UNRAVELING SNP VARIATION IN SILKWORM GENOTYPES THROUGH GENOTYPING BY SEQUENCING

Lakshmi Bharathi S[1], S Manthira Moorthy[2], K.P.Arunkumar[2], Kusuma L[2], R Sumathy[2], S Pooja[1], Vidya Niranjan[1*]

Author-1 S Lakshmi Bharathi, Department of Biotechnology, RV College of Engineering, Bangalore, Karnataka, India.

Author-2 Manthira Moorthy, Department of Silkworm breeding and molecular biology, Central Sericulture Research and Training Institute, Mysore, Karnataka, India.

Author-3 KP Arunkumar, Department of Molecular Genetics, Central Sericulture Research and Training Institute, Mysore, Karnataka, India.

Author-4 Kusuma L, Department of Silkworm breeding and molecular biology, Central Sericulture Research and Training Institute, Mysore, Karnataka, India.

Author-5 R Sumathy, Department of Bioinformatic Centre, Central Sericulture Research and Training Institute, Mysore, Karnataka, India.

Author-6 S Pooja, Department of Biotechnology, RV College of Engineering, Bangalore, Karnataka, India.

Author 7 Vidya Niranjan Professor and Head of Department of Biotechnology, RV College of Engineering, Bangalore, Karnataka, India*

Corresponding Author- Vidya Niranjan Professor and Head of Department of Biotechnology, RV College of Engineering, Bangalore, Karnataka, India*

ABSTRACT

Silkworm (*Bombyx mori*) is a highly significant insect in the economy, renowned for its production of high-quality silk. Among the various silk types, Mulberry and Eri varieties stand out as the major exports. Silk-based products, including Silk Fabrics, Silk Garments, Silk Waste, Silk Carpets, and Natural Silk Yarn, contribute significantly to international markets, driving the need for increased silk yield and productivity.

To advance silk production, we collected over 100 diverse silkworm instar larvae for RAD Sequencing, employing high-throughput next-generation sequencing to analyze genetic loci within the genome. Through our study, we successfully identified the SNP and QTL of silkworms, laying the foundation for future investigations into the relationship between SNP positions and phenotypes. Such knowledge will empower sericulturists in selecting the most suitable breeds and pave the way for targeted genetic modifications aimed at enhancing silk yield and quality. This research holds the promise of elevating the sericulture industry to new heights of success.

Keywords: *Bombyx mori,* RAD Sequencing, Single Nucleotide Polymorphism (SNP), Quantitative Trait Locus (QTL)

1.0 INTRODUCTION

Sericulture, centered around the silkworm (Bombyx mori), plays a pivotal role in India, the second-largest silk producer globally, with an estimated annual output of around 349 metric tonnes. Several states, including Andhra Pradesh, Assam, Bihar, Gujarat, Jammu & Kashmir, Karnataka, Chhattisgarh, Maharashtra, Tamil Nadu, Uttar Pradesh, and West Bengal, significantly contribute to the nation's silk production.(1)

Silk production encompasses four main subtypes: Mulberry, Tasar, Eri, and Muga. According to the Central Silk Board of India, Mulberry and Eri silk types dominate the country's silk exports. Prominent among the exported silk products are Silk Fabrics, Silk Garments, Silk Waste, Silk Carpets, and Natural Silk Yarn, which witness high demand in international markets.(2)

The aim objective of the study was to collect the three different instar cycles of silkworm larvae for SNP and QTL mapping. Over 100 breeds of silkworm were collected and a high-throughput sequencing technique was followed for phenotype detection.

RAD Seq is a cutting-edge, high-throughput next-generation sequencing (NGS) technique utilized for the comprehensive analysis of numerous genetic loci spanning the genome.(3) Remarkably, it employs selective targeting of specific regions of interest within the genome, enabling the simultaneous examination of thousands of these loci. The process commences by fragmenting genomic DNA through the precise action of restriction enzymes, which cleave at specific recognition sites, yielding short DNA fragments with overhangs. These fragments are then skilfully coupled with adapters, adeptly facilitating the amplification of regions adjacent to the restriction sites. Subsequently, the power of next-generation sequencing (NGS) is harnessed to decode the sequences of these fragments. Consequently, this technique empowers the discerning comparison of genetic variation among individuals or populations, precisely scrutinizing the identified genetic loci (4,5).

The RAD-Seq technology was employed to map the regions of interest related to yield in both domestic Xiafang (D_XF) and wild silkworms (W_AK) genomes. For this purpose, a total of 100 BC1 individuals were subjected to RAD-Seq sequencing, resulting in an average of 2,230,620 RAD tags per individual. The number of RAD tags varied between 720,477 and 4,622,071 across the sequenced individuals. Subsequently, a linkage map was constructed using parental mapping information from W_AK and D_XF strains. Through this analysis, 11 Quantitative Trait Loci (QTLs) associated with pupal weight (PW), cocoon shell ratio (CSR), whole cocoon weight (WCW), and cocoon shell weight (CSW) characteristics were identified. These QTLs were found to be located on 7 chromosomes. (6)

The sequencing by synthesis (SBS) method enables instruments to collect data in sync with enzymatic synthesis by adding nucleotides using various enzymes and detection methods. The limitation of SBS lies in the increasing noise during successive incorporation and imaging cycles, which impacts the length of sequence readings. While Sanger reads still surpass SBS reads in length, SBS-based sequencing systems employ either direct fluorescence detection or indirect sensing through nucleotide incorporation products.(7)

Mutations lead to polymorphism, with different types identified based on the mutation's nature. Single base mutations cause the most basic form, known as "Single Nucleotide Polymorphisms" (SNPs). SNPs are abundant throughout the genome, offering opportunities to discover new genes related to traits or diseases. Currently used in whole genome linkage studies, SNPs are expected to become primary markers for exploring population evolutionary history due to their widespread presence, variation, and easy screening potential.(8,9)

Quantitative Trait Loci (QTL) mapping in silkworms holds immense significance for the sericulture industry. As a crucial economic insect producing high-quality silk, understanding the genetic basis of important traits like pupal weight, cocoon shell ratio, and cocoon weight is essential for enhancing silk production and quality. QTL mapping allows researchers to identify specific genomic regions linked to these traits, providing valuable insights into their genetic control.(10)

By pinpointing the genes and markers associated with desirable traits, breeders can selectively choose silkworm strains with favorable QTLs, leading to improved yields and quality. Moreover, QTL mapping facilitates targeted genetic modifications, enabling the development of superior silkworm breeds suited for specific environmental conditions or market demands. This powerful tool not only boosts silk productivity but also opens doors to innovative advancements in sericulture, contributing to the sustainable growth of the silk industry.(11)

This advancement in understanding genetic variations and loci in sericulture could have valuable implications for improving silk production and quality in the future.

2.0 RESEARCH METHODOLOGY

2.1 DNA ISOLATION AND QC

Genomic DNA isolation was performed on late instar larvae or pupa from 100 selected silkworm breeds. The CTAB technique was employed to isolate the DNA, and its purity was carefully assessed. Subsequently, each of the 100 samples underwent analysis. The total DNA content of each sample was measured and evaluated using Agarose Gel Electrophoresis and nano-drop readings.

2.2 LIBRARY PREPARATION FOR SEQUENCING

Genomic DNA samples from all sources were subjected to in-silico evaluation to determine the appropriate restriction enzymes-1 for digestion. After digestion, fragments were ligated with barcoded adapters possessing compatible sticky ends with the primary digestion enzymes and illumina P5/P7 sequences. Multiple PCR amplifications were conducted, followed by pooling and size selection of the samples to construct the GBS (Genotyping by Sequencing) library.

For optimization of enzyme sets and fragment size, the genomic assembly of species with fully sequenced genomes and closely related species was subjected to in-silico digestion analysis.

This analysis considered data production, genome coverage, evenness, reduction of repeated regions, and enzymatic features. The integration of experimental digestion assays with computational methods allowed for dependable and repeatable findings across a range of species.

