Transgene Detection with Next-Generation Sequencing – and a Supervised Learning Approach?

Chris Siu Yeung Chow (SID 46696120)

ABSTRACT

The detection and classification of genetically modified organisms (GMOs) play a crucial role in ensuring the safety and compliance of agricultural products. This report explores the application of Next-Generation Sequencing (NGS) technology and possible improvements with supervised learning for transgene detection in sorghum for scientific advancements.

I. INTRODUCTION

Challenges and uncertainties in transgene experiments

Transgenes, or artificially introduced genes with vectors called plasmids, have become a subject of increasing interest and concern in the field of genetic engineering and biotechnology. Genetically-modified (GM) organisms are created to express better traits and functionalities for commercial and academic purposes. However, transgene technologies are not perfect.

For example, particle bombardment, a transformation technique that is achieved by shooting plasmids with the gene of interest into cells along with tiny particles, is limited in its ability to accurately deliver the gene to the target site in the genome. It could also damage the genetic material to due kinetic damage and cause partial integration of the transgene.

In addition, unsuccessful attempts had been made to produce non-GM cattle through gene technologies. In GM hornless cattle, the backbone of a plasmid was unintendedly integrated into the genome with bacterial resistance genes [9]. Also, orange petunias being sold on the market were found to contain Maize genes [2]. It was suspected that GM plants were used in the breeding programmes for more colour varieties.

In Australia, transgenic organisms are classified as GM with as few as 7 base pairs of foreign DNA unless the inserted material belongs to the host species or if the insertion occurs naturally [12]. Hence, it is possible to create non-GM edits in the genome with technologies such as CRISPR while adhering to the regulations. For example, if the Cas9 transgene and selectable markers of the plasmid are no longer present in the genome, then gene edited plants are considered non-GM and can be grown in the field. However, gene-editing technologies are not fully mature. Recent studies have shown their unintended effects in the creation of new proteins and large deletions with complex rearrangements of DNA [5, 11]. Also, they have the ability to target regions that are highly resistant to mutation in nature. The potential consequences could be disastrous to nature and society if used in the wrong hands. Therefore, the implications of a robust way for regulators to detect transgenes is not just of ethical concern, but also of economic, environmental and health.

Introduction to Next-Generation Sequencing (NGS) as a potential tool for transgene detection

Next-Generation Sequencing (NGS) is a promising way to detect GM organisms for that purpose. A study has pointed out that the associated costs and sequencing runs only increase when samples consist of less than 1% DNA of GM organisms [13]. This implies that NGS may be suitable for regulatory purposes as it is cost-efficient to detect transgenes in bulk (Figure 12).

Also, it is theoretically possible to detect partial plasmid integration. The transgenes are inserted with the error-prone particle bombardment with gene gun. Hence this project investigates that aspect.

Research objective and overview of the report

In this project, I have examined the use of NGS data for the detection of transgenes introduced using particle bombardment in sorghum. This report describes the reproducible bioinformatic pipeline for mapping pair-end NGS reads, and the procedure to optimise it for the classification of 26 sorghum samples on whether they are GM.

Tx430_glasshouse_21_embryo_season___A3	CPDI2-2___A6
Tx430_TC_control___A1	CPDI2-2___B6
Tx430_2020___D3	CPDI2-2___C6
y-kaf9-1___B10	CPDI2-2___F5
y-kaf9-2___A10	CPDI2-2___G5
y-kaf9-2___D9	CPDI2-2___H5
y-kaf9-2___E9	CPDI2-5___A5
y-kaf9-2___F9	CPDI2-5___B5
y-kaf9-2___G9	CPDI2-5___H4
y-kaf9-2___H9	CPDI3-1-1___D6
y-kaf9-2___C10	CPDI3-1-1___E6
y-kaf9-3___D10	CPDI3-1-1___F6
	CPDI3-2-3___C5
	CPDI3-2-3___D5
	CPDI3-2-3___E5

Figure 1. A List of the Names for 26 Sorghum Samples. The figure displays a list of 26 samples, with the PCR-tested transgenic samples highlighted in blue. The initial analyses were conducted on the first 12 samples (Left). For further details regarding the samples, please refer to the supplementary data.

II. METHODOLOGY AND EXPERIMENTAL WORKFLOW

Overview of library preparation

My project starts from the sequenced NGS data. For an individual or combined library of DNA samples, detailed protocols for sampling and library preparation are executed. Once extracted, the DNA goes through enzyme treatments like Tn5 transposase for fragmentation and the addition of ligation adaptors.

The prepared library samples are loaded onto a lane in a Next-Generation Sequencing machine like Illumina's Novaseq platform to generate paired-end reads, stored as fastq files.

Bioinformatic Pipeline

Afterwards, the bioinformatic scripts developed by Dr. Peter Crisp were executed on Bunya, the High-Performance Computing (HPC) Cluster, at the University of Queensland. They were modified for more functionality in the later stages of the project. The original versions of these scripts can be accessed publicly at the GitHub repository: https://github.com/pedrocrisp/crisplab_epigenomics. The modified scripts are available at: https://github.com/phycochow/crisplab_wgs.

