Journal of Balkan Science and Technology
journal homepage: www.jbst.nku.edu.tr/

A Representative DNA Sequence Analysis with High-Throughput Whole Genome Resequencing Data

Çiğdem KANSU[1*], Jason A. HOLLIDAY[2]

[1] *Tekirdağ Namık Kemal University, Faculty of Arts and Sciences, Department of Biology, Tekirdağ, TÜRKİYE*
[2] *Department of Forest Resources and Environmental Conservation, Virginia Tech, Blacksburg, Virginia, USA*

Technical Note

Keywords:
Next-generation sequencing, Whole-genome resequencing, SNP calling

Received: 21.08.2024
Accepted: 04.10.2024
Published: 30.12.2024

DOI: 10.55848/jbst.2024.44

ABSTRACT

Rapid advances in DNA sequencing technologies made it cost-effective to get high-throughput genomic sequence data. Applications in whole genome resequencing presented new approaches to investigate natural variation and adaptation of species. The genus *Populus* in the Salicaceae family includes economically and ecologically critical species that are dioecious, wind-pollinated, and predominantly distributed in river ecosystems. Poplar serves as a model for the research on the evolutionary genomics of forest tree species. In this study, we aimed to present an example of a comprehensive SNP calling pipeline, demonstrating its application in elucidating genetic diversity and population structure from whole genome resequencing data. The results discovered genetic variants from whole genome resequencing data of poplar species. The pipeline involved sequence preprocessing, alignment to the reference genome, and variant calling.

1. Introduction

Next-generation sequencing (NGS) technologies have greatly transformed genomics by offering exceptional possibilities for efficient and affordable sequencing of whole genomes [1], [2]. The demand for NGS in scientific research has significantly increased due to its capacity to rapidly and precisely produce large quantities of genetic data. The increased information because of numerous complete genome sequences from thousands of species will revolutionize our comprehension of the genetic diversity in natural populations. Whole genome resequencing (WGR) is a process that involves sequencing the complete genetic material of numerous individuals from a certain species, and it is a highly effective method for studying genetic variation throughout an entire genome [3]. The method allows for a thorough identification of variations such as single nucleotide polymorphisms (SNPs), insertions, deletions, and other structural variants [4]. This detailed information is beneficial for population genomics, where understanding genetic diversity and structure both within and between populations is crucial.

One of the most frequently used methods in whole genome resequencing (WGR) is short-read sequencing, which is also called short-read NGS, or second-generation sequencing. It is a massive parallel sequencing method that allows for the rapid and high-throughput sequencing of DNA. This approach is highly efficient and cost-effective, making it a popular choice for large-scale genomic studies. Additionally, short-read sequencing provides high accuracy and coverage, which is essential for detecting genetic variations and understanding complex genomes.

The data obtained from short-read sequencing requires comprehensive bioinformatics analysis to be effectively utilized. Short-read sequencing generates massive amounts of data that include millions of short DNA fragments. Advanced bioinformatics tools and computational methods are used to convert these raw sequences into useful biological interpretations. These analyses involve quality control, alignment to reference genome, and variant calling to identify single nucleotide polymorphisms (SNPs), insertions, deletions, and other structural variants.

Genus Populus in the Salicaceae family has economically and ecologically critical key species that are dioecious, wind-pollinated, and distributed especially in river ecosystems. This genus is used as a model for research studies on the evolution and ecological dynamics of forest tree species and tree form and function research as well [5]. The members of the genus have a crucial worldwide potential in producing industrial wood and raw material. As well as wide phenotypic variation and interspecies hybridization, according to the current adopted classification, the genus is represented by 6 sections (Abaso, Turanga, Leucoides, Aigeiros, Tacamahaca, and Populus) having 29 species [6].

Populus trichocarpa Torrey and Grey, the black cottonwood or western-balsam poplar, is a native North American poplar in the Tacamahaca section. It is the first tree

[*] *Tekirdağ Namık Kemal University, Faculty of Arts and Sciences, Department of Biology, Kampüs Cd. No:1, 59030 Tekirdağ, TÜRKİYE*
E-mail address: ckansu@nku.edu.tr

species whose genome is sequenced [5]. It is accepted as a model tree species owing to its desirable biological characteristics, such as relatively small genome size, straightforward genetic transformation, the tendency of vegetative propagation, quick responses to biotic and abiotic stress, and shorter generation time compared to other forest tree species [7]. *Populus balsamifera* L., balsam poplar, is another species in the Tacamahaca section. These sister species, which morphologically resemble, are quite different ecologically in the way that they show adaptation to contrasting environmental conditions. *P. trichocarpa* is distributed in a rather humid and temperate range from northern California to southern Alaska while *P. balsamifera* is a cold-tolerant boreal species whose range extends from Alaska to Maine [8].

