

A Thesis Submitted for the Degree of PhD at the University of Warwick

Permanent WRAP URL:

<http://wrap.warwick.ac.uk/131745>

Copyright and reuse:

This thesis is made available online and is protected by original copyright.

Please scroll down to view the document itself.

Please refer to the repository record for this item for information to help you to cite it.

Our policy information is available from the repository home page.

For more information, please contact the WRAP Team at: wrap@warwick.ac.uk

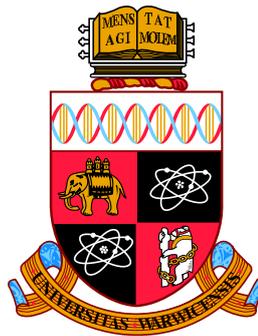
**Insights into Doubled Haploid Breeding Programs
from a Genome-Wide Transcriptome and Methylome
Analysis of *Brassica oleracea***

Jonathan Lewis Price

This thesis is submitted in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

University of Warwick, Department of Life Sciences



October 2018

Acknowledgements

I would like to extend my gratitude to my supervisor, Dr. Jose Gutierrez-Marcos, for giving me this opportunity and allowing me to develop my knowledge and skills under his guidance on so many projects. You really have been a source of inspiration and great ideas.

I would also like to thank my advisory panel, Dr. Sascha Ott and Dr. Graham Teakle for their generous time and advice. My thanks also goes to everyone in the Marcos group over the last 4 years many of whom have kept me sane with sound advice in work and evenings at the local. Without you the four years would have been even harder.

I also need to thank my parents, Dean and Linda Price, for their support over the last 28 years. Particularly the encouragement to go to University, without that, this would not have been possible. I would also like to thank my partner in crime, Grace Watkins who has been a source of calm and understanding throughout, it really wouldn't be the same without you. Also for those late nights of correcting my atrocious spelling, so on that note I would also like to thank Ruth Watkins for her literary skill and punctuation policing.

Lastly, I would like to thank the BBSRC and the MIBTP program, I have really developed throughout the 4 years of my PhD and it would not have been possible without the funding and the training received from the MIBTP program.

Publications

The work herein has not been published in a peer reviewed journal. However during my PhD I was able to contribute significantly to a number of projects. Although not assessed for this degree, the funding received from the BBSRC has allowed this research to be conducted.

- Price, J., Harrison, M., Hammond, R., Adams, S., Gutierrez-marcos, J., Mallon, E. (2018). Alternative splicing associated with phenotypic plasticity in the bumble bee *Bombus terrestris*. *Molecular Ecology*.
- Wibowo, A., Becker, C., Durr, J., Price, J., Papareddy, R., Santain, Q., Spaepen, S., Hilton, S., Bending, G., Schulze-Lefert, P., Weigel, D., and Gutierrez-Marcos, J. (2018). Incomplete reprogramming of cell-specific epigenetic marks during asexual reproduction leads to heritable phenotypic variation in plants. *PNAS*.
- Wibowo, A., Becker, C., Marconi, G., Durr, J., Price, J., Hagmann, J., Papreddy, R., Putra, H., Kageyama, J., Becker, J., Weigel, D., Gutierrez-marcos, J. (2016). Hyperosmotic stress memory in *Arabidopsis* is mediated by distinct epigenetically labile sites in the genome and is restricted in the male germline by DNA glycosylase activity. *eLife*.

Declarations

This thesis was submitted to the University of Warwick for the degree of Doctor of Philosophy. The work presented here is original and has not been considered for any other award. All work here in was performed by myself with few exceptions outlined below.

Mr. Robert Maple, Dr. Yang Seok, Mr. Ranjith Papareddy (University of Warwick)

Assisted in the growth and harvesting of samples along with the extraction of RNA and DNA for sequencing.

Max Plank Institute, Tübingen

The sequencing facility at MPI Tübingen performed the library preparation and the sequencing of the libraries.

Warwick Crop Centre

Also assisted in the growth of the samples, management of the greenhouses and catalogued and housed the seed stock used in this study.

Insights into Doubled Haploid Breeding Programs from a Genome-Wide Transcriptome and Methylome Analysis of *Brassica oleracea*

Abstract

With global food insecurity on the rise and increased pressures on habitat conservation it is vital to pursue increased efficiencies in plant breeding. New understanding of genetics in the last five decades has provided significant advances in yield gains in most cultivated crops. However, new advances in genomics have opened the door to better design breeding programs; the major strategy to develop new cultivars, in order to exploit novel sources of genetic variation and increase selection efficiency. Many studies have focused on the efficient use of molecular markers to speed up selection processes, yet the use of omics studies in plant breeding programs is currently under utilised as they could be used to elucidate the mechanisms underpinning the inheritance of traits currently exploited by plant breeders. This study has generated whole-genome transcriptome and methylome data to uncover the mechanisms implicated in the inheritance of traits generated during a doubled haploid (DH) breeding program. Our analysis reveals the existence of a number of predictive elements that explain the molecular variation present in DH breeding. The major elements contributing to this variation are the level of dominance in hybrids and the contribution of each parental genome in individual DH lines. Collectively, our works demonstrate that genomic and epigenomic studies can provide insights into genome regulation and more importantly, aid the design of new plant breeding programs.

Abbreviations

- **DNA** Deoxyribonucleic Acid
- **RNA** Ribonucleic Acid
- **mRNA** Messenger ribonucleic Acid
- **siRNA** Small ribonucleic Acid
- **siRNA** Silencing ribonucleic Acid
- **miRNA** Micro ribonucleic Acid
- **ncRNA** Non-coding ribonucleic Acid
- **MAB** Marker assisted breeding
- **MAS** Marker assisted selection
- **NGS** Next generation sequencing
- **RNASeq** Ribonucleic Acid sequencing
- **BSSeq** Bisulphite Sequencing
- **QTL** Quantitative trait loci
- **RFLP** Restriction fragment length polymorphism
- **SSR** Small sequence repeat
- **SNP** Single nucleotide polymorphism
- **F1** Filial 1

-
- **DH** Doubled haploid
 - **GCA** General combining ability
 - **SCA** Specific combining ability
 - **GS** Genomic selection
 - **MPV** Mid parent value
 - **HP** High parent
 - **LP** Low parent
 - **TF** Transcription factor
 - **TE** Transposable element
 - **5meC** 5 methyl cytosine
 - **GS** Genomic selection
 - **gbM** Gene body methylation
 - **FDR** False discovery rate
 - **DEG** Differentially expressed gene
 - **phDEG** Parent hybrid differentially expressed gene
 - **dhDEG** Doubled haploid differentially expressed gene
 - **DMR** Differentially methylated region
 - **phDMR** Parent hybrid differentially methylated region

-
- **dhDMR** Doubled haploid differentially methylated region
 - **MR** Methylated region
 - **ELD** Expression level dominance
 - **MLD** Methylation level dominance
 - **PCR** Polymerase chain reaction

Contents

1	Introduction	1
1.1	Plant Breeding	1
1.1.1	From Breeding to Molecular Breeding	1
1.1.2	Hybridisation	5
1.1.3	Doubled Haploids	7
1.1.4	Typical Modern Breeding Strategies	9
1.2	Genome Regulation in Plants	13
1.2.1	Gene Expression	13
1.2.2	DNA Methylation	15
1.3	Merging Plant Genomes	19
1.4	Genomic Analyses in <i>Brassica</i> Species	21
1.5	Aims and Hypothesis	23
2	Methods	25
2.1	Creation of F1 and Doubled Haploid Lines	25
2.2	Selection of Samples and Plant Growth	26
2.3	RNA Extraction Data Generation	27
2.4	RNASEq Data Processing	28
2.5	DNA Extraction	29

2.6	Bisulphite Data Processing	29
2.7	Single Cytosine Analysis	30
2.8	Methylated Regions	30
2.9	Differentially Methylated Regions	31
2.10	Parent and Hybrid Differential Expression and Methylation	32
2.11	Homologous Recombination Site Detection	35
2.12	Differential Expression and Methylation in the DHLs	38
2.13	Gene Ontology Analysis	38
2.14	Intersection of DMRs and Genes and Transposons	38
3	Transcriptome and Methylome Analysis of A12Dhd and GDDH33	
	Parental Lines	42
3.1	Introduction	42
3.2	Chapter Aims and Hypothesis	46
3.3	Mapping of Parental Accessions to the Reference <i>B. oleracea</i> Genome	47
3.4	Gene Expression Differences Between A12Dhd and GDDH33	47
3.5	Methylation Analysis in the Parental Lines	51
	3.5.1 Analysis of Single Cytosines	51
	3.5.2 Differentially Methylated Regions between A12Dhd and GDDH33	53
3.6	Discussion	55
4	Transcriptome and Methylome Analysis of the F1 Hybrids	57
4.1	Introduction	57
	4.1.1 Transcriptomic Studies of F1 Hybrids and Polyploids	58

4.1.2	Epigenetic Studies of F1 Hybrid	61
4.2	Chapter Aims and Hypothesis	65
4.3	Gene Expression in the Parent Hybrid Cross	66
4.3.1	Gene Expression Dynamics in the F1 Hybrid	66
4.3.2	Differentially Expressed Genes	69
4.4	Methylation in the Parent Hybrid Cross	72
4.4.1	Analysis of Single Cytosines	72
4.4.2	DMR Dynamics in the F1 Hybrid	76
4.4.3	Location of Methylation and Differential Methylation	80
4.5	Discussion	83
5	Transcriptome and Methylome Analysis of the Doubled Haploid Lines	87
5.1	Introduction	87
5.2	Chapter Aims and Hypothesis	90
5.3	Genotyping Doubled Haploid Lines - Understanding the Parental Genome Contribution	92
5.4	Gene Expression Dynamics in the DH lines	101
5.5	Differentially Expressed Genes in the DH Lines	105
5.6	Methylation Dynamics in the DH Lines	107
5.7	Correlation Between DNA Methylation and Gene Expression in <i>B.</i> <i>oleracea</i> DH Lines	114
5.8	Discussion	120
6	Discussion	124

6.1	Parental Regulome Divergence Governs Parental Combining Ability for mRNA Dosage Dependant Traits	127
6.2	Exploiting the Expression-level Dominance in Plant Breeding Pro- grams for mRNA Dosage Dependant Traits	129
6.3	Utilising Parental Genome Contribution in Homozygous Line Selec- tion	131
6.4	Future Directions	132
6.4.1	Parental Genome Contributions in Doubled haploid breeding .	132
6.4.2	Studies of Genome Mergers	135
	Appendices	160
	A Read Processing Numbers	161
	B Appendix for Chapter 5	166

List of Figures

- 1.1 **Simplified schematic of typical breeding strategies.** Arrows indicate the movement of potential cultivars through the breeding program. 12

- 2.1 **Schematic of samples in this study along with their method of creation.** Samples in bold indicate those with whole genome bisulphite sequencing data and whole genome RNA sequencing data. Line names or numbers are shown underneath each line and the arrows indicate each sample's method of creation. 28

2.2 Schematic showing how the d/a and parental d/a ratios are calculated and plotted. The ratios are used to show how the dynamics of a DMR or gene in F1 hybrid relate to the parental methylation or expression. a) The calculation of the ratios. Firstly, three pairwise comparisons are performed (A12Dhd - F1, GDDH33 - F1 and A12Dhd - GDDH33). Then for each of these genes or DMRs shown to be significant in at least one comparison, two ratios are calculated. The d/a ratio and the parental d/a ratio. b) Displays the meaning of the ratios. The d/a ratio (left histogram) describes the methylation of the DMR or expression of the gene in the F1 according to the high or low parent (parent with highest or lowest expression). The parental d/a ratio (right histogram) describes the methylation of the DMR or expression of the gene in the F1 according to the expression of the maternal parent (A12Dhd) or the paternal parent (GDDH33). The histograms show the thresholds imposed on these ratios that decide the expression or methylation category (additive, parental-level dominance or above / below parental levels. c) Plotting and display of the ratios and categories. In the top plot, each genes ratios are plotted, the d/a ratio on the y-axis and the parental d/a ratio is plotted on the x-axis. Plotting in this way, each differentially expressed feature can be categorised according to both the high / low parent and the maternal / paternal parent. The bottom table of c) shows this categorisation. Roman numerals show the categories as they are commonly described (Yoo et al., 2013). Underneath the Roman numerals in the table, there is a graphic displaying the expression or methylation pattern of this category for the 3 genotypes (A12Dhd - maternal, GDDH33 - paternal and F1). . . . 34

- 2.3 **Schematic of SNP genotyping pipeline.** a) From vcf files of each parental genome (A12DHd - yellow, GDDH33 - blue), positions in each genome are only kept if their allele frequency is equal to 1 and their coverage is greater than 10. b) Positions between the parental genome are compared, they are kept if the base call in each parental genome is different and both genomes have received greater than 10 coverage. c) This is the set of distinguishing positions that can separate the DH genomes. d) The final step of the program introduces the vcf file from a DH line and looks for positions with an allele frequency of 1, it then categorises each position according to the set of parental SNPs identified earlier. The program identifies a HR site if there is a switch in the parental inheritance with 10 SNPs on either side of the HR site. 37
- 2.4 **Three examples of possible intersections from the GFF intersector program.** a) The case in which there is a gene which is overlapped by 5 DMRs, this becomes an intersect regions containing 1 gene and 5 DMRs. b) The intersect regions contains 2 genes and 4 DMRs, the two genes are contained within one region because they are less than the user supplied flanking region (f) apart and both gene has at least 1 DMR intersecting. c) In this example the second gene is not included in the intersect region because it has no associated DMRs within f of the TSS or TES. Yellow boxes show user defined regions e.g DMRs. Blue boxes show exons of genes, the TES and TSS are defined in the graphics with stars. 40

- 2.5 **Schematic showing the process of correlating gene expression and DMR methylation within regions identified using the GFFIntersector program.** a) Shows the first step of correlation program in which DMRs of the same sequence context that directly overlap are combined into DMR blocks. b) Correlation of the DMR blocks and genes. Correlations are made between all DMR blocks and any gene within the flanking region of the DMR block, this maybe more than one gene. For each gene and DMR block comparison (5 in the above example), the coverage in methylation in each sample is assessed to ensure 5 reads are covering the region in at least 5 samples. Then the gene is checked for expression in at least one sample and a 1.2 fold difference. Then for each comparison identified, Spearman’s rank correlation is performed. From all comparisons those with are FDR of less than 0.01 are considered significant. 41
- 3.1 **Genetic distance between *B oleracea* varieties.** Dendrogram, modified from Golicz et al. (2016) shows the genetic distance between the 9 lines used by Golicz et al. (2016). Numbers in green show genes that are present in the varieties below that node but not present in the others. The blue numbers represent the number genes that are not present in the lines below that node but present in the other lines. The scale indicates the number of nucleotide substitutions per site. 45

3.2	Parental differentially expressed genes. a) Scatter plot showing the average of each replicates expression of each gene for A12Dhd and GDDH33 with significant pDEGs appearing in red. b) Heatmap displaying expression of the 3216 parental DEGs with hierarchical clustering.	48
3.3	Global cytosine methylation in A12Dhd and GDDH33 in each methylation context. CG (top), CHG (middle) and CHH (bottom). a) Histogram displaying the proportion of sites exhibiting methylation ratios of 0-100% the panel in the corner of each plot displays a zoomed view of the distribution of sites with 1-100% methylation. b) The average methylation percentage of all cytosines.	52
3.4	Average methylation across genomic features for CG, CHG and CHH methylation. Genes (Left), DNA and RNA transposable elements (Right). For each feature and context the 2 kb flanking regions for each feature are an average methylation value for a particular position for each feature in the genome. Across the feature body each feature is split into 100 bins and the methylation is averaged over these bins and then averaged across all features for these bins.	53

3.5 **Numbers and location of pDMRs and pMRs.** a) Barplot of the numbers of DMRs between A12Dhd and GDDH33 in each sequence context. DMRs with higher methylation in A12Dhd are shown in yellow and DMRs with higher methylation in GDDH33 are shown in blue.

b) Location of these DMRs within genomic features, each base of a set of DMRs is assigned to the feature that it overlaps with. Then the results are displayed as a percentage of the total bases in that set. For each sequence context both A12Dhd MRs and GDDH33 MRs are shown. Then the DMRs between these two genotypes are split into DMRs with higher methylation in A12Dhd (A12) and DMRs with higher methylation in GDDH33 (GD). WG refers to the assignment of all the bases in the reference genome when assigned to a feature. This is done in a hierarchical fashion to account for overlapping features (gene, transposon, upstream, downstream, intergenic: in order of decreasing importance) 54

- 4.1 **Gene expression dynamics in F1 hybrid.** a) Venn diagram showing parental DEGs (blue) from the comparison between A12Dhd and GDDH33. Then the parental and hybrid DEGs (brown) from all 3 comparisons (A12Dhd - F1, GDDH33 - F1 and A12Dhd - GDDH33). This plot shows there is little novel differential expression in the F1 hybrid. b) Dominant-to-additive plot showing expression dynamic of phDEGs in F1 hybrid relative to the parental expression. Each phDEGs ratios are plotted, the d/a ratio on the y-axis and the parental d/a ratio is plotted on the x-axis. Plotting in this way, each phDEG can be categorised according to both the high / low parent and the maternal / paternal parent as shown by the numbers in the quadrants of the graph. c) Shows the categorisation of each gene. Roman numerals show the categories as they are commonly described (Yoo et al., 2013). Underneath the Roman numerals in the table, there is a graphic displaying the expression or methylation pattern of this category for the 3 genotypes (A12Dhd - maternal, GDDH33 - paternal and F1) then underneath that are the proportions of the phDEGs belonging to 12 mutually exclusive expression patterns. 67
- 4.2 **Heatmaps of phDEGs. The F1 genotype forms a clade with A12Dhd for additively and non-additively expressed genes.** a) Heatmap of additive phDEGs. b) Heatmap of non-additively expressed phDEGs. Scale displayed is $\log_2(\text{DESeq2 normalised expression levels})$. Hierarchical clustering was performed and displayed as a dendrogram on the top of the heatmaps. 68

4.3 **Global cytosine methylation in A12Dhd, GDDH33 and F1 in each methylation context.** CG (top), CHG (middle) and CHH (bottom).
a) Histogram displaying the proportion of sites exhibiting methylation ratios of 0-100% the panel in the corner of each plot displays a zoomed view of the distribution of sites with 1-100% methylation. b) The average methylation percentage of all cytosines. 74

4.4 **Methylation average across genomic features for CG, CHG and CHH methylation.** Genes (Left), DNA and RNA transposable elements (Right). For each feature and context the 2 kb flanking regions for each feature are an average methylation value for a particular position for each feature in the genome. Across the feature body each feature is split into 100 bins and the methylation is averaged over these bins and then averaged across all features for these bins 75

4.5 **Venn diagrams showing overlap between pDMRs (blue) and phDMRs (brown). The F1 has little novel methylation at CG sites but more novel methylation at CHG and CHH sites.** For each context separately the pDMRs are obtained from the comparison between A12Dhd and GDDH33 and the phDMRs are obtained from the three comparisons (A12Dhd - GDDH33, A12Dhd - F1 and GDDH33 - F1. 77

- 4.6 **Methylation dynamics for CG, CHG and CHH context phDMRs.**
- a) Dominant-to-additive plots showing methylation dynamics of phDMRs in F1 hybrid relative to the parental methylation. Each phDMRs ratios are plotted, the d/a ratio on the y-axis and the parental d/a ratio is plotted on the x-axis. Plotting in this way, each phDMR can be categorised according to both the high / low parent and the maternal / paternal parent as shown by the numbers in the quadrants of the graph. b) Shows the categorisation of each phDMR. Roman numerals show the categories as they are commonly described (Yoo et al., 2013). Underneath the Roman numerals in the table, there is a graphic displaying the expression or methylation pattern of this category for the 3 genotypes (A12Dhd - maternal, GDDH33 - paternal and F1) then underneath that are the proportions of the phDMRs belonging to 12 mutually exclusive expression patterns in each sequence context. 78
- 4.7 **Heatmaps of phDMRs. Methylation of the F1 is most similar to A12Dhd for both additive and non-additive phDMRs.** a) Additive phDMRs. b) Non-additive phDMRs. For each sequence context; CG, CHG and CHH. Scale shows the methylation rate of the DMR. 79

4.8 Location of phMRs and phDMRs at CG, CHG and CHH context within genomic features. A - Additive phDMRs, N - Non-additive phDMRs, T - Transgressive phDMRs. Mr - MRs from relevant context. WG - The reference genome if all bases were assigned to a feature. All bases of a particular feature set are assigned to a genomic feature and then plotted as a percentage of the total number of bases in that feature set. This is done in a hierarchical fashion to account for overlapping features (gene, transposon, upstream, downstream, intergenic: in order of decreasing importance).	81
4.9 Locations of phDMRs and phMRs in different transposon types. In each case it shows the proportion of bases in each feature type that occupies the different types of transposons.	82

5.1 Large example of circos plots for whole genome comparisons.

Whole genome comparisons between any DH line and either parent are invalid, illustrates the need to only compare inherited regions between the parents and DH lines. Results from all DH lines tested can be seen in Figures 5.2, 5.3, although smaller they are just there to illustrate the point mentioned above. Whole genome comparison of the DH line 2134 and GDDH33 in blue. Then whole genome comparison of DH line 2134 and A12Dhd in yellow. a) Predicted genotype of the DH line. b) Magnitude of DEGs between the DH line and both parents. c) Magnitude of CG DMRs between the DH line and both parents. d) Magnitude of CHG DMRs between the DH line and both parents. e) Magnitude of CHH DMRs between the DH line and both parents. The outer ideogram displays the chromosome coordinates in megabases along with the centromere positions as a white band. 93

- 5.2 **Circos plots for whole genome comparisons. Whole genome comparisons between any DH line and either parent are invalid, illustrates the need to only compare inherited regions between the parents and DH lines.** DH line vs GDDH33 in blue and DH line vs A12Dhd in yellow. a) Predicted genotype of DH lines. b) Magnitude of DEGs between the DH lines and both parents. c) Magnitude of CG DMRs between the DH lines and both parents. d) Magnitude of CHG DMRs between the DH lines and both parents. e) Magnitude of CHH DMRs between the DH lines and both parents. The outer ideogram displays the chromosome coordinates in megabases along with the centromere positions as a white band. 94
- 5.3 **Circos plots for whole genome comparisons. Whole genome comparisons between any DH line and either parent are invalid, illustrates the need to only compare inherited regions between the parents and DH lines.** DH line vs GDDH33 in blue and DH line vs A12Dhd in yellow. a) Predicted genotype of DH lines. b) Magnitude of DEGs between the DH lines and both parents. c) Magnitude of CG DMRs between the DH lines and both parents. d) Magnitude of CHG DMRs between the DH lines and both parents. e) Magnitude of CHH DMRs between the DH lines and both parents. The outer ideogram displays the chromosome coordinates in megabases along with the centromere positions as a white band. 95

5.4	Plots of results from SNP genotyping and epigenotyping. For each chromosome the observed parent of origin is shown as 4 bars (yellow - A12Dhd, blue - GDDH33). Bars from top to bottom - SNP data genotype, 60kb epigenotyping, 70kb epigenotyping and 150kb epigenotyping.	98
5.5	Plots of results from SNP genotyping and epigenotyping. For each chromosome the observed parent of origin is shown as 4 bars (yellow - A12Dhd, blue - GDDH33). Bars from top to bottom - SNP data genotype, 60kb epigenotyping, 70kb epigenotyping and 150kb epigenotyping.	99
5.6	Plots of results from SNP genotyping and epigenotyping. For each chromosome the observed parent of origin is shown as 4 bars (yellow - A12Dhd, blue - GDDH33). Bars from top to bottom -SNP data genotype, 60kb epigenotyping, 70kb epigenotyping and 150kb epigenotyping.	100
5.7	Genotypes of DH lines. a) Percentage of parental genome inheritance to each DH line. b) Circos plot. Each ring represents the diploid genome of a DH line. Each chromosome is coloured according to it's inheritance.	100

- 5.8 Dynamics of dhDEGs in each of the parentally inherited genomes from each DH line.** a) There are more dhDEGs on GDDH33 inherited genomes compared to A12Dhd inherited genomes (T-test ($t = -2.047$, $p\text{-value} = 0.03174$)). b) Venn diagram with the parental and hybrid DEGs identified in Chapter 4 and the DH DEGs from each line, split by parental inheritance and their overlapping genes. c) Shows the F1 expression dynamics of the phDEGS that overlap with dhDEGs (A = phDEGs and A12Dhd inherited dhDEGs, G = phDEGs and GDDH33 inherited dhDEGs) and the phDEGs that recover in the DH lines (R) these are sections shown in the venn diagrams in panel b. There is a significant association between F1 expression and dynamics and the category of DEG - X^2 ($df = 4$, $N = 3254$) = 145.7, $p\text{-value} < 0.001$. . . 102
- 5.9 Parental dominant-to-additive ratios of the dhDEGs, for each inherited genome dhDEGs tend to display expression dynamics similar to that of the other parental genome .** From left to right; Top - 2069, 3088, 3238, Middle - 1047, 5071, 1003, Bottom - 5119, 2134, 3013. For each line their A12Dhd inherited dhDEGs are shown in yellow and the GDDH33 inherited dhDEGs are shown in blue. The x-axis displays the parental d/a ratio, a ratio of 1 would mean a gene has equal expression to the gene in A12Dhd and a ratio of -1 means the gene would have equal expression to the GDDH33 parent. 103

5.10 **There is a negative relationship between relative gene expression change and whole genome inheritance in the DH lines.** For each inherited genome in each DH line the relative gene expression change (dhDEGs per gene inherited) is plotted against the amount of genome inherited from that parent. The significant relationships are shown as lines calculated by linear regression. 104

5.11 **Combined GO analysis for dhDEGs, each node represents an enriched GO term.** (FDR <0.05). Size of node and label represents the number of DH lines a given terms is enriched in. 106

5.12 **Comparison of CG, CHG and CHH dhDMRs with the phDMRs.**
 a) Venn diagram with the parental and hybrid DMRs identified in Chapter 4 and the dhDMRs from each line, split by parental inheritance and their overlapping DMRs. b) Shows the F1 expression dynamics of the phDMRs that overlap with dhDMRs (A = phDMRs and A12Dhd inherited dhDMRs, G = phDMRs and GDDH33 inherited dhDMRs) and the phDMRs that recover in the DH lines (R) these are sections shown in the venn diagrams in panel a. There is a significant association between F1 methylation dynamics and the category of DMR -(CG = X^2 (df = 4, N = 22807) = 465.5, p-value <0.001), (CHG = X^2 (df = 4, N = 11946) = 483.6, p-value <0.001), (CHH = X^2 (df = 4, N = 20199) = 2509.5, p-value <0.001) 109

5.13 Number of dhDMRs in each line split by parental inheritance for each sequence context. There are more CG dhDMRs on A12Dhd inherited genome sections than GDDH33 genome sections (CG - T-test ($t = -2.224$, $p\text{-value} = 0.0485$)). CHG and CHH inherited sections do not show significant differences (CHG - ($t = 0.601$, $p\text{-value} = 0.5583$), CHH - ($t = 0.743$, $p\text{-value} = 0.4689$)) 110

5.14 Relationship between parental genome dosage and epigenetic changes in DH lines. For each inherited genome in each DH line the relative gene methylation change (dhDMRs per MR inherited) is plotted against the amount of genome inherited from that parent. The significant relationships are shown as lines calculated by linear regression. Non significant relationships are shown as a faint line. 111

5.15 Locations of dhDMRs in different genomic features. Each base of the DMRs are assigned to the genomic feature that it overlaps with in the annotation, these are displayed as a proportion of the total bases of that DMR category. This is done in a hierarchical fashion to account for overlapping features (gene, transposon, upstream, downstream, intergenic: in order of decreasing importance) 112

5.16 Locations of dhDMRs in different transposon types. Each base of the DMRs is assigned to the transposon type that it overlaps with in the annotation, these are displayed as a proportion of the total bases of that DMR category. 113

5.17 **The methylation contexts of the dhDMRs differ in their distribution across genes with which they overlap.** Density plot showing the distribution of dhDMRs in each sequence context over genes. For each feature and context the 2 kb flanking regions show the number of DMRs that are assigned to a particular position for each feature in the genome. Across the gene body each gene is split into 100 bins and the plot displays the number of DMRs that reside in that bin from all features. 115

5.18 **Methylation and expression of the Agamous-like locus.** The left hand side shows single base resolution methylome of the available samples for this region. The right hand side shows the expression of this locus in the available samples. Bars are coloured to show the genotype of this locus. 116

5.19 **Aligned reads for Agamous-like locus does not follow the Parkin et al. (2014) annotation.** This plot shows aligned reads from all samples for this locus. It shows that there is little evidence for the exon boundaries found in the official annotation. However there are transcripts produced from this locus. The annotation for Agamous-like is shown in blue at the bottom. The coordinates for chromosome 6 are shown above and then the aligned reads for all samples are shown in red with the number of reads from these that actually support the annotation shown as red numbers. 117

5.20	Methylation and expression of the FAS4 locus. The left hand side shows single base resolution methylome of the available samples for this region. The right hand side shows the expression of this locus in the available samples. Bars are coloured to show the genotype of this locus.	118
5.21	Methylation and expression of the TIL locus. The left hand side shows single base resolution methylome of the available samples for this region. The right hand side shows the expression of this locus in the available samples. Bars are coloured to show the genotype of this locus.	119
6.1	Simplified schematic of typical breeding strategies. Arrows indicate the movement of potential cultivars through the breeding program. Red processes indicate suggestions from this thesis.	126

List of Tables

3.1	Phenotypic measurements of A12Dhd and GDDH33 made by Ngwako (2003) , raw data was unavailable and so actual measurements are approximated here. Parent in bold shows the parent with highest value of that trait at that time point. The measurements are given below are the average measurement. (GDDH33 - A12Dhd)	44
3.2	GO analysis of 1726 genes with higher expression in A12Dhd compared to GDDH33 . All terms achieved FDR <0.05.	49
3.3	GO analysis of 1490 genes with higher expression in GDDH33 compared to A12Dhd . All terms achieved FDR <0.05.	50
4.1	GO analysis of 2234 additively expressed F1 genes . All terms achieved FDR < 0.05	70
4.2	GO analysis of 1119 non additively expressed F1 genes . All terms achieved FDR < 0.05	71
A.1	Table displaying the raw read numbers and the number after processing of the reads and alignment to the reference genomes. RNASeq libraries - Original old batch.	162

A.2	Table displaying the raw read numbers and the number after processing of the reads and alignment to the reference genomes. RNASeq libraries - New batch.	163
A.3	Table displaying the raw read numbers and the number after processing of the reads and alignment to the reference genomes. BS-Seq libraries - Original old batch.	164
A.4	Table displaying the raw read numbers and the number after processing of the reads and alignment to the reference genomes. BS-Seq libraries - New batch.	165
B.1	Regression statistics. Relative gene expression change and genome ownership.	167
B.2	Regression statistics. Relative CG and CHG methylation change and genome ownership.	168
B.3	Regression statistics. Relative CHH methylation change and genome ownership.	169

Chapter 1

Introduction

1.1 Plant Breeding

1.1.1 From Breeding to Molecular Breeding

The earliest evidence of human driven selection on plants dates from around 12,000 years ago. In Göbekli Tepe, Turkey it is believed that *Triticum monococcum* was selected over the natural einkorn wheat because its tough spindle allowed easy harvest and threshing. This has now been shown to have a monogenic basis, so these prehistoric humans selected this gene which, without human requisite, would be disadvantageous (Schlegel, 2018). This selection of beneficial traits by humans has occurred in every civilisation and has resulted in the wide array of edible plants available today. The scientific documentation of agriculture and horticulture was taken up in the ancient cultures and even before 1 AD there was knowledge of grafting, clonal propagation and selection. However, plant breeding in a systematic way was only undertaken after the discovery of gender in plants by Rudolf Camerarius in 1694, which allowed directed pollination and the documenting of plant pedigree. The first hybrid cross of

Dianthus caryophyllus and *Dianthus barbatus* was reported by Thomas Fairchild in 1717. However, in these early times, the concepts of genes and chromosomes were not available and so it was an empirical science based on fine observations and "breeders experience" without a theoretical background. The theory was not applied to these methods until 1900, when the 'Experiments on Plant Hybridization' was confirmed, having originally been written in 1865 by Gregor Mendel. This meant an explosion of new varieties and the birth of quantitative genetics.

The next large advance in breeding came in the 'green revolution' in the mid-twentieth century. At this time, advances in health care meant that the population was increasing. In Asian countries, much of the suitable agricultural land was already in use (Khush, 2001). This necessity led to the development of new technologies for plant breeding, particularly in rice, maize and wheat. The gains came from systematically targeting many different characteristics including; yield stability, stress resistance (biotic and abiotic), environmental adaptation and quality. This was combined with an increase in knowledge of mechanical farming, genetics, chemical applications and more stable irrigation. This, and a forceful social reform strategy and propaganda campaign led by the West resulted in increased yields world-wide. The food and agricultural agency reported that in the two decades between 1965 and 1985, crop yields per hectare improved by 56 percent. The techniques of artificial crossing, hybridisation, induced mutation and tissue culture were all used during this time and created many cultivars, such as IR8; developed in 1967 by the Indian rice breeder Nekkanti Subba. It offered yields of 10 tons per hectare where it was common to expect only 1.5 tons per hectare (Schlegel, 2018). But after the gold rush of gains, yield growth slowed. This is because monogenic traits can easily be exploited with dedicated phe-

notypic observation. Quantitative traits on the other hand, are controlled by multiple loci and do not segregate in a mendelian fashion often resulting in continuous phenotypes. This makes isolating the underlying genetic factors very difficult (Walley et al., 2012).

Now arguably we are in the next agricultural era: the era of post-genomic science. The discovery of restriction fragment length polymorphism (RFLP) in the 1980s allowed the use of molecular markers to improve selection efficiency in plant breeding programs (Botstein et al., 1980). Later the use of PCR meant that simple sequence repeat (SSR) markers took over (Mullis et al., 1986). These genomic markers were highly automatable and cheaper per data point. For example; Monsanto switched to a fully automated single nucleotide polymorphism (SNP) based genotyping system. Then from 2000 to 2006 data production increased 40-fold whilst the cost per data point decreased 6-fold (Eathington et al., 2007).

With next generation sequencing becoming common place, more ways of exploiting the genetic code are becoming apparent and now, marker assisted breeding (MAB) forms the basis for many modern breeding programs. The techniques are required because many important polygenic traits cannot be harnessed with mendelian methods. This is because quantitative traits are controlled by multiple loci, known as quantitative trait loci (QTLs). The first stage of MAB programs rely on identifying DNA markers closely linked to the phenotypes of interest. These techniques rely on the fact that the distance between two loci causing a segregating phenotype is proportional to the pattern of segregation of the alleles in later generations. After a genetic map has been constructed or the genome has been sequenced, different methods can be used to

identify the markers needed for selection. QTL analysis exploits these principles by genotyping; segregating populations nearly inbred lines, recombinant inbred lines and doubled haploid lines to identify markers that co-segregate with the phenotype. New methods like expressionQTL, epiQTL and proteinQTL are also ways of selecting the cause of a trait (Moose and Mumm, 2008). Along with identifying the underlying genetic cause of traits, DNA genotyping from young plants is much less work than phenotyping older plants (Butruille et al., 2015). However, these methods also suffer from many false positives and the population size required is large. In bigger crop species with long life cycles, growth space and maintenance is a big cost, along with the labour required for the phenotyping studies. Being able to genotype quickly increases the number of breeding cycles possible per year, meaning that even with the advanced technology required it is still more cost effective.

The next generation sequencing (NGS) platforms allow single base resolution of any genome. Now many genomes are sequenced providing the groundwork for omics studies across many taxa. Better resolution of techniques means that we are able to have a deeper understanding of the mechanisms underpinning the phenotypic observations seen in plant breeding. Omics studies can be focussed at the DNA, but also at the gene level with transcriptional studies and at the epigenetic level. Using these methods alongside traditional ones gives greater understanding of the actual causes of the phenotypes. This has allowed many functions of genes to be identified and has elucidated the genetic mechanisms of many traits. This knowledge is vital to moving forward with knowledge based methods of plant breeding which can allow invaluable prediction.

1.1.2 Hybridisation

F1 hybrids, a result of cross-fertilisation between two different varieties or even species, have been exploited world-wide for hundreds of years and are now an integral part of the plant breeder's toolbox (Chen, 2013). The reason for their commercial value is two-fold. Firstly, hybrids of inbred varieties benefit from a wide range of transgressive phenotypes. Often the F1s growth rate, seed size, biomass and fertility is better than that of the two parents, this is commonly known as heterosis (Schlegel, 2018). Secondly, they can be used for combination breeding; where desirable traits from two parents are combined in the progeny, hopefully both conferring all of the desired characteristics. As early as 1760 it was discovered that tobacco hybrids experienced growth vigour relative to their parents and detasseling for hybridisation in maize has been reported from as early as 1830. However, the man generally credited with inventing detasseling is Willian James Beal (1833-1910). Since this date, maize yields have increased more than 6-fold. Other species also benefit from hybridisation and now most varieties of maize, cabbage, radish and pepper are F1 hybrids. It is hard to underestimate their influence as a source of new varieties and yield increases.

Even though this phenomenon was widely exploited, the mechanisms underpinning it were, and still are, unclear. Early theories for heterosis include; dominance, overdominance, epistasis and pseudo-dominance. Epistasis states that gene functions rely on the presence of alleles at other loci, and so each gene relies on the overall genetic background (Powers, 1944). Dominance attributes the superior phenotype to the suppression of deleterious recessive alleles by dominant alleles from the other (Davenport, 1908). Overdominance states that combinations of alleles which in a ho-

mozygous state are deleterious become advantageous and pseudo dominance (which is similar to overdominance) allows the complementation of recessives to come from other loci on the homologous chromosome. Even though some success has come by studying heterosis in these terms, the models all fall short of explaining the full heterotic phenotype. Much of the research conducted in the area may only partially agree with one or more of these classical hypotheses. In addition, many of these models are nearly 100 years old and lack the molecular understanding of today. To this end, it has been suggested that these terms be abandoned in search of a more combinatorial explanation of F1 growth vigour taking into account modern genetic knowledge (Birchler et al., 2003; Chen, 2013).

Another caveat in understanding heterosis is the heterosis definition itself. Generally heterosis is used to refer to growth vigour of a hybrid individual, but different crops have different traits that make them desirable and so, sometimes hybrid vigour is used to describe the vigour of a particular trait. Here, heterosis is used to describe the specific phenomena of increased growth rate, biomass and fertility. Whereas hybrid vigour is used to describe the increase of a specific trait in an F1 hybrid relative to the two parents e.g oil production in oil palm (Jin et al., 2017). This way of looking at hybrid vigour as a series of increased traits with underlying mechanisms is currently more achievable than a unifying theory for heterosis. Due to the advent of next generation sequencing, various omics analyses have given promising insights into the genomic impact of F1 hybridisation at the DNA level, transcriptomic level and epigenetic level (Chen, 2013). Only from understanding the genomic impacts of hybridisation can an explanation for these complex phenomenon be found (Birchler et al., 2003).

1.1.3 Doubled Haploids

Haploid plants can be produced spontaneously in nature or induced by a number of different techniques. This was first described by Blakeslee et al. (1922) using the species *Datura stramonium*. Then similar reports followed for tobacco and many other species (Thomas et al., 2003). Typically, the observation of this phenomenon led to exploitation and integration with plant breeding programs. But this did not happen until 1964, when protocols for haploid embryo formation via *in vitro* culture was developed for *Datura* anthers (Guha and Maheshwari, 1964). However, haploid plants are unable to pair chromosomes during meiosis and are infertile. This makes them useless for plant breeding technologies and led to the discovery of chromosome doubling, which again can happen spontaneously (Murovec and Bohanec, 2012). However, chromosome doubling can also be induced through chemicals such as colchicine. Doubling of a haploid chromosome set produces doubled haploids (DH). They are used routinely in elite cultivar development for a variety of crop species because of their ability to produce homozygous lines in one generation (Ferrie and Möllers, 2011). Traditionally, up to 8 generations of selfing has been used to obtain plant lines with 99.2% homozygosity. However, DHs can achieve 100% homozygosity in only one generation allowing the easy fixation of traits in homozygosis. In self-pollinating species, these DHs can directly become cultivars or be used as parents in further crosses. This can reduce the time to cultivar release to less than seven years (Yan et al., 2017). These benefits were attractive and spurred on research for protocols for more than 250 species (Ferrie and Möllers, 2011). DH technology is widely applicable to many of the molecular breeding techniques introduced above due to their homozygosity. They

allow access to recessive alleles, the introgression of useful agronomic traits and facilitate mapping of QTLs (Filiault et al., 2017; Bakhtiar et al., 2014). Studies using DH lines also benefit from the number of lines that have associated genetic maps (Cogan et al., 2001).

A typical breeding program utilising DHs starts with the crossing of two genotypes. This creates hybrids with heterozygous genomes containing both parental chromosomes. Recombination in this hybrid during meiosis creates novel combinations of both of the parental genomes. Then through DH production these novel combinations are fixed in a homozygous state by doubling of the number of chromosomes. These lines can then be propagated as true breeding lines for phenotypic selection and scoring over multiple generations. DH homozygosity make phenotypic selection more efficient because their recessive alleles are directly expressed. They can also be used in a recurrent selection program which involves using the best DH lines as parents for future hybridisation and DH production. Multiple cycles of this can give gradual improvements. The two major steps; DH production and plant growth with selection are complex and require much time and resources. To this end many studies have focussed on the improvement of DH induction. This process requires specialist equipment and training, can typically take 4 weeks, and is often very inefficient. This is made harder by the fact that protocols are unique to each species and often even between genotypes of a species the same protocols will not work. Many studies have focussed on improving the embryo to plant ratio during in vitro culture. Studies have looked at the parental genotypes, the donor environment, pretreatments, different media adjustments, the culture environment and stage of microspore development with much success (Ferrie and Möllers, 2011). However, studies of improved selection methods

are few and far between even though this a more time and resource consuming step.