For restriction enzyme digestion, 0.3-0.6µg DNA was fully digested using an optimized set of restriction enzymes to achieve the desired marker density. Following digestion, both ends of the fragments were ligated with P1&P2 barcoded adapters, respectively. PCR enrichment was carried out for tags containing both P1 & P2 adapters, and DNA fragments from different samples were pooled. The desired fragments of DNA were recovered through gel electrophoresis. High-throughput DNA sequencing was performed using illumina technology. The qualified DNA libraries were pooled based on effective concentration and expected data production. The paired-end sequencing was carried out with a read length of 144bp on each end. (Figure-1)

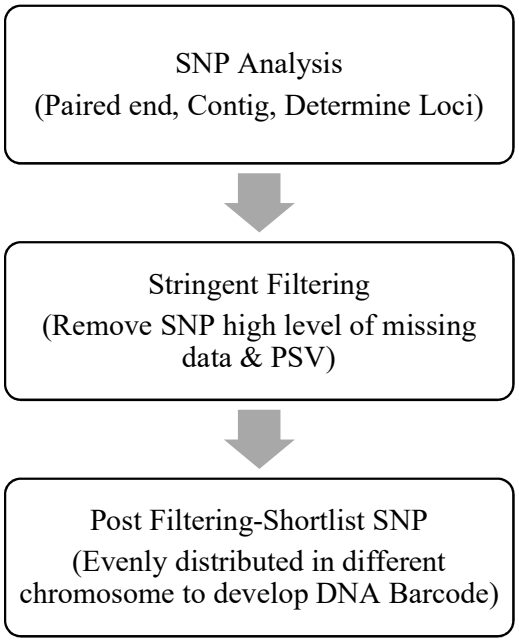


Figure-1: Flowchart for filtering SNPs

3.0 RESULTS AND DISCUSSION

3.1 DNA ISOLATION AND QC

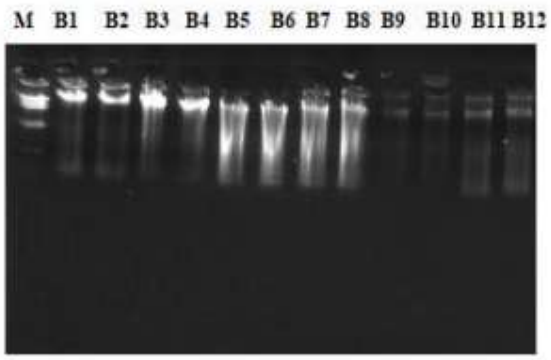
Late instar larvae or pupa from all 100 silkworm breeds underwent genomic DNA isolation using the CTAB method. The resulting DNA samples were quantified and assessed for quality using Agarose Gel Electrophoresis and nano-drop reading. Out of the 100 samples, 46 successfully passed the quality control (QC) evaluation, while 54 samples did not meet the QC criteria. (Table-1, Figure-2)

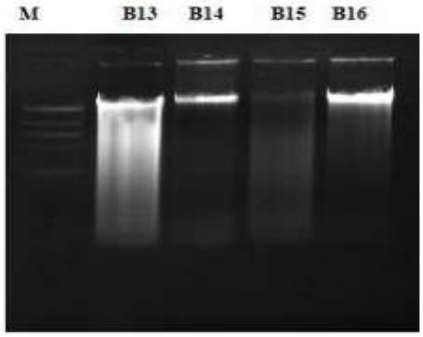
Table-1: DNA samples and their respective QC

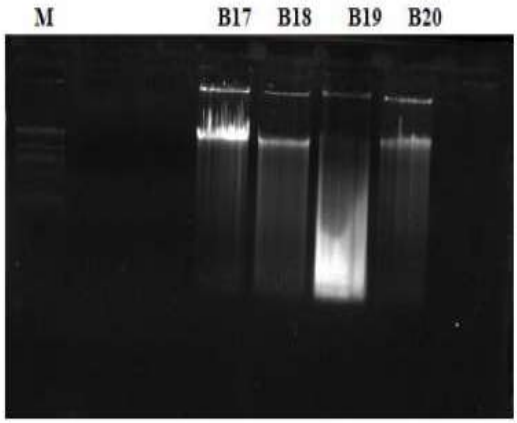
Sno	Sample ID	Nanodrop concentration (ng/ml)	260/280	260/230	Volume	Quantity (µg/ml)	Remarks
1	B1	1697.8	1.87	0.63	50	84.89	QC pass
2	B2	1973.5	1.98	0.42	50	98.675	QC pass
3	B3	182.3	1.84	0.22	50	9.115	QC pass
4	B4	814.4	1.88	0.41	50	40.72	QC pass
5	B5	227.5	2.02	2.3	50	11.375	QC pass
6	B6	25.6	1.84	1.23	50	1.28	QC pass
7	B7	3779	1.93	2.29	30	113.37	QC pass
8	B8	1357.6	1.96	2.45	30	40.728	QC pass
9	B9	435.2	1.98	2.27	30	13.056	QC pass
10	B10	143.2	1.99	2.29	50	7.16	QC pass
11	B11	1493	1.84	0.79	50	74.65	QC pass
12	B12	493.3	1.76	0.82	50	24.665	QC pass
13	B13	305.5	1.92	0.6	50	15.275	QC pass
14	B14	529.6	1.99	0.76	50	26.48	QC pass
15	B15	133.5	2.03	0.79	50	6.675	QC pass
16	B16	1041.2	1.67	0.89	50	52.06	QC pass
17	B17	467.3	1.87	1.23	30	14.019	QC pass
18	B18	560.5	1.75	0.98	30	16.815	QC pass
19	B19	234.5	1.97	0.67	30	7.035	QC pass
20	B20	328.4	1.66	1.86	30	9.852	QC pass
21	B21	225.6	1.98	0.26	60	13.536	QC fail
22	B22	1362.4	1.98	0.74	60	81.744	QC fail
23	B23	1562.4	1.92	0.76	60	93.744	QC fail
24	B24	752.6	1.89	0.45	60	45.156	QC fail
25	B25	55.1	1.82	1.1	60	3.306	QC fail
26	B26	1454.4	1.87	0.79	60	87.264	QC fail
27	B27	1255.2	1.99	0.69	60	75.312	QC fail
28	B28	400.3	1.88	0.43	60	24.018	QC fail
29	B29	123.4	2.01	0.76	60	7.404	QC fail
30	B30	303.2	1.77	0.98	60	18.192	QC fail
31	B31	1272.9	1.98	0.71	60	76.374	QC fail
32	B32	236.6	1.09	0.29	60	14.196	QC fail
33	B33	114.4	0.7	0.13	60	6.864	QC fail
34	B34	45	1.43	0.34	60	2.7	QC fail
35	B35	1568.4	1.68	0.79	60	94.104	QC fail
36	B36	828.9	1.09	0.6	60	49.734	QC fail
37	B37	1234.1	1.64	0.75	60	74.046	QC fail
38	B38	831.1	1.53	0.55	60	49.866	QC fail
39	B39	1946.7	1.62	0.72	60	116.802	QC fail

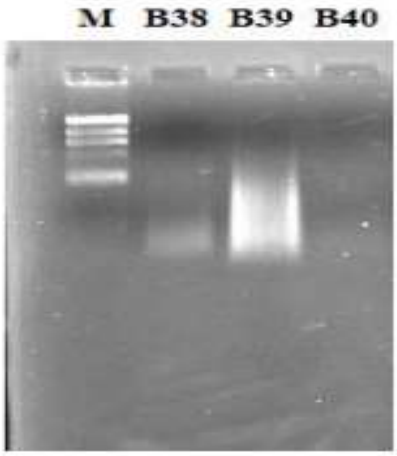
40	B40	1192.7	1.61	0.7	60	71.562	QC fail
41	B41	167.2	1.88	0.2	60	10.032	QC pass
42	B42	104.4	1.82	0.12	60	6.264	QC pass
43	B43	570.3	1.99	0.73	60	34.218	QC pass
44	B44	234.1	1.84	2.21	60	14.046	QC pass
45	B45	783.5	1.85	0.41	60	47.01	QC pass
46	B46	1006.9	1.93	0.74	60	60.414	QC pass
47	B47	886.7	1.95	0.42	60	53.202	QC pass
48	B48	345.2	1.91	0.98	60	20.712	QC pass
49	B49	239.3	1.96	0.34	60	14.358	QC pass
50	B50	400.1	1.81	0.56	60	24.006	QC pass
51	B51	620.4	1.83	0.67	60	37.224	QC pass
52	B52	781.2	1.78	0.54	60	46.872	QC pass
53	B53	2371.6	1.79	0.82	60	142.296	QC pass
54	B54	1993.8	1.8	0.7	60	119.628	QC pass
55	B55	537.6	1.83	0.54	60	32.256	QC pass
56	B56	2150.2	1.79	0.92	60	129.012	QC pass
57	B57	541.8	1.88	0.59	60	32.508	QC pass
58	B58	1304.3	1.93	0.72	60	78.258	QC pass
59	B59	1490.6	1.96	0.68	60	89.436	QC pass
60	B60	1378.7	1.99	0.67	60	82.722	QC pass
61	M1	59.1	1.91	1.99	60	3.546	QC fail
62	M2	384.2	1.92	2.17	60	23.04	QC fail
63	M3	50.4	0.41	0.38	60	3.024	QC fail
64	M4	213.6	0.36	0.33	60	12.816	QC fail
65	M5	134.8	2.02	2.15	60	8.088	QC fail
66	M6	31.9	1.15	1.18	60	1.914	QC fail
67	M7	126.8	0.53	0.34	60	7.608	QC fail
68	M8	45.1	1.14	0.54	60	2.706	QC fail
69	M9	34.1	1.89	0.54	60	2.046	QC fail
70	M10	17.2	2.13	1.23	60	1.032	QC fail
71	M11	1213.3	1.87	0.82	60	72.798	QC pass
72	M12	179.4	1.97	0.25	60	10.764	QC pass
73	M13	1229.5	1.78	0.58	60	73.77	QC pass
74	M14	108	1.82	0.14	60	6.48	QC pass
75	M15	813.9	1.88	0.47	60	48.834	QC pass
76	M16	123.3	1.91	0.15	60	7.398	QC pass
77	M17	411.4	1.92	0.48	60	24.684	QC fail
78	M18	1180.3	1.97	0.93	60	70.818	QC fail
79	M19	166.5	0.81	0.19	60	9.99	QC fail
80	M20	142.8	0.7	0.15	60	8.568	QC fail
81	M21	144.4	0.93	0.19	60	8.664	QC fail

