Benchmark of Common Variant Calling Pipelines vs. the OmicsBox Approach

Enrique Presa Díez[1], Adolfo López-Cerdán[1], Stefan Götz[1]
[1]BioBam Bioinformatics S.L., Valencia, Spain

biobam BIOINFORMATICS SOLUTIONS

Introduction

Motivation

- **Genetic Variant Detection** plays a prominent role in diverse areas, such as biomedical research or plant breeding.
- By comparing the DNA sequences of different individuals, researchers can determine genetic variants and **associate them with a phenotype**.
- As a wide range of tools becomes increasingly accessible, **numerous pipelines** are available for analysis. This study aims to assess the **performance** of different pipelines in comparison to **the cloud-based Genetic Variation Pipeline of Omicsbox**.

Approach

- To compare these pipelines, we evaluate accuracy and runtime in a **GBS dataset**.
- **Accuracy** is determined by comparing the number of consistent genotypes in a **Genotyping-by-Sequencing (GBS)** dataset and a **Whole-Genome Sequencing (WGS)** dataset from the same samples**.**
- Time of execution is calculated considering the **number of CPUs** and **amount of memory used**.

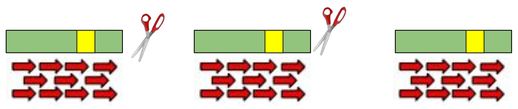
Methods

Input Data

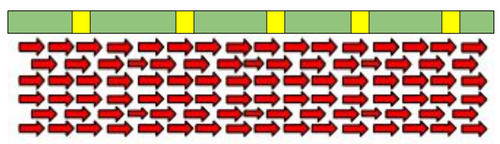
- A total of **24 soybean samples** from the Canadian line were analyzed. The study data consists of 24 single-end FASTQ files for the **GBS dataset** (average size: 82 MB) and 24 paired-end FASTQ files for the **WGS dataset** (average size: 2.32 GB).
- Soybean has a **medium-sized diploid genome** (2n = 40 and 1.1 Gb).
- All samples are **pure lines:** as soybean is an autogamous species, these samples are the result of auto fecundation, so they are **highly homozygous**.

Two Different Sequencing Protocols

Genotype-By-Sequencing (GBS)



Whole-Genome-Sequencing (WGS)



Variant Calling Pipelines

Tool/Pipeline	TASSEL-GBS v2	Stacks	IGST	Fast-GBS	OmicsBox Pipeline
Aligner	BWA-aln	BWA-mem	BWA-aln	BWA-mem	BWA-mem
Variant Caller	DiscoverySNP Caller v2	pstacks	BCFtools	Platypus	BCFtools

Assessment

- **Accuracy**. The accuracy is determined by comparing the agreement of genotypes between a Genotyping-by-Sequencing (GBS) dataset and a Whole-Genome Sequencing (WGS) dataset obtained from the same samples. This parameter **depends on the Variant Calling tools**.
- **Time of execution.** This variable mainly depends on the **computational resources** (CPU and memory) and the **algorithm architecture** (parallelization, distributed computing, load balancing, etcetera).

Short-Read Preprocessing with Trimmomatic

- **Reads adapters** are removed.
- **3' ends** with Phred quality lower than 30 are trimmed.
- Reads with **average Phred quality** lower than 25 are filtered out.

DNA-Seq Alignment with BWA

- Reads are aligned to the Williams 82 reference genome from NCBI.
- Default parameters in OmicsBox are used.

Variant Calling with BCFtools

- The **coefficient for downgrading mapping quality** for reads containing excessive mismatches is 50.
- The **minimum mapping quality** for an alignment to be used is 20.
- The **minimum base quality** for a base to be considered is 20.
- A probabilistic realignment for the computation of base alignment quality **BAQ** is run in all reads.

Variant Filtering

Quality Filtering

- **Depth**: 10
- **Phred Quality**: 20
- **Mapping Quality**: 40
- **Multiple Alleles** are removed.
- **Indels** are also removed.

Genotypic Filtering

- **Genotype Depth**: 2
- **Genotype Quality**: 6
- **Minimum Allele Frequency**: 0.04
- **Maximum heterozygosity**: 0.5

Results

GBS Pipelines	TASSEL-GBS v2	Stacks	IGST	Fast-GBS	OmicsBox Pipeline
SNPs	28158	18941	25650	34953	24284
Heterozygotes (%)	5.7	4.4	5.9	3.4	3.54
CPUs	10	10	10	10	8
Memory (GB)	18	14	240	27	24
Runtime (h:m)	4:16	3:30	12:59	1:47	0:45
Accuracy	92.3%	93.2%	98.4%	98.7%	93.84%

- OmicsBox Pipeline can notably decrease execution time **(2 to 17-fold)** compared to the most common Variant Calling pipelines, while maintaining a **similar number of discovered SNPs**.
- The **proportion of heterozygotes** is analogous to that identified by other algorithms.
- The accuracy of the OmicsBox Pipeline is on par with other pipelines. Nonetheless, since accuracy depends on the Variant Calling tools and algorithm settings, it could be enhanced by incorporating a **more appropriate variant filtering step** to eliminate variants with unreliable genotypes, particularly heterozygous genotypes.

Final Insights

Conclusions

- OmicsBox Genetic Variation Pipeline shows **reduced execution time**, memory usage, and **CPU consumption.**
- The **cloud-based and parallelized design** of OmicsBox Pipeline boosts its efficiency.
- The combined utilization of BWA-mem, BCFtools, and a filtering stage yields accuracy **comparable** to other frequently employed pipelines in variant calling.

Future Perspectives

- Investigate ways to mitigate the issue of calling **heterozygous genotypes**, particularly in the context of the OmicsBox pipeline.
- Optimize the Variant Filtering step to generate a default **filtering profile** that selects only the most trustworthy variants from a **GBS experiment**.

References:

1. Torkamaneh, D., Laroche, J., & Belzile, F. (2016). Genome-wide SNP calling from genotyping by sequencing (GBS) data: a comparison of seven pipelines and two sequencing technologies. PloS one, 11(8), e0161333.
2. OmicsBox - Bioinformatics made easy. BioBam Bioinformatics (Version 2.2.4). September 8, 2022. www.biobam.com/omicsbox.

Powered by: OmicsBox