1.1.4 Typical Modern Breeding Strategies

A modern plant breeding program serves a large customer base of individual agricultural businesses. To increase the penetration of these markets it must achieve benefits in a wide variety of characteristics including yield, flavour, secondary metabolites and resistance to disease, pests and abiotic stresses. To achieve this, the programs will incorporate many of the above mentioned techniques. Even so, many breeding programs follow a similar strategy (Shimelis and Laing, 2012). This can be divided into two categories of work; pre-breeding and cultivar development. Pre-breeding is the process of introducing beneficial genetic variation into breeding programs by choosing "parental lines". These can be chosen from natural landraces; these populations are naturally heterogeneous and often, having been cultivated by farmers, possess the traits that the farmer wants as well as being adapted to the local environment. Parents can also be selected from previous cross performance or from being successful cultivars themselves with a desired trait. In cross-pollinating species, the most common route is to perform test crosses between these selected lines, the aim of which is to identify parents with high combining ability. This can be measured as an increase in the desired trait above that of the parental lines (Fasahat et al., 2016). Parents can be scored according their general combining ability (GCA) and their specific combining ability (SCA). This is a very resource-demanding step in the plant breeding process. Fasahat et al. (2016) have reported that up to 80% of resources in breeding programs are used on crosses that do not become cultivars. This is because, in the most thorough methods of hybrid scoring, $(n * n) * 2$ (where n is the number of parents) crosses

are needed to assess every interaction in a full diallel cross and phenotypic selection requires the growth of many individuals. Other crosses which reduce the need to cross every parent are available, however, even with these techniques, the workload is huge. This pre-breeding program is usually a continuous effort from a breeding company which runs alongside the cultivar development (Figure 1.1).

Once hybrids have been scored they can be released as superior F1 hybrid seeds, go onto cultivar development, or be used for MAB technologies. In non-molecular breeding strategies, to achieve true breeding, lines must be homozygous. As mentioned earlier this was traditionally done with selfing, but with more protocols available, DH lines are now the popular choice reducing this process by up to 8 growing cycles. This allows for the saved resources to be directed toward selection of more beneficial lines which can become cultivars or parents for future pre-breeding crosses. If marker-assisted breeding is being applied then hybrids are used for backcrossing and introgression; or can be selected based on them containing a marker at a young age, removing the need to grow and select negative plants (Figure 1.1). These processes are very lengthy and without MAB, no less than 8 but often more than 20 breeding cycles are required, which in the case of annual crops could be 20 years. With new MAB technologies the process can be as little as 2 years (Takagi et al., 2015) but a two-fold increase is expected.

Improvements in these processes are always required because we constantly require increased efficiencies from the same land or need new varieties that can grow where crops have previously been unsuitable. However, the process of cultivar development through traditional and modern genetic methods is time consuming and re-

source demanding and this is magnified in crops with longer life cycles. With increasing demand from emerging economies, the global population increase and heightened concerns over habitat conservation, improvements to cultivar development programs and insights into genome regulation are vital.

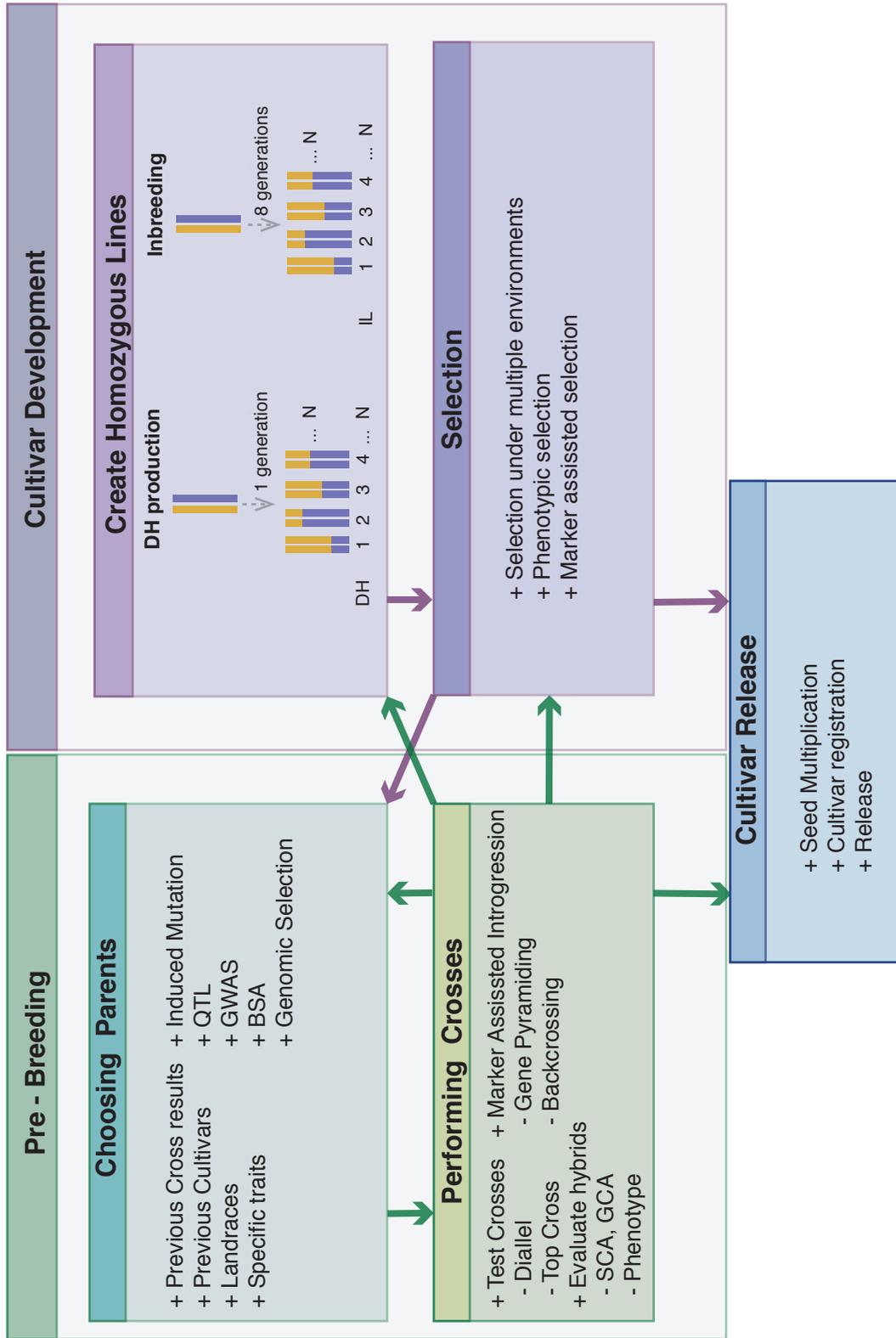


Figure 1.1: Simplified schematic of typical breeding strategies. Arrows indicate the movement of potential cultivars through the breeding program.

1.2 Genome Regulation in Plants

1.2.1 Gene Expression

Each cell is part of a constantly changing environment, both at the macro and micro level. To ensure that all signals are acted upon and growth can be regulated properly, each genome is finely tuned. At the phenotype level, this results in the plants adapting to their external environment and developing correctly. But at the cellular level, it is an orchestrated network of responses which are usually a series of well-ordered events, often with redundancies and some stimuli cause responses from multiple gene "networks" (Vihervaara et al., 2018). For example, the wounding response in *Arabidopsis* is controlled via MYC2 upon receiving signals from phytohormones. MYC2 then induces expression of ANAC019, ANAC055 and ANAC072. These transcription factors effect the expression of many downstream genes (Kazan and Manners, 2013). However, in the presence of ethylene and phytohormones, EIL1 and EIN3 are activated which in turn activate ERF1, PDF1.2 and ORA59 leading to a pathogen response. This is a typical complex cellular response and a similar story occurs even with normal cellular functions, redundant pathways and dosage dependant signal cascades. Even though signals can be detected and acted upon in different ways, the major driving force in many of these cellular responses is changes in levels of transcription.

All protein coding genes are transcribed by RNA POLYMERASE II (POLII). POLII cannot act alone and requires various factors to initiate binding to specific DNA sequences. Thousands of such factors have been identified that participate in regulated transcription. These are mainly other proteins, but also include RNA (Fuda et al., 2009). The main contributors to transcriptional regulation are the regulatory DNA

sequences existing at each gene locus and the transcription factors that are currently acting at these loci. This combination of specific transcription factors acting at specific DNA sequences dictates the resulting temporal, spacial and the magnitudinal gene expression. Multiple factors can act on a single gene and so the DNA regulatory elements can also have multiple elements. This is usually described by the core promoter, the proximal regions of DNA either side and the more distal enhancer regions. The core promoter is bound by general transcription factors (GTF) and governed by the different core promoter sequences forming distinct preinitiation complexes (Juven-Gershon et al., 2008). Specific TFs bind to promoter proximal and enhancer regions; these factors can change the level of transcription of genes by directly interacting with POLII, the GTFs or by re-organising local chromatin (Fuda et al., 2009). Then, the actual process of transcription can in turn be regulated in a number of ways: promoter opening, initiation, promoter-proximal pausing, elongation, co-transcriptional processing, termination and machinery re-cycling (Juven-Gershon et al., 2008). Regulation at so many steps of transcription shows the wide array of transcriptional responses that are required by a cell.

As well as these regulatory mechanisms consisting of the direct DNA sequence associated proteins, gene expression can be regulated by other external factors. These include: the epigenetic state of the localized DNA (chromatin and methylation), RNAs of various types including siRNA and miRNA, as well as local TEs. However, high throughput RNA sequencing has now become the tool of choice for analysing the levels of transcription in the whole genome (Wang et al., 2009). This is because it avoids the need for a priory of genes that are present in the organism which are required in older techniques such as the microarray. This means that with RNA sequencing one

can catalogue the different transcript species (mRNA, ncRNA and sRNA), identify the structure of genes and other transcripts and quantify the levels of these transcripts under a particular condition. As mentioned the cell is a constantly changing environment and RNA sequencing (RNA-Seq) garners a snapshot of the lysed mRNA in a particular tissue at that timepoint. In many cases this is sufficient to infer phenotype as for the most part, transcriptional levels and protein levels correlate well.

1.2.2 DNA Methylation

DNA methylation is conserved across animals and plants and most commonly occurs as 5 methyl cytosine (5meC), where a methyl group is added to the 5 position of the cytosine ring. In plants, methylation occurs at all cytosines regardless of the subsequent nucleotides, but due to understanding gained in *Arabidopsis thaliana* we will explore its properties grouped by the three commonly described sequence contexts; CG, CHG and CHH (where H is a T, G or A). This is because they are deposited and maintained by different mechanisms (Law and Jacobsen, 2011). Because of the different inheritance mechanisms it has been found in most plants that methylation levels within a tissue vary between the different contexts in a consistent manner. CG methylation is generally fully methylated or unmethylated, whereas CHG and CHH methylation have lower levels of methylation, suggesting a more mosaic distribution within the cells of a tissue. At each cytosine, the level of methylation is defined by the interplay between methylation and demethylation. Methylation is controlled by methyltransferases and de-methylation is controlled through the action of 5meC DNA glycosylases.

Methylation at CG sites is inherited during DNA replication by the action of

METHYLTRANSFERASE 1 (MET1). Hemi-methylated daughter strands are recognised by the VIM family proteins during mitotic DNA replication and recruit MET1 which subsequently methylates the newly synthesised DNA using the old strand as a template (Kawashima and Berger, 2014). Maintenance of the other symmetrical mark, CHG is maintained by a negative feedback loop involving CMT3, SUVH 4, 5, 6 and to a lesser extent CMT2 (Zhang et al., 2018). CMT3 binds to H3K9me2 histone modifications causing methylation to take place and the SRA domain of SUVH4 binds to CHG methylation to methylate the histones (Ebbs, 2006; Du et al., 2014). The asymmetrical CHH methylation is controlled by the most complex mechanism of the three contexts; the RNA directed DNA methylation (RdDM) pathway, relying on sRNA to guide DRM1 and DRM2 methyltransferases (Zhang et al., 2018). The methylation of CHH sites (and other cytosine contexts) starts with the binding of SHH1 to H3K9me2. This recruits RDR2, this polymerase transcribes 24nt sRNA loci. The sRNAs produced are sequestered by argonaught proteins and then direct DRM1 and 2 to the target sites (Kawashima and Berger, 2014). Further to this the CLASSY gene family also direct these methylation marks (Zhou et al., 2018). In addition, CHH methylation can be maintained through the action of DRM2 or CMT2 and DDM1. As for methylation removal, this can happen passively; methylation is not actively added to the sites. Or enzymatically; through the actions of the base excision pathway. In *Arabidopsis*, active demethylation is controlled by ROS1, DME, DML2 and DML3 which can excise bases from each of the sequence contexts.

Considering that 24%, 7% and 2% of CG, CHG and CHH sites have some methylation within the *Arabidopsis* genome it is reasonable to assume that there is an evolutionary advantage to applying and maintaining these types of methylation. At the

genome wide level, methylation in all contexts is highest at the pericentromeric regions and is generally lower in the gene rich portions of the genome. At the feature level, many studies have shown that methylation in all contexts is highest at transposable elements (Zhang et al., 2018). This high concentration of methylated cytosines across repetitive features such as TEs suggests that suppressing these elements is one of DNA methylation's primary roles. In plants such as maize, TEs can make up more than 70% with much of this having the ability to transpose to other locations and cause changes to the genomic DNA. This means that not only are they a source of variation for the plants, which is linked to their selection, but they need to be tightly regulated to ensure there is no damage. Methylation of transposable elements can prevent transposition because it can impair the transcriptional machinery (Zhang et al., 2018). However mutants of the methylation machinery often only produce a small number of transpositions owing to other post transcriptional silencing mechanisms (Mirouze et al., 2009; Kato et al., 2003). Even though DNA methylation is enriched for the peri-centromeric regions, there is still a large amount of DNA methylation found within genes but it has a more complicated relationship with gene expression.

DNA methylation at promoter regions usually results in decreased expression but in certain cases the opposite is true e.g in the cases of ROS1 and many fruit ripening genes in tomato (Lang et al., 2017; Lei et al., 2015). This can be because the methylation blocks the transcriptional machinery, or recruits transcriptional repressors, or can influence localised chromatin conformation and affect gene expression indirectly (Zhu et al., 2016). In *A. thaliana*, 5% of genes have promoter methylation. Usually this is a result of methylation spreading from nearby TEs, so it is reasonable to assume that other crop plants such maize would have a larger influence from pro-

moter methylation. This could be why defects in methylation genes are lethal for these species (Zhang et al., 2018). Methylation in the gene body (gbM) is also observed in most plant species: in *A. thaliana* 30% of genes are methylated (Zhang et al., 2006). Generally gbM is more associated with CG methylation and is only rarely directly correlative with transcription. Met-1 mutants lacking much of the gbM do not show increased or decreased expression overall and do not correlate with expression in the *A. thaliana* natural accessions (Kawakatsu et al., 2016). When looking at direct interactions it is perhaps more useful to look at a gene by gene basis. Examples include IBM1 (Rigal et al., 2016), RSM1 (Wibowo et al., 2018), MYB2, CIN1 (Wibowo et al., 2016). Few consistencies arise from the examples so the major signature is correlation between gene expression and methylation of a region nearby. However, methylation may have a role indirectly in the regulation of gene expression. DNA methylation is tightly linked to changes in chromatin: in *A. thaliana* reduced levels of H3.3 caused decreased gbM leading to reduced H1 linker histones and changed chromatin accessibility (Duan et al., 2017). gbM may also prevent aberrant transcripts by blocking POLII entries (Neri et al., 2017). Some genes also have nested TEs, these are usually heavily methylated and in some cases cause mis-expression of the gene if their methylation state changes. An intriguing example is the mantled phenotype in oil palm which is caused by the demethylation during clonal propagation of a TE inside the DEFICIENS gene which regulates floral development (Ong-Abdullah et al., 2015).

The genome is very complex regulatory network and for the most part, phenotype is a combinatorial effect from many inputs. Methylation in plants and most animal species forms a physical structure on the DNA and is a key part of this regulatory network. This is why methylation as a source of variation for plant breeding is now

more popular, although the stability of these marks makes them hard to harness in an effective way. The different contexts of methylation have different behaviours and mechanisms of deposition and maintenance. This makes the contexts of methylation have different effects and associations as epialleles (Niederhuth and Schmitz, 2017). The location of the methylation within the genome also affects the way the methylation is regulated (Sigman and Slotkin, 2016). DNA methylation has a well established role in repeat silencing. However, DNA methylation has a more complicated role with regulation of gene expression.

1.3 Merging Plant Genomes

The central dogma was first announced by Francis Crick in 1958 and revamped in a 1970 Nature paper simply stating that DNA is converted to mRNA then to protein (Crick, 1970). This was amazingly simple, but an idea of great importance that we have been building on ever since. In the 50 years since its inception we have advanced greatly in the tools we use to study genetics allowing us greater resolution and access to unparalleled amounts of information from the genome. We now understand that the genome is a complex regulatory system. Comprising many levels and layers of regulation that operate on DNA, mRNA and protein. These regulatory elements whether they be sRNAs, histone modifications, methylated cytosines or ubiquitins have evolved through selection over a long period of time and have evolved to be a redundant system often having multiple inputs controlling the same outputs. Through natural and artificial selection, desirable phenotypes and the DNA underlying those phenotypes accumulates preferentially in a population. Over much time this can cause vast differences in the DNA sequences between the organisms and speciation.

In nature this can happen naturally such as geographical isolation of parts of a population, but it can also happen in the process of cultivar development. As discussed earlier, in plant breeding it is common to make hybrids. A genome merger such as in a polyploid or filial 1 hybrid (F1) can result in unexpected changes to the transcriptome and epigenome of plants and studying this variation has led to insights into genome regulation, evolution and advanced plant breeding (Hu et al., 2016; Chen, 2013; Rigal et al., 2016; Kawanabe et al., 2016).

The molecular mechanisms underpinning this phenomenon have been studied at the transcriptomic and epigenomic level and the terms genome shock, transcriptome shock and epigenome shock (Rigal et al., 2016) describe the outcome of combining two genomes vividly. It has been examined in F1 crosses in a number of species including; maize, cotton, *Arabidopsis* and others (Greaves et al., 2014; Lauss et al., 2018; Groszmann et al., 2015; Hegarty et al., 2008) and also in polyploids (Qi et al., 2012; Yoo et al., 2013) with some consistent and varying results. However, the phenomenon has not been studied in doubled haploids. In an F1 cross the genomes of two parents come together in a heterozygous structure and undergo widespread change. The evolutionary distances of the parents is considered to be one of the main factors, with more divergence between the parents leading to more unexpected changes to the transcriptome and the methylome (Greaves et al., 2016; Chen, 2013). This has been exemplified in many species where additive and non-additive changes in the F1 are enriched for changes that already exist between the parental lines (Groszmann et al., 2015). Expression level dominance and homologous expression bias is also widely observed in F1 hybrids of different species. This is where the F1 preferentially assumes the expression or methylation levels of one parent. Studies show that the majority

of these changes are non sex-linked (Rigal et al., 2016; Yoo et al., 2013) although some examples do show the paternal and maternal effects for some genes (Kirkbride et al., 2015). An extreme example of this type of response is nucleolar dominance. This results in one parental set of rRNA genes being silenced preferentially (Chen, 2013). Further to these studies of F1 individuals an allopolyploidization event results in genome shock in a different way. In this scenario the doubling of all chromosomes results in additional gene dosage effects. Studies have been conducted in *Arabidopsis*, senecio, wheat, cotton (Yoo et al., 2013). In general allopolyploids have been shown to have less gene expression changes than their hybrids (Yoo et al., 2013). But there are similarities between these two forms of genome merger, in particular, many species show parental expression and methylation genome dominance.

The majority of these studies mentioned, utilise whole genome RNA sequencing as it garners a snapshot of mRNA abundance. mRNA levels of differentially expressed genes have been shown to correlate with changes at the protein level, and many studies have shown phenotypic effects associated with differences in gene expression (Kousounadis et al., 2015). However, the non-mendelian nature of the changes reported suggests an epigenetic involvement and indeed, many of the studies demonstrate altered DNA methylation upon genome confrontation (Rigal et al., 2016; Greaves et al., 2016).

1.4 Genomic Analyses in *Brassica* Species

Brassicaceae are part of the *cruciferae* family and encompass many species which are important crops for consumption by humans and animals (Lanner-Herrera et al., 1996). In fact, they may have been an important food crop as early as the neolithic period

(2650 to 2230 BCE). They are described by the triangle of U which is a cytological classification of the genomes of the *Brassica* species (Nagaharu and Nagaharu, 1935). It contains three main genomes of *B. rapa* (AA; $2n=20$), *Brassica oleracea* (CC; $2n = 18$) and *B. nigra* (BB; $2n = 16$) along with their allotetraploid hybrids (*B. juncea* AABB; $2n=36$), (*B. carinata* BBCC; $2n=34$) and (*B. napus* AACC; $2n= 38$). The *B. oleracea* species described by the CC genomes is arguable the most important *Brassica* for human consumption. This is because of the wide variety morphology that breeders have managed to produce from this genome. There are 14 different cultivator types available in *B. oleracea*, each with a unique selling point (USP) (Kays and Dias, 1995). They include; Cabbage, Kale, Broccoli, Brussel Sprouts and Kholrabi. They also provide health benefits through the production of glucosinolates, a secondary metabolite that is almost exclusively produced by this plant family (Higdon et al., 2007). Because of their commercial value, the *B. oleracea* species now benefits from a large variety of genetic tools, including many genetic marker maps (Cogan et al., 2001; Gao et al., 2007) and comprehensive reference genomes (Parkin et al., 2014; Golicz et al., 2016).

Many recent omic's studies have been undertaken in *B. oleracea* reviewed by Witzel et al. (2015). These include; studies of DNA methylation, QTL analysis, transcriptomics, proteomics and metabolomics (Gao et al., 2014; Parkin et al., 2014; Walley et al., 2012). Genomics studies have been centred around the exploitation of DNA markers to elucidate QTLs. QTLs for flowering time, developmental genes, stress response and secondary metabolites have been identified in different *B. oleracea* cultivars (Li et al., 2013). The methylome of *B. oleracea* leaves has also been studied (Parkin et al., 2014). Although RNASeq has been around for more than 10 years,

transcriptomic studies of the *B. oleracea* species are fairly limited, this is probably because the reference genomes of this species have only just become available. The transcriptomic studies that have been performed have been mainly focussed on abiotic stress responses and not related to plant breeding directly (Witzel et al., 2015). The deeper the understanding we have over the genomes of these species, the more we can apply and exploit the principles in plant breeding.

1.5 Aims and Hypothesis

As discussed, DH lines are used extensively in the commercial sector and often many resources are wasted during DH line selection and parental selection for crosses. Along with this, DH lines are also another form of the genome merger which has been extensively studied in F1 diploid hybrids and polyploids. To this end, the overall aim of this thesis is to identify any variation in the transcriptome or methylome that exists in the F1 or DH lines in these samples of *B. oleracea*. Then assess whether any of this variation discovered can be used to improve current DH breeding programs or improve current understanding of genome regulation in the light of the genome merger. These aims will be addressed with whole genome bisulphite sequencing and whole genome RNA-Seq data from a doubled haploid breeding program consisting of two parental lines, the F1 hybrid and nine DH lines. The general hypotheses being addressed in the thesis are as follows:

- Are there any detectable transcriptional differences between the parental lines, the F1 hybrid or the DH lines within this data set?
- Are there any detectable differences in genomic methylation between the two

parental lines, the F1 hybrid or the DH lines within this data set?

- If any changes are detected, are any of these changes applicable to the study of genome mergers or useful in current DH breeding programs?
- Is there any direct correlation between gene expression and methylation?

Chapter 2

Methods

2.1 Creation of F1 and Doubled Haploid Lines

The population for this experiment was selected from a larger mapping population initiated by Sebastian et al. (2000) which was originally designed as an RFLP mapping population. Two doubled haploid parents *B. oleracea ssp. italica* (GDDH33: D.J. Keith, John Innes Centre, Norwich) and *B. oleracea ssp. alboglabra* (A12DHd: D.J. Keith, John Innes Centre, Norwich) were crossed to create a set of F1 hybrids with A12DHd as the female parent. Then, as described in Chuong and Beversdorf (1985). Young flower buds from the main raceme and lateral branches were washed with 5.25% (v/v) solution of sodium hydrochlorite for 10 minutes to achieve surface sterilization and further washed with deionized water three times. Yellow-green anthers from buds where the petals were approximately half the length of the anthers were extracted for microspore culture. The anthers were then macerated in a solution of B5 media (G0209, Duchefa Biochemie) (supplemented with 135 (w/v) sucrose) and the resulting solution filtered through a 44 um nylon screen and centrifuged for 3

minutes at 100 X g before collection and washed 3 times with washing media. These microspores were then cultured for 4 weeks in darkness in NN media (N0223, Duchefa Biochemie) with some alteration (Devoid of Difco potato extract, but containing 13% sucrose, 0.05 mg/1 BA and 1.00 mg/1 NAA, titrated to pH 6.0 and filter sterilized) first for 3 days at 35 degrees and then for remaining time at 25 degrees. After 4 weeks the resulting embryos (when the embryos reached 0.5 mm in length) were transferred to solid B5 media lacking growth regulators and NN media but with the addition of 2% (w/v) sucrose and 0.8% (w/v) agar. These embryos were then maintained in the media with 16h photoperiod at 25 degrees for a further 4 weeks for plantlets to develop. Haploid and doubled haploid plants were then multiplied via cuttings to increase the population of the G0 plants. This population is housed at the Warwick crop centre under the accession BolAGDH where they have further selfed and catalogued these lines.

2.2 Selection of Samples and Plant Growth

The population BolAGDH described above benefits from marker map in which many DH lines have been genotyped. However, when choosing samples, we were limited by the lines in which seeds were available for the early generations. Generation 0 was not considered as this generation is the original DH and has been subject to severe stress including heat treatment, hormones, and long periods of culture (Bohuon et al., 1996). This has been shown to affect the genome of the initial doubled haploids in Barley (Li et al., 2007) and wheat (Machczyn et al., 2012) along with this, the G0 generation are single plants and so do not offer replication or the chance to re-grow lines. The samples analysed were grown in two batches. The first batch contains both

parents (A12Dhd and GDDH33) the F1 hybrids and then 3 DH lines (3238, 1047 and 1003) these lines were grown in greenhouse conditions (planted at the end of June). Following this pilot a further 6 DH lines were selected based on their whole genome contribution from each parent utilizing the genetic maps. This resulted in DH lines being chosen that display a full range of genome contributions from both parents. The second batch (consisting of both parents and lines 2069, 3088, 2134, 5119, 5071 and 3013) were grown also in greenhouse conditions (planted at the end of June). In each batch, plants were grown for 5 weeks and then 5 plants from each line were taken for analysis. Leaves 4-6 were chosen based on visual inspection of any damage, then from the middle of the chosen leaf from each plant to the apex was removed using scissors. The 5 leaves from each group of 5 plants were then combined and leaf material was flash frozen in liquid nitrogen macerated in a pestle and mortar and divided into two for RNA and DNA extraction. This resulted in technical replicates for all DH samples and no biological replicates for either the bisulphite sequencing or RNA sequencing. A12Dhd, GDDH33 has two biological replicates, but these were grown in the two separate batches. Then F1 whose two biological replicates were both grown in the first batch (Figure A.4, A.3, A.2, A.1).

2.3 RNA Extraction Data Generation

Total RNA was extracted from the leaf material using RNeasy Plant Mini Kit (Qiagen) according to manufacture instructions. Libraries were created using the Illumina Stranded total RNA library kit. These were amplified and sequenced as 150bp reads on an Illumina Hiseq 2000.

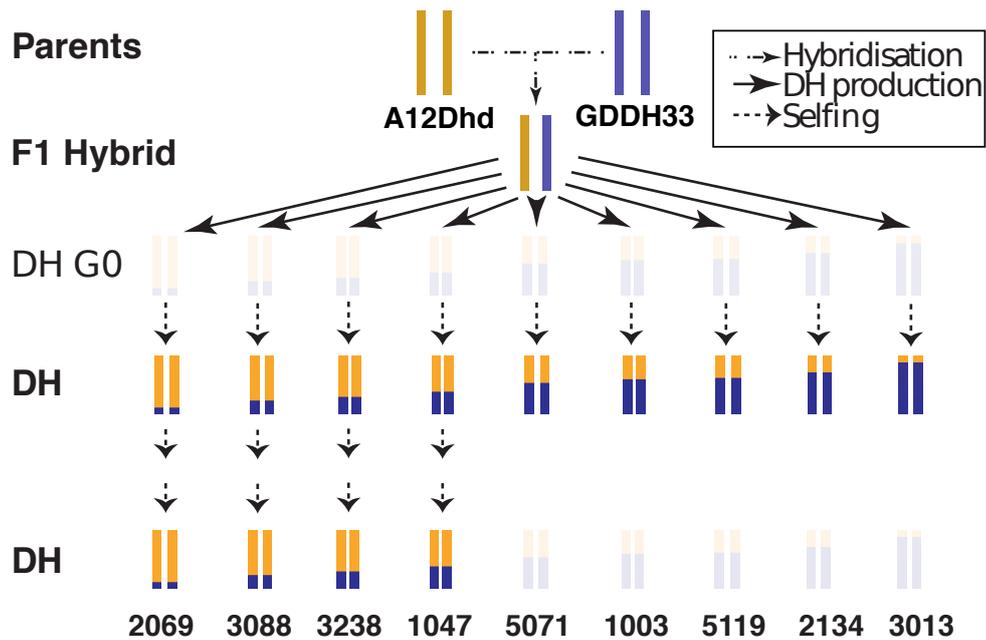


Figure 2.1: **Schematic of samples in this study along with their method of creation.** Samples in bold indicate those with whole genome bisulphite sequencing data and whole genome RNA sequencing data. Line names or numbers are shown underneath each line and the arrows indicate each sample's method of creation.

2.4 RNASeq Data Processing

The libraries were assessed for quality using FastQC (Andrews, 2010) and low quality reads were trimmed using Trimmomatic (Bolger et al., 2014) (Parameters - *ILLUMINACLIP : adapters.fa : 2 : 30 : 10*, *HEADCROP : 6*, *LEADING : 3*, *TRAILING : 3*, *SLIDINGWINDOW : 4 : 15*, *MINLEN : 3*). SortmeRNA Kopylova et al. (2012) was then used to remove remaining rRNA contamination and reads were then aligned to the kale-like TO1000DH reference genome (Parkin et al. 2014) using Tophat 2 with default parameters (Trapnell et al., 2012). Raw gene counts were obtained from the python package HTSeq-count (Anders et al., 2015). Differential gene expression was analysed using DESeq2 (Love et al., 2014) and a gene was considered differentially expressed if it experienced a log₂ fold change

>0.5 and an FDR-corrected p-value <0.05.

2.5 DNA Extraction

Genomic DNA from leaf material was extracted using the DNAeasy Plant Kit (Qiagen). Libraries were then created with the Illumina TruSeq Nano kit (Illumina, CA, USA) according to the manufacturers instructions. Fragments of 300bp were then size selected. Bisulphite conversion was performed with Epitect Plus DNA bisulphite Conversion Kit (Qiagen, Hilden, Germany) after adaptor ligation. Library enrichment was performed using the Kapa Hifi Uracil+ DNA polymerase (Kapa Biosystems, MA, USA) according to the manufacturers instructions. and sequenced as 100bp reads paired end reads on an Illumina HiSeq 2000 instrument.

2.6 Bisulphite Data Processing

Reads were first assessed for quality using FastQC (Andrews, 2010) and then trimmed for low quality sequences using Trimmomatic (v0.33) (Parameters *ILLUMINACLIP : adapters.fa : 2 : 30 : 10, HEADCROP : 6 , LEADING : 3 , TRAILING : 3 , SLIDINGWINDOW : 4 : 15 , MINLEN : 3*) (Bolger et al., 2014). Bismark (Krueger and Andrews, 2011) (parameters *-n2, -l28, -p3*) was used to align all reads to the kale-like TO1000DH reference genome (Parkin et al., 2014). Duplicates were removed using GATK (McKenna et al., 2010) and then -CX report files were generated using bismark methylation extractor with (parameters *--bedGraph, --CX, -cytosine_report*). Statistics from single cytosine methylation were parsed from these files and they are also the substrate for calling differentially methylated regions (DMRs) and methylated regions (MRs).

2.7 Single Cytosine Analysis

Analysis of single cytosines was performed for the parents and F1 hybrid. To assess the distribution of methylation at all cytosines, an RScript was created that first filters cytosines for a coverage of at least 5 reads. Then the average methylation value for each cytosine is calculated and the average and histogram of these values is plotted. For the methylation over TEs and genes, a Perl script was created that takes an input the CX report file and outputs a data frame which contains the relative position of the cytosine to the nearest TSS (If within 2Kb) https://github.com/PriceJon/thesis_scripts/blob/master/split_cytosines_on_GFF.pl. Cytosines are removed with less than 5 reads covering and if they overlap with more than one feature e.g gene and 2Kb upstream from another gene. These values are then plotted using ggplot with the geom smooth function.

2.8 Methylated Regions

To call methylated regions, the original computeDMR.R scripts in the DMRCaller package (Catoni et al., 2018) was altered such that it allows the program to call regions that are methylated from each sample. The script first bins the genome and then filters based the same criteria of DMRCaller. Using the same parameters as computeDMR.R this script can be used to identify all methylated regions that may or may not be differentially methylated regions. Used in conjunction with DMRCaller can be used to identify enrichment within the DMRs. parameters; MeC required for MR; CG = 0.6, CHG = 0.35, CHH = 0.2. Other parameters; Bin size = 200, minCyt = 4, minReads = 4, minGap = 150, pValueThreshold = 0.01.

2.9 Differentially Methylated Regions

DMRs were called with R package, DMRCaller (Catoni et al., 2018). To keep track of regions and to allow for direct comparisons with methylated regions (MRs) the bin method was chosen. This algorithm splits the genome into equal size bins. Then the algorithm looks for regions that have enough cytosines with enough coverage. Then check that the difference in methylation percentage is big enough and that the statistical test produces a significant result. The program uses a score test for significance: if the number of reads for a bin in condition one and two is denoted by n_1 and n_2 . And the number of cytosines from those reads with methylation is denoted by m_1 and m_2 . The percentage of methylation are given by $p_1 = m_1/n_1$ and $p_2 = m_2/n_2$. Then the total percent p and the coverage index v can be given by:

$$p = \frac{m_1 + m_2}{n_1 + n_2}$$

and

$$v = \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

Then the Z score:

$$Z = \frac{(p_1 - p_2)v}{\sqrt{p(1-p)}}$$

The Z score created is then represented as a p-value assuming a normal distribution and two-sided t-test. As shown in this study and others, methylation in the three sequence contexts display different distributions. This means that looking for the same methylation difference is not applicable for the 3 different contexts. From the analysis of the distributions, the required methylation difference was set for each sequence

context as CG = 0.6, CHG = 0.35, CHH = 0.2. Other parameters; Bin size = 200, minCyt = 4, minReads = 4, minGap = 150, pValueThreshold = 0.01.

2.10 Parent and Hybrid Differential Expression and Methylation

DMRs and DEGs were called pairwise between the three parent and hybrid genotypes. These three comparisons comprise all significantly differentially expressed genes or methylated regions that exist between the parents or either parent and the F1 hybrid. For each of these genes the d/a ratio and the modified ratio described in Guo et al., (2013) is calculated. The d/a ratio describes the expression of the F1 relative to the parent with high or low expression. If \bar{x} is the average expression between A12Dhd and GDDH33. Then:

$$d = ExpF1 - \bar{x}$$

and

$$a = ExpHP - \bar{x}$$

Expression in the F1 can then be viewed in terms of the high or low parent with the ratio d/a (Parent with highest or lowest expression at that locus). A gene with expression most similar to the parent with higher gene expression would have a d/a ratio of 1, similarly a d/a ratio of -1 would mean expression was identical to the expression of the lowly expressed parent. If the F1 has expression equal to the MPV then the d/a ratio would be 0. In the modified parental d/a ratio the expression of the hybrid is described according to the A12Dhd or GDDH33 parent (Figure 2.2). Here:

$$d = ExpF1 - \bar{x}$$

and

$$p'a = ExpA12Dhd - \bar{x}$$

The only difference here is that a ratio of 1 or -1 now shows the F1 as having expression most similar to A12Dhd or GDDH33 respectively. Using the d/a and the d/p'a ratios together allows for the clustering of phDEGs into 12 mutually exclusive categories which are commonly used to describe gene expression (Yoo et al., 2013). See Figure 2.2 for a schematic of this analysis.

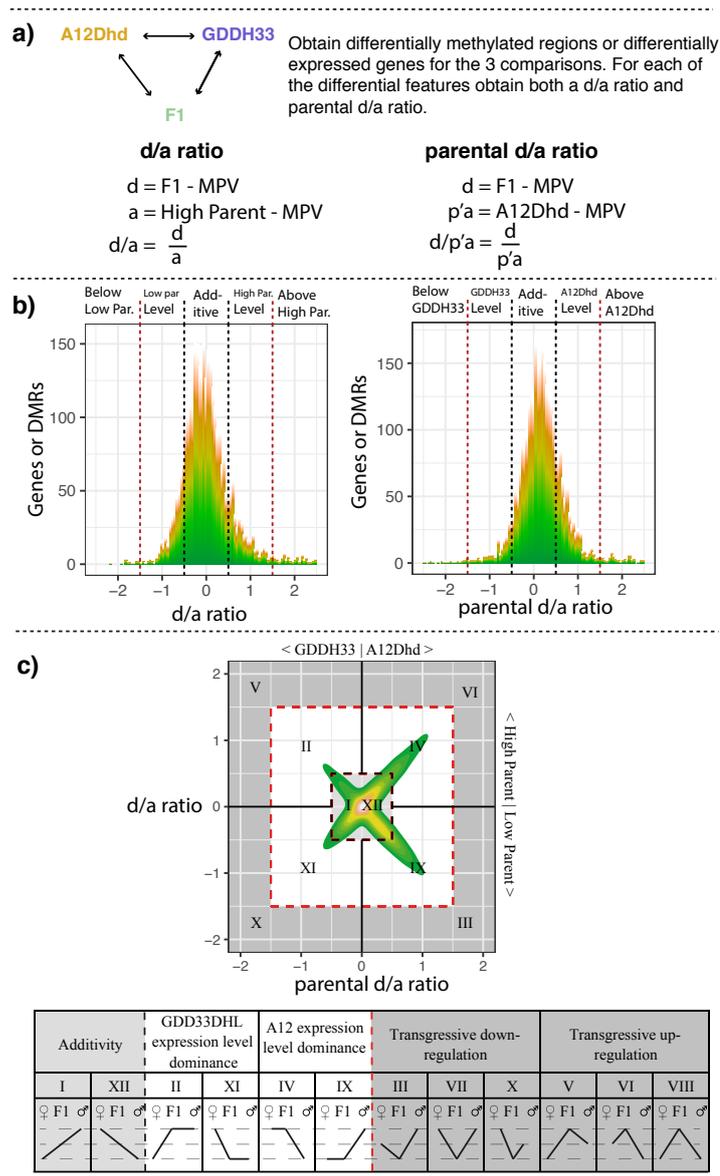


Figure 2.2: **Schematic showing how the d/a and parental d/a ratios are calculated and plotted. The ratios are used to show how the dynamics of a DMR or gene in F1 hybrid relate to the parental methylation or expression.** a) The calculation of the ratios. Firstly, three pairwise comparisons are performed (A12Dhd - F1, GDDH33 - F1 and A12Dhd - GDDH33). Then for each of these genes or DMRs shown to be significant in at least one comparison, two ratios are calculated. The d/a ratio and the parental d/a ratio. b) Displays the meaning of the ratios. The d/a ratio (left histogram) describes the methylation of the DMR or expression of the gene in the F1 according to the high or low parent (parent with highest or lowest expression). The parental d/a ratio (right histogram) describes the methylation of the DMR or expression of the gene in the F1 according to the expression of the maternal parent (A12Dhd) or the paternal parent (GDDH33). The histograms show the thresholds imposed on these ratios that decide the expression or methylation category (additive, parental-level dominance or above / below parental levels. c) Plotting and display of the ratios and categories. In the top plot, each genes ratios are plotted, the d/a ratio on the y-axis and the parental d/a ratio is plotted on the x-axis. Plotting in this way, each differentially expressed feature can be categorised according to both the high / low parent and the maternal / paternal parent. The bottom table of c) shows this categorisation. Roman numerals show the categories as they are commonly described (Yoo et al., 2013). Underneath the Roman numerals in the table, there is a graphic displaying the expression or methylation pattern of this category for the 3 genotypes (A12Dhd - maternal, GDDH33 - paternal and F1).

2.11 Homologous Recombination Site Detection

The bisulphite sequencing data from the parents and DH lines facilitates two genotyping methods; genotyping with SNPs and genotyping using DNA methylation (epigenotyping). Using both methods together can give the most accurate view of the crossover landscape in the doubled haploid lines. SNP genotyping utilises single nucleotide polymorphisms and the homozygous structure of the parents and DHLs to identify the parent of origin of the DHLs. I developed a pipeline in the Perl language capable of SNP genotyping through bisulphite sequencing https://github.com/PriceJon/thesis_scripts/tree/master/SNP_DH_genotyping_pipeline. The program requires as input .vcf files of the two parents and at least one DH line, these .vcf files can be obtained from the program MethylKit (Akalın et al., 2012). The program first identifies positions in both of the parental lines which have an allele frequency of 1.00 and have sufficient read coverage (10 reads). The program then checks each of these positions in the other parent for its basecall. If the position in the parent is also homozygous, has enough coverage and is different to the other parent. This position is then considered a distinguishing SNP and can be used to genotype the DH lines. Once the full SNP landscape has been found, the second part of the program concentrates on identifying HR sites. Using .vcf files for each of the DH lines it categorises the bases according to the parent of origin. After the DH lines genome has been binned into A12Dhd or GDDH33 the program then looks for switches in parental SNPs throughout the genome. To ensure that each HR site was a genuine, switches are only chosen by the program if there is at least 10 high quality markers on each side of the HR site

(Figure 2.3). To further verify the HR sites identified by the developed pipeline, we used epigenotyping. Epigenotyping exploits the differences in the methylomes of the two parents and categorises the chromosome segments of the doubled haploid lines according to its methylation status, we used the pipeline outlined in Hofmeister et al. (2017) with a few alterations; we used only CG methylation, we used altered class weights (Mother-0.5, Mid-parent value-0, Father-0.5) and lastly we used bin sizes of 150kb, 70kb and 60kb. These epiHR sites and the original SSR markers were then used to verify the placement of the snpHR sites.

