

Additional file 3

Fig S1: exSTRa plots of EGA and simulated genomes. a. Empirical distribution function (ECDF; left) and t-sum (right) plots of Isaac-aligned data. In the ECDF plots, samples with a known expansion at the tested STR locus were colored red and the rest of the samples were colored black. The t-sum plots show the samples identified by exSTRa to harbor a significant expansion at the given locus. The title on top of each plot indicates the STR locus, the number of repeats in the hg19 reference genome, and the pathogenic lower bound number of repeats of the STR. **b.** ECDF (left) and t-sum (right) plots of BWA-aligned data.

Fig S1

a

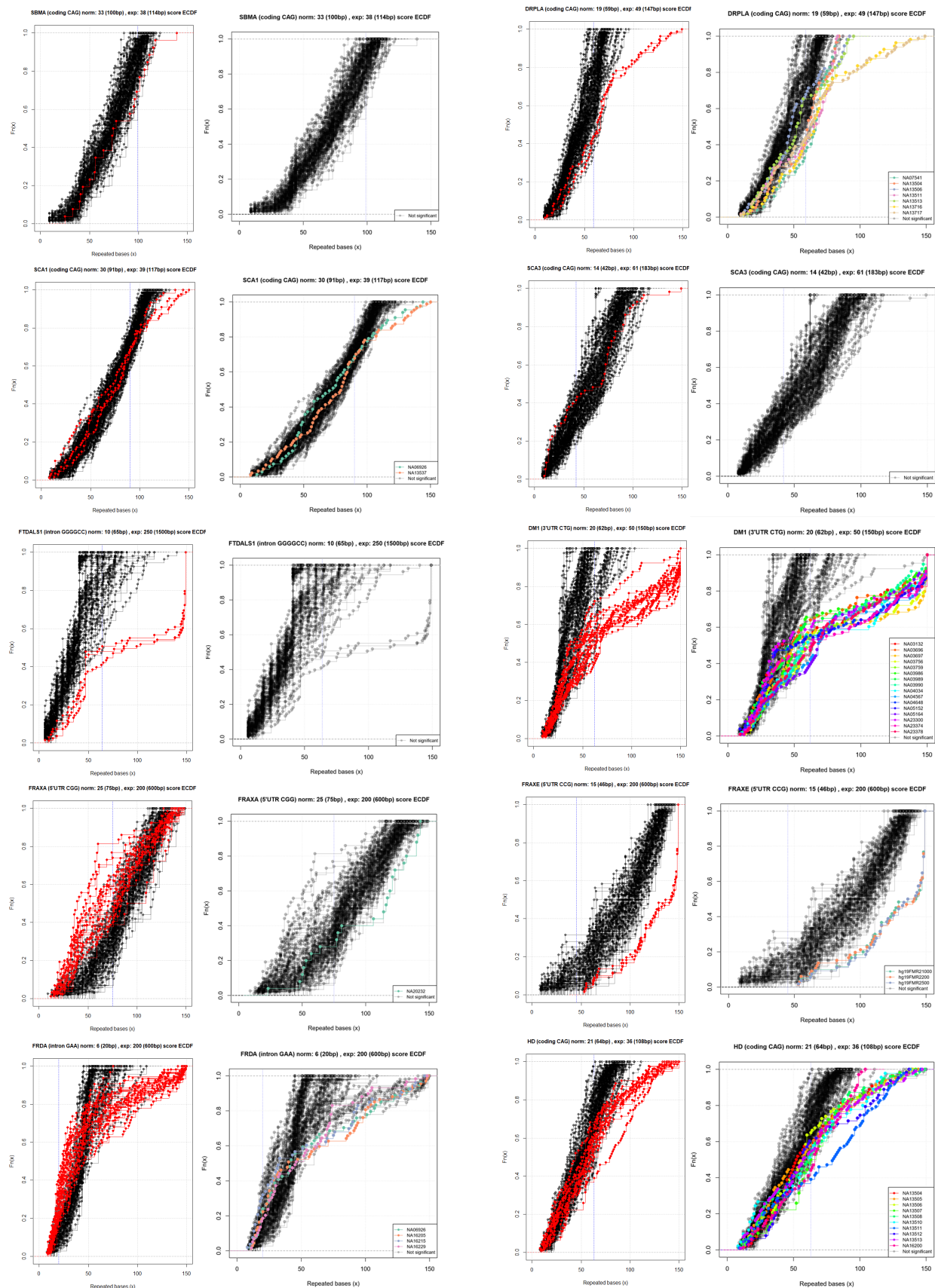


Fig S1

b

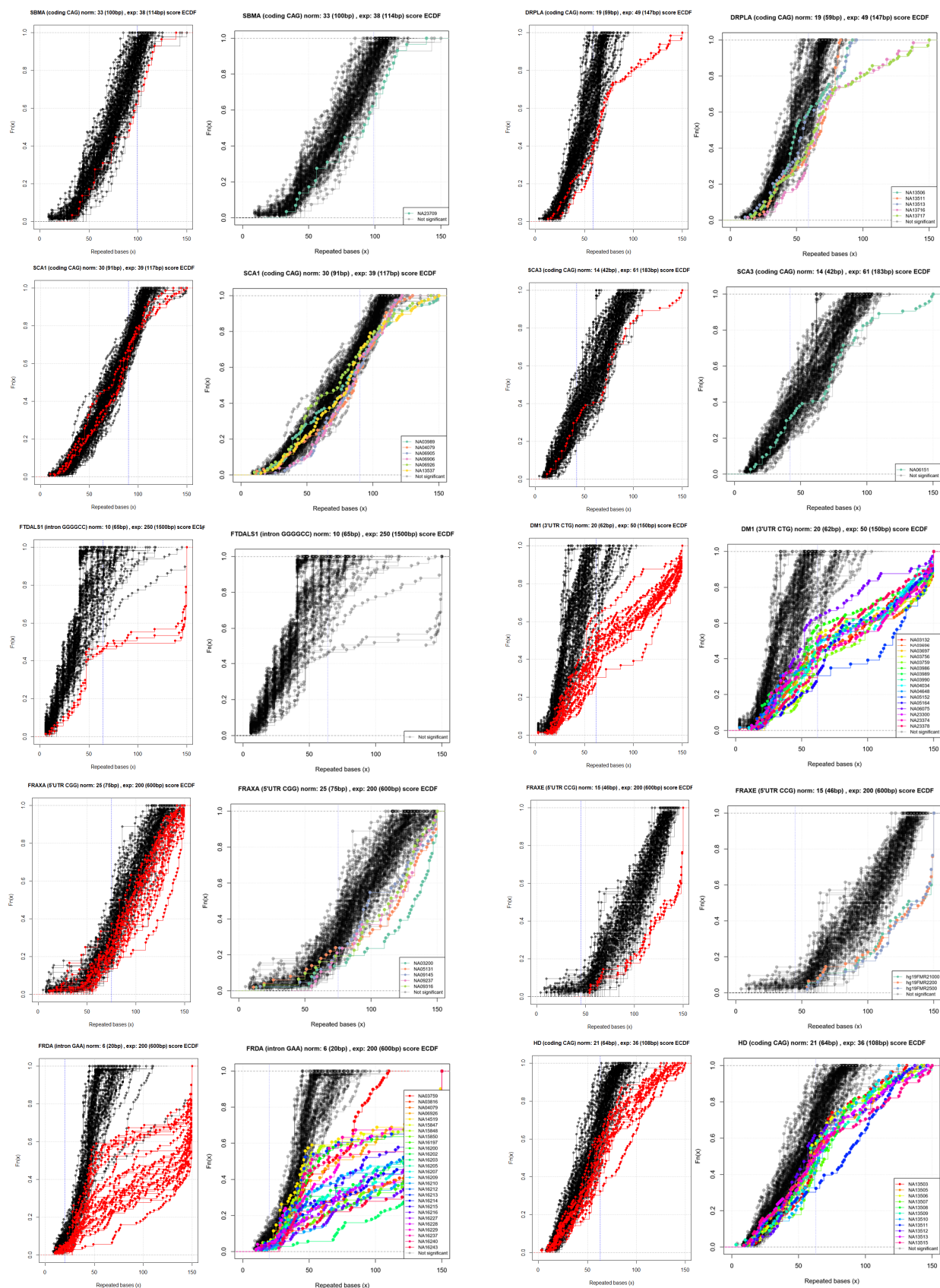


Fig S2: Decision tree model of the default analysis of Isaac-aligned EGA genomes on the training dataset. Decision tree generated by the classifier on the training dataset (n=942). Node #0 at the top of the tree is the root node. Each node lists an STR tool (feature). The “samples” number represents the total number of data points present within a particular node, and “value” shows the number of expanded (full-mutation, FM) and non-expanded (non-FM) data points. The shade of the colour of each node reflects the proportion of expanded to non-expanded data points, with deeper orange and blue meaning more non-expanded and expanded data points, respectively. Gini index shows the impurity at each node. The terminal nodes or leaves with a Gini value of 0 have data points belonging entirely to either the expanded or the non-expanded class. EHv2: ExpansionHunter version 2; Expanded and Not_Expanded imply the presence and absence of FM, respectively.

