Genome analysis of the repetitive DNA organisation in the 45S rDNA of *Musella lasiocarpa*

Student Number: 209005657

Supervisor: Prof. Heslop-Harrison

Supervisor Department: Genetics and Genome Biology

Word Count: 5798

Contents

List of figures

List of tables

Contents

Abstract

Background

In recent history genome analysis has been central to the categorisation of species. While it served for other uses such as locating genes that control the phenotypic aspects of an organism. However, with the highly conservative nature of the functional genes it has been necessary to use exploratory analysis of the genomes to effectively differentiate species contained within a family. The *Musaceae* family has been of significant inconvenience when classifying. In this studied Nanopore technology was used to analyase the 45S rDNA looking for a higher order and possible pattern, due to variation, to the tandem repeats found in the sub-monomers of the 45S rDNA. Further exploratory data analysis was undertaken to look for any defining features that could be used for future research or further categorisation.

The *Musaceae* family is of substantial importance to the economy as it contains the species of both the banana and plantain. Bananas are the most consumed fruit on the planet as well as supporting several of the sub-Saharan and tropical economically (Robinson and Saúco, 2010) While many of the species with in this family have been sequenced and analysed *Musella lasiocarpa* has yet to be fully studied. With plethora of uses, including medicinal, its genome could be key to helping humans overcome future struggles.

Findings

Using the novel technique of Nanopore technology and a bioinformatics programme (geneious) the 45S monomer of *Musella lasiocarpa* was found to consist of sub monomers with average lengths of 8,992bps regularly occurring with 8 tandem repeats with an overall GC richness of 56.7%. Insertions and deletions occurred causing variance to the regular sub monomer. No trend was found in relation to the variance in the tandem repeat number concluding no higher order was present. A sub monomer length of 8.9kb is fairly regular for the 45S rDNA in plants. Exploratory analysis exposed a high number of insertions that were fairly regular throughout the 45S sequence which averaged 58.8kb in length and more commonly appeared before the tandem repeats.

Conclusions

Further investigation is needed to determine any possible significance to the insertions regularity and length. Though this study unveiled new knowledge on the DNA sequence of *Musella lasiocarpa* that has previously never undergone such detailed analysis and can be used toward full sequencing of the genome. This specific information can be used in the taxonomy of the *Musaceae* family.

1.0 Introduction

Throughout the years the study of genomes has been essential for characterising and measuring biodiversity and it origins during evolution. The utilisation of species DNA in the analysis of intricate ecosystems have been crucial in identifying taxonomic families, separating species and genera, and finding their phylogenetic relationships. Understanding the genome has also allowed evolutionary mechanisms to be determined and genetic diversity to be exploited for breeding programs that have helped to improve crop varieties through utilizing various genes. This fundamental study has provided the essential building blocks necessary to protect and preserve species as well as identifying genetic adaptations that can be advantageously exploited for crop production, medicinal uses and finding genes and genotypes with disease resistances.

1.1 The *Musaceae* family

The *Musaceae* family, containing the banana and its wild relatives, includes vital crops for local consumption and global export trade. *Musaceae* is a major part of some tropical and subtropical ecosystems. Within the small *Musaceae* family seen in Figure 1, there are only three genera of *Musella*, *Ensete* and the true bananas within *Musa*. *Musella* is a monotypic genus, with only the species *M. lasiocarpa,* the target of this project*. Musella lasiocarpa* (MLA) is known by many colloquial names such as Chinese dwarf banana, golden lotus banana or Chinese yellow banana. Originally MLA was found at altitudes of 2500m in the Chinese mountains of the Yunnan Province. MLA has 2n =18 chromosomes, while some other species in the family are 2n=22 (Šimoníková *et al*., 2022).

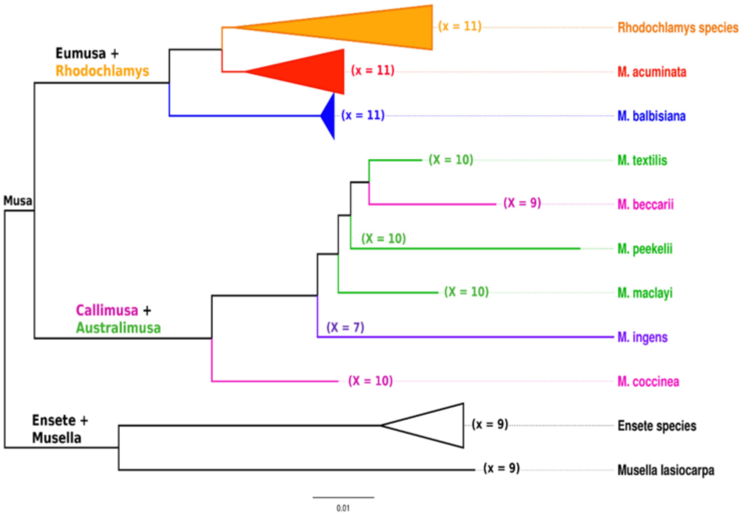

Figure 1 – Phylogenetic tree of the *Musaceae* family from Šimoníková *et al* (2022). This tree shows the chromosome number (x) as well as all the family relations.

1.2 Genome structure and assemblies

Several of the *Musaceae* species have complete assemblies of their DNA sequence at the chromosome-scale (reported in Banana Genome Hub (BGN); Rouard et al. 2022). Within the University of Leicester laboratory, an end-to-end haploptype assembly of the two chromosome sets in the diploid species *Musella lasiocarpa*, the most complete of any *Musaceae* species, is being completed and was used in this project. The assembly is currently being uploaded and analysis completed on the Banana Genome Hub.

The plant genome consists of both genes and large numbers of repeats with some organisms containing as little as 5% of their genome composed of nucleotide sequences that only appear once, including the genes (Michael and VanBuren, 2020). This repetitive DNA has become very important for the study of evolutionary genetics and taxonomy due to the copious differences and rapid evolution displayed in the repetitive regions. Importantly these differences are even present between closely related species. The genes of organisms are known to have high synteny across taxonomic groups therefore revealing very little information on the general chromosomal organisation of the organism that is primarily composed of repetitive DNA.

Long, single molecule DNA sequencing with Nanopore technology was used as it has the capability of sequencing huge lengths of DNA; such as the repetitive DNA sequences found in plant chromosomes where repetitive motifs will be spanned by single reads so each monomer can be examined, unlike use of short reads. The long-molecule sequencing technique has been commercialised and it is often referred to as Oxford Nanopore technology (ONT) and shorthand as Nanopore. In the case of repetitive DNA, which consists of short monomers in tandem repeats, long-molecule sequences, Nanopore has for the first time enabled the variation and organisation to be analysed thoroughly in stretches of repetitive DNA. Compared to other techniques such as illumina sequencing, which only has a 150-base read, it is far more effective and appropriate for this project. Even analysis of large insert clones (eg BACs >50kb) has proved difficult with chaemerism and mis-assembly of short reads. Nanopore can read 500-2.3 million base pairs (Heidelberg, 2005) at one time though it does have a relatively high error rate of 8%. With such large scale sequencing the errors are not a hindrance. For a lower error rate the Nanopore sequence can be compared the 150 reads of 2llumine sequencing to line the smaller reads into order while also checking for any errors in the Nanopore sequence.

In this project, I specifically looked at the tandem repeats within the 45S rDNA. The 45S rDNA region is highly conserved DNA. This region is made up of several units and can be seen in Figure 2. These units are made up of genes coding for 18S 5.8S and 26S rRNA of which are all separated by internal transcribed spacers. Each 45S unit is separated by non-transcribed intergenic spacers, with the three genes relatively conserved across all eukaryotes and much more variation in the spacer regions. As 45S rDNA sub monomer is around 9kb in the *Musella* species length (Hřibová et al., 2011), the Nanopore was essential to analyse multiple repeat monomers.

