GENOME-WIDE CHARACTERIZATION OF SMALL RNA, GENE EXPRESSION AND DNA METHYLATION CHANGES IN RESPONSE TO SALT STRESS IN *Musa acuminata*

GUDIMELLA RANGANATH

FACULTY OF SCIENCE
UNIVERSITY OF MALAYA
KUALA LUMPUR

2018

GENOME-WIDE CHARACTERIZATION OF SMALL RNA, GENE EXPRESSION AND DNA METHYLATION CHANGES IN RESPONSE TO SALT STRESS IN *Musa acuminata*

GUDIMELLA RANGANATH

THESIS SUBMITTED IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

INSTITUTE OF BIOLOGICAL SCIENCES
FACULTY OF SCIENCE
UNIVERSITY OF MALAYA
KUALA LUMPUR

2018

UNIVERSITY OF MALAYA
ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: GUDIMELLA RANGANATH

Matric No: SHC100039

Name of Degree: DOCTOR OF PHILOSOPHY

Title of Project Paper/Research Report/Dissertation/Thesis ("this Work"):

GENOME-WIDE CHARACTERIZATION OF SMALL RNA, GENE

EXPRESSION AND DNA METHYLATION CHANGES IN RESPONSE TO

SALT STRESS IN *Musa acuminata*

Field of Study: BIOTECHNOLOGY

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;
(2) This Work is original;
(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
(4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
(5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
(6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature   Date:

Subscribed and solemnly declared before,

Witness's Signature   Date:

Name:

Designation:

GENOME-WIDE CHARACTERIZATION OF SMALL RNA, GENE EXPRESSION AND DNA METHYLATION CHANGES IN RESPONSE TO SALT STRESS IN *Musa acuminata*

ABSTRACT

Banana, a commercially important crop which serves as a staple food in several countries worldwide, faces threats from abiotic stress especially related to soil and water salinity due to climate change. Most banana cultivars are salt sensitive, which results in low productivity and fruit of low quality. Physiological responses to salt stress are regulated by underlying gene expression which is influenced by microRNA, small interfering RNA and methylations of genic regions. This study integrated data from transcriptomes, small RNA transcriptomes, degradomes and methylomes using high-throughput sequencing of RNA and DNA extracted from the roots of salt-stressed and non-salt-stressed banana plantlets. Various bioinformatics approaches were adopted for analysis of multi-omics data, for miRNA prediction using small RNA transcriptome a customized pipeline was designed using miRDeep2, miRNA target validation using degradomes was performed by cleaveland4 tool, methylomes were analysed using Bismark and MethPipe tools. Data integration for small RNA and degradome data was performed using network mapping by cytoscape tool. Similarly, data integration for small RNA, transcriptome and methylomes was performed by using statistical approach by custom scripts and visualized data using genome browser. Genome-wide microRNAs were annotated using small RNA transcriptome data and the most recent banana genome sequence. A total of 180 mature miRNAs belonging to 20 orthologous miRNA families and 39 *Musa*-specific miRNA families were identified. Candidate microRNA targets genes were predicted using bioinformatics tools and validated using degradome data. Profiling of transcription factor binding sites (TFBS) motifs across miRNA promoter regions showed that transcription factors belonging to TCP, AP2; ERF, GATA, NF-YB, DOF, B3, bZIP, trihelix, ZF-HD,

bHLH and Dehydrin are likely abundant in the *Musa acuminata* genome. A putative miRNA-mediated regulatory network is proposed for miR156, miR164, miR166, miR171, miR319 miR396, miR528, mac-miR6, mac-miR-new14 and mac-miR-new20 and their respective transcription factor targets. Genome-wide association between DNA methylation, expression of genes and of 21nt and 24nt small RNAs in response to salt stress was determined using methylome, transcriptome and small RNA transcriptome libraries. DNA methylation in genic regions showed transcriptional repression in several stress-responsive gene candidates such as DRE2, DHN1, AP2, ion-transport related genes, i.e. calcium permeable stress-gated cation channel 1-like and cation/H+ antiporter 20-like, and peroxidases (PER1, PER67 and PNC1), which are ROS-related antioxidants during salt stress. Salt-stressed root samples displayed symmetric CG methylation and CHH demethylation adjacent to differentially expressed genes, while 21 and 24nt siRNA clusters on genomic loci showed increased methylation levels in CG, CHG and CHH contexts. This research contributes *Musa-* specific miRNA"ome" and small RNA-targeted differentially methylated genic regions which serve as molecular and epigenetic markers to support improvement of banana to address cultivation in salinized soil. *Musa-* specific genomic markers will serve as an important knowledge base for crop improvement and plant breeding programs.

Keywords: banana, salt stress, methylation, microRNA, transcription factors

PENCIRIAN LUAS GEN RNA KECIL, EKSPRESI GEN DAN PERUBAHAN METILASI DNA SEBAGAI TINDAK BALAS KEPADA TEKANAN GARAM DALAM *Musa acuminata*

ABSTRAK

Pisang merupakan tanaman komersil penting yang berfungsi sebagai makanan ruji di beberapa negara di seluruh dunia, menghadapi ancaman dari tekanan abiotik terutamanya yang berkaitan dengan salinitas tanah dan air akibat perubahan iklim. Kebanyakan kultivar pisang adalah peka-garam yang mengakibatkan produktiviti rendah dan penghasilan buah-buahan berkualiti rendah. Tindak balas fisiologi terhadap tekanan garam dikawal oleh ekspresi gen asas yang dipengaruhi oleh mikroRNA, gangguan RNA kecil dan metilasi kawasan genetik. Kajian ini menyatukan data daripada transkriptom, transkriptom RNA kecil, degradom dan metilom dengan menggunakan celusan tinggi penjujukan RNA dan DNA yang diekstrak dari akar tumbuhan pisang yang adanya tekanan garam dan tanpa tekanan garam. Pelbagai pendekatan bioinformatik telah digunakan untuk menganalisis data multi-omiks, untuk ramalan MiRNA dengan menggunakan transkrip RNA kecil saluran paip tersuai yang direka menggunakan miRDeep2, pengesahan sasaran miRNA menggunakan degradom dilakukan oleh alat cleaveland4, metilom dianalisis dengan menggunakan alat Bismark dan MethPipe. Integrasi data untuk RNA dan data degradom kecil dilakukan dengan menggunakan pemetaan rangkaian oleh alat sitoskap. Begitu juga dengan penyepaduan data untuk RNA kecil, transkriptom dan metilom dilakukan dengan menggunakan pendekatan statistik oleh skrip adat dan data divisualisasi dengan menggunakan penanda genom . MikroRNAs genom-luas telah diberi penjelasan dengan menggunakan data transkrip RNA kecil dan urutan genom pisang yang paling terkini. Sejumlah 180 miRNA matang yang terdiri daripada 20 keluarga miRNA *orthologous* dan 39 keluarga miRNA *Musa*-spesifik telah dikenal pasti. Calon gen sasaran mikroRNA dijangka menggunakan alat bioinformatik

dan disahkan menggunakan data degradom. Penyusuk tapak ikatan faktor transkripsi (TFBS) motif di seluruh kawasan promoter miRNA menunjukkan bahawa faktor-faktor transkripsi milik TCP, AP2; ERF, GATA, NF-YB, DOF, B3, bZIP, trihelix, ZF-HD, bHLH dan Dehydrin mungkin banyak dalam genom *Musa acuminata*. Rangkaian pengawalseliaan mediasi miRNA yang disangka telah dicadangkan untuk miR156, miR164, miR166, miR171, miR319 miR396, miR528, mac-miR-new14 dan mac-miR-new20 serta sasaran-sasaran faktor transkripsi masing-masing. Pertalian genom-luas antara metilasi DNA, ekspresi gen serta 21nt dan 24nt RNA kecil sebagai tindak balas terhadap tekanan garam ditentukan dengan menggunakan metilom, transkriptom dan juga pustaka transkriptom RNA yang kecil. Metilasi DNA di kawasan genetik menunjukkan bahawa penindasan transkrip dalam beberapa calon gen yang responsif di bawah tekanan seperti DRE2, DHN1, AP2, gen yang berkaitan dengan ion pengangkutan, iaitu kalsium kation yang telap tahan tekanan saluran "1-*like*" dan kation/H + antiporter "20-*like*" serta peroksidas (PER1, PER67 dan PNC1) yang merupakan antioksidan berkaitan dengan ROS semasa wujudnya tekanan garam. Sampel akar garam yang ditekankan menunjukkan simetrik metilasi CG dan demetilasi CHH bersebelahan dengan gen yang berbeza-beza, sementara kelompok-kelompok 21 dan 24nt siRNA pada loci genomik menunjukkan peningkatan tahap metilasi dalam konteks CG, CHG dan CHH. Penyelidikan ini memberi sumbangan, iaitu *Musa*-khusus miRNA "ome" dan RNA kecil yang disasarkan di kawasan-kawasan genetik bermetil yang berfungsi sebagai penanda molekul dan epigenetik untuk menyokong peningkatan pisang untuk penanaman di tanah bergaram. Penanda genomik *Musa*-khusus akan menjadi asas pengetahuan penting dalam penambahbaikan tanaman dan program pembiakbakaan tumbuhan.

Kata kunci: pisang, tekanan garam, metilasi, mikroRNA, faktor transkripsi

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude towards Professor Dr. Jennifer Ann Harikrishna for believing in my abilities and giving this opportunity to pursue doctorate study under her guidance. I would like to thank her for encouragement and immense support throughout the period of study. I would also like thank my co-supervisor Professor Dr. Norzulaani Khalid for her guidance on plant tissue culture and always encouraging the idea behind my research. I would like to acknowledge Ministry of Education (MOE), Malaysia and University of Malaya for the research grant, sponsorship and funding for my doctorate study. I'm very thankful to Dr. Martti Tammi for his mentorship for my PhD which helped me develop critical skills in writing and implementing research. I want to acknowledge Dr. Lee Wansin, Dr. Purabi Mazumdar and Dr. Pooja Singh for their help in performing laboratory analysis and mentoring me with wet lab knowledge for my study. I would like to thank my colleagues from BGM and CEBAR, Hui Li, Su Ee, Tyson, Gwo Ron Wong for all their moral support and friendship. I'm immensely thankful to Dr. Arif Anwar and all my colleagues from Sengenics for their constant support and understanding throughout my PhD. I thank all my friends in Malaysia for being moral support and optimistic towards my determination to pursue my study with several constraints. With heartful of gratitude, I want to acknowledge my Mom, Dad, Wife, Brother and Sister for being my strength, confidence and hope in every situation, so that I can achieve my passion and dream of pursuing my PhD. I would like to thank my brother Mr. Muralinath for being a very big support to achieve my research goals. I would like to express my love and gratitude towards my wife Deepika who has made my achievements into her goals, improving my self-determination the time I needed the most. I want to dedicate my research to MY DAD, Mr. Gopinath as his lifetime achievement.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS AND ABBREVIATIONS

%	:	Percentage
A genome	:	*Musa acuminata* genome
AP2	:	Apetala 2
ARF	:	Auxin response factor
B genome	:	*Musa balbisiana* genome
bHLH	:	Basic helix-loop-helix
bp	:	Base pairs
BS-seq	:	Bisulphite sequencing
bZIP	:	Basic leucine zipper
cDNA	:	Complementary DNA
CG	:	CpG sites
CHG/CHH	:	H corresponds to A, T or C
CRISPR	:	Clustered Regularly Interspaced Short Palindromic Repeats
CTAB	:	Cetyltrimethylammonium bromids
CTR	:	control
DCL	:	Dicer like
DMR	:	Differentially methylated region
DNA	:	Deoxyribonucleic acid
dS m^{-1}	:	deciSiemens per meter
dsRNA	:	double-stranded Ribonucleic acid
DSS	:	Dispersion shrinkage for sequencing data
ERF	:	Ethylene response factor
FAO	:	Food and Agriculture Organization
GRF	:	Growth response factor

Hc-siRNA	:	Heterochromatin siRNA
HOX	:	Homeobox
HSFB	:	Heat stress transcription factor B
LINE	:	Long interspersed nuclear elements
LTR	:	Long terminal repeats
miRNA	:	MicroRNA
miRNA*	:	Star strand of mature miRNA
ml	:	microlitre
Mya	:	Million years ago
NaCl	:	Sodium chloride
PKW	:	Pisang Klutuk Wulung
PMRD	:	Plant MicroRNA Database
Pre-miRNA	:	Precursor miRNA
Pri-miRNA	:	Primary miRNA
PTGS	:	Post transcriptional gene silencing
RdDM	:	RNA directed DNA methylation
RdRp	:	RNA dependent RNA polymerase
RNA	:	Ribonucleic acid
RNA-seq	:	Transcriptome sequencing
rRNA	:	Ribosomal RNA
SCL	:	Scarecrow like protein
siRNA	:	Small interfering RNA
SPL	:	SQUAMOSA promoter-binding protein like
sRNA	:	Small RNA
sRNA-seq	:	Small RNA sequencing
Ta-siRNA	:	Transacting siRNA

TF	:	Transcription factors
TFBS	:	Transcription factor binding sites
TGS	:	Transcriptional gene silencing
TPM	:	Transcripts per million
TR100	:	100 mM NaCl treatment
TR300	:	300 mM Nacl treatment
TSS	:	Transcription start sites
ZF-HD	:	Zinc finger homeodomain

LIST OF APPENDICES

CHAPTER 1: INTRODUCTION

Banana is fourth most important crop after rice, wheat and maize in terms of its importance as a source of staple starch crop (Perrier et al., 2011). It is considered as an iconic fruit with numerous health benefits and more than 85% of produced banana within a country are consumed locally (Sharrock & Frison, 1998). Banana (*Musa* spp.) are giant perennial monocotyledonous herbs of the order Zingerberales, a sub group of the widely-studied poales, which also includes staple food crops like rice. Most of the commercial cultivars of banana are triploid (*2n = 3x =33*) and are sterile with fruit development by parthenocarpy. There are hybrid varieties between two diploids (2n = 2x = 22) species *Musa acuminata* and *Musa balbisiana* with A and B genomes, respectively (D'Hont et al., 2000).

Bananas originated in India, China and South-east Asia regions, where wild varieties of *M. acuminata* (AA genome) and *M. balbisiana* (BB genome) are found (Simmonds, 1962). The centre of diversity of banana has been reported as Malaysia or Indonesia (Daniells, 2001) and bananas are distributed across tropical rainforests in these countries. Studies on banana domestication based on nuclear and cytoplasmic markers showed that *M. acuminata* subspecies *malaccensis* is widely spread across the Malay peninsula (Perrier et al., 2011). About 50% of the banana growing area in Malaysia is cultivated by popular commercial cultivars *i.e. Pisang Berangan* and Cavendish types (AAA genome) with a total harvesting area of around 29,000 ha (Mokhtarud-din & William, 2011). In Malaysia, bananas are considered second in terms of production and fourth in terms of export revenue from fruits (Kayat et al., 2016). Banana production in Malaysia has declined around 40% since 2004 (FAOSTAT, 2015), which may be due to the spread of Panama (Fusarium wilt) and Moko (Bacterial wilt) diseases (Mokhtarud-din & William, 2011).

Bananas usually have a shallow root system and permanent green canopy which requires an abundant supply of water for fruit yield and production (Turner, 2007; Van Asten et al., 2011). Many biotic and abiotic factors use roots as the entry point to the plant and affect banana plantations and fruit production. Biotic factors such as soil-borne pathogens and pests, including biotic stress factors such as soil moisture stress, water stress, salinity stress and dehydration (Reviewed in Ravi and Vaganan (2016). Due to depleting water and soil conditions worldwide (FAO, 2017), the high water-loving crop like banana will also reduce in yield (Wairegi et al., 2010). Most of the commercial cultivars of banana belong to *M. acuminata* (AA) genotype which are sensitive to abiotic stress, while *M. balbisiana* (BB) genotypes are considered to be more resistant to abiotic stress (Vanhove et al., 2012) possibly, due to its domestication in extreme climatic conditions which has influenced its genetic structure. The above factors increase the need to study genetic and molecular level changes caused by abiotic stress in banana cultivars.

A banana genome sequencing project was undertaken by the Global *Musa* Genomics Consortium in 2012, which published genome of the DH-Pahang (A genome) (doubled-haploid Cavendish) cultivar consisting of 472.2Mb in length with 11 chromosomes annotated with 36,542 protein coding gene models and 235 microRNA families (D'Hont et al., 2012). Later in 2013, a collaboration of scientists from the University of Malaya with the University of Leuven, Belgium, led to publication of a genome sequence for *M. balbisiana* (B genome) variety Pisang Klutuk Wulung (PKW) (Davey et al., 2013). Banana genome data can improve the analysis of the transcriptional and post transcriptional changes which influence gene expression triggered by abiotic stress. Transcriptional and post-transcriptional gene regulation include microRNA (miRNA) - based gene silencing, transcription factors (TFs)- mediated gene regulation and small interfering RNA (siRNA) based DNA methylation. Elucidating genome wide miRNA

regulatory networks and sites of siRNA-based *de novo* methylation associated with abiotic stress exposure in banana is the main aim of this thesis.

The primary objectives of this thesis research were:

1.   to compare genome wide microRNA (miRNA) sequences within the banana A genome (*M. acuminata)* and B genome (*M. balbisiana)* using high-throughput sequencing small RNA datasets.

2.   to predict targets of banana miRNA towards elucidating the role of miRNA and miRNA-target genes in salinity-stressed banana roots based on analysis of high-throughput sequencing small RNA and degradome datasets.

3.   to determine association of transcription factors binding sites (TFBS) on salt stress-responsive miRNA promoter regions and miRNA target transcription factors in banana.

4.   to determine association of DNA methylation with expression of genes and siRNA in salinity-stressed banana roots.

CHAPTER 2: LITERATURE REVIEW

2.1   Bananas

2.1.1   Bananas and plantains

Bananas and plantains belong to the order of Zingiberales and the family of Musaceae (Simmonds, 1962). The two genera in this family are *Musa* and *Ensete*. The genus *Musa* is divided into five main series on the basis of chromosome numbers, orientation and arrangement of flowers in the inflorescence. The five series are Musa ($X = 11$), Rhodochlamy ($X = 11$), Callimusa ($X = 10$ or 9), Australiamusa ($X = 10$) and Ingentimusa ($X = 14$) (Heslop-Harrison & Schwarzacher, 2007; Simmonds & Weatherup, 1990).

The modern method of classifying edible bananas was devised by Simmonds and Shepherd (1955). Most modern edible bananas originally came from two wild, seeded species, the Malaysian origin *Musa acuminata* Colla (A genome) and the Indochina origin *Musa balbisiana* Colla (B genome) (Perrier et al., 2011; Simmonds & Shepherd, 1955). However, a few other cultivars may have arisen from hybridization with *Musa schizocarpa* (S genome) and at least one Philippine clone may have come from ancient hybridization between *Musa balbisiana* and *Musa textilis* (T genome). Interspecific hybridization between *Musa acuminata* $\times$ *Musa balbisiana* produced polyploidy clones with different combinations of A and B genomes (Saraswathi et al., 2011; Vanhove et al., 2012). The establishment of these hybrid clones would have occurred in prehistoric times, and the earliest records of cultivation are from India about 2500 years ago (Heslop-Harrison & Schwarzacher, 2007). These hybrids conferred a measure of hardiness and drought tolerance as a result of the introduction of genes from species adapted to such conditions (Heslop-Harrison & Schwarzacher, 2007). Furthermore, the *M. balbisiana* genes induced greater disease resistance, improved nutritional value, increased

starchiness and provided hybrids suitable for cooking in comparison to *M. acuminata* genes (Robinson & Saúco, 2010).

2.1.2 Abiotic stress tolerance in banana

Abiotic stresses are caused by non-living factors including light (high light, UV and darkness), water (deficit and flooding), salt, temperature (frost, low and heat), nutrient imbalance, oxidation stress, hypoxia and physical factors (wind). Tolerance to such stress depends on the developmental stage and cultivar of the plant. Plants adopt stress resistance mechanisms such as avoidance (prevents stress exposure), tolerance (withstand stress condition) and acclimation (alteration of physiological responses).

Banana crops naturally grow in habitats such as warm and hot climates and only survive within a limited range of temperatures. Banana cultivars are restricted to sub-tropical and tropical areas between 30° north and 30° south with mean temperatures of 27°C, while root growth occurs between 22-25°C and lower temperatures will slow down the growth. Optimal banana growth conditions are at least 25 mm of water per week and an annual average rainfall of 2000–2500 mm throughout the year (Vanhove et al., 2012). Adequate water supply and sufficient nutrients during the early and late vegetative phases are crucial and determine the growth and yield of banana plants (Turner, 2007). Lack of sufficient irrigation practices such as low-quality water, limited water supply, long dry seasons and extreme temperatures will hinder banana growth and expansion of banana cultivation. Major abiotic stress factors effecting banana crop are drought, soil moisture deficit, salt and temperature stress which seem to be to overlooked by current studies/technologies which are available for increasing banana production and plantation (Wairegi et al., 2010). Hence, there is a research gap and need of understanding tolerance levels of banana towards abiotic stresses in present changing climatic conditions which greatly effects productivity of economically important crop like banana.

2.1.2.1 Salt stress tolerance in banana

Salt stress is one of the major abiotic stress factors effecting banana productivity (Ravi & Vaganan, 2016). Salt stress is estimated to affect 20% of total cultivated land worldwide and 33% of irrigated land (Shrivastava & Kumar, 2015). Salinity-related problems arise due to dry climates, saline soils and low-quality irrigation water. Plants can survive concentrations of salinity of up to around 4 desiSemens per meter (dS/m), (~40 mmol) NaCl, but most plants show stress symptoms even with lower levels of salinity, which leads to reduction of the yield (Gao et al., 2007). In bananas, high salt concentrations i.e. greater than 4 dS/m will promote fast deterioration of the banana root system (Gauggel et al., 2005). Salt stress effects in banana appear in leaf margins showing necrosis (Shapira et al., 2009), reduces pseudo stem thickness and also causes delay in flowering (sometimes by more than 2-3 months) (Ravi & Vaganan, 2016). Salinity causes dehydration and osmotic stress which influence fruit physical parameters including fruit length, circumference, fruit pulp, peel weight, volume and density which are important parameters to determine quality and price of banana (Mahouachi, 2007; Ravi & Vaganan, 2016). Banana cultivars are also shown to be salt sensitive crops that on exposure show poor plant production and reduction in the yield (Israeli et al., 1986; Yano-Melo et al., 2003).

2.1.3 Banana genomes

The banana nuclear genome is relatively small ~600Mbp and it was estimated that 55% of the genome consists of DNA repeats (D'Hont et al., 2012; Hribova et al., 2007; Hribova et al., 2010; Novak et al., 2014). The banana A-genome (*Musa acuminata* var. DH-Pahang) and B-genome (*Musa balbisiana* var. 'Pisang Klutuk Wulung') were sequenced in separate genome projects in years 2012 and 2013, respectively. The sequencing project undertaken by the Global *Musa* Genomics Consortium published the genome sequence of the *Musa acuminata* var. DH-Pahang (doubled-haploid Cavendish).

Assembled *Musa acuminata* genome length was reported as 473Mb which represents ~90% of the total estimated *Musa acuminata* genome i.e. 523 Mb. Genomic assembly also reported 11 chromosomes annotated with 36,542 protein coding gene models (D'Hont et al., 2012). A complete genome sequence for the banana B genome based on *Musa balbisiana* var. 'Pisang Klutuk Wulung' ('PKW', B-genome) (Davey et al., 2013) was assembled using the A-genome (D'Hont et al., 2012) as a reference. The B-genome was reported as 341.4 Mb length containing 36,638 predicted functional gene sequences. Recently, banana version-2 genome was reported with improved genome assembly and annotations (Martin et al., 2016).

2.2   MicroRNA

2.2.1   MicroRNA (miRNA) biogenesis

Since the first report of plant miRNA (Reinhart et al., 2002), there have been considerable advances in understanding its functional role and origin. In plants, several miRNAs are highly conserved as well as more recently evolved, suggesting link between the evolutionary conservation of plant miRNAs and the mechanisms underlying the miRNA biogenesis (Chorostecki et al., 2017; D'Ario et al., 2017). miRNAs are synthesized as primary (Pri)-miRNA transcripts of RNA polymerase II, Pri-miRNA is single-stranded polyadenylated RNA molecules which fold into hairpin-like structures. Pri-miRNA are then cleaved by the RNAse III enzyme, Dicer like 1 (DCL1) into shorter hairpin structures, known as precursor miRNA (pre-miRNA) (Figure 2.1A). Pre-miRNA are again processed by DCL1 into 20-22nt length mature miRNA duplexes consisting of a mature miRNA strand and a star miRNA strand (complementary of mature miRNA). Relatively longer miRNA (23-25 length) were first detected in Arabidopsis and rice and result from processing by another RNAse III enzyme, Dicer like 3 (DCL3) that can potentially function in transcriptional gene silencing (TGS) (Fukudome & Fukuhara, 2017).
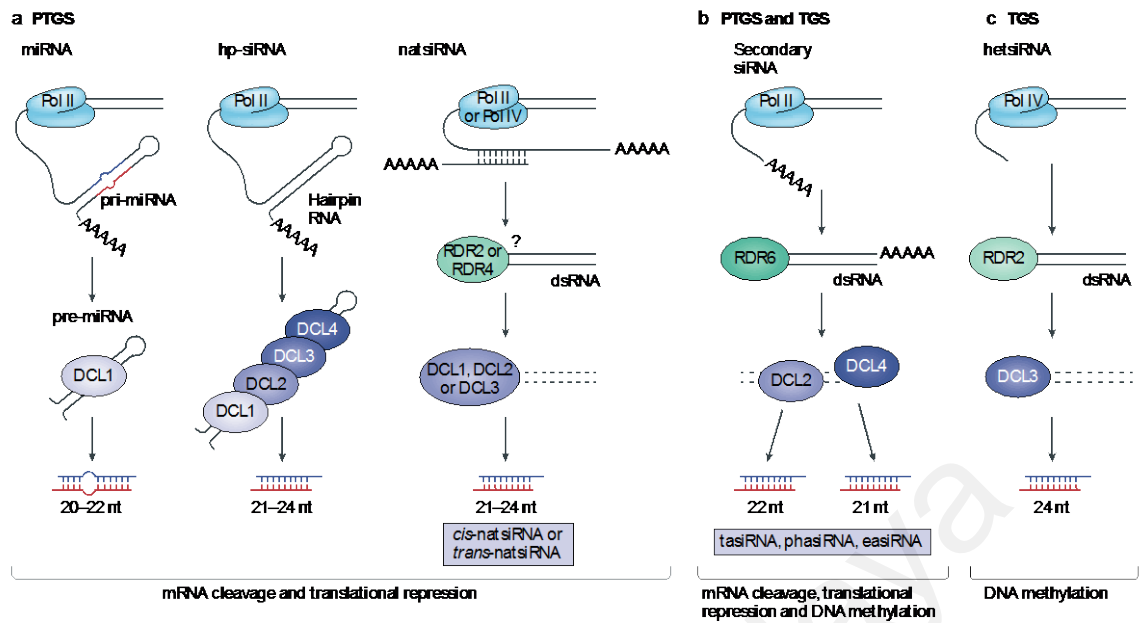
Figure 2.1: Endogenous small RNA biogenesis cascades (Borges & Martienssen, 2015). A) Post Transcriptional Gene Silencing (PTGS) by Precursor miRNA (pre-miRNA), Hairpin-siRNA (hp-siRNA), Natural antisense siRNA (nat-siRNA) B) Secondary siRNA are categorized into trans-acting siRNA(ta-siRNA), phased siRNA (phasiRNA) and epigenetically active siRNA (ea-siRNA) C) 24nt siRNA derived from pericentromeric chromatin regions are termed as heterochromatin siRNA(het-siRNA). Reprinted by permission from Springer Nature.

2.2.2 Transcription factors and their binding sites

Transcription factors (TFs) are DNA-binding proteins which bind to short DNA sequences and regulate transcription of eukaryotic genes by activating or blocking the recruitment of RNA polymerase at transcription start sites (TSS) (Weake & Workman, 2010) TFs bind sequence-specifically with cis-regulatory sequences located in promoter regions of the target genes, termed as transcription factor binding sites (TFBS). TFBS are cis-regulatory elements include transcriptional enhancers which bound to multiple TFs to activate expression of genes (Kolovos et al., 2012) but also may act as silencers of gene expression. Understanding the role of such transcriptional enhancers and silencers in plants involves exploring multiple cis-regulatory elements upstream of TSS or coding regions of genes (Weber et al., 2016). However, miRNA biogenesis also driven by such transcriptional enhancers for example, cell division cycle 5 (CDC5) transcription factor

from Arabidopsis interacts with miRNA promoters and DNA-dependent RNA polymerase II and serves as positive regulator for miRNA accumulation (Zhang et al., 2013). In contrast, miRNA is influenced by active 5' splice sites for Pri-miRNA precursors (Bielewicz et al., 2013). The role of 5' splice sites is demonstrated in miR402 in Arabidopsis, where inactivation of the 5' splice sites at close proximity to pre-miRNA of miR402-hosting intron revealed significant accumulation of mature miRNA (Knop et al., 2016). Hence, apart from TFBS, active 5' splice sites and polyadenylation sites can also influence miRNA biogenesis.

TFs are classified into different families based on the structure of their DNA-binding domains (Gonzalez, 2016). In plants, transcription factors form signalling cascades that govern developmental processes and environmental stress responses by regulating gene expression levels. Transcriptional regulation may play a more important role in plants than animals, given the large number of transcription factors in plant genomes which range from 6% to 10% of the total number of genes (Riechmann et al., 2000). The banana genome was reported to have the highest number of putative TFs (3,155 predicted TF genes) of all sequenced plant genomes (D'Hont et al., 2012). Genome-wide transcriptional regulatory code can determine networks formed from different transcriptional elements contribute to global gene expression (Harbison et al., 2004). Yeast one-hybrid (Y1H) system enables detecting in vivo regulatory interactions between TFs and DNA binding sites (Reece-Hoyes & Marian Walhout, 2012). For example, using Y1H, TF-miRNA promoter interactions between eight miRNA promoters and 15 TFs were predicted in Arabidopsis roots (Brady et al., 2011). Other high throughput sequencing method, i.e. Chromatin immunoprecipitation (ChiP-seq) also allows genome-wide *de novo* discovery of TFBS and in vivo interactions with TFs (Kaufmann et al., 2010). High-throughput in vitro techniques such as protein binding microarray (PBMs) yielded genome wide TFBS in Arabidopsis which showed functional relevance between TFBS and target TF

(Weirauch et al., 2014). Such co-regulation activity of TFs was also observed with miRNA genes that establish a regulatory feedback loop where miRNA is involved in controlling another component (either TF or non-TF protein coding genes) forming small genetic circuits.

2.2.3 miRNA and transcription factor co-regulation in plants

miRNA-directed gene expression is regulated by transcription factors (TFs) which determine cellular fate specification (Guo et al., 2016; Hobert, 2004). Promoters also determine the specificity, direction and efficiency of transcription mechanisms of downstream miRNA genes in response to any biological event in plants (Chen et al., 2016). TFs are reported to bind to the cis-regulatory elements (motifs) on the pre-miRNA genes and interact with the transcription start site (TSS) thereby activating or repressing the miRNA genes (Arora et al., 2013). In plants, genome wide target prediction has shown that a majority of stress responsive miRNAs target TFs (Zhang, 2015). Banana miRNAs have been experimentally validated to regulate TFs, including the miR156d target SPL, miR166b target SRPK4, miR319m target GAMYB, miR399a target WRKY and miR4995 target F-box (Chai et al., 2015; Lee et al., 2015). The annotation of miRNA genes and prediction of the cognate miRNA targets based on computational methods has identified useful candidates for the study of gene expression regulation in plants in response to various factors.

2.2.4 miRNA mediated networks in plants

miRNA-associated regulatory networks in plant genomes indicate that miRNA which target transcription factors (TFs) may be regulated by the same or related TF, to co-regulate gene expression (Arora et al., 2013; Qiu et al., 2010). Such co-regulation of miRNAs and TFs in a biological response can establish different types of regulatory networks. According to Megraw et al. (2016), miRNA-TF containing networks include

lock-on switches (involving self-regulation of TFs and miRNA), feedback loops (involving miRNA-repressing TF and TF-inducing miRNA) and miRNA-mediated networks (involving both miRNA and TF in controlling another component which is either a TF or a non-TF protein coding gene) (Figure 2.2).
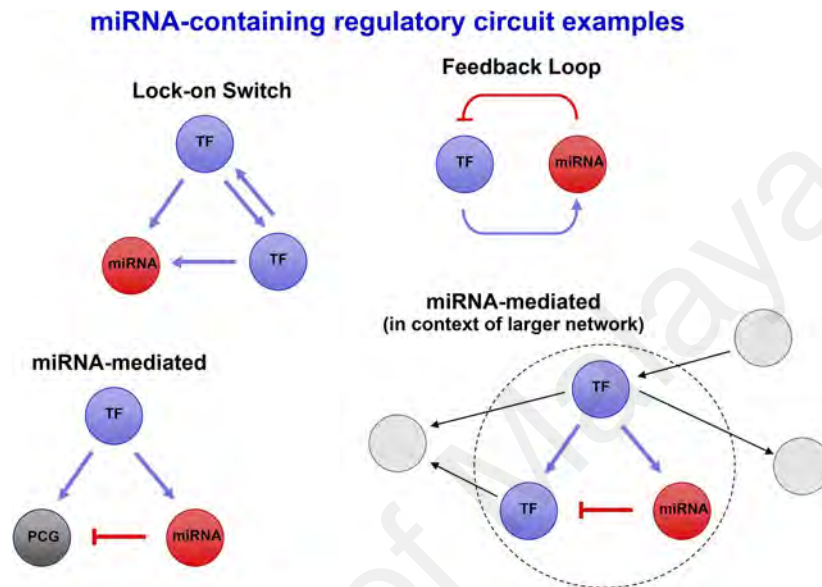
Figure 2.2: miRNA regulatory circuits (Megraw et al., 2016). Examples shown here include Lock-on Switch (involving self-regulation of TFs and miRNA), Feedback Loop (involving miRNA-repressing TF and TF-inducing miRNA), miRNA-mediated (involving both miRNA and TF in controlling another component which is either a TF or a non-TF protein coding gene). miRNA-mediated in larger contexts have role in regulatory cascades and can be part of signal processing. Reprinted by permission from American society of plant physiologists.

2.2.5 miRNA in banana

The first report of the complete sequence of the banana A genome identified 37 miRNA families which represents 235 miRNA precursors with nine conserved families (D'Hont et al., 2012). Among the eight conserved families in poales (Lee et al., 1993) miR437, miR441, miR444, miR528, miR818, miR821, miR1435 and miR2275, only the miR528 family was found in the *Musa* genome. Later, Chai et al. (2015) predicted 244 miRNA: target pairs using bioinformatics approach and validated tissue-specific

expression levels of miR156d, miR166b, miR319m, miR399a, miR4995 and miR5538 in roots, leaves, flowers, and fruits tissues.

2.3 Small Interfering RNA (siRNA)

2.3.1 siRNA biogenesis

Small interfering RNA (siRNA) are generated from exogenous RNA (such as viruses) or endogenous RNA. If single-stranded, these RNA molecules are converted into long double-stranded RNA (dsRNA) by RNA-dependent RNA polymerases (RdRp) which process into different types of siRNAs targeting specific endogenous loci (Willmann et al., 2011). dsRNAs are cleaved by Dicer like (DCLs) proteins to generate small RNA of different sizes ranging from 21 to 24nt. Similar to miRNAs, siRNAs are loaded into AGO1 of RNA-induced silencing complex (RISC) which guides post transcriptional gene regulation by specific pathway, i.e. RNA directed DNA methylation (RdDM) (Chinnusamy & Zhu, 2009; Matzke et al., 2009; Matzke et al., 2015). Based on the origin of the dsRNA, siRNA can be classified into repeat associated siRNA (ra-siRNAs) usually 24nt in size (Matzke et al., 2009), trans-acting siRNAs (ta-siRNAs) usually 21nt in size (Allen et al., 2005) and natural antisense transcript-derived siRNA (NAT-siRNAs) (Figure 2..2B). Similarly, NAT-siRNA and ta-siRNA are shown to be actively expressed during different biotic and abiotic stress conditions in plants (Khraiwesh et al., 2012; Sunkar et al., 2007). Repeat associated siRNA (24nt small RNA) guide *de novo* DNA methylation which is involved in genome stability, heterochromatin maintenance and stress-triggered pathways (Khraiwesh et al., 2012; Yao et al., 2010).

siRNA usually form duplexes which are processed from different kinds of precursors by various Dicer like enzymes i.e. DCL2, DCL3 and DCL4 (Figure 2.1). Precursors for siRNA include overlapping regions of natural antisense pair transcripts, long single-stranded hairpins from inverted repeat (IR) (Kasschau et al., 2007) or double-stranded

RNA (dsRNA) synthesized from RNA-dependent RNA polymerase (Dunoyer et al., 2010; Zhang et al., 2007) and intron regions which silence host genes (Chen et al., 2011; Meng et al., 2013). Transposon-derived 24nt siRNA are also common among plant genomes and trigger DNA methylation and chromatin modification events (Kasschau et al., 2007). It is also evident that plant anti-viral defence mechanisms generate secondary siRNA from transgene/viral RNA to promote resistance (Wang & Smith, 2016).

2.3.2   Functional Role of siRNA in plants

Small interfering RNA (siRNA) are generated either from endogenous RNA or exogenous RNA (such as viruses). siRNAs are processed by RNA-dependent RNA polymerases (RdRp) and cleaved into single-stranded siRNA of different sizes by Dicer like (DCLs). Heterochromatic siRNA has role in chromatin maintenance, DNA methylation and retro element expression (Borges & Martienssen, 2015). Major class of heterochromatic siRNA i.e. 24-nt siRNA are associated with RdDM, transcriptional gene silencing and silencing active transposable elements(Fultz et al., 2015; Matzke et al., 2015). However, studies suggest association of 21nt and 24nt siRNA in deposition of DNA methylation leading to transcriptional silencing (Nuthikattu et al., 2013). Small RNAs of 24nt size are associated with DNA methylation at thousands of sites genome-wide in Arabidopsis and are predominant in the non-CG context (CHG and CHH) methylation (Lewsey et al., 2016). Tissue specific enrichment of 24nt siRNA is shown to be associated with substantial increase in CHH methylation effecting gene expression in Arabidopsis (Erdmann et al., 2017) and *Brassica rapa* (Liu et al., 2017; Takahashi et al., 2018). In contrast, 24nt siRNA from transposable retro elements transiently decrease in abundance in causing reduction of transposon expression in callus subcultures in maize (Alejandri-Ramirez et al., 2018).

siRNA have also shown crucial role in regulating gene expression in response to abiotic stress in plants (Khraiwesh et al., 2012). In Arabidopsis, 24nt siRNA from SRO5 mRNA which targets P5CDH, leads to mRNA degradation which triggers accumulation of the osmoprotectant proline as part of salt stress toleraance (Borsani et al., 2005). In another study on Arabidopsis, 24-nt siRNAs targets 500 bp upstream region of AtMYB74 which is heavily methylated and upon promoter deletion revealed siRNA target region is necessary to maintain AtMYB74 expression patterns (Xu et al., 2015). Similarly in wheat seedlings, 21nt siRNA responsive to cold, heat, salt or drought stress were subjected to RT-PCR to identify expression changes of four siRNA (Yao et al., 2010).

2.3.3 Endogenous siRNA in banana

The banana genome that was first completely sequenced harbours copies of a banana streak virus (BSV), i.e. endogenous BSV (eBSV), a plant pararetrovirus which integrates into the host genomes (D'Hont et al., 2012). Endogenous BSV has an evolution history of integration into different banana cultivars as viral DNA which can exist as an episomal form infecting plant cells (Iskra-Caruana et al., 2014). BSV derived DNA serve as retro elements in banana genome that generates endogenous siRNA with antiviral activity (Gayral et al., 2008). Studies in banana show the prevelance of such virus derived siRNA, i.e. small RNA sequencing for complete siRNA profiles were generated from six BSV species (BSOLV, BSGFV, BSIMV, BSMYV, BSVNV and BSCAV), shown to be persistant in *Musa acuminata* triploid (AAA) banana plants (Rajeswaran et al., 2014). siRNA profiles show BSV infection induces 21nt , 22nt and 24nt viral siRNA which can be associated with AGOs to target the viral genome. Abundance of 24nt siRNA is high in these plants and covers the entire circular viral DNA genomes in sense and anti-sense strands. In contrast to siRNA abundance no cytosine methylation was observed on viral DNA , thus BSV evades silencing in banana plants by avoiding siRNA-directed DNA methylation and transcriptional silencing (Rajeswaran et al., 2014). Intronic hairpin RNA

produce diverse set of endogenous siRNA were demonstrated to have coexpressed with their host genes in rice (Chen et al., 2011). Such intron derived siRNA target vital fungal genes which could show effective resistance towards *Fusarium oxysporumf.sp.cubense* (FOC) in transgenic banana plants (Ghag et al., 2014).

2.4   DNA Methylation

Epigenetic mechanisms include DNA methylation, histone modifications and noncoding RNAs which provides plants with multlayered and robust mechanisms to fine-tune gene expression patterns (Pikaard & Mittelsten Scheid, 2014). In plants, DNA methylation occurs by addition of methyl group at C_5 position of cytosine, in CG and non-CG contexts. Non-CG methylation occurs in symmetrical and assymetrical regions in CHG and CHH contexts respectively (H= A, T or C) (Law & Jacobsen, 2010; Wassenegger et al., 1994). These modifications are often temporary and in plants a change to a normal phenotype is common, while sometimes the change may be transferred to subsequent generations by sexual propagation (Brettell & Dennis, 1991). Cytosine bases are often extensively methylated in the genome of higher plants (Gehring & Henikoff, 2007) with the level of cytosine modification ranging from 6% to 30% of the Cs in the genome (Chen & Li, 2004). In Arabidopsis, cytosine methylation occurs primarily in CG dinucleotides (24%), but CNG and CNN (where N =Adenine, Cytosine or Thymine) have also been found, occurring at the levels of 6.7% and 1.7%, respectively (Cokus et al., 2008). DNA methylation in plants is species-, tissue-, organelle-, and age-specific (Vanyushin, 2006). Genome-wide cytosine methylation and the sequencing of bisulphite-converted DNA were used to map the distribution of cytosine methylation in the entire genome of Arabidopsis (Zhang et al., 2006b; Zilberman et al., 2007). The cytosine-methylated proportion of the Arabidopsis genome is composed primarily of localized tandem or inverted repeats, transposons and dispersed repeats that are concentrated within or around centromeric regions (Zhang et al., 2010a). In plants,

epigenetics can act as memory for resetting plant processes during stress recovery which is directed by RNA metabolism, post transcriptional gene silencing and RNA directed DNA methylation (Crisp et al., 2016).

2.4.1   RNA directed DNA methylation (RdDM)

The rapid development and improvement of DNA sequencing methods has helped in the analysis of complex plant genomes as well as to determine gene expression levels, i.e. transcriptomes and small non-coding RNA sequences. The small expressed sequences include microRNA (miRNA) and small interfering RNA (siRNA), which are involved in transcriptional gene silencing (TGS) and post transcriptional gene silencing (PTGS). RdDM is a *de novo* DNA methylation pathway in plants which is largely guided by dicer independent non-coding RNAs, and that siRNA are required to maintain DNA methylation at particular loci on genome (Yang et al., 2016). RdDM is a nuclear process in which siRNAs direct the cytosine methylation of DNA sequences that are complementary to 24nt siRNAs (Chinnusamy & Zhu, 2009). Recent investigations reveal non-canonical RdDM pathway either mediated by miRNA or RDR6 (21 and 22nt primary siRNA) which provide different insights into RdDM (Cuerda-Gil & Slotkin, 2016). RdDM is a well studied small RNA-directed epigenetic pathway which is guided by 24nt small interfering RNA (siRNA) (Du et al., 2015; Matzke et al., 2015; Zhai et al., 2015).

In plants, biogenesis of 24nt siRNA is directed by RNA polymerase IV(Pol IV) and Pol V in close partnership with RNA dependent RNA polymerase (RDR2) (Wendte & Pikaard, 2017). Moreover, 24nt siRNAs tend to perfectly match with 5' end or 3' end of precursor RNAs, suggesting that individual precursors give rise to siRNAs by single DCL3 cleavage events (Blevins et al., 2015). The complementary base-pairing between AGO4 bound siRNA and scaffold RNAs produced by RNA polymerase V (Pol V) triggers domains rearranged methyltransferase (DRM2) and *de novo* methylation (Matzke

& Mosher, 2014; Matzke et al., 2015; Wierzbicki et al., 2012) (Figure 2.3). However, RdDM is not always associated with accumulation of siRNA and instead can be activated by other small RNA or long RNA (Dalakouras & Wassenegger, 2013) or any viriod derived small RNAs (vd-sRNAs) (Dalakouras et al., 2013).

DNA methyltransferases are necessary for cytosine methylation in plants and methyltransferases maintain DNA methylation. DNA methylatransferases are required to maintain different contexts of methylations i.e. symmetrical CG methylation is maintained by Methyltransferase1 (MET1) (To et al., 2011), CHG methylation is maintained by Chromomethylase3 (CMT3) (Enke et al., 2011) and asymmetrical CHH methylation is maintained by different methyltransferases DRM2 (Cao & Jacobsen, 2002) and Chromomethylase2 (CMT2) (Stroud et al., 2014). CMT2 has a main role in maintaining CHH methylations at heterochromatic regions and long transposons (Stroud et al., 2014; Zemach et al., 2013). Components of RdDM along with small RNA promotes heterochromatin formation and transcriptional gene silencing (TGS) at transposible elements (TEs) and repeats (Holoch & Moazed, 2015).
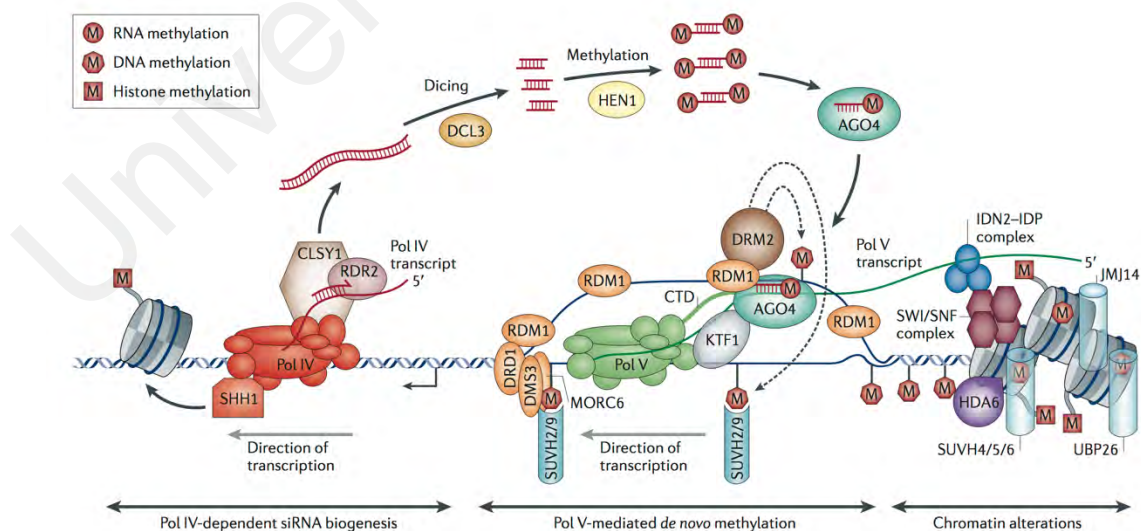
Figure 2.3: Canonical RdDM pathway mediated by Pol-IV and Pol-V (Matzke & Mosher, 2014). Reprinted by permission from Springer Nature.

17

2.4.2 Genome-wide methylation in plants

Genome-wide cytosine methylation landscapes regulate and maintain normal plant development and so investigation of the epigenome reveals the interplay between gene expression and small RNA. In Arabidopsis, observing the methylome, transcriptome and small RNA transcriptome reveal direct strand-specific DNA methylation at RNA-DNA homology and altered transcript abundance of genes and transposons upon modification of DNA methylation (Hofmann, 2012; Lister et al., 2008). A similar study in maize described the role of epigenetic marks i.e. H3K27me3 and DNA methylation in tissue specific manner in association with decreased level of 21nt miRNAs and 24nt siRNAs (Wang et al., 2009). Silencing of transposable elements (TE) mediated by small RNA and DNA methylation has been shown in wheat (Cantu et al., 2010). A study of a soybean epigenome also highlights the RdDM functionality which shows small RNA abundance was positively correlated with hypermethylated regions and a portion of hypomethylated regions were correlated with high gene expression changes among various tissues (Song et al., 2013b).

2.5 Next generation sequencing technologies and "omics"

Exploring the genetic material underlying biological processes is a basic and necessary step of biological research. The past few years' advancements in next-generation sequencing technologies (NGS) has facilitated the exploration of genetic components in high throughput and high resolution with scalability and efficiency (Table 2.1) (Esposito et al., 2016). Application of NGS along with bioinformatics software in agriculture related research has allowed genome-wide scanning of variants, binding site motifs, epigenetics, transcriptomics, marker regions and small RNA with high-resolution mapping in less time and with lower cost (Yu et al., 2017). Such technologies enable researchers to address fundamental questions about plant biology and plant sustainability.

Table 2.1: Main NGS technologies used in omics studies (Ohashi et al., 2015).

Technology	Read Length (bp)	Yield (Reads per run)	"Omics"
Roche 454	700	~700 thousand	Transcriptomics
Illumina HiSeq 2000/2500	300	~300billion	Transcriptomics, Genomics, Epigenomics
SOLiD	100	~200billion	Transcriptomics, Genomics
Ion Torrent	200	~60 billion	Transcriptomics, Genomics
PacBio RS II	14,000	~47 thousand	Transcriptomics, Genomics

2.5.1 Illumina sequencing

Illumina DNA sequencing is considered as second-generation sequencing and has proven to be effective with short and long read nucleotide sequencing. Since the first draft of the human genome, several next generation sequencing technologies for genome sequencing have been created and correspondingly the bioinformatics field has expanded to manage the large-scale data generated by these methods (Levy & Myers, 2016). The first genome analyser with sequencing by synthesis on a glass solid phase surface was reported by Fedurco et al. (2006). Sequencing by synthesis (SB) technology was commercialized by Illumina as the Genome Analyser and Hi-Seq systems. SBS Library preparation involves random fragmentation of template DNA and ligation with oligonucleotide adaptors. Amplification of DNA uses a method described as bridge PCR (Adessi et al., 2000; Fedurco et al., 2006). Each nucleotide is labelled with a chemically cleavable fluorescent reporter group at the 3′-OH end which allows a single base

incorporation in each sequencing cycle (Cao et al., 2017) which has proven to be cost competitive (Reviewed in Liu et al. (2012). The major disadvantage of PCR based sequencing methods is the possible introduction of bias in read distribution, ultimately affecting coverage. Third generation sequencing methods using single molecule read sequencing-SMRT such as from Pacific Biosciences and Oxford Nanopore Technologies (Mikheyev & Tin, 2014) have proven to be effective in avoiding amplification bias and reduce error rate in sequencing (Eid et al., 2009). However, single molecule reads sequencing generates error-prone long reads and errors are corrected by using short, high-fidelity sequences from Illumina to achieve >99.9% base-call accuracy leading to better assemblies than other sequencing strategies (Koren et al., 2012).

2.5.2   Transcriptome (RNA-seq) sequencing

Regulation of RNA transcription and processing directly affects protein synthesis and mediate cellular functions. Sequencing RNA provides the abundance and sequence of the RNA transcripts. Illumina based RNA-seq methods are based on the use of random hexamer priming to reverse transcribe poly(A)-selected mRNA (Figure S1A) (Illumina, 2017a). However, this method might introduce primer bias, which influence the uniformity of the location of reads along expressed transcripts (Hansen et al., 2010). Such non-uniform read distribution are taken into account before determining transcript abundance in algorithms such as Cufflinks (Trapnell et al., 2010). RNA sequencing can involve single-end (SE) or paired-end (PE) reads, longer the PE reads improves mappability to genome or perform *de novo* transcriptome assembly to facilitates quantification RNA expression among the datasets (Garber et al., 2011). RNA-seq has many applications such as alternative splicing, fusion transcripts and small RNA expression (miRNA and siRNA). With good experimental design and by understanding technical variability of RNA-seq data, several bioinformatics data analysis approaches are available to obtain biologically meaningful results (Conesa et al., 2016).

2.5.3   Degradome (PARE-seq) sequencing

Deep sequencing of 5' ends of polyadenylated products of miRNA-mediated mRNA decay resulted in identifying several novel-miRNA-target RNA pairs in Arabidopsis (German et al., 2008). Such sequencing of RNA degraded products can be achieved by Parallel analysis of RNA Ends (PARE) sequencing. PARE-seq is performed by ligating 5' adapters containing an Mme I restriction site to degraded uncapped mRNA and these mRNA are reverse-transcribed (Figure S1B). Resulting cDNA fragments are digested with Mme I, purified, ligated to 3' adapters, and PCR-amplified. Sequencing of PCR-amplified CDNA provides sequences of transcripts that undergo degradation. However, PCR amplification might lead to biases and errors caused by polymerase will result in incorrect sequences (Illumina, 2017a).

2.5.4   Bisulphite (BS-seq) sequencing

Genome-wide DNA methylation profiling can be possible at single-nucleotide resolution by sequencing. A common procedure of DNA methylation profiling involves fragmentation of genomic DNA by restriction enzyme digestion and bisulphite conversion. The current study utilises reduced-representation bisulphite sequencing (RRBS-seq), which uses one or multiple restriction enzymes on the genomic DNA to produce sequence-specific fragmentation (Meissner et al., 2005). The fragmented genomic DNA is treated with bisulphite and sequenced. RRBS provides genome-wide coverage of CpGs at single base resolution and covers CG methylation in dense regions of genome such as promoters and repeat regions (Figure S1C) (Illumina, 2017b). Some disadvantages of RRBS are that restriction enzymes cut at specific sites providing biased sequence selection and a lack of coverage at intergenic and distal regulatory elements (Illumina, 2017b; Yong et al., 2016).

2.6 MicroRNA (miRNA) prediction

2.6.1 Bioinformatics prediction of miRNA in plants

miRNA was first discovered in *Caenohabditis elegans* where lin-4 encodes anti-sense small RNA which negatively regulates lin-14 gene (Lee et al., 1993). Later, high throughput sequencing revolution has allowed the study of miRNA's genome wide role in plants and animals. In plants, miRNAs are involved in developmental, cellular, hormonal, physiological and stress responsive pathways (D'Ario et al., 2017). Tissue and stage specific miRNA have also been discovered in plants which are involved in maintaining and affecting developmental processes (Chen et al., 2012; Sunkar et al., 2012). Along with high throughput methods, bioinformatics algorithms for predicting plant miRNA have facilitated in exploring conserved and plant-specific miRNA (Lu et al., 2005; Unver et al., 2009). Large scale predictions of miRNA resulted in several validated and putative miRNA families which led to the development of specific databases for miRNA: miRBase (Kozomara & Griffiths-Jones, 2014) and PMRD (Zhang et al., 2010b) are highly accessible databases for plant-specific miRNA. miRBase (www.mirbase.org) release 21, contains 28,645 entries representing hairpin precursor miRNA, 35,828 mature miRNA products related to 223 species. PMRD includes 28,214 entries specific to 166 species of plants, this database has been upgraded with other non-coding RNAs into Plant Non-Coding RNA Database (PNRD) based on literature mining (Yi et al., 2015).

With abundant information of miRNA sequences and due to their high sequence and structure conservation, bioinformatics approaches offer robust methods to identify orthologous miRNA and plant-specific miRNA (Gomes et al., 2013; Unver et al., 2009). The most commonly used miRNA prediction methods based on NGS data are the plant version of miRDeep2 (Friedlander et al., 2012; Meyers et al., 2008; Thakur et al., 2011), UEA sRNA workbench (Stocks et al., 2012) and miRanalyzer (Hackenberg et al., 2011).

Bioinformatics algorithms predict known miRNAs from the related plant species along with novel or plant-specific miRNAs. Novel or plant-specific miRNA are detected based on the predicted capacity of sequences to form a qualifying duplex, the presence of both miRNA: miRNA* duplex sequences, presence of candidate precursors that are unique to novel miRNA and a hairpin structure conformation without large bulges in the terminal loop (Friedlander et al., 2012; Meyers et al., 2008).

Genome-wide prediction of conserved and plant-specific miRNA in non-model plants are based on utilization of sequenced raw reads. Predicting miRNA on genome with raw reads is influenced by number of unique mapping reads on genome and depth of the reference genome assembly (Budak & Kantar, 2015; Kurtoglu et al., 2014). Chromosome-based conservation of miRNA precursors in hexaploid wheat genomes (Deng et al., 2014; Kurtoglu et al., 2013) shows evolutionary conservation of miRNA in polyploid wheat. miRNA families such as miR156, miR159, miR160, miR166, miR171, miR408, miR390 and miR395 are highly conserved plant miRNA families and also linked with developmental or stress responses across embryophyta, a most populous sub kingdom of green plants (Cuperus et al., 2011). Due to fast progress in miRNA functional studies, novel or plant-specific miRNA especially in non-model plant species, may represent highly promising targets for research in the future towards exploring biological functions of plant-specific miRNA (Qin et al., 2014).

2.6.2   Validation of miRNA target pairs by degradome

To establish miRNA-mediated networks, validated miRNA-target pairs are necessary. Using high-throughput NGS technology RNA ends can be sequenced by parallel analysis of RNA ends (PARE) for degradome sequencing in plants (German et al., 2008). Degradome reads can be used to predict miRNA-target pairs by sophisticated tools such as Cleaveland4 (Brousse et al., 2014) and sPARTA (Kakrana et al., 2014). With analysis

of degradome reads, both conserved and novel miRNA-target pairs can be identified which can be sample or tissue-specific.

2.6.3 miRNA promoter prediction in plants

TFs bind to the cis-element, or transcription factor binding sites (TFBS) on the promoter region of miRNA genes and interact with the transcription start site (TSS). TFBS cis-elements are positioned upstream of miRNA genes and control transcription (Lee et al., 2007). TFBS are conserved across TF families and are present in clusters known as homotypic clusters on the promoter sites (Lifanov et al., 2003; Singh et al., 2015). In plants, TFBS motifs in the promoter region of miRNA genes were first reported in Arabidopsis (Megraw et al., 2006). Later, TFBS motifs were reported to be species specific in Arabidopsis and rice (Zhou et al., 2007). The TFBS motifs were also found to be conserved in miRNA promoter regions and reported to play key role in regulating the miRNA genes in response to abiotic stress in rice (Devi et al., 2013). miRNA promoters are located within the upstream regions of the gene bodies encoding primary transcripts (Figure 2.4). Precursors miRNA(pre-miRNAs) can be used to predict TSS and promoter region on 5' upstream region (Megraw & Hatzigeorgiou, 2010; Meng et al., 2011)
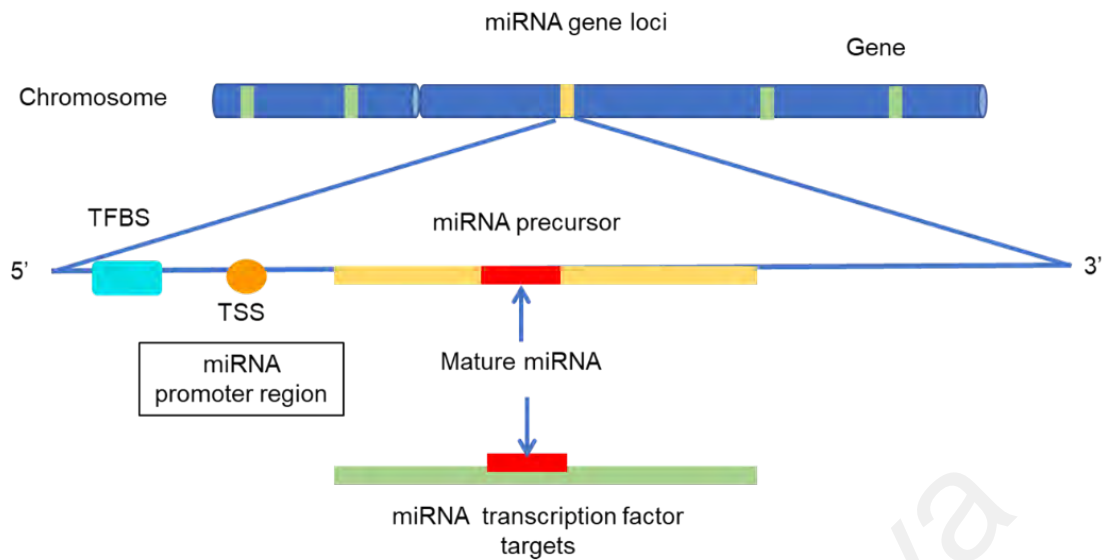
Figure 2.4: Schematic representation of miRNA gene, transcription start site and its promoter region.

2.7   Role of bioinformatics in crop improvement

Next generation sequencing data with high performance computing and bioinformatics tools revolutionized data collection, organization and integration in the field of plant breeding and genetics (Bhadauria, 2017). Bioinformatics evolved in terms of computational tools to implement multifaceted algorithms for analysing omics data. Such tools allow analysis of high-throughput omics data allowing to explore multiple omics data at single interface (Yu et al., 2017). The number of completely sequenced plant genomes has rapidly grown since the year 2000 along with relevant transcriptomic, epigenome and metagenome data (Esposito et al., 2016) (Figure 2.5). NGS data are usually in raw fragmented read format which has to be pre-processed and cleaned for downstream analysis by assembly, predictions and comparisons with reference databases (Leipzig, 2017). Such analysis will define structures, feature identification, putative functions and taxonomic assignments to further elucidate the data. A standard list of plant bioinformatics databases and list of highly accessible bioinformatics tools are available in Appendix A (Table S1 and Table S2).
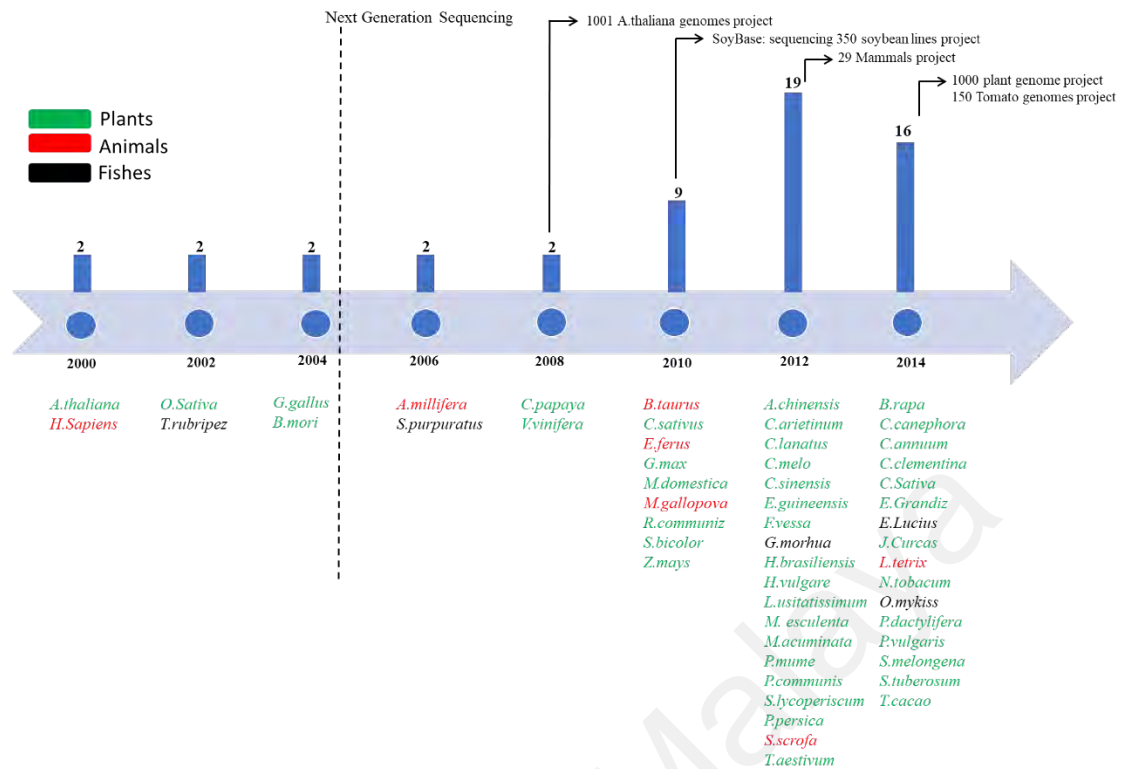
Figure 2.5: Timeline of completely sequenced plant genomes. Timeline from 2000 to 2014 in indicating the rapid growth of completely sequenced genomes of plants (green), animals (red) and fishes (black). The dashed line represents the start of the NGS era establishing massive sequencing projects. Adapted from Esposito et al. (2016).

2.8   Multi-omics approach for crop improvement.

The combination of omics approaches (multi-omics) has been an efficient approach to determine the overview of cellular and molecular activity (Ohashi et al., 2015). Agronomically important traits are always associated with understanding of genotype and phenotype variations among plants. Omics such as genomics, transcriptomics and epigenomics contribute to agriculture spans in identification and manipulation of genes associated to specific phenotypic trait (Mochida & Shinozaki, 2011) (Figure 2.6). Genomics helps in genome assisted breeding by selecting molecular markers that map within specific genes or QTLs known to be associated with traits or phenotypes (Varshney et al., 2014). Small RNA transcriptomics reports miRNA and siRNA which have emerged as an attractive tool for deciphering plant function and to support the development of

improved and novel traits by manipulation of gene expression (Kamthan et al., 2015). Such small RNA manipulations have shown desirable results in plants especially in biotic stress tolerance (Zhou et al., 2013). Similarly, epigenomics is closely associated with traits that help in plant resistance towards environmental stresses. DNA methylation patterns alter gene expression in environmental stress conditions making the plant withstand the biological changes. Such differentially methylated patterns can be termed as methylated or epi-QTLs which could be used in crop breeding programs (Pandey et al., 2016). Some of these differentially methylated patterns might be heritable and transferred to subsequent generations, such inheritance could be promising in crop improvement. Multi-omics approaches allow us to select high value functional alleles or markers for traits to develop cost effective fixed SNP genotyping arrays assisting in efficiency of plant breeding programs.
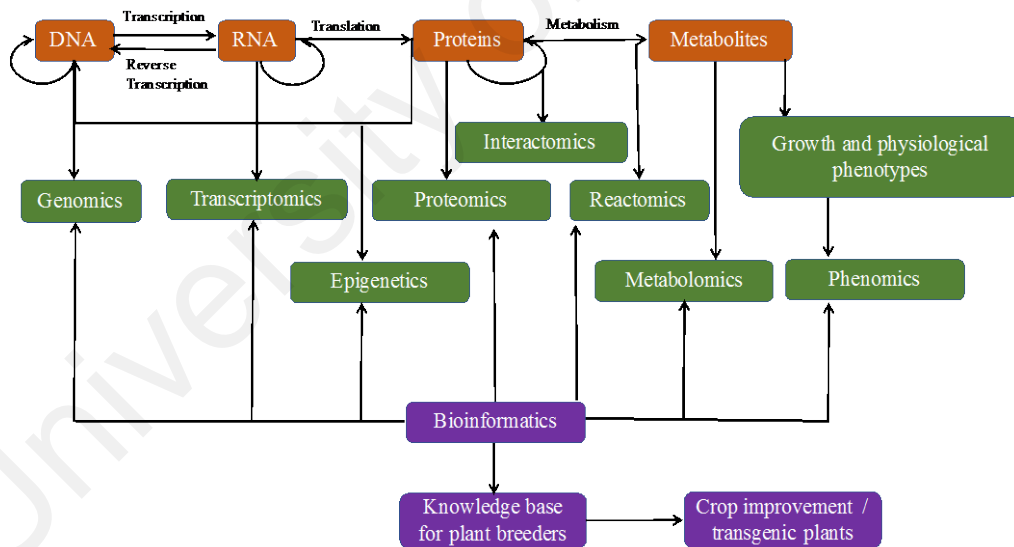
Figure 2.6: Multi-omics allowing data integration for sustainable agriculture. Major "omics" studies involving next generation sequencing (NGS) and bioinformatics in agriculture crop improvement. Adapted from Mochida and Shinozaki (2011).

CHAPTER 3: MATERIALS AND METHODS

3.1   Plant Materials and treatment

Clonal tissue cultured plantlets of *Musa acuminata* cultivar Berangan (AAA genome) were purchased from Felda Agricultural Service Sdn Bhd, Malaysia. Plantlets of 6-8 cm in height with healthy roots were selected and exposed to salt treatments by transfer to Murashige and Skoog (MS) medium (Murashige & Skoog, 1962) with no added NaCl (Control) or supplemented with 100mM NaCl (TR100) and 300mM NaCl (TR300). Root tissues were pooled from 3-5 plantlets for RNA and DNA isolation. After 48h, roots were harvested from plantlets were frozen in liquid nitrogen followed by RNA isolation. Banana embryogenic suspension cell samples were prepared from embryogenic callus induced from immature male flowers of 'Berangan' as described by Jalil et al. (2003). Newly initiated suspension cells were used for RNA isolation.

3.2   RNA isolation

Total nucleic acids were isolated from banana root tissues and embryogenic suspension cells using a modified CTAB nucleic acid isolation method (Kiefer et al., 2000). Concentrations of purified nucleotides were determined at 260 nm using a NanoDrop 2000 Spectrophotometer (Nanodrop Technologies LLC, Wilmington, DE, USA) and purity assessed at an absorbance ratio of 260/280 nm and 260/230 nm. RNA integrity was confirmed by agarose gel electrophoresis and on an Agilent 2100 Bioanalyzer (Agilent Technologies, Inc.). RNA was treated with DNase I (Invitrogen™, Life Technologies, Thermo Fisher Scientific Corporation, Waltham, MA, USA). Only samples with high RNA integrity number (RIN $\geq$ 8) were used for RNA sequencing. RNA extractions for somatic embryogenic callus were performed by Dr. Lee Wan Sin[1]. Small

---

[1] As stated in manuscript author's contribution, LWS carried out the small RNA library preparation.

28

RNA transcriptome and Degradome libraries for three samples (Control, TR100 and TR300) without replicates were sent for sequencing.

3.3 DNA isolation

DNA was isolated from banana roots using a modified CTAB-based method described by Stewart (1997). DNA extraction was performed in 2-ml scale and chloroform-isoamylalcohol (24:1) extraction step was repeated three times to obtain DNA with high purity. Absorbance at 260 nm and 280 nm was measured spectrophotometrically to determine DNA purity and Qubit® 2.0 Fluorometry assay (Invitrogen, Life Technologies, Carlsbad, CA, USA) was performed to determine DNA concentrations. Only samples with A260/A280 ratio between 1.8-2.2, A230/A260 ratio higher than 1 and concentrations 50 ng/µl and above with total amount of DNA not lower than 5 µg were used for bisulphite conversion and library construction prior to high-throughput bisulphite sequencing (BS-seq). BS-Seq was performed for two (Control & TR300) samples without replicates.

3.4 RNA sequencing

3.4.1 Library construction and small RNA sequencing

Small RNA library construction for Illumina sequencing was carried out using Illumina's kit according to the manufacturer's recommendations. In brief, small RNAs of 16 to 28 nt were recovered from high resolution gel (15% (w/v) PAGE) and then ligated with 5' and 3' Illumina adapters using T4 RNA ligase. The small RNA-adapters ligation products were amplified by RT-PCR using Illumina's small RNA primer set and the cDNA was sequenced using Illumina HiSeq 2000 II (Illumina Inc., San Diego, CA). platform by service provider Beijing Genome institute (BGI), Shenzhen, China.

3.4.2   Library construction and Degradome sequencing

Degradome cDNA libraries using sliced ends of polyadenylated transcripts were constructed based on a method described by Addo-Quaye et al. (2008) Approximately, 150 ng of poly(A)+ RNA was used as input RNA and annealing with biotinylated random primers and captured RNA fragments by streptavidin. Libraries were sequenced using the 5' adapter only after reverse transcription and PCR, resulting in the sequencing of the first 50 nucleotides of the inserts that represented the 5' ends of the original RNAs. Single-end sequencing (50 bp) was performed on an Illumina Hiseq2000 II (German et al., 2009) (Illumina Inc., San Diego, CA). by service provider Beijing Genome institute (BGI), Shenzhen, China.

3.5   DNA sequencing

3.5.1   Library construction and bisulphite sequencing

Bisulphite sequencing of DNA from banana samples was performed by reduced representation of bisulphite sequencing (RRBS) method. Experiment was performed by fragmenting genomic DNA by restriction enzyme and adding DNA-end repair, 3'-dA over hang to each fragment. Fragments of 40-220bp were selected for bisulphite treatment by ZYMO EZ DNA Methylation-Gold kit. Further, DNA library is qualified by PCR before proceeding for sequencing. RRBS was performed on an Illumina HiSeq 2000 platform (Illumina Inc., San Diego, CA) by service provider Beijing Genomics institute (BGI), Shenzhen,China.

3.6   Bioinformatics analysis of next generation sequencing (NGS) data

3.6.1   Small RNA and degradome data pre-processing

FastQC (Andrews, 2010) was used for quality control and removing reads with a phred score below 20. Adaptor sequences were identified and trimmed from each read using fastx-clipper (http://hannonlab.cshl.edu/fastx_toolkit/). Small RNA and degradome reads

were trimmed to remove adapter sequences and low-quality reads. Reads were then mapped to the Rfam database (Kalvari et al., 2018) using Bowtie (Langmead et al., 2009). Matches with tRNA, rRNA, small nucleolar RNA and sequences below 19 or above 24 nucleotides in length were not considered for further analysis. Small RNA reads were aligned to banana genomes using Bowtie2 (Langmead & Salzberg, 2012) and only reads uniquely mapped to banana genomes were considered for furthur analysis (Table S6 and Table S7).

3.6.2 Small RNA dataset preparation for results in section 4.1

A non-redundant query set of small RNA reads was compiled from root (Control) and embryogenic cell suspension and included all 235 miRNA sequences reported for the *Musa* 'Pahang' doubled haploid A-genome retrieved from the banana genome database (D'Hont et al., 2012), and publicly available small RNA data from *Musa acuminata* 'calcutta 4' leaf, flower and fruit tissues (sequenced within the framework of a NSF project and downloaded from http://smallrna.udel.edu) . As there were no available resources for the B genome and based on the known close genetic relatedness of *Musa* A and B genomes, small RNA datasets from the *Musa acuminata* cultivar were used as a reference set to predict miRNA in *Musa balbisiana* genome sequence.

3.6.3 miRNA prediction from small RNA datasets

miRNA prediction was performed using miRDeep2 tool (Friedlander et al., 2012) using scripts modified according to the criteria set for plant genomes (Meyers et al., 2008). The query miRNA data set was mapped separately to the A-genome (D'Hont et al., 2012) and the draft PKW B-genome (Davey et al., 2013). Regions of 300nt surrounding the matched position of each read were excised from the genome sequences, and then RNAfold software (Hofacker, 2003) used to predict sequences able to form stem-loop structures, using default options, whilst 'Randfold' (Bonnet et al., 2004) was

used to calculate p-values for potential miRNA precursors predicted by the 'miRDeep2' algorithm. The candidate miRNA precursors selected had the following features; a predicted stem loop structure of 75 nt and a bulge-loop size of less than 6nt; the mature miRNA was within the stem region of the precursor; less than four mismatches were allowed between the mature miRNA: miRNA* duplex; miRNA and miRNA* were on opposite arms of the precursor forming a duplex with 3' overhangs; the predicted minimum folding energy (MFE) was between -15kcal/mol to -47.2kcal/mol. Orthologous miRNA sequences/homologues were annotated by BLASTn comparison (Altschul et al., 1990) to mature and stem loop miRNA sequences from miRBase v19 (Kozomara & Griffiths-Jones, 2014). Predicted miRNA were considered novel if they had no match (allowing for a maximum of 2 mismatches i.e. n/n, (n-1)/n, (n-2)/n nucleotide matches, n=length of mature miRNA) to any entry in miRBase (release 19) and PMRD (accessed, February 2013). *Musa*-specific miRNA sequences not present in either miRBase or PMRD databases, were arbitrarily named starting at '1' and using the miRBase species-based name format. For miRNA families observed to be present in both A- and B-genomes, paralogous miRNA loci count in each *Musa* genome were estimated based on the 300nt precursor regions predicted by miRDeep2. miRNA targets were predicted with 'psRNAtarget' online server (http://plantgrn.noble.org/psRNATarget/) (Dai & Zhao, 2011) with default options. MiRNA prediction pipeline adopted for this chapter is represented as flowchart in Appendix B (Figure S2) of this thesis.

3.6.3.1   miRNA annotation and nomenclature

MiRNA nomenclature followed the definitions at miRBase (Budak et al., 2016; Kozomara & Griffiths-Jones, 2014) according to the criteria: a) Mature miRNA sequences with an exact match to the orthologous mature miRNA were named with the same family and sub classification. b) Mature miRNA sequences which were distinct and from the same precursor were named as isomirs, with subsequent numbering based on the

number of occurrences. c) Mature miRNA sequences from the 5' strand of the precursor miRNA were named with a 5p suffix and those on the 3' strand on the same precursor were named with a 3p suffix. d) Mature miRNA sequences from the same family but from different precursors were given the suffix ".1" (e.g. mac-miR156.1) and with subsequent numbering based on the number of occurrences. Novel miRNAs, and those matching miRNAs that have previously been reported only in banana were termed as "*Musa*-specific" miRNA and numbering is based on the number of occurrences. *Musa*-specific miRNA predicted in this study have a prefix of "mac-miR-new" and miRNA which match our previous study (Lee et al., 2015) are used with the same name (E.g.: mac-miR6).

3.6.3.2   miRNA promoter prediction

Two kb regions, 5'-upstream of each predicted miRNA stem loop locus on the banana genome (version-2) were retrieved using a custom python script (Appendix C Script 2). Transcription start sites (TSS) and TATA box were predicted by TSSPlant v1.2016 (Shahmuradov et al., 2017) and only miRNA with both TSS and TATA box within the upstream 2kb regions of miRNA stem loop loci were considered for promoter prediction. The 800bp upstream region for single TSS sites and sequences between multiple TSS were considered as putative miRNA promoter sequences (Megraw & Hatzigeorgiou, 2010) (Commands used to retrieve promoter sequences are in Appendix C Script 3).

3.6.3.3   Transcription factor binding site (TFBS) prediction

MiRNA promoter sequences were scanned for transcription factor binding sites (TFBS) at the PlantPAN 2.0 server (Chow et al., 2016). The frequency of all associated transcription factor families for all miRNAs were calculated using an R script (Script 4). Highly enriched TFBS motifs among miRNA promoters were filtered using a threshold of > 4 in each miRNA. miRNA, TF and TFBS networks were constructed using Cytoscape v3.6 (Shannon et al., 2003) using only the miRNA with TF targets (based on

degradome analysis). Gene ontology annotation of TFBS motifs was performed by GOMO prediction tool (Buske et al., 2010). Flowchart representing complete pipeline adopted for analysis of miRNA, degradome and scanning of TFBS represented in Appendix B (Figure S3).

3.6.4 Degradome analysis

Cleaved miRNA sites from degradome reads were predicted using the Cleaveland-4 tool (Brousse et al., 2014) with additional cleavage sites between the 9th and 10th or the 11th and 12th positions of miRNA sequences (Zheng et al., 2012). Degradome reads were mapped to *Musa acuminata* gene coding regions by Bowtie version-1 (Langmead et al., 2009) to predict mRNA-miRNA pairs and which specifies sites of cleavage. Cleaveland-4 categorizes five types of read-matches based on the abundance of the degradome reads matching to the cleavage site, to determine miRNA targets. Category-0 denotes > 1 raw read at a position with complementarities to the mature miRNA, and the abundance at that position equals to the maximum (where "maximum" means the maximum number of matches of any mature miRNA on the transcript and only one maximum on the transcript). Category-1 denotes > 1 raw read, with abundance at the position equal to the maximum on the transcript, and more than one maximum position on the transcript. Category 2 denotes >1 raw read, above the average depth (where "average" is from all positions that have at least one read), but not the maximum on the transcript. Category 3 denotes >1 read, but below or equal to the average depth of coverage on the transcript. Category 4 denotes only one read at mapping to position. To obtain high confidence miRNA targets from Cleaveland-4 results, miRNA targets were filtered by MFE ratio ≥ 0.65, P-value ≤ 0.05, Category 0–3. MiRNA-targets were annotated as belonging to gene ontology (GO) categories using the biomart tool from the plant Ensembl database (http://plants.ensembl.org/biomart/martview) (Kinsella et al., 2011).

3.6.5   Small RNA clusters on genome

Small RNA reads of 21nt and 24nt length were filtered from control (UT) and 300mM NaCl (ST) samples of small RNA-seq which were aligned to banana genome using Bowtie2 (Langmead & Salzberg, 2012) and only reads uniquely mapped to banana genome were considered for furthur analysis. Small RNA from each sample were normalized to transcripts per million (TPM) by formula: TPM = (siRNA abundance/Total abundance) * 1,000,000. Coverage of small RNA on genome using generic feature format (GFF3) by coverageBed utility in BEDtools (Quinlan & Hall, 2010) and graphs were generated using "ggbio" (Yin et al., 2012) from bioconductor (www.bioconductor.org) package in R. siRNA clusters were generated by merging all reads mapped within 100bp regions all over the genome. Differential expression of clusters were calculated by DEGSeq (version 1.15, http://bioinfo.au.tsinghua.edu.cn/software/degseq) (Wang et al., 2010), a method based on assumptions for the read count, using a multiple averaging (MA) plot based random sampling model (MARS). False discovery rate (FDR)< 0.05 with log2Ratio >2 & < -2 were set as threshold values.

3.6.6   Analysis of Bisulphite sequencing (RRBS) reads.

BS-seq reads were mapped to reference banana genome using bismark version 0.12.3 (Krueger & Andrews, 2011) with default parameters. Uniquely mapped reads were considered for further analysis. The mapped reads were converted to "mr" format for further analysis using MethPipe v0.16.3 (Song et al., 2013a) tool. To compute the single-site methylation levels, methcounts program from MethPipe was used to determine the methylation levels in CG, CHG and CHH contexts. Coverage of methylation level on genome were plotted by "ggplot2" (Wickham, 2009) and "ggbio" (Yin et al., 2012) from bioconductor (www.bioconductor.org) package in R. To calculate DMR regions, the following steps were performed 1) prefiltering for each position to select regions with number of reads mapping ≥ 10 and number of methylated reads ≥ 10 (for methylation) or

= 0 (for loss of methylation); 2) DMR calculation is done by DSS method (Park & Wu, 2016) which estimates the dispersion parameter from Gamma-Poisson or Beta-Binomial distributions with parameters delta=0.1, p value threshold=0.05, minimum length of DMR = 100, minimum number of CG/CHG/CHH methylated reads = 10 and merge DMR within 100bp distance (Appendix C Script 8).

3.6.7 Analysis of transcriptome reads

Transcriptome raw reads were also obtained from NCBI's SRA database with accession numbers SRR1339507 and SRR1339902 for control (UT) and 300 mM NaCl (ST) treatment respectively (Lee et al., 2015). SRA reads were converted into fastq format by using SRA-toolkit (Leinonen et al., 2011) before preprocessing and filtering. FastQC (Andrews, 2010) was used for quality control and removing reads < 20 phred score. Adaptor sequences were identified and trimmed from each read using fastx-clipper (http://hannonlab.cshl.edu/fastx_toolkit/). Transcriptome clean reads were aligned to banana genome using Bwa software (Li & Durbin, 2009). Transcriptome reads uniquely mapped to banana genome were considered for counting reads on each gene. Gene counts were normalized to transcripts per million (TPM) by formula: TPM=(RNA abundance/Total abundance) * 1,000,000. An R package, DEGSeq (version 1.15, http://bioinfo.au.tsinghua.edu.cn/software/degseq) (Wang et al., 2010) was used to identify differentially expressed genes based on binomial assumptions for the read count, using a multiple averaging (MA) plot based random sampling model (MARS). FDR < 0.05 and log2Ratio >2 & < -2 were set as threshold values to select for differentially expressed transcripts (Appendix C Script 9).

3.6.8 Data availability

Data generated from banana root salt stress treatment are available under NCBI Bioproject accession PRJNA246442. Small RNA datasets are available under NCBI SRA

accession numbers SRR1524838 (Control), SRR1340394 (TR100), SRR1524839 (TR300) which were used for miRNA and siRNA annotation. Degradome sequencing reads are available under NCBI SRA with accession numbers SRR5337783 (Control), SRR5337782 (TR100), SRR5337781(TR300) which were used for miRNA target prediction. Bisulphite sequencing reads will also be available at same NCBI Bioproject accession once accepted for publication. *Musa balbisiana* genome published as part of current study is available for download at http://banana-genome-hub.southgreen.fr/organism/Musa/balbisiana.

3.6.9   Data sources

Repetitive parts of the banana genome (DNA_hAT, LINE, LTRGypsy, LTRCopia, unclassified RE, rDNA_Satellite, clDNA) consisting of 1902 sequences were retrived from a published report (Hribova et al., 2010). *Musa acuminata* version-1 and *Musa acuminata* version-2 reference genomes and gene-coding sequences were downloaded from the banana genome hub at http://banana-genome-hub.southgreen.fr (Droc et al., 2013). Supplementary datasets for Appendix D and Appendix E are available in CD provided with this thesis or openly accessible at GitHub (https://github.com/ranga85/Thesis-Datasets).

3.6.10   Gene and Repeat annotations in *Musa* A-and B-genomes

Protein coding gene sequences in the B-genome were predicted by *ab initio* gene prediction using FGENESH software (http://linux1.softberry.com/all.htm). Repeats or transposable elements (TE) were detected in the *Musa* A- and B-genome sequences using Repeat Masker tool (http://www.repeatmasker.org).

CHAPTER 4: RESULTS

4.1 Comparative genomics of banana A- and B-genomes

4.1.1 miRNA prediction on banana A and B genomes

A non-redundant set of plant miRNAs, which included the 37 miRNA families (representing 234 precursors) previously reported for the *Musa acuminata* 'Pahang' A-genome (D'Hont et al., 2012) was used to predict miRNA precursors and families within both A and B *Musa* genomes (Figure 4.1). The results show a slightly larger number of predicted orthologous miRNA precursors for the B-genome (270 miRNA precursors compared to 266 for the A genome), but the diversity of orthologous miRNA families was lower, with 42 families predicted for the B-genome compared to 47 families for the A-genome. All of the known miRNA families detected in the B-genome were also found to be present in the A-genome. Overall, 10 additional miRNA families were found compared to those reported by D'Hont et al. (2012) .The additional miRNA families detected were miR415, miR529, miR1134, miR5021, miRf10125, miRf10576, miRf11033, miRf11036, miRf11143 and miRf11357. Of these, only miR415 and miRf11036 were not detected in the B-genome. These new families may be due to additional entries added to the PMRD database since the analysis by D'Hont and colleagues, also may be due to the new *M. acuminata* small RNA sequence data used in our query dataset. These were large (averaging >11 million Illumina sRNA-seq "clean reads" per library) libraries, and derived from several different banana tissues (leaf, root, flower, fruit and somatic embryogenic cultures) (refer to Table S3 for dataset statistics).
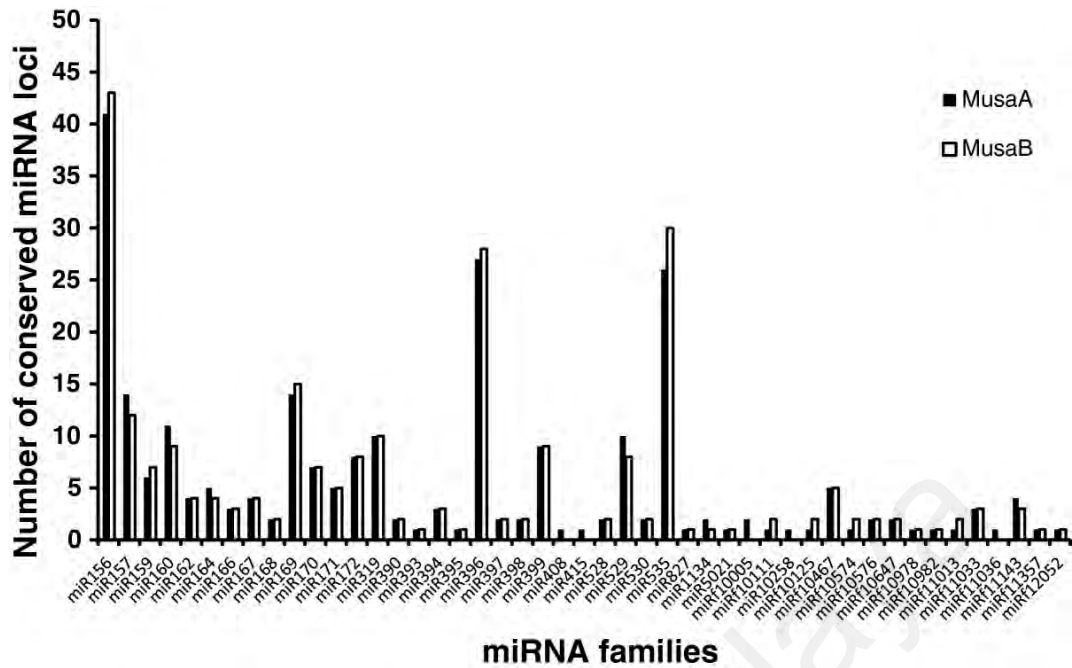
Figure 4.1 : Overview of numbers of conserved miRNA families present in the Musa A- and B-genomes (Davey et al., 2013). In addition to the orthologous miRNA families, there were also 32 *Musa*-specific miRNA precursors predicted, that belong to 28 *Musa*-specific miRNA families with no significant match to any previously reported mature miRNA sequence (Table 4.1). These include sequences that were unique to either *Musa* A- or B-genomes in addition to four families common to both genomes. According to authors' contributions in manuscript, I have carried out small RNA data generation and analysis. Hence my contribution for the data and analysis shown in this figure is 100%.

Table 4.1: Predicted *Musa*-specific miRNA in *Musa* A and B genomes.

miRNA family[a]	Locus A-genome	Locus B-genome	Mature-miRNA sequence 5' – 3'
mba-miR1		chr11: 8494985..8495047	AGAAACUUUUGUUGGAGAGGAAC
mac-miR2 mba-miR2	chr5:6134793..6134884	chr5: 5463327..5463417	CCGCAGGAGAGAUGAUGCCGCU
mba-miR3		chrUn_random:810706..810748	UACCGUACUGUACCGGCGUUU
mba-miR4		chr10: 20915746..20915798	CCUGAUUUGCUAAGUAGAUUU
mba-miR5		chr7: 16539311..16539413	UGGUUGAUGACGAUGUCGGCC
mac-miR6 mba-miR6	chr7:1377251..1377323	chr7: 1254647..1254718	UAGGAGAGAUGACACCGGCU
mac-miR7 mba-miR7	chr1:10560884..10561005	chr1: 9229340..9229458	AAACUAGUGCUAAGACCCAAUCUC
mba-miR8		chr1: 10667301..10667378	GGUGGUCUGGAUGAGGAUGCC
mac-miR9 mba-miR9	chr7:4569849..4569909	chr7: 4219274..4219333	UGGCUGAUGAUGAGUGAUCUU
mba-miR10		chr8: 20785187..20785267	CUUUGGCUUCUGGGUAGACGUA
mba-miR11		chr9: 8635028..8635089	UGUACGGAUAUGGUAGAGGGGCGU
mba-miR12		chr4: 17987962..17988016	AUCCCCGAGUGGGGUCGGUCGGAC
mba-miR13		chr8: 23338135..23338275	CUCGAGAUAUAUGAGUGUGGACA
mba-miR14		chrUn_random: 21488242..21488315	GGCACCUCGAUGUCGGCUC
mba-miR15		chr9: 25179371…25179434	GAGGAGGAGAAGAAAUGGAUCUG

[a]miRNA family name with prefix "mac" and "mba" represents miRNA family from Musa A genome and B genomes. Locus represents genomic position (chromosome: start..end) of miRNA precursor on Musa A and B   genomes.

Table 4.1, continued.

mba-miR16		chr6: 3503608..3503675	GAAGAGGAAGGAGAAGUCG
mba-miR17		chr1: 17563902..17563986	CAGAAGUAGAAUACAUAAC
mba-miR18		chrUn_random: 135937842..135937988	UCCUUUUAGACCGUUGACGA
mac-miR19	chr4:22573796..22573893		UCCAGGAGAGAUGACACCAAC
mac-miR20	chr1:4998969..4999025		GGCGAUGAUGAUUGGUGAAUGU
mac-miR21	chrUn_random:15968301..15968390		GGAGAGAUGGCUGAGUGGACUAAA
mac-miR22	chrUn_random:23270434..23270480		CGAGGUGUAGCGCAGUCUGG
mac-miR23	chr8:31825102..31825180		UGGGAAGAAGACAAGGACAACAUG
mac-miR24	chr6:8249043..8249103		GAUCUCUGACCGAGCGGACUCC
mac-miR25	chr4:345377..345475		CAACGAUGAUGAGCCUACUAGACC
mac-miR26	chr11:15733314..15733359		AGAUGAGGUAAAGUAGUGCGA
mac-miR27	chr6:9168756..9168839		CAGCGACCUAAGGAUAACU
mac-miR28	chr7:28606179..28606234		GCGGAUGUGGCCAAGUGGU
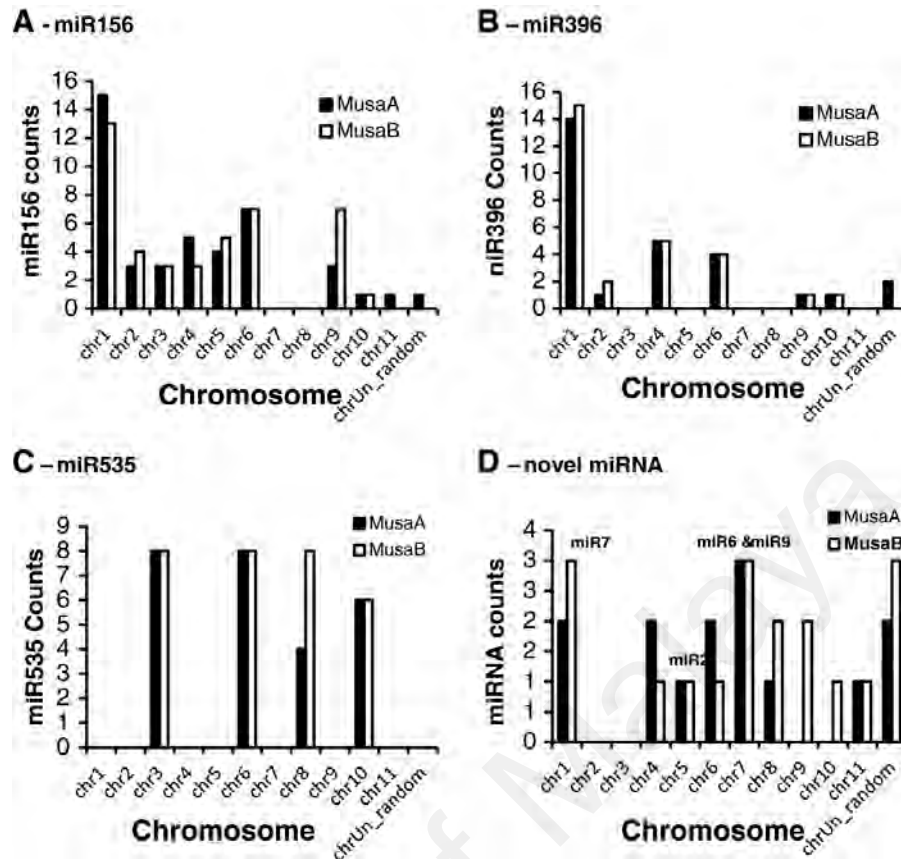
Figure 4.2: Distribution of known and novel (Musa-specific) miRNA families. Overview of the distribution of known and novel (*Musa*-specific) miRNA families over the *Musa* A- and B-genomes. A. miR156, B. miR396, C. miR535, D. novel miRNA RNA families. (Davey et al., 2013). According to authors' contributions in manuscript, I have carried out small RNA data generation and analysis. Hence my contribution for the data and analysis shown in this figure is 100%.

4.1.2 Genome distribution of miRNA precursors

The three orthologous miRNA families that were most highly represented in both *Musa* A- and B-genomes showed similar patterns of distribution across the chromosomes of both genomes, (Figure 4.2) suggesting synteny of A- and B-genomes. However, as the B-genome was assembled using the A-genome as reference, the gene order in this draft are preliminary and validation by FISH or similar methods will be needed to confirm this. As would be expected, the more recently evolved, *Musa*-specific miRNA families that are unique to either the *M. acuminata* or *M. balbisiana* genome are distributed evenly

across the genomes, with the exception of A-Chr2, B-Chr2, A-Chr3, B-Chr3, and A-Chr9 and A-Chr10, which lack any these sequences (Figure 4.2D).

4.1.3 Comparison of *Musa* A and B genome gene annotation

Protein coding gene sequences in the B-genome were predicted by *ab initio* gene prediction using FGENESH software (http://linux1.softberry.com/all.htm). This resulted in the identification of 39,914 unique gene models. This number is higher than the 36,483 predicted from direct transfer of annotated regions of the A-genome to the B-genome and suggests that the gene count has been overestimated to the extent of around 109%. The higher predicted gene count is largely due to a nearly 2-fold higher number of gene models located within the concatenated contig set 'B_chrUn_random', relative to the chrUn_random of the A-genome (Table 4.2). The set of PKW B-genome gene models were descriptively annotated online using the Blast2Go software (Conesa & Gotz, 2008). Blast results against the NCBI non-redundant protein database show that 38,886 (97.4%) of the sequences had a positive hit, of which 30,541 had an e-value of '0'. Following annotation steps, GO terms could be assigned to 37,367 (93.6%) sequences and 34,044 (85.3%) were annotated by InterProScan. Based on the annotations assigned here, we can see that the B-genome contains 3,276 transposable elements (TE's), of which 1,470 are located in the B-unChr_random sequence. If these TS's are removed, the final functional B-genome gene count is 36,638, which is almost identical to the A-genome count of 36,542. Gene annotation data in gff3 format is available for download at http://banana-genome-hub.southgreen.fr/organism/Musa/balbisiana.

4.1.4 Repeat detection and annotation in *Musa* A and B genomes

Repeats or transposable elements (TE) were detected in the *Musa* A- and B-genome sequences using RepeatMasker tool (http://www.repeatmasker.org). In total, repetitive regions were found to occupy 26.85% (108.1 Mbp) of the PKW consensus B-genome

(Table 4.3), which is similar to the 27.76% reported by D'Hont et al. (2012) for the A-genome. Annotation of the repetitive sequences of the B-genome showed that overall numbers of repeat elements is slightly higher in the B-genome and the Ty1/copia and Ty3/Gypsy repeats dominate, representing 18.8% and 6.3% of the genome respectively (Table 4.3). Whilst the numbers of Non-LTR transposons (LINE), DNA transposons (clDNA and DNAhat) and Satellite repeats (Type 1 and Type 2) are similar in both A- and B-genomes and represent less than 1% of the total consensus B-genome sequence, the LTR transposons (Ty1/copia and Ty3/Gypsy) are more abundant in the B-genome.

Table 4.2: Comparison of the *Musa* A- and B-genome annotations.

	Musa acuminata (A genome)				
	Length (bp)	**CDS**	**Gene**	**mRNA**	**miRNA**
Chr1	27,573,629	2,835	2,942	2,835	31
Chr2	22,054,697	2,327	2,384	2,327	16
Chr3	30,470,407	3,251	3,337	3,251	21
Chr4	30,051,516	3,367	3,465	3,367	39
Chr5	29,377,369	2,971	3,057	2,971	22
Chr6	34,899,179	3,698	3,794	3,698	16
Chr7	28,617,404	2,765	2,834	2,765	14
Chr8	35,439,739	3,454	3,536	3,454	14
Chr9	34,148,863	3,109	3,193	3,110	22
Chr10	33,665,772	3,155	3,233	3,155	12
Chr11	25,514,024	2,677	2,762	2,678	13
chrUn_random	141,147,818	2,927	3,054	2,927	15
Total	472,960,417	36,536	37,591	36,538	235
	Musa balbisiana (B genome)				
	Length	**CDS**	**Gene**	**mRNA**	**miRNA**
Chr1	22,038,404	2,832	2,827	2,827	57
Chr2	17,349,238	2,331	2,330	2,330	32
Chr3	24,161,952	3,217	3,216	3,216	27
Chr4	24,656,528	3,301	3,298	3,298	23
Chr5	23,648,591	3,115	3,114	3,114	20

Table 4.2, continued.

Chr6	27,831,592	3,761	3,757	3,757	26
Chr7	22,212,853	2,933	2,929	2,929	17
Chr8	27,665,716	3,600	3,594	3,594	17
Chr9	25,900,723	3,284	3,283	3,283	18
Chr10	25,230,959	3,345	3,340	3,340	15
Chr11	20,721,546	2,718	2,714	2,714	7
chrUn_random	141,129,053	5,515	5,512	5,511	5
Total	402,547,155	39,952	39,914	39,913	264

Table 4.3: Overview and classification of the repeats present in the *Musa* A and *Musa* B genomes.

	Musa acuminata (A genome)			*Musa balbisiana* (B genome)		
Class	**Count**	**Bp**	**%**	**Count**	**Bp**	**%**
Ty1/Copia	5,606	3,158,199	0.67%	5,616	2,760,972	0.69%
copia/Angela	32,073	20,697,639	4.38%	32,056	19,380,064	4.81%
Copia/ SIRE1Maximus	90,910	62,820,929	13.28%	97,868	49,333,251	12.26%
Copia/Tnt1	4,191	5,137,617	1.09%	4,377	4,320,053	1.07%
Ty3/Gypsy	6,236	6,717,506	1.42%	6,542	5,554,874	1.38%
Gypsy/CRM	1,051	1,124,528	0.24%	973	1,016,030	0.25%
Gypsy/ Galadriel	1,992	2,997,110	0.63%	2,244	2,739,827	0.68%
Gypsy/Galadriel-lineage	1	28	0.00%	2	296	0.00%
Gypsy/Reina	16,445	11,955,226	2.53%	15,882	10,331,187	2.57%
Gypsy/Tekay	9,234	7,545,095	1.60%	9,245	5,851,644	1.45%
LINE	2,868	1,824,495	0.39%	2,544	1,580,226	0.39%
RE	14,494	5,415,085	1.14%	13,794	3,525,005	0.88%
Satellite/Type1	274	523,572	0.11%	299	484,199	0.12%
Satellite/Type2	68	103,955	0.02%	41	24,429	0.01%
clDNA	2,434	517,168	0.11%	2,491	487,816	0.12%
DNA/hAT	1,952	764,792	0.16%	1,818	675,712	0.17%
Total	189,829	131,302,944	27.76%	195,792	108,065,585	26.85%

4.1.5   Targets of novel B-genome miRNA

Using stringent cut-off expect value of 2.0, 18 predicted *Musa*-specific miRNA families were found in the B-genome (Table 4.4). Of these, seven (mba-miR3, mba-miR5, mba-miR8, mba-miR12, mba-miR13, mba-miR15 and mba-miR18) were predicted to have targets within coding regions of the B-genome sequence (Table 4.4). None of these predicted miRNA families were present in the A-genome and thus are presumed to be B-genome specific in function and to have evolved after the divergence of the *M. balbisiana* and *M. acuminata* species ~4.6 Mya (Lescot et al., 2008). Predicted targets of these B-genome candidate miRNAs correlate with a range of functions across plant development and metabolism and notably, several are proposed to be involved in tolerance or responses to biotic and abiotic stress. Several of the B-genome specific miRNA predicted targets, are proteins with unknown function, and appear to be only present in the B-genome and thus may represent B-genome distinct functional networks.

Table 4.4: Novel (*Musa*-specific) miRNA targets in *Musa* B genome.

Musa-specific miRNA	Gene ID	Gene Description
Mba-miR3	ITC1587_Bchr2_P03722	ferredoxin-dependent glutamate synthase
	ITC1587_Bchr9_P25070	casein kinase
Mba-miR5	ITC1587_Bchr5_G12598	alpha-amylase precursor
	ITC1587_Bchr7_G20691	alpha-amylase precursor
	ITC1587_Bchr5_G12404	alpha-amylase precursor
	ITC1587_Bchr5_G12400	kinase interacting family protein
	ITC1587_Bchr10_G31545	alpha-amylase precursor
	ITC1587_Bchr8_G21814	histone deacetylase 19-like
	ITC1587_BchrUn_random_G34767	histone deacetylase 19-like
Mba-miR8	ITC1587_Bchr1_G01417	multidrug resistance
	ITC1587_Bchr10_G30154	multidrug and toxin extrusion protein 1-like
	ITC1587_Bchr3_G07121	multidrug and toxin extrusion protein 2-like
	ITC1587_Bchr4_G10476	multidrug resistance
	ITC1587_chr11_G31983	multidrug resistance
	ITC1587_chr2_G03828	sal1 phosphatase-like
	ITC1587_chr6_G16462	protein

Table 4.4, continued.

Mba-miR8	ITC1587_chr2_G03989	multidrug and toxin extrusion protein 2-like
	ITC1587_chr8_G22505	multidrug resistance
	ITC1587_chr11_G32521	protein
	ITC1587_chr6_G17311	sigma factor sigb regulation protein rsbq
Mba-miR12	ITC1587_chr9_G26340	retrotransposon ty3-gypsy subclass
	ITC1587_chr1_G01861	retrotransposon unclassified
Mba-miR13	ITC1587_chr4_G11666	transcription factor bhlh123-like
	ITC1587_chr7_G18824	hydroxysteroid 11-beta-dehydrogenase 1-like protein
Mba-miR15	ITC1587_Bchr2_P03946	leucine-rich repeat receptor-like protein kinase at2g19210-like
	ITC1587_Bchr9_P26004	dirigent-like protein
	ITC1587_BchrUn_random_P35259	No annotation
	ITC1587_Bchr6_P16374	helicase domain-containing protein
	ITC1587_Bchr10_P31031	protein
	ITC1587_Bchr7_P20281	hypothetical protein VITISV_041103
Mba-miR18	ITC1587_Bchr10_P28401	hypothetical protein VITISV_043980
	ITC1587_Bchr8_P23495	retrotransposon-like protein
	ITC1587_BchrUn_random_P35706	hypothetical protein VITISV_006955
	ITC1587_Bchr9_P26834	retrotransposon ty1-copia subclass
	ITC1587_BchrUn_random_P34702	frigida-like protein
	ITC1587_Bchr9_P26967	PREDICTED: uncharacterized protein LOC101218085
	ITC1587_Bchr8_P23107	PREDICTED: uncharacterized protein LOC101218085
	ITC1587_chrUn_random_G35193	No annotation
	ITC1587_chr7_G19362	PREDICTED: uncharacterized protein LOC101218085
	ITC1587_chr7_G19340	hypothetical protein VITISV_043746
	ITC1587_chr9_G26937	retrotransposon-like protein
	ITC1587_chr10_G28978	hypothetical protein VITISV_041588
	ITC1587_chrUn_random_G37729	hypothetical protein VITISV_022540
	ITC1587_chrUn_random_G36021	hypothetical protein VITISV_030841
	ITC1587_chr8_G23163	opie1 pol protein
	ITC1587_chr3_G06388	retrotransposon-like protein
	ITC1587_chr7_G20383	phytoalexin-deficient 4-2 protein
	ITC1587_chrUn_random_G37096	PREDICTED: uncharacterized protein LOC101218085

4.2 Salt stress responsive miRNA and miRNA targets in banana roots.

4.2.1 miRNA promoter prediction

To predict miRNA promoter regions, we examined the distribution of transcription start sites (TSS) and TATA boxes in 5'-upstream 2 kb region of miRNA stem loop encoding loci. Among the 114 miRNA upstream regions, 96 miRNAs have both TSS and TATA box sites and of these, 68 have multiple TSS and 28 have single TSS. The distance between TSS to the first nucleotide of miRNA stem loop encoding loci was calculated (Figure 4.3A). It was noted that no TSS were detected in the 200bp upstream from the first nucleotide of the miRNA stem loop encoding loci. Positional distribution of TATA boxes shows most located between -30 to -35 bp from the respective TSS sites (Figure 4.3B).
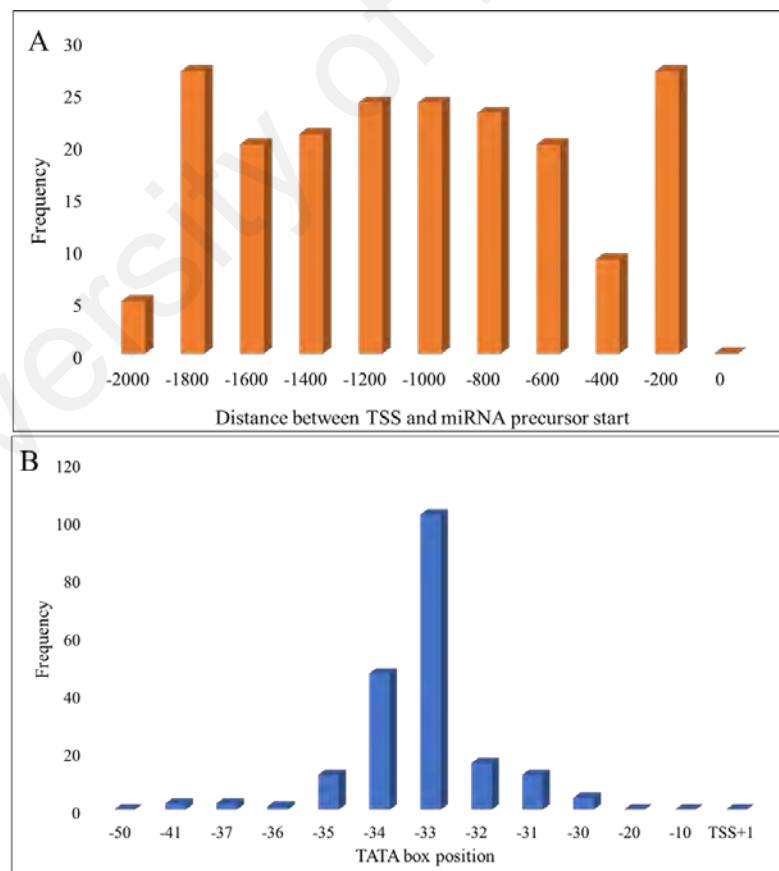

Figure 4.3: TSS and TATA box distribution on miRNA promoter region. A) Frequency distribution of TSS in region 2kb upstream of miRNA precursor start sites. B) Frequency distribution of position of TATA box with respect to TSS+1.

4.2.2   miRNA distribution on the banana genome version-2

Annotation of miRNA from small RNA datasets yielded 180 mature miRNAs. These comprised 143 orthologous miRNAs belonging to 20 orthologous miRNA families and 39 *Musa*-specific miRNAs, of which 36 were not previously reported (Appendix D Table1) (refer to section 3.6.9 for data source). A total of 114 non-redundant miRNA precursors were identified on the *Musa acuminata* genome version-2 for orthologous and *Musa*-specific miRNA. MiR156 is the most represented orthologous miRNA family across the genome with 29 mature miRNAs (including isomiRs) and 14 precursors distributed across 7 of the 11 chromosomes (Figure 4.4). At least one *Musa*-specific miRNA precursor was present on each chromosome of *Musa acuminata* genome version-2 with a total of 39 *Musa*-specific miRNAs (Figure 4.4).
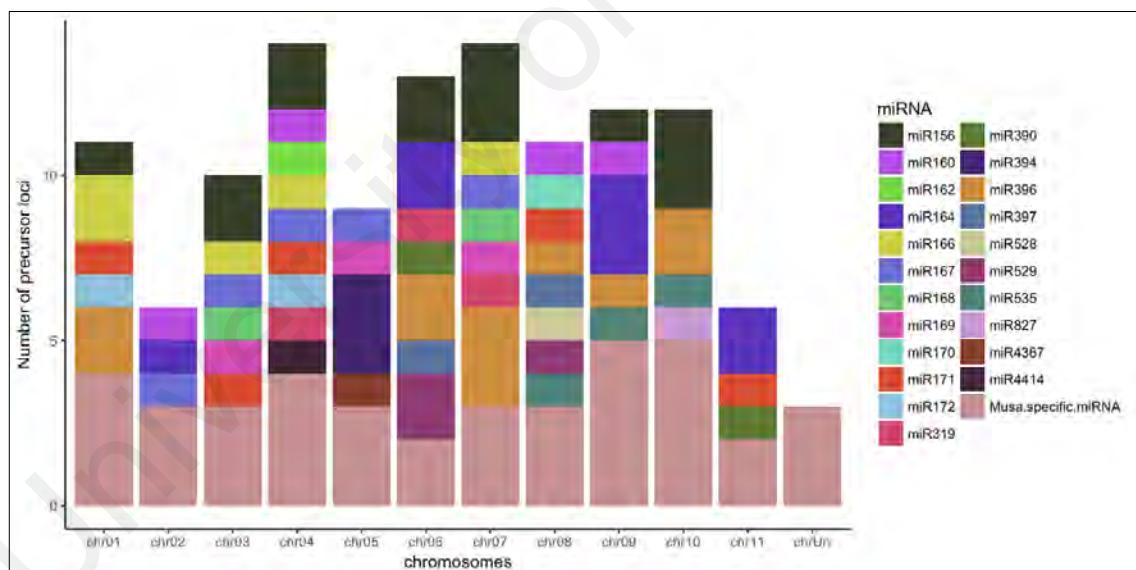

Figure 4.4: miRNA precursor distribution in the banana genome version 2. Bar height represents the number of precursor loci for each miRNA family.

4.2.3 Identification of TFBS within miRNA promoter region

Analysis of transcription factor binding site (TFBS) motifs in the sequences 800 bp upstream of TSS of 96 miRNAs, identified 30,461 potential TFBS belonging to 38 TF families (Figure 4.5). TFBS motifs in miRNA promoter regions are highly abundant for the TF families TCP (15.85%), AP2-ERF (12.61%), GATA; tify (12.23%), NF-YB; NF-YA; NF-YC (8.96%), Dof (7.66%), B3 (7.01%), BZIP (6.96%), Trihelix (5.76%), ZF-HD (3.78%), Dehydrin (3.52%) and bHLH (2.69%) (Appendix D Table2) (refer to section 3.6.9 for data source). In addition to TFBS motifs, it was observed that the signature motifs of 301 cis-elements occur across the 96 miRNA promoter regions, with the 10 most abundant cis-elements being IBOXCORE, TAAAGSTKST1, WBOXATNPR1, WBOXHVISO1, SURECOREATSULTR11, MYBST1, BIHD1OS ABRELATERD1, MYBCOREATCYCB1 and SORLIP1AT (Figure S5). Gene Ontology (GO) annotation of highly abundant TFBS revealed the motifs belonging to TCP, AP2; ERF, NF-Y, B3 and Dehydrin have role in ribosome biogenesis, while, ZF-HD and bHLH motifs show significant matches to several biotic and abiotic responses such as response to gibberellin, jasmonic acid, cold, ABA, wounding and water deprivation (Table 4.5).
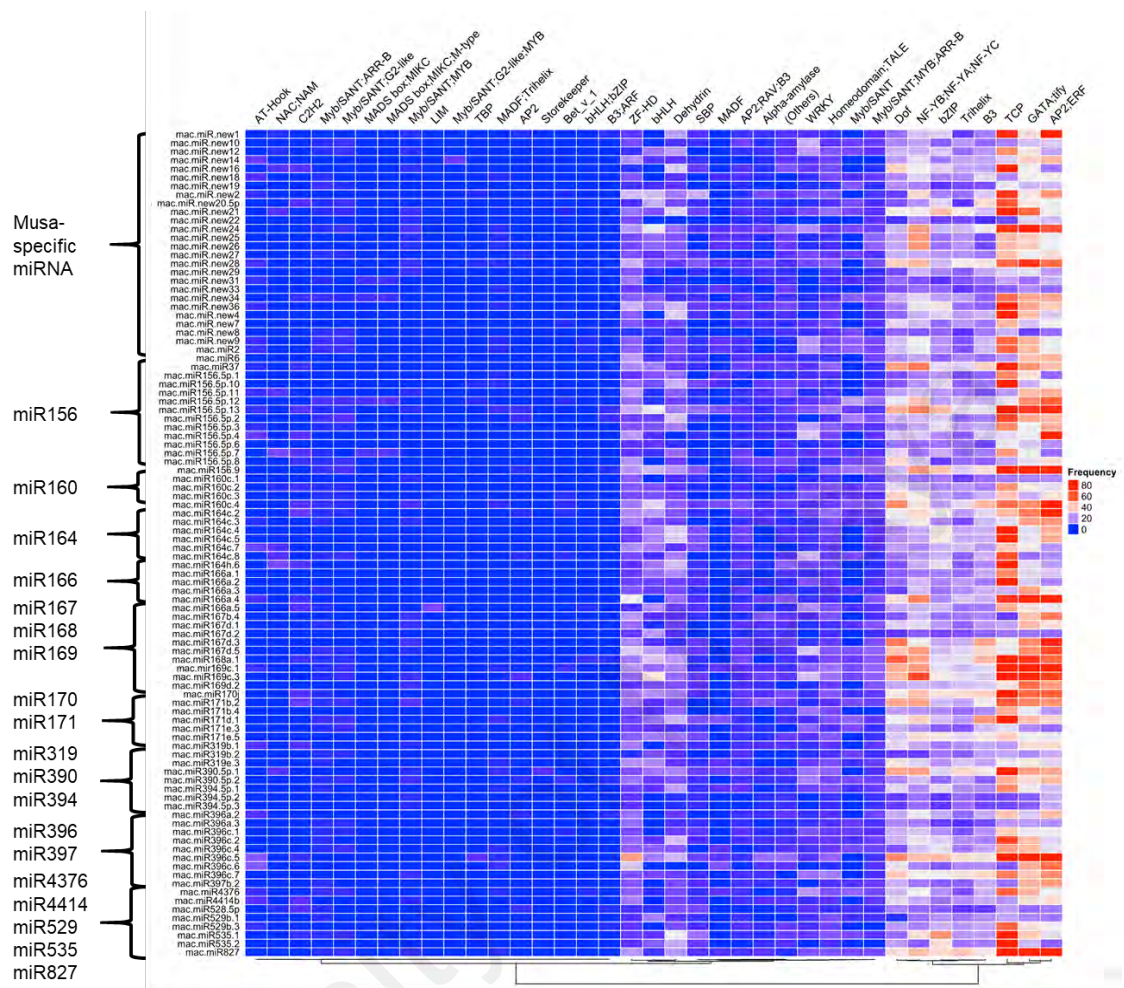
Figure 4.5: TFBS motif frequencies within miRNA promoter sequences. miRNA promoter regions are 800 bp region located 5'-upstream of predicted TSS as explained in Methods Section 3.6.3.2. TFBS motifs scanned across each promoter region were clustered by their frequencies based on Ward's clustering method and Euclidean distance using R library "ComplexHeatmap". Red represents high frequency of TFBS within a miRNA promoter region and blue represents low frequency to absence of TFBS.

Table 4.5: TFBS motif gene ontology annotations from GOMO prediction tool.

Motif	Consensus Logo[a]	Top 5 Predictions[b]	q-value[c] (<0.05)
TCP		MF structural constituent of ribosome CC cytosolic large ribosomal subunit CC nucleolus BP ribosome biogenesis MF nucleotide binding	7.61E-05
AP2; ERF		BP translation MF structural constituent of ribosome CC chloroplast stroma CC mitochondrial inner membrane CC cytosolic large ribosomal subunit	8.257e-05
GATA;tify		CC chloroplast CC mitochondrion	1.659e-04

[a]Consensus Logos generated by GOMO prediction which represents the consensus motif for each TF family.
[b]GOMO prediction considered all TFBS motifs from the promoter regions of 96 banana miRNA.
[c]The q-values are calculated using the method proposed by Benjamini and Hochberg (1995).

Table 4.5, continued.
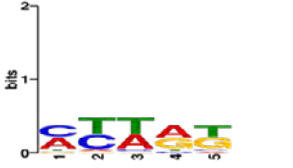
NF-YB; NF-YA; NF-YC		MF structural constituent of ribosome BP translation BP ribosome biogenesis CC cytosolic large ribosomal subunit CC light-harvesting complex	3.326e-04
Dof		No predictions	
B3		BP translation MF structural constituent of ribosome CC chloroplast thylakoid membrane CC chloroplast envelope CC mitochondrial inner membrane	1.104e-04
bZIP		CC chloroplast thylakoid membrane CC chloroplast stroma CC chloroplast envelope CC mitochondrial inner membrane MF transcription factor activity	3.309e-04

Table 4.5, continued.

ZF-HD		MF transcription factor activity CC endomembrane system BP regulation of transcription, DNA-dependent BP response to gibberellin stimulus BP response to jasmonic acid stimulus	1.990e-04
bHLH		MF nutrient reservoir activity BP response to abscisic acid stimulus BP response to cold BP response to wounding BP response to water deprivation	9.936e-04
Dehydrin		MF structural constituent of ribosome CC chloroplast envelope CC chloroplast stroma CC chloroplast thylakoid membrane CC mitochondrial inner membrane	1.098e-04

4.2.4　miRNA target genes determined by degradome sequencing

Degradome sequencing of banana roots treated with 100mM NaCl (TR100), 300mM (TR300) NaCl and control (CTR) yielded 11.75, 11.88 and 18.39 million clean reads, respectively. Of these 56.05%, 41.04% and 61.13% of the non-redundant set of degradome clean reads were mapped to gene coding sequences from the banana genome. The miRNA cleavage sites within miRNA target mRNA sequences, were predicted by Cleaveland-4 (criteria of selection of miRNA targets is elaborated in section 3.6.4), to identify 128 miRNA target genes at a threshold of Category 3 and p-value ≤ 0.05. Around 60% i.e. 79 miRNA targets were predicted to be in Category 0 i.e. high abundance of reads targeting only one position within a cleavage site (Figure S6). Of the 128 predicted targets 105 were for miRNA that have orthologous targets in other plants (Appendix D Table3) (refer to section 3.6.9 for data source), and 22 were for 12 *Musa*-specific miRNAs, of which three miRNAs (mac-miR2, mac-miR6, mac-miR37) are previously reported (Lee et al., 2015) (Table 4.6).

Table 4.6: *Musa*-specific miRNA targets in the banana genome.

Musa-specific miRNA	miRNA sequence (5' – 3')	Target Gene-ID	Function Description	miRNA target site (5'-3')
mac-miR37	UAAAGCUGCCAGCAUGAUCUGAUCU	Ma02_t19010.1	Tubulin alpha chain (TUBA6)	GGAUUCAGGUCGGCAACGCCUGCUGGGAGCUUUA
mac-miR6	UAGGAGAGAUGACACCGGC	Ma05_t00530.1	Auxin response factor 17 (ARF17)	UGAGAUCAGGCUGGCAGCUUGU
		Ma07_t29010.1	Sugar transport protein 13 (STP13)	GCCGGUGGUCAUCAUCUCCUG
		Ma11_t24110.2	Auxin response factor 6	UGAGAUCAGGCUGGCAGCUUGU
mac-miR-new3	AGUGGUGGAGGGUCGAUGAAGAGG	Ma07_t12970.1	Putative Glucan endo-1,3-beta-glucosidase 14	AGUGGUGGAGGGUCGAUGAAGAGG
		Ma06_t37480.1	Putative Probable F-box protein At1g60180	AGUGGUGGAGGGUCGAUGAAGAGG
mac-miR-new11	AGGGUCGUACCUCUGACGGCGUC	Ma08_t02370.1	CAX-interacting protein 4	AG-GGUCGU-ACCUCUGACGGCGUC
mac-miR-new14	UAGGAAUCUAGGAUGACAAGG	Ma10_t09730.1	Putative Probable WRKY transcription factor 39	UAGGAAUCUAG--GAUGACAAGG
		Ma11_t04080.1	Scarecrow like protein 27, putative, expressed	CGAUUGAGCCGUGCCAAUAUCG

Table 4.6, continued.

Musa-specific miRNA	miRNA sequence (5' – 3')	Target Gene-ID	Function Description	miRNA target site (5'-3')
mac-miR-new20-3p	UUUCCAAUACCUCCCAUGCCAAUGG	Ma05_t11770.1	Putative Disease resistance protein RPS2	UUUCCAAUACCUCCCAUGCCAAUGG
		Ma06_t33710.1	Putative disease resistance protein RGA1	UUUCCAAUACCUCCCAUGCCA---AUGG
		Ma07_t22920.1	Putative Disease resistance protein RGA2	UUUCCAAUACCUCCCAUGCCAAUGG
		Ma08_t34150.1	Putative Pathogenesis-related protein 1	UCGUUGAGUGCAGC-GU---UGAUCA
		Ma11_t01940.1	probable lysine-specific demethylase (ELF6)	GGUAUGGGGCUGGUUUGGAAAAC
mac-miR-new23	UCCAGAGACAUGAUAGCAACACGG	Ma01_t08050.2	ABC transporter G family member 36	UAUGUUUGGAUCUAGAAUGGUAUG
mac-miR-new31	AAGUAUGAUUGUGGCAUGUAUUGG	Ma01_t16810.1	Glutathione S-transferase U17	UCCCUAUGAUGAUCAUCG
		Ma01_t16820.1	Glutathione S-transferase U17	UCCCUAUGAUGAUCAUCG
		Ma11_t19380.1	Aquaporin TIP2-2	UCCCUAUGAUGA--UC-AUCG
mac-miR-new32	AGCUGAGGUAACAACUUUUGAGAA	Ma01_t16810.1	Glutathione S-transferase U17	UCCCUAUGAUGAUCAUCG
		Ma06_t01650.1	Ankyrin repeat-containing protein At5g02620	CCAUGGCUAUAGGUGCCACAAUCAUGCUG
mac-miR-new35	AAACUCAACGAGACCGAGCGGCG	Ma11_t10950.1	NADH dehydrogenase (ubiquinone)	GAUGCUCUCAAAGCUGUUCAGUCUGGACUG
mac-miR-new36	UCAUUCGAUCUGUCCUUUUU	Ma11_t20360.1	Probable protein phosphatase 2C 9	CGCCGUCUCGGUCUGGCCGGGUUU
		Ma06_t14330.1	Putative expressed protein	GAGGAGGACAGAAUCGAAGGA

4.2.5 Network mapping of miRNA and TF in response to salt stress

To understand the interplay between miRNA, transcription factors (TF) and TFBS, a miRNA regulatory network was constructed by making connections between i) miRNA to TF which were their predicted miRNA targets. Among predicted miRNA targets, 26 orthologous miRNA families belonging to mac-miR156, mac-miR160, mac-miR164, mac-miR166, mac-miR171, mac-miR319, mac-miR396 mac-miR529 and mac-miR528 have TF targets (SPL, ARF, NAC, HOX, SCL, MYB, GRF and HSFB1) that are conserved with those reported in other plants. Similarly, two Musa-specific miRNAs have predicted TF targets i.e. mac-miR-new14 targets WRKY21, mac-miR-new20 targets ELF6 transcription factors. ii) miRNA with predicted TF targets have respective TFBS for example, mac-miR156 target SPL have a SPL binding site (TNCGTACAA motif) in the miRNA promoter. Consensus logos were generated for TF target specific binding sites within the promoter regions of miRNA (Table 4.7). The resulting network (Figure 4.6) using miRNA, TF and respective TFBS suggests interaction between 20 TFs targeted by 26 orthologous miRNAs and 2 Musa-specific miRNAs.
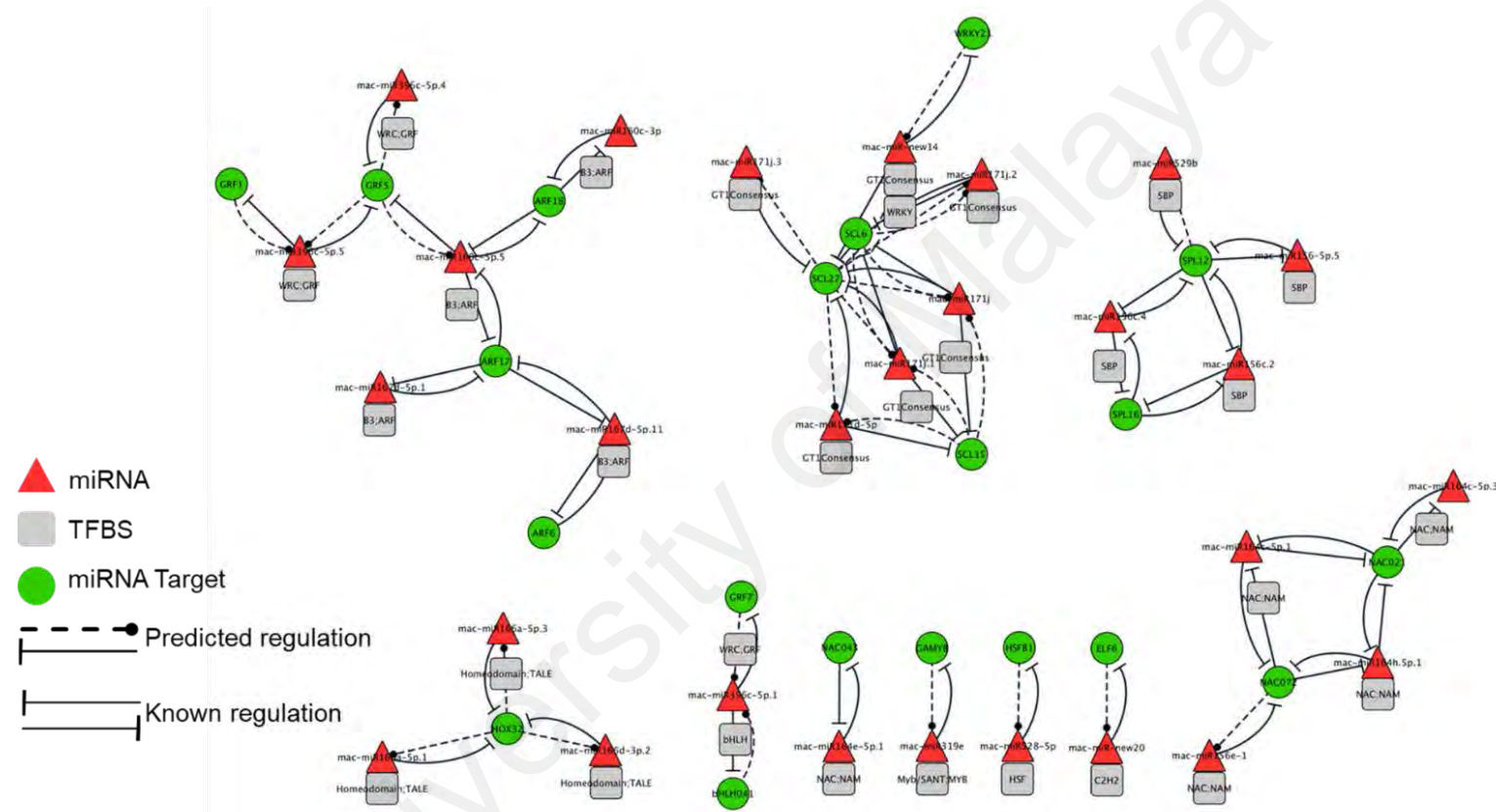
Figure 4.6: Regulatory circuits involving miRNA, TFBS in miRNA promoters and miRNA-targeted transcription factors. miRNA mediated regulatory network of target transcription factors (TF) and respective transcription factor binding sites (TFBS). Solid lines with horizontal bars indicate repression of miRNA targets predicted from our study and supported by reports in the literature for various plant species (Known regulation) and dotted lines with circles indicate predicted regulation only from the data in the current study.

Table 4.7:miRNA and miRNA TF target specific TFBS motifs. miRNA TF targets are identified from degradome analysis.

miRNA Family	TF-Target	TFBS	PlantPANID[a]	TFBS Logo
mac-miR156	SPL12, SPL16	SBP	TF_motif_seq_0508	
mac-miR171	SCL15, SCL27, SCL6	GT1Consensus	TF_motif_seq_0321	
mac-miR-new14	WRKY21	WRKY	TF_motif_seq_0270	

[a] PlantPanID represents the database accession number and TFBS logos generated by MEME Suite based on the motif sequences from banana for all members within a miRNA family.

Table 4.7, continued.

mac-miR160	ARF18, ARF17, ARF6	B3; ARF	TF_motif_seq_0335	
mac-miR164	NAC021, NAC072, NAC043	NAC; NAM	TFmatrixID_ 0386	
mac-miR166	HOX32	Homeodomain; TALE	TF_motif_seq_0246	
mac-miR319	GAMYB	Myb/SANT; Myb	TF_motif_seq_0376	

Table 4.7, continued.

mac-miR396	GRF7, GRF5, GRF1	WRC; GRF	TFmatrixID_ 0441	
mac-miR396	bHLH041	bHLH	TF_motif_seq_0300	
mac-miR528	HSFB1	HSF	TF_motif_seq_0010	
mac-miR-new20	ELF6	C2H2	TFmatrixID_ 0211	

4.3 Association of DNA methylation with expression of genes and siRNA in salinity-stressed banana roots

4.3.1 DNA and RNA extraction

Total RNA was extracted from roots of plants grown in control and NaCl supplemented media using TRIzol (Invitrogen, USA) (Figure 4.7). Quantitative analysis of the RNA samples performed by Agilent 2100 Bioanalyzer, showed the RNA to be of suitable quality for high-throughput sequencing (Figure 4.8).
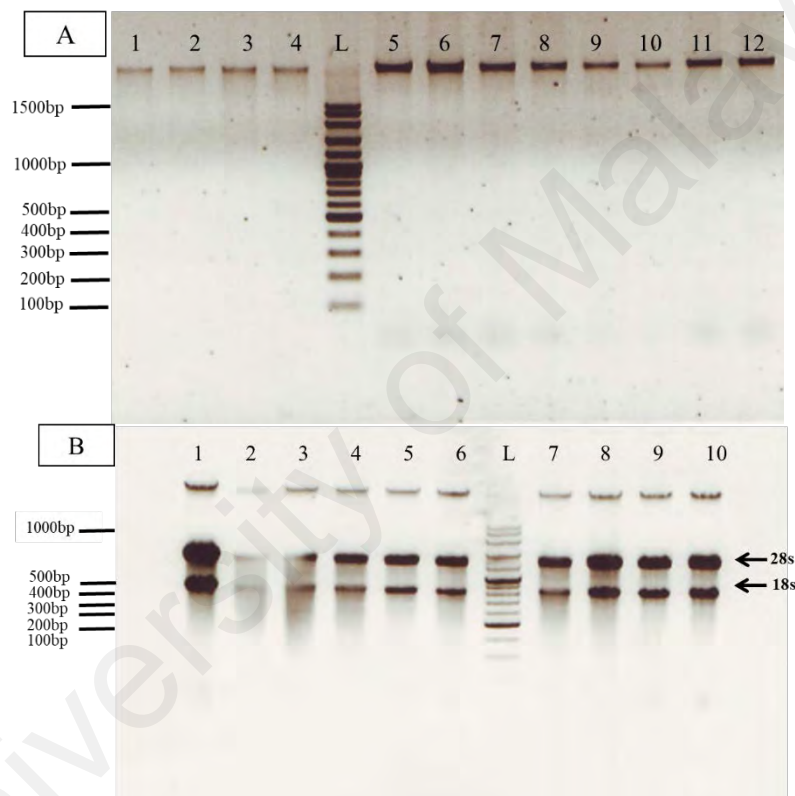
Figure 4.7: DNA and RNA extraction gel electrophoresis. A) Lanes 1 ,2, 5-8: DNA from untreated control replicates. Lanes 3,4 ,9-12: DNA from plantlets after treatment with 300mM NaCl. B) Lanes 1 to 6: RNA from untreated control replicates. Lanes 8 to 11: RNA from plantlets after treatment with 300mM NaCl. Lane L: 100bp plus DNA Marker (Fermentas). 4µl of DNA/RNA loaded on 10 % gel for all samples.

Control(UT)   300mM NaCl treated (ST)
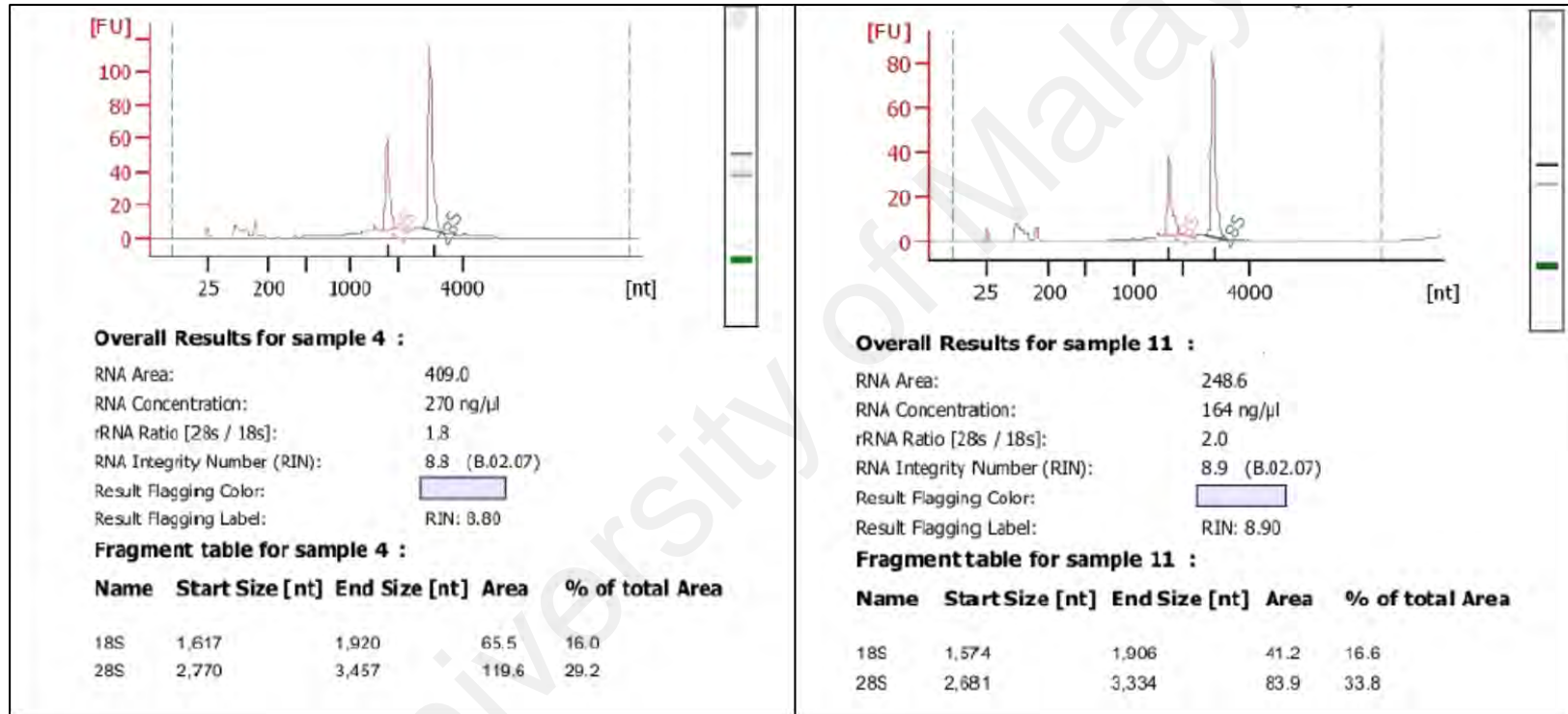
Figure 4.8: Agilent 2100 Bioanalyzer result for RNA quantification. Agilent 2100 Bioanalyzer result for the samples chosen for small RNA sequencing untreated Control (UT) sample and 300mM NaCl treated (ST) samples.

Analysis of small RNA (sRNA) datasets from banana root tissues subjected to salt stress (sample "ST" ~28 million reads) and non-salt stress (sample "UT" ~15 million reads) (Table S8) showed a difference in size distribution with 21 nucleotide (21nt) sRNA showing relatively equal expression in both conditions, whilst there were fewer 24 nucleotide (24nt) sRNA following salt treatment (Figure 4.9). The 24nt class is of particular interest in this study as these are reported to be associated with methylation related epigenetic changes in plants resulting in gene silencing (known as Transcriptional Gene Silencing or "TGS"). In the current data sets 24nt sRNA comprise 23.07% and 18.85% of sRNA in non-salt treated and salt-treated samples, respectively (Table S8).
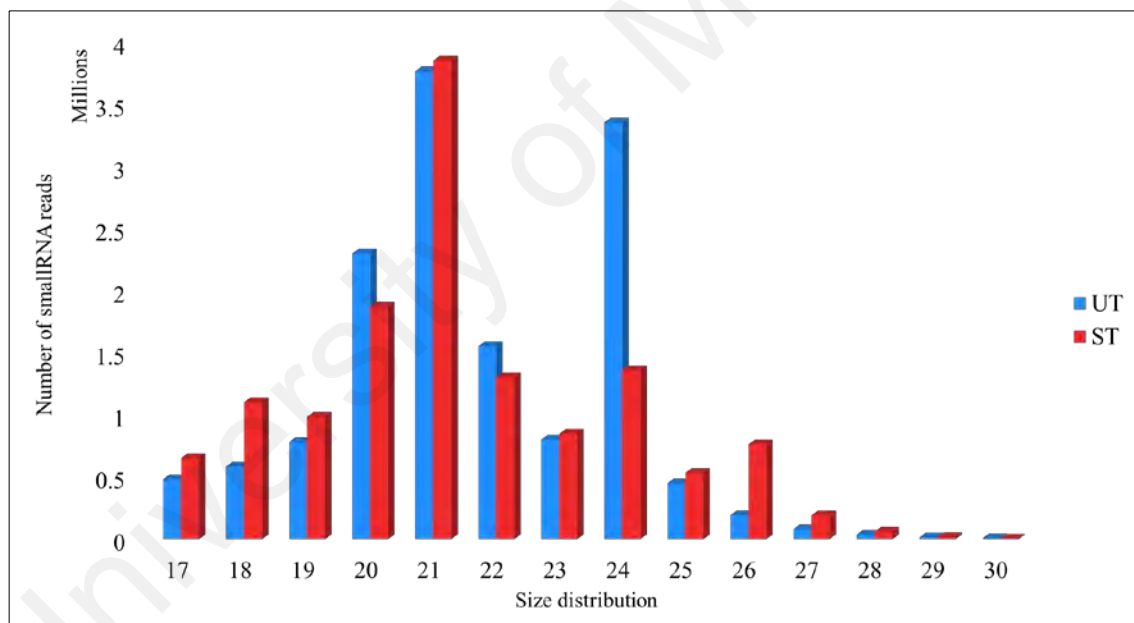
Figure 4.9: Small RNA size distribution. Data distribution is based on small RNA transcriptome with two replicates. UT: 0mM NaCl treatment (Non-salt stress) ST: 300mM NaCl treatment (Salt stress).

4.3.2   Genome wide DNA methylation changes following salt stress in banana

To observe genome-wide DNA methylation changes caused by salt stress in banana roots, we adopted whole genome bisulphite sequencing of root tissues from in vitro plantlets exposed to salt stress (ST) (300ml NaCl) and control, non-salt-stress conditions (UT). A total, 216,286,990 and 216,016,372 paired end bisulphite reads from single data set (n=1) were sequenced from non-salt stress and salt-stress exposed banana roots, respectively (Table S9). Around 19 gigabases from non-salt stress (UT) and salt stress (ST) methylome datasets were mapped to the reference genome of *Musa acuminata* DH pahang genome version-2. More than 70% of the reads from salt stress and non-salt stress datasets mapped to the genome sequence. Only uniquely mapped reads (60.3% for salt stress and 62.3% for non-salt stress) were considered for cytosine methylation analysis at single-base resolution (Table S10). Overall, there were more number of methylated cytosines in salt stress samples, i.e. 8.22% compared to 7.83% for control samples (Table S11). Distribution of methylated cytosine levels across the chromosomes showed positive correlation with repeat associated loci on all chromosomes (Figure 4.10). Distribution of methylated cytosines levels across genomic regions, show repeats and introns are relatively highly methylated in both non-salt stress and salt stress compared to other genic regions i.e. exons, upstream 2kb of gene and downstream 2kb of gene (Figure 4.11) .
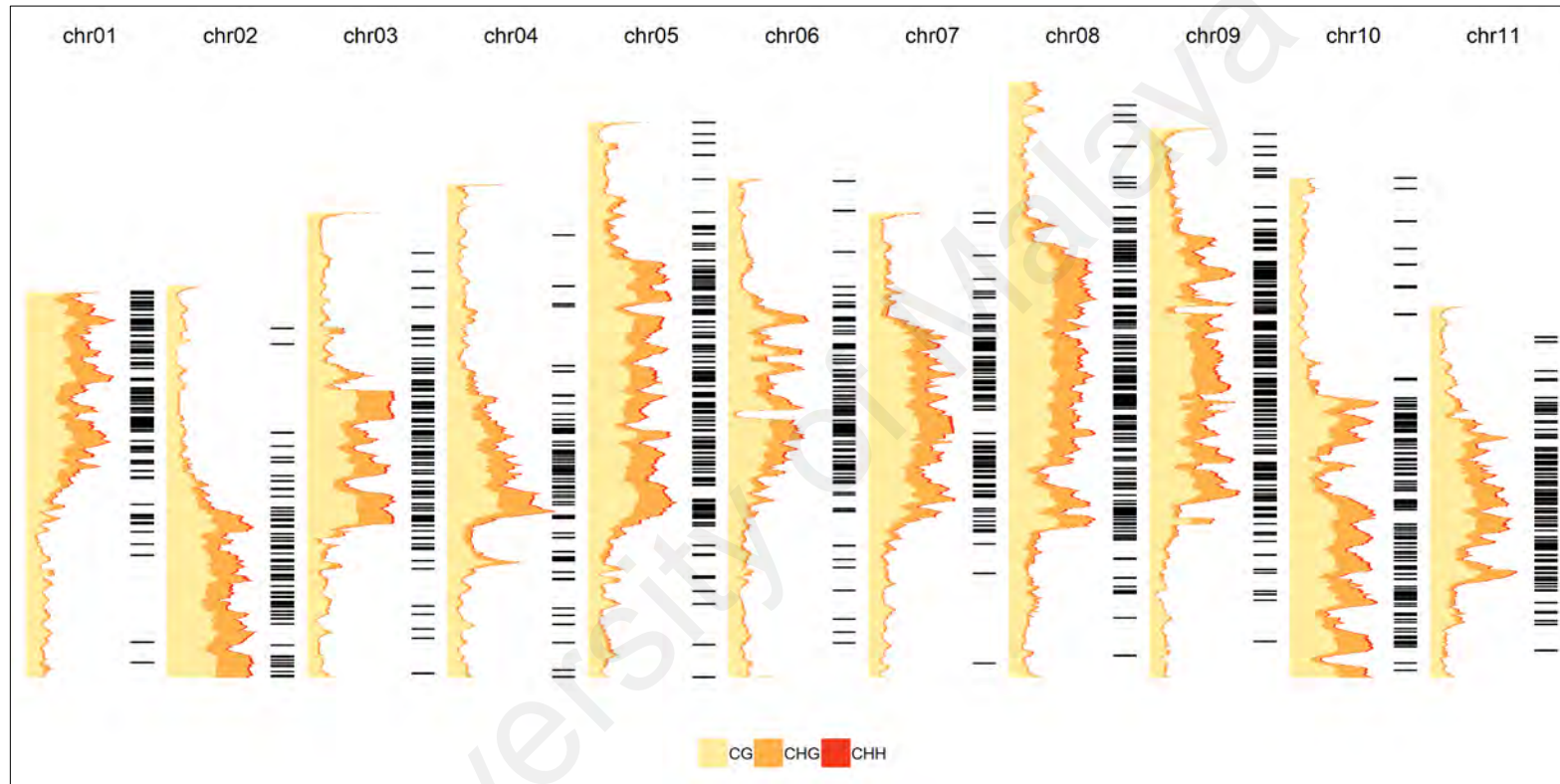
Figure 4.10: Chromosomal overview of DNA methylome. Black horizontal bars: repeat localization; Yellow: CG methylation; Orange: CHG methylation; Red: CHH methylation. Analysis used a sliding window of 500kb with100kb step size. X-axis represents average methylation ratio (methylated cytosine/total number of cytosine within the mapped reads across each window) and Y-axis represents chromosomal positions. Methylation ratios was based on a single data set (n = 1).
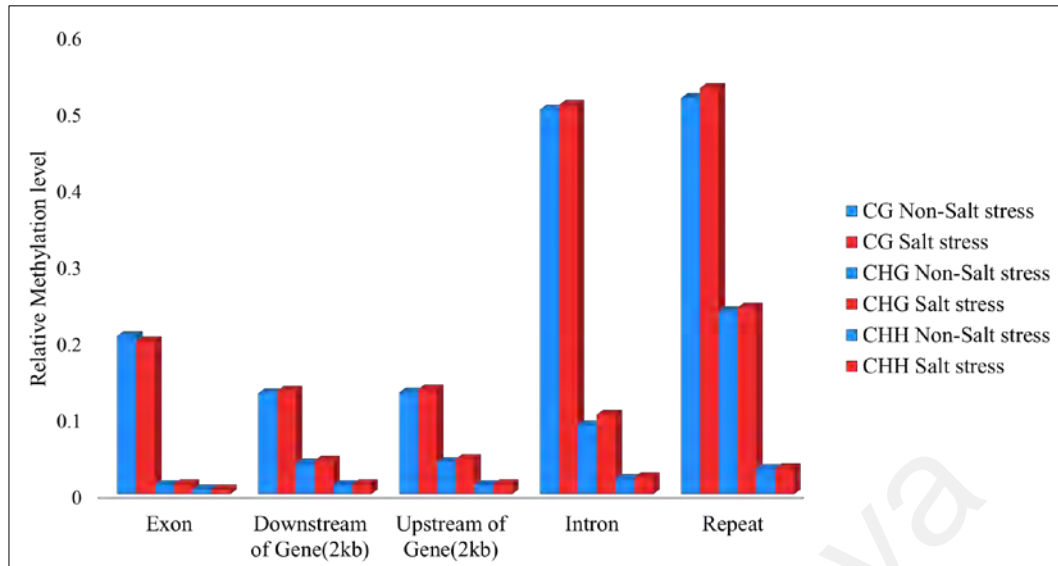
Figure 4.11: Average Methylation level across genomic regions. X-axis shows the genome regions and Y-axis shows the average methylation ratio (methylated cytosine divided by total reads mapped to exons, introns, Upstream 2kb region of gene, Downstream 2kb of the gene and repeat regions). Methylation ratios was based on a single data set (n = 1).

4.3.3   21nt and 24nt siRNA guided methylation during salt stress

Small RNA sequencing from non-salt stress (0 mM NaCl) and salt stress (300 mM NaCl) resulted in 14.4 million and 13.5 million clean reads with slightly higher levels of 21nt and 24nt small RNA observed in samples from control roots (Table S8). Over 60% of 21nt and 24nt small RNA reads mapped to version-2 genome of *Musa acuminata* DH pahang genome. siRNA clusters were created by merging overlapping reads with 100 bp region of the genome for 21nt and 24nt small RNA, two separate clusters were built for 21nt and 24nt siRNA to identify its association with methylation in CG, CHG and CHH contexts. After merging 21nt siRNA, 21,090 and 12,886 clusters were obtained on the genome from non-salt stress and salt stress small RNA datasets, respectively (Table 4.8). Similarly, for 24nt siRNA clusters of 86,788 and 67,568 were built using non-salt stress and salt stress small RNA datasets. Filtering of siRNA clusters with > 10 reads results in 12,052 and 62,304 siRNA clusters for 21nt and 24nt, respectively (Table 4.8). Differential

expression between non-salt stress and salt stress siRNA clusters based on abundance in each cluster resulted in 489 clusters and 889 clusters, respectively.

Table 4.8: siRNA clustering statistics. Statistics are based on small RNA transcriptome with two replicates.

	21nt siRNA		24nt siRNA	
	Non-salt stress	**salt stress**	**Non-salt stress**	**salt stress**
Clusters (100bp merge)	21,090	12,886	86,788	67,568
Filtered clusters (>10 reads)	12,052		62,304	
Average small RNA abundance	271.51	283.94	59.46	51.33
Differentially expressed clusters (FDR[a] <0.05)	488		889	

[a]False discovery rate (FDR) determines adjusted p-values for each test and controls number of false discoveries in those tests to attain significant results.

To observe association between 21nt and 24nt siRNA abundance with methylation patterns, differentially expressed siRNA clusters were overlapped with methylated cytosines levels from salt stress and non-salt stress methylomes. Both upregulated and downregulated 21nt and 24nt siRNA clusters showed higher cytosine methylations in CG, CHG and CHH contexts in the methylome of salt stressed roots compared to those of the control (Figure 4.12). In non-salt stress sample, differentially expressed 21nt and 24nt siRNA clusters show higher methylation levels only in the CHH context. (Figure 4.12). Distribution of differentially expressed 21nt and 24nt siRNA clusters over genomic regions show 43% and 69% of siRNA clusters overlap with intergenic regions, respectively (Figure 4.13). Among genic regions, 10.37% of 21nt siRNA clusters overlap with exons while 24nt siRNA show only 3.62% overlaps (Table S13).
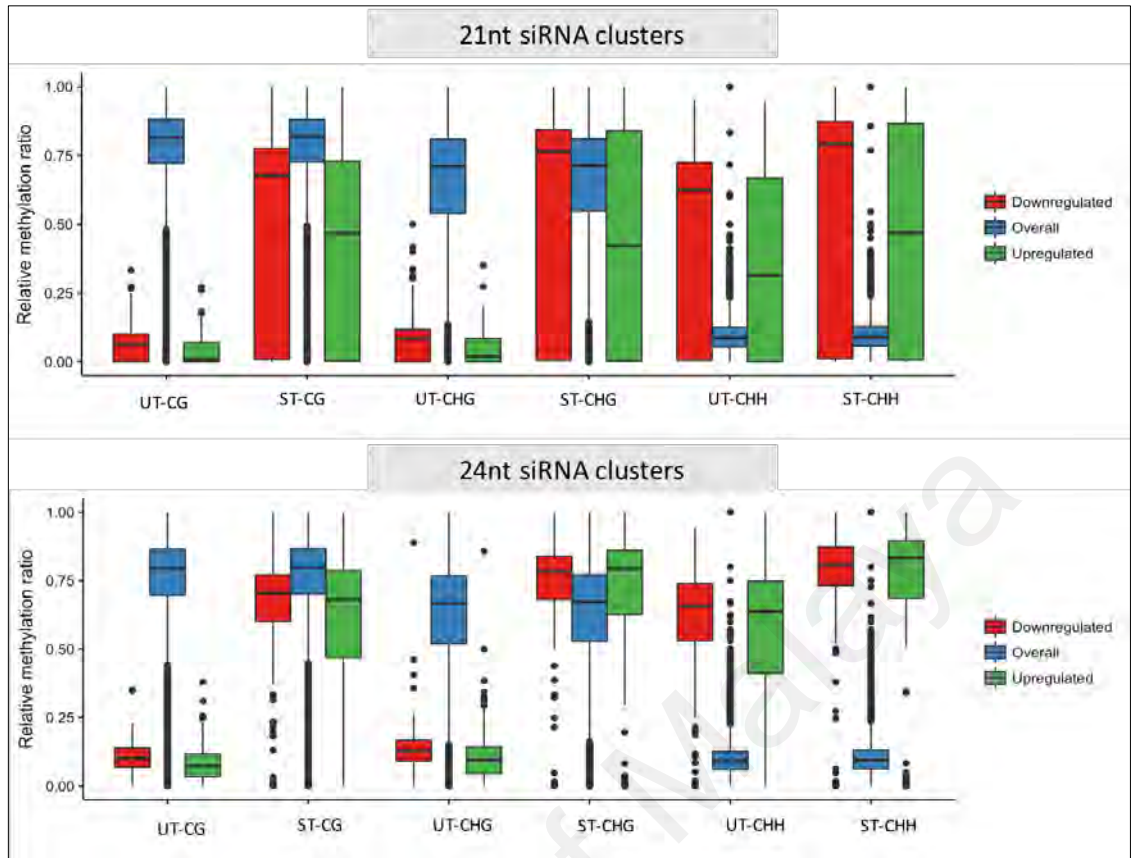
Figure 4.12: Association between siRNA clusters and overlapping methylation coverage on genomic loci. The relative methylation ratio is total number of methylated cytosine divided by total number of cytosines with the reads mapping to an siRNA cluster locus. Methylation ratios were calculated based on a single data set (n = 1). Boxplot representation of methylation ratio on differentially expressed siRNA clusters on banana genome between salt stress (ST) and control (UT). Methylation ratios were plotted for upregulated (Green) and downregulated (Red) siRNA clusters along with overall methylation ratios among all siRNA clusters (Blue) were arranged in order of UT-CG, ST-CG, UT-CHG, ST-CHG, UT-CHH and ST-CHH where UT is Non-salt stress and ST is salt stress.
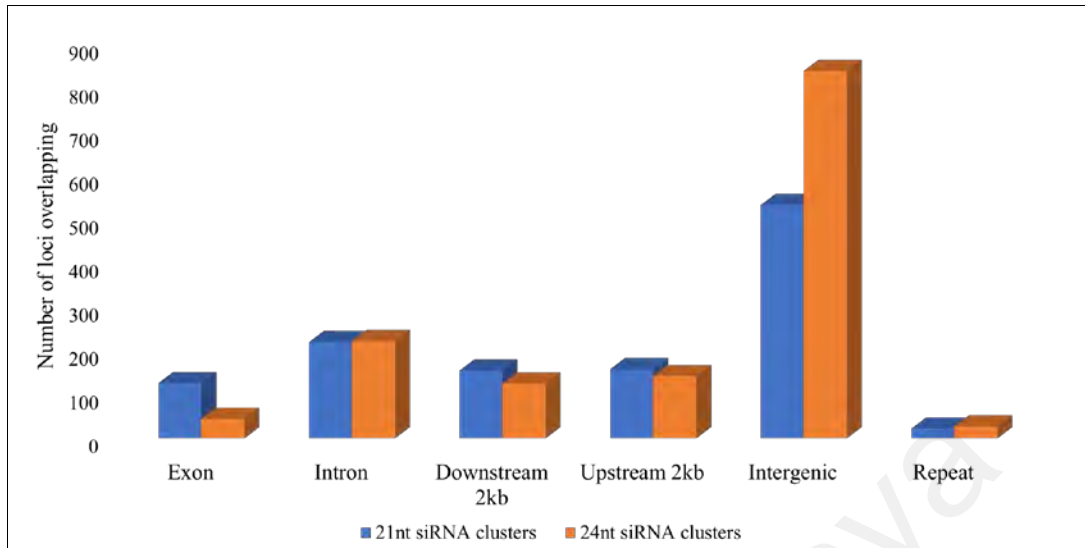
Figure 4.13: Distribution of siRNA clusters across genomic regions. Differentially expressed siRNA clusters (21 and 24nt siRNA) overlapping with genomic regions.

4.3.4 Differentially methylated regions (DMR) and gene expression responding to salt stress

To identify stress related differences in mRNA expression associated with different levels of methylation, transcriptome data from previous study (Lee et al., 2015) on salt stress and non-salt stress samples was used (Table S14). Transcriptome reads from both the samples were mapped to genome which both datasets show > 78% mapping to the version-2 of banana genome.(Table S15). Uniquely mapped reads to genome were considered for differential expression analysis, 1,076 genes were observed to be differentially expressed between salt-stress and non-salt stress conditions. Differentially methylated regions (DMR) are genomic loci with hyper- or hypomethylation in salt stress relative to non-salt stress methylome. DMRs were identified from single-base resolution mappings and pre-filtered with methylated reads $\geq$ 10 and DMR loci with minimum length of 100bp were considered using binomial distribution method. DMR prediction in detail is explained in the methods section 3.6.6. A total of 11,162 , 3,375 and 13,357 DMR on genome were identified between salt stress and non-salt stress methylomes in CG, CHG and CHH contexts (Figure 4.14A). Most of the CG DMRs (91.8%) are

hypermethylated, almost half of CHG DMRs (53.3%) are hyper methylated in salt stress methylome. Interestingly, 89.2% of CHH DMRs are hypomethylated in salt stress. DMR loci distribution among genic regions on genome shows exons, introns and intergenic regions are hypermethylated in CG context and hypomethylated in CHH context (Figure 4.14B). To identify regulation of gene expression based on DMR loci, we considered upstream 2kb , downstream 2kb and intronic regions of the genes on the genome. DMR loci to gene association revealed genes related to biological processes such as transcription, nucleic acid binding, transported activity and response to stress (Figure S8). An overview of DMR loci association with 1,076 differentially expressed genes (significant difference between levels expressed in salt-stress and in non-salt stress) and non-differentially expressed genes indicates a role of methylation in regulating gene expression. Positive correlation is observed between gene upregulation in salt stress and presence of DMR in genic loci (Figure 4.15). Interestingly, 256 differentially expressed genes were associated with DMR at their genic regions, i.e. gene upstream 2kb, intronic or gene downstream 2kb regions. (Appendix E Table1) (refer to section 3.6.9 for data source). Genes upregulated in salt stress are majorly associated with hypermethylation and hypomethylation in CG and CHH contexts, respectively . Among the DMR associated genes, transcription factors such as MYB (Ma00_g04960, Ma06_g11270, Ma01_g11890 and Ma06_g11270), heat stress transcription (Ma05_g18920), WRKY71(Ma10_g03640), ERF (Ma09_g12570) are among those upregulated in salt stress which are associated with hypermethylation in CG context. Hypermethylation in CG context is also associated with stress responsive genes such as Cation/H(+) antiporter 20 (Ma05_g06950) and Calmodulin (Ma05_g08710) which are also upregulated in salt stress condition.
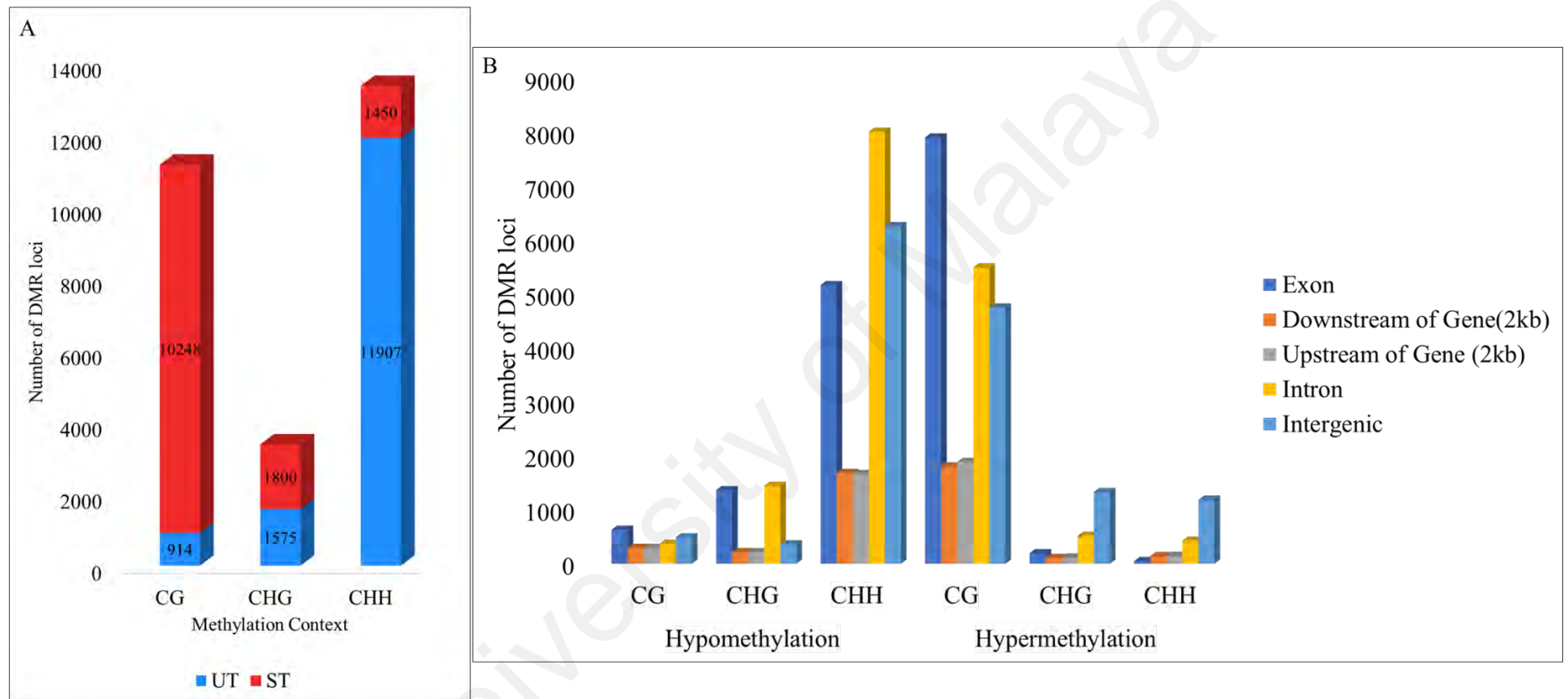
Figure 4.14: Distribution of differentially methylated regions (DMR) across genomic regions. A) Distribution of differentially methylated regions (DMR) loci on banana genome in CG, CHG and CHH methylation contexts in salt stress (ST) and non-salt stress (UT) methylomes. Bars represent the number of DMR loci overlapping regions on the genome i.e. Exon, Gene Downstream (2kb) and Gene Upstream (2kb), Intron and Intergenic in CG, CHG and CHH contexts. Calling DMR regions was based on a single data set (n = 1).
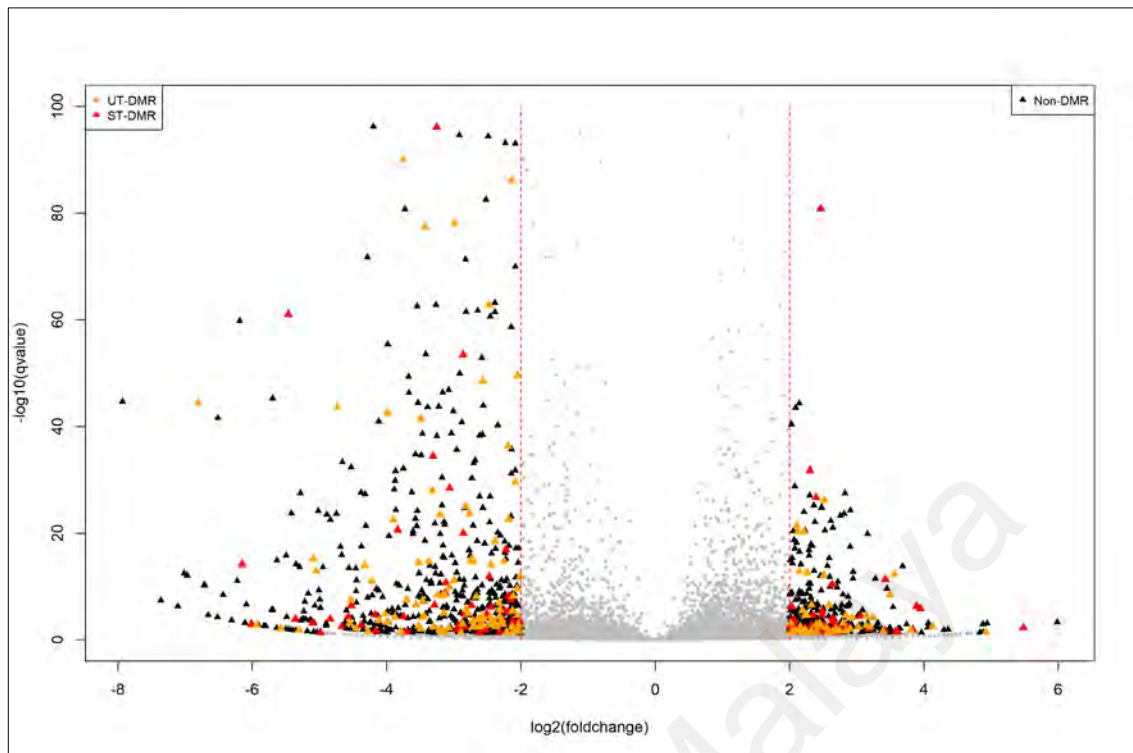
Figure 4.15: Differentially methylated regions (DMR) and association with gene expression. Volcano plot showing all the gene expression in salt stress (ST) and non-salt stress (UT) transcriptomes. Genes with DMR association (2kb –upstream, intronic and 2kb-downstream regions) are highlighted in the volcano plot, i.e. Red colour for salt stress and orange for non-salt stress. X-axis represents log2 foldchange (non-salt stress/salt stress). Y-axis represents -log10 of q-value or FDR. Red dotted lines represent the cut-off of log2foldchange > 2 or < -2. Differentially expression was calculated based on a single data set (n = 1).

4.3.5   Association between 24nt siRNA clusters and DNA methylation

Among 256 differentially expressed genes associated with DMR loci in upstream 2kb, intronic or downstream 2kb regions, 78 genes overlap with 24nt siRNA clusters (Appendix E Table2) (refer to section 3.6.9 for data source). At DMR overlap regions 24nt siRNA abundance show no linear correlation with hypermethylation of CG,CHG and CHH contexts in salt stress condition (Figure S9). Strikingly, eleven genes (14.1%, 11/78) show *de novo* methylation sites where hypermethylation is observed in CG context relative to with hypomethylation in CHH context (Figure 4.16). Among eleven genes include salt stress responsive genes such as DRE2 (Ma04_g30480), Cation/H antiporter 20 (Ma05_g06950), UDP-glucose 4-epimerase (Ma05_g08910), serine/threonine-protein

kinase (Ma05_g24750) and pyruvate decarboxylase isozyme 2 (Ma11_g23090) (Figure 4.16A-E). Similarly, 14 genes (17.9%, 14/78) show CG *de novo* methylations i.e. hypermethylation only in CG context which includes floral homeotic protein APETALA 2 (Ma09_g04220) and Putative Dehydrin COR410 (Ma11_g15310) (Figure 4.16F & G). Only five genes (6.4%, 5/78) were associated with hypermethylation in CHH context which includes peroxidase 47 (Ma01_g21520), peroxidase (Ma05_g15710) (Figure 4.16H) , beta-D-xylosidase 4-like (Ma05_g25760) , queuine tRNA-ribosyltransferase-like (Ma10_g24770) and uncharacterized LOC103991602 (Ma07_g13600). Lastly, hypermethylation were also identified in CHG context among six genes (7.7%, 6/78) which includes molybdate transporter 1-like (Ma03_g28120), chromosome-associated kinesin KIF4A (Ma01_g00510), UDP-glucose 4-epimerase GEPI48 I (Ma05_g08910), calcium permeable stress-gated cation channel (Ma01_g05470), probable protein phosphatase 2C (Ma01_g01270).
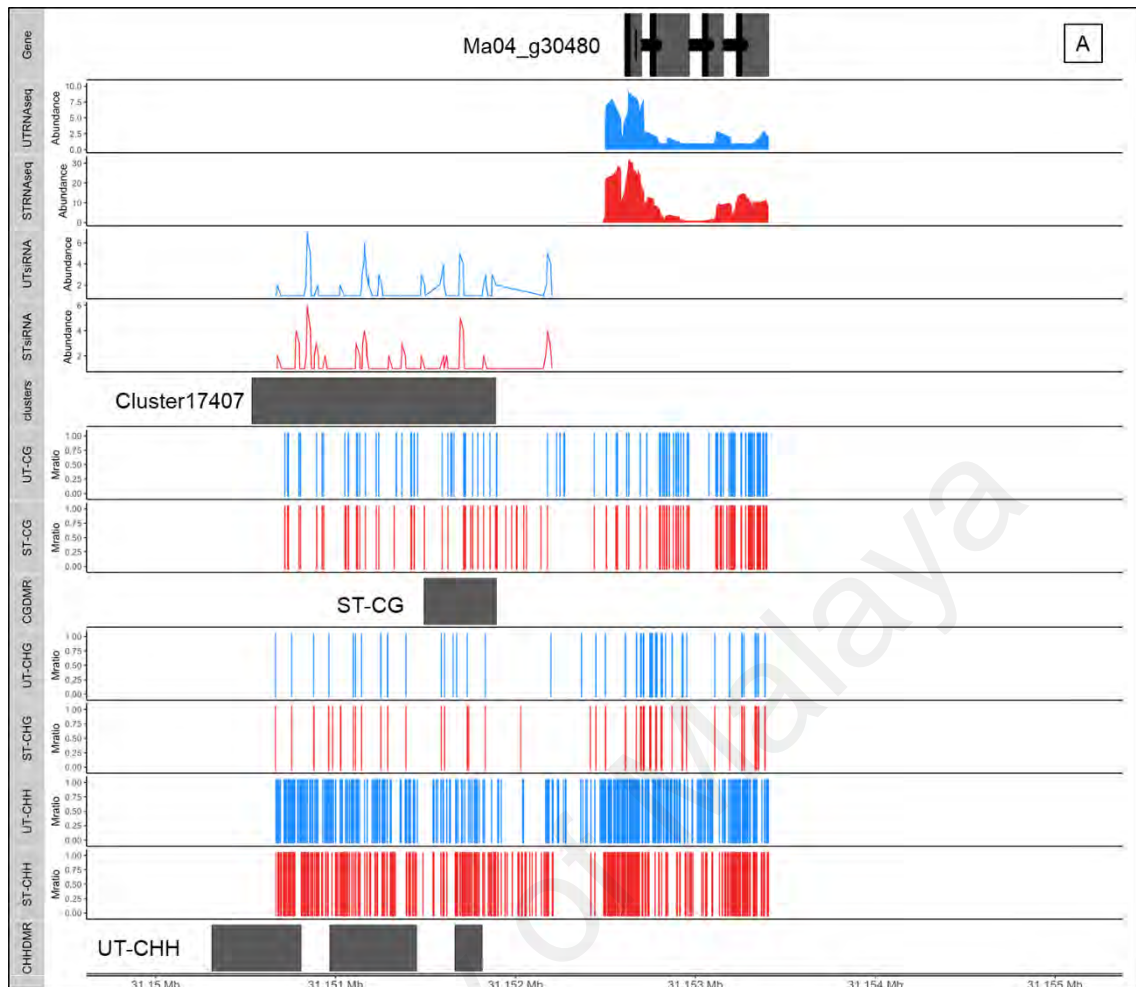
Figure 4.16: Genome browser view of overlapping DMR, siRNA adjacent to differentially expressed genes. Musa acuminata V2 genome landscapes of the differentially expressed genes which overlap with DMR loci and 24nt siRNA clusters. Track annotations are: Gene : Musa acuminata v2 genome genes; UT: Non-salt stress datasets; ST-Salt stress datasets UT-RNA-seq /ST-RNA-seq: gene coverage from transcriptomes ; UT-24nt-sRNA/ST-24nt-sRNA: 24nt small RNA coverage on genome; 24nt siRNA clusters: siRNA cluster loci on genome (refer to methods section:3.6.5); UT-CG/ST-CG/UT-CHG/ST-CHG/UT-CHH/ST-CHH: Methylation levels across the genome regions for all contexts ; CG-DMR/CHG-DMR/CHH-DMR. Differentially methylated regions for all contexts, RNA-seq, small RNA and DNA methylation data shown in the browser based on a single data set (n = 1). A) DRE2 (Ma04_g30480). B) UDP-glucose 4-epimerase (Ma05_g08910). C) Cation/H antiporter 20 (Ma05_g06950). D) Serine/threonine-protein kinase (Ma05_g24750). E) Pyruvate decarboxylase isozyme 2 (Ma11_g23090). F) AP2 (Ma09_g04220). G) Putative Dehydrin COR410 (Ma11_g15310). H) Peroxidase (Ma05_g15710).
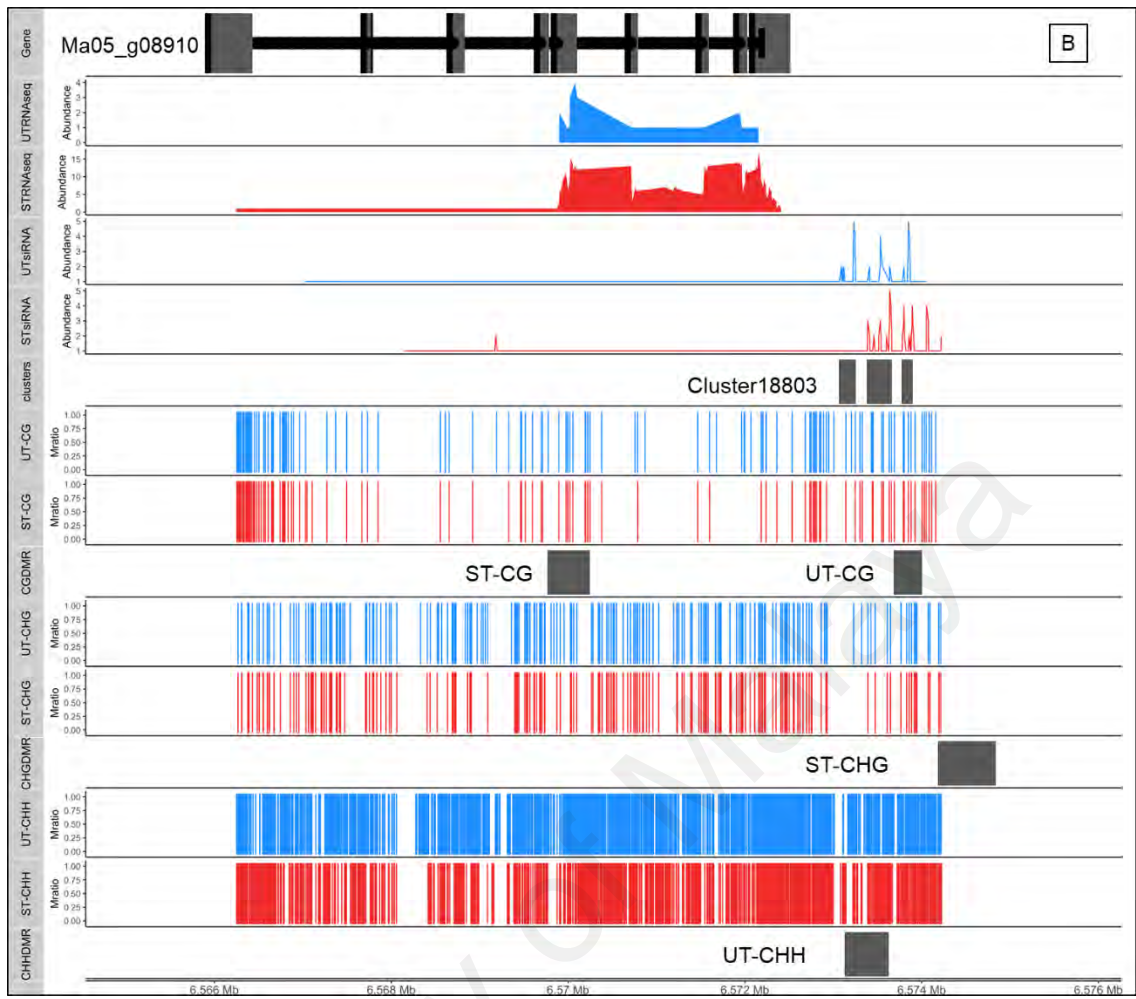
Figure 4.16, continued.
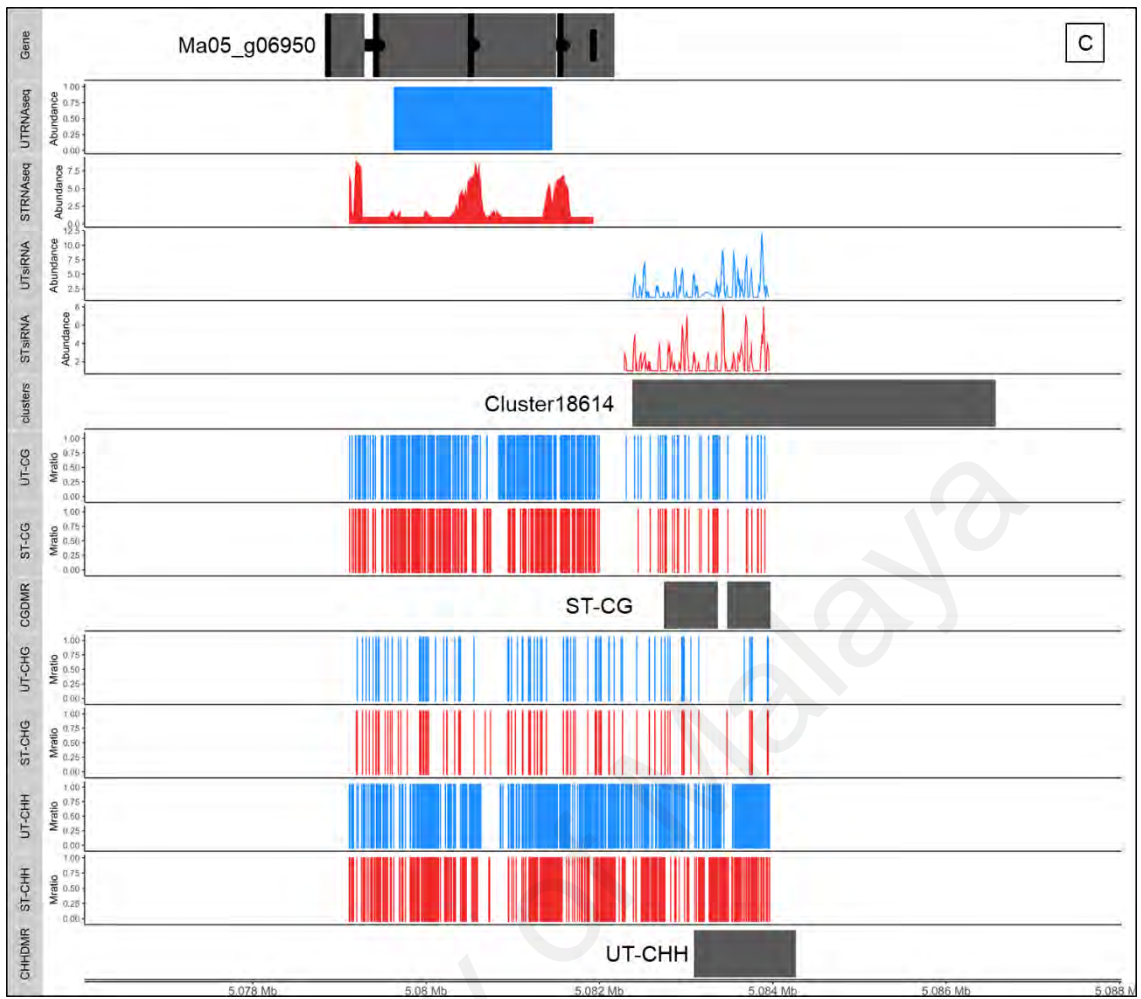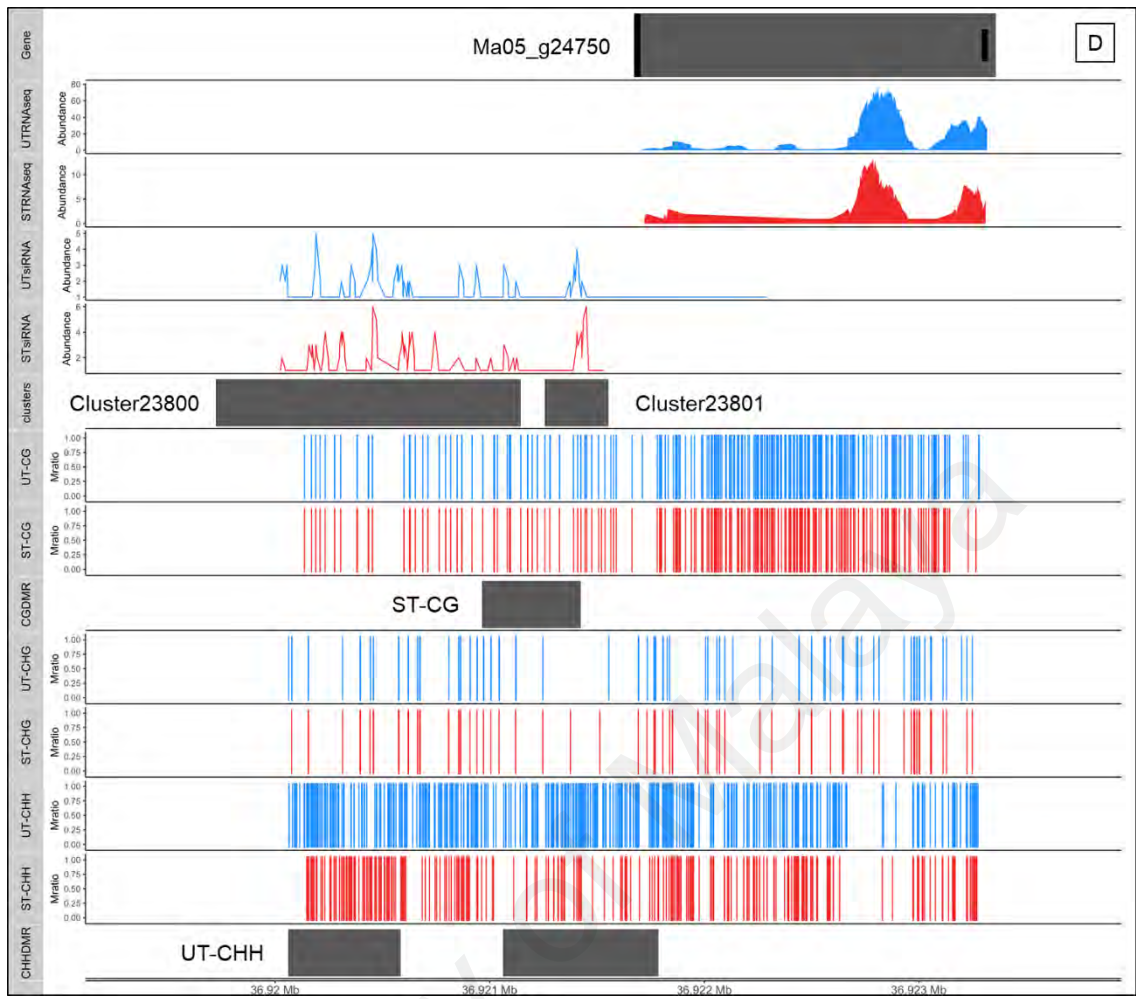
Figure 4.16, continued.

Figure 4.16, continued.
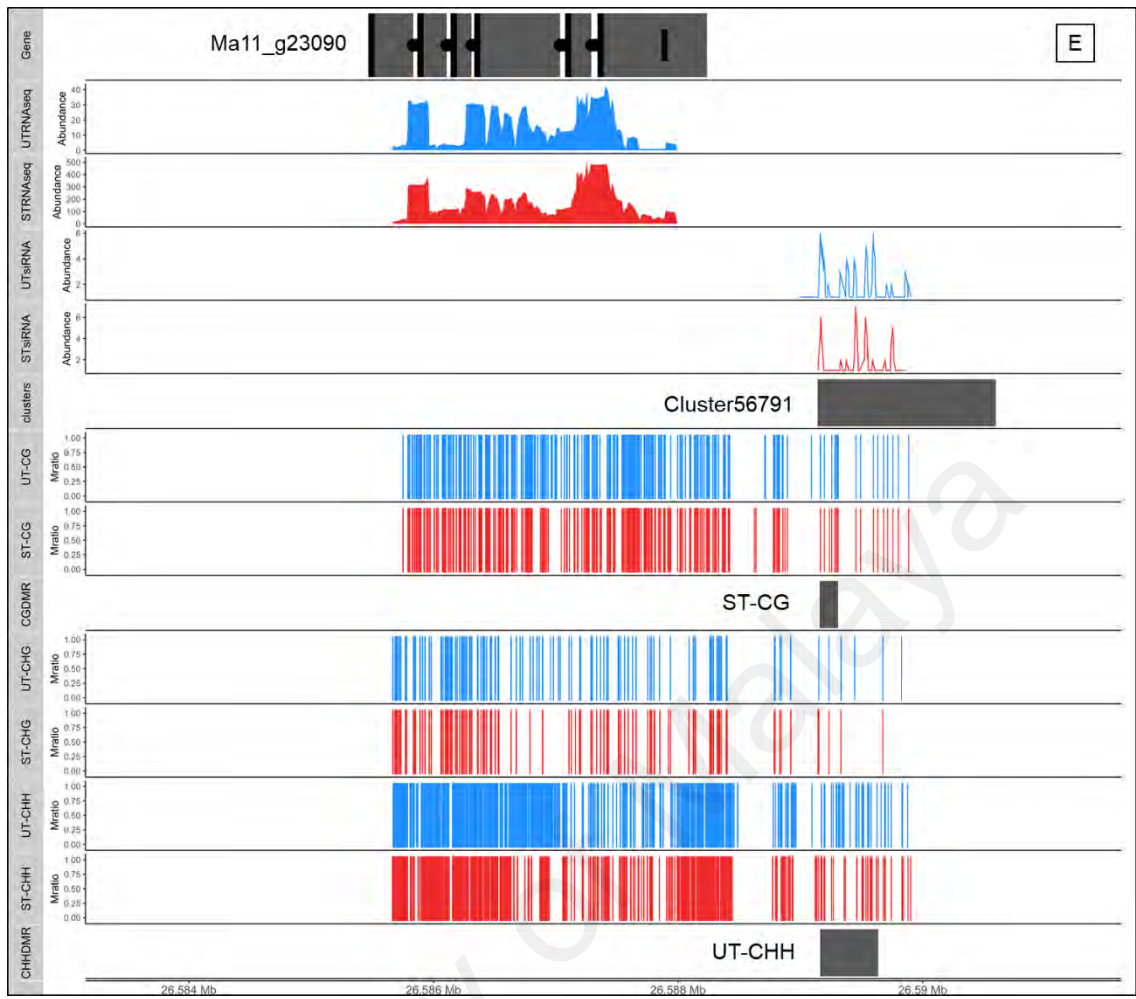
Figure 4.16, continued.
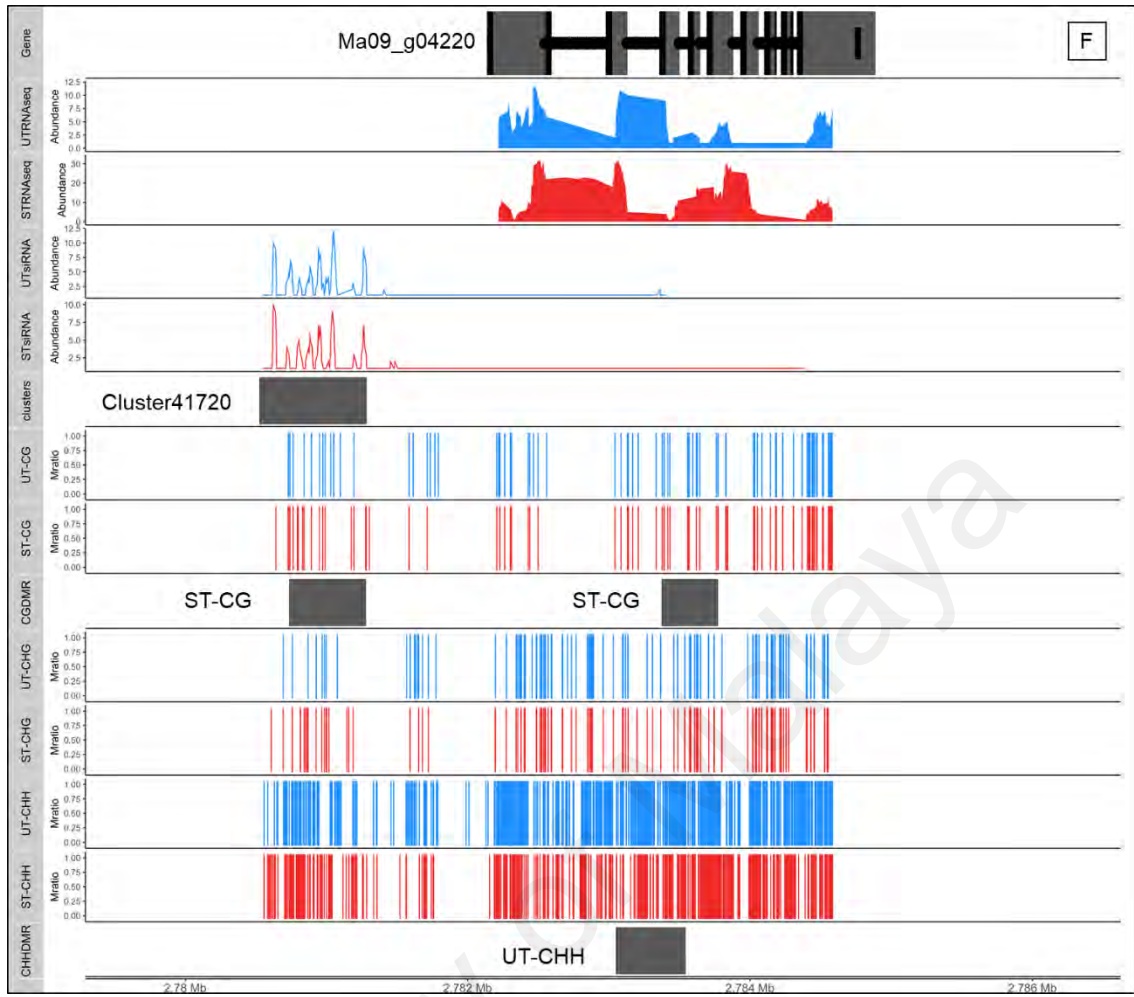
Figure 4.16, continued.
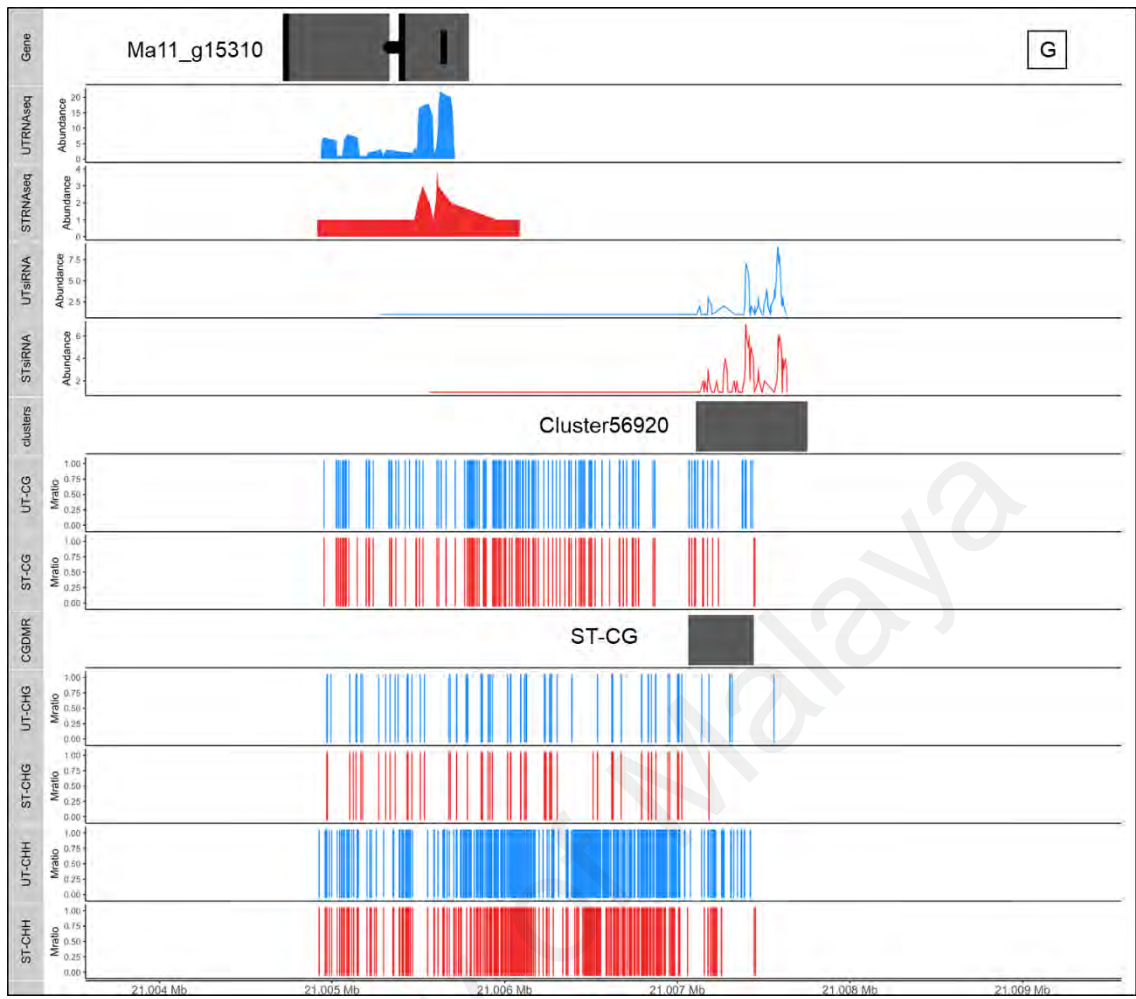
Figure 4.16, continued.
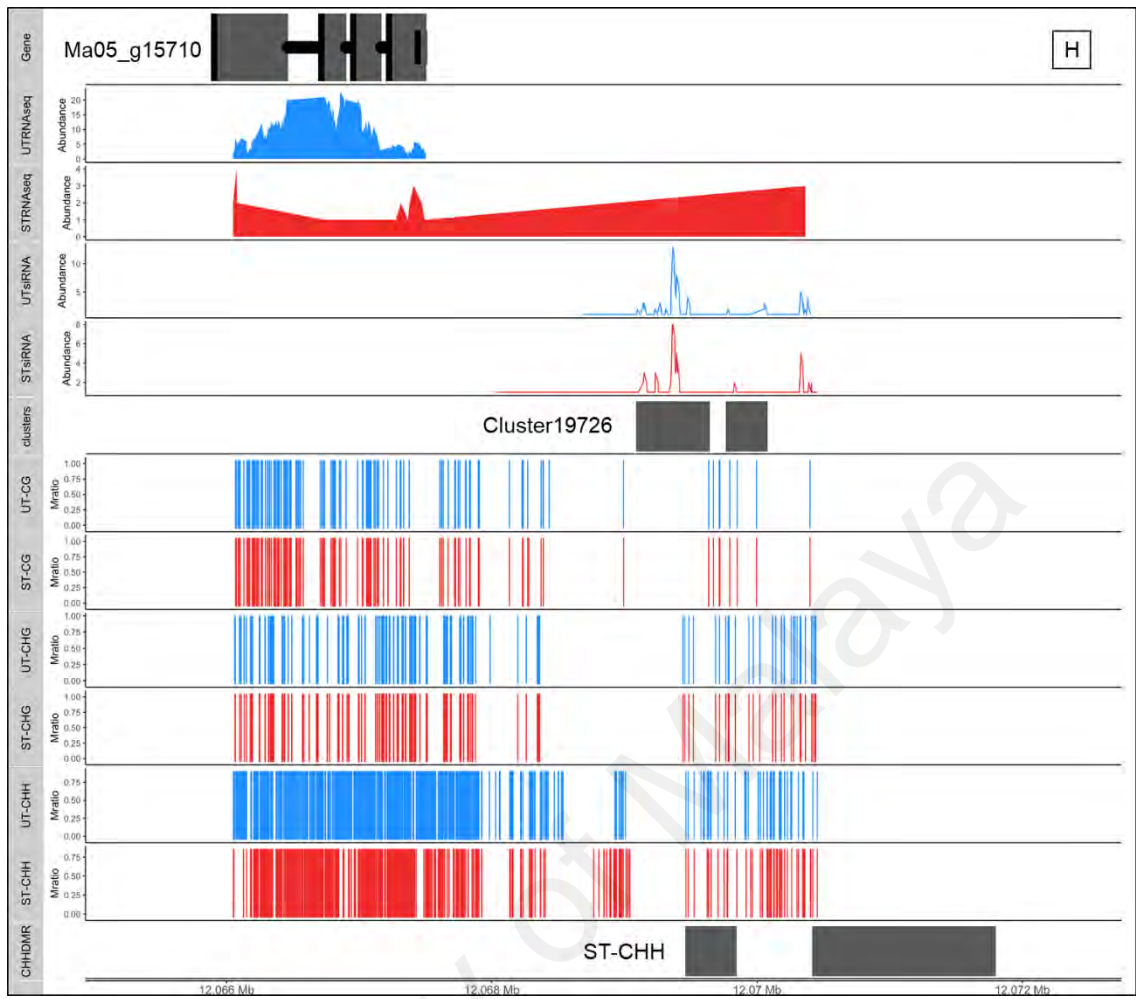
Figure 4.16, continued.

4.3.6 Repeat associated methylation changes associated with salt stress

The *Musa acuminata* genome comprises of ~55% of repetitive content (D'Hont et al., 2012; Hribova et al., 2007) which is mostly made up of Maximus/SIRE and Angela lineages of Ty1/copia long terminal repeat(LTR) and Ty3/gypsy elements (14 to 34.5% of the genome) (Novak et al., 2014). Classification of banana transposable elements (TE) includes 1,902 TEs from Hribova et al. (2010) was used to annotate repeat loci on the banana version-2 genome. Genome wide differentially methylated regions (DMR) were used to study methylation profiling on the repeat associated loci. Profiling across repeat loci on banana genome reveal overlapping of hypermethylated loci in all three contexts CG (218) ,CHG (55) and CHH (281). Majority of Long tearminal repeat (LTR) loci are associated with hypermethylation in CG and hypomethylation in CHH contexts. More than 34% of CG hypermethylated loci are associated with copia/SIRE1Maximus followed by Gypsy(9.17%), Gypsy/Renia (5.5%) and Gypsy/Tekay (5.96%). Satellite repeats, i.e. 45srRNA, 5srRNA and clDNA uniquely show CG hypermethylation (Figure 4.17A). Hypermethylation in CHG context show a similar trend of distribution among LTR elements i.e. >25% in Copia elements and >14% in Gypsy elements. (Figure 4.17B). Hypomethylation in CHH context is observed across all the repeat elements except 5SrRNA (Figure 4.17C). Hypomethylation in CHH context might be demethylation of LTR elements and other centromeric repeats such as Long ,interspersed elements (LINE) due to salt stress.
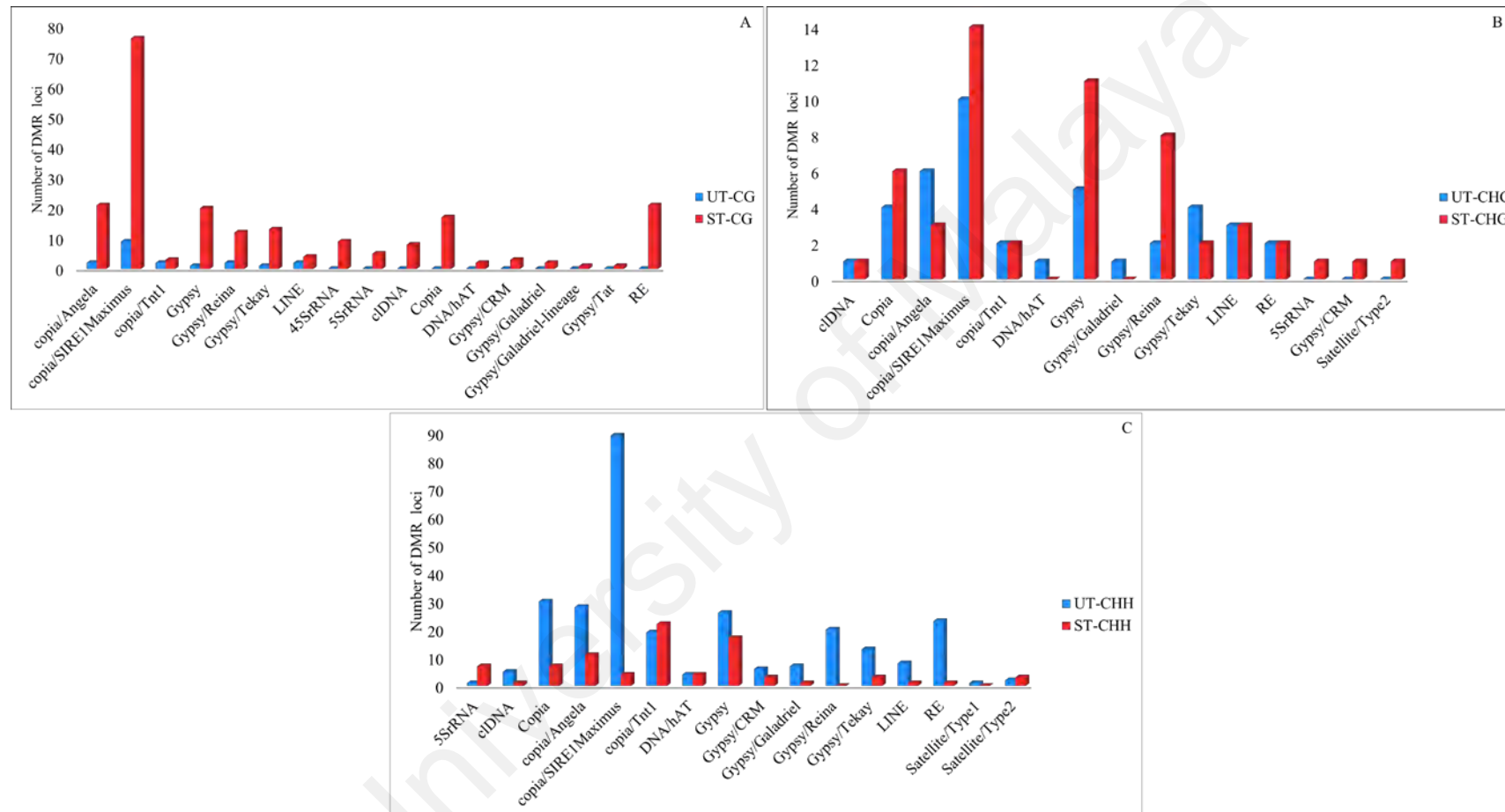
Figure 4.17: Distribution of DMR across repeat loci on banana genome. Number of loci with of DMR's associated (repeat loci, repeat 2kb upstream and 2kb downstream regions) with repeats classes. Plots are drawn for each methylation contexts CG(A), CHG(B) and CHH(C) to highlight the association of DMR in salt stress (UT) and non-salt stress (ST) methylomes.

CHAPTER 5: DISCUSSION

5.1   Comparative miRNA profiles in Musa A- and B-genomes

Among predicted orthologous miRNA families between *Musa* A and B genomes, 42 miRNA families are common to both genomes and are also shared among common ancestors of *Musa* i.e. embrophytes, angiosperms and poales (D'Hont et al., 2012). For example, the miR528 family which had previously been reported only for poales genomes (Friedlander et al., 2012), was shown by D'Hont et al. (2012) to be present in the A-genome and confirmed to be in the B-genome based on the analysis included in this thesis. Later banana miRNA expression studies revealed involvement of miR528 in fruit ripening (Bi et al., 2015), salt stress responses (Lee et al., 2015), cold stress (Jingyi et al., 2017) and fusarium-wilt (Song, 2016). Among the newly-predicted known miRNA families present in both *Musa* genomes, miR1134 has been reported as being abiotic stress - related and is found in the monocots *Triticum aestivum* (Wheat) (Eren et al., 2015) and *Festuca arundinacea* (Tall fescue) (Unver et al., 2010); the miR5021 reported to be involved in abiotic stress in *Brassica juncea* (Singh et al., 2017) and Arabidopsis (Izadi et al., 2017). whilst the families miRf10125, miRf10576, miRf11033, miRf11143 and miRf11357 are all of unknown function but were also computationally predicted from the Arabidopsis, poplar and rice genome sequences (Zhang et al., 2010b). The higher number of miRNA loci in the B-genome may be related to a higher number of transposable elements (transposons and retrotransposons) present as these are thought to have contributed to the generation of species-specific miRNA genes in plants (Nozawa et al., 2012). The differences in retroelements and miRNA on homologous chromosomes suggest that some of these miRNAs have arisen after the whole genome duplication events, since chromosomes 9 and 10 *Musa* block 2 of D'Hont et al. (2012) are among the regions thought not to have been involved in the paleopolyploidisation (D'Hont et al., 2012).

Among *Musa*-specific miRNA targets predicted, casein kinase (ITC1587_Bchr2_P03722), a predicted target of mba-miR3 was previously reported to affect multiple developmental and stress response pathways in Arabidopsis (Vilela et al., 2015). Dirigent (ITC1587_Bchr9_P26004), a predicted target of mba-miR15, which are a family of proteins associated with lignification, biotic and abiotic stress responses in rice (Liao et al., 2017) and *Medicago sativa* L (Behr et al., 2015). Mba-miR8 targets the multidrug and toxic compound extrusion (MATE) family (ITC1587_Bchr1_G01417), which in plants are associated with citric acid efflux and metal tolerance in *Medicago trancatula* (Wang et al., 2017) and zinc tolerance in Arabidopsis (Pineau et al., 2012) and aluminium tolerance in sorghum (Magalhaes, 2010). An additional predicted target of mba-miR8, Sal-1 phosphatase (ITC1587_chr2_G03828), has been reported to be inactivated by oxidative stress in chloroplast of which controls accumulation of substrate, as a plant stress signal (Chan et al., 2016). Given the wider stress and disease resistance reported for banana B genomes (Safar et al., 2004; Tripathi et al., 2008), further functional validation of these miRNA and target genes, and in particular those of unknown function, is of particular interest.

5.2   Genome-wide salt stress responsive miRNA

MicroRNA (miRNA) co-regulate gene expression by acting at a post-transcriptional level, in association with TF to play a major role in plant development and responses to stress (Guo et al., 2016). The release of version-2 of the banana genome (Martin et al., 2016) with improved assembly of intergenic regions of the genome, allowed the identification of several additional *Musa*-specific miRNA loci and sequences. From the genome sequence data, together with new degradome data, it was possible to identify transcriptional regulatory elements in promoter regions of banana miRNA and to predict interactions between miRNA and the TF that either are targets or regulators of miRNA.

5.2.1 Highly represented TFBS motifs in miRNA gene promoter regions

Genome wide scanning showed the TFBS motifs for some TF families to be located within the promoter regions of most of the banana miRNAs (Figure 4.5). Over-representation of certain "miRNA-preferred" TFBS motifs was first reported in the miRNA promoters of Arabidopsis (Megraw et al., 2006) and significantly conserved motifs were identified within putative miRNA promoter regions of Arabidopsis and rice (Zhou et al., 2007). Highly abundant TFBS identified in banana miRNA promoters included binding sites for TCP, which has been associated with leaf development and growth hormone signalling in Arabidopsis (Li, 2015), NF-Y, which is involved in drought tolerance and flower development in Arabidopsis (Petroni et al., 2012; Siriwardana et al., 2016) and AP2 which is a regulator of floral development in Arabidopsis (Xie et al., 2015) and nodule formation in common bean (Nova-Franco et al., 2015). Other highly represented TFBS belong to bZIP, bHLH, ZF-HD and Dehydrin which are involved in tolerance of drought and salt stress in banana (Muthusamy et al., 2016; Shekhawat et al., 2011). Such highly abundant TFBS motifs likely represent conservation of regulatory elements in the promoters of banana miRNA genes, together with the regulatory networks of these TF families common among higher plants.

5.2.2 Orthologous miRNA target auxin signalling, redox homeostasis and developmental specific genes

Several miRNAs and their targets are conserved across plant families, as reported for miR156/SPLs, miR164/NACs, miR166/HOX, miR167/ARFs, miR171/SCLs, miR319/GAMYB, miR396/GRFs, miR528/SOD (Morea et al., 2016; Zhang et al., 2006a). Orthologous miRNA in the banana genome also showed conserved targets based on degradome data (Appendix D Table3) (refer to section 3.6.9 for data source). Gene ontology of predicted miRNA targets, showed predominance for auxin activated signalling, oxidation-reduction processes (redox) and developmental processes

(Appendix D Table4) (refer to section 3.6.9 for data source). In banana, auxin signalling has been reported to be involved in the progression of fruit development and ripening, and in responses to abiotic stresses (Hu et al., 2015). Stress responses in plants are associated with the generation of reactive oxygen species (ROS), which interact with redox homeostasis and are regulated by miRNA-targeted transcription factors in plants (reviewed in Sewelam et al. (2016)). In banana, redox homeostasis genes are reported to be differentially expressed in plants exposed to cold (Yang et al., 2015), ethylene and high temperature treatments (Du et al., 2016). miRNA-targeted redox homeostasis genes such as AIF2 (target of mac-miR156), laccase-25 and laccase -17 (targets of mac-miR397b-3p.1), superoxide dismutase (target of mac-miR528-5p), cytochrome p450 (target of mac-miR396c-5p.1) were observed in the current study.

5.2.3 Targets of Musa-specific miRNA have functions associated with root development and salt stress responses

Musa-specific miRNAs targets identified in the current study (Table 4.6) have functions related to plant development, signal transduction and homeostasis, including auxin signalling, protein dephosphorylation, gravitropism, calcium ion transport, sugar transport and oxidation reduction (redox). In plant roots, miRNAs are involved in auxin signalling to promote root cap formation, lateral root development, adventitious rooting and primary root growth (Gutierrez et al., 2009; Meng et al., 2010). Some of the *Musa*-specific miRNA targets are also involved in biotic and abiotic stress responses: RPS2 has been shown to be involved in biotic stress responses in Arabidopsis (Wilton et al., 2010), rice (Xu et al., 2014) and potato (Song et al., 2003); GSTU17 and PP2C are responsive to salinity and drought stress in Arabidopsis (Chen et al., 2012; Krzywinska et al., 2016). Other *Musa*-specific miRNA targets have potential roles in plant development, based on their Arabidopsis orthologs e.g. CXIP4 (Manohar et al., 2011), ELF6 (Yu et al., 2008) and PDR3 (Ticconi et al., 2009). Mac-miR6 and mac-miR37 were validated to be salt

stress responsive miRNA in our previous study (Lee et al., 2015). Mac-miR6 targets STP13 (sugar transport protein 13) which plays an active role in fungal defence in Arabidopsis (Lemonnier et al., 2014). Mac-miR37 targets tubulin-alpha-1 chain (TUBA6) which is associated with forming tubulin heterodimer important for cell wall assembly, tissue patterning and root growth in Arabidopsis (Buschmann et al., 2009). *Musa*-specific mac-miR6, was first identified in the version-1 genome of *Musa acuminata* (Davey et al., 2013) and is predicted to target stress responses and membrane transport.

5.2.4   Network mapping of miRNA and TF targets in banana suggest feedback regulation as an important regulatory module

Integrating miRNA and TF into genome scale gene regulatory networks has revealed reciprocal regulation and coordinated regulation of shared genes (Martinez & Walhout, 2009). In plants, such regulatory networks are classified into lock-on switch (involves self-regulating of TFs and miRNA), feedback loop (involves miRNA repressing TF and TF inducing miRNA) and miRNA-mediated networks (involves both miRNA and TF in controlling another component either TF or non-TF protein coding genes) (Martinez & Walhout, 2009; Megraw et al., 2016). In banana, based on our data presented in this thesis, it is proposed that there are feedback loops between miRNA (orthologous and *Musa*-specific), TF targets and respective TFBS on miRNA promoters (Figure 4.6). Feedback loop in which a miRNA-regulated transcription factor regulates the transcription of its cognate miRNA. For example, miR156 has the predicted target TF SPL, and also has a TFBS for SPL in the promoter region, which forms a positive feedback loop involving miR156 and SPLs reinforces the proper timing for lateral root development progression (Yu et al., 2015). Feedback loops predictions for four miRNAs (mac-miR156 with target SPLs; mac-miR160 and mac-miR167 with targets ARF17 and ARF6; mac-miR164 with target NACs) can be supported based on studies of miRNA-TF orthologues that have been associated with root development in plants. In another instance, miR164 family targets

NAC transcription factors in banana, which also forms a feedback regulatory network to temporally control lateral root formation in maize (Li et al., 2012). Similarly, miR160 and miR167 which target both ARF17 and ARF6, form a regulatory feedback network to develop adventitious roots in Arabidopsis (Gutierrez et al., 2009). Interestingly, miR396 regulates transition of Arabidopsis root stem cells by transit-amplifying cells (transient amplifying cell divisions to ensure enough cells for organ growth) by forming a regulatory circuit by repressing GRF (Rodriguez et al., 2015) and current study predicts an auto-regulatory feedback circuit between mac-miR396 and its target GRF in banana roots. Also, proposed putative feedback loops involving *Musa*-specific miRNA (mac-miR-new14 targets WRKY21 and mac-miR-new20 targets ELF6) which, based on the function of the predicted mRNA targets, may fine tune the regulation of root development, biotic and abiotic responses in banana. Based on current analysis, a model specific to banana miRNA and TF regulation is proposed (Figure 5.1). Similar auto-regulatory feedback loops were also proposed by Arora et al. (2013) in drosophila. These auto-regulatory loop can behave as unilateral or reciprocal negative feedback loops and double negative feedback loops (Krol et al., 2010). Negative and positive feedback loop models regulated by miRNA were also proposed in calcineurin/NFAT signalling pathway in humans (Kannambath, 2016). In plants such feedback loop models have been well reviewed by Megraw et al. (2016), defined as miRNA-containing regulatory genetic circuits. The model proposed in banana is based on literature supported concepts of feedback loops and it may serve as working model to further validate such feedback regulations directed by miRNA. Other modes of miRNA containing feedback regulation involves larger context of networks where TF and miRNA directly or indirectly regulate other non-TF protein coding genes. In banana, both orthologous miRNA and *Musa*-specific miRNA target TFs and also non-TF protein coding genes which might form a larger regulatory network influencing several processes. In banana, our study identified miRNA and TF

regulatory networks which suggests future works into understanding their role in phenotype, physiology and response to various environmental stresses.
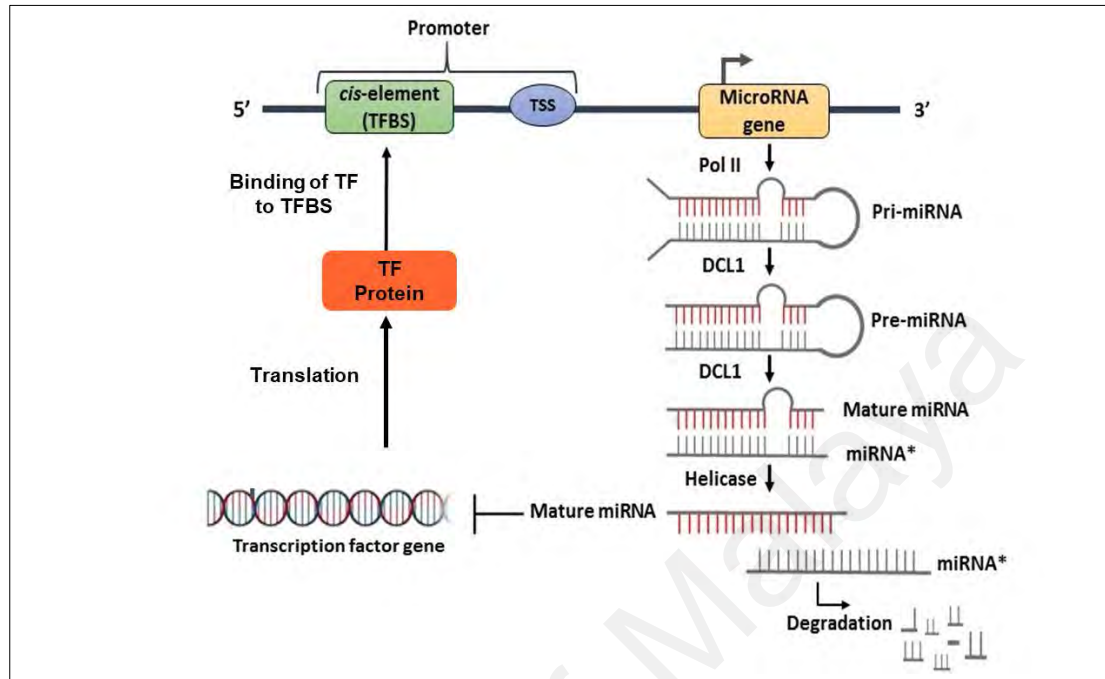
Figure 5.1: Proposed model for miRNA mediated feedback regulation in banana. Model for feedback regulatory loop of miRNA and transcription factor in banana. For simplicity, all possible mechanisms are not presented. TFBS: Transcription factor binding site; TSS: Transcription start site; Pol II: RNA Polymerase II; DCL1: DICER LIKE family 1; Pri-miRNA: Primary microRNA; Pre-miRNA: Precursor microRNA; mRNA: Messenger RNA.

5.3   Dynamics of DNA methylation in response to salt stress in banana

In plants, DNA methylation is defined by CG and non-CG contexts, but non-CG methylation plays an important role which involves silencing DNA via RdDM pathway mainly in monocots and dicots (Law & Jacobsen, 2010; Matzke & Mosher, 2014). In salt stress, plants show dynamic methylation changes by demethylation or loss of cytosine methylation (Baek et al., 2011; Song et al., 2012), transcriptional activation of genic regions (Song et al., 2012) and hypermethylations (Dyachenko et al., 2006; Wang et al., 2015). However, the current study didn't examine histone modifications such as methylation and acetylation which are known to effect transcription (Pikaard & Mittelsten

Scheid, 2014). Genome-wide DNA methylation profiling on banana genome associated with salt stress, provides a useful resource of an overall epigenome map and stress related methylation dynamics which influence gene expressions and transposon methylation.

5.3.1 Banana methylomes

DNA methylation levels in non-salt stress methylome are directly proportional to the repeat loci on the genome (Figure 4.10) which is common pattern observed in other plants such as Arabidopsis (Lister et al., 2008), two maize inbred lines (Regulski et al., 2013), tomato (Zhong et al., 2013), Soyabean (Schmitz et al., 2013) and Brachypodium (Eichten et al., 2016). Enrichment of DNA methylated levels at repeat loci in banana genome may also signify the location of centromeric heterochromatic regions where repeats are highly localized (Cizkova et al., 2013; D'Hont et al., 2012). Average DNA methylation percentages for CG (31.1%), CHG (12.55%) and CHH (2.34%) contexts in banana genome follow similar trend of methylation prominent in other plants where CG methylation make up largest proportion of total DNA methylation (Niederhuth et al., 2016). Gene body methylation is dominated by CG methylation compared to upstream and downstream regions of the genes (Figure 4.11) which is consistent with observations in other plants (Mirouze & Vitte, 2014; Song et al., 2013b). In contrast, genome-wide CHH methylation is low compared to CG and CHG methylation i.e. ~ 2% of total methylated cytosines. Such lower levels of methylated CHH (~1.4 - 5.8%), both in terms of total DNA methylation level and as proportion of total methylated sites are also reported in poaceae (grass) family and clonally propagated species (Niederhuth et al., 2016). Prevelance of CHH islands in transposons and gene flanking regions have been observed in maize and other grass monocots but role of CHH islands influencing gene expression is unclear (Gehring, 2016; Gent et al., 2013). Nevertheless, in banana non-CG context i.e CHG and CHH show hypermethylation in intronic and repeat loci, might be associated with origin of 21-24nt siRNA which guide gene methylations in plants (Chen

et al., 2011; Qin et al., 2015). Overall proportion of cytosine methylation is induced due to salinity stress in banana and is majorly contributed by non-CG context (CHG and CHH) methylation. Such salt-induced alterations in methylation in non-CG context were also observed in *Hordeum vulgare* (Konate et al., 2018) and *Medicago tranculata* (Yaish et al., 2018). Average methylation ratios in CG context showed no statistically significant difference between salt stress and non-salt stress methylomes (Table S11). However, targetted analysis on genome by identifying siRNA loci and differentially methylated regions across genome have shown several hyper- and hypo-methylations induced by salt stress.

5.3.2   siRNA role in influencing salt stress associated methylations

Small RNA abundance suggests repression of 21 and 24nt siRNA during salt stress (Figure 4.9). Small RNA (21-24nt) can be classified into microRNA (miRNA), transacting siRNA (ta-siRNA), repeat associated siRNA (ra-siRNA) and natural antisense transcripts siRNA (nat-siRNA) (Axtell, 2013). As observed in banana datasets, 24nt siRNA also show low abundance corresponding to salt stress in Arabidopsis where 24nt siRNA from SRO5 mRNA which targets P5CDH leads to mRNA degradation which triggers proline accumulation to tolerate excess salt (Borsani et al., 2005). Similarly, 24nt siRNA influence in response to salt stress is also shown in wheat seedling (Yao et al., 2010) and Arabidopsis where 24nt siRNAs accumulation is decreased in wild type plants (Xu et al., 2015). Distribution of 21 and 24nt siRNA among repeats of banana, show high abundance of small RNA in tandem repeats i.e rDNA satellites (45sRNA, 5sRNA and other satellites) (Figure S7). Interestingly, abundance of 24nt siRNA is more than 2-fold high during salt stress in rDNA satellites while abundance is repressed in LTR elements Copia and Gypsy. Tandem repeats have been shown to be involved in epigenetic silencing by RNA interference because of siRNA generated by RNA-Dependent RNA polymerase and Dicer (Martienssen, 2003). Tandem repeats are also associated with gene silencing

phenomenon which become target for DNA methylation via RdDM pathway (Chan et al., 2006; Matzke et al., 2007). Differentially expressed siRNA clusters associate with high methylation for CG, CHG and CHH in salt stress (Figure 4.12). siRNA clusters show low methylation in CG and CHG but high methylation only in CHH context in non-salt stress condition (Figure 4.12). RdDM targeted regions have been shown to have highest CHH methylation levels which primarily corresponds to transposons and also form CHH islands (Stroud et al., 2014; Zemach et al., 2013). A study on several plant methylomes also showed 24nt siRNA to map to regions of higher CHH methylation which are smaller in size, in comparison to regions of low or intermediate DNA methylations (Niederhuth et al., 2016). This suggests smaller CHH methylated loci overlapping siRNA clusters in banana genome might driven by RdDM for silencing transposons and genes.

5.3.3　　Transcriptional and methylation profiling without replicates

Next generation sequencing technologies have been widely applied to study multiple biological "omics" such as transcriptome and methylome.To detect biological effect caused by multiple "omics", biological replicates are considered as essential part of the experimental design (Fang & Cui, 2011). However, limited numbers of replicates are possible within most available budgets which will often lead to unstable estimates. In this typical situation there is a need to adopt methods that use biological variation from sequencing data to establish prior distribution of data. DEGseq is one such method used to identify differentially expressed transcripts from transcriptome data sequencing for the current study. DEGseq is applicable for biologically pooled average experiments where number of reads per gene are sampled independently, estimating underlying binomial distribution by random sampling model (MA-plot) (Wang et al., 2010). Among model-based estimates with simulation of prior distribution, DEGseq was shown to perform well without replicates (Lee et al., 2011). In contrast, performance of DEGseq is relatively poor when using large replicate data (seven technical replicates for each of 96 biological

replicates) and in comparison, to other tools such as DESeq, DEseq2, EBseq, EdgeR and Limma (Schurch et al., 2016). Likewise, to identify differentially methylated regions (DMR) from methylomes, a statistical method DSS-single was utilised. Method characterises the count data of methylation which accounts for spatial correlation and sequence depth to estimate biological variation even without replicates (Wu et al., 2015). Overall read depth and pre filtered methylated cytosine counts per position as considered as key factors for calling DMR regions (refer to section 3.6.6 for detailed methods), the DSS method is considered as robust with desirable distribution (Park & Wu, 2016). DSS method can be efficiently applied for analyzing RRBS data (refer to section 3.5.1), it considers densely clustered CpG sites for spatial correlations (Wu et al., 2015). However, use of statistical methods such as DEGseq and DSS should be completely understood and applied only in suitable studies and alternative validations of such findings is necessary to avoid false positives (Fang & Cui, 2011). For each of the above analysis and predictions, there are concerns and limitations so needs to be some validation using alternative approaches such as qualitative real time PCR (qRT-PCR) for gene expression and Methylation-specific PCR (MS-PCR) for differentially methylated sites to detemine the effectiveness of these approches for the current study.

5.3.4 Gene expression might be influenced by adjacent DMR and siRNA loci

DMR detected using the DSS method showed a hypermethylation in CG context and hypomethylation in CHH context relative to salt stress (Figure 4.14A). Exon and intronic regions are hypermethylated in CG context which corresponds to hypomethylation of CHH (Figure 4.14B). This phenomenon might be result of demethylation of CHH across genomic loci due to effect of salt stress. Such CHH demethylation induced by heat stress was reported to activate stress responsive genes in Arabidopsis (Popova et al., 2013). In wheat, osmotic and salinity stresses induce demethylation in the promoter region of the TaGAPC1 gene (Fei et al., 2017). In tobbaco, cold stress induces DNA demethylation in

the coding sequence of NtGPDL gene which correlates with gene expression (Choi & Sano, 2007). Demethylation may also be directed by RdDM pathway where ROS3 (RNA recognition motif-containing protein) binds to small RNAs to perform sequence specific demethylation (Zheng et al., 2008). Demethylation may be one of the roles of predicted siRNA loci which coincides with DMR regions on genic regions in banana. DMR association with gene expression in banana reveals genes may be regulated by upstream, downstream or intronic methylations (Figure 4.15). Majority of genes (82.1%, 64/78) which coincides with DMR and siRNA loci at are either CG hypermethylations and CHH hypomethylations (Figure 4.16 A-H). Such dynamic methylation changes were also observed in drought and salinity responses in rice cultivars where gene expression is influenced by proximal transposable elements via methylation or demethylation (Garg et al., 2015). In leaf tissues of Maize abiotic stress plantlets, 24nt siRNA loci were specificially enriched in the upstream region of the most highly expressed genes (Lunardon et al., 2016). Similarly in switchgrass leaf tissue, abundance of siRNAs at upstream promoter regions of genes show positive correlation with CHH methylation and methylation show positive association with gene expression (Yan et al., 2018). In contrast, negative association was also observed between upstream promoter CG methylation and gene expression levels induced by heat stress in leaf tissues of *Brassica rapa* (Liu et al., 2018). Similar phenomenon is also observed in current study, hypermethylated CG show low abundance of siRNA in upstream gene regions which coincides with upregulation of gene expression.While hypomethylated CHH show high abundance of siRNA in upstream gene regions which coincides with upregulation of gene expression. Although this pattern is not consistent all over the genome, siRNA cluster overlap CG and CHH methylation might be *de novo* DNA methylations or demethylations in regulating genes as response to salt stress in banana. To summarise methylation dynamics observed in salt stress banana roots, a hypothetical model has been proposed to

show how DNA methylation modulate gene expression in association with siRNA in response to salt stress (Figure 5.2). Analogous models showing dynamics of DNA methylations were proposed in other plants, for example, Dowen et al. (2012) shows how DMR at proximity of the gene in association with 21nt siRNA regulate gene expression in response to biotic stress in infected leaf tissues of Arabidopsis. A schematic representation of events related to dynamic DNA methylations, show similar events where hypermethylation and demethylation control gene expression during environmental stress imposition in plant cells (Sahu et al., 2013). The model proposed for current study might serves as framework to further elucidate the dynamic repsonses due to salt stress in banana. The current study reports the genome wide interplay between multiple omics from salt stress, i.e. transciptome, small RNA "ome" and methylome. The data were used to identify salt stress responsive genes in adjacent to differentially methylated regions that were overlapping 24nt siRNA clusters across the banana genome. Such genomic loci with DMR and siRNA clusters evidences the prevelance of RdDM and support a role in regulating stress responses in banana. It must be considered, however that a lack of replicates weakens the statistical strength of the current data which might result in false positives. Hence, to confirm the observations in the study will require more replicates for further validations to reinforce the biological meaning of the findings.
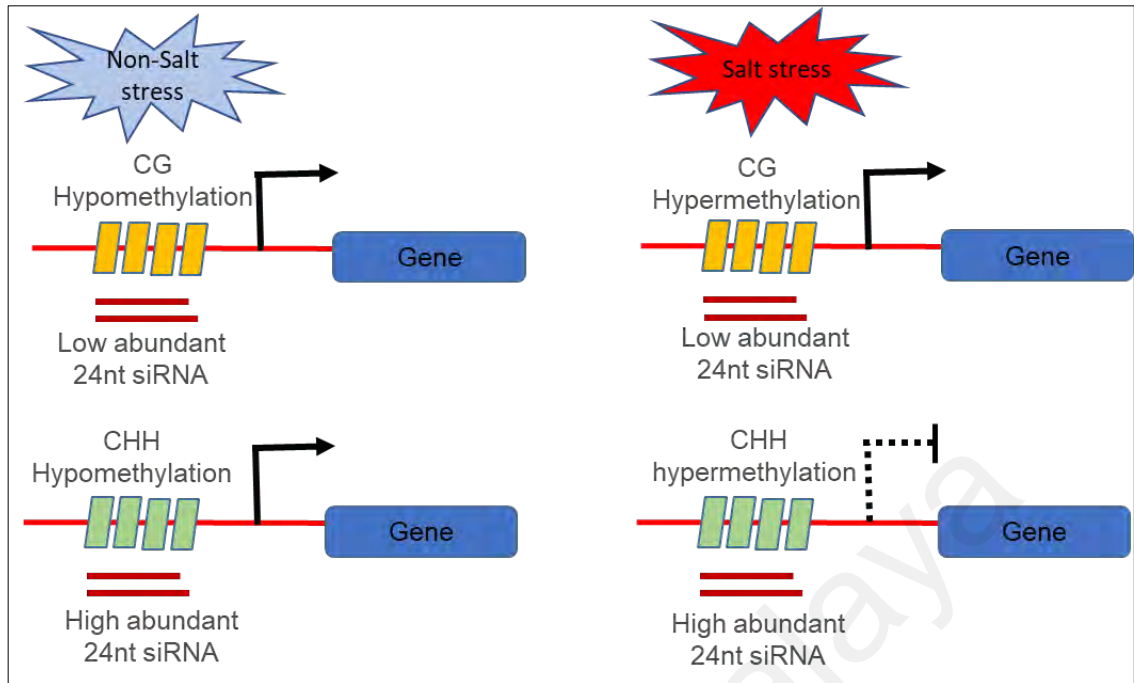
Figure 5.2: Proposed model on dynamic DNA methylation changes observed in banana methylomes. Model shows association of DNA methylation with gene expression and 24nt siRNA abundance in non-salt stress and salt-stress conditions.

CHAPTER 6: CONCLUSION

Study reported in this thesis is majorly data driven and analysis is combination of right choice of bioinformatics tools and customized scripts. Such data driven analysis contributes a sucessful model into the field of plant omics and agriculture to define biological meaning to the data. Futher integration of omics data resource and plant phenotype will connect the genotype and phenotype to understand the fundamental processes in plants.

With sequenced banana genomes of *Musa acuminata* and *Musa balbisiana*, current research identifies genome-wide miRNA loci, using high-throughput sequencing small RNA datasets. Both Musa A and Musa B genomes encode 28 *Musa*-specific miRNA families predicted to target multiple developmental and stress responsive genes.

In this study, miRNA profiles among salt-stressed banana roots were predicted using small RNA transcriptome and degradome high-throughput sequencing datasets. A total of 180 mature miRNA belonging to 20 orthologous miRNA families and 39 *Musa*-specific families were identified on the most recently published banana genome sequence using small RNA transcriptome data from salt stressed banana roots. Degradome data was used to confirm 128 predicted targets from both orthologous miRNA and *Musa*-specific miRNA families. miRNA targets of *Musa*-specific miRNAs were transcription factors mainly related to auxin signalling, redox homeostasis and root development. With predicted miRNA genomic loci, putative banana miRNA promoter regions were identified which show transcription factor binding sites that are highly represented in banana miRNA promoter regions. To integrate miRNA, miRNA transcription factor targets and transcription factor binding sites, a miRNA and transcription factor regulatory circuits were successfully constructed with miRNA target specific binding sites for

orthologous miRNA and *Musa*-specific miRNA. Along with known miRNA and transcription factor feedback regulation, putative novel regulations were also identified in this study.

To further explore genome-wide regulation of small RNA in banana roots following exposure to salt stress, this study established association between DNA methylation, expression of genes and 21nt and 24nt small RNA using methylome, transcriptome and small RNA transcriptome datasets. Salt-stressed root samples displayed symmetric CG methylation and CHH demethylation adjacent to differentially expressed genes. Profiling of 21 and 24nt siRNA clusters on genomic loci showed increased methylation levels in CG, CHG and CHH contexts in response to salt stress. With successful integration of gene expression, small RNA and DNA methylation, 78 salt stress responsive genes with differential methylation of their loci and corresponding changes in associated 24nt siRNA were identified in this study. Genes include DRE2, DHN1, AP2, calcium permeable stress-gated cation channel 1-like and cation/H+ antiporter 20-like, and peroxidases (PER1, PER47) which are ROS-related antioxidants, indicating a role of siRNA directed DNA methylation in epigenetic regulation of stress signalling and response in the banana root. Epigenetic regulation involves not only DNA methylation, thus study other types of epigenetic marks such as remodeling of histones and their modification might show inherited stress memory which allows reprogramming of banana adaptation to salt stress.

This study introduced a novel approach of data integration to leverage the predicted functional connections between different omics related data associated with salt-stress in banana. The current study identified miRNA regulatory circuits that can be used to build larger regulatory networks and show the complexity of genetic regulation underlying a dynamic and responsive biological system for stress response. The study has also revealed genome-wide epigenetic markers that can be used to further explore the genetic and

phenotypic changes in banana plant due to salt stress. While the putative miRNA regulatory circuits and genome-wide epigenetic associations with salt stress reported in this study may be uncertain due to the limitations such as lack of replicates and complementary experimental evidence, this study can be considered as a demonstration of approaches to identify strong gene candidates for future functional validation and provides an example of a multi-omics data approach for understanding a complex biological network, such as that explored here towards a better understanding of salt stress responses in banana (Figure 6.1).
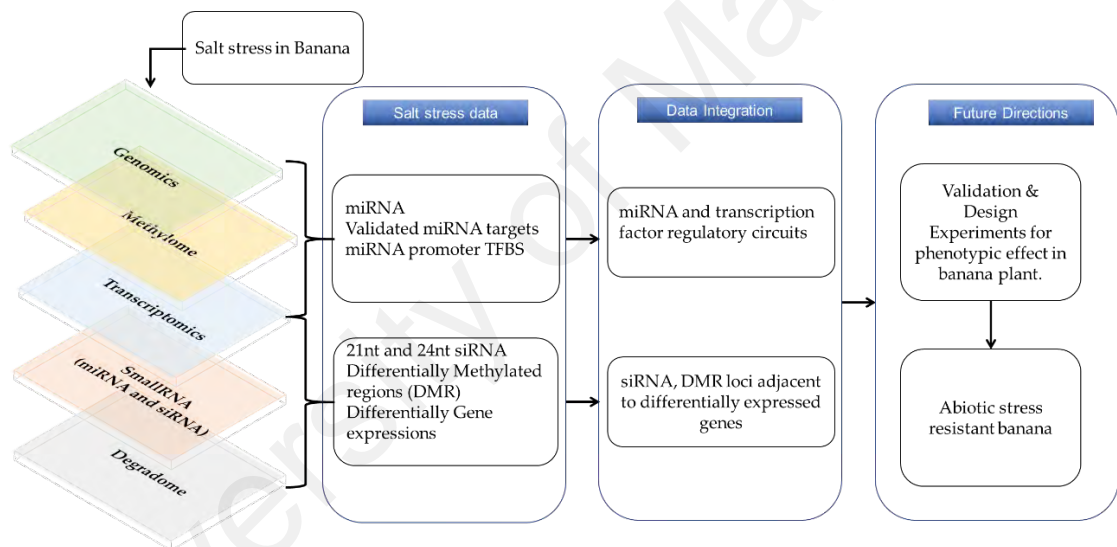
Figure 6.1: Schematic workflow showing current study and future directions. Integration of multi-omics data from the current study can be applied to develop strategies which combat abiotic stress in banana plants.

REFERENCES

Addo-Quaye, C., Eshoo, T. W., Bartel, D. P., & Axtell, M. J. (2008). Endogenous siRNA and miRNA targets identified by sequencing of the Arabidopsis degradome. *Current Biology, 18*(10), 758-762.

Adessi, C., Matton, G., Ayala, G., Turcatti, G., Mermod, J. J., Mayer, P., & Kawashima, E. (2000). Solid phase DNA amplification: Characterisation of primer attachment and amplification mechanisms. *Nucleic Acids Research, 28*(20), E87.

Alejandri-Ramirez, N. D., Chavez-Hernandez, E. C., Contreras-Guerra, J. L., Reyes, J. L., & Dinkova, T. D. (2018). Small RNA differential expression and regulation in Tuxpeno maize embryogenic callus induction and establishment. *Plant Physiology and Biochemistry, 122*, 78-89.

Allen, E., Xie, Z., Gustafson, A. M., & Carrington, J. C. (2005). microRNA-directed phasing during trans-acting siRNA biogenesis in plants. *Cell, 121*(2), 207-221.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology, 215*(3), 403-410.

Andrews, S. (2010). FastQC: A quality control tool for high throughput sequence data. . Retrieved 05 February 2013 from http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Arora, S., Rana, R., Chhabra, A., Jaiswal, A., & Rani, V. (2013). miRNA-transcription factor interactions: A combinatorial regulation of gene expression. *Molecular Genetics and Genomics, 288*(3-4), 77-87.

Axtell, M. J. (2013). Classification and comparison of small RNAs from plants. *Annual Review of Plant Biology, 64*, 137-159.

Baek, D., Jiang, J. F., Chung, J. S., Wang, B. S., Chen, J. P., Xin, Z. G., & Shi, H. Z. (2011). Regulated AtHKT1 Gene expression by a distal enhancer element and DNA Methylation in the promoter plays an important role in salt tolerance. *Plant and Cell Physiology, 52*(1), 149-161.

Behr, M., Legay, S., Hausman, J. F., & Guerriero, G. (2015). Analysis of cell wall-related genes in organs of *Medicago sativa* L. under different abiotic stresses. *International Journal of Molecular Sciences, 16*(7), 16104-16124.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological), 57*(1), 289-300.

Bhadauria, V. (2017). *Next-generation sequencing and bioinformatics for plant science*. Norfolk, UK: Caister Academic Press.

Bi, F., Meng, X., Ma, C., & Yi, G. (2015). Identification of miRNAs involved in fruit ripening in Cavendish bananas by deep sequencing. *BMC Genomics, 16*, 776.

Bielewicz, D., Kalak, M., Kalyna, M., Windels, D., Barta, A., Vazquez, F., . . . Jarmolowski, A. (2013). Introns of plant pri-miRNAs enhance miRNA biogenesis. *EMBO Reports, 14*(7), 622-628.

Blevins, T., Podicheti, R., Mishra, V., Marasco, M., Wang, J., Rusch, D., . . . Pikaard, C. S. (2015). Identification of Pol IV and RDR2-dependent precursors of 24 nt siRNAs guiding de novo DNA methylation in Arabidopsis. *Elife, 4*, e09591.

Bonnet, E., Wuyts, J., Rouze, P., & Van de Peer, Y. (2004). Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. *Bioinformatics, 20*(17), 2911-2917.

Borges, F., & Martienssen, R. A. (2015). The expanding world of small RNAs in plants. *Nature Reviews: Molecular Cell Biology, 16*(12), 727-741.

Borsani, O., Zhu, J., Verslues, P. E., Sunkar, R., & Zhu, J. K. (2005). Endogenous siRNAs derived from a pair of natural cis-antisense transcripts regulate salt tolerance in Arabidopsis. *Cell, 123*(7), 1279-1291.

Brady, S. M., Zhang, L., Megraw, M., Martinez, N. J., Jiang, E., Yi, C. S., . . . Benfey, P. N. (2011). A stele-enriched gene regulatory network in the Arabidopsis root. *Molecular Systems Biology, 7*, 459.

Brettell, R. I., & Dennis, E. S. (1991). Reactivation of a silent Ac following tissue culture is associated with heritable alterations in its methylation pattern. *Molecular and General Genetics, 229*(3), 365-372.

Brousse, C., Liu, Q. K., Beauclair, L., Deremetz, A., Axtell, M. J., & Bouche, N. (2014). A non-canonical plant microRNA target site. *Nucleic Acids Research, 42*(8), 5270-5279.

Budak, H., & Kantar, M. (2015). Harnessing NGS and big data optimally: Comparison of miRNA prediction from assembled versus non-assembled sequencing data-The case of the grass *Aegilops tauschii* complex genome. *OMICS: A Journal of Integrative Biology, 19*(7), 407-415.

Budak, H., Bulut, R., Kantar, M., & Alptekin, B. (2016). MicroRNA nomenclature and the need for a revised naming prescription. *Briefings in Functional Genomics, 15*(1), 65-71.

Buschmann, H., Hauptmann, M., Niessing, D., Lloyd, C. W., & Schaffner, A. R. (2009). Helical growth of the Arabidopsis mutant tortifolia2 does not depend on cell division patterns but involves handed twisting of isolated cells. *Plant Cell, 21*(7), 2090-2106.

Buske, F. A., Bodén, M., Bauer, D. C., & Bailey, T. L. (2010). Assigning roles to DNA regulatory motifs using comparative genomics. *Bioinformatics, 26*(7), 860-866.

Cantu, D., Vanzetti, L. S., Sumner, A., Dubcovsky, M., Matvienko, M., Distelfeld, A., . . . Dubcovsky, J. (2010). Small RNAs, DNA methylation and transposable elements in wheat. *BMC Genomics, 11*, 408.

Cao, X., & Jacobsen, S. E. (2002). Locus-specific control of asymmetric and CpNpG methylation by the DRM and CMT3 methyltransferase genes. *Proceedings of the National Academy of Sciences of the United States of America, 99 Suppl 4*, 16491-16498.

Cao, Y., Fanning, S., Proos, S., Jordan, K., & Srikumar, S. (2017). A review on the applications of next generation sequencing technologies as applied to food-related microbiome studies. *Frontiers in Microbiology, 8*, 1829.

Chai, J., Feng, R. J., Shi, H. R., Ren, M. Y., Zhang, Y. D., & Wang, J. Y. (2015). Bioinformatic identification and expression analysis of banana MicroRNAs and their targets. *PLoS ONE, 10*(4), e0123083.

Chan, K. X., Mabbitt, P. D., Phua, S. Y., Mueller, J. W., Nisar, N., Gigolashvili, T., . . . Pogson, B. J. (2016). Sensing and signaling of oxidative stress in chloroplasts by inactivation of the SAL1 phosphoadenosine phosphatase. *Proceedings of the National Academy of Sciences of the United States of America, 113*(31), E4567-4576.

Chan, S. W. L., Zhang, X. Y., Bernatavichute, Y. V., & Jacobsen, S. E. (2006). Two-step recruitment of RNA-directed DNA methylation to tandem repeats. *PLoS Biology, 4*(11), 1923-1933.

Chen, D., Meng, Y., Yuan, C., Bai, L., Huang, D., Lv, S., . . . Chen, M. (2011). Plant siRNAs from introns mediate DNA methylation of host genes. *RNA, 17*(6), 1012-1024.

Chen, J. H., Jiang, H. W., Hsieh, E. J., Chen, H. Y., Chien, C. T., Hsieh, H. L., & Lin, T. P. (2012). Drought and salt stress tolerance of an Arabidopsis glutathione S-transferase U17 knockout mutant are attributed to the combined effect of glutathione and abscisic acid. *Plant Physiology, 158*(1), 340-351.

Chen, M., Wei, M., Dong, Z. H., Bao, H., & Wang, Y. W. (2016). Genomic identification of microRNA promoters and their cis-acting elements in Populus. *Genes & Genomics, 38*(4), 377-387.

Chen, T., & Li, E. (2004). Structure and function of eukaryotic DNA methyltransferases. *Current Topics in Developmental Biology, 60*, 55-89.

Chinnusamy, V., & Zhu, J. K. (2009). RNA-directed DNA methylation and demethylation in plants. *Science in China. Series C: Life Sciences, 52*(4), 331-343.

Choi, C. S., & Sano, H. (2007). Abiotic-stress induces demethylation and transcriptional activation of a gene encoding a glycerophosphodiesterase-like protein in tobacco plants. *Molecular Genetics and Genomics, 277*(5), 589-600.

Chorostecki, U., Moro, B., Rojas, A. M. L., Debernardi, J. M., Schapire, A. L., Notredame, C., & Palatnik, J. F. (2017). Evolutionary footprints reveal insights into plant MicroRNA biogenesis. *Plant Cell, 29*(6), 1248-1261.

Chow, C. N., Zheng, H. Q., Wu, N. Y., Chien, C. H., Huang, H. D., Lee, T. Y., . . . Chang, W. C. (2016). PlantPAN 2.0: An update of plant promoter analysis navigator for reconstructing transcriptional regulatory networks in plants. *Nucleic Acids Research, 44*(D1), D1154-D1160.

Cizkova, J., Hribova, E., Humplikova, L., Christelova, P., Suchankova, P., & Dolezel, J. (2013). Molecular analysis and genomic organization of major DNA satellites in banana (*Musa* spp.). *PLoS ONE, 8*(1), e54808.

Cokus, S. J., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C. D., . . . Jacobsen, S. E. (2008). Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature, 452*(7184), 215-219.

Conesa, A., & Gotz, S. (2008). Blast2GO: A comprehensive suite for functional analysis in plant genomics. *International Journal of Plant Genomics, 2008*, 12.

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., . . . Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology, 17*, 13.

Crisp, P. A., Ganguly, D., Eichten, S. R., Borevitz, J. O., & Pogson, B. J. (2016). Reconsidering plant memory: Intersections between stress recovery, RNA turnover, and epigenetics. *Science Advances, 2*(2), e1501340.

Cuerda-Gil, D., & Slotkin, R. K. (2016). Non-canonical RNA-directed DNA methylation. *Nature Plants, 2*(11), 16163.

Cuperus, J. T., Fahlgren, N., & Carrington, J. C. (2011). Evolution and functional diversification of MIRNA genes. *Plant Cell, 23*(2), 431-442.

D'Ario, M., Griffiths-Jones, S., & Kim, M. (2017). Small RNAs: Big impact on plant development. *Trends in Plant Science, 22*(12), 1056-1068.

D'Hont, A., Denoeud, F., Aury, J. M., Baurens, F. C., Carreel, F., Garsmeur, O., . . . Wincker, P. (2012). The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants. *Nature, 488*(7410), 213-217.

D'Hont, A., Paget-Goy, A., Escoute, J., & Carreel, F. (2000). The interspecific genome structure of cultivated banana, *Musa* spp. revealed by genomic DNA in situ hybridization. *Theoretical and Applied Genetics, 100*(2), 177-183.

Dai, X., & Zhao, P. X. (2011). psRNATarget: A plant small RNA target analysis server. *Nucleic Acids Research, 39*(Web Server issue), W155-159.

Dalakouras, A., Dadami, E., & Wassenegger, M. (2013). Viroid-induced DNA methylation in plants. *Biomolecular Concepts, 4*(6), 557-565.

Dalakouras, A., & Wassenegger, M. (2013). Revisiting RNA-directed DNA methylation. *RNA Biology, 10*(3), 453-455.

Daniells, J. (2001). *Musalogue: A catalogue of Musa germplasm. Diversity in the genus Musa*. Montpellier (FRA): INIBAP.

Davey, M. W., Gudimella, R., Harikrishna, J. A., Sin, L. W., Khalid, N., & Keulemans, J. (2013). "A draft *Musa balbisiana* genome sequence for molecular genetics in polyploid, inter- and intra-specific *Musa* hybrids". *BMC Genomics, 14*, 683.

Deng, P. C., Nie, X. J., Wang, L., Cui, L. C., Liu, P. X., Tong, W., . . . Song, W. N. (2014). Computational identification and comparative analysis of miRNAs in Wheat group 7 chromosomes. *Plant Molecular Biology Reporter, 32*(2), 487-500.

Devi, S. J., Madhav, M. S., Kumar, G. R., Goel, A. K., Umakanth, B., Jahnavi, B., & Viraktamath, B. C. (2013). Identification of abiotic stress miRNA transcription factor binding motifs (TFBMs) in Rice. *Gene, 531*(1), 15-22.

Dowen, R. H., Pelizzola, M., Schmitz, R. J., Lister, R., Dowen, J. M., Nery, J. R., . . . Ecker, J. R. (2012). Widespread dynamic DNA methylation in response to biotic stress. *Proceedings of the National Academy of Sciences of the United States of America, 109*(32), E2183-2191.

Droc, G., Lariviere, D., Guignon, V., Yahiaoui, N., This, D., Garsmeur, O., . . . Bocs, S. (2013). The banana genome hub. *Database: The Journal of Biological Databases and Curation, 2013*, bat035.

Du, J., Johnson, L. M., Jacobsen, S. E., & Patel, D. J. (2015). DNA methylation pathways and their crosstalk with histone methylation. *Nature Reviews: Molecular Cell Biology, 16*(9), 519-532.

Du, L., Song, J., Forney, C., Palmer, L. C., Fillmore, S., & Zhang, Z. (2016). Proteome changes in banana fruit peel tissue in response to ethylene and high-temperature treatments. *Horticulture Research, 3*, 16012.

Dunoyer, P., Brosnan, C. A., Schott, G., Wang, Y., Jay, F., Alioua, A., . . . Voinnet, O. (2010). An endogenous, systemic RNAi pathway in plants. *EMBO Journal, 29*(10), 1699-1712.

Dyachenko, O. V., Zakharchenko, N. S., Shevchuk, T. V., Bohnert, H. J., Cushman, J. C., & Buryanov, Y. I. (2006). Effect of hypermethylation of CCWGG sequences in DNA of *Mesembryanthemum crystallinum* plants on their adaptation to salt stress. *Biochemistry-Moscow, 71*(4), 461-465.

Eichten, S. R., Stuart, T., Srivastava, A., Lister, R., & Borevitz, J. O. (2016). DNA methylation profiles of diverse *Brachypodium distachyon* align with underlying genetic diversity. *Genome Research, 26*(11), 1520-1531.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., . . . Turner, S. (2009). Real-time DNA sequencing from single polymerase molecules. *Science, 323*(5910), 133-138.

Enke, R. A., Dong, Z., & Bender, J. (2011). Small RNAs prevent transcription-coupled loss of histone H3 lysine 9 methylation in *Arabidopsis thaliana*. *PLoS Genetics, 7*(10), e1002350.

Erdmann, R. M., Satyaki, P. R. V., Klosinska, M., & Gehring, M. (2017). A small RNA pathway mediates allelic dosage in endosperm. *Cell Reports, 21*(12), 3364-3372.

Eren, H., Pekmezci, M., Okay, S., Turktas, M., Inal, B., Ilhan, E., . . . Unver, T. (2015). Hexaploid wheat (*Triticum aestivum*) root miRNome analysis in response to salt stress. *Annals of Applied Biology, 167*(2), 208-216.

Esposito, A., Colantuono, C., Ruggieri, V., & Chiusano, M. L. (2016). Bioinformatics for agriculture in the Next-Generation sequencing era. *Chemical and Biological Technologies in Agriculture, 3*(1), 9.

Fang, Z., & Cui, X. (2011). Design and validation issues in RNA-seq experiments. *Briefings in Bioinformatics, 12*(3), 280-287.

Fedurco, M., Romieu, A., Williams, S., Lawrence, I., & Turcatti, G. (2006). BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. *Nucleic Acids Research, 34*(3), e22.

Fei, Y., Xue, Y., Du, P., Yang, S., & Deng, X. (2017). Expression analysis and promoter methylation under osmotic and salinity stress of TaGAPC1 in wheat (*Triticum aestivum* L). *Protoplasma, 254*(2), 987-996.

Friedlander, M. R., Mackowiak, S. D., Li, N., Chen, W., & Rajewsky, N. (2012). miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Research, 40*(1), 37-52.

Fukudome, A., & Fukuhara, T. (2017). Plant dicer-like proteins: Double-stranded RNA-cleaving enzymes for small RNA biogenesis. *Journal of Plant Research, 130*(1), 33-44.

Fultz, D., Choudury, S. G., & Slotkin, R. K. (2015). Silencing of active transposable elements in plants. *Current Opinion in Plant Biology, 27*, 67-76.

Gao, J. P., Chao, D. Y., & Lin, H. X. (2007). Understanding abiotic stress tolerance mechanisms: Recent studies on stress response in Rice. *Journal of Integrative Plant Biology, 49*(6), 742-750.

Garber, M., Grabherr, M. G., Guttman, M., & Trapnell, C. (2011). Computational methods for transcriptome annotation and quantification using RNA-seq. *Nature Methods, 8*(6), 469-477.

Garg, R., Narayana Chevala, V. V. S., Shankar, R., & Jain, M. (2015). Divergent DNA methylation patterns associated with gene expression in Rice cultivars with contrasting drought and salinity stress response. *Scientific Reports, 5*, 14922.

Gauggel, C. A., Sierra, F., & Arévalo, G. (2005). *The problem of banana root deterioration and its impact on production: Latin America's experience.* Montpellier.

Gayral, P., Noa-Carrazana, J. C., Lescot, M., Lheureux, F., Lockhart, B. E., Matsumoto, T., . . . Iskra-Caruana, M. L. (2008). A single banana streak virus integration event in the banana genome as the origin of infectious endogenous pararetrovirus. *Journal of Virology, 82*(13), 6697-6710.

Gehring, M., & Henikoff, S. (2007). DNA methylation dynamics in plant genomes. *Biochimica et Biophysica Acta (BBA) - Bioenergetics, 1769*(5-6), 276-286.

Gehring, M. (2016). Prodigious plant methylomes. *Genome Biology, 17*(1), 197.

Gent, J. I., Ellis, N. A., Guo, L., Harkess, A. E., Yao, Y., Zhang, X., & Dawe, R. K. (2013). CHH islands: De novo DNA methylation in near-gene chromatin regulation in Maize. *Genome Research, 23*(4), 628-637.

German, M. A., Pillay, M., Jeong, D.-H., Hetawal, A., Luo, S., Janardhanan, P., . . . German, R. (2008). Global identification of microRNA–target RNA pairs by parallel analysis of RNA ends. *Nature Biotechnology, 26*(8), 941.

German, M. A., Luo, S., Schroth, G., Meyers, B. C., & Green, P. J. (2009). Construction of Parallel Analysis of RNA Ends (PARE) libraries for the study of cleaved miRNA targets and the RNA degradome. *Nature Protocols, 4*(3), 356-362.

Ghag, S. B., Shekhawat, U. K., & Ganapathi, T. R. (2014). Host-induced post-transcriptional hairpin RNA-mediated gene silencing of vital fungal genes confers efficient resistance against Fusarium wilt in banana. *Plant Biotechnology Journal, 12*(5), 541-553.

Gomes, C. P., Cho, J. H., Hood, L., Franco, O. L., Pereira, R. W., & Wang, K. (2013). A review of computational tools in microRNA discovery. *Frontiers in Genetics, 4*, 81.

Gonzalez, D. H. (2016). *Plant transcription factors: Evolutionary, structural and functional aspects* (Vol. 1). Amsterdam: Academic Press, an imprint of Elsevier.

Guo, M., Liu, J. H., Ma, X., Luo, D. X., Gong, Z. H., & Lu, M. H. (2016). The plant heat stress transcription factors (HSFs): structure, regulation, and function in response to abiotic stresses. *Frontiers in Plant Science, 7*(114).

Gutierrez, L., Bussell, J. D., Pacurar, D. I., Schwambach, J., Pacurar, M., & Bellini, C. (2009). Phenotypic plasticity of adventitious rooting in Arabidopsis is controlled by complex regulation of AUXIN RESPONSE FACTOR transcripts and microRNA abundance. *Plant Cell, 21*(10), 3119-3132.

Hackenberg, M., Rodriguez-Ezpeleta, N., & Aransay, A. M. (2011). MiRanalyzer: An update on the detection and analysis of microRNAs in high-throughput sequencing experiments. *Nucleic Acids Research, 39*(Web Server issue), W132-W138.

Hansen, K. D., Brenner, S. E., & Dudoit, S. (2010). Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Research, 38*(12), e131.

Harbison, C. T., Gordon, D. B., Lee, T. I., Rinaldi, N. J., Macisaac, K. D., Danford, T. W., . . . Young, R. A. (2004). Transcriptional regulatory code of a eukaryotic genome. *Nature, 431*(7004), 99-104.

Heslop-Harrison, J. S., & Schwarzacher, T. (2007). Domestication, genomics and the future for banana. *Annals of Botany, 100*(5), 1073-1084.

Hobert, O. (2004). Common logic of transcription factor and microRNA action. *Trends in Biochemical Sciences, 29*(9), 462-468.

Hofacker, I. L. (2003). Vienna RNA secondary structure server. *Nucleic Acids Research, 31*(13), 3429-3431.

Hofmann, N. R. (2012). A global view of hybrid vigor: DNA methylation, small RNAs, and gene expression. *Plant Cell, 24*(3), 841.

Holoch, D., & Moazed, D. (2015). RNA-mediated epigenetic regulation of gene expression. *Nature Reviews: Genetics, 16*(2), 71-84.

Hribova, E., Dolezelova, M., Town, C. D., Macas, J., & Dolezel, J. (2007). Isolation and characterization of the highly repeated fraction of the banana genome. *Cytogenetic and Genome Research, 119*(3-4), 268-274.

Hribova, E., Neumann, P., Matsumoto, T., Roux, N., Macas, J., & Dolezel, J. (2010). Repetitive part of the banana (*Musa acuminata*) genome investigated by low-depth 454 sequencing. *BMC Plant Biology, 10*, 204.

Hu, W., Zuo, J., Hou, X., Yan, Y., Wei, Y., Liu, J., . . . Jin, Z. (2015). The auxin response factor gene family in banana: Genome-wide identification and expression analyses during development, ripening, and abiotic stress. *Frontiers in Plant Science, 6*, 742.

Illumina, I. (2017a). RNA SEQUENCING METHODS COLLECTION. Retrieved 5 September 2018 from https://www.illumina.com/content/dam/illumina-marketing/documents/products/research_reviews/rna-sequencing-methods-review-web.pdf.

Illumina, I. (2017b). DNA SEQUENCING METHODS COLLECTION. Retrieved 5 September 2018 from https://www.illumina.com/content/dam/illumina-marketing/documents/products/research_reviews/dna-sequencing-methods-review-web.pdf.

Iskra-Caruana, M. L., Chabannes, M., Duroy, P. O., & Muller, E. (2014). A possible scenario for the evolution of banana streak virus in banana. *Virus Research, 186*, 155-162.

Israeli, Y., Lahav, E., & Nameri, N. (1986). The effect of salinity and sodium adsorption ratio in the irrigation water, on growth and productivity of bananas under drip irrigation conditions. *Fruits, 41*(5), 297-302.

Izadi, F., Nikfekr, R., & Soorni, J. (2017). Transcription factor-microRNAs regulatory network in response to multiple stresses in'*Arabidopsis thaliana*'. *Plant Omics, 10*(4), 183.

Jalil, M., Khalid, N., & Othman, R. Y. (2003). Plant regeneration from embryogenic suspension cultures of *Musa acuminata* cv. Mas (AA). *Plant Cell, Tissue and Organ Culture, 75*(3), 209-214.

Jingyi, W., Juhua, L., Caihong, J., Hongxia, M., Jianbin Zhang, Z. W., Biyu, X., & Zhiqiang, J. (2017). Cold stress responsive microRNAs and their targets in *Musa balbisiana*. *Frontiers of Agricultural Science and Engineering, 3*(4), 335-345.

Kakrana, A., Hammond, R., Patel, P., Nakano, M., & Meyers, B. C. (2014). sPARTA: A parallelized pipeline for integrated analysis of plant miRNA and cleaved mRNA data sets, including new miRNA target-identification software. *Nucleic Acids Research, 42*(18), e139-e139.

Kalvari, I., Argasinska, J., Quinones-Olvera, N., Nawrocki, E. P., Rivas, E., Eddy, S. R., . . . Petrov, A. I. (2018). Rfam 13.0: Shifting to a genome-centric resource for non-coding RNA families. *Nucleic Acids Research, 46*(D1), D335-D342.

Kamthan, A., Chaudhuri, A., Kamthan, M., & Datta, A. (2015). Small RNAs in plants: Recent development and application for crop improvement. *Frontiers in Plant Science, 6*, 208.

Kannambath, S. (2016). Micro-RNA feedback loops modulating the calcineurin/NFAT signaling pathway. *Noncoding RNA, 2*(2), 3.

Kasschau, K. D., Fahlgren, N., Chapman, E. J., Sullivan, C. M., Cumbie, J. S., Givan, S. A., & Carrington, J. C. (2007). Genome-wide profiling and analysis of Arabidopsis siRNAs. *PLoS Biology, 5*(3), e57.

Kaufmann, K., Muino, J. M., Osteras, M., Farinelli, L., Krajewski, P., & Angenent, G. C. (2010). Chromatin immunoprecipitation (ChIP) of plant transcription factors followed by sequencing (ChIP-SEQ) or hybridization to whole genome arrays (ChIP-CHIP). *Nature Protocols, 5*(3), 457-472.

Kayat, F., Ayob, M. A., Idris, A., Ibrahim, M., JM, S., Yusuf, M. A., . . . Zulariff, A. (2016). *Study of the intention of banana growers in improving the production in Jeli, Kelantan*. Paper presented at the Second Asia Pacific Conference on Advanced Research (APCAR) 2016, Hotel Grand Chancellor, Melbourne, Australia.

Khraiwesh, B., Zhu, J. K., & Zhu, J. (2012). Role of miRNAs and siRNAs in biotic and abiotic stress responses of plants. *Biochimica et Biophysica Acta (BBA) - Bioenergetics, 1819*(2), 137-148.

Kiefer, E., Heller, W., & Ernst, D. (2000). A simple and efficient protocol for isolation of functional RNA from plant tissues rich in secondary metabolites. *Plant Molecular Biology Reporter, 18*(1), 33-39.

Kinsella, R. J., Kahari, A., Haider, S., Zamora, J., Proctor, G., Spudich, G., . . . Flicek, P. (2011). Ensembl BioMarts: A hub for data retrieval across taxonomic space. *Database: The Journal of Biological Databases and Curation, 2011*, bar030.

Knop, K., Stepien, A., Barciszewska-Pacak, M., Taube, M., Bielewicz, D., Michalak, M., . . . Szweykowska-Kulinska, Z. (2016). Active 5' splice sites regulate the biogenesis efficiency of Arabidopsis microRNAs derived from intron-containing genes. *Nucleic Acids Research, 45*(5), 2757–2775.

Kolovos, P., Knoch, T. A., Grosveld, F. G., Cook, P. R., & Papantonis, A. (2012). Enhancers and silencers: An integrated and simple model for their function. *Epigenetics Chromatin, 5*(1), 1.

Konate, M., Wilkinson, M., Mayne, B., Pederson, S., Scott, E., Berger, B., & Rodriguez Lopez, C. (2018). Salt stress induces non-CG methylation in coding regions of Barley seedlings (*Hordeum vulgare*). *Epigenomes, 2*(2), 12.

Koren, S., Schatz, M. C., Walenz, B. P., Martin, J., Howard, J. T., Ganapathy, G., . . . Adam, M. P. (2012). Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nature Biotechnology, 30*(7), 693-700.

Kozomara, A., & Griffiths-Jones, S. (2014). miRBase: Annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Research, 42*(D1), D68-73.

Krol, J., Loedige, I., & Filipowicz, W. (2010). The widespread regulation of microRNA biogenesis, function and decay. *Nature Reviews: Genetics, 11*(9), 597-610.

Krueger, F., & Andrews, S. R. (2011). Bismark: A flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics, 27*(11), 1571-1572.

Krzywinska, E., Kulik, A., Bucholc, M., Fernandez, M. A., Rodriguez, P. L., & Dobrowolska, G. (2016). Protein phosphatase type 2C PP2CA together with ABI1 inhibits SnRK2.4 activity and regulates plant responses to salinity. *Plant Signaling & Behavior, 11*(12), e1253647.

Kurtoglu, K. Y., Kantar, M., Lucas, S. J., & Budak, H. (2013). Unique and conserved microRNAs in Wheat chromosome 5D revealed by next-generation sequencing. *PLoS ONE, 8*(7), e69801.

Kurtoglu, K. Y., Kantar, M., & Budak, H. (2014). New Wheat microRNA using whole-genome sequence. *Functional & Integrative Genomics, 14*(2), 363-379.

Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology, 10*(3), R25.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods, 9*(4), 357-359.

Law, J. A., & Jacobsen, S. E. (2010). Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nature Reviews: Genetics, 11*(3), 204-220.

Lee, J., Li, Z. H., Brower-Sinning, R., & John, B. (2007). Regulatory circuit of human microRNA biogenesis. *PLoS Computational Biology, 3*(4), 721-732.

Lee, J., Ji, Y., Liang, S., Cai, G., & Muller, P. (2011). On differential gene expression using RNA-Seq data. *Cancer Informatics, 10*, 205-215.

Lee, R. C., Feinbaum, R. L., & Ambros, V. (1993). The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell, 75*(5), 843-854.

Lee, W. S., Gudimella, R., Wong, G. R., Tammi, M. T., Khalid, N., & Harikrishna, J. A. (2015). Transcripts and microRNAs responding to salt stress in *Musa acuminata* Colla (AAA Group) cv. Berangan roots. *PLoS ONE, 10*(5), e0127526.

Leinonen, R., Sugawara, H., & Shumway, M. (2011). The sequence read archive. *Nucleic Acids Research, 39*(Database issue), D19-21.

Leipzig, J. (2017). A review of bioinformatic pipeline frameworks. *Briefings in Bioinformatics, 18*(3), 530-536.

Lemonnier, P., Gaillard, C., Veillet, F., Verbeke, J., Lemoine, R., Coutos-Thevenot, P., & La Camera, S. (2014). Expression of Arabidopsis sugar transport protein STP13 differentially affects glucose transport activity and basal resistance to Botrytis cinerea. *Plant Molecular Biology, 85*(4-5), 473-484.

Lescot, M., Piffanelli, P., Ciampi, A. Y., Ruiz, M., Blanc, G., Leebens-Mack, J., . . . Town, C. D. (2008). Insights into the *Musa* genome: Syntenic relationships to Rice and between *Musa* species. *BMC Genomics, 9*.

Levy, S. E., & Myers, R. M. (2016). Advancements in next-generation sequencing. *Annual Review of Genomics and Human Genetics, 17*, 95-115.

Lewsey, M. G., Hardcastle, T. J., Melnyk, C. W., Molnar, A., Valli, A., Urich, M. A., . . . Ecker, J. R. (2016). Mobile small RNAs regulate genome-wide DNA methylation. *Proceedings of the National Academy of Sciences of the United States of America, 113*(6), E801-810.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics, 25*(14), 1754-1760.

Li, J., Guo, G., Guo, W., Guo, G., Tong, D., Ni, Z., . . . Yao, Y. (2012). miRNA164-directed cleavage of ZmNAC1 confers lateral root development in Maize (*Zea mays* L.). *BMC Plant Biology, 12*, 220.

Li, S. (2015). The *Arabidopsis thaliana* TCP transcription factors: A broadening horizon beyond development. *Plant Signaling & Behavior, 10*(7), e1044192.

Liao, Y., Liu, S., Jiang, Y., Hu, C., Zhang, X., Cao, X., . . . Zhu, J. (2017). Genome-wide analysis and environmental response profiling of dirigent family genes in Rice (*Oryza sativa*). *Genes & Genomics, 39*(1), 47-62.

Lifanov, A. P., Makeev, V. J., Nazina, A. G., & Papatsenko, D. A. (2003). Homotypic regulatory clusters in Drosophila. *Genome Research, 13*(4), 579-588.

Lister, R., O'Malley, R. C., Tonti-Filippini, J., Gregory, B. D., Berry, C. C., Millar, A. H., & Ecker, J. R. (2008). Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell, 133*(3), 523-536.

Liu, G., Xia, Y., Liu, T., Dai, S., & Hou, X. (2018). The DNA methylome and association of differentially methylated regions with differential gene expression during heat stress in *Brassica rapa*. *International Journal of Molecular Sciences, 19*(5).

Liu, H., Ma, L., Yang, X., Zhang, L., Zeng, X., Xie, S., . . . Shen, Y. (2017). Integrative analysis of DNA methylation, mRNAs, and small RNAs during Maize embryo dedifferentiation. *BMC Plant Biology, 17*(1), 105.

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., . . . Law, M. (2012). Comparison of next-generation sequencing systems. *BioMed Research International, 2012*.

Lu, C., Tej, S. S., Luo, S., Haudenschild, C. D., Meyers, B. C., & Green, P. J. (2005). Elucidation of the small RNA component of the transcriptome. *Science, 309*(5740), 1567-1569.

Lunardon, A., Forestan, C., Farinati, S., Axtell, M. J., & Varotto, S. (2016). Genome-wide characterization of Maize small RNA loci and their regulation in the required to maintain repression6-1 (rmr6-1) mutant and long-term abiotic stresses. *Plant Physiology, 170*(3), 1535-1548.

Magalhaes, J. V. (2010). How a microbial drug transporter became essential for crop cultivation on acid soils: aluminium tolerance conferred by the multidrug and toxic compound extrusion (MATE) family. *Annals of Botany, 106*(1), 199-203.

Mahouachi, J. (2007). Growth and mineral nutrient content of developing fruit on banana plants (*Musa acuminata* AAA,'Grand Nain') subjected to water stress and recovery. *Journal of Horticultural Science & Biotechnology, 82*(6), 839-844.

Manohar, M., Shigaki, T., Mei, H., Park, S., Marshall, J., Aguilar, J., & Hirschi, K. D. (2011). Characterization of Arabidopsis Ca2+/H+ exchanger CAX3. *Biochemistry, 50*(28), 6189-6195.

Martienssen, R. A. (2003). Maintenance of heterochromatin by RNA interference of tandem repeats. *Nature Genetics, 35*(3), 213-214.

Martin, G., Baurens, F. C., Droc, G., Rouard, M., Cenci, A., Kilian, A., . . . D'Hont, A. (2016). Improvement of the banana "*Musa acuminata*" reference sequence using NGS data and semi-automated bioinformatics methods. *BMC Genomics, 17*(1), 243.

Martinez, N. J., & Walhout, A. J. (2009). The interplay between transcription factors and microRNAs in genome-scale regulatory networks. *Bioessays, 31*(4), 435-445.

Matzke, M., Kanno, T., Huettel, B., Daxinger, L., & Matzke, A. J. (2007). Targets of RNA-directed DNA methylation. *Current Opinion in Plant Biology, 10*(5), 512-519.

Matzke, M., Kanno, T., Daxinger, L., Huettel, B., & Matzke, A. J. (2009). RNA-mediated chromatin-based silencing in plants. *Current Opinion in Cell Biology, 21*(3), 367-376.

Matzke, M. A., & Mosher, R. A. (2014). RNA-directed DNA methylation: An epigenetic pathway of increasing complexity (vol 15, 394, 2014). *Nature Reviews Genetics, 15*(8).

Matzke, M. A., Kanno, T., & Matzke, A. J. (2015). RNA-Directed DNA Methylation: The evolution of a complex epigenetic pathway in flowering plants. *Annual Review of Plant Biology, 66*, 243-267.

Megraw, M., Baev, V., Rusinov, V., Jensen, S. T., Kalantidis, K., & Hatzigeorgiou, A. G. (2006). MicroRNA promoter element discovery in Arabidopsis. *RNA, 12*(9), 1612-1619.

Megraw, M., & Hatzigeorgiou, A. G. (2010). MicroRNA promoter analysis. *Methods in Molecular Biology, 592*, 149-161.

Megraw, M., Cumbie, J. S., Ivanchenko, M. G., & Filichkin, S. A. (2016). Small genetic circuits and MicroRNAs: Big players in polymerase II transcriptional control in plants. *Plant Cell, 28*(2), 286-303.

Meissner, A., Gnirke, A., Bell, G. W., Ramsahoye, B., Lander, E. S., & Jaenisch, R. (2005). Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Research, 33*(18), 5868-5877.

Meng, Y., Ma, X., Chen, D., Wu, P., & Chen, M. (2010). MicroRNA-mediated signaling involved in plant root development. *Biochemical and Biophysical Research Communications, 393*(3), 345-349.

Meng, Y., Shao, C., & Chen, M. (2011). Toward microRNA-mediated gene regulatory networks in plants. *Briefings in Bioinformatics, 12*(6), 645-659.

Meng, Y., Shao, C., Ma, X., & Wang, H. (2013). Introns targeted by plant microRNAs: A possible novel mechanism of gene regulation. *Rice (N Y), 6*(1), 8.

Meyers, B. C., Axtell, M. J., Bartel, B., Bartel, D. P., Baulcombe, D., Bowman, J. L., . . . Zhu, J. K. (2008). Criteria for annotation of plant MicroRNAs. *Plant Cell, 20*(12), 3186-3190.

Mikheyev, A. S., & Tin, M. M. (2014). A first look at the Oxford Nanopore MinION sequencer. *Molecular Ecology Resources, 14*(6), 1097-1102.

Mirouze, M., & Vitte, C. (2014). Transposable elements, a treasure trove to decipher epigenetic variation: insights from Arabidopsis and crop epigenomes. *Journal of Experimental Botany, 65*(10), 2801-2812.

Mochida, K., & Shinozaki, K. (2011). Advances in omics and bioinformatics tools for systems analyses of plant functions. *Plant & Cell Physiology, 52*(12), 2017-2038.

Mokhtarud-din, H., & William, R. (2011). Status of banana cultivation and disease incidences in Malaysia, 5. Retrieved 28 September 2017 from http://www.itfnet.org/Download/Banana_Workshop2011/1.pdf.

Morea, E. G., da Silva, E. M., e Silva, G. F., Valente, G. T., Barrera Rojas, C. H., Vincentz, M., & Nogueira, F. T. (2016). Functional and evolutionary analyses of the miR156 and miR529 families in land plants. *BMC Plant Biology, 16*, 40.

Murashige, T., & Skoog, F. (1962). A revised medium for rapid growth and bio assays with tobacco tissue cultures. *Physiologia Plantarum, 15*(3), 473-497.

Muthusamy, M., Uma, S., Backiyarani, S., Saraswathi, M. S., & Chandrasekar, A. (2016). Transcriptomic changes of drought-tolerant and sensitive banana cultivars exposed to drought stress. *Frontiers in Plant Science, 7*, 1609.

Niederhuth, C. E., Bewick, A. J., Ji, L. X., Alabady, M. S., Kim, K. D., Li, Q., . . . Schmitz, R. J. (2016). Widespread natural variation of DNA methylation within angiosperms. *Genome Biology, 17*(1), 194.

Nova-Franco, B., Iniguez, L. P., Valdes-Lopez, O., Alvarado-Affantranger, X., Leija, A., Fuentes, S. I., . . . Hernandez, G. (2015). The micro-RNA72c-APETALA2-1 node as a key regulator of the common bean-*Rhizobium etli* nitrogen fixation symbiosis. *Plant Physiology, 168*(1), 273-291.

Novak, P., Hribova, E., Neumann, P., Koblizkova, A., Dolezel, J., & Macas, J. (2014). Genome-wide analysis of repeat diversity across the family Musaceae. *PLoS ONE, 9*(6), e98918.

Nozawa, M., Miura, S., & Nei, M. (2012). Origins and evolution of microRNA genes in plant species. *Genome Biology and Evolution, 4*(3), 230-239.

Nuthikattu, S., McCue, A. D., Panda, K., Fultz, D., DeFraia, C., Thomas, E. N., & Slotkin, R. K. (2013). The initiation of epigenetic silencing of active transposable elements is triggered by RDR6 and 21-22 nucleotide small interfering RNAs. *Plant Physiology, 162*(1), 116-131.

Ohashi, H., Hasegawa, M., Wakimoto, K., & Miyamoto-Sato, E. (2015). Next-generation technologies for multiomics approaches including interactome sequencing. *BioMed Research International, 2015*, 104209.

Pandey, G., Sharma, N., Sahu, P. P., & Prasad, M. (2016). Chromatin-based epigenetic regulation of plant sbiotic stress response. *Current Genomics, 17*(6), 490-498.

Park, Y., & Wu, H. (2016). Differential methylation analysis for BS-seq data under general experimental design. *Bioinformatics, 32*(10), 1446-1453.

Perrier, X., De Langhe, E., Donohue, M., Lentfer, C., Vrydaghs, L., Bakry, F., . . . Denham, T. (2011). Multidisciplinary perspectives on banana (*Musa* spp.) domestication. *Proceedings of the National Academy of Sciences of the United States of America, 108*(28), 11311-11318.

Petroni, K., Kumimoto, R. W., Gnesutta, N., Calvenzani, V., Fornari, M., Tonelli, C., . . . Mantovani, R. (2012). The promiscuous life of plant nuclear factor y transcription factors. *Plant Cell, 24*(12), 4777-4792.

Pikaard, C. S., & Mittelsten Scheid, O. (2014). Epigenetic regulation in plants. *Cold Spring Harbor Perspectives in Biology, 6*(12), a019315.

Pineau, C., Loubet, S., Lefoulon, C., Chalies, C., Fizames, C., Lacombe, B., . . . Richard, O. (2012). Natural Variation at the FRD3 MATE transporter locus reveals cross-talk between Fe homeostasis and Zn tolerance in *Arabidopsis thaliana*. *PLoS Genetics, 8*(12), e1003120.

Popova, O. V., Dinh, H. Q., Aufsatz, W., & Jonak, C. (2013). The RdDM pathway is required for basal heat tolerance in Arabidopsis. *Molecular Plant, 6*(2), 396-410.

Qin, J. P., Ma, X. X., Yi, Z. L., Tang, Z. H., & Meng, Y. J. (2015). Intronic regions of plant genes potentially encode RDR (RNA-dependent RNA polymerase)-dependent small RNAs. *Journal of Experimental Botany, 66*(7), 1763-1768.

Qin, Z. R., Li, C. L., Mao, L., & Wu, L. (2014). Novel insights from non-conserved microRNAs in plants. *Frontiers in Plant Science, 5*(586).

Qiu, C., Wang, J., Yao, P., Wang, E., & Cui, Q. (2010). microRNA evolution in a human transcription factor and microRNA regulatory network. *BMC Systems Biology, 4*, 90.

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics, 26*(6), 841-842.

Rajeswaran, R., Seguin, J., Chabannes, M., Duroy, P. O., Laboureau, N., Farinelli, L., . . . Pooggin, M. M. (2014). Evasion of short interfering RNA-directed antiviral silencing in *Musa acuminata* persistently infected with six distinct banana streak pararetroviruses. *Journal of Virology, 88*(19), 11516-11528.

Ravi, I., & Vaganan, M. M. (2016). Abiotic stress tolerance in banana. In S. K. Rao N., Laxman R. (Ed.), *Abiotic Stress Physiology of Horticultural Crops* (pp. 207-222). New Delhi: Springer.

Reece-Hoyes, J. S., & Marian Walhout, A. J. (2012). Yeast one-hybrid assays: A historical and technical perspective. *Methods, 57*(4), 441-447.

Regulski, M., Lu, Z. Y., Kendall, J., Donoghue, M. T. A., Reinders, J., Llaca, V., . . . Martienssen, R. A. (2013). The maize methylome influences mRNA splice sites and reveals widespread paramutation-like switches guided by small RNA. *Genome Research, 23*(10), 1651-1662.

Reinhart, B. J., Weinstein, E. G., Rhoades, M. W., Bartel, B., & Bartel, D. P. (2002). MicroRNAs in plants. *Genes & Development, 16*(13), 1616-1626.

Riechmann, J. L., Heard, J., Martin, G., Reuber, L., Jiang, C. Z., Keddie, J., . . . Yu, C. L. (2000). Arabidopsis transcription factors: Genome-wide comparative analysis among eukaryotes. *Science, 290*(5499), 2105-2110.

Robinson, J. C., & Saúco, V. G. (2010). *Bananas and plantains* (2nd ed. Vol. 19). UK: Cab International.

Rodriguez, R. E., Ercoli, M. F., Debernardi, J. M., Breakfield, N. W., Mecchia, M. A., Sabatini, M., . . . Palatnik, J. F. (2015). MicroRNA miR396 regulates the switch between stem cells and transit-amplifying cells in Arabidopsis roots. *Plant Cell, 27*(12), 3354-3366.

Safar, J., Noa-Carrazana, J. C., Vrana, J., Bartos, J., Alkhimova, O., Sabau, X., . . . Piffanelli, P. (2004). Creation of a BAC resource to study the structure and evolution of the banana (*Musa balbisiana*) genome. *Genome, 47*(6), 1182-1191.

Sahu, P. P., Pandey, G., Sharma, N., Puranik, S., Muthamilarasan, M., & Prasad, M. (2013). Epigenetic mechanisms of plant stress responses and adaptation. *Plant Cell Reports, 32*(8), 1151-1159.

Saraswathi, M., Uma, S., Prasanya Selvam, K., Ramaraj, S., Durai, P., & Mustaffa, M. (2011). Assessing the robustness of IRAP and RAPD marker systems to study intra-group diversity among Cavendish (AAA) clones of banana. *The Journal of Horticultural Science and Biotechnology, 86*(1), 7-12.

Schmitz, R. J., He, Y. P., Valdes-Lopez, O., Khan, S. M., Joshi, T., Urich, M. A., . . . Ecker, J. R. (2013). Epigenome-wide inheritance of cytosine methylation variants in a recombinant inbred population. *Genome Research, 23*(10), 1663-1674.

Schurch, N. J., Schofield, P., Gierlinski, M., Cole, C., Sherstnev, A., Singh, V., . . . Barton, G. J. (2016). How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? *RNA, 22*(6), 839-851.

Sewelam, N., Kazan, K., & Schenk, P. M. (2016). Global plant stress signaling: Reactive oxygen species at the cross-road. *Frontiers in Plant Science, 7*, 187.

Shahmuradov, I. A., Umarov, R. K., & Solovyev, V. V. (2017). TSSPlant: A new tool for prediction of plant Pol II promoters. *Nucleic Acids Research, 45*(8), e65.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., . . . Ideker, T. (2003). Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research, 13*(11), 2498-2504.

Shapira, O., Khadka, S., Israeli, Y., Shani, U., & Schwartz, A. (2009). Functional anatomy controls ion distribution in banana leaves: significance of Na+ seclusion at the leaf margins. *Plant Cell and Environment, 32*(5), 476-485.

Sharrock, S., & Frison, E. (1998). *Musa* production around the world-trends, varieties and regional importance. *INIBAP annual report*, 42-47.

Shekhawat, U. K., Srinivas, L., & Ganapathi, T. R. (2011). MusaDHN-1, a novel multiple stress-inducible SK(3)-type dehydrin gene, contributes affirmatively to drought- and salt-stress tolerance in banana. *Planta, 234*(5), 915-932.

Shrivastava, P., & Kumar, R. (2015). Soil salinity: A serious environmental issue and plant growth promoting bacteria as one of the tools for its alleviation. *Saudi Journal of Biological Sciences, 22*(2), 123-131.

Simmonds, N., & Weatherup, S. (1990). Numerical taxonomy of the cultivated bananas. *Tropical Agriculture, 67*(1), 90-92.

Simmonds, N. W., & Shepherd, K. (1955). The taxonomy and origins of the cultivated bananas. *Botanical Journal of the Linnean Society, 55*(359), 302-312.

Simmonds, N. W. (1962). *The evolution of the bananas*. London: Longmans, Green.

Singh, I., Smita, S., Mishra, D. C., Kumar, S., Singh, B. K., & Rai, A. (2017). Abiotic stress responsive miRNA-target network and related markers (SNP, SSR) in *Brassica juncea. Frontiers in Plant Science, 8*, 1943.

Singh, P., Sweta, K., Mohanapriya, A., Sudandiradoss, C., Siva, R., Gothandam, K. M., & Babu, S. (2015). Homotypic clustering of OsMYB4 binding site motifs in promoters of the Rice genome and cellular-level implications on sheath blight disease resistance. *Gene, 561*(2), 209-218.

Siriwardana, C. L., Gnesutta, N., Kumimoto, R. W., Jones, D. S., Myers, Z. A., Mantovani, R., & Holt, B. F., 3rd. (2016). NUCLEAR FACTOR Y, subunit a (NF-YA) proteins positively regulate flowering and act through FLOWERING LOCUS T. *PLoS Genetics, 12*(12), e1006496.

Song, J., Bradeen, J. M., Naess, S. K., Raasch, J. A., Wielgus, S. M., Haberlach, G. T., . . . Jiang, J. (2003). Gene RB cloned from *Solanum bulbocastanum* confers broad spectrum resistance to potato late blight. *Proceedings of the National Academy of Sciences of the United States of America, 100*(16), 9128-9133.

Song, Q., Decato, B., Hong, E. E., Zhou, M., Fang, F., Qu, J., . . . Smith, A. D. (2013a). A reference methylome database and analysis pipeline to facilitate integrative and comparative epigenomics. *PLoS ONE, 8*(12), e81148.

Song, Q. X., Lu, X., Li, Q. T., Chen, H., Hu, X. Y., Ma, B., . . . Zhang, J. S. (2013b). Genome-wide analysis of DNA methylation in Soybean. *Molecular Plant, 6*(6), 1961-1974.

Song, S., Chen, X., Huang, D., Xu, Y., Zeng, H., Hu, X., & Wang, W. (2016). Identification of miRNAs differentially expressed in Fusarium wilt-resistant and susceptible banana varieties. *South African Journal of Botany*, *106*, 244-249.

Song, Y. G., Ji, D. D., Li, S., Wang, P., Li, Q., & Xiang, F. N. (2012). The dynamic changes of DNA methylation and histone modifications of salt responsive transcription factor genes in Soybean. *PLoS ONE, 7*(7), e41274.

Stewart, C. N. (1997). Rapid DNA extraction from plants. In B. R. Micheli M.R. (Ed.), *Fingerprinting methods based on arbitrarily primed PCR* (pp. 25-28). Berlin, Heidelberg: Springer.

Stocks, M. B., Moxon, S., Mapleson, D., Woolfenden, H. C., Mohorianu, I., Folkes, L., . . . Moulton, V. (2012). The UEA sRNA workbench: A suite of tools for analysing and visualizing next generation sequencing microRNA and small RNA datasets. *Bioinformatics, 28*(15), 2059-2061.

Stroud, H., Do, T., Du, J., Zhong, X., Feng, S., Johnson, L., . . . Jacobsen, S. E. (2014). Non-CG methylation patterns shape the epigenetic landscape in Arabidopsis. *Nature Structural & Molecular Biology, 21*(1), 64-72.

Sunkar, R., Chinnusamy, V., Zhu, J. H., & Zhu, J. K. (2007). Small RNAs as big players in plant abiotic stress responses and nutrient deprivation. *Trends in Plant Science, 12*(7), 301-309.

Sunkar, R., Li, Y. F., & Jagadeeswaran, G. (2012). Functions of microRNAs in plant stress responses. *Trends in Plant Science, 17*(4), 196-203.

Takahashi, S., Osabe, K., Fukushima, N., Takuno, S., Miyaji, N., Shimizu, M., . . . Fujimoto, R. (2018). Genome-wide characterization of DNA methylation, small RNA expression, and histone H3 lysine nine di-methylation in *Brassica rapa* L. *DNA Research, 25*(5), 511-520.

Thakur, V., Wanchana, S., Xu, M., Bruskiewich, R., Quick, W. P., Mosig, A., & Zhu, X. G. (2011). Characterization of statistical features for plant microRNA prediction. *BMC Genomics, 12*, 108.

Ticconi, C. A., Lucero, R. D., Sakhonwasee, S., Adamson, A. W., Creff, A., Nussaume, L., . . . Abel, S. (2009). ER-resident proteins PDR2 and LPR1 mediate the developmental response of root meristems to phosphate availability. *Proceedings of the National Academy of Sciences of the United States of America, 106*(33), 14174-14179.

To, T. K., Kim, J. M., Matsui, A., Kurihara, Y., Morosawa, T., Ishida, J., . . . Seki, M. (2011). Arabidopsis HDA6 regulates locus-directed heterochromatin silencing in cooperation with MET1. *PLoS Genetics, 7*(4), e1002055.

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., . . . Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology, 28*(5), 511-515.

Tripathi, L., Odipio, J., Tripathi, J. N., & Tusiime, G. (2008). A rapid technique for screening banana cultivars for resistance to Xanthomonas wilt. *European Journal of Plant Pathology, 121*(1), 9-19.

Turner, D. W., Fortescue, Jeanie A., & Thomas, Dane S. (2007). Environmental physiology of the bananas (*Musa* spp.). *Brazilian Journal of Plant Physiology, 19(4)*, 463-484.

Unver, T., Namuth-Covert, D. M., & Budak, H. (2009). Review of current methodological approaches for characterizing microRNAs in plants. *International Journal of Plant Genomics, 2009*, 262463.

Unver, T., Bakar, M., Shearman, R. C., & Budak, H. (2010). Genome-wide profiling and analysis of *Festuca arundinacea* miRNAs and transcriptomes in response to foliar glyphosate application. *Molecular Genetics and Genomics, 283*(4), 397-413.

Van Asten, P., Fermont, A., & Taulya, G. (2011). Drought is a major yield loss factor for rainfed East African highland banana. *Agricultural Water Management, 98*(4), 541-552.

Vanhove, A. C., Vermaelen, W., Panis, B., Swennen, R., & Carpentier, S. C. (2012). Screening the banana biodiversity for drought tolerance: can an in vitro growth model and proteomics be used as a tool to discover tolerant varieties and understand homeostasis. *Frontiers in Plant Science, 3*, 176.

Vanyushin, B. F. (2006). DNA methylation in plants. *Current Topics in Microbiology and Immunology, 301*, 67-122.

Varshney, R. K., Terauchi, R., & McCouch, S. R. (2014). Harvesting the promising fruits of genomics: applying genome sequencing technologies to crop breeding. *PLoS Biology, 12*(6), e1001883.

Vilela, B., Najar, E., Lumbreras, V., Leung, J., & Pages, M. (2015). Casein kinase 2 negatively regulates abscisic acid-activated SnRK2s in the core abscisic acid-signaling module. *Molecular Plant, 8*(5), 709-721.

Wairegi, L. W. I., van Asten, P. J. A., Tenywa, M. M., & Bekunda, M. A. (2010). Abiotic constraints override biotic constraints in East African highland banana systems. *Field Crops Research, 117*(1), 146-153.

Wang, J., Hou, Q., Li, P., Yang, L., Sun, X., Benedito, V. A., . . . Zhao, J. (2017). Diverse functions of multidrug and toxin extrusion (MATE) transporters in citric acid efflux and metal homeostasis in *Medicago truncatula. Plant Journal, 90*(1), 79-95.

Wang, L., Feng, Z., Wang, X., Wang, X., & Zhang, X. (2010). DEGseq: An R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics, 26*(1), 136-138.

Wang, M. B., & Smith, N. A. (2016). Satellite RNA pathogens of plants: impacts and origins-an RNA silencing perspective. *Wiley Interdisciplinary Reviews: RNA, 7*(1), 5-16.

Wang, W. S., Huang, F., Qin, Q., Zhao, X. Q., Li, Z. K., & Fu, B. Y. (2015). Comparative analysis of DNA methylation changes in two Rice genotypes under salt stress and subsequent recovery. *Biochemical and Biophysical Research Communications, 465*(4), 790-796.

Wang, X., Elling, A. A., Li, X., Li, N., Peng, Z., He, G., . . . Deng, X. W. (2009). Genome-wide and organ-specific landscapes of epigenetic modifications and their relationships to mRNA and small RNA transcriptomes in Maize. *Plant Cell, 21*(4), 1053-1069.

Wassenegger, M., Heimes, S., Riedel, L., & Sanger, H. L. (1994). RNA-directed de novo methylation of genomic sequences in plants. *Cell, 76*(3), 567-576.

Weake, V. M., & Workman, J. L. (2010). Inducible gene expression: Diverse regulatory mechanisms. *Nature Reviews: Genetics, 11*(6), 426-437.

Weber, B., Zicola, J., Oka, R., & Stam, M. (2016). Plant enhancers: A call for discovery. *Trends in Plant Science, 21*(11), 974-987.

Weirauch, M. T., Yang, A., Albu, M., Cote, A. G., Montenegro-Montero, A., Drewe, P., . . . Hughes, T. R. (2014). Determination and inference of eukaryotic transcription factor sequence specificity. *Cell, 158*(6), 1431-1443.

Wendte, J. M., & Pikaard, C. S. (2017). The RNAs of RNA-directed DNA methylation. *Biochimica et Biophysica Acta (BBA) - Bioenergetics, 1860*(1), 140-148.

Wickham, H. (2009). *ggplot2: Elegant graphics for data analysis.* (2nd ed.). New York: Springer-Verlag

Wierzbicki, A. T., Cocklin, R., Mayampurath, A., Lister, R., Rowley, M. J., Gregory, B. D., . . . Pikaard, C. S. (2012). Spatial and functional relationships among Pol V-associated loci, Pol IV-dependent siRNAs, and cytosine methylation in the Arabidopsis epigenome. *Genes & Development, 26*(16), 1825-1836.

Willmann, M. R., Endres, M. W., Cook, R. T., & Gregory, B. D. (2011). The functions of RNA-Dependent RNA polymerases in Arabidopsis. *Arabidopsis Book, 9*, e0146.

Wilton, M., Subramaniam, R., Elmore, J., Felsensteiner, C., Coaker, G., & Desveaux, D. (2010). The type III effector HopF2Pto targets Arabidopsis RIN4 protein to promote *Pseudomonas syringae* virulence. *Proceedings of the National Academy of Sciences of the United States of America, 107*(5), 2349-2354.

Wu, H., Xu, T. L., Feng, H., Chen, L., Li, B., Yao, B., . . . Conneely, K. N. (2015). Detection of differentially methylated regions from whole-genome bisulfite sequencing data without replicates. *Nucleic Acids Research, 43*(21), e141.

Xie, W., Huang, J., Liu, Y., Rao, J., Luo, D., & He, M. (2015). Exploring potential new floral organ morphogenesis genes of *Arabidopsis thaliana* using systems biology approach. *Frontiers in Plant Science, 6*, 829.

Xu, R., Wang, Y. H., Zheng, H., Lu, W., Wu, C. G., Huang, J. G., . . . Zheng, C. C. (2015). Salt-induced transcription factor MYB74 is regulated by the RNA-directed DNA methylation pathway in Arabidopsis. *Journal of Experimental Botany, 66*(19), 5997-6008.

Xu, X., Hayashi, N., Wang, C.-T., Fukuoka, S., Kawasaki, S., Takatsuji, H., & Jiang, C.-J. (2014). Rice blast resistance gene Pikahei-1 (t), a member of a resistance gene cluster on chromosome 4, encodes a nucleotide-binding site and leucine-rich repeat protein. *Molecular Breeding, 34*(2), 691-700.

Yaish, M. W., Al-Lawati, A., Al-Harrasi, I., & Patankar, H. V. (2018). Genome-wide DNA Methylation analysis in response to salinity in the model plant caliph medic (*Medicago truncatula*). *BMC Genomics, 19*(1), 78.

Yan, H., Bombarely, A., Xu, B., Frazier, T. P., Wang, C., Chen, P., . . . Huang, L. (2018). siRNAs regulate DNA methylation and interfere with gene and lncRNA expression in the heterozygous polyploid switchgrass. *Biotechnology for Biofuels, 11*, 208.

Yang, D. L., Zhang, G., Tang, K., Li, J., Yang, L., Huang, H., . . . Zhu, J. K. (2016). Dicer-independent RNA-directed DNA methylation in Arabidopsis. *Cell Research, 26*(11), 1264.

Yang, Q. S., Gao, J., He, W. D., Dou, T. X., Ding, L. J., Wu, J. H., . . . Yi, G. J. (2015). Comparative transcriptomics analysis reveals difference of key gene expression between banana and plantain in response to cold stress. *BMC Genomics, 16*, 446.

Yano-Melo, A. M., Saggin, O. J., & Maia, L. C. (2003). Tolerance of mycorrhized banana (*Musa* sp cv. Pacovan) plantlets to saline stress. *Agriculture Ecosystems & Environment, 95*(1), 343-348.

Yao, Y., Ni, Z., Peng, H., Sun, F., Xin, M., Sunkar, R., . . . Sun, Q. (2010). Non-coding small RNAs responsive to abiotic stress in Wheat (*Triticum aestivum* L.). *Functional & Integrative Genomics, 10*(2), 187-190.

Yi, X., Zhang, Z., Ling, Y., Xu, W., & Su, Z. (2015). PNRD: A plant non-coding RNA database. *Nucleic Acids Research, 43*(Database issue), D982-989.

Yin, T., Cook, D., & Lawrence, M. (2012). ggbio: An R package for extending the grammar of graphics for genomic data. *Genome Biology, 13*(8), R77.

Yong, W. S., Hsu, F. M., & Chen, P. Y. (2016). Profiling genome-wide DNA methylation. *Epigenetics Chromatin, 9*, 26.

Yu, D., Ma, X., Zuo, Z., Shao, W., Wang, H., & Meng, Y. (2017). Bioinformatics resources for deciphering the biogenesis and action pathways of plant small RNAs. *Rice (N Y), 10*(1), 38.

Yu, N., Niu, Q. W., Ng, K. H., & Chua, N. H. (2015). The role of miR156/SPLs modules in Arabidopsis lateral root development. *Plant Journal, 83*(4), 673-685.

Yu, X., Li, L., Li, L., Guo, M., Chory, J., & Yin, Y. (2008). Modulation of brassinosteroid-regulated gene expression by Jumonji domain-containing proteins ELF6 and REF6 in Arabidopsis. *Proceedings of the National Academy of Sciences of the United States of America, 105*(21), 7618-7623.

Zemach, A., Kim, M. Y., Hsieh, P. H., Coleman-Derr, D., Eshed-Williams, L., Thao, K., . . . Zilberman, D. (2013). The Arabidopsis nucleosome remodeler DDM1 allows DNA methyltransferases to access H1-containing heterochromatin. *Cell, 153*(1), 193-205.

Zhai, J., Bischof, S., Wang, H., Feng, S., Lee, T. F., Teng, C., . . . Jacobsen, S. E. (2015). A one precursor One siRNA model for Pol IV-Dependent siRNA biogenesis. *Cell, 163*(2), 445-455.

Zhang, B., Pan, X., Cannon, C. H., Cobb, G. P., & Anderson, T. A. (2006a). Conservation and divergence of plant microRNA genes. *Plant Journal, 46*(2), 243-259.

Zhang, B. (2015). MicroRNA: A new target for improving plant tolerance to abiotic stress. *Journal of Experimental Botany, 66*(7), 1749-1761.

Zhang, M., Kimatu, J. N., Xu, K., & Liu, B. (2010a). DNA cytosine methylation in plant development. *Journal of Genetics and Genomics, 37*(1), 1-12.

Zhang, S., Xie, M., Ren, G., & Yu, B. (2013). CDC5, a DNA binding protein, positively regulates posttranscriptional processing and/or transcription of primary microRNA transcripts. *Proceedings of the National Academy of Sciences of the United States of America, 110*(43), 17588-17593.

Zhang, X., Yazaki, J., Sundaresan, A., Cokus, S., Chan, S. W., Chen, H., . . . Ecker, J. R. (2006b). Genome-wide high-resolution mapping and functional analysis of DNA methylation in Arabidopsis. *Cell, 126*(6), 1189-1201.

Zhang, X., Henderson, I. R., Lu, C., Green, P. J., & Jacobsen, S. E. (2007). Role of RNA polymerase IV in plant small RNA metabolism. *Proceedings of the National Academy of Sciences of the United States of America, 104*(11), 4536-4541.

Zhang, Z., Yu, J., Li, D., Zhang, Z., Liu, F., Zhou, X., . . . Su, Z. (2010b). PMRD: Plant microRNA database. *Nucleic Acids Research, 38*(Database issue), D806-813.

Zheng, X., Pontes, O., Zhu, J., Miki, D., Zhang, F., Li, W. X., . . . Zhu, J. K. (2008). ROS3 is an RNA-binding protein required for DNA demethylation in Arabidopsis. *Nature, 455*(7217), 1259-1262.

Zheng, Y., Li, Y. F., Sunkar, R., & Zhang, W. X. (2012). SeqTar: An effective method for identifying microRNA guided cleavage sites from degradome of polyadenylated transcripts in plants. *Nucleic Acids Research, 40*(4), e28.

Zhong, S. L., Fei, Z. J., Chen, Y. R., Zheng, Y., Huang, M. Y., Vrebalov, J., . . . Giovannoni, J. J. (2013). Single-base resolution methylomes of tomato fruit development reveal epigenome modifications associated with ripening. *Nature Biotechnology, 31*(2), 154-159.

Zhou, M., Li, D., Li, Z., Hu, Q., Yang, C., Zhu, L., & Luo, H. (2013). Constitutive expression of a miR319 gene alters plant development and enhances salt and drought tolerance in transgenic creeping bentgrass. *Plant Physiology, 161*(3), 1375-1391.

Zhou, X., Ruan, J., Wang, G., & Zhang, W. (2007). Characterization and identification of microRNA core promoters in four model species. *PLoS Computational Biology, 3*(3), e37.

Zilberman, D., Gehring, M., Tran, R. K., Ballinger, T., & Henikoff, S. (2007). Genome-wide analysis of *Arabidopsis thaliana* DNA methylation uncovers an interdependence between methylation and transcription. *Nature Genetics, 39*(1), 61-69.

LIST OF PUBLICATIONS AND PAPERS PRESENTED

Publications

Gudimella, R., Singh, P., Mazumdar, P., Wong, G. R., Lau, S. E., & Harikrishna, J. A. (2018). Genome-wide regulatory network mapping of miRNA and transcription factors in banana roots. *Tropical Plant Biology*, 1-13.

Davey, M. W., **Gudimella, R.**, Harikrishna, J. A., Sin, L. W., Khalid, N., & Keulemans, J. (2013). "A draft *Musa balbisiana* genome sequence for molecular genetics in polyploid, inter- and intra-specific *Musa* hybrids". *BMC Genomics, 14*, 683.

Conference paper

Harikrishna, J. A., Lee, W. S., **Gudimella, R**., Khalid, N., & Davey, M. W. (2014, August). The banana salt stress transcriptome: large and small RNA expression in stressed banana. *In XXIX International Horticultural Congress on Horticulture: Sustaining Lives, Livelihoods and Landscapes (IHC2014): IX 1114* (pp. 125-132).

Poster presentation:

Gudimella, R., Tammi, M., Khalid N., and Harikrishna, J.A. Computational identification of functional siRNA in *Musa acuminata*. *In 10th International conference of Bioinformatics (INCOB-2011),30th November – 2nd December,2011 at Kuala Lumpur, Malaysia.*

Lee W. S., **Gudimella, R.,** Tammi, M., Khalid N., and Harikrishna, J.A. Transcriptome and sRNAome: Gene expression dynamics in abiotic- stressed banana plants. Poster presentation. *In The University of Malaya's Researchers' Conference 2013, 19th-20th November 2013, Research Management and Innovation Complex, University of Malaya*, *Kuala Lumpur, Malaysia.*

CrossMark

Genome-Wide Regulatory Network Mapping of miRNA and Transcription Factors in Banana Roots

Ranganath Gudimella[1,2] · Pooja Singh[2] · Purabi Mazumdar[2] · Gwo Rong Wong[1,2] · Su-Ee Lau[2] · Jennifer Ann Harikrishna[1,2]

Abstract

MicroRNA (miRNA) are important regulators of gene expression. Plant miRNA have been reported both to target and to be regulated by transcription factors, however, genomic distribution of miRNA, transcription factor targets for miRNA, transcription factor binding sites (TFBS) of miRNA promoters and their regulatory networks have not been systematically mapped in banana. In this study, genome-wide annotation of miRNA in the most recently published banana genome sequence was used to predict miRNA promoter regions and to map TFBS of miRNA genes. A total of 183 mature miRNAs, comprising 144 orthologous miRNA and 39 Musa-specific miRNA were predicted. Following this, banana root degradome data was used to confirm miRNA targets and the transcription factor targets were placed into a predicted network together with their targeting miRNA using cytoscape. Gene ontology of the 20 transcription factors among the predicted miRNA targets, showed predominance for auxin-activated signalling and developmental processes. Profiling of TFBS motifs across miRNA promoter regions showed that binding site motifs for TCP, AP2/ERF, GATA, NF-YB, DOF, B3, bZIP, tribelix, ZF-HD, bHLH and Dehydrin transcription factor families are abundant in the *Musa acuminata* genome. Finally, we propose a regulatory network for the miRNA families miR156, miR164, miR166, miR171, miR319, miR396, miR528, mac-miR-new14 and mac-miR-new20 and their respective transcription factor targets.

Keywords miRNA · Banana · Cis-elements · Transcription factors · Binding sites · Regulatory networks

Abbreviations

GO	Gene Ontology
miRNA	MicroRNA
Redox	Oxidation-reduction
TF	Transcription factor
TFBS	Transcription factor binding site
TSS	Transcription start site

✉ Jennifer Ann Harikrishna
  jennihari@um.edu.my

[1] Centre for Research in Biotechnology for Agriculture (CEBAR), University of Malaya, 50603 Kuala Lumpur, Malaysia

[2] Institute of Biological Sciences, Faculty of Science, University of Malaya, 50603 Kuala Lumpur, Malaysia

Introduction

MicroRNAs (miRNA) are important negative regulators of gene expression. The genomic distribution of miRNAs shows them to be both genic (within the genes) and intergenic (between two genes) and to be encoded singly or in clusters (Cui et al. 2009; Zhang et al. 2009). Several miRNAs show differential expression in response to various metabolic processes, to environmental stress and as part of normal plant development (Chen et al. 2016). Comparison of miRNA families across plant species including banana; *Musa acuminata* (D'Hont et al. 2012) and *Musa balbisiana* (Davey et al. 2013) show that plant genomes encode both conserved and species-specific miRNA. Further investigation of miRNA-associated regulatory networks in plant genomes indicate that miRNA which target transcription factors (TFs) may be regulated by the same or related TF, to co-regulate gene expression (Qiu et al. 2010; Guo et al. 2016; Samad et al. 2017). Such co-regulation of miRNAs and TFs in a biological response can establish different types of regulatory

🖺 Springer

"A draft *Musa balbisiana* genome sequence for molecular genetics in polyploid, inter- and intra-specific *Musa* hybrids"

Mark W Davey[1*], Ranganath Gudimella[2], Jennifer Ann Harikrishna[2], Lee Wan Sin[2], Norzulaani Khalid[2] and Johan Keulemans[1]

Abstract

Background: Modern banana cultivars are primarily interspecific triploid hybrids of two species, *Musa acuminata* and *Musa balbisiana*, which respectively contribute the A- and B-genomes. The *M. balbisiana* genome has been associated with improved vigour and tolerance to biotic and abiotic stresses and is thus a target for *Musa* breeding programs. However, while a reference *M. acuminata* genome has recently been released (Nature 488:213–217, 2012), little sequence data is available for the corresponding B-genome.

To address these problems we carried out Next Generation gDNA sequencing of the wild diploid *M. balbisiana* variety 'Pisang Klutuk Wulung' (PKW). Our strategy was to align PKW gDNA reads against the published A-genome and to extract the mapped consensus sequences for subsequent rounds of evaluation and gene annotation.

Results: The resulting B-genome is 79% the size of the A-genome, and contains 36,638 predicted functional gene sequences which is nearly identical to the 36,542 of the A-genome. There is substantial sequence divergence from the A-genome at a frequency of 1 homozygous SNP per 23.1 bp, and a high degree of heterozygosity corresponding to one heterozygous SNP per 55.9 bp. Using expressed small RNA data, a similar number of microRNA sequences were predicted in both A- and B-genomes, but additional novel miRNAs were detected, including some that are unique to each genome. The usefulness of this B-genome sequence was evaluated by mapping RNA-seq data from a set of triploid AAA and AAB hybrids simultaneously to both genomes. Results for the plantains demonstrated the expected 2:1 distribution of reads across the A- and B-genomes, but for the AAA genomes, results show they contain regions of significant homology to the B-genome supporting proposals that there has been a history of interspecific recombination between homeologous A and B chromosomes in *Musa* hybrids.

Conclusions: We have generated and annotated a draft reference *Musa* B-genome and demonstrate that this can be used for molecular genetic mapping of gene transcripts and small RNA expression data from several allopolyploid banana cultivars. This draft therefore represents a valuable resource to support the study of metabolism in inter- and intraspecific triploid *Musa* hybrids and to help direct breeding programs.

Keywords: (a)biotic stress, Banana, Fe'i, Genetic diversity, microRNA (miRNA), Molecular breeding, *Musa acuminata*, *Musa balbisiana*, Pisang Klutuk Wulung, Plantain, Polyploidy, Wild banana

* Correspondence: mark.davey@biw.kuleuven.be
[1]Laboratory of Fruit Breeding and Biotechnology, Division of Crop Biotechnics, Department of Biosystems, Katholieke Universiteit Leuven, Willem de Croylaan 42, box 2427B-3001, Heverlee, Leuven, Belgium
Full list of author information is available at the end of the article

A tale of two banana sequences ⌄

 Anne Vézina ⊙ Thursday, 07 November 2013

The recent publication by Belgian and Malaysian scientists of a draft *Musa balbisiana* genome sequence in BMC Genomics went largely unnoticed, at least compared to the media attention that surrounded last year's publication of the *Musa acuminata* genome sequence. The media may have a limited appetite for banana genomics but not *Musa* scientists. They knew from the beginning that the so-called A genome was not the whole story. No sooner had its sequence been released that French and Chinese scientists were discussing joining forces to produce a reference sequence for the edible banana's other founding genome, the B genome donated by *Musa balbisiana*, which is often associated with tolerance to abiotic stresses. It looks as if they have been beaten to the finish line, but the fact is that the two scientific teams were pursuing different strategies.

Both teams used a genebank accession called Pisang Klutuk Wulung (PKW), characterized by a black pseudostem. But while the Belgian-Malaysian team extracted the DNA directly from plantlets obtained from the ITC, the French-Chinese team extracted the DNA from a doubled haploid (DH) PKW plant produced by scientists at the French Agricultural Research Centre for Development (CIRAD).

When sequencing the first genome of a species, scientists often use a homozygous derivative (a doubled haploid) to facilitate the assembling of the reference sequence. By doing so, however, they forego capturing allelic diversity (the alternative forms – alleles – of the same gene). This typically comes later during resequencing efforts that compare closely related genomes to the reference one, which is how the draft *balbisiana* sequence was obtained. The Belgian-Malaysian scientists used the *acuminata* reference sequence generated from DH Pahang as a template onto which they aligned their fragments of *balbisiana* DNA.

In addition to predicting 36,638 genes, the scientists found more than 18 million SNPs (DNA markers that have a single nucleotide difference between individuals), which by definition cannot be detected in DH plants. They also looked for microRNA (small non-coding RNA molecule that are encoded by DNA and play a role in the regulation of gene expression) and found new microRNA families, some of which are unique to the B genome. All the sequence data have been uploaded to The Banana Genome Hub where they are available for browsing and downloading by the scientific community to further annotate and investigate.

Meanwhile, the French-Chinese team is continuing its work on the reference sequence which will go one step further and position the genes on the *balbisiana* chromosomes by using the genetic map being developed by CIRAD scientists as part of RTB's genotyping activities. Knowing where the genes and transposable elements are located on the chromosomes is important in understanding how the B genome is organized and to what extent it is structurally similar to the A genome.

The *balbisiana* reference sequence anchored to the species' 11 chromosomes is being produced by the CATAS' Institute of Tropical Bioscience and Biotechnology in collaboration with the Beijing Genomics Institute and CIRAD, in the framework of the Global *Musa* Genomics Consortium.

The draft *balbisiana* sequence is the result of a collaboration between KU Leuven's Laboratory of Fruit Breeding and Biotechnology and the University of Malaya's Centre for Research in Biotechnology for Agriculture and Institute of Biological Sciences.

129

Table S1: List of bioinformatics databases as plant genome resources.

Database	URL
TAIR (the Arabidopsis information resource)	www.arabidopsis.org/
RGAP (rice genome annotation project)	rice.plantbiology.msu.edu/
RAP-DB (the rice annotation project database)	rapdb.dna.affrc.go.jp/
Phytozome	phytozome.jgi.doe.gov/pz/portal.html
miRBase (the microRNA database)	www.mirbase.org/
PLncDB (plant long non-coding RNA database)	chualab.rockefeller.edu/gbrowse2/homepage.html
PNRD (a plant non-coding RNA database)	(structuralbiology.cau.edu.cn/PNRD)
GEO (gene expression omnibus)	www.ncbi.nlm.nih.gov/geo/
SRA (sequence read archive)	www.ncbi.nlm.nih.gov/sra/
DDBJ	trace.ddbj.nig.ac.jp/dra/index_e.html
Next-Gen Sequence Databases	mpss.danforthcenter.org/index.php
ASRP (Arabidopsis small RNA project)	asrp.danforthcenter.org/
CSRDB (cereal small RNAs database)	sundarlab.ucdavis.edu/smrnas/
PlantNATsDB (plant natural antisense transcripts database)	bis.zju.edu.cn/pnatdb/
mirEX (Arabidopsis pri-miRNA expression atlas)	www.combio.pl/mirex1/
PmiRKB (plant microRNA knowledge base)	bis.zju.edu.cn/pmirkb/
AVT (AtGenExpress visualization tool)	jsp.weigelworld.org/expviz/expviz.jsp
Arabidopsis eFP Browser	bar.utoronto.ca/efp_arabidopsis/cgi-bin/efpWeb.cgi
PceRBase (plant ceRNA database)	bis.zju.edu.cn/pcernadb/index.jsp
Arabidopsis epigenome maps	neomorph.salk.edu/epigenome/epigenome.html
The SIGnAL Arabidopsis Methylome Mapping Tool	signal.salk.edu/cgi-bin/methylome
UCSC Genome Browser on Arabidopsis thaliana (2004)	epigenomics.mcdb.ucla.edu/cgi-bin/hgTracks?clade = plant&org = A. + thaliana
UCSC Genome Browser on Arabidopsis thaliana (2009)	genomes.mcdb.ucla.edu/cgi-bin/hgTracks?db = araTha2
Rice epigenome maps	plantgenomics.biology.yale.edu
Plant Methylome DB	epigenome.genetics.uga.edu/PlantMethylome/

Table S2: List of bioinformatics tools and software for plant specific analysis.

Software	URL	Description
PlantCARE (a plant cis-acting regulatory element database)	bioinformatics.psb.ugent.be/webtools/plantcare/html/	Plant gene promoter analysis
PLACE (a database of plant cis-acting regulatory DNA elements)	www.dna.affrc.go.jp/htdocs/PLACE/	
JASPAR (an open-access database for eukaryotic transcription factor binding profiles)	jaspar.genereg.net	
The MEME suite (containing motif-based sequence analysis tools)	meme-suite.org	
Bowtie	bowtie-bio.sourceforge.net/index.shtml	An ultrafast, memory-efficient short read aligner
Bowtie 2	bowtie-bio.sourceforge.net/bowtie2/index.shtml	An ultrafast and memory-efficient tool for aligning relatively long sequencing reads to long reference sequences
ShortStack	github.com/MikeAxtell/ShortStack/releases/	A Perl program for comprehensive annotation and quantification of small RNA genes
NATpipe	www.bioinfolab.cn/NATpipe/NATpipe.zip	Natural antisense transcript prediction
RNAfold webserver	rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi	RNA secondary structure prediction
RNAshapes	bibiserv.cebitec.uni-bielefeld.de/download/tools/rnashapes.html	
miTRATA (microRNA truncation and tailing analysis)	wasabi.ddpsc.org/~apps/ta/index.php	3′ modification analysis of plant small RNAs
WebLogo	weblogo.threeplusone.com/	Search for the conserved sequence motifs
psRNATarget (a plant small RNA target analysis server)	plantgrn.noble.org/psRNATarget/	Target prediction tools for plant small RNAs
Small RNA Target Prediction	wasabi.ddpsc.org/~apps/tp/	
TAPIR (target prediction for plant microRNAs)	bioinformatics.psb.ugent.be/webtools/tapir/	Not only target prediction, also target mimic prediction for plant microRNAs

Table S2, continued.

comPARE (PARE validated miRNA targets)	mpss.danforthcenter.org/tools/mirna_apps/comPARE.php	Degradome-seq data-based validation for plant microRNA—target pairs
sPARTA-Web (small RNA-PARE target analyzer)	mpss.danforthcenter.org/tools/mirna_apps/sparta.php	Degradome-seq data-based validation for plant small RNA—target pairs
CleaveLand4	github.com/MikeAxtell/CleaveLand4/releases	A Perl program for degradome-seq data-based validation for plant small RNA—target pairs
agriGO (a GO analysis toolkit for the agricultural community)	bioinfo.cau.edu.cn/agriGO/index.php	Functional analysis of target genes based on Gene Ontology annotations

Table S3: Small RNA libraries statistic.

sRNA libraries	Total read count	Unique read count
Root	14,494,041	3,488,872
Flower	4,339,904	1,608,519
Fruit	4,926,076	2,526,944
Leaves	2,976,266	1,008,359
Somatic Embryogenic cultures	30,689,659	7,734,891
Reads from all libraries	57,425,946	16,367,585
Non-redundant set (size trimmed to 19-24nt)	15,364,143	12,961,473

Table S4: Classification of Small RNA reads based on annotations.

sRNA classification	Redundant Reads	Unique reads
rRNA	344	86
tRNA	20	14
snRNA	179	129
Known miRNA	941699	386
Exon RNA	493559	388322
repeat region	11225	2806
Unannotated	11514447	12569730
Total	12961473	12961473

Table S5: Mapping statistics of small RNA reads on banana genome version-1.

Chromosome	*Musa* A genome		*Musa* B genome	
	Total Mapped	Uniquely Mapped	Total Mapped	Uniquely Mapped
chr1	224830	140812	151368	83913
chr2	164902	94982	118764	62712
chr3	231298	142711	159417	88107
chr4	233421	155286	158459	93491
chr5	252833	151250	177244	93074
chr6	284841	177187	196639	109779
chr7	219348	127701	160110	86210
chr8	302794	174427	210837	106079
chr9	282219	160320	199418	101217
chr10	275995	142232	207700	94381
chr11	216091	130484	159538	90250
chrUn_random	1387552	427214	1038236	294533
Total Reads	4076124	2024606	2937730	1303746
% reads mapped	31.44%	15.60%	22.60%	10.05%

Table S6: Statistics of small RNA sequencing.

Type	Read count		
	Control	**TR100**	**TR300**
Raw reads	18,189,390	18,396,868	19,641,136
High quality reads	15,911,422	16,039,100	15,190,814
Clean reads*	14,420,971	14,747,201	13,560,353
Reads after collapse**	1,828,161	1,671,987	842,589
Reads aligned to genome***	1,370,670(74.9%)	1,243,425(74.3%)	677,347(80.3%)

Table S7: Statistics of degradome sequencing.

Type	Read count		
	Control	**TR100**	**TR300**
Raw reads	864,578,160	552,588,541	558,682,796
Clean data*	864,429,640	552,428,882	558,635,843
Reads after collapse	18,392,120	11,753,806	11,885,869
Unique clean tags**	5,530,664	3,056,001	4,882,475
Tags aligned to genome***	3100736(56.06%)	1254219(41.04%)	2984652(61.13%)

*After filtering of contaminant sequences including sequencing adapters and poly(A).

**Non-redundant reads.

***Non-redundant reads mappable to the reference *Musa acuminata* genome version- 2

Table S8: Small RNA sequencing and mapping statistics.

Type	Small RNA statistics	
	Non-salt stress	**Salt stress**
Raw reads	18,189,390	19,641,136
High quality reads	15,911,422	15,190,814
Clean reads*	14,420,971	13,560,353
21nt small RNA	1,107,912	8,28,197
	7.68%	6.10%
24nt small RNA	3,328,219	2,556,916
	23.07%	18.85%
21nt aligned to genome***	768,262	593,333
24nt aligned to genome***	2,155,893	1, 690,156

*After filtering of contaminant sequences including sequencing adapters and poly(A).

***Non-redundant reads mappable to the reference *Musa acuminata* genome version- 2

Table S9: Bisulphite sequencing output of salt-stressed banana roots.

	Non-salt stress (UT)	**Salt stress (ST)**
Total Number of Reads	216, 286,990	216,016,372
Read Length (bp)	90	90
Total Number of Bases	19,465,829,100 bases	19,441,473,480 bases

Table S10: Mapping statistics of bisulphite sequencing reads to banana genome.

	Non-salt stress (UT)	Salt stress (ST)
Total Reads	108,143,495	108,008,186
Paired-end alignments with a unique best hit	67,388,376	65,091,889
	62.30%	60.30%
Pairs that did not map uniquely (Multiple Alignments)	13,832,101	13,716,147
	12.80%	12.70%
Mapped reads to *Musa acuminata* genome version- 2	81,220,477	78,808,036
	75.10%	72.90%
Pairs without alignments	26,923,018	29,200,150
	24.90%	27%

Table S11: Cytosine methylation statistics in CG, CHG and CHH contexts.

	Non-salt stress (UT)			Salt Stress (ST)		
	Methylated	%*	Un methylated	Methylated	%*	Un methylated
CG	96,206,963 4.27%	31.1%	213,296,872 9.46%	97,702,241 4.46%	31.1%	216,324,252 9.89%
CHG	44,080,483 1.95%	12.2%	317,068,020 14.06%	46,044,191 2.10%	12.9%	310,995,575 14.12%
CHH	36,274,793 1.61%	2.28%	1,548,299,732 (68.65%)	36,221,844 1.65%	2.4%	1,479,879,490 67.66%
Total	176,562,239 7.83%		2,078,664,624 92.17%	179,968,276 8.22%		2,007,199,317 91.77%
	2,255,226,863			2,187,167,593		

* Calculated by % methylation (context) = 100 * methylated Cs (context) / methylated Cs (context) + unmethylated Cs (context).

Table S12: Distribution of Methylated Cytosine among regions on genome.

	CG		CHG		CHH	
	Non-Salt stress	Salt stress	Non-Salt stress	Salt stress	Non-Salt stress	Salt stress
Exon	0.207	0.2	0.012	0.013	0.006	0.006
Downstream 2kb	0.132	0.135	0.04	0.044	0.012	0.013
Upstream2 Kb	0.133	0.137	0.042	0.046	0.012	0.013
Intron	0.503	0.509	0.09	0.104	0.02	0.022
Repeat	0.518	0.531	0.239	0.244	0.033	0.034

Table S13: Statistics of Differentially expressed siRNA clusters overlapping genomic regions.

	No. of 21nt siRNA clusters	No. of 24nt siRNA clusters	21nt siRNA clusters (%)	24nt siRNA clusters (%)
Exon	126	44	10.37	3.62
Intron	220	222	18.11	18.27
Downstream 2kb	155	125	12.76	10.29
Upstream 2kb	158	142	13	11.69
Intergenic	534	841	43.95	69.22
Repeat	22	27	1.81	2.22
Total	1215	1401	100%	100%

Table S14: Paired-end transcriptome sequencing (RNA-Seq) statistics.

	Non-salt stress	Salt stress
Total number of reads	12,279,060	11,274,596
Average read length (bp)	90	90
Total number of HQ reads*	11,352,903	10,504,837
Percentage of HQ reads	92.46%	93.17%
Total number of bases	1,105,115,400	1,014,713,640
Total number of HQ bases**	1,037,037,224	956,651,654
Percentage of HQ Bases	93.84%	94.28%

* >70% of bases in a read with Phred Quality Score > 20

** Bases with Phred Quality Score > 20.

**HQ: high quality.

Table S15: Mapping statistics of paired-end transcriptome sequencing reads (RNA-Seq) to reference *Musa* genome.

|	**Non-salt stress**	**Salt stress**
Total Transcriptome Reads	12,279,060	11,274,596
Mapped to Genome*	9,680,571(78.8%)	8,815,511(78.2%)
Multiple Alignments	573700(4.6%)	57,8775(5.1%)

*Sequence mapped to a reference *Musa acuminata* DH Pahang version-2 genome (Martin et al., 2016).

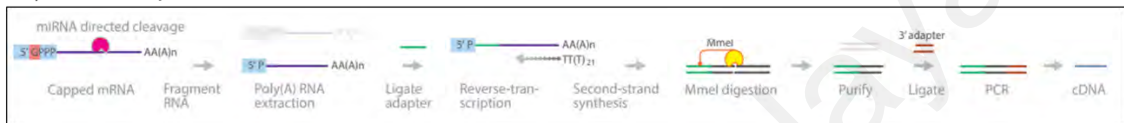**Table S16:** Distribution of DMR loci across genomic regions.

	Non-salt stress			**Salt stress**		
	CG	CHG	CHH	CG	CHG	CHH
Exon	629	1364	5169	7908	193	47
Downstream	292	215	1682	1804	108	139
Upstream	292	213	1665	1887	115	145
Intron	370	1445	8016	5497	516	426
Intergenic	490	358	6262	4752	1329	1185
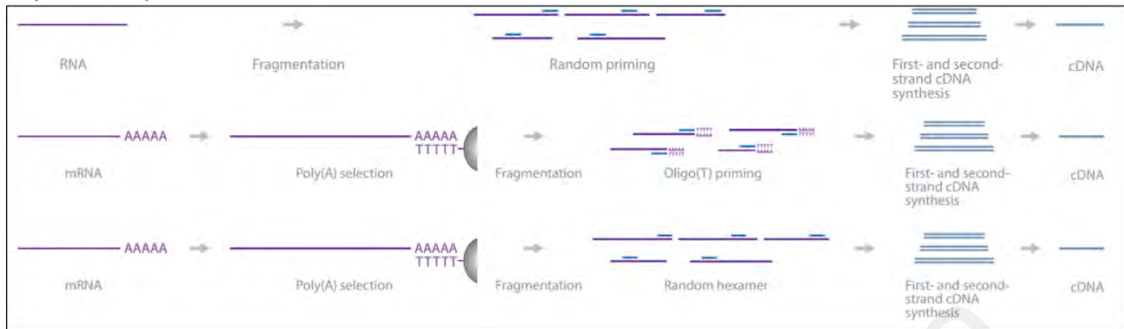
Figure S1: Schematic overview of sequencing method for A) RNA sequencing (RNA-seq), B) Parallel analysis of RNA ends (PARE-seq) and C) Reduced Representation of Bisulphite Sequencing (RRBS-seq) used in current study (Illumina, 2017a; Illumina, 2017b)
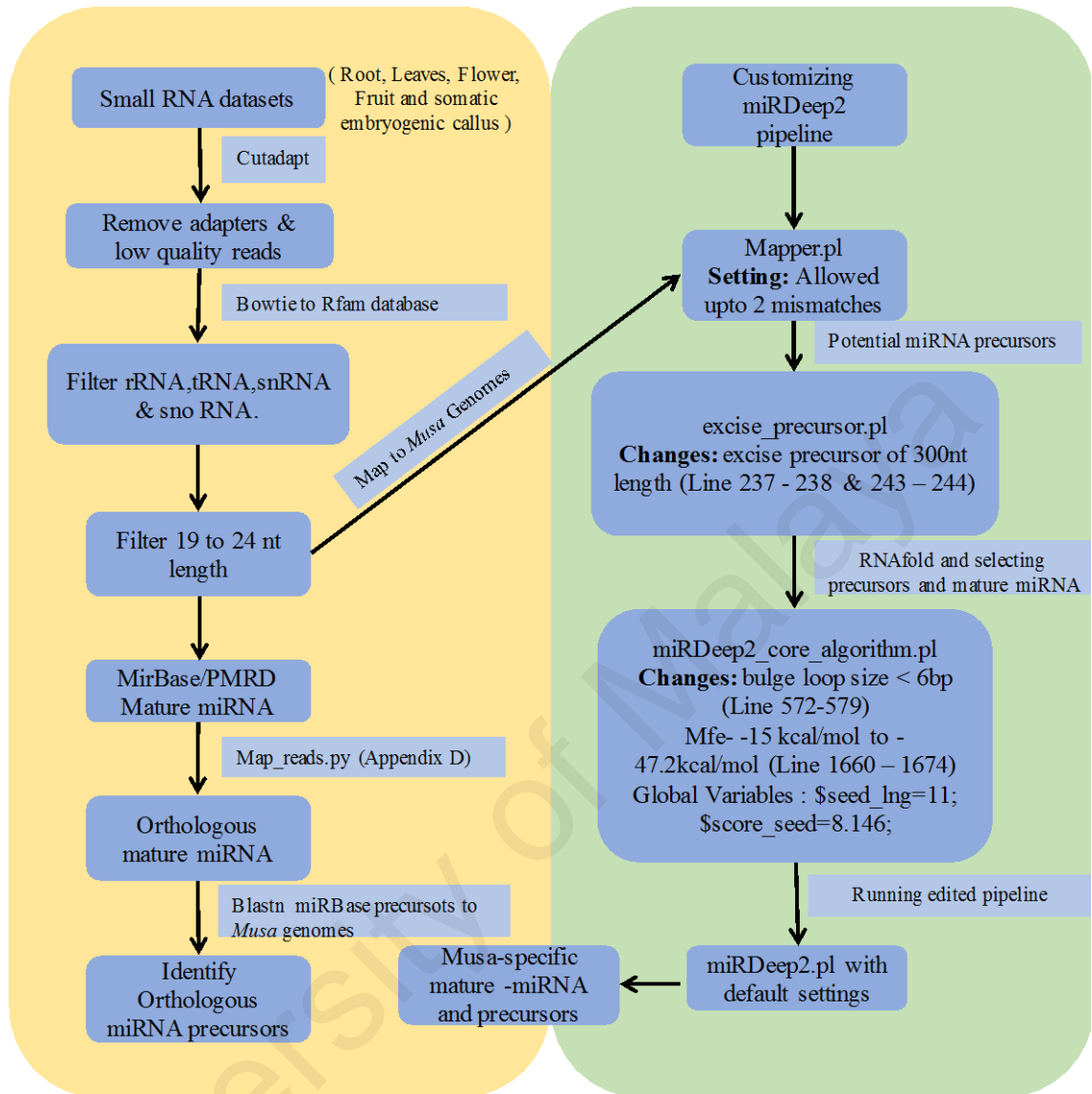
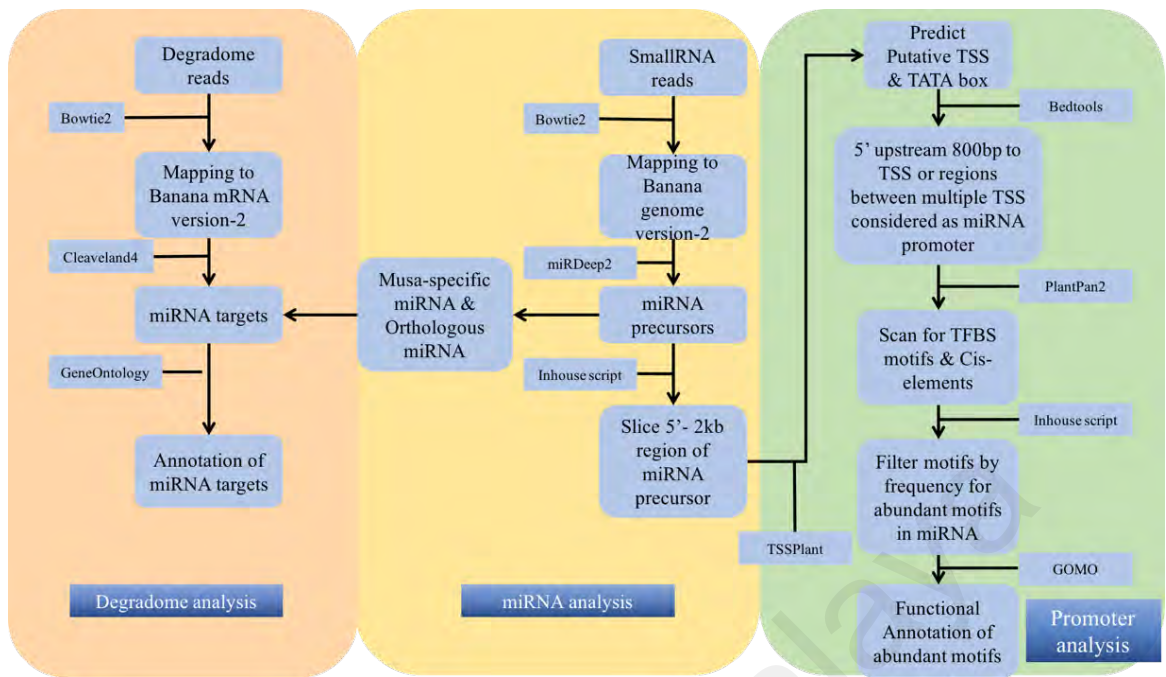**Figure S2**: Customized pipeline for predicting miRNA for Musa genomes.

Figure S3: Flow chart of methods used for miRNA, degradome and TF predictions on banana version 2 genome.
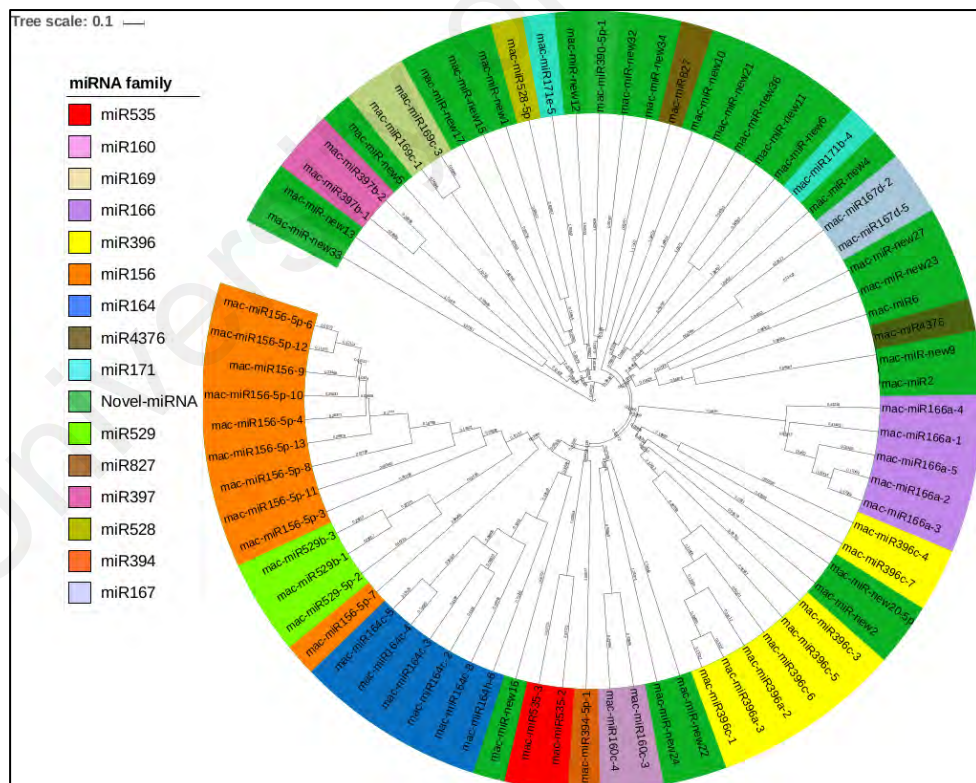

Figure S4: Phylogeny tree constructed by multiple sequence alignment of banana precursor miRNA sequences showing conservation of orthologous miRNA.
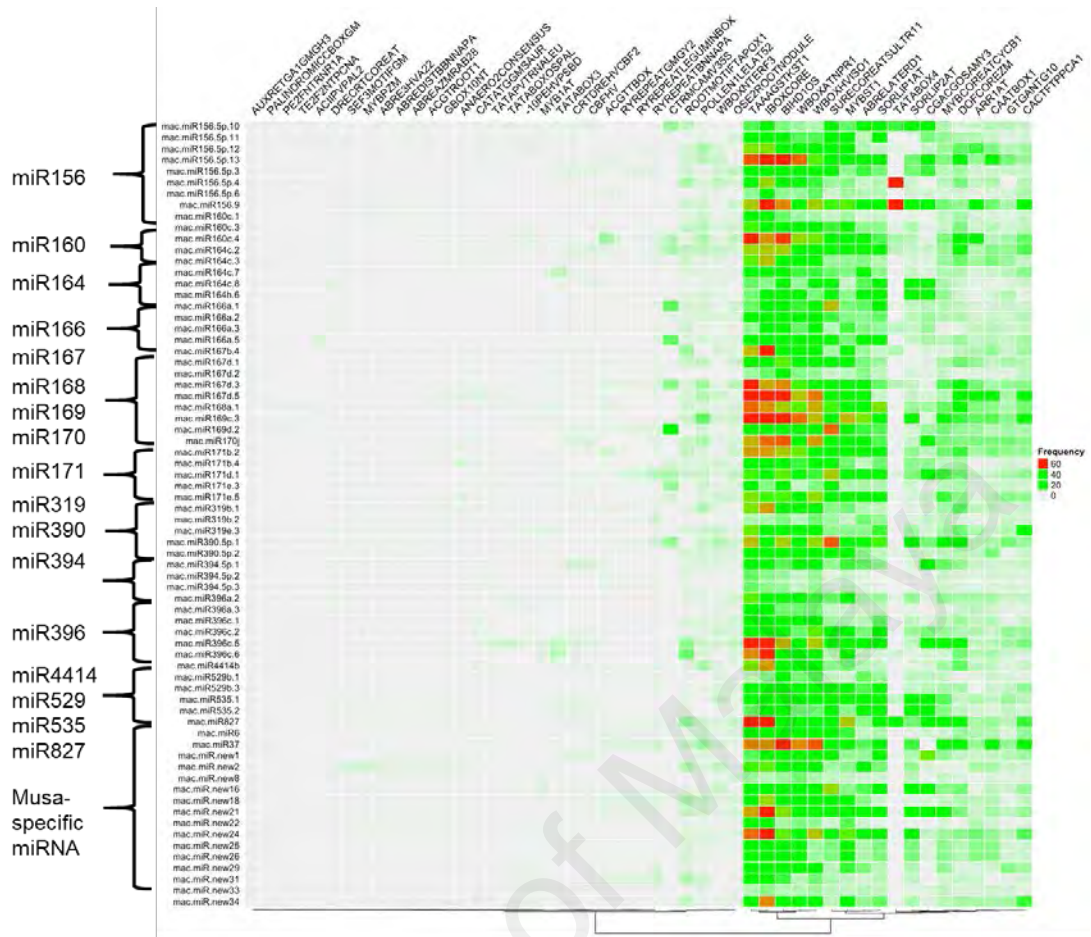
Figure S5: *Cis-elements*/Non-TFBS frequencies within miRNA promoter sequences. Clustering performed by calculating Ward's clustering method and Euclidean distance using R library "ComplexHeatmap". Red represents higher frequency of TFBS within a miRNA promoter region and blue represents low abundance to absence of TFBS.
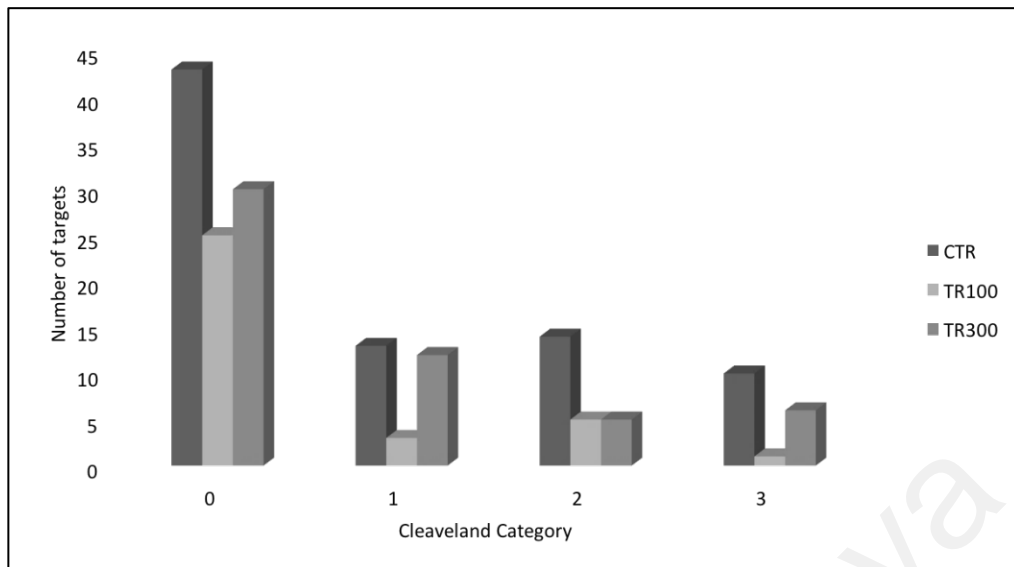
Figure S6: Number of targets predicted by Cleaveland categories showing high number of targets are from category '0' i.e highly precise.

Figure S7: 24nt small RNA distribution among banana repeats.
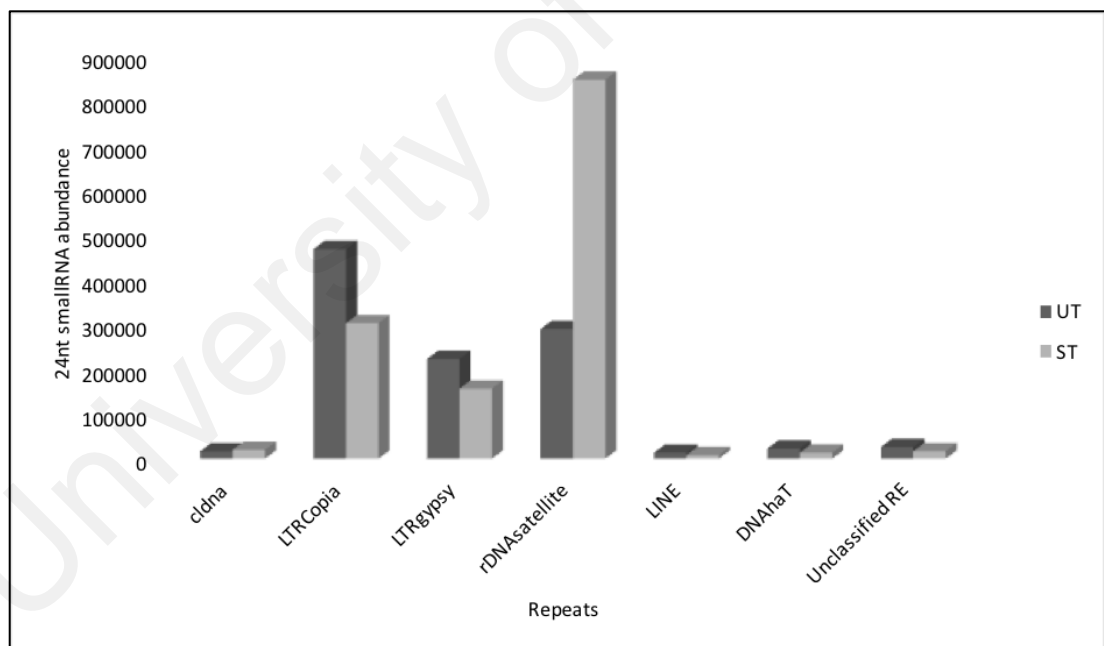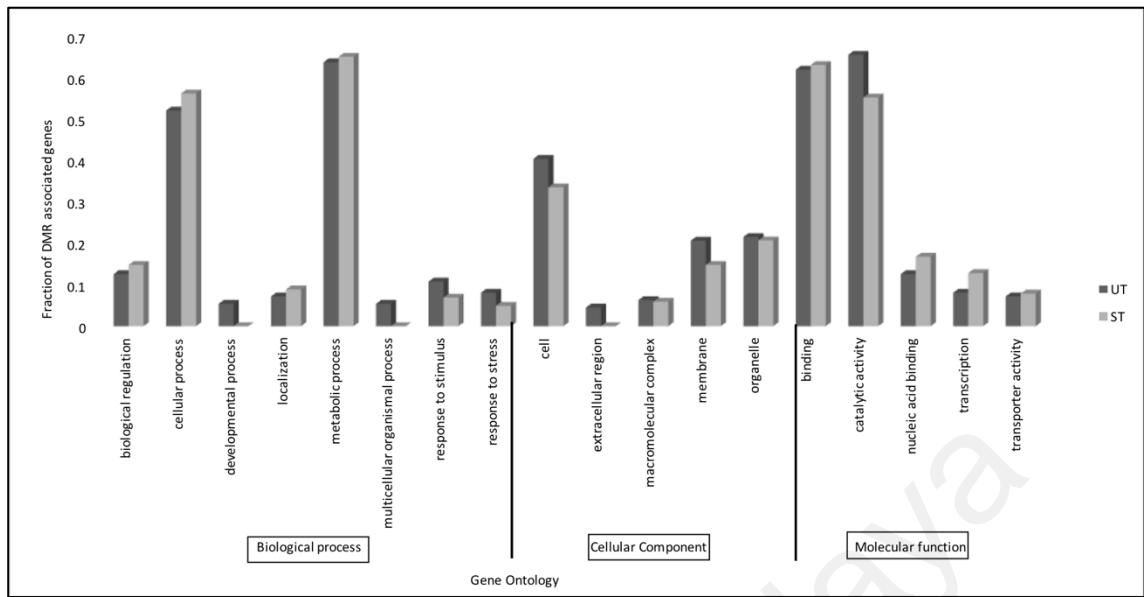
Figure S8: Gene ontology level-2 GO terms of DMR associated genes are between non-salt stress (UT) and salt-stress (ST).
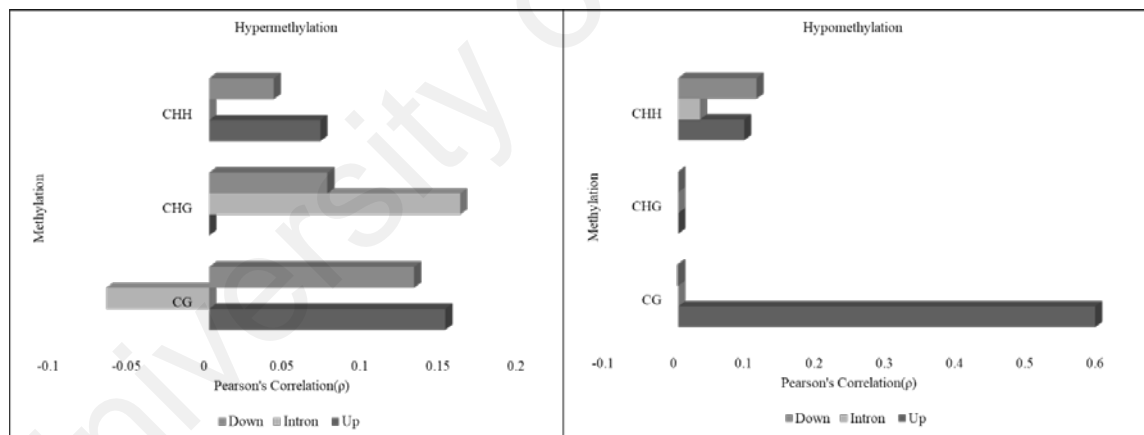

Figure S9: Pearson correlation between small RNA abundance and methylation ratio among 78 genes with DMR and siRNA association.
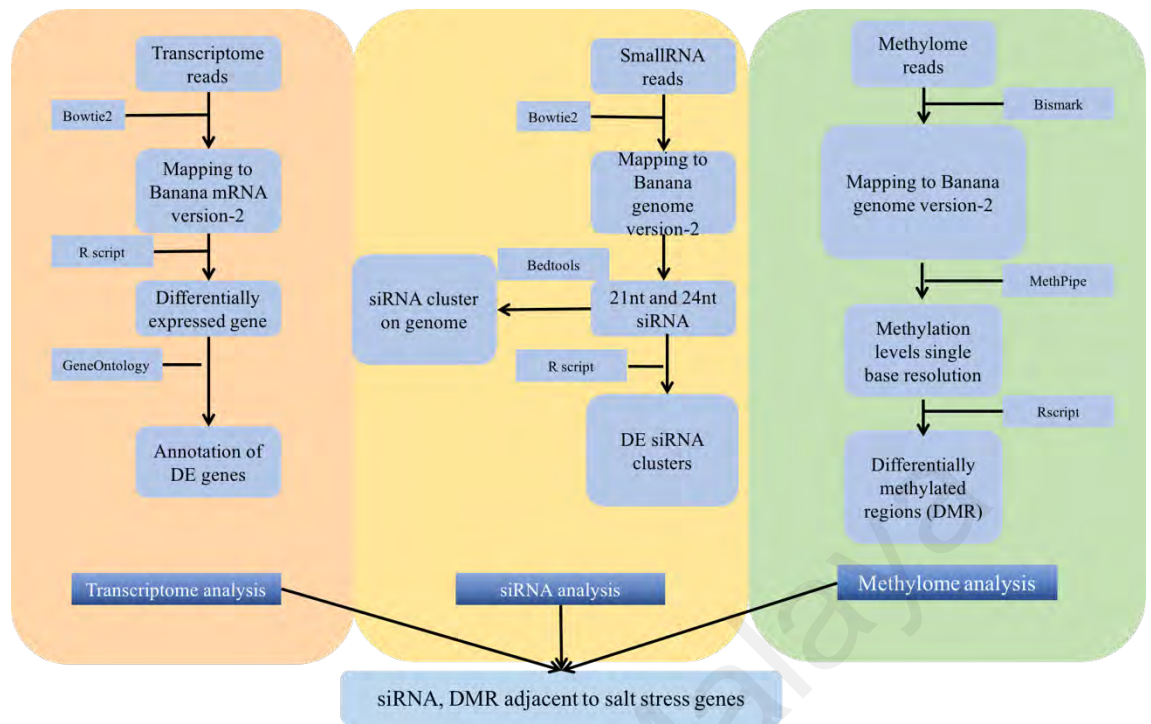
Figure S10: Flow chart of methods used for transcriptome, siRNA and methylome on banana version 2 genome.

APPENDIX C (SCRIPTS USED FOR ANALYSIS AND GENERATE FIGURES)

Script 1: Map smallReads to miRBase for known miRNA.

```
# Usage : map_reads_script.py input file > output.fasta
# input file = Length,abundance and sequence tab delimited format
# Output file = miRNA matches in fasta format
#
###############################################################################
#
from Bio import SeqIO
import csv
import sys

readFile = sys.argv[1]          # Read input from command line

it = SeqIO.parse(open('Non-redundant miRBase/PMRD miRNA in fasta format'),'fasta')
#

nrD = (Megraw et al., 2016)
for rec in it:
    nrD[rec.seq.tostring()] = rec.id          # Creating ids for matching reads and identifi
es 5' and 3' pairing.
    nrD[rec.seq.tostring()[1:]] = rec.id+'_5pdg'
    nrD[rec.seq.tostring()[:-1]] = rec.id+'_3pdg'

linkedMature = ' '.join(nrD.keys())
print linkedMature

reader = csv.reader(open(readFile,'rU'), delimiter='\t')          # Read input file

for line in reader:
    length,  count,  seq = line
    seq = seq.replace('T', 'U')
    if seq in linkedMature:

        try:
            print '>%s %s'%(nrD[seq], count)          # print match sequence in fasta f
ormat
            print seq
        except KeyError:
            pass
```

Script 2: Retrieve 5' 2kb upstream region of miRNA precursor

```
# Usage : mac_promoter_extraction_genome.py
# input file = Length,abundance and sequence tab delimited format
# Output file = miRNA matches in fasta format
#
################################################################################
#
from Bio import SeqIO
from Bio.Seq import Seq
outfile=open("miRNA_promoter_using_genome_v2.fasta","w") # the output file

for rec1 in SeqIO.parse("miRNA-precursors.fa","fasta"):#Input file precursor file name
in fasta format
    T=True
    print rec1.description
    for rec2 in SeqIO.parse("musa_acuminata_v2_pseudochromosome.fna","fasta"): #the
genome file name and format

        if str(rec1.seq).upper() in str(rec2.seq).replace("T","U"):
            T=False
            n1=str(rec2.seq).replace("T","U").find(str(rec1.seq).upper())
            n2=n1+len(str(rec1.seq))
            outfile.write(">")
            outfile.write(str(rec1.description))
            outfile.write(" Promoter:")
            #outfile.write(str(rec2.description))
            outfile.write(str(n1-2000))
            outfile.write("..")
            outfile.write(str(n1))
          #  outfile.write("(+)")
            outfile.write("\n")
            promoter = str(rec2.seq)[n1-2000:n1]
          #  print promoter
            outfile.write(str(promoter))
            outfile.write("\n")
            print "found on +"

        rStrand=str(rec2.seq.complement()).replace("T","U")[::-1]
        if str(rec1.seq).upper() in rStrand:
            T=False
            n1=rStrand.find(str(rec1.seq).upper())
            n2=n1+len(str(rec1.seq))
            n11=len(rStrand)-n2
            n22=len(rStrand)-n1
            outfile.write(">")
            outfile.write(str(rec1.description))
            outfile.write(" Promoter:")
            #outfile.write(str(rec2.description))
            outfile.write(str(n22))
            outfile.write("..")
            outfile.write(str(n22+2000))
```

149

```
        outfile.write("\n")
        promoter = rStrand[n1-2000:n1]
        #print promoter.replace("U","T")
        outfile.write(str(promoter).replace("U","T"))
        outfile.write("\n")
        print "found on -"

    if T:
        print "Not found! ******************* Not found!"
outfile.close()
```

Script 3: Bedtools commands to retrieve miRNA promoter sequences.

 # to run TSSP on command line

bedtools slop -s -i "Input bedfile with miRNA TSS positions" -g "Input chromosome siz
es in chr\tlength format" -l 800 -r 0 > "slop-output-file.bed" *# Command to read the 5' u
pstream 800bp region for each TSS.*

bedtools getfasta -fi "chromosomal sequences in fasta file" -bed "slop-output-file.bed" -
s -fo "output.fasta" -name -fullHeader *# Retrieve fasta format sequences of 800bp miRN
A promoter region.*

Script 4: Reads PlantPan output for each miRNA to calculate motif frequency.

```
library(plyr)
library(dplyr)

out_directory <- "Input folder name from with files from plantpan promoter analysis" #
Read all files from the folder
setwd(out_directory)
file_list <- list.files()
#list.data <- lapply(file_list, read.delim,header=FALSE)
list.data <- lapply(file_list,function(i){read.table(i, header=FALSE,sep = "\t") })# Re
ad each file in directory
names(list.data)<- as.character(file_list) # Assigns filenames
PA <- ldply(list.data, data.frame) # Arranges all files data into one dataframe
motif.count <- count(PA,V1,V3) # Counts based on miRNA and TF family motifs
#motif.filter <- subset(motif.count,n >= 4) # filters motifs with more than 4 frequency (
Run if required)
motif.freq <- aggregate(n ~ V1 + V3, data = motif.count, sum) # Sums the motif freque
ncy based on TF family and miRNA

# Splits data frame based on miRNA and assigns frequency to each TF family
motif.split <- split(motif.freq , f = motif.freq$V1)
motif.lap <- lapply(motif.split,function(x) x[(names(x) %in% c("V3", "n"))])

merge.all <- function(x, y) {
  merge(x, y, all=TRUE, by="V3")
```

```
}
# Merges all data frame into single dataframe and gives column names
motif.merge <- Reduce(merge.all, motif.lap)
col_list <- c("Motif",file_list)
colnames(motif.merge)[1:ncol(motif.merge)]<- col_list
motif.merge[is.na(motif.merge)] <- 0

# Write output in tab delimited format
write.table(motif.merge,file="Path of output file",sep="\t",quote = FALSE,row.names
= FALSE)
```

Script 5: Stacked barplots with multi-color palette (Figure 4.4).

```
install.packages(ggplot2)
install.packages(rwantshue)
install.packages(reshape)

library("ggplot2")
library('rwantshue')
library("reshape")
# read file tab delimited file
miRNApre <- read.table(file="",sep="\t",header=TRUE)

# Reformat dataframe for plotting
miRNApre.m <- melt(miRNApre)

# Assign colors for all the variables
cols <- colorRampPalette(brewer.pal(12, "Dark2"))
myPal <- cols(length(unique(miRNApre.m$variable)))

# create a color scheme object
scheme <- iwanthue()
# generate a new color palette (vector of hex values) with presets...
scheme$hex(21)

#plot bar chart
miRNA.plot <- ggplot(data=miRNApre.m, aes(x=X, y=value,fill=variable)) +
 geom_bar(stat="identity") +
 theme_classic() +
 scale_fill_manual(values = scheme$hex(21),name="miRNA") + xlab("chromosomes
") + ylab("Number of precursor loci")

# Save figure in tiff format
tiff(filename = "",10,5,res=300,compression="lzw",units="in")
miRNA.plot
dev.off()
```

Script 6: Heatmap representation of TFBS frequencies (Figure 4.5).

```r
install.packages("ComplexHeatmap")
install.packages("preprocessCore")
install.packages("circlize")

library(ComplexHeatmap)
library(preprocessCore)
library(circlize)

TF.freq <- as.matrix(read.table(file="",sep="\t",header=TRUE,row.names = 1))
miRNA.class <- read.table(file="",sep="\t",header = TRUE)
TF.class <- read.table(file="",sep="\t",header = TRUE)

TF.norm <- normalize.quantiles(TF.freq,copy = TRUE)
colnames(TF.norm) <- colnames(TF.freq)
rownames(TF.norm) <- rownames(TF.freq)

hmap <- Heatmap(TF.freq,cluster_columns = FALSE,cluster_rows = FALSE,name="F
requency",
          clustering_method_rows = "complete",clustering_distance_rows = "pearson",
          row_names_gp = gpar(fontsize = 12),column_names_gp =  gpar(fontsize = 1
2), bottom_annotation = hcol,row_names_side = "left")

hcol <- columnAnnotation(df=miRNA.class,width = unit(1, "cm"))
hrow <- rowAnnotation(df=TF.class,width = unit(1, "cm"))

png(filename = "",20,18,res=300,units="in")
hmap + hrow
dev.off()
```

Script 7: Shell script to run bisulphite sequence analysis.

```bash
# Mapping to genome using Bismark
~/NGS/aligner/bismark_v0.16.3/bismark -q -p 4 -B SENR2001004 ~/projects/banana
-WGBS/banana-genome-v2/genome/ -1 ~/projects/banana-WGBS/trim-fastq/SENR200
1004_1_val_1.fq.gz -2 ~/projects/banana-WGBS/trim-fastq/SENR2001004_2_val_2.fq.
gz

#Post bismark mapping MethPipe commands to generate single base methylation levels
#Script allows reading of multiple bam files to generate seperate outputs

for sample in `ls /home/alignment/*.bam` # specify the path for all bam files
do
dir="/mnt/sdb/banana-WGBS"  # specify the path to use as base name
base=$(basename $sample ".bam"  )
echo $dir/$base
to-mr -o $(Leinonen et al.).mr -m bismark $dir/$(Leinonen et al.).bam # Converts Bam
file to mr format
LC_ALL=C sort -k 1,1 -k 2,2n -k 3,3n -k 6,6 -o $(Leinonen et al.)-sorted.mr $(Leinone
```

```
n et al.).mr # Sort mr format files
duplicate-remover -S $(Leinonen et al.)-dremove-stat.txt -o $(Leinonen et al.)-deremo
ve.mr $(Leinonen et al.)-sorted.mr # Remove duplicates from sorted mr format files
methcounts -c "/Path/banana-genome-v2/genome" -o $(Leinonen et al.).meth $(Leinon
en et al.)-deremove.mr # # Generate meth counts files from mr format files
done
```

Script 8: Calculate differentially methylated regions using BS-seq data.

```
# Usage: Rscript DMR-Bsseq-script.R
# Input: Tab delimited file of methylation counts (chr,position,number of reads mapped,
number of methylated cytosines)
# Output: Tab delimited file i.e DMR regions
# Installation:
# source("https://bioconductor.org/biocLite.R")
# biocLite("DSS")
# biocLite("bsseq")
#########################################################################

library(DSS)
require(bsseq)
dat1 <- read.table("Input file with for UT sample for CG,CHG and CHH seperately", h
eader=TRUE)
dat2 <- read.table("Input file with for ST sample for CG,CHG and CHH seperately", he
ader=TRUE)

# read BS seq data into bsseq object
BSobj.CG <- makeBSseqData( list(dat1, dat2),c("UT","ST") )
BSobj.CHG <- makeBSseqData( list(dat1, dat2),c("UT","ST") )
BSobj.CHH <- makeBSseqData( list(dat1, dat2),c("UT","ST") )

# Test for differentially methylated loci

dmlTest.CG <- DMLtest(BSobj.CG, group1=c("UT"), group2=c("ST"),smoothing = T
RUE)
head(dmlTest.CG)
dmlTest.CHG <- DMLtest(BSobj.CHG, group1=c("UT"), group2=c("ST"),smoothing
= TRUE)
head(dmlTest.CHG)
dmlTest.CHH <- DMLtest(BSobj.CHH, group1=c("UT"), group2=c("ST"),smoothing
= TRUE)
head(dmlTest.CHH)

# Call differentially methylated regions
dmrs2.CG <- callDMR(dmlTest.CG, delta=0.1, p.threshold=0.05,minlen = 100,minCG
= 10,dis.merge = 100)
dmrs2.CHG <- callDMR(dmlTest.CHG, delta=0.1, p.threshold=0.05,minlen = 100,min
CG = 10,dis.merge = 100)
dmrs2.CHG <- callDMR(dmlTest.CHH, delta=0.1, p.threshold=0.05,minlen = 100,min
CG = 10,dis.merge = 100)
```

```
# write result to output file in tab delimited format
write.table(dmrs2.CG,"Path to write DMR output",quote=FALSE,sep = "\t",row.name
s = FALSE)
```

Script 9: Differential Expression of RNA-seq and siRNA.

```
# Usage: Rscript DEGseq-script.R
# Input: Tab delimited file of raw counts (Geneid,sample1,sample2)
# Output: Tab delimited file i.e output_score.txt and output.html
# installation: install.packages("DEGseq")
#####################################################################
library(DEGseq)

####### Read the tab delimited raw counts file

miRNA.new<-read.table("",row.names=1,header=TRUE)
# Convert the dataframe into matrix
miRNA.mat <- as.matrix(miRNA.new)
# Normalize raw counts by transcripts per million (TPM)
CTR.norm <- miRNA.mat[,1]*1000000/sum(miRNA.mat[,1])
TR300.norm <- miRNA.mat[,2]*1000000/sum(miRNA.mat[,2])
# Convert into dataframe
namemir <- row.names(miRNA.exp)
musa.miRNA <- data.frame(miRNA.exp$TR300,miRNA.exp$TR100,row.names = na
memir)
write.table(musa.miRNA,file="",sep="\t") # write normalize counts into seperate file
############### Differential expression calculation ##################

geneExpFile <- "Path of file"  # read the normalize counts into DEGseq program
CTR <- readGeneExp(file=geneExpFile, geneCol=1, valCol=2)
TR300 <- readGeneExp(file=geneExpFile, geneCol=1, valCol=3)
mapResultBatch1 <- c(CTR)
mapResultBatch2 <- c(TR300)
DEGexp( geneExpMatrix1=CTR, geneCol1=1, expCol1=2, groupLabel1="CTR",
     geneExpMatrix2=TR300, geneCol2=1, expCol2=2, groupLabel2="TR300",method
="MARS",qValue=0.05,thresholdKind=3,foldChange = 2, rawCount= FALSE,outputDi
r='Directorypath')
```

Script 10: Chromosomal view of methylomes (Figure 4.10).

```r
library(rtracklayer)
library(ggplot2)
library(ggbio)
library(IRanges)
library(GenomicRanges)
library(Rsamtools)
library(gridExtra)

bed.ctr <- read.table(file="Path of input file",sep="\t", col.names = c("chrom", "chrom
Start", "chromEnd", "score","type"))
musarepgff <- import.bed("Repeats bed file")
musarepgff.gr <- as(musarepgff,"GRanges")




ctrbed.gr <- GRanges(seqnames = bed.ctr$chrom,ranges = IRanges(start=bed.ctr$chro
mStart,end = bed.ctr$chromEnd,),score=bed.ctr$score,type=bed.ctr$type)


final.plot <- ggbio() + layout_karyogram(ctrbed.gr, geom="area", aes(x=start, y=score
,fill=type),ylim=c(0,50)) +
  scale_fill_brewer(palette="Dark2") +
  layout_karyogram(musarepchr1, geom = "rect",ylim = c(70, 100)) +
  theme_genome() +
  coord_flip() +
  theme(legend.position="bottom",axis.text.x = element_blank(),axis.ticks.x = element
_blank(),strip.text.x = element_text(colour = "black",size = 12),strip.background = ele
ment_blank(),axis.text.y = element_text(colour = "black",size = 10),legend.title = ele
ment_blank()) +
  labs(x=NULL,y=NULL)


pdf("path to PDF output",12,6)
final.plot
dev.off()
```

Script 11: Boxplot representation of methylation ratio (Figure 4.12).

```r
install.packages(ggplot2)
install.packages(reshape)

library("ggplot2")
library("reshape")

one.dat <- read.table(file="Path to inputfile",sep="\t",header=TRUE)

one.m <- melt(one.dat)

one.plot <- ggplot(data = one.m,aes(x=variable,y=value,fill=X)) + geom_boxplot(posit
ion = "dodge") +
  theme_classic() + xlab("") + ylab("Relative methylation ratio") + theme(axis.title.x=
element_blank(),legend.title=element_blank(),axis.text.x = element_text(angle = 45,
```

```
hjust = 1)) + scale_fill_brewer(palette = "Set1")

png(filename = "Path to output png file",10,3.5,res=300,units="in")
one.plot
dev.off()
```

Script 12: Volcano plot for differentially expressed genes (Figure 4.15).

```
fcdmr <- read.table("Input File", header=TRUE)
fcdmr0 <- subset(fcdmr,fcdmr$FC!="NA")
#plot(fcdmr)
#boxplot(fcdmr0)

library(ggplot2)
library(reshape)
fcdmr1 <- melt(fcdmr)
fcdmr2 <- subset(fcdmr1,fcdmr1$variable!="Non.DMR")
#ggplot(fcdmr1,aes(x=variable,y=value)) + geom_boxplot()

###############
x <- c(2,2,2,2)
y <- c(0,20,40,100)
a <- c(-2,-2,-2,-2)# represents p = 0.1
b <- c(0,20,40,100)
line1 <- data.frame(x,y)
line2 <- data.frame(a,b)

png("output.png",12,8,res = 300,units = "in")
par(mfrow=c(1,1))
with(fcdmr, plot(FC,-log10(q.value), pch=20, cex = 0.7, col = "gray",ylim=range(0,1
00),
          ylab = "-log10(qvalue)", xlab = "log2(foldchange)"))
with(subset(fcdmr,type == "UT"), points(FC,-log10(q.value), pch = 20, cex = 0.4, col
= "darkgray"))
with(subset(fcdmr,type == "ST"), points(FC,-log10(q.value), pch = 20, cex = 0.4, col
= "black"))

with(subset(fcdmr,q.value <= 0.05 & FC >= 2 & type == "ST"), points(FC,-log10(q.
value), pch = 17, cex = 1, col = "red"))
with(subset(fcdmr,q.value <= 0.05 & FC <= -2 & type == "ST"), points(FC,-log10(q
.value), pch = 17, cex = 1, col = "red"))

with(subset(fcdmr,q.value <= 0.05 & FC >= 2 & type == "UT"), points(FC,-log10(q.
value), pch = 17, cex = 1, col = "orange"))
with(subset(fcdmr,q.value <= 0.05 & FC <= -2 & type == "UT"), points(FC,-log10(q
.value), pch = 17, cex = 1, col = "orange"))

lines(line1, col = "red", lwd = 1, lty=2)
lines(line2, col = "red", lwd =1, lty =2 )
legend("topleft", legend=c("UT-DMR", "ST-DMR"), col=c("orange", "red"), pch
= 17, cex=0.8,title = "DE-genes")
```

```
legend("topright", legend=c("UT-DMR", "ST-DMR","Non-DMR"), col=c("darkg
ray", "black","gray"), pch = 20, cex=0.8,title = "Non-DE genes")
dev.off()
```

Script 13: Additional Script (Filter small RNA reads after BLAST).

```perl
use strict;
use Bio::SeqIO;
use Bio::SearchIO;
print "Usage: filtersRNAreads.pl sRNAreads.fa rfamblastoutputfile(-m 9)","\n";

my $sRNA = $ARGV[0];
my $rfamout = $ARGV[1];
#my $mbout = $ARGV[2];
#my $filterset = $ARGV[3];
my %queryid;

my $queries = Bio::SeqIO->new (-file=>$sRNA,-format=>"fasta");
my $in = new Bio::SearchIO(-format => 'blasttable', -file   => $rfamout);
#my $outseq = Bio::SeqIO->new(-file => ">$sRNA.filter", -format => "fasta");

while (my $query = $queries->next_seq()) {
   my $queryid  = $query->display_id();
   my $seqstr =  $query->seq();
while  (my $result = $in->next_result($query)){
while  (my $hit = $result->next_hit){
   my $hsp = $hit->next_hsp ;
   if($hsp->length('total') == 24 && $hsp->percent_identity >= 95) {
      my $rquery = $result->query_name;
      print $rquery,"\t",$hit->name,"\t",$hit->significance, "\t",$hit->num_hsps,"
\n";
        }

   }
 }
}
```