NGS Data Processing and Analysis

Büşra Özdemir, 150200036
Aslı Yel, 150200054

***Abstract*—This project report presents an in-depth exploration of Next-Generation Sequencing (NGS) technology and the bioinformatics tools used for NGS data processing and analysis. Various topics, including DNA sequencing, variant calling, alignment, mapping, and reference genomes, are covered. CoSAP, a comparative sequencing analysis platform, was employed to analyze NGS data and evaluate the performance of different pipelines. The report includes detailed comparisons of pipeline configurations, highlighting the importance of selecting the appropriate configuration for NGS data analysis. Evaluation of pipeline performance utilized metrics like precision, recall, F1 score, and accuracy. Additionally, the report explores the use of principal component analysis, Jaccard similarity, heatmap, and clustering techniques for data visualization and analysis. Overall, this project provides valuable insights into NGS data processing and analysis, and the bioinformatics tools used in this field.**

***Index Terms*—Bioinformatics, NGS, Pipeline, DNA Sequencing, Data Processing, Dataset, Variant Calling, Mutect, SomaticSniper Aligment, Mapping, CoSAP, Reference Genome, FASTQ, BAM, BED, Presicion, Recall, F1 Score, Accuracy, Principle Component, Jaccard, Heatmap, Clustering**

I. Introduction

A. DNA Sequencing

DNA sequencing is a common laboratory technique used to determine the precise order of nucleotides(adenine, thymine, cytosine, and guanine), or bases, in a DNA molecule. The arrangement of these bases (commonly denoted by the first letters of their chemical names: A, T, C, and G) contains the biological information essential for cell development and functioning. Establishing the DNA sequence is essential for understanding the functions of genes and other elements within the genome. Presently, various methods exist for DNA sequencing, each possessing distinct characteristics, and ongoing genomics research actively explores the development of additional sequencing techniques. Next-Generation Sequencing (NGS), represents a significant advancement in DNA sequencing technology.[1]

Next-Generation Sequencing (NGS): NGS is a general term referring to various massively parallel and high-throughput sequencing technologies. In comparison to alternative sequencing methods, NGS is significantly faster and more cost-effective, but it also relies on sophisticated bioinformatics tools for data analysis.[2]

B. NGS Data Processing Algorithms: Bioinformatics Pipelines

NGS produces vast datasets, demanding the execution of numerous computationally intensive procedures to ensure proper analysis. The execution of bioinformatics algorithms in a predefined sequence to process NGS data is commonly known as the NGS bioinformatics pipeline. A bioinformatics pipeline systematically guides and processes extensive sequence data along with their corresponding metadata through a sequence of modifications, utilizing various software components, databases, and operational environments. NGS bioinformatics pipelines are commonly designed for specific platforms and can be adapted to suit the particular requirements of a laboratory.[2] The major steps of an NGS bioinformatics pipeline are as follows.

Quality Control: Assess the quality of raw sequencing data to identify and remove low-quality reads. Quality control is a critical step in an NGS bioinformatics pipeline, involving the assessment and assurance of data quality to ensure the reliability of downstream analyses.

Alignment/Mapping: Mapping or alignment is the process of aligning short DNA sequences (reads) obtained from NGS to a reference genome.

Variant Calling: Variant calling involves identifying variations, such as single nucleotide polymorphisms (SNPs) or insertions/deletions (Indels), by comparing aligned reads to a reference genome.

Variant Annotation: Variant annotation is the process of adding biological information to detected variants to understand their potential functional impact.

C. Comparative Sequencing Analysis Platform (CoSAP)

Biotools are typically written by experts, as commercializing bioinformatics programs is uncommon. This is mainly due to the limited user base with highly specific demands. A platform, namely CoSAP, with numerous biotools is used along this research.

CoSAP is a simple yet feature-rich tool for creating pipelines for NGS data. It offers reproducibility and, by letting users compare the output of several pipelines, seeks to provide better knowledge about the capabilities and constraints of the available technologies.[3]

Read trimming, read mapping, duplicate removal and base calibration, variant calling, and variant annotation are the processes that make up a typical variant calling pipeline. CoSAP offers multiple tool choices for every one of these stages.[3]

II. Materials and Methods

A. Hardware Specifications

- **Operating System:** Windows 10 Pro
- **Model:** Monster Abra A7 V11.2
- **RAM:** 16 GB
- **Storage:** 500 GB

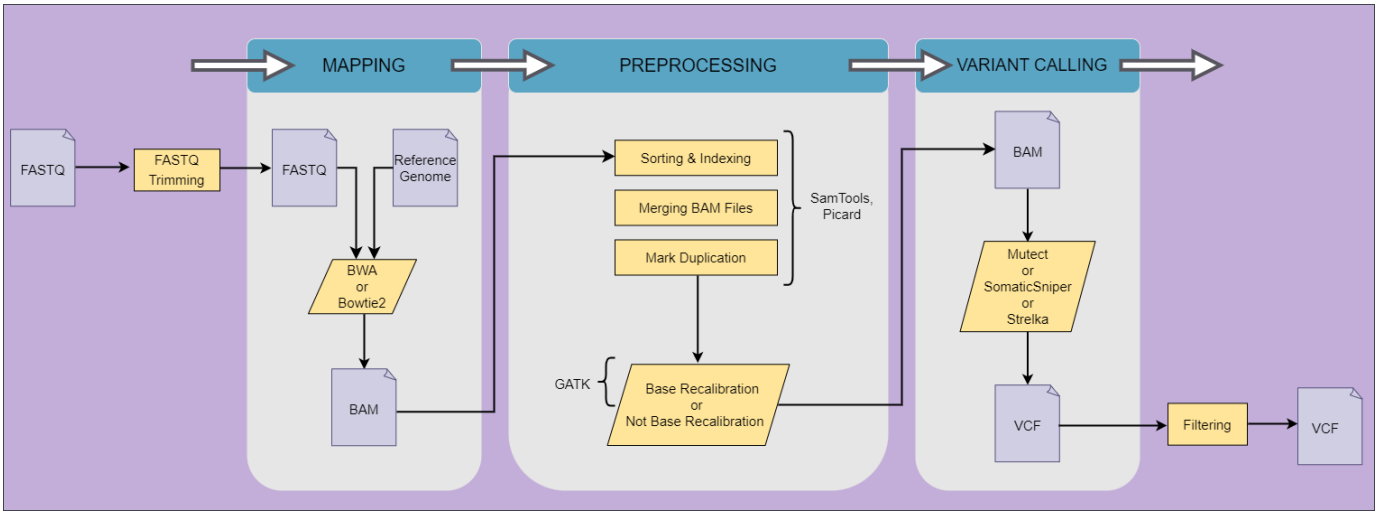


Fig. 1: NGS Bioinformatics Pipeline Structure

B. Used Dataset

- **FASTQ Files:** For storing NGS data, FASTQ files have become the standard format. For this NGS data processing and analysis project, 2 specific FASTQ files are downloaded as normal-tumor pair data: SRR7890851 as normal and SRR7890850 as tumor.

