20,251 **b)** CMImpute, **c)** species baseline, **d)** logistic regression, **e)** tissue baseline, and **f)** global baseline-imputed combination mean samples with the species signal removed (as shown in Fig. 6d). Samples are ordered based on hierarchical clustering of the difference between the full imputed samples and the species baseline samples followed by optimal leaf ordering.

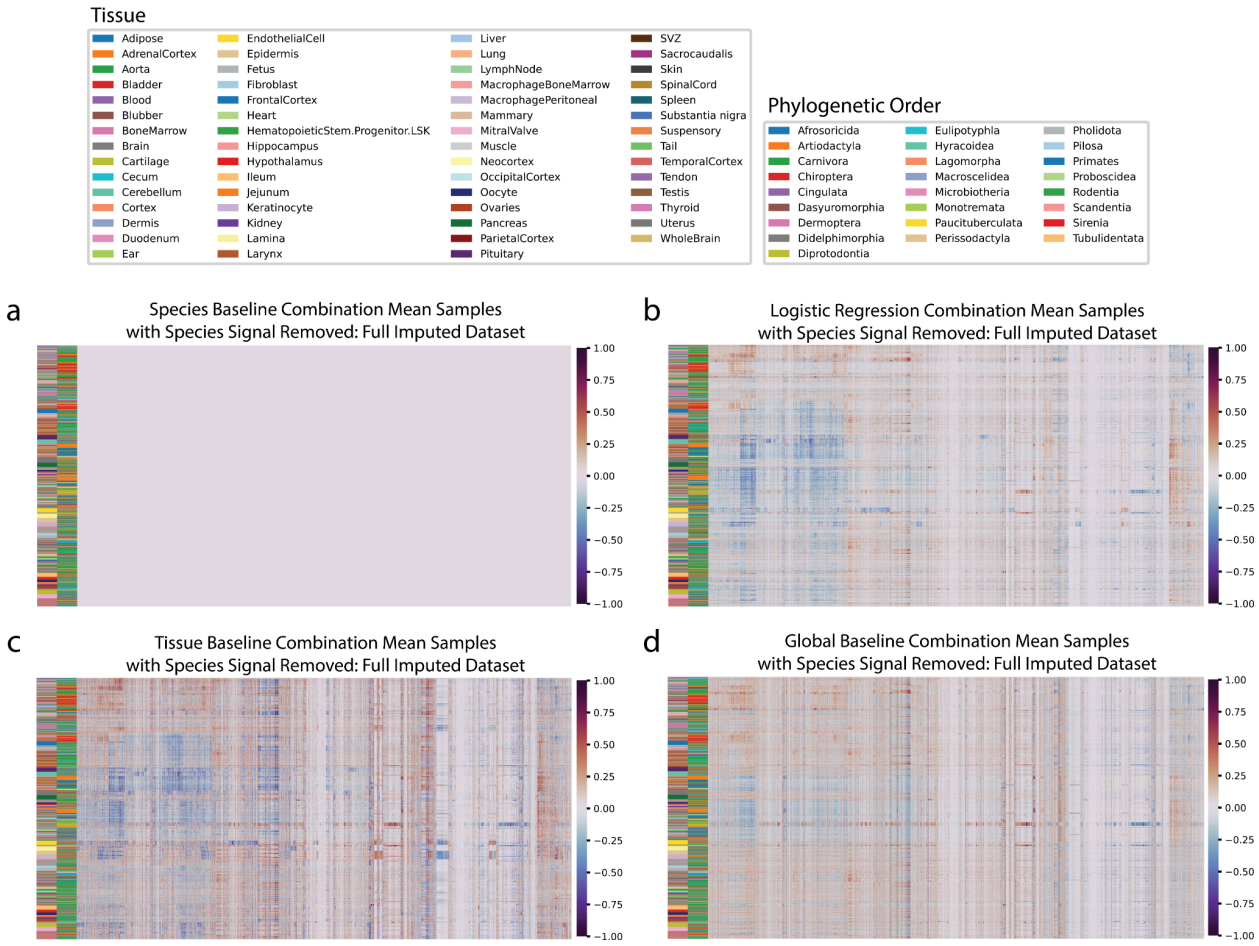


Figure S23. Visualization of baseline-imputed combination mean samples with the species signal removed. a-d) Heatmaps of **a)** species baseline, **b)** logistic regression, **c)** tissue baseline, and **d)** global baseline-imputed combination mean samples with the species signal removed. Sample and probe order, color bar labels, and color scale corresponds to Fig. 6d.

