SNP Calling with GATK in the Discovery Environment

Put together by Jacob Landis for the CyVerse Workshop: "AG2PI: Introduction to SNP Data Analysis" on May 20th, 2021

After signing into CyVerse (cyverse.org) click Launch the Discovery Environment. You will need to sign in again. On the left you should see several buttons including: "Home", "Data", "Apps", and "Analyses".

Click the "Data" button. Your username should be listed on top. Click on the "analyses" folder. Click "+Folder" icon to create a folder called "AG2PI_SNP" in your local directory. This will be our main repository for this walkthrough. All of the files that we will need for this have previously been uploaded to a shared Community Data folder called "AG2PI_Workshop_May2021". Specifically, what you need is here:

Reference genome:
/iplant/home/shared/AG2PI_Workshop_May2021/Utricularia_gibba_PNAS2017.fasta

Sequencing reads (14 gzipped fastq files):
/iplant/home/shared/AG2PI_Workshop_May2021/Ugibba_Illumina/

For most of the analyses we can just leave the input files in the Community folder. However, to make things easier, it will be best to download the genome file and upload that to your current working folder. This can be done by clicking on the box next to the fasta file, click the three little dots in the top right, and download. When the download is complete, navigate to your working directory in your analyses folder, and upload the file by clicking the "Upload" button.

Each app will create a subfolder by default with all the log files and resulting files. However, we can also specify the output directory to our current working directory so that we do not need to worry about transferring any other files.

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

Name ↑	Integrated By
BWA index 0.7.4	Matthew Vaughn

Analysis Name: BWA index
Output directory keep with our default for the tutorial:
/iplant/home/*username*/analyses/AG2PI_SNP
Select fast file: Utricularia_gibba_PNAS2017.fasta or
/iplant/home/shared/AG2PI_Workshop_May2021/Utricularia_gibba_PNAS2017.fasta if you didn't move the fast file to your current folder yet.

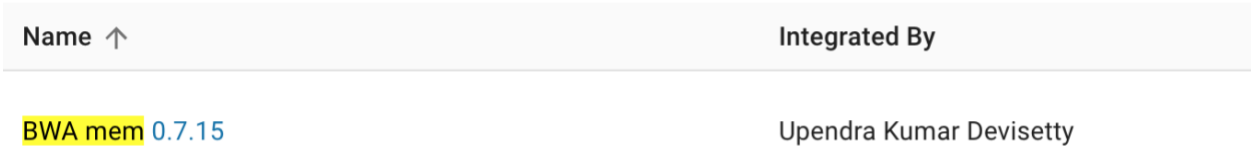
BWT construction algorithm: Leave with "Auto"
Uncheck the box "Index files with new BWA 0.6x+ naming scheme"

Minimum resource requirements:
Minimum CPU cores: 8
Minimum memory: 2 GiB
Minimum Disk Space: 2 GiB

Launch analysis
You should be taken to a list of all your Analyses. You can see when this index run finishes. Should take 3-5 minutes.

Once the analysis finishes, you should see 5 files in the output folder.

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



Analysis Name: BWA_mem_0.7.15

Inputs
Left Read file: select AG2PI_Workshop_May2021/Ugibba_Illumina/Ugibba_bladderSample1_subset.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 "Next"

Minimum resource requirements:
Minimum CPU cores: 8
Minimum memory: 2 GiB
Minimum Disk Space: 2 GiB

Click -> *Launch analysis*

This part can take a while, especially with a large data set. The file subsets we are using only have 250,000 reads, therefore the mapping will only take about 5 minutes per file. The full files of these samples would take about 2 hours each to map. One major issue that we need to sort is how the output files are named. The default is “bwa_output.sam”, so we need to make sure that we rename each output file before moving on. We can do this by going to “Data” and selecting the box to the left of the file. Then Click on the three dots in the top right and hit rename. Make sure to rename it to match the input file, such as “Ugibba_bladderR1.sam”

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

Name ↑	Integrated By
Samtools 1.7 SAM to sorted BAM	Amanda Cooksey

Analysis Name: Samtools 1.07 SAM to sorted BAM

Input file:
Select your SAM file that you just created, in this case Ugibba_bladder_Sample1.sam.
Output file name: Ugibba_bladder_Sample1_sorted.bam
Output file format: BAM
Options: sort by reference coordinates
Next

Minimum resource requirements:
Minimum CPU cores: 8
Minimum memory: 2 GiB
Minimum Disk Space: 2 GiB

Click -> *Launch analysis*

Once run finishes, you should see a sorted BAM file in your directory.

