Step-by-Step Guide: Variant Calling

Author: Itunuoluwa Isewon PhD

Email: itunu.isewon@covenantuniversity.edu.ng

You have been provided with a text file which contains six sample IDs, upload by clicking on the upload button and dropping the file.

Data: Click this LINK to access the data, then **import history** accordingly **[change history name to ACEMFS_Your-Name].**

Quality Control

N.B.: Skip "Obtain Fastq files" if you click the data LINK above.

Obtain Fastq files:

1. In the toolbar, click on **Get Data**
2. Choose "**Download and Extract Reads in FASTA/Q format from NCBI SRA**"
3. Change the select input type to "**List of SRA accession',** then choose your sample ID file and run the tool
4. In this tutorial, we'll use six datasets.

Sample ID	Condition
SRR15044361	test
SRR15044360	test
SRR15044359	test
SRR15044358	control
SRR15044357	control
SRR15044356	control

Perform QC:

1. In the search bar, type **fastqc**
2. Choose the desired fastq files (paired end) in the raw read tab.
3. Leave all other tabs unchanged
4. Once it runs, two files are generated: a raw data file and a Webpage file
5. View the result by clicking on the webpage file produced
6. Repeat for the second data set and compare their results.

Multiqc

Why: It helps us to obtain a more intuitive comparison

1. In the search bar, type **multiqc**
2. On the "**Which tool was used to generate logs?**" tab, choose **Fastqc**
3. Then click on "**Insert FastQC output**"
4. Type of output is raw data
5. Add the raw data files generated earlier
6. Leave all other parameters at the default

7. Run tool
8. View the result by clicking on the webpage file produced

Variant Calling

Mapping

1. Search for **Map with BWA-MEM** in the tool search bar, choose the options for longer reads
2. We would be using a built-in genome
3. Choose **Aspergillus flavus NRRL3357** as the reference genome
4. Leave other parameters as default

Descriptive statistics

1. Search for **Samtools flagstat** in the tool search bar, and choose the options for longer reads
2. Select the file generated from the BWA-MEM and leave the output format as txt
3. Run tool
4. view results

Generate genotype likelihoods

1. Search for **bcftools mpileup** in the tool search bar
2. We are using a single BAM alignment input
3. Select the file generated from the BWA-MEM
4. The reference genome is **Aspergillus flavus NRRL3357**
5. Output format is uncompressed VCF
6. Run tool

Variant calling

1. Search for **bcftools call** in the tool search bar, choose the options for longer reads
2. Select the file generated from the **bcftools mpileup**
3. Leave all other parameters default
4. Output format is uncompressed VCF
5. Run tool
6. View result

Remove homologous variants and variants with missing phenotype

1. Search for **Filter data on any column using simple expressions** in the tool search bar
2. Select the file generated from the **bcftools call**
3. Supply the condition c10 != '0/0': sample genotype information is on the tenth column, != means not equal to, '0/0' represents homologous variants (portions of the genome not different from the reference)
4. Run tool
5. View result
6. Search for **Filter data on any column using simple expressions** in the tool search bar
7. Select the file generated from the last step
8. Supply the condition c10 != './.' : './.' denotes missing data
9. Run tool

10. View result

Sorting

1. Find sort in the search bar, choose “Sort data in ascending or descending order”
2. Sort on column 6: Quality
3. Keep every other parameter as the default.
4. Variants with high quality are now on top.

Variant Annotation with Ensembl Fungi VEP

1. Search for Variant Effect Predictor (VEP) in the Ensembl tools search bar.
2. Select the input file: choose the VCF file generated from the sorting step.
3. Species selection: Set species to your organism (e.g., Aspergillus flavus, Saccharomyces cerevisiae, etc.) from the Ensembl Fungi database.
4. Input format: keep as VCF.
5. Output options: Keep default output as tab-delimited text or select VCF with annotations if you prefer an annotated VCF.
6. Annotations to include (keep defaults, but you can also enable if available):
   - Gene symbol
   - Consequence terms (missense, synonymous, stop gained, etc.)
   - Protein domains (Pfam, InterPro)
   - SIFT/PolyPhen predictions (if available for your species)
   - Transcript ID and biotype

In the meantime, a list of variants has been provided for you to use. Run each of these using the “ RUN VEP! For this line option”.

AAIH03000093.1	2709654	.	G	A	3.02336
AAIH03000170.1	1814273	.	T	A	3.02336
AAIH03000282.1	926657	.	G	T	3.02336
AAIH03000072.1	3023139	.	T	C	3.02501
AAIH03000170.1	1417818	.	C	T	3.02996
AAIH03000103.1	103700	.	taaaaaaa	taaaaaa	3.03091
AAIH03000103.1	598760	.	T	C	3.03539
AAIH03000235.1	829594	.	G	A	3.04327
AAIH03000072.1	2160575	.	A	C	3.04541
AAIH03000226.1	603606	.	T	A	3.05565
AAIH03000226.1	3304340	.	T	A	3.06291
AAIH03000011.1	401846	.	G	A	3.07192
AAIH03000072.1	4101557	.	T	C	3.07192
AAIH03000173.1	1293019	.	tc	t	3.07736
AAIH03000072.1	3410716	.	C	T	3.08215