HISTONE MODIFICATION MARKS IMPROVE IDENTIFICATION OF OIL PALM TRANSCRIPTION START SITES

SARPAN, N1; TATARINOVA, T V2; LOW, E-T L1; ONG-ABDULLAH, M1; SAPIAN, I S3 and OOI, S-E1*

ABSTRACT

Epigenetic regulation involves modifications of chromatin components such as post-translational modifications of histone proteins, methylation of cytosines in deoxyribonucleic acid (DNA), the involvement of small RNA and chromatin remodeling. Numerous methods have been established to understand the epigenetic control of agronomically important traits. Chromatin immunoprecipitation with sequencing (ChIP-Seq) is widely used to identify the binding sites of transcription factors or modified histones on a genome-wide scale. Here, ChIP-Seq targeting H3K4me3 and H3K27me3 marks in oil palm spears were conducted to examine genomic regions enriched with these histone modifications. Due to low DNA amounts from ChIP experiments, the data analysis workflow was optimised based on ChIP-Seq workflows on other plants. Mapping to specific target regions revealed that the histone mark peak positions were located close to predicted transcription start sites (TSS). This agrees with H3K4me3 and H3K27me3 profiles in other plants where H3K4me3 marks are generally associated with active genes and promoter regions while H3K27me3 marks are linked to repressed genes. Gene-wide mapping for low coverage ChIP-Seq data showed that H3K4me3 and H3K27me3 profiles on the oil palm genome corresponded to consensus histone profiles in other plants. This is the first ChIP-Seq analysis workflow reported for oil palm spears, which can be used to develop future oil palm ChIP-Seq studies.

Keywords: ChIP-Seq, Elaeis guineensis, epigenetics, H3K4me3, H3K27me3.

Received: 19 October 2020; Accepted: 2 February 2021; Published online: 19 May 2021.

INTRODUCTION

Epigenetics mechanisms lead to spatial-temporal gene expression changes that are not mediated by changes in the underlying deoxyribonucleic acid (DNA) sequence (Wollmann and Berger, 2012). Unlike the genome, the epigenome is dynamically altered by environmental factors (Baulcombe and Dean, 2014). In oil palm, crop improvement programmes involve numerous breeding trials, which are then selected for massive multiplication through tissue culture (Zulkifli et al., 2017). However, the production of elite clonal materials has yet to be maximised due to the occurrence of mantled fruits that causes oil and yield losses. Through epigenome-wide association studies, the loss of DNA methylation at the EgDEF1’s Karma epi-allele in tissue culture regenerants was associated with the mantling phenotype (Ong-Abdullah et al., 2015). Subsequently, an epi-fingerprinting assay for in vitro cultured palms was developed to discriminate true-to-types from off-types (Ong-Abdullah et al., 2016). This step serves as quality control (QC) to improve the efficiency of the tissue culture process and thus, marks the beginning of applied epigenetics for the
oil palm industry through the identification of low-risk planting materials.

Chromatin, which is built by a nucleosome array containing 147 bp of DNA and pairs of histones H2A, H2B, H3 and H4, acts as a target for epigenetic modifiers comprising methylation of cytosines in the DNA, modifications at the N-terminal tails of histone proteins and small non-coding ribonucleic acid (RNA) molecules (Lu et al., 2020). These mechanisms affect the packaging of chromatin and accessibility of the transcriptional machinery, thus, influencing gene expression. The pattern of modifications on histone tails, known as the histone code, contribute to shaping chromatin structure into either condensed heterochromatin or open euchromatin (Prakash and Fournier, 2018). H3K4me3 and H3K27me3 are two widely investigated histone marks in plants. The wide distribution of H3K4me3 marks in promoter regions is associated with active transcription, while H3K27me3 marks are usually associated with the Polycomb Repressive Complex involved in gene silencing (Gan et al., 2015; Zhang et al., 2009). In plants, histone modification signatures are seen for several developmental processes, including response to stress (Zeng et al., 2019), flower morphogenesis (Engelhorn et al., 2017) and embryogenesis (Berenguer et al., 2017). Studies on histone modifications often involve locating their binding sites through a method called chromatin immunoprecipitation with sequencing (ChIP-Seq) (Liu et al., 2010). In ChIP-Seq, genomic DNA is first fragmented, followed by the capturing of histone-bound fragments. The non-histone fractions are then discarded, leaving only the enriched fraction that will be sequenced. Subsequent bioinformatics analysis of ChIP-Seq datasets generally includes QC checks of sequencing reads, aligning reads to the reference genome and finding peaks corresponding to ChIP-enriched regions. Reported ChIP-Seq data usually involves at least 20 million mapped reads (Zhang et al., 2015; Zong et al., 2013), but the chromatin immunoprecipitation step generally provides low yields in the nanogram range (Ranawaka et al., 2020). Thus, reads duplication is more apparent when insufficient DNA from inefficient immunoprecipitation is involved in constructing sequencing library and becomes more abundant when the library is deeply sequenced (Landt et al., 2012).

In this study, the histone modifications H3K4me3 and H3K27me3 in oil palm leaf spear tissue were investigated using ChIP-Seq. These two histone modifications were selected based on their influence on transcriptional regulation, whereby the H3K4me3 is associated with active transcription while the H3K27me3 is linked to repressed transcription at gene-rich regions. The DNA bound to histones enriched with these two histone modifications was immunoprecipitated and sequenced. The low DNA yields obtained through ChIP led to low coverage ChIP-Seq data. To circumvent this limitation and still gain preliminary insights on these histone marks in oil palm, a gene-wide mapping approach was able to demonstrate that H3K4me3 and H3K27me3 profiles in oil palm are in concordance with their consensus distribution patterns on gene regions in other plants. To our knowledge, this is the first report of a ChIP-Seq workflow described for the oil palm.