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

Name ↑	Integrated By
GATK-CreateSequenceDictionary v4.18.1	REETU TUTEJA

Analysis Name: GATK-CreateSequenceDictionary

Input:

Reference Sequence: Utricularia_gibba_PNAS2017.fasta
Output: Utricularia_gibba_PNAS2017.dict

Minimum resource requirements:
Minimum CPU cores: 4
Minimum memory: 2 GiB
Minimum Disk Space: 2 GiB
Click -> *Launch analysis*

Once run finishes, you should see a resulting .dict file.

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

Name ↑	Integrated By
Index Fasta file (Samtools 1.7 faidx)	Amanda Cooksey

Analysis Name: Index Fasta file Samtools faidx

Reference input:
Select the reference fasta file: Utricularia_gibba_PNAS2017.fasta

Minimum resource requirements:
Minimum CPU cores: 4
Minimum memory: 2 GiB
Minimum Disk Space: 2 GiB
Click -> *Launch analysis*

Once the run finishes, you should see a .fai file in your analyses 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. **However, we are going to skip this step in our currently analyses.**

Name ↑	Integrated By
GATK-MarkDuplicates v4.1.8.1	REETU TUTEJA

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.

Name	Integrated By
GATK-AddOrReplaceReadGroups v1.4.8.1	REETU TUTEJA

Analysis Name: GATK-AddOrReplaceReadGroups

Input:
Input file: Ugibba_bladder_Sample1.bam
LB: 1
PL: Illumina
PU: rnaseq
SM: Ugibba_bladder_Sample1
Output:
Output File: Ugibba_bladder_Sample1_sorted.RG.bam

Minimum resource requirements:
Minimum CPU cores: 8
Minimum memory: 4 GiB
Minimum Disk Space: 4 GiB
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+sample.

8. **Samtools Index BAM file** – Index the BAM file containing ReadGroups for SNP calling

Name	Integrated By
Samtools 1.7 Index BAM file	Amanda Cooksey

Analysis Name: Samtools 1.07 Index BAM

Inputs:
Select a BAM file to index: Ugibba_bladder_Sample1.RG.bam

Minimum resource requirements:
Minimum CPU cores: 8
Minimum memory: 4 GiB
Minimum Disk Space: 4 GiB

Click -> *Launch analysis*

Once the run finishes, you should now have a .bai file in the analyses folder.

9. **GATK HaplotypeCaller** – Do initial SNP calling for each sample with HaplotypeCaller. This part can take several hours for each sequencing sample, though our subset should finish in 10-15 minutes.

Name ↑	Integrated By
GATK-HaplotypeCaller v4.1.8.1	REETU TUTEJA

Analysis Name: GATK-HaplotypeCaller

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.RG.bam
Input File Index: Ugibba_bladderR1_sorted.RG.bam.bai
Emit-ref-confidence: GVCF
Output: Ugibba_bladderR1.g.vcf.gz

Minimum resource requirements:
Minimum CPU cores: 8
Minimum memory: 4 GiB
Minimum Disk Space: 4 GiB

Click -> *Launch analysis*

Once the run finishes, you should see the resulting g.vcf.gz and g.vcf.gz.tbi files.

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.

Name ↑	Integrated By
GATK-combineGCVFs v4.1.8.1	REETU TUTEJA

Analysis Name: GATK-combineGVCFs

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

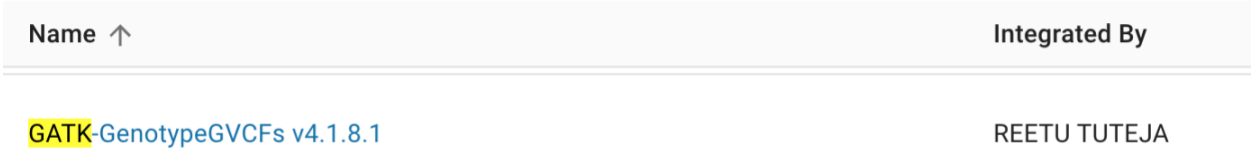
VCF file(s): Select all the g.vcf.gz files that need to be included
Indexed input files: Select all the g.vcf.gz.tbi files that need to be included (needs to match the vcf files included)
Output File: Ugibba_combined.g.vcf.gz

Minimum resource requirements:
Minimum CPU cores: 8
Minimum memory: 4 GiB
Minimum Disk Space: 4 GiB

Click -> *Launch analysis*

Once the run finishes, you will have the resulting g.vcf.gz and g.vcf.gz.tbi file in your current 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.