Fig S2

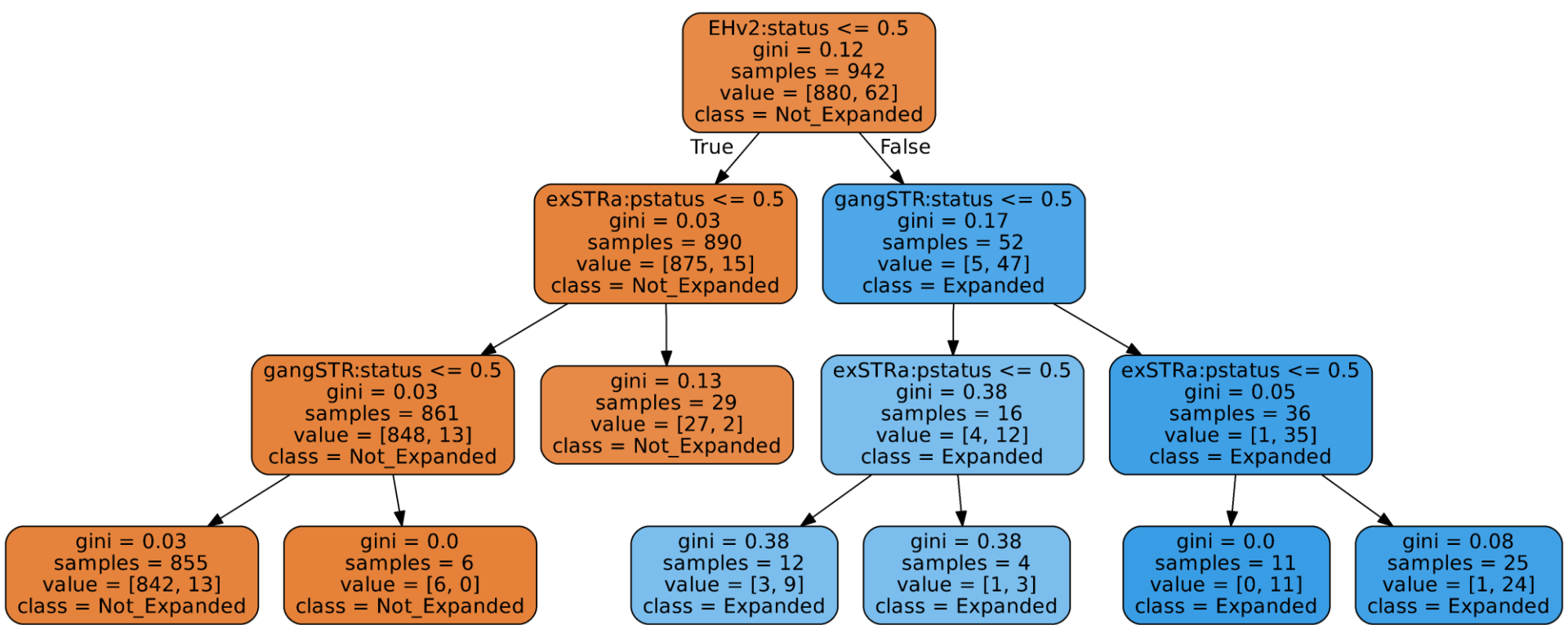


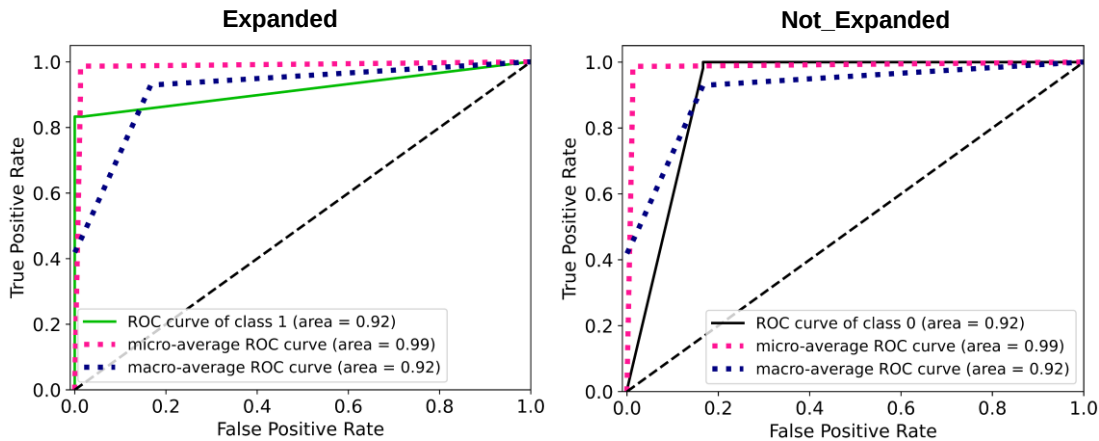
Fig S3: Performance metrics of the decision tree model in the default analysis of Isaac-aligned EGA genomes on the test dataset. **a.** Table showing the number of expanded (class 1) and non-expanded (class 0) STR alleles (column “Support”) present in the test dataset and several metrics, including precision, recall, and F1-score in determining expanded and non-expanded alleles, accuracy, and macro-, and weighted-averages of these metrics. Macro average (Macro avg): unweighted mean of the performance metrics calculated for the Expanded and Not_Expanded class labels; and Weighted average (Weighted avg): weighted mean of the performance metrics calculated for the Expanded and Not_Expanded class labels (accounts for label imbalance). **b.** Receiver Operating Characteristics (ROC) curve with False-Positive Rate on x-axis and True-Positive Rate on y-axis showing the performance of the decision tree model on the test dataset (n=236) and Precision-recall curve showing recall on x-axis and precision on y-axis for the test dataset. **c.** Confusion matrix showing the number of expanded and non-expanded alleles along the blue-shaded diagonal that were rightly-classified and three expanded alleles that were misclassified as non-expanded. **d.** Feature importances showing STR tools and their normalized (Gini) importance on x- and y-axis, respectively.

Fig S3

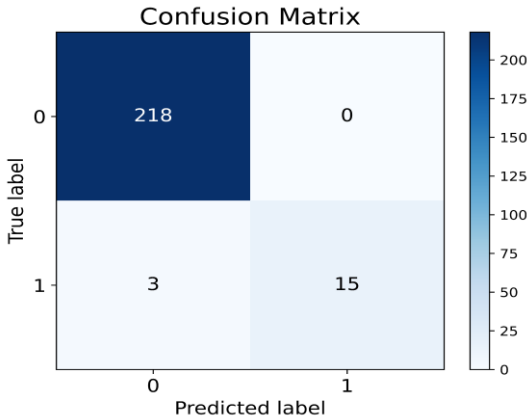
a

Class	Precision	Recall	F1-score	Support
1 (Expanded)	1.00	0.83	0.91	18
0 (Not_Expanded)	0.99	1.00	0.99	218
Accuracy			0.99	236
Macro avg	0.99	0.92	0.95	236
Weighted avg	0.99	0.99	0.99	236

b



c



d

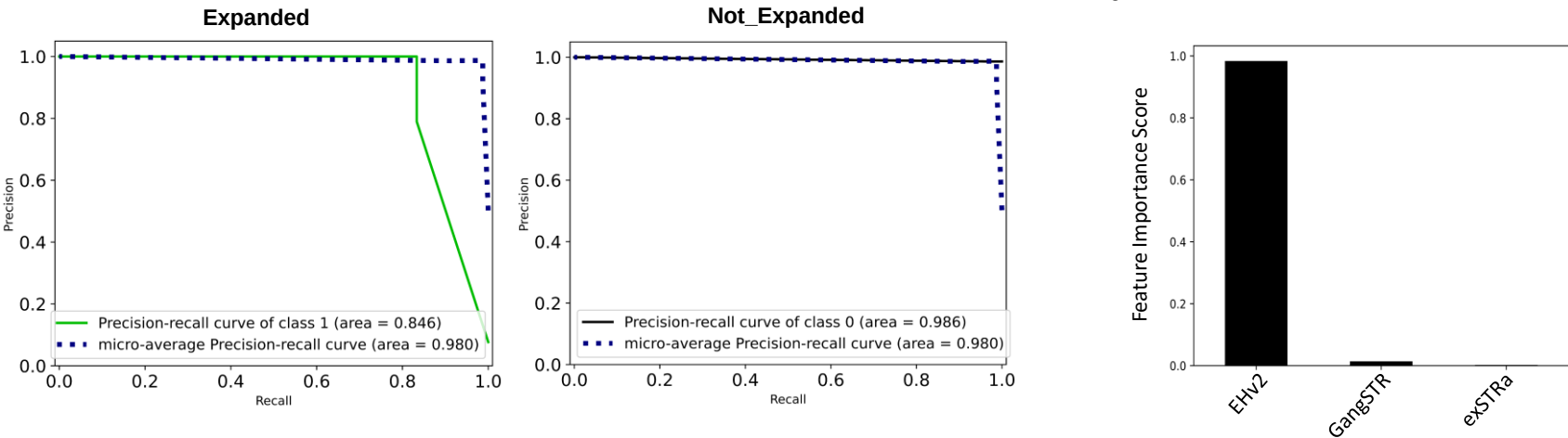


Fig S4: Decision tree model of the default analysis of BWA-aligned EGA genomes on the training dataset. Decision tree generated ~~by the classifier~~ on the training dataset (n=940). EHv3: ExpansionHunter version 3; TRED: TREDPARSE; Expanded and Not_Expanded imply the presence and absence of full-mutation, respectively.