  SRR7890850 downloaded from here and SRR7890851 from here.

- **Reference Genome:** Human reference genome hg38 is downloaded from Google Cloud and is used to perform alignments.

- **Mapper Indices:** Index files to be used for mapping operations are downloaded from here(for Bowtie mapper) and here(for BWA mapper).

- **High Confidence Bed Files:** After obtaining VCF files, a further filtering process is needed. For this purpose, high confidence bed files are downloaded from here. Also the WES data is downloaded from here.

C. Configuring The Pipelines

In constructing a bioinformatics pipeline, pre-existing tools or programming utilities are oftenly integrated into the workflow. Consequently, the challenge can be reduced to just scripting the execution of these established tools in an appropriate sequence where the output of one step is the input of the following step. Therefore, the pivotal role of a bioinformatics pipeline lies in accurately implementing the necessary tools and coordinate their work.[4]

For the scope of this NGS data processing project, a workflow as in Fig. 1 is followed for the implementation of distinct pipelines. The major stages are Mapping, Preporcessing and Variant Calling. For all pipelines FASTQ Trimming is done before Mapping. Two different mapper options called BWA and Bowtie2 are in hand for Mapping Stage. In Preprocessing, Mark Duplication parameter is fixed for all pipelines and there are two options for Base Recalibration parameter(using base recalibration or not). On the other hand, for the Variant Calling stage, there are three different callers, namely Mutect, SomaticSniper and Strelka to be used. After Variant Calling, a further filtering process is done.

In total, there are 12 different pipeline configurations (2 recaliration options x 2 mappers x 3 variant callers) to be implemented and executed on the sample mentioned at Sec. II-B.

Considering all these, there are some fixed parameters like FASTQ Trimming and Mark Duplication and some variable parameters like Mapper library, presence of Base Calibration and Caller library. Since CoSAP allows the users to configure these parameters by adding, deleting or rearranging some steps, one can create pipelines by writing a basic Python script utilizing the cosap library.

III. Results

A. Principle Component Analysis (PCA)

Principal Component Analysis (PCA) is a technique for reducing dimensionality, commonly employed to decrease the number of variables in extensive datasets. This involves transforming a substantial set of variables into a more compact one that retains the majority of the information present in the original set. Reducing the number of variables in a dataset inevitably results in a sacrifice of accuracy. However, the strategy in dimensionality reduction is to trade a bit of accuracy for simplicity. Smaller datasets are more manageable for exploration and visualization, facilitating quicker analysis of data without the burden of redundant variables.[5]

Principal components are new variables derived as linear combinations or mixtures of the original ones. These combinations are created in a manner that ensures the new variables, known as principal components, are uncorrelated, and the majority of information from the initial variables is best represented in the first components. In essence, for an n-dimensional dataset, n principal components can be obtained and PCA aims to compress the maximum information into the first component, followed by the maximum remaining information in the second component, and so forth.[5] For the case, n is 13 and the scree plot that shows explained variability by each principal component is shown in Fig. 2.

In order to visualize the data, a scatter plot of the first and second principal components which carry the most of the variance in the original data is plotted as in Fig. 3. Each point in this plot represents a variant list(12 for pipelines and one for ground truth). The closer the points are, the more similar the variant lists are, the plot indicates.

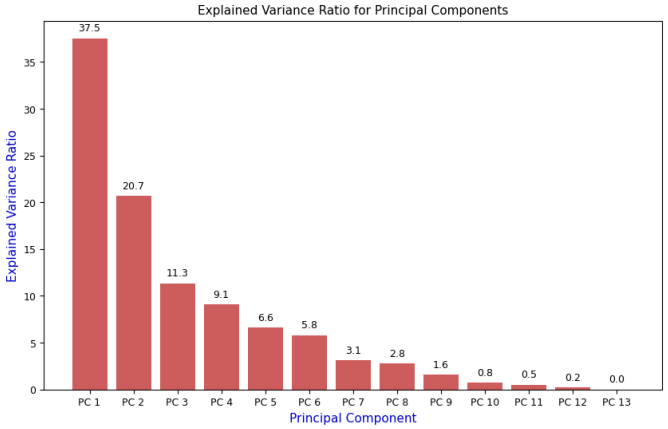


Fig. 2: Explained Variance Ratio Chart for Principal Components

It can clearly be seen from Fig. 3 that there is a clustering between the pipelines with same variant caller, indicating pipelines with same variant caller shows similarity. This demonstrates that the effect of the variant caller choice on the result is considerably higher than the choices for mapper and base recalibration.

The data frame constructed on which PCA is applied consists of information about not only pipelines to be tested but also ground truth variant list(hc_bed_filtered.vcf). Therefore, there are 13 data points on the scatter plot, instead of 12. This allows one to see the similarity of each pipeline's variant list with ground truth list. When the chart is examined in this sense, it is easily noticed that Mutect gives the most proper results independent of the other parameters like mapper and base calibration.

