機械学習による低品質バリアント領域を推定する指標(UNMETスコア)の導出とその評価

長浜バイオ大学
土方敦司

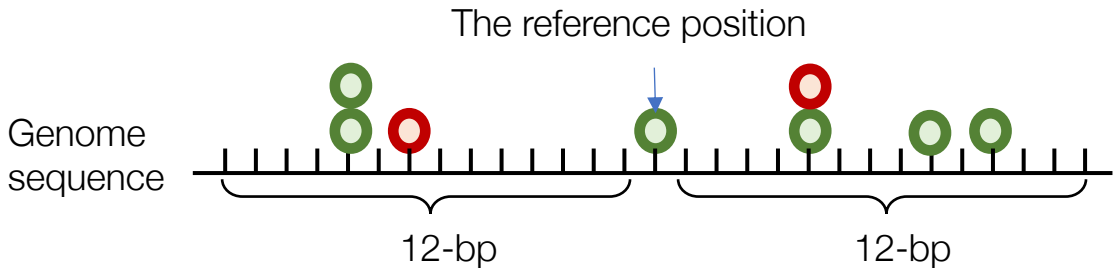
NGSでバリアント解析が困難なゲノム領域はどこか？

- 高品質のバリアントが検出できない領域
- ショートリードNGSで解析が難しいとされる領域

  ＝マッピングエラー・シークエンスエラーの起きやすい領域
  - ゲノム重複
  - リピート配列（単調リピート、タンデムリピート）
  - 低複雑度配列（Low complexity region）
  - コピー数多型（CNV）
  - 構造多型（Structural variation）
- gnomADのバリアントデータから特定できるか？