MATERIALS AND METHODS

Plant Material

Spear leaf tissues of 10 clonal palms at the MPOB Bagan Datuk Research Station, Perak, Malaysia were sampled, frozen in liquid nitrogen and stored at -80°C until further use. The palms, labeled from A to J (Table 1), originated from the same tissue used in in vitro cloning and are thus, grouped as biological replicates.

Chromatin Immunoprecipitation and Sequencing

Chromatin was isolated following a protocol described by Kaufmann et al. (2010), with modifications, as described by Sarpan et al. (2018). Briefly, the spears were ground in liquid nitrogen, added with lysis buffer and cold-incubated. The mixture was filtered and centrifuged to collect the pellet. The pellet underwent a few series of washing steps prior to shearing. Chromatin was sheared using a Covaris M220 (Covaris, USA) for 25 min in a screw-cap microtube at 4°C with the following settings: 5% duty factor and 200 cycles/burst. ChIP-validated antibodies, anti-H3K4me3 (ab12209) and anti-H3K27me3 (ab6002, Abcam, United Kingdom), as used by Sarpan et al. (2018) and EpiSeeker ChIP Kit-Plants (Abcam, United Kingdom) were used for immunoprecipitation in this study. Sequencing libraries were prepared from ChIP DNA using the TruSeq® ChIP library preparation kit (Illumina, USA). The amount of input DNA (non-immunoprecipitated) used was 10 ng, while ChIP DNA starting amounts were varied between 1 and 5 ng. A total of 10 H3K27me3 ChIP-Seq libraries were prepared but only nine H3K4me3 ChIP-Seq libraries were generated as library preparation failed for one of the samples. Libraries were sequenced paired-end at a read-length of 125 bp on a HiSeq2500 Illumina platform with a depth of at least 20 million reads.

ChIP-Seq Analysis

The quality of sequencing reads was assessed using FastQC v0.11.3 software (Babraham Bioinformatics). Raw reads were trimmed using Flexbar v3.4.0 (Dodt et al., 2012) to filter out...
| Replicate | ChIP-SEQ Datasets Mapped to Genome-Wide |
|-----------|-----------------------------------------|
|           | H3K4me3 control | H3K4me3 ChIP | H3K27me3 control | H3K27me3 ChIP |
| A         | 35,984,910 | 35,669,134 | 35,435,466 | 27,756,510 |
| B         | 50,059,260 | 51,665,290 | 44,625,124 | 1,334,086 |
| C         | 41,616,024 | 40,797,260 | 39,660,246 | 11,028,820 |
| D         | 41,616,024 | 40,797,260 | 39,660,246 | 11,028,820 |
| E         | 41,616,024 | 40,797,260 | 39,660,246 | 11,028,820 |
| F         | 41,616,024 | 40,797,260 | 39,660,246 | 11,028,820 |
| G         | 41,616,024 | 40,797,260 | 39,660,246 | 11,028,820 |
| H         | 41,616,024 | 40,797,260 | 39,660,246 | 11,028,820 |
| I         | 41,616,024 | 40,797,260 | 39,660,246 | 11,028,820 |
| J         | 41,616,024 | 40,797,260 | 39,660,246 | 11,028,820 |

**Table 1:**

| Library Type | Biological replicate | Raw reads | Trimmed reads | Mapped reads | Non-duplicate reads | Total bases (all positions) | Bases with 0 reads | Bases with ≥1 reads | Genome coverage (%) | Read coverage per base |
|--------------|----------------------|-----------|---------------|--------------|-------------------|-----------------------------|------------------|-------------------|---------------------|----------------------|
| A            | control              | 35,984,910 | 35,669,134 | 35,435,466 | 27,756,510        | 809,330,882                | 52.76            | 4.15              |                     |                      |
| B            | H3K4me3              | 50,059,260 | 51,665,290 | 44,625,124 | 1,334,086         | 487,620,203                | 61.76            | 5.71              |                     |                      |
| C            | H3K4me3              | 41,616,024 | 40,797,260 | 39,660,246 | 11,028,820        | 683,486,435                | 55.45            | 4.22              |                     |                      |
| D            | H3K4me3              | 41,616,024 | 40,797,260 | 39,660,246 | 11,028,820        | 683,486,435                | 55.45            | 4.22              |                     |                      |
| E            | H3K4me3              | 41,616,024 | 40,797,260 | 39,660,246 | 11,028,820        | 683,486,435                | 55.45            | 4.22              |                     |                      |
| F            | H3K4me3              | 41,616,024 | 40,797,260 | 39,660,246 | 11,028,820        | 683,486,435                | 55.45            | 4.22              |                     |                      |
| G            | H3K4me3              | 41,616,024 | 40,797,260 | 39,660,246 | 11,028,820        | 683,486,435                | 55.45            | 4.22              |                     |                      |
| H            | H3K4me3              | 41,616,024 | 40,797,260 | 39,660,246 | 11,028,820        | 683,486,435                | 55.45            | 4.22              |                     |                      |
| I            | H3K4me3              | 41,616,024 | 40,797,260 | 39,660,246 | 11,028,820        | 683,486,435                | 55.45            | 4.22              |                     |                      |
| J            | H3K4me3              | 41,616,024 | 40,797,260 | 39,660,246 | 11,028,820        | 683,486,435                | 55.45            | 4.22              |                     |                      |