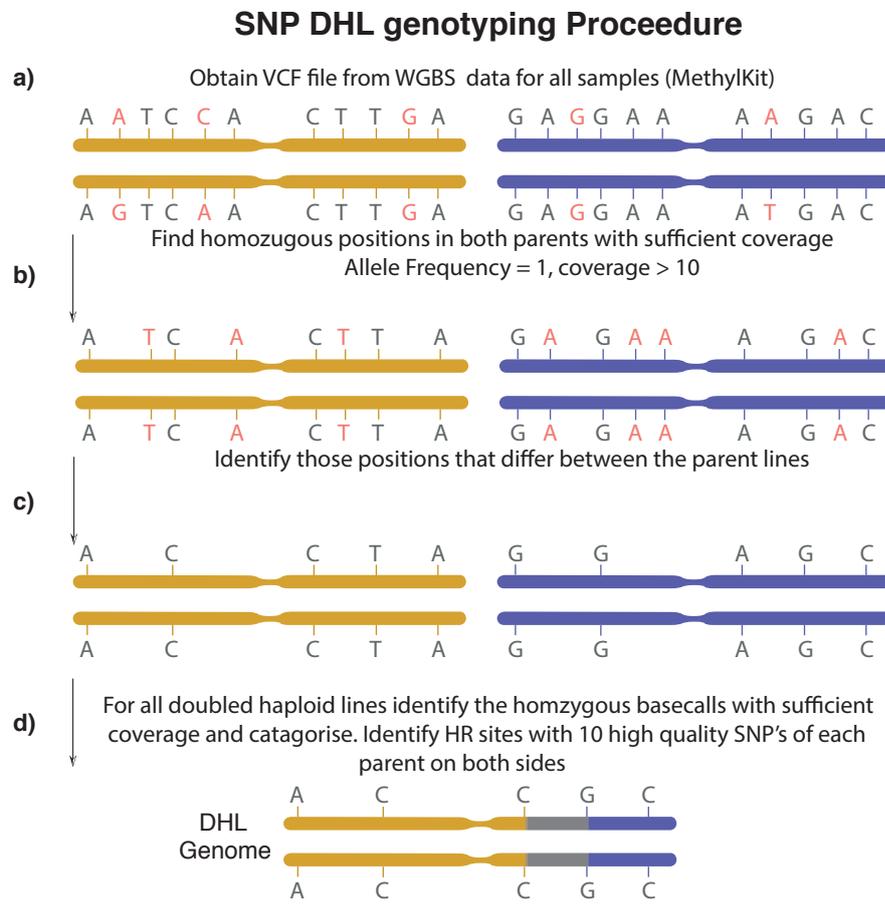


Figure 2.3: **Schematic of SNP genotyping pipeline.** a) From vcf files of each parental genome (A12DHd - yellow, GDDH33 - blue), positions in each genome are only kept if their allele frequency is equal to 1 and their coverage is greater than 10. b) Positions between the parental genome are compared, they are kept if the base call in each parental genome is different and both genomes have received greater than 10 coverage. c) This is the set of distinguishing positions that can separate the DH genomes. d) The final step of the program introduces the vcf file from a DH line and looks for positions with an allele frequency of 1, it then categorises each position according to the set of parental SNPs identified earlier. The program identifies a HR site if there is a switch in the parental inheritance with 10 SNPs on either side of the HR site.

2.12 Differential Expression and Methylation in the DHLs

Each DH lines genome was split into A12Dhd inherited or GDDH33 inherited from the HR sites detected. Then comparisons of DMRs of DEGs were only drawn between directly inherited regions between the parental lines and the DH lines. Relative gene expression change in these genome segments is termed as the percentage of genome inherited / number of DMRs or DEGs. Linear regression was performed on these values (percentage ownership relative change) using the `lm` function in R.

2.13 Gene Ontology Analysis

Gene ontology analysis was performed with BiNGO (Maere et al., 2005) with the custom GO database option. The GO database for *B. oleracea* was downloaded from PLAZA4.0 (Van Bel et al., 2018). In the case of the parents and parent hybrid different lists of DEGs were entered into BINGO via cytoscape then a hypergeometric test was used to identify overrepresented terms (FDR <0.01). In the case of the DH lines, DEGs for each line were entered into BINGO via cytoscape and then the merge network function was used to make a consolidated network. Then common enriched terms were identified in each of the DHL lines and coloured in cytoscape. GO terms only enriched in one line were removed if they did not have a child process that was enriched in more than one line.

2.14 Intersection of DMRs and Genes and Transposons

Proportion of DMRs residing within different features is calculated by taking each base of the DMRs in a particular list and assigning them to one feature type. This is done in a hierarchical fashion to account for overlapping features (gene, transposon,

upstream, downstream, intergenic: in order of decreasing importance). Further to this, I developed the GFFintersector program for looking into the gene intersections further (www.github.com/PriceJon/GFF_Intersector). Utilising GRanges, the program can efficiently identify all overlapping regions with a user supplied flanking region and visualise the results. The program identifies all genes that have at least one DMR within the gene region plus and minus the user supplied flanking regions. See Figure 2.4 for 3 examples. Figure 2.4 a shows the simple case in which there is a gene which is overlapped by 5 DMRs, this becomes an intersect regions containing 1 gene and 5 DMRs. In Figure 2.4 b, the intersect regions contains 2 genes and 4 DMRs, the two genes are contained within one region because they are less than the user supplied flanking region (f) apart and both genes have at least 1 DMR intersecting. In the last case, Figure 2.4 c shows a similar example, but in this example the second gene is not included in the intersect region because it has no associated DMRs within f of the TSS or TES. After identifying all intersect regions, the program is capable of visualising results genome wide and at the intersect region level.

The output of this program is a file showing all intersect regions. These intersections were then taken and a further R program was developed that can correlate the methylation and expression values. The first step is to identify DMR blocks, these are regions that contain at least one DMR within an intersect region Figure 2.5. These DMRs are collapsed if of the same context and in the same place in the genome. Once collapsed, the comparisons that need to be made are identified before filtering. For each comparison, the DMR block is checked to see that there is greater than 5 reads on average covering each cytosine in at least 5 samples. Then the genes are checked to ensure that expression is observed in at least one sample and there is at least a 1.2 fold

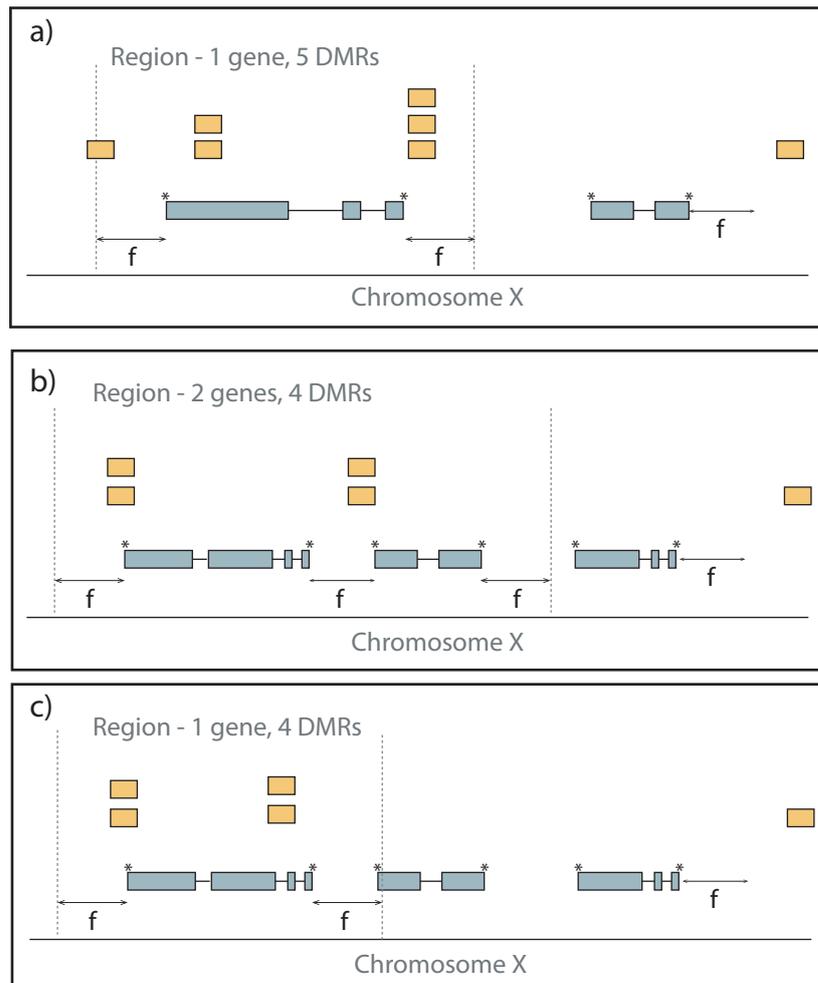


Figure 2.4: **Three examples of possible intersections from the GFF intersector program.** a) The case in which there is a gene which is overlapped by 5 DMRs, this becomes an intersect regions containing 1 gene and 5 DMRs. b) The intersect regions contains 2 genes and 4 DMRs, the two genes are contained within one region because they are less than the user supplied flanking region (f) apart and both gene has at least 1 DMR intersecting. c) In this example the second gene is not included in the intersect region because it has no associated DMRs within f of the TSS or TES. Yellow boxes show user defined regions e.g DMRs. Blue boxes show exons of genes, the TES and TSS are defined in the graphics with stars.

difference in expression between the highest and lowest sample. Then, using spearman's rank correlation, the program produces an output with the significance values and statistics for each region. From here, an FDR is calculated from the p-value and any results with $FDR < 0.01$ are considered significant. This candidate list was then manually investigated.

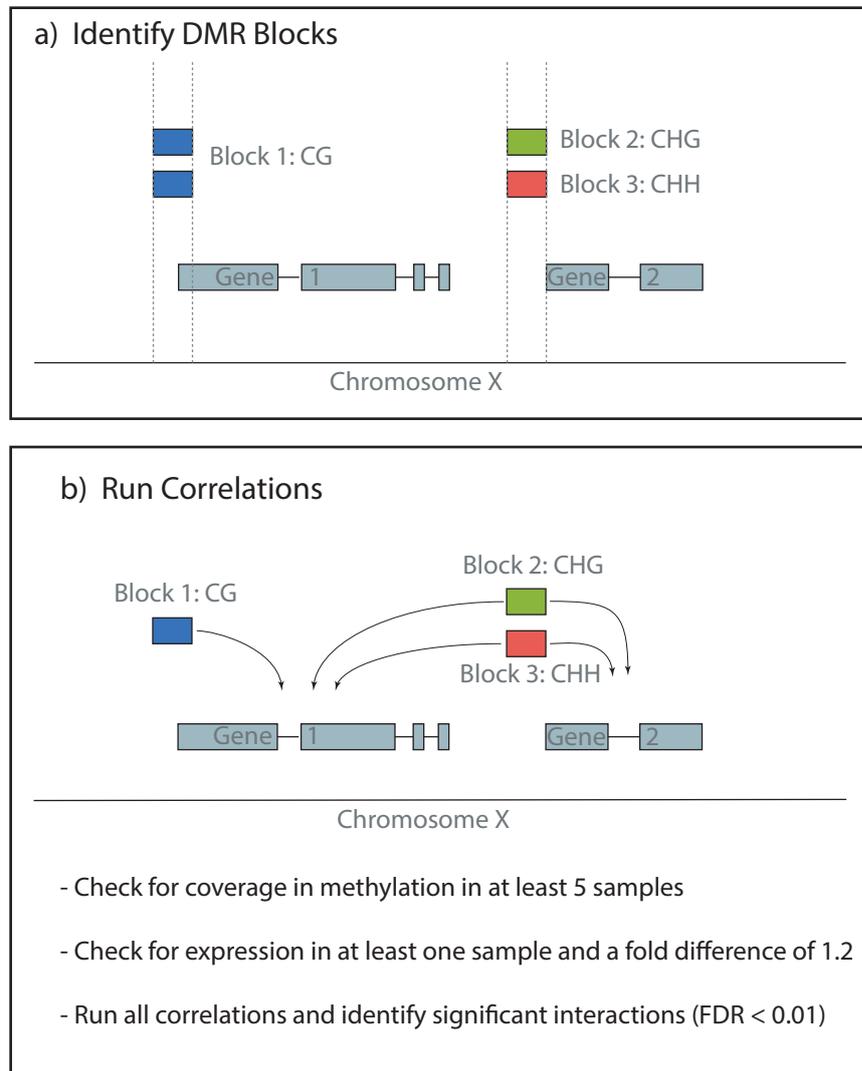


Figure 2.5: **Schematic showing the process of correlating gene expression and DMR methylation within regions identified using the GFFIntersector program.**

a) Shows the first step of correlation program in which DMRs of the same sequence context that directly overlap are combined into DMR blocks. b) Correlation of the DMR blocks and genes. Correlations are made between all DMR blocks and any gene within the flanking region of the DMR block, this maybe more than one gene. For each gene and DMR block comparison (5 in the above example), the coverage in methylation in each sample is assessed to ensure 5 reads are covering the region in at least 5 samples. Then the gene is checked for expression in at least one sample and a 1.2 fold difference. Then for each comparison identified, Spearman's rank correlation is performed. From all comparisons those with are FDR of less than 0.01 are considered significant.

Chapter 3

Transcriptome and Methylome

Analysis of A12Dhd and GDDH33

Parental Lines

3.1 Introduction

This study is focussed on the analysis of DH lines of *B. oleracea*, but to understand the changes that occur in the transcriptomes and methylomes of the DH lines, it is important to understand the parental lines and how they differ. The two parents in this study are themselves isogenic lines generated by DH technology (Sebastian et al., 2000). They were originally chosen for DH populations because they are polymorphic and would provide a good marker set for MAB technologies (Bohuon et al., 1996; Sebastian et al., 2000). Indeed they did and the marker set and DH population created has been used in a number of QTL studies with success (Walley et al., 2012). As well as their sequence divergence, the two parents differ in their morphology; The

A12Dhd parent is a Chinese Kale (*B. oleracea ssp. alboglabra*), from now on referred to A12Dhd. This parent has been selected for high leaf growth and fast generation cycling time. The GDDH33 parent is a broccoli (*B. oleracea ssp. italica*) now referred to as GDDH33 and has been selectively bred for its flowers. Because these species have the benefit of a molecular marker map they have been studied in a number of QTL studies. These include; flowering time, leaf morphology, calcium and magnesium variation, seedling vigour and secondary metabolites (Walley et al., 2012). Ngwako (2003) took measurements of various morphological features at different time-points for both parents in the hope of identifying QTLs. They showed that A12Dhd has larger leaves and a larger fresh weight at harvest than GDDH33 (Table 3.1). Many QTL for these traits were identified and it was suggested that a number of QTL may control multiple phenotypes through cross talk in regulatory networks. However they were not able to fully explain the variance in the leaf traits through QTL analysis (Ngwako, 2003). Bohuon et al. (1998) showed that A12Dhd has a much shorter flowering time but this trait was attributed to possibly more than 5 QTL including CONSTANTS. Another trait that has been analysed is the glucosinolate content, these secondary metabolites are important for human health and have received widespread interest. Issa et al., (2010) show that the leaves of A12Dhd and GDDH33 have different glucosinolate profiles. This is due to the different biosynthetic pathways that exist in these species. They managed to explain the majority of these differences by presence and absence of genes in either parental progenitor.

These experiments have only been possible because of the resources available for *B. oleracea* and the dedicated maintenance and documentation of *Brassica* lines by institutes such as Warwick Crop Centre (Walley et al., 2012). Since these experiments,

Chapter 3. Transcriptome and Methylome Analysis of A12Dhd and GDDH33 Parental Lines

Table 3.1: **Phenotypic measurements of A12DHd and GDDH33 made by Ngwako (2003)**, raw data was unavailable and so actual measurements are approximated here. Parent in bold shows the parent with highest value of that trait at that time point. The measurements are given below are the average measurement. (GDDH33 - A12Dhd)

	# of Leaves	Largest leaf length (cm)	Largest leaf width (cm)	Largest leaf petiole (cm)	Flowering Time (DAS)	Fresh Weight (g)
40 DAS	A12DHd	A12DHd	A12DHd	A12DHd		
	6.5 - 7	8.1 - 12.2	7.0 - 8.0	4.1 - 5.1		
67 DAS	A12DHd	A12DHd	A12DHd	A12DHd		
		22.0 - 24.0	12.0 - 15.0	7.8 - 9.8		
First Flower					GDDH33	
					85 - 60	
116 DAS						A12DHd
						300 - 500

sequencing has become widespread and reference genomes have been made for *B. oleracea*. The reference genome was created from T1000DH (*B. oleracea* var. *al-boglabra*), this is a Chinese Kale, very similar to A12Dhd (Figure 3.1) (Parkin et al., 2014). Then in 2016, the pan-genome of *B. oleracea* was made. It was built upon the Parkin et al. (2014) reference but included transcriptomic and genetic information from 9 more lines (Golicz et al., 2016). They show that the majority of the different *Brassica* lines share the same core genome (81%). Only 2% of the genes are unique to one line. Out of the 9 lines, TO1000 has the the second largest amount of genes (500) missing that are present in the other lines and the broccoli has 195 which are not present in TO1000 (Figure 3.1). The actual number of different genes between these genotypes will be larger than this value because 7000 genes were not included in this analysis. These particular *B. oleracea* lines are popular for QTL studies because of their marker map therefore the transcriptome and methylome sequencing of these lines will not only provide a basis for the analysis of the F1 hybrid and DH lines in this study but be a useful resource for other studies utilising these lines.

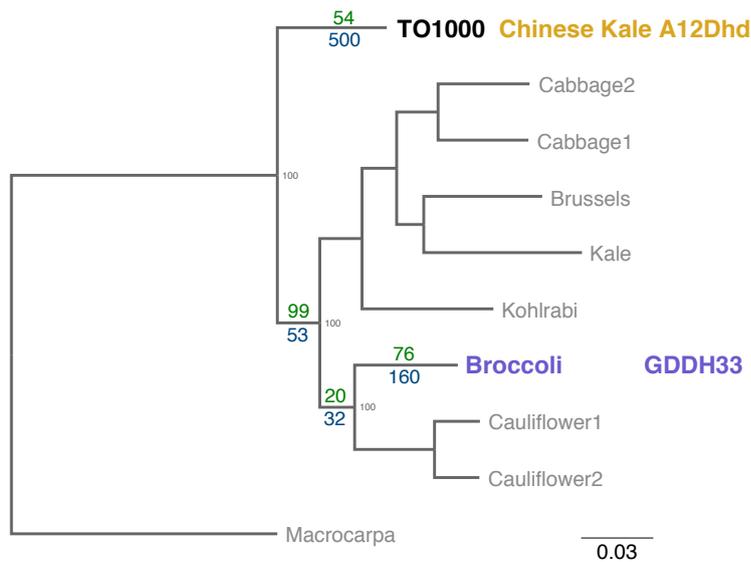


Figure 3.1: **Genetic distance between *B. oleracea* varieties.** Dendrogram, modified from Golicz et al. (2016) shows the genetic distance between the 9 lines used by Golicz et al. (2016). Numbers in green show genes that are present in the varieties below that node but not present in the others. The blue numbers represent the number genes that are not present in the lines below that node but present in the other lines. The scale indicates the number of nucleotide substitutions per site.

3.2 Chapter Aims and Hypothesis

The overall aim of this chapter is to set out and understand the transcriptomic and epigenetic variation that exists between the parental lines A12Dhd and GDDH33. This aim will be addressed by the analysis of whole genome RNA-Seq by looking at the differential expression of genes and whole genome bisulphite sequencing of the two parental genotypes, looking at both single cytosine positions and then DMRs. The hypotheses being asked in the chapter are as follows:

- Are there any detectable transcriptional between the parental lines within this data set?
- Are there any detectable differences in genomic methylation between the two parental lines within this data set?

By setting out the differences between the parental lines these can be used as a point of reference for changes that occur in the F1 hybrid and DH lines, whilst also providing a resource for others studying these lines.

3.3 Mapping of Parental Accessions to the Reference *B. oleracea* Genome

The samples in this study are being mapped to the same reference genome but they are evolutionary diverged species, it is therefore important to understand any bias in the mapping of the parental accessions to the reference genomes. The bisulphite sequencing reads mapped with a very similar efficiency in all lines. This ranged from 33 - 51 million paired end reads per sample mapped to the reference genome after removal of PCR duplicates (Table A.3, A.4). This is a minimum coverage of 15 X over the Parkin et al. (2014) reference genome. Non conversion rates were also very low for all libraries (<0.02%). Looking at the cytosines over all we find that 70% of all cytosines are covered by 10 or more reads.

The RNA sequencing reads also showed no bias between the parental lines in their mapping efficiency. However the mapping efficiency of the RNASeq was quite low and variable between samples (36% - 60%) (Table A.2, A.1). In the original library batch this meant that 6-12 million paired end reads were uniquely mapped in different samples and in the new batch there were between 12- 24 million mapped paired end reads. The reference contain 59,225 genes. In our data we managed to map to 35,941 genes for A12Dhd and 33,900 for GDDH33. When looking at all samples together, we found 51,616 genes had at least 1 uniquely mapped read.

3.4 Gene Expression Differences Between A12Dhd and GDDH33

To understand the transcriptomic changes resulting from hybridisation, it is important to understand the differences that exist between the parental lines. A compari-

Chapter 3. Transcriptome and Methylome Analysis of A12Dhd and GDDH33 Parental Lines

son between A12Dhd and GDDH33 revealed 3216 parental differentially expressed genes (pDEGs) this equates to 6.2% of considered genes (3216/51616). 1490 of these pDEGs have higher expression in GDDH33 and 1726 pDEGs have higher expression in A12Dhd (Figure 3.2). The two different groups of pDEGs have different enrichment profiles. The pDEGs with higher expression in A12Dhd show enrichment for gene involved in nucleotide binding (~200 genes) and primary metabolic process (540 genes). Whereas the GDDH33 higher pDEGs have a more varied enrichment profile. These genes are enriched for functions involved in the defence response, ribosomes, nucleosomes and cell wall.

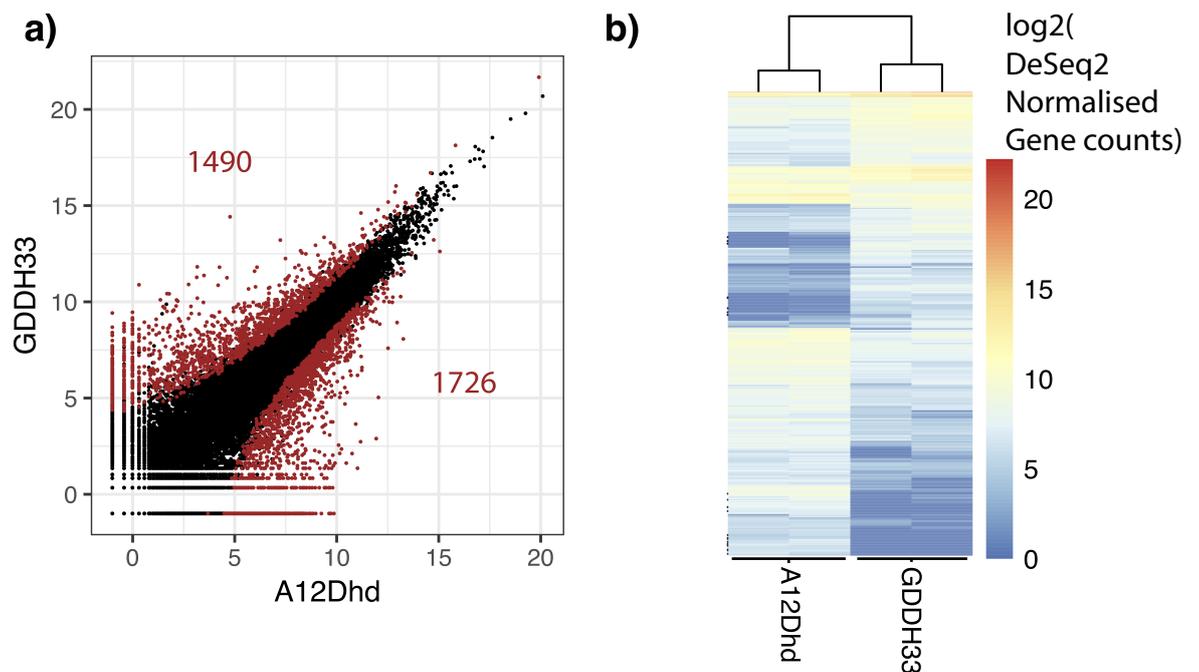


Figure 3.2: **Parental differentially expressed genes.** a) Scatter plot showing the average of each replicates expression of each gene for A12Dhd and GDDH33 with significant pDEGs appearing in red. b) Heatmap displaying expression of the 3216 parental DEGs with hierarchical clustering.

Table 3.2: GO analysis of 1726 genes with higher expression in A12Dhd compared to GDDH33. All terms achieved FDR <0.05.

GO-ID	GO Description	# List	# Genome	FDR
GO:0043531	ADP binding	43	571	1.02E-03
GO:0032559	adenyl ribonucleotide binding	208	4958	2.60E-02
GO:0030554	adenyl nucleotide binding	214	5216	3.70E-02
GO:0001883	purine nucleoside binding	214	5225	3.72E-02
GO:0001882	nucleoside binding	214	5252	4.04E-02
GO:0032553	ribonucleotide binding	226	5427	2.28E-02
GO:0032555	purine ribonucleotide binding	226	5427	2.28E-02
GO:00017076	purine nucleotide binding	232	5688	3.18E-02
GO:0000166	nucleotide binding	240	6029	4.83E-02
GO:0043170	macromolecule metabolic process	414	11088	4.83E-02
GO:0005515	protein binding	459	12019	1.61E-02
GO:0005737	cytoplasm	480	13003	4.04E-02
GO:0044238	primary metabolic process	540	14869	4.98E-02

Table 3.3: **GO analysis of 1490 genes with higher expression in GDDH33 compared to A12Dhd.** All terms achieved FDR <0.05.

GO-ID	GO Description	# List	# Genome	FDR
GO:0030529	ribonucleoprotein complex	79	1248	1.95E-08
GO:0045454	cell redox homeostasis	21	302	8.67E-03
GO:0005576	extracellular region	73	1467	4.52E-04
GO:0005618	cell wall	78	1451	1.62E-05
GO:0030312	external encapsulating structure	78	1453	1.65E-05
GO:0006091	gener. of precursor metab. and energy	41	574	1.33E-05
GO:0051707	response to other organism	79	1758	4.75E-03
GO:0050896	response to stimulus	366	9108	2.11E-09
GO:0032991	macromolecular complex	169	3601	6.36E-08
GO:0009719	response to endogenous stimulus	146	3279	1.76E-05
GO:0010033	response to organic substance	166	3902	4.00E-05
GO:0009725	response to hormone stimulus	135	3076	9.54E-05
GO:0009628	response to abiotic stimulus	156	3968	3.84E-03
GO:0019752	carboxylic acid metabolic process	78	1601	4.62E-04
GO:0005730	nucleolus	48	944	6.57E-03
GO:0015934	large ribosomal subunit	22	295	2.56E-03
GO:0022625	cytosolic large ribosomal subunit	21	263	1.56E-03
GO:0009058	biosynthetic process	245	6112	1.04E-05
GO:0046483	heterocycle metabolic process	44	772	1.16E-03
GO:0008152	metabolic process	632	18170	5.99E-07
GO:0044237	cellular metabolic process	498	14085	1.37E-05
GO:0009987	cellular process	667	19372	7.27E-07
GO:0044249	cellular biosynthetic process	235	5856	1.65E-05
GO:0008150	biological_process	941	28566	1.51E-08
GO:0044271	cellular nitrogen compound biosynth. Proc.	53	969	5.94E-04
GO:0044283	small molecule biosynthetic process	76	1616	1.71E-03
GO:0044281	small molecule metabolic process	152	3302	1.49E-06
GO:0005840	ribosome	68	954	2.82E-09
GO:0003735	structural constituent of ribosome	58	796	3.09E-08
GO:0005198	structural molecule activity	66	990	6.36E-08
GO:0006412	translation	74	1056	8.85E-10
GO:0033279	ribosomal subunit	35	476	4.26E-05
GO:0022626	cytosolic ribosome	47	563	1.96E-08
GO:0000287	magnesium ion binding	19	227	1.80E-03
GO:0006952	defense response	83	1680	1.56E-04

3.5 Methylation Analysis in the Parental Lines

3.5.1 Analysis of Single Cytosines

Bisulphite sequencing grants the ability to look at the single base resolution methylome of each of the parents: A12Dhd and GDDH33. Overall, they have similar global distributions of cytosine methylation in the 3 methylation contexts. CG has a bimodal distribution, CHG uniform and CHH sites show a left skewed distribution (Figure 3.3). This highlights the way in which CG methylated sites are faithfully copied to the new DNA during replication and result in homogeneous methylation throughout the studied tissue and that non-CG sites have a more mosaic pattern throughout the cells of the tissue (Figure 3.3).

This trend continues when looking at the methylation level accross genomic features. Methylation is highest over transposon features with genes accumulating the lowest amount of methylation in each context (Figure 3.4). At gene loci there is very little difference between the parental genotypes at non-CG sites and the methylation is very low. At transposable elements the opposite trend is seen, here CG sites show very little difference in their methylation between the genotypes. At CHG sites there is a small difference over TEs, but most striking here is the CHH methylation, where GDDH33 has higher methylation than A12Dhd over both RNA and DNA TEs.

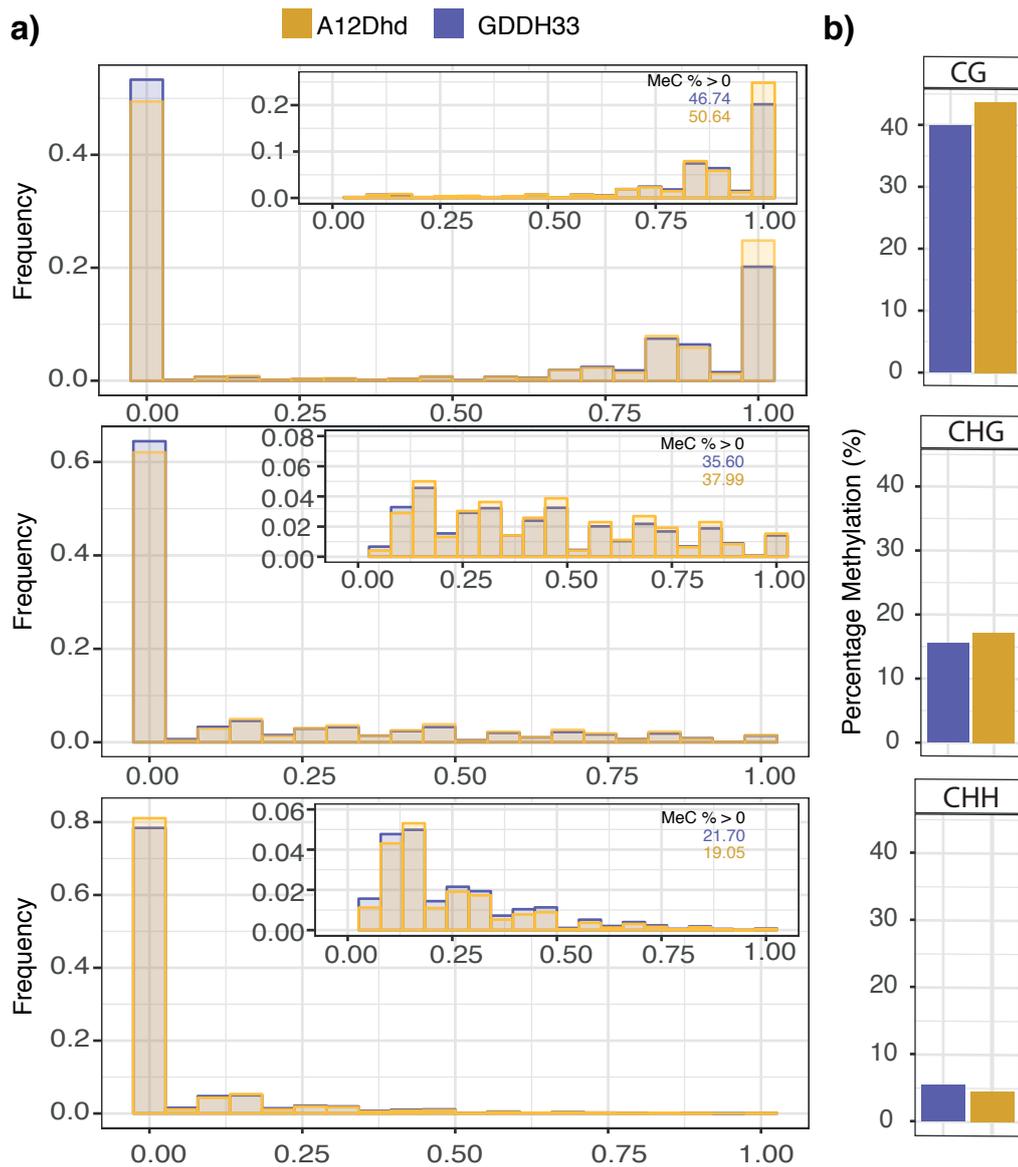


Figure 3.3: **Global cytosine methylation in A12Dhd and GDDH33 in each methylation context.** CG (top), CHG (middle) and CHH (bottom). a) Histogram displaying the proportion of sites exhibiting methylation ratios of 0-100% the panel in the corner of each plot displays a zoomed view of the distribution of sites with 1-100% methylation. b) The average methylation percentage of all cytosines.

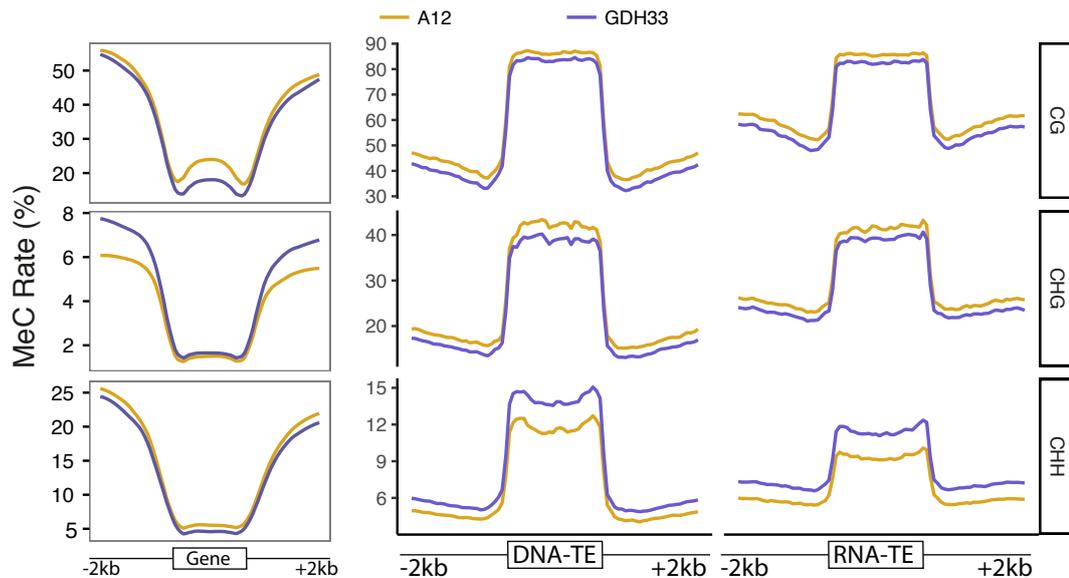


Figure 3.4: **Average methylation across genomic features for CG, CHG and CHH methylation.** Genes (Left), DNA and RNA transposable elements (Right). For each feature and context the 2 kb flanking regions for each feature are an average methylation value for a particular position for each feature in the genome. Across the feature body each feature is split into 100 bins and the methylation is averaged over these bins and then averaged across all features for these bins.

3.5.2 Differentially Methylated Regions between A12Dhd and GDDH33

As discussed in the methods, the three contexts of methylation were compared separately for DMRs between the two parents A12Dhd and GDDH33. The main differences occur at CG sites followed by CHH sites and then CHG. In total there were 23264, 8905 and 13009 parental DMRs (pDMRs) in CG, CHG and CHH context respectively. The location of the parental lines MRs and DMRs within genomic features agrees with the analysis of single cytosines in these lines. The MRs in each parent are located mainly in the transposons (54% - 67%) (Figure 3.5). In agreement with the single cytosine analysis, this is true for all sequence context MRs. The pDMRs also display the trend seen in the single cytosine analysis. CHH pDMRs have a similar

distribution to the CHH pMRs they reside mainly within the transposon sequences (61%). CG and CHG pDMRs do differ from the distribution of CG and CHH pMRs. CHG pDMRs are more prevalent in the intergenic regions and the genic regions. The CG pDMRs are just highly enriched for the genic regions showing that the difference in CG methylation between the parents exists mainly within the gene body.

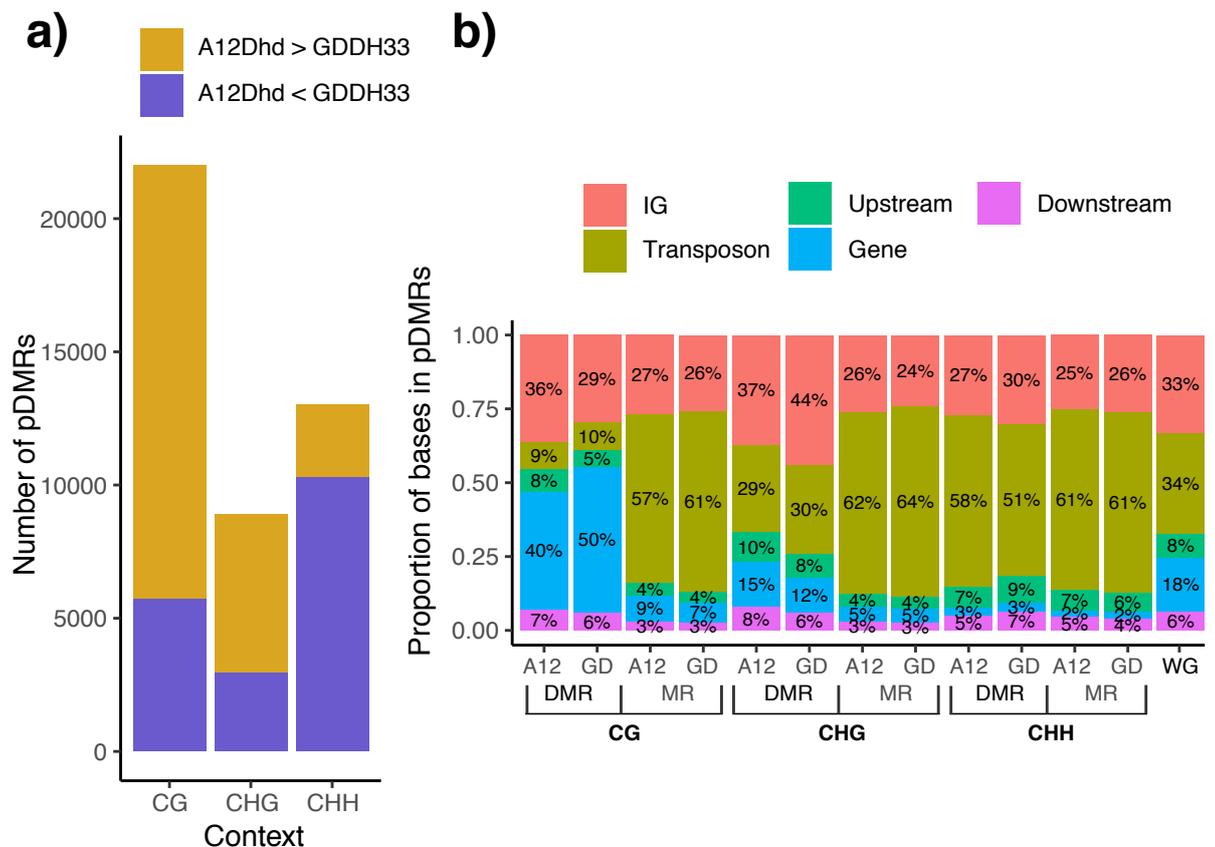


Figure 3.5: **Numbers and location of pDMRs and pMRs.** a) Barplot of the numbers of DMRs between A12Dhd and GDDH33 in each sequence context. DMRs with higher methylation in A12Dhd are shown in yellow and DMRs with higher methylation in GDDH33 are shown in blue. b) Location of these DMRs within genomic features, each base of a set of DMRs is assigned to the feature that it overlaps with. Then the results are displayed as a percentage of the total bases in that set. For each sequence context both A12Dhd MRs and GDDH33 MRs are shown. Then the DMRs between these two genotypes are split into DMRs with higher methylation in A12Dhd (A12) and DMRs with higher methylation in GDDH33 (GD). WG refers to the assignment of all the bases in the reference genome when assigned to a feature. This is done in a hierarchical fashion to account for overlapping features (gene, transposon, upstream, downstream, intergenic: in order of decreasing importance)

3.6 Discussion

The parental lines are the foundation for this breeding program, therefore to understand the genomic impacts of hybridisation and DH production in this program it is vital to have an understanding of the transcriptome and methylome of the parents. All samples in this experiment are aligned to the reference genome of T01000DH (Parkin et al., 2014). Because there is a different reference for either parental species, there is a possibility that the mapping of reads could be affected. However the mapping efficiencies of all lines in this study are consistent. In single cell experiments, as little as 50000 reads can be used to quantify genes with higher expression. But it has been suggested that up to 100 million reads should be used to quantify lowly expressed transcripts (Conesa et al., 2016). The range of library sizes is sufficient for quantification in both the RNASeq (Tarazona et al., 2011) and bisulphite sequencing (Ziller et al., 2015). Although some lowly expressed genes maybe ignored in the RNASeq analysis (Conesa et al., 2016). The genes we could not detect in any samples could be tissue specific genes that are not expressed in the leaf samples (Parkin et al., 2014).

