Deliverable 1.3 BOOSTER maize pancistrome generation

Generating MOA- and eMetylation-seq data for maize F1 hybrids of the SeqOccIn project

Deliverable 1.3 includes the pancistrome TF footprint data generated for the maize germplasm within Booster. This data was generated on leaf blade tissue of nested F1 omni-hybrid lines. The nested F1 hybrid population was generated by crossing a common tester line (Oh43) as mother with 29 diverse paternal lines of the “SeqOccIn” population. INRAE's "Sequencage Occitanie Innovation" (or SeqOccIn) project, jointly led by Genotoul's GeT and Bioinfo platforms, is part of the AP01 "Stimulating innovation" axis of the "Regional Research and Innovation Platforms" call for projects of the Occitanie region on the FEDER-ESF MIDI-PYRENEES ET GARONNE 2014-2020 Operational Programme. Part of the SeqOccIn project includes generating high-quality genome assemblies of diverse European maize germplasm. BOOSTER partner UDUS was granted pre-publication access to both germplasm and genome assembly data of the SeqOccIn project under MTA. This includes the ability to generate phenotyping data (Drought-response), MOA-seq, and eMethylation-seq data using SeqOccIn project resources funded in a separate grant. This specifically includes datasets DS.01.10, DS.01.06, DS.01.10, and DS.01.13a (see D6.2, Data Management Plan, and its updated version, D6.3), which have been generated for BOOSTER. Due to the 3rd party funding and pre-publication access of these SeqOccIn project resources, the public release of BOOSTER datasets DS.01.10, DS.01.06, DS.01.10, and DS.01.13a will be under embargo as described in the BOOSTER data management plan (D6.3, Appendix).

Plant material

To facilitate the generation of the maize pancistrome, we created 29 F1 hybrid lines. The parental lines used for these crosses are listed in D1.1. The genotype for the maternal genome was chosen based on drought susceptibility and available high quality genome assembly. In contrast the paternal genotype was based on genetic diversity, available high quality assembled genome, and lastly contrasting drought susceptibility.

The F1 hybrids were grown as described in D1.2 the trial experiment using the established conditions. Briefly all seeds of F1 hybrids were pre-germinated for 72h to assure all plants shared a comparable developmental stage. All pots were automatically filled to a similar level. Each replicate (pool of 4 pots and 4 plants per pot) was grown in a randomized block design, and the experiment repeated 3 times to yield the final three replicates.

Plants were grown side by side in greenhouses, under long-day conditions (16h day/8h night, 28-30℃) for approximately 26 days until 75% of the plants per pot showed the formation of the leaf 4 auricle. Plants were then randomized and 12 pots per treatment were grown with or without periodic watering through a bottom drench system for 120h (slightly longer than in the trial with 113h). Plants were then harvested and the leaf blades of the oldest leaf without a yet formed auricle were immediately frozen in liquid nitrogen.

Relative water content and soil capacity measurements

Relative water content (RWC) measurements were conducted in the same way as for the pilot experiment (see D1.2): In brief, at harvest a 3 cm piece of rolled leaves was cut from the stalk between leaf 3 and leaf 4. Of this piece, the fresh weight (FW), turgor weight (TW) and dry weight (DW) were determined and the RCW was calculated as (FW-DW)/(TW-DW)*100. Considering all three replicates, all well-watered samples had a RWC median of around 90% or higher, while we observed a diverse response in the drought samples, with RWC median values from around 77% to 55%. For all lines we found the difference in RWC between well-watered and drought treated samples to be significant (ANOVA followed by Tukey post-hoc test, $p < 0.05$). All RWC data has been deposit in DS.01.06.

MOA-seq assay control and quality evaluation

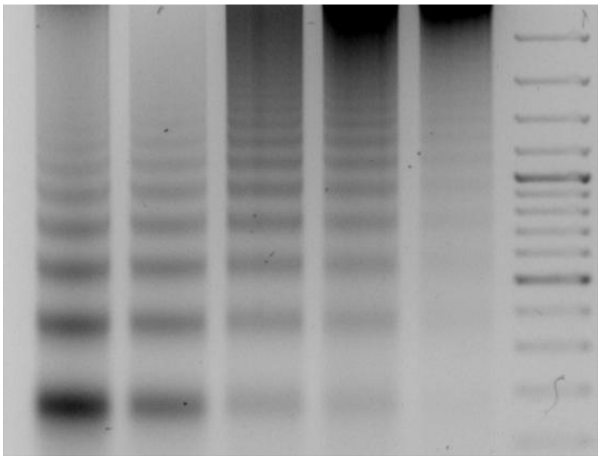

Figure 1: Standard example for differential MOA assay on maize B73 leaf samples used to standardize MOA samples. Decreasing MNase concentrations from left to right: 50U, 25U, 12.5U, 6.25U, 3.125U.

