Variant calling from NGS data of two accessions of *Lablab purpureus*

Group-06 members; Immaculate Nahereza, Jane Njeri, Winfred Gatua, Nsangi Olga Tendo, Davis Kiberu, Nsubuga Moses, Eneza Yoel

Supervisors; Jean-Baka Domelevo Entfellner, Oluwaseyi Shorinola, Peter Emmrich

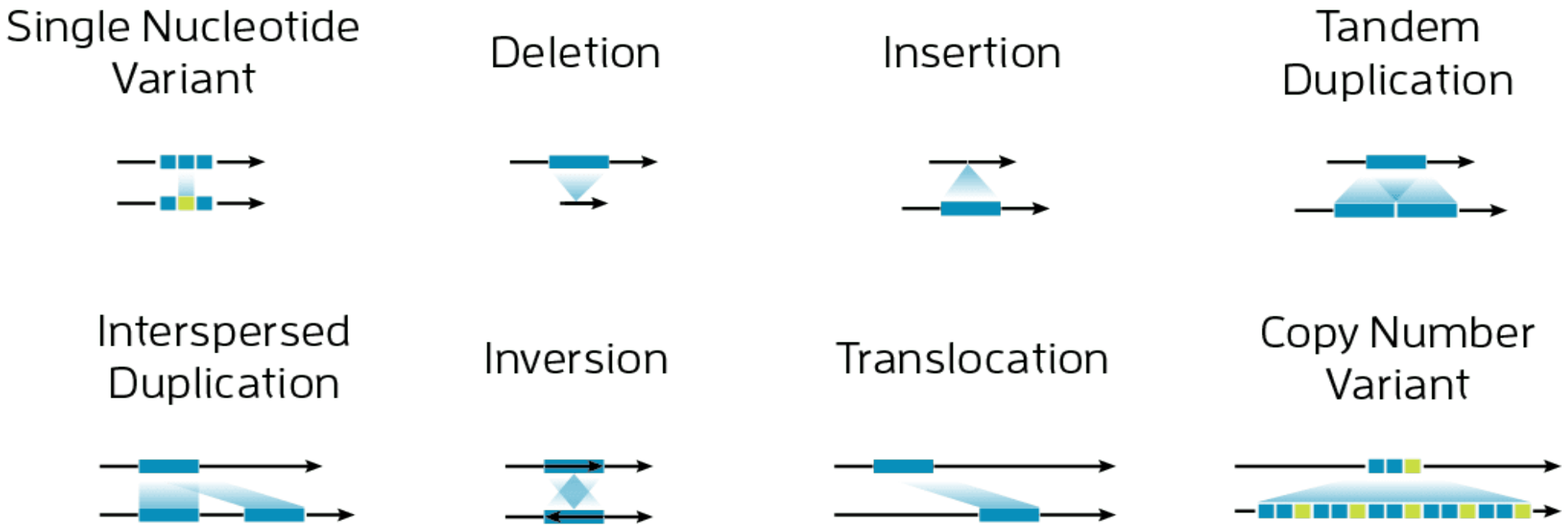
BACKGROUND

- *Lablab purpureus* is a bean (family Fabaceae) commonly known as “lablab” which is native to Africa and widely cultivated in East Africa.

- It is called “Njahi” or black beans in Kenya where it is an important part of the daily diet.

- Lablab is, however, still an orphan crop with limited genomics and genetics resources.