Global methylation for these lines is in the same range as those previously published for *B. oleracea* (Parkin et al., 2014). *A. thaliana* shows lower overall methylation levels, this increased methylation in *B. oleracea* could have arisen from the increased silencing of the *Brassica* subgenome after genome duplication (Parkin et al., 2014). This could also reflect the increase of transposable elements within the *B. oleracea* genome when compared to *A. thaliana*. The distributions of methylation are unique to each sequence context, this has been demonstrated in many other species

and highlights the way in which methylation is propagated by different mechanisms in each sequence context (Feng et al., 2010). Because CG methylation is copied through DNA replication it manifests as a binary signal with cytosines generally being not methylated or fully methylated. In contrast CHG and CHH methylation have few cytosines with 100% methylation suggesting a more mosaic pattern of methylated cytosines within the tissue assayed. These non CG sites may have a more temporal and spacial regulatory influence. The majority of methylation resides within transposon sequences regardless of the sequence context, this is the case for many other studies and demonstrates methylations primary role in transposon silencing (Law and Jacobsen, 2011). The differences between the parents mainly appear as changes in CG and CHH methylation. The location of the different context means that gene body methylation is higher in A12Dhd and transposon methylation is highest in GDDH33. The two parents differ genetically by approximately 1 SNP every 1500bp, given that a DMR requires many changed cytosines over 100bp, it is unlikely these methylation changes are a result of DNA polymorphisms between the parents and represent real methylation changes.

These particular *Brassica* lines are popular because of their associated marker map. Having transcriptome and methylation sequencing available for these lines will be useful to other studies utilising these lines. Further to this, understanding the methylation and gene expression of the parental lines is vital to observing how these features change through F1 hybridisation and DH production.

Chapter 4

Transcriptome and Methylome

Analysis of the F1 Hybrids

4.1 Introduction

As highlighted in the main introduction, classical genetic models have only managed, in rare cases, to explain the phenotypic heterosis of F1 hybrids. There are also very few examples of single gene heterosis in plants (Jin et al., 2017). Further to this, heterosis is not the only benefit of F1 hybrids, as they are also used to combine parental traits. So, due to the commercial value of F1 hybrids and their prominent role in speciation, F1 hybrid transcriptomes and epigenomes have been the subject of more than 40 research articles and reviews within the last 10 years. With many of these studies of the transcriptome, methylome, sRNAome and chromatome producing consistent results even across evolutionarily distinct taxa. The terms transcriptome shock (Adams and Wendel, 2005) and epigenome shock (Rigal et al., 2016) have been coined to describe the outcomes of genome mergers. However, this is a very complex event which has

wide implications for the resulting F1 hybrid from the genome level to the gross phenotype. And so, over the last 15 years, many different transcriptomic and epigenetic signatures of F1 genome mergers have been described (Chen, 2013).

4.1.1 Transcriptomic Studies of F1 Hybrids and Polyploids

Transcriptomic studies in F1s of many species including rice, maize, *Arabidopsis* species, cotton, oil palm, and tomato have been conducted. They have shown consistent genome-wide changes in gene expression resulting from hybridisation and have given insight into regulatory networks and expression dynamics that contribute to heterotic F1 phenotypes (Yoo et al., 2013; Jin et al., 2017; Groszmann et al., 2015). Hypothetically, if all genes in a given genome are not regulated by any external factor then the combining of two haploid parental genomes in the same nucleus would result in additive gene expression for every gene. That is where the expression of both alleles in the F1 hybrid is closest to the mid-parent value MPV $(ExpP1 + ExpP2)/2$. However, it has been demonstrated from early studies using microarrays (Guo et al., 2006; Wang et al., 2005; Stupar et al., 2007) to more recent studies (Zhang et al., 2016; Greaves et al., 2016) that non-additive gene expression is prevalent in F1 hybrids. Non-additive expression patterns are any pattern that differs from the mid parent value. These patterns have been categorised differently in different studies but most agree on two major expression categories; the expression of the F1 hybrid is most similar to one parent (expression-level dominance) and the expression of the F1 is outside of the parental range (transgressive expression). To categorise the expression of genes in this way, the dominant-to-additive ratio (d/a) is commonly used. This ratio describes the expression of genes in the F1 according the parent with the highest or

lowest expression at that gene. Then a modified ratio described in Guo et al. (2006), the parental d/a (d/p'a). In this modified ratio, the expression of the hybrid is described according to the expression of the A12Dhd or GDDH33 parent (Chapter 2 - Methods). Using the two ratios together allows for the clustering of DEGs into 12 mutually exclusive categories possible from a hybrid and inbred parent cross, first exemplified by Yoo et al. (2013).

Firstly, the transcriptomes of hybrids have a very small proportion of genes with expression outside of the parental range of expression (Hu et al., 2016; Yoo et al., 2013; Groszmann et al., 2015). This was noted by Yoo et al. (2013) where they show that > 95% of the differentially expressed genes between cotton polyploids and their parents are already differentially expressed between the parents. Also, in *Arabidopsis* and other species, the proportion of transgressively expressed genes is generally low at around 5% (Lauss et al., 2018; Hu et al., 2016). These trends have now been shown to occur in multiple studies of F1s and seems to be a ubiquitous feature of hybridisation. This means that the transcriptional variation witnessed in the F1 is almost solely reliant on the transcriptional variation already existing between the parents. This could contribute to the fact that in many species, more distinct parental genotypes result in more additive and non-additively expressed genes in their resulting F1s (Jackson and Chen, 2010; Chen, 2013). Further to this, increased disruption in transcription has been associated with heterotic growth for both additively expressed genes (Guo et al., 2006) and non additively expressed genes (Greaves et al., 2016). Because of this, it has been shown that some interspecific F1 hybrids are more heterotic than the intraspecific F1 hybrids in *Arabidopsis* and maize. However this is not always the case and F1 hybrids between Col and Ler demonstrate this point, they are genetically

similar but have a heterotic phenotype (Groszmann et al., 2014).

Secondly, It has been reported in allopolyploids and F1 hybrids of maize (Swanson-Wagner et al., 2006), cotton (Yoo et al., 2013), senecio (Hegarty et al., 2008) and others (Chen, 2013) that both alleles of a gene in an F1 can assume the expression pattern of only one of the parental lines. At the genome-wide level, it is common to see more of the non-additively expressed genes being expressed at the level of one of the parents. This phenomenon was coined as genomic dominance by Rapp et al. (2009) but has now been termed expression-level dominance (ELD) to avoid confusion with other similar terms (Grover et al., 2012; Yoo et al., 2013). This phenomenon has been linked to the phenotype of F1s by Van Gioi et al. (2017) where they show that dominant expression patterns confer drought stress to the F1 hybrid. Another study in maize by Bi et al. (2014) show that expression-level dominance of a maize parent SRG200 with increased nitrogen usage efficiency also confers this phenotype to the offspring. Due to the complexity of ELD and the fact that it affects many genes, it has been attributed to many causes. In *Brassica rapa* it was shown that subgenome dominance could be influenced by the differing TE load of the parental genome, it was shown that lowly expressed homologues have a higher density of flanking TEs that are targeted by sRNAs (Cheng et al., 2016; Bottani et al., 2018) furthered these theories in allopolyploids by introducing cis-trans mis-regulation and genome size. By altering the regulatory landscape in the F1 the differing TFs can cause transcriptional change which can be fixed with epigenetic mechanisms. Further to this, they hypothesise that genome size will have an effect on TE efficiency with larger genomes requiring more specific binding of their TFs. They also suggest that similar parental genomes in any given crosses will produce similar expression level dominance outcomes in the

allopolyploid offspring.

And lastly, many of the transcriptomic studies have identified common enriched functions within the non-additively expressed genes that, in some cases, are responsible for heterosis. In *Arabidopsis*, Groszmann et al. (2015) have shown negative correlations between the expression of salicylic acid (SA) pathways and auxin pathways and that by reducing SA levels it is possible to mimic the heterotic phenotype in some individuals. Others have shown similar consequences of misregulation of the defence response pathway including Hegarty et al. (2008) and Chen (2013) which discuss many studies that report changes in defence response pathways. Along with defence, Fujimoto et al. (2012) also show changes in metabolomic genes and chloroplast related genes leading to increased photosynthetic capacity of the F1. Many of these studies refer to many genes but some studies have found single genes that confer heterosis. In oil palm, Jin et al. (2017) show that higher expression of WRINKLED1 in the F1 results in more oil production. The transcriptional landscape of one individual is complex, and so combining two genomes in a F1 compounds the issue. Different genes confer different traits in different studies but common dynamic of gene expression emerge, which, when understood, can be used to exploit the inheritance of complex traits and accelerate plant breeding.

4.1.2 Epigenetic Studies of F1 Hybrid

Much of the transcriptional change previously reported is a of a non-mendelian nature. Therefore, many studies have focussed on the epigenetic impact of genome mergers and it's role in transcriptional regulation and heterosis. Changes to sRNA levels and DNA methylation have been reported in many species (Greaves et al., 2012;

Zhang et al., 2016; Greaves et al., 2016) but chromatin (Greaves et al., 2015), miRNAs (Shivaprasad et al., 2012) and transposable elements (Cheng et al., 2016) are also implicated in the mis-regulation of genes in genome mergers. Direct inheritable epigenetic changes have been identified by Greaves et al. (2016) where they found multiple DMR associated genes with correlated expression and methylation that persist into the F2 generation. Rigal et al. (2016) also describe the IBM1 locus which is regulated by a combination of DNA methylation and an intronic DMR.

The importance of 21-24nt sRNAs and their roles in modulating gene expression is well documented. It is no surprise then that both 22nt microRNAs and 24nt siRNAs have been implicated in the misregulation of genes in F1 hybrids. Shivaprasad et al. (2012) show in tomatoes that mis-regulation of miRNAs that are responsive to the F1 genome merger can affect phenotypes relating to key agronomic traits. Additionally in *Arabidopsis* where microRNAs that control secondary metabolites responsible for the defence response, are mis-regulated upon genome combination (Ng et al., 2012). However it has not been observed in all F1 crosses as Li et al. (2012) observed no changes in miRNA levels in reciprocal hybrids of *Arabidopsis* species. The other types of sRNA, 21nt and 24nt siRNAs have also shown contribution to heterotic gene expression and phenotype. As discussed in the introduction, siRNAs through the actions of the RdDM pathway, regulate de-novo methylation in plants. Therefore, changes in sRNAs are inherently linked to changes seen in DNA methylation, in particular, changes that occur at CHG and CHH context cytosines. Most but not all studies across many taxa have reported reduction in sRNA levels in F1 hybrids (Barber et al., 2012; He et al., 2010; Li et al., 2012). This reduction in the levels of sRNA levels leads to implications for methylome of the resulting hybrid. Changes in methylation

in F1 hybrids can be described in terms of trans-chromosomal methylation (TCM) and trans-chromosomal de-methylation (TCdM). Greaves et al. (2016) and Kirkbride et al. (2015) have described the influence of sRNA on TCM / TCdM. They show that in the case of little sequence divergence, sRNAs produced by one parent can methylate both alleles and cause TCM. Opposingly, in the case of sequence divergence or in the case one sRNA allele fails to make the expression threshold for methylation, TCdM can occur (Zhang et al., 2016). It is unsurprising then that many studies find that CHG and CHH parent-hybrid DMRs in F1 hybrids are highly associated with sRNAs; Greaves et al. (2016) show that 80% of sRNAs are associated with DMRs. Zhang et al. (2016) also show that 80% of interacting DMRs are associated with sRNA locus and that by using polIV mutants, methylation interaction can be abolished. However, Zhang et al. (2016) also show that heterotic phenotypes are not abolished in these mutants with no methylation interactions. Rigal et al. (2016) coined the term epigenome shock and showed that in the extreme case of hybrids between Met-1 and Wild Type *Arabidopsis*. A reduction in CHH methylation causes activation of transposons and misregulation of genes and Greaves et al. (2016) also show transgenerational inheritance of correlated methylation and gene expression states. These findings suggest that methylation interactions via the RDdM pathway are not vital for heterotic phenotypes but can produce heritable gene expression changes and heritable phenotypic variation.

A big predictor of the magnitude of DNA methylation change is the difference in methylome between the two parents (Greaves et al., 2016). Much like with divergent gene expression, levels of additive and non-additive changes to methylation and siRNA expression in the F1 are enriched for changes that already exist between the parental lines (Groszmann et al., 2015). This is more severe in some studies and less

severe in others, such as Zhang et al. (2016) where they show that there are thousands of loci that experience transgressive methylation in the F1 hybrids of *Arabidopsis* at regions where the parents have similar methylation levels. Although, in that study all contexts of methylation were combined so it is possible that CG and non-CG methylation behave differently. It was also shown by Greaves et al. (2012) that the magnitude of difference in the parents for a particular DMR is positively correlated with the probability of non-additive change in an F1 hybrid. This supports the theory that more evolutionary divergent parental lines could produce more novel methylation states in an F1 hybrid. Another consistent predictive factor in methylation changes in F1 hybrids is methylation-level dominance. This phenomenon is linked thematically to expression level dominance but applies to DNA methylation. It has been observed in many species and results in genomic regions of the F1 hybrid preferentially assuming the methylation levels of one parent. Because many studies show a bias toward one parent, it can be seen that the F1s methylome is overall more similar to one parent over the other. Studies show that the majority of these changes are non sex-linked and are repeatable regardless of the direction of the cross (Rigal et al., 2016; Yoo et al., 2013).

4.2 Chapter Aims and Hypothesis

Given the previous chapters analysis of the parental lines, this chapters aim is to understand how the parental transcriptomes and methylomes are altered in the heterozygous F1 genotype. This aim will be addressed by the analysis of whole genome RNA-Seq by looking at the differential expression of genes and whole genome bisulphite sequencing, looking at both single cytosine positions and then DMRs, of the three genotypes in the half diallel cross (A12Dhd, GDDH33 and the F1). The hypotheses being asked in the chapter are as follows:

- In other studies of F1 hybrids, common signatures in the methylomes and transcriptomes of these hybrids have been identified. Do these same signatures exist in this F1 hybrid in *B. oleracea*?
- Do any genes differ in their expression in the F1 hybrid when compared to the parental lines?
- Do any regions differ in their methylation in the F1 hybrid when compared to the parental lines?

Describing how this *B. oleracea* hybrid responds in relation to current knowlege can provide more evidence for common responses in F1 hybrids which are a vital resource for plant breeders. Along with this, any changes identified here can be compared to changes that exists within the DH lines and can be used to identify differences and commonalities in these two different genome mergers. Finally this data can also provide a valuable resource for others using these lines.

4.3 Gene Expression in the Parent Hybrid Cross

4.3.1 Gene Expression Dynamics in the F1 Hybrid

Whole genome RNASeq data from the 3 genotypes (A12Dhd, GDDH33 and F1) was used for the analysis of the transcriptional dynamics in the F1 hybrid. Firstly, 3 pairwise comparisons for differentially expressed genes were made between the three genotypes in the half-diallel cross. There were 3216 DEGs found between A12Dhd and GDDH33 (pDEGs). From the other comparisons, 444 and 905 genes were found to be differentially expressed between the F1 and A12Dhd and GDDH33 respectively. This already indicates that the F1 is transcriptionally more similar to A12Dhd than GDDH33. Combining the DEGs from each of the three comparisons together, they total 3353 parent and hybrid DEGs (phDEGs). This reveals that only 137 phDEGs are significantly different in the hybrid but not differentially expressed between the two parental lines (Figure 4.1 a). The expression of these phDEGs in the F1 hybrid can be explained in terms of its relationship to expression in the parental lines. Two dominance-to-additive ratios were applied to these phDEGs (Figure 4.1 b). From this analysis, 66.6% (2234) of the phDEGs show the expected additive expression in the F1. The other 33.3% (1119) of genes have non-additive and unexpected expression patterns. The majority of the 1119 non-additively expressed phDEGs show expression level dominance (expression most similar to one of the parents) with only 282 phDEGS showing transgressive expression (expression outside of parental range). There is a large bias in the non-additively expressed phDEGs for A12Dhd expression level dominance (843/1119), this bias is independent of the direction of the difference in the parents and follows the expression of the A12Dhd parent whether it originally

had higher or lower expression than GDDH33 (Figure 4.1 b, c). This is corroborated by hierarchical clustering of the additive and non-additive phDEGs (Figure 4.2). In this analysis, for both non-additively and additively expressed genes a clade forms between A12Dhd and the F1, highlighting their more similar expression patterns.

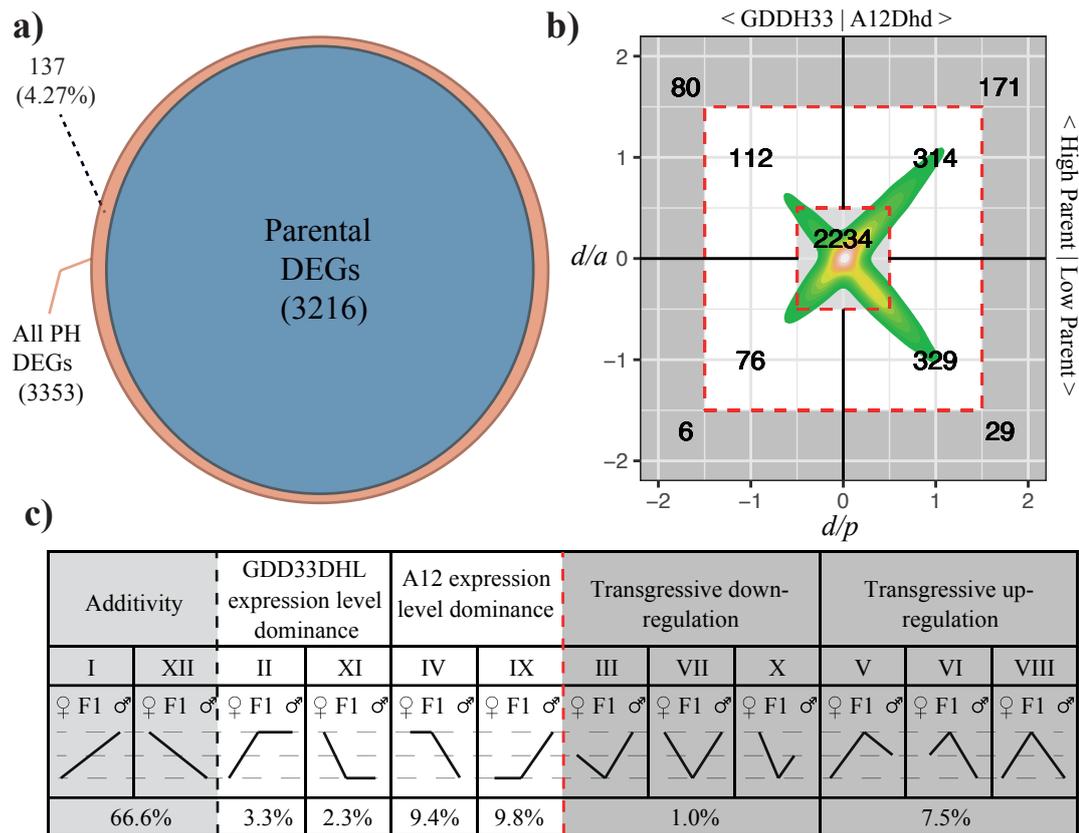


Figure 4.1: Gene expression dynamics in F1 hybrid. a) Venn diagram showing parental DEGs (blue) from the comparison between A12Dhd and GDDH33. Then the parental and hybrid DEGs (brown) from all 3 comparisons (A12Dhd - F1, GDDH33 - F1 and A12Dhd - GDDH33). This plot shows there is little novel differential expression in the F1 hybrid. b) Dominant-to-additive plot showing expression dynamic of phDEGs in F1 hybrid relative to the parental expression. Each phDEGs ratios are plotted, the d/a ratio on the y-axis and the parental d/a ratio is plotted on the x-axis. Plotting in this way, each phDEG can be categorised according to both the high / low parent and the maternal / paternal parent as shown by the numbers in the quadrants of the graph. c) Shows the categorisation of each gene. Roman numerals show the categories as they are commonly described (Yoo et al., 2013). Underneath the Roman numerals in the table, there is a graphic displaying the expression or methylation pattern of this category for the 3 genotypes (A12Dhd - maternal, GDDH33 - paternal and F1) then underneath that are the proportions of the phDEGs belonging to 12 mutually exclusive expression patterns.

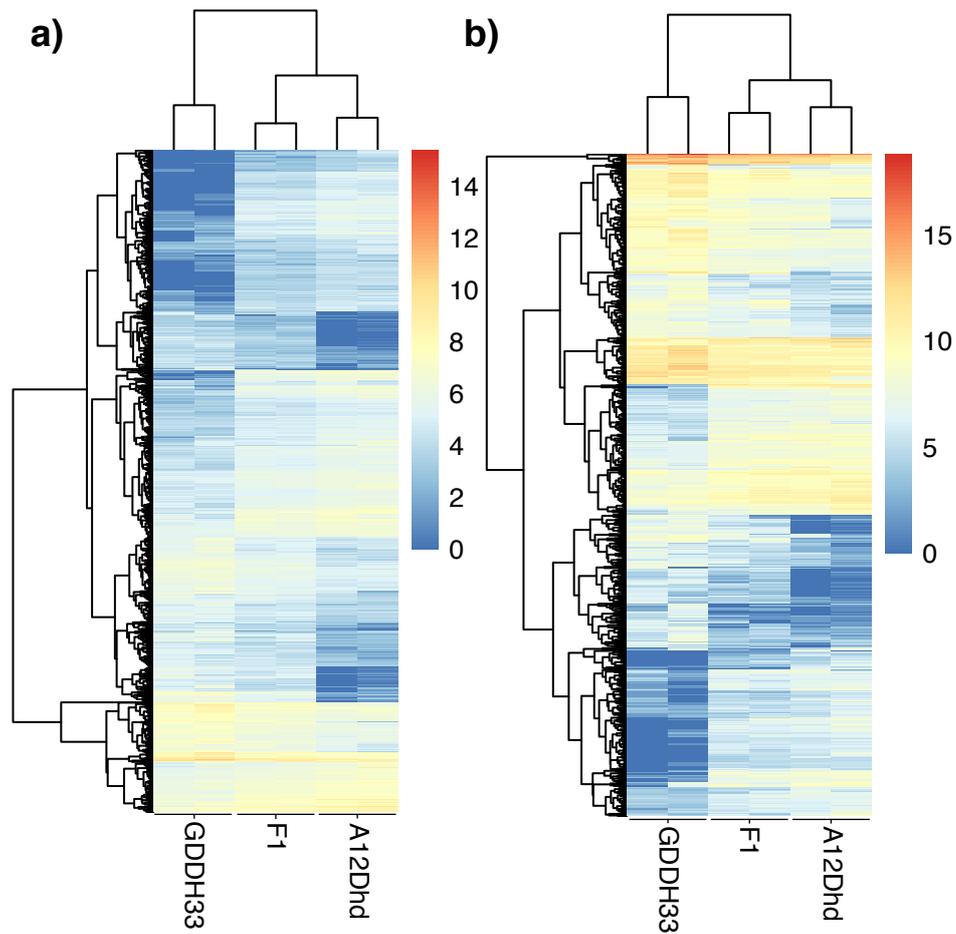


Figure 4.2: **Heatmaps of phDEGs. The F1 genotype forms a clade with A12Dhd for additively and non-additively expressed genes.** a) Heatmap of additive phDEGs. b) Heatmap of non-additively expressed phDEGs. Scale displayed is $\log_2(\text{DESeq2 normalized expression levels})$. Hierarchical clustering was performed and displayed as a dendrogram on the top of the heatmaps.

4.3.2 Differentially Expressed Genes

GO enrichment analysis was performed on the different sets of DEGs identified in the F1 hybrid. Differentially expressed genes in the F1 hybrid that were not differentially expressed between the parents (137 genes) are not enriched for any particular function. The same is true for the 282 differentially expressed genes that are outside of the parental range in the F1 hybrid. This suggests there is no discrimination for these types of genes and they represent a similar proportion of functions to that of the whole genome. However, the 2234 additively expressed genes and the 1119 non-additively expressed genes both have enriched gene ontology terms. In the additively expressed genes the main terms are involved in nucleotide binding and ADP binding (Table 4.1). These terms also appear in the enrichment of the pDEGs. The genes displaying expression-level dominance in the F1 have a different enrichment profile, they show terms relating to the chloroplast and anatomical structure development (Table 4.2). Although many of these terms are also enriched within the pDEGs it shows that in this cross some gene functions are more likely to be additively expressed and others are more likely to be non-additively expressed.

Table 4.1: **GO analysis of 2234 additively expressed F1 genes.** All terms achieved $FDR < 0.05$

GO-ID	GO Description	# List	# Genome	FDR
GO:0071211	protein targeting vacuole involv. autophagy	3	3	7.58E-03
GO:0009626	plant-type hypersensitive response	13	102	2.51E-02
GO:0034050	host program cell death induced by symbiont	13	103	2.71E-02
GO:0042254	ribosome biogenesis	30	356	2.35E-02
GO:0022613	ribonucleoprotein complex biogenesis	35	441	2.35E-02
GO:0043531	ADP binding	51	571	1.01E-04
GO:0044428	nuclear part	106	1787	2.35E-02
GO:0016491	oxidoreductase activity	157	2779	1.26E-02
GO:0032991	macromolecular complex	207	3601	5.50E-04
GO:0032559	adenyl ribonucleotide binding	270	4958	9.61E-04
GO:0030554	adenyl nucleotide binding	280	5216	1.70E-03
GO:0001883	purine nucleoside binding	281	5225	1.47E-03
GO:0001882	nucleoside binding	281	5252	2.02E-03
GO:0032553	ribonucleotide binding	294	5427	5.71E-04
GO:0032555	purine ribonucleotide binding	294	5427	5.71E-04
GO:0044267	cellular protein metabolic process	296	5696	8.64E-03
GO:0017076	purine nucleotide binding	304	5688	1.02E-03
GO:0000166	nucleotide binding	326	6029	2.17E-04
GO:0019538	protein metabolic process	341	6802	2.49E-02
GO:0044446	intracellular organelle part	384	6835	3.46E-07

Table 4.2: **GO analysis of 1119 non additively expressed F1 genes.** All terms achieved FDR < 0.05

GO-ID	GO Description	# List	# Genome	FDR
GO:0010177	methylthioalkylmalate synth.	2	3	2.69E-02
GO:0016138	glycoside biosynthetic process	8	106	1.60E-02
GO:0044272	sulfur compound biosynth. process	12	208	9.41E-03
GO:0009505	plant-type cell wall	23	637	1.62E-02
GO:0005198	structural molecule activity	31	990	1.88E-02
GO:0009941	chloroplast envelope	34	1116	1.71E-02
GO:0009526	plastid envelope	37	1161	6.21E-03
GO:0005576	extracellular region	40	1467	3.36E-02
GO:0009416	response to light stimulus	44	1387	2.16E-03
GO:0031967	organelle envelope	46	1705	2.32E-02
GO:0019752	carboxylic acid metabolic proc.	49	1601	1.96E-03
GO:0042180	cellular ketone metabolic proc.	50	1644	1.96E-03
GO:0009653	anatomical structure morphogenesis	51	1748	3.50E-03
GO:0044434	chloroplast part	62	2156	1.40E-03
GO:0044435	plastid part	65	2200	5.66E-04
GO:0005829	cytosol	84	3553	1.53E-02
GO:0048856	anatomical structure development	96	4311	2.80E-02
GO:0032502	developmental process	113	5227	2.80E-02
GO:0009507	chloroplast	116	4321	3.06E-05
GO:0009536	plastid	125	4550	3.07E-06
GO:0044249	cellular biosynthetic process	125	5856	2.55E-02
GO:0009058	biosynthetic process	129	6112	2.80E-02
GO:0044422	organelle part	144	6838	1.88E-02
GO:0044444	cytoplasmic part	246	12031	1.45E-03
GO:0005737	cytoplasm	262	13003	1.81E-03
GO:0044237	cellular metabolic process	278	14085	2.99E-03
GO:0043231	intracell membr. organelle	280	14803	2.66E-02
GO:0043227	membr.-bounded organelle	280	14808	2.68E-02
GO:0043229	intracellular organelle	297	15521	9.94E-03
GO:0043226	organelle	297	15535	1.04E-02
GO:0044424	intracellular part	330	17443	8.28E-03
GO:0005622	intracellular	335	17747	8.28E-03
GO:0008152	metabolic process	347	18170	2.62E-03
GO:0009987	cellular process	357	19372	2.07E-02
GO:0044464	cell part	470	26187	9.01E-03
GO:0005623	cell	472	26293	8.30E-03
GO:0005575	cellular_component	479	26720	8.04E-03

4.4 Methylation in the Parent Hybrid Cross

4.4.1 Analysis of Single Cytosines

Bisulphite sequencing grants the ability to look at the single base resolution methylome of each of the 3 genotypes; A12Dhd, GDDH33 and the F1 hybrid. Overall, the three genotypes have similar global distributions of cytosine methylation in the 3 methylation contexts, CG has a bimodal distribution, CHG uniform and CHH sites have a left skewed distribution. This highlights the way in which CG methylated sites are faithfully copied to the new DNA during replication and result in homogeneous methylation throughout the studied tissue and that non-CG sites have a more mosaic pattern throughout the cells of the tissue. However, at this genome-wide level the F1 shows a mid-parent value (MPV) for CG and CHG methylation while experiencing lower methylation levels than both parents at CHH sites (Figure 4.3).

This trend continues when looking at the methylation level across genomic features. Methylation is highest in each context over transposon features with genes accumulating the lowest amount of methylation in each context. At genes there is very little difference between the parental genotypes at non-CG sites and the methylation is very low. At genic CG sites there is a larger difference between the parental lines with the F1 showing a MPV. At transposable elements the opposite trend is seen, here CG sites show very little difference in their methylation between the genotypes. At CHG sites there is a small difference but the F1 takes a MPV. But most striking here is the CHH methylation over transposon features. The difference observed in the whole genome global distributions can be seen mainly to reside within the transposon portion of the genome, where the F1 is transgressively demethylated over both DNA

and RNA transposons (Figure 4.4).

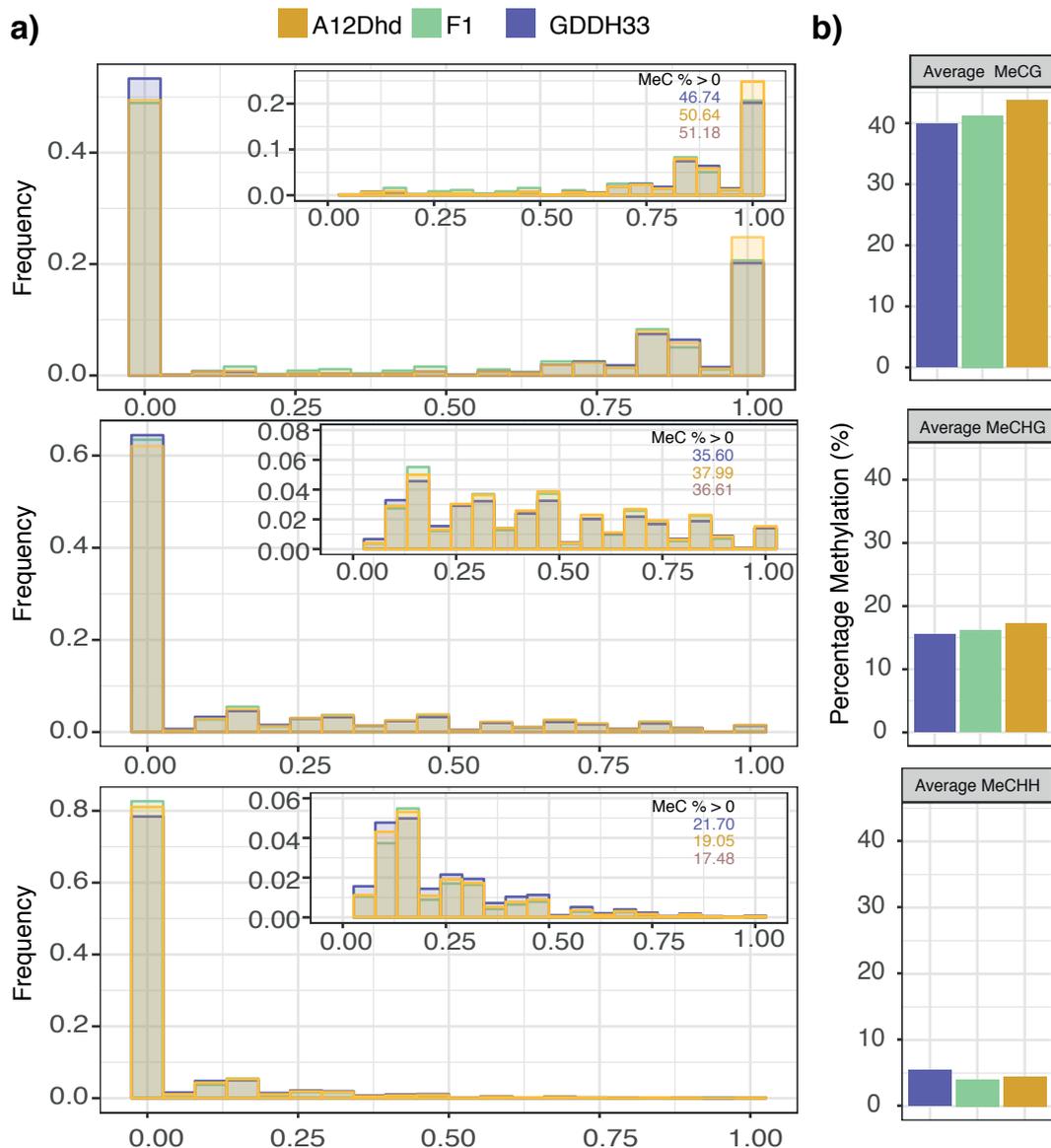


Figure 4.3: **Global cytosine methylation in A12Dhd, GDDH33 and F1 in each methylation context.** CG (top), CHG (middle) and CHH (bottom). a) Histogram displaying the proportion of sites exhibiting methylation ratios of 0-100% the panel in the corner of each plot displays a zoomed view of the distribution of sites with 1-100% methylation. b) The average methylation percentage of all cytosines.

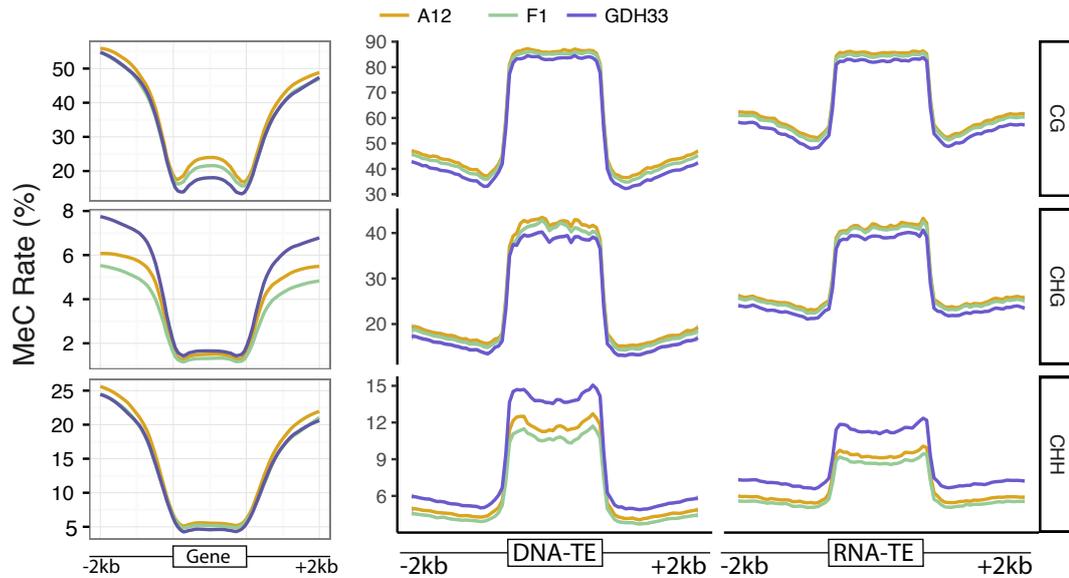


Figure 4.4: Methylation average across genomic features for CG, CHG and CHH methylation. Genes (Left), DNA and RNA transposable elements (Right). For each feature and context the 2 kb flanking regions for each feature are an average methylation value for a particular position for each feature in the genome. Across the feature body each feature is split into 100 bins and the methylation is averaged over these bins and then averaged across all features for these bins

4.4.2 DMR Dynamics in the F1 Hybrid

The change seen in the single cytosine analysis are also present when analysing DMRs between the 3 genotypes. Combining the DMRs in each context from all three comparisons in the half diallel cross; 23264 (CG), 12624 (CHG) and 22050 (CHH) DMRs were identified that differ significantly in at least one of the three comparisons made between the parents or hybrid (phDMRs). At CG context phDMRs there are very few novel changes, this is exemplified by most CG phDMRs already existing as pDMRs (Figure 4.5). This shows a reliance on epigenetic distance in the parents for CG methylation changes in the hybrid, a similar trend to that described with the phDEGs. In contrast, CHG and CHH context phDMRs display more novel changes in the F1 hybrid (CHG - 3719-29%, CHH - 9041-41%) (Figure 4.5). These are phDMRs that differ in the F1 but have similar methylation in the parent lines.

The methylation of the phDMRs differ significantly in at least one comparison between the 3 genotypes (A12Dhd, GDDH33 and F1) but to understand the methylation interactions occurring in the F1 we project the findings in the same manner as the phDEGs, by using the dominant-to-additive ratio and the parental-dominant-to-additive ratios and further categorising the DMRs into 12 mutually exclusive types of methylation patterns. Figure 4.6 displays the ratios graphically, along with the different expression categories. At CG phDMRs, additivity is the major category (63.5%, 14786 / 23264). At non symmetrical CHG and CHH phDMRs the proportion of additivity is only 37.3% and 20.3% respectively. Non-additive methylation is also widespread in the F1 hybrid. In each context there is a bias for non-

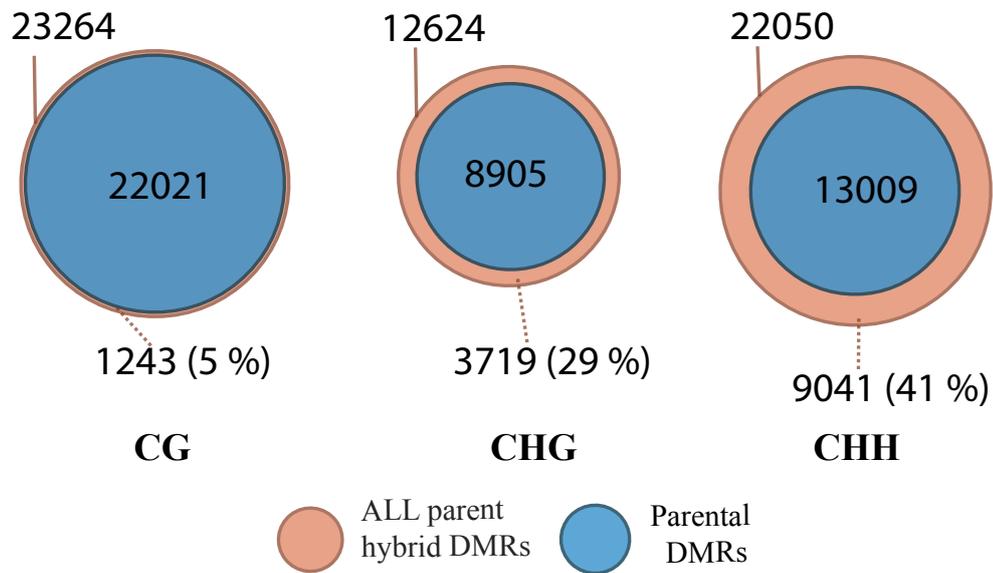


Figure 4.5: **Venn diagrams showing overlap between pDMRs (blue) and phDMRs (brown). The F1 has little novel methylation at CG sites but more novel methylation at CHG and CHH sites.** For each context separately the pDMRs are obtained from the comparison between A12Dhd and GDDH33 and the phDMRs are obtained from the three comparisons (A12Dhd - GDDH33, A12Dhd - F1 and GDDH33 - F1).

additive phDMRs to follow the methylation of the A12Dhd parent (CG - (6144/7649, 80%), CHG (2642/4360, 60%), CHH - (5895/8281, 71.1%)). This follows the original methylation of the parents, where at CG phDMRs there are more transchromosomal methylation events (TCM) bringing the methylation to the level of A12Dhd but at CHH sites there are more trans-chromosomal de-methylation events (TCdM). Further to this and in agreement with the phDEGs, hierarchical clustering of the phDMRs (Figure 4.7) shows A12Dhd forming a clade with the F1 hybrid and GDDH33 clustering as an outgroup. The same trend is seen for additive phDMRs (Figure 4.7 a) and for non-additive phDMRs (Figure 4.7 b).