First, the raw reads were subjected to quality checks using FastQC and MultiQC [1, 3]. Then, Trim Galore was applied to remove low-quality sequences and eliminate adaptor sequences from the reads [6]. Finally, the remaining reads were aligned to a reference using bowtie2 with the pre-set parameters in the existing scripts [7]. Subsequently, the BAM and Bed files were generated using the BigWig module in deeptools [10]. To visualise the data visualization, the Integrative Genome Viewer (IGV) was employed (Figure 3, 4, 5).

The current reference genome for sorghum is BTx623, an inbred genotype characterized by its short stature and early maturation. For a comprehensive analysis, the BTx623 genome was merged with the two plasmids sequences to act as the mapping reference for Bowtie2. The plasmid sequences were treated as two extra chromosomes. For the specific purpose of detecting the transgene, a combined file with the plasmid sequences was also used as the reference. This would be discussed in the later sections.

Iterative software to apply pipeline

In the advanced stages of the project, modifications were made to the existing scripts, and new scripts were developed to streamline the entire bioinformatic pipeline with more functionality. It enabled the iterative execution of Crisp's scripts and information extraction from the log files for extensive results. Throughout this process, various issues were encountered, including handling file types, navigating directories, and managing disk space. These factors were carefully considered during the algorithm design phase to ensure optimal performance and robustness. Raw reads were subsampled at odd percentages ranging from 1% to 99% to assess the impact of varying sequencing depths on the analysis.

Additionally, I briefly tested an idea to apply supervised learning for the classification of GM samples. The decision tree model was developed using the scikit-learn (sklearn) library [4]. Model evaluation was performed to customize parameters using cross-validation and F1 score from the sklearn.metrics module. The best model was selected and trained on the training data, with predictions made on the test set. The decision tree graph was generated using the export_graphviz function from the sklearn.tree module.

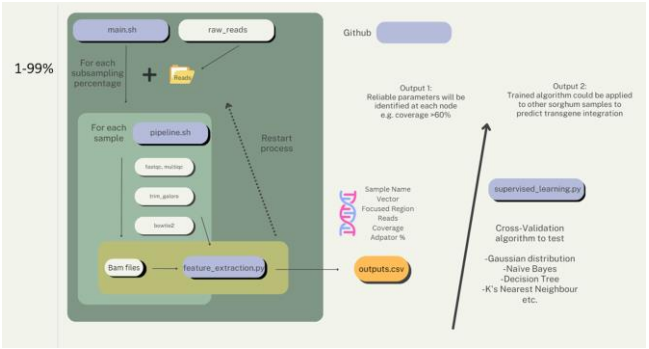

Figure 2. A Graphical Illustration of the Software Architecture to Extract Features from NGS Data. The automated script is depicted in the green box that iteratively submits bioinformatic pipeline jobs in a queue to the SLURM scheduler on the HPC. The extracted information is stored in a csv file shown in orange. It is fed into a supervised learning software as described in methods.

III. Results

Limitations of directly mapping reads to plasmid sequences

Challenges arise when trimmed reads are directly mapped to plasmid sequences. An example analysis demonstrates that mapping raw reads solely to the plasmids leads to numerous spikes in the number of reads mapped to specific positions, particularly for the NptII vector (Figure 3). These spikes can be attributed to elements such as plant promoters and gRNA, causing false positives in the mapping process.

Merging sorghum and plasmid sequence

To overcome this, an alternative approach is employed. The sorghum and plasmid sequences are merged to create a combined file. This integration of sequences results in the noise generated by the sorghum regions, which is subsequently rejected by the algorithm. By adopting this merged approach, the detection of transgenes in sorghum using Next-Generation Sequencing (NGS) becomes possible with negligible false positives, providing a more accurate and reliable means of transgene identification.

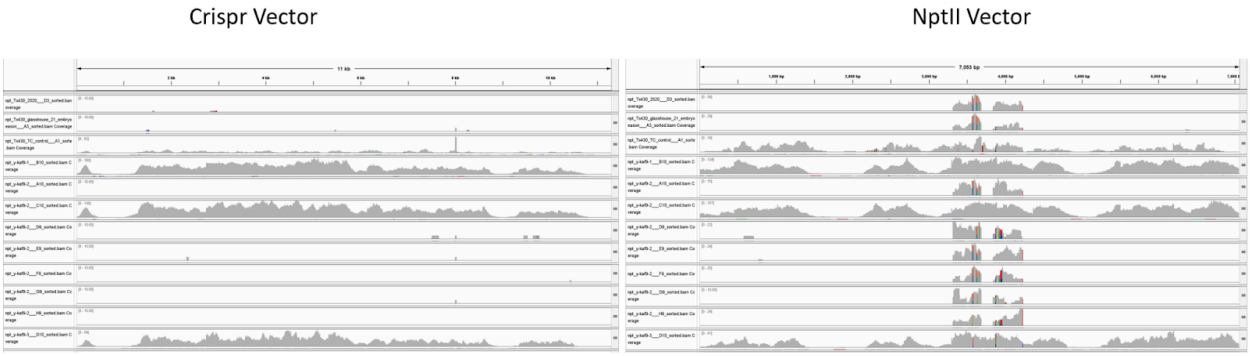

Figure 3. Results of Mapping the Raw Reads to the Plasmids. The figure shows an alignment of raw reads from the 12 samples to the plasmids containing the transgene. The y-axis represents the read depth, while the x-axis represents the position of the plasmid. The majority of the plasmids are shown. The mapped reads are depicted in blue. The Cas9 gene starts from approximately 2kb to 6kb in the CRISPR vector. The positive results are A1, B10, C10, D10, E9, D9.