VARIANT TYPES



Types of Variants

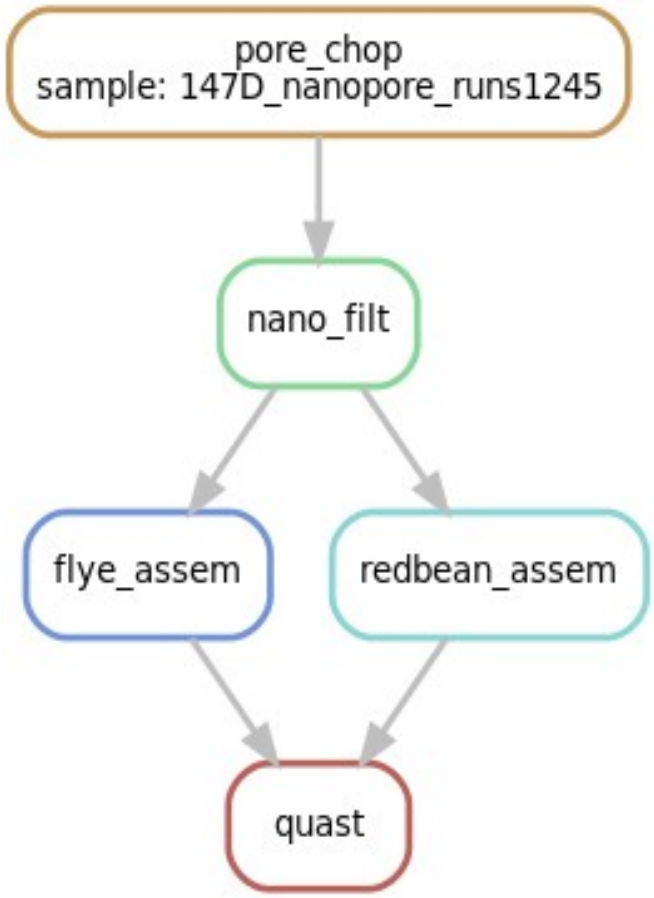
Variant calling is widely used genetics as a way of identifying variants associated with a specific trait, population or hereditary diseases.

OBJECTIVES

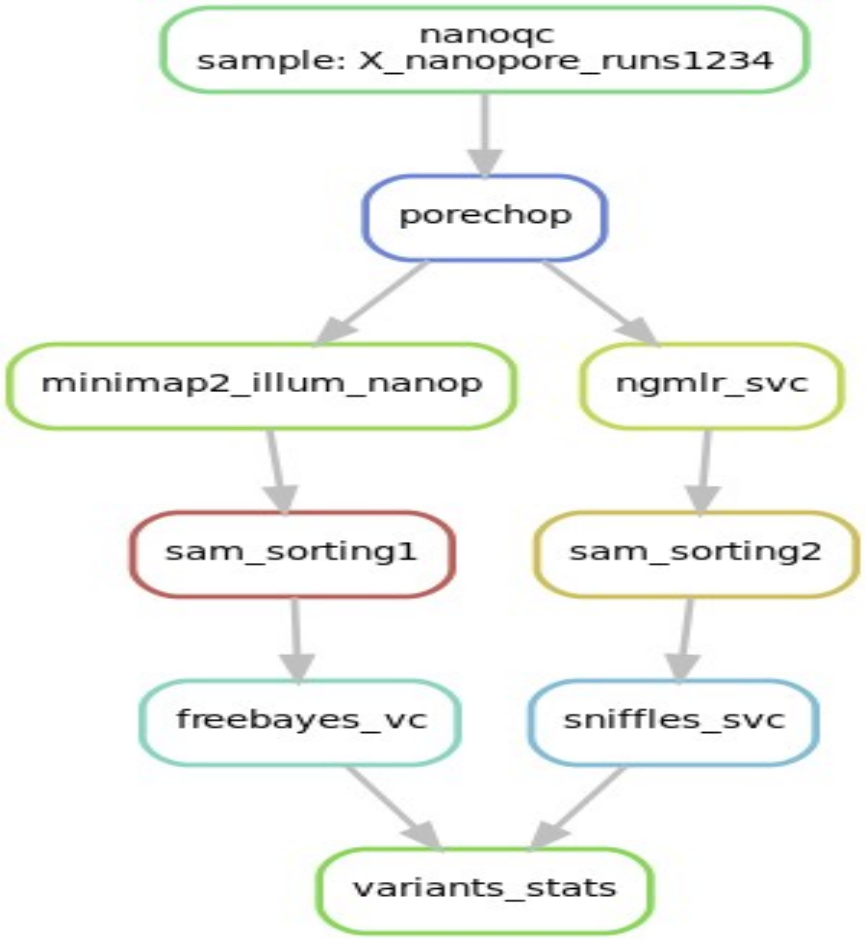
- To do denovo assembly using Illumina based and Oxford nanopore based 147D reads.
- To create a variant calling pipeline for *Lablab purpureus*
- To convert the pipeline to a reproducible and portable snakemake workflow.