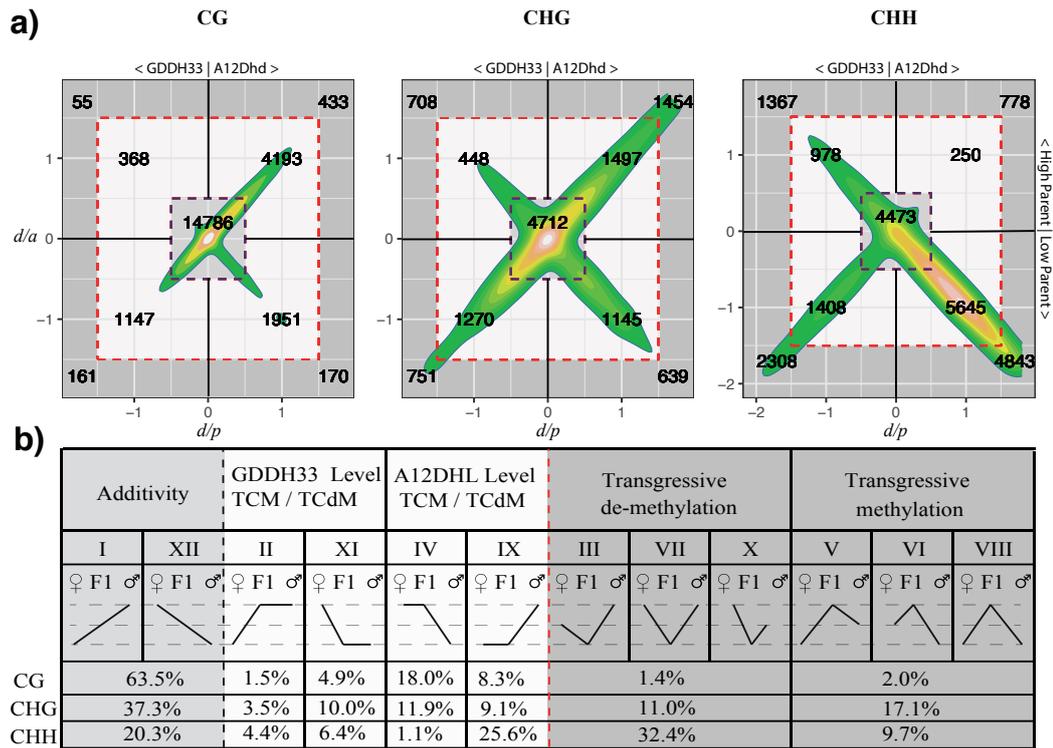


Figure 4.6: **Methylation dynamics for CG, CHG and CHH context phDMRs.** a) Dominant-to-additive plots showing methylation dynamics of phDMRs in F1 hybrid relative to the parental methylation. Each phDMRs ratios are plotted, the d/a ratio on the y-axis and the parental d/a ratio is plotted on the x-axis. Plotting in this way, each phDMR can be categorised according to both the high / low parent and the maternal / paternal parent as shown by the numbers in the quadrants of the graph. b) Shows the categorisation of each phDMR. Roman numerals show the categories as they are commonly described (Yoo et al., 2013). Underneath the Roman numerals in the table, there is a graphic displaying the expression or methylation pattern of this category for the 3 genotypes (A12Dhd - maternal, GDDH33 - paternal and F1) then underneath that are the proportions of the phDMRs belonging to 12 mutually exclusive expression patterns in each sequence context.

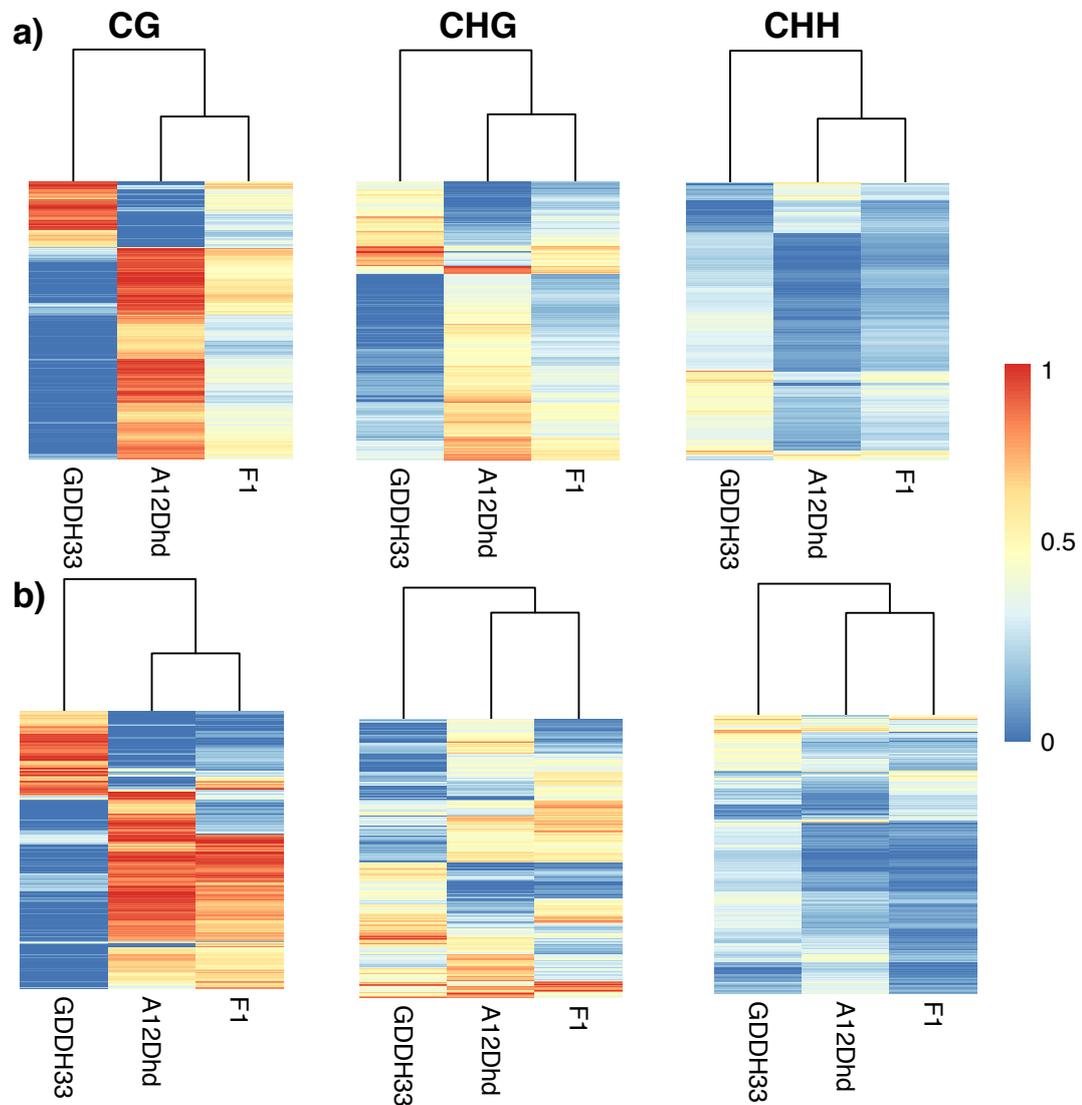


Figure 4.7: **Heatmaps of phDMRs. Methylation of the F1 is most similar to A12Dhd for both additive and non-additive phDMRs.** a) Additive phDMRs. b) Non-additive phDMRs. For each sequence context; CG, CHG and CHH. Scale shows the methylation rate of the DMR.

Even considering the large proportion of A12Dhd dominant hypomethylation at CHH phDMRs (71%) we found that the F1 undergoes widespread transgressive hypomethylation when compared to both parents (Figure 4.6). This is seen in the global single cytosine analysis where there is an increase of CHH sites without methylation, a lower average methylation and lower methylation throughout the non-genic areas of the genome (Figure 4.3). In the phDMRs we see this trend in more detail with 7151 of the 22050 CHH phDMRs experiencing transgressive de-methylation.

4.4.3 Location of Methylation and Differential Methylation

The MRs of the parents and hybrid in each sequence context show a very similar distribution over genomic features. These MRs show where the methylation resides within the genomes. We find that in each context approximately 60% of the MRs exist in transposable elements. The TE annotation occupies 34% of the genome showing that there is an enrichment of methylation in the TE fraction of the genomes (Figure 4.8). This can be also seen from the analysis of single cytosines in Chapter 3 (Figure 4.4) where methylation in all sequence contexts is higher in the pericentromeric regions and highest over TE annotated regions. However, when looking at the phDMRs there are stark differences between CG, CHG and CHH phDMRs. CHH phDMRs reside mainly within the transposon sequences (60%). This distribution does not deviate from where the methylation usually resides. So it can be said that CHH methylation is transgressively demethylated in this F1 hybrid but the methylation is indiscriminate and happens across all CHH sites but these CHH sites generally reside within transposon features. As for CG sites we find stark differences between the normal locations of MRs and the locations of DMRs with DMRs being preferentially associated with

the genic regions.

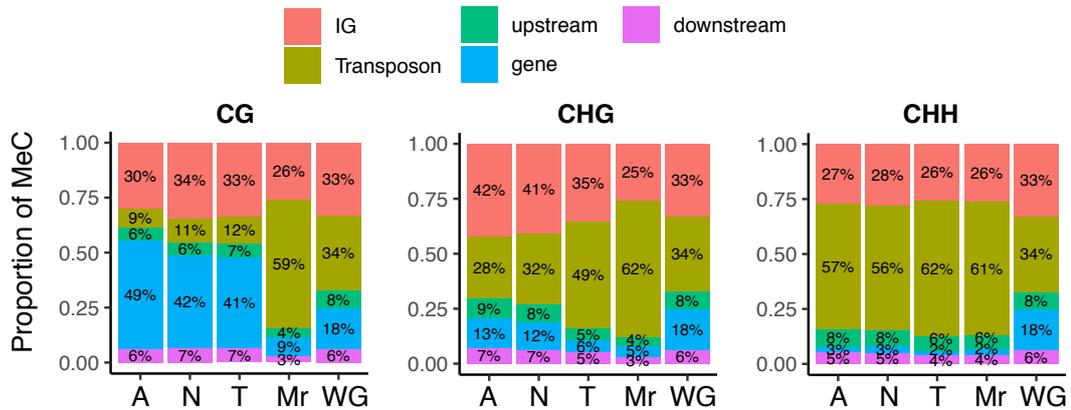


Figure 4.8: **Location of phMRs and phDMRs at CG, CHG and CHH context within genomic features.** A - Additive phDMRs, N - Non-additive phDMRs, T - Transgressive phDMRs. Mr - MRs from relevant context. WG - The reference genome if all bases were assigned to a feature. All bases of a particular feature set are assigned to a genomic feature and then plotted as a percentage of the total number of bases in that feature set. This is done in a hierarchical fashion to account for overlapping features (gene, transposon, upstream, downstream, intergenic: in order of decreasing importance).

From the single cytosine analysis, all transposon types are demethylated at CHH loci when compared to both parents (Figure 4.4). A similar trend was observed in the CHH phDMRs. These DMRs that intersect with transposons can be viewed in terms of the transposons that they overlap with. In each context, additive, non additive and transgressive phDMRs behave similarly. Both CG and CHG phDMRs have a similar transposon profile, occupied by half DNA half RNA transposons. But, CHH phDMRs show a slight increase for DNA transposons. Considering these are hypomethylated DMRs this could indicate a preference for hypomethylation DNA TEs (Figure 4.9).

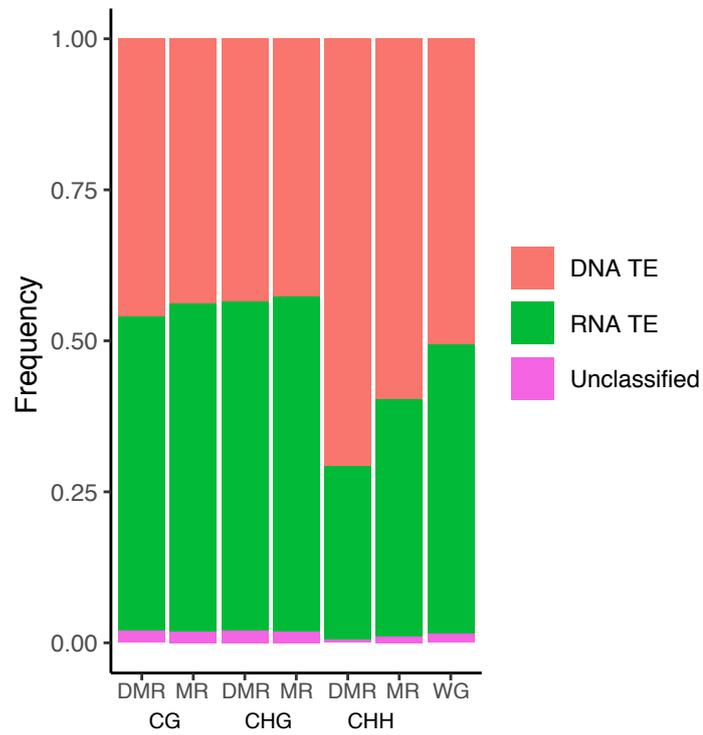


Figure 4.9: **Locations of phDMRs and phMRs in different transposon types.** In each case it shows the proportion of bases in each feature type that occupies the different types of transposons.

4.5 Discussion

Many efforts have been made to categorise and understand the consequences of F1 genome mergers on the transcriptome and epigenome. As discussed in this chapters introduction, many signatures have been described that are common between crosses of different species (Chen, 2013). In a similar vein, in *B. oleracea* we found many of the same signatures in the transcriptome and methylome of the F1. By analysing the global dynamics of phDMRs and phDEGs in the same way, we show that there are many similarities between the global dynamics of the phDEGs and CG phDMRs. Whereas the CHG and CHH phDMRs display different dynamics. This extends to: the proportion of additive features over non-additive features, the lack of non-parental patterns and presence of expression-level and methylation-level dominance. Most of the phDEGs and CG phDMRs display the expected additive expression and methylation patterns. This trend has been shown in both F1 hybrids and polyploids of cotton, along with F1s of *Arabidopsis* species and maize (Stupar et al., 2008; Lauss et al., 2018; Yoo et al., 2013). Greaves et al. (2016) show that the majority of this additive methylation results from the F1 retaining the methylation state of the parents on both alleles. Non-CG methylation however displays less additive patterns, this exemplifies the different mechanisms of methylation maintenance in plants (Law and Jacobsen, 2011). In most species, cis regulatory elements dominate gene expression regulation and the majority of CG methylation is copied faithfully along with these *cis* regulatory elements (Zhang et al., 2018; Chen and Rajewsky, 2007). Therefore, additive F1 patterns are expected for these features, CHG and CHH methylation on the other hand are mainly regulated in *trans* (Zhang et al., 2018), because of this reliance on regulatory elements from both

alleles, these features experience more non additive patterns. This default for additive expression and CG methylation in the F1 is linked to the finding that changes in CG methylation and gene expression are governed by the differences already existing in the parental lines. This shows there is not much ability for a gene to have transgressive expression patterns. Over evolutionary time if two genes or methylated regions have not diverged in their expression or methylation it is reasonable to assume that whatever regulated these genes or regions has also not changed (Chen and Rajewsky, 2007). This means that only where regulatory elements have diverged between the parental species can gene expression in this F1 be affected. However, we see that non-CG methylation shows more novel patterns. These are mainly controlled by the de-novo methylation pathway and show that genes and methylated regions regulated in *trans* are more affected by heterozygosity. In which cases the amount of parental divergence can lead to increased heterosis remains unclear (Zhang et al., 2016), but being able to predict the amount of additive and non-additive gene expression can be very useful in hybrid breeding.

The F1 hybrid also displays unbalanced expression-level dominance and methylation-level dominance. At the genome-wide level, this results in the F1 hybrids transcriptome and methylome becoming more similar to A12Dhd than the GDDH33 parent. Dominance like this was initially identified in the transcriptome by Rapp et al. (2009) but more recently studied by Van Gioi et al. (2017) and Bottani et al. (2018). However, it has not been described in the same way for methylation. In this study, the extent of methylation-level dominance is mainly confined to CG sites. CHG phDMRs do show methylation-level dominance but to a lesser degree. These expression patterns could be caused by genome size, TE density in proximity to genes and mis-

matches of trans-effectors (Bottani et al., 2018). In this case, the genomes of the two parents are of identical size and so the expression level dominance is likely due to the TE density and / or the mis-matching of trans regulators. Expression and methylation level dominance is an important phenomenon for genome evolution and plant speciation. Bi et al. (2014) showed that resistance to nitrogen limitation can be conferred to hybrid offspring through the action of expression-level dominance. Van Gioi et al. (2017) also show that dominance in drought stress genes can directly confer phenotype to drought stress through this dominance in the F1. They also show that similar genotype crosses result in similar dominance expression patterns and drought stress phenotypes in the F1s, this was also postulated by Bottani et al. (2018). This means that in some cases, ELD of certain functions of genes can confer phenotypes to F1 hybrids and in these cases it relies on the specific parental combination to ensure dominance of these genes. CHH phDMRs do show methylation-level dominance along with the other sequence contexts. But they also undergo wide spread transgressive demethylation. Because A12Dhd has a lower methylation status than GDDH33 at CHH sites, this transgressive demethylation could be construed as MLD. But the reduction in CHH methylation has also been shown in many studies and is likely caused by the effect of hybridisation on sRNAs (Greaves et al., 2012). Some studies do show increased methylation such as crosses of certain *Arabidopsis* ecotypes (Shen et al., 2012). Suggesting that in the cases of little sequence divergence, sRNAs from both parents are able to methylate each parental genome in the F1 and lead to increased methylation. Conversely, in cases such as this one, where there is sequence divergence; sRNAs may fail to methylate the target on the other parental genome and cause demethylation.

Our data supports the view that the divergence between the parents is a major driving force for changes in the F1 hybrid. In a wider cross, the regulatory landscape between the two parental species will be more divergent. This becomes the source of additive and non-additive expression in F1 hybrids. This increase in transgressive expression pattern can influence phenotype and drive cultivar development (Greaves et al., 2016). Further to this, the specific parental combination then affects the dominance in gene expression and probably CG methylation. In some cases, this dominance can confer desirable phenotypes to the F1 (Van Gioi et al., 2017; Bi et al., 2014). Already, breeders have advanced phenotype data for many crosses but understanding the causes of these inherited traits would allow better predictive power in breeding programs.

Chapter 5

Transcriptome and Methylome

Analysis of the Doubled Haploid Lines

5.1 Introduction

Doubled Haploid (DH) lines have been a valuable tool in the toolbox of plant breeders since their discovery in the 1980s. Doubled haploids allow the creation of homozygous lines in one generation leading to increased genetic gain for many plant breeding programs. (Ferrie and Möllers, 2011). They are also vital for MAB technologies and because of their haploid stage are now becoming popular for the production of transgenic plants, this is because a mutation made at the haploid stage before chromosome doubled will be fixed in homozygosis (Ferrie and Möllers, 2011). As well as these benefits, the genomes of DH lines provide an interesting extension to the studies of genome mergers in F1 hybrids and polyploids. An F1 hybrid is heterozygous at many loci, containing a haploid chromosome set from both parents. A tetraploid hybrid contains a full diploid set from both parents. This leads to many implications for the

genome of the resulting hybrid which are set out in Chapter two. DH lines are another form of these mergers, they are diploid and homozygous but have sections of the genome that belong to both of the parents (Figure 1.1). F1s and polyploids have been the subject of many research articles to understand the genomic consequences of merging genomes but DH lines have received little attention.

Transcriptomic studies of DH lines are very limited, many of them are concerned with the embryogenesis itself and extract RNA from the micropores or the early embryo in the aid of identifying genes responsible for the embryonic induction (Seifert et al., 2016). Transcriptomic studies of adult DHs have only focussed on the difference between two DH lines, only trying to understand the gene expression differences underlying differing phenotypes (Abdelrahman et al., 2015; Jung et al., 2014). Likewise epigenetic studies have looked at methylation mutations during microspore culture and causes for gametoclonal variation and off-type plants (Solís et al., 2015; Machczyńska et al., 2014). However these studies do not draw direct comparisons to the parental genomes and because of this, miss out on some information about the genomic consequences of this genome merger.

Using DH lines in breeding programs can be divided into two distinct steps; DH production and the selection of advantageous lines. The DH production step has received much attention in the effort to increase the efficiencies of DH production (Ferrie and Möllers, 2011). The process can take 3 months in some species to 1 year in others and is often very inefficient. To this end, studies have dissected each stage of the protocols for DH production looking for efficiencies (Ferrie and Caswell, 2011). Even still, DH production is heavily dependant on the genotype of the donor plant

and often protocols are not translatable between different species or even cultivars. DH production is also a very stressful process. Long periods of culture can introduce genomic changes and often, DH lines are inferior to lines produced via traditional in-breeding methods. This can be because of somaclonal variation from changes arising during in vitro culture as well as abnormalities from induced chromosome doubling. It is therefore surprising that little is known about lasting transcriptomic and epigenetic changes that are introduced in this step and whether it can be harnessed for plant breeding. The second step of selecting lines is arguably more resource demanding. Often requiring growth space for hundreds or thousands of lines. The selection methods have improved if the breeder is using a marker assisted approach as genotyping at a young age removes many undesirable lines before resources are wasted. But little other research has been done into predicting which lines will be more advantageous before phenotypic selection.

5.2 Chapter Aims and Hypothesis

The aims of this chapter are to define the impact on the transcriptome and methylome of merging the parental genomes in the DH lines. This aim will be addressed by the analysis of whole genome RNA-Seq by looking at the differential expression of genes and whole genome bisulphite sequencing looking at both single cytosine positions and then DMRs. This chapter will be using comparisons between the data from the 9 DH lines in this study and the parental lines. The hypotheses being asked in the chapter are as follows:

- Studies of genome mergers in F1 diploid hybrids and polyploids have been conducted and common signatures have been identified. Do any similar or different signatures in the transcriptome and methylome exist in the genomes of the DH lines in this study?
- Do any genes differ in their expression in the DH lines when compared to the parental lines?
- Do any regions differ in their methylation in the DH lines when compared to the parental lines?
- In some studies methylation has been shown to have direct correlations with gene expression. Do any genes show correlation with methylated regions in this study?

By answering these hypotheses it will be possible to extend current knowledge of genome mergers in F1 hybrids and polyploids. Also this information could be useful

to plant breeders and improve plant breeding programs. Finally this data can also provide a valuable resource for others using these lines.

5.3 Genotyping Doubled Haploid Lines - Understanding the Parental Genome Contribution

The doubled haploid lines are homozygous but are a mosaic of both parental genomes caused by recombination of the parental genomes during meiosis in the F1. As shown in Chapter 2 the DH parents do differ in their transcriptome and methylome. Because of this, whole genome comparisons between the DH lines and either parent would result in comparisons between parts of the DH genomes being compared to the other parental genotype. In the first instance, whole genome comparisons were performed between the parents and the doubled haploid lines for both gene expression and methylation in the 3 contexts. When this whole genome comparison is performed, the magnitude of CG DMRs and DEGs is highly correlated with the predicted genome structure. Regions inherited from A12Dhd have many more DMRs when compared to GDDH33 and the opposite is true for regions inherited from GDDH33. CHG DMRs do follow this trend, however it is less clear with this methylation context. The trend is even less visible again with the CHH DMRs. This can be seen clearly in the large example in Figure 5.1 from line 2134, this exemplifies the need to only compare directly inherited genomic regions. Results from all DH lines tested can be seen in Figures 5.2, 5.3, although smaller they are just there to illustrate the point mentioned above.

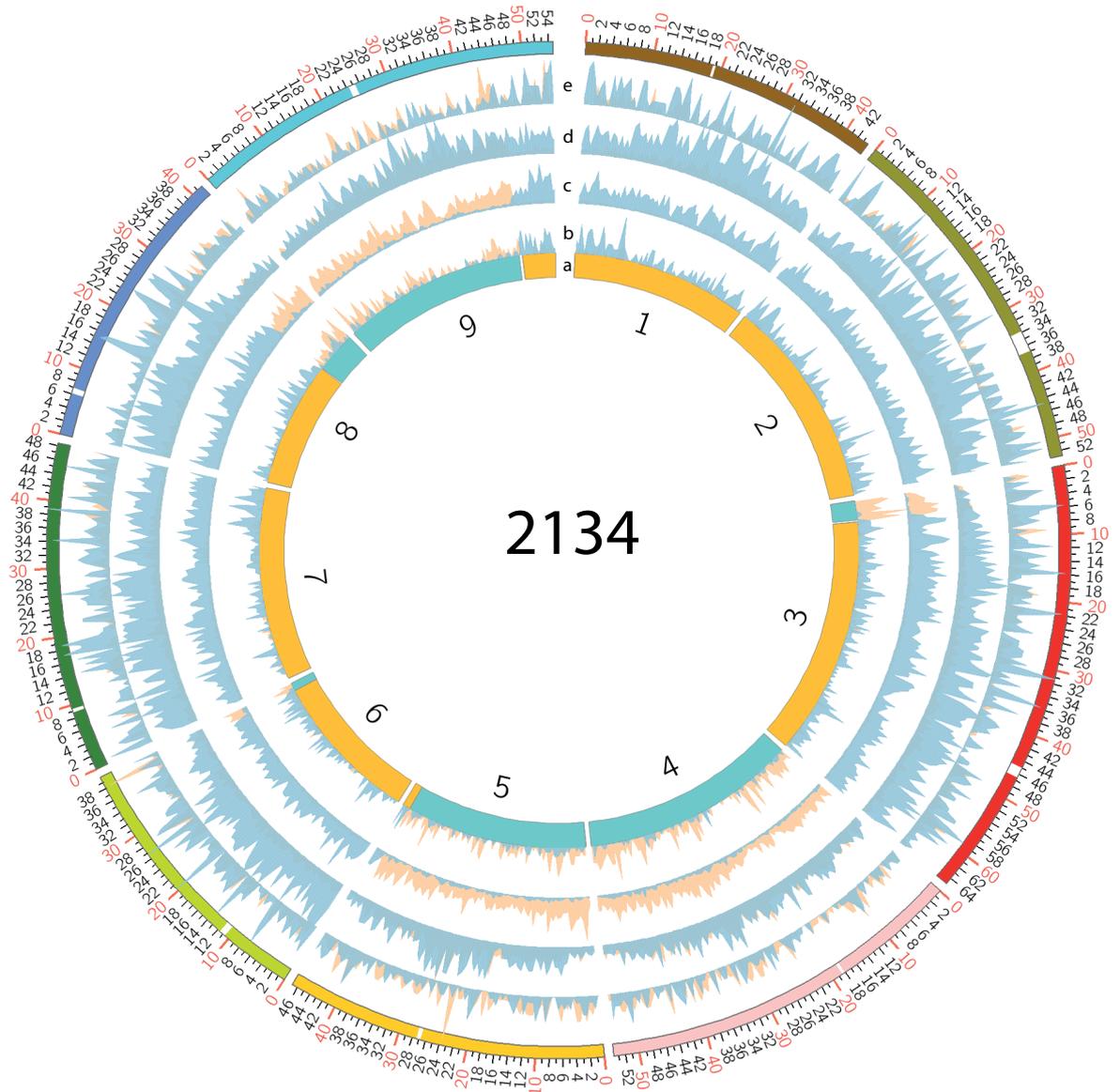


Figure 5.1: Large example of circos plots for whole genome comparisons. Whole genome comparisons between any DH line and either parent are invalid, illustrates the need to only compare inherited regions between the parents and DH lines. Results from all DH lines tested can be seen in Figures 5.2, 5.3, although smaller they are just there to illustrate the point mentioned above. Whole genome comparison of the DH line 2134 and GDDH33 in blue. Then whole genome comparison of DH line 2134 and A12Dhd in yellow. a) Predicted genotype of the DH line. b) Magnitude of DEGs between the DH line and both parents. c) Magnitude of CG DMRs between the DH line and both parents. d) Magnitude of CHG DMRs between the DH line and both parents. e) Magnitude of CHH DMRs between the DH line and both parents. The outer ideogram displays the chromosome coordinates in megabases along with the centromere positions as a white band.

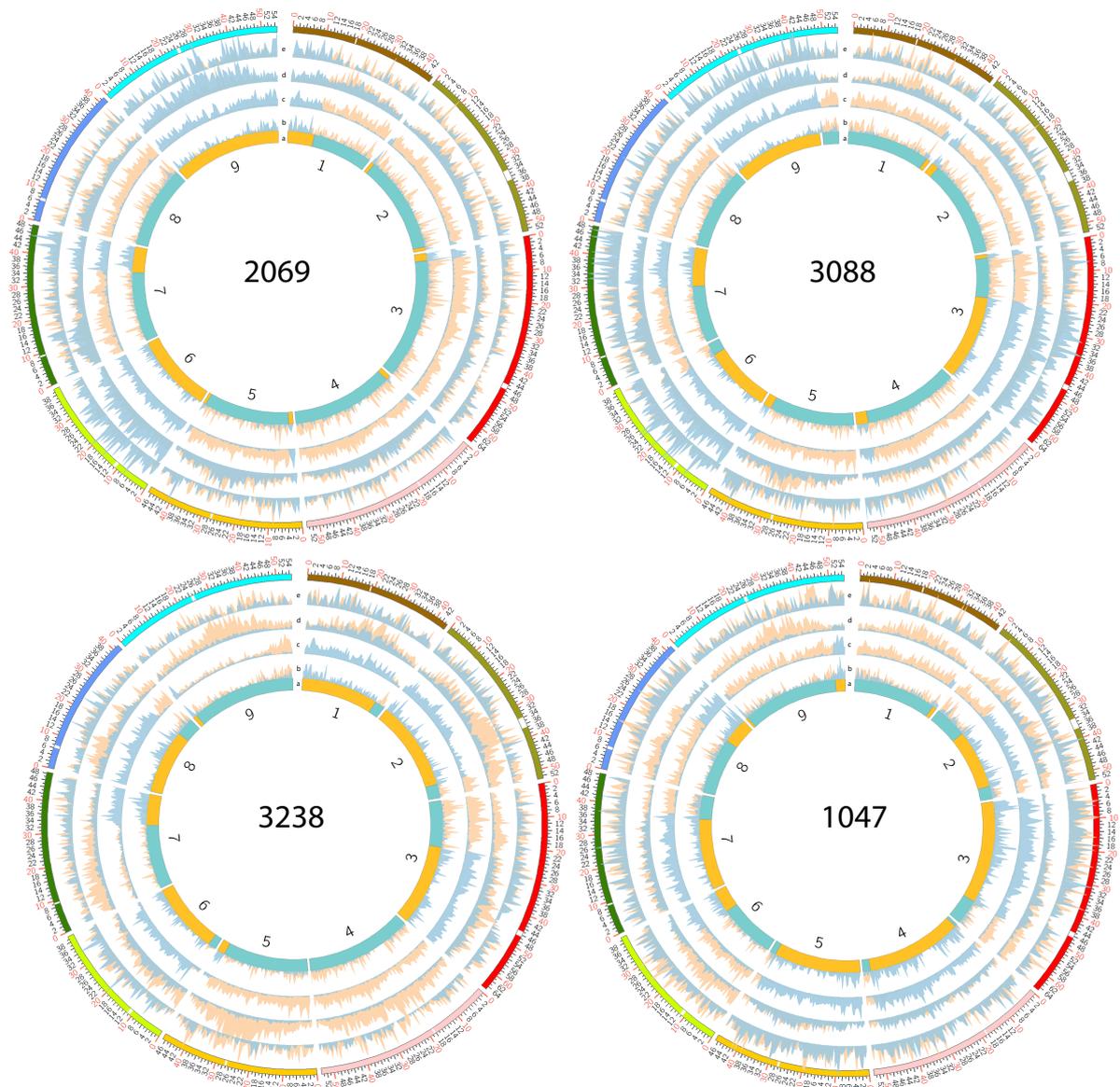


Figure 5.2: **Circos plots for whole genome comparisons. Whole genome comparisons between any DH line and either parent are invalid, illustrates the need to only compare inherited regions between the parents and DH lines.** DH line vs GDDH33 in blue and DH line vs A12Dhd in yellow. a) Predicted genotype of DH lines. b) Magnitude of DEGs between the DH lines and both parents. c) Magnitude of CG DMRs between the DH lines and both parents. d) Magnitude of CHG DMRs between the DH lines and both parents. e) Magnitude of CHH DMRs between the DH lines and both parents. The outer ideogram displays the chromosome coordinates in megabases along with the centromere positions as a white band.

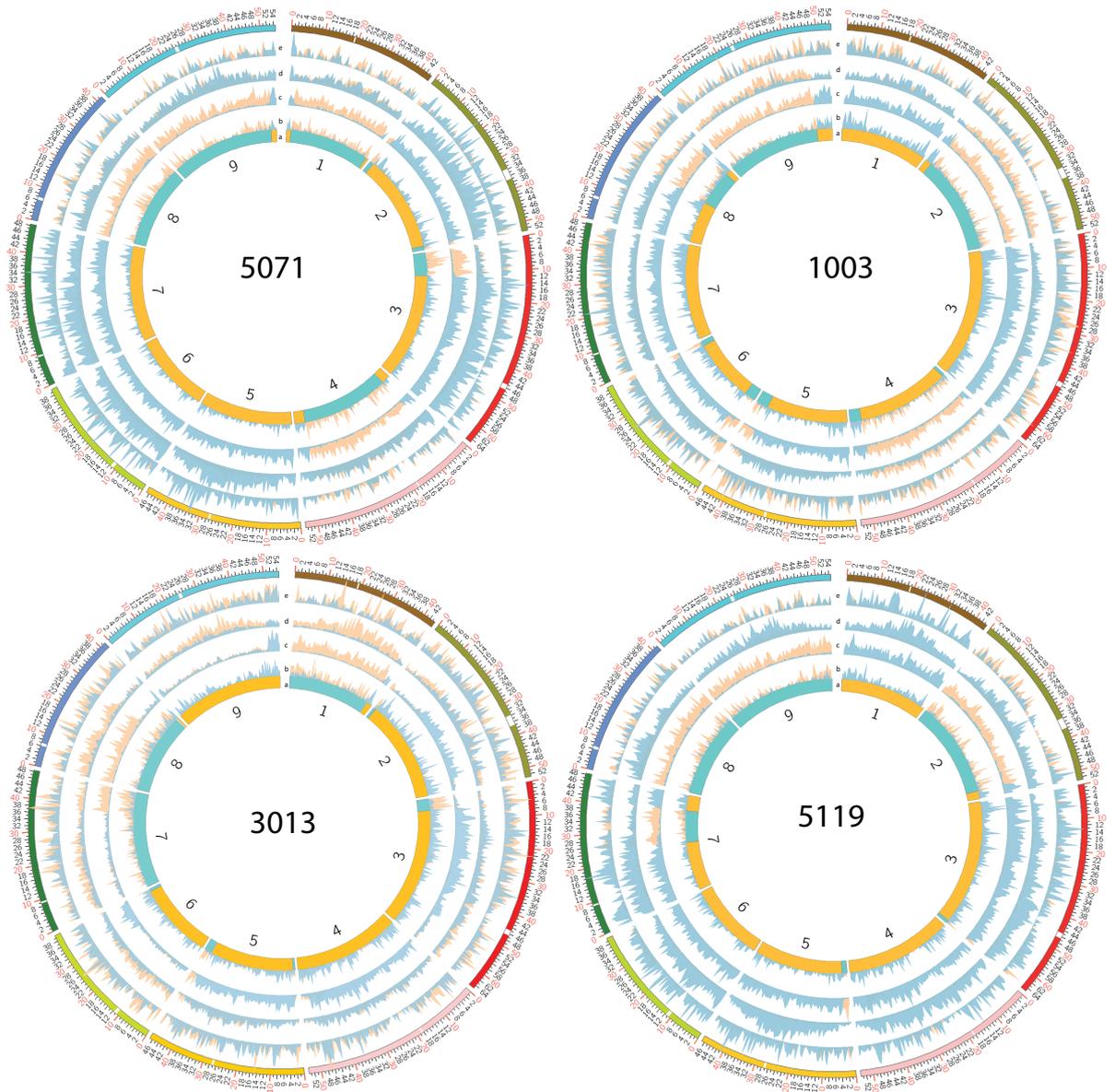


Figure 5.3: **Circos plots for whole genome comparisons. Whole genome comparisons between any DH line and either parent are invalid, illustrates the need to only compare inherited regions between the parents and DH lines.** DH line vs GDDH33 in blue and DH line vs A12Dhd in yellow. a) Predicted genotype of DH lines. b) Magnitude of DEGs between the DH lines and both parents. c) Magnitude of CG DMRs between the DH lines and both parents. d) Magnitude of CHG DMRs between the DH lines and both parents. e) Magnitude of CHH DMRs between the DH lines and both parents. The outer ideogram displays the chromosome coordinates in megabases along with the centromere positions as a white band.

This means that whole genome comparison between DH lines and either parent are confounded by how much of the parental genome was inherited. For this reason, it was vital to have coordinates for the HR sites, and with good resolution to ensure that the minimal amount of data was discarded and then only compare directly inherited regions between the DH lines and the relevant parent. The bisulphite sequencing data contains information about each samples methylome, and DNA sequence. This allows the genotyping of individuals through SNPs and epialleles. SNPs are likely to be the most reliable of the two methods due to the reduced mutation rate observed in DNA compared to DNA methylation. Using the developed SNP genotyping pipeline, 320000 homozygous genome positions were identified that can distinguish the two parental genomes. This equates to approximately one SNP every 1500bp. We could isolated a HR site to 130036 bp with the smallest resolution being 2364 and the largest being 807229. On average there are 0.88 HR sites per chromosome per line and each line has on average 8 HR sites. The most HR sites are on chromosome 2 (15) the most common was 1 HR site (42) and 24 chromosomes had no HR sites. Further to this, the 9 lines in this study vary in the whole genome contribution from both parents. This varies from 70% GDDH33 in line 2069 to 70% A12Dhd in line 3013 (Figure 5.7). The SNP identified HR sites agreed 100% with every HR sites from the SSR marker being identified. But the resolution was greatly increased from 90 markers to 320000 markers.

The whole genome bisulphite sequence data also allows the parental genomes to be distinguished by their methylation state. As shown in the circos figures, CG is the most reliable context for epigenotyping, originally this method was tried with all methylation contexts but results were inconclusive and so only CG methylated sites

were used for the verification of the SNP HR sites. A pipeline was implemented described in Hofmeister et al. (2017) for this verification, which uses the methylation state to identify HR sites. The pipeline works on a very similar principle to SNP genotyping. After running the pipeline for all samples, results from SNP genotyping and epigenotyping were combined to check their agreement (Figure 5.4, 5.5, 5.6). An epiHR site can either be; in agreement with the location of the snpHR site, in disagreement of the location with snpHR site, missing in snpHRs data or missing in epiHRs data. We found that out of the 81 chromosomes and 71 HR sites surveyed. There were 5 sites where the epiHR sites were in disagreement with the snpHR sites but within 1Mb. There were 3 sites shown to exist in the epiHRs but not in the snpHRs and 3 sites that were identified by the SNP data but not found in the epi data. Of note here is that all 3 of the extra HR sites originate from one DH line, 3088. For the majority of lines it would be possible to accurately predict HR sites from epigenetic variation alone. However some false positives and false negatives could occur. For this chapter's analysis, the data from the SNPs was used because it was verified by the marker map in 100% of cases and represents the inherited DNA sequence.

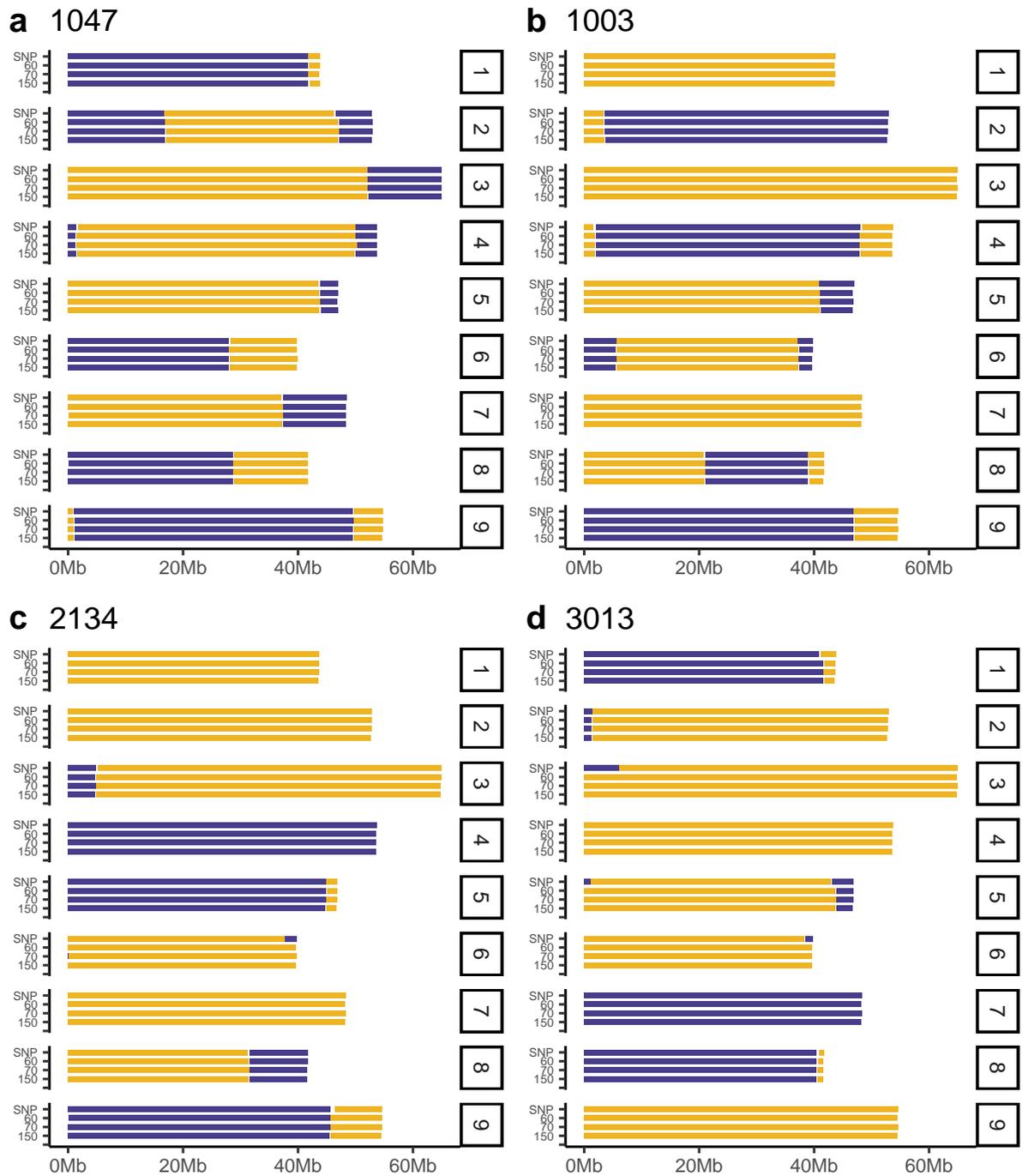


Figure 5.4: **Plots of results from SNP genotyping and epigenotyping.** For each chromosome the observed parent of origin is shown as 4 bars (yellow - A12Dhd, blue - GDDH33). Bars from top to bottom - SNP data genotype, 60kb epigenotyping, 70kb epigenotyping and 150kb epigenotyping.