The goal of this paper is to provide a guideline for researchers to gain an understanding of DNA sequencing analysis. For this purpose, the whole genome resequencing of poplar species was used. The whole genome resequencing data of *P. trichocarpa* was used together with its widely distributed relative *P. balsamifera* to perform a representative population genomics analysis. The analyses performed for this study were presented as an example guideline for SNP calling and an example of population genomics. Sequence reads for the whole project have been archived on NCBI (PRJNA996882). The detailed analysis of the project results was presented in Bolte et al. [9].

2. Material And Method

2.1. DNA extraction and Sequencing

Vegetative branch cuttings were collected from 576 trees from seven latitudinal transects in *Populus trichocarpa* and *Populus balsamifera* distribution range (from 40° N to 65° N and −100° W to −150° W). They were propagated in greenhouse conditions. Young leaf tissue was sampled and DNA was extracted from approximately 100 mg of leaf tissue using the Qiagen plant DNeasy kit with a modified method [9].

Genomic DNA libraries were constructed at the Duke University Center for Genomic and Computational Biology, using an Illumina DNA Prep kit (Illumina Inc., San Diego, USA). The whole-genome resequencing had been performed on an Illumina NovaSeq 6000 instrument with the S4 flow cell in 2x150bp format.

2.2. Bioinformatics

All the analyses carried out on the Virginia Tech Advanced Research Computing systems (ARC). ARC is a high-performance computing (HPC) resource and service for Virginia Tech researchers. It offers a variety of CPU, GPU, and high-throughput computing clusters (https://arc.vt.edu/). This is basically a combination of computing power to obtain a greater performance than a typical desktop or workstation. HPC uses clusters of powerful processors, which are working in parallel, for big datasets and tries to solve complex problems at extremely high speeds.

2.2.1. DNA Sequencing Pipeline

After sequencing, first analysis is typically conducted by the instrument's software immediately. This process includes base-calling for each clonally amplified DNA fragment. During this phase, quality control steps, such as read filtering and trimming, are also performed. Sequence data is recorded along with the quality scores (Phred values) and the result is a FASTQ format (Fig. 1).

As the first phase of the study, it was aimed to discover the variants in the results of whole genome resequencing for the samples. This includes a hierarchical pipeline for the SNP discovery consisting of stepwise analysis of Next Generation Sequencing (NGS) data. The general view of the pipeline for the SNP calling from fastq files is given in Fig. 2.

Sequence preprocessing was done with a custom pipeline for de-indexing, QC (Quality Control), and trimming adapter sequences. The resulting fastq files were aligned to the reference genome (Ptrichocarpa_533_v4.0.fa.gz, in Phytozome database) using Burrows-Wheeler Aligner (BWA-MEM). A representative script for this reference genome alignment is given in Fig. 3. This is also an example of a slurm script used to submit a job to an HPC cluster. All the analyses done in this study were submitted as a slurm job to the cluster. Slurm is a cluster management and job scheduling system (for details see webpage of slurm workload manager).

The resulting SAM files were converted to BAM format with Samtools (Fig. 4)[10], [11]. Samtools is a software dealing with high-throughput sequencing data. SAM stands for Sequence Alignment/Map. It is a human-readable text format to store alignment data in a series of tab-delimited ASCII columns. It is then converted into BAM, which is SAM file's binary equivalent and stands for Binary Alignment/Map. BAM files are compressed versions and are not human-readable. After reference genome alignment, the resulting BAM files should be sorted according to the physical location of the sequences and then indexed. This indexing could be thought of as an index of a book. The subsequent genotyping tools in the pipeline should have info on the positions of the sequences in the BAM file so that they can call the genotypic variants present in the files. The BAI format created after indexing is the indexed form of BAM files. This sorting and indexing are done with Samtools.

Fig. 1 An example of fastq file format.