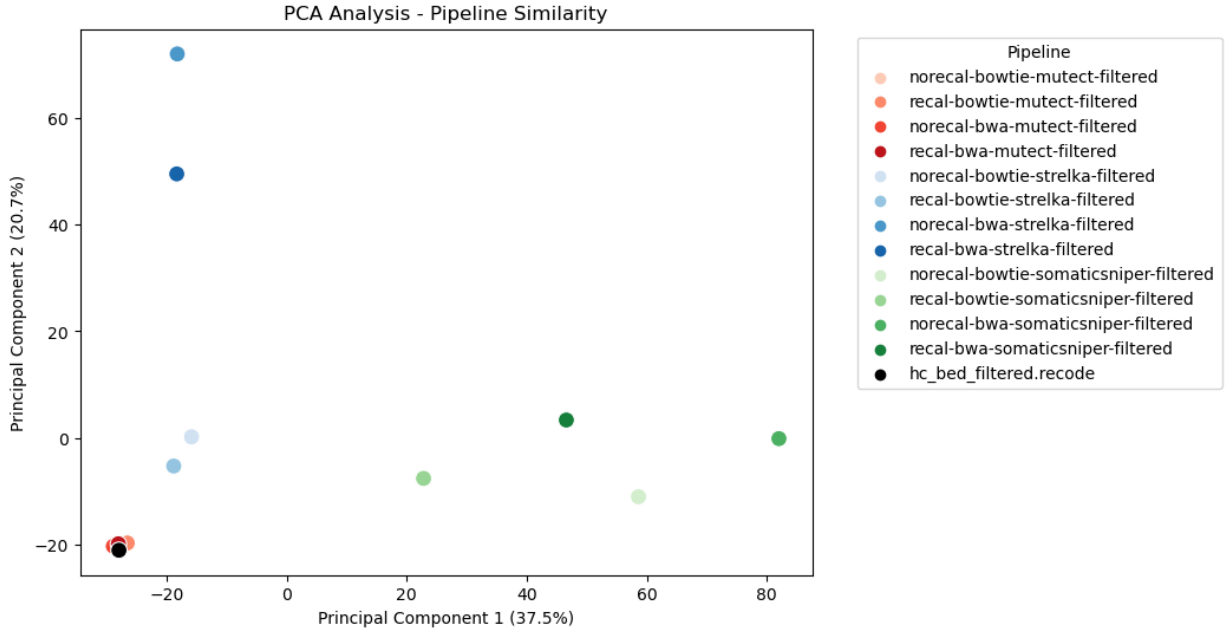
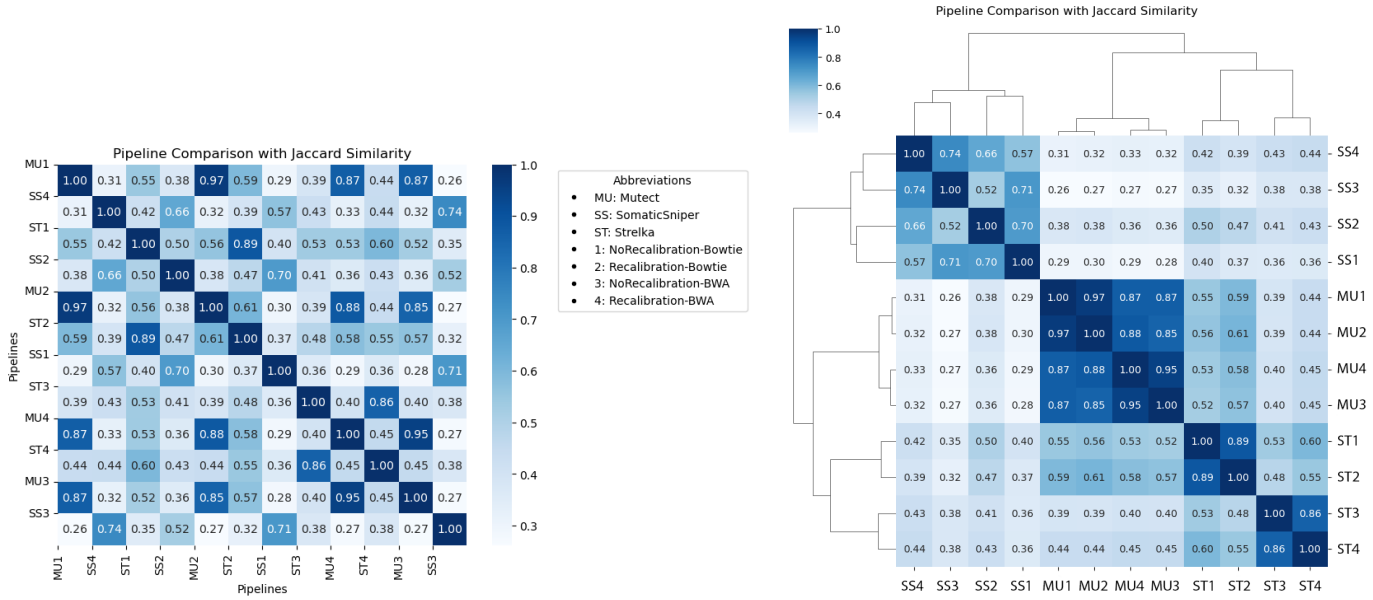


Fig. 3: PC1-PC2 Scatter Plot

B. Heatmap and Clustering



(a) Shuffled Similarity Matrix Heatmap

(b) Clustered Similarity Matrix Heatmap

Fig. 4: Heatmap on Jaccard Similarity of Pipelines

Jaccard similarity and Jaccard distance are widely used as a statistic for similarity and dissimilarity measurement.

In order to compare 12 pipelines subject to this research, Jaccard similarity of each pipeline pair is calculated, and put into a matrix. A heatmap, which is a visualization technique for representing data in a tabular format where values are represented by colors, is created out of this matrix, which can be seen from Fig. 4(a). However, this heatmap does not highlight underlying patterns or similarities that could be more apparent with clustering.

Clustering is the process of grouping similar items or data points together based on certain characteristics and can have a significant impact on the interpretation and visualization of a heatmap. Clustering in heatmaps helps to highlight patterns and relationships within the data by rearranging rows and/or columns based on similarity. That's why a new heatmap with dendograms that shows the clustering of pipelines is created, can be seen from Fig. 4(b). The difference is obvious.

Comparing the pipelines in terms of the similarity matrix heatmap provided in Fig.5, one can observe the similarity between the different pipeline configurations based on the variants detected. The heatmap shows the Jaccard similarity score between the pipelines, with darker colors indicating higher similarity.