Figure 5.5: **Plots of results from SNP genotyping and epigenotyping.** For each chromosome the observed parent of origin is shown as 4 bars (yellow - A12Dhd, blue - GDDH33). Bars from top to bottom - SNP data genotype, 60kb epigenotyping, 70kb epigenotyping and 150kb epigenotyping.

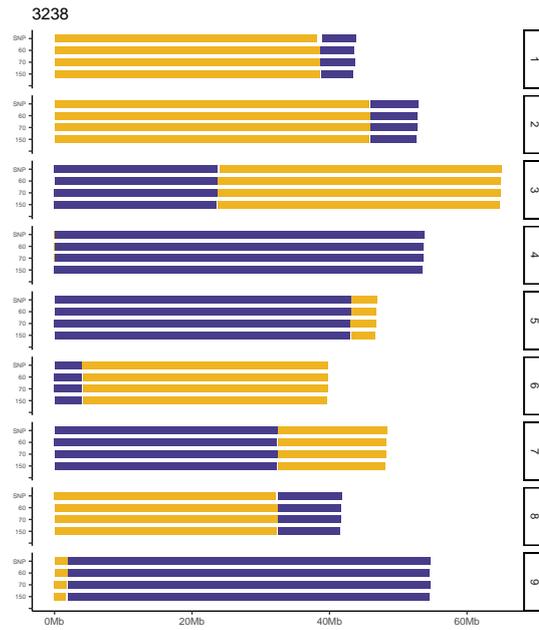


Figure 5.6: **Plots of results from SNP genotyping and epigenotyping.** For each chromosome the observed parent of origin is shown as 4 bars (yellow - A12Dhd, blue - GDDH33). Bars from top to bottom -SNP data genotype, 60kb epigenotyping, 70kb epigenotyping and 150kb epigenotyping.

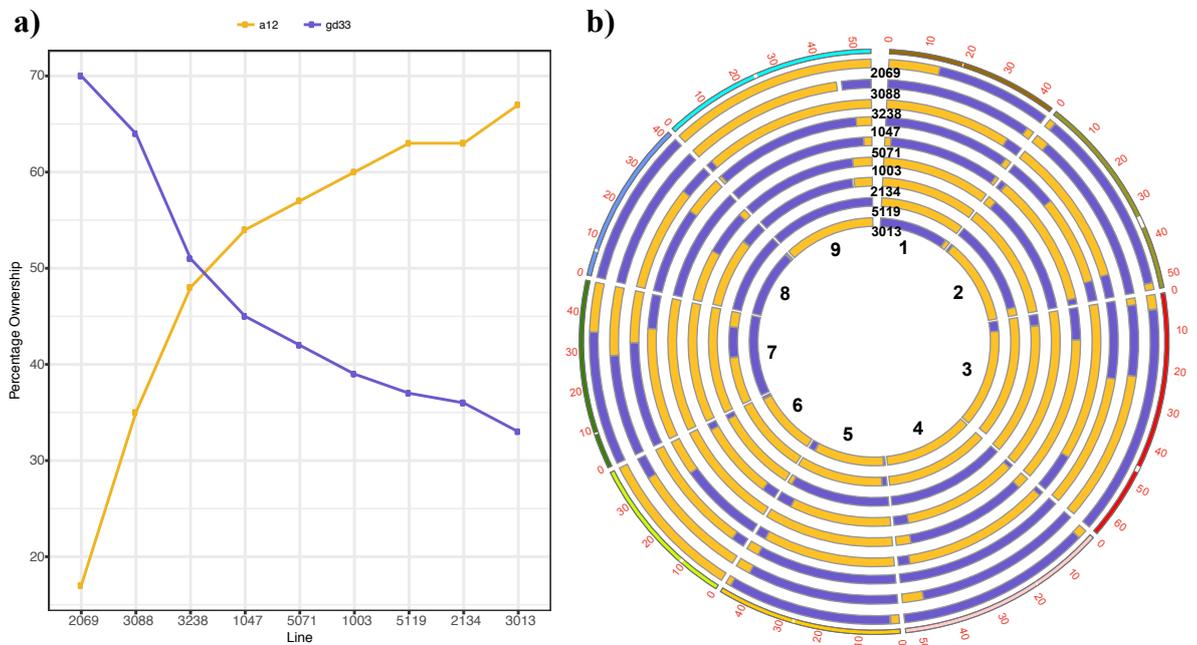


Figure 5.7: **Genotypes of DH lines.** a) Percentage of parental genome inheritance to each DH line. b) Circos plot. Each ring represents the diploid genome of a DH line. Each chromosome is coloured according to its inheritance.

5.4 Gene Expression Dynamics in the DH lines

The transcriptomes of the DH lines were split into A12Dhd inherited or GDDH33 inherited using the identified HR sites. Then pairwise comparisons for differentially expressed genes were drawn between genes in the DH lines and the relevant parent. From these comparisons we identified 1820 dhDEGs (ranging from 156 - 736). This equates to 0.3% - 1.4% of the transcriptome that experiences significant change in the doubled haploid lines. Comparing these dhDEGs to the phDEGs identified earlier, 79% of the phDEGs have normal parental-level expression in all DH lines (Figure 5.8 b). The phDEGs with recovered expression levels have an overrepresentation of genes that were additively expressed in the F1 hybrid (Figure 5.8 b) (X^2 (df = 4, N = 3254) = 145.7, p-value <0.001). The remaining 694 phDEGs still persist in at least one DH line and represent 38% of the dhDEGs. Of these, the dhDEGs inherited from the A12Dhd parent are enriched for genes that displayed GDDH33 expression level dominance in the F1 and conversely, the GDDH33 dhDEGs are enriched for genes that displayed A12Dhd expression level dominance in the F1 (Figure 5.8 a, b). As well as these expression level dominance (ELD) effects that are inherited from the hybrid, the DH lines also show dominance throughout the novel dhDEGs. This is exemplified by the fact that there are more GDDH33 inherited dhDEGs than A12Dhd inherited dhDEGs (Figure 5.8) T-test (t = -2.047, p-value = 0.03174). Additionally, the majority of the dhDEGs display expression-level dominance. This is where protein coding genes inherited from one parent assume expression patterns most similar to that of the other parent. The finding applies to dhDEGs inherited from both parents but is

most prominent with dhDEGs inherited from GDDH33 (Figure 5.9).

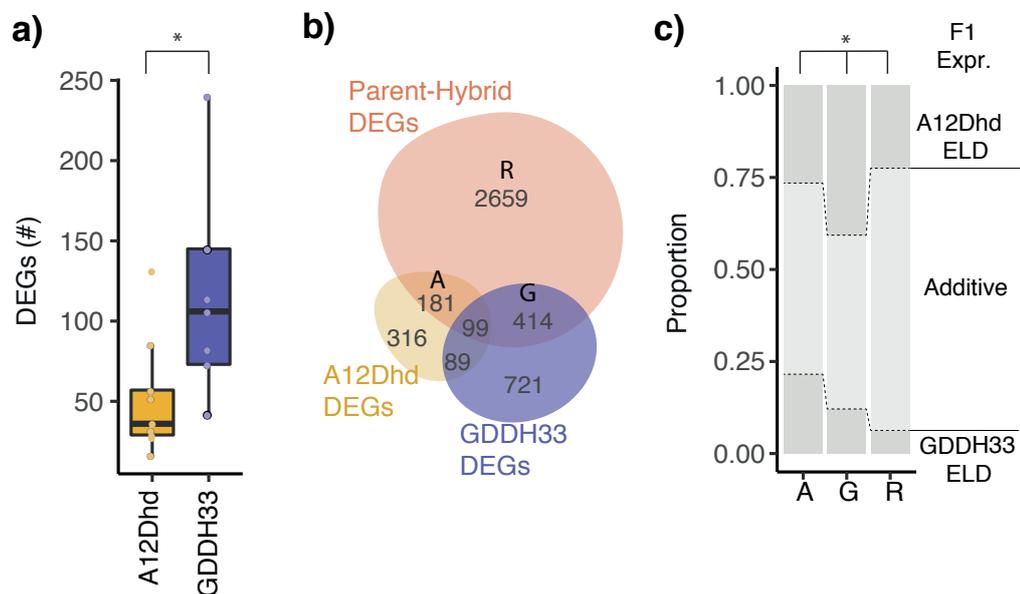


Figure 5.8: **Dynamics of dhDEGs in each of the parentally inherited genomes from each DH line.** a) There are more dhDEGs on GDDH33 inherited genomes compared to A12Dhd inherited genomes (T-test ($t = -2.047$, $p\text{-value} = 0.03174$)). b) Venn diagram with the parental and hybrid DEGs identified in Chapter 4 and the DH DEGs from each line, split by parental inheritance and their overlapping genes. c) Shows the F1 expression dynamics of the phDEGS that overlap with dhDEGs (A = phDEGs and A12Dhd inherited dhDEGs, G = phDEGs and GDDH33 inherited dhDEGs) and the phDEGs that recover in the DH lines (R) these are sections shown in the venn diagrams in panel b. There is a significant association between F1 expression and dynamics and the category of DEG - X^2 ($df = 4$, $N = 3254$) = 145.7, $p\text{-value} < 0.001$.

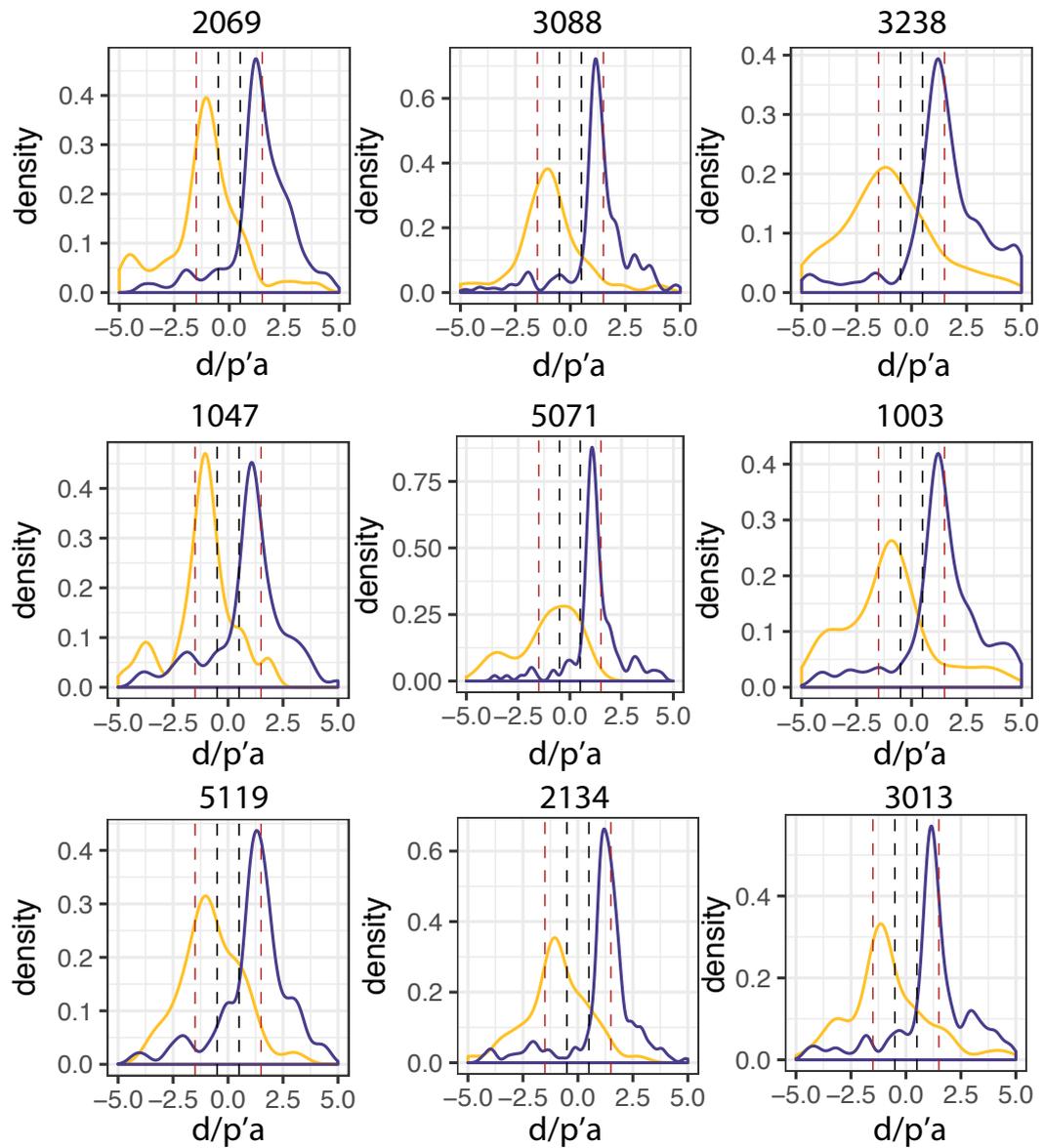


Figure 5.9: **Parental dominant-to-additive ratios of the dhDEGs, for each inherited genome dhDEGs tend to display expression dynamics similar to that of the other parental genome .** From left to right; Top - 2069, 3088, 3238, Middle - 1047, 5071, 1003, Bottom - 5119, 2134, 3013. For each line their A12Dhd inherited dhDEGs are shown in yellow and the GDDH33 inherited dhDEGs are shown in blue. The x-axis displays the parental d/a ratio, a ratio of 1 would mean a gene has equal expression to the gene in A12Dhd and a ratio of -1 means the gene would have equal expression to the GDDH33 parent.

In addition to the parent and hybrid influence and the expression level dominance of the dhDEGs, the total quantity of genome inherited from either parent also affects the transcriptomic stability of the doubled haploid lines (Figure 5.10). We find that there is a negative relationship between the amount of genome inherited and amount of relative gene expression change for both A12Dhd and GDDH33 inherited dhDEGs (Table B.1). This means that a DH line inheriting only 20% of its genome from A12Dhd could experience up to three times more relative gene expression changes in its A12Dhd inherited genes than another DH line which inherits 80% of its genome from A12Dhd.

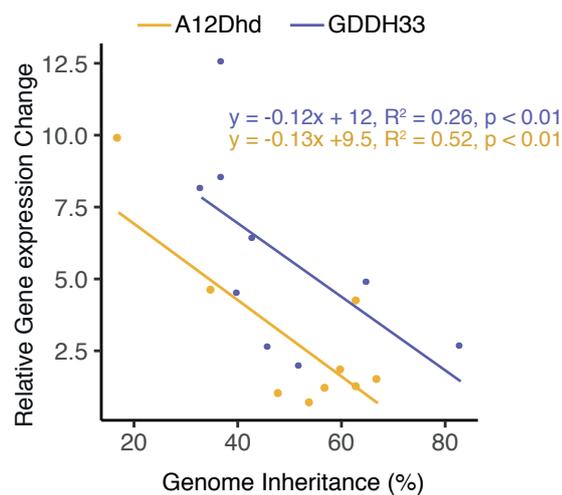


Figure 5.10: **There is a negative relationship between relative gene expression change and whole genome inheritance in the DH lines.** For each inherited genome in each DH line the relative gene expression change (dhDEGs per gene inherited) is plotted against the amount of genome inherited from that parent. The significant relationships are shown as lines calculated by linear regression.

5.5 Differentially Expressed Genes in the DH Lines

To assess if the DH lines have some common response in spite of their differing genomes we performed gene ontology enrichment analysis with each of the DH lines differentially expressed genes. Then enriched terms (FDR 0.05) from each DH line were combined to identify common enriched terms between the different lines (Figure 5.11). There were many terms found to be enriched in more than 6 of the 9 DH lines. These include terms relating mainly to the stress response and the cell wall but also include developmental processes and metabolic processes. This common transcriptional response is echoed when looking at the common dhDEGs between lines. We find that the DH lines share 75% of their DEGs with at least one other line. But unlike the GO terms only 16 dhDEGs are shared between more than 6 DH lines showing that even though a lot of the DH line have a common enriched terms, their response at the transcriptional level is largely unique.

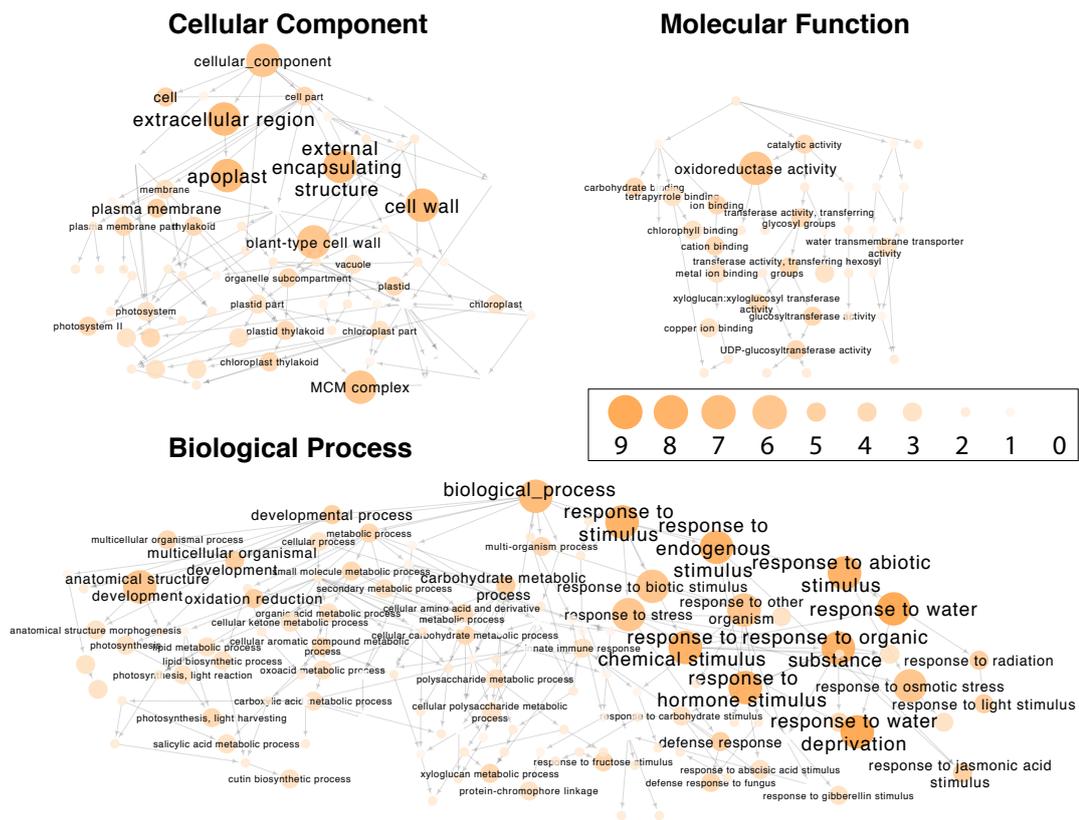


Figure 5.11: **Combined GO analysis for dhDEGs, each node represents an enriched GO term. (FDR <0.05). Size of node and label represents the number of DH lines a given terms is enriched in.**

5.6 Methylation Dynamics in the DH Lines

In the same way as for gene expression, the DH line's genomes were split into A12Dhd inherited and GDDH33 inherited using the previously identified HR sites. DMR comparisons were drawn only between directly inherited regions in the parents and DH lines. From these comparisons we identified 22233, 50111 and 54880 DMRs in CG, CHG and CHH context respectively (range; CG-1911,6431; CHG-3771,9430; CHH-5224,12021). There are twice as many dhDMRs in non-CG context and there are other clear distinctions between CG and non-CG dhDMRs. In agreement with the CG phDMRs, the CG dhDMRs display methylation dynamics similar to gene expression. The GDDH33 inherited CG regions are more sensitive to perturbation and experience 1.5 fold more CG dhDMRs, T-test ($t = -2.224$, $p\text{-value} = 0.0485$). These regions and particularly the GDDH33 inherited regions favourably associate with the genic portions of the genome and the data further shows that there is more hypomethylation of CG dhDMRs in the gene body in GDDH33 inherited genes than A12Dhd inherited genes (Figure 5.15). Furthermore we found that the majority of CG phDMRs (72%) regained normal parental methylation levels in all lines. This, coupled with the fact that there are ~5 - fold less CG-dhDMRs than CG-phDMRs shows that the genome merger in the DH lines has a less dramatic and different effect on the CG methylome than the F1 hybrid merger. We also find that CG-dhDMRs are affected by the amount of genome inherited from either parent (Table B.2,B.3). At low contributions from either parent (20%) a DH line there could have up to 3-fold more CG-dhDMRs on those inherited regions.

With non-CG dhDMRs we also find that there is a different response to that of

the F1 hybrid with most of the non-CG phDMRs returning to normal levels (Figure 5.12). These regions that are inherited from the hybrid are inherited from GDDH33 are enriched for regions that displayed A12Dhd dominance in the F1 and the opposite is true for the A12Dhd regions inherited (Figure 5.12) ($CG = X^2$ (df = 4, N = 22807) = 465.5, p-value <0.001), ($CHG = X^2$ (df = 4, N = 11946) = 483.6, p-value <0.001), ($CHH = X^2$ (df = 4, N = 20199) = 2509.5, p-value <0.001). But these regions only account for 10% of the non-CG dhDMRs. We find that there are different signatures within these non-CG DMRs. Firstly, there are more than twice as many non-CG changes compared to CG changes (Figure 5.12). We find that the non-CG DMRs do not show a preference for the GDDH33 genome which has been seen in every other analysis and that the magnitude in methylation is much less affected by the parental genome contributions of the DH lines (Figure 5.14, Table B.2, B.3). Interestingly, the largest fraction of genome occupied by these non-CG dhDMRs is in transposon sequences which account for around 60% of non-CG DMRs (Figure 5.15).

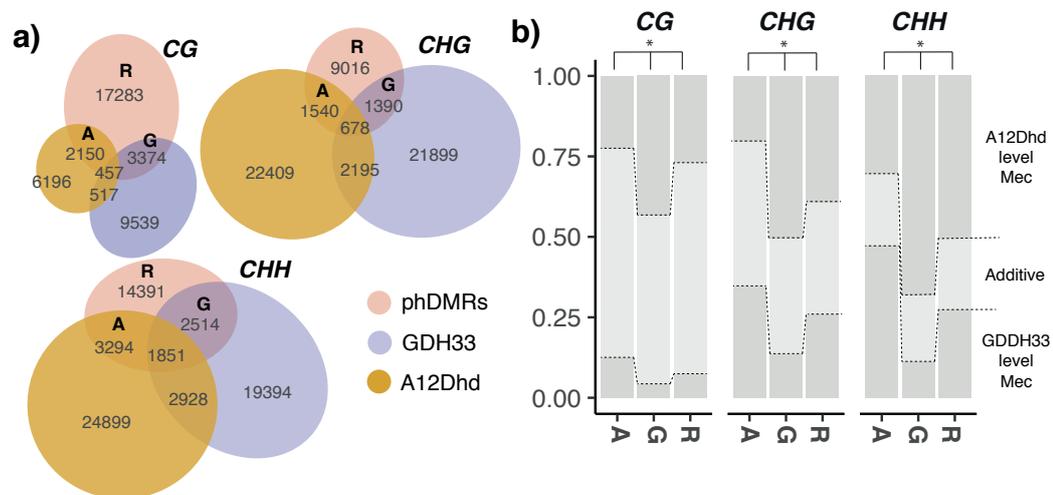


Figure 5.12: Comparison of CG, CHG and CHH dhDMRs with the phDMRs.
a) Venn diagram with the parental and hybrid DMRs identified in Chapter 4 and the dhDMRs from each line, split by parental inheritance and their overlapping DMRs.
b) Shows the F1 expression dynamics of the phDMRs that overlap with dhDMRs (A = phDMRs and A12Dhd inherited dhDMRs, G = phDMRs and GDDH33 inherited dhDMRs) and the phDMRs that recover in the DH lines (R) these are sections shown in the venn diagrams in panel a. There is a significant association between F1 methylation dynamics and the category of DMR -(CG = X^2 (df = 4, N = 22807) = 465.5, p-value <0.001), (CHG = X^2 (df = 4, N = 11946) = 483.6, p-value <0.001), (CHH = X^2 (df = 4, N = 20199) = 2509.5, p-value <0.001) .

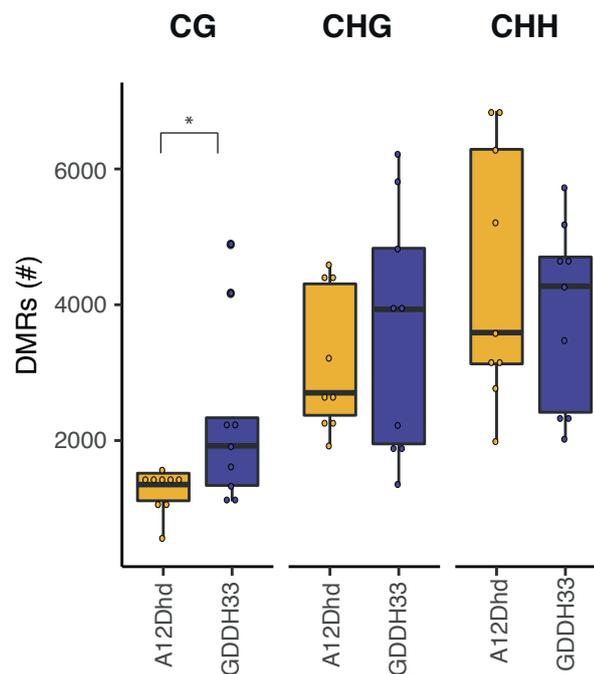


Figure 5.13: **Number of dhDMRs in each line split by parental inheritance for each sequence context.** There are more CG dhDMRs on A12Dhd inherited genome sections than GDDH33 genome sections (CG - T-test ($t = -2.224$, $p\text{-value} = 0.0485$)). CHG and CHH inherited sections do not show significant differences (CHG - ($t = 0.601$, $p\text{-value} = 0.5583$), CHH - ($t = 0.743$, $p\text{-value} = 0.4689$))

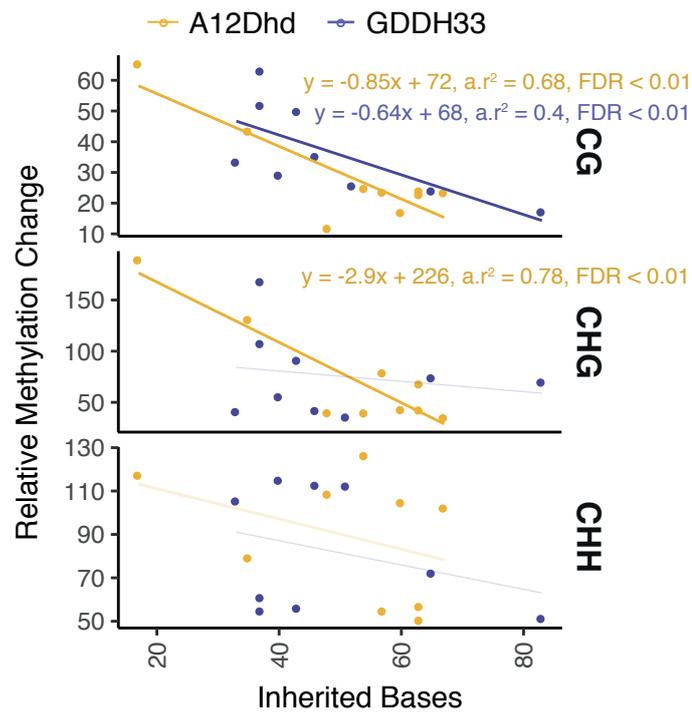


Figure 5.14: **Relationship between parental genome dosage and epigenetic changes in DH lines.** For each inherited genome in each DH line the relative gene methylation change (dhDMRs per MR inherited) is plotted against the amount of genome inherited from that parent. The significant relationships are shown as lines calculated by linear regression. Non significant relationships are shown as a faint line.

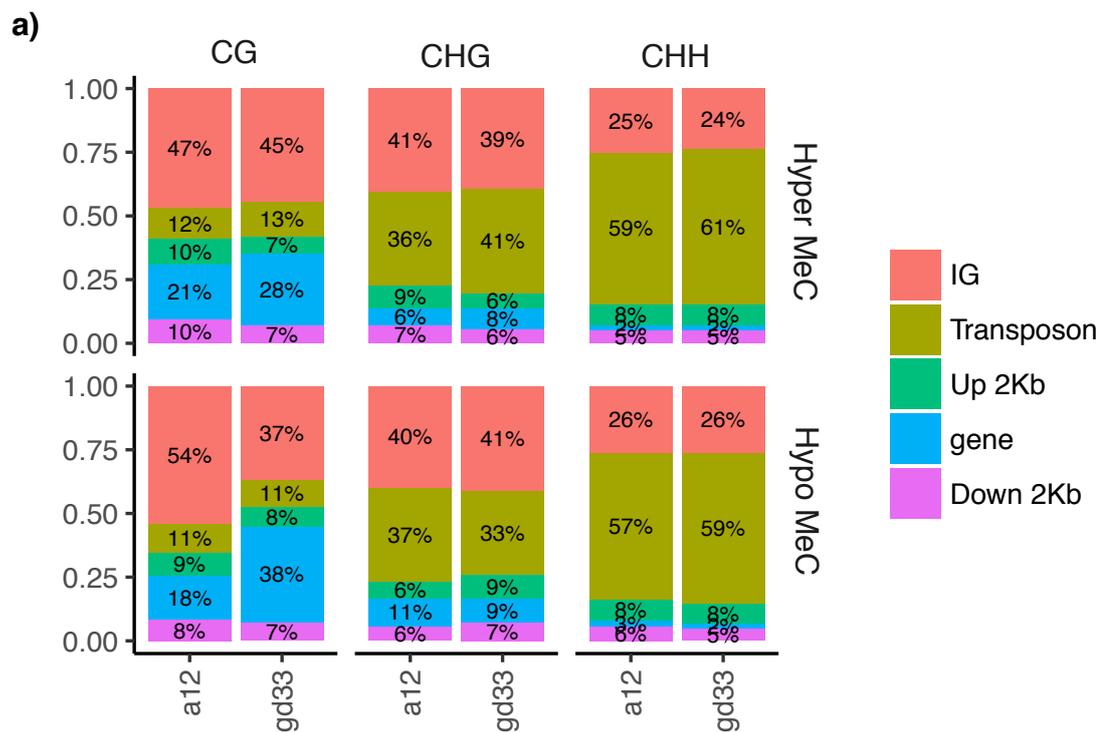


Figure 5.15: **Locations of dhDMRs in different genomic features.** Each base of the DMRs are assigned to the genomic feature that it overlaps with in the annotation, these are displayed as a proportion of the total bases of that DMR category. This is done in a hierarchical fashion to account for overlapping features (gene, transposon, upstream, downstream, intergenic: in order of decreasing importance)

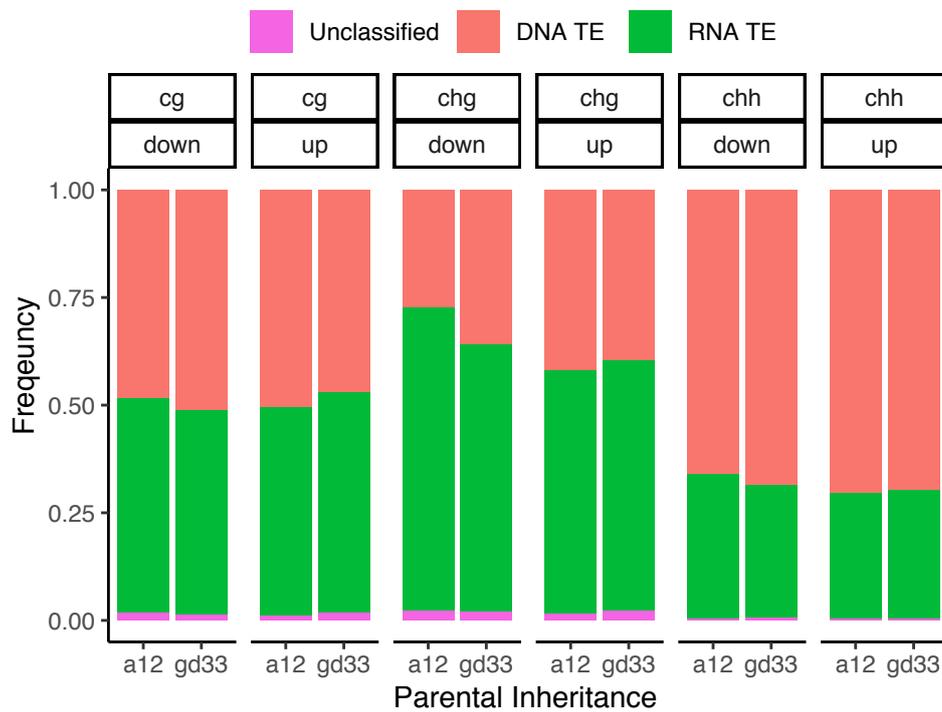


Figure 5.16: **Locations of dhDMRs in different transposon types.** Each base of the DMRs is assigned to the transposon type that it overlaps with in the annotation, these are displayed as a proportion of the total bases of that DMR category.

5.7 Correlation Between DNA Methylation and Gene Expression in *B. oleracea* DH Lines

To identify dhDMRs that could have a detectable functional link, I developed a program (GFFIntersector) that is capable of intersecting genomic coordinates of genes and DMRs. All genes in the reference genome were included as possible intersections. This total genic space including exons and introns in the annotated reference is 20.35% (99401595 bp). In total, there are 151648 dhDMRs, these occupy 25,979,919 bp (5.31%) of the *B. oleracea* reference genome. Then, using GFFIntersector, both sets of coordinates were intersected to find intersecting regions. In total we identified 19,240 dhDMRs that physically intersect with 12,875 genes. The different context dhDMRs have different locations within the genic features. CG dhDMRs are more abundant in the gene body. Whereas, CHG and CHH dhDMRs are most abundant within the up and downstream regions, but this is most prominent with CHH methylation (Figure 5.17).

Those regions of methylation that have a conserved function in directly regulating gene expression will be expected to have correlating methylation and gene expression patterns in all samples. Therefore we correlated gene expression values and methylation values for each gene and intersecting DMR block to identify significant interactions (FDR <0.01). In total we identified 247 genes with sufficient data in at least ten samples and with significant correlations. This is 2% of the actual genes that have associated dhDMRs. Forty of these are uncharacterised in the annotation and five are annotated as retro-transposons. During the analysis we selected genes with DMRs that have significant correlation between a gene and at least 1 DMR. Upon manual in-

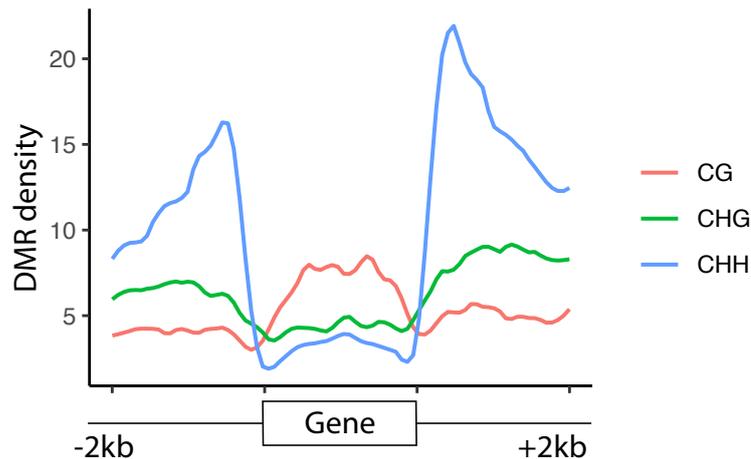


Figure 5.17: **The methylation contexts of the dhDMRs differ in their distribution across genes with which they overlap.** Density plot showing the distribution of dhDMRs in each sequence context over genes. For each feature and context the 2 kb flanking regions show the number of DMRs that are assigned to a particular position for each feature in the genome. Across the gene body each gene is split into 100 bins and the plot displays the number of DMRs that reside in that bin from all features.

Inspection of these regions we found that many genes have multiple associated dhDMRs which do not all have methylation that correlate with the expression of the gene. This makes assigning functional relevance to DMRs in some of these regions difficult. To this end we assessed each region manually for its clarity of methylation. After this manual inspection we have chosen to investigate *agamous*-like (Bo6g014360), *fas4* (Bo9g121160) and *TIL* (Bo9g134760) further because of their defined gene functions, GO terms which appear enriched in the dhDEGs and clear DMR signature.

In the case of *Agamous*-like (BO6G014360), we find that a DMR exists in the 3rd exon and 3rd intron of the gene (Figure 5.18). Further to this we find that there is also a transposon within the 3rd exon of the gene, exactly overlapping with the DMR. Ten DH samples comprising of 6 DH lines had sufficient RNA and methylation data to perform the correlations along with both parent and F1. All of these DHs inherited this region from A12Dhd. However, 4 of these lines display methylation and expression

values most similar to the GDDH33 parent (Figure 5.18). Upon closer inspection of the RNASeq reads for this region it was found that the exon boundaries are not observed by the locus. This could indicate that this gene is no longer a protein coding gene or that the reference annotation for this locus is incorrect (Figure 5.19).

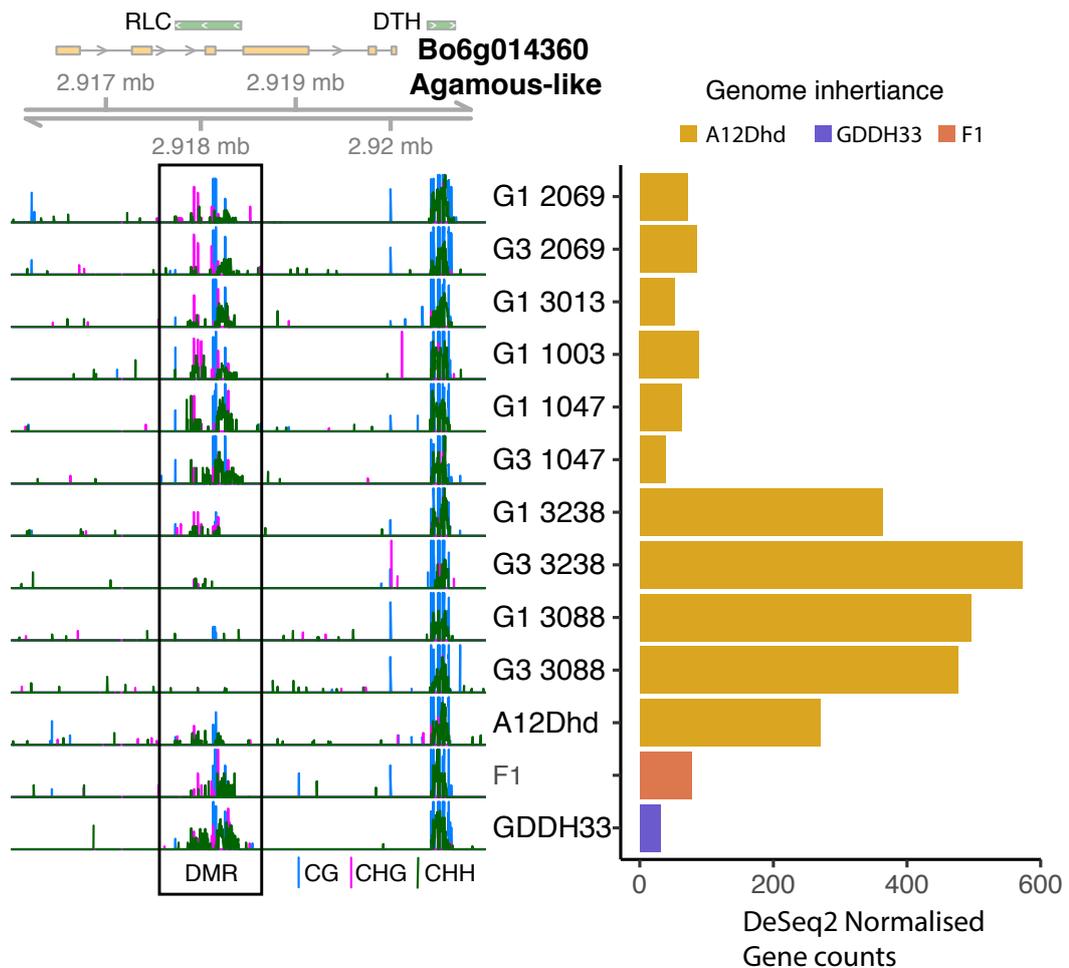


Figure 5.18: **Methylation and expression of the Agamous-like locus.** The left hand side shows single base resolution methylome of the available samples for this region. The right hand side shows the expression of this locus in the available samples. Bars are coloured to show the genotype of this locus.

