Refining Bulk Segregant Analyses: Ontology-Mediated Discovery of Flowering Time Genes in *Brassica oleracea*

Rutger A. Vos[1,2,*], Karin van Veen[2], M. Eric Schranz[3], Klaas Vrieling[2], Peter G. L. Klinkhamer[2], Frederic Lens[1,2]

1. Naturalis Biodiversity Center, Leiden, The Netherlands
2. Institute of Biology Leiden, Leiden University, Leiden, The Netherlands
3. Biosystematics Group, Wageningen University and Research, Wageningen, The Netherlands

[*]denotes corresponding author

ABSTRACT

Background

Bulk segregant analysis (BSA) can help identify quantitative trait loci (QTLs), but this may result in substantial bycatch of functionally irrelevant genes. Here we develop a Gene Ontology-mediated approach to zoom in on specific markers implicated in flowering time from among QTLs identified by BSA of the giant woody Jersey kale phenotyped in four bulks of flowering onset.

Results

Our BSA yielded tens of thousands of candidate genes. We reduced this by two orders of magnitude by focusing on genes annotated with terms contained within relevant subgraphs of the Gene Ontology. A further enrichment test led to the pathway for circadian rhythm in plants. The genes that enriched this pathway are attested from previous research as regulating flowering time. Some of these genes were also identified as having functionally significant variation compared to *Arabidopsis*

Conclusions

We validated and confirmed our ontology-mediated results through a more targeted, homology-based approach. However, our ontology-mediated approach produced additional genes of putative importance, showing that the approach aids in exploration and discovery. We view our method as potentially applicable to the study of other complex traits and therefore make our workflows available as open-source code and a reusable Docker container.

Keywords: Giant woody Jersey kale; Bulk segregant analysis; Quantitative trait locus; Flowering time; *Brassica oleracea*; Gene Ontology; Pathway analysis; Singular enrichment analysis; Graph topology; SNP effects

BACKGROUND

Identifying the genes that underlie quantitative trait variation is one of the main challenges in genetics and, to the extent that this is attainable *in silico*, in bioinformatics. One appealingly straightforward approach to discovering candidate loci involved in quantitative trait differences is to sort individuals of a segregating, crossed population into 'bulks' defined by extremes in trait values and then interrogating the genetic contrasts between these bulks, i.e. bulk segregant analysis (BSA [1, 2]). High-throughput sequencing of pooled DNA of bulks has made it possible to quickly generate haystacks of data at low cost, within which are the genetic needles (genomic regions, specific genes, and finally SNPs) that caused the salient differences between the bulks.

Several statistics have been developed to aid in the discovery of candidates of these needles. For each SNP in a sequenced bulk, metrics exist that express its relative coverage compared to other bulks (the Δ(SNP-index) *sensu* Takagi et al. [3]) or whether its allele frequency deviates from the expectation (the modified G' statistic of Magwene et al. [4]). Then, having defined a threshold value for the metric and using a sliding window approach, regions of (more or less) contiguous SNPs in whose metric values the bulks differ can be found, resulting in putative quantitative trait loci (QTLs) in the form of genomic regions. If the analysis is performed using a sufficiently annotated reference genome to map against, genes and SNPs that regulate the trait and intersect with the intervals should be directly locatable. However, this general approach is somewhat imprecise (and more so with low thresholds or large window sizes), resulting in a lot of 'bycatch' of irrelevant genes. Here, we present an approach to remove such bycatch and obtain more refined result sets by traversing and pruning subgraphs of Gene Ontology [5] annotations and KEGG pathways [6] enriched by the

initial QTL finding. We apply and validate this approach using bulks determined by contrasting flowering time in a *Brassica oleracea* cross.

B. oleracea (family Brassicaceae, tribe Brassiceae) is a diploid plant species that gave rise to many vegetable crops rich in essential nutrients such as vitamin C and anti-carcinogens such as sulforaphane [7]. The human-induced selection for diverse edible parts of *B. oleracea* has resulted in a remarkable phenotypic diversity that exceeds the variation in most wild species. Based on the specialized morphology of the edible structures, the species is often classified into kales (leaves), cabbages (leaves), cauliflower (inflorescences), broccoli (inflorescences), Brussels sprouts (axillary buds), and kohlrabi (stems and leaves). Furthermore, *B. oleracea* (n = 9), together with its close relative *B. rapa* (n = 10), are the parental species that gave rise to *B. napus* (n = 19), arguably the most economically valuable species of the tribe Brassicaceae (canola or rapeseed).

Research in *Brassica* has been advanced by the release of genomes from various species (e.g. [8–10]), including two reference genomes from *B. oleracea*: the rapid cycling line TO1000DH3 [11] and *B. oleracea* var. *capitata* [12]. All the *Brassica* genomes show evidence of a Brassicaceae lineage-specific whole-genome triplication since their divergence from the *Arabidopsis* lineage [13], followed by diploidization that involved substantial genome reshuffling and gene losses [8, 14, 15]. This makes *Brassica* an ideal model to study polyploid genome evolution, duplicated gene loss mechanisms, neo- and subfunctionalization, and its associated impact on the vast intraspecific morphological diversity [16].