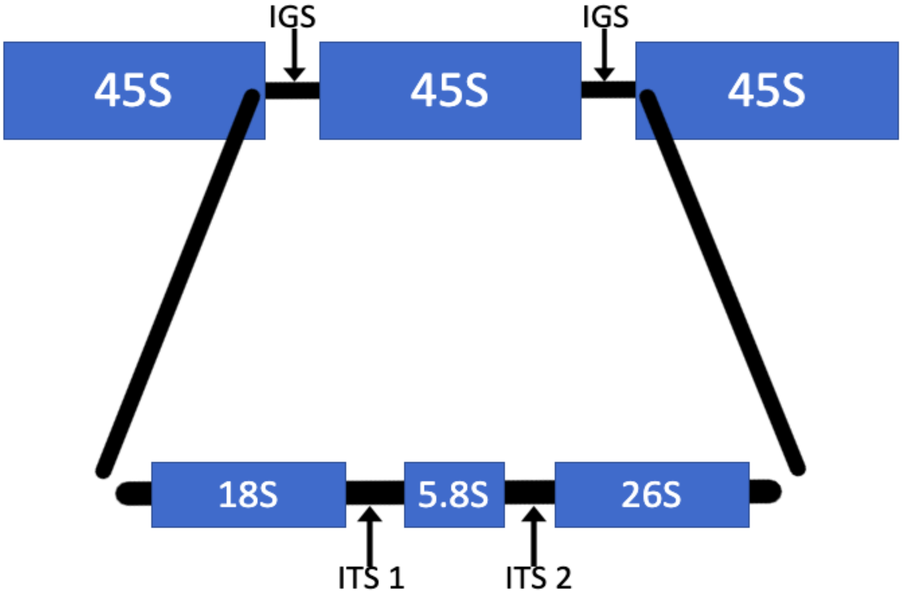

Figure 2 - 45S rDNA structure (adapted from project presentation). The 45S transcription units are separated by non-transcribed intergenic spacers (IGS). Within the 45S rDNA locus contains genes for 18S, 5.8S and 26S rRNA and are separated by internal transcribed spacers (ITS1 and ITS2).

1.3 Overview and aims

There are two different sections of my project. The 'dry' lab using bioinformatics and the 'wet' lab conducting FISH (fluorescent in situ hybridization) experiments and using DNA in solutions. Two approaches were used to obtain the necessary data for analysis. When looked at as a whole these add up to dot plots to compare DNA sequences and references (in this case, made using the program Geneious). When these plots were created certain variability was investigated while analysing the monomers.

My aim was to identify variability and possible higher orders of tandem repeats within the 45S rDNA. In conjunction with this I intended to undergo exploratory analysis of conservation, homogenization and variation within the 45S rDNA monomers. This was done by using novel techniques exploiting long-read sequences to find a possible higher order within the tandem repeats.

2.0 MATERIALS AND METHODS

2.1 Wet lab method

The methods used in my project have been adapted from Schwarzacher and Heslop-Harrison (2000) and Schwarzacher (2016) for preparation of plant chromosomes and fluorescent in situ hybridisation. Bioinformatics generally built on approaches developed within the project laboratory and used by Zang et al. 2022, D'Hont et al. 2012 and Zang et al. 2023.

1. Plant material preparation and root tip fixation

Plants of *Musella lasiocarpa* (MLA) were grown in large 10litre pots in the Adrian roof glasshouse at 28°C. Newly growing roots were collected by pulling the potted MLA plant out and removing the root tips, the whiter end of small and medium new roots. Clean forceps and scissors were used to place the root tips into clean Bijou tubes containing about 4ml 2mM 8-hydroxyquinoline to accumulate metaphases; they were left for 2 hours at room temperature (19-25°C) and then transferred to 4°C for 4-6 hours.

Seeds of barley (*Hordeum vulgare*) cv Fabiola were germinated by wetting filter paper (with bottled water) in a petri dish and incubating at 25°C in the dark. The new root growth was then cut off and placed in aerated ice water for 24 hours to amass metaphases. After metaphase arresting, roots were fixed in 3:1 ethanol: acetic acid solution. Fixations were stored for 2-16 hours at room temperature (19-25°C) and then transferred to -20°C.

2. Chromosome preparations

Fixed roots were transferred into a sterilised petri dish using sterilised forceps. The roots were then washed 2 to 3 times in an enzyme buffer (10x enzyme buffer of pH4.6 dilute with distilled water to 1x; 10x: 60ml 100mM of tri-sodium citrate and 40ml 100mM citric acid) for approximately 10 mins. Buffer is pipetted out and then the petri dish is refilled with the buffer to submerge the roots.

It was ensured that the roots remained intact and the tip unbroken. Any dirt remaining was then removed and any excess root was cut off (1cm was all that was needed). Once the roots had sunk they were transferred into the enzyme solution (for MLA: 3.2% cellulase C1184; 13% pectinase P4716; 0.4 % Onozuka cellulase and 8% viscozyme; for barley half this strength) for digestions at 37 °C for 60-90 mins for barley, and 4 hours for MLA. When the roots were soft, the enzyme was removed and replaced by the enzyme buffer. The roots were left for several hours at room temperature or overnight at 4°C.

Roots were moved onto a sterilised Superfrost microscope slide using clean forceps. A drop of acetic acid (60%) was then placed onto the root in the middle of the microscope slide which was then placed under a stereo microscope. Sterilised forceps together with a syringe (dissecting needle or any fine sharp appendage can be used in its place) were used to extract the meristem. The root cap and surrounded material was removed and discarded. A cover slip was then cautiously placed over the sample carefully to avoid air bubbles; these were removed by lightly tapping on the cover slip using a syringe. Excess fluid was then dabbed away using filter paper that was pressed over the edges of the cover slip. Folded filter paper was positioned round the slide and cover slip and pressed against a flat surface with force to squash the cells. After this, the slide was examined under the phase contrast microscope to locate chromosomes in metaphase and then moved to dry ice for 5-10 mins. The cover slip was then removed from the slide using a sharp razor blade to flick it off the cover slip. After air drying, the slides were stored between 4°C to -20°C for the duration of the experiment and extracted when needed.

3. Fluorescent in situ hybridization

Pre-treatments

Post fixation was then undertaken using a solution of ethanol and acetic acid at a (3:1) ratio for 10-30 minutes. Then the slides were treated with 100% ethanol for 5 minutes which was then repeated and left to air-dry. Slides were then treated with RNase in 2xSSC (diluted DNase-free RNase stock at 1mg/ml), 1:10, final concentration: 100/ml). For this, 200µl of RNase were added to each slide and a plastic coverslip was placed on top. Slides were incubated at 37°C for 1 hour, and then washed in 2xSSC for 2 minutes, and then again for 10 minutes.

The slides were then exposed to pepsin treatment and incubated for 2 minutes with 0.01M of HCL applied. Pepsin stock (1mg/ml) was then diluted with HCL (0.01M) at a 1:200 ratio to form 5µg/ml of Pepsin. This was then applied to the slide (200µg) then a cover slip (plastic) was placed on top of the slide. Once incubated for 10 minutes (37°C) the slide was rinsed (with distilled water) and washed in 2xSSC for at least 5 minutes. Using premade paraformaldehyde solution, the slides were placed in the fume hood for incubation in the solution for 10 minutes at room temperature. Then the slides were washed twice, first for 2 minutes, then for 10 minutes in 2xSSC.