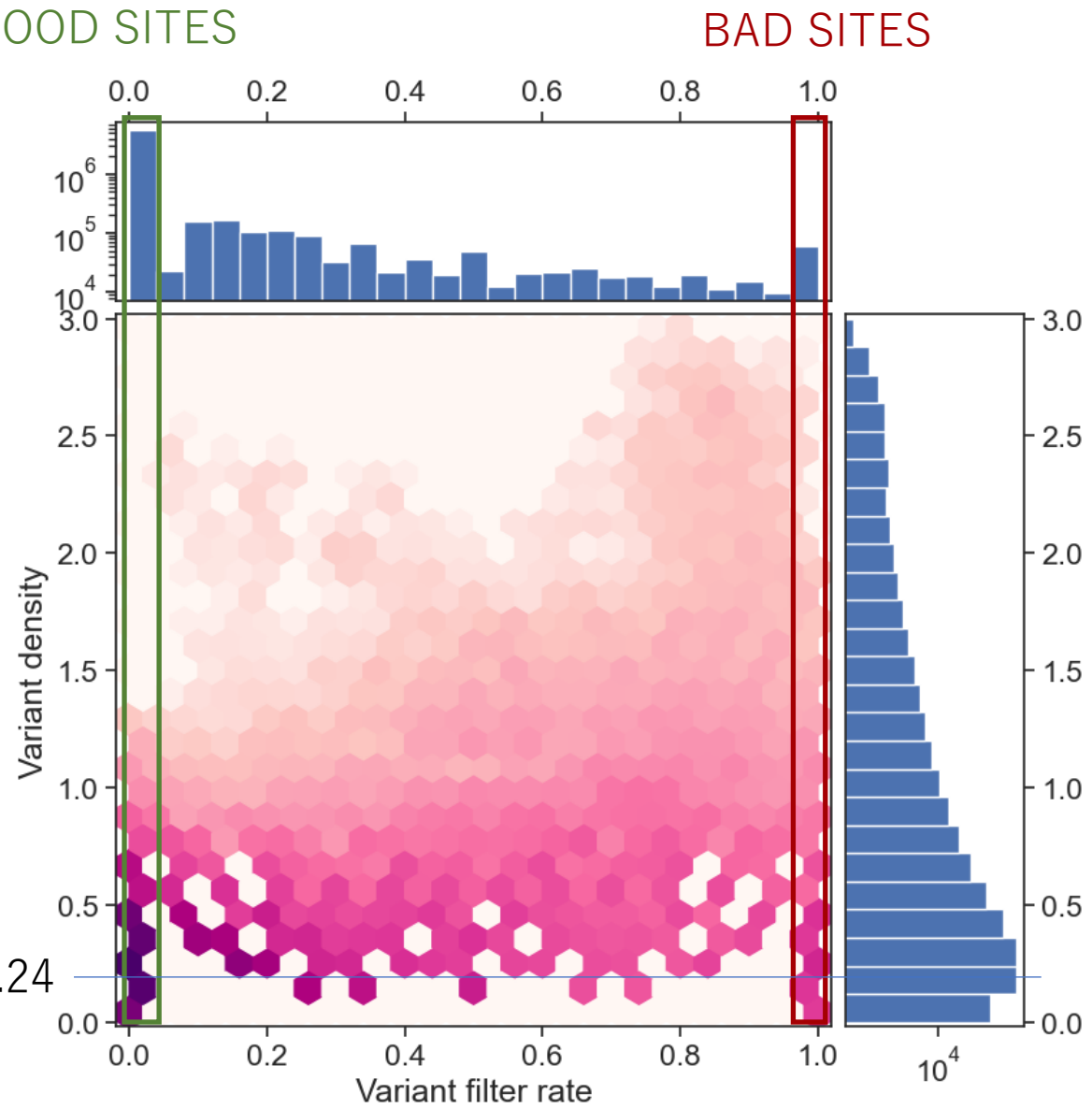
gnomAD v3.1バリアントデータにおける低品質バリアントサイトの特定

Variant densityとVariant filter rateの定義



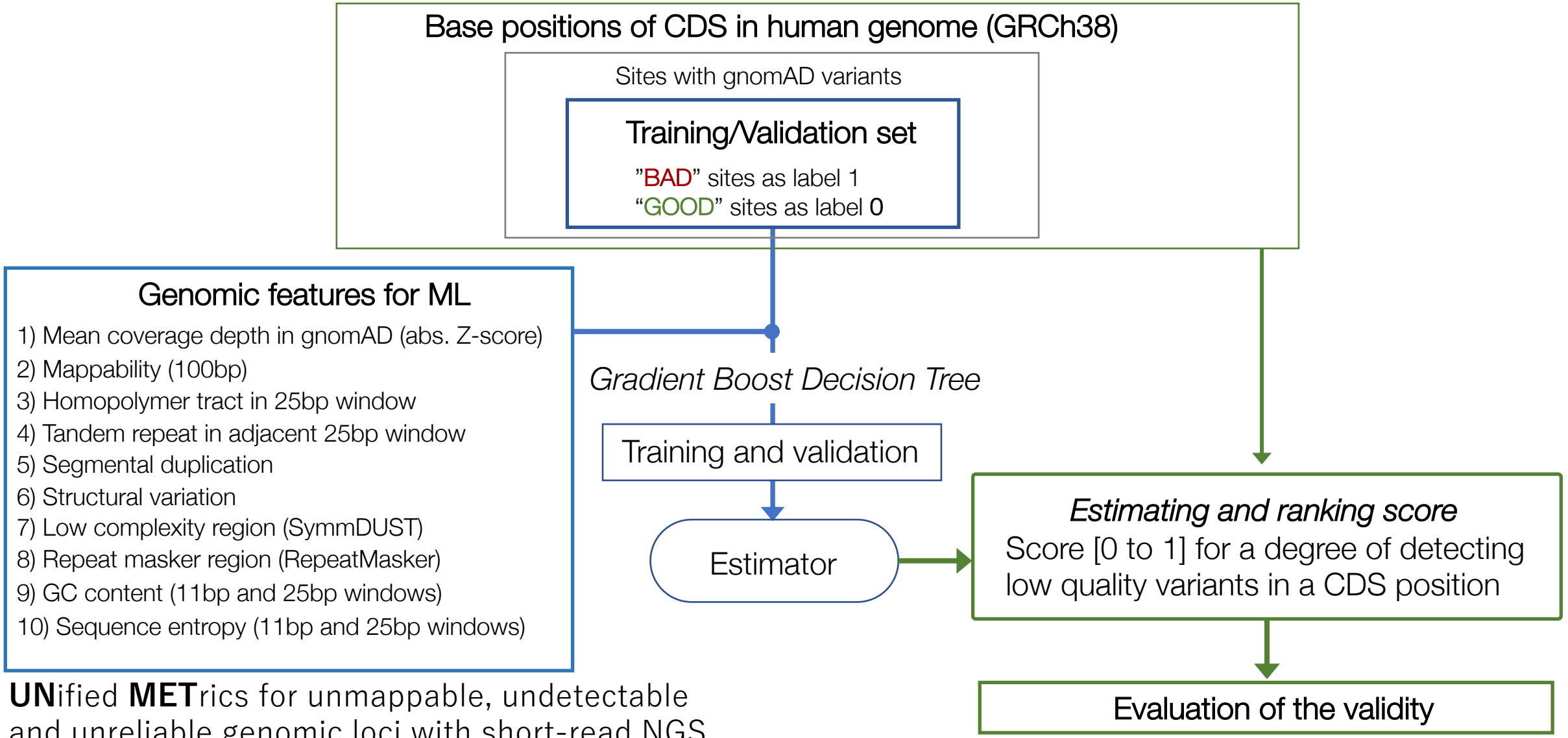