A critical agronomic trait that shows remarkable variation in *B. oleracea* is flowering time. For example, whereas broccoli is a short-lived annual that flowers in the year it was planted,

cabbage is biannual, needing a cold period to induce flowering (i.e. vernalization) [17]. The genetics that underlies variation in flowering time within and among *Brassica* species is reasonably well characterized [18]. The FLOWERING LOCUS T (FT) locus, its transcriptional repressor FLOWERING LOCUS C (FLC), and its transcriptional activator CONSTANS (CO) all play a central role both in *B. oleracea*, *B. rapa* and in *Arabidopsis*. However, the way FT expression is modulated differs between *Brassica* and *Arabidopsis*. The overall flowering time pathway is much more complex in all cases, involving over two dozen other genes in multiple, divergent copies scattered across the genome [18]. As the exact locations of these copies are mostly known, sufficient background information is available to validate and interpret the results of the analysis we present here and assess its potential for applicability in less well-characterized traits. To provide further context for our results, we also resequenced the novel genome of the late-flowering, heterozygous, giant woody walking stick kale native to Jersey Island (cultivar *B. oleracea* convar. *acephala* var. *viridis*), one of the two parents of the BSA crosses (the other being the rapid cycling, homozygous line TO1000DH3, which has been sequenced before [11]). For this Jersey kale genome, we assess the impact of SNPs within known flowering time pathway genes and compare and contrast these with the genes discovered through our BSA analysis.

METHODS

Plant material, crosses, genotyping and phenotyping

We crossed the homozygous doubled haploid *B. oleracea* kale-like alboglabra line TO1000DH3 [11] with the giant woody walking stick kale (*B. oleracea* convar. *acephala* var. *viridis*) native to Jersey (Channel Islands, UK [19, 20]), the latter grown from seeds ordered from Mr and Mrs Johnson, who own a company making artisanal walking sticks (Homestill, La Grande Route de St. Jean, St. Helier, Jersey, Channel Islands). We selected TO1000DH3

for its rapid flowering time and short generation time (approx. 65 days). In contrast, the Jersey kale is extremely late flowering, has a much longer generation time (at least six months), and requires a vernalization period. We crossed the two parents reciprocally, resulting in F1 seeds from both parents, which we established in tissue culture and potted in soil. We genotyped this F1 population with an allele-specific assay (KASP) on our in-house high throughput SNP genotyping platform and phenotyped the plants on time till first flowering based on two individuals per genotype, distinguishing early (EF), intermediate (IF), late (LF) and non-flowering (NF, at time of DNA extraction). We set the boundaries between these different cohorts such that we obtained bulks of roughly equal numbers of individuals (around ten per bulk; Fig. 1).

DNA extraction and sequencing data pre-processing

We performed genomic DNA extractions on a KingFisher Flex magnetic particle processor robot (Thermo Scientific) using a NucleoMag® 96 Plant kit (Macherey-Nagel GmbH & Co.). We used a volume of 150 µl for elution. We measured DNA concentrations on a Dropsense (TRINEAN NV) using a DropPlate 96-S. Based on these measurements, we pooled the DNAs of the same phenotype (EF, IF, LF and NF) equimolarly to create four DNA bulks for sequencing. We prepared libraries according to the protocol of Macrogen, containing random fragmentation of the DNA sample followed by 5' and 3' adapter ligation, amplification of the adapter-ligated fragments using unique index primers and gel purification. From this, we sent 400ng DNA aliquot to Macrogen for paired-end sequencing on the Illumina HiSeq X platform (read length 150bp) on a shared run. We used the BWA-MEM [21] and SAMtools [22] toolchain to map each bulk's reads against the *B. oleracea* TO1000DH3 reference genome v2.1 of EnsemblPlants release 39, which we filtered so that we mapped against chromosomes

only. We then used GATK HaplotypeCaller [23, 24] for variant (i.e. SNP and indel) calling, yielding the results summarized in Table 1.

Bulk genotyping and QTL region analysis

Given that we phenotyped the F1s by flowering time binned in four bulks, there are six pairs of contrasts (i.e. EF ↔ IF, EF ↔ LF, EF ↔ NF; IF ↔ LF, IF ↔ NF; and LF ↔ NF). We performed joint genotyping for these contrasts using the GATK CombineGVCFs/GenotypeGVCFs workflow. We then filtered these genotypes further, excluding low coverage per sample (<40x), low coverage across the pair of merged samples (<100x), unusually high coverage (>400x, e.g. repeats), low values for the GATK Genotype Quality score (<99), and low values for the frequency of the reference allele (<0.2, a conservative value as TO1000DH3 is homozygous). We calculated smoothed G statistics (G', see [4]) over a sliding window 1Mb wide, filtering outliers by Δ(SNP-index) [3] and retaining all SNPs with G'>2.5 for further analysis. We then performed a QTL-seq analysis [3] to identify candidate QTL regions by simulation using 10k replicates and a two-sided 95% confidence interval (see Fig. 2). For the G' and QTL-seq calculations and simulations, we used the R package QTLseqr [25]. Based on our inferred QTL regions and smoothed G' values, we scanned the mapped assembly of each bulk for genes that fall within QTL regions and have non-synonymous SNPs with high G'. Gene coordinates were based on the annotation of the TO1000DH3 (i.e. the *B. oleracea* GFF3 release v2.1.39 of EnsemblPlants; [11]). To cross-reference the products of these genes with other information resources, we then mapped the *B. oleracea* genes to the curated and machine-predicted proteomics identifiers of UniProtKB/TrEMBL [26] using BioMart [27].