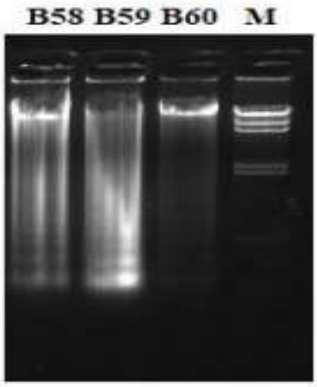
82	M22	945.7	1.63	0.7	60	56.742	QC fail
83	M23	887.6	1.58	0.62	60	53.256	QC fail
84	M24	1129	1.7	0.76	60	67.74	QC fail
85	M25	355.6	1.34	0.42	60	21.336	QC fail
86	M26	83.3	0.71	0.11	60	4.998	QC fail
87	M27	475.9	1.44	0.54	60	28.554	QC fail
88	M28	456	1.49	0.54	60	27.36	QC fail
89	M29	259.7	1.17	0.31	60	15.582	QC fail
90	M30	156.3	0.93	0.2	60	9.378	QC fail
91	M31	948	1.18	0.44	60	56.88	QC fail
92	M32	797	1.31	0.45	60	47.82	QC fail
93	M33	2816.4	1.89	1.33	60	168.984	QC fail
94	M34	719.5	1.5	0.59	60	43.17	QC fail
95	M35	874.1	1.76	0.97	60	52.446	QC fail
96	M36	1239.7	1.73	0.83	60	74.382	QC fail
97	M37	413.3	1.46	0.5	60	24.798	QC fail
98	M38	48.5	1.32	0.32	60	2.91	QC fail
99	M39	865.1	1.59	0.6	60	51.906	QC fail
100	M40	39.1	1.92	1.4	60	2.346	QC fail

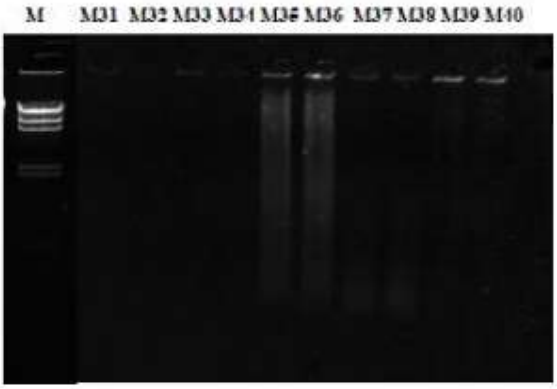


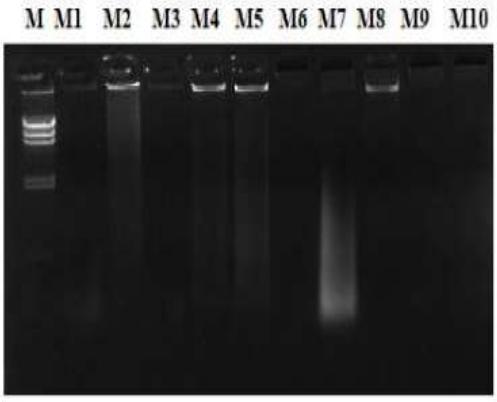


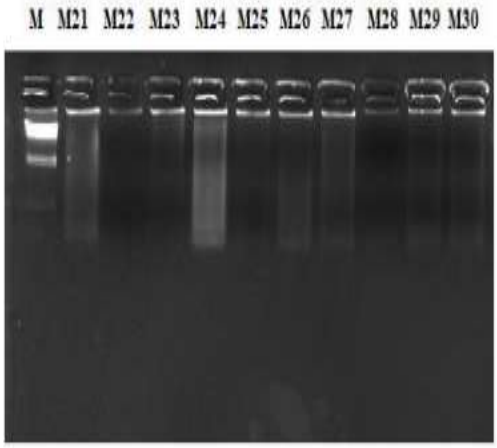

M
B17
B18
B19
B20


M
B38
B39
B40


B58
B59
B60
M


M
M31
M32
M33
M34
M35
M36
M37
M38
M39
M40


M
M1
M2
M3
M4
M5
M6
M7
M8
M9
M10


M
M21
M22
M23
M24
M25
M26
M27
M28
M29
M30

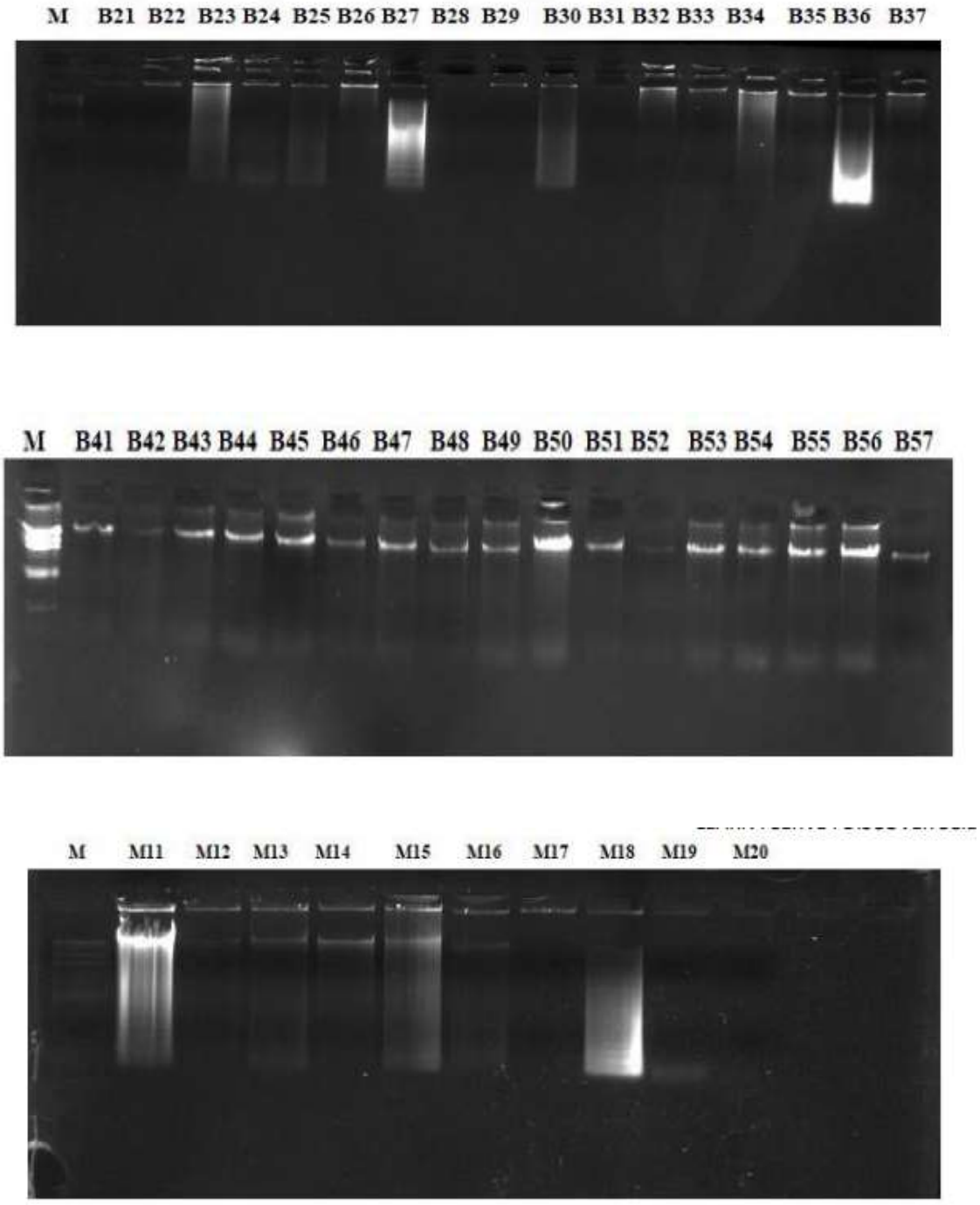


Figure-2: Consist of all the DNA samples obtained on 1XTAE Gel images

3.2 DNA RE-ISOLATION AND QC

DNA extarction of the samples which initially failed the extarction was re-isolated using CTAB method after which initally QC was perfoemed and for quantified and qualified using Agarose Gel Electrophoresis and nano-drop reading. The results were concluded that 76 sampes passed the primary QC. (Table-2,Figure-3)