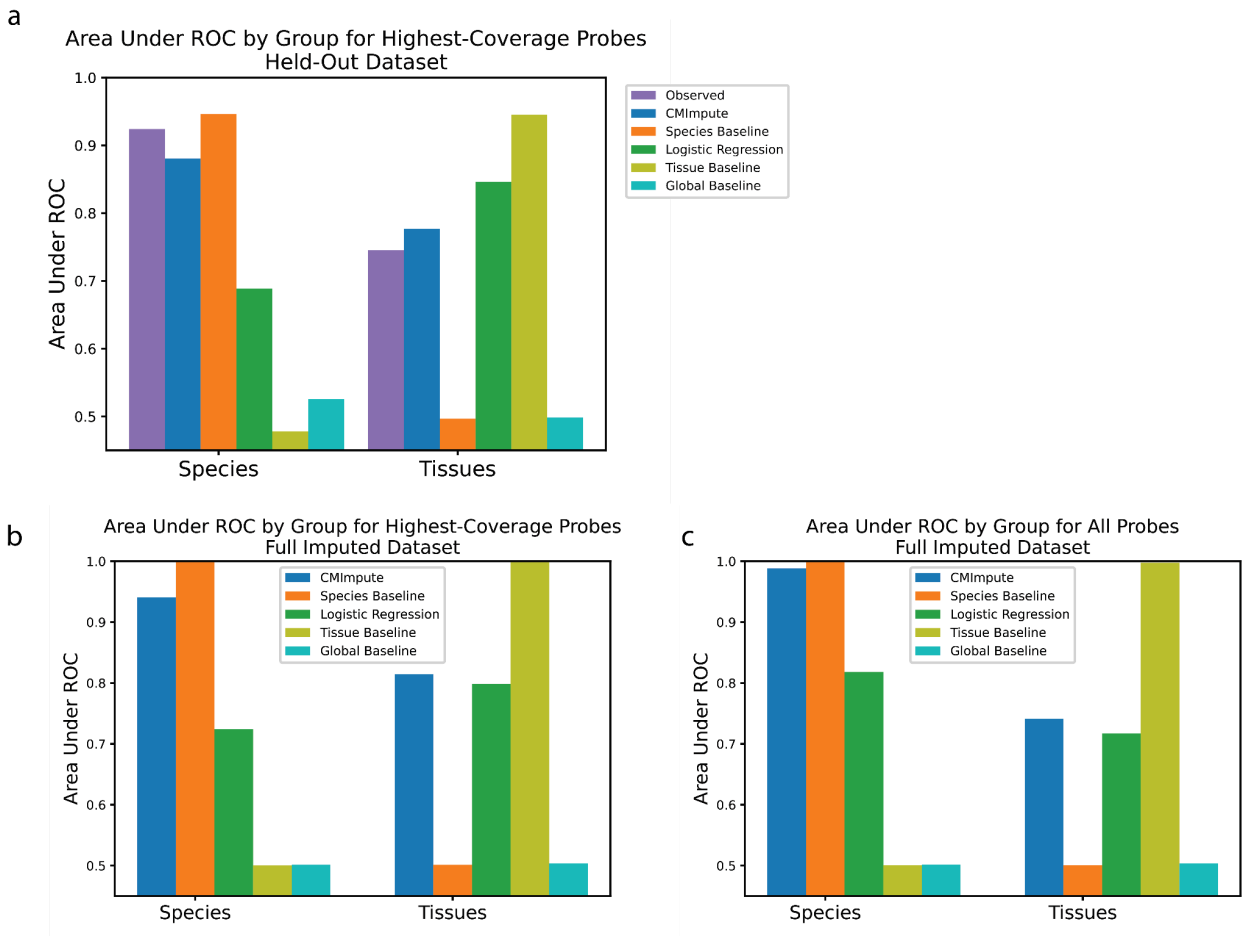


Figure S24. Species and tissue signal for the subset of highest-coverage probes and the full imputed dataset. a) Area Under ROC values for predicting whether samples within the cross-validation dataset are from the same species (left) or tissue (right) based on their pairwise correlations for the subset of highest-coverage probes. Results when considering all probes shown in Fig. 7. **b-c)** Area Under ROC values for predicting whether samples within the full 20,251 combination mean sample imputed dataset are from the same species (left) or tissue (right) based on their pairwise correlations when considering **b)** the subset of highest-coverage probes and **c)** all probes.