Finally, the slides were dehydrated beginning with a wash in 70% ethanol for 2 minutes; followed by two more washes that increase in percentage of ethanol (85% then 100%) for 2 minutes each. The slides were then left to air-dry.

Hybridisation

The hybridisation mix was calculated as shown in Table 1 which was then used in the calculation and production of the probe mixes as seen in Table 2 using probes as detailed in Table 3. This was then added to the slides that had been pre-treated, in the centre where the material had been fixed. With a cover slip (plastic) these slides were placed into a thermal cycler and denatured. The barley was set at 68.1°C for 10 minutes and then reduced to 37°C to cool down and left overnight (17 hours). MLA was set to 70.5°C for 3 minutes, reduced to 37.5 for 5 minutes and then left overnight (17 hours).

Table 1- Calculations for hybridisation mix.

Mixture Element	Stock Concentration	Final desired concentration	Final concentration in hybridisation mix	Amount for one slide	Mastermix 30% FA 2xSSC
Final volume	40µl	40mM			
No. of slides +1					20
100% formamide	100	30	30%	12	240
20xSSC	20	2	2x	4	80
50% Dextran sulphate	50	10	10%	8	160
Salmon sperm 1 µg/µl	1	0.025	0.025µg	1	20
100mM EDTA	100	1.25	1.25mM	0.5	10
10% SDS	10	0.125	0.125%	0.5	10
Total mastermix (µl)				26	520
Probe and water (µl)				14	280
Total Hybridisation mix (µl)				40	800

Table 2- Calculations for hybridisation of probes. The below table gives the calculation for the slides processed by the group, my slides are highlighted in bold.

		A	B	C
Red		45S rDNA	5S rDNA	5S M
Green		5S rDNA	45S rDNA	5S rDNA
Species/slides		Fabiola Wilma Triticum **MLA**	**Fabiola** Wilma Triticum EGL	Fabiola Wilma Triticum
No. of slides		4	4	3
Master mix	26 per slide	30%	30%	30%
		104 µl	104 µl	78 µl
5SR-Cy3 (µl)			8	
5SR-Cy3 (µl)			8	6
5SL-Cy3 (µl)			8	
18SMR-Cy3 (µl)		10		
45S(7)-Cy3 (µl)		10		
5.8S-Cy3 (µl)		10		
5SR-FAM (µl)		12		9
5SL-FAM (µl)		12		9
18SM-FAM (µl)			10	
45S(6)-FAM (µl)			10	
45S(5)-FAM(µl)			10	
Sum (µl)		128	158	102
Water (µl)		32	2	18
Total Volume (µl)	40	160	160	120

Table 3 - Details of oligonucleotide probes synthesised with fluorchrome cyanine 3(CY3) or FAM at the 5'end:

Name	probe	Sequence	Label
5SR-FAM	5S rDNA	(6FAM)AGTACTAGGATGGGTGACCCCCT GGGAAGTCCTCGTGTTGC	FAM (green)
5SL-FAM	5S rDNA	(6FAM)GCGATCATACCAGCACTAAAGCA CCGGATCCCATCAGAACTCC	FAM (green)
18SM-FAM	45S rDNA	(6FAM)GAGCCTGAGAAACGGCTACCACA TCCAAGGAAGGCAGCAGG	FAM (green)
45S(6)-FAM	45S rDNA	(6FAM)GTCAACGCGAGCTGATGACTCGC GCTTACTAGGAATTCCTCG	FAM (green)
45S(5)-FAM	45S rDNA	(6FAM)ACGAGCTCCAGCTATCCTGAGGG AAACTTCGGAGGGAACCAG	FAM (green)
5SR-Cy3	5S rDNA	(Cyanine3)AGTACTAGGATGGGTGACC CCCTGGGAAGTCCTCGTGTTGC	CY3 (red)
5SM-Cy3	5S rDNA	(Cyanine3)TCAGAACTCCGAAGTTAAG CGTGCTTGGGCGAGAGTAGTAC	CY3 (red)
5SL-Cy3	5S rDNA	(Cyanine3)GCGATCATACCAGCACTAA AGCACCGGATCCCATCAGAACTCC	CY3 (red)
18SMR-Cy3	45SrDNA	(Cyanine3)CAAGAACGAAAGTTGGGGG CTCGAAGACGATCAGATACCGTCC	CY3 (red)
45S(7)-Cy3	45SrDNA	(Cyanine3)GGCATCACAGACCTGTTAT TGCCTCAAACTTCCGTGGCCTAG	CY3 (red)
5.8S-Cy3	45SrDNA	(Cyanine3)CCGTGAACCATCGAGTCTT TGAACGCAAGTTGCGCCCGAGGCC	CY3 (red)

Post-hybridisation washing

While all performed in a fume hood, 500ml of 2xSSC and 200ml of 0.1xSSC was made and placed in water baths of 45°C. Detection buffer was made alongside these solutions consisting of 4xSSC with 0.2% Tween. The slides were then taken from their incubation and added into a bath of 2xSSC until the coverslips became loose and floated off at 35°C-40°C. The slides were then washed again in 2xSSC at temperatures between 42-45°C for 2 minutes. Then washed twice in 0.1xSSC for 5 mins, first at 44°C then at 42°C. The final two washes were then completed in2xSSC for 5 minutes each and transferred into the detection buffer (5 mins) at room temperature.

4. Staining and mounting slides

200ul of DAPI solution (2 µ□g/ml DAPI (4',6-diamidino-2- phenylindole) diluted in McIlvaines buffer) was used to cover the slide and then a cover slip was placed over the slide. It was left for 30 mins at room temperature in a dark place. The slide was then washed with a fluorochrome buffer after removing the cover slip and left out to dry.

After washing slides a couple of drops (1 to 2) of antifade solution were added to each slide and a large and thin coverslip (40mmx24mm No.0), suitable for microscopy, was applied carefully avoiding any bubbles. The slide was placed between a piece of folded filter paper applying light pressure to remove excess antifade. Slides were stored for later use at 4°C.

5. Microscopy

A fluorescent microscope (Nikon 80i Eclipse) was used to analyse the slides. The shutters for the lighting were should be shut when initiating this process to not expose the slides unnecessarily and therefore fade the staining colour. The slides were used at room temperature
Just a drop of immersion oil was applied to the coverslip, taking care to cover the area of chromosome prep. The objective lenses were then moved round and out of the way in an effort to make loading the slide easier. Once the slide had been carefully loaded onto the stage of the fluorescent microscope the lenses were slowly moved into place with the 20x objective lenses used as the initial focal lens. This was done with the utmost care to not damage the slide if the stage is too high. The filter wheel is then set to DAPI and the shutter opened. The manual focusing knob is used to focus the image and the slide was scanned for a chromosome in metaphase. Once found the objective lenses were moved onto a higher objective (100x).

The image then needed to be refocused manually so that a photo could be taken, using DS-QiMc monochrome camera, and NIS-Elements v.2.34 (Nikon, Tokyo, Japan). We photographed Cy3 first, then FAM and then DAPI and saved these as nd2 files. The images were then enhanced and saved as jpg images.

2.2 Dry Lab Method

The dry lab was bioinformatics exploratory data analysis using nanopore sequences. In the dry lab, we used the programme Geneious using data that was obtained using the Nanopore reads. When the 6.1 chromosome (chromosome 6, haplotype 1) was laid over the 45S repeats where they matched, a dot is created. This generates the dotplots shown in Figure 3. These dot plots highlight the areas of tandem repeats as seen in panel B of Fig 1 and I used these to count the number of repeats. When analysing, there were certain variations that appeared regularly while analysing the monomers. The Fig3 shows a regular monomer. The box-like shape is the tandem repeats here; this one shows 8.