Fig S4

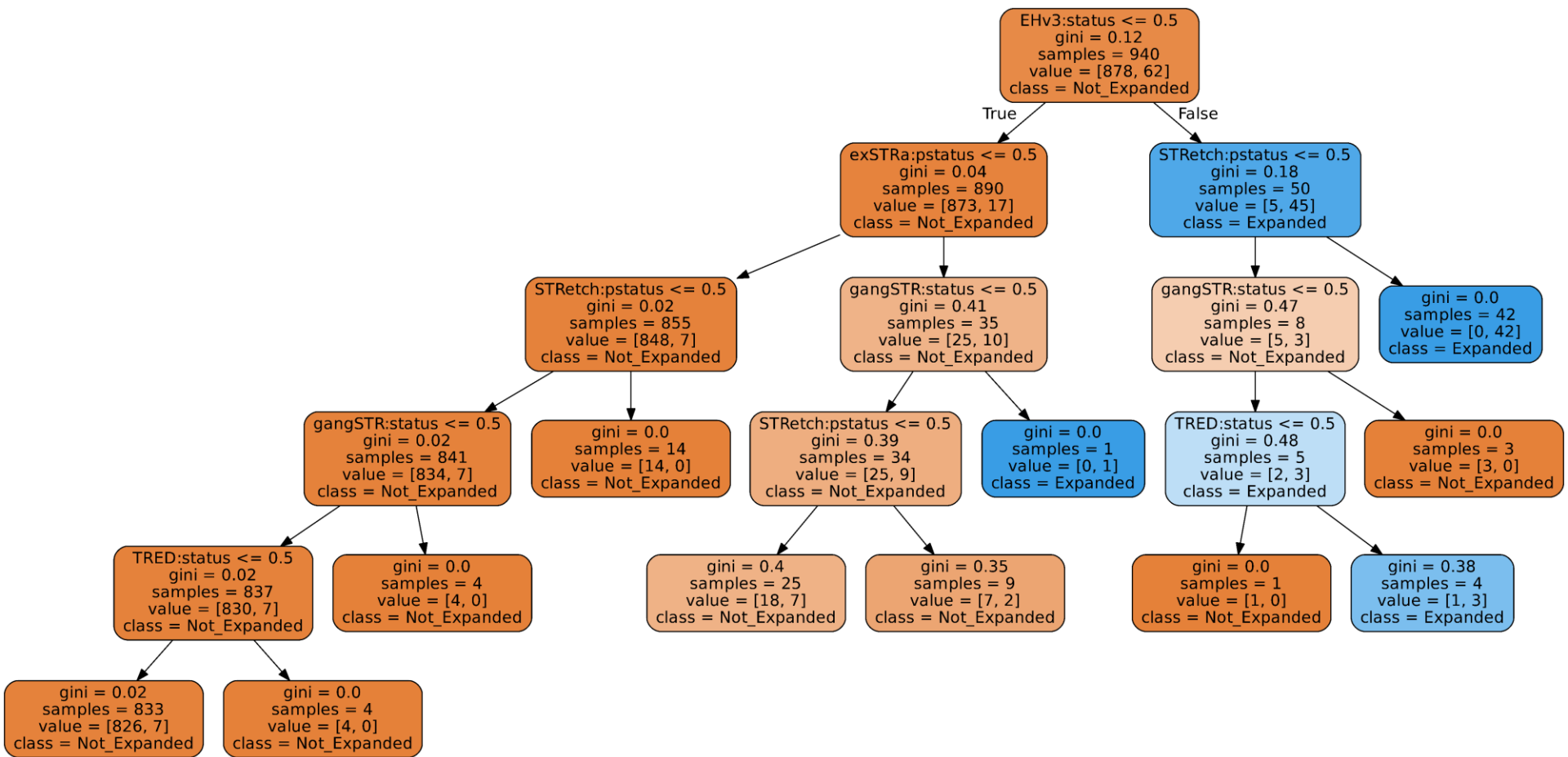


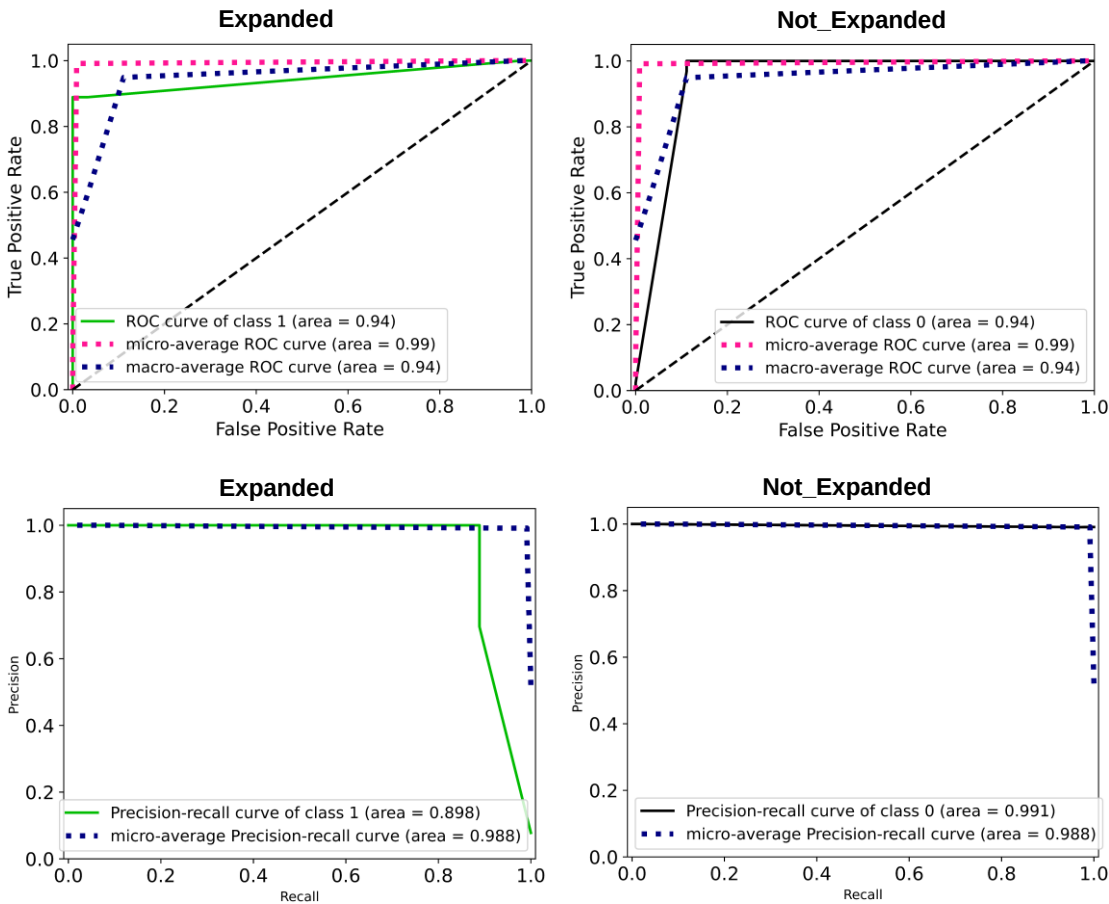
Fig S5: Performance metrics of the decision tree model in the default analysis of BWA-aligned EGA genomes on the test dataset. **a.** Table showing the number of expanded (class 1) and non-expanded (class 0) STR alleles (column “Support”) present in the test dataset and several metrics, including precision, recall, and F1-score in determining expanded and non-expanded alleles, accuracy, and macro-, and weighted-averages of these metrics. **b.** Receiver Operating Characteristics (ROC)-curve with False-Positive Rate on x-axis and True-Positive Rate on y-axis showing the performance of the decision tree model on the test dataset (n=236) and Precision-recall curve showing recall on x-axis and precision on y-axis for the test dataset. **c.** Confusion matrix showing the number of expanded and non-expanded alleles along the blue-shaded diagonal that were rightly-classified and two expanded alleles that were misclassified as non-expanded. **d.** Feature importances showing STR tools and their normalized (Gini) importance on x- and y-axis, respectively.

Fig S5

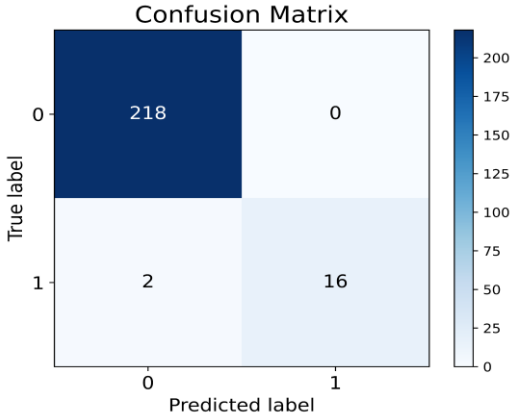
a

Class	Precision	Recall	F1-score	Support
1 (Expanded)	1.00	0.89	0.94	18
0 (Not_Expanded)	0.99	1.00	1.00	218
Accuracy			0.99	236
Macro avg	1.00	0.94	0.97	236
Weighted avg	0.99	0.99	0.99	236

b



c



d

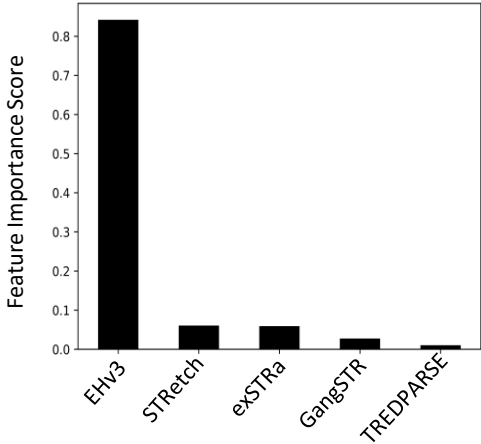


Fig S6: exSTRa plots of EGA genomes analysed with 100 controls. a. Isaac alignments. **b.** BWA alignments. ECDF (left) and t-sum (right) plots are shown.

