SNP Calling with GATK in the Discovery Environment

Put together by Jacob Landis for the CyVerse Webinar: "Hit the Ground Running with SNP analyses" on September 11, 2020

After signing into CyVerse (cyverse.org) click Launch the Discovery Environment. You will need to sign in again probably. On the left you should see three buttons: "Data", "Apps", and "Analyses". For this tutorial we will mostly be using the top two, but you can click the "Analyses" button at any time to see what is currently running on your account.

Click the "Data" button. Your username should be listed on top. Click on that, then the "analyses" folder. Click File ->New Folder to create a folder called Tutorial. This will be our main repository for this walkthrough. All of the files that you will need for this have previously been uploaded to a shared Community Data folder called "Botany2020NMGWorkshop". Specifically, what you need is here:

Reference genome:
/iplant/home/shared/Botany2020NMGWorkshop/Genome_assembly/Completed_assemblies/Utricularia_gibba_PNAS2017.fasta

Sequencing reads (14 gzipped fastq files):
/iplant/home/shared/Botany2020NMGWorkshop/raw_data/Ugibba/transcriptome

To get the necessary files into the correct folder, there are two options. The first, which takes much longer, is to download each file that is needed onto your local machine and then uploaded those to a new folder in your account and analyses. The easier option is to leave the files where they are, and when we run the first step (indexing the reference genome fasta file), check the option to "Retain inputs". This will put a copy into the folder you specify. This is my recommendation, unless you want to upload some of your own data. However, this pipeline run in the Discovery Environment only accommodates small data sets since the maximum allocations each person can get is 8 CPUs and 15 GB of RAM.

For each of the steps below, you can find the necessary program by clicking "Apps" and type in the name as shown below in the Search Apps tool bar. If you are going to be using the same apps time-and-time again, I would suggest clicking on the three dots on the right side of the app and click "Add to Favorites". Each app will create a subfolder with all the log files if you want to check anything out. After each run we will move the resulting output files back to our main Tutorial folder.

1. **Index the reference with BWA** – This is for mapping reads

Analysis Name: BWA index

Keep the analysis name the same but change the output folder to Tutorial in your analyses folder. Everything will be put here to keep us organized.
Check the “Retain inputs” option to move a copy of the reference fasta file to our analysis directory.
Options:
Select a FASTA file to index: Utricularia_gibba_PNAS2017.fasta
Keep BWT construction algorithm: Auto
Uncheck the option “Index files with new BWA 0.6x+ naming scheme”
Click -> *Launch analysis*

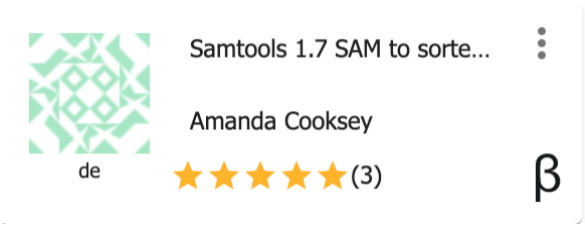
Once the analysis finishes, we will need to move the output files into our main directory so we don’t get bogged down in folders. For all the 6 files in the folder “BWA_index…”, click the little box next to the file name. When all are selected, click Edit -> Move, and select our main analysis folder.

2. **BWA mem 0.7.15** – Map reads from each individual to the reference genome

Analysis Name: BWA_mem_0.7.15
Can keep the name the same. For output folder select our Tutorial folder.
Inputs
Left Read file: select Ugibba_bladderR1.fastq.gz (You will have to do this separately for each sample in the current format of the app). For this data set, we only have single end reads. If you had paired end, you would specify the appropriate partner read in the Right Read File.
Reference Genome
Select the reference genome we just indexed in the previous step (Utricularia_gibba_PNAS2017.fasta).
Alignment options
Keep all as default
Output options
Keep as default
Click -> *Launch analysis*

Once the run is complete, move the resulting sam file to our main folder. You should rename it Ugibba_bladderR1.sam. After mapping the reads to the genome, we do not need to worry about the fastq.gz files any more.