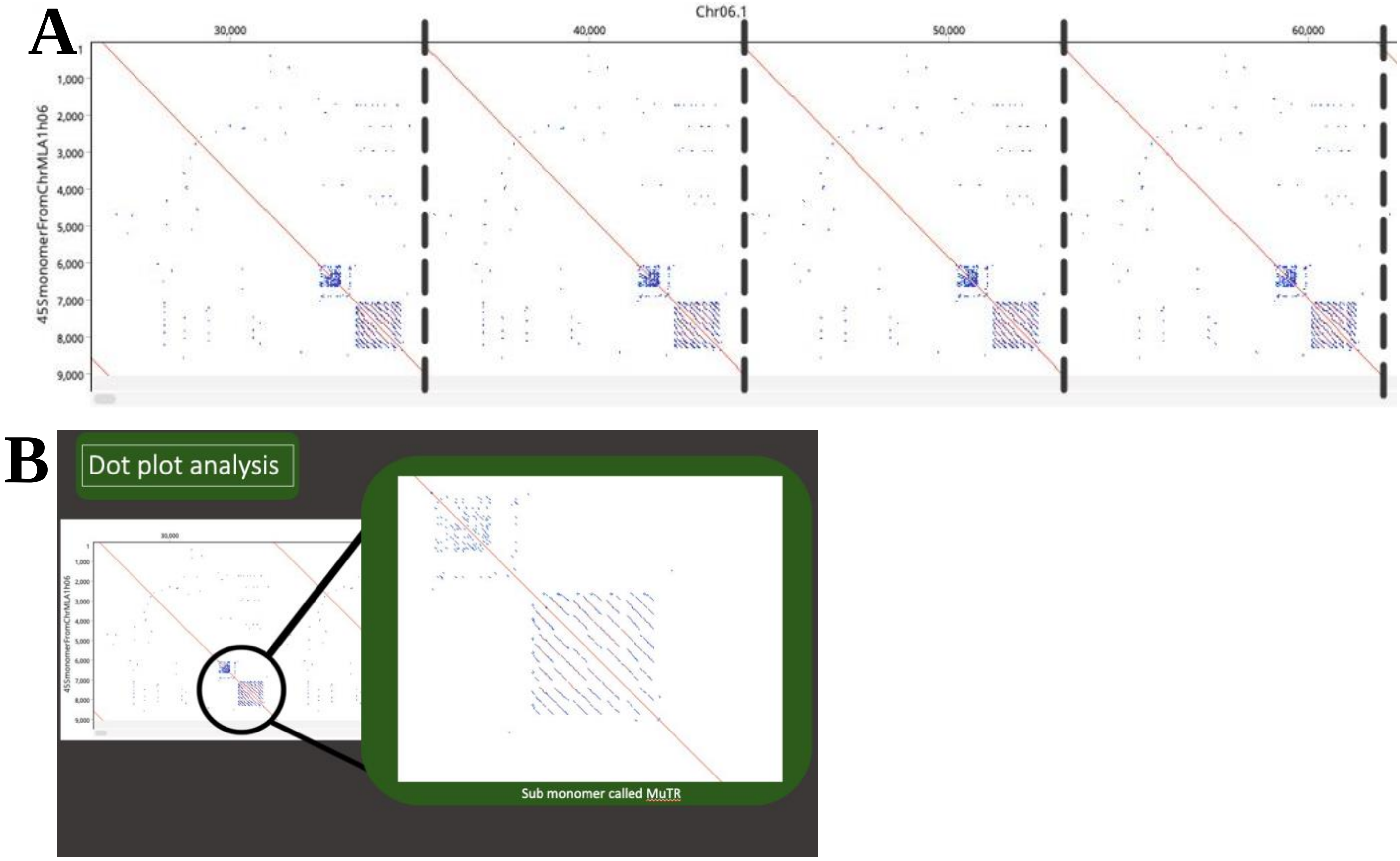

Figure 3– Generated dot plots of the 45S sub-monomer tandem repeats (MuTR; the consensus sequence, previously calculated, including 8 copies of an internal tandem repeat). (Used in project presentation). Panel A shows four monomers from the chromosome assembly, indicated by vertical dotted lines that were superimposed to show my interpretation of monomers from the chromosome assembly of the 45S rDNA. Chromosome 6.1 (assembly of chromosome 6 from haplotype 1) is shown on the X axis illustrating the number nucleotide bases while the Y axis shows also the bps but of the 45S monomer. Panel B shows an enlarged section of the monomer that highlights the MuTR. In this case the MuTR has 8 repeats as counted across the top.

Both insertions and deletions were found of the MuTR monomer, a tandem sub-repeat within the 45S monomer. These could occur before, after and within the tandem repeats of the monomer. This can be seen in Figure 4 with panel A showing the result of an insertion and deletion before the tandem repeats. When the insertions or deletions where found in the MuTR the number of repeats could be affected like in panel B and C from a deletion and insertion respectively. Though insertions could produce no changes in the repeat number like in panel D.

