Supplemental Material-Performance Metrics

Nate Olson

August 18, 2014

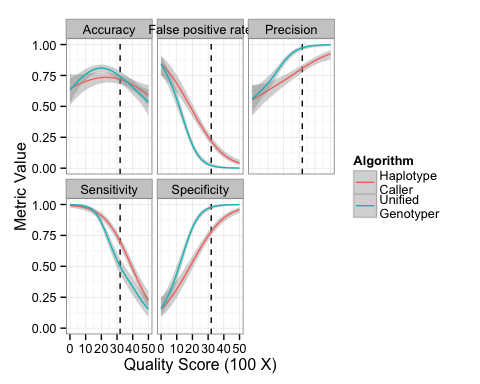
To demonstrate the different performance metrics that can be used to evaluate SNP calling pipeline we used variant calls for sequence datasets for the NIST *S. enterica* genomic DNA reference material. While we are in the process of characterizing the materials the results presented here and in the main manuscript are for demonstration purposes and the variants identified as true variants were defined solely for demonstration purposes and not represent a known biological variant. To generate the demonstartion data sequence data for eight replicate vials of the *S. enterica* NIST microbial genomic DNS reference material was mapped to a *Salmonella enterica* serovar Typhimurium LT2 reference genome in the GenBank database (NC\_003197 and NC\_003277) using the BWA MEM mapping algorithm (<http://bio-bwa.sourceforge.net/>). Variants were called using the GenomeAnalysisTK UnifiedGenotyper and Haplotype caller, variant calling algorithm (<https://www.broadinstitute.org/gatk/>). The union of the variant call sets was defined as the variant population. Variants with median quality scores less then 2500 were defined as false positives. This value was selected as it provided approximately equal numbers of positive (n = 48 and negative (n = 41) variants in the truth set. Additionally, for demonstration purposes quality scores were cut off at 5000 and for individual variant call sets where variants in the truth set were not called a quality score of zero was assigned.

The ROCR package was used to calcuate the performance metrics and ggplot2 was used to plot the performance metrics. See the github site for code use to generate this document and the generate the demonstration dataset

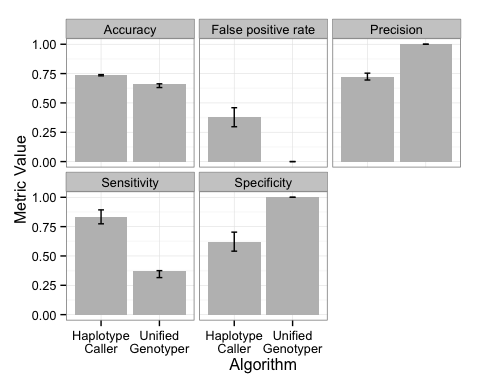
## Single Peformance Metric

Shaded region indicates the 95% confidence interval calculated from the eight replicate datasets.

### Continuous Parameter

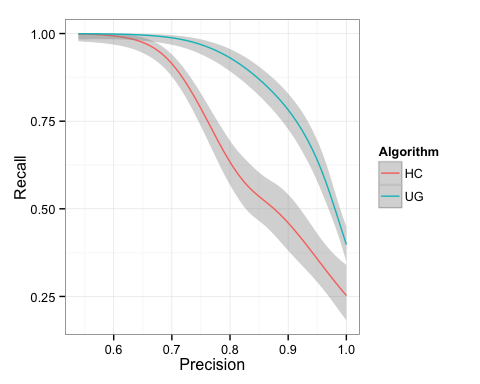


### Discrete Parameter

Bar plots with 95% confidence intervals (2.5-97.5 quantiles) are presented below, in the manuscript boxplots are used presents a single performance metric relative to a discrete parameter. 

## Two performance metrics

### Continuous Parameter: precision/recall curve



## Discrete Parameter