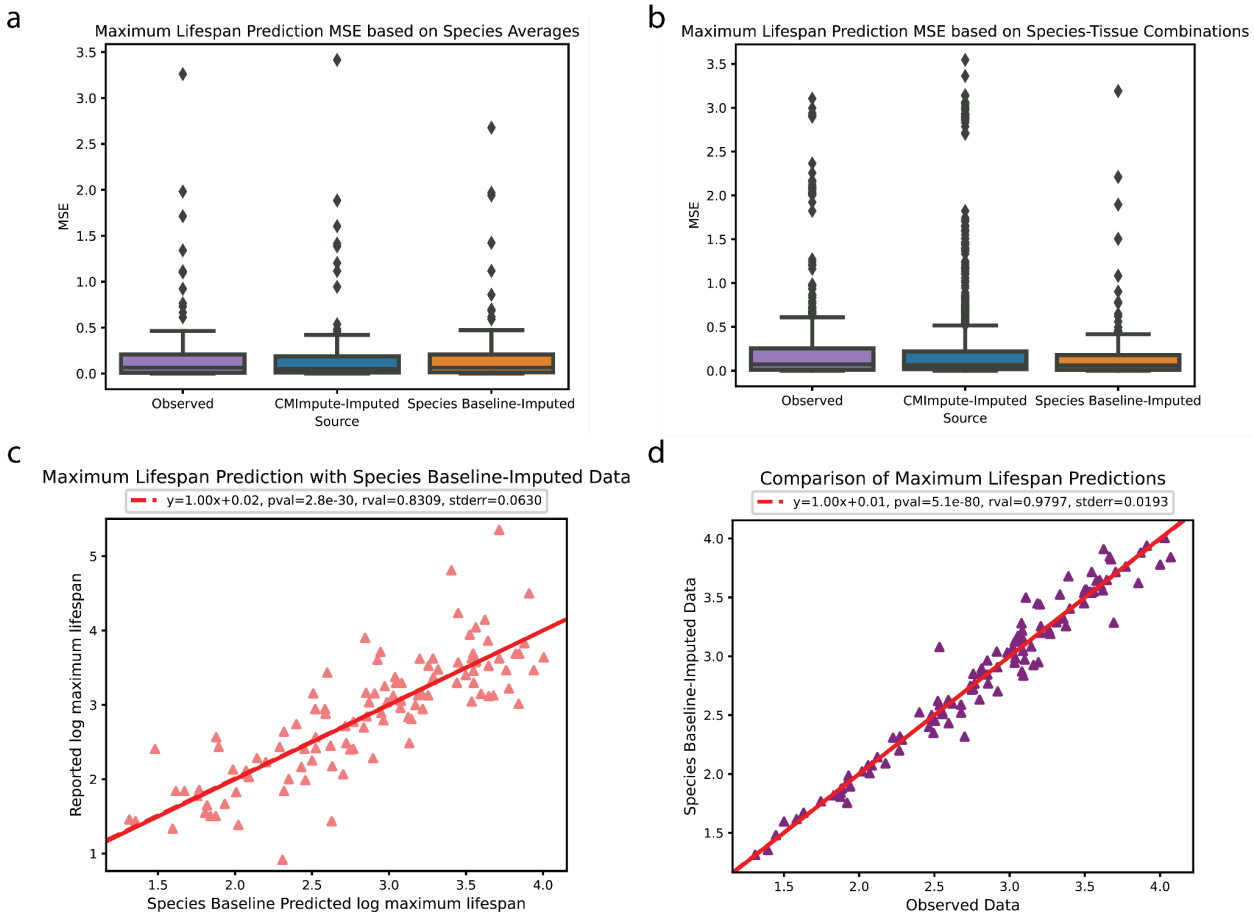


Figure S25. Prediction of species' maximum lifespan using combination mean samples from different tissue types. a) Boxpot comparing the log-maximum lifespan prediction MSE using observed CMImpute-imputed, and species baseline-imputed averages of combination mean samples within each species. Both the imputed and experimentally-profiled combination mean samples span the same 114 species. **b)** Same as a) but comparing the log-maximum lifespan prediction mean squared error using observed, CMImpute, and species baseline species-tissue combination mean samples. The 441 observed combination mean samples span all observed tissues and the 6,285 CMImpute and species baseline-imputed combination mean samples span all non-observed tissues across the 114 species. **c)** Leave-one-species-out linear regression analysis using species-average samples similar to Fig. 8b. Average methylation calculated over

species baseline-imputed species-tissue combination mean samples. Predicted log-maximum lifespan (x-axis) plotted against the reported log-maximum lifespan (y-axis). **d)** Comparison of maximum lifespan predictions based on average species methylation samples between using observed and species baseline-imputed data similar to Fig. 8c.