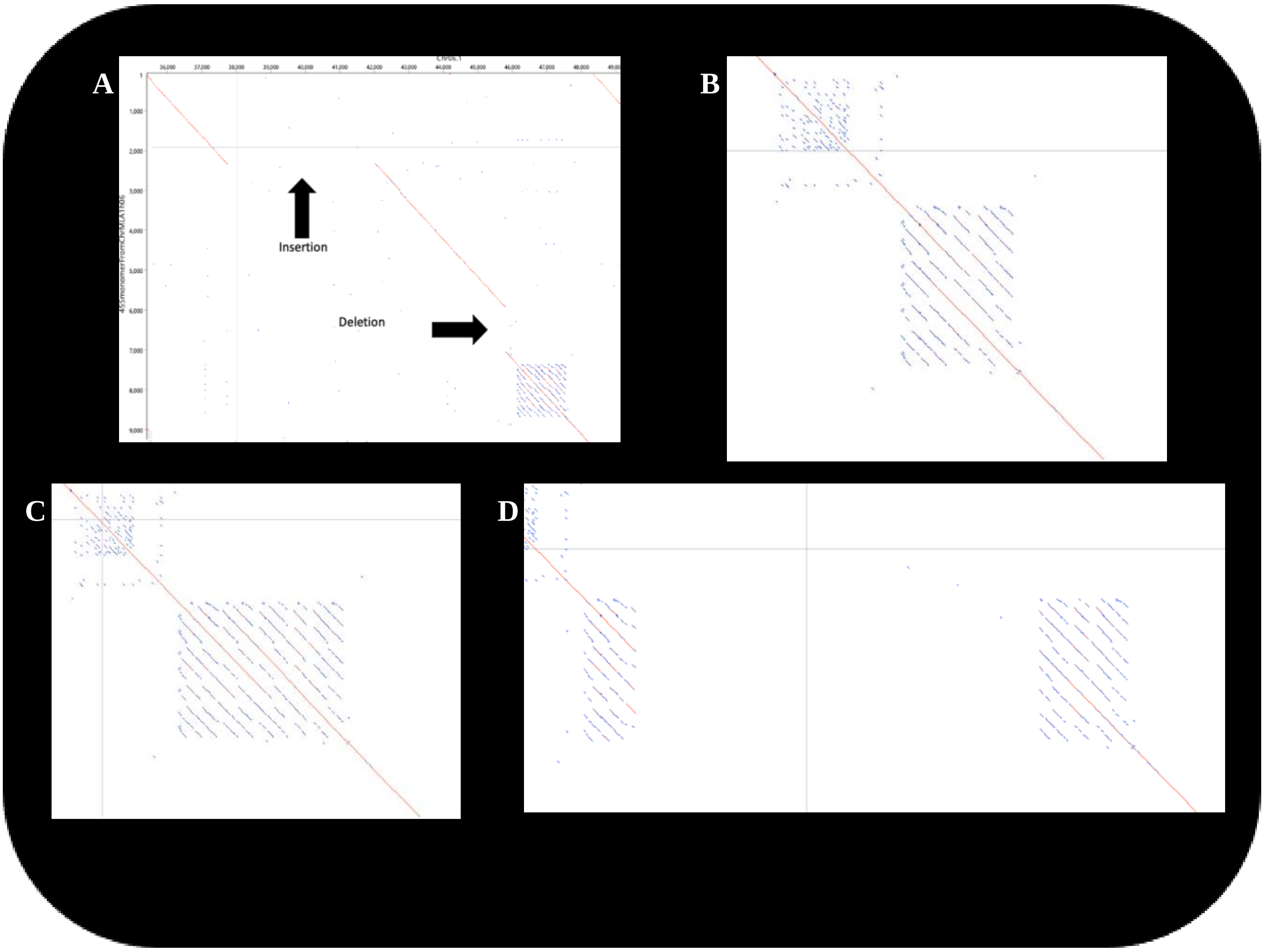

Figure 4– Variations found in the whole 45S sequence (A) and of the MuTR of the 45S monomer of the MLA chromosomes bioinformatics (Adapted from project presentation). Panel A shows an insertion before the tandem repeat as well as an insertion. Panel B and C show a deletion and insertion correspond with changes to the MuTR tandem repeat number. Panel D shows and insertion within the MuTR tandem repeat array that resulted in no change of the tandem repeat number.

3.0 RESULTS

3.1 Dry lab results

The dry lab revolved around conducting exploratory data analysis that lead to several points of interest with a focus on the sub-repeat within the monomer MuTR (the consensus sequence, previously calculated, including 8 copies of a tandem repeat). Initially, the number of repeats were recorded and consistently a repeat number of 8 was observed. Though there was some variation, as seen in Figure 5, allowing for the continuation into the exploratory data analysis in search for any possible higher order in the event that the MuTRs, which were not found to contain 8 repeats, appeared in any pattern. To visualise this, I put my findings into a graph seen in Figure 4. This graph showed no sign of a higher order but provides strong evidence for the repeat number being 8 in the majority of the 45S rDNA repeat units.

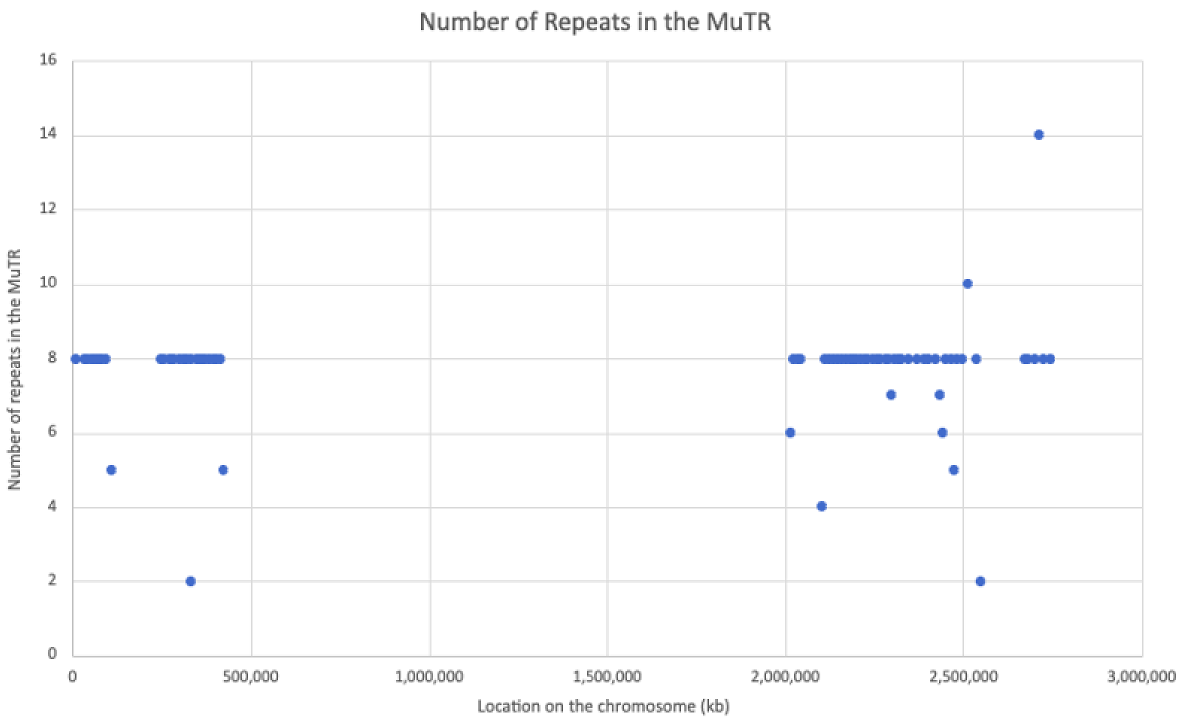

Figure 5– Scatter graph of the number of repeats found in the MuTR (tandem repeat) of the 45S region of an MLA chromosome (Used in project presentation). This graph shows the number of repeats that were found in 100 data entries of 45S monomers. No higher order to the number of repeats in the MuTR as it stays consistently at 8. Though there are changes in this there is no visible pattern of the other numbers of repeats.
MuTR – the previously calculated consensus sequence including copies of a tandem repeat

Throughout the analysis variations of the monomer consistently appeared throughout the data. Insertions were far more common and regularly appeared before the MuTR. This directed the exploration to focus on the length of insertion found within the MuTR.

Consequently, it was found that these insertions had a total average length of 58,879 bps. Subsequently, the data was compiled into scatter graphs of the lengths of the insertions that can be seen in Figure 6 panel A. This visual aid helped to highlight three significant data points. These three points where concluded as obvious outliers and were confirmed once the raw data revealed measurements of 54900bps, 130464 bps, 1582332bps. Compared to the other data with the highest measurement of 11962bp the smallest outlier was over double that of the rest of the data cohort. In panel B, you can see these outliers have been removed and revealed insertions had a general trend hovering just under 4500bps.

