Variant Calling with GATK- Day 2

- GATK Tutorial
  - Variant_Discovery_Tutorial.pdf

Variant Calling with GATK- Day 2

- ## GATK Tutorial – Variant_Discovery_Tutorial.pdf
  - Skip Section 1 and SubSection 2.2
    - Except note the reference genome to choose is **Human (1kg, b37+decoy)**
  - So, go through these sections…
    - **2.1 DIAGNOSING UNKNOWN BAMS /View header and check read groups**
    - **3 VARIANT DISCOVERY**

- **IMPORTANT**: you will run the GATK commands on hoffman, but visualize the results on your local computer (using IGV and R).
  - The files transferred earlier from hoffman to your local computer include the output files you're about to create, so you don't have to transfer them, just look for the file with the same name.
  - They are in gatkWorkshop/1702/data_bundle/data/**outputs/**
  (not sandbox)

Run HaplotypeCaller using the hoffman queue

- Use the queue to run long jobs
- Use the queue to run many jobs in parallel

- Here we'll create a file with all of the commands we just ran, but with new output file names, and submit it to the queue.

  – Call the file whatever you want, maybe runHC.sh.
  – Make the first line "#!/bin/bash" (optional but good practice)
  – Add each command on it's own line (change the output file names)...

```
#!/bin/bash
java –Xmx1g  -jar $GATK -T HaplotypeCaller -R ref/ref.fasta -I bams/mother.bam \
-o sandbox/mother.q.g.vcf -ERC GVCF
java –Xmx1g -jar $GATK -T HaplotypeCaller -R ref/ref.fasta -I bams/father.bam \
-o sandbox/father.q.g.vcf -ERC GVCF
java –Xmx1g -jar $GATK -T HaplotypeCaller -R ref/ref.fasta -I bams/son.bam \
-o sandbox/son.q.g.vcf -ERC GVCF
java –Xmx1g -jar $GATK -T GenotypeGVCFs -R ref/ref.fasta -V sandbox/mother.q.g.vcf \
-V sandbox/father.q.g.vcf -V sandbox/son.q.g.vcf -o sandbox/trio.q.GGVCF.vcf
```

Submit runHC.sh to the hoffman2 queuing system

```
qsub –cwd –V –M joebruin@ucla.edu –m bea –l h_data=4G,h_rt=24:00:00 runHC.sh
```

Here's what each part means
qsub   - submit
-cwd   - run from the current working directory (so it will find files there)
-V   - use current environmental variables
-M   - give your email address if you want to receive email updates
-m   - when to email, b = beginning of job, e = end of job, a = abort
-l   - resource request, here 4GB memory and 24 hours.

This may run for a while, you can leave it overnight.

Check if running with

```
qstat |grep joebruin
```

When done, look for output files:
runHC.sh.o<jobNum>, runHC.sh.e<jobNum> and outputs from commands in runHC.sh