Fig S6

a

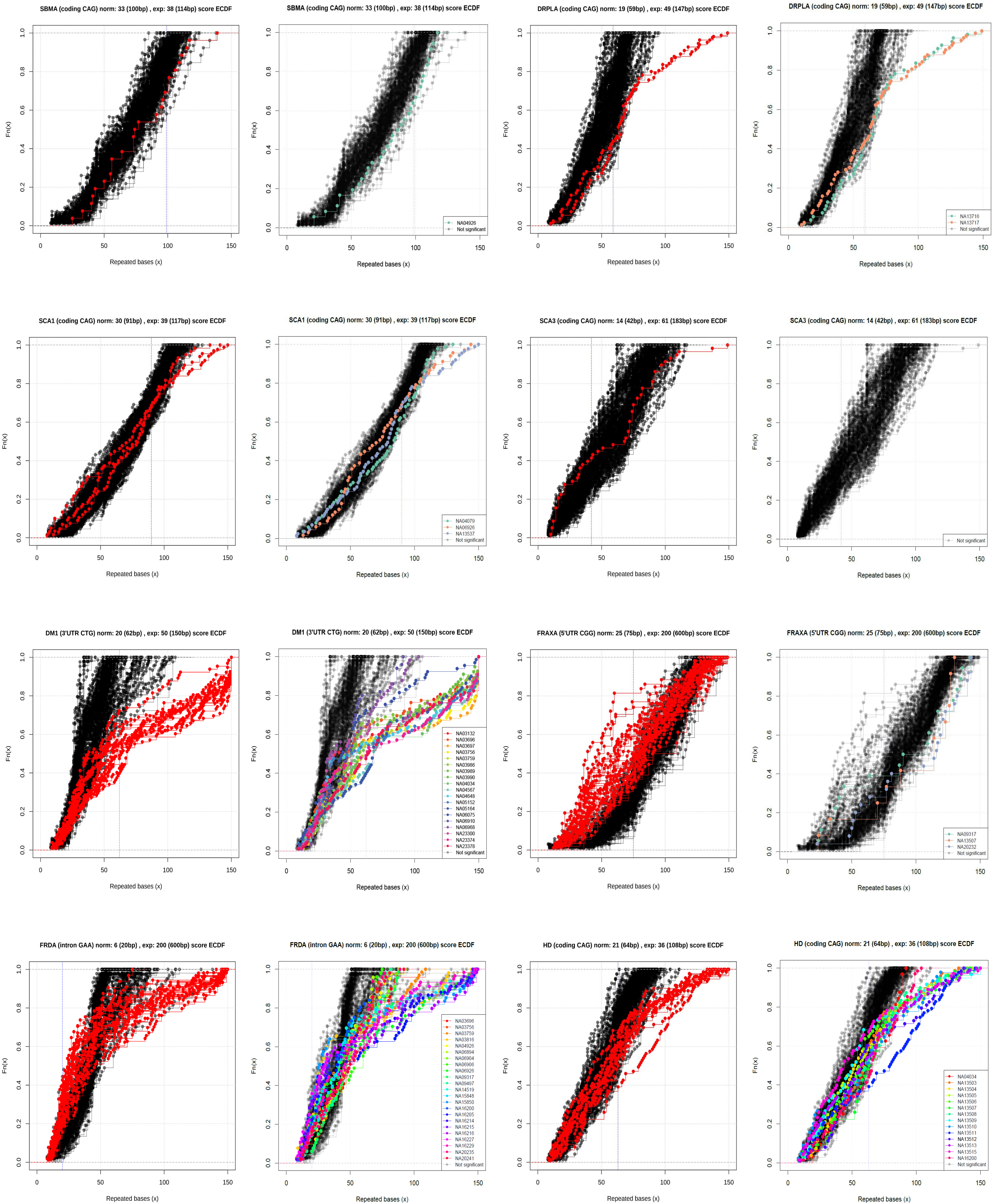


Fig S6

b

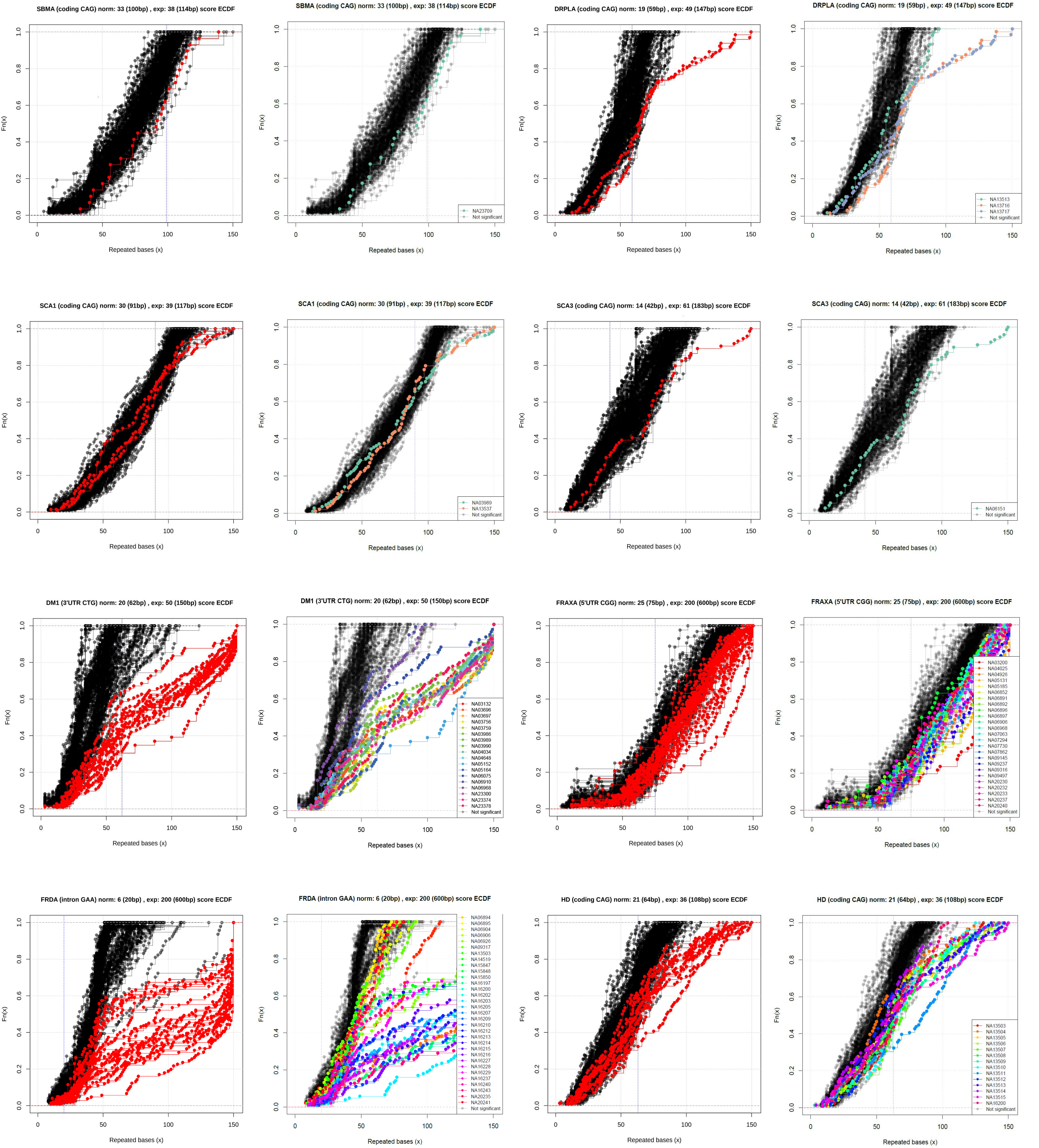


Fig S7: Decision tree model of the modified analysis of Isaac-aligned EGA genomes on the training dataset. Decision tree generated ~~by the classifier~~ on the training dataset (n=942). EHv3: ExpansionHunter version 3; TRED: TREDPARSE; Expanded and Not_Expanded imply the presence and absence of full-mutation, respectively.