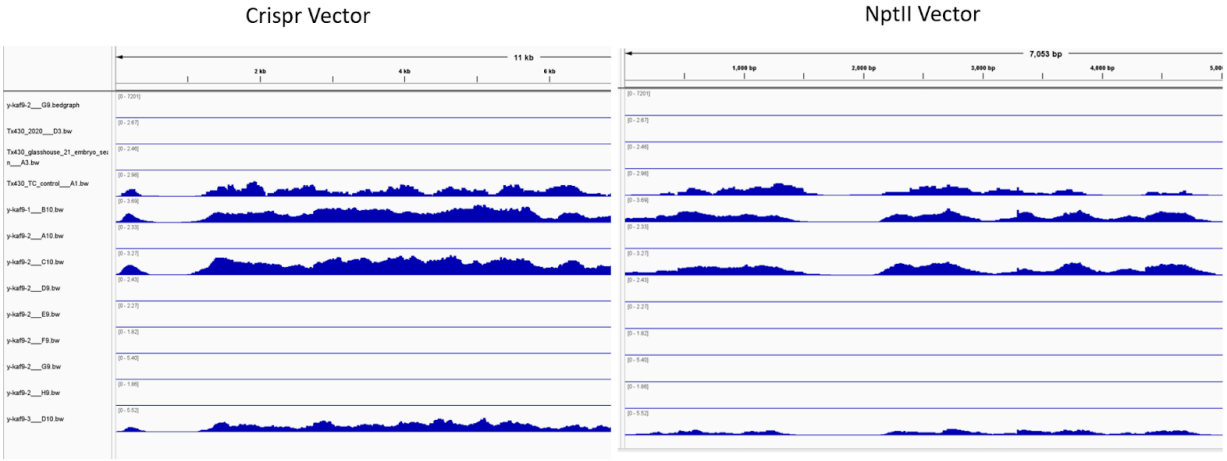

Figure 4. Results of Mapping of Raw Reads to both the Reference Genome and Plasmids. Alignment of raw reads from the 12 samples to the plasmids containing the transgene. The y-axis represents the read depth, while the x-axis represents the position of the plasmid. The majority of the plasmids are shown. The mapped reads are depicted in blue. The Cas9 gene starts from approximately 2kb to 6kb in the CRISPR vector. The positive results are A1, B10, C10, D10.

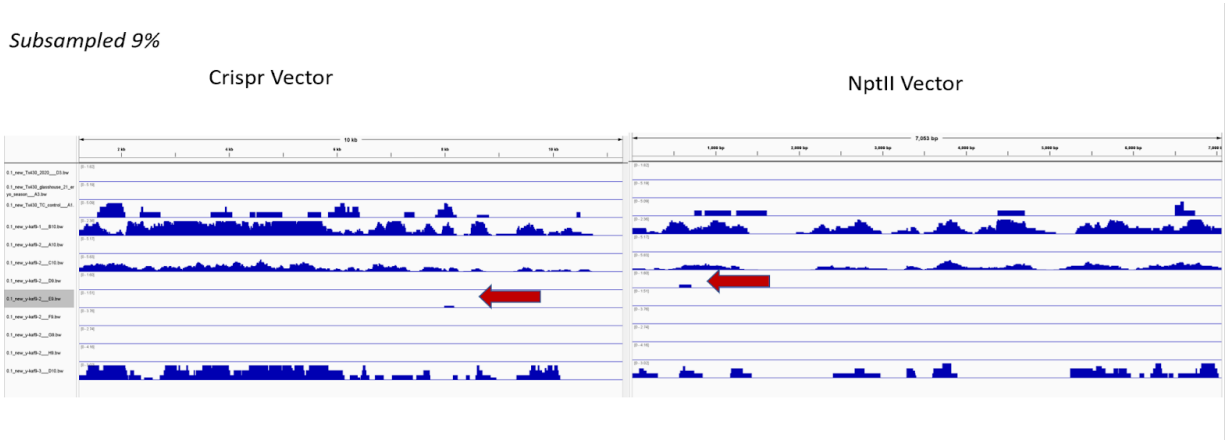

Figure 5. Results of Mapping 9% of Raw Reads to both the Reference Genome and Plasmids. Alignment of raw reads from the 12 samples to the plasmids containing the transgene. The y-axis represents the read depth, while the x-axis represents the position of the plasmid. The majority of the plasmids are shown. The mapped reads are depicted in blue. The Cas9 gene starts from approximately 2kb to 6kb in the CRISPR vector. The positive results are A1, B10, C10, D10, E9, D9. The red arrows point to a small proportion of mapped reads in E9 (Left) and D9 (Right).

Successful detection of transgenes in sorghum

The feasibility of using Next-Generation Sequencing (NGS) data to detect transgenes is demonstrated in this project. The analysis reveals successful detection of the two transgenic plants (A1 and B10) along with the corresponding plasmids in all positive samples. Results also show that with a sufficient number of reads (e.g., 65%), the plasmids can still be reliably detected. However, reducing the read percentage to 9% leads to decreased detection quality, as evidenced by the presence of chunky reads in both transgenic and non-transgenic samples. This phenomenon is attributed to the mapping software bowtie2, which struggles to classify certain reads as repetitive DNA when the sample size of reads is insufficient.