Table-2: DNA Re-isolation of samples and their respective QC

S no	Sample ID	Nanodrop concentration (ng/ul)	260/280	260/230	Volume	Quantity (µg/ml)	Remarks
1	B21	125.6	1.87	0.26	60	7.536	QC Pass
2	B22	162.4	1.86	0.74	60	9.744	QC Pass
3	B23	156.4	1.84	0.76	60	9.384	QC Pass
4	B24	752.6	1.89	0.45	60	45.156	QC Pass
5	B25	55.1	1.82	1.1	60	3.306	QC Pass
6	B26	144.4	1.87	0.79	60	8.64	QC Pass
7	B27	125.2	1.99	0.69	60	7.512	QC Pass
8	B28	40.3	1.88	0.43	60	2.4	QC Pass
9	B31	127.9	1.98	0.71	60	7.62	QC Pass
10	B32	236.6	1.87	0.29	60	14.16	QC Pass
11	B34	45	1.88	0.34	60	2.7	QC Pass
12	B35	158.4	1.68	0.79	60	9.48	QC Pass
13	B36	800.9	1.85	0.6	60	4.8	QC Pass
14	B37	124.1	1.87	0.75	60	7.44	QC Pass
15	B38	431.1	1.86	0.55	60	25.86	QC Pass
16	B40	192.7	1.88	0.7	60	11.562	QC Pass
17	M1	159.1	1.91	1.99	60	9.546	QC fail
18	M2	284.2	1.92	2.17	60	17.05	QC Pass
19	M3	150.4	1.87	0.38	60	9.024	QC Pass
20	M5	134.8	1.86	2.15	60	8.08	QC Pass
21	M6	71.9	1.89	1.18	60	4.314	QC Pass
22	M8	145.1	1.9	0.54	60	8.706	QC Pass
23	M9	134.1	1.89	0.54	60	8.046	QC Pass
24	M22	545.7	1.98	0.7	60	32.74	QC Pass
25	M23	687.6	188	0.62	60	41.25	QC Pass
26	M24	129	1.87	0.76	60	7.74	QC Pass
27	M25	355.6	1.84	0.42	60	21.33	QC Pass
28	M36	139.7	1.89	0.83	60	8.382	QC Pass
29	M37	413.3	1.82	0.5	60	24.798	QC Pass
30	M39	265.1	1.9	0.6	60	15.906	QC Pass
31	M40	139.1	1.92	1.4	60	8.34	QC Pass

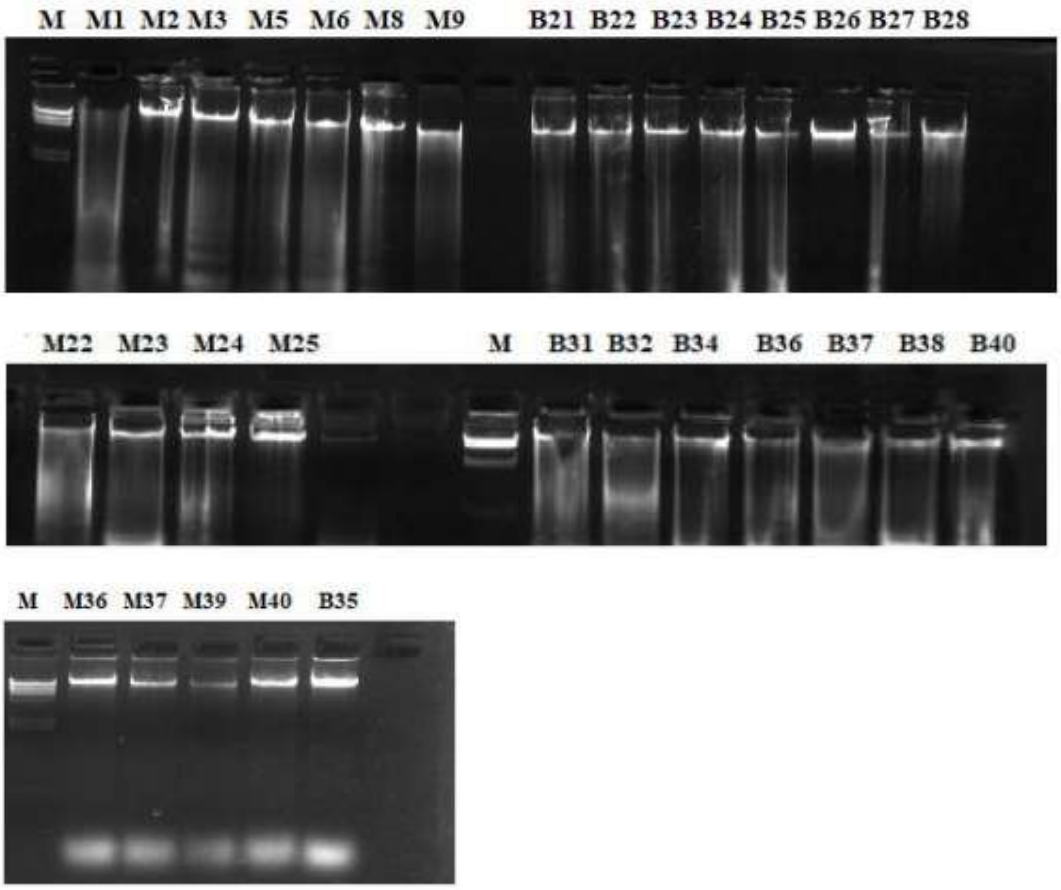


Figure-3: DNA Re-isolation gel electrophoresis

CONCLUSION: 76 samples in total passed the first QC.

3.3 LIBRARY PREPARATION FOR SEQUENCING

The genomic DNA samples underwent digestion with specific restriction enzymes-1, selected based on in-silico analysis. The resulting fragments were ligated with two types of barcoded adapters: one with compatible sticky ends with primary digestive enzymes and the illumina P5/P7 sequence, and the other without compatibility. Following multiple PCR amplifications, the samples were pooled, and suitable fragments were chosen to complete the construction of the GBS library (Figure-4).

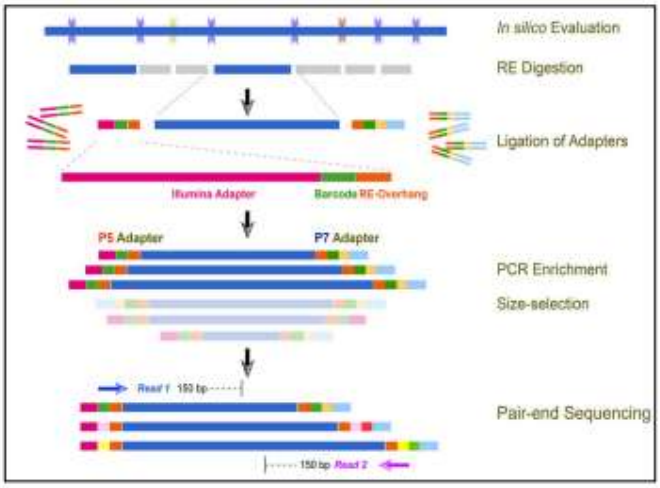


Figure-4 Overview of the Library Preparation

CONCLUSION: Out of the 76 samples, 66 samples passed and 10 failed the Library QC. The 66 passed samples were proceeded with Sequencing

3.3.1 DNA SEQUENCING AND RAW DATA

Qualified DNA libraries were combined based on projected data output and concentration. Illumina performed paired-end sequencing with 144 bp read lengths. Image files transformed to FASTQ for analysis.

3.3.2 SEQUENCING QUALITY DISTRIBUTION

The Phred Score (QPhred) representing base quality is calculated using QPhred = 10 log (e), where 'e' is the sequencing error rate. Table-3 shows the correlation between Phred scores from Casava version 1.8 and illumina sequencing quality. (Figure-5)

Table-3: The error rate distribution

Phred Score	Error Rate	Correct Rate	Q-Score
10	1/10	90.0%	Q10
20	1/100	99.0%	Q20
30	1/1000	99.9%	Q30
40	1/10000	99.99%	Q40

