# ExomePLUS Benchmark Steps

**Happy Metrics by Sample**

1. Apply GATK PASS/FAIL hard filters

bcftools filter -s FAIL -i '(TYPE="snp" & INFO/QD>=2 & INFO/FS<=60 & INFO/MQ>=40 & INFO/SOR<=3 & (INFO/MQRankSum="." | INFO/MQRankSum>=-12.4) & (INFO/ReadPosRankSum="." | INFO/ReadPosRankSum>=-8.0)) | (TYPE~"indel" & INFO/QD>=2 & INFO/FS<=200 & INFO/SOR<=10 & (INFO/ReadPosRankSum="." | INFO/ReadPosRankSum>=-20))' $vcf -Oz -o $output

1. Run ha.py

python hap.py --threads 3 $truthset $query -f $confidence\_bed -r $MUGQIC\_INSTALL\_HOME/genomes/species/Homo\_sapiens.GRCh37/genome/Homo\_sapiens.GRCh37.fa -o $dir/prefix

1. Aggregate .summary.csv data into indel\_all, indel\_pass, snp\_all, snp\_pass tables
   1. metric\_tables.R
      1. NOTE: you will need to manually enter your desired X axis (mean coverage, protocol name, etc.)
      2. input\_dir is directory containing all .summary.csvs of interest
2. Plot metrics based on manually defined X-axis
   1. metric\_figure.R
      1. input\_dir is directory containing output of step 4

**Gene**

* Location of CDS and UTR bed files by chromosome:
  + ~/projects/def-vmooser/clarewei/by\_gene
* Make bam files from vcf:

bcftools mpileup -f /cvmfs/soft.mugqic/CentOS6/genomes/species/Homo\_sapiens.GRCh37/genome/Homo\_sapiens.GRCh37.fa -E -q 20 -Q 20 --ff 0x0704 -d 8000 $bam -a FORMAT/DP | bcftools annotate -x INFO,FORMAT/PL -Oz -o $new\_vcf\_name

* Split by chromosomes and then into smaller chunks
  + Rscript findOverlap.R $small\_’chunk' $bed $output\_dir
  + Rscript combine\_by\_chrom.R $chr\_csv\_dir $output\_dir
* Rscript plot\_by\_gene.R
  + Will need to manually enter sample IDs and desired plotting order