PIPELINE

- Assembly



- Variant calling



OBTAINED STATISTICS

```
Location                          : /home/user7/Mini-pr
Failed Filters                    : 3438
Passed Filters                    : 22868
SNPs                              : 0
MNPs                              : 0
Insertions                        : 0
Deletions                         : 411
Indels                            : 10938
Structural variant breakends      : 8061
Symbolic structural variants      : 258
Same as reference                 : 3200
SNP Transitions/Transversions     : - (0/0)
Total Het/Hom ratio               : 1.65 (12256/7412)
SNP Het/Hom ratio                 : - (0/0)
MNP Het/Hom ratio                 : - (0/0)
Insertion Het/Hom ratio           : - (0/0)
Deletion Het/Hom ratio            : 10.11 (374/37)
Indel Het/Hom ratio               : 17.70 (10353/585)
Breakend Het/Hom ratio            : 0.20 (1335/6726)
Symbolic SV Het/Hom ratio         : 3.03 (194/64)
Insertion/Deletion ratio          : 0.00 (0/411)
Indel/SNP+MNP ratio               : - (11349/0)
```

ONGOING ANALYSIS

```
-bash-4.2$  sacct -X
       JobID    JobName      User    Account      State    CPUTime  AllocCPUS  Partition                      NodeList
------------ ---------- --------- ---------- ---------- ---------- ---------- ---------- ------------------------------
702028        freebayes     user6       ilri    RUNNING 19-16:20:44          4      batch compute05
702061       freebayes+     user6       ilri    RUNNING 31-03:48:56          8      batch compute05
702129          Variant     user6       ilri  COMPLETED 1-17:35:36           8      batch compute05
702130        snakeflow     user6       ilri     FAILED   00:00:56           8      batch compute05
702131        snakeflow     user6       ilri     FAILED   00:01:12           8      batch compute05
702132        snakeflow     user6       ilri     FAILED   00:02:24           8      batch compute05
702133          Variant     user6       ilri     FAILED   00:00:00         120      batch None assigned
702134        snakeflow     user6       ilri     FAILED   00:01:44           8      batch compute05
702135          Variant     user6       ilri     FAILED   00:00:00         120      batch None assigned
702136        snakeflow     user6       ilri     FAILED   00:00:52           4      batch compute05
702137          Variant     user6       ilri     FAILED   00:00:00         120      batch None assigned
702138        snakeflow     user6       ilri CANCELLED+   02:19:12           4      batch compute05
702139       snakejob.+     user6       ilri CANCELLED+   00:00:24           8      batch compute05
702140        snakeflow     user6       ilri     FAILED   00:00:08           4      batch compute05
702141        snakeflow     user6       ilri    RUNNING 1-05:33:12          4      batch compute05
702142       snakejob.+     user6       ilri  COMPLETED   16:35:20           4      batch compute05
702145       snakejob.+     user6       ilri    RUNNING   12:57:28           4      batch compute05
-bash-4.2$
```

ONGOING ANALYSIS

```
-bash-4.2$  ls
freebayes_147X_ont_reads.vcf  freebayes_flye_147X_ont_reads.vcf
-bash-4.2$  grep -v '##' freebayes_flye_147X_ont_reads.vcf | less -S

[4]+  Stopped                 grep --color=auto -v '##' freebayes_flye_147X_ont_reads.vcf | less -S
-bash-4.2$  grep -v '##' freebayes_147X_ont_reads.vcf | less -S

[5]+  Stopped                 grep --color=auto -v '##' freebayes_147X_ont_reads.vcf | less -S
-bash-4.2$ ls -alh
total 1.1G
drwxrwxr-x. 2 user6 user6 4.0K Aug 23 21:33 .
drwxrwxr-x. 6 user6 user6 4.0K Aug 24 00:50 ..
-rw-rw-r--. 1 user6 user6 763M Aug 27 19:01 freebayes_147X_ont_reads.vcf
-rw-rw-r--. 1 user6 user6 337M Aug 27 18:54 freebayes_flye_147X_ont_reads.vcf
-bash-4.2$
```