**Notes:**
- H3K4me3 and H3K27me3 marks improve identification of oil palm transcription start sites.
- The table includes biological replicates for each library type and replicate.
- The coverage metrics are based on total reads mapped to the genome.
contaminated adapter sequences and low-quality reads. The trimmed reads were aligned to (1) *Elaeis guineensis* P5-build reference genome (NCBI BioProject accessions PRJNA192219, accessions ASJS0000000) (Singhet al., 2013) and, (2) gene models (Sanusi et al., 2018) ±1 kb flanking regions (define as ‘gene-wide’), using Bowtie2 v2.2.5 (Langmead and Salzberg, 2012) with --very-sensitive-local settings. Removal of duplicate reads and read coverage assessments were performed using SAMtools v1.9 (Li et al., 2009). Mean coverage was calculated in R. Histone-bound regions were identified using MACS2 v2.1.1.20160309 (Zhang et al., 2008) with options --q 0.05 specified for H3K4me3, while --broad --broad-cutoff 0.05 was conducted for H3K27me3. Gene-wide distribution patterns of pooled H3K4me3 and H3K27me3 peaks was plotted based on the average pileup reads of the peaks calculated for a 2 kb window around gene translation start positions and visualised using R. Average pileup is calculated by adding the pileup for individual genes at all positions of the detected peak and then dividing by the number of peaks. Prediction of TSS was made using SoftBerry tools. ChIP-Seq data can be accessed under GEO Accession No. GSE159142.

### RESULTS

Raw sequencing reads in fastq format was analysed with FastQC, which provides a quality assessment of the raw sequencing data. For all datasets, quality scores per base were generally good with a slight decrease towards the end of the reads. Approximately 1%-2% of the sequence reads were eventually discarded during the read trimming process due to the presence of adapter sequences and low-quality calls (Table 1).

**Low ChIP Recovery for ChIP-Seq Lead to Low Sequence Coverage and High Duplication**

In standard bioinformatics pipelines, sequencing reads are made informative by mapping to a reference genome using an aligner software. In this study, the trimmed reads were initially aligned to the P5-build of the *E. guineensis* reference genome (Singh et al., 2013). This step generated an alignment result in the Sequence Alignment/Map (SAM) format, which assigns specific genomic locations to each read. H3K4me3 and H3K27me3 ChIP-Seq sets in this study comprised the ChIP and their respective non-immunoprecipitated (denoted hereafter as control) sequencing data. While biological replicates A-J in the control group showed a consistent mapping percentage of more than 99%, mapping efficiencies of the immunoprecipitated biological replicates A-J were lower, ranging from 95%-99% for H3K4me3 and 88%-99% for H3K27me3 (Table 2). The
overall read mapping performance was acceptable (Figure 1), but duplication was high in each ChIP data compared to their corresponding control (Figure 2 and Table 2). The high amount of duplicates was probably due to the low amounts of ChIP DNA used to prepare the ChIP-Seq libraries. Duplicates also reflect the redundant reads resulting from polymerase chain reaction (PCR) overamplification, leading to an over-representation of ChIP-Seq reads at some genomic regions (Loh et al., 2017). The duplicates are technically identified as read pairs with their 5' ends mapped to the same genomic coordinates. Removal of duplicates was deemed necessary in this study due to the high duplication observed; otherwise peak calling or statistical analyses would be affected by the high duplication. Removal of duplicates thus, led to significant losses in reads, leaving only <20% of data for further analysis (Figure 2). The number of remaining reads after duplicates removal did not correlate with the amounts of starting material used to prepare the ChIP-Seq libraries (Tables 1 and 2). While control data mapped to 45%-62% of the genome, only 4%-60% of the genome was covered by ChIP reads (Table 1). On average, about 4-6 mapped reads per base were observed in control data, but only 2-4 and 2-5 read coverage was observed for H3K4me3- and H3K27me3-ChIP data respectively (Table 1). Overall, the immunoprecipitated ChIP-Seq data had high duplication rates and reduced coverage, which was most likely due to the low starting DNA input for library preparation, and is therefore not suitable for genome-wide analysis. However, it may still be useful for a targeted working hypothesis limited to selective regions, e.g., genes, repeats and intergenic regions, as reported in the following section.

![Figure 1](image1.png)

Figure 1. Average mapping percentages of ChIP-Seq data aligned to genome-wide and gene-wide E. guineensis pisifera reference databases. Error bars indicate standard deviation (H3K4me3; n=9 and H3K27me3; n=10). Values below indicate the mapping percentage of each bar.