GOOD vs. BAD をゲノム配列の特徴から区別できるか？

機械学習による低品質サイトの
定量的かつ統一的な指標の作成



機械学習によるUNMETスコアの算出

1）機械学習により、BADとGOODを区別する学習モデルの作成
2）学習モデルから、各塩基の”BAD”らしさを定量化（UNMETスコア）

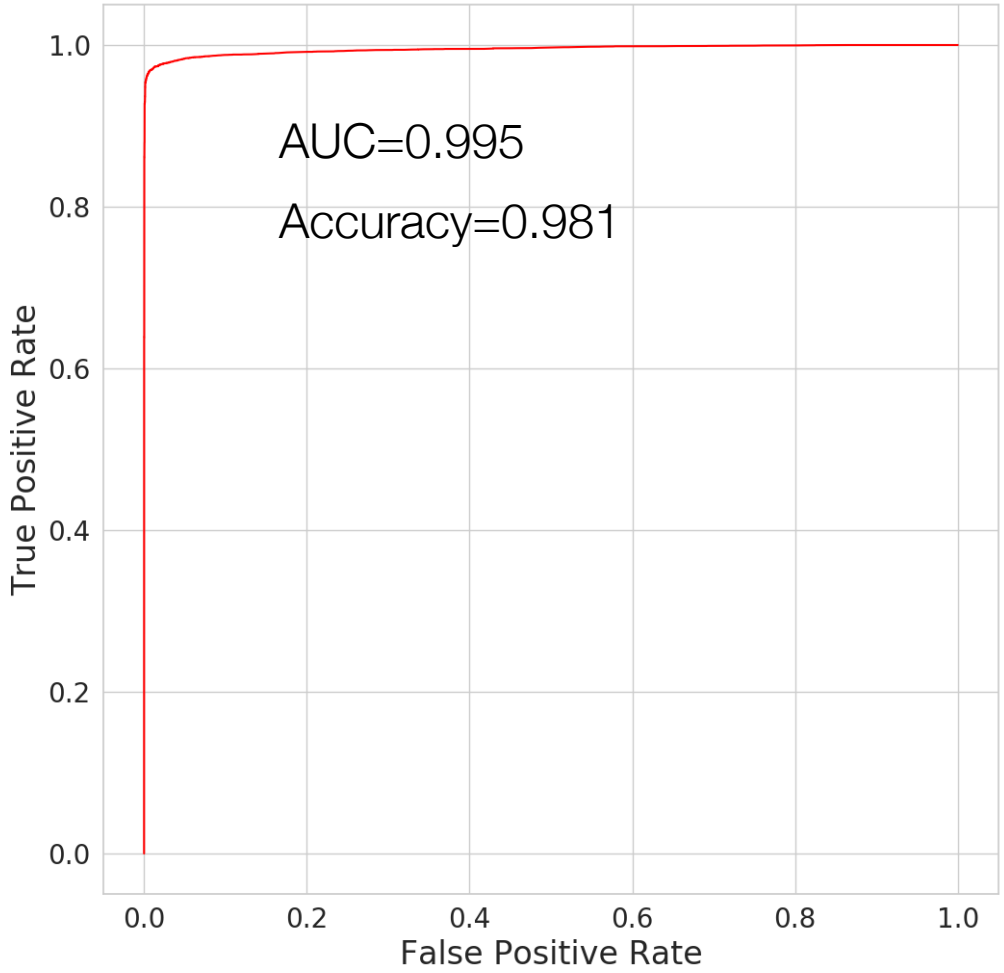


UNified **MET**rics for unmappable, undetectable and unreliable genomic loci with short-read NGS.