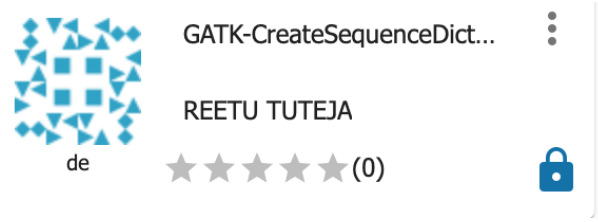
3. **Samtools SAM to sorted BAM** – Convert from SAM to sorted BAM to save computational resources



Analysis Name: Samtools SAM to sorted BAM
Keep the analysis name the same but make sure to change the output folder to "Tutorial"
Input file:
Select your SAM file that you just moved to the main analysis folder, in this case Ugibba_bladderR1.sam.
Output file:
Output File Name: Ugibba_bladderR1_sorted.bam
Output file format: BAM
Options:
Keep as default
Click -> *Launch analysis*

Once run finishes, move the resulting BAM file to the main folder.

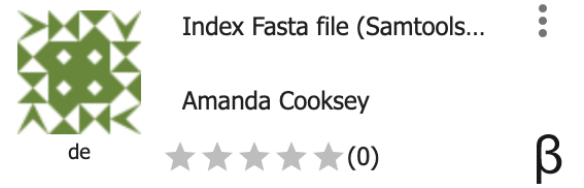
4. **GATK CreateSequenceDictionary** – Create a dictionary of the reference genome



Analysis Name: GATK-CreateSequenceDictionary
Keep the analysis name the same but make sure to change the output folder to "Tutorial"
Input:
Reference Sequence: Utricularia_gibba_PNAS2017.fasta
Output: Utricularia_gibba_PNAS2017.dict
Click -> *Launch analysis*

Once run finishes, move the resulting .dict file to the main folder.

5. **Index fasta file (Samtools)** – Index the reference genome for SNP calling



Analysis Name: Index Fasta file Samtools faidx

Keep the analysis name the same but make sure to change the output folder to “Tutorial”
Reference input:
Select the reference fasta file: Utricularia_gibba_PNAS2017.fasta
Click -> *Launch analysis*

Once the run finishes, move the resulting .fai file to the main folder.

6. **GATK MarkDuplicates** – Mark PCR duplicates. Should do this for any data that was PCR amplified. Would recommend not doing this for RAD-Seq data though.

Analysis Name: GATK-MarkDuplicates
Keep the analysis name the same but make sure to change the output folder to “Tutorial”
Input data:
Select the sorted BAM file you want to mark, in this case it is Ugibba_bladderR1_sorted.bam
Output:
Output file: Ugibba_bladderR1_sorted.duplicates.bam
Metrics File: Ugibba_bladderR1_sorted.duplicates.metrics
Click -> *Launch analysis*

Once the run finishes, move the resulting .bam file to the main folder.

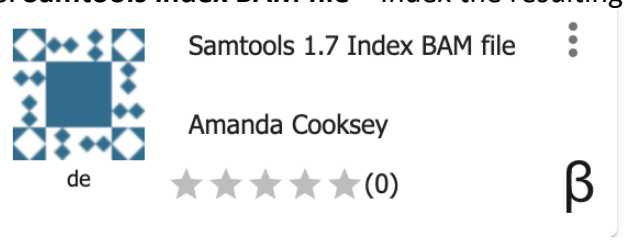
7. **GATK AddOrReplaceReadGroups** – Add ReadGroup for each sample. In the command line tutorial we do this in the BWA MEM step, but the current app does not allow that.

Analysis Name: GATK-AddOrReplaceReadGroups
Keep the analysis name the same but make sure to change the output folder to “Tutorial”
Input:
Input file: Ugibba_bladderR1_sorted.duplicates.bam
LB: 1
PL: Illumina
PU: rnaseq
SM: Ugibba_bladderR1
Output:
Output File: Ugibba_bladderR1_sorted.duplicates.RG.bam

Click -> *Launch analysis*

Once the run finishes, move the resulting .bam file to the main folder. For all other samples in this data set, you can keep LB, PL, and PU with the exact same information. Need to make sure that SM is a unique sample name, in this tutorial include the name of the files which has species+organ+replicate.