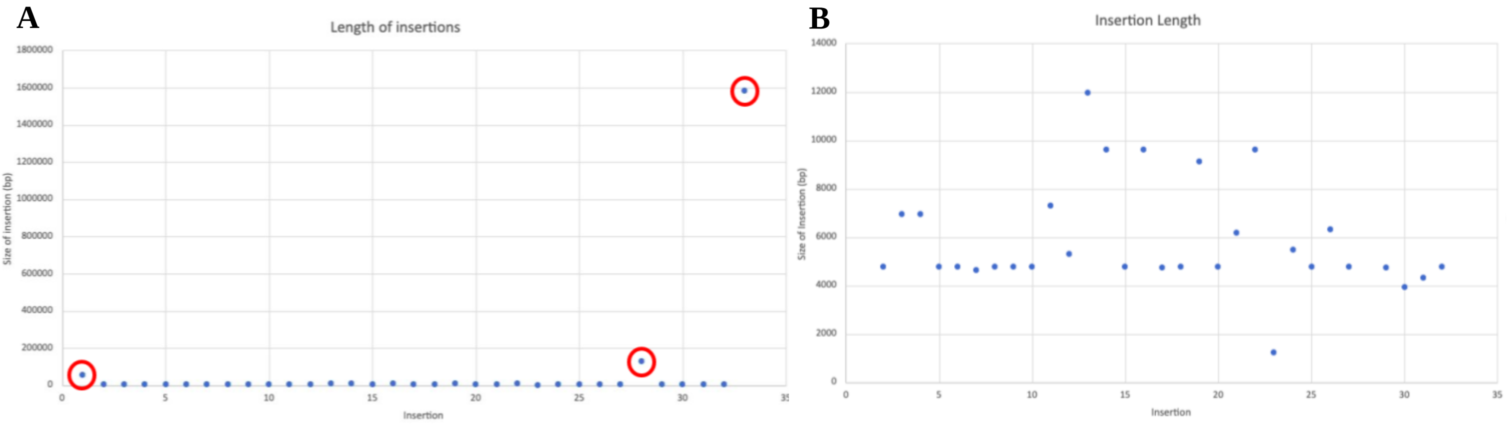

Figure 6– Scatter graph showing the length of insertions in the 45S monomer found on the 6.1 chromosome of *Musella lasiocarpa* (MLA) (Adapted from project presentation). Graph A shows the raw data of the length of insertion found. There are three outliers that have been circled in red. Once taken out the data for length of insertions in this MLA monomer can be seen in graph B. Once the outliers were removed panel B shows a general trend of insertions that are around 4500bps in length.

Finally, the overall length of the monomer was investigated. Efforts were concentrated on the regular monomers removing insertion lengths. This uncovered the average length of the monomer to be 8992 bps. Once placed into the bioinformatics programme it was also noted the sequence had a GC richness of 56.7%.

3.2 Wet lab results

The FISH experiments resulted in the photos seen in Figure 7 showing both the 5S region tagged with FAM (green) and 45S region tagged with CY3 (red). Both the control of barley and MLA showed both the 5S and the 45S regions. Though there are chromosomes missing from the MLA sample and there is not great separation from panel A Fabiola. Nevertheless, these visualisations show both probes were effective and produced clear signals for both the 5S and 45S regions on the chromosomes. This provides evidence of the presence of these repeat regions in the MLA genome, with only the sites detected in the genome assembly. However, there are missing sites as the barley should show 8 sites (Leitch and Heslop-Harrison, 1993) of the 5S sequence and 10 sites of the 45S sequences (Leitch and Heslop-Harrison, 1992) which are not all present in the panels A and B in Figure 7. Panel C does show a full number of chromosomes (2n=18) panel A is thought to have 2n=14 but this is hard to differentiate as does B.

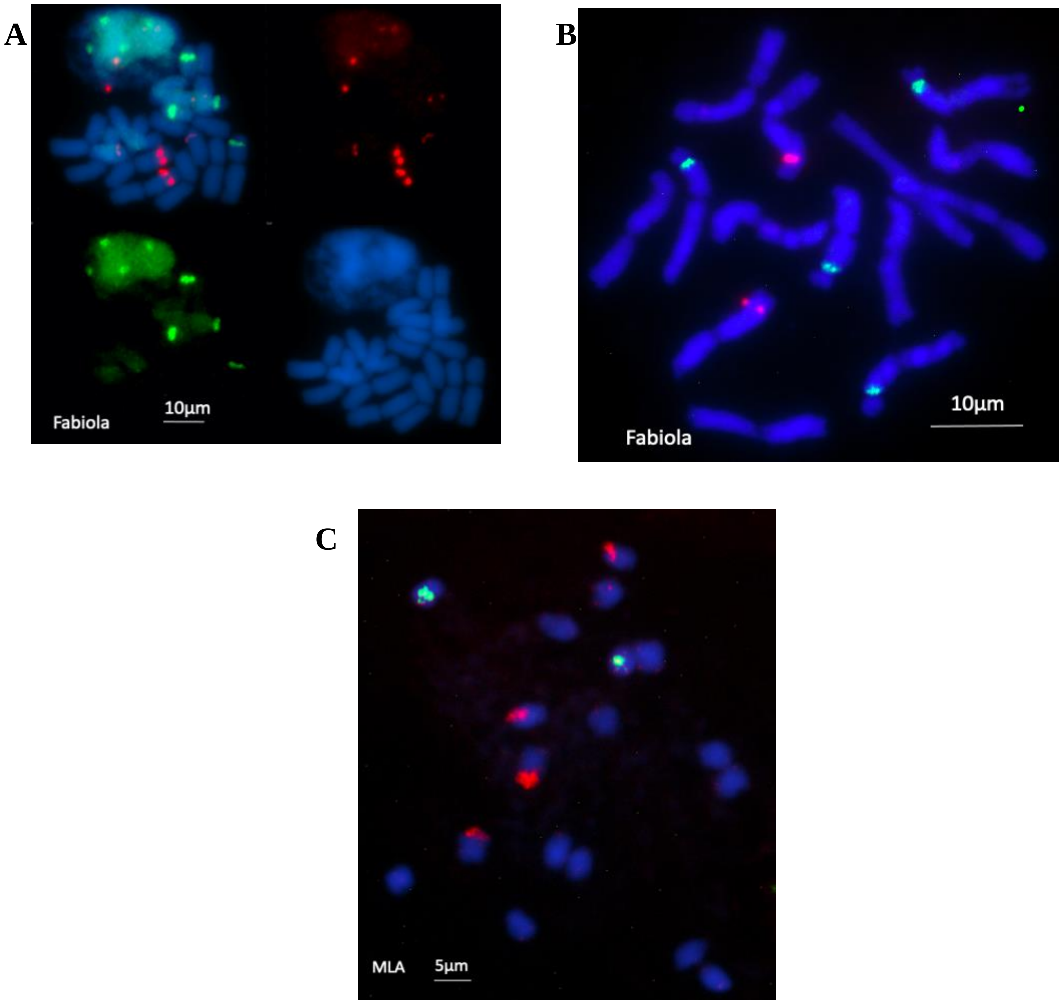

Figure 7– Photographs of Barley and *Musella lasiocarpa* (MLA) chromosomes in metaphase (Used in project presentation). All panels show the 5S and 45S regions that have probes that have been tagged with florochromes. The 5S region has been tagged with FAM and the 45S region has been tagged with CY3. Both panels A and B show Fabiola that is German spring barley and panel C shows *Musella lasiocarpa* (MLA) chromosomes.

4.0 Discussion

4.1 Wider Context

This information and knowledge of the banana family is not arbitrary and can be used in the wider community. The Banana family is a very important family with its huge contributions to humans. Bananas, the fruit, have been fundamental in many cultures, diets and economies. It is the most eaten fruit in the world when the species *Musa sapientum* and *Musa paradisiaca* are combined. After commodities such as coffee, cocoa and sugar these *Musa* species are considered the 5th most important in international trade of agricultural products (Josling and Taylor, 2003). Fully understanding the genomes of the surrounding species can be used in the modern world faced with difficulties such as climate change, famine and plant diseases. Knowing MLA is prevalent at high altitudes and adverse terrain could indicate a genetic advantage that could be utilised in the species that are more ubiquitous in our society. As described by (Janssens *et al.*, 2016). This leads into the concept of pan genomes which is also a big contributor to the motivation behind my project. As described by J. Wang *et al.* 2023 the pan genome is used as a way to enhance genetic analysis and moderate the biases formed when using a core genome for analysis. Having a broader view of the family's genomes allows for a more holistic view of family and intricacies of each species.