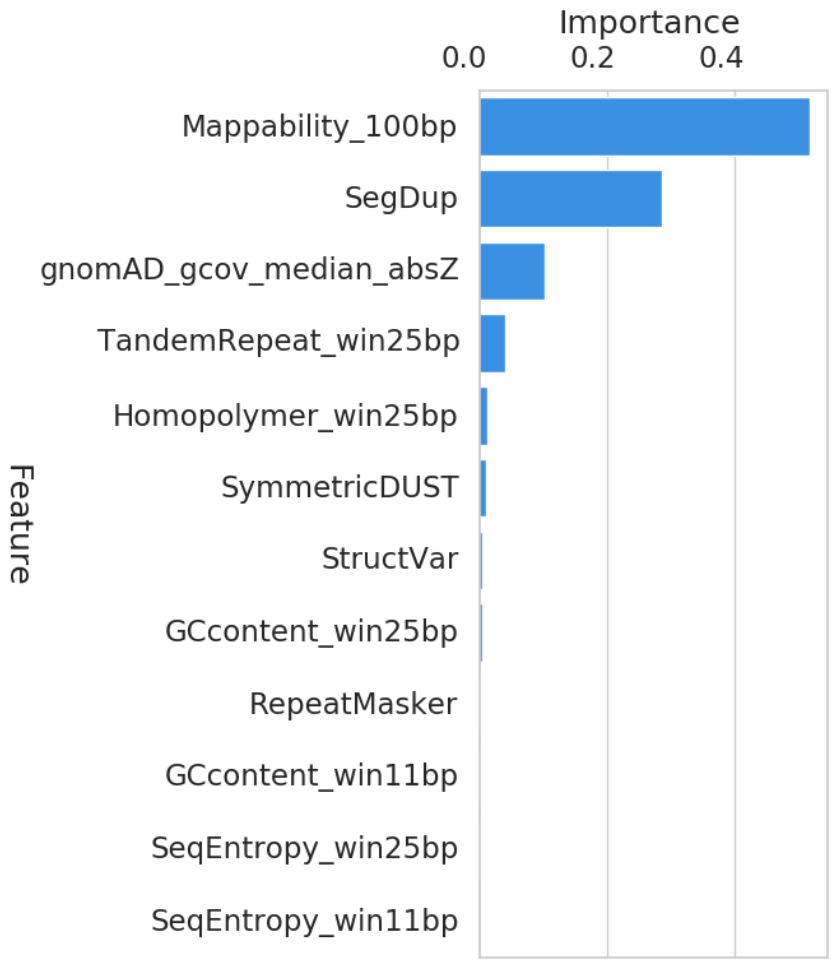
機械学習によるGOOD vs. BADの分類性能評価

・テストデータにおける精度は98.1% (誤分類率は約1.9%)
・重要度の最も高い特徴量はMappability

GBDTモデルによる分類性能評価

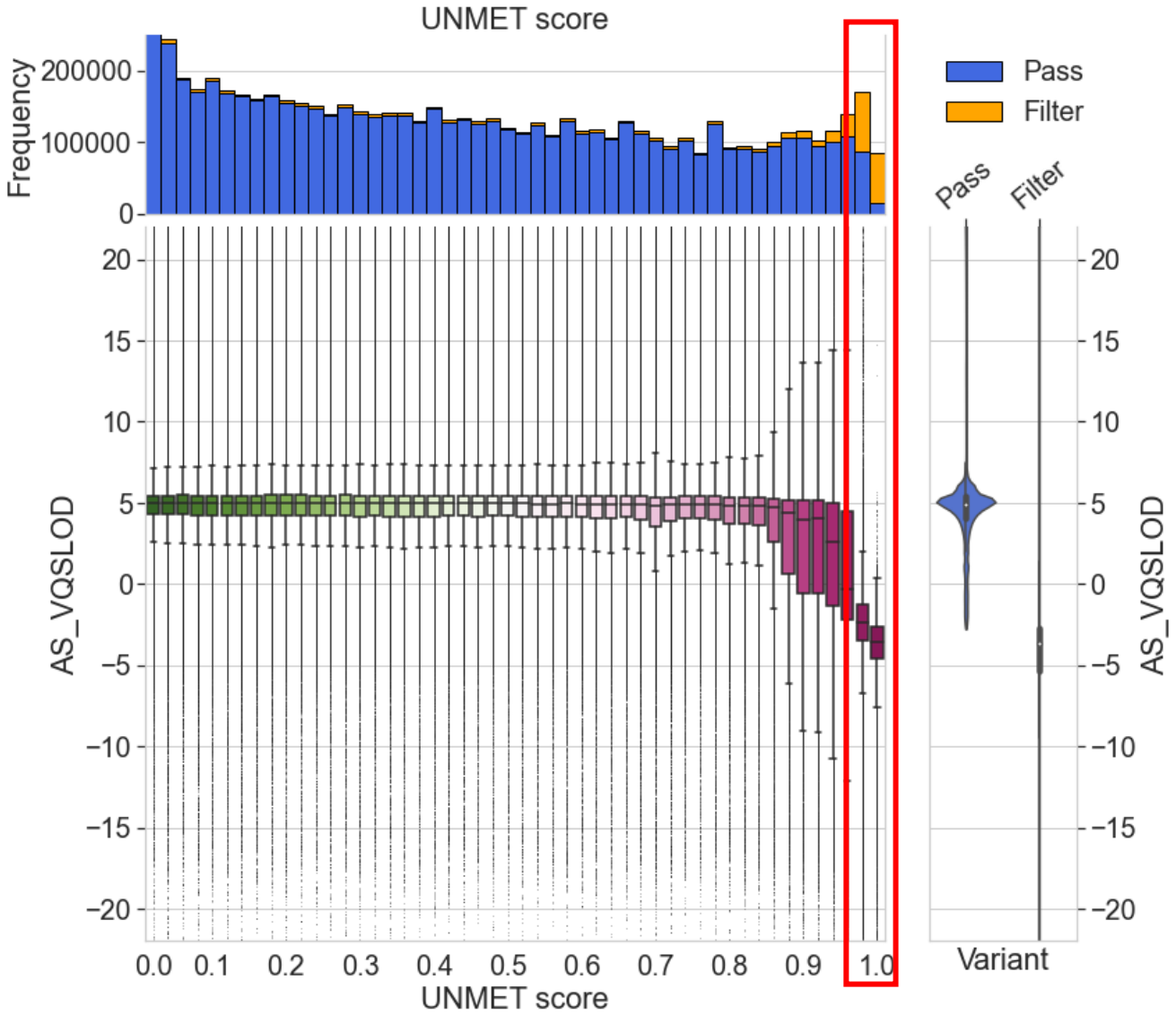