Impact of reads on detection efficiency

It was hypothesised that the number of reads has a significant impact on the efficiency of transgene detection. To determine the required number of reads, a model developed in previous studies was applied to sorghum [13]. The findings revealed that approximately 150,000 reads were necessary for reliable detection of transgenes. Moreover, to achieve comprehensive insights on insertion site and statistical significance, a substantially larger number of reads, up to 100 times more, were required (Figure 6). These highlight the importance of read depth in NGS to ensure accurate and conclusive results.

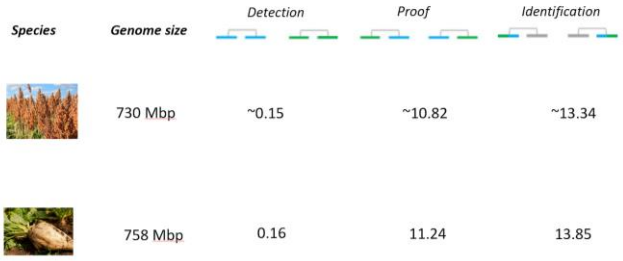

Figure 6. Number of Total Reads Needed to Achieve 95% Confidence for Transgene Detection Events in Sorghum (Upper) and Sugar Beet (Lower). Detection refers to the detection of transgenes; Proof refers to proof of transgene integration into the genome; Identification refers to the identification of the insertion site. The numbers are in millions.

In the analysis of the relationship between the number of raw reads and coverage, we observed a significant exponential growth in coverage as the number of reads increased (Figure 7, 8). This trend was consistent across both vectors mapped. The figures will be further discussed.

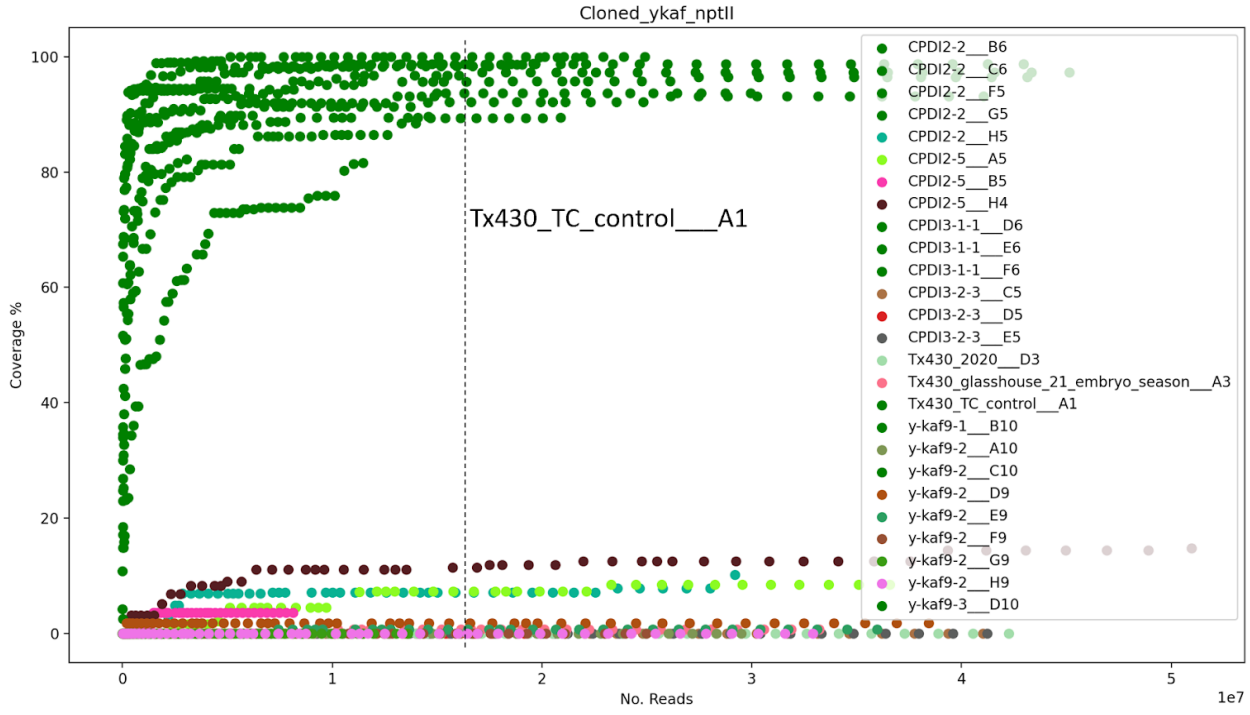

Figure 7. Relationship between Number of Raw Reads and Coverage % for different samples in the Cloned_ykaf_nptII vector. The samples represented in dark green include the transgenic samples and C10, D10 that tested positive in earlier tests. Other samples are plotted in random colours. Each dot represents the plasmid coverage at a certain amount of raw reads. The vertical dotted black line represents the convergence threshold determined visually for all samples. The sample name in the middle represents the scatter-dot line with lowest converged coverage. Plotted with matplotlib module in Python.