Fig S7

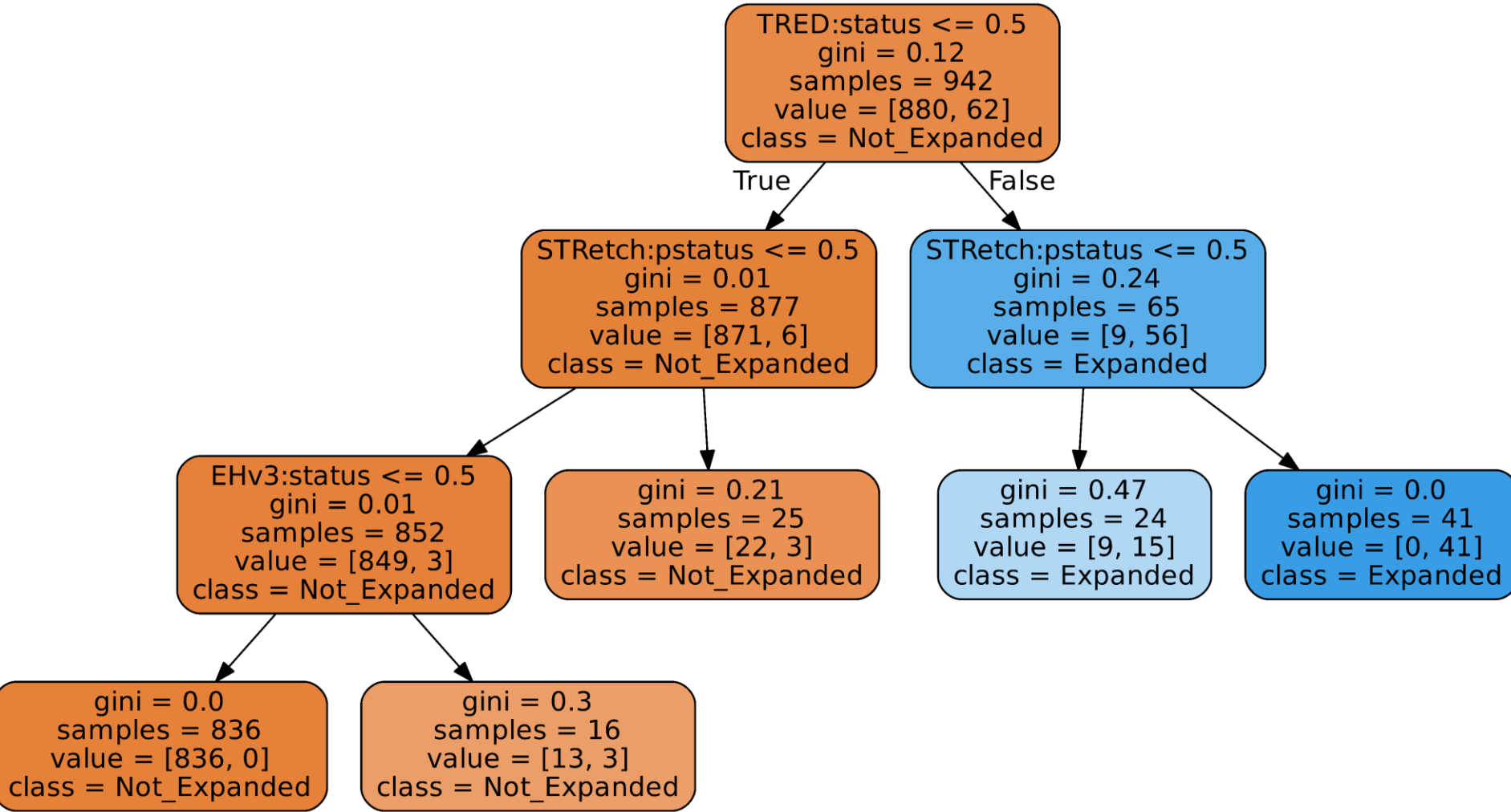


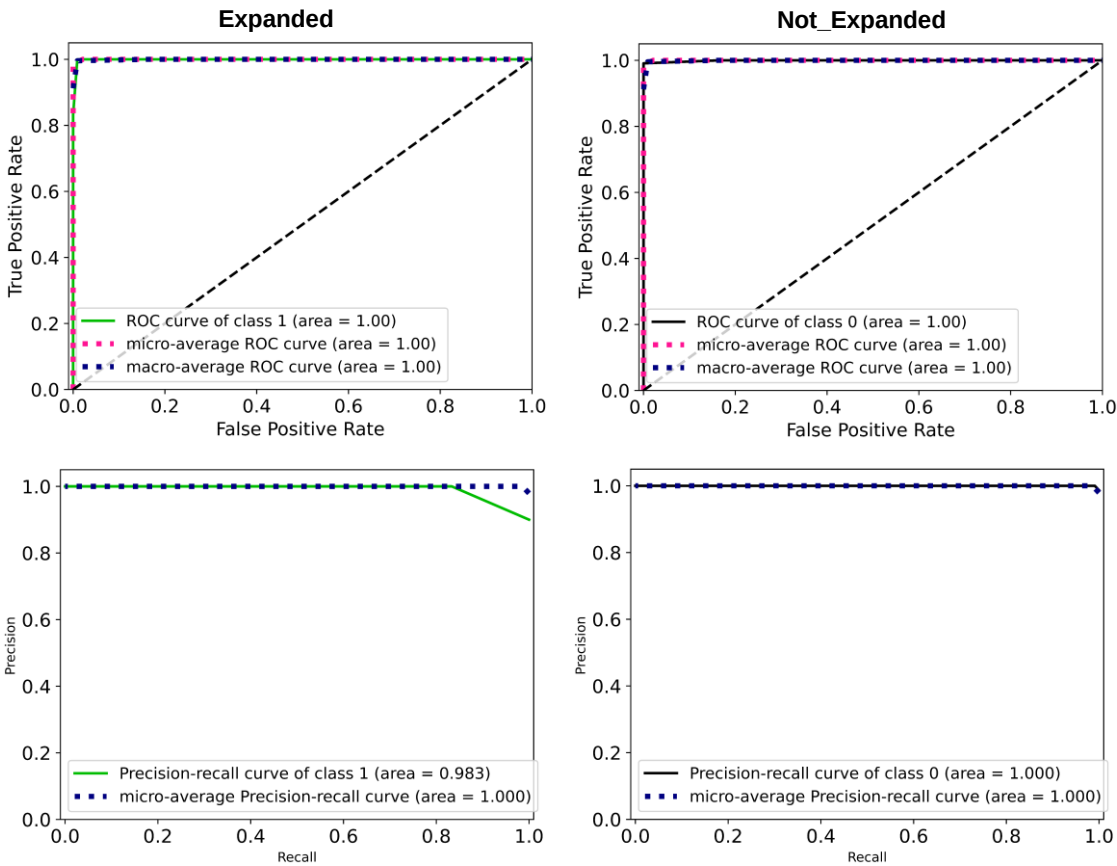
Fig S8: Performance metrics of decision tree model in modified analysis of Isaac-aligned EGA test dataset. a. Classification report showing the number of expanded (class 1) and non-expanded (class 0) STR alleles (column “Support”) present in the test dataset and several metrics, including precision, recall, and F1-score in determining expanded and non-expanded alleles, accuracy, and macro-, and weighted-averages of these metrics. **b.** Receiver Operating Characteristics (ROC) curve with False-Positive Rate on x-axis and True-Positive Rate on y-axis showing the performance of decision tree model on test dataset (n=236) and Precision-recall curve showing recall on x-axis and precision on y-axis for test dataset. **c.** Confusion matrix showing the number of expanded and non-expanded alleles along the blue-shaded diagonal that were rightly-classified and two non-expanded alleles that were misclassified as expanded. **d.** Feature importances showing STR tools and their normalized (Gini) importance on x- and y-axis, respectively.

Fig S8

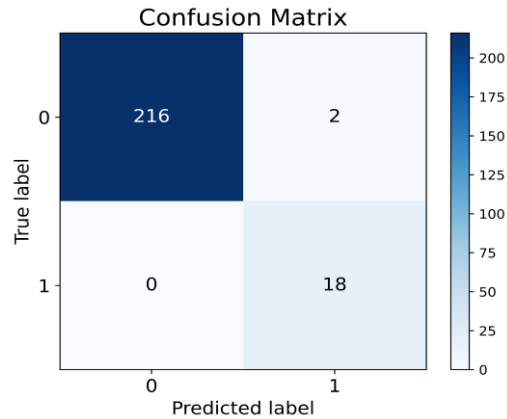
a

Class	Precision	Recall	F1-score	Support
1 (Expanded)	0.90	1.00	0.95	18
0 (Not_Expanded)	1.00	0.99	1.00	218
Accuracy			0.99	236
Macro avg	0.95	1.00	0.97	236
Weighted avg	0.99	0.99	0.99	236