![Figure 2](image2.png)

Figure 2. Presence of sequence duplicates in ChIP-Seq data aligned to genome-wide or gene-wide E. guineensis pisifera reference databases. Error bars indicate standard deviation (H3K4me3; n=9 and H3K27me3; n=10). Values below indicate the percentage of duplication for each bar.
TABLE 2. SUMMARY OF ChIP-SEQ DATASETS

| Histone modification | Library type | Biological replicate | ^aLibrary starting material | ^bMapping (%) | ^cDuplication (%) |
|----------------------|-------------|----------------------|----------------------------|---------------|-------------------|
|                      |             |                      | Genomewide | Gene-wide | Genomewide | Gene-wide |
| H3K4me3              | control     | A                    | 10 | 99.34 | 76.49 | 22.24 | 25.68 |
|                      |             | B                    | 10 | 99.25 | 75.76 | 13.59 | 18.30 |
|                      |             | C                    | 10 | 99.24 | 76.09 | 32.62 | 35.55 |
|                      |             | D                    | 10 | 99.37 | 77.03 | 47.31 | 50.59 |
|                      |             | E                    | 10 | 99.32 | 76.51 | 28.55 | 31.80 |
|                      |             | G                    | 10 | 99.22 | 76.21 | 27.62 | 30.39 |
|                      |             | H                    | 10 | 99.27 | 76.07 | 5.30  | 11.09 |
|                      |             | I                    | 10 | 99.32 | 78.40 | 25.23 | 32.84 |
|                      |             | J                    | 10 | 99.25 | 75.64 | 4.45  | 9.50  |
| H3K4me3              | ChIP        | A                    | 5.79 | 99.23 | 76.43 | 83.54 | 85.58 |
|                      |             | B                    | 2.10 | 99.18 | 77.12 | 84.60 | 87.28 |
|                      |             | C                    | 1.43 | 98.47 | 76.30 | 81.98 | 83.85 |
|                      |             | D                    | 3.81 | 96.86 | 74.65 | 97.23 | 97.79 |
|                      |             | E                    | 5.38 | 98.61 | 75.96 | 87.85 | 89.57 |
|                      |             | G                    | 4.01 | 98.53 | 76.21 | 97.05 | 97.70 |
|                      |             | H                    | 2.23 | 98.37 | 76.59 | 94.87 | 95.84 |
|                      |             | I                    | 7.29 | 99.05 | 76.50 | 70.12 | 72.60 |
|                      |             | J                    | 6.21 | 94.51 | 72.84 | 91.22 | 92.31 |
| H3K27me3             | control     | A                    | 10  | 99.31 | 75.48 | 5.63  | 12.00 |
|                      |             | B                    | 10  | 99.25 | 75.72 | 5.97  | 12.73 |
|                      |             | C                    | 10  | 99.24 | 75.70 | 6.63  | 12.14 |
|                      |             | D                    | 10  | 99.33 | 75.63 | 5.51  | 11.73 |
|                      |             | E                    | 10  | 99.33 | 75.93 | 7.69  | 12.94 |
|                      |             | F                    | 10  | 99.32 | 75.96 | 13.30 | 18.51 |
|                      |             | G                    | 10  | 99.29 | 75.78 | 5.35  | 11.77 |
|                      |             | H                    | 10  | 99.30 | 76.12 | 24.72 | 28.78 |
|                      |             | I                    | 10  | 99.32 | 75.75 | 8.23  | 13.25 |
|                      |             | J                    | 10  | 99.33 | 75.96 | 10.29 | 15.08 |
| H3K27me3             | ChIP        | A                    | 2.38 | 92.68 | 71.76 | 96.81 | 97.25 |
|                      |             | B                    | 2.68 | 96.17 | 74.07 | 90.53 | 91.74 |
|                      |             | C                    | 3.70 | 98.53 | 75.95 | 78.47 | 80.47 |
|                      |             | D                    | 2.71 | 87.90 | 67.79 | 83.23 | 83.25 |
|                      |             | E                    | 4.74 | 95.96 | 73.88 | 97.01 | 97.65 |
|                      |             | F                    | 10.10 | 99.24 | 76.32 | 23.39 | 28.48 |
|                      |             | G                    | 10.10 | 98.11 | 76.41 | 95.65 | 96.40 |
|                      |             | H                    | 1.12 | 97.96 | 75.93 | 84.44 | 86.19 |
|                      |             | I                    | 4.35 | 98.81 | 76.51 | 82.70 | 84.90 |
|                      |             | J                    | 3.81 | 98.35 | 76.25 | 90.34 | 91.79 |

Note: ^aDNA recovered from immunoprecipitation reactions used to prepare ChIP-Seq library.
^bPercentage of reads mapped to the reference databases using Bowtie2.
^cDuplicates presence in the mapped reads.
Gene-wide as an Alternative to Genome-wide Alignment