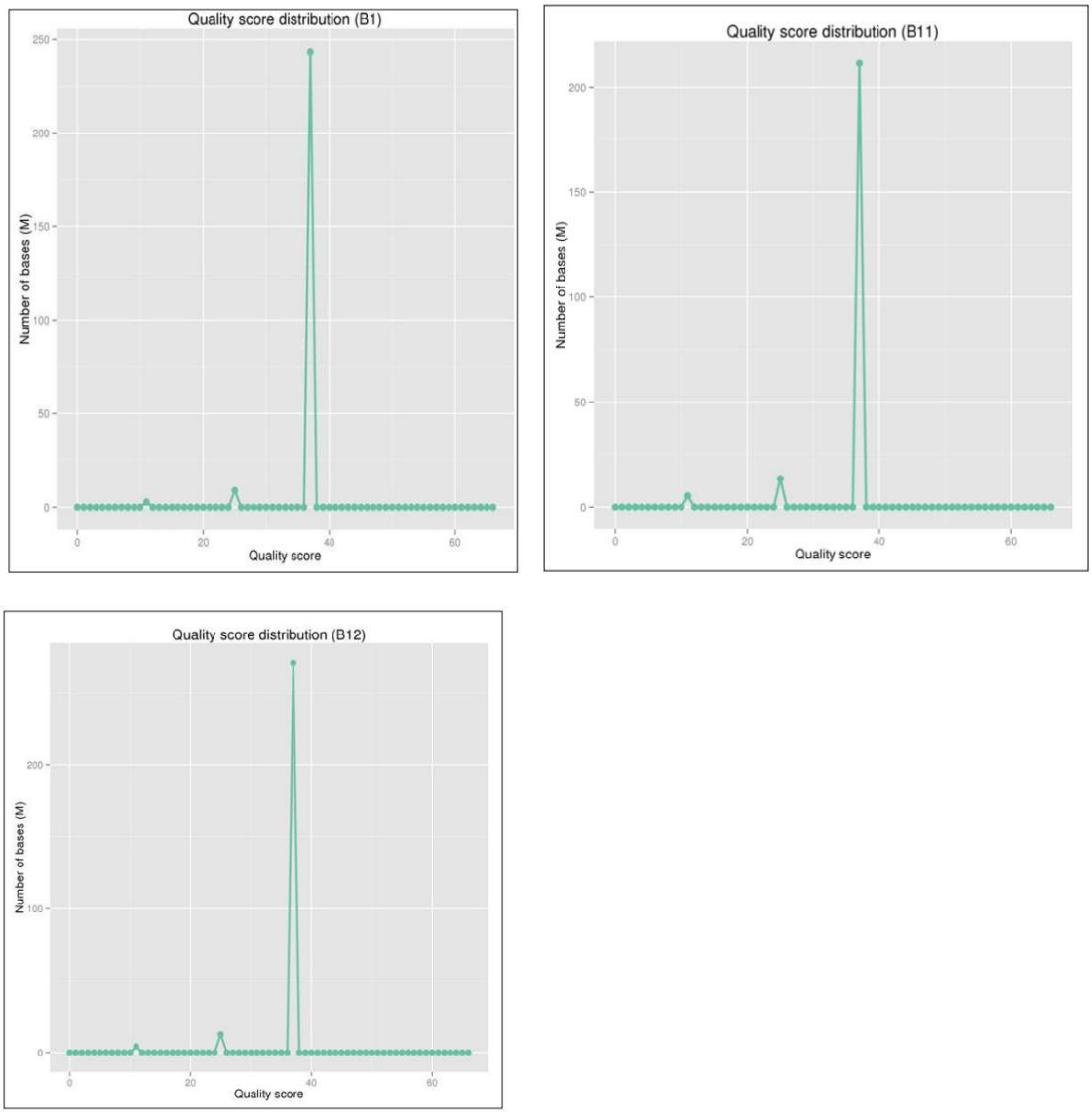


Figure-5: Describes the Distribution of Sequencing Quality

3.3.3 DISTRIBUTION OF SEQUENCING ERROR

Sequencing error is intricately linked to the inherent quality of the acquired sequence, influenced by factors like the sequencing platform, chemical reactants, and sample quality (Figure-6). Next-generation sequencing (NGS) utilizing sequencing-by-synthesis exhibits two common characteristics in its error rate distribution:

1. Error rate increases with longer sequencing read lengths due to chemical reagent consumption, DNA template degradation from laser irradiation, and potential error accumulation during sequencing cycles, a feature found in all Illumina high-throughput sequencing systems.
2. The initial bases experience higher sequencing error rates, likely attributed to reading errors during the first few cycles after optical instrument calibration.

To identify sites with unusually high error rates, where erroneous bases may be overrepresented, the entire sequence length is analysed for each sequence.

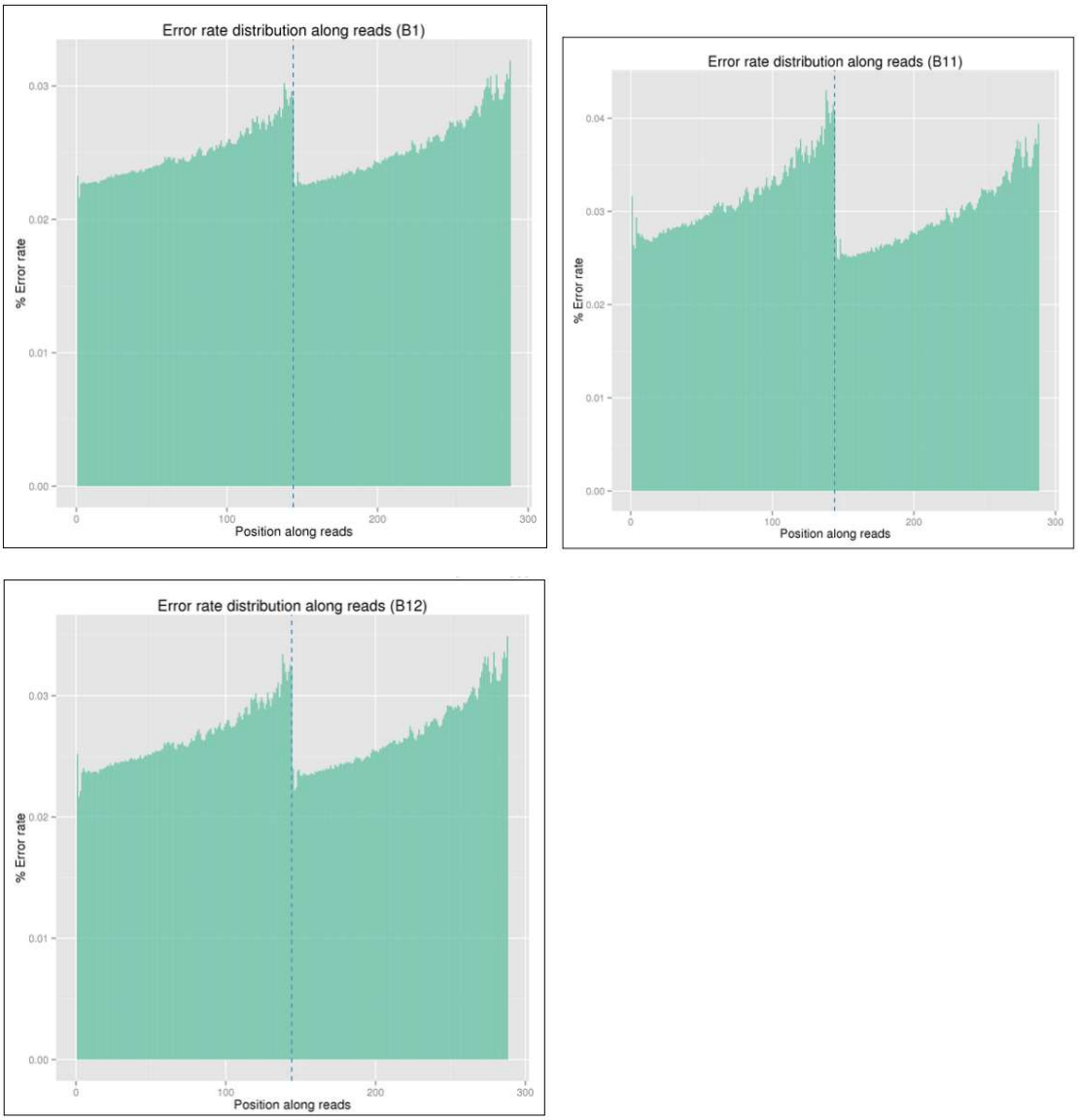


Figure-6: Distribution of sequencing error rate

3.3.4 GC CONTENT DISTRIBUTION

GC content distribution analysis can detect AT or GC separation. A balanced A to T and C to G ratio, following DNA base-pairing principles, is expected in a significant number of double-strand DNA sequences. GC concentration variations between species are evident in the base distribution, revealing traits associated with libraries. The sequencing library type and level directly impact the distribution pattern. N content, indicating incorrectly called bases during base calling, is indicative of sequencing quality. The presence of restriction enzyme cut sites in Read1 and Read2 for GBS affects GC content randomness, potentially leading to slight GC separation or disturbance. A thorough assessment of library construction and sequencing

quality involved calculating A, T, C, G, and N content and their distribution across sequence reads. (Figure-7)