Functional enrichment, refinement, and pathway analysis

We performed singular enrichment analyses (SEA, [28]) separately for all six contrasts using the agriGO web service [29], which uses the Blast2GO [30] results for *B. oleracea* compiled by the Blast2GO Functional Annotation Repository (B2G-FAR, [31]) to establish a reference list against which to assess term enrichment by way of a hypergeometric test corrected for multiple comparisons using the Benjamini–Yekutieli method [32]. To determine the overlap between our SEAs, we merged their results in a cross-comparison (SEACOMPARE, [29]), which showed congruence in enriching numerous terms related to reproduction across all contrasts. For each of the SEA result sets, we pruned the enriched (FDR<0.05) subgraph by retaining only those terms that are reproductive developmental processes, i.e. that are subtended by the upper-level term *developmental process involved in reproduction* (GO:0003006) from the domain *biological process* of the Gene Ontology [5]. Within the pruned subgraph (Fig. 3), three out of the top-level terms are related to flower development or morphogenesis, one to seed maturation, and one (GO:0010228) is defined as:

"The process involved in transforming a meristem that produces vegetative structures, such as leaves, into a meristem that produces reproductive structures, such as a flower or an inflorescence."

As this developmental process precedes those defined by the other top-level terms in the subgraph, we took the list of genes annotated to these terms. We used this as the input for a pathway enrichment analysis as implemented in g:Profiler [33]. This yielded an alternative view in the extent to which the genes enrich other GO terms, as well as any KEGG [6] pathways.

Genome analysis of the Jersey kale

To gain more background insight into the genome of the giant woody Jersey kale as a potential model in general, and with an eye on differences with TO1000DH3 in flowering time loci in particular, we also sequenced the genome of a specimen of this cultivar. We followed the same protocols for DNA extraction, sequencing, genome assembly, and variant calling described for the bulks in the section *DNA extraction and sequencing data pre-processing*. However, as there was no pooling of multiple individuals (i.e. no BSA), the coverage for this single individual was commensurately higher (approx. 100x coverage). For this Jersey kale genome, we then used SnpEff [34] to assess the impact of SNPs within the loci previously identified as participating in the flowering time pathway of *Arabidopsis* [18].

RESULTS

The F1 seeds from our crosses resulted in 42 distinct genotypes that we successfully established in tissue culture and transferred in soil. We found a 420-day time lag between the earliest and the latest flowering F1 genotypes, presumably owing to the heterozygosity of the Jersey kale parent: the first two F1 genotypes (genotype numbers 17135, 17136) started to flower 113 days after potting, while more than a year later, at day 533, the last F1 flowered (genotype number 17109). The first bulk, early flowering EF, comprised 11 F1 genotypes that flowered between 113-135 days after potting. The second bulk, flowering at intermediate age IF, included eight F1s that flowered from 154-164 days after potting. The third bulk, late-flowering LF, represented 11 F1 genotypes that started to flower from 176-239 days after potting. The fourth bulk, non-flowering (NF) at the time of DNA extraction (day 294), included 9 F1s that only flowered after DNA extraction, up to 533 days after potting. Phenotyping results are summarized in Fig. 1 and detailed in Supplementary Table 1.

Sequencing resulted in yields per bulk between 54.3×10^9 and 60.3×10^9 bases, observed in 359.6×10^6 through 399.4×10^6 reads (which are 150bp on HiSeq X). Given the size of the reference genome (approx. 488Mbp [11]), this corresponds with a depth in the range of 111x-123x per bulk, or about 10x per individual. We retained most of the estimated raw coverage in the assemblies, yielding an average mapped coverage of 99x-108x (more sequencing statistics are listed in Supplementary Table 2). Following variant calling and joint genotyping of the six pairwise comparisons of phenotype bulks, our G' sliding window analysis produced fairly consistent results across all comparisons. We found regions with windows of G'>2.5 on all chromosomes but C1. Still, regions that featured in the majority of pairwise comparisons were restricted to the q arm of C3 (spanning, for example, the locus of one of the CO copies), C6 (spanning one of the FT loci), C7 (spanning a FUL copy) and C8 (spanning a PHYB copy). A comparative view of these G' regions is shown in Fig. 2. At this stage of the analysis, the QTL regions intersected with the genomic coordinates of 14,257 genes, out of which 10,469 had non-synonymous SNPs.

The term enrichment analyses (SEA, [28]) yielded between 812 and 1409 significantly enriched terms for the six bulk contrasts. Interestingly, the number of terms appears to covary somewhat with the magnitude of the trait differences, in that contrasts between the contiguous cohorts EF ↔ IF and LF ↔ NF enriched the lowest numbers of terms (1086 and 812, respectively) while those between the non-contiguous cohorts IF ↔ NF and EF ↔ NF returned the most terms (1409 and 1280). Nevertheless, the different analyses' results overlapped extensively, as the total number of distinct terms returned overall was 1544. This was confirmed by SEACOMPARE ([29]), which also indicated extensive overlap across the six comparisons (shown in Supplementary Table 3).

Each of the comparisons enriched a subgraph of the GO topology, which we pruned further to retain only those parts subtended by GO:0003006 (*developmental process involved in reproduction*). Across the comparisons, this resulted in a consensus subgraph that spanned 14 enriched terms, shown in Fig. 3. Among these 14 terms were five upper-level terms that have an implied ordering in time and specificity with respect to their contribution to the onset and further development of flowering:

1. GO:0010228 - *vegetative to reproductive phase transition of meristem*
2. GO:0048439 - *flower morphogenesis*
3. GO:0048437 - *floral organ development*
4. GO:0048438 - *floral whorl development*
5. GO:0010431 - *seed maturation*

Terms 2-5 can only start to play a role once the transition of meristem from vegetative to reproductive has commenced (and likewise, seed maturation can only happen in a fully developed flower). Hence, we then zoomed in on only those genes that are (transitively) annotated with GO:0010228 and used these as input for g:Profiler [33], whose results are shown in Fig. 4.