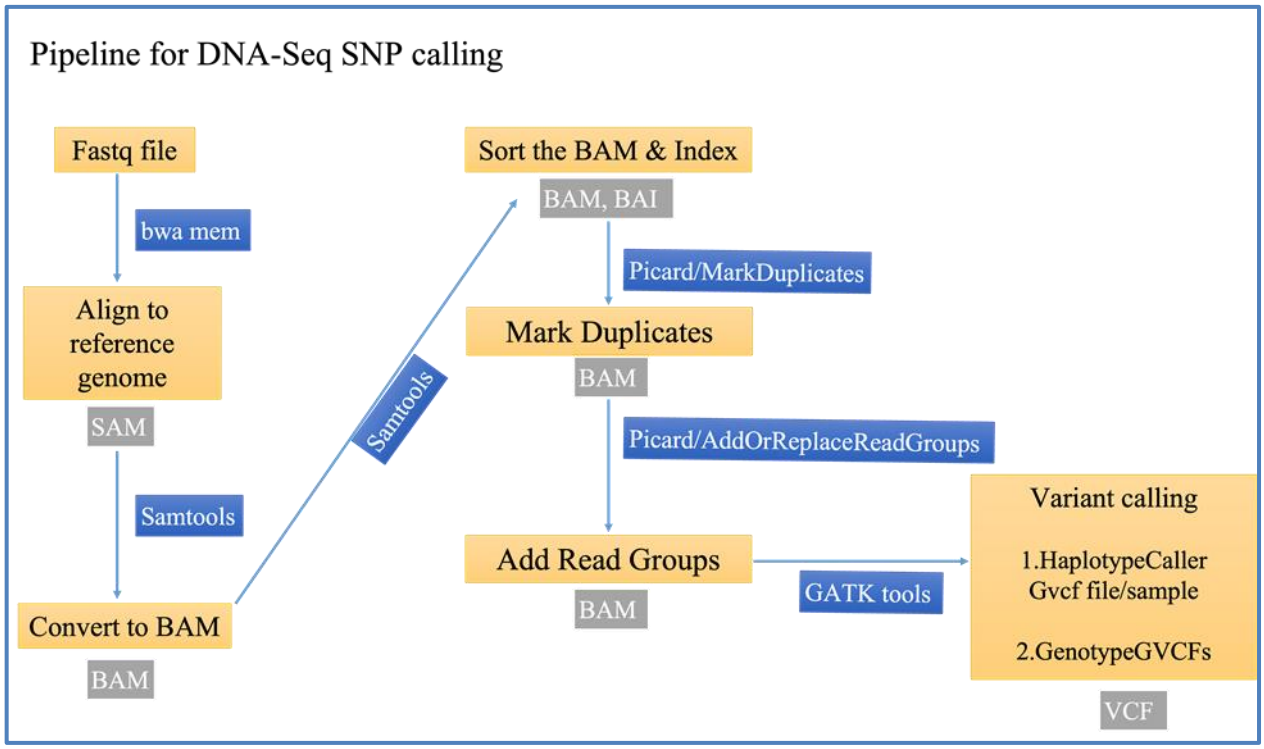

Fig. 2 The pipeline for SNP calling with whole-genome resequencing data.

```
#!/bin/bash

#SBATCH -J bwa_align
#SBATCH -t 6:00:00
#SBATCH --nodes=1
#SBATCH --ntasks-per-node=1
#SBATCH --cpus-per-task=16
#SBATCH --mem=12gb
#SBATCH -p normal_q
#SBATCH -A Poplar
#SBATCH --mail-type=begin
#SBATCH --mail-type=end
#SBATCH --mail-user=cigdemkansu@vt.edu
#SBATCH --output=bwa_align_%j.out
#SBATCH --error=bwa_align_%j.err

ID="Ptri"
SM="Sample"
LB="Lib1"
PU="unit1"
PL="ILLUMINA"
PROJ_PATH="/projects/jah1_lab/cigdem"
IND="/projects/jah1_lab/cigdem/poplar_genome_jason/Ptrichocarpa_533_v4.0.fa"
R1="/projects/jah1_lab/cigdem/results/Ptri01_R1.fastq"
R2="/projects/jah1_lab/cigdem/results/Ptri01_R2.fastq"

cd /projects/intro2gds/IndividualFolder/cigdem/software/bwa-0.7.17
export PATH=$PATH:/projects/jah1_lab/cigdem/software/bwa-0.7.17
bwa mem -R '@RG\tID:$ID\tSM:$SM\tLB:$LB\tPU:$PU\tPL:$PL' -t 16 ${IND} ${R1} ${R2} > ${PROJ_PATH}/results/Ptri01_RG.sam
```

Slurm commands: slurm is an open source cluster management and job scheduling system for Linux clusters

Fig. 3 A representative BWA-MEM slurm script for the reference genome alignment of paired-end sequencing fastq files for an individual.

Polymorphisms (SNPs and INDELs) were called using the Genome Analysis Toolkit (GATK) HaplotypeCaller algorithm [12], [13], with the resulting VCF file filtered to yield a final high-quality dataset. The scripts for the following analysis were presented in Fig. 5. The duplicates generated during the library construction or sequencing [14] were removed by MarkDuplicates (Picard), which is a Genome Analysis Tool Kit (GATK) tool used to determine the duplicates and to remove them if needed [15]. The call of SNPs in the BAM files was done with the HaplotypeCaller, one of the GATK tools [16]. All HaplotypeCaller GVCF files were combined with another GATK tool, CombineGVCFs [17]. As a last step in the SNP calls is the genotyping of the samples pre-called with HaplotypeCaller. This is performed by GenotypeGVCFs, another tool from GATK [18]. The bash script for this step is shown in Fig. 6.

As a result of this step in the pipeline, a VCF file was obtained. VCF stands for Variant Call Format, and it is a tab-delimited text file used to store genetic variation in sequencing results [19]. It contains a header consisting of information on the organism, genome build version, etc., as well as definitions of the annotations used in the VCF file. After the header, each line represents a single variant with columns including variant information describing SNPs, indels, and other variation types as well (fig 7).