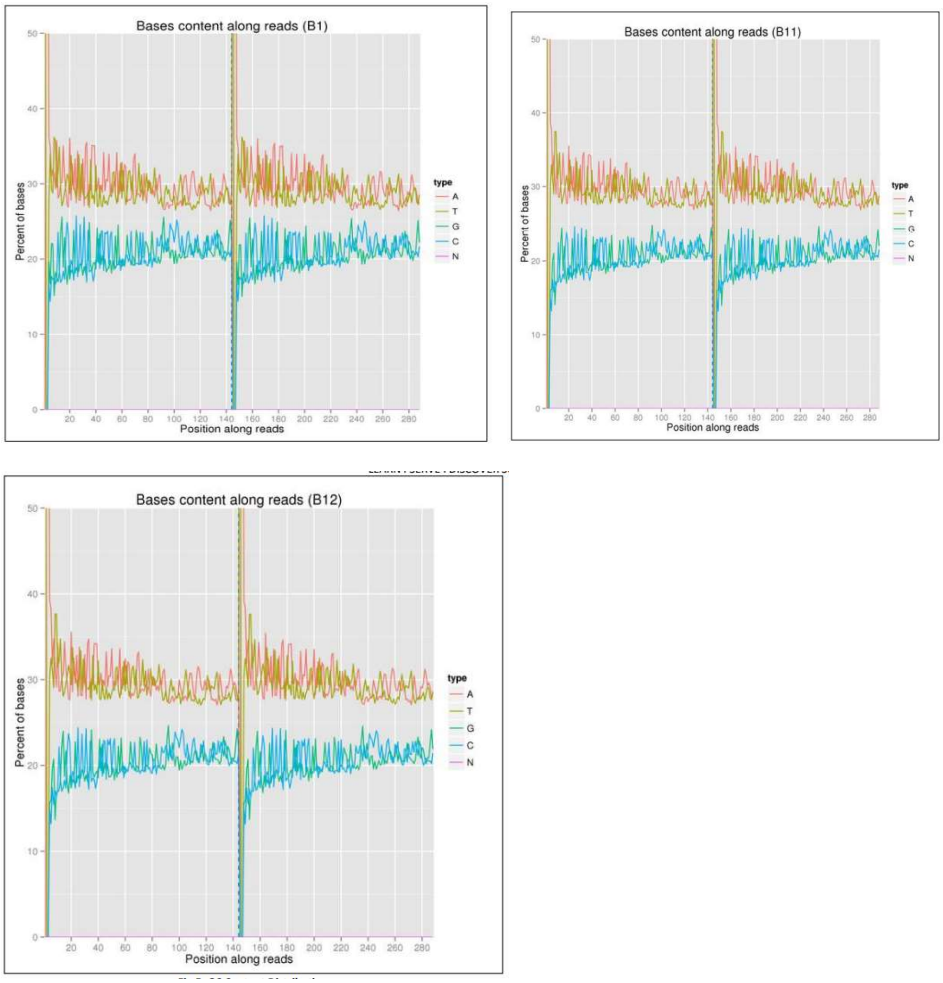


Figure-7: GC Content Distribution

3.3.5 SEQUENCING DATA FILTRATION

The collected raw data from sequencing is subjected to quality control to identify and eliminate low-quality reads and adapter contamination, which may complicate subsequent analyses. Cleaning the data ensures that only high-quality readings are used for further research (Figure-8). The quality control process involves the following steps:

1. Paired reads are discarded if either read contains adapter contamination.
2. Paired reads are discarded if uncertain nucleotides (N) constitute more than 10 percent of either read.
3. Paired reads are discarded if low-quality nucleotides (base quality less than 5, $Q \leq 5$) make up more than 50 percent of either read.

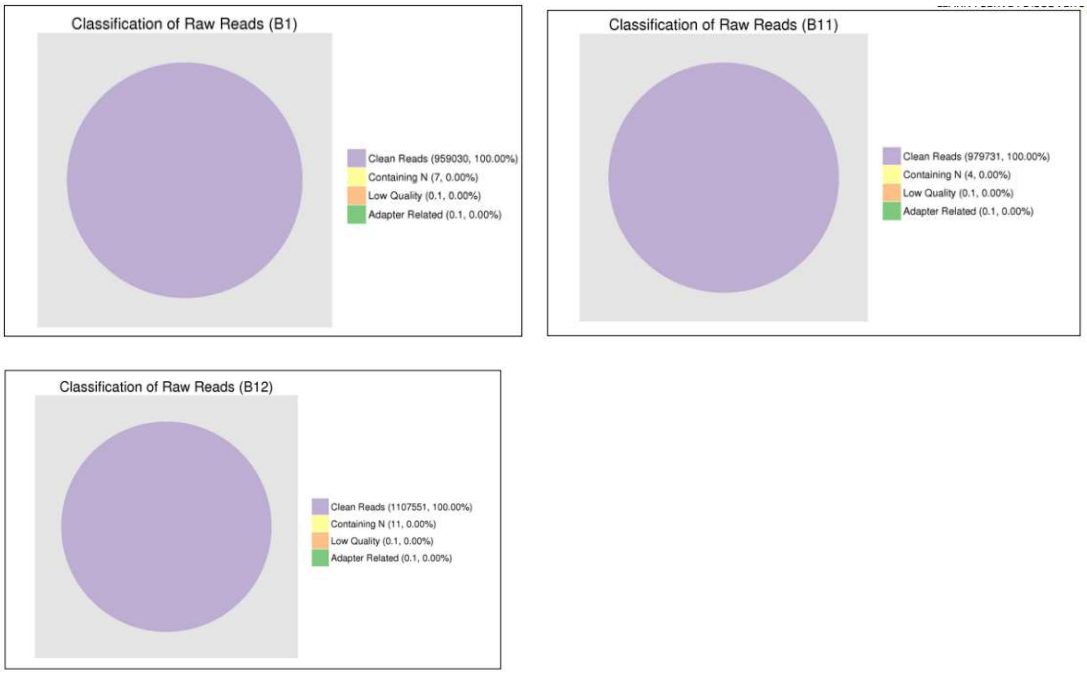


Figure-8: Classification of the sequenced reads

3.3.6 STATISTICS SUMMARY OF SEQUENCING DATA

The removal of low-quality data, this run resulted in 20.335G clean data out of a total of 20.335G raw data, encompassing 66 samples. The output range of raw data for each sample, from 216.634 M to 401.907 M, indicated sufficient data generation. The sequencing quality met the necessary analytical standards, achieving Q20 and Q30 scores of 94.56 percent and 85.55 percent, respectively. Additionally, the GC content ranging from 37.89 to 41.01 percent fell within the typical distribution range, ensuring the required level of quality. (Table-4) In conclusion, the library construction and sequencing procedures are successful and highly reliable.

Table-4: Statistics of the sequencing data

Sample	Raw Base(bp)	Clean Base(bp)	Effective Rate(%)	Error Rate(%)	Q20(%)	Q30(%)	GC Content (%)
B1	276202656	276200640	100	0.04	98.07	93.87	40.38
B11	282163680	282162528	100	0.04	96.22	89.16	40.05
B12	318977856	318974688	100	0.04	97.55	92.38	39.86
B13	216633888	216626112	100	0.04	96.6	90.06	40.02
B14	254157120	254149344	100	0.06	94.56	85.55	39.48
B15	293330880	293324832	100	0.04	97.66	92.71	40.17
B16	354966336	354963168	100	0.04	97.62	92.59	39.73
B17	371805120	371795328	100	0.04	97.42	92.09	39.67
B18	279913536	279905760	100	0.04	97.47	92.21	40.34
B19	395159328	395157312	100	0.04	97.13	91.36	39.9

B2	316273248	316272384	100	0.04	98.2	94.25	40.7
B20	331985376	331984512	100	0.04	97.67	92.71	40.05
B21	269582112	269582112	100	0.04	98.23	94.31	40.45
B22	303031584	303031584	100	0.04	98.41	94.76	40.49
B23	313735968	313728192	100	0.04	97.67	92.75	40.69
B24	285175584	285175584	100	0.04	97.91	93.41	40.16
B25	299424672	299416320	100	0.04	97.3	91.74	40.53
B26	316579968	316579968	100	0.04	96.82	90.53	40.09
B27	311710176	311702112	100	0.04	97.84	93.14	40.03
B28	350870688	350870688	100	0.04	97.25	91.63	40.96
B3	294225408	294213888	100	0.04	97.34	91.88	40.4
B31	363436416	363436416	100	0.05	95.86	88.1	41.01
B32	363296736	363286656	100	0.04	97.81	93.09	39.71
B34	264968352	264968352	100	0.04	97.92	93.4	40.3
B35	401907456	401907456	100	0.04	97.63	92.68	39.37
B36	252710208	252703296	100	0.04	97.67	92.77	40.32
B37	302855328	302855328	100	0.04	97.47	92.22	40.13
B38	303121728	303118560	100	0.04	97.73	92.96	38.91
B4	279381312	279380160	100	0.04	97.52	92.38	40.39
B40	268028640	268025184	100	0.04	97.89	93.37	40.33
B43	351966816	351954144	100	0.04	97.56	92.43	39.07
B45	400137984	400135968	100	0.04	97.43	92.1	40.39
B46	232714944	232709472	100	0.04	96.52	89.81	40.06
B47	390070944	390068928	100	0.04	98.14	93.99	40.56
B48	339139872	339137280	100	0.04	97.62	92.61	40.38
B50	321079968	321071040	100	0.04	97.88	93.36	40.54
B51	310729824	310723776	100	0.04	97.69	92.82	40.06
B52	290256768	290252448	100	0.04	97.19	91.44	39.33
B53	331274016	331268544	100	0.04	96.61	90.07	38.58
B54	228877920	228876480	100	0.04	96.74	90.39	40.28
B55	342912096	342908640	100	0.04	96.63	90.09	40.36
B56	375653664	375649344	100	0.04	97.43	92.13	40.02
B57	352452096	352442880	100	0.04	97.95	93.56	40.45
B58	315848736	315840960	100	0.04	97.58	92.5	40.22
B59	328479840	328470048	100	0.04	97.5	92.34	40.53
B60	314679744	314677728	100	0.04	98.1	93.97	40.48
B9	316032192	316030464	100	0.04	96.66	90.22	40.72
M11	272935008	272922912	100	0.04	97.77	93.03	40.35
M12	329647680	329645088	100	0.04	96.28	89.28	39.06
M13	322914816	322905024	100	0.04	96.95	90.91	40.06
M14	363389472	363387744	100	0.04	97.42	92.13	40.37
M15	316363968	316357920	100	0.04	96.72	90.34	40.09