The g:Profiler results show enrichment both for terms above GO:0010228 and those below it. The terms shown with gray background in Fig. 4 are inevitably enriched because we restricted the input gene list to those whose annotations descend from GO:0010228 – and therefore also from all upper terms 'above' it. More interesting are the terms below it, some of which are more specific and shed light on what, according to these annotation sets, triggers the phase transition: light and photoperiodism. However, this analysis step's most salient result is the discovery of KEGG [6] pathway 04712 *Circadian rhythm – plant*, based on the presence of

Bo4G006930 (CIRCADIAN CLOCK ASSOCIATED 1, CCA1), Bo6G099320 (FLOWERING TIME, FT) and Bo6G120900 (TWIN SISTER OF FT, TSF).

Genome sequencing of the Jersey kale yielded 1,092,319,676 forward and reverse reads (150bp). Mapped against the TO1000DH3, this resulted in an assembly covering 87.6% of the reference genome with an average depth of 170.6x. Variant calling on this assembly produced a total of about 7.5×10^6 raw variants of all types (i.e. including indels and polymorphisms longer than 1bp). The SnpEff analysis, which assessed the impact of variants in *Arabidopsis* flowering time pathway homologs, returned results of comparable magnitude as obtained in previous research in the *B. oleracea* cultivar 'Kashirka' [18]. For example, most gene copies were affected by, at least, non-synonymous SNPs, and much fewer of those by splicing variation or indels causing frameshifts. Similarly to 'Kashirka', one copy of FT and one of FLC had indels, while the CO copies had non-synonymous SNPs but no indels. The detailed results of the analysis are available in Supplementary Table 4.

DISCUSSION

Our analysis was somewhat complicated by the proliferation of flowering time cohorts (four bulks, where BSA studies typically consider only the two extremes of the trait value distribution) and the commensurate increase in bulk contrasts to consider ($n(n-1)/2$ working out to six contrasts for four bulks). Another complication was the right censoring in waiting time till flowering for bulk NF ("non-flowering"): as three plants in this bulk flowered after the point of DNA extraction while others never did, the bulk is a mixture of very late flowering and non-flowering genotypes. Nevertheless, our results showed consistency across all comparisons in the discovered QTL regions and the GO terms the genes in these regions enriched.

A modestly interesting result was that the extremes in the numbers of enriched GO terms loosely corresponded with the magnitude of the difference in trait values between bulks in a comparison: greater differences in flowering time enriched more terms, smaller differences fewer. For our present purposes, the increase in the number of enriched terms with greater trait differences constituted a loss in precision: as bulks are more different in flowering time, differences in the onset of contingent developmental processes (e.g. seed maturation) have a chance to manifest as well, clouding the picture and yielding more GO terms.

Our key finding was that the iterative pruning of the enriched GO subgraphs substantially reduced the number of candidate QTLs and genes in the result set. We first focused on the upper-level term GO:0003006 (*developmental process involved in reproduction*) and then zoomed in further on nodes subtended by its descendant GO:0010228 (*vegetative to reproductive phase transition of meristem*). This progression was discovered from the data and should therefore be transferrable to other systems without prior knowledge of the underlying genetics or annotations. As a result of this approach, we reduced the number of candidate genes from 10,469 to a final set of 29 genes resulting from the g:Profiler analysis. Considering that these genes include those previously established as key in regulating flowering time (both in their being homologous to those in *Arabidopsis* and in their variation between *Brassica* cultivars [18]), we view our approach as a powerful complement to existing workflows in processing BSA results.

A potential weakness is that the technique's usefulness hinges on the quality of genome annotations and KEGG pathways: without the combination of good functional characterization (often inferred from homology) and a known background against which to

perform the hypergeometric tests, gene set enrichment analyses cannot work. In practice, this means that the technique will be most applicable to well-studied model organisms or reasonably close relatives of *Arabidopsis*.

Previous research in flowering time in *B. oleracea* cultivars used the pathway in *Arabidopsis* as the backbone on which to map participating gene copies and their variants [18], making different cultivars and species more easily comparable. As a result, we established similar patterns of variation in the Jersey kale genome as previously have been found in the 'Kashirka' cultivar. However, with this approach, the importance of additional genes outside the homologous pathway is never discovered. In contrast, our approach also uncovered the role of Bo6G120900 (TWIN SISTER OF FT, TSF). As such, the technique we present here is at least complementary, and especially useful in exploring less well-characterized pathways and discovering participating genes.

CONCLUSIONS

We performed a bulk segregant analysis (BSA) across four cohorts of a cross between the Jersey kale and the *B. oleracea* model TO1000DH3 phenotyped on flowering time. The data we collected consisted of high throughput sequencing reads, which we analyzed using standard tools for identifying QTLs in pairwise BSA comparisons. This resulted in numerous regions throughout the genome, though concentrated at loci known from previous research in flowering time regulation and consistent across pairwise comparisons. To reduce the set of candidate loci to more manageable dimensions, we developed an approach that limits the result set by focusing on genes annotated with terms contained within relevant subgraphs of the Gene Ontology. This reduced the resulting gene set from tens of thousands to dozens of candidate genes. A further enrichment test led to the pathway for circadian rhythm in plants.

The genes that enriched this pathway are attested from previous research as being involved in regulating flowering time, and some of these genes were also identified as having functionally significant variation compared to *Arabidopsis*. As such, we validated and confirmed our ontology-mediated results through a more targeted, homology-based approach. However, the ontology-mediated approach produced additional genes of putative importance, showing that the approach aids in exploration and discovery. We view our method as potentially applicable to the study of other complex traits and therefore make our workflows available as open-source code and a reusable Docker container.