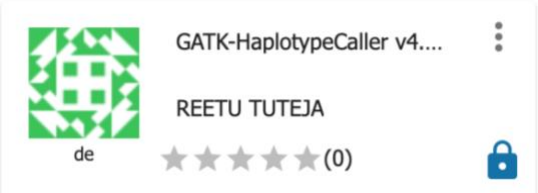
8. **Samtools Index BAM file** – Index the resulting Duplicates Marked BAM file for SNP calling



Analysis Name: Samtools Index BAM
Keep the analysis name the same but make sure to change the output folder to "Tutorial"
Inputs:
Select a BAM file to index: Ugibba_bladderR1_sorted.duplicates.RG.bam
Click -> *Launch analysis*

Once the run finishes, move the resulting .bai file to the main folder.

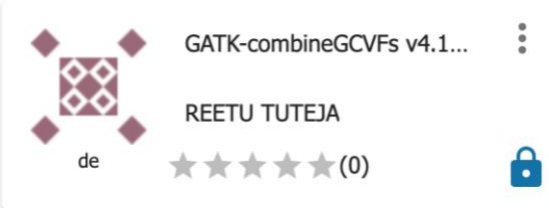
9. **GATK HaplotypeCaller** – Do initial SNP calling for each sample with HaplotypeCaller. This part will take several hours for each sequencing sample.



Analysis Name: GATK-HaplotypeCaller
Keep the analysis name the same but make sure to change the output folder to "Tutorial"
Input data:
Reference sequence file: Utricularia_gibba_PNAS2017.fasta
Reference genome index file: Utricularia_gibba_PNAS2017.fasta.fai
Reference genome dict file: Utricularia_gibba_PNAS2017.dict
Input File: Ugibba_bladderR1_sorted.duplicates.RG.bam
Input File Index: Ugibba_bladderR1_sorted.duplicates.RG.bam.bai
Emit-ref-confidence: GVCF
Ouput:
Output File: Ugibba_bladderR1.g.vcf.gz
Click -> *Launch analysis*

Once the run finishes, move the resulting g.vcf.gz and g.vcf.gz.tbi file to a new folder titled "GVCF" within the Tutorial folder.

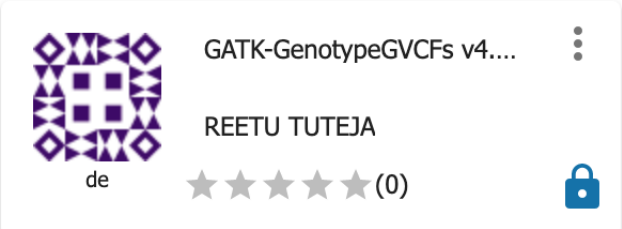
10. **GATK CombineGVCFs** – Combine all the GVCF files from each HaplotypeCaller step into one file so that we can do Joint Genotyping next which incorporates data from all of our samples to determine what is a variant.



Analysis Name: GATK-combineGVCFs
Keep the analysis name the same but make sure to change the output folder to the new folder you just made "GVCF"
Input data:
Reference sequence file: Reference sequence file: Utricularia_gibba_PNAS2017.fasta
VCF file(s): Click Add, and select all the g.vcf.gz files that you have created.
Output:
Output File: Ugibba_combined.g.vcf.gz
Click -> *Launch analysis*

Once the run finishes, move the resulting g.vcf.gz and g.vcf.gz.tbi file to the "GVCF" folder.

11**. GATK GenotypeGVCFs** – Final step in SNP calling. Resulting file is a vcf.gz file which can be used for SNP filtering and downstream analyses. This part may take several hours, especially with larger data sets.



Analysis Name: GATK-GenotypeGVCFs
Keep the analysis name the same but make sure to change the output folder to the new folder you just made "GVCF"
Input data:
Reference sequence file: Utricularia_gibba_PNAS2017.fasta
Reference genome index file: Utricularia_gibba_PNAS2017.fasta.fai
Reference genome dict file: Utricularia_gibba_PNAS2017.dict
VCF file(s): Ugibba_combined.g.vcf.gz
Indexed input files: Ugibba_combined.g.vcf.gz.tbi
Output:
Output File: Ugibba_initial_SNP_calls.vcf.gz

You are now done with SNP calling and can move onto filtering. Currently there is not an App set up to do that in the Discovery Environment, but if you follow the SNP_filltering.sh script on GitHub you can do it on your local machine.