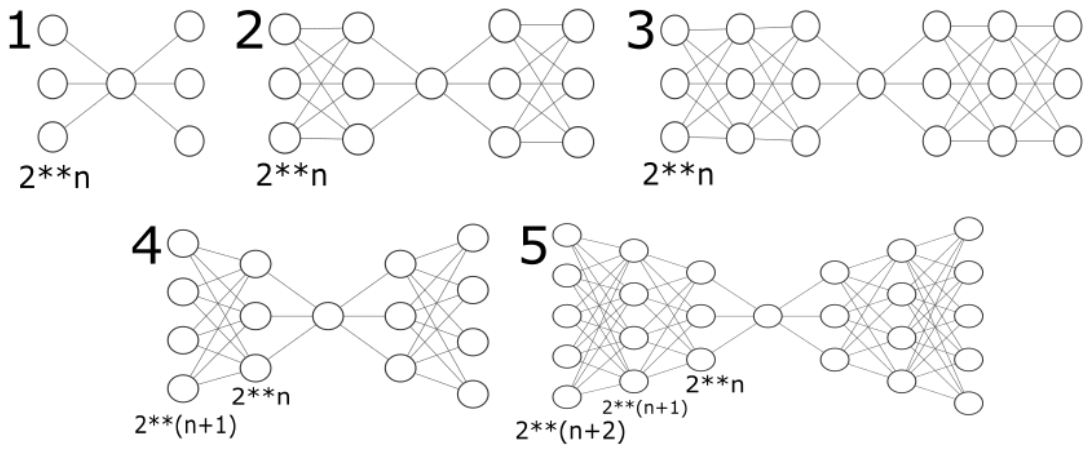


Figure S26. Potential network architectures. Five network architectures that CMImpute considered during the hyperparameter grid search. Each architecture is symmetrical around the latent space with equal numbers of hidden layers in the encoder and decoder. The layout options consist of one through three hidden layers in the encoder and decoder and both tapered (each layer closer to the latent space gets smaller) and equal dimension (each layer in the encoder and decoder is the same dimension) for two and three hidden layer options. Each option (1-5) corresponds to the layouts parameter and n corresponds to the hidden layer dimension parameter (n) from Supplementary Table 2. The final network layout, along with all other hyperparameters, are selected via hyperparameter grid search. Notation of 2**n represents 2 to the power of n.