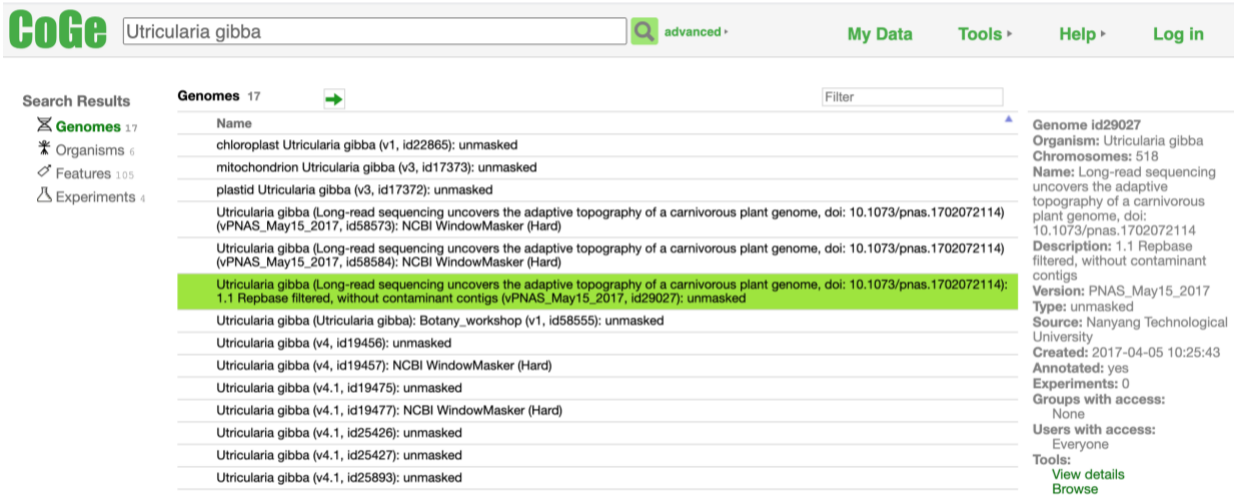
Analysis Name: GATK-GenotypeGVCFs

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 File: Ugibba_initial_SNP_calls.vcf.gz

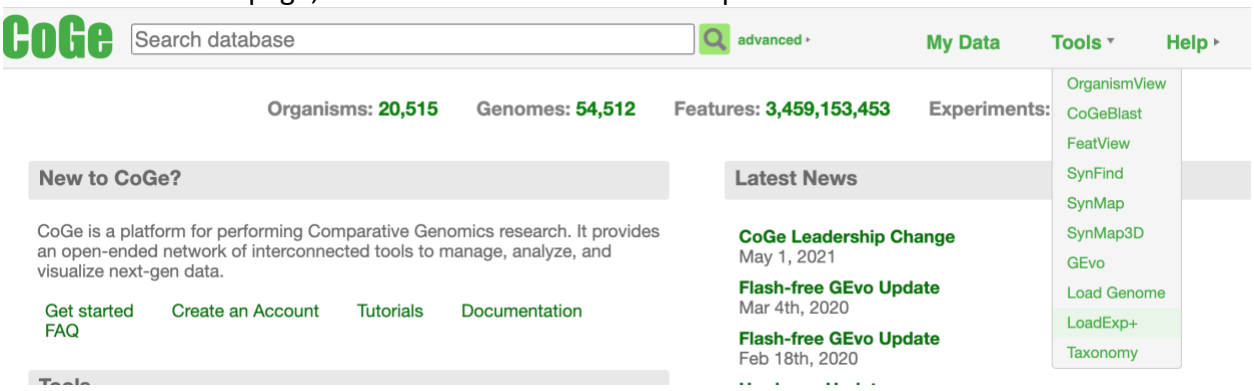
You are now done with SNP calling. Normally we would need to do some filtering to make sure we retain only high-quality SNPs, but there is not a good App on CyVerse to do that currently. The easiest approach is to download VCFtools on your local machine and follow the SNP_filltering.sh script on GitHub. For now, we are going to take a created VCF file with the samples we dealt with today and visualize the SNPs in CoGe, which is also part of CyVerse.