Supplementary Materials

The supplementary files referred to in the manuscript body are contained in an archive that additionally holds an extended methods section with links to all other outputs of this project. The archive is located at 10.5281/zenodo.5182315

Acknowledgements

The authors are grateful for the institutional support and funding provided to this project by Naturalis Biodiversity Center through internal project grant 077 LENS and the Institute of Biology Leiden of Leiden University. Lab support for DNA sequencing was provided by Elza Duijm of Naturalis. Some of the SnpEff analyses were performed by Esther Kockelmans, Rik Frijmann, Nino Vrolijk and Daphne van Ginneken of the University of Applied Sciences Leiden.

REFERENCES

1. Michelmore RW, Paran I, Kesseli R V. Identification of markers linked to disease-resistance genes by bulked segregant analysis: A rapid method to detect markers in specific genomic regions by using segregating populations. Proc Natl Acad Sci U S A. 1991;88:9828–32. doi:10.1073/pnas.88.21.9828.

2. Giovannoni JJ, Wing RA, Ganal MW, Tanksley SD. Isolation of molecular markers from specific chromosomal intervals using DNA pools from existing mapping populations. Nucleic Acids Res. 1991;19:6553–68. doi:10.1093/nar/19.23.6553.

3. Takagi H, Abe A, Yoshida K, Kosugi S, Natsume S, Mitsuoka C, et al. QTL-seq: rapid mapping of quantitative trait loci in rice by whole genome resequencing of DNA from two bulked populations. Plant J. 2013;74:174–83. doi:10.1111/tpj.12105.

4. Magwene PM, Willis JH, Kelly JK. The statistics of bulk segregant analysis using next generation sequencing. PLoS Comput Biol. 2011;7.

5. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: Tool for the unification of biology. Nature Genetics. 2000;25:25–9.

6. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. The KEGG resource for deciphering the genome. Nucleic Acids Res. 2004;32 DATABASE ISS.:D277–80. doi:10.1093/nar/gkh063.

7. Fisher ML, Ciavattone N, Grun D, Adhikary G, Eckert RL. Sulforaphane reduces YAP/ΔNp63α signaling to reduce cancer stem cell survival and tumor formation. Oncotarget. 2017;8:73407–18. doi:10.18632/oncotarget.20562.

8. Wang X, Wang H, Wang J, Sun R, Wu J, Liu S, et al. The genome of the mesopolyploid crop species Brassica rapa. Nat Genet. 2011;43:1035–40.

9. Chalhoub B, Denoeud F, Liu S, Parkin IAP, Tang H, Wang X, et al. Early allopolyploid evolution in the post-neolithic Brassica napus oilseed genome. Science (80-). 2014;345:950–

1 3.
2 10. Yang J, Liu D, Wang X, Ji C, Cheng F, Liu B, et al. The genome sequence of
3 allopolyploid Brassica juncea and analysis of differential homoeolog gene expression
4 influencing selection. Nat Genet. 2016;48:1225–32. doi:10.1038/ng.3657.
5 11. Parkin IAP, Koh C, Tang H, Robinson SJ, Kagale S, Clarke WE, et al. Transcriptome and
6 methylome profiling reveals relics of genome dominance in the mesopolyploid Brassica
7 oleracea. Genome Biol. 2014;15.
8 12. Liu S, Liu Y, Yang X, Tong C, Edwards D, Parkin IAP, et al. The Brassica oleracea
9 genome reveals the asymmetrical evolution of polyploid genomes. Nat Commun.
10 2014;5:3930. doi:10.1038/ncomms4930.
11 13. Lysak MA, Koch MA, Pecinka A, Schubert I. Chromosome triplication found across the
12 tribe Brassiceae. Genome Res. 2005;15:516–25.
13 14. Mun JH, Kwon SJ, Yang TJ, Seol YJ, Jin M, Kim JA, et al. Genome-wide comparative
14 analysis of the Brassica rapa gene space reveals genome shrinkage and differential loss of
15 duplicated genes after whole genome triplication. Genome Biol. 2009;10.
16 15. Town CD, Cheung F, Maiti R, Crabtree J, Haas BJ, Wortman JR, et al. Comparative
17 genomics of Brassica oleracea and Arabidopsis thaliana reveal gene loss, fragmentation, and
18 dispersal after polyploidy. Plant Cell. 2006;18:1348–59. doi:10.1105/tpc.106.041665.
19 16. Schranz EM, Mohammadin S, Edger PP. Ancient whole genome duplications, novelty and
20 diversification: The WGD Radiation Lag-Time Model. Curr Opin Plant Biol. 2012;15:147–
21 53. doi:10.1016/j.pbi.2012.03.011.
22 17. Osborn TC, Kole C, Parkin IAP, Sharpe AG, Kuiper M, Lydiate DJ, et al. Comparison of
23 flowering time genes in Brassica rapa, B. napus and Arabidopsis thaliana. Genetics.
24 1997;146:1123–9.
25 18. Schiessl S V., Huettel B, Kuehn D, Reinhardt R, Snowdon RJ. Flowering time gene

variation in Brassica species shows evolutionary principles. Front Plant Sci. 2017;8:1742. doi:10.3389/fpls.2017.01742.