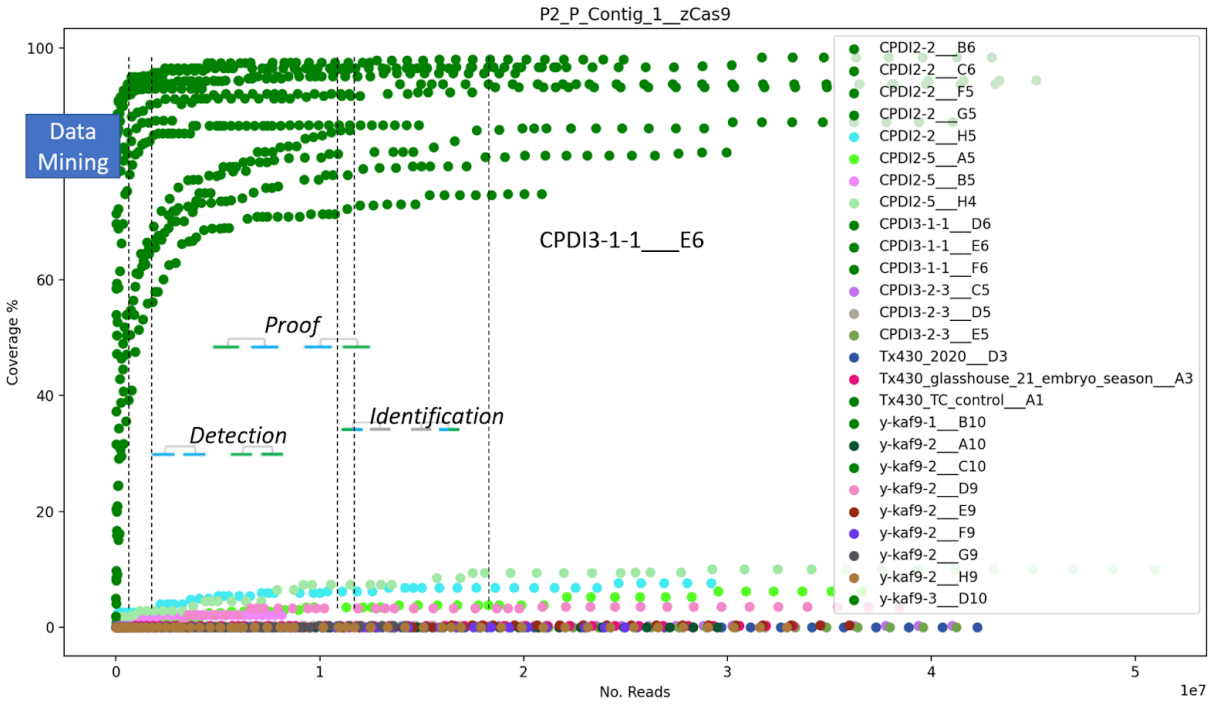

Figure 8. Relationship between Number of Raw Reads and Coverage % for different samples in the P2_P_Contig_1_zCas9 vector. The samples represented in dark green include the transgenic samples and C10, D10 that tested positive in earlier tests. Other samples are plotted in random colors. Each dot represents the plasmid coverage at a certain amount of raw reads. The right-most vertical dotted black line represents the convergence threshold determined visually for all samples. The middle vertical dotted black lines represents the statistical threshold to identify the insertion site (right) and to prove transgene integration (left). The one to the left represents the statistical threshold to detect the transgene in the samples. The left-most vertical dotted black line represents the estimated raw read count required for utilising a trained model based on supervised learning. The sample name in the middle represents the scatter-dot line with lowest converged coverage. Plotted with matplotlib module in Python.