各特徴量の重要度



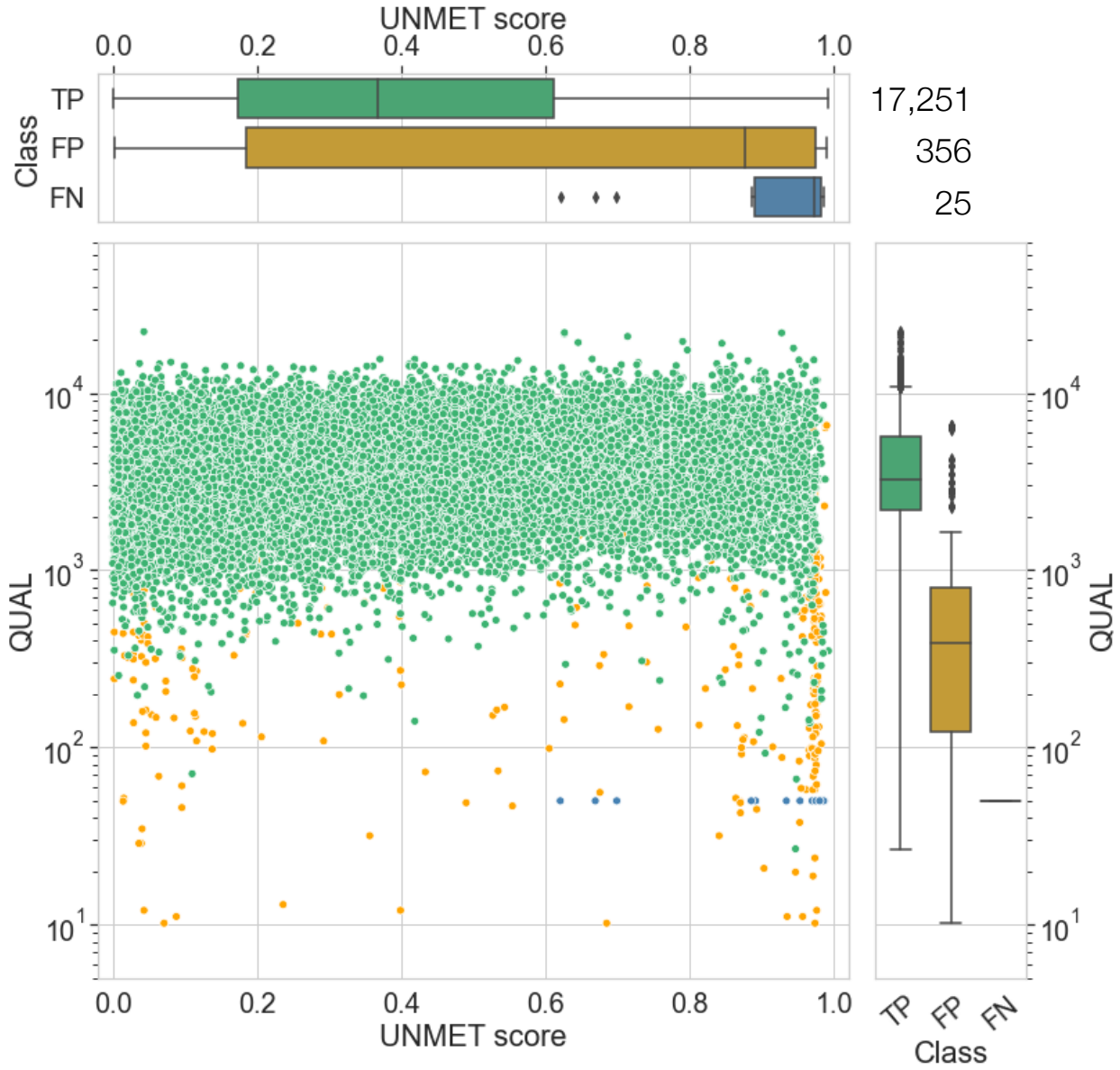
UNMETスコアの評価（１）AS_VQSLODスコアとの比較

- gnomADのバリアント評価スコア（AS_VQSLOD）とランクスコアは負の相関（Spearman: -0.223）
- 特にランクスコアが0.96を超えると急激にAS_VQSLODの値は低下する