Read mapping on a genome-wide scale is a common practice in next-generation sequencing analysis and is widely acknowledged as a reliable approach when biases and limitations are minimal. With the limited number of unique ChIP reads, an alternative is to map onto a targeted region. In this study, a gene-wide database was built from genomic areas consisting of gene models from their translation start to stop ±1 kb of their flanking upstream and downstream regions. The upstream 1 kb region would contain part of the promoter regions important for gene regulation, including the transcription start sites located around 50-200 bp upstream of the start codon (Shahmuradov et al., 2017). These genomic, upstream, and downstream regulatory regions are usually covered by H3K4me3 and H3K27me3 modifications in other plants such as Arabidopsis, peach, and potato (de la Fuente et al., 2015; Yang et al., 2018; Zeng et al., 2019). This database generated for the oil palm was ~236 Mb, comprising 24,927 genes with their upstream and downstream regions. Predicted TSS are available for 13,016 genes. As a small portion of genes have multiple predicted TSS, a total of 14,353 TSS and 16,405 transcription termination sites (TTS) positions were identified. Gene-wide mapping rates were only slightly lower compared to genome-wide (Figure 1), with the majority of reads in ChIP and control datasets (~75%) mapping to genomic regions (Table 2). The remaining 25%, therefore, most likely mapped to intergenic regions. This observation also suggested that most H3K4me3 and H3K27me3 marks are focused on or near genic regions. These observations are in agreement with the properties of H3K4me3 and H3K27me3 marks reported for plants such as Arabidopsis, peach, and potato (de la Fuente et al., 2015; Yang et al., 2018; Zeng et al., 2019), i.e., both marks are highly associated with regulatory elements of genes and transcription. These results also suggested that the low coverage ChIP-Seq data could provide preliminary insights into H3K4me3 and H3K27me3 profiles on targeted oil palm genic regions. The current version of the oil palm genome assembly is a preliminary one (Chan et al., 2017); there are 40,360 genomic scaffolds, ranging in length from 1992 to 22,100 nt. By focusing on the relevant regions (genic and regulatory) of well-annotated genes, we avoid ‘dilution’ of the signal from spurious mapping of reads to incorrectly assembled regions. This approach has improved the mapping precision of ChIP-Seq reads on target genic regions. However, the ChIP method on oil palm spears will require further optimisation to increase the DNA yield recovery, most likely by pooling an increased number of immunoprecipitation reactions.

Histone Marks Highly Associate with Predicted TSS of Genes

After mapping analysis, genomic regions with ChIP enrichment over background noise were detected as peaks using the Model-based Analysis of ChIP-Seq (MACS2) peak-calling software. These ‘called peaks’ represent biologically relevant histone modification sites. Peak detection parameters were optimised based on the binding properties of the specific histone modification, i.e., a ‘sharp’ feature for H3K4me3 and a ‘broad’ feature for H3K27me3 at 0.05 false discovery rate (FDR) cut-off. Using the MACS2 peak caller, 300 to 4000 peaks associated with H3K4me3 and H3K27me3 were identified, albeit inconsistently among the biological replicates, likely due to their low coverage (Table 3). The H3K4me3 and H3K27me3 peak distribution profiles indicated that their abundance was predominant near the predicted TSS of genes (Figure 3). The oil palm H3K4me3 profile showed a narrow distribution with its peak centered at the predicted TSS (located on average 100 bp upstream of translation start positions) (Figure 3a), while H3K27me3 peak profile was much broader and was distributed over the gene region from the TSS (Figure 3b). These profiles were generally consistent with H3K4me3 and H3K27me3 patterns in other plants (de la Fuente et al., 2015; Yang et al., 2018; Zeng et al., 2019). These profiles suggest that low coverage ChIP-Seq data could still reveal prominent characteristics of these histone modification profiles in oil palm.

| Histone modification | Biological replicate | No. of ChIP-Seq peaks |
|----------------------|----------------------|-----------------------|
| H3K4me3             | A                    | 2356                  |
|                     | B                    | 3623                  |
|                     | C                    | 504                   |
|                     | D                    | 743                   |
|                     | E                    | 1584                  |
|                     | G                    | 296                   |
|                     | H                    | 1168                  |
|                     | I                    | 3910                  |
|                     | J                    | 580                   |
| H3K27me3            | A                    | 739                   |
|                     | B                    | 707                   |
|                     | C                    | 1227                  |
|                     | D                    | 1436                  |
|                     | E                    | 727                   |
|                     | F                    | 1581                  |
|                     | G                    | 472                   |
|                     | H                    | 762                   |
|                     | I                    | 1348                  |
|                     | J                    | 918                   |
DISCUSSION

In this study, H3K4me3 and H3K27me3 ChIP-Seq libraries were prepared with low amounts of DNA from pooled ChIP reactions on oil palm leaf chromatin. The initial low DNA amounts generated ChIP-Seq data with low numbers of unique mapped reads at low coverage. This was even after pooling the DNA from six to eight ChIP reactions, which yielded between 1 and 10 ng DNA. It thus, appears that the number of ChIP reactions per sample will need to be scaled up significantly. ChIP yields obtained were lower than those obtained for maize, where a pool of three ChIP reactions yielded about 50 ng DNA (Oka et al., 2017). As the purification of immunoprecipitated DNA is a necessary step before library preparation, assessing the efficiency of purification agents could help in maximising DNA recovery without compromising the quality. Among several commercial purification kits and reagents, DNA recovery was higher when phenol-chloroform was used (Zhong et al., 2017).

Nevertheless, using targeted regions in analysing such imperfect datasets has been demonstrated to be a practical approach when a genome-wide interrogation is not feasible. This approach is in line with published ChIP-Seq guidelines when dealing with low-quality data validated based on several quantitative quality metrics (Landt et al., 2012; Nakato and Shirahige, 2017). About 75% of the data could be mapped to a database comprising genes and their upstream and downstream regions, suggesting that these histone marks are enriched mostly at the genic areas of oil palm. Although read mapping to a smaller database does not always increase alignment sensitivity and specificity, it still holds several advantages. In practice, assigning reads to only selected biologically relevant positions could provide a quick estimate when addressing urgent biological problems, a QC step in a pilot study, or in any setting when focusing only on differentially expressed genes is justified.