4.2 Importance of *Musella lasiocarpa*

The uses humans have for *Musella* have not always been documented well. Though it has been used in Chinese civilizations for centuries, predominantly for medicinal uses, it has sorely lacked documentation throughout ethnobotanical literature. Liu and Kress (2003) state the flowering appendage of the *Musella* plant has been used to decrease bleeding and infection when crushed and dressed on a wound. It has also been used as an antidote to the poisonous flower Aconitum (wolf's bane/monkshood) and reducing intoxication due to alcohol. When boiled it has been found to alleviate constipation, inflammation (enteritis) and other diseases. Yet, this is not the only way that MLA has been used as the pseudostem, formed of closely packed leaf sheaths, has been prepared in multiple ways for consumption. The 'waste' products from these uses are often harnessed for fodder; boiled leaves, pseudostem and rhizome are popular pig's food. Due to its natural preference for cool, arid and mountainous environments often *Musella* is the only greenery lasting through the winter months. Though it is highly cultivated for its many uses and decorative status it is limited in its wild habitat. These uses and ecological features are reasons for the investigation into its wider genome. Deeper understanding of the genome could lead to a deeper understanding of the abilities of these uses.

4.3 Evolution and importance of 45S rDNA

The 45S rDNA locus contains the genes for three Ribosomal ribonucleic acid (rRNA), 18S, 5.8S and 26S, that are all transcribed as one unit. On the contrary, the 5S rDNA only encodes for 5S rRNA genes. 45S rDNA genes are spliced after transcription. The genes within the 45S rDNA are separated by ITS1 and ITS2 which are internally transcribed spacers. ITS have been fantastic markers of evolution due to their comparatively fast evolution to the neighbouring stable rRNA locus (Hřibová *et al.*,2011). Ribosomal genes arranged in tandem repeats, such as the 45S rDNA in MLA, are thought to have closer relation to each other than other species (concerted evolution) as explained by Elder and Turner (1995). The evolution of these genes tends to be incredibly similar and occur in time possibly due to the high conservation of genes and the imbalanced crossing over occurrences. Comparatively, ITS are under significantly less selection pressure causing the increase in evolutionary pace. While the ITS of the 45S rDNA are not included in the final composition of the mature rRNA they are required for the signalling needed for the correct transcription of rRNA. Both 5.8S and ITS2 are highly conserved sequences and this conservation is the feature that is most useful in the identification of differences in paralogs of ITS spacers that result in pseudogenes. These aspects of the ITS regions make them of particular interest to phylogenetic studies and hence why this study looks into the 45S region in more detail.

4.4 Nanopore technology

Nanopore technology was used in this project due to its many benefits. It one of the only systems that can offer rapid insights into the subject genome with full sequencing of both small fragments of a genome and extremely long full-scale sequences. It is the technology uses protein pores (Nano scale) placed in an electro resistant membrane. The DNA is then able to pass through the holes with the individual nucleotides small enough to pass through the Nano pores. The slight charges associated with each of the different biological molecules change the current in the membrane and this potential change in the membrane is recorded. Each of the nucleotide bases causes a different change in the membrane potential and the bases are recognised and coded in real time. It is a fourth-generation sequencing technology that uses the advantages of both computer and gene engineering technology. This fusion created the Oxford Nanopore (ONT) that Lin, Hui and Mao (2021) describe as both powerful and innovative built off the back of second (Illumina) and third (PacBio) generation technology. In comparison to ONT can read far more base pairs in one read than any other method reaching 10 million base pairs (Bharagava *et al.*, 2019). It is also extremely cost effective with the lowest US dollar per Gb at 21-42 in comparison to Illumina of 50-63 US dollars (Lin, Hui and Mao, 2021) and even more expensive PacBio. It has a high input and output rate as well as having a low initial capital cost, the only portable system and requires no sample preparations such as chemical labelling or PCR amplification. There is a short fall with over all accuracy of the Nanopore sequencing having the lowest accuracy rate of 87-98% accuracy across the models. Both Illumina and PacBio have 99% accuracy. All these factors were considered when determining the needs of our experiment as further explained in section 4.5.

4.5 ‘Dry’ lab method advantages and disadvantages

This project allowed for further knowledge of the MLA genome to be uncovered. There has yet been a full exploration of the genome but using novel techniques such as the Nanopore to analyse the genome has been fundamental to the continuation of understanding the Musaceae family. Nanopore was instrumental in this study. Not only was its compact nature useful in the initial collection but it allows for extremely long reads which previously has been unattainable with quick turnaround times. This enabled access to reads of almost 9000bps, of the 45S sub-monomers, tandem repeats within a sequence of over 4 million base pairs. Compared to the 150bp of read ability of Illumina (closest competitor) the Nanopore technology is an impeccable asset. Though the Nanopore has a comparatively lower accuracy rate with possible error rates of 8% when sequencing such large reads this has little effect on the final result. However, in future Illumina reads could be used in conjunction with the Nanopore reads, while illumine can produce very accurate short reads the Nanopore read can be used as a reference genome to line all the Illumina reads into the correct order resulting in accurate full sequence. Although this would be relatively time consuming. A similar technique could be used with PacBio with longer reads than the Illumina and higher accuracy than the Nanopore could be a future venture. There is also the possibility of more Nanopores assemblies being made. This could produce repeats that could more conclusively decipher the full and accurate sequence of the 45S ribosome.

4.6 Species comparison

This bioinformatics yielded new knowledge of the 45S MuTR consisting of 8 repeats and confirming the length of the monomer to be 8.9kb. This information can be used in the comparison and categorisation of the species in the *Musaceae* family. This banana family has been controversial in the attempts to classify the species and genus. Largely this can be attributed to the lack of complementary nucleotide information in relation to the morphological information that is currently possessed. When trying to categorise the different genera in the banana family looking at the ribosomal DNA has been essential as it tends to evolve much slower than the rest of the genome. On the contrary, the repetitive sequences tend to be considered fast evolving components. This enables the species to be traced back and linked to each other with more accuracy and then isolate genomes with the repetitive units. For example, when differentiating between *Musella*, *Ensete* and *Musa* in concluding the relatedness of the different genera centromere repeats were used. Both *Musella* and *Ensete* contained Egcen (a centromeric) sequences, which are tandemly repeated satellites yet Musa did not contain these repeats. Suggesting that though they were all related, with all three containing long interspersed elements like Nancia, *Musella* and *Ensete* have a closer evolutionary relationship than with *Musa;* the rest of the family genus. It also provided more evidence for the divergences of these genera from *Musa* that contains the vast majority of the species within the family of *Musaceae.* Consequently, this indicates the understanding of tandem repeats is essential to the specification of these species and genera.

Another facet of the 45S organisation that can be used in the differentiation of species is the overall monomer length. Other publications such as Z.-F. Wang *et al.* (2023) that looked at the genome assembly of *Musa beccarii* uncovered the length of the consensus sequence of the 45S monomer to be 10.4kbs. This is important data that can be used in evaluations between the species of the *Musaceae* family. In contrast to that of the *Musella* with the MuTR length being 8.9kb indicating a relatively considerable difference in length. Alongside this there is also variance in the GC richness of the two sequences, however, this is much less significant with *Musa beccarii* containing 60% GC richness whereas MLA holds 56.7%. One of the spices and genera that is concidered to be closer in relation to that of MLA is *Ensete glaucum*. This organism has also been of interest in recent studies such as (Wang *et al.*, 2022) stating the monomer length of the 45S rDNA to be 9.9kb. This length is comparatively closer to the organisation of the 45S rDNA monomer of MLA. Similarly, when looking at MLA and identifying the pattern of 8 repeats compared to that of *Ensete glaucum* with 20 to 46 repeats is fundamental in defining each categorisation. Variations in the genome size of each of these species could have been Affected by the methods and equipment used in each of the various studies. However these methods may have affected, the studies there is conclusive evidence to suggest the need for further research into have the repetitive regions of the genome for phylogenetic categorisation.