The resulting the vcf file in this study includes *P. trichocarpa* (interior and coastal populations), *P. balsamifera*, *Populus angustifolia*, *Populus deltoides* and hybrids of these in the contact zones. This mixed-species data set was first subsetted to 10,000 random SNPs using the awk program in bash scripting for the ease of computation in downstream analyses as will be presented as an example pipeline.

```
#!/bin/bash
#bwa-mem alignment(just for 1 individual)
ID="Ptri"
SM="Sample"
LB="Lib1"
PU="unit1"
PL="ILLUMINA"
PROJ_PATH="/projects/jah1_lab/cigdem"
IND="/projects/jah1_lab/cigdem/poplar_genome_jason/Ptrichocarpa_533_v4.0.fa"
R1="/projects/jah1_lab/cigdem/results/Ptri01_R1.fastq"
R2="/projects/jah1_lab/cigdem/results/Ptri01_R2.fastq"

cd /projects/intro2gds/IndividualFolder/cigdem/software/bwa-0.7.17
export PATH=$PATH:/projects/jah1_lab/cigdem/software/bwa-0.7.17
bwa mem -R '@RG\tID:$ID\tSM:$SM\tLB:$LB\tPU:$PU\tPL:$PL' -t 16 \
${IND} ${R1} ${R2} > ${PROJ_PATH}/results/Ptri01_RG.sam

#samtools convert to bam file
module load SAMtools/1.11-GCC-10.2.0
cd /projects/jah1_lab/cigdem/results
samtools view -S -b Ptri01_RG.sam > Ptri01_RG.bam

#samtools sorting the alignment
module load SAMtools/1.11-GCC-10.2.0
cd /projects/jah1_lab/cigdem/results
samtools sort Ptri01_RG.bam -o Ptri01_RG_sorted.bam

#samtools indexing bam files
module load SAMtools/1.11-GCC-10.2.0
cd /projects/jah1_lab/cigdem/results
samtools index Ptri01_RG_sorted.bam Ptri01_RG_sorted.bai
```

Fig. 4 The bash scripts of the variant calling pipeline for 1 individual, including alignment, bam file conversion and indexing.

2.2.2. Population genomics with R

For the partial population genomics analysis, the VCF file was read with the vcfR package. The format of the R object was converted to different data structures that are compatible with different packages for population genomics analyses. The summary statistics and genetic differentiation were calculated with the hierfstat package in R [20]. To screen the genetic differentiation between species' populations, F_{ST} values were compared among species with the hierfstat package. To infer population structure by determining the number of clusters, a multivariate method, Discriminant Analysis of Principal Components (DAPC), was used without prior knowledge of the structuring [21] with the adegenet package in R [22].

```
#picard marking duplicates
module load picard/2.21.6-Java-11
cd /projects/jah1_lab/cigdem/results

java -Xmx2g -XX:ParallelGCThreads=5 -jar $EBROOTPICARD/picard.jar MarkDuplicates \
      INPUT=Ptri01_RG_sorted.bam \
      OUTPUT=Ptri01_RG_sorted_mardup.bam \
      METRICS_FILE=Ptri01_mardup_metrics.txt \
          ASSUME_SORTED=True

#samtools indexing again after marking duplicates
module load SAMtools/1.11-GCC-10.2.0
cd /projects/jah1_lab/cigdem/results
samtools index Ptri01_RG_sorted_mardup.bam Ptri01_RG_sorted_mardup.bai

#gatk haplotype calling-getting gvcf files (only for 1 file, otherwise use parallel)
PROJ_PATH="/projects/jah1_lab/cigdem"
REF="/projects/jah1_lab/cigdem/poplar_genome_jason/Ptrichocarpa_533_v4.0.fa"
BAM="/projects/jah1_lab/cigdem/results/Ptri01_RG_sorted_mardup.bam"
export PATH="/projects/jah1_lab/cigdem/software/gatk-4.2.6.1/:$PATH"

gatk --java-options "-Xmx5g -XX:+UseParallelGC" HaplotypeCaller \
-R ${REF} \
-I ${BAM} \
-O ${PROJ_PATH}/results/Ptri01_variants.g.vcf \
    -ERC GVCF

#gatk combining gvcf files(needed if >1 individuals)
PROJ_PATH="/projects/jah1_lab/cigdem"
REF="/projects/jah1_lab/cigdem/poplar_genome_jason/Ptrichocarpa_533_v4.0.fa"
GVCF="/projects/jah1_lab/cigdem/results/Ptri01_variants.g.vcf"
export PATH="/projects/jah1_lab/cigdem/software/gatk-4.2.6.1/:$PATH"

gatk --java-options "-Xmx7g" CombineGVCFs \
    -R $REF \
    -V $GVCF \
    -O $PROJ_PATH/results/cohort_Ptri01_variants.g.vcf.gz
```

Fig. 5 The bash scripts of the variant calling pipeline for 1 individual, including marking duplicates, calling variants and creating a combined gvcf file.

```
#gatk genotyping gvcfs-getting vcf files
PROJ_PATH="/projects/jah1_lab/cigdem"
REF="/projects/jah1_lab/cigdem/poplar_genome_jason/Ptrichocarpa_533_v4.0.fa"
CGVCF="/projects/jah1_lab/cigdem/results/cohort_Ptri01_variants.g.vcf.gz"

export PATH="/projects/jah1_lab/cigdem/software/gatk-4.2.6.1/:$PATH"

gatk --java-options "-Xmx7g" GenotypeGVCFs \
    -R $REF \
    -V $CGVCF\
    -O $PROJ_PATH/results/Ptri01_variants.vcf.gz
```

Fig. 6 The bash script of genotyping the combined gvcf files for all samples to obtain a vcf file.