Besides, with large plant genome sizes rich with repeats, the use of a smaller functional database reduces the time for computational analysis and saves space for the storage of processing files. The different ChIP-Seq read coverage among biological replicates contributes to variation in numbers and regions of ChIP enrichment, thus, compromising reproducibility. However, enrichment of H3K4me3 and H3K27me3 marks near TSS is supported by the similar characteristics of these profiles in Arabidopsis, peach and potato (de la Fuente et al., 2015; Yang et al., 2018; Zeng et al., 2019), indicating that the ChIP assay worked for oil palm even though sequence coverage was low. Scaling up the immunoprecipitation reactions or improving on the ChIP yield recovery would be required in future ChIP-Seq studies. In Arabidopsis, predicted enhancers were characterised and validated using open chromatin signatures, which include H3K27ac and H3K27me3 histone modification patterns and non-coding RNA with GUS-based reporter assays (Zhu et al., 2015). Therefore, histone enrichment profiles could be used to verify predicted TSS locations and improve the annotation of the oil palm genome.

The number of sequencing reads needed to reach a reasonable coverage varies depending on the types of histone binding patterns of the various histone modification types. At least 20 million reads are required for sharp binding patterns, while 60 million reads are suggested for broad binding patterns (Landt et al., 2012; Nakato and Shirahige, 2017). Read coverage is deemed sufficient when the saturation point could be established or when the same histone enrichment sites are repeatedly identified from additional sequencing. A way to improve read coverage is by evaluating the antibody’s quality to capture histone-bound regions. The efficiency of the antibody can be tested by conducting a titration of chromatin experiment to a fixed amount of antibody. One can evaluate whether diluting chromatin would improve the precipitation efficiency, thus, assessing the

![Figure 3. Gene-wide distribution pattern of ChIP-Seq peaks near the transcription start sites. (a) narrow distribution of H3K4me3 profile; (b) broad distribution of H3K27me3 profile. ‘0’ on the x-axis denotes the ATG start codon.](image-url)
presence of inhibitory factors within the chromatin (Haring et al., 2007). Monoclonal and polyclonal antibodies, which respectively recognise single and multiple epitopes, affect the degree of enrichment and background noise differently. Testing out several commercial antibodies is also advisable as different brands may exhibit different efficiencies. The number of individual ChIP reactions needs to be increased to achieve the required amount for library construction. However, low amounts of ChIP DNA may provide useful sequencing data if an appropriate sequencing library preparation method explicitly for low DNA input is used. Recently, as low as 0.01 ng DNA, which is ≥100x lower than the starting amount used in this study, can be used for constructing ChIP-sequencing libraries (reviewed by Dahl and Gilfillan, 2015). The reduction of tagging (Schmidl et al., 2007). Monoclonal and polyclonal antibodies, which respectively recognise single and multiple epitopes, affect the degree of enrichment and background noise differently. Testing out several commercial antibodies is also advisable as different brands may exhibit different efficiencies. The number of individual ChIP reactions needs to be increased to achieve the required amount for library construction. However, low amounts of ChIP DNA may provide useful sequencing data if an appropriate sequencing library preparation method explicitly for low DNA input is used. Recently, as low as 0.01 ng DNA, which is ≥100x lower than the starting amount used in this study, can be used for constructing ChIP-sequencing libraries (reviewed by Dahl and Gilfillan, 2015). The reduction of tagging (Schmidl et al., 2015). The reduction of library preparation steps compared to published ChIP-Seq protocols therefore allows for samples with low cell input, as yield losses throughout the procedure is also reduced. With emerging new technologies, the ChIP-Seq assay is now even feasible at the single-cell level (Clark et al., 2016).

CONCLUSION

This is the first ChIP-Seq report on histone modification patterns for the oil palm E. guineensis. Although preliminary, the primary limitation in this study was the low coverage and high duplication of the H3K4me3 and H3K27me3 ChIP-Seq data due to low ChIP recovery. However, the data was still able to provide some preliminary insights on these histone profiles in oil palm through a gene-wide mapping approach, followed by peak calling. The traditional genome-wide mapping approach used for analysing ChIP-Seq data is not suitable for partially assembled genomes combined with the low read yield situation. Restriction to the functionally relevant areas in oil palm genome resulted in H3K4me3 and H3K27me3 profiles consistent with the profiles of these histone modifications in plants such as Arabidopsis, peach and potato. This study’s findings and recommendations may be useful for other plant species with incomplete genome sequence information to accurately identify histone profiles without requiring intensive computational power.

ACKNOWLEDGEMENT

We are grateful to the Director-General of MPOB for permission to publish this study. We thank members of the Epigenetics Group, Advanced Biotechnology and Breeding Centre, MPOB for their invaluable technical support. We also thank Chan Kuang Lim and Nik Shazana Nik Sanusi from the Bioinformatics Unit, MPOB. Our appreciation also goes to the Clonal Propagation Group, MPOB for the clonal palms used in this study. This study was funded by the MPOB.

REFERENCES

Baulcombe, D C and Dean, C (2014). Epigenetic regulation in plant responses to the environment. Cold Spring Harb. Perspect. Biol., 6: 1-19.

Berenguer, E; Bárány, I; Solís, M-T; Pérez-Pérez, Y; Risueño, M C and Testillano, P S (2017). Inhibition of histone H3K9 methylation by BIX-01294 promotes stress-induced microspore totipotency and enhances embryogenesis initiation. Front. Plant Sci., 8: 1-19.