The symmetry observed in the heatmap is a result of the construction of the Jaccard similarity matrix before generating the heatmap. The Jaccard similarity matrix is symmetric by definition because the Jaccard similarity between pipeline A and pipeline B is the same as the similarity between pipeline B and pipeline.

Also, the diagonal represents the self-similarity of each pipeline, meaning each pipeline is compared with itself. As a result, the diagonal values are all equal to 1, indicating a perfect match when a pipeline is compared to itself. This is a natural outcome of the Jaccard similarity metric, where the intersection of a set with itself is the set itself, leading to a similarity score of 1.

Based on the heatmap, the following observations can be made:

1. The pipelines "Norecal-Bowtie-Mutect" and "Recal-Bowtie-Mutect" are the most similar, with a similarity score of 0.97. This indicates that these pipelines detected a similar set of variants.

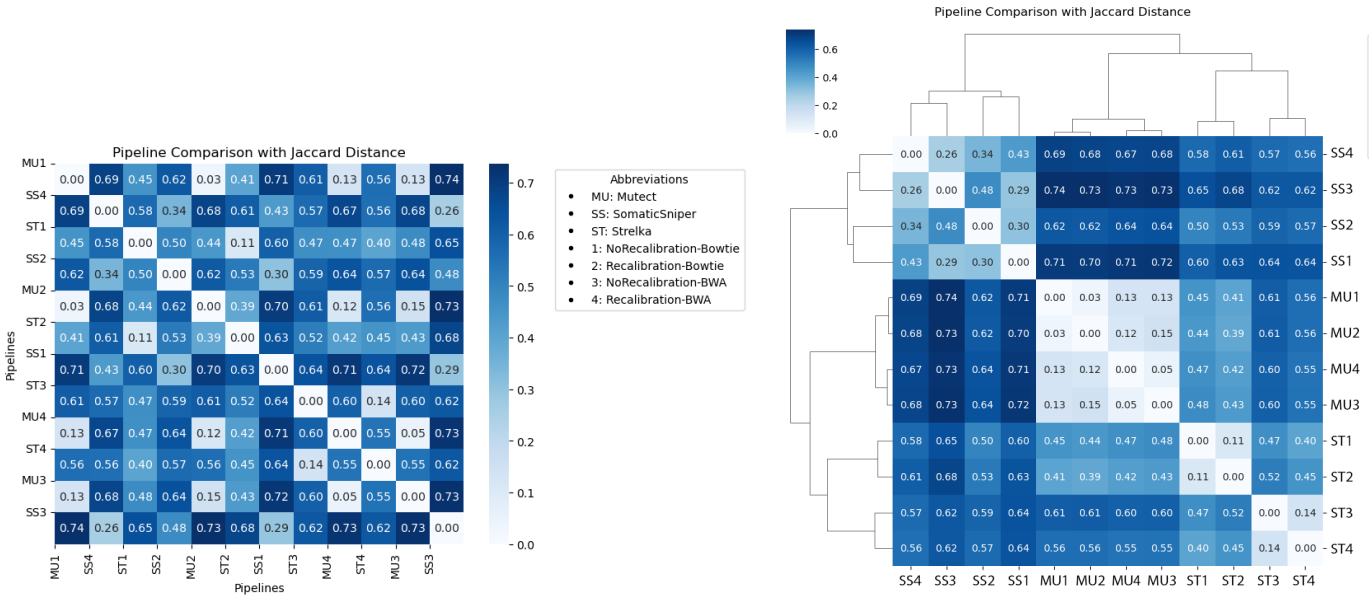
2. The pipelines "Norecal-Bwa-Mutect" and "Recal-Bwa-Mutect" are also highly similar, with a similarity score of 0.95. This suggests that these pipelines detected a similar set of variants.

3. The pipelines "Norecal-Bowtie-Strelka" and "Recal-Bowtie-Strelka" are similar, with a similarity score of 0.89. This indicates that these pipelines detected a similar set of variants.

4. The pipelines "Norecal-Bwa-SomaticSniper" and "Recal-Bwa-SmoticSniper" are also similar, with a similarity score of 0.74. This suggests that these pipelines detected a similar set of variants.

5. The pipelines "Norecal-Bwa-SomaticSniper" and "Norecal-Bowtie-Mutect" are the least similar, with a similarity score of 0.26. This indicates that these pipelines detected a different set of variants.

Overall, the similarity matrix heatmap highlights the importance of selecting the appropriate pipeline configuration based on the specific research question and dataset. It also suggests that the choice of mapper and variant caller has a greater impact on the similarity of the detected variants than the choice of base recalibration.



(a) Shuffled Distance Matrix Heatmap

(b) Clustered Distance Matrix Heatmap

Fig. 5: Heatmap on Jaccard Distance of Pipelines

This time jaccard distance of each pipeline pair is calculated, put into a matrix and a heatmap is created. After that a heatmap with clustering is created. Again, the effect of clustering can be seen from Fig. 5(a) and Fig. 5(b).

Comparing the pipelines in terms of the distance matrix heatmap provided in Fig. 5(b), one can observe the distance between the different pipeline configurations based on the variants detected. The darker the color, the higher the distance between the pipelines.

Based on the heatmap, the following observations can be made:

1. The pipelines ”Norecal-Bowtie-Mutect” and ”Recal-Bwa-Strelka” are the closest, with a distance score of 0.03, indicating that these pipelines detected a similar set of variants.

2. The pipelines ”Norecal-Bwa-Mutect” and ”Recal-Bwa-Mutect” are also very close, with a distance score of 0.05, indicating that these pipelines detected a similar set of variants.

3. The pipelines ”Norecal-Bowtie-Strelka” and ”Recal-Bowtie-Strelka” are also very close, with a distance score of 0.11, indicating that these pipelines detected a similar set of variants.

4. The pipelines ”Norecal-Bwa-SomaticSniper” and ”Recal-Bwa-SmoticSniper” are close, with a distance score of 0.26, indicating that these pipelines detected a similar set of variants.

5. The pipelines ”Norecal-Bwa-SomaticSniper” and ”Norecal-Bowtie-Mutect” are the furthest apart, with a distance score of 0.74, indicating that these pipelines detected a different set of variants.