To test the optimal MNase digestion level we performed a partial MNase digestion on B73 leave samples (75 mg finely homogenized sample, Fig. 1) and sequenced and analysed all digestion levels individually. We specifically aimed to measure the range of MNase concentrations with the highest signal-to-noise ratio for optimal conditions (see also Savadel and Hartwig et al. 2021). To compare single-to-noise ratio of the different digest levels we compared MOA coverage at peak locations (same peak location for each sample) vs. coverage at randomly selected gene exon locations. We determined the optimal digestion level at between 12.5U (M4) and 6.25U (M3). We aimed for similar digest levels in all samples (see Fig.2 for an example).

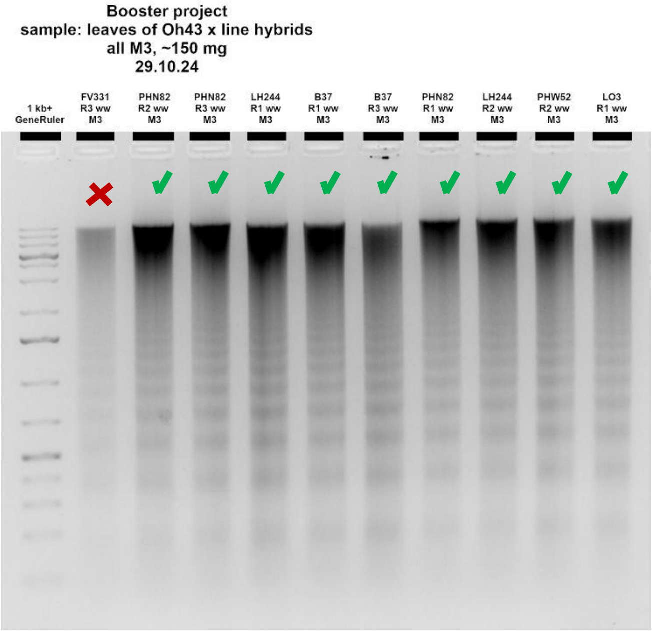

Figure 2: Example picture of DNA form MOA digestion used for library production and sequencing separated on a 1.5% TAE agarose gel. Ten MOA samples from F1 hybrids are shown. The lane on the very contains 1ul of a 1 Kb size standard (New England Biolabs, stronger bands correspond to 0.5 Kb, 1Kb and 3kb).

MOA-seq sample preparation and sequencing

To perform the MOA assay, we used finely homogenized (150 mg) input material, performed MNase digestions with 6.25U as the default starting concentration, and selected samples with digest pattern comparable to the test examples (M3-M4 level) for library production and sequencing. Appart from these optimisations, MOA-seq sample and library preparation was performed using following the procedure described in Savadel et al. 2021. A total of 180 library samples we generated so far (KWS F1 hybrid analysis not yet finished). This includes 30 lines, 3 replicates, 2 conditions (well-watered, drought stressed), and 1 assay (MOA-seq). For each library, only samples which yielded at least 1 ng/ul pre-pooling concentration, and which passed spot-checking with bioanalyzer were used (others were repeated).

MOA-seq data analysis

Data analysis was performed as described in our proof-of-concept study (Engelhorn et al. biorxiv) with minor modification: Reads were filtered using SeqPurge (v2022-07-15) with parameters "-min_len 20 -qcut 0" (Sturm et al. 2016). Due to the short fragment length in MOA, read pairs almost completely overlapped. MOA-seq paired-end reads were merged into single-end reads, including base quality score correction, using flash (v1.2.11) (Magoc and Salzberg, 2011) with default parameters. Diploid genomes were created by concatenating the Oh43 genome with the respective paternal genome (SeqOccIn genomes), with the addition of two maize US-NAM genome (CML52, CML228, Hufford et al. 2021). Reads were mapped to the diploid genome using STAR (v2.7.7a). As STAR was originally designed to map RNA, we set the flag --alignIntronMax 1 for DNA (no introns allowed) as well as parameters "--outSAMmultNmax 2, --winAnchorMultimapNmax 100, and -outBAMsortingBinsN 5 (Dobin et al. 2013). We generated two data-sets, one where reads were only allowed to map one in the diploid genome (mapping quality 255, used to generate MPs and AMPs data) and one where we allowed reads to map exactly twice but not in more instances with double mapping reads being randomly assigned one of the two positions (used for visualization and overall peak coverage data). Format conversion and the calculation of the average mapped fragment length (AMFL) was done using SAMtools (v1.9) (Li et al. 2009). The effective genome size was calculated using unique-kmers.py (https://github.com/dib-lab/khmer/) with AFML and respective genome fasta as inputs. The deeptools (v 3.5.0) function bamCoverage was used to generate normalized (reads per genome coverage, RPGC) bedgraph files of full length read data.
To generate fragment-center tracks, bam files were converted to bed format using bamToBed (Quinlan and Hall 2010) and each mapped read was shortened to 20 bp centered around the middle of the read. Read shortening was performed using awk: for reads with uneven number of bases, the middle base was taken, and then the read was extended 10 bp to each site. For reads with even numbers of bases, one of the two middle bases was chosen randomly and the reads were extended 10 bp to each site. The function genomeCoverageBed of Bedtools (v2.29.0) was then used to convert the bed files to bedgraph, scaled by the quotient of the effective genome size and the number of uniquely mapped reads (corresponding to RPGC of deeptools bamCoverage). BigWig files for visualization were generated using bedGraphToBigWig (Kent et al. 2010).