ONGOING ANALYSIS

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL SQL CONSOLE 1: ssh

```
##fileformat=VCFv4.1
##fileDate=20200823
##source=freeBayes v1.0.2-dirty
##reference=/home/user6/Lablab_ref/isaac_147D_flye2.7_hypo-polished_assembly_2020.fa
##phasing=none
##commandline="freebayes -f /home/user6/Lablab_ref/isaac_147D_flye2.7_hypo-polished_assembly_2020.fa /home/user6/Result/Alignment/sorted_flye
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of samples with data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total read depth at the locus">
##INFO=<ID=DPB,Number=1,Type=Float,Description="Total read depth per bp at the locus; bases in reads overlapping / bases in haplotype">
##INFO=<ID=AC,Number=A,Type=Integer,Description="Total number of alternate alleles in called genotypes">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##INFO=<ID=AF,Number=A,Type=Float,Description="Estimated allele frequency in the range (0,1]">
##INFO=<ID=RO,Number=1,Type=Integer,Description="Reference allele observation count, with partial observations recorded fractionally">
```

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	unknown
contig_1	114	.	C	T	5.83425e-14	.			AB=0;ABP=0;AC=0;AF=0;AN=2;AO=275;CIGAR=1X;DP=1249;DPB=1249;DPRA=0;EPP
contig_1	140	.	CTT	CT	1.77436e-13	.			AB=0;ABP=0;AC=0;AF=0;AN=2;AO=249;CIGAR=1M1D1M;DP=1110;DPB=1044.67;DPR
contig_1	246	.	ACC	AC	0	.		AB=0;ABP=0;AC=0;AF=0;AN=2;AO=301;CIGAR=1M1D1M;DP=1252;DPB=1167;DPRA=0;EPP=234	
contig_1	304	.	GAAAAAG	GAAAAG	3323.97	.		AB=0.340094;ABP=287.741;AC=1;AF=0.5;AN=2;AO=436;CIGAR=1M1D5M;DP=1282;DPB=1242	
contig_1	368	.	C	T	0	.		AB=0;ABP=0;AC=0;AF=0;AN=2;AO=363;CIGAR=1X;DP=1209;DPB=1209;DPRA=0;EPP=462.005	
contig_1	374	.	AGG	AG	0	.		AB=0;ABP=0;AC=0;AF=0;AN=2;AO=290;CIGAR=1M1D1M;DP=1254;DPB=1168.67;DPRA=0;EPP=	
contig_1	442	.	AGG	AG	384.568	.		AB=0.346535;ABP=209.623;AC=1;AF=0.5;AN=2;AO=350;CIGAR=1M1D1M;DP=1010;DPB=904.	
contig_1	481	.	AGGGGGA	AGGGGA	375.529	.		AB=0.265748;ABP=487.266;AC=1;AF=0.5;AN=2;AO=270;CIGAR=1M1D5M;DP=1016;DPB=916;	
contig_1	514	.	GAG	GG	1.03184e-13	.		AB=0;ABP=0;AC=0;AF=0;AN=2;AO=264;CIGAR=1M1D1M;DP=1109;DPB=1044.33;DPR	
contig_1	518	.	G	A	0	.		AB=0;ABP=0;AC=0;AF=0;AN=2;AO=335;CIGAR=1X;DP=1131;DPB=1131;DPRA=0;EPP=379.492	
contig_1	534	.	C	T	2.16733e-13	.		AB=0;ABP=0;AC=0;AF=0;AN=2;AO=280;CIGAR=1X;DP=1158;DPB=1158;DPRA=0;EPP	
contig_1	709	.	CTTTTTC	CTTTTC	107.228	.		AB=0.240614;ABP=687.921;AC=1;AF=0.5;AN=2;AO=282;CIGAR=1M1D5M;DP=1172;DPB=1126	

ONGOING ANALYSIS

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL SQL CONSOLE 1: ssh

```
##fileformat=VCFv4.1
##fileDate=20200822
##source=freeBayes v1.0.2-dirty
##reference=/home/user6/Lablab_ref/Lablab_purpureus_147D_AOCC.fa
##phasing=none
##commandline="freebayes -f /home/user6/Lablab_ref/Lablab_purpureus_147D_AOCC.fa --genotype-qualities /home/user6/Result/Alignment/sorted_147
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of samples with data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total read depth at the locus">
##INFO=<ID=DPB,Number=1,Type=Float,Description="Total read depth per bp at the locus; bases in reads overlapping / bases in haplotype">
##INFO=<ID=AC,Number=A,Type=Integer,Description="Total number of alternate alleles in called genotypes">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##INFO=<ID=AF,Number=A,Type=Float,Description="Estimated allele frequency in the range (0,1]">
##INFO=<ID=RO,Number=1,Type=Integer,Description="Reference allele observation count, with partial observations recorded fractionally">
```