The second example, Fas4 homolog (Bo9g121160) contains a large DMR that

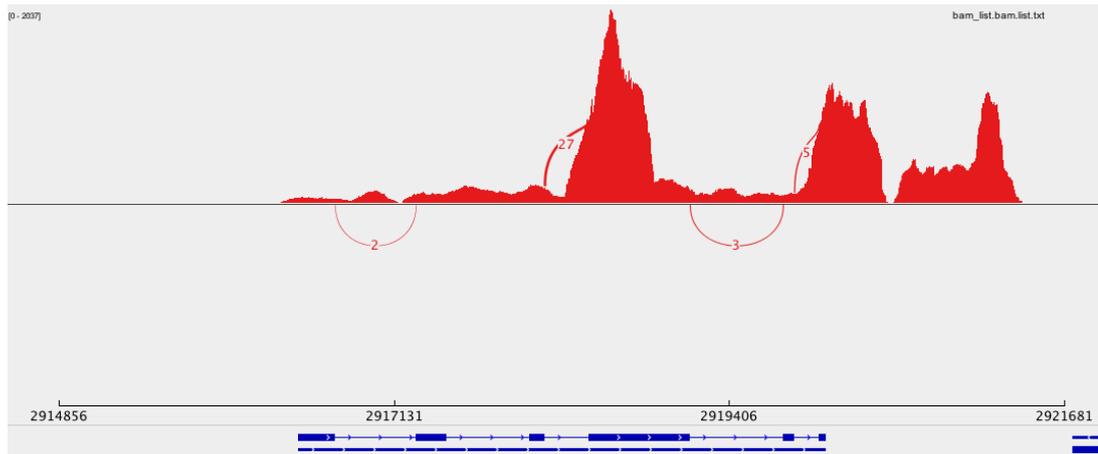


Figure 5.19: **Aligned reads for Agamous-like locus does not follow the Parkin et al. (2014) annotation.** This plot shows aligned reads from all samples for this locus. It shows that there is little evidence for the exon boundaries found in the official annotation. However there are transcripts produced from this locus. The annotation for Agamous-like is shown in blue at the bottom. The coordinates for chromosome 6 are shown above and then the aligned reads for all samples are shown in red with the number of reads from these that actually support the annotation shown as red numbers.

spans the first five exons of the gene. At this region, the methylation and the expression of the gene are correlated across seven DH lines and the parents and hybrid (Figure 5.20). Interestingly, a transposon at the TSS is heavily methylated in GDDH33 parent and the lines inheriting this region, but this methylation does not exist in A12Dhd. The lines inheriting this region from A12Dhd also lack methylation at this region, with exception of line 3238. Line 3238 displays very similar methylation to the GDDH33 parent and also shares GDDH33's gene expression pattern. Further to this the methylation and expression is not recovered after two generations of selfing in line 3238.

The third example, temperature-induced lipocalin (Bo9g134760) also contains a large gene body DMR that spans the length of the gene. Seven DH lines, both parents and the F1 had sufficient data to correlate the methylation and expression values

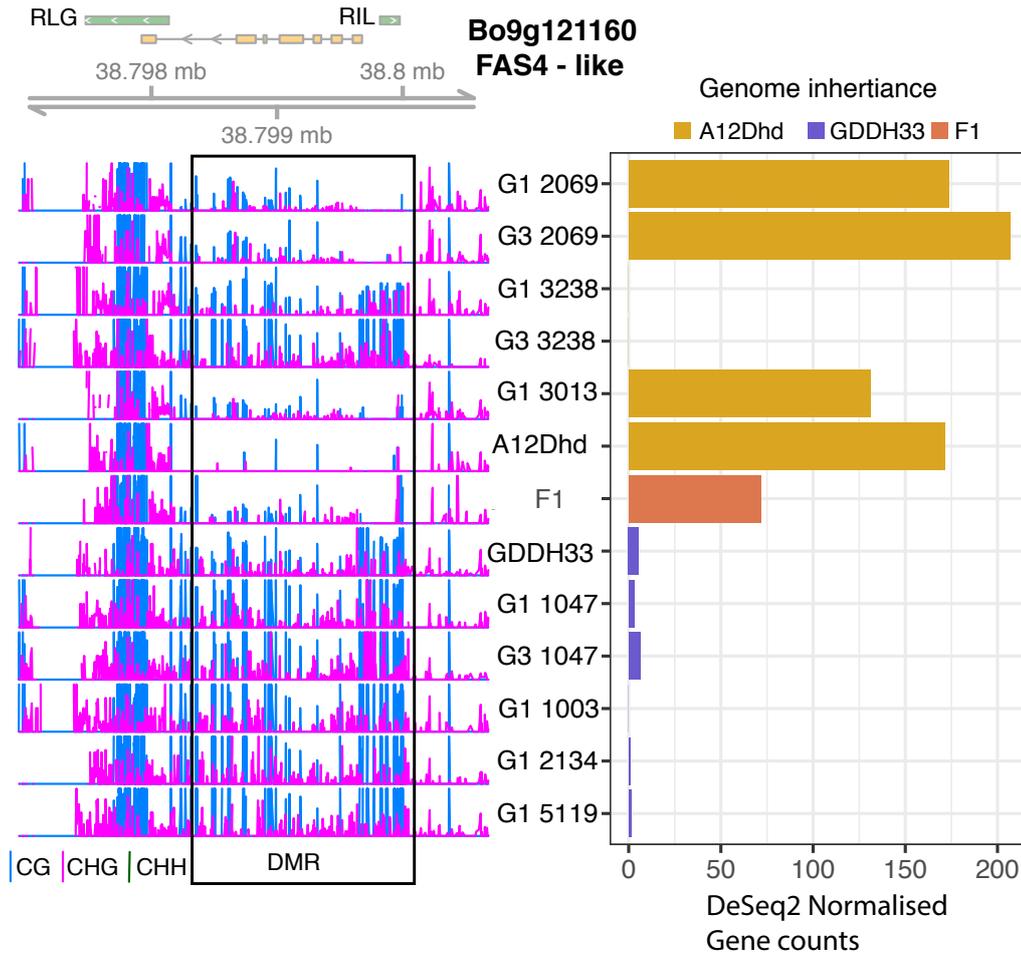


Figure 5.20: **Methylation and expression of the FAS4 locus.** The left hand side shows single base resolution methylome of the available samples for this region. The right hand side shows the expression of this locus in the available samples. Bars are coloured to show the genotype of this locus.

(Figure 5.21). We find that the DH lines inheriting the region from GDDH33 all have GDDH33 expression levels combined with low methylation levels also seen in the GDDH33 parent. But from the lines that are A12Dhd inherited for this region we find one line (3238) that has GDDH33 methylation and expression in generation 1. However when looking in this line (3238) after two generations of selfing some of the methylation has recovered, along with the gene expression level.

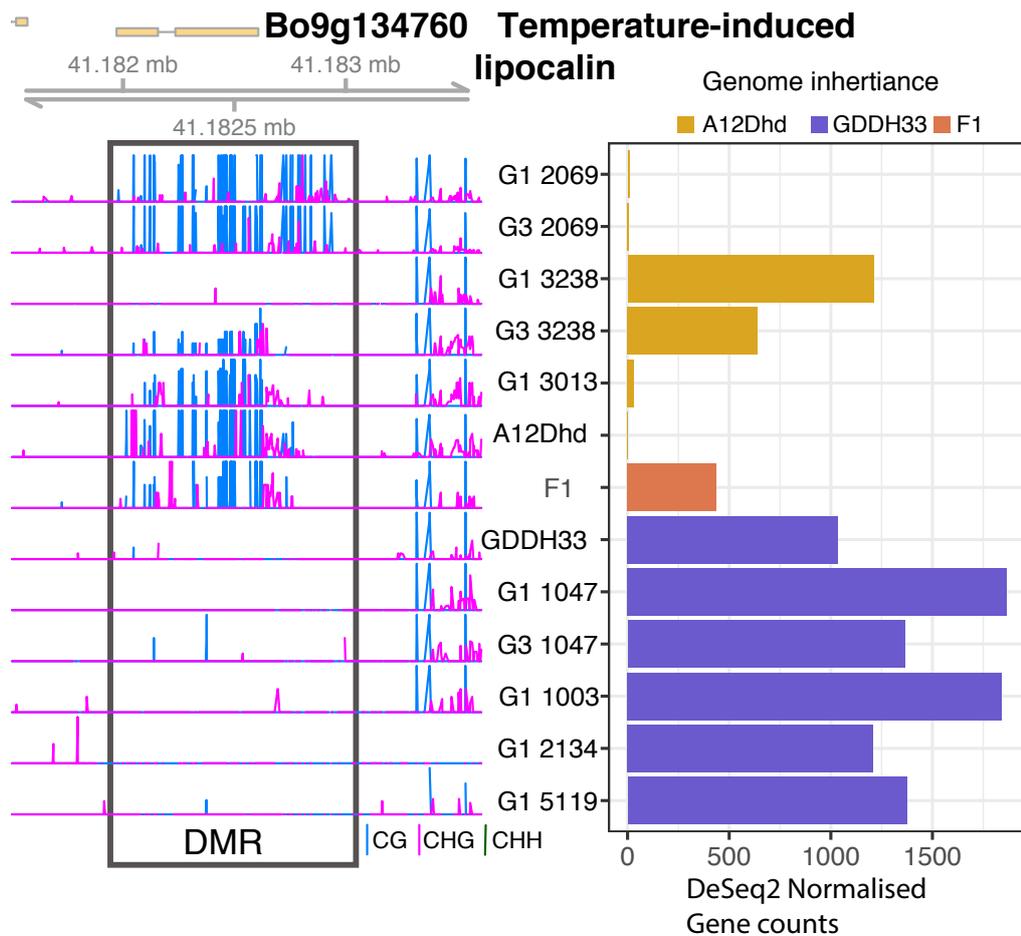


Figure 5.21: **Methylation and expression of the TIL locus.** The left hand side shows single base resolution methylome of the available samples for this region. The right hand side shows the expression of this locus in the available samples. Bars are coloured to show the genotype of this locus.

5.8 Discussion

As shown in previous studies, the offspring of a genome merger can experience many different forms of gene expression and epigenetic change relative to their parents (Swanson-Wagner et al., 2006; Yoo et al., 2013). However these studies have only been focussed on the genome mergers of F1 hybrids and polyploids (Chen, 2013). Our data shows that the DH lines also show altered methylation and expression patterns when compared to their parents, yet the perturbations in the DH lines are largely distinct and of a lesser magnitude than those occurring in the F1 hybrid. This means that many of the changes in gene expression witnessed in the F1 revert to normal parent levels in the DH lines and that the change in regulatory landscape of the DH lines is different to that of an F1 hybrid. This amelioration of F1 gene expression changes has also been reported after polyploidisation from F1 hybrids in *Senecio cambrensis*, here we show that the same effect is present in DH lines (Hegarty et al., 2008). However, in the DH lines this diminution in the effect of F1 gene expression change is simply due to the regaining of a homozygous genome. Whereas the *Senecio* polyploids described by Hegarty et al. (2008) also have a full complementation of both parental alleles. Meaning that it is not a full set of alleles but a homozygous genome structure that is responsible for this reduction in change in this case.

Even though the F1 and DHs have distinct responses, there are similarities between these genome mergers. One of the most prominent forms of gene expression changes shown to occur in F1s and polyploids is expression-level dominance (ELD) (Otto, 2007; Song and Chen, 2015). Our results show that these effects are not only a signature of genome confrontation in many F1s and polyploids but DH lines as well

(Chen, 2013; Bottani et al., 2018). Further to this, in agreement with results from the F1, we also find that this phenomenon extends to methylation as well, and in particular CG methylation. We find that there is a bias for changes, particularly gene expression and CG methylation, for the GDDH33 paternal genome. And that DEGs and DMRs inherited from one parent often display dynamics similar to that of the other parent. It has been shown that the TE load of the parental genomes is responsible in some cases, where spreading of methylation can occur to nearby genes (Cheng et al., 2016). Bottani et al. (2018) also elegantly described the effect that different genome sizes between the parents could have on TF affinity and cellular concentrations of proteins as well as the effect of mis matches in trans regulators. In this study, the genome sizes of the two parents are identical and so at least in this example, ELD and MLD can occur even with similar genome sizes. Further to this, at least in the case of 3 genes, DNA methylation could also be correlated with these changes. This could be a secondary change to the transcriptional response and the changing of heterochromatin states or a causative factor (Zhang et al., 2016). In the majority of cases we find no significant correlations between gene expression and methylation, showing that the real relationship between gene expression and methylation is complicated and often indirect.

This transcriptional and epigenetic change is often linked to chemical and physiological changes within the cell and phenotype (Greaves et al., 2016; Lauss et al., 2018). In fact, it is these effects resulting from hybridisation and polyploidisation that is often attributed to the majority of plant evolution and why these tools are used ubiquitously by plant breeders (Otto, 2007). However there are many choices to be made during plant breeding and trial and error wastes human and commercial resources (Langridge

and Fleury, 2011). A marker assisted breeding program will typically involve growing many hundreds DH lines with selected markers and assessing for the phenotype of interest (Langridge and Fleury, 2011). The ability to predict, even in part, the stability of the transcriptomes and methylomes of a DH line before extensive phenotypic analysis can streamline current processes. Our data shows a number of factors that are linked to the magnitude of variation in the transcriptomes and methylomes of these DH lines. These include; changes persisting from F1 hybrid, A12Dhd dominance and the combination of both parent genomes to the DH line. The changes persisting from the F1 have the smallest effect on DH perturbations. As in other species, these changes mainly rely on differences already existing between the parents. And some studies have shown that more diverse parents of F1 crosses have have a larger shock effect to the epigenome and transcriptome (Greaves et al., 2012). This, combined with the A12Dhd dominance discussed earlier comprises the predictive elements of the DH lines perturbations relating to the parents. These changes can be specific to specific parent combinations and different combinations of parents can produce different shock effects in their progeny and result in more or less perturbations in the offspring genome (Van Gioi et al., 2017). The total genome contribution of each parent to the DH line is also correlated with the magnitude of change in the DH lines which can also be another useful predictor for early selection of lines.

Previous studies focussed on improving doubled haploid breeding strategies have been largely centred around the efficiencies of DH production (Ferrie and Caswell, 2011). But DH lines also provide a chance to study an interesting genetic phenomenon, the genome merger. Currently this has been addressed in F1 hybrids and polyploids with some success (Chen, 2013). However genome mergers have not been

Chapter 5. Transcriptome and Methylome Analysis of the Doubled Haploid Lines

studied in the context of DHs. The data presented here represents the first genome wide transcriptomic and epigenetic analysis of DH lines and shows that DHs provide an additional model for studying genome confrontation that allows the prediction of genetic and epigenetic variation that can be used during plant breeding.

Chapter 6

Discussion

Plants are the basis of most food chains and life on earth. Because of this, plant science is a fast moving field with considerable investment from both governmental and commercial sources. The knowledge gained from this investment has a history of being adopted quickly by plant breeders, often leading to increased yield (Eriksson and Ammann, 2017). Currently, two major advances have been observed. Firstly, the birth of quantitative genetics in the early 1900's and secondly in the green revolution of the 1960's. Now, omics technologies are becoming common place, they are set to reshape plant breeding once again. We have already seen the start of this with MAB technologies; with seed chipping and genotyping for MAS, breeders are able to remove plants during line selection before resources are wasted (Butruille et al., 2015). This means more phenotypic selection can be applied to these complex traits which may only be present in certain growing conditions or growth stages and may even require destruction of the material to measure. Even still, with the population set to increase to 9 billion in 2050, yield will have to double to accommodate this population

size (Hickey et al., 2017). Therefore, there needs to be continued development in all areas of plant breeding programs to meet this demand (Glenn et al., 2017).

For an individual cultivar, its creation could be considered a linear process of descent. On the other hand, large scale plant breeding programs are ongoing cyclical processes, with new genetic variation going in and superior cultivars coming out (Figure 6.1). During this process, genetic gain can come from introducing new variation into the program, or from speeding up selection processes allowing varieties to be produced faster (Langridge and Fleury, 2011). Introducing new genetic variation is part of the pre-breeding stage, desirable traits and / or markers are selected and these are introduced into the breeding program through crossing (Hybrid crossing or back-crossing). This process can be very inefficient and wastes many resources on crosses that do not perform well or markers that do not behave as expected (Witcombe et al., 2013). The other way to improve genetic gain is through increased selection efficiencies; line selection occurs at a number of stages throughout the breeding program. This can be in selecting advantageous hybrids or in selection of advantageous homozygous lines (Figure 6.1). Currently, this selection process relies on phenotypic study and often the presence of a genetic marker or markers. The data discussed here shows that techniques such as RNA sequencing that provide greater resolution of the genome can provide insights into the mechanisms underpinning these inherited traits. By understanding the mechanisms it can be possible to increase efficiencies when choosing parents for crosses, selecting hybrids, selecting recombinant homozygous lines and even marker assisted selected lines.

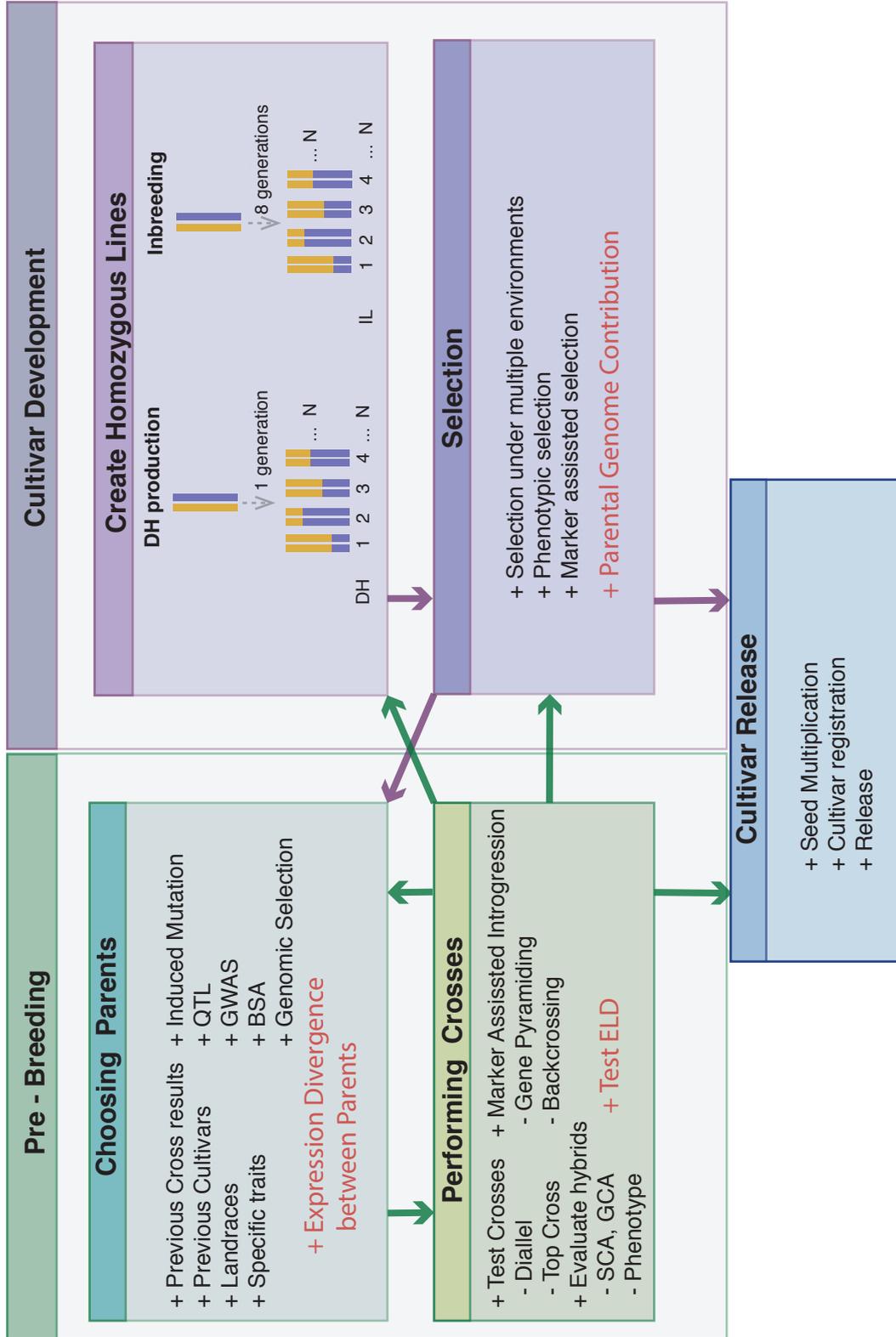


Figure 6.1: Simplified schematic of typical breeding strategies. Arrows indicate the movement of potential cultivars through the breeding program. Red processes indicate suggestions from this thesis.

6.1 Parental Regulome Divergence Governs Parental Combining Ability for mRNA Dosage Dependant Traits

Parental selection is arguably the most important step in plant breeding programs, because the genetic variability here determines the variability in the breeding program (Fasahat et al., 2016). However, current methods of choosing parental lines for breeding programs are inefficient and it has been suggested that more than 80% of resources are used on crosses that fail to produce useful varieties (Witcombe et al., 2013). This is because choosing parents that hybridise to produce the desired phenotype(s) is a non-trivial task. It is currently based on phenotypic, molecular and pedigree information obtained from large numbers of crosses and is very difficult to predict outcomes based on the information held from the parents alone, requiring these crosses to be performed (Bertan et al., 2007).

The parents are then scored according to their general combining ability (GCA) and specific combining ability (SCA) leading to the development of heterotic groups (Langridge and Fleury, 2011). These heterotic groups reduce the SCE to GCE ratios and provide efficiencies by reducing trial and error. However, creating heterotic groups takes a long time and requires a meticulous crossing strategy and so, many species do not benefit from these resources (Glenn et al., 2017). After the scoring of parents, the information is generally only applicable to the parents and if new parents are required, a new set of crosses are also required. Without knowledge of the underlying mechanisms behind combining ability it is difficult to predict for a larger or different populations. Another interesting caveat to the choosing of parents is the outcome required by the plant breeder, this is usually increased yield but could also be

enhanced resistance, secondary metabolites or species specific traits such as increased oil production (Jin et al., 2017). These different phenotypes are controlled by different genetic mechanisms and therefore SCA and GCA for a specific trait may not work for other traits.

Often, these phenotypes have complex underlying genetic causes decided by more than one locus and each locus could be contributing to the phenotype in a different way (Walley et al., 2012). In some cases, non-synonymous polymorphisms between individuals can result in altered protein function and differing phenotype (Papini-Terzi et al., 2003). However, many genes are also regulated in a dosage dependant manner with the cellular concentration of their protein dictating their efficiency (Shi et al., 2015). The major driving force of this is the transcriptional level of the gene which in turn can be controlled by many regulatory factors (Vogel and Marcotte, 2012). So, combining genomes such as in an F1 hybrid complicates this further with regulatory elements from both species coming together in one genome. In the extreme case of two identical lines being combined it would be expected that the gene expression levels of all genes would be the same in both species and therefore upon combining in the F1 the genes would maintain their expression levels. However, in the opposite case, where regulatory elements from the two parents have diverged, gene expression in the two parents would be different and it is this difference in the parents that can provide mRNA dosage novelty in the resulting hybrid.

To this end, we have found that the transcriptome and methylome of F1 hybrids have little ability to show transgressive patterns and only where regulatory factors differ between the parents can additive and non-additive expression occur. Many studies

try to correlate heterotic effects and genetic distance, but often strong environmental effects are found (Betrán et al., 2003) and in some cases there is no correlation at all, as is the case for Col X Ler hybrids of *Arabidopsis* (Groszmann et al., 2014). Our data suggests that in some cases, where there is a dosage dependence, it is not the genetic distance *per se* that controls the heterotic effects but the divergence of specific regulatory elements between the parents. This divergence causes transcriptional change in the F1 hybrid and in some cases alters phenotype. Future effort should be directed towards the understanding of transcriptomic and epigenetic involvement in combining ability. Only certain traits will be controlled by mRNA dosage. So understanding for which traits, differences in regulatory elements will provide variance in the offspring, is vital to moving forward with predictive plant breeding. I propose that a combinatorial model of genetic, transcriptomic and epigenetic influence is fundamental to understand how specific parental contributions underpin different traits.

6.2 Exploiting the Expression-level Dominance in Plant Breeding

Programs for mRNA Dosage Dependant Traits

Expression level dominance is a ubiquitous feature of genome mergers in diploid and polyploid hybrids having been discovered in a multitude of different species (Yoo et al., 2013; Hegarty et al., 2008; Chen, 2013). ELD results in the expression of protein coding genes in the hybrid being expressed at the level of only one of the parents. Our data shows that this phenomenon occurs also in this cross of *B. oleracea*. Considering these expression patterns are a category of non-additively expressed genes, these types of gene expression change are also governed by the regulatory divergence between the parental lines. However, the increased or decreased expression of specific genes

in hybrids can have phenotypic impacts in useful agronomic traits (Van Gioi et al., 2017).

Although many studies have identified this phenomenon, only a few have realised the agronomic importance of this. In polyploids, it is mainly used to explain biased fractionalisation through genome dominance (Renny-Byfield and Wendel, 2014; Otto, 2007). But in F1 hybrids it has been shown that the dominance of certain of genes can confer parental traits. In maize it has been shown that ELD dominance of certain genes can confer both nitrogen use efficiency and drought resistance (Van Gioi et al., 2017; Bi et al., 2014). In tobacco, it has also been shown that hybrid vigour of nicotine content can be affected by ELD (Tian et al., 2018). It is also true to say that the dominance of these genes is not parent specific but rather the specific combination of both of the parents that determines the dominance with different parental combinations producing different ELD signatures (Bottani et al., 2018; Van Gioi et al., 2017). This is a very interesting phenomenon which we have already been exploiting unknowingly; this is often why QTL studies are directed to regulatory parts of the genome (Jang et al., 2006) and could explain certain cases where combining ability is dependant on the environment (Samir et al., 2015). Because each gene has it's own regulatory landscape, there could be multiple causes of ELD and it could even be different for different genes. It has been shown that differing TE load and their proximity to genes in the parental genomes can explain ELD in polyploids through the spreading of silencing methylation into nearby genes (Cheng et al., 2016). It has also been suggested that differing parental genome sizes can affect the required binding efficiencies of TFs causing biased ELD, however here the genomes of the parents are of equal size and so, at least in this case, ELD can occur with similar genome sizes (Golicz et al., 2016;

Bottani et al., 2018). A mis-match of trans effectors and transcription factors could also be the cause (Bottani et al., 2018). However, an overarching theory for ELD has not been confirmed.

There are many traits that have been shown to be mRNA dosage dependant and so, by mining the literature and other QTL results it is possible to increase genetic gain by having a greater understanding of how some traits are conferred through ELD in the hybrid. Future effort should be directed at understanding whether this phenomena does have a singular explanation or whether this does in fact happen on a gene by gene basis. By understanding the mechanism underpinning ELD it could be possible to target specific ELD based traits, possibly with only the genome sequence of the parental line.

6.3 Utilising Parental Genome Contribution in Homozygous Line Selection

Growing DHs for phenotypic selection is another resource hungry process in plant breeding programs, often many hundreds of lines will be grown with detailed phenotypic analysis being performed, and this is only for one set of parental crosses (Langridge and Fleury, 2011). Even a MAB program will end up selecting many lines to grow that contain the desired marker or markers. Many of these lines will be discarded, with only the best being selected for further rounds of phenotypic testing in multiple environments. This growth and then discarding of lines wastes human and commercial resources and slows down breeding programs (Fasahat et al., 2016). Therefore any reduction in this waste of resources can allow them to be placed elsewhere, increasing genetic gain through faster selection (Glenn et al., 2017).

To this end, our data suggests that transcriptional stability in the DH lines is affected by the amount of genome inherited from either parent. With more transcriptional changes occurring on the parental genome of a DH line where the genome contribution is small. Further to this we find that the effect extends to CG methylation. It may also extend to CHG and CHH methylation changes but in our experiment the effects are confounded by environmental differences. Even in new genomic selection programs little thought is given to the genome contributions of the DH lines as most efforts focus on the selection of markers in the DH lines (Crossa et al., 2017). This could explain why introgression of desired traits does not always produce the desired phenotype and why some DHs containing the desired marker are more successful than others (Langridge and Fleury, 2011). This could provide a valuable model for predicting DH variability which could be exploited easily by plant breeders that already have genetic markers.

6.4 Future Directions

6.4.1 Parental Genome Contributions in Doubled haploid breeding

This study shows that the transcriptional and epigenetic stability of a DH line is, in part, governed by the amount of the parental genome that was inherited in that DH line. This means that in the case where an unequal amount of both parental genomes are inherited in a DH line, the parental genome that is underrepresented in the DH line would experience more relative change to the transcriptome and methylome. In this study we only surveyed nine DH lines from one cross and further to this, the nine DH lines in this study are lacking biological replicates. Therefore it is hard to conclude how widely applicable this phenomena is. DH lines are used extensively in

the commercial sector and little weight is given to the contribution of either parent to the DH lines. Often, in a molecular breeding program, the breeder only considers the inclusion of a specific genetic marker or set of genetic markers. Often, many hundreds of DH lines are created for extensive phenotypic selection. Then many of these DH lines that are selected for phenotypic evaluation are discarded because of poor performance. Any improvement in the selection efficiency of these lines can allow resources to be directed elsewhere, including more detailed phenotyping of the selected lines.

Future effort should be directed towards understanding how applicable the use of genomic contributions can be in DH plant breeding processes. By modelling the relationships of transcriptional and epigenetic change, parental genome contribution and phenotypic change in more detail it would be possible to build a model of estimated molecular change to the genome and the likelihood of this affecting the plant phenotype. This measure could then be used by plant breeders as additional means of selection.

Unfortunately it was not possible in this experiment to have a larger population due to seed availability and cost of sequencing. However, it would be possible to model the relationship more effectively with a larger population. The wider the range of DH genome contributions, the better the chances are of capturing the real relationships. The DH population would benefit from having a full range of genome contributions as well as having different DH lines with the same whole genome contributions but different individual chromosome make-ups from either parent.

Another caveat to the experiment in this thesis is the impossibility of separating the

genomic effects of the genome merger and the effects that the *in vitro* culture has on the genome. To this end, controls of the parental lines should also be created, these should be the parental lines that have been subjected to the same conditions as the DH lines. That is, microspores isolated from either parent, *in vitro* culture to form haploid plants and then chromosome doubling of the haploids. In this way, when they are compared to the parents and DH lines, changes arising from the culture itself can be seen in both parental genomes without the effects of the hybridisation and genome merger.

These samples: the parents, the parents after culture and many DH lines should be grown together in a controlled environment. From these samples, whole genome RNA-Seq and whole genome bisulphite data should be created. Then, using methods developed in this thesis, the genomes of the DH lines should be compared back to the parents and DEGs and DMRs from each of the inherited genomes should be found. From here it is a case of modelling the relative transcriptional and epigenetic change against the amount of genome inherited. From analysing the limited data here, it is possible that there will be an exponential decay relationship between the amount of the parental genome inherited and the amount of epigenetic and transcriptional change for that inherited genome. Further to this, one of the parental genomes will be more affected than the other genome.

Further to this, it would be interesting to see how this relationship changes with the genetic distance of the parents, it may be that this relationship is more prominent with a greater genetic distance between the parental lines. As discussed earlier, certain genes will have specific regulatory elements that, when disturbed, will cause changes

to gene expression and in some cases phenotype. Having an estimated transcriptional change based on the genome contribution could be very helpful and once this is understood in more detail could inform plant breeders during selection procedures.

6.4.2 *Studies of Genome Mergers*

The genome merger is a very interesting genomic phenomenon that already occurs in nature. However, studying this process can provide insights into genome regulation and currently this has only been addressed in F1 diploids and polyploids. The DH line is another form of this genome merger and this study shows that more interesting insights into genome regulation can come from the study of the genome mergers using DH lines. This study shows that F1 heterozygosity causes more extreme changes to the genome than that of the DH lines, this has also been shown in subsequent polyploids of F1 hybrids. However, these three types of merger have not been studied together.

The three forms of genome mergers, F1 diploid hybrids, polyploid hybrids and DH lines have different effects on the resulting offspring genomes. The F1 is a completely heterozygous genome structure and the resulting offspring genome relies on one allele from each parent and so there is a mismatch of regulatory elements. The polyploid has two sets of alleles from both parents and so the problems are more concerned with dosage than missing regulatory elements. Then the DH; in this scenario the genome is diploid and homozygous but contains a mosaic from both parental genomes. Therefore, there are missing regulatory elements but the diploid genome structure differs from the heterozygous F1.

It would therefore be interesting to analyse a population where these three types of genome mergers have been created from the same parents. Understanding how

different genes respond to these different genomic conditions can give insights into gene and genome regulation.

Bibliography

- Abdelrahman, M., Sawada, Y., Nakabayashi, R., Sato, S., Hirakawa, H., El-Sayed, M., Hirai, M. Y., Saito, K., Yamauchi, N. and Shigyo, M. (2015), 'Integrating transcriptome and target metabolome variability in doubled haploids of *Allium cepa* for abiotic stress protection', *Molecular Breeding* **35**(10), 195.
- Adams, K. L. and Wendel, J. F. (2005), 'Novel patterns of gene expression in polyploid plants', *Trends in Genetics* **21**(10), 539–543.
- Akalin, A., Kormaksson, M., Li, S., Garrett-Bakelman, F. E., Figueroa, M. E., Melnick, A. and Mason, C. E. (2012), 'MethylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles', *Genome Biology* **13**(10), R87.
- Anders, S., Pyl, P. T. and Huber, W. (2015), 'HTSeq-A Python framework to work with high-throughput sequencing data', *Bioinformatics* **31**(2), 166–169.
- Andrews, S. (2010), 'FastQC: A quality control tool for high throughput sequence data.', **[Online]**(Accessed: October 2016), <http://www.bioinformatics.babraham.ac.uk/projects/>.
- Bakhtiar, F., Afshari, F., Najafian, G. and Mohammadi, M. (2014), 'Backcross-

- breeding and doubled-haploid facilitated introgression of stripe rust resistance in bread wheat', *Archives of Phytopathology and Plant Protection* **47**(14), 1675–1685.
- Barber, W. T., Zhang, W., Win, H., Varala, K. K., Dorweiler, J. E., Hudson, M. E. and Moose, S. P. (2012), 'Repeat associated small RNAs vary among parents and following hybridization in maize', *Proceedings of the National Academy of Sciences* **109**(26), 10444–10449.
- Bertan, I., Carvalho, F. I. F. D. and Oliveira, A. A. C. D. (2007), 'Parental Selection Strategies in Plant Breeding Programs', *Journal of Crop Science and Biotechnology* **10**(4), 211–222.
- Betrán, F. J., Ribaut, J. M., Beck, D. and de León, D. G. (2003), 'Genetic Diversity, Specific Combining Ability, and Heterosis in Tropical Maize under Stress and Nonstress Environments', *Crop Science* **43**(3), 797.
- Bi, Y. M., Meyer, A., Downs, G. S., Shi, X., El-kereamy, A., Lukens, L. and Rothstein, S. J. (2014), 'High throughput RNA sequencing of a hybrid maize and its parents shows different mechanisms responsive to nitrogen limitation', *BMC Genomics* **15**(1), 1–12.
- Birchler, J. A., Auger, D. L. and Riddle, N. C. (2003), 'In search of the molecular basis of heterosis.', *The Plant Cell* **15**(10), 2236–9.
- Blakeslee, A. F., Belling, J., Farnham, M. E. and Bergner, A. D. (1922), 'A haploid mutant in the jimson weed, datura stramonium.', *Science* **55**(1433), 646–7.
- Bohuon, E. J. R., Keith, D. J., Parkin, I. A. P., Sharpe, A. G. and Lydiate, D. J. (1996),

- ‘Alignment of the conserved C genomes of *Brassica oleracea* and *Brassica napus*’, *Theoretical and Applied Genetics* **93-93**(5-6), 833–839.
- Bohuon, E. J. R., Ramsay, L. D., Craft, J. A., Arthur, A. E., Marshall, D. F., Lydiate, D. J. and Kearsley, M. J. (1998), ‘The Association of Flowering Time Quantitative Trait Loci with Duplicated Regions and Candidate Loci in *Brassica oleracea*’, *Genetics* **150**(1), 393–401.
- Bolger, A. M., Lohse, M. and Usadel, B. (2014), ‘Trimmomatic: A flexible trimmer for Illumina sequence data’, *Bioinformatics* **30**(15), 2114–2120.
- Botstein, D., White, R. L., Skolnick, M. and Davis, R. W. (1980), ‘Construction of a genetic linkage map in man using restriction fragment length polymorphisms.’, *American Journal of Human Genetics* **32**(3), 314–31.
- Bottani, S., Zabet, N. R., Wendel, J. F. and Veitia, R. A. (2018), ‘Gene Expression Dominance in Allopolyploids: Hypotheses and Models.’, *Trends in Plant Science* **23**(5), 393–402.
- Butruille, D. V., Birru, F. H., Boerboom, M. L., Cargill, E. J., Davis, D. A., Dhungana, P., Dill, G. M., Dong, F., Fonseca, A. E., Gardunia, B. W., Holland, G. J., Hong, N., Linnen, P., Nickson, T. E., Polavarapu, N., Pataky, J. K., Popi, J. and Stark, S. B. (2015), Maize Breeding in the United States: Views from Within Monsanto, *in* ‘Plant Breeding Reviews: Volume 39’, Vol. 39, John Wiley & Sons, Inc., Hoboken, New Jersey, pp. 199–282.
- Catoni, M., Tsang, J. M., Greco, A. P. and Zabet, N. (2018), ‘DMRcaller: a versatile

- R/Bioconductor package for detection and visualization of differentially methylated regions in CpG and non-CpG contexts', *Nucleic Acids Research* **46**(19), e114.
- Chen, K. and Rajewsky, N. (2007), 'The evolution of gene regulation by transcription factors and microRNAs', *Nature Reviews Genetics* **8**(2), 93–103.
- Chen, Z. J. (2013), 'Genomic and epigenetic insights into the molecular bases of heterosis', *Nature Reviews Genetics* **14**(7), 471–482.
- Cheng, F., Sun, C., Wu, J., Schnable, J., Woodhouse, M. R., Liang, J., Cai, C., Freeling, M. and Wang, X. (2016), 'Epigenetic regulation of subgenome dominance following whole genome triplication in *Brassica rapa*', *New Phytologist* **211**(1), 288–299.
- Chuong, P. V. and Beversdorf, W. D. (1985), 'High frequency embryogenesis through isolated microspore culture in *Brassica napus* L. and *B. Carinata* braun', *Plant Science* **39**(3), 219–226.
- Cogan, N., Harvey, E., Robinson, H., Lynn, J., Pink, D., Newbury, H. and Puddephat, I. (2001), 'The effects of anther culture and plant genetic background on *Agrobacterium rhizogenes*-mediated transformation of commercial cultivars and derived doubled-haploid *Brassica oleracea*', *Plant Cell Reports* **20**(8), 755–762.
- Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., Mcpherson, A., Szcześniak, W., Gaffney, D. J., Elo, L. L., Zhang, X. and Mortazavi, A. (2016), 'A survey of best practices for RNA-seq data analysis', *Genome Biology* **17**(1), 1–19.
- Crick, F. (1970), 'Central Dogma of Molecular Biology', *Nature* **227**(5258), 561–563.

- Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., de los Campos, G., Burgueño, J., González-Camacho, J. M., Pérez-Elizalde, S., Beyene, Y., Dreisigacker, S., Singh, R., Zhang, X., Gowda, M., Roorkiwal, M., Rutkoski, J. and Varshney, R. K. (2017), 'Genomic Selection in Plant Breeding: Methods, Models, and Perspectives', *Trends in Plant Science* **22**(11), 961–975.
- Davenport, C. B. (1908), 'Recent Advances in the Theory of Heredity', *Journal of Heredity* **4**(1), 355–357.
- Du, J., Johnson, L. M., Groth, M., Feng, S., Hale, C. J., Li, S., Vashisht, A. A., Gallego-Bartolome, J., Wohlschlegel, J. A., Patel, D. J. and Jacobsen, S. E. (2014), 'Mechanism of DNA methylation-directed histone methylation by KRYPTONITE', *Molecular Cell* **55**(3), 495–504.
- Duan, C.-G., Wang, X., Zhang, L., Xiong, X., Zhang, Z., Tang, K., Pan, L., Hsu, C.-C., Xu, H., Tao, W. A., Zhang, H. and Zhu, J.-K. (2017), 'A protein complex regulates RNA processing of intronic heterochromatin-containing genes in Arabidopsis', *Proceedings of the National Academy of Sciences* **114**(35), 7377–7384.
- Eathington, S. R., Crosbie, T. M., Edwards, M. D., Reiter, R. S., Bull, J. K., Reiter, R., Bull, J., Co, M., Lindbergh Blvd, N. and Coeur, C. (2007), 'Molecular Markers in a Commercial Breeding Program', *Crop Science* **47**(3), 154–163.
- Ebbs, M. L. (2006), 'Locus-Specific Control of DNA Methylation by the Arabidopsis SUVH5 Histone Methyltransferase', *The Plant Cell* **18**(5), 1166–1176.
- Eriksson, D. and Ammann, K. H. (2017), 'A Universally Acceptable View on the

- Adoption of Improved Plant Breeding Techniques’, *Frontiers in Plant Science* **7**(1), 1999.
- Fasahat, P., Rajabi Javad Mohseni Rad, A. and Derera, J. (2016), ‘Principles and Utilization of Combining Ability in Plant Breeding’, *Biometrics and Biostatistics International Journal* **4**(1), 1–24.
- Feng, S., Cokus, S. J., Zhang, X., Chen, P.-Y., Bostick, M., Goll, M. G., Hetzel, J., Jain, J., Strauss, S. H., Halpern, M. E., Ukomadu, C., Sadler, K. C., Pradhan, S., Pellegrini, M. and Jacobsen, S. E. (2010), ‘Conservation and divergence of methylation patterning in plants and animals’, *Proceedings of the National Academy of Sciences* **107**(19), 8689–8694.
- Ferrie, A. M. and Möllers, C. (2011), ‘Haploids and doubled haploids in Brassica spp. for genetic and genomic research’, *Plant Cell, Tissue and Organ Culture* **104**(3), 375–386.
- Ferrie, A. M. R. and Caswell, K. L. (2011), ‘Isolated microspore culture techniques and recent progress for haploid and doubled haploid plant production’, *Plant Cell, Tissue and Organ Culture* **104**(3), 301–309.
- Filiault, D. L., Seymour, D. K., Maruthachalam, R. and Maloof, J. N. (2017), ‘The Generation of Doubled Haploid Lines for QTL Mapping’, **1610**(1), 39–57.
- Fuda, N. J., Ardehali, M. B. and Lis, J. T. (2009), ‘Defining mechanisms that regulate RNA polymerase II transcription in vivo INSIGHT REVIEW’, *Nature* **461**(1), 186–192.