Quality control evaluation

Quality control was performed using FastQC (Andrews S (2010)."FastQC: a quality control tool for high throughput sequence data", http://www.bioinformatics.babraham.ac.uk/projects/fastqc) and summarised using MultiQC (Ewels et al., 2016). We aimed at a sequencing depth from at least 85 to an optimal 100 million raw read-pairs per replicate (300 million per F1) passing quality filters (mean phred >20). We recently demonstrated in a proof-of-concept study that 3x100 million per F1 lines is sufficient, both to achieve a high enough sequencing depth for sensitive peak called using merged read files, and also to have enough coverage for the individual replicates to validate potential TF binding bias to specific haplotypes in the F1 background. We observed that we achieved the required sequencing depth for all samples, with the majority of samples reaching optimal sequence depth levels (Fig. 3). It should be noted that the duplication rate for MOA-seq reads indicated by the QC algorithm is not a sign of low quality but a characteristic of MOA-seq data. Read-pairs which share the same start and stop positions when comparing both read-mates are normally considered potential PCR duplicates. However, MOA reads naturally stack at similar positions as they represent TF footprint regions, and due to the short-read length of MOA fragments both read-mates overlap in almost all MOA reads. The MOA library protocol uses only a minimal amount of PCR cycles (4-5 total cycles) which makes PCR duplicates intrinsically rare when pooling a somewhat large number of libraries. As such we consider the total read number, not only unique reads for the QC analysis of MOA sequencing.

