**Variance analysis**

Gradient Boosted Tree: Boosting is the process of building a large, additive decision tree by fitting a sequence of smaller decision trees, called layers. The tree at each layer consists of a small number of splits. The tree is fit based on the residuals of the previous layers, which allows each layer to correct the fit for bad fitting data from the previous layers. The final prediction for an observation is the sum of the predictions for that observation over all of the layers. The column contributions were estimated in the model fitting, which is based on the total number of instances over all of the trees when the specific column is used to split the data. The proportion of the contribution of each factor was calculated as sum of squares attributed to the factor divided by the total sum of squares. In the model fitting, 70% of the data were used as training set while 30% were used as testing set. The key parameters, the splits per tree, number of layers, and learning rate, were set as 5, 100, and 0.1 respectively.

Boosted Tree JMP code:

Boosted Tree (

Y( :concordance ),

X( :aligner, :GATK, :caller, :individual, :location ),

Validation Portion( 0.3 ),

Method( "Boosted Tree" ),

Plot Actual by Predicted( 1 ),

Column Contributions( 1 ),

Splits per Tree( 5 ),

Number of Layers( 100 ),

Learning Rate( 0.1 ),

Go

)

Variance Components Analysis: Variance component analysis is to obtain estimates of the contributions that different factors make to the overall variability of the data, as expressed by their variance. These contributions are called variance components. Variance components were estimated through fitting a random effect model as following:

,

where denotes an vector of response (Jaccard index), is the design matrix for the random effects, is a vector of unknown random effects with design matrix . Both and are assumed following normal distribution with means at 0. and are the variance components that need to be estimated. The ratio of the contribution of each factor to the overall variability was calculated by the variance component of each factor divided by the total.

Random effect model JMP code:

Fit Model(

Y( :concordance ),

Effects,

Random Effects(

:aligner \* :caller,

:GATK \* :caller,

:aligner \* :GATK,

:aligner \* :location,

:caller \* :location,

:GATK \* :location,

:individual \* :location,

:aligner,

:GATK,

:caller,

:individual,

:location

),

NoBounds( 1 ),

Personality( "Standard Least Squares" ),

Method( "REML" ),

Emphasis( "Effect Leverage" ),

Run(

:concordance << {Summary of Fit( 1 ), Analysis of Variance( 0 ),

Parameter Estimates( 1 ), Scaled Estimates( 0 ),

Plot Actual by Predicted( 1 ), Plot Regression( 0 ),

Plot Residual by Predicted( 1 ), Plot Studentized Residuals( 0 ),

Plot Effect Leverage( 1 ), Plot Residual by Normal Quantiles( 0 )}

)

);