19. Parker S, Cox GS. The giant cabbage of the Channel Islands. Guernsey hist Monogr. 1970;10.

20. Prendergast HD V., Rumball N. Walking sticks as seed savers - The case of the Jersey kale [Brassica oleracea L. Convar. Acephala (DC.) Alef. Var. Viridis L.]. Econ Bot. 2000;54:141–3.

21. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 2013. http://arxiv.org/abs/1303.3997. Accessed 3 Sep 2019.

22. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.

23. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20:1297–303.

24. Depristo MA, Banks E, Poplin R, Garimella K V., Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43:491–501.

25. Mansfeld BN, Grumet R. QTLseqr: An R Package for Bulk Segregant Analysis with Next-Generation Sequencing. Plant Genome. 2018;11:180006. doi:10.3835/plantgenome2018.01.0006.

26. Schneider M, Lane L, Boutet E, Lieberherr D, Tognolli M, Bougueleret L, et al. The UniProtKB/Swiss-Prot knowledgebase and its Plant Proteome Annotation Program. J Proteomics. 2009;72:567–73.

27. Smedley D, Haider S, Ballester B, Holland R, London D, Thorisson G, et al. BioMart – biological queries made easy. BMC Genomics. 2009;10:22. doi:10.1186/1471-2164-10-22.

28. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2009;37:1–13. doi:10.1093/nar/gkn923.

29. Du Z, Zhou X, Ling Y, Zhang Z, Su Z. agriGO: A GO analysis toolkit for the agricultural community. Nucleic Acids Res. 2010;38 SUPPL. 2.

30. Conesa A, Götz S. Blast2GO: A comprehensive suite for functional analysis in plant genomics. Int J Plant Genomics. 2008;2008:1–12. doi:10.1155/2008/619832.

31. Götz S, Arnold R, Sebastián-León P, Martín-Rodríguez S, Tischler P, Jehl MA, et al. B2G-FAR, a species-centered GO annotation repository. Bioinformatics. 2011;27:919–24. doi:10.1093/bioinformatics/btr059.

32. Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. Ann Stat. 2001;29:1165–88.

33. Reimand J, Arak T, Adler P, Kolberg L, Reisberg S, Peterson H, et al. g:Profiler—a web server for functional interpretation of gene lists (2016 update). Nucleic Acids Res. 2016;44:W83–9. doi:10.1093/nar/gkw199.

34. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin). 2012;6:80–92. doi:10.4161/fly.19695.

1 **TABLES**

Phenotype	Bulk size	Total read bases (bp)	Total reads	GC (%)	Q20 (%)	Q30 (%)	Coverage a, b	Variants
EF	11	60,303,831,724	399,363,124	36.93	95.02	89.25	123, 108	40,224,519
IF	8	56,804,209,216	376,186,816	36.92	94.84	88.94	116, 103	43,785,856
LF	11	54,587,863,530	361,509,030	37.14	96.00	91.34	112, 100	42,852,937
NF	9	54,296,890,456	359,582,056	37.13	96.84	92.81	111, 99	42,213,427

2

3 **Table 1**. Summary results of the sequencing of bulks of early (EF), intermediate (IF), late

4 (LF) and "non" flowering (NF, actually not flowering at time of DNA extraction) phenotypes.

5 Bulk size refers to the number of individuals pooled for that phenotype. Coverage is given as

6 a) total read bases divided by reference genome size; and b) average mapped coverage.

7

FIGURES

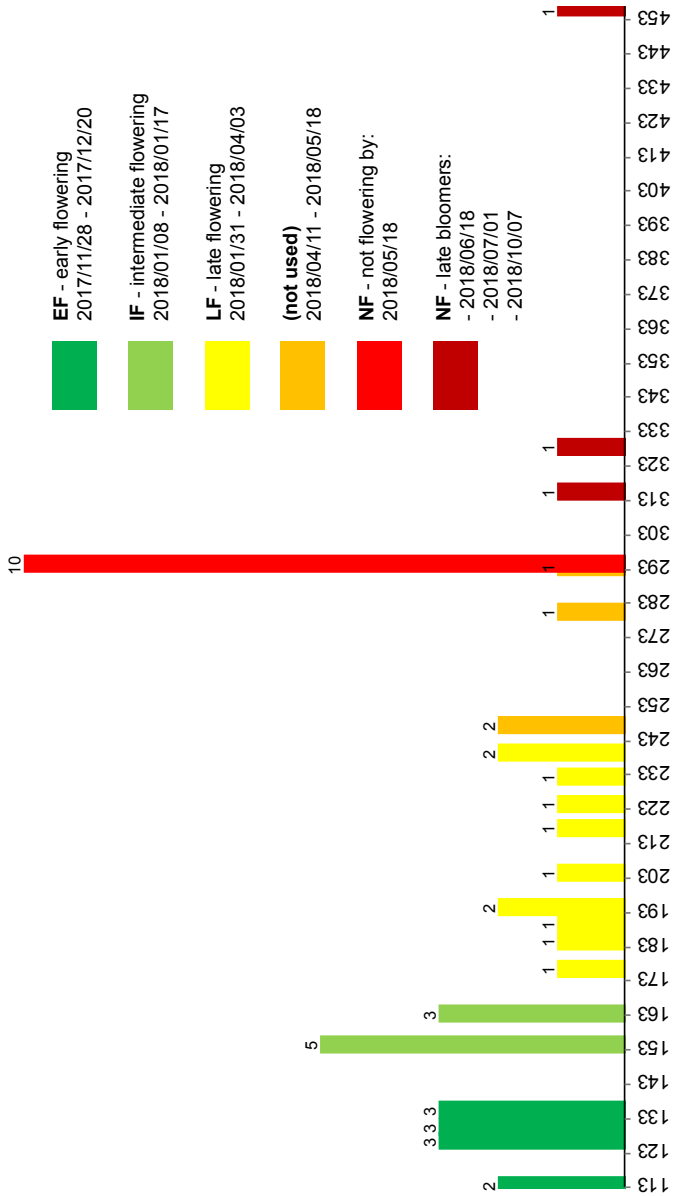

Figure 1 – phenotyping results and assignment to bulks. F1 individuals were assigned to one of four bulks: Early Flowering (EF), Intermediate Flowering (IF), Late Flowering (LF) or Non-Flowering (NF). Of the latter, three individuals (late bloomers) flowered anyway, after DNA extraction.