3. Results and Discussion

According to admixture analysis (data not shown), there were four genetic clusters identified. Thus, in the remainder of this paper, the results will be discussed on four taxa (*P. balsamifera*, interior *P. trichocarpa*, coastal *P. trichocarpa*, a population including *Populus angustifolia* and *Populus deltoides*) and their possible hybrids.

Estimating heterozygosity is one of the ways that provides insights into the genetic diversity within and between populations. Heterozygosity, defined as the proportion of individuals in a population that are heterozygous at a given

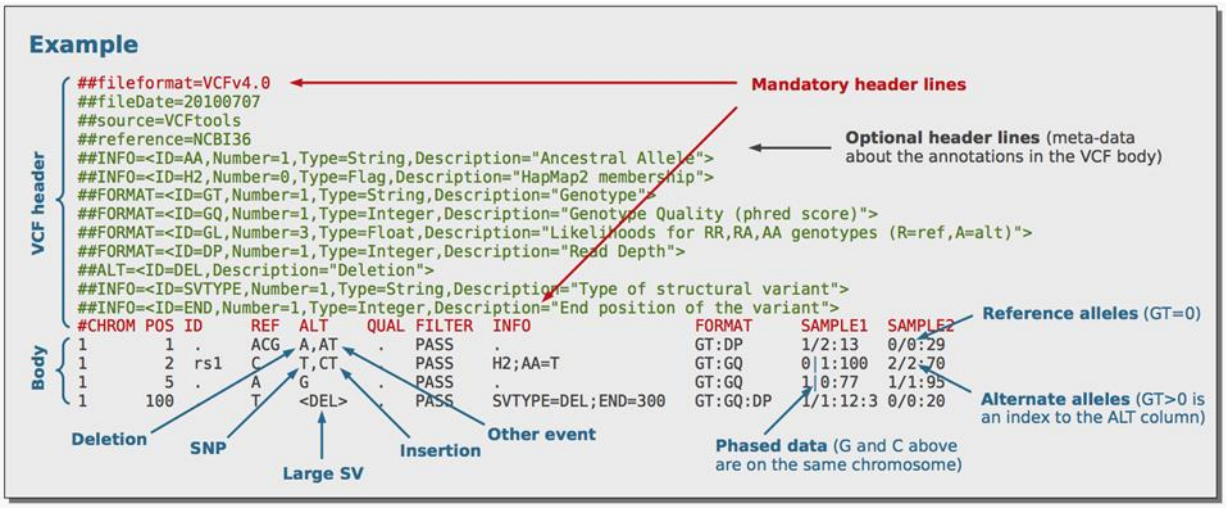

Fig. 7 An example of a vcf file format. Reprinted from [19].

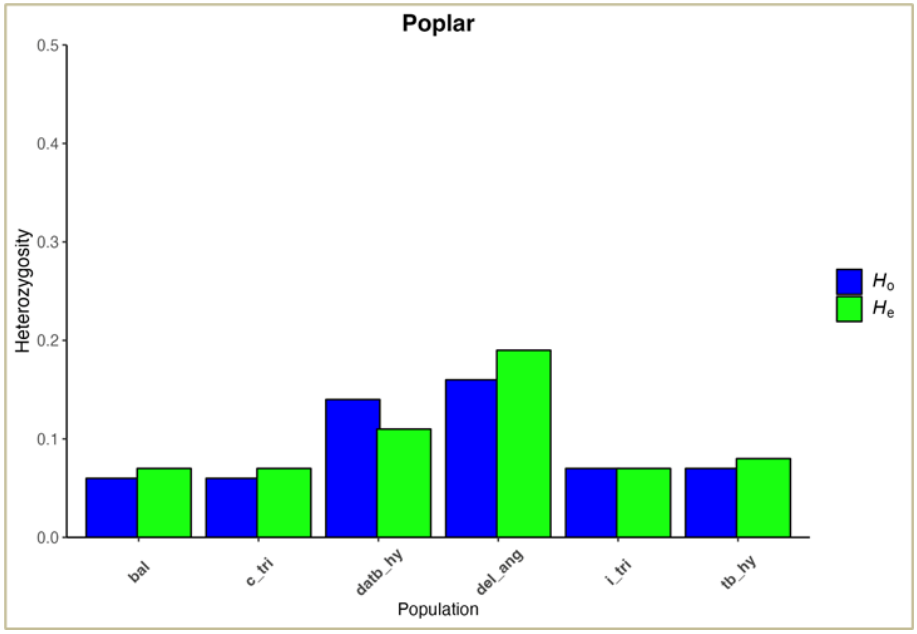

Fig. 8 The histogram of H_o (observed heterozygosity) and H_e (expected heterozygosity) values for each taxon. (c_tri: coastal *P. trichocarpa*, i_tri: interior *P. trichocarpa*, bal: *P. balsamifera*, del_ang: *P.deltoides* and *P.angustifolia*, tb_hy: *P. trichocarpa* x *balsamifera* hybrids, datb_hy: possible hybrids of all species).

