**Workflow genome project in files filtering at first step**

* **Raw HG002 callset (from AWS DRAGEN):**  
  /home/q/project/HG002/data/raw/HG002.GRCh38.noprefix.vcf.gz

**Intermediate files (produced in Step 1)**

1. **Primary contigs only (Step 1.0)**
   * Command: bcftools view -r 1..22,X,Y,MT
   * Output:  
     /home/q/project/HG002/data/preproc/HG002.primary.GRCh38.vcf.gz
2. **Normalized + split multiallelics (Step 1.1)**
   * Command: bcftools norm -f GRCh38.fa -m -both
   * Input: HG002.primary.GRCh38.vcf.gz
   * Output:  
     /home/q/project/HG002/data/preproc/HG002.norm.split.GRCh38.vcf.gz
3. **First-pass QC (site-level PASS + QUAL≥30) (Step 1.2)**
   * Command: bcftools view -f PASS -i 'QUAL>=30'
   * Input: HG002.norm.split.GRCh38.vcf.gz
   * Output:  
     /home/q/project/HG002/data/preproc/HG002.qc1.passq30.GRCh38.vcf.gz
   * Stats file: results/qc\_stats\_passq30.txt

**Final cleaned callset (Step 1.6)**

1. **Genotype-level QC (DP≥10, GQ≥20, GT != mis)**
   * Command: bcftools view -i 'GT!="mis" && FMT/DP>=10 && FMT/GQ>=20'
   * Input: HG002.qc1.passq30.GRCh38.vcf.gz
   * **Final output:**  
     /home/q/project/HG002/data/preproc/HG002.qc2.genofilt.GRCh38.vcf.gz (+ .tbi)
   * Stats file: results/qc\_stats\_q30\_dp10\_gq20.txt