Overall, the distance matrix heatmap suggests that the choice of mapper and variant caller has a greater impact on the similarity of the detected variants than the choice of base recalibration. The heatmap also highlights the importance of selecting the appropriate pipeline configuration based on the specific research question and dataset.

C. Statistical Metrics

Performances of pipelines are analyzed and compared based on metrics such as Precision, Recall, F1 Score, and Accuracy. Before defining them some concepts should be understood in the context of variant calling pipelines.

True Positive (TP): It is a variant that was detected by the pipeline configured and is one that exists in the high-confidence list of variants.

False Positive (FP): It is a variant that was detected by the pipeline configured but is one that does not exist in the high-confidence list of variants.

True Negative (TN): It is a variant that was not detected by the pipeline configured and is one that does not exist in the high-confidence list of variants.

False Negative (FN): It is a variant that was not detected by the pipeline configured but is one that does exist in the high-confidence list of variants.

$$precision = \frac{true\ positives}{true\ positives + false\ positives}$$

$$recall = \frac{true\ positives}{true\ positives + false\ negatives}$$

$$F1 = 2 \times \frac{precision \times recall}{precision + recall}$$

$$accuracy = \frac{true\ positives + true\ negatives}{true\ positives + true\ negatives + false\ positives + false\ negatives}$$

Fig. 6: Calculations of Metrics

Precision: Indicates the proportion of true positive predictions out of all positive predictions. It measures the accuracy of the positive predictions.

Recall: Represents the proportion of true positive predictions out of all actual positives. It measures the ability of the model to identify all relevant instances.

F1 Score: The harmonic mean of precision and recall, provides a balance between the two metrics.

Accuracy: Measures the overall correctness of the predictions.

Precision is also known as positive predictive value and it is the fraction of relevant instances among the retrieved instances. Recall is also known as sensitivity and it is the fraction of relevant instances that were retrieved. Precision can be seen as a measure of quality, and recall as a measure of quantity. Higher precision means that an algorithm returns more relevant results than irrelevant ones, and high recall means that an algorithm returns most of the relevant results (whether or not irrelevant ones are also returned).[6]

Based on the metrics values, it's evident that the choice of pipeline configuration, including the use of different mappers and variant callers, has a significant impact on the performance metrics. For instance, the precision, recall, F1 score, and accuracy vary across different pipeline configurations, highlighting the importance of selecting the appropriate combination of tools for variant calling.

The metrics values also demonstrate that the recalibration of the pipeline configurations generally leads to improvements in performance metrics, as seen in the higher precision, recall, F1 score, and accuracy values for the recalibrated pipelines compared to the non-recalibrated ones.

Furthermore, the choice of variant caller and mapper also influences the performance metrics, as evidenced by the differences in precision, recall, F1 score, and accuracy across the different combinations of variant callers and mappers.

Overall, this data underscores the significance of carefully selecting and configuring the components of variant calling pipelines to achieve optimal performance in terms of precision, recall, F1 score, and accuracy.

Comparing the pipelines in terms of metrics such as precision, recall, accuracy, and F1 score, the following observations can be made:

1. Precision:

- The pipeline "Recal-Bowtie-Mutect" achieved the highest precision of 0.92, followed closely by "Norecal-Bowtie-Mutect" with a precision of 0.89. These pipelines demonstrated the highest precision among all the configurations tested.
- The pipeline "Norecal-Bwa-SomaticSniper" had the lowest precision of 0.30, indicating a higher rate of false positives in variant calling.

2. Recall:

- The pipelines "Norecal-Bwa-Strelka" and "Recal-Bwa-Strelka" achieved the highest recall of 0.81, followed closely by "Norecal-Bwa-SomaticSniper" with a precision of 0.75, indicating a lower rate of false negatives in variant calling.
- The pipeline "Recal-Bowtie-SomaticSniper" had the lowest recall of 0.67, suggesting a higher rate of false negatives in variant calling.

3. F1 Score:

- The pipelines "Recal-Bowtie-Mutect" and "Recal-Bwa-Mutect" achieved the highest F1 score of 0.82, indicating a good balance between precision and recall.
- The pipeline "Norecal-Bwa-SomaticSniper" had the lowest F1 score of 0.43, indicating a lower balance between precision and recall.

4. Accuracy:

- The pipelines "Recal-Bowtie-Mutect" and "Recal-Bwa-Mutect" achieved the highest accuracy of 0.69, followed closely by "Norecal-Bowtie-Mutect" and "Norecal-Bwa-Mutect" with accuracies of 0.68 and 0.67, respectively.
- The pipeline "Norecal-Bwa-SomaticSniper" had the lowest accuracy of 0.27, indicating lower overall correctness in variant calling.

These comparisons highlight the varying performance of the pipelines in terms of precision, recall, F1 score, and accuracy, emphasizing the importance of selecting the appropriate pipeline configuration for NGS data analysis.

Norecal-Bowtie-SomaticSniper:
- Variants only in hc-bed-filtered: 357
- Variants only in norecal-bowtie-somaticsniper: 1602
- Variants present in both: 804
- Precision: 0.33
- Recall: 0.69
- F1 Score: 0.45
- Accuracy: 0.29

Recal-Bowtie-SomaticSniper
- Variants only in hc-bed-filtered: 380
- Variants only in recal-bowtie-somaticsniper: 971
- Variants present in both: 781
- Precision: 0.45
- Recall: 0.67
- F1 Score: 0.54
- Accuracy: 0.37

Norecal-Bwa-SomaticSniper
- Variants only in hc-bed-filtered: 291
- Variants only in norecal-bwa-somaticsniper: 2019
- Variants present in both: 870
- Precision: 0.30
- Recall: 0.75
- F1 Score: 0.43
- Accuracy: 0.27

Recal-Bwa-SomaticSniper
- Variants only in hc-bed-filtered: 307
- Variants only in recal-bwa-strelka: 1458
- Variants present in both: 854
- Precision: 0.37
- Recall: 0.74
- F1 Score: 0.49
- Accuracy: 0.33

Norecal-Bowtie-Strelka
- Variants only in hc-bed-filtered: 330
- Variants only in norecal-bowtie-strelka: 482
- Variants present in both: 831
- Precision: 0.63
- Recall: 0.72
- F1 Score: 0.67
- Accuracy: 0.51