CoGe analyses (https://genomevolution.org/coge/)

We need to upload a VCF file that was called against a genome already present in CoGe. For this tutorial, we used the following genome:



On the CoGe main page, click “Tools” and then “LoadExp+”

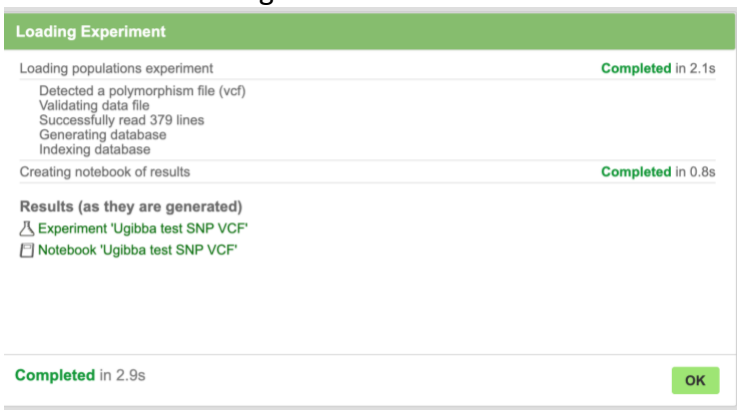


You can either select the VCF file from the CyVerse Data Store, either your resulting GATK VCF from the analyses folder, or the supplied VCF in the AG2PI_Workshop_May2021 Community Data folder. You can also upload a VCF by clicking the upload tab. Add the data to a new notebook.

Title: “Ugibba test SNP VCF”
Version: 1.1
Source: Utricularia gibba RNA-Seq
Genome: Utricularia gibba (Long-read sequencing uncovers the adaptive topography of a carnivorous plant genome, doi: 10.1073/pnas.1702072114): 1.1 Repbase filtered, without contaminant contigs (vPNAS_May15_2017, id29027): unmasked

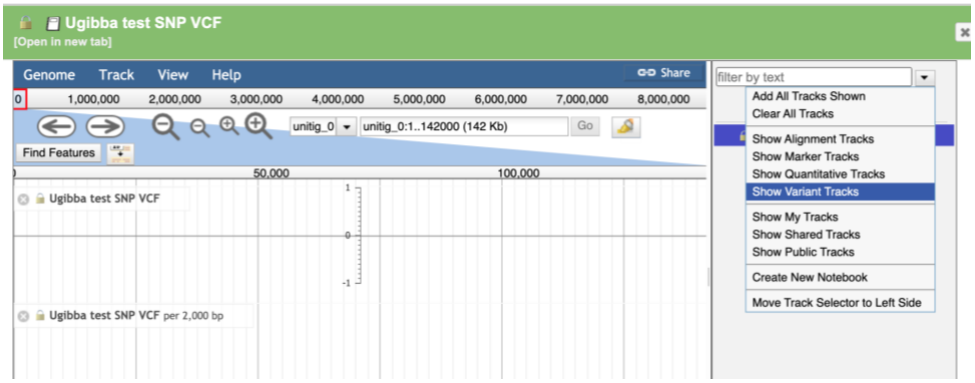
Click Next, then Start.

If all works, you should see the following screen:

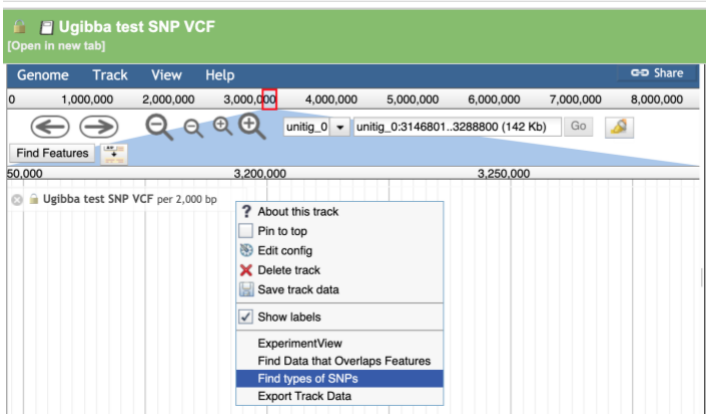


Click “My Data”, then “Notebooks” tab. Double click the Ugibba test SNP VCF, which should open a new box. Then click “Browse”.

Along the top right, click the dropdown arrow, the select “Show Variant Tracks”



To find particular SNPs, Click the Ugibba test SNP VCF per 2,000 bp link in the main window, then “Find types of SNPs”



Select any types that you want to see, then use the arrows to scan through all the SNPs on the current contig/chromosome.