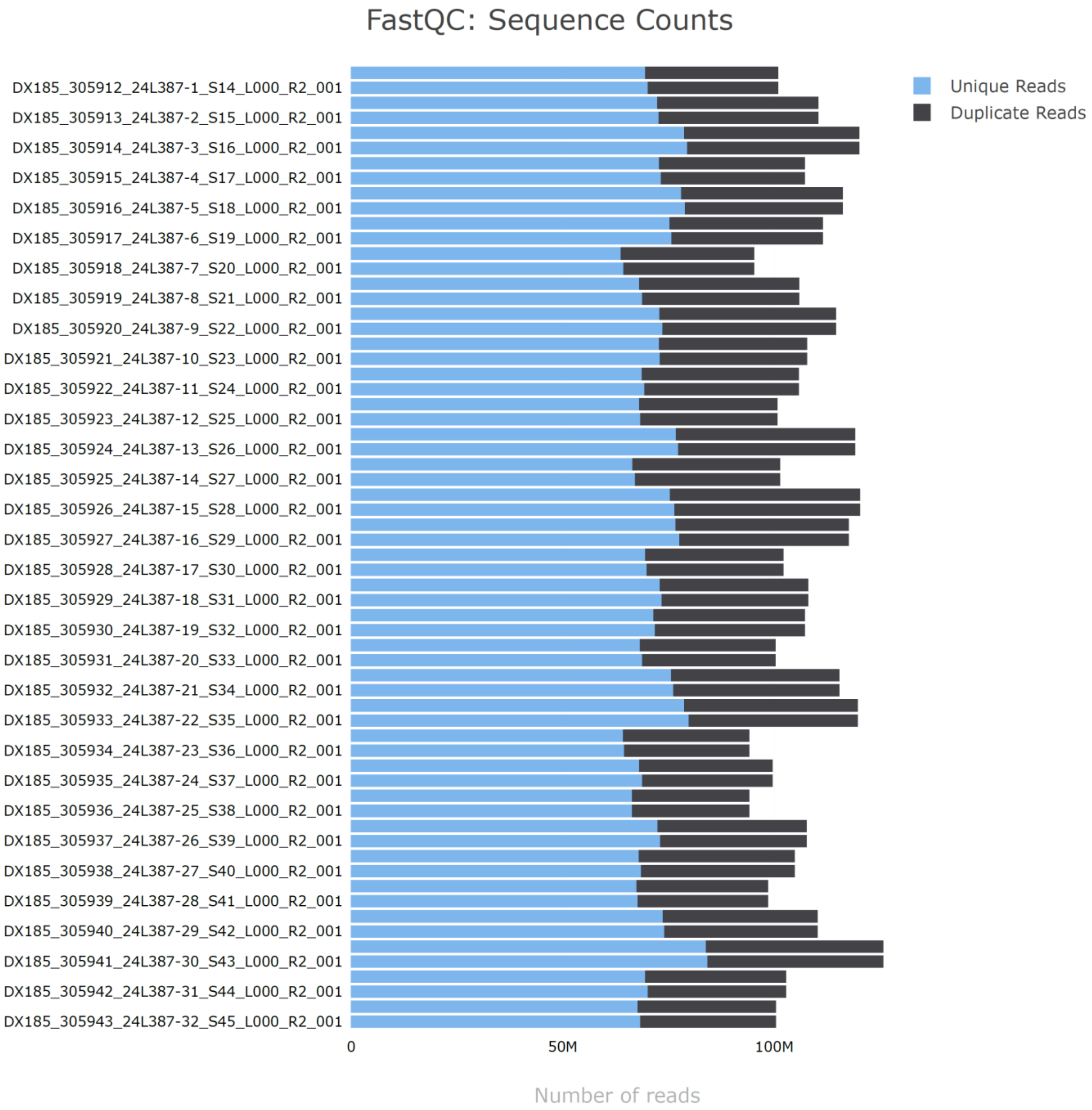

Figure 3: Example of MOA-seq sequencing depth results for a representative Illumina NovaSeq S4 lanes with 30 samples of all the sequenced MOA libraries. A total of 30 out of 30 samples reached the calculated minimum cutoff of >80 million raw reads per replicate, and 28 out of 30 samples reach the calculated optimal sequencing depth of >100 million raw read pairs.

As expected for accurately quantified pooled input library, we observed very high sequencing quality far above the required cutoff for Illumina NovaSeq platform. We measured consistent base quality of above a phred value of > 30 (99.9% likelihood of correct base call) for more than 95% of all MOA samples (see Fig. 4 for a representative example).

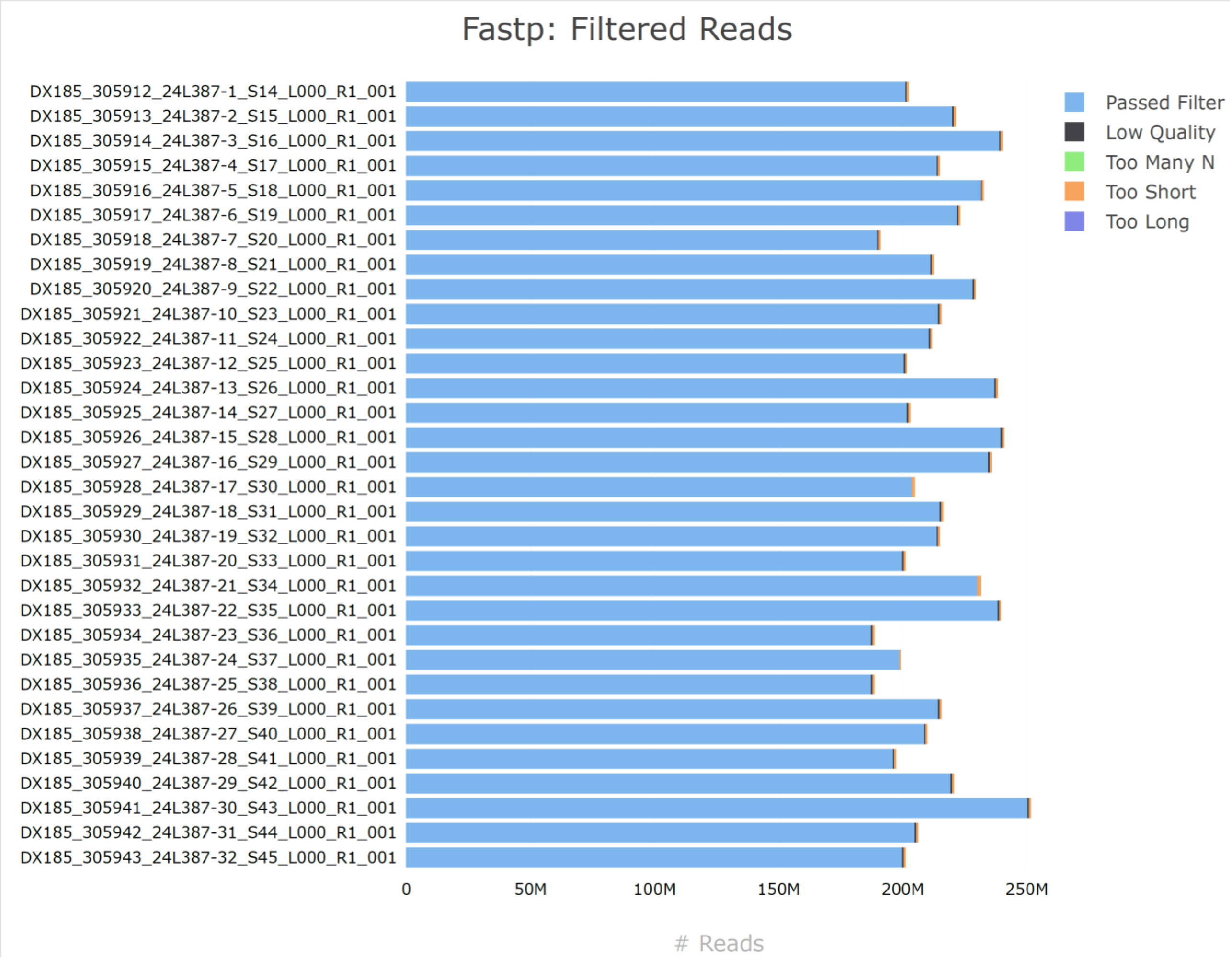

Figure 4: Example of read numbers remaining after quality filtering for sequenced MOA libraries of one replicate (numbers of read mates 1 and read 2 added together). Over 95% of reads are above the phred score of 30 (blue coloured, bases with 99.9% likelihood of being correct).