b



c



d

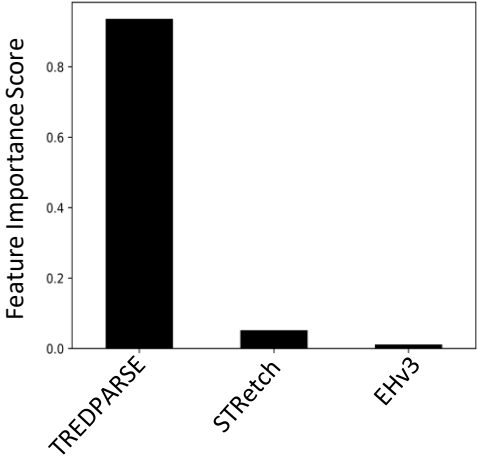
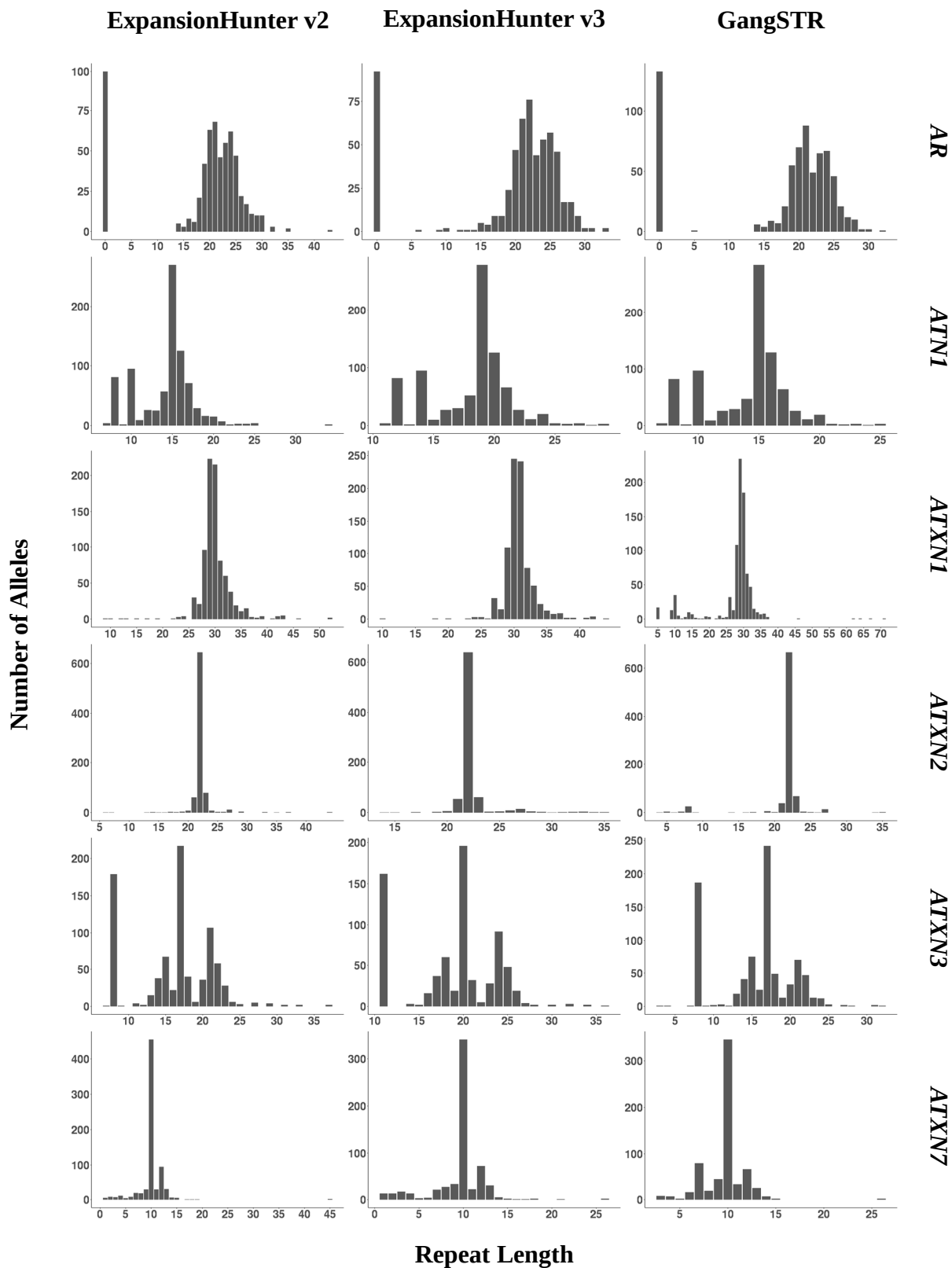
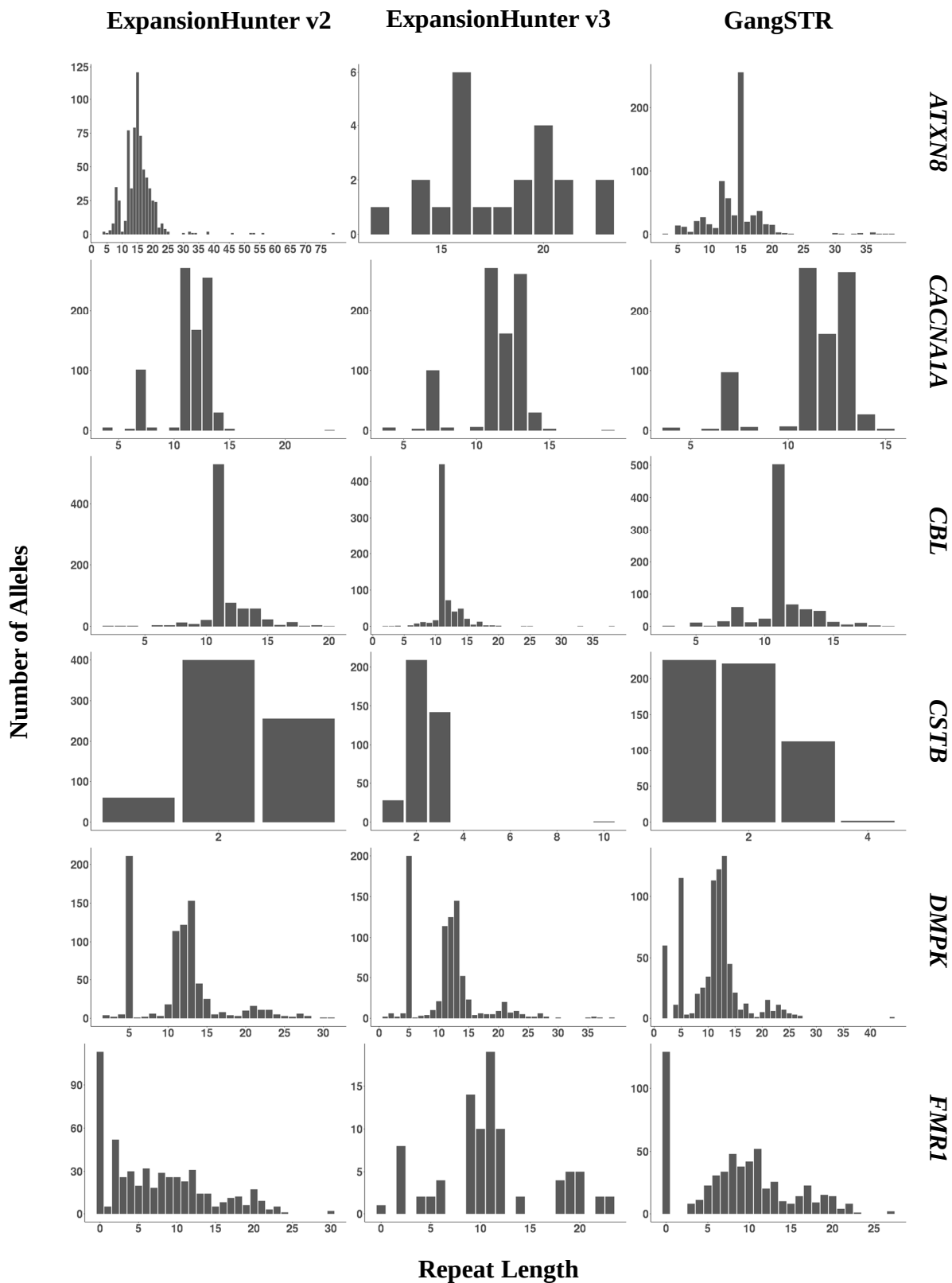
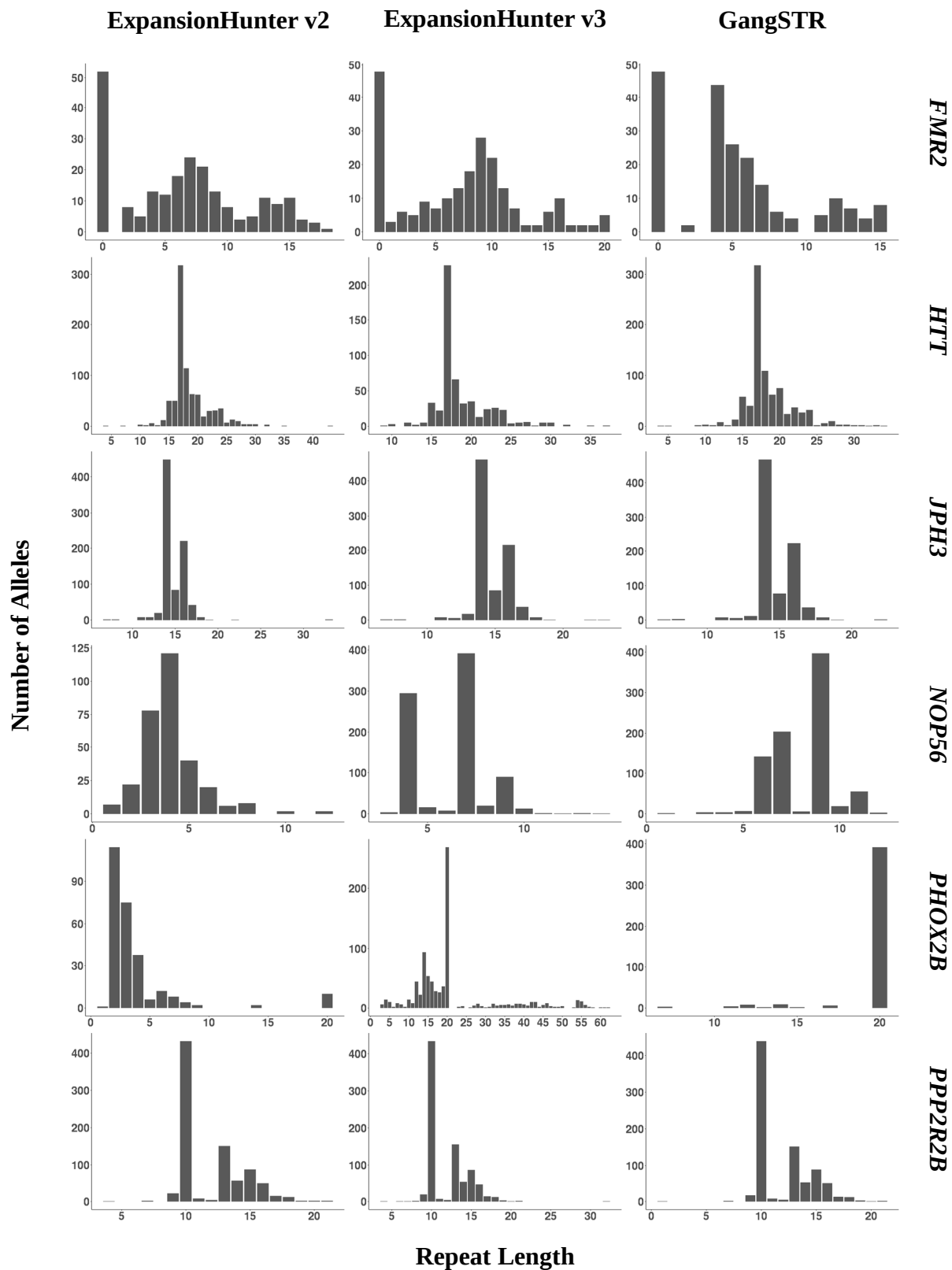


Fig S9: Allele frequency distribution of analysed disease short tandem repeat loci in the CAUSES exomes. Genotype calls of ExpansionHunter versions 2 and 3 and GangSTR were used to generate the distribution barplots. The x-axis shows the repeat count of the genotyped alleles and the y-axis shows the frequency i.e., the number of alleles with a particular repeat count. The 0 repeat alleles in the plots for X chromosome short tandem repeats (*AR*, *FMR1*, and *FMR2*) are that of hemizygous males in whom the second allele's repeat length was set to 0 and do not reflect the actual genotype calls of the tools.