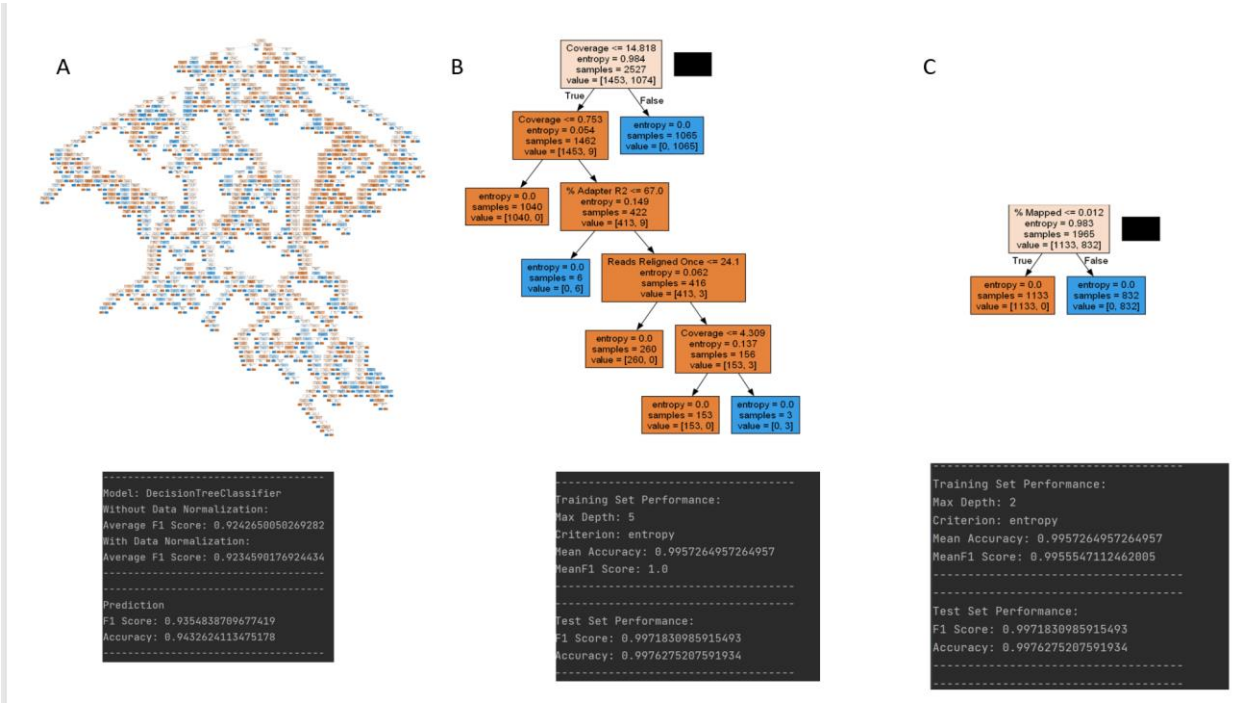

Figure 10. Number of Total Reads Needed to Achieve 95% Confidence for Transgene Detection Events in Sorghum (Upper) and Sugar Beet (Lower). Detection refers to the detection of transgenes; Proof refers to proof of transgene integration into the genome; Identification refers to the identification of the insertion site. The numbers are in millions.

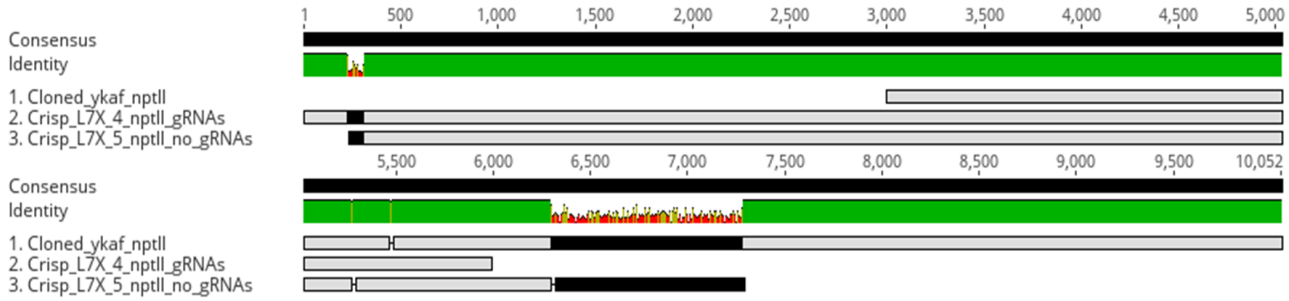

Figure 11. Alignment of Resequenced Cloned_ykaf_nptII Vector. The alignment displays three sequences: Sequence 1 represents the original sequence of the vector, Sequence 2 represents an alternative variety of the vector, and Sequence 3 represents the updated sequence of the vector. The matching regions between the sequences are depicted in grey. Aligned with in-built algorithm in Genious Prime.

Impact of reads on detection efficiency

The decision tree models (Figure 11) were developed to predict transgene presence in sorghum samples using Next-Generation Sequencing (NGS) data. The models were trained using a specific parameter, max_depth, which varied across the trees until they reached high levels of accuracy and F1 score.

These models offer an example approach for classifying sorghum samples as transgenic or non-transgenic based on key features, such as coverage values, extracted during the bioinformatic pipeline (Figure 10). By evaluating these parameters at each decision node, the models suggest reliable parameters regarding transgene integration. For instance, a threshold of coverage >14.818% and a mapped percentage <0.012% may be indicative of non-GM.

IV. DISCUSSION

Overview of the Gene Technology Regulation, 2001

NGS technology offers significant advantages in addressing the criteria outlined by the Gene Technology Regulation [12] to determine whether an organism should be classified as a genetically modified organism (GMO).

One key criterion is the identification of heritable changes within the organism. NGS allows for a comprehensive analysis of the organism's genetic material, providing a high-resolution view of its entire genome. By sequencing the DNA, NGS can accurately detect and identify any heritable changes present, enabling the assessment of whether genetic modifications have occurred.

Another important criterion is the introduction of foreign DNA from a different species. NGS plays a vital role in addressing this criterion by facilitating the detection of foreign DNA fragments or transgenes within the organism. By sequencing and comparing the organism's DNA to reference genomes, NGS can identify variations and differences that indicate the presence of foreign genetic material, aiding in the determination of whether the organism contains DNA from a different species.

NGS offers a comprehensive approach to analyzing the entire genome, including both coding and non-coding regions. This capability allows for the detection of specific genetic alterations introduced through genetic engineering techniques, such as the insertion of transgenes. By examining the complete genetic makeup of the organism, NGS ensures a thorough assessment of potential genetic modifications.

Reasons for incomplete coverage of plasmids

The incomplete coverage of plasmids can be attributed to several factors, including mapping errors, limitations of the technology, and errors in the plasmid reference sequence.