- Fujimoto, R., Taylor, J. M., Shirasawa, S., Peacock, W. J. and Dennis, E. S. (2012), 'Heterosis of Arabidopsis hybrids between C24 and Col is associated with increased photosynthesis capacity', *Proceedings of the National Academy of Sciences* **109**(18), 7109–7114.
- Gao, J., Yu, X., Ma, F. and Li, J. (2014), 'RNA-seq analysis of transcriptome and glucosinolate metabolism in seeds and sprouts of broccoli (*Brassica oleracea* var. *italica*)', *PLoS ONE* **9**(2), e88804.
- Gao, M., Li, G., Yang, B., Qiu, D., Farnham, M. and Quiros, C. (2007), 'High-density *Brassica oleracea* linkage map: Identification of useful new linkages', *Theoretical and Applied Genetics* **115**(2), 277–87.
- Glenn, K. C., Alsop, B., Bell, E., Goley, M., Jenkinson, J., Liu, B., Martin, C., Parrott, W., Souder, C., Sparks, O., Urquhart, W., Ward, J. M. and Vicini, J. L. (2017), 'Bringing New Plant Varieties to Market: Plant Breeding and Selection Practices Advance Beneficial Characteristics while Minimizing Unintended Changes', *Crop Science* **57**(6), 2906.
- Golicz, A. A., Bayer, P. E., Barker, G. C., Edger, P. P., Kim, H., Martinez, P. A., Kit, C., Chan, K., Severn-ellis, A., McCombie, W. R., Parkin, I. A. P., Paterson, A. H., Pires, J. C., Sharpe, A. G., Tang, H., Teakle, G. R., Town, C. D., Batley, J. and Edwards, D. (2016), 'The pangenome of an agronomically important crop plant *Brassica oleracea*', *Nature Communications* **7**(1), 1–8.
- Greaves, I. K., Eichten, S. R., Groszmann, M., Wang, A., Ying, H., Peacock, W. J. and Dennis, E. S. (2016), 'Twenty-four nucleotide siRNAs produce heritable trans-

- chromosomal methylation in F1 Arabidopsis hybrids', *Proceedings of the National Academy of Sciences* **113**(44), 6895–6902.
- Greaves, I. K., Gonzalez-Bayon, R., Wang, L., Zhu, A., Liu, P.-C., Groszmann, M., Peacock, W. J. and Dennis, E. S. (2015), 'Epigenetic Changes in Hybrids', *Plant Physiology* **168**(4), 1197–1205.
- Greaves, I. K., Groszmann, M., Wang, A., Peacock, W. J. and Dennis, E. S. (2014), 'Inheritance of Trans Chromosomal Methylation patterns from Arabidopsis F1 hybrids', *Proceedings of the National Academy of Sciences* **111**(5), 2017–2022.
- Greaves, I. K., Groszmann, M., Ying, H., Taylor, J. M., Peacock, W. J. and Dennis, E. S. (2012), 'Trans Chromosomal Methylation in Arabidopsis hybrids', *Proceedings of the National Academy of Sciences* **109**(9), 3570–3575.
- Groszmann, M., Gonzalez-Bayon, R., Greaves, I. K., Wang, L., Huen, A. K., Peacock, W. J. and Dennis, E. S. (2014), 'Intraspecific Arabidopsis hybrids show different patterns of heterosis despite the close relatedness of the parental genomes.', *Plant physiology* **166**(1), 265–80.
- Groszmann, M., Gonzalez-Bayon, R., Lyons, R. L., Greaves, I. K., Kazan, K., Peacock, W. J. and Dennis, E. S. (2015), 'Hormone-regulated defense and stress response networks contribute to heterosis in Arabidopsis F1 hybrids', *Proceedings of the National Academy of Sciences* **112**(46), 6397–6406.
- Grover, C. E., Gallagher, J. P., Szadkowski, E. P., Yoo, M. J., Flagel, L. E. and Wendel, J. F. (2012), 'Homoeolog expression bias and expression level dominance in allopolyploids', *New Phytologist* **196**(4), 966–971.

- Guha, S. and Maheshwari, S. C. (1964), 'In vitro Production of Embryos from Anthers of *Datura*', *Nature* **204**(4957), 497–497.
- Guo, M., Rupe, M. A., Yang, X., Crasta, O., Zinselmeier, C., Smith, O. S. and Bowen, B. (2006), 'Genome-wide transcript analysis of maize hybrids: allelic additive gene expression and yield heterosis', *Theoretical and Applied Genetics* **113**(5), 831–845.
- He, G., Zhu, X., Elling, A. A., Chen, L., Wang, X., Guo, L., Liang, M., He, H., Zhang, H., Chen, F., Qi, Y., Chen, R. and Deng, X.-W. (2010), 'Global Epigenetic and Transcriptional Trends among Two Rice Subspecies and Their Reciprocal Hybrids', *The Plant Cell* **22**(1), 17–33.
- Hegarty, M. J., Barker, G. L., Brennan, A. C., Edwards, K. J., Abbott, R. J. and Hiscock, S. J. (2008), 'Changes to gene expression associated with hybrid speciation in plants: further insights from transcriptomic studies in *Senecio*', *Philosophical Transactions of the Royal Society B: Biological Sciences* **363**(1506), 3055–3069.
- Hickey, J. M., Chiurugwi, T., Mackay, I., Powell, W., Hickey, J. M., Chiurugwi, T., Mackay, I., Powell, W., Eggen, A., Kilian, A., Jones, C., Canales, C., Grattapaglia, D., Bassi, F., Atlin, G., Gorjanc, G., Dawson, I., Rabbi, I., Ribaut, J.-M., Rutkoski, J., Benzie, J., Lightner, J., Mwacharo, J., Parmentier, J., Robbins, K., Skot, L., Wolfe, M., Rouard, M., Clark, M., Amer, P., Gardiner, P., Hendre, P., Mrode, R., Sivasankar, S., Rasmussen, S., Groh, S., Jackson, V., Thomas, W., Beyene, Y. and Beyene, Y. (2017), 'Genomic prediction unifies animal and plant breeding programs to form platforms for biological discovery', *Nature Genetics* **49**(9), 1297–1303.
- Higdon, J., Delage, B., Williams, D. and Dashwood, R. (2007), 'Cruciferous veg-

- etables and human cancer risk: epidemiologic evidence and mechanistic basis', *Pharmacological Research* **55**(3), 224–236.
- Hofmeister, B. T., Lee, K., Rohr, N. A., Hall, D. W. and Schmitz, R. J. (2017), 'Stable inheritance of DNA methylation allows creation of epigenotype maps and the study of epiallele inheritance patterns in the absence of genetic variation', *Genome Biology* **18**(1), 1–16.
- Hu, X., Wang, H., Diao, X., Liu, Z., Li, K., Wu, Y., Liang, Q., Wang, H. and Huang, C. (2016), 'Transcriptome profiling and comparison of maize ear heterosis during the spikelet and floret differentiation stages', *BMC Genomics* **17**(1), 1–18.
- Jackson, S. and Chen, Z. J. (2010), 'Genomic and expression plasticity of polyploidy', *Current Opinion in Plant Biology* **13**(2), 153–159.
- Jang, C. S., Kamps, T. L., Skinner, D. N., Schulze, S. R., Vencill, W. K. and Paterson, A. H. (2006), 'Functional Classification, Genomic Organization, Putatively cis-Acting Regulatory Elements, and Relationship to Quantitative Trait Loci, of Sorghum Genes with Rhizome-Enriched Expression', *Plant Physiology* .
- Jin, J., Sun, Y., Qu, J., Syah, R., Lim, C. H., Alfiko, Y., Rahman, N. E. B., Suwanto, A., Yue, G., Wong, L., Chua, N. H. and Ye, J. (2017), 'Transcriptome and functional analysis reveals hybrid vigor for oil biosynthesis in oil palm', *Scientific Reports* **7**(1), 1–12.
- Jung, H.-J., Dong, X., Park, J.-I., Thamilarasan, S. K., Lee, S. S., Kim, Y.-K., Lim, Y.-P., Nou, I.-S. and Hur, Y. (2014), 'Genome-Wide Transcriptome Analysis of

- Two Contrasting *Brassica rapa* Doubled Haploid Lines under Cold-Stresses Using Br135K Oligomeric Chip', *PLoS ONE* **9**(8), e106069.
- Juven-Gershon, T., Hsu, J.-Y., Theisen, J. W. and Kadonaga, J. T. (2008), 'The RNA polymerase II core promoter the gateway to transcription', *Current Opinion in Cell Biology* **20**(3), 253–259.
- Kato, M., Miura, A., Bender, J., Jacobsen, S. E. and Kakutani, T. (2003), 'Role of CG and non-CG methylation in immobilization of transposons in Arabidopsis.', *Current Biology* **13**(5), 421–6.
- Kawakatsu, T., Huang, S.-s. C., Jupe, F., Sasaki, E., Schmitz, R. J., Urich, M. A., Castanon, R., Nery, J. R., Barragan, C., He, Y., Chen, H., Dubin, M., Lee, C.-R., Wang, C., Bemm, F., Becker, C., O'Neil, R., O'Malley, R. C., Quarless, D. X., Schork, N. J., Weigel, D., Nordborg, M., Ecker, J. R., Alonso-Blanco, C., Andrade, J., Becker, C., Bemm, F., Bergelson, J., Borgwardt, K., Chae, E., Dezwaan, T., Ding, W., Ecker, J. R., Expósito-Alonso, M., Farlow, A., Fitz, J., Gan, X., Grimm, D. G., Hancock, A., Henz, S. R., Holm, S., Horton, M., Jarsulic, M., Kerstetter, R. A., Korte, A., Korte, P., Lanz, C., Lee, C.-R., Meng, D., Michael, T. P., Mott, R., Mulyati, N. W., Nägele, T., Nagler, M., Nizhynska, V., Nordborg, M., Novikova, P., Picó, F. X., Platzer, A., Rabanal, F. A., Rodriguez, A., Rowan, B. A., Salomé, P. A., Schmid, K., Schmitz, R. J., Seren, Ü., Sperone, F. G., Sudkamp, M., Svardal, H., Tanzer, M. M., Todd, D., Volchenboum, S. L., Wang, C., Wang, G., Wang, X., Weckwerth, W., Weigel, D. and Zhou, X. (2016), 'Epigenomic Diversity in a Global Collection of *Arabidopsis thaliana* Accessions', *Cell* **166**(2), 492–505.
- Kawanabe, T., Ishikura, S., Miyaji, N., Sasaki, T., Wu, L. M., Itabashi, E., Takada,

- S., Shimizu, M., Takasaki-Yasuda, T., Osabe, K., Peacock, W. J., Dennis, E. S. and Fujimoto, R. (2016), 'Role of DNA methylation in hybrid vigor in *Arabidopsis thaliana*', *Proceedings of the National Academy of Sciences* **113**(43), 6704–6711.
- Kays, S. J. and Dias, J. C. S. (1995), 'Common names of commercially cultivated vegetables of the world in 15 languages', *Economic Botany* **49**(2), 115–152.
- Kazan, K. and Manners, J. M. (2013), 'MYC2: The Master in Action', *Molecular Plant* **6**(3), 686–703.
- Khush, G. S. (2001), 'Green revolution: the way forward', *Nature Reviews Genetics* **2**(10), 815–822.
- Kirkbride, R. C., Yu, H. H., Nah, G., Zhang, C., Shi, X. and Chen, Z. J. (2015), 'An Epigenetic Role for Disrupted Paternal Gene Expression in Postzygotic Seed Abortion in *Arabidopsis* Interspecific Hybrids', *Molecular Plant* **8**(12), 1766–1775.
- Kopylova, E., Noé, L. and Touzet, H. (2012), 'SortMeRNA: Fast and accurate filtering of ribosomal RNAs in metatranscriptomic data', *Bioinformatics* **28**(24), 3211–3217.
- Koussounadis, A., Langdon, S. P., Um, I. H., Harrison, D. J. and Smith, V. A. (2015), 'Relationship between differentially expressed mRNA and mRNA-protein correlations in a xenograft model system', *Scientific Reports* **5**(1), 1–9.
- Krueger, F. and Andrews, S. R. (2011), 'Bismark: A flexible aligner and methylation caller for Bisulfite-Seq applications', *Bioinformatics* **27**(11), 1571–1572.
- Lang, Z., Wang, Y., Tang, K., Tang, D., Datsenka, T., Cheng, J., Zhang, Y., Handa,

- A. K. and Zhu, J.-K. (2017), 'Critical roles of DNA demethylation in the activation of ripening-induced genes and inhibition of ripening-repressed genes in tomato fruit', *Proceedings of the National Academy of Sciences* **114**(1), 4511–4519.
- Langridge, P. and Fleury, D. (2011), 'Making the most of omics' for crop breeding', *Trends in Biotechnology* **29**(1), 33–40.
- Lanner-Herrera, C., Gustafson, M., Filt, A. S. and Bryngelsson, T. (1996), 'Diversity in natural populations of wild Brassica oleracea as estimated by isozyme and RAPD analysis', *Genetic Resources and Crop Evolution* **43**(1), 13–23.
- Lauss, K., Wardenaar, R., Oka, R., van Hulten, M. H., Guryev, V., Keurentjes, J. J., Stam, M. and Johannes, F. (2018), 'Parental DNA methylation states are associated with heterosis in epigenetic hybrids', *Plant Physiology* **176**(2), 1627–1645.
- Law, J. A. and Jacobsen, S. E. (2011), 'Establishing, maintaining and modifying DNA methylation patterns in plants and animals', *Nature Reviews Genetics* **11**(3), 204–220.
- Lei, M., Zhang, H., Julian, R., Tang, K., Xie, S. and Zhu, J.-K. (2015), 'Regulatory link between DNA methylation and active demethylation in Arabidopsis', *Proceedings of the National Academy of Sciences* **112**(11), 3553–3557.
- Li, X., Ramchiary, P., Dhandapani, P., Choi, P. and Lim, Y. (2013), Omics Applications in Brassica Species, in 'OMICS Applications in Crop Science', CRC Press, pp. 163–190.
- Li, Y., Varala, K., Moose, S. P. and Hudson, M. E. (2012), 'The Inheritance Pattern of

- 24 nt siRNA Clusters in Arabidopsis Hybrids Is Influenced by Proximity to Transposable Elements', *PLoS ONE* **7**(10), e47043.
- Love, M. I., Anders, S. and Huber, W. (2014), 'Differential analysis of count data - the DESeq2 package', *Genome Biology* **15**(12), 550.
- Machczyńska, J., Orłowska, R., Mańkowski, D. R., Zimny, J. and Bednarek, P. T. (2014), 'DNA methylation changes in triticale due to in vitro culture plant regeneration and consecutive reproduction', *Plant Cell, Tissue and Organ Culture* **119**(2), 289–299.
- Maere, S., Heymans, K. and Kuiper, M. (2005), 'BiNGO: A Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks', *Bioinformatics* **21**(16), 3448–3449.
- McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M. and DePristo, M. A. (2010), 'The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data', *Genome Research* **20**(1), 1297–1303.
- Mirouze, M., Reinders, J., Bucher, E., Nishimura, T., Schneeberger, K., Ossowski, S., Cao, J., Weigel, D., Paszkowski, J. and Mathieu, O. (2009), 'Selective epigenetic control of retrotransposition in Arabidopsis', *Nature* **461**(7262), 427–430.
- Moose, S. P. and Mumm, R. H. (2008), 'Molecular Plant Breeding as the Foundation for 21st Century Crop Improvement 1', **147**(1), 969–977.
- Mullis, K., Faloona, F., Scharf, S., Saiki, R., Horn, G. and Erlich, H. (1986), 'Specific

- enzymatic amplification of DNA in vitro: the polymerase chain reaction.’, *Cold Spring Harbor Symposia on Quantitative Biology* **51 Pt 1**, 263–73.
- Murovec, J. and Bohanec, B. (2012), *Haploids and Doubled Haploids in Plant Breeding*, Vol. 39, inTech.
- Nagaharu, U. and Nagaharu, N. (1935), ‘Genome analysis in Brassica with special reference to the experimental formation of *B. napus* and peculiar mode of fertilisation’, *Journal of Japanese Botany* **8**(1), 389–452.
- Neri, F., Rapelli, S., Krepelova, A., Incarnato, D., Parlato, C., Basile, G., Maldotti, M., Anselmi, F. and Oliviero, S. (2017), ‘Intragenic DNA methylation prevents spurious transcription initiation’, *Nature* **543**(7643), 72–77.
- Ng, D. W.-K., Lu, J. and Chen, Z. J. (2012), ‘Big roles for small RNAs in polyploidy, hybrid vigor, and hybrid incompatibility’, *Current Opinion in Plant Biology* **15**(2), 154–161.
- Niederhuth, C. E. and Schmitz, R. J. (2017), ‘Biochimica et Biophysica Acta Putting DNA methylation in context : from genomes to gene expression in plants ’, *Gene Regulatory Mechanisms* **1860**(1), 149–156.
- 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., Malike, 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 Brunt, A., Wischmeyer, C., Beil, M., Hogan, M., Lakey, N., Lim, C.-C., Arulandoo, 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**(7570), 533–7.
- Otto, S. P. (2007), 'The Evolutionary Consequences of Polyploidy', *Cell* **131**(1), 452–462.
- Papini-Terzi, F. S., Galhardo, R. D. S., Farias, L. P., Menck, C. F. M. and Van Sluys, M. A. (2003), 'Point mutation is responsible for Arabidopsis tz-201 mutant phenotype affecting thiamin biosynthesis', *Plant and Cell Physiology* **44**(8), 856–860.
- Parkin, I. A., Koh, C., Tang, H., Robinson, S. J., Kagale, S., Clarke, W. E., Town, C. D., Nixon, J., Krishnakumar, V., Bidwell, S. L., Denoeud, F., Belcram, H., Links, M. G., Just, J., Clarke, C., Bender, T., Huebert, T., Mason, A. S., Chris Pires, J., Barker, G., Moore, J., Walley, P. G., Manoli, S., Batley, J., Edwards, D., Nelson, M. N., Wang, X., Paterson, A. H., King, G., Bancroft, I., Chalhoub, B. and Sharpe, A. G. (2014), 'Transcriptome and methylome profiling reveals relics of genome dominance in the mesopolyploid Brassica oleracea', *Genome Biology* **15**(6), 1–18.
- Powers, L. (1944), 'An Expansion of Jones's Theory for the Explanation of Heterosis', *The American Naturalist* **78**(776), 275–280.
- Qi, B., Huang, W., Zhu, B., Zhong, X., Guo, J., Zhao, N., Xu, C., Zhang, H., Pang, J., Han, F. and Liu, B. (2012), 'Global transgenerational gene expression dynamics in two newly synthesized allohexaploid wheat (*Triticum aestivum*) lines', *BMC Biology* **10**(1), 3.

- Rapp, R. A., Udall, J. A. and Wendel, J. F. (2009), 'Genomic expression dominance in allopolyploids', *BMC Biology* **7**(1), 18.
- Renny-Byfield, S. and Wendel, J. F. (2014), 'Doubling down on genomes: Polyploidy and crop plants', *American Journal of Botany* **101**(10), 1711–1725.
- Rigal, M., Becker, C., Pélissier, T., Pogorelcnik, R., Devos, J., Ikeda, Y., Weigel, D. and Mathieu, O. (2016), 'Epigenome confrontation triggers immediate reprogramming of DNA methylation and transposon silencing in *Arabidopsis thaliana* F1 epihybrids', *Proceedings of the National Academy of Sciences* **113**(14), 2083–2092.
- Samir, P., Rahul, Slaughter, J. C. and Link, A. J. (2015), 'Environmental interactions and epistasis are revealed in the proteomic responses to complex stimuli', *PLoS ONE* **10**(8), e0134099.
- Schlegel, R. H. J. (2018), *History of Plant Breeding*, 1 edn, CRC Press.
- Sebastian, R. L., Howell, E. C., King, G. J., Marshall, D. F. and Kearsey, M. J. (2000), 'An integrated AFLP and RFLP *Brassica oleracea* linkage map from two morphologically distinct doubled-haploid mapping populations', *TAG Theoretical and Applied Genetics* **100**(1), 75–81.
- Seifert, F., Bössow, S., Kumlehn, J., Gnad, H. and Scholten, S. (2016), 'Analysis of wheat microspore embryogenesis induction by transcriptome and small RNA sequencing using the highly responsive cultivar Svilena', *BMC Plant Biology* **16**(1), 97.

- Shen, H., He, H., Li, J., Chen, W., Wang, X., Guo, L., Peng, Z., He, G., Zhong, S., Qi, Y., Terzaghi, W. and Deng, X. W. (2012), ‘Genome-Wide Analysis of DNA Methylation and Gene Expression Changes in Two Arabidopsis Ecotypes and Their Reciprocal Hybrids’, *The Plant Cell* **24**(3), 875–892.
- Shi, X., Zhang, C., Ko, D. K. and Chen, Z. J. (2015), ‘Genome-Wide Dosage-Dependent and -Independent Regulation Contributes to Gene Expression and Evolutionary Novelty in Plant Polyploids’, *Molecular Biology and Evolution* **32**(9), 2351–2366.
- Shimelis, H. and Laing, M. (2012), ‘Timelines in conventional crop improvement: pre-breeding and breeding procedures’, *American Journal of Crop Science* **6**(11), 1542–1549.
- Shivaprasad, P. V., Dunn, R. M., Santos, B. A., Bassett, A. and Baulcombe, D. C. (2012), ‘Extraordinary transgressive phenotypes of hybrid tomato are influenced by epigenetics and small silencing RNAs’, *The EMBO Journal* **31**(2), 257–266.
- Sigman, M. J. and Slotkin, R. K. (2016), ‘The First Rule of Plant Transposable Element Silencing : Location , Location , Location’, *Plant Cell* **28**(2), 304–313.
- Solís, M.-T., El-Tantawy, A.-A., Cano, V., Risueño, M. C. and Testillano, P. S. (2015), ‘5-azacytidine promotes microspore embryogenesis initiation by decreasing global DNA methylation, but prevents subsequent embryo development in rapeseed and barley’, *Frontiers in Plant Science* **6**(1), e1–16.
- Song, Q. and Chen, Z. J. (2015), ‘Epigenetic and developmental regulation in plant polyploids’, *Current Opinion in Plant Biology* **24**(1), 101–109.

- Stupar, R. M., Gardiner, J. M., Oldre, A. G., Haun, W. J., Chandler, V. L. and Springer, N. M. (2008), 'Gene expression analyses in maize inbreds and hybrids with varying levels of heterosis', *BMC Plant Biology* **8**(1), 1–19.
- Stupar, R. M., Hermanson, P. J. and Springer, N. M. (2007), 'Nonadditive expression and parent-of-origin effects identified by microarray and allele-specific expression profiling of maize endosperm.', *Plant Physiology* **145**(2), 411–25.
- Swanson-Wagner, R. A., Jia, Y., DeCook, R., Borsuk, L. A., Nettleton, D. and Schnable, P. S. (2006), 'All possible modes of gene action are observed in a global comparison of gene expression in a maize F1 hybrid and its inbred parents', *Proceedings of the National Academy of Sciences* **103**(18), 6805–6810.
- Takagi, H., Tamiru, M., Abe, A., Yoshida, K., Uemura, A., Yaegashi, H., Obara, T., Oikawa, K., Utsushi, H., Kanzaki, E., Mitsuoka, C., Natsume, S., Kosugi, S., Kanzaki, H., Matsumura, H., Urasaki, N., Kamoun, S. and Terauchi, R. (2015), 'MutMap accelerates breeding of a salt-tolerant rice cultivar', *Nature Biotechnol* **33**(1), 445–449.
- Tarazona, S., Garcia-Alcalde, F., Dopazo, J., Ferrer, A. and Conesa, A. (2011), 'Differential expression in RNA-seq: A matter of depth', *Genome Research* **21**(12), 2213–2223.
- Thomas, W. T. B., Forster, B. P. and Gertsson, B. (2003), Doubled haploids in breeding, in 'Doubled Haploid Production in Crop Plants', Springer Netherlands, Dordrecht, pp. 337–349.
- Tian, M., Nie, Q., Li, Z., Zhang, J., Liu, Y., Long, Y., Wang, Z., Wang, G. and Liu,

- R. (2018), 'Transcriptomic analysis reveals overdominance playing a critical role in nicotine heterosis in *Nicotiana tabacum* L', *BMC Plant Biology* **18**(1), 48.
- Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., Pimentel, H., Salzberg, S. L., Rinn, J. L. and Pachter, L. (2012), 'Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks', *Nature Protocols* **7**(3), 562–578.
- Van Bel, M., Diels, T., Vancaester, E., Kreft, L., Botzki, A., Van De Peer, Y., Coppens, F. and Vandepoele, K. (2018), 'PLAZA 4.0: An integrative resource for functional, evolutionary and comparative plant genomics', *Nucleic Acids Research* **46**(1), D1190–D1196.
- Van Gioi, H., Mallikarjuna, M. G., Shikha, M., Pooja, B., Jha, S. K., Dash, P. K., Basappa, A. M., Gadag, R. N., Rao, A. R. and Nepolean, T. (2017), 'Variable Level of Dominance of Candidate Genes Controlling Drought Functional Traits in Maize Hybrids.', *Frontiers in Plant Science* **8**(1), 940.
- Vihervaara, A., Duarte, F. M. and Lis, J. T. (2018), 'Molecular mechanisms driving transcriptional stress responses', *Nature Reviews Genetics* **19**(6), 385–397.
- Vogel, C. and Marcotte, E. M. (2012), 'Insights into the regulation of protein abundance from proteomic and transcriptomic analyses', *Nature Reviews Genetics* **13**(4), 227–232.
- Walley, P. G., Teakle, G. R., Moore, J. D., Allender, C. J., a.C. Pink, D., Buchanan-Wollaston, V. and Barker, G. C. (2012), 'Developing genetic resources for pre-

- breeding in *Brassica oleracea* L.: an overview of the UK perspective', *Journal of Plant Biotechnology* **29**(1), 62–68.
- Wang, J., Tian, L., Lee, H.-S., Wei, N. E., Jiang, H., Watson, B., Madlung, A., Osborn, T. C., Doerge, R. W., Comai, L. and Chen, Z. J. (2005), 'Genomewide Nonadditive Gene Regulation in *Arabidopsis* Allotetraploids', *Genetics* **172**(1), 507–517.
- Wang, Z., Gerstein, M. and Snyder, M. (2009), 'RNA-Seq: a revolutionary tool for transcriptomics.', *Nature Reviews Genetics* **10**(1), 57–63.
- Wibowo, A., Becker, C., Durr, J., Price, J., Spaepen, S., Hilton, S. and Putra, H. (2018), 'Partial maintenance of organ-specific epigenetic marks during plant asexual reproduction leads to heritable phenotypic variation', *Proceedings of the National Academy of Sciences* **115**(1), 9145–9152.
- Wibowo, A., Becker, C., Marconi, G., Durr, J., Price, J., Hagmann, J., Papareddy, R., Putra, H., Kageyama, J., Becker, J., Weigel, D. and Gutierrez-Marcos, J. (2016), 'Hyperosmotic stress memory in *Arabidopsis* is mediated by distinct epigenetically labile sites in the genome and is restricted in the male germline by DNA glycosylase activity', *eLife* **5**(1), 1–27.
- Witcombe, J. R., Gyawali, S., Subedi, M., Virk, D. S. and Joshi, K. D. (2013), 'Plant breeding can be made more efficient by having fewer, better crosses', *BMC Plant Biology* **13**(1), 1.
- Witzel, K., Stützel, H., Alvarez, J. B., Hena, A., Kamal, M., Baldermann, S., Neugart, S., Ruppel, S., Schreiner, M. and Wiesner, M. (2015), 'Recent progress in the

- use of 'omics technologies in brassicaceous vegetables', *Frontiers in Plant Science* **14**(1), 6–244.
- Yan, G., Liu, H., Wang, H., Lu, Z., Wang, Y., Mullan, D., Hamblin, J. and Liu, C. (2017), 'Accelerated Generation of Selfed Pure Line Plants for Gene Identification and Crop Breeding', *Frontiers in Plant Science* **8**(1), 1786.
- Yoo, M. J., Szadkowski, E. and Wendel, J. F. (2013), 'Homoeolog expression bias and expression level dominance in allopolyploid cotton', *Heredity* **110**(2), 171–180.
- Zhang, H., Lang, Z. and Zhu, J.-K. (2018), 'Dynamics and function of DNA methylation in plants', *Nature Reviews Molecular Cell Biology* **19**(8), 489–506.
- Zhang, Q., Wang, D., Lang, Z., He, L., Yang, L., Zeng, L., Li, Y., Zhao, C., Huang, H., Zhang, H., Zhang, H. and Zhu, J.-K. (2016), 'Methylation interactions in Arabidopsis hybrids require RNA-directed DNA methylation and are influenced by genetic variation', *Proceedings of the National Academy of Sciences* **113**(29), 4248–4256.
- Zhang, X., Yazaki, J., Sundaresan, A., Cokus, S., Chan, S. W.-L., Chen, H., Henderson, I. R., Shinn, P., Pellegrini, M., Jacobsen, S. E. and Ecker, J. (2006), 'Genome-wide High-Resolution Mapping and Functional Analysis of DNA Methylation in Arabidopsis', *Cell* **126**(6), 1189–1201.
- Zhou, M., Palanca, A. M. S. and Law, J. A. (2018), 'Locus-specific control of the de novo DNA methylation pathway in Arabidopsis by the CLASSY family', *Nature Genetics* **50**(6), 865–873.
- Zhu, H., Wang, G. and Qian, J. (2016), 'Transcription factors as readers and effectors of DNA methylation', *Nature Reviews Genetics* **17**(9), 551–565.

Ziller, M. J., Hansen, K. D., Meissner, A. and Aryee, M. J. (2015), 'Coverage recommendations for methylation analysis by whole-genome bisulfite sequencing', *Nature Methods* **12**(3), 230–232.

Appendices

Appendix A

Read Processing Numbers

Table A.1: Table displaying the raw read numbers and the number after processing of the reads and alignment to the reference genomes. RNASeq libraries - Original old batch.

Line	Generation	Biological replicate	Technical Replicate	Raw #	Mapped #	Mapped %
1003	1	1	1	19,485,465	7,070,787	36.3
1003	1	1	2	20,045,221	8,307,178	41.4
1003	1	1	3	20,592,231	8,535,057	41.4
1047	1	1	1	19,804,332	9,874,589	49.9
1047	1	1	2	18,517,430	7,795,277	42.1
1047	1	1	3	22,337,535	9,582,890	42.9
1047	3	1	1	20,024,367	9,119,477	45.5
1047	3	1	2	17,093,002	7,595,148	44.4
1047	3	1	3	26,353,015	9,972,018	37.8
3238	1	1	1	15,723,491	6,675,917	42.5
3238	1	1	2	18,291,895	7,310,723	40.0
3238	1	1	3	14,697,614	5,767,028	39.2
3238	3	1	1	29,353,738	10,939,706	37.3
3238	3	1	2	17,391,307	6,315,011	36.3
3238	3	1	3	17,890,302	6,705,101	37.5
5147	1	1	1	18,534,713	7,721,070	41.7
5147	1	1	2	20,632,881	7,531,542	36.5
5147	1	1	3	17,708,251	6,783,923	38.3
A12Dhd	NA	1	1	12,223,378	5,675,928	46.4
A12Dhd	NA	1	2	20,116,959	12,117,454	60.2
A12Dhd	NA	1	3	18,331,032	10,049,594	54.8
F1	NA	1	1	18,293,910	7,503,724	41.0
F1	NA	1	2	23,210,740	8,878,928	38.3
F1	NA	1	3	19,544,238	8,134,939	41.6
F1	NA	2	1	22,486,649	10,368,657	46.1
F1	NA	2	2	13,573,655	6,226,319	45.9
F1	NA	2	3	13,469,501	5,758,625	42.8
GDDH33	NA	1	1	17,352,959	11,282,020	65.0
GDDH33	NA	1	2	24,816,004	8,717,480	35.1

Table A.2: Table displaying the raw read numbers and the number after processing of the reads and alignment to the reference genomes. RNASeq libraries - New batch.

Line	Generation	Biologica		Raw #	Mapped #	Mapped %
		1 Replicate	Technical Replicate			
2069	1	1	1	35442575	17013019	48.0
2069	1	1	2	31588577	22230144	70.4
2069	3	1	1	24548432	11873245	48.4
2069	3	1	2	28028907	13921540	49.7
2134	1	1	1	43968637	24099674	54.8
2134	1	1	2	35766746	17847998	49.9
3013	1	1	1	41953635	24354035	58.0
3013	1	1	2	36235605	20122688	55.5
3088	1	1	1	43437120	21385975	49.2
3088	1	1	2	42260859	22126652	52.4
3088	3	1	1	41254728	19749624	47.9
3088	3	1	2	37294777	17473873	46.9
5071	1	1	1	34995520	18393598	52.6
5071	1	1	2	36177880	19297895	53.3
5119	1	1	1	44315624	23045561	52.0
5119	1	1	2	34245296	17278144	50.5
A12	NA	1	1	23596186	13808836	58.5
A12	NA	1	2	33200332	20007036	60.3
GDH33	NA	1	1	43045101	36865720	85.6
GDH33	NA	1	2	45541926	20016104	44.0

Table A.3: Table displaying the raw read numbers and the number after processing of the reads and alignment to the reference genomes. BS-Seq libraries - Original old batch.

Line	Generation	Pair	Raw #	Mapped /		
				Deduplicated #	Mapped %	non-con
1003	1	R1	71,020,689	41,804,163.00	58.86	0.00
1003	1	R2	71,020,689	41,804,163.00	58.86	0.00
1047	1	R1	72,027,542	50,113,508.00	69.58	0.00
1047	1	R2	72,027,542	50,113,508.00	69.58	0.00
1047	3	R1	65,334,820	38,854,860.00	59.47	0.00
1047	3	R2	65,334,820	38,854,860.00	59.47	0.00
3013	1	R1	81,837,092	50,242,548.00	61.39	0.00
3013	1	R2	81,837,092	50,242,548.00	61.39	0.00
3238	1	R1	77,002,939	47,376,609.00	61.53	0.00
3238	1	R2	77,002,939	47,376,609.00	61.53	0.00
3238	3	R1	56,276,181	39,403,975.00	70.02	0.00
3238	3	R2	56,276,181	39,403,975.00	70.02	0.00
5147	1	R1	74,724,157	44,050,595.00	58.95	0.00
5147	1	R2	74,724,157	44,050,595.00	58.95	0.00
A12Dhd	NA	R1	66,412,067	39,937,379.00	60.14	0.00
A12Dhd	NA	R2	66,412,067	39,937,379.00	60.14	0.00
F1	NA	R1	50,432,876	33,986,936.00	67.39	0.00
F1	NA	R2	50,432,876	33,986,936.00	67.39	0.00
GDDH33	NA	R1	85,862,308	44,717,675.00	52.08	0.00
GDDH33	NA	R2	85,862,308	44,717,675.00	52.08	0.00

Table A.4: Table displaying the raw read numbers and the number after processing of the reads and alignment to the reference genomes. BS-Seq libraries - New batch.

Line	Generation	Pair	Raw #	Mapped /		non-con
				Deduplicated #	Mapped %	
2069	1	R1	68,036,443	33,800,390	49.68	0.01
2069	1	R2	68,036,443	33,800,390	49.68	0.01
2069	3	R1	65,499,786	38,183,883	58.30	0.00
2069	3	R2	65,499,786	38,183,883	58.30	0.00
2134	1	R1	54,452,031	40,777,296	74.89	0.01
2134	1	R2	54,452,031	40,777,296	74.89	0.01
3013	1	R1	70,973,523	42,983,542	60.56	0.00
3013	1	R2	70,973,523	42,983,542	60.56	0.00
3088	1	R1	63,897,139	44,222,778	69.21	0.01
3088	1	R2	63,897,139	44,222,778	69.21	0.01
3088	3	R1	63,250,333	44,227,131	69.92	0.01
3088	3	R2	63,250,333	44,227,131	69.92	0.01
5071	1	R1	64,761,317	44,991,062	69.47	0.01
5071	1	R2	64,761,317	44,991,062	69.47	0.01
5119	1	R1	73,329,928	46,780,007	63.79	0.01
5119	1	R2	73,329,928	46,780,007	63.79	0.01
A12	NA	R1	75,022,317	51,741,117	68.97	0.01
A12	NA	R2	75,022,317	51,741,117	68.97	0.01
GDH33	NA	R1	66,037,976	46,119,242	69.84	0.00
GDH33	NA	R2	66,037,976	46,119,242	69.84	0.00

Appendix B

Appendix for Chapter 5

Table B.1: **Regression statistics.** Relative gene expression change and genome ownership.

A12Dhd inherited dhDEGs

Call:

```
lm(formula = change ~ percentage, data = ownership[ownership$parent ==
"a12", ])
```

Residuals:

```
   Min      1Q  Median      3Q      Max
-2.3800 -1.3814 -0.0531  0.9181  3.6200
```

Coefficients:

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.2958      1.4427   5.750 9.16e-05 ***
percentage  -0.1162      0.0297  -3.911 0.00207 **
```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.932 on 12 degrees of freedom
Multiple R-squared: 0.5603, Adjusted R-squared: 0.5237
F-statistic: 15.29 on 1 and 12 DF, p-value: 0.00207

GDDH33 inherited dhDEGs

Call:

```
lm(formula = change ~ percentage, data = ownership[ownership$parent ==
"gd33", ])
```

Residuals:

```
   Min      1Q  Median      3Q      Max
-3.475 -2.388  0.348  1.261  5.272
```

Coefficients:

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.05523   3.30116   3.652 0.00816 **
percentage  -0.12792   0.06501  -1.968 0.08977 .
```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.971 on 7 degrees of freedom
Multiple R-squared: 0.3562, Adjusted R-squared: 0.2642
F-statistic: 3.873 on 1 and 7 DF, p-value: 0.08977

Table B.2: **Regression statistics.** Relative CG and CHG methylation change and genome ownership.

A12Dhd Inherited DH genome CG	A12Dhd Inherited DH genome CHG
<p>Call: lm(formula = change ~ percentage, data = x[x\$parent == "a12" & x\$context == "cg",])</p> <p>Residuals: Min 1Q Median 3Q Max -19.7119 -1.5772 0.7978 5.3423 8.1828</p> <p>Coefficients: Estimate Std. Error t value Pr(> t) (Intercept) 72.7797 10.6656 6.824 0.000248 *** percentage -0.8573 0.1984 -4.321 0.003475 ** --- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1</p> <p>Residual standard error: 9.066 on 7 degrees of freedom Multiple R-squared: 0.7273, Adjusted R-squared: 0.6884 F-statistic: 18.67 on 1 and 7 DF, p-value: 0.003475</p>	<p>Call: lm(formula = change ~ percentage, data = x[x\$parent == "a12" & x\$context == "chg",])</p> <p>Residuals: Min 1Q Median 3Q Max -44.725 -6.322 6.441 13.286 27.623</p> <p>Coefficients: Estimate Std. Error t value Pr(> t) (Intercept) 226.555 29.028 7.805 0.000107 *** percentage -2.949 0.540 -5.462 0.000944 *** --- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1</p> <p>Residual standard error: 24.67 on 7 degrees of freedom Multiple R-squared: 0.8099, Adjusted R-squared: 0.7828 F-statistic: 29.83 on 1 and 7 DF, p-value: 0.000944</p>
GDDH33 Inherited DH genome CG	GDDH33 Inherited DH genome CHG
<p>Call: lm(formula = change ~ percentage, data = x[x\$parent == "gd33" & x\$context == "cg",])</p> <p>Residuals: Min 1Q Median 3Q Max -13.206 -8.647 -1.878 7.844 19.033</p> <p>Coefficients: Estimate Std. Error t value Pr(> t) (Intercept) 68.0420 12.9859 5.24 0.0012 ** percentage -0.6470 0.2557 -2.53 0.0392 * --- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1</p> <p>Residual standard error: 11.69 on 7 degrees of freedom Multiple R-squared: 0.4777, Adjusted R-squared: 0.4031 F-statistic: 6.402 on 1 and 7 DF, p-value: 0.03922</p>	<p>Call: lm(formula = change ~ percentage, data = x[x\$parent == "gd33" & x\$context == "chg",])</p> <p>Residuals: Min 1Q Median 3Q Max -42.774 -35.182 6.306 12.361 86.200</p> <p>Coefficients: Estimate Std. Error t value Pr(> t) (Intercept) 100.7463 48.9663 2.057 0.0787 . percentage -0.5033 0.9664 -0.521 0.6186 --- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1</p> <p>Residual standard error: 44.09 on 7 degrees of freedom Multiple R-squared: 0.0373, Adjusted R-squared: -0.1002 F-statistic: 0.2712 on 1 and 7 DF, p-value: 0.6186</p>

Table B.3: **Regression statistics.** Relative CHH methylation change and genome ownership.

CHH				
Call: lm(formula = change ~ percentage, data = x[x\$parent == "a12" & x\$context == "chh",])				
Residuals: Min 1Q Median 3Q Max -30.452 -24.135 4.316 21.609 39.095				
Coefficients: Estimate Std. Error t value Pr(> t) (Intercept) 124.9946 33.7757 3.701 0.00765 ** percentage -0.6964 0.6283 -1.108 0.30428 --- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				
Residual standard error: 28.71 on 7 degrees of freedom Multiple R-squared: 0.1493, Adjusted R-squared: 0.0278 F-statistic: 1.229 on 1 and 7 DF, p-value: 0.3043				
CHH				
Call: lm(formula = change ~ percentage, data = x[x\$parent == "gd33" & x\$context == "chh",])				
Residuals: Min 1Q Median 3Q Max -33.881 -27.718 -0.765 28.031 31.511				
Coefficients: Estimate Std. Error t value Pr(> t) (Intercept) 109.5198 31.8044 3.444 0.0108 * percentage -0.5600 0.6277 -0.892 0.4019 --- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				
Residual standard error: 28.64 on 7 degrees of freedom Multiple R-squared: 0.1021, Adjusted R-squared: -0.02617 F-statistic: 0.796 on 1 and 7 DF, p-value: 0.4019				