In order to further evaluate the quality of MOA-seq reads, we analysed the mapping statistics of the MOA-reads from all samples to their concatenated F1 hybrid genome, which included both parental haplotypes of the F1 (for example the reads from the Oh43 x GF111 F1 were mapped to the concatenated Oh43 genome combined with the GF 111 genome). We found a unique mapping rate between 20-25% for all samples. For example, for the Oh43 x GF111 well-watered replicate 3 from a total of 93,800,458 input read pairs, 21,904,231 (23.55%) of reads mapped uniquely (Table 1). This number matches the expected mapping rate achieved for maize F1 hybrids in our previous proof-of-concept study (19.5-26.5%, Engelhorn et al. biorxiv)). We note that relatively low unique mapping rate is due to the F1 nature of the target genome, containing to mostly conserved haplotypes. This is illustrated by the high number of reads mapping to multiple loci (1-2 identical mapping locations) of 24.9%.

Oh43 x GF111 ww, rep3	
Number of input reads |	93800458
Average input read length |	54
UNIQUE READS:	
Uniquely mapped reads number |	21904231
Uniquely mapped reads % |	23.35%
Average mapped length |	58.63
Number of splices: Total |	0

Number of splices: Annotated (sjdb) |	0
Number of splices: GT/AG |	0
Number of splices: GC/AG |	0
Number of splices: AT/AC |	0
Number of splices: Non-canonical |	0
Mismatch rate per base, % |	0.04%
Deletion rate per base |	0.00%
Deletion average length |	1
Insertion rate per base |	0.00%
Insertion average length |	1.27
MULTI-MAPPING READS:	
Number of reads mapped to multiple loci |	23356819
% of reads mapped to multiple loci |	24.90%
Number of reads mapped to too many loci |	24907338
% of reads mapped to too many loci |	26.55%
UNMAPPED READS:	
Number of reads unmapped: too many mismatches |	0
% of reads unmapped: too many mismatches |	0.00%
Number of reads unmapped: too short |	232405
% of reads unmapped: too short |	0.25%
Number of reads unmapped: other |	23399665
% of reads unmapped: other |	24.95%
CHIMERIC READS:	
Number of chimeric reads |	0
% of chimeric reads |	0.00%

Table 1: Representative example of MOA mapping statistics. Reads were mapped to the concatenated Oh43 x GF111 genome using the STAR mapping suite (for parameters see method details) for the Oh43 x GF111 well-watered replicate 3.

Furthermore, early evaluation of the MOA-seq results indicates that the MOA data are of high quality. We observed good overlap of our MOA footprint results, with previous known binding sites of TFs, including the major Brassinosteroid phytohormone TF BZR1 (Hartwig et al., 2023). A representative example of such an overlap of MOA footprint and the BZR1 ChIP-seq binding data are shown in Figure 5a-b.

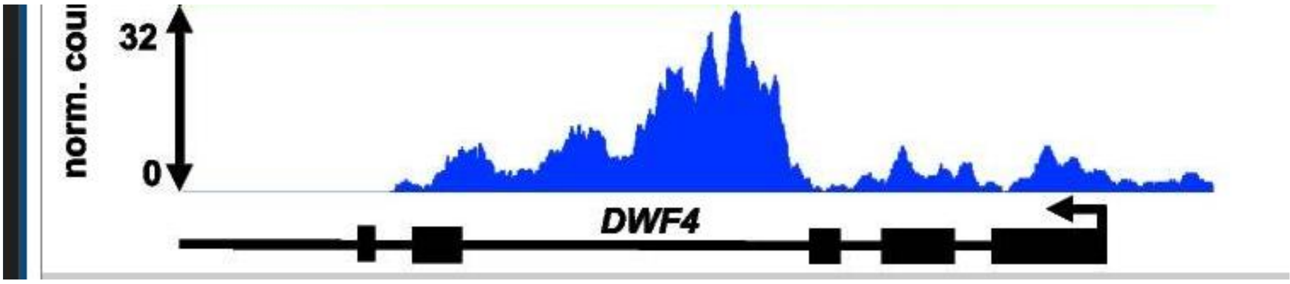

Figure 5a: BZR1 ChiP-seq results of the ZmDWF4 gene analyzed in B73. Shown are exons 1-5 of the maize brassinosteroid biosynthesis gene DWF4. The gene is known to be feedback regulated through the action of the major BR TF BZR1. The figure was adapted from Hartwig et al., 2023.