Aside from partial plasmid integration, which was not carefully examined in this project, mapping errors can occur during the alignment process when the reads from the sequencing data are mapped to the reference genome or plasmid sequence. These errors can arise due to the presence of repetitive elements or regions with high sequence similarity, which can make it challenging for the mapping algorithms to accurately assign the reads to their correct positions. In that case, Bowtie2 would treat them as noise and delete the reads [7]. As a result, some regions of the plasmid may remain unmapped or have incomplete coverage.

The limitations of the sequencing technology itself can also contribute to incomplete coverage. Certain sequencing platforms may have limitations in read length, accuracy, or depth, which can impact the ability to capture the entire plasmid sequence. For example, pair-end reads are limited to detect large indels. Additionally, variations in library preparation techniques or sequencing biases can lead to uneven coverage across the plasmid, resulting in gaps or regions with lower read representation.

Errors in the plasmid reference sequence can further contribute to incomplete coverage. Moreover, the varieties of the sorghum sample are Tx430, but the reference genome was Tx623. If the reference sequence used for mapping contains inaccuracies or incomplete information

about the plasmid, it can affect the alignment process and lead to missing or misaligned reads. These errors can arise from issues in the assembly of the plasmid sequence or from variations between the reference sequence and the actual plasmid being analysed.

To address that concern, we have attempted to resequence the vectors, only the Cas9 vector was successful. By comparing the updated sequence, it appears there are large indels at 1 to 3000 bp and 7500 to 10000 bp (Figure 11). The plasmid design [8] is not included here due to lack of software and time constraints, this could be investigated in the future.

Feasibility of NGS for Transgene Detection

It was determined that a total of 150k reads are required for successful detection, while an additional 100 times that amount is needed for the remaining analysis (Figure 6). Notably, the data revealed that as the number of raw reads increased, the coverage exhibited a steep upward trajectory, reaching higher levels with greater sequencing depth. This exponential growth pattern was only observed for the transgenic samples, indicating that it is primarily driven by the accumulation of reads from the target plasmid rather than random noise. These findings highlight the critical role of the number of reads in accurately capturing and quantifying plasmid sequences within the tested samples.

Classifying transgenicity using NGS data poses several challenges. Firstly, there is currently no standardised approach for classifying transgenic plants using NGS data, making it difficult to establish a universally accepted classification system. This lack of standardisation introduces ambiguity and variability in the interpretation of NGS results, hindering accurate transgene detection.

Another challenge lies in the potential for false positives. The presence of background noise, such as sequencing errors or alignment artifacts, can lead to the incorrect identification of transgenes, resulting in false positive classifications. This highlights the need for rigorous quality control measures and robust bioinformatic pipelines to minimize false positives and ensure reliable transgene classification.

In this project, to deal with that situation, we created a combined file containing both the sorghum genome and plasmid samples. However, that increases the computational space and time during the mapping step, making the process less scalable for large-scale studies, and applications to other transgenes.

Furthermore, utilising NGS data for transgene classification requires proficiency in programming languages like Bash and familiarity with specific tools like Bunya. This presents a barrier for researchers without programming expertise, making it challenging to effectively utilise existing scripts and tools developed by experts in the field.

Addressing these challenges requires the establishment of standardised guidelines for transgene classification, the development of robust bioinformatic tools accessible to researchers of varying programming backgrounds, and the optimization of computational workflows to improve efficiency and scalability in transgene detection using NGS data.

Hence, I tested a supervised learning approach. In theory, the machine could learn characteristics of transgenic samples from NGS data. It would be able to generalise the key parameters to classify GM. Although I achieved prediction accuracies up to 99.8%, I believe it was overfitted because of the limited number of sample sizes. The iterative software to run the scripts was developed in the last week of the project. Hence, the training dataset only contained features of 55 percentages x 26 samples x 2 vectors bioinformatic runs. Still, with the identified key parameters like coverage >14.818%, this could greatly reduce the amount of total reads required to achieve high levels of confidence. Thus, libraries could be prepared with more samples for transgene detection with NGS, enabling the regulation of transgenic non-GM crops in the field.

V. IMPLICATIONS AND FUTURE DIRECTIONS

Significance of NGS-based transgene detection

NGS is a powerful tool that enables the generation of an extensive number of reads, facilitating comprehensive analysis and statistical inferences. In the context of this study, an important finding was the determination of a confidence threshold of 14 million total raw reads. This threshold ensured an accurate representation of the transgene insertion site through a diverse range of paired-end reads.

To this project, the identification of this threshold was essential as it marked the point at which the coverage began to converge. It means the transgenic sample could be diluted in the form of a combined library that contains at least 1 GM sorghum in 10 samples. It indicates the reliability and accuracy of the NGS approach for regulatory agencies to use it as a bulk method to detect GM sorghums in the field as the cost for sequencing decreases with more samples.

Furthermore, although not specifically investigated in this project, the findings support the feasibility of using NGS to identify the precise insertion site of a transgene. The results obtained in this study reveal that the coverage of the transgene insertion site converges near its maximum level before reaching the threshold required for site

identification. This convergence of coverage indicates the applicability of the developed model to sorghum samples, showcasing the potential of NGS to offer researchers a rich dataset of informative information. By harnessing this technology, researchers can gain deeper insights into the behaviour of transgenes integration and genetic modifications, empowering them to draw robust conclusions and advance our understanding in this field.