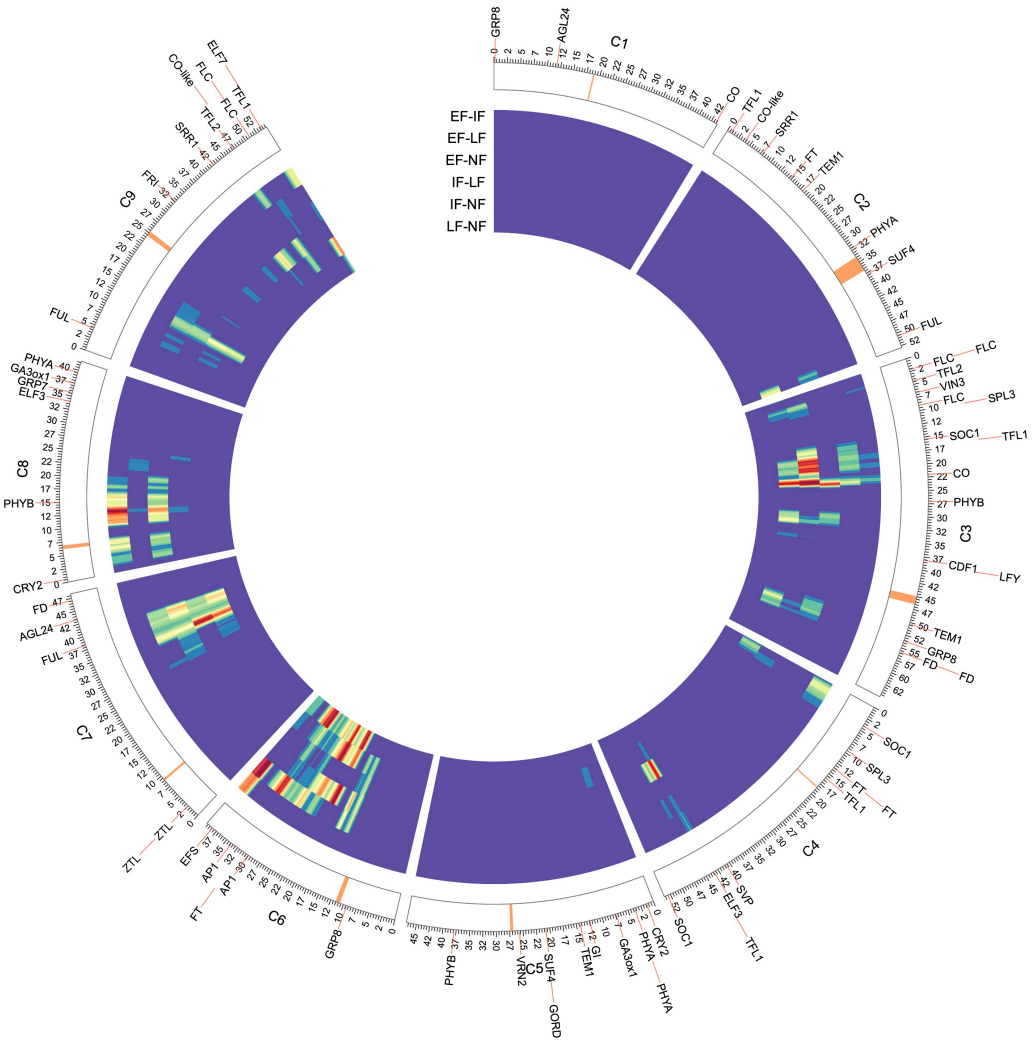

Figure 2 – ‘circos’ plot of QTL regions. Six concentric heatmaps show QTLs identified for the six possible contrasts among the four bulks. The QTLs are mapped on the annotated TO1000DH3 reference, showing the locations of flowering time genes previously identified for *B. oleracea* [18]. Centromeres indicated in orange.