locus, serves as an indicator of genetic variation. Barplots of the distribution of observed and expected heterozygosity for each group was given in Fig. 8. The greatest heterozygosity values were observed for *P. deltoides* and *P. angustifolia* group, indicating substantial genetic variation. *P. balsamifera* exhibited lowest heterozygosity compared to others. When we compared interior and coastal *P. trichocarpa*, they showed a comparable level of heterozygosity to their hybrids with *P. balsamifera*. Hybrid populations typically show increased heterozygosity compared to their parental populations because they inherit diverse genetic material from both parent species. However, backcrosses to their parent species can reduce heterozygosity.

F_{ST} is a measure of population differentiation based on genetic structure. As the value of F_{ST} decreases, the differentiation between the populations scales down. The F_{ST} values ranged between 0.047 and 0.467 (Fig. 9). It was shown that the most differentiated populations are interior *P. trichocarpa* and *P. deltoides* and *P. angustifolia* group as expected because their distribution range is distant than the others. On the other hand, the least differentiated ones are coastal *P. trichocarpa* and *P. balsamifera*. It was intriguing that the hybrids of *P. trichocarpa* and *P. balsamifera* are more differentiated from the interior *P. trichocarpa* population. The more maritime climate on the western side of the Rocky Mountains in the contact zone of the two species favors *P. trichocarpa* but interior *P. trichocarpa* populations have a southern, interior distribution. Thus there could be a geographic isolation reducing the gene flow from interior *P. trichocarpa* populations and together with the genetic drift, this would allow for greater differentiation as the hybrids continue to accumulate genetic differences.

It is suggested to use model-free methods like DAPC as a more convenient approach for populations that are clonal or partially clonal [21]. DAPC is particularly advantageous because it does not rely on assumptions about Hardy-Weinberg equilibrium or linkage equilibrium. In this approach, basically, variance in the samples is partitioned into a between-group and within-group component. This maximizes discrimination between groups. DAPC transforms data using a principal components analysis (PCA) and then it determines the clusters by using discriminant analysis (DA). The resulting DAPC plot is shown in Fig 10.

Axis 1 (47.9% of the variance) captured the majority of the genetic differentiation. *P. deltoides* and *P. angustifolia* group is significantly separated from all other clusters, indicating it is genetically distinct. Possible hybrids of all species are well-separated from other hybrids and *P. trichocarpa* and *P. balsamifera*, indicating distinct genetic differentiation. The overlap between *P. trichocarpa* x *balsamifera* hybrids and coastal and interior *P. trichocarpa* suggested close relation as expected.

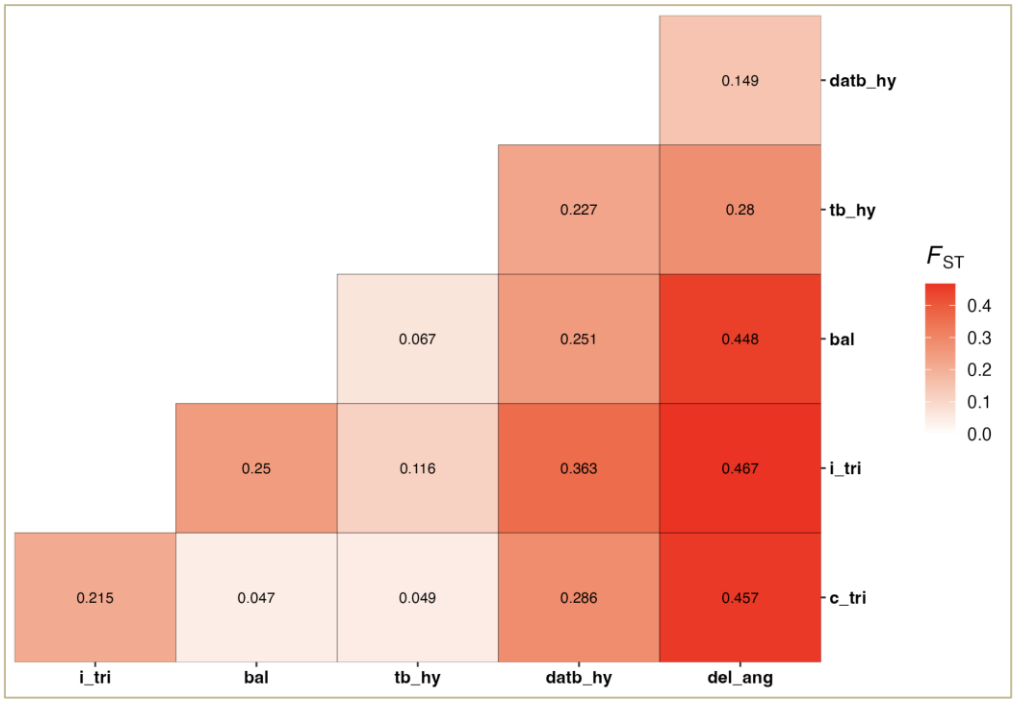

Fig. 9 Pairwise F_{ST} heatmap. (c_tri: coastal *P. trichocarpa*, i_tri: interior *P. trichocarpa*, bal: *P. balsamifera*, del_ang: *P.deltoides* and *P.angustifolia*, tb_hy: *P. trichocarpa* x *balsamifera* hybrids, datb_hy: possible hybrids of all species).