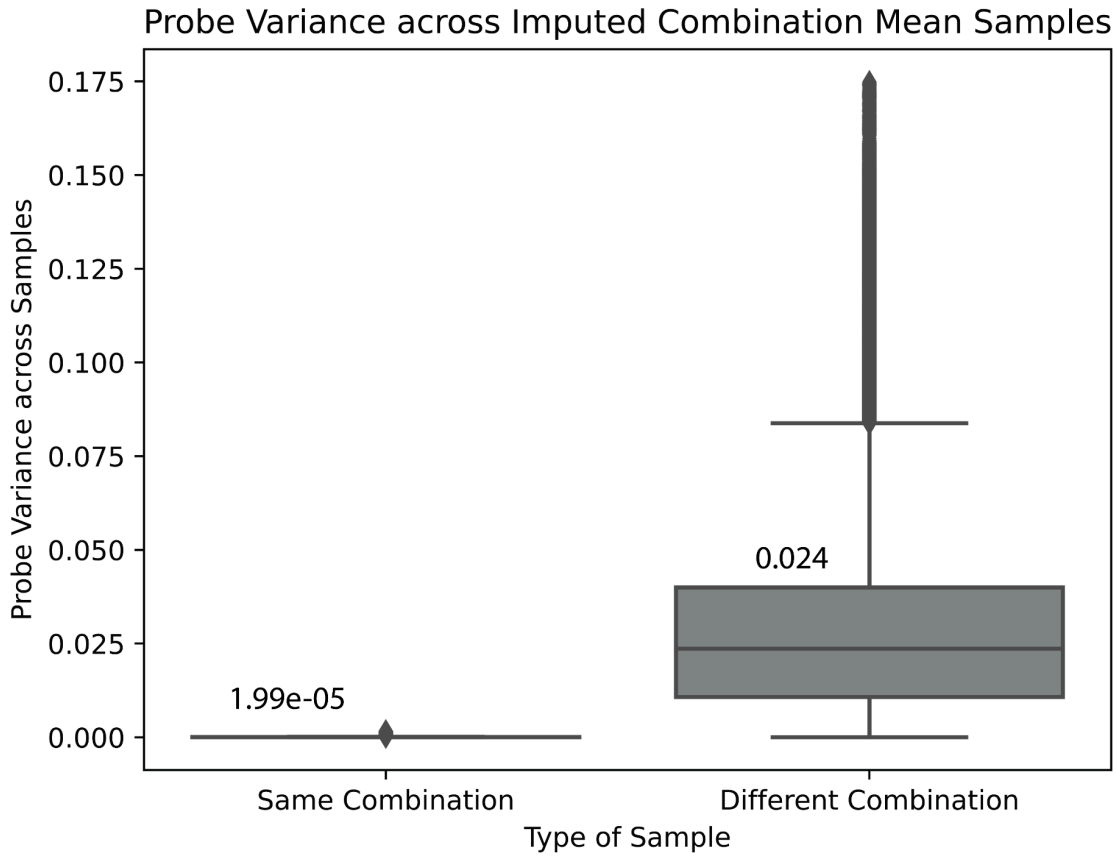


Figure S27. Impact of the random normal latent space sampling on the final imputation result. For a random fold, each held-out combination is imputed 20 times using a different random normal sampling for the latent representation. For each species-tissue combination, the variance across the 20 samples with different random normal latent samplings was calculated. Additionally, the variance across samples from different species-tissue combinations was calculated. The boxplot shows the difference in probe variance between samples of the same combination but different latent samplings and samples of different combinations. Each box is labeled with its median probe variance.

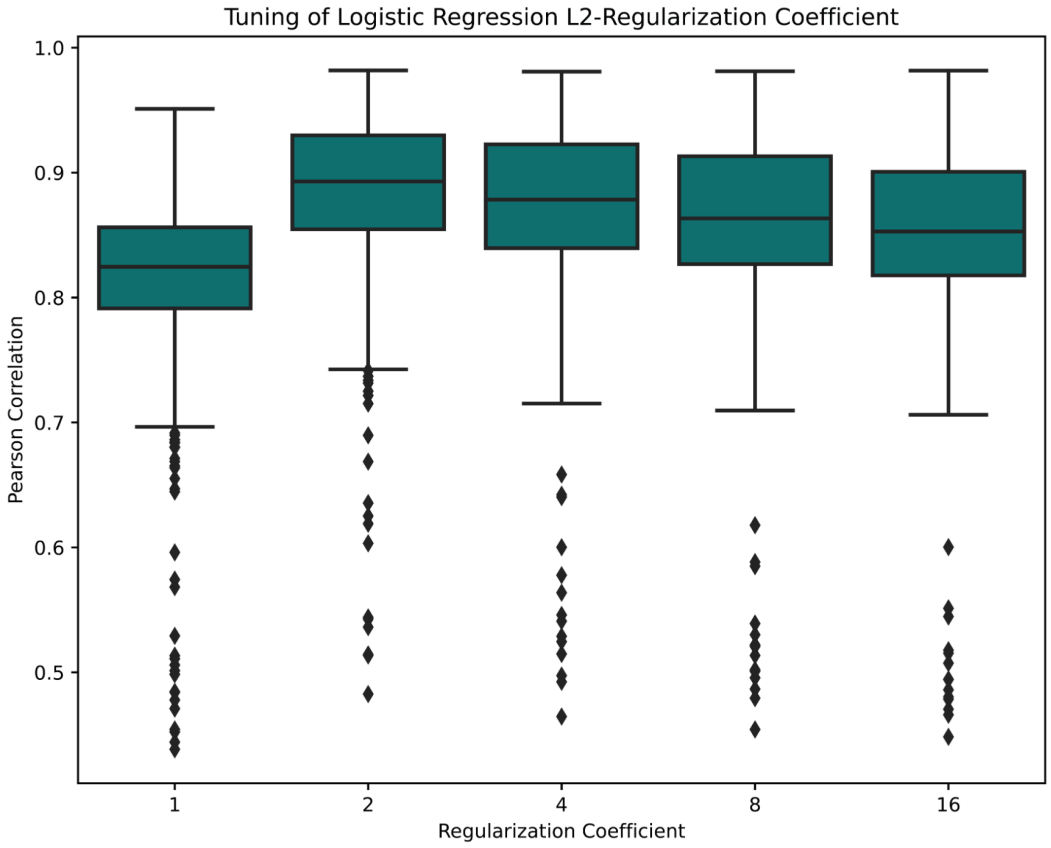


Figure S28. Tuning of the logistic regression baseline L_2-regularization coefficient. For each coefficient, a logistic regression model was trained for each probe on each of the five cross-validation folds (Methods). The boxplot shows the validation sample-wise Pearson correlation with held-out observed values for regression coefficients of 1, 2, 4, 8, and 16.