Chan, K-L; Tatarinova, T V; Rosli, R; Amiruddin, N; Azizi, N; Halim, M A A; Sanusi, N S N M; Jayanthi, N; Ponomarenko, P; Triska, M; Solovyev, V; Firdaus-Raah, M; Sambanthamurthi, R; Murphy, D and Low, E-T L (2017). Evidence-based gene models for structural and functional annotations of the oil palm genome. Biol. Direct, 12(1): 21. DOI: 10.1186/s13062-017-0191-4.

Clark, S J; Lee, H J; Smallwood, S A; Kelsey, G and Reik, W (2016). Single-cell epigenomics: Powerful new methods for understanding gene regulation and cell identity. Genome Biol., 17: 1-10.

Dahl, J A and Gilfillan, G D (2018). How low can you go? Pushing the limits of low-input ChIP-seq. Brief. Funct. Genomics, 17: 89-95.

de la Fuente, L; Conesa, A; Lloret, A; Badenes, M L and Ríos, G (2015). Genome-wide changes in histone H3 lysine 27 trimethylation associated with bud dormancy release in peach. Tree Genet. Genomes, 11: 45. DOI: 10.1007/s11295-015-0869-7.

Dodt, M; Roehr, J T; Ahmed, R and Dieterich, C (2012). FLEXBAR-flexible barcode and adapter processing for next-generation sequencing platforms. Biology (Basel), 1: 895-905.

Engelhorn, J; Blanvillain, R; Kröner, C; Parrinello, H; Rohmer, M; Posé, D; Ott, F; Schmid, M and Carles, C (2017). Dynamics of H3K4me3 chromatin marks prevails over H3K27me3 for gene regulation during flower morphogenesis in Arabidopsis thaliana. Epigenomes, 1: 8. DOI: 10.3390/epigenomes1020008.
Gan, E-S; Xu, Y and Ito, T (2015). Dynamics of H3K27me3 methylation and demethylation in plant development. *Plant Signal. Behav.*, 10(9): e1027851.

Haring, M; Offermann, S; Danker, T; Horst, J; Peterhansel, C and Stam, M (2007). Chromatin immunoprecipitation: Optimization, quantitative analysis and data normalization. *Plant Methods*, 3: 11. DOI: 10.1186/1746-4811-3-11.

Kaufmann, K; Muñoz, J M; Østerås, M; Farinelli, L; Krajevski, P and Angenent, G C (2010). Chromatin immunoprecipitation (ChiP) of plant transcription factors followed by sequencing (ChiP-SEQ) or hybridization to whole genome arrays (ChiP-CHIP). *Nat. Protoc.*, 5(3): 457-472.

Landt, S G; Marinov, G K; Kundaje, A; Kheradpour, P; Pauli, F; Batzoglou, S; Bernstein, B E; Bickel, P; Brown, J B; Cautyting, P; Chen, Y; DeSalvo, G; Epstein, C; Fisher-Aylor, K; Euskinicro, G; Gerstein, M; Gertz, J; Hartemink, A J; Hoffman, M M; Iyer, V R; Jung, Y L; Karmakar, S; Kellis, M; Kharchenko, P V; Li, Q; Liu, T; Liu, S; Ma, L; Milosavljevic, A; Myers, R M; Park, P J; Pazzin, M J; Perry, M D; Raha, D; Reddy, T E; Rozowsky, J; Shores, N; Sidow, A; Slattery, M; Stamatoyannopoulos, J A; Tolstorukov, M Y; White, K P; Xi, S; Farnham, P J; Lieb, J D; Wold, B J and Snyder, M (2012). ChiP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.*, 22: 1813-1831.

Langmead, B and Salzberg, S L (2012). Fast gapped-read alignment with Bowtie2. *Nat. Methods*, 9: 357-359.

Li, H; Handsaker, B; Wysoker, A; Fennell, T; Ruan, J; Homer, N; Marth, G; Abecasis, G and Durbin, R (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, 25: 2078-2079.

Liu, E T; Pott, S and Huss, M (2010). Q&A: ChIP-seq technologies and the study of gene regulation. *BMC Biol.*, 8: 56. DOI: 10.1186/1741-7007-8-56.

Loh, Y-H E; Feng, J; Nestler, E and Shen, L (2017). Bioinformatic analysis for profiling drug-induced chromatin modification landscapes in mouse brain using ChIP-seq data. *Bio Protoc.*, 7(3): e2123. DOI: 10.21769/BioProtoc.2123.

Lu, Y; Zhou, D-X and Zhao, Y (2020). Understanding epigenomics based on the rice model. *Theor. Appl. Genet.*, 133: 1345-1363.

Nakato, R and Shirahige, K (2017). Recent advances in ChIP-seq analysis: From quality management to whole-genome annotation. *Brief. Bioinform.*, 18: 279-290.

Oka, R; Zicola, J; Weber, B; Anderson, S N; Hodgman, C; Gent, J I; Wesselingk, J-J; Springer, N M; Hoeofsloot, H C J; Turck, F and Stam, M (2017). Genome-wide mapping of transcriptional enhancer candidates using DNA and chromatin features in maize. *Genome Biol.*, 18(1): 137. DOI: 10.1186/s13059-017-1273-4.