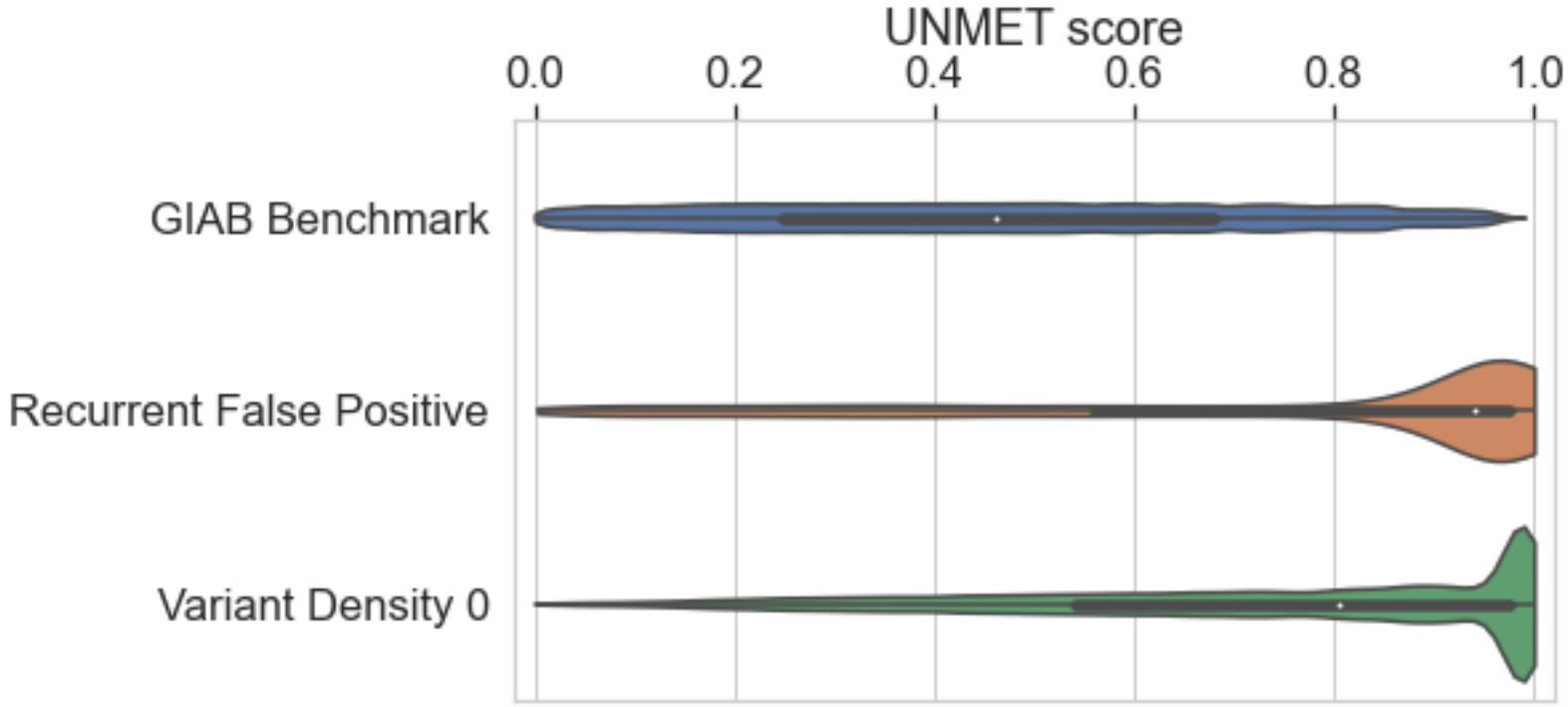
UNMETスコアの評価（２）Platinumゲノムサンプルのエクソームデータ

- エクソームデータ（HG0005、かずさでシークエンス）のQUAL値との比較
- UNMETスコアの分布はFPはTPに比べ高値に分布。FNはさらに高値に分布



UNMETスコアの評価（３）他のデータセットとの比較

１）Benchmark領域（バリアントコールの精度評価に用いられる領域）の中央値は0.5付近
２）誤検出されやすい領域(Recurrent false positive)、バリアントが検出できない領域は高いUNMETスコア



GIAB Benchmark: Intersection of the benchmark regions in 7 individual samples (HG0001-HG0007)

Recurrent False Positive: Field *et al.* (2019) BMC Genomics, **20**:546

Variant Density 0: The region of which the variant density is 0.