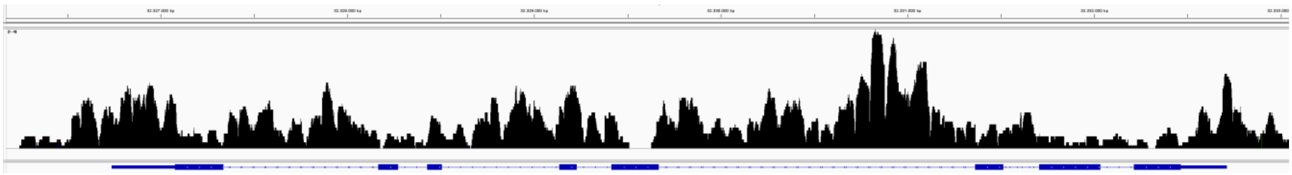

Figure 5b: The results show the MOA data from the merged Oh43xGF111 well-watered sample. The MOA data also showed a significant peak (MACS3, q < 0.05) peak downstream of exon 3, overlapping with the BZR1 ChIP-seq binding site.

DNA methylation analysis protocol for BOOSTER parental lines (SeqOccIn parental lines)

This analysis protocol aims to closely follow the procedure performed for the American nested association mapping (NAM) population methylation data presented in Hufford et al., 2021. Plant material was generated by the laboratory of Clementine Vitte (Quantitative Genetics and Evolution research unit - Le Moulon, Gif-sur-Yvette, France)

Plant material

The photoperiod was 16 h of light and 8 h of dark. The temperature was approximately 25 °C during light hours. The relative humidity was approximately 54%. Seedlings were grown for approximately 6 days and harvested 4–6 h after Zeitgeber Time 0 (ZT0, lights on). Seedlings were harvested when the first leaf had emerged 2–3 cm above the apical tip of the coleoptile. The seedlings were cut 3 mm above the coleoptile–mesocotyl boundary, excluding the shoot apical meristem, and the second leaf was removed from within the sheath of the first leaf. Only the inner second leaves, which contained the third and fourth leaves sheathed inside, were used for experiments. For each genotype, two replicates were made, each containing a pool of second leaves collected from 5 to 6 plants.

Plant material for the 29 SeqOccIn inbred lines was harvested as described in Ricci et al. (2019): Seedlings were grown in long day conditions of 16h light and 8h darkness for approximately 6 days (when the first leaf had emerged 2–3 cm above the apical tip of the coleoptile). At this stage, the second leaf and any eventual present third leaves were harvested by cutting just above the shoot apical meristem. Each replicate included at least five leaves, and two replicates were collected.

DNA extraction and enzyme methyl (EM)-seq

DNA extraction, library preparation and sequencing were performed as described in Hartwig et al., 2023,

specifically:

Tissues were homogenized in liquid nitrogen, and DNA was isolated with the DNeasy Plant Mini Kit (Qiagen). Libraries were prepared using the NEBNext Enzymatic Methyl-seq Kit (NEB) following the protocol for large DNA inserts. Therefore, 200ng genomic DNA was combined with 0.002 ng CpG methylated pUC19 DNA and 0.04 ng unmethylated lambda DNA. Fragmentation was done by using the Diagenode Bioruptor NGS in three rounds, 30s on, 90s of. Agilent Technologies 4200 Tape Station was used to determine the size distribution and concentration of the libraries.

Methylation ratio analysis

1. Quality control

Quality control was performed using FastQC (Andrews S (2010)."FastQC: a quality control tool for high throughput sequence data", http://www.bioinformatics.babraham.ac.uk/projects/fastqc) and summarised using MultiQC (Ewels et al., 2016). We aimed at a sequencing depth comparable to the one for Oh43 produced by the NAM consortium (232 million read pairs and ~72% unique reads), with Oh43 being the tester line used in our approach. Total read numbers for the SeqOccIn lines are all above 300 million read pairs (Readnumbers_EMseq_Booster.xlsx) with 60% to 72% unique reads (Fig. 6). Sequence quality is uniform and very high throughout the reads (Fig. 7).