Ong-Abdullah, M; Ordway, J M; Jiang, N; Ooi, S-E; Kok, S-Y; Sarpan, N; Azimi, N; Hashim, A T; Ishak, Z; Rosli, S K; Malikke, F A; Bakar, N A A; Marjuni, M; Abdullah, N; Yaakub, Z; Amiruddin, M D; Nookiah, R; Singh, R; Low, E-T L; Chan, K-L; Azizi, N; Smith, S W; Bacher, B; Budiman, M A; Van Brunst, A; Wischneyer, C; Beil, M; Hogan, M; Lakey, N; Lim, C-C; Arulando, X; Wong, C-K; Choo, C-N; Wong, W-C; Kwan, Y-Y; Alwee, S S R S; Sambanthamurthi, R and Martienssen, R A (2015). Loss of Karma transposon methylation underlies the mantled somaclonal variant of oil palm. *Nature*, 525: 533-537.

Ong-Abdullah, M; Ordway, J M; Jiang, N; Ooi, S-E; Mokri, A; Kok, S Y; Sarpan, N; Azimi, N; Hashim, A T; Ishak, Z; Rosli, S K; Nookiah, R; Singh, R; Low, E-T L; Sachdeva, M; Smith, S W; Lakey, N; Martienssen, R A and Sambanthamurthi, R (2016). Tissue culture and epigenetics. *The Planter*, 92: 741-749.

Prakash, K and Fournier, D (2018). Evidence for the implication of the histone code in building the genome structure. *Biosystems*, 164: 49-59.

Ranawaka, B; Tanurdzic, M; Waterhouse, P and Nain, F (2020). An optimised chromatin immunoprecipitation (ChiP) method for starchy leaves of *Nicotiana benthamiana* to study histone modifications of an allotetraploid plant. *Mol. Biol. Rep.*, 47: 9499-9509.

Sanusi, N S N M; Rosli, R; Halim, M A A; Chan, K-L; Nagappan, J; Azizi, N; Amiruddin, N; Tatarinova, T and Naim, F (2020). An optimised chromatin immunoprecipitation (ChIP) protocol for histone modification in oil palm. *J. Oil Palm Res.*, 30: 242-250.

Schmidl, C; Rendeiro, A F; Sheffield, N C and Bock, C (2015). ChIPmentation: Fast, robust, low-input ChIP-seq for histones and transcription factors. *Nat. Methods*, 12: 963-965.

Shahmuradov, I A; Umarov, R K and Solovyev, V V (2017). TSSPlant: A new tool for prediction of plant Pol II promoters. *Nucleic Acids Res.*, 45(8): e65.
K-L; Halim, M A; Azizi, N; Nagappan, N; Bacher, B; Lakey, N; Smith, S W; He, D; Hogan, M; Budiman, M A; Lee, E K; DeSalle, R; Kudrna, D; Goicoechea, J L; Wing, R A; Wilson, R K; Fulton, R S; Ordway, J M; Martienssen, R A and Sambanthamurthi, R (2013). Oil palm genome sequence reveals divergence of interfertile species in old and new worlds. Nature, 500: 335-339.

Wollmann, H and Berger, F (2012). Epigenetic reprogramming during plant reproduction and seed development. Curr. Opin. Plant Biol., 15: 63-69.

Yang, Z; Qian, S; Scheid, R N; Lu, L; Chen, X; Du, X; Lv, X; Boersma, M D; Scalf, M and Smith, L M (2018). EBS is a bivalent histone reader that regulates floral phase transition in Arabidopsis. Nat. Genet., 50: 1247-1253.

Zeng, Z; Zhang, W; Marand, A P; Zhu, B; Buell, C R and Jiang, J (2019). Cold stress induces enhanced chromatin accessibility and bivalent histone modifications H3K4me3 and H3K27me3 of active genes in potato. Genome Biol., 20: 1-17.

Zhang, W; Garcia, N; Feng, Y; Zhao, H and Messing, J (2015). Genome-wide histone acetylation correlates with active transcription in maize. Genomics, 106(4): 214-220.

Zhang, X; Bernatavichute, Y V; Cokus, S; Pellegrini, M and Jacobsen, S E (2009). Genome-wide analysis of mono-, di- and trimethylation of histone H3 lysine 4 in Arabidopsis thaliana. Genome Biol., 10(6): R62. DOI: 10.1186/gb-2009-10-6-r62.

Zhang, Y; Liu, T; Meyer, C A; Eeckhoute, J; Johnson, D S; Bernstein, B E; Nussbaum, C; Myers, R M; Brown, M; Li, W and Liu, X S (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol., 9(9): R137. DOI: 10.1186/gb-2008-9-9-r137.

Zhong, J; Ye, Z; Lenz, S W; Clark, C R; Bharucha, A; Farrugia, G; Robertson, K D; Zhang, Z; Ordog, T and Lee, J-H (2017). Purification of nanogram-range immunoprecipitated DNA in ChIP-seq application. BMC Genomics, 18(1): 985. DOI: 10.1186/s12864-017-4371-5.

Zhu, B; Zhang, W; Zhang, T; Liu, B and Jiang, J (2015). Genome-wide prediction and validation of intergenic enhancers in Arabidopsis using open chromatin signatures. Plant Cell, 27: 2415-2426.

Zong, W; Zhong, X; You, J and Xiong, L (2013). Genome-wide profiling of histone H3K4-trimethylation and gene expression in rice under drought stress. Plant Mol. Biol., 81: 175-188.

Zulkifli, Y; Norziha, A; Naqiuddin, M H; Fadila, A M; Nor Azwani, A B; Suzana, M; Samsul, K R; Ong-Abdullah, M; Singh, R; Parveez, G K A and Kushairi, A (2017). Designing the oil palm of the future. J. Oil Palm Res., 29: 440-455.