Recal-Bowtie-Strelka
- Variants only in hc-bed-filtered: 335
- Variants only in recal-bowtie-strelka: 347
- Variants present in both: 826
- Precision: 0.70
- Recall: 0.71
- F1 Score: 0.71
- Accuracy: 0.55

Norecal-Bwa-Strelka
- Variants only in hc-bed-filtered: 218
- Variants only in norecal-bwa-strelka: 1258
- Variants present in both: 943
- Precision: 0.43
- Recall: 0.81
- F1 Score: 0.56
- Accuracy: 0.39

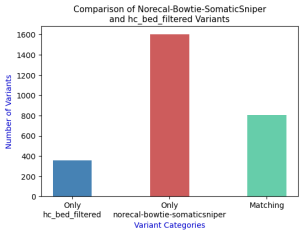
Recal-Bwa-Strelka
- Variants only in hc-bed-filtered: 225
- Variants only in recal-bwa-strelka: 988
- Variants present in both: 936
- Precision: 0.49
- Recall: 0.81
- F1 Score: 0.61
- Accuracy: 0.44

Norecal-Bowtie-Mutect
- Variants only in hc-bed-filtered: 297
- Variants only in norecal-bowtie-mutect: 101
- Variants present in both: 864
- Precision: 0.89
- Recall: 0.74
- F1 Score: 0.81
- Accuracy: 0.68

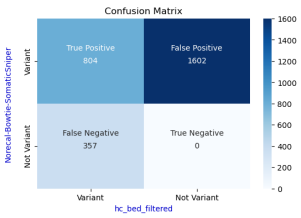
Recal-Bowtie-Mutect
- Variants only in hc-bed-filtered: 301
- Variants only in recal-bowtie-mutect: 78
- Variants present in both: 860
- Precision: 0.92
- Recall: 0.74
- F1 Score: 0.82
- Accuracy: 0.69

Norecal-Bwa-Mutect
- Variants only in hc-bed-filtered: 259
- Variants only in norecal-bwa-mutect: 167
- Variants present in both: 902
- Precision: 0.84
- Recall: 0.78
- F1 Score: 0.81
- Accuracy: 0.67

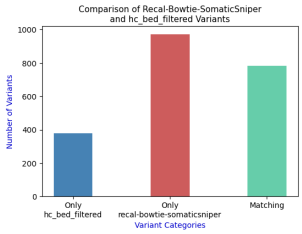
Recal-Bwa-Mutect
- Variants only in hc-bed-filtered: 260
- Variants only in recal-bwa-mutect: 145
- Variants present in both: 901
- Precision: 0.86
- Recall: 0.78
- F1 Score: 0.82
- Accuracy: 0.69