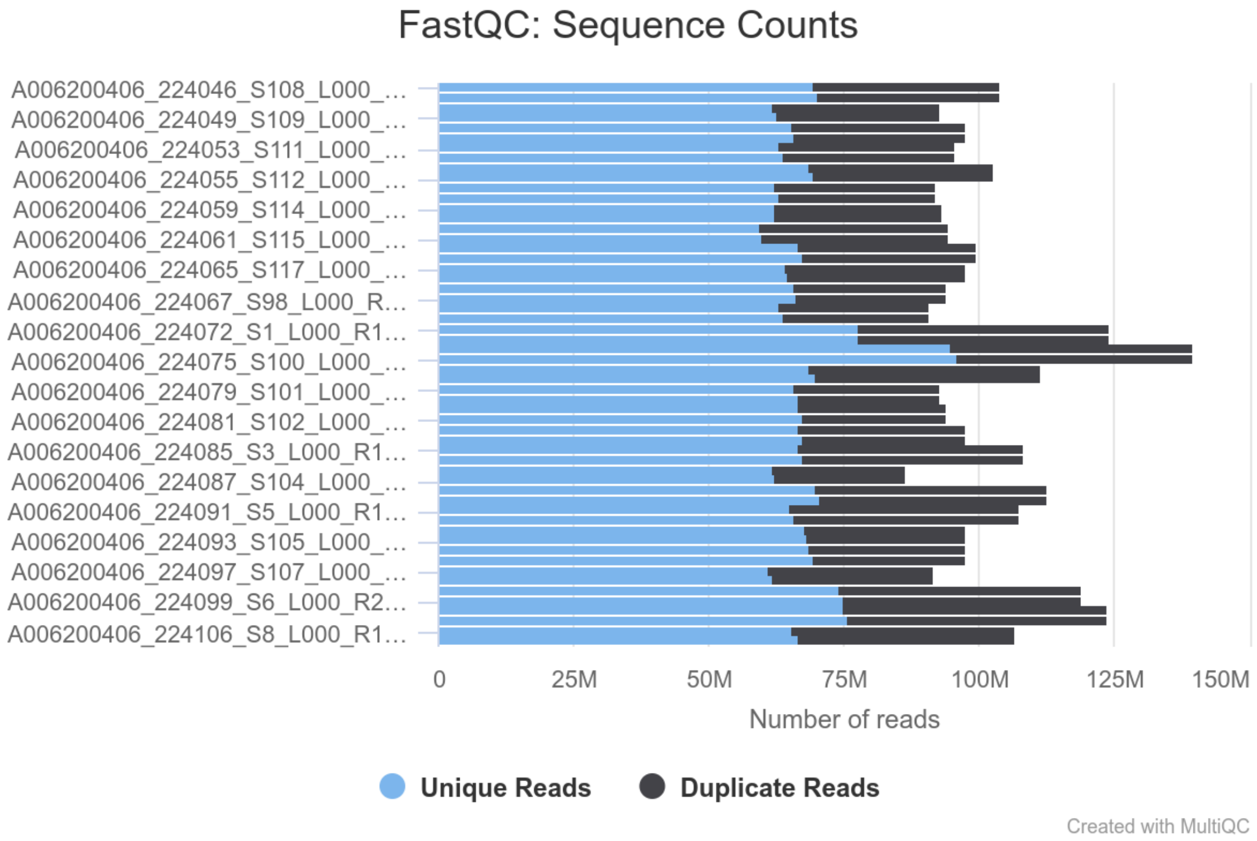

Figure 6: Unique and duplicated read counts for the SeqOccIn lines (results of one lane shown here).

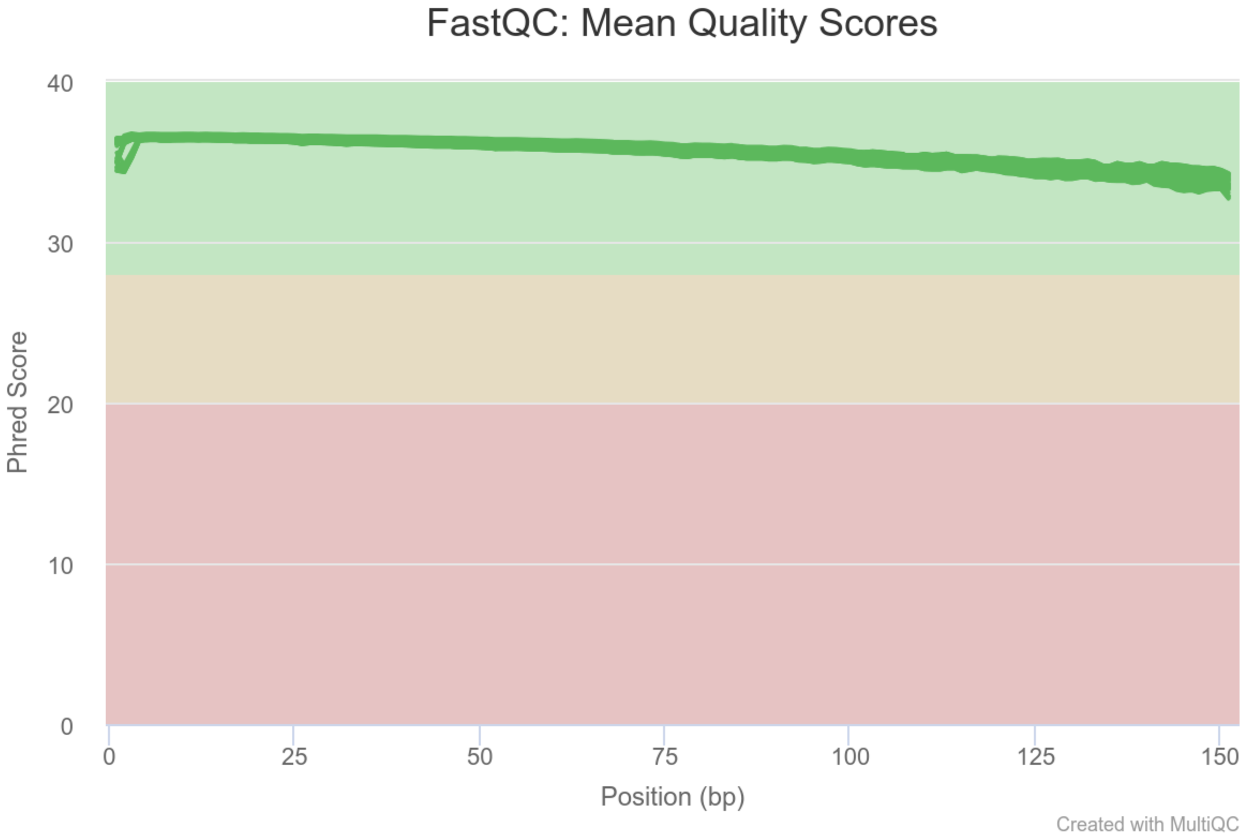

Figure 7: Per base sequence quality for the SeqOccIn lines (results of one lane shown here).