Fig S9







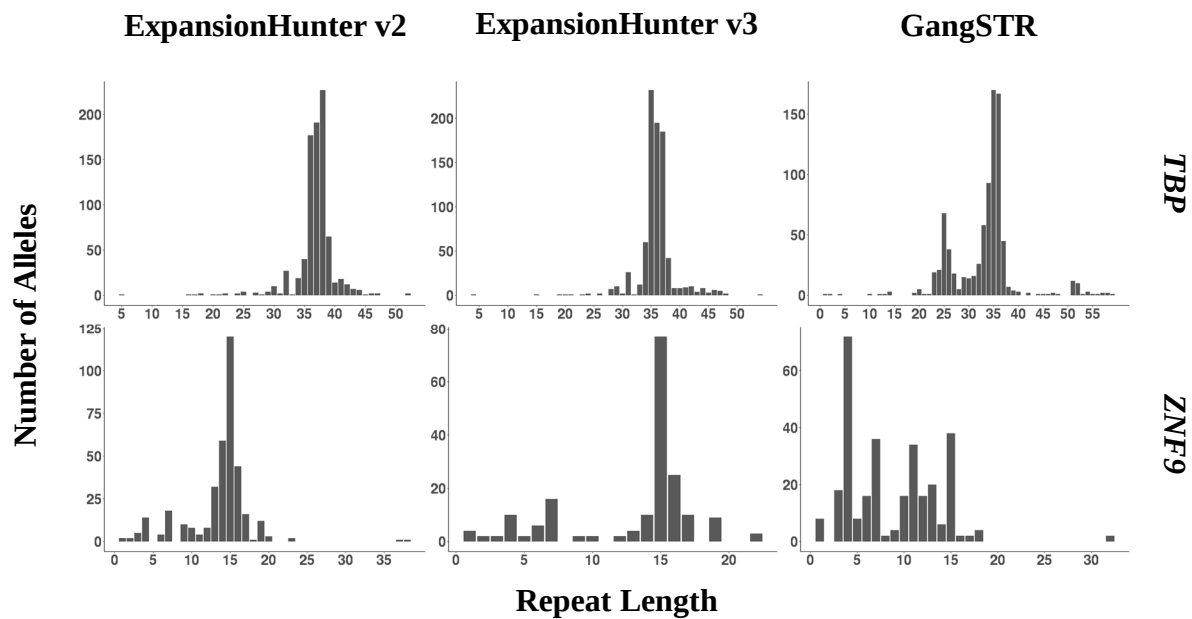
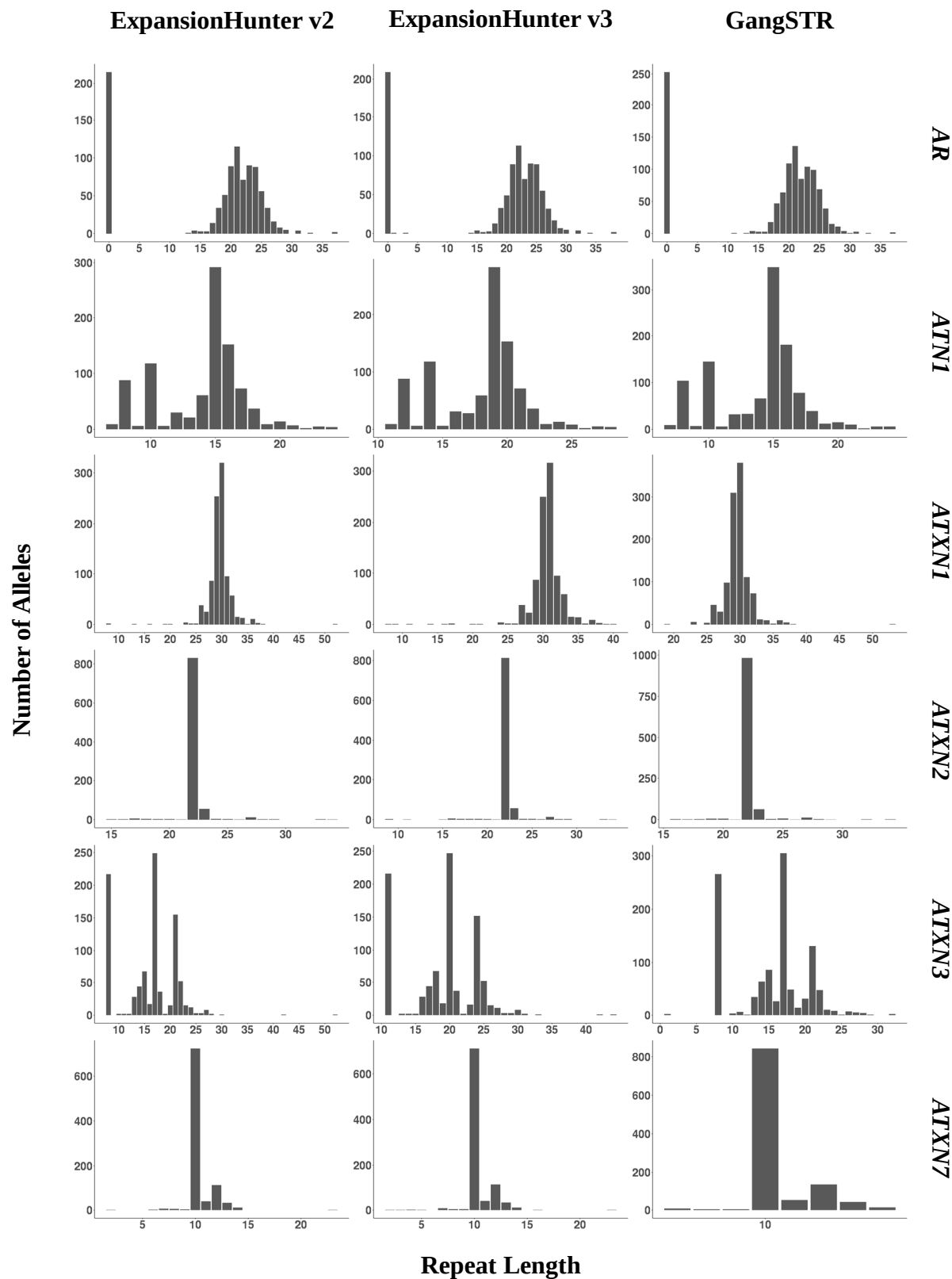
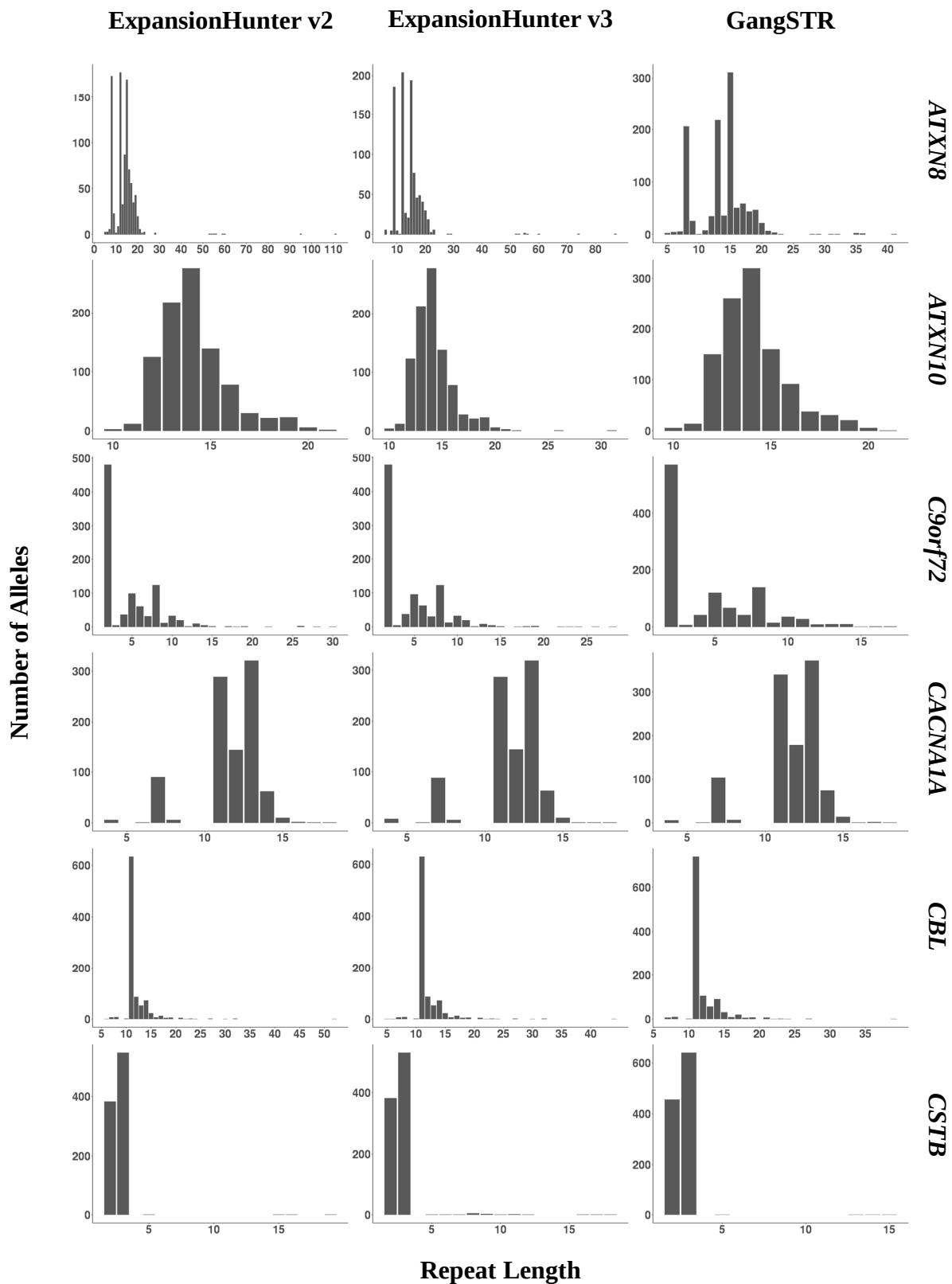
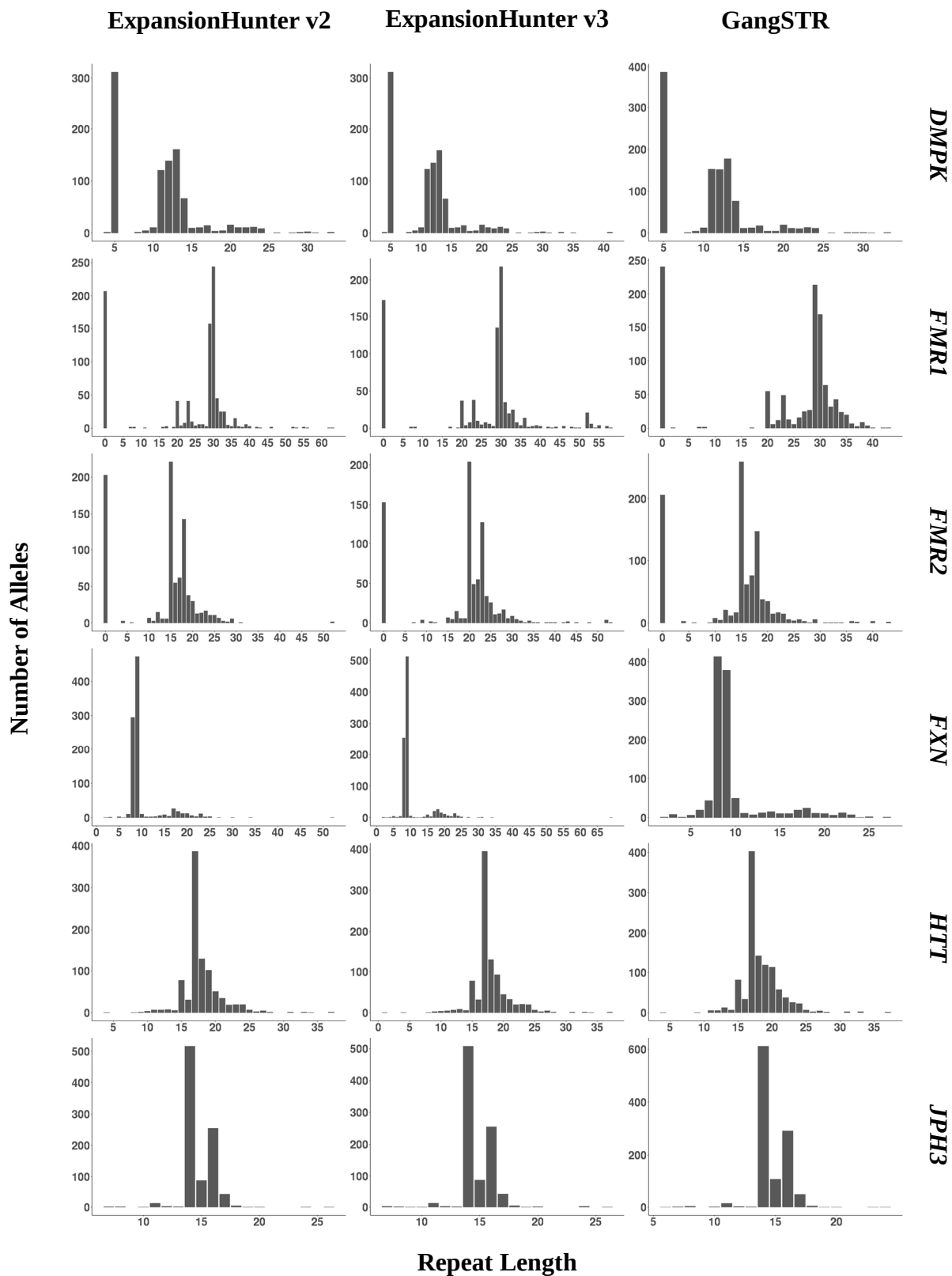