4.7 'Wet' lab method advantages and disadvantages

In our wet experiments, we used the control of barley, specifically Fabiola which is a German spring Barley. It is also a monocote like MLA as well as being very well studied and understood making this a great model species. This enabled us to effectively troubleshoot and test the probes ensuring the probes were picking up the correct regions. These probes were used instead of clones to reduce cost. Clones are large and cost comparatively more than a probe which is already directly labelled. Saving time and money in the process allowing for a more achievable timeline. Other model species were considered such as wheat, however, this has far too many chromosomes. There were some limitations to our wet lab work. While we were able to relatively easily locate and obtain the barley control chromosomes this was not replicated in the sampling of MLA. In comparison, the roots were far more delicate and resulted in the genetic material being lost, when removing the outer root material, in the collection process (seen in 2.0 methods and materials 2. Chromosome preparation). This difficulty was further amplified when applying pressure to the slide to squash the cells exposing the chromosomes. This was difficult in both the barley and MLA samples as with too much acetic acid the material would again be lost when dispersed by the pressure. This could have resulted in the lack of expected 5S and 45S sites in the control barley. The dispersal method can cause chromosomes to be washed away as well as many cells chromosomes being mixed together. Far more repetition of the FISH experiments need to be under taken to get more accurate photos of full chromosome view. Even with the copious amount of repeat that were undertaken a very limited number of slides amounted to viable results. The number of chromosomes expected for the MLA was 2n=18 (Šimoníková *et al*., 2022) which was also recorded in the photographs that were obtained during this project. However, only two 5S sites were seen in the MLA slide and 4 sites of 45S, due to the inaccuracy of the controls this data would need more repeats to assure the accuracy. With more time and practice the method could be perfected for more consistent slide replications.

4.8 Conclusion

This study has unveiled novel data that can now be used in the full genome analysis of an incredibly important species of *Musaceae* family. The novel Nanopore and system was successfully used in the bioinformatics portion, providing the sequences used for the final analysis. The information garnered from these sequences can be used in the construction of the Pangenome of the *Musaceae* family and in future phylogenetic categorisation. Though more work will have to be done to grasp the full importance of the information the knowledge of the 45S rDNA length, insertion length, MuTR tandem repeat number and lack of higher order the study has yielded new and important data.

5.0 Acknowledgments

I would like to thank both Professor Heslop-Harrison and Dr Schwarzacher for their constant support and guidance throughout my project.

6.0 References

Bharagava, R.N., Purchase, D., Saxena, G. and Mulla, S.I., 2019. Applications of metagenomics in microbial bioremediation of pollutants: from genomics to environmental cleanup. In *Microbial diversity in the genomic era* (pp. 459-477). Academic Press.

D'hont, A., Denoeud, F., Aury, J.M., Baurens, F.C., Carreel, F., Garsmeur, O., Noel, B., Bocs, S., Droc, G., Rouard, M. and Da Silva, C., 2012. The banana (Musa acuminata) genome and the evolution of monocotyledonous plants. *Nature*, *488*(7410), pp.213-217.

Droc, G., Martin, G., Guignon, V., Summo, M., Sempéré, G., Durant, E., Soriano, A., Baurens, F.C., Cenci, A., Breton, C. and Shah, T., 2022. The banana genome hub: a community database for genomics in the Musaceae. *Horticulture Research*, *9*.

Elder Jr, J.F. and Turner, B.J., 1995. Concerted evolution of repetitive DNA sequences in eukaryotes. *The Quarterly Review of Biology*, *70*(3), pp.297-320.

Hřibová, E., Čížková, J., Christelová, P., Taudien, S., de Langhe, E. and Doležel, J., 2011. The ITS1-5.8 S-ITS2 sequence region in the Musaceae: structure, diversity and use in molecular phylogeny. *Plos one*, *6*(3), pp.17863.

Janssens, S.B., Vandelook, F., De Langhe, E., Verstraete, B., Smets, E., Vandenhouwe, I. and Swennen, R., 2016. Evolutionary dynamics and biogeography of M usaceae reveal a correlation between the diversification of the banana family and the geological and climatic history of Southeast Asia. *New Phytologist*, *210*(4), pp.1453-1465.

Josling, T.E. and Taylor, T.G. eds., 2003. *Banana wars: the anatomy of a trade dispute*. CABI.

Leitch, I.J. and Heslop-Harrison, J.S., 1992. Physical mapping of the 18S–5.8 S–26S rRNA genes in barley by in situ hybridization. *Genome*, *35*(6), pp.1013-1018.

Leitch, I.J. and Heslop-Harrison, J.S., 1993. Physical mapping of four sites of 5S rDNA sequences and one site of the α-amylase-2 gene in barley (Hordeum vulgare). *Genome*, *36*(3), pp.517-523.

Lin, B., Hui, J. and Mao, H., 2021. Nanopore technology and its applications in gene sequencing. *Biosensors*, *11*(7), p.214.

Liu, A.Z., Kress, W.J. and Long, C.L., 2003. The ethnobotany of Musella lasiocarpa (Musaceae), an endemic plant of southwest China. *Economic botany*, *57*(2), pp.279-281.

Martin, G., Baurens, F.C., Droc, G., Rouard, M., Cenci, A., Kilian, A., Hastie, A., Doležel, J., Aury, J.M., Alberti, A. and Carreel, F., 2016. Improvement of the banana "Musa acuminata" reference sequence using NGS data and semi-automated bioinformatics methods. *BMC genomics*, *17*(1), pp.1-12.

Michael, T.P. and VanBuren, R., 2020. Building near-complete plant genomes. *Current Opinion in Plant Biology*, *54*, pp.26-33.

Rijzaani, H., Bayer, P.E., Rouard, M., Doležel, J., Batley, J. and Edwards, D., 2022. The pangenome of banana highlights differences between genera and genomes. *The Plant Genome*, *15*(1).

Robinson, J.C. and Saúco, V.G., 2010. *Bananas and plantains* (Vol. 19). Cabi.

Springer-Verlag Berlin Heidelberg, 2005. Published online: 28 July 2005 Crosslinking of Vinylidene Fluoride-Containing Fluoropolymers. *Crosslinking in Materials Science: Technical Applications*, *184*, pp.127-211.

Wang, J., Yang, W., Zhang, S., Hu, H., Yuan, Y., Dong, J., Chen, L., Ma, Y., Yang, T., Zhou, L. and Chen, J., 2023. A pangenome analysis pipeline provides insights into functional gene identification in rice. *Genome Biology*, *24*(1), pp.1-22.

Wang, Z., Rouard, M., Biswas, M.K., Droc, G., Cui, D., Roux, N., Baurens, F.C., Ge, X.J., Schwarzacher, T., Heslop-Harrison, P. and Liu, Q., 2022. A chromosome-level reference genome of Ensete glaucum gives insight into diversity and chromosomal and repetitive sequence evolution in the Musaceae. *GigaScience*, *11*.

Wang, Z.F., Rouard, M., Droc, G., Heslop-Harrison, P. and Ge, X.J., 2023. Genome assembly of Musa beccarii shows extensive chromosomal rearrangements and genome expansion during evolution of Musaceae genomes. *GigaScience*, *12*.