2. Read mapping and determination of methylation counts at base-pair resolution

To allow for parallel processing, fastq files were split into 10 mio read pair files using famas (https://github.com/andreas-wilm/famas) and subsequent analyses were parallelised with task-spooler (https://github.com/justanhduc/task-spooler).

Adapter removal and mapping of reads was performed using BS-Seeker2 v2.1.8 (Guo et al., 2013) with one mismatch allowed, a maximum insert size of 1000 and the adapter sequence specified as AGATCGGAAGAGC. Observed mapping efficiencies were 61-65% with around 80%,59% and 1.9% of C methylated in CG, CHG and CHH contexts, respectively.
Duplicated reads were removed using picard tools (https://broadinstitute.github.io/picard/) and methylation counts in the three contexts were extracted using the bs_seeker2-call_methylation.py tool for further analysis.

Data Information:
All additional data (raw, QC, and transformed) associated with D1.3 as described in the data management plan (D6.3, Appendix) were deposited at the following databases:

Maize drought resilience data:
Samples: 29 F1 hyrbids, 3x biological replicates, 2 conditions (well-watered, drought stressed)
Upload repository: DataPlant; https://git.nfdi4plants.org/thartwig/booster
Output data: https://git.nfdi4plants.org/thartwig/booster/assays/RWC/dataset/Output

DMP (Deliverable 6.3) data included:

- DS.01.06 (PUBLIC/EMBARGO according to the info provided in Appendix of D6.3)

MOA-seq
Samples: 29 F1 hyrbids, 3x biological replicates, 2 conditions (well-watered, drought stressed)

Upload repository: DataPlant; https://git.nfdi4plants.org/thartwig/booster

Raw sequencing input data: https://git.nfdi4plants.org/thartwig/booster/assays/MOA-seq/dataset\Input
Output data: https://git.nfdi4plants.org/thartwig/booster/assays/MOA-seq/dataset/Output
DMP (Deliverable 6.3) data included:

- DS.01.10 (PUBLIC/EMBARGO, according to the info provided in Appendix of D6.3)

E-Methylation-seq
Samples: 29 SeqOccIn Inbred lines (see D1.1 for line details); 1 Sample per line pooled from 2 biological replicate prior to sequencing; all sequenced on 3 NovaSeq S4 lanes
Upload repository: DataPlant; https://git.nfdi4plants.org/thartwig/booster

Raw sequencing input data: https://git.nfdi4plants.org/thartwig/booster/assays/eMethyl-seq/dataset/Input
Output data: https://git.nfdi4plants.org/thartwig/booster/assays/eMethyl-seq/dataset/Output
DMP (Deliverable 6.3) data included:

- DS.01.13a (PUBLIC/EMBARGO, according to the info provided in Appendix of D6.3)

References:

Hufford M B et al.,De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes.Science373,655-662(2021).

Savadel S, Hartwig T, Turpin Z et al. 2021 The native cistrome and sequence motif families of the maize ear. PLOS Genetics.

Hartwig T, Banf M, Prietsch GP, et al. 2023. Hybrid allele-specific ChIP-seq analysis identifies variation in brassinosteroid-responsive transcription factor binding linked to traits in maize. Genome Biol. 24: 108.

Ricci WA, Lu, Z, Ji L et al. Widespread long-range cis-regulatory elements in the maize genome. Nat. Plants 5, 1237–1249 (2019).

Ewels P, Magnusson M, Lundin S, Käller M, MultiQC: summarize analysis results for multiple tools and samples in a single report, Bioinformatics, Volume 32, Issue 19, October 2016, Pages 3047–3048,

Guo, W, Fiziev, P, Yan, W et al. BS-Seeker2: a versatile aligning pipeline for bisulfite sequencing data. BMC Genomics 14, 774 (2013).

Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D. 2010. BigWig and BigBed: enabling browsing of large distributed datasets. Bioinformatics 26: 2204–2207.

Dobin A, Davis CA, Schlesinger F, et al.2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29: 15–21.

Magoč, T, & Salzberg, S L (2011). FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics (Oxford, England)*, *27*(21), 2957–2963.

Li H, Handsaker B, Wysoker A, et al.2009. The Sequence Alignment/Map format and SAMtools. Bioinforma. Oxf. Engl. 25: 2078–2079.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26: 841–842.

Engelhorn J, et al. Genetic variation at transcription factor binding sites largely explains phenotypic heritability in maize
bioRxiv 2023.08.08.551183; doi: https://doi.org/10.1101/2023.08.08.551183