FINE MAPPING OF MALE-STERILITY *ms-1* GENE USING WHOLE GENOME RESEQUENCING OF MUSKMELON (*Cucumis melo* L.)

Thesis

Submitted to the Punjab Agricultural University
in partial fulfillment of the requirement
for the degree of

MASTER OF SCIENCE
in
BIOTECHNOLOGY
(Minor Subject: Plant Breeding and Genetics)

By

Simranjot Kaur
(L-2018-A-177-M)

School of Agricultural Biotechnology
College of Agriculture

©PUNJAB AGRICULTURAL UNIVERSITY
LUDHIANA -141 004

2021

CERTIFICATE I

This is to certify that the thesis entitled, **"FINE MAPPING OF MALE-STERILITY *ms-1* GENE USING WHOLE GENOME RESEQUENCING OF MUSKMELON (*Cucumis melo* L.)"** submitted for the degree of **Master of Science,** in the subject of **Biotechnology** (Minor subject: **Plant Breeding and Genetics**) of the Punjab Agricultural University, Ludhiana, is a bonafide research work carried out by **Simranjot Kaur (L-2018-A-177-M)** under my supervision and that no part of this thesis has been submitted for any other degree.

The assistance and help received during the course of investigation have been fully acknowledged.

_____

(Dr Navraj Kaur Sarao)
Major Advisor
Principal Biotechnologist
School of Agricultural Biotechnology
Punjab Agricultural University
Ludhiana-141 004 (India)

CERTIFICATE II

This is to certify that the thesis entitled, "**FINE MAPPING OF MALE-STERILITY *ms-1* GENE USING WHOLE GENOME RESEQUENCING OF MUSKMELON (*Cucumis melo* L.)**" submitted by **Simranjot Kaur (L-2018-A-177-M)** to the Punjab Agricultural University, Ludhiana, in partial fulfillment of the requirements for the degree of **M.Sc.** in the subject of **Biotechnology** (Minor subject: **Plant Breeding and Genetics**) has been approved by the Student's Advisory Committee along with the External Examiner after an oral examination on the same.

(Dr Navraj Kaur Sarao)
Major Advisor

(Dr Manisha Mangal)
External Examiner
Sr. Scientist, Vegetable Division
IARI, New Delhi

(Dr Parveen Chhuneja)
Head of the Department

(Dr Jaskarn Singh Mahal)
Dean Postgraduate Studies

ACKNOWLEDGEMENT

At the very outset I must bow with all the humility before the Divine Power who has bestowed me with the requisite intelligence, health, and above all the will to carry out the consummation of the stupendous task of this research work. I also thank him for bringing me in contact with learned team of guides, all of whom have contributed to my success.

Though the debt of learning can never be repaid, it is my sovereign privilege to express my profound sense of gratitude and reverence to my Major Advisor ***Dr Navraj Kaur Sarao****, Principal Biotechnologist, School of Agricultural Biotechnology, for her judicious and esteemed guidance and constant encouragement throughout the course of this investigation and preparation of this manuscript.*

I express my sincere thanks to ***Dr Parveen Chunneja****, Senior Molecular Geneticist-cum-Head, School of Agricultural Biotechnology for providing me with the desired facilities for my research work.*

I personally feel obliged to the members of my advisory committee ***Dr Sat Pal Sharma****, Principle Vegetable Breeder, Department of Vegetable Science,* ***Dr Parveen Chunneja****, Senior Molecular Geneticist-cum-Head, School of Agricultural Biotechnology, and* ***Dr Yogesh Vikal****, Dean PGS nominee, School of Agricultural Biotechnology for their intellectual guidance throughout and the successful execution of the work. I would like to express my gratitude to* ***Dr Deepak Singla****, Bioinformatician, School of Agricultural Biotechnology, who helped a lot throughout my research work.*

My vocabulary utterly fails in expressing my profound gratitude to my loving and supporting parents ***Mr. Hardeep Singh*** *&* ***Mrs. Baljinderjeet Kaur*** *grandparents,* ***Mr. Gurmail Singh*** *&* ***Late Mrs. Mohinder Kaur****, and* ***Uncle Sukhdev Singh*** *&* ***Aunt Arvinder Kaur*** *for their unconditional love, constant encouragement, and unwavering belief in me. No choice of words will suffice to adequately register my gratitude towards my loving sisters* ***Amritpal Kaur*** *and* ***Karamveer Kaur*** *for their love and moral upliftment during the period of distress.*

"To hear Complaints with Patience, even when Complaints are Vain, It is one of the Duties of Friendship". I am indebted to my friends ***Sushmita, Ankita, Simrandeep, Simranpreet, Sagar****, and* ***Amandeep*** *who made this journey much more pleasant. Your friendship makes my life a wonderful experience.*

I fully acknowledge the co-operation rendered by the laboratory staff with special thanks to research fellows, my seniors and juniors; ***Dr Gurpreet, Dr Shabda, Dr Pankaj, Jaideep, Mamta, Jasmeen, Nidhi, Dechen, Karamvir, Vishal*** *and* ***Rupinder****, for sharing their valuable knowledge with me during the course of my research work.*

I feel proud and thankful to be a part of Punjab Agricultural University, Ludhiana where I learnt a lot and cherished beautiful and unforgettable memories of my life. Last but not least, I duly acknowledge my sincere thanks to all who showed love and care for me. Everyone may not have been mentioned but none is forgotten.

Date:

Place: Ludhiana *(Simranjot Kaur)*

Title of the Thesis	:	Fine mapping of male-sterility *ms-1* gene using whole genome resequencing of muskmelon (*Cucumis melo* L.)
Name of the student and Admission No.	:	Simranjot Kaur L-2018-A-177-M
Major Subject	:	Biotechnology
Minor Subject	:	Plant Breeding and Genetics
Name and Designation of Major Advisor	:	Dr Navraj Kaur Sarao Principal Biotechnologist
Degree to be Awarded	:	M.Sc.
Year of award of degree	:	2021
Total pages in thesis	:	77 + VITA
Name of the University	:	Punjab Agricultural University, Ludhiana – 141 004, Punjab, India

ABSTRACT

Muskmelon (*Cucumis melo* L.) is a warm season crop, cultivated in tropical and temperate regions of the world. Five types of male-sterility genes have been reported in this crop, out of which *ms-1* and *ms-5* has been utilized in hybrid seed production programmes. The identification of molecular markers linked to these male-sterile genes may find practical applications in hybrid seed production and marker-assisted backcross (MABC) programmes. The previously found SSR markers linked to *ms-1* gene are quite at a distance to be considered as tightly-linked. Therefore, the current study was aimed at fine mapping the genomic loci harbouring the *ms-1* gene using whole genome resequencing approach. A cross between male-sterile MS-1 and male-fertile KP_4HM-15 was used to raise an F_2 mapping population for the construction of two contrasting bulks namely, male-sterile bulk (MSB) and male-fertile bulk (MFB). These bulks along with both the parents were sequenced and analyzed using SNP-index method for identifying the *ms-1* associated region. A 400-kb candidate region on chromosome six was identified using this method which lied within the previously mapped 3.2 Mb region using SSR markers. This candidate region harboured nine highly-significant SNPs and 28 annotated genes. The nine highly-significant SNPs can be used to develop KASP markers for the further improvement of genetic map. It was observed that out of 28 annotated genes, six genes were involved in imparting male-sterility in other crops and therefore these six genes were regarded as candidate putative genes. It was also noticed that the genes fasciclin-like arabinogalactan protein 1 (FLA) and ABC transporter G family member 26, each harboured a highly-significant SNP. Also, one of gene named protein-disulfide isomerase (PDI) contained a missense variation from 'A' to 'T' at 36,921,314 bp position. This missense mutation in the PDI gene might be responsible for the male-sterility in *ms-1* plants.

Keywords: Whole genome resequencing, *ms-1* gene, muskmelon, BSA

Signature of the major advisor **Signature of the student**

ਖੋਜ ਗ੍ਰੰਥ ਦਾ ਸਿਰਲੇਖ	:	ਖਰਬੂਜੇ (*Cucumis melo* L.) ਦੀ ਪੂਰਨ ਜੀਨੋਮ ਪੁਨ:ਅਨੁਕ੍ਰਮਣ ਵਿਧੀ ਦੀ ਵਰਤੋਂ ਕਰਕੇ ਨਰ ਜੀਵਾਣੂ-ਰਹਿਤ *ms-1* ਜੀਨ ਦੀ ਫਾਈਨ ਮੈਪਿੰਗ
ਵਿਦਿਆਰਥੀ ਦਾ ਨਾਮ ਅਤੇ ਦਾਖਲਾ ਨੰ.	:	ਸਿਮਰਨਜੋਤ ਕੌਰ ਐਲ-2018-ਏ-177-ਐਮ
ਪ੍ਰਮੁੱਖ ਵਿਸ਼ਾ	:	ਬਾਈਓਟੈਕਨੋਲਜੀ
ਸਹਿਯੋਗੀ ਵਿਸ਼ਾ	:	ਪਲਾਂਟ ਬ੍ਰੀਡਿੰਗ ਅਤੇ ਜੈਨੇਟਿਕਸ
ਮੁੱਖ ਸਲਾਹਕਾਰ ਦਾ ਨਾਮ ਅਤੇ ਅਹੁਦਾ	:	ਡਾ ਨਵਰਾਜ ਕੌਰ ਸਰਾਓ ਪ੍ਰਿੰਸੀਪਲ ਬਾਈਓਟੈਕਨੋਲੋਜਿਸਟ
ਡਿਗਰੀ	:	ਐਮ.ਐੱਸ ਸੀ.
ਡਿਗਰੀ ਮਿਲਣ ਦਾ ਸਾਲ	:	2021
ਖੋਜ ਪੱਤਰ ਵਿੱਚ ਕੁੱਲ ਪੰਨੇ	:	77 + ਵੀਟਾ
ਯੂਨੀਵਰਸਿਟੀ ਦਾ ਨਾਮ	:	ਪੰਜਾਬ ਐਗਰੀਕਲਚਰਲ ਯੂਨੀਵਰਸਿਟੀ, ਲੁਧਿਆਣਾ - 141 004, ਪੰਜਾਬ, ਭਾਰਤ

ਸਾਰ ਅੰਸ਼

ਖਰਬੂਜਾ (*Cucumis melo* L.) ਇੱਕ ਗਰਮ ਰੁੱਤ ਦੀ ਫ਼ਸਲ ਹੈ ਜਿਸਦੀ ਕਾਸ਼ਤ ਸੰਸਾਰ ਦੇ ਉਸ਼ਮ ਕਟੀਬੰਦ ਅਤੇ ਗਰਮ ਖੇਤਰਾਂ ਵਿੱਚ ਕੀਤੀ ਜਾਂਦੀ ਹੈ। ਇਸ ਫ਼ਸਲ ਵਿੱਚ ਪੰਜ ਤਰ੍ਹਾਂ ਦੇ ਨਰ ਜੀਵਾਣੂਰਹਿਤ ਜੀਨਾਂ ਦਾ ਪਤਾ ਚੱਲਿਆ ਹੈ, ਜਿਹਨਾਂ ਵਿੱਚੋਂ *ms-1* ਅਤੇ *ms-5* ਜੀਨਾਂ ਦੀ ਵਰਤੋਂ ਹਾਈਬ੍ਰਿਡ ਬੀਜ ਉਤਪਾਦਨ ਪ੍ਰੋਗ੍ਰਾਮਾਂ ਵਿੱਚ ਕੀਤੀ ਗਈ ਹੈ। ਇਹਨਾਂ ਨਰ-ਜੀਵਾਣੂਰਹਿਤ ਜੀਨਾਂ ਨਾਲ ਸੰਬੰਧਤ ਆਣਵਿਕ ਮਾਰਕਰਾਂ ਦੀ ਪਹਿਚਾਣ ਨਾਲ ਹਾਈਬ੍ਰਿਡ ਬੀਜ ਉਤਪਾਦਨ ਅਤੇ ਮਾਰਕਰ-ਅਧਾਰਿਤ ਬੈਕਰਾਸ (MABC) ਪ੍ਰੋਗ੍ਰਾਮਾਂ ਵਿੱਚ ਮਦਦ ਮਿਲ ਸਕਦੀ ਹੈ। *ms-1* ਨਾਲ ਸਬੰਧਤ ਪਹਿਲਾਂ ਲੱਭੇ ਗਏ SSR ਮਾਰਕਰਾਂ ਨੂੰ ਕਾਫੀ ਦੂਰੀ ਤੇ ਮੌਜੂਦ ਹੋਣ ਕਾਰਨ ਵਧੇਰੇ ਜੁੜੇ ਹੋਏ ਜੀਨ ਵਜੋਂ ਨਹੀਂ ਮੰਨਿਆ ਜਾ ਸਕਦਾ। ਇਸ ਲਈ, ਮੌਜੂਦਾ ਅਧਿਐਨ, ਪੂਰਨ ਜੀਨੋਮ ਪੁਨ:ਅਨੁਕ੍ਰਮਣ ਦ੍ਰਸ਼ਿਟੀਕੋਣ ਦੀ ਵਰਤੋਂ ਕਰਕੇ *ms-1* ਜੀਨ ਨੂੰ ਪਨਾਹ ਦੇਣ ਵਾਲੇ ਫਾਈਨ ਮੈਪਿੰਗ ਜੀਨੋਮਿਕ ਲੋਕਾਇ ਉਪਰ ਕੇਂਦਰਤ ਸੀ। MSB ਅਤੇ MFB ਬਨਾਉਣ ਲਈ ਨਰ ਜੀਵਾਣੂਰਹਿਤ MS-1 ਅਤੇ ਨਰ-ਜੀਵਾਣੂਯੁਕਤ KP_4HM-15 ਵਿੱਚ ਕਰਾਸਿੰਗ ਕਰਕੇ F_2 ਵੰਸ਼ਾਵਲੀ ਦੇ ਪੌਦੇ ਵਿਕਸਤ ਕੀਤੇ ਗਏ। *ms-1* ਨਾਲ ਸਬੰਧਤ ਖੇਤਰ ਦਾ ਪਤਾ ਲਗਾਉਣ ਲਈ ਦੋਨਾਂ ਮਾਪਿਆਂ ਸਹਿਤ ਇਹਨਾਂ ਬਲਕਸ ਦਾ ਅਨੁਕ੍ਰਮਣ ਕੀਤਾ ਗਿਆ ਅਤੇ SNP-ਇੰਡੇਕਸ ਵਿਧੀ ਦੀ ਵਰਤੋਂ ਕਰਕੇ ਇਹਨਾਂ ਦਾ ਮੁਲਾਂਕਣ ਕੀਤਾ ਗਿਆ। ਇਸ ਵਿਧੀ ਦੀ ਵਰਤੋਂ ਨਾਲ ਗੁਣਸੂਤਰ ਛੇ ਉਪਰ 400-kb ਕੇਂਡੀਡੇਟ ਖੇਤਰ ਦੀ ਪਹਿਚਾਣ ਹੋਈ ਜੋਕਿ ਪਹਿਲਾਂ SSR ਮਾਰਕਰਾਂ ਰਾਹੀਂ ਲੱਭੇ ਗਏ 3.2 Mb ਖੇਤਰ ਦਰਮਿਆਨ ਮੌਜੂਦ ਸੀ। ਕੇਂਡੀਡੇਟ ਖੇਤਰ ਉਪਰ ਨੌਂ ਬਹੁਤ ਵਧੇਰੇ ਅਰਥਪੂਰਨ SNPs ਅਤੇ 28 ਵਿਖਿਆਤ ਜੀਨ ਮੌਜੂਦ ਸਨ। ਅਨੁਵਾਂਸ਼ਿਕੀ ਮੈਪ ਦੇ ਸੁਧਾਰ ਲਈ KASP ਮਾਰਕਰ ਵਿਕਸਤ ਕਰਨ ਲਈ ਬਹੁਤ ਵਧੇਰੇ ਅਰਥਪੂਰਨ SNPs ਦੀ ਵਰਤੋਂ ਕੀਤੀ ਜਾ ਸਕਦੀ ਹੈ। ਇਹ ਦੇਖਿਆ ਗਿਆ ਕਿ 28 ਵਿਖਿਆਤ ਜੀਨਾਂ ਵਿੱਚੋਂ, ਛੇ ਜੀਨ ਦੂਜੀਆਂ ਫ਼ਸਲਾਂ ਵਿੱਚ ਨਰ-ਜੀਵਾਣੂਰਹਿਤਤਾ ਪ੍ਰਦਾਨ ਕਰਦੇ ਸਨ ਅਤੇ ਇਸ ਲਈ ਇਹਨਾਂ ਛੇ ਜੀਨਾਂ ਨੂੰ ਕੇਂਡੀਡੇਟ ਪੁਟੇਟਿਵ ਜੀਨਾਂ ਮੰਨਿਆ ਗਿਆ। ਅਧਿਐਨ ਦੌਰਾਨ ਇਹ ਵੀ ਦੇਖਿਆ ਗਿਆ ਕਿ ਐਰਾਬੀਨੋਗਲੈਕਟਨ ਪ੍ਰੋਟੀਨ 1 (FLA) ਵਰਗੇ ਜੀਨਾਂ ਅਤੇ ABC ਟ੍ਰਾਂਸਪੋਰਟਰ G ਪਰਿਵਾਰਕ ਮੈਂਬਰ 26 ਨੇ ਬਹੁਤ ਵਧੇਰੇ ਅਰਥਪੂਰਨ SNP ਨੂੰ ਪਨਾਹ ਦਿੱਤੀ। ਪ੍ਰੋਟੀਨ ਡਾਈਸਲਫਾਈਡ ਆਈਸੋਮੀਰੇਜ਼ (PDI) ਨਾਮਕ ਇੱਕ ਹੋਰ ਜੀਨ ਨੇ 36,921,314 bp ਉਪਰ 'A' ਤੋਂ 'T' ਤੱਕ ਵਿਭਿੰਨਤਾ ਦਰਸਾਈ। PDI ਵਿੱਚ ਗਤਲ ਮੁਟੇਸ਼ਨ *ms-1* ਪੌਦਿਆਂ ਵਿੱਚ ਨਰ-ਜੀਵਾਣੂਰਹਿਤਤਾ ਲਈ ਜ਼ਿੰਮੇਵਾਰ ਹੋ ਸਕਦੀ ਹੈ।

ਮੁੱਖ ਸ਼ਬਦ: ਪੂਰਨ ਜੀਨੋਮ ਪੁਨ:ਅਨੁਕ੍ਰਮਣ, *ms-1* ਜੀਨ, ਖਰਬੂਜਾ, BSA

ਪ੍ਰਮੁੱਖ ਸਲਾਹਕਾਰ ਦੇ ਹਸਤਾਖਰ | **ਵਿਦਿਆਰਥੀ ਦੇ ਹਸਤਾਖਰ**

CONTENTS

CHAPTER	TITLE	PAGE NO.
I	INTRODUCTION	1 – 3
II	REVIEW OF LITERATURE	4 – 16
III	MATERIAL AND METHODS	17 – 28
IV	RESULTS AND DISSCUSSION	29 – 67
V	SUMMARY	68 – 70
	REFERENCES	71– 77
	VITA	

LIST OF TABLES

Table No.	Title	Page No.
3.1	Composition of 2X CTAB extraction buffer	20
3.2	Composition of RNase buffer for preparing 100ml RNase solution	20
3.3	Quality scores and base calling accuracy (Table reconstructed from https://www.illumina.com/documents/products/technotes/technote_Q-Scores.pdf)	24
4.1	Phenotypic segregation pattern of F_2 population	33
4.2	The description of the sequence identifier (Line1) contents in a .fastq file	34
4.3	Statistics of the data obtained after sequencing.	35
4.4	Summary of sequencing data- raw input reads and reads obtained after trimming	37
4.5	Table of intervals (1 Mb sliding window) with mean ΔSNP-index values greater than the statistically calculated ΔSNP-index values at $P \leq 0.01$	40
4.6	Table of intervals (1 Mb sliding window) with mean ΔSNP-index values greater than the statistically calculated ΔSNP-index values at $P \leq 0.05$	40-41
4.7	Positions of highly significant SNPs having ΔSNP-index = 1 ($P \leq 0.01$)	42
4.8	Blast2GO annotations for 43 genes in the 400-kb candidate region	50-51
4.9	List of six putative candidate genes and their known functions in male-sterility	55
4.10	Chromosome-wise distribution and rate of variants	56
4.11	Number of variants by type	57
4.12	Number of effects by impact	57
4.13	Number of effects by functional class	57
4.14	Number of effects by type and region	58-59
4.15	Variant annotation and impact of nine highly significant SNPs on chromosome six	63

LIST OF FIGURES

Fig No.	Title	Page No.
2.1	(a) Maintenance of recessive male-sterile plants and (b) their utilization in the hybrid seed production (Kumar *et al* 2000)	6
2.2	Backcrossing and self-pollinating all the progenies simultaneously to identify segregating (*Msms*) and non-segregating (*MsMs*) individuals for a recessive male-sterile gene (Kaul 1998)	7
2.3	Principle of SNP-index method for identifying the genomic region associated with the trait of interest (Takagi *et al* 2013)	12
3.1	Flow-chart for the development of F_2 mapping population and their F_3 progenies	18
3.2	Samples used for whole genome resequencing	22
4.1	Anthers as observed under light microscope (a) indehiscent anthers observed in MS-1 inbred line and (b) anthers of KP_4HM-15 covered with visible pollen grains	30
4.2	(a) Female flowers of MS-1 with sterile anthers, (b) Male flowers of KP_4HM-15 with fertile anthers. The pollens from the male flowers of KP_4HM-15 were used to pollinate the female flowers of MS-1	30
4.3	F_2 plants raised from MS-1 X KP_4HM15 in green house	31
4.4	Polymorphism between MS-1 and KP_4HM-15 for fruit traits. (a) MS-1 fruit with netted skin, dark-green rind and oval shape, (b) KP_4HM-15 fruit with non-netted skin, light-green rind and round shape, (c) orange flesh of MS-1, and light-green flesh of KP_4HM-15	31
4.5	Polymorphism among the fruits of F_2 segregates with respect to external fruit characters like netting, rind color and shape (a), (b), (c), and flesh color (d)	32
4.6	(a) Fertile anthers flooded with pollens and (b) sterile anthers devoid of pollens	32
4.7	An image displaying the contents of the MS-1.fastq file	35
4.8	FastQC results before (a,c,e) and after trimmomatic (b,d,f). (a) and (b) shows the per base sequence quality, (c) and (d) shows the per base sequenc content, and (e) and (f) displays the presence of adapter content	36
4.9	Contents of a .vcf file generated after variant-calling using BCFtools	38
4.10	Contents of the SNP_inde.tsv file	40
4.11	Contents of the sliding_window.tsv file	40
4.12	SNP-index plot between male-sterile bulk (MSB) and MS-1 assembly	44
4.13	SNP-index plot between male-fertile bulk (MFB) with MS-1 assembly	45

Fig No.	Title	Page No.
4.14	ΔSNP-index plot (SNP-index MFB – SNP-index MSB). Green and orange lines represent the 95% and 99% confidence interval of simulated ΔSNP-index	46
4.15	Comparison of the mapped region of *ms-1* gene using QTL-seq (present study) and SSR markers (Singh *et al* 2019). The previously mapped 3.2 Mb region (d) was narrowed down to 400 kb region (a, b, and c) on chromosome six	47
4.16	Distribution of variants used in Qtl-seq software	48
4.17	A bar graph representing the distribution of genes undergone each step in Blast2GO	49
4.18	Functional annotation for the 43 genes present within the 400 kb candidate region	52
4.19	Proportion of SNPs and InDels in the 400-kb candidate region	59
4.20	Density of variants (SNPs and InDels) within the 28 annotated genes	60
4.21	Distribution of variant impacts in the 400-kb candidate region	61
4.22	The distribution of different variant effects caused by the variants present within the six candidate genes	62
4.23	IGV genome browser view of the missense SNP position in the PDI gene	62
4.24	Positions of highly significant SNPs within (a) gene id MELO3C014103.2, (b) gene id MELO3C014107.2 (c) intergenic region MELO3C014111.2 to MELO3C014112.2, (d) gene id MELO3C014115.2, (e) intergenic region MELO3C014117.2- MELO3C014119.2, (f) intergenic region MELO3C014117.2-MELO3C014119.2, (g) gene id MELO3C014121.2, (h) MELO3C014122.2, and (i) intergenic region MELO3C031975.2-MELO3C014126.2 as viewed under IGV software	64-66

CHAPTER I

INTRODUCTION

Muskmelon (*Cucumis melo* L.) is a diploid crop (2n = 2x = 24) belonging to the family *Cucurbitaceae,* which is also known as the gourd family. Besides muskmelon, cucumber (*Cucumis sativus* L.), watermelon (*Citrullus lanatus*), pumpkin (*Cucurbita maxima*), gourd (*Cucurbita* spp.), and squash (*Cucurbita* spp.) are some of the commercially important species of this family. Melons display tremendous variation with respect to fruit size, shape, color, taste, and texture; and therefore hold a great nutritional, economical, and biological significance (Kirkbride 1993, Whitaker and Davis 1962). The total genome size of muskmelon was estimated to be around 454 Mb (Arumuganathan & Earle 1991), out of which 417 Mb has been successfully assembled by sequencing the double haploid line, DHL92 (Garcia-Mas *et al* 2012 and Ruggieri *et al* 2018).

Muskmelon is a summer season crop and is grown in tropical and sub-tropical regions of the world. 90% of the fruit contains water and the rest 10% is carbohydrate. The fruits of muskmelon are appreciated and savored all over the world owing to their textured flesh, musky aroma, high sugar content, and health-promoting phytonutrients. The nutritional profile of muskmelon comprises vitamin-A (β-carotene), vitamin-C (Ascorbic acid), folic acid, and potassium (Lester 2008). Low amounts of cholesterol, sodium, and fat combined with therapeutic properties like anti-inflammatory, anti-microbial, anti-cancer, anti-oxidant, diuretic, analgesic, etc., have led to increased demand for fruit consumption worldwide (Parle & Singh 2011). Moreover, the seeds of muskmelon are used as soup thickening agent, for extraction of vegetable oil, and also as desserts and roasted snacks (Mallek-Ayadi *et al* 2018).

Currently, the world production of melons is 27 million tons annually, with Asia and America being the largest producers (Anonymous, 2019). China is the major producer with an annual production rate of 13.48 million tons followed by Turkey, and India. In India, the area under muskmelon cultivation is around 57 thousand hectares with a total production of 1.2 million tons and an average yield of 24.48 tons ha^{-1} (Anonymous 2019). Within India, Uttar Pradesh is the highest producer with 548.67 thousand tons followed by Andhra Pradesh (314.39 thousand tons), Madhya Pradesh (127.48 thousand tons) and Punjab (99.10 thousand tons) (Anonymous 2017-18). In Punjab, muskmelon was cultivated on 6.34 thousand hectares land, with a total production of 114.68 thousand tons and an average yield of 19.94 tons ha^{-1} (Anonymous 2020).

Muskmelon exhibits enormous diversity with respect to agro-morphological as well as economical traits and the availability of such wide diversity allows for the selection and breeding of melons for increased yield and quality. The target traits for producing superior

melon varieties/hybrids are early maturing, high yielding, resistance to biotic & abiotic stresses, along with various fruit characters like uniform shape and size of fruits, tough netted skin, small seed cavity, high TSS (total soluble salts), and flesh texture, color, and thickness (Choudhary & Pandey 2010). The availability of pollination control mechanisms like genic male-sterility (GMS), monoecy, and gynoecy in muskmelon crop have rendered hybrid production as an easy job.

Muskmelon commonly produces monoecious and andromonoecious types of flowers and hence is predominantly a cross-pollinated crop. Cross-pollination is achieved naturally by bees and manually by hands after emasculations (Pitrat 2008). Commercial production of pure hybrid seeds requires large-scale hand emasculations and pollinations which is a costly venture. Furthermore, the hybrid seeds so produced cost twelve to thirty times higher as compared to open-pollinated varieties (McCreight & Elmstrom 1984). The existence of GMS, monoecy, and gynoecy in melons has been successfully exploited for commercial production of hybrid seeds (Munshi & Alvarez 2005). Male-sterility, being a breeder's tool provides a perfect outcrossing system by eliminating the tedium of hand emasculation during hybrid production, thereby reducing the cost and labour.

The presence of GMS genes in muskmelon has proved to be of immense importance for obtaining pure muskmelon hybrids (McCreight & Elmstrom 1984). Till now five male-sterile genes have been reported in muskmelon, viz. *ms-1* (Bohn & Whitaker 1949), *ms-2* (Bohn & Principe 1964), *ms-3* (McCreight & Elmstrom 1983a), *ms-4* (Lozanov 1983), and *ms-5* (Lecouviour *et al* 1990). All these male-sterile genes are recessive and non-allelic in nature (Nagaraju *et al* 2017). The MS-1 genotype was introduced in India in 1978 and since then this genotype has been used to produce various hybrids like Punjab Hybrid (Nandpuri *et al* 1982), Punjab Anmol (Lal *et al* 2007), and MH-27 (PAU 2018). Besides these, MH-51 hybrid has recently been developed using *ms-5* male-sterile gene (PAU 2018).

A recessive male-sterile (*msms*) gene containing line is maintained by crossing it with heterozygous dominant male-fertile plant (*Msms*). A progeny from such a cross will segregate into male-sterile (*msms*) and male-fertile (*Msms*) genotypes in 1:1 ratio. The male-sterile (*msms*) plants so obtained are used in the hybrid production while the rest heterozygous male-fertile (*Msms*) plants are used as maintainer/restorer lines. Mostly the genotypes in which the male-sterile gene arises are not of agronomic importance and hence require the transfer of the gene from the source to the desirable genotypes (Kaul 1998). This is done through backcrossing in which the donor parent is the line containing male-sterile gene and the recipient or recurrent parent is a cultivated line containing all the agronomically valuable traits.

The transfer of recessive male-sterile genes using classical Mendelian approach of backcrossing is a very tedious process. The progeny from such a backcross segregates into homozygous (*MsMs*) and heterozygous (*Msms*) male-fertile genotypes, which are phenotypically indistinguishable. In such cases, all the plants are backcrossed to the recurrent parent and at the same time, self-pollinated to detect the heterozygous male-sterile (*Msms*) plants. The backcross progeny from *MsMs* plants are then discarded (Kaul 1998).

The efficiency of the traditional backcrossing method for gene transfer can be improved by identifying tightly-linked markers to the trait of interest. These markers allow for the identification, mapping, and hence indirect selection of the target traits. Although morphological markers have been reported for *ms-1* and *ms-2* genes, the linkage is loose (McCreight & Elmstrom 1983b and Pitrat 1991). Recently, researchers have utilized DNA markers and have successfully mapped *ms-3* (Park & Crossby 2004), *ms-5* (Sheng *et al* 2017), and *ms-1* (Singh *et al* 2020) genes. Sheng *et al* (2017) were even able to identify the candidate gene (*ABORTED MICROSPORES*) underlying the *ms-5* phenotype.

Singh *et al* (2020) mapped the *ms-1* gene in an F_2 population obtained from a cross between male-sterile MS-1 and male-fertile KP_4HM-15 using SSR (Simple Sequence Repeat) markers at Punjab Agricultural University (PAU), Ludhiana. They were able to map the *ms-1* gene to a 20.3 cM region on chromosome six between two SSR markers namely, DM0187 and DM0038. The physical distance enclosed by these markers corresponds to a 3.2 Mb region which is still a very large genomic region for considering these as tightly-linked markers.

Keeping the above situation in view, the current study aims at narrowing down this 3.2 Mb *ms-1* associated region using whole-genome resequencing approach. This narrowed-down region can be further used to identify and clone the gene responsible for *ms-1* phenotype. The cloning of the underlying gene will make it easier for the production of the new GMS-line, thereby eliminating the time-consuming method of gene transfer via backcrossing. Thus the present investigation was proposed with the following objectives:

i. Whole-genome resequencing and identification of SNPs

ii. Fine mapping of genetic male-sterile *ms-1* gene

CHAPTER II

REVIEW OF LITERATURE

Muskmelon (*C. melo*) is an annual vegetable crop belonging to family *Cucurbitaceae*. It is cultivated in tropical and temperate climatic conditions. 90% of melon contains water and the rest 10% is carbohydrate. They are also rich in dietary fiber, vitamin-A and vitamin-C (Munshi & Alvarez 2005). Sweet melons are savored as desserts whereas the non-sweet types are either consumed raw, pickled, or cooked. Fruits of muskmelon exhibit high morphological and genetic variations, and are of immense biological importance (Sageret 1826 and Stepansky *et al* 1999). Heterosis for various economic traits has been observed in muskmelon (Choudhary & Pandey 2010). Hybrid varieties exhibiting marketable traits like early maturity, high yield, uniform and better fruit quality have high demand, particularly in Mediterranean and East Asian countries, and hence are of great economic value (Garcia-Mas *et al* 2012). Various pollination control mechanisms like genic male-sterility (GMS), monoecy, and gynoecy can be used as breeder's tools to increase the production of melon hybrids (Munshi & Alvarez 2005). Five types of recessive nuclear male-sterile genes have been reported in muskmelon. The utilization of GMS as a natural out-crossing system has been exploited extensively in muskmelon for hybrid seed production. GMS acts as a genetic means for emasculation and facilitates the development of pure F_1 hybrids. New GMS lines can also be developed by transferring male-sterile gene through backcross techniques. However, no GMS line has been developed so far because the conventional backcross (BC) technique of transferring a recessive gene is laborious and time-demanding. The efficiency of GMS can be improved by using a closely linked molecular marker which can be used in marker-assisted back-cross (MABC), facilitating the transfer of male-sterile gene into the desired superior inbred line. The closely linked genetic marker can also be used to detect sterile plants so that fertile plants can be rouged out from the hybrid block at an early stage.

2.1 Male-sterility and its types

Male-sterility is a condition in which male gametes (microgametophytes or pollen grains) are either absent or non-functional whereas the female gametes (megagametophytes or embryo sac and eggs) function normally. Male-sterility is the result of mutations in the DNA, either nuclear or cytoplasmic or both. On the basis of genotype, male-sterility is classified into three types (Kaul 1998), viz. Genic male-sterility (GMS), Cytoplasmic male-sterility (CMS), and Genic-cytoplasmic male-sterility (GCMS).

i. **Genic male-sterility:** This type of male-sterility is controlled by nuclear genes and thus follows Mendelian inheritance and expression patterns. In the majority of plants, GMS is recessive in nature, but there are few exceptions (e.g cabbage and broccoli) that show

dominant behavior. GMS is caused by the occurrence of mutation in any gene(s) controlling either microsporogenesis (stamen development process) or microgametogenesis (pollen development process).

ii. **Cytoplasmic male-sterility (CMS):** This type of male-sterility is governed by the male-sterility inducing cytoplasm, designated as sterile (S) cytoplasm. The normal cytoplasm which does not confer male-sterility is designated as normal fertile (N) cytoplasm. Since the zygote carries the cytoplasm from female parent, CMS shows maternal inheritance.

iii. **Genic-cytoplasmic male-sterility (GCMS):** The interaction of nuclear genes (*fr* genes) and S-cytoplasm govern the sterility in GCMS system. The plant species exhibiting GCMS system possess both *Fr* (male-fertile) and *fr* (male-sterile) nuclear genes as well as N- and S-cytoplasm types. S-cytoplasm possesses specific c-genes which interact with nuclear *fr* genes to produce male-sterile phenotype so that only S_{frfr} individuals are male-sterile. The nuclear *fr* genes are ineffective in N-cytoplasm and similarly S-cytoplasm is ineffective in presence of *Fr* nuclear genes. Such genotypes namely, $N_{Fr_}$, N_{frfr} and $S_{Fr_}$ are male-fertile.

2.2 Genic male-sterility and its breeding value

Genic male-sterility (GMS) is the result of mutations, either occurring naturally or induced after mutagen treatments, in the nuclear genes. In both cases, majority of the mutations are recessive and monogenic in nature. However, dominant as well as polygenic control of GMS has also been reported in crops (Kaul 1998). These mutations in the nuclear genes disrupt the development and phenotype of anthers. Therefore, male-sterility can also be classified into structural, sporogenous and functional types. Structural sterility is when the differentiation and development of androecium (stamens) is altered and as a result no viable pollen is formed. In such cases, stamens may be absent, aborted, deformed or even transformed into other floral structures. In sporogenous types of male-sterility, differentiation and development of microsporogenous tissue are impaired by the action of male-sterile genes. In this, the male-sterile genes may act either at the pre-meiotic stage (transformation of PMC to meiotic cells), meiotic stage (tetrad formation), or post-meiotic stage (degenerated microspores and abnormal tapetum). In the functional form of male-sterility, viable and functional pollen is formed but the male-sterile gene action inhibits anther dehiscence and hence pollen is not released. Of the three types of male-sterility, sporogenous type is more frequent in GMS.

GMS finds its utility in commercial hybrid seed production sector, as it acts as a genetic tool to overcome the tedium of artificial hand emasculation and pollination. The presence of stable GMS ensures the development of pure F_1 hybrid seeds, as it eliminates the

chance of contamination by self-pollination. Thus, the first step of GMS utilization in breeding programs is the detection and maintenance of stable GMS lines. The dominant GMS gene is relatively easy to maintain, as the cross involving dominant GMS gene produces 50% individuals with male-sterile genotype (Rao *et al* 1989). However, in case of recessive condition (*msms*), 50% male-sterile individuals are only formed when crossed with heterozygous fertile plant. Therefore, in hybrid seed production field, 50% male-fertile segregants (*Msms*) need to be identified and removed before they shed pollen (Fig. 2.1)

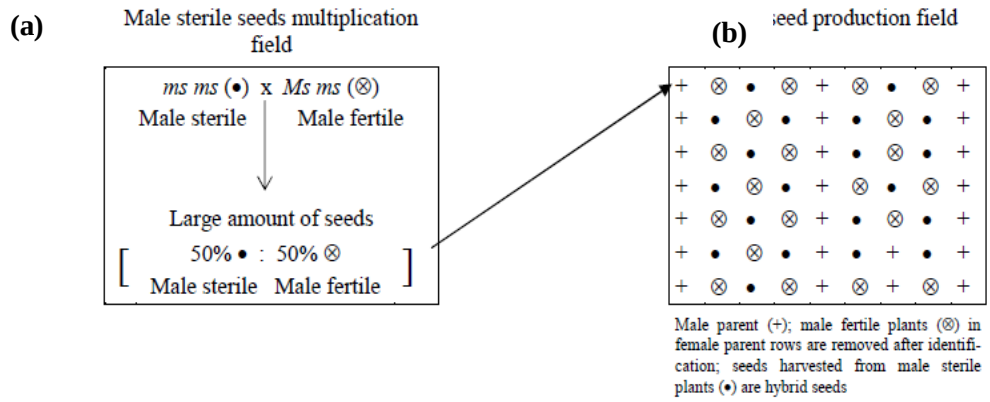


Fig. 2.1: (a) Maintenance of recessive male-sterile plants and (b) their utilization in the hybrid seed production (Kumar *et al* 2000)

Furthermore, the breeding value of GMS system can be enhanced by developing new GMS lines possessing desirable horticultural traits, which can be further utilized in the hybrid seed production programmes. These lines are generated by transferring male-sterile gene from the source to desirable genotypes via backcrossing (BC) technique. The dominant male-sterile gene can be easily transferred using a usual backcross procedure. However, it takes up-to five to six BCs to transfer the male-sterile gene and recover the recipient's genotype, followed by the maintenance of lines using pedigree or bulk methods. On the other hand, the transfer of recessive male-sterile gene using BC is a complex procedure and requires a longer generation time as compared to the dominant one. This is because the progeny from such genes segregate into *MsMs* and *Msms* which are difficult to distinguish phenotypically. To distinguish between the two, either the progeny is first self-pollinated to find the segregating individuals, which are then selected for backcrossing; or by backcrossing and self-pollinating all the progenies at the same time and identifying the segregating individuals (*Msms*) which are carried forward while the non-segregating ones (*MsMs*) are discarded (Fig. 2.2).

2.3 Genic male-sterility in Muskmelon

Muskmelons predominantly produce andromonoecious flowers, but gynomonoecious and monoecious types also occur, as a result muskmelon is mainly a cross pollinated crop. The existence of GMS, monoecy and gynoecy as pollination control mechanisms has been

successfully exploited for commercial seed production in melons (Nandpuri *et al* 1982 and Lal *et al* 2007), thereby reducing the hybrid seed production costs (Munshi & Alvarez 2005).

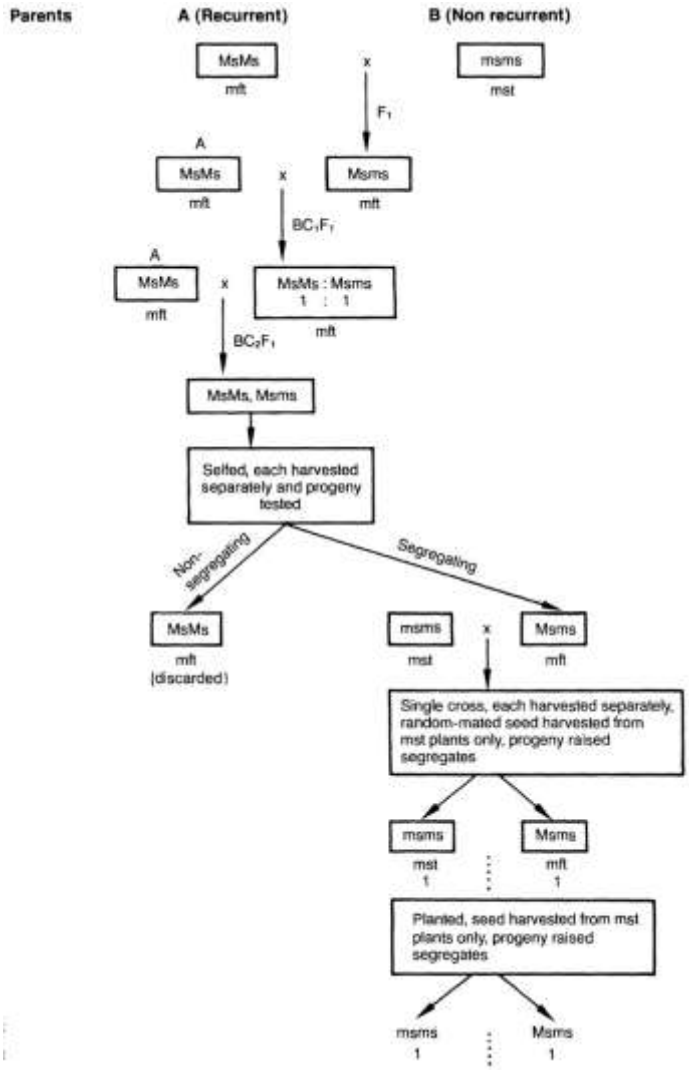


Fig. 2.2: Backcrossing and self-pollinating all the progenies simultaneously to identify segregating (*Msms*) and non-segregating (*MsMs*) individuals for a recessive male-sterile gene (Kaul 1998)

The use of male-sterility in hybrid production has been reported in many crops including chilli (Hundal and Khurana 1993), cabbage (Fang *et al* 1997), pepper (Shifriss 1997), onion (Pathak CS 2000), and tomato (Sawhney 2004).

The only type of male-sterility reported in muskmelon till date is nuclear or genic male-sterility. GMS in muskmelon is governed by five recessive nuclear genes namely, *ms-1* (McCreight & Whitaker 1949), *ms-2* (Bohn & Principe 1962), *ms-3* (McCreight & Elmstrom 1984), *ms-4* (Lozanov 1983) and *ms-5* (Lecouviour *et al* 1990). All these genes are non-allelic and located on different linkage groups (Pitrat 1991). Each male-sterile gene exhibits unique phenotype (Pitrat 2002). The *ms-1* mutant gene displays empty pollen walls at tetrad stage

and produces non-dehiscent anthers (Bohn & Whitaker in 1949). The *ms-2* gene was observed by Bohn and Principe in 1964 while cultivating the powdery mildew resistant cantaloupe line, La Jolla 40460. The stamens of *ms-2* mutant plants are smaller than the normal plants. The third male-sterile gene, *ms-3* was identified in line PI 321005 (Georgia 47 x Smith's Perfect) by McCreight and Elmstrom in 1983 at Leesburg, Florida. The anthers of *ms-3* mutant plants are dull, waxy, and translucent and the phenotype is evident to the naked eye (McCreight & Elmstrom 1984). The *ms-4* mutant was identified by Lozanov in 1983 and the phenotype is easy to detect in the field as the male-sterile flower abort at bud stage. Lastly, the *ms-5* mutant was spotted by Lecouviour in 1990 while developing the powdery mildew resistant line PMR 45. This mutant was later used by the Clause Seed Company, United States for producing F_1 hybrids.

Despite the presence of five male-sterile genes in muskmelon, only *ms-1* has proved its potential in hybrid production. Punjab Hybrid (Nandpuri *et al* 1982), Punjab Anmol (Lal *et al* 2007) and MH-27 (PAU 2018) are the commercially developed hybrids using MS-1 genotype. Recently *ms-5* has also been utilized to develop MH-51 (PAU 2018) hybrid. The lack of inclination towards the use of GMS in hybrid breeding programs may be due to the difficulty in identifying and rouging out the male-fertile plants before pollinations. Moreover, most of the male-sterile gene carrying lines do not possess desirable horticultural traits and transfer of the male-sterile gene via conventional backcrossing method is time-consuming. However, with the construction of genetic maps and availability of tightly-linked molecular markers, conventional breeding programs can be accelerated.

2.4 Mapping genes using molecular markers

Genetic mapping refers to the determination of relative positions of genes on a DNA molecule. Akin to a road map, genetic map represents a relative genetic distance between markers or genes along chromosomes (Collard *et al* 2005). The principle of genetic mapping is based on the fact that genes or markers segregate during recombination at meiosis, thereby allowing their analysis in the progeny (Paterson 1996). Mendel's second law of independent assortment of alleles holds true for genes located on different chromosomes or located far away from each other on the same chromosome. But genes located close to each other on the same chromosome do not assort independently. This is because the distance between two genes is directly related to the frequency of recombination events between them, that is, lesser the distance between the two genes, less are the chances for the recombination event to take place between them. As a result, these two genes are more likely to be inherited together in the progeny (Semagn *et al* 2006). Such genes which are frequently inherited together are called linked genes and this linkage of marker sequences with the genes is used to map or tag

the genes. The construction of linkage or genetic maps requires the development of an appropriate mapping population for collecting phenotypic data and application of molecular markers for generating genotypic data, which are then analyzed to produce the linkage or genetic maps.

The selection of plants based on the morphological characters has long been known. Gregor Mendel used morphological characters in his experiments to study the inheritance patterns in pea (*Pisum sativum*). He randomly selected the contrasting traits in his experiments, which were not influenced by the environmental conditions and hence showed stable expression. These traits were later termed as *qualitative* traits. On the other hand, the *quantitative* traits exhibit continuous variation in their phenotype due to environmental influences. Therefore, phenotype based selections for these traits often goes in vain. The genetic investigations for these traits led to the identification of genetic linkage with various other morphological markers (Sax 1923 and Thoday 1961). The utility of these morphological markers was realized as they aided in the indirect selection of desirable traits. Since then these markers have evolved from morphological markers to biochemical markers (isozymes) and finally, to molecular markers (DNA-based).

DNA-based markers are nucleotide sequences with known chromosomal locations and are closely linked to the gene of interest (Nadeem *et al* 2018). They act as 'signs' or 'flags' and help in locating genes at specific genomic positions within the chromosomes (Collard *et al* 2005). On the basis of detection methods, DNA markers are classified into,

i. Hybridization-based markers like RFLP,

ii. Polymerase chain reaction (PCR)-based markers such as RAPD, AFLP, ISSR, EST, CAPS, SSR etc., and

iii. Sequencing-based markers like SNPs.

The development of molecular markers greatly facilitated research in various disciplines like taxonomy, phylogeny, ecology, genetics and plant breeding. The co-segregation data of markers and traits/phenotypes in a defined population is used to generate saturated molecular maps. SSR markers and SNP markers have been the markers of choice for mapping genes in the field of genetics and plant breeding. Simple Sequence Repeats (SSRs) or microsatellites are short tandem repeats of 1-6 bp length and are uniformly distributed throughout the genomes of both eukaryotes and prokaryotes. The polymorphism at SSR loci is exclusively due to the variation in the number of repeat units. The unique sequences flanking the SSR loci are conserved within species and even across species. The polymorphism is thus detected by amplifying the SSR loci of organisms, using specific primers derived from the conserved flanking sequences. SSR markers have been the most

preferred markers over the last two decades because of their high levels of polymorphism, co-dominant nature, and uniform distribution (Mason 2014). Also, the assay is easy to perform, exhibit high reproducibility and is amenable to high throughput (Singh & Singh 2015).

Recently, high-throughput sequencing-based markers gained popularity, shifting the attention from SSRs to SNP-based markers. SNP or single nucleotide polymorphism is a single base variation and constitute approximately about 90% of the genetic variation in any organism (Singh & Singh 2015). With the advent of high-throughput (HTP) next generation sequencing (NGS) technologies it has been possible to mine or identify millions of SNPs in genomes. In contrast with other molecular markers, SNPs discovery relies on the comparison of homologous sequences between different genotypes to identify sequence level variations (Paux *et al* 2011). Also, because of bi-allelic nature of SNPs, these require less time for large-scale genotyping than those by the earlier-generation molecular markers. The NGS technology offers significant amount of inexpensive and accurate genome sequence information, and is also effective in producing SNP markers than any other genetic DNA fingerprinting methods.

2.5 Bulked Segregant Analysis

BSA or bulked segregant analysis is a rapid and efficient method for identification of trait-linked markers (Michelmore *et al* 1991 and Giovannoni *et al* 1991). The technique involves preparation of two bulks or pools by combining the DNA samples from individuals contrasting for a particular trait (e.g., tall vs. short plants or resistant vs. susceptible plants) in a population. The individuals within the bulk are genetically identical for a selected region (trait loci) but arbitrary at all other regions. The alleles for the target gene gets enriched within each pool because of bulking, that is one bulk would be enriched with allele *A* while the contrasting bulk would be enriched with allele *a*. Whereas, the alleles for all other genes would be randomly segregated within the two bulks (Zhang and Panthee 2020). Because of deliberate selection of individuals for preparing contrasting bulks, the genomic region harboring the target gene would differ considerably between the two pools, which thus provide the basis for the selection of polymorphic markers. The polymorphic markers obtained using BSA, are more likely to be linked to the target region or trait as compared to the conventional analysis of entire population. The BSA therefore simplifies the procedure as well as reduces the scale and cost of trait-linked marker identification or gene mapping.

Michelmore *et al* (1991) introduced bulked segregant analysis as a rapid method for detecting molecular markers linked to any target gene or genomic region of interest. Two DNA bulks are constructed from segregating population advanced from a single cross. Each bulk consists of individuals that are identical for a specific genomic region or trait but random

at all other regions. Thus, the bulks are genetically dissimilar in the target region and seemingly heterozygous at all unlinked regions. The principle underlying BSA is that pooling of known individuals allows the analysis of a specific genomic region against an arbitrary genetic background. Markers within 25-centimorgan on either side of the target locus can be reliably identified using this technique. This approach was expected to have extensive application in self-pollinated as well as obligatorily outbreeding species. BSA is simple and rapid as compared to the use of near-isogenic lines as only a segregating population from a single cross is required.

2.6 BSA-seq analysis

BSA coupled with whole genome sequencing (BSA-Seq) accelerates the identification of the regions associated with the trait of interest. The technique involves the whole genome sequencing of two extreme bulks followed by the BSA-Seq data analysis. Either SNP index method (Takagi *et al* 2013) or G-statistic method (Magwene *et al* 2018) is followed to analyze the BSA-Seq data. Both these methods quantify reference base or alternate base enrichment of a single SNP to identify the SNP-trait associations (Zhang & Panthee 2020). SNP index method is the widely used method for analyzing BSA-Seq data and has been used to map genes in rice (Yang *et al* 2013), pigeonpea (Singh *et al* 2016), soybean (Song *et al* 2017), watermelon (Dou *et al* 2018), pepper (Zhang *et al* 2018), cucumber (Pujol *et al* 2019), tomato (Liu *et al* 2019), groundnut (Kumar *et al* 2020), sunflower (Liu *et al* 2020), and many more.

The principle underlying the BSA-seq data analysis via SNP-index method is shown in Fig. 2.3 (Takagi *et al* 2013). BSA-seq involves the development of two contrasting bulks by pooling 10-20 individuals from the segregating population (F_2, recombinant inbred lines and double haploids), followed by their sequencing along with either of the one parent or both. In Fig. 2.3, two contrasting bulks for plant height namely; ‘highest bulk’ for high stature and ‘lowest bulk’ for the low stature were developed and sequenced along with the parents, cultivar A (low stature) and cultivar B (high stature). The sequenced reads of the bulks were then aligned to the reference genome of one of the parental cultivar, in this case cultivar A. It is expected that the bulks contain genomic regions contributed equally by both the parents for majority of the genome except the region containing the gene of interest. This region will contain unequal distribution of parental genomes with respect to the phenotypic difference between the highest and the lowest bulks. The proportion of the parental genomes in the bulks were quantified by evaluating the number of short reads corresponding to each of the two parental genomes differentiated by a SNP between the two. After aligning the bulks’ sequenced reads to the reference genome of cultivar A, SNPs were identified and the short

reads containing SNP allele different than the reference allele were counted (k). The proportion of short reads containing alternate allele (k) to the total number of short reads (n) for a particular genomic region is referred to as the SNP-index (kn^{-1}, Fig. 2.3b).

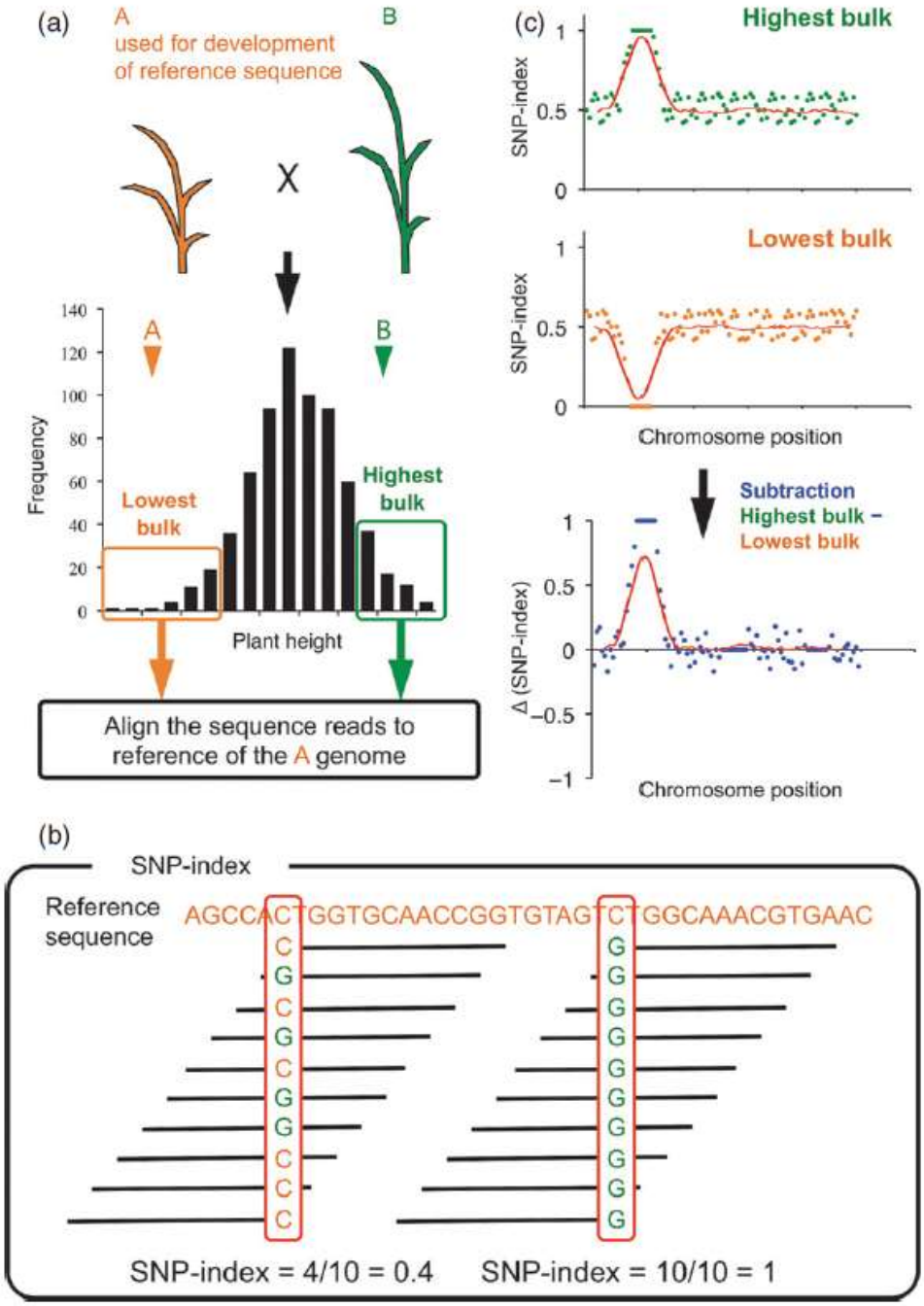


Fig. 2.3: Principle of SNP-index method for identifying the genomic region associated with the trait of interest (Takagi *et al* 2013)

A SNP-index of 0.5 indicates that the SNP containing reads for a given genomic fragment are equally contributed by both the parents whereas the SNP-index of 0 and 1 indicates that the SNP containing read is contributed by reference parent (cultivar A) and the

non-reference parent (cultivar B), respectively. Similarly, SNP-index values for all the SNPs are calculated and plotted with respect to their genomic positions (Fig. 2.3c). Sliding window analysis is used to visualize the graphs, where the average of SNP-index values within a sliding window of a particular size are plotted and depicted as a red line in the SNP-index graph (Fig. 2.3c). It is expected that the graphs of the two bulks to be identical for all the genomic regions except the ones corresponding to the phenotypic differences between the two bulks. The graphs for both the bulks at this region appear as mirror image to each other with respect to the line of SNP-index = 0.5. Such regions have high probability of harbouring genes associated with the trait of interest. Furthermore, the ΔSNP-index values are also calculated along with the threshold values (confidence intervals) which are calculated via computer simulations at 99% and 95% level of significance. The values for each SNP (Fig. 2.3c) are calculated by subtracting the SNP-index values of bulk 1 from the SNP-index values of bulk 2 (SNP-index of *highest bulk* – SNP-index of *lowest bulk*). The greater the absolute values of ΔSNP-index, higher are the chances that the SNP contributes to the trait of interest.

2.7 Mapping of genetic male-sterile genes in muskmelon

The male-sterile gene, *ms-3* was identified by McCreight and Elmstrom in 1984. Mapping of *ms-3* gene was carried out by Park and Crossby (2004) using RAPD (Randomly Amplified Polymorphic DNA) markers. The most tightly linked RAPD marker so obtained was then converted to SCAR (Sequence Characterized Amplified Region) marker. An F_2 population obtained from the cross between MS-3 (male-sterile) and 'TAM Dulce' (male-fertile) was used as mapping population. A total of 680 RAPD primers were analyzed using Bulked Segregant Analysis (BSA) approach. Out of 680 RAPD primers, two markers were found to be linked to *ms-3* gene and among them OAM08.650 was closest to *ms-3* gene at 2.1 cM. The forward and reverse sequences of this RAPD fragment were used to develop 20-mer primer pair for the development of SCAR marker, SOAM08.644. They confirmed the linkage of these linked RAPD and SCAR markers on an F_2 population from a different cross (MS-3 × 'Mission') and found them consistently linked to the *ms-3* gene at 5.2 cM. The reason for the difference in linkage distances in the two F_2 populations may be because of marker sampling variation or chromosomal differences and varying recombination frequencies. The presence of these markers was also validated in 22 heterozygous fertile F_1 plants.

The presence of another male-sterile gene *ms-5* was reported by Lecouviour *et al* (1999), and this gene was mapped by Sheng *et al* (2017) using SNP markers. They used SLAF-seq (Specific Length Amplified Fragment Sequencing) approach for sequencing the whole genome of parental lines, MS-5 (male-sterile) and HM1-1 (male-fertile), and bulks to

obtain large number of SNPs. The sequenced reads were then aligned with the reference genome and based on the sequence depth, 3,870 markers were selected for association analysis. Four regions in melon's genome were regarded as association regions. Out of these four regions mapped on chromosome 09 in the melon's genome, a region spanning 4,304,282 to 5,036,784 was selected to further fine the map using CAPS markers. Two markers were found to be linked to *ms-5* at 0.2 cM (BSA16) and 0.1 cM (BSA3-3). The results from linkage map analysis mapped *ms-5* to a 32-kb region located on scaffold 000048 on chromosome nine (from 2,522,791 to 2,555,104) of the muskmelon line DHL92 (reference genome). Among the 32-kb region, fourteen genes were screened, but only six genes were annotated. The expression of these six genes was analyzed at two different bud stages using RT-PCR analysis. The results of qRT-PCR showed higher expression of all the candidate genes in MS-5 plants than in HM1-1 plants, with the exception that *C. melo* transcription factor ABORTED MICROSPORES (AMS) exhibited a higher expression level in HM1-1 compared with MS-5 plants. The genetic variation in the expression of AMS between MS-5 and HM1-1 was analyzed by aligning the DNA sequences with that of DHL92 reference genome and an 8-bp insertion in the promoter region of *ms-5* was obtained. Hence, it was concluded that AMS is the possible candidate gene for *ms-5*. Dai *et al* (2019) also confirmed the differential expression of AMS gene in the buds (1mm and 2mm diameter) of *ms-5* male-sterile and male-fertile plants.

Recently, Singh *et al* (2020) also mapped another nuclear male-sterile gene, *ms-1* using SSR (Simple Sequence Repeat) markers. They made a cross of 'MS-1 × KP_4HM-15' and developed an F_2 mapping population which was then used to develop two contrasting bulks for male-sterile and male-fertile. A total of 506 SSR markers were first genotyped on parents to find polymorphic markers. The resulting 104 parental polymorphic markers were then used for Bulked Segregant Analysis (BSA) and it was found that only seven markers differentiated the sterile and fertile bulks. These seven polymorphic markers were then genotyped on individual plant constituting the sterile and fertile bulks, discarding the two SSR markers which were not in conformity with the phenotypic segregation data. The remaining 5 SSR markers were then genotyped on the mapping population comprising of 150 individuals. The linkage analysis conducted using the marker genotypic data and phenotypic data found three markers (DM0187, DM0038 and TJ14) linked to *ms-1* gene on chromosome six covering a total genetic distance of 35.6 cM. The marker DM0187 was closest to the *ms-1* gene at a genetic distance of 4.8 cM.

2.8 Mapping of genes in muskmelon using BSA-seq

The *MePhyto* gene conferring resistance to phytophthora blight in muskmelon line ZQK9 was mapped by Wang *et al* (2020) using BSA-seq approach combined with chromosome walking. An F_2 mapping population comprising of 498 individuals was developed by crossing the susceptible line E31 (female parent) to the resistant line ZQK9 (male parent). Two parental pools and two F_2 pools with extreme phenotypes were constructed namely RP (ZQK9), SP (E31), F2R (resistant), and F2S (susceptible) and sequenced using Illumina Nova-seq sequencing platform. The sequenced reads were aligned to reference melon genome, DHL92 version 3.5.1 using BWA, resulting in ~ 80% coverage with at least 10x depth, followed by SNP and Indel mining using GATK software. A total of 2.5 million SNPs were identified between the pools. The SNP index and ΔSNP index values for each SNP were calculated and plotted against the genomic positions. The variants were filtered before plotting based on depth (>5x), SNP-index (≥ .3), SNP depth (>7) and SNP genotype (homozygous i.e., 0/0 or 1/1). The association analysis using SNP index method mapped the *MePhyto* gene to a genomic region from 22.0 Mb to 25.3 Mb on chromosome 12. A total of 5,545 indels were detected in this region and annotated using snEff software. 2027 Indel primers were designed to precisely map the locus containing MePhyto gene. A total of twenty markers were identified to be associated with the resistant gene in ZQK9 and the linkage analysis mapped the *MePhyto* gene to a 0.6 cM genetic region between indel 63 and indel 82 corresponding to the physical location of 22.8-22.9 Mb on chromosome 12. The 52.44 kb genomic interval carried 8 genes out of which wall-associated receptor kinase (WAK) was considered to contribute to phytophthora blight resistance in melon. The qRT PCR analysis confirmed the differential expression of WAK.

Another QTL controlling the stigma color in *C. melo* was mapped by Qiao *et al* (2020) using an F_2 mapping population derived from the cross between MR-1 (green stigma) and M1-32 (yellow stigma). Two contrasting DNA bulks with highest and lowest chlorophyll content were prepared by pooling 25 F_2 plants each with green and yellow stigmas respectively. The pooled samples were sequenced and the clean reads were aligned to the melon reference genome using BWA. The Unified Genotyper Module (GATK) was used to detect the SNP and indels which were filtered using variant filtration and annotated with ANNOVAR. The genomic regions associated with the stigma color were analyzed using SNP-index method. Significant peaks were observed on chromosome 6 and 8 suggesting the presence of the genes controlling the stigma color on these chromosomes. The SNPs distributed uniformly across these chromosomes, having polymorphic restriction sites between two parents, were identified to develop 106 CAPs markers using SNP2CAPS and

analyzed on 150 F_2 individuals. The genotypic data from 32 primer pairs were used to construct the linkage map using QTL Icimapping software. A genetic map of 304.17 cM consisting of two linkage groups named chromosome 6 (165.40cM) and 8 (138.77cM) were constructed containing 16 CAPs markers each with an average distance of 10.33 cM and 8.67 cM, respectively. The major qtl(GS8.1) was detected on chromosome 6 located between marker H-90 and A-6 covering a genetic distance of 37.62 cM. The genetic locus harboring major QTL GS8.1 was further narrowed down by developing new polymorphic CAPS marker to a 4.13 cM which corresponds to a 268 kb interval on the physical map. This region consists of 27 putative genes out which 8 genes carried non-synonymous SNPs. The Genome annotation identified 3 possible genes *MELO3C003149, MELO3C003158,* and *MELO3C003165.*

CHAPTER III

MATERIALS AND METHODS

The present study entitled, "Fine mapping of male-sterility *ms-1* gene using whole genome resequencing of muskmelon (*Cucumis melo* L.)" was conducted at the School of Agricultural Biotechnology and the Department of Vegetable Sciences, Punjab Agricultural University (PAU), Ludhiana, during 2018-2021. PAU is located at 30.9019° N latitude and 75.8078° E longitude.

The experimental materials used and the methods adopted to carry out this study are discussed in this chapter. The investigation was divided primarily into two experiments,

3.1 Whole genome resequencing and identification of SNPs

3.2 Fine mapping of genetic male-sterile *ms-1* gene

3.1 Whole genome resequencing and identification of SNPs

3.1.1 Location of work

School of Agricultural Biotechnology, PAU, Ludhiana and Vegetable Research Farm, Department of Vegetable Science, PAU, Ludhiana

3.1.2 Experimental methodology

3.1.2.1 Plant material

In the present investigation, the male-sterile line, MS-1, and the male-fertile line, KP_4HM-15 were crossed (MS-1 × KP_4HM-15) to develop an F_2 mapping population. MS-1 is a male-sterile inbred line, introduced in India in 1978 and since then it is being maintained and refined at Punjab Agricultural University, Ludhiana. MS-1 carries a recessive genic male-sterile (GMS) gene, *ms-1* which results in the production of indehiscent anthers (Bohn and Whitaker 1949). On the other hand, KP_4HM-15 is a fully fertile inbred line with promising fruit characteristics and fusarium wilt resistance. KP_4HM-15 was developed by transferring the fusarium wilt resistance genes from a local snapmelon line KP_4 or Kariam Phut (*Cucumis melo var momordica*) into cv. Hara Madhu (*Cucumis melo var reticulatus*) (Vashisht *et al* 2019). In this study, a cross between male-sterile line, MS-1 and male-fertile line, KP_4HM-15 was used to generate an F_2 mapping population.

3.1.2.2 Generation of F_2 mapping population

The F_1 plants, obtained from the cross 'MS-1 × KP_4HM-15', were grown in the farm area of Department of Vegetable Science. These F_1 plants were self-pollinated to obtain the seeds for the next generation during February 2019. The seeds from the F_1 fruits were directly sown in soil for developing an F_2 mapping population. The F_2 population along with the

parents was raised in off-season during September 2019 to January 2020 in green-house at School of Agricultural Biotechnology, PAU. The crop was raised by following the standard agronomical practices recommended by PAU, Ludhiana.

The male-fertile F_2 plants were further self-pollinated to generate $F_{2:3}$ progenies in order to distinguish between segregating (*Msms*) and non-segregating (*MsMs*) F_2 male-fertile types. The flow-chart describing the development of F_2 mapping population is shown in Fig. 3.1.

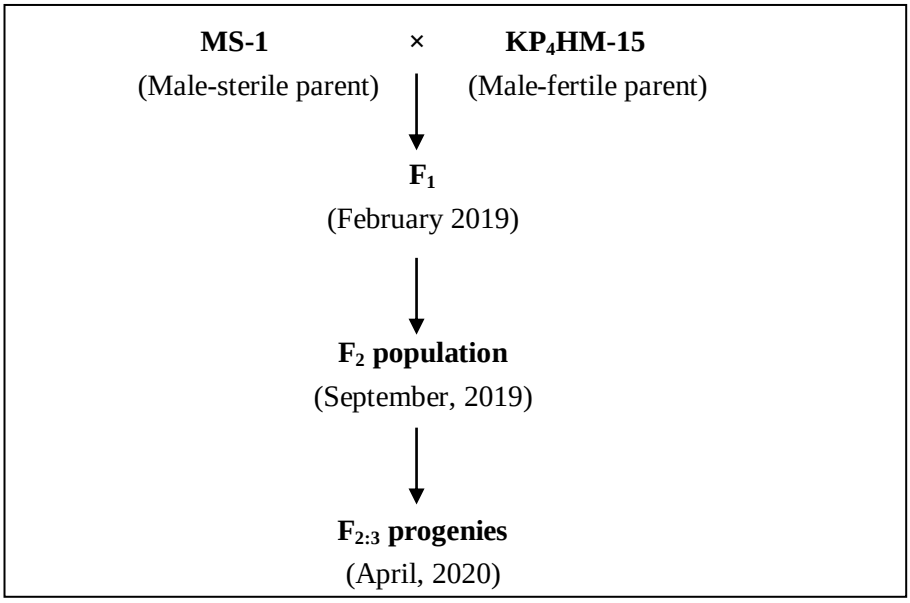


Fig. 3.1: Flow-chart for the development of F_2 mapping population and their F_3 progenies

3.1.2.3 Phenotyping of F_2 population

The phenotype for *ms-1* i.e., male-sterility or fertility, was assessed by examining the staminate flowers for the absence or presence of pollens at the time of anthesis, between 6:00-9:00 am. The absence/presence of pollens as yellow dust was determined by touching or rubbing the anthers against a black surface. A plant was tagged as a male-fertile (*MsMs/Msms*) as soon as a single male-fertile flower was identified, whereas a plant was tagged as a male-sterile (*msms*) only after scoring at least 8-10 flowers per plant as suggested by Dhillon and Kumar (2008). It has been reported that the phenotype of *ms-1* male-fertile plants is unstable and requires the screening of at least 8-10 flowers before tagging a plant as male-sterile (McCreight 1980 and Dhillon & Kumar 2008).

To distinguish homozygous (*MsMs*) F_2 male-fertile plants from heterozygous (*Msms*) ones, F_2 population was advanced to raise $F_{2:3}$ progenies. These progenies were sown in the fields during April 2020. The F_2 plants with segregating progenies for *ms-1* gene were tagged as heterozygous male-fertiles (*Msms*) and the plants with 100 percent male-fertile progenies were tagged as homozygous male-fertiles (*MsMs*).

3.1.2.4 Genomic DNA isolation

Genomic DNA from young leaf tissues of 152 F_2 individuals along with both the parents, MS-1 and KP_4HM15 was isolated in October 2019. Cetyltrimethylammonium bromide (CTAB) method for rapid DNA extraction, given by Doyle and Doyle (1987), was used and the protocol followed is discussed below.

a) Sample preparation

- Young leaf samples (70-100mg) from individual plants were collected in separate butter-paper bags and stored at -80˚C till further use.
- Tissue samples were snap-freezed by placing them into liquid nitrogen filled mortars to make the samples brittle.
- Snap-freezed tissue samples were then crushed to fine powder using pestle-mortar and were transferred to properly labeled 2ml polypropylene tubes.

b) CTAB extraction

- After grinding, 800µl of pre-heated (65°C) 2X CTAB extraction buffer was added to the crushed samples. The composition of the extraction buffer is given in Table 3.1.
- The powdered sample was suspended in the extraction buffer by gently inverting the tubes.
- The tubes were incubated in the water bath for 45 minutes at 65°C with intermittent shaking.

c) Solvent extraction

- 800µl of chloroform: isoamyl alcohol (24:1) was added to each sample and tubes were placed on a shaker for 45 min for the uniform and constant mixing of the contents.
- Tubes were centrifuged at 10,000 rpm for 15 min.
- The clear supernatant was transferred into the fresh 1.5ml polypropylene tubes.

d) DNA precipitation

- 600µl of chilled isopropanol was added to the supernatant followed by gentle inversions until the DNA precipitated and stored at -20°C for 20 min.
- Tubes were centrifuged at 10,000 rpm for 15 min to pellet out the DNA.
- Supernatant was discarded from each sample leaving behind the DNA pellet.

e) Re-suspension of DNA

- DNA pellet was rinsed with ethanol by adding 200µl ethanol (70%) followed by centrifugation at 10,000 rpm for 10 min.

- Supernatant was decanted from each sample and again washed with 100µl ethanol (70%).
- Samples were air dried by inverting the tubes on a clean filter paper after discarding the supernatant.
- DNA pellet was re-suspended in 100µl of 1X TE and stored at -20°C.

f) RNase treatment

- 5-7µl of RNase solution (10mg/ml) was added to a final concentration of 10µg/ml to each sample.
- The samples were then incubated at 37°C in water bath for 30 minutes.
- RNase solution was prepared by mixing 10mg of RNase powder in 1ml of RNase buffer.
- The composition of RNase buffer for preparing 50ml and 100ml of RNase solution is given in Table 3.2

Table 3.1: Composition of 2X CTAB extraction buffer

Components	Final Concentration
CTAB	2.0%
1M Tris (pH 8.0)	100mM
NaCl	1.4 M
0.5M EDTA	20mM
β-Mercaptoethanol*	2.0%
ddH_2O	-

*β-Mercaptoethanol was added to CTAB just before the use

Table 3.2: Composition of RNase buffer for preparing 100ml RNase solution

Components	100ml
Tris-HCl	0.150g
NaCl	8.700g

3.1.2.5 Quality and quantity assessment of isolated DNA

The quality and quantity of the isolated DNA was estimated by agarose gel electrophoresis and also by UV-Spectrophotometer (Thermo Scientific Nanodrop^{TM}8000 Spectrophotometer, USA).

a) Agarose Gel Electrophoresis

To check the quality and quantity of DNA, 0.8% of agarose gel was used and viewed under UV trans-illuminator.

1. 0.8% agarose gel was prepared by dissolving 0.8g of agarose in 100ml of 0.5X TBE gel electrophoresis buffer (45mM Tris base, 45mM Boric Acid and 1mM EDTA) and heated till agarose dissolved completely *i.e.* till the solution became transparent.
2. 5µl of ethidium bromide (10 mg/ml) was added to the gel when the temperature cooled down to 55-60°C, to make the final concentration of 0.5 µg/ml buffer.
3. The gel solution was then transferred to the already prepared gel mould containing combs and was then left to solidify for 30-40 min.
4. The DNA samples were prepared for loading by adding Bromophenol blue (6X loading dye containing 0.25% w/v bromophenol blue, 0.4% w/v sucrose or 30% Glycerol in sterile water) to DNA such that the final concentration of the loading dye was 1X.
5. After the solidification of the agarose gel, the gel along with the mould was transferred to the electrophoresis tank and combs were removed to form the wells for DNA loading.
6. The prepared DNA samples were loaded into the wells using micropipette. Along with the DNA samples, DNA of known concentrations (100ng/µl, 200ng/µl, 400ng/µl, 500ng/µl and 1000ng/µl) was also added.
7. After loading the DNA samples, the gel was run for 1-2 hr at a constant voltage of 5V/cm.
8. The gel was then visualized under UV trans-illuminator and photographed using VisionWorksLS Analysis software.
9. The intensity of fluorescence of each sample was compared with that of known DNA concentrations and then DNA concentration of each sample was ascertained.
10. The quality of DNA samples was judged based on whether the DNA formed the single high molecular weight band (good quality) or a smear (degraded or poor quality).

b) DNA quantification using Spectrophotometer

DNA quality and quantity was checked using Thermo Scientific Nanodrop™8000 Spectrophotometer, USA following the procedure as under:

1. A blank measurement was made with 1µl of 1X TE.
2. DNA quantity was estimated by pipetting out 1µl of each sample onto the end of a fiber optic cable (the receiving fiber). Quality was determined in terms of 260/280 absorbance ratio and the data was logged in an archive file on the PC.
3. On the basis of quantification, bulks were prepared so that the final concentration of the bulked DNA was >50ng/µl.

3.1.2.6 Preparation of contrasting bulks

Two contrasting bulks were developed by selecting homozygous male-sterile (*msms*) and homozygous male-fertile (*MsMs*) individuals from the F_2 population of 152 plants. Genotypic selection in F_2 was done after confirmation in $F_{2:3}$ progenies. Ten DNA samples from plants each of *msms* genotype and *MsMs* genotype were selected to prepare male-sterile bulk (MSB) and male-fertile bulk (MFB), respectively. Each bulk was constructed by mixing equal concentrations of DNA from individual plant sample to make a final DNA concentration of more than 50ng/µl.

3.1.2.7 Whole genome resequencing

The genomic DNA of parents (MS-1, KP_4HM-15) and male-sterile bulk (MSB) and male-fertile bulk (MFB) were out-sourced for library preparation and whole genome re-sequencing to NGB Diagnostics Pvt. Ltd. (New Delhi, India). 1-2µg of DNA from each sample was first sheared using NEBNext® dsDNA Fragmentase, followed by construction of genomics libraries using NEBNext® Ultra™ DNA Library Prep Kit. The QC of the final library was analysed on the Bioanalyzer. Later, the sequencing of these libraries was carried out on Illumina HiSEQ 4000 to generate 151 bases long, paired end (PE) reads with Q30>80%.

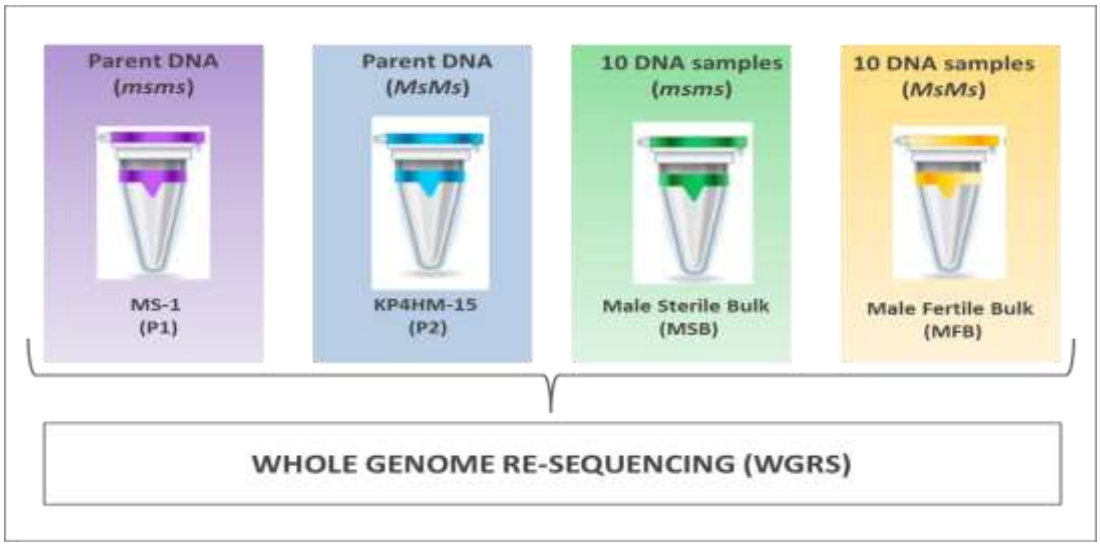


Fig. 3.2: Samples used for whole genome resequencing

3.1.2 Observations recorded

- Number of male-fertile (*MsMs/Msms*) plants in F_2 population.
- Number of male-sterile (*msms*) plants in F_2 population.
- Number of homozygous (*MsMs*) and heterozygous (*Msms*) male-fertile plants in F_2 population.

- Total amount of data and coverage obtained after resequencing of MS-1 (male-sterile parent), KP4HM-15 (male-fertile parent), MSB (male-sterile bulk) and MFB (male-fertile bulk).

3.1.3 Segregation Analysis

To validate the recessive inheritance of *ms-1* gene, goodness of fit of the F_2 population as per the expected phenotypic ratio of 3:1 was worked out using the chi-square analysis. The chi-square value was calculated as follows:

$$\chi^2_{(n-1)\,df} = \sum \{(O-E)^2/E\}$$

where, O = Observed number of plants

E = Expected number of plants

n = number of classes

df = Degree of freedom

3.1.4 Calculation of genomic coverage

Sequencing coverage is the number of reads that align to (or cover) the reference genome. It describes how often a reference sequence is covered by bases form the reads. It was calculated by the following formula,

$$C = LN/G$$

where, C = coverage

L = read length

N = number of reads

G = haploid genome length

3.2 Fine mapping of genetic male-sterile *ms-1* gene

3.2.1 Location of work

Bioinformatics Laboratory, School of Agricultural Biotechnology, PAU, Ludhiana.

3.2.2 Experimental methodology

3.2.2.1 Experimental material

To fine map the genetic male-sterile gene *ms-1*, the sequenced reads from male-sterile parent (MS-1), male-sterile bulk (MSB) and male-fertile bulk (MFB) were analyzed *in-silico* using bioinformatics softwares.

3.2.2.2 Quality check (QC) of the sequenced reads

During sequencing processes, certain bases are called incorrectly because of the instrument's internal technical limitations. Depending upon the type of sequencing platform, different types and amount of errors are generated which can greatly impact the downstream

analysis steps. Therefore, it is necessary to identify and filter these low quality reads, and generate a file with clean reads for further analysis. The raw sequence files obtained after sequencing are in fastq format, which contains nucleotide sequences along with their corresponding Phred quality scores or Q-scores. A Q-score of a base indicates the probability of that base being called incorrectly by the sequencer. Therefore, Q-score is used as a measure to determine the accuracy of the sequencer. Mathematically, Q-score is logarithmically related to the base calling error probabilities (P),

$$Q = -10 \log_{10} P$$

The Phred quality score of 30 or Q30 indicates the probability of a base called incorrectly in 1000 calls, which is equivalent to 99.9% accuracy of a base being called correctly.

Table 3.3: Quality scores and base calling accuracy (Table reconstructed from https://www.illumina.com/documents/products/technotes/technote_Q-Scores.pdf)

Phred Quality Score	Probability of Incorrect Base Call	Base Call Accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%

The raw sequence files obtained for the parents and the bulks were subjected to quality control (QC) checks using FastQC (Andrews 2010). This software offers a set of analysis that not only identifies errors originating in the sequencer but also those arising during the library preparation. After downloading and installation of FastQC software, raw sequence files (fastq format) were uploaded as input files and the QC report was generated and viewed on the main interactive display. This QC report included the summary and details of the analysis modules that were run. Per base sequence quality (shows quality value range of each base), per sequence quality score (plots quality values of a subset of sequences), per sequence GC content (measures GC content and compares it to the modeled normal distribution of GC content), adapter content (detects the presence of read-through adapter sequences) are some of the analysis modules among others, which FastQC software runs to analyze the quality of raw sequences. The QC reports so obtained were saved as an HTML file for future reference.

3.2.2.3 Read trimming and adapter clipping

The bad quality reads and undesirable sequences present within the raw sequence files, poses a problem while performing the downstream analysis steps. As a consequence, these sequences need to be removed to obtain results with precision and reliability. These undesirable sequences in the sequence files of MS-1, KP_4HM-15, MSB and MFB were removed using Trimmomatic version 0.39 (Bolger *et al* 2014) software. Trimmomatic is a command line tool used to trim bad quality reads and clip away the undesirable sequences like adapter sequences. It supports the trimming of both single-end as well as paired-end reads and works with Illumina fastq files having either +33 or +64 Phred quality scores. Trimmomatic offers a variety of useful trimming tasks which are provided in a particular order in the command line. The trimming steps used to remove the unwanted reads in raw sequence files of MS-1, KP_4HM-15, MSB and MFB along with the parameters (within parenthesis) used were as follows,

1. ILLUMINACLIP: Clips illumina-specific sequences (adapters). (ILLUMINACLIP: TruSeq3-PE.fa:2:30:10)
2. LEADING: Chops the bases at the start of the read if the Q score is less than threshold. (LEADING:3)
3. TRAILING: Chops the bases at the end of the reads if the Q score is less than threshold. (TRAILING:3)
4. HEADCROP: Removes a specified number of bases at the start of the read. (HEADCROP:10)
5. SLIDINGWINDOW: Trims the reads with average quality score below a threshold within a specified sliding window. (SLIDINGWINDOW:4:15)
6. MINLEN: Discards the read if the read is below a specified length. (MINLEN:36)

Files containing forward and reverse sequences were used as input files and four output files were specified in the command line. The four output files consist of paired and unpaired files for both forward and reverse input files. The ‘paired’ output file contains both forward and reverse reads which survived processing while the ‘unpaired’ output file included only those reads where the partner did not survived the processing. The paired files obtained after trimming were again checked for quality and then used for downstream analysis.

3.2.2.4 Development of reference-guided MS-1 genome assembly

The reference genome of muskmelon, DHL92 v3.6.1 (Ruggieri V *et al* 2018) was used to develop reference-guided genome assembly for one of the parent. The male-sterile

parent MS-1, carrying the *ms-1* gene, was selected for developing the MS-1 assembly. The MS-1 assembly was developed by,

1. Aligning the clean reads of MS-1 to the reference genome DHL92 v3.6.1 using Bowtie2 v2.4.2 (Langmead & Salzberg 2012), in order to determine the reads' point of origin with respect to the reference.
2. The aligned reads were then used to call the variants between MS-1 and the reference genome using SAMtools (Li *et al* 2009) & BCFtools (Li 2011).
3. The variants so obtained were then used to develop the MS-1 assembly by replacing the detected SNPs in the MS-1 with those from the DHL92 reference genome.

The MS-1 assembly thus generated was used as a reference to call the variants in the bulks, MSB and MFB and calculate their corresponding SNP-index values.

3.2.2.5 Narrowing down the genomic locus carrying *ms-1* gene using SNP-index method

To narrow down the genomic locus harbouring *ms-1* gene, QTL-seq pipeline (Sugihara *et al* 2020) was used. This pipeline is based on the calculation of SNP-index values for the identification of both qualitative and quantitative trait loci (Abe *et al* 2012 and Takagi *et al* 2013). QTL-seq pipieline is a python-based script developed at Iwate Biotechnology Research Centre, Japan and is available at https://github.com/YuSugihara/QTL-seq. QTL-seq utilizes the sequences from two extreme bulks (e.g. male-sterile bulk and male-fertile bulk) and a whole genome resequencing of either of the two parental cultivars.

In this study, the clean reads of male-sterile bulk (MSB) and male-fertile bulk (MFB), along with those of male-sterile parent, MS-1 were used. The MS-1 genome assembly developed previously was used as a reference with respect to which the variants were identified in the MSB, MFB and MS-1 parent. The steps in the Qtl-seq pipeline to fine map the *ms-1* locus were as follows,

1. The reads form three files (MSB, MFB and MS-1) were first aligned to the MS-1 assembly using the in-built Burrows-Wheeler Alignertool (Li & Durbin, 2009).
2. The reads after alignment were then piped to SAMtools for filtering the improperly aligned reads.
3. The properly aligned reads were used to call the variants using BCFtools.
4. The VCF file thus generated was then filtered based on mapping quality and strand bias.
5. The filtered SNPs were then used to calculate SNP-index and ΔSNP-index values. For each SNP, within the male-sterile bulk (MSB) and male-fertile bulk (MFB), the SNP-index values were calculated according to the method given by Abe *et al* 2012.

6. Subsequently, the ΔSNP-index values for each SNP were calculated by subtracting the SNP-index values of MSB from the SNP-index values of MFB (SNP-index MSB - SNP-index MFB).
7. Under the hypothesis of no QTL, the expected ΔSNP-index values forming the confidence-intervals were obtained by performing computer simulations at 99% and 95% level of significance (Sugihara *et al* 2020).
8. To identify the key genomic positions associated with the *ms-1* trait, genome-wide SNP-index values for both MSB and MFB along with ΔSNP-index values were analyzed using a sliding window of 1 Mb and a step-size of 100-kb.
9. The standalone program QTL-plot, available within the QTL-seq pipeline, calculated and simulated the expected ΔSNP-index values and also generated scatter-plots for SNP-index and ΔSNP-index values with respect to their genomic positions.

All these steps, from alignment of the reads to the generation of the scatter-plots, are connected via a python script within the Qtl-seq pipeline. The default values for various parameters within the pipeline were used for each step, which are: sliding window size (1Mb), step size (100-kb), minimum depth of variants (8), maximum depth of variants (250), minimum mapping quality (40), and minimum base quality (18).

3.2.2.6 Functional annotation

The candidate region delimited after the Qtl-seq implementation was further screened for the presence of candidate gene(s). The function of all the genes within the candidate genomic interval was assessed using Blast2GO or B2G software (Götz *et al* 2008). Blast2GO software functionally annotates the DNA or protein sequences based on the Gene Ontology vocabulary and is available at https://www.biobam.com/download-omicsbox/. Blast2GO functionally annotates the genes in 4 steps, which include: BLAST searching the query sequences, mapping or fetching GO terms for the obtained BLAST hits, annotation by assigning the appropriate GO term and an additional annotation for domain/motif information using InterPro Scan.

To functionally annotate the genes within the mapped region, the corresponding protein fasta sequences were extracted from CM4.0_protein.fasta file available at https://www.melonomics.net/melonomics.html#/download. These sequences were then uploaded into Blast2GO and searched against the NCBI database to find similar sequences using Basic Alignment Search Tool (BLAST). The homologue sequences obtained after BLAST were then mapped to retrieve the GO terms and appropriate enzyme codes were

obtained. The final GO term from the pool of GO terms, obtained during mapping, was assigned by applying the Blast2GO annotation algorithm. The InterPro domains and motifs were directly queried at the InterProScan web service. Blast2GO also allows the visualization of the combined annotation of all the sequences through the generation of charts and graphs. The multilevel pie-charts based on the GO terms (molecular function, biological process and cellular component) were generated to visualize the annotations. These annotated genes were then searched against the available literature for their role in the male-sterility.

3.2.2.7 Variant annotation

The variants within the candidate region were also analyzed to predict their effect on the annotated genes using snpEff software (Cingolani *et al* 2012). snpEff is a variant annotation and effect predicting tool available at https://pcingola.github.io/SnpEff/download/. This software annotates and predicts the effects of variants based on their genomic location (such as intronic, 5' UTR, 3' UTR, upstream, downstream or intergenic regions) and the resultant change in amino acid (such as synonymous or non-synonymous SNPs, gain or loss of start codon, gain or loss of stop codon).

The genome-wide variants were first annotated using snpEff, and only those variants were selected which were used for SNP-index calculations. Before providing the filtered .vcf file to snpEff, Muskmelon database was also build using 'snpEff build' option, as muskmelon's database was not provided within the software. After the genome-wide annotation of the variants, the positions of highly-significant SNPs and variants present within the candidate gene(s) were extracted and viewed using the IGV genome browser.

3.2.3 Observations recorded

- Quality of the sequenced reads before and after trimming.
- Number of reads that survived and dropped after trimming.
- The genomic region associated with the *ms-1* trait.
- Candidate gene(s) within the narrowed-down genomic region

CHAPTER IV

RESULTS AND DISCUSSION

The present investigation entitled "Fine mapping of male-sterility *ms-1* gene using whole genome resequencing of muskmelon (*Cucumis melo* L.)" involved the generation of F_2 mapping population. Two contrasting parents for the *ms-1* trait namely, MS-1 inbred line (male-sterile) and KP_4HM-15 inbred line (male-fertile) were used to generate the mapping population. The anthers in MS-1 were small and indehiscent (Fig. 4.1a) while the anthers of KP_4HM-15 were flooded with pollens (Fig. 4.1b). The cross was made by pollinating female flowers of MS-1 (Fig. 4.2a), with the pollens collected from male flowers of KP_4HM-15 (Fig. 4.2b). An F_2 mapping population, comprising of 152 individuals, was raised at Green-house no. 3, School of Agricultural Biotechnology (SOAB), PAU (Fig. 4.3) during the winters of 2019. Besides *ms-1*, these inbred lines are polymorphic with respect to other traits also e.g., the fruits of MS-1 have netted skin with dark green rind color and orange flesh, while KP_4HM-15 have non-netted skin with light green rind color and light green flesh (Fig. 4.4). The segregation of these traits in F_2 population is shown in Fig. 4.5.

The first experiment in this study was aimed at developing the F_2 mapping population for fine mapping the *ms-1* gene. Based on the genotype of F_2 individuals, two contrasting bulks namely, male-sterile bulk (MSB) and male-fertile bulk (MFB) were constructed by pooling the DNA from ten individuals each of homozygous male-sterile (*msms*) and homozygous male-fertile (*MsMs*) genotypes, respectively. Both the bulks along with their parents, MS-1 and KP_4HM-15 were sequenced and used in the second experiment. The second experiment was designed to analyze these raw sequenced reads *in silico* using bioinformatic tools to fine map the *ms-1* gene. The narrowed down genomic locus was further analyzed for the candidate gene(s) underlying the *ms-1* trait. The results obtained during the present investigation are discussed in this chapter.

4.1 Whole genome resequencing and identification of SNPs

4.1.1 Assessment of mapping population for male-sterility/fertility

The male flowers of F_2 and F_3 individuals were analyzed visually to ascertain the genotype of the F_2 plants. F_2 plants were sown in Green-house of School of Agricultural Biotechnology, which started flowering in October 2019. The male-sterile and male-fertile plants were distinguished by rubbing the anthers from the staminate flowers on the black surface and visually looking for the absence/presence of pollens (Fig. 4.6). The plants with flowers shedding yellow dust on the black surface were tagged as male-fertile plants and the plants with at least 8-10 flowers showing no yellow dust were tagged as male-sterile plants. The flowers of the male-fertile (*MsMs/Msms*) plants had anther lobes flooded with dense mass of pollen grains while there was hardly any pollen visible on the anthers of male-sterile (*msms*) flowers.

Fig. 4.1: Anthers as observed under light microscope (a) indehiscent anthers observed in MS-1 inbred line and (b) anthers of KP_4HM-15 covered with visible pollen grains

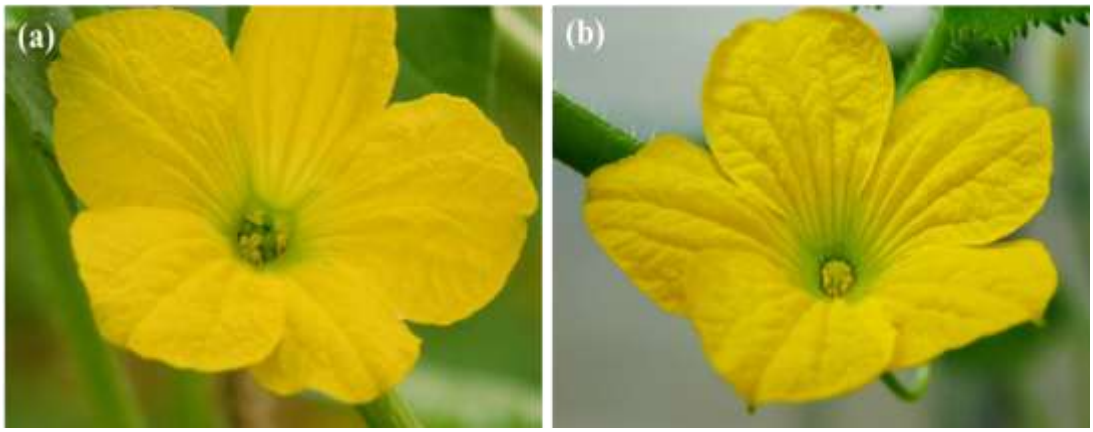


Fig. 4.2: (a) Female flowers of MS-1 with sterile anthers, (b) Male flowers of KP_4HM-15 with fertile anthers. The pollens from the male flowers of KP_4HM-15 were used to pollinate the female flowers of MS-1

Fig. 4.3: F_2 plants raised from MS-1 X KP_4HM15 in green house

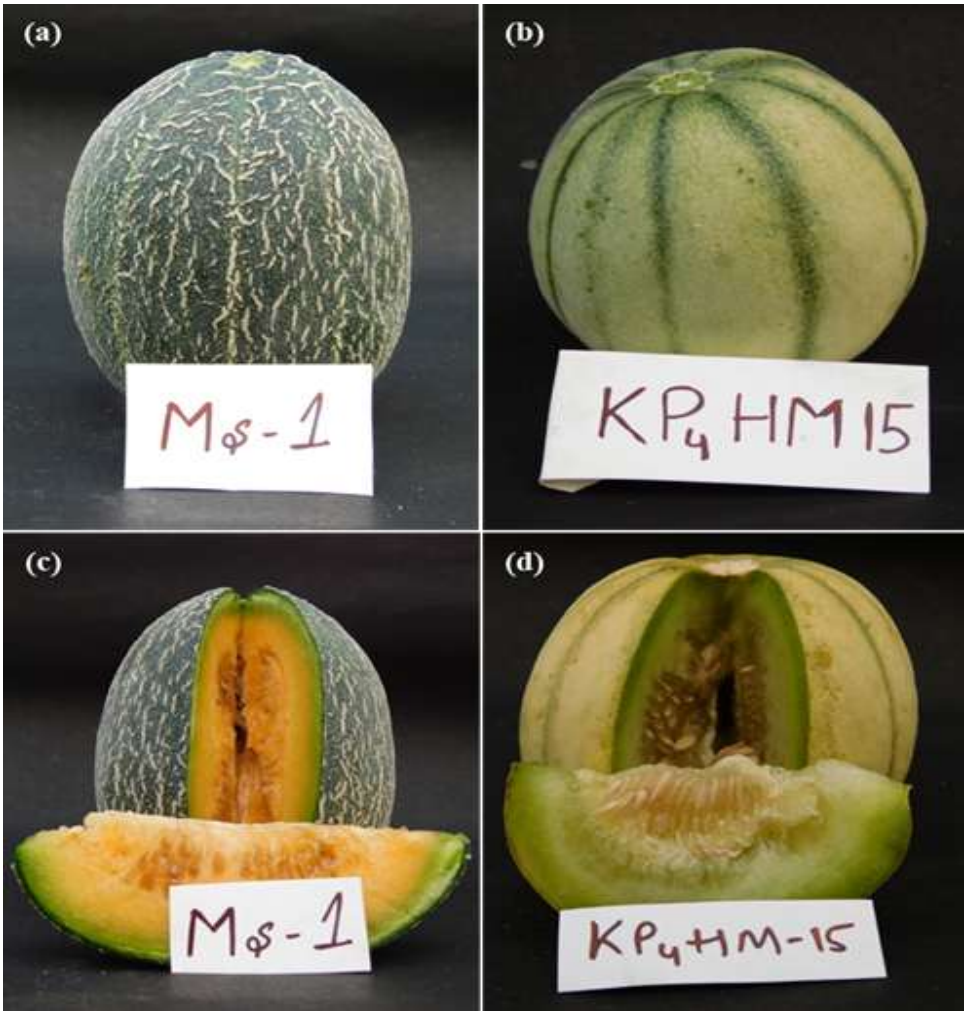


Fig. 4.4: Polymorphism between MS-1 and KP_4HM-15 for fruit traits. (a) MS-1 fruit with netted skin, dark-green rind and oval shape, (b) KP_4HM-15 fruit with non-netted skin, light-green rind and round shape, (c) orange flesh of MS-1, and light-green flesh of KP_4HM-15

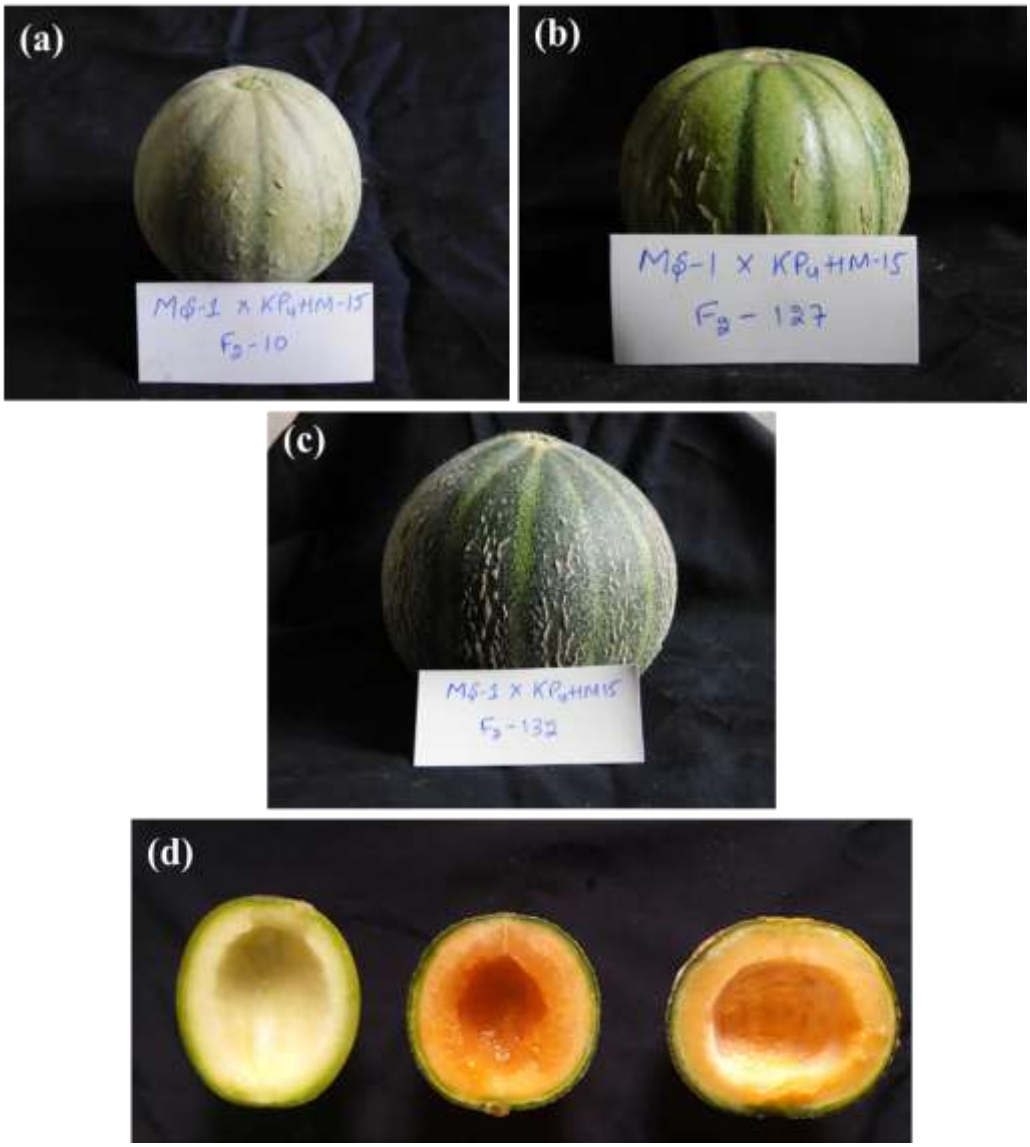


Fig. 4.5: Polymorphism among the fruits of F_2 segregates with respect to external fruit characters like netting, rind color and shape (a), (b), (c), and flesh color (d)

Fig. 4.6: (a) Fertile anthers flooded with pollens and (b) sterile anthers devoid of pollens

4.1.2 Segregation analysis of *ms-1* gene

The F_2 mapping population segregated for *ms-1* gene into: (i) dominant male-fertile (*MsMs*), (ii) heterozygous male-fertile (*Msms*), and (iii) recessive male-sterile (*msms*) genotypes. A total of 152 F_2 individuals were screened for the segregation of male-sterility, out of which 112 plants were identified as male-fertile (*MsMs/Msms*) and 40 plants were male-sterile (*msms*). The chi-square analysis was conducted to test the inheritance of single recessive gene, '*ms-1*' in F_2 population. The 'goodness of fit' for the observed phenotypic data to the expected phenotypic ratio of 3:1 was tested.

Table 4.1: Phenotypic segregation pattern of F_2 population

Cross	Fertility Status	Observed	Expected	$(O-E)^2/E$	$\sum\{(O-E)^2/E\}$
MS-1× KP_4HM-15	Male-fertile	112	114	0.035	0.14
	Male-sterile	40	38	0.105	

The table value of chi-square at 5% level of significance and 1 degree of freedom is,

$$\chi^2_{1df,\ 0.05\alpha} = 3.841$$

The calculated chi-square value obtained was 0.14 which is found to be non-significant at 5% level of significance and 1 degree of freedom (Table 4.1). Thus, the non-significant chi-square value showed that the observed segregation of F_2 population did not deviate from the expected 3:1 ratio. The segregation analysis confirmed that male-sterility in melon is governed by a single recessive gene *ms-1*.

Out of 112 F_2 male-fertile plants (*MsMs/Msms*), 65 plants were self-pollinated to generate $F_{2:3}$ progenies. Each F_3 progeny comprised of 15-20 plants. Flowers of every plant within each progeny were carefully examined to determine the segregation of *ms-1* and thus the genotype of their respective F_2 plants. In 26 F_3 progenies, no male-sterile flower was found. The corresponding F_2 plants giving rise to these progenies were designated as homozygous dominant (*MsMs*) male-fertile. Similarly, the remaining 39 progenies which segregated into male-sterile and male-fertile plants were considered to have originated from heterozygous (*Msms*) male-fertile F_2 plants. Thus, a total of 105 F_2 plants (65 male-fertile and 40 male-sterile) were successfully genotyped. The genotype of the remaining 47 F_2 plants could not be determined either due to lack of fruit set or poor germination of the seed obtained from these individual F_2 plants. These remaining plants were either homozygous dominant or heterozygous male-fertile plants.

Based on the genotype of these 105 F_2 plants two contrasting bulks were generated by pooling the DNA from ten plants each with homozygous male-sterile genotype (*msms*) and heterozygous male-fertile genotype (*Msms*) for constructing male-sterile bulk (MSB) and male-fertile bulk (MFB), respectively.

4.1.3 Whole genome resequencing of parents and bulks

Illumina HiSEQ4000 platform was used for whole genome resequencing of four DNA samples namely, MS-1 (male-sterile parent), KP_4HM-15 (male-fertile parent), MSB (male-sterile bulk) and MFB (male-fertile bulk). The Illumina sequencing generated a total of 40 GB paired-end (PE) data in .fastq file format. The .fastq file stores nucleotide sequence data along with their corresponding Phred quality score values which are encoded in ASCII characters. The .fastq file generated for the male-sterile parent, MS-1 is shown in Fig. 4.7 as an example. The file stores the sequence information for each fragment/read sequenced in the following format,

(i) Line 1: Illumina sequence identifier beginning with @ symbol;

(ii) Line 2: Raw sequence letters;

(iii) Line 3: A separator, usually + symbol;

(iv) Line 4: Quality values for sequence in the Line 2.

The contents of the sequence identifier are explained in Table 4.2 using the first sequence in the MS-1.fastq file as an example (Fig. 4.7).

Table 4.2: The description of the sequence identifier (Line1) contents in a .fastq file

Contents of Line 1	Description
@E00537	Instrument name
307	Run Id
HCVT5CCX2	Flowcell Id
7	Flowcell lane
1101	Number within the flowcell lane
8563	‘x’-coordinate of the cluster within the tile
150112	‘y’-coordinate of the cluster within the tile
1	Member of a sequence read pair, 1 or 2 (for PE only)
N	Passed read after filtering (otherwise Y)
0	Number indicating that the control bit is on; otherwise it is an even number
NTCGAACA+NGANCTCG NTCGAACA	Index sequence

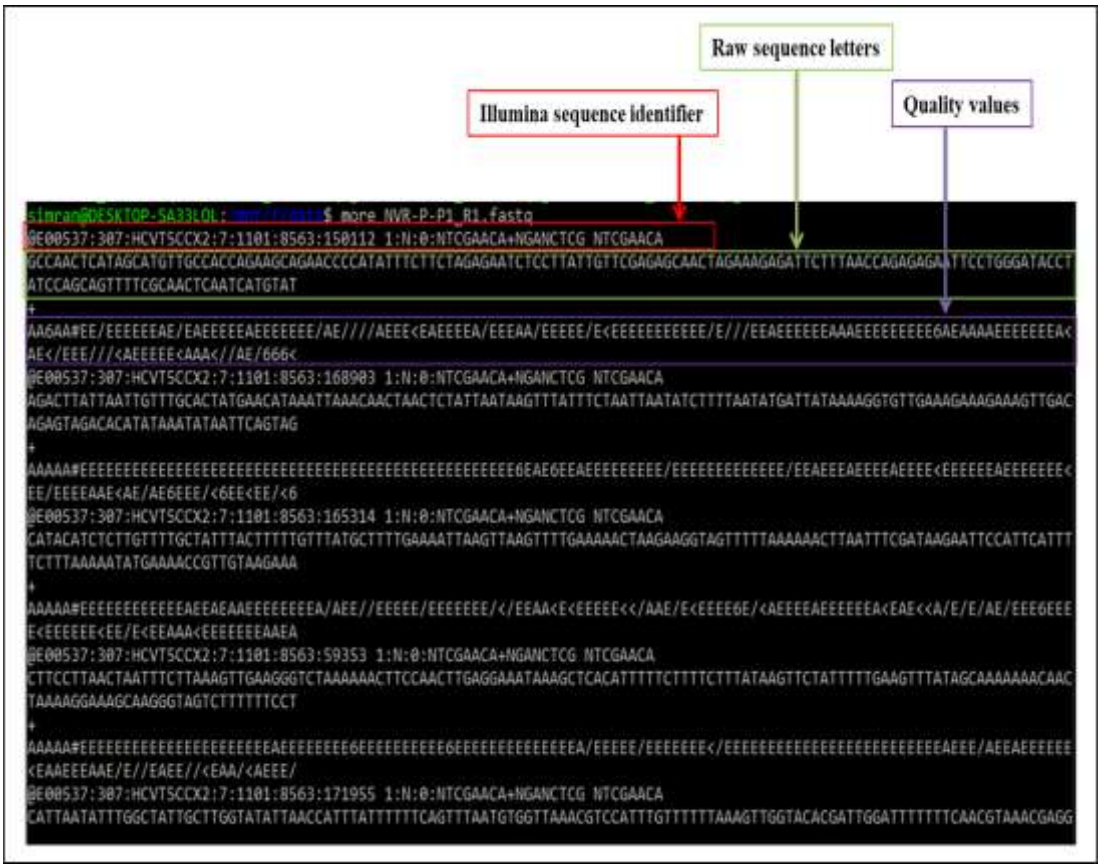


Fig. 4.7: An image displaying the contents of the MS-1.fastq file

The highest amount of sequenced data was generated for MFB comprising of 12.7 GB data, followed by MSB, KP_4HM-15 and MS-1 with 9.62, 9.37 and 9.31 GB data, respectively. The average read-length for the four samples was 151 base pairs (bp). The number of reads generated for MS-1, KP4HM-15, MSB and MFB are 53.7, 55.0, 55.5, and 72.8 million reads, respectively. The sequencing coverage calculated was 52.74x for MFB, 40.25x for MSB, 39.85x for KP4HM-15 and 38.92x for MS-1. The GC content for all samples was between 33-36%.

Table 4.3: Statistics of the data obtained after sequencing.

Genotype	Sequenced data (GB)	Sequenced reads	Sequencing coverage
MS-1	9.31	53,753,767	38.92x
KP4HM-1	9.37	55,025,107	39.85x
MSB	9.62	55,577,927	40.25x
MFB	12.7	72,827,049	52.74x

4.2 Fine mapping of genetic male-sterile *ms-1* gene

4.2.1 Quality control (QC) analysis and pre-processing of the sequenced reads

The raw reads obtained after sequencing (Illumina HiSEQ 4000) were assessed for quality and adapter contamination using FastQC v0.11.9. It was observed that the bases towards the end of the reads were of bad quality with phred quality score less than 30 (Q30), Fig. 4.8a. The proportions of Adenine (A) and Guanine (G) bases at the start of the reads (10 bases) varied greatly with respect to the Thymine (T) and Cysteine (C) contents respectively, Fig. 4.8c. The presence of Illumina universal adapter sequences was also reported in the FastQC report, Fig. 4.8e.

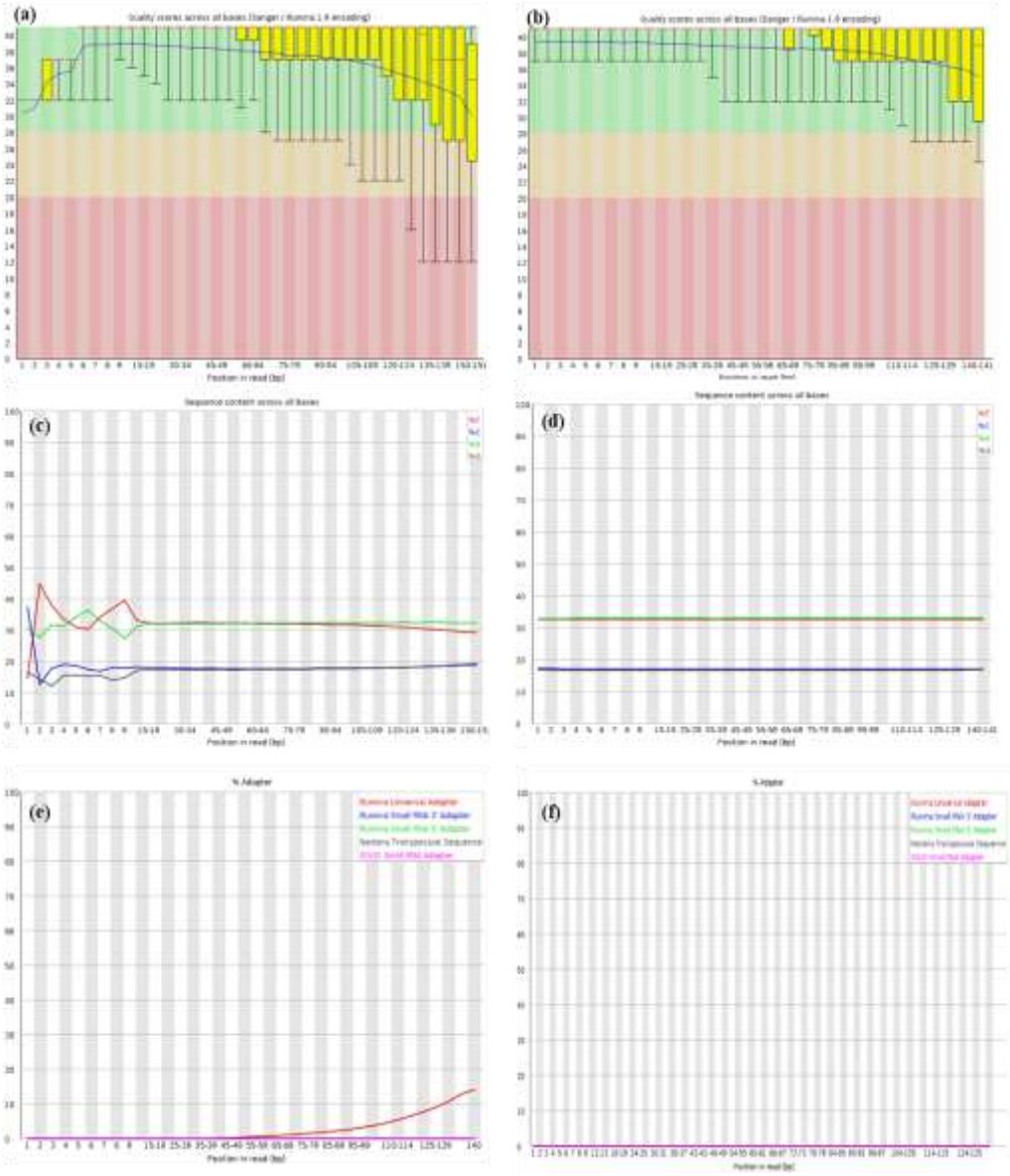


Fig. 4.8: FastQC results before (a,c,e) and after trimmomatic (b,d,f). (a) and (b) shows the per base sequence quality, (c) and (d) shows the per base sequenc content, and (e) and (f) displays the presence of adapter content

All of these undesirable bases and sequences were removed using Trimmomatic version 0.38. Trimmomatic is a read trimming and adapter clipping tool which was used to remove the poor quality reads, and trim the undesirable bases and adapter sequences. After trimming, two output files (paired and unpaired files) were generated for both forward and reverse sequences. The paired files contain sequence reads for which both the forward and reverse reads survived the processing whereas the unpaired files contain the sequence reads whose partner reads didnot survive the processing. The files obtained after trimming were again subjected to FastQC to check for any remaining undesirable reads.

Trimmomatic removed 8,899, 10,240, 296,229, and 275,781 undesirable reads from MS-1, KP4HM-15, MSB and MFB respectively. And a total of 52.4 million clean reads for MS-1, 53.5 million clean reads for KP_4HM-15, 43.9 million clean reads for MSB, and 62.4 million clean reads for MFB were achieved. The clean reads thus obtained had (i) quality score more than 30 (> Q30) (Fig. 4.7 b); (ii) equal proportion of A to T and G to C (Fig. 4.7 d); and (iii) no adapter sequences (Fig. 4.7f). The statistics after the trimming of the reads is shown in the Table 4.4.

Table 4.4: Summary of sequencing data- raw input reads and reads obtained after trimming

Genotype	Input read pairs	Reads after trimming			
		Both surviving	Forward only surviving	Reverse only surviving	Dropped
MS-1	53,753,767	52,409,420 (97.50%)	1,253,084 (2.33%)	82,364 (0.15%)	8,899 (0.02%)
KP_4HM-15	55,025,107	53,509,970 (97.25%)	1,412,135 (2.57%)	92,762 (0.17%)	10240 (0.02%)
MSB	55,577,927	43,962,282 (79.10%)	11,148,220 (20.06%)	171,196 (0.31%)	296,229 (0.53%)
MFB	72,827,049	62,429,792 (85.72%)	9,893,945 (13.59%)	22,753 (0.31%)	275,781 (0.38%)

4.2.2 SNP-index method for fine mapping *ms-1* genomic locus

The clean reads from MS-1 were aligned to DHL92 (v3.6.1) reference genome using Bowtie2 v2.4.2 (Langmead & Salzberg 2012). A genomic coverage of about 96.85% was achieved for MS-1. The reference-guided MS-1 assembly was developed by substituting the variant alleles of reference DHL92 in place of MS-1 alleles using SAMtools & BCFtools (Li *et al* 2009 & Li 2011). The resultant MS-1 reference assembly was used in the QTL-seq

pipeline (Sugihara *et al* 2020) to identify the genomic locus linked to *ms-1*. The qtl-seq pipeline outputs the results into a directory which further contains five sub-directories namely, 10_ref (indexed reference files), 20_bam (filtered alignment files), 30_vcf (variant called files), 40_qtlseq (SNP-index files and graphs), and log (log files).

The clean reads of MS-1, MSB, and MFB were aligned to the MS-1 reference assembly using the in-built BWA mem program within the QTL-seq pipeline. The BWA index files for the MS-1 reference assembly were output in the 10_ref directory and the alignment files were saved in the 20_bam after filtering the incorrectly aligned reads using SAMtools. The alignment files for three samples were then piled-up using BCFtools followed by variant-calling to generate a variant-called file (.vcf) file. The variants were called for the three samples (MS-1, MSB and MFB) with respect to the MS-1 reference assembly, developed previously. The variant file thus generated was saved in the 30_vcf directory and the contents of the .vcf file are shown in Fig. 4.9. The .vcf file contains 3 parts:

i. Meta-information lines, beginning with double-hash symbol (##) and contains the lines describing the INFO, FILTER and FORMAT entries used in the body of the .vcf file.

ii. Header line, beginnig with a single-hash symbol (#) and contains header for 8 mandatory columns (CHROM, POS, ID, REF, ALT, QUAL, FILTER and INFO) along with an additional FORMAT column

iii. Data lines, contains the data for corresponding to header fields for each entry i.e. for each variant called.

Fig. 4.9: Contents of a .vcf file generated after variant-calling using BCFtools

A total of 68,400,494 variants were called in the .vcf file which were then filtered based on the default filtering criteria set within the QTL-seq pipeline. 3,190,703 variants were

retained after the filtering step and were used for calculating the SNP-index values for both the bulks (Fig. 4.16 b,c). Two types of tab-separated files were generated after SNP-index calculations: snp_index.tsv and sliding_window.tsv, and were saved in the 40_qtlseq directory. The snp_index.tsv file contained the calculated SNP-index and ΔSNP-index values for each SNP along with the computer-simulated expected values for ΔSNP-index at 95% and 99% level of significance forming the confidence intervals within which no variant is associated with the *ms-1*. On the other hand, sliding_window.tsv file contained the mean SNP-index and ΔSNP-index values calculated for the variants present in the 1MB interval along with the mean threshold values for ΔSNP-index at 95% and 99% level of significance. The contents of the snp_index.tsv and sliding_window.tsv file are shown in Fig. 4.10 and Fig. 4.11, respectively. Two additional files each for snp_index and sliding_window were also generated, which contained the variants having SNP-index values lying outside the statistically calculated confidence-interval at 99% and 95% level of significance. These files were snp_index.p99.tsv, snp_index.p95.tsv, sliding_window.p99.tsv, and sliding_ window.p99.tsv.

The SNP-index values for each bulk corresponds to the frequency of each parental allele in the population of the bulked samples. The SNP-index of 0.5 indicates equal contribution of both the parents. On the other hand, the SNP-index of <0.5 or >0.5 indicates that the SNP is most likely to be contributed by the male-sterile parent or male-fertile parent, respectively. It was observed that the ΔSNP-index values in the region 36,800,00 to 37,200,00 on chromosome six lied outside the statistically calculated confidence interval at 99% level of significance. While at 95% level of significance, it was observed that two regions on chromosome six (26,200,000 bp to 31,100,00 bp and 35,300,000 bp to 37,200,000 bp) lied outside the calculated confidence interval. It was also noted that the mean SNP-index values in these regions were less than 0.5 for MSB, indicating that the region is contributed by the sterile-parent MS-1 (used as reference) while the mean SNP-index values more than 0.5 for MFB, indicate the contribution of male-fertile parent, KP_4HM-15 (not used as reference) in these regions. The Table 4.5 and 4.6 shows the calculated and expected SNP-index and ΔSNP-index values for the *ms-1* associated regions at 99% and 95% level of significance respectively.

To vizualize the SNP-index data, a plot for SNP-index and ΔSNP-index values were generated by QTL-seq. The Fig. 4.12 and 4.13 displays the graphs obtained after plotting the SNP-index values for MSB and MFB with respect to their genomic locations. The green and yellow dots represent the SNP-index value for each variant in MSB and MFB, respectively and the red line indicates the average of these SNP-index values within 1Mb sliding window.

It is observed that the mean SNP-index values lies close to 0.5 at all the genomic locations or have same values for both the bulks except at chromosome six. The mean SNP-index values on chromosome six deviates significantly from 0.5 in opposite directions between the bulks. Most of the SNP-index values are less than 0.5 in MSB and more than 0.5 in MFB, indicating that the genomic locus associated with *ms-1* is on chromosome six. This data correlated with the previous study on *ms-1* mapping by Singh *et al* (2019), where they found SSR markers linked to *ms-1* gene on chromosome six.

1	CHROM	POSI	VARIANT	DEPTH 1	DEPTH 2	MEAN p99	MEAN p95	MEAN SNP-index 1	MEAN SNP-index 2	MEAN DELTA SNP-index
2	chr00	405	indel	13	15	0.5897	0.4821	0.4615	0.6667	0.2051
3	chr00	408	indel	16	15	0.6083	0.4792	0.3125	0.4667	0.1542
4	chr00	3164	snp	23	23	0.5217	0.4348	0.6522	0.6522	0
5	chr00	4769	snp	9	20	0.6389	0.4889	0.6667	0.5	-0.1667
6	chr00	6470	snp	21	25	0.5371	0.4248	0.6667	0.56	-0.1067
7	chr00	6911	snp	10	21	0.6143	0.4714	0.5	0.3333	-0.1667
8	chr00	7045	snp	16	15	0.6083	0.4792	0.75	0.5333	-0.2167
9	chr00	12010	snp	22	27	0.5337	0.4141	0.2273	0.5556	0.3283
10	chr00	17055	indel	16	45	0.5236	0.4125	0.3125	0.2667	-0.0458
11	chr00	18631	snp	8	18	0.6389	0.5139	0.25	0.3333	0.0833
12	chr00	19878	snp	8	23	0.6576	0.5217	0.25	0.3913	0.1413
13	chr00	21308	snp	12	16	0.6042	0.4792	0.5833	0.5625	-0.0208
14	chr00	24734	snp	10	32	0.5875	0.4625	0.8	0.4375	-0.3625
15	chr00	25417	indel	10	16	0.625	0.4875	0.8	0.25	-0.55
16	chr00	25502	snp	9	19	0.6257	0.5088	0.4444	0.4211	-0.0234
17	chr00	27735	snp	13	22	0.6014	0.451	0.3077	0.5455	0.2378
18	chr00	80170	indel	15	13	0.5897	0.4821	0.6667	0.6923	0.0256
19	chr00	80238	snp	10	13	0.6462	0.5154	0.2	0.3846	0.1846
20	chr00	80928	snp	39	51	0.4887	0.371	0.2051	0.3137	0.1086
21	chr00	81070	snp	30	39	0.5	0.3923	0.3	0.359	0.059
22	chr00	81197	snp	9	17	0.6536	0.4902	0.7778	0.5294	-0.2484
23	chr00	92749	snp	13	20	0.5731	0.4462	0.5385	0.3	-0.2385

Fig. 4.10: Contents of the SNP_inde.tsv file

1	CHROM	POSI	MEAN p99	MEAN p95	MEAN SNP-index 1	MEAN SNP-index 2	MEAN DELTA SNP-index
2	chr00	1000000	0.5968	0.4662	0.4614	0.3917	-0.0697
3	chr00	1100000	0.5967	0.4659	0.4562	0.3862	-0.07
4	chr00	1200000	0.5947	0.4643	0.4463	0.3818	-0.0645
5	chr00	1300000	0.5955	0.4651	0.4439	0.3816	-0.0623
6	chr00	1400000	0.5952	0.4647	0.4417	0.3827	-0.059
7	chr00	1500000	0.5959	0.4652	0.4521	0.3879	-0.0642
8	chr00	1600000	0.5961	0.4654	0.4514	0.3885	-0.063
9	chr00	1700000	0.598	0.467	0.4484	0.3936	-0.0548
10	chr00	1800000	0.5958	0.4656	0.4406	0.3932	-0.0475
11	chr00	1900000	0.5962	0.4661	0.4447	0.4004	-0.0442
12	chr00	2000000	0.5921	0.4627	0.4421	0.4068	-0.0353
13	chr00	2100000	0.5922	0.4628	0.4423	0.4059	-0.0364
14	chr00	2200000	0.592	0.4622	0.4272	0.4153	-0.0119
15	chr00	2300000	0.592	0.4624	0.4254	0.4148	-0.0105
16	chr00	2400000	0.5991	0.4685	0.4376	0.4289	-0.0087
17	chr00	2500000	0.5987	0.4685	0.426	0.4449	0.0188
18	chr00	2600000	0.5984	0.4679	0.4362	0.4423	0.0061
19	chr00	2700000	0.5977	0.4661	0.4378	0.4324	-0.0054
20	chr00	2800000	0.5754	0.4523	0.4488	0.3993	-0.0495
21	chr00	2900000	0.5141	0.4019	0.3745	0.3155	-0.059
22	chr00	3000000	0.5111	0.3995	0.3729	0.3241	-0.0489
23	chr00	3100000	0.5008	0.3902	0.3684	0.334	-0.0343

Fig. 4.11: Contents of the sliding_window.tsv file

Table 4.5: Table of intervals (1 Mb sliding window) with mean ΔSNP-index values greater than the statistically calculated ΔSNP-index values at $P \leq 0.01$

CHROM	POSI	mean_p99	mean_p95	Mean SNP-index (MSB)	Mean SNP-index (MFB)	Mean ΔSNP-index
chr06	36800000	0.6076	0.4753	0.0447	0.653	0.6084
chr06	36900000	0.6084	0.4758	0.0433	0.6538	0.6105
chr06	37000000	0.6097	0.477	0.0419	0.663	0.6211
chr06	37100000	0.6101	0.4772	0.0458	0.6711	0.6253
chr06	37200000	0.6098	0.4771	0.0476	0.6741	0.6265

Table 4.6: Table of intervals (1 Mb sliding window) with mean ΔSNP-index values greater than the statistically calculated ΔSNP-index values at $P \leq 0.05$

CHROM	POSI	mean_p99	mean_p95	Mean SNP-index (MSB)	Mean SNP-index (MFB)	Mean ΔSNP-index
chr06	26200000	0.6142	0.4798	0.1067	0.5908	0.4841
chr06	26300000	0.614	0.4792	0.1024	0.5834	0.481
chr06	28000000	0.6119	0.4786	0.1216	0.6096	0.488
chr06	28100000	0.6125	0.4793	0.1166	0.6146	0.498
chr06	28200000	0.6126	0.4793	0.1183	0.6152	0.4968
chr06	28400000	0.6123	0.4791	0.1274	0.6122	0.4848
chr06	28500000	0.6127	0.4796	0.1313	0.6115	0.4802
chr06	28700000	0.6138	0.4803	0.1298	0.6113	0.4815
chr06	28800000	0.614	0.4805	0.1313	0.6118	0.4806
chr06	28900000	0.6139	0.4805	0.1261	0.6164	0.4903
chr06	29000000	0.6142	0.4808	0.1243	0.6193	0.495
chr06	29100000	0.6141	0.4806	0.128	0.6192	0.4913
chr06	30500000	0.6094	0.4765	0.1223	0.6044	0.482
chr06	30600000	0.6091	0.4759	0.1043	0.6122	0.5079
chr06	30700000	0.6061	0.4738	0.0951	0.616	0.5209
chr06	30800000	0.6073	0.475	0.0879	0.6156	0.5276
chr06	30900000	0.6069	0.4746	0.0876	0.614	0.5264
chr06	31000000	0.6069	0.4745	0.0989	0.6048	0.5059
chr06	31100000	0.6066	0.4743	0.1157	0.5952	0.4795

CHROM	POSI	mean_p99	mean_p95	Mean SNP-index (MSB)	Mean SNP-index (MFB)	Mean ΔSNP-index
chr06	35300000	0.6031	0.4707	0.0912	0.5654	0.4742
chr06	35400000	0.6037	0.4713	0.0847	0.5659	0.4812
chr06	35500000	0.6041	0.4717	0.0765	0.569	0.4925
chr06	35600000	0.6046	0.4721	0.0723	0.5774	0.505
chr06	35700000	0.6046	0.4721	0.0725	0.5786	0.5061
chr06	35800000	0.6052	0.4726	0.0711	0.5843	0.5132
chr06	35900000	0.6052	0.4727	0.0669	0.5867	0.5198
chr06	36000000	0.6054	0.473	0.062	0.5978	0.5358
chr06	36100000	0.6059	0.4736	0.0591	0.6149	0.5557
chr06	36200000	0.6059	0.4736	0.0594	0.6189	0.5595
chr06	36300000	0.6062	0.4737	0.056	0.6233	0.5674
chr06	36400000	0.607	0.4744	0.0547	0.6308	0.5762
chr06	36500000	0.6068	0.4744	0.0526	0.6351	0.5825
chr06	36600000	0.6073	0.475	0.05	0.6395	0.5895
chr06	36700000	0.6068	0.4745	0.0462	0.6499	0.6037
chr06	36800000	0.6076	0.4753	0.0447	0.653	0.6084
chr06	36900000	0.6084	0.4758	0.0433	0.6538	0.6105
chr06	37000000	0.6097	0.477	0.0419	0.663	0.6211
chr06	37100000	0.6101	0.4772	0.0458	0.6711	0.6253
chr06	37200000	0.6098	0.4771	0.0476	0.6741	0.6265

To vizualize the genomic region(s) harbouring *ms-1* gene, the SNP-index graphs for both the bulks were combined by subtracting the SNP-index values of MSB form those of MFB and the resultant ΔSNP-index values were plotted against their genomic positions (Fig. 4.14 and Fig. 4.16). The blue dots represent the individual ΔSNP-index values and the red line is the mean of these values in a 1 Mb genomic interval. The ΔSNP-index values within the orange lines and green lines represent the 99% and 95% confidence intervals of no QTL, respectively. The mean ΔSNP-index values crossing these intervals are *ms-1* linked genomic regions and it was observed that the regions between 36,800,000 bp and 37,200,000 bp lies outside the maximum confidence interval ($P \leq 0.01$). The two genomic regions crossing 95% confidence interval were also seen after plotting the ΔSNP-index values.

It is inferred from the QTL-seq results, that the highest absolute ΔSNP-index values were obtained in the region 36.8 Mb to 37.2 Mb on chromosome six (Table 4.5 and Fig.

4.14). The ΔSNP-index values in this region lied outside the maximum confidence interval calculated at $P \leq 0.01$, thereby rejecting the null hypothesis of no QTL and accepting the alternate hypothesis for presence of trait-linked QTL within the 400-kb interval (36.8 Mb to 37.2 Mb) on chromosome six. This 400-kb region on chromosome six was designated as candidate region and lied within the previously mapped region using SSR markers (Fig. 4.15). The reccent work done by Singh *et al* (2019), delimited the *ms-1* gene on chromosome six within the SSR markers, DM0038 and DM0187. The physical locations of these markers, with respect to DHL92 (v3.6.1) reference genome, is from 34.4 Mb (DM0038) to 37.6 Mb (DM0187). Thus, using whole genome resequencing approach in the present study, *ms-1* associated region was narrowed down from the previously mapped 3.2 Mb region to 400-kb region on chromosome six (Fig. 4.15).

This 400-kb candidate region harboured 720 variants comprising of 513 SNPs and 207 InDels. Of these 513 SNPs, 9 SNPs were highly significant, having ΔSNP-index value 1 at $P \leq 0.01$ (Table 4.7) and minimum depth of ≤ 8. The SNP-index (MSB) values for these highly significant SNPs were 0, indicating the contribution of alleles entirely from male-sterile parent MS-1, whereas the SNP-index (MFB) values were 1, indicating the contribution of alleles entirely from male-fertile parent KP4HM-15. These highly significant SNPs opens up the prospect for the development of allele-specific makers which can be used to improve the previously developed genetic map by Singh *et al* (2020) and thus for the efficient development of *ms-1* male-sterile lines.

Table 4.7: Positions of highly significant SNPs having ΔSNP-index = 1 (P ≤ 0.01)

CHROM	POSI	DEPTH (MSB)	DEPTH (MFB)	p99	p95	SNP-index (MSB)	SNP-index (MFB)	ΔSNP-index
chr06	36946327	15	21	0.581	0.4476	0	1	1
chr06	36960682	10	8	0.7	0.55	0	1	1
chr06	37015656	15	10	0.6333	0.5	0	1	1
chr06	37058969	13	15	0.5897	0.4821	0	1	1
chr06	37087976	9	13	0.6581	0.5128	0	1	1
chr06	37088025	8	11	0.6932	0.5455	0	1	1
chr06	37109800	12	11	0.6515	0.4848	0	1	1
chr06	37115410	10	13	0.6462	0.5154	0	1	1
chr06	37168159	11	10	0.7	0.5273	0	1	1

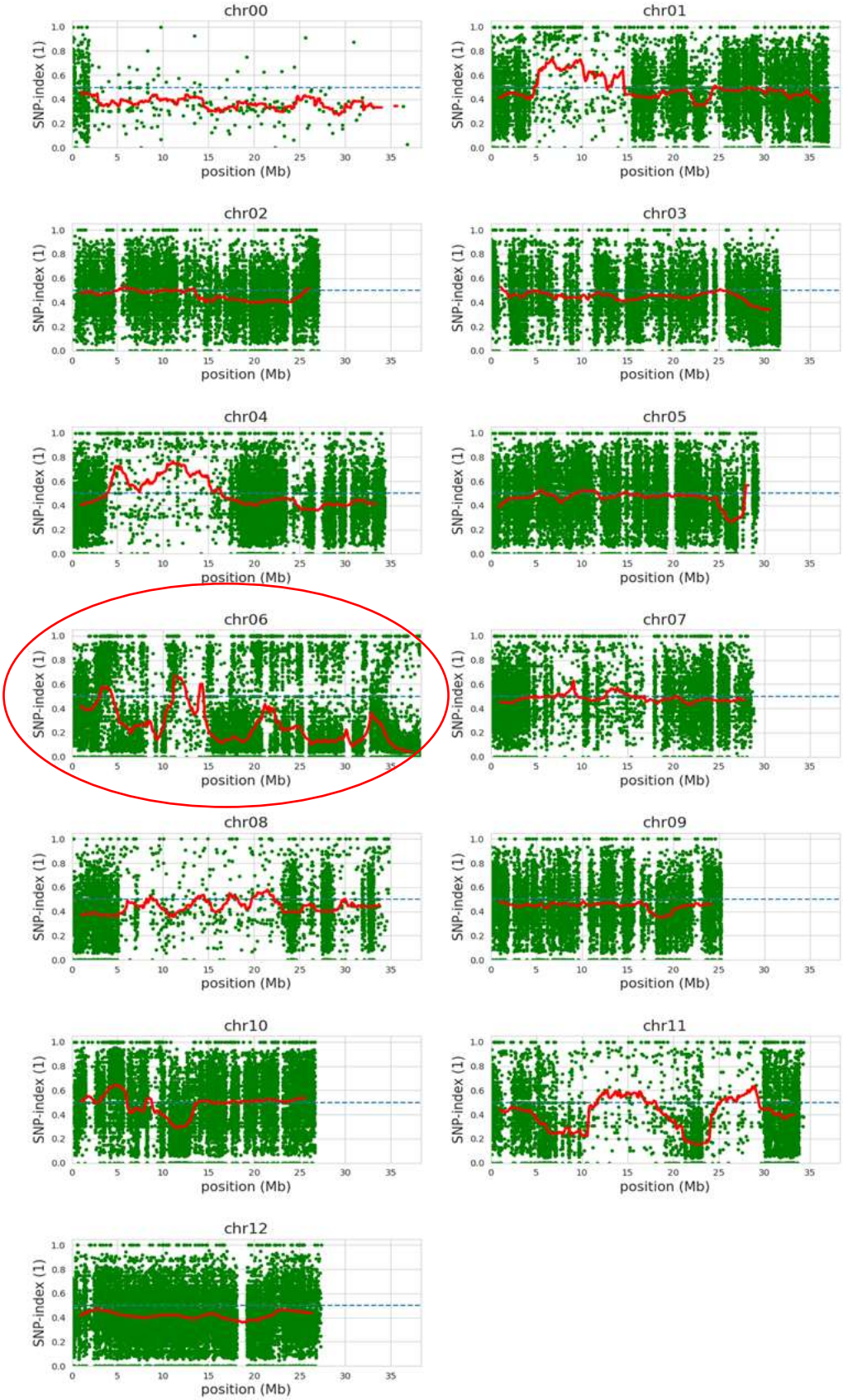


Fig. 4.12: SNP-index plot between male-sterile bulk (MSB) and MS-1 assembly

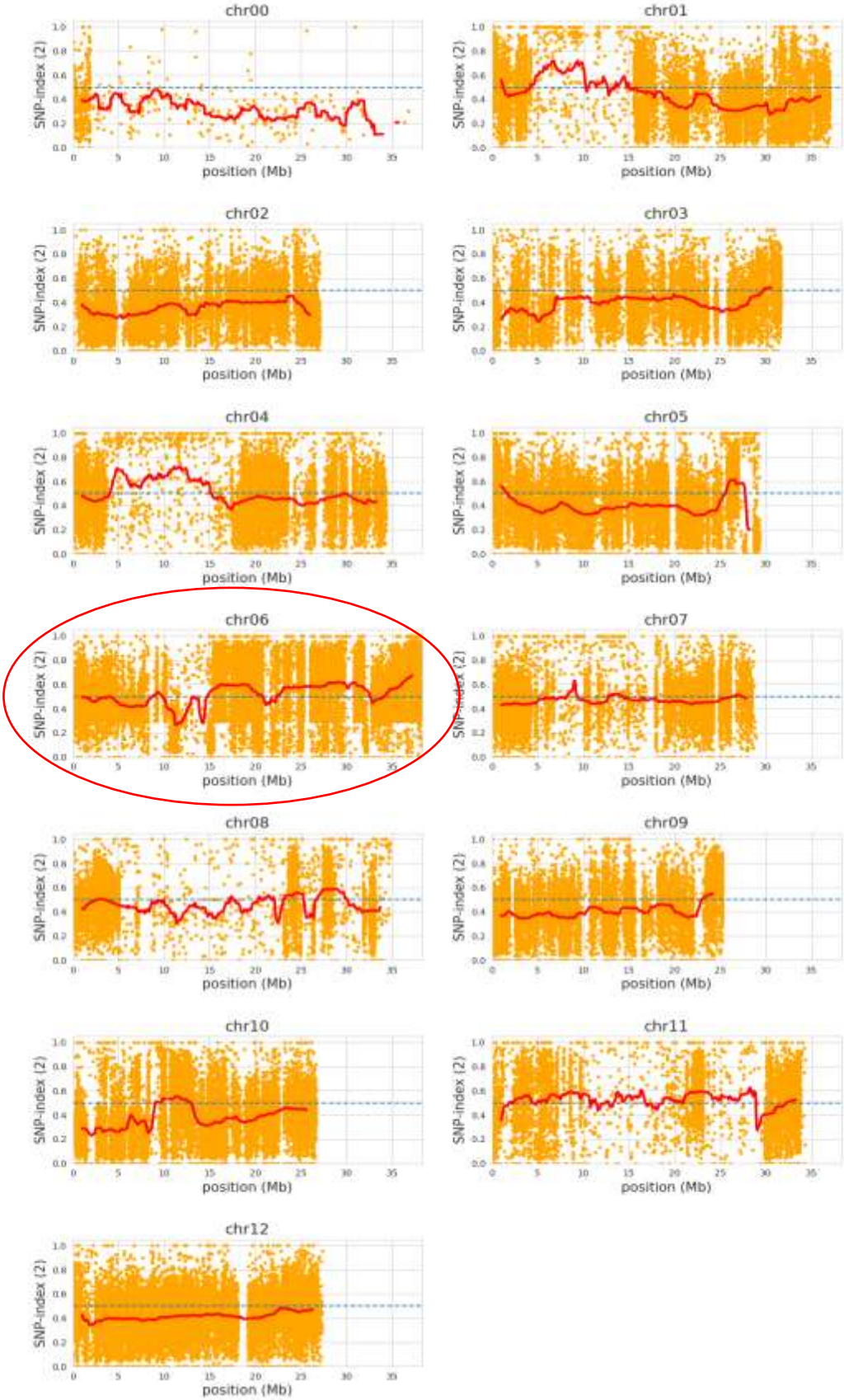


Fig. 4.13: SNP-index plot between male-fertile bulk (MFB) with MS-1 assembly

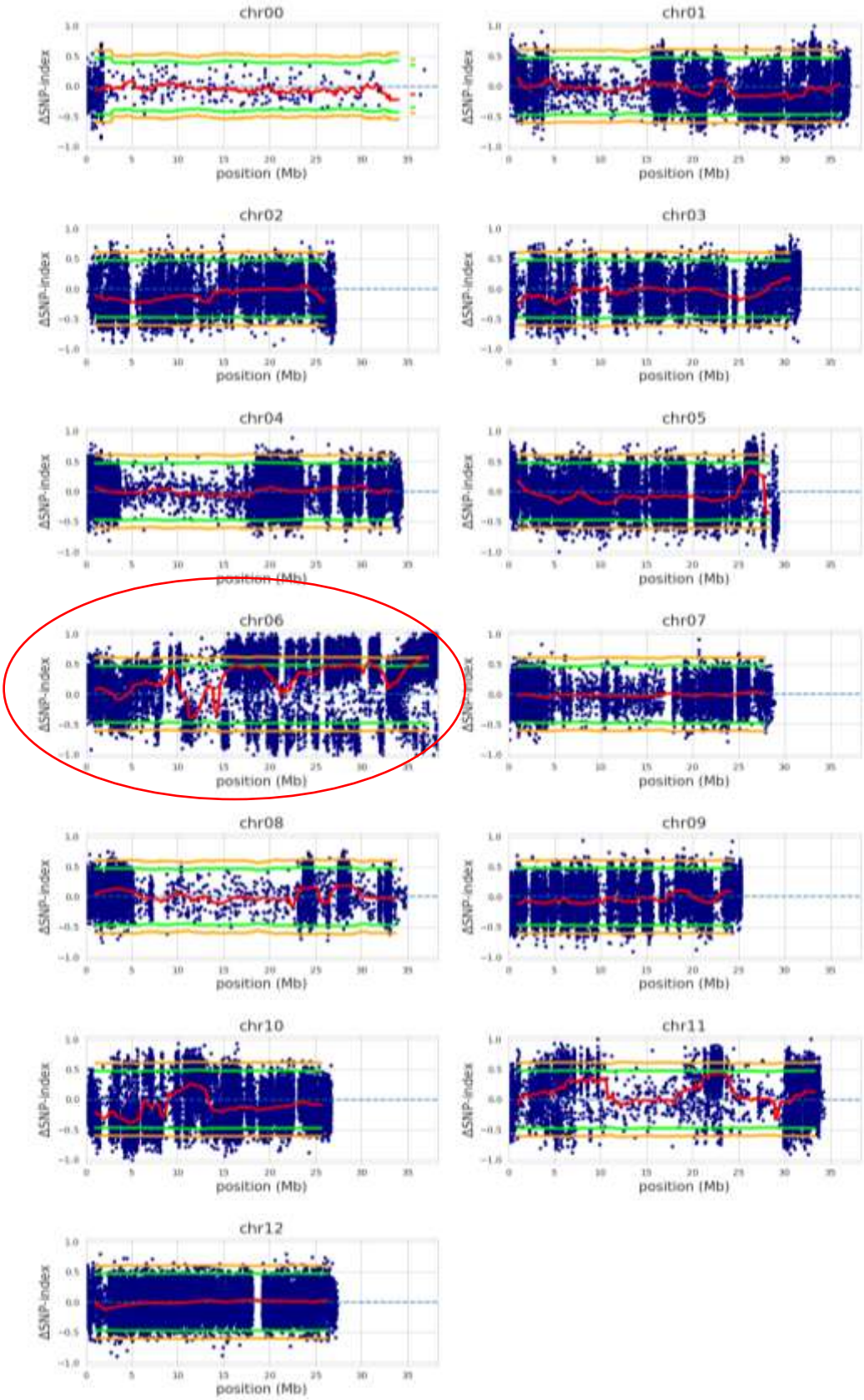


Fig. 4.14: ΔSNP-index plot (SNP-index MFB – SNP-index MSB). Green and orange lines represent the 95% and 99% confidence interval of simulated ΔSNP-index

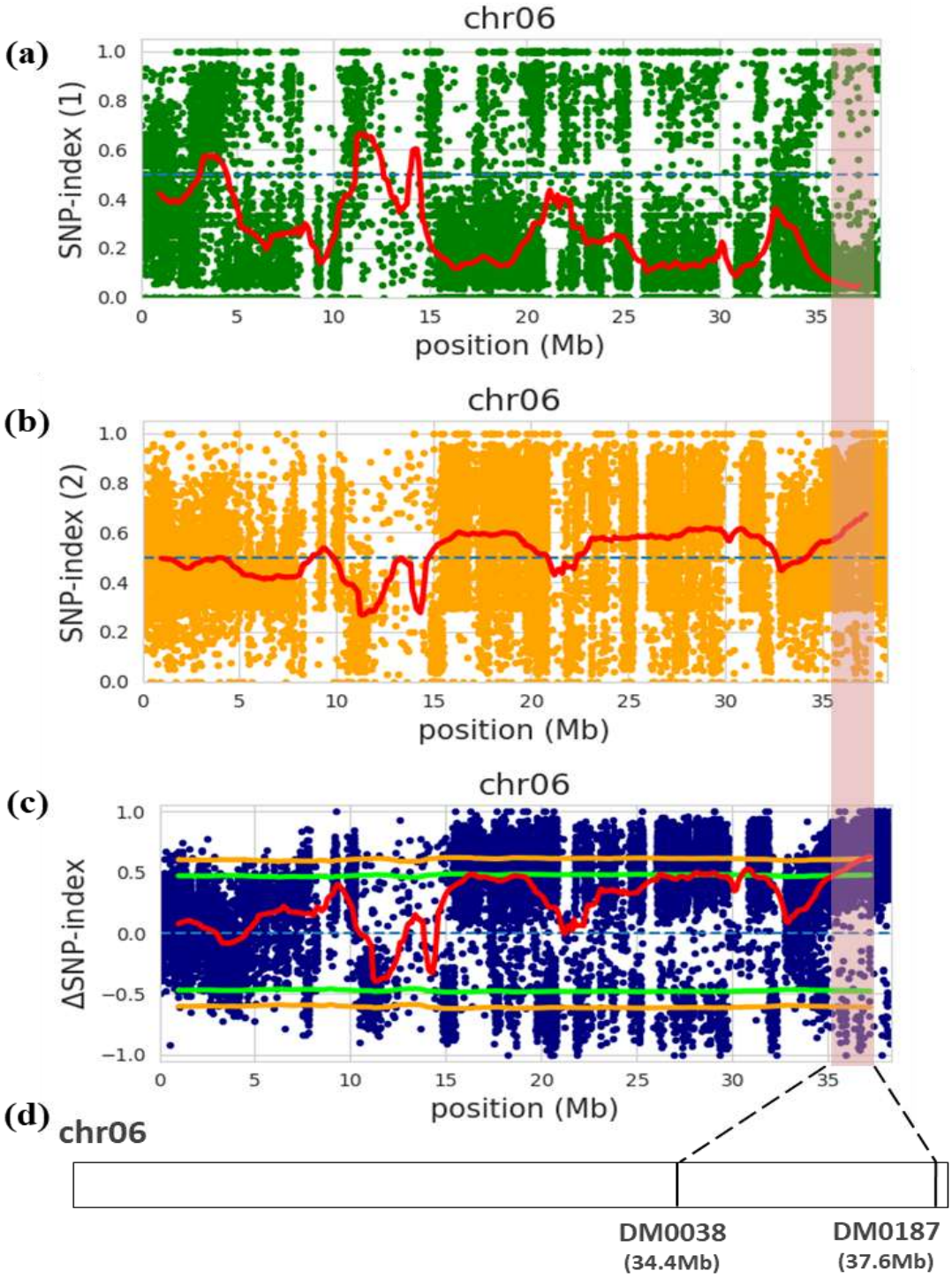


Fig. 4.15: Comparison of the mapped region of *ms-1* gene using QTL-seq (present study) and SSR markers (Singh *et al* 2019). The previously mapped 3.2 Mb region (d) was narrowed down to 400-kb region (a, b, and c) on chromosome six

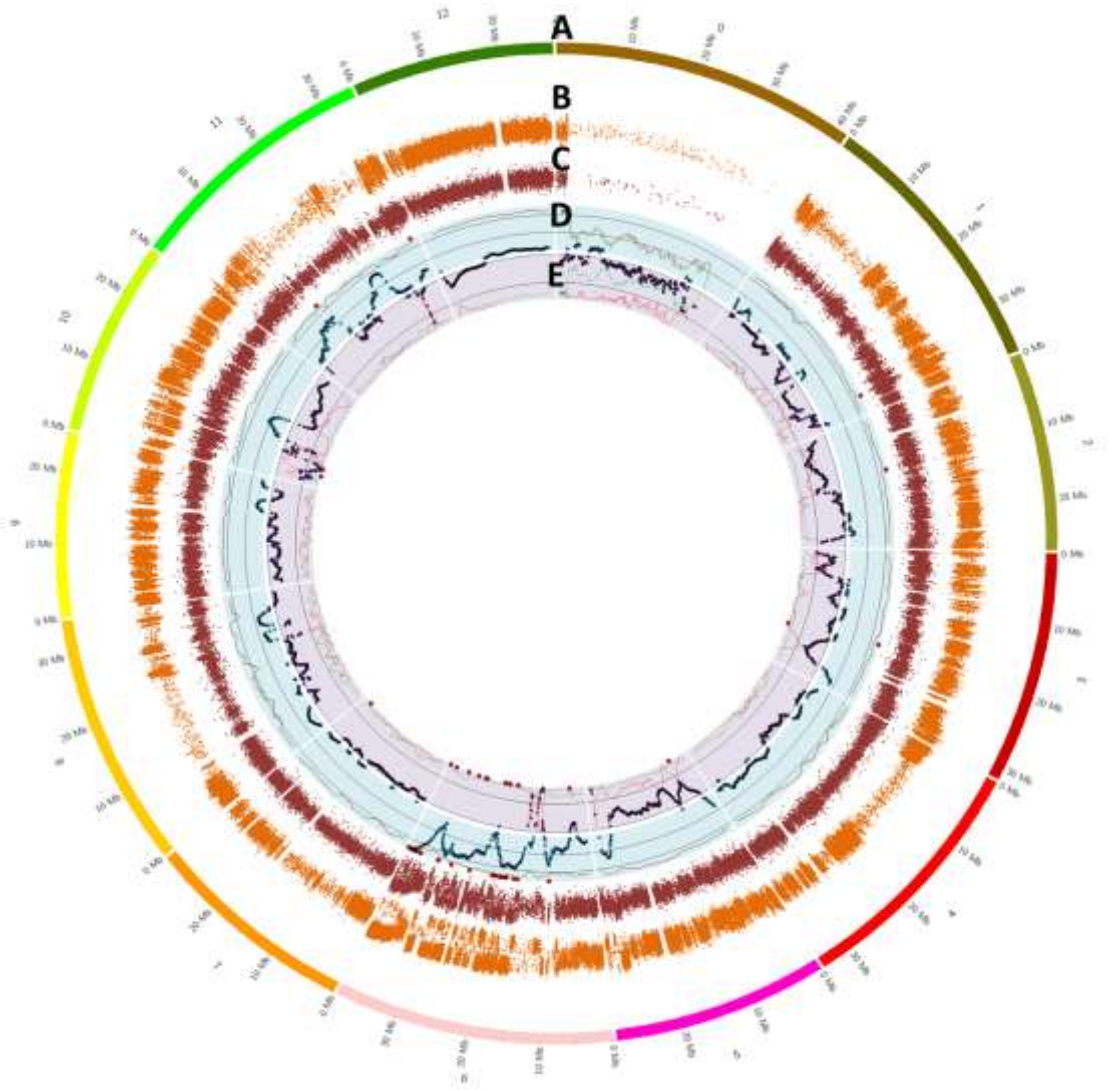


Fig. 4.16: Distribution of variants used in Qtl-seq software. (a) Track A represents the pseudomolecules of reference genome DHL92 (v3.6.1). (b) Track B and C displays the number of SNPs and InDels used to calculate ΔSNP-index values, respectively. (c) Track D and E plots the mean ΔSNP-index values ranging from 1 to 0 and 0 to -1 respectively, calculated using 1 Mb window size and 100-kb step size. (d) Blue coloured dots depicts the variants with ΔSNP-index values between 1 to 0, which represents the presence alleles from male-sterile parent, MS-1. (e) Purple coloured dots depicts the variants with ΔSNP-index values between 0 to -1, which represents the presence of alleles from male-fertile parent, KP_4HM-15. (f) Red and green lines in Track D and E represents the expected mean ΔSNP-index values at 99% and 95% level of significance, respectively. (g) The red dots represents the highly-significant SNPs having ΔSNP-index value of 1.

4.2.3 Functional annotation using Blast2GO

The 400-kb candidate region obtained after QTL-seq was further screened for the presence of genes. It was found that the region carried 43 genes. The function of these genes was assessed using Blast2GO software. It was observed that out of these 43 genes, a total of 35 genes returned hits with similar sequences after the Blast search in NCBI database and 28 of them were successfully annotated (Fig. 4.17). Blast2GO annotated these genes by assigning the appropriate GO term (Table 4.8). The detailed statistics of these annotations based on the three aspects of Gene Ontology i.e., biological process, cellular component, and molecular function, are represented graphically in Fig. 4.18.

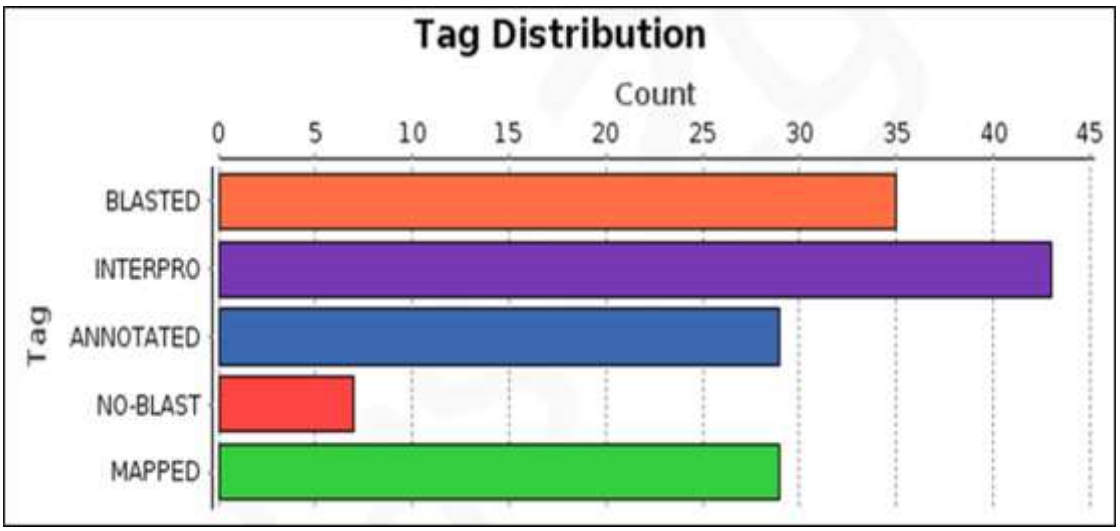


Fig. 4.17: A bar graph representing the distribution of genes undergone each step in Blast2GO

The following inferences were made from these graphs:

i. Based on the biological processes (Fig. 4.18a), most of the gene products were involved in the transcriptional regulation (16%), NAD biosynthesis (13%), methylation (13%) and protein transport (9%).

ii. Based on the cellular component (Fig. 4.18b), majority of the gene products were localized in the cell membrane (46%) and chloroplasts (17%).

iii. Based on molecular function (Fig. 4.18c), it was observed that a wide range of diversity existed amongst these genes.

The Blast2GO annotated genes were further searched online for their role in plant male-sterility. It was observed that six genes out of these 28 annotated genes were involved in the growth and development of the microspore or pollen, and the defect in these genes can causes male-sterility. These six genes were designated as putative candidate genes and are listed in Table 4.9. The role of these genes in the growth and development of male gametocytes are discussed as follows:

Table 4.8: Blast2GO annotations for 43 genes in the 400-kb candidate region

	Start	Stop	Gene ID	Blast2GO annotation
gene	36812195	36812420	MELO3C031969.2	---NA---
gene	36818464	36818990	MELO3C014093.2	---NA---
gene	36822091	36822408	MELO3C031970.2	CACTA en-spm transposon protein
gene	36826629	36826958	MELO3C014094.2	caffeic acid 3-O-methyltransferase-like
gene	36834645	36834890	MELO3C014095.2	xanthine dehydrogenase 1-like isoform X1
gene	36835038	36835199	MELO3C031971.2	xanthine dehydrogenase 1-like isoform X1
gene	36838016	36838170	MELO3C014096.2	LINE-1 retrotransposable element ORF2 protein
gene	36839006	36845090	MELO3C031972.2	---NA---
gene	36855378	36855632	MELO3C014097.2	---NA---
gene	36866774	36872102	MELO3C014098.2	caffeic acid 3-O-methyltransferase-like
gene	36887490	36888207	MELO3C014099.2	PREDICTED: uncharacterized protein LOC103491021
gene	36888237	36889250	MELO3C014100.2	chitin-binding lectin 1-like
gene	36904547	36904777	MELO3C014101.2	---NA---
gene	36919774	36922174	MELO3C032285.2	protein disulfide-isomerase-like
gene	36923822	36926060	MELO3C032286.2	---NA---
gene	36926106	36932783	MELO3C014102.2	glutamic acid-rich protein-like
gene	36946833	36947294	MELO3C014103.2	nuclear transcription factor Y subunit B-3-like
gene	36950934	36951793	MELO3C014104.2	gibberellin-regulated protein 14-like
gene	36954661	36958280	MELO3C014105.2	inactive beta-amylase 4, chloroplastic
gene	36959014	36959917	MELO3C014106.2	protein TIC 20-v, chloroplastic
gene	36961275	36962122	MELO3C014107.2	fasciclin-like arabinogalactan protein 1
gene	36965282	36968859	MELO3C014108.2	coiled-coil domain-containing protein 22 isoform X2

	Start	Stop	Gene ID	Blast2GO annotation
gene	36970564	36974285	MELO3C014109.2	nucleotide-sugar uncharacterized transporter 2 isoform X1
gene	36978464	36982360	MELO3C014110.2	oligopeptide transporter 1-like
gene	36990736	37009759	MELO3C014111.2	DNA-directed RNA polymerases I and III subunit RPAC1
gene	37026817	37031855	MELO3C014112.2	phosphatidylinositol/phosphatidylcholine transfer protein SFH2
gene	37032007	37036034	MELO3C014113.2	WEB family protein At5g55860
gene	37038667	37041938	MELO3C014114.2	protein NOI4-like
gene	37045965	37062736	MELO3C014115.2	E3 ubiquitin-protein ligase UPL1
gene	37072987	37077115	MELO3C014116.2	poly(A)-specific ribonuclease PARN isoform X1
gene	37076910	37079120	MELO3C014117.2	probable L-type lectin-domain containing receptor kinase S.7
gene	37096346	37100532	MELO3C014119.2	nicotinamide/nicotinic acid mononucleotide adenylyltransferase isoform X1
gene	37108875	37109569	MELO3C014120.2	nuclear transcription factor Y subunit B-9-like
gene	37114700	37118143	MELO3C014121.2	ABC transporter G family member 26
gene	37118681	37123688	MELO3C014122.2	protein CWC15 homolog
gene	37125425	37132392	MELO3C014123.2	uncharacterized WD repeat-containing protein C2A9.03-like
gene	37137089	37137912	MELO3C014124.2	photosystem I reaction center subunit V, chloroplastic-like
gene	37150274	37150511	MELO3C031974.2	acetylajmalan esterase-like
gene	37154257	37154999	MELO3C031975.2	GDSL esterase/lipase At5g03980-like
gene	37177893	37178060	MELO3C014126.2	---NA---
gene	37181695	37190730	MELO3C014127.2	actin-related protein 8
gene	37190859	37196408	MELO3C032287.2	ty3-gypsy retrotransposon protein
gene	37197902	37200295	MELO3C014128.2	probable calcium-binding protein CML21

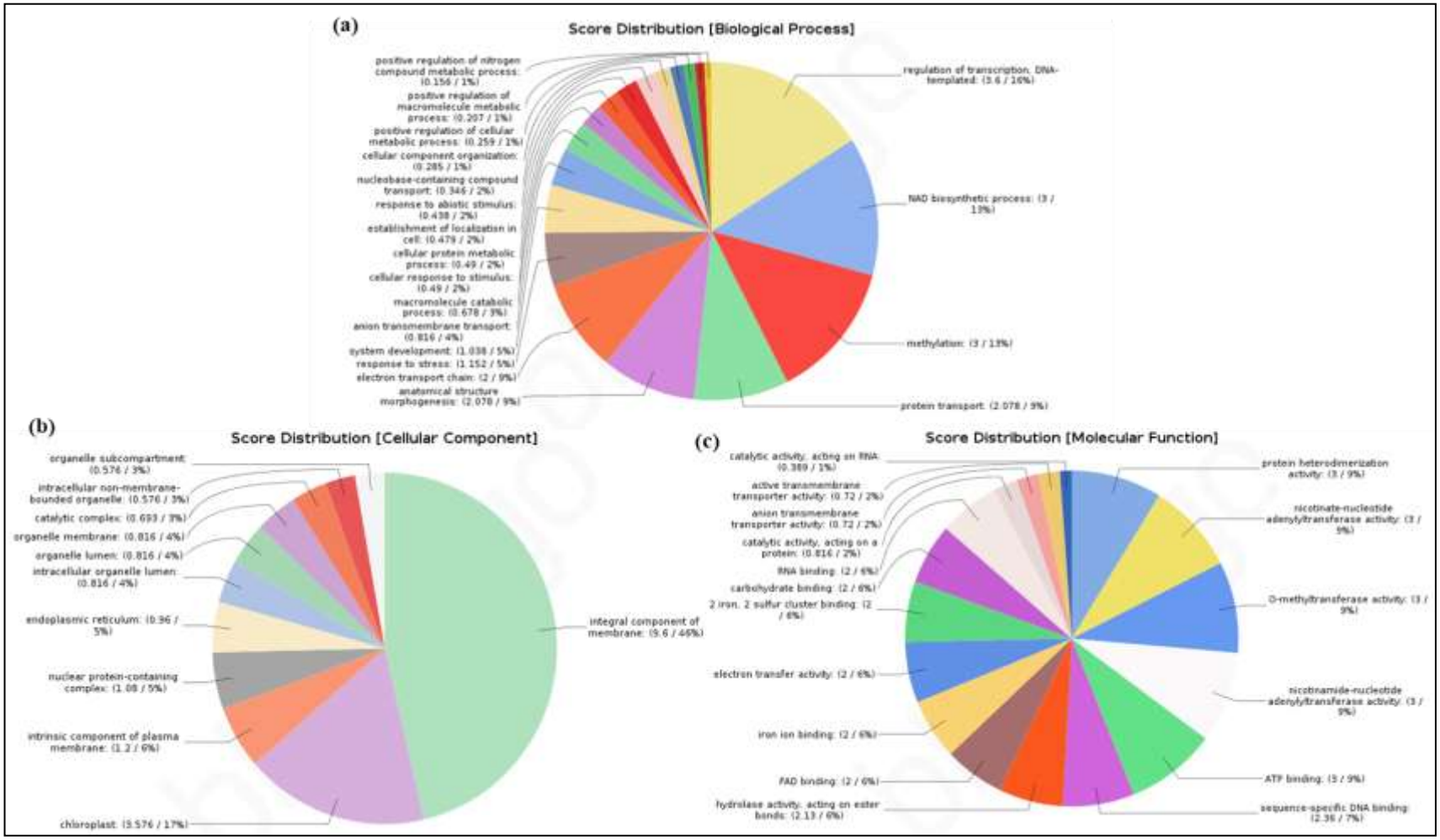


Fig. 4.18: Functional annotation for the 43 genes present within the 400 kb candidate region

i. Protein disulfide-isomerases (PDIs)

PDIs have been reported to be involved in providing thermo-tolerance to the pollens during heat-stress (Feldevard *et al* 2020) by acting as chaperon proteins and thereby facilitating the folding of nascent and misfolded proteins in the secretory pathway. Feldevard *et al* (2020) reported the up-regulation of PDI9 gene in heat-stressed Arabidopsis plant, leading to its increased accumulation in the pollen grains. The transfer-DNA (T-DNA) insertional mutations in PDI9 gene resulted in non-viable pollens exhibiting disruptions in the reticulated pattern of the exine along with an increased adhesion of pollen grains under heat stress.

ii. Fasciclin-like arabinogalactan proteins (FLAs):

The FLAs are abundant in reproductive tissues and are speculated to be involved in gametophytic cell differentiation, pollination and pollen tube growth (Cheung & Wu 1999 and Coimbra *et al*., 2007). The down-regulation of two AGP genes namely, AGP6 and AGP11 by RNAi and artificial microRNA resulted in pollen grain and pollen tube defects (Levitin *et al* 2008 and Coimbra *et al* 2009). Using the RNAi-mediated down-regulation of FLA3 gene in arabidopsis, Li *et al* 2010 reported the occurrence of 50% abnormal pollen grains with shrunken and wrinkled phenotypes. The transmission electron microscopy (TEM) revealed abnormalities in the intine layer of *FLA3*-RNAi plants.

iii. Nucleotide/sugar transporter family protein (NSTPs):

Reyes *et al (*2010) reported two NSTPs in Arabidopsis namely, AtUTr1 and AtUTr3, which are required for the incorporation of uridine 5'-disphosphate (UDP)-glucose into the ER and is essential for the pollen development. Transcriptomic analysis at different stages of pollen development has shown that AtUTr1 and AtUTr3 are expressed specifically in pollen (Bock *et al* 2006). The studies on the double mutant *atutr1 atutr3* pollen suggests that the UDP-glucose transport into the ER is necessary for late events in pollen development and could also be related to early steps during the tube growth.

iv. ATP-binding cassette (ABC) transporter:

ABC transporter proteins belong to the ATP-binding protein family that uses ATP-derived energy to transport molecules across cell membranes. The role of ABC transporter protein in pollen development has been widely reported (Kuromori *et al* 2011, Choi *et al* 2011, and Chang *et al* 2018). The abnormalities in the pollen exine deposition were observed in the T-DNA mutant line for *AtABCG26,* indicating the role of AtABCG26 in the transfer of exine substrates, such as sporopollenin from the tapetum to microspore (Kuromori *et al* 2011). Similar observations were reported in rice where a loss-of-function of OsABCG3 gene

resulted in abnormal degradation of tapetum thereby producing shriveled pollens lacking nexine II (exine layer) and intine layer (Chang 2018).

v. Transducin/WD40 repeat-like superfamily protein:

WD40 are 40–60 amino acids long motifs ending with a tryptophan and aspartic acid (WD) residues (Migliori *et al* 2012). Han *et al* (2006) reported a five WD40-repeat containing protein RICE IMMATURE POLLEN 1 (RIP1), which is necessary for pollen maturation and germination. *rip1* mutant pollen displayed delayed development along with delayed formation of starch granules and the intine layer. Ju *et al* (2016) reported inhibition of pollen germination due to overexpression of WD40 domain protein JINGUBANG (JGB) in Arabidopsis. Another WD40 domain containing protein, germinating modula-tor of rice pollen (GORI), was reported to be essential for pollen germination (Kim *et al* 2021).

vi. GDSL esterase/lipase proteins (GELPs):

GELPS are hydrolytic enzymes and the defects in these genes results in the delayed degradation of anther wall layers resulting in abnormal pollen exine formation (Zhang 2020a, Huo *et al* 2020, Yuan *et al* 2020, An *et al* 2020, and Zhang *et al* 2020b). An *et al* (2019) reported that a loss-of-function mutation in maize GDSL lipase encoding gene, ZmMs30 resulted in defective anther cuticle, irregular foot layer of pollen exine, and complete male sterility. Zhao *et al* (2020) characterized another GDSL lipase protein encoding gene in rice, Rice Male Sterile2 (RMS2) that is required for anther development and pollen fertility by causing tapetal degradation, the formation of anther cuticle and pollen exine, and central vacuole development. Huo *et al* (2020) reported the characterization of a GDSL lipase encoding male-sterile mutant *irregular pollen exine2 (ipe2)* of maize (*Zea mays*), which displays shrunken anthers and no starch accumulation in mature pollen grains.

Also, Table 4.10 shows the list of genes harbouring nine highly significant SNPs and it was observed that five of these SNPs were located within the genes, annotated as nuclear transcription factor Y subunit B-3-like (MELO3C014103.2.1), fasciclin-like arabinogalactan protein 1 (MELO3C014107.2), E3 ubiquitin-protein ligase UPL1 (MELO3C014115.2), ABC transporter G family member 26 (MELO3C014121.2), and protein CWC15 homolog (MELO3C014122.2) and the rest four highly significant SNPs were present in the intergenic region. As previously discussed, fasciclin-like arabinogalactan protein 1 (FLAs) and ABC transporter G family member 26 are involved in the male gametophyte development.

Table 4.9: List of six putative candidate genes and their known functions in male-sterility

Gene Id	Gene Description	Function	References
MELO3C032285.2 (36919774-36922174)	protein disulfide-isomerase-like	Provides pollen tolerance to heat-stress	Feldeverd *et al* 2020
MELO3C014107.2 (36961275-36962122)	Fasciclin-like arabinogalactan protein	Affects intine formation by altering cellulose deposition	Levitin *et al* 2008, Coimbra *et al* 2009, and Li *et al* 2010
MELO3C014109.2 (36970564-36974285)	Nucleotide/sugar transporter family protein	Supply nucleotides and sugar to developing pollen.	Reyes *et al* 2010
MELO3C014121.2 (37114700-37118143)	ABC transporter G family member 26	Involved in transport of sporopollenin monomers (exine component) from tapetum to microspores.	Kuromori *et al* 2011, Choi *et al* 2011, and Chang *et al* 2018
MELO3C014123.2 (37125425-37132392)	Transducin/WD40 repeat-like superfamily protein	Overexpression of this gene resulted in delayed growth and male-sterile phenotype	Han *et al* (2006), Ju *et al* 2016, and Kim *et al* 2021
MELO3C031975.2 (37154257-37154999)	GDSL esterase/lipase At5g03980-like	Involved in lipid metabolism during degradation of tapetal cells and middle layer affecting exine formation	Yuan *et al* 2020 and An *et al* 2019

4.2.4 Variant annotation using snpEff

The genome-wide variants were also analyzed using snpEff software (Cingolani *et al* 2012), which annotated and predicted the effecct of these variants on genes and gene products. A .vcf file containing genome-wide variants was used as an input file in snpEff software. In total 68.4 million variants were called in the .vcf file, out of which 3.19 million variants were used in the Qtl-seq software for fine mapping the *ms-1* associated region. These 3.19 million variants were present on all chromosomes with an average rate of 1 variant every 130 bases. The chromosome-wise distribution of these variants is given in Table 4.10. Most of these variants were present on chromosome 0, which contains the unassigned scaffolds or chromosomal regions. The majority of these variants were deletions, followed by single nucleotide polymorphisms (SNPs), insertions and mixed types (Table 4.11). These variants were then used to annotate and predict their effects on the genes and gene products.

Table 4.10: Chromosome-wise distribution and rate of variants

Chromosome	Length	Variants	Variants rate
1	37,037,532	351,453	105
2	27,064,691	242,606	111
3	31,666,927	278,526	113
4	34,318,044	324,470	105
5	29,324,171	61,276	478
6	38,297,372	304,226	125
7	28,958,359	160,802	180
8	34,765,488	257,077	135
9	25,243,276	140,278	179
10	26,663,822	329,971	80
11	34,457,057	367,806	93
12	27,563,660	360,514	76
0	41,641,883	11,698	3,559
Total	**417,002,282**	**3,190,703**	**130**

The snpEff software predicted a total of 6,353,469 effects for 3,190,703 variants provided and categorized these effects on the basis of their impact (Table 4.12) and functional class (Table 4.13). The majority of these variants (99.36%) had negligible (MODIFIER) effect on the genes i.e. these were either non-coding variants or variants affecting the non-coding regions of the genes, where predictions are difficult or there is no evidence of impact. Examples of variants with MODIFIER impact are exon variant and downstream gene variant. The HIGH impact variants which are assumed to have disruptive impact on the proteins were only 0.376% of the total variants and these HIGH impact variants may cause either gain or

loss of stop codon or frame-shift mutations. The LOW impact variants which are mostly harmless and unlikely to change protein behavior were 0.114% and MODERATE impact variants with non-disruptive impact that might change protein effectiveness were 0.15% of the total variants.

It was also observed that about half of the variants (51.83%) produced silent mutations and 47 % produced missense mutations while 0.99% variants produced nonsense mutations. The snpEff software aslo generated the variant effect data on the basis of type and region (Table 4.14) and it was observed that most of the variants were intergenic region variant (38%) followed by upstream variant (21.76%), downstream variant (17.87%), intronic region variant (10.51%) and intragenic region variant (10.07%).

Table 4.11: Number of variants by type

Variant Type	Total
Single Nucleotide Polymorphism (SNP)	163,570
Multiple Nucleotide Polymorphism (MNP)	0
Insertions (INS)	50,161
Deletions (DEL)	2,976,969
MIXED	3
Total	3,190,703

Table 4.12: Number of effects by impact

Impact Type (alphabetical order)		Count	Percent
HIGH		23,868	0.376%
LOW		7,255	0.114%
MODERATE		9,530	0.15%
MODIFIER		6,312,816	99.36%

Table 4.13: Number of effects by functional class

Functional Class Type (alphabetical order)		Count	Percent
MISSENSE		4,255	47.168%
NONSENSE		90	0.998%
SILENT		4,676	51.835%

Table 4.14: Number of effects by type and region

	Type (alphabetical order)	Count	Percent
TYPE	3_prime_UTR_variant	44,967	0.707%
	5_prime_UTR_premature_start_codon_gain_variant	251	0.004%
	5_prime_UTR_variant	25,329	0.398%
	bidirectional_gene_fusion	1	0%
	conservative_inframe_deletion	767	0.012%
	conservative_inframe_insertion	174	0.003%
	disruptive_inframe_deletion	4,254	0.067%
	disruptive_inframe_insertion	127	0.002%
	downstream_gene_variant	1,135,739	17.845%
	frameshift_variant	19,407	0.305%
	intergenic_region	2,414,459	37.936%
	intragenic_variant	641,480	10.079%
	intron_variant	674,797	10.602%
	missense_variant	4,220	0.066%
	non_coding_transcript_variant	73	0.001%
	splice_acceptor_variant	94	0.001%
	splice_donor_variant	4,235	0.067%
	splice_region_variant	6,768	0.106%
	start_lost	54	0.001%
	start_retained_variant	2	0%
	stop_gained	110	0.002%
	stop_lost	28	0%
	stop_retained_variant	8	0%
	synonymous_variant	4,680	0.074%
	upstream_gene_variant	1,382,511	21.722%

	Type (alphabetical order)	Count	Percent
REGION	DOWNSTREAM	1,135,738	17.876%
	EXON	33,681	0.53%
	GENE	1	0%
	INTERGENIC	2,414,459	38.002%
	INTRON	668,291	10.519%

REGION	Type (alphabetical order)	Count	Percent
	SPLICE_SITE_ACCEPTOR	92	0.001%
	SPLICE_SITE_DONOR	4,231	0.067%
	SPLICE_SITE_REGION	2,400	0.038%
	TRANSCRIPT	641,553	10.098%
	UPSTREAM	1,382,511	21.76%
	UTR_3_PRIME	44,948	0.707%
	UTR_5_PRIME	25,564	0.402%

After obtaining genome-wide variant annotation file, the variant effects and annotations in the 400-kb candidate region were extracted and analyzed. It was observed that the candidate region comprised of 207 SNPs and 513 InDels (Fig. 4.19) and their distribution within the 28 annotated genes in shown in Fig. 4.20.

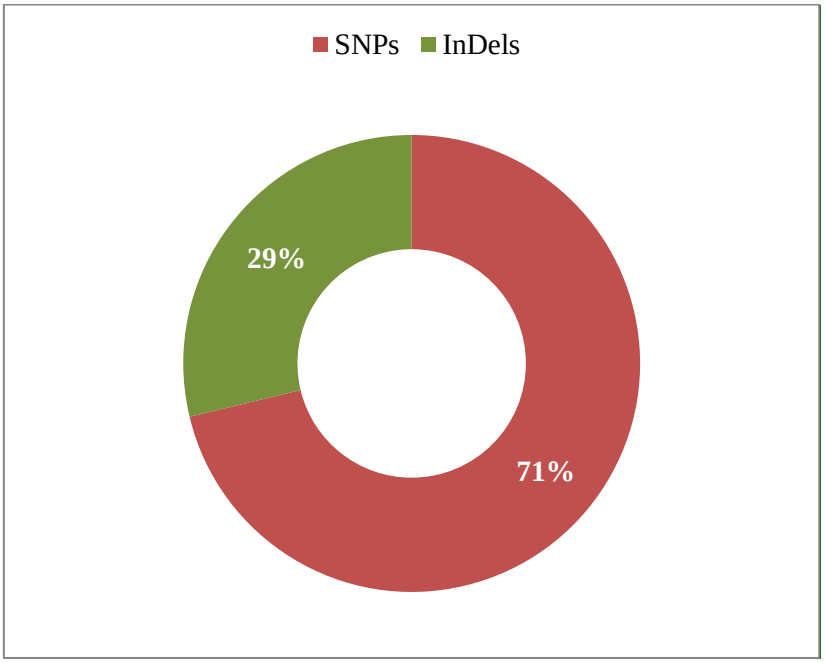


Fig. 4.19: Proportion of SNPs and InDels in the 400-kb candidate region

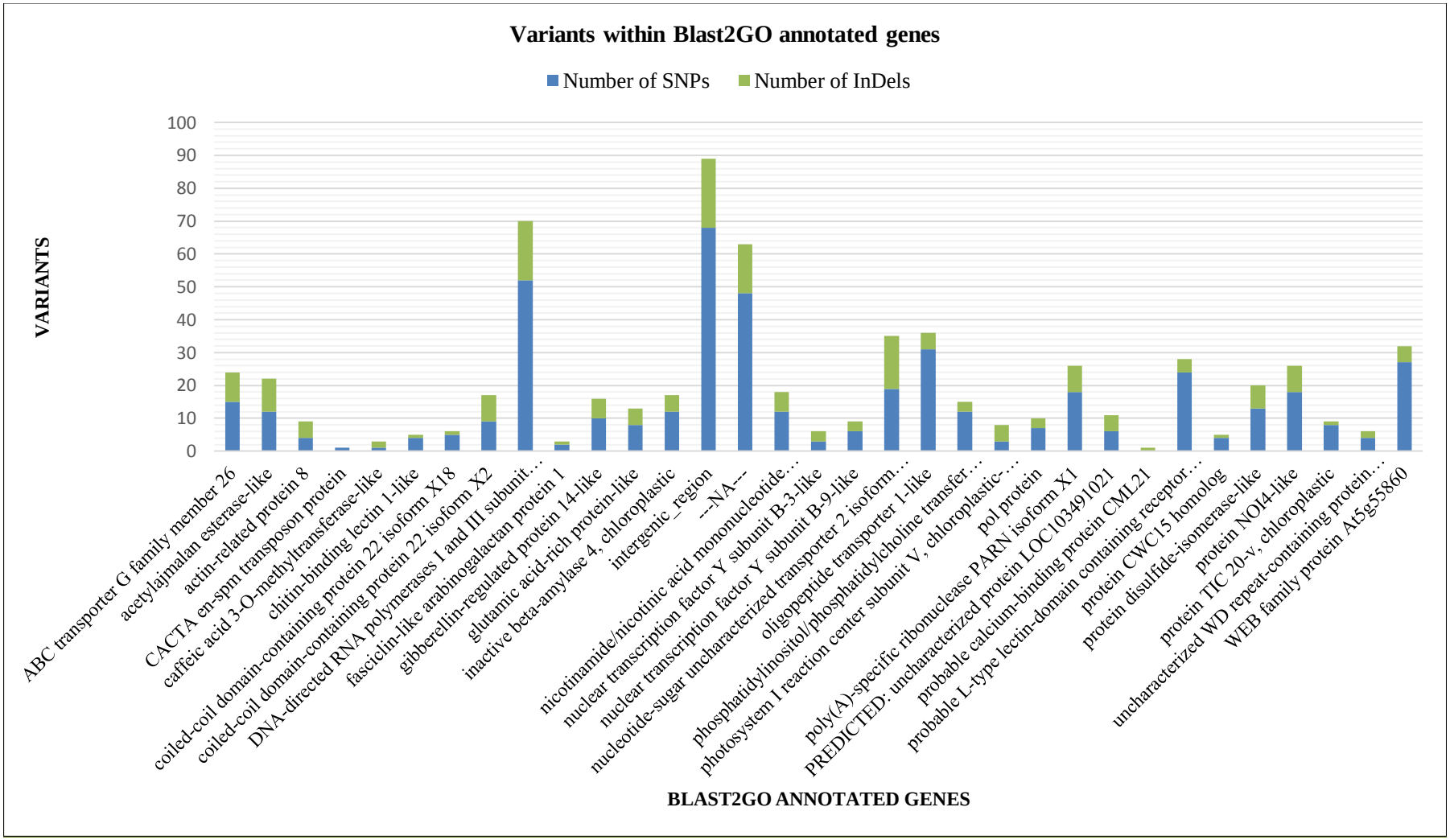


Fig. 4.20: Density of variants (SNPs and InDels) within the 28 annotated genes

Of the variants in the candidate region, the main emphasis was placed on the variants present in the six candidate genes identified in the Blast2GO step and nine highly significant SNPs found using Qtl-seq. It was noticed that most of the variants in the six candidate genes were of MODIFIER effect such as 3' UTR variant, 5' UTR variant, downstream variant and intron variant (Fig. 4.21). Within these six putative candidate genes, four genes had variants with LOW impact: PDI and Arabinogalactose (synonymous variant), ABCG26 (splice site variant) and GDSL-like protein (both synonymous variant and splice site variant) (Fig. 4.22). There was no HIGH impact variant within any of the six candidate genes but one MODERATE impact variant was reported in the PDI gene having a missense variant (Fig. 4.22) at position 36,921,314 bp. The variant alleles at this position between male-sterile bulk (MSB), male-fertile bulk (MFB) and male-sterile parent (MS-1) were then viewed in IGV genome browser with MS-1assembly as reference genome and it was observed that MS-1 and MSB had allele 'T' while the MSB had allele 'A' at the variant position (Fig. 4.23). This observation suggests that the missense mutation at 36,921,314 bp position in the PDI gene might be responsible for the male-sterility (*ms-1*) phenotype in the MS-1 genotype. Additionally, all the nine highly significant SNPs lying in the candidate region had MODIFIER effect and were located in the intergenic, upstream, and downstream region, except one SNP which was a synonymous variant having a LOW impact (Table 4.15). The positions of the five highly significacnt SNPs lying within the genes along with the four highly significant SNPs residing in the intergenic region were also viewed using IGV software (Fig. 24).

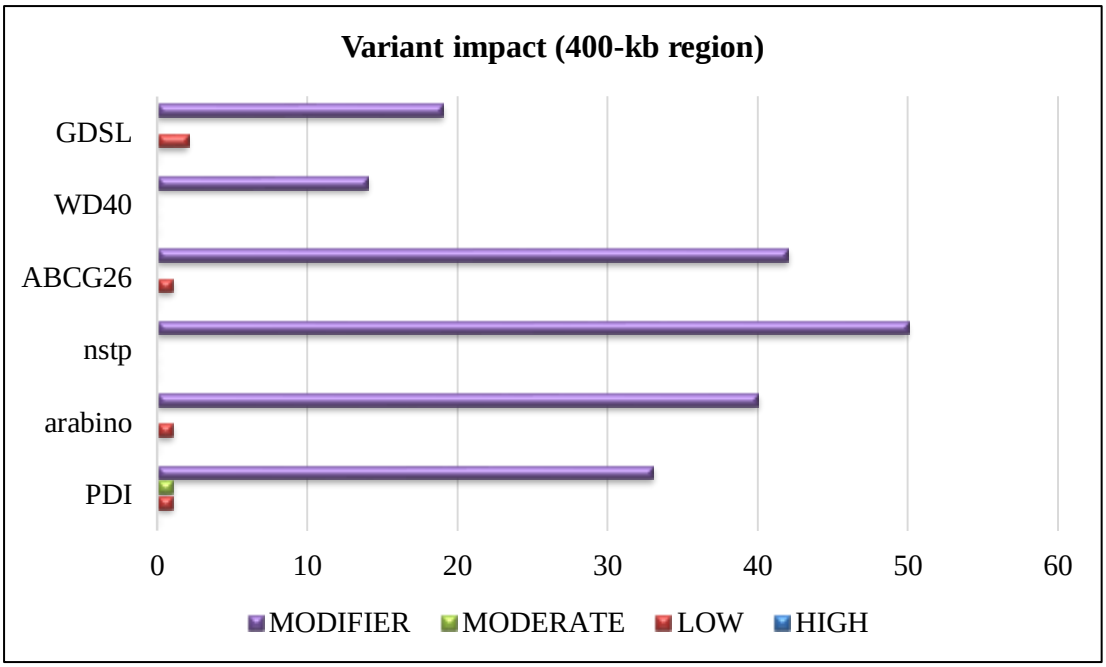


Fig. 4.21: Distribution of variant impacts in the 400-kb candidate region

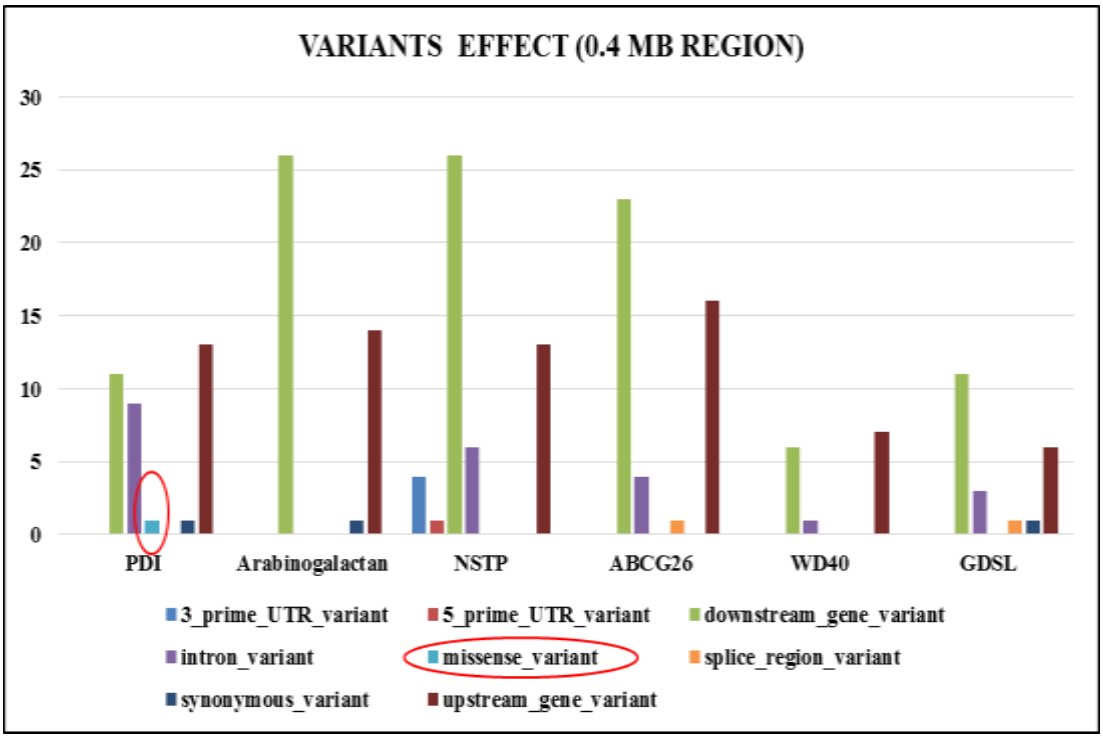


Fig. 4.22: The distribution of different variant effects caused by the variants present within the six candidate genes

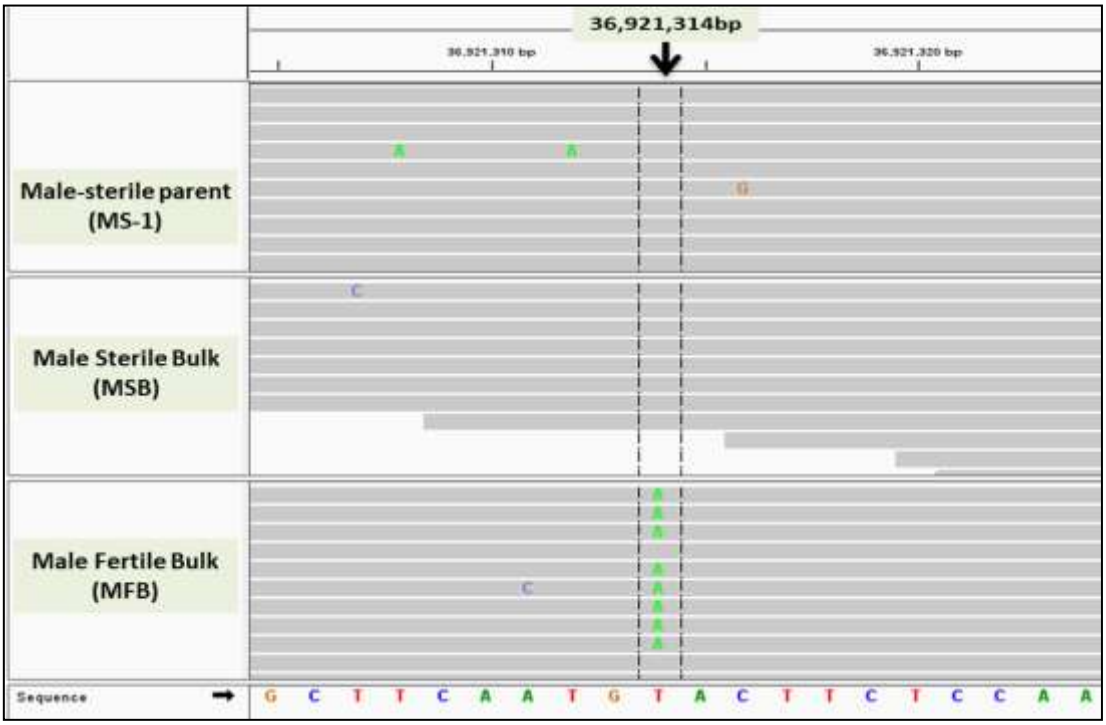
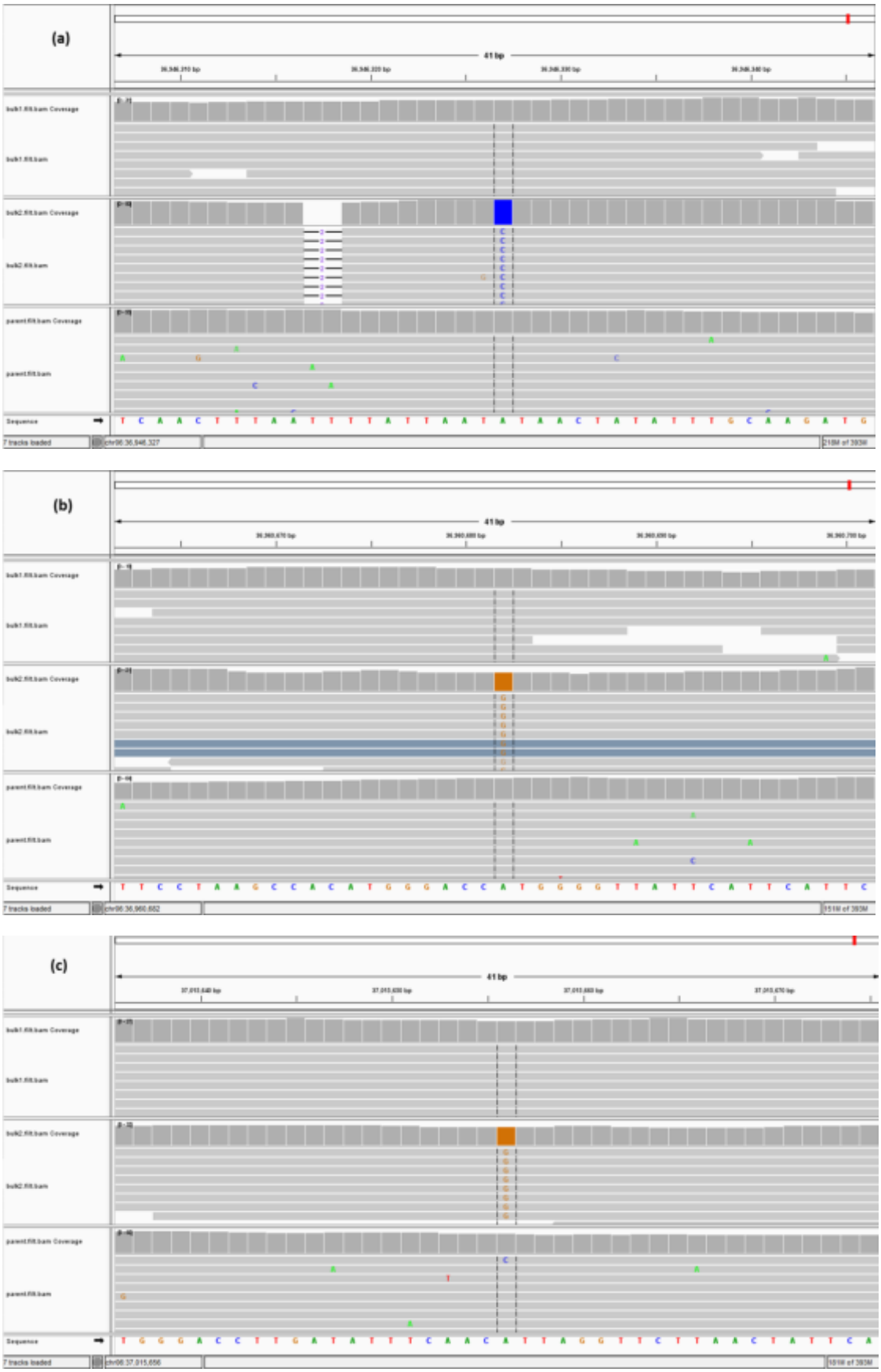


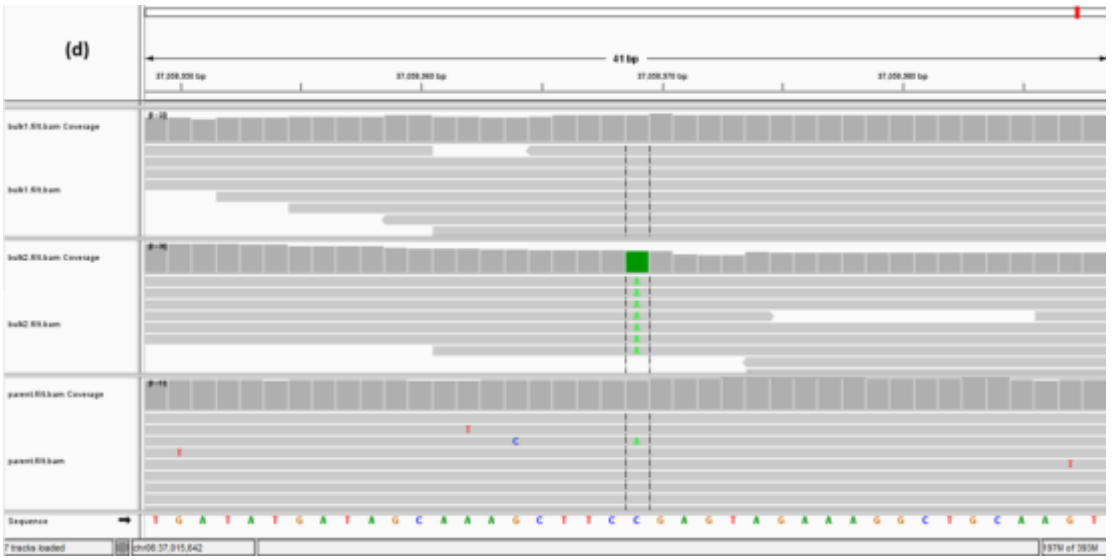
Fig. 4.23: IGV genome browser view of the missense SNP position in the PDI gene

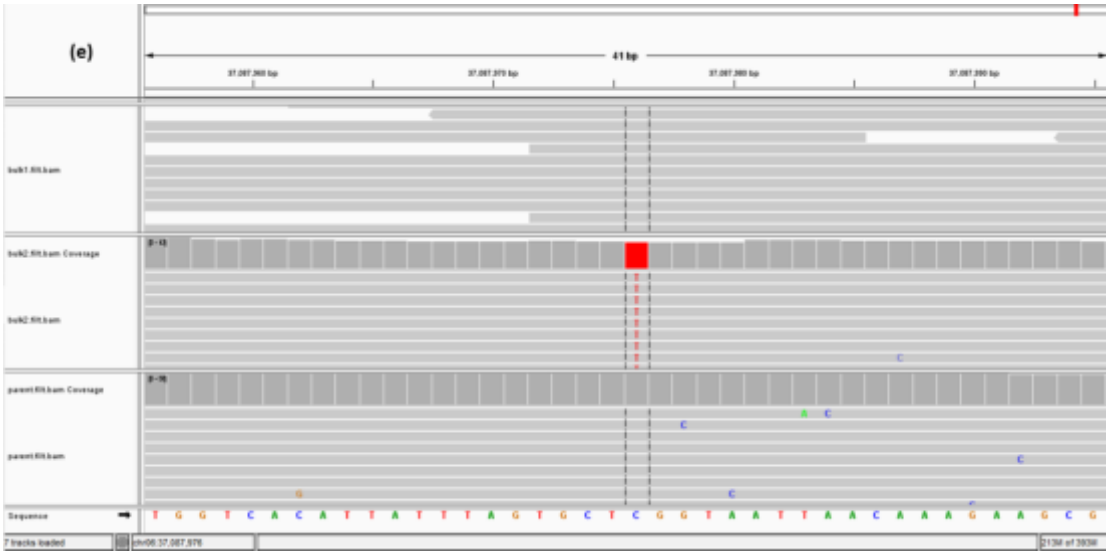
Table 4.15: Variant annotation and impact of nine highly significant SNPs on chromosome six

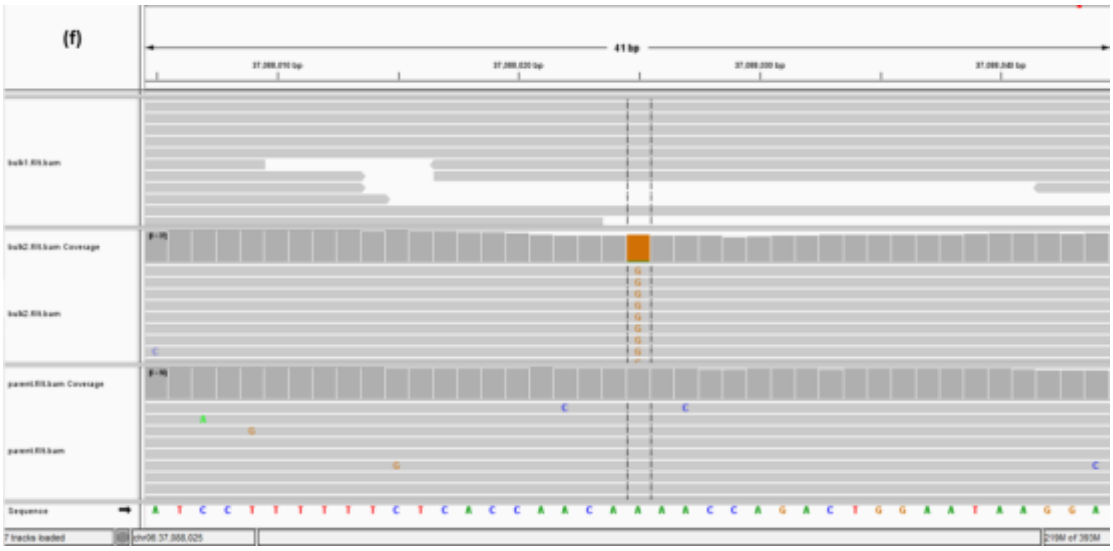
POSI	GENE ID	BLAST2GO ANNOTATION	REF ALLELE	ALT ALLELE	ANNOTATION	VARIANT IMPACT
36946327	MELO3C014103.2	nuclear transcription factor Y subunit B-3-like	A	C	upstream_gene_variant	MODIFIER
36960682	MELO3C014107.2	Fasciclin-like arabinogalactan protein	A	G	upstream_gene_variant	MODIFIER
37015656	MELO3C014111.2-MELO3C014112.2	-	A	G	intergenic_region	MODIFIER
37058969	MELO3C014115.2	E3 ubiquitin-protein ligase UPL1	C	A	synonymous_variant	LOW
37087976	MELO3C014117.2-MELO3C014119.2	-	C	T	intergenic_region	MODIFIER
37088025	MELO3C014117.2-MELO3C014119.2	-	A	G	intergenic_region	MODIFIER
37109800	MELO3C014121.2	ABC transporter G family member 26	A	G	upstream_gene_variant	MODIFIER
37115410	MELO3C014122.2	protein CWC15 homolog	G	C	downstream_gene_variant	MODIFIER
37168159	MELO3C031975.2-MELO3C014126.2	-	C	T	intergenic_region	MODIFIER



Contd…


(d)
41 bp
bulk1.filt.bam Coverage
bulk1.filt.bam
bulk2.filt.bam Coverage
bulk2.filt.bam
parent.filt.bam Coverage
parent.filt.bam
Sequence
7 tracks loaded


(e)
41 bp
bulk1.filt.bam
bulk2.filt.bam Coverage
bulk2.filt.bam
parent.filt.bam Coverage
parent.filt.bam
Sequence
7 tracks loaded


(f)
41 bp
bulk1.filt.bam
bulk2.filt.bam Coverage
bulk2.filt.bam
parent.filt.bam Coverage
parent.filt.bam
Sequence
7 tracks loaded

Contd…

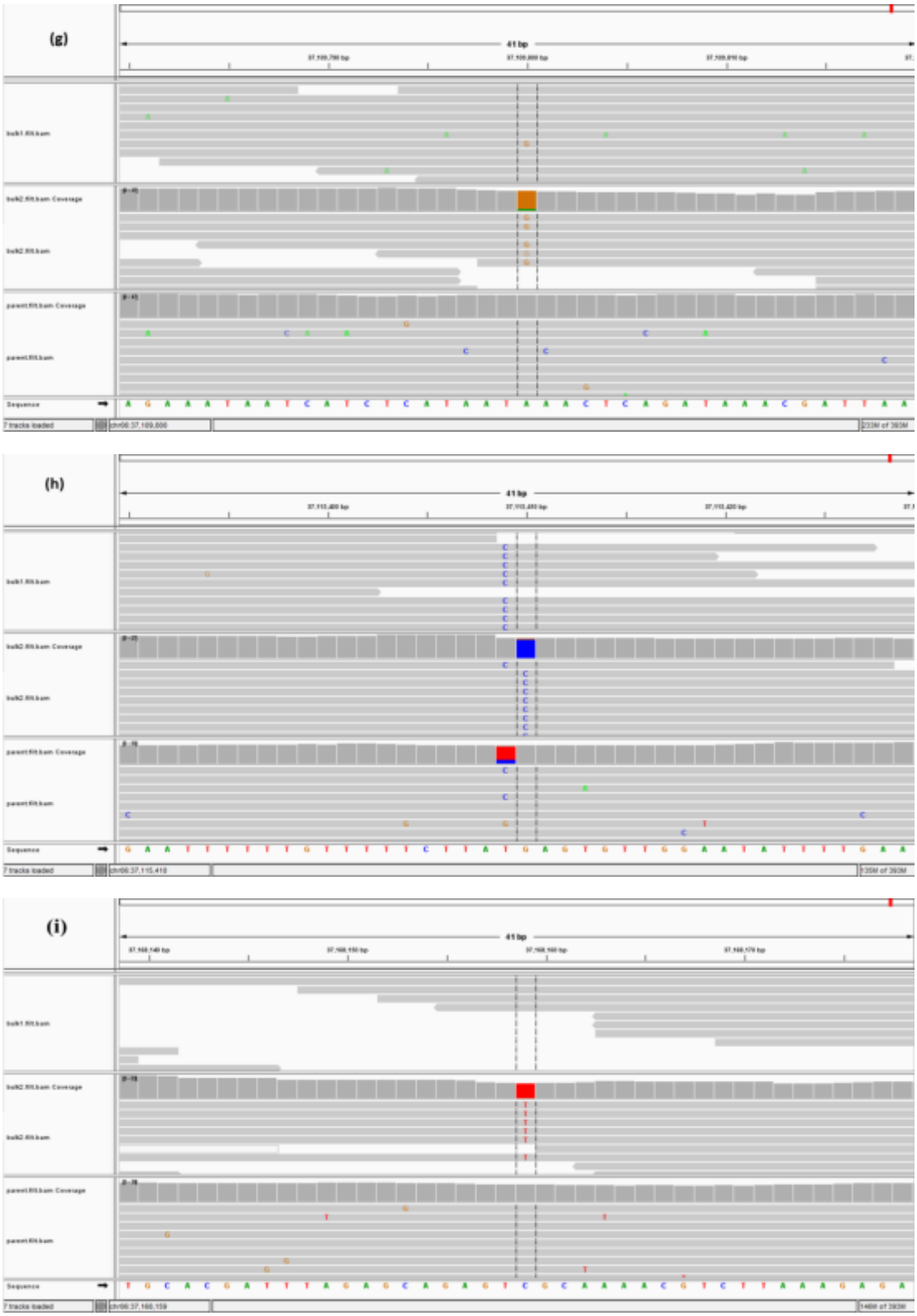


Fig. 4.24: Positions of highly significant SNPs within (a) gene id MELO3C014103.2, (b) gene id MELO3C014107.2 (c) intergenic region MELO3C014111.2 to MELO3C014112.2, (d) gene id MELO3C014115.2, (e) intergenic region MELO3C014117.2- MELO3C014119.2, (f) intergenic region MELO3C014117.2-MELO3C014119.2, (g) gene id MELO3C014121.2, (h) MELO3C014122.2, and (i) intergenic region MELO3C031975.2- MELO3C014126.2 as viewed under IGV software

Muskmelon being a summer season's crop is grown under high temperature conditions and therefore the microspores/pollens must have gained mechanisms to tolerate such increased temperatures. It can be theorized that the chaperon proteins like PDI might be involved in protecting the proteins from unfolding under higher temperatures and thereby allowing their transfer to the developing pollens via a secretory pathway. Thus any defect in the PDI gene, like the missense mutation detected in the current study, may lead to the abnormal development of the pollens.

The present study concludes that the *ms-1* governing gene lies in the 400-kb region on chromosome six which harbours six putative candidate genes and nine highly significant SNPs. Out of these nine highly significant SNPs, two lie within the predicted putative candidate genes; fasciclin-like arabinogalactan protein 1 (FLAs) and ABC transporter G family member 26. These highly significant SNPs can be used to develop KASP markers in order to improve the genetic map. Out of the six putative candidate genes, one gene protein disulfide isomerase (PDI) contains a missense mutation of nucleotide 'A' to 'T' at 36,921,314 bp position. On the basis of *in silico* assessment, the missense mutation in the PDI gene is likely the underlying cause for the male-sterile phenotype of *ms-1* gene. The validation of this result needs to be performed in wet-lab.

CHAPTER V

SUMMARY

Muskmelon (*Cucumis melo* L.) is a eudicot diploid crop (2n = 2x = 24), grown in temperate and tropical regions of the world. It belongs to *Cucurbitaceae* family and contains a genome size of about 454 Mb (Arumuganathan and Earle 1991). Ruggieri *et al* (2018) were able to assemble 417 Mb of the total melon genome, *in silico*. Muskmelon is well known for its fleshy fruits and musky aroma along with various nutritional and therapeutic advantages (Parle & Singh 2011). The world production of melons is around 27 million tons with China being the major producer (Anonymous 2019). In India melon is cultivated in Uttar Pradesh, Andhra Pradesh, Madhya Pradesh, and Punjab.

The fruits of muskmelon display a wide range of diversity, which have been exploited commercially for the production of superior and high-quality hybrids via heterosis breeding (Stepansky *et al* 1999 and Choudhary & Pandey 2010). Muskmelon is mainly an andromonecious plant (having both hermaphrodite and staminate flowers), thus requiring the removal of anthers (emasculation) a day prior to anthesis, followed by pollination after anthesis for hybrid production. Manual emasculation and pollination for producing F_1 hybrids is a costly venture and increases the chances of contamination through self-pollination. The availability of natural out-crossing systems like genic male-sterility (GMS), gynoecy and monoecy in muskmelon have found application in the cost-effective production of hybrids.

In muskmelon, five recessive and non-allelic nuclear male-sterile types have been reported (*ms-1* to *ms-5*). Out of these, only *ms-1* has been exploited commercially for the production of superior quality hybrids (Punjab Hybrid 1982, Punjab Anmol 2007, and MH-27 2018). The male-sterile plants of MS-1 genotype produce flowers with small and indehiscent anthers having empty pollen walls at tetrad stage (Bohn & Whitaker 1949). The markers linked to three male-sterile genes namely, *ms-3* (Park & Crossby 2004), *ms-5* (Sheng *et al* 2017), and *ms-1* (Singh *et al* 202) have been reported previously. Singh *et al* (2020) located the gene underlying *ms-1* on chromosome six between two SSR markers (DM0187 and DM0038) covering a genetic distance of 20.3 cM. The physical distance corresponding to these SSR markers is 3.2 Mb which is still a very large distance to consider these markers as tightly linked. Thus to reduce this distance the present investigation was conducted to fine map the *ms-1* gene using a whole-genome resequencing approach of male-sterile and male-fertile bulks.

In the present study, an F_2 mapping population from the cross of male-sterile MS- 1 and male-fertile KP_4HM-15 (MS-1 × KP_4HM-15) was used to develop two contrasting bulks namely, male-sterile bulk (MSB) and male-fertile bulk (MFB). A total of 152 F_2 plants were

screened to assess the genotype of these plants. It was found that F_2 population comprised of 112 male-fertile individuals (*MsMs*/*Msms*) and 40 male-sterile individuals (*msms*) which fits perfectly with the Mendelian ratio of 3:1 for the recessive gene, confirming the recessive nature of *ms-1*. The genotype for 112 male-fertile individuals were ascertained by the sowing the F_3 progenies (15-20 plants each) from these individuals. The genotype of 65 F_2 fertile plants was successfully determined with 39 individuals producing segregating progenies for *ms-1* (*Msms*) and the remaining 26 plants not segregating for *ms-1* gene (*MsMs*). The genotype of 37 individuals out of 112 total male-fertile plants could not be assessed due to lack of seed set and poor seed germination. Thus, the genotype of a total 105 F_2 plants was successfully determined.

Based on the genotype of these F_2 plants, two contrasting bulks namely, male-sterile bulk (MSB) and male-fertile bulk (MFB) were constructed by pooling the high-quality DNA from ten individuals each of homozygous male-sterile genotype (*msms*) and homozygous male-fertile genotype (*MsMs*), respectively. The DNA samples from the bulks (MSB and MFB) along with the parents (MS-1 and KP4HM-15) were sequenced using Illumina HiSeq 4000 platform yielding a total of 40 GB paired-end sequenced data. The sequencing coverage ranged from 38.85x to 52.74x. The raw sequence files were then subjected to quality check using FastQC followed by removal of low quality reads, sequences with disproportionate amount of bases, and the adapter sequences using Trimmomatic. The pre-processing of reads using Trimmomatic retained 79% to 97% of original reads which were of high quality and devoid of any adapter contamination. These clean reads were then further utilized to identify the genomic region associated with the *ms-1* gene.

SNP-index method (Abe *et al* 2012 and Takagi *et al* 2013) implemented within Qtl-seq pipeline (Sugihara *et al* 2020) was used to fine map the *ms-1* region. The clean reads of male-sterile parent MS-1 and two bulks (MSB and MFB) were used in this pipeline along with the MS-1 assembly as a reference. The MS-1 reference assembly was developed by calling the variants between MS-1 genotype and the muskmelon reference genome, DHL92 followed by the replacement of the detected SNPs in the MS-1 with those from the DHL92 reference genome. Qtl-seq yielded a total of 68.4 million variants out of which 3.1 million variants were used to calculate the SNP-index values for both the bulks followed by the calculation of ΔSNP-index values. The expected ΔSNP-index values were calculated by computer simulations at 99% and 95% level of significance to identify the corresponding confidence intervals. It was found that the 400 kb region on chromosome six ranging from 36.8 Mb to 37.2 Mb lied outside the confidence interval calculated at 99% level of significance indicating the presence of gene underlying *ms-1* within this region. This 400 kb

candidate region lied within the previously mapped 20.3 cM distance (3.2 Mb physical distance) delimited by two SSR markers, namely DM0187 and DM0038 (Singh *et al* 2020). This region also harboured nine highly-significant SNPs which can be used to construct KASP markers in order to further refine the genetic map for *ms-1* gene. Thus, the findings from the current study corroborated with the previous results and further resulted in narrowing down the region from 3.2 Mb to 400 kb candidate region on chromosome six.

Further screening of the genes within the 400 kb candidate interval revealed the presence of 43 genes out of which 28 genes were functionally annotated using Blast2GO (Götz *et al* 2008) and it was found that six of these total annotated genes were related directly or indirectly to the male-sterile phenotype in various crops and were designated as putative candidate genes. These six putative candidate genes were protein disulfide-isomerase-like, fasciclin-like arabinogalactan protein, nucleotide/sugar transporter family protein, ABC transporter G family member 26, transducin/WD40 repeat-like superfamily protein, and GDSL esterase/lipase At5g03980-like. The effect of variants present within these six genes was predicted by a variant annotation tool, snpEff (Cingolani *et al* 2012). It was observed that most of the genes had MODIFIER and LOW impact variants, which donot cause any significant changes in the genes or gene products. No gene had HIGH or MODERATE impact variant except, protein disulfide-isomerase (PDI) gene which harboured a missense variation (MODERATE impact) of 'A' to 'T' at 36,921,314 bp position which lies within the CDS (coding sequence) region of the PDI gene.

PDI gene encodes an endoplasmic reticulum (ER)-localized chaperone protein named protein disulfide-isomerase which catalyzes the folding of nascent or misfolded proteins during ER stress via UPR mediated stress response. It has been reported that pollens' tolerance to high temperatures might be due to the protein folding activities of these PDI proteins (Feldeverd *et al* 2020). The defect in this chaperone protein might result in misfolded proteins thereby halting the secretory pathways carrying the precursor molecules for the developing pollen. Hence, from the current studies it is concluded that the *ms-1* associated region lies in the 400 kb region on chromosome six which harbours a missense mutation of 'A' to 'T' in the PDI gene at 36,921,314 bp position. Thus, the misense mutation in the PDI gene might be responsible for the altered phenotype of pollens in the *ms-1* male sterile plants.

REFERENCES

Abe A, Kosugi S, Yoshida K, Natsume S, Takagi H, Kanzaki H and Terauchi R (2012) Genome sequencing reveals agronomically important loci in rice using MutMap. *Nat Biotechnol* **30**(2): 174.

An X, Dong Z, Tian Y, Xie K, Wu S, Zhu T and Wan X (2019) ZmMs30 encoding a novel GDSL lipase is essential for male fertility and valuable for hybrid breeding in maize. *Mol Plant* **12**(3): 343-359.

Andrews S (2010) FastQC: a quality control tool for high throughput sequence data.

Anonymous (2017-18) State-wise area and production of muskmelon in India. Available at http://apeda.in/agriexchange/India%20Production/India_Productions.aspx?cat=Vegetables&hscode=1078

Anonymous (2019) Production/Area for melons including cantaloupes. Available at http://www.fao.org/faostat/en/#data/QC

Anonymous (2020) *Package of practices for cultivation of vegetables*. Pp 3. Punjab Agricultural University, Ludhiana.

Arumuganathan K and Earle E D (1991) Nuclear DNA content of some important plant species. *Plant Mol Biol Rep* **9**(3): 208-18.

Bohn G W and Principe J A (1964) A second male-sterility gene in the muskmelon. *J Hered* **55**: 211-15.

Bohn G W and Whitaker T W (1949) A gene for male sterility in the muskmelon (*Cucumis melo* L.). *Proc Am Soc Hort Sci* **53**: 309-14.

Bolger A M, Lohse M and Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**(15): 2114-20.

Chang Z, Jin M, Yan W, Chen H, Qiu S, Fu S and Tang X (2018) The ATP-binding cassette (ABC) transporter OsABCG3 is essential for pollen development in rice. *Rice* **11**(1), 1-15.

Cheung A Y and Wu H M (1999) Arabinogalactan proteins in plant sexual reproduction. *Protoplasma* **208**(1): 87-98.

Choi H, Jin J Y, Choi S, Hwang J U, Kim Y Y, Suh M C and Lee Y (2011) An ABCG/WBC-type ABC transporter is essential for transport of sporopollenin precursors for exine formation in developing pollen. *Plant J* **65**(2): 181-93.

Choudhary B R and Pandey S (2010) Breeding of F1 hybrids in muskmelon: Accomplishment and prospects. *Indian J Arid Hort* **5**(1-2), 1-5.

Cingolani P, Platts A, Wang L L, Coon M, Nguyen T, Wang L and Ruden D M (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly* **6**(2): 80-92.

Coimbra S, Almeida J, Junqueira V, Costa M L and Pereira L G (2007) Arabinogalactan proteins as molecular markers in Arabidopsis thaliana sexual reproduction. *J Expt Bot* **58**(15-16): 4027-35.

Coimbra S, Costa M, Jones B, Mendes M A and Pereira L G (2009) Pollen grain development is compromised in Arabidopsis agp6 agp11 null mutants. *J Expt Bot* **60**(11): 3133-42.

Collard B C Y, Jahufer M Z Z, Brouwer J B and Pang E C K (2005) An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: The basic concepts. *Euphytica* **142**: 169-96.

Dai D, Xiong A, Yuan L, Sheng Y, Ji P, Jin Y and Luan F (2019) Transcriptome analysis of differentially expressed genes during anther development stages on male sterility and fertility in *Cucumis melo* L. line. *Gene* **707**: 65-77.

Dhillon N P S, Kumar J and Pitrat M (2008) Assessment of stability of expression of various male-sterile genes in muskmelon in sub-tropical field conditions. *Trop Pest Manage* **37**: 421-28.

Dou J, Zhao S, Lu X, He N, Zhang L, Ali A and Liu W (2018) Genetic mapping reveals a candidate gene (ClFS1) for fruit shape in watermelon (*Citrullus lanatus* L.). *Theoretical Appl Genetics* **131**(4): 947-58.

Doyle J J and Doyle J L (1987) A rapid DNA isolation procedure for small quantities of fresh leaf tissue (No. RESEARCH).

Fang Z, Sun P, Liu Y, Yang L, Wang X, Hou A and Bian C (1997) A male sterile line with dominant gene (Ms) in cabbage (*Brassica oleracea* var. *capitata*) and its utilization for hybrid seed production. *Euphytica* **97**(3): 265-68.

Feldeverd E, Porter B W, Yuen C Y, Iwai K, Carrillo R, Smith T and Christopher D A (2020) The arabidopsis protein disulfide isomerase subfamily M isoform, PDI9, localizes to the endoplasmic reticulum and influences pollen viability and proper formation of the pollen exine during heat stress. *Frontiers Plant Sci* **11**: 15-18.

Garcia-Mas J, Benjak A, Sanseverino W, Bourgeois M, Mir G, González V M, Henaff E, Camara F, Cozzuto L, Lowy E, Alioto T, Capella-Gutierrez S, Blanca J, Cañizares J, Ziarsolo P, Gonzalez-Ibeas D, Rodríguez-Moreno L, Droege M, Du L and Baulcombe D C (2012) The genome of melon (Cucumis melo L.). *P Natl Acad Sci USA* **109**: 11872-77.

Giovannoni J J, Wing R A, Ganal M W and Tanksley S D (1991) Isolation of molecular markers from specific chromosomal intervals using DNA pools from existing mapping populations. *Nucleic Acids Res* **19**(23): 6553-68.

Götz S, García-Gómez J M, Terol J, Williams T D, Nagaraj S H, Nueda M J and Conesa A (2008) High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res* **36**(10): 3420-35.

Han M J, Jung K H, Yi G, Lee D Y and An G (2006) Rice Immature pollen 1 (RIP1) is a regulator of late pollen development. *Plant Cell Physiol* **47**(11): 1457-72.

Hundal J S and Khurana D S (1993) CH-1-A new hybrid of chilli. *Progressive Farming* **29**: 11-13.

Huo Y, Pei Y, Tian Y, Zhang Z, Li K, Liu J and Liu J (2020) IRREGULAR POLLEN EXINE2 Encodes a GDSL Lipase Essential for Male Fertility in Maize. *Plant Physiol* **184**(3): 1438-54.

Jegadeesan S, Chaturvedi P, Ghatak A, Pressman E, Meir S, Faigenboim A and Firon N (2018) Proteomics of heat-stress and ethylene-mediated thermotolerance mechanisms in tomato pollen grains. *Frontiers Plant Sci* **9**: 1558.

Ju Y, Guo L, Cai Q, Ma F, Zhu Q Y and Zhang Q (2016) Arabidopsis JINGUBANG is a negative regulator of pollen germination that prevents pollination in moist environments. *Plant Cell* **28**(9): 2131-46.

Kaul M L H (1988) *Male Sterility in Higher Plants: Monographs on Theoretical and Applied Genetics*. Vol 10. Springer-Verlag, Berlin, Germany.

Kim Y J, Kim M H, Hong W J, Moon S, Kim E J, Silva J and Jung K H (2021) GORI, encoding the WD40 domain protein, is required for pollen tube germination and elongation in rice. *Plant J* **105**(6): 1645-64.

Kirkbride J H (1993) *Biosystematic monograph of the genus Cucumis (Cucurbitaceae): botanical identification of cucumbers and melons*. Parkway Publishers, Inc.

Kumar R, Janila P, Vishwakarma M K, Khan A W, Manohar S S, Gangurde S S and Varshney R K (2020) Whole-genome resequencing-based QTL-seq identified candidate genes and molecular markers for fresh seed dormancy in groundnut. *Plant Biotechnol J* **18**(4): 992-1003.

Kuromori T, Ito T, Sugimoto E and Shinozaki K (2011) Arabidopsis mutant of AtABCG26, an ABC transporter gene, is defective in pollen maturation. *J Plant Physiol* **168**(16): 2001-05.

Lal T, Vashisht V K and Dhillon N P S (2007) Punjab Anmol-a new hybrid of muskmelon (*Cucumis melo* L.). *J Res Punjab Agric Univ* **44**: 83.

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

Lecouviour M, Pitrat M and Risser G (1990) A fifth gene for male sterility in Cucumis melo. *Cucurbit Genet Coop Rep* **13**: 34-35.

Lester G E (2008) Antioxidant, sugar, mineral, and phytonutrient concentrations across edible fruit tissues of orange- fleshed Honeydew melon (*Cucumis melo* L.). *J Agr Food Chem* **56**: 3694-98.

Levitin B, Richter D, Markovich I and Zik M (2008) Arabinogalactan proteins 6 and 11 are required for stamen and pollen function in Arabidopsis. *Plant J* **56**(3): 351-63.

Li H (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**(21): 2987-93.

Li H and Durbin R (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**(14): 1754-60.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N and Durbin R (2009) The sequence alignment/map format and SAMtools. *Bioinformatics* **25**(16): 2078-79.

Li J, Yu M, Geng L L and Zhao J (2010) The fasciclin-like arabinogalactan protein gene, FLA3, is involved in microspore development of Arabidopsis. *Plant J* **64**(3), 482-97.

Liu G, Zhao T, You X, Jiang J, Li J and Xu X (2019) Molecular mapping of the Cf-10 gene by combining SNP/InDel-index and linkage analysis in tomato (*Solanum lycopersicum*). *BMC Plant Biology* **19**(1): 1-11.

Liu S, Wang P, Liu Y and Wang P (2020) Identification of candidate gene for resistance to broomrape (*Orobanche cumana*) in sunflower by BSA-seq. *Oil Crop Sci* **5**(2): 47-51.

Lozanov P (1983) Selekcija na mazkosterilni roditelski komponenti za ulesnjavana na proizvodstvoto na hibridni semena ot papesi. Dokl. na parva naucna konferncija po genetika i selekapa Razgrad.

Lu D P and Christopher D A (2008) Endoplasmic reticulum stress activates the expression of a sub-group of protein disulfide isomerase genes and AtbZIP60 modulates the response in Arabidopsis thaliana. *Mol Genet Genomics* **280**(3): 199-210.

Magwene P M, Willis J H and Kelly J K (2011) The statistics of bulk segregant analysis using next generation sequencing. *PLoS Comput Biol* **7**(11): e1002255.

Mallek-Ayadi S, Bahloul N and Kechaou N (2018) Chemical composition and bioactive compounds of *Cucumis melo* L. seeds: Potential source for new trends of plant oils. *Process Safety Environ Prot* **113**: 68-77.

MasonA S (2015) SSR genotyping. Plant Genotyping. pp 77-89, Humana Press, New York, NY.

McCreight J D (1980) Male Sterile-1: Stability of Expression. *Cucurbit Genet Coop Rep* **3**: 31.

McCreight J D and Elmstrom G W (1984) A third muskmelon malesterile gene. Hort *Sci* **19**: 268-70.

McCreight J D and Elmstrom G W 1983a. A third male sterile gene in muskmelon. *Cucurbit Genet Coop Rep* **6**: 46.

McCreight J D and Elmstrom G W 1983b. Male sterile-Leesburg: The third muskmelon male sterile gene. *Hort Sci* **18**: 564.

Michelmore R W, Paran I and Kesseli R V (1991) Identification of markers linked to disease-resistance genes by bulked segregant analysis: a rapid method to detect markers in specific genomic regions by using segregating populations. *Proc Nat Acad Sci* **88**(21): 9828-32.

Migliori V, Mapelli M and Guccione E (2012) On WD40 proteins: propelling our knowledge of transcriptional control? *Epigenetics* **7**(8): 815-22.

Munshi A D and Alvarez J M (2005) Hybrid melon development. *J New Seeds* **6**: 321-60.

Nadeem M A, Nawaz M A, Shahid M Q, Doğan Y, Comertpay G, Yıldız M, Hatipoğlu R, Ahmad F, Alsaleh A, Labhane N, Özkan H, Chung G and Baloch F S (2018) DNA molecular markers in plant breeding: current status and recent advancements in genomic selection and genome editing. *Biotechnol Biotec Eq* **32**: 261-85.

Nagaraju M M, Thomson T, Rao G K and Siva M (2017) Role of Male Sterility in Vegetable Hybrid Seed Production. *Int J Curr Microbiol App Sci* **6**: 134-41.

Nandapuri K S, Singh S and Lal T (1982) 'Punjab Hybrid' a variety of muskmelon. *Prog Farming* **18**: 3-4.

Park S O, Crosby K M, Huang R and Mirkov T E (2004) Identification and confirmation of RAPD and SCAR markers linked to the ms-3 Gene controlling male sterility in melon (*Cucumis melo* L.). *J Am Soc Hortic Sci* **129**: 819-25.

Parle M and Singh K (2011) Musk Melon is eat-Must Melon. *Int Res J Pharmacy* **2**(8): 52-57

Paterson, A. H, 1996. Making genetic maps. In: Paterson A H (Ed.) *Genome Mapping in Plants*, pp. 23-39. R. G. Landescompany, San Diego, California; Academic Press, Austin, Texas

Pathak C S (2000) Hybrid seed production in onion. *J New Seeds* **1**(3-4): 89-108.

Paux E, Faure S, Choulet F, Roger D, Gauthier V, Martinant J P, Sourdille P, Balfourier F, Le Paslier M C, Chauveau A, Cakir M, Gandon B and Feuillet C (2010) Insertion site-based polymorphism markers open new perspectives for genome saturation and marker-assisted selection in wheat. *Plant Biotechnol J* **8**: 196-210.

Pitrat M (1991) Linkage groups in *Cucumis melo* L. *J Hered* **82**: 406-11.

Pitrat M (2002) 2002 gene list for melon. *Cucurbit Genet Coop Rep* **25**: 76-93.

Pitrat M (2008) Melon. In: Prohens-Tomás J and Nuez F (Eds.) *Vegetables I: Asteraceae, Brassicaceae, Chenopodicaceae, and Cucurbitaceae.* Pp. 283-315. Springer, New York.

Pitrat M (2008) Melon. In: *Vegetables I*, pp. 283-315. Springer, New York, NY.

Pujol M, Alexiou K G, Fontaine A S, Mayor P, Miras M, Jahrmann T and Aranda M A (2019) Mapping cucumber vein yellowing virus resistance in cucumber (*Cucumis sativus* L.) by using BSA-seq analysis. *Frontiers Plant Sci* **10**: 1583.

Qiao A, Fang X, Liu S, Liu H, Gao P and Luan F (2021) QTL-seq identifies major quantitative trait loci of stigma color in melon. *Hort Plant J* **15**: 318-21.

Rao M K, Devi K U and Arundhati A (1990) Applications of genie male sterility in plant breeding. *Plant Breeding* **105**(1): 1-25.

Reyes F, León G, Donoso M, Brandizzí F, Weber A P and Orellana A (2010) The nucleotide sugar transporters AtUTr1 and AtUTr3 are required for the incorporation of UDP-glucose into the endoplasmic reticulum, are essential for pollen development and are needed for embryo sac progress in Arabidopsis thaliana. *Plant J* **61**(3): 423-35.

Ruggieri V, Alexiou K G, Morata J, Argyris J, Pujol M, Yano R and Garcia-Mas J (2018). An improved assembly and annotation of the melon (*Cucumis melo* L.) reference genome. *Scientific Rep* **8**(1), 1-9.

Sageret A (1826) (Considerations on the production of hybrids, variants and varieties in general and those of the Cucurbitaceae family in particular). *Ann Sci Nat Bot Biol* **8**: 294-314.

Sawhney V K (2004) Photoperiod-sensitive male-sterile mutant in tomato and its potential use in hybrid seed production. *J Hort Sci Biotechnol* **79**(1): 138-41.

Sax K (1923) The association of size differences with seed-coat pattern and pigmentation in *Phaseolus vulgaris*. *Genetics* **8**(6): 552.

Sheng Y, Wang Y, Jiao S, Jin Y, Ji P and Luan F (2017) Mapping and preliminary analysis of ABORTED MICROSPORES (AMS) as the candidate gene underlying the male sterility (ms-5) mutant in melon (*Cucumis melo* L.). *Front Plant Sci* **8**: 902-16.

Shifriss C (1997) Male sterility in pepper (*Capsicum annuum* L.). *Euphytica* **93**(1): 83-88.

Singh B D and Singh A K (2015) *Marker-assisted plant breeding: principles and practices.* New Delhi: Springer India.

Singh M, Sharma S P, Sarao N K, Kaur S and Chhuneja P (2020) Molecular mapping of nuclear male-sterility gene ms-1 in muskmelon (*Cucumis melo* L.). *J Hort Sci Biotechnol* **95**(2): 162-68.

Singh V K, Khan A W, Saxena R K, Kumar V, Kale S M, Sinha P and Varshney R K (2016) Next-generation sequencing for identification of candidate genes for Fusarium wilt and sterility mosaic disease in pigeonpea (*Cajanus cajan*). *Plant Biotechnol J* **14**(5): 1183-94.

Song J, Li Z, Liu Z, Guo Y and Qiu L J (2017) Next-generation sequencing from bulked-segregant analysis accelerates the simultaneous identification of two qualitative genes in soybean. *Frontiers Plant Sci* **8**: 919.

Stepansky A, Kovalski I and Perl-Treves R (1999) Intraspecific classification of melons (*Cucumis melo* L.) in view of their phenotypic and molecular variation. *Plant Syst Evol* **217**: 313-32.

Sugihara Y, Young L, Yaegashi H, Natsume S, Shea D J, Takagi H and Abe A (2020) High-performance pipeline for MutMap and QTL-seq. *bioRxiv*.

Takagi H, Abe A, Yoshida K, Kosugi S, Natsume S, Mitsuoka C and Terauchi R (2013) QTL-seq: rapid mapping of quantitative trait loci in rice by whole genome resequencing of DNA from two bulked populations. *Plant J* **74**(1): 174-83.

Thoday J M (1961) Location of polygenes. *Nature* **191**(4786): 368-70.

Vashisht V K, Lal T, Sharma S P and Thind T S (2015). KP4HM-15: an inbred line of muskmelon resistant to Fusarium wilt. In: *III International Symposium on Underutilized Plant Species 1241*, pp. 481-86.

Wang P, Xu X, Zhao G, He Y, Hou C, Kong W and Xu Z (2020) Genetic mapping and candidate gene analysis for melon resistance to *Phytophthora capsici*. *Scientific Rep* **10**(1): 1-11.

Wang X, Chen J, Liu C, Luo J, Yan X, Ai A and Peng X (2019) Over-expression of a protein disulfide isomerase gene from *Methanothermobacter thermautotrophicus*, enhances heat stress tolerance in rice. *Gene* **684**: 124-30.

Whitaker T W and Davis D W (1962) *Cucurbits: botany, cultivation and utilization*. pp 162-74. Interscience, New York.

Yang Z, Huang D, Tang W, Zheng Y, Liang K, Cutler A J and Wu W (2013) Mapping of quantitative trait loci underlying cold tolerance in rice seedlings via high-throughput sequencing of pooled extremes. *Plos one* **8**(7): e68433.

Yuan G, Zou T, Zhang X, Liu M, Luo T, He Z and Li S (2020) A rice GDSL esterase/lipase protein (GELP) is required for anther and pollen development. *Mol Breeding* **40**(9): 1-15.

Zhang H, Wang M, Li Y, Yan W, Chang Z, Ni H and Tang X (2020) GDSL esterase/lipases OsGELP34 and OsGELP110/OsGELP115 are essential for rice pollen development. *J Int Plant Biol* **62**(10): 1574-93.

Zhang J and Panthee D R (2020) PyBSASeq: a simple and effective algorithm for bulked segregant analysis with whole-genome sequencing data. *BMC Bioinformatics* **21**(1): 1-10.

Zhang W (2020) A GDSL lipase is required for anther and pollen development. *Plant Physiol* **15**: 1810-11.

Zhang X, Wang G, Chen B, Du H, Zhang F, Zhang H and Geng S (2018) Candidate genes for first flower node identified in pepper using combined SLAF-seq and BSA. *PLoS One* **13**(3): e0194071.

Zhao J, Long T, Wang Y, Tong X, Tang J, Li J and Zhang J (2020) RMS2 encoding a GDSL lipase mediates lipid homeostasis in anthers to determine rice male fertility. *Plant Physiol* **182**(4): 2047-64.

VITA

Name of the Student	:	Simranjot Kaur
Father's name	:	Hardeep Singh
Mother' name	:	Baljinderjeet Kaur
Nationality	:	Indian
Date of Birth	:	27th October, 1996
Permanent home Address	:	H. No. 260 Ranchi Colony, Tharike road, Ludhiana

EDUCATIONAL QUALIFICATION

Bachelor degree	:	B.Sc. Biotechnology (Hons.)
University	:	Punjab Agricultural University, Ludhiana, India
Year of Award	:	2018
OCPA	:	7.77/10.00
Master's degree	:	M.Sc. Biotechnology
University	:	Punjab Agricultural University, Ludhiana, India
OCPA	:	7.81/10.00
Title of Master's thesis	:	Fine mapping of male-sterility *ms-1* gene using whole genome resequencing of muskmelon (*Cucumis melo* L.)
Award/Distinction/Fellowships/ Scholarships	:	-