Fig S10: Allele frequency distribution of analysed disease short tandem repeat loci in the CAUSES and IMAGINE genomes.

Fig S10







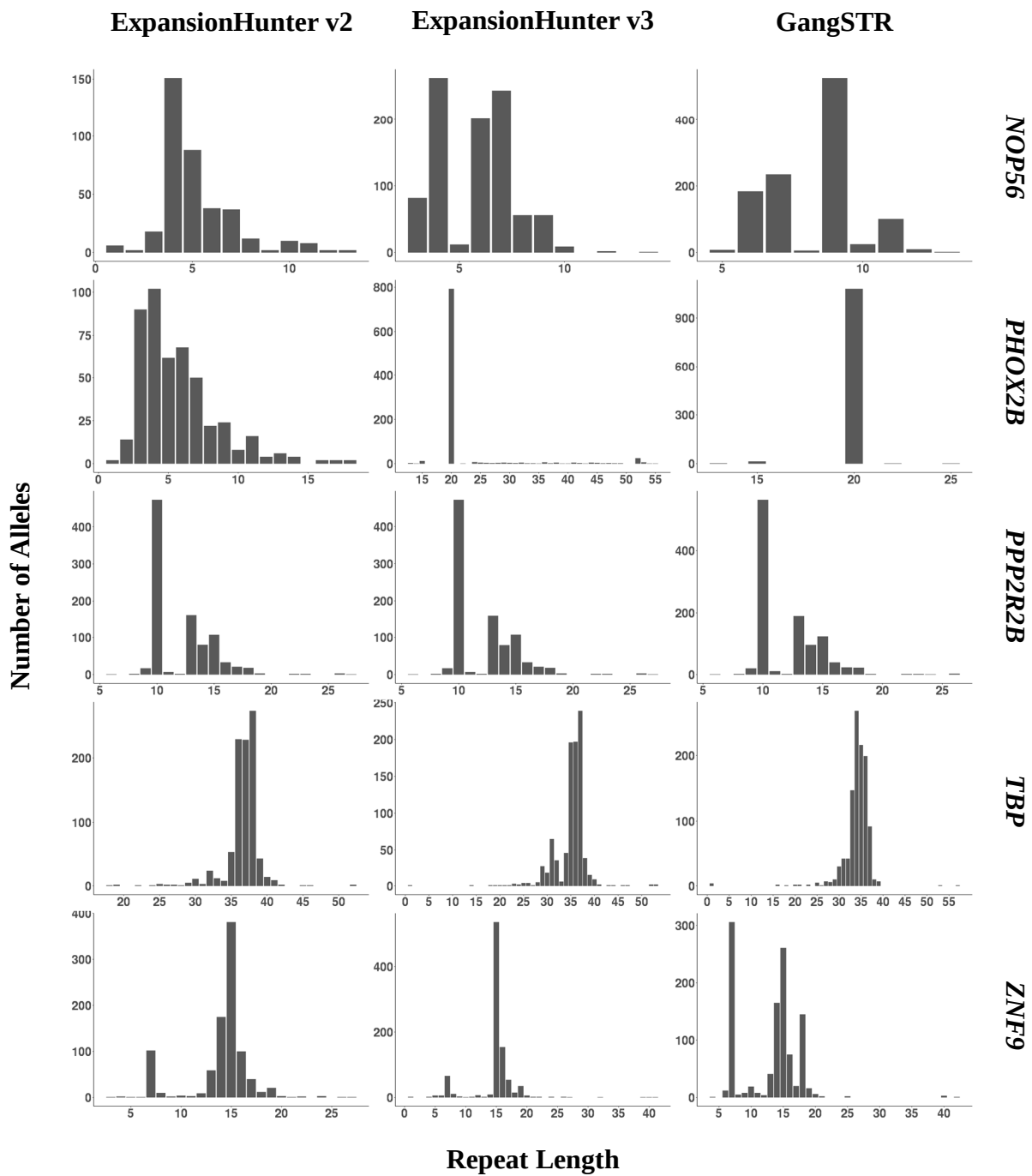


Fig S12: Analysis of the *DMPK* locus by ExpansionHunter version 2 with and without off-target sites in the EGA dataset. Results of Isaac- (top panels) and BWA-aligned (bottom panels) genomes. Scatter plots on the left show the estimated repeat lengths of the *DMPK* locus in the EGA genomes by ExpansionHunter version 2 with off-target (OT) sites (wOT) and without OT (nOT) sites alongside the consensus repeat size of known *DMPK* expansion-positive samples. The horizontal red-dotted line define the lower-bound repeat lengths of *DMPK* full-mutations (50 repeats). Plots on the right show the Pearson's correlation between estimated and consensus repeat lengths of full-mutation alleles wOT and nOT.