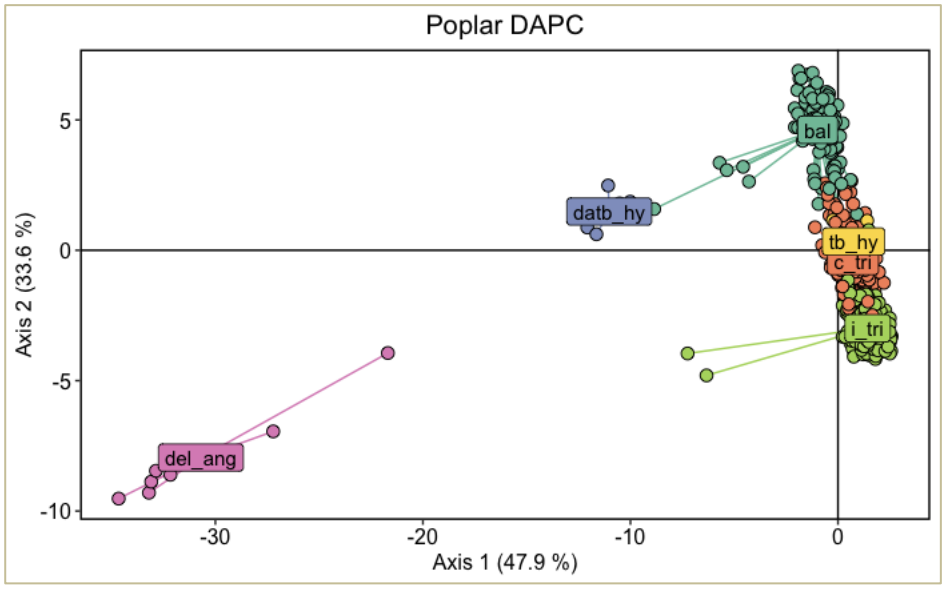

Fig. 10 DAPC plot of Poplar populations. (c_tri: coastal *P. trichocarpa*, i_tri: interior *P. trichocarpa*, bal: *P. balsamifera*, del_ang: *P.deltoides* and *P.angustifolia*, tb_hy: *P. trichocarpa* x *balsamifera* hybrids, datb_hy: possible hybrids of all species).

4. Conclusion

In summary, this study presented an example pipeline of SNP calling for a population genomics study. The advancement in next-generation sequencing leads to increase in whole-genome resequencing studies, which in turn enables generating high-throughput genomics data. Efficient and effective sequencing data analysis requires a straight-forward approach, and the pipelines should include appropriate software. Although there exists several tools and platforms for genomic analysis pipelines, knowledge of the steps in the bioinformatics has great importance in genomics studies.

Declaration

Author Contribution: Conceive– J.A. Holliday; Design– Ç. Kansu; Experimental Performance, Data Collection and Processing– Ç. Kansu; Analysis and Interpretation– Ç. Kansu; Literature Review– Ç. Kansu, J.A. Holliday; Writer– Ç. Kansu, J.A. Holliday; Critical Review– Ç. Kansu, J.A. Holliday.

Conflict of interests: The authors have declared no conflicts of interest.

Acknowledgements

This work was supported by National Science Foundation grant PGR-1856450, USDA National Institute of Food and Agriculture project and Hatch Appropriations (PEN04809 and Accession 7003639), NIFA project VA-136641, Schatz Center for Tree Molecular Genetics. Çiğdem Kansu was funded by the TUBITAK 2219 International Postdoctoral Research Fellowship.

Orcid-ID

Çiğdem Kansu https://orcid.org/0000-0002-0921-2881
Jason A. Holliday https://orcid.org/0000-0002-2662-8790

References

[1] M. Mardis, "Next-generation DNA sequencing methods," Annual Review of Genomics and Human Genetics, vol. 9, pp. 387-402, 2008. https://doi.org/10.1146/annurev.genom.9.081307.164359

[2] E. van Dijk, H. Auger, Y. Jaszczyszyn, and C. Thermes, "Ten years of next-generation sequencing technology," Trends in Genetics, vol. 30, no. 9, pp. 418-426, 2014. https://doi.org/10.1016/j.tig.2014.07.001

[3] J. Cao, K. Schneeberger, S. Ossowski, et al., "Whole-genome sequencing of multiple Arabidopsis thaliana populations," Nature Genetics, vol. 43, pp. 956-963, 2011. https://doi.org/10.1038/ng.911