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	unknown
scaffold39_cov66	63	.	T	A	0.112701		.		AB=0;ABP=0;AC=2;AF=1;AN=2;AO=2;CIGAR=1X;DP=2;DPB=2;DPRA=0;EPP
scaffold16_cov104	311	.	GTTTTTA	GTTTTA	19.358	.			AB=0;ABP=0;AC=2;AF=1;AN=2;AO=2;CIGAR=1M1D5M;DP=2;DPB=1.71429;DPRA=0;E
scaffold44_cov101	108	.	T	C	1.51034	.			AB=0;ABP=0;AC=2;AF=1;AN=2;AO=2;CIGAR=1X;DP=2;DPB=2;DPRA=0;EPP=7.35324
scaffold44_cov101	123	.	G	A	3.00319	.			AB=0;ABP=0;AC=2;AF=1;AN=2;AO=2;CIGAR=1X;DP=2;DPB=2;DPRA=0;EPP=7.35324
scaffold44_cov101	294	.	T	C	0.303114		.		AB=0;ABP=0;AC=2;AF=1;AN=2;AO=2;CIGAR=1X;DP=2;DPB=2;DPRA=0;EPP
scaffold46_cov71	280	.	A	G	0.516909		.		AB=0;ABP=0;AC=2;AF=1;AN=2;AO=2;CIGAR=1X;DP=2;DPB=2;DPRA=0;EPP
scaffold46_cov71	286	.	TAGAA	TA	1.08481	.			AB=0;ABP=0;AC=2;AF=1;AN=2;AO=2;CIGAR=1M3D1M;DP=2;DPB=0.8;DPRA=0;EPP=3
scaffold46_cov71	299	.	AC	AGC	0.621414		.		AB=0;ABP=0;AC=2;AF=1;AN=2;AO=2;CIGAR=1M1I1M;DP=2;DPB=3;DPRA=0
scaffold46_cov71	331	.	ACG	AG	0.981255		.		AB=0;ABP=0;AC=2;AF=1;AN=2;AO=2;CIGAR=1M1D1M;DP=2;DPB=1.33333;
scaffold46_cov71	400	.	GAAAAAC	GAAAC	0.168958		.		AB=0;ABP=0;AC=2;AF=1;AN=2;AO=2;CIGAR=1M2D4M;DP=2;DPB=1.42857;
scaffold21_cov94	542	.	A	T	1.09823	.			AB=0;ABP=0;AC=0;AF=0;AN=2;AO=2;CIGAR=1X;DP=3;DPB=3;DPRA=0;EPP=3.0103;
scaffold17_cov129	28	.	T	C	10.886	.			AB=0;ABP=0;AC=2;AF=1;AN=2;AO=2;CIGAR=1X;DP=3;DPB=3;DPRA=0;EPP=7.35324

SNAKEMAKE WORKFLOW

- Create an environment with all the packages needed

  $conda env create –-name variant --file Config.yaml (exported env)

- Do a dry run snakemake

  $Snakemake –np

- Run Snakemake

  $Snakemake -j

➢ https://github.com/enezermjema/Mini-project-group-06

LESSONS

- Different tools that are used for Oxford Nanopore reads
- Working with long reads is computationally intensive
- Minimap2 alignment not compatible with sniffles

CHALLENGES

- Computational resources
- Error correction prolonged due to working remotely

Pwani UNIVERSITY

WE BUILD FOR THE FUTURE

Group members

Supervisors; Jean-Baka Domelevo Entfellner,

Oluwaseyi Shorinola,

Peter Emmrich

EANBIT
Eastern Africa Network for Bioinformatics Training

icipe

ILRI
INTERNATIONAL LIVESTOCK RESEARCH INSTITUTE
CGIAR

Thank you!