Potential applications in quality control, experimental validation, and regulatory compliance.

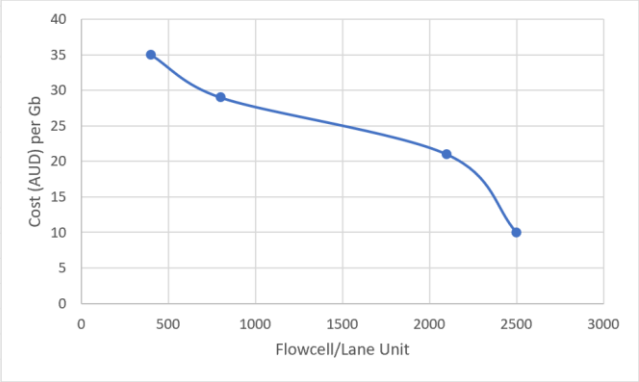

Figure 12. Relationship between Sequencing Cost and Lane Unit with Illumina's NovaSeq. The data is provided by Dr. Peter Crisp.

The findings of this project have significant implications for various applications in quality control, experimental validation, and regulatory compliance in the field of GM organisms. The cost-effectiveness of NGS-based transgene detection at large-scale sequencing (Figure 12), coupled with its ability to provide comprehensive insights into the genetic makeup of organisms, positions it as a valuable tool for bulk GMO detection by regulatory agencies.

One potential application is in quality control processes, where NGS can be utilised to assess the presence of transgenes in large-scale samples. By implementing the proposed supervised learning models or the identified threshold of 150,000 reads for detection, regulatory agencies can streamline their processes and handle a larger number of samples within a shorter time frame. This scalability enhances the efficiency of regulatory assessments and encourages advancement in gene-editing technology.

NGS-based transgene detection also holds promise for experimental validation purposes. Researchers can leverage the high-resolution view of the genome provided by NGS to validate the successful integration and precise insertion site of transgenes in genetically modified organisms. This allows for accurate characterization and verification of genetic modifications, supporting research efforts and enhancing the understanding of transgenic traits and their effects on organisms. With the proposed machine learning approach, the technical barrier to use the technology is also reduced.

Moving forward, future research and development efforts should focus on standardising the classification of transgenic plants using NGS data. The establishment of consistent guidelines and protocols for data interpretation will enhance the reliability and comparability of results across different laboratories and research studies. Additionally, the development of user-friendly bioinformatic tools and workflows, accessible to researchers with varying levels of programming expertise, will further democratise the application of NGS for transgene detection and analysis.

VI. CONCLUSION

In conclusion, the cost-effectiveness of NGS, coupled with its ability to provide comprehensive genomic insights, offers promising potential in the realms of quality control, experimental validation, and regulatory compliance for GMO detection. By harnessing the power of NGS technology, regulatory agencies can efficiently detect and monitor transgenic organisms, ensuring the integrity and safety of agricultural products and facilitating compliance with GMO regulations in the field.

VII. REFERENCES

1. Andrews S, Others. FastQC: a quality control tool for high throughput sequence data. Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom; 2010.

2. Bashandy H, Teeri TH. Genetically engineered orange petunias on the market. Planta. 2017;246: 277–280.

3. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 2016;32: 3047–3048.

4. Garreta R, Moncecchi G. Learning scikit-learn: Machine Learning in Python. Packt Publishing Ltd; 2013.

5. Kosicki M, Tomberg K, Bradley A. Repair of double-strand breaks induced by CRISPR-Cas9 leads to large deletions and complex rearrangements. Nat Biotechnol. 2018;36: 765–771.

6. Krueger F. Trim Galore: a wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files, with some extra functionality for …. URL http://www bioinformatics babraham ac uk.

7. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9: 357–359.

8. McGuffie MJ, Barrick JE. pLannotate: engineered plasmid annotation. Nucleic Acids Res. 2021;49: W516–W522.

9. Norris AL, Lee SS, Greenlees KJ, Tadesse DA, Miller MF, Lombardi HA. Template plasmid integration in germline genome-edited cattle. Nat Biotechnol. 2020;38: 163–164.

10. Ramírez F, Dündar F, Diehl S, Grüning BA, Manke T. deepTools: a flexible platform for exploring deep-sequencing data. Nucleic Acids Res. 2014;42: W187–91.

11. Tuladhar R, Yeu Y, Tyler Piazza J, Tan Z, Rene Clemenceau J, Wu X, et al. CRISPR-Cas9-based mutagenesis frequently provokes on-target mRNA misregulation. Nat Commun. 2019;10: 4056.

12. Victorian Government - Office of the Parliamentary Counsel. Gene Technology Regulations 2001 153/2001. Victorian Government - Office of the Parliamentary Counsel; 2001.

13. Willems S, Fraiture M-A, Deforce D, De Keersmaecker SCJ, De Loose M, Ruttink T, et al. Statistical framework for detection of genetically modified organisms based on Next Generation Sequencing. Food Chem. 2016;192: 788–798.

VIII. ACKNOWLEDGEMENTS

Special thanks to Dr. Peter Crisp, Dr Karen Massel, and Prof Ian Godwin for this amazing experience.

VIIII. SUPPLEMENTARY DATA