[4] B. Hunter, K.M. Wright, K. Bomblies, " Short read sequencing in studies of natural variation and adaptation" Curr Opin Plant Biol, vol. 16, no. 1, pp. 85-91, 2013. https://doi.org/10.1016/j.pbi.2012.10.003

[5] J. Tuskan, S. DiFazio, S. Jansson, et al., "The genome of black cottonwood, Populus trichocarpa (Torr. & Gray)," Science, vol. 313, pp. 1596-1604, 2006. https://doi.org/10.1126/science.1128691

[6] J. Eckenwalder, "Systematics and evolution of Populus," in Biology of Populus and its implications for management and conservation, R. F. Stettler, H. D. Bradshaw, P. E. Heilman, and T. M. Hinckley, Eds. Ottawa, ON, Canada: NRC Research Press, 1996, pp. 7-32.

[7] Ellis, B, S Jansson, SH Strauss, GA Tuskan, " Why and how Populus became a “Model Tree" in Genetics and Genomics of Populus, S. Jansson, R. Bhalerao, A. Groover, Eds. Plant Genetics and Genomics: Crops and Models, vol 8. Springer, New York, NY, 2010,pp. 1025-1041. https://doi.org/10.1007/978-1-4419-1541-2_10

[8] J. Richardson, J.G. Isebrands, " Ecology and physiology of poplars and willows," in Poplars and willows: trees for society and the environment, JG Isebrands, J Richardson, Eds. Oxfordshire, UK: CABI, 2014,pp. 92–123. https://doi.org/10.1079/9781780641089.0092

[9] C.E. Bolte, T. Phannareth, M. Zavala-Paez, et al., " Genomic insights into hybrid zone formation: The role of climate, landscape, and demography in the emergence of a novel hybrid lineage " Molecular Ecology, 33:e17430, 2024. https://doi.org/10.1111/mec.17430

[10] H. Li and R. Durbin, "Fast and accurate short read alignment with Burrows-Wheeler transform," Bioinformatics, vol. 25, no. 14, pp. 1754-1760, 2009. https://doi.org/10.1093/bioinformatics/btp324

[11] H. Li, B. Handsaker, A. Wysoker, et al., "The Sequence Alignment/Map format and SAMtools," Bioinformatics, vol. 25, no. 16, pp. 2078-2079, 2009. https://doi.org/10.1093/bioinformatics/btp352

[12] A. McKenna, M. Hanna, E. Banks, et al., "The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data," Genome Research, vol. 20, no. 9, pp. 1297-1303, 2010. https://doi.org/10.1101/gr.107524.110

[13] R. Poplin, V. Ruano-Rubio, M. A. DePristo, et al., "Scaling accurate genetic variant discovery to tens of thousands of samples," bioRxiv, 2017. https://doi.org/10.1101/201178

[14] J. R. Ebbert, M. A. Wadsworth, M. H. Staley, et al., "Evaluating the necessity of PCR duplicate removal from next-generation sequencing data and a comparison of approaches," BMC Bioinformatics, vol. 17, no. 1, pp. 239, 2016. https://doi.org/10.1186/s12859-016-1097-3

[15] "Webpage of GATK tool MarkDuplicates," GATK Website. [Online]. Available: https://gatk.broadinstitute.org/hc/en-us/articles/360036227892-MarkDuplicates. [Accessed: 14-Aug-2024].

[16] "Webpage of HaplotypeCaller," GATK Website. [Online]. Available: https://gatk.broadinstitute.org/hc/en-us/articles/360036227912-HaplotypeCaller. [Accessed: 14-Aug-2024].

[17] "Webpage of CombineGVCFs," GATK Website. [Online]. Available: https://gatk.broadinstitute.org/hc/en-us/articles/360036227852-CombineGVCFs. [Accessed: 14-Aug-2024].

[18] "Webpage of GenotypeGVCFs," GATK Website. [Online]. Available: https://gatk.broadinstitute.org/hc/en-us/articles/360036227832-GenotypeGVCFs. [Accessed: 14-Aug-2024].

[19] Tang, D. (2024). Learning vcf file. GitHub. [Online]. Available: https://github.com/davetang/learning_vcf_file?tab=readme-ov-file#learning-the-vcf-format. [Accessed: 14-Aug-2024].

[20] J. Goudet, "Hierfstat, a package for R to compute and test hierarchical F-statistics," Molecular Ecology Notes, vol. 5, no. 1, pp. 184-186, 2005. https://doi.org/10.1111/j.1471-8286.2004.00828.x

[21] T. Jombart, S. Devillard, and F. Balloux, "Discriminant analysis of principal components: a new method for the analysis of genetically structured populations," BMC Genetics, vol. 11, no. 1, pp. 94, 2010. https://doi.org/10.1186/1471-2156-11-94

[22] T. Jombart, "adegenet: a R package for the multivariate analysis of genetic markers," Bioinformatics, vol. 24, no. 11, pp. 1403-1405, 2008.