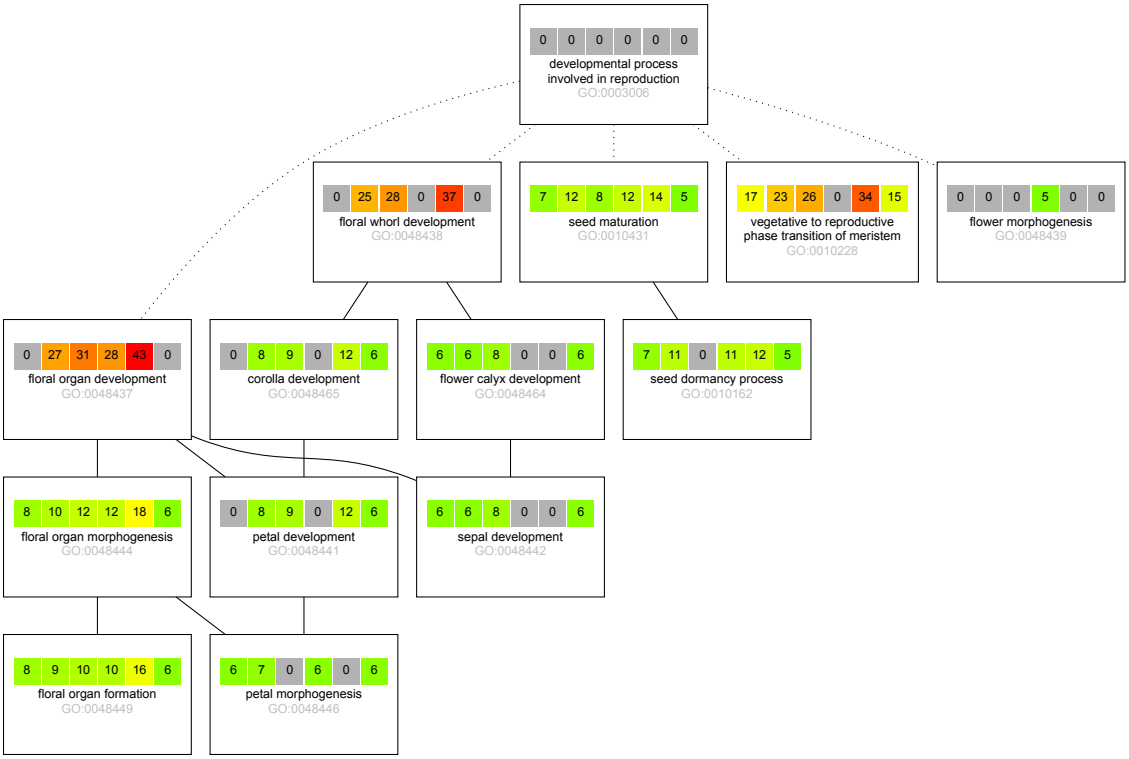

Figure 3 – genes annotated to GO subgraph terms. Each box represents a GO term subtended by GO:0003006. Within each box, the six bulk contrasts for four bulks (i.e. EF ↔ IF, EF ↔ LF, EF ↔ NF; IF ↔ LF, IF ↔ NF; and LF ↔ NF) are shown as squares. Each square shows the number of genes annotated with that term for that contrast. The color coding corresponds with the significance of the enrichment as an inverse heatmap, i.e. the smaller *p*, the hotter the tint.

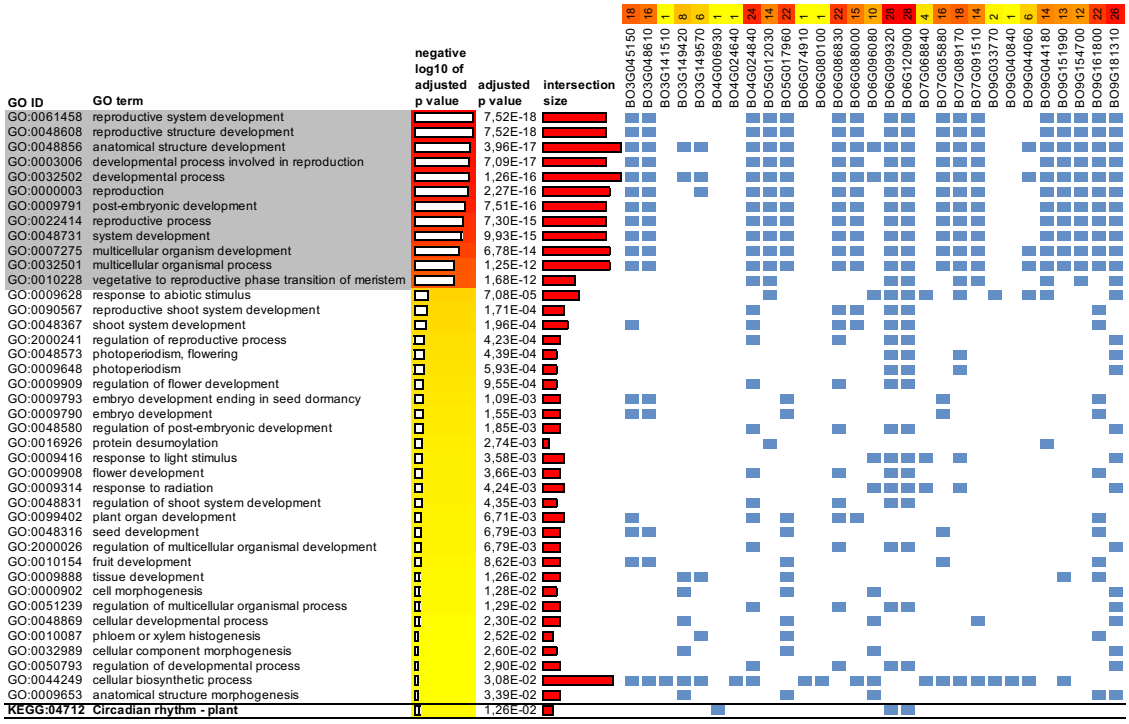

Figure 4 – g:Profiler results for the pruned subgraph. In this analysis, the genes (the columns labeled with BO gene IDs) annotated by GO:0010228 are used as input. In consequence, the GO terms above this term (both in this table and topologically in the GO graph, here shown with gray background) are heavily enriched. Below it are descendant terms that are also significantly enriched. Blue cells indicate which genes contribute to the enrichment. g:Profiler also returns enriched pathways. In this case this is only – but importantly – "KEGG:04712, Circadian rhythm – plant".