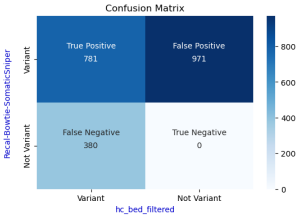
(a) Norecal Bowtie



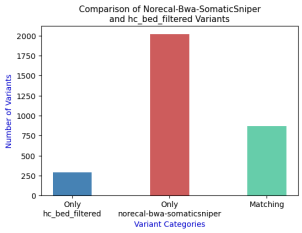
(b) Norecal Bowtie



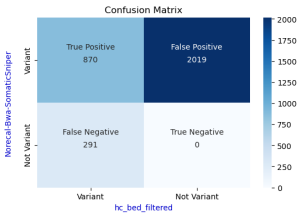
(c) Recal Bowtie



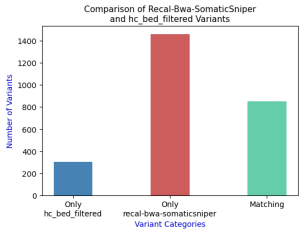
(d) Recal Bowtie



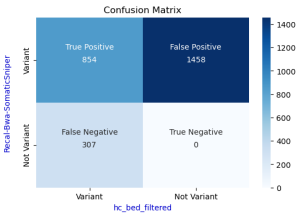
(e) Norecal Bwa



(f) Norecal Bwa

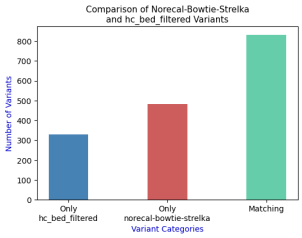


(g) Recal Bwa

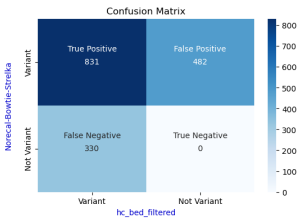


(h) Recal Bwa

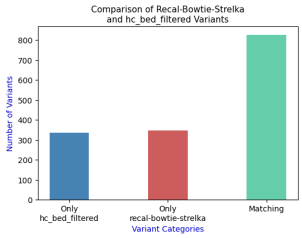
Fig. 7: SomaticSniper



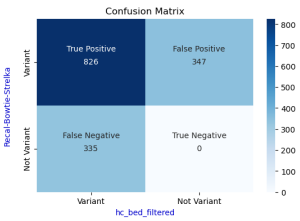
(a) Norecal Bowtie



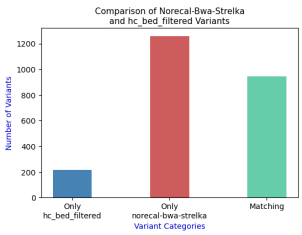
(b) Norecal Bowtie



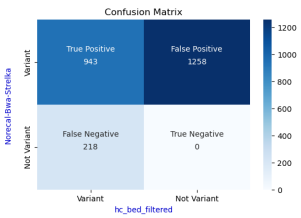
(c) Recal Bowtie



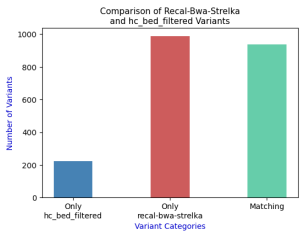
(d) Recal Bowtie



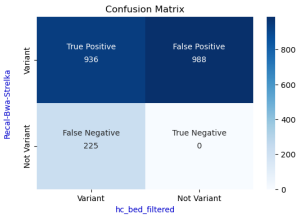
(e) Norecal Bwa



(f) Norecal Bwa

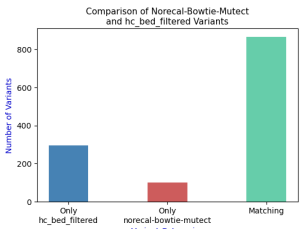


(g) Recal Bwa

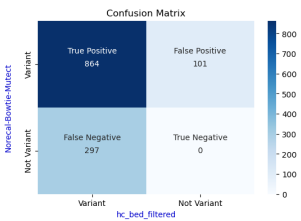


(h) Recal Bwa

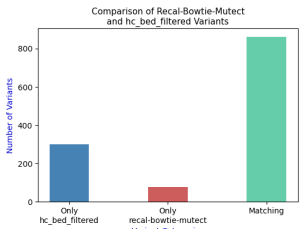
Fig. 8: Strelka



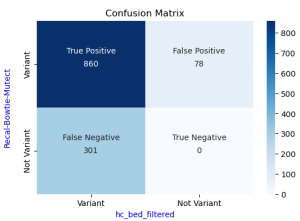
(a) Norecal Bowtie



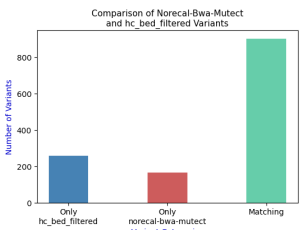
(b) Norecal Bowtie



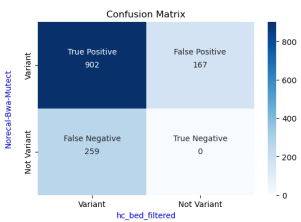
(c) Recal Bowtie



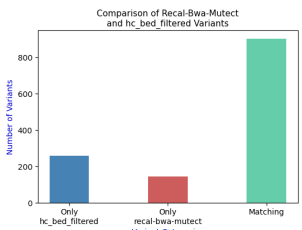
(d) Recal Bowtie



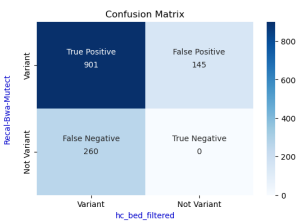
(e) Norecal Bwa



(f) Norecal Bwa



(g) Recal Bwa



(h) Recal Bwa

Fig. 9: Mutect

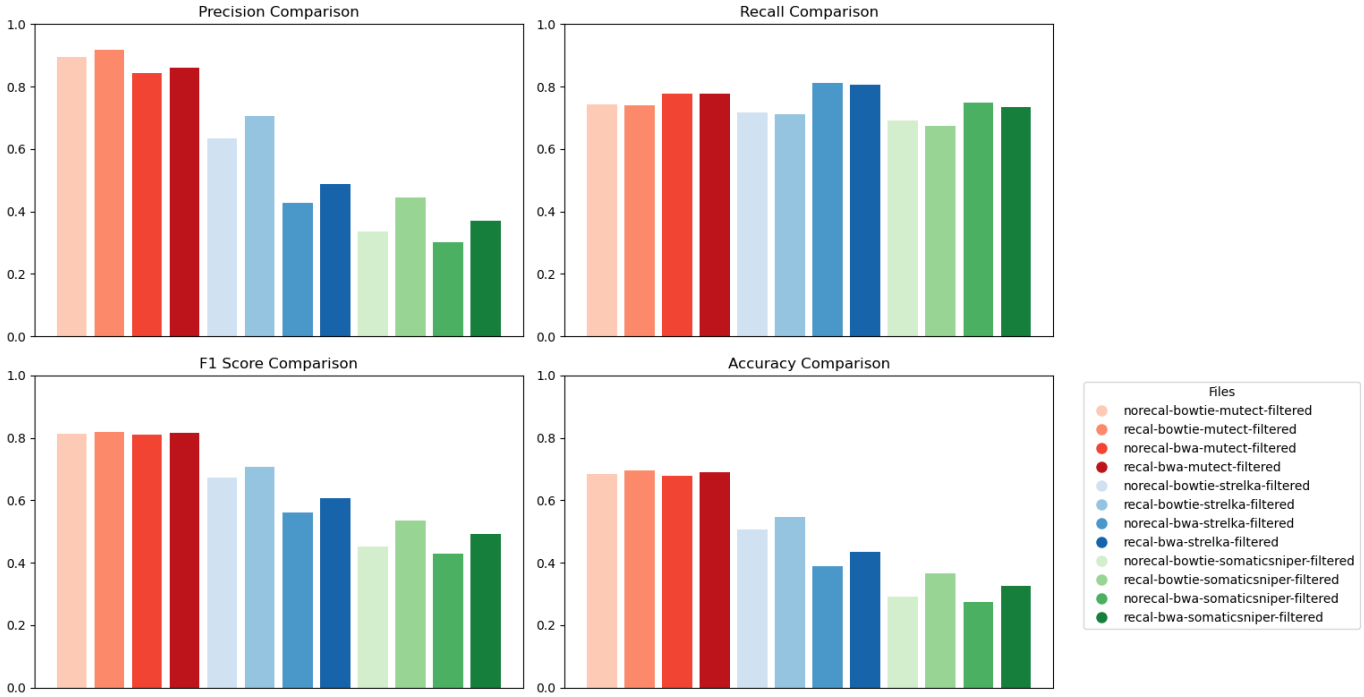


Fig. 10: Metrics Comparison

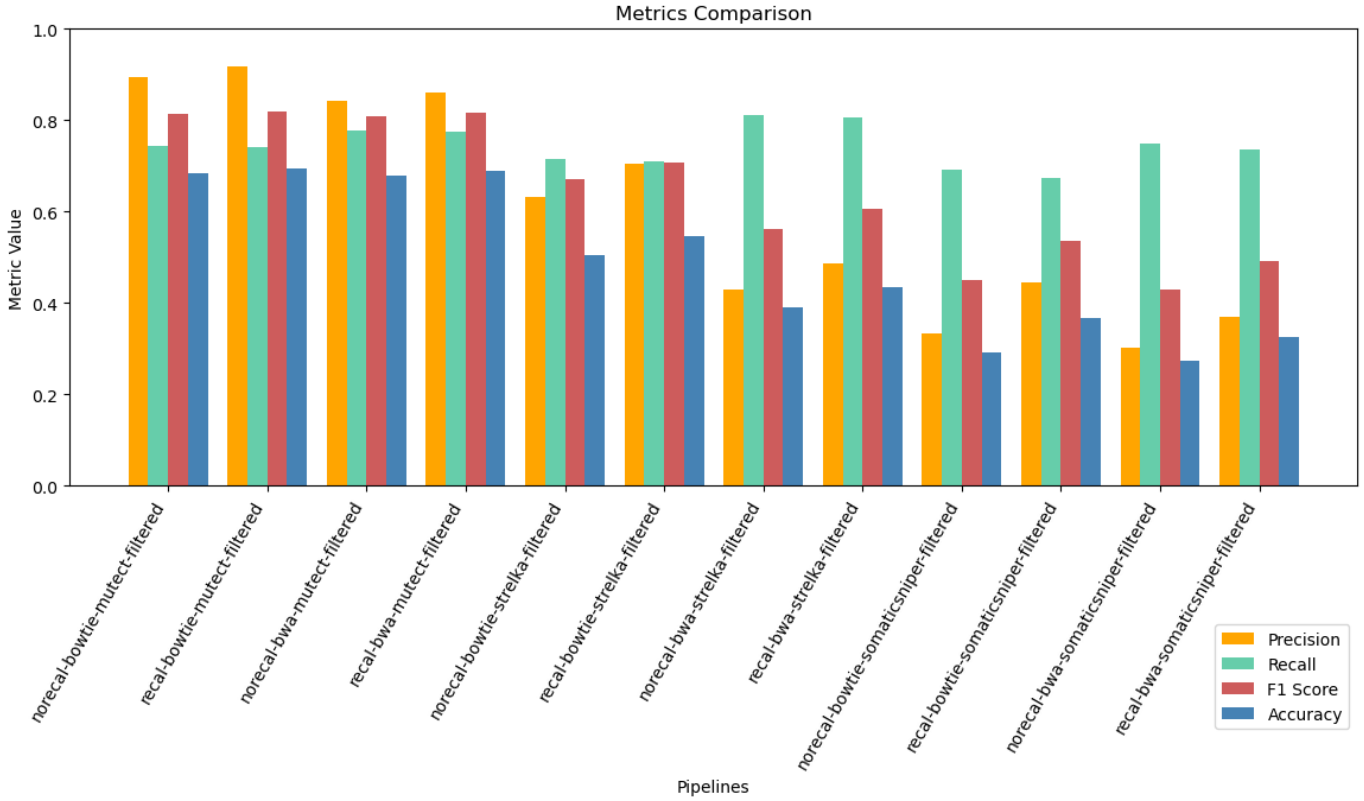


Fig. 11: Metrics Comparison

D. Execution Times

Bowtie mapper for normal: 535 minutes
Bowtie mapper for tumor: 439 minutes

BWA mapper for normal: 410 minutes
BWA mapper for tumor: 382 minutes

- BWA mapper tooks less time than Bowtie mapper.
- Both BWA and Bowtie aligns normal sample in a longer time than tumor sample.

Mark Duplication after BWA mapper: 430 minutes
Mark Duplication after Bowtie mapper: 439 minutes

- Mark Duplication is performed in approximately the same time both after BWA and Bowtie.

(After NoRecalibration and Bowtie mapper)
SomaticSniper variant caller: 20 minutes
Strelka variant caller: 17.5 minutes
Mutect variant caller: 30 minutes

(After Recalibration and Bowtie mapper)
SomaticSniper variant caller: 16 minutes
Strelka variant caller: 13.5 minutes
Mutect variant caller: 42 minutes

(After NoRecalibration and BWA mapper)
SomaticSniper variant caller: 14 minutes
Strelka variant caller: 13.5 minutes
Mutect variant caller: 49 minutes

(After Recalibration and BWA mapper)
SomaticSniper variant caller: 17 minutes
Strelka variant caller: 15 minutes
Mutect variant caller: 31 minutes

- Independent of the Base Calibration and Mapper choices, the execution time of variant callers have an descending order for Mutect, SomaticSniper and Strelka respectively.

IV. Discussion

In this research, the CoSap environment is employed to create Next Generation Sequencing (NGS) pipelines. Two distinct aligners/mappers, three variant callers and two Base Calibration options are utilized to assess the optimal performance among the 12 total pipelines on a human exome, which was instrumental in generating the list of variants detected through NGS. The evaluation of pipeline performance includes the analysis of metrics such as Precision, Recall, F1-Score, and Accuracy. Visualization tools such as heatmaps, PCA plots, and Precision-Recall charts are employed for a comprehensive analysis.

REFERENCES

[1] N. H. G. R. Institute. "Dna sequencing - genetics home reference." Accessed: April 28, 2024. (2024), [Online]. Available: https://www.genome.gov/genetics-glossary/DNA-Sequencing.

[2] S. Roy, C. Coldren, A. Karunamurthy, *et al.*, "Standards and guidelines for validating next-generation sequencing bioinformatics pipelines: A joint recommendation of the association for molecular pathology and the college of american pathologists," *The Journal of Molecular Diagnostics*, vol. 20, no. 1, pp. 4–27, 2018, ISSN: 1525-1578. DOI: https://doi.org/10.1016/j.jmoldx.2017.11.003. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1525157817303732.

[3] MBaysanLab. "Cosap: Comparative sequencing analysis platform." Last updated: November 28, 2023. Accessed: April 28, 2024. (2023), [Online]. Available: https://github.com/MBaysanLab/cosap.

[4] Hyperskill. "Title of the specific web page/step (if available)." Accessed: January 3, 2024. (n.d.), [Online]. Available: https://hyperskill.org/learn/step/22933.

[5] Built In. "Step-by-step explanation of principal component analysis." Accessed: January 3, 2024. (n.d.), [Online]. Available: https://builtin.com/data-science/step-step-explanation-principal-component-analysis.

[6] Wikipedia. "Precision and recall." Accessed: January 3, 2024. (n.d.), [Online]. Available: https://en.wikipedia.org/wiki/Precision_and_recall.