M2	239485824	239480640	100	0.04	97.77	93.02	40.13
M22	306230112	306230112	100	0.04	97.41	92.06	39.55
M23	386918784	386912448	100	0.04	97.28	91.74	40.16
M24	365744160	365736672	100	0.04	97.21	91.53	40.81
M25	247569408	247569408	100	0.04	97.27	91.64	40.52
M3	295843968	295839936	100	0.04	96.81	90.52	37.89
M36	270252576	270252576	100	0.04	97.78	93.03	40.34
M37	294470496	294470496	100	0.04	98.16	94.15	40.41
M39	258729984	258729984	100	0.04	97.91	93.41	40
M40	233424288	233415936	100	0.04	97.87	93.31	39.44
M5	241623072	241623072	100	0.04	98.29	94.43	40.26
M6	265958208	265950432	100	0.04	97.91	93.34	40.27
M8	257656608	257650272	100	0.04	98.08	93.87	40.36
M9	288266400	288260352	100	0.04	97.92	93.44	40.05

3.3.7 ENZYMATIC DIGESTION EVALUATION

Table-5.1: Sample Proportions in Library- FGBS21H000093-1

lib name	sample name	bases num	rateInOneLib
FGBS21H000093-1	B16	354966336	0.032
FGBS21H000093-1	B17	371805120	0.034
FGBS21H000093-1	B14	254157120	0.023
FGBS21H000093-1	B15	293330880	0.026
FGBS21H000093-1	B12	318977856	0.029
FGBS21H000093-1	B13	216633888	0.02
FGBS21H000093-1	B11	282163680	0.025
FGBS21H000093-1	B18	279913536	0.025
FGBS21H000093-1	B19	395159328	0.036
FGBS21H000093-1	M11	272935008	0.025
FGBS21H000093-1	B58	315848736	0.028
FGBS21H000093-1	B59	328479840	0.03
FGBS21H000093-1	B52	290256768	0.026
FGBS21H000093-1	B53	331274016	0.03
FGBS21H000093-1	B50	321079968	0.029
FGBS21H000093-1	B51	310729824	0.028
FGBS21H000093-1	B56	375653664	0.034
FGBS21H000093-1	B57	352452096	0.032
FGBS21H000093-1	B54	228877920	0.021
FGBS21H000093-1	B55	342912096	0.031
FGBS21H000093-1	M14	363389472	0.033
FGBS21H000093-1	B60	314679744	0.028
FGBS21H000093-1	B45	400137984	0.036

FGBS21H000093-1	B47	390070944	0.035
FGBS21H000093-1	B46	232714944	0.021
FGBS21H000093-1	B43	351966816	0.032
FGBS21H000093-1	B48	339139872	0.031
FGBS21H000093-1	B20	331985376	0.03
FGBS21H000093-1	B4	279381312	0.025
FGBS21H000093-1	M13	322914816	0.029
FGBS21H000093-1	M12	329647680	0.03
FGBS21H000093-1	B1	276202656	0.025
FGBS21H000093-1	B2	316273248	0.029
FGBS21H000093-1	B3	294225408	0.027
FGBS21H000093-1	B9	316032192	0.028

Table-5.2: Sample Proportions in Library- FGBS21H000094-1

lib name	sample name	bases num	rateInOneLib
FGBS21H000094-1	M39	258729984	0.028
FGBS21H000094-1	M37	294470496	0.032
FGBS21H000094-1	M36	270252576	0.029
FGBS21H000094-1	M5	241623072	0.026
FGBS21H000094-1	M6	265958208	0.029
FGBS21H000094-1	M3	295843968	0.032
FGBS21H000094-1	M2	239485824	0.026
FGBS21H000094-1	M9	288266400	0.031
FGBS21H000094-1	M8	257656608	0.028
FGBS21H000094-1	B38	303121728	0.033
FGBS21H000094-1	B34	264968352	0.029
FGBS21H000094-1	B35	401907456	0.044
FGBS21H000094-1	B36	252710208	0.027
FGBS21H000094-1	B37	302855328	0.033
FGBS21H000094-1	B31	363436416	0.039
FGBS21H000094-1	B32	363296736	0.039
FGBS21H000094-1	M24	365744160	0.04
FGBS21H000094-1	M25	247569408	0.027
FGBS21H000094-1	M40	233424288	0.025
FGBS21H000094-1	M22	306230112	0.033
FGBS21H000094-1	M23	386918784	0.042
FGBS21H000094-1	B40	268028640	0.029
FGBS21H000094-1	B28	350870688	0.038
FGBS21H000094-1	B27	311710176	0.034
FGBS21H000094-1	B26	316579968	0.034
FGBS21H000094-1	B25	299424672	0.032

FGBS21H000094-1	B24	285175584	0.031
FGBS21H000094-1	B23	313735968	0.034
FGBS21H000094-1	B22	303031584	0.033
FGBS21H000094-1	B21	269582112	0.029
FGBS21H000094-1	M15	316363968	0.034

3.3.8 ENZYMATIC DIGESTION SUMMARY

Reads that lack the recognition sequence of the primary or additional restriction enzyme(s) are categorized as completely cut, while reads with the exact conserved sequence of the first restriction enzyme at the beginning and ends of both Read1 and Read2 are considered successfully enzyme-captured reads among paired clean reads. The enzyme digestion proportion in this project ranges from 62.9 percent to 98.0 percent, while the percentage of enzyme-captured reads ranges from 98.1 percent to 99.7 percent. (Table-6)

Table-6: Enzymatic Digestion Summary

sample	total_PE_cleanReads	total_PE_enzyme CatchReads	total_PE_enzyme CutCompletely Reads	enzymeCatch Ratio(%)	enzymeCut Completely Ratio(%)
B1	959030	955116	930864	99.6	97.5
B11	979731	973717	926656	99.4	95.2
B12	1107551	1102868	1024446	99.6	92.9
B13	752174	747570	716290	99.4	95.8
B14	882463	870483	807310	98.6	92.7
B15	1018489	1014987	983034	99.7	96.9
B16	1232511	1227345	1184943	99.6	96.5
B17	1290956	1285931	1229947	99.6	95.6
B18	971895	967294	938370	99.5	97
B19	1372074	1363262	1316776	99.4	96.6
B2	1098168	1094820	1072631	99.7	98
B20	1152724	1148820	1114625	99.7	97
B21	936049	933006	910742	99.7	97.6
B22	1052193	1048930	1026540	99.7	97.9
B23	1089334	1084573	1055682	99.6	97.3
B24	990193	985727	957593	99.5	97.1
B25	1039640	1032718	1000989	99.3	96.9
B26	1099236	1091622	1037261	99.3	95
B27	1082299	1078139	1015616	99.6	94.2
B28	1218301	1213054	1177477	99.6	97.1
B3	1021576	1016194	984223	99.5	96.9
B31	1261932	1247673	1182312	98.9	94.8

B32	1261412	1255826	1191790	99.6	94.9
B34	920029	916501	888577	99.6	97
B35	1395512	1388775	1274207	99.5	91.8
B36	877442	872973	843721	99.5	96.6
B37	1051581	1046183	1005432	99.5	96.1
B38	1052495	1046310	977557	99.4	93.4
B4	970070	967531	941221	99.7	97.3
B40	930643	927633	888396	99.7	95.8
B43	1222063	1217654	1058811	99.6	87
B45	1389361	1380893	1343644	99.4	97.3
B46	808019	801538	765536	99.2	95.5
B47	1354406	1350783	1317811	99.7	97.6
B48	1177560	1173902	1143117	99.7	97.4
B50	1114830	1106354	1077656	99.2	97.4
B51	1078902	1071857	1037433	99.3	96.8
B52	1007821	1002081	951591	99.4	95
B53	1150238	1140142	1056977	99.1	92.7
B54	794710	785167	754536	98.8	96.1
B55	1190655	1182594	1145866	99.3	96.9
B56	1304338	1297403	1249052	99.5	96.3
B57	1223760	1220233	1192107	99.7	97.7
B58	1096670	1093431	1060274	99.7	97
B59	1140521	1134107	1102155	99.4	97.2
B60	1092631	1084890	1060083	99.3	97.7
B9	1097328	1090615	1059261	99.4	97.1
M11	947649	944580	918754	99.7	97.3
M12	1144601	1131736	1075106	98.9	95
M13	1121198	1099672	981680	98.1	89.3
M14	1261763	1245117	1202080	98.7	96.5
M15	1098465	1090699	1047221	99.3	96
M2	831530	827899	800231	99.6	96.7
M22	1063299	1056323	1012882	99.3	95.9
M23	1343446	1335009	1265534	99.4	94.8
M24	1269919	1257668	1221842	99	97.2
M25	859616	854650	826012	99.4	96.6
M3	1027222	1012149	636711	98.5	62.9
M36	938377	935132	906092	99.7	96.9
M37	1022467	1018695	993208	99.6	97.5
M39	898368	895274	862098	99.7	96.3
M40	810472	806402	757712	99.5	94
M5	838969	836489	817748	99.7	97.8
M6	923439	920004	888525	99.6	96.6

M8	894619	891587	869648	99.7	97.5
M9	1000904	994060	963751	99.3	97

3.3.9 MAPPING STATISTICS WITH REFERENCE GENOME

The sample mapping rates show how closely each sample resembles the reference genome. Indicators of evenness and homology with the reference genome include depth and coverage. (Table-7)

Table-7: Statistics of mapping rate, depth and coverage

sample	clean reads	mapped reads	mapping rate(%)	average depth(X)	coverage at least 1X(%)	coverage at least 4X(%)
B1	1861728	1802451	96.82	4.36	12.35	5
B11	1853312	1808720	97.59	4.44	12.2	4.94
B12	2048892	1995448	97.39	4.54	13.09	5.29
B13	1432580	1391845	97.16	3.41	12.13	4
B14	1614620	1573315	97.44	4.84	9.71	4.16
B15	1966068	1923581	97.84	4.57	12.63	5.23
B16	2369886	2316786	97.76	6.31	11	5.34
B17	2459894	2402791	97.68	6.48	11.08	5.46
B18	1876740	1831361	97.58	4.42	12.41	5.04
B19	2633552	2564628	97.38	6.42	11.95	5.98
B2	2145262	2063994	96.21	5.24	11.77	5.33
B20	2229250	2176607	97.64	5.67	11.49	5.49
B21	1821484	1782891	97.88	4.14	12.92	4.9
B22	2053080	2010230	97.91	4.24	14.21	5.47
B23	2111364	2048665	97.03	4.54	13.49	5.39
B24	1915186	1862843	97.27	3.49	15.93	5.38
B25	2001978	1958376	97.82	4.1	14.33	5.41
B26	2074522	2030701	97.89	5.29	11.51	4.96
B27	2031232	1985012	97.72	4.22	14.1	5.4
B28	2354954	2299672	97.65	5.28	13.01	5.66
B3	1968446	1923327	97.71	4.38	13.11	5.31
B31	2364624	2309280	97.66	5.99	11.52	5.23
B32	2383580	2331397	97.81	5.83	11.97	5.22
B34	1777154	1687420	94.95	3.73	13.51	4.62
B35	2548414	2491369	97.76	6.97	10.67	4.99
B36	1687442	1647246	97.62	4.11	11.94	4.48
B37	2010864	1966988	97.82	4.97	11.88	4.99
B38	1955114	1910223	97.7	6.34	8.98	4.14
B4	1882442	1835319	97.5	4.21	13.03	5.2
B40	1776792	1735847	97.7	4.08	12.72	4.76

B43	2117622	2068057	97.66	6.27	9.87	4.39
B45	2687288	2624690	97.67	6.11	12.82	6.31
B46	1531072	1493182	97.53	3.65	12.07	4.21
B47	2635622	2571532	97.57	5.96	12.94	6.27
B48	2286234	2235134	97.76	5.95	11.23	5.59
B50	2155312	2107405	97.78	4.99	12.63	5.51
B51	2074866	2025545	97.62	4.67	12.94	5.5
B52	1903182	1859192	97.69	5.22	10.63	4.76
B53	2113954	2064187	97.65	6.49	9.51	4.57
B54	1509072	1470312	97.43	3.56	12.34	4.17
B55	2291732	2214531	96.63	5.63	11.76	5.67
B56	2498104	2438654	97.62	5.59	13.03	6.01
B57	2384214	2329893	97.72	5.21	13.36	6.05
B58	2120548	2069101	97.57	5.19	11.91	5.4
B59	2204310	2153909	97.71	5.22	12.29	5.58
B60	2120166	2074949	97.87	5.03	12.33	5.47
B9	2118522	2037376	96.17	5.21	11.68	5.32
M11	1837508	1799582	97.94	3.86	14.07	5.26
M12	2150212	2110219	98.14	6.56	9.72	5.04
M13	1963360	1906524	97.11	4.58	12.56	5.25
M14	2404160	2357277	98.05	4.62	15.42	6.51
M15	2094442	2058374	98.28	6.46	9.61	5.07
M2	1600462	1568900	98.03	3.31	14.3	4.51
M22	2025764	1997319	98.6	6.25	9.71	4.71
M23	2531068	2494447	98.55	5.77	13.15	5.97
M24	2443684	2392061	97.89	5.35	13.48	5.97
M25	1652024	1620662	98.1	3.55	13.76	4.6
M3	1273422	1239258	97.32	4.71	7.96	3.27
M36	1812184	1778999	98.17	3.69	14.57	5.07
M37	1986416	1945666	97.95	4.15	14.13	5.37
M39	1724196	1685804	97.77	3.59	14.11	4.83
M40	1515424	1481979	97.79	3	14.81	4.07
M5	1635496	1606013	98.2	3.7	13.11	4.56
M6	1777050	1744171	98.15	3.89	13.5	4.89
M8	1739296	1709364	98.28	4.24	12.2	4.78
M9	1927502	1896785	98.41	4.29	13.39	5.19

3.3.10 SUMMARY OF MAPPING RESULTS

The mapping rate of each sample to the 459,881,487 bp reference genome varies from 94.95% to 98.6%. The average depth on the reference genome (excluding Ns) ranges from 3.0X to

6.97X, with over 7.96% having more than 1X coverage. These results fall within the qualified normal range and are suitable for subsequent variation detection and related analyses.

4.0 CONCLUSION

In India, the sericulture industry encompasses more than 500 silkworm genotypes, including multivoltine and bivoltine breeds. Studying the distribution of genetic variation and diversity in silkworms is essential due to the vast germplasm and lack of information on favorable traits for selecting breeding parents to improve yield. To preserve the genetic variety with minimal loss of diversity and redundancy, a core collection representing the entire collection has been proposed. This core collection would include a small group of accessions with maximum allelic diversity in the least amount of material.

At CSRTI, Mysore, silkworm breeds were collected from various locations and subjected to three inbreeding cycles. A diverse set of approximately 100 genotypic silkworm breeds, consisting of 60 bivoltine and 40 multivoltine breeds, was identified for Genotyping by Sequencing (GBS) analysis. Phenotypic data focused on various traits such as pupation percentage, cocoon weight, shell weight, shell percentage, thermotolerance, disease tolerance (NPV), filament length, Reelability percentage, raw silk percentage, neatness, and evenness.

5.0 FUNDING

The project was funded by Department of Biotechnology (DBT), Ministry of Science and Technology, Government of India. The project titled Assessment of SNP Variation in Silkworm (Bombyx mori) by RAD Sequencing and Genome-wide Association Mapping of Important Commercial Traits and the sanction order number is BT/PR15059/TDS/121/9/2015.

6.0 ACKNOWLEDGEMENT

We would like to thank Department of Biotechnology (DBT), Ministry of Science and Technology, Government of India for the funds and Central Sericulture Research & Training Institute, Mysore for rearing silkworms. We would like to show our sincere gratitude to Dr. Anala MR Professor, Department of Information Science and Engineering, R V College of Engineering, Bangalore, India for providing us Titan X GPU for high-performance computing. The authors would like to specially thank all the professor and staffs of R.V. College of Engineering for their support for completing the project.

7.0 CONFLICT OF INTEREST

The authors declare there is no conflict of interest

8.0 REFERENCE

1. Biology analysis group, Xia Q, Zhou Z, Lu C, Cheng D, Dai F, et al. A Draft Sequence for the Genome of the Domesticated Silkworm (*Bombyx mori*). Science. 2004 Dec 10;306(5703):1937–40.

2. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience. 2015 Dec;4(1):7.

3. Tong X, Han MJ, Lu K, Tai S, Liang S, Liu Y, et al. High-resolution silkworm pan-genome provides genetic insights into artificial selection and ecological adaptation. Nat Commun. 2022 Sep 24;13(1):5619.

4. Grattapaglia D, de Alencar S, Pappas G. Genome-wide genotyping and SNP discovery by ultra-deep Restriction-Associated DNA (RAD) tag sequencing of pooled samples of E. grandis and E. globulus. BMC Proc. 2011 Dec;5(S7):P45, 1753-6561-5-S7-P45.

5. Yangkun W, Yan H, Tianzhen Z. Current status and perspective of RAD-seq in genomic research: Current status and perspective of RAD-seq in genomic research. Hered Beijing. 2014 Apr 25;36(1):41–9.

6. Fang SM, Zhou QZ, Yu QY, Zhang Z. Genetic and genomic analysis for cocoon yield traits in silkworm. Sci Rep. 2020 Mar 30;10(1):5682.

7. Uchiyama H, Takasu Y, Moriyama M, Yoshitake K, Uchiyama H, Iizuka T, et al. A novel monocarboxylate transporter involved in 3-hydroxykynurenine transport for ommochrome coloration [Internet]. Genetics; 2023 Jun [cited 2023 Jul 25]. Available from: http://biorxiv.org/lookup/doi/10.1101/2023.06.01.543243

8. Schork NJ, Fallin D, Lanchbury JS. Single nucleotide polymorphisms and the future of genetic epidemiology: SNPs and genetic epidemiology. Clin Genet. 2000 Oct;58(4):250–64.

9. Brumfield RT, Beerli P, Nickerson DA, Edwards SV. The utility of single nucleotide polymorphisms in inferences of population history. Trends Ecol Evol. 2003 May;18(5):249–56.

10. Members of the Complex Trait Consortium. The nature and identification of quantitative trait loci: a community's view. Nat Rev Genet. 2003 Nov;4(11):911–6.

11. Dhingani RM, Umrania VV, Tomar RS, Parakhia MV. Introduction to QTL mapping